Predicting Used Car Prices-- Statistical Modeling Approach for Business Decision Support

Yuerong Shu, Ziyi Zhang & Ruiyan Tang

Introduction and background:

For used car businesses to achieve maximum profitability and competitive positioning they require precise assessments of each vehicle's market value. (Sokoya, 2018) For traditional used car dealers, they often assess the value of a car by their experience and intuition. However, a data-driven approach can provide a better and precise estimation for all used cars in the market. Rohan who owns a used car business, wants to develop a data-driven strategy to identify important features that affect used car sales prices. He asks for our help to create a predictive model using data from 90 vehicles to support pricing decisions and enhance business outcomes. The dataset provided features multiple vehicle specifications and our project aims to achieve two main research goals:

Our main goal focuses on what vehicle features like mileage, engine size, brand and help determine how much a car sells for. It explores how each factor influences the final sale price.

The second goal is to understand how these features can help create a strong prediction model by showing which variables are most important. This makes it easier to estimate car prices accurately based on their characteristics.

A comprehensive statistical analysis was performed to address these questions. We divided the dataset into two parts with 72 observations for developing the model and 18 observations for testing predictive accuracy through random splitting.

This report presents the data exploration process and predictor identification steps followed by model development and validation and concludes with key findings and recommendations. To understand the relationship between quantitative explanatory variables, we examined scatterplots between Price and several quantitative variables using the training set. Notice that all prices are scaled in 100,000s for readability.

Graph 1B: The graph overall shows no strong or clear linear relationship between mileage per gallon (fuel efficiency) and price. We can tell that there are cheap cars regardless of mileage level, and even some cars with low mpg (fuel efficiency) have high prices. We can separate the data into two clusters, Cluster 1: Mileage < 18 km/l and Cluster 2: Mileage $\geq$ 18 km/l. In cluster 1, the points are more vertically spread out, they prices from 5 to 40. We cannot see a clear linearity. In cluster 2, most prices are concentrated below 15, there is a slightly negative linear relationship that higher mileage cars tend to be cheaper. The observed points are sparser in this cluster.

Graph 1D: This graph shows a clear positive relationship,, however, an exponential trend appears, as power increases, price increases at a faster rate. Higher power output is generally associated with higher prices. The data might suggest a strong correlation, with a few outliers among high power cars.

We can tell from Graph 1E that only the distribution of mileage is nearly normal, the rest of the distributions of the numeric variables are right-skewed, suggesting most cars have relatively low mileage, low engine size and low power. This distribution indicates that most cars in the market are economic cars, while there are a small portion of luxury cars or sport cars.

To understand these categorical variables, we constructed six box plots. We summarized these categorical variables into two groups: Important categorical predictors and moderate predictors. Important categorical predictors show clear separation in price distribution with distinct medians and ranges across categories, thus these important predictors suggest a strong influence on car pricing. In contrast, moderate predictors show more overlap and less distinct medians in box plots, suggesting a more moderate effect on price.

In graph 2E, Luxury brands (Audi, BMW, Mercedes) have high median means prices(around 25-30) and wider interquartile ranges. Among mid-range brands(Ford, Toyota, Volkswagen), Ford has a wider spread, the price range form around 5 to 30. Toyota and VW are more consistent, they have moderate medians and less variation compared to Ford. Cheap brands(Honda, Hyundai, Mahindra, Maruti, Tata) cluster at the lower part of the price scale between 5-10. Maruti and Mahindra have tight IQRs, suggesting these brands only manufacture budget cars.

In graph 2F, we can tell that Automatic cars tend to sell for significantly more expensive than manual cars, with a higher median and wider spread.

Statistical Methods and Analysis: After summarizing and analyzing observation toward the variables of observations, using the 72-object training dataset, three variable selection approaches were implemented to select statistically significant variables for building an effective and efficient prediction model. Introduction of what methods we used and what model we decide to adapt was attached below. Before applying selection, two important changes were made:

The Year variable was revised to Age variable using formula 2025-Year=Age, this variable describes the age of the car object.

For Model variable, we can see clearly that there is a big difference between the prices of models of cars in graph2G. However, since the train dataset only consists 72 objects, most models only appear once, which does not provide reliable data distribution range to analyze and conclude. In another word, the distribution of model's corresponding price could be very extreme or non-representative. Another observation is that through checking dataset, the model seems to be unique for each brand: no different brands share same model in the dataset. Statement in brand official site proves the guess. (reference1) Because of these, we consider Model to be a risky variable that is very related to variable brand. Including the Model variable into variable selection could cause model produce unreliable and unrepresentative prediction. Therefore, Model variable in dataset is not used in all selection.

As shown in graph3A, Akaike's Information Criteria I(AIC) stepwise selection method yields Power, Brand, Fuel_Type, Mileage, Transmission and Kilometers_Driven as important predictors. Compared to BIC selection method, AIC generates model with more precise results.

As shown in graph3B, Bayesian's Information Criteria (BIC) stepwise selection method yields Power, Brand, Fuel_Type, Mileage, and Transmission as important predictors. Compared to model given by AIC stepwise selection, the variable Kilometers_Driven is dropped. Considering BIC penalizes complex model more heavily than AIC, it is very possible that Kilmeters_Driven is not selected because the improvement the variable brings in terms of increasing precision does not counteract the negative effect of increasing complexity of the model.

With Best Subset approach, we only consider the top one or best models among k-predictor variable models since there were too many predictors for R to check. (See results on graph3C) We applied all selection criteria available for the best subset approach, including model's value of R2 (graph3D1), adjusted R2 (graph3D2) and AIC (graph3D3). Unsurprisingly, we see the model that contains the most variable has lowest AIC and highest value of R2 and adjusted R2, however, one thing to notice is that the fourth best model on all values containing BrandBMW, BrandMercedes, Transmission manual, and Seats also obtain a pretty good indicator value with only 4 variables. Considering the model containing 4 variables loose little criteria value compared to the 8 predictor variables, we decided to take this model as our best subset model.

For different selection approaches, the decision for the final model depends on testing accuracy for the models by using the residual and the histogram with AIC, BIC and best subset. Based on the histogram without logarithms, all approaches have a skewed distribution (Graph AIC), (Graph BIC), (Graph sub). However, by transforming the predictor power, the distribution improved more for the best subset selection method (Graph sub_log). Compared with AIC and BIC, both logged power histogram still shows some degree of skewness to the right (Graph AIC_log), (Graph BIC_log). The centers of residual distributions also leave 0 and move to the right. Therefore, the final decision chosen will be the logged power model in the best subset approach.

Conclusion: Based on above statement, the selected model is:

$$\hat{Price} = -5151665 + 1331982 \cdot log(Power) + 103876 \cdot BrandBMW - 779089 \cdot BrandFord - 870159 \cdot BrandHonda - 1054686 \cdot BrandHyundai - 989824 \cdot BrandMahindra - 821630 \cdot BrandMaruti + 126517 \cdot BrandMercedes - 783899 \cdot BrandTata - 748247 \cdot BrandToyota - 697566 \cdot BrandVolkswagen - 174881 \cdot Transmission:Manual + 104265 \cdot Seat$$

For how to decide value of categorical variables, see table1 and table2 as interpretation:

As graph 2E and selection results(graph3D1, graph3D2 and graph3D3) show, variable Brand plays a significant role in estimating the price of a vehicle as a factor with grouped levels or dummy variables. In the subset method, we can see all 3 selection criteria indicate that only BrandBMW and BrandMercedes

are considered important variables. However, we decided to put all brands into the prediction model. We made this decision based on two ideas: First, there is a large difference between different brand price's variance and average. Only including two brand variables will cause large prediction errors for cars with brands other than BMW and Mercedes. Second, acquiring data for brand of car does not cost time or extra expenditure, in this case, adding other brand variables would improve precision without harming efficiency or cost.

For checking validity assumption, from Graph sub_v based on the residual plot, all points are scattered evenly between the zero horizontal line and no sign of violation for linearity. Since we are using random seed codes for the datasets, we are randomly selecting the samples from the datasets which guarantee independence. For the constant variables, the graph has a slightly megaphone shape from the standardized residuals plot suggesting the error variance is increasing. For normality, from the QQ plot, we can see that error distribution has thicker and longer tails than the normal distribution.

Discussion:

Limitation: There are two limitations that appear when building the model: having small datasets of 18 samples lead to model selection hard and the data transformation when selecting the model.

The dataset itself has a problem of small amounts of data included which affect the test set when the testing for accuracy. After picking predictors using 3 selection methods, we use residual plot and histogram which use actual value from the test set subtract the predicted value to check for precision. Due to the lack of data, all histograms with different selection methods show subtle differences(Graph AIC), (Graph BIC), (Graph sub), (Graph AIC_log_both), (Graph BIC_log_both), (Graph sub_log_both), (Graph AIC_log), (Graph BIC_log), (Graph sub_log) which makes the model decision hard.

After plotting scatterplots with price and numeric variables, two scatterplots from the output which are the with price versus engine (Graph1C) and price versus power (Graph1D) show clear nonlinearity and curvature. However, using AIC, BIC and best subset selections for the predictors, we decided not to include engine as a predictor. So, we transform the scatterplot with price and power by logarithm for further consideration.

With only power transformed, the graph still shows some nonlinearity (Graph 1D_log_power), but it is much improved. While with both price and power logged (Graph 1D_log_both), the graph shows a strong positive relationship. However, by using the data in test set, the precision for all the selection methods shown in the residual plot and histogram for logging both price and power have a strong right skewed shape and the center deviates from the zeros (Graph AIC_log_both), (Graph BIC_log_both), (Graph sub_log_both) more than only logging the variable power (Graph AIC_log), (Graph BIC_log), (Graph sub_log). Hence, the final decision for the predictor used in the model will be logged power instead of logged price and power.

Future directions: Considering 72 objects is a relatively small number for building a model, in future, if the company wants to improve the model by providing more datasets and objects should be collected for more precise predictions. As the sample size is large enough, the predictor Model in the dataset can also be taken into consideration and the testing data will also be large enough to help the model to improve.

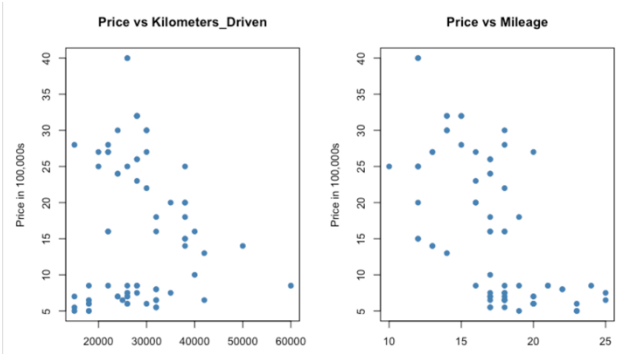
Table1:

Example car brand	BrandBMW	BrandFord	BrandHonda	BrandMercedes	BrandTata	BrandHyundai
BMW	1	0	0	0	0	0
Ford	0	1	0	0	0	0
Tata	0	0	0	0	1	0

And so on...

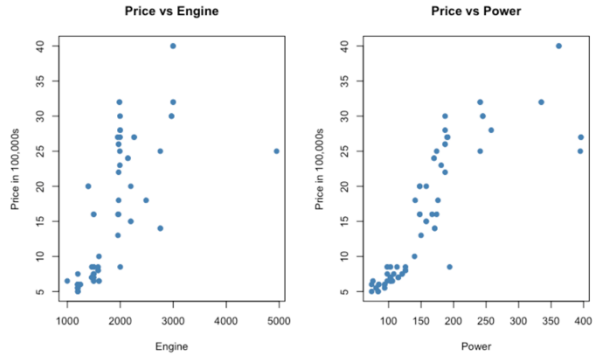
Table2:

Example car transmission type	Transmission:Manual
Manual	1
Automatic	0



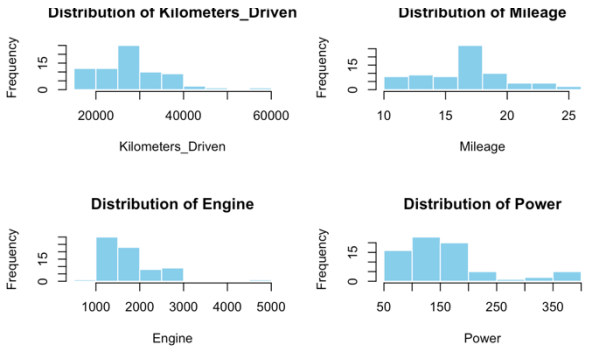
Graph 1A

Graph 1B

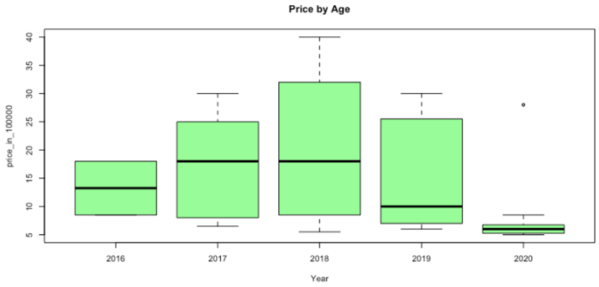


Graph 1C

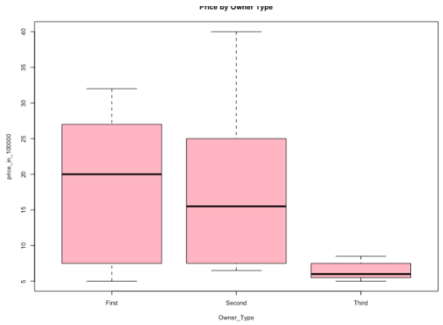
Graph 1D



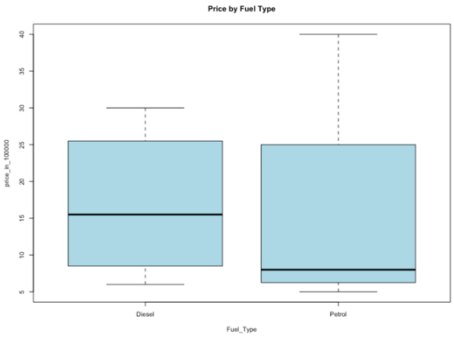
Graph 1E



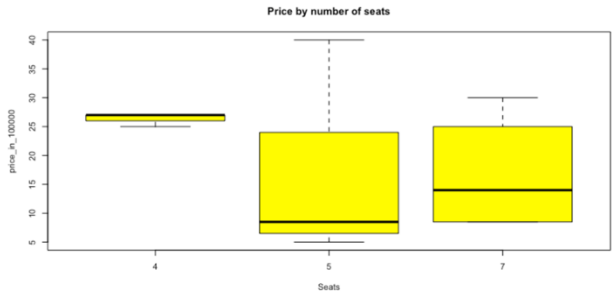
Graph 2A



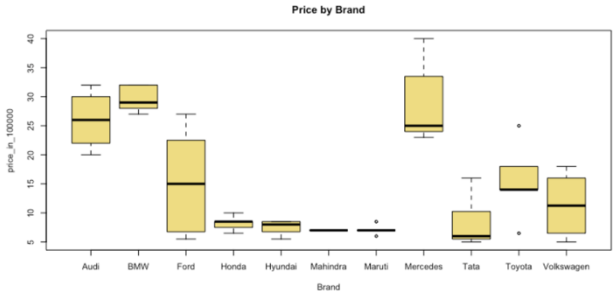
Graph 2B



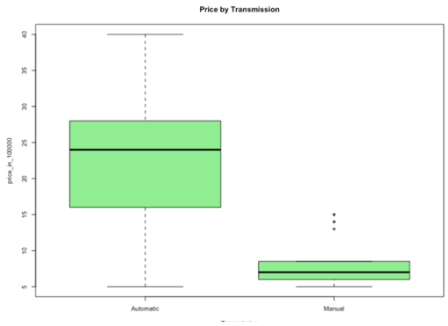
Graph 2C



Graph 2D

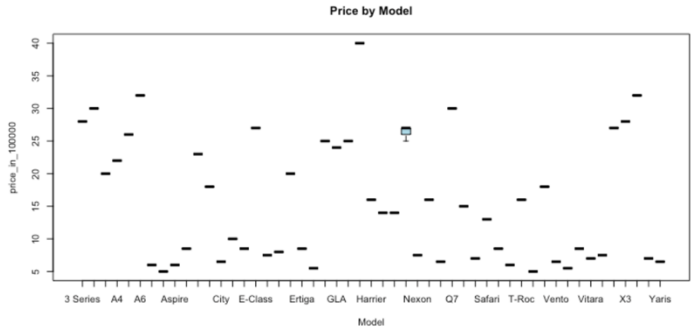


Graph 2E

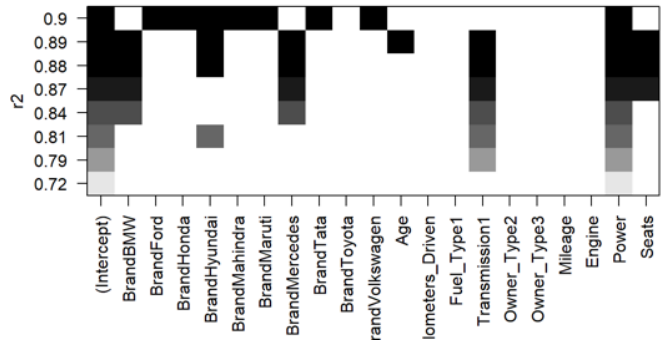


graph2G

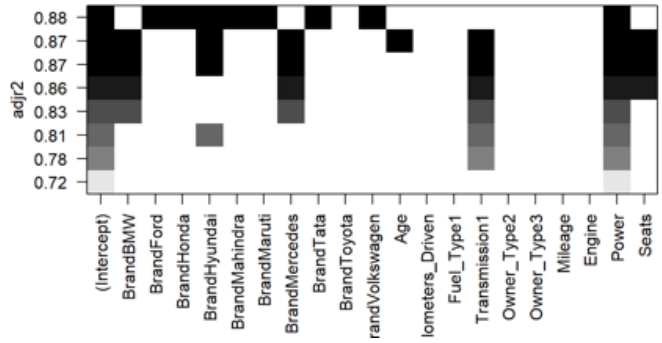
Graph 2F



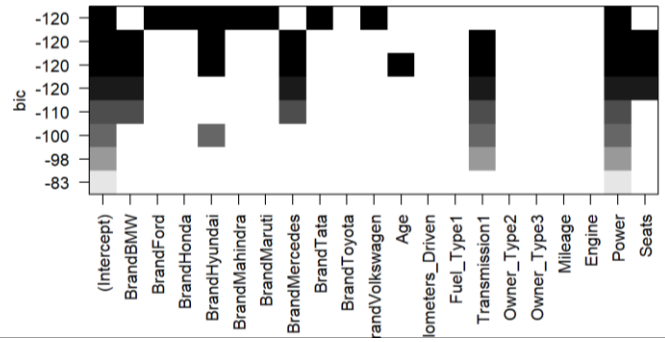
graph3D1



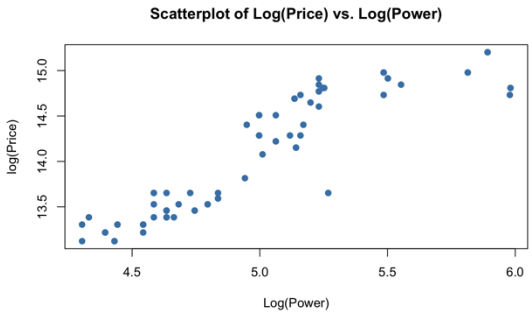
Graph3D2



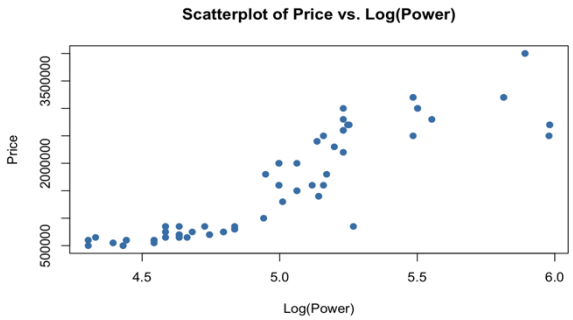
Graph3D3



Graph 1D_log_both

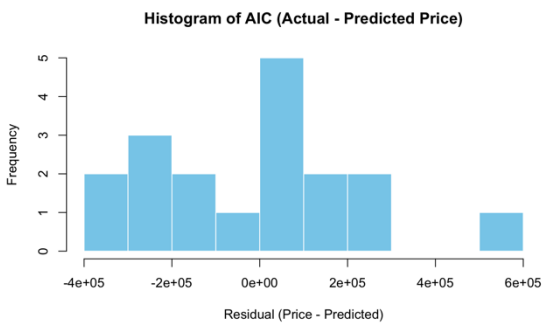


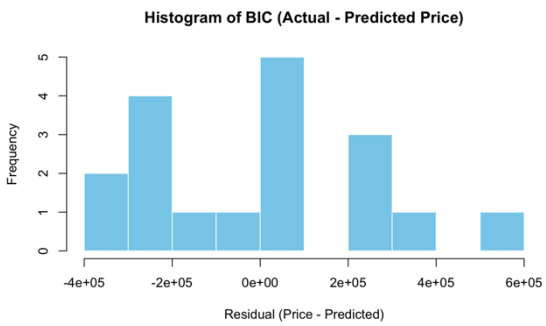
Graph 1D_log_power



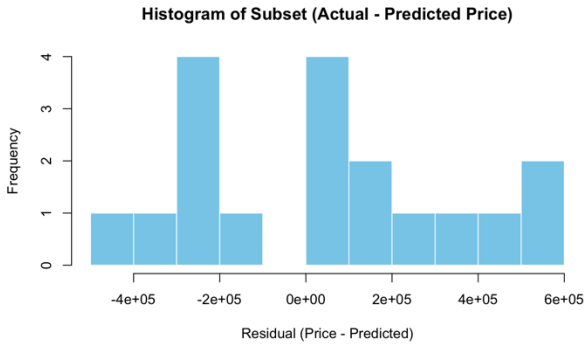
Graph AIC

Graph BIC

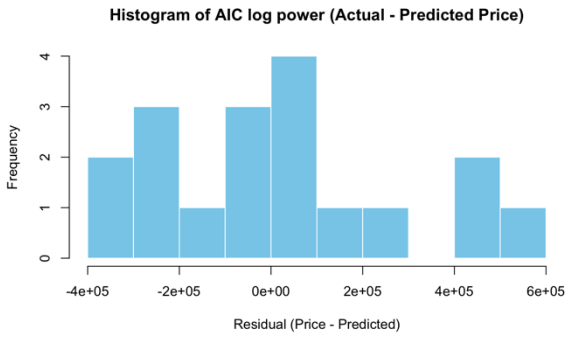




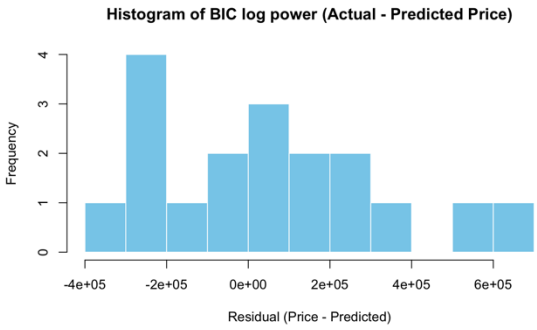
Graph sub



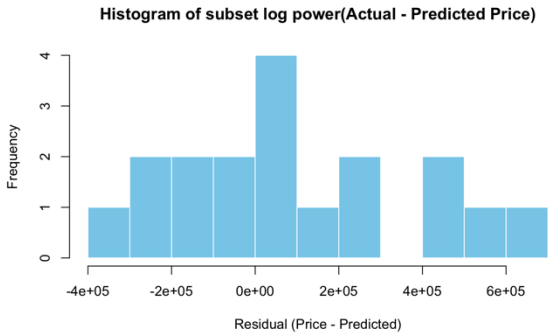
Graph AIC_log



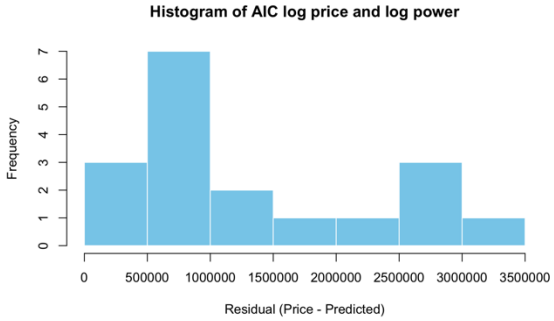
GraphBIC_log



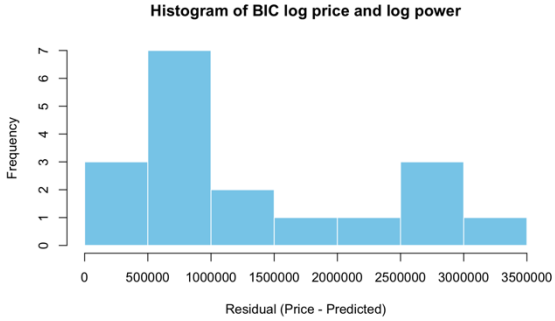
Graph sub_log



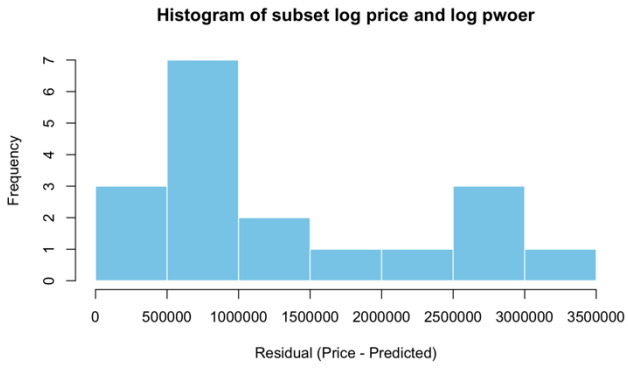
Graph AIC_log_both



Graph BIC_log_both



Graph sub_log_both



graph3A

```
Start:  AIC=1990.5
Price ~ 1
```

```
                    Df  Sum of Sq        RSS    AIC
+ Power              1 5.1251e+13 1.9823e+13 1900.6
+ Brand             10 5.2312e+13 1.8762e+13 1914.6
+ Engine             1 3.5792e+13 3.5282e+13 1942.1
+ Transmission       1 3.5181e+13 3.5892e+13 1943.3
+ Mileage            1 2.7868e+13 4.3206e+13 1956.7
+ Owner_Type         2 1.0999e+13 6.0075e+13 1982.4
+ Age                1 5.5270e+12 6.5547e+13 1986.7
<none>                            7.1074e+13 1990.5
+ Fuel_Type          1 6.2500e+11 7.0449e+13 1991.9
+ Seats              1 1.3638e+10 7.1060e+13 1992.5
+ Kilometers_Driven  1 6.0844e+09 7.1068e+13 1992.5


Step:  AIC=1900.57
Price ~ Power

                    Df  Sum of Sq        RSS    AIC
+ Brand             10 1.4829e+13 4.9940e+12 1821.3
+ Transmission       1 4.6105e+12 1.5213e+13 1883.5
+ Fuel_Type          1 1.7123e+12 1.8111e+13 1896.1
+ Seats              1 1.2818e+12 1.8541e+13 1897.8
<none>                            1.9823e+13 1900.6
+ Owner_Type         2 9.7582e+11 1.8847e+13 1900.9
+ Mileage            1 3.9791e+11 1.9425e+13 1901.1
+ Age                1 1.4236e+11 1.9681e+13 1902.0
+ Engine             1 9.9061e+10 1.9724e+13 1902.2
+ Kilometers_Driven  1 2.7399e+09 1.9820e+13 1902.6
- Power              1 5.1251e+13 7.1074e+13 1990.5


Step:  AIC=1821.31
Price ~ Power + Brand

                    Df  Sum of Sq        RSS    AIC
+ Fuel_Type          1 9.4813e+11 4.0459e+12 1808.2
+ Seats              1 6.1340e+11 4.3806e+12 1813.9
+ Transmission       1 4.6484e+11 4.5292e+12 1816.3
+ Mileage            1 2.4633e+11 4.7477e+12 1819.7
+ Engine             1 1.5615e+11 4.8379e+12 1821.0
<none>                            4.9940e+12 1821.3
+ Kilometers_Driven  1 7.5671e+10 4.9183e+12 1822.2
+ Age                1 3.0485e+10 4.9635e+12 1822.9
+ Owner_Type         2 1.1992e+11 4.8741e+12 1823.6
- Brand             10 1.4829e+13 1.9823e+13 1900.6
- Power              1 1.3768e+13 1.8762e+13 1914.6


Step:  AIC=1808.15
Price ~ Power + Brand + Fuel_Type

                    Df  Sum of Sq        RSS    AIC
+ Mileage            1 2.6867e+11 3.7772e+12 1805.2
+ Transmission       1 2.3113e+11 3.8148e+12 1805.9
+ Seats              1 1.9533e+11 3.8505e+12 1806.6
<none>                            4.0459e+12 1808.2
+ Age                1 3.4376e+10 4.0115e+12 1809.5
+ Kilometers_Driven  1 1.5652e+10 4.0302e+12 1809.9
+ Engine             1 1.2654e+09 4.0446e+12 1810.1
+ Owner_Type         2 3.6448e+10 4.0094e+12 1811.5
- Fuel_Type          1 9.4813e+11 4.9940e+12 1821.3
```

```
- Brand             10 1.4065e+13 1.8111e+13 1896.1
- Power              1 1.4380e+13 1.8426e+13 1915.3


Step:  AIC=1805.2
Price ~ Power + Brand + Fuel_Type + Mileage


                    Df  Sum of Sq        RSS    AIC
+ Transmission       1 3.2901e+11 3.4482e+12 1800.6
+ Kilometers_Driven  1 1.1473e+11 3.6625e+12 1805.0
<none>                             3.7772e+12 1805.2
+ Age                1 9.9348e+10 3.6779e+12 1805.3
+ Seats              1 5.4295e+10 3.7229e+12 1806.2
+ Engine             1 4.1425e+10 3.7358e+12 1806.4
- Mileage            1 2.6867e+11 4.0459e+12 1808.2
+ Owner_Type         2 1.4600e+10 3.7626e+12 1808.9
- Fuel_Type          1 9.7046e+11 4.7477e+12 1819.7
- Power              1 7.9986e+12 1.1776e+13 1885.1
- Brand             10 1.3729e+13 1.7507e+13 1895.6


Step:  AIC=1800.64
Price ~ Power + Brand + Fuel_Type + Mileage + Transmission


                    Df  Sum of Sq        RSS    AIC
+ Kilometers_Driven  1 1.0769e+11 3.3405e+12 1800.3
+ Seats              1 1.0459e+11 3.3436e+12 1800.4
<none>                             3.4482e+12 1800.6
+ Age                1 5.9766e+10 3.3884e+12 1801.4
+ Engine             1 2.3657e+10 3.4245e+12 1802.1
+ Owner_Type         2 3.2029e+10 3.4162e+12 1804.0
- Transmission       1 3.2901e+11 3.7772e+12 1805.2
- Mileage            1 3.6655e+11 3.8148e+12 1805.9
- Fuel_Type          1 6.9992e+11 4.1481e+12 1811.9
- Power              1 4.5691e+12 8.0173e+12 1859.4
- Brand             10 9.9217e+12 1.3370e+13 1878.2


Step:  AIC=1800.35
Price ~ Power + Brand + Fuel_Type + Mileage + Transmission +
    Kilometers_Driven


                    Df  Sum of Sq        RSS    AIC
<none>                             3.3405e+12 1800.3
+ Seats              1 8.0771e+10 3.2597e+12 1800.6
- Kilometers_Driven  1 1.0769e+11 3.4482e+12 1800.6
+ Engine             1 2.9204e+10 3.3113e+12 1801.7
+ Age                1 1.9443e+09 3.3386e+12 1802.3
+ Owner_Type         2 6.0638e+10 3.2799e+12 1803.0
- Transmission       1 3.2197e+11 3.6625e+12 1805.0
- Mileage            1 4.6644e+11 3.8070e+12 1807.8
- Fuel_Type          1 8.0407e+11 4.1446e+12 1813.9
- Power              1 4.4575e+12 7.7980e+12 1859.4
- Brand             10 9.2120e+12 1.2552e+13 1875.7


Call:
lm(formula = Price ~ Power + Brand + Fuel_Type + Mileage + Transmission +
    Kilometers_Driven, data = traincars)


Coefficients:
```

```
      (Intercept)              Power           BrandBMW           BrandFord
BrandHonda       BrandHyundai      BrandMahindra
        1.993e+06          5.657e+03          4.238e+04         -9.910e+05         -
8.947e+05         -1.166e+06         -1.194e+06
      BrandMaruti      BrandMercedes           BrandTata        BrandToyota
BrandVolkswagen         Fuel_Type1            Mileage
       -9.918e+05          2.926e+03         -8.977e+05         -7.333e+05         -
8.508e+05          2.632e+05         -4.236e+04
    Transmission1  Kilometers_Driven
        2.207e+05         -6.367e+00
```

graph3B

```
Start:  AIC=1992.78
Price ~ 1

                    Df  Sum of Sq        RSS    AIC
+ Power              1 5.1251e+13 1.9823e+13 1905.1
+ Brand             10 5.2312e+13 1.8762e+13 1939.7
+ Engine             1 3.5792e+13 3.5282e+13 1946.6
+ Transmission       1 3.5181e+13 3.5892e+13 1947.9
+ Mileage            1 2.7868e+13 4.3206e+13 1961.2
+ Owner_Type         2 1.0999e+13 6.0075e+13 1989.2
+ Age                1 5.5270e+12 6.5547e+13 1991.2
<none>                            7.1074e+13 1992.8
+ Fuel_Type          1 6.2500e+11 7.0449e+13 1996.4
+ Seats              1 1.3638e+10 7.1060e+13 1997.0
+ Kilometers_Driven  1 6.0844e+09 7.1068e+13 1997.0

Step:  AIC=1905.12
Price ~ Power

                    Df  Sum of Sq        RSS    AIC
+ Brand             10 1.4829e+13 4.9940e+12 1848.6
+ Transmission       1 4.6105e+12 1.5213e+13 1890.3
+ Fuel_Type          1 1.7123e+12 1.8111e+13 1902.9
+ Seats              1 1.2818e+12 1.8541e+13 1904.6
<none>                            1.9823e+13 1905.1
+ Mileage            1 3.9791e+11 1.9425e+13 1907.9
+ Age                1 1.4236e+11 1.9681e+13 1908.9
+ Engine             1 9.9061e+10 1.9724e+13 1909.0
+ Kilometers_Driven  1 2.7399e+09 1.9820e+13 1909.4
+ Owner_Type         2 9.7582e+11 1.8847e+13 1910.0
- Power              1 5.1251e+13 7.1074e+13 1992.8

Step:  AIC=1848.63
Price ~ Power + Brand

                    Df  Sum of Sq        RSS    AIC
+ Fuel_Type          1 9.4813e+11 4.0459e+12 1837.7
+ Seats              1 6.1340e+11 4.3806e+12 1843.5
+ Transmission       1 4.6484e+11 4.5292e+12 1845.9
<none>                            4.9940e+12 1848.6
+ Mileage            1 2.4633e+11 4.7477e+12 1849.3
+ Engine             1 1.5615e+11 4.8379e+12 1850.6
```

```
+ Kilometers_Driven  1 7.5671e+10 4.9183e+12 1851.8
+ Age                1 3.0485e+10 4.9635e+12 1852.5
+ Owner_Type         2 1.1992e+11 4.8741e+12 1855.4
- Brand             10 1.4829e+13 1.9823e+13 1905.1
- Power              1 1.3768e+13 1.8762e+13 1939.7


Step:  AIC=1837.74
Price ~ Power + Brand + Fuel_Type


                    Df  Sum of Sq        RSS    AIC
+ Mileage            1 2.6867e+11 3.7772e+12 1837.1
<none>                            4.0459e+12 1837.7
+ Transmission       1 2.3113e+11 3.8148e+12 1837.8
+ Seats              1 1.9533e+11 3.8505e+12 1838.5
+ Age                1 3.4376e+10 4.0115e+12 1841.4
+ Kilometers_Driven  1 1.5652e+10 4.0302e+12 1841.7
+ Engine             1 1.2654e+09 4.0446e+12 1842.0
+ Owner_Type         2 3.6448e+10 4.0094e+12 1845.7
- Fuel_Type          1 9.4813e+11 4.9940e+12 1848.6
- Brand             10 1.4065e+13 1.8111e+13 1902.9
- Power              1 1.4380e+13 1.8426e+13 1942.6


Step:  AIC=1837.07
Price ~ Power + Brand + Fuel_Type + Mileage


                    Df  Sum of Sq        RSS    AIC
+ Transmission       1 3.2901e+11 3.4482e+12 1834.8
<none>                            3.7772e+12 1837.1
- Mileage            1 2.6867e+11 4.0459e+12 1837.7
+ Kilometers_Driven  1 1.1473e+11 3.6625e+12 1839.1
+ Age                1 9.9348e+10 3.6779e+12 1839.4
+ Seats              1 5.4295e+10 3.7229e+12 1840.3
+ Engine             1 4.1425e+10 3.7358e+12 1840.6
+ Owner_Type         2 1.4600e+10 3.7626e+12 1845.3
- Fuel_Type          1 9.7046e+11 4.7477e+12 1849.3
- Brand             10 1.3729e+13 1.7507e+13 1904.7
- Power              1 7.9986e+12 1.1776e+13 1914.7


Step:  AIC=1834.79
Price ~ Power + Brand + Fuel_Type + Mileage + Transmission


                    Df  Sum of Sq        RSS    AIC
<none>                            3.4482e+12 1834.8
+ Kilometers_Driven  1 1.0769e+11 3.3405e+12 1836.8
+ Seats              1 1.0459e+11 3.3436e+12 1836.8
- Transmission       1 3.2901e+11 3.7772e+12 1837.1
- Mileage            1 3.6655e+11 3.8148e+12 1837.8
+ Age                1 5.9766e+10 3.3884e+12 1837.8
+ Engine             1 2.3657e+10 3.4245e+12 1838.6
+ Owner_Type         2 3.2029e+10 3.4162e+12 1842.7
- Fuel_Type          1 6.9992e+11 4.1481e+12 1843.8
- Brand             10 9.9217e+12 1.3370e+13 1889.6
- Power              1 4.5691e+12 8.0173e+12 1891.3


Call:
lm(formula = Price ~ Power + Brand + Fuel_Type + Mileage + Transmission,
    data = traincars)
```

```
Coefficients:
    (Intercept)            Power         BrandBMW        BrandFord       BrandHonda
BrandHyundai    BrandMahindra      BrandMaruti
        1689226             5715            83466          -969348          -935963
-1204552         -1167995          -965189
  BrandMercedes        BrandTata      BrandToyota  BrandVolkswagen       Fuel_Type1
Mileage    Transmission1
          33754          -901201          -755428          -845437           222150
-35033           223043


graph3C
BrandMercedes BrandTata BrandToyota BrandVolkswagen Age
1  ( 1 ) " "       " "        " "          " "            " "             " "
" "             " "        " "          " "            " "
2  ( 1 ) " "       " "        " "          " "            " "             " "
" "             " "        " "          " "            " "
3  ( 1 ) " "       " "        " "          "*"            " "             " "
" "             " "        " "          " "            " "
4  ( 1 ) "*"       " "        " "          " "            " "             " "
"*"             " "        " "          " "            " "
5  ( 1 ) "*"       " "        " "          " "            " "             " "
"*"             " "        " "          " "            " "
6  ( 1 ) "*"       " "        " "          "*"            " "             " "
"*"             " "        " "          " "            " "
7  ( 1 ) "*"       " "        " "          "*"            " "             " "
"*"             " "        " "          " "            "*"
8  ( 1 ) " "       "*"        "*"          "*"            "*"             "*"
" "             "*"        " "          "*"            " "
          Kilometers_Driven Fuel_Type1 Transmission1 Owner_Type2 Owner_Type3
Mileage Engine Power Seats
1  ( 1 ) " "                " "         " "            " "            " "          "
"       " "     "*"    " "
2  ( 1 ) " "                " "         "*"            " "            " "          "
"       " "     "*"    " "
3  ( 1 ) " "                " "         "*"            " "            " "          "
"       " "     "*"    " "
4  ( 1 ) " "                " "         "*"            " "            " "          "
"       " "     "*"    " "
5  ( 1 ) " "                " "         "*"            " "            " "          "
"       " "     "*"    "*"
6  ( 1 ) " "                " "         "*"            " "            " "          "
"       " "     "*"    "*"
7  ( 1 ) " "                " "         "*"            " "            " "          "
"       " "     "*"    "*"
8  ( 1 ) " "                " "         " "            " "            " "          "
"       " "     "*"    " "
```

VIF

	GVIF	Df	GVIF^(1/(2*Df))
Power	3.484163	1	1.866591
Brand	6.596049	10	1.098915

```
Fuel_Type       1.531354  1       1.237479
Mileage         3.084286  1       1.756214
Transmission    2.670094  1       1.634042
Kilometers_Driven 1.834917  1       1.354591
                GVIF Df GVIF^(1/(2*Df))
Power       3.469040  1       1.862536
Brand       4.861174 10       1.082274
Fuel_Type   1.253487  1       1.119592
Mileage     2.684819  1       1.638542
Transmission 2.669180  1       1.633762
                GVIF Df GVIF^(1/(2*Df))
Brand       3.664279 10       1.067086
Transmission 2.499584  1       1.581007
Power       2.276120  1       1.508682
Seats       1.364545  1       1.168137
```

Appendix:

Graph 1A: This plot suggests a weak negative relationship between kilometers driven and price. There are a large number of cars clustered at the lower price points, suggesting that kilometers driven might not be an important explanatory variable among cheap cars. There seems to be two clusters among cars with higher value and lower value.

Graph 1C: There is a noticeable positive linear relationship between engine capacity and price. There is a leverage point at the 5000 cc level, which possibly stands for a performance car or a truck that is different that daily used cars. There seems to be an exponential curve that we might need to consider data transformation in our following steps.

In graph 2A, 2018 cars have the highest median and spread, possibly due to high-value and large amount vehicles sold that year. 2020 cars are cheapest, it is possibly because the random seed we use and COVID can be potential factors.

In graph 2B, we can tell first-hand cars have higher median price and more variation, then the second-hand cars' price drops notably. Third-hand cars are consistently low-priced with narrow spread. To summarize, the owner type is a strong predictor.

In graph 2C, diesel cars have a higher mean than petrol cars but the petrol cars have a wider distribution. This is possibly because of diesel cars are generally SUVs which are larger and bit more expansive than sedans. However, high value cars like luxury or sport cars will not be powered by petrol.

In graph 2D, 5-seat cars are dominant but vary widely in price. 7-seat cars have a higher mean than 5-seat cars. 4-seat cars have the highest mean but a small IQR, this is possibly because of the lack of enough sales record for 4-seat cars.

In order to make sure there is no multicollinearity appearing among the variables, we calculate the VIF to avoid happening. Due to the variables containing categorical variables in it, the R output shows GVIF for

the table and by researching it is used for categorical variables. For all the predictors included all of them have no multicollinearity appears which help us to build the model with individual variables.

Reference:

1. Audi A5 official site The new Audi A5 models | Audi MediaCenter
2. Sokoya, G. O. (2018). *A quantitative study of the relationship between leadership styles and employee job satisfaction* (Doctoral dissertation, Walden University). Walden Dissertations and Doctoral Studies. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=7228&context=dissertations