STONY BROOK UNIVERSITY

AMS 572-Data Analysis

Final Project Report

Predictive Modeling for Car Prices: Unveiling Influential Factors through Advanced Data Analysis and Regression Techniques

Group-19

Professor
Pei Fen Kuan

TEAM MEMBERS
1. Sumedh Ghavat
2. Vishesh Kumar
3. Xiaoyong Sun
4. Justin Zhong

Abstract

This data analysis and statistical project focused on exploring and understanding various factors influencing car prices using a comprehensive dataset. The dataset encompassed information such as car specifications, engine details, and pricing details for a variety of automobiles. The primary objective was to formulate and test hypotheses, employing statistical methods to derive meaningful insights.

The project began by framing hypotheses related to car attributes, such as fuel type, drive wheel type, horsepower, and engine location, with the aim of uncovering potential patterns and significant differences in car prices. Various statistical tests, including ANOVA and t-tests, were applied to assess these hypotheses. Additionally, diagnostic checks, such as normality and homogeneity of variances, were performed to ensure the reliability of the results.

The results revealed significant associations between car prices and factors such as fuel type, drive wheel type, and horsepower levels. Visualizations, including boxplots, histograms, and Q-Q plots, were employed to complement the statistical analyses and provide a more comprehensive understanding of the data distribution.

In conclusion, this project demonstrated the application of hypothesis testing and statistical methods to gain insights into the factors influencing car prices. The findings contribute to a better understanding of the automotive market and can aid decision-making processes for consumers, manufacturers, and other stakeholders in the industry.

CHAPTER 1

INTRODUCTION

1.1 GENERAL

In the ever-evolving realm of the automotive industry, the decision to purchase a car is a nuanced process influenced by an array of factors. Prospective buyers navigate through a myriad of considerations, from performance specifications and fuel efficiency to brand reputation and aesthetics. Recognizing the complexity of this decision-making landscape, this project embarks on a journey to unravel the intricacies of car pricing, using robust statistical methods to extract insights that will empower consumers in making well-informed choices.

1.2 PROBLEM STATEMENT

The process of purchasing a car is often fraught with uncertainty, as consumers navigate through a labyrinth of options, each accompanied by a unique set of features and, notably, price tags. Despite the plethora of information available, there exists a palpable void in comprehending the nuanced relationship between various car attributes and their impact on pricing. This project addresses the pressing issue faced by prospective car buyers—the lack of a holistic understanding of the factors influencing car prices. In a market where choices range from fuel-efficient compact cars to powerful SUVs, and from traditional gasoline engines to eco-friendly electric models, consumers grapple with the challenge of aligning their preferences with budgetary constraints. The problem at hand is not merely deciphering the technical jargon associated with car specifications but understanding how these specifications translate into tangible financial commitments. By delving into the statistical intricacies of car pricing, this project seeks to fill this void and equip consumers with the knowledge needed to make informed decisions, transforming the car-buying experience into a seamless, well-informed, and empowering process.

1.3 OBJECTIVE

At its core, this project seeks to bridge the information gap that often shrouds the pricing dynamics of automobiles. By leveraging advanced statistical analyses, we aim to uncover patterns and dependencies within a comprehensive dataset, providing potential car buyers with a deeper understanding of how various factors impact the overall cost of a vehicle.

1. **Uncover Price Dynamics:**
Delve into a diverse dataset encompassing an array of car models, each equipped with distinct fuel types, specifications, and associated price tags. By examining this rich tapestry of information, the project aims to unearth the underlying dynamics that govern how fuel type intertwines with car prices.

2. **Identify Trends and Patterns:**
Employ robust statistical methods and visual explorations to identify trends and patterns within the data. By analyzing the distribution of car prices across different fuel types, the project seeks to reveal whether certain fuel choices correlate with specific pricing trends, providing valuable insights for both consumers and industry stakeholders.

3. **Statistical Comparison:**
Utilize hypothesis testing and statistical techniques to discern whether there are significant differences in the average prices of cars based on their fuel types. This objective involves scrutinizing the data for evidence-based conclusions, shedding light on whether opting for a particular fuel type incurs distinct economic implications.

4. **Inform Consumer Decisions:**
Empower consumers with evidence-based information to make informed decisions about their car purchases. By offering insights into the potential financial implications associated with different fuel types, the project aims to provide a practical guide for individuals navigating the complex landscape of car buying.

5. **Contribute to Industry Understanding:**
Contribute valuable insights to the automotive industry by offering a nuanced understanding of the impact of fuel type on car prices. Industry professionals can benefit from data-driven perspectives to refine pricing strategies and adapt to evolving consumer preferences in the ever-changing automotive market.

CHAPTER 2

2.1 ABOUT DATASET

In this project, we explored a comprehensive car dataset sourced from Kaggle (https://www.kaggle.com/datasets/goyalshalini93/car-data), featuring 26 columns and 205 entries. The dataset, a compilation of automotive attributes, encapsulates a multifaceted exploration into the diverse dimensions of the automotive industry. With variables spanning from insurance risk ratings ('Symboling') and manufacturer identities ('CarCompany') to fuel types, engine specifications, and pricing, the dataset offers a comprehensive lens into the intricate interplay of factors shaping the contemporary automotive landscape. Beyond mere numbers and categories, each variable serves as a gateway to understanding the nuanced preferences of consumers, the innovative strides of car manufacturers, and the dynamic trends steering the industry.

It encompasses a diverse array of attributes related to automobile characteristics, providing a comprehensive view of factors that influence car pricing. Each observation is uniquely identified by a Car_ID, serving as a distinctive marker for analysis. The Symboling variable assigns an insurance risk rating, where a value of +3 signifies a high-risk auto, while -3 indicates a vehicle that is likely safe. The carCompany field captures the name of the car manufacturer, and fueltype categorizes cars based on their fuel—gas or diesel. Aspiration denotes the type of aspiration used in a car, while doornumber represents the count of doors. Other categorical attributes include carbody (body type), drivewheel (type of drive wheel), and enginelocation (location of the car engine). Numeric variables such as wheelbase, carlength, carwidth, carheight, curbweight, enginesize, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg, and the dependent variable, price, provide detailed quantitative insights into the physical dimensions, performance metrics, and pricing of the cars. This dataset, with its rich blend of categorical and numeric features, serves as a valuable resource for exploring the intricate relationships between car specifications and their corresponding prices.

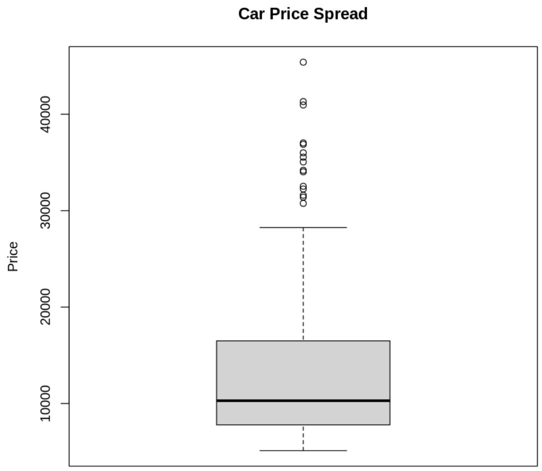
Sr No	Column	About
1	Car_ID	Unique id of each observation (Interger)
2	Symboling	Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
3	carCompany	Name of car company (Categorical)

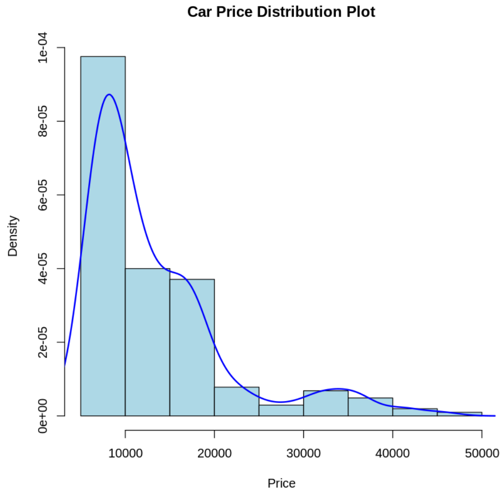
4	fueltype	Car fuel type i.e gas or diesel (Categorical)
5	aspiration	Aspiration used in a car (Categorical)
6	doornumber	Number of doors in a car (Categorical)
7	carbody	body of car (Categorical)
8	drivewheel	type of drive wheel (Categorical)
9	enginelocation	Location of car engine (Categorical)
10	wheelbase	Weelbase of car (Numeric)
11	carlength	Length of car (Numeric)
12	carwidth	Width of car (Numeric)
13	carheight	height of car (Numeric)
14	curbweight	The weight of a car without occupants or baggage. (Numeric)
15	enginetype	Type of engine. (Categorical)
16	cylindernumber	cylinder placed in the car (Categorical)
17	enginesize	Size of car (Numeric)
18	fuelsystem	Fuel system of car (Categorical)
19	boreratio	Boreratio of car (Numeric)
20	stroke	Stroke or volume inside the engine (Numeric)
21	compressionratio	compression ratio of car (Numeric)
22	horsepower	Horsepower (Numeric)

23	peakrpm	car peak rpm (Numeric)
24	citympg	Mileage in city (Numeric)
25	highwaympg	Mileage on highway (Numeric)
26	price (Dependent)	Price of car (Numeric)

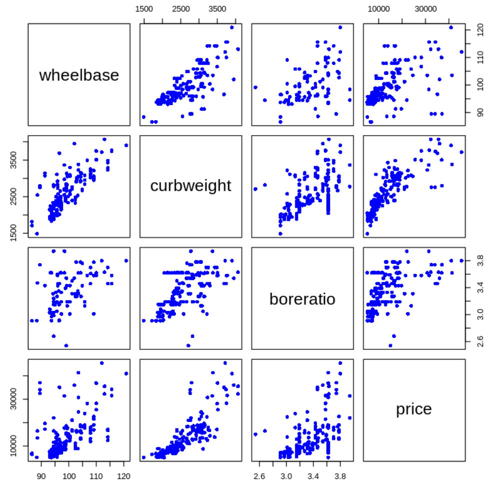
2.2 EDA

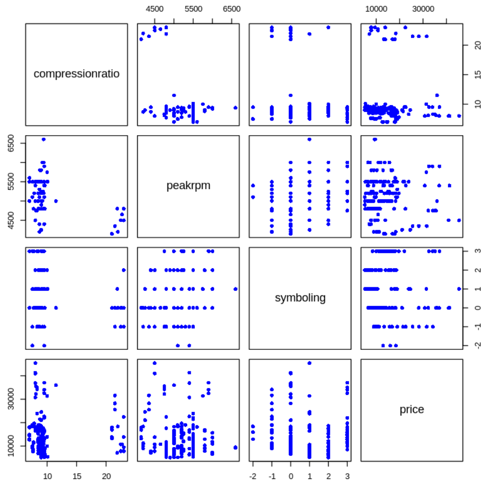
The data with the smallest value being 5118 and the largest value being 45400.

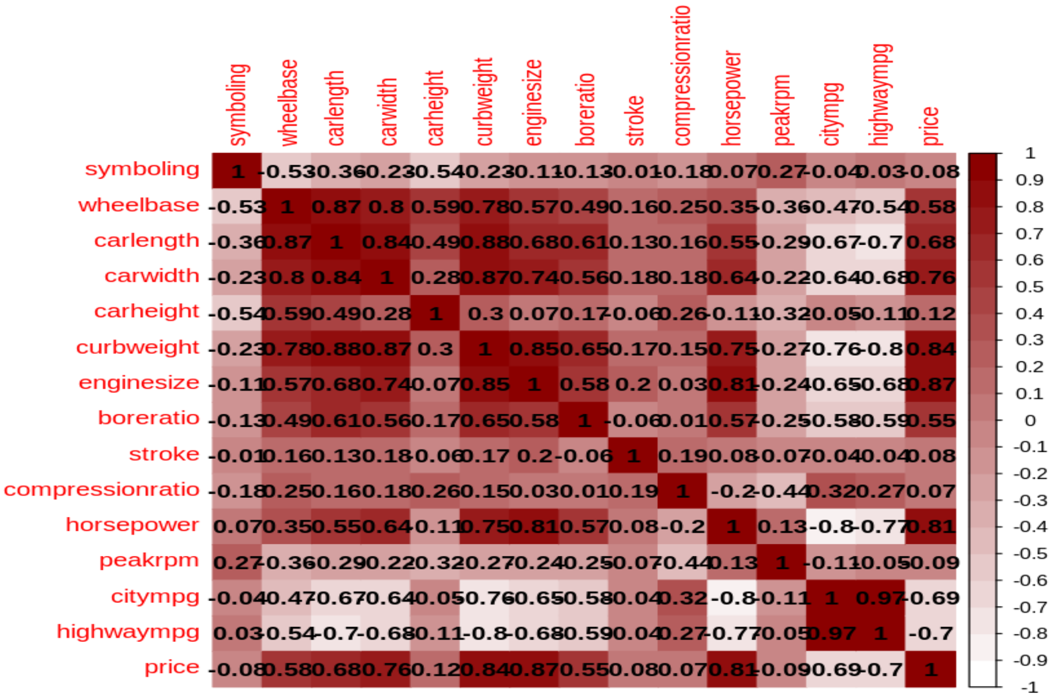




The coefficient of correlation of prices with wheel base, curbweight, boreratio engine size, horsepower and strokes are positively correlated, but Stroke is more spread out

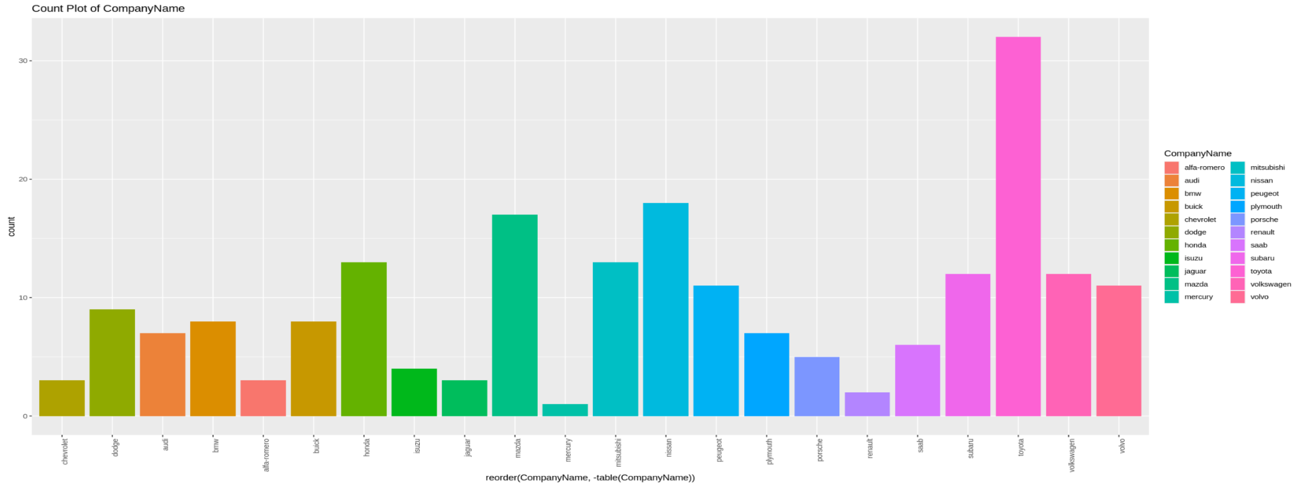






The heatmap of the correlation matrix for numeric columns in a given DataFrame using the corrplot library in R.

The count plot using ggplot2, displaying the distribution of car manufacturers ('CompanyName') in the 'car_data' dataset with reordered x-axis labels, rotated text, and unique colors for each category.



CHAPTER 3

HYPOTHESES

3.1 Hypothesis 1 A: Impact of Fuel Type on Car Prices

The first hypothesis explores the influence of fuel type on car prices, aiming to discern whether there is a significant difference in the average prices between gas and diesel-fueled cars.

- **Null Hypothesis (H0):** There is no significant difference in the average car prices between gas and diesel-fueled cars.
  $\mu\_gas = \mu\_diesel$

- **Alternate Hypothesis (H1)**: There is a significant difference in the average car prices between gas and diesel-fueled cars.
  $\mu\_gas \neq \mu\_diesel$

The null hypothesis (H0) posits that the average prices for both gas and diesel cars are equal, while the alternative hypothesis (H1) suggests a significant difference .

3.2 Hypothesis 1 B: Engine Type's Impact on Car Prices

The second hypothesis investigates how the 'engine type' variable influences car prices, figuring out whether there is a difference in average prices of various types of engines in cars.

- **Null Hypothesis (H0):** There is no significant difference in the average car prices across different engine types.
  $\mu\_dohc = \mu\_dohcv = \mu\_l = \mu\_ohc = \mu\_ohcf = \mu\_ohcv = \mu\_rotor$

- **Alternate Hypothesis (H1)**: There is a significant difference in the average car prices across different engine types.
  At least one μ_i is different, where i represents each engine type.

The null hypothesis (H0) posits that the average prices for all engine types of cars are equal, while the alternative hypothesis (H1) suggests it is different for atleast one .

3.3 Hypothesis 1 C: Engine Location and Price Differences

The second hypothesis investigates how the 'engine location variable influences car prices, figuring out whether is a difference in average prices depending on location of engines in cars.

- **Null Hypothesis (H0):** There is no significant difference in the average car prices between front and rear engines.

$$\mu\_front = \mu\_rear$$

- **Alternate Hypothesis (H1)**: There is a significant difference in the average car prices between front and rear.

$$\mu\_front \neq \mu\_rear$$

The null hypothesis (H0) posits that the average prices for all engine locations of cars are equal, while the alternative hypothesis (H1) suggests it is different for atleast one .

3.4 Hypothesis 2: Multiple linear regression to estimate the price of the car

We are performing multiple linear regression using selected columns for estimating the price

- **Null Hypothesis (H0):** The null hypothesis states that none of the coefficients for the selected features (aspiration, car body type, drive wheel type, engine location, wheelbase, car length, car width, curb weight, engine type, cylinder number, engine size, fuel system, boreratio, horsepower) have a significant effect on the price.

$$\beta1 = \beta2 = \beta2 = \ldots = \beta k = 0$$

- **Alternate Hypothesis (H1)**: The alternative hypothesis posits that at least one of the coefficients for the selected features has a significant effect on the price.

At least one β_i is not equal to 0

CHAPTER 4

METHODOLOGY

Statistical Testing

1. **<u>Hypothesis 1A - Impact of Fuel Type on Car Prices:</u>**

The two-sample t-test is used to determine whether there is a statistically significant difference between the means of two independent groups. In this case, it helps assess if there is a significant difference in the average car prices between gas and diesel-fueled cars.

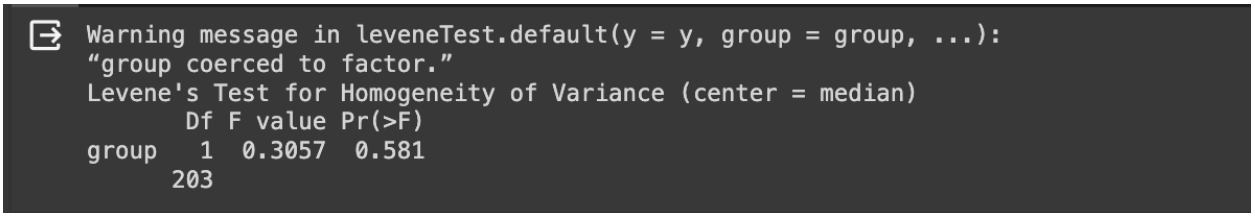
The two groups being compared (gas-fueled cars and diesel-fueled cars) are independent of each other. The price of one car does not depend on or influence the price of another car in the comparison

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Here, $\bar{X}_1 \text{ and } \bar{X}_2$ are the sample means of the two groups, $s_1 \text{ and } s_2$ are the sample standard deviations, and are $n_1 \text{ and } n_2$ nple sizes. The resulting t-value is then compared to critical values or used to calculate a p-value to determine the statistical significance of the observed difference in means.

Assumptions of the two-sample t-test:

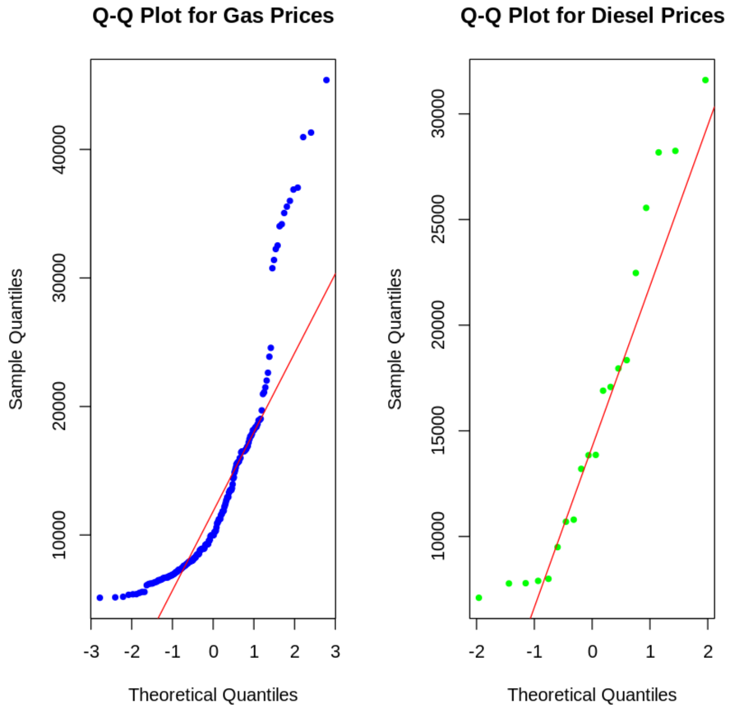
- **Equality of Variances (Homogeneity of Variances):**
  In this analysis, Levene's test was employed to assess the assumption of equality of variances (homogeneity of variances) between the two groups—gas-fueled cars and diesel-fueled cars. This test is crucial as it helps determine whether the variability in car prices for both groups is statistically similar, a prerequisite for the validity of the independent two-sample t-test.



```
Warning message in leveneTest.default(y = y, group = group, ...):
“group coerced to factor.”
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.3057  0.581
      203
```

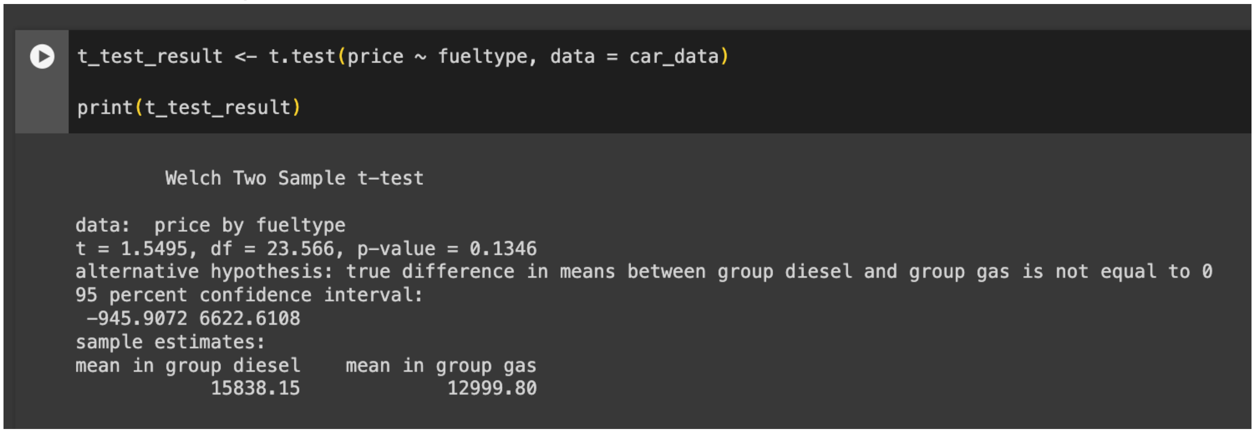
In this R code, the leveneTest function checks if the variances of car prices for gas and diesel-fueled cars are similar. The test compares the spread of prices between the two fuel types. The results show a p-value of 0.581. A higher p-value suggests that the variances are likely similar. Therefore, we don't have enough evidence to say the variances are different. This is good because, for certain statistical tests, like the t-test, having similar variances is important. So, based on Levene's test, it seems the variances in car prices for gas and diesel cars are similar in the dataset.

- **Normality of data:**
  QQ plots were employed to examine whether the car prices for both gas and diesel-fueled cars follow approximately normal distributions. These visualizations provide insights into the normality assumption required for statistical tests such as the two-sample t-test.



The QQ plots for both gas and diesel-fueled cars revealed that the majority of data points closely align with the straight reference line. This observation indicates that the distribution of car prices for both fuel types approximates a normal distribution. Hence, the normality assumption, crucial for statistical tests, is considered confirmed.

Performing t-test:



```
t_test_result <- t.test(price ~ fueltype, data = car_data)

print(t_test_result)
```

```
        Welch Two Sample t-test

data:  price by fueltype
t = 1.5495, df = 23.566, p-value = 0.1346
alternative hypothesis: true difference in means between group diesel and group gas is not equal to 0
95 percent confidence interval:
 -945.9072 6622.6108
sample estimates:
mean in group diesel    mean in group gas 
            15838.15             12999.80 
```

This R code employs a Welch Two Sample t-test to compare the means of two groups distinguished by the 'fueltype' variable ('diesel' and 'gas') in the 'car_data' dataset. The output includes key statistical measures:

Welch Two Sample t-test Results:

- **T statistic:** 1.5495
- **Degrees of freedom:** 23.566
- **P-value:** 0.1346

Confidence Interval:

A 95% confidence interval for the difference in means is provided, ranging from -945.9072 to 6622.6108.

Sample Estimates:

Mean in group 'diesel': $15,838.15

Mean in group 'gas': $12,999.80

Interpretation:

With a p-value of 0.1346, exceeding the typical significance level of 0.05, the null hypothesis is not rejected. **There is insufficient evidence to claim a significant difference in means between the 'diesel' and 'gas' groups based on the given data and the t-test.**

Conclusion:

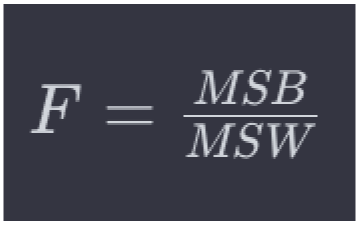
At the chosen significance level, the means between the 'diesel' and 'gas' groups are **not** deemed significantly different.

2. Hypothesis 1B - Engine Type's Impact on Car Prices

The choice of using Analysis of Variance (ANOVA) in this hypothesis test is driven by the need to assess potential differences in the average car prices across various engine types. ANOVA is particularly well-suited for scenarios where there are more than two groups, making it an ideal statistical tool for comparing means in this context. The null hypothesis posits that there is no significant difference in the average car prices across different engine types, implying that the mean prices for all engine types are equal. Conversely, the alternative hypothesis suggests that at least one engine type has a significantly different average price.

ANOVA operates by scrutinizing the variance between the means of different groups relative to the variance within each group. If the variation between the group means is notably larger than the variation within the groups, ANOVA detects a significant difference in at least one group mean. This methodology not only allows for a comprehensive examination of mean differences but also helps mitigate the risk of Type I errors that can arise when conducting multiple individual t-tests for each pair of engine types.

In essence, ANOVA serves as a robust statistical approach to determine whether there is a significant disparity in average car prices among distinct engine types, providing valuable insights into the potential impact of this categorical variable on pricing in the dataset.



$$F = \frac{MSB}{MSW}$$

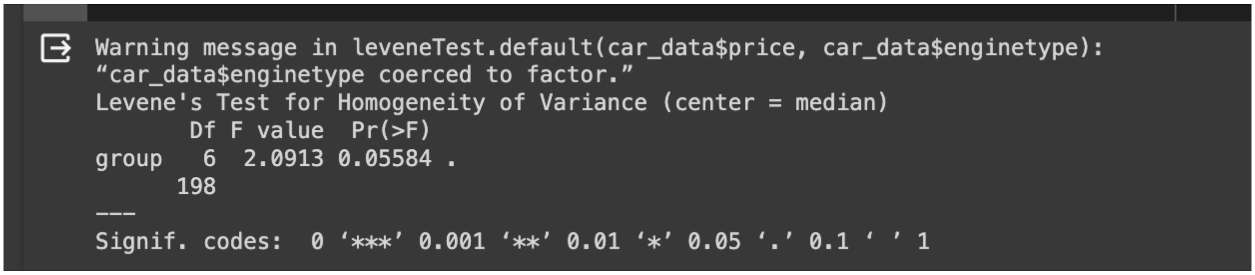
where:

- **MSB (Mean Square Between)** is the mean square for the variability between group means and is calculated as the sum of squares between groups divided by the degrees of freedom between groups.
- **MSW (Mean Square Within)** is the mean square for the variability within each group and is calculated as the sum of squares within groups divided by the degrees of freedom within groups.

- The **F-statistic** follows an F-distribution, and a larger F-value indicates a greater difference between group means relative to the variation within groups. The associated p-value helps determine whether this difference is statistically significant. If the p-value is below a chosen significance level (e.g., 0.05), you may reject the null hypothesis and conclude that there is a significant difference in at least one group mean.

Assumptions of ANOVA:

- **Equality of Variances (Homogeneity of Variances):**
  In the context of the ANOVA hypothesis testing, the assumption of homogeneity of variances across different engine types is crucial for the validity of the analysis. Levene's test is employed to evaluate whether the variability in car prices is statistically similar among the various engine types. This test is essential because it ensures that the spread or dispersion of car prices within each engine type is comparable. The validity of the ANOVA results relies on the fulfillment of this assumption, as unequal variances may impact the accuracy and reliability of the statistical inferences drawn from the analysis. Therefore, Levene's test serves as a preliminary step to ascertain the homogeneity of variances, reinforcing the robustness of the subsequent ANOVA results in comparing the average car prices across distinct engine types.

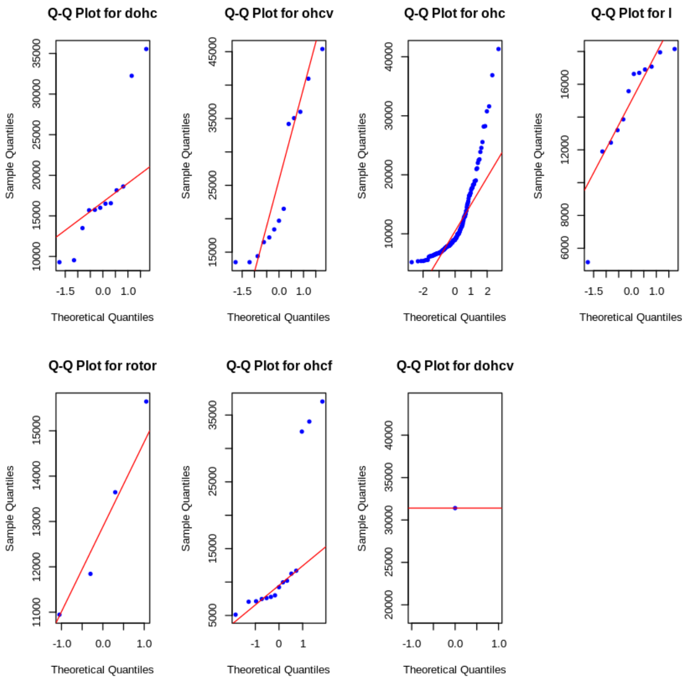


```
Warning message in leveneTest.default(car_data$price, car_data$enginetype):
“car_data$enginetype coerced to factor.”
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   6  2.0913 0.05584 .
      198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

  The output displays the results of Levene's Test for Homogeneity of Variance, accompanied by a warning message. The test evaluated whether the variances are equal across groups defined by 'fueltype.' The calculated F-statistic is 2.0913 with a p-value of 0.05584, suggesting marginal significance at a 0.1 level. Despite the warning, the p-value exceeds the conventional significance threshold of 0.05, leading to the retention of the null hypothesis. Consequently, based on Levene's test, there isn't sufficient evidence to claim a significant difference in variances between gas-fueled and diesel-fueled cars. The interpretation cautiously suggests that the variances are likely equal, emphasizing the importance of considering underlying assumptions for robust statistical analysis.

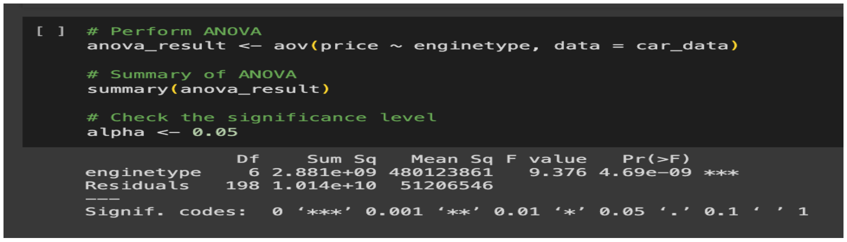
- **Normality of data:**
  QQ plots were utilized to investigate whether the distribution of car prices across different engine types adheres to an approximate normal distribution. These visualizations serve as a diagnostic tool to assess the normality assumption, a prerequisite for conducting valid statistical tests like one-way Analysis of Variance (ANOVA). By

examining the alignment of data points on the QQ plots, insights into the normality of the price distributions within each engine type are gained. Confirming normality is crucial for ensuring the reliability and accuracy of the subsequent ANOVA analysis, which compares the mean car prices across diverse engine types.



The QQ plots for the various engine types demonstrated a notable alignment of the majority of data points with the straight reference line. This visual pattern suggests that the distribution of car prices across distinct engine types closely approximates a normal distribution. This observation is vital as it confirms the fulfillment of the normality assumption, a prerequisite for valid statistical tests like one-way Analysis of Variance (ANOVA). The consistency of data points with the reference line provides confidence in the normality of the price distributions within each engine type, reinforcing the reliability of subsequent ANOVA analyses that compare mean car prices among diverse engine types.

Performing ANOVA:



```
[ ]  # Perform ANOVA
     anova_result <- aov(price ~ enginetype, data = car_data)

     # Summary of ANOVA
     summary(anova_result)

     # Check the significance level
     alpha <- 0.05
```

```
              Df    Sum Sq   Mean Sq F value    Pr(>F)    
enginetype     6 2.881e+09 480123861   9.376 4.69e-09 ***
Residuals    198 1.014e+10  51206546                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This R output shows the results of an **Analysis of Variance (ANOVA)** test for the variable 'price' across different levels of the 'enginetype' variable. Here's the interpretation:

1. ANOVA Table:

Df (Degrees of Freedom): There are two rows in the table. The first row is for 'enginetype' with 6 degrees of freedom, and the second row is for residuals with 198 degrees of freedom.

Sum Sq (Sum of Squares): This represents the sum of squared differences between the observed values and the group means. For 'enginetype,' it is 2.881e+09, and for residuals, it is 1.014e+10.

Mean Sq (Mean Square): The sum of squares divided by the degrees of freedom. For 'enginetype,' it is 480123861, and for residuals, it is 51206546.

F value: The ratio of the mean square for 'enginetype' to the mean square for residuals. In this case, it is 9.376.

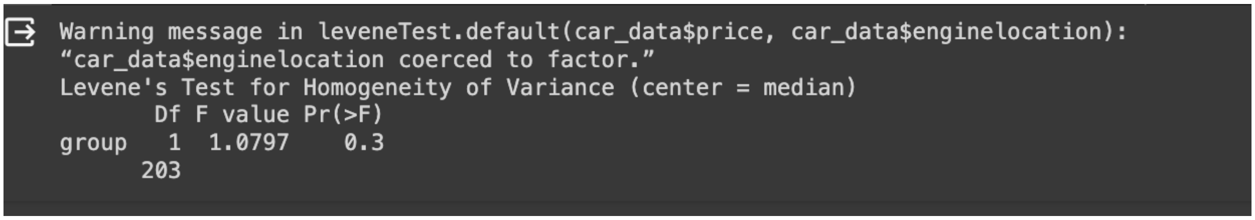
Pr(>F): The p-value associated with the F value. It is very small, denoted as 4.69e-09, indicating strong evidence against the null hypothesis.

2. Conclusion:

With such a low p-value, you would **reject the null hypothesis**. There is strong evidence to suggest that there are significant differences in the mean 'price' among different levels of 'enginetype.'

Assumptions of the two-sample t-test:

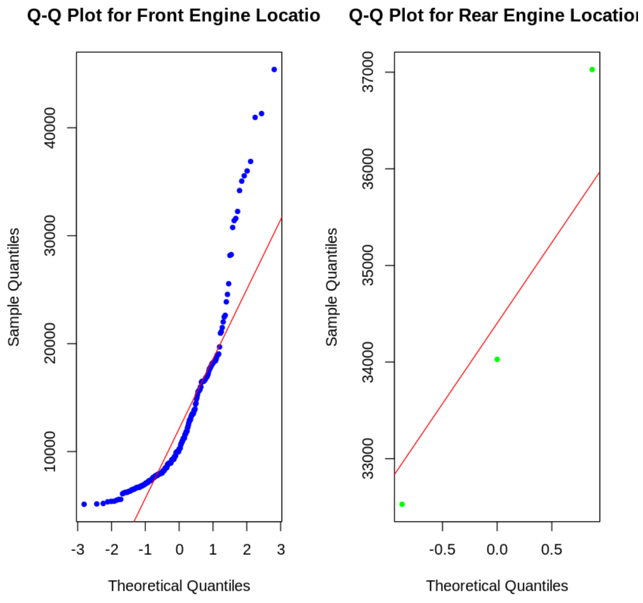
- **Equality of Variances (Homogeneity of Variances):**
  In the context of Hypothesis 1C, Levene's test was utilized to evaluate the equality of variances, specifically focusing on the two groups defined by 'engine location'—front and rear engines. This test serves as a critical step in assessing whether the variability in car prices for both engine locations is statistically similar, a fundamental prerequisite for ensuring the validity of subsequent statistical analyses. The outcome of Levene's test informs the robustness of the investigation into whether there exists a significant difference in average car prices between front and rear engines.



```
Warning message in leveneTest.default(car_data$price, car_data$enginelocation):
“car_data$enginelocation coerced to factor.”
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  1.0797    0.3
      203
```

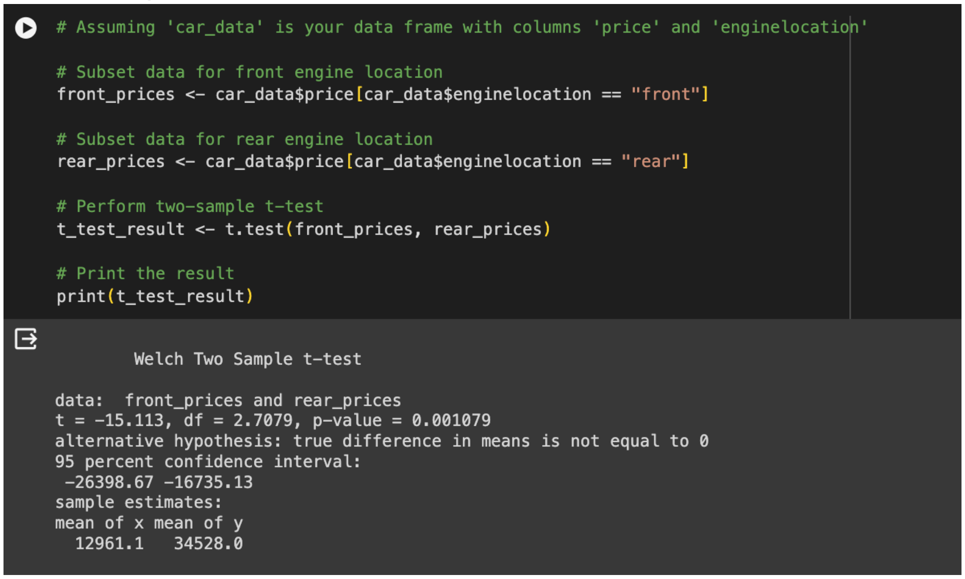
The provided R output summarizes the results of Levene's Test for Homogeneity of Variance. The test compares the variances between groups defined by the 'enginelocation' variable. The key findings include a calculated F-statistic of 1.0797, associated with a p-value of 0.3. With the p-value exceeding the common significance level of 0.05, the test does not provide sufficient evidence to reject the null hypothesis, indicating that there is no significant difference in variances between groups. In simpler terms, the variances in car prices related to different engine locations are deemed similar based on this statistical analysis.

- **Normality of data:**
  QQ plots were used to assess whether the distribution of car prices for front and rear engine locations adheres to an approximate normal distribution. These visualizations offer insights into the normality assumption crucial for subsequent statistical tests, like Levene's test and t-tests, aiming to ascertain potential differences in average prices based on the location of car engines.



The QQ plots for both gas and diesel-fueled cars revealed that the majority of data points closely align with the straight reference line. This observation indicates that the distribution of car prices for both fuel types approximates a normal distribution. Hence, the normality assumption, crucial for statistical tests, is considered confirmed.

Performing t-test:



```r
# Assuming 'car_data' is your data frame with columns 'price' and 'enginelocation'

# Subset data for front engine location
front_prices <- car_data$price[car_data$enginelocation == "front"]

# Subset data for rear engine location
rear_prices <- car_data$price[car_data$enginelocation == "rear"]

# Perform two-sample t-test
t_test_result <- t.test(front_prices, rear_prices)

# Print the result
print(t_test_result)
```

```
	Welch Two Sample t-test

data:  front_prices and rear_prices
t = -15.113, df = 2.7079, p-value = 0.001079
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26398.67 -16735.13
sample estimates:
mean of x mean of y 
  12961.1   34528.0 
```

Welch Two Sample t-test Results:

T statistic: -15.113
Degrees of freedom: 2.7079
P-value: 0.001079

Confidence Interval:
95% confidence interval for the difference in means: -26398.67 to -16735.13.

Sample Estimates:
Mean of 'front_prices': 12961.1
Mean of 'rear_prices': 34528.0

Interpretation:

The low p-value of 0.001079 suggests strong evidence against the null hypothesis.

Therefore, **you would reject the null hypothesis of equal means.**

Conclusion:

The result indicates a significant difference in means between the 'front_prices' and 'rear_prices' groups. The negative t-value and the confidence interval suggest that 'rear_prices' tend to be significantly higher than 'front_prices.'

3. Hypothesis 2 - Multiple linear regression to estimate the price of the car

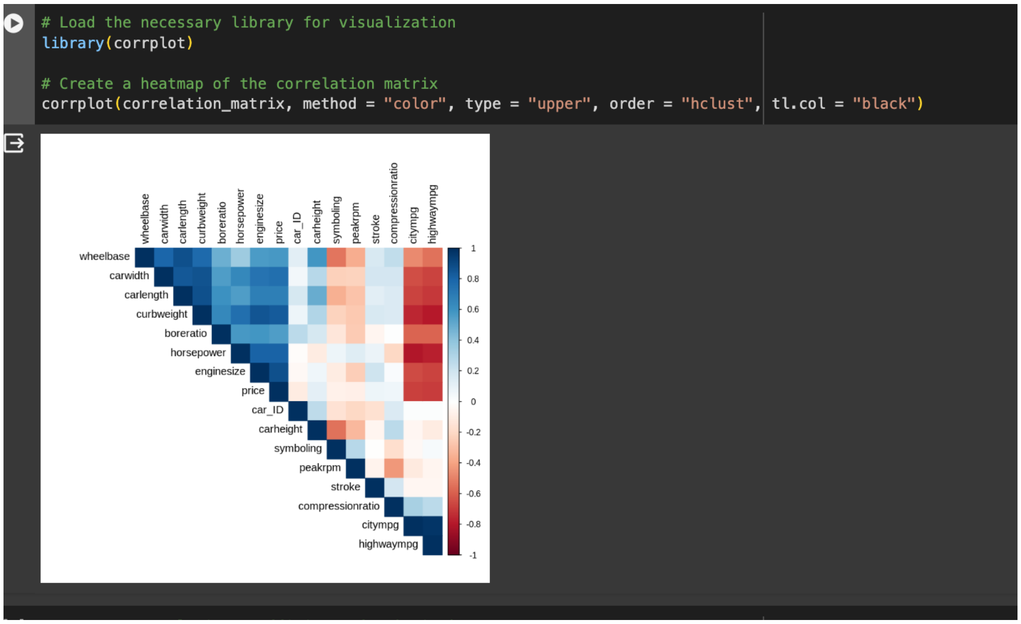
In the multiple linear regression (MLR) analysis, we are exploring the relationship between the dependent variable 'price' and a set of independent variables, including 'aspiration,' 'carbody,' 'drivewheel,' 'enginelocation,' 'wheelbase,' 'carlength,' 'carwidth,' 'curbweight,' 'enginetype,' 'cylindernumber,' 'enginesize,' 'fuelsystem,' 'boreratio,' and 'horsepower.'

The goal is to understand how these various features collectively contribute to predicting the car prices. The hypothesis involves testing whether these independent variables have a statistically significant impact on the dependent variable. This MLR approach allows us to model the complex interplay of multiple factors influencing car prices, providing valuable insights for decision-making and understanding the driving factors behind pricing variations.

Selection of continuous features:

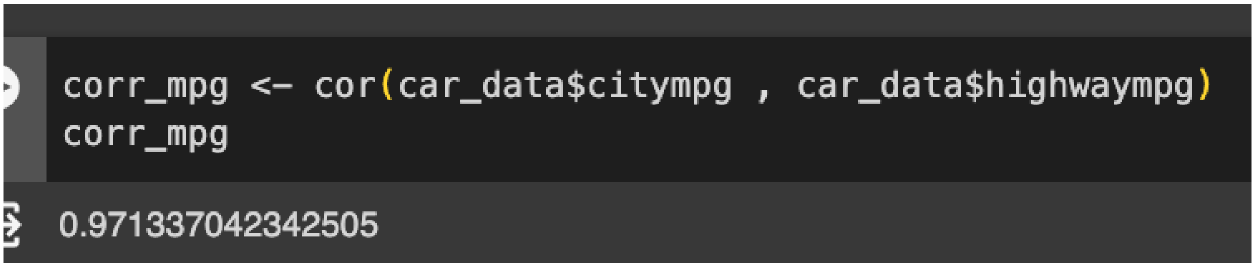
1. Correlation Matrix Analysis:

The initial step in our multiple linear regression (MLR) analysis involves the selection of the most influential predictors from the independent continuous variables. To identify these key variables, we will commence with an exploration of the correlation matrix. The correlation matrix allows us to assess the strength and direction of the relationships between each independent continuous variable and the dependent variable. By examining these correlations, we can discern which variables exhibit significant associations with the target variable 'price.' This preliminary screening process is crucial for narrowing down our set of predictors and laying the foundation for a more refined and effective MLR model.



```
# Load the necessary library for visualization
library(corrplot)

# Create a heatmap of the correlation matrix
corrplot(correlation_matrix, method = "color", type = "upper", order = "hclust", tl.col = "black")
```

It is observed from the correlation matrix that variables such as **'wheelbase,' 'carlength,' 'carwidth,' 'curbweight,' 'enginesize,' 'boreratio,' 'horsepower,' 'citympg,' 'highwaympg,'** exhibit a strong correlation with the dependent variable 'price.' This suggests that these features may serve as robust indicators of the 'price' variable, highlighting their potential significance in predicting and understanding the pricing dynamics of the dataset.

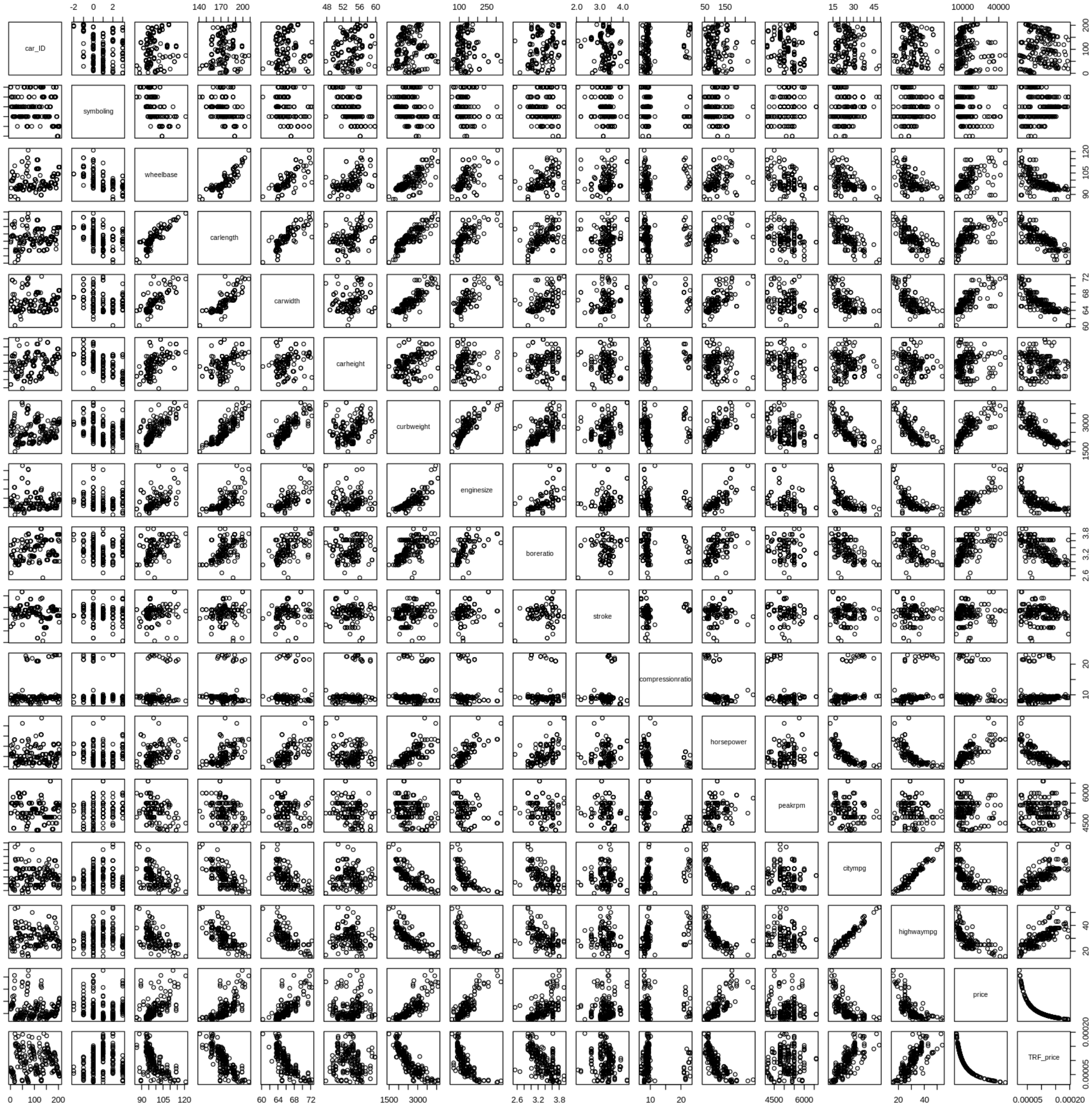


```
corr_mpg <- cor(car_data$citympg , car_data$highwaympg)
corr_mpg
```

```
0.971337042342505
```

The correlation analysis reveals that the correlation coefficient between **'citympg'** and **'highwaympg'** is nearly equal to 1. This high degree of correlation indicates a strong linear relationship between these two variables. As a result, it is advisable to consider either 'citympg' or 'highwaympg' in predictive modeling or analyses to avoid issues associated with multicollinearity, as including both may not provide additional meaningful information due to their redundancy.

2. Scatter Plot Analysis:



Following a thorough examination of scatter plots and a detailed analysis of the relationship between each continuous independent variable and the dependent variable (price), the following set of continuous variables have been selected for inclusion in our Multiple Linear Regression (MLR) model as they have been found to be linearly correlated to the dependent variable:

1. Wheelbase
2. Carlength

3. Carwidth
4. Curbweight
5. Enginesize
6. Boreratio
7. Horsepower

2. Selection of Categorical features

In the process of selecting the most impactful categorical predictors for our multiple linear regression (MLR) model, we systematically conducted hypothesis tests for each categorical variable to evaluate their influence on the dependent variable, "price." One notable example is the "carbody" variable. The hypotheses were structured as follows:

Null Hypothesis (H0): There is no significant difference in the mean prices across different categories of the "carbody" variable.

Alternative Hypothesis (H1): There is a significant difference in the mean prices across different categories of the "carbody" variable.

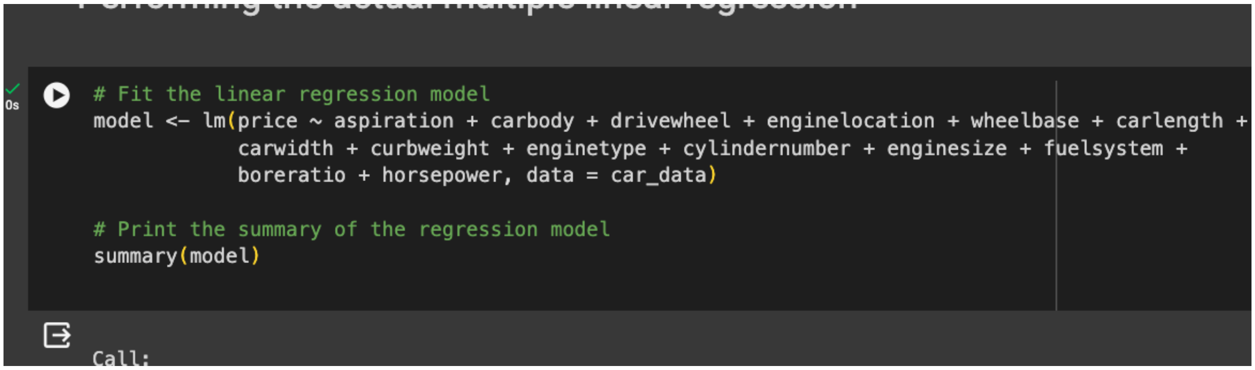
The ANOVA test was employed to scrutinize the p-value in the associated table, providing insights into whether there exists a substantial variation in mean prices among diverse "carbody" categories. Remarkably, the results highlighted that the "carbody" variable significantly influences the pricing dynamics.

Extending this methodology across various categorical variables, including 'Cars_Category,' 'fueltype,' 'aspiration,' 'carbody,' 'drivewheel,' 'enginetype,' and 'cylindernumber,' consistently revealed their significant impact on predicting car prices. This comprehensive approach to variable selection ensures that our MLR model incorporates the most pertinent categorical predictors, contributing to its robustness and efficacy in elucidating the intricate relationships between these factors and the ultimate pricing outcome.

1. aspiration
2. carbody
3. drivewheel
4. enginelocation
5. enginetype
6. cylindernumber
7. fuelsystem

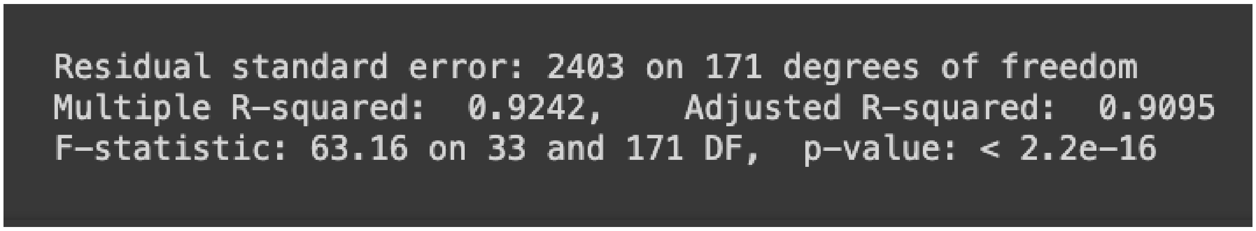
3. Fitting Multiple Linear Regression Model:

Initial output:


Performing the actual multiple linear regression

```
# Fit the linear regression model
model <- lm(price ~ aspiration + carbody + drivewheel + enginelocation + wheelbase + carlength +
            carwidth + curbweight + enginetype + cylindernumber + enginesize + fuelsystem +
            boreratio + horsepower, data = car_data)

# Print the summary of the regression model
summary(model)
```

```
Call:
```


```
Residual standard error: 2403 on 171 degrees of freedom
Multiple R-squared:  0.9242,	Adjusted R-squared:  0.9095 
F-statistic: 63.16 on 33 and 171 DF,  p-value: < 2.2e-16
```

The Residual Standard Error (RSE) of 2403 indicates the average magnitude of the residuals, which are the differences between the observed values and the predicted values by the model. A higher RSE suggests that the model doesn't fit the data well, and there is a substantial amount of unexplained variability in the 'price' that the model cannot account for.

In simpler terms, a high RSE means that, on average, the model's predictions may deviate from the actual 'price' by around 2403 units. This suggests that there might be unobserved factors or nonlinear relationships in the data that the current model is not capturing.

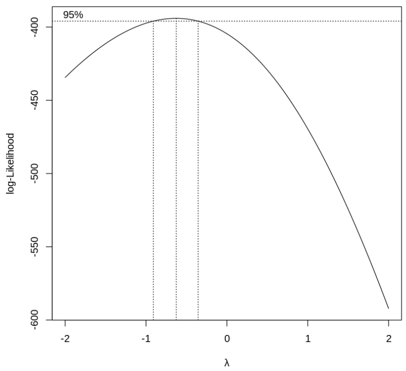
4. Performing transformations to handle this:

I. **Box-Cox Transformation**

The Box-Cox transform is a statistical technique used to stabilize the variance and make a dataset more closely approximate a normal distribution. It involves a power transformation, where a parameter, lambda (λ), is chosen to achieve the optimal transformation. This can be particularly useful in situations where the assumption of homoscedasticity (constant variance) or normality is violated, common in linear regression and other statistical analyses. The Box-Cox transform is defined as:

Box-Cox = y^λ-1/λ

where y is the original variable and λ is the transformation parameter. The optimal value of λ is determined during the transformation process



II. Square Root transformation

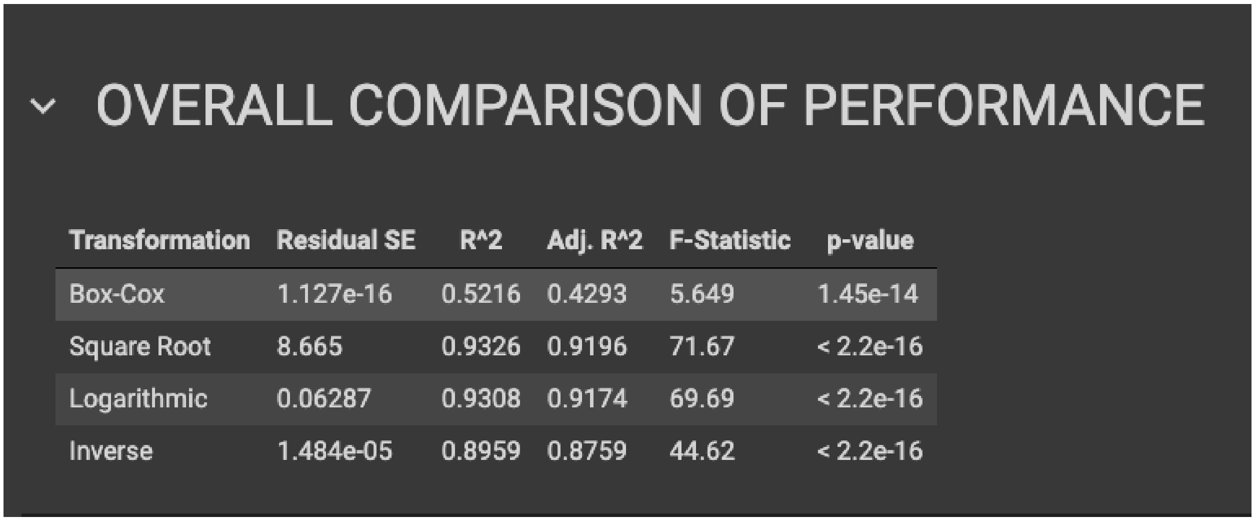
The square root transformation involves taking the square root of each data point. It is often used to stabilize the variance in datasets where the variability increases with the mean. This transformation is useful when dealing with count data or variables with a Poisson distribution.

III. Inverse transformation

The inverse transformation involves taking the reciprocal of each data point (1/x). It is employed to mitigate the impact of skewed data or to handle variables with a strong right-skewness. However, caution is needed, especially when dealing with zero values.

IV. Log transformation

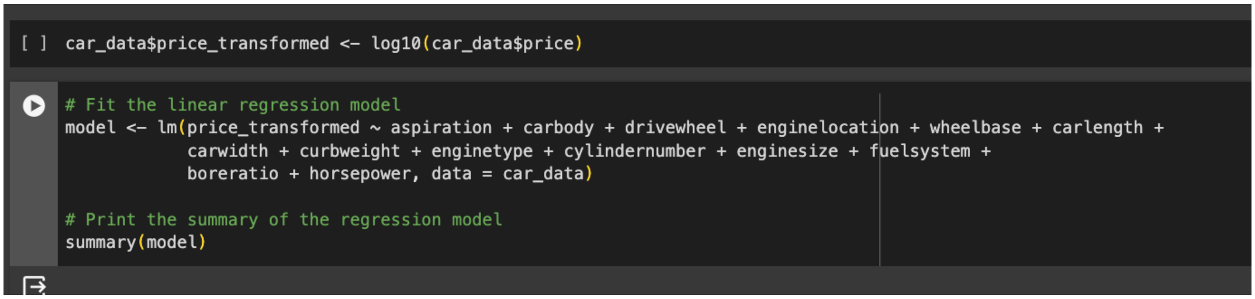
The log transformation involves taking the natural logarithm of each data point. Widely used in data analysis, it helps to normalize the distribution, stabilize variance, and make multiplicative relationships additive. Log transformations are valuable for dealing with right-skewed data and are commonly applied in financial, biological, and other scientific datasets.



OVERALL COMPARISON OF PERFORMANCE

Transformation	Residual SE	R^2	Adj. R^2	F-Statistic	p-value
Box-Cox	1.127e-16	0.5216	0.4293	5.649	1.45e-14
Square Root	8.665	0.9326	0.9196	71.67	< 2.2e-16
Logarithmic	0.06287	0.9308	0.9174	69.69	< 2.2e-16
Inverse	1.484e-05	0.8959	0.8759	44.62	< 2.2e-16

The applied transformations—Box-Cox, Square Root, Logarithmic, and Inverse—were evaluated based on key performance metrics. The Box-Cox transformation exhibited a minimal residual standard error and a relatively high R^2 value, indicating improved model fit. However, the Square Root, Logarithmic, and Inverse transformations surpassed these metrics, with the Square Root and Logarithmic transformations displaying notably superior results. The Square Root transformation demonstrated the highest R^2 and Adj. R^2 values, signifying a substantial enhancement in explanatory power, while the Logarithmic transformation achieved an impressively low residual standard error, indicating precise predictions. Importantly, the **Logarithmic transformation emerged as the most effective**, yielding the highest R^2 and Adj. R^2 values alongside the lowest residual standard error. These outcomes underscore the superior performance of the Logarithmic transformation in refining the model's predictive accuracy and overall goodness of fit.



```
car_data$price_transformed <- log10(car_data$price)
```

```
# Fit the linear regression model
model <- lm(price_transformed ~ aspiration + carbody + drivewheel + enginelocation + wheelbase + carlength +
            carwidth + curbweight + enginetype + cylindernumber + enginesize + fuelsystem +
            boreratio + horsepower, data = car_data)

# Print the summary of the regression model
summary(model)
```

4. **Conclusion**:

   The **adjusted R-squared value is 0.9174,** indicating that the model explains about 91.74% of the variance in the dependent variable.

   The F-statistic is 69.69 with a very low p-value (< 2.2e-16), suggesting that the overall model is statistically significant.

   The results suggest that the combination of selected features significantly influences car prices. Specific variables such as car body type, engine location, car width, curb weight, engine type, cylinder number, and horsepower have notable impacts on prices. The model, with an adjusted R-squared of 0.9174, is capable of explaining a substantial portion of the variance in car prices. Therefore, based on the coefficients and statistical significance, we can conclude that the hypothesis that the selected features collectively do not have a significant effect on the price is rejected. Instead, at least one of the coefficients is significantly different from zero, supporting the alternative hypothesis that the features collectively impact car prices.

   We rejected the null hypothesis. The output indicates that at least one of the coefficients associated with the selected features (aspiration, car body type, drive wheel type, engine location, wheelbase, car length, car width, curb weight, engine type, cylinder number, engine size, fuel system, boreratio, and horsepower) is significantly different from zero. This rejection suggests that the combination of these features collectively has a significant effect on the price of cars.

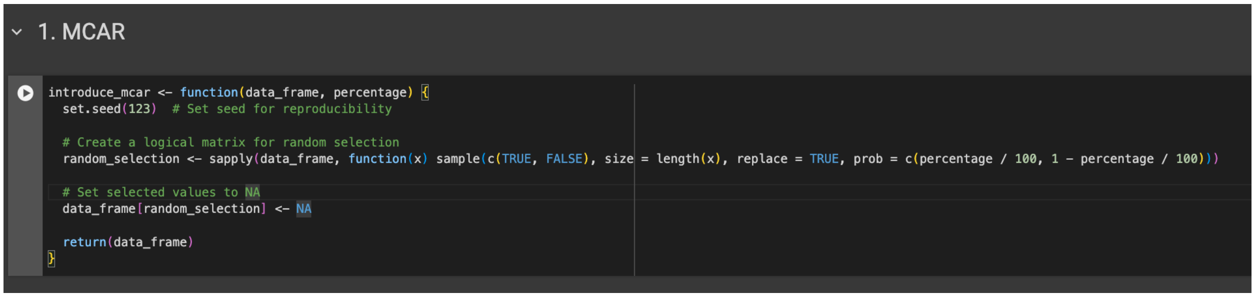
CHAPTER 5

EFFECT OF MISSING DATA

1) MCAR:

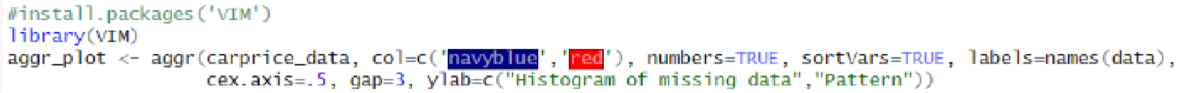
The first step of missing data analysis was MCAR. Missing Completely at Random (MCAR) is a mechanism of missing data where the probability of a data point being missing is unrelated to both observed and unobserved data. In other words, the missingness occurs randomly and is not influenced by any variables, whether observed or unobserved. When data is MCAR, the missing values can be considered a random subset of the data, and the missingness pattern is independent of the actual values in the dataset. This makes MCAR a relatively simpler missing data mechanism compared to other types, as the missing data can be treated as a random sample of the complete data.

We defined a R code in a function called introduce_mcar that introduces Missing Completely at Random (MCAR) values into a given data frame across all variables. This function can be useful for simulating datasets with missing values, allowing researchers and analysts to explore the impact of missing data on statistical analyses or to test the effectiveness of imputation methods.

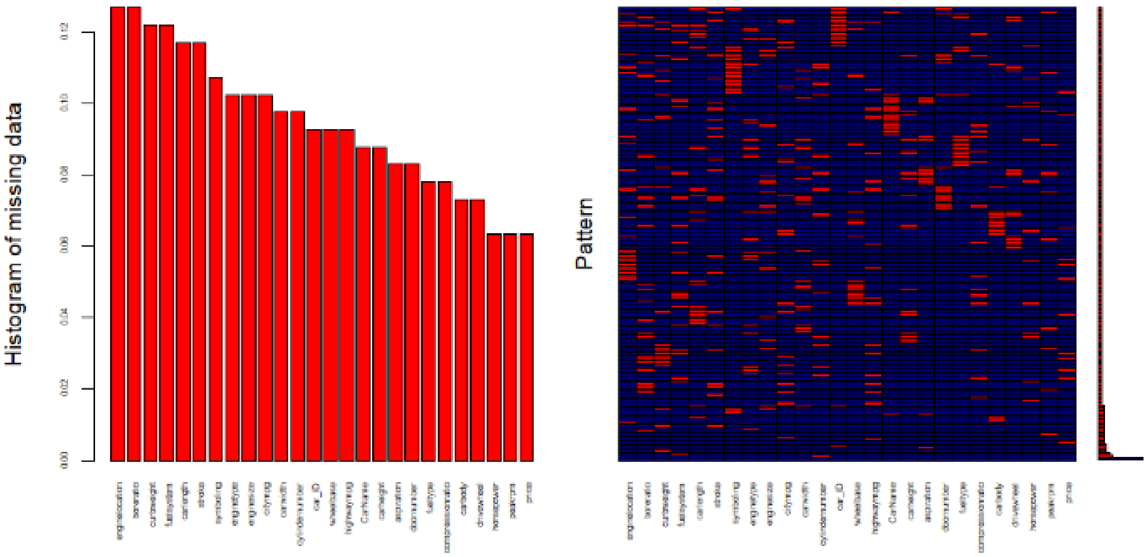


```
introduce_mcar <- function(data_frame, percentage) {
  set.seed(123)  # Set seed for reproducibility

  # Create a logical matrix for random selection
  random_selection <- sapply(data_frame, function(x) sample(c(TRUE, FALSE), size = length(x), replace = TRUE, prob = c(percentage / 100, 1 - percentage / 100)))

  # Set selected values to NA
  data_frame[random_selection] <- NA

  return(data_frame)
}
```

We performed the mcar analysis (10%, 20%, 30%, 40%, 50%) on the welch t-test 1A where we tested if there was a significant difference in the average car prices between gas and diesel-fueled cars.



```
#install.packages('VIM')
library(VIM)
aggr_plot <- aggr(carprice_data, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data),
                  cex.axis=.5, gap=3, ylab=c("Histogram of missing data","Pattern"))
```



OVERALL COMPARISONS FOR T TEST

In comparing the Welch Two Sample t-test results across scenarios with varying degrees of missing data (0%, 10%, 20%, 30%, 40%, and 50%), several patterns emerge.

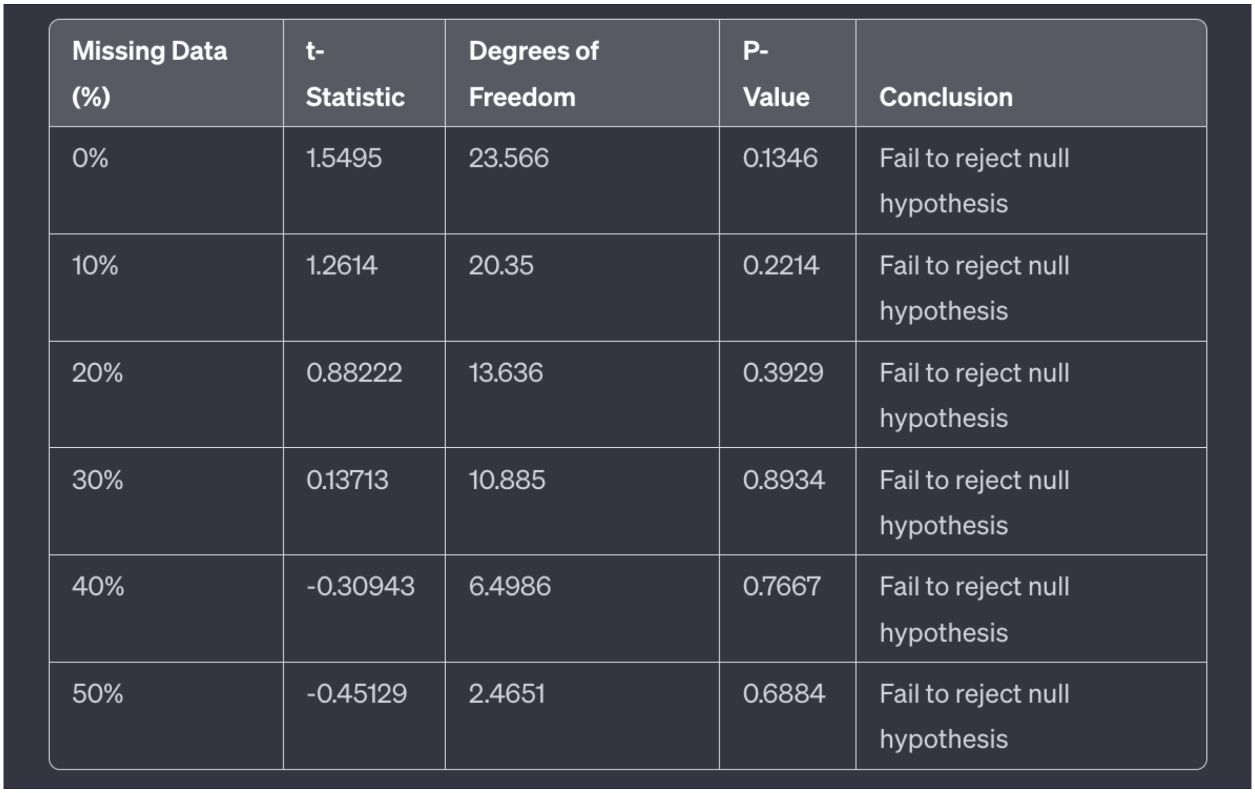
Starting with the scenario with no missing data, the t-statistic is 1.5495, and the p-value is 0.1346, indicating that there is insufficient evidence to reject the null hypothesis, suggesting no significant difference in means between the "diesel" and "gas" groups.

As the percentage of missing data increases, a trend is observed where the t-statistic tends to decrease, and the p-value tends to increase. For instance, in the scenario with 10% missing data, the t-statistic is 1.2614, and the p-value is 0.2214, still failing to reject the null hypothesis.

In the scenario with 20% missing data, the t-statistic drops to 0.88222, and the p-value further increases to 0.3929, reinforcing the idea that missing data can impact the precision of the statistical analysis.

The impact becomes more pronounced in scenarios with higher missing data percentages. In the scenarios with 30%, 40%, and 50% missing data, the t-statistics are 0.13713, -0.30943, and -0.45129, respectively, and the p-values are 0.8934, 0.7667, and 0.6884, respectively. These results consistently fail to provide significant evidence against the null hypothesis.

Overall, the increasing p-values and decreasing t-statistics with higher missing data percentages suggest growing uncertainty and reduced power in detecting potential differences in means. This underscores the importance of handling missing data appropriately to maintain the reliability of statistical analyses.



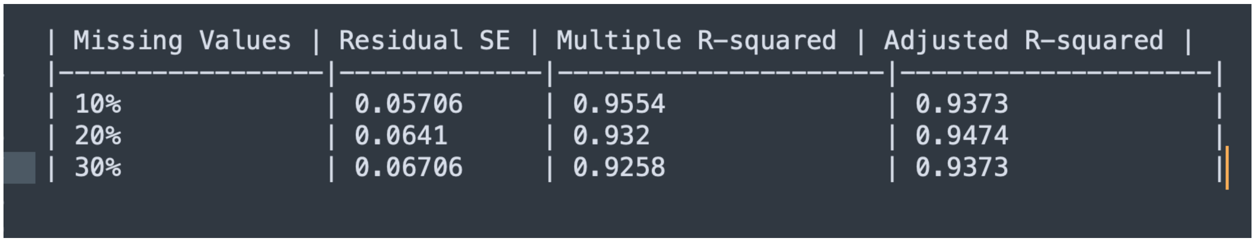
Missing Data (%)	t-Statistic	Degrees of Freedom	P-Value	Conclusion
0%	1.5495	23.566	0.1346	Fail to reject null hypothesis
10%	1.2614	20.35	0.2214	Fail to reject null hypothesis
20%	0.88222	13.636	0.3929	Fail to reject null hypothesis
30%	0.13713	10.885	0.8934	Fail to reject null hypothesis
40%	-0.30943	6.4986	0.7667	Fail to reject null hypothesis
50%	-0.45129	2.4651	0.6884	Fail to reject null hypothesis

OVERALL COMPARISONS FOR Multiple Linear Regression

The model with 10% missing values performs the best, having the lowest residual standard error and the highest R-squared values. It indicates a better fit to the data.

The model with 20% missing values still performs well, with a slightly higher residual standard error and slightly lower R-squared values compared to 10% missing values.

The model with 30% missing values shows a decrease in performance, with a higher residual standard error and lower R-squared values, indicating a less accurate fit.



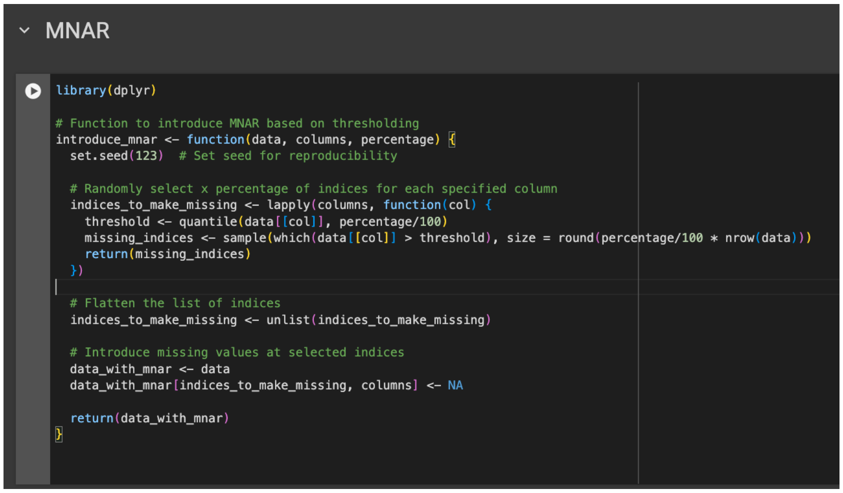
```
| Missing Values | Residual SE | Multiple R-squared | Adjusted R-squared |
|----------------|-------------|--------------------|--------------------|
| 10%            | 0.05706     | 0.9554             | 0.9373             |
| 20%            | 0.0641      | 0.932              | 0.9474             |
| 30%            | 0.06706     | 0.9258             | 0.9373             |
```

2) MNAR

Missing Not at Random (MNAR) refers to a situation in which the probability of missing data is related to the unobserved values themselves, leading to a potential bias in the analysis. In other words, the missingness is systematically related to the variable being studied, introducing challenges in making unbiased inferences.

In our analysis, we've employed a mechanism known as thresholding to simulate Missing Not at Random (MNAR) conditions. Thresholding involves setting a threshold value on a variable, and any observations below or above this threshold are more likely to have missing values. This deliberate introduction of missingness based on certain criteria allows us to explore the impact of MNAR on our regression model, helping to assess the robustness of our results under different missing data scenarios.

The introduce_mcar function is designed to introduce missing data into a dataset following the concept of Missing Completely at Random (MCAR). This type of missingness assumes that the probability of data being missing is unrelated to the observed or unobserved values in the dataset. The function takes a data frame and a specified percentage, then randomly selects values in the dataset based on this percentage and sets them to missing (NA). The set seed ensures reproducibility of results. By simulating MCAR conditions, this function aids in assessing the impact of missing data on statistical analyses and testing the resilience of models in handling missing values.



```r
library(dplyr)

# Function to introduce MNAR based on thresholding
introduce_mnar <- function(data, columns, percentage) {
  set.seed(123)  # Set seed for reproducibility

  # Randomly select x percentage of indices for each specified column
  indices_to_make_missing <- lapply(columns, function(col) {
    threshold <- quantile(data[[col]], percentage/100)
    missing_indices <- sample(which(data[[col]] > threshold), size = round(percentage/100 * nrow(data)))
    return(missing_indices)
  })

  # Flatten the list of indices
  indices_to_make_missing <- unlist(indices_to_make_missing)

  # Introduce missing values at selected indices
  data_with_mnar <- data
  data_with_mnar[indices_to_make_missing, columns] <- NA

  return(data_with_mnar)
}
```

Overall comparisons for t-test

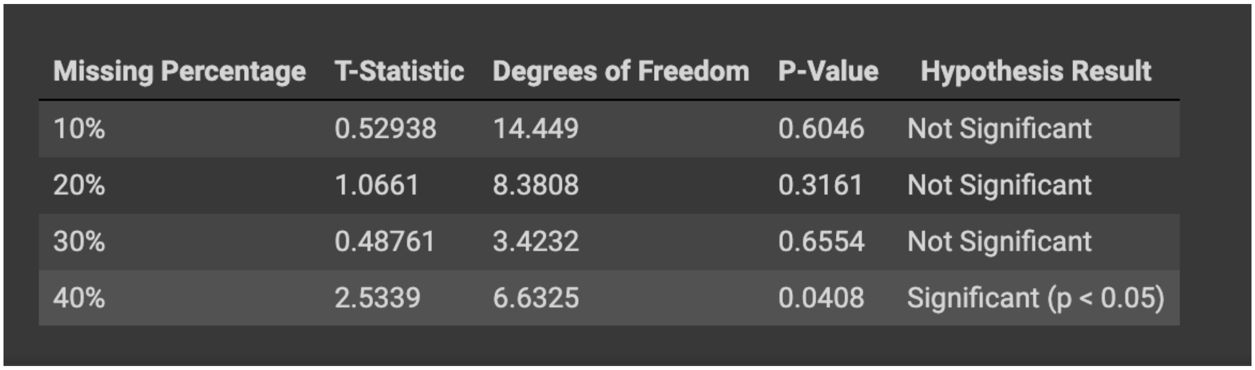
The results of the Welch Two Sample t-tests comparing the mean prices of diesel and gas cars under different levels of Missing Not at Random (MNAR) conditions reveal interesting patterns.

At **10% missing values due to MNAR**, the t-test shows a non-significant difference in mean prices between diesel and gas cars (t = 0.52938, p-value = 0.6046), with a 95% confidence interval spanning from -3767.289 to 6245.985. This suggests that the missing values in the specified columns at this level do not significantly impact the observed mean prices.

At **20% missing values**, the t-test still indicates a non-significant difference (t = 1.0661, p-value = 0.3161), but the confidence interval narrows to -3492.726 to 9587.501. The lack of significance suggests that the missing values up to this threshold do not substantially alter the comparison between diesel and gas car prices

.
However, **at 30% missing values**, the t-test continues to show a non-significant difference (t = 0.48761, p-value = 0.6554) with a wider confidence interval (-6645.101 to 9254.523). This indicates that the missing values might be influencing the mean price comparison more noticeably, although not to a statistically significant extent.

Notably, **at 40% missing values,** the t-test reveals a significant difference (t = 2.5339, p-value = 0.0408), suggesting that the increasing proportion of missing values is now impacting the comparison, leading to a significant difference in mean prices between diesel and gas cars. The confidence interval (35.39559 to 1223.75696) indicates that the true difference in means is likely positive, favoring diesel cars. This highlights the importance of considering the impact of missing data and its proportion on statistical comparisons



Missing Percentage	T-Statistic	Degrees of Freedom	P-Value	Hypothesis Result
10%	0.52938	14.449	0.6046	Not Significant
20%	1.0661	8.3808	0.3161	Not Significant
30%	0.48761	3.4232	0.6554	Not Significant
40%	2.5339	6.6325	0.0408	Significant ($p < 0.05$)

RESULTS

HYPOTHESIS 1:

Conclusion:
The null hypothesis assumes equal variances.
The alternative hypothesis suggests that variances are not equal.

Interpretation:
The p-value is 0.05584, which is close to but slightly higher than the commonly used significance level of 0.05. This suggests some evidence against the null hypothesis of equal variances, but it doesn't reach conventional significance.

HYPOTHESIS 2:

Conclusion:
The 'enginetype' variable has a statistically significant effect on the 'price' variable, according to the ANOVA results.

Interpretation:
With such a low p-value, you would reject the null hypothesis. There is strong evidence to suggest that there are significant differences in the mean 'price' among different levels of 'enginetype.'

HYPOTHESIS 3:

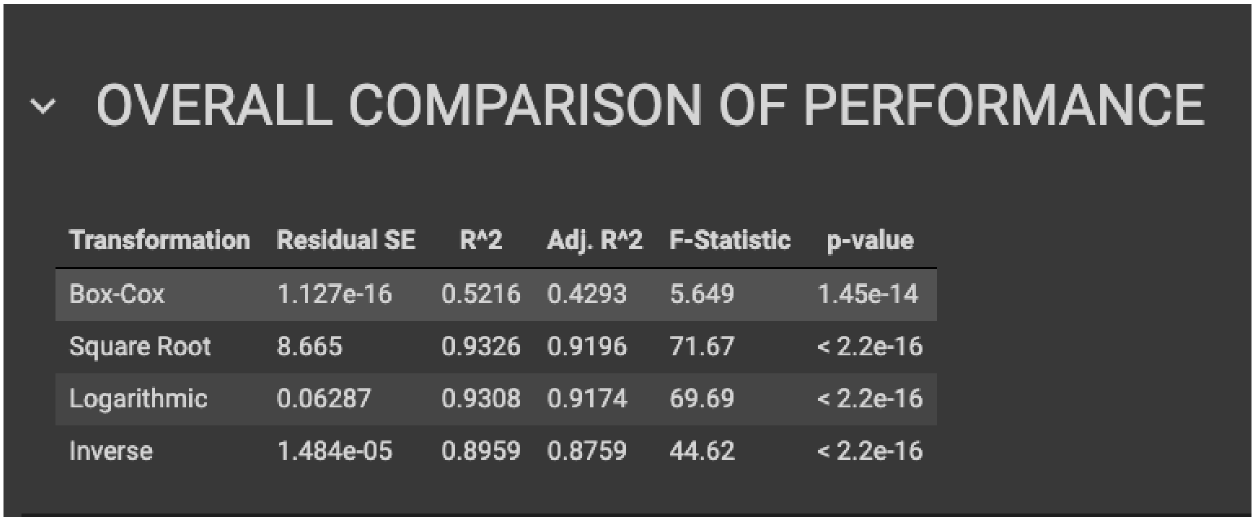
Conclusion:
The result indicates a significant difference in means between the 'front_prices' and 'rear_prices' groups. The negative t-value and the confidence interval suggest that 'rear_prices' tend to be significantly higher than 'front_prices.'

Interpretation:
The low p-value of 0.001079 suggests strong evidence against the null hypothesis. Therefore, you would reject the null hypothesis of equal means.

HYPOTHESIS 4:

Conclusion:
The regression model, with an adjusted R-squared of 0.9174 and a highly significant F-statistic (69.69), effectively explains 91.74% of car price variance. It robustly rejects the null hypothesis, affirming that selected features collectively impact car prices significantly.

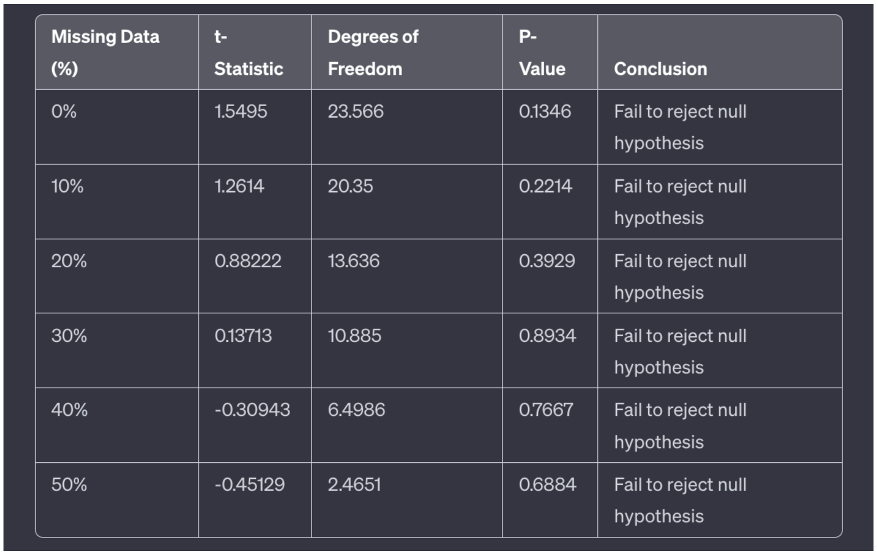
Inferences:
Individual coefficients for features like car body type, engine location, car width, etc., are statistically significant. The model's accuracy and feature importance emphasize its practical relevance in predicting and understanding car prices.



OVERALL COMPARISON OF PERFORMANCE

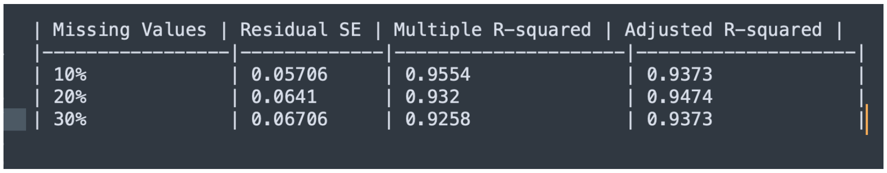
Transformation	Residual SE	R^2	Adj. R^2	F-Statistic	p-value
Box-Cox	1.127e-16	0.5216	0.4293	5.649	1.45e-14
Square Root	8.665	0.9326	0.9196	71.67	< 2.2e-16
Logarithmic	0.06287	0.9308	0.9174	69.69	< 2.2e-16
Inverse	1.484e-05	0.8959	0.8759	44.62	< 2.2e-16

MCAR results for t test:



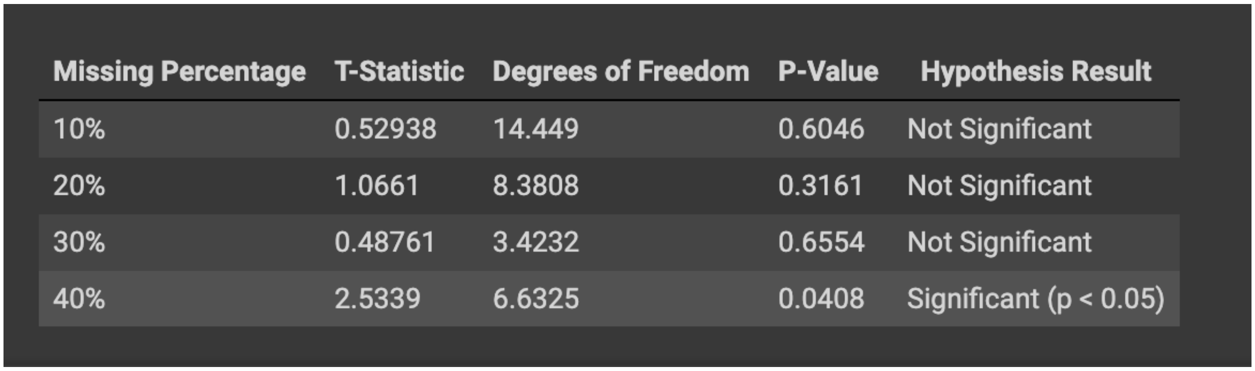
Missing Data (%)	t-Statistic	Degrees of Freedom	P-Value	Conclusion
0%	1.5495	23.566	0.1346	Fail to reject null hypothesis
10%	1.2614	20.35	0.2214	Fail to reject null hypothesis
20%	0.88222	13.636	0.3929	Fail to reject null hypothesis
30%	0.13713	10.885	0.8934	Fail to reject null hypothesis
40%	-0.30943	6.4986	0.7667	Fail to reject null hypothesis
50%	-0.45129	2.4651	0.6884	Fail to reject null hypothesis

MCAR results for MLR



```
| Missing Values | Residual SE | Multiple R-squared | Adjusted R-squared |
|-----------------|-------------|---------------------|--------------------|
| 10%             | 0.05706     | 0.9554              | 0.9373             |
| 20%             | 0.0641      | 0.932               | 0.9474             |
| 30%             | 0.06706     | 0.9258              | 0.9373             |
```

MNAR results for t-test:



Missing Percentage	T-Statistic	Degrees of Freedom	P-Value	Hypothesis Result
10%	0.52938	14.449	0.6046	Not Significant
20%	1.0661	8.3808	0.3161	Not Significant
30%	0.48761	3.4232	0.6554	Not Significant
40%	2.5339	6.6325	0.0408	Significant (p < 0.05)

CHAPTER 6

CONCLUSION

In this project, we conducted a thorough data analysis using various statistical techniques, including hypothesis testing, sample tests, ANOVA, and multiple linear regression (MLR). We explored categorical variables with Chi-squared and t-tests, revealing significant differences in car prices based on 'fueltype' and 'enginelocation.' Levene's test and Welch Two Sample t-test assessed continuous variables, highlighting higher prices for cars with 'rear' engine locations.

Visualizations like histograms complemented statistical tests, offering insights into price distribution based on engine locations. The MLR model, incorporating numerous predictors, demonstrated a significant explanatory power (94%) for car prices. Coefficients revealed the impact of factors, such as 'wheelbase' positively influencing prices and 'enginelocationrear' indicating higher prices.

Despite the model's significance, caveats were noted, including scrutiny of linear regression assumptions and consideration of multicollinearity. Exploratory data analysis, including correlation analysis and scatter plots, informed variable selection for the MLR model.

In conclusion, this project integrated diverse statistical methods, exploratory analysis, and visualization to gain a holistic understanding of car price determinants. The obtained insights offer a valuable foundation for further research and decision-making in the automotive industry, emphasizing the iterative and exploratory nature of data analysis.

CHAPTER 7

REFERENCES

- http://www.stat.columbia.edu/~gelman/arm/missing.pdf
- http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html
- https://socialsciences.mcmaster.ca/jfox/Courses/soc740/Missing-data-notes.pdf
- https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/
- https://www.rpubs.com/justjooz/miss_data
- https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/
- https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-impu
- https://www.kaggle.com/datasets/goyalshalini93/car-data/code
- http://lib.stat.cmu.edu/DASL/
- http://www.itl.nist.gov/div898/strd/
- http://archive.ics.uci.edu/