Michigan Technological University

STATISTICAL METHODS

MA 5701

PROJECT REPORT

Project Title : Analysis of Factors Affecting Vehicle Selling Price

Group Name : Stats Squad

1. Aishwarya Kotra
2. Surya Gowtham Vakkalagadda
3. Ganesh Vannam

Analysis Of Factors Affecting Vehicle Selling Price

Introduction

Our project undertakes a detailed exploration into the various factors that influence the pricing dynamics of the used car market. This sector, known for its vibrancy and complexity, is influenced by a multitude of variables that can significantly impact the valuation of vehicles. Economic shifts and evolving consumer preferences have led to an increased demand for used vehicles, emphasizing the need for a comprehensive understanding of the elements that determine car prices. Through this project, we aim to provide deeper insights into these factors, which will enable both buyers and sellers to navigate the market more effectively and make informed decisions.

The burgeoning interest in the used car market is driven by several key factors: the rapid depreciation of new vehicles, the increased reliability of older models, and shifts in consumer spending behaviour towards more cost-effective automotive options. These trends underscore the critical need for detailed empirical analysis to help stakeholders understand price determinations more clearly. Our analysis is centred on how specific attributes such as the age of the vehicle, mileage, type of fuel used, and engine size influence the market price of used cars. A precise understanding of these factors is essential for transparent and fair-trading practices within the industry.

To tackle these questions, we employ multiple linear regression analysis, a robust statistical technique that allows us to examine the impact of various car features on their selling prices. This method helps us identify which attributes play a significant role in car valuation and quantify their exact impact on pricing. Such detailed insights are invaluable to car dealers aiming to refine their pricing strategies and to consumers who are keen to ensure they are receiving or paying a fair price for a vehicle.

In addition to helping individual buyers and sellers, the predictive model developed from our analysis will also serve as a crucial tool for a broader range of stakeholders, including automotive dealerships and policymakers. This model will aid in refining business strategies across the industry, from adjusting inventory levels to fine-tuning consumer pricing models, ensuring that all transactions are conducted with a high degree of transparency and equity.

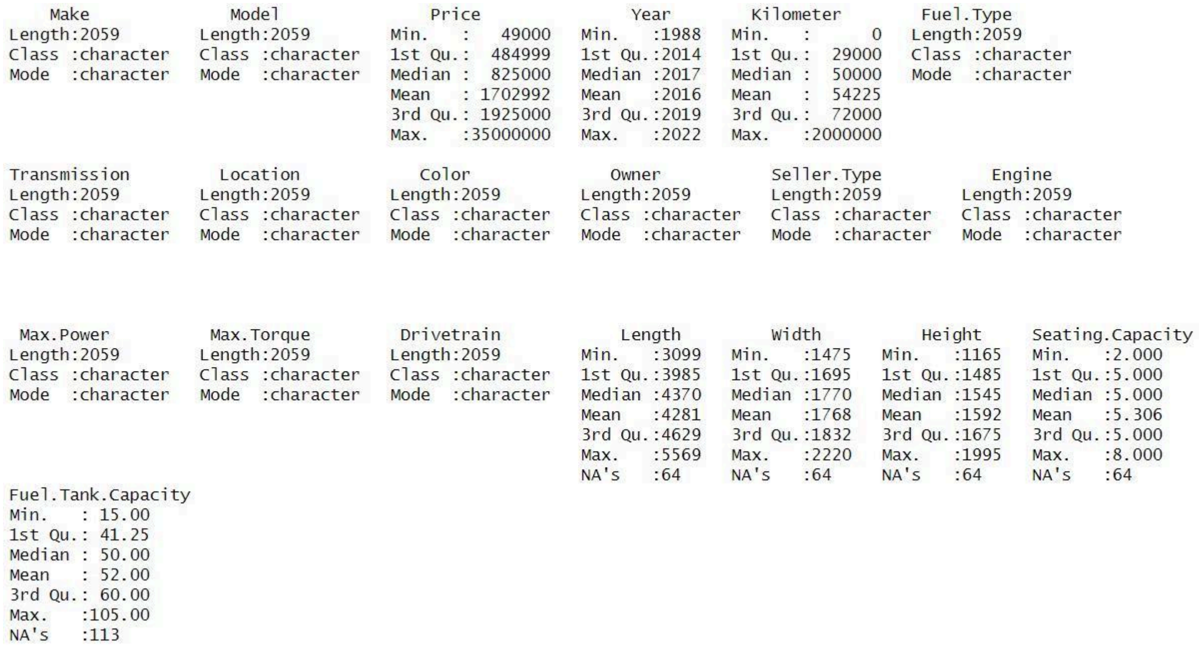
The project's broader implications are likely to extend even further, potentially leading to a more standardized approach to pricing within the used car industry. By providing a clearer view of the factors that influence vehicle pricing, our research will help streamline market operations and promote fair business practices.

In conclusion, this project is designed to dissect and understand the complex pricing dynamics in the used car market by identifying and quantifying the influence of various factors. Through thorough data analysis and the application of rigorous statistical methods, we aim to deliver insights that will not only enhance academic understanding but also provide practical tools and frameworks for market participants. By doing so, we ensure that our findings will serve as a reliable resource for stakeholders across the used car market, thereby facilitating informed decision-making and promoting transparency and fairness in market practices. This in-depth exploration will undoubtedly enrich the body of knowledge in the automotive industry and serve as a benchmark for future research and practical applications in market analysis.

Methods

Data

Our study utilized a dataset from a reputable online repository featuring 2059 observations of used cars, capturing diverse vehicle attributes relevant to their market pricing. This dataset includes 20 variables detailing the make and model, sale price, year of manufacture, kilometres travelled, fuel type, transmission, and other specifications such as engine details and physical dimensions.



```
     Make               Model                Price               Year        Kilometer          Fuel.Type
 Length:2059        Length:2059        Min.   :   49000   Min.   :1988   Min.   :      0   Length:2059
 Class :character   Class :character   1st Qu.:  484999   1st Qu.:2014   1st Qu.:  29000   Class :character
 Mode  :character   Mode  :character   Median :  825000   Median :2017   Median :  50000   Mode  :character
                                       Mean   : 1702992   Mean   :2016   Mean   :  54225
                                       3rd Qu.: 1925000   3rd Qu.:2019   3rd Qu.:  72000
                                       Max.   :35000000   Max.   :2022   Max.   :2000000

 Transmission         Location            Color              Owner             Seller.Type           Engine
 Length:2059        Length:2059        Length:2059        Length:2059        Length:2059        Length:2059
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


  Max.Power          Max.Torque          Drivetrain            Length         Width          Height      Seating.Capacity
 Length:2059        Length:2059        Length:2059        Min.   :3099   Min.   :1475   Min.   :1165   Min.   :2.000
 Class :character   Class :character   Class :character   1st Qu.:3985   1st Qu.:1695   1st Qu.:1485   1st Qu.:5.000
 Mode  :character   Mode  :character   Mode  :character   Median :4370   Median :1770   Median :1545   Median :5.000
                                                          Mean   :4281   Mean   :1768   Mean   :1592   Mean   :5.306
                                                          3rd Qu.:4629   3rd Qu.:1832   3rd Qu.:1675   3rd Qu.:5.000
                                                          Max.   :5569   Max.   :2220   Max.   :1995   Max.   :8.000
                                                          NA's   :64     NA's   :64     NA's   :64     NA's   :64
 Fuel.Tank.Capacity
 Min.   : 15.00
 1st Qu.: 41.25
 Median : 50.00
 Mean   : 52.00
 3rd Qu.: 60.00
 Max.   :105.00
 NA's   :113
```

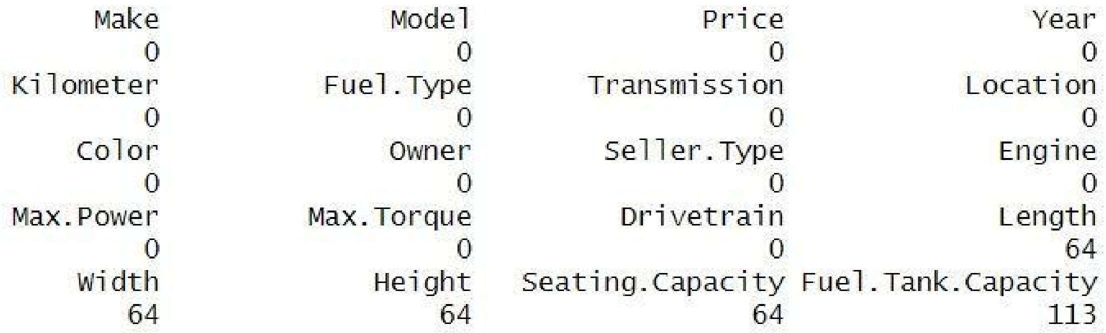
Summary statistics of used car dataset variables

Summary statistics were generated to gain insights into the dataset, revealing a price range from 49,000 to 3,500,000, indicating a mix of budget-friendly and luxury vehicle options. The manufacturing dates range from 1988 to 2022, allowing for an analysis of how a car's age correlates with its value. The mileage shows significant variation, pointing to a wide usage spectrum. Detailed engine and power specifications provide a basis for evaluating performance-based valuations. Categorical variables such as fuel type and transmission offer additional perspectives on vehicle performance and market preferences.

To investigate the factors influencing vehicle sale prices, we selected multiple linear regression as our statistical method. This approach is particularly effective for assessing the impact of several independent variables on a single outcome variable, which in this study is the selling price. The versatility of multiple linear regression is particularly suited to our data, which contains multiple dimensions that potentially affect the final selling price. This statistical method will enable us to ascertain the individual contribution of each factor, such as the vehicle's mileage, age, and engine details, while controlling for the presence of other variables.

Data Cleaning

In the data cleaning phase, we identified missing values across various variables in our dataset. A visual representation of the missing values for each variable is provided in the image below:



```
             Make             Model             Price              Year
                0                 0                 0                 0
        Kilometer         Fuel.Type      Transmission          Location
                0                 0                 0                 0
            Color             Owner       Seller.Type            Engine
                0                 0                 0                 0
        Max.Power        Max.Torque        Drivetrain            Length
                0                 0                 0                64
            Width            Height  Seating.Capacity Fuel.Tank.Capacity
               64                64                64               113
```

Figure: Missing Values Overview

This image illustrates the extent of missing data in our dataset, allowing us to prioritize our cleaning efforts effectively.

After identifying missing values, we employed several techniques to handle them and ensure the integrity of our dataset. These techniques included:

1. Imputation of Missing Numerical Values: We replaced missing numerical values with the median, a robust measure of central tendency.
2. Imputation of Missing Categorical Values: Missing categorical values were filled with the mode, representing the most frequently occurring category in each column.
3. Removal of Duplicate Entries: Duplicate rows were identified and removed to maintain the uniqueness of each observation.

Feature Engineering

In the feature engineering phase of our analysis, we focused on preparing the dataset to ensure compatibility with our modelling techniques. This involved several crucial steps designed to refine and optimize the data for analysis, which included handling missing values, encoding categorical variables, extracting numeric values from strings, and ensuring data integrity.

Handling Missing Values:

We addressed missing values comprehensively, recognizing that such gaps can significantly impact the quality of our analysis. For numerical variables, we opted to impute missing data using the median value of each variable, a method favoured for its robustness against outliers. In contrast, for categorical variables, we used the mode to fill in missing entries, ensuring that the most frequently occurring category was used to maintain the existing data distribution.

Encoding Categorical Variables:

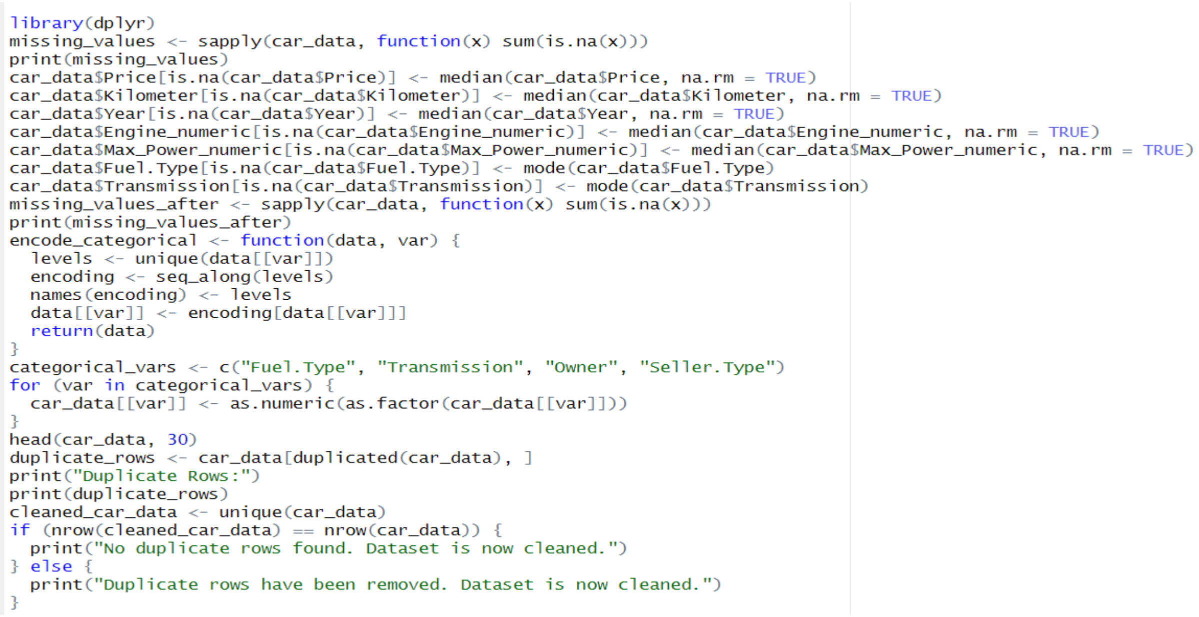
Transforming categorical variables into a numeric format is essential. We systematically converted variables such as fuel type, transmission, owner details, and seller type into numeric codes. This transformation not only facilitated the application of statistical models but also helped in maintaining the integrity and interpretability of the variables.

Extracting Numeric Data from Strings:

Our dataset included variables where numerical values were embedded within strings, such as engine capacity and max power output. We extracted these numeric parts for use in our quantitative analyses. This step was critical in ensuring that all relevant variables were accurately represented in a format suitable for regression analysis.

Ensuring Data Integrity:

To further enhance the dataset's quality, we identified and removed duplicate entries, ensuring that each data point represented a unique observation. This process was vital for preventing any bias that duplicates might introduce into our analysis.



```
library(dplyr)
missing_values <- sapply(car_data, function(x) sum(is.na(x)))
print(missing_values)
car_data$Price[is.na(car_data$Price)] <- median(car_data$Price, na.rm = TRUE)
car_data$Kilometer[is.na(car_data$Kilometer)] <- median(car_data$Kilometer, na.rm = TRUE)
car_data$Year[is.na(car_data$Year)] <- median(car_data$Year, na.rm = TRUE)
car_data$Engine_numeric[is.na(car_data$Engine_numeric)] <- median(car_data$Engine_numeric, na.rm = TRUE)
car_data$Max_Power_numeric[is.na(car_data$Max_Power_numeric)] <- median(car_data$Max_Power_numeric, na.rm = TRUE)
car_data$Fuel.Type[is.na(car_data$Fuel.Type)] <- mode(car_data$Fuel.Type)
car_data$Transmission[is.na(car_data$Transmission)] <- mode(car_data$Transmission)
missing_values_after <- sapply(car_data, function(x) sum(is.na(x)))
print(missing_values_after)
encode_categorical <- function(data, var) {
  levels <- unique(data[[var]])
  encoding <- seq_along(levels)
  names(encoding) <- levels
  data[[var]] <- encoding[data[[var]]]
  return(data)
}
categorical_vars <- c("Fuel.Type", "Transmission", "Owner", "Seller.Type")
for (var in categorical_vars) {
  car_data[[var]] <- as.numeric(as.factor(car_data[[var]]))
}
head(car_data, 30)
duplicate_rows <- car_data[duplicated(car_data), ]
print("Duplicate Rows:")
print(duplicate_rows)
cleaned_car_data <- unique(car_data)
if (nrow(cleaned_car_data) == nrow(car_data)) {
  print("No duplicate rows found. Dataset is now cleaned.")
} else {
  print("Duplicate rows have been removed. Dataset is now cleaned.")
}
```

Code for feature Engineering

Data Visualization

In our analysis of used car pricing dynamics, we utilized a range of visualization techniques to illuminate the relationship between different attributes and the selling prices of vehicles. Our exploration began with a box plot depicting the "Variation in Selling Price by Fuel Type." Each box in the plot represents the selling price distribution for a specific fuel type, with the x-axis delineating the fuel types and the y-axis denoting the selling prices. From this visualization, we can discern whether certain fuel types correlate with higher or lower selling prices.

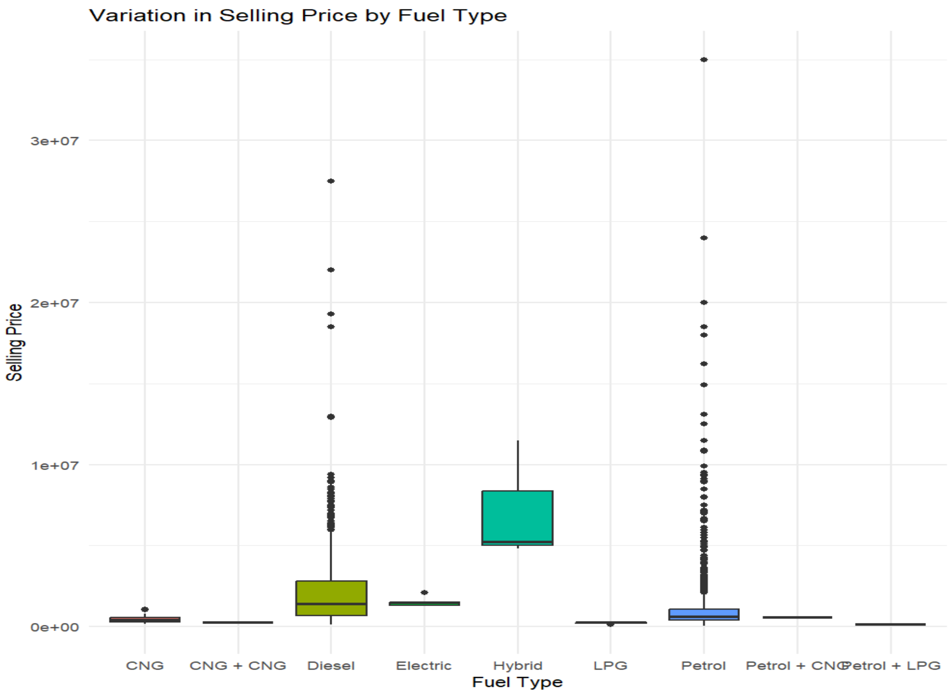


Figure 1: Boxplot of Variation in Selling Price by Fuel Type

Additionally, the "Distribution of Cars by Age" was explored using a histogram. This visualization offers insights into the age distribution of cars within the dataset. The x-axis indicates the age of cars, calculated as the difference between the current year and the year of manufacture, while the y-axis displays the frequency of cars within each age interval. By examining this histogram, the prevalence of cars across different age groups in the market can be observed, and how age influences selling prices can be discerned.

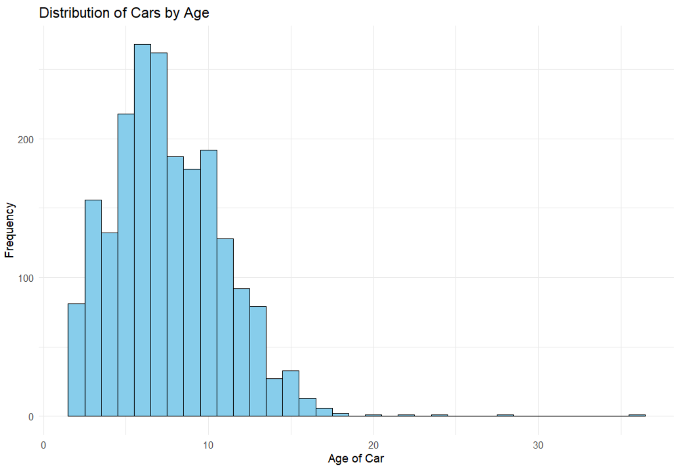
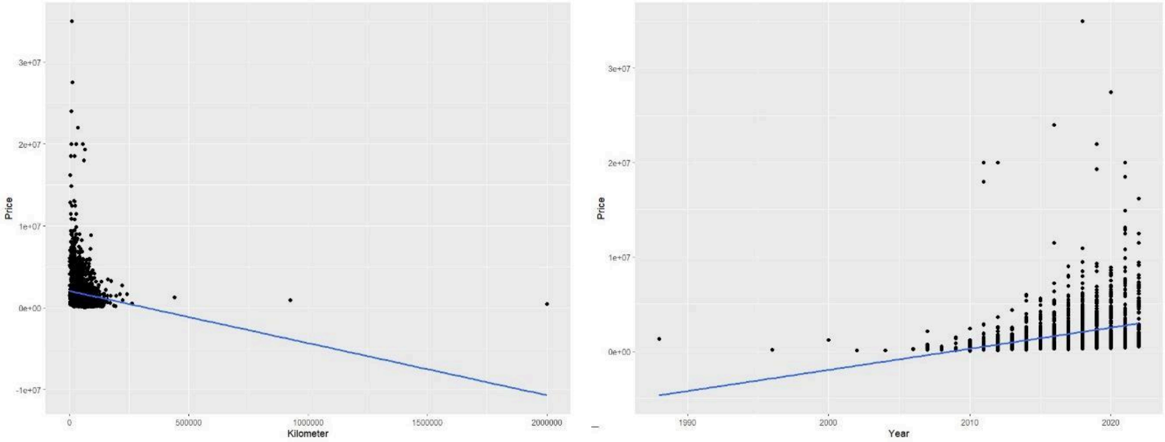


Figure 2: Histogram: Distribution of Cars by Age

Our statistical exploration encompassed creating scatterplots to investigate the linearity between the selling price and each predictor, such as mileage, vehicle age, engine size, power output, and various other factors. The plots revealed trends and relationships, illustrating, for instance, a general decrease in price with an increase in mileage, while newer vehicles seemed to fetch higher prices. Engine size and power output were also depicted, with both showing a positive trend, suggesting that vehicles with larger engines and more power tend to be priced higher.



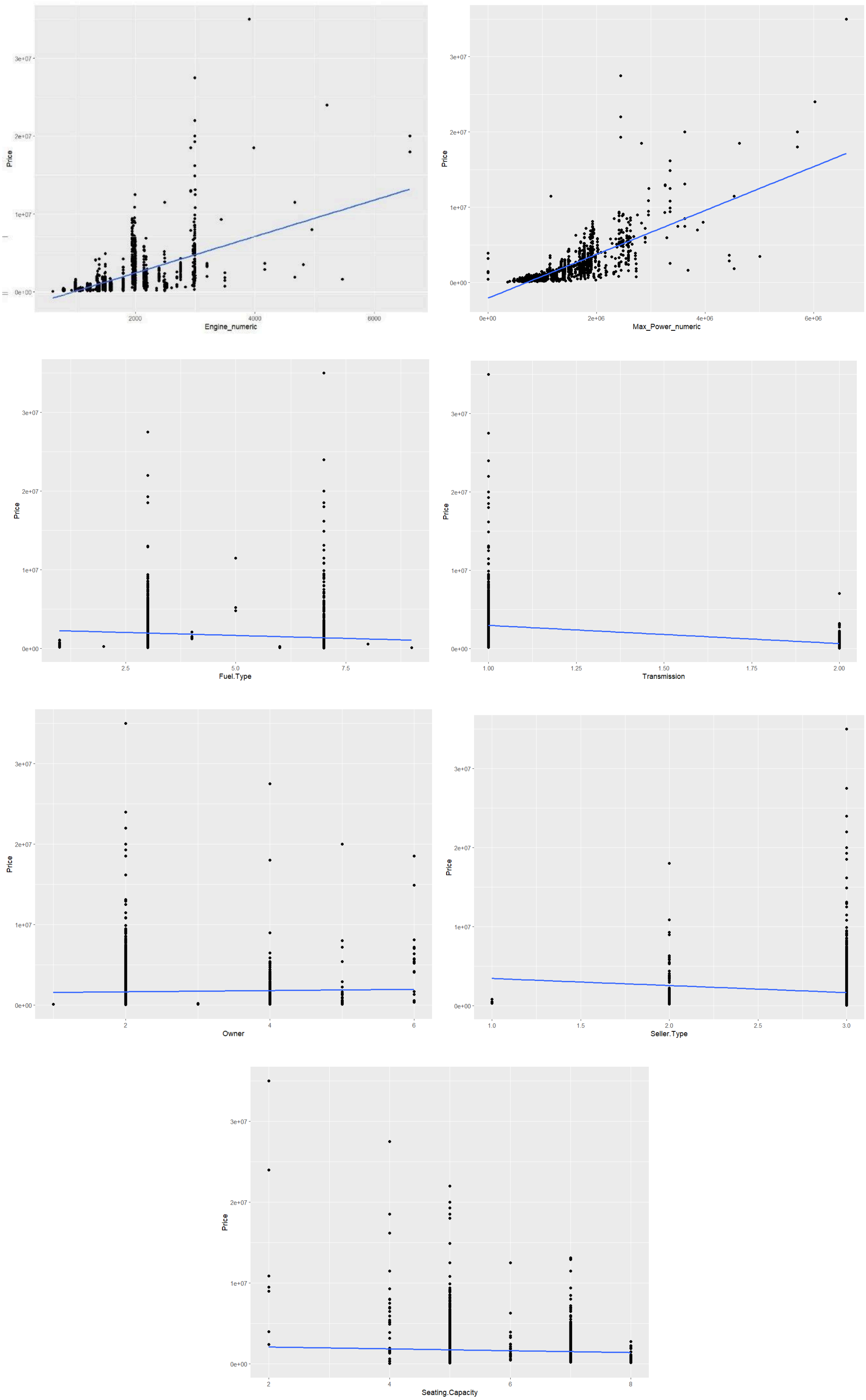


Figure 3: Scatterplot Series Illustrating Vehicle Selling Price Relationships with Various Predictors

These scatterplots are crucial for a visual assessment of the relationships and patterns within our data, offering a glimpse into the potential factors that influence vehicle pricing. The visualization of the data spread and the fitted trend lines are pivotal for interpreting the underlying trends, readying the data for a more in-depth regression analysis. These visual tools will help anchor our analysis, providing clear, visual context to the complex data patterns we have uncovered.

These visualizations serve as indispensable tools for uncovering trends, patterns, and outliers within the dataset. By leveraging these insights, informed decisions can be made and meaningful conclusions regarding the factors influencing used car prices can be derived, thereby enhancing understanding of the dynamics at play in the automotive market.

Data Normalization and Winsorization

Before normalization and Winsorization, the dataset may contain outliers that skew the distribution of numerical features, potentially leading to biased model outcomes. Outliers, represented by extreme values, can significantly impact statistical analysis and machine learning algorithms by exerting undue influence on results.

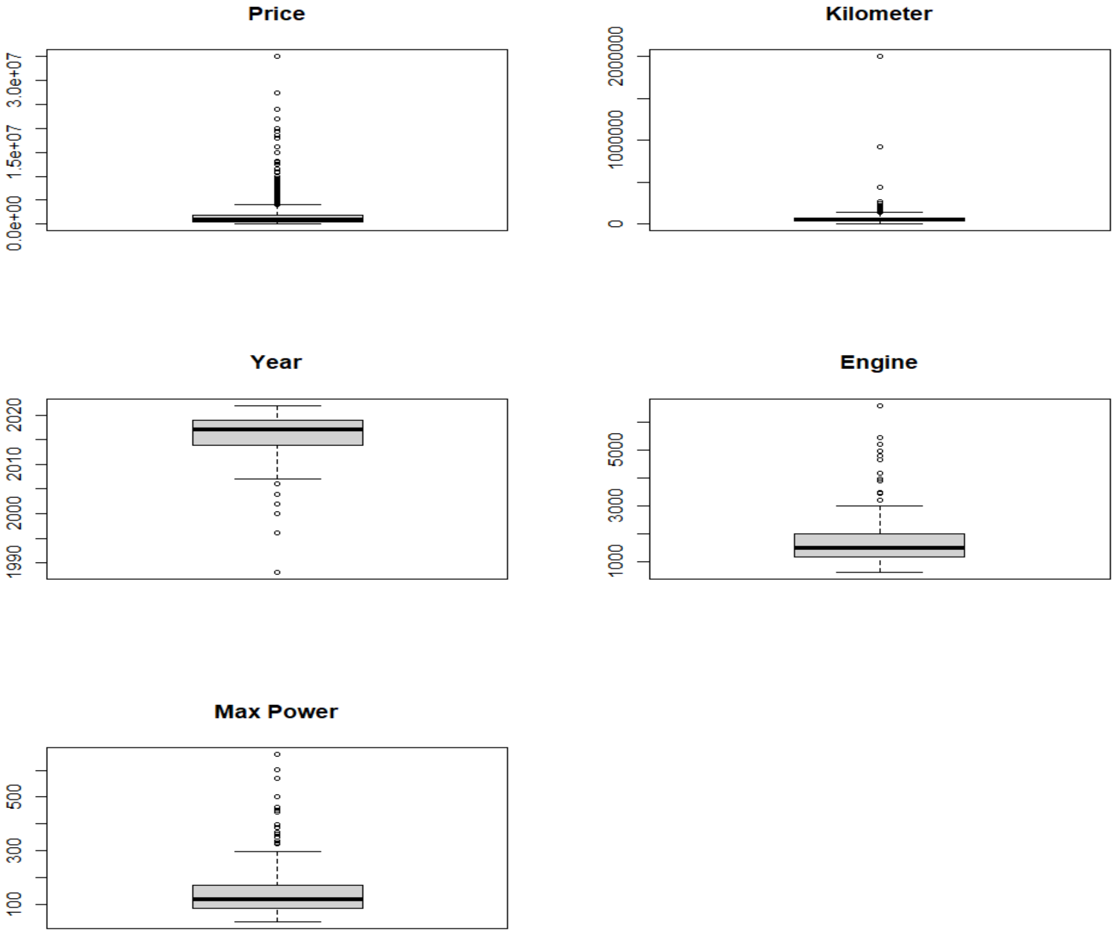


Figure 4: Boxplot of Selling Prices before Winsorization

After applying normalization and Winsorization techniques, outliers are addressed to ensure more robust data analysis. Normalization rescales numerical features to a standard scale, mitigating the influence of outliers and ensuring that all features contribute equally to the analysis. Winsorization involves capping extreme values at specified percentiles, effectively limiting their impact on the dataset. For instance, the Winsorization process might cap the top and bottom 5% of values, replacing them with values corresponding to the 5th and 95th percentiles.

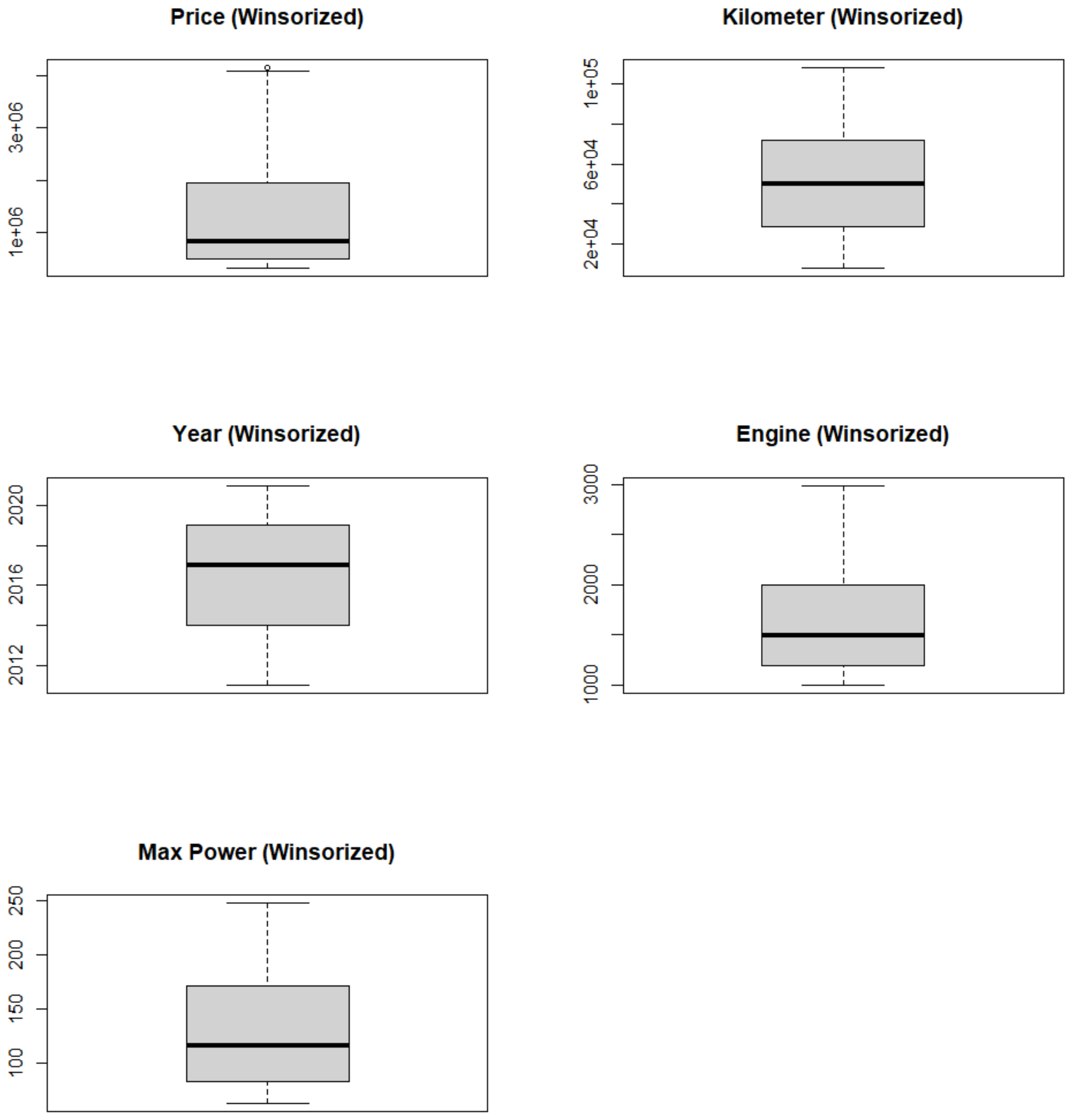


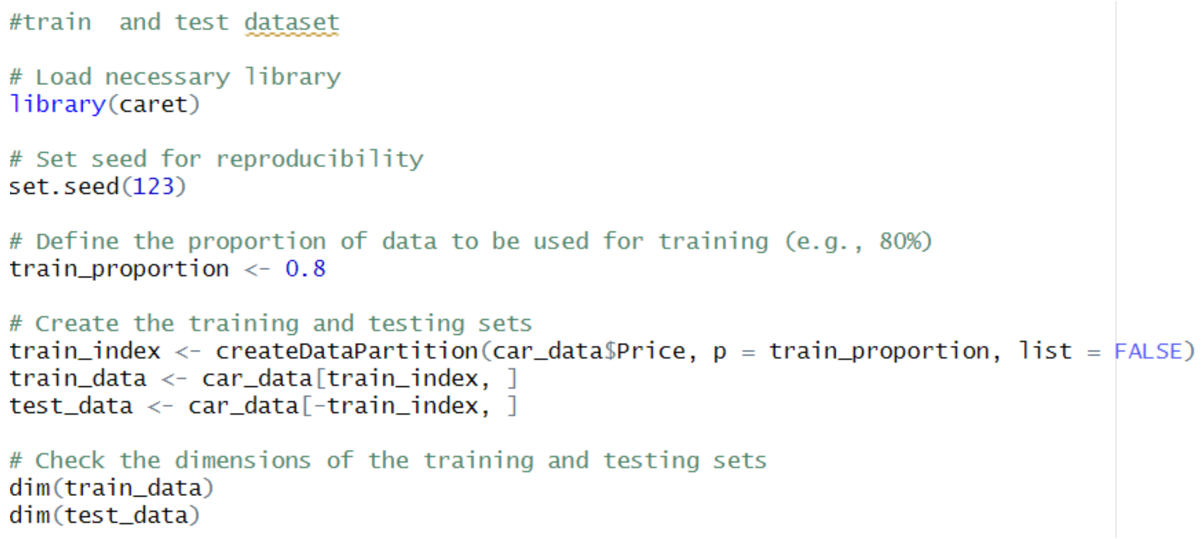
Figure 5: Boxplot of Selling Prices after Winsorization

By normalizing and Winsorizing the data, the stability of the dataset improves, outliers are attenuated, and distributions become more symmetrical. These preprocessing steps enhance the performance and interpretability of statistical models, leading to more reliable insights and predictions in data analysis.

Training and Test Data Segmentation

Introduction to Data Division: In predictive analytics, separating our dataset into training and test sets is critical for evaluating the accuracy of our models. This split allows us to train the model with one subset of the data and test its performance on another subset that it has not seen before. This procedure helps verify that our assessments are impartial and indicative of the model's potential real-world performance.

Data Splitting Methodology:



```
#train  and test dataset

# Load necessary library
library(caret)

# Set seed for reproducibility
set.seed(123)

# Define the proportion of data to be used for training (e.g., 80%)
train_proportion <- 0.8

# Create the training and testing sets
train_index <- createDataPartition(car_data$Price, p = train_proportion, list = FALSE)
train_data <- car_data[train_index, ]
test_data <- car_data[-train_index, ]

# Check the dimensions of the training and testing sets
dim(train_data)
dim(test_data)
```

We utilized the **caret** package in R to facilitate our modeling process, offering tools to efficiently manage predictive modeling workflows. Initially, we established a random seed to ensure that our process is reproducible. This is an essential practice that allows others to replicate our work precisely by generating the same random data splits.

We designated 80% of our dataset for training purposes—a standard practice that offers a good balance between having enough data to train our models robustly and reserving a sufficient portion for validation purposes. Using the **createDataPartition** function, we sorted the dataset into training and testing sets based on the **Price** attribute, ensuring a consistent distribution of this variable across both datasets.

Composition of Training and Testing Sets:

Following our data partitioning strategy, the training dataset ended up with approximately 1649 records, while the testing dataset comprised about 410 records, each containing 23 variables. This distribution verifies that our split aligns with the intended 80/20 ratio. The training set is utilized for model development, allowing the model to learn from the data patterns, whereas the testing set acts as an independent evaluator to assess how well the model performs with new information.

Importance of Separate Training and Testing Phases:

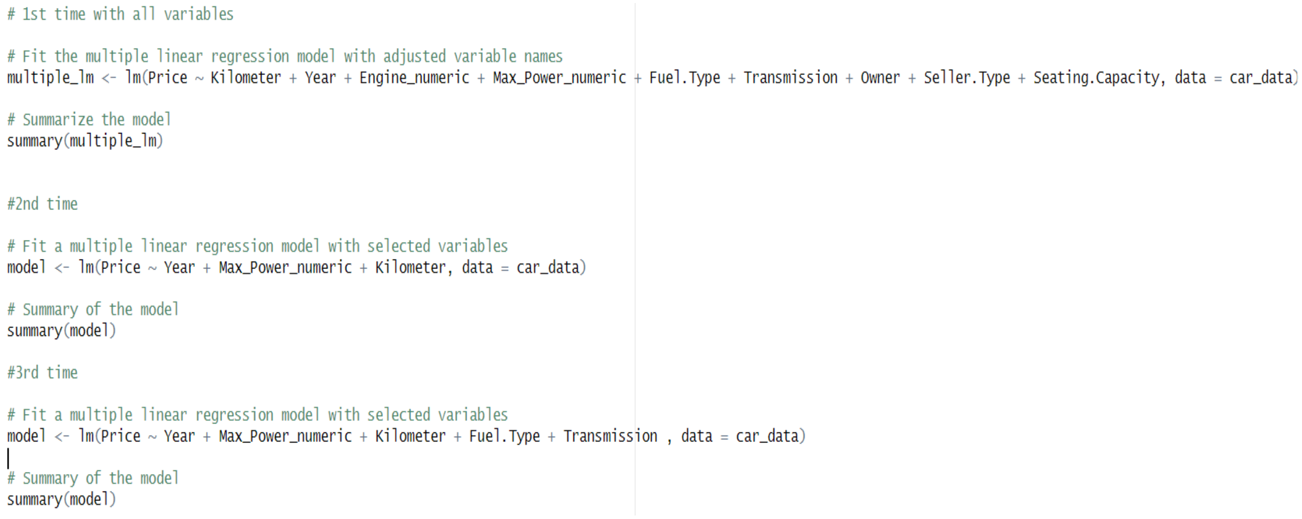
This structured approach of using distinct datasets for training and testing is essential. It ensures that the model is not only tuned with a comprehensive dataset but also validated against an unbiased dataset to test its predictive strength. This practice boosts the credibility of our model's predictive capabilities, providing confidence that our insights and subsequent recommendations are based on rigorously tested models.

This strategy is fundamental to our analytical framework, ensuring that the final model is both accurate and effective in practical applications.

Model:

Multiple Linear Regression (MLR) is a statistical technique that models the relationship between two or more independent variables (predictors) and a dependent variable (response) by fitting a linear equation to observed data. Multiple linear regression is used in scenarios where several variables influence a single outcome and where understanding the impact of each variable is crucial. The reasons for using MLR in this analysis of car prices are:

- MLR can predict the outcome variable (car prices) based on multiple input features, such as year, mileage, engine size, and others. This is particularly useful in pricing models where several factors collectively influence the final price
- MLR helps quantify the influence of each factor on the price. For example, it can estimate how much the price decreases for every kilometer increased in mileage or how much the price goes up with every year newer the car model is.



```
# 1st time with all variables

# Fit the multiple linear regression model with adjusted variable names
multiple_lm <- lm(Price ~ Kilometer + Year + Engine_numeric + Max_Power_numeric + Fuel.Type + Transmission + Owner + Seller.Type + Seating.Capacity, data = car_data)

# Summarize the model
summary(multiple_lm)


#2nd time

# Fit a multiple linear regression model with selected variables
model <- lm(Price ~ Year + Max_Power_numeric + Kilometer, data = car_data)

# Summary of the model
summary(model)

#3rd time

# Fit a multiple linear regression model with selected variables
model <- lm(Price ~ Year + Max_Power_numeric + Kilometer + Fuel.Type + Transmission , data = car_data)
|
# Summary of the model
summary(model)
```

Here we developed three multiple linear regression models with different combinations of predictors.

Model 1: Full model with all available predictors

Model 2: Reduced model focusing on key predictors

Model 3: Reduced model with key predictors plus categorical variables

Model Summaries

Model 1: Comprehensive Model

Predictors: Kilometer, Year, Engine_numeric, Max_Power_numeric, Fuel.Type, Transmission, Owner, Seller.Type, Seating.Capacity

Significant Predictors: Kilometer, Year, Max_Power_numeric, Fuel.Type, Transmission

Adjusted R-squared: 0.8314 , Indicates that 83.14% of the variance in car prices is explained by the model.

Key Findings: This model includes a mix of significant and non-significant predictors, suggesting potential redundancy in some variables.

This model includes a wide range of variables. Despite its complexity, not all included variables significantly predict the car price, as some have high p-values indicating their non-significant impact and these variables are Engine_numeric, Seller.Type, Seating.Capacity, and Owner (Borderline Significance, p-value around 0.071) .These variables showed high p-values, indicating that they are not statistically significant predictors of car price.

Model 2: Simplified Model

Predictors: Year, Max_Power_numeric, Kilometer

Adjusted R-squared: 0.8156 — This model explains 81.56% of the variance in car prices.

Key Findings: Despite its simplicity, this model retains high explanatory power with all predictors showing strong statistical significance.

Model 3: Enhanced Simplified Model

Predictors: Year, Max_Power_numeric, Kilometer, Fuel.Type, Transmission

Adjusted R-squared: 0.8313 — Close to the full model, explaining 83.13% of the variance.

Key Findings: The inclusion of categorical variables (Fuel.Type, Transmission) adds valuable information, enhancing the model's explanatory power without overcomplicating it.

In the third model, two categorical variables (Fuel.Type and Transmission) are added back to the predictors used in the second model. We did this to see if these categorical variables improve the model’s predictive power or explain additional variance in Price.

Results:

The regression model predicts Price using the predictors: Year, Max_Power_numeric, Kilometer, Fuel.Type, and Transmission.

Coefficients Interpretation:

Intercept (β_0): This is the estimated price when all other predictors are at zero, which is not a practical scenario in this context, but mathematically, it's where the regression line crosses the Y-axis.

Year (β_1 = 92,400): For each additional year, the price of the car increases by approximately 92,400, assuming other factors remain constant. This suggests newer cars are more valuable.

Max_Power_numeric (β_2 = 16,410): For each additional unit increase in max power, the price increases by about 16,410, reflecting that more powerful cars are priced higher.

Kilometer (β_3 = -6,397): For each additional kilometer driven, the car price decreases by approximately 6,397. This indicates that the more kilometers driven by the car, the less car value.

Fuel.Type (β_4 = -67,050): This coefficient likely compares different fuel types against a baseline (diesel). Cars with this fuel type (petrol) are priced 67,050 lower on average than the baseline, showing fuel type's impact on pricing.

Transmission (β_5 = -302,600): This suggests that cars with this type of transmission (manual, baseline automatic is) are significantly cheaper by 302,600 compared to cars with the baseline transmission type.

Model Fit:

Residual Standard Error (515,400): Represents the typical deviation of the observed values from the regression line. It's quite large, indicating variability in car prices not captured by the model.

R-squared (0.8317): About 83.17% of the variability in car prices is explained by the model, which is fairly high, suggesting a good fit.

Adjusted R-squared (0.8313): This is adjusted for the number of predictors in the model and is very close to the R-squared, indicating the model isn't overly penalized by unnecessary predictors.

F-Test:

The F-statistic is used in the context of testing the overall significance of a regression model. It tests the null hypothesis that all regression coefficients are equal to zero, which implies that the predictors have no linear relationship with the outcome variable.

Components of the F-statistic in the Model:

F-statistic value: 1950

Degrees of Freedom for the Model: 5

Degrees of Freedom for the Residuals: 1973

p-value: < 2.2e-16

F-statistic Value (1950): This value measures the ratio of two variances,the variance explained by the model and the variance of errors.

A high F-value (as in model) suggests that the model explains a significant amount of variability in the dependent variable, much more than what would be expected by chance.

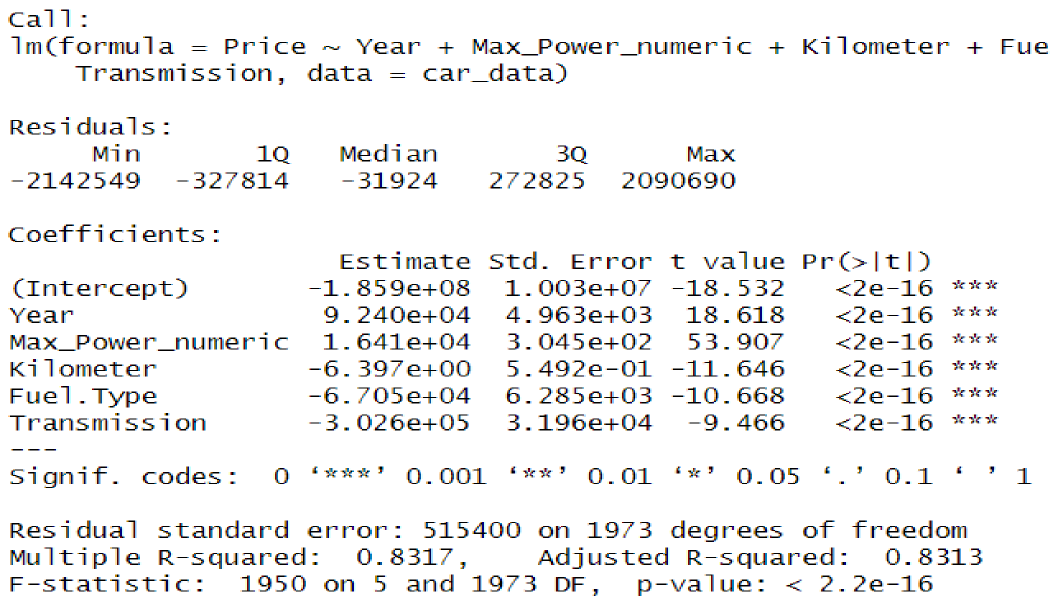
Degrees of Freedom for the Model (5): This represents the number of predictors in the model. It is crucial as it impacts how the MSR is calculated.

Degrees of Freedom for the Residuals (1973): This represents the total observations minus the number of predictors minus one. It is used to calculate the MSE.

p-value (< 2.2e-16): The p-value associated with the F-statistic tests the null hypothesis that all coefficients in the model are zero (i.e., no predictor has an effect on the outcome).

A very small p-value (in the model) indicates strong evidence against the null hypothesis. Thus, we reject the null hypothesis, concluding that at least some of the predictors have a non-zero effect on the dependent variable.

A significant F-statistic confirms that the model is statistically valid, implying that it effectively captures the relationship between predictors and the dependent variable.



```
Call:
lm(formula = Price ~ Year + Max_Power_numeric + Kilometer + Fuel.Type + 
    Transmission, data = car_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2142549  -327814   -31924   272825  2090690 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1.859e+08  1.003e+07 -18.532   <2e-16 ***
Year               9.240e+04  4.963e+03  18.618   <2e-16 ***
Max_Power_numeric  1.641e+04  3.045e+02  53.907   <2e-16 ***
Kilometer         -6.397e+00  5.492e-01 -11.646   <2e-16 ***
Fuel.Type         -6.705e+04  6.285e+03 -10.668   <2e-16 ***
Transmission      -3.026e+05  3.196e+04  -9.466   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 515400 on 1973 degrees of freedom
Multiple R-squared:  0.8317,	Adjusted R-squared:  0.8313 
F-statistic:  1950 on 5 and 1973 DF,  p-value: < 2.2e-16
```

Assumption of Multilinear Regression:

Linearity:

The relationship between the predictors and the dependent variable should be linear.

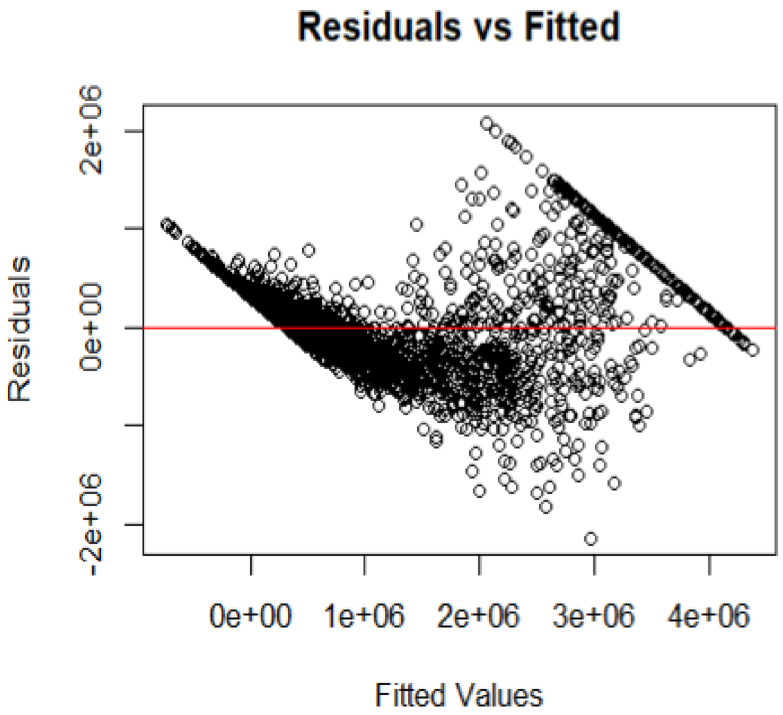


Figure 6 : Residuals vs Fitted

The Residuals vs. Fitted plot shows a pattern where residuals fan out as the fitted values increase, creating a cone-like or funnel shape. This pattern suggests a couple of potential issues with the linear regression model's assumptions

The assumption of linearity is violated because the residuals should be randomly scattered around the horizontal axis (the red line at 0), without any systematic pattern. Here, the residuals appear to systematically increase or decrease over the range of fitted values, suggesting that a linear model may not be the best fit for the data.

Partial Regression (Added Variable) Plots:

Year Plot:

There's a linear trend indicating that the variable 'Year' is positively associated with the price, controlling for the other variables in the model. The pattern suggests that as cars are newer (with a higher 'Year'), the price tends to be higher.

Max_Power_numeric Plot:

This plot shows a strong, positive linear relationship between 'Max_Power_numeric' and the price, indicating that cars with more power are associated with higher prices, controlling for the other variables. The linearity of the relationship is a good sign for the validity of including 'Max_Power_numeric' in a linear regression model.

Kilometer Plot:

The relationship between 'Kilometer' and the price seems to be negative, suggesting that as the number of kilometers increases (suggesting a used or older car), the price tends to decrease. The linear relationship is less clear compared to 'Max_Power_numeric', but the negative trend aligns with expectations.

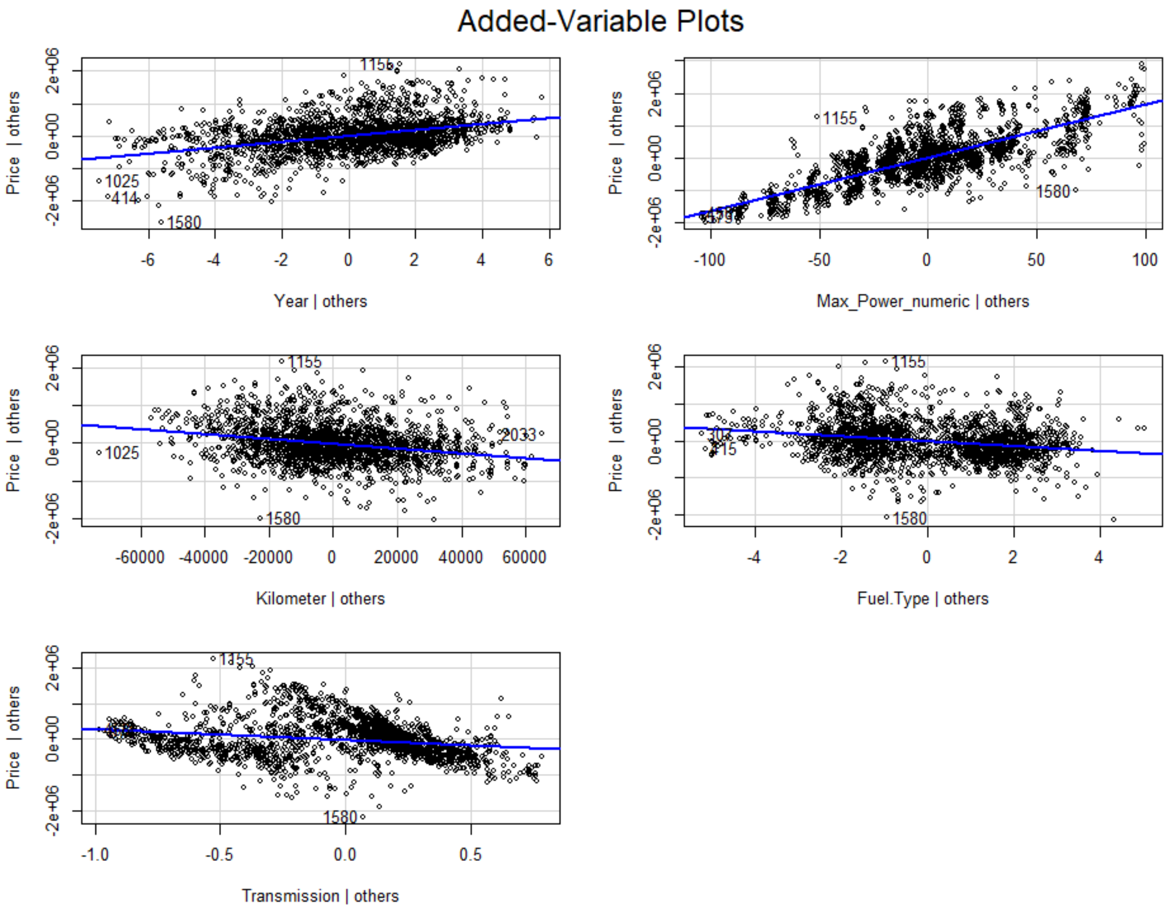


Figure 7: Variable plots

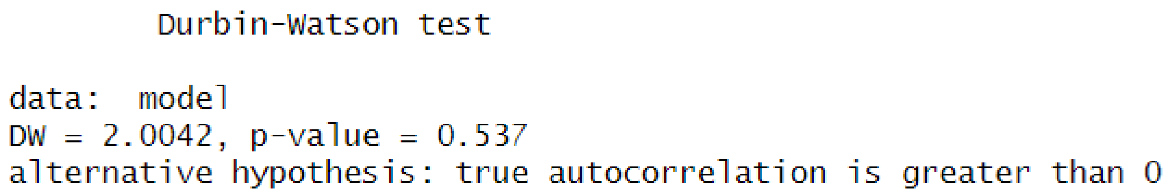
Homoscedasticity:

The widening spread of residuals as the fitted values increase is indicative of heteroscedasticity. This means the variance of the error terms is not constant across all levels of the independent variables. Homoscedasticity (constant variance) is an assumption of linear regression, and this plot suggests that assumption is not met

Independence of Errors

Residuals (errors) should be independent of each other, which is crucial for getting unbiased estimates of standard errors and confidence intervals.

The independence of the residuals was assessed using the Durbin-Watson statistic, which yielded a value of 2.0042. This value is very close to the benchmark value of 2, suggesting that there is no substantial autocorrelation in the residuals of our regression model (p-value = 0.537). Therefore, we do not have evidence to suggest that the residuals are dependent on one another, supporting one of the key assumptions of the multiple linear regression model. The p-value associated with the DW test (p-value = 0.537) is not below the common alpha level of 0.05, so we fail to reject the null hypothesis of no autocorrelation.



```
        Durbin-Watson test

data:  model
DW = 2.0042, p-value = 0.537
alternative hypothesis: true autocorrelation is greater than 0
```

Normality of Residuals:

- Q-Q Plot of Residuals:

  The Q-Q (quantile-quantile) plot compares the quantiles of the residuals to the quantiles of a normal distribution, where residuals are normally distributed, the points should fall approximately along the reference line (red line).

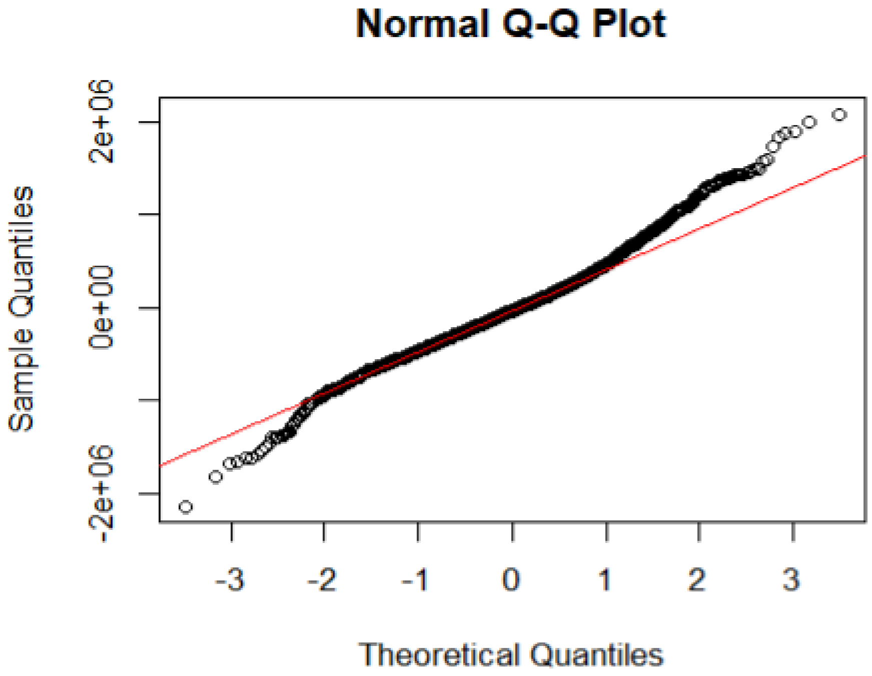


Figure 8: Normal Q-Q plot

Center of the Plot: The points in the center of the plot (around the 0 of the theoretical quantiles) seem to follow the red line fairly closely, which indicates that the residuals are approximately normally distributed in the middle of the distribution.

Tails of the Plot: The points at the extremes (tails) of the distribution deviate from the line. This suggests that the residuals have heavier tails than the normal distribution. In other words, there are more extreme values (both low and high) than what would be expected in a normal distribution.

The deviations in the tails might suggest potential issues with the presence of non-linearity that hasn't been captured by the model. Given the central tendency's conformity to normality and considering the robustness of the normality assumption in larger samples, these deviations are not necessarily a cause for concern. So we can declare the Normality assumption is met.

Multicollinearity

Predictors should not be too highly correlated with each other as it can inflate the variance of the coefficient estimates and make the model unstable.

- Pearson Correlation Matrix:

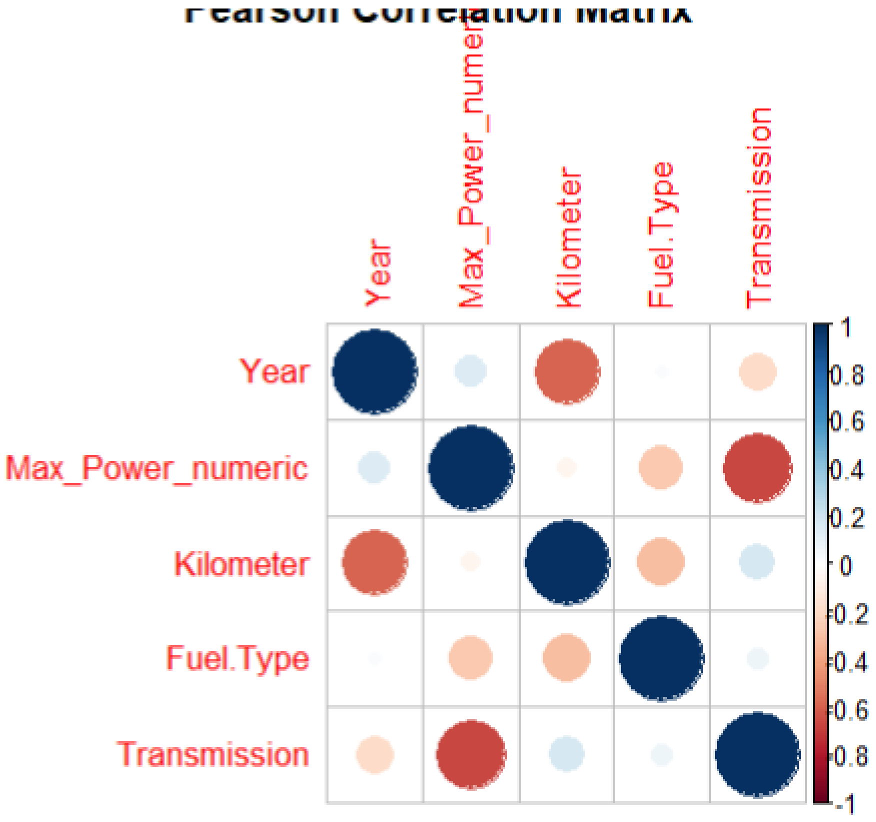
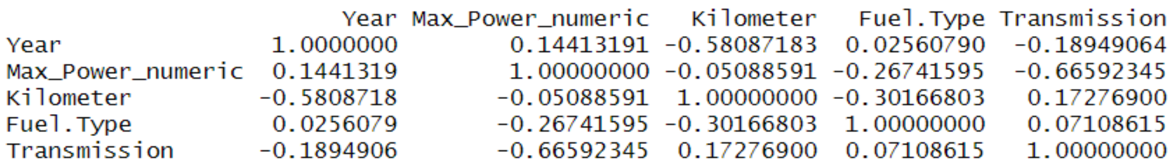


Figure 9: Pearson Correlation Matrix

Year and Max_Power_numeric: There is a positive correlation, indicating that newer car models tend to have more power, which is plausible due to advancements in car technology.

Year and Kilometer: There appears to be a negative correlation, suggesting that newer cars have fewer kilometers on them, which makes sense because older cars typically have been driven more.



```
                        Year Max_Power_numeric   Kilometer   Fuel.Type Transmission
Year               1.0000000        0.14413191 -0.58087183  0.02560790  -0.18949064
Max_Power_numeric  0.1441319        1.00000000 -0.05088591 -0.26741595  -0.66592345
Kilometer         -0.5808718       -0.05088591  1.00000000 -0.30166803   0.17276900
Fuel.Type          0.0256079       -0.26741595 -0.30166803  1.00000000   0.07108615
Transmission      -0.1894906       -0.66592345  0.17276900  0.07108615   1.00000000
```

.

There is a moderate negative correlation between Year and Kilometer (0.58), which is expected as newer cars tend to have fewer kilometers.

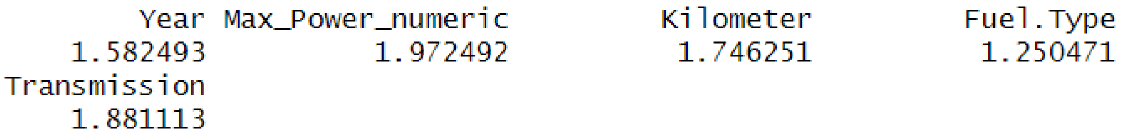
Max_Power_numeric and Transmission have a significant negative correlation (0.67), suggesting that cars with higher power are less likely to have a certain type of transmission (manual, we assumed that higher power is associated with automatic transmission).

Fuel.Type has a moderate negative correlation with Kilometer (0.30) and Max_Power_numeric (0.27), which might suggest that fuel type is related to both the age and power of the car.

No other variables show a high degree of correlation that would typically be concerning for multicollinearity (generally, a threshold of |r|>0.7 or >0.8 (|r|>0.8 is considered high)

Variance Inflation Factor (VIF): VIF values greater than 10 (some sources say 5) are indicative of multicollinearity issues.

All VIF values are below the common threshold of 5, which is generally considered a sign that multicollinearity may not be a concern



```
             Year Max_Power_numeric         Kilometer         Fuel.Type 
         1.582493          1.972492          1.746251          1.250471 
     Transmission 
         1.881113 
```

From the given results, we can conclude that there is no significant multicollinearity affecting the regression model. The predictors do not show excessively high correlations with one another, and the VIF values are within acceptable limits. This implies that each predictor is providing unique information.

.

Absence of High Leverage Points:

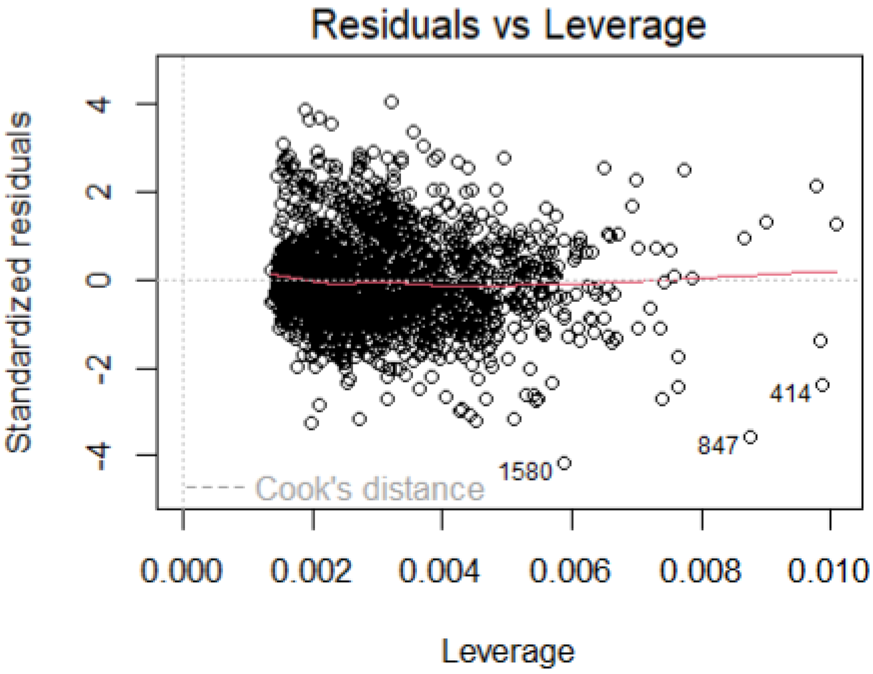


Figure 10: Residuals vs Leverage Plot

In evaluating the influence of individual data points on our regression model, we constructed a Residuals vs. Leverage plot, complemented by Cook's Distance contours. The analysis indicated that while there are observations with high leverage, such as points 414 and 847, these do not correspond with large residuals and hence do not significantly distort the overall model fit. However, observations like point 1580 warrant closer examination due to their proximity to the Cook's Distance threshold, which suggests they could be influential.

This gives us a balanced assessment, highlighting potential areas for further investigation without necessarily suggesting that the model is incorrect or invalid.

Cook’s Distance:

Cook's Distance Values: Observations with a Cook's Distance larger than 1 are generally considered highly influential.

Influence on the Model: Most of the data points in your plot have a Cook's Distance close to zero, indicating that they have little influence on the model. A few spikes exceed the common threshold, suggesting these points could be influential.

In the process of assessing the influence of individual data points on our linear regression model, we examined Cook's Distance for each observation. The Cook's Distance plot revealed that the vast majority of points have minimal influence on the model, with Cook's

Distance values close to zero. But, there are a select few observations, identifiable by their higher peaks in the plot, that show a greater level of influence. While not exceeding the common threshold of 1, these points may still merit additional scrutiny to ensure that our

The model is not overly sensitive to these particular data points. No single observation appears to be excessively influential to the point of distorting model predictions significantly, but the presence of these higher Cook's Distance values suggests that a review of potential outliers or leverage points could be beneficial to confirm the robustness of our model.

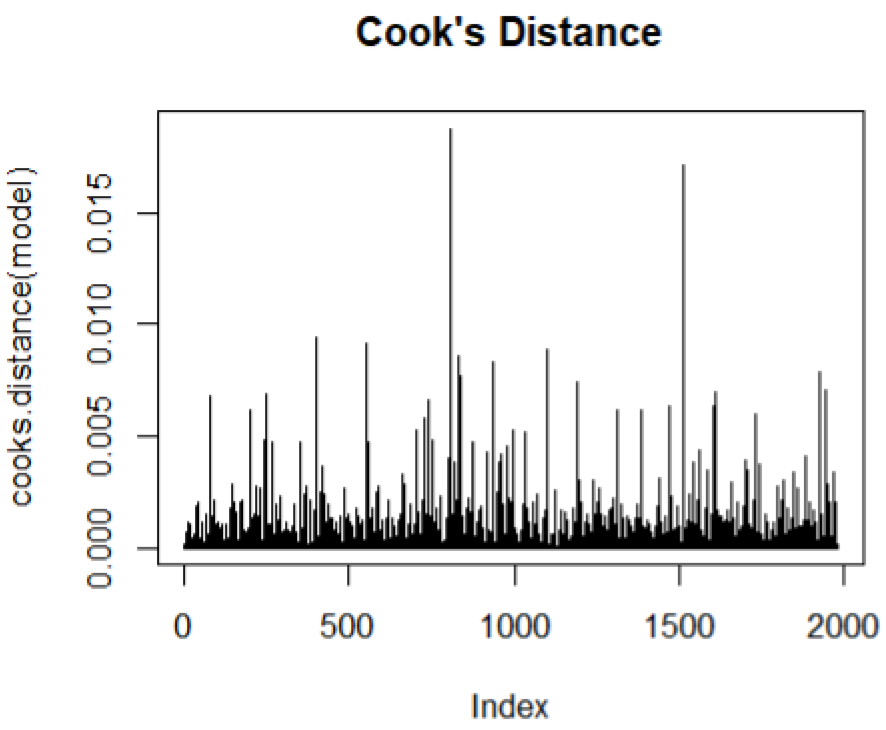
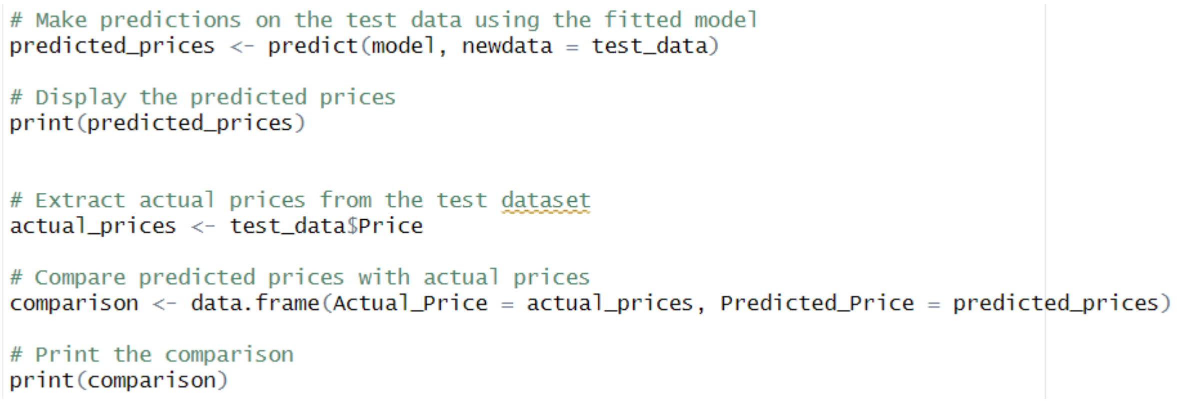


Figure 11: Cook's Distance

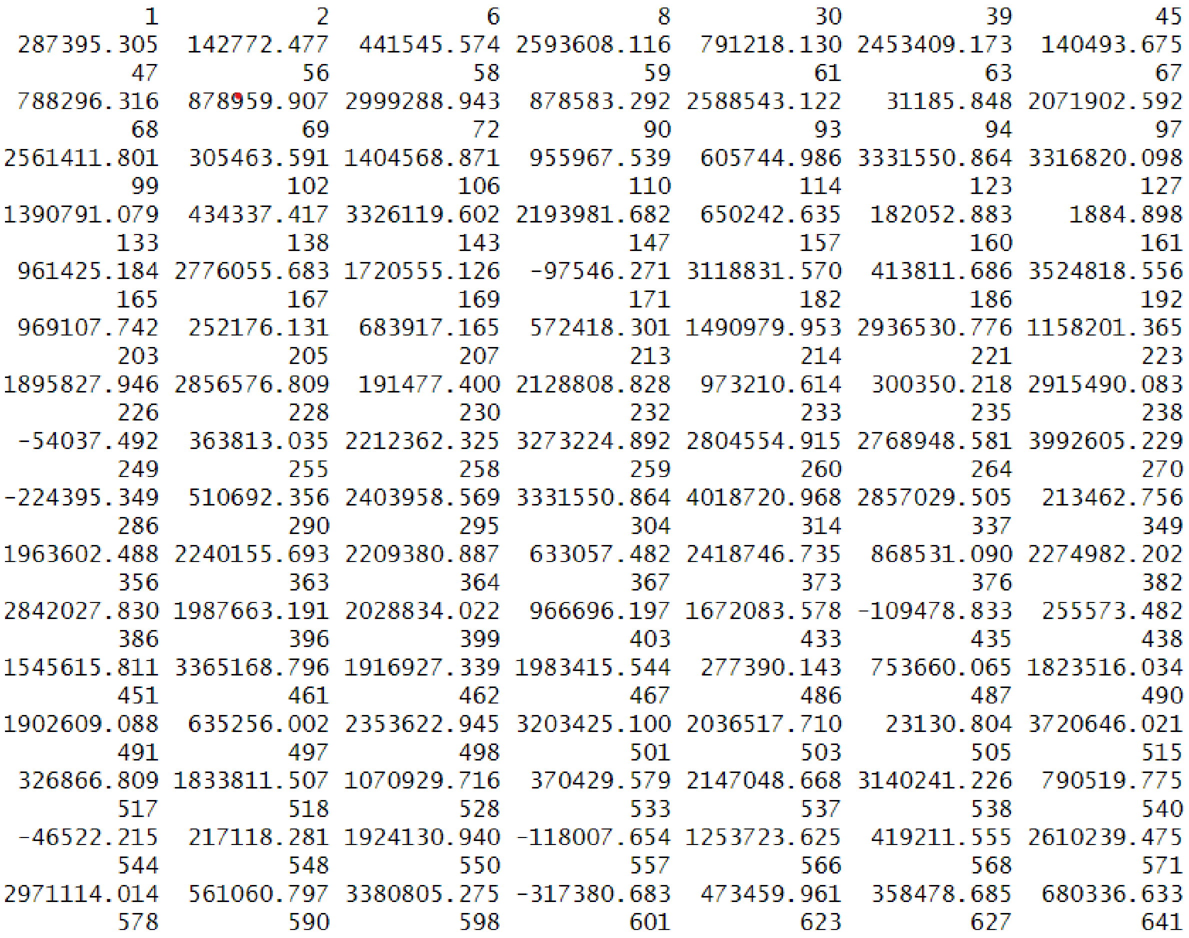
Testing:

Now , we have tested the unseen part of the data with the trained model of linear regression and predicted prices. And we also did comparison to the actual prices and predicted prices to see how well the model is performing.

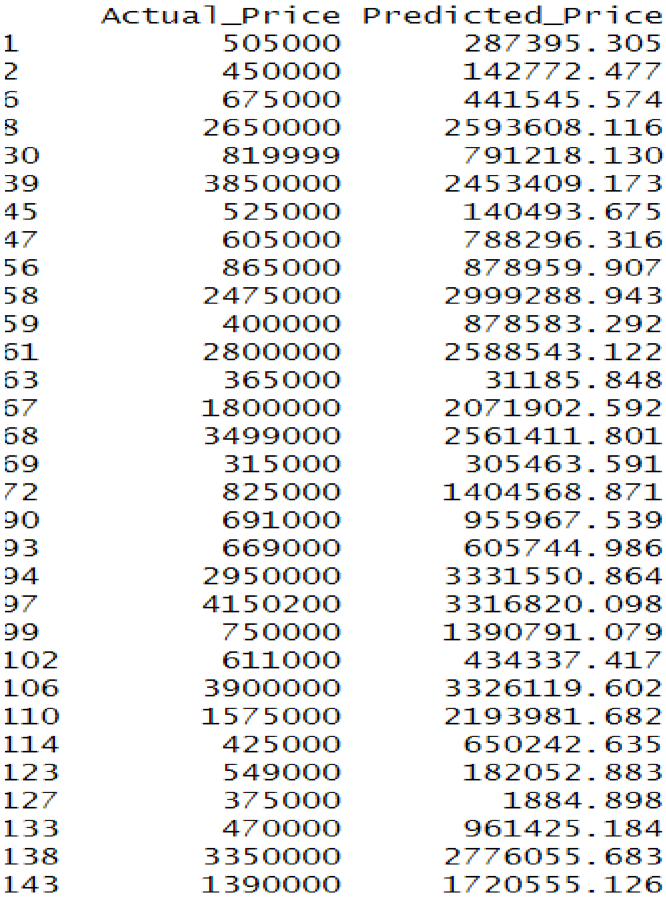


```
# Make predictions on the test data using the fitted model
predicted_prices <- predict(model, newdata = test_data)

# Display the predicted prices
print(predicted_prices)


# Extract actual prices from the test dataset
actual_prices <- test_data$Price

# Compare predicted prices with actual prices
comparison <- data.frame(Actual_Price = actual_prices, Predicted_Price = predicted_prices)

# Print the comparison
print(comparison)
```

these are the values we got after predicting prices:-

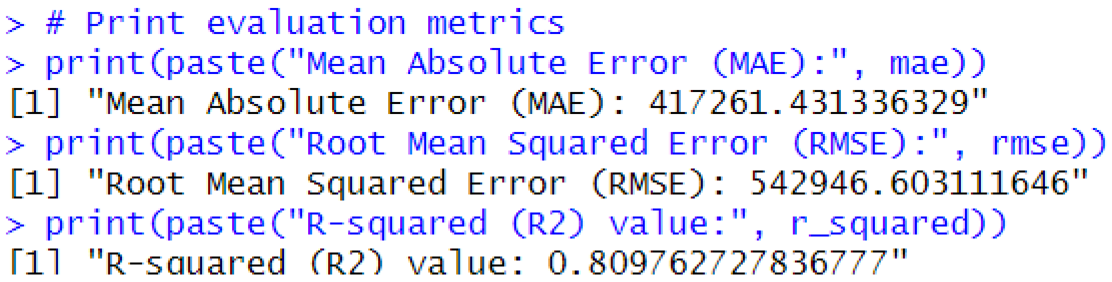


```
          1           2           6           8          30          39          45 
 287395.305  142772.477  441545.574 2593608.116  791218.130 2453409.173  140493.675 
         47          56          58          59          61          63          67 
 788296.316  878959.907 2999288.943  878583.292 2588543.122   31185.848 2071902.592 
         68          69          72          90          93          94          97 
2561411.801  305463.591 1404568.871  955967.539  605744.986 3331550.864 3316820.098 
         99         102         106         110         114         123         127 
1390791.079  434337.417 3326119.602 2193981.682  650242.635  182052.883    1884.898 
        133         138         143         147         157         160         161 
 961425.184 2776055.683 1720555.126  -97546.271 3118831.570  413811.686 3524818.556 
        165         167         169         171         182         186         192 
 969107.742  252176.131  683917.165  572418.301 1490979.953 2936530.776 1158201.365 
        203         205         207         213         214         221         223 
1895827.946 2856576.809  191477.400 2128808.828  973210.614  300350.218 2915490.083 
        226         228         230         232         233         235         238 
 -54037.492  363813.035 2212362.325 3273224.892 2804554.915 2768948.581 3992605.229 
        249         255         258         259         260         264         270 
-224395.349  510692.356 2403958.569 3331550.864 4018720.968 2857029.505  213462.756 
        286         290         295         304         314         337         349 
1963602.488 2240155.693 2209380.887  633057.482 2418746.735  868531.090 2274982.202 
        356         363         364         367         373         376         382 
2842027.830 1987663.191 2028834.022  966696.197 1672083.578 -109478.833  255573.482 
        386         396         399         403         433         435         438 
1545615.811 3365168.796 1916927.339 1983415.544  277390.143  753660.065 1823516.034 
        451         461         462         467         486         487         490 
1902609.088  635256.002 2353622.945 3203425.100 2036517.710   23130.804 3720646.021 
        491         497         498         501         503         505         515 
 326866.809 1833811.507 1070929.716  370429.579 2147048.668 3140241.226  790519.775 
        517         518         528         533         537         538         540 
 -46522.215  217118.281 1924130.940 -118007.654 1253723.625  419211.555 2610239.475 
        544         548         550         557         566         568         571 
2971114.014  561060.797 3380805.275 -317380.683  473459.961  358478.685  680336.633 
        578         590         598         601         623         627         641 
```



```
    Actual_Price Predicted_Price
1         505000      287395.305
2         450000      142772.477
6         675000      441545.574
8        2650000     2593608.116
30        819999      791218.130
39       3850000     2453409.173
45        525000      140493.675
47        605000      788296.316
56        865000      878959.907
58       2475000     2999288.943
59        400000      878583.292
61       2800000     2588543.122
63        365000       31185.848
67       1800000     2071902.592
68       3499000     2561411.801
69        315000      305463.591
72        825000     1404568.871
90        691000      955967.539
93        669000      605744.986
94       2950000     3331550.864
97       4150200     3316820.098
99        750000     1390791.079
102       611000      434337.417
106      3900000     3326119.602
110      1575000     2193981.682
114       425000      650242.635
123       549000      182052.883
127       375000        1884.898
133       470000      961425.184
138      3350000     2776055.683
143      1390000     1720555.126
```

To evaluate the performance of your model,we measured the difference between the predicted prices and the actual prices. Common metrics for this purpose include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE).



```
> # Print evaluation metrics
> print(paste("Mean Absolute Error (MAE):", mae))
[1] "Mean Absolute Error (MAE): 417261.431336329"
> print(paste("Root Mean Squared Error (RMSE):", rmse))
[1] "Root Mean Squared Error (RMSE): 542946.603111646"
> print(paste("R-squared (R2) value:", r_squared))
[1] "R-squared (R2) value: 0.809762727836777"
```

Mean Absolute Error (MAE): 417,261.43

An MAE of 417,261.43 suggests that, on average, the model's predictions are about 417,261.43 units (likely the currency in which the car prices are measured) away from the actual price.

Root Mean Squared Error (RMSE): 542,946.60

An RMSE of 542,946.60, which is higher than the MAE, indicates that there are some larger errors between the predicted and actual values.

R-squared (R2) Value: 0.8098

The R-squared value is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination.

An R-squared of 0.8098 suggests that approximately 80.98% of the variation in the car price can be explained by the model's inputs. This is a relatively high R-squared value, implying good model performance.

- The model's predictive performance, highlighting the high MAE and RMSE values, indicating substantial errors in predictions.
- Despite the large errors, the model still captures a significant portion of the variance in the dependent variable, as indicated by the relatively high R-squared value.

Analysis of the Project

Throughout the course of this project, we aimed to develop a regression model capable of predicting car prices based on various predictor variables. We undertook a comprehensive approach that included model building, diagnostic testing, and evaluation. Here is an overall conclusion that encapsulates the entirety of the work performed:

Model Development and Diagnostics:

We began by fitting a multiple linear regression model using a set of predictors deemed relevant to car prices. The initial model included variables such as Year, Kilometer, Max_Power_numeric, Fuel.Type, and Transmission. Diagnostic plots and statistical tests were employed to assess the model's underlying assumptions:

Linearity: Partial regression plots indicated a generally linear relationship between each of the predictors and the response variable, supporting the linearity assumption of the model.

Independence of Residuals: The Durbin-Watson test produced a statistic of 2.0042, which suggests no significant autocorrelation in the residuals.

Normality of Residuals: The Normal Q-Q Plot revealed that while residuals were normally distributed in the center of the distribution, there were deviations in the tails, hinting at the presence of outliers or non-normality in the extreme values.

Homoscedasticity: The Residuals vs. Fitted Values plot indicated potential heteroscedasticity, as evidenced by a cone-shaped pattern, which could imply that variance stabilizing transformations might be needed.

Multicollinearity: Pearson correlation coefficients and Variance Inflation Factor (VIF) scores were within acceptable ranges, indicating that multicollinearity was not a significant issue in the model.

Influential Points: Cook's Distance measures pointed out a few potentially influential cases, but not to an extent that would unduly compromise the model.

Model Performance:

Upon prediction of car prices using the test data, the model demonstrated good predictive power, as indicated by an R-squared value of approximately 0.81, suggesting that about 81% of the variability in car prices was explained by the model. The MAE and RMSE values, while sizable, provided a quantitative measure of the model's average and individual errors, respectively.

Conclusion:

The regression model built during this project serves as a substantial predictor of car prices, capturing a significant portion of the variance within the dataset. However, certain limitations have been identified, such as potential outliers affecting the predictions and indications of heteroscedasticity, which could be addressed in further iterations of the model. The model's current performance suggests utility in a real-world application, with the proviso that further enhancements and validations may be warranted to refine its predictive accuracy. Future work may involve exploring non-linear models, robust regression techniques, or machine learning algorithms to handle the complex patterns in the data that a linear model may not adequately capture.