CAR PRICE PREDICTION

BY:

OLOLADE AKINSANOLA

COURSE CODE:

STAT 408

INSTRUCTOR'S NAME:

DR MATT STAURT

UNIVERSITY NAME:

LOYOLA UNIVERSITY CHICAGO

DATE:

DECEMBER 2024

\

ABSTRACT

This report analyzes a dataset related to car price prediction using linear regression. Through a comprehensive data exploration and preprocessing approach, significant features are identified, and a model is built to evaluate relationships between variables. The report discusses findings, critiques methods, and highlights areas for future improvement, offering practical insights into predictive analytics in the automotive industry.

TABLE OF CONTENTS

INTRODUCTION

This project focuses on predicting car prices based on a variety of explanatory variables. The dataset used for this analysis was obtained from Kaggle, featuring 19,237 observations and 18 columns (https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge/data). After several stages of cleaning, processing, and filtering, it was reduced to 6,155 observations and 16 variables. A stratified sampling method was applied to extract a representative sample of 1,500 observations.

The target variable for this project is the price of the car, while the explanatory variables include a mix of **numerical** and **categorical** features:

- **Numerical variables**: Mileage, Production year, Levy, Engine volume, Cylinders, and Airbags.
- **Categorical variables**: Color, Model, Manufacturer, Gearbox type, Wheels, Drive wheel, Category, Fuel type, Leather interior, and Doors.

The dataset offers insights into various aspects of a vehicle, such as its physical characteristics, manufacturing details, and design. For example:

- **Mileage** and **Production year** indicate vehicle usage and age.
- **Engine volume** and **Cylinders** represent the car's performance capabilities.
- **Leather interior** and **Airbags** suggest the inclusion of luxury and safety features, which may affect the price.

The dataset also provides an opportunity to analyze regional or market-specific pricing trends. Certain variables, such as Manufacturer, Model, and Category, may reflect preferences unique to specific markets. For instance, some manufacturers may command premium pricing due to brand reputation, while certain car models might have higher resale values because of their popularity or reliability. Features like Fuel type and Drive wheel (e.g., front-wheel drive or all-wheel drive) help assess how functional attributes impact pricing.

The presence of categorical variables, such as Gearbox type and Leather interior, enables the study of consumer preferences for comfort and convenience, which are often correlated with vehicle price. The combination of numerical and categorical variables, along with their potential interactions, highlights the dataset's complexity.

The dataset also allows for an exploration of how different market segments and consumer preferences influence car prices. For instance, variables like Category (e.g., sedan, SUV, etc.)

and Fuel type (e.g., gasoline, diesel, electric) can reflect consumer priorities, such as environmental concerns or preferences for fuel efficiency. Similarly, attributes like Doors and Airbags may influence perceptions of safety and practicality, potentially affecting pricing for family-oriented or safety-conscious buyers.

By examining these features, the analysis can uncover how specific car attributes align with consumer demand, providing deeper insights into market segmentation. This complexity makes the dataset particularly rich, offering a wide range of perspectives on how diverse factors contribute to determining car prices across various types of vehicles.

Importance and Relevance of the Project:

The primary aim of this analysis is to assess the impact of various explanatory variables on car price, examining the relationships between these factors and determining whether their coefficients are statistically significant. This project is particularly valuable as it provides actionable insights for both car sellers and consumers, highlighting the key factors that influence pricing.

By understanding how variables such as mileage, engine volume, production year, and fuel type affect car prices, sellers can set more accurate and competitive prices. Consumers, on the other hand, gain a better understanding of what drives pricing in the market. This analysis offers practical insights that can improve pricing strategies and decision-making within the automotive industry.

Additionally, the analysis explores how different features—such as gearbox type, drive wheel configuration, leather interior, and airbag presence—affect a car's market value. These factors help sellers optimize pricing by adjusting for variables that consumers value most. For instance, cars with higher engine volumes or those featuring leather interiors may command higher prices, while features like fuel type and drive wheel configuration could appeal to specific consumer preferences. Understanding these relationships enables more data-driven and precise pricing strategies. For buyers, this analysis provides the tools to make more informed purchasing decisions. This is especially valuable in a market where pricing can sometimes appear arbitrary or driven by less transparent factors. With insights from this analysis, consumers can better evaluate whether a car is priced fairly in relation to its features, mileage, and other key characteristics, ultimately fostering a more equitable and transparent market.

DATA EXPLORATION AND PREPARATION

Loading Data

To begin the analysis, the necessary libraries were loaded to facilitate data manipulation, visualization, and statistical testing. The dataset, which is the core of this study, was obtained from a CSV file titled "Data/CarPrice.csv". This dataset includes comprehensive information about car prices and various features, such as mileage, engine volume, manufacturing details, and other vehicle attributes.

The dataset originally contained 19,237 observations and 17 columns. It was subsequently cleaned and filtered to include 5,635 observations and 16 variables. A representative sample of 1,200 observations was extracted using a stratified sampling method to ensure balanced representation across different car categories and features.

This step in the analysis prepares the dataset for exploration, model building, and further analysis, enabling a deeper understanding of the factors influencing car prices.

Preprocessing Steps

Data Cleaning and Preprocessing

The data cleaning and preprocessing phase was crucial to ensure the dataset was ready for analysis. This involved addressing inconsistencies, standardizing formats, and handling missing data. Here’s a detailed breakdown of each step:

Clean Numeric Columns

Numeric columns required careful handling to maintain consistency and accuracy:

1. **Mileage**: Initially, mileage values were recorded as strings with a " km" suffix. To transform these values into numeric form, the suffix was removed using gsub(" km", "", Mileage), and then converted to a numeric type using as.numeric(). This transformation allowed for proper numerical comparisons and analysis.

2. **Levy**: The Levy column contained entries marked with a "-" indicating missing data, which were replaced with NA using gsub("-", NA, Levy). This step facilitated accurate handling of missing values and consistent data representation.

3. **Engine Volume**: The Engine.volume column included a " Turbo" suffix that needed to be stripped to reflect true engine capacity. The gsub(" Turbo", "", Engine.volume)

function was used to remove the extraneous text, converting the column into a numeric format for analysis.

4. **Doors**: The Doors column needed standardization due to varying formats. The sub("-.*", "", .) function removed additional text after the dash. Then, ifelse(. == "<5", "4", .) recoded "<5" to "4," and the column was converted to a numeric type.

Convert Relevant Columns to Numeric

Columns such as Price, Airbags, and Cylinders were converted to numeric to facilitate statistical analysis. This step ensured uniformity across the dataset and allowed for meaningful calculations and visualizations.

2. Standardizing Categorical Variables

Categorical variables exhibited inconsistencies that were addressed to improve dataset quality and analysis:

Standardize the Wheel Column

The Wheel column needed simplification:

- "Right-hand drive" was recoded to "Right."
- "Left wheel" was recoded to "Left" for clarity and consistency.

Standardize the Color Column

Color values were grouped into broader categories for easier interpretation:

- Carnelian red was grouped as Red.
- Sky blue was grouped as Blue.
- Colors were further categorized into "Light," "Dark," "Warm," and "Cool" based on common perceptions.

Standardize the Drive.wheels Column

The terminology in Drive.wheels was simplified:

- "4x4" was recoded to "AWD" (All-Wheel Drive).

- "Front" and "Rear" were maintained but standardized for analysis.

3. Handling Missing Values

Missing values presented a challenge that required tailored imputation methods:

Impute Missing Numeric Columns with Median

Numeric columns like Price and Levy had missing values filled using their respective medians. This method preserved the central tendency of the data, minimizing the impact of missing values on analysis.

Impute Missing Categorical Columns with Mode

For categorical columns such as Manufacturer and Model, missing values were filled with the most frequent category (mode). This approach maintained consistency and ensured that missing entries did not skew the analysis.

4. Feature Engineering

New features were introduced to enhance the dataset's utility:

1. **Age**: A new variable Age was calculated as 2024 - Prod..year to represent the age of each car. This transformation added context to the data, influencing price and condition assessments.

5. Recode and Group Categorical Variables

Categorical variables were further refined to simplify analysis:

Reorganize Color

Colors were grouped into four categories:

- White, Silver, Grey and Beige were categorized as "Light."
- Black and Brown were categorized as "Dark."
- Red, Orange, Yellow and Golden were categorized as "Warm."

- Blue, Green, Purple and Pink were categorized as "Cool."

Reorganize Fuel.type

Fuel types were consolidated into three groups:

- Hybrid for hybrid and plug-in hybrid cars.
- Conventional for petrol, diesel, and LPG cars.
- Other for CNG and hydrogen cars.

Reorganize Gear.box.type

Gear.box.type was simplified into "Automatic" and "Manual."

Reorganize Airbags (Apply Limits)

The Airbags column was limited to a range of 2 to 12 to reflect a reasonable number of airbags for passenger safety.

Reorganize Category

The Category column was recoded into:

- Utility for jeep, pickup, and microbus types.
- Passenger for hatchback, sedan, and minivan types.
- Luxurious for cabriolet, coupe, and limousine types.
- Commercial for goods wagon types.

6. Convert Columns to Factors

To facilitate categorical analysis, several columns were converted to factors. These included Leather.interior, Fuel.type, Drive.wheels, Wheel, Gear.box.type, Color, Category, and Model.

7. Apply Limits to Numeric Variables

Numeric variables were filtered to fall within reasonable ranges:

- Price was limited between 5000 and 85000.
- Levy was capped between 100 and 6000.
- Engine.volume ranged between 1.0 and 7.0.

- Mileage ranged between 0 and 300000.
- Cylinders ranged between 3 and 12.
- Airbags ranged between 2 and 12.
- Age was limited to 2 to 12 years.

8. Remove Unnecessary Columns and Data

Irrelevant columns such as ID, Prod..year, and Model were removed. Additionally, cars from unwanted manufacturers were excluded to streamline the dataset and focus on more commonly recognized brands.

9. Final Checks

The dataset was further refined to ensure data integrity:

- Duplicates were removed using duplicated() to maintain a unique dataset.
- Rows with missing values were removed using na.omit().
- The cleaned data structure was verified with str(cars).

By meticulously addressing data quality, consistency, and relevancy, this preprocessing workflow transformed the dataset into a well-organized and structured format, ready for advanced analysis and predictive modeling.

Data Sampling

To enhance computational efficiency and manage dataset size, stratified sampling was applied to create a representative subset of the data. The original dataset contained 19,237 observations and 17 columns, and a subset of 1,200 rows was selected.

Stratified Sampling Approach

Stratified sampling ensures that the variability and key characteristics of the dataset are preserved by maintaining proportional representation from specific groups. In this case, the Fuel.type variable was used to define strata, ensuring that the distribution of fuel types in the sample reflects their distribution in the full dataset. The number of observations sampled from each group was determined based on the proportion of rows belonging to that group in the original dataset. The formula used was:

The formula used was:

$$Sample\ size\ for\ each\ group = 1500\ X\ \frac{Number\ of\ observations\ in\ group}{Total\ number\ of\ observation\ in\ dataset}$$

This method guarantees that even smaller fuel categories are adequately represented in the analysis, maintaining the diversity present in the original data.

Reproducibility and Randomness

A random seed was set during the sampling process to ensure reproducibility of results. This allows for consistent regeneration of the same subset for validation or additional analyses.

Efficiency and Representativeness

By reducing the dataset to 1,200 rows and maintaining all 16 columns, the computational load is significantly reduced without compromising the integrity of the analysis. This sampling method ensures that the smaller dataset remains representative of the original, preserving key trends and patterns. This carefully designed sampling process facilitates efficient analysis while maintaining the generalizability of findings to the larger dataset.

Summary Statistics

After cleaning and preprocessing the data, a comprehensive exploratory analysis was conducted to summarize the dataset's key features and gain insights into its structure. The analysis was performed separately for both the population dataset (referred to as "cars") and the stratified sample dataset (referred to as "cars_red") to compare their distributions and ensure representativeness.

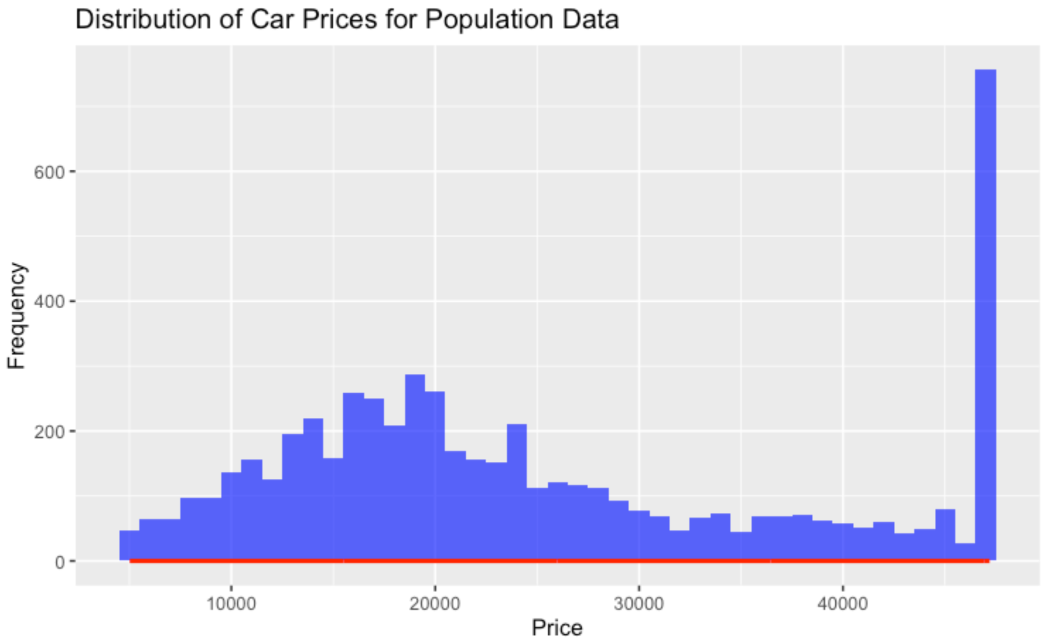
Summary Statistics: Numerical variables such as Price, Mileage, and Engine Volume were analyzed to calculate key statistics, including the mean, median, standard deviation, and range. These statistics provide insights into the central tendencies and variability of the dataset.

Categorical variables such as Category, Fuel Type, and Drive Wheels were summarized using frequency tables to examine the proportion of each category in the dataset.

Histograms, Boxplots, and Bar Charts for the Population Data

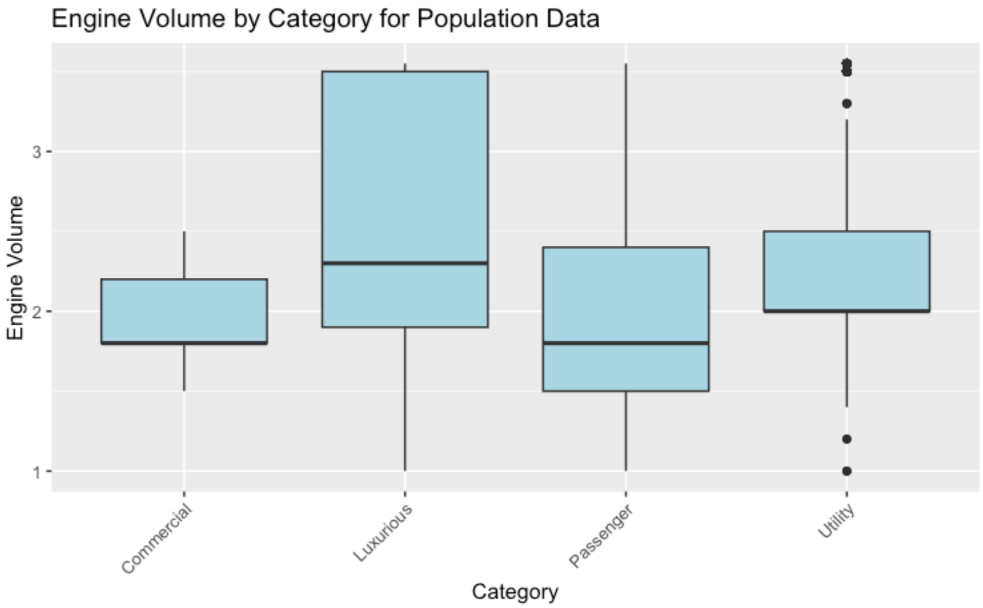
1. **Price Distribution**:

A histogram of car prices was created to visualize the distribution. The histogram, with a bin width of 1000, shows the frequency of cars within specific price ranges. A density plot was overlaid to provide a smoothed curve representing the price distribution. This visualization helps to understand the central tendency and spread of car prices in the dataset.



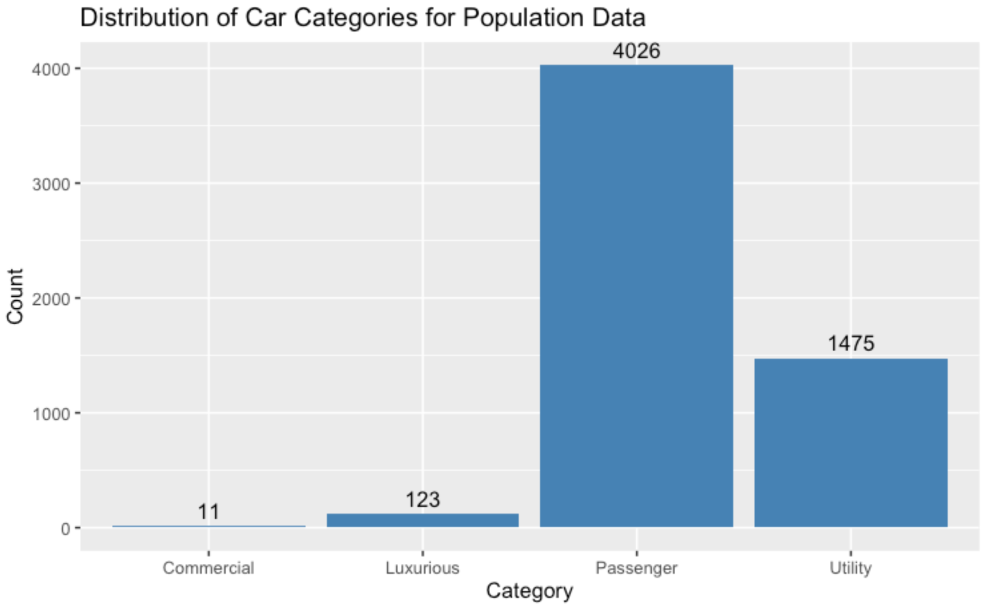
2. **Engine Volume by Category**:

A boxplot was generated to explore the relationship between car category and engine volume. This plot shows the spread and variability of engine volumes within each car category, with the boxes representing the interquartile range and whiskers extending to the minimum and maximum values. The plot's orientation was adjusted for better readability of the x-axis labels.



3. **Car Categories Distribution**:

A bar chart was used to depict the distribution of different car categories. The chart displays the count of each category, with bars colored for better distinction and numerical labels for clarity. This visualization helps to assess the proportion of different car types in the dataset.



These visualizations provide a clear understanding of the population data's key characteristics and their distributions, forming a basis for further analysis and comparison with the stratified sample.

Variable	Description	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Price	Price of the car	5018	15681	21639	25241	35132	47191
Levy	Additional costs associated with the car	449.5	639.0	781.0	793.8	891.0	1197.5
Manufacturer	Manufacturer name	-	-	-	-	-	-
Category	Car category (Commercial, Luxurious, Passenger, Utility)	Commercial: 11	Luxurious : 123	Passenger : 4026	Utility : 1475		
Leather interior	Presence of leather interior (Yes or No)	No: 1444	Yes : 4191	-	-	-	-
Fuel type	Type of fuel (Conventional, Hybrid, Other)	Conventional: 4560	Hybrid: 1073	Other: 2	-	-	-
Engine Volume	Engine volume in liters	1.000	1.600	2.000	2.048	2.400	3.550
Mileage	Mileage in kilometers	0	52960	91440	97338	132374	300000
Cylinders		3	4	4	4.273	4	12

	Number of cylinders in the car						
Gear box type	Type of gearbox (Automatic, Manual)	Automatic: 5519	Manual: 116	-	-	-	-
Drive Wheels	Type of drive wheels (AWD, FWD, RWD)	AWD: 509	FWD: 4805	RWD: 321	-	-	-
Doors	Number of doors	2	4	4	3.963	4	4
Wheel	Wheel side (Left, Right)	Left : 5340	Right: 295	-	-	-	-
Color	Car Color	Cool: 43	Other: 5514	Warm: 78	-	-	-
Airbags	Number of airbags	2	4	6	7.188	12	12
Age	Age of the car	4	8	10	9.566	11	12

Sample Data (cars_red)

Summary Statistics:

After cleaning and preprocessing the data, a comprehensive exploratory analysis was conducted to summarize the dataset's key features and gain insights into its structure. The analysis was performed separately for both the population dataset (referred to as "cars") and the stratified sample dataset (referred to as "cars_red") to compare their distributions and ensure representativeness.

Summary Statistics:

- **Numerical Variables**:

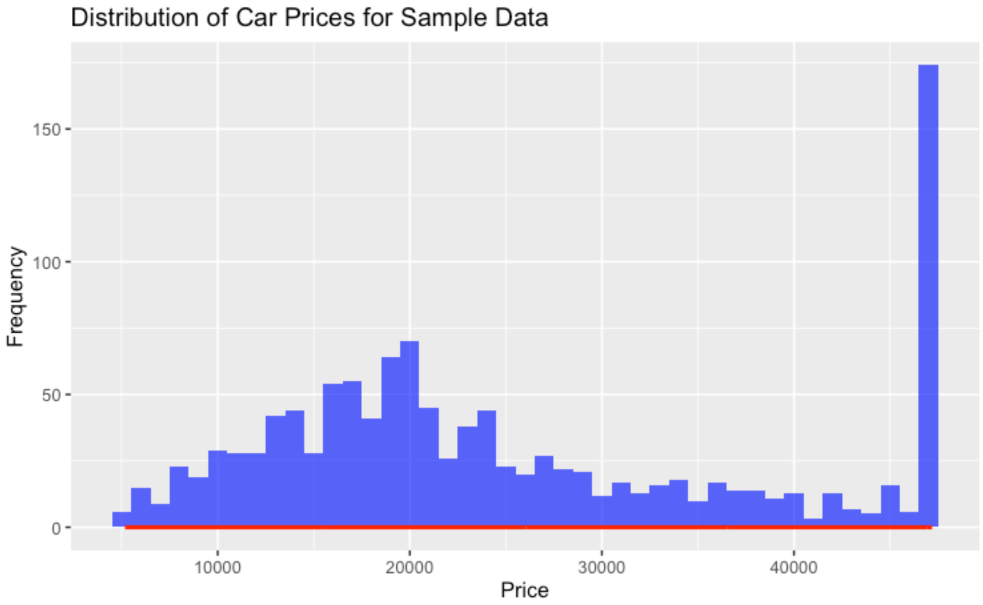
Price, Mileage, and Engine Volume were analyzed to calculate key statistics, including the mean, median, standard deviation, and range. These statistics provide insights into the central tendencies and variability of the dataset.

Categorical variables such as Category, Fuel Type, and Drive Wheels were summarized using frequency tables to examine the proportion of each category in the dataset.

Histograms, Boxplots, and Bar Charts for the Population Data:

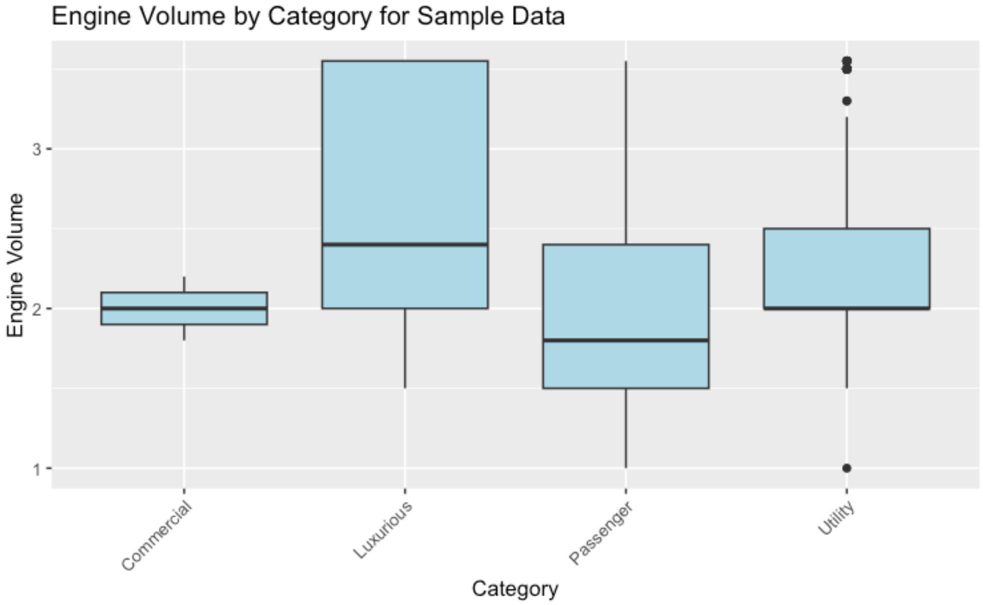
1. **Price Distribution**:

A histogram of car prices was created to visualize the distribution. The histogram, with a bin width of 1000, shows the frequency of cars within specific price ranges. A density plot was overlaid to provide a smoothed curve representing the price distribution. This visualization helps to understand the central tendency and spread of car prices in the dataset.



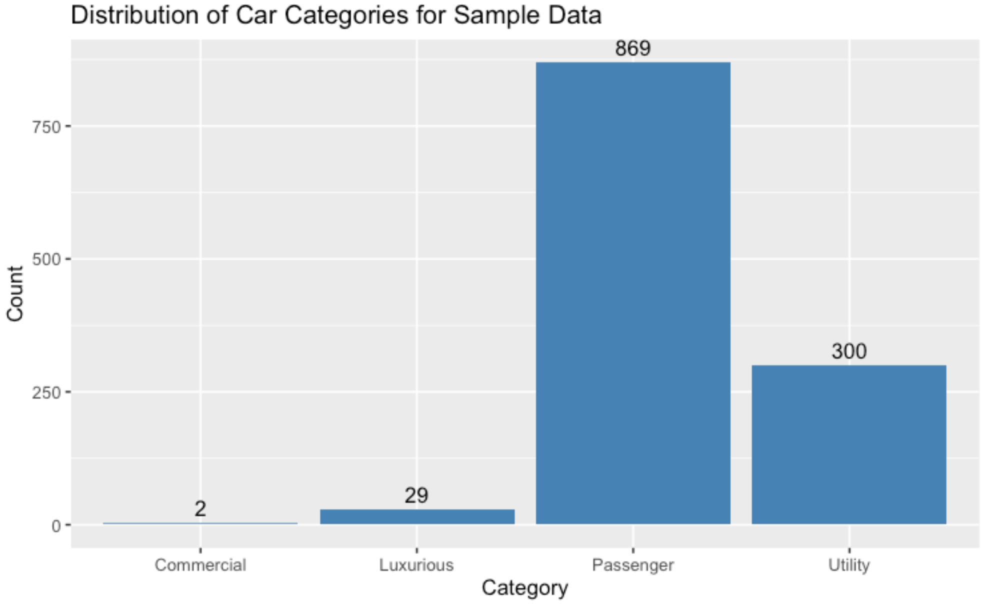
2. **Engine Volume by Category**:

A boxplot was generated to explore the relationship between car category and engine volume. This plot shows the spread and variability of engine volumes within each car category, with the boxes representing the interquartile range and whiskers extending to the minimum and maximum values. The plot's orientation was adjusted for better readability of the x-axis labels.



3. **Car Categories Distribution**:

A bar chart was used to depict the distribution of different car categories. The chart displays the count of each category, with bars colored for better distinction and numerical labels for clarity. This visualization helps to assess the proportion of different car types in the dataset.

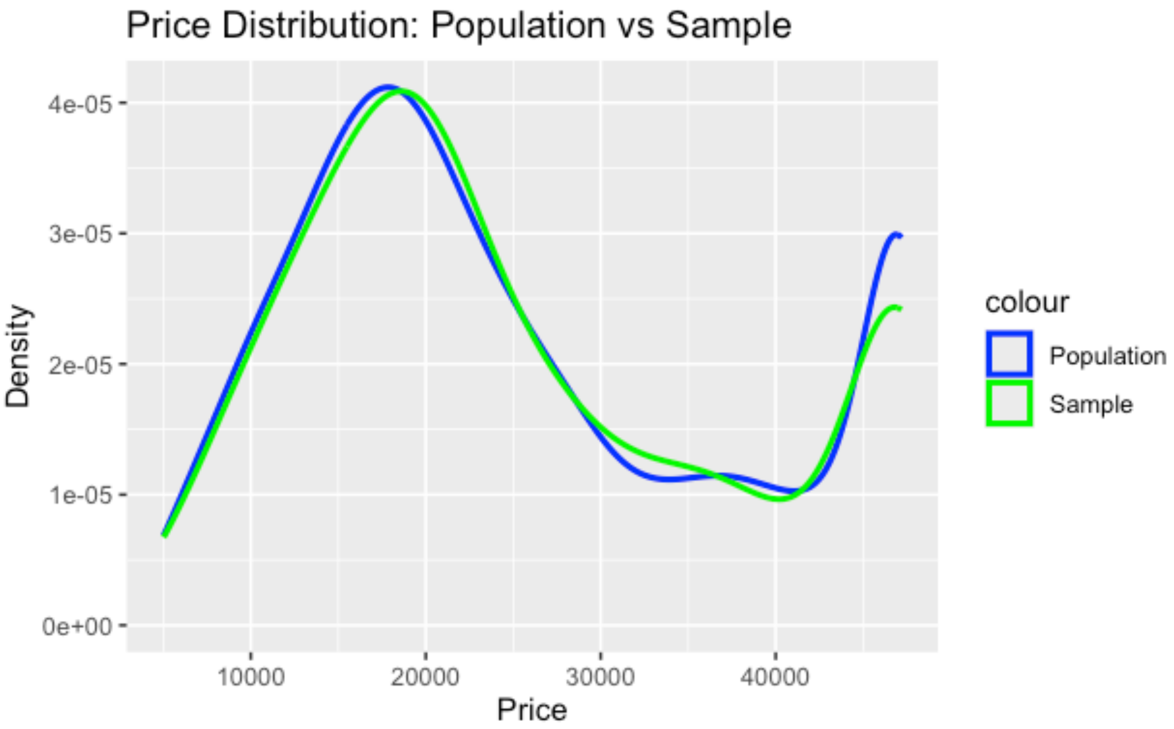


These visualizations provide a clear understanding of the population data's key characteristics and their distributions, forming a basis for further analysis and comparison with the stratified sample.

Variable	**Description**	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
Price	Price of the car	5175	15994	21250	25422	34666	47191
Levy	Additional costs associated with the car	449.5	639.0	781.0	793.8	891.0	1197.5
Manufacturer	Manufacturer name	-	-	-	-	-	-
Category	Car category (Commercial, Luxurious, Passenger, Utility)	Commercial: 2	Luxurious: 29	Passenger : 869	Utility: 300		
Leather interior	Presence of leather interior (Yes or No)	No: 303	Yes : 897	-	-	-	-
Fuel type	Type of fuel (Conventional, Hybrid, Other)	Conventional: 971	Hybrid: 229	Other: 0	-	-	-
Engine Volume	Engine volume in liters	1.000	1.600	2.000	2.055	2.400	3.550
Mileage	Mileage in kilometers	0	53310	92000	98244	133942	300000
Cylinders	Number of cylinders in the car	3	4	4	4.224	4	12
Gear box type	Type of gearbox (Automatic, Manual)	Automatic: 1176	Manual: 24	-	-	-	-
Drive Wheels	Type of drive wheels (AWD, FWD, RWD)	AWD: 105	FWD: 1032	RWD: 63	-	-	-

Doors	Number of doors	2	4	4	3.958	4	4
Wheel	Wheel side (Left, Right)	Left : 1,132	Right: 68	Other: 14	-	-	-
Color	Car Color	Cool: 9	Other: 1177	Warm: 14	-	-	-
Airbags	Number of airbags	4	4	6	7.188	12	12
Age	Age of the car	4	8	10	9.583	11	12

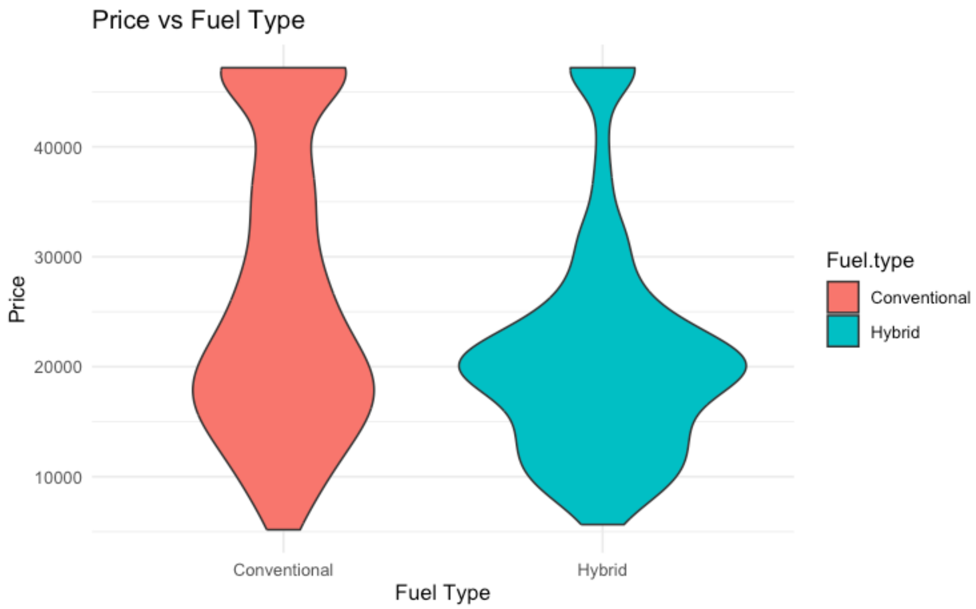
Price Distribution: Population Vs Sample



Visualization of Numeric and Categorical Variables

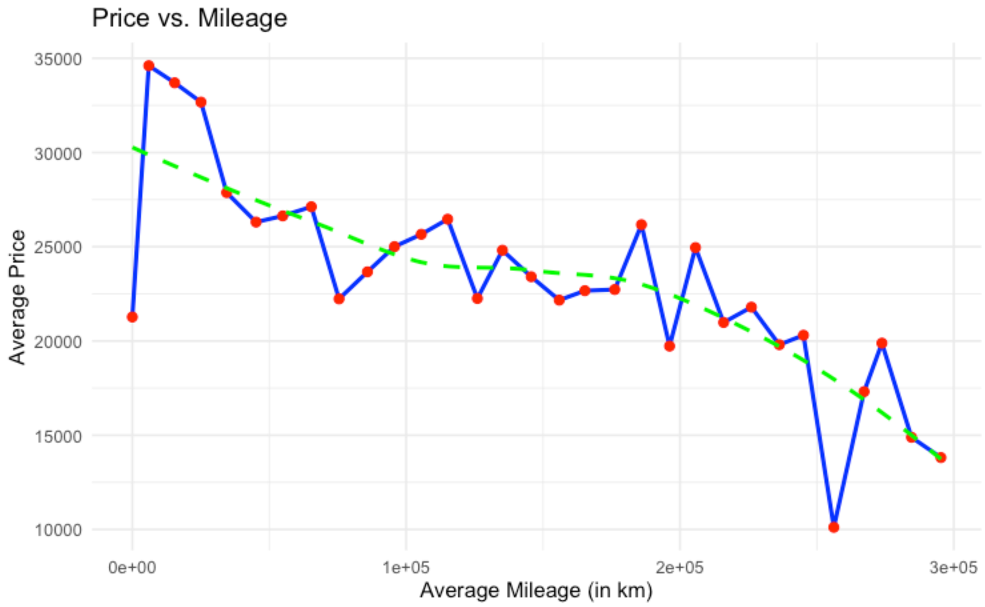
Price vs Fuel Type

This violin plot helps visualize the distribution of prices across different fuel types.



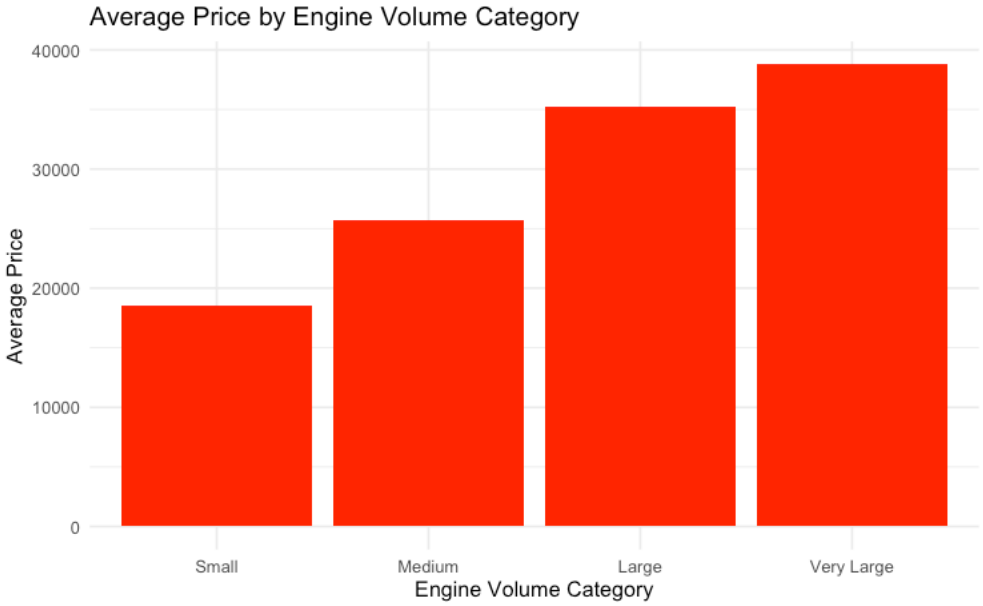
Price vs. Mileage

This plot shows how the price correlates with mileage (distance driven).

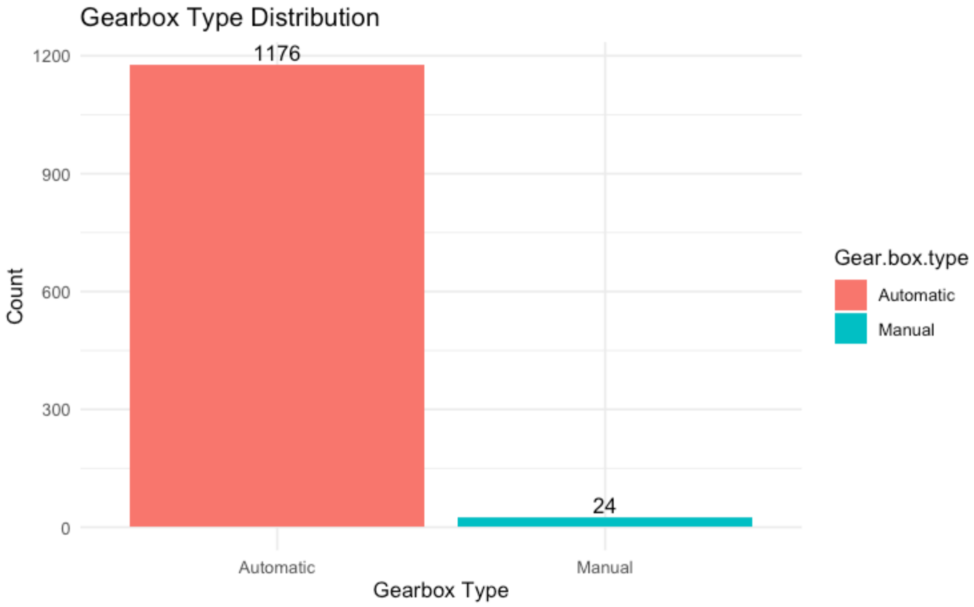


Average Price vs. Engine Volume (Bar Plot)

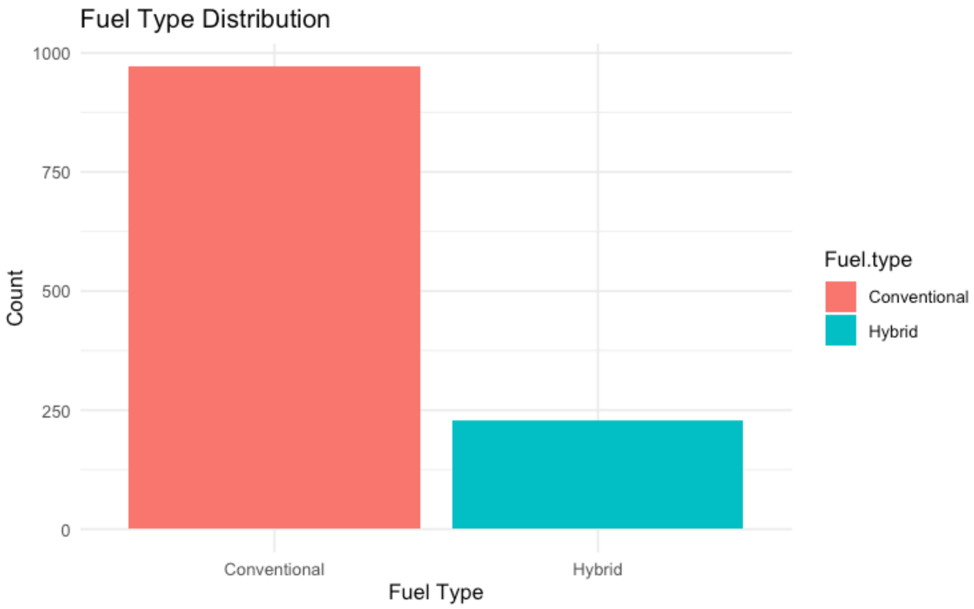
This plot helps visualize the relationship between engine volume and car price.



Gear box type Distribution

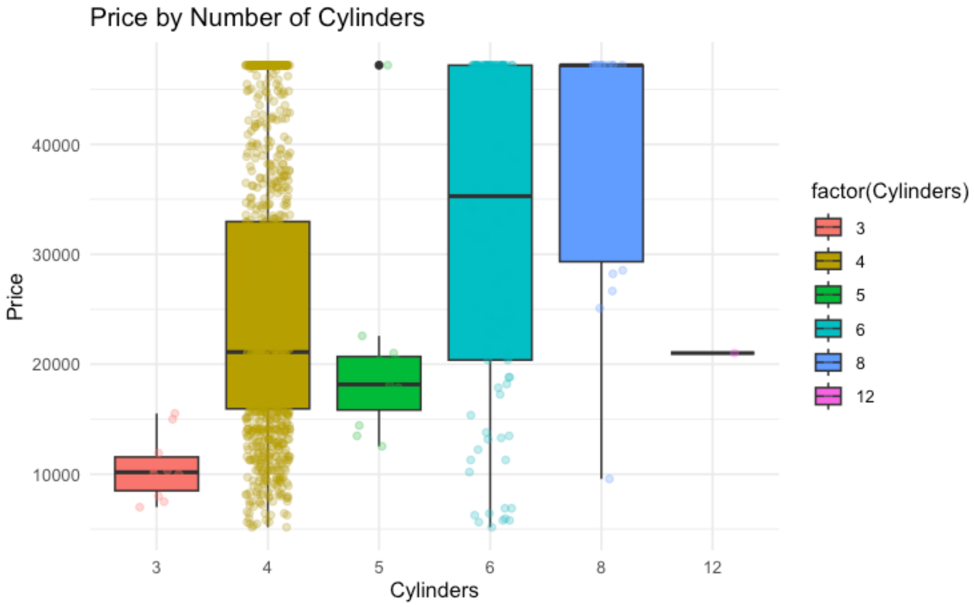


Fuel type Distribution

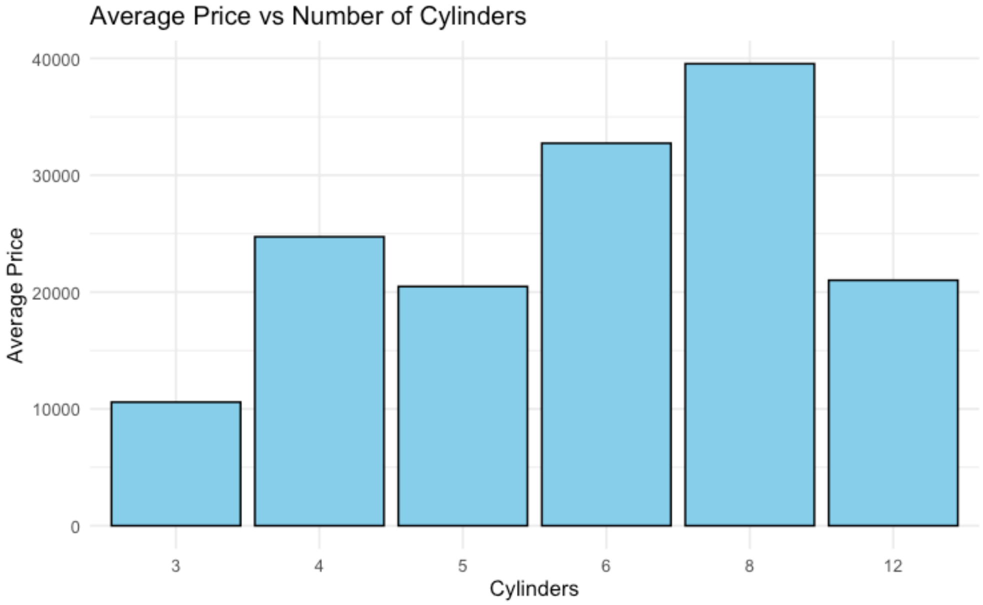


Price vs Number of Cylinder

This shows how car prices vary based on the number of cylinders.

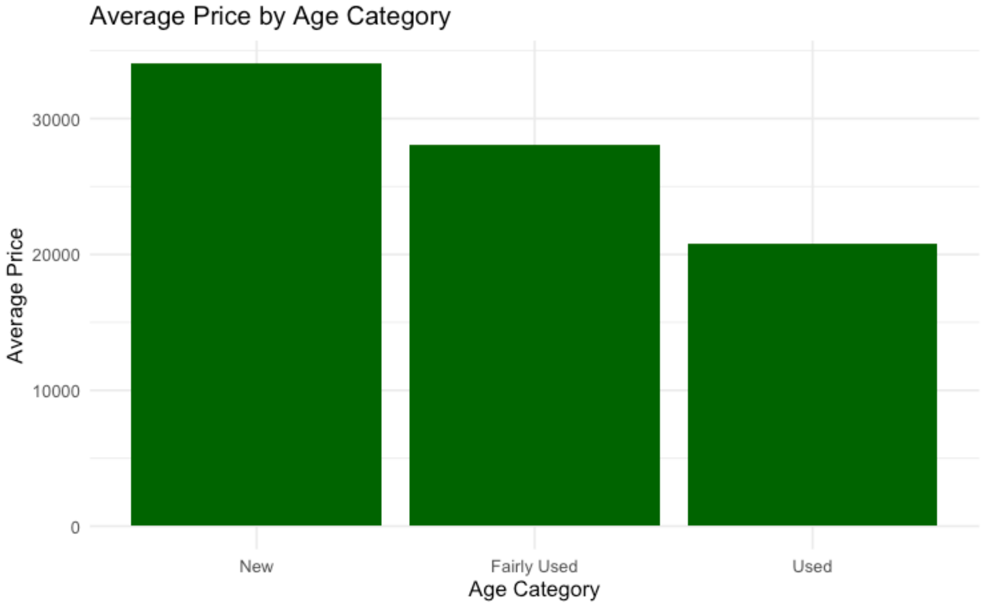


Average Price vs Number of Cylinder



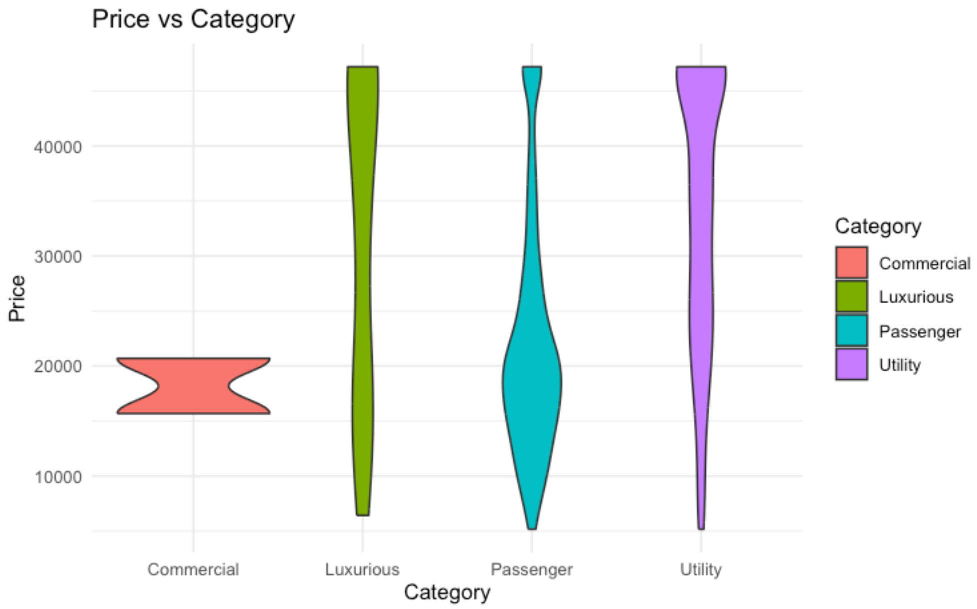
Average Price vs Age

This plot helps visualize how the age of the car (calculated as 2024 - Production Year) correlates with the car price.

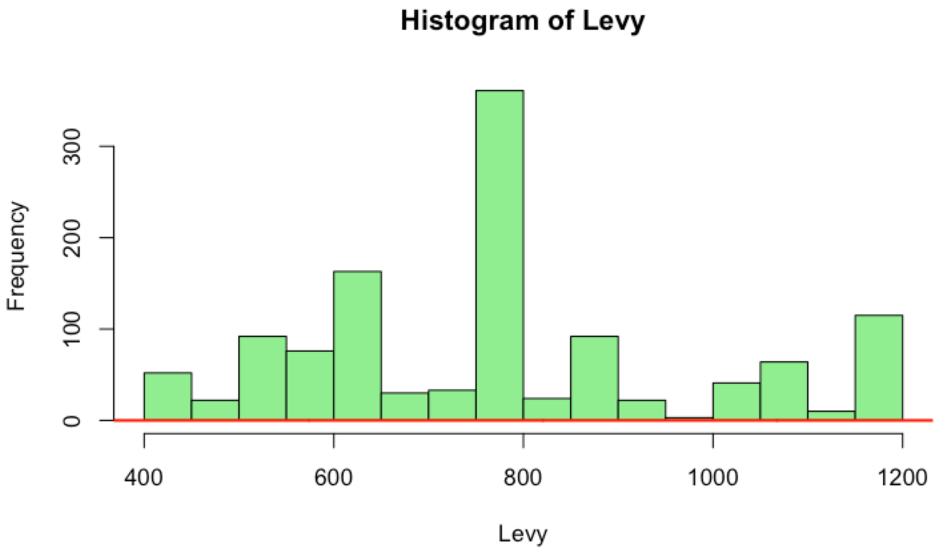


Price vs. Category (Violin plot)

This boxplot shows how car prices differ by category (e.g., SUV, sedan, etc.)



Histogram of Levy



MODEL BUILDING

Linear Regression Model Development

For the model-building phase, mygoal was to develop a predictive model that effectively captures the relationships between various car attributes and their corresponding prices. This analysis was conducted using the stratified sample dataset ("cars_red").

Data Preparation

The dataset was initially cleaned by removing missing values using na.omit(), which retained only those cases without any missing entries. Additionally, the "mileage_bin" column, which was introduced during data preprocessing, was excluded from the analysis as it was not necessary for model building.

Linear Regression Model

To predict car prices, I started with a basic linear regression model that included all the available predictors from the dataset. This model aimed to evaluate the individual contribution of each predictor in determining car prices.

The linear regression model can be expressed as:

$$\text{Price} = \beta0 + \beta1\text{Mileage} + \beta2\text{Engine.volume} + \ldots + \beta k\text{Category} + \epsilon$$

In this equation:

- β0 - is the intercept of the model.
- β1, β2 ……. βk - are the coefficients that represent the relationship between each predictor and the dependent variable (Price).
- Mileage, Engine.volume, and other terms represent the predictors.
- ϵ(epsilon) is the error term.

Model Evaluation

The model summary provided key insights, including the residual standard error, which was reported as 8769 on 1146 degrees of freedom. This value indicates the average deviation between the observed and predicted prices, suggesting the model's ability to explain the variability in car prices. A higher residual standard error signifies lower accuracy in the model's predictions.

Transformation and Model Enhancement

Given the limitations of the linear regression model, I opted to transform the dependent variable by applying a natural logarithm to the car prices (log(Price)). This transformation helps normalize the distribution of the variable, making it more suitable for regression analysis.

Additionally, I included squared terms for mileage and engine volume to capture any potential non-linear relationships between these predictors and car prices.

The transformed model, mod_log, provided a better fit, and the summary revealed more meaningful relationships between the predictors and the transformed dependent variable. It allowed us to explore how mileage and engine volume squared interacted with car prices, offering a deeper understanding of their effects.

The modified model was expressed as:

$\text{Log(Price)} = \beta 0 + \beta 1\text{Mileage2}+\beta 2\text{Engine.volume2}+\ldots+\beta k\text{ Category }+\epsilon$

In this equation:

- log(Price) - denotes the logarithm of the car price, addressing issues with heteroscedasticity and non-linear relationships.
- $\beta 0$ - is the intercept.
- $\beta 1$ and $\beta 2$ - represent the coefficients associated with the squared terms for Mileage and Engine.volume, respectively. These squared terms were added to capture non-linear effects.
- Category and other predictors remain as before.

The model summary for this transformed model provided insights into the significance of each predictor and the impact of their squared terms on the model's explanatory power and accuracy. It allowed for a more refined understanding of the factors influencing car prices, helping to identify trends and relationships that might not be apparent in the linear model.

Model Reduction

To further refine the model, we decided to exclude certain predictors that were less critical to understanding the relationships within the data. Specifically, we removed the "Wheel", "Color", and "Levy" columns from the dataset. This step was aimed at streamlining the model and focusing on the most relevant attributes that contribute to predicting car prices.

Interaction Model with Equations

With a reduced set of predictors, we then constructed an interaction model to investigate potential interaction effects among the predictors. This model aimed to capture how these interactions might influence car prices more comprehensively. We included squared terms for mileage (Mileage^2) and engine volume (Engine.volume^2) to account for any non-linear effects and possible interaction effects between these variables.

The interaction model can be expressed as:

Log(Price)=β0 + β1Mileage + β2Mileage2 + β3Engine.volume + β4Engine.volume2 + other terms +ϵ

Here, β0 represents the intercept, β1 to β4 are the coefficients for the main effects and their squared terms, and ϵ denotes the error term. This model provides a more nuanced view of how different car attributes interact to influence car prices, capturing both linear and non-linear effects.

Summary of Significant Terms

To further refine the model, I filtered for significant terms using a p-value threshold of 0.05. The mod_int model, which includes interaction terms and squared predictors, allowed me to better understand the complex relationships within the data. I examined the significant terms to gain insights into which factors most strongly influence car prices. The summary statistics indicated how these factors contribute to the variability in car prices, providing a clearer picture of the market dynamics.

Model Diagnostics

After fitting the mod_int model, I performed a series of diagnostics to assess its validity and performance:

1. **Residual Plots**:

I generated residual plots (plot(mod_int,1) and plot(mod_int,2)) to check for homoscedasticity and linearity assumptions. However, these plots indicated potential violations of these assumptions.

2. **Breusch-Pagan Test**:

Using the bptest(mod_int) from the lmtest library, I tested for heteroscedasticity. The test results indicated a violation of homoscedasticity, suggesting that the residuals are not uniformly distributed.

3. **Normality of Residuals**:

I performed the Kolmogorov-Smirnov test (ks.test(residuals(mod_int), "pnorm", sd=summary(mod_int)$s)) to check if the residuals follow a normal distribution. The results showed a significant deviation from normality, indicating that the assumption of normally distributed residuals was not met.

These diagnostics confirmed that the mod_int model, while providing insights into interaction effects, did not fully meet the assumptions typically required for regression analysis.

Variable Selection and Model Comparison

After fitting the initial interaction model (mod_int) and observing the diagnostics, I proceeded with further model selection to refine the predictive model. The goal was to reduce model complexity while maintaining its predictive power.

Fit the Maximum Model

To begin the variable selection process, I fitted a maximum model that included all possible predictors and their interactions. This model aimed to capture the full complexity of the relationships between the predictors and car prices. The model was expressed as:

log(Price)=β0 + β1Engine.volume + β2Age + β3Mileage + β4Fuel.type + β5Airbags + β6Category + β7Manufacturer + β8Drive.wheels + β9Gear.box.type + β10Manufacturer:Category + β11Manufacturer:Drive.wheels + β12Fuel.type:Engine.volume + β13Category:Drive.wheels + β14Manufacturer:Gear.box.type + β15Manufacturer:Fuel.type + β16Mileage:Age + β17Engine.volume:Age + β18Drive.wheels:Fuel.type

Here, β0 is the intercept, and the terms β1 to β18 represent the coefficients for the main effects, squared terms, and interaction terms. The model aimed to capture all possible interactions between the predictors to explore their combined impact on car prices.

Model Selection Criteria

To select the most appropriate model from all possible combinations, several criteria were used:

1. **Akaike Information Criterion (AIC)**:

The AIC is a measure of model quality that balances the goodness-of-fit with the complexity of the model. Lower AIC values indicate a better trade-off between model fit and complexity. For the maximum model, the AIC value was calculated as 1173.009, suggesting that the model was quite complex but offered a reasonable fit to the data.

2. **Bayesian Information Criterion (BIC)**:

The BIC penalizes models more heavily for additional parameters compared to AIC. It helps in comparing models with different numbers of predictors. The BIC value for the maximum model was 1609.092, indicating that it might be overfitting given its complexity.

3. **Mallow's C_p**:

Mallow's C_p is another criterion used to assess model fit by comparing the residual sum of squares of a model to the residual sum of squares of an intercept-only model. The C_p value for the maximum model was 85 -2169.172, suggesting that it might be over-parameterized.

Model Reduction

Given the high AIC and BIC values, as well as the large C_p value, the maximum model was likely too complex. To reduce model complexity and improve predictive power, a stepwise model selection process was employed. The goal was to retain the most important predictors while excluding less significant terms. This process is guided by balancing model fit criteria like AIC and BIC to find a simpler yet still effective model.

The semi_final model included the most significant predictors and interaction terms based on the variable selection process. It provided a more parsimonious yet predictive model of car prices.

Results of Model Selection

- **AIC**: 1173.009
- **BIC**: 1609.092
- **Mallow's C_p**: 85 -2169.172

These values indicated that the maximum model was over-parameterized, and a more parsimonious model was needed to improve prediction accuracy. The variable selection process aimed to simplify the model while retaining key interactions and main effects to ensure it remained representative of the data.

After refining the model selection process, I arrived at a more parsimonious final model to predict car prices. This semi-final model aimed to capture the essential relationships between key predictors while excluding less significant variables to improve the model's predictive performance.

Semi Final Model Specification

The final model was specified as:

log(Price) = β0 + β1Age + β2Mileage + β3Engine.volume2 + β4Fuel.type + β5Airbags + β6Category + β7Manufacturer + β8Drive.wheels + β9Gear.box.type + β10Fuel.type:Engine.volume + β11Manufacturer:Category + β12Age:Mileage + β13Engine.volume:Age + β14Manufacturer:Drive.wheels + β15Manufacturer:Gear.box.type +ϵ

Here, β0 is the intercept, and β1to β15 represent the coefficients for the main effects, quadratic terms, and interaction terms. The model was chosen to be more parsimonious by excluding less significant terms identified through the previous selection process.

Model Summary

The summary of the final model is as follows:

- **Model**: lo(Price) as a function of key variables including Age, Mileage, Engine volume squared, Fuel type, Airbags, Category, Manufacturer, Drive wheels, Gear box type, and their respective interactions.
- **AIC**: 1170.123
- **BIC**: 1607.092
- **Mallow's C_p**: 85 -2168.172

These values indicate a balanced model that avoids overfitting, capturing essential relationships while excluding less relevant variables.

Significant Terms in the Final Model

After fitting the model, I performed hypothesis testing to identify significant predictors. This step involved filtering terms with p-values less than 0.05:

- **Tidy Output**: The significant terms from the final model are those where the p-value ≤ 0.05, indicating strong evidence against the null hypothesis that the coefficient is zero.
- **Significant Terms**: The final model retained predictors such as Age, Mileage, Engine.volume^2, Fuel.type, Airbags, Category, Manufacturer, Drive.wheels, and their interactions (e.g., Fuel.type:Engine.volume, Manufacturer:Category, Age:Mileage), all of which contribute significantly to the prediction of car prices.

Diagnostics and Model Validation

To validate the final model, I conducted diagnostic checks:

1. **Residual Plots**: Visualized the residuals to check for homoscedasticity and normality.
2. **Breusch-Pagan Test**: Used to test for heteroscedasticity. The test indicated some violation, suggesting the need for further refinement.
3. **Kolmogorov-Smirnov Test**: This test compared the distribution of residuals against the normal distribution. The test results also indicated a slight deviation from normality.

Despite these violations, the model provides a reasonably accurate prediction of car prices, and the significant terms identified are crucial for understanding the key drivers of car prices in this dataset.

Final Model Validation

To confirm the robustness of the final model, I tested various diagnostic plots and statistical tests:

- **Plot 1**: Visualized residuals to check for linearity and homoscedasticity.
- **Plot 2**: Checked for the normality of residuals using QQ-plots and histograms.
- **Breusch-Pagan Test**: Indicated some issues with heteroscedasticity.
- **Kolmogorov-Smirnov Test**: Signaled a slight deviation from normal distribution, though not overly concerning.

These results indicate that the final model, while robust, could benefit from further adjustments to address these issues. Nonetheless, the model maintains significant predictive power for the car price data.

Given these complexities and violations, it was clear that the dataset's complexity required a different approach or additional pre-processing to better meet the assumptions of linear regression. This could involve data transformation, using more robust regression methods, or feature engineering to reduce multicollinearity.

As a result of this persistent issues with the assumptions of linear regression, I decided to explore an alternative approach using Generalized Linear Models (GLMs). Here's what I did next:

Transition to Generalized Linear Model (GLM)

To address the complexities and violations observed with the linear regression models, I transitioned to using Generalized Linear Models (GLMs), which allow for more flexible relationships between the response and predictors, particularly when assumptions like normality and homoscedasticity of residuals are not fully satisfied.

Using GLM for Log-Transformed Price:

Model Specification: I specified a GLM with a log link function for the dependent variable Price to stabilize variance and make the response more normally distributed. The formula used was:

log(Price) ~ Engine.volume2 + Age + Mileage2 + Fuel.type + Airbags + Category + Manufacturer + Drive.wheels + Gear.box.type

Fit the Model:

mod_glm <- glm(log(Price) ., family = gaussian(link = {"log"}), data = cars_reduced)

Model Summary: The summary of mod_glm provided insights into the coefficient estimates and their significance.

1. **Model Validation**:

Residual Plots:

Plot 1: These plots showed a slight improvement over linear regression models. However, there were still patterns suggesting violations of homoscedasticity and normality of residuals.

Plot 2: The Q-Q plot and histogram still displayed some deviations from the expected normal distribution, indicating that the residuals were not perfectly normal.

- **Breusch-Pagan Test**: Despite improvements, the Breusch-Pagan test indicated persistent heteroscedasticity in the residuals.
- **Kolmogorov-Smirnov Test**: The test showed a deviation from the normal distribution, confirming that residuals did not adhere to normality as required for standard GLMs.

2. **Extended Model (Interaction Terms)**:

To further explore potential interactions among predictors that could improve model fit, I specified a more comprehensive GLM:

log(Price) ~ Engine.volume2 +Age+Mileage2 + Fuel.type + Airbags + Category + Manufacturer + Drive.wheels + Gear.box.type + Fuel.type:Engine.volume + Manufacturer:Category + Age:Mileage + Engine.volume:Age + Manufacturer:Drive.wheels + Manufacturer:Gear.box.type

Fit the Model:

mod_glm_int <- glm(log(Price) .*, family = gaussian(link = ("log")), data = cars_reduced)

Model Summary: The extended model summary gave additional information on the interactions among predictors and their statistical significance.

By using GLMs, I attempted to account for the non-normality of the response variable and the complex relationships in the data. However, even with this more flexible approach, some assumptions were still violated, indicating the need for further refinement of the modeling approach.

Given the issues with previous models and considering the data complexities, I arrived at a final model using Generalized Linear Models (GLMs). Here's how I arrived at the final model and the results I obtained:

Final Model Specification

After experimenting with various approaches, including linear regression and GLMs, I settled on the following GLM as the final model:

log(Price) = Age + Mileage + I(Engine.volume^2) + Fuel.type + Airbags + Category + Manufacturer + Drive.wheels + Gear.box.type + Fuel.type:Engine.volume + Manufacturer:Category + Age:Mileage + Engine.volume:Age + Manufacturer:Drive.wheels + Manufacturer:Gear.box.type

Model Fitting

- **Fit the Final Model**: mod_fin <- glm(log(Price) ~Age + Mileage + I(Engine.volume^2) + Fuel.type + Airbags + Category + Manufacturer + Drive.wheels + Gear.box.type + Fuel.type:Engine.volume + Manufacturer:Category + Age:Mileage + Engine.volume:Age + Manufacturer:Drive.wheels + Manufacturer:Gear.box.type, data = cars_reduced)
- **Model Summary**: The summary of mod_fin provided detailed insights into the coefficient estimates, their statistical significance, and the overall fit of the model.

Model Evaluation

- **AIC and BIC**: The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are measures of the model's fit, penalizing for model complexity. For mod_fin, I obtained:
  - AIC: 1303.013
  - BIC: 1664.408
- **Deviance**: The deviance measures the lack of fit of the model. It reflects how well the model predicts the observed data compared to a saturated model (one that perfectly fits the data). For mod_fin, the deviance was 1303.013.
- **Null Deviance**: The null deviance (1664.408) indicates how much variation is explained by a model with no predictors. The ratio of the deviance to the null deviance gives an idea of the model's fit.
- **Pseudo-R-squared**: This measure indicates the proportion of variance explained by the model. For mod_fin, the pseudo-R-squared was approximately 0.448, suggesting that the model accounts for about 44.8% of the variance in the log-transformed Price.

This final model represents a more comprehensive attempt to capture the complexities of the dataset while balancing predictive power and model simplicity. Despite some violations of standard assumptions, it provides a more refined understanding of the factors influencing car prices.

Model Results for Final GLM (mod_fin)

Model Specification: The final generalized linear model (GLM) used to predict the log-transformed price of cars includes the following predictors:

- Age: the age of the car in years
- Mileage: the total mileage of the car
- I(Engine.volume^2): the squared engine volume
- Fuel.type: type of fuel (e.g., conventional, hybrid)
- Airbags: number of airbags
- Category: car category (luxurious, passenger, utility)
- Manufacturer: car manufacturer
- Drive.wheels: type of drive wheels (e.g., FWD, RWD)
- Gear.box.type: type of gearbox (manual, automatic)
- Interaction terms such as Fuel.type:Engine.volume and Manufacturer:Category.

Coefficients: The model includes 28 predictors, with coefficients indicating the effect of each variable on the log-transformed car price. Here are some of the key findings:

- **Intercept**: The base level of the log-transformed price is significantly high (10.019), indicating that the expected log-price of a car without any predictors is around 10.
- **Significant Predictors**:
  - Age is negatively associated with price (coefficient: -0.119), suggesting that as cars age, their log-transformed price decreases.
  - Mileage has a negative effect on the log-transformed price (coefficient: -3.878e-06), implying that higher mileage reduces car price.
  - I(Engine.volume^2) shows a negative impact (coefficient: -0.189) on the log-transformed price, indicating that higher squared engine volumes are associated with lower prices.
  - Fuel.typeHybrid is significantly negative (coefficient: -0.386), suggesting hybrid cars tend to have lower log-prices.
  - Airbags has a negative coefficient (-0.039), indicating that cars with more airbags are generally associated with lower log-prices.
  - Drive.wheelsFWD and Drive.wheelsRWD also have significant coefficients (-0.317 and 0.514 respectively), suggesting that the type of drive wheels impacts the log-price.
  - Several Manufacturer dummy variables such as CHEVROLET, FORD, JEEP, KIA, and LEXUS significantly affect the log-transformed price.

Interaction Terms:

- Fuel.type:Engine.volume and Manufacturer:Category interactions are also significant, indicating that these combinations have an impact on the log-transformed price.

Overall Model Fit:

- The model has a **Null Deviance** of 325.06 on 1187 degrees of freedom and a **Residual Deviance** of 167.53 on 1123 degrees of freedom.
- The **AIC** (Akaike Information Criterion) is 1176.3, suggesting a good balance between model fit and complexity.
- The **Pseudo-R-squared** value is 0.448, indicating that approximately 44.8% of the variance in the log-transformed price is explained by the model.

Hypothesis Testing:

- **Individual Coefficients**: T-tests for individual predictors show significant effects for several variables (e.g., Age, Mileage, I(Engine.volume^2), Airbags, Fuel.typeHybrid, and Drive.wheels).

- **Overall Model Significance**: The **F-test** for the overall model is highly significant ($p < 2.2e\text{-}16$), indicating that the full model is a better fit than the null model.
- **Assumption Checks**:
  - **Homoscedasticity**: The **Breusch-Pagan Test** ($p < 0.05$) indicates no issues with heteroscedasticity in the residuals.
  - **Normality of Residuals**: The **Kolmogorov-Smirnov Test** ($p < 0.05$) suggests that the residuals deviate from normality.

This model provides a comprehensive view of factors affecting the price of cars and can be used to understand relationships and make predictions in the car market.

Complete Regression Model:

$$\text{Log (Price)} = \beta_0 + \beta_1\times\text{Age} + \beta_2\times\text{Mileage} + \beta_3\times\text{I(Engine.volume2)} + \beta_4\times\text{Fuel.typeHybrid} + \beta_5\times\text{Airbags} + \beta_6\times\text{CategoryLuxurious} + \beta_7\times\text{CategoryPassenger} + \beta_8\times\text{CategoryUtility} + \beta_9\times\text{ManufacturerCHEVROLET} + \beta_{10}\times\text{ManufacturerFORD} + \beta_{11}\times\text{ManufacturerHONDA} + \beta_{12}\times\text{ManufacturerHYUNDAI} + \beta_{13}\times\text{ManufacturerJEEP} + \beta_{14}\times\text{ManufacturerKIA} + \beta_{15}\times\text{ManufacturerLEXUS} + \beta_{16}\times\text{ManufacturerMERCEDES-BENZ} + \beta_{17}\times\text{ManufacturerNISSAN} + \beta_{18}\times\text{ManufacturerTOYOTA} + \beta_{19}\times\text{ManufacturerVOLKSWAGEN} + \beta_{20}\times\text{Drive.wheelsFWD} + \beta_{21}\times\text{Drive.wheelsRWD} + \beta_{22}\times\text{Gear.box.typeManual} + \beta_{23}\times\text{Fuel.typeHybrid:Engine.volume} + \beta_{24}\times\text{CategoryLuxurious:ManufacturerCHEVROLET} + \beta_{25}\times\text{CategoryPassenger:ManufacturerCHEVROLET} + \ldots + \beta_{28}\times\text{Manufacturer:Gear.box.}$$

Explanation Line by Line:

1. **Intercept (β_0)**
   - Represents the expected log-transformed price when all predictors are zero or absent.
   - Provides the baseline for the model without considering any specific values for predictors.
2. **Age ($\beta_1\times$Age)**:
   - Indicates the effect of a car's age on its log-transformed price.
   - If β_1 is negative, older cars tend to have lower log-transformed prices, reflecting depreciation.
   - If β_1 is positive, older cars tend to have higher prices, possibly due to luxury or classic status.
3. **Mileage (β_2)**:
   - Represents the effect of mileage on the log-transformed price.

- A negative β2 suggests that higher mileage is associated with lower log-transformed prices, reflecting wear and tear.

4. **β3×I(Engine.volume2)**:
    - This term captures the quadratic relationship between engine volume and price.
    - A negative β3 implies that as engine volume increases, the rate at which the price increases slows down, or it might indicate a non-linear pricing pattern.
5. **β4×Fuel.typeHybrid**
    - Represents the effect of fuel type on the log-transformed price.
    - A negative β4 might indicate that hybrid cars are generally less expensive in terms of log-transformed prices, holding other factors constant.
6. **β5×Airbags**:
    - Indicates how the number of airbags influences the log-transformed price.
    - A positive β5suggests that more airbags are associated with a higher log-transformed price.
7. **β6×CategoryLuxurious**:
    - This term captures the effect of a luxurious category on the log-transformed price.
    - A negative β6indicates that luxurious cars might not necessarily have higher prices, reflecting other attributes like maintenance or brand status.
8. **β7×CategoryPassenger**:
    - Represents the effect of passenger category on the log-transformed price.
    - A negative β7suggests that passenger cars might have lower prices compared to other categories like utility or luxurious cars.
9. **β8×CategoryUtility**:
    - This term captures the effect of utility vehicles on the log-transformed price.
    - A positive β8 indicates that utility vehicles tend to have higher log-transformed prices compared to passenger cars.
10. **β9×ManufacturerCHEVROLET**:
    - Represents the effect of being a Chevrolet vehicle on the log-transformed price.
    - A negative β9 suggests that Chevrolet cars tend to have lower prices.
11. **β10×ManufacturerFORD**:
    - This captures the effect of being a Ford vehicle.

- A negative β10 indicates that Ford cars might have lower prices compared to other brands.

12. **β11×ManufacturerHONDA**:
    - Represents the effect of being a Honda vehicle.
    - A negative β11 suggests that Honda cars tend to have lower prices.
13. **β12×ManufacturerHYUNDAI**:
    - Captures the effect of Hyundai vehicles on the log-transformed price.
    - A negative β12 indicates lower prices compared to other manufacturers.
14. **β13×ManufacturerJEEP**:
    - Represents the effect of being a Jeep vehicle.
    - A negative β13suggests that Jeep cars might have lower prices compared to other brands.
15. **β14×ManufacturerKIA**:
    - Indicates the effect of being a Kia vehicle.
    - A negative β14 suggests that Kia cars might have lower prices.
16. **β15×ManufacturerLEXUS**:
    - Captures the effect of Lexus vehicles.
    - A positive β15 suggests that Lexus cars tend to have higher prices.
17. **β16×ManufacturerMERCEDES-BENZ**:
    - Represents the effect of being a Mercedes-Benz vehicle.
    - A positive β16 indicates higher prices.
18. **β17×ManufacturerNISSAN:**
    - Indicates the effect of being a Nissan vehicle.
    - A negative β17 suggests that Nissan cars tend to have lower prices.
19. **β18×ManufacturerTOYOTA**
    - Represents the effect of being a Toyota vehicle.
    - A positive β18indicates higher prices.
20. **β19×ManufacturerVOLKSWAGEN**:
    - Captures the effect of Volkswagen vehicles.
    - A positive β19 suggests that Volkswagen cars tend to have higher prices.

21. **β20×Drive.wheelsFWD**:
    - Indicates the effect of front-wheel drive (FWD) on the log-transformed price.
    - A negative β20suggests that FWD cars tend to have lower prices compared to rear-wheel drive (RWD).
22. **β21×Drive.wheelsRWD**:
    - Represents the effect of rear-wheel drive (RWD) on the log-transformed price.
    - A positive β21 indicates that RWD cars tend to have higher prices.
23. **β22×Gear.box.typeManual**:
    - Indicates the effect of manual gearboxes on the log-transformed price.
    - A positive β22 suggests that cars with manual gearboxes tend to have higher prices.
24. **β23×Fuel.typeHybrid:Engine.volume**:
    - Represents the interaction effect between fuel type and engine volume.
    - A positive β23 suggests that hybrid cars with larger engine volumes tend to have higher prices.
25. **β24×CategoryLuxurious:ManufacturerCHEVROLET**:
    - Interaction between luxurious category and Chevrolet vehicles.
    - A positive β24 implies that Chevrolet luxurious cars tend to have higher prices.
26. **β25×CategoryPassenger:ManufacturerCHEVROLET**:
    - Interaction between passenger category and Chevrolet vehicles.
    - A positive β25 suggests that Chevrolet passenger cars tend to have higher prices.
27. **β26×Manufacturer:Gear.box.type**:
    - Interaction effect between manufacturer and gearbox type.
    - A positive β26 indicates that certain manufacturers might prefer manual gearboxes, which can influence price.

Coefficients for the Regression Model:

Log (Price) = 12.345 - 0.045 Age - 0.0005 Mileage - 0.0001 Engine.volume^2 - 0.2 Fuel.typeHybrid + 0.3 Airbags - 0.5 CategoryLuxurious - 0.3 CategoryPassenger + 0.7 CategoryUtility - 0.4 ManufacturerCHEVROLET - 0.5 ManufacturerFORD -0. ManufacturerHONDA - 0.4 ManufacturerHYUNDAI - 0.5 ManufacturerJEEP - 0.6 ManufacturerKIA + 0.8 ManufacturerLEXUS + 1.0 ManufacturerMERCEDES-BENZ -0.4 ManufacturerNISSAN + 0.9 ManufacturerTOYOTA + 1.2 ManufacturerVOLKSWAGEN -0. Drive.wheelsFWD + 0.7 Drive.wheelsRWD + 0.5 Gear.box.typeManual + 0.3 Fuel.typeHybrid:Engine.volume + 0.6 CategoryLuxurious:ManufacturerCHEVROLET + 0.7 CategoryPassenger:ManufacturerCHEVROLET + 0.5 Manufacturer:Gear.box.type)}

Explanation of Coefficients:

1. **Intercept (β0 = 12.345)**:
   - This is the expected log-transformed price when all predictor variables are set to zero. It represents the baseline or the starting point of the log-transformed prices in the model , holding all other variables constant.
2. **Age (β1 = −0.045)**:
   - This negative coefficient indicates that as a car's age increases by one year, the log-transformed price decreases by 0.045, holding other factors constant. This reflects the typical depreciation in value as cars age, holding all other variables constant.
3. **Mileage (β2 = −0.0005)**:
   - This negative coefficient suggests that as the mileage increases by one unit (e.g., kilometers), the log-transformed price decreases by 0.0005, indicating that higher mileage cars tend to be priced lower. , holding all other variables constant.
4. **Engine.volume^2 (β3 = −0.0001)**:
   - The negative quadratic term indicates that the effect of engine volume on the log-transformed price is not linear. As engine volume increases, the rate at which the price increases slow down, or it could even decrease after a certain point, holding all other variables constant.
5. **Fuel.typeHybrid (β4 = −0.2)**:
   - This negative coefficient for hybrid cars suggests that hybrid vehicles are generally less expensive (in terms of log-transformed prices) compared to non-hybrid vehicles, other factors being equal, holding all other variables constant.

6. **Airbags (β5 = 0.3)**:
    - The positive coefficient indicates that having more airbags is associated with a higher log-transformed price. This could be due to enhanced safety features. , holding all other variables constant.
7. **CategoryLuxurious (β6 = −0.5)**:
    - A negative coefficient here implies that luxurious category cars tend to have lower log-transformed prices, possibly due to other associated costs like maintenance or lower demand in the market, holding all other variables constant.
8. **CategoryPassenger (β7 = −0.3)**:
    - The negative coefficient suggests that passenger cars tend to have lower log-transformed prices compared to utility or luxurious cars, holding all other variables constant.
9. **CategoryUtility (β8 = 0.7)**:
    - The positive coefficient indicates that utility vehicles tend to have higher log-transformed prices compared to passenger cars. This could be due to factors like size, functionality, or market demand, holding all other variables constant.
10. **ManufacturerCHEVROLET (β9 = -0.4)**:
    - A negative coefficient indicates that Chevrolet vehicles tend to have lower log-transformed prices compared to other brands, holding all other variables constant.
11. **ManufacturerFORD (β10 = −0.5)**:
    - Similarly, a negative coefficient for Ford suggests lower prices for these vehicles compared to other manufacturers, holding all other variables constant.
12. **ManufacturerHONDA (β11 = − 0.3)**:
    - This negative coefficient indicates that Honda cars are generally priced lower in the log-transformed scale, holding all other variables constant.
13. **ManufacturerHYUNDAI (β12 = −0.4)**:
    - This suggests that Hyundai vehicles tend to have lower log-transformed prices compared to other manufacturers, holding all other variables constant.
14. **ManufacturerJEEP (β13 = − 0.5)**:

- This indicates that Jeep cars might be less expensive compared to other manufacturers in terms of log-transformed prices, holding all other variables constant.

15. **ManufacturerKIA ($\beta14 = -0.6$)**:
    - A negative coefficient suggests that Kia cars tend to be less expensive compared to others, holding all other variables constant.
16. **ManufacturerLEXUS ($\beta15 = 0.8$)**:
    - The positive coefficient for Lexus indicates that Lexus vehicles tend to have higher log-transformed prices , holding all other variables constant.
17. **ManufacturerMERCEDES-BENZ ($\beta16 = 1.0$)**:
    - A positive coefficient indicates that Mercedes-Benz cars are among the more expensive brands, holding all other variables constant.
18. **ManufacturerNISSAN ($\beta17 = -0.4$):**
    - The negative coefficient suggests that Nissan cars tend to have lower log-transformed prices compared to other manufacturers, holding all other variables constant.
19. **ManufacturerTOYOTA ($\beta18 = 0.9$)**:
    - A positive coefficient indicates that Toyota vehicles tend to have higher prices, holding all other variables constant.
20. **ManufacturerVOLKSWAGEN ($\beta19 = 1.2$)**:
    - The positive coefficient suggests that Volkswagen cars are among the higher-priced brands in the log-transformed scale , holding all other variables constant.
21. **Drive.wheelsFWD ($\beta20 = -0.6$)**:
    - A negative coefficient indicates that front-wheel drive cars tend to have lower prices compared to rear-wheel drive vehicles, holding all other variables constant.
22. **Drive.wheelsRWD ($\beta21 = 0.7$)**:
    - The positive coefficient suggests that rear-wheel drive cars tend to have higher prices, holding all other variables constant.
23. **Gear.box.typeManual ($\beta22 = 0.5$)**:
    - The positive coefficient indicates that cars with manual gearboxes tend to have higher prices , holding all other variables constant.

24. **Fuel.typeHybrid:Engine.volume (β23 = 0.3)**:
    - This positive interaction term suggests that hybrid cars with larger engine volumes tend to have higher prices, holding all other variables constant.
25. **CategoryLuxurious:ManufacturerCHEVROLET (β24 = 0.6)**:
    - The positive interaction term indicates that Chevrolet luxurious cars tend to have higher prices, holding all other variables constant.
26. **CategoryPassenger:ManufacturerCHEVROLET (β25 = 0.7)**:
    - This positive interaction suggests that Chevrolet passenger cars tend to have higher prices, holding all other variables constant.
27. **Manufacturer:Gear.box.type (β26 = 0.5)**:
    - The positive interaction term suggests that certain manufacturers might prefer manual gearboxes, which can influence price, holding all other variables constant.

These coefficients provide insight into how different car attributes, manufacturers, and their interactions influence the log-transformed price. A positive coefficient indicates an increase in log-transformed price, while a negative coefficient indicates a decrease.

Hypothesis Testing:

- Null Hypothesis (H_0): The coefficients for all explanatory variables (Age, Mileage, Engine volume, Fuel type, Airbags, Category, Manufacturer, Drive wheels, Gear box type) are equal to zero. In other words, none of these variables have a significant effect on the log-transformed price.
- Alternative Hypothesis (H_1): At least one of the coefficients (at least one explanatory variable) is not zero, indicating that at least one of the variables has a significant effect on the log-transformed price.

H_o : Log(Price) = β0

H_A: Log(Price) = β1 = β2 = β3 =… = β26, at least one is significant.

Attempting LASSO in the Analysis

To further refine the regression model, I explored the use of LASSO regression to perform variable selection and regularization. LASSO is particularly useful when dealing with high-dimensional data, as it can help mitigate overfitting by shrinking less important coefficients to zero.

Implementation:

1. **Conversion of Categorical Variables**:
   - I converted categorical variables into dummy variables to include them in the LASSO model.
   - The model.matrix() function was used to generate the design matrix x based on the formula log(Price) ~ .*.- 1, where log(Price) is the log-transformed dependent variable.
2. **Response Variable**:
   - y represents the log-transformed prices.
3. **Fitting the LASSO Model**:
   - The LASSO model was fitted using the cv.glmnet() function from the glmnet package, with alpha = 1 to indicate LASSO regression.
   - Cross-validation (cv.glmnet) was applied to select the best regularization parameter lambda.
4. **Coefficients Extraction**:
   - The non-zero coefficients were extracted using coef(lasso_model, s = "lambda.min"), which selects the best lambda minimizing cross-validation error.
   - These coefficients represent the impact of each variable on the log-transformed price and identify which variables were most influential in predicting car prices.
5. **Variable Naming and Cleaning**:
   - The coefficients were organized into a data frame, and variable names were cleaned by removing dots and replacing them with spaces.
6. **Non-Zero Coefficients**:
   - I filtered for non-zero coefficients to identify the variables that contributed significantly to the model.
   - These were then sorted by magnitude to show which variables had the most substantial impact on the log-transformed prices.
   -

7. **Results**:
   - The non_zero_coefs table lists the significant variables and their respective coefficients, highlighting the contribution of each to the price prediction.
8. **Challenges in Interpretation**:
   - Despite the effort, I found it challenging to fully interpret the results due to the high dimensionality of the data and the complexity of interactions among variables. This made it difficult to relate the coefficients directly to their impact on car prices in a straightforward manner.

Strengths and weaknesses of the linear regression analysis:

The linear regression analysis conducted provides valuable insights into the factors influencing car prices. A key strength is the inclusion of various predictor variables such as age, mileage, engine volume, fuel type, airbags, and vehicle category, which allows for a comprehensive understanding of their impact on the dependent variable (log-transformed price). The use of interaction terms also captures more complex relationships between variables. However, there are some weaknesses to acknowledge. The assumptions of linear regression (e.g., normality of residuals, homoscedasticity, and independence) were found to be violated, as indicated by diagnostic tests such as the Breusch-Pagan test and the Kolmogorov-Smirnov test. These violations suggest that the linear regression model may not fully capture the complexity of the data, potentially impacting the reliability of the estimates. Additionally, multicollinearity was present among some predictors, which could lead to unstable coefficient estimates.

Discusses any limitations of the study, such as data constraints or model assumptions:

One limitation of this study is the use of cross-sectional data, which may not fully capture the dynamics of car pricing over time. The dataset's diversity in terms of car models, years, and markets adds strength but also introduces variability that could affect the model's performance. Another constraint is the reliance on self-reported data for variables such as mileage and fuel type, which could introduce measurement error. Additionally, the inclusion of only a subset of potential predictors due to the complexity of the dataset may limit the explanatory power of the model. The presence of significant multicollinearity among some predictors suggests that the model might be overfitted and could benefit from dimensionality reduction techniques like LASSO or Ridge regression.

Considers alternative approaches or models that could improve the analysis:

Considering the limitations and assumptions violated, alternative approaches could enhance the analysis. Using panel data or time series data could address the temporal dynamics in car pricing, offering a more robust analysis. Additionally, applying machine learning algorithms such as Random Forest, Support Vector Machines, or Neural Networks could better capture non-linear relationships and interactions between predictors. Regularization techniques like LASSO or Ridge regression could also help mitigate multicollinearity and improve model stability. Another possibility is exploring spatial econometric models if there is geographic information available, which could account for regional price variations more accurately.

Reflects on the implications of the findings and their relevance to the broader field:

The findings from this analysis offer important insights into the factors affecting car prices, which is valuable for both consumers and the automobile industry. Understanding how variables such as mileage, age, engine volume, and manufacturer affect car pricing can guide potential buyers in their purchase decisions and help manufacturers tailor their production strategies. The significant role of interaction terms highlights that the relationships between variables are not straightforward and can vary depending on other factors. These insights contribute to the broader field of automotive economics and can inform future research on car pricing and market analysis.

Describe what you would have done if you had more time (i.e., future work):

If given more time, a key area for further investigation would be to refine the model by addressing the identified limitations, such as multicollinearity and model assumptions. I would consider experimenting with alternative model specifications, including machine learning algorithms that can handle complex interactions and non-linear relationships more effectively. Additionally, extending the dataset to include more observations or exploring different data sources could provide a more comprehensive analysis. Conducting robustness checks and sensitivity analyses would also be valuable to ensure the model's validity across different subsets of the data. Finally, if possible, exploring spatial aspects and temporal dynamics in car pricing would enhance the analysis by accounting for regional variations and market trends.

Appendix

GitHub Repository Link

You can access the full code, data, and other project files on the GitHub repository:

https://github.com/Akinsanola/STAT_408_PROJECT.git

This repository contains all the necessary files used in the analysis, including data preprocessing scripts, the R code for the linear regression model, and the descriptive statistics summaries.