**Multi Linear Regression**

**Problem Statement:** – To predict the price of a used Toyota Corolla considering the only features as stated in business problem.

Consider only the below columns and prepare a prediction model for predicting Price.

“Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight"

**Dataset**: – ToyotaCorolla

**Algorithm Used** – Multi Linear Regression Model

**Assumptions made while preparing a Multi linear regression is that:**

1. All features have a linear relationship with the outcome or output variable (L).
2. Features should not be dependent on each other, if so the phenomenon of multi-collinearity will exist and there by impacting the o/p prediction or estimates.
3. Multiple regression assumes that the residuals are normally distributed.
4. This assumption states that the variance of error terms are similar across the values of the independent variables.

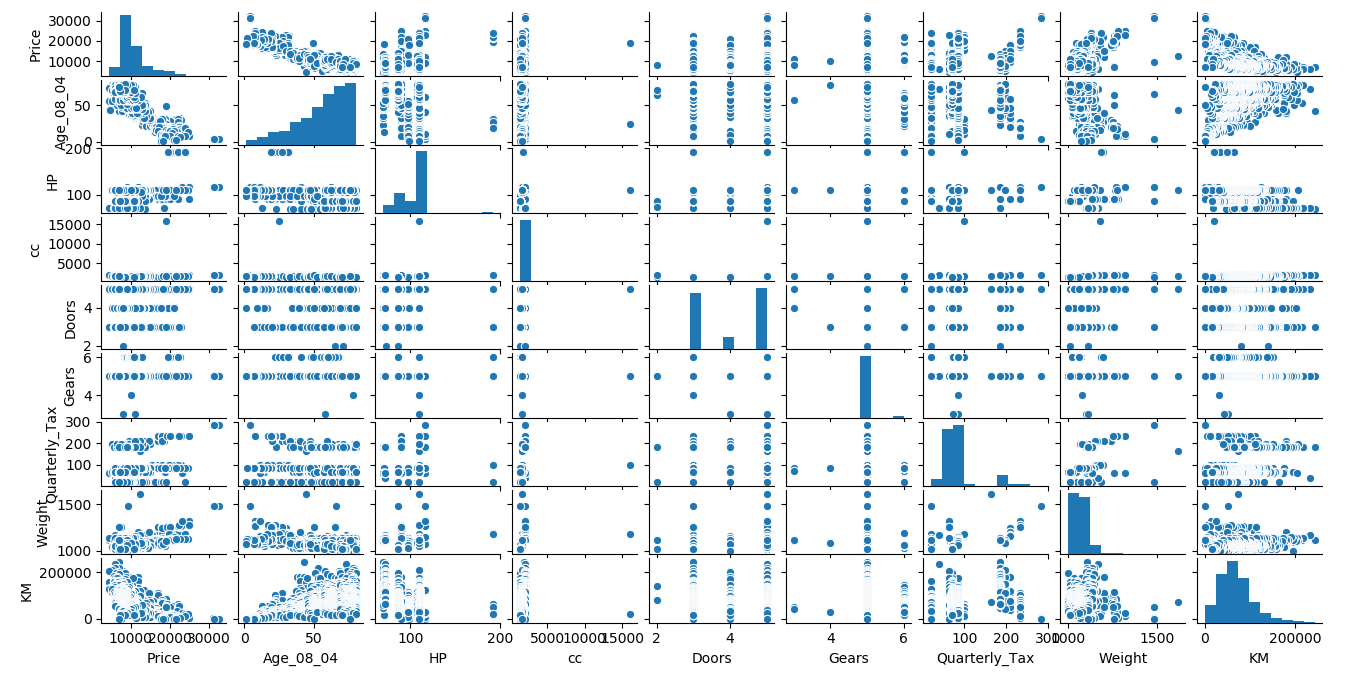
**Feature Description: -**

1. Model -- model of the car
2. Price -- Offer Price in EUROs
3. Age\_08\_04 -- Age in months as in August 2004
4. Mfg\_Month -- Manufacturing month (1-12)
5. Mfg\_Year -- Manufacturing Year
6. KM -- Accumulated Kilometers on odometer
7. Fuel\_Type -- Fuel Type (Petrol, Diesel, CNG)
8. HP -- Horse Power
9. Met\_Color -- Metallic Color? (Yes=1, No=0)
10. Color -- Color (Blue, Red, Grey, Silver, Black, etc.)
11. Automatic -- Automatic ( (Yes=1, No=0)
12. cc -- Cylinder Volume in cubic centimeters
13. Doors -- Number of doors
14. Cylinders -- Number of cylinders
15. Gears -- Number of gear positions
16. Quarterly\_Tax -- Quarterly road tax in EUROs
17. Weight -- Weight in Kilograms
18. Mfr\_Guarantee -- Within Manufacturer's Guarantee period (Yes=1, No=0)
19. BOVAG\_Guarantee -- BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0)
20. Guarantee\_Period -- Guarantee period in months
21. ABS -- Anti-Lock Brake System (Yes=1, No=0)
22. Airbag\_1 -- Driver\_Airbag (Yes=1, No=0)
23. Airbag\_2 -- Passenger Airbag (Yes=1, No=0)
24. Airco -- Airconditioning (Yes=1, No=0)
25. Automatic\_airco -- Automatic Airconditioning (Yes=1, No=0)
26. Boardcomputer -- Boardcomputer (Yes=1, No=0)
27. CD\_Player -- CD Player (Yes=1, No=0)
28. Central\_Lock -- Central Lock (Yes=1, No=0)
29. Powered\_Windows -- Powered Windows (Yes=1, No=0)
30. Power\_Steering -- Power Steering (Yes=1, No=0)
31. Radio -- Radio (Yes=1, No=0)
32. Mistlamps -- Mistlamps (Yes=1, No=0)
33. Sport\_Model -- Sport Model (Yes=1, No=0)
34. Backseat\_Divider -- Backseat Divider (Yes=1, No=0)
35. Metallic\_Rim --Metallic Rim (Yes=1, No=0)
36. Radio\_cassette -- Radio Cassette (Yes=1, No=0)
37. Tow\_Bar -- Tow Bar (Yes=1, No=0)

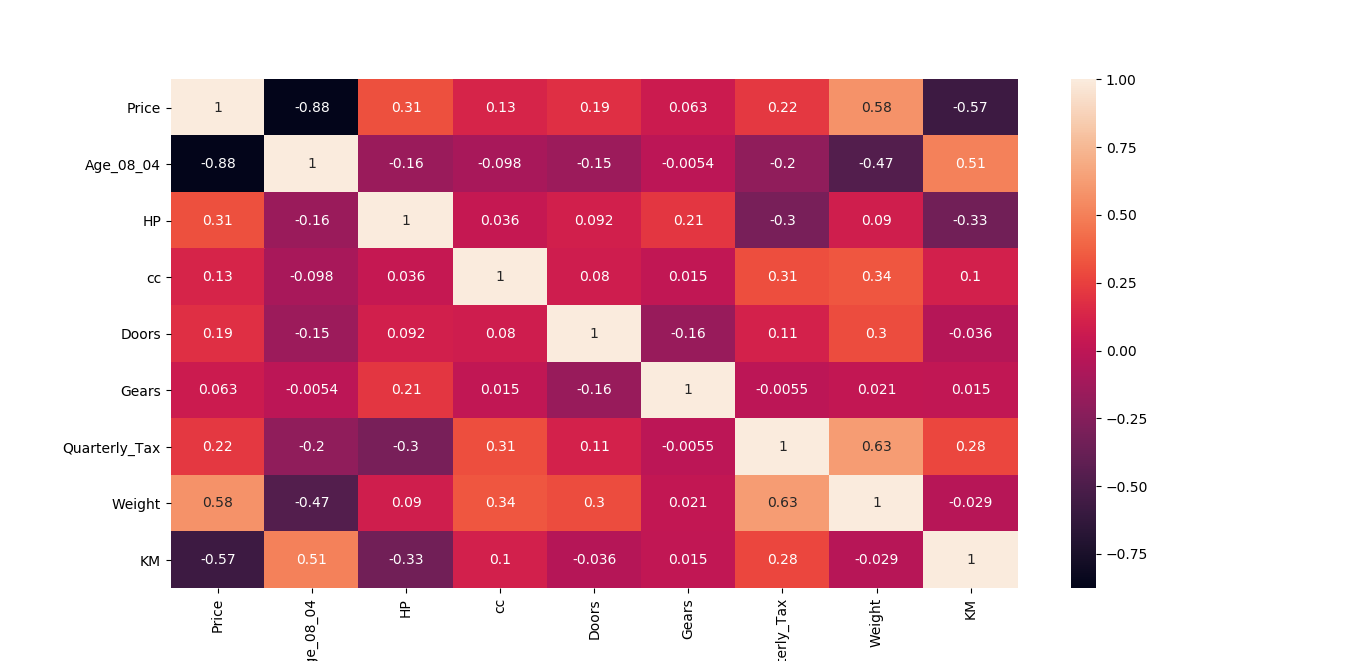
**Modelling Process followed –**

1. Reading the data (using pandas – pd)
2. EDA or Exploratory Data Analysis (Involves finding influencers, treating the outliers, removal of the same and data cleansing)
3. Model input – to test the data after cleansing the data
4. Splitting the data – To split the data into test and train
5. Checking the statistics and accuracy of the model.
6. Final model presentation

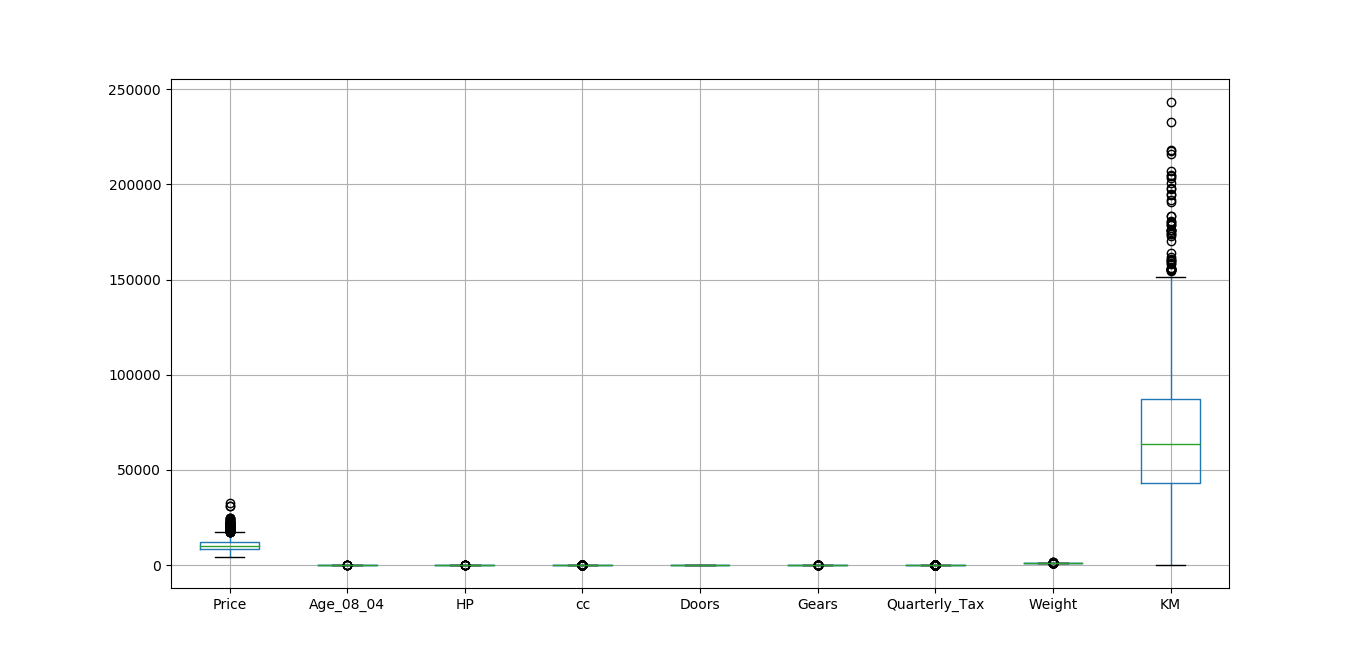
**EDA:**

****

**Correlation Plot:**



Boxplot – for checking the outlier presence in each feature.



After building the model taking raw input into account (without any data wrangling), we get the below features.

**Regression Results:**

R-squared: 0.890

OLS Adj. R-squared: 0.889

F-statistic: 1440.

Prob (F-statistic): 0.00

Log-Likelihood: -12186.

AIC: 2.439e+04

BIC: 2.444e+04

**Significance stats:**

coef std err t P>|t| [0.025 0.975]

-------------------------------------------------------------------------------------------------

Intercept 2.116e+04 3009.107 7.030 0.000 1.53e+04 2.71e+04

Age\_08\_04 -110.6511 2.433 -45.484 0.000 -115.423 -105.879

KM -0.0177 0.001 -15.010 0.000 -0.020 -0.015

HP 37.1339 2.771 13.402 0.000 31.699 42.569

cc -5837.8651 456.462 -12.789 0.000 -6733.277 -4942.454

Doors -135.9248 36.703 -3.703 0.000 -207.923 -63.927

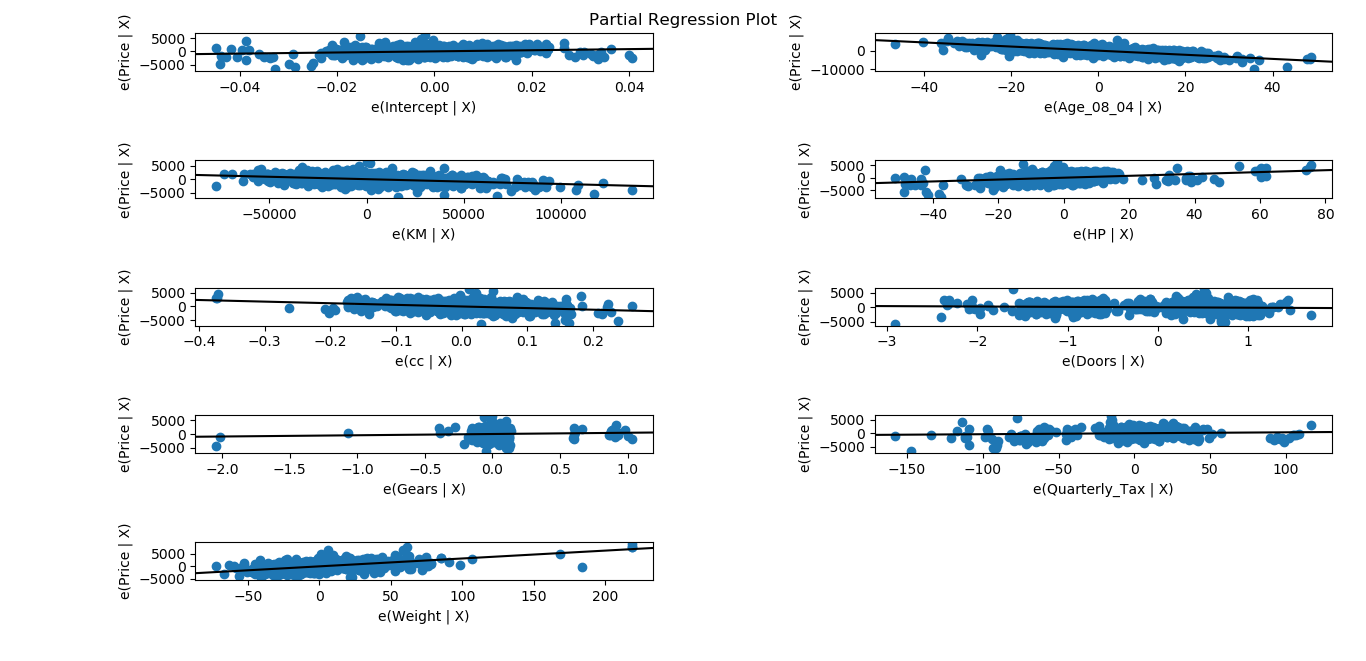
Gears 448.7112 177.172 2.533 0.011 101.166 796.257

Quarterly\_Tax 3.5187 1.349 2.609 0.009 0.873 6.164

Weight 31.8228 1.260 25.251 0.000 29.351 34.295

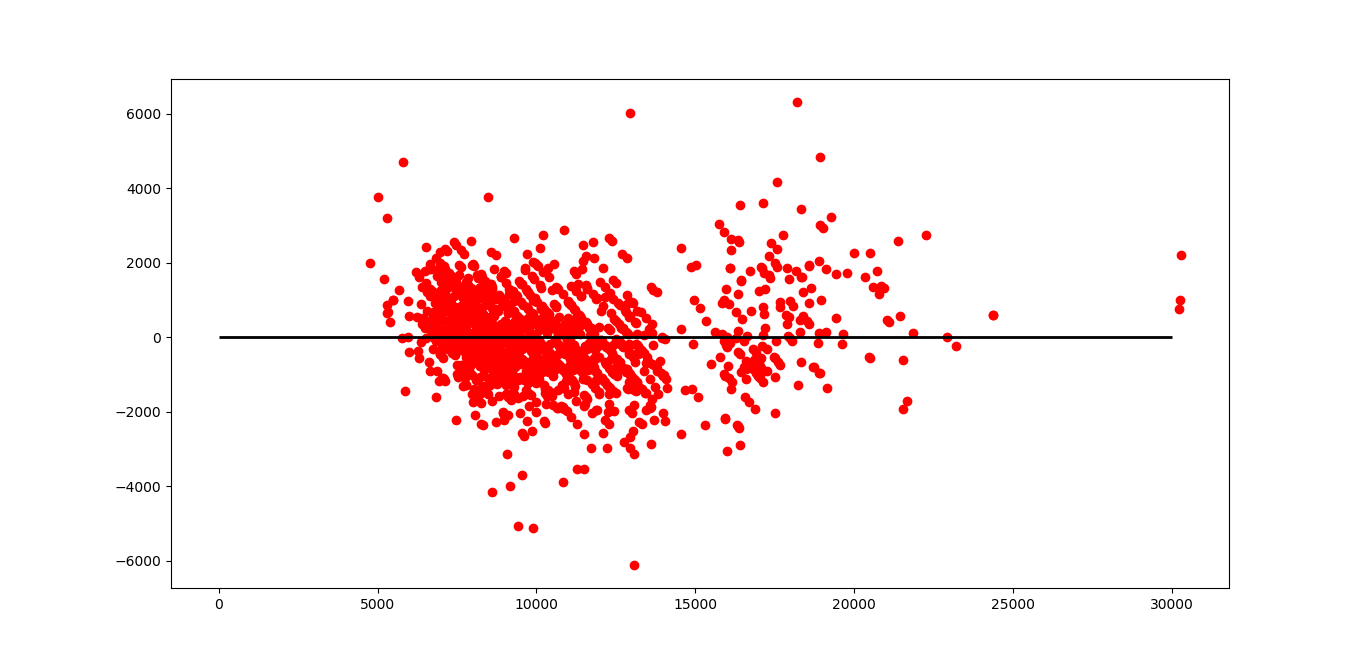
From above stats it can be observed that all the variables have significance on the output and hence cannot be ignored. Moreover the adj R^2 is found to be 0.889 which can still be considered as a valid R score.

Partial Regression Plot



Error Distribution box plot – The plot shows that the error distribution is normal and the same can be identified by the mean/median/mode which lies at the centre.

We got to know that the mean of the residue (y-y\_pred) is almost equal to 0 (-5.373520e-10)



Plotting the actual vs fitted values in the hyperplane