Practical MVA - Car Price Estimation

Goktug Cengiz
Henry Qiu

June 2019

Contents

1 Introduction

The aim of this project is on estimating car price. The data used in this project was provided by the UCI Machine Learning Repository. One of the most significant question we wanted to answer with this project was to find how much of an impact does the factors play a role in determining the price of a car using multivariate analysis techniques.

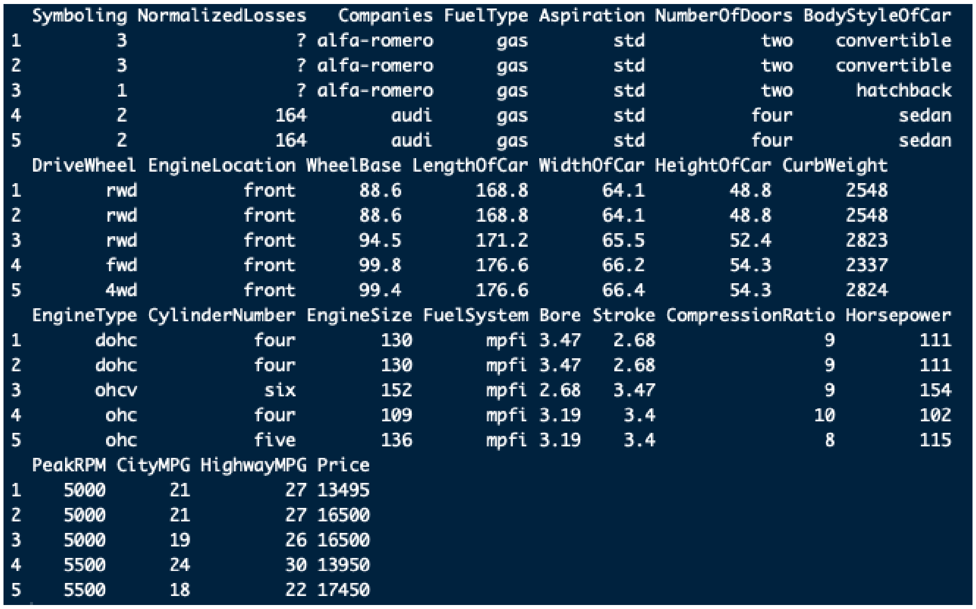
2 Data Understanding

Attribute Information:

- **Symboling:** -3, -2, -1, 0, 1, 2, 3. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.
- **NormalizedLosses:** continuous from 65 to 256.
- **Companies:** Name of car company (alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo).
- **FuelType:** Car fuel type (diesel, gas).
- **Aspiration:** Aspiration used in a car (std, turbo).
- **NumberOfDoors:** Number of doors in a car (four, two).
- **BodyStyleOfCar:** Body of car (hardtop, wagon, sedan, hatchback, convertible).
- **DriveWheel:** type of drive wheel (4wd, fwd, rwd).
- **EngineLocation:** Location of car engine (front, rear).
- **WheelBase:** Distance between wheels of car (continuous from 86.6 120.9). distance between wheels of car.
- **LengthOfCar:** Length of Car (continuous from 141.1 to 208.1).
- **WidthOfCar:** Width of Car (continuous from 60.3 to 72.3.)
- **HeightOfCar:** Height of Car (continuous from 47.8 to 59.8).
- **CurbWeight:** The weight of a car without occupants or baggage. (continuous from 1488 to 4066).

- **EngineType:** Type of engine (dohc, dohcv, l, ohc, ohcf, ohcv, rotor).
- **CylinderNumber:** Number of cylinders placed in the car (eight, five, four, six, three, twelve, two).
- **EngineSize:** Size of engine (continuous from 61 to 326.)
- **FuelSystem:** 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
  bbl - BBL means the amount of holes that air enters the engine the technical name is barrels. When you take off the air filter there will be either 1 hole or 2 holes side by side or 4 holes 2 in the front and 2 in the rear. Hence the values 1bbl, 2bbl, 4bbl
  idi - Indirect injection. Fuel is not directly injected into the combustion chamber port injection
  spfi - Sequential Port fuel injection
  mpfi - Multipoint Port Fuel Injection
- **Bore:** (continuous from 2.54 to 3.94.) Bore is the diameter of each cylinder.
- **Stroke:** (continuous from 2.07 to 4.17.) Stroke is the length that it travels when moving from bottom position to the top position.
- **CompressionRatio:** (continuous from 7 to 23.) the ratio of the maximum to minimum volume in the cylinder of an internal combustion engine.
- **Horsepower:** (continuous from 48 to 288.) The power an engine produces is called horsepower.
- **PeakRPM:** (continuous from 4150 to 6600.) RPM stands for revolutions per minute, and it's used as a measure of how fast any machine is operating at a given time. In cars, rpm measures how many times the engine's crankshaft makes one full rotation every minute, and along with it, how many times each piston goes up and down in its cylinder.
- **CityMPG:** Mileage in city (continuous from 13 to 49.)
- **HighwayMPG:** Mileage on highway (continuous from 16 to 54.)
- **Price:** Price of car (continuous from 5118 to 45400.)

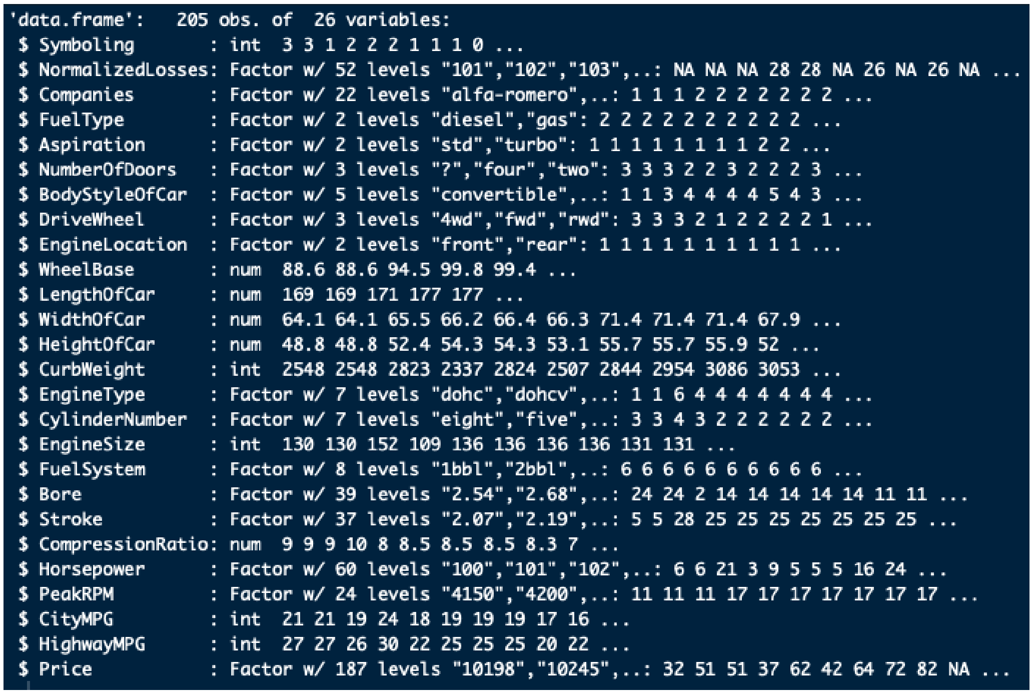
After reading the data in the .csv format we downloaded, we examined the data in the first 5 rows thanks to the head function. As you can see below, there are 2 types of variables, categorical and continuous. In addition, there are missing values in the data set, which are represented by a question mark. In the recoding section, NA values will be assigned to missing values represented by question mark.



```
  Symboling NormalizedLosses   Companies FuelType Aspiration NumberOfDoors BodyStyleOfCar
1         3                ? alfa-romero      gas        std           two    convertible
2         3                ? alfa-romero      gas        std           two    convertible
3         1                ? alfa-romero      gas        std           two      hatchback
4         2              164        audi      gas        std          four          sedan
5         2              164        audi      gas        std          four          sedan
  DriveWheel EngineLocation WheelBase LengthOfCar WidthOfCar HeightOfCar CurbWeight
1        rwd          front      88.6       168.8       64.1        48.8       2548
2        rwd          front      88.6       168.8       64.1        48.8       2548
3        rwd          front      94.5       171.2       65.5        52.4       2823
4        fwd          front      99.8       176.6       66.2        54.3       2337
5        4wd          front      99.4       176.6       66.4        54.3       2824
  EngineType CylinderNumber EngineSize FuelSystem Bore Stroke CompressionRatio Horsepower
1       dohc           four        130       mpfi 3.47   2.68                9        111
2       dohc           four        130       mpfi 3.47   2.68                9        111
3       ohcv            six        152       mpfi 2.68   3.47                9        154
4        ohc           four        109       mpfi 3.19    3.4               10        102
5        ohc           five        136       mpfi 3.19    3.4                8        115
  PeakRPM CityMPG HighwayMPG Price
1    5000      21         27 13495
2    5000      21         27 16500
3    5000      19         26 16500
4    5500      24         30 13950
5    5500      18         22 17450
```

Figure 1: Head of Data Set

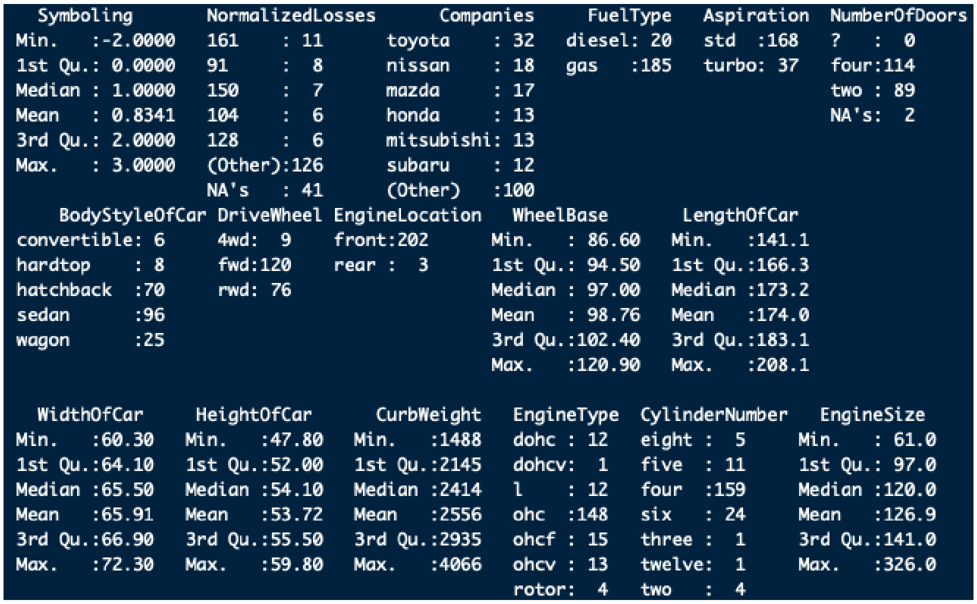
According to Figure 2, it can be considered as the data set consist of 205 observations and 26 variables. Moreover, we can see what the data types of data set are. However, we have to change some of variables' such as price due to the fact that price is continuous variable and can not represented as factor.



```
'data.frame':   205 obs. of  26 variables:
 $ Symboling       : int  3 3 1 2 2 2 1 1 1 0 ...
 $ NormalizedLosses: Factor w/ 52 levels "101","102","103",..: NA NA NA 28 28 NA 26 NA 26 NA ...
 $ Companies       : Factor w/ 22 levels "alfa-romero",..: 1 1 1 2 2 2 2 2 2 2 ...
 $ FuelType        : Factor w/ 2 levels "diesel","gas": 2 2 2 2 2 2 2 2 2 2 ...
 $ Aspiration      : Factor w/ 2 levels "std","turbo": 1 1 1 1 1 1 1 1 2 2 ...
 $ NumberOfDoors   : Factor w/ 3 levels "?","four","two": 3 3 3 2 2 3 2 2 2 3 ...
 $ BodyStyleOfCar  : Factor w/ 5 levels "convertible",..: 1 1 3 4 4 4 4 5 4 3 ...
 $ DriveWheel      : Factor w/ 3 levels "4wd","fwd","rwd": 3 3 3 2 1 2 2 2 2 1 ...
 $ EngineLocation  : Factor w/ 2 levels "front","rear": 1 1 1 1 1 1 1 1 1 1 ...
 $ WheelBase       : num  88.6 88.6 94.5 99.8 99.4 ...
 $ LengthOfCar     : num  169 169 171 177 177 ...
 $ WidthOfCar      : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
 $ HeightOfCar     : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
 $ CurbWeight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
 $ EngineType      : Factor w/ 7 levels "dohc","dohcv",..: 1 1 6 4 4 4 4 4 4 4 ...
 $ CylinderNumber  : Factor w/ 7 levels "eight","five",..: 3 3 4 3 2 2 2 2 2 2 ...
 $ EngineSize      : int  130 130 152 109 136 136 136 136 131 131 ...
 $ FuelSystem      : Factor w/ 8 levels "1bbl","2bbl",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ Bore            : Factor w/ 39 levels "2.54","2.68",..: 24 24 2 14 14 14 14 14 11 11 ...
 $ Stroke          : Factor w/ 37 levels "2.07","2.19",..: 5 5 28 25 25 25 25 25 25 25 ...
 $ CompressionRatio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
 $ Horsepower      : Factor w/ 60 levels "100","101","102",..: 6 6 21 3 9 5 5 5 16 24 ...
 $ PeakRPM         : Factor w/ 24 levels "4150","4200",..: 11 11 11 17 17 17 17 17 17 17 ...
 $ CityMPG         : int  21 21 19 24 18 19 19 19 17 16 ...
 $ HighwayMPG      : int  27 27 26 30 22 25 25 25 20 22 ...
 $ Price           : Factor w/ 187 levels "10198","10245",..: 32 51 51 37 62 42 64 72 82 NA ...
```

Figure 2: Data Types of Data Set

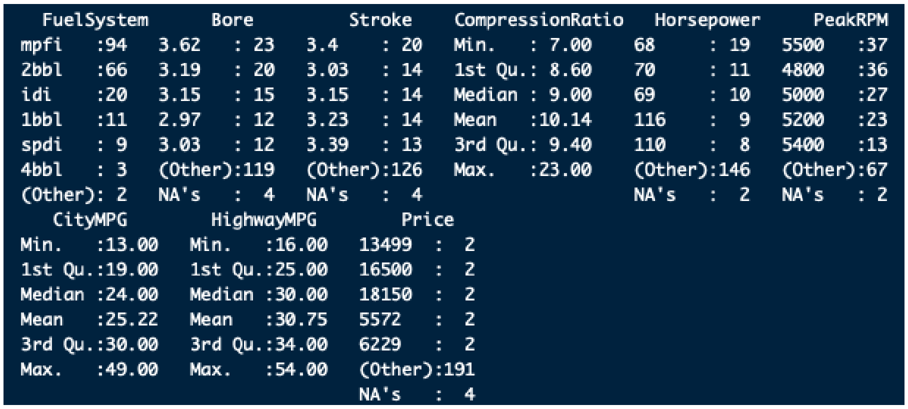
As you see below, we summarized data, so we have knowledge about variables some of statistical values such as minimum, maximum, mean, medium etc.



```
  Symboling        NormalizedLosses       Companies     FuelType   Aspiration  NumberOfDoors
Min.   :-2.0000   161     : 11      toyota    : 32   diesel: 20   std  :168    ?   :  0
1st Qu.: 0.0000   91      :  8      nissan    : 18   gas   :185   turbo: 37    four:114
Median : 1.0000   150     :  7      mazda     : 17                             two : 89
Mean   : 0.8341   104     :  6      honda     : 13                             NA's:  2
3rd Qu.: 2.0000   128     :  6      mitsubishi: 13
Max.   : 3.0000   (Other):126      subaru    : 12
                  NA's    : 41      (Other)   :100
    BodyStyleOfCar DriveWheel EngineLocation   WheelBase       LengthOfCar
convertible: 6     4wd:  9    front:202      Min.   : 86.60   Min.   :141.1
hardtop    : 8     fwd:120    rear :  3      1st Qu.: 94.50   1st Qu.:166.3
hatchback  :70     rwd: 76                   Median : 97.00   Median :173.2
sedan      :96                               Mean   : 98.76   Mean   :174.0
wagon      :25                               3rd Qu.:102.40   3rd Qu.:183.1
                                             Max.   :120.90   Max.   :208.1

  WidthOfCar      HeightOfCar       CurbWeight    EngineType  CylinderNumber    EngineSize
Min.   :60.30   Min.   :47.80   Min.   :1488   dohc : 12   eight :  5     Min.   : 61.0
1st Qu.:64.10   1st Qu.:52.00   1st Qu.:2145   dohcv:  1   five  : 11     1st Qu.: 97.0
Median :65.50   Median :54.10   Median :2414   l    : 12   four  :159     Median :120.0
Mean   :65.91   Mean   :53.72   Mean   :2556   ohc  :148   six   : 24     Mean   :126.9
3rd Qu.:66.90   3rd Qu.:55.50   3rd Qu.:2935   ohcf : 15   three :  1     3rd Qu.:141.0
Max.   :72.30   Max.   :59.80   Max.   :4066   ohcv : 13   twelve:  1     Max.   :326.0
                                               rotor:  4   two   :  4
```

Figure 3: Summary of Data Set



```
  FuelSystem       Bore          Stroke    CompressionRatio   Horsepower      PeakRPM
mpfi    :94   3.62    : 23   3.4     : 20   Min.   : 7.00    68      : 19   5500   :37
2bbl    :66   3.19    : 20   3.03    : 14   1st Qu.: 8.60    70      : 11   4800   :36
idi     :20   3.15    : 15   3.15    : 14   Median : 9.00    69      : 10   5000   :27
1bbl    :11   2.97    : 12   3.23    : 14   Mean   :10.14    116     :  9   5200   :23
spdi    : 9   3.03    : 12   3.39    : 13   3rd Qu.: 9.40    110     :  8   5400   :13
4bbl    : 3   (Other):119   (Other):126   Max.   :23.00    (Other):146   (Other):67
(Other): 2   NA's    :  4   NA's    :  4                    NA's    :  2   NA's   : 2
   CityMPG         HighwayMPG          Price
Min.   :13.00   Min.   :16.00   13499   :  2
1st Qu.:19.00   1st Qu.:25.00   16500   :  2
Median :24.00   Median :30.00   18150   :  2
Mean   :25.22   Mean   :30.75   5572    :  2
3rd Qu.:30.00   3rd Qu.:34.00   6229    :  2
Max.   :49.00   Max.   :54.00   (Other):191
                                NA's    :  4
```

Figure 4: Summary of Data Set

3 Data Pre-processing

3.1 Feature Selection

The ’Symboling’ and ’Normalized Losses’ columns were removed as these were introduced for insurance risk assessment purposes and are not meaningful to this study.

"EngineLocation" variable was excluded and will be explained why in visualization section.

"Stroke" variable was excluded and will be explained why in visualization section.

3.2 Recoding

Some of variables which are called as "Bore", "Stroke", "HorsePower", "PeakRPM" and "Price" were converted to character, then numeric or integer type.

3.3 Data Cleaning

Missing and outlier values cause many setbacks in our model. That’s why we’ll get rid of them.

3.3.1 Missing Value Detection

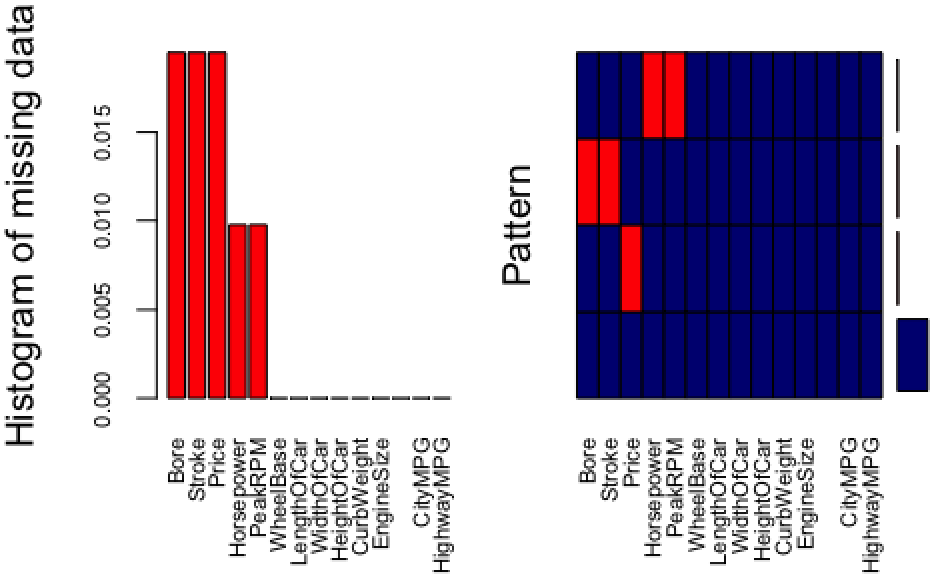


Figure 5: Missing Value Detection

We can interpret Figure 5 as the variables which are called as "Bore", "Stroke", "HorsePower", "PeakRPM" and "Price" have missing values.

3.3.2 Missing Value Imputation

Missing values were imputed thanks to missForest packege in R. 'missForest' is used to impute missing values particularly in the case of mixed-type data. It can be used to impute continuous and/or categorical data including complex interactions and nonlinear relations. It yields an out-of-bag (OOB) imputation error estimate. Moreover, it can be run parallel to save computation time. Consequently, we have both of continuous and categorical variables to be imputed that's why we use it.

3.3.3 Univariate Outlier Detection

In univariate outlier detection, outliers were detected by Box plot. If a value exceeds the box plot boundary line, it is considered as an outlier. In this case, the variables which are called as "WheelBase", "LengthOfCar", "WidthOfCar", "EngineSize", "Stroke", "CompressionRatio", "Horsepower", "PeakRPM", "CityMPG", "HighwayMPG" were accepted as outlier values.

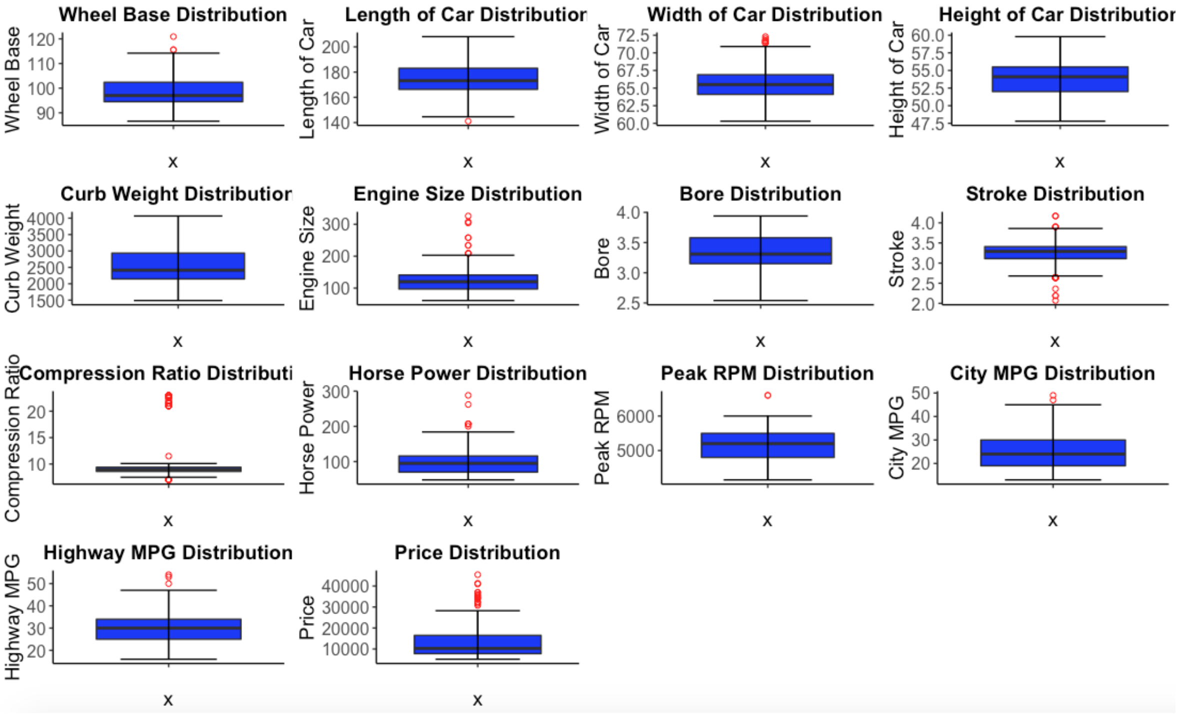


Figure 6: Outlier Detection by Box Plot

3.3.4 Multivariate Outlier Detection

The Mahalanobis distance was used of each data instance against the distribution of the whole data set to have an outlier score for each point. We can see how many points are considered outliers with each distance given a cutoff point with confidence level % 97.5.

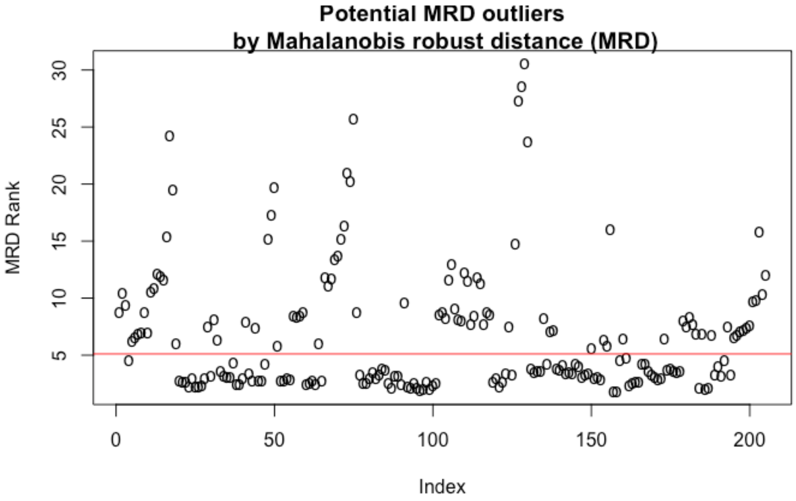
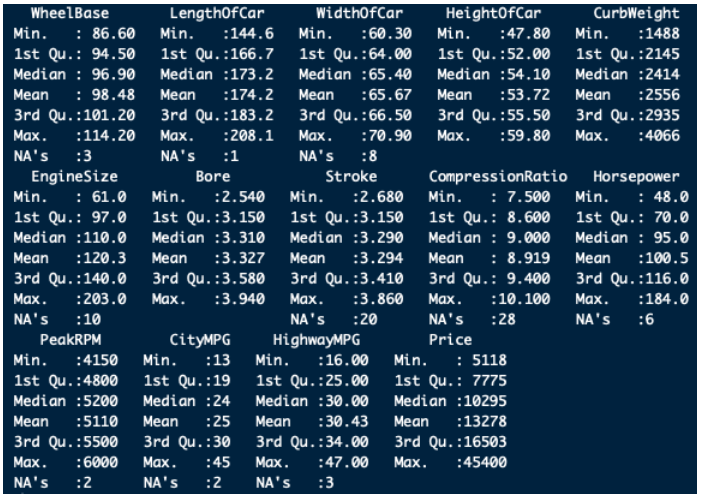


Figure 7: Mahalanobis Robust Distance

3.3.5 Outlier Treatment

NA values were assigned to the detected outliers and treated as if they were missing. Again, using Random Forest, missing values were estimated. Thus, outlier values have been treated.



```
   WheelBase       LengthOfCar      WidthOfCar      HeightOfCar      CurbWeight
 Min.   : 86.60   Min.   :144.6   Min.   :60.30   Min.   :47.80   Min.   :1488
 1st Qu.: 94.50   1st Qu.:166.7   1st Qu.:64.00   1st Qu.:52.00   1st Qu.:2145
 Median : 96.90   Median :173.2   Median :65.40   Median :54.10   Median :2414
 Mean   : 98.48   Mean   :174.2   Mean   :65.67   Mean   :53.72   Mean   :2556
 3rd Qu.:101.20   3rd Qu.:183.2   3rd Qu.:66.50   3rd Qu.:55.50   3rd Qu.:2935
 Max.   :114.20   Max.   :208.1   Max.   :70.90   Max.   :59.80   Max.   :4066
 NA's   :3        NA's   :1       NA's   :8
   EngineSize          Bore           Stroke       CompressionRatio   Horsepower
 Min.   : 61.0   Min.   :2.540   Min.   :2.680   Min.   : 7.500   Min.   : 48.0
 1st Qu.: 97.0   1st Qu.:3.150   1st Qu.:3.150   1st Qu.: 8.600   1st Qu.: 70.0
 Median :110.0   Median :3.310   Median :3.290   Median : 9.000   Median : 95.0
 Mean   :120.3   Mean   :3.327   Mean   :3.294   Mean   : 8.919   Mean   :100.5
 3rd Qu.:140.0   3rd Qu.:3.580   3rd Qu.:3.410   3rd Qu.: 9.400   3rd Qu.:116.0
 Max.   :203.0   Max.   :3.940   Max.   :3.860   Max.   :10.100   Max.   :184.0
 NA's   :10                      NA's   :20      NA's   :28       NA's   :6
    PeakRPM         CityMPG      HighwayMPG         Price
 Min.   :4150   Min.   :13   Min.   :16.00   Min.   : 5118
 1st Qu.:4800   1st Qu.:19   1st Qu.:25.00   1st Qu.: 7775
 Median :5200   Median :24   Median :30.00   Median :10295
 Mean   :5110   Mean   :25   Mean   :30.43   Mean   :13278
 3rd Qu.:5500   3rd Qu.:30   3rd Qu.:34.00   3rd Qu.:16503
 Max.   :6000   Max.   :45   Max.   :47.00   Max.   :45400
 NA's   :2      NA's   :2    NA's   :3
```

Figure 8: NA Values instead of Outliers

If you notice the Figure 9, the price variable still has outliers. The reason for this is the availability of expensive cars. Therefore, we did not accept and treat values exceeding the limit in the price variable as outliers.

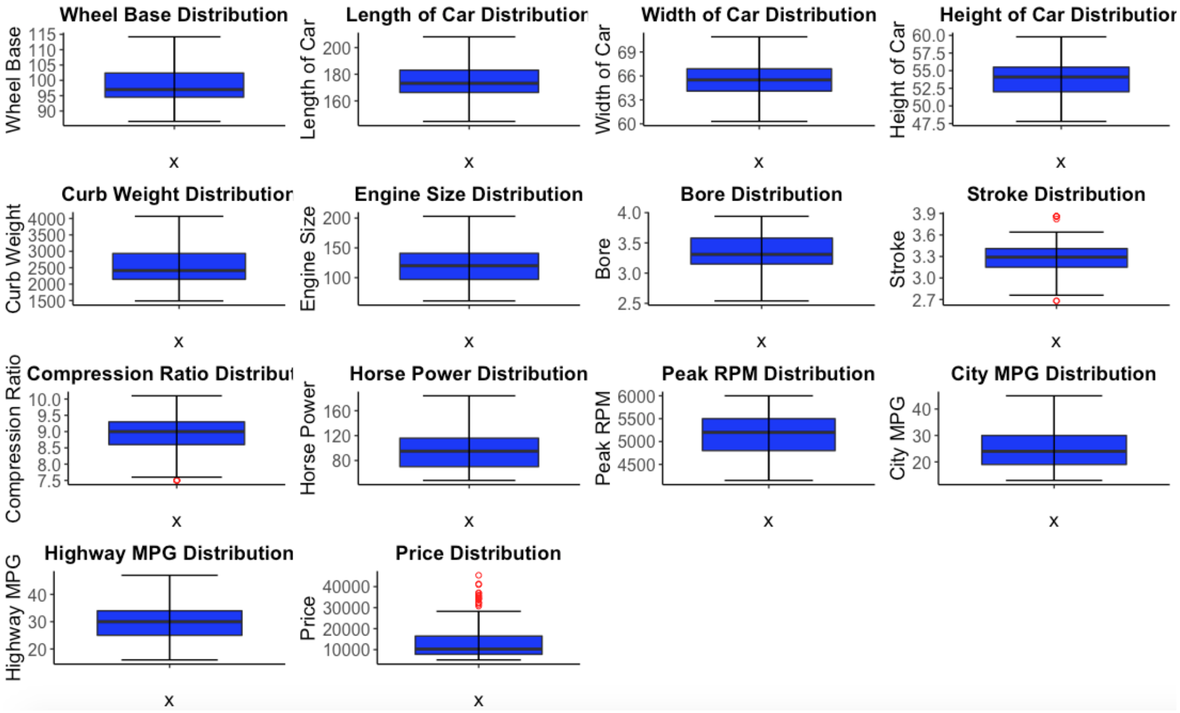


Figure 9: After Outlier Treatment

4 Descriptive Analytics

We used some of methods such as visualization, clustering, correspondence analysis, and multiple correspondence analysis etc for the interpretation of latent concepts.

4.1 Visualization

In this section, our cleaned data set will be visualized by means of pie charts, box plots, histograms. Moreover, we can examine how variables are distributed and decide which variables will be used in prediction analytics. Finally, the relationship will be observed between independent variables (continuous/categorical) and dependent variable (price).

When we consider all pie charts plots (i.e Figure 10, Figure 11, Figure 12, Figure 13), we can say that companies have broad distribution. On the other hand, Engine Location has narrow distribution and it will be excluded from our data set. We could not observe any problem about other variables.

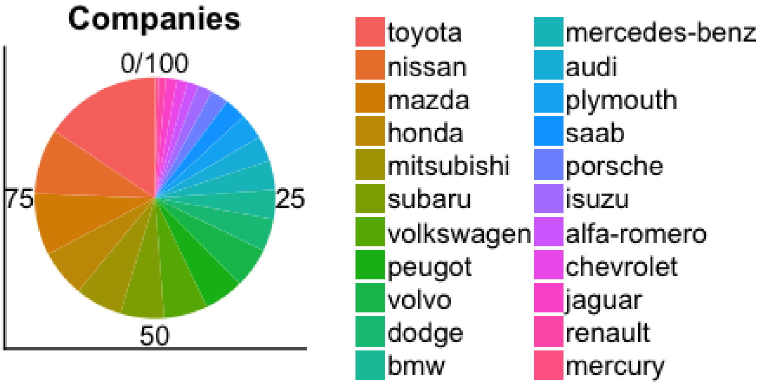


Figure 10: Pie Chart for Categorical Variables

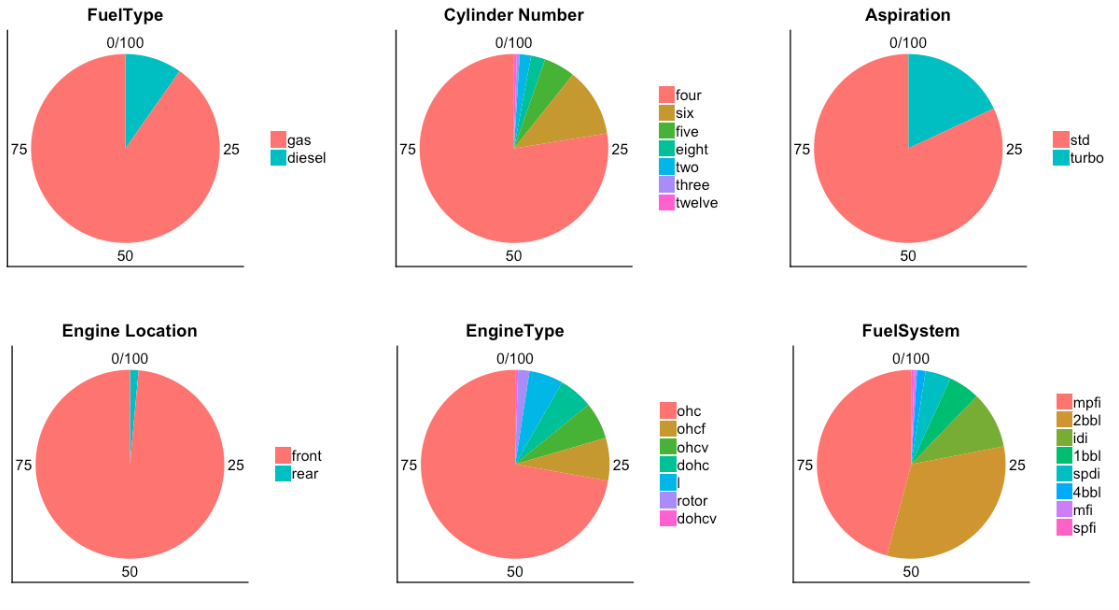


Figure 11: Pie Chart for Categorical Variables

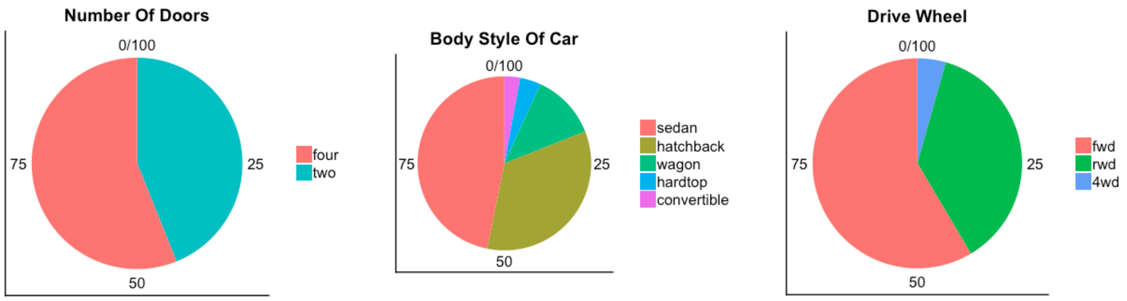


Figure 12: Pie Chart for Categorical Variables 2

How does the categorical variables impact price ? To answer this question we visualized relationship between them by Box plot.

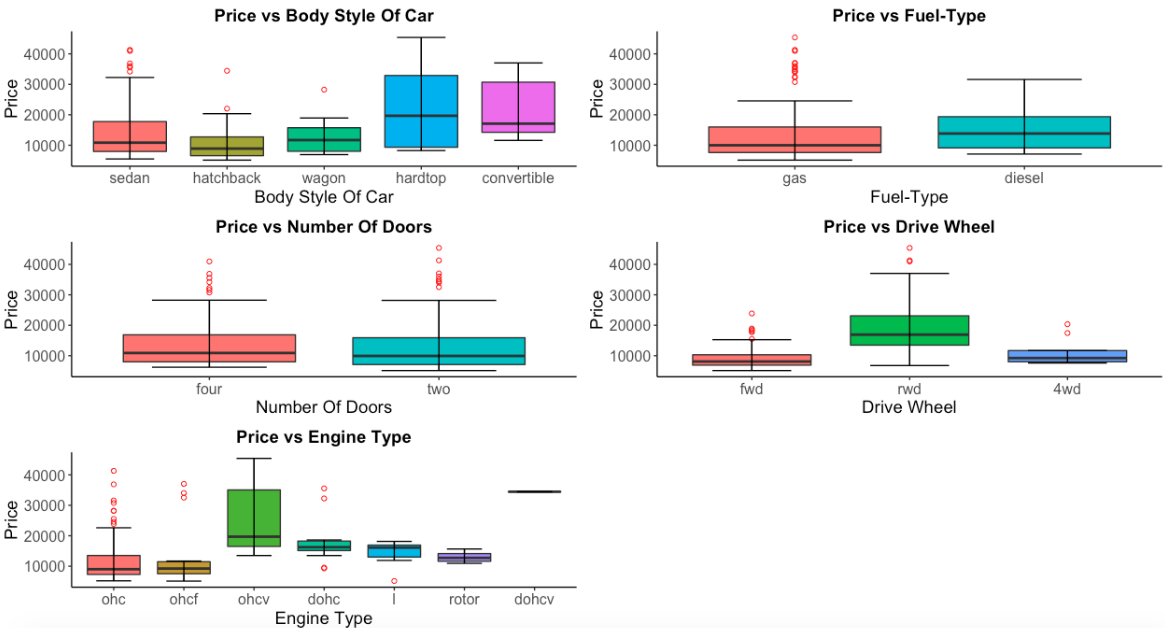


Figure 13: Bivariate description for Categorical Variables

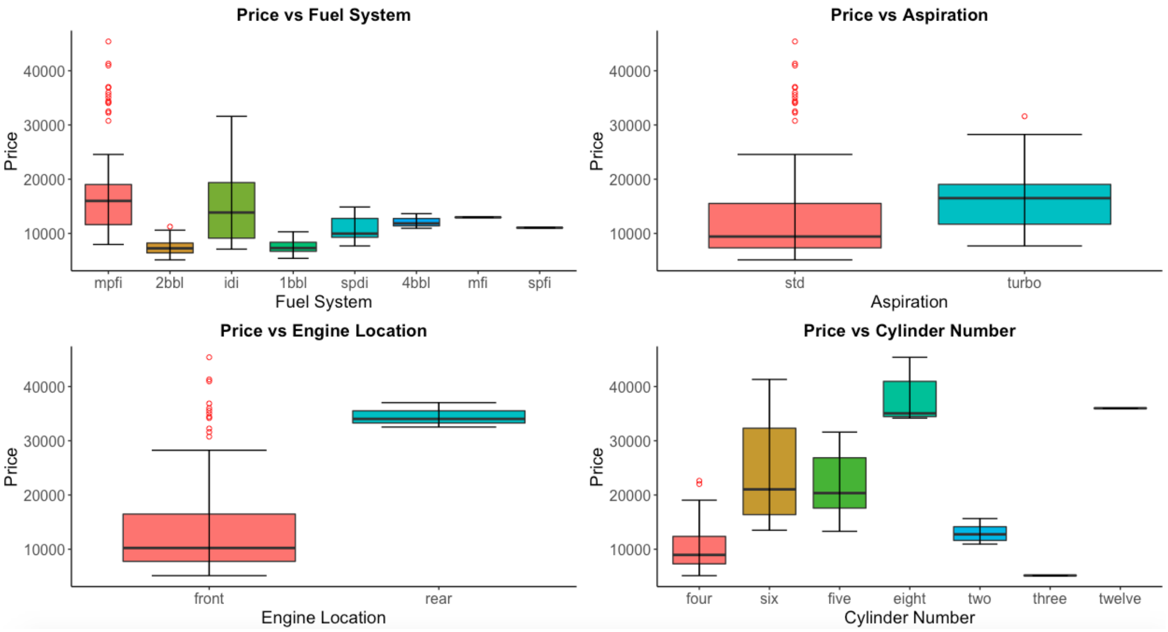


Figure 14: Bivariate description for Categorical Variables 2

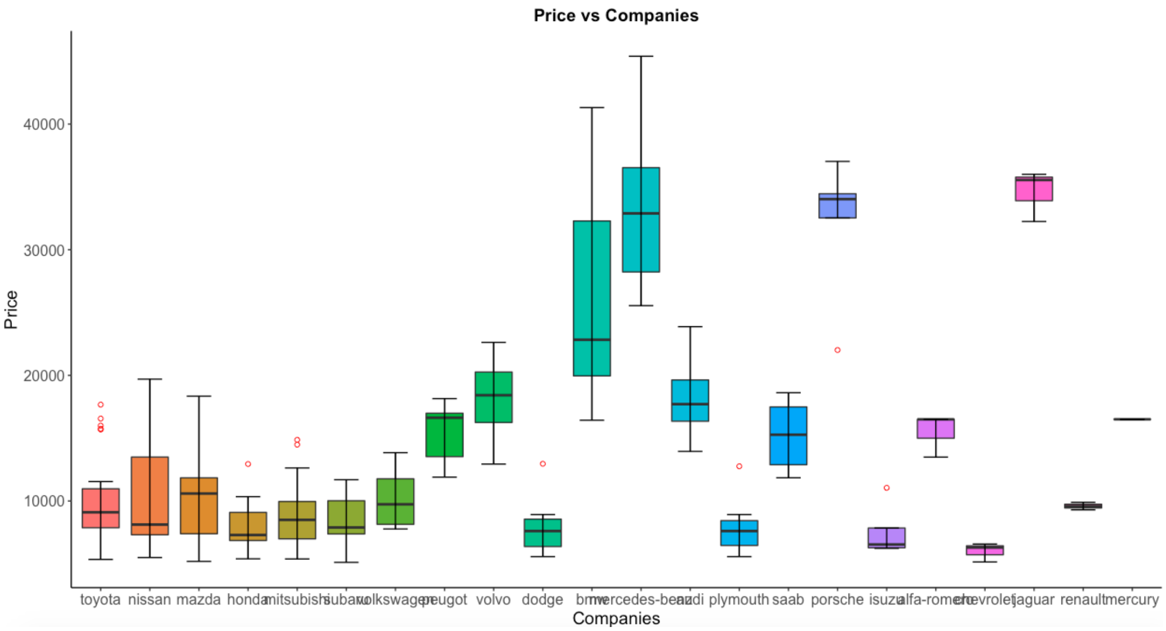


Figure 15: Bivariate description for Companies

When Figures 14-15-16 are examined, many comments can be made. To exemplify, BMW - Jaguar and Mercedes Benz have high price cars. In addition, the cars which have turbo aspiration are more expensive than the cars which have std aspiration. Finally, the cars have four cylinders are most cheapest.

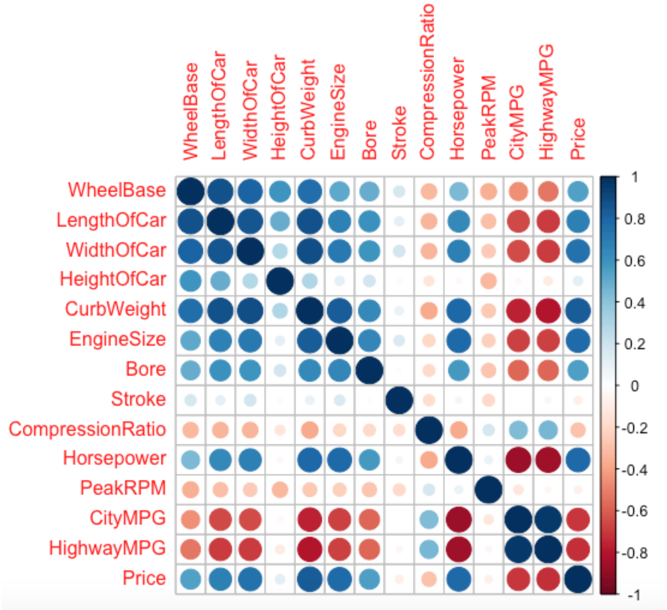


Figure 16: Correlations between continuous variables

The table above shows the correlation between the continuous variables. There is a strong correlation between other variables except Stroke. It will be excluded from our data set.

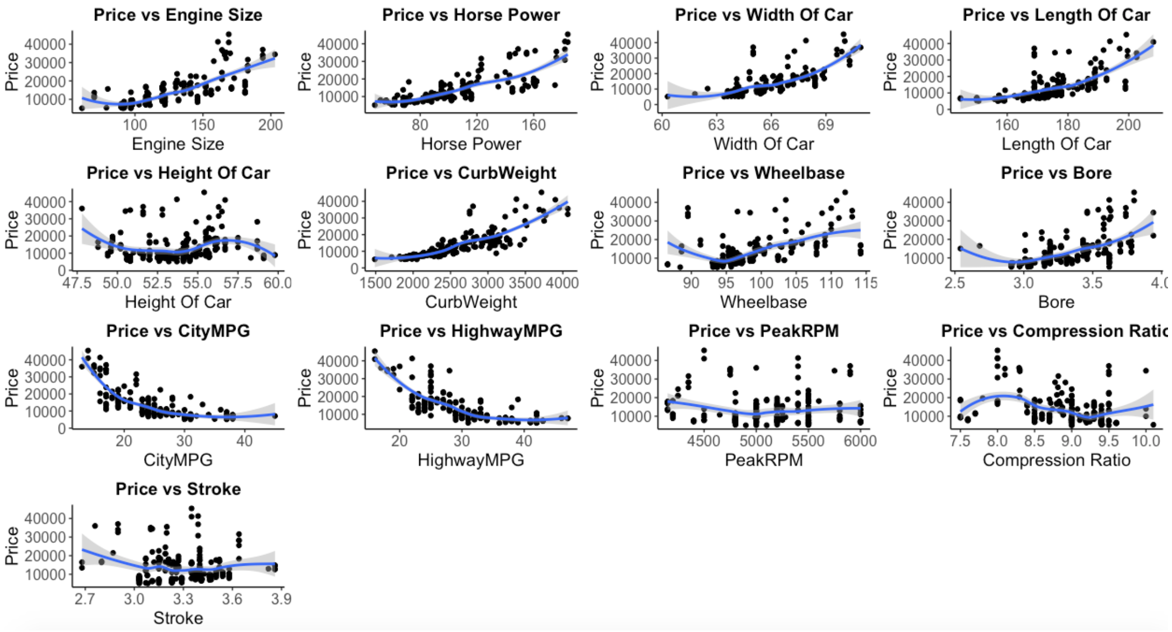


Figure 17: Bivariate description for Continuous Variables

When we examine Figure 21, we can see some of values far from our nonlinear line and consider if they are outlier or not. The reason for this is that, as we mentioned before, we did not consider the values in the price as outlier because of the fact that expensive cars may not be outlier.

4.2 Principal Component Analysis

To discover the latent concepts, we have performed the Principal Component Analaysis over the numerical attributes in our data set (with no significance attributes removed).

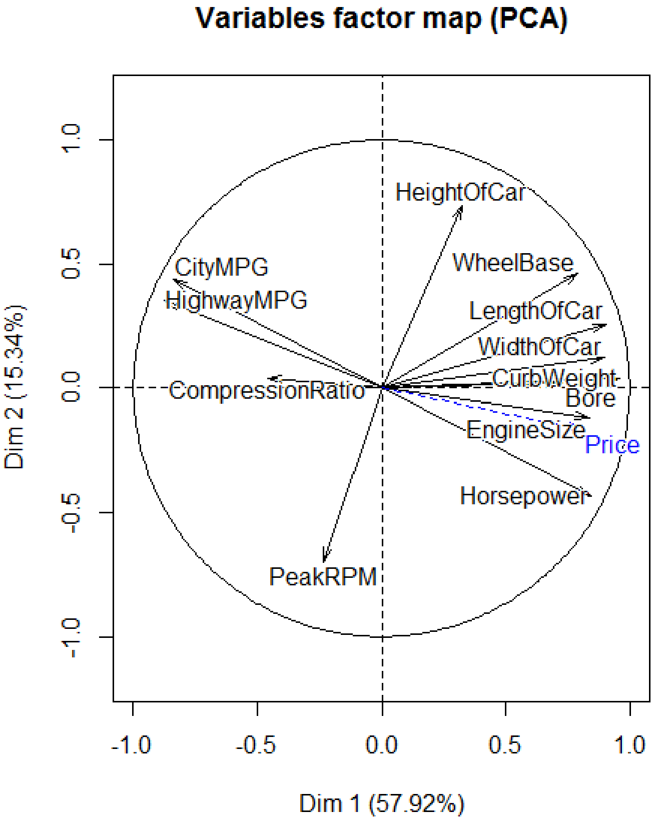


Figure 18: Variables Factor Map

The principal component analysis is performed over the supplementary variable Price (marked in blue). In the variables factor map we can see the contribution of the variables to the principal components. We see that the first dimension explains a very big part (58%) of the information and the second dimension a 16%.

We can also observe that the variables more related to the first dimension are HeightOfCar(in the first quadrant) and PeakRPM(in the third quadrant), because their projection into the first dimension axis is very high. By the other hand, the variables CurbWeight, Bore, EngineSize, HorsePower, WidthOfCar, LengthOfCar and WheelBase (in the first and fourth quadrant) and CityMPG and HighwayMPG(in the third quadrant) has very high value of projection over the second dimension.

As we can also see, our supplementary variable Price is located in the fourth quadrant, with high value of projection over the second dimension. We can observe that EnginSize seems to be the most correlated variable with Price, because their direction and projection are really close.

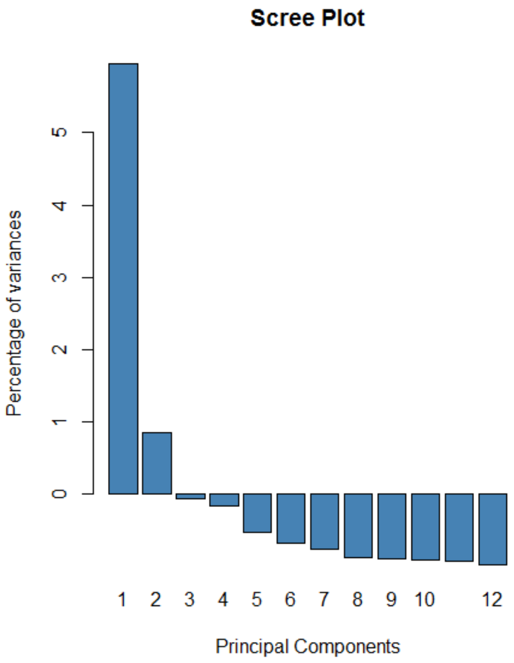


Figure 19: Barplot of Eigen values of each dimension respect to the average of Eigen values

To see the significant dimensions we have computed the difference between the eigen values of each dimension with the average eigen value. As we can observe in the chart, the only significance eigenvalues (over 0) are the first two dimension. We can say the significant dimension is 2.

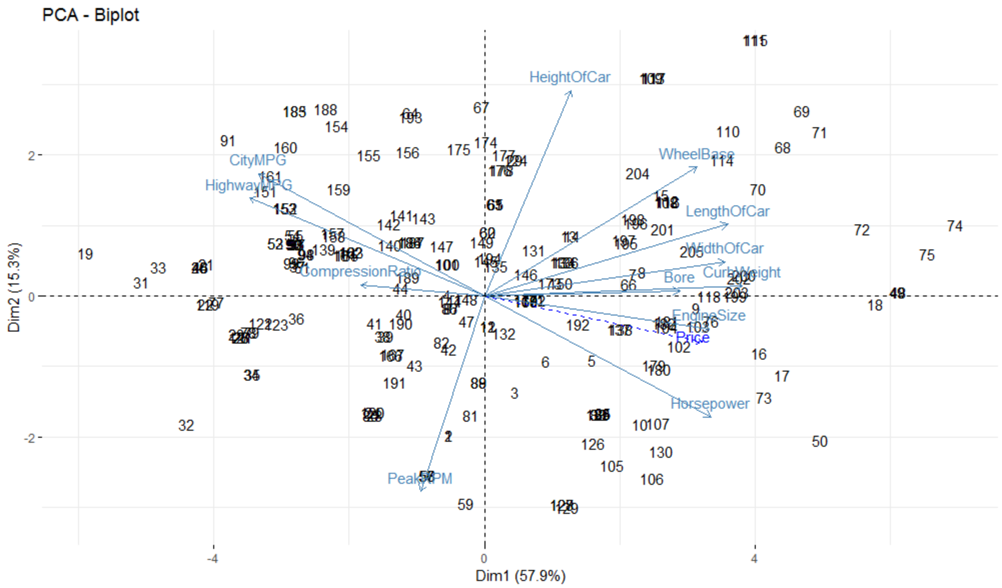


Figure 20: Individuals Factor Map

In the figure above we can observe the projection of the individuals, it shows the contribution of these individuals to the principal components together with the variables. The identifiers of the cars does not gives us enough explanatory significance, but we can interpret that individuals located closed to an arrow of a variable has high value in that attribute and contributes more.

4.3 Clustering

We are going to perform clustering over our data set to observe what kind of clusters it exists in the data, that is which individuals shares same features and what differences are there between individuals in different in clusters.

Taking into account that in our data set we have mixed type of variables, we are going to use k-means as clustering method.

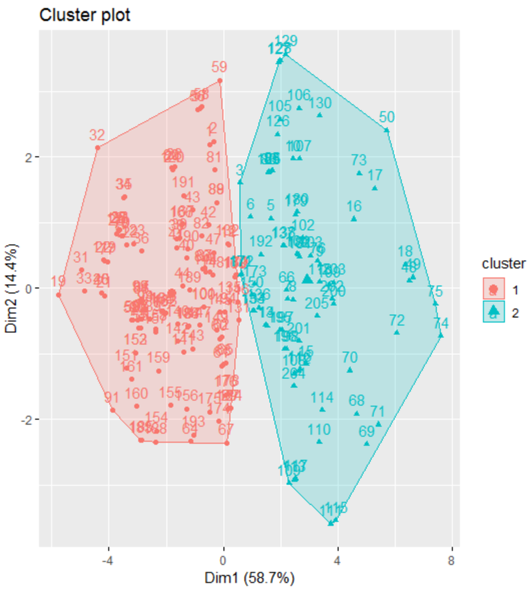


Figure 21: Pruned Tree

In the picture above we can observe the clusters obtained. We have obtained two clusters separated by the dimension 1, which is the most contributing dimension. However we cannot extract any information about the individuals.

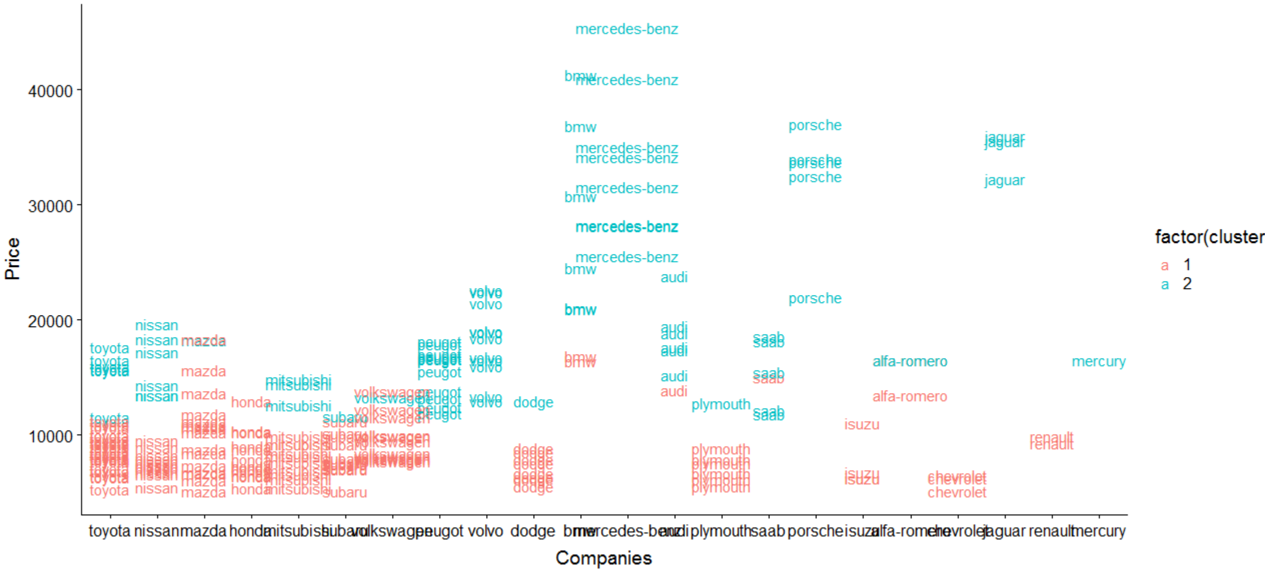


Figure 22: Pruned Tree

Now in this figure we can see the brand of the cars distributed, we can see cars like bmw, mercedes-benz, are clearly separated to mazda, honda, volswaggen, etc, which makes us think their price are quite different. We can see there are some brands repeated (some of BWM in class blue and some in red), this might be the cars with different capacities, blue ones seems to be the more powerful and red the less. This clustering makes us thinks that the clusters means more luxury or cars with more potential or power and cars weakers.

5 Data Partitioning

Thanks to createDataPartition function, a new index is created. Eighty percent of the data was called the train, the rest was called the test set.

6 Predictive Analytics

We are going to build predictive analytics to perform learning the dataset model and prediction of the price of the cars so that we can achieve our objective that is the prediction of the price of the cars taking into account all the attriutes.

6.1 Decision Tree

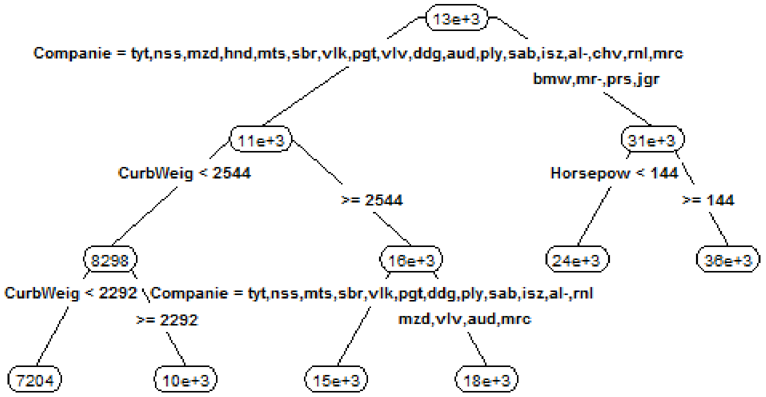


Figure 23: Pruned Tree

We can see our first model of decision tree using variable price as supplementary variable, over this tree we are going to perform a pruning to obtain the optimal tree.

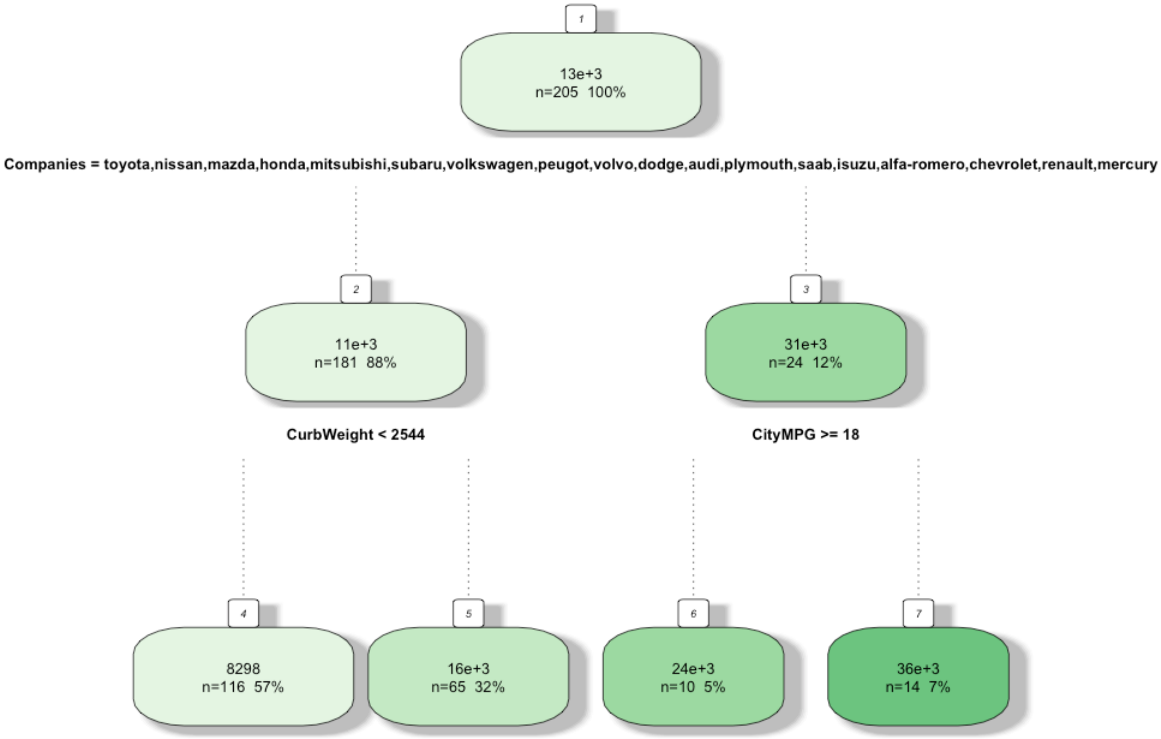

1
13e+3
n=205 100%
Companies = toyota,nissan,mazda,honda,mitsubishi,subaru,volkswagen,peugot,volvo,dodge,audi,plymouth,saab,isuzu,alfa-romero,chevrolet,renault,mercury
2
11e+3
n=181 88%
3
31e+3
n=24 12%
CurbWeight < 2544
CityMPG >= 18
4
8298
n=116 57%
5
16e+3
n=65 32%
6
24e+3
n=10 5%
7
36e+3
n=14 7%

Figure 24: Pruned Tree

From the tree we can obtain rules such as cars of the class as toyota, nissan, mazda, and so on with CurWeight smaller than 2544 should be in the class of cheap cars (with a price closed to 8298).

We can also see the percentage of cars in different classes. The "best" ones with 7% only.

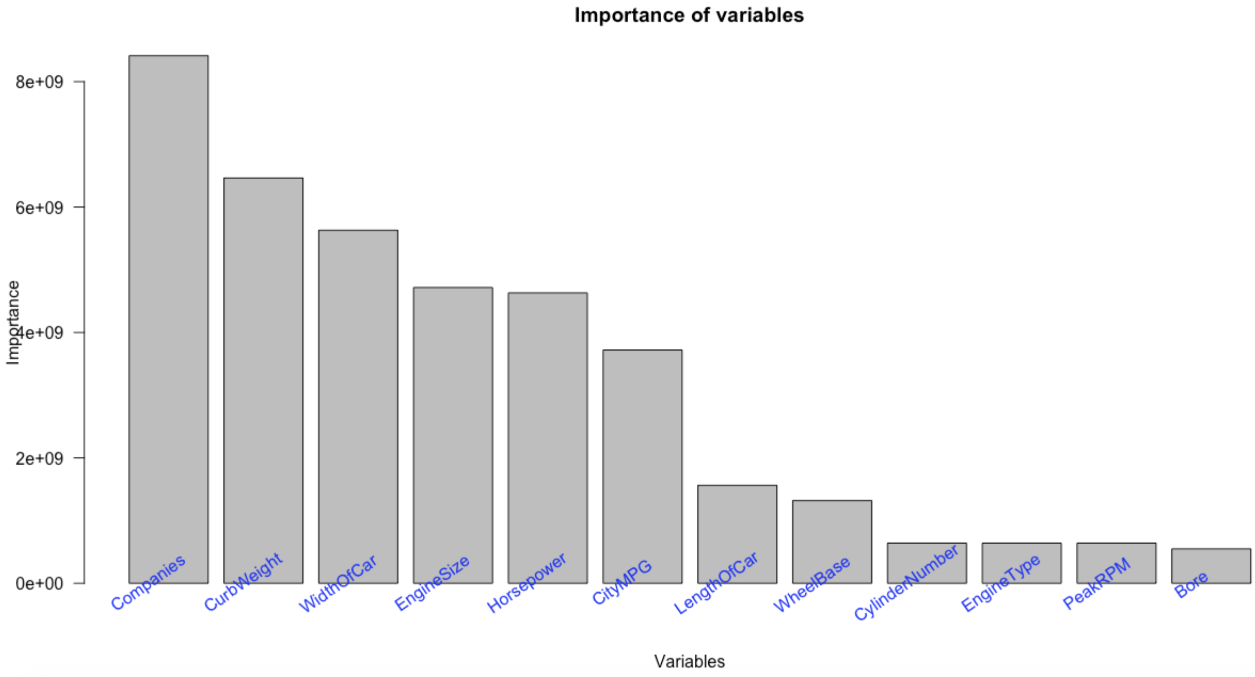


Figure 25: Importance Variable of Data Set

In the figure above we can see the importance of the variables in the decision of the price of a car. The most decisive variables is the company, that completely make sense because famous cars such as BMW, Benz, Audi are famous because of the brand. CurbWeight, WidthOfCar, EnineSize and Horsepower has also a clearly difference respect to other variables, because they explain the capacity of the cars, with best capacity, more expensive is a car, that completely make sense.

7 Conclusions

Car price estimation was made by analyzing automobile data set with the methods used in multivariate data analysis course. With the analysis we observed that our data can be splitted into two clusters, mainly the most luxury and potential cars respect to the less luxury and weaker cars. With the prediction model we can obtain the conclusion that the most relevant variables that determine a car's price is the company or brand, more than the variables of capacity, although these also affects.