R Data Analysis Project Report

-Amir Sajjad Mohammadi

Introduction

This report presents a detailed analysis of customer service data. Techniques from statistics and machine learning were used to uncover patterns in repair, cost, and service behavior.

1. Load Packages and Data

Essential libraries are loaded, and the dataset is read using `read.csv()`

```r
library(dplyr)
library(ggplot2)
library(arules)
library(arulesViz)

df <- read.csv(file.choose())
```

---

2. Initial Data Exploration

We examine the structure and summary of the dataset and rename one column for clarity.

```r
str(df)
head(df)
summary(df)
names(df)
colnames(df)[colnames(df) == "Repair_Action_Desc"] <- "Action"
```

---

3. Handling Missing Values

We check and remove `Product_Date` since more than 50% of its values are missing.

```r
df$Product_Date[df$Product_Date == ""] <- NA
if (mean(is.na(df$Product_Date)) > 0.5) {
  df <- df %>% select(-Product_Date)
}
```

---

4. Repair Speed Metrics

We calculate the average time to assign and complete repairs. Result: TAT01 = 1.29 days, TAT02 = 6.65 days.

```r
mean_TAT01 <- mean(df$TAT01, na.rm = TRUE)
mean_TAT02 <- mean(df$TAT02, na.rm = TRUE)
```

---

5. Return Rate Calculation

Duplicate serial numbers indicate repeat repairs. Result: 194 duplicate serials found. Return rate = 0.06 (6%) indicates moderate repeat service occurrence.

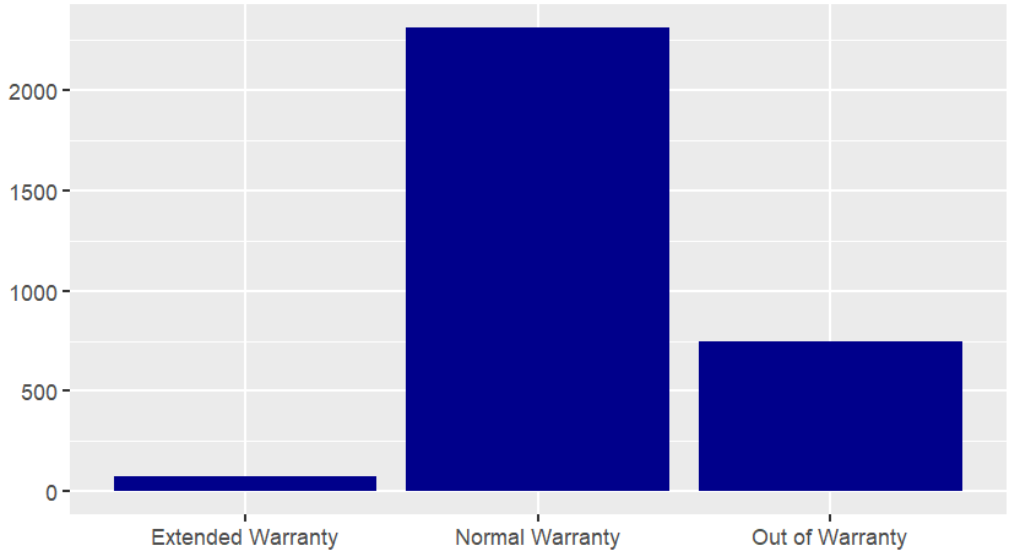
```r
duplicated_serials <- df[duplicated(df$Serial_No) |
duplicated(df$Serial_No, fromLast = TRUE), ]
return_rate <- nrow(duplicated_serials) / nrow(df)
```

---

6. Warranty Type Distribution

Bar chart showing the frequency of different warranty types.

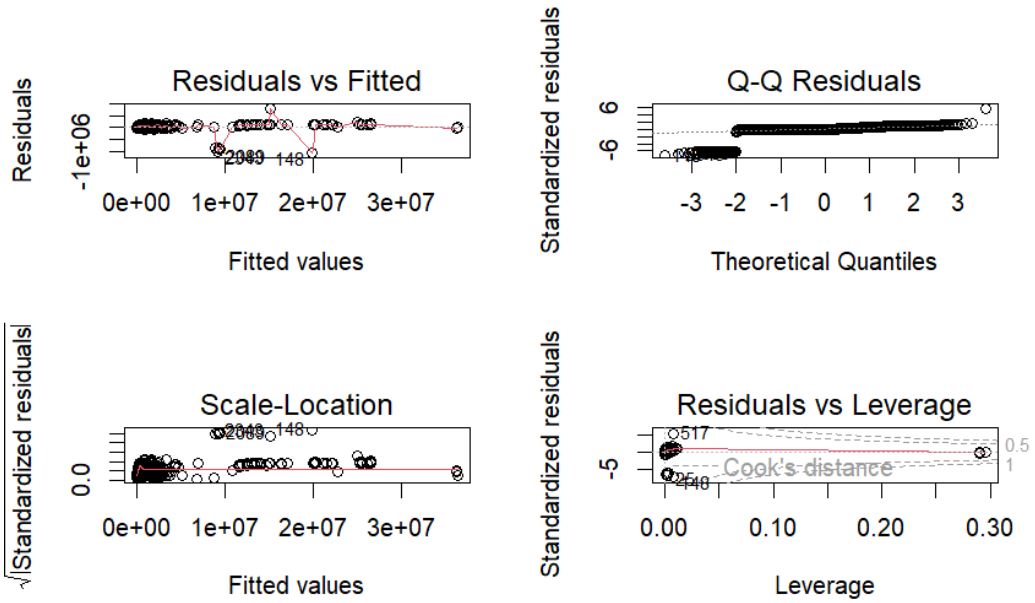
```r
ggplot(df, aes(x = Cost_Type)) +
 geom_bar(fill = "darkblue") +
 labs(title = "Distribution of Warranty Types", x = "Warranty Type",
y = "Count")
```



7. Linear Regression - Cost Prediction

Multiple linear regression shows that cost is strongly related to parts, discount, and labor. The regression analysis showed that parts cost, discounts, and labor charges had a strong and statistically significant relationship with total invoice amount (R^2 = 0.999). Service type did not have a significant effect.

```r
model <- lm(Total_Invoice_Amount ~ Parts_Amount +
Discount_Amount + Labor_Charge_Amount + Service_type, data = df)
summary(model)
par(mfrow = c(2, 2))
plot(model)
```

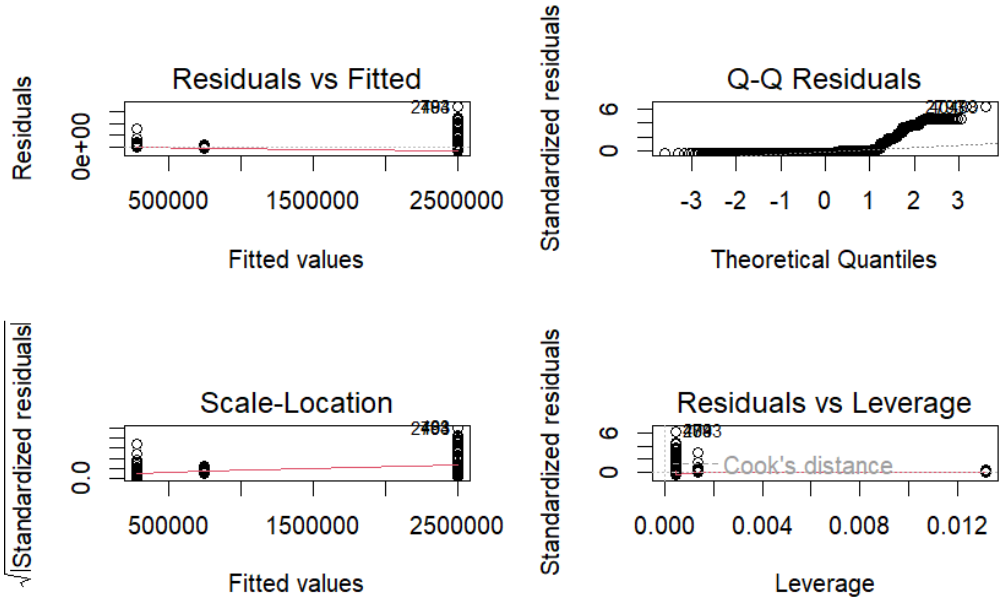


8. Linear Regression - Warranty Type Effect

The model revealed that warranty type only explains 3% of the variation in total invoice amount. Only 'Normal Warranty' was statistically significant among warranty types.

```r
df$Cost_Type <- as.factor(df$Cost_Type)
model2 <- lm(Total_Invoice_Amount ~ Cost_Type, data = df)
```

```
summary(model2)
plot(model2)
```



9. Association Rules - Apriori Algorithm

Apriori analysis uncovered strong patterns. For example, items with 'Extended Warranty' and 'Repair Refusal' were almost always associated with 'HARD DISK DRIVE' (lift > 30). Also, worn-out MP3 players frequently led to 'SOFTWARE UPGRADE'. Such rules can help anticipate issues and optimize repair strategies.

```
r
df_apriori <- df[, c("Cost_Type", "Product_Group", "Defect_Des",
"Action")]
df_apriori[] <- lapply(df_apriori, as.factor)
trans <- as(df_apriori, "transactions")
rules <- apriori(trans, parameter = list(supp = 0.01, conf = 0.5,
minlen = 2))
inspect(sort(rules, by = "lift")[1:10])
```

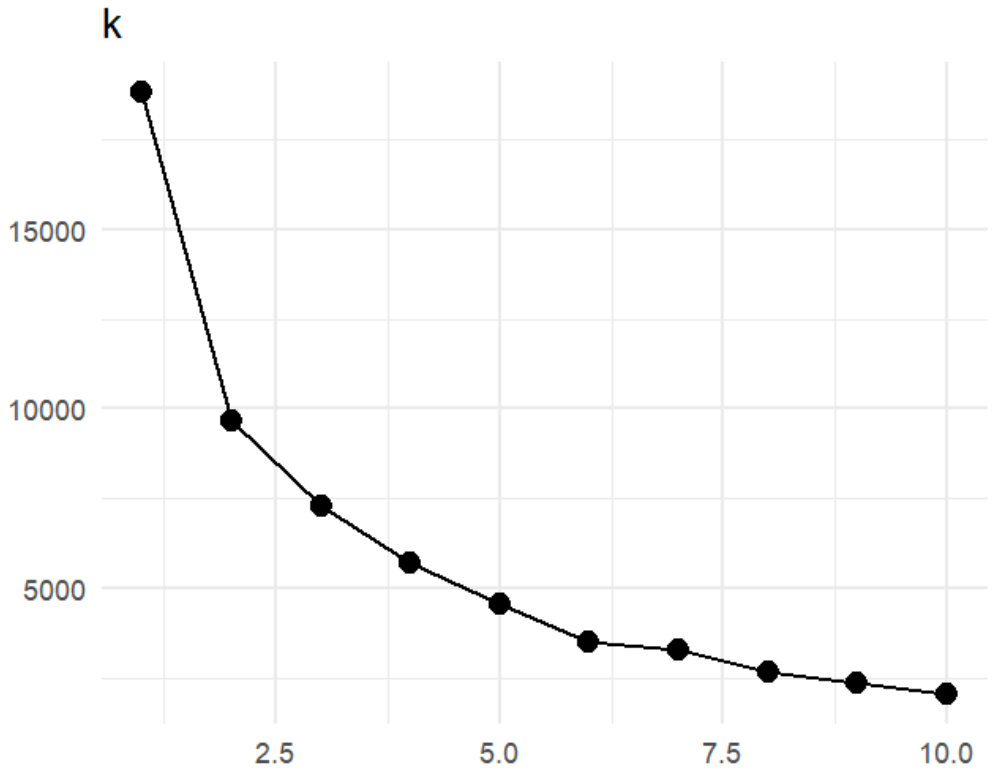
10. Clustering - KMeans

We segment customers using repair and cost features.

10.1 Elbow Method

The optimal number of clusters is found to be 3 using the Elbow method.

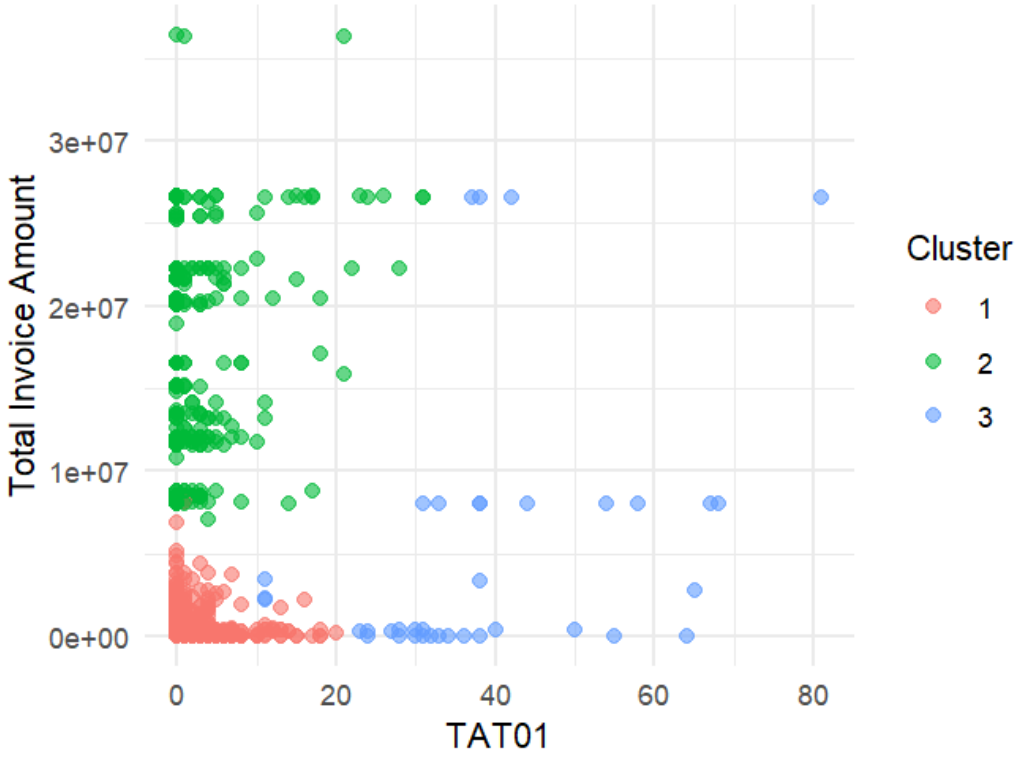
```r
elbow <- sapply(1:10, function(k) {
  kmeans(df_scaled, centers = k, nstart = 10)$tot.withinss
})

ggplot(data = data.frame(k = 1:10, elbow = elbow), aes(x = k, y = elbow)) +
  geom_line() +
  geom_point(shape = 19, size = 3) +
  labs(x = "Number of Clusters", y = "Within-Cluster Sum of Squares", title = "Elbow Method") +
  theme_minimal()
```



10.2 KMeans Clustering Execution

Three clusters are identified and plotted to observe group characteristics.

```r
km <- kmeans(df_scaled, centers = 3, nstart = 25)
df$Cluster <- as.factor(km$cluster)
ggplot(df, aes(x = TAT01, y = Total_Invoice_Amount, color = Cluster))
+
 geom_point(alpha = 0.6, size = 2) +
 labs(title = "Customer Clustering Based on Repair Time and Cost", x
= "TAT01", y = "Total Invoice Amount")
```



Conclusion

This project combined EDA, regression, association mining, and clustering to understand repair services. Insights revealed drivers of cost, frequent issue patterns, and customer segments.