Q-1 Identify the Categorical Variables

1. Model
2. Fuel Type (Petrol, Diesel, CNG)
3. Color (Blue, Red, Grey, Silver, Black, etc.)
4. Met_Color (Yes=1, No=0)
5. Mfr_Month
6. Mfr_Year
7. Automatic ((Yes=1, No=0)
8. Mfr_Guarantee (Yes=1, No=0)
9. BOVAG_Guarantee
10. ABS (Yes=1, No=0)
11. Airbag_1 (Yes=1, No=0)
12. Airbag_2 (Yes=1, No=0)
13. Airco (Yes=1, No=0)
14. Automatic_airco (Yes=1, No=0)
15. Boardcomputer (Yes=1, No=0)
16. CD_Player (Yes=1, No=0)
17. Central_Lock (Yes=1, No=0)
18. Powered_Windows (Yes=1, No=0)
19. Power_Steering (Yes=1, No=0)
20. Radio (Yes=1, No=0)
21. Mistlamps (Yes=1, No=0)
22. Sport_Model (Yes=1, No=0)
23. Backseat_Divider (Yes=1, No=0)
24. Metallic_Rim (Yes=1, No=0)
25. Radio_cassette (Yes=1, No=0)
26. Parking assistance system (Yes=1, No=0)
27. Tow_Bar (Yes=1, No=0)

Q-2 Explain the relationship between a categorical variable and the series of binary dummy variables derived from it.

- A Dummy variable or Indicator Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels.
- A variable with N categories will be transformed into N or N-1 dummy variables ,each dummy will represent if a certain category is present or not.
- For example:- Petrol, Diesel, and CNG. If we convert fuel type to dummy variables,we get 3 dummy variables: Fuel_Type_Petrol (if the fuel type is Petrol then Fuel_Type_Petrol=1, otherwise Fuel_Type_Petrol=0), Fuel_Type_Diesel (if the fuel type is Diesel then Fuel_Type_Diesel=1, otherwise Fuel_Type_Diesel=0) and same for CNG.

- We create N-1 dummy variables when we have categorical variables which can be divided into Yes or No respectively so it can be converted into 1 and 0 respectively.

Q-3 How many dummy binary variables are required to capture the information in a categorical variable with N categories.

- N or N-1 dummy binary variables are required to capture the information in a categorical variable with N categories .
- In some situations like Linear Regression, use of all N dummies will cause failure because the nth variable contains redundant information and can be expressed as a linear combination of the others.
- In places where only 2 categories are there, N-1 variables should be used as they contain all the available information about the variable from which they were derived.
- For eg:-A car has Central Lock or not .It will be in two categories Yes or No (1or0),So dummy variable will be has central lock=1 so at this point not have central lock=0 ,car does not have central lock can be represented by N-1=(2-1=1).
- Both the '0' and '1' does not indicate numerical values but are strings representing something valuable information about the variable.

Q-4 Use R to convert the categorical variables in this dataset into dummy variables, and explain in words, for one record, the values in the derived binary dummies.

- I used the package named "dummies" to convert categorical variables into dummy variables.
- I used the code dummy.data.frame()
- EXAMPLE:- The Fuel_Type variable contains three categories - Petrol, Diesel, CNG. we get three different dummy variable namely Fuel_Type_Diesel, Fuel_Type_Petrol, Fuel_Type_CNG.
- Now, for the 1st record Fuel type is **Diesel**. Values in dummy variables are as follows: Fuel_Type_Diesel = 1, Fuel_Type_Petrol = 0, Fuel_Type_CNG = 0. In other words, the Fuel_Type_Diesel variable is '1' and other variable against the same row are '0'.

Q-5 Use R to produce a correlation matrix and matrix plot. Comment on the relationships among variables.

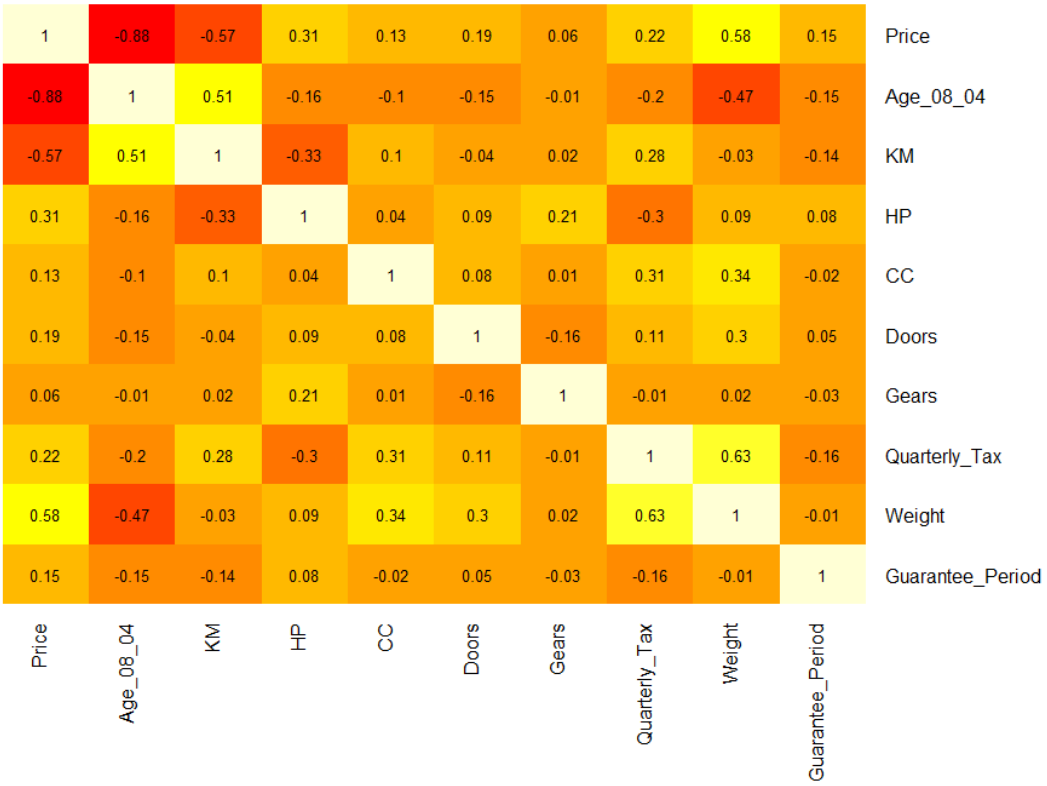
Relationship after analysing **Correlation Matrix and Matrix Plot**

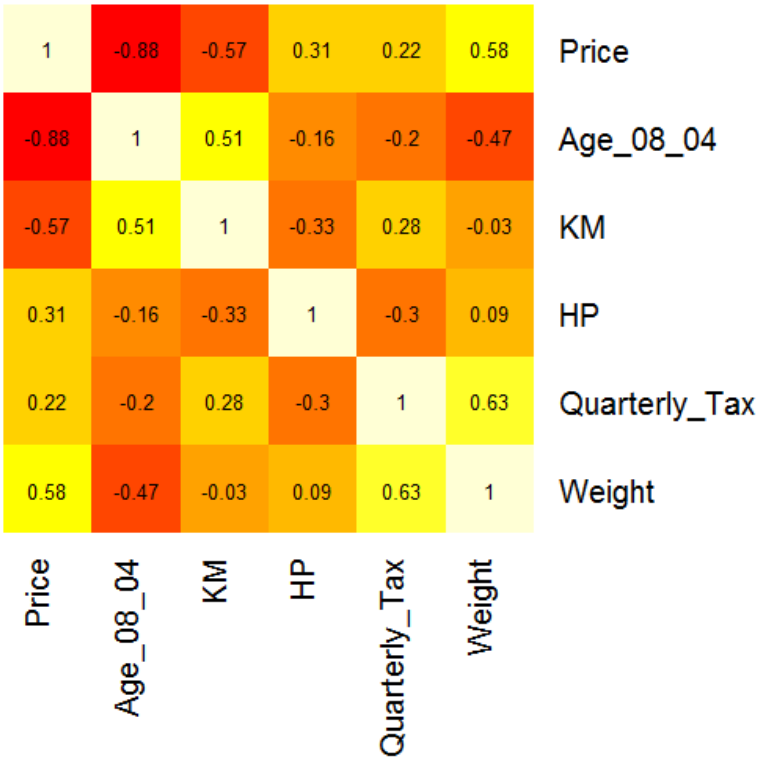
- Age_08_04 is negatively correlated with price (-0.88): This means when the age of the car is more the price of the car goes down.
- KM is negatively correlated with price(-0.57): When the km increases similar to age the price of the car decreases.
- Weight is positively correlated with price(0.58): Weight of the Car is positively correlated with the Price (higher price higher weight)
- KM is positively correlated with Age_08_04(0.51): Other than price, KM and Age of Car is positively stating a increasing relationship for both the factors when one increases.
- Weight is positively correlated with quarterly tax(0.63) : Quarterly Road Tax collected is positively related to Weight stating higher Car weight higher the Quarterly road tax.

Other observations useful for obtaining result:

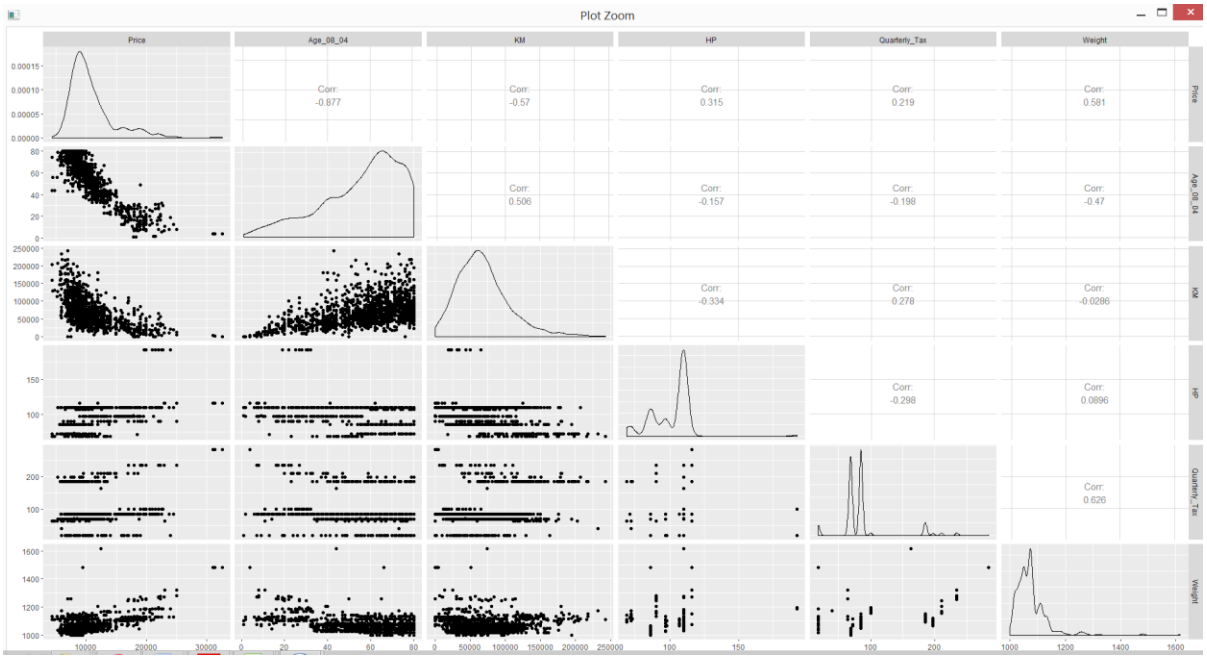
- Radio and Radio Cassette had a high correlation of 0.99 which helped me in dropping one of the variable.
- CNG dummy variable was removed as only 17 out of 1436 car were consumers of CNG.
- The number of Cylinders(4) were same across the dataset so was dropped while predicting prices.
- Mfg_Month and Mfg_Year are categorical variables but were creating a lot of dummy variables that is 12 for months alone and many others for year. So, Age in months of Car was taken to avoid overlapping information as it only resulted in minimal amount of loss of information.

Heatmap for all the continuous variables :





*CC, Doors, Gears, Guarantee period not considered as those are not significant for interest of study



Heatmap for analysing all the variables to avoid irrelevant variables