Department of Computer Science & Engineering – Fall 2024

STA 301 – Foundations of Statistics for Data Science – Dr. Hana Sulieman

Predicting Price and Year: Insights from the Australian Vehicle Dataset

Sunday, 8th December 2024

Abstract

The purpose of this project was to apply the statistical and analytical techniques learned throughout the semester to build predictive models for vehicle pricing and manufacturing year using a dataset of Australian vehicles. Exploratory data analysis revealed significant relationships between price and key variables such as mileage, engine capacity, and transmission type. Inferential techniques, including hypothesis testing, ANOVA, and correlation analysis, were employed to validate these relationships and guide model development. A multiple regression model was built to predict vehicle prices, achieving an R^2 of 0.7437 after addressing non-normality through Box-Cox transformations and outlier removal. Additionally, classification models, including multinomial logistic regression and random forests, were implemented to predict vehicle year ranges, with random forests achieving the highest accuracy of 80%. However, further analysis was done on a simpler logistic regression model that achieved 68% accuracy, which allowed for an in-depth look at the model equation and coefficients. The project demonstrates the integration of various statistical methods to uncover insights and build models capable of making accurate predictions, while highlighting the challenges of data preprocessing, feature selection, and balancing model interpretability with performance.

Section 1: Data Set Information

The Australian Vehicle Prices dataset contains 16,734 car listings from Australia in 2023, capturing essential car attributes and price information. The dataset contains 19 columns that allow for analyzing the factors influencing vehicle prices in the Australian market.

Features:

1. Brand: Categorical variable indicating the car manufacturer (e.g., Toyota, BMW).
   - Scale: Nominal
2. Year: Numeric variable indicating the manufacturing year of the vehicle.
   - Scale: Interval
3. Model: Categorical variable for the specific vehicle model.
   - Scale: Nominal
4. Car/SUV: Categorical variable indicating the type of vehicle (e.g., SUV, Hatchback).
   - Scale: Nominal
5. Title: Title or description of the car.
   - Scale: Nominal
6. UsedOrNew: Categorical variable specifying whether the vehicle is used, new, or demo.
   - Scale: Nominal
7. Transmission: Categorical variable specifying the type of transmission (e.g., Automatic, Manual).
   - Scale: Nominal
8. Engine: Categorical variable detailing engine configuration (e.g., "4 cyl, 2.2 L").
   - Scale: Nominal
9. DriveType: Categorical variable for the drivetrain type (e.g., AWD, FWD, RWD).
   - Scale: Nominal
10. FuelType: Categorical variable for the type of fuel used (e.g., Diesel, Unleaded, Premium).
    - Scale: Nominal
11. FuelConsumption: Categorical variable specifying fuel consumption (e.g., "6.7 L / 100 km").
    - Scale: Nominal
12. Kilometers: Categorical variable representing the vehicle mileage.
    - Scale: Ordinal (but stored as a string).
13. ColourExtInt: Categorical variable for exterior and interior colors (e.g., "White / Black").
    - Scale: Nominal
14. Location: Categorical variable specifying the location of the vehicle listing.
    - Scale: Nominal
15. CylindersinEngine: Categorical variable for the number of cylinders in the engine (e.g., "4 cyl").
    - Scale: Nominal
16. BodyType: Categorical variable describing the body type of the vehicle (e.g., SUV, Coupe).
    - Scale: Nominal
17. Doors: Categorical variable representing the number of doors.
    - Scale: Ordinal (but stored as a string).
18. Seats: Categorical variable representing the number of seats.

    - Scale: Ordinal (but stored as a string).
19. Price: Numeric variable indicating the price of the car in Australian dollars.
    - Scale: Ratio

Target Variables:

- Price: Target variable for regression model
- Year: Target variable for classification model

Section 2

Before implementing models for classification and regression, we performed several preprocessing steps to clean and prepare the dataset. The steps are detailed below:

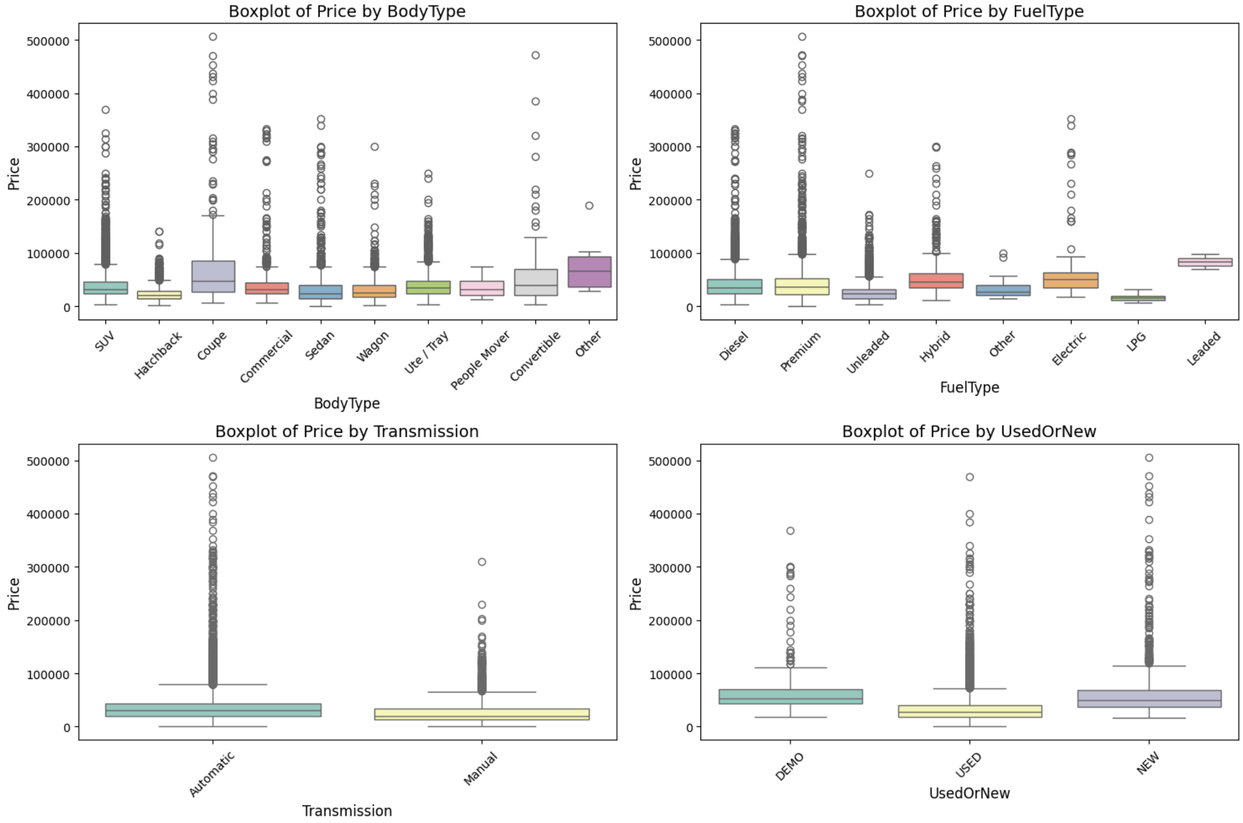
1. **General Cleaning:**
    - Dropped rows with all missing values (NaN).
    - Converted Year from float to int for consistency.
    - Replaced invalid characters (- and /) with NaN to standardize the dataset.
2. **Feature-Specific Preprocessing:**
    - **Car/SUV:** Dropped this column as it contained inaccurate data and duplicated information already provided by CarType.
    - **CylindersInEngine:** Retained only the numeric value (e.g., removed the "cyl" string) and converted it to an ordinal and categorical variable.
    - **EngineCapacity:** Removed the cylinder count, keeping only the engine size in liters as a float.
    - **FuelConsumption:** Extracted the numeric value, renamed the column to FuelConsumptionPer100km to reflect that all values are in per-100km units.
    - **Kilometres:** Converted this column to numeric format.
    - **ColourExtInt:** Split into two columns: ColourExt (external color) and ColourInter (interior color) and dropped ColurInter since it had more than 7000 missing values with multiple incorrect colours (e.g. 2Lle21).
    - **ColurExt:** Changed - to NaN and then replaced invalid colors (like 5 years, 3 years) with the 'Other" value.
    - **Doors and Seats:** Removed text like "door" and "seat," converted the values to numeric, and treated them as ordinal and categorical variables.
    - **Price:** Converted "POA" ("Price on Application") to NaN and then to numeric.
    - **Title:** Dropped this column as it overlapped with information in Model, Brand, or Year, and its high cardinality (over 4,500 unique values) was likely to negatively impact the model.
    - **Year:** Binned into four ranges: "2020-2023," "2017-2019," "2012-2016," and "Before 2012." The new column, YearRanges, was created as an ordinal categorical variable to balance the data while maintaining meaningful groupings.
3. **Imputation of Missing Values:**
    - **Dropped rows** where City, Model, or Brand had fewer than 5 occurrences to remove sparse or unreliable data.
    - **Categorical Variables:** Missing values in Transmission, FuelType, BodyType, CylindersInEngine, Doors, and Seats were imputed by:
        - Grouping by Brand and Model and assigning the most frequent value.

        - If both Brand and Model were NaN, the most frequent value from the entire column was used.
    - **Numerical Variables:** Missing values in EngineCapacity, FuelConsumptionPer100km, and Price were imputed by:
        - Grouping by Brand and Model and calculating the mean.
        - If both Brand and Model were NaN, the mean of the entire column was used.
    - **Kilometres:** Imputed by grouping by UsedOrNew and YearRanges since mileage is closely tied to these factors.
    - **ColourExt, City, and State:** Missing values were imputed with the category "Other."
4. **Final Dataset:** After preprocessing, the cleaned and imputed dataset contained **15,779 rows and 18 columns**. Below is a summary of the dataset's features:

Feature	Type	Possible Values
Brand	Categorical, nominal	49
Model	Categorical, nominal	358
UsedOrNew	Categorical, nominal	3
Transmission	Categorical, binary	2
DriveType	Categorical, nominal	5
FuelType	Categorical, nominal	8
Kilometers	Numerical, continuous	-
CylindersInEngine	Categorical, ordinal	8
BodyType	Categorical, nominal	10
Doors	Categorical, ordinal	4
Seats	Categorical, ordinal	12
Price	Numerical, continuous	-
EngineCapacity	Numerical, continuous	-
FuelConsumptionPer100km	Numerical, continuous	-
ColourExt	Categorical, nominal	18
City	Categorical, nominal	511
State	Categorical, nominal	9
YearRanges	Categorical, ordinal	4

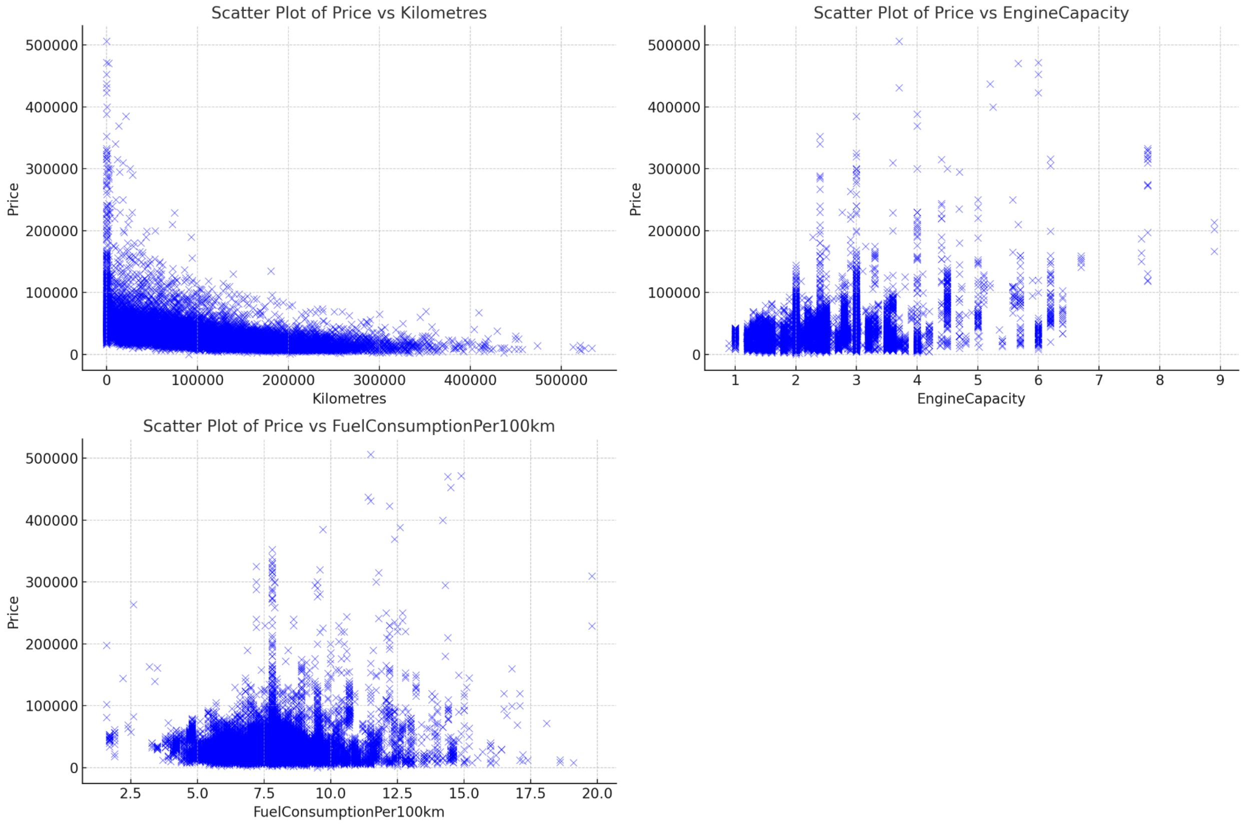
Subsection 2.1: *Exploratory data analysis: Graphs and Summary Statistics*

Boxplots for important categorical variables against Price:



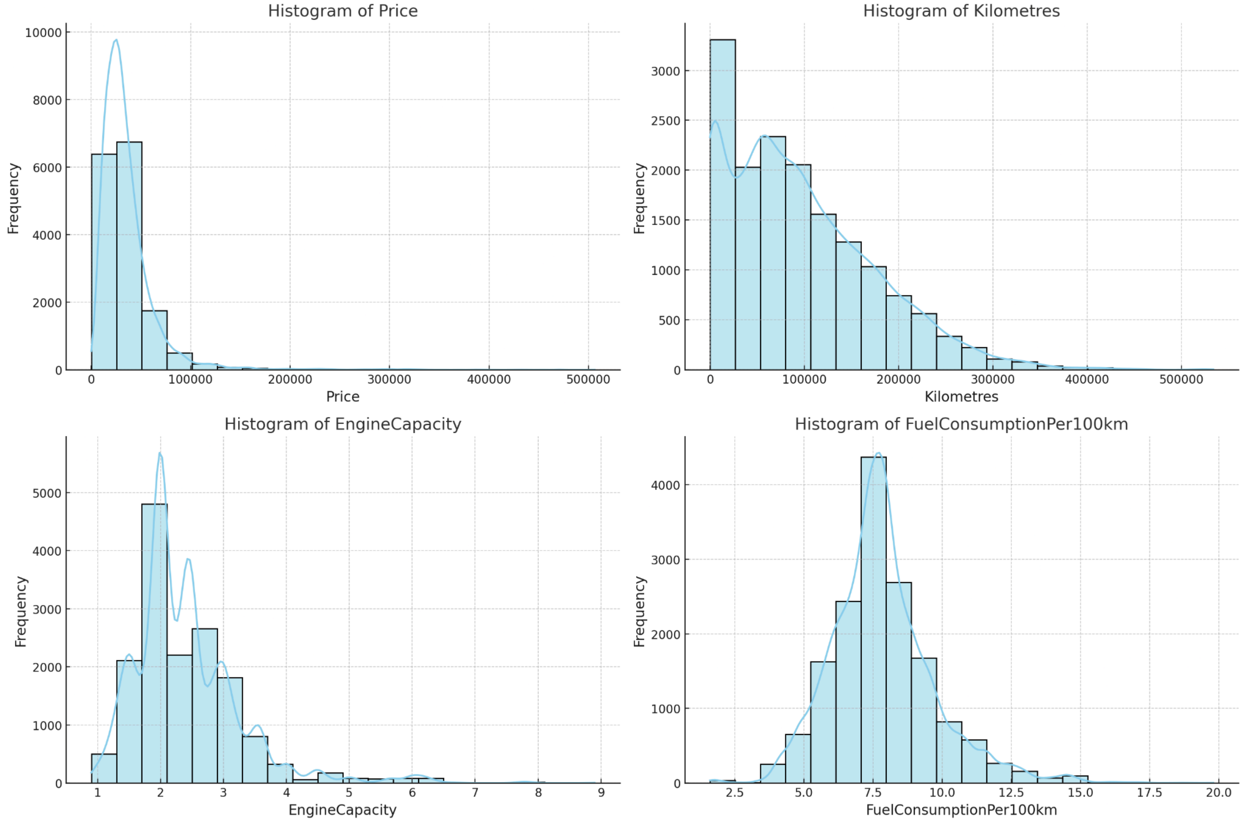
The boxplots illustrate how Price varies across several categorical variables: BodyType, FuelType, Transmission, and UsedOrNew. For BodyType, SUVs and sedans exhibit higher median prices compared to hatchbacks and commercial vehicles, while coupes and convertibles show a wider price range. In FuelType, electric and hybrid vehicles tend to have higher prices, reflecting advanced technology and market trends, while unleaded and diesel vehicles are more moderately priced. The Transmission box plot indicates that automatic vehicles generally have slightly higher prices than manual ones. Lastly, the UsedOrNew plot shows that new vehicles command the highest prices, followed by demo vehicles, with used vehicles having the lowest median price and broader variability.

Scatter plot between Price and the other numerical variables:



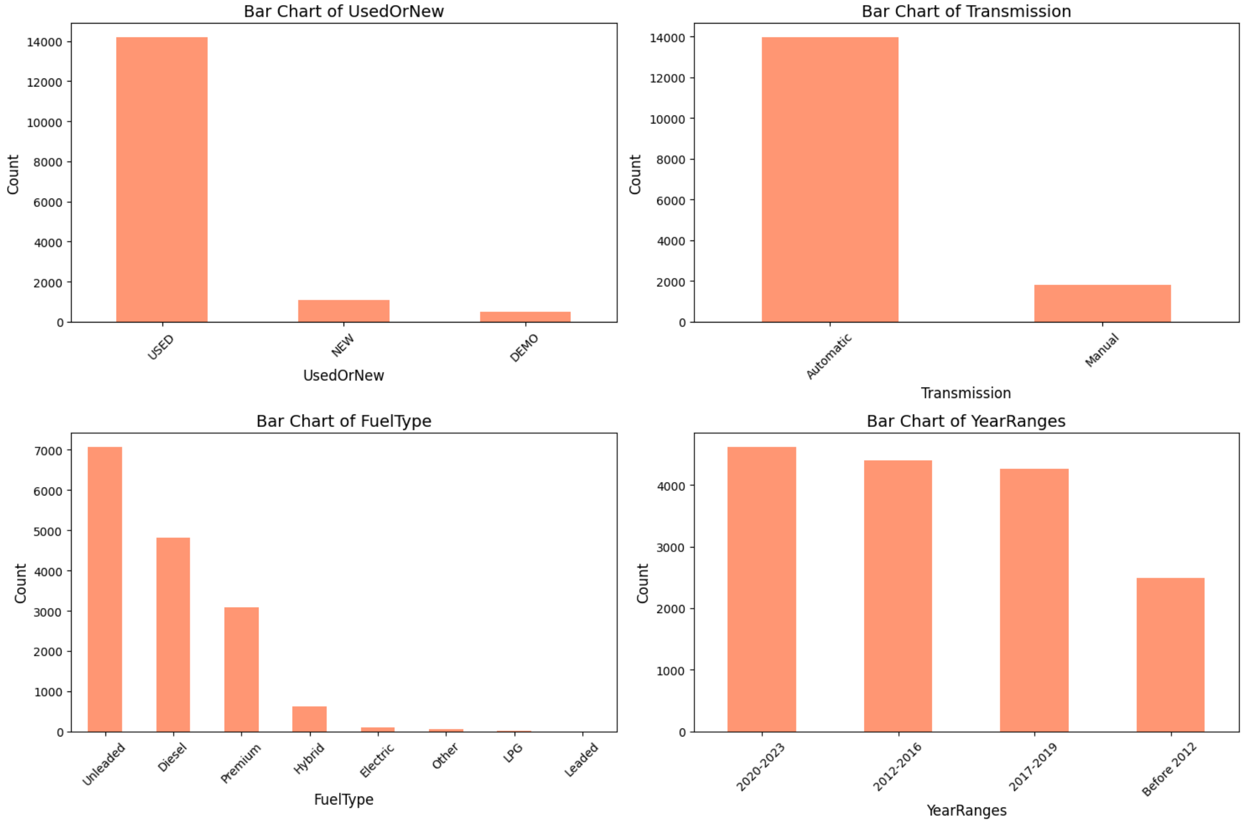
The first scatter plot shows a clear negative relationship between Price and Kilometres, with vehicles having higher mileage generally priced lower. The second plot demonstrates a weak positive trend between Price and EngineCapacity, where vehicles with larger engines tend to have slightly higher prices, although the relationship is not strongly linear. Lastly, the scatter plot for Price versus FuelConsumptionPer100km reveals no clear linear trend, though vehicles with higher fuel consumption occasionally appear at higher price ranges, indicating other factors may influence pricing in these cases.

Histograms for numerical variables:



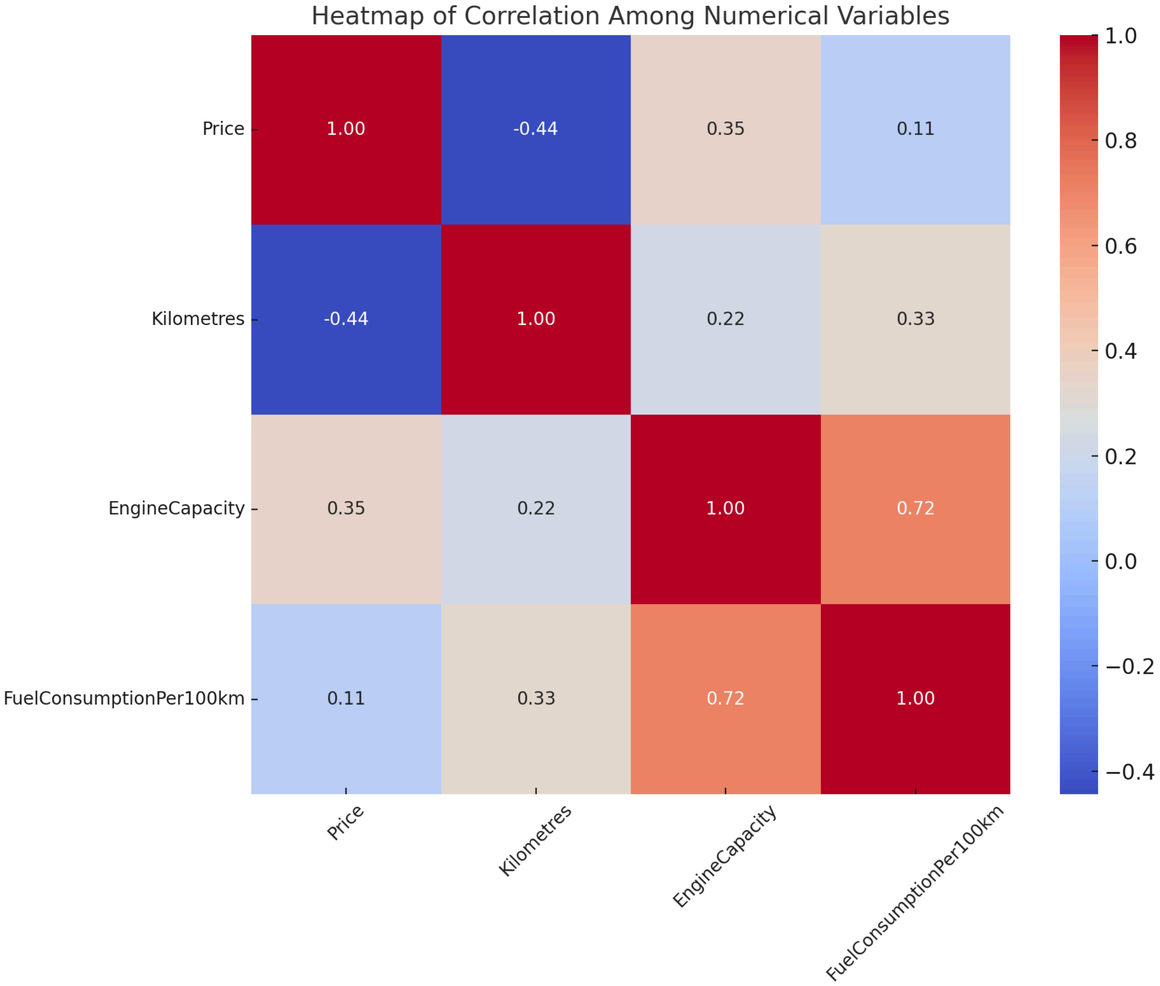
The distribution of Price is positively skewed, with the majority of vehicles priced in the lower range and a smaller number of high-priced vehicles creating a long tail. Kilometres shows a similar right-skewed pattern, where most vehicles have lower mileage, but a few outliers have significantly higher values. The EngineCapacity variable demonstrates a roughly uniform distribution, with no dominant engine sizes but a slight clustering around commonly preferred capacities. Lastly, FuelConsumptionPer100km is normally distributed.

Bar charts for some important categorical variables:



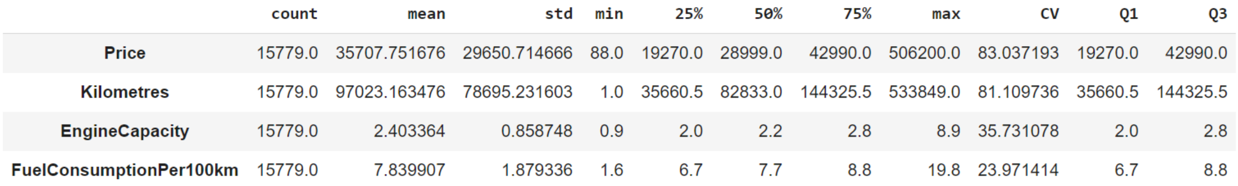
For UsedOrNew, the majority of vehicles are used, while new and demo vehicles form smaller proportions. The Transmission chart shows a strong preference for automatic vehicles, with manual transmissions being far less common. In the FuelType chart, unleaded fuel is the most prevalent, followed by diesel and premium, while hybrid and electric vehicles make up a small fraction. The YearRanges chart indicates that vehicles from the ranges 2020–2023, 2012–2016, and 2017–2019 are similarly distributed, with fewer vehicles from before 2012.

Heatmap for numerical variables:



A strong positive correlation is evident between EngineCapacity and FuelConsumptionPer100km, indicating that vehicles with larger engines tend to consume more fuel. Kilometres shows a weak negative correlation with Price, suggesting that vehicles with higher mileage are generally less expensive. Other relationships, such as those involving EngineCapacity and Kilometres, exhibit weaker correlations, reflecting minimal linear dependence.

Summary statistics for numerical variables:



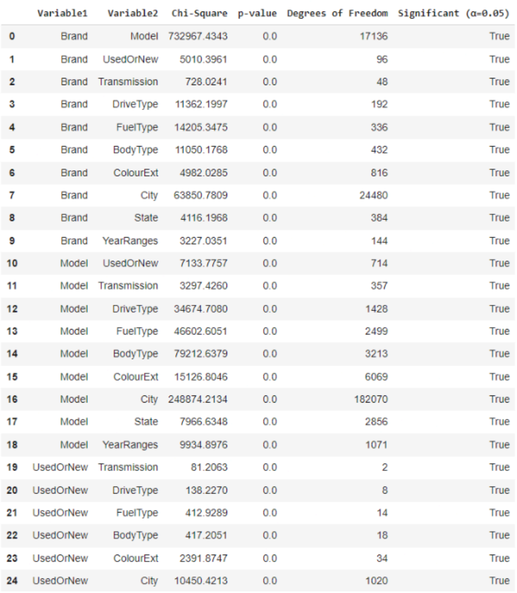
	count	mean	std	min	25%	50%	75%	max	CV	Q1	Q3
Price	15779.0	35707.751676	29650.714666	88.0	19270.0	28999.0	42990.0	506200.0	83.037193	19270.0	42990.0
Kilometres	15779.0	97023.163476	78695.231603	1.0	35660.5	82833.0	144325.5	533849.0	81.109736	35660.5	144325.5
EngineCapacity	15779.0	2.403364	0.858748	0.9	2.0	2.2	2.8	8.9	35.731078	2.0	2.8
FuelConsumptionPer100km	15779.0	7.839907	1.879336	1.6	6.7	7.7	8.8	19.8	23.971414	6.7	8.8

The mean and median values show central tendencies, with Price and Kilometres displaying a right-skewed distribution due to higher means compared to medians. Variability is evident in the standard deviation, with Price showing significant dispersion due to the wide range of vehicle prices, as reflected by a maximum of $506,200. The coefficient of variation (CV) indicates relative variability, with Kilometres and Price having high CVs, highlighting their broad distributions. Quartiles (Q1 and Q3) further emphasize data spread, particularly for Kilometres, where vehicles range from low mileage to over 144,325. Fuel efficiency, as measured by FuelConsumptionPer100km, shows less variation, suggesting more consistent data within this variable.

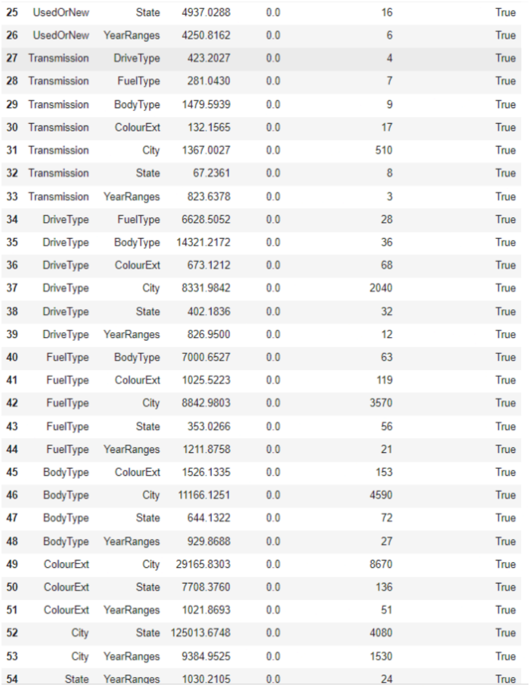
The chosen descriptive techniques effectively summarize and explore the dataset's diverse characteristics, which include both numerical and categorical variables. For numerical variables, summary statistics such as mean, median, standard deviation, and quartiles provide a clear understanding of central tendency and variability, while visualizations like histograms and boxplots reveal data distributions and potential outliers. Heatmaps highlight correlations, offering insights into relationships between variables such as EngineCapacity and FuelConsumptionPer100km. For categorical variables, bar charts display proportions and frequencies, making it easy to interpret distributions across categories like UsedOrNew and FuelType. Grouped boxplots bridge numerical and categorical data, allowing for comparisons across categories, such as price differences between new and used vehicles.

Subsection 2.2: *Inferential Statistics*

Chi-Square test for independence:



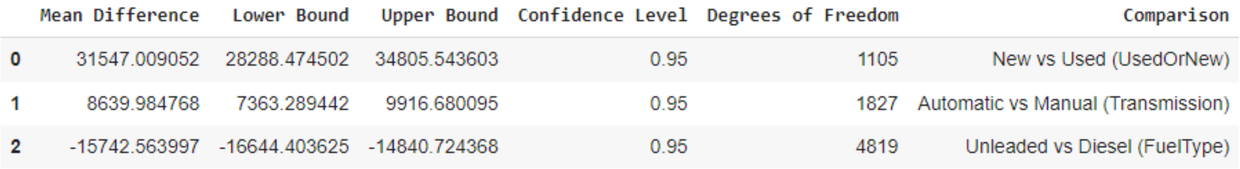
	Variable1	Variable2	Chi-Square	p-value	Degrees of Freedom	Significant (α=0.05)
0	Brand	Model	732967.4343	0.0	17136	True
1	Brand	UsedOrNew	5010.3961	0.0	96	True
2	Brand	Transmission	728.0241	0.0	48	True
3	Brand	DriveType	11362.1997	0.0	192	True
4	Brand	FuelType	14205.3475	0.0	336	True
5	Brand	BodyType	11050.1768	0.0	432	True
6	Brand	ColourExt	4982.0285	0.0	816	True
7	Brand	City	63850.7809	0.0	24480	True
8	Brand	State	4116.1968	0.0	384	True
9	Brand	YearRanges	3227.0351	0.0	144	True
10	Model	UsedOrNew	7133.7757	0.0	714	True
11	Model	Transmission	3297.4260	0.0	357	True
12	Model	DriveType	34674.7080	0.0	1428	True
13	Model	FuelType	46602.6051	0.0	2499	True
14	Model	BodyType	79212.6379	0.0	3213	True
15	Model	ColourExt	15126.8046	0.0	6069	True
16	Model	City	248874.2134	0.0	182070	True
17	Model	State	7966.6348	0.0	2856	True
18	Model	YearRanges	9934.8976	0.0	1071	True
19	UsedOrNew	Transmission	81.2063	0.0	2	True
20	UsedOrNew	DriveType	138.2270	0.0	8	True
21	UsedOrNew	FuelType	412.9289	0.0	14	True
22	UsedOrNew	BodyType	417.2051	0.0	18	True
23	UsedOrNew	ColourExt	2391.8747	0.0	34	True
24	UsedOrNew	City	10450.4213	0.0	1020	True



25	UsedOrNew	State	4937.0288	0.0	16	True
26	UsedOrNew	YearRanges	4250.8162	0.0	6	True
27	Transmission	DriveType	423.2027	0.0	4	True
28	Transmission	FuelType	281.0430	0.0	7	True
29	Transmission	BodyType	1479.5939	0.0	9	True
30	Transmission	ColourExt	132.1565	0.0	17	True
31	Transmission	City	1367.0027	0.0	510	True
32	Transmission	State	67.2361	0.0	8	True
33	Transmission	YearRanges	823.6378	0.0	3	True
34	DriveType	FuelType	6628.5052	0.0	28	True
35	DriveType	BodyType	14321.2172	0.0	36	True
36	DriveType	ColourExt	673.1212	0.0	68	True
37	DriveType	City	8331.9842	0.0	2040	True
38	DriveType	State	402.1836	0.0	32	True
39	DriveType	YearRanges	826.9500	0.0	12	True
40	FuelType	BodyType	7000.6527	0.0	63	True
41	FuelType	ColourExt	1025.5223	0.0	119	True
42	FuelType	City	8842.9803	0.0	3570	True
43	FuelType	State	353.0266	0.0	56	True
44	FuelType	YearRanges	1211.8758	0.0	21	True
45	BodyType	ColourExt	1526.1335	0.0	153	True
46	BodyType	City	11166.1251	0.0	4590	True
47	BodyType	State	644.1322	0.0	72	True
48	BodyType	YearRanges	929.8688	0.0	27	True
49	ColourExt	City	29165.8303	0.0	8670	True
50	ColourExt	State	7708.3760	0.0	136	True
51	ColourExt	YearRanges	1021.8693	0.0	51	True
52	City	State	125013.6748	0.0	4080	True
53	City	YearRanges	9384.9525	0.0	1530	True
54	State	YearRanges	1030.2105	0.0	24	True

For all combinations of categorical variables, the p-values are less than 0.05, leading us to reject the null hypothesis and conclude that there is a statistically significant association between all pairs of variables. The strongest association is observed between Brand and Model, as expected, since brands often have unique models. Additionally, Brand is significantly associated with UsedOrNew, Transmission, DriveType, and FuelType, indicating that a vehicle's brand strongly influences its condition, transmission type, drivetrain, and fuel type. These associations suggest that brand plays a pivotal role in determining these characteristics.

Confidence intervals for difference in means:



	Mean Difference	Lower Bound	Upper Bound	Confidence Level	Degrees of Freedom	Comparison
0	31547.009052	28288.474502	34805.543603	0.95	1105	New vs Used (UsedOrNew)
1	8639.984768	7363.289442	9916.680095	0.95	1827	Automatic vs Manual (Transmission)
2	-15742.563997	-16644.403625	-14840.724368	0.95	4819	Unleaded vs Diesel (FuelType)

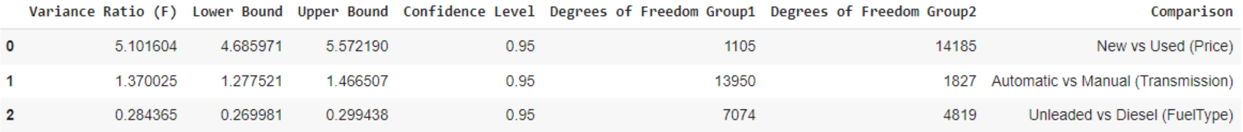
The variables for comparison were chosen based on their practical relevance and ability to provide meaningful insights about pricing trends. For New vs Used vehicles, the confidence interval for the difference in mean prices is [28,288.47, 34,805.54], with a mean difference of 31,547.00. This comparison is crucial because the condition of a vehicle (new or used) is one of the primary factors influencing its price, and understanding this difference is important for buyers and sellers in the automotive market.

For the Automatic vs Manual transmission comparison, the confidence interval for the mean price difference is [7,363.29, 9,916.68], with a mean difference of 8,639.98. Transmission type significantly impacts vehicle pricing, as automatic transmissions are generally preferred in many markets due to convenience, making this an insightful comparison.

Finally, for the Unleaded vs Diesel fuel type comparison, the confidence interval is [-16,644.40, -14,840.72], with a mean difference of -15,742.56. Fuel type is another key characteristic that affects pricing because it reflects differences in vehicle technology, efficiency, and target markets. Diesel vehicles are typically more expensive due to better fuel efficiency and higher engine longevity.

In all three cases, the confidence intervals do not contain zero, leading to the rejection of the null hypothesis. This confirms significant differences in mean prices between the respective groups, and the chosen variables allow us to highlight essential pricing trends in the dataset.

Confidence intervals for ratio of variances:



	Variance Ratio (F)	Lower Bound	Upper Bound	Confidence Level	Degrees of Freedom Group1	Degrees of Freedom Group2	Comparison
0	5.101604	4.685971	5.572190	0.95	1105	14185	New vs Used (Price)
1	1.370025	1.277521	1.466507	0.95	13950	1827	Automatic vs Manual (Transmission)
2	0.284365	0.269981	0.299438	0.95	7074	4819	Unleaded vs Diesel (FuelType)

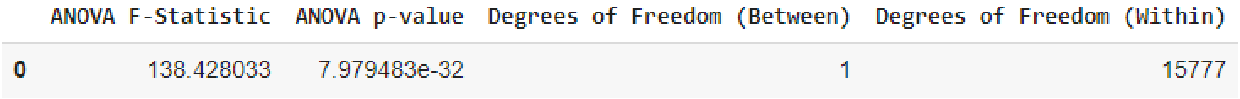
For the comparison of New vs Used (Price), the variance ratio is 5.10, with a 95% confidence interval of [4.69, 5.57]. This indicates that the price variability for new vehicles is significantly higher than for used vehicles, as the confidence interval does not include 1. This comparison was chosen because understanding the variability in pricing between new and used vehicles provides key insights into market behavior and pricing trends.

For Automatic vs Manual (Transmission), the variance ratio is 1.37, with a confidence interval of [1.28, 1.47]. This suggests that the price variability for vehicles with automatic transmissions is slightly higher than for those with manual transmissions. This variable was chosen as transmission type is a critical factor in pricing, and its variability reflects differences in consumer preferences and technology costs.

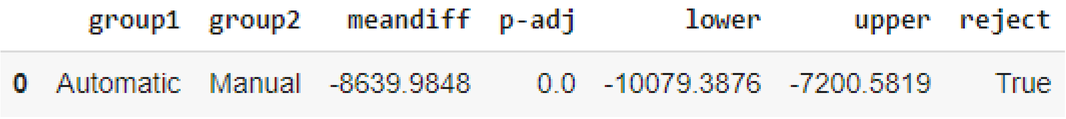
Finally, for Unleaded vs Diesel (FuelType), the variance ratio is 0.28, with a confidence interval of [0.27, 0.30]. This result indicates that price variability for diesel vehicles is significantly lower than for unleaded vehicles. This comparison is important because fuel type is a major determinant of vehicle price, and understanding its variability helps in analyzing market preferences and technological impacts.

In all cases, the confidence intervals do not include 1, leading to the rejection of the null hypothesis and confirming significant differences in price variability between the respective groups. These variables were specifically chosen for their relevance in influencing vehicle pricing and their practical significance in understanding market dynamics.

One-way ANOVA:



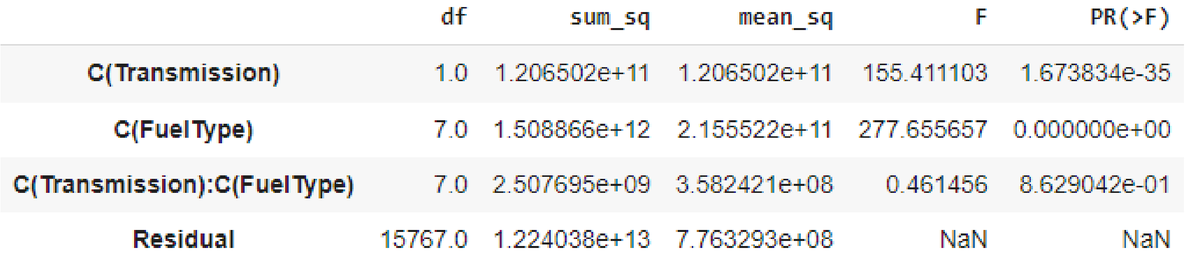
	ANOVA F-Statistic	ANOVA p-value	Degrees of Freedom (Between)	Degrees of Freedom (Within)
0	138.428033	7.979483e-32	1	15777



	group1	group2	meandiff	p-adj	lower	upper	reject
0	Automatic	Manual	-8639.9848	0.0	-10079.3876	-7200.5819	True

The one-way ANOVA results show an F-statistic of 138.43 and a p-value of 7.98 × 10^{-32}, which is far below the significance level of 0.05. This indicates a statistically significant difference in mean Price between automatic and manual vehicles. The Tukey HSD results further confirm that the mean difference in Price between automatic and manual vehicles is -8,639.98, with a 95% confidence interval of [-10,079.39, -7,200.58]. Since the confidence interval does not contain zero, we can confidently reject the null hypothesis and conclude that vehicles with automatic transmissions are significantly more expensive than those with manual transmissions.

Two-way ANOVA:

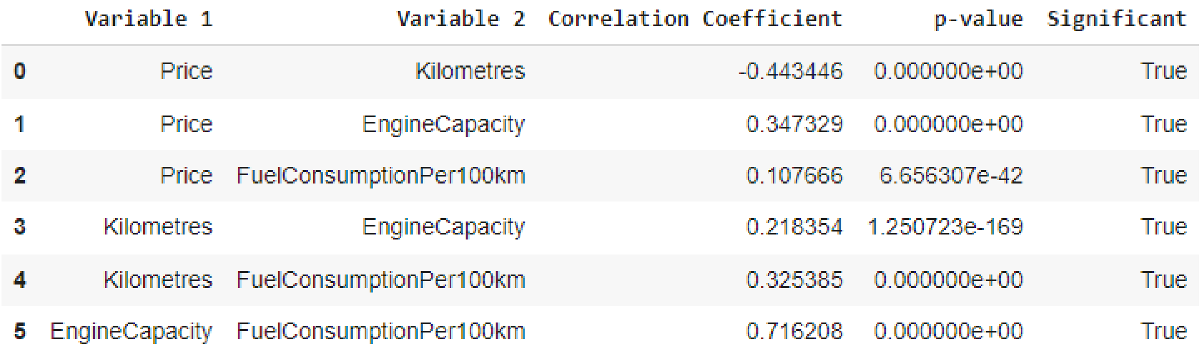


	df	sum_sq	mean_sq	F	PR(>F)
C(Transmission)	1.0	1.206502e+11	1.206502e+11	155.411103	1.673834e-35
C(FuelType)	7.0	1.508866e+12	2.155522e+11	277.655657	0.000000e+00
C(Transmission):C(FuelType)	7.0	2.507695e+09	3.582421e+08	0.461456	8.629042e-01
Residual	15767.0	1.224038e+13	7.763293e+08	NaN	NaN

The main effect of Transmission is significant with an F-value of 155.41 and a p-value of 1.67 × 10^{-35}, indicating that the mean Price differs significantly between vehicles with automatic and manual transmissions.

The main effect of FuelType is highly significant with an F-value of 277.66 and a p-value of 0.000, showing that the mean Price varies significantly across different fuel types.
The interaction effect of Transmission and FuelType is not significant, with an F-value of 0.46 and a p-value of 0.86, suggesting that the combined influence of Transmission and FuelType on Price is not statistically significant.

Correlation analysis:



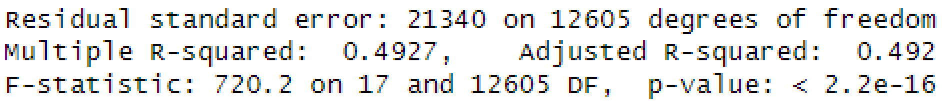
	Variable 1	Variable 2	Correlation Coefficient	p-value	Significant
0	Price	Kilometres	-0.443446	0.000000e+00	True
1	Price	EngineCapacity	0.347329	0.000000e+00	True
2	Price	FuelConsumptionPer100km	0.107666	6.656307e-42	True
3	Kilometres	EngineCapacity	0.218354	1.250723e-169	True
4	Kilometres	FuelConsumptionPer100km	0.325385	0.000000e+00	True
5	EngineCapacity	FuelConsumptionPer100km	0.716208	0.000000e+00	True

1. **Price and Kilometres:** The correlation coefficient is -0.443, indicating a moderate negative relationship. As the kilometres increase, the price tends to decrease, likely reflecting depreciation with higher usage. The p-value is extremely small ($p < 0.05$), confirming this relationship is statistically significant.
2. **Price and EngineCapacity:** The correlation coefficient is 0.347, suggesting a weak to moderate positive relationship. Vehicles with larger engine capacities tend to have higher prices, and the significant p-value supports this.
3. **Price and FuelConsumptionPer100km:** The correlation coefficient is 0.108, indicating a weak positive relationship. Higher fuel consumption is slightly associated with higher prices, and the relationship is statistically significant.
4. **Kilometres and EngineCapacity:** The correlation coefficient is 0.218, showing a weak positive relationship. Vehicles with higher kilometres tend to have slightly larger engine capacities, possibly reflecting older models or high-mileage vehicles. This relationship is also significant.
5. **Kilometres and FuelConsumptionPer100km:** The correlation coefficient is 0.325, indicating a weak positive relationship. Vehicles with higher mileage tend to have slightly higher fuel consumption, possibly due to wear and tear or older technology.
6. **EngineCapacity and FuelConsumptionPer100km:** The correlation coefficient is 0.716, reflecting a strong positive relationship. Larger engines are strongly associated with higher fuel consumption, as expected, and the relationship is statistically significant.

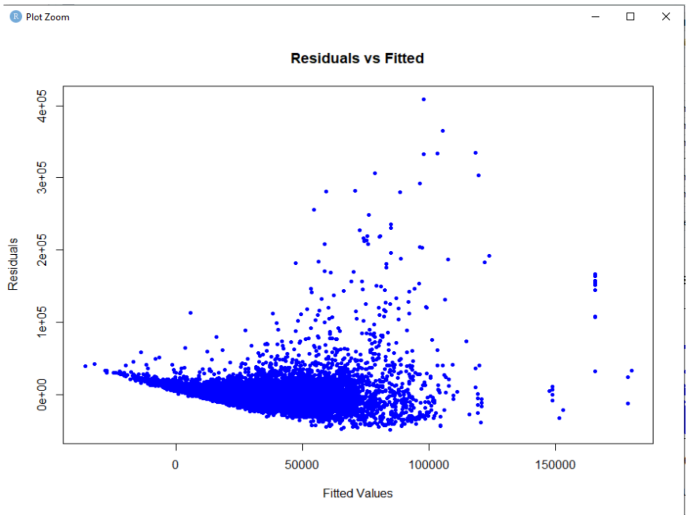
Section 3: Model Building

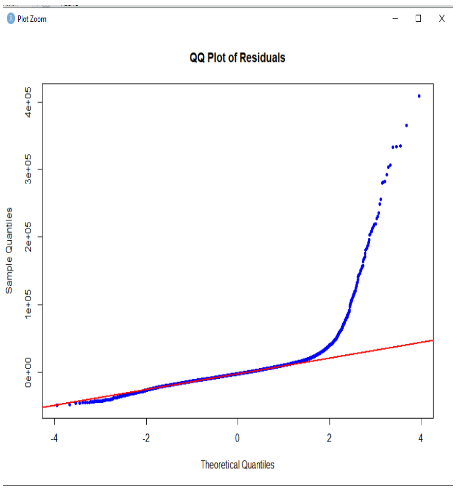
Subsection 3.1: *Generalized Linear regression model building*

The target variable Price is a numerical continuous variable and we have 17 predictors consisting of both numerical and categorical variables. Before building a multiple regression model we split the training and test dataset in the ratio of 80% and 20%. We further numerically encoded the categorical variables. The initial Linear Regression model fitted on the training data without any box cox or removing outlier procedures gave us an R2 of 0.4927 and multiple R2 of 0.492.

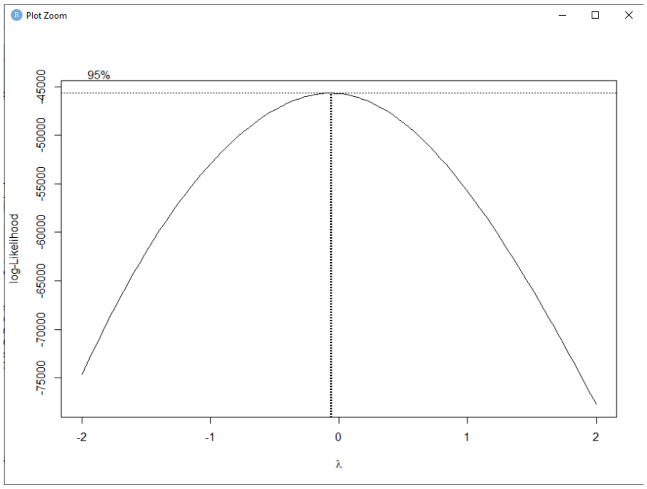


```
Residual standard error: 21340 on 12605 degrees of freedom
Multiple R-squared:  0.4927,    Adjusted R-squared:  0.492
F-statistic: 720.2 on 17 and 12605 DF,  p-value: < 2.2e-16
```

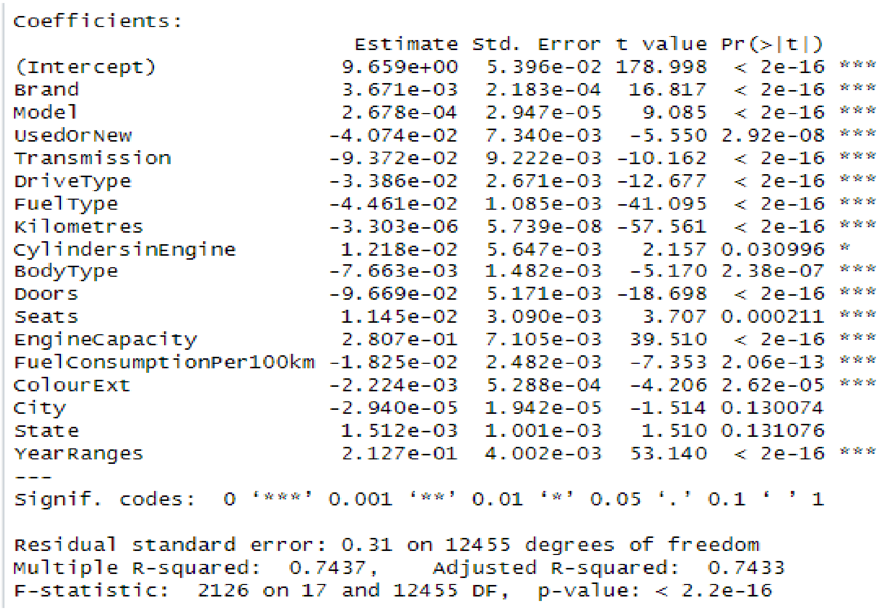




After generating the residual vs fitted and qq plots for the same, we observed that the variance wasn’t constant as the pattern wasn't random and was fanning out as the values increased. Similarly for the qq plot there were a large number of outliers present and the distribution appeared to be skewed to the left rather than matching a normal distribution. To tackle such issues we did box cox and removed outliers using Z-score.

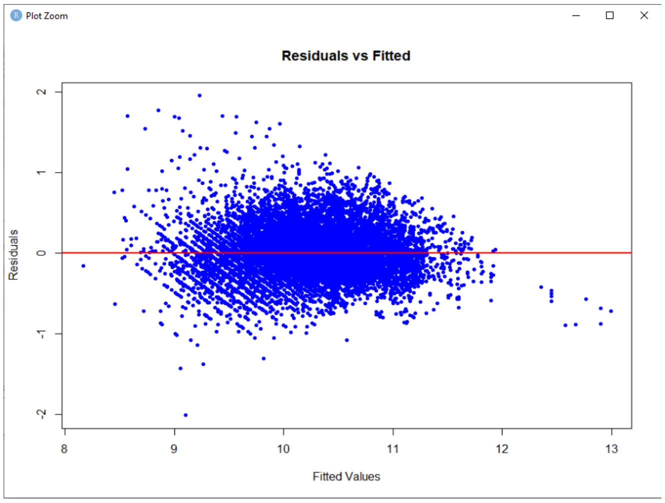


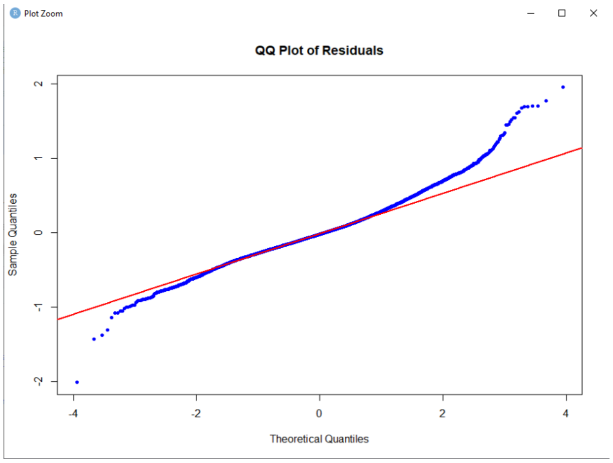
The optimal Lambda was closer to 0, hence we opted for a log transformation of Price and using a Z score threshold of 3 we removed 200 outliers from the training data.



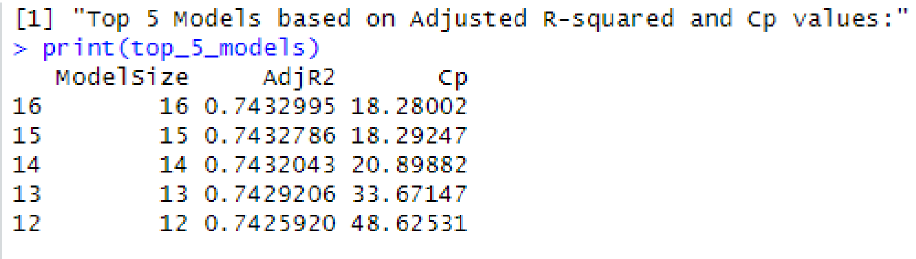
```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             9.659e+00  5.396e-02 178.998  < 2e-16 ***
Brand                   3.671e-03  2.183e-04  16.817  < 2e-16 ***
Model                   2.678e-04  2.947e-05   9.085  < 2e-16 ***
UsedOrNew              -4.074e-02  7.340e-03  -5.550 2.92e-08 ***
Transmission           -9.372e-02  9.222e-03 -10.162  < 2e-16 ***
DriveType              -3.386e-02  2.671e-03 -12.677  < 2e-16 ***
FuelType               -4.461e-02  1.085e-03 -41.095  < 2e-16 ***
Kilometres             -3.303e-06  5.739e-08 -57.561  < 2e-16 ***
CylindersinEngine       1.218e-02  5.647e-03   2.157 0.030996 *  
BodyType               -7.663e-03  1.482e-03  -5.170 2.38e-07 ***
Doors                  -9.669e-02  5.171e-03 -18.698  < 2e-16 ***
Seats                   1.145e-02  3.090e-03   3.707 0.000211 ***
EngineCapacity          2.807e-01  7.105e-03  39.510  < 2e-16 ***
FuelConsumptionPer100km -1.825e-02  2.482e-03  -7.353 2.06e-13 ***
ColourExt              -2.224e-03  5.288e-04  -4.206 2.62e-05 ***
City                   -2.940e-05  1.942e-05  -1.514 0.130074    
State                   1.512e-03  1.001e-03   1.510 0.131076    
YearRanges              2.127e-01  4.002e-03  53.140  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.31 on 12455 degrees of freedom
Multiple R-squared:  0.7437,	Adjusted R-squared:  0.7433 
F-statistic:  2126 on 17 and 12455 DF,  p-value: < 2.2e-16
```

After doing the above, we again fitted a multiple regression model after doing Boxcox and removing the required outliers. This time we can see that the R2 and adj R2 improved significantly. The R2 jumped to 0.7437 and multiple R2 jumped to around 0.7433. Also we can observe high p values for the variables City & State.



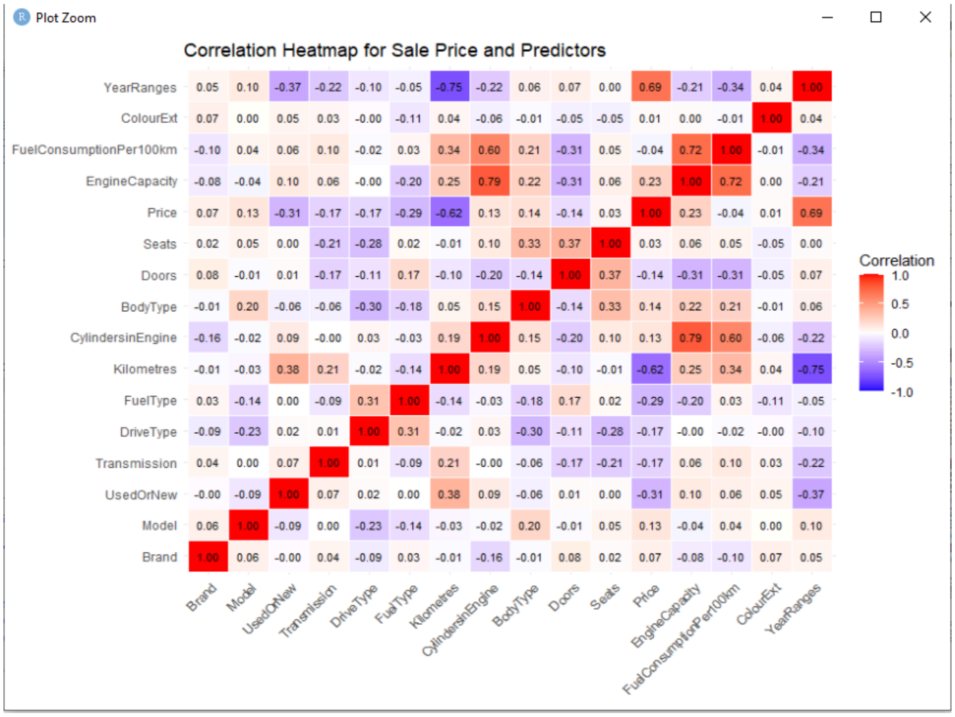


Similarly we can see in the residuals vs fitted model, the pattern is random , hence the variance is constant and for the qq plot the distribution closely resembles a normal distribution.



```
[1] "Top 5 Models based on Adjusted R-squared and Cp values:"
> print(top_5_models)
   ModelSize     AdjR2       Cp
16        16 0.7432995 18.28002
15        15 0.7432786 18.29247
14        14 0.7432043 20.89882
13        13 0.7429206 33.67147
12        12 0.7425920 48.62531
```

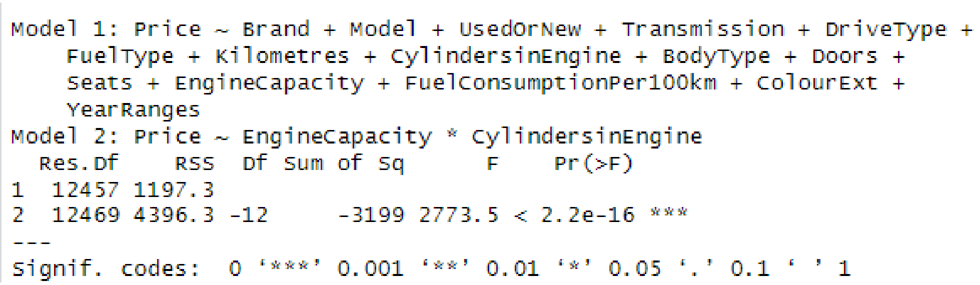
We further conducted a best subsets regression excluding the full model to see which variables can be dropped in the model with 16 and 15 predictors. We observed that city and state were not considered and as the R2 is almost similar for predictor size 15 and 16, we decided to drop City and State columns. Previously as well City and State had high p values and were not significant in predicting the price of the car.



Correlation Heatmap for Sale Price and Predictors

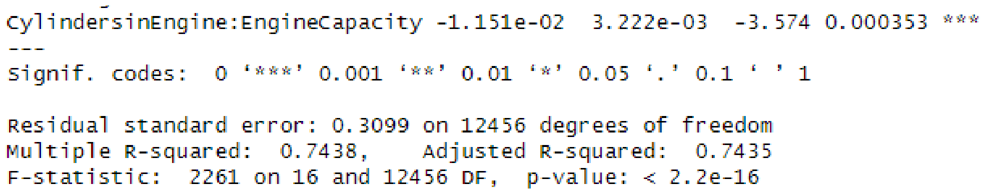
	Brand	Model	UsedOrNew	Transmission	DriveType	FuelType	Kilometres	CylindersinEngine	BodyType	Doors	Seats	Price	EngineCapacity	FuelConsumptionPer100km	ColourExt	YearRanges
YearRanges	0.05	0.10	-0.37	-0.22	-0.10	-0.05	-0.75	-0.22	0.06	0.07	0.00	0.69	-0.21	-0.34	0.04	1.00
ColourExt	0.07	0.00	0.05	0.03	-0.00	-0.11	0.04	-0.06	-0.01	-0.05	-0.05	0.01	0.00	-0.01	1.00	0.04
FuelConsumptionPer100km	-0.10	0.04	0.06	0.10	-0.02	0.03	0.34	0.60	0.21	-0.31	0.05	-0.04	0.72	1.00	-0.01	-0.34
EngineCapacity	-0.08	-0.04	0.10	0.06	-0.00	-0.20	0.25	0.79	0.22	-0.31	0.06	0.23	1.00	0.72	0.00	-0.21
Price	0.07	0.13	-0.31	-0.17	-0.17	-0.29	-0.62	0.13	0.14	-0.14	0.03	1.00	0.23	-0.04	0.01	0.69
Seats	0.02	0.05	0.00	-0.21	-0.28	0.02	-0.01	0.10	0.33	0.37	1.00	0.03	0.06	0.05	-0.05	0.00
Doors	0.08	-0.01	0.01	-0.17	-0.11	0.17	-0.10	-0.20	-0.14	1.00	0.37	-0.14	-0.31	-0.31	-0.05	0.07
BodyType	-0.01	0.20	-0.06	-0.06	-0.30	-0.18	0.05	0.15	1.00	-0.14	0.33	0.14	0.22	0.21	-0.01	0.06
CylindersinEngine	-0.16	-0.02	0.09	-0.00	0.03	-0.03	0.19	1.00	0.15	-0.20	0.10	0.13	0.79	0.60	-0.06	-0.22
Kilometres	-0.01	-0.03	0.38	0.21	-0.02	-0.14	1.00	0.19	0.05	-0.10	-0.01	-0.62	0.25	0.34	0.04	-0.75
FuelType	0.03	-0.14	0.00	-0.09	0.31	1.00	-0.14	-0.03	-0.18	0.17	0.02	-0.29	-0.20	0.03	-0.11	-0.05
DriveType	-0.09	-0.23	0.02	0.01	1.00	0.31	-0.02	0.03	-0.30	-0.11	-0.28	-0.17	-0.00	-0.02	-0.00	-0.10
Transmission	0.04	0.00	0.07	1.00	0.01	-0.09	0.21	-0.00	-0.06	-0.17	-0.21	-0.17	0.06	0.10	0.03	-0.22
UsedOrNew	-0.00	-0.09	1.00	0.07	0.02	0.00	0.38	0.09	-0.06	0.01	0.00	-0.31	0.10	0.06	0.05	-0.37
Model	0.06	1.00	-0.09	0.00	-0.23	-0.14	-0.03	-0.02	0.20	-0.01	0.05	0.13	-0.04	0.04	0.00	0.10
Brand	1.00	0.06	-0.00	0.04	-0.09	0.03	-0.01	-0.16	-0.01	0.08	0.02	0.07	-0.08	-0.10	0.07	0.05

We generated a heatmap to scout for possible interaction terms, here in the heatmap we observe high correlation between EngineCapacity and CylindersinEngine & EngineCapacity and FuelConsumptionPer100km. We decided to test the significance of the interaction terms by doing an anova comparison which revealed that the term EngineCapacity*CylindersinEngine was significant. However, while conducting measures such as AIC and BIC revealed the term was adding higher complexity to the linear model than it was helpful in the predicting the price of automobile.



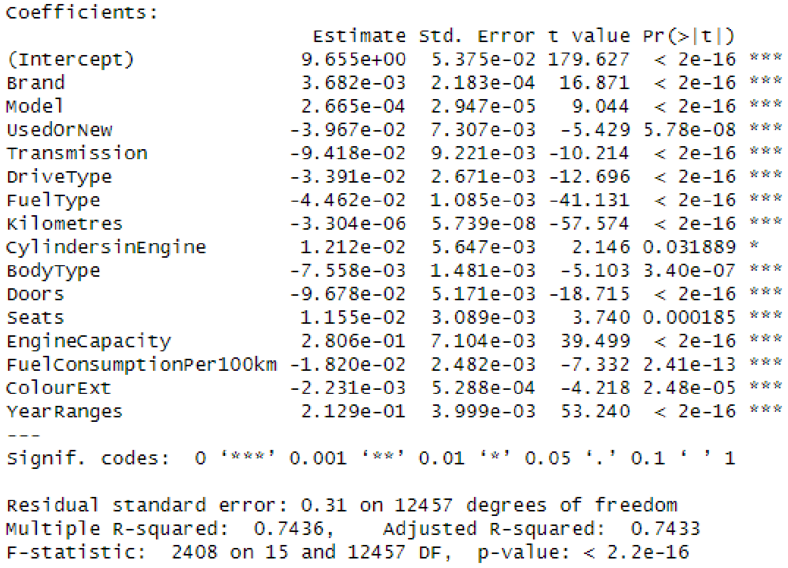
```
Model 1: Price ~ Brand + Model + UsedOrNew + Transmission + DriveType +
    FuelType + Kilometres + CylindersinEngine + BodyType + Doors +
    Seats + EngineCapacity + FuelConsumptionPer100km + ColourExt +
    YearRanges
Model 2: Price ~ EngineCapacity * CylindersinEngine
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1  12457 1197.3
2  12469 4396.3 -12     -3199 2773.5 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To clarify the same we fitted the linear regression model including the interaction term and generated anova table below. Here we observed there was a minimal change in R2 and adj R2. R2 and adj R2 increased to 0.7438 and 0.7435 respectively. As the change was insignificant we decided to drop the extra interaction term EngineCapacity*CylindersinEngine for model interpretability.



```
CylindersinEngine:EngineCapacity -1.151e-02  3.222e-03  -3.574 0.000353 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3099 on 12456 degrees of freedom
Multiple R-squared:  0.7438,    Adjusted R-squared:  0.7435 
F-statistic:  2261 on 16 and 12456 DF,  p-value: < 2.2e-16
```

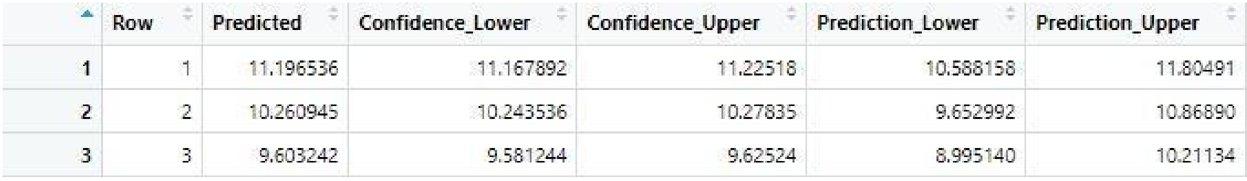
Final model:



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              9.655e+00  5.375e-02 179.627  < 2e-16 ***
Brand                    3.682e-03  2.183e-04  16.871  < 2e-16 ***
Model                    2.665e-04  2.947e-05   9.044  < 2e-16 ***
UsedOrNew               -3.967e-02  7.307e-03  -5.429 5.78e-08 ***
Transmission            -9.418e-02  9.221e-03 -10.214  < 2e-16 ***
DriveType               -3.391e-02  2.671e-03 -12.696  < 2e-16 ***
FuelType                -4.462e-02  1.085e-03 -41.131  < 2e-16 ***
Kilometres              -3.304e-06  5.739e-08 -57.574  < 2e-16 ***
CylindersinEngine        1.212e-02  5.647e-03   2.146 0.031889 *  
BodyType                -7.558e-03  1.481e-03  -5.103 3.40e-07 ***
Doors                   -9.678e-02  5.171e-03 -18.715  < 2e-16 ***
Seats                    1.155e-02  3.089e-03   3.740 0.000185 ***
EngineCapacity           2.806e-01  7.104e-03  39.499  < 2e-16 ***
FuelConsumptionPer100km -1.820e-02  2.482e-03  -7.332 2.41e-13 ***
ColourExt               -2.231e-03  5.288e-04  -4.218 2.48e-05 ***
YearRanges               2.129e-01  3.999e-03  53.240  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.31 on 12457 degrees of freedom
Multiple R-squared:  0.7436,    Adjusted R-squared:  0.7433 
F-statistic:  2408 on 15 and 12457 DF,  p-value: < 2.2e-16
```

Hence based on our training data we narrowed down a multiple regression model of 15 variables for predicting the Price of the automobile.

Prediction Intervals (PI) and Confidence Intervals (CI):



	Row	Predicted	Confidence_Lower	Confidence_Upper	Prediction_Lower	Prediction_Upper
1	1	11.196536	11.167892	11.22518	10.588158	11.80491
2	2	10.260945	10.243536	10.27835	9.652992	10.86890
3	3	9.603242	9.581244	9.62524	8.995140	10.21134

This table shows the predicted Price values along with the corresponding CI and PI for the first three rows of the dataset based on the multiple regression model. The CI represents the range where the true mean Price is expected to lie for a given input, while the PI provides a broader range that captures individual predictions, accounting for variability in the data.

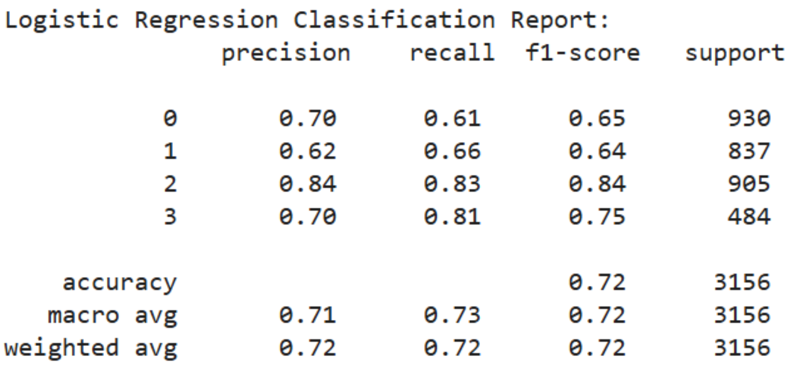
Subsection 3.2: *Classification model through Logistic regression*

In this section, we applied several classification models on the dataset to predict the target variable YearRanges, which is ordinal and multiclass and can take on four possible values: Before 2012, 2012-2016, 2017-2019, and 2020-2023. Due to the variety of features and the potential complexity of their relationships, we tested a range of models including Naive Bayes, Random Forest, and Logistic Regression, comparing their performance and interpretability.

Additionally, further preprocessing steps were performed that helped improve the performance of the models. Non-numerical categorical variables were numerically encoded, and the target variable YearRanges was label encoded to represent each category as an integer. Furthermore, numerical columns were standardized using standard scaling to ensure a uniform scale.

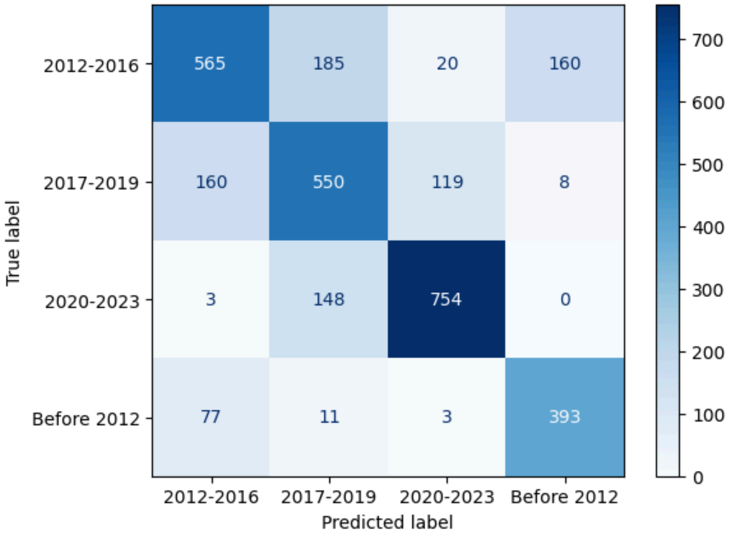
First, we trained Naive Bayes and Random Forest models to obtain a baseline performance for the classification task on this dataset. Naive Bayes achieved an accuracy of 65% while Random Forest achieved an accuracy of 80%. The poor performance of Naive Bayes can be attributed to the fact that the data violates the main assumption of Naive Bayes, which is that the features are independent. Random Forest performed better because it can model nonlinear relationships between the features. The precision, recall, F1-scores, and confusion matrices (which can be found in ClassificationModels.ipynb) indicated that the models struggled with 2012-2016 and 2017-2019 classes while performing better in Before 2019 and 2020-2023 classes. However, for the sake of interpretability, we chose to proceed with several Logistic Regression models since they give us direct access to the coefficients and regression equations, allowing us to understand the effects of each feature.

The next model we trained was a multinomial Logistic Regression with added polynomial features. Typically, Logistic Regression fits a linear decision boundary, which may not capture the complex and nonlinear relationships of the data. By adding polynomial features of degree 2, we introduced squared terms and pairwise interaction terms among the features, allowing the model to build more flexible decision boundaries. When the polynomial features were added to the data, the dimensionality increased significantly with 170 variables. Although the added complexity can improve performance, it makes the model harder to interpret and more prone to overfitting. We trained the model with L2 regularization to prevent overfitting, and the achieved accuracy was 72%. This was an improvement over Naive Bayes and the basic Logistic Regression model (which is discussed later), but less accurate than Random Forest.

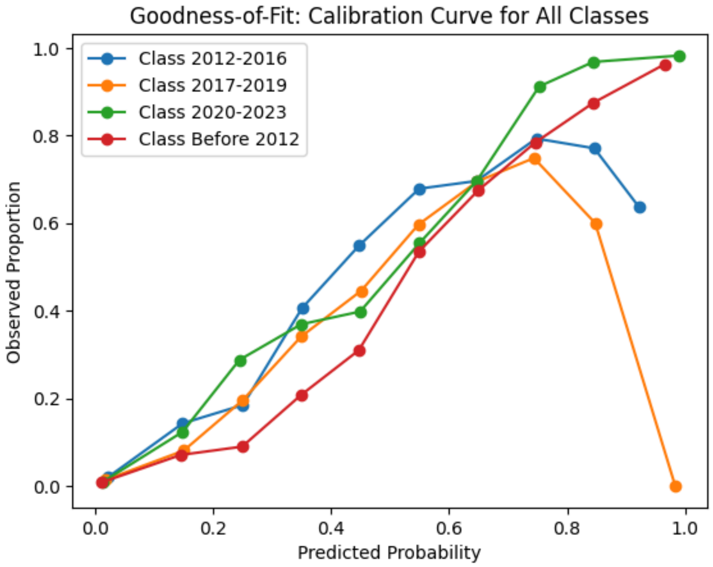


```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.61      0.65       930
           1       0.62      0.66      0.64       837
           2       0.84      0.83      0.84       905
           3       0.70      0.81      0.75       484

    accuracy                           0.72      3156
   macro avg       0.71      0.73      0.72      3156
weighted avg       0.72      0.72      0.72      3156
```

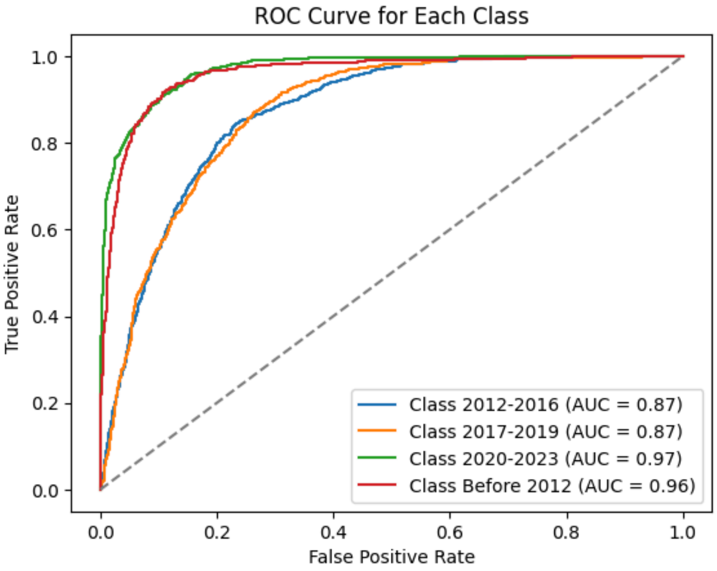
The classification report showed decent precision and recall for most YearRanges categories. The model performed relatively well for the Before 2012 (class 3) and 2020-2023 (class 2) classes, but it still faced difficulty in classifying the intermediate categories. Overall, the precision, recall, and F1-scores were worse than the Random Forest classifier.



The confusion matrix provides further details on the misclassifications that occurred. For example, several 2012-2016 vehicles were misclassified as 2017-2019. This suggests that adding quadratic terms and interactions did not fully enable the model to distinguish between adjacent year ranges.



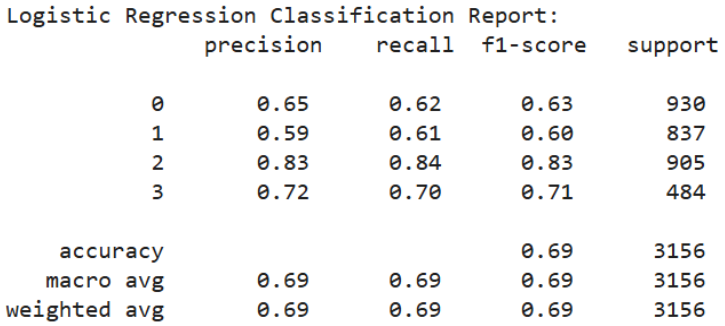
We plotted the calibration curves to assess whether the predicted probabilities matched the observed probabilities. Ideally, the points should lie along the diagonal line. Here, the calibration curves suggest that the model isn't perfectly calibrated. It can be overconfident or underconfident for each class, making the predicted probabilities unreliable.



Looking at the ROC curve, we found that the model can identify the Before 2012 and 2020-2023 classes very well, rarely confusing them with other categories, with an AUC above 0.96. As for the 2012-2016 and 2017-2019 categories, the model still performed reasonably well with an AUC of 0.87, but not for extreme values, indicating that these intermediate classes share common features with other categories.

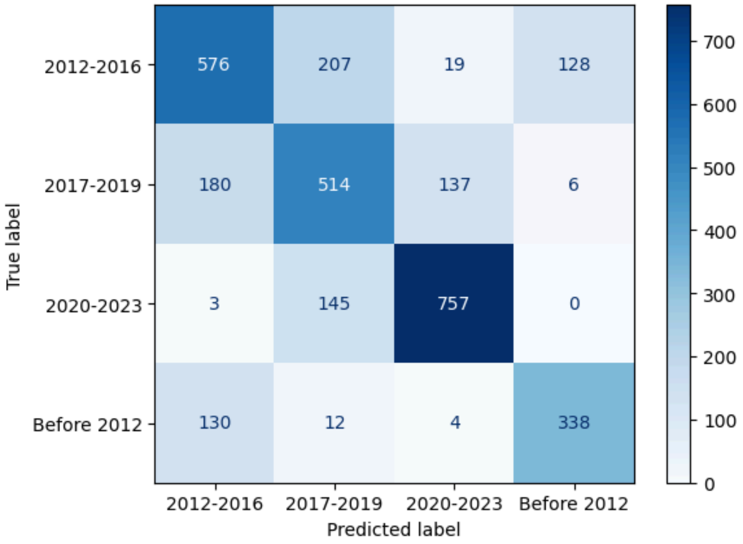
Examining the model's coefficients highlighted the main drawback of adding polynomial features, which is the interpretability of the model equation. With 170 variables, it becomes impossible to interpret each coefficient meaningfully. We attempted to obtain a summary of the model using statsmodels, however, due to the complexity and high dimensionality, the summary did not provide any p-values. Without the p-values, we could not identify which terms are significant, preventing us from simplifying the model by eliminating insignificant terms. Therefore, we decided it would be better to use a simpler Logistic Regression model without added polynomial terms if the performance is not too different.

The next model we trained was a multinomial Logistic Regression model without added polynomial terms. This approach makes the model more interpretable, though it may perform worse in capturing nonlinear relationships. After applying the model on the test set, we observed an accuracy of 69%, which is lower than Random Forest, slightly lower than the polynomial Logistic Regression, and higher than the Naive Bayes model.

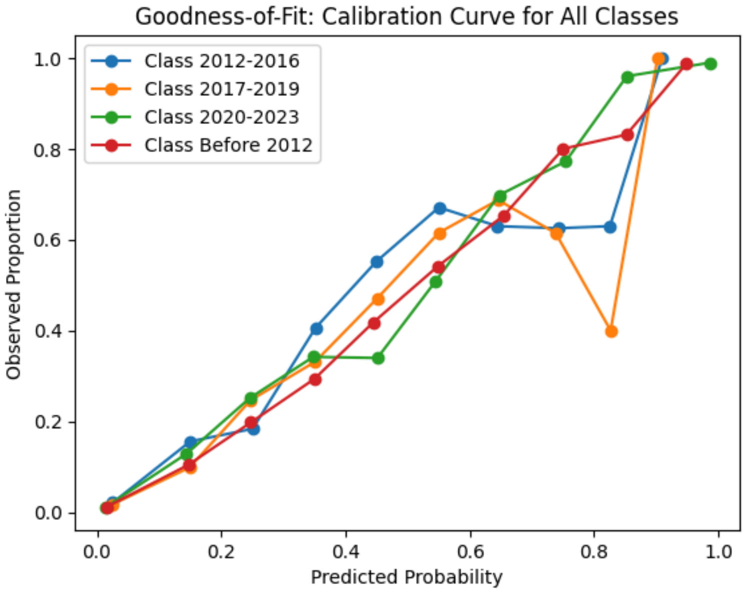


```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.62      0.63       930
           1       0.59      0.61      0.60       837
           2       0.83      0.84      0.83       905
           3       0.72      0.70      0.71       484

    accuracy                           0.69      3156
   macro avg       0.69      0.69      0.69      3156
weighted avg       0.69      0.69      0.69      3156
```

The classification report showed that the precision and recall varied by class. The model performed best in identifying 2020-2023 (class 2) vehicles with a higher precision and recall. This indicates that the newer vehicles have more distinct characteristics. Classifying intermediate categories such as 2012-2016 (class 0) and 2017-2019 (class 1) remained challenging as seen by the precision and recall. As for the F1-scores, they ranged from 0.60 to 0.83 depending on the class, with the 2020-2023 category having the highest score, highlighting the model's reliability in identifying this class.



The confusion matrix shows that the model struggled with intermediate classes 2012-2016 and 2017-2019 while performing better at identifying the extreme classes Before 2012 and 2020-2023. This suggests that the model is capable of distinguishing the oldest and most recent vehicles while it struggles for intermediate ages.



The calibration curve shows us that the model probability estimates are reasonable even if not perfect. Compared to the polynomial Linear Regression calibration curve, the 2017-2019 curve is closer to the diagonal, indicating that the model estimates the probability more accurately.

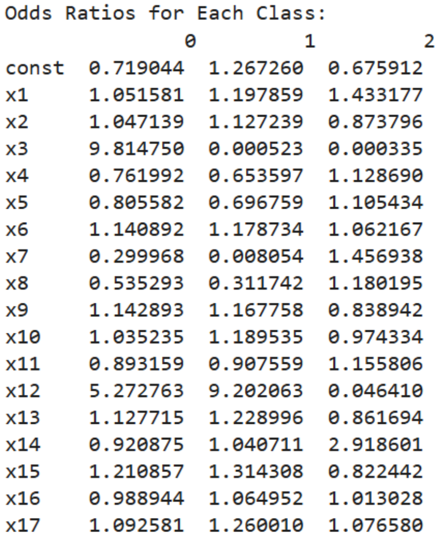
Since this is a 4-class multinomial model, it has 3 regression equations (versus the baseline class), which are shown below:

logit(P(Class=0)) = -0.3298 + 0.0503*X1 + 0.0461*X2 + 2.2839*X3 + -0.2718*X4 + -0.2162*X5 + 0.1318*X6 + -1.2041*X7 + -0.6249*X8 + 0.1336*X9 + 0.0346*X10 + -0.1130*X11 + 1.6626*X12 + 0.1202*X13 + -0.0824*X14 + 0.1913*X15 + -0.0111*X16 + 0.0885*X17

logit(P(Class=1)) = 0.2369 + 0.1805*X1 + 0.1198*X2 + -7.5552*X3 + -0.4253*X4 + -0.3613*X5 + 0.1644*X6 + -4.8216*X7 + -1.1656*X8 + 0.1551*X9 + 0.1736*X10 + -0.0970*X11 + 2.2194*X12 + 0.2062*X13 + 0.0399*X14 + 0.2733*X15 + 0.0629*X16 + 0.2311*X17

logit(P(Class=2)) = -0.3917 + 0.3599*X1 + -0.1349*X2 + -8.0025*X3 + 0.1211*X4 + 0.1002*X5 + 0.0603*X6 + 0.3763*X7 + 0.1657*X8 + -0.1756*X9 + -0.0260*X10 + 0.1448*X11 + -3.0702*X12 + -0.1489*X13 + 1.0711*X14 + -0.1955*X15 + 0.0129*X16 + 0.0738*X17

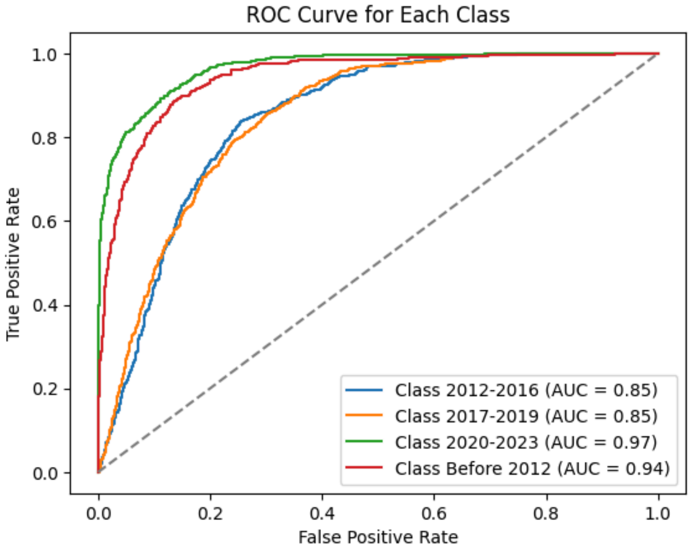
The advantage of this simpler mode is that the equation and coefficients are more easily interpretable. Using the coefficients, we calculated the odds ratios and found the following:



```
Odds Ratios for Each Class:
              0         1         2
const  0.719044  1.267260  0.675912
x1     1.051581  1.197859  1.433177
x2     1.047139  1.127239  0.873796
x3     9.814750  0.000523  0.000335
x4     0.761992  0.653597  1.128690
x5     0.805582  0.696759  1.105434
x6     1.140892  1.178734  1.062167
x7     0.299968  0.008054  1.456938
x8     0.535293  0.311742  1.180195
x9     1.142893  1.167758  0.838942
x10    1.035235  1.189535  0.974334
x11    0.893159  0.907559  1.155806
x12    5.272763  9.202063  0.046410
x13    1.127715  1.228996  0.861694
x14    0.920875  1.040711  2.918601
x15    1.210857  1.314308  0.822442
x16    0.988944  1.064952  1.013028
x17    1.092581  1.260010  1.076580
```

An odds ratio above 1 indicates that the predictor raises the odds of that class (relative to the baseline class), while an odds ratio below 1 indicates a decrease in odds for that class. These are hard to interpret due to the encoding and scaling, however, we can get a general idea of how a change in one of the features influences the likelihood of each class.

As for the regression equations, each one expresses the log-odds of belonging to that class (relative to the baseline class). Compared to the polynomial Logistic Regression equation (170 variables), this one is much simpler with 17 variables. We were able to generate a summary of the model, which included the p-values for each coefficient. This allowed us to identify which predictors lacked statistical significance, and we used this to train a simplified model, as will be discussed shortly. The model's R-squared value was found to be 0.46, indicating that this model explains only 46% of the variability.

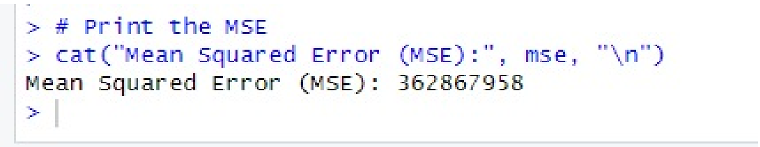


Looking at the ROC curves the AUC values were generally good, but slightly worse than the polynomial Logistic Regression model. However, since all the AUC values are above 0.7, it indicates that the model has decent discriminative ability, especially for the newest and oldest vehicles.

While evaluating this model, we found that several predictors were statistically insignificant. This led us to develop the final model, which simplified Logistic regression after dropping insignificant terms. We found the performance to be comparable after keeping only significant predictors, with the benefits of a simpler model outweighing the drawbacks in performance, which is why it was selected as our final model (mainly due to its interpretability). This model is discussed in detail in Section 4.

Section 4: Model validation

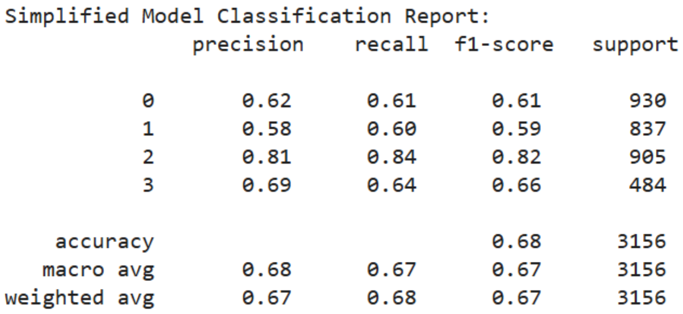
After fitting the final base multiple regression model we obtained and raised the predicted values to exponential as we had earlier applied a log transformation to the training data (Box-cox). We calculated the average of (predicted-actual prices)^2. The MSE gave us a value of 362867958. This indicates that, on average, the squared difference between the predicted and actual vehicle prices is substantial, reflecting an RMSE of approximately 19,048, which suggests significant prediction errors for higher-priced vehicles due to the wide range of prices in the dataset.



```
> # Print the MSE
> cat("Mean Squared Error (MSE):", mse, "\n")
Mean Squared Error (MSE): 362867958
> |
```

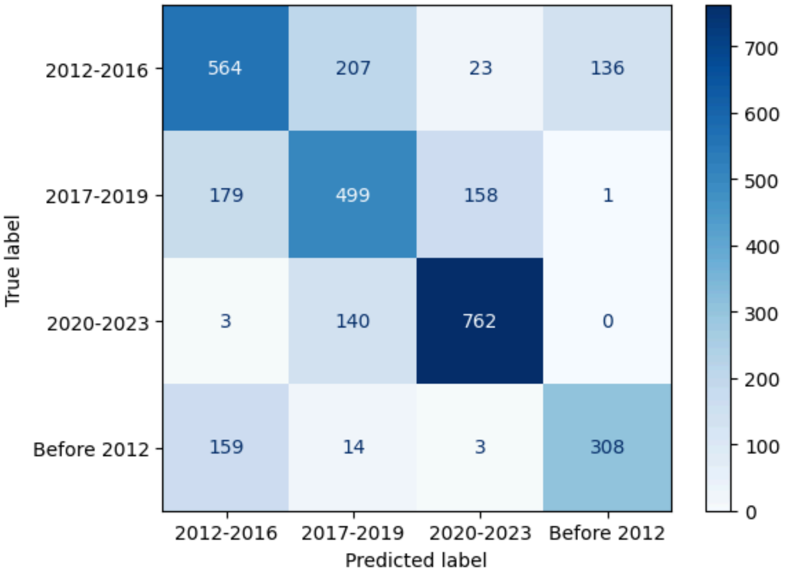
After examining several classification models, we selected the simplified Logistic Regression model for our final analysis. This model was chosen despite its lower predictive accuracy because the simplicity and interpretability provide valuable insights into the features and how they influence the probability of belonging to each YearRanges class.

Using the p-values of the coefficients (which can be found in ClassificationModels.ipynb), we identified features x4, x5, x7, x8, x9, x12, x15, and x17 as significant due to having a p-value less than 0.05. All other variables were removed due to their insignificance, leaving us with a model with the following features: Transmission, DriveType, Kilometres, CylindersinEngine, BodyType, Price, ColourExt, and State.

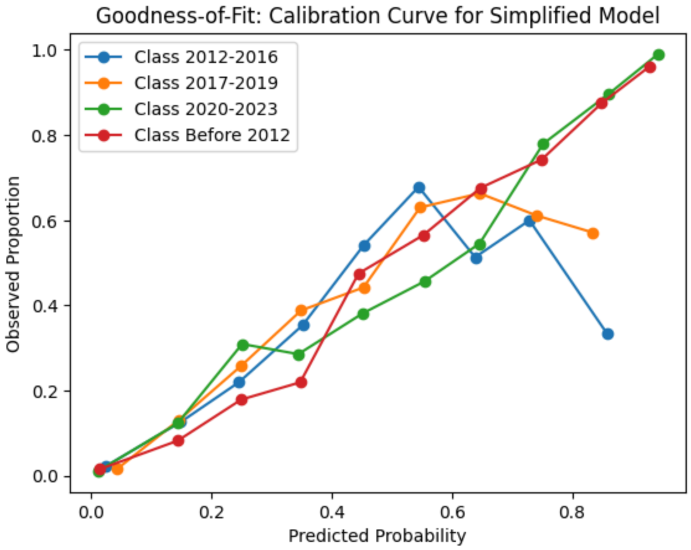


```
Simplified Model Classification Report:
              precision    recall  f1-score   support

           0       0.62      0.61      0.61       930
           1       0.58      0.60      0.59       837
           2       0.81      0.84      0.82       905
           3       0.69      0.64      0.66       484

    accuracy                           0.68      3156
   macro avg       0.68      0.67      0.67      3156
weighted avg       0.67      0.68      0.67      3156
```

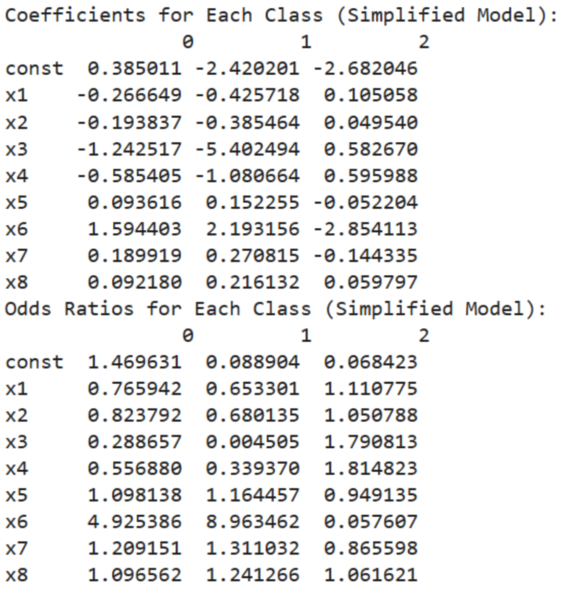
This new model had an accuracy of 68%, which is comparable to the other two logistic regression models. The classification report shows that 2020-2023 (class 2) maintained a high F1-score,indicating that the simplified model can capture the distinctiveness of newer vehicles. The intermediate classes 2012-2016 (class 0) and 2017-2019 (class 1) had moderate precision and recall, indicating a difficulty in distinguishing them. The performance of Before 2012 (class 3) was also moderate, with lower metrics than class 2. Overall, the model had a balanced performance across classes, and while not very high, it is reasonable given the focus on interpretability and simplicity.



The confusion matrix shows us that the misclassifications mostly occurred between closely related year ranges. The model found it challenging to separate between intermediate classes while it correctly identifies a substantial portion of Before 2012 and 2020-2023.



The calibration curves were not perfectly aligned with the diagonal, but they indicated that the model's probability estimates were not too far from the actual values.



```
Coefficients for Each Class (Simplified Model):
              0         1         2
const  0.385011 -2.420201 -2.682046
x1    -0.266649 -0.425718  0.105058
x2    -0.193837 -0.385464  0.049540
x3    -1.242517 -5.402494  0.582670
x4    -0.585405 -1.080664  0.595988
x5     0.093616  0.152255 -0.052204
x6     1.594403  2.193156 -2.854113
x7     0.189919  0.270815 -0.144335
x8     0.092180  0.216132  0.059797
Odds Ratios for Each Class (Simplified Model):
              0         1         2
const  1.469631  0.088904  0.068423
x1     0.765942  0.653301  1.110775
x2     0.823792  0.680135  1.050788
x3     0.288657  0.004505  1.790813
x4     0.556880  0.339370  1.814823
x5     1.098138  1.164457  0.949135
x6     4.925386  8.963462  0.057607
x7     1.209151  1.311032  0.865598
x8     1.096562  1.241266  1.061621
```

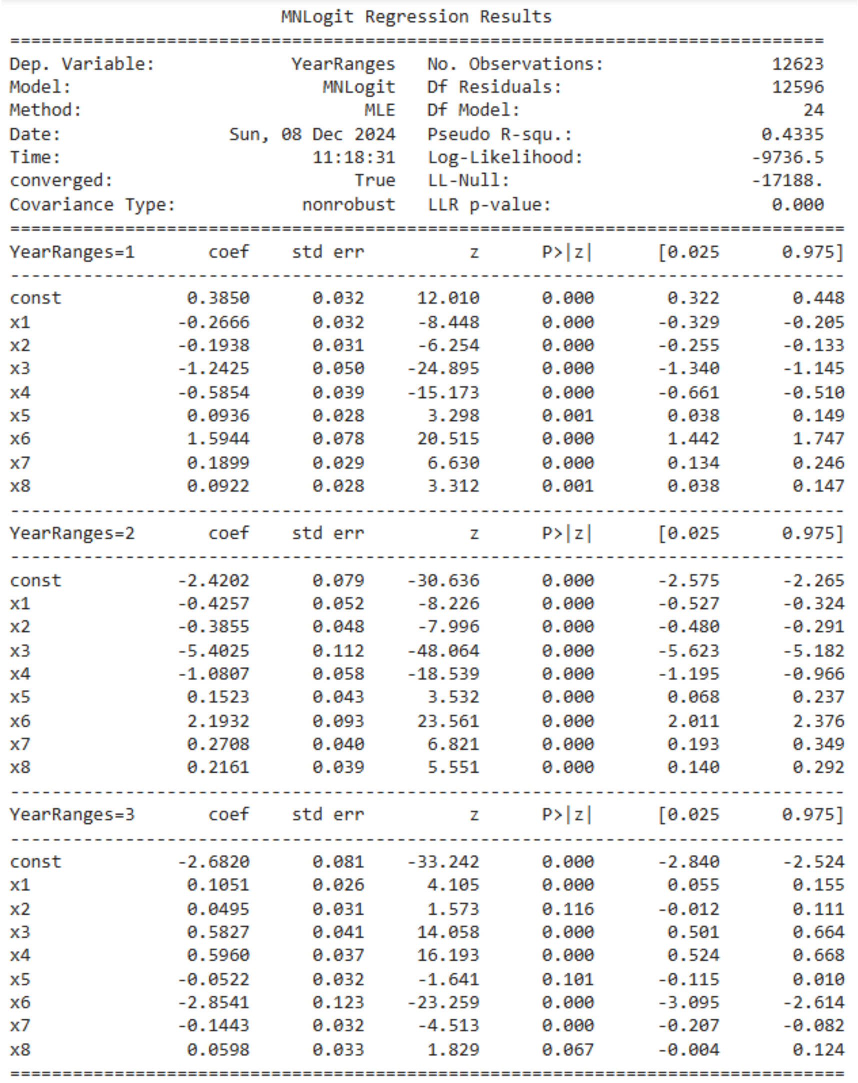
The main advantage of this simple model is the interpretability of its coefficients and odds ratios. The coefficients and odds ratios allow us to observe the effect of each variable on the probability of each class. With fewer features (8 features compared to the previous model's 17 features), we can easily identify which features are actually influential. For example, x6 significantly increases the odds of being in class 1 and it significantly decreases the odds of being in class 2.

logit(P(Class=0)) = 0.3850 + -0.2666*X1 + -0.1938*X2 + -1.2425*X3 + -0.5854*X4 + 0.0936*X5 + 1.5944*X6 + 0.1899*X7 + 0.0922*X8

logit(P(Class=1)) = -2.4202 + -0.4257*X1 + -0.3855*X2 + -5.4025*X3 + -1.0807*X4 + 0.1523*X5 + 2.1932*X6 + 0.2708*X7 + 0.2161*X8

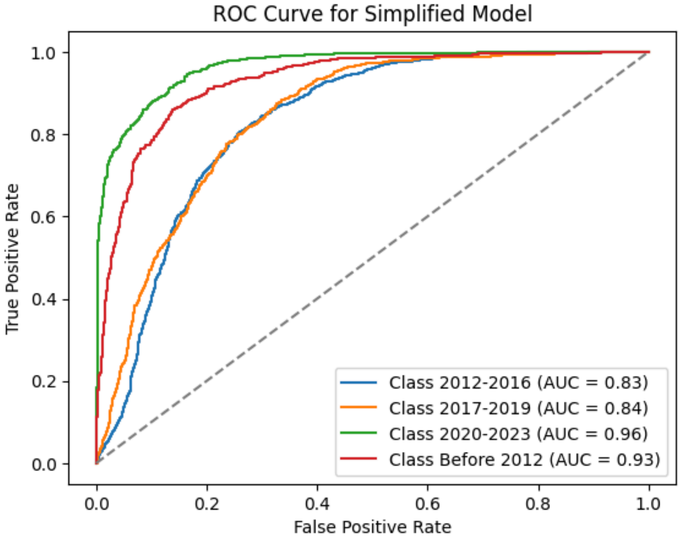
logit(P(Class=2)) = -2.6820 + 0.1051*X1 + 0.0495*X2 + 0.5827*X3 + 0.5960*X4 + -0.0522*X5 + -2.8541*X6 + -0.1443*X7 + 0.0598*X8

The simplified regression equations also help us interpret how changing a feature affects the log odds of a particular class. We can confirm that X6 contributes positively to class 1 and negatively to class 2.



```
                          MNLogit Regression Results                          
==============================================================================
Dep. Variable:             YearRanges   No. Observations:                12623
Model:                        MNLogit   Df Residuals:                    12596
Method:                           MLE   Df Model:                           24
Date:                Sun, 08 Dec 2024   Pseudo R-squ.:                  0.4335
Time:                        11:18:31   Log-Likelihood:                -9736.5
converged:                       True   LL-Null:                       -17188.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================
YearRanges=1       coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.3850      0.032     12.010      0.000       0.322       0.448
x1              -0.2666      0.032     -8.448      0.000      -0.329      -0.205
x2              -0.1938      0.031     -6.254      0.000      -0.255      -0.133
x3              -1.2425      0.050    -24.895      0.000      -1.340      -1.145
x4              -0.5854      0.039    -15.173      0.000      -0.661      -0.510
x5               0.0936      0.028      3.298      0.001       0.038       0.149
x6               1.5944      0.078     20.515      0.000       1.442       1.747
x7               0.1899      0.029      6.630      0.000       0.134       0.246
x8               0.0922      0.028      3.312      0.001       0.038       0.147
--------------------------------------------------------------------------------
YearRanges=2       coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -2.4202      0.079    -30.636      0.000      -2.575      -2.265
x1              -0.4257      0.052     -8.226      0.000      -0.527      -0.324
x2              -0.3855      0.048     -7.996      0.000      -0.480      -0.291
x3              -5.4025      0.112    -48.064      0.000      -5.623      -5.182
x4              -1.0807      0.058    -18.539      0.000      -1.195      -0.966
x5               0.1523      0.043      3.532      0.000       0.068       0.237
x6               2.1932      0.093     23.561      0.000       2.011       2.376
x7               0.2708      0.040      6.821      0.000       0.193       0.349
x8               0.2161      0.039      5.551      0.000       0.140       0.292
--------------------------------------------------------------------------------
YearRanges=3       coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -2.6820      0.081    -33.242      0.000      -2.840      -2.524
x1               0.1051      0.026      4.105      0.000       0.055       0.155
x2               0.0495      0.031      1.573      0.116      -0.012       0.111
x3               0.5827      0.041     14.058      0.000       0.501       0.664
x4               0.5960      0.037     16.193      0.000       0.524       0.668
x5              -0.0522      0.032     -1.641      0.101      -0.115       0.010
x6              -2.8541      0.123    -23.259      0.000      -3.095      -2.614
x7              -0.1443      0.032     -4.513      0.000      -0.207      -0.082
x8               0.0598      0.033      1.829      0.067      -0.004       0.124
================================================================================
```

The model summary shows that the R-squared value is 43%, which is not much lower than the previous model. This tells us that dropping the insignificant features did not majorly reduce the percentage of variance that is explained by the model. However, it indicates that there are missing features or interactions that could improve the model performance. As for the p-values, they are all statistically significant for at least two of the classes.



The ROC curves still show solid AUC values despite being slightly lower than the complex models. The 2020-2023 still has the highest AUC, indicating that even with fewer features, the model can distinguish newer vehicles from older ones with ease. Other classes have lower AUCs, but they are still in a reasonable range.

Section 5: Discussion and Conclusions

The primary goal of this project was to build predictive models that accurately estimate vehicle prices and classify manufacturing year ranges using a dataset of Australian vehicles. Throughout the project, we applied a variety of statistical tools and techniques, combining exploratory data analysis, inferential statistics, and predictive modeling. The results highlighted both the strengths and limitations of the methodologies employed.

The final multiple regression model achieved reasonable accuracy after addressing data challenges such as non-normality through Box-Cox transformations (log transformation) and careful handling of outliers using Z-score. We also dropped a few insignificant terms after careful consideration and visualized residual plots to refine the model overall. However, the relatively high MSE and the RMSE indicated that the model's performance could vary across the wide range of vehicle prices, particularly for high-priced vehicles. These challenges emphasized the importance of understanding the scale and variability of the data when evaluating model performance.

The classification models for predicting manufacturing year ranges demonstrated the utility of machine learning techniques. Among the tested models, random forests provided the highest accuracy, outperforming traditional multinomial logistic regression. This outcome underscored the value of non-linear and ensemble methods in capturing complex relationships within the data. Despite this, balancing interpretability and performance remained a key consideration, as more advanced models like random forests often lack the transparency of simpler methods.

Throughout this project, several challenges were encountered. Data preprocessing was a significant hurdle, particularly in managing missing values, handling high-cardinality categorical variables with many levels, and addressing outliers. Additionally, interpreting the results of statistical tests and models required careful attention to assumptions, such as homoscedasticity and normality, which were not always met. Another notable difficulty was the selection of the appropriate transformations and metrics for model evaluation, given the wide variability in vehicle prices and features. Furthermore, for classification, a major challenge involved dealing with an unbalanced target variable. This was remedied by binning the target variable to classes that were as close to balanced as possible while still having meaningful interpretation. These challenges highlighted the importance of iterative model refinement and critical evaluation of results.

This project provided invaluable insights into the application of statistical methods to real-world data. It reinforced the importance of exploratory data analysis in uncovering key relationships and informed the use of inferential statistics to validate findings. Moreover, the process of model building demonstrated the trade-offs between simplicity and predictive accuracy, emphasizing the need to tailor approaches to the specific context of the data and research objectives. Overall, this research project has deepened our understanding of statistical modeling and analytics, equipping us with the skills to address similar challenges in future endeavors.

All codes are attached with the submission