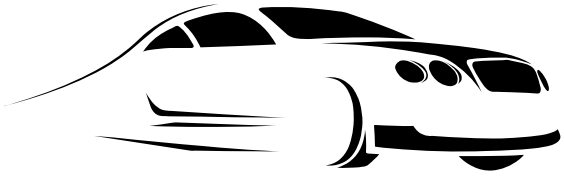

MGSC - 661 FINAL PROJECT REPORT
AUTOMOBILE RISK RATING ANALYSIS

BY

Xiaorong Tian

PART 1: INTRODUCTION

In modern society, it is mandatory for almost all vehicles to secure third-party insurances prior to operating on public roads, providing coverage in case of accidents. For the insurances providers, the most crucial thing for them is to figure out the optimal price for different vehicles. As we all know, the quality, configuration, and safety performance of a vehicle can significantly influence the severity and frequency of the accidents it may involve in, which will eventually affect the amount of insurance compensation. Hence, it is very important to conduct a risk rating analysis to provide necessary information for insurance pricing. The objective of this project is to identify and analyze the key factors that significantly influence the risk rating of a vehicle.

PART 2: DATA DESCRIPTION

2.1. Feature Engineering

Before entering the model building stage, I did some feature engineering to better show the performance and safety performance of a vehicle. I used the existing features in the dataset to create new features, including features that depict the performance of a vehicle, such as power to weight ratio, engine efficiency and total displacement. As for the safety performance of a vehicle, the dataset is lack of information about safety measures (such as ABS, electronic stability control, etc.). Therefore, I used the ratio between a vehicle's length, width, and height to indicate the structural integrity of it. You can find detailed information bellow.

a. Power-to-weight ratio: This ratio is calculated by dividing the horsepower by the curb weight of the vehicle. Generally speaking, a higher power-to-weight ratio indicates faster acceleration and fuel efficiency.
b. Engine efficiency: This feature is calculated by dividing the horsepower by the engine size. This feature gives us an idea of how efficient the engine converts fuel into power.
c. Total displacement: The calculation of total of displacement is a bit more complicated. The formula is:

$$Total\ Displacement = \frac{\pi}{4} \times Bore^{2} \times Stroke \times Number\ of\ Cylinders$$

The total displacement describes the total volume of air/fuel mixture an engine can draw in a complete engine cycle, which indicates the engine power potential and fuel consumption of an engine. A higher total displacement rate means an engine can burn more fuel and produce more power in each cycle.

d. Length to width ratio / length to height ratio / width to height ratio: The length to width ratio gives us ideas of a vehicle's aspect ratio, influencing aerodynamics and handling. Meanwhile, the length to height ratio tells us the information about a vehicle's stability and also aerodynamics. Last, the width to height ratio provides insight into the center of gravity and rollover risk. In conclusion, these three factors give us the general idea of the safety performance of a specific vehicle.

2.2. Exploratory Data Description

Now, to gather additional information essential for constructing the model, a comprehensive data analysis is required. The initial focus is on the 'price' variable, chosen due to its intuitive correlation with a vehicle's quality and design sophistication. Based on my research, there is a huge gap between the prices of vehicles. Most of the vehicles share a price lower than 20,000 dollars, while only several of them have a price higher than 30,000 dollars. You can see the visualization of distribution in image 2.2.1.

Secondly, the 'horsepower' variable is also a key indicator of the performance of the vehicles. As we can see from image 2.2.2, the distribution of horsepower follows a similar pattern to price. Most vehicles have horsepower lower than 125, while a few have horsepower exceeding 150. This distribution suggests that the majority of the vehicles in this dataset are designed for moderate performance and efficiency, catering to everyday use and practicality. In contrast, the smaller group with higher horsepower likely represents performance-oriented models

After getting to know the overall distribution of the price and performance of vehicles in the dataset, now let us dive into the correlation between predictors. In image 2.2.3, you can find the visualization of all the numerical predictors. There are several pairs of them that show interesting patterns. The first one is the close correlation between different metrics that describe the overall dimensions and capacity of the vehicle. For instance, curb weight, which is the total weight of the vehicle, is highly correlated with width, height, length, as well as wheelbase and engine size. This indicates that larger vehicles tend to be heavier and have bigger engines.

Meanwhile, the variables that detailedly describe the performance of a vehicle are also closely related. For example, the total displacement shows a significant negative correlation with city/highway MPG. This relationship is intuitive as higher total displacement generally means more fuel consumption per engine cycle, leading to lower fuel efficiency. At the same time, total displacement is significantly positively correlated with horsepower, suggesting that vehicles with larger engine displacements tend to have higher power output.

By looking at this correlation information, we can understand how different aspects of vehicle design and engineering are interrelated. It also helps in predicting how changes in one feature might impact others, which is crucial for us to find meaningful pattern in predicting risk.

PART 3: MODEL BUILDING & RESULTS

In order to effectively predict the risk rating of a vehicle, I implemented two algorithms. The first one is logistic regression, and the second one is random forest model. I will thoroughly explain how did I build these two models in the following text.

3.1. Logistic Regression

The reason for choosing this model is that it aligns well with the requirements of our data and the research question at hand. The risk rating, which ranges from -2 to 3, can be evenly divided into two categories, high risk and low. Then we can see the general impact of each predictors on these two categories.

Before running the model, we need to eliminate multicollnearity and conduct dimension reduction due to the high number of predictors. The PCA is a perfect we can utilize under this circumstance. You can see the scree plot in image 3.1.1 I used to determine the perfect number of PCA to use. The optimal number of PCA is 3. You can see the visualization of PCA in 3.1.2 as well.

To understand which predictors may lead to a higher or lower risk rating, we can include the most and least risky vehicles in one PCA graph and observe the distinctions. As depicted in image 3.1.3, vehicles classified as low risk typically exhibit greater wheel base, length-to-width ratio, curb weight, and notably, a higher price point. Conversely, vehicles deemed high risk tend to demonstrate a higher power-to-weight ratio, engine efficiency, and horsepower, as well as a greater number of cylinders. From these observations, we can conclude that vehicles with more robust and performance-oriented features, such as increased horsepower and a higher number of cylinders, are often associated with a higher risk rating. In contrast, more substantial and potentially more luxurious vehicles, which often come with a heftier price tag, are likely to be categorized as lower risk. These insights can be instrumental for manufacturers and consumers alike, as they highlight key characteristics that contribute to the risk assessment of a vehicle.

Now it is time to use PCA in our logistic regression. After running, I found the logistic regression model displays robust predictive performance with an accuracy of 90.32%, substantiated by a Kappa statistic of 0.8075, indicating strong agreement beyond chance. Sensitivity stands at 81.25%, reflecting a high true positive rate. The balanced accuracy is 0.9062, suggesting a consistent performance across both classes. These key metrics highlight the model's effectiveness, particularly its capacity to correctly classify negative instances and its precision in positive predictions. You can see the original result in image 3.1.4.

3.2. Random Forest

The random forest model constructed provides a comprehensive classification framework for the prediction of automobile risk ratings (symboling), a crucial factor in automotive safety and insurance underwriting. This model is trained on a substantial dataset with 2000 trees and considers five variables at each decision split, ensuring a robust and diversified decision-making process. You can see the result of this random forest model in image 3.2.1.

Analyzing the out-of-bag (OOB) error rate, which is an estimate of the model performance on unseen data, we find it to be 11.95%. This relatively low OOB error rate is indicative of the model's strong generalization capabilities. The confusion matrix further elucidates the model's performance across different risk categories, ranging from -2 to 3. Here, we see a varied distribution of classification errors, with the model performing exceptionally well for certain risk levels (e.g., -1, 1), while having room for improvement in others (e.g., 2, 3).

The importance of each predictor variable in the random forest model is evaluated based on two criteria: MeanDecreaseAccuracy and MeanDecreaseGini. The variable 'make' stands out with the highest MeanDecreaseAccuracy, suggesting the manufacturer's pivotal role in the prediction of the risk rating. This may be attributed to the brand's reputation and its correlation with safety standards and features. 'Wheel_base', 'length', and 'width' also emerge as significant variables, implying that a vehicle's dimensions are influential in determining its risk classification.

Other notable variables with considerable importance scores include 'price', 'engine_size', and 'curb_weight'. These suggest that more expensive, larger, and heavier vehicles might be associated with lower risk ratings. It’s interesting to note the substantial importance of 'Length_to_Width_Ratio', 'Length_to_Height_Ratio', and 'Width_to_Height_Ratio', highlighting the model's sensitivity to the vehicle's proportions, which could be correlated with its stability and therefore, safety.

PART 4: CONCLUSION

Starting from exploratory data analysis, it revealed several critical insights, particularly the influence of vehicle price and horsepower on risk categorization, which aligns with industry expectations. The observed correlation patterns among various vehicle metrics provided a deeper understanding of vehicle design and its implications on safety and performance.

The logistic regression model demonstrated a high level of predictive accuracy, with an impressive balance across classes, as evidenced by the high Kappa statistic and sensitivity values. This model's results, supported by PCA, have provided a reliable method for classifying vehicle risk. At the same time, the random forest model further complemented these findings,

showing a strong capability to generalize with a low OOB error rate. The importance measures extracted from this model placed significant emphasis on the make, wheelbase, length, and width, indicating these as pivotal factors in risk rating predictions.

Based on my findings, vehicles that are likely to have higher risk ratings often showcase not only a higher power-to-weight ratio, indicative of superior performance and speed, but also larger total displacement, which correlates with greater engine power and potential for aggressive driving. In contrast, vehicles with lower risk ratings typically feature higher engine efficiency, signaling a balance between fuel consumption and power output, and are usually designed with a focus on practicality and safety rather than high performance. Other significant factors influencing risk ratings include the manufacturer of the vehicle, reflecting the brand's safety standards and reputation, and physical dimensions like wheelbase, length, and width, which are linked to the vehicle's stability and structural integrity. These combined elements paint a comprehensive picture of a vehicle's risk profile, with performance-oriented models generally attracting higher risk ratings, while more luxury, utilitarian, efficiency-focused vehicles are often deemed lower risk.

Managerial implications of this research are profound. For insurance providers, understanding the variables that contribute most significantly to risk can inform better pricing strategies, allowing for more tailored and equitable insurance premiums that reflect the true risk profile of different vehicle types.

Appendix

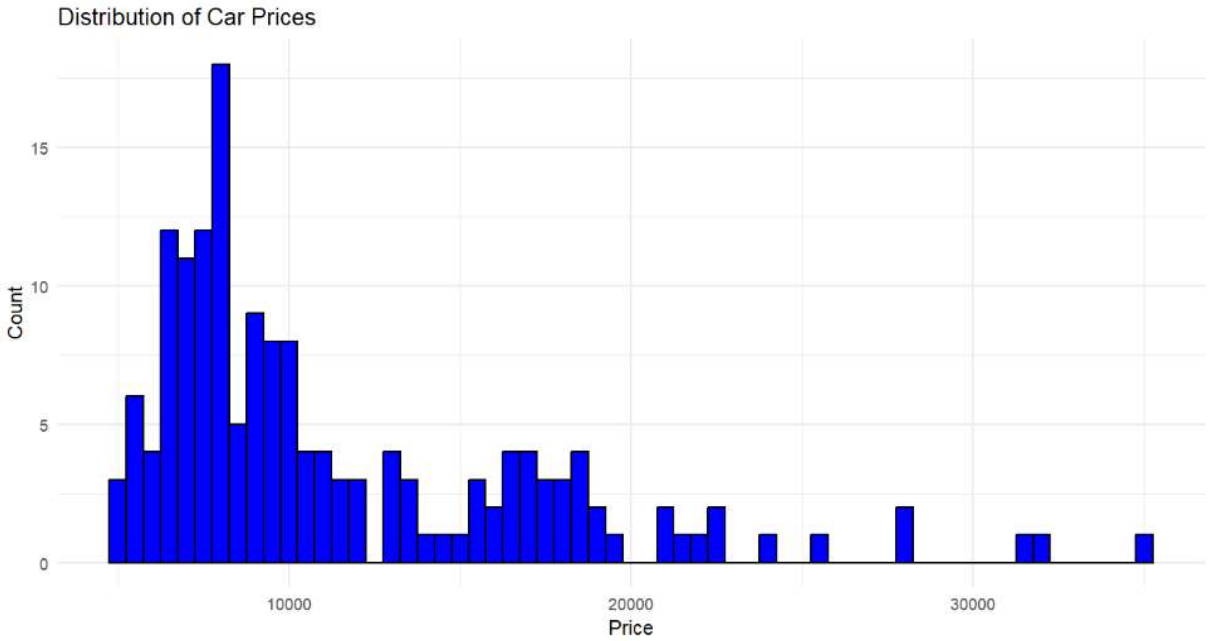

Image 2.2.1

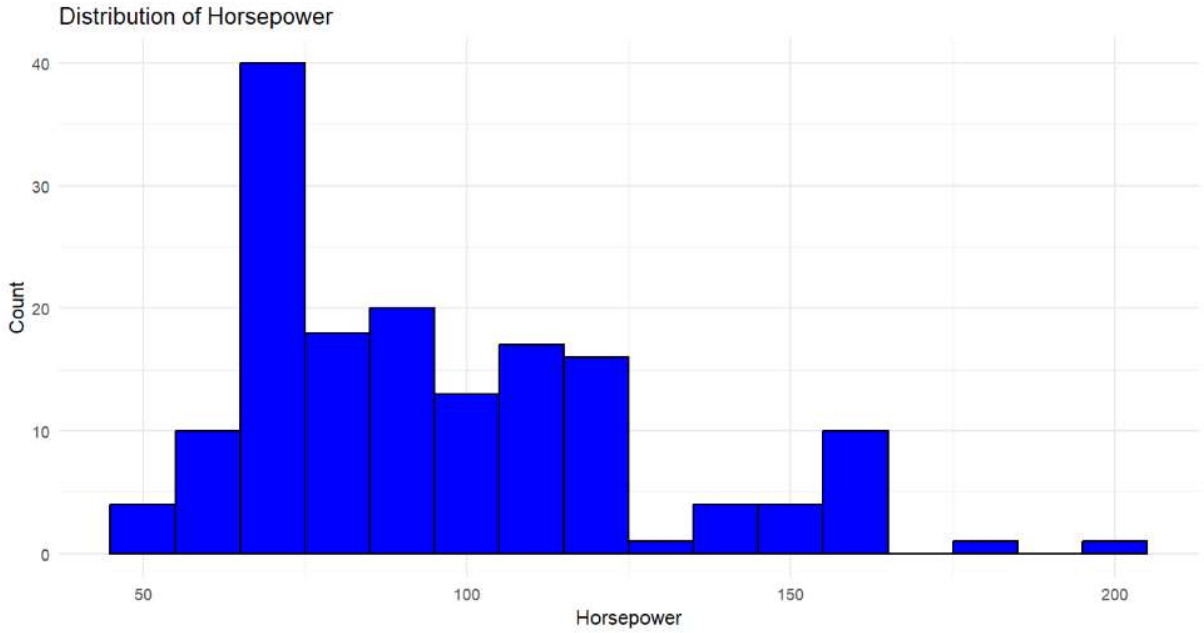

Image 2.2.2

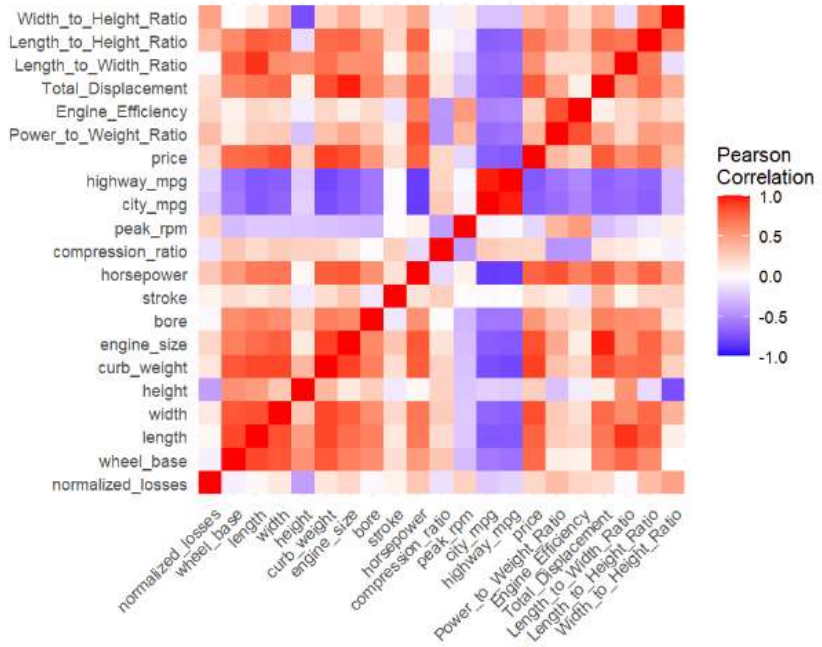

Image 2.2.3

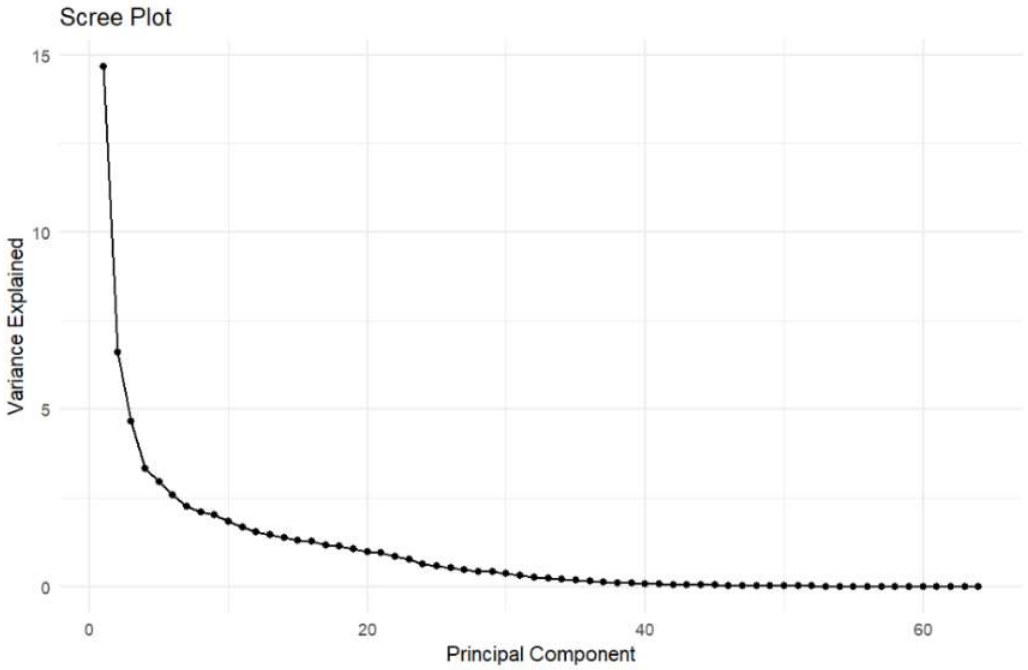

Image 3.1.1

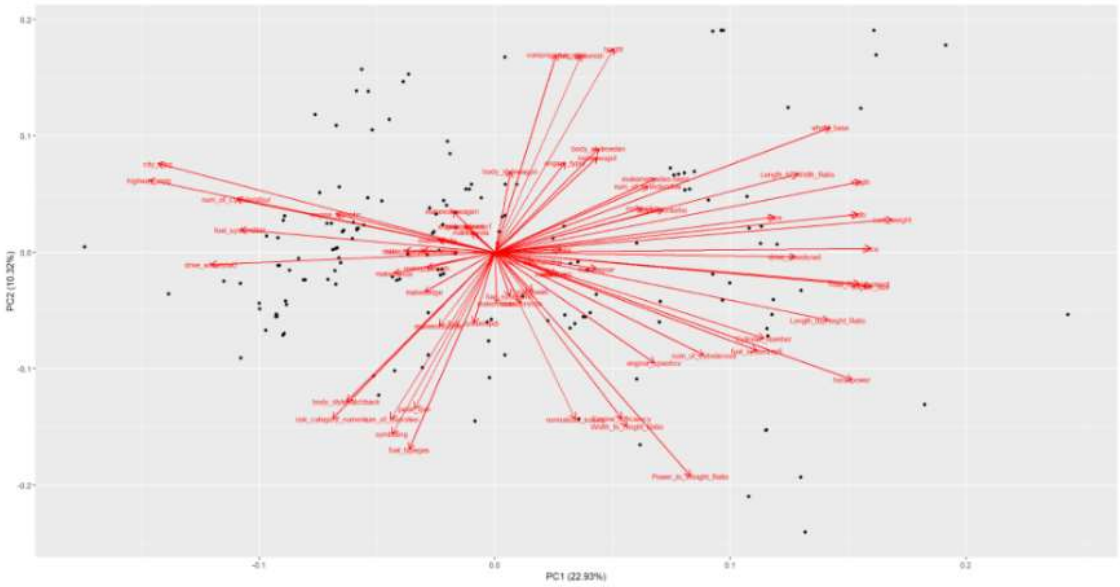

Image 3.1.2

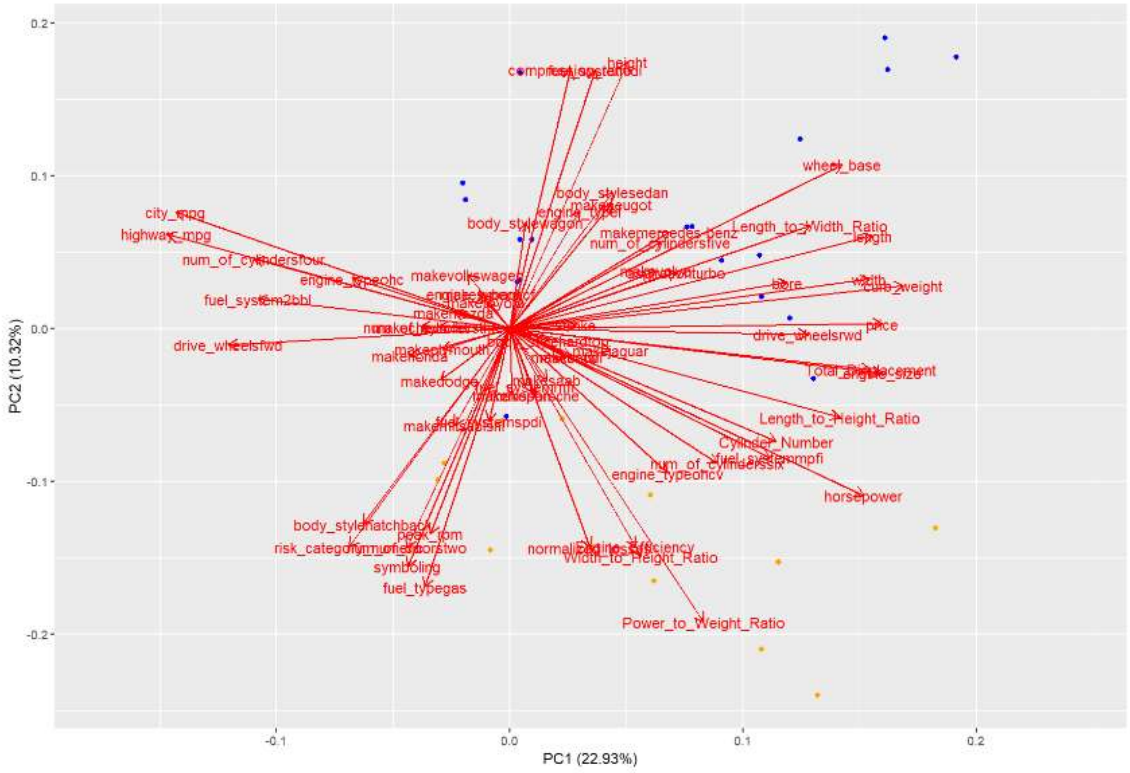

Image 3.1.3 (High risk: orange; Low risk: blue)

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 13  0
         1  3 15

               Accuracy : 0.9032
                 95% CI : (0.7425, 0.9796)
    No Information Rate : 0.5161
    P-Value [Acc > NIR] : 5.161e-06

                  Kappa : 0.8075

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 0.8125
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8333
             Prevalence : 0.5161
         Detection Rate : 0.4194
   Detection Prevalence : 0.4194
      Balanced Accuracy : 0.9062

       'Positive' Class : 0
```

Image 3.1.4

```
                Type of random forest: classification
                      Number of trees: 2000
No. of variables tried at each split: 5

        OOB estimate of  error rate: 11.95%
Confusion matrix:
   -2 -1  0  1  2  3 class.error
-2  2  1  0  0  0  0  0.33333333
-1  0 18  1  1  0  0  0.10000000
0   0  2 43  2  1  0  0.10416667
1   0  1  1 43  0  1  0.06521739
2   0  0  2  4 23  0  0.20689655
3   0  0  0  2  0 11  0.15384615
```

Image 3.2.1

Code

```
# MGSC - 661 Final Project
auto = read.csv('C:/Users/95675/OneDrive/桌面/R/Final Project/Dataset
5 -- Automobile data_Processed.csv')
attach(auto)

auto$make=as.factor(auto$make)
auto$fuel_type=as.factor(auto$fuel_type)
auto$aspiration=as.factor(auto$aspiration)
auto$num_of_doors=as.factor(auto$num_of_doors)
auto$body_style=as.factor(auto$body_style)
auto$drive_wheels=as.factor(auto$drive_wheels)
auto$engine_type=as.factor(auto$engine_type)
auto$fuel_system=as.factor(auto$fuel_system)
auto$num_of_cylinders=as.factor(auto$num_of_cylinders)
# Part 1: Exploratory Data Description
# 1.1. Key indicators distribution
# Price
library(ggplot2)
ggplot(auto, aes(x = price)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Car Prices",
       x = "Price",
       y = "Count")

# Horsepower
# Load ggplot2 package
library(ggplot2)

# Create a histogram of the horsepower variable
ggplot(auto, aes(x = horsepower)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Horsepower",
       x = "Horsepower",
       y = "Count")
```

```r
# 1.2. Correlation between predictors
library(dplyr)
library(ggplot2)
library(reshape2)

selected_auto <- auto %>%
  select(normalized_losses, wheel_base, length, width, height,
curb_weight,
         engine_size, bore, stroke, horsepower,
         compression_ratio, peak_rpm, city_mpg, highway_mpg, price,
         Power_to_Weight_Ratio, Engine_Efficiency,
Total_Displacement,
         Length_to_Width_Ratio, Length_to_Height_Ratio,
Width_to_Height_Ratio)

# Calculate the correlation matrix
cor_matrix <- cor(selected_auto, use = "complete.obs")  # using
'complete.obs' to handle missing values

# Melt the correlation matrix for ggplot
melted_cor_matrix <- melt(cor_matrix)

# Plotting the matrix graph
ggplot(data = melted_cor_matrix, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title = element_blank()) +
  coord_fixed()
print(cor_matrix)



# Part 1: Logistic Regression
library(dplyr)
library(caret)
```

```r
library(dplyr)

# Convert categorical variables to numeric using one-hot encoding
auto_categorical = model.matrix(~ make + fuel_type + aspiration +
num_of_doors + body_style + drive_wheels + engine_type + fuel_system
+ num_of_cylinders - 1, data = auto)
auto_numeric = auto[, sapply(auto, is.numeric)]
# Standardize the data
auto_combined = cbind(auto_numeric, auto_categorical)

# Standardize the combined data
auto_scaled = scale(auto_combined)
# Perform PCA
pca_result = prcomp(auto_scaled, center = TRUE, scale. = TRUE)

# Summary of PCA
summary(pca_result)
single_level_factors = sapply(auto, function(x) is.factor(x) &&
length(levels(x)) < 2)

# Print out the names of these variables
cat("Single-level factors:",
names(single_level_factors[single_level_factors]), "\n")


# Scree plot to help decide the number of components to retain
scree_plot = data.frame(Comp = 1:length(pca_result$sdev), Var =
pca_result$sdev^2)
ggplot(scree_plot, aes(x = Comp, y = Var)) +
  geom_line() +
  geom_point() +
  theme_minimal() +
  labs(title = "Scree Plot", x = "Principal Component", y = "Variance
Explained")

# Load the ggfortify package
library(ggfortify)
autoplot(pca_result, data = auto, loadings = TRUE, loadings.label =
TRUE, loadings.label.size = 3)
```

```
colors = ifelse(auto$symboling == 3, "orange", ifelse(auto$symboling
== -1, "blue", "transparent"))

# Create PCA plot with points colored based on symboling values
p = autoplot(pca_result, data = auto, loadings = TRUE,
             col = colors, loadings.label = TRUE) +
  scale_colour_manual(values=c("orange", "blue", "transparent")) +
  theme(legend.position="top") # To hide legend if not needed

# Print the plot
print(p)
# Extract the first three principal components
pca_scores = pca_result$x[, 1:3]

# Create a new data frame with the PCA scores
auto_pca = data.frame(auto, PCA1 = pca_scores[,1], PCA2 =
pca_scores[,2], PCA3 = pca_scores[,3])

# Create the new binary variable for symboling
auto_pca$risk_category = ifelse(auto_pca$symboling >= 1 &
auto_pca$symboling <= 3, "low_risk", "high_risk")

# Convert the new variable to a binary numeric format
auto_pca$risk_category_numeric = ifelse(auto_pca$risk_category ==
"low_risk", 1, 0)

# Splitting the dataset into training and testing sets
set.seed(123)  # For reproducibility
indexes = createDataPartition(auto_pca$risk_category_numeric, p =
0.8, list = FALSE)
train_data = auto_pca[indexes, ]
test_data = auto_pca[-indexes, ]

# Building the logistic regression model using the first three
principal components
logit_model = glm(risk_category_numeric ~ PCA1 + PCA2 + PCA3, data =
train_data, family = binomial())
```

```
# Summary of the model
model_summary = summary(logit_model)

# Print the summary to the console
print(model_summary)

# Perform ANOVA to test the significance of the predictors
anova_logit_model = anova(logit_model, test="Chisq")

# Print the ANOVA results to the console
print(anova_logit_model)

# Predicting on test data
predictions = predict(logit_model, newdata = test_data, type =
"response")
predicted_class = ifelse(predictions > 0.5, 1, 0)

# Load the caret package for confusion Matrix
library(caret)

# Evaluating model performance
confusionMatrix(factor(predicted_class),
factor(test_data$risk_category_numeric))


# Part 2: Random Forest with initial data and processed data with new
columns
auto$symboling=as.factor(auto$symboling)
library(randomForest)
myforest2=randomForest(symboling~normalized_losses+make+fuel_type+asp
iration+num_of_doors+body_style+drive_wheels+wheel_base+length+width+
height+curb_weight+engine_type+num_of_cylinders+engine_size+fuel_syst
em+bore+stroke+compression_ratio+horsepower+peak_rpm+city_mpg+highway
_mpg+price+Power_to_Weight_Ratio+Engine_Efficiency+Total_Displacement
+Length_to_Width_Ratio+Length_to_Height_Ratio+Width_to_Height_Ratio,
ntree=2000, data=auto, importance=TRUE,  na.action = na.omit)
myforest2
importance(myforest2)
```