DSCI 5260 Project
An Analysis on Used Cars in the United Kingdom

Chintan Rajesh | Jonathan Moncrief | Aziz Haryani | Sreshta Budeti

Contents

Executive Summary

This project is analyzing the key factors determining the price of cars in the United Kingdom. The file we have chosen to consist of over 390,000 datapoints since it covers a over a span from 1970s till 2020 and beyond. In order to come with key indicators first we first had to load the data in order to filter out missing data and plot the missing data in terms of the areas where the missing data such as year, mileage, transmission etc. Once we have plot the missingness, we went to find the skewness in the data via boxplot to determine the outliers and see where is the datapoint more concentrated and how much it is spread across.

Now in order to perform EDA (exploratory data analysis) we need to have a target entity, which is going to be price and we will use the price of cars as the entity used to compare other factors such as mileage, year, mpg and tax etc. now we know that since we are dealing with a numeric based problem and doing predictions in terms of determining the price of a car, we performed a linear regression based data analysis on the dataset followed by decision tree, lasso regression and ridge regression.

Problem Statement

Through this project, we attempted to build Machine learning models to predict the price of around 400,000 used cars and determining the key factors that led to the price points of the cars.

Introduction

For the process of determining what dataset our group was going to use for this project, there were two important aspects that we required for the project. First, we wanted to analyze information that all of our group had an interest in. Secondly, the dataset that would be chosen needed to have enough information to be able to effectively create a model for this project. In our discussions and research for this project, we came across the dataset, Used Car Data Set on the Kaggle website.

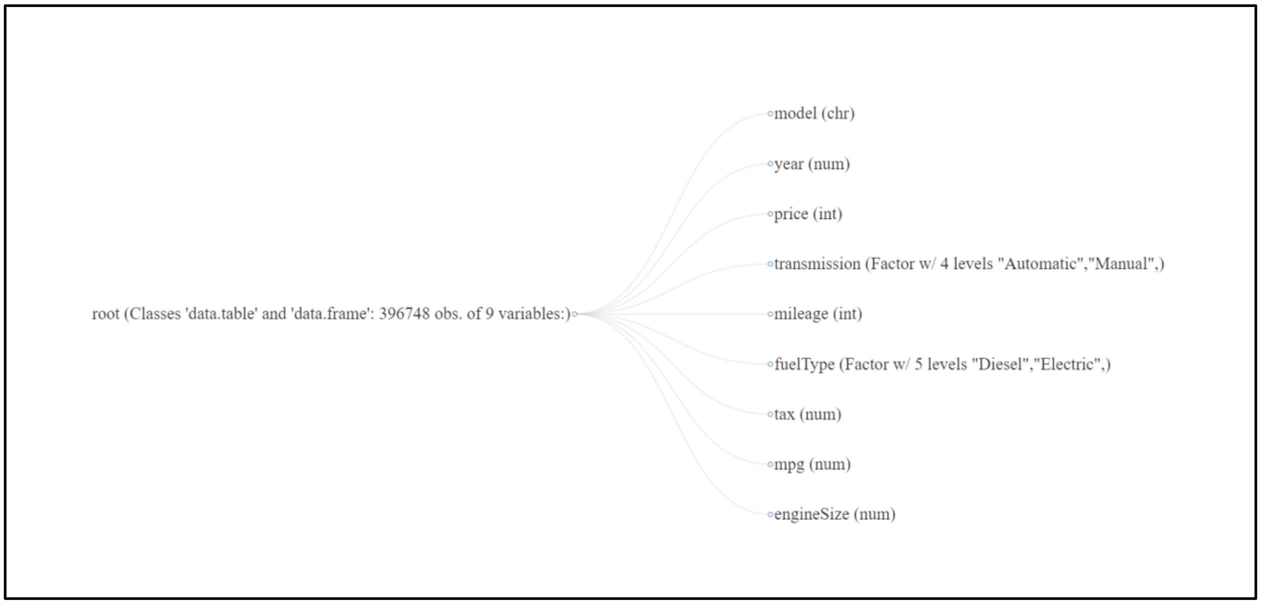
Data Description

In regard to the dataset having enough information, the combined dataset consists of a total of 396,748 observations of used cars that were built from 1970 to 2020 with a total of 10 different dimensions/variables for each observation. The variables consist of the model of the vehicle, vehicle year, price of the vehicle, transmission type, total mileage, fuel type, taxes, miles per gallon, and engine size. Depicted below is an example of the dataset.

```
##   model year price transmission mileage fuelType tax  mpg engineSize tax.Â..
## 1    A1 2017 12500       Manual   15735   Petrol 150 55.4        1.4      NA
## 2    A6 2016 16500    Automatic   36203   Diesel  20 64.2        2.0      NA
## 3    A1 2016 11000       Manual   29946   Petrol  30 55.4        1.4      NA
## 4    A4 2017 16800    Automatic   25952   Diesel 145 67.3        2.0      NA
## 5    A3 2019 17300       Manual    1998   Petrol 145 49.6        1.0      NA
## 6    A1 2016 13900    Automatic   32260   Petrol  30 58.9        1.4      NA
```

Transforming the Data

Therefore, the dataset appears to have enough observations and variables to analyze and potentially create a model to test this dataset. As previously discussed, there are 10 dimensions and 396,748 rows. Additionally, we determined that the transmission size and fuel types were a characters. After further analysis, we decided to transform the transmission and fuel type were transformed to factor. So we designated target values such as 0 for manual transmission, 1 for automatic transmission and 2 for other transmission. We further ran Principal Component Analysis, and decided to not use those features and instead continued with other numeric features.



Outliers

Additionally, we observed that a year built for one of the data points was 2060. This appears to be an incorrect input; thus, was imputed with the mean of the dimension, which is 2017.

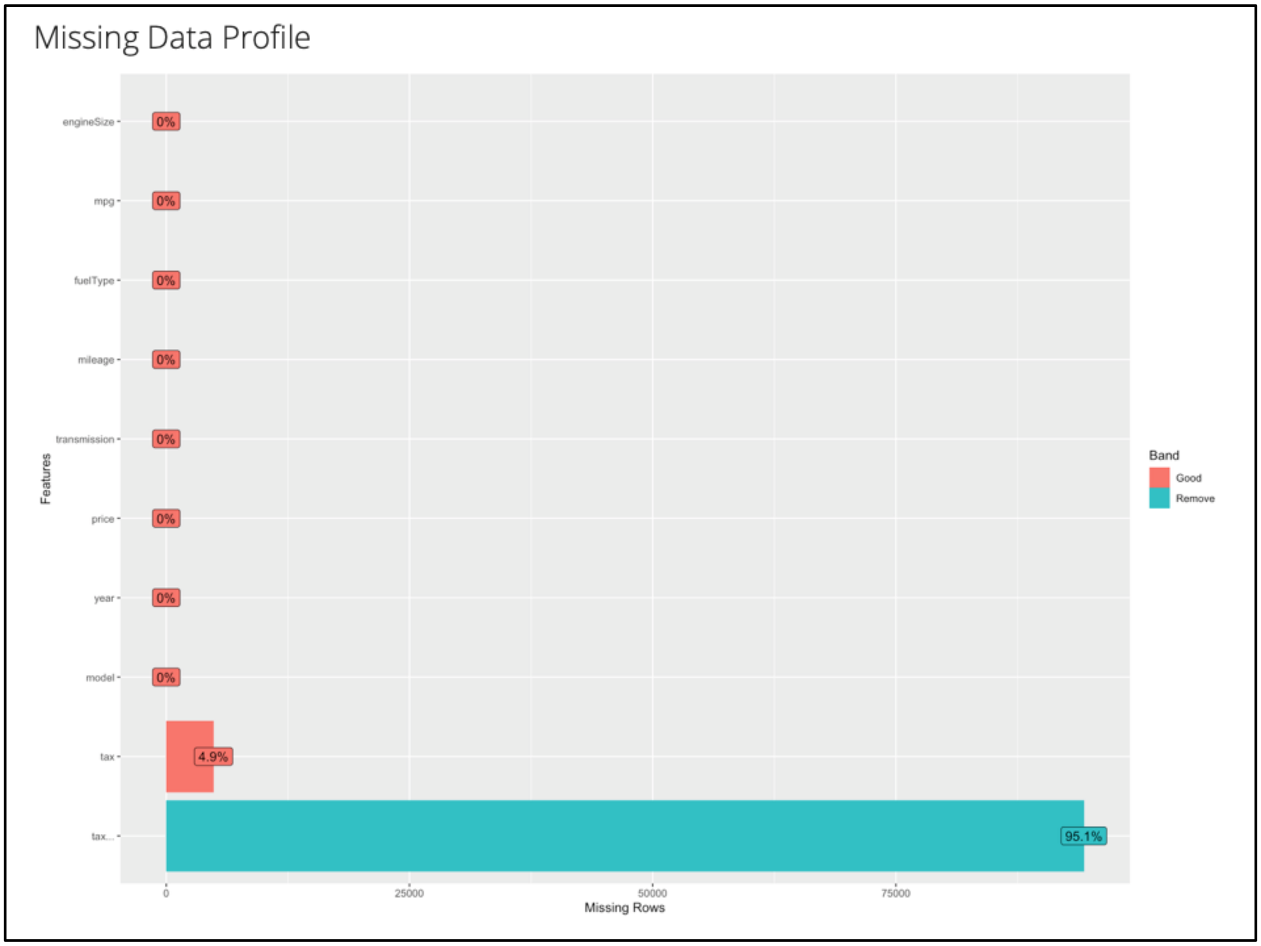
```
cardata$year[cardata$year == 2060] <- 2017

#Looking at the summary of year and MPG
summary(cardata$year)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1970    2016    2017    2017    2019    2020
```

Mean, Median, and Mode

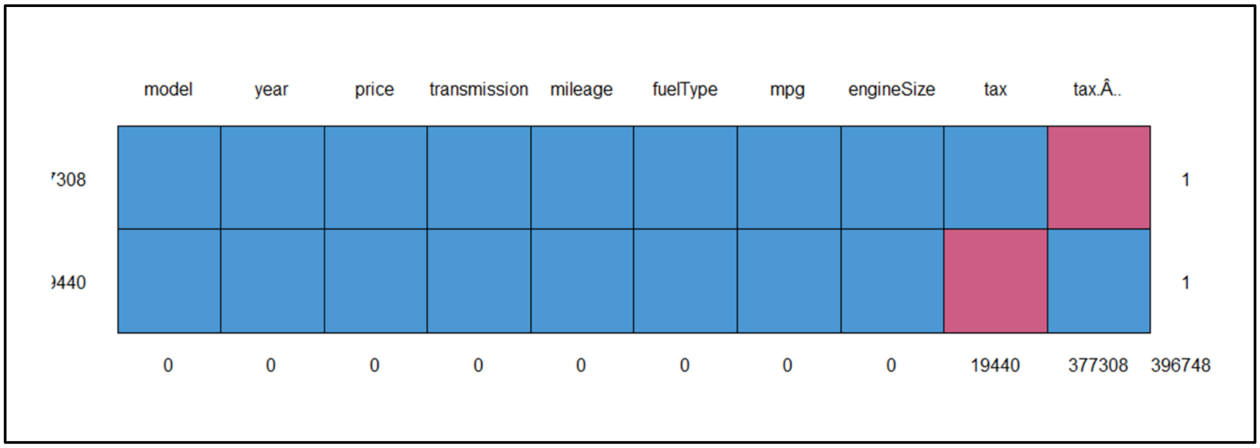
```
##     model               year          price          transmission
##  Length:396748      Min.   :1970   Min.   :    450   Automatic: 80224
##  Class :character   1st Qu.:2016   1st Qu.:   9999   Manual   :225780
##  Mode  :character   Median :2017   Median : 14495   Other    :     36
##                     Mean   :2017   Mean   : 16805   Semi-Auto: 90708
##                     3rd Qu.:2019   3rd Qu.: 20870
##                     Max.   :2020   Max.   :159999
##     mileage          fuelType          tax              mpg
##  Min.   :      1   Diesel  :163712   Min.   :  0.0   Min.   :  0.30
##  1st Qu.:   7424   Electric:     24   1st Qu.:120.3   1st Qu.: 47.10
##  Median : 17460   Hybrid  : 12312   Median :145.0   Median : 54.30
##  Mean   : 23059   Other   :    988   Mean   :120.3   Mean   : 55.17
##  3rd Qu.: 32340   Petrol  :219712   3rd Qu.:145.0   3rd Qu.: 62.80
##  Max.   :323000                      Max.   :580.0   Max.   :470.80
##    engineSize
##  Min.   :0.000
##  1st Qu.:1.200
##  Median :1.600
##  Mean   :1.663
##  3rd Qu.:2.000
##  Max.   :6.600
```

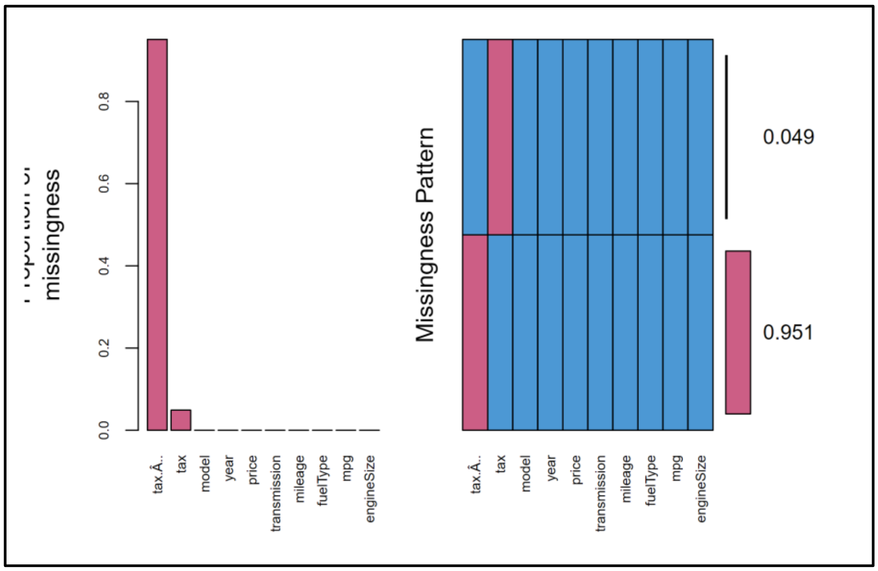
After the outlier has been corrected, we analyzed the mean, median, and mode. The MPG appears to have vehicles that can achieve over 470 miles per gallon. This appeared to be another outlier. However, after another analysis of the dataset, it appears that are other vehicles by the same make and model can achieve higher miles gallon. Thus, the dataset was not adjusted to remove these observations.

Missingness of the Data

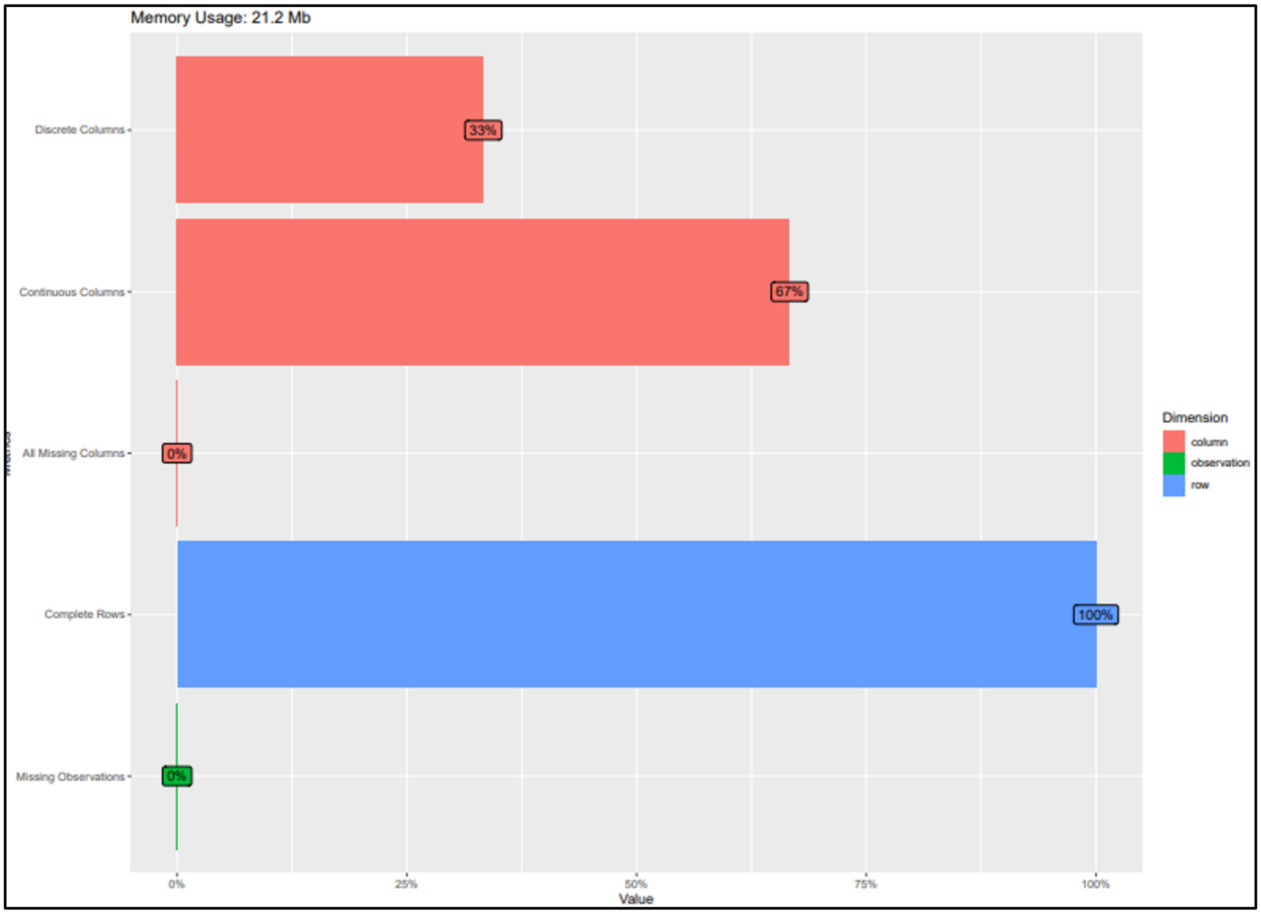



```
p <- function(x) {sum(is.na(x))/length(x)*100}
apply(cardata, 2, p)


##        model         year        price transmission      mileage     fuelType
##     0.000000     0.000000     0.000000     0.000000     0.000000     0.000000
##          tax          mpg   engineSize       tax.Â..
##     4.899836     0.000000     0.000000    95.100164

md.pattern(cardata, plot = TRUE)
```





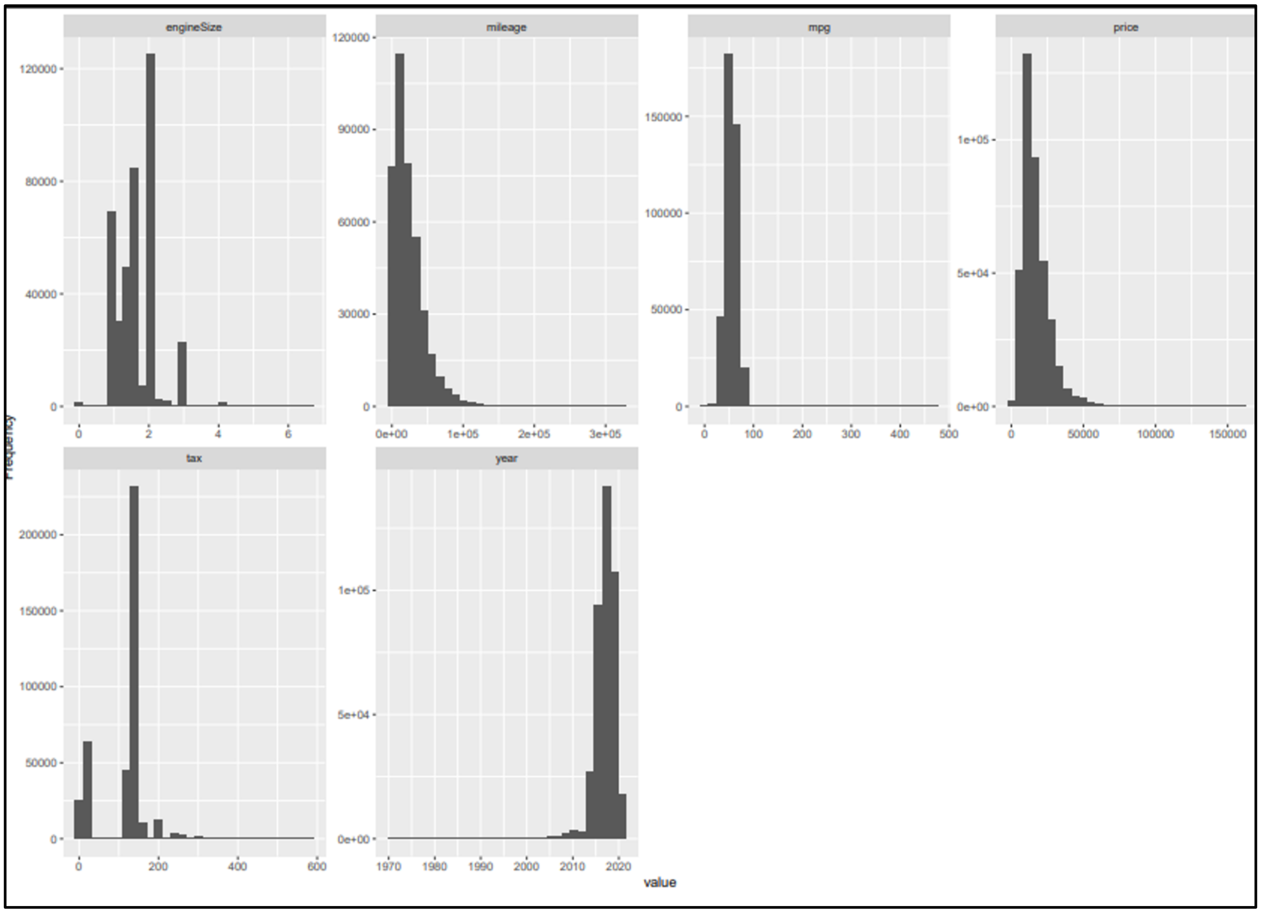
We observed during analysis that 95 percent of the observations were missing the Tax. A variable. This dimension appears to be collected with less than 5 percent of the observations. Additionally, the Tax.A dimension appeared to be a conversion of the tax into euro from the dollar. Which was not necessary for the dataset, given that over 95 percent of the taxes were reported in dollars. Therefore, this column was removed from the dataset.



Depicted above shows the dataset after the Tax.A column was removed the dataset, the dataset now has no missingness with any observations.

Measure of Skewness

We also performed an analysis of the data structure a univariate analysis to understand the distribution of values for a single variable for the dataset.



The histogram graph of all the variables tells that the data is not normally distributed. The data is either distributed the left, or right, or it is unevenly distributed.

Engine Size - As mentioned how uneven the data is we can see in engine size how the data is distributed. The min datapoint is 0 and the max datapoint is 6.6. The 1st quartile and 3rd quartile is 1.2 and 2.0 respectively. Based on these datapoints we can see the engine size data is skewed towards the left and mostly concentrated between 1.2 to 2.0

Mileage - The min datapoint is 1 and the max datapoint is 323,000. The 1st quartile and 3rd quartile is 7424 and 32,340 respectively. Based on these datapoints we can see the mileage data is skewed towards the left and mostly concentrated between 0 and 3rd quartile.

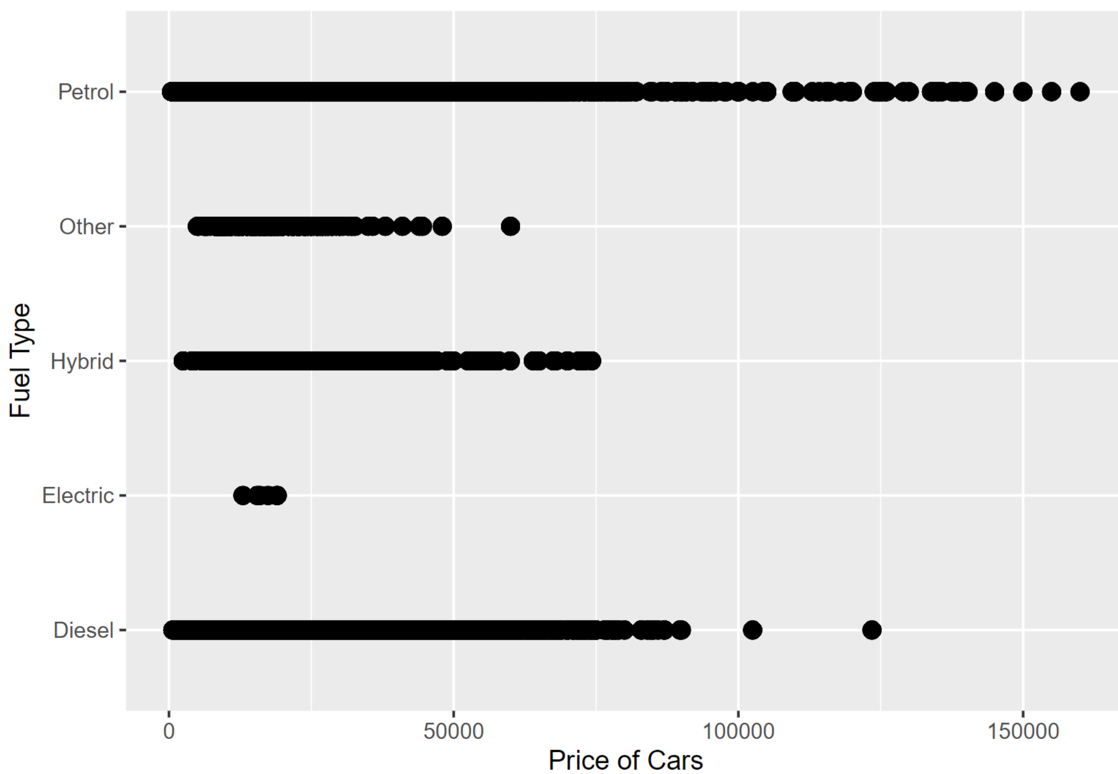
MPG - The min datapoint is 0.30 and the max datapoint is 470.80. The 1st quartile and 3rd quartile is 47.10 and 62.8 respectively. Based on these datapoints we can see the MPG data is skewed towards the left and mostly concentrated between 1st and 3rd quartile.

Price - This is our target variable. The min datapoint is $450 and the max datapoint is $159,999. The 1st quartile and 3rd quartile is $9,999 and $20,870 respectively. Based on these datapoints we can see the price data is skewed towards the left and mostly concentrated between 1st and 3rd quartile.

Tax - The min datapoint is $0 and the max datapoint is $580. The 1st quartile and 3rd quartile is $120.30 and $145 respectively. Based on these datapoints we can see the tax data is skewed towards the left and spread unevenly.

Year - The oldest car in the dataset is from 1970, which is our outlier, and the most recent car is from 2020. The 1st quartile and 3rd quartile is 2016 and 2019 respectively. Based on these datapoints we can see the year data is skewed towards the right and mostly concentrated between 1st and 3rd quartile.

Relationship between fuel type and price

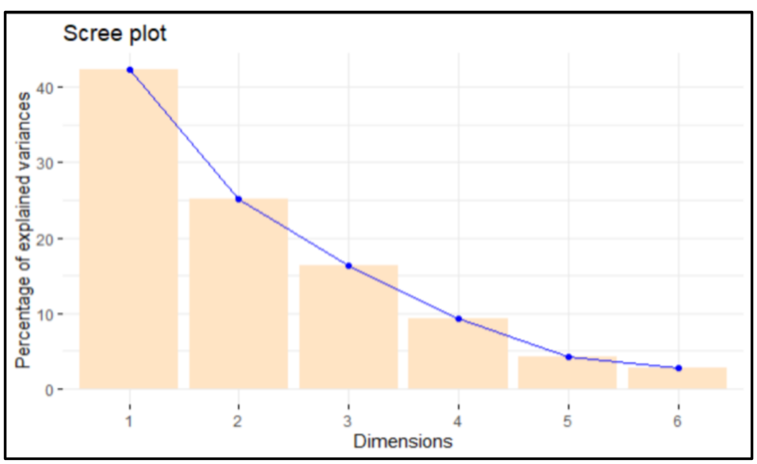


Based on the chart presented, we can conclude that the greatest number of cars in the dataset are Petrol followed by Diesel and so on. Additionally, we can also observe that the most expensive cars are Petroleum based.

Principal Component Analysis

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.5929 1.2294 0.9881 0.74530 0.50087 0.41044
Proportion of Variance 0.4229 0.2519 0.1627 0.09258 0.04181 0.02808
Cumulative Proportion  0.4229 0.6748 0.8375 0.93011 0.97192 1.00000
                  PC1        PC2        PC3         PC4          PC5
year        0.4417368 -0.4879535  0.1168747 -0.01570584 -0.682697466
price       0.5226547  0.1461787  0.4365598  0.06677394 -0.001391164
mileage    -0.4282324  0.5129753  0.0344070 -0.02500506 -0.718493010
tax         0.3727545  0.2413916 -0.5236387 -0.72335580 -0.032582468
mpg        -0.3466827 -0.2677013  0.5758340 -0.68621182  0.072407347
engineSize  0.2986350  0.5894537  0.4344993 -0.02321716  0.106703271
                   PC6
year       -0.29459773
price       0.71443592
mileage     0.18823389
tax         0.06543257
mpg         0.02080436
engineSize -0.60220200
```

We used **Principal Component Analysis** (PCA) as a non-parametric statistical technique primarily to check whether which features with most importance. We can notice that PC4 and PC5 have the lowest variance. However, since it did not make a huge difference to our model, we determined that the components in PCA were not significant enough; hence, we decided to drop the PCA.



Prediction Models

Predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data.

For our Data, we used 4 different machine learning models to predict the price. For this project, we used the following models:

1)LINEAR REGRESSION

2) DECISION TREE

3) RIDGE REGRESSION

4) LASSO REGRESSION

Linear Regression

Linear regression is one of the most used predictive modelling techniques .It is represented by an equation $Y = a + bX + e$, where a is the intercept, b is the slope of the line and e is the error term. This equation can be used to predict the value of a target variable based on given predictor variable.

For pricing prediction, Linear Regression is the most thought after model.

```
Call:
lm(formula = price ~ year + mpg + engineSize + mileage, data =
traindata)

Residuals:
   Min     1Q Median     3Q    Max
-42297  -3095   -436   2307 107832

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.205e+06  1.357e+04 -236.18   <2e-16 ***
year         1.589e+03  6.722e+00  236.42   <2e-16 ***
mpg         -2.573e+01  6.250e-01  -41.16   <2e-16 ***
engineSize   1.178e+04  1.796e+01  655.70   <2e-16 ***
mileage     -1.078e-01  6.869e-04 -156.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5385 on 316904 degrees of freedom
Multiple R-squared:  0.7023,    Adjusted R-squared:  0.7023
F-statistic: 1.869e+05 on 4 and 316904 DF,  p-value: < 2.2e-16
```

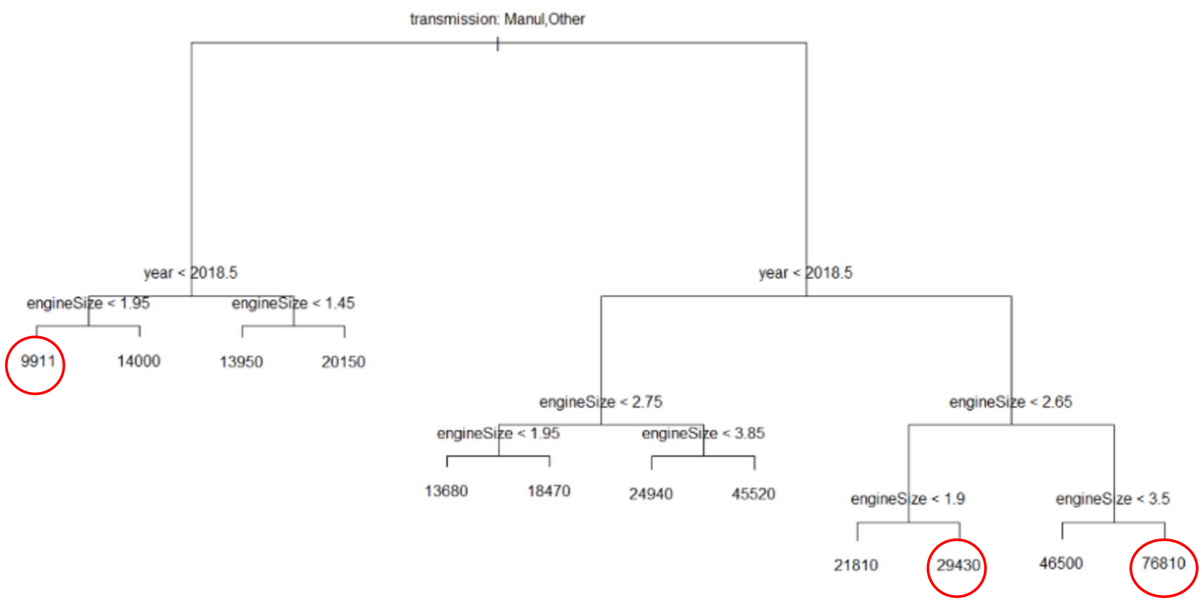
We got Adjusted R-squared as 70.23% which implies it is a good prediction model. Based on the linear regression model, we get the following linear equation:

Y=-3.21 +1.59(year)-2.57(mpg)+1.17(Engine Size)-1.08(mileage)

Decision Tree

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

The variables used in the model are Transmission, Year and Engine Size.



The rules of the decision tree are

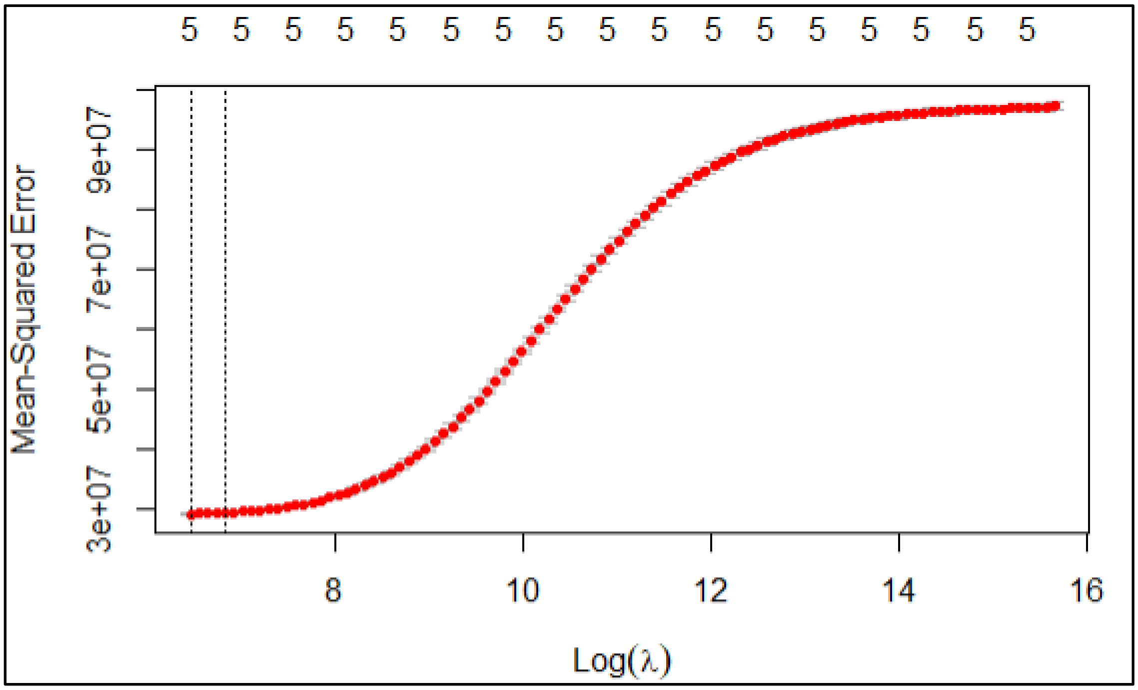
1) Lowest Price Vehicle
   When Transmission = manual AND Year < 2018.5 AND Engine Size < 1.95, then according to the prediction model the price of the car is $9,911.

2) Moderately Priced vehicle
   When Transmission = Other AND Year > 2018.5 AND Engine Size < 2.65 AND Engine Size > 1.9 , then according to the prediction model the price of the car is $29,430.

3) Most Expensive Vehicle
   When Transmission = Other AND Year > 2018.5 AND Engine Size > 3.5, then according to the prediction model the price of the car is $76,810.

Ridge Regression

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values. The variables that were utilized for this prediction model are year, engine size, tax, mileage, and mpg.



As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

$$\lambda = 629.8693$$

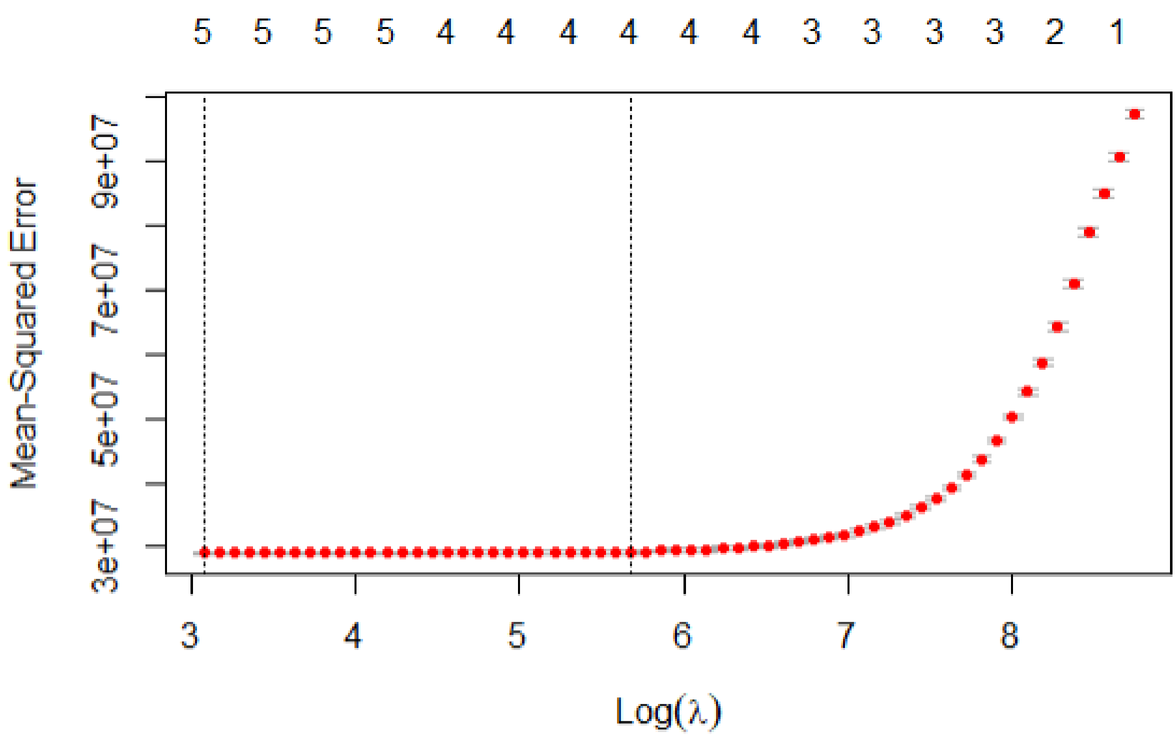
This is optimal lambda value that minimizes the mean squared error.

```
6 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept) -3.033279e+06
year         1.505136e+03
engineSize   1.099846e+04
tax          1.680694e-02
mileage     -1.040993e-01
mpg         -3.264531e+01
```

Y=-3.03 +1.51(year)+1.10(enginesize)+1.68(tax)+-1.04(mileage)+-3.26(mpg)

Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).



As λ decrease, the mean squared error decreases, leading to decreased variance but increased bias.

$$\lambda = 21.60747$$

This is optimal lambda value that minimizes the mean squared error.

```
6 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept) -3.033279e+06
year         1.505136e+03
engineSize   1.099846e+04
tax          1.680694e-02
mileage     -1.040993e-01
mpg         -3.264531e+01
```

Conclusion

Based on the project and all the analysis that best prediction model is Linear Regression Model with 54% accuracy with Adjusted R-Square as 70.23%. The reason behind such low accuracy is linked with skewedness of data and overfitting problem.

The Decision Tree Model is the next best model with prediction price as it uses variables and gives the accuracy of 53%.

References

https://rstudio-pubs-static.s3.amazonaws.com/248952_706edc85cfa84a369dfe401a763d32fc.html

www.Kaggle.com

https://www.statology.org/ridge-regression-in-r/

https://www.statology.org/lasso-regression-in-r/