Statistical analysis

HOUSING PRICE

Kiarash Hossein Zade

Subject : Statistical Analysis Of Boston Housing Price Dataset

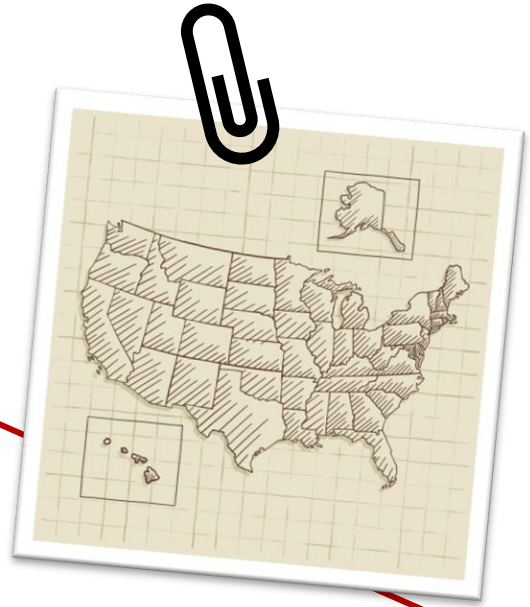
Introduction : This File Includes My Work Of Examining The Boston Housing Dataset By Conducting Statistical Analysis.

contact :

09382960100

LinkedIn Profile

INTRODUCTION:



The Boston housing price dataset is a well-know dataset used for regression analysis.

The dataset contains information on 506 housing units in the Boston area, with 14 attributes or features such as crime rate, number of rooms, and distance to employment centers.
The goal is to predict the median value of owner-occupied homes (MEDV) using these features.

In this project I am going to examine the Boston housing price dataset in terms of statistical analysis and linear algebra.

PROJECT OVERVIEW:

This project exemplifies the practical application of statistical and linear algebraic concepts using Excel. Throughout this program, I used different skills like data manipulation, analysis, and visualization.

THE PROJECT INVOLVES :

. Examining the type of each variable of the dataset

. Examining the distribution and descriptive statistics each variable

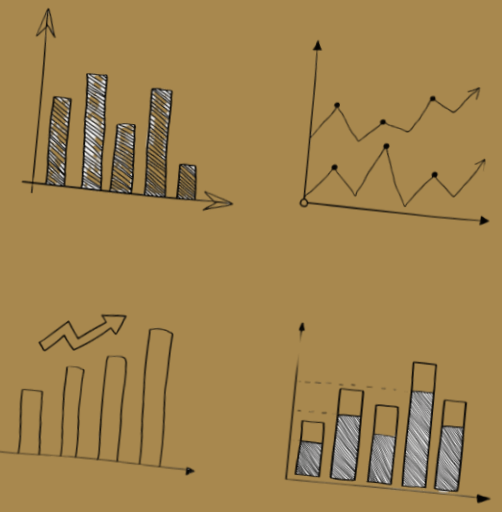
. Conducting statistical tests like (Normality test, Leven's Test, T-student test, Anova test and so on)

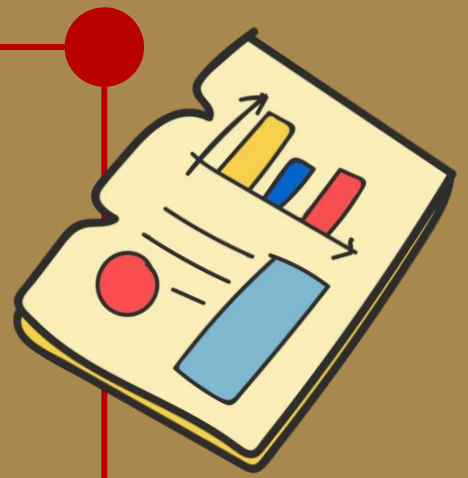
. Finding outliers with box-plot, z-score, and Grubbs methods
. Data transformation (Z-transformation, Box-cox transformation, Normalization)

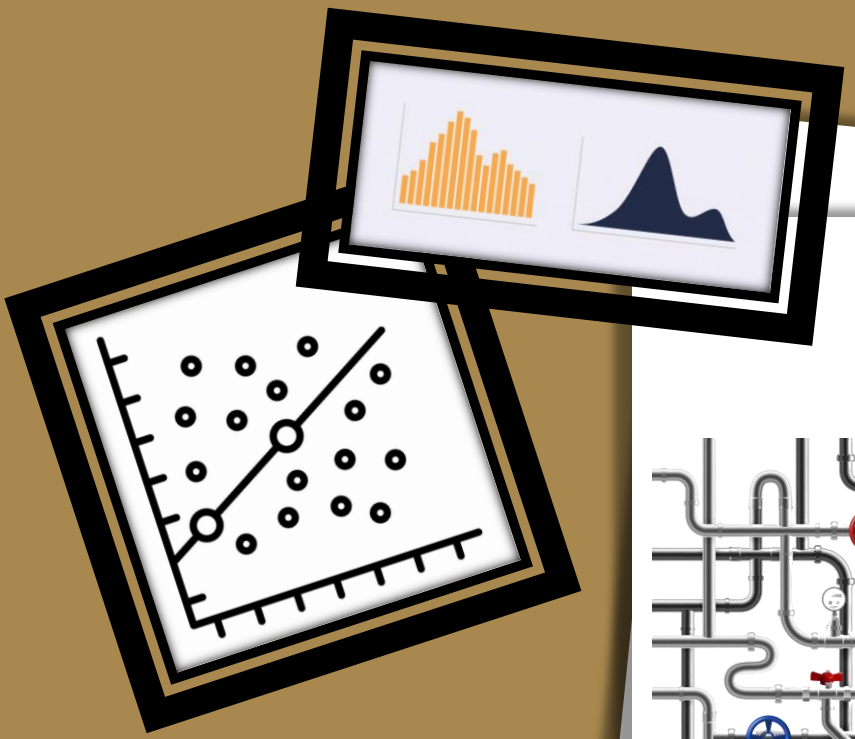
. Modeling the target variable with linear regression method

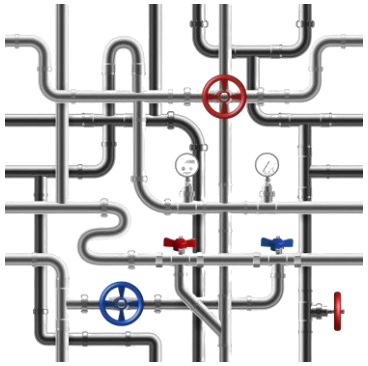
. Examining the correlation between different variable (using Pearson, Spearman, Point-Biserial methods and other methods)

. Equality test of means and variances of two samples









WORKFLOW:

In this section I am going to briefly explain the order and the pipeline of my work to prepare you for what you are going to face with.

PIPELINE:

1st step. Examining the type of each variable
2nd step. Dealing with missing values of each attribute
3rd step. Examining each variable in terms of statistics
- . Examining the distribution of variable
- . Examining the descriptive statistics of variable
- . Conducting normality test on the variable
- . Data transformation
- . Outlier detecting
- . Examining the correlation with target variable
- . Scatter plot of variable with target variable
- . Equality test of means and variances of samples
- . Distribution fitting

4th step. Modeling the target variable with linear regression method

Introduction to Boston Housing Dataset: Key Variables & Their Types

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CRIM	ZN (%)	INDUS (%)	CHAS	NOX (PPM)	RM	AGE (%)	DIS (Miles)	RAD	TAX (10,000$)	PTRATIO	B	LSTAT (%)	MEDV (1000$)
2	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
3	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
4	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
5	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4

Variable	Description	Type	Variable	Description	Type
CRM	Per capita crime rate by town	Quantitative Continuous	**DIS**	Weighted distances to five Boston employment centers	Quantitative Continuous
ZN	Proportion of residential land zoned for lots over 25,000 sq. ft.	Quantitative Continuous	**RAD**	Index of accessibility to radial highways	Qualitative Ordinal
INDUS	Proportion of non-retail business acres per town	Quantitative Continuous	**TAX**	Full-value property tax rate per $10,000	Quantitative Discrete
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)	Quantitative Discrete Binary	**PTRATIO**	Pupil-teacher ratio by town	Quantitative Continuous
NOX	Nitric oxide concentration (parts per 10 million)	Quantitative Continuous	**B**	Proportion of population that is Black	Quantitative Continuous
RM	Average number of rooms per dwelling	Quantitative Continuous	**LSTAT**	Percentage of lower status of the population	Quantitative Continuous
AGE	Proportion of owner-occupied units built before 1940	Quantitative Continuous	**MEDV**	Median value of owner-occupied homes in $1,000	Quantitative Continuous

Data Cleansing

Handling Missing Values Of CRIM (Crime Rate) Feature

	A
1	CRIM
44	0.1415
45	0.15936
46	0.12269
47	0.17142
48	0.18836
49	0.22927
50	0.25387
51	0.21977
52	0.08873
53	0.04337
54	0.0536
55	NA
56	0.0136
57	0.01311
58	0.02055

Detecting Missing values:

As we can see, Crime Rate column contains some missing values that we should firstly detect them, then decide how to deal with them.

	A
1	CRIM
55	NA
117	NA
185	NA
193	NA
194	NA
198	NA
231	NA
238	NA
243	NA
264	NA
290	NA
304	NA
371	NA
395	NA
396	NA
411	NA
428	NA
432	NA
462	NA
467	NA

Count Of Missing Values:

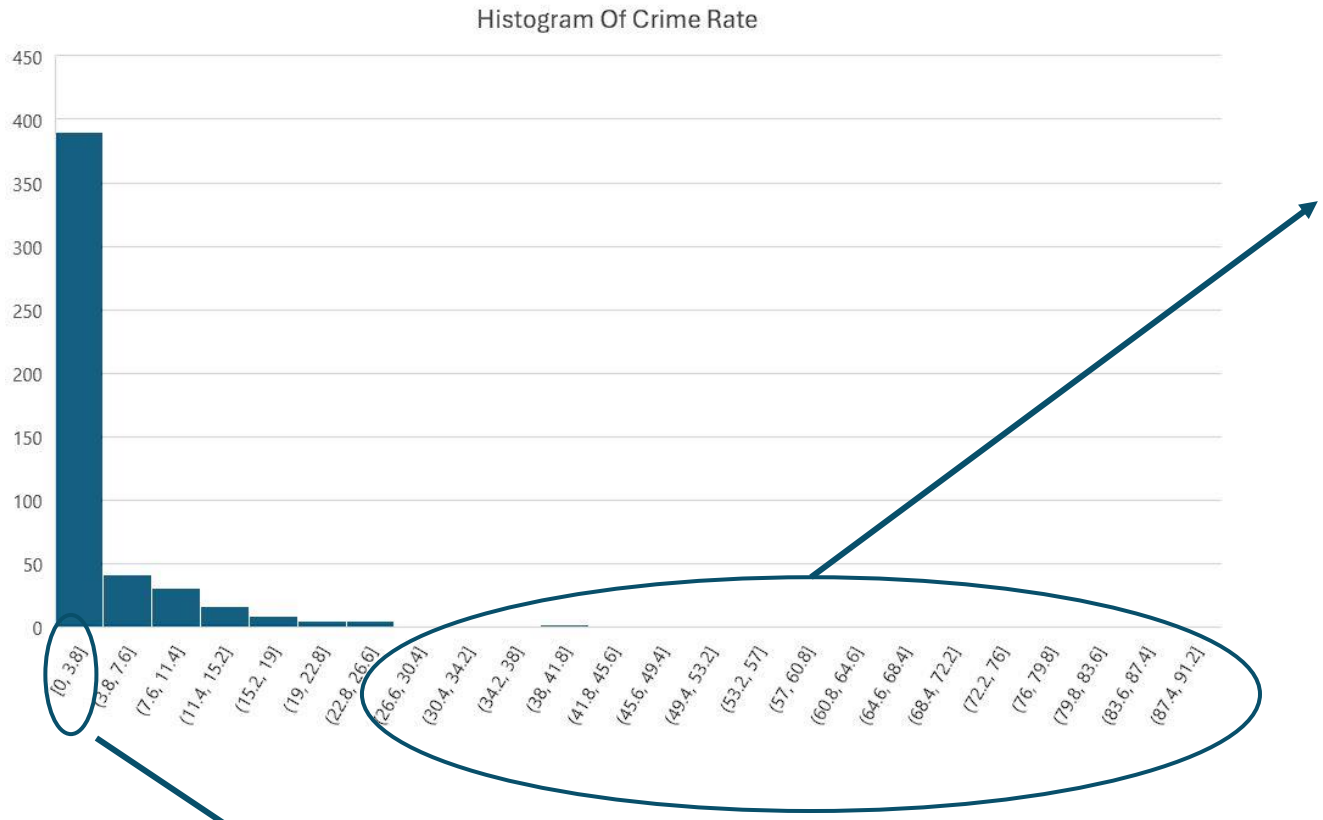
As we can see, there are 20 missing values in Crime Rate column, that would approximately be 4% of our sample.

There are several approaches that we can take to solve this problem

- Replacing it by average of Crime Rate feature
- Replacing it with mode
- Replacing it with median
- Impute the missing values by a model
- Removing missing values

Data Cleansing

Handling Missing Values Of CRIM (Crime Rate) Feature



Replacing Missing Values With Average:

If we replace the missing values of Crime Rate with average, the marked proportion of data will affect the replacement.
So, It does not seem wise if we replace it with average, because the distributions is highly and positively skewed

The Best Approach:

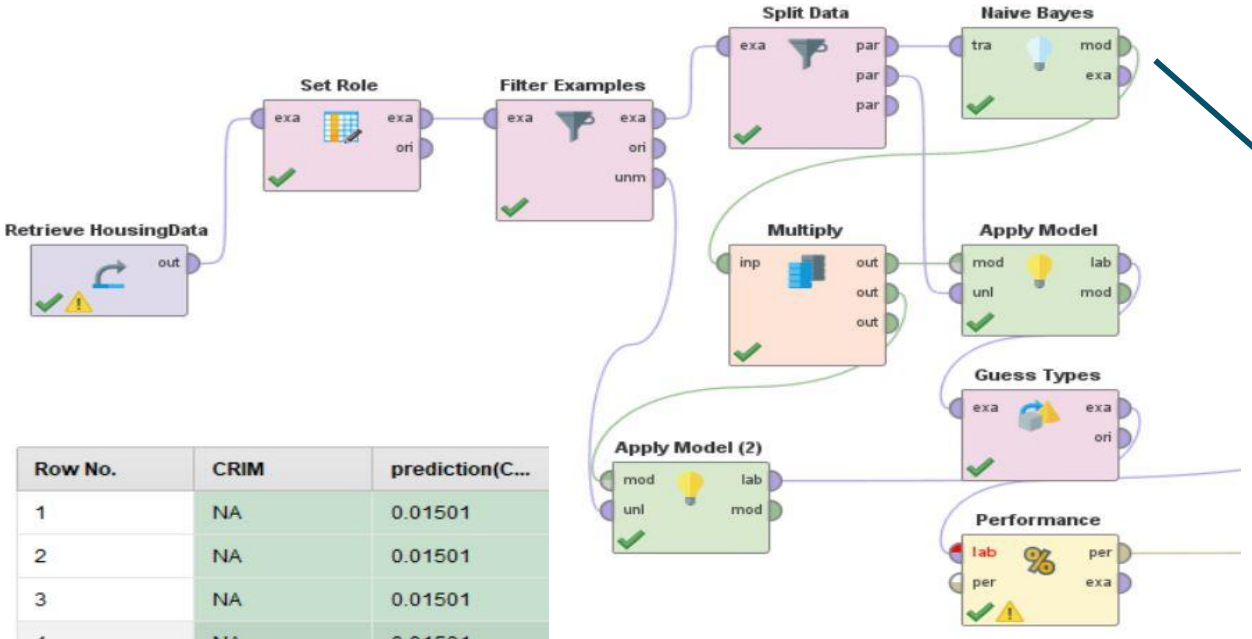
Based on the reasons I have mentioned and by paying attention to the nature of Crime Rate feature, imputing the missing values might be a better approach.
Crime Rate in a area might be influenced by the areas around, and we do not know that those specific areas that we do not have the crime rate for, are near to which one of these areas (those with low crime rate or those with high crime rate) so I believe that imputing by a model would a better solution.

Replacing Missing Values With Mode and Median:

Replacing missing values of this feature with mode or median, seems a better approach than replacing them with average; but If we replace the missing values of this feature with mode or median, we might make a mistake because the Crime Rate in those areas may not follow the majority and by replacing mode or median, we are making them to follow the majority.

Data Cleansing

Handling Missing Values Of CRIM (Crime Rate) Feature



Row No.	CRIM	prediction(C...
1	NA	0.01501
2	NA	0.01501
3	NA	0.01501
4	NA	0.01501
5	NA	0.01501
6	NA	0.01501
7	NA	0.01501
8	NA	0.01501
9	NA	0.01501
10	NA	0.01501
11	NA	0.01501
12	NA	0.01501
13	NA	14.3337
14	NA	14.3337
15	NA	14.3337
16	NA	14.3337
17	NA	14.3337

Imputing missing values with Naive Bayes Algorithm:

I created different models for predicting the missing values of Crime Rete feature and I compared the root mean squared error and other metrics of each of them to find the best one for prediction.
This model gave me the root mean squared error of 4 which was relatively good.

Predictions Of Model:

As we can see, the predicted values for missing values of Crime Rate, is so close to the mode of this feature, so we can replace the missing values by the mode with a good confidence level

average	3.611874
median	0.253715
mode	0.01501

Data Cleansing

Handling Missing Values Of ZN (Zoned) Feature

ZN (%)
0
0
0
0
0
0
0
0
NA
28
28
0
0
0

Detecting Missing values:

As we can see, ZN column contains some missing values that we should firstly detect them, then decide how to deal with them.

ZN (%)
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA

Count Of Missing Values:

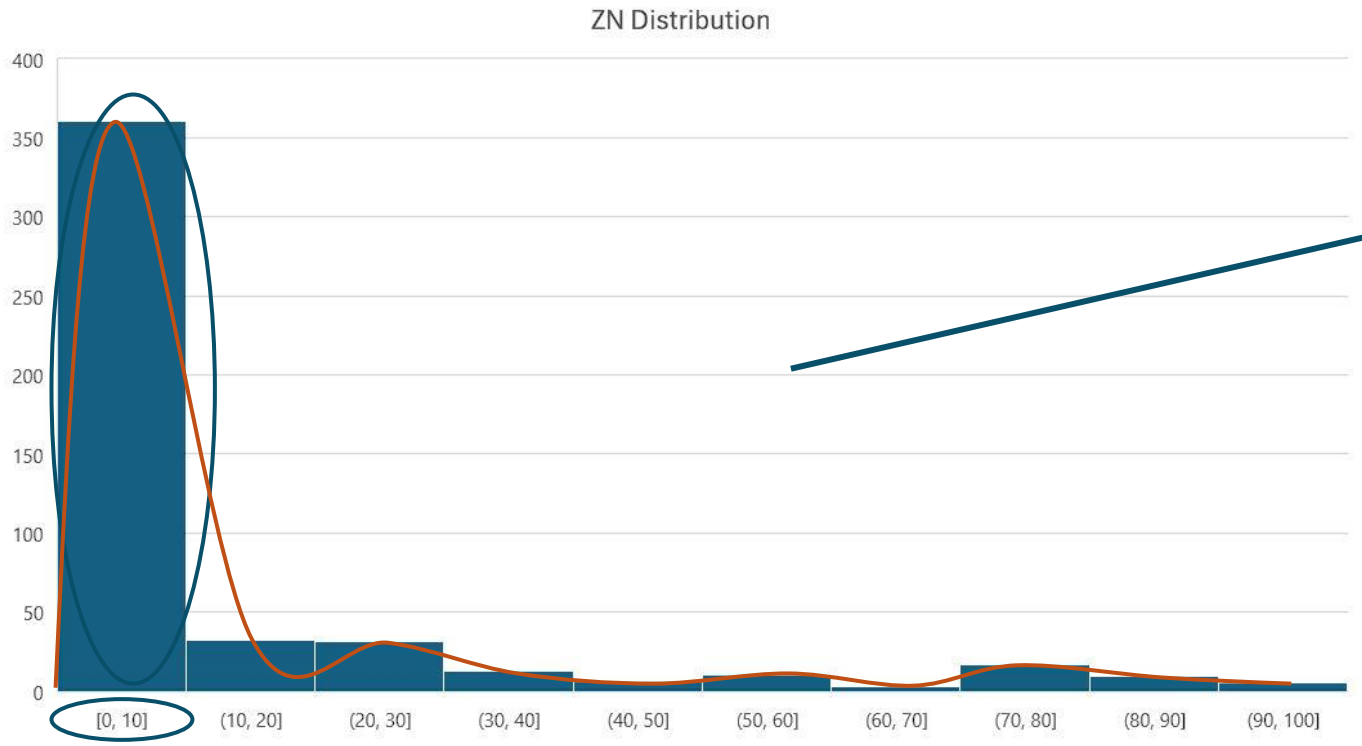
As we can see, there are 20 missing values in ZN column, that would approximately be 4% of our sample.

There are several approaches that we can take to solve this problem

- Replacing it by average of ZN feature
- Replacing it with mode
- Replacing it with median
- Impute the missing values by a model
- Removing missing values

Data Cleansing

Handling Missing Values Of ZN (Zoned) Feature



Histogram Of ZN Feature:

Paying attention to the histogram chart of ZN feature indicates that this feature is highly positively skewed and the mode of this feature must be a value between 0 to 10, the chart suggests us that replacing the missing values of this feature with the mode, might be the best method of handling missing values of this attribute.

So, next step for us for facing the missing values of this feature is to find the mode. We do it we use of descriptive statistics option of XLSTAT add-in of excel.

Descriptive statistics (Qualitative data):

Variable\ Statistic	Nbr. of observations	Nbr. of missing values	Sum of weights	Nbr. of categories	Mode	Mode frequency
ZN (%)	506	0	506	27	0	360

Mode:

The descriptive statistics of this feature tells us the mode of ZN attribute is 0 and the frequency of this occurring is 360 times between 506 records.

Data Cleansing

Handling Missing Values Of INDUS (Industrial) Feature

INDUS (%)
8.14
8.14
5.96
NA
5.96
5.96
2.95
2.95
6.91
6.91
6.91
6.91
6.91
6.91
NA
6.91
6.91
5.64
NA
5.64
5.64

Detecting Missing values:

As we can see, INDUS column contains some missing values that we should firstly detect them, then decide how to deal with them.

INDUS (%)
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA

Count: 20 Min: 0 Max: 0 Sum: 0

Count Of Missing Values:

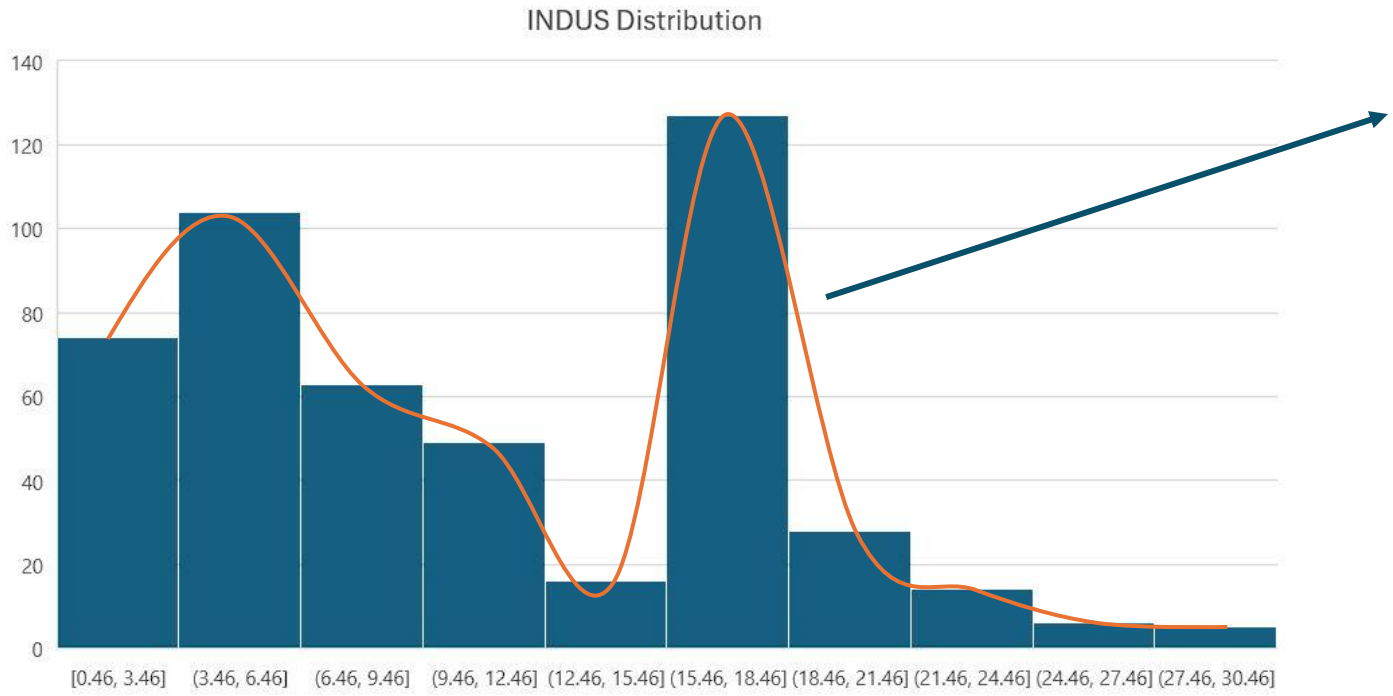
As we can see, there are 20 missing values in INDUS column, that would approximately be 4% of our sample.

There are several approaches that we can take to solve this problem

- Replacing it by average of Crime Rate feature
- Replacing it with mode
- Replacing it with median
- Impute the missing values by a model
- Removing missing values

Data Cleansing

Handling Missing Values Of INDUS (Industrial) Feature



Histogram Of INDUS Feature:

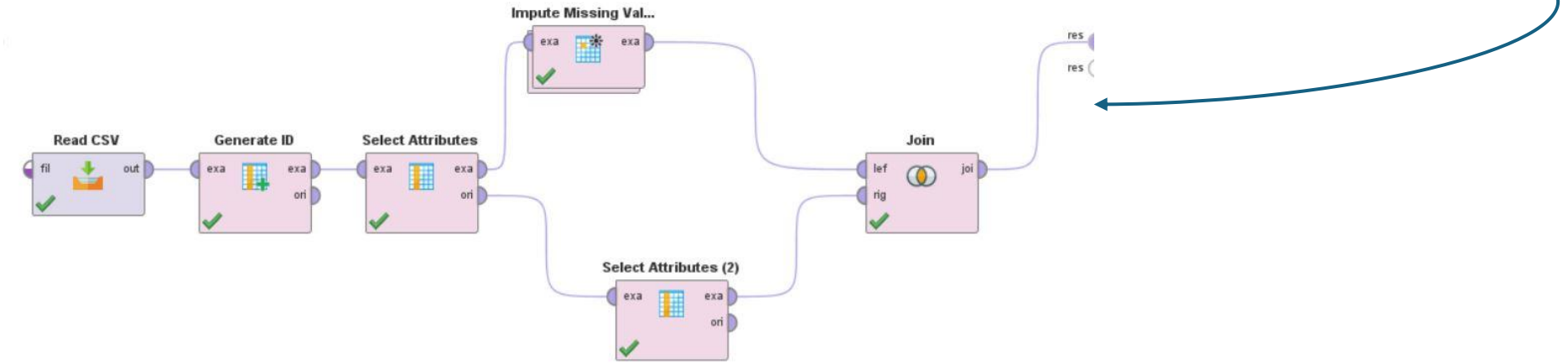
Paying attention to the histogram chart of INDUS feature indicates that this feature is multi-modal, so replacing the missing values with mode does not seem wise.

Cause this feature does not follow a normal distribution, replacing missing values of this attribute with average would not seem a valid approach.

On the other hand, this features seems to be highly influenced by other features, because INDUS feature shows the industrial proportion of each sample which can be related to population, air pollution and so on.

So, I believe that imputing the missing values of INDUS feature would be the best approach that we can take.

For this purpose, I used RapidMiner software to model INDUS feature and impute the missing values of it.



Data Cleansing

Handling Missing Values Of INDUS (Industrial) Feature

I imputed the missing values of INDUS feature with three different algorithms and kept the results to compare .
As we can see, they seem so similar so I decided to take the average of them and replace the missing values of INDUS feature with these averages.

Row No.	id	CRIM	ZN (%)	INDUS (%)	indus-missing
37	37	0.097	0	3.410	-1000
48	48	0.229	0	7.849	-1000
52	52	0.043	21	5.436	-1000
124	124	0.150	0	25.650	-1000
134	134	0.330	0	21.890	-1000
148	148	2.369	0	19.580	-1000
149	149	2.331	0	19.580	-1000
174	174	0.092	0	4.050	-1000
178	178	0.054	0	4.050	-1000
220	220	0.114	0	13.890	-1000
246	246	0.191	22	5.994	-1000
293	293	0.036	80	4.950	-1000
298	298	0.141	0	5.994	-1000
306	306	0.055	33	1.960	-1000

Imputed Values With The Use Of Decision Tree Algorithm

Identifier For Keeping Track Of Missing Values Of INDUS

neural	knn	decision tree	average
7.177981174	8.366428418	3.41	6.32
7.619970752	6.304594138	7.849090909	7.26
5.425517185	5.615101361	5.436363636	5.49
18.35297293	21.57738663	25.65	21.86
21.8108098	21.89	21.89	21.86
20.73114454	17.83131089	19.58	19.38
19.97324677	19.58	19.58	19.71
6.628441253	5.807149303	4.05	5.50
6.328765706	7.389840069	4.05	5.92
8.660299333	12.42342015	13.89	11.66
4.604502162	7.026207327	5.993636364	5.87
3.90917989	3.556979045	4.95	4.14
6.6697242	9.916911436	5.993636364	7.53
5.116005872	5.019441472	1.96	4.03
2.422701462	2.957447165	2.03	2.47
17.69900087	18.1	18.1	17.97
18.93715042	18.1	18.1	18.38
18.21334844	18.1	18.1	18.14
18.40577776	18.1	18.1	18.20
17.83716379	18.1	18.1	18.01

Data Cleansing

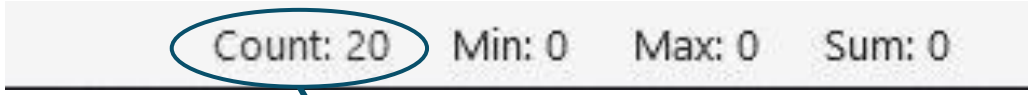
Handling Missing Values Of CHAS (Charles River) Feature

CHAS
0
0
0
0
0
0
NA
0
0
NA
0
0
0
0
NA
0
0
0

Detecting Missing values:

As we can see, CHAS column contains some missing values that we should firstly detect them, then decide how to deal with them.

CHAS
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA



Count Of Missing Values:

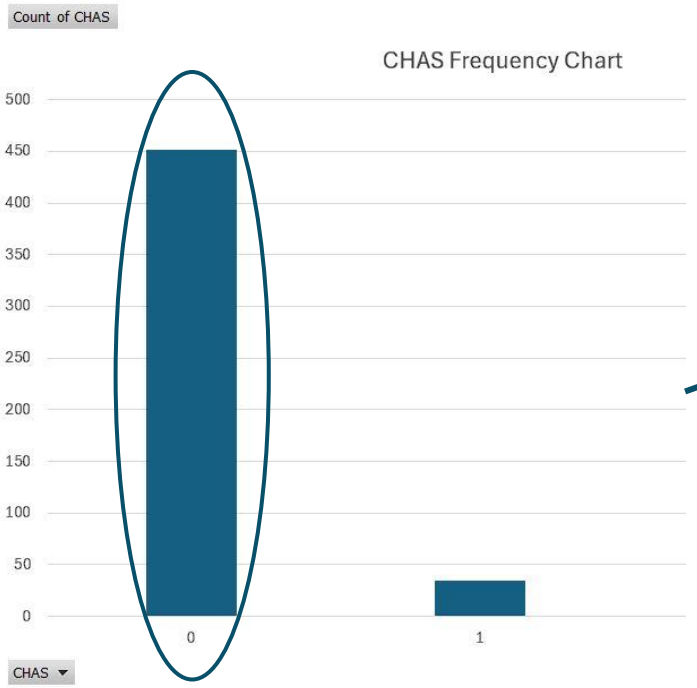
As we can see, there are 20 missing values in CHAS column, that would approximately be 4% of our sample.

There are several approaches that we can take to solve this problem

- Replacing it by average of Crime Rate feature
- Replacing it with mode
- Replacing it with median
- Impute the missing values by a model
- Removing missing values

Data Cleansing

Handling Missing Values Of CHAS (Charles River) Feature



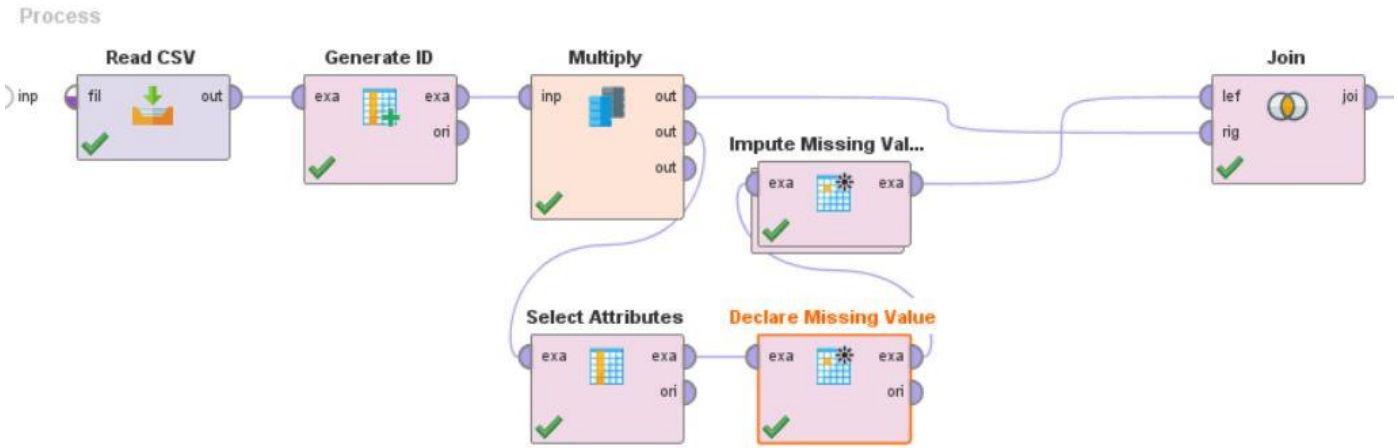
Row Labels	Count of CHAS
0	452
1	34
Grand Total	**486**

Frequency Chart Of CHAS Feature:

Considering the frequency chart of CHAS feature, remembering that this is a binomial feature, and paying attention to difference between 0 (which indicates that specific record of us is not near Charles river) and 1 (which indicates that specific record of us is near Charles river) lead us to this approach to replace the missing values of CHAS feature with mode (which is 0).

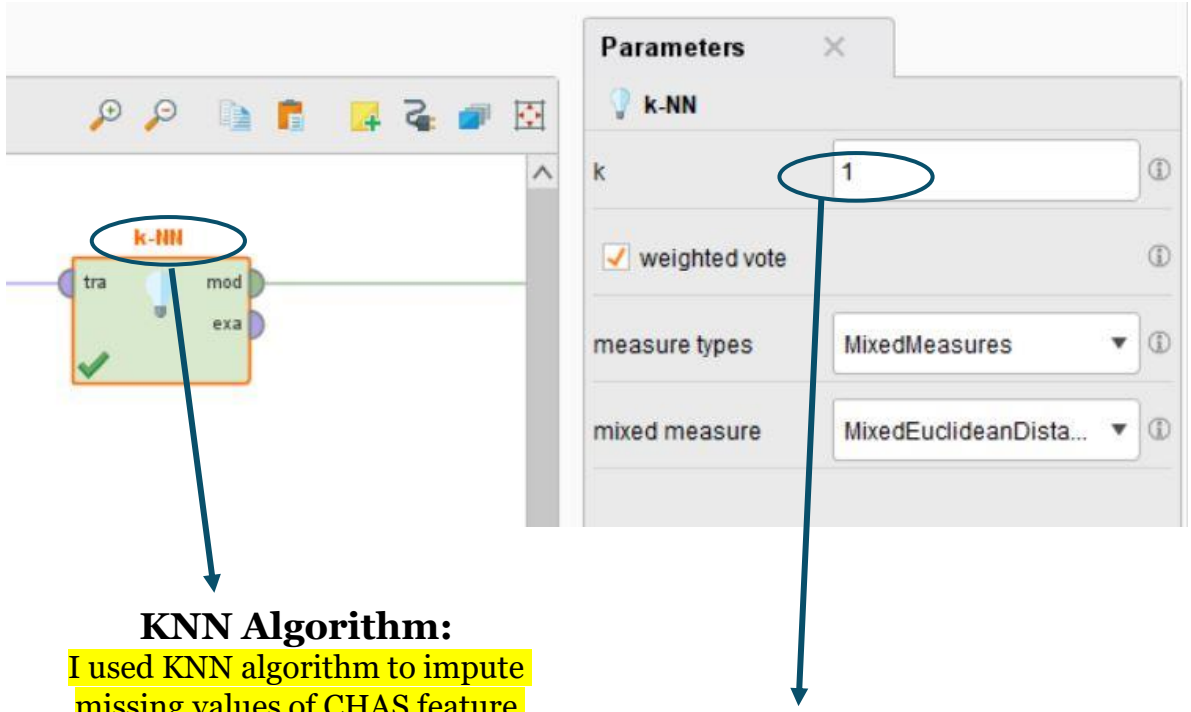
But for being more accurate, I would prefer to impute the missing values with an algorithm. Cause it would have an effect on different aspects of that area (like business or population of the area) .

So, I created a model by the use of RapidMiner to impute the missing values of this feature.



Data Cleansing

Handling Missing Values Of CHAS (Charles River) Feature



KNN Algorithm:

I used KNN algorithm to impute missing values of CHAS feature considering that it indicates that an area is near Charles river or not, I think that KNN algorithm would be helpful because it is a distance-base algorithm

Number Of Neighbors:

I set the number of neighbors to 2 and then to 1 and got same results.

identifier:

I created this feature to keep the track of the missing values of CHAS feature.

Row No.	id	CRIM	ZN (%)	INDUS (%)	CHAS	chas-mi...
7	7	0.088	12.500	7.870	0	1000
10	10	0.170	12.500	7.870	0	1000
15	15	0.638	0	8.140	0	1000
44	44	0.159	0	6.910	0	1000
120	120	0.145	0	10.010	0	1000
153	153	1.127	0	19.580	0	1000
207	207	0.230	0	10.590	1	1000
241	241	0.113	30	4.930	0	1000
246	246	0.191	22	5.870	0	1000

Results:

As expected, it imputed all the missing values to be 0 (it is the mode of this feature); but only imputed the 207th sample to be 1.

NOTE:

Cause our samples are imbalance, we could have made a mistake imputing the missing values like this. To be more accurate, we could first balance them. But considering our purpose, that would not be necessary.

Data Cleansing

Handling Missing Values Of AGE Feature

AGE (%)
31.1
21.4
36.8
33
6.6
17.5
7.8
6.2
NA
45
74.5
45.8
53.7
36.6
33.5

Detecting Missing values:

As we can see, AGE column contains some missing values that we should firstly detect them, then decide how to deal with them.

AGE (%)
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA
NA

Count Of Missing Values:

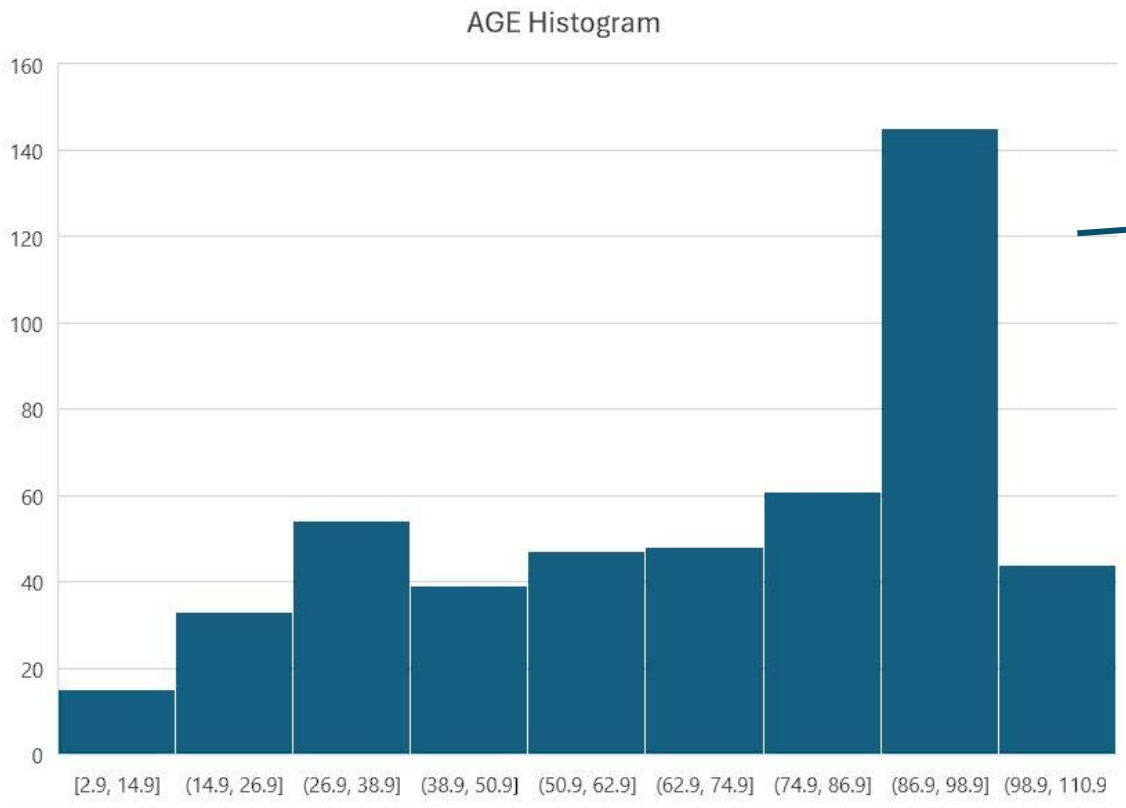
As we can see, there are 20 missing values in AGE column, that would approximately be 4% of our sample.

There are several approaches that we can take to solve this problem

- Replacing it by average of Crime Rate feature
- Replacing it with mode
- Replacing it with median
- Impute the missing values by a model
- Removing missing values

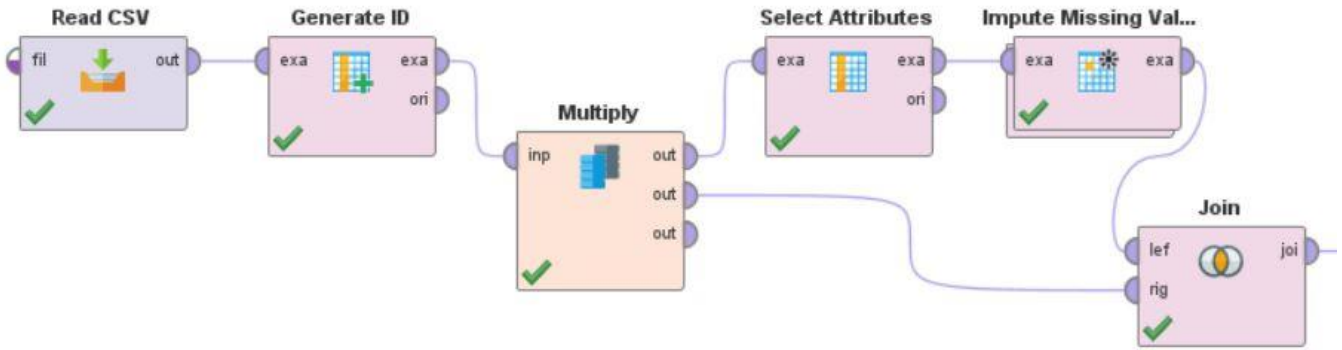
Data Cleansing

Handling Missing Values Of AGE Feature



Histogram Of AGE Feature:

The histogram chart of this variable obviously does not follow a normal distribution and it is negatively skewed.
Considering the chart, replacing missing variables of this feature with mode or average would not be my preferred method.
I prefer to go with imputing the missing values and using ensembled method to deal with missing values of this attribute.
So; like before, I used RapidMiner for this purpose and created a model.



Data Cleansing

Handling Missing Values Of AGE Feature

Algorithms Which I Used:

As we can see, I used three different algorithms for imputing the missing values of AGE feature and kept the results of each algorithm to use in ensemble method that I wanted to apply.

KNN	Decision Tree	Neural Network	Average
52.2	49.7	35.2	45.7
85.9	89.3	83.8	86.3
91.8	97.2	99.9	96.3
96.0	98.9	104.2	99.7
64.8	98.0	104.4	89.1
92.1	97.4	108.4	99.3
95.0	96.1	99.8	97.0
69.6	63.1	79.9	70.9
26.2	32.6	32.9	30.6
56.5	27.3	54.8	46.2
68.3	38.8	81.0	62.7
52.6	51.4	38.3	47.5
25.2	38.5	56.7	40.1
87.4	69.4	90.8	82.5
47.5	23.2	43.3	38.0
94.9	98.0	72.5	88.5
85.1	88.7	96.4	90.1
87.7	82.6	95.5	88.6
91.3	91.8	95.5	92.8
66.9	69.4	88.6	75.0

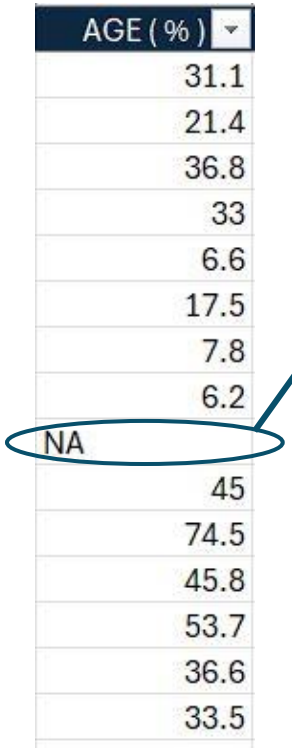
Ensemble Method:

Considering that AGE variable is a continuous variable, I used the averaging method based on these three algorithms to deal with missing values of AGE feature.

At the end, I replaced each missing values with this corresponding value draw from averaging method.

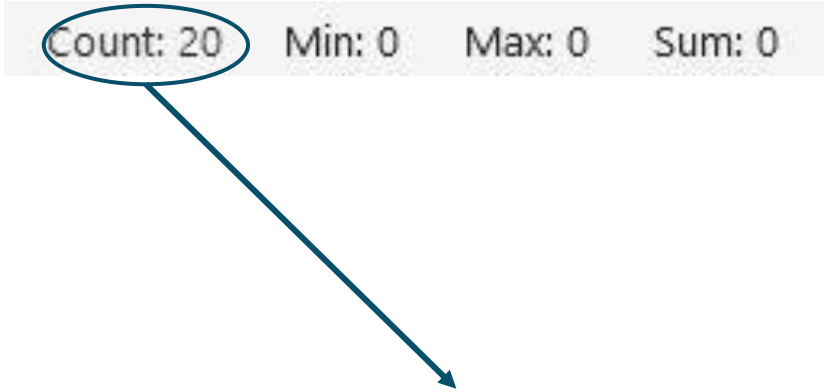
Data Cleansing

Handling Missing Values Of LSTAT (low-status) Feature



Detecting Missing values:

As we can see, LSTAT column contains some missing values that we should firstly detect them, then decide how to deal with them.



Count Of Missing Values:

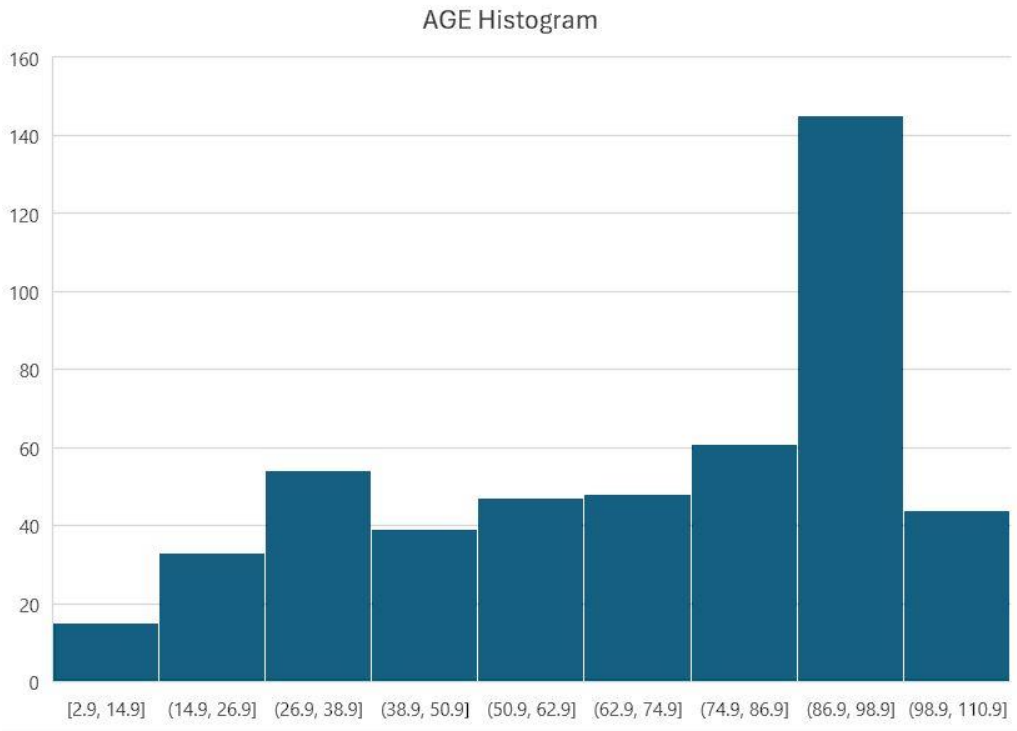
As we can see, there are 20 missing values in LSTAT column, that would approximately be 4% of our sample.

There are several approaches that we can take to solve this problem

- Replacing it by average of Crime Rate feature
- Replacing it with mode
- Replacing it with median
- Impute the missing values by a model
- Removing missing values

Data Cleansing

Handling Missing Values Of LSTAT (low-status) Feature

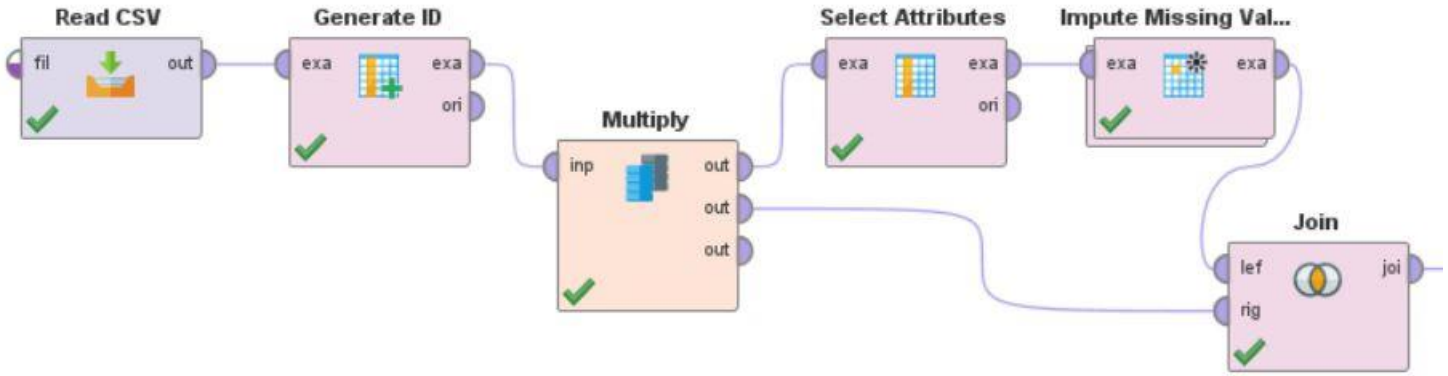


Histogram Of LSTAT Feature:

Histogram of this feature indicates that this variable is uni-modal, so one approach would be replacing the missing values with mode, but the point is that; LSTAT (which indicates the proportion of low-status people in an area) would be guessable based on that area's characteristics.

so, like before I preferer to impute the missing values not to replace them.

So, I created a model like this model below.



Data Cleansing

Handling Missing Values Of LSTAT (low-status) Feature

Algorithms Which I Used:

As we can see, I used three different algorithms for imputing the missing values of LSTAT feature and kept the results of each algorithm to use in ensemble method that I wanted to apply.

KNN	Decision Tree	Neural Network	Average
52.2	49.7	35.2	45.7
85.9	89.3	83.8	86.3
91.8	97.2	99.9	96.3
96.0	98.9	104.2	99.7
64.8	98.0	104.4	89.1
92.1	97.4	108.4	99.3
95.0	96.1	99.8	97.0
69.6	63.1	79.9	70.9
26.2	32.6	32.9	30.6
56.5	27.3	54.8	46.2
68.3	38.8	81.0	62.7
52.6	51.4	38.3	47.5
25.2	38.5	56.7	40.1
87.4	69.4	90.8	82.5
47.5	23.2	43.3	38.0
94.9	98.0	72.5	88.5
85.1	88.7	96.4	90.1
87.7	82.6	95.5	88.6
91.3	91.8	95.5	92.8
66.9	69.4	88.6	75.0

Ensemble Method:

Considering that LSTAT variable is a continuous variable, I used the averaging method based on these three algorithms to deal with missing values of AGE feature.
At the end, I replaced each missing values with this corresponding value draw from averaging method.

Data Cleansing

Handling Missing Values

All Missing Values Are Replaced Or Imputed:

As we can see in the table below, all missing values of all features of our dataset are either replaced or imputed now.

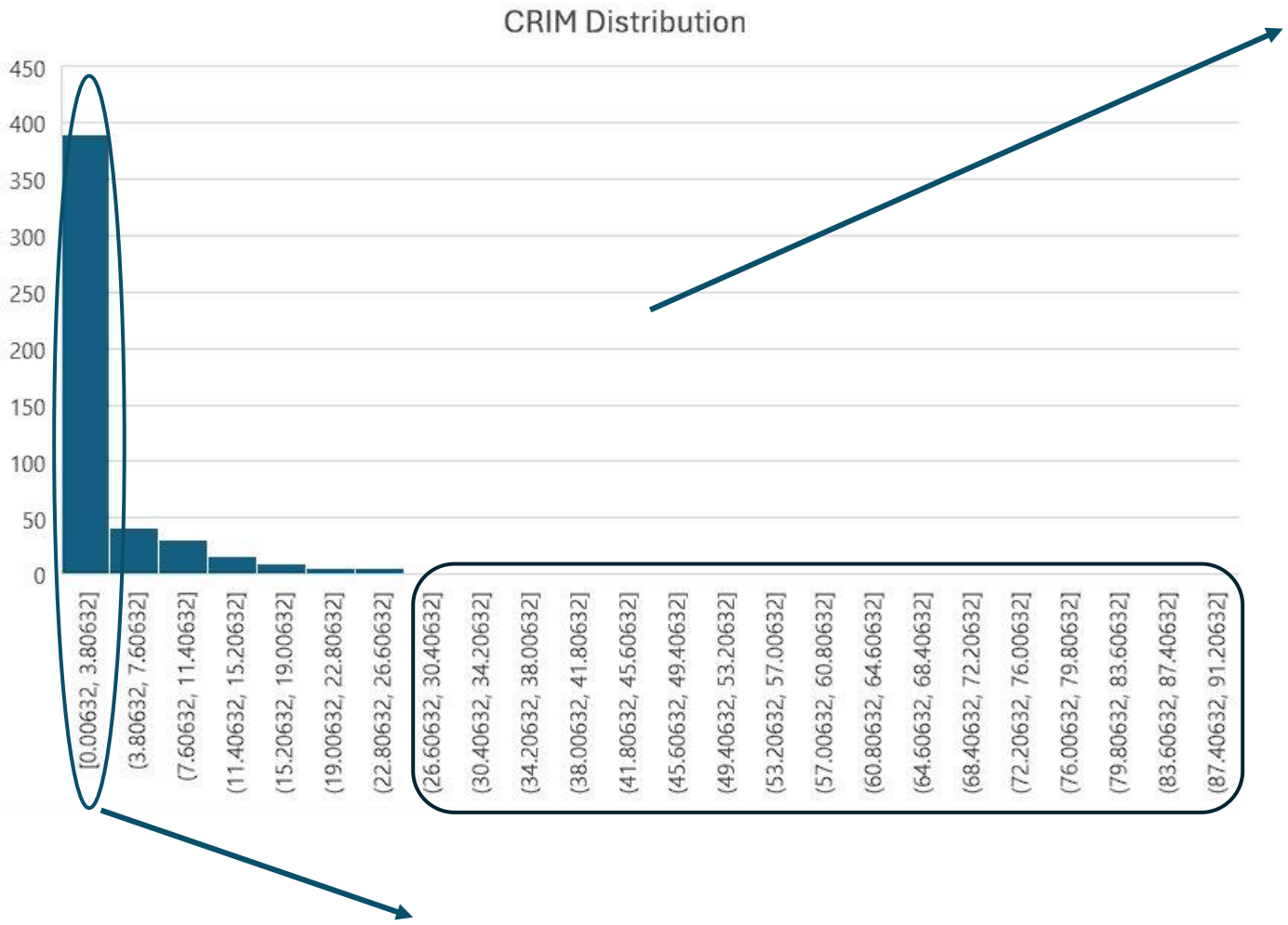
We do not have any missing values anymore.

Now there would be the best time to go for the next step of our analysis, that would be examining the distribution of each variable and their descriptive statistics.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CRIM	ZN (%)	INDUS (%)	CHAS	NOX (PPM)	RM	AGE (%)	DIS (Miles)	RAD	TAX (1,000$)	PTRATIO	B	LSTAT (%)	MEDV (1,000$)
2	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
3	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
4	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
5	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
6	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	6.29	36.2
7	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
8	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
9	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
10	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
11	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
12	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
13	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
14	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
15	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
16	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
17	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9
18	1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307	21	386.85	6.58	23.1
19	0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307	21	386.75	14.67	17.5
20	0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307	21	288.99	11.69	20.2
21	0.7258	0	8.14	0	0.538	5.727	69.5	3.7965	4	307	21	390.95	11.28	18.2
22	1.25179	0	8.14	0	0.538	5.57	98.1	3.7979	4	307	21	376.57	21.02	13.6
23	0.85204	0	8.14	0	0.538	5.965	89.2	4.0123	4	307	21	392.53	13.83	19.6
24	1.23247	0	8.14	0	0.538	6.142	91.7	3.9769	4	307	21	396.9	18.72	15.2
25	0.98843	0	8.14	0	0.538	5.813	100	4.0952	4	307	21	394.54	19.88	14.5
26	0.75026	0	8.14	0	0.538	5.924	94.1	4.3996	4	307	21	394.33	16.3	15.6
27	0.84054	0	8.14	0	0.538	5.599	85.7	4.4546	4	307	21	303.42	16.51	13.9

Statistical Analyses (CRIM Variable)

Examining The Distribution



Mode:

The mode of CRIM feature must be something in this range.

CRIM Histogram Chart:

The chart shows us that Crime rate variable is highly and positively skewed. Crime rates above 26% are rare and majority of different towns of Boston have crime rates below this number.

Being this much skewed, might have an effect on our statistical analyses. So; we need to apply some kind of transformation (log transformation, Box-Cox transformation, etc.) to deal with this problem, and then, we can apply our analyses.

The chart also suggests that this variable might have many outliers that we will talk about later on.

Statistical Analyses (CRIM Variable)

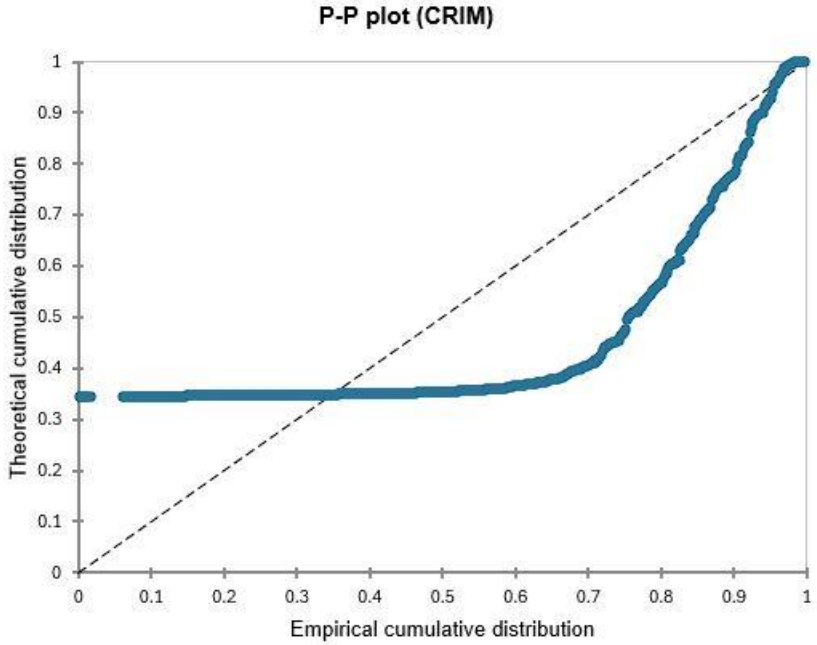
Examining The Descriptive Statistics

Statistic	CRIM
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	0.006
Maximum	88.976
Freq. of minimum	1
Freq. of maximum	1
Range	88.970
1st Quartile	0.069
Median	0.225
3rd Quartile	2.809
Sum	1755.671
Mean	3.470
Variance (n)	73.377
Variance (n-1)	73.522
Standard deviation (n)	8.566
Standard deviation (n-1)	8.575
Skewness (Pearson)	5.296
Kurtosis (Pearson)	37.600
Lower bound on mean (95%)	2.721
Upper bound on mean (95%)	4.219
Lower bound on variance (95%)	65.234
Upper bound on variance (95%)	83.505

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen only 1 time among all of the records
- Maximum value of this variable can be seen only 1 time among all of the records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A skewness value of 5.29 is considered very high and indicates that your distribution is highly positively skewed. The distribution has a long tail on the right side. This means that the majority of the data points are concentrated on the left side, but there are some extremely high values that stretch out to the right.
- A kurtosis of 37.6 suggest that the peak of the distribution of this variable is extremely sharp
- The mean of the population of this variable must be something between 2.7 and 4.2 with confidence level of 95%
- The variance of the population of this variable must be something between 65.2 and 83.5 with confidence level of 95%

Statistical Analyses (CRIM Variable)

Normality Test (Anderson-Darling Method)



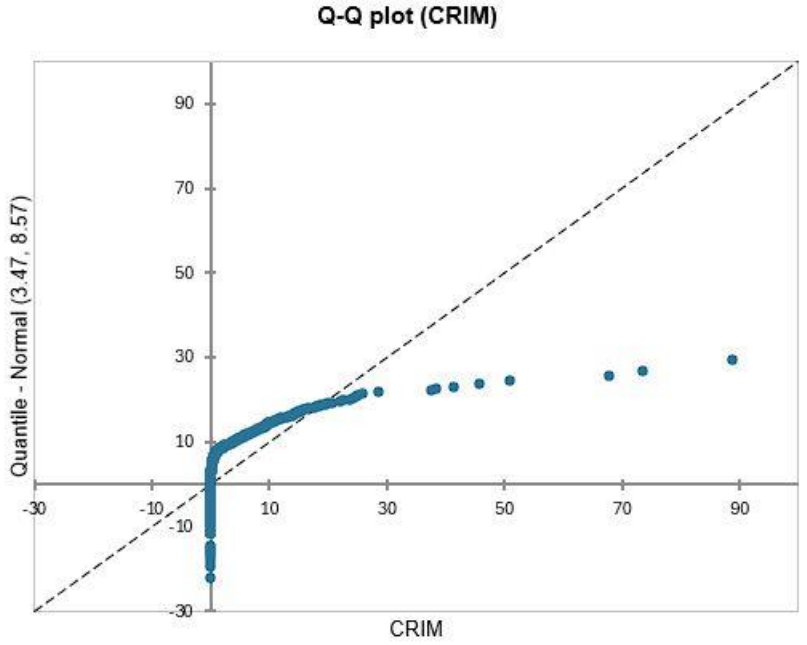
Normality Test Result :

p-value is less than alpha, so we should reject the null hypothesis. So; CRIM variable (as we saw before) does not follow a normal distribution. Considering the p-value, it is not even close to alpha, and we might never convert this variable's distribution, to a normal one. Even by transforming methods or removing outliers.

Anderson-Darling test (CRIM):	
A^2	+Inf
p-value (Two-tailed)	**<0.0001**
alpha	0.050

P-P & Q-Q plot:

Paying attention to these two charts, as we mentioned before, it does not seem that we can make this variable to follow a normal distribution. It is far away from a normal distribution



Statistical Analyses (CRIM Variable)

Outliers Detecting (Variable Transformation)

Raw data:

This is the raw data of CRIM variable without any transformations.

Z-Score Normalization:

In this column, I transformed the data of CRIM variable with use of excel functions.
I create a function like this :

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

XLSTAT Check (Z-Score Normalization):

In this column, I transformed the data of CRIM variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

	A	B	C	D	E
1	CRIM	CRIM (Z transformation)	CRIM (Normalization)	standardize (n-1)	0 to 1
2	0.00632	-0.403916175	0	-0.403916175	0
3	0.02731	-0.401468223	0.000235923	-0.401468223	0.000235923
4	0.02729	-0.401470556	0.000235698	-0.401470556	0.000235698
5	0.03237	-0.400878102	0.000292796	-0.400878102	0.000292796
6	0.06905	-0.39660031	0.00070507	-0.39660031	0.00070507

Normalization:

In this column, I normalized the data of CRIM variable with use of excel functions.
I create a function like this :

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

XLSTAT Check (Normalization):

In this column, I normalized the data of CRIM variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (CRIM Variable)

Outliers Detecting (Box-Plot Method)

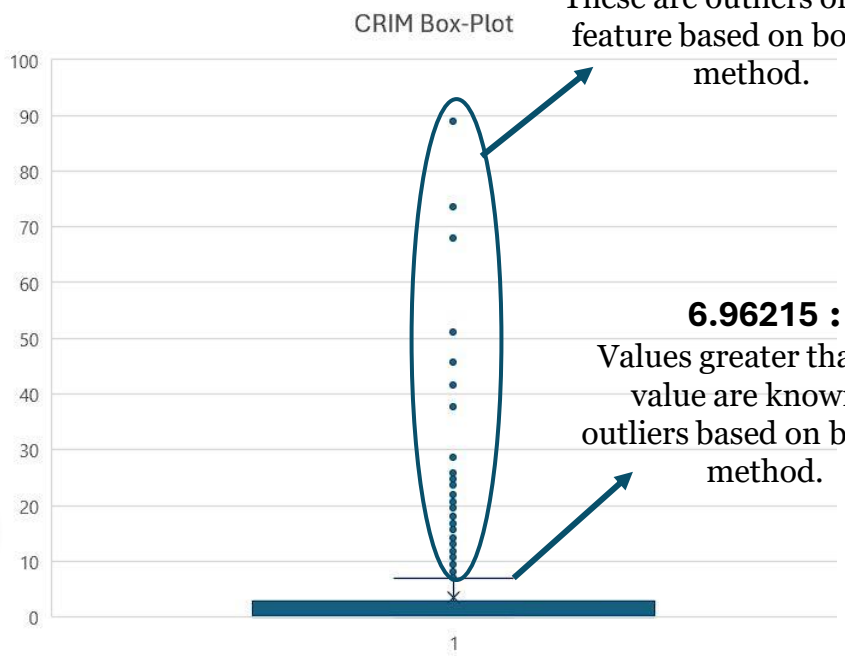
CRIM
8.98296
13.5222
9.2323
8.26725
11.1081
18.4982
19.6091
15.288
9.82349
23.6482
17.8667
88.9762
15.8744
9.18702
7.99248
20.0849
16.8118
24.3938
22.5971
14.3337

A sample Of Outliers:

In this table, we can see some of the outliers of CRIM feature, based on Box-Plot method.

Box-plot chart:

I inserted a box-plot chart for CRIM variable, and got the chart as it is here.
As we can see, there are a lot of outliers based on box-plot method.
I also applied a conditional formatting on CRIM for values greater than 6.96215 (that I got from the chart) to filter the feature and to find out that how many outliers are detected based on box-plot method.



Outliers:

These are outliers of CRIM feature based on box-plot method.

6.96215 :

Values greater than this value are known as outliers based on box-plot method.

Average: 17.45211163 Count: 80 Min: 7.02259 Max: 88.9762 Sum: 1396.16893

80 Outliers:

80 outliers are detected based on box-plot method

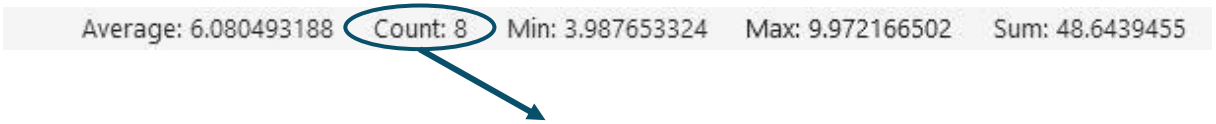
Statistical Analyses (CRIM Variable)

Outliers Detecting (Z-Score Method)

CRIM	CRIM (Z transformation)
88.9762	9.972166502
38.3518	4.068112682
41.5292	4.438675901
67.9208	7.516587477
51.1358	5.559042432
45.7461	4.930470461
73.5341	8.171236723
37.6619	3.987653324

Outliers With Z-Score Method:

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.
So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I got the result as you can see in the table.



8 Outliers:

8 outliers are detected based on Z-Score method.
While, the number of outliers which were detected based on box-plot method was 80.
As it is obvious, there is a significant different between these two methods.

Box-Plot VS Z-Score :

When we compare the results of these two methods for detecting outliers for CRIM feature, there is a significant different.
With Box-Plot method we got 80 outliers
With Z-Score method we got 8 outliers
If we want to decide outliers of which method should rely on, I prefer to go with Z-Score method, cause each outlier detected by this method is also detected as outlier with box-plot method.
On the other hand, number of outliers with box-plot method for this feature are too much, approximately 15.8% of our samples. So; it does not seem wise to go with box-plot method in this situation.

Statistical Analyses (CRIM Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, CRIM variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.

So; in first step, I removed the outliers which were detected by Z-Score method, from the dataset, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.

As the second step, I transformed CRIM variable by Box-Cox method to see if it would follow a normal distribution after the transformation and the answer to this question was also negative.

Also, we could guess these results by paying attention to the histogram chart, P-P and Q-Q charts of this variable.

So, as the conclusion, we find it out that we cannot convert CRIM variable to normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

Summary statistics:

Variable	Observations	Obs. with missing	Obs. without	Minimum	Maximum	Mean	Std. deviation
CRIM	498	0	498	0.006	28.656	2.632	5.017

Anderson-Darling test (CRIM):

A^2	81.379
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Step1.Normality Test After Removing Outliers:

CRIM variable did not follow a normal distribution after removing its outliers.

CRIM (Box_Cox Transformation)
0.00632
0.02731
0.02729
0.03237
0.06905
0.02985
0.08829
0.14455
0.21124
0.17004
0.22489
0.11747
0.09378
0.62976
0.63796
0.62739

Summary statistics:

Variable	Observations	Obs. with missing	Obs. without	Minimum	Maximum	Mean	Std. deviation
CRIM	498	0	498	0.006	28.656	2.632	5.017

Anderson-Darling test (CRIM):

A^2	81.379
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Step2.Normality Test After Box-Cox Transformation:

CRIM variable did not follow a normal distribution after applying Box-Cox transformation.

Statistical Analyses (CRIM Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:
I am going to check the correlation between CRIM variable and target variable which is MEDV, these are both continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	CRIM	MEDV (1,000$)
CRIM	**1**	**-0.384**
MEDV (1,000$)	**-0.384**	**1**

Values in bold are different from 0 with a significance level alpha=0.05

Weak And Inverse Correlation:
The correlation matrix and the value of -0.38 tells us that there is an inverse correlation between these 2 variables.
Meaning that if one of the increase, the other one will decrease.
On the other hand, the absolute value would be 0.38, which indicates that the correlation is relatively weak.

p-values (Pearson):

Variables	CRIM	MEDV (1,000$)
CRIM	**0**	**<0.0001**
MEDV (1,000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:
The value is <0.0001 suggests that the correlation between CRIM and MEDV is statistically significant and it is not due to random changes.

Coefficients of determination (Pearson):

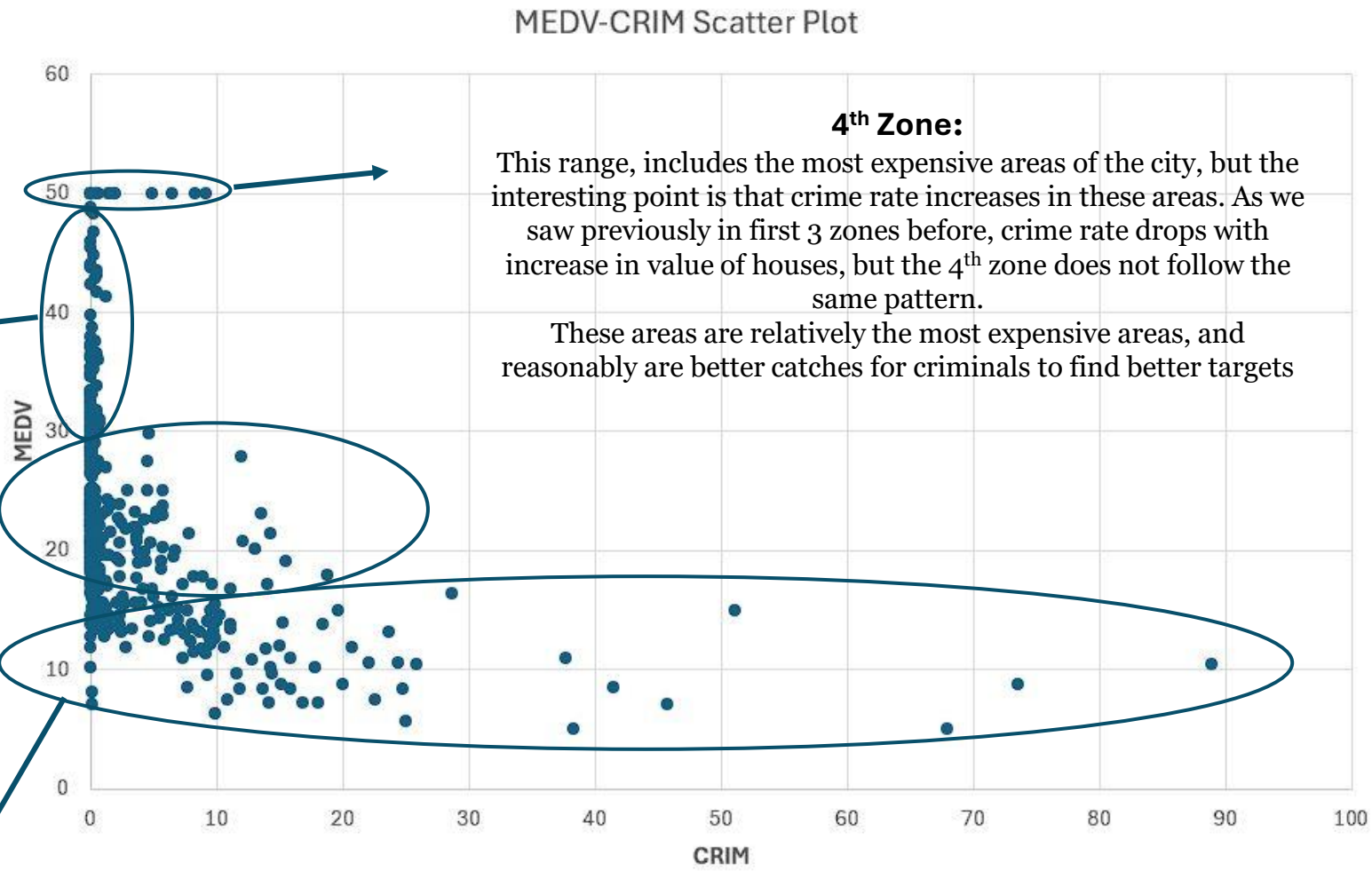
Variables	CRIM	MEDV (1,000$)
CRIM	**1**	0.148
MEDV (1,000$)	0.148	**1**

Power Of Prediction:
The value of 0.148 in this table, indicates that only 14.8% of the variance in target variable (MEDV) can be explained by the variance in CRIM variable.

Statistical Analyses (CRIM Variable)

Scatter Plot With The Target Variable

4 Zones:
As we see on the chart, we can divide the city into 4 zones based on median value of houses in each area and crime rates of each area.
This gives us an interesting insight as we can interpreter as following :



4th Zone:
This range, includes the most expensive areas of the city, but the interesting point is that crime rate increases in these areas. As we saw previously in first 3 zones before, crime rate drops with increase in value of houses, but the 4th zone does not follow the same pattern.
These areas are relatively the most expensive areas, and reasonably are better catches for criminals to find better targets

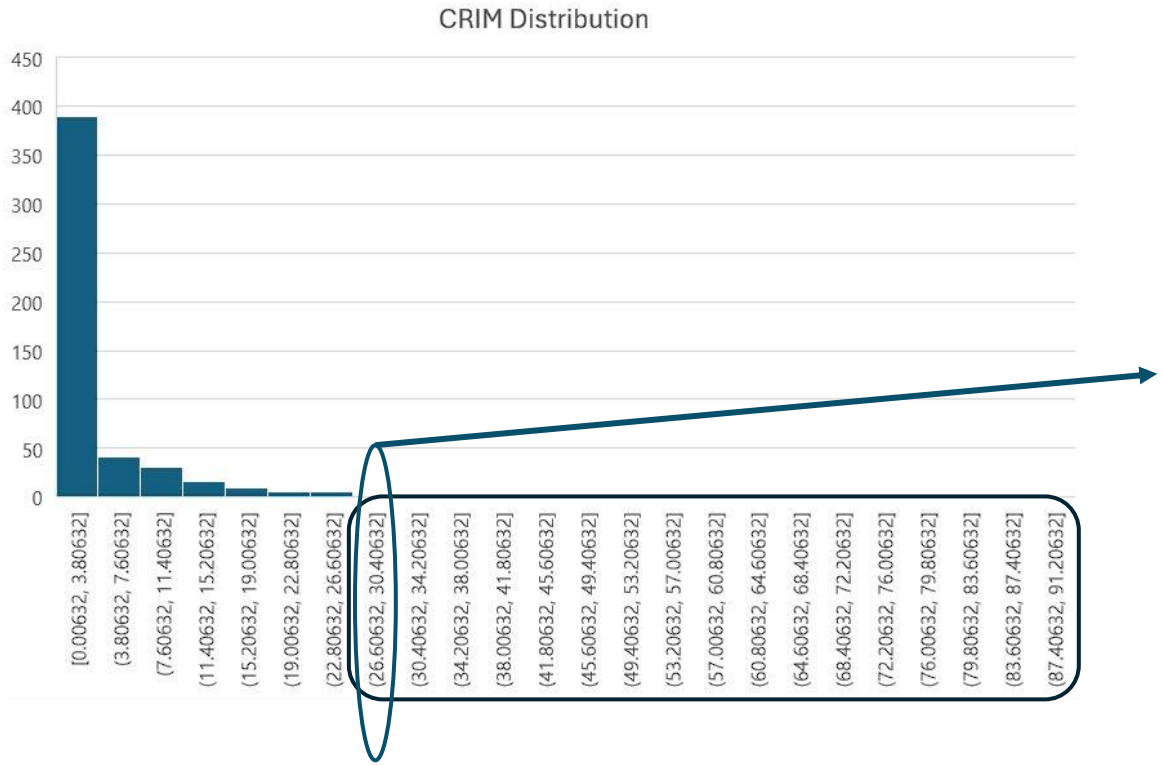
3rd Zone:
In this range of values, the crime rate is at its minimum value. This range is relatively more expensive than 1st and 2nd zones, but interestingly it does not include the most expensive houses.

2nd Zone:
In this range of values, the crime rate is relatively lower than 1st zone, but it stills some areas which crime rate in them is relatively high (not as high as 1st zone)

1st Zone:
In this range of values, the crime rate is relatively high. And there is no difference between areas in this range, in term of crime rates

Statistical Analyses (CRIM Variable)

Question : Is There Any Difference Between Average Of House Prices Based On Crime Rates?



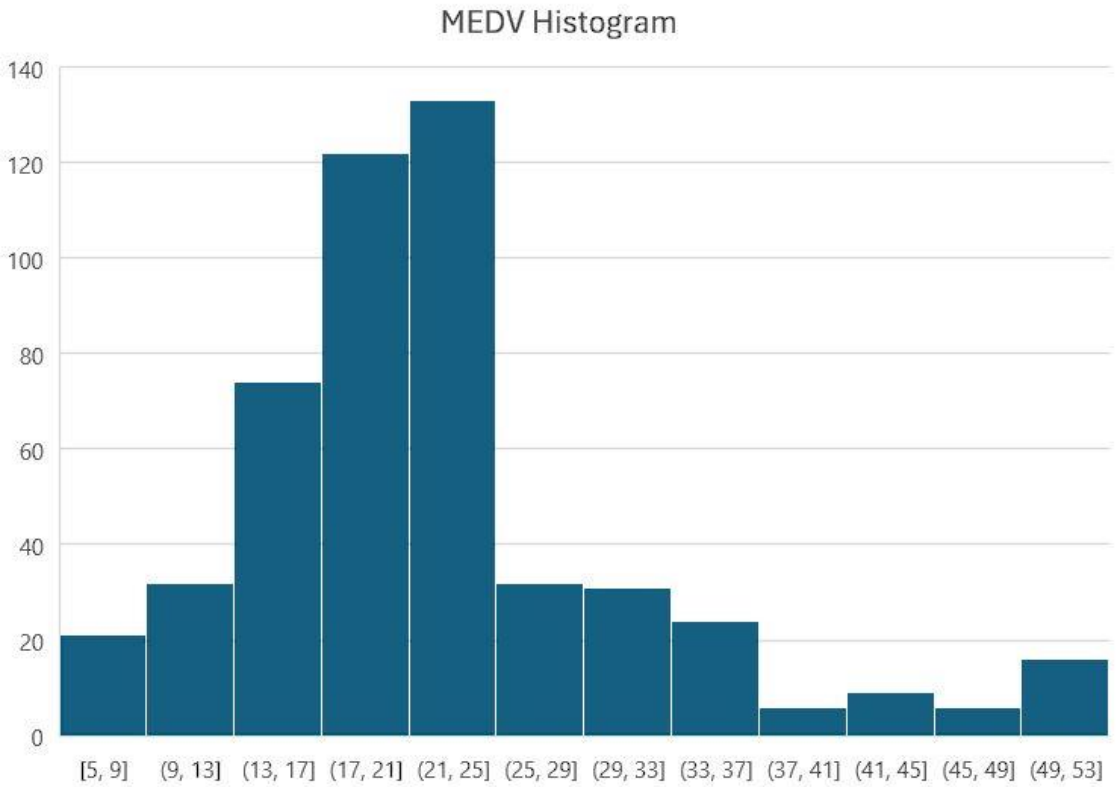
Crime Rate Of 26.6%:

I am going to create a new feature based on CRIM.
This feature is going to be 0 for areas which crime rate in them is less than 26.6%
And is going to be 1 for areas which crime rate in them is above 26.6%
And I am going to compare the average of house prices between these 2 classes.
The reason of choosing the value of 26.6% is the distribution of CRIM feature.
Cause as we can see on the chart, areas with crime rates above this value are relatively rare and can be considered as dangerous areas.
So; it would be wise if we compare areas as we consider dangerous and those which are not considered dangerous (based on our definition of being dangerous) in terms of house prices

	A	B	C
1	CRIM	CRIM Binary Classification	MEDV
2	0.00632	=IF(A2>26.6,1,0)	24
3	0.02731	0	21.6
4	0.02729	0	34.7
5	0.03237	0	33.4
6	0.06905	0	36.2
7	0.02985	0	28.7
8	0.08829	0	22.9

Statistical Analyses (CRIM Variable)

MEDV Normality Test (Anderson-Darling Method)



Anderson-Darling test (MEDV):	
A^2	11.822
p-value (Two-tailed)	**<0.0001**
alpha	0.050

1st Step. Normality Test:

The first step of comparing the average house prices between these two classes that we mentioned previously, is to see if MEDV feature which is our target feature does or does not follow a normal distribution.

As we see on the chart and based on the Anderson-Darling test's result, MEDV feature does not follow a normal distribution.

So; now we should compare the variance of MEDV feature of areas with crime rates above 26.6% (class 1) and areas with crime rates of less than this value (class 0).

And the point is that cause MEDV feature does not follow a normal distribution, So; we should run Leven's test for comparing variances

Statistical Analyses (CRIM Variable)

Variances Equality Test (Leven's Method)

2nd Step. Variances Equality Test:

As we can see, P-value is greater than alpha, so; we should accept the null hypothesis. Meaning that variance of house prices with class 1 (those with crime rates above 26.6%) , is equal to variance of house prices with class 0 (those with crime rates less than 26.6%).

So; for comparing the average of house prices between these two classes, with should assume the equality of variances.

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	2.733
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	0.099
alpha	0.050

Test interpretation:

H0: The variances are identical.

Ha: At least one of the variances is different from another.

As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of CRIM variable, we should use the appropriate method which is Leven's test.

If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (CRIM Variable)

Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:	
95% confidence interval on the difference between the means:	
[7.136, 19.085]	
Difference	13.110
t (Observed value)	4.311
|t| (Critical value)	1.965
DF	504
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Higher Average Of House Prices:

This tells us that areas of class 0, have higher house prices on average, comparing to areas of class 1

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have crime rates less than 26.6% (class 0) is not equal to the average of house prices for those areas which have crime rates greater than 26.6% (class 1).
As we could guess.

Statistical Analyses (CRIM Variable)

The Best Fitting Distribution

Automatic fit summary:

Distribution	p-value
Chi-square	<0.0001
Fisher-Tippett (1)	<0.0001
Fisher-Tippett (2)	<0.0001
Gamma (1)	<0.0001
Gamma (2)	<0.0001
GEV	<0.0001
Gumbel	<0.0001
Log-normal	**<0.0001**
Logistic	<0.0001
Normal	<0.0001
Student	<0.0001
Weibull (1)	<0.0001
Weibull (2)	<0.0001

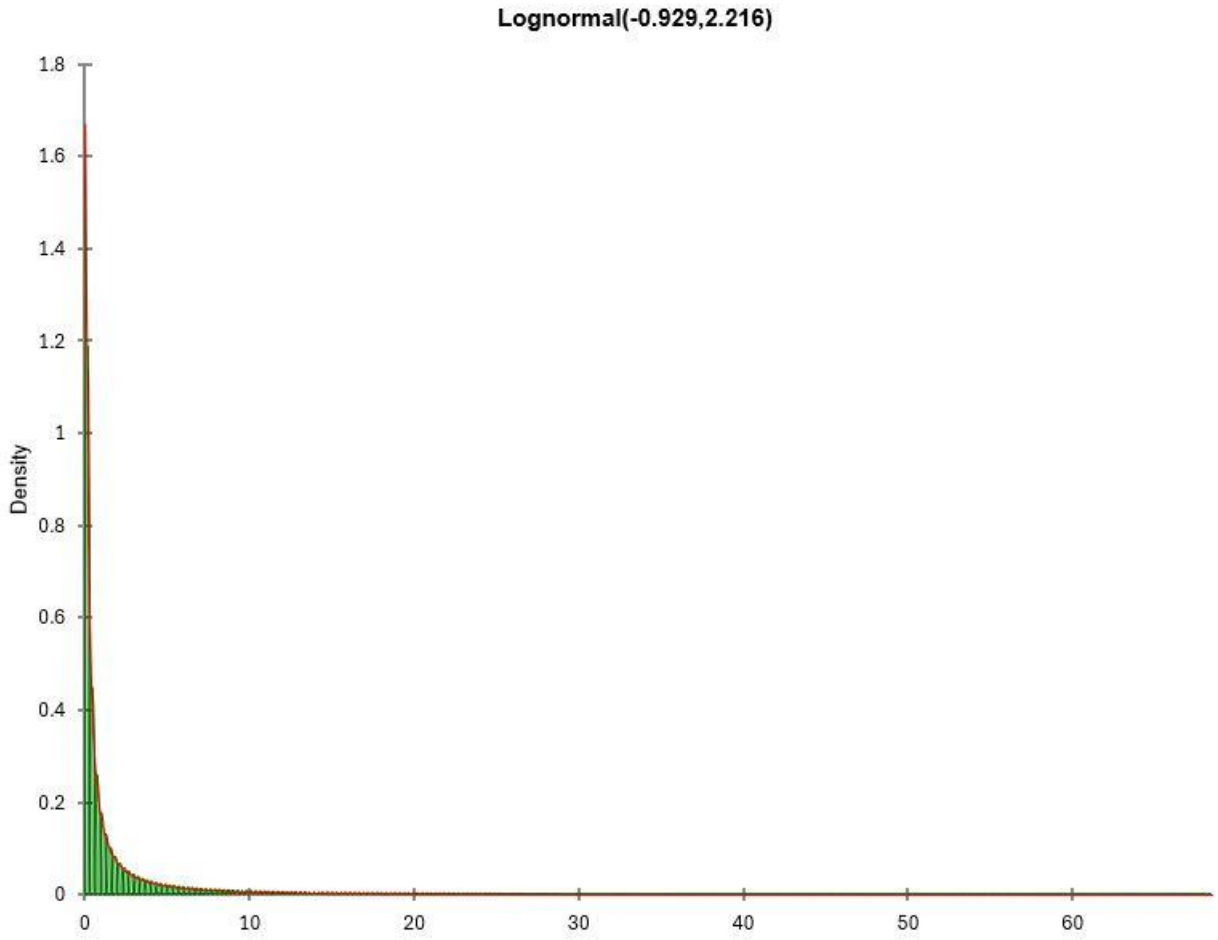
Log-Normal Distribution :

With use of XLSTAT, I found out that the best fitting distribution for CRIM variable, is log-normal distribution with given parameters as below (μ & σ)

Then again, with use of XLSTAT I plot the distribution with these parameters and their corresponding values and I got the chart which you can see on the right, which seems so suit for CRIM variable considering this variable's distribution.

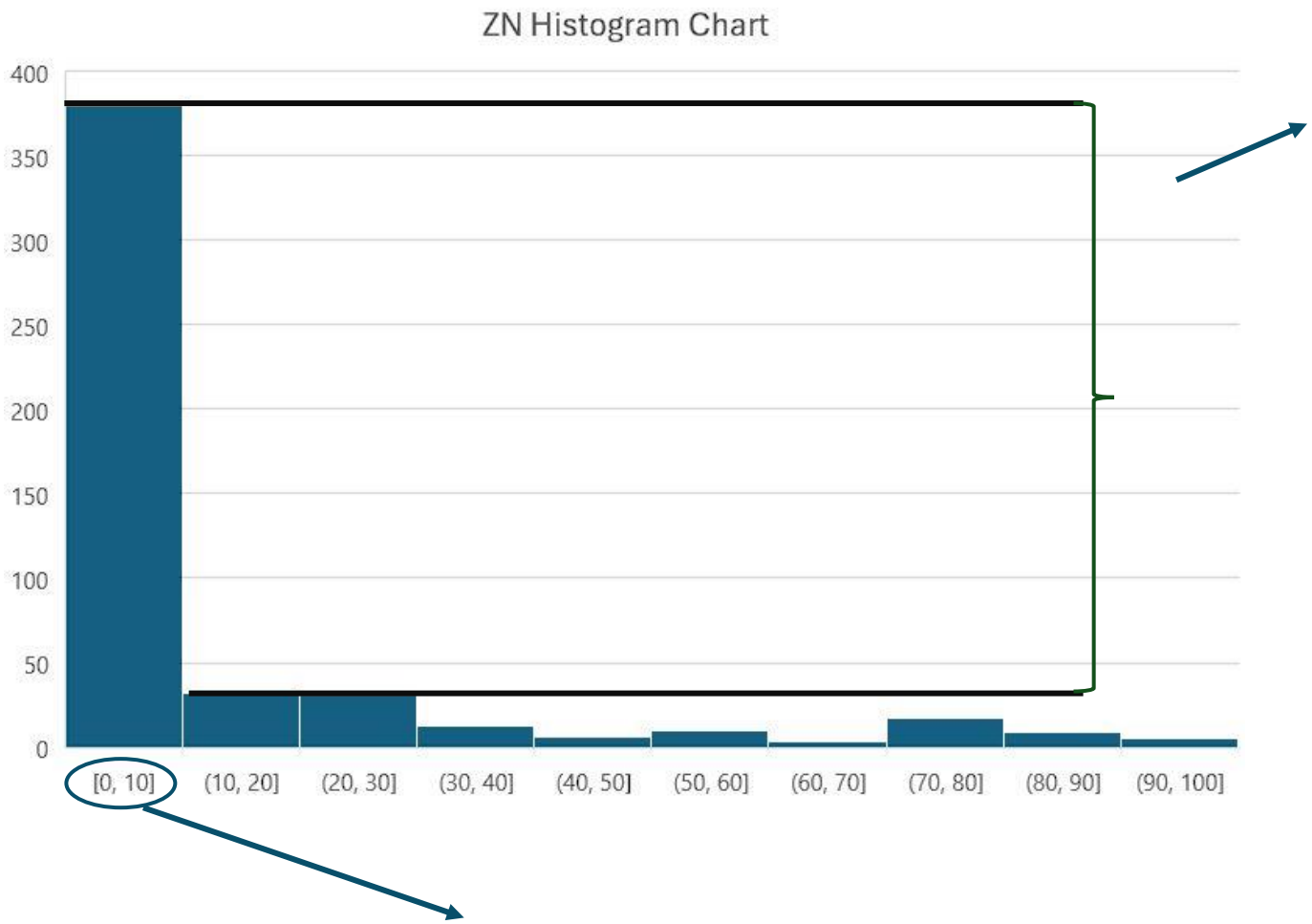
Estimated parameters (Log-normal):

Parameter	Value	Standard error
μ	-0.929	0.099
sigma	2.216	0.070



Statistical Analyses (ZN Variable)

Examining The Distribution



Mode:

The mode of ZN feature must be something in this range.

ZN Histogram Chart:

The chart shows us that ZN variable is positively skewed and it has a right tail.

The high skewness and kurtosis values suggest the presence of outliers and extreme values on the higher end of the distribution. Some areas have significant proportions of land zoned for large lots, but these are rare.

The mode must be something between 0 to 10, and the number of areas with ZN equal to mode, must be much more than areas with ZN not equal to mode of this variable.
As we can see on the chart, the gap is so great.

Obviously; this variable is not normally distributed and it has a log-like distribution.

Statistical Analyses (ZN Variable)

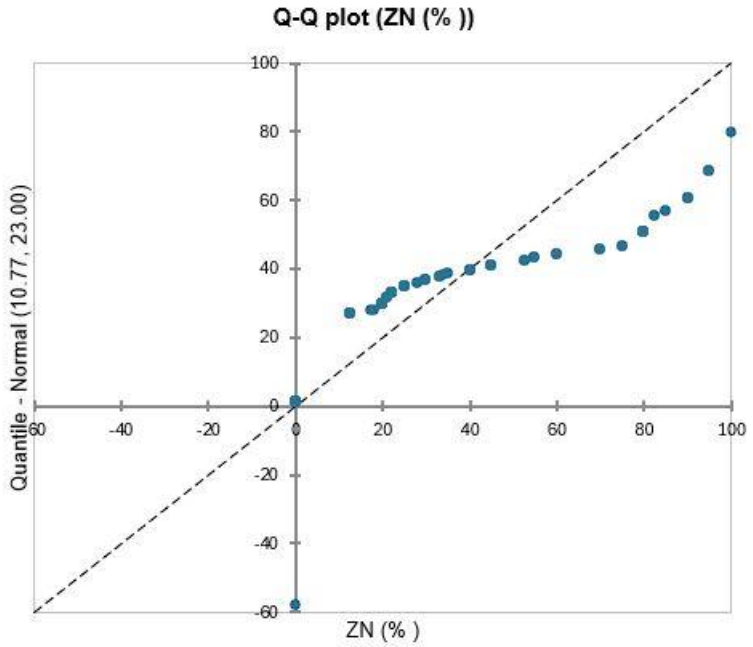
Examining The Descriptive Statistics

Statistic	ZN (%)
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	0.000
Maximum	100.000
Freq. of minimum	380
Freq. of maximum	1
Range	100.000
1st Quartile	0.000
Median	0.000
3rd Quartile	0.000
Sum	5449.000
Mean	10.769
Variance (n)	529.109
Variance (n-1)	530.156
Standard deviation (n)	23.002
Standard deviation (n-1)	23.025
Skewness (Pearson)	2.318
Kurtosis (Pearson)	4.415
Lower bound on mean (95%)	8.758
Upper bound on mean (95%)	12.780
Lower bound on variance (95%)	470.392
Upper bound on variance (95%)	602.143

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable is repeated 380 times and it obviously is the mode
- Maximum value of this variable can be seen only 1 time among all of the records
- Maximum - Minimum
- The fact that Q1, Q2, and Q3 are all zero indicates that at least 75% of the observations have a ZN value of zero. This suggests that a large proportion of the houses are in areas with no zoning for large residential lots.
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A skewness of 2.3 is considered quite high. This indicates a substantial departure from symmetry. In practical terms, the majority of your data points are concentrated on the left side, but there are some significantly higher values that create this positive skew.
- The distribution has a sharper peak compared to a normal distribution.
- The mean of the population of this variable must be something between 8.7 and 12.7 with confidence level of 95%
- The variance of the population of this variable must be something between 470.3 and 602.1 with confidence level of 95%

Statistical Analyses (ZN Variable)

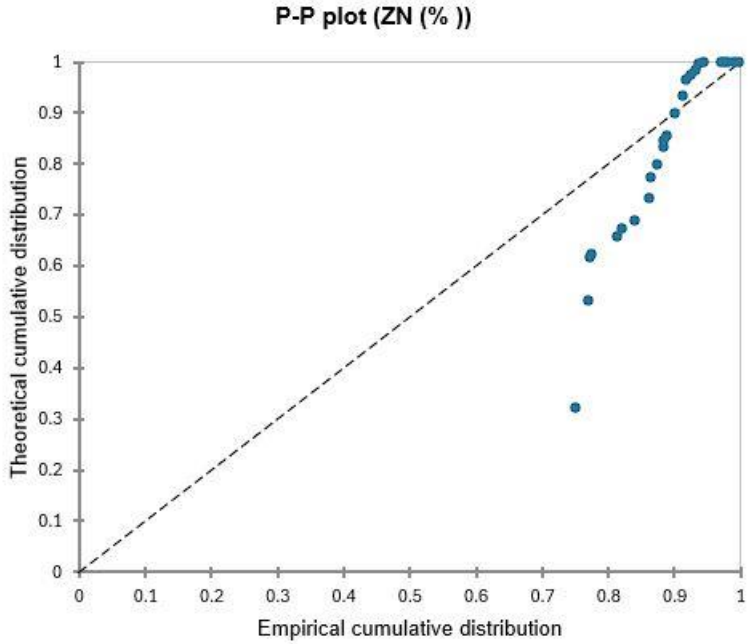
Normality Test (Anderson-Darling Method)



Normality Test Result :

p-value is less than alpha, so we should reject the null hypothesis. So; ZN variable (as we saw before) does not follow a normal distribution.
Considering the p-value, it is not even close to alpha, and we might never convert this variable's distribution, to a normal one. Even by transforming methods or removing outliers.

Anderson-Darling test (ZN (%)):	
A^2	103.413
p-value (Two-tailed)	**<0.0001**
alpha	0.050



P-P & Q-Q plot:

Paying attention to these two charts, as we mentioned before, it does not seem that we can make this variable to follow a normal distribution.

Statistical Analyses (ZN Variable)

Outliers Detecting (Variable Transformation)

Raw data:
This is the raw data of ZN variable without any transformations.

Z-Score Normalization:
In this column, I transformed the data of ZN variable with use of excel functions.
I create a function like this : $X_{standardized} = \frac{X - \mu}{\sigma}$

XLSTAT Check (Z-Score Normalization):
In this column, I transformed the data of ZN variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

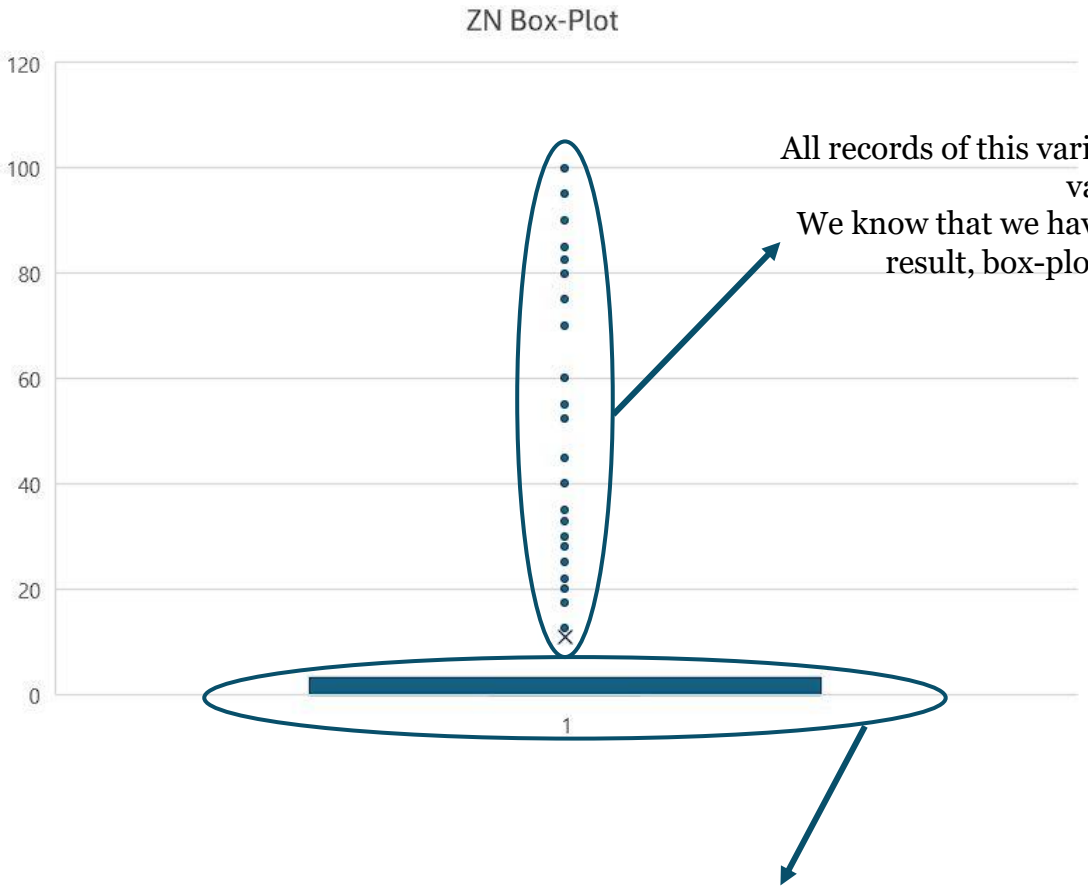
	A	B	C	D	E
1	ZN (%)	ZN (Z transformation)	ZN (Normalization)	Standardized (n-1)	0 to 1
2	18	0.314058041	0.18	0.314058041	0.18
3	0	-0.467696711	0	-0.467696711	0
4	0	-0.467696711	0	-0.467696711	0
5	0	-0.467696711	0	-0.467696711	0
6	0	-0.467696711	0	-0.467696711	0

Normalization:
In this column, I normalized the data of ZN variable with use of excel functions.
I create a function like this : $X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$

XLSTAT Check (Normalization):
In this column, I normalized the data of ZN variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (ZN Variable)

Outliers Detecting (Box-Plot Method)



Outliers :

All records of this variable, except zeros, are detected as outliers for this variable with box-plot method.
We know that we have 506 records and 380 of the are zeros, so as the result, box-plot method tells us that we have 126 outliers

Whiskers & Box :

As we saw on this variable descriptive statistics table, 1st quartile, median and 3rd quartile of this variable are all zero.
So; IQR (= 3rd quartile – 1st quartile) is also zero.
As result, whisker lines (3rd quartile + 1.5 IQR , 1st quartile – 1.5 IQR) would also be zero.
So; with box-plot method, all the values of this variable which are anything except zero, would be detected as outliers for this variable.

Conclusion :

Box-plot method for detecting outliers of ZN variable, considers all of values (except the mode) as outliers.
It does not seem wise if we rely on this method for detecting the outliers of this variable.
So; it would be better to try other method and rely on them instead of box-plot method.
I guess that z-score method would be a better approach for this purpose.

Statistical Analyses (ZN Variable)

Outliers Detecting (Z-Score Method)

	A	B
1	ZN (%	ZN (Z transformation
57	90	3.441077046
58	85	3.223922949
59	100	3.875385242
67	80	3.006768851
68	80	3.006768851
197	80	3.006768851
198	80	3.006768851
199	80	3.006768851
200	80	3.006768851
201	95	3.658231144
202	95	3.658231144
203	82.5	3.1153459
204	82.5	3.1153459
205	95	3.658231144
206	95	3.658231144
256	80	3.006768851
257	80	3.006768851
258	90	3.441077046
285	90	3.441077046
286	90	3.441077046
288	80	3.006768851
292	80	3.006768851
293	80	3.006768851
294	80	3.006768851
349	85	3.223922949
350	80	3.006768851
355	90	3.441077046
356	80	3.006768851

Average: 3.231678452 Count: 28 Min: 3.006768851 Max: 3.875385242 Sum: 90.48699666

Outliers With Z-Score Method:

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.

So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I got the result as you can see in the table.

28 Outliers:

28 outliers are detected based on Z-Score method.

While, the number of outliers which were detected based on box-plot method was 126.

As it is obvious and we mentioned on previous slide, Z-score method seems more reliable.

Box-Plot VS Z-Score :

When we compare the results of these two methods for detecting outliers for CRIM feature, there is a significant different.

With Box-Plot method we got 126 outliers

With Z-Score method we got 28 outliers

If we want to decide outliers of which method should rely on, I prefer to go with Z-Score method, cause each outlier detected by this method is also detected as outlier with box-plot method.

On the other hand, number of outliers with box-plot method for this feature are too much, approximately 25% of our samples. So; it does not seem wise to go with box-plot method in this situation.

Statistical Analyses (ZN Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, ZN variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.

So; I removed the outliers which were detected by Z-Score method, from the dataset, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.

So, as the conclusion, we find it out that we cannot convert the ZN variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

Anderson-Darling test (ZN (%)):	
A^2	105.130
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test After Removing Outliers :

This is the result of normality test of ZN variable after removing its outliers which were found by Z-score method.

As we can see, p-value is less than alpha, so; this variable does not follow a normal distribution even after removing its outliers.

Statistical Analyses (ZN Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:
I am going to check the correlation between ZN variable and target variable which is MEDV, these are both continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	ZN (%)	MEDV
ZN (%)	**1**	**0.362**
MEDV	**0.362**	**1**

Weak And Direct Correlation:
The correlation matrix and the value of 0.362 tells us that there is a direct correlation between these 2 variables.
Meaning that if one of the increase, the other one will increase too.
On the other hand, the absolute value would be 0.36, which indicates that the correlation is relatively weak.

p-values (Pearson):

Variables	ZN (%)	MEDV
ZN (%)	**0**	**<0.0001**
MEDV	**<0.0001**	**0**

Statistical Significance Of The Correlation:
The value is <0.0001 suggests that the correlation between ZN and MEDV is statistically significant and it is not due to random changes.

Coefficients of determination (Pearson):

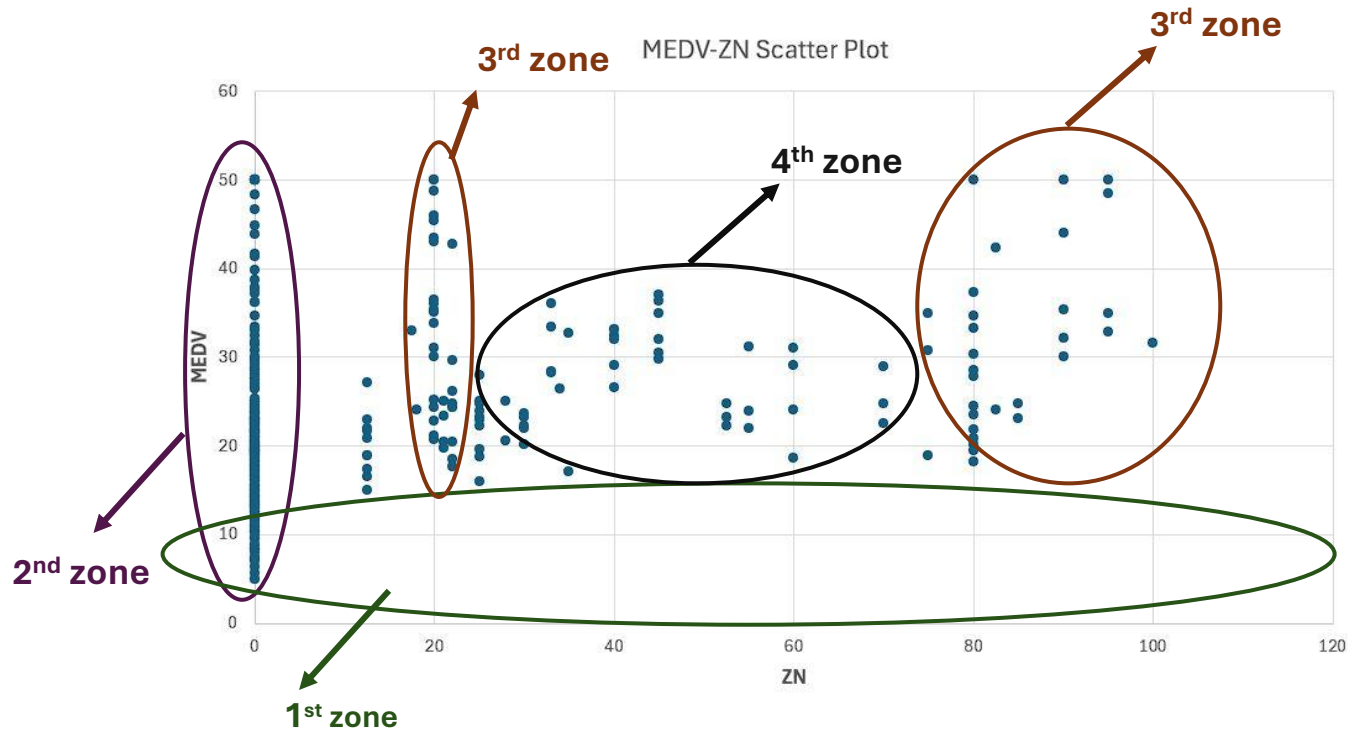
Variables	ZN (%)	MEDV
ZN (%)	**1**	0.131
MEDV	0.131	**1**

Power Of Prediction:
The value of 0.131 in this table, indicates that only 13.1% of the variance in target variable (MEDV) can be explained by the variance in ZN variable.

Statistical Analyses (ZN Variable)

Scatter Plot With The Target Variable

4 Zones:
As we see on the chart, we can divide the city into 4 zones based on median value of houses in each area and ZN rates of each area.
This gives us an interesting insight as we can interpreter as following :



4th Zone:
This zone contains ZN values between 20 to 80, and this zone cannot include the most expensive or the cheapest houses.
This zone must be for middle-class families

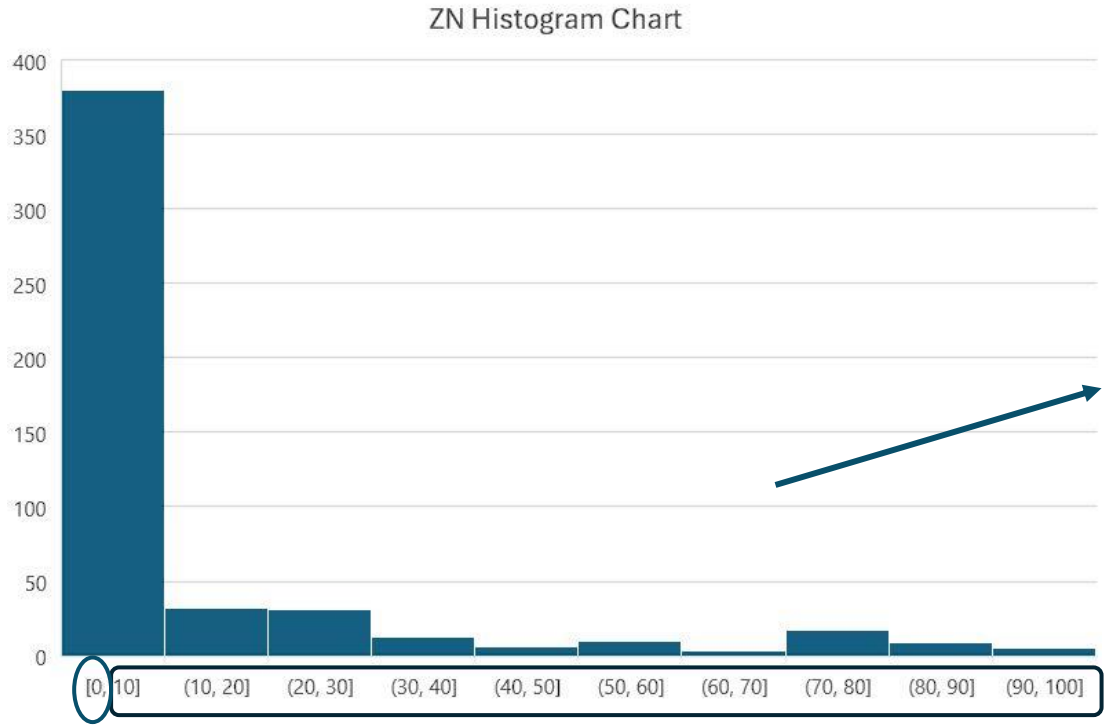
3rd Zone:
For areas with ZN rates above zero, only these two zones can include the most expensive houses.
For areas that ZN variable in them is exactly 20% or more than 80%, in addition to 1st zone, house prices can be the most expensive ones and include the richest families

2nd Zone:
This zone shows us that house prices in areas which ZN rate in them is equal to zero, can vary from the cheapest ones to the most expensive ones.
None of the other zones with different ZN rates, have this variety.
As soon as, ZN variable increases, house prices cannot be less than a minimum value.

1st Zone:
This zone contains the cheapest houses. Interesting point is that, in this range of house prices ZN variable does not extend from zero.
And only this range have this characteristic.
Meaning that, for lower than a specific value of MEDV variable, ZN variable cannot be anything except zero.
So; in areas that house prices are so cheap, there are not any lots over 25,000 sq. ft.
These areas are probably crowded with small houses.

Statistical Analyses (ZN Variable)

Question : Is There Any Difference Between Average Of House Prices Based On ZN Rates?



ZN Rate Above Zero :

I am going to create a new feature based on ZN.
This feature is going to be 0 for areas which ZN rate in them is equal to zero
And is going to be 1 for areas which ZN rate in them is above zero
And I am going to compare the average of house prices between these 2 classes.
The reason of choosing the value of zero is the distribution of ZN feature.
Cause as we can see on the chart, areas with ZN rates above this value are relatively rare and can be considered as less crowded areas.
So; it would be wise if we compare areas as we consider crowded and those which are not considered crowded (based on our definition of being crowded) in terms of house prices

	A	B	C
1	ZN (%)	ZN Binary Classification	MEDV
2	18	1	24
3	0	=IF(A3=0,0,1)	21.6
4	0	0	34.7
5	0	0	33.4
6	0	0	36.2
7	0	0	28.7
8	12.5	1	22.9

Statistical Analyses (ZN Variable)

Variances Equality Test (Leven's Method)

Variances Equality Test:

As we can see, P-value is greater than alpha, so; we should accept the null hypothesis. Meaning that variance of house prices with class 1 (those with ZN rates above zero) , is equal to variance of house prices with class 0 (those with ZN rates is equal to zero).

So; for comparing the average of house prices between these two classes, with should assume the equality of variances.

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	2.645
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	0.105
alpha	0.050

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of ZN variable, we should use the appropriate method which is Leven's test.

If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (CRIM Variable)

Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[-9.662, -6.213]

Difference	-7.937
t (Observed value)	-9.041
|t| (Critical value)	1.965
DF	504
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Higher Average Of House Prices:

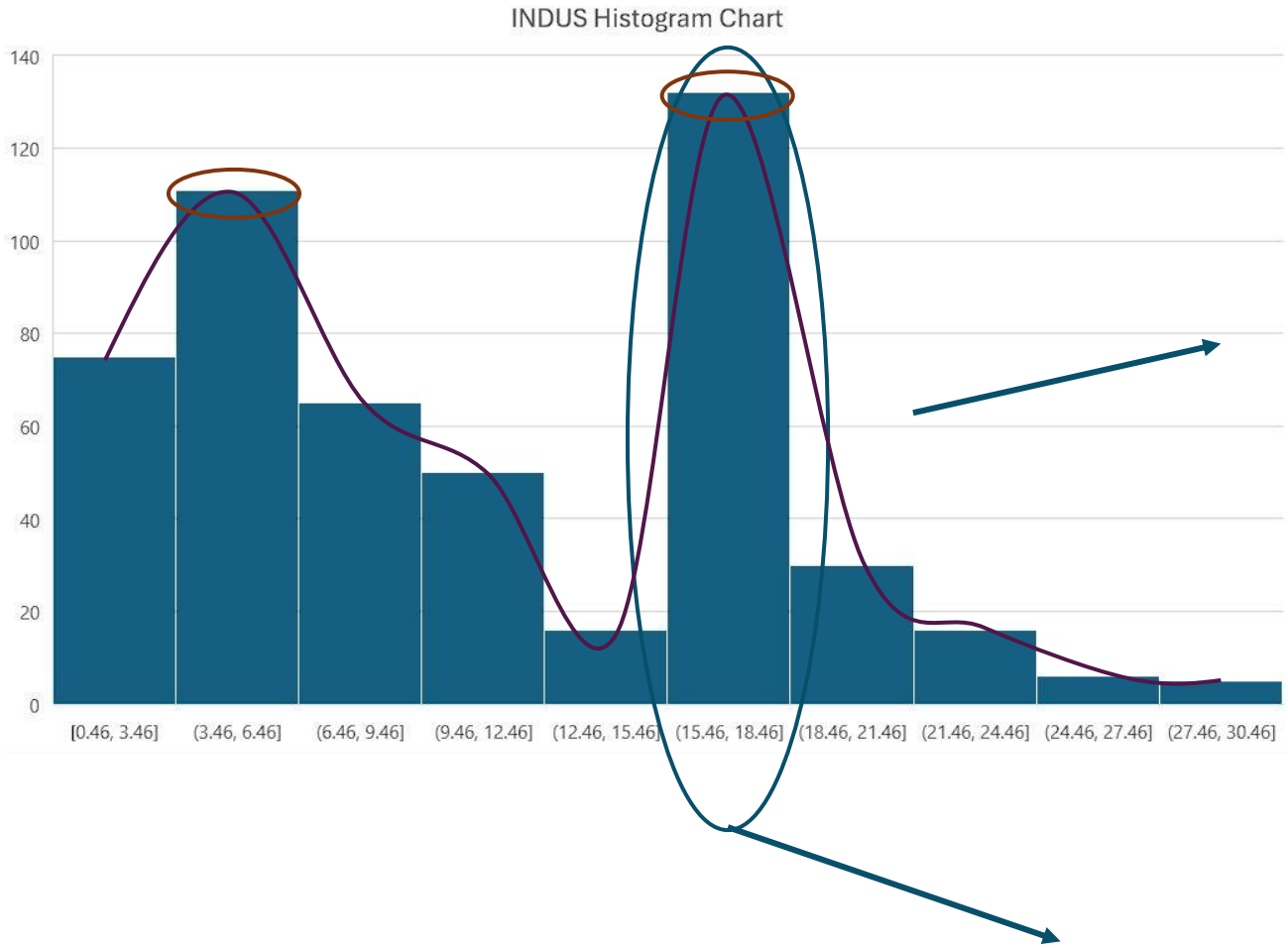
This tells us that areas of class 1, have higher house prices on average, comparing to areas of class 0

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have ZN rate equal to zero (class 0) is not equal to the average of house prices for those areas which have ZN rates greater than zero (class 1).
As we could guess.

Statistical Analyses (INDUS Variable)

Examining The Distribution



INDUS Histogram Chart:

The chart shows us that INDUS variable is multi-modal.
One problem can be finding the reason which makes this variable multi-modal.
If we could find the reason for this, we could make divide this variable into two, and then we had two distributions, which both of the were positively skewed.
Another insight that we can draw from this chart, is that we do not have any areas in Boston city which has more than 30.5% of industrial land.
On the other hand, areas with more than 21% of industrial land are relatively rare.

Mode:

The mode of INDUS feature must be something in this range.

Statistical Analyses (INDUS Variable)

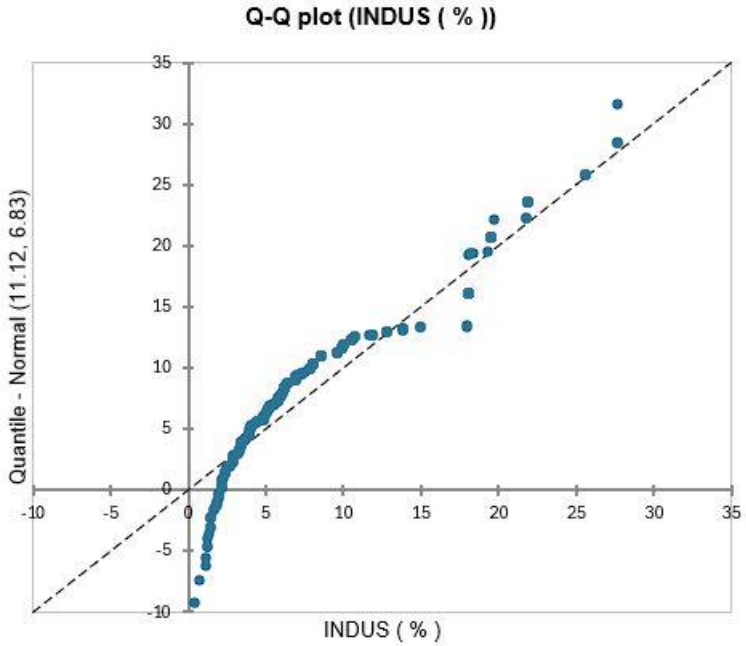
Examining The Descriptive Statistics

Statistic	INDUS (%)
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	0.460
Maximum	27.740
Freq. of minimum	1
Freq. of maximum	5
Range	27.280
1st Quartile	5.190
Median	9.690
3rd Quartile	18.100
Sum	5626.520
Mean	11.120
Variance (n)	46.693
Variance (n-1)	46.785
Standard deviation (n)	6.833
Standard deviation (n-1)	6.840
Skewness (Pearson)	0.295
Kurtosis (Pearson)	-1.241
Lower bound on mean (95%)	10.522
Upper bound on mean (95%)	11.717
Lower bound on variance (95%)	41.511
Upper bound on variance (95%)	53.138

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen only 1 time among all of the records
- Maximum value of this variable can be seen 5 times among all records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A skewness of 0.295 is close to 0, suggesting that the distribution is approximately symmetrical with only a slight positive skew. Since the value is greater than 0, it indicates a slight positive skew, meaning the right tail of the distribution is a bit longer or fatter than the left tail.
- Since the excess kurtosis (calculated as kurtosis - 3) is negative, your distribution has flatter tails and a broader peak compared to a normal distribution.
- The mean of the population of this variable must be something between 10.5 and 11.7 with confidence level of 95%
- The variance of the population of this variable must be something between 41.5 and 53.1 with confidence level of 95%

Statistical Analyses (INDUS Variable)

Normality Test (Anderson-Darling Method)

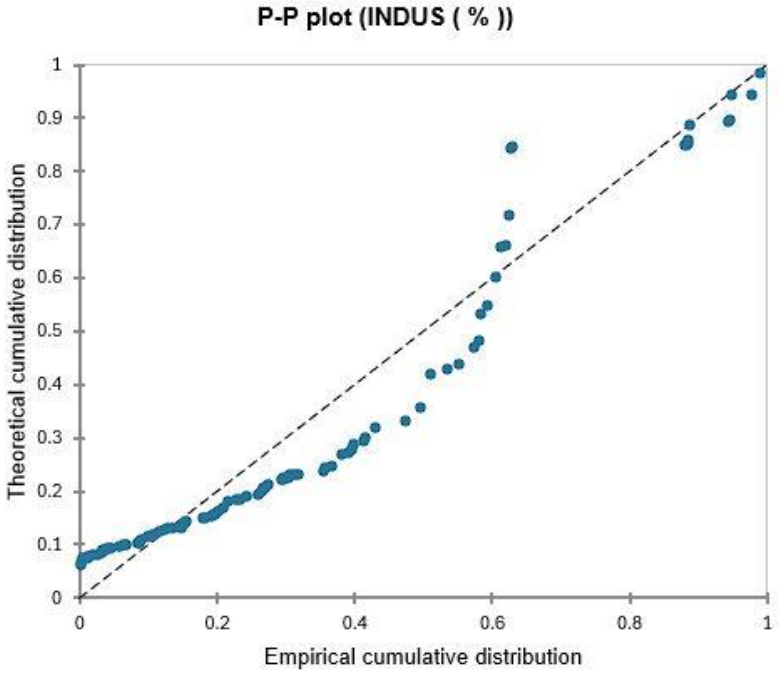


Normality Test Result :

p-value is less than alpha, so we should reject the null hypothesis. So; INDUS variable does not follow a normal distribution.

Considering the p-value, it is not even close to alpha, and we might never convert this variable's distribution, to a normal one. Even by transforming methods or removing outliers.

Anderson-Darling test (INDUS (%)):	
A^2	22.372
p-value (Two-tailed)	**<0.0001**
alpha	0.050



P-P & Q-Q plot:

Paying attention to these two charts, as we mentioned before, it does not seem that we can make this variable to follow a normal distribution.

Statistical Analyses (INDUS Variable)

Outliers Detecting (Variable Transformation)

Raw data:
This is the raw data of INDUS variable without any transformations.

Z-Score Normalization:
In this column, I transformed the data of INDUS variable with use of excel functions.
I create a function like this :

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

XLSTAT Check (Z-Score Normalization):
In this column, I transformed the data of INDUS variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

	A	B	C	D	E
1	INDUS (%)	INDUS (Z transformation)	INDUS (Normalization)	Standardized	0 to 1
2	2.31	-1.287961151	0.067815249	-1.287961151	0.067815249
3	7.07	-0.592050806	0.242302053	-0.592050806	0.242302053
4	7.07	-0.592050806	0.242302053	-0.592050806	0.242302053
5	2.18	-1.306967106	0.063049853	-1.306967106	0.063049853
6	2.18	-1.306967106	0.063049853	-1.306967106	0.063049853

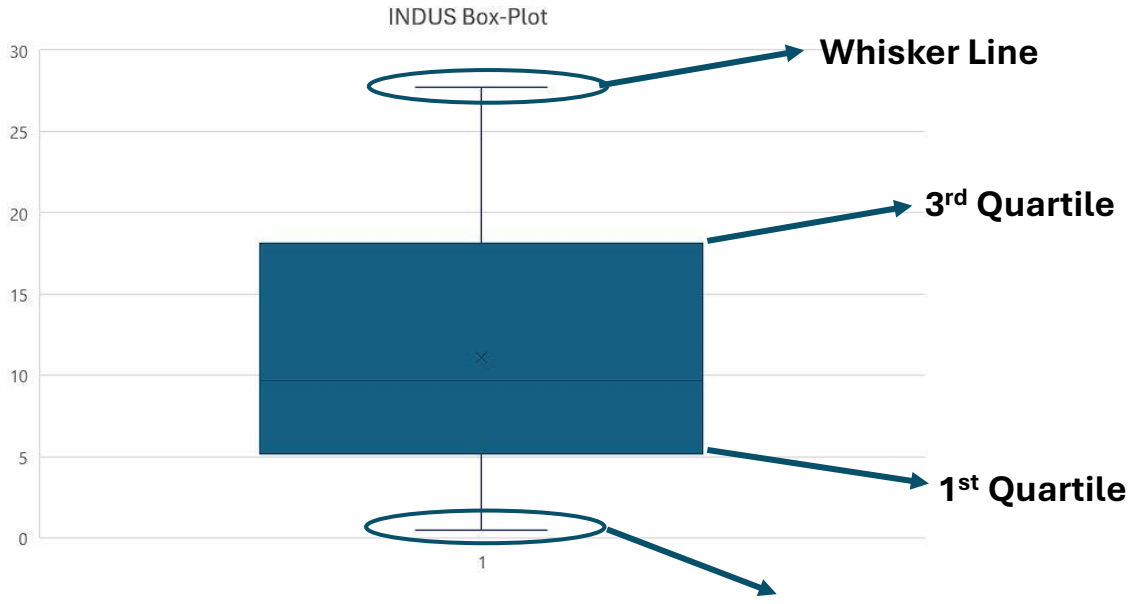
Normalization:
In this column, I normalized the data of INDUS variable with use of excel functions.
I create a function like this :

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

XLSTAT Check (Normalization):
In this column, I normalized the data of INDUS variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (INDUS Variable)

Outliers Detecting (Box-Plot Method)



1st Q	5.19
Median	9.69
3rd Q	18.1
IQR = (3rd Q - 1st Q)	12.91
3rd Q + 1.5IQR	37.465
1st Q - 1.5IQR	-14.175

Above Maximum:
As we can see, this value, which is a limit line, and any value above it should be considered as outlier; is greater than the maximum of INDUS variable
So, we would not have outliers between high values of INDUS variable

Below Minimum:
This value is another limit line, and any value below it, should be considered as outlier; is less than the minimum of INDUS feature
So, we would not have outliers between low values of INDUS feature

Box-plot chart:
In the analysis of the INDUS variable, representing the proportion of non-retail business acres per town, a box-plot chart was utilized to identify potential outliers. Interestingly, the box-plot did not detect any outliers for this variable. This indicates that the data points for INDUS are relatively consistent and fall within the expected range, without significant deviations that could be classified as outliers. The absence of outliers suggests a uniform distribution of non-retail business acres across the towns in the dataset, which implies stability and predictability in this feature. Consequently, the INDUS variable demonstrates a more homogeneous pattern compared to other variables that may exhibit higher variability and outliers.

Statistical Analyses (INDUS Variable)

Outliers Detecting (Z-Score Method)

INDUS (Z transformation)
-1.287961151
-0.592050806
-0.592050806
-1.306967106
-1.306967106
-1.306967106
-0.475091084
-0.475091084
-0.475091084
-0.475091084
-0.475091084
-0.475091084
-0.475091084
-0.475091084
-0.435617178
-0.435617178

No Outliers With Z-Score Method :

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.
So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I found out that there is no value between transformed data to be greater than 3 or less than -3

Box-Plot VS Z-Score :

In the analysis of the INDUS variable, which represents the proportion of non-retail business acres per town, both the Box-Plot and Z-Score methods were employed to detect potential outliers. Intriguingly, neither method identified any outliers in this variable, demonstrating a consistent result across different statistical approaches. The Box-Plot, which visually represents the data distribution and highlights outliers based on the interquartile range (IQR), showed no values beyond the whiskers. Similarly, the Z-Score method, which quantifies the number of standard deviations a data point is from the mean, confirmed that all INDUS values were within the typical range, with no Z-Scores exceeding the common threshold of 3. This congruence between the Box-Plot and Z-Score methods reinforces the stability and uniformity of the INDUS variable, indicating that the distribution of non-retail business acres per town is relatively homogeneous and free from significant anomalies. This reliable pattern across different detection techniques underscores the robustness of the INDUS variable in the dataset.

Statistical Analyses (INDUS Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, INDUS variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.

So; I transformed this variable with box-cox method, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.

So, as the conclusion, we find it out that we cannot convert the INDUS variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

Box-Cox transformation
1.008450961
3.071808319
3.071808319
0.926355882
0.926355882
0.926355882
3.326936276
3.326936276
3.326936276
3.326936276
3.326936276
3.326936276
3.326936276

Transformed data of INDUS Variable With Box-Cox Method

Normality Test After Box-Cox Transformation :

As we can see, the result of the normality test of transformed data (with box-cox method), INDUS variable still does not follow a normal distribution.

Anderson-Darling test (Box-Cox transformation):	
A^2	15.152
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Statistical Analyses (INDUS Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:
I am going to check the correlation between INDUS variable and target variable which is MEDV, these are both continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	INDUS (%)	MEDV
INDUS (%)	**1**	**-0.484**
MEDV	**-0.484**	**1**

Relatively Strong And Inverse Correlation:
The correlation matrix and the value of -0.48 tells us that there is an inverse correlation between these 2 variables.
Meaning that if one of the increase, the other one will decrease.
On the other hand, the absolute value would be 0.48, which indicates that the correlation is relatively strong.

p-values (Pearson):

Variables	INDUS (%)	MEDV
INDUS (%)	**0**	**<0.0001**
MEDV	**<0.0001**	**0**

Statistical Significance Of The Correlation:
The value is <0.0001 suggests that the correlation between INDUS and MEDV is statistically significant and it is not due to random changes.

Coefficients of determination (Pearson):

Variables	INDUS (%)	MEDV
INDUS (%)	**1**	0.235
MEDV	0.235	**1**

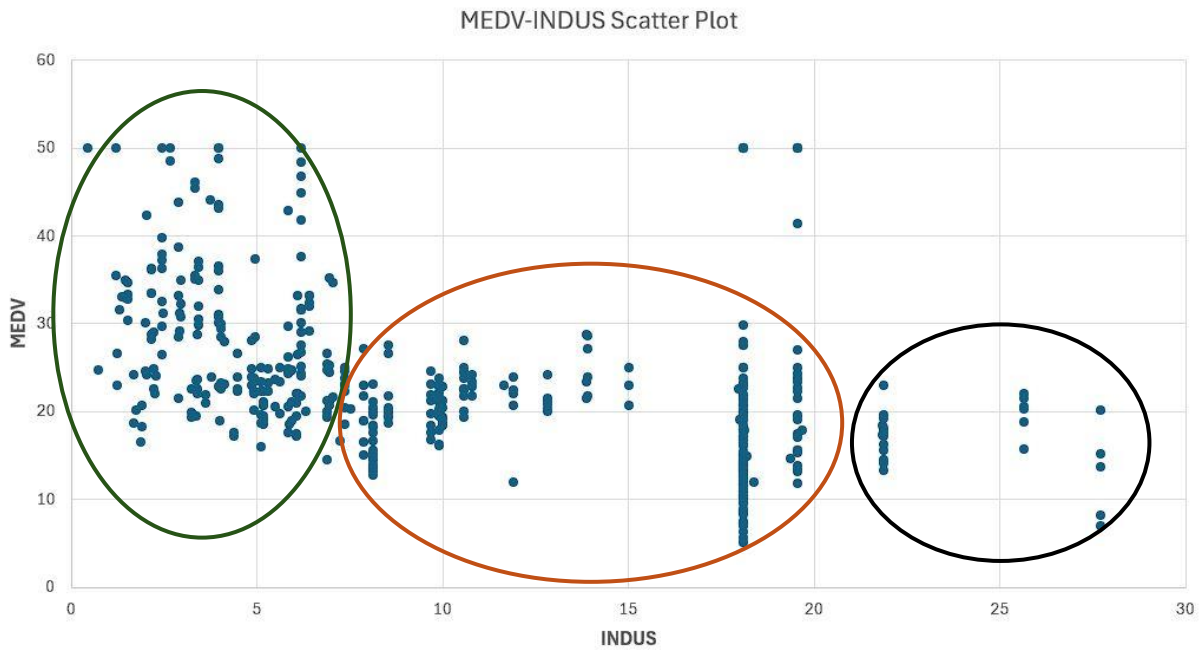
Power Of Prediction:
The value of 0.235 in this table, indicates that only 23.5% of the variance in target variable (MEDV) can be explained by the variance in INDUS variable.

Statistical Analyses (INDUS Variable)

Scatter Plot With The Target Variable

3 Zones:

If we pay attention to scatter chart of MEDV and INDUS variables, we can divide the chart into three zones.
We can extract some insights based on this scatter plot that we are going to talk about.



Green Zone:

This zone includes some records which are in areas that have the lowest values of INDUS variable.
Meaning that in these areas, proportion of land which are dedicated to industrial purposes is relatively less than the other two zones, that why we call this zone as green zone.
On the other hand, if we look at the chart carefully, we can see that the records of this zone, have higher house prices relatively to other zones.
This fact suggest us that houses with high values are mostly in areas which are not industrial.

Orange Zone:

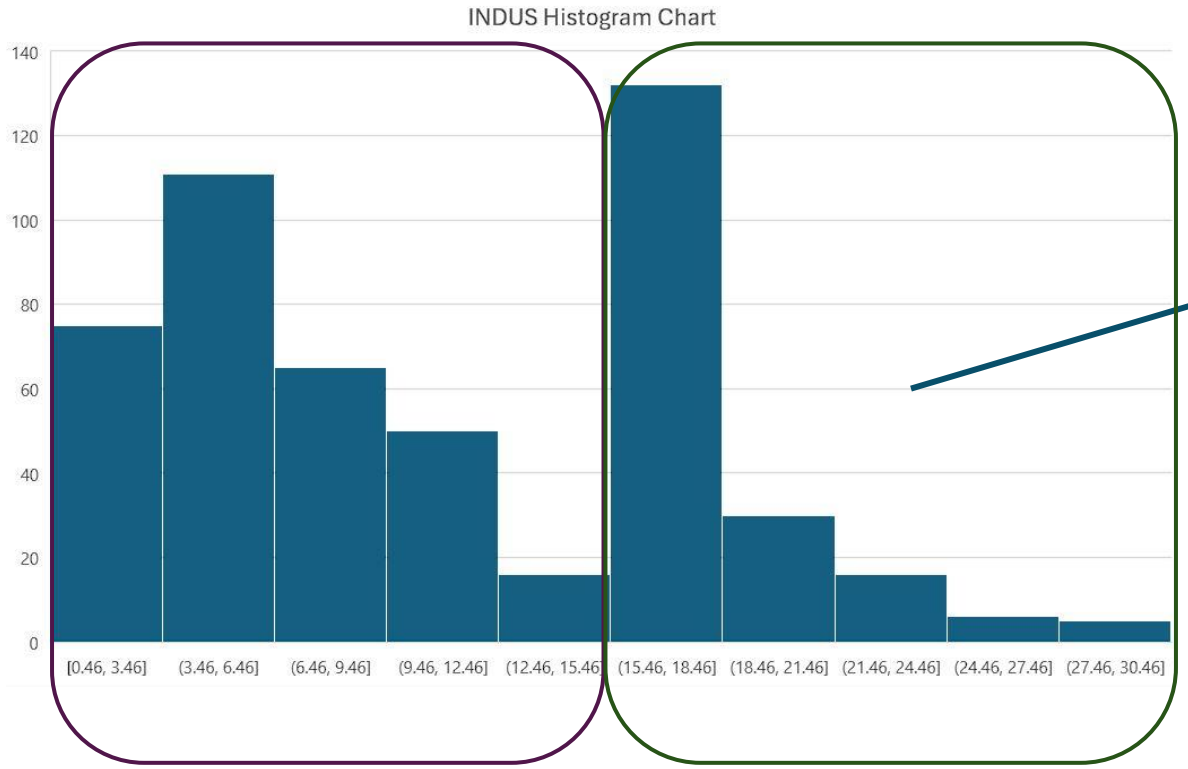
This zone includes areas which are not highly industrial
Proportion of industrial lands in these areas is lower than the black zone and greater than green zone.
As we can see on the chart, as a result of this fact, house prices in this zone is relatively lower than the house prices of the green zone, and there are some areas in this zone which have higher house prices than the black zone.

Black Zone:

This zone has the highest proportion of industrial land.
As a result; the concentration of records in this zone is less than the two others.
Also, the house prices is relatively lower in this zone in comparison to other zones.
We guess that NOX rate (which is another feature of our dataset) in this zone should be greater than the other areas.

Statistical Analyses (INDUS Variable)

Question : Is There Any Difference Between Average Of House Prices Based On INDUS Ratea?



INDUS Rate Above Zero :

I am going to create a new feature based on INDUS.
This feature is going to be 0 for areas which INDUS rate in them is less than 15.46 And is going to be 1 for areas which INDUS rate in them is above 15.46
And I am going to compare the average of house prices between these 2 classes.
The reason of choosing the value of 15.46 is the distribution of INDUS feature.
Cause as we can see on the chart, it seems that INDUS variable is made of two distributions.
And it might be interesting and includes some insights if we divide the INDUS variable into these two distributions.

	A	B
1	INDUS (%)	INDUS Binary Classification
2	2.31	=IF(A2>15.46,1,0)
3	7.07	0
4	7.07	0
5	2.18	0
6	2.18	0

Statistical Analyses (INDUS Variable)

Variances Equality Test (Leven's Method)

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	0.060
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	0.807
alpha	0.050

Variances Equality Test:

As we can see, P-value is greater than alpha, so; we should accept the null hypothesis. Meaning that variance of house prices with class 1 (those with INDUS rates above 15.46) , is equal to variance of house prices with class 0 (those with INDUS rates below 15.46).
So; for comparing the average of house prices between these two classes, with should assume the equality of variances.

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of INDUS variable, we should use the appropriate method which is Leven's test.
If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (INDUS Variable)

Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[6.319, 9.348]

Difference	7.833
t (Observed value)	10.163
|t| (Critical value)	1.965
DF	504
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Higher Average Of House Prices:

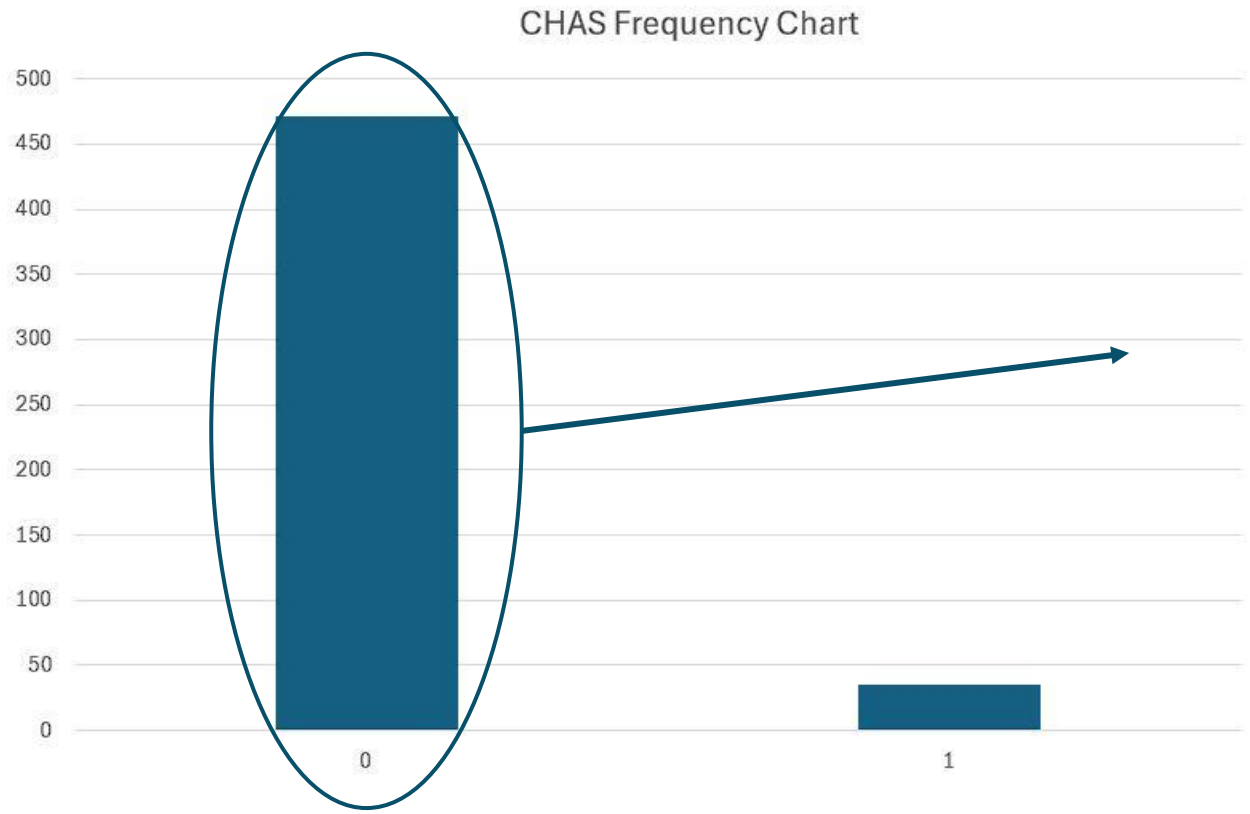
This tells us that areas of class 0, have higher house prices on average, comparing to areas of class 1

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have INDUS rates below 15.46 (class 0) is not equal to the average of house prices for those areas which have INDUS rates greater than 15.46 (class 1).

Statistical Analyses (CHAS Variable)

Examining The Distribution



CHAS Frequency Chart:

The charts shows us that only a few areas of Boston are near to Charles river.
As it was obvious.
This feature is imbalance so it will have an effect on our modeling purpose, unless; we make it balance.

Statistical Analyses (CHAS Variable)

Examining The Descriptive Statistics

Variable\ Statistic	Nbr. of observations	Nbr. of missing values	Sum of weights	Nbr. of categories	Mode	Mode frequency	Categories	Frequency per category	Rel. frequency per category (%)	Lower bound on frequencies (95%)	Upper bound on frequencies (95%)	Proportion per category	Lower bound on proportions (95%)	Upper bound on proportions (95%)
CHAS	506	0	506	2	0	471	0	471.000	93.083	90.872	95.294	0.931	0.909	0.953
							1	35.000	6.917	4.706	9.128	0.069	0.047	0.091

We are examining "CHAS" variable

There are 506 observations when we look at the "CHAS" column in our dataset

All of 506 rows of this column are filled with data and there are not any missing values

Sum of weights for a qualitative variable like CHAS does not give any meaningful information, because "CHAS" is not quantitative.

This shows us that "CHAS" variable contains 2 categories As we new before: 0, for areas near Charles river and 1, for areas which are not near Charles river

Category 0 is the mode, so its frequency is greater than 1

The frequency of the mode (which was for class 0) is 471

There are two categories in this variable column class 0 and class 1

There are 471 areas which are not near Charles river and 35 areas near Charles river

93% of areas are not near Charles river and 7% of areas are near Charles river

With a confidence level of 95%, we can say that on the population, areas of class 0, will make something between 90.8% to 95.29% of the population and this number for areas with class 1, will be something between 4.7% to 9.1%

These three columns don't give us any new information, they just give the same information of previous last three columns

Statistical Analyses (CHAS Variable)

Correlation Test With The Target Variable (Point-Biserial Method)

Why Point-Biserial Method:

I am going to check the correlation between CHAS variable and target variable which is MEDV, CHAS is a binary variable and MEDV is a continuous one, and because of this reason I should use appropriate corresponding method; which for checking the correlation between a binary and a continuous variable is Point-Biserial method.

Dividing The Target Variable :

I divide the target variable into two columns as you can see. One column shows the MEDV variable's values which their associated CHAS is equal to 0, and another one, where their associated CHAS is equal to 1.

	A	B
1	MEDV for CHAS = 0	MEDV for CHAS = 1
2	24	13.4
3	21.6	17
4	34.7	15.6
5	33.4	27
6	36.2	50

Calculating Point-Biserial Correlation :

For calculating point-biserial correlation, I used the formula as following :

$$r_{pb} = \frac{(M_1 - M_0) \cdot \sqrt{p \cdot q}}{s}$$

Mean of MEDV for **CHAS** =0 (μ0)	22.07452
Mean of MEDV for **CHAS** =1 (μ1)	28.7
proportion of cases where **CHAS** = 1 (p)	0.069
proportion of cases where **CHAS** = 0 (q)	0.931
standard deviation of **MEDV** (s)	9.197104
r (ponit-biserial)	0.182585

Interpretation Of R (Point-Biserial Correlation) :

The positive sign indicates that there is a positive relationship between the two variables. This means that homes closer to the Charles River (where CHAS = 1) tend to have higher median values compared to those further away (where CHAS = 0).

A coefficient of 0.1825 suggests a weak positive correlation. While there is some association between proximity to the Charles River and higher median home values, it is not a strong relationship. Other factors may also be influencing home values.

The positive but weak correlation might suggest that while proximity to the Charles River has some impact on home values, it is not a major determinant. Other features, such as overall location, amenities, and neighborhood characteristics, might also play crucial roles.

Biserial correlation (MEDV / CHAS) / Two-tailed test:	
r	0.183
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Statistical Significance Of The Correlation:

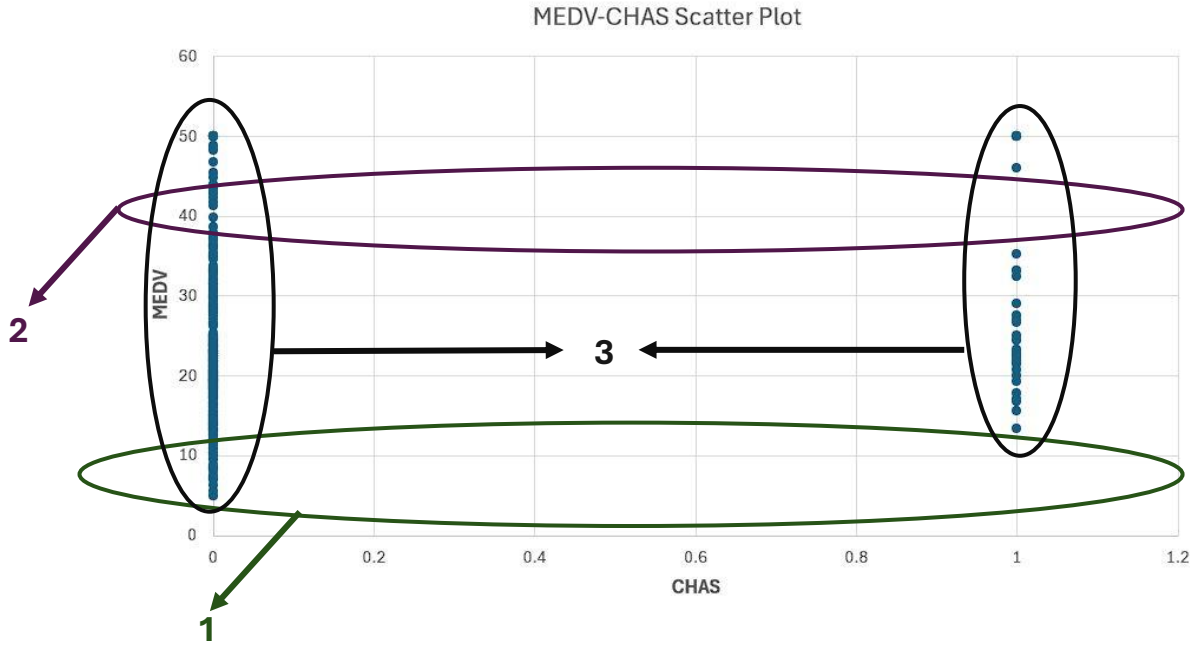
The value is <0.0001 suggests that the correlation between CHAS and MEDV is statistically significant and it is not due to random changes.

Statistical Analyses (CHAS Variable)

Scatter Plot With The Target Variable

Comparing Areas Near Charles River And Areas Which Are Not Near Charles River :

With paying attention to the scatter plot, we can draw some insights from it as are mentioned below :



1. Not Cheaper Than A Minimum :

If we look at the area which is marked with number 1, we can see that median value of house prices for those houses which are near the Charles river, have a minimum, which is higher than the minimum of median house prices of those houses which are not near the Charles river.

2. Houses Near The Charles River Do Not have This Range of Price :

If we look at the area which is marked with number 2, we can see that for a specific range of median value of house prices (around 40K), houses which are near the Charles rivers, experience a gap.

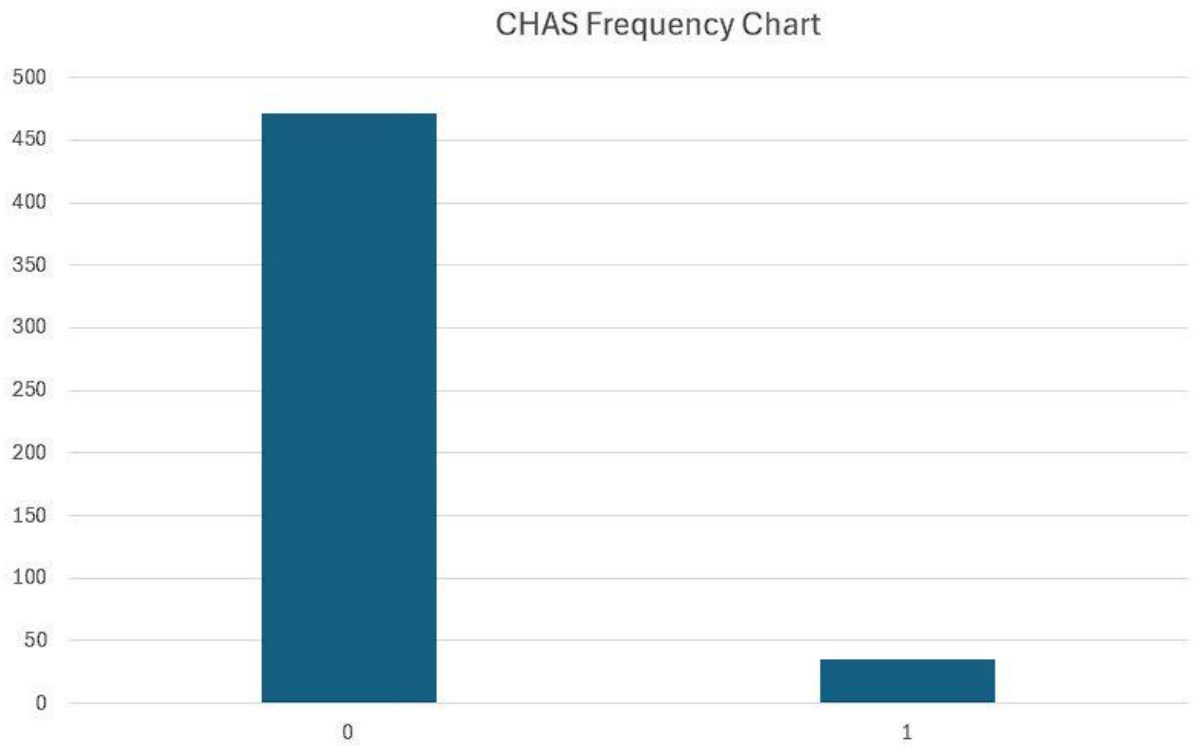
3. Price Concentration:

If we look at the areas which are marked with number 3, we can see that price concentration and variety for houses which CHAS feature for them is equal to 0 (they are not near the Charles river) is much higher than houses which CHAS feature for them is equal to 1 (they are near the Charles river).

Meaning that if you are looking for a house to buy, houses which are not near the Charles river, give you more options to choose among

Statistical Analyses (CHAS Variable)

Question : Is There Any Difference Between Average Of House Prices Based On CHAS Classification ?



CHAS Classification :

Comparing the mean MEDV (median value of owner-occupied homes) for tracts where CHAS = 1 (tract bounds the Charles River) and CHAS = 0 (tract does not bound the river) using parametric tests in XLSTAT is beneficial for several reasons. Firstly, this comparison helps us understand the impact of proximity to the Charles River on property values. By analyzing the differences in mean home values, we can gain insights into whether being near the river contributes to higher or lower housing prices. This information is valuable for urban planners, real estate investors, and potential homebuyers as it highlights the significance of location in real estate valuation. Secondly, identifying such correlations enables policymakers to make informed decisions regarding zoning and development around the river. It provides empirical evidence that can guide strategic planning and investment in infrastructure to enhance property values and community development. Lastly, from an analytical perspective, using parametric tests ensures that the results are statistically robust and reliable, offering a clear and quantifiable understanding of the relationship between the location of homes relative to the Charles River and their market values.

Statistical Analyses (CHAS Variable)

Variances Equality Test (Leven's Method)

Variances Equality Test:

As we can see, P-value is less than alpha, so; we should reject the null hypothesis. Meaning that variance of house prices with class 1 (CHAS = 1) , is not equal to variance of house prices with class 0 (CHAS = 2). So; for comparing the average of house prices between these two classes, with should not assume the equality of variances.

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	7.777
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	**0.005**
alpha	0.050

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of CHAS variable, we should use the appropriate method which is Leven's test.
If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (CHAS Variable)

Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[-10.689, -2.562]

Difference	-6.625
t (Observed value)	-3.304
|t| (Critical value)	2.026
DF	36.981
p-value (Two-tailed)	**0.002**
alpha	0.050

Higher Average Of House Prices:

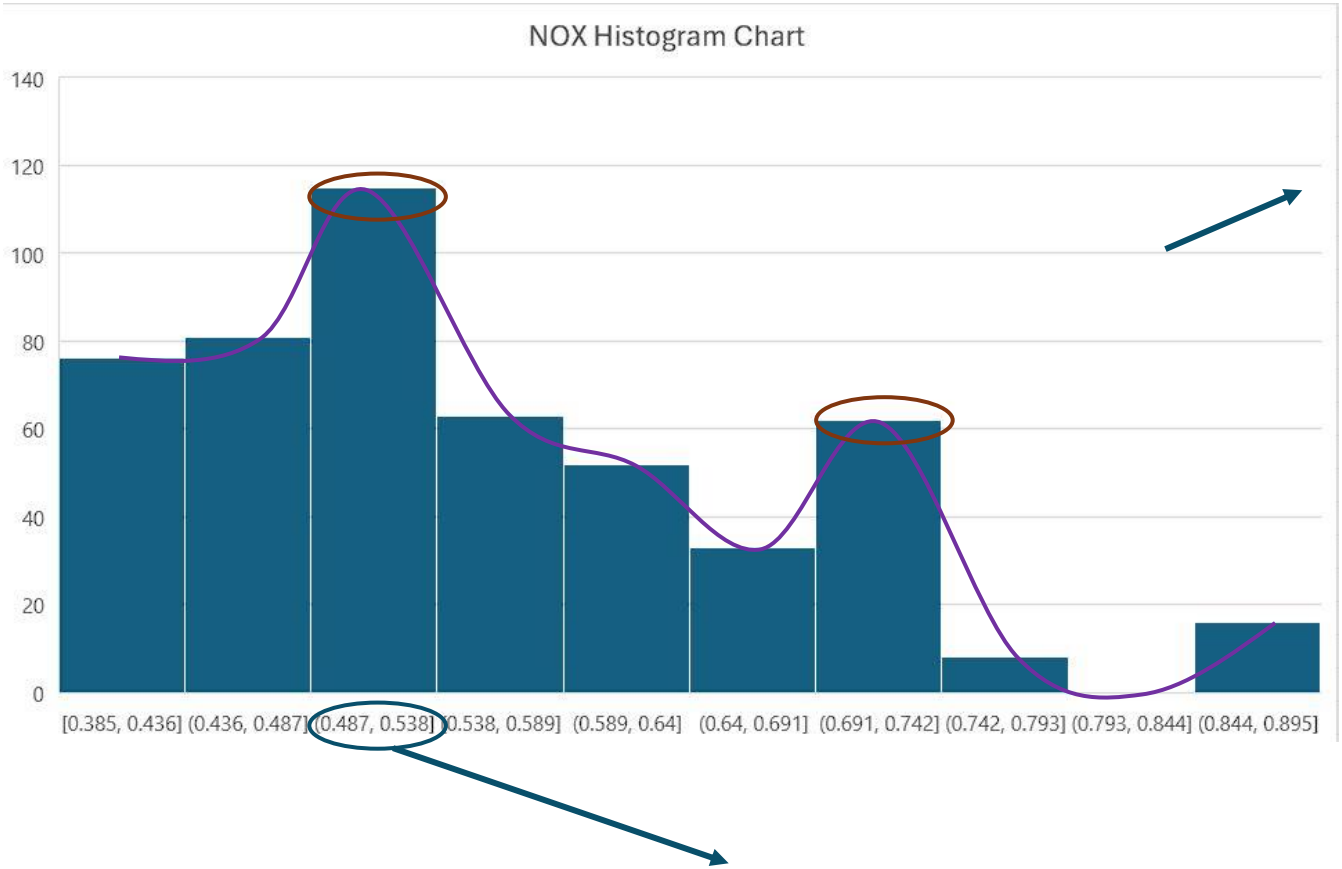
This tells us that areas of class 1, have higher house prices on average, comparing to areas of class 0

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have CHAS = 0 (class 0) is not equal to the average of house prices for those areas which have CHAS = 1 (class 1).

Statistical Analyses (NOX Variable)

Examining The Distribution



NOX Histogram Chart:

The NOX feature in the Boston Housing dataset represents the concentration of nitrogen oxides in the air, measured in parts per 10 million. The histogram of the NOX feature reveals a distribution that is multi-modal and slightly right-skewed, with most values clustered towards the lower end of the scale. This indicates that the majority of the areas in the dataset have relatively low levels of nitrogen oxides, which is a positive environmental indicator. However, there is a gradual tail extending towards higher values, reflecting some areas with elevated NOX concentrations. This slight positive skewness suggests the presence of outliers or specific regions with higher pollution levels. Understanding the distribution of the NOX feature is crucial for environmental assessment and urban planning, as it highlights areas that may require targeted pollution control measures. The histogram provides a clear visual representation, allowing for an immediate grasp of the data's central tendency and variability, essential for making informed decisions and analyses regarding air quality and its impact on housing values.

Mode:

The mode of ZN feature must be something in this range.

Statistical Analyses (NOX Variable)

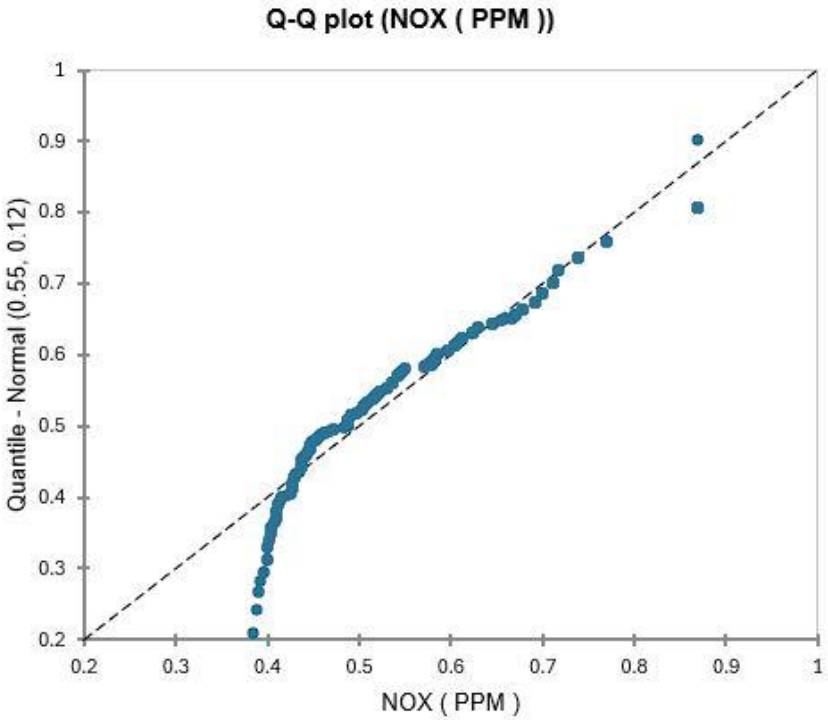
Examining The Descriptive Statistics

Statistic	NOX (PPM)
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	0.385
Maximum	0.871
Freq. of minimum	1
Freq. of maximum	16
Range	0.486
1st Quartile	0.449
Median	0.538
3rd Quartile	0.624
Sum	280.676
Mean	0.555
Variance (n)	0.013
Variance (n-1)	0.013
Standard deviation (n)	0.116
Standard deviation (n-1)	0.116
Skewness (Pearson)	0.727
Kurtosis (Pearson)	-0.076
Lower bound on mean (95%)	0.545
Upper bound on mean (95%)	0.565
Lower bound on variance (95%)	0.012
Upper bound on variance (95%)	0.015

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen only 1 time among all of the records
- Maximum value of this variable can be seen 16 times among all of the records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- The distribution has more data points concentrated on the left side, with a tail extending towards higher values on the right side.
- The distribution has a peak that is slightly less sharp compared to a normal distribution.
- The mean of the population of this variable must be something between 0.54 and 0.56 with confidence level of 95%
- The variance of the population of this variable must be something between 0.012 and 0.015 with confidence level of 95%

Statistical Analyses (NOX Variable)

Normality Test (Anderson-Darling Method)



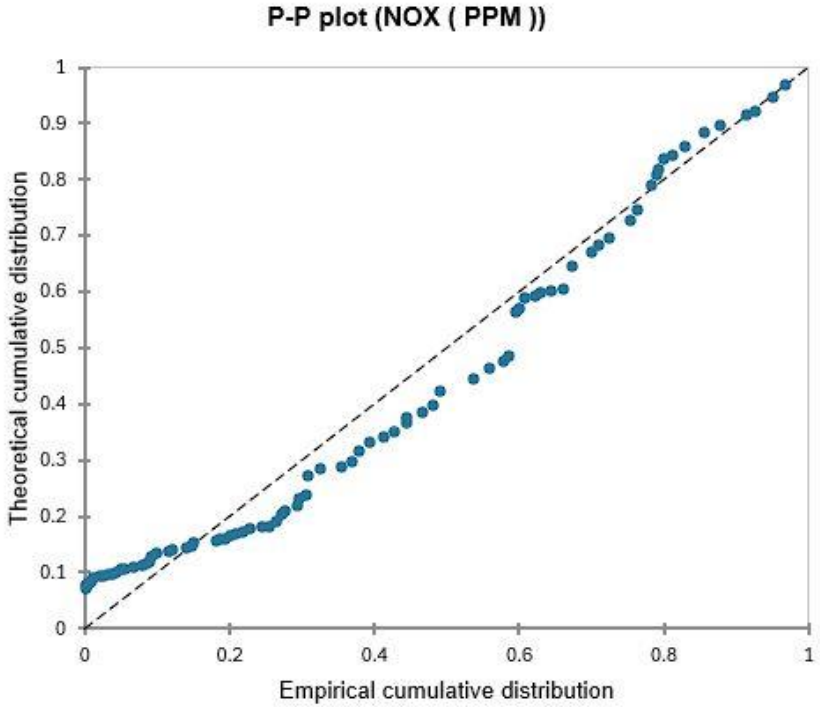
Normality Test Result :

p-value is less than alpha, so we should reject the null hypothesis. So; NOX variable does not follow a normal distribution.

Anderson-Darling test (NOX (PPM)):	
A^2	8.338
p-value (Two-tailed)	**<0.0001**
alpha	0.050

P-P & Q-Q plot:

Paying attention to these two charts, there might be some hope for converting this variable to a normally distributed variable.



Statistical Analyses (NOX Variable)

Outliers Detecting (Variable Transformation)

Raw data:

This is the raw data of NOX variable without any transformations.

Z-Score Normalization:

In this column, I transformed the data of NOX variable with use of excel functions.
I create a function like this : $X_{standardized} = \frac{X - \mu}{\sigma}$

XLSTAT Check (Z-Score Normalization):

In this column, I transformed the data of NOX variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

	A	B	C	D	E
1	NOX (PPM)	NOX (Z transformation)	NOX (Normalization)	Standardized (n-1)	0 to 1
2	0.538	-0.144074855	0.314814815	-0.144074855	0.314814815
3	0.469	-0.739530361	0.172839506	-0.739530361	0.172839506
4	0.469	-0.739530361	0.172839506	-0.739530361	0.172839506
5	0.458	-0.83445805	0.150205761	-0.83445805	0.150205761
6	0.458	-0.83445805	0.150205761	-0.83445805	0.150205761

Normalization:

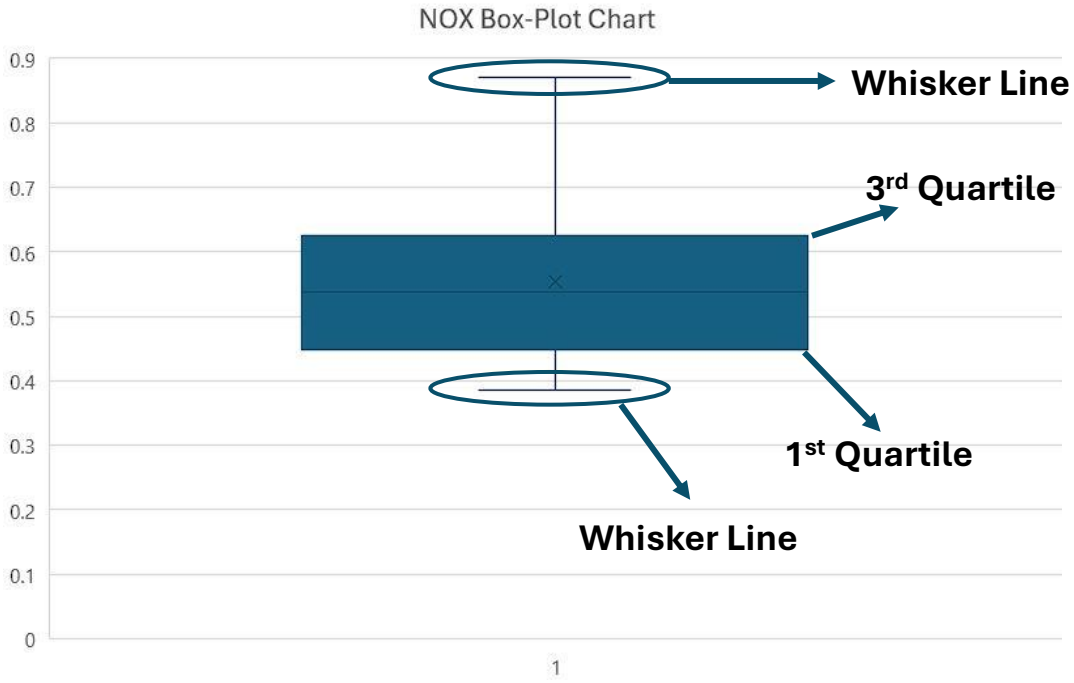
In this column, I normalized the data of NOX variable with use of excel functions.
I create a function like this : $X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$

XLSTAT Check (Normalization):

In this column, I normalized the data of NOX variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (NOX Variable)

Outliers Detecting (Box-Plot Method)



1st Q	0.449
Median	0.538
3rd Q	0.624
IQR = (3rd Q - 1st Q)	0.175
3rd Q + 1.5IQR	0.8865
1st Q - 1.5IQR	0.1865

Above Maximum:
As we can see, this value, which is a limit line, and any value above it should be considered as outlier; is greater than the maximum of NOX variable
So, we would not have outliers between high values of NOX variable

Below Minimum:
This value is another limit line, and any value below it, should be considered as outlier; is less than the minimum of NOX feature
So, we would not have outliers between low values of NOX feature

Box-plot chart:
In our analysis of the NOX variable, which quantifies the concentration of nitrogen oxides in the atmosphere, we used a box-plot chart to identify potential outliers. The resulting box-plot revealed no outliers, suggesting a consistent and stable distribution of NOX levels across the dataset. This lack of outliers indicates that most observations fall within a predictable range, reflecting uniform air quality within the surveyed areas. Such a result is beneficial for environmental monitoring, as it demonstrates minimal extreme values in nitrogen oxide concentrations.

Statistical Analyses (NOX Variable)

Outliers Detecting (Z-Score Method)

NOX (Z transformation)
-0.144074855
-0.739530361
-0.739530361
-0.83445805
-0.83445805
-0.83445805
-0.264891914
-0.264891914
-0.264891914
-0.264891914
-0.264891914
-0.264891914

No Outliers With Z-Score Method :

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.

So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I found out that there is no value between transformed data to be greater than 3 or less than -3

Box-Plot VS Z-Score :

In our analysis of the NOX variable, representing the concentration of nitrogen oxides in the air, both the Box-Plot and Z-Score methods were employed to identify potential outliers. Remarkably, neither method detected any outliers for this variable. The Box-Plot, which visually displays the spread and central tendency of the data using quartiles, confirmed that all NOX values fell within the whiskers, indicating no significant deviations. Similarly, the Z-Score method, which measures how many standard deviations a data point is from the mean, found all NOX values to be within the common threshold (typically, $Z < 3$), reinforcing the absence of outliers. This agreement between the Box-Plot and Z-Score methods highlights the consistency and reliability of the NOX data, suggesting a uniform distribution of nitrogen oxides concentrations. Such findings are crucial for environmental monitoring and urban planning, as they indicate stable air quality levels across the surveyed areas, allowing for more focused and effective pollution control strategies.

Statistical Analyses (NOX Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, NOX variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.

So; I transformed this variable with box-cox method, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.

So, as the conclusion, we find it out that we cannot convert the NOX variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

Box-Cox transformation
-0.834416374
-1.09240764
-1.09240764
-1.140399632
-1.140399632
-1.140399632
-0.881493567

Transformed data of NOX Variable With Box-Cox Method

Anderson-Darling test (Box-Cox transformation):	
A^2	4.186
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test After Box-Cox Transformation :

As we can see, the result of the normality test of transformed data (with box-cox method), NOX variable still does not follow a normal distribution.

Statistical Analyses (NOX Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:

I am going to check the correlation between NOX variable and target variable which is MEDV, these are both continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	NOX (PPM)	MEDV (1000$)
NOX (PPM)	**1**	**-0.427**
MEDV (1000$)	**-0.427**	**1**

Relatively Strong And Inverse Correlation:

The correlation matrix and the value of -0.42 tells us that there is an inverse correlation between these 2 variables.
Meaning that if one of the increase, the other one will decrease.
On the other hand, the absolute value would be 0.42, which indicates that the correlation is relatively strong.

p-values (Pearson):

Variables	NOX (PPM)	MEDV (1000$)
NOX (PPM)	**0**	**<0.0001**
MEDV (1000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:

The value is <0.0001 suggests that the correlation between NOX and MEDV is statistically significant and it is not due to random changes.

Coefficients of determination (Pearson):

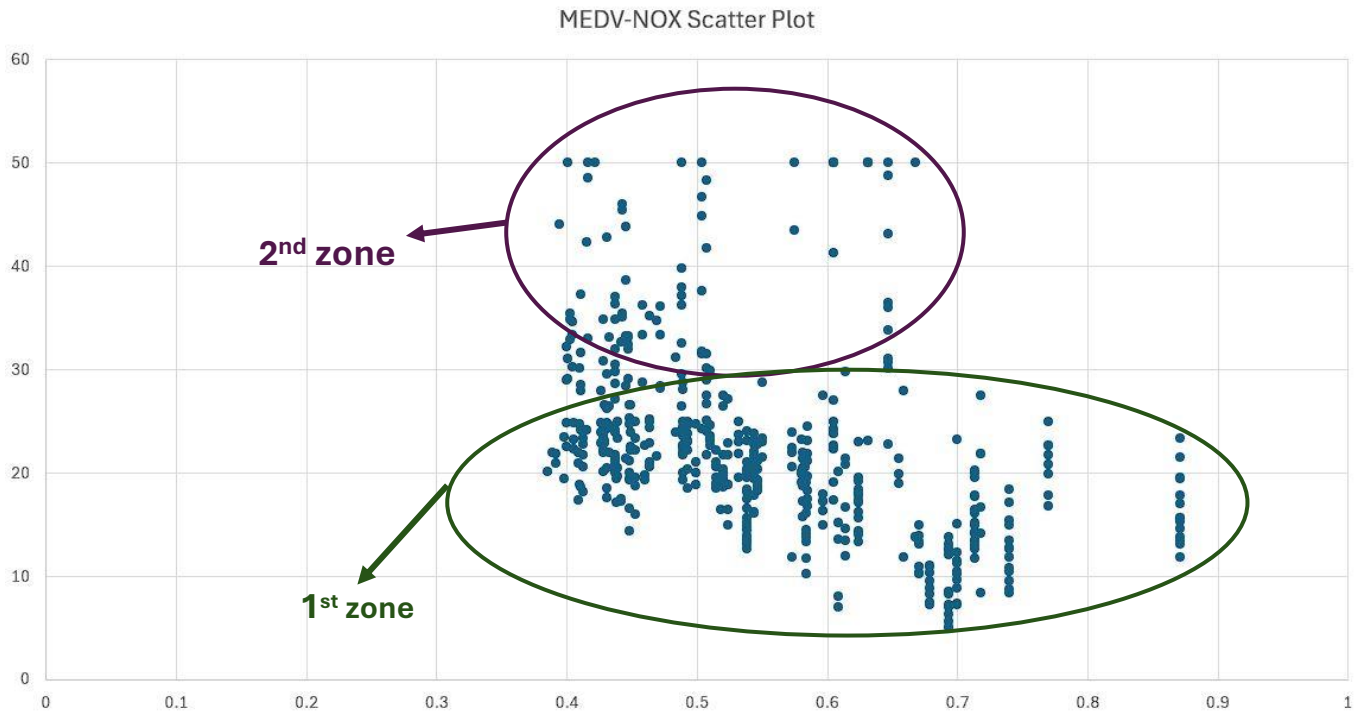
Variables	NOX (PPM)	MEDV (1000$)
NOX (PPM)	**1**	0.183
MEDV (1000$)	0.183	**1**

Power Of Prediction:

The value of 0.183 in this table, indicates that only 18.3% of the variance in target variable (MEDV) can be explained by the variance in NOX variable.

Statistical Analyses (NOX Variable)

Scatter Plot With The Target Variable



2 Zones:
As we see on the chart, we can divide the city into 2 zones based on median value of houses in each area and NOX rates of each area.
This gives us an interesting insight as we can interpreter as following :

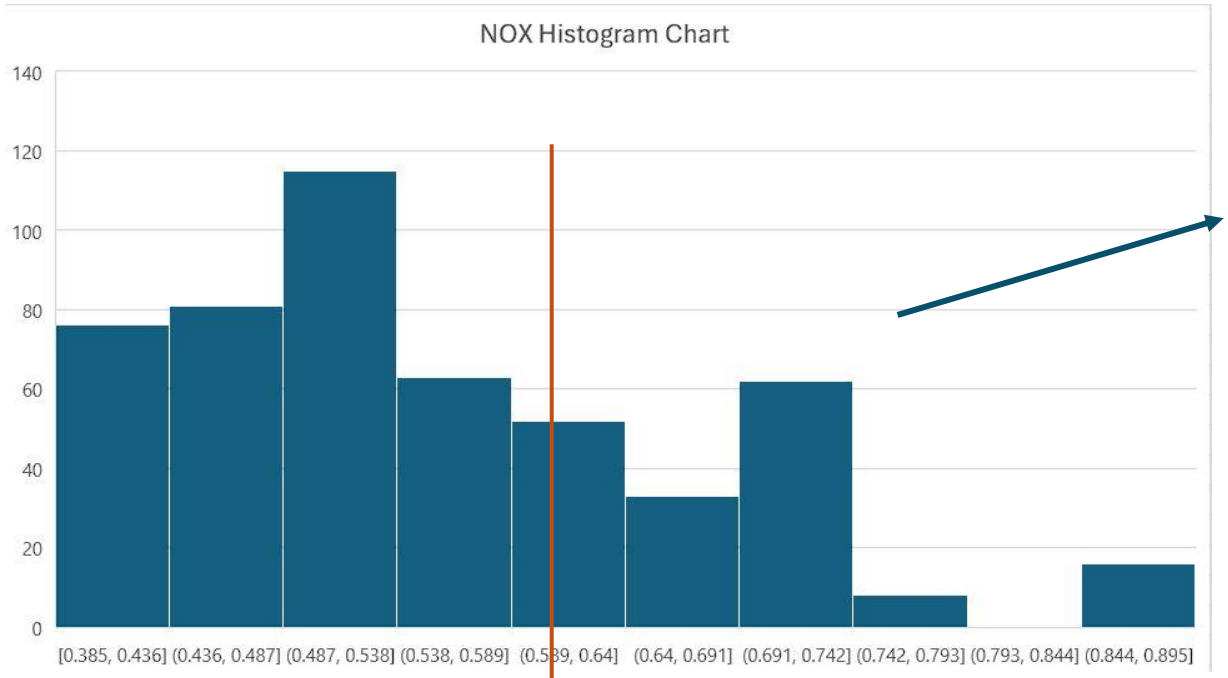
1st Zone:
This zone shows the MEDVs below a specific amount (around 30K), as we can see on the chart, for this zone, NOX variable can range from its minimum to its maximum.
Meaning that houses which are relatively cheap, can be found in different areas with different NOX rates.

2nd Zone:
This zone shows the MEDVs above a specific amount (around 30K), as we can see on the chart, for this zone, NOX variable cannot accept any values and it does not extend from a specific NOX rate (around 0.65).
Meaning that houses which are relatively expensive, are in areas which the NOX rete for them is lower, and they are cleaner areas.

Correlation Between NOX & MEDV:
As we can see on the chart, there is an inverse correlation between these two variables.
We can see that, while NOX rate increases, MEDV is decreasing.

Statistical Analyses (NOX Variable)

Question : Is There Any Difference Between Average Of House Prices Based On NOX Rates?



Average Of NOX Feature :

I am going to create a new feature based on NOX.
This feature is going to be 0 for areas which NOX rate in them is less than average
And is going to be 1 for areas which NOX rate in them is above the average
And I am going to compare the average of house prices between these 2 classes.
It would be a good practice if we divide the MEDV feature into these two classes.
We can consider the class 0, as clean areas, and class 1 , as areas which are not relatively clean

Average :

This line shows the average value for the NOX feature which is 0.555

	A	B
1	NOX (PPM)	NOX Binary Classification
2	0.538	=IF(A2>0.555,1,0)
3	0.469	0
4	0.469	0
5	0.458	0
6	0.458	0

Statistical Analyses (NOX Variable)

Variances Equality Test (Leven's Method)

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	4.871
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	**0.028**
alpha	0.050

Variances Equality Test:

As we can see, P-value is less than alpha, so; we should reject the null hypothesis. Meaning that variance of house prices with class 1 (those with NOX rate above the average) , is not equal to variance of house prices with class 0 (those with NOX rate below the average).

So; for comparing the average of house prices between these two classes, with should not assume the equality of variances.

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of NOX variable, we should use the appropriate method which is Leven's test.

If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (NOX Variable)

Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[4.671, 7.903]

Difference	6.287
t (Observed value)	7.651
|t| (Critical value)	1.966
DF	365.081
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Higher Average Of House Prices:

This tells us that areas of class 0, have higher house prices on average, comparing to areas of class 1
Meaning that on average, houses which are in more clean areas, have higher prices.

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have NOX rates below the average (class 0) is not equal to the average of house prices for those areas which have NOX rates above the average (class 1).

Statistical Analyses (NOX Variable)

The Best Fitting Distribution

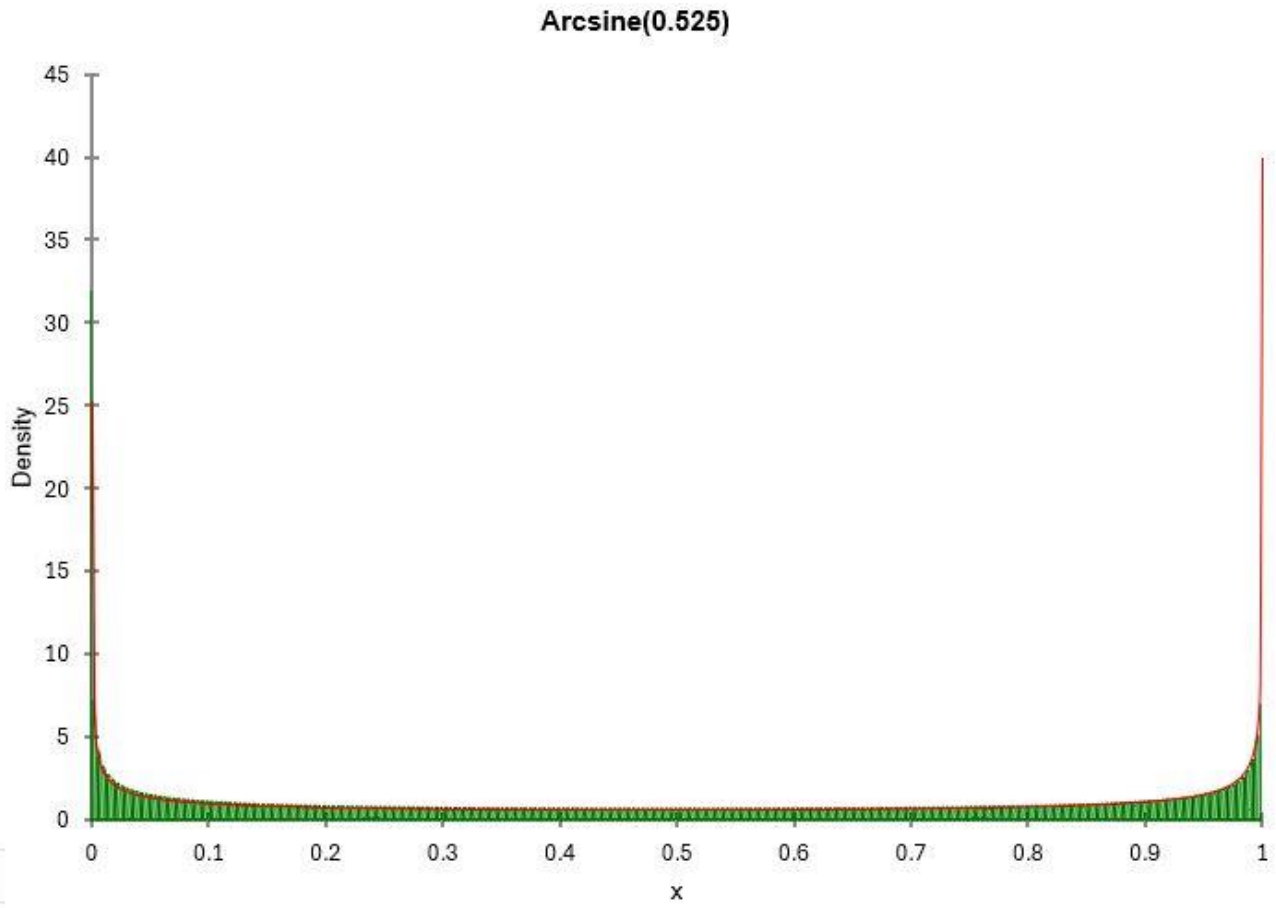
Distribution	p-value
Arcsine	**1.000**
Beta	<0.0001
Chi-square	<0.0001
Erlang	<0.0001
Fisher-Tippett (1)	<0.0001
Fisher-Tippett (2)	0.001
Gamma (2)	0.002
GEV	<0.0001
Gumbel	<0.0001
Log-normal	0.002
Logistic	0.001
Normal	<0.0001
Student	<0.0001
Weibull (1)	<0.0001
Weibull (2)	<0.0001

Arcsine Distribution :

With use of XLSTAT, I found out that the best fitting distribution for NOX variable, is arcsine distribution with given parameter as below (α) Then again, with use of XLSTAT I plot the distribution with this parameters and its corresponding value and I got the chart which you can see on the right, which seems so suit for NOX variable considering this variable's distribution.

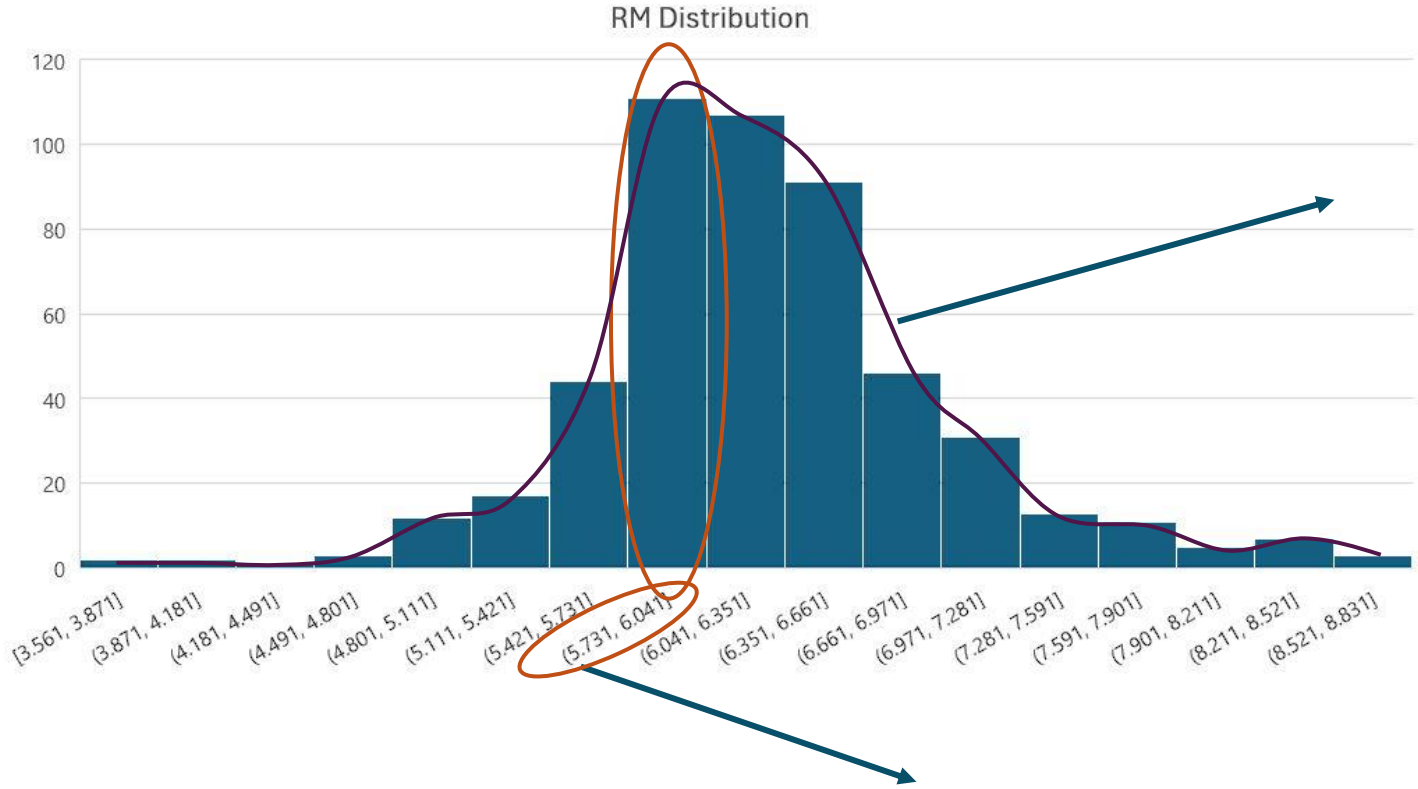
Estimated parameter (Arcsine):

Parameter	Value	Standard error
alpha	0.525	0.014



Statistical Analyses (RM Variable)

Examining The Distribution



Mode:

The mode of RM feature must be something in this range.

RM Histogram Chart:

It seems that RM variable follows a normal distribution.
Outliers might be from both sides of this distribution.
3 middle bars are so much more frequent than the others, showing that number of rooms which are in this range, is for typical houses. Most of houses have same number of rooms.
This feature can be a good indicator of house areas.
Right tail is thicker than the left one
Showing that in comparison of these two categories, houses which have relatively more number of rooms than a normal, are more frequent than those which have less number of rooms than normal.
The distribution is uni-modal
The single peak represents the central tendency of the dataset
The distribution also looks moderately symmetric

Statistical Analyses (RM Variable)

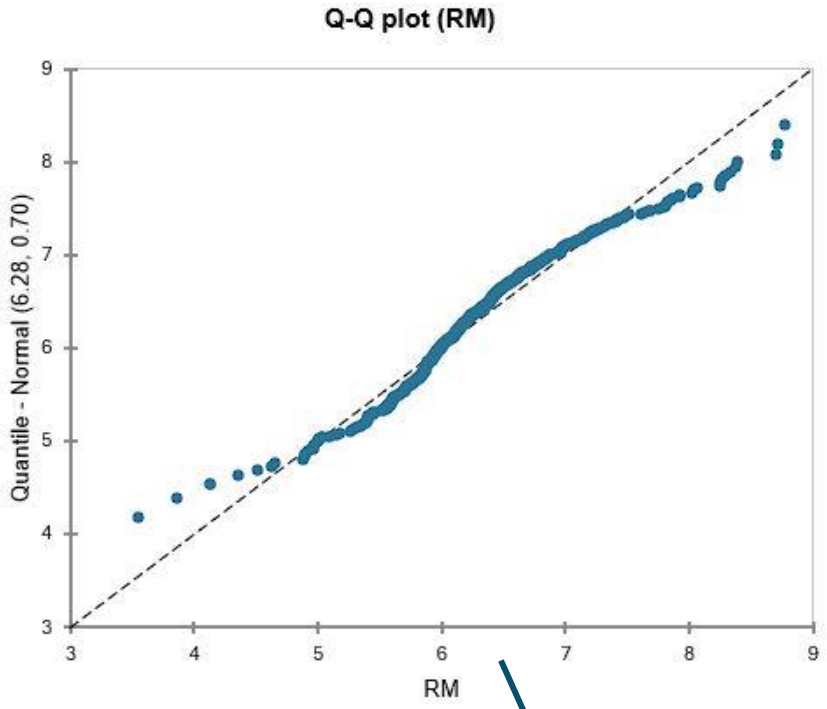
Examining The Descriptive Statistics

Statistic	RM
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	3.561
Maximum	8.780
Freq. of minimum	1
Freq. of maximum	1
Range	5.219
1st Quartile	5.886
Median	6.209
3rd Quartile	6.624
Sum	3180.025
Mean	6.285
Variance (n)	0.493
Variance (n-1)	0.494
Standard deviation (n)	0.702
Standard deviation (n-1)	0.703
Skewness (Pearson)	0.402
Kurtosis (Pearson)	1.861
Lower bound on mean (95%)	6.223
Upper bound on mean (95%)	6.346
Lower bound on variance (95%)	0.438
Upper bound on variance (95%)	0.561

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen only 1 time among all of the records
- Maximum value of this variable can be seen only 1 time among all records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A value of 0.402 indicates a slight positive skew, meaning the data is not perfectly symmetrical but leans slightly towards higher values..
- The distribution has a sharper peak around the mean, indicating a higher concentration of values near the center. The distribution also has fatter tails, suggesting more extreme values or outliers than a normal distribution.
- The mean of the population of this variable must be something between 6.2 and 6.3 with confidence level of 95%
- The variance of the population of this variable must be something between 0.43 and 0.56 with confidence level of 95%

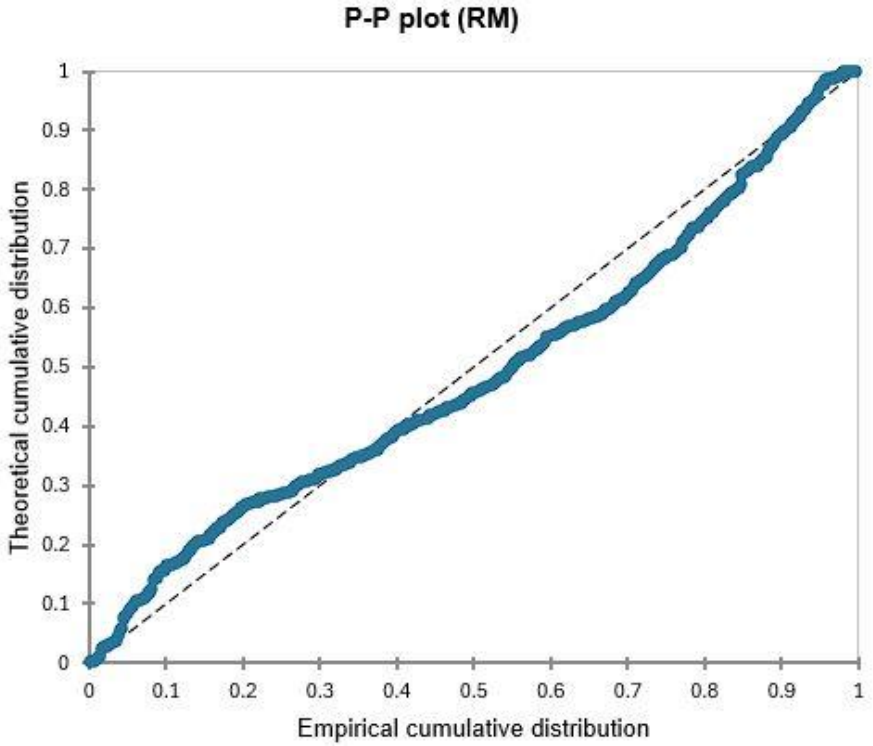
Statistical Analyses (RM Variable)

Normality Test (Anderson-Darling Method)



Anderson-Darling test (RM):	
A^2	6.118
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test Result :
p-value is less than alpha, so we should reject the null hypothesis. So; RM variable does not follow a normal distribution.



P-P & Q-Q plot:
Paying attention to these two charts, it seems that there is hope for converting RM variable to a normally distributed one.
Probably with removing outliers or with box-cox transformation we can achieve a normal distribution.

Statistical Analyses (RM Variable)

Outliers Detecting (Variable Transformation)

Raw data:
This is the raw data of RM variable without any transformations.

Z-Score Normalization:
In this column, I transformed the data of RM variable with use of excel functions.
I create a function like this : $X_{standardized} = \frac{X - \mu}{\sigma}$

XLSTAT Check (Z-Score Normalization):
In this column, I transformed the data of RM variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

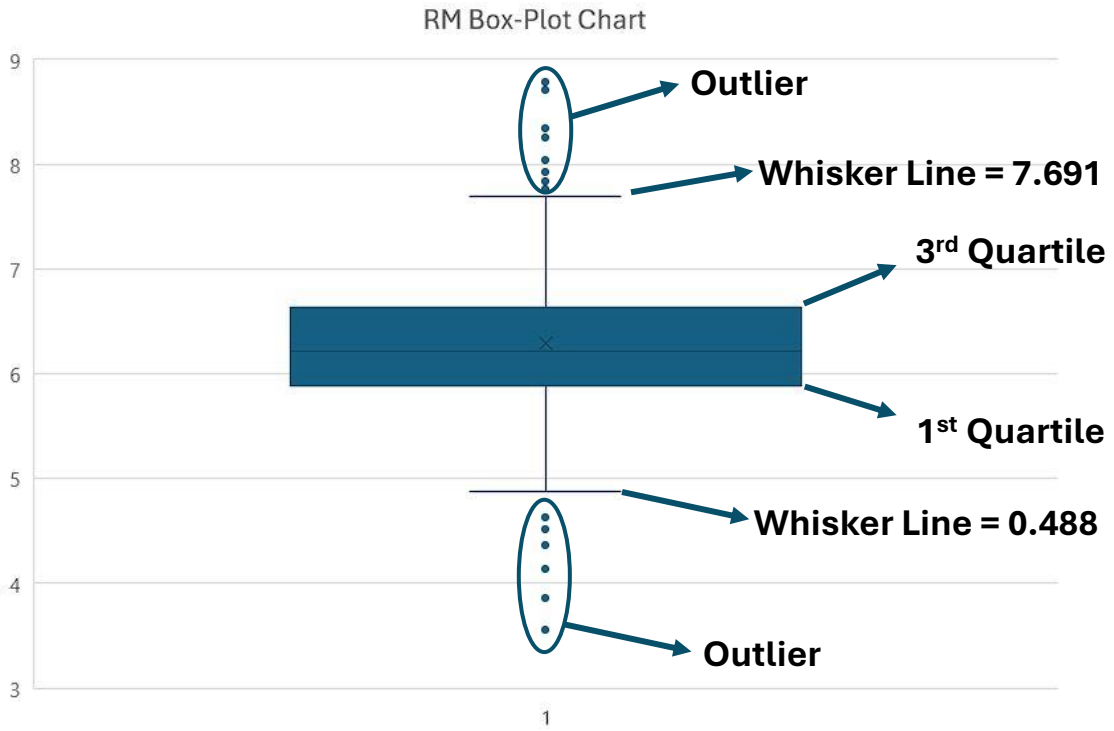
	A	B	C	D	E
1	RM	RM (Z transformation)	RM (Normalization)	Standardized (n-1)	0 to 1
2	6.575	0.41326292	0.577505269	0.41326292	0.577505269
3	6.421	0.194082387	0.547997701	0.194082387	0.547997701
4	7.185	1.281445551	0.694385898	1.281445551	0.694385898
5	6.998	1.015297761	0.658555279	1.015297761	0.658555279
6	7.147	1.227362043	0.687104809	1.227362043	0.687104809

Normalization:
In this column, I normalized the data of RM variable with use of excel functions.
I create a function like this : $X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$

XLSTAT Check (Normalization):
In this column, I normalized the data of RM variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (RM Variable)

Outliers Detecting (Box-Plot Method)



Outliers :

As we can see on box-plot chart of RM variable, this variable has outliers both sides of its distribution (we guessed this before when we were examining the RM distribution)

All values of RM which are lower than 0.48 or those which are greater than 7.69 are detected as outliers based on box-plot method for RM variable.

Another thing that we should find out, is the number of outliers which were detected by this method for RM variable, so; I applied conditional formatting on RM, with criteria that we talked about, and I got the number 30, that you can see below as “count”

Outliers on the Lower Side: These outliers represent dwellings with an unusually low number of rooms compared to the average. This could indicate smaller homes or apartments that are uncommon in the dataset.

Outliers on the Upper Side: These outliers represent dwellings with an unusually high number of rooms compared to the average. These might be large homes or multi-family units that are also less common.

Whiskers & Box :

IQR (= 3rd quartile – 1st quartile)

whisker lines :
3rd quartile + 1.5 IQR = 7.691
1st quartile – 1.5 IQR = 0.488

Outliers:
Values of RM which are above 7.691
Values of RM which are below 0.488

Conclusion :

We saw the distribution of the RM variable, and we saw that it has a good potential of converting to a normal distribution.
We detected 30 outliers with box-plot method for RM variable.
There might be a good chance for converting RM, to a normal one, by removing its outliers that we just detected.
On the other hand, outliers were detected both sides of RM distribution, it implies that the samples are so dense around the mode.

Statistical Analyses (RM Variable)

Outliers Detecting (Z-Score Method)

	A	B
1	RM	RM (Z transformation)
227	8.725	3.473250881
259	8.704	3.443362627
264	8.398	3.007848061
366	8.78	3.551529643
367	3.561	-3.876413226
369	3.863	-3.446591661
376	4.138	-3.055197852
408	4.138	-3.055197852

Outliers With Z-Score Method:

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.
So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I got the result as you can see in the table.

Average: 0.005323828 Count: 8 Min: -3.876413226 Max: 3.551529643 Sum: 0.042590622

8 Outliers:

8 outliers are detected based on Z-Score method.
While, the number of outliers which were detected based on box-plot method was 30.
As it is obvious, there is a significant different between these two methods.

Box-Plot VS Z-Score :

When we compare the results of these two methods for detecting outliers for RM feature, there is a significant different.
With Box-Plot method we got 30 outliers
With Z-Score method we got 8 outliers
If we want to decide outliers of which method should rely on, I prefer to go with Z-Score method, cause each outlier detected by this method is also detected as outlier with box-plot method.
On the other hand, number of outliers with box-plot method for this feature are too much, approximately 6% of our samples. So; it does not seem wise to go with box-plot method in this situation.

Statistical Analyses (RM Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, RM variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.

So; I removed the outliers which were detected by Z-Score method once, and once I removed the outliers which were detected by box-plot method , from the dataset, and applied a normality test for both of these conditions to see if now, it follows a normal distribution, and the answer was negative to this question.

So; I applied box-cox transformation on the RM variable (once on values which were not detected as outliers based on box-plot method, and once on values which were not detected as outliers based on z-score method)

And then I applied normality test again and the result was that neither of these methods worked.

So, as the conclusion, we find it out that we cannot convert the RM variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

Anderson-Darling test (Box-Cox transformation):

A^2	0.825
p-value (Two-tailed)	**0.033**
alpha	0.050

Normality Test After Removing Outliers :

This is the result of normality test of RM variable after removing its outliers and after transforming it with box-cox method.

As we can see, p-value is less than alpha, so; this variable does not follow a normal distribution even after removing its outliers.

Statistical Analyses (RM Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:
I am going to check the correlation between RM variable and target variable which is MEDV, these are both continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	RM	MEDV (1000$)
RM	**1**	**0.695**
MEDV (1000$)	**0.695**	**1**

Relatively Strong And Direct Correlation:
The correlation matrix and the value of 0.69 tells us that there is a direct correlation between these 2 variables.
Meaning that if one of the increase, the other one will increase too.
On the other hand, the absolute value would be 0.069, which indicates that the correlation is relatively strong.

Coefficients of determination (Pearson):

Variables	RM	MEDV (1000$)
RM	**1**	0.484
MEDV (1000$)	0.484	**1**

p-values (Pearson):

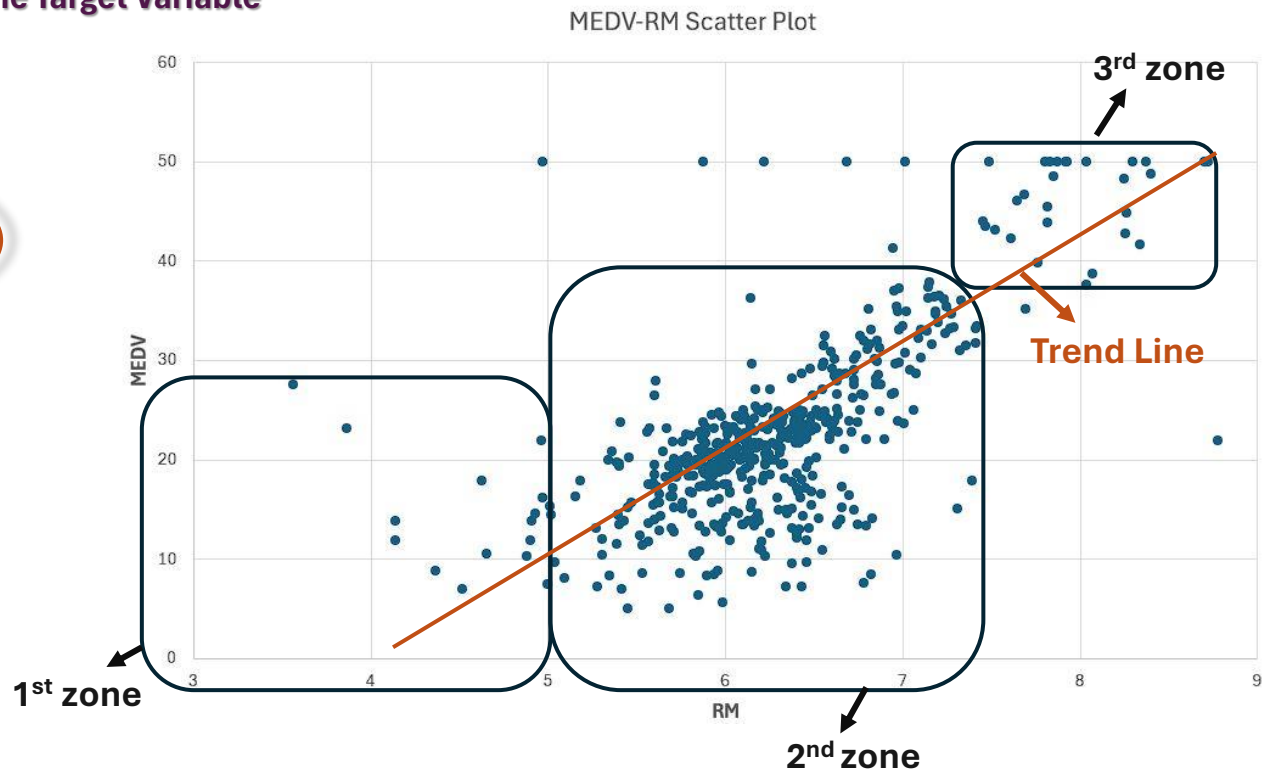
Variables	RM	MEDV (1000$)
RM	**0**	**<0.0001**
MEDV (1000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:
The value is <0.0001 suggests that the correlation between RM and MEDV is statistically significant and it is not due to random changes.

Power Of Prediction:
The value of 0.484 in this table, indicates that 48.8% of the variance in target variable (MEDV) can be explained by the variance in RM variable.

Statistical Analyses (RM Variable)

Scatter Plot With The Target Variable



3 Zones:

As we see on the chart, we can divide the city into 3 zones based on median value of houses in each area and RM rates of each area.
This gives us an interesting insight as we can interpreter as following :

1st Zone:

This zone contains the low rates of RM comparing to other zones. As we can see, house prices won't get upper than a specific value (around 30K).
This zone probably is mostly made of low-status families.
Samples are not concentrated, the gap between them is relatively huge, meaning that in this zone, the variety option is less than the other zones.

2nd Zone:

This zone is probably for middle-class families. Number of rooms is in a normal range in this zone, while MEDV can vary from a low value (around 8K) to a relatively high value (around 40K).
This shows that this zone is mostly made of normal houses and middle-class families.
On the other hand, the concentration of samples in this area is much more than the other 2 zones, meaning that we can consider the 2nd zone as the yardstick.

Trend Line :

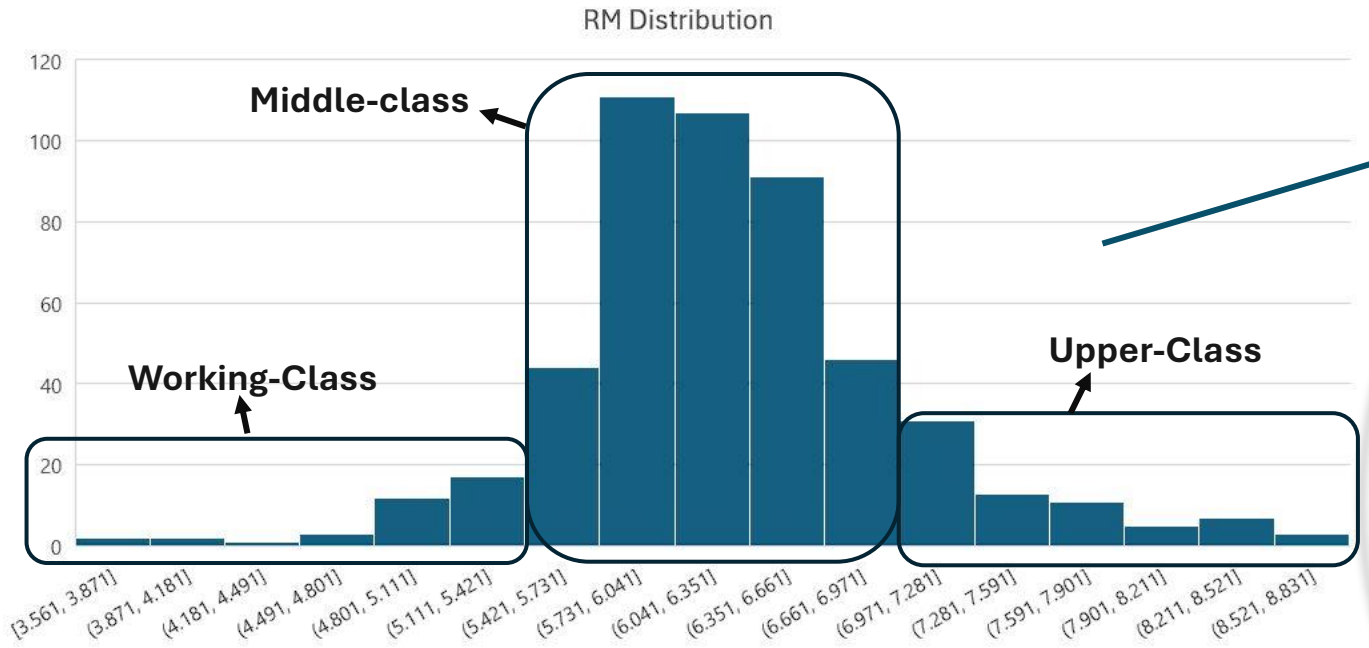
As we can see on the chart, the trend line shows a direct correlation between RM and MEDV variables.
When RM increases, MEDV will also increase.
In each zones, samples which are below this line, and they are more on the right side, are probably better options to choose.

3rd Zone:

This zone seems to be for upper-class families, the number of rooms in this zone is much more than the other zones and also the house prices in this zone is much higher than the other two zones.
This zone contains the most expensive houses. The minimum of house prices in this zone is the maximum of house prices of 2nd zone.

Statistical Analyses (RM Variable)

Question : Is There Any Difference Between Average Of House Prices Based On RM Rates?



RM 3 Classes :

I am going to create a new feature based on RM feature.
This feature is going to be 1 for areas which RM rate in them is greater than 6.971.
And is going to be 2 for areas which RM rate in them is between 5.421 and 6.971.
And is going to be 3 for areas which RM rate in them is less than 5.421.
I considered these 3 classes to be for upper-class families, middle-class families and low-status families respectively.
And I am going to compare the average of house prices between these 3 classes.
The reason for dividing the RM variable into these three categories is the distribution of RM variable
It seems that there is a connection between this distribution and the financial status of families who live in these areas.
So, I am going to examine if the average of MEDV variable of these 3 classes are equal to each other or not.
Our guess is that they are not similar, the most expensive one must be class 1, then class 2 and then class 3
So, lets find it out

	A	B	C
1	RM	RM Classification	MEDV (1000$)
2	6.575	=IF(A2>6.971, 1, IF(A2>=5.421, 2, 3))	
3	6.421	2	21.6
4	7.185	1	34.7
5	6.998	1	33.4
6	7.147	1	36.2

Statistical Analyses (RM Variable)

Question : Is There Any Difference Between Average Of House Prices Based On RM Rate? (ANOVA Test Results, 1st Page)

Goodness of fit statistics (MEDV (1000$)):	
Observations	506
Sum of weights	506
DF	503
R^2	0.467
Adjusted R^2	0.465
MSE	45.231
RMSE	6.725
MAPE	27.604
DW	0.899
Cp	3.000
AIC	1931.748
SBC	1944.427
PC	0.539

R Squared:
With variance of RM variable, we can anticipate 46.7% of variance of the target variable which is MEDV

Correlation matrix:				
	RM Classification-1	RM Classification-2	RM Classification-3	MEDV (1000$)
RM Classification-1	**1**	-0.774	-0.113	0.668
RM Classification-2	-0.774	**1**	-0.542	-0.425
RM Classification-3	-0.113	-0.542	**1**	-0.219
MEDV (1000$)	0.668	-0.425	-0.219	**1**

Correlations Between Different RM Classes :
Values which are in this right triangle, show the correlation between different classes of RM.
As we can see all the values show inverse correlations.
And Class 1 has a strong , inverse correlation with class 2

Correlations Between Different RM Classes With Target Variable :
These values in blue box show the correlation between different RM classes with the target variable
As we can see, there is a relatively strong correlation between target variable and 1st class of RM variable. As we mentioned before, this class was associated with upper-class families based on our definition.
Target variable has an inverse correlation with other 2 classes, it may indicate that selling big houses in those areas may be a little bit harder. People who live in those areas (class 2 & class 3) are less likely to buy big houses, maybe due to their financial status.
We should not forget that before running this ANOVA test, we did not remove the outliers of RM variable, there might be an effect on these results. Maybe we could have better understanding of what is actually happening if we removed those outliers.
However, by now, these results can satisfy our purposes.

Statistical Analyses (RM Variable)

Question : Is There Any Difference Between Average Of House Prices Based On RM Rate? (ANOVA Test Results, 2nd Page)

Analysis of variance (MEDV (1000$)):

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	2	19965.344	9982.672	220.707	**<0.0001**
Error	503	22750.951	45.231		
Corrected Total	505	42716.295			

There Is Difference Between RM Different Classes :

This number here, tells us that there is a meaningful difference between different classes of RM variable in terms of MEDV.
Meaning that if want to created a model for MEDV variable, it should include RM variable.
RM is an effective element on MEDV variable.

Model parameters (MEDV (1000$)):

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	15.359	1.106	13.892	**<0.0001**	13.187	17.532
RM Classification-1	22.493	1.367	16.455	**<0.0001**	19.808	25.179
RM Classification-2	5.151	1.156	4.457	**<0.0001**	2.880	7.422
RM Classification-3	0.000	0.000				

Reliable Coefficients :

All of the values, are less than alpha, meaning that all of the coefficients which are used in the model below, are reliable.

Equation of the model (MEDV (1000$)):

MEDV (1000$) = 15.3594594594593+22.4933976833979*RM Classification-1+5.1508162297638*RM Classification-2

Model :

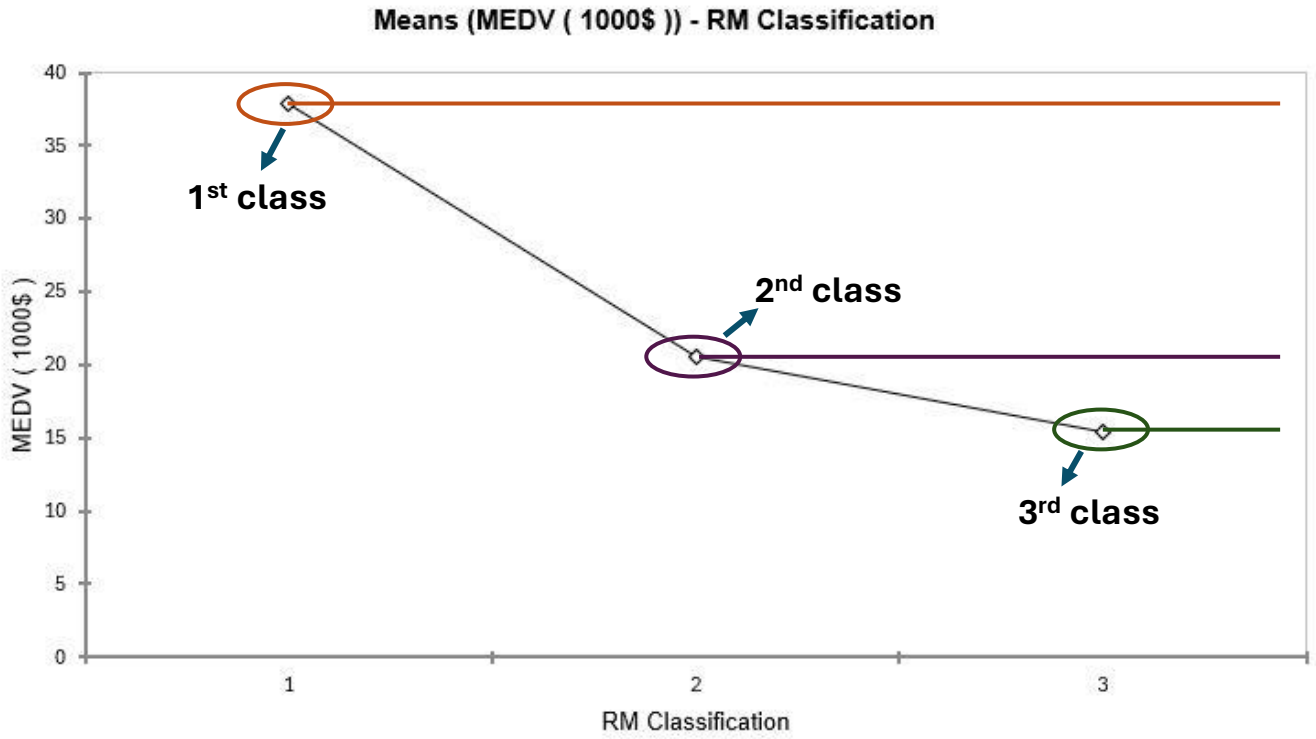
XLSTAT created a model for us.
The dependent variables are different classes of RM and the dependent variable is MEDV.
There is also an interception which is estimated to be 15.35
It is a linear model.
We won't rely on this model for anticipating MEDV, because it only contains RM variable

Statistical Analyses (RM Variable)

Question : Is There Any Difference Between Average Of House Prices Based On RM Rate? (ANOVA Test Results, 3rd Page)

RM Classification / Tukey (HSD) / Analysis of the differences between the categories with a confidence interval of 95% (MEDV (1000$)):

Contrast	Difference	Standardized difference	Critical value	Pr > Diff	Significant
1 vs 3	22.493	16.455	2.351	**<0.0001**	Yes
1 vs 2	17.343	19.900	2.351	**<0.0001**	Yes
2 vs 3	5.151	4.457	2.351	**<0.0001**	Yes
Tukey's d critical value:			3.324		



Tukey (HSD) Table Interpretation :

All of the results are “yes” meaning that there is a meaningful difference between different classes of RM.
Neither of them are similar to each others.
As we guessed before on previous slides, it seems that we correctly divided RM variable into these 3 classes and we correctly assigned them to different families in terms of financial status.

1st class :

As we can see, the average of MEDV for this class is much higher than the other two classes as we could guess before
We assigned this class to upper-class families who have better financial status than the other two classes.
The difference between this class and other two classes are huge

2nd class :

This class is far away from the first class, and a little bit closer to the 3rd class in terms of average of MEDV
Meaning that on average, houses in this class, are cheaper than houses of the 1st class and more expensive than houses of the 3rd class
The chart shows us that this class has more similarities with 3rd class
We assigned this class to middle-class families

3rd class :

This class is assigned to low-status families.
The chart shows us that on average, houses in this class, are the cheapest ones.
On average, they are cheaper than houses of other two classes
This class of houses are probably smaller than houses of other two classes because RM variable can be used as an indicator of house area

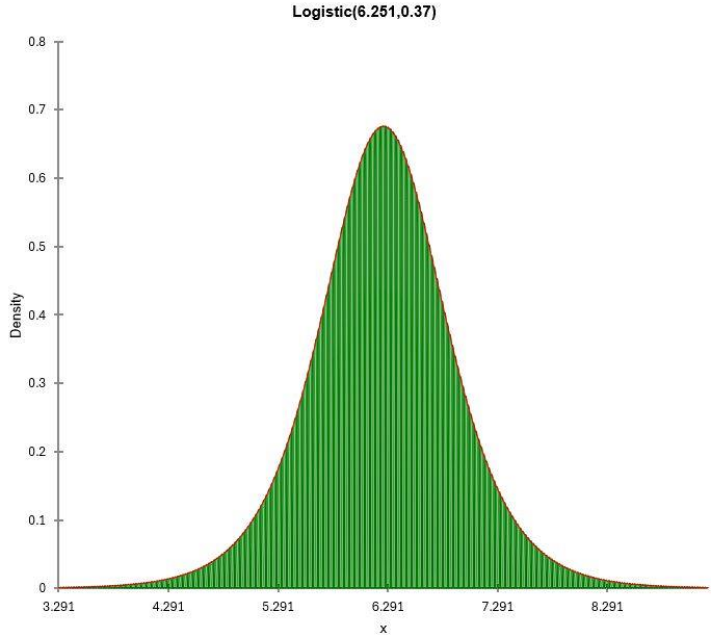
Statistical Analyses (RM Variable)

The Best Fitting Distribution

Distribution	p-value
Beta4	0.001
Chi-square	<0.0001
Erlang	0.010
Fisher-Tippett (1)	<0.0001
Fisher-Tippett (2)	<0.0001
Gamma (2)	0.010
GEV	<0.0001
Gumbel	<0.0001
Log-normal	0.005
Logistic	**0.167**
Normal	0.002
Student	<0.0001
Weibull (1)	<0.0001
Weibull (2)	<0.0001

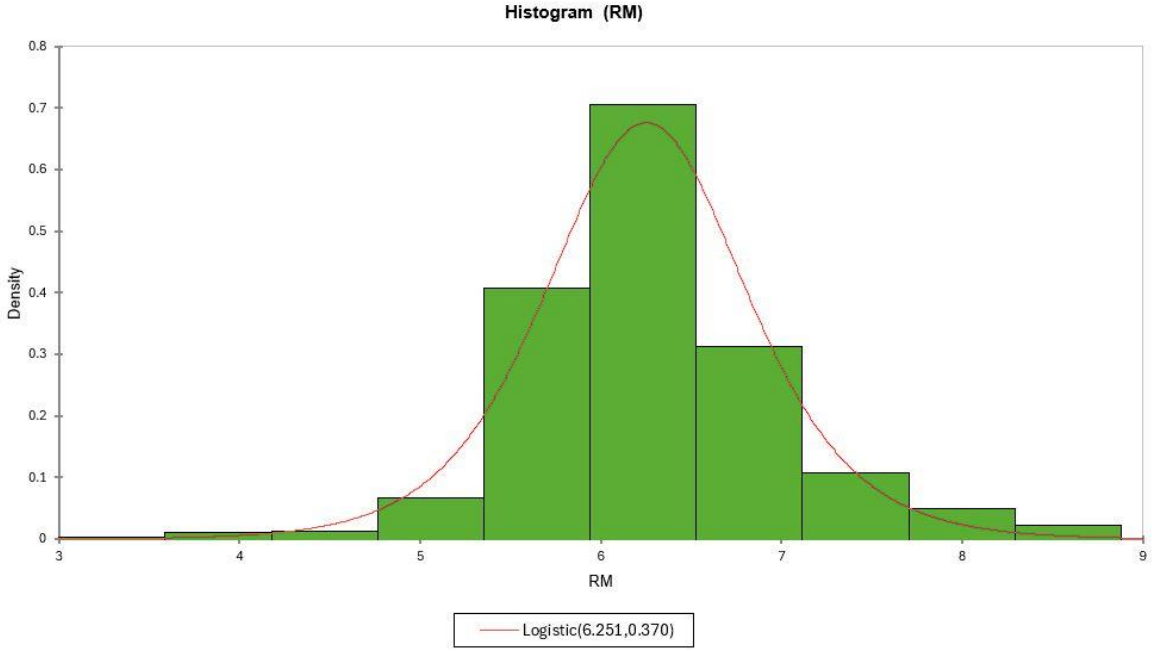
Logistics Distribution :

With use of XLSTAT, I found out that the best fitting distribution for RM variable, is logistic distribution with given parameter as below (μ & σ)
Then again, with use of XLSTAT I plot the distribution with these parameters and its corresponding value and I got the chart which you can see on the right, which seems so suit for NOX variable considering this variable's distribution.



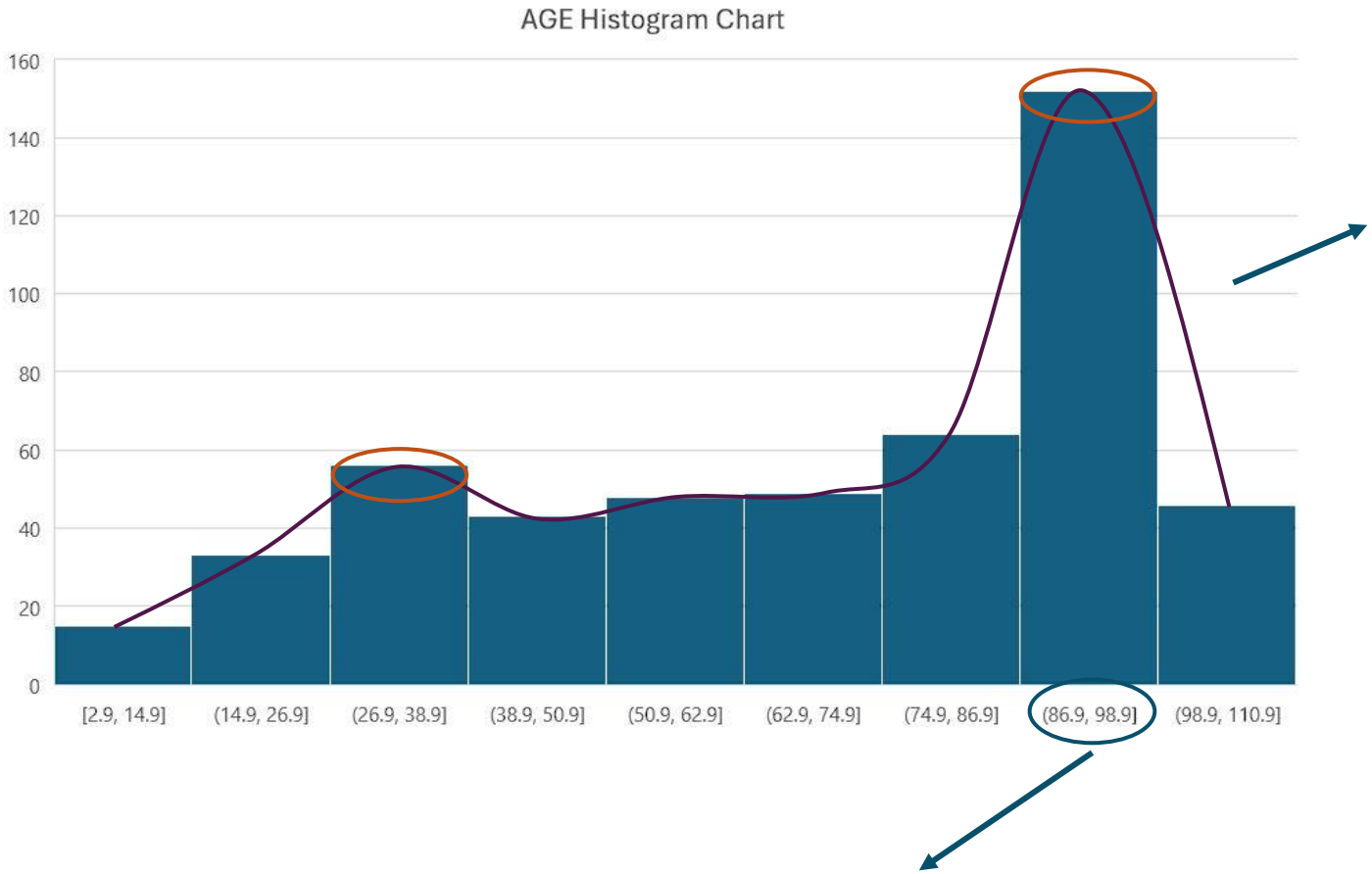
Estimated parameters (Logistic):

Parameter	Value	Standard error
μ	6.251	0.029
s	0.370	0.014



Statistical Analyses (AGE Variable)

Examining The Distribution



AGE Histogram Chart:

The chart shows us that the AGE variable is moderately negatively skewed.
The distribution has more data points concentrated on the right side, with a tail extending towards the lower values on the left.
The mean of the distribution is likely less than the median, as the lower (left) tail pulls the mean downwards.
The majority of the data points are clustered towards higher ages, but there are enough lower ages to create a noticeable leftward skew.
This feature is also multi-modal, meaning that as we can see on the chart, it has two peaks.
One of our problems that must be solved is that we find the reason that causes AGE feature to be multi-modal.

Mode:

The mode of AGE feature must be something in this range.

Statistical Analyses (AGE Variable)

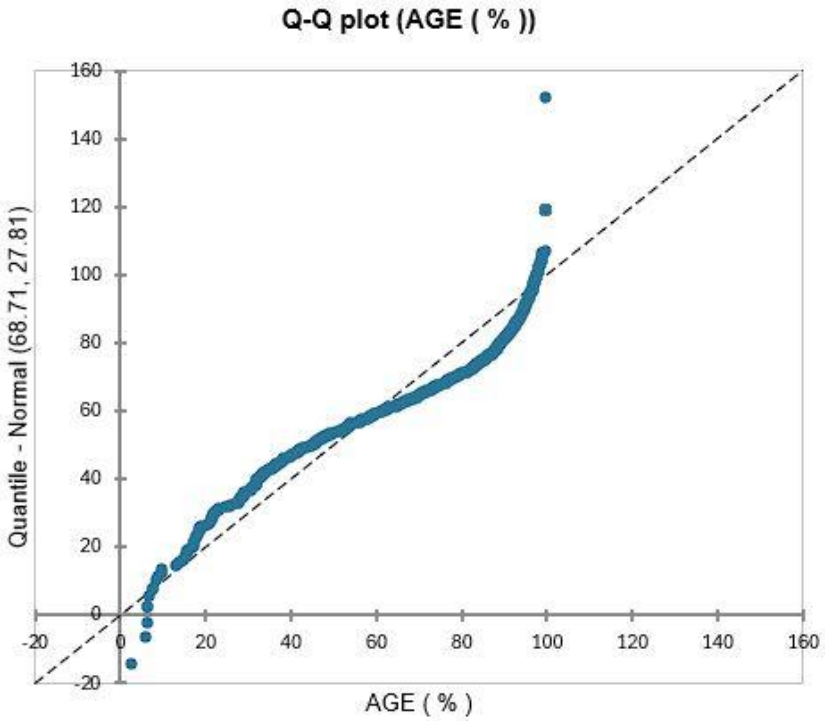
Examining The Descriptive Statistics

Statistic	AGE (%)
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	2.900
Maximum	100.000
Freq. of minimum	1
Freq. of maximum	42
Range	97.100
1st Quartile	45.625
Median	76.950
3rd Quartile	93.875
Sum	34766.900
Mean	68.709
Variance (n)	773.168
Variance (n-1)	774.699
Standard deviation (n)	27.806
Standard deviation (n-1)	27.833
Skewness (Pearson)	-0.586
Kurtosis (Pearson)	-0.974
Lower bound on mean (95%)	66.278
Upper bound on mean (95%)	71.140
Lower bound on variance (95%)	687.369
Upper bound on variance (95%)	879.891

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen only 1 time among all of the records
- Maximum value of this variable can be seen 42 times among all records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A skewness value of -0.58 for the AGE variable indicates a moderate negative skewness, meaning that while most ages cluster around higher values, there is a noticeable stretch towards lower values, influencing the mean.
- A kurtosis value of -0.97 for the AGE variable indicates a platykurtic distribution, meaning it has a flatter peak and lighter tails, with fewer extreme values. This suggests a more evenly distributed set of ages around the central value.
- The mean of the population of this variable must be something between 66.2 and 77.1 with confidence level of 95%
- The variance of the population of this variable must be something between 687.3 and 879.8 with confidence level of 95%

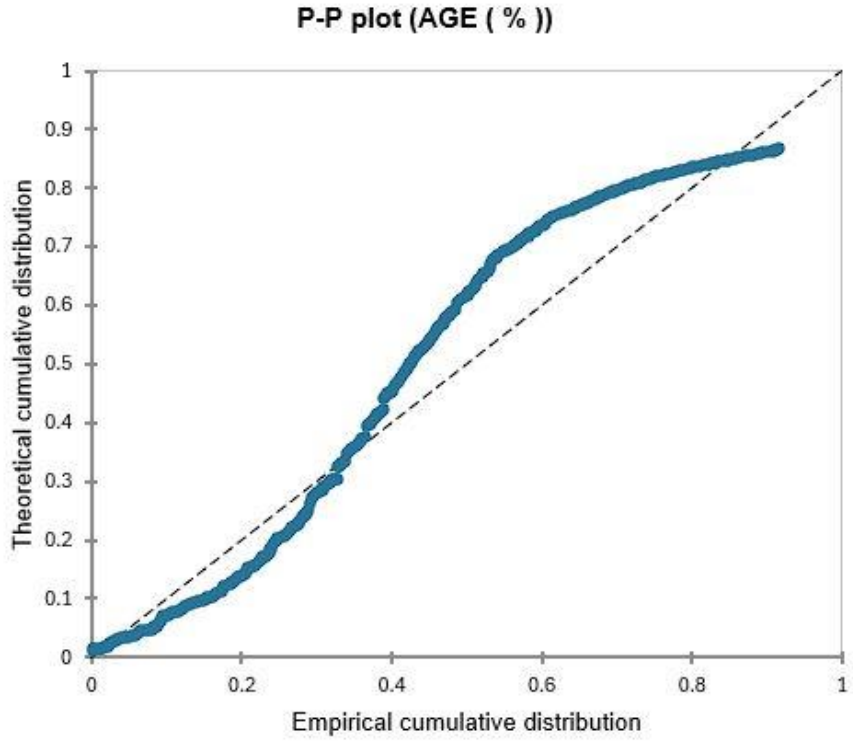
Statistical Analyses (AGE Variable)

Normality Test (Anderson-Darling Method)



Anderson-Darling test (AGE (%)):	
A^2	18.284
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test Result :
p-value is less than alpha, so we should reject the null hypothesis. So; AGE variable does not follow a normal distribution.



P-P & Q-Q plot:
Paying attention to these two charts, it seems that there is no hope for AGE variable to be converted to a normally distributed variable.

Statistical Analyses (AGE Variable)

Outliers Detecting (Variable Transformation)

Raw data:
This is the raw data of AGE variable without any transformations.

Z-Score Normalization:
In this column, I transformed the data of AGE variable with use of excel functions.
I create a function like this :

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

XLSTAT Check (Z-Score Normalization):
In this column, I transformed the data of AGE variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

	A	B	C	D	E
1	AGE (%)	AGE (Z transformation)	AGE (Normalization)	Standardized (n-1)	0 to 1
2	65.2	-0.126081818	0.641606591	-0.126081818	0.641606591
3	78.9	0.366132171	0.782698249	0.366132171	0.782698249
4	61.1	-0.273386734	0.59938208	-0.273386734	0.59938208
5	45.8	-0.823085569	0.441812564	-0.823085569	0.441812564
6	54.2	-0.52129013	0.528321318	-0.52129013	0.528321318

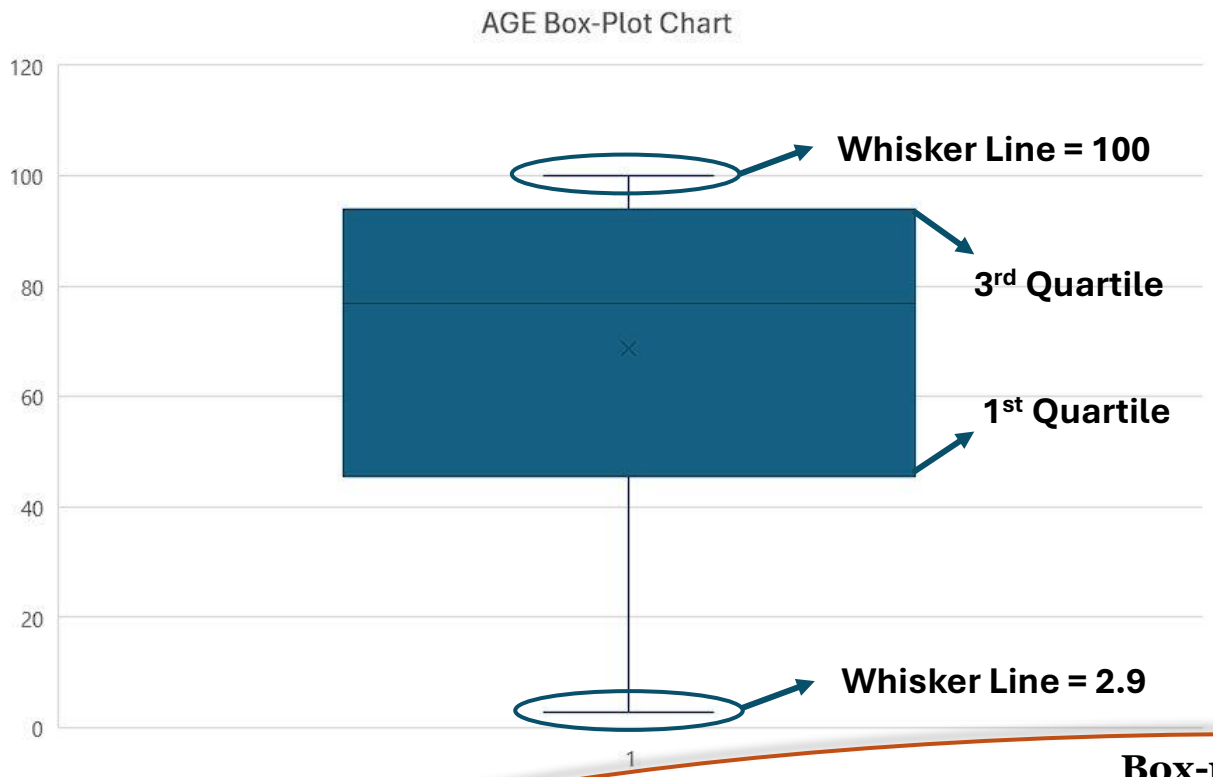
Normalization:
In this column, I normalized the data of AGE variable with use of excel functions.
I create a function like this :

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

XLSTAT Check (Normalization):
In this column, I normalized the data of AGE variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (AGE Variable)

Outliers Detecting (Box-Plot Method)



1st Q	45.625
Median	76.95
3rd Q	93.875
IQR	48.25
3rd Q + 1.5IQR	166.25
1st Q - 1.5IQR	-26.75

Above Maximum:

As we can see, this value, which is a limit line, and any value above it should be considered as outlier; is greater than the maximum of AGE variable
So, we would not have outliers between high values of AGE variable

Below Minimum:

This value is another limit line, and any value below it, should be considered as outlier; is less than the minimum of AGE feature
So, we would not have outliers between low values of AGE feature

Box-plot chart:

In our examination of the AGE variable, representing the age distribution of the sample population, a box-plot chart was utilized to detect potential outliers. Remarkably, the box-plot revealed no outliers within this dataset. This outcome indicates a consistent and homogeneous age distribution.
The absence of outliers highlights a lack of extreme ages that deviate significantly from the overall pattern, providing a stable and reliable dataset for further analysis. Consequently, this uniform distribution allows for more accurate and representative statistical insights, making it a robust indicator for demographic studies and related analyses.
How ever, there are some values in this feature which may not be statistically outlier, but the raise our curiosity to think of them. For example, existence of the value 100, seems not right. It means that there are an area in the Boston, that all of the houses in this area are built before 1940, with even not an exception.
On the other hand, because we did not find any outliers with box-plot method, we guess that we cannot find any, with z-score method as well.

Statistical Analyses (AGE Variable)

Outliers Detecting (Z-Score Method)

	A	B
1	AGE (%)	AGE (Z transformation)
2	65.2	-0.126081818
3	78.9	0.366132171
4	61.1	-0.273386734
5	45.8	-0.823085569
6	54.2	-0.52129013
7	58.7	-0.359614003
8	66.6	-0.075782578

No Outliers With Z-Score Method :

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.

So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I found out that there is no value between transformed data to be greater than 3 or less than -3

Box-Plot VS Z-Score :

In our comprehensive analysis of the AGE variable, we utilized both the Box-Plot and Z-Score methods to detect potential outliers. Remarkably, both methods consistently revealed the absence of outliers within the dataset.

The Box-Plot method, a graphical tool, displayed no data points beyond the whiskers, indicating that all ages fell within the expected range. This visual confirmation was corroborated by the Z-Score method, a statistical approach that measures the number of standard deviations each data point is from the mean. All AGE values exhibited Z-Scores within the common threshold (typically $Z < 3$), further affirming the lack of significant deviations from the central tendency.

The agreement between the Box-Plot and Z-Score methods underscores the reliability and consistency of the AGE data, suggesting a uniform distribution across the sample. This uniformity is crucial for demographic analysis, as it ensures that the dataset accurately represents the population without extreme values skewing the results.

Statistical Analyses (AGE Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, AGE variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.

So; I transformed this variable with box-cox method, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.

So, as the conclusion, we find it out that we cannot convert the AGE variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

Box-Cox transformati
205.8631787
266.4062375
188.5505891
127.6274138
160.3255525
178.6005074
211.8632202

Transformed data of AGE Variable With Box-Cox Method

Anderson-Darling test (Box-Cox transformation):	
A^2	17.056
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test After Box-Cox Transformation :

As we can see, the result of the normality test of transformed data (with box-cox method), AGE variable still does not follow a normal distribution.

Statistical Analyses (AGE Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:

I am going to check the correlation between AGE variable and target variable which is MEDV, these are both continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	AGE (%)	MEDV (1000$)
AGE (%)	**1**	**-0.382**
MEDV (1000$)	**-0.382**	**1**

Weak And Inverse Correlation:

The correlation matrix and the value of -0.38 tells us that there is an inverse correlation between these 2 variables.
Meaning that if one of the increase, the other one will decrease.
On the other hand, the absolute value would be 0.38, which indicates that the correlation is relatively weak.

Coefficients of determination (Pearson):

Variables	AGE (%)	MEDV (1000$)
AGE (%)	**1**	0.146
MEDV (1000$)	0.146	**1**

p-values (Pearson):

Variables	AGE (%)	MEDV (1000$)
AGE (%)	**0**	**<0.0001**
MEDV (1000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:

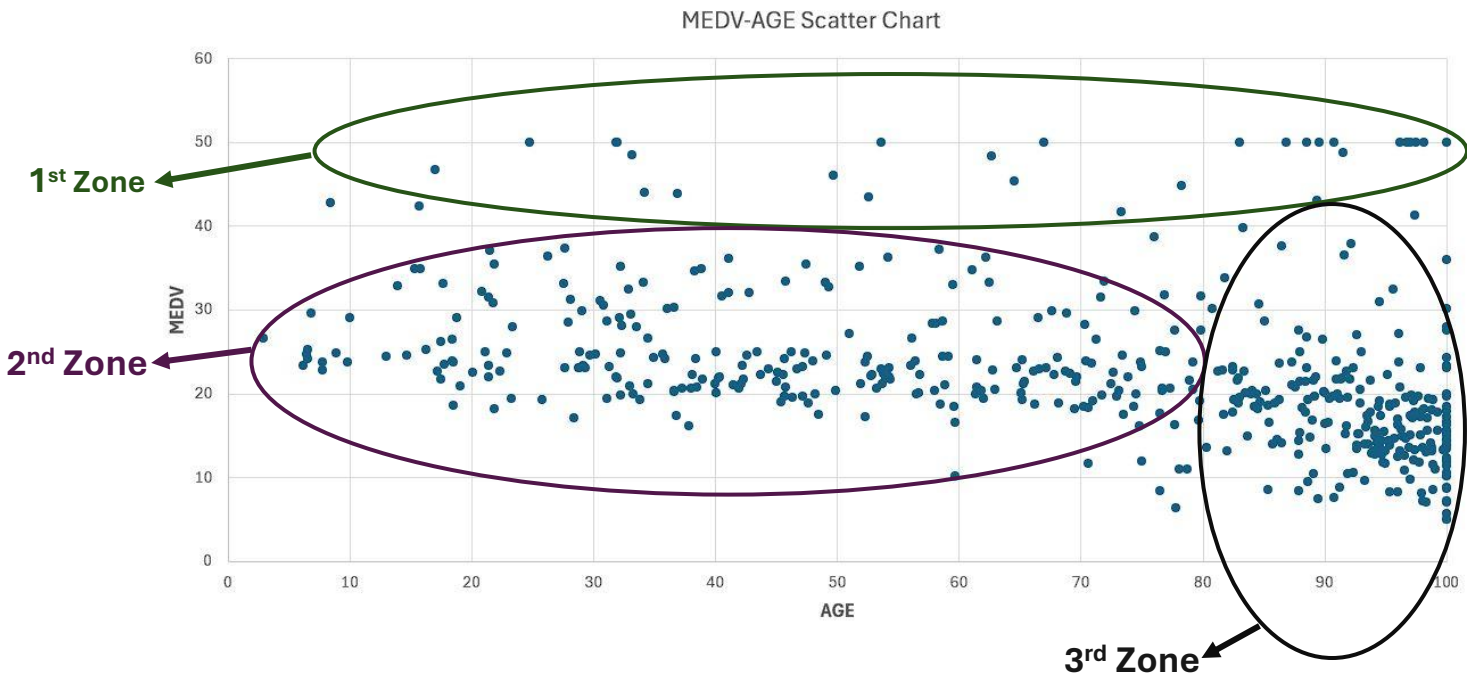
The value is <0.0001 suggests that the correlation between AGE and MEDV is statistically significant and it is not due to random changes.

Power Of Prediction:

The value of 0.146 in this table, indicates that only 14.6% of the variance in target variable (MEDV) can be explained by the variance in AGE variable.

Statistical Analyses (AGE Variable)

Scatter Plot With The Target Variable



3 Zones:

As we see on the chart, we can divide the city into 3 zones based on median value of houses in each area and AGE rates of each area.
This gives us an interesting insight as we can interpreter as following :

1st Zone:

This zone shows that the highest values of MEDV variable can be existed in different areas with different rates of AGE variable.
AGE variable, with the given definition (proportion of houses in an area, which are built before 1940) seems not be a good factor for anticipating the MEDV feature (as we saw, the correlation was weak)
As we can see on the chart, for the highest values of MEDV, AGE variable can change from the low values of AGE, to the highest value of AGE

2nd Zone:

This zone contains areas which have AGE rates lower than 80% . Meaning that at least 20% of each of areas which are in this zone, are built after 1940 .
As we can see MEDVs in this zone, has a minimum and cannot be anything below this minimum.

3rd Zone:

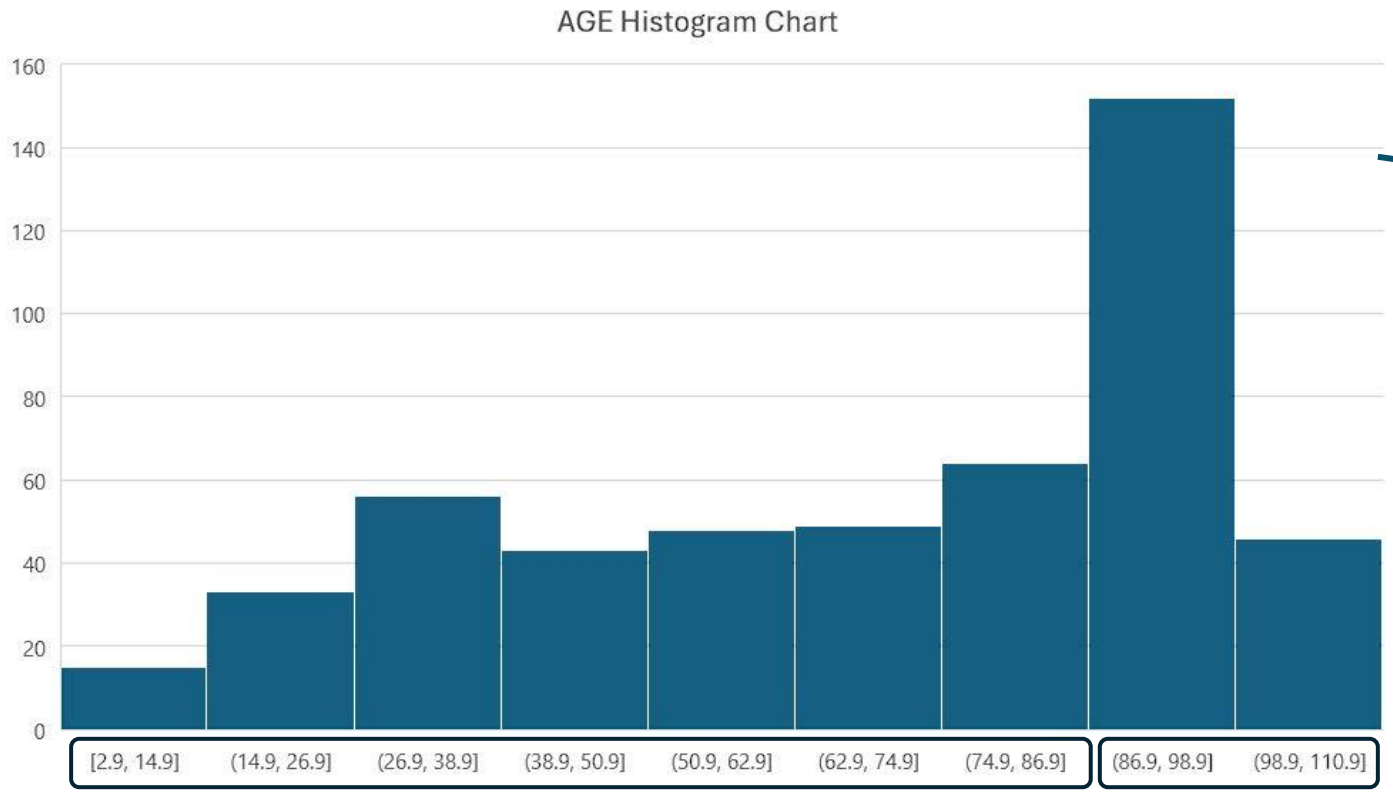
Samples in this zone include the highest values of AGE variable, and as it was guessable, they include the lowest levels of MEDV feature.
It indicates that for areas, that proportion of houses which were built before 1940 is above 80% , the MEDV drops to the lowest levels of itself.
To simplify it, house prices in these areas can be cheaper than the other areas.

Correlation Between AGE & MEDV:

As we can see on the chart, there is an inverse and weak correlation between these two variables.
We can see that, while AGE rate increases, MEDV is decreasing.
But the point is that this correlation is so weak that it cannot create a clear pattern for us.
That was why that I mentioned in 1st zone's description, that if our definition of AGE variable was different, it could be more beneficial to us.

Statistical Analyses (AGE Variable)

Question : Is There Any Difference Between Average Of House Prices Based On AGE Rates?



AGE 2 Classes :

I am going to create a new feature based on AGE feature. This feature is going to be 0 for areas which AGE rate in them is lower than 86.9.
And is going to be 1 for areas which AGE rate in them is greater than 86.9.
I chose these 2 classes because of the distribution of the AGE variable. As we saw before, there is not a strong correlation between AGE and MEDV variables to create a clear pattern for us. But it seems to me interesting to examine if there is any difference between these two categories; because category labeled as class 1, has extreme values of age rate and it probably should be different with other areas in term of house prices

	A	B	C
1	AGE (%)	AGE Binary Classification	MEDV (1000$)
2	65.2	=IF(A2>86.9,1,0)	24
3	78.9	0	21.6
4	61.1	0	34.7
5	45.8	0	33.4
6	54.2	0	36.2

Statistical Analyses (AGE Variable)

Variances Equality Test (Leven's Method)

Variances Equality Test:

As we can see, P-value is greater than alpha, so; we should accept the null hypothesis. Meaning that variance of house prices with class 1 (samples with age rates above 86.9) , is equal to variance of house prices with class 0 (samples with age rates less 86.9).

So; for comparing the average of house prices between these two classes, with should assume the equality of variances.

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	1.019
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	0.313
alpha	0.050

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of AGE variable, we should use the appropriate method which is Leven's test.

If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (AGE Variable)

Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[5.090, 8.174]

Difference	6.632
t (Observed value)	8.450
|t| (Critical value)	1.965
DF	504
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Higher Average Of House Prices:

This tells us that areas of class 0, have higher house prices on average, comparing to areas of class 1 Meaning that on average, houses which are in areas that proportion of houses which are built before 1940 in those areas are less than 80%, have higher prices.

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have AGE rates below 86.9 (class 0) is not equal to the average of house prices for those areas which have AGE rates above 86.9 (class 1).

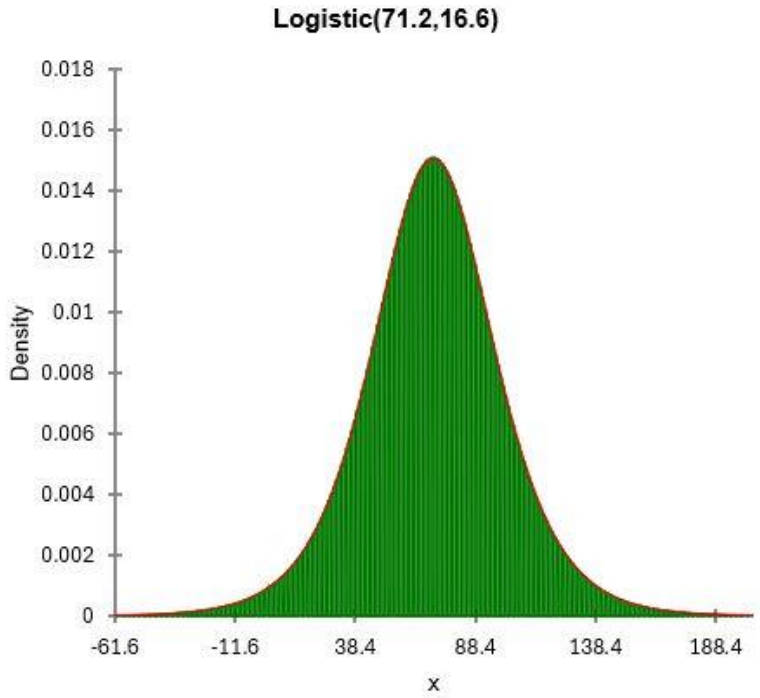
Statistical Analyses (AGE Variable)

The Best Fitting Distribution

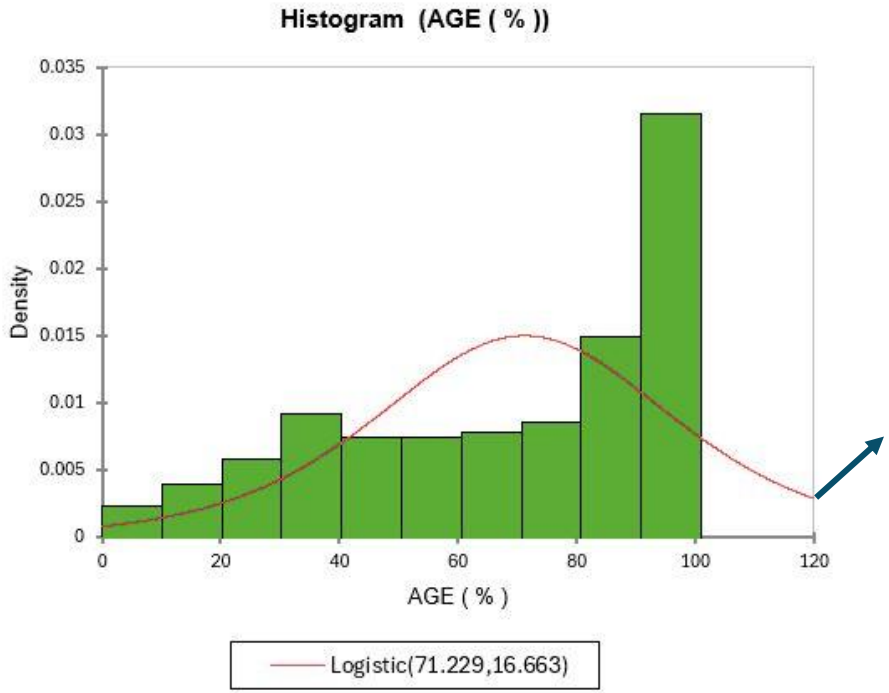
Distribution	p-value
Chi-square	<0.0001
Erlang	<0.0001
Exponential	<0.0001
Fisher-Tippett (1)	<0.0001
Fisher-Tippett (2)	<0.0001
Gamma (1)	<0.0001
Gamma (2)	<0.0001
GEV	<0.0001
Gumbel	<0.0001
Log-normal	<0.0001
Logistic	**<0.0001**
Student	<0.0001
Weibull (2)	<0.0001

Logistics Distribution :

With use of XLSTAT, I found out that the best fitting distribution for AGE variable, is logistic distribution with given parameter as below (μ & σ) Then again, with use of XLSTAT I plot the distribution with these parameters and its corresponding value and I got the chart which you can see on the right.



Parameter	Value	Standard error
μ	71.229	0.107
s	16.663	0.070

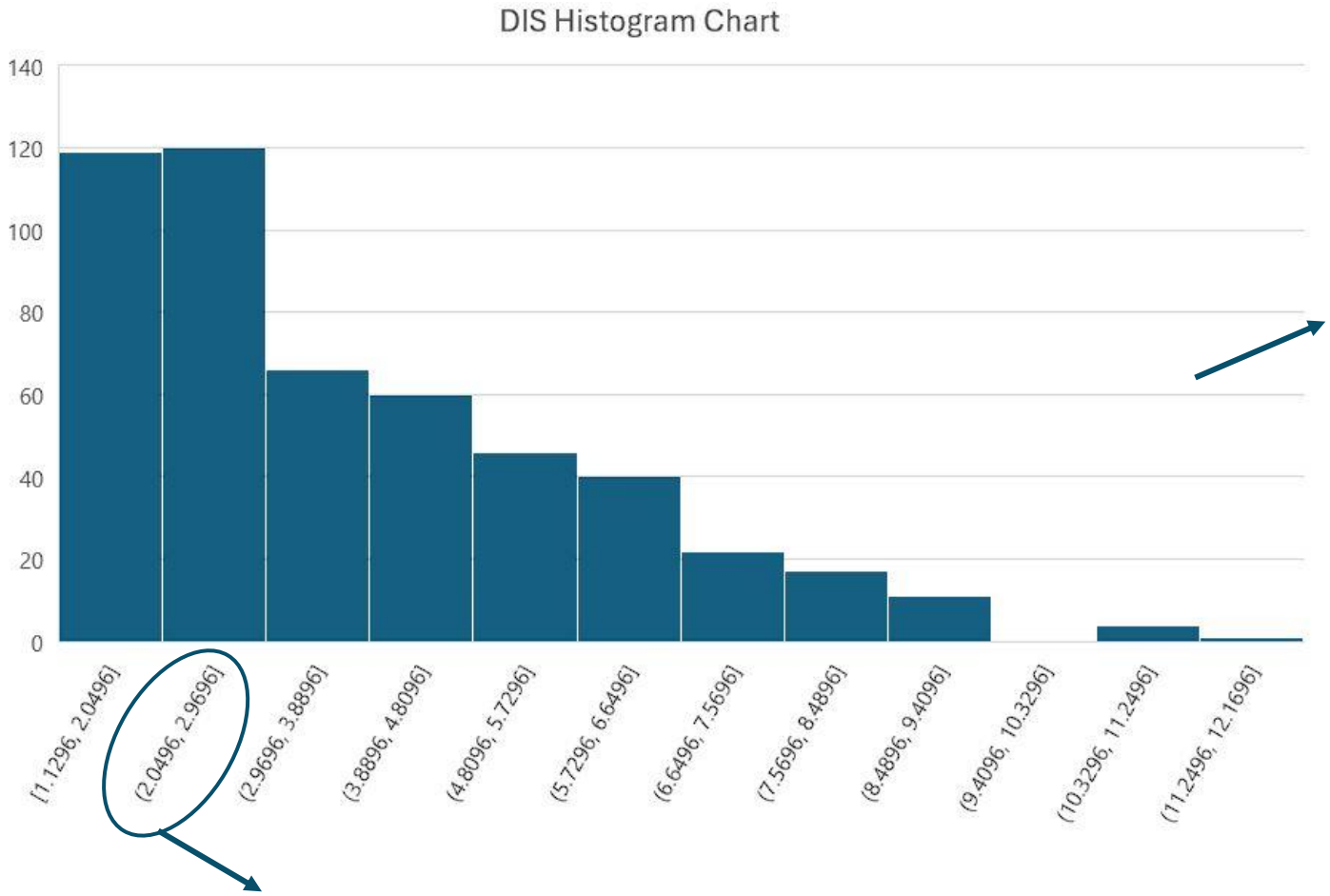


Not Perfectly Fitted :

Red line on the left chart, shows the best fitting estimated distribution for the AGE variable.
As we can see, it is not perfectly fitted, but it is the best we could do with actual distribution of the AGE variable.

Statistical Analyses (DIS Variable)

Examining The Distribution



Mode:

The mode of DIS feature must be something in this range.

DIS Histogram Chart:

The DIS variable, which measures the weighted distance to five Boston employment centers, reveals some intriguing characteristics in its distribution.

Upon examining the histogram chart, it becomes evident that the DIS variable exhibits a positive skewness, indicating that the distribution has a longer right tail. This skewness suggests that while most values are clustered towards the lower end, there are some higher values that extend further to the right, representing neighborhoods situated at varying distances from employment hubs.

Additionally, the kurtosis signifies a relatively flat peak compared to a normal distribution, indicating a broader, more spread-out distribution with fewer extreme values. This moderately platykurtic nature of the distribution implies a wider range of distances with a balanced spread around the mean.

The histogram chart visually confirms these statistical insights, highlighting the prevalence of shorter distances with a gradual decline in frequency as distances increase. Understanding these distribution characteristics is essential for urban planning and accessibility analysis, as it highlights the diversity in distance to employment centers across different neighborhoods.

I expect that DIS variable has an inverse correlation with the target variable which is MEDV.

Statistical Analyses (DIS Variable)

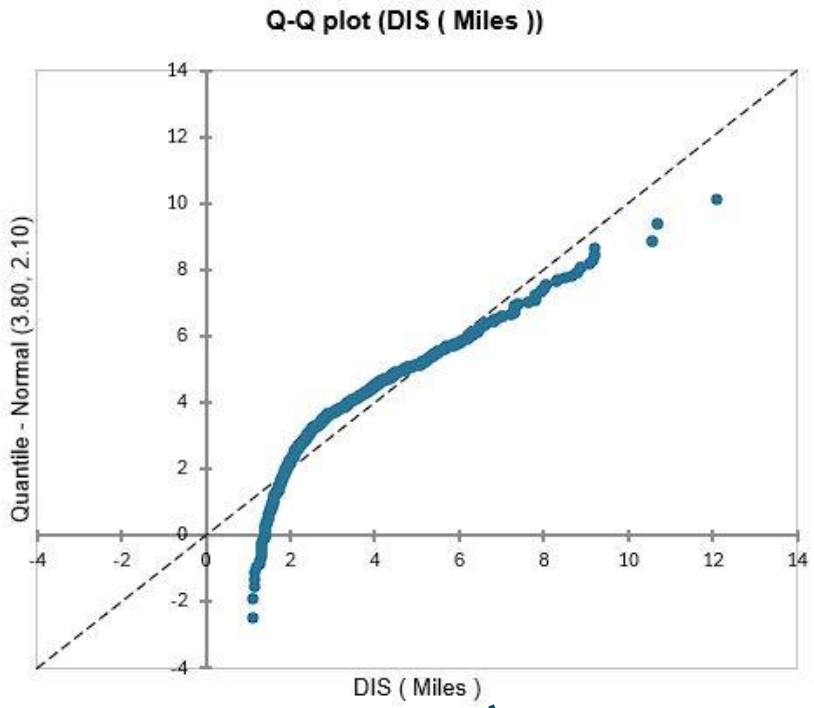
Examining The Descriptive Statistics

Statistic	DIS (Miles)
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	1.130
Maximum	12.127
Freq. of minimum	1
Freq. of maximum	1
Range	10.997
1st Quartile	2.100
Median	3.207
3rd Quartile	5.188
Sum	1920.292
Mean	3.795
Variance (n)	4.425
Variance (n-1)	4.434
Standard deviation (n)	2.104
Standard deviation (n-1)	2.106
Skewness (Pearson)	1.009
Kurtosis (Pearson)	0.471
Lower bound on mean (95%)	3.611
Upper bound on mean (95%)	3.979
Lower bound on variance (95%)	3.934
Upper bound on variance (95%)	5.036

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen only 1 time among all of the records
- Maximum value of this variable can be seen only 1 time among all records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A skewness value of 1.009 for the DIS variable indicates a moderate positive skewness, meaning that while most distances are clustered towards shorter values, there are a few larger values that pull the distribution to the right.
- A kurtosis value of 0.47 for the DIS variable indicates a platykurtic distribution, meaning it has a flatter peak and lighter tails, with fewer extreme values. This suggests a more evenly distributed set of distances around the central value.
- The mean of the population of this variable must be something between 3.6 and 3.9 with confidence level of 95%
- The variance of the population of this variable must be something between 3.9 and 5.03 with confidence level of 95%

Statistical Analyses (DIS Variable)

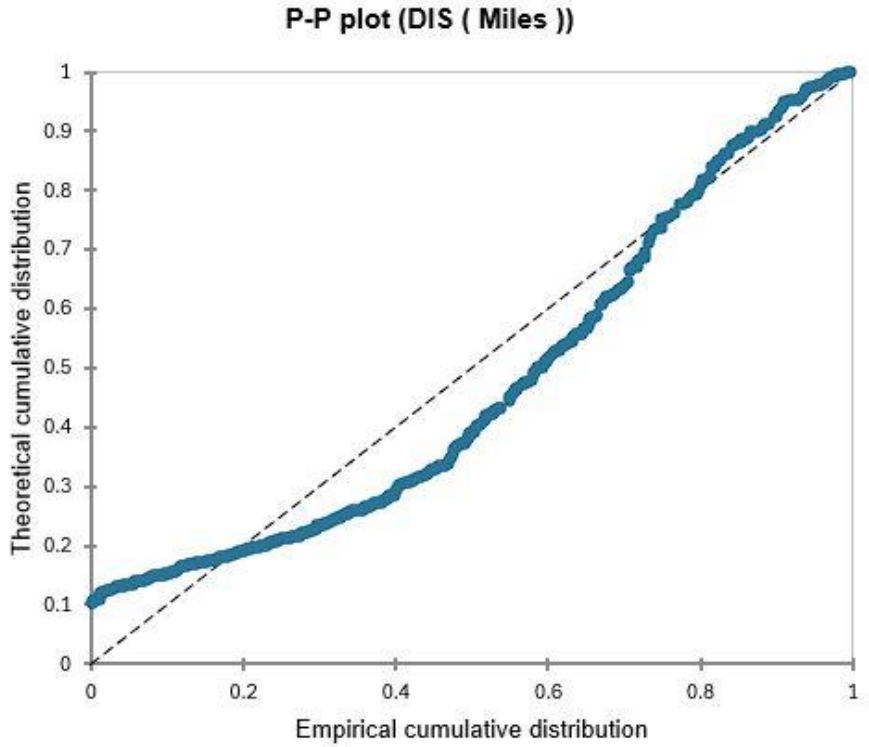
Normality Test (Anderson-Darling Method)



Normality Test Result :
p-value is less than alpha, so we should reject the null hypothesis. So; DIS variable does not follow a normal distribution.

Anderson-Darling test (DIS (Miles)):	
A^2	15.059
p-value (Two-tailed)	**<0.0001**
alpha	0.050

P-P & Q-Q plot:
These plots show us that there is a difference between DIS variable's distribution and a normal distribution as the normality test's result unveiled this fact to us.



Statistical Analyses (DIS Variable)

Outliers Detecting (Variable Transformation)

Raw data:

This is the raw data of DIS variable without any transformations.

Z-Score Normalization:

In this column, I transformed the data of DIS variable with use of excel functions.
I create a function like this :

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

XLSTAT Check (Z-Score Normalization):

In this column, I transformed the data of DIS variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

	A	B	C	D	E
1	DIS (Miles)	DIS (Z transformation)	DIS (Normalization)	Standardized (n-1)	0 to 1
2	4.09	0.140074984	0.269203139	0.140074984	0.269203139
3	4.9671	0.55660905	0.34896198	0.55660905	0.34896198
4	4.9671	0.55660905	0.34896198	0.55660905	0.34896198
5	6.0622	1.076671135	0.44854459	1.076671135	0.44854459
6	6.0622	1.076671135	0.44854459	1.076671135	0.44854459

Normalization:

In this column, I normalized the data of DIS variable with use of excel functions.
I create a function like this :

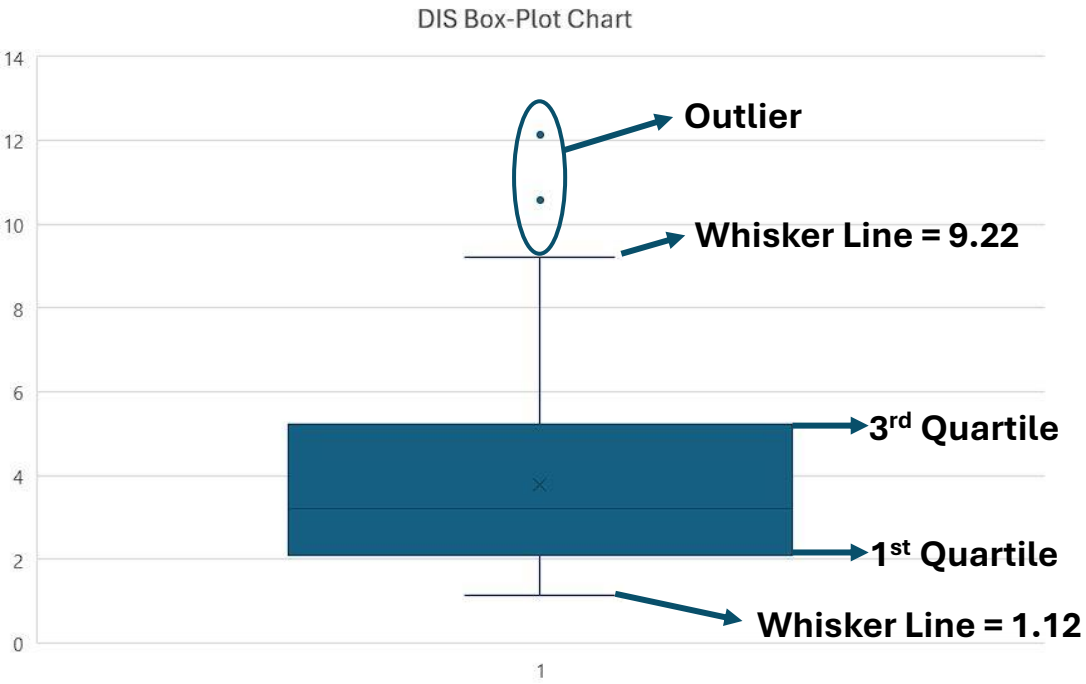
$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

XLSTAT Check (Normalization):

In this column, I normalized the data of DIS variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (DIS Variable)

Outliers Detecting (Box-Plot Method)



Outliers :

As we can see on box-plot chart of DIS variable, this variable has outliers only right side of its distribution (As we knew from examining the distribution of this variable)

All values of DIS which are greater than 9.2229 are detected as outliers based on box-plot method for DIS variable.

Another thing that we should find out, is the number of outliers which were detected by this method for DIS variable, so; I applied conditional formatting on DIS, with criteria that we talked about, and I got the number 5, that you can see below as “count”

Outliers on the Upper Side: These outliers represent dwellings which are relatively far away from five Boston employment centers. These dwellings are may be for rich families as their second house on countryside, or for parents who are retired, and so on. Anything these areas are, they are far from employment centers.

Whiskers & Box :

IQR (= 3rd quartile – 1st quartile)

whisker lines :
3rd quartile + 1.5 IQR = 9.2229
1st quartile – 1.5 IQR = 1.12

Outliers:
Values of DIS which are above 9.2224
Values of DIS which are below 1.12

Conclusion :

We detected 5 outliers for DIS variable. The number of outliers with Z-Score method is probably lesser than this number of 5.
Anyway, we can rely on this method of box-plot, the number of outliers detected with this method is relatively less, comparing to the numbers of samples we have (which is 506)

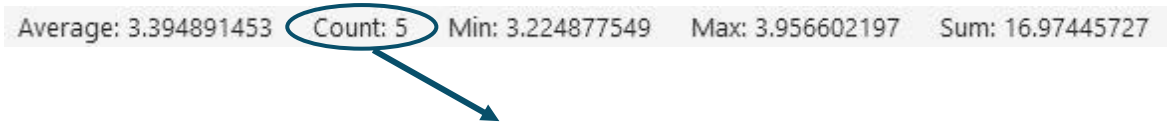
Statistical Analyses (DIS Variable)

Outliers Detecting (Z-Score Method)

	A	B
1	DIS (Miles)	DIS (Z transformation)
353	10.7103	3.284049986
354	10.7103	3.284049986
355	12.1265	3.956602197
356	10.5857	3.224877549
357	10.5857	3.224877549

Outliers With Z-Score Method:

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.
So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I got the result as you can see in the table.



5 Outliers:

5 outliers are detected based on Z-Score method.
While, the number of outliers which were detected based on box-plot method was also 5.

Box-Plot VS Z-Score :

When we compare the results of these two methods for detecting outliers for DIS feature, there is no difference between these two methods.
With Box-Plot method we got 5 outliers
With Z-Score method we got 5 outliers
Any record of DIS variable which were detected as outlier which Z-Score method, was also detected as outlier with Box-Plot method.
This ensures us enough that we can rely on these five samples to be known as outliers.

Statistical Analyses (DIS Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, DIS variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.

So; I transformed this variable with box-cox method, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.

On the second step, I removed the outliers of this variable and again applied a normality test to see if it now follows a normal distribution and the answer to this question was also negative.

So, as the conclusion, we find it out that we cannot convert the DIS variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

Box-Cox transformation
1.264869763
1.418585024
1.418585024
1.571460133
1.571460133

Transformed data of AGE Variable With Box-Cox Method

Anderson-Darling test (Box-Cox transformation):	
A^2	3.773
p-value (Two-tailed)	**≤0.0001**
alpha	0.050

Normality Test After Box-Cox Transformation :

As we can see, the result of the normality test of transformed data (with box-cox method), DIS variable still does not follow a normal distribution.

Anderson-Darling test (Box-Cox transformation):	
A^2	4.074
p-value (Two-tailed)	**≤0.0001**
alpha	0.050

Normality Test After Removing Outliers :

As we can see, even after removing the outliers of the DIS variable and conducting a normality test again, this variable is not following a normal distribution.

Statistical Analyses (DIS Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:

I am going to check the correlation between DIS variable and target variable which is MEDV, these are both continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	DIS (Miles)	MEDV (1000$)
DIS (Miles)	**1**	**0.250**
MEDV (1000$)	**0.250**	**1**

Weak And Direct Correlation:

The correlation matrix and the value of 0.25 tells us that there is a direct correlation between these 2 variables.
Meaning that if one of the increase, the other one will increase as well.
On the other hand, the absolute value would be 0.25, which indicates that the correlation is relatively weak.

Coefficients of determination (Pearson):

Variables	DIS (Miles)	MEDV (1000$)
DIS (Miles)	**1**	0.062
MEDV (1000$)	0.062	**1**

p-values (Pearson):

Variables	DIS (Miles)	MEDV (1000$)
DIS (Miles)	**0**	**<0.0001**
MEDV (1000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:

The value is <0.0001 suggests that the correlation between DIS and MEDV is statistically significant and it is not due to random changes.

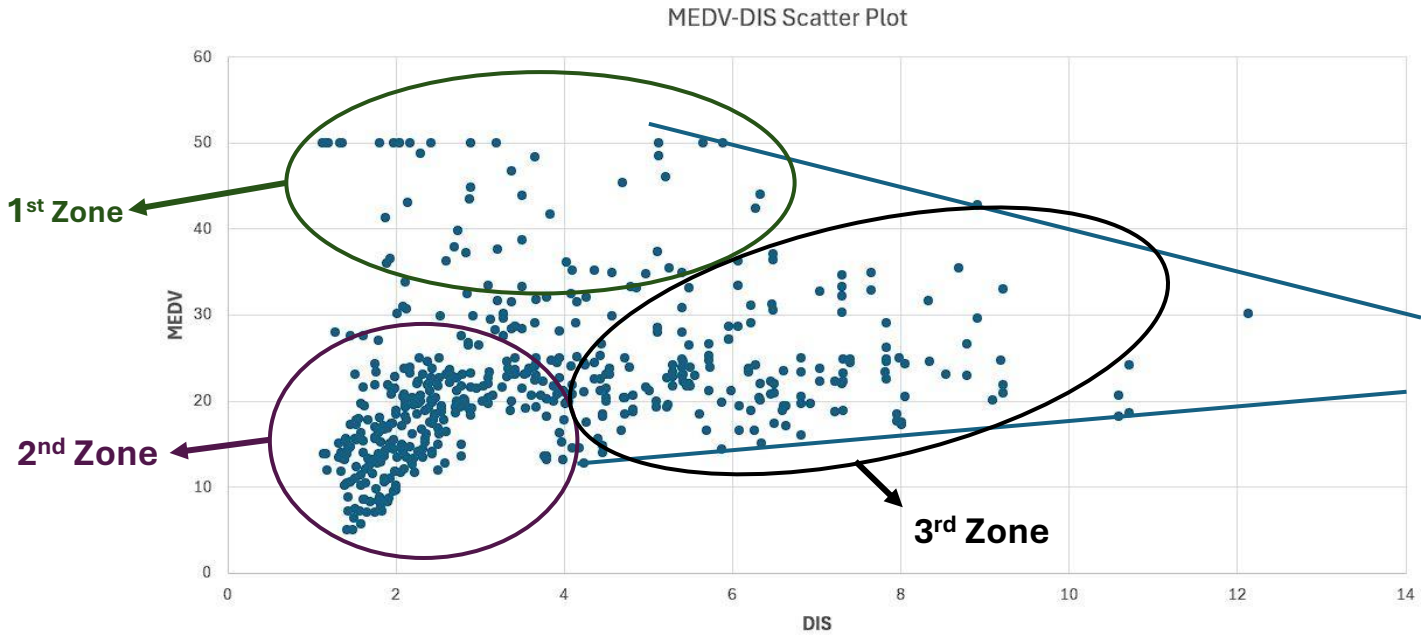
Power Of Prediction:

The value of 0.062 in this table, indicates that only 6.2% of the variance in target variable (MEDV) can be explained by the variance in DIS variable.

Statistical Analyses (DIS Variable)

Scatter Plot With The Target Variable

3 Zones:
As we see on the chart, we can divide the city into 3 zones based on median value of houses in each area and DIS rates of each area.
This gives us an interesting insight as we can interpreter as following :



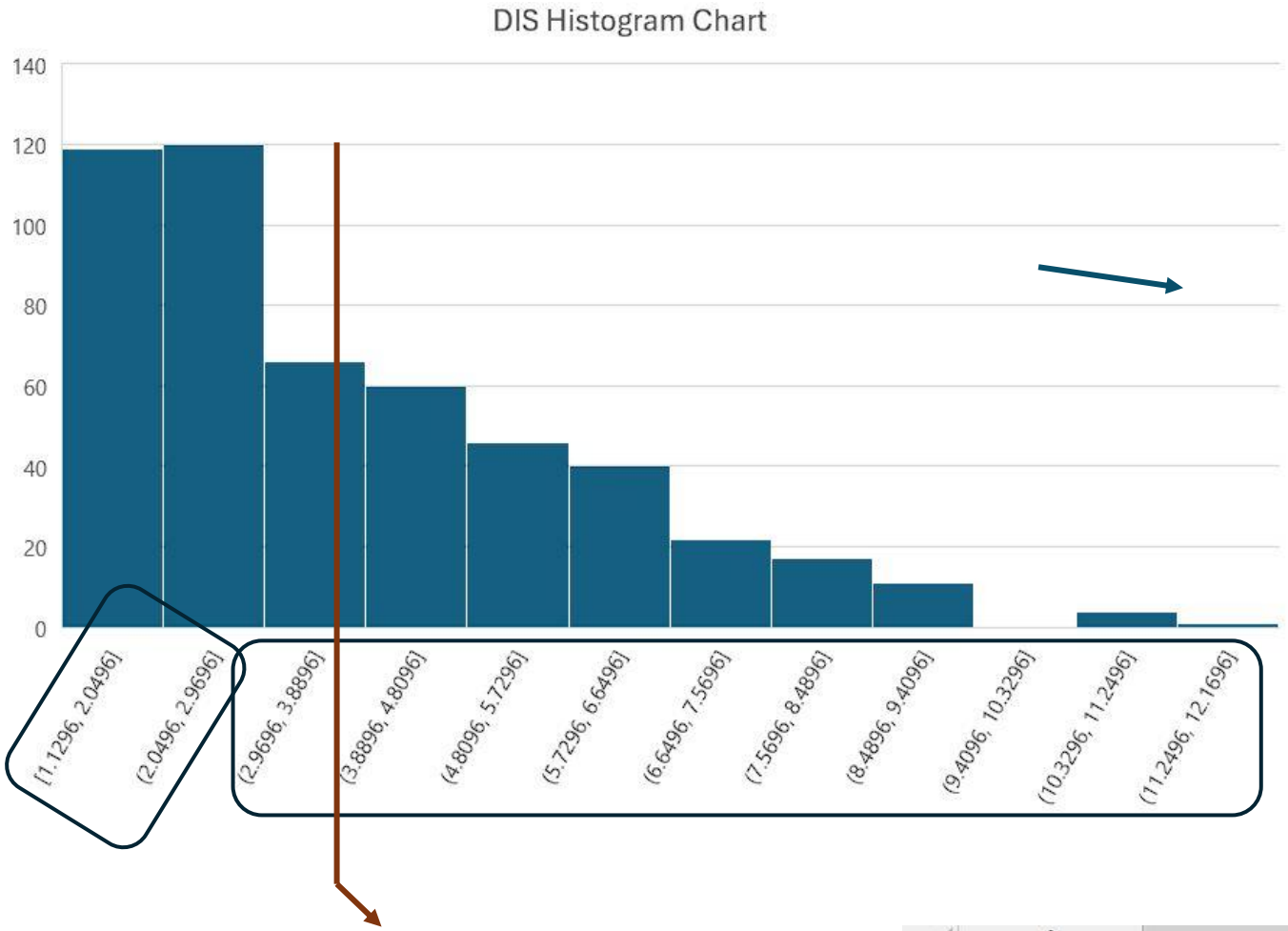
1st Insight:
Records which are included in this zone, have the highest values of MEDV variable.
Also; the concentration of records in this zone, is relatively lower than the other zones.
We can attribute records of this zone to upper-class families.
Weighted distance of these areas to 5 employment centers of Boston is less than 3rd zone and greater than 2nd zone.
Most business-owners may probably be in this zone.
Records of this zone have a logical weighted distance to 5 employment centers of Boston, not so close, not so far.

2nd Insight:
Records of this zone are so concentrated and they have the lowest values of DIS variable.
Meaning that they are the nearest areas to 5 employment centers of Boston.
This zone probably is mostly made of working-class families which prefer to be so close to their workplace and the also want houses which are as cheapest as possible.

3rd Insight:
Records of this zone are not as concentrated as 2nd zone and also more concentrated than 1st zone.
My guess is that this zone is mostly made of middle-class families. House prices is this zone is relatively higher than the 2nd zone.
Records of this zone have the highest values of DIS variable, meaning that they are far away from 5 employment centers of Boston.
This zone can be a potential option for retired parents.

Statistical Analyses (DIS Variable)

Question : Is There Any Difference Between Average Of House Prices Based On DIS Rates?



AGE 2 Classes :

I am going to create a new feature based on DIS feature.
This feature is going to be 0 for areas which DIS rate in them is lower than average.
And is going to be 1 for areas which DIS rate in them is greater than average.
I chose these 2 classes because of the distribution of the DIS variable. This variable is highly and positively skewed.
The first two bars of its histogram have much more frequencies than the other one. Meaning that the weighted distance of the most of areas of Boston is in the same range.
On the other hand, the average of this variable is so close to these two bars.
As a result, I am going to examine that if there is any difference between the average of MEDV, between these two classes.

Average :

The average of the DIS variable is equal to 3.795

	A	B	C
1	DIS (Miles)	DIS Binary Classification	MEDV (1000$)
2	4.09	=IF(A2>3.795,1,0)	24
3	4.9671	1	21.6
4	4.9671	1	34.7
5	6.0622	1	33.4
6	6.0622	1	36.2

Statistical Analyses (DIS Variable)

Variances Equality Test (Leven's Method)

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	10.397
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	**0.001**
alpha	0.050

Variances Equality Test:

As we can see, P-value is less than alpha, so; we should reject the null hypothesis. Meaning that variance of house prices with class 1 (DIS variable for them is above its average) , is not equal to variance of house prices with class 0 (DIS variable for them is below the average).
So; for comparing the average of house prices between these two classes, with should not assume the equality of variances.

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of NOX variable, we should use the appropriate method which is Leven's test.
If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (DIS Variable)

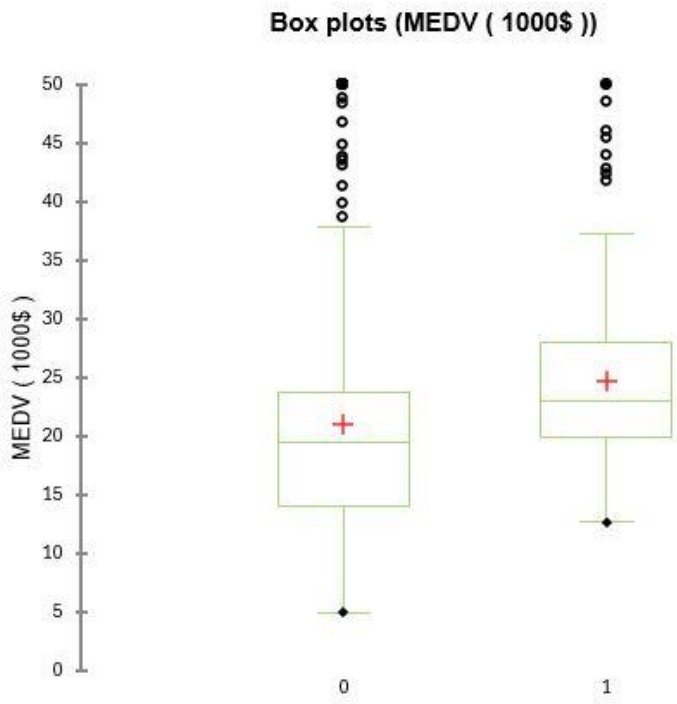
Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[-5.221, -2.189]

Difference	-3.705
t (Observed value)	-4.801
|t| (Critical value)	1.965
DF	503.114
p-value (Two-tailed)	**<0.0001**
alpha	0.050



Higher Average Of House Prices:

This tells us that areas of class 1, have higher house prices on average, comparing to areas of class 0
Meaning that on average, houses which have DIS rate above the average, have higher prices

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have DIS rates below the average (class 0) is not equal to the average of house prices for those areas which have DIS rates above the average (class 1).

Statistical Analyses (DIS Variable)

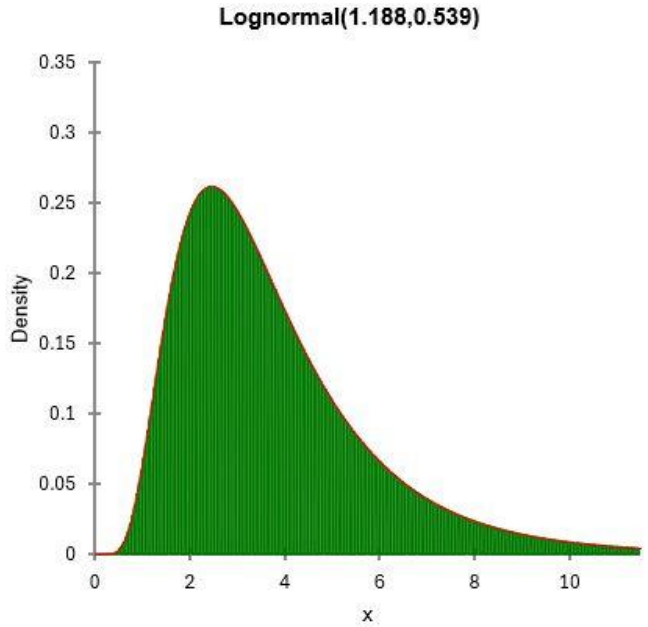
The Best Fitting Distribution

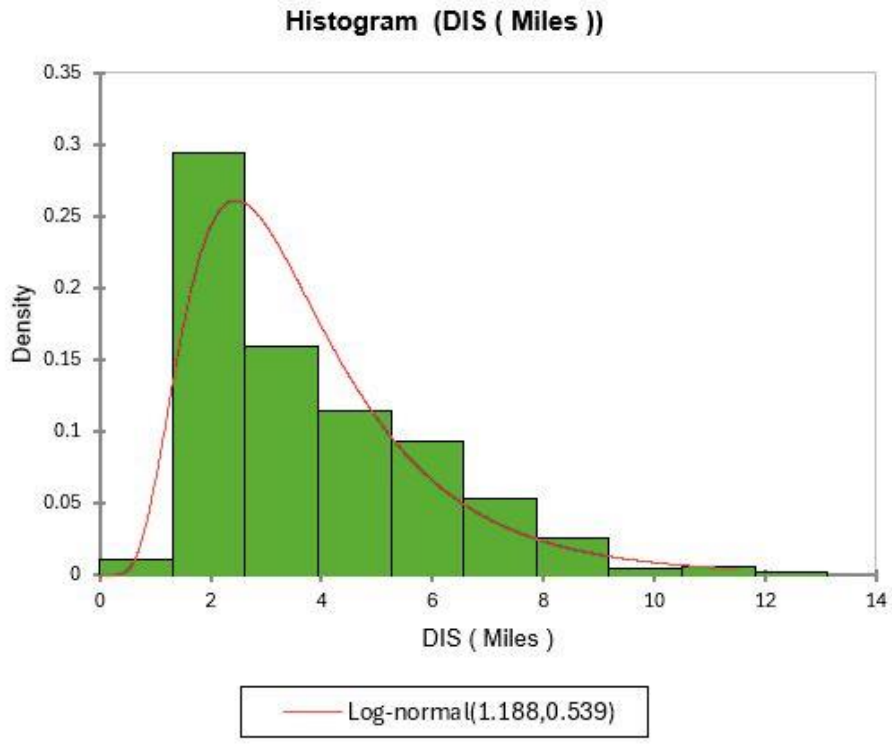
Distribution	p-value
Chi-square	<0.0001
Erlang	0.000
Exponential	<0.0001
Fisher-Tippett (1)	<0.0001
Gamma (1)	<0.0001
Gamma (2)	0.001
GEV	<0.0001
Gumbel	<0.0001
Log-normal	**0.014**
Logistic	<0.0001
Normal	<0.0001
Student	<0.0001
Weibull (1)	<0.0001
Weibull (2)	0.000

Log-Normal Distribution :

With use of XLSTAT, I found out that the best fitting distribution for DIS variable, is log-normal distribution with given parameter as below (μ & σ) Then again, with use of XLSTAT I plot the distribution with this parameters and its corresponding value and I got the chart which you can see on the right, which seems so suit for DIS variable considering this variable's distribution.

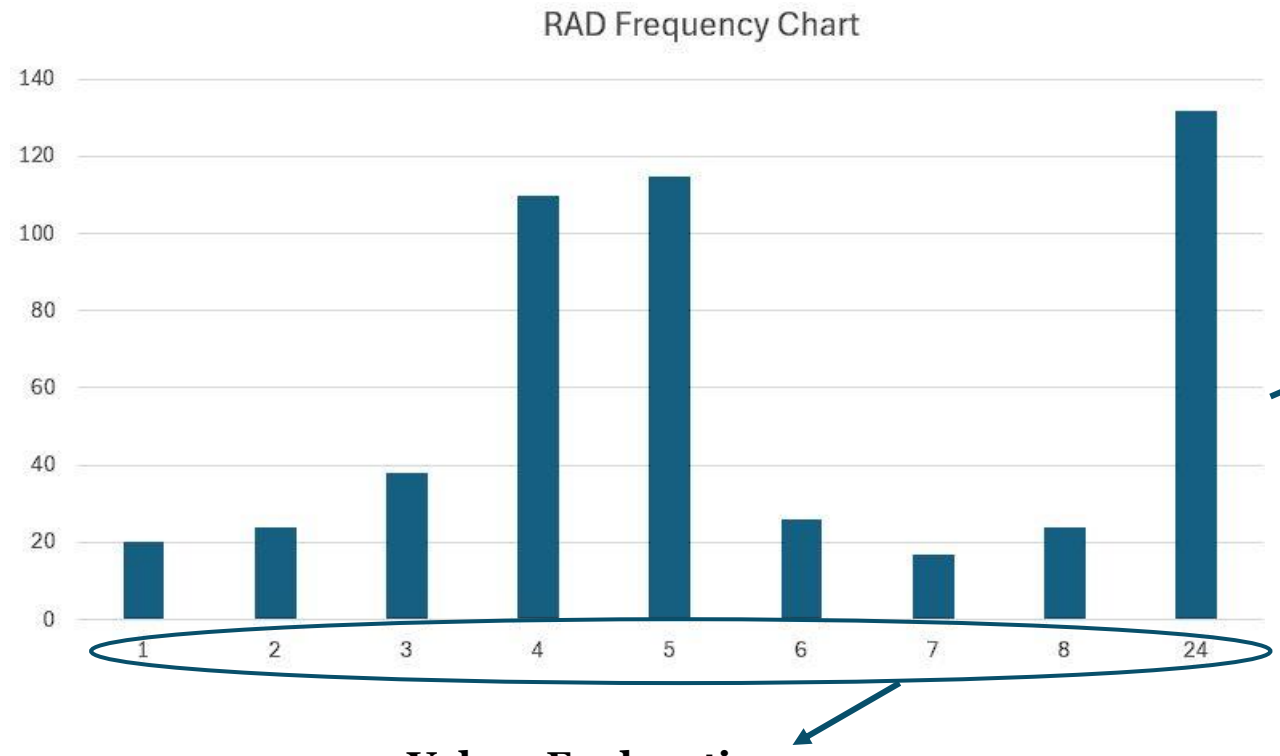
Parameter	Value	Standard error
μ	1.188	0.024
sigma	0.539	0.017





Statistical Analyses (RAD Variable)

Examining The Distribution



RAD Frequency Chart:

The RAD variable in the Boston Housing dataset, which represents the index of accessibility to radial highways, exhibits an interesting distribution pattern. The frequency chart for RAD reveals distinct clustering among certain values, reflecting the variation in accessibility levels across different neighborhoods.

The dataset contains RAD values of 1, 2, 3, 4, 5, 6, 7, 8, and an outlier value of 24. The frequency table indicates that the majority of neighborhoods have RAD values within the range of 1 to 8, highlighting varying degrees of highway accessibility. Notably, the value 24 stands out as it is significantly higher, representing areas with exceptional access to radial highways. This high value suggests the presence of major transportation hubs or highly connected regions, which can have substantial implications for urban mobility and property values.

This distribution provides valuable insights into the connectivity of different neighborhoods, underscoring the importance of transportation infrastructure in urban planning and real estate dynamics. By examining these frequencies, we can better understand the spatial distribution of accessibility and its impact on community development.

Below is the frequency table that visually represents the distribution of the RAD variable.

Values Explanation :

The values of this variable, range from 1 to 24, with higher values indicating better accessibilities to radial highways.

Anomaly At 24 :

The value 24 stands out as it's significantly higher than the other values. This indicates that locations with a RAD value of 24 have exceptionally good access to radial highways, possibly representing a major hub or highly connected area.

Row Labels	Count of RAD
1	20
2	24
3	38
4	110
5	115
6	26
7	17
8	24
24	132
Grand Total	**506**

Statistical Analyses (RAD Variable)

Examining The Descriptive Statistics

Variable\ Statistic	Nbr. of observations	Nbr. of missing values	Sum of weights	Nbr. of categories	Mode	Mode frequency	Categories	Frequency per category	Rel. frequency per category (%)	Lower bound on frequencies (95%)	Upper bound on frequencies (95%)	Proportion per category	Lower bound on proportions (95%)	Upper bound on proportions (95%)
RAD	506	0	506	9	24	132	1	20.000	3.953	2.255	5.650	0.040	0.023	0.057
							2	24.000	4.743	2.891	6.595	0.047	0.029	0.066
							3	38.000	7.510	5.214	9.806	0.075	0.052	0.098
							4	110.000	21.739	18.145	25.333	0.217	0.181	0.253
							5	115.000	22.727	19.076	26.379	0.227	0.191	0.264
							6	26.000	5.138	3.215	7.062	0.051	0.032	0.071
							7	17.000	3.360	1.790	4.930	0.034	0.018	0.049
							8	24.000	4.743	2.891	6.595	0.047	0.029	0.066
							24	132.000	26.087	22.261	29.913	0.261	0.223	0.299

There are 506 observations when we look at the “RAD” column in our dataset

All of 506 rows of this column are filled with data and there are not any missing values

Sum of weights for a qualitative variable like RAD does not give any meaningful information, because “RAD” is not quantitative.

This shows us that “RAD” variable contains 9 categories As we knew before. Higher values indicate better accessibility to radial highways

Category 24 is the mode, so its frequency is greater than other categories

The frequency of the mode (which was for class 24) is 132

There are 9 categories in this variable column

This column shows the frequency of each category of RAD variable

This column shows the frequency of each category of RAD variable in percent

With a confidence level of 95%, we can say that on the population, different categories of RAD variable, will have given proportions.

These three columns don’t give us any new information, they just give the same information of previous last three columns

Statistical Analyses (RAD Variable)

Correlation Test With The Target Variable (Spearman Method)

Why Spearman Method:

I am going to check the correlation between RAD variable and target variable which is MEDV. MEDV is a continuous variable and RAD is an ordinal variable, and because of this reason I should use appropriate corresponding method; which for checking the correlation between these two kinds of variables is Spearman method.

Correlation matrix (Spearman):

Variables	RAD	MEDV (1000$)
RAD	**1**	**-0.347**
MEDV (1000$)	**-0.347**	**1**

Moderate And Inverse Correlation:

The correlation matrix and the value of -0.34 tells us that there is an inverse correlation between these 2 variables.
In other words, homes closer to these radial highways may have lower median values compared to those farther away.
Meaning that if one of the increase, the other one will decrease.
On the other hand, the absolute value of 0.34 indicates a moderate relationship.

p-values (Spearman):

Variables	RAD	MEDV (1000$)
RAD	**0**	**<0.0001**
MEDV (1000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:

The value is <0.0001 suggests that the correlation between RAD and MEDV is statistically significant and it is not due to random changes.

Coefficients of determination (Spearman):

Variables	RAD	MEDV (1000$)
RAD	**1**	0.120
MEDV (1000$)	0.120	**1**

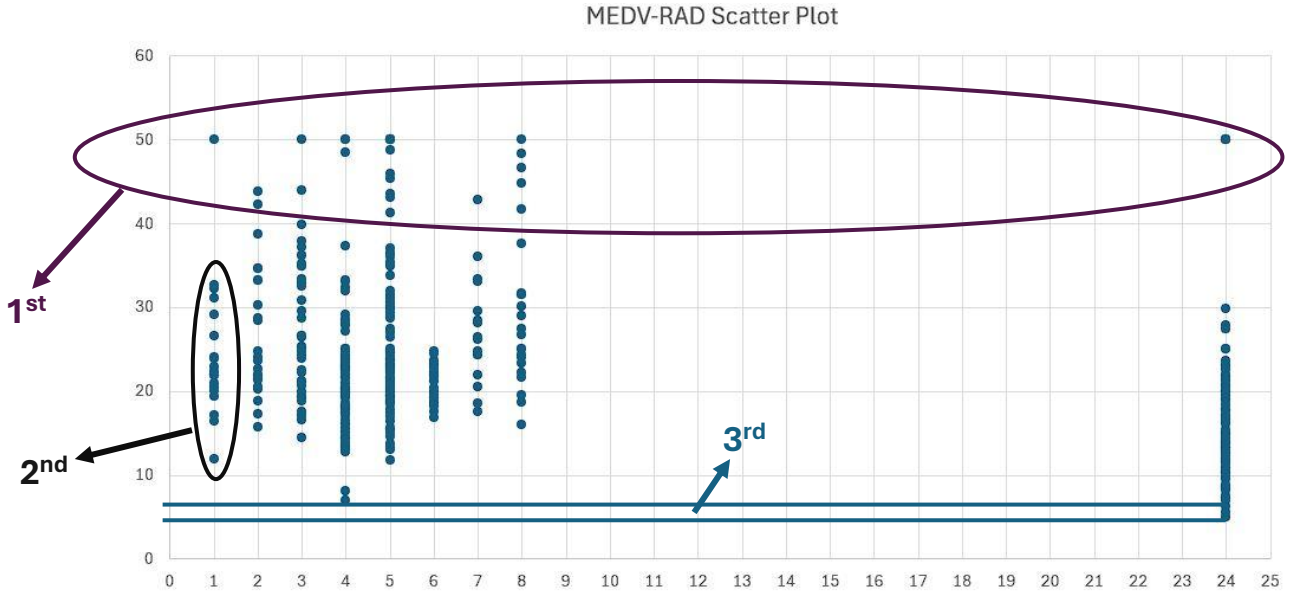
Power Of Prediction:

The value of 0.12 in this table, indicates that only 12% of the variance in target variable (MEDV) can be explained by the variance in RAD variable.

Statistical Analyses (RAD Variable)

Scatter Plot With The Target Variable

Comparing Different Areas With Different Weighted Distances From Radial Highways:
With paying attention to the scatter plot, we can draw some visual insights from it as are mentioned below :



1st Insight :

It seems that for the highest values of MEDV, accessibility is not an issue. Although the concentration of samples are different between different categories; but as we can see on the chart, for each different levels of accessibility, there are some areas that have higher MEDVs than the other samples.

2nd Insight :

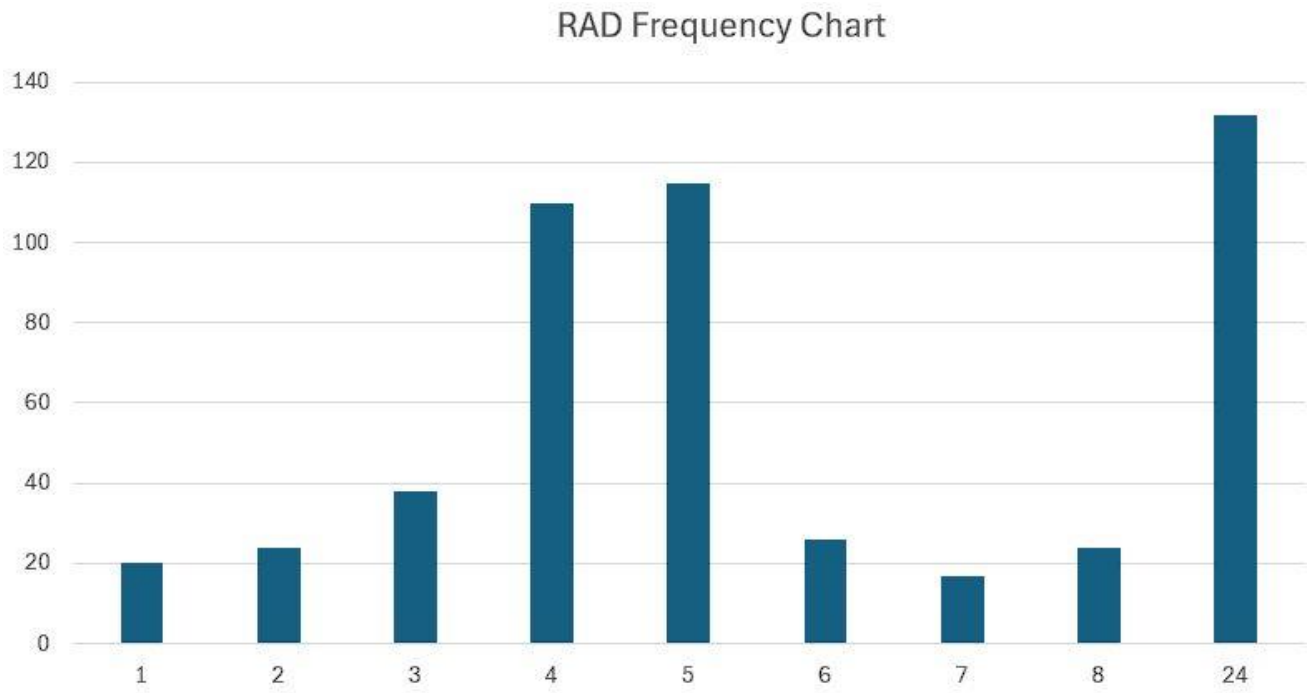
If we look at the area which is marked with number 2, we can see that for category number 1 (which has the worst accessibility to radial highways), the MEDV (median value of owner-occupied homes) tends to vary within a narrower range compared to the other categories.
Meaning that, samples in this category, are not popular choices, whether for upper-class or middle-class or working-class families.

3rd Insight :

There are some areas in category number 24, which have the lowest MEDVs. These probably are areas which are so close to radial highways which this characteristic of them, make them not desirable for anyone, and make the cheap.

Statistical Analyses (RAD Variable)

Question : Is There Any Difference Between Average Of House Prices Based On RAD Classification ?



RAD Classification :

To explore the relationship between housing values and accessibility to radial highways, we examined the average MEDV (median value of owner-occupied homes) across different RAD categories. The RAD variable, which measures the index of accessibility to radial highways, is categorized into several groups, each representing different levels of connectivity to major roadways.

Our analysis aims to determine whether there are significant differences in the average MEDV among these RAD categories. By comparing the mean housing values across the various RAD groups, we can gain insights into how proximity to radial highways impacts property values. This examination is crucial for urban planners and real estate analysts, as it highlights the potential influence of transportation infrastructure on housing market dynamics.

Preliminary results suggest variability in median home values depending on the level of highway accessibility. These findings could inform decisions related to urban development, zoning, and investment strategies, ensuring a balanced approach to infrastructure and residential planning

Statistical Analyses (RAD Variable)

Question : Is There Any Difference Between Average Of House Prices Based On RAD Categories? (ANOVA Test Results, 1st Page)

Observations	506
Sum of weights	506
DF	497
R^2	0.229
Adjusted R^2	0.216
MSE	66.296
RMSE	8.142
MAPE	29.268
DW	0.711
Cp	9.000
AIC	2131.147
SBC	2169.186
PC	0.799

Adjusted R Squared:
With variance of RAD variable, we can anticipate 21.6% of variance of the target variable which is MEDV

Correlation matrix:

	RAD-1	RAD-2	RAD-3	RAD-4	RAD-5	RAD-6	RAD-7	RAD-8	RAD-24	MEDV (1000$)
RAD-1	1	-0.045	-0.058	-0.107	-0.110	-0.047	-0.038	-0.045	-0.121	0.040
RAD-2	-0.045	1	-0.064	-0.118	-0.121	-0.052	-0.042	-0.050	-0.133	0.104
RAD-3	-0.058	-0.064	1	-0.150	-0.155	-0.066	-0.053	-0.064	-0.169	0.167
RAD-4	-0.107	-0.118	-0.150	1	-0.286	-0.123	-0.098	-0.118	-0.313	-0.066
RAD-5	-0.110	-0.121	-0.155	-0.286	1	-0.126	-0.101	-0.121	-0.322	0.187
RAD-6	-0.047	-0.052	-0.066	-0.123	-0.126	1	-0.043	-0.052	-0.138	-0.039
RAD-7	-0.038	-0.042	-0.053	-0.098	-0.101	-0.043	1	-0.042	-0.111	0.093
RAD-8	-0.045	-0.050	-0.064	-0.118	-0.121	-0.052	-0.042	1	-0.133	0.190
RAD-24	-0.121	-0.133	-0.169	-0.313	-0.322	-0.138	-0.111	-0.133	1	-0.396
MEDV (1000$)	0.040	0.104	0.167	-0.066	0.187	-0.039	0.093	0.190	-0.396	1

Correlations Between Different RAD Categories :
Values which are in this right triangle, show the correlation between different Categories of RAD variable.
As we can see all the values show inverse correlations.
And category number 3 has the strongest , inverse correlation with category number 2

Correlations Between Different RAD Categories With Target Variable :
These values in blue box show the correlation between different RAD categories with the target variable
As we can see, there is a relatively strong and inverse correlation between target variable and category number 24 of RAD variable. As we mentioned before, this category includes areas with the best accessibility to radial highways.
Meaning that for those areas which are relatively closer to radial highways, MEDV will drop.
Target variable has a direct and relatively stronger correlation with category number 8, it indicates that in these areas (which have good accessibility to radial highways, but not so close to them) MEDV will slightly increase as a result of this characteristic of these areas.

Statistical Analyses (RAD Variable)

Question : Is There Any Difference Between Average Of House Prices Based On RAD Categories? (ANOVA Test Results, 2nd Page)

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	8	9767.260	1220.907	18.416	**<0.0001**
Error	497	32949.036	66.296		
Corrected Tota	505	42716.295			

There Is Difference Between RAD Different Categories :

This number here, tells us that there is a meaningful difference between different categories of RAD variable in terms of MEDV.
Meaning that if want to created a model for MEDV variable, it should include RAD variable.
RAD is an effective element on MEDV variable.

Parameters Of The Model :

This column shows all of the parameters which are used in the model to anticipate the MEDV based on RAD categories.

Model parameters (MEDV (1000$)):

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	16.404	0.709	23.147	**<0.0001**	15.011	17.796
RAD-1	7.961	1.954	4.075	**<0.0001**	4.123	11.800
RAD-2	10.430	1.807	5.772	**<0.0001**	6.880	13.979
RAD-3	11.525	1.499	7.689	**<0.0001**	8.580	14.470
RAD-4	4.983	1.051	4.741	**<0.0001**	2.918	7.049
RAD-5	9.303	1.039	8.957	**<0.0001**	7.263	11.344
RAD-6	4.573	1.747	2.618	**0.009**	1.141	8.006
RAD-7	10.702	2.098	5.101	**<0.0001**	6.580	14.824
RAD-8	13.955	1.807	7.723	**<0.0001**	10.405	17.504
RAD-24	0.000	0.000				

Lower & Upper Bonds :

Represent the confidence interval for each coefficient estimate.
A confidence interval provides a range within which we expect the true value of the coefficient to fall, with a certain level of confidence (95%)

Coefficients Of The Model :

This column shows the corresponding coefficients for each parameters of the model.

Reliable Coefficients :

All of the values, are less than alpha, meaning that all of the coefficients which are used in the model created by the use of linear regression, are reliable.

Standard Error :

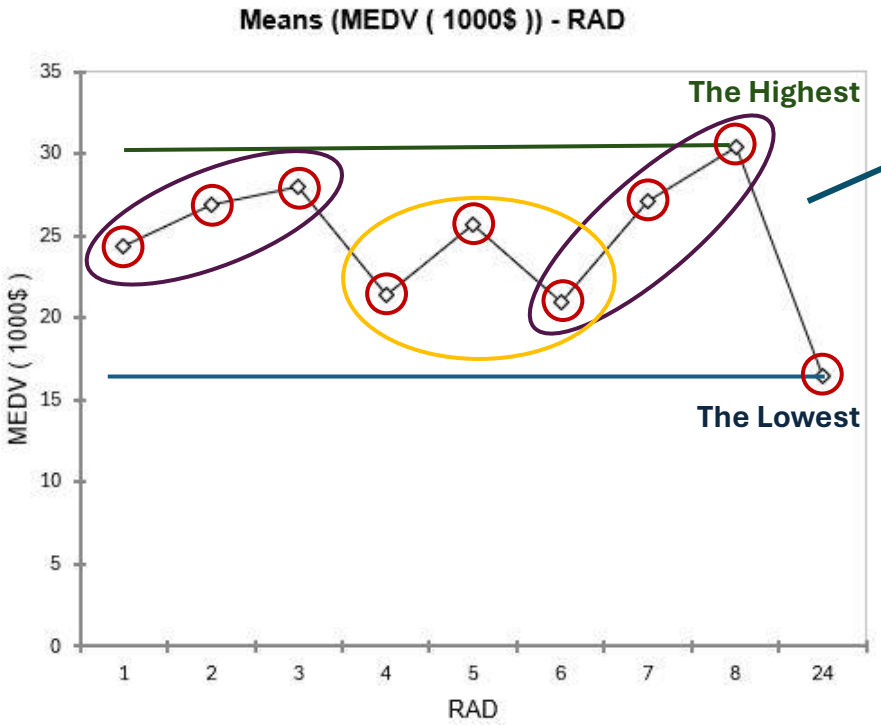
The Standard Error (often abbreviated as SE) tells us how much the estimated value of a coefficient might vary if you repeated your analysis with different samples of data.
It shows the precision of the coefficient estimate.
Smaller standard errors indicate more precise estimates.

T-Statistics :

The t-statistic is a measure used in hypothesis testing to determine whether a coefficient is significantly different from zero.
It is calculated as the coefficient estimate divided by its standard error.
High Absolute Value: A high absolute value of the t-statistic (either positive or negative) suggests that the corresponding predictor is significantly different from zero, implying a significant effect on the dependent variable.
Low Absolute Value: A low absolute value suggests that the predictor is not significantly different from zero, implying it might not have a significant effect.

Statistical Analyses (RAD Variable)

Question : Is There Any Difference Between Average Of House Prices Based On RAD Categories? (ANOVA Test Results, 3rd Page)



Comparing The Average Of MEDVs Of Different RAD Categories:

This chart shows are the average of MEDVs of each category of RAD variable.

As we can see, category number 8 has the highest value of MEDVs on average and the category number 24 has the lowest value of MEDVs on average.

Consider the interval between category number 1 to category number 3, also the interval between category number 6 to category number 8 (purple ovals)
In these two intervals, there is a direct correlation between categories and average of MEDVs of each category, as we can see; as accessibility gets better, average of MEDV goes higher (in these two mentioned intervals)

On the other hand, consider the interval between category number 4 to category number 6.(yellow oval)
Something happens in this interval which does not obey the rule and is destroying our pattern.
It can be considered as a case study to get deeper in.

And finally, category number 24, this category has the best accessibility to radial highways, and has the lowest average of MEDVs. It can probably be due to its short distance to highways which make areas of this category noisy and crowded.

Summary of all pairwise comparisons for RAD (Tukey (HSD)):

Category	LS means(MEDV (1000$))	Groups		
8	30.358	A		
3	27.929	A		
7	27.106	A	B	
2	26.833	A	B	
5	25.707	A	B	
1	24.365	A	B	
4	21.387		B	C
6	20.977		B	C
24	16.404			C

Grouping Categories Of The RAD Variable :

According to the table on the left, we can group different categories of the RAD variable.
We can divide them in 4 different groups.
Categories which are in same group, approximately have same average of MEDVs, so; we can group them.
4 groups are labeled as : A, AB, BC, C
Comparing these groups to each other is another issue that we will talk about in the next slide.
By now, it is enough for us to know, which categories have approximately same average of MEDVs.

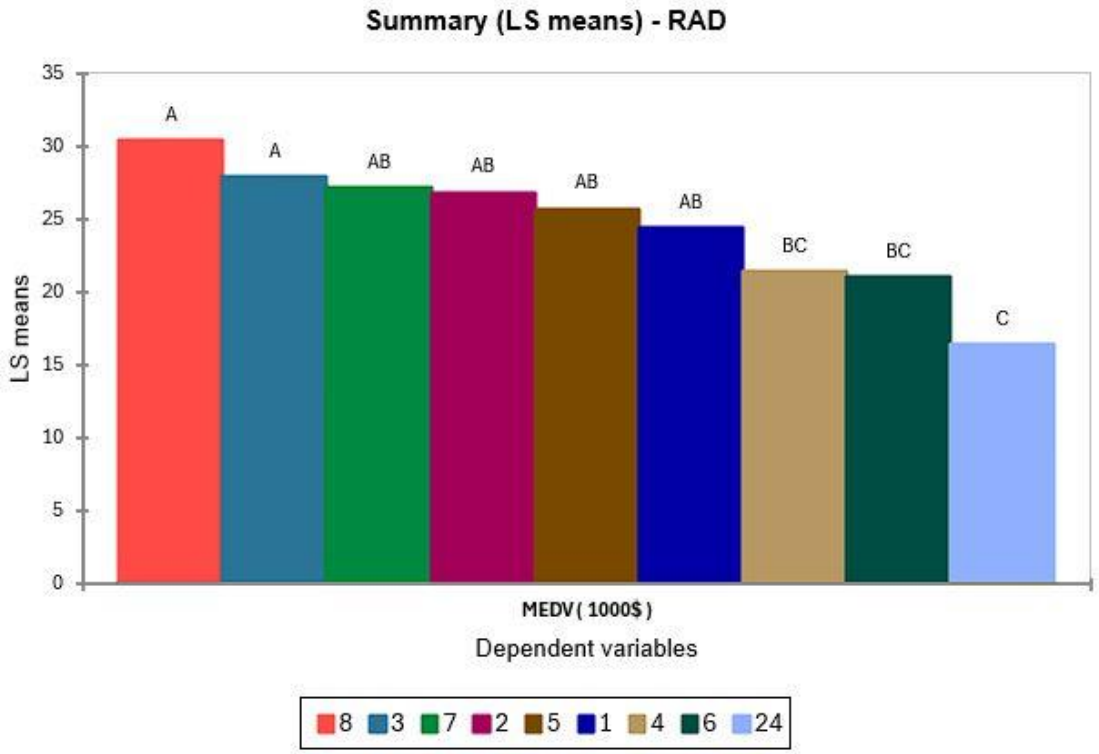
Statistical Analyses (RAD Variable)

Question : Is There Any Difference Between Average Of House Prices Based On RAD Categories? (ANOVA Test Results, 4th Page)

	MEDV (1000$)
8	30.358
3	27.929
7	27.106
2	26.833
5	25.707
1	24.365
4	21.387
6	20.977
24	16.404

The Highest To The Lowest :

Table on the right, ranks the different categories of RAD variable, from the highest (category number 8) to the lowest (category number 21) in terms of average of their MEDVs.
On the previous slide, we compare them on the “Means” chart, and now we have each category and its corresponding average of its MEDVs.

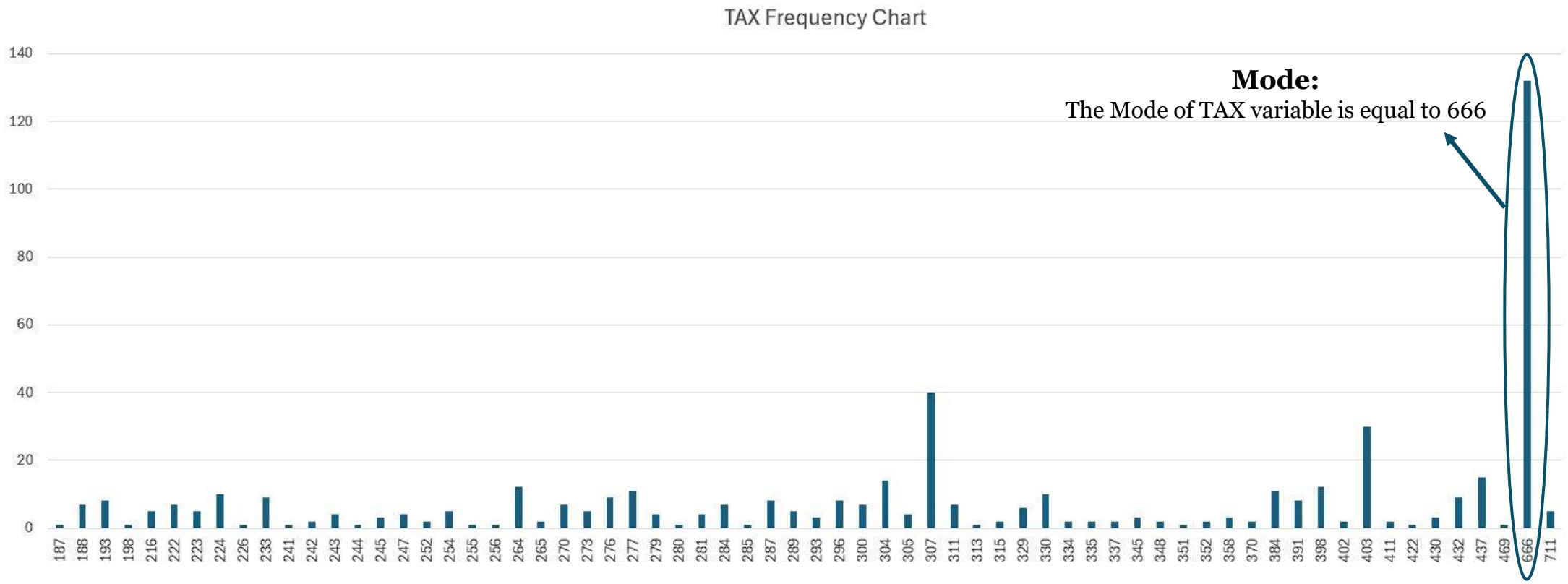


Group Comparing :

On the previous slide, we saw that we can group some categories of the RAD variable together, that was because that the average of their MEDVs was approximately the same.
And we said that comparing these groups can be an issue that we can talk about.
On the left, we have a bar chart that compares these groups.
As we can see, group A (which includes the categories number 8 and 3)has the highest average of MEDVs and group C (which only includes the category number 24) has the lowest.
This grouping method can help us with different purposes.
Whether it is how for planning to buy a house or urban planning and so on.

Statistical Analyses (TAX Variable)

Examining The Distribution



TAX Frequency Chart:

The TAX variable in the Boston Housing dataset, representing the full-value property tax rate per $10,000, exhibits a distinct distribution pattern across the dataset. The frequency chart provides a visual representation of how the TAX rates are distributed among the different samples. By examining the frequency chart, we can observe that certain tax rates are more common than others, indicating clusters where the majority of properties fall within specific tax brackets.

Notably, the chart highlights several peaks, suggesting that specific tax rates occur more frequently, possibly reflecting local policies or regional tax regulations that influence the property tax rates in those areas. Understanding the distribution of the TAX variable is crucial for analyzing the financial burden on property owners and its potential impact on housing values and market dynamics.

This analysis provides valuable insights into the taxation landscape within the dataset, offering a clearer picture of how property tax rates vary and what patterns emerge from the data.

Statistical Analyses (TAX Variable)

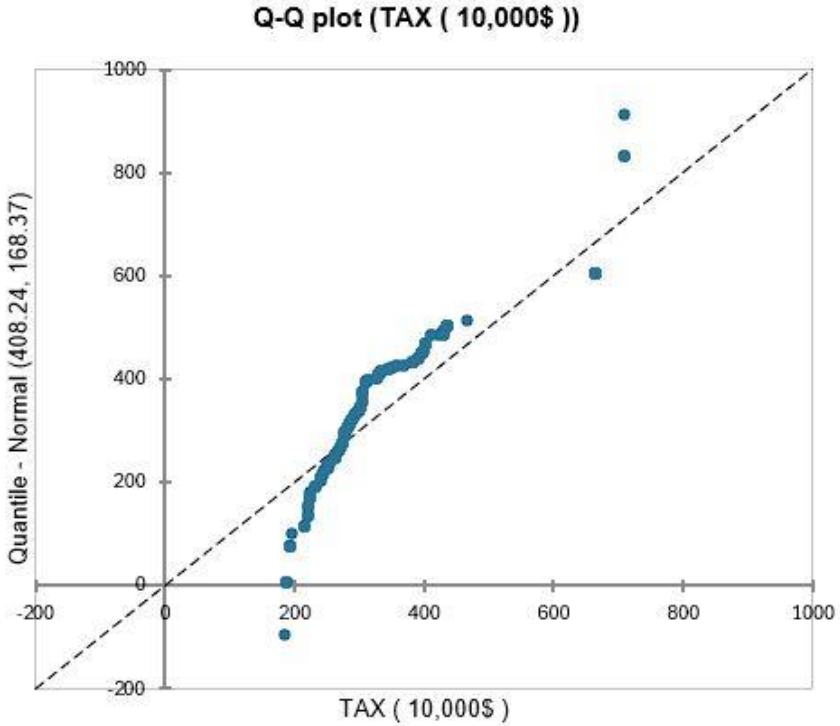
Examining The Descriptive Statistics

Statistic	TAX (10,000$)
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	187.000
Maximum	711.000
Freq. of minimum	1
Freq. of maximum	5
Range	524.000
1st Quartile	279.000
Median	330.000
3rd Quartile	666.000
Sum	206568.000
Mean	408.237
Variance (n)	28348.624
Variance (n-1)	28404.759
Standard deviation (n)	168.370
Standard deviation (n-1)	168.537
Skewness (Pearson)	0.668
Kurtosis (Pearson)	-1.143
Lower bound on mean (95%)	393.517
Upper bound on mean (95%)	422.957
Lower bound on variance (95%)	25202.727
Upper bound on variance (95%)	32261.674

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen only 1 time among all of the records
- Maximum value of this variable can be seen 5 times among all records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A skewness value of 0.66 for the DIS variable indicates a moderate positive skewness, meaning that while most distances are clustered towards shorter values, there are a few larger values that pull the distribution to the right.
- The peak of the distribution is lower and broader than that of a normal distribution. there are fewer extreme outliers in the data. In other words, the TAX values are more evenly spread out without significant high or low extremes.
- The mean of the population of this variable must be something between 393.5 and 422.9 with confidence level of 95%
- The variance of the population of this variable must be something between 25202.7 and 32261 with confidence level of 95%

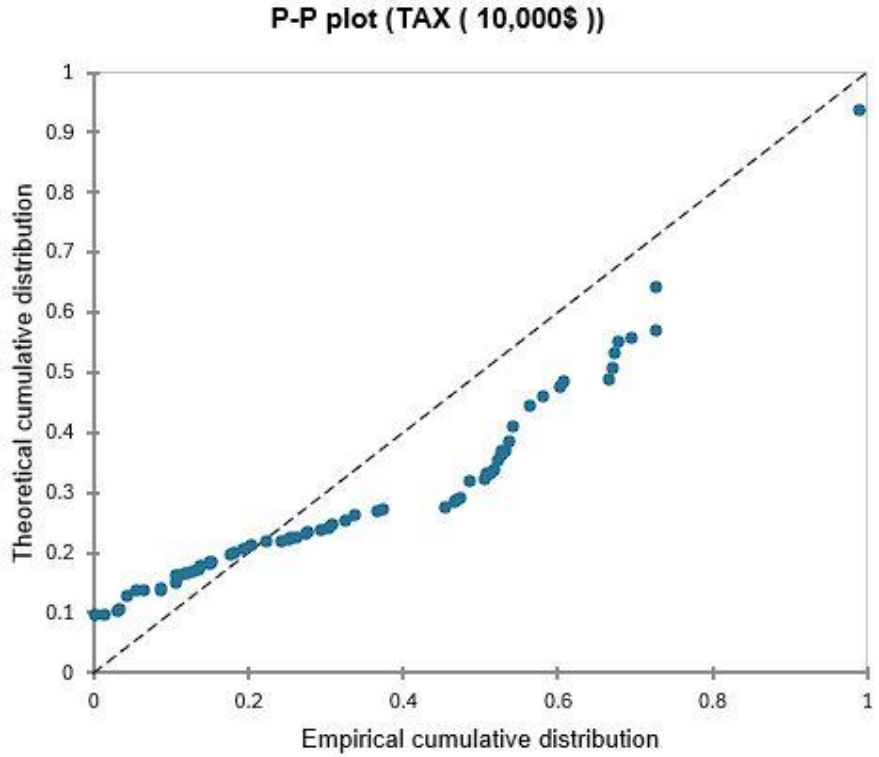
Statistical Analyses (TAX Variable)

Normality Test (Anderson-Darling Method)



Anderson-Darling test (TAX (10,000$)):	
A^2	39.351
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test Result :
p-value is less than alpha, so we should reject the null hypothesis. So; TAX variable does not follow a normal distribution.



P-P & Q-Q plot:
These plots show us that there is a difference between TAX variable's distribution and a normal distribution as the normality test's result unveiled this fact to us.

Statistical Analyses (TAX Variable)

Outliers Detecting (Variable Transformation)

Raw data:

This is the raw data of TAX variable without any transformations.

Z-Score Normalization:

In this column, I transformed the data of TAX variable with use of excel functions.
I create a function like this :

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

XLSTAT Check (Z-Score Normalization):

In this column, I transformed the data of TAX variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

	A	B	C	D	E
1	TAX (10,000$)	TAX (Z transformation)	TAX (Normalization)	Standardized (n-1)	0 to 1
2	296	-0.665949179	0.208015267	-0.665949179	0.208015267
3	242	-0.98635338	0.104961832	-0.98635338	0.104961832
4	242	-0.98635338	0.104961832	-0.98635338	0.104961832
5	222	-1.105021603	0.066793893	-1.105021603	0.066793893
6	222	-1.105021603	0.066793893	-1.105021603	0.066793893

Normalization:

In this column, I normalized the data of TAX variable with use of excel functions.
I create a function like this :

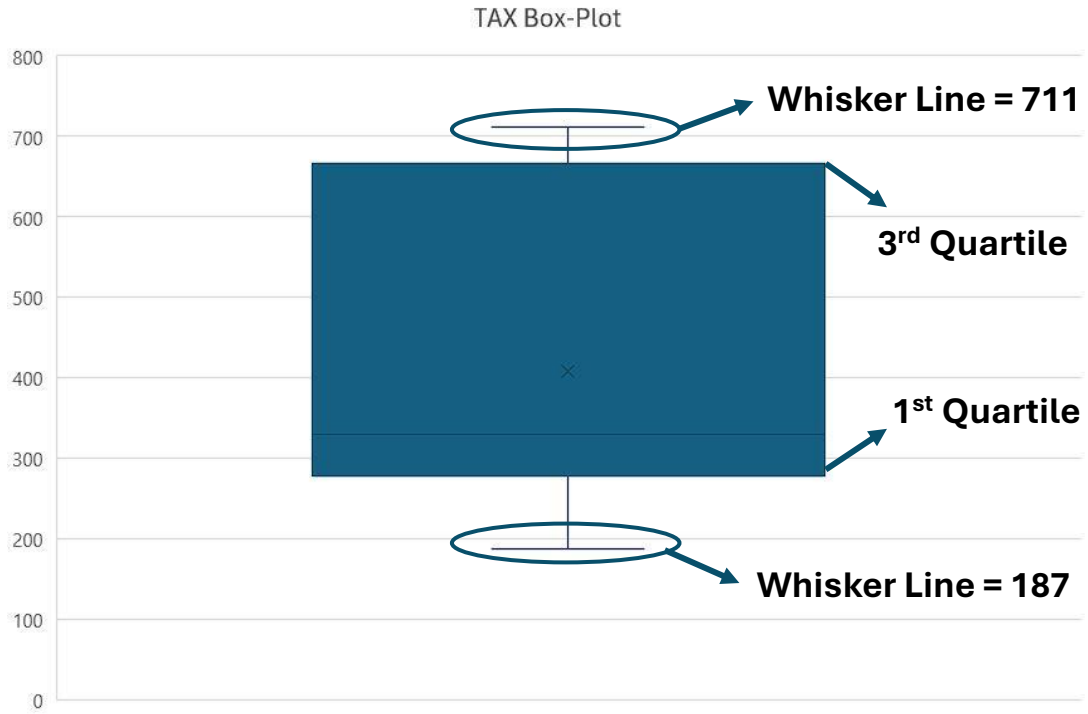
$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

XLSTAT Check (Normalization):

In this column, I normalized the data of TAX variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (TAX Variable)

Outliers Detecting (Box-Plot Method)



1st Q	279
Median	330
3rd Q	666
IQR = 3rd Q - 1st Q	387
1st Q - 1.5IQR	-301.5
3rd Q + 1.5IQR	1246.5

Below Minimum:
This value is another limit line, and any value below it, should be considered as outlier; is less than the minimum of TAX feature
So, we would not have outliers between low values of TAX feature

Above Maximum:
As we can see, this value, which is a limit line, and any value above it should be considered as outlier; is greater than the maximum of TAX variable
So, we would not have outliers between high values of TAX variable

Box-plot chart:
In our analysis of the TAX variable using a box-plot chart, we observed that there are no outliers present. This indicates that the property tax rates in our dataset are relatively consistent, with no extreme values that deviate significantly from the central distribution. The absence of outliers suggests a uniform application of tax rates across the neighborhoods, reflecting a stable and predictable tax environment. This finding is important for understanding the overall distribution of property tax rates and reinforces the reliability of our data for further analysis.

Statistical Analyses (TAX Variable)

Outliers Detecting (Z-Score Method)

	A	B
1	TAX (10,000$)	TAX (Z transformation)
2	296	-0.665949179
3	242	-0.98635338
4	242	-0.98635338
5	222	-1.105021603
6	222	-1.105021603
7	222	-1.105021603
8	311	-0.576948013
9	311	-0.576948013

No Outliers With Z-Score Method :

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.
So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I found out that there is no value between transformed data to be greater than 3 or less than -3

Box-Plot VS Z-Score :

In our examination of the TAX variable, both the box-plot and the Z-score method consistently revealed the absence of outliers. The box-plot chart confirmed that there are no values significantly deviating from the central distribution, indicating a uniform spread of property tax rates. Similarly, the Z-score method, which assesses the number of standard deviations a data point is from the mean, corroborated this finding by showing no values falling outside the typical threshold for outliers (commonly set at Z-scores beyond ±3). This consistency between the two methods strengthens our confidence in the data's reliability and the uniformity of tax rates across the dataset, providing a solid foundation for further analysis and interpretation.

Statistical Analyses (TAX Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, TAX variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.

So; I transformed this variable with box-cox method, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.

So, as the conclusion, we find it out that we cannot convert the TAX variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

TAX Box-Cox transformation
1.80725775
1.796576819
1.796576819
1.791645347
1.791645347
1.791645347
1.809710806
1.809710806

Transformed data of TAX Variable With Box-Cox Method

Anderson-Darling test (TAX Box-Cox transformation):	
A^2	15.579
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test After Box-Cox Transformation :

As we can see, the result of the normality test of transformed data (with box-cox method), TAX variable still does not follow a normal distribution.

Statistical Analyses (TAX Variable)

Correlation Test With The Target Variable (Spearman Method)

Why Spearman Method:

I am going to check the correlation between TAX variable and target variable which is MEDV.
MEDV is a continuous variable and TAX is a discrete variable, and because of this reason I should use appropriate corresponding method; which for checking the correlation between a continuous variable and a discrete variable is Spearman method.

Correlation matrix (Spearman):

Variables	TAX (10,000$)	MEDV (1000$)
TAX (10,000$)	**1**	**-0.562**
MEDV (1000$)	**-0.562**	**1**

Relatively Strong And Inverse Correlation:

The correlation matrix and the value of -0.56 tells us that there is an inverse correlation between these 2 variables.
Meaning that if one of the increase, the other one will decrease.
On the other hand, the absolute value would be 0.56, which indicates that the correlation is relatively strong.

p-values (Spearman):

Variables	TAX (10,000$)	MEDV (1000$)
TAX (10,000$)	**0**	**<0.0001**
MEDV (1000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:

The value is <0.0001 suggests that the correlation between TAX and MEDV is statistically significant and it is not due to random changes.

Coefficients of determination (Spearman):

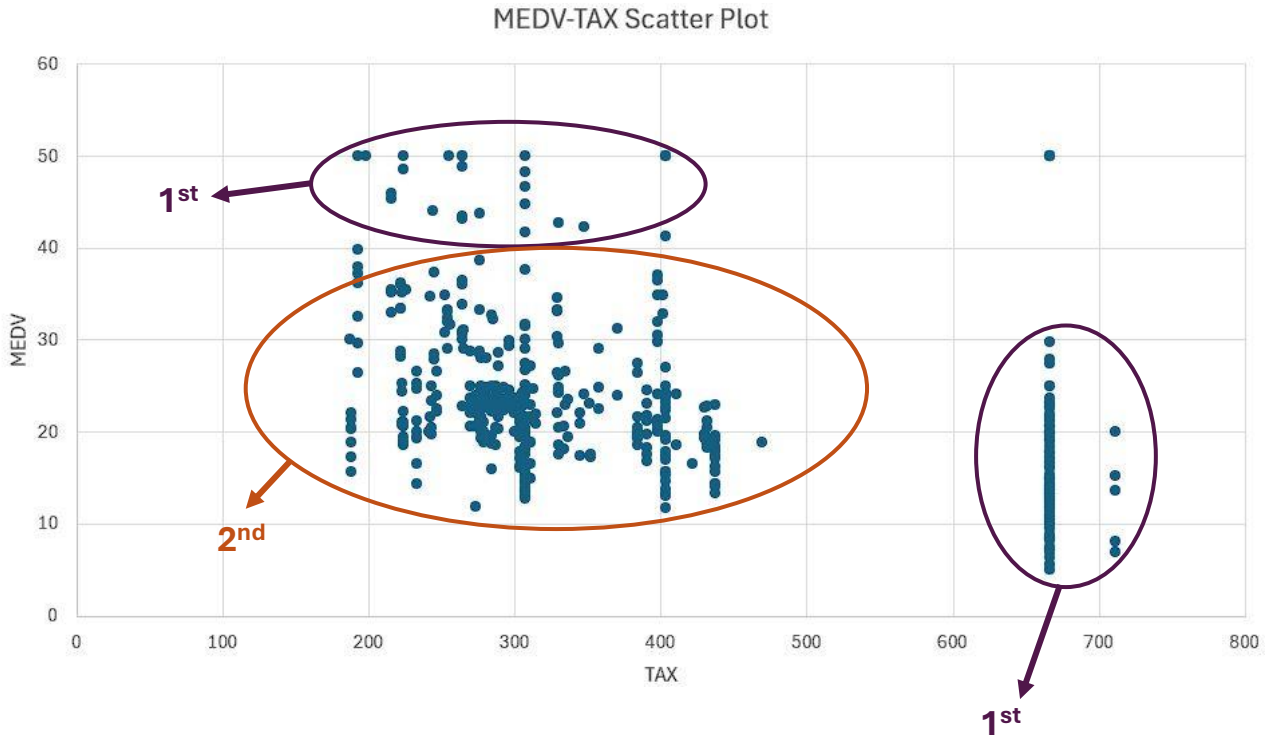
Variables	TAX (10,000$)	MEDV (1000$)
TAX (10,000$)	**1**	0.316
MEDV (1000$)	0.316	**1**

Power Of Prediction:

The value of 0.316 in this table, indicates that only 31.6% of the variance in target variable (MEDV) can be explained by the variance in TAX variable.

Statistical Analyses (TAX Variable)

Scatter Plot With The Target Variable



Insights :

As we see on the chart, we can draw some insights from the scatter plot of the MEDV and TAX variables.
Some of the insights we can discuss about are motioned below
Clearly, we can extract other hints from the plot, but by now, these would satisfy our purposes here

1st Insight:

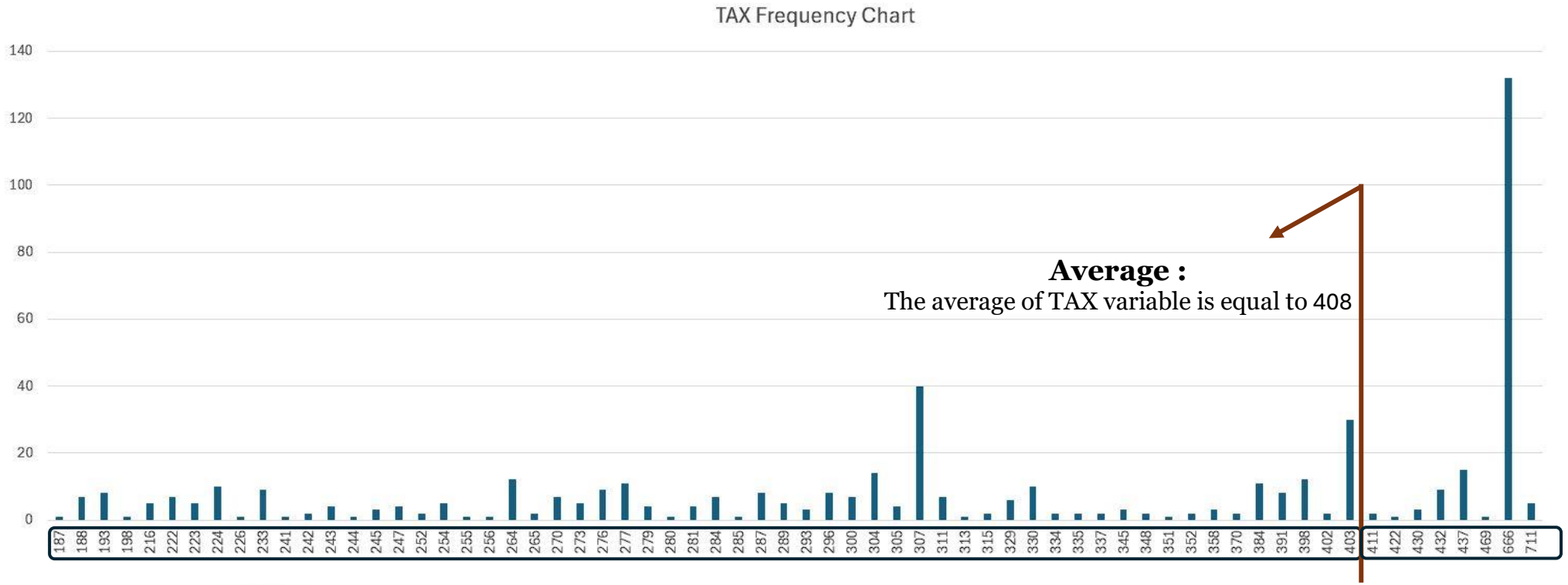
If we look at the zones which are marked as "1st" with purple ovals, we can extract a pattern from them.
Distribution of samples in these zones, show us that there is an inverse correlation between TAX and MEDV variables.
Meaning that as TAX variable increase, the MEDV will decrease.
This fact can imply that houses with high values of TAX are cheaper than the other ones, and houses with low values of TAX can be probably more expensive than others
(it becomes important here to notice that houses with lower values of TAX can vary in a wider range. They can have more variety in terms of house prices)

2nd Insight:

If we look at the zone which is marked as "2nd" with orange oval, we can extract some insights about the areas which are included in this zone.
It seems that this zone, can be attributed to middle-class families.
Samples in this zone have relatively lower values of TAX and also lower values of MEDVs.
The concentration and the variety of samples in this zone is much more than the others.
Samples in this zone, seem to make the majority of samples and the can range from relatively low value of MEDV to high values of it.

Statistical Analyses (TAX Variable)

Question : Is There Any Difference Between Average Of House Prices Based On TAX Variable?



TAX 2 Classes :

I am going to create a new feature based on TAX feature. This feature is going to be 0 for areas which TAX value in them is lower than the average.
And is going to be 1 for areas which TAX value in them is greater than the average.
I chose these 2 classes because it seems to me interesting and logical to compare these two classes in terms of the average of their MEDVs.

	A	B	C
1	TAX (10,000$)	TAX (Binary Classification)	MEDV (1000$)
2	296	=IF(A2>408,1,0)	24
3	242	0	21.6
4	242	0	34.7
5	222	0	33.4
6	222	0	36.2
7	222	0	28.7
8	311	0	22.9

Statistical Analyses (TAX Variable)

Variances Equality Test (Leven's Method)

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	5.085
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	**0.025**
alpha	0.050

Variances Equality Test:

As we can see, P-value is less than alpha, so; we should reject the null hypothesis. Meaning that variance of house prices with class 1 (TAX value for them is above the average) , is not equal to variance of house prices with class 0 (TAX value for them is below the average).
So; for comparing the average of house prices between these two classes, with should not assume the equality of variances.

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of TAX variable, we should use the appropriate method which is Leven's test.
If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (TAX Variable)

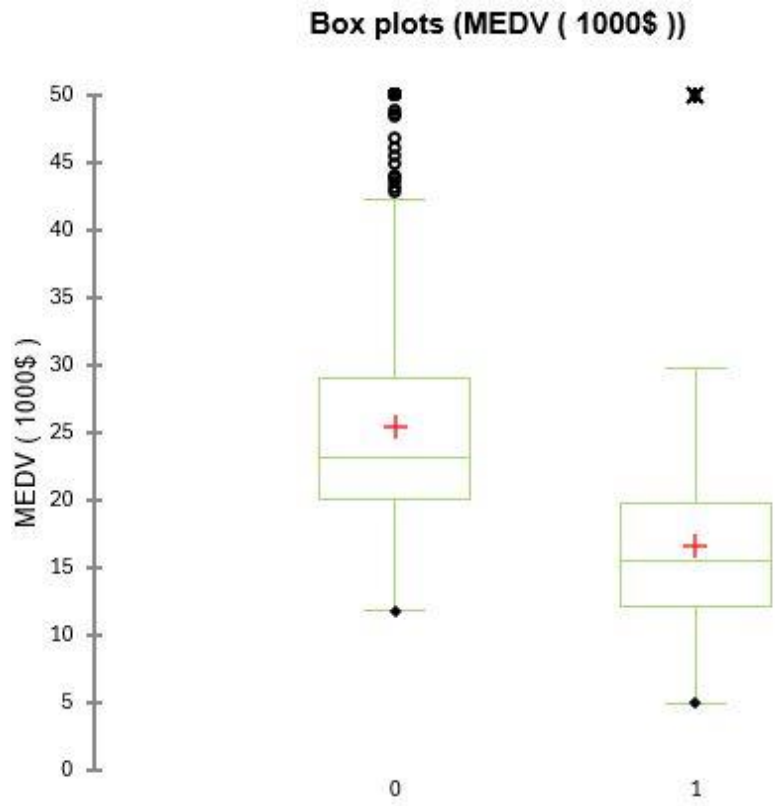
Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[7.273, 10.242]

Difference	8.758
t (Observed value)	11.602
|t| (Critical value)	1.967
DF	359.678
p-value (Two-tailed)	**<0.0001**
alpha	0.050



Higher Average Of House Prices:

This tells us that areas of class 0, have higher house prices on average, comparing to areas of class 1 Meaning that on average, houses which have lower TAX values are more expensive than the others.

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have TAX values below the average (class 0) is not equal to the average of house prices for those areas which have TAX value higher than the average (class 1).

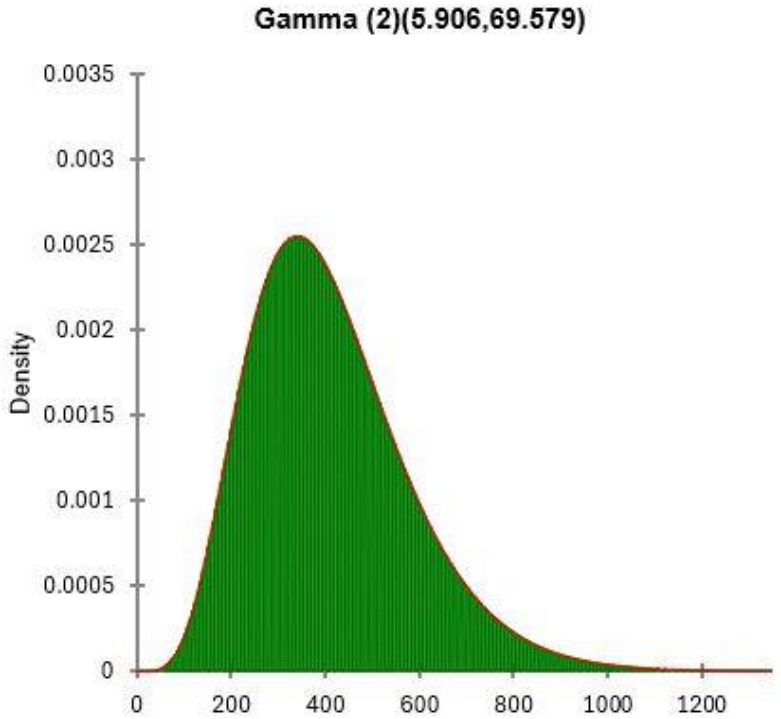
Statistical Analyses (TAX Variable)

The Best Fitting Distribution

Distribution	p-value
Negative binomial (1)	<0.0001
Negative binomial (2)	<0.0001
Erlang	<0.0001
Exponential	<0.0001
Fisher-Tippett (1)	<0.0001
Fisher-Tippett (2)	<0.0001
Gamma (1)	<0.0001
Gamma (2)	**<0.0001**
GEV	<0.0001
Gumbel	<0.0001
Log-normal	<0.0001
Logistic	<0.0001
Normal	<0.0001
Student	<0.0001
Weibull (2)	<0.0001

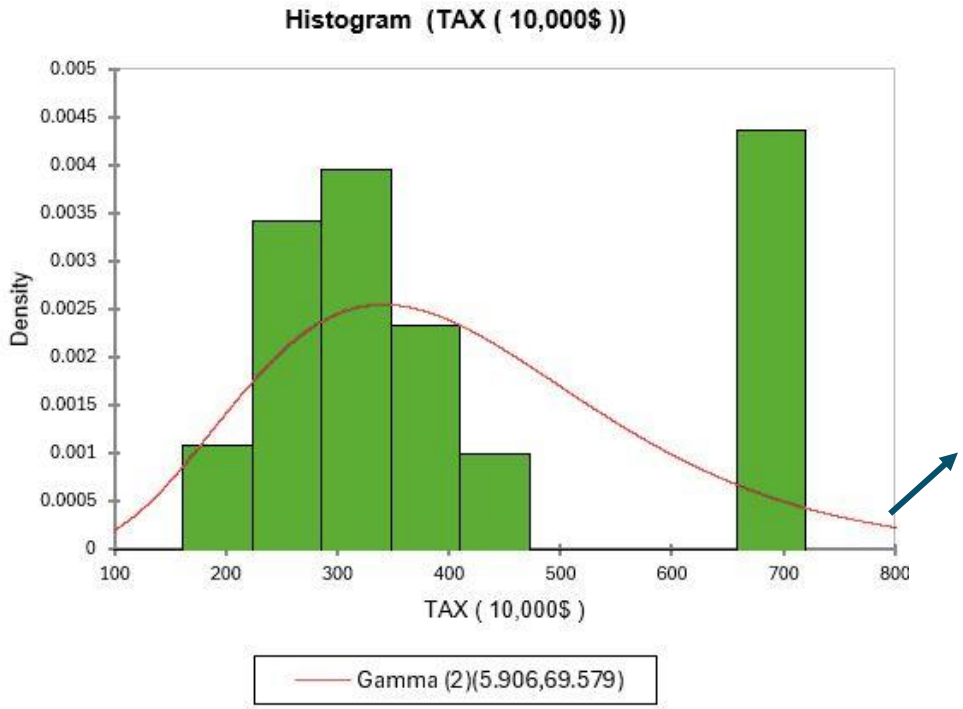
Gamma (2) Distribution :

With use of XLSTAT, I found out that the best fitting distribution for TAX variable, is Gamma (2) distribution with given parameter as below (K & β) Then again, with use of XLSTAT I plot the distribution with these parameters and its corresponding value and I got the chart which you can see on the right.



Estimated parameters (Gamma (2)):

Parameter	Value	Standard error
k	5.906	0.405
beta	69.579	4.998

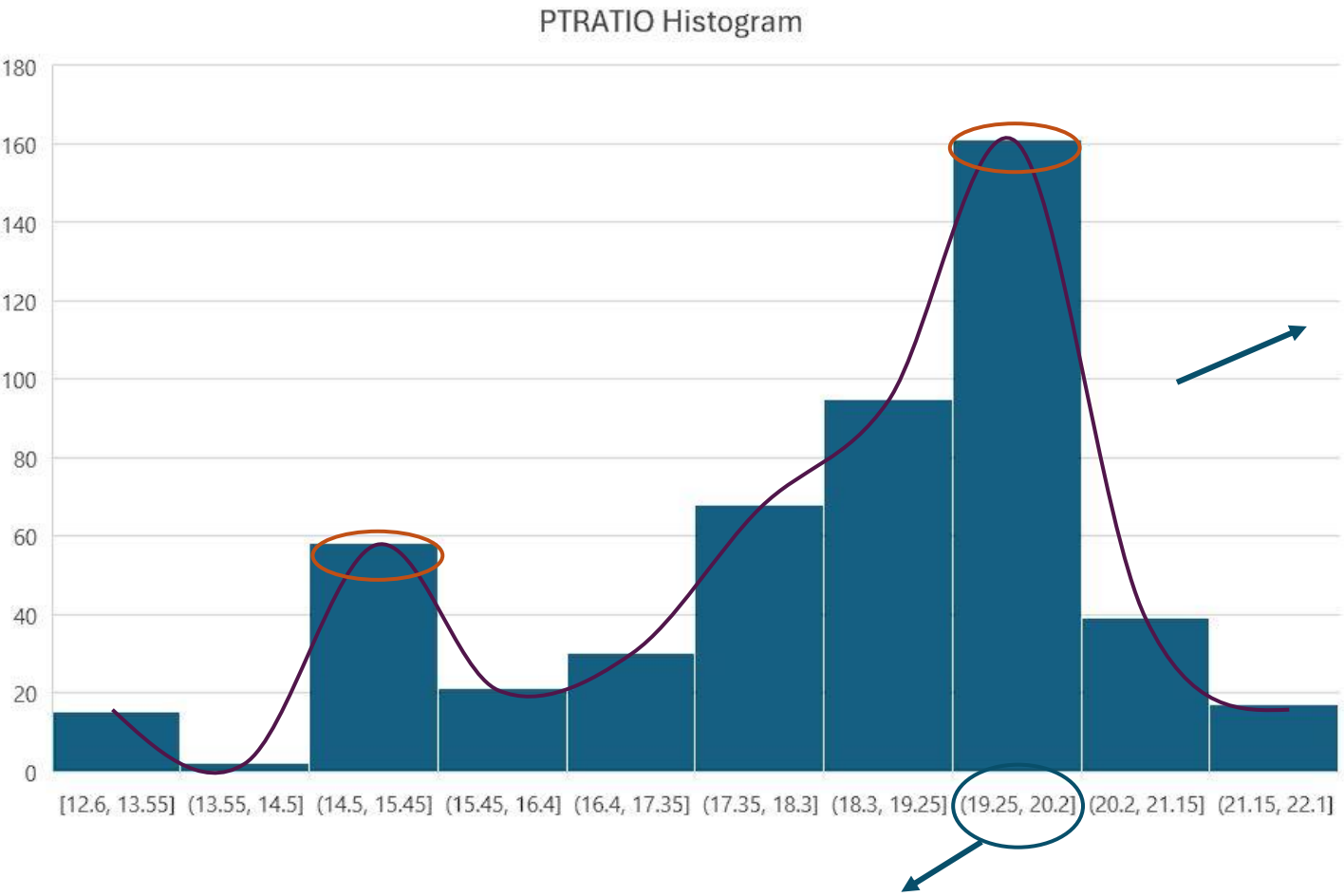


Not Perfectly Fitted :

Red line on the left chart, shows the best fitting estimated distribution for the TAX variable.
As we can see, it is not perfectly fitted, but it is the best we could do with actual distribution of the TAX variable.

Statistical Analyses (PTRATIO Variable)

Examining The Distribution



Mode:

The mode of PTRATIO feature must be something in this range.

PTRATIO Histogram Chart:

The histogram chart for the PTRATIO variable, representing the pupil-teacher ratio by town, reveals several important characteristics about its distribution.

Firstly, the distribution is negatively skewed, indicating that most towns have higher pupil-teacher ratios, with fewer towns enjoying lower ratios. This skewness suggests that while some areas have favorable student-teacher ratios, the majority face higher ratios.

Additionally, the histogram exhibits two prominent peaks. These peaks likely represent distinct groups of towns with similar pupil-teacher ratios, highlighting variations in educational environments.

Understanding the distribution of PTRATIO is crucial for analyzing how the educational resources in different neighborhoods impact both property values and residents' quality of life.

Statistical Analyses (PTRATIO Variable)

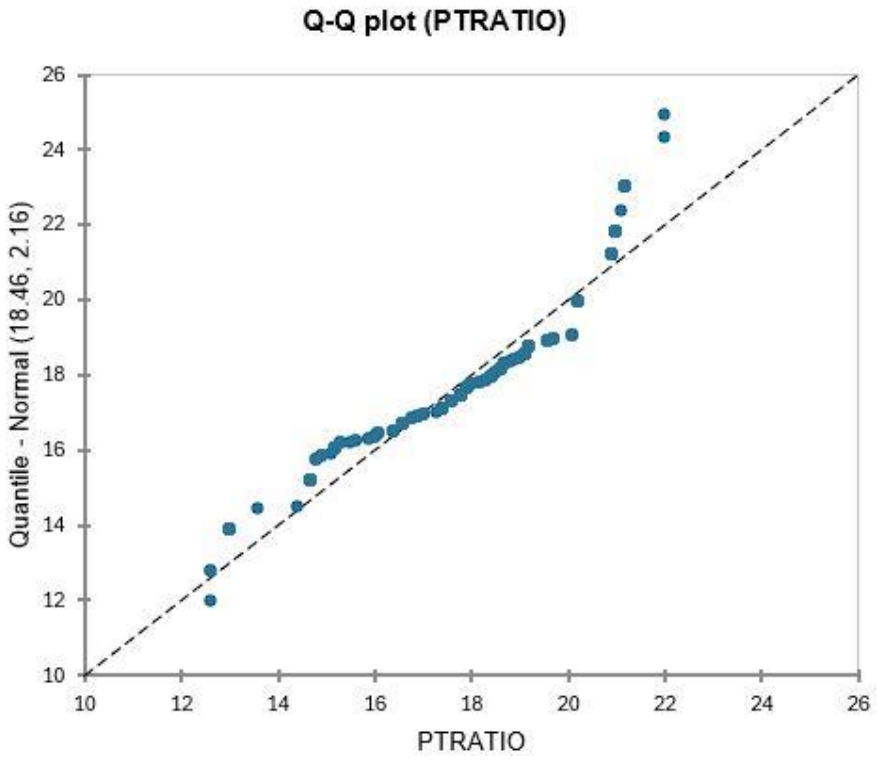
Examining The Descriptive Statistics

Statistic	PTRATIO
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	12.600
Maximum	22.000
Freq. of minimum	3
Freq. of maximum	2
Range	9.400
1st Quartile	17.400
Median	19.050
3rd Quartile	20.200
Sum	9338.500
Mean	18.456
Variance (n)	4.678
Variance (n-1)	4.687
Standard deviation (n)	2.163
Standard deviation (n-1)	2.165
Skewness (Pearson)	-0.800
Kurtosis (Pearson)	-0.294
Lower bound on mean (95%)	18.266
Upper bound on mean (95%)	18.645
Lower bound on variance (95%)	4.159
Upper bound on variance (95%)	5.323

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen 3 times among all of the records
- Maximum value of this variable can be seen 2 times among all records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A skewness of -0.8 suggests that the distribution has a slight left skew, with most of the data clustered towards the higher end but with a tendency for some lower values.
- A kurtosis value of 0.29 suggests that the distribution is not perfectly normal but has moderately fatter tails and a sharper peak. This provides insights into the data's tendency to have slightly more extreme values than a normal distribution.
- The mean of the population of this variable must be something between 18.2 and 18.6 with confidence level of 95%
- The variance of the population of this variable must be something between 4.1 and 5.3 with confidence level of 95%

Statistical Analyses (PTRATIO Variable)

Normality Test (Anderson-Darling Method)



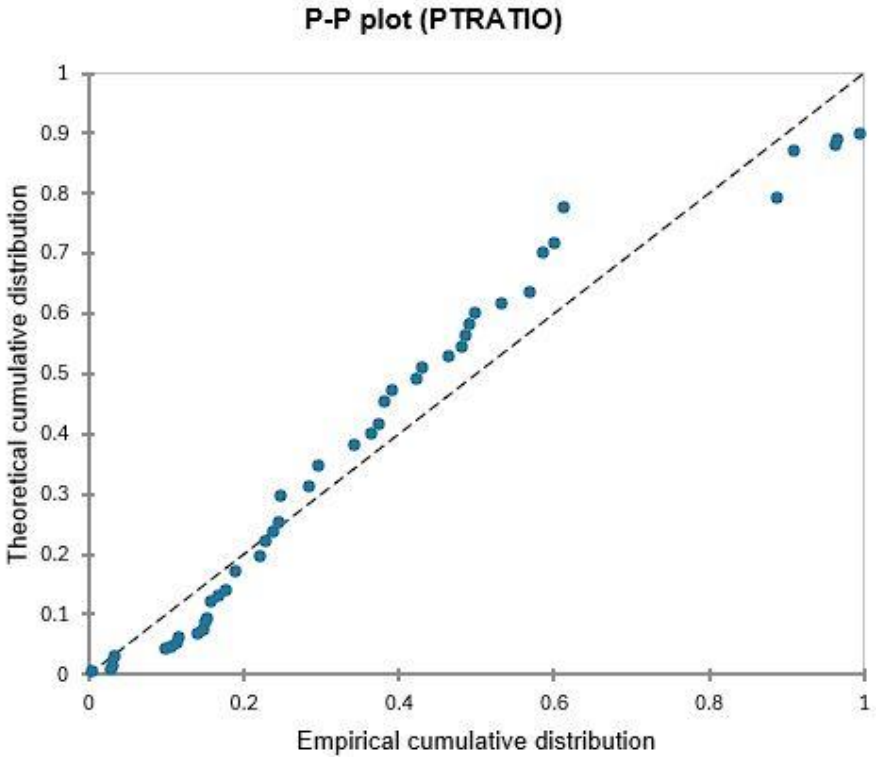
Normality Test Result :

p-value is less than alpha, so we should reject the null hypothesis. So; PTRATIO variable does not follow a normal distribution.

Anderson-Darling test (PTRATIO):	
A^2	17.411
p-value (Two-tailed)	**<0.0001**
alpha	0.050

P-P & Q-Q plot:

These plots show us that there is a difference between PTRATIO variable's distribution and a normal distribution as the normality test's result unveiled this fact to us.



Statistical Analyses (PTRATIO Variable)

Outliers Detecting (Variable Transformation)

Raw data:

This is the raw data of PTRATIO variable without any transformations.

Z-Score Normalization:

In this column, I transformed the data of PTRATIO variable with use of excel functions.
I create a function like this :

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

XLSTAT Check (Z-Score Normalization):

In this column, I transformed the data of PTRATIO variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

	A	B	C	D	E
1	PTRATIO	PTRATIO (Z transformation)	PTRATIO (Normalization)	Standardized (n-1)	0 to 1
2	15.3	-1.457557967	0.287234043	-1.457557967	0.287234043
3	17.8	-0.3027945	0.553191489	-0.3027945	0.553191489
4	17.8	-0.3027945	0.553191489	-0.3027945	0.553191489
5	18.7	0.112920349	0.64893617	0.112920349	0.64893617
6	18.7	0.112920349	0.64893617	0.112920349	0.64893617

Normalization:

In this column, I normalized the data of PTRATIO variable with use of excel functions.
I create a function like this :

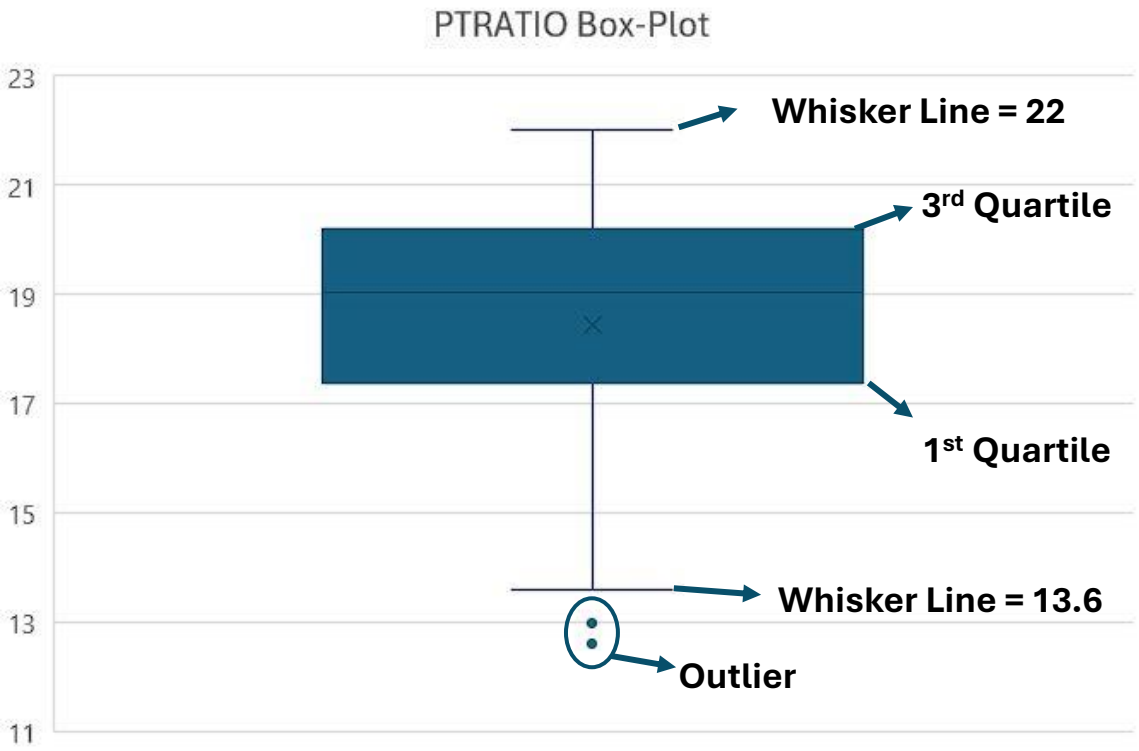
$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

XLSTAT Check (Normalization):

In this column, I normalized the data of PTRATIO variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (PTRATIO Variable)

Outliers Detecting (Box-Plot Method)



Outliers :

As we can see on box-plot chart of PTRATIO variable, this variable has outliers are at the left side of its distribution (as we could guess from the histogram chart of this variable)

Values which are less than 13.6, are detected as outliers for PTRATIO
I applied a conditional formatting rule on PTRATIO column to find the values which are less than 13.6 to find out how many outliers are detected with box-plot method and the result (as you can see below) was 15.

Meaning that there are 15 samples between our records which are detected as outliers and they belong to areas that teacher/student ratio in them is relatively low. Meaning that there are fewer numbers of students per teacher in those areas. These areas (which are detected as outliers) probably have better educational status to offer and probably are expensive and attributed to either upper-class families, or they are in some small towns which are not crowded.

Whiskers & Box :

IQR (= 3rd quartile – 1st quartile)

whisker lines :
3rd quartile + 1.5 IQR = 22
1st quartile – 1.5 IQR = 13.6

Outliers:
Values of PTRATIO which are above 22
Values of PTRATIO which are below 13.6

Conclusion :

With box-plot method we detected 15 outliers for PTRATIO variable, on the next slide we are going to detected the outliers of this variable with z-score method.
We guess that the number of outliers with z-score method is less than the number of outliers with box-plot method.
Anyway, 15 outliers out of 506 records makes approximately 3% of our sample and it is relatively low that we can rely on.

Statistical Analyses (PTRATIO Variable)

Outliers Detecting (Z-Score Method)

	A	B
1	PTRATIO	PTRATIO (Z transformation)
2	15.3	-1.457557967
3	17.8	-0.3027945
4	17.8	-0.3027945
5	18.7	0.112920349
6	18.7	0.112920349

No Outliers With Z-Score Method :

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.
So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I found out that there is no value between transformed data to be greater than 3 or less than -3

Box-Plot VS Z-Score :

With box-plot method we detected 15 outliers while with z-score method we found non.
These outliers which we found with box-plot method may help us to convert PTRATIO variable to a normally distributed one.
Finding no outliers with z-score method could be guessable, because this method is more sensitive and more selective than box-plot method.
Anyway, we only can rely on box-plot method because that's all we have for detecting outliers of PTRATIO variable.

Statistical Analyses (PTRATIO Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, PTRATIO variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.

So; I transformed this variable with box-cox method, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.

On the second step, I removed the outliers of this variable and again applied a normality test to see if it now follows a normal distribution and the answer to this question was also negative.

So, as the conclusion, we find it out that we cannot convert the PTRATIO variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

PTRATIO Box-Cox transformation
32745.31273
63253.42694
63253.42694
78392.21901
78392.21901
78392.21901

Transformed data of PTRATIO Variable With Box-Cox Method

Anderson-Darling test (PTRATIO Box-Cox transformation):	
A^2	11.683
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test After Box-Cox Transformation :

As we can see, the result of the normality test of transformed data (with box-cox method), PTRATIO variable still does not follow a normal distribution.

Anderson-Darling test (PTRATIO Box-Cox transformation):	
A^2	11.925
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test After Removing Outliers :

As we can see, even after removing the outliers of the PTRATIO variable and conducting a normality test again, this variable is not following a normal distribution.

Statistical Analyses (PTRATIO Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:

I am going to check the correlation between PTRATIO variable and target variable which is MEDV, these are both continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	PTRATIO	MEDV (1000$)
PTRATIO	**1**	**-0.508**
MEDV (1000$)	**-0.508**	**1**

Relatively Strong And Inverse Correlation:

The correlation matrix and the value of -0.50 tells us that there is an inverse correlation between these 2 variables.
Meaning that if one of the increase, the other one will decrease.
On the other hand, the absolute value would be 0.50, which indicates that the correlation is relatively strong.
meaning that, when PTRATIO gets higher (the number of students per teacher gets higher) , as a result, educational quality decreases, and the MEDV variable will also decrease

p-values (Pearson):

Variables	PTRATIO	MEDV (1000$)
PTRATIO	**0**	**<0.0001**
MEDV (1000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:

The value is <0.0001 suggests that the correlation between PTRATIO and MEDV is statistically significant and it is not due to random changes.

Coefficients of determination (Pearson):

Variables	PTRATIO	MEDV (1000$)
PTRATIO	**1**	0.258
MEDV (1000$)	0.258	**1**

Power Of Prediction:

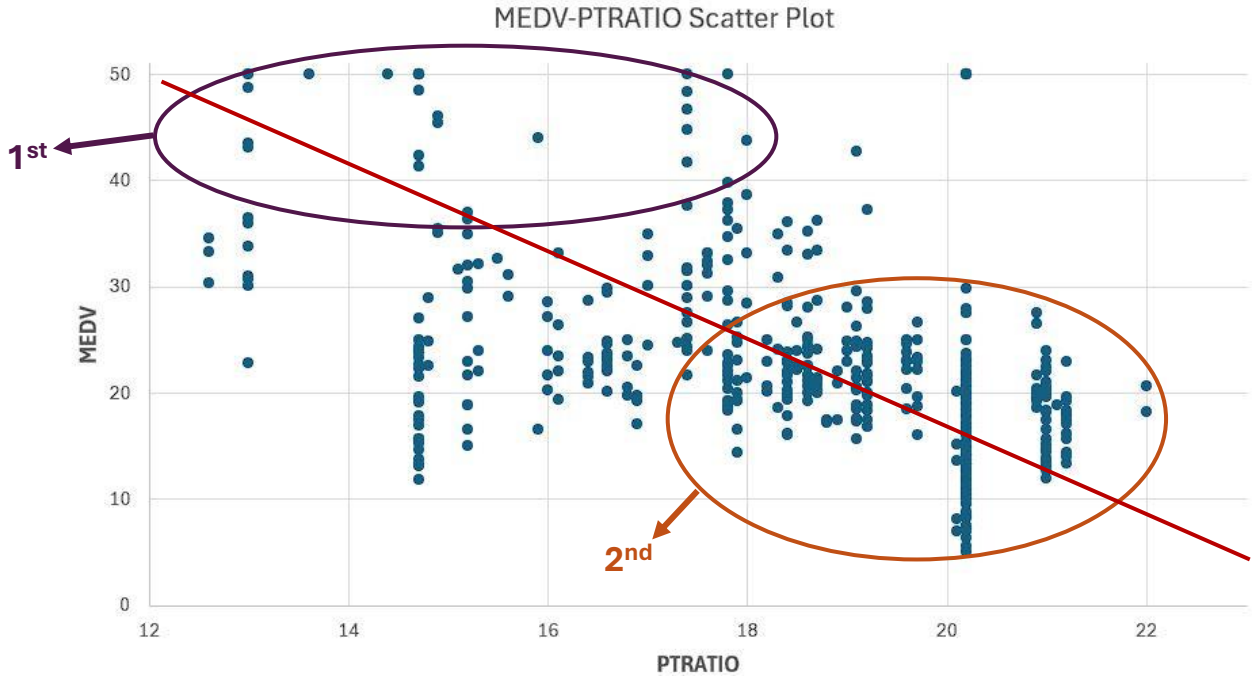
The value of 0.258 in this table, indicates that only 25.8% of the variance in target variable (MEDV) can be explained by the variance in PTRATIO variable.

Statistical Analyses (PTRATIO Variable)

Scatter Plot With The Target Variable

Insights :

As we see on the chart, we can draw some insights from the scatter plot of the MEDV and PTRATIO variables.
Some of the insights we can discuss about are motioned below
Clearly, we can extract other hints from the plot, but by now, these would satisfy our purposes here



1st Insight:

This zone contains the most expensive houses.
We can see that the PTRATIO variable is relatively low for these houses.
Suggesting that these two variables have an inverse correlation (as it shown with red line)
On the other hand, the concentration of records in this zone shows us lower variety in comparison to the other zone that we have marked as 2nd .
Suggesting that there are fewer records in our dataset that PTRATIO in them is relatively low.
This zone probably contains upper-class families and the LSTAT rate (the other feature in our dataset which shows the proportion of low-status families) may be lower in this zone than the other one.

2nd Insight:

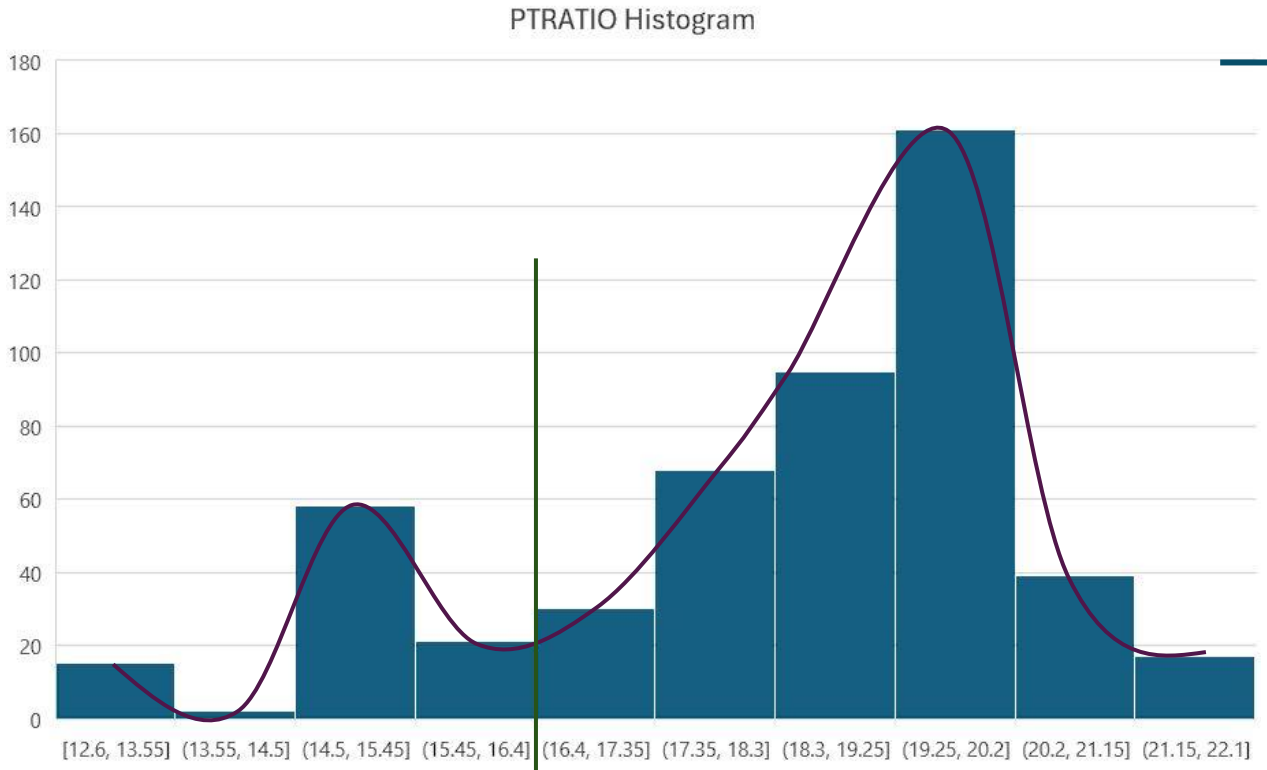
This zone is made of areas that PTRATIO in them is relatively higher than the other areas.
The dense concentration of records in this zone suggests that most of areas of our dataset share the same characteristics in terms of PTRATIO variable.
On the other hand, house prices in this zone is relatively lower than the other areas which one more time shows an inverse correlation between these two variables.

Trend Line:

The red line in the chart, shows the relationship between MEDV and PTRATIO variables.
As we can see, the correlation is inverse.
Meaning that as one of these variables increases, the other one will decrease.
This fact seems logical, because rich families usually looking for better educations and areas that have fewer number of students per teacher have probably better educational system.

Statistical Analyses (PTRATIO Variable)

Question : Is There Any Difference Between Average Of House Prices Based On PTRATIO values?



Average Of NOX Feature :

I am going to create a new feature based on PTRATIO.

this feature is going to be 0 for areas which have PTRATIOs less than 16.4

And is going to be 1 for areas which have PTRATIOs greater than 16.4

And I am going to compare the average of house prices between these 2 classes.

The reason of choosing the value of 16.4, is the histogram chart of the PTRATIO variable.
If we look at this variable's histogram, it seems that PTRATIO variable is made of two different distributions
We can talk about this issue later but by now, it seems interesting if we could divide the PTRATIO into these two classes.

Dividing Wall:
This green line, separates two distributions which seem to be existed in the histogram chart of PTRATIO variable.

	A	B	C
1	PTRATIO	PTRATIO Binary Classification	MEDV (1000$)
2	15.3	=IF(A2>14.6,1,0)	24
3	17.8	1	21.6
4	17.8	1	34.7

Statistical Analyses (PTRATIO Variable)

Variances Equality Test (Leven's Method)

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	1.062
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	0.303
alpha	0.050

Variances Equality Test:

As we can see, P-value is greater than alpha, so; we should accept the null hypothesis. Meaning that variance of house prices with class 1 (those with PTRATIO greater than 16.4) , is equal to variance of house prices with class 0 (those with PTRATIO less than 16.4).

So; for comparing the average of house prices between these two classes, with should assume the equality of variances.

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of PTRATIO variable, we should use the appropriate method which is Leven's test.

If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (PTRATIO Variable)

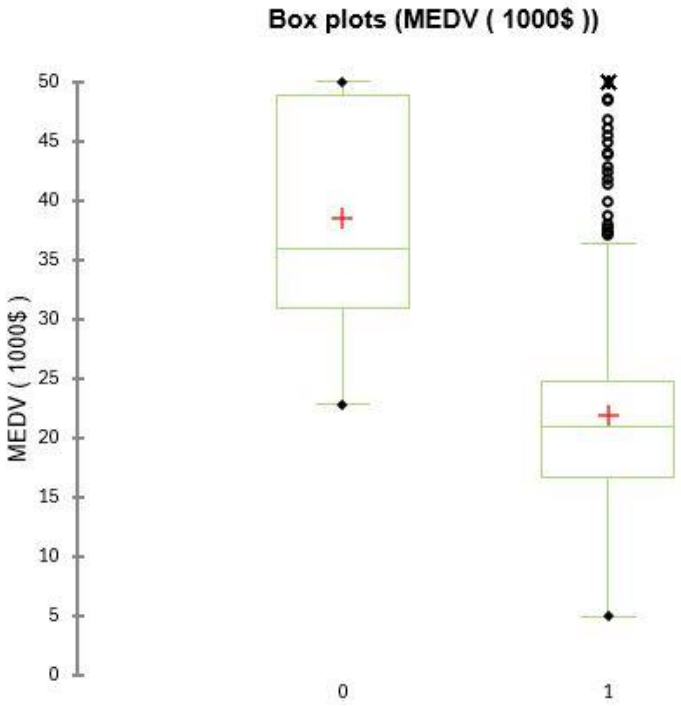
Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[12.301, 20.744]

Difference	16.522
t (Observed value)	7.689
|t| (Critical value)	1.965
DF	504
p-value (Two-tailed)	**<0.0001**
alpha	0.050



Higher Average Of House Prices:

This tells us that areas of class 0, have higher house prices on average, comparing to areas of class 1 Meaning that on average, houses which are in areas with better educational quality (there are fewer number of students per teacher) are more expensive than the others.

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have PTRATIO values below 16.4 (class 0) is not equal to the average of house prices for those areas which have PTRATIO value higher than 16.4 (class 1).

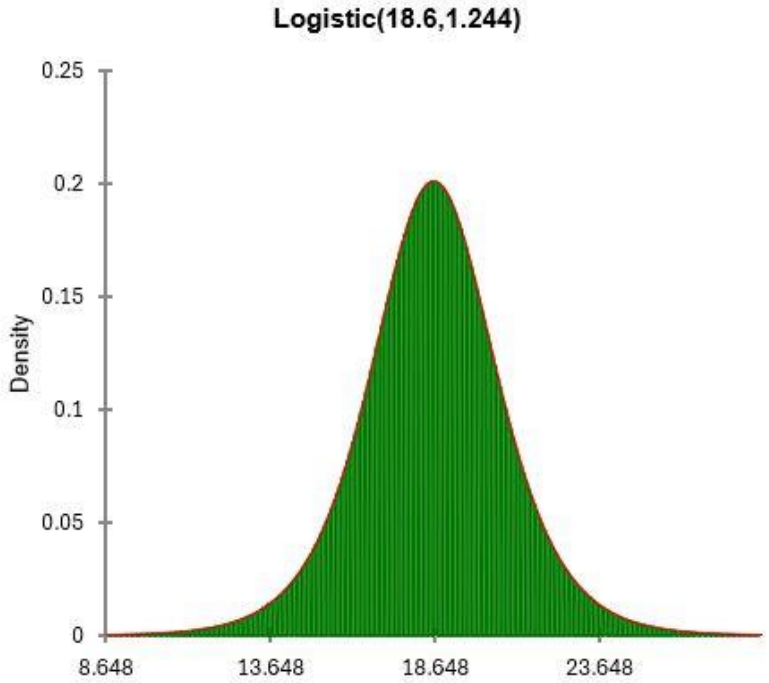
Statistical Analyses (PTRATIO Variable)

The Best Fitting Distribution

Distribution	p-value
Beta4	<0.0001
Chi-square	<0.0001
Erlang	<0.0001
Fisher-Tippett (1)	<0.0001
Fisher-Tippett (2)	<0.0001
Gamma (2)	<0.0001
GEV	<0.0001
Gumbel	<0.0001
Log-normal	<0.0001
Logistic	**<0.0001**
Normal	<0.0001
Student	<0.0001
Weibull (2)	<0.0001

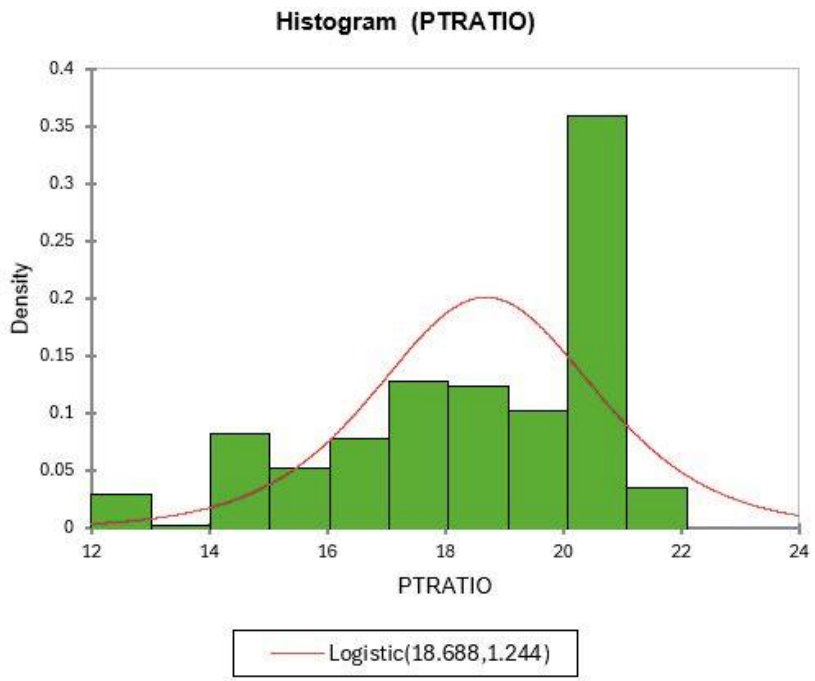
Logistic Distribution :

With use of XLSTAT, I found out that the best fitting distribution for PTRATIO variable, is logistic distribution with given parameter as below (μ & σ) Then again, with use of XLSTAT I plot the distribution with these parameters and its corresponding value and I got the chart which you can see on the right.



Estimated parameters (Logistic):

Parameter	Value	Standard error
μ	18.688	0.074
s	1.244	0.020



Not Perfectly Fitted :

Red line on the left chart, shows the best fitting estimated distribution for the PTRATIO variable. As we can see, it is not perfectly fitted, but it is the best we could do with actual distribution of the PTRATIO variable.

Statistical Analyses (DIS Variable)

Examining The Distribution

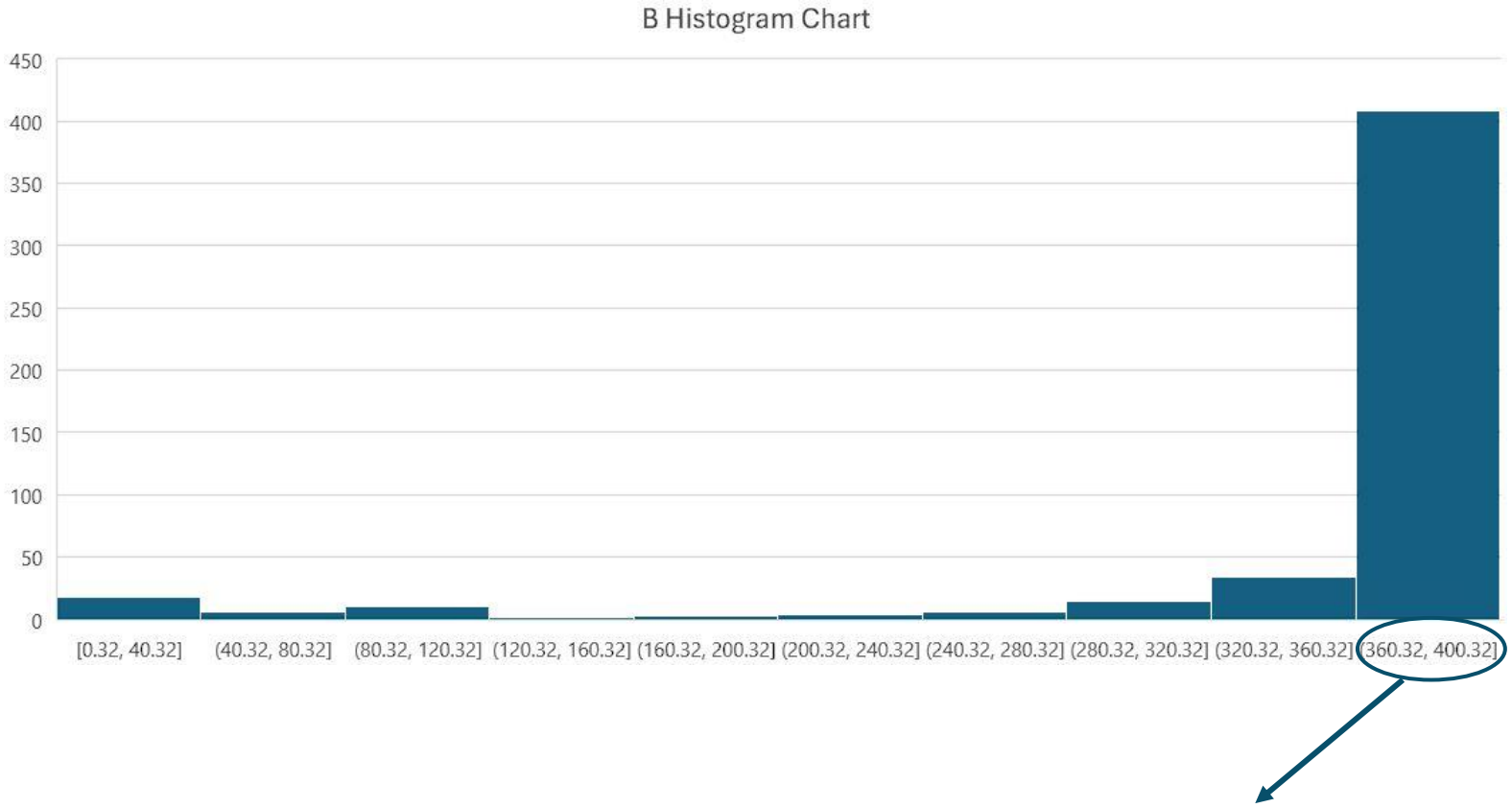
DIS Histogram Chart:

The histogram chart for the B variable, representing the proportion of people of African descent in each town, offers insightful observations about its distribution.

The distribution is highly negatively skewed, suggesting that the majority of towns have higher proportions of this demographic. This significant skewness reveals that while a large number of towns exhibit high values, fewer towns have lower values, creating a pronounced left tail in the histogram. Such a pattern underscores the demographic concentrations within the dataset, providing a clearer picture of community compositions.

Understanding the histogram and distribution of the B variable is crucial for analyzing the socio-economic dynamics across different neighborhoods. The skewness points to potential areas of focus for policy-making, social services, and urban planning.

By acknowledging these demographic patterns, stakeholders can better address community needs and foster inclusive growth. This analysis enriches our comprehension of the dataset and aids in deriving meaningful insights for further research and application.



Mode:

The mode of B feature must be something in this range.

Statistical Analyses (B Variable)

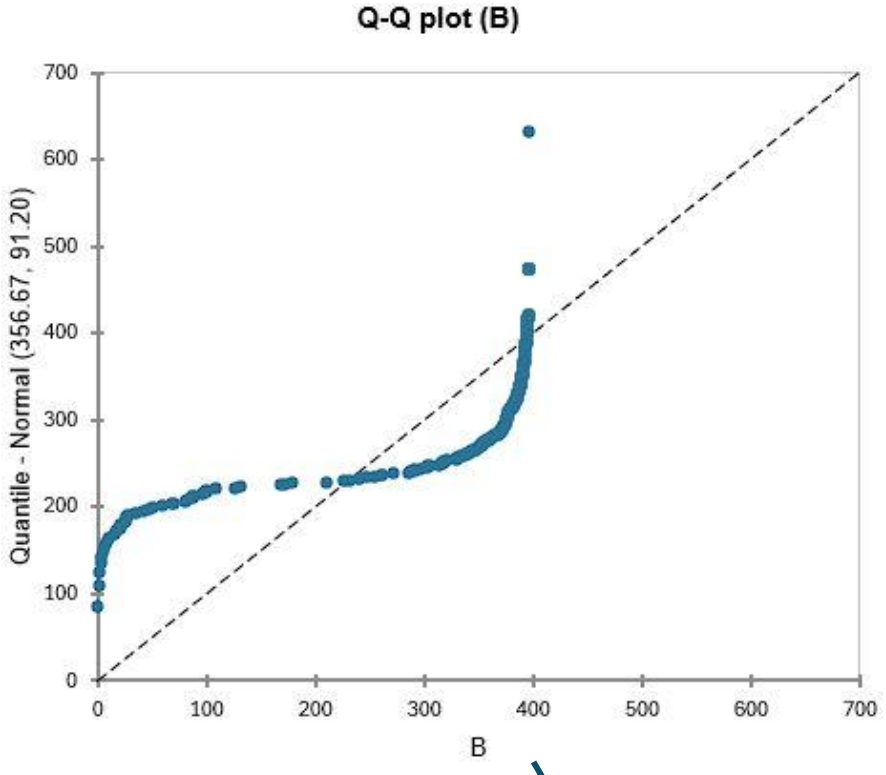
Examining The Descriptive Statistics

Statistic	B
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	0.320
Maximum	396.900
Freq. of minimum	1
Freq. of maximum	121
Range	396.580
1st Quartile	375.378
Median	391.440
3rd Quartile	396.225
Sum	180477.060
Mean	356.674
Variance (n)	8318.280
Variance (n-1)	8334.752
Standard deviation (n)	91.205
Standard deviation (n-1)	91.295
Skewness (Pearson)	-2.882
Kurtosis (Pearson)	7.144
Lower bound on mean (95%)	348.700
Upper bound on mean (95%)	364.648
Lower bound on variance (95%)	7395.186
Upper bound on variance (95%)	9466.479

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen only 1 time among all of the records
- Maximum value of this variable can be seen 121 times among all records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A skewness of -2.88 suggests a substantial asymmetry with a heavy left tail, indicating that lower values are more extreme and less common, while higher values are more frequent. This highlights the presence of outliers and gives insight into the overall shape of the data.
- A kurtosis of 7.14 suggests that the distribution has a very sharp peak and heavy tails, indicating a high probability of extreme values and outliers.
- The mean of the population of this variable must be something between 348.7 and 364.6 with confidence level of 95%
- The variance of the population of this variable must be something between 7395.1 and 9466.4 with confidence level of 95%

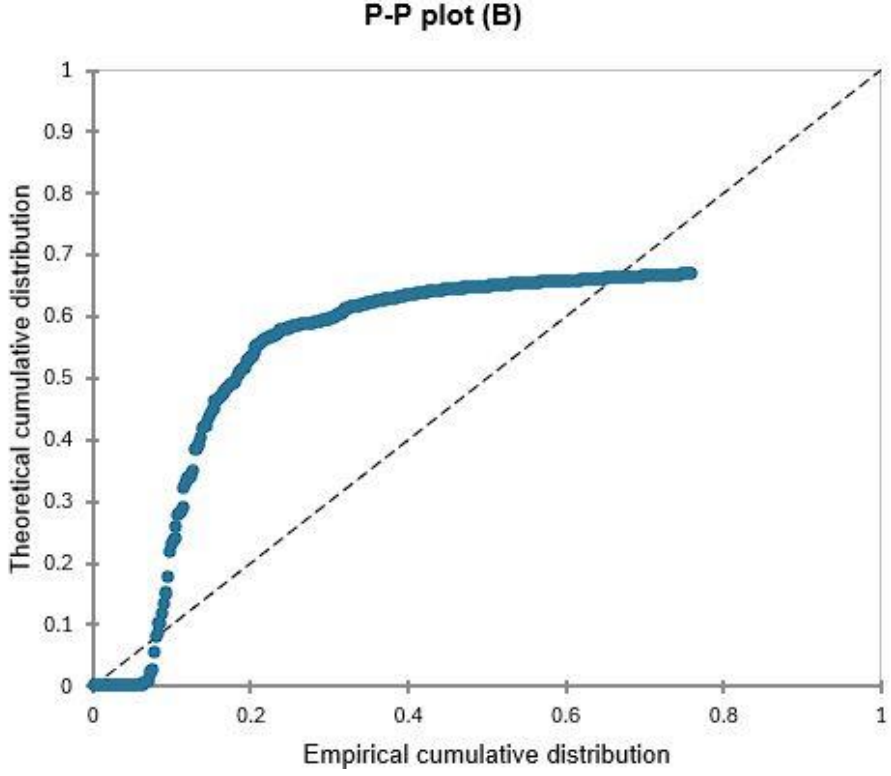
Statistical Analyses (B Variable)

Normality Test (Anderson-Darling Method)



Anderson-Darling test (B):	
A^2	109.288
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test Result :
p-value is less than alpha, so we should reject the null hypothesis. So; B variable does not follow a normal distribution.



P-P & Q-Q plot:
These plots show us that there is a difference between B variable's distribution and a normal distribution as the normality test's result unveiled this fact to us.

Statistical Analyses (B Variable)

Outliers Detecting (Variable Transformation)

Raw data:
This is the raw data of B variable without any transformations.

Z-Score Normalization:
In this column, I transformed the data of B variable with use of excel functions.
I create a function like this :

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

XLSTAT Check (Z-Score Normalization):
In this column, I transformed the data of B variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

	A	B	C	D	E
1	B	B (Z-transformation)	B (Normalization)	Standardized (n-1)	0 to 1
2	396.9	0.440615895	1	0.440615895	1
3	396.9	0.440615895	1	0.440615895	1
4	392.83	0.396035074	0.989737254	0.396035074	0.989737
5	394.63	0.415751408	0.99427606	0.415751408	0.994276
6	396.9	0.440615895	1	0.440615895	1

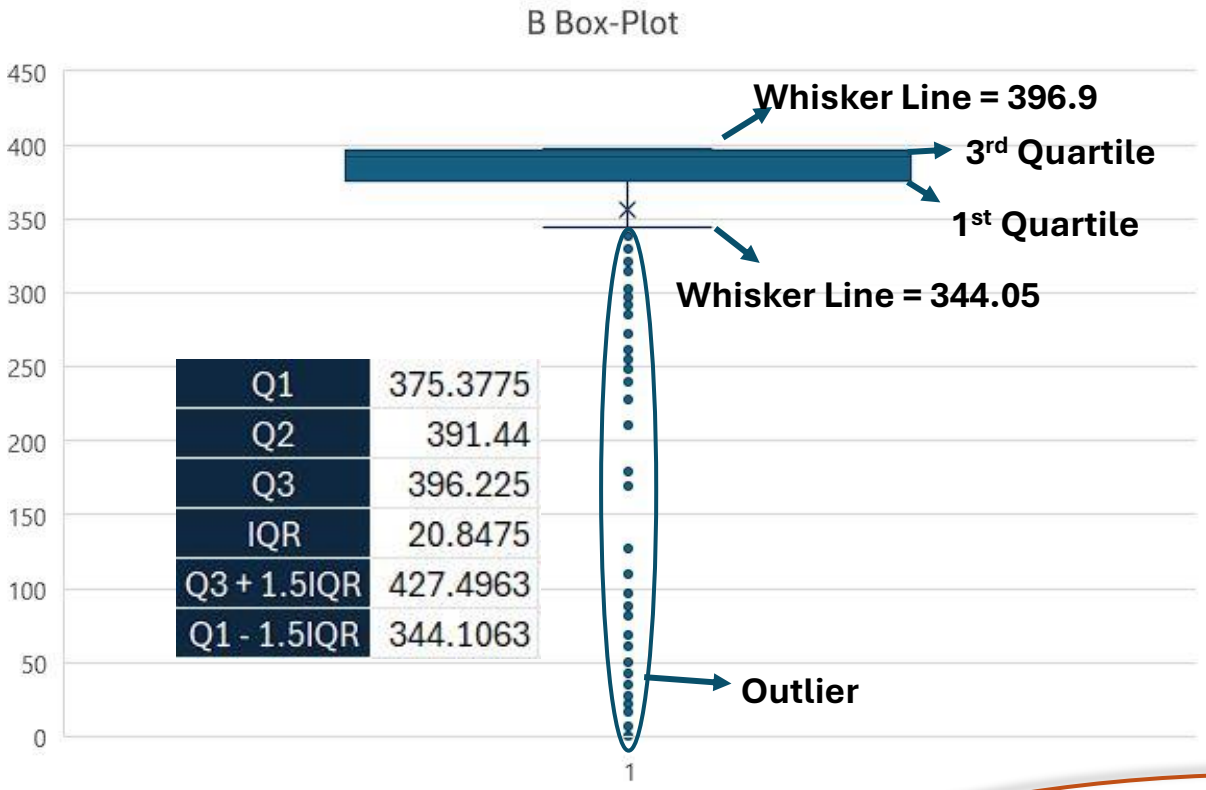
Normalization:
In this column, I normalized the data of B variable with use of excel functions.
I create a function like this :

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

XLSTAT Check (Normalization):
In this column, I normalized the data of B variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (B Variable)

Outliers Detecting (Box-Plot Method)



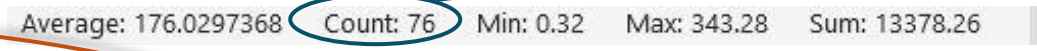
Outliers :

As we can see on box-plot chart of B variable, this variable has outliers are at the left side of its distribution (as we could guess from the histogram chart of this variable)

Values which are less than 344.05, are detected as outliers for B variable.
I applied a conditional formatting rule on B column to find the values which are less than 344.05 to find out how many outliers are detected with box-plot method and the result (as you can see below) was 76.

Meaning that there are 76 samples between our records which are detected as outliers and they belong to areas that proportion of African-American residentials is relatively less than the others areas.

These areas can be considered as old-fashion areas that racist beliefs are still popular in these areas (note that we said “can be considered”, we did not claim this as a fact. This low proportion of African-American people in these areas may have some other reasons.)



Whiskers & Box :

IQR (= 3rd quartile – 1st quartile)

whisker lines :
3rd quartile + 1.5 IQR = 396.9 = maximum
1st quartile – 1.5 IQR = 344.05

Outliers:
Values of B which are below 344.05

Conclusion :

With box-plot method we detected 76 outliers for B variable, on the next slide we are going to detected the outliers of this variable with z-score method.
We guess that the number of outliers with z-score method is less than the number of outliers with box-plot method.
Anyway, 76 outliers out of 506 records makes approximately 15% of our sample, which is so high.
We cannot rely on box-plot method for deciding which records are outliers, because they are too man, it is better to go with z-score method.

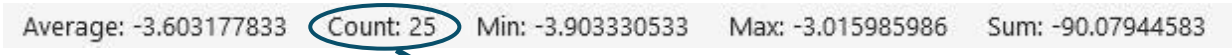
Statistical Analyses (B Variable)

Outliers Detecting (Z-Score Method)

	A	B
1	B	B (Z-transformation)
104	70.8	-3.131326538
412	2.6	-3.87835651
413	35.05	-3.52291483
414	28.79	-3.591483857

Outliers With Z-Score Method:

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.
So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I got the result as you can see in the table.



25 Outliers:

25 outliers are detected based on Z-Score method.
While, the number of outliers which were detected based on box-plot method was 76.

Box-Plot VS Z-Score :

When we compare the results of these two methods for detecting outliers for B feature, there is a big difference between these two methods.
With Box-Plot method we got 76 outliers
With Z-Score method we got 25 outliers
The number of outliers detected with box-plot method is much more than the number of outliers detected with z-score method.
It would be better if we rely on z-score method because the number of outliers detected by this method is less and any records which is know as outlier with z-score method is also know as outlier with box-plot method.

Statistical Analyses (B Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, B variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it.
So; I transformed this variable with box-cox method, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.
On the second step, I removed the outliers of this variable and again applied a normality test to see if it now follows a normal distribution and the answer to this question was also negative.
So, as the conclusion, we find it out that we cannot convert the B variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

Box-Cox transformation
905317226.9
905317226.9
871767208.3
886491488.4
905317226.9

Transformed data of B Variable With Box-Cox Method

Anderson-Darling test (Box-Cox transformation):	
A^2	71.125
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test After Box-Cox Transformation :

As we can see, the result of the normality test of transformed data (with box-cox method), B variable still does not follow a normal distribution.

Anderson-Darling test (Box-Cox transformation):	
A^2	62.306
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test After Removing Outliers :

As we can see, even after removing the outliers of the B variable and conducting a normality test again, this variable is not following a normal distribution.

Statistical Analyses (B Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:

I am going to check the correlation between B variable and target variable which is MEDV.
Both are continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	B	MEDV (1000$)
B	**1**	**0.333**
MEDV (1000$)	**0.333**	**1**

Relatively Weak And Direct Correlation:

The correlation matrix and the value of 0.33 tells us that there is a direct correlation between these 2 variables.
Meaning that if one of the increase, the other one will also increase.
On the other hand, the absolute value would be 0.33, which indicates that the correlation is relatively weak.

Coefficients of determination (Pearson):

Variables	B	MEDV (1000$)
B	**1**	0.111
MEDV (1000$)	0.111	**1**

p-values (Pearson):

Variables	B	MEDV (1000$)
B	**0**	**<0.0001**
MEDV (1000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:

The value is <0.0001 suggests that the correlation between B and MEDV is statistically significant and it is not due to random changes.

Power Of Prediction:

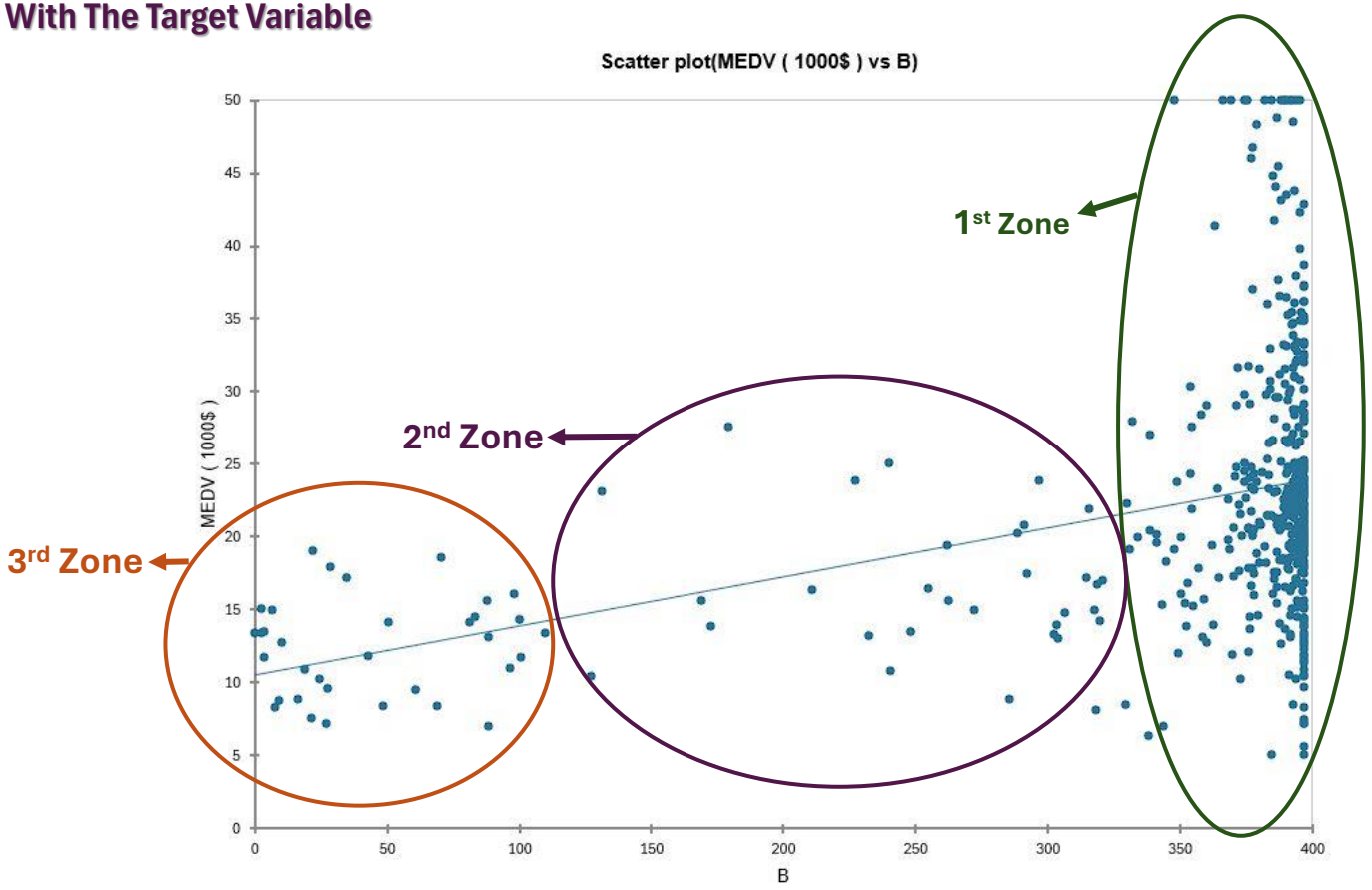
The value of 0.111 in this table, indicates that only 11.1% of the variance in target variable (MEDV) can be explained by the variance in B variable.

Statistical Analyses (B Variable)

3 Zones:

As we see on the chart, we can divide the city into 3 zones based on median value of houses in each area and B value of each area.
This gives us an interesting insight as we can interpreter as following :

Scatter Plot With The Target Variable



1st Zone:

This zone hast the most concentrated distribution of the records (as we saw before on the histogram chart of B variable that the most of the records were clustered on the right side of the histogram chart.)
Meaning that there are many areas in our dataset which have high values of B variable. In other words, there are many areas which have high proportion of the African-American residential.
This fact may suggest that the distribution of African-American does not follow a normal distribution.
This fac can imply that there may be some areas that racism is still popular in them.
Note that we did not claim that it is the cause, we just say that it can be possible.
On the other hand, another interesting fact is that the MEDV variable in this zone can range from the lowest to the highest of itself, while this does not happen for the other two zones.

2nd Zone:

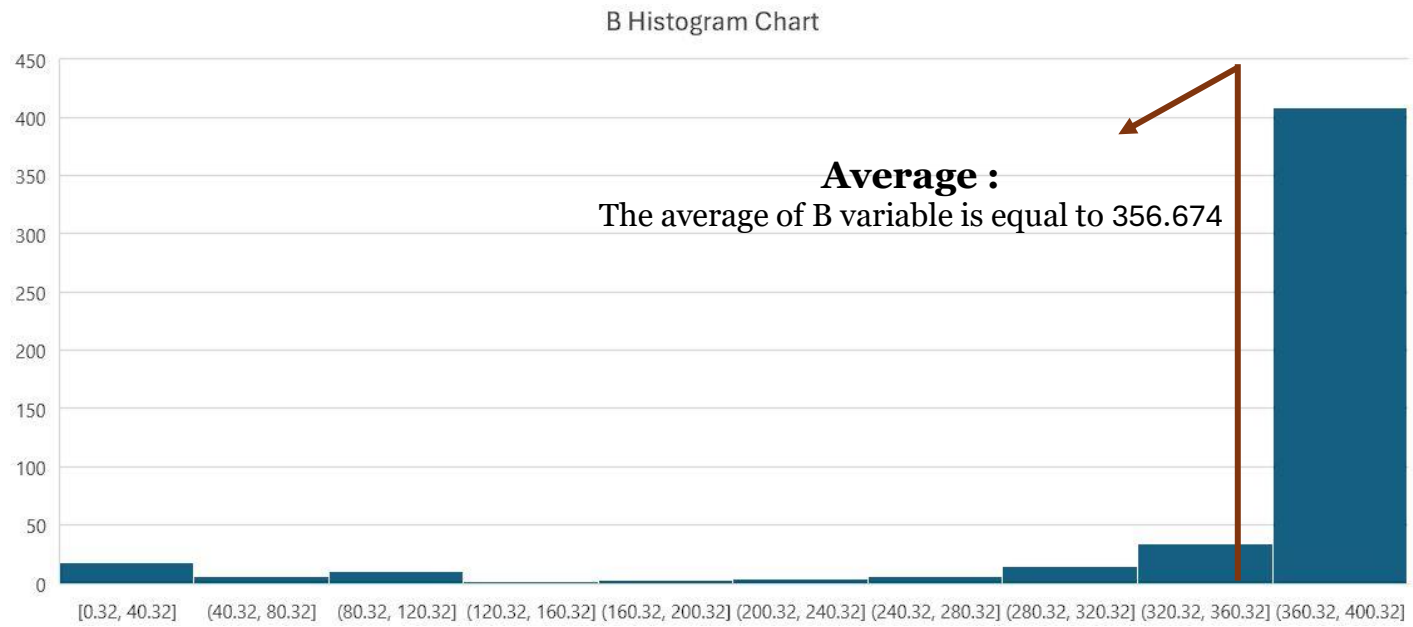
This zone includes areas which have lower proportion of African-American residentials, and have also lower values of MEDVs.
Areas in this zone, has a minimum in terms of MEDVs and has a maximum which are not absolute minimum and maximum of our dataset.
Meaning that areas in this range do not have the most expensive or the cheapest houses of Boston in them. Probably, this zone is mostly made of middle-class families.
On the other hand, the concentration of records which are included in this zone is lower than the other zones.
This can suggest that there are fewer areas in Boston that share the same characteristics as areas in this zone.

3rd Zone:

Areas which are included in this zone have relatively lower proportion of African-American residentials and also the have lower values of MEDV variable in comparison to the other two zones.
If we consider that there are some areas that racism in them is still popular, and this is the reason of B variable is not following a normal distribution, these areas are more probable to be in this zone.
Also, we can attribute these areas to working-class families due to house prices in this zone, which are lower than the other two zones.

Statistical Analyses (B Variable)

Question : Is There Any Difference Between Average Of House Prices Based On B Variable?



Average :
The average of B variable is equal to 356.674

"B" 2 Classes :
I am going to create a new feature based on B feature.
This feature is going to be 0 for areas which have B values lower than the average.
And is going to be 1 for areas which have B values greater than the average.
I chose these 2 classes because it seems to me interesting and logical to compare these two classes in terms of the average of their MEDVs.

	A	B	C
1	B	B Binary Classifcation	MEDV (1000$)
2	396.9	1	24
3	396.9	1	21.6
4	392.83	1	34.7
5	394.63	1	33.4
6	396.9	1	36.2

Statistical Analyses (B Variable)

Variances Equality Test (Leven's Method)

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	8.196
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	**0.004**
alpha	0.050

Variances Equality Test:

As we can see, P-value is less than alpha, so; we should reject the null hypothesis. Meaning that variance of house prices with class 1 (those with B values greater than the average) , is not equal to variance of house prices with class 0 (those with B values less than the average).

So; for comparing the average of house prices between these two classes, with should not assume the equality of variances.

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of "B" variable, we should use the appropriate method which is Leven's test.

If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (B Variable)

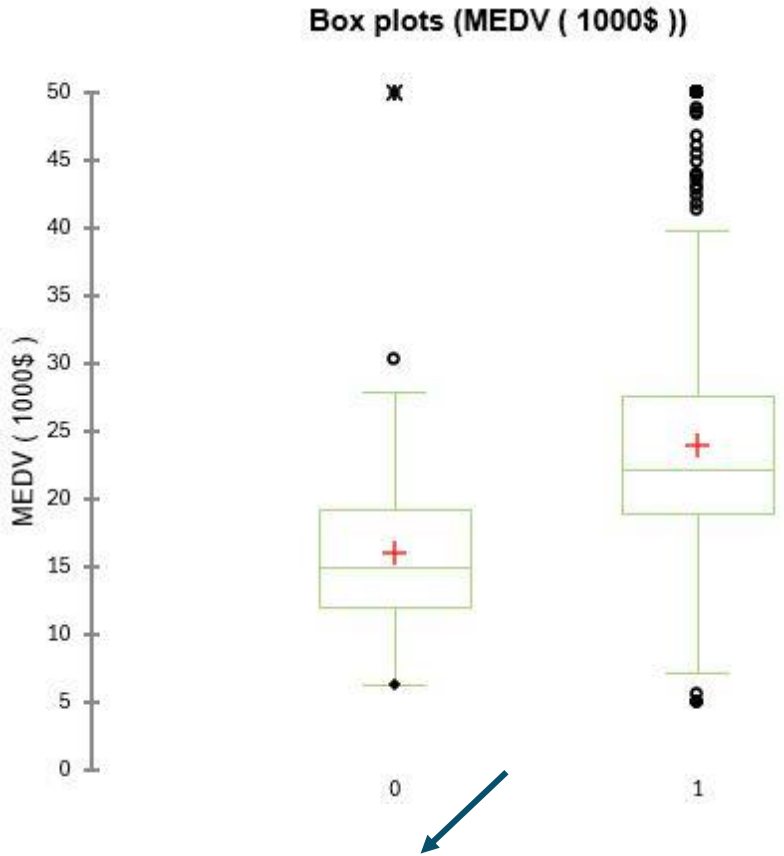
Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[-9.551, -6.352]

Difference	-7.951
t (Observed value)	-9.811
|t| (Critical value)	1.973
DF	180.225
p-value (Two-tailed)	**<0.0001**
alpha	0.050



Higher Average Of House Prices:

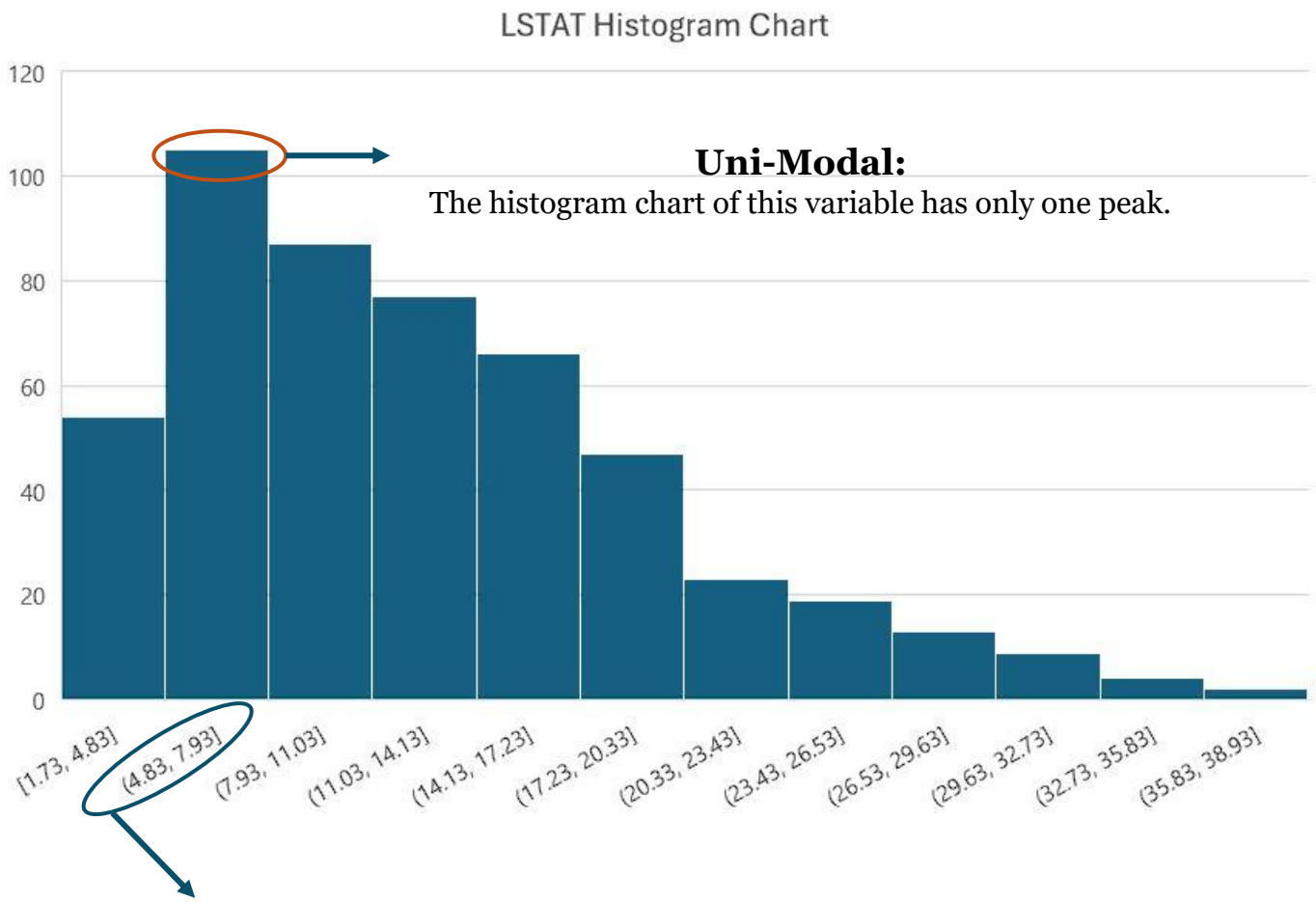
This tells us that areas of class 1, have higher house prices on average, comparing to areas of class 0 Meaning that on average, houses which are in areas with more African-American residentials are more expensive than the others.

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have B values less than the average (class 0) is not equal to the average of house prices for those areas which have B values greater than the average (class 1).

Statistical Analyses (LSTAT Variable)

Examining The Distribution



Uni-Modal:

The histogram chart of this variable has only one peak.

Mode:

The mode of LSTAT feature must be something in this range.

LSTAT Histogram Chart:

The histogram chart for the LSTAT variable, representing the percentage of the lower status population, reveals several critical characteristics.

The distribution is uni-modal, indicating that there is a single, prominent peak where the majority of the values are concentrated. This suggests that there is a common level of lower status across many neighborhoods, with one predominant cluster of values.

Additionally, the histogram shows a positively skewed distribution, where most of the values are concentrated towards the lower end, with fewer values stretching towards the higher end. This positive skewness signifies that while many neighborhoods have a relatively lower percentage of lower-status individuals, there are some areas with significantly higher percentages that extend the distribution to the right.

Analyzing the LSTAT variable provides valuable insights into the socio-economic conditions within the dataset. The uni-modal nature of the distribution suggests a commonality in lower status across neighborhoods, while the positive skewness highlights the presence of certain areas with notably higher lower-status populations. This information can be crucial for urban planning, policy-making, and socio-economic studies, as it helps to identify areas that may require more attention and resources.

Understanding the distribution of the LSTAT variable aids in a better comprehension of the overall population dynamics and their implications on various aspects of community life.

Statistical Analyses (LSTAT Variable)

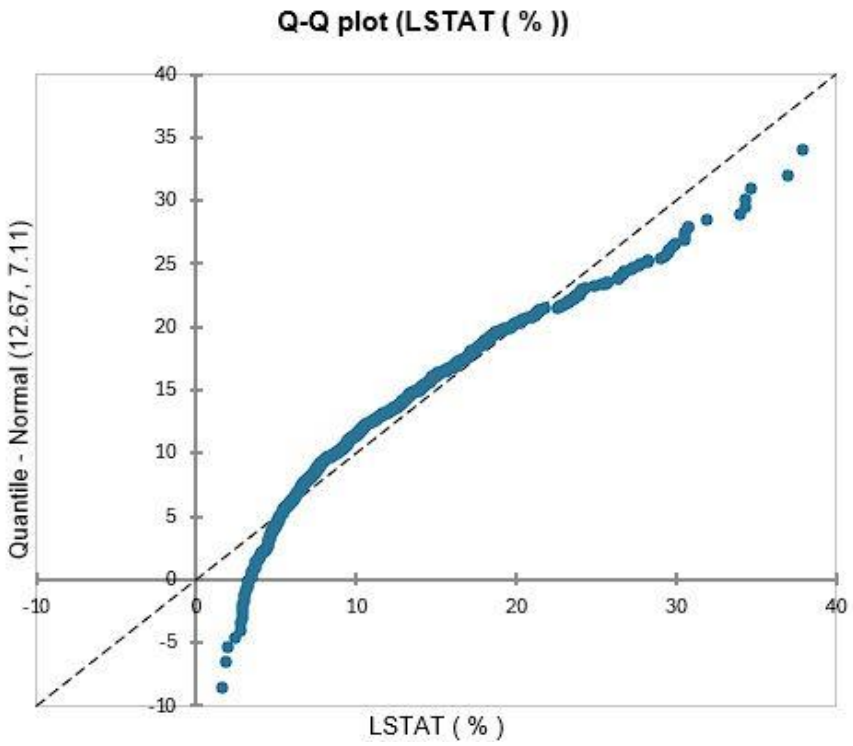
Examining The Descriptive Statistics

Statistic	LSTAT (%)
Nbr. of observations	506
Nbr. of missing values	0
Obs. without missing data	506
Minimum	1.730
Maximum	37.970
Freq. of minimum	1
Freq. of maximum	1
Range	36.240
1st Quartile	7.125
Median	11.330
3rd Quartile	16.930
Sum	6411.220
Mean	12.670
Variance (n)	50.482
Variance (n-1)	50.582
Standard deviation (n)	7.105
Standard deviation (n-1)	7.112
Skewness (Pearson)	0.916
Kurtosis (Pearson)	0.516
Lower bound on mean (95%)	12.049
Upper bound on mean (95%)	13.292
Lower bound on variance (95%)	44.880
Upper bound on variance (95%)	57.450

- There are 506 observations in this variable's column
- there are not any missing values for this variable
- All of the records are filled with data
- Minimum value of this variable
- Maximum value of this variable
- Minimum value of this variable can be seen only 1 time among all of the records
- Maximum value of this variable can be seen only 1 time among all records
- Maximum - Minimum
- 25% of data of this variable are below this value and 75% of our data are greater than this number
- 50% of data of this variable are below this value and 50% of our data are greater than this number
- 75% of data of this variable are below this value and 25% of our data are greater than this number
- Sum of all values in this variable's column
- Average of our sample
- The variance of the population for this variable
- The variance of the sample for this variable
- The standard deviation of the population for this variable
- The standard deviation of the sample for this variable
- A skewness of 0.91 suggests a moderate positive skew, with most data points clustered towards the lower end and some higher values extending the right tail. This helps understand the distribution pattern and identify the tendency towards higher values.
- A kurtosis value of 0.51 suggests that the distribution is not perfectly normal but has slightly fatter tails and a sharper peak.
- The mean of the population of this variable must be something between 12.04 and 13.2 with confidence level of 95%
- The variance of the population of this variable must be something between 44.8 and 57.4 with confidence level of 95%

Statistical Analyses (LSTAT Variable)

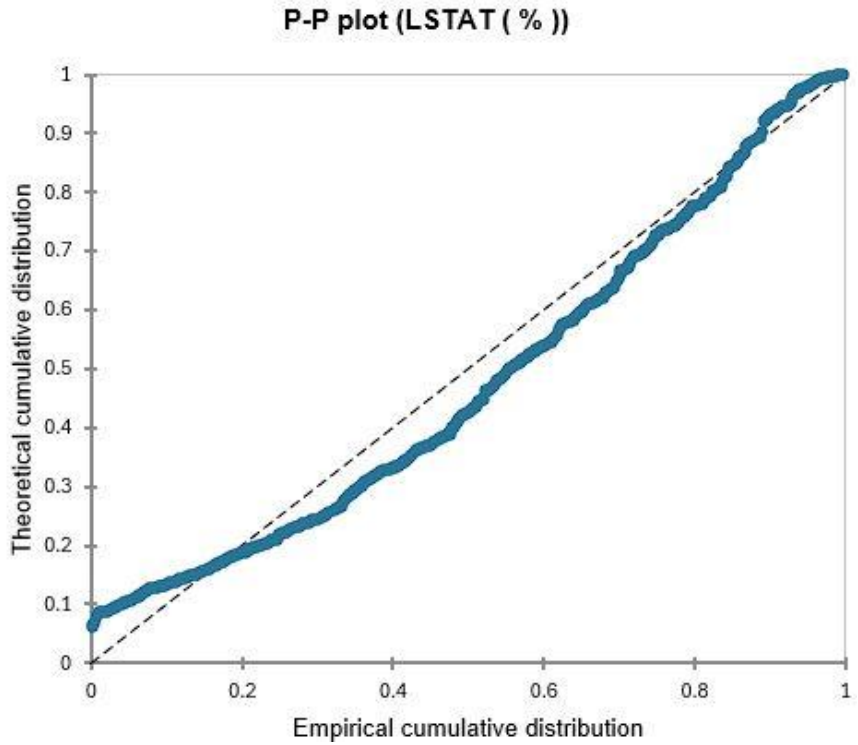
Normality Test (Anderson-Darling Method)



Anderson-Darling test (LSTAT (%)):	
A^2	7.999
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test Result :
p-value is less than alpha, so we should reject the null hypothesis. So; LSTAT variable does not follow a normal distribution.

P-P & Q-Q plot:
These plots show us that there is a difference between LSTAT variable's distribution and a normal distribution as the normality test's result unveiled this fact to us.



Statistical Analyses (LSTAT Variable)

Outliers Detecting (Variable Transformation)

Raw data:
This is the raw data of LSTAT variable without any transformations.

Z-Score Normalization:
In this column, I transformed the data of LSTAT variable with use of excel functions.
I create a function like this :

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma}$$

XLSTAT Check (Z-Score Normalization):
In this column, I transformed the data of LSTAT variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

	A	B	C	D	E
1	LSTAT (%)	LSTAT (Z-transformation)	LSTAT (Normalization)	Standardized (n-1)	0 to 1
2	4.98	-1.081311813	0.089679912	-1.081311813	0.089679912
3	9.14	-0.496392964	0.204470199	-0.496392964	0.204470199
4	4.03	-1.214887031	0.063465784	-1.214887031	0.063465784
5	2.94	-1.368147017	0.033388521	-1.368147017	0.033388521
6	6.29	-0.897118618	0.125827815	-0.897118618	0.125827815

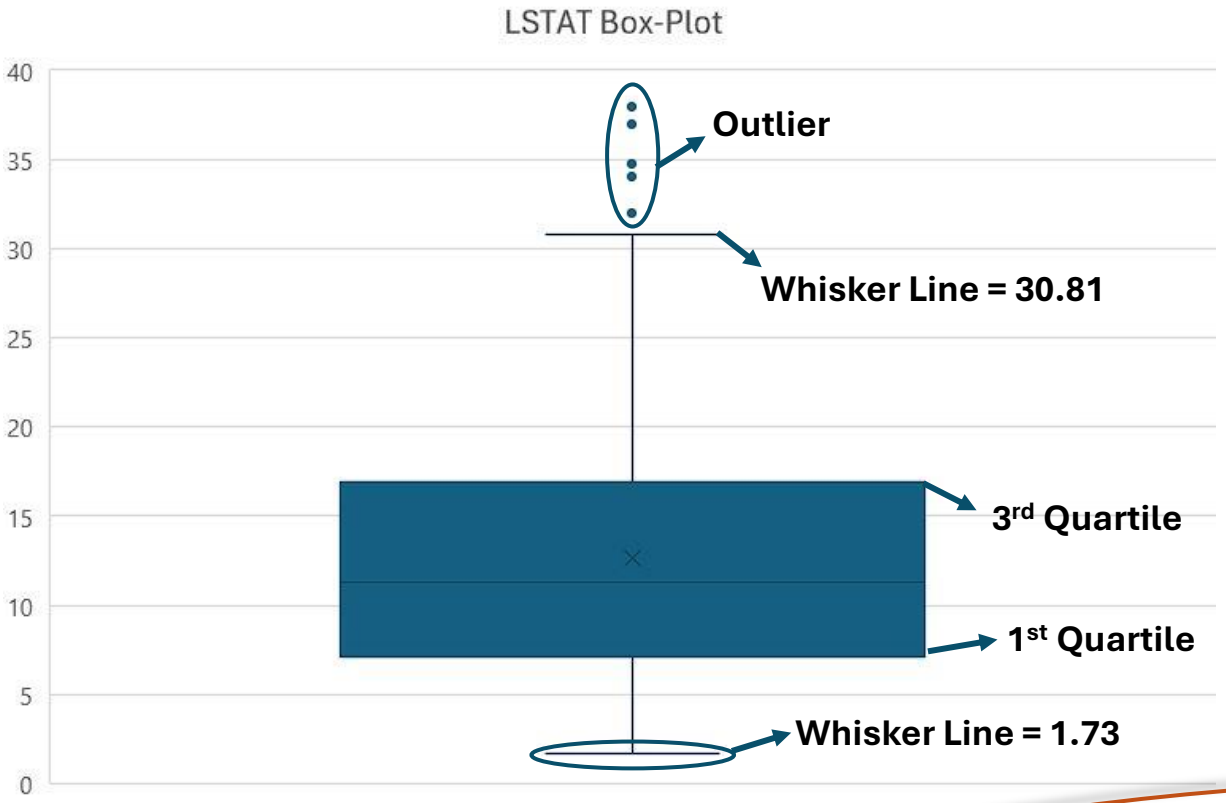
Normalization:
In this column, I normalized the data of LSTAT variable with use of excel functions.
I create a function like this :

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

XLSTAT Check (Normalization):
In this column, I normalized the data of LSTAT variable with use of XLSTAT transformation option to doublecheck my transformation with excel functions.

Statistical Analyses (LSTAT Variable)

Outliers Detecting (Box-Plot Method)



Outliers :

In the analysis of the LSTAT variable, representing the percentage of lower status population, using the box-plot method, we identified seven outliers. These outliers are exclusively found at the higher values of the LSTAT distribution. This indicates that there are several neighborhoods with significantly higher percentages of lower status individuals compared to the rest of the dataset.

The presence of these high-value outliers highlights areas that may require additional attention and resources to address socio-economic disparities. Understanding these outliers is crucial for identifying and supporting communities with higher needs, ensuring a more equitable approach in policy-making and resource allocation.

Average: 34.93 Count: 7 Min: 31.99 Max: 37.97 Sum: 244.51

Whiskers & Box :

IQR (= 3rd quartile – 1st quartile)

whisker lines :
3rd quartile + 1.5 IQR = 30.81
1st quartile – 1.5 IQR = 1.73

Outliers:
Values of LSTAT which are above 30.81

Conclusion :

In conclusion, the analysis of the LSTAT variable provides valuable insights into the socio-economic distribution within the dataset. The identification of seven high-value outliers using the box-plot method highlights neighborhoods with significantly higher percentages of lower-status individuals. These outliers are critical for understanding areas that may require additional attention and resources to address socio-economic disparities. The uni-modal and positively skewed distribution further underscores the commonality and concentration of lower status across many neighborhoods. By examining these patterns, we can better inform policy-making, urban planning, and resource allocation, ensuring a more equitable and supportive approach to community development. This comprehensive analysis of the LSTAT variable enriches our understanding of the population dynamics and their implications for the overall well-being of the neighborhoods.

Statistical Analyses (LSTAT Variable)

Outliers Detecting (Z-Score Method)

	A	B
1	LSTAT (%)	LSTAT (Z-transformation)
143	34.41	3.056707832
375	34.77	3.107325809
376	37.97	3.557263385
414	34.37	3.051083612
416	36.98	3.418063948
440	34.02	3.00187169

Outliers With Z-Score Method:

As we know, in Z-Score method for detecting outliers, values which are greater than (average + 3 x standard deviation) and less than (average – 3 x standard deviation) are known as outliers.
So; after standardizing the variable, I applied a conditional formatting on this variable to detect values which are greater than 3 or less than -3, to keep the track of the outliers of this feature based on Z-Score method and I got the result as you can see in the table.

Average: 3.198719379 Count: 6 Min: 3.00187169 Max: 3.557263385 Sum: 19.19231628

6 Outliers:

6 outliers are detected based on Z-Score method.
While, the number of outliers which were detected based on box-plot method was 7.

Box-Plot VS Z-Score :

In addition to the box-plot method, the Z-score method was used to identify outliers in the LSTAT variable, representing the percentage of the lower status population. This method revealed six outliers, indicating neighborhoods with significantly higher percentages of lower-status individuals compared to the overall dataset. The consistency of outliers found with both methods underscores the reliability of these observations. These high-value outliers are crucial for pinpointing areas with greater socio-economic challenges, enabling targeted interventions and resource allocation to support these communities effectively.

Statistical Analyses (LSTAT Variable)

Outliers Detecting (Grubbs Method)

Concept:

As we saw before, LSTAT variable is not normally distributed, and as we know, for detecting outliers with Grubbs method, our variable must follow a normal distribution otherwise we cannot apply Grubbs test on it. So; I transformed this variable with box-cox method, and applied a normality test again to see if now, it follows a normal distribution, and the answer was negative to this question.

On the second step, I removed the outliers of this variable and again applied a normality test to see if it now follows a normal distribution and the answer to this question was also negative.

So, as the conclusion, we find it out that we cannot convert the LSTAT variable to a normally distributed variable. So as the result, we cannot apply Grubbs method for detecting the outliers of this variable.

Box-Cox transformation
1.934242913
2.868599242
1.637460147
1.220727213
2.27853734

Transformed data of LSTAT Variable With Box-Cox Method

Anderson-Darling test (Box-Cox transformation):	
A^2	0.771
p-value (Two-tailed)	**0.045**
alpha	0.050

Normality Test After Box-Cox Transformation :

As we can see, the result of the normality test of transformed data (with box-cox method), LSTAT variable still does not follow a normal distribution.

Anderson-Darling test (Box-Cox transformation):	
A^2	1.099
p-value (Two-tailed)	**0.007**
alpha	0.050

Normality Test After Removing Outliers :

As we can see, even after removing the outliers of the LSTAT variable and conducting a normality test again, this variable is not following a normal distribution.

Statistical Analyses (LSTAT Variable)

Correlation Test With The Target Variable (Pearson Method)

Why Pearson Method:

I am going to check the correlation between LSTAT variable and target variable which is MEDV.
Both are continuous variables, and because of this reason I should use appropriate corresponding method; which for checking the correlation between two continuous variables is Pearson method.

Correlation matrix (Pearson):

Variables	LSTAT (%)	MEDV (1000$)
LSTAT (%)	**1**	**-0.737**
MEDV (1000$)	**-0.737**	**1**

Strong And Inverse Correlation:

The correlation matrix and the value of -0.73 tells us that there is an inverse correlation between these 2 variables.
Meaning that if one of the increase, the other one will decrease.
On the other hand, the absolute value would be 0.73, which indicates that the correlation is strong.
LSTAT feature hast the strongest correlation with target variable among all of features of this dataset.

Coefficients of determination (Pearson):

Variables	LSTAT (%)	MEDV (1000$)
LSTAT (%)	**1**	0.543
MEDV (1000$)	0.543	**1**

p-values (Pearson):

Variables	LSTAT (%)	MEDV (1000$)
LSTAT (%)	**0**	**<0.0001**
MEDV (1000$)	**<0.0001**	**0**

Statistical Significance Of The Correlation:

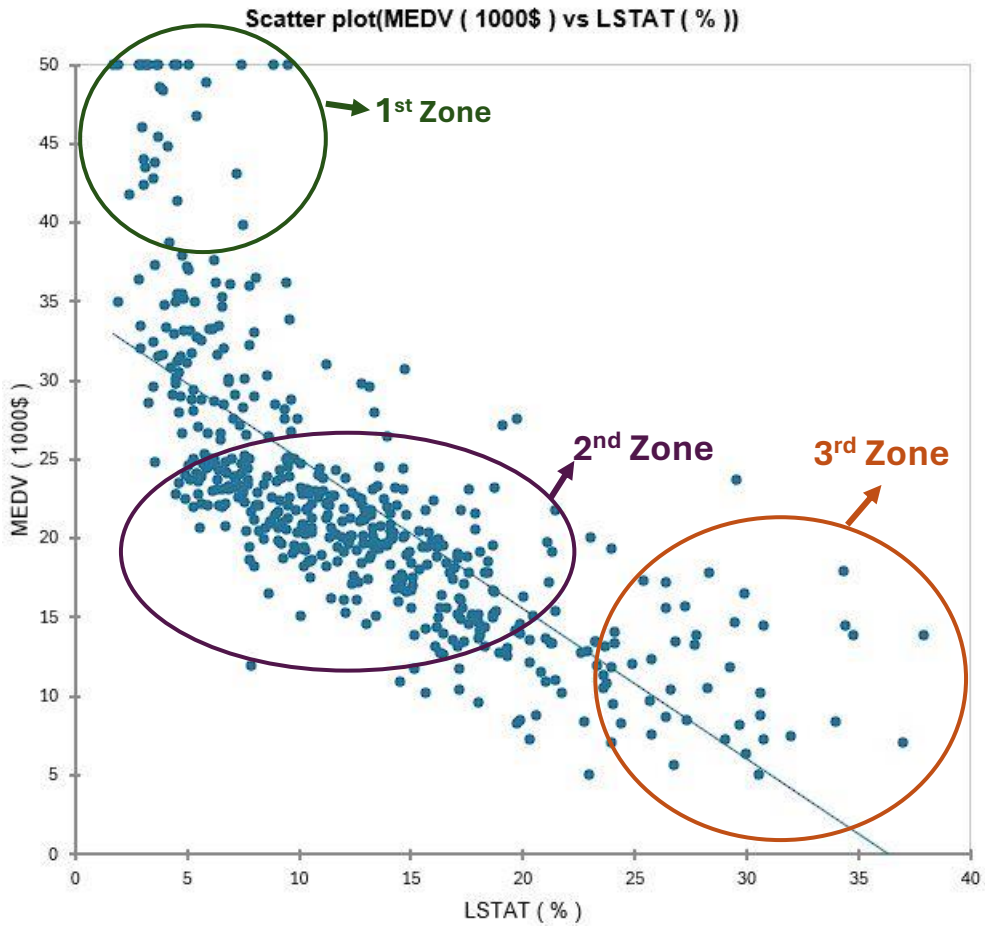
The value is <0.0001 suggests that the correlation between LSTAT and MEDV is statistically significant and it is not due to random changes.

Power Of Prediction:

The value of 0.543 in this table, indicates that only 54.3% of the variance in target variable (MEDV) can be explained by the variance in LSTAT variable.

Statistical Analyses (LSTAT Variable)

Scatter Plot With The Target Variable



3 Zones:

As we see on the chart, we can divide the city into 3 zones based on median value of houses in each area and LSTAT value of each area.
This gives us an interesting insight as we can interpreter as following :

1st Zone:

Areas which are included in this zone, can be attributed to upper-class families.
Records of this zone, have the highest values of MEDV.
Houses prices in this zone are higher than the others, and as a result, LSTAT variable have the lowest values of itself in these areas.

2nd Zone:

This zone can be attributed to middle-class families.
As we can see on the chart, the number of records and concentration of them in this zone is much more higher than the other two zones.

3rd Zone:

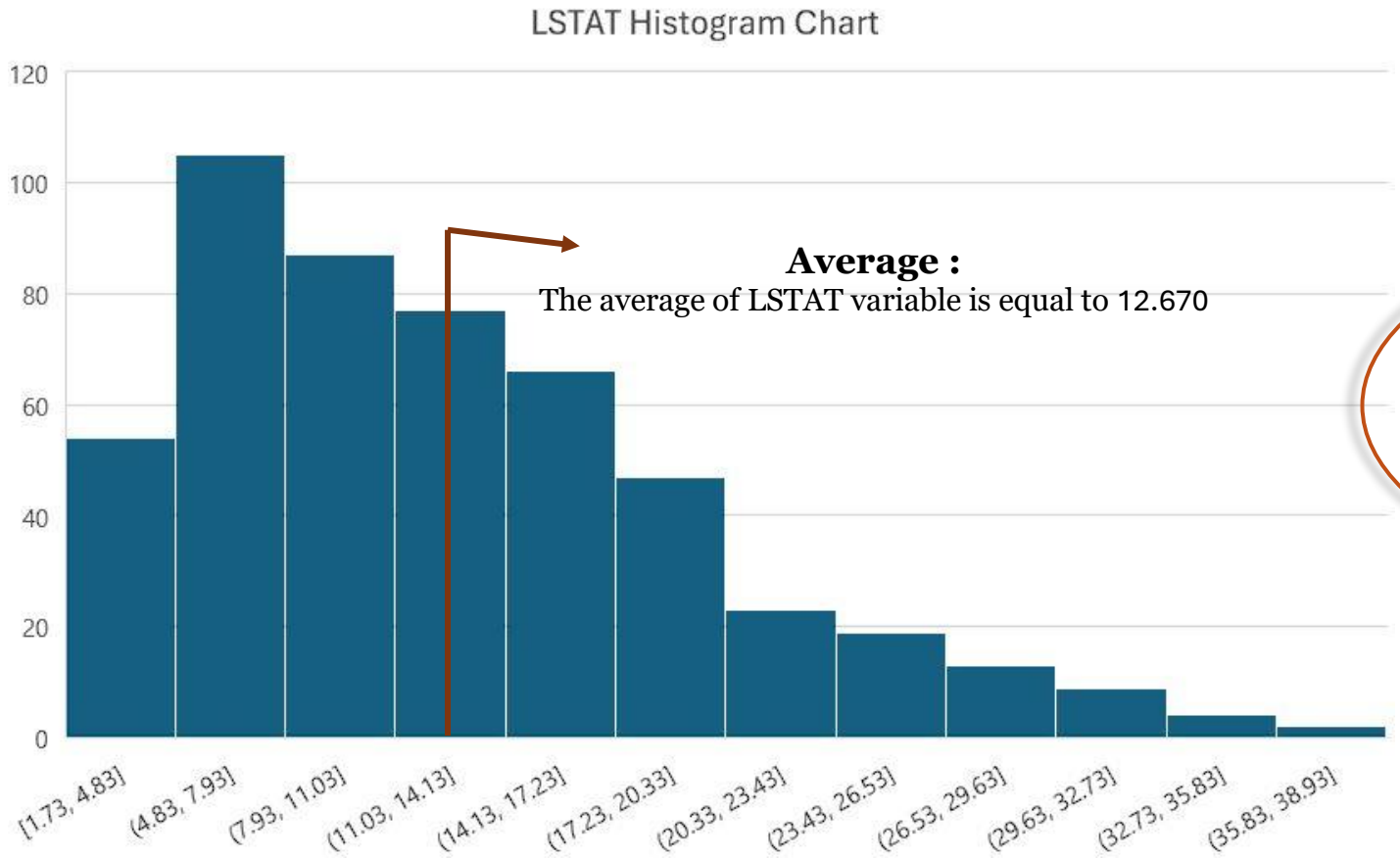
This zone can be attributed to working-class families.
As we can see on the chart, this zone includes areas that LSTAT value in them is higher than the other two zones.
Meaning that there are more families in this zone which experience bad economical conditions.
House prices in areas of this zone is relatively lower than house prices in any other areas.
Also, the concentration of records in this zone is lower than the concentration of records in other zones.
This suggest that majority of people in Boston, have a normal financial conditions.

Strong And Inverse Correlation:

Trend line on the chart, shows a strong and inverse correlation between LSTAT and MEDV variables.
Meaning that, as one of the increases, the other one will decrease.
And the slope of the line shows that this inverse correlation is so strong.
This fact suggests (as it would be obvious), that areas of this town, that house price in the Is so high, have less residentials with financial problem.
And areas which the majority of its residentials of it, are facing financial difficulties, have low house prices.

Statistical Analyses (LSTAT Variable)

Question : Is There Any Difference Between Average Of House Prices Based On LSTAT Variable?



Average :
The average of LSTAT variable is equal to 12.670

LSTAT 2 Classes :
I am going to create a new feature based on LSTAT feature.
This feature is going to be 0 for areas which have LSTAT values lower than the average.
And is going to be 1 for areas which have LSTAT values greater than the average.
I chose these 2 classes because it seems to me interesting and logical to compare these two classes in terms of the average of their MEDVs.

	A	B	C
1	LSTAT (%)	LSTAT (Binary Classification)	MEDV (1000$)
2	4.98	=IF(A2>12.67,1,0)	24
3	9.14	0	21.6
4	4.03	0	34.7
5	2.94	0	33.4

Statistical Analyses (LSTAT Variable)

Variances Equality Test (Leven's Method)

Levene's test (Mean) / Two-tailed test:	
F (Observed value)	52.068
F (Critical value)	3.860
DF1	1
DF2	504
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Variances Equality Test:

As we can see, P-value is less than alpha, so; we should reject the null hypothesis. Meaning that variance of house prices with class 1 (those with LSTAT values greater than the average) , is not equal to variance of house prices with class 0 (those with LSTAT values less than the average).
So; for comparing the average of house prices between these two classes, with should not assume the equality of variances.

Why Leven's Method :

As we saw before, MEDV variable is not normally distributed. So; for checking the equality of variances of MEDV variable based on different categories of LSTAT variable, we should use the appropriate method which is Leven's test.
If MEDV was normally distributed, Fisher's test must be conducted

Statistical Analyses (LSTAT Variable)

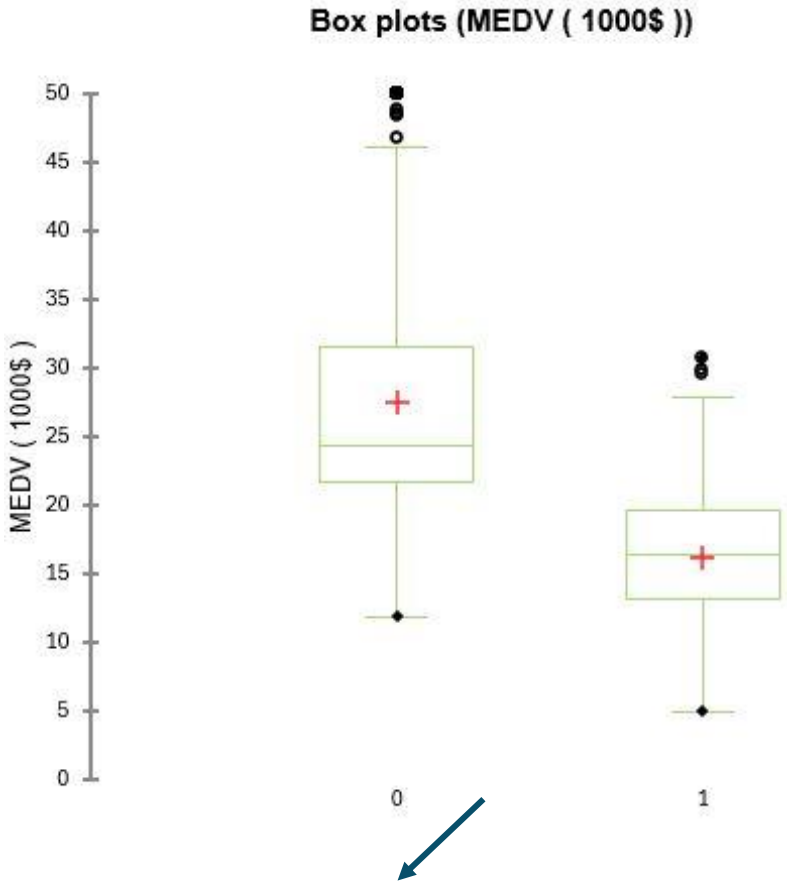
Average Equality Test (T- test)

t-test for two independent samples / Two-tailed test:

95% confidence interval on the difference between the means:

[10.047, 12.464]

Difference	11.256
t (Observed value)	18.302
|t| (Critical value)	1.965
DF	459.124
p-value (Two-tailed)	**<0.0001**
alpha	0.050



Higher Average Of House Prices:

This tells us that areas of class 0, have higher house prices on average, comparing to areas of class 1
Meaning that on average, houses which are in areas with less residentials which have financial difficulties, are more expensive than the others

Not Equal:

P-value is less than alpha, so; we should reject the null hypothesis. Meaning that the average of house prices for those areas which have LSTAT values less than the average (class 0) is not equal to the average of house prices for those areas which have LSTAT values greater than the average (class 1).

Statistical Analyses (PTRATIO Variable)

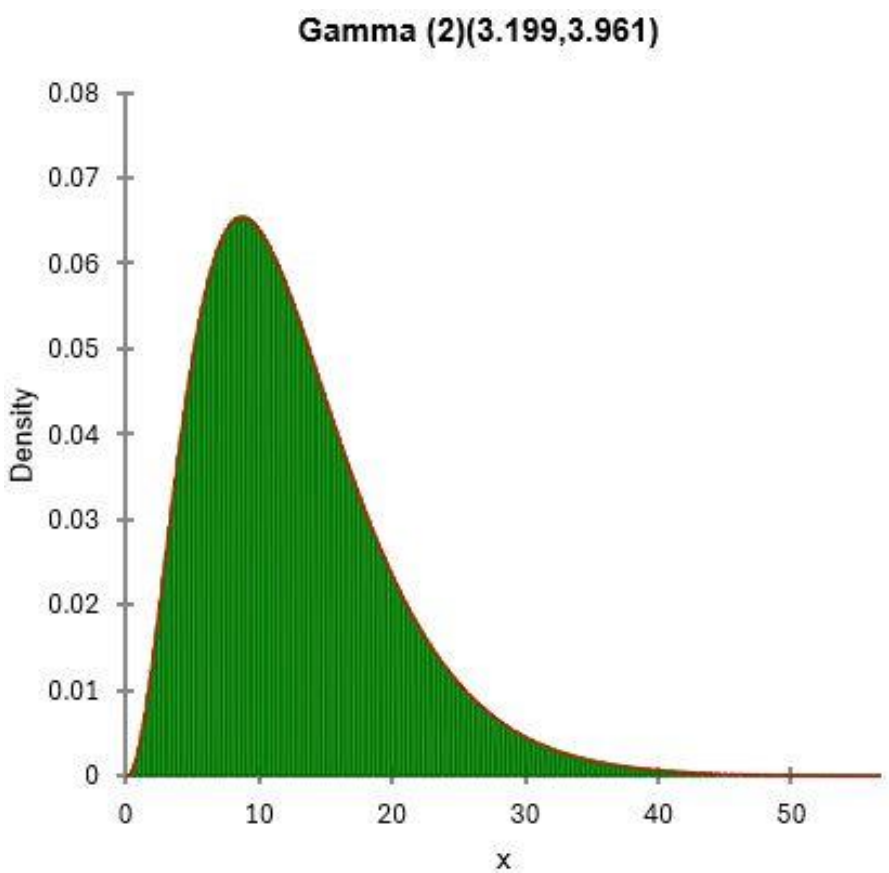
Distribution	p-value
Chi-square	<0.0001
Erlang	0.052
Fisher-Tippett (1)	<0.0001
Fisher-Tippett (2)	0.152
Gamma (1)	<0.0001
Gamma (2)	**0.503**
GEV	0.017
Gumbel	<0.0001
Log-normal	0.101
Logistic	0.001
Normal	0.001
Student	<0.0001
Weibull (2)	0.353

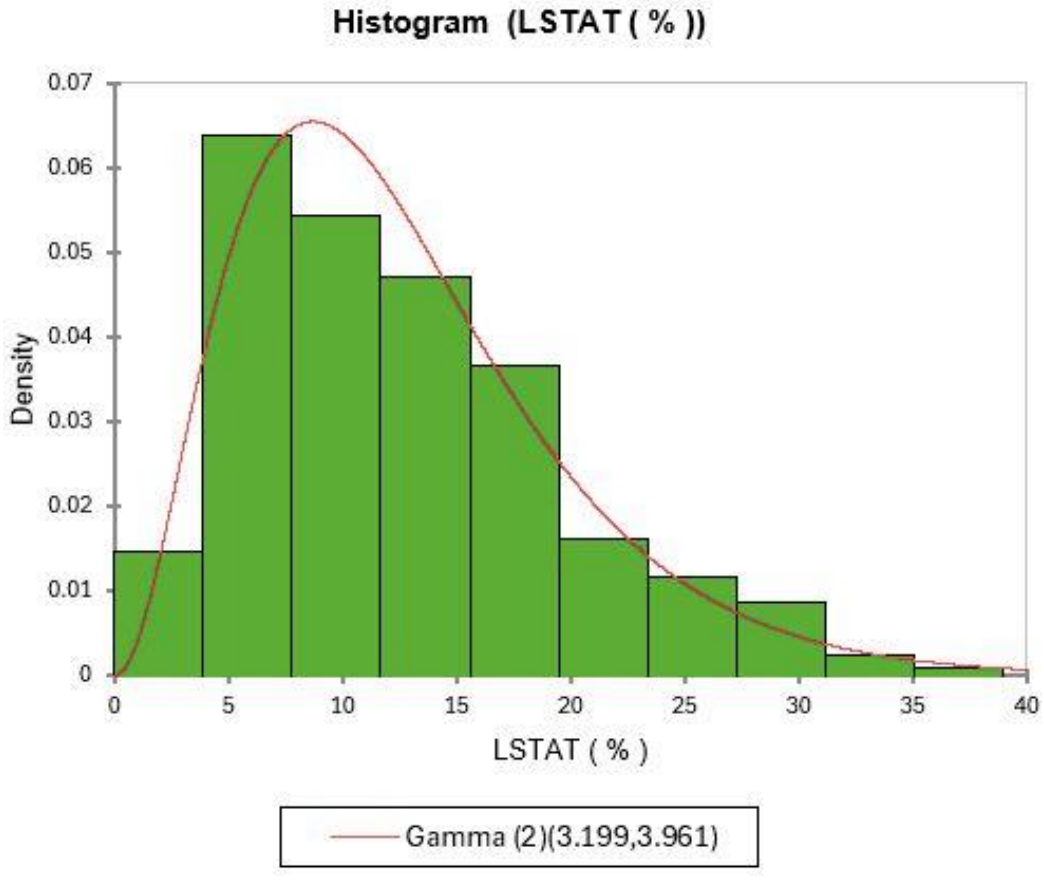
The Best Fitting Distribution

Gamma (2) Distribution :

With use of XLSTAT, I found out that the best fitting distribution for LSTAT variable, is gamma (2) distribution with given parameter as below (K & β) Then again, with use of XLSTAT I plot the distribution with these parameters and its corresponding value and I got the chart which you can see on the right.

Estimated parameters (Gamma (2)):		
Parameter	Value	Standard error
k	3.199	0.194
beta	3.961	0.259





Linear Regression Modeling Of Target Variable

ANCOVA Technique (1st Page, Summary Statistics)

Summary statistics (Quantitative data):

Variable	Observations	Obs. with missing	Obs. without	Minimum	Maximum	Mean	Std. deviation
MEDV (1000$)	506	0	506	5.000	50.000	22.533	9.197
CRIM	506	0	506	0.006	88.976	3.470	8.575
ZN (%)	506	0	506	0.000	100.000	10.769	23.025
INDUS (%)	506	0	506	0.460	27.740	11.120	6.840
NOX (PPM)	506	0	506	0.385	0.871	0.555	0.116
RM	506	0	506	3.561	8.780	6.285	0.703
AGE (%)	506	0	506	2.900	100.000	68.709	27.833
DIS (Miles)	506	0	506	1.130	12.127	3.795	2.106
TAX (10,000$)	506	0	506	187.000	711.000	408.237	168.537
PTRATIO	506	0	506	12.600	22.000	18.456	2.165
B	506	0	506	0.320	396.900	356.674	91.295
LSTAT (%)	506	0	506	1.730	37.970	12.670	7.112

Summary statistics (Qualitative data):

Variable	Categories	Counts	Frequencies	%
CHAS	0	471	471	93.083
	1	35	35	6.917
RAD	1	20	20	3.953
	2	24	24	4.743
	3	38	38	7.510
	4	110	110	21.739
	5	115	115	22.727
	6	26	26	5.138
	7	17	17	3.360
	8	24	24	4.743
	24	132	132	26.087

Summary Statistics Tables:

Conducting ANCOVA technique with the use of XLSTAT, we got 2 tables as summary statistics of our variables.
Our dataset includes both qualitative and quantitative variable, so; we got the summary of the in 2 different tables because of their different natures.

Variables' Names: This column contains the name of quantitative variables in our dataset.

Number Of Records: This column contains the number of records of each variable.

Number Of Missing Values: This column contains the number of missing values of each variable.

Number Of Complete Values: This column contains the number of records which are completed. (they are not missing values)

Minimum Value: This column contains the minimum value of each variable.

Maximum Value: This column contains the maximum value of each variable.

Average Value: This column contains the average value of each variable.

Standard Deviation Value: This column contains the standard deviation value of each variable.

Variables' names: This column contains the name of qualitative variables in our dataset.

Categories' Names: This column contains the name of different categories of each qualitative variable.

Number Of Records: These two columns contain the number of records of each category of each variable.

Proportion Of Each Category: This column contains the proportion of each category of each variable by percentage.

Linear Regression Modeling Of Target Variable

ANCOVA Technique (2nd Page, Correlation Matrix)

	CRIM	ZN (%)	INDUS (%)	NOX (PPM)	RM	AGE (%)	DIS (Miles)	TAX (10,000$)	PTRATIO	B	LSTAT (%)	CHAS-0	CHAS-1	RAD-1	RAD-2	RAD-3	RAD-4	RAD-5	RAD-6	RAD-7	RAD-8	RAD-24	MEDV (1000$)
CRIM	1	-0.186	0.394	0.411	-0.220	0.345	-0.366	0.560	0.278	-0.365	0.451	0.058	-0.058	-0.081	-0.088	-0.112	-0.189	-0.176	-0.090	-0.072	-0.082	0.607	-0.384
ZN (%)	-0.186	1	-0.519	-0.499	0.312	-0.546	0.632	-0.304	-0.395	0.170	-0.401	0.036	-0.036	0.224	0.094	0.070	0.068	0.004	0.011	0.097	-0.044	-0.278	0.362
INDUS (%)	0.394	-0.519	1	0.767	-0.391	0.651	-0.712	0.725	0.386	-0.359	0.602	-0.051	0.051	-0.180	-0.054	-0.279	-0.034	-0.108	-0.100	-0.163	-0.169	0.607	-0.484
NOX (PPM)	0.411	-0.499	0.767	1	-0.302	0.733	-0.769	0.668	0.189	-0.380	0.591	-0.066	0.066	-0.161	-0.135	-0.252	-0.229	0.076	-0.080	-0.183	-0.120	0.604	-0.427
RM	-0.220	0.312	-0.391	-0.302	1	-0.242	0.205	-0.292	-0.356	0.128	-0.614	-0.106	0.106	0.078	0.116	0.076	-0.114	0.084	-0.060	0.096	0.212	-0.222	0.695
AGE (%)	0.345	-0.546	0.651	0.733	-0.242	1	-0.746	0.508	0.262	-0.278	0.605	-0.082	0.082	-0.164	-0.031	-0.199	-0.145	0.018	-0.072	-0.192	-0.013	0.448	-0.382
DIS (Miles)	-0.366	0.632	-0.712	-0.769	0.205	-0.746	1	-0.534	-0.232	0.292	-0.498	0.089	-0.089	0.215	0.032	0.183	0.160	-0.025	0.025	0.239	0.065	-0.490	0.250
TAX (10,000$	0.560	-0.304	0.725	0.668	-0.292	0.508	-0.534	1	0.461	-0.442	0.544	0.041	-0.041	-0.141	-0.196	-0.274	-0.226	-0.246	-0.049	-0.115	-0.142	0.910	-0.469
PTRATIO	0.278	-0.395	0.386	0.189	-0.356	0.262	-0.232	0.461	1	-0.177	0.375	0.107	-0.107	-0.084	-0.120	-0.038	0.166	-0.479	-0.069	-0.004	-0.050	0.479	-0.508
B	-0.365	0.170	-0.359	-0.380	0.128	-0.278	0.292	-0.442	-0.177	1	-0.367	-0.053	0.053	0.073	0.073	0.112	0.151	0.074	0.078	0.065	0.070	-0.447	0.333
LSTAT (%)	0.451	-0.401	0.602	0.591	-0.614	0.605	-0.498	0.544	0.375	-0.367	1	0.055	-0.055	-0.150	-0.083	-0.140	-0.037	-0.152	-0.013	-0.123	-0.142	0.496	-0.737
CHAS-0	0.058	0.036	-0.051	-0.066	-0.106	-0.082	0.089	0.041	0.107	-0.053	0.055	1	-1.000	0.015	0.061	0.019	-0.026	-0.038	0.063	0.051	-0.122	0.020	-0.183
CHAS-1	-0.058	-0.036	0.051	0.066	0.106	0.082	-0.089	-0.041	-0.107	0.053	-0.055	-1.000	1	-0.015	-0.061	-0.019	0.026	0.038	-0.063	-0.051	0.122	-0.020	0.183
RAD-1	-0.081	0.224	-0.180	-0.161	0.078	-0.164	0.215	-0.141	-0.084	0.073	-0.150	0.015	-0.015	1	-0.045	-0.058	-0.107	-0.110	-0.047	-0.038	-0.045	-0.121	0.040
RAD-2	-0.088	0.094	-0.054	-0.135	0.116	-0.031	0.032	-0.196	-0.120	0.073	-0.083	0.061	-0.061	-0.045	1	-0.064	-0.118	-0.121	-0.052	-0.042	-0.050	-0.133	0.104
RAD-3	-0.112	0.070	-0.279	-0.252	0.076	-0.199	0.183	-0.274	-0.038	0.112	-0.140	0.019	-0.019	-0.058	-0.064	1	-0.150	-0.155	-0.066	-0.053	-0.064	-0.169	0.167
RAD-4	-0.189	0.068	-0.034	-0.229	-0.114	-0.145	0.160	-0.226	0.166	0.151	-0.037	-0.026	0.026	-0.107	-0.118	-0.150	1	-0.286	-0.123	-0.098	-0.118	-0.313	-0.066
RAD-5	-0.176	0.004	-0.108	0.076	0.084	0.018	-0.025	-0.246	-0.479	0.074	-0.152	-0.038	0.038	-0.110	-0.121	-0.155	-0.286	1	-0.126	-0.101	-0.121	-0.322	0.187
RAD-6	-0.090	0.011	-0.100	-0.080	-0.060	-0.072	0.025	-0.049	-0.069	0.078	-0.013	0.063	-0.063	-0.047	-0.052	-0.066	-0.123	-0.126	1	-0.043	-0.052	-0.138	-0.039
RAD-7	-0.072	0.097	-0.163	-0.183	0.096	-0.192	0.239	-0.115	-0.004	0.065	-0.123	0.051	-0.051	-0.038	-0.042	-0.053	-0.098	-0.101	-0.043	1	-0.042	-0.111	0.093
RAD-8	-0.082	-0.044	-0.169	-0.120	0.212	-0.013	0.065	-0.142	-0.050	0.070	-0.142	-0.122	0.122	-0.045	-0.050	-0.064	-0.118	-0.121	-0.052	-0.042	1	-0.133	0.190
RAD-24	0.607	-0.278	0.607	0.604	-0.222	0.448	-0.490	0.910	0.479	-0.447	0.496	0.020	-0.020	-0.121	-0.133	-0.169	-0.313	-0.322	-0.138	-0.111	-0.133	1	-0.396
MEDV (1000$	-0.384	0.362	-0.484	-0.427	0.695	-0.382	0.250	-0.469	-0.508	0.333	-0.737	-0.183	0.183	0.040	0.104	0.167	-0.066	0.187	-0.039	0.093	0.190	-0.396	1

Correlation Between Independent Variables:

Values in the right triangle, show the correlation between the independent variables of our dataset.
We can use the to detect pair of independent variable which have strong correlation with each other and remove one of them from our model.
Why should we do that is because of the multicollinearity phenomena.
Having two independent variables which are correlated in our model does not make sense because one of them would be effective and the other one is just increasing the complexity of our model.

Correlations Between Independent Variables And The Dependent Variable:

Values in the blue rectangle, show the correlations between independent variables of our dataset and our target variable which is MEDV.
The absolute value of these number range from 0 to 1 and as it gets closer to one, it shows more strong correlation which can be helpful for us to create our model base on.
As a result, we can use the values in this box to choose the most beneficial independent variables for creating our model.

Linear Regression Modeling Of Target Variable

ANCOVA Technique (3rd Page, The Best Model)

Simple Linear Regression:

If we wanted to use only one variable in our linear regression model to predict the target variable, we could use LSTAT feature which shows the proportion of low-status people in each area of our sample.
We could predict the target variable by the use of LSTAT variable with 54% of accuracy, but this number is less than 70% and that's why we are not interested in.

Regression of variable MEDV (1000$):

Summary of the variables selection MEDV (1000$):

Nbr. of variables	Variables	MSE	R^2	Adjusted R^2	Mallows' Cp	Akaike's AIC	Schwarz's SBC	Amemiya's PC
1	LSTAT (%)	38.742	0.543	0.542	381.317	1852.397	1860.851	0.459
2	RM / LSTAT (%)	30.761	0.638	0.636	199.961	1736.672	1749.352	0.365
3	RM / PTRATIO / LSTAT (%)	27.424	0.678	0.676	124.797	1679.569	1696.476	0.326
4	RM / DIS (Miles) / PTRATIO / LSTAT (%)	26.490	0.689	0.687	104.378	1663.019	1684.152	0.316
5	NOX (PPM) / RM / DIS (Miles) / PTRATIO / LSTAT (%)	25.005	0.707	0.704	71.584	1634.810	1660.169	0.299
6	NOX (PPM) / RM / DIS (Miles) / PTRATIO / LSTAT (%) / I	24.285	0.720	0.713	62.523	1627.877	1687.049	0.294
7	NOX (PPM) / RM / DIS (Miles) / PTRATIO / B / LSTAT (%	23.540	0.729	0.722	46.863	1613.069	1676.467	0.286
8	NOX (PPM) / RM / DIS (Miles) / PTRATIO / B / LSTAT (%	23.059	0.735	0.727	37.140	1603.595	1671.220	0.281
9	ZN (%) / NOX (PPM) / RM / DIS (Miles) / PTRATIO / B / L	22.593	0.741	0.733	27.791	1594.234	1666.085	0.275
10	CRIM / ZN (%) / NOX (PPM) / RM / DIS (Miles) / PTRATI(	22.195	0.746	0.738	19.976	1586.198	1662.276	0.271
11	CRIM / ZN (%) / NOX (PPM) / RM / DIS (Miles) / TAX (1C	22.018	0.749	**0.740**	17.075	1583.111	1663.415	0.270
12	CRIM / ZN (%) / INDUS (%) / NOX (PPM) / RM / DIS (Mil	22.061	0.749	0.739	19.031	1585.065	1669.596	0.271
13	CRIM / ZN (%) / INDUS (%) / NOX (PPM) / RM / AGE (%	22.105	0.749	0.739	21.000	1587.033	1675.790	0.272

My Suggested Selection Of Variables:

I suggest this selection of independent variable.
This selection is made of 5 independent variables (6 variables less than the recommended selection by XLSTAT), while its power of prediction is only 4% less than the XLSTAT recommended selection.
On the other hand, we can predict the target variables by this selection of independent variables by 70% accuracy which is an acceptable power of prediction.
I am going to model my target variable with this selection of independent variables.

More Variables, Less Accuracy :

It is so interesting that what have happened here.
If we look at the row number 11, which has 11 independent variables selected for predicting the target variable, the accuracy of its prediction is 74%, while if we look at the number 13, which has all the available variables selected, the accuracy of its prediction is less than the row number 11.
Also, the difference is not huge and it may be occurred due to calculation reasons, it reminds us that more variables selected, does not necessarily mean a better combination you have for prediction purposes.

The Best Suggested Selection By XLSTAT:

This selection of independent variables which is made of 11 features is the best selection which is suggested by XLSTAT for modeling our dependent variable (MEDV).
As we see on the chart, by bringing these 11 independent variables in our linear model, we can predict the target variable by 74% of accuracy.
But the number of variables is too much and this will increase the complexity of our model so; we are always looking for ways of reducing the number of independent variables which are going to be included in our model.
With paying attention to this point, I will look for other selections which have less selected features.

Linear Regression Modeling Of Target Variable

ANCOVA Technique (3rd Page, Metrics Of Regression)

Goodness of fit statistics (MEDV (1000$)):	
Observations	506
Sum of weights	506
DF	500
R^2	0.707
Adjusted R^2	0.704
MSE	25.005
RMSE	5.000
MAPE	17.949
DW	0.966
Cp	6.000
AIC	1634.810
SBC	1660.169
PC	0.300

Degree Of Freedom :

In a linear regression model, the degrees of freedom are typically calculated as the number of observations minus the number of parameters being estimated (including the intercept).

R-Squared :

R-squared, also known as the coefficient of determination, is a statistical measure in a regression model that represents the proportion of the variance in the dependent variable that is predictable from the independent variables. In other words, it indicates how well the regression model fits the observed data.

Because we used 5 variables in our model, we do not consider this metric as a yardstick, we use adjusted R-squared.

An R-squared value of 0.707 means that approximately 70.7% of the variance in the dependent variable can be explained by the independent variables in your regression model. This indicates that your model has a good level of explanatory power.

Adjusted R-Squared :

Adjusted R-squared is a modified version of R-squared that takes into account the number of predictors in your regression model. While R-squared always increases with the addition of more predictors, Adjusted R-squared only increases if the new predictor improves the model more than would be expected by chance.

An Adjusted R-squared value of 0.704 means that approximately 70.4% of the variance in the dependent variable can be explained by the independent variables in your regression model, after adjusting for the number of predictors.

MSE :

MSE stands for Mean Squared Error. It's a common metric used to evaluate the performance of regression models. The MSE measures the average squared difference between the observed actual outcomes and the predicted values by the model.

When the Mean Squared Error (MSE) is equal to 25, it means that, on average, the squared differences between the predicted values and the actual observed values are 25 units squared.

RMSE :

RMSE stands for Root Mean Squared Error. It's a standard way to measure the error of a model in predicting quantitative data. RMSE is the square root of the mean squared error (MSE), and it gives you an idea of how well your model's predictions compare to the actual data.

When the Root Mean Squared Error (RMSE) is equal to 5, it indicates that, on average, the differences between the predicted values by your regression model and the actual observed values are about 5 units.

MAPE :

MAPE stands for Mean Absolute Percentage Error. It's a metric used to measure the accuracy of a forecasting or regression model. MAPE expresses the error as a percentage, making it easier to understand and interpret.

A MAPE (Mean Absolute Percentage Error) value of 17.9 indicates that, on average, the prediction error of your model is 17.9%. In simpler terms, the predictions made by your model are off by approximately 17.9% from the actual values.

Linear Regression Modeling Of Target Variable

ANCOVA Technique (4th Page, Analysis Of Variance)

Degree Of Freedom:
Degrees of Freedom are the number of independent values or quantities that can vary in the analysis without breaking any constraints.

Sum Of Squares:
Sum of Squares (SS) measures the total variability in the data and helps break down this variability into different components

Between-Groups Sum Of Squares (SSB):
Represents the variation due to differences between group means.
Calculated as the sum of the squared differences between each group mean and the overall mean, weighted by the number of observations in each group.

Mean Squares :
Mean Squares (MS) are used in statistical analyses such as ANOVA (Analysis of Variance) to measure the average variability within data. They are calculated by dividing the Sum of Squares (SS) by their respective Degrees of Freedom (DF).

Between-Groups:
This row shows between-groups characteristics.

Withing-Groups:
This row shows withing-groups characteristics.

Total:
This row shows the totals.

Analysis of variance (MEDV (1000$)):

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	5	30213.967	6042.793	241.667	<0.0001
Error	500	12502.328	25.005		
Corrected Total	505	42716.295			

F :
F refers to the F-statistic. The F-statistic is used to determine whether there are significant differences between the means of the groups being compared. It's calculated by dividing the Mean Square Between (MSB) by the Mean Square Within (MSW):

Within-Groups Sum Of Squares (SSW):
A smaller p-value (typically less than 0.05) suggests that the observed differences between group means are statistically significant, and you can reject the null hypothesis (which states that there are no differences between group means).A larger p-value (greater than 0.05) indicates that the observed differences are not statistically significant, and you fail to reject the null hypothesis.

Between-Groups Degrees Of Freedom:
This is the number of groups minus 1.

Within-Groups Degrees of Freedom:
This is the total degrees of freedom minus the between-groups degrees of freedom.

Total Degrees of Freedom:
This is the total number of observations minus 1.

Total Sum Of Squares:
Represents the overall variability in the dependent variable.
Calculated as the sum of the squared differences between each observation and the overall mean.

Within-Groups Sum Of Squares (SSW):
Represents the variation within each group that is not explained by the model.
Calculated as the sum of the squared differences between each observation and its group mean.

Linear Regression Modeling Of Target Variable

ANCOVA Technique (5th Page, Model Parameters)

Parameters Of The Model :

This column shows all of the parameters which are used in the model to anticipate the MEDV based on selected independent variables.

Coefficients Of The Model :

This column shows the corresponding coefficients for each parameters of the model.

Model parameters (MEDV (1000$)):

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	37.465	4.620	8.110	**<0.0001**	28.389	46.541
NOX (PPM)	-18.103	3.264	-5.546	**<0.0001**	-24.515	-11.690
RM	4.177	0.412	10.127	**<0.0001**	3.367	4.988
DIS (Miles)	-1.187	0.169	-7.036	**<0.0001**	-1.518	-0.855
PTRATIO	-1.045	0.114	-9.187	**<0.0001**	-1.268	-0.821
LSTAT (%)	-0.581	0.048	-12.051	**<0.0001**	-0.676	-0.486

Lower & Upper Bonds :

Represent the confidence interval for each coefficient estimate.
A confidence interval provides a range within which we expect the true value of the coefficient to fall, with a certain level of confidence (95%)

Reliable Coefficients :

All of the values, are less than alpha, meaning that all of the coefficients which are used in the model created by the use of linear regression, are reliable.

Standard Error :

The Standard Error (often abbreviated as SE) tells us how much the estimated value of a coefficient might vary if you repeated your analysis with different samples of data.
It shows the precision of the coefficient estimate.
Smaller standard errors indicate more precise estimates.

T-Statistics :

The t-statistic is a measure used in hypothesis testing to determine whether a coefficient is significantly different from zero.
It is calculated as the coefficient estimate divided by its standard error.
High Absolute Value: A high absolute value of the t-statistic (either positive or negative) suggests that the corresponding predictor is significantly different from zero, implying a significant effect on the dependent variable.
Low Absolute Value: A low absolute value suggests that the predictor is not significantly different from zero, implying it might not have a significant effect.

Linear Regression Modeling Of Target Variable

ANCOVA Technique (6th Page, Normality Test On Residuals & Final Model)

Test on the normality of the residuals (Shapiro-Wilk) (MEDV (1000$)):	
W	0.902
p-value (Two-tailed)	**<0.0001**
alpha	0.050

Normality Test On Residuals:

Normality test on the residuals of our model is conducted, P-value is less than alpha which shows that the residuals of our model do not follow a normal distribution.
It lessen the reliability of the model, but does not have a significant effect.

Equation of the model (MEDV (1000$)):

MEDV (1000$) = 37.4651380769412-18.1025582854832*NOX (PPM)+4.17725911207922*RM-1.1868444586705*DIS (Miles)-1.04460521654236*PTRATIO-0.580929495022922*LSTAT (%)

Created Model:

This equation is the finial equation for predicting the target variable based on the independent variables which are available in our dataset.
As we see before, we can predict the target variable by this equation with accuracy of 70%