Final Report:

Data Cops - Unraveling Australian Vehicle Prices

Submitted to:

Professor Thiru Ramaraj

Report Prepared By:

Naveen Kumar Reddy Veeramreddy

Baaghi Sai Keerthi

Uttam Sai Mohan Kammila

Sai Ram Reddy Kathi

Table of Contents

1. Introduction

This report includes comprehensive analysis of Australian vehicles prices data by analyzing the features of the data and estimating the prices of the vehicles using various analysis and regression techniques.

The Australian automotive market represents a dynamic and diverse landscape shaped by various factors such as consumer preferences, economic dynamics, and the interplay of features influencing vehicle prices. Stakeholders in this complex market, whether consumers, sellers, or analysts, face challenges in understanding and navigating through the intricate web of factors affecting vehicle valuations.

Our project, "Data Cops - Unraveling Australian Vehicle Prices," seeks to unravel these complexities through a rigorous data analysis approach. By employing regression analysis and statistical methods, we aim to uncover the relationships between vehicle prices and key features, including brand reputation, year, model specifications, and overall condition. This endeavor is not just about predicting prices; it's a journey towards unveiling the underlying patterns and insights that drive the Australian vehicle market.

2. Dataset Explanation

The dataset chosen for this study is the Kaggle Australian Vehicle Prices dataset, a collection of over 16,000 records representing car prices in Australia for the year 2023. Sourced from various online platforms, this dataset encapsulates a diverse range of brands, models, and features, providing a comprehensive snapshot of the Australian automotive market.

Key features in the dataset include brand, year, model, car/SUV designation, title (used/new), transmission type, engine specifications, fuel type, fuel consumption, kilometers, exterior/interior color, location, cylinders in the engine, body type, doors, seats, and, most importantly, the price. With such a rich set of attributes, the dataset offers an extensive foundation for analysis, allowing us to delve into market trends, feature analysis, and price prediction with a holistic understanding of the Australian vehicle market dynamics.

In summary, "Data Cops - Unraveling Australian Vehicle Prices" is not merely about predicting prices; it's a deep dive into understanding the nuances of the Australian automotive market, leveraging a dataset that reflects its diversity and complexity. The insights derived from this analysis will not only contribute to accurate price predictions but also offer valuable perspectives on market trends and feature influences.

3. Data Cleaning

Data cleaning is a crucial step in the data analysis process to ensure accuracy, reliability, and consistency. Through systematic data cleaning, these issues are addressed by handling missing values, removing duplicates, standardizing formats, normalizing scales, and addressing outliers. Cleaned data not only facilitates accurate interpretation of results but also enhances the performance of machine learning models, aligning the dataset with the assumptions of statistical methods.

In our data cleaning process, we implemented the following steps:

Dropping Columns

We dropped *Model, Car.Suv, Title, CylindersinEngine, ColourExtInt* columns to focus on relevant features.

Splitting the Data

We had split *Engine* and *FuelConsumption* columns into relevant sub-columns for better analysis. This step aimed to break down complex data into more manageable and insightful components.

Dropping Unwanted Columns

We dropped *KmsforLitersper100km* column for efficiency in subsequent modeling.

Removing Unwanted Characters

Numeric columns like *Cylinders, EngineDisplacement, Litersper100km, Doors,* and *Seats* were extracted from non-numeric characters. This ensured that these columns contained only numeric values, facilitating accurate analysis and modeling.

Handling Null Values

We dropped rows with NA values and rows with '-' values to ensure data quality.

Converting Character to Numerical Columns

A function was applied to convert columns to numeric types where applicable.

Remove Inverted Commas

Inverted commas were removed from the *Price* and *Kilometres* columns, converting them to numeric types.

The cleaned dataset provides a reliable foundation for exploratory data analysis and subsequent modeling. The next steps will involve exploratory data analysis, identifying significant factors influencing vehicle prices, and developing predictive models to estimate car prices accurately. The study will contribute to informed decision-making in the Australian automotive market.

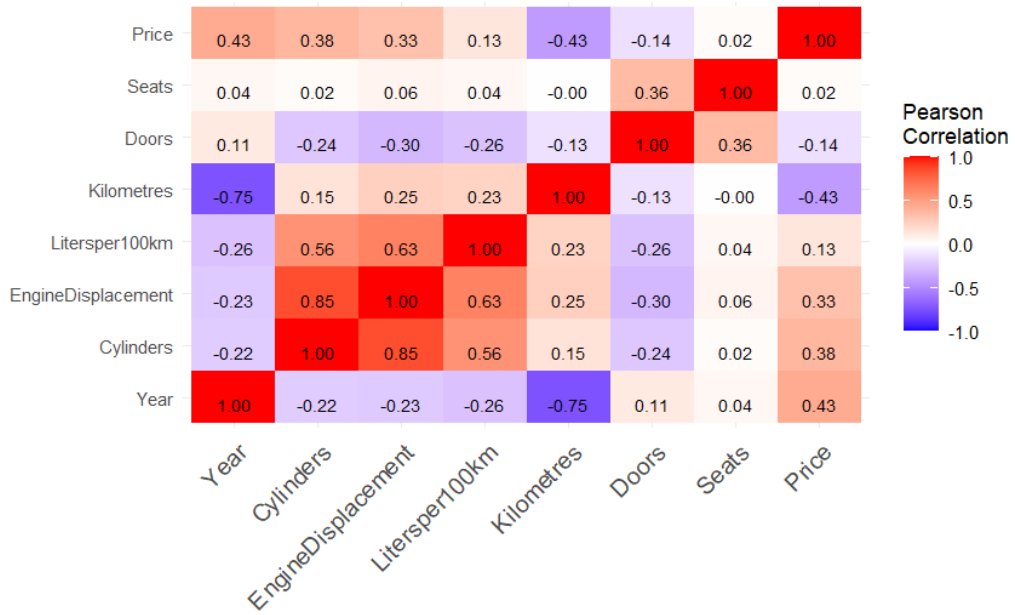
4. Exploratory Data Analysis

In this section, we delve into the analysis of the data to understand the relation between the variables. All the plots are attached in the appendix for reference.

4.1 Box plots

Box plots reveal the price variability across different categories, such as transmission types and the number of doors and seats. These plots are instrumental in discerning distinct pricing tiers, which may significantly influence our predictive models.

4.2 Heat Map



A heatmap of the correlation matrix vividly illustrates inter-variable relationships. It uncovers the strength and direction of correlations, like the pronounced positive correlation between engine displacement and cylinder count, hinting at potential multicollinearity.

4.3 Bar Plots

Bar plots for categorical variables like 'UsedOrNew', 'Transmission', and 'FuelType' elucidate the distribution frequencies within each category. These plots accentuate the dominance of used vehicles and automatic transmissions in the dataset, and such prevalent categories might require careful consideration during dummy variable creation.

5. Linear Regression Modeling

This section outlines the process of building linear regression models to predict vehicle prices based on selected features. The steps include constructing a simple linear regression model, checking for multicollinearity, manual variable elimination, and employing different selection methods to enhance model accuracy.

5.1 Simple Linear Regression Model

A baseline model was established using simple linear regression. The model includes various predictors such as brand, year, transmission type, and others. The summary statistics for this model are presented below:

Model Summary:

Residual Standard Error: 15070 on 14099 degrees of freedom

Multiple R-squared: 0.7163, Adjusted R-squared: 0.7144

F-statistic: 367.1 on 97 and 14099 DF, p-value: < 2.2e-16

Interpretation:

The model explains approximately 71.63% of the variability in vehicle prices.

A significant F-statistic suggests that the overall model is statistically significant.

5.2 Checking Multicollinearity

In assessing multicollinearity within our dataset, the Generalized Variance Inflation Factor (GVIF) was employed due to the presence of categorical variables. The GVIF values were subsequently adjusted for the degrees of freedom associated with each predictor. An adjusted GVIF value threshold of approximately 2.5 was set as the benchmark for concern; values below

this threshold indicate a negligible level of multicollinearity. Analysis revealed that all variables, except for Engine Displacement, exhibited adjusted GVIF values well below our threshold, indicating minimal multicollinearity. Engine Displacement presented a moderate level of multicollinearity but remained within acceptable limits for our predictive modeling purposes. Thus, our data shows that multicollinearity is unlikely to substantially affect the integrity of the regression analysis.

5.3 Manual Variable Elimination

The manual variable elimination process has proven to be effective in refining the linear regression model. By removing variables with low significance, such as those identified through p-values, the model has undergone a streamlined improvement. This contributes not only to enhanced model interpretability but also helps in eliminating noise and focusing on the most influential predictors. In the context of the identified multicollinearity issues, it may be beneficial to consider the removal of ‘Doors' variable, among others, to further address potential collinearity concerns and model complexity issues.

5.4 Forward Selection Method

5.4.1 Steps Taken

The initial model (fitStart) starts with an intercept-only model.

Variables are iteratively added to the model based on their significance (direction="forward").

The chosen model includes Year, Cylinders, FuelType, Kilometres, DriveType, Doors, UsedOrNew, State, EngineDisplacement, and Transmission.

5.4.2 Key Results

The final model (fitForward) includes 25 predictors including the categorical variables which has levels.

The coefficients of the model indicate the estimated effect of each predictor on the vehicle price.

The Adjusted R-squared is 0.5318, indicating that approximately 53.18% of the variability in vehicle prices is explained by the selected predictors.

The F-statistic is 646 with a very low p-value (< 2.2e-16), suggesting that the model is statistically significant.

5.4.3 Coefficients Interpretation

The coefficients for each predictor provide information on how a one-unit change in that predictor relates to the change in vehicle price.

For example, the coefficient for the "Year" variable is 1.672e+03, indicating that, on average, a one-year increase in the vehicle's manufacturing year is associated with an increase of $1,672 in the price.

5.4.4 Significance of Predictors

The significance codes (***, **, *, etc.) help identify predictors that are statistically significant in predicting vehicle prices. For instance, "Year," "Cylinders," "FuelType," "Kilometres," etc., are highly significant.

5.4.5 Residuals

The residuals (model errors) provide insights into how well the model fits the data. The minimum and maximum residuals are shown.

5.5 Backward Elimination Method

5.5.1 Steps Taken

The initial model (fitFull) starts with all predictors included.

Variables are iteratively removed from the model based on their significance (direction="backward").

The chosen model includes Brand, Year, UsedOrNew, Transmission, Cylinders, EngineDisplacement, DriveType, FuelType, Litersper100km, Kilometres, State, BodyType, and Seats.

5.5.2 Key Results

The final model (fitBackward) includes 81 predictors.

The coefficients of the model indicate the estimated effect of each predictor on the vehicle price.

The Adjusted R-squared is 0.7144, suggesting that approximately 71.44% of the variability in vehicle prices is explained by the selected predictors.

The F-statistic is 370.9 with a very low p-value (< 2.2e-16), indicating that the model is statistically significant.

5.5.3 Coefficients Interpretation

The coefficients for each predictor provide information on how a one-unit change in that predictor relates to the change in vehicle price.

For example, the coefficient for the "Year" variable is 2.038e+03, indicating that, on average, a one-year increase in the vehicle's manufacturing year is associated with an increase of $2,038 in the price.

5.5.4 Significance of Predictors

The significance codes (***, **, *, etc.) help identify predictors that are statistically significant in predicting vehicle prices. For instance, "Year," "Cylinders," "FuelType," "Kilometres," etc., are highly significant.

5.5.5 Residuals

The residuals (model errors) provide insights into how well the model fits the data. The minimum and maximum residuals are shown.

6. Interaction Terms and Quadratic Terms Modeling

This section explores modeling with interaction terms and quadratic terms to capture more complex relationships in the context of vehicle pricing.

6.1 Basic Quadratic Terms

In Model 1, basic quadratic terms are introduced to capture potential non-linear relationships between the response variable (Price) and some of the predictor variables. The model includes quadratic terms for the 'Year', and 'Kilometres.' Additionally, other predictor variables such as 'Brand,' 'DriveType,' 'Cylinders,' 'FuelType,' 'BodyType,' 'UsedOrNew,' 'Transmission,' 'Litersper100km,' 'Seats,' and 'State' are included in the linear regression model.

The summary statistics for Model 1 indicate a high coefficient of determination (Multiple R-squared: 0.7362), suggesting that the model explains a substantial portion of the variability in the response variable. The inclusion of quadratic terms allows the model to capture non-linear patterns in the relationships.

6.2 Selected Interactions

In Model 2, specific interaction terms are incorporated to capture nuanced relationships between variables. The selected interactions include 'EngineDisplacement:FuelType' and 'Kilometres:Year.' These interactions aim to account for potential dependencies between the engine displacement and fuel type, as well as the interaction between the age of the vehicle and its mileage.

The summary statistics for Model 2 reveal insights into the impact of these selected interactions on the model. The coefficients associated with these interactions help understand how the relationship between the predictors and the response variable changes when considering their joint effects with a Adjusted R-squared of 0.7438.

6.3 Quadratic Terms with Key Interactions

Model 3 combines quadratic terms with key interactions to model complex relationships while accounting for non-linear patterns and nuanced interactions. The quadratic terms for 'Year,' 'EngineDisplacement,' and 'Kilometres' are retained, and interactions like 'EngineDisplacement:Kilometres' and 'Year:Kilometres' are introduced.

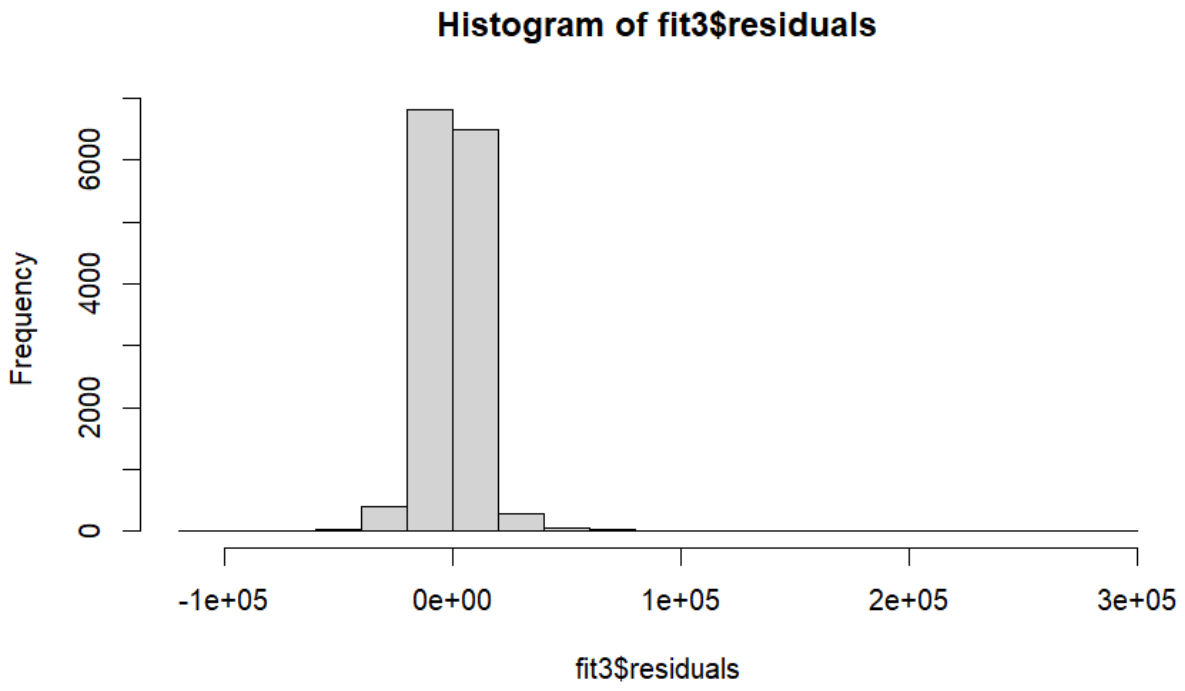
The summary statistics for Model 3 provide a comprehensive overview of the combined impact of quadratic terms and key interactions on the model's performance with a Adjusted R squared of 0.7552. This approach allows for a more flexible representation of the relationships within the data.

7. Residual Analysis

This section delves into a detailed examination of the residuals from the regression model to provide insights into model performance and potential areas for refinement.

7.1 Residuals and Histogram

Certainly, analyzing the distribution of residuals is an important step in assessing the performance of a regression model. The histogram you've provided is a visualization of the distribution of the residuals in the fit3 model. Here's a breakdown of what the histogram shows:



- The x-axis represents the value of the residuals.
- The y-axis represents the frequency of each residual value. For example, the tallest bar around 0 on the x-axis indicates that there are more residuals around 0 than any other value.

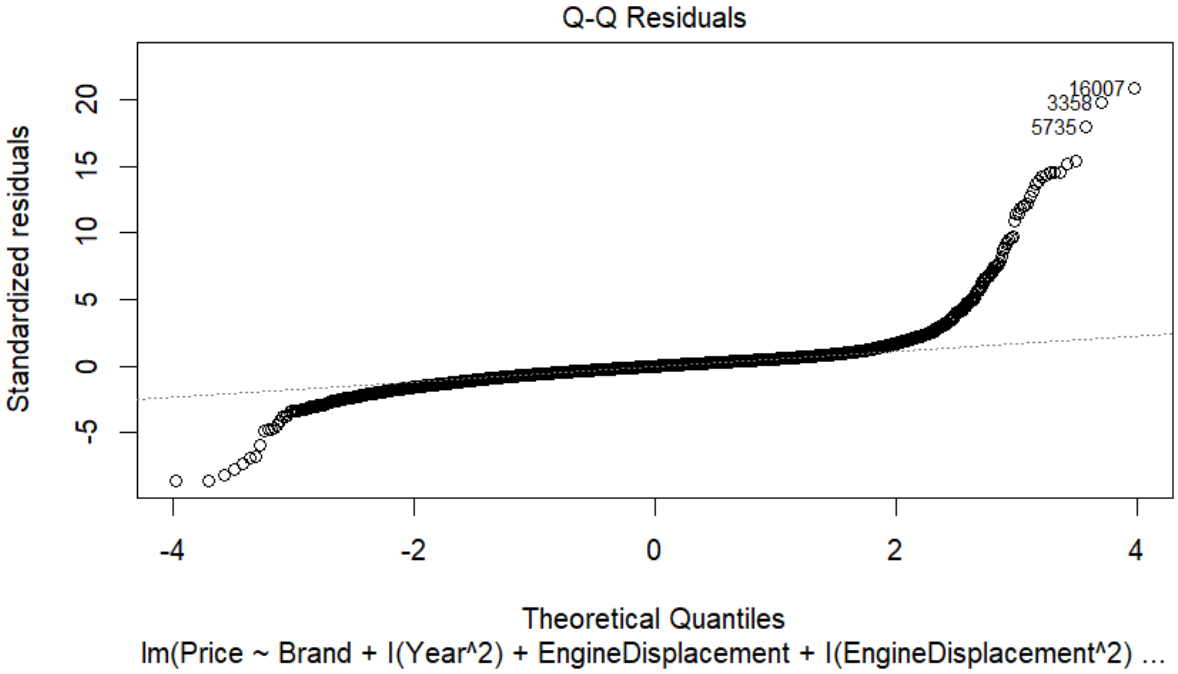
By looking at the distribution of the residuals, we can gain insights into the model's accuracy and potential for bias. Ideally, the residuals should be normally distributed around zero. This would indicate that the model is fitting the data well and there are no systematic errors.

The histogram appears centered around zero is a good sign. There also don't appear to be any outliers on the far left or right side of the histogram.

7.2 Model Plots

The x-axis represents the theoretical quantiles, and the y-axis represents the standardized residuals.

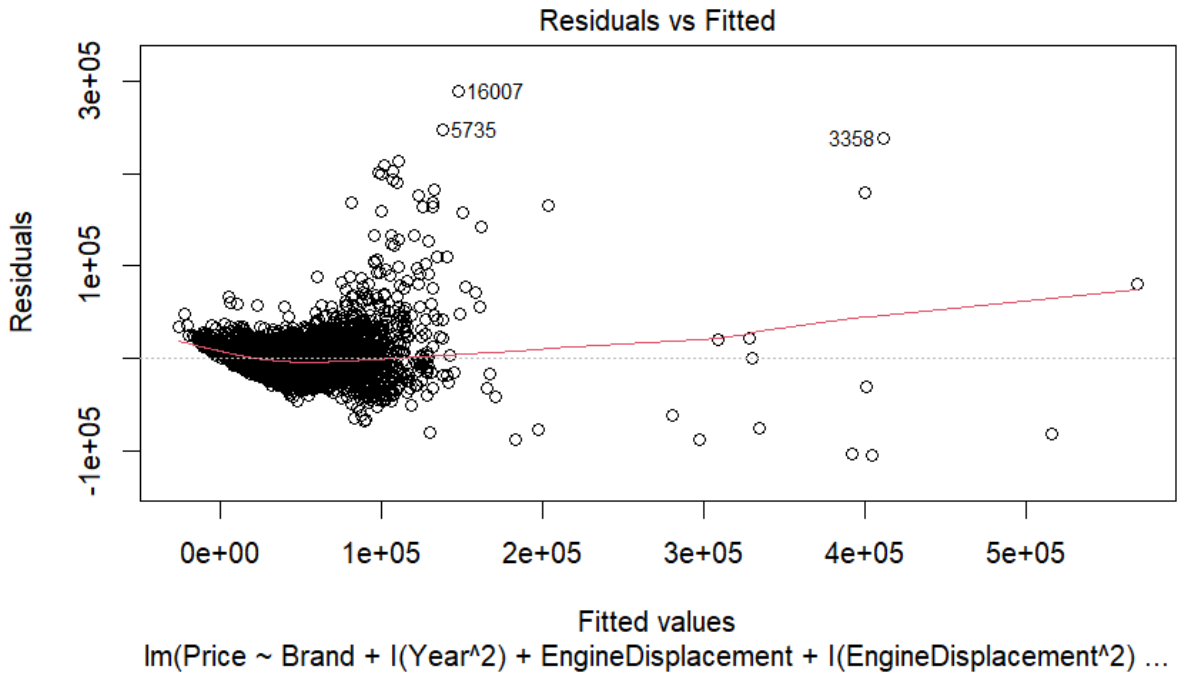
Here's a breakdown of what the plot shows:



- The points in the plot represent the residuals. Ideally, the points should fall roughly along a straight diagonal line. This would indicate that the model is fitting the data well and there are no systematic errors.
- In this specific case, the points do not fall exactly on a straight line, which suggests that there may be some non-normality in the residuals. However, the deviation from the straight line is relatively minor, so it is likely not a major concern.

Normal QQ plots, on the other hand, are used to compare the distribution of the residuals to a normal distribution. They are created by plotting the quantiles of the residuals on the y-axis versus the quantiles of a normal distribution on the x-axis. If the residuals are normally distributed, the points will fall roughly along a straight line.

The residual plot suggests that there may be some non-normality which is handled by logarithmic transformation of the Price column.



This plot is used to assess the assumption of homoscedasticity, which is the idea that the variance of the errors is constant across all levels of the independent variable.

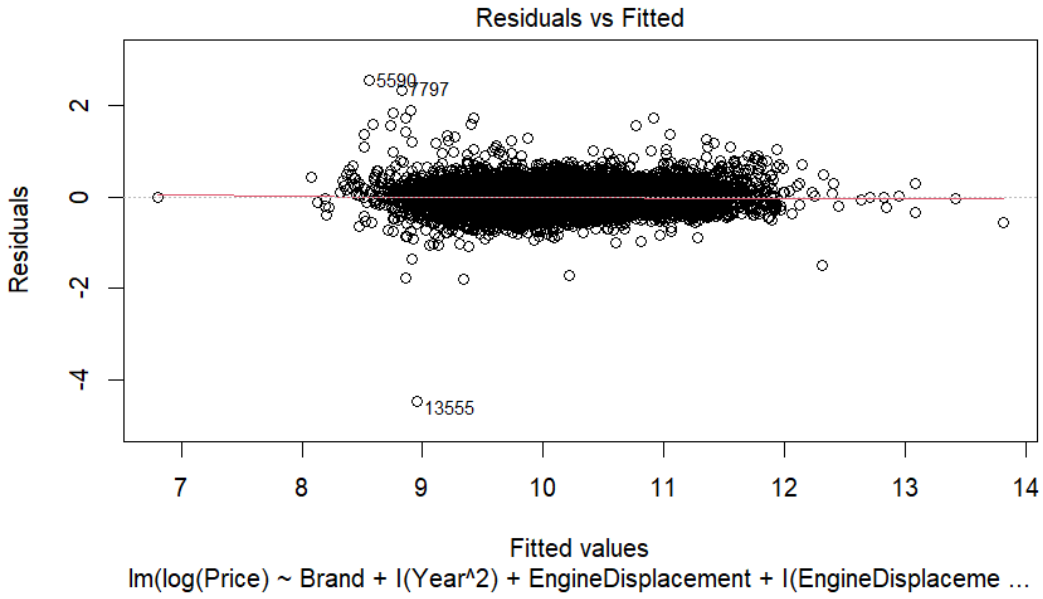
Let's break down what the plot shows:

- The x-axis represents the fitted values, which are the predicted values from the regression model.
- The y-axis represents the standardized residuals, which are the residuals (the difference between the actual and fitted values) divided by the standard deviation of the residuals.

Ideally, in a scenario with homoscedasticity, the residuals would be randomly scattered around zero on the y-axis, regardless of the fitted value on the x-axis. This would indicate that the variance of the errors is constant across all levels of the independent variable.

The standardized residuals appear to be more spread out for higher fitted values. This suggests that the variance of the errors is **not** constant across all levels of the independent variable, and there may be a violation of the homoscedasticity assumption.

Finally, based on residual analysis applying the log transformation on the Price column the plot turned to be good by adhering to the assumptions.



8. Dummy Variables and Model Evaluation

In this section, we delve into the application of dummy variables and the assessment of regression model performance.

8.1 Creating Dummy Variables

The creation of dummy variables is detailed, enabling the incorporation of categorical data into the regression models.

8.2 Model Evaluation Metrics

To comprehensively assess model accuracy and fit, various metrics are calculated.

Root Mean Squared Error (RMSE): The RMSE value of 16098.20 suggests that, on average, the model's predictions deviate by approximately $16,098.20 from the actual prices. Lower RMSE values indicate better model performance.

Mean Squared Error (MSE): The MSE of 259,151,995.25 represents the average squared difference between predicted and actual prices. As MSE is sensitive to outliers, its magnitude should be interpreted in the context of the data.

Mean Absolute Error (MAE): The MAE of 8,398.04 indicates the average absolute difference between predicted and actual prices. Lower MAE values signify better model accuracy.

R-squared on Test Set: The value of 0.6416 indicates that the model explains approximately 64.16% of the variance in the target variable on the test set.

Adjusted R-squared on Test Set: The adjusted R-squared of 0.6289 considers the number of predictors in the model, providing a slightly more conservative estimate of model fit.

9. Lasso and Ridge Regression

In this section, we introduce Lasso and Ridge regression as regularization techniques designed to enhance model generalization and address multicollinearity issues.

9.1 Model Fitting

Lasso and Ridge regression models are fitted to the data, and the regularization parameter is selected.

9.2 Predictions and Metrics

Predictions from the Lasso and Ridge regression models are evaluated using the previously mentioned metrics, providing insights into their performance compared to linear regression models.

Performance metrics are as follows:

Model	RMSE	MAE	R-Squared	Adjusted R-Squared
Lasso	16,856.38	11,222.77	0.5185	0.2844
Ridge	16,806.29	11,108.41	0.5214	-0.0140
Elastic Net	16,854.71	11,220.12	0.5186	0.2731

10. Conclusion

Our comprehensive study, "Data Cops - Unraveling Australian Vehicle Prices," has provided valuable insights into the Australian automotive market through meticulous data cleaning, exploratory analysis, and advanced regression techniques. Our analysis journey highlighted the significance of vehicle attributes on price variability and the necessity for sophisticated modeling to capture the intricate dynamics of vehicle valuations.

The employment of linear regression modeling, enhanced with polynomial and interaction terms, has allowed us to elucidate complex relationships within the dataset, evidenced by substantial R-squared values. Through the adept handling of multicollinearity and the strategic use of dummy variables, we have fostered model stability and interpretability. Our foray into regularization techniques such as Lasso and Ridge regression has further refined our predictive capabilities, ensuring robustness against overfitting and enhancing generalizability.

Learnings:

- The value of rigorous data pre-processing cannot be overstated, as it directly correlates to the quality of subsequent analysis.
- Visual data exploration is instrumental in understanding underlying patterns and guiding feature engineering efforts.
- The interplay of model complexity and interpretability is delicate, necessitating careful selection and transformation of variables.

Future Work:

- Enhance Complexity: Explore a variety of interaction and quadratic terms to refine the model's accuracy.
- Global Expansion: Incorporate data from multiple countries to develop a globally applicable vehicle pricing model.
- Data Augmentation: Increase the dataset size and integrate more diverse features for improved generalizability.
- Dig deeper and conduct a comprehensive analysis of outliers detected in the process.

11. References

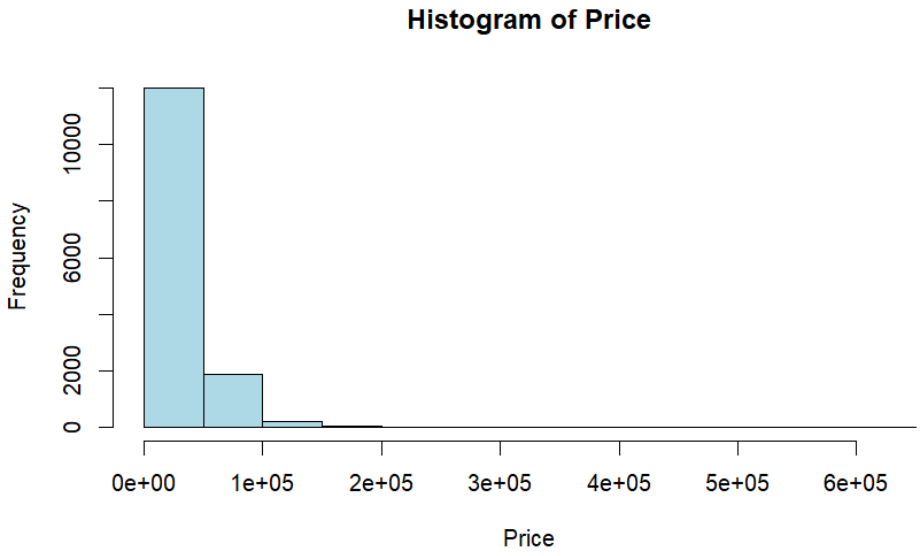
ELGIRIYEWITHANA, N. (2023) Australian Vehicle Prices https://www.kaggle.com/datasets/nelgiriyewithana/australian-vehicle-prices/data

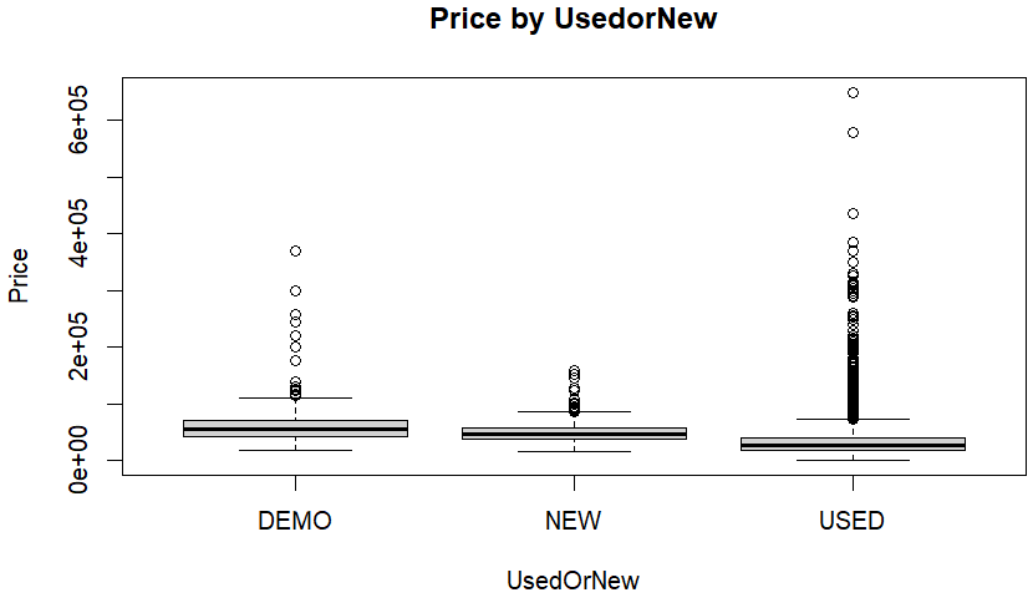
GeeksforGeeks. (2023) **Data Preprocessing in R https://www.geeksforgeeks.org/data-preprocessing-in-r/**

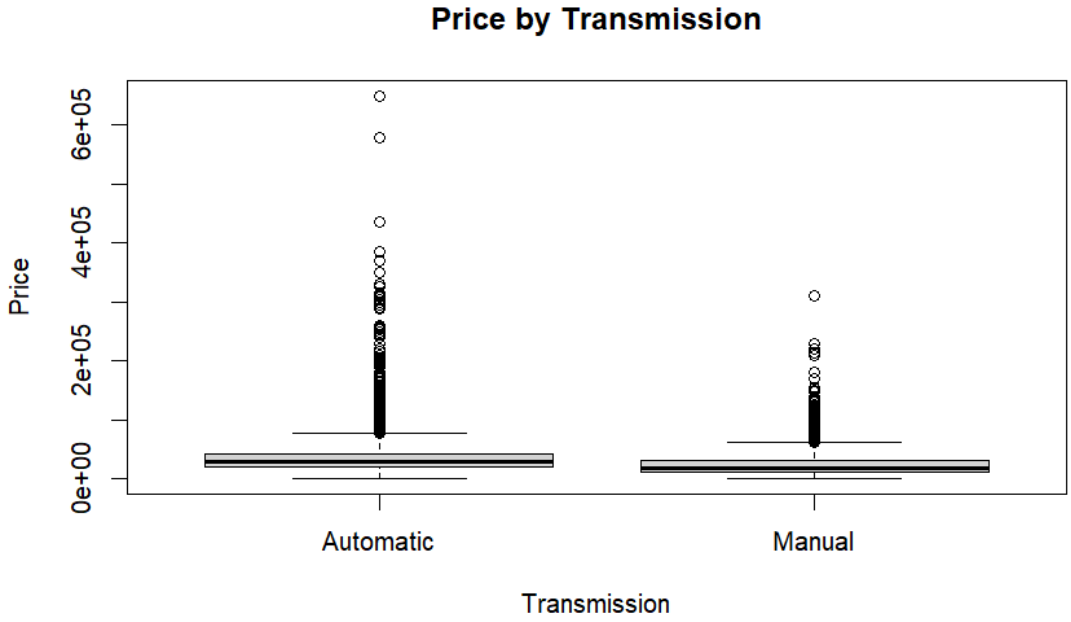
John Fox & Georges Monette (1992) Generalized Collinearity Diagnostics, Journal of the American Statistical Association, 87:417, 178-183, DOI: 10.1080/01621459.1992.10475190

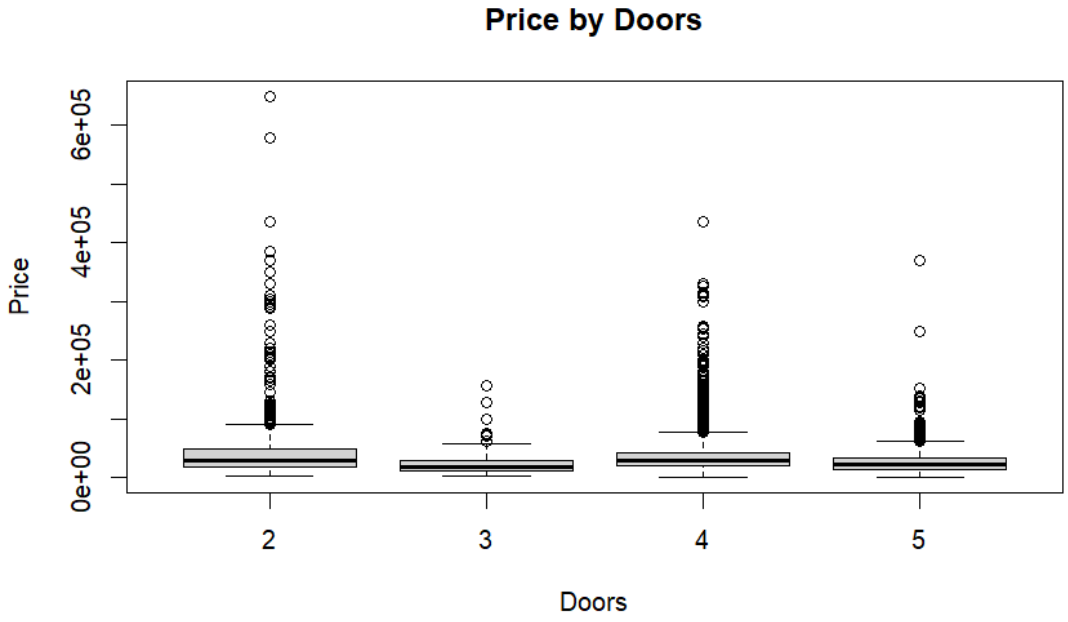
Long, J. (2021) **Exploring interactions with continuous predictors in regression models** https://cran.r-project.org/web/packages/interactions/vignettes/interactions.html

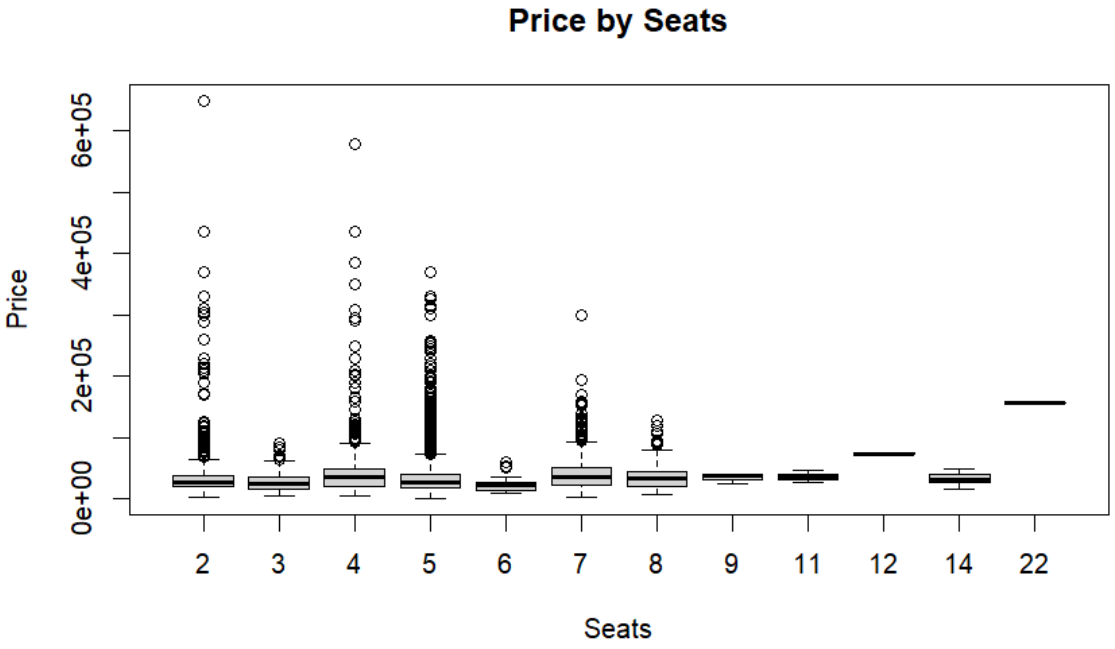
12. Appendix

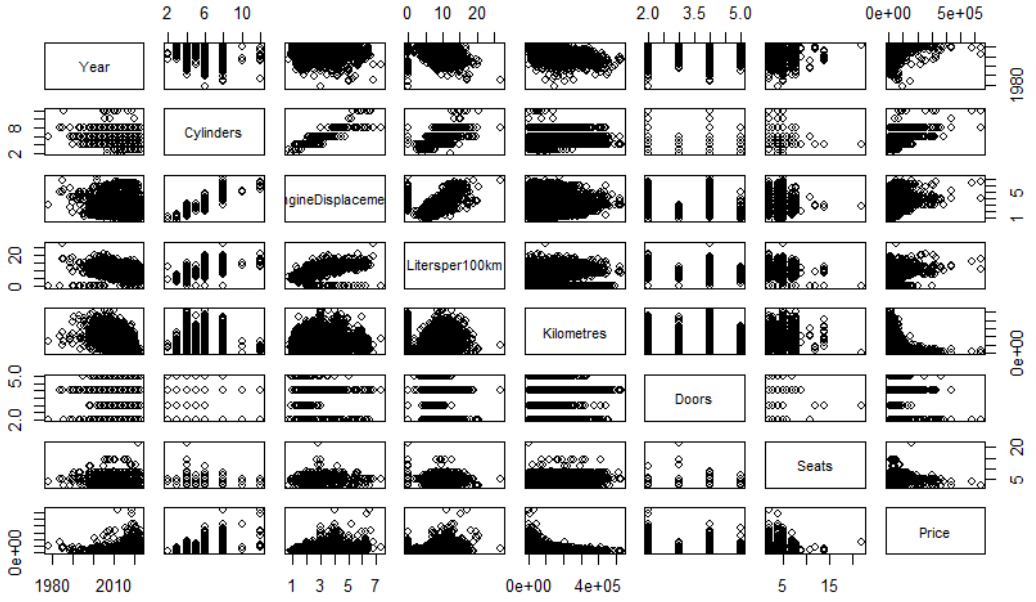

Histogram of Price
Frequency
0
2000
6000
10000
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
Price

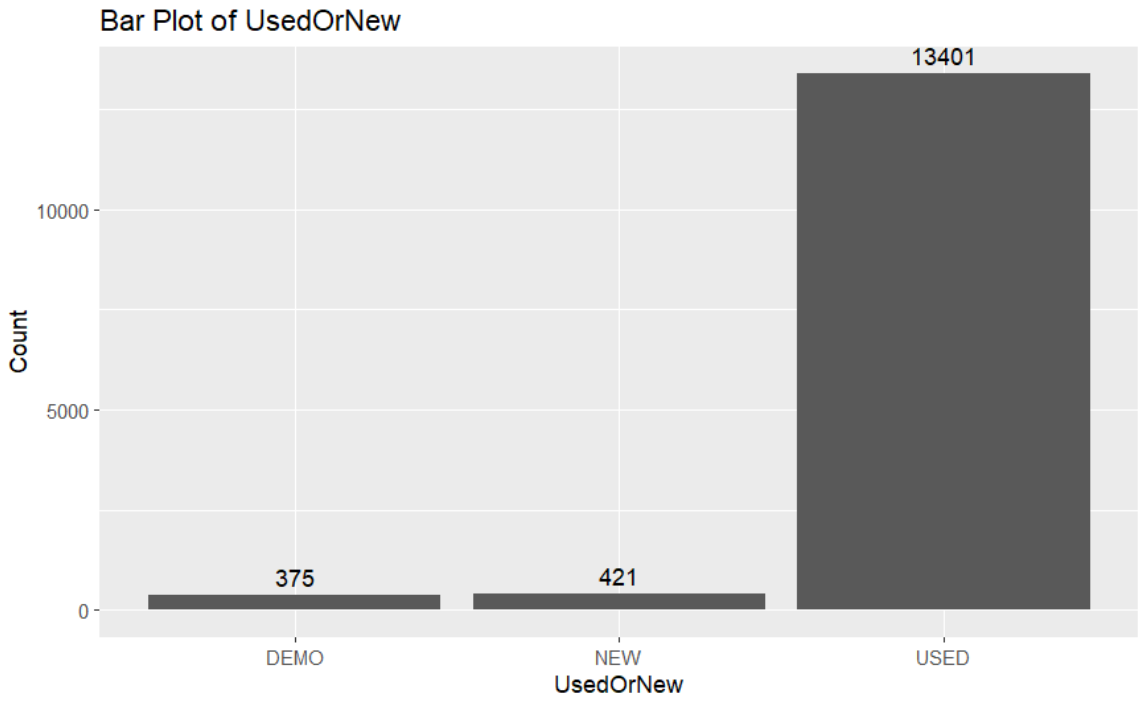

Price by UsedorNew
Price
0e+00
2e+05
4e+05
6e+05
DEMO
NEW
USED
UsedOrNew

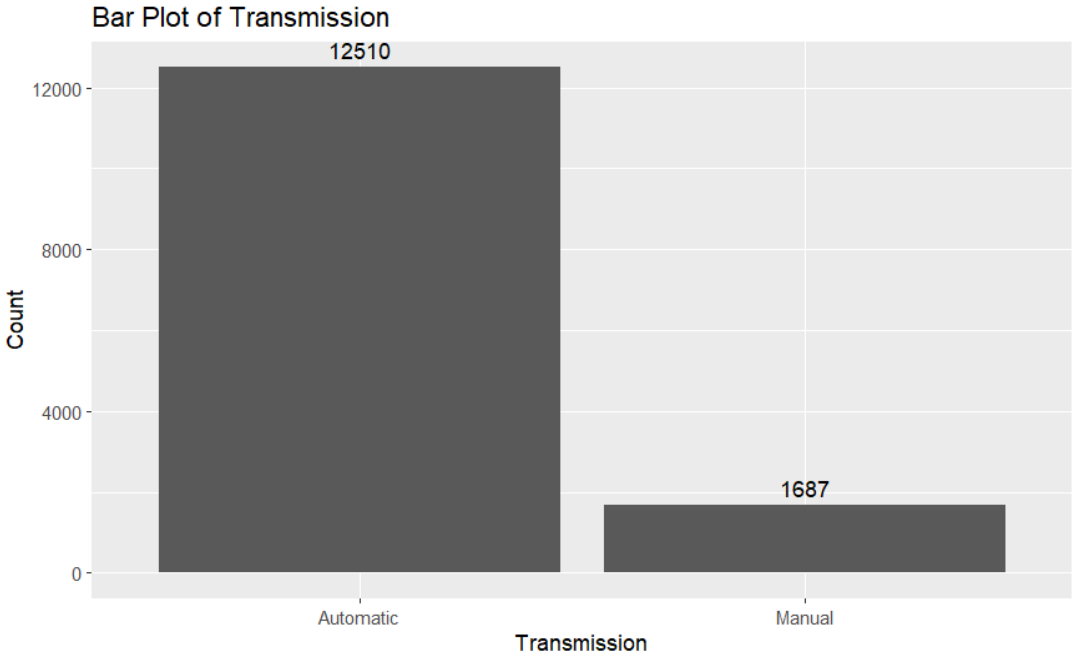

Price by Transmission
Price
0e+00
2e+05
4e+05
6e+05
Automatic
Manual
Transmission

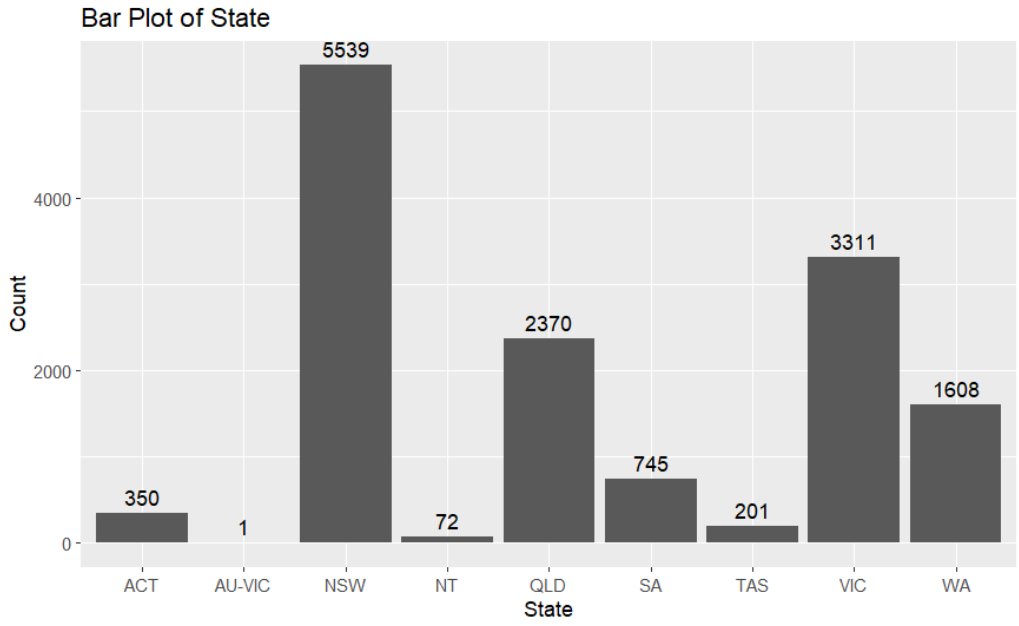

Price by Doors
Price
0e+00
2e+05
4e+05
6e+05
2
3
4
5
Doors

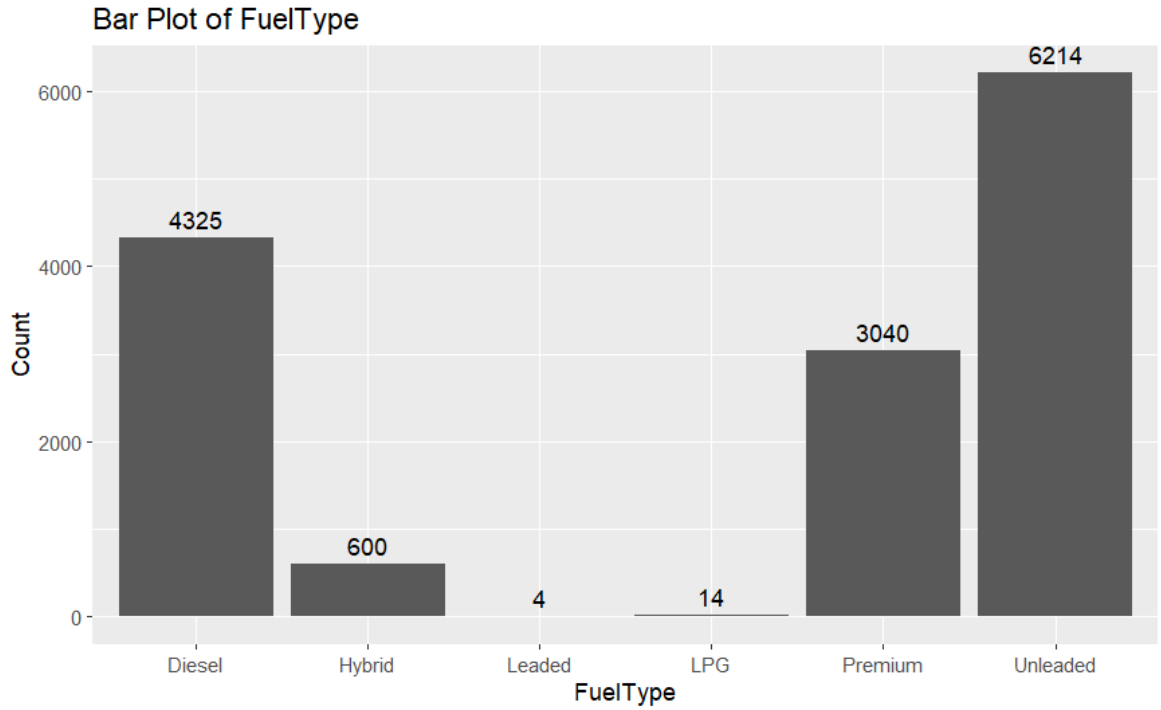

Price by Seats
Price
0e+00
2e+05
4e+05
6e+05
2
3
4
5
6
7
8
9
11
12
14
22
Seats

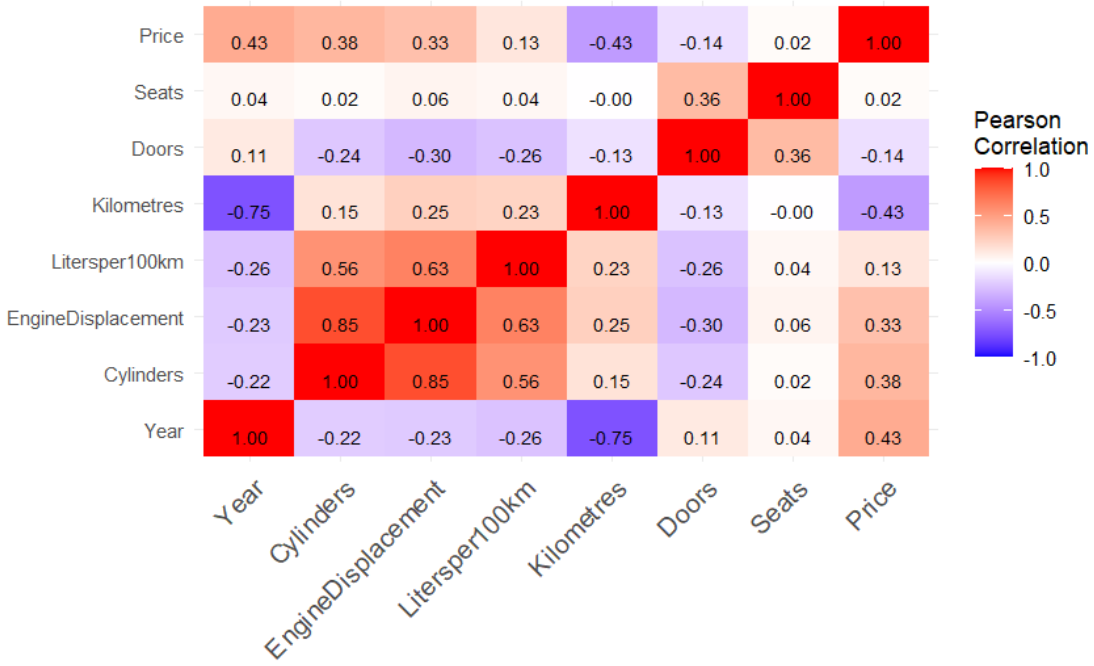

Year
Cylinders
ngineDisplaceme
Litersper100km
Kilometres
Doors
Seats
Price
1980
2010
1
3
5
7
0e+00
4e+05
5
15


Bar Plot of UsedOrNew
13401
10000
Count
5000
375
421
0
DEMO
NEW
USED
UsedOrNew


Bar Plot of Transmission
12510
12000
8000
Count
4000
1687
0
Automatic
Manual
Transmission


Bar Plot of State
5539
4000
3311
Count
2370
2000
1608
745
350
201
1
72
0
ACT
AU-VIC
NSW
NT
QLD
SA
TAS
VIC
WA
State


Bar Plot of FuelType
6214
6000
4325
4000
3040
Count
2000
600
4
14
0
Diesel
Hybrid
Leaded
LPG
Premium
Unleaded
FuelType


Price
Seats
Doors
Kilometres
Litersper100km
EngineDisplacement
Cylinders
Year
0.43 0.38 0.33 0.13 -0.43 -0.14 0.02 1.00
0.04 0.02 0.06 0.04 -0.00 0.36 1.00 0.02
0.11 -0.24 -0.30 -0.26 -0.13 1.00 0.36 -0.14
-0.75 0.15 0.25 0.23 1.00 -0.13 -0.00 -0.43
-0.26 0.56 0.63 1.00 0.23 -0.26 0.04 0.13
-0.23 0.85 1.00 0.63 0.25 -0.30 0.06 0.33
-0.22 1.00 0.85 0.56 0.15 -0.24 0.02 0.38
1.00 -0.22 -0.23 -0.26 -0.75 0.11 0.04 0.43
Year
Cylinders
EngineDisplacement
Litersper100km
Kilometres
Doors
Seats
Price
Pearson Correlation
1.0
0.5
0.0
-0.5
-1.0