**About the dataset:**

- Toyota Corolla dataset has 39 columns and 1436 rows in total. We saw that most of the data types are binary and a few of them contain numeric and categorical variables.
- No missing values were found in the dataset, so there was no need for omission or imputation to be done in the dataset before the analysis.
- In order to do the summary statistics and to ensure appropriate handling of the dataset in the modeling, we converted the three categorical variables namely Fuel_Type, Color and Model into factors, where each character is converted into levels.

**An overview of the summary statistics:**

- A wide range of car values is shown by the price, which ranges from €4,350 to €32,500 with an average of about €10,731.
- Age & KM: Most of the cars are 56 months old and have about 68,000 kilometers on them on an average.
- Fuel Type: Most of the vehicles in the datasets are petrol-powered, followed by diesel and a few CNG vehicles.

**Visual Perspectives**:

- Fuel Type Distribution: Petrol has highest count in the fuel type, according to a bar chart.
- Price by Fuel Type: If we see the boxplot, compared to gasoline or CNG, diesel vehicles typically have higher median prices and greater variability.
- Correlation: There are many significant relationships between variables. For example, Age_08_04 has a strong positive with KM at 0.87, possibly meaning that older cars have more kilometers and vice versa, or HP has a strong positive with quarterly_tax at 0.74 as more horse powers, more taxes got to pay. Furthermore, Age_08_04 has a strong negative with guarantee_period at -0.76 as new cars have longer guarantee period while old ones have less period. Moreover Mfg_year and price also show strong correlation of 0.89.
- For the plot distributions, we can clearly see that most distributions have only 2 columns at the values of 0 or 1 due to binary type. Only some plots that contain numeric data type have a distribution. For example, KM, Price, and Age_08_04. KM and Price have the right skew distribution while Age_08_04 has a left skew distribution. All of them have outliers either at the end of the left or at the end of the right.

***Fig1: Summary Statistics***

```
       Id                                                  Model        Price          Age_08_04
 Min.   :   1.0   TOYOTA Corolla 1.6 16V HATCHB LINEA TERRA 2/3-Doors:109   Min.   : 4350   Min.   : 1.00
 1st Qu.: 361.8   TOYOTA Corolla 1.3 16V HATCHB LINEA TERRA 2/3-Doors: 84   1st Qu.: 8450   1st Qu.:44.00
 Median : 721.5   TOYOTA Corolla 1.6 16V LIFTB LINEA LUNA 4/5-Doors  : 80   Median : 9900   Median :61.00
 Mean   : 721.6   TOYOTA Corolla 1.6 16V LIFTB LINEA TERRA 4/5-Doors : 71   Mean   :10731   Mean   :55.95
 3rd Qu.:1081.2   TOYOTA Corolla 1.4 16V VVT I HATCHB TERRA 2/3-Doors: 54   3rd Qu.:11950   3rd Qu.:70.00
 Max.   :1442.0   TOYOTA Corolla 1.6 16V SEDAN LINEA TERRA 4/5-Doors : 43   Max.   :32500   Max.   :80.00
                  (Other)                                           :995
   Mfg_Month        Mfg_Year          KM          Fuel_Type          HP            Met_Color          Color
 Min.   : 1.000   Min.   :1998   Min.   :     1   CNG   :  17   Min.   : 69.0   Min.   :0.0000   Grey   :301
 1st Qu.: 3.000   1st Qu.:1998   1st Qu.: 43000   Diesel: 155   1st Qu.: 90.0   1st Qu.:0.0000   Blue   :283
 Median : 5.000   Median :1999   Median : 63390   Petrol:1264   Median :110.0   Median :1.0000   Red    :278
 Mean   : 5.549   Mean   :2000   Mean   : 68533                 Mean   :101.5   Mean   :0.6748   Green  :220
 3rd Qu.: 8.000   3rd Qu.:2001   3rd Qu.: 87021                 3rd Qu.:110.0   3rd Qu.:1.0000   Black  :191
 Max.   :12.000   Max.   :2004   Max.   :243000                 Max.   :192.0   Max.   :1.0000   Silver :122
                                                                                                  (Other): 41
   Automatic            CC            Doors          Cylinders     Gears        Quarterly_Tax       Weight
 Min.   :0.00000   Min.   : 1300   Min.   :2.000   Min.   :4   Min.   :3.000   Min.   : 19.00   Min.   :1000
 1st Qu.:0.00000   1st Qu.: 1400   1st Qu.:3.000   1st Qu.:4   1st Qu.:5.000   1st Qu.: 69.00   1st Qu.:1040
 Median :0.00000   Median : 1600   Median :4.000   Median :4   Median :5.000   Median : 85.00   Median :1070
 Mean   :0.05571   Mean   : 1577   Mean   :4.033   Mean   :4   Mean   :5.026   Mean   : 87.12   Mean   :1072
 3rd Qu.:0.00000   3rd Qu.: 1600   3rd Qu.:5.000   3rd Qu.:4   3rd Qu.:5.000   3rd Qu.: 85.00   3rd Qu.:1085
 Max.   :1.00000   Max.   :16000   Max.   :5.000   Max.   :4   Max.   :6.000   Max.   :283.00   Max.   :1615

 Mfr_Guarantee    BOVAG_Guarantee  Guarantee_Period       ABS            Airbag_1         Airbag_2
 Min.   :0.0000   Min.   :0.0000   Min.   : 3.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.: 3.000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000
 Median :0.0000   Median :1.0000   Median : 3.000   Median :1.0000   Median :1.0000   Median :1.0000
 Mean   :0.4095   Mean   :0.8955   Mean   : 3.815   Mean   :0.8134   Mean   :0.9708   Mean   :0.7228
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 3.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :36.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000

     Airco        Automatic_airco   Boardcomputer      CD_Player       Central_Lock     Powered_Windows
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
 Median :1.0000   Median :0.00000   Median :0.0000   Median :0.0000   Median :1.0000   Median :1.000
 Mean   :0.5084   Mean   :0.05641   Mean   :0.2946   Mean   :0.2187   Mean   :0.5801   Mean   :0.562
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.000
 Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000

 Power_Steering       Radio          Mistlamps       Sport_Model     Backseat_Divider  Metallic_Rim
 Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
 Median :1.0000   Median :0.0000   Median :0.000   Median :0.0000   Median :1.0000   Median :0.0000
 Mean   :0.9777   Mean   :0.1462   Mean   :0.257   Mean   :0.3001   Mean   :0.7702   Mean   :0.2047
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000

 Radio_cassette   Parking_Assistant     Tow_Bar
 Min.   :0.0000   Min.   :0.000000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.0000
 Median :0.0000   Median :0.000000   Median :0.0000
 Mean   :0.1455   Mean   :0.002786   Mean   :0.2779
 3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.000000   Max.   :1.0000
```

*Fig2: Fuel Type Bar Plot*

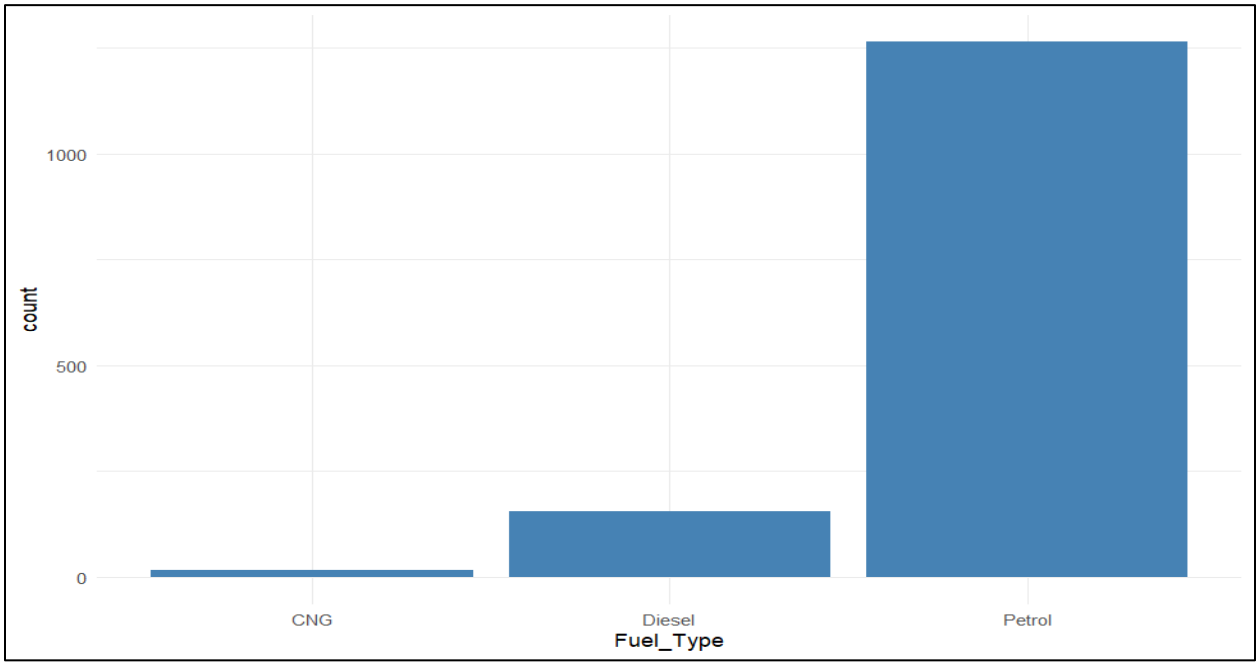



*Fig3: Price vs Fuel Type Boxplot*

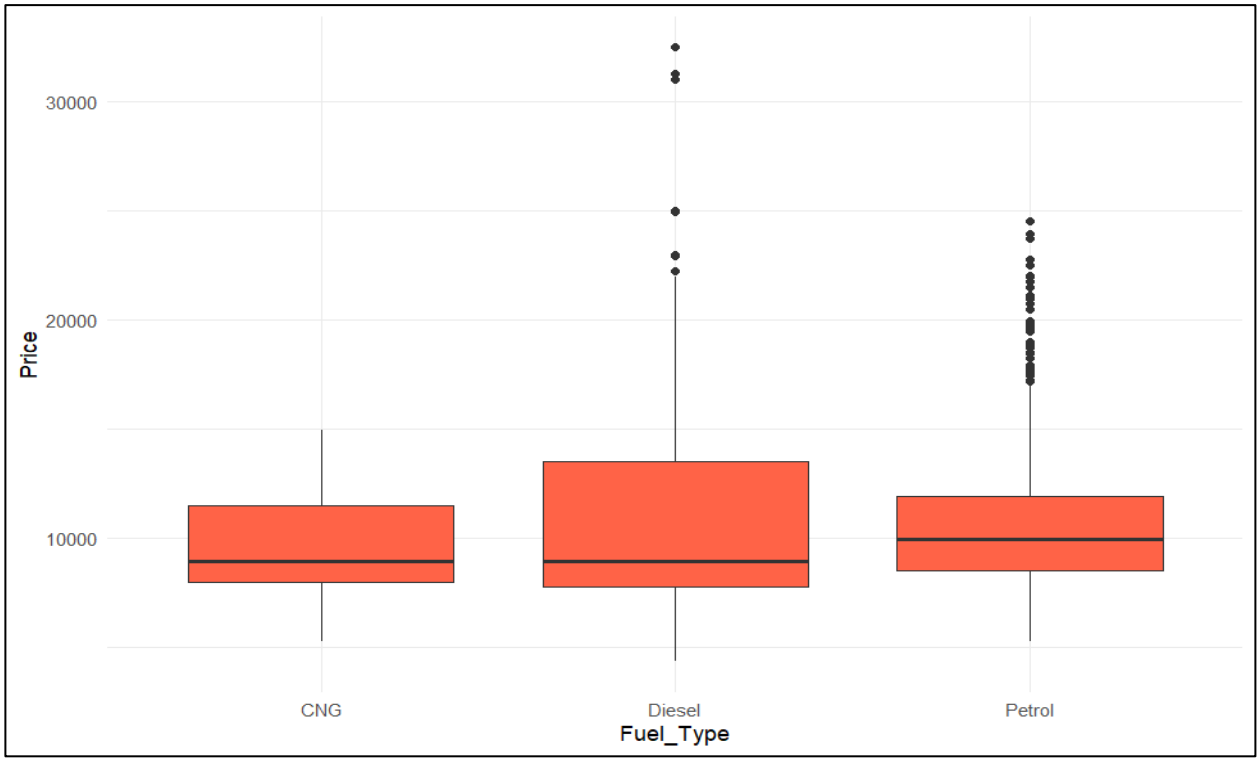


*Fig4: Heatmap*

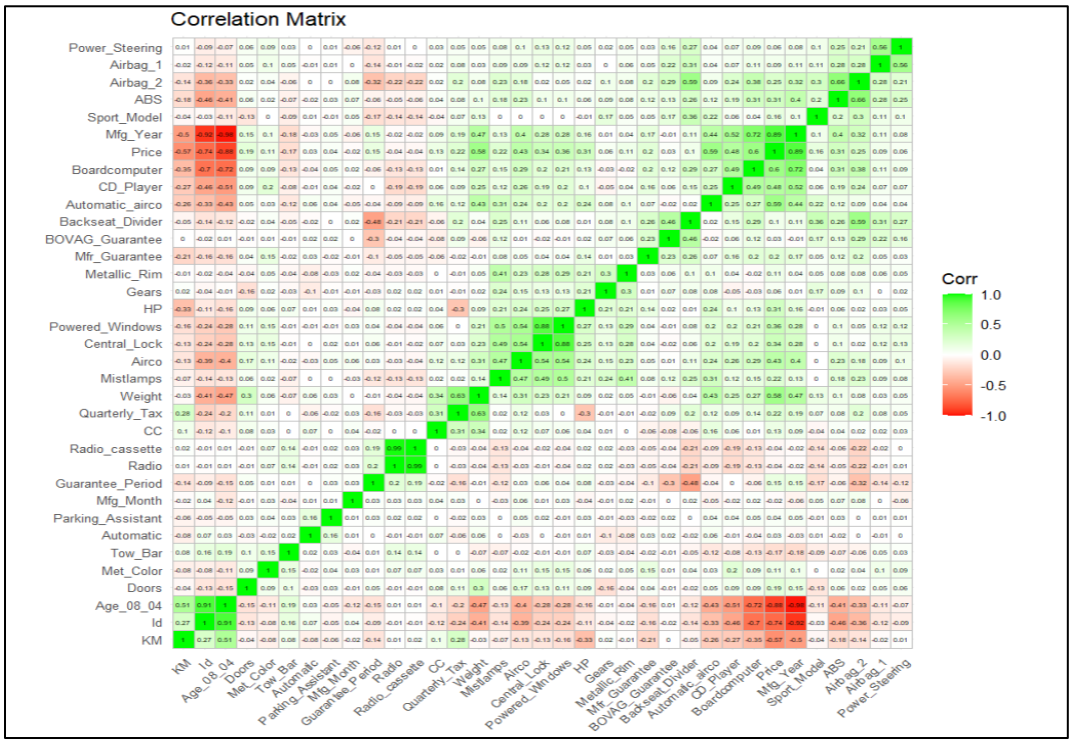



*Fig5: Histogram*

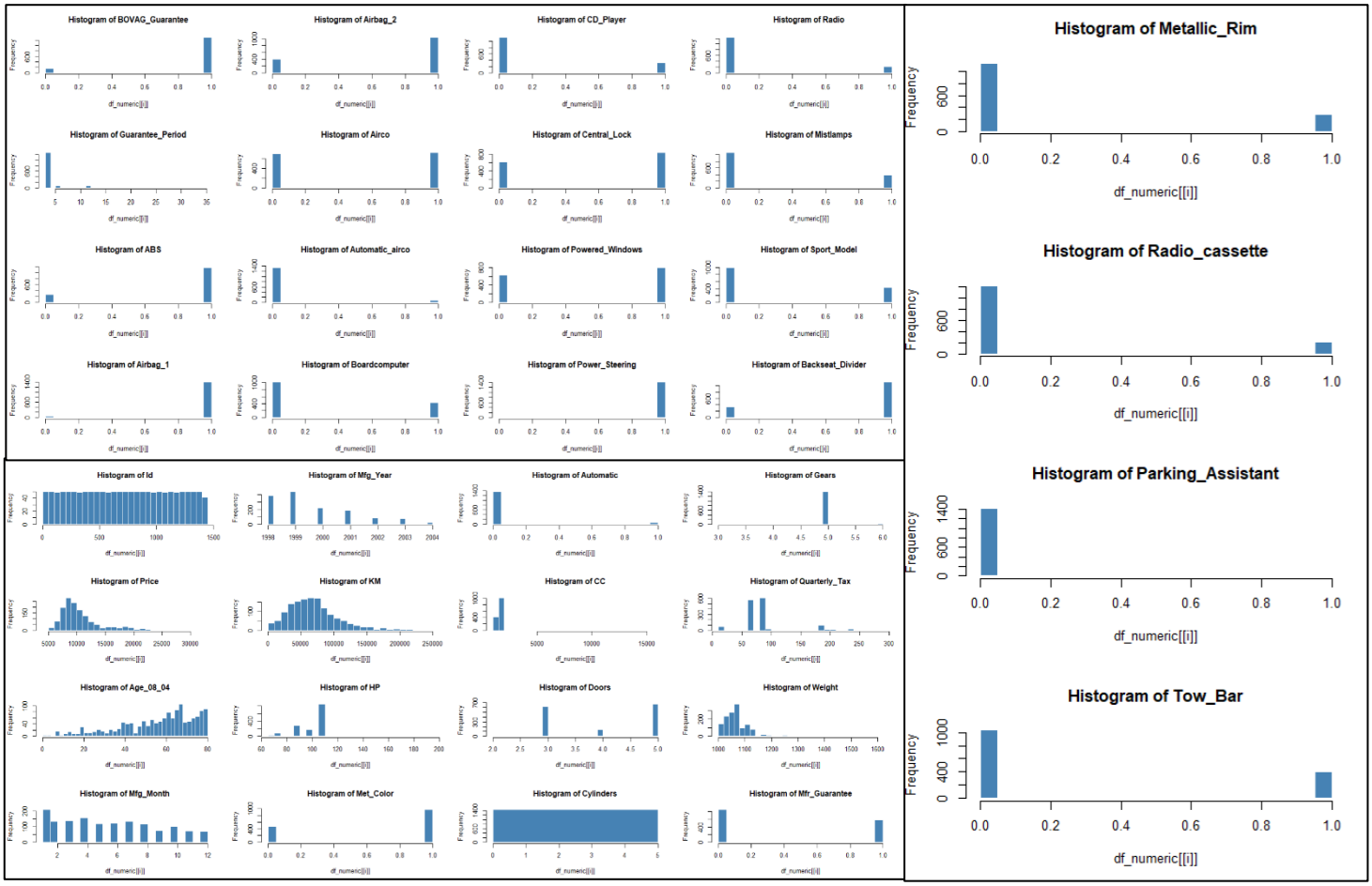
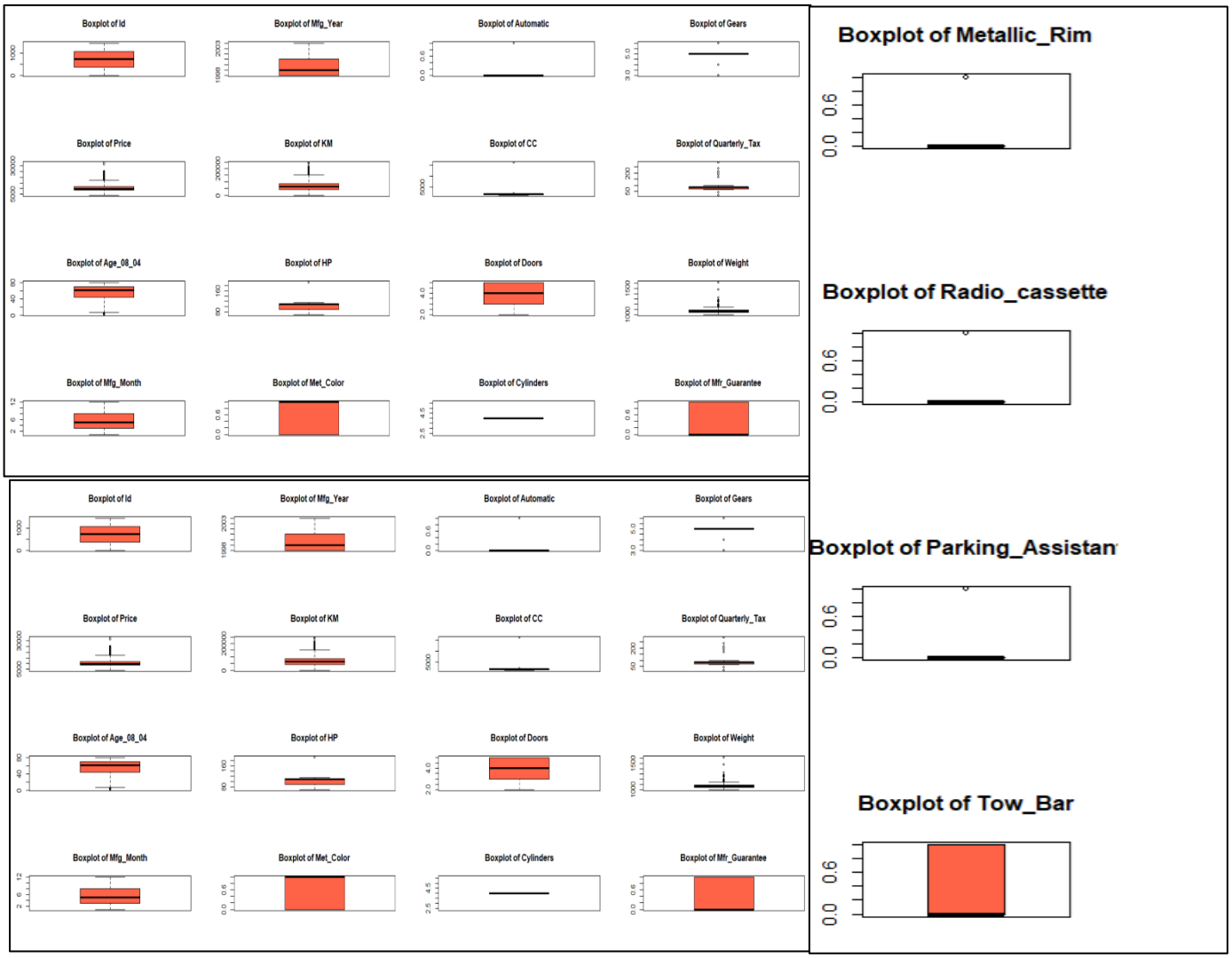


*Fig6: Boxplot*




**2.**

In this part, we have split the dataset into 60% training and 40% validation sets using seed 16. A multiple linear regression was built using 16 predictors to understand which ones are explaining the variability in target variable at its best. The model achieved the R -squared value of 88%, which means that the predictors explain 88% variation in car prices. Most of the predictors were found to be statistically significant with low p-values and high t-values.

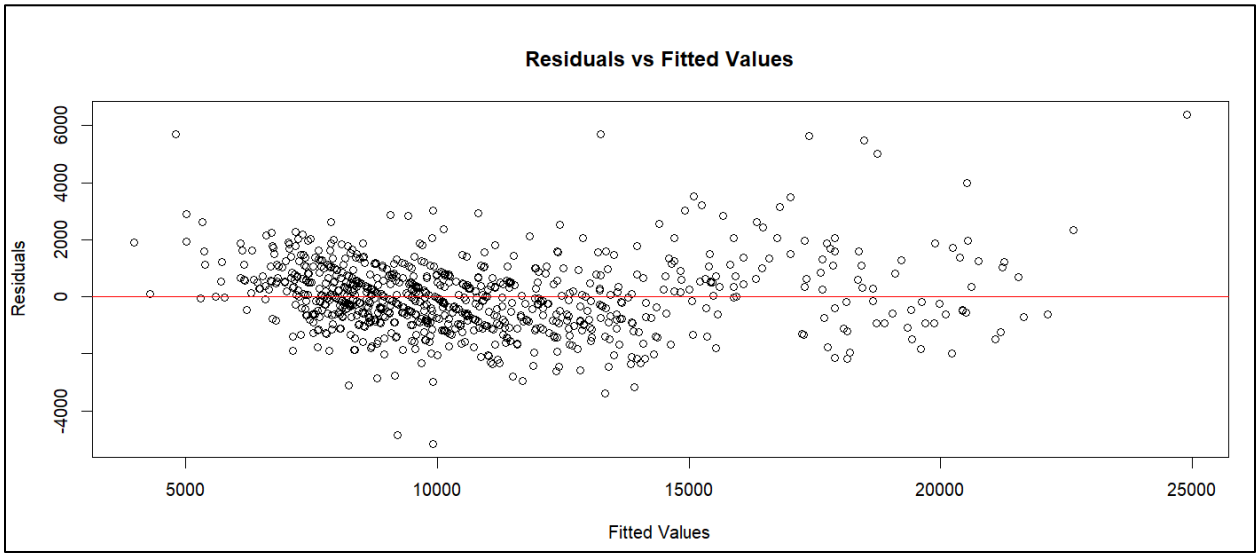
**3.**

**a.** Top 4 most important car specifications for the car's price are Age_08_04, KM, Automatic_airco, and HP. We can identify that by looking at the largest absolute t values and lowest p values for each variable. For example, Age_08_04 has a t value of -30.918 and p value is smaller

than 2e-16 or Automatic_airco has t value of 13.524 and p value is also smaller than 2e-16. Same thing with other variables.

**b.** Based on the visual assessment, residuals seem to spread more as fitted values increase, leading to not perfectly constant variance. This means that the data appear heteroscedastic. Using Breusch-Pagan test, BP = 154, df = 16, and p value is almost zero, showing the evidence of heteroscedastic.

***Fig7: Plot for Residuals***




**c.** RMSE and MAE are very close to the training set, indicating that the model accuracy is good and shows no sign of overfitting. Additionally, ME and MPE are close to zero, meaning that there is no significant bias in over prediction or under prediction. MAPE also shows strong accuracy.

**d.** Accuracy of complex models is more as the metrics RMSE, MAE, MAPE are lower as compared to simpler models. But the simpler model is easier to interpret due to a smaller number of predictors. So, if our goal is to maximize predictive accuracy, we should go with the complex model and if we want a more interpretable model, the simpler one is good to go for instead.

Our recommendation would be to go for Model with 16 predictors (complex model) as it has higher R-squared and adjusted R-squared value with feature richness. Higher accuracy and lower error metrics i.e., RMSE, MAE and MAPE compared to simple model. As the model training and validation errors are consistent, it shows that there is no overfitting when including large number of predictors.