**CHAPTER : DATA ANALYSIS AND MODEL PREDICTION**

This Chapter covers the step by step analysis of the Data and how different Techniques of Data mining and analysis were applied in order to extract clean and useful data for prediction modeling.

Before Diving into the analysis part, the first step is to understand our data.

So now we will first discuss the different attributes in our Dataset

1. dateCrawled – date when the ad first crawled, datatype is date

2. name – string consisting of car name, brand, model etc.

3. seller – nature of seller either private or commercial

4. offerType – whether the car is on offer or has the buyer requested for an offer

5. price – price on the ad to sell the car ($)

6. abtest – two versions of ad

7. vehicleType – type of cars such as suv, cabrio etc

8. yearOfRegistration – year in which it was first registered

9. gearbox – manual or auto

10. powerPS – power of the car (HP)

11. model – model type of the car

12. kilometer – number of kilometers the car has travelled

13. monthOfRegistration – month of registration

14. fuelType – types of fuel

15. brand – make of the car

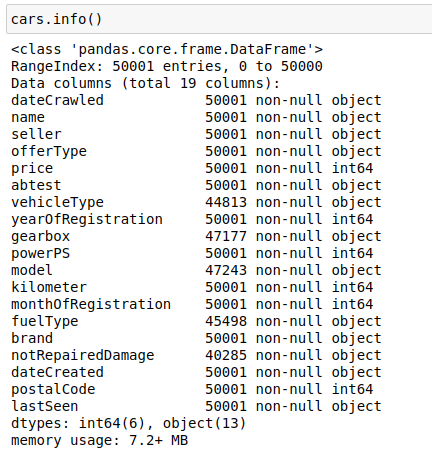
16. notRepairedDamage – status of repair for damages

17. dateCreated – date at which the ad at storm Motors was created

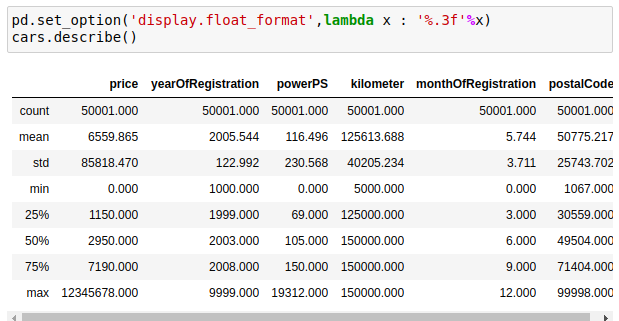
18. postalCode – postalCode of the seller

19. lastSeen – when the crawler saw this ad last online

So now we have an overview about the dataset, we will begin our analysis by first importing the dataset and getting info about the columns in the dataset, as shown below :



Next we will describe the Numerical attributes upto 3 decimal places as shown below :



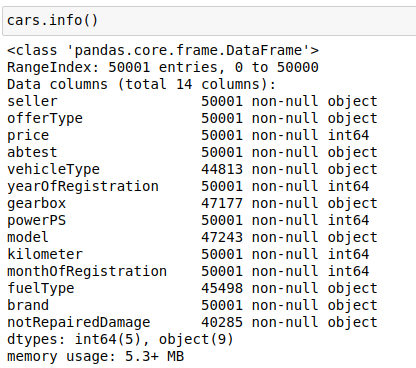
As visible from above image, there is a huge difference between mean and the median of price column. This means that price is very very skewed.

Now we will be removing unnecessary columns from the dataset such as

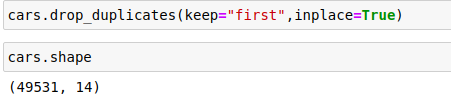
‘name' , 'dateCrawled' , 'dateCreated' , 'postalCode' and 'lastSeen as shown below :



Checking the Update info about the cars dataset



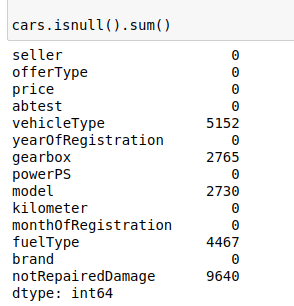
Now we will check for any duplicate entries in the dataset, and if there is any , then we will remove the same.



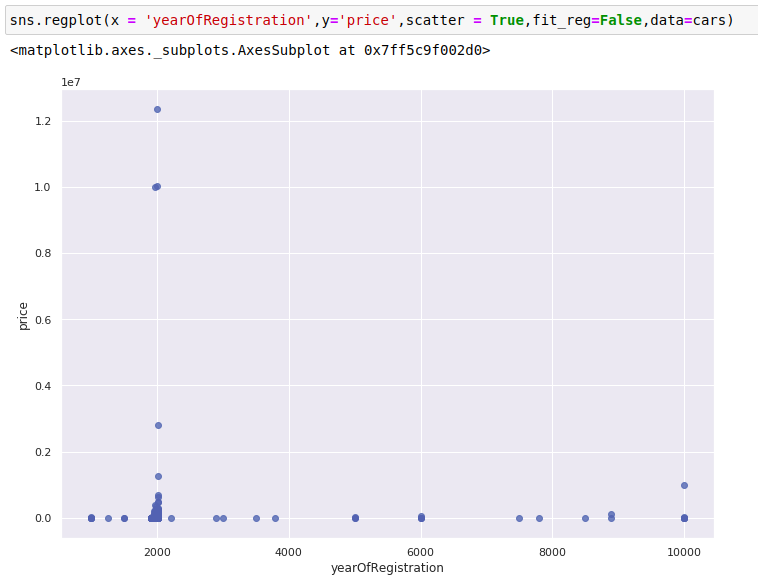
It can be seen that the number of rows decreased from 50001 to 49531 after removing the duplicate rows from the data set.

Next we will focus on Data cleaning part, that is understanding the null values in the data and will work upon them in order to get a clean and consistent dataset for further model predicition and analysis.

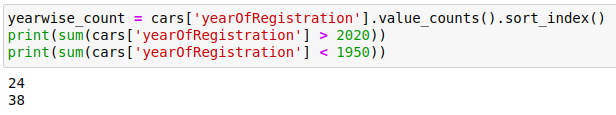
Below we check for any null values in the data set



We would like to plot the year of registration against the price of the car, in order to analyse the trend of price when yearOfRegistration is recent or old.

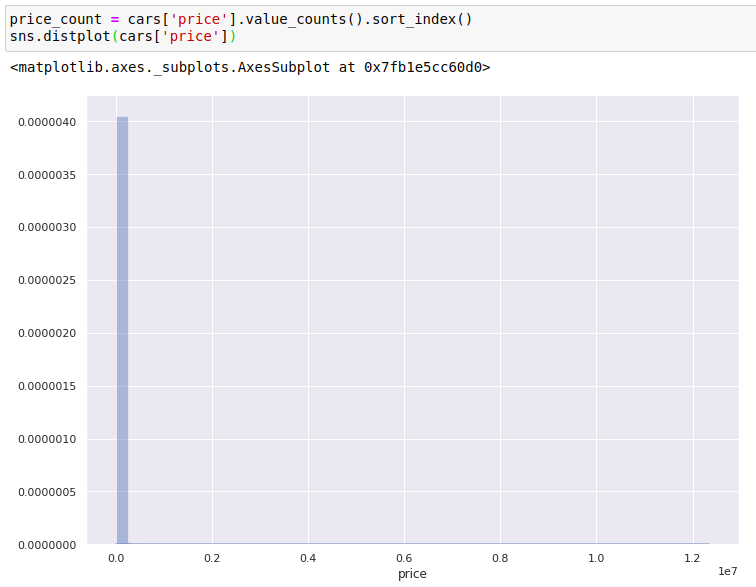


It is visible that there are a lot of outliers in the data. We are living in 2020 and there are rows with data much later than that. For eg we have an entry with yearOfRegistration as 10000 which is not possible. So we need to handle these outliers, for this we will check how many cars are there with registraion after 2020 and before 1950.



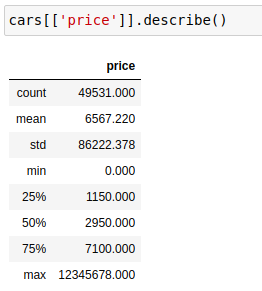
Here we get to know that we have 24 cars with registration after 2020 and 38 cars with registration before 1950.

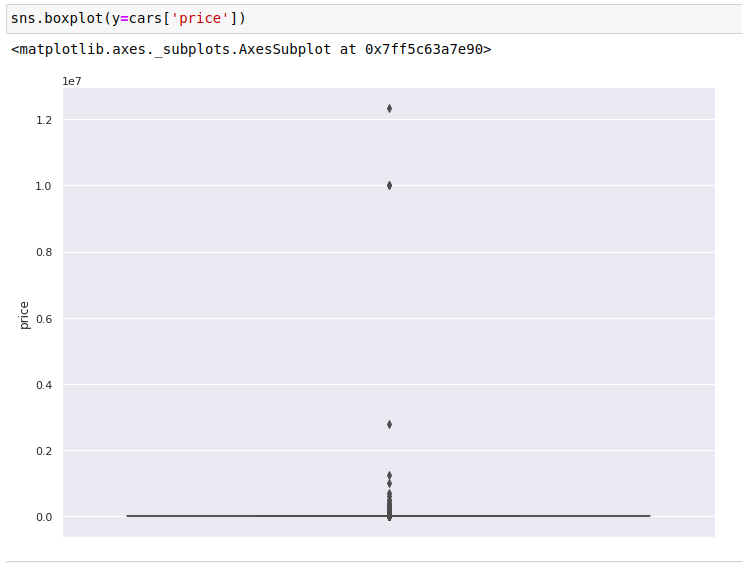
Next we try plotting the histogram of the price column and the result is shown below.



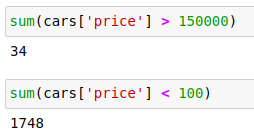
It was expected to be like this as studied above the price column is very skewed in nature.

For this, the price column is redescribed below and and also the boxplot for price column is studied below



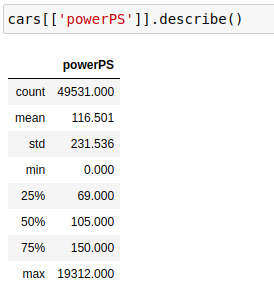


It is visible from the above boxplot that there are a lot of outliers in the price column, so after some hit and trials , a suitable range was achieved.

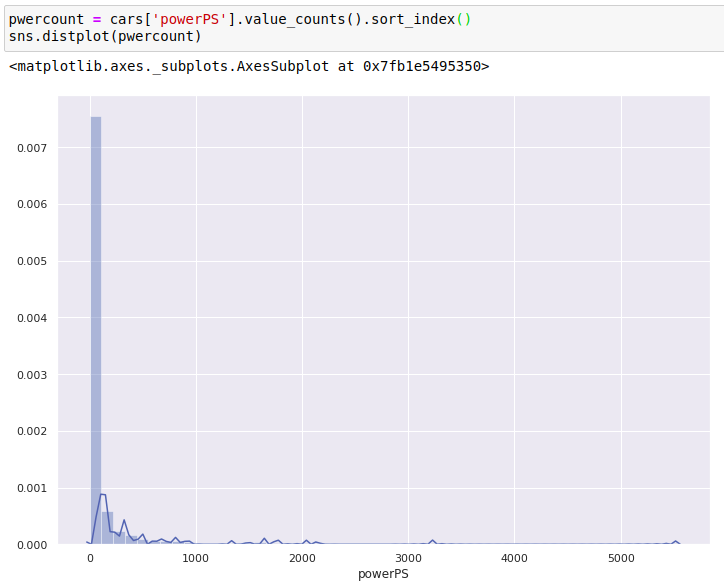


Now according to this data, there were 1748 cars having price below 100 dollars and 34 cars having price above 150000 dollars which is leading to all those ouliers in our data.

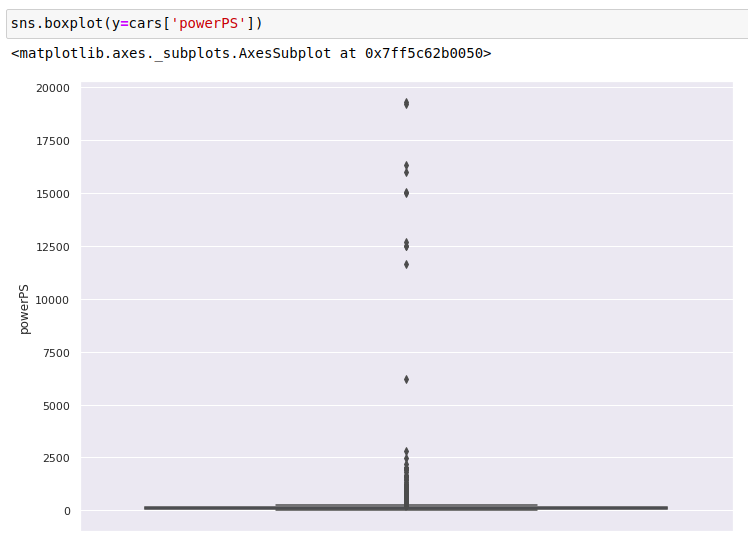
Similarly we observe for powerPS that there are a lot of outliers in the column as described below :



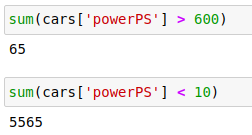
Furthermore, the histogram for the powerPS is plotted and shown below :



The boxplot of powerPS shows the outliers in the column :

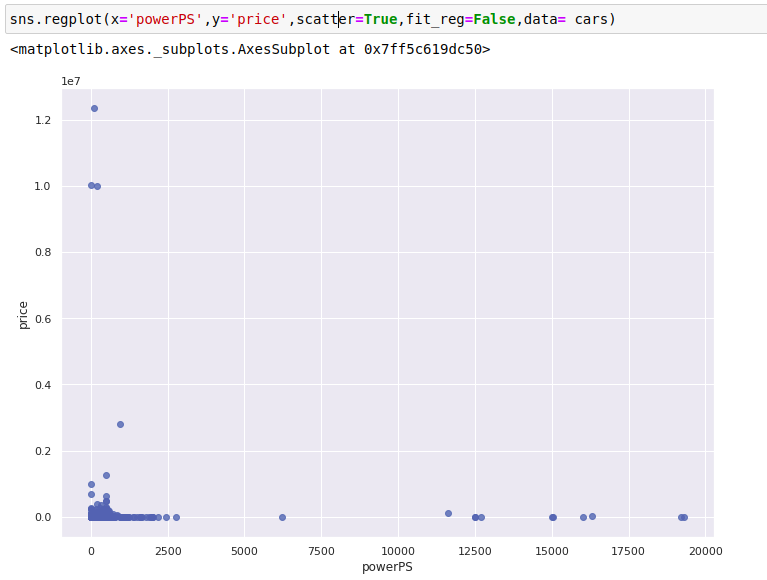


By hit and trial , we get to have the range for powerPS as shown :



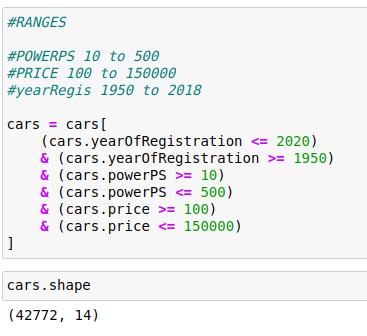
This shows that there are 5565 cars with power below 10 which is absurd and there are only 65 cars with power greater than 600 out of approx 50k cars.

Next we study how the graph varies when we plot together the price vs the powerPS graph in order to see a relation between the two.



Now we will modify our data in order to remove these outliers in the original data set.

The ranges and code snippet for the same are shown below :

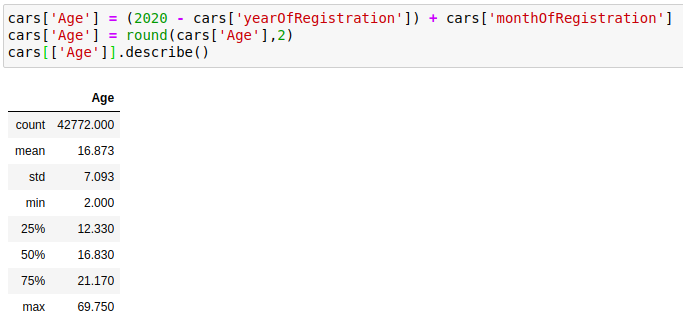


It can be observed that we have year of registration and month of registration. So instead of having 2 different features we can combine it into 1 single feature under the name “Age”.

For this, first we will convert months into years and then we will calculate the age of respective cars in years.



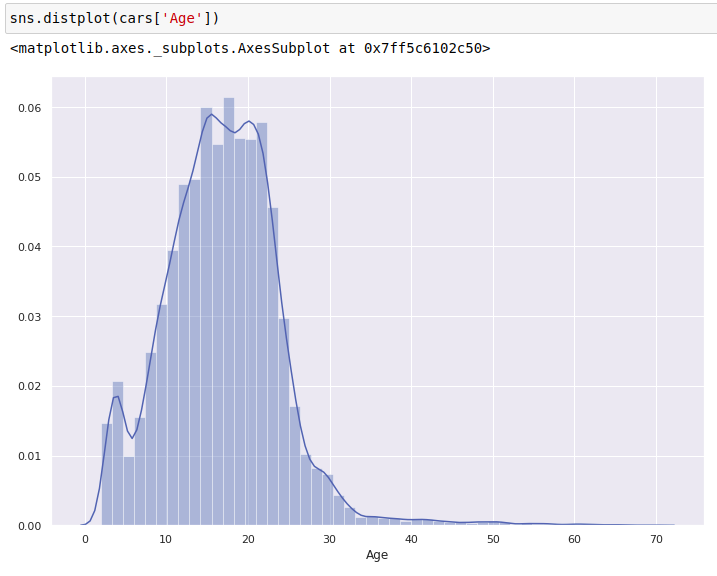
After this Age is calculated and described below :

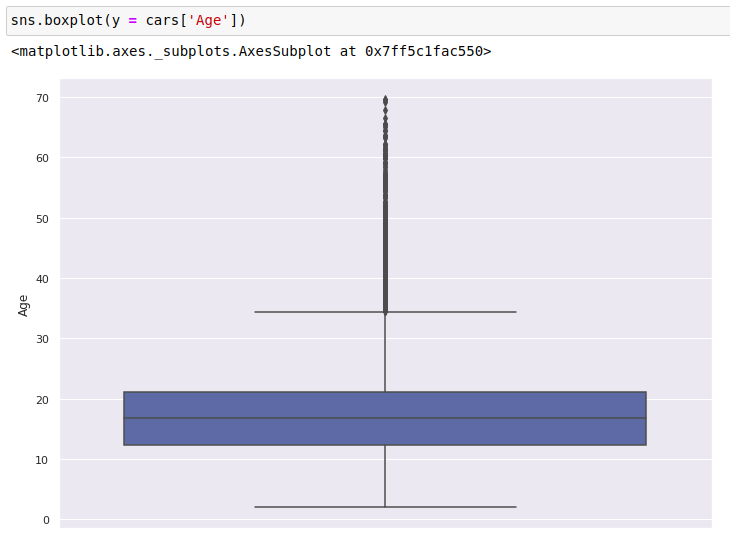


Now after this has been done, the two columns that are yearOfRegistration and monthOfRegistration can be removed from our dataset as the column Age is appended to the data set.

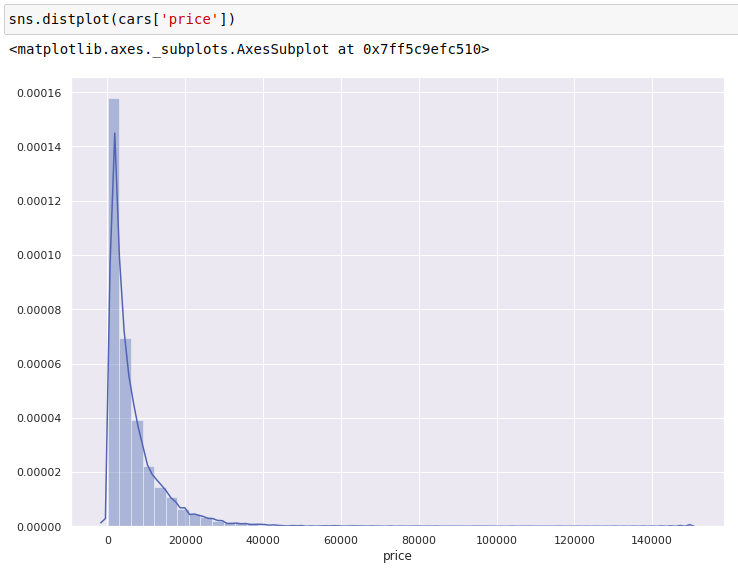


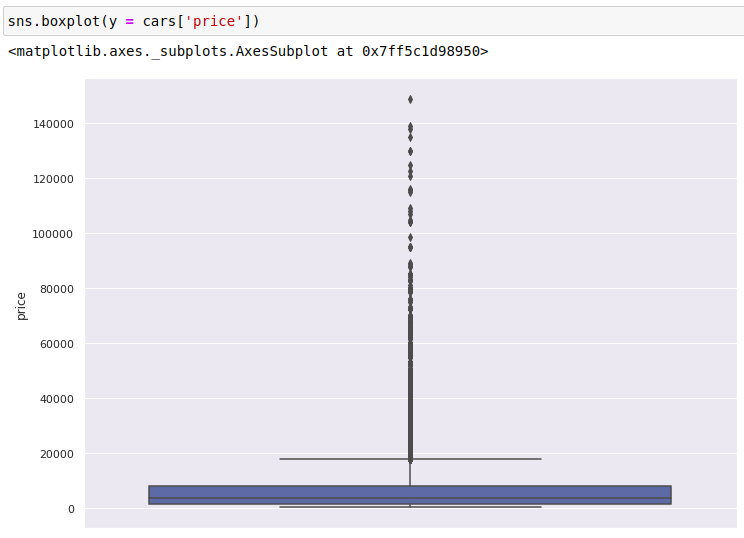
Now the column Age is analyzed by plotting its histogram and Box plots as shown



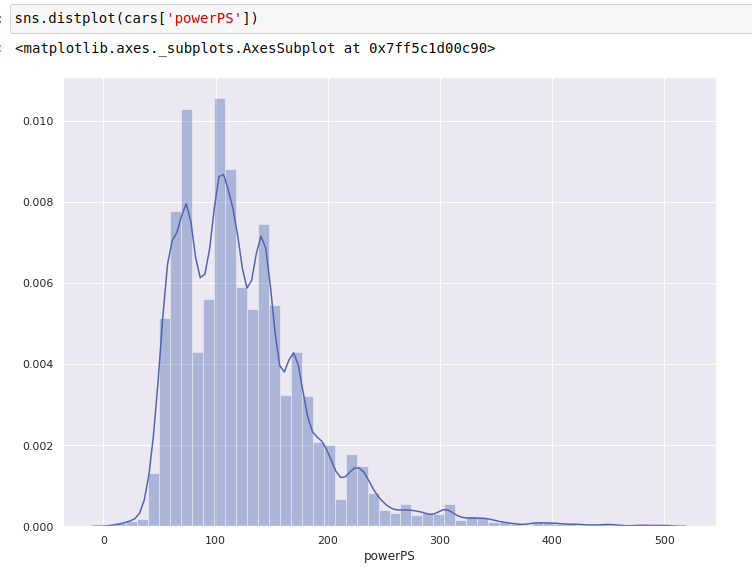


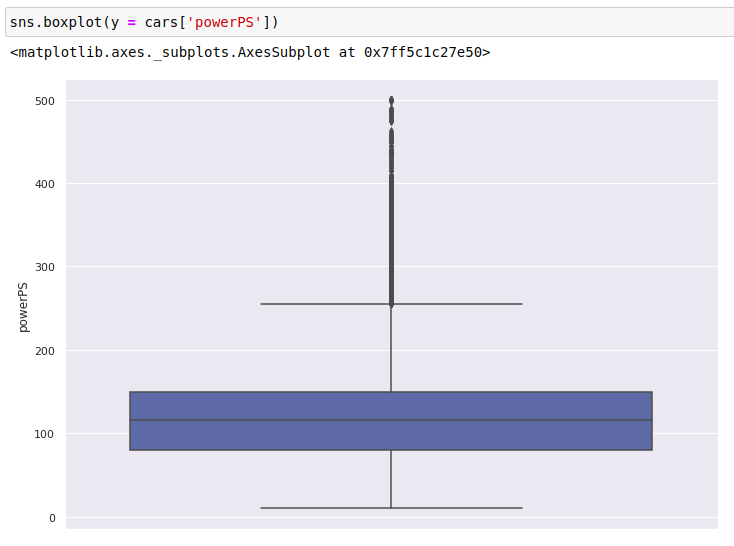
Next after Updating our dataset , we will see how Price is behaving after data reduction



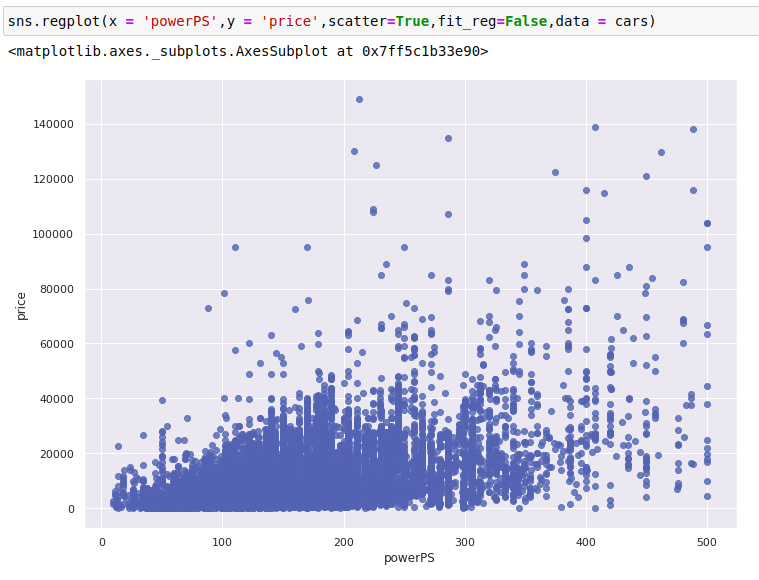


Next we will study the powerPS after Updation



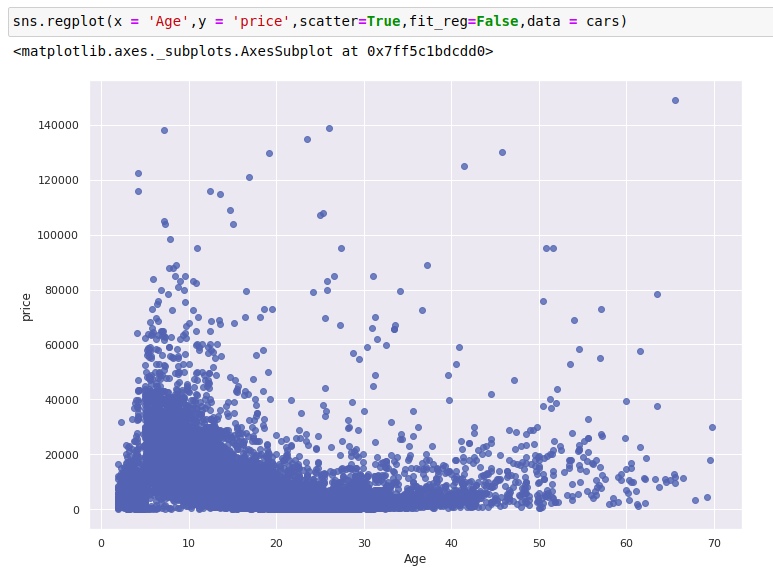


Now we will again study the relation between price and powerPS by plotting a Scatter Plot



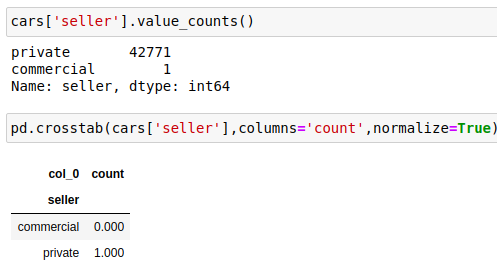
It can be seen that the price slightly increses with the increase in power of the car.

Similarly we can check the relation of price with Age of the Car by plotting the scatter plot of Age against price as shown below :



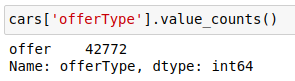
It is visible from the above scaatter plot that as the car is in new condition, or if the age of the car is less, then the price is more and as the age increases, the price of the car decreases.

Next we consider the column ‘seller’. The count of values in this attribute is done as shown below :

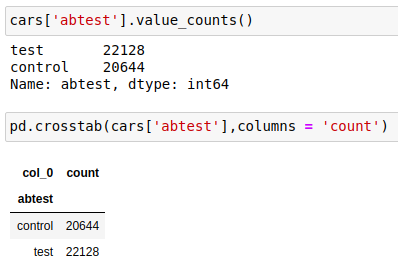


I can be seen the probability of driver being private is 1 and being commercial is 0.

Next we see the column offerType, which can have 2 values offer or request and in this current case, we only have the values as ‘offer’ as shown below :

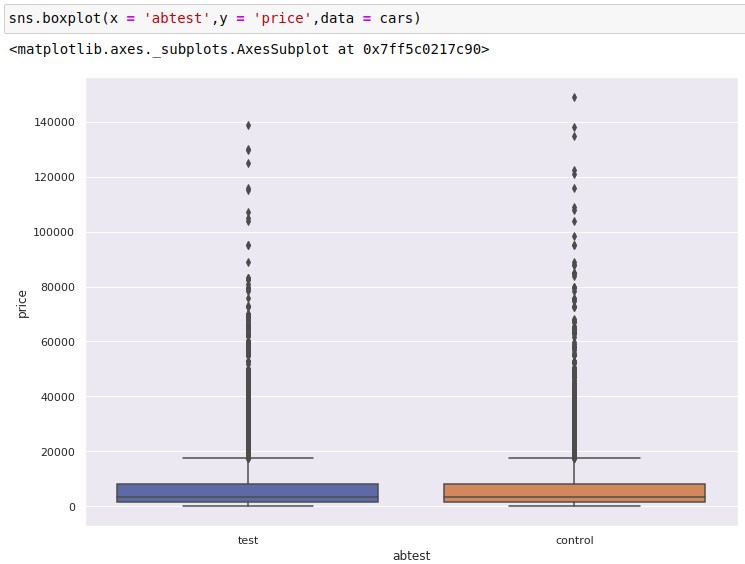


Next we see the count values of the column abtest as shown below :

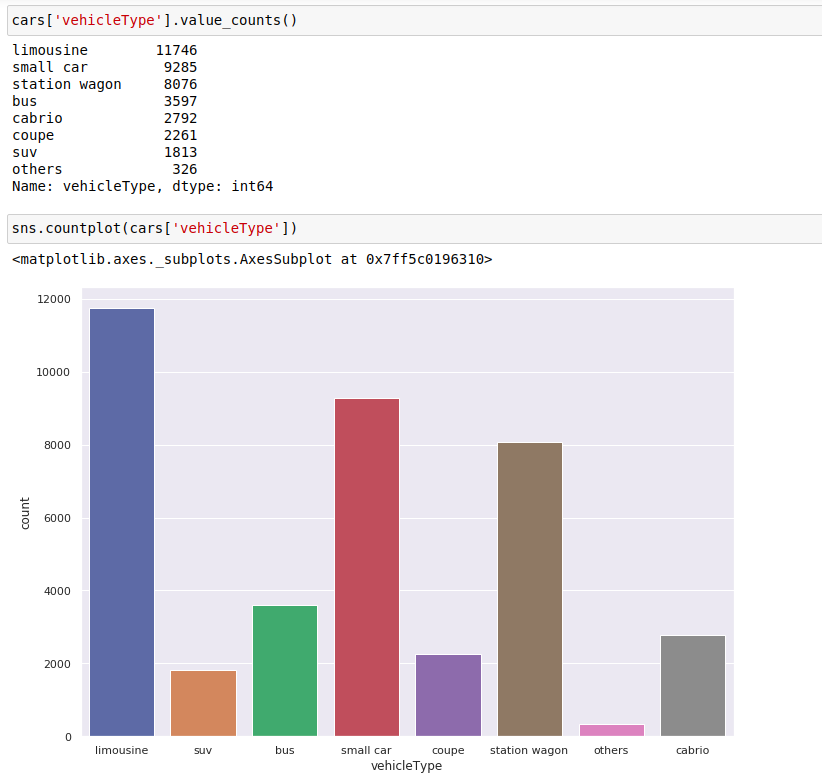


The countplot and boxplot for this column is shown below :





Next we try to see how many different types of vehicles are present in the dataset



This shows the count of different vehicleTypes in the dataset

The data set consists of 8 vehicleTypes namely

1. limousine

2. suv

3. bus

4. small car

5. coupe

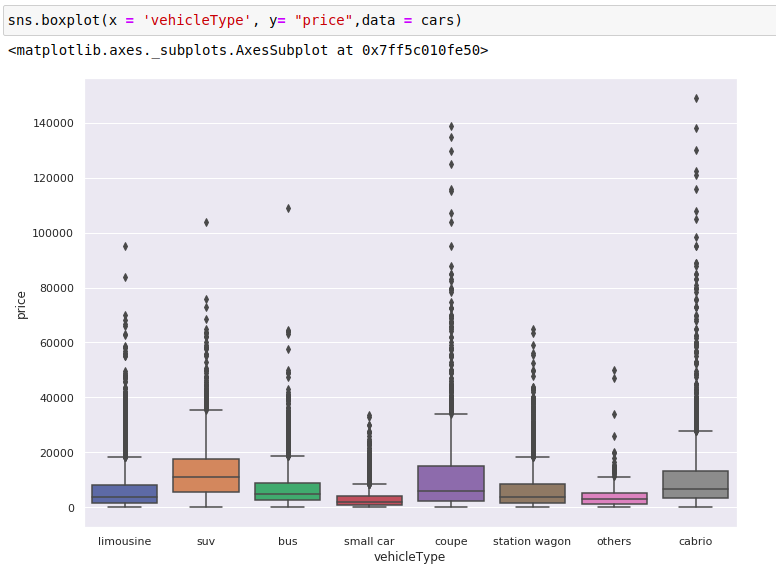
6. station wagon

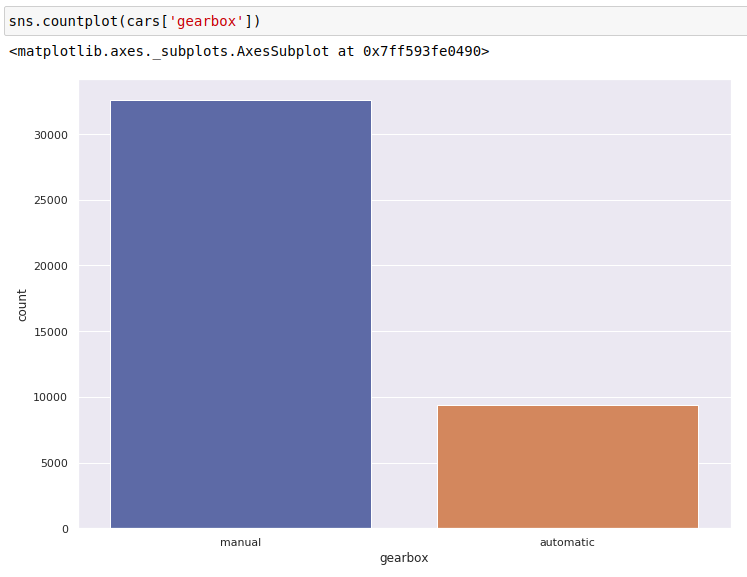
7. others

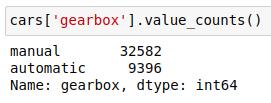
8. cabrio

We will see how different vehicleTypes act against the price of the cars.

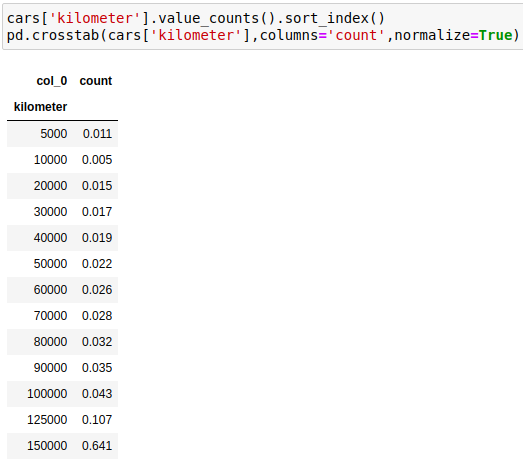
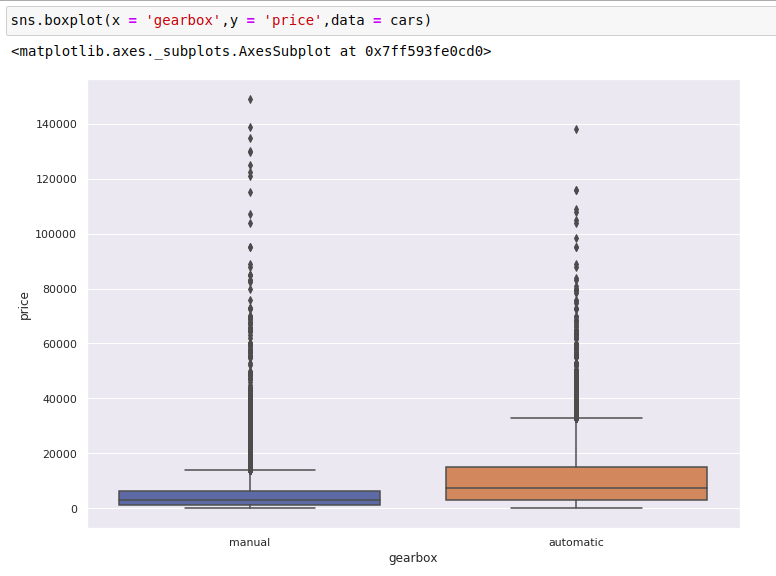
For this we will plot the boxplot for different vehicleTypes against price as shown below :

Next column comes the type of gearbox transmission, either manual or automatic.

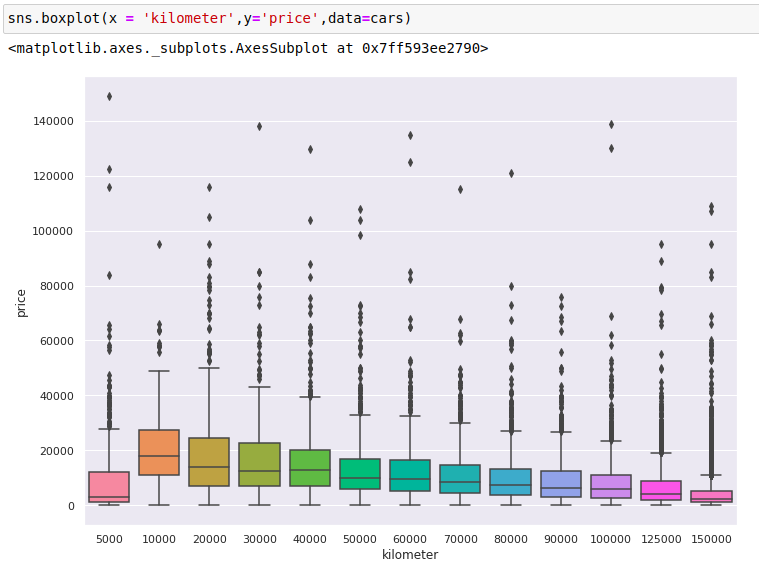




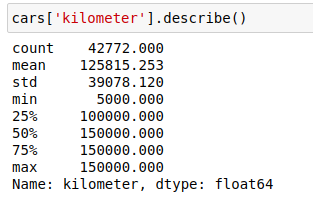
The dataset contains 32582 cars with manual transmission and 9396 cars with automatic transmission. Next we see how price varies with the type of gearbox via a boxplot.

Next we see the Kilometers travelled.

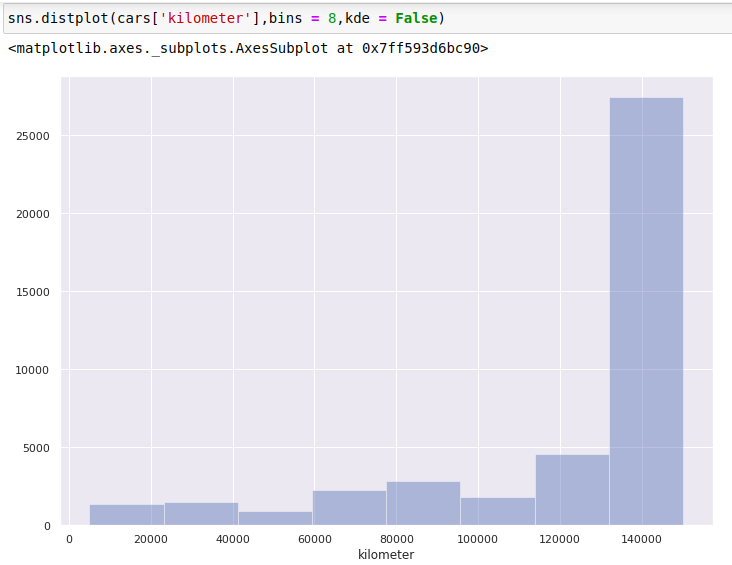
We will see the Box plot for Kilometers against price of the cars as shown below :



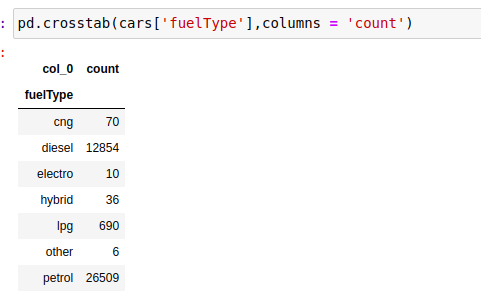
We describe the Kilometers column as :



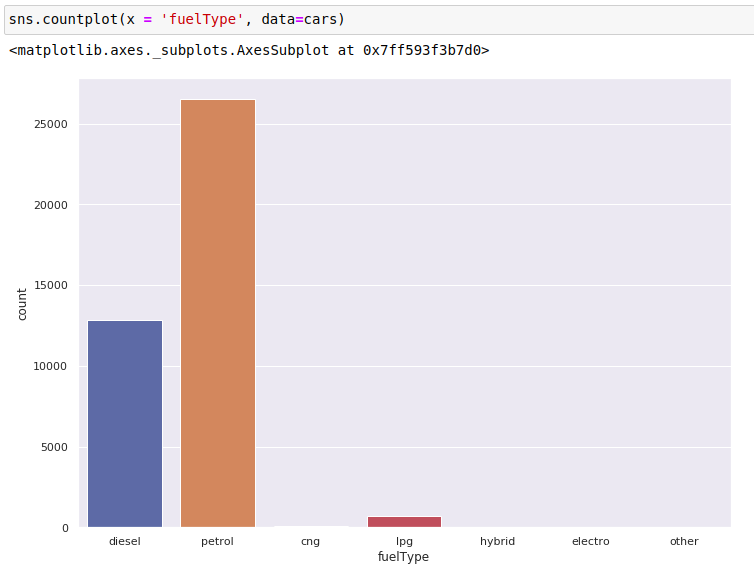
Histogram for Kilometers :



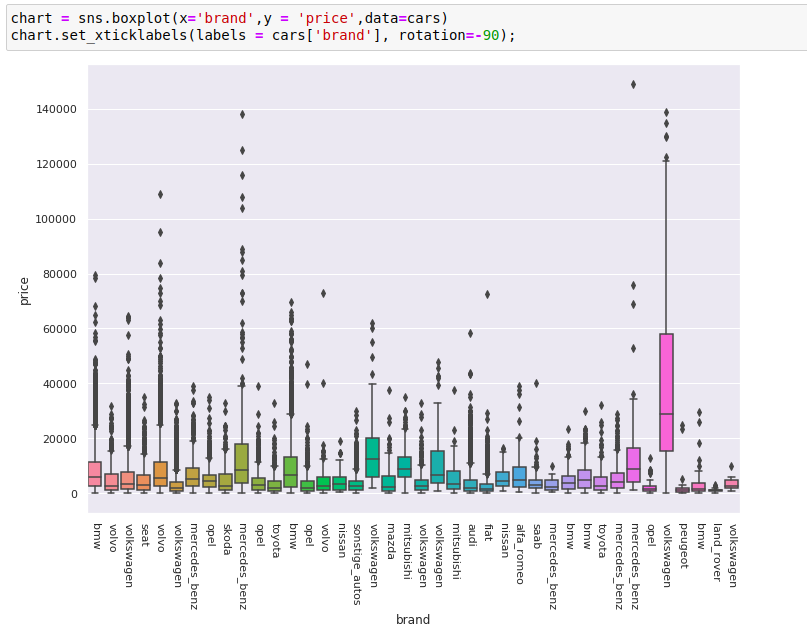
Next we move onto the different fueltypes and there count as shown below :



The countplot for the different fuelTypes is shown as :



Next we see how the price varies with the brand type in the given data set. This is shown with the help of box plots for different brands :





In the above screenshot we consider the column of notRepairedDamage

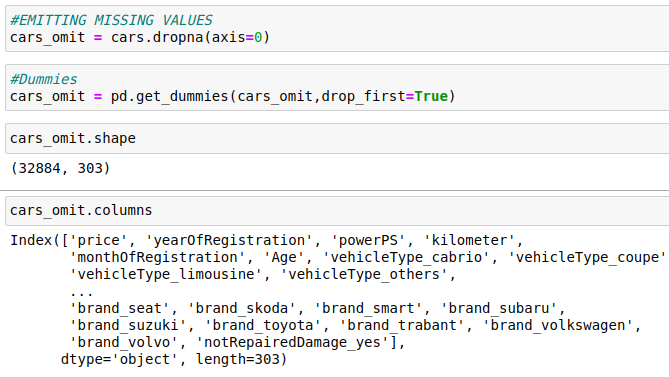
The count of repaired as yes or no is described and also the joint probability is described

Next we drop the other un necessary columns such as seller, offerType, abtest from the dataset as they are not required for prediction

Further we consider the correlation of other columns excluding object type columns with the price column and the result is described in the above screenshot.

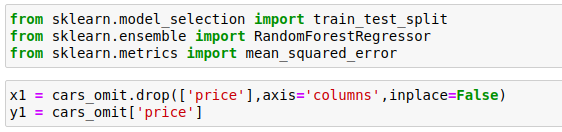
Now that we have analyzed the columns in the dataset, we will now drop the null values if there are left any.

Attributes with multiple values will be encoded into 0 and 1 by creating separate columns for each unique value and all this is shown as below :



Next we will move onto building a prediction model

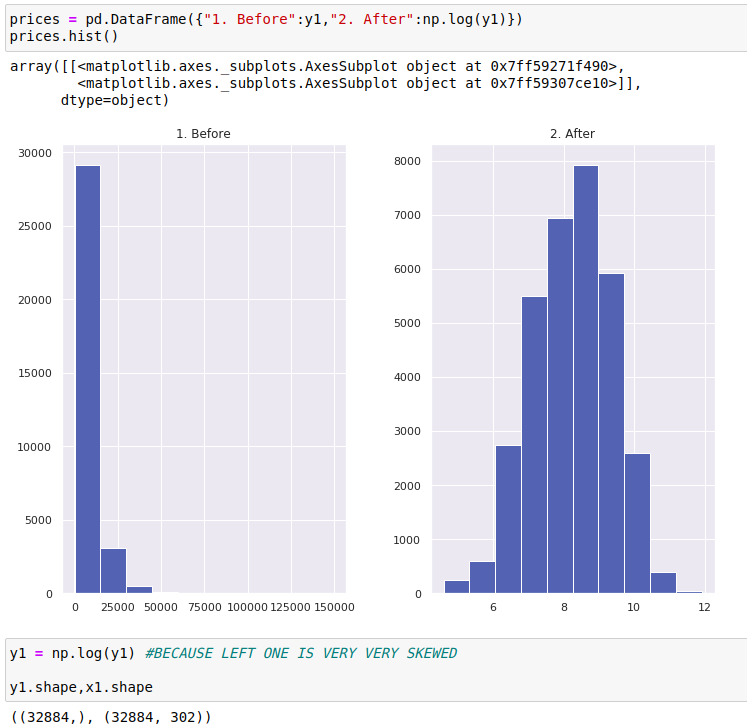
First we will get the data ready by removing the price feature from X labels and making the y variable equal to the price of the cars in the dataset.



Next we know that the price column is very skewed so taking the current column wont result into a good regression model and hence further predictions will not be matching with the actual values.

So in order to remove this issue we can change the value of the price column by taking the log of each cell and after prediction we can raise it to exponent to get the actual price value.

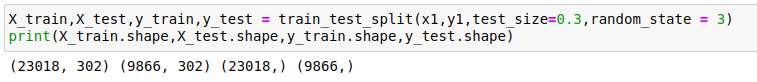
This step done is explained by plotting the histogram in both the cases as shown below :



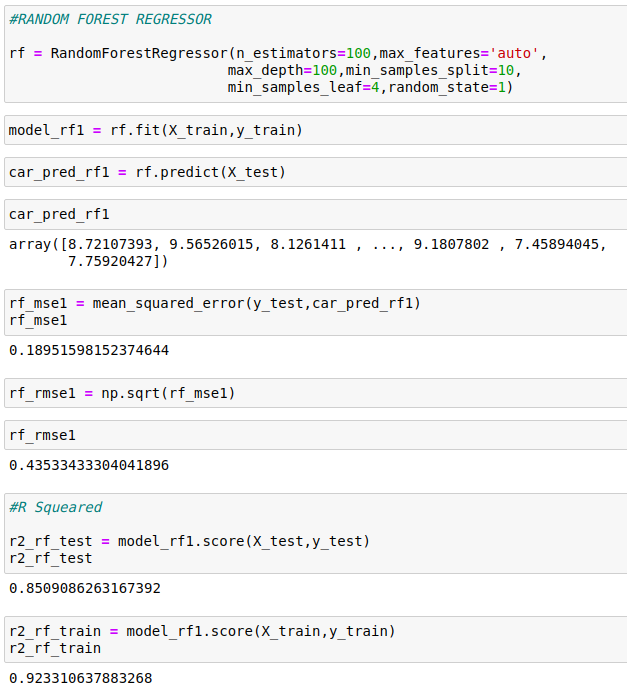
Next we will split the dataset into 2 different values

1. Training Set – for preparing the model

2. Testing Set – for checking the model precision



Next we will build the Random forest Regressor from sklearn library as shown below :



Here we get the R Square score as 0.85 for the Test data and 0.92 for the Train Data

Next we will plot the residual graph to check how the predicted and actual values vary

This is shown as below on the next page :

