Predicting Used Car Prices using Regression Analysis

Satkar Karki
University of South Dakota
DSCI 724 : Data Mining for Managers
Submitted to: Dr. Thomas Tiahrt
December 8, 2024

Table of Contents

1. Introduction

The goal of this project is to build a predictive model that accurately estimates the price of used cars based on key attributes such as year of manufacture, mileage, engine size, and other relevant factors. Using multiple regression analysis, the model provides a fair price estimate by evaluating these predictors. In essence, this tool assists both buyers seeking a reasonable market price and sellers aiming to set a fair value for their used cars.

2. Business Problem Description (what problem does the predictive model solve)

A core challenge in the used car industry is determining fair pricing. As buyers seek affordable deals and sellers strive for competitive pricing, the pricing for used cars is often inconsistent. The demand for used cars is large due to cost-effectiveness, availability, and value retention. This creates an opportunity for a data-driven pricing solution to maintain fairness in valuations. A predictive model incorporating factors like the car's age, condition, mileage, and other relevant factors would provide balanced pricing that benefits both buyers and sellers. The motivation for this project stems from the widespread sentiment among consumers in the used car market, where pricing inconsistencies create frustration.

3. Data Source (from Project Proposal)

The dataset is sourced from Kaggle, a widely used platform for data-science projects. The dataset comprises 10,668 observations of used Audi car listings scraped from various multiple car listing websites and includes 9 variables. The dataset doesn't have any missing values which have been verified by using the "*sum(is.na())"* function in R software. This dataset is appropriate for the project as it captures a wide range of variables known to affect car prices and are mostly numeric.

URL to the Dataset: https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes

4. Variable Description

- **Price**: The response variable for this analysis is the price of the used Audi cars, measured in Pound Sterling (GBP). This variable represents the selling price listed for each used

Audi car in the dataset. Since the main objective is to predict the price of used cars, this variable will be central to the regression model.

- **Model**: This variable represents the model of the Audi car and is categorical. There are numerous models captured in the dataset as different models will affect customer perception about the car.
- **Year**: This variable indicates the year the car was manufactured and is continuous. As depreciation over time plays a crucial role in determining the current value of a used car, this variable will be significant as a predictor in the model.
- **Transmission**: This variable is the second categorical variable in the dataset which denotes whether the car has manual, semi-automatic, or automatic transmission.
- **Mileage**: The total number of miles traveled by the vehicle is represented by mileage which is a continuous variable. It is expected that the value of a used car with higher mileage to depreciate and will be an important predictor in this model.
- **Fuel Type**: The variable captures the type of fuel that the car uses (e.g., petrol, diesel) and is categorical. Fuel efficiency is crucial in determining a car's desirability which will certainly prove to be a significant predictor of a used car's price.
- **Tax**: The tax variable reflects the yearly road tax that the car incurs and is based on the car's CO2 emissions, engine size, and such. An important factor while considering the price for a used car, this variable is continuous.
- **MPG**: Miles per gallon (MPG) is the indicator of a vehicle's fuel efficiency and is a continuous variable. Cars with high MPG are expected to have a higher value and will be a significant variable in this predictive model.
- **Engine Size**: Engine size, a continuous variable, reflects the car's engine capacity in liters. This will also be a significant predictor as larger engines generally provide more power but can also be less fuel-efficient.

5. Model Specification

The model is based on a multiple regression framework, adhering to the generic formula.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k + e$$

where:

- y is the response variable.
- x represents the predictor variables.
- β_k represents the coefficient for k predictors.
- β_0 is the intercept of the regression line.
- e is the error term capturing the variability.

However, the response variable, which is the price of the listed used cars, was log-transformed. This transformation was applied primarily to address proportional differences between the predictor values and response variable, as well as to manage the skewness caused by high-value outliers from expensive car listings. The decision to log-transform the response variable was further validated by examining the residual plots before and after the transformation. The residual plots showed a more stabilized and homoscedastic pattern after applying the log transformation, indicating a better model fit and improved interpretability.

The mathematic formula for the model has been provided below:

$$\log(price) = \beta_0 + \beta_1 (Year) + \beta_2 (Transmission) + \beta_3 (Model) + \beta_4 (FuelType) + \beta_5 (MPG) + \beta_6 (Mileage) + \beta_7 (EngineSize) + \beta_8 (Tax) + e$$

In the regression model, categorical variables were encoded using dummy variables, where one category (the reference level) serves as the baseline with its effect being captured in the intercept (β_0). For instance, the variable *Transmission* has three categories out of which "Automatic" was set as the reference in this model. The co-efficient for other two categories were included in the model indicating how these transmission types influence the log-transformed price relative to automatic vehicles with the variability associated with the reference level accounted for by the

intercept term. All the variables in the model had very small p-values and were retained in the final model.

6. Predictive Results

6.1. Interaction Terms

This model does not contain interaction terms despite attempts to evaluate potential interactions, such as between *"mpg"* and *"EngineSize"*. These tests did not improve either the model metrics or its predictive accuracy. Further, exclusion of interaction terms help avoids unnecessary complexity in the model.

6.2. Non-linear Predictor Transformations

The three categorical variables – *"model", "transmission",* and *"fuelType"* were transformed because R identified them as character variables. The categorical variables were converted into factors using the *as.factor()* function in R. By turning them into factors, the model could understand that these variables represent different categories, like different car models, types of transmission, or fuel types. Due to this, the overall interpretability and accuracy of the model in predicting the price of user cars improved. An attempt to log-transform the *mileage* variable was made, however, it did not improve the model accuracy.

6.3. Descriptive Statistics

6.3.1 Correlation Matrix

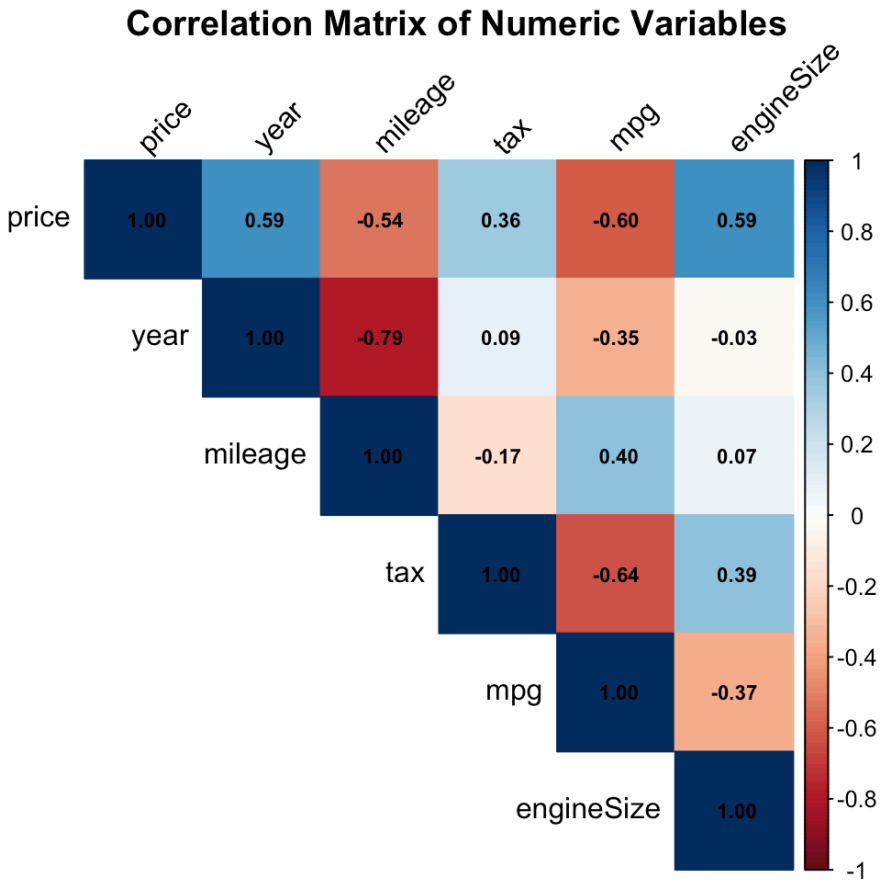

FIGURE 1 *The correlation matrix.*

The correlation matrix suggest that price is positively correlated with year and engine size which is quite expected because newer cars and those with larger engines tend to be expensive. Conversely, price is negatively correlated with mileage and mpg as cars with higher mileage and better fuel efficiency are typically priced lower. Among the predictor variables, mileage and year exhibit a strong negative correlation (-0.79), suggesting a potential multicollinearity issue. However, the subsequently calculated VIF results confirm that multicollinearity is not a concern in the model.

6.3.2 Means and standard deviations

Variable	Mean	Standard Deviation
year	2017.1	2.17
price	22896.69	11714.84
mileage	24827.24	23505.26
tax	126.01	67.17
mpg	50.77	12.95
engineSize	1.93	0.6
log(price)	9.93	0.47

TABLE 1 *Means and standard deviations for variables.*

6.3.3 Graphs and Box plots

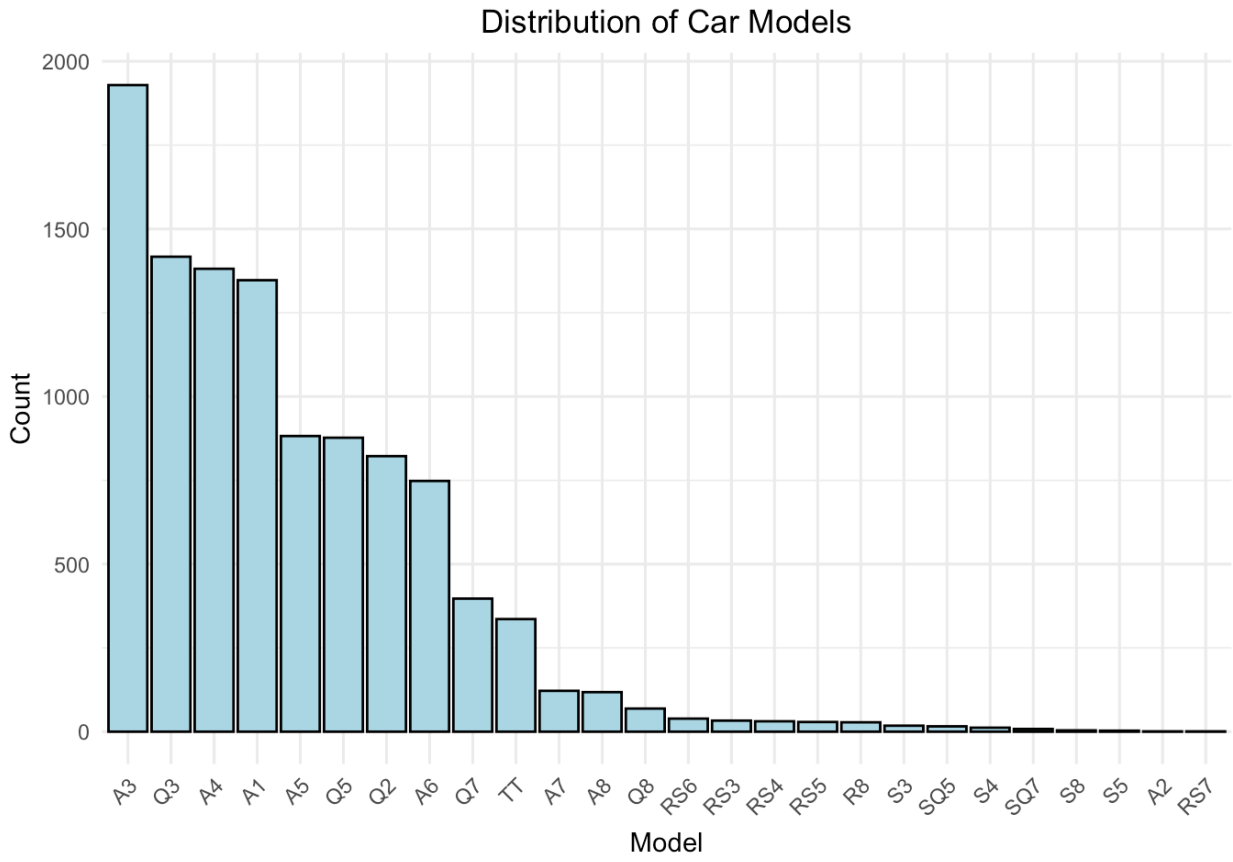

FIGURE 2 *Bar-chart displaying the distribution of car models in the Audi dataset.*

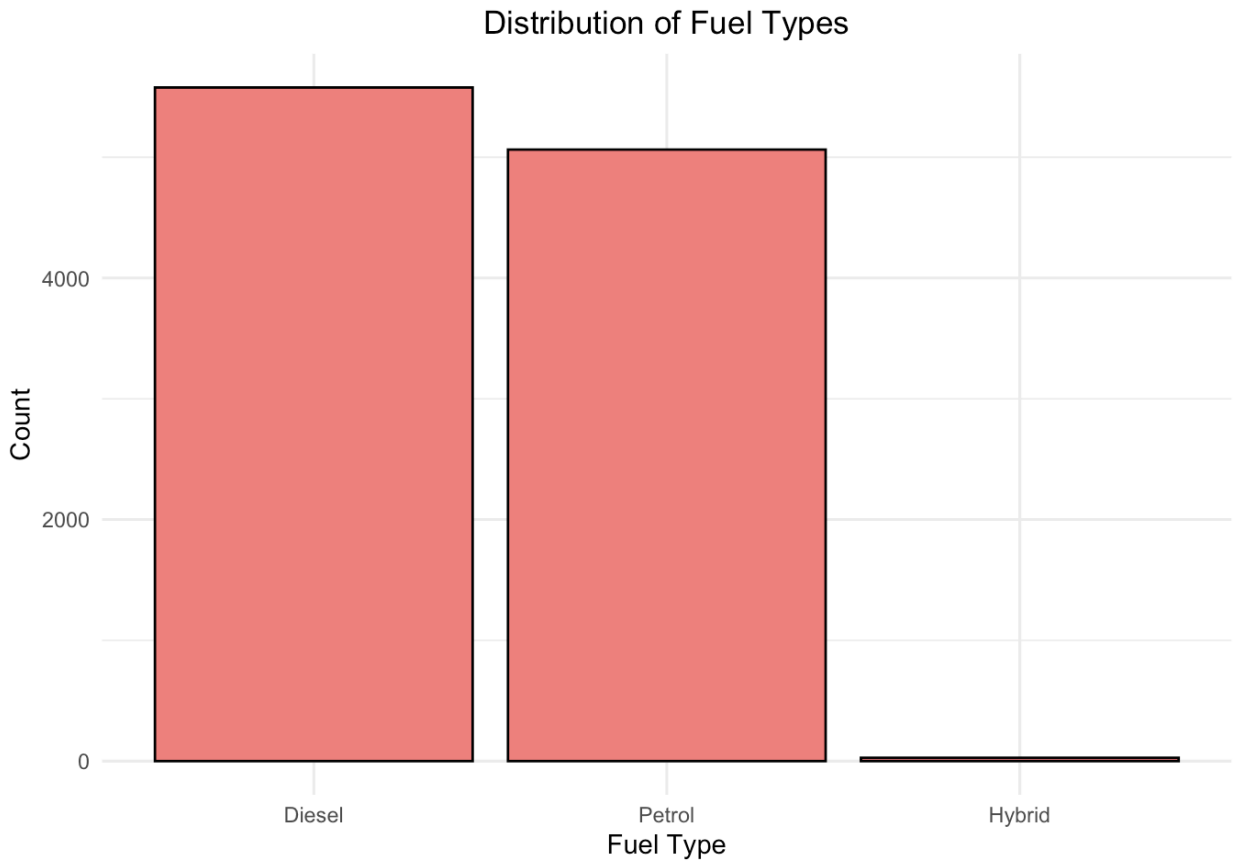

FIGURE 3 *Bar-chart displaying the distribution of fuel type for cars listed in the Audi dataset.*

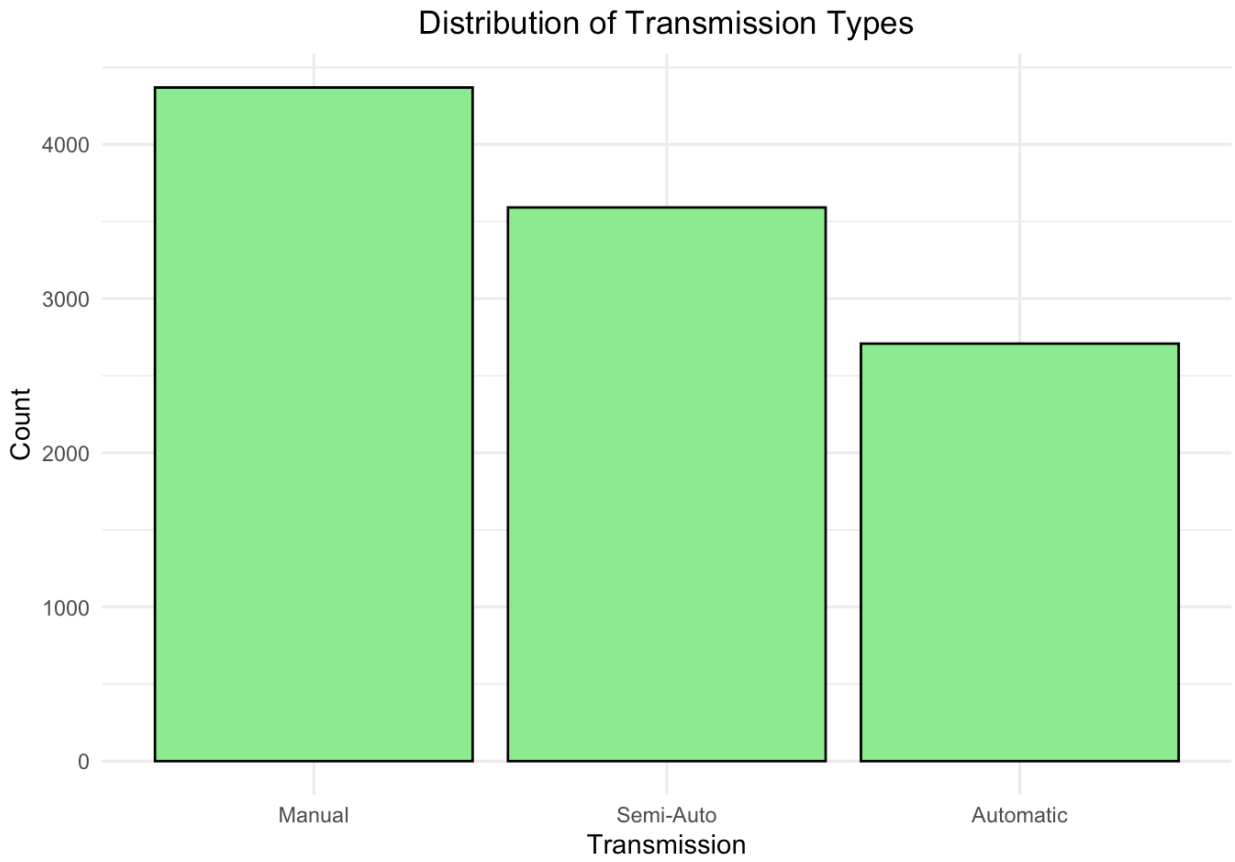

FIGURE 4 *Bar-chart displaying the distribution of transmission types among cars listed in the Audi dataset.*

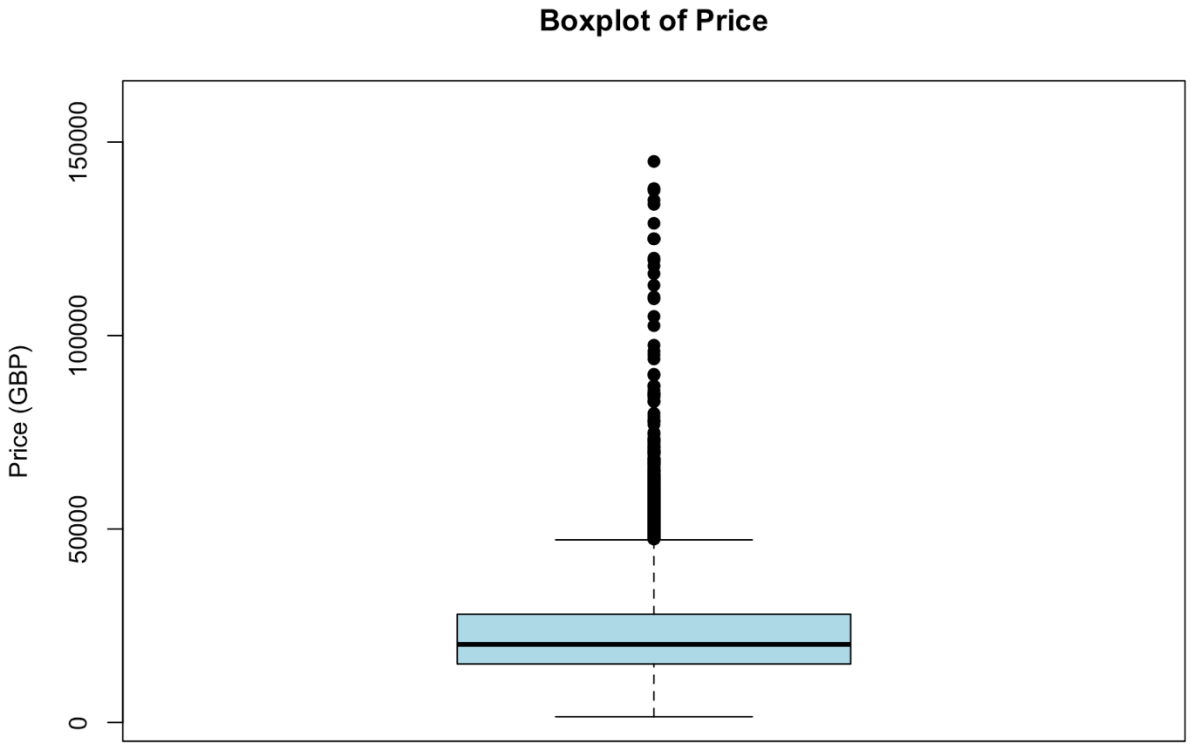

FIGURE 5 *Boxplot of car prices in the Audi dataset.*

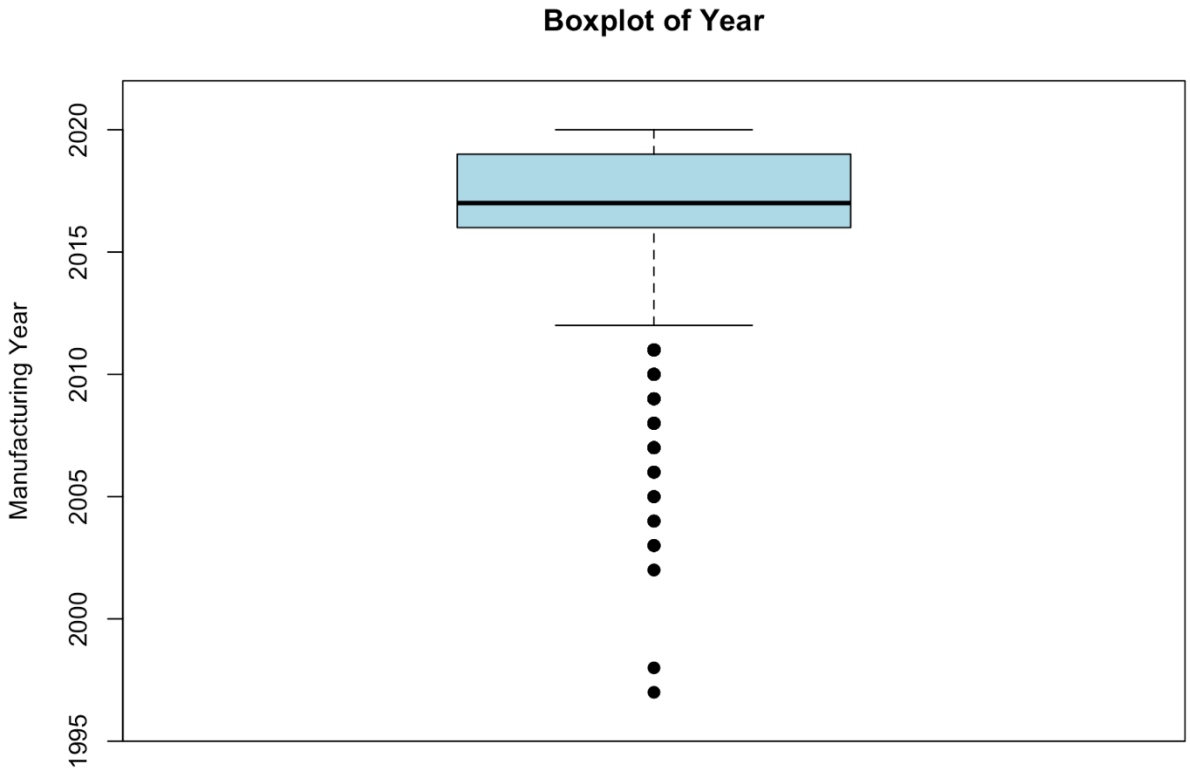

FIGURE 6 *Boxplot of car manufacturing year.*

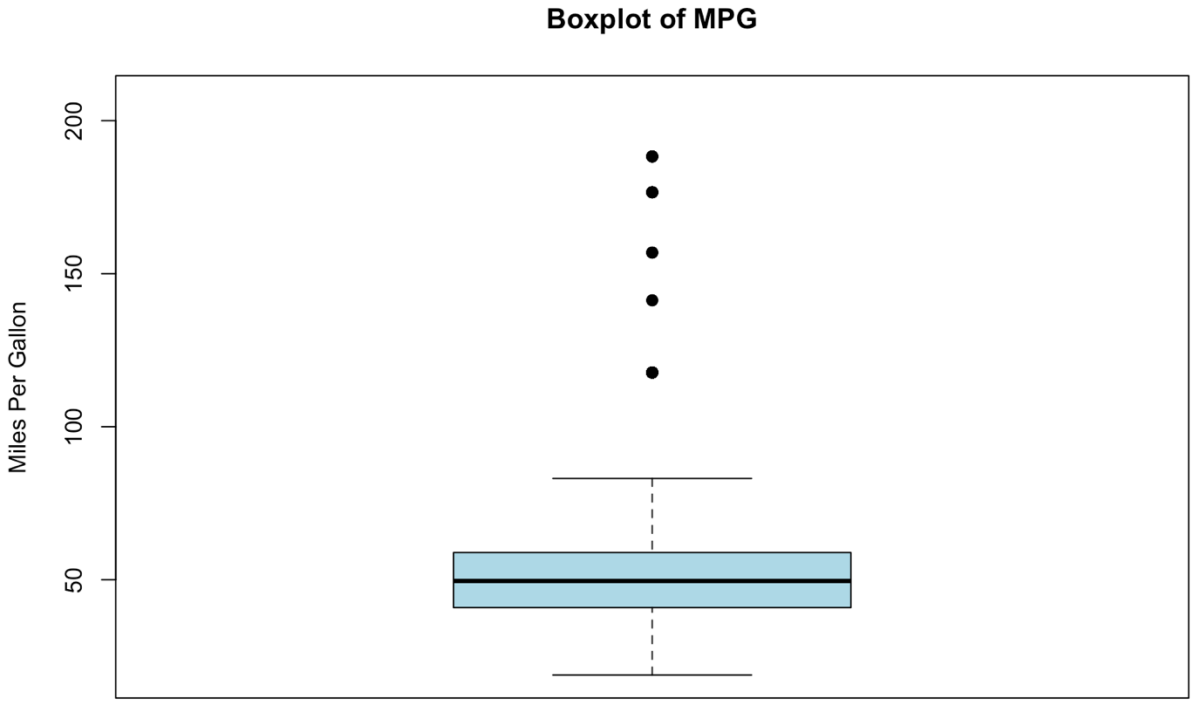

FIGURE 7 *Boxplot of fuel efficiency (MPG).*

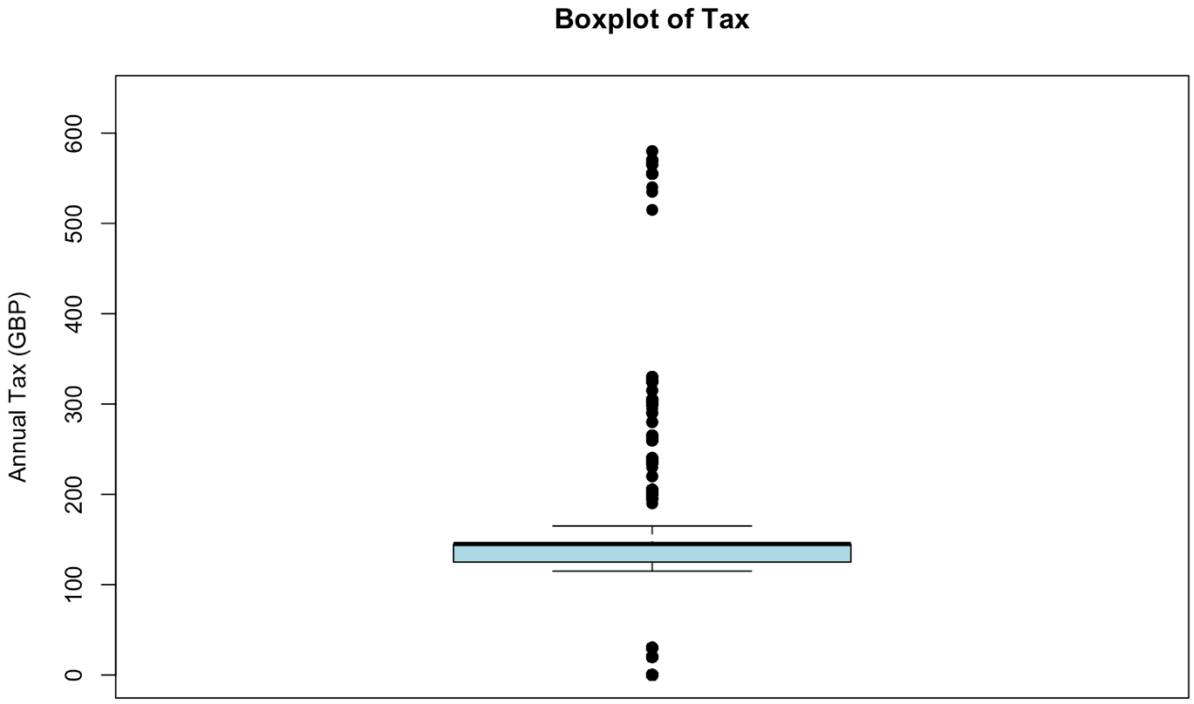

FIGURE 8 *Boxplot of annual road tax.*

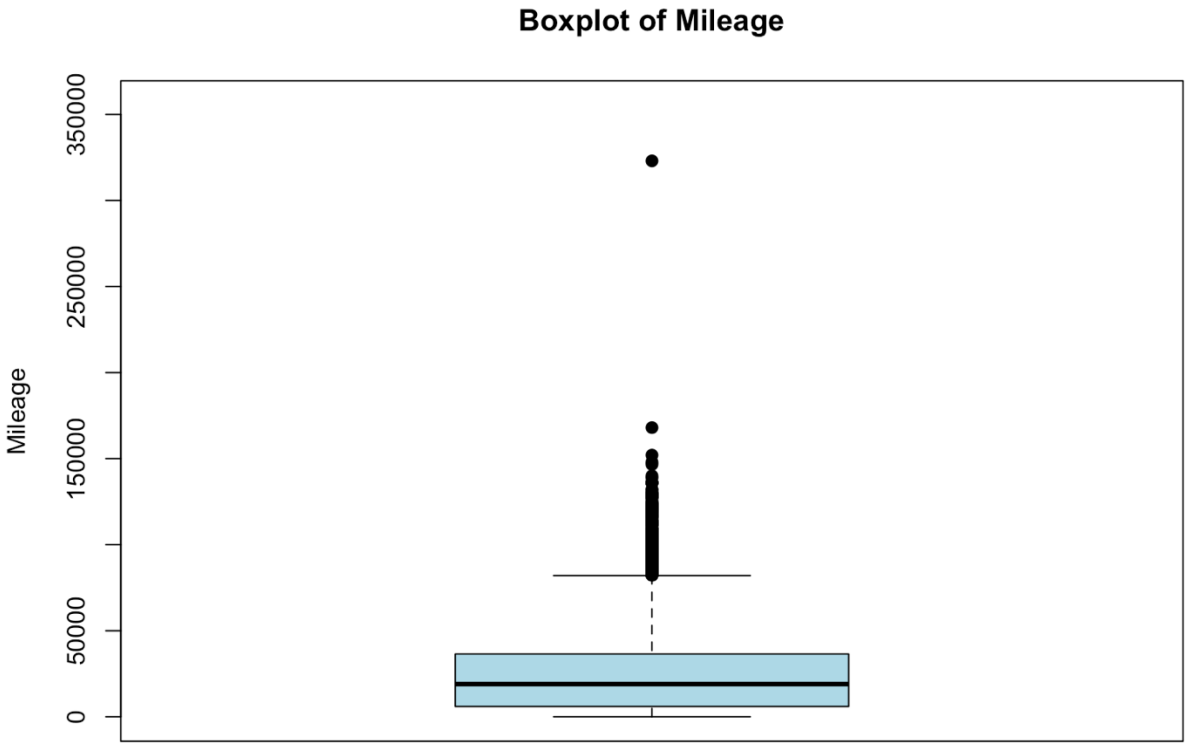

FIGURE 9 *Boxplot of car mileage.*

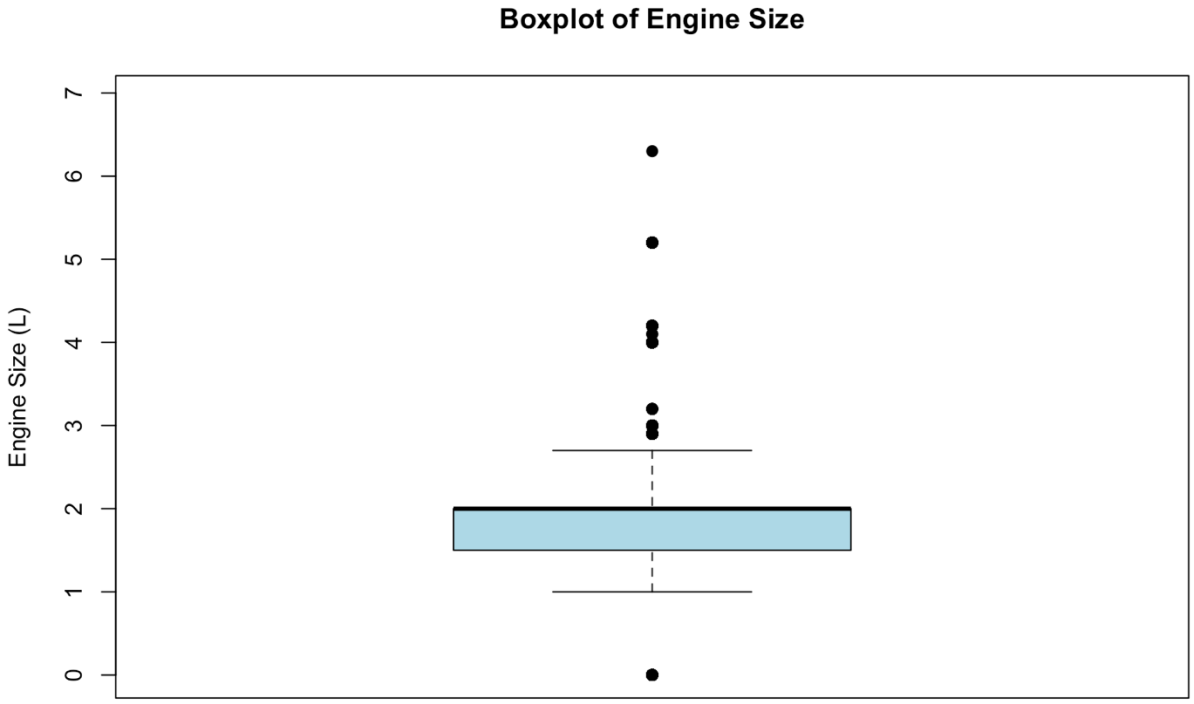

FIGURE 10 *Boxplot of engine size.*

6.4. Summary of Evaluation Metrics

Evaluation Metric	Value
R-Squared (R^2)	0.9374 *(adjusted: 0.9372)*
Mean Absolute Error (MAE)	2056.97
Root Mean Squared Error (RMSE)	3067.83

TABLE 1 *Evaluation Metrics Summary*

The selected predictors can explain approximately 93.74% of the variance caused in used car prices. Furthermore, there is not much difference between the R^2 and the adjusted R^2 value suggesting that all the included predictors contribute meaningfully to the model. The MAE value of this model explains that on average the model's predicted price deviate from its true prices by about $2,056.79. RMSE also explains the variance between predicted values and actual values, but it assigns more weight to larger errors as residuals are squared before averaging. The RMSE value suggest that the average difference between predicted and actual values of the used cars is approximately $3067.83. Given the range of used car prices in the dataset, both the MAE and RMSE values are relatively low, indicating a good level of predictive accuracy. Overall, the combination of these metrics suggest that the model is robust and reliable for predicting used car prices, with a high degree of accuracy.

6.5. Plots of individual relationships

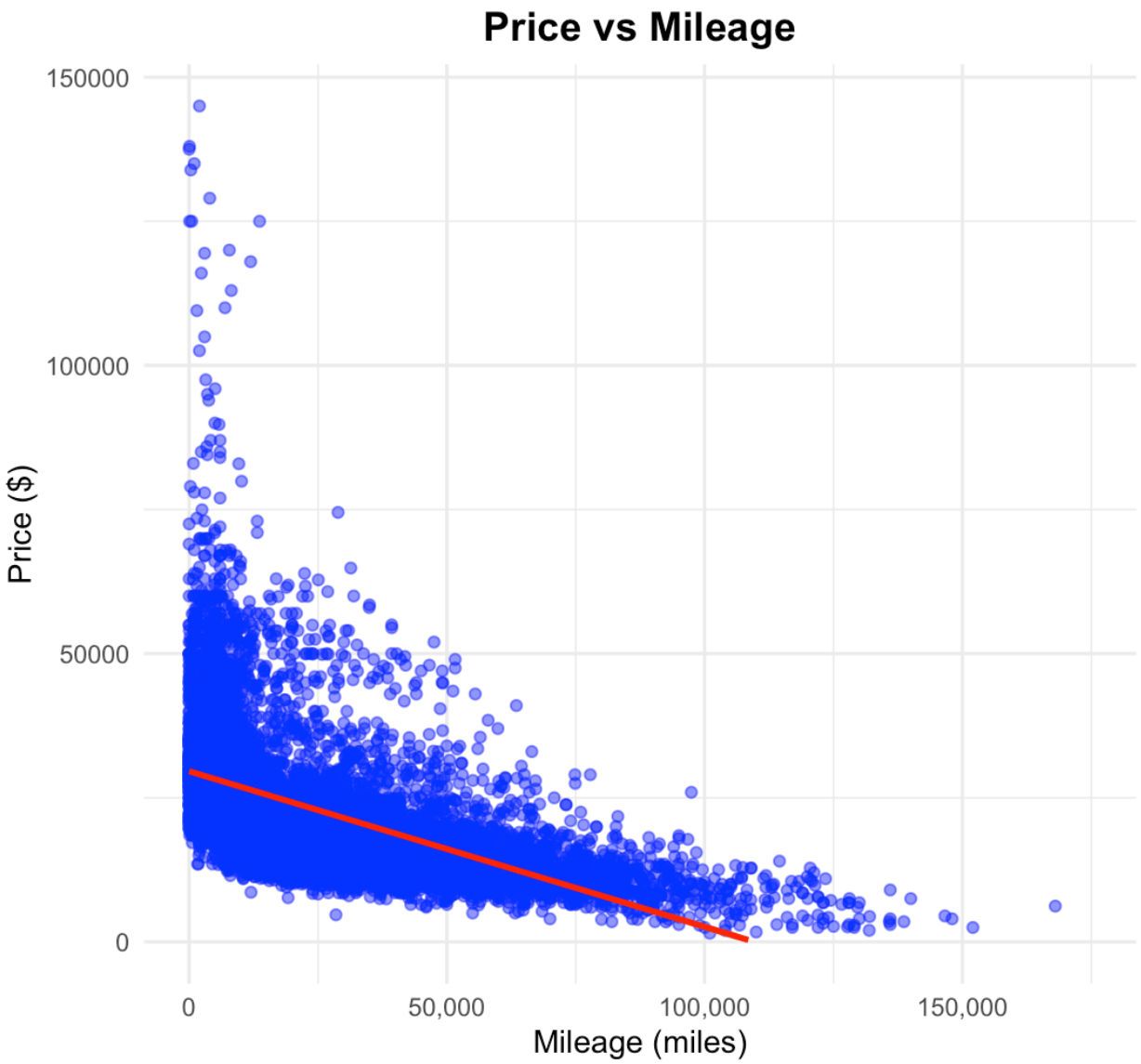

FIGURE 11 *Scatter plot showing the relationship between price and mileage.*

The downward trend of the red regression line indicates that the price of used cars decreases as mileage increases. Most of the cars in the dataset have mileage below 100,000, but their prices vary widely, suggesting the influence of other factors such as model of fuel efficiency. Beyond 100,000 miles, the data points become sparse, with most cars in this range priced under $25,000.

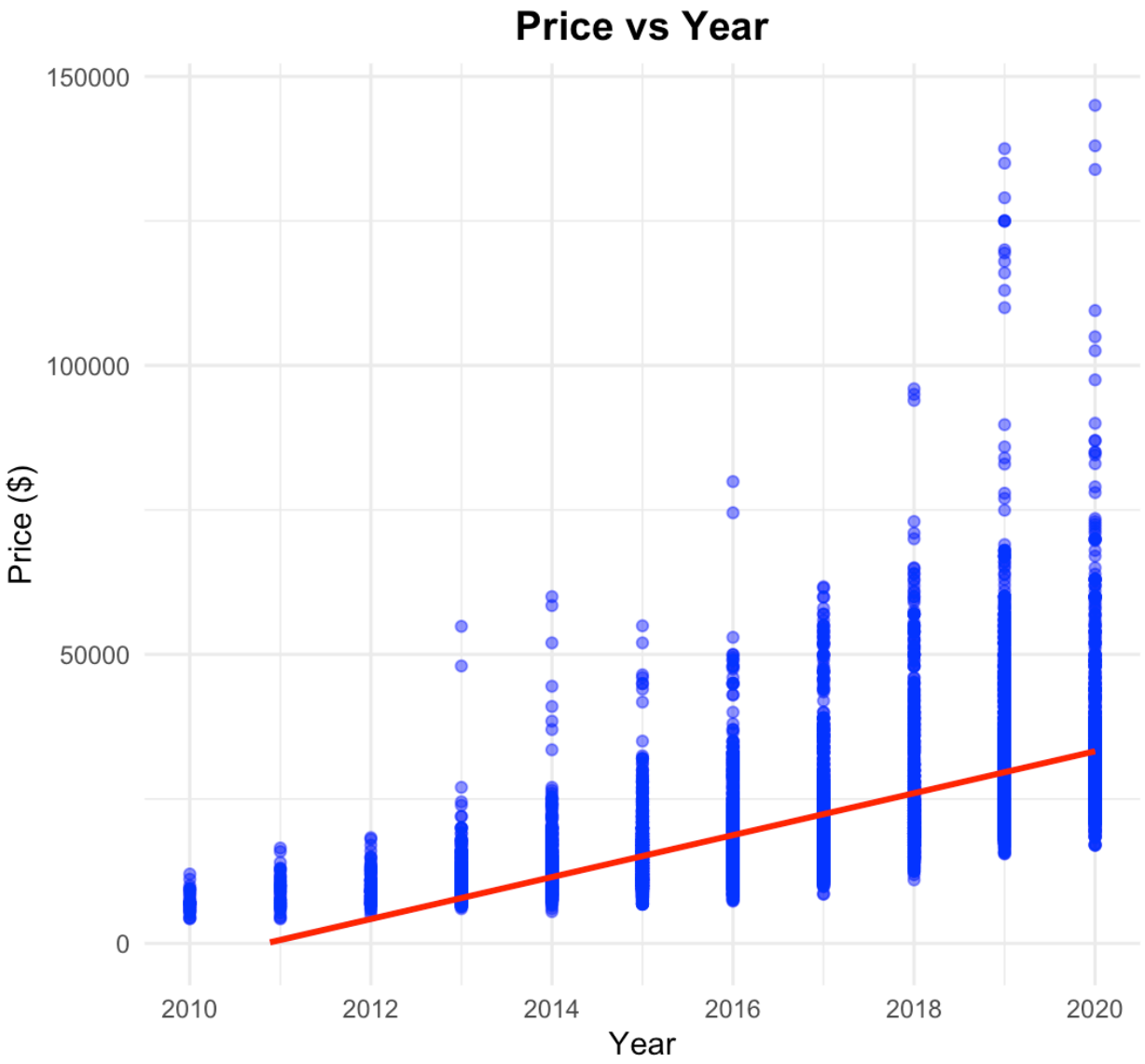

FIGURE 12 *Scatter plot showing the relationship between price and year.*

The upwards trend of the regression line indicates a positive correlation between price and year of manufacture. Cars manufactured closer to 2020 are priced higher than those from 2010. Most data points are clustered around newer years (2016 – 2020), suggesting lower representation of older models in the resale market. While many data points cluster within specific price ranges, newer models show greater price variation, with some priced significantly higher than others.

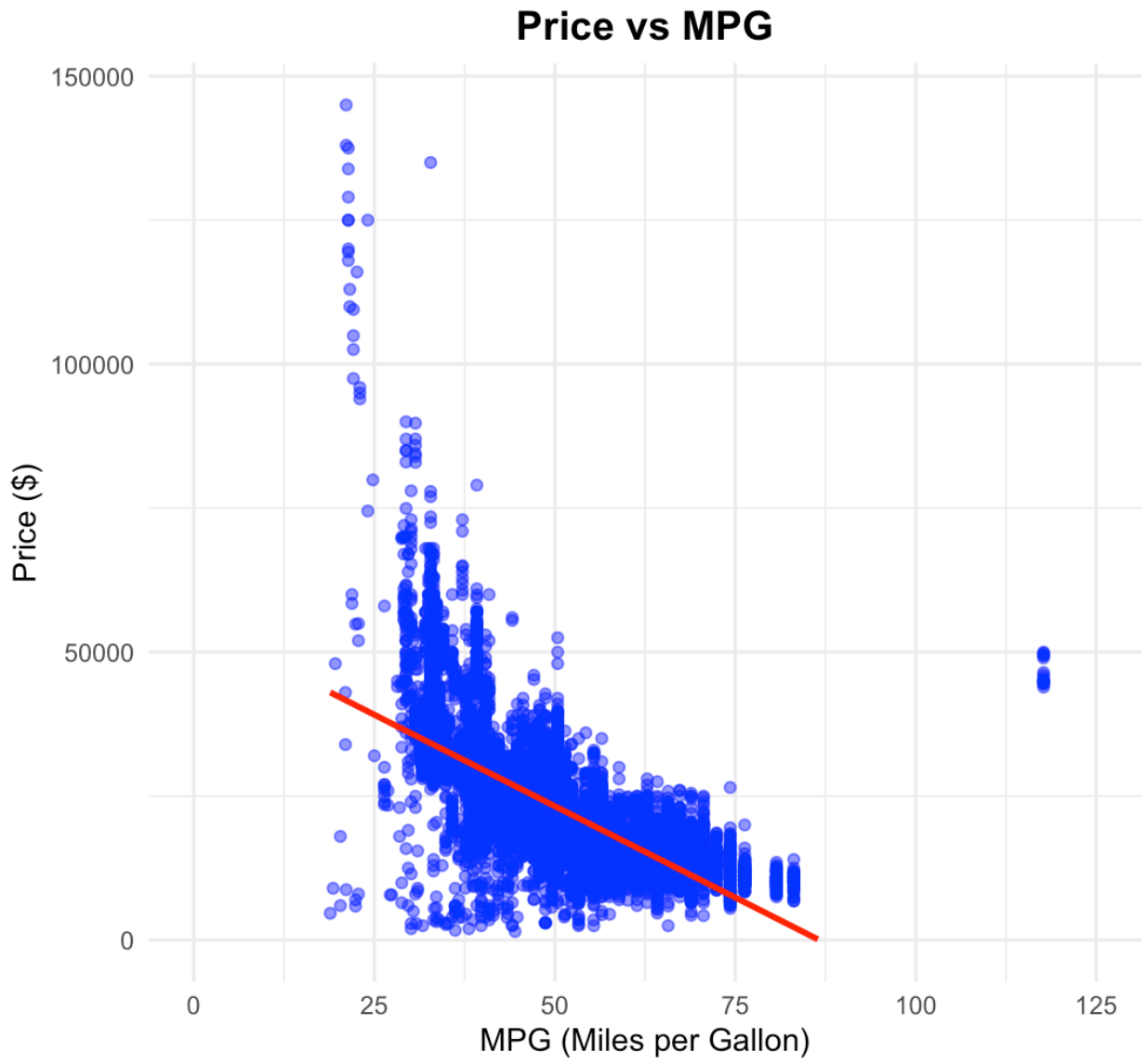

FIGURE 13 *Scatter plot showing the relationship between price and MPG.*

The scatterplot shows a negative correlation between price and MPG, indicating that cars with higher MPG are generally priced lower. High-MPG cars are typically designed and marketing for economy consumers and are priced accordingly. In contrast, expensive cars with powerful engines tend to have lower fuel efficiency but maintain higher resale values. The chart also highlights a few outliers exceeding 100 MPG, likely representing hybrid models captured in the dataset.

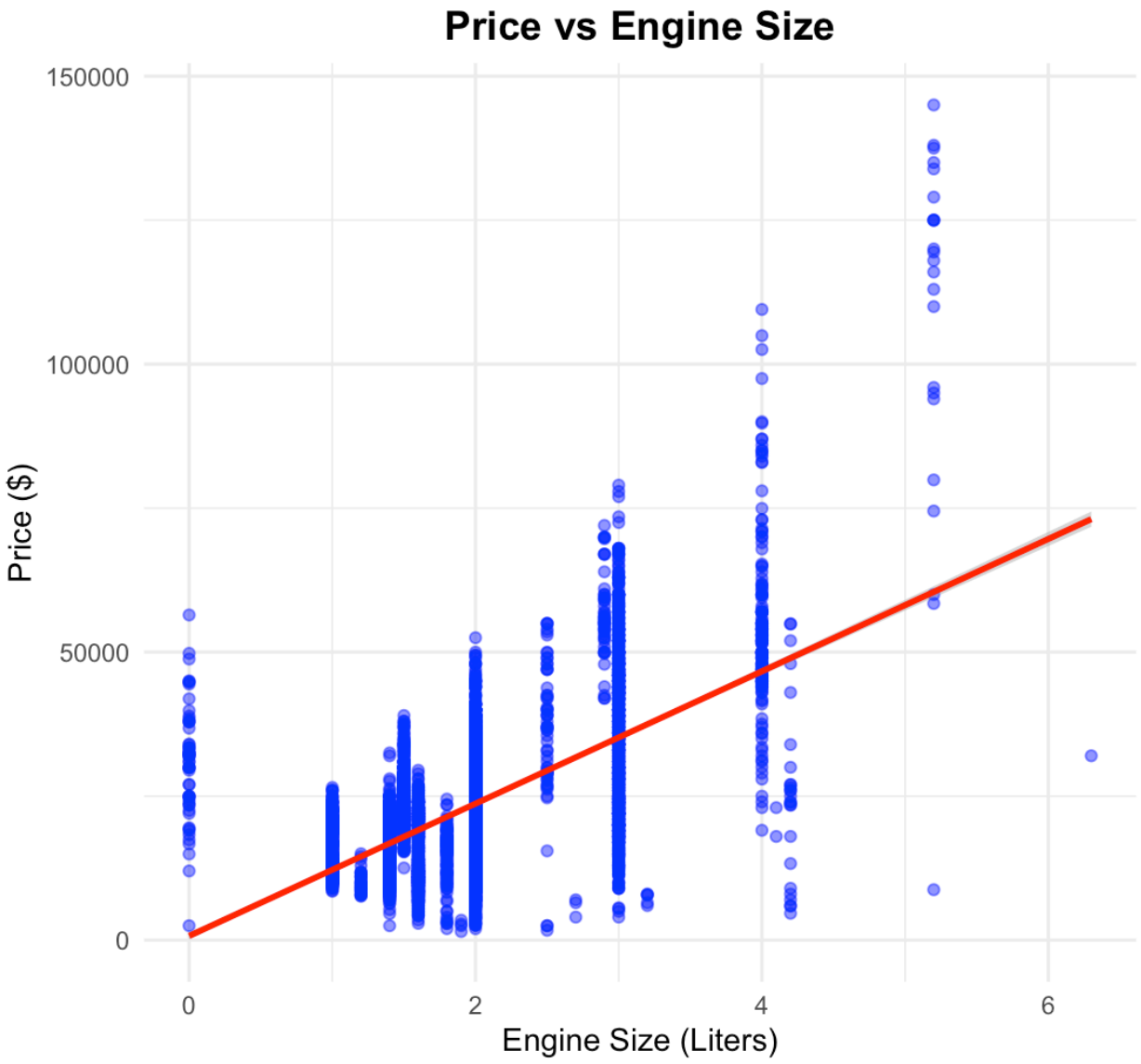

FIGURE 14 *Scatter plot showing the relationship between price and engine size.*

The upward trend of the regression line indicates that car prices increase with engine size. Luxury and premium models typically have more powerful engines, hence commanding a higher price. Most data points cluster around an engine size of 3, suggesting this is the average for listed cars. Beyond an engine size of 3, prices become widely dispersed, whereas smaller engine sizes show densely clustered price ranges.

FIGURE 15 *Scatter plot showing the relationship between price and tax.*

The relationship between price and tax is positively correlated, as indicated by the upward trend of the regression line. Higher taxes are typically associated with more expensive cars, though values become sparse beyond $400. Most data points are concentrated around lower tax values, suggesting a focus on mid-range and economy cars in the used car market.

FIGURE 16 *Box plot showing the relationship between price and transmission type.*

Manual transmission is the cheapest among the three types, as shown by its lower median line. All transmission types have outliers, with automatic and semi-automatic vehicles showing some highly priced ones. Additionally, automatic and semi-automatic cars have a wider price range, indicating greater variability in prices.

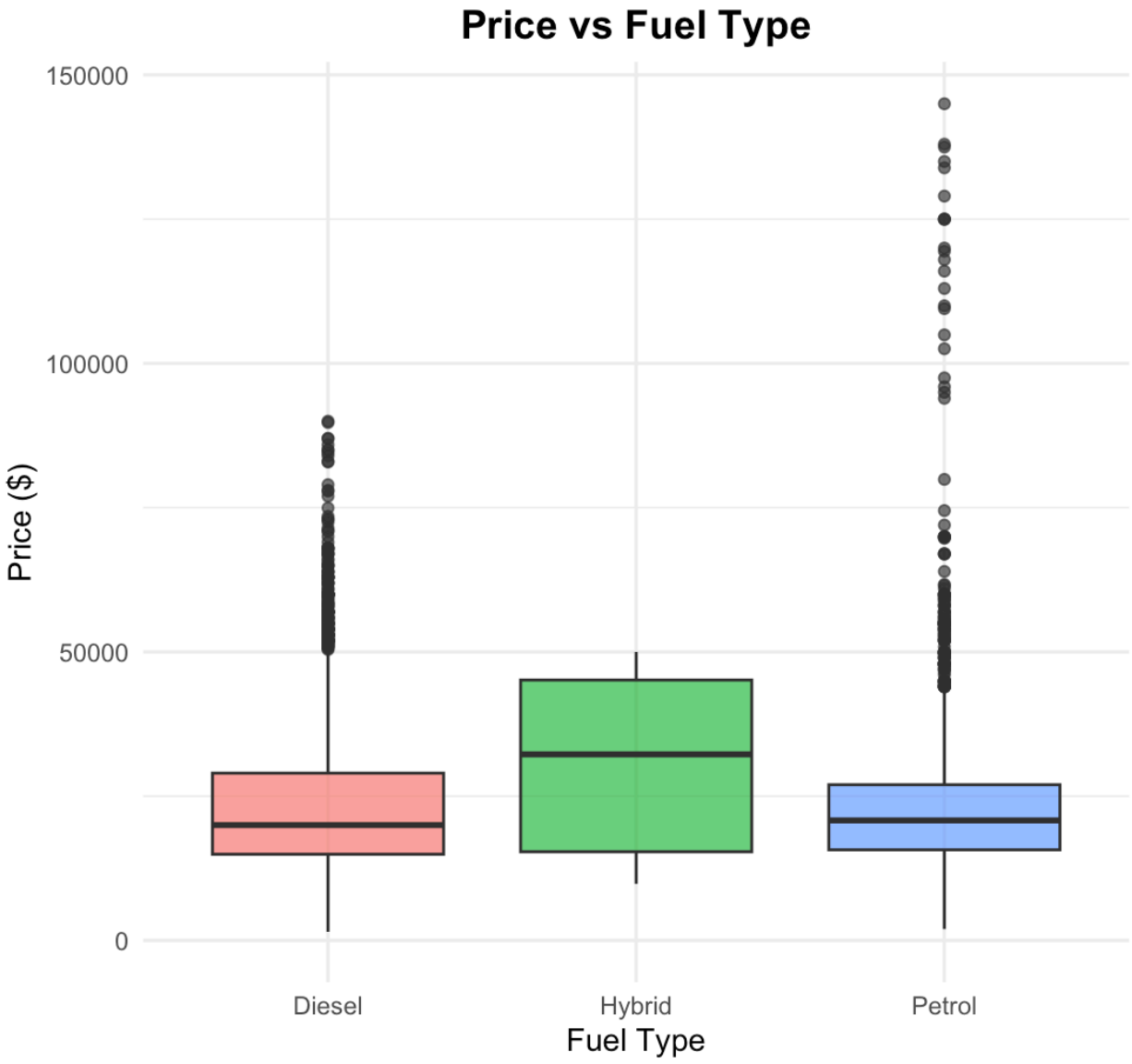

FIGURE 17 *Box plot showing the relationship between price and fuel type.*

Hybrid cars have the highest median price, reflecting their higher cost compared to fuel-based vehicles. Diesel and petrol cars show significant high-priced outliers, especially petrol, while hybrids have fewer outliers due to their limited representation and newer models.

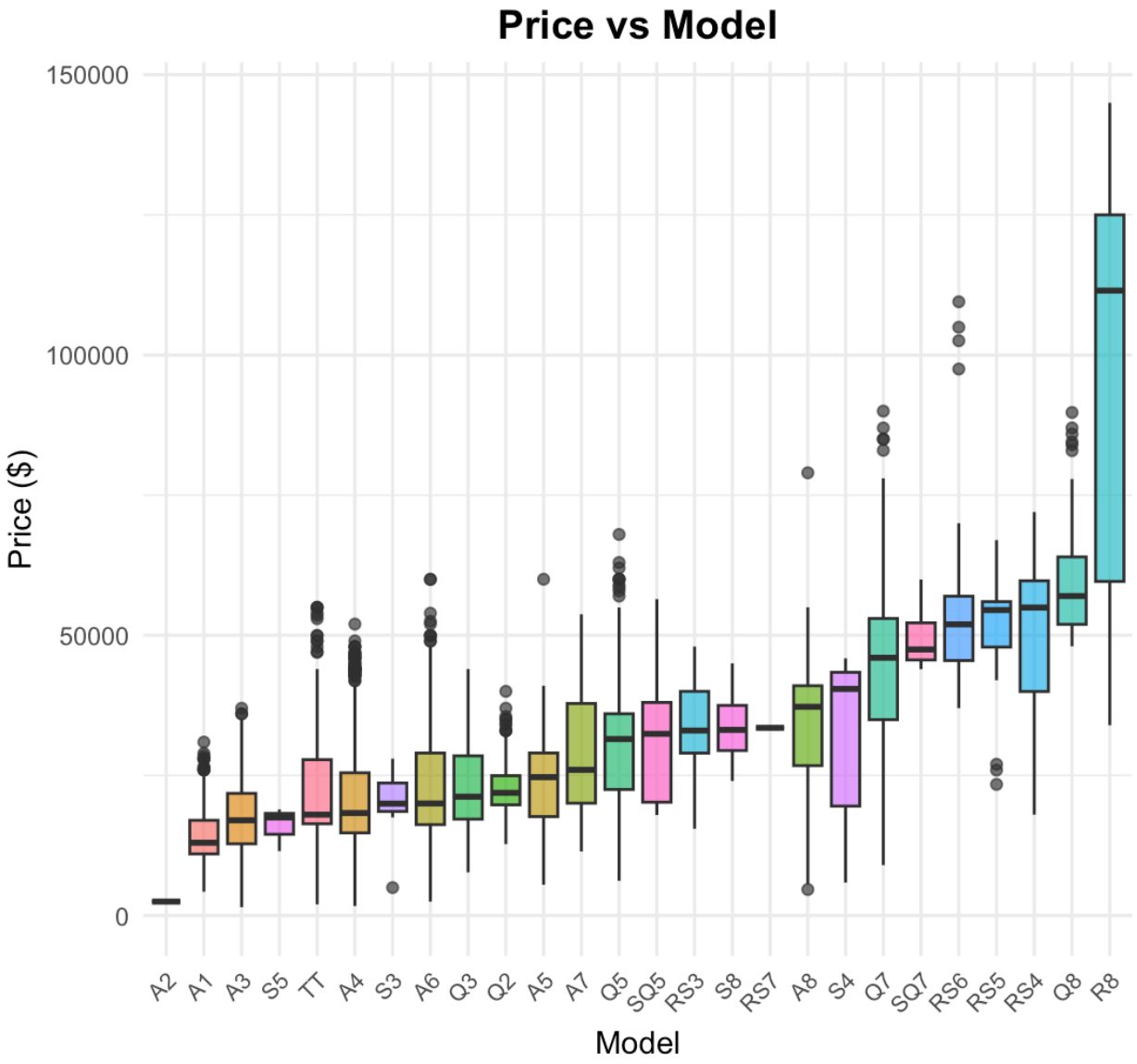

FIGURE 18 *Box plot showing the relationship between price and model.*

The box plot shows how car prices vary across different models, with luxury models like R8 and RS6 having the highest median prices and wide price ranges, reflecting their premium status. Economy models like A1 and A2 have lower prices with less variation, indicating consistent affordability. Outliers are common across most models, especially in higher-end models, likely due to special editions or additional features.

7. Diagnostic Plots of the Final Model

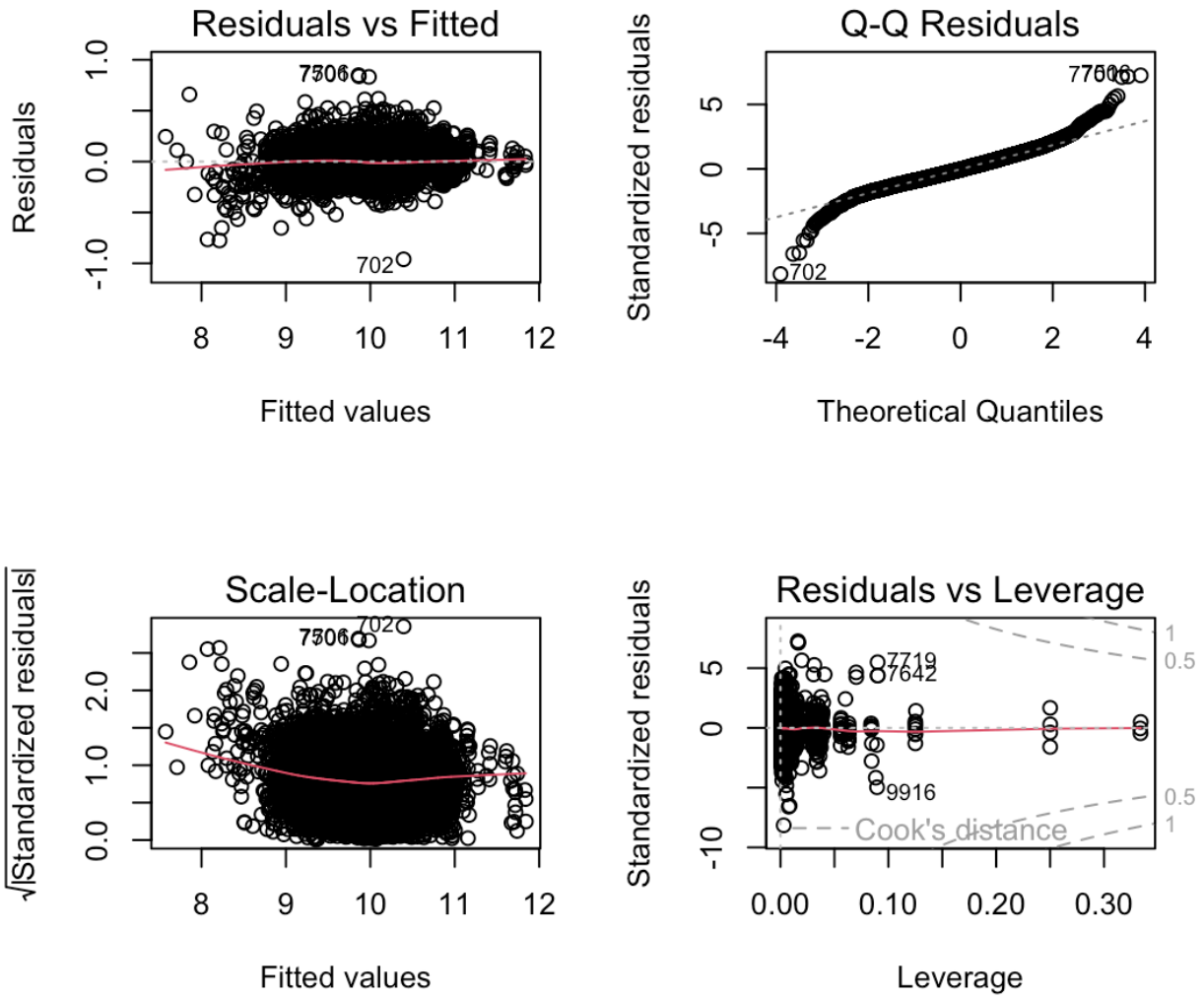

FIGURE 19 *Diagnostic Plots.*

7.1. Residuals vs. Fitted

This chart checks if the model's predictions are accurate across all price ranges. The chart shows that the values are scattered around the line which is good. However, there seems to be clustering across some areas which may represent common price ranges in the dataset. However, the clustering is a minor concern as the residuals should be randomly scattered around the horizontal line at zero.

7.2. Normal Q-Q

This chart evaluates whether the error values of the model follow a normal pattern. All the errors follow a normal pattern except for the few dots at both the ends. This implies that the model struggles with accurately predicting prices of used cars which are either cheap or expensive. It is acceptable because the dataset contains fewer instances of very cheap or very expensive cars, which is a common occurrence in real-life scenarios.

7.3. Scale-Location

It is ideal for the points to be evenly scattered around the horizontal line in a scale-location chart. While the red line remains relatively flat, there is some uneven spread of points, implying minor variations in error variance across different price ranges. As the model doesn't capture every possible predictor while buying used cars, an even spread can't be expected.

7.4. Residuals vs. Leverage

All the values fall under the Cook's distance lines as seen in the Residuals v/s Leverage chart. While there are some data points which are closer to the boundaries, these are most-likely rare listings. The chart also confirms that the model might have some trouble with predicting very cheap or expensive cars but will work well for most car prices.

8. Variable Inflation Factor (VIF) Values

Variable	VIF
Year	1.71
Transmission	1.12
Model	1.04
FuelType	1.36

MPG	1.96
Mileage	1.74
EngineSize	1.98

TABLE 3 *Variable Inflation Factor (VIF) Summary*

The VIF values are computed to detect any potential multi-collinearity issues among the predictor variables. The most common threshold for VIF is set at 5, and there is no concern as all the values appear to be less than 2. While there seems to be a mild correlation with predictors such as engine size, mpg, and mileage, it is not severe. The VIF values in summary suggest that the predictors contribute independently as well as significantly to the model.

9. Summary

A multiple regression model was used to predict the price of used Audi cars based on key attributes like year, mileage, and engine size. The log transformation of the response variable ensured linearity between predictors and the log-transformed price, while diagnostic plots confirmed that residuals were approximately normally distributed and homoscedastic. Additionally, multicollinearity checks showed no significant issues, validating the model's adherence to the fundamental assumptions of linear regression.

10. Appendix

10.1. Work Plan Tasks

Work Plan Tasks (Estimated vs/ Actual Time)				
Phase	**Task**	**Description**	**Estimated Time**	**Actual Time**
Business Understanding	Define Project Objectives	Clearly define the goal and the scope of the data mining project as applied to business decision-making contexts.	1 hour	1 hour
	Business Problem Statement	Finalize the business problem that the predictive model will solve to include in the final report.	1 hour	1 hour
Data Understanding	Data Collection and Overview	Explore dataset repositories and select the "Audi Car" dataset from Kaggle upon initial observation and verification.	2 hours	2 hours
	Data Quality Assessment	Check for potential missing values, duplicates or data inconsistencies.	1 hour	1 hour
Data Preparation	Exploratory Data Analysis	Conduct the exploratory data analysis including summary statistics, visualization for each variable. Compute correlation matrix to address potential issues related to multi-collinearity.	5 hours	6 hours
	Data Cleaning	Handle potential outliers in the dataset using data reduction techniques such as data	4 hours	3 hours

		transformation which will ensure consistency for model input.		
	Test Features	Modify or create new features in the model through non-linear transformation or adding up interaction terms for model compatibility.	3 hours	3 hours
	Convert Categorical Variables	Convert the categorical variables as factors (e.g. transmission, and fuel-type) for model compatibility.	2 hours	1 hour
	Data Splitting	Divide the data into training and test datasets for model evaluation.	1 hour	1 hour
Modelling	Data Tuning and Modeling	Train the model on the training data and run different iterations testing different versions to capture the most effective predictive model.	6 hours	5 hours
	Model Evaluation	Select the best-performing version of the predictive model by evaluating metrics such as R-squared, MAE, MSE, etc.	3 hours	3 hours
Evaluation	Diagnostic Plots	Generate the four diagnostic plots and ensure that the model fits the assumption of the model's residuals such as linearity, homoscedasticity, presence of large residuals, etc.	3 hours	4 hours
	VIF Calculation	Perform the VIF calculation to identify potential collinearity issues.	1 hour	1 hour

	Consultation	Consult with Dr. Tiahrt about the overall predictive model performance and work in feedback.	3 hours	2 hours
Deployment	Model Execution	Run the model and compare the actual values.	1 hour	1 hour
	Report Drafting	Work on the first and final draft of the report by following the project rubric instructions.	6 hours	8 hours
	Report Compilation	Organize the final draft into a structured report for clarity.	1 hour	2 hours
	Proofreading	Proofread the finalized version of the report for potential grammatical errors and ensure a structured layout to the final report.	2 hours	2 hours
	Upload	Upload the final report.	<1 hour	<1 hour

TABLE 5 *Work plan tasks.*

10.2. Predictive Results

The table below presents a comparison of the model-predicted prices and actual prices, based on a dataset split into 70% training and 30% testing sets.

Actual Price	Predicted Price
11000	14724.61
16800	16579.61
13900	15020.13
11750	12216.1
12000	15300.23
19000	23294.24
22500	25112.54
17500	20486.38
15800	15987.25
16000	18187.82

TABLE 6 *Comparison of Actual Prices and Model-Predicted Prices*