Analysis the factors influence the price of used Toyota cars in UK

Introduction

Navigating and negotiating prices in the second-hand car market is often more difficult than in the first-hand car market, where pricing is typically more transparent. Therefore, in this report, I will analyze the factors that impact the pricing of second-hand Toyota cars in UK. This analysis can benefit both buyers and sellers, as buyers can make more informed decisions and obtain reasonable prices, while sellers may consider maintaining their cars well if they plan to sell them at a higher price in the future.

Numerous studies have been conducted on the used car market in recent years. For example, Ozgur C explored seven explanatory variables and used SAS software to develop various linear regression models, concluding that mileage and liter were the most significant factors for predicting the price. Opera C predicted the price of a car using linear regression and identified the most relevant factors that can affect the price of the car. These two studies had similar goals to this report, with a slight difference in the chosen dataset.
In contrast, Chen C collected over 100,000 used car records throughout China and conducted empirical analysis on a thorough comparison of two algorithms: linear regression and random forest. He concluded that the effect of the linear regression model becomes better with the increase of sample size. He focused more on different models' performance on different datasets, whereas in this report, we only focus on multiple linear regression.

Methods

The data utilized in this report was obtained from Exchange and Mart and has been preprocessed by Aditya. The dataset comprises various attributes such as price, transmission, mileage, model, year, fuel type, road tax, miles per gallon (mpg), and engine size.
During the data preprocessing step, categorical variables, namely transmission, fuel type, and model, were translated into numeric variables for ease of use. To identify any influential observations that could have a large impact on the model, I utilized cook's distance to find leverage points. If bad leverage points are detected, they will be deleted. To ensure the assumptions for multiple linear regression such as linearity, independence, homoscedasticity, and normality were met, I performed exploratory data analysis using simple linear regression, correlation , residual plot, and Q-Q plot. Then, I used VIF to check multicollinearity on all predictors and set the cut-off be 5. If multicollinearity is detected, I will remove at least one of these predictors from the model.

The primary statistical technique employed to reach the research question's conclusion was multiple linear regression. To determine which variables to use in the model, I evaluated the selections made by AIC, BIC, and lasso independently. The data was subsequently divided into

ten sections, and I utilized cross-validation to forecast the efficiency of each approach. The ultimate model was created using the method that performed the best on the test set, and the conclusion about the most significant factors that impact a second-hand car's price was determined based on the model's chosen variables.

Results

For categorical variables in the dataset, we can visualize the features by summary table (Figure 1), which shows the differences between prices within each subgroup.

Summary table for 3 categorical variables in the dataset

FuelType	Price.mean	Price.sd	count
Diesel	15697.807	10208.564	503
Hybrid	17185.473	5485.107	2043
Other	14121.162	8010.610	105
Petrol	9759.538	4134.284	4087

transmission	Price.mean	Price.sd	count
Automatic	16582.829	6331.004	2657
Manual	9551.497	3623.472	3826
Other	12795	NA	1
Semi - Auto	14797.138	11872.082	254

model	Price.mean	Price.sd	count
Auris	12507.912	3094.1599	712
Avensis	9884.357	3376.0616	115
Aygo	7905.415	1662.7941	1961
C-HR	20651.541	3470.3608	479
Camry	26910.091	1434.0645	11
Corolla	20942.734	4791.3148	267
GT86	19908.849	5459.3619	73
Hilux	21504.593	5699.0426	86
IQ	4247.250	1286.0205	8
Land Cruiser	36487.157	13103.3875	51
PROACE VERSO	28680.200	6307.1038	15
Prius	18998.845	4696.9918	232
RAV4	18161.059	6577.1932	473
Supra	50741.000	3294.2781	12
Urban Cruiser	4617.500	476.1215	4
Verso	12169.158	3025.6527	114
Verso-S	5746.667	1273.1883	3
Yaris	10553.084	2570.6885	2122

Figure 1

During the data preprocessing stage, Cook's Distance identified 238 leverage points out of 4719 samples. None of these leverage points were considered to be bad, and removing them would have led to bias in our model. Therefore, we decided to retain all leverage points.

Next, in the exploratory data analysis (EDA) step, I examined the linearity of the dataset by fitting simple linear regressions. This was necessary because it is difficult to determine linearity using multiple linear regression visualization. Additionally, the correlation map was used to visualize the correlation between predictors.

Scatter plot of each predictor vs price

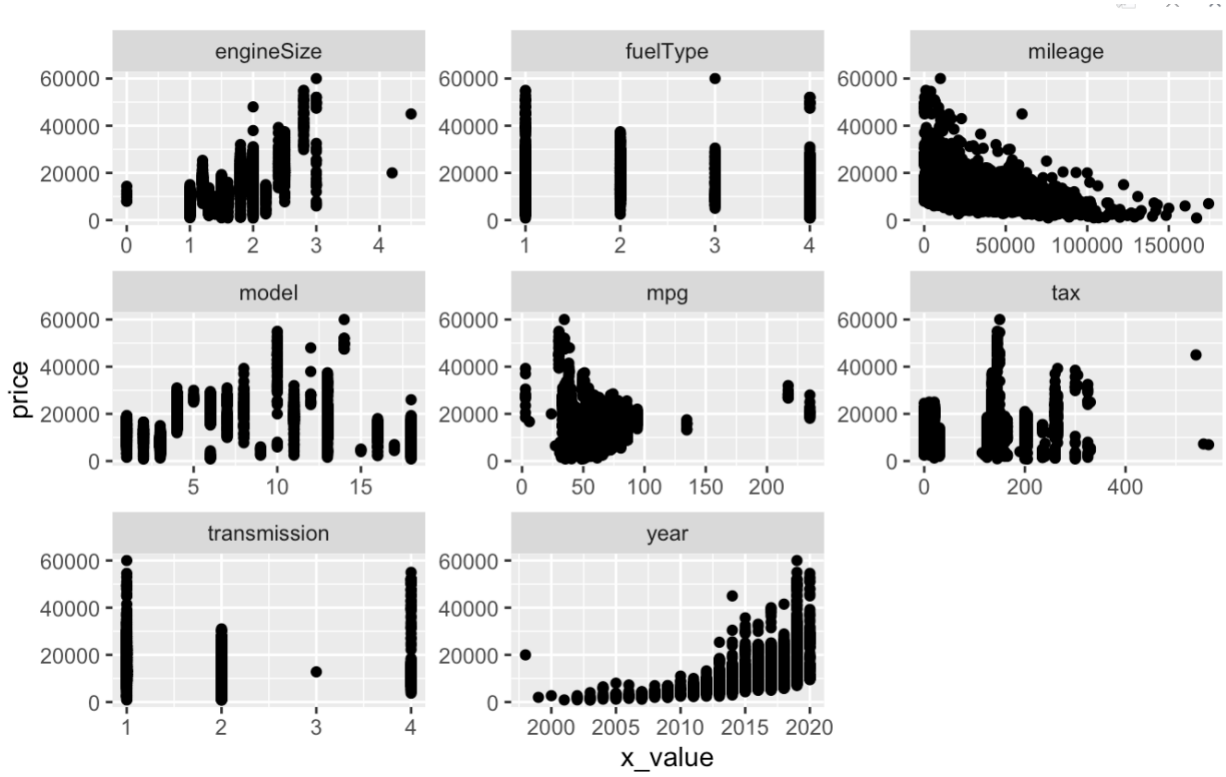


Figure 2

To satisfy the homoscedasticity assumption, it is necessary for the residual value to be independent of the x value. In Figure 3, the residual plot for each predictor shows no distinct pattern for the residual value, indicating that the equal variance property is maintained.

Additionally, we utilized Q-Q plots to verify the normality of each predictor. Although normality is not essential for linear regression, we chose to maintain this property for ease of calculation. The majority of points on the 45-degree line indicate that the predictors are normally distributed.

Taken together, the conclusions drawn from the EDA demonstrate that the assumptions of multiple linear regression have been met, indicating that it will be a suitable model.

Residual plot for each predictor

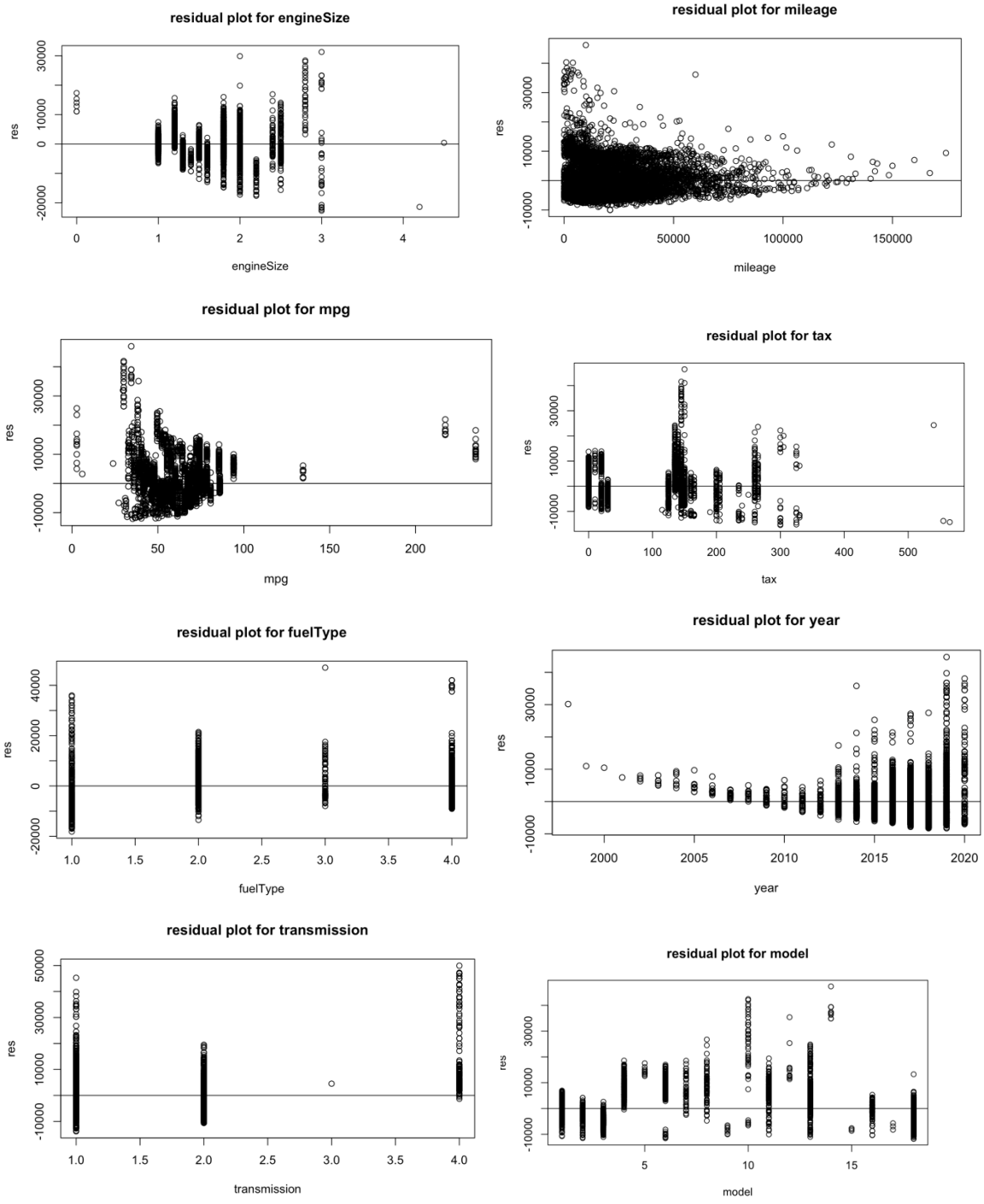


Figure 3

The presence of multicollinearity can cause inaccuracies in multiple linear regression. To check for this issue in our model, we utilized the variance inflation factor (VIF) and set a cutoff value of 5. Predictors with high VIF values are considered to exhibit multicollinearity. In Figure 4, we found that none of the predictors exceeded the cutoff value, indicating that multicollinearity is not an issue in our model.

VIF for each predictor

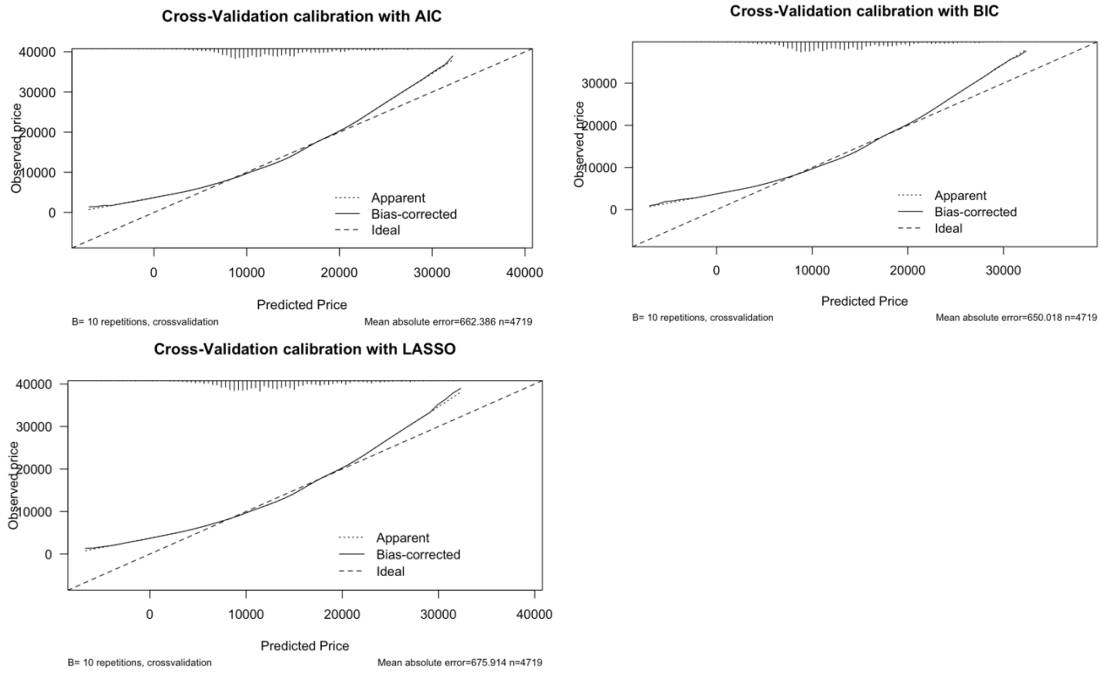
model	year	transmission	mileage	fuelType	tax
1.121967	2.265529	1.508867	2.318190	3.248213	1.417497

mpg	engineSize
1.804856	3.170478

Figure 4

Finally, in order to select variables for the model, we employed the AIC, BIC, and Lasso methods and evaluated their performance on the test set. To obtain a more accurate evaluation of different models, we also used cross-validation and split the dataset into 10 groups. AIC selected all variables except for fuel Type, while BIC did not select mpg when compared to AIC. The Lasso method did not select tax and mpg. AIC performed the best on both the training and test sets, indicating that the most important factors influencing the price of a used Toyota car in the UK are model, transmission, mileage, year, mpg, tax, and engineSize. By comparing the mean square error on training and test set, we can see the model is slightly overfitting.

Model selection with 3 methods



Model selection method	Prediction error on training set	Prediction error on test set
AIC	8609463	9241610
BIC	8614303	9275259
Lasso	8622912	9293036

Figure 5

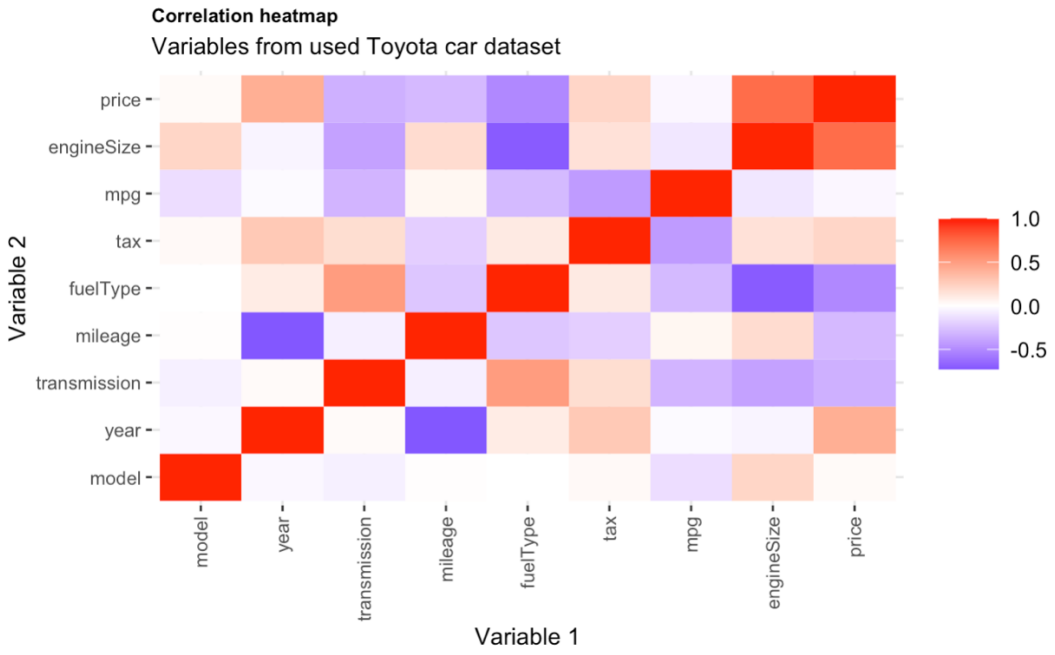
Discussion

The findings indicate that despite high errors in all models, there is still a discernible pattern in how each predictor affects the price. For instance, the coefficient for mileage is -0.0822, implying that for every one unit increase in mileage, the estimated price decreases by 0.0822 units. Additionally, tax and mileage have a negative correlation with the price, while mpg and engine size have a positive correlation. Furthermore, automatic and hybrid cars are likely to fetch higher prices in the market. These results align with our expectations that newer and better-conditioned cars are more expensive. Therefore, the research question can be answered using the fitted model.

It is a common occurrence in the used car market for two cars in identical conditions to be sold at different prices, owing to the lack of transparency in the second-hand market. To improve the accuracy of future research, additional factors such as the seller's type (e.g., used car platform or owner) could be incorporated into the dataset.

Appendix

Correlation heatmap



Q-Q plot

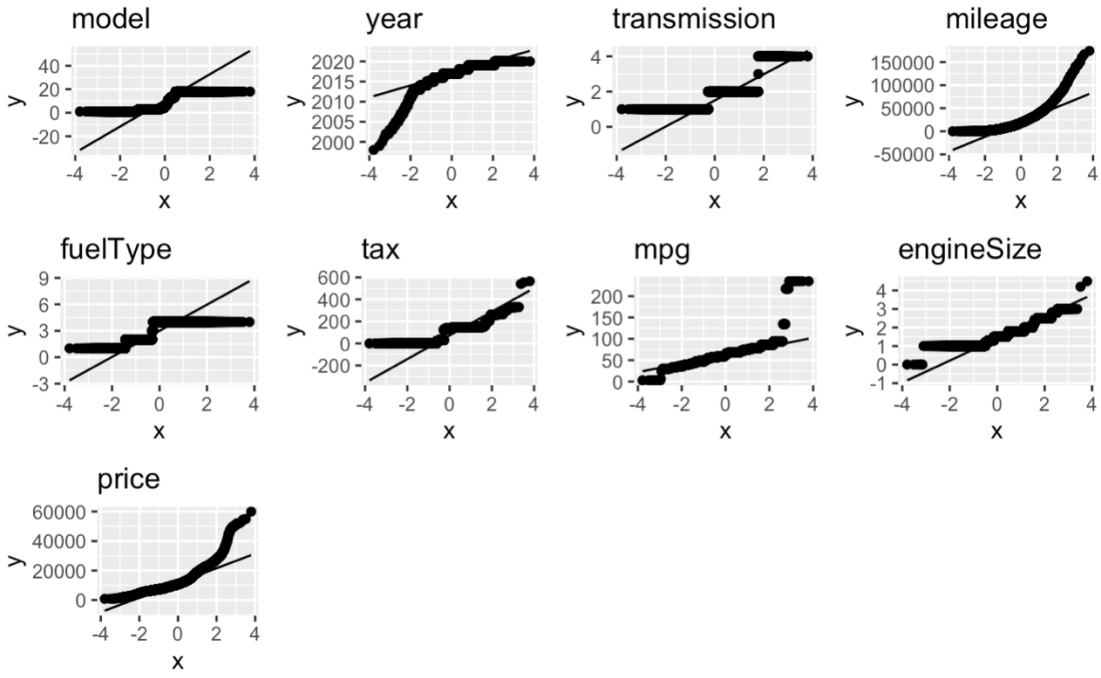


Figure 2

Lasso method

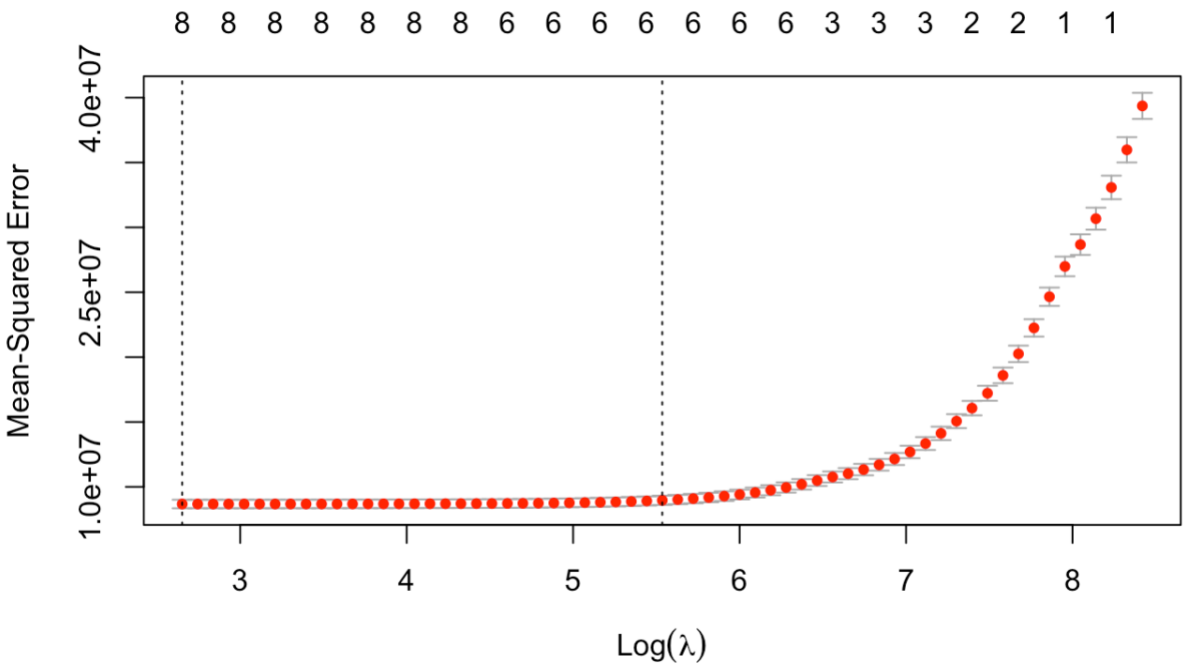


Figure 3

References

Aditya. (2020, July 4). 100,000 UK used car data set. Kaggle. Retrieved March 18, 2023, from https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes

Chen, C., Hao, L., & Xu, C. (2017). Comparative analysis of used car price evaluation models. AIP Conference Proceedings. https://doi.org/10.1063/1.4982530

OPREA, C. (2010). MAKING THE DECISION ON BUYING SECOND-HAND CAR MARKET USING DATA MINING TECHNIQUES. Oprea. Retrieved March 18, 2023, from http://www.annals.seap.usv.ro/index.php/annals/article/view/317/326

Ozgur, C., Hughes, Z., Rogers, G., & Parveen, S. (2016, August). Multiple linear regression applications automobile pricing - Ijmsi.org. Retrieved March 18, 2023, from https://www.ijmsi.org/Papers/Volume.4.Issue.6/A040601010.pdf