

“A PROJECT REPORT ON

PRICE PREDICTION OF USED CARS”



**SUBMITTED BY**

**HIMAJA IJJADA**

**ACKNOWLEDGMENT**

I express my sincere gratitude to FlipRobo Technologies for giving me the opportunity to work on the price prediction of used cars project using machine learning algorithms. I would also like to thank FlipRobo Technologies for providing me with the requisite knowledge to webscrape the datasets to work with. And I would like to express my gratitude to Mr. Mohd Kashif (SME FlipRobo) and Ms. Sapna Verma (SME FlipRobo) for being of a great help in completion of the project.

Most of the concepts used to predict the Prices of used cars project are learned from Data Trained Institute and below documentations.

* https://scikit-learn.org/stable/
* https://seaborn.pydata.org/
* https://www.scipy.org/

**CONTENTS**

* **Introduction**
* Business Problem Framing
* Conceptual Background of the Domain Problem
* Review of Literature
* Motivation for the Problem Undertaken
* **Analytical problem framing**
* Mathematical/ Analytical Modeling of the Problem
* Data Sources and their formats
* Data Preprocessing Done
* Data Inputs- Logic- Output Relationships
* Assumptions
* Hardware and Software Requirements and Tools Used
* **Model/s Development and evaluation**
* Visualizations
* Identification of possible problem-solving approaches (methods)
* Testing of Identified Approaches (Algorithms)
* Run and Evaluate selected models
* Key Metrics for success in solving problem under consideration
* Interpretation of the Results
* **Conclusion**
* Key Findings and Conclusions of the Study
* Learning Outcomes of the Study in respect of Data Science
* Limitations of this work and Scope for Future Work

**Introduction**

* **Business Problem Framing**

Predicting the price of used cars is an important and interesting problem. Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (model), the origin of the car (the original location of the manufacturer), its mileage (the number of kilometers it has run) and its horsepower (amount of power that an engine produces). Due to rising fuel prices, fuel economy is also of prime importance. Unfortunately, in practice, most people do not know exactly how much fuel their car consumes for each km driven. Other factors such as the type of fuel it uses, the interior style, the braking system, acceleration, engine displacement, the volume of its cylinders (measured in cc), its size, number of doors, paint color, weight of the car, consumer reviews, prestigious awards won by the car manufacturer, its physical state, whether it is a sports car, whether it has cruise control, whether it is automatic or manual transmission, whether it belonged to an individual or a company and other options such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator all may influence the price as well. Some special factors which buyers attach importance is the local of previous owners, whether the car had been involved in serious accidents. The look and feel of the car certainly contribute a lot to the price. As we can see, the price depends on a large number of factors. Unfortunately, information about all these factors are not always available and the buyer must make the decision to purchase at a certain price based on few factors only. In this work, we have considered only a small subset of the factors which are more important.

With the Covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to Covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

**Business goal:** The main aim of this project is to predict the price of used car based on various features. Machine Learning is a field of technology developing with immense abilities and applications in automating tasks. So, we will deploy an ML model for car selling price prediction and analysis. This kind of system becomes handy for many people. This model will provide the approximate selling price for the car based on different features like fuel type, transmission, and price, weight, running in kms, engine displacement, mileage etc. and this model will help the client to understand the price of used cars.

* **Conceptual Background of the Domain Problem**

Car Price Prediction is really an interesting machine learning problem as there are many factors that influence the price of a car in the second-hand market. In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party – usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a pre-defined number of months/years to the seller/financer. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e., its expected resale value. Thus, it is of commercial interest to seller/financers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value is under-estimated by the seller/financer at the beginning, the instalments will be higher for the clients who will certainly then opt for another seller/financer. If the residual value is over-estimated, the instalments will be lower for the clients but then the seller/financer may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Thus, we can see that estimating the price of used cars is of very high commercial importance as well.

Here we are trying to help the client works with small traders, who sell used cars to understand the price of the used cars by deploying machine learning models. These models would help the client/sellers to understand the used car market and accordingly they would be able to sell the used car in the market.

* **Review of Literature**

Literature review covers relevant literature with the aim of gaining insight into the factors that are important to predict the price of used cars in the market. In this study, we discuss various applications and methods which inspired us to build our supervised ML techniques to predict the price of used cars in different locations. We did a background survey regarding the basic ideas of our project and used those ideas for the collection of data information by doing web scraping from [www.cardekho.com](http://www.cardekho.com/) **website which is a** web platform where seller can sell their used car.

This project is more about data exploration, feature engineering and pre-processing that can be done on this data. Since we scrape huge amount of data that includes more car related features, we can do better data exploration and derive some interesting features using the available columns. Different techniques like ensemble techniques, k-nearest neighbors, and decision trees have been used to make the predictions.

The goal of this project is to build an application which can predict the car prices with the help of other features. In the long term, this would allow people to better explain and reviewing their purchase with each other in this increasing digital world.

* **Motivation for the Problem Undertaken**

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, engine displacement, running, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately.

So, the main aim is to use machine learning algorithms to develop models for predicting used car prices.

* To build a supervised machine learning model for forecasting value of a vehicle based on multiple attributes.
* The system that is being built must be feature based i.e., feature wise prediction must be possible.
* Providing graphical comparisons to provide a better view

**Analytical problem framing**

* **Mathematical/ Analytical Modeling of the Problem**

We need to develop an efficient and effective Machine Learning model which predicts the price of a used cars. So, “Car\_Price” is our target variable which is continuous in nature. Clearly it is a Regression problem where we need to use regression algorithms to predict the results.

This project is done on two phases:

1. **Data Collection Phase: I have done web scraping to collect the data of used cars from the well-known website**[**www.cardekho.com**](http://www.cardekho.com)**where I found more features of cars compared to other websites and I fetch data for different locations. As per the requirement of our client we need to build the model to predict the prices of these used cars.**
2. **Model Building Phase:** After collecting the data, I built a machine learning model. Before model building, have done all data pre-processing steps. The complete life cycle of data science that I have used in this project are as follows:

* Data Cleaning
* Exploratory Data Analysis
* Data Pre-processing
* Model Building
* Model Evaluation
* Selecting the best model
* **Data Sources and their formats**

We have collected the dataset from the website [**www.cardekho.com**](http://www.cardekho.com)**which is a web platform where** seller can sell their used car. The data is scrapped using Web scraping technique and the framework used is Selenium. We scrapped nearly 12600 of the data and fetched the data for different locations and collected the information of different features of the car and saved the collected data in excel format. The dimension of the dataset is 12608 rows and 20 columns including target variable “Car\_Price”.

****

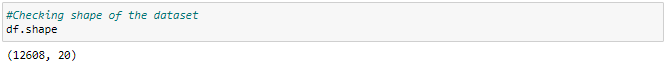
* **Data Preprocessing Done**

Data pre-processing is the process of converting raw data into a well-readable format to be used by Machine Learning model. Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre-process our data before feeding it into our model. I have used following pre-processing steps:

Dropping unnecessary column

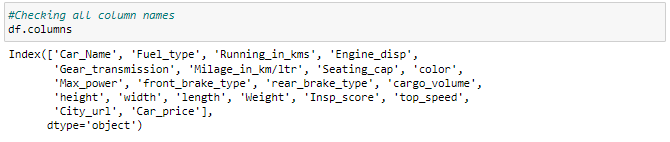


Checking shape of the dataset



The dataset has 12608 rows and 20 columns.

Checking all column names

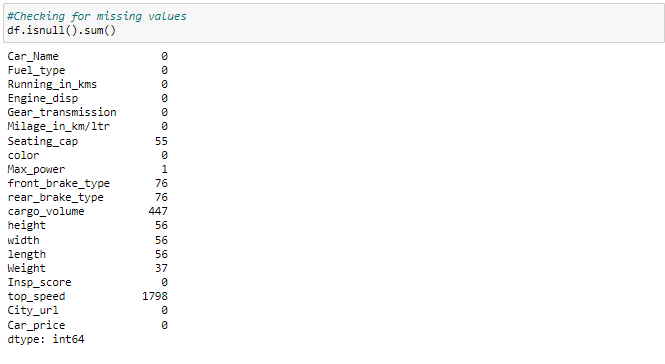


Above are the list of column names in the dataset.

**Feature description:**

* Car\_Name : Name of the car with Year
* Fuel\_type : Type of fuel used for car engine
* Running\_in\_kms : Car running in kms till the date
* Engine\_disp : Engine displacement/engine CC
* Gear\_transmission : Type of gear transmission used in car
* Milage\_in\_km/ltr : Overall milage of car in Km/ltr
* Seating\_cap : Availability of number of seats in the car
* color : Car color
* Max\_power : Maximum power of engine used in car in bhp
* front\_brake\_type : type of brake system used for front-side wheels
* rear\_brake\_type : type of brake system used for back-side wheels
* cargo\_volume : the total cubic feet of space in a car's cargo area.
* height : Total height of car in mm
* width : Width of car in mm
* length : TOtal length of the car in mm
* Weight : Gross weight of the car in kg
* Insp\_score : inspection rating out of 10
* top\_speed : Maximum speed limit of the car in km per hours
* City\_url : Url of the page of cars from a particular city
* Car\_price : Price of the car

Checking for missing values

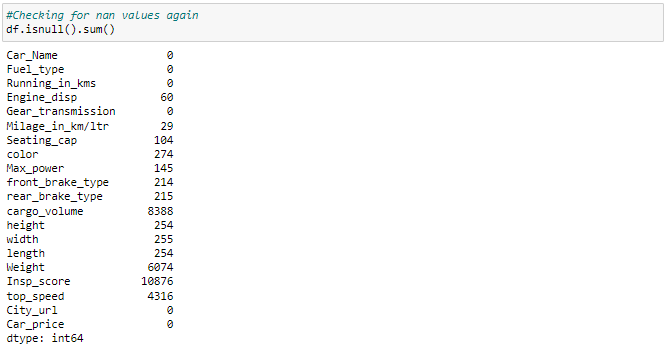


There are some entries like '-' and 'null' so let's replace these with nan.

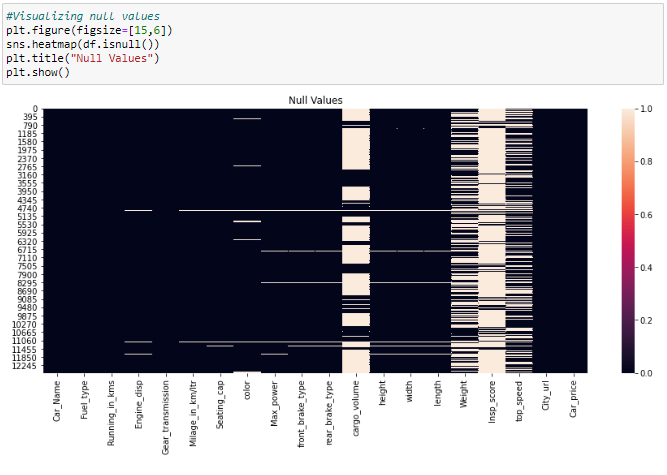
Replacing unnecessary entries with nan



Checking for nan values again



Visualizing null values



## Observations

By visualization of null values we can clearly say that

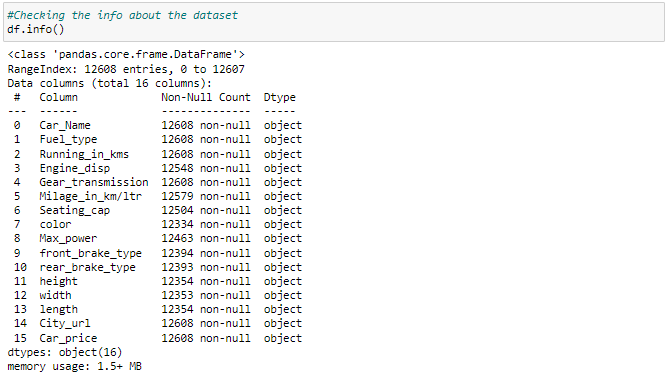
* There are huge null values in the following columns of the dataset-
  + cargo\_volume
  + Weight
  + Insp\_score
  + top\_speed
* There are few missing values in the following columns
  + Engine\_disp
  + Milage\_in\_km/ltr
  + Seating\_cap
  + color
  + Max\_power
  + front\_brake\_type
  + rear\_brake\_type
  + height
  + width
  + length
* We can drop those columns with more than 50% missing values and others with less than 50% missing values can be replaced using iteration methods

Dropping columns with more than 50% of missiing values



We have dropped the above columns

Checking the info about the dataset

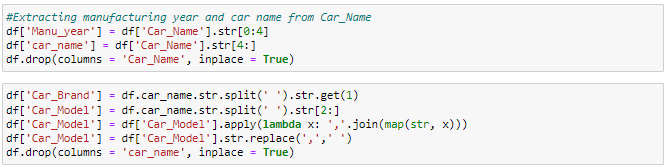


## Observations

Above is the info about the dataset and we shall fill these missing values in the dataset using imputation methods.

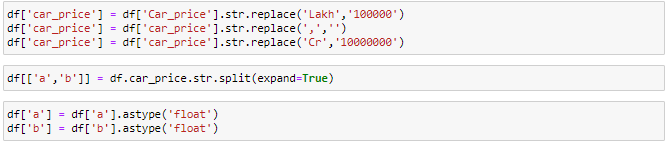
## Car\_Name:

As the Car\_name column has year of manufacture, car model and car name all together so we shall extract them for better seperated values



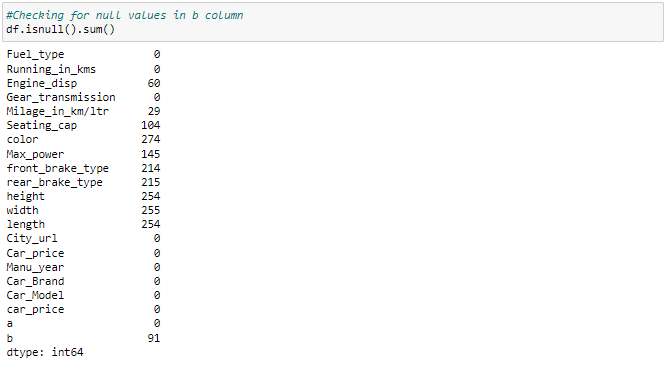
## Car\_Price:

Since Car\_Price is our target, it shall be in the continuous data format. So we have to change the car\_price column from lakhs and crores to integer format.



Now we have separated the column with alpha numeric data to replace the values with float type data

Checking for null values in b column





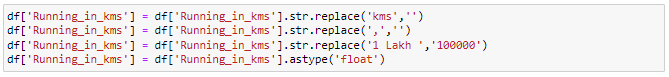
We have filled the nan values in the column



Now we have converted the data in car\_price to float type data

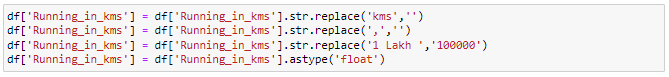


We have dropped the columns as they are redundant

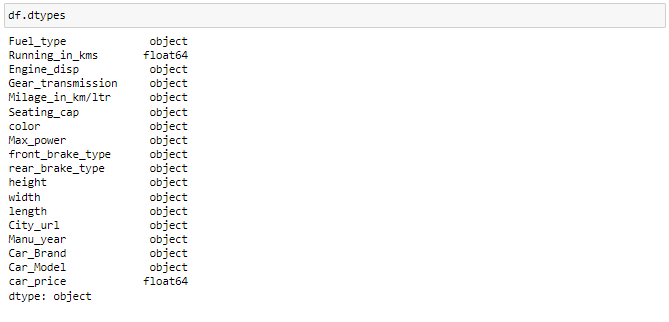


## Running\_in\_kms:

Since this column should be int datatype but it has some string values and ',' in between so let's replace them.



We have replaced the values in the above column with number of km and converted into the float type data



## Observations

* Most of the columns are in object type data.

## Engine\_disp:

The column 'Engine\_disp' should be continuous column so we will convert it to float datatype.



we have converted the data type of 'Engine\_disp' to float datatype.

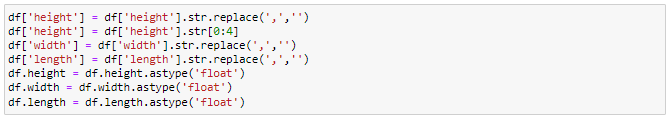
## Milage\_in\_km/ltr:

In Milage\_in\_km/ltr column the data type is object so we have to change this to float type.



We have converted the data type of 'Milage\_in\_km/ltr' to float datatype.

## Converting the data type of columns height, width and length to float datatype:

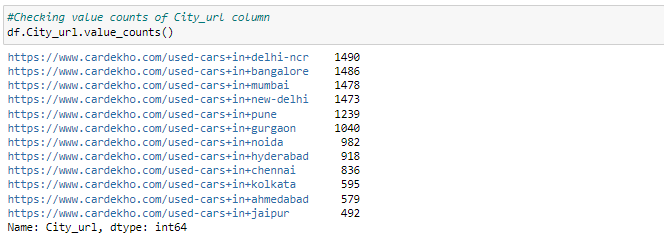


we have converted the data type of columns height, width and length to float datatype

## City\_url:

Let's extract city name from city url column.

Checking value counts of City\_url column



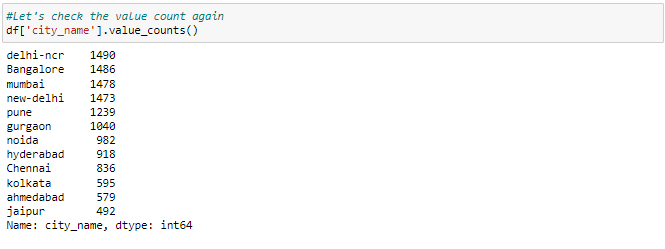
The above data shows the no of cars from each of the cities

Replacing city names from city urls



We have replaced the city names with city urls

Let's check the value count again



Since we have extracted city names let's drop City\_url.



We have dropped the city url column as it is redundant

## Seating\_cap:

Let's change the data type of seating\_cap to float type.



## Manu\_Year:

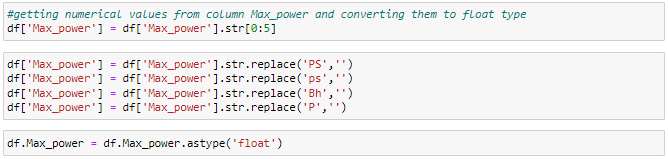
Let's extract car age from manufactured year.



## Max\_power:

We have to change the datatype of Max\_power column to float datatype.

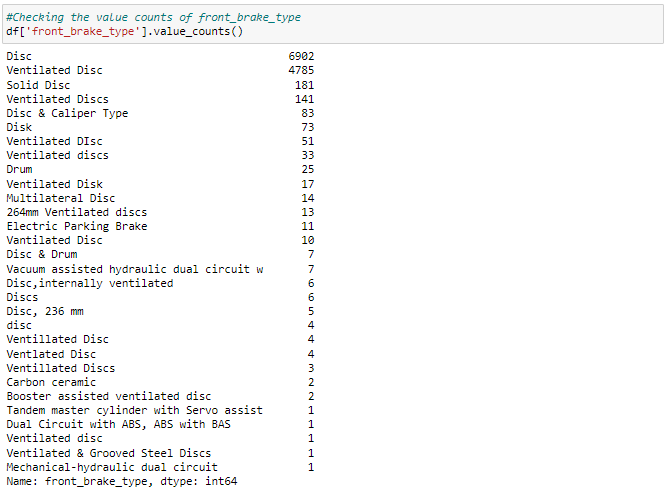
Getting numerical values from column Max\_power and converting them to float type



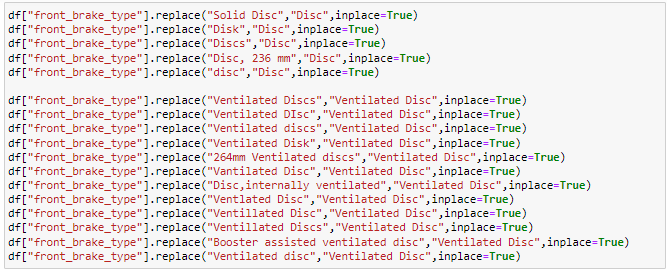
We have changed the datatype of Max\_power column to float datatype.

## front\_brake\_type:

Let's group the similar entries in this column.

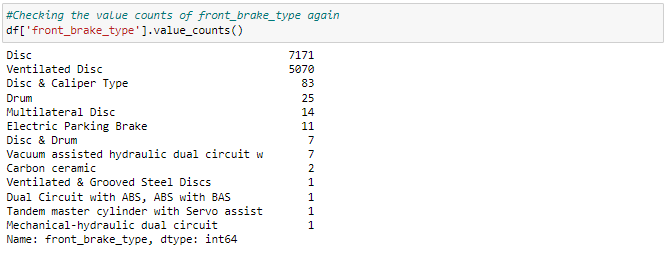


The 'front\_brake\_type' feature has few similar entries which shall be grouped for easy evaluation and better understanding



We have grouped all the similiar entries into one entry

Checking the value counts of front\_brake\_type again

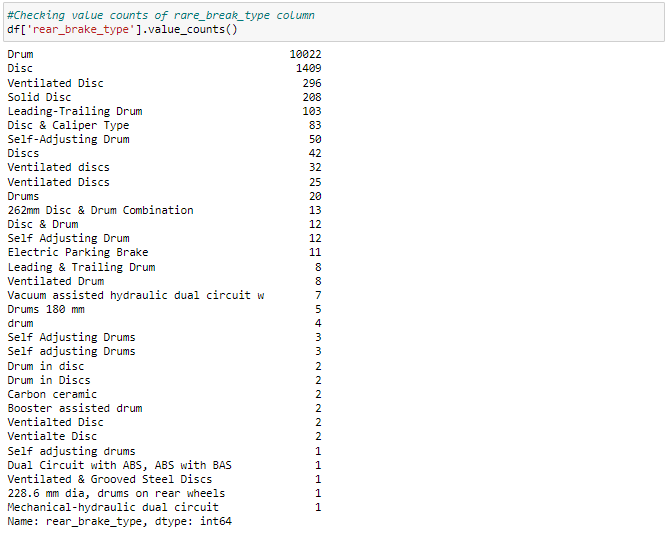


We can see that the value counts has reduced due to grouping of similar entries

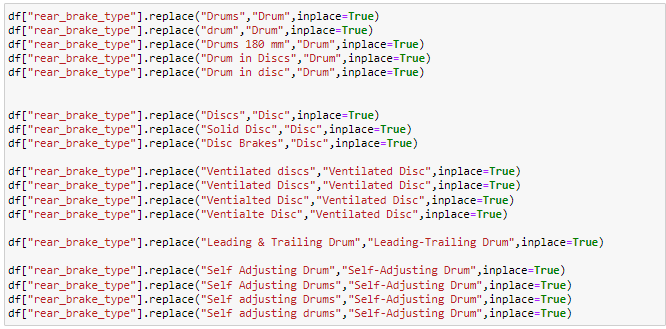
## rare\_brake\_type:

Let's group the similar entries in this column.

Checking value counts of rare\_break\_type column

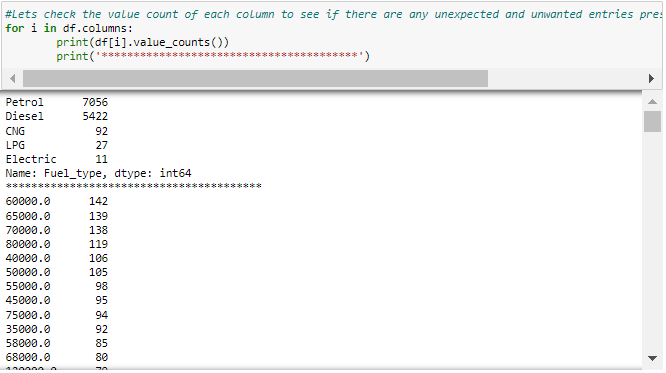


The 'rear\_brake\_type' feature has few similar entries which shall be grouped for easy evaluation and better understanding



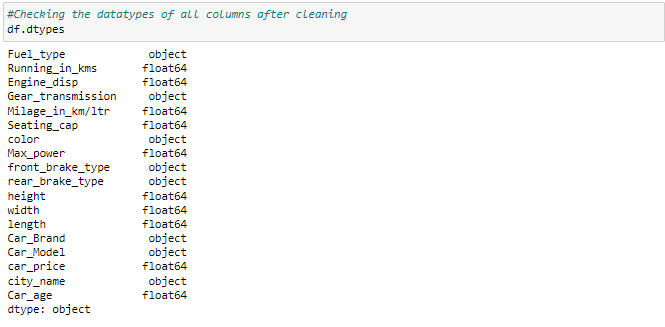
We have grouped all the similiar entries into one entry

checking the value count of each column



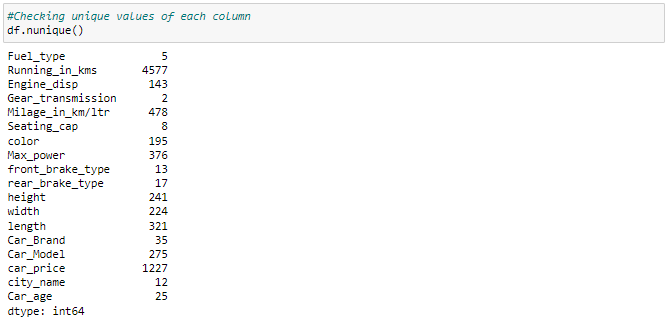
We can see that the value counts has reduced due to grouping of similar entries

Checking the datatypes of all columns after cleaning



Now in the dataset we have two types of data - Float and object type data.

Checking unique values of each column

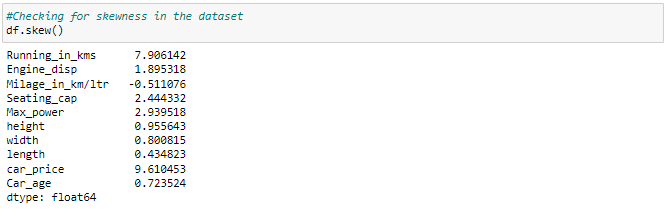


Above are the unique value counts of each column. We don't find anything unnecessary so let's proceed.

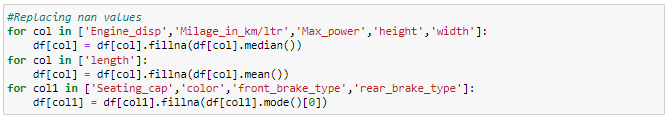
# Imputation technique to replace nan values



We have to replace the nan values in continuous columns by there mean and categorical columns with it's mode.

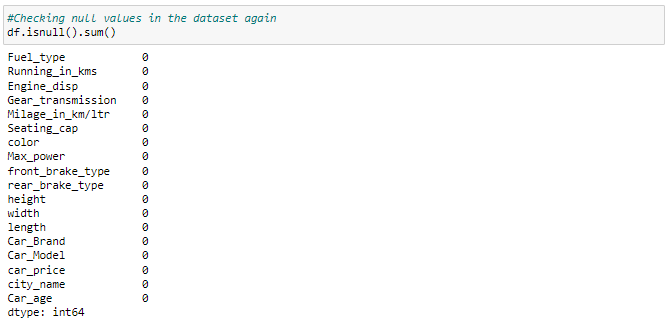


In the numerical columns with skewness null values has to be replaced by median.

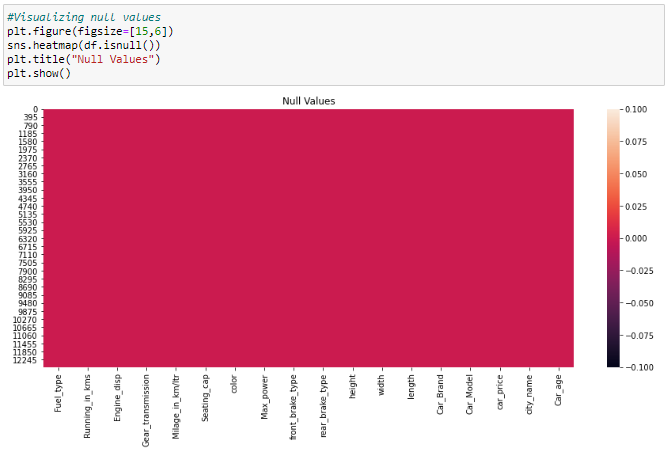


We have replaced all the necessary values

Checking null values in the dataset again



Visualizing null values

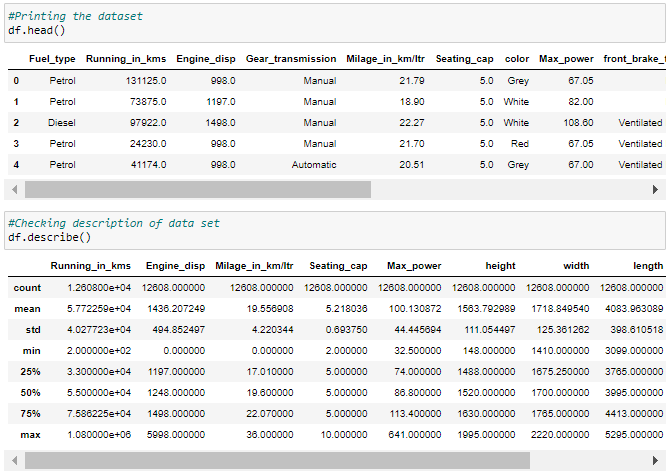


## Observations

As we have replaced all null values successfully, there are no null values present in the dataset

* **Data Inputs- Logic- Output Relationships**

Printing the dataset



* **Assumptions**
* All the columns have mean higher than the standard deviation. Standard deviation is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean.
* Only the 'Milage\_in\_km/ltr' column have the 50% percentile higher than the mean, which says that the distribution is negatively skewed.
* Al the other columns have mean higher than the 50th percentile. When the mean is greater than the median, the distribution is positively skewed.
* There is a significant difference between the max value and 75th percentile for all the columns, which indicates presence of outliers.
* **Hardware and Software Requirements and Tools Used**

To build the machine learning projects it is important to have the following hardware and software requirements and tools.

**Hardware required:**

* Processor: core i5 or above
* RAM: 8 GB or above
* ROM/SSD: 250 GB or above

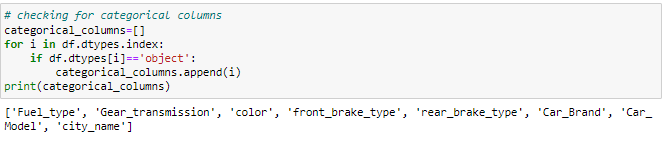
**Software required:**

* Distribution: Anaconda Navigator
* Programming language: Python
* Browser based language shell: Jupyter Notebook
* Chrome: To scrape the data

**Model/s Development and evaluation**

* **Visualizations**

# Univariate Analysis:

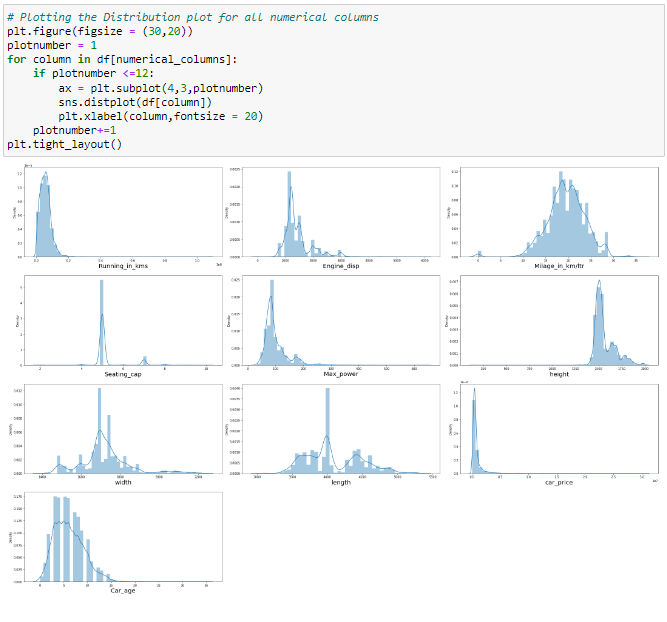


Above is the list of categorical columns.



Above is the list of numerical columns.

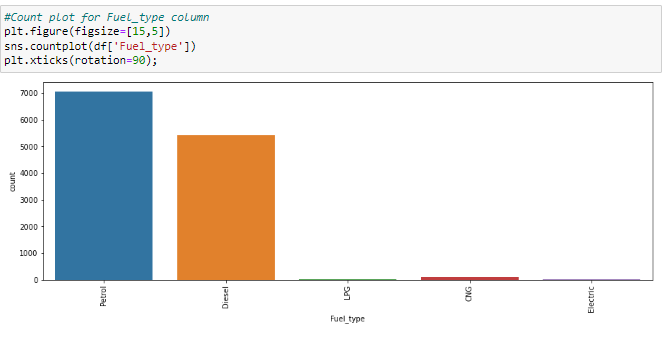
Plotting the Distribution plot for all numerical columns



## Observations

There is skewness in almost all the numerical columns.

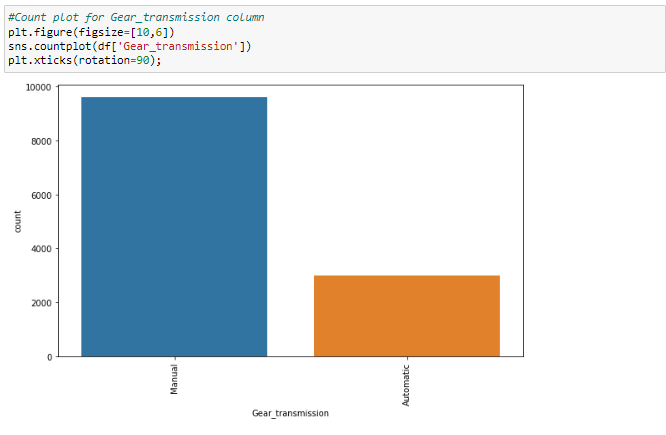
Count plot for Fuel\_type column



## Observations

* Maximum cars are petrol driven and diesel driven cars stand in next position.
* LPG, Electric and CNG are least used fuels for the cars in the given dataset

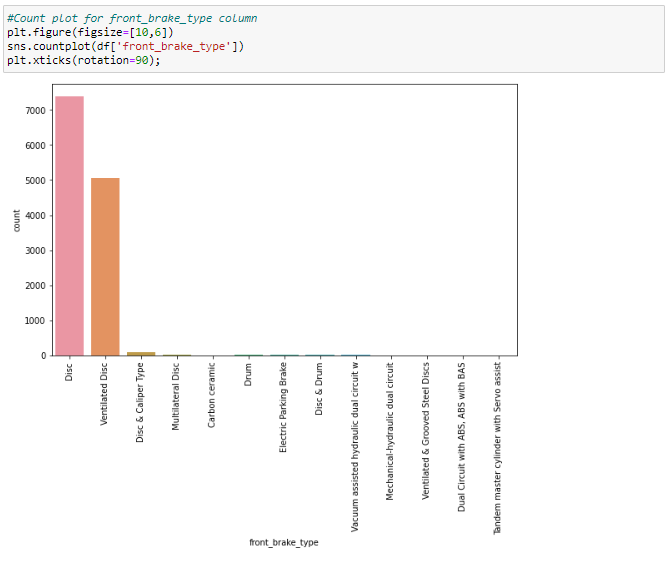
Count plot for Gear\_transmission column



## Observations

* Most of the cars have Manual gear transmission.
* Only few of the cars have automatic gear transmission compared to the manual transmission.

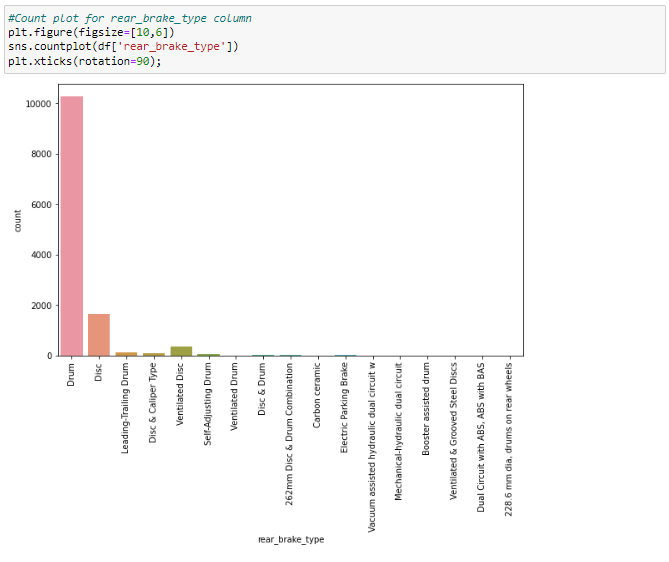
Count plot for front\_brake\_type column



## Observations

* Disc front brake cars are more in number followed by Ventilated Disc.

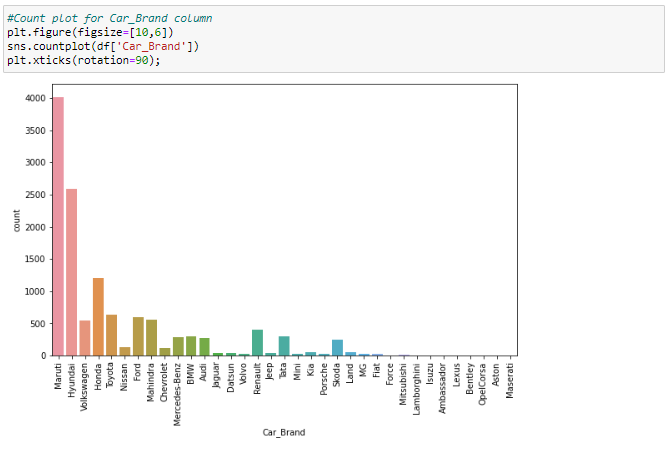
Count plot for rear\_brake\_type column



## Observations

Drum rare break cars are more in number followed by the Disc rear breaks

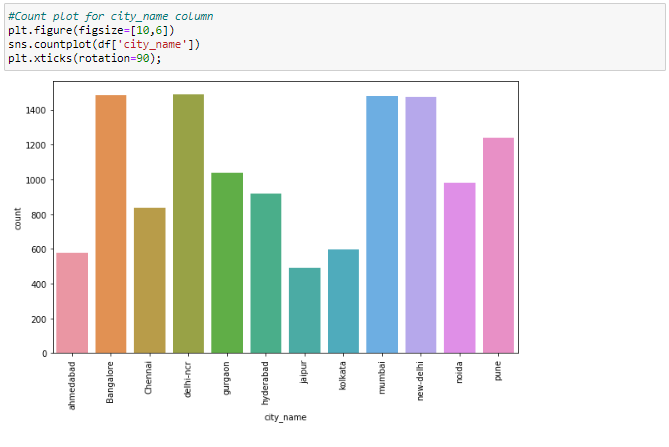
Count plot for Car\_Brand column



## Observations

* Maximum cars under sale in the daatset are Maruti followed by Hyundai.
* Foreign company cars are fewer than the indian made brands

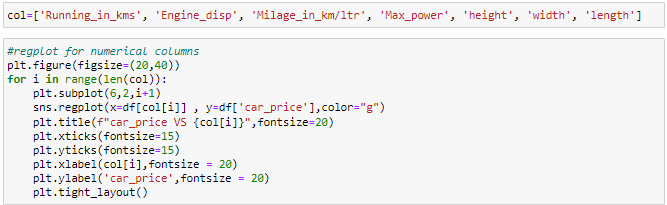
Count plot for city\_name column

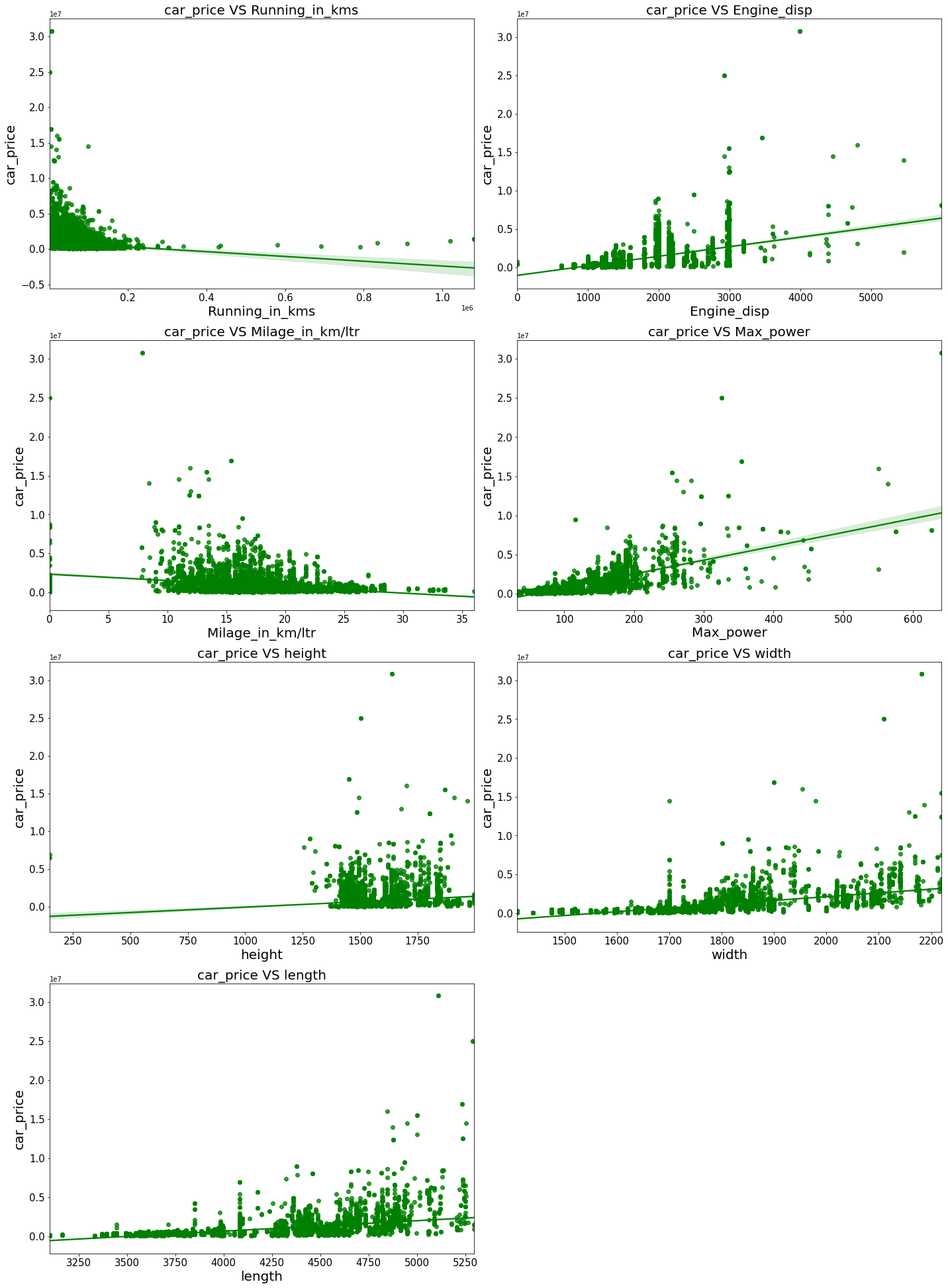


## Observations

* In Bangalore,delhi-ncr,mumbai and new-delhi we can find maximum cars for sale. Since these are densly populated places.
* Jaipur and ahmedabad have least number of cars in the dataset

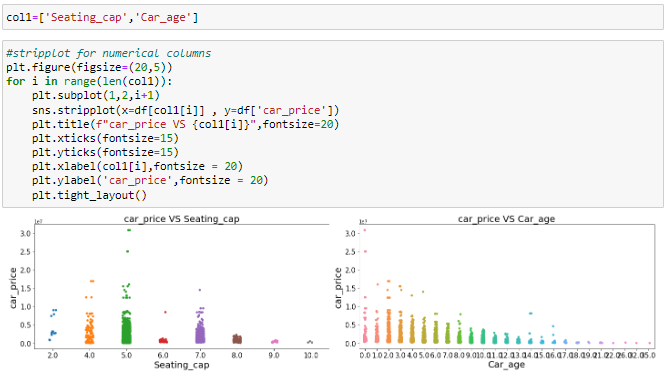
# Bivariate Analysis:





## Observations

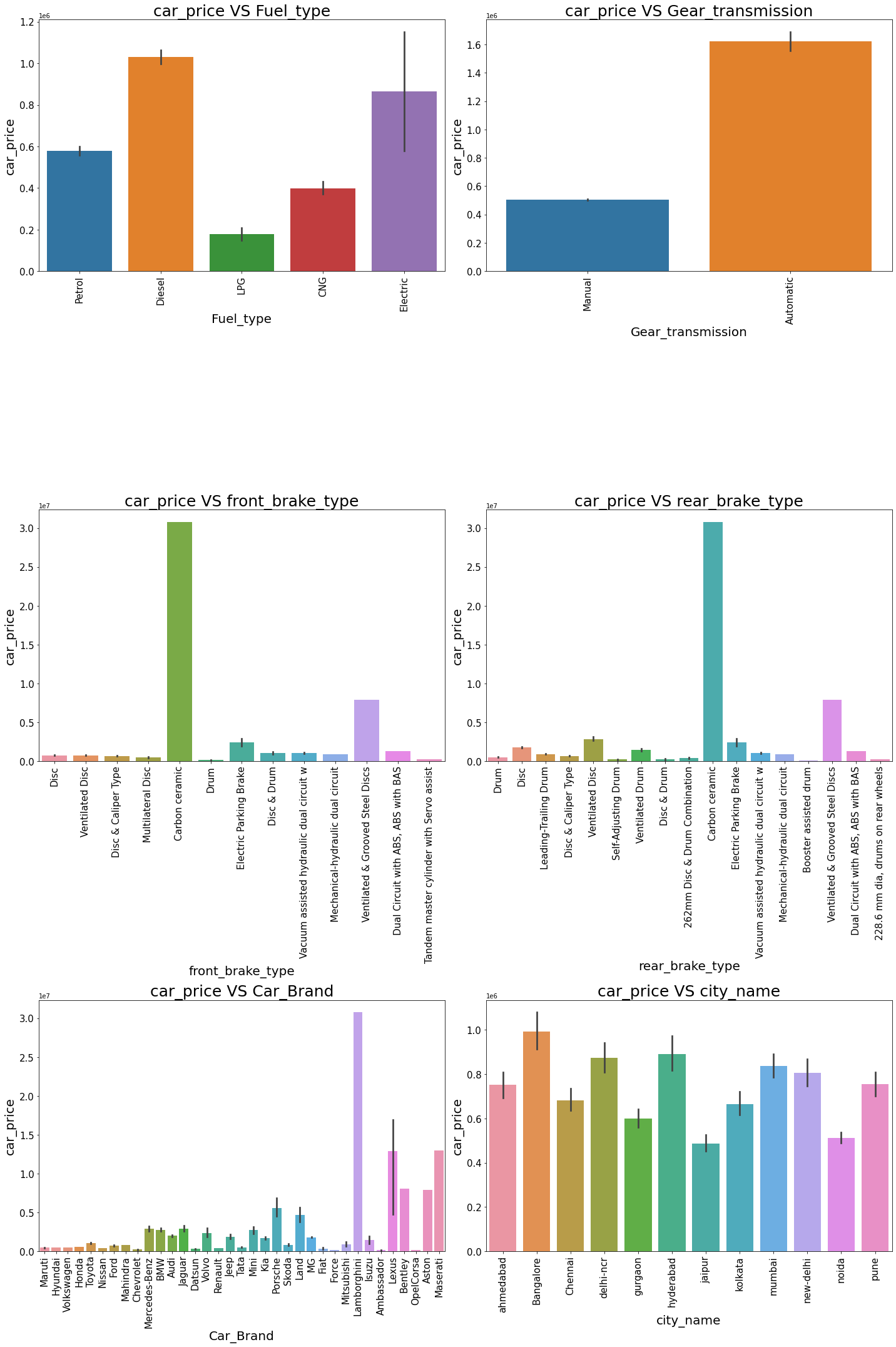
* Maximum cars are having below 20k driven kms. And car price is high for less driven cars.
* Maximum cars are having 1000-3000 Endine\_disp. And car price is high for 3000 Endine\_disp.
* Maximum cars are having milage of 10-25kms. And ,milage has no proper relation with car price.
* As Max\_power is increasing car price is also increasing.
* Car\_price has no proper relation with height.
* As the width is increasing car price is also increasing.
* As length is increasing car price is also increasing.
* Weight also has linear relationship with car price.
* As top\_speed is increasing car price is also increasing.

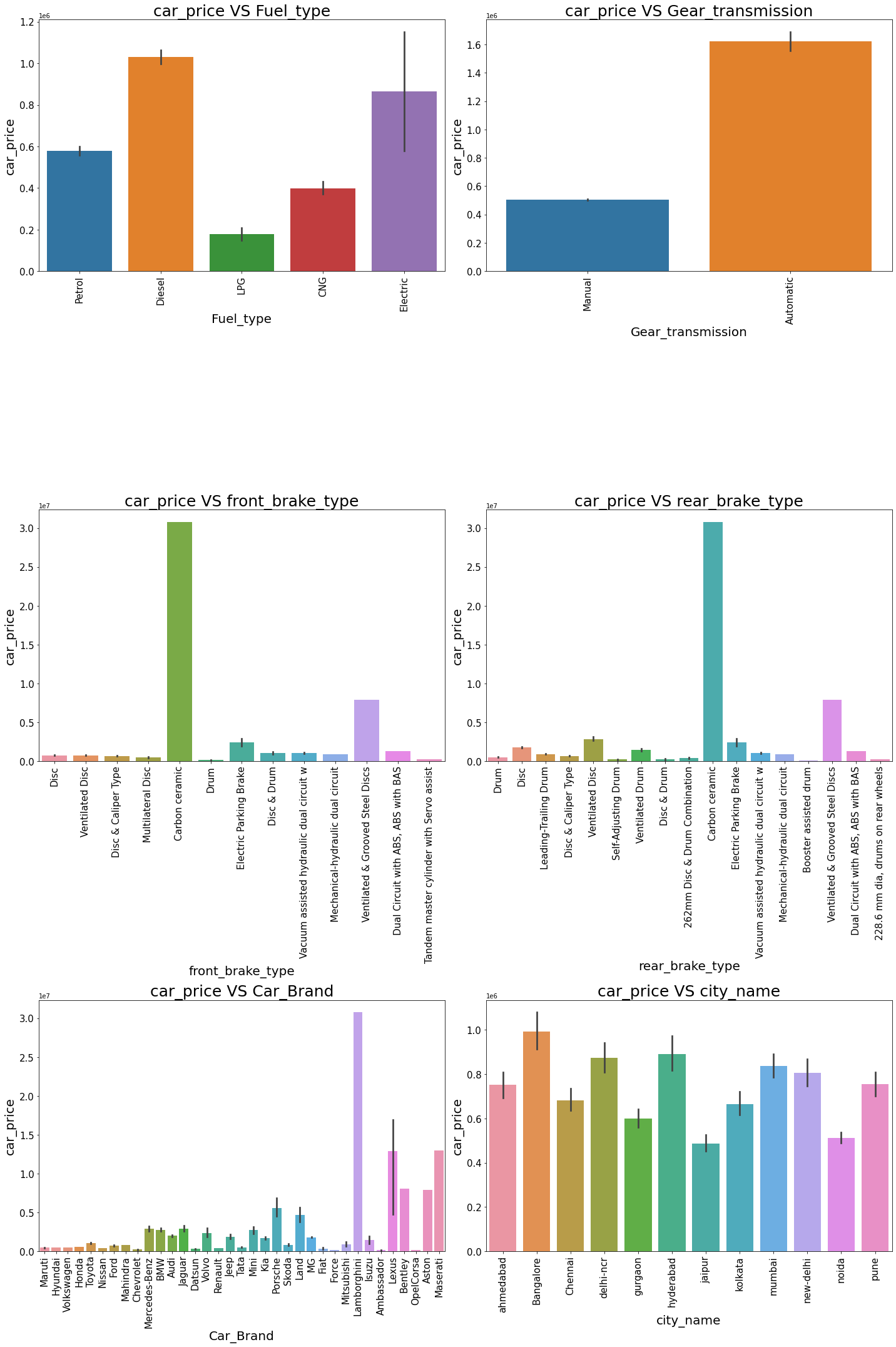


## Observations

* Cars with 5 and 4 seats are having highest price.
* 5 seater cars have higher prices than the other cars.
* As the age of the car increases the car price decreases. Lesser the age of the car, higher the price.

Bar plot for all categorical columns

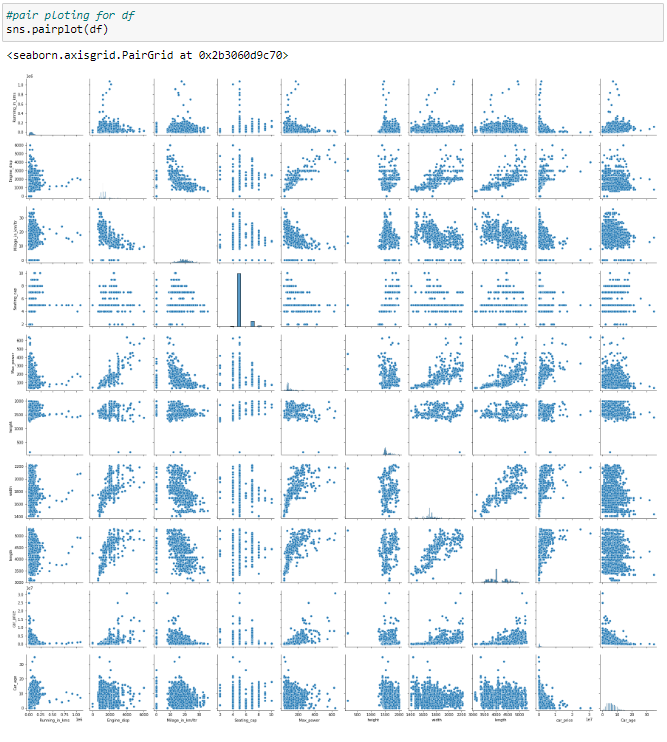




## Observations

* For Diesel and Electric cars the price is high compared to Petrol,LPG and CNG.
* Cars with automatic gear are costlier than manual gear cars.
* Cars with Carbon Ceramic front break are costlier compared to other cars.
* Cars with carbon Ceramic rear braek are costlier compared to other cars.
* Lamborghini brand cars are having highset sale price.
* In Bangalore, Hyderabad and delhi-ncr the car prices are high as they are highly populated cities.

Pair plotting for df

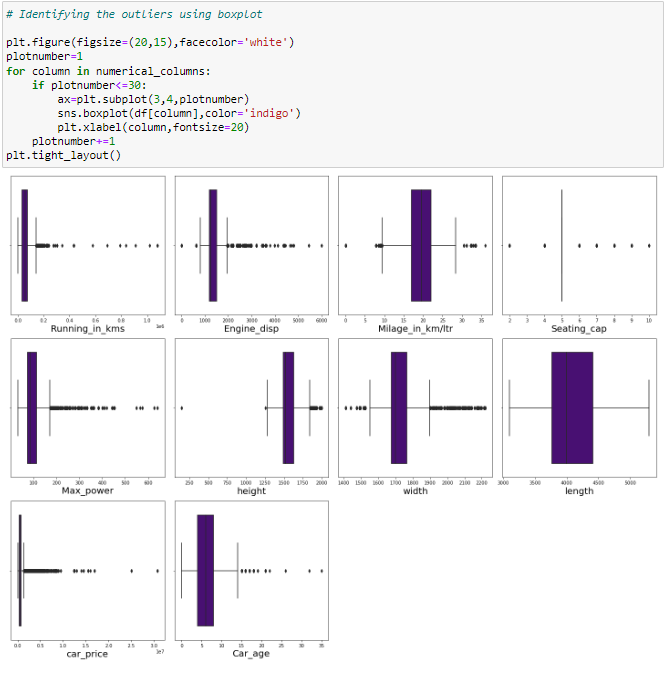


## Observations

By looking into the pair plot of pair of features we can notice the presence of outliers in each plot so we have to deal with this in the later steps.

* **Identification of possible problem-solving approaches (methods)**

# Checking for outliers:

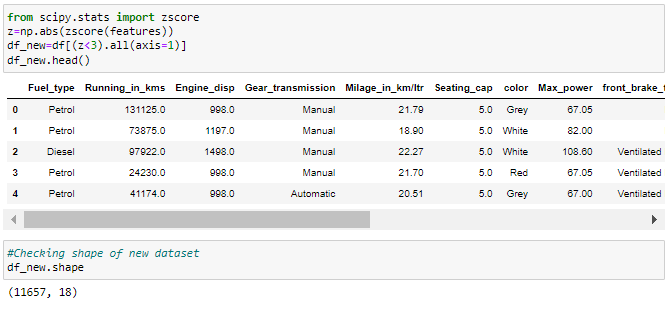
****

## Observations

* There are outliers in all columns except length.
* Since car\_price is our target we should not remove outliers from it.
* Removing Outliers:

****

Above are the list of columns with outliers in the dataset.

****

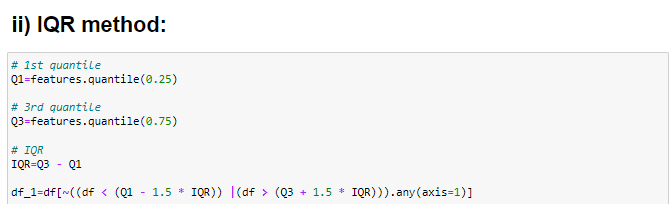
The new dataset has 11657 rows and 18 columns.

****

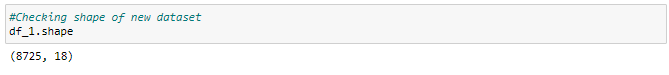
Previously the dataset has 12608 rows and 18 columns.

Checking dataloss in zscore method

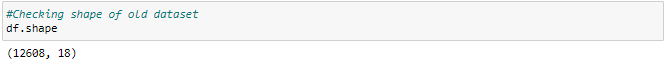
****

****

We have removed the skewness of the dataset using IQR method.

****

The new dataset has 8725 rows and 18 columns.

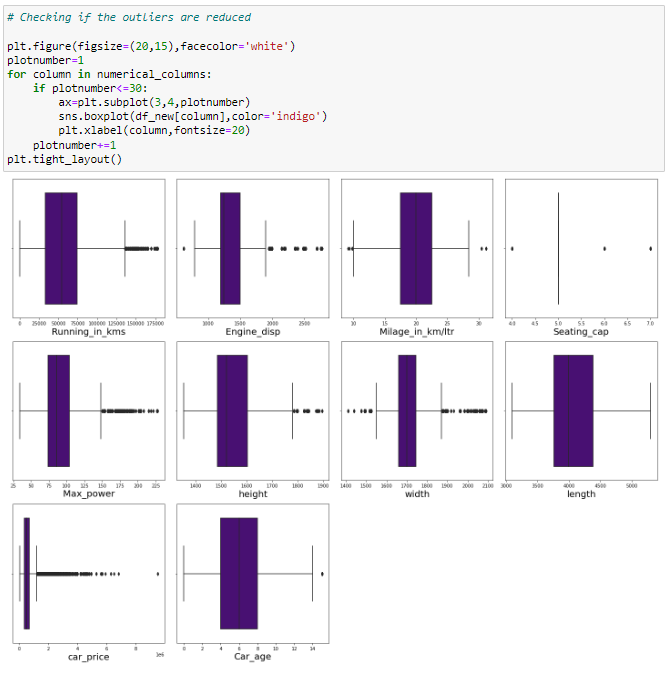
****

Previously the dataset has 12608 rows and 18 columns.

****

In IQR method the data loss is more than 10%. So let us continue with the dataset obtained after removing the outliers using Z-score method

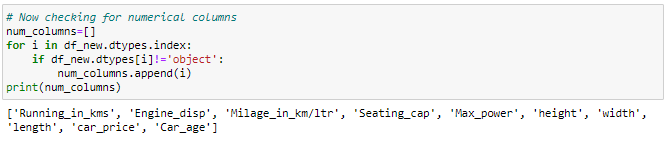
Checking the outliers again

****

## Observations

Outliers has been reduced in all the columns.

# Checking for skewness:

****

# Checking for skewness in the dataset

****

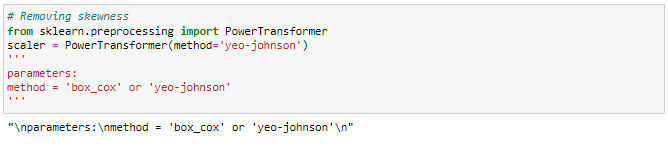
## Observations

We can notice there is skewness in all the numerical columns except Milage\_in\_km/ltr,width,length and Car\_age. So we have to remove this skewness. Since car\_price is my target no need to remove skewness in this column.

# Removing skewness using yeo-johnson method:

****

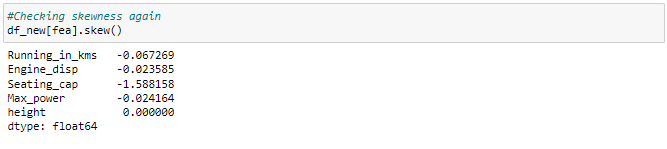
Taking a list as fea with all the columns with skewness.

****

Using yeo\_johnson method for removing the skewness.

****

Got removed from skewness.

****

## Observations

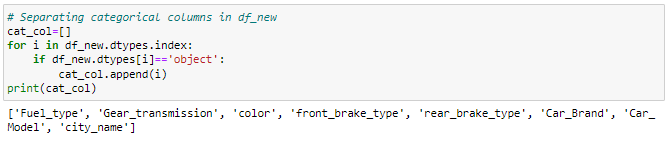
In all the columns skewness has reduced and in height column skewness is zero after removing which means this column has single entry throught out. So let us drop this column as it has no impact on model building.

Dropping height column

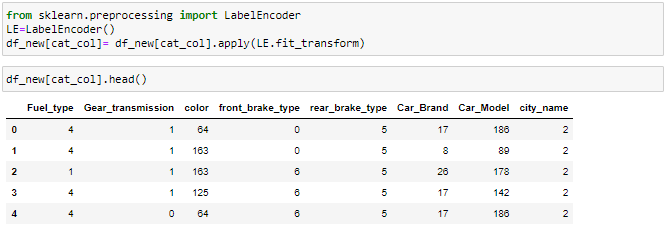
****

# Label Encoding:

Separating categorical columns in df\_new

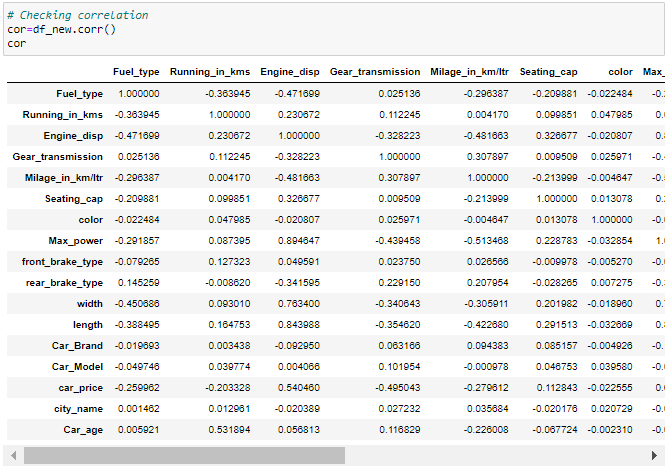
****

Above are the list of categorical columns in df\_new.

****

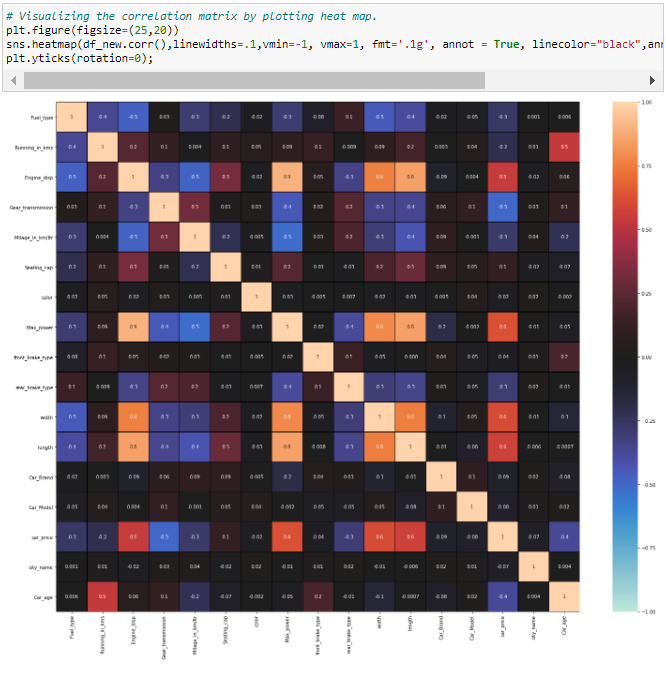
Using label encoder we have encoded the categorical columns.

# Correlation

****

## Observations

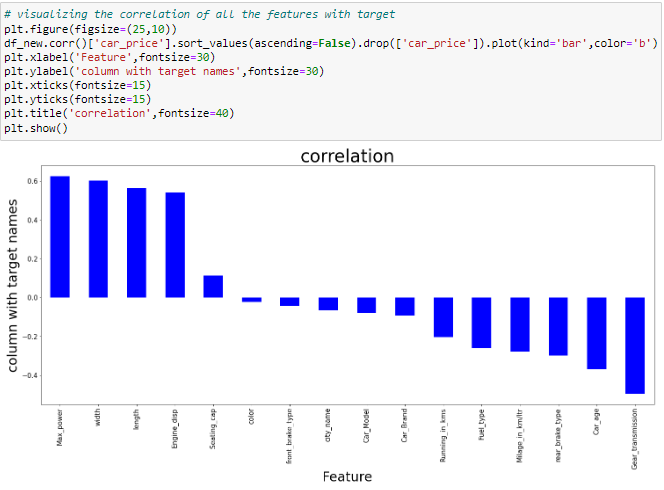
Above are the correlations of all the pair of features.To get better visualization on the correlation of features, let us plot it using heat map.

****

## Observations

* Most of the columns in the darker shades are near by zero, which says they donot have much impact on the dataset.
* We can notice there is multicolinearity issue in the dataset. So we have to use VIF to remove multicolinearity.
* Width, lngth, and max power are highly correlated with the engine\_disp column which causes multicollinearity in the model
* Our target column 'Car\_price' is highly correlated with Width, lngth, and max power
* 'Gear\_transmission' is highly negatively correlated with the target column 'Car\_price'.
* The columns 'color', 'front\_brake\_type', 'Car\_brand', 'Car\_model', 'city\_name' have near zero correlation with the target colunn which says there is no impact on the target column due to any change in these columns

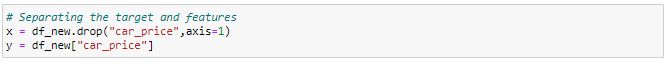
Let's visualize the correlation of all the features with target to get better insight.

****

## Observations

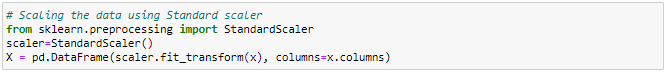
Color is less correlated with target. But will keep it and continue.

# Separating Features and Target:

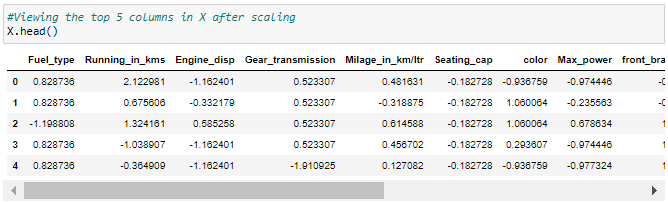
****

We have separated my target and independent columns.

# Scaling

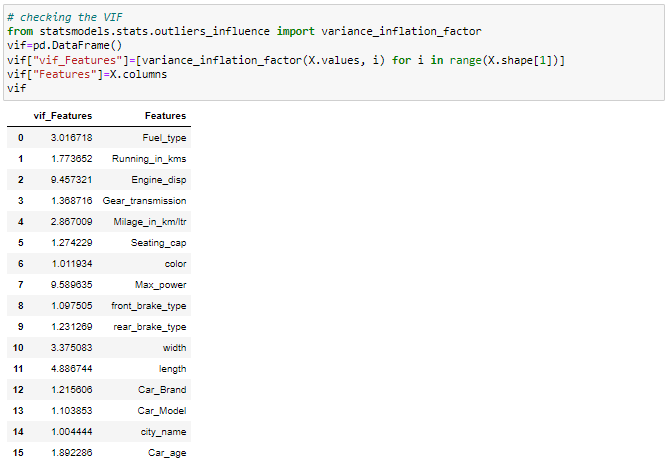
****

I have scaled my data using Standard scaler.

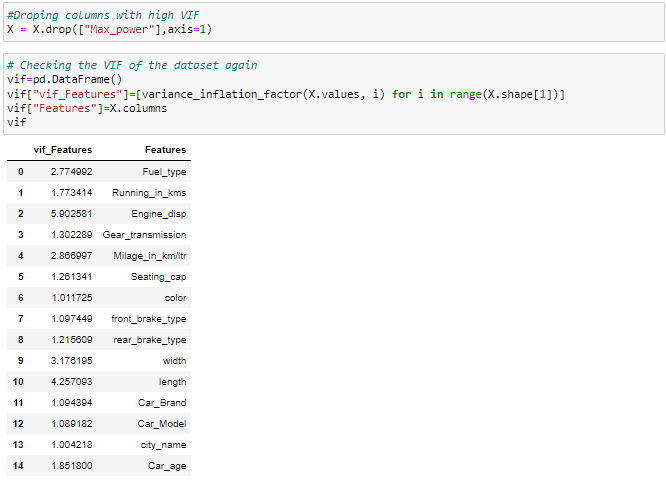
****

This is the data of independent variables after scaling.

# Checking for multicollinearity

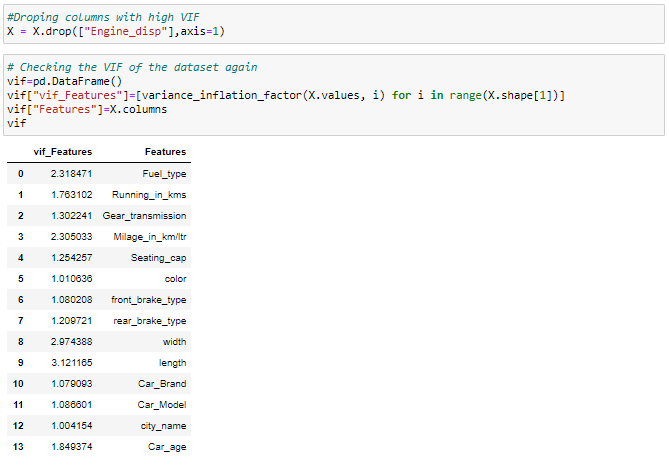
****

Dropping columns with high VIF

****

We can see that the "Engine\_disp" has high VIF, Hence we shall drop this column for better model building

Dropping columns with high VIF

****

Now we can see that the multicollinearity issue is solved.

# Finding Best Random State and Accuracy:

****

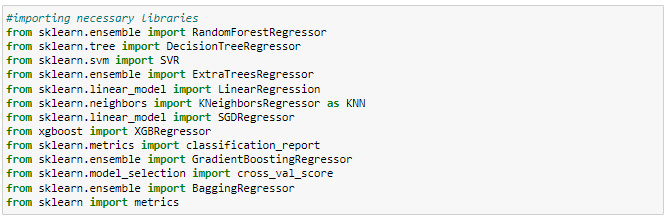
We have got the best accuracy and random state.

## Creating Train test split

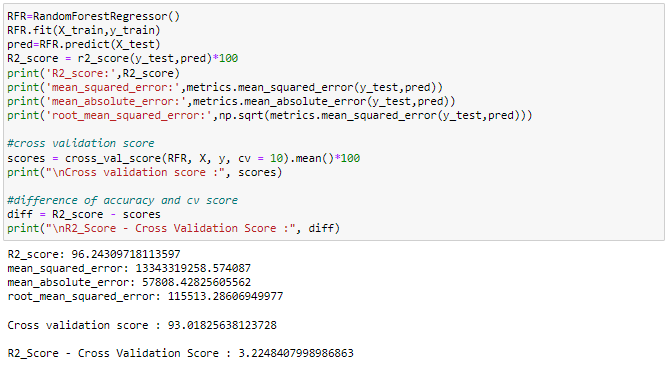
****

Created train test split.

# Regression Algorithms:

****

# i) RandomForestRegressor:

****

RFR is giving 96.24% r2\_score.

# ii) XGBRegressor:

****

XGBRegressor is giving 97.10% r2\_score.

# iii) GradientBoostingRegressor:

****

GradientBoostingRegressor is giving 93.12% r2\_score.

# iv) DecisionTreeRegressor:

****

DecisionTreeRegressor is giving 93.17% r2\_score.

# v) Bagging Regressor:

****

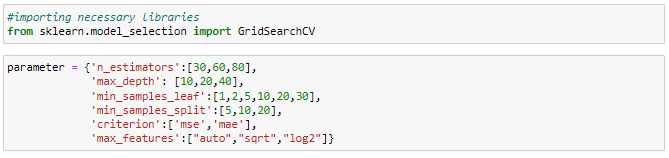
Bagging Regressor is giving 94.17% r2\_score.

The XGBReggressor has 97.10 as r2\_score but has a higher difference between the r2\_score and the Cross validation score than the Random forest regressor

**By evaluating based on the difference of model accuracy i.e., r2\_score and cross validation score we can say that RandomForestRegressor as the best model with 96.24% r2\_score.**

* **Testing of Identified Approaches (Algorithms)**

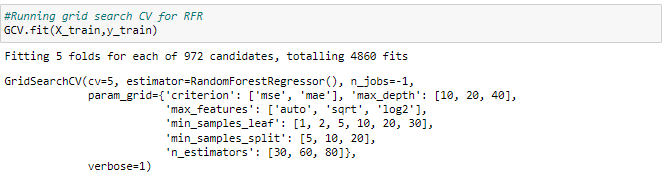
# Hyper parameter tuning for best model:

****

Giving RandomForestRegressor parameters.

****

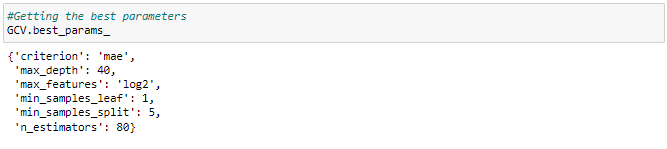
Running grid search CV for RandomForestRegressor.

****

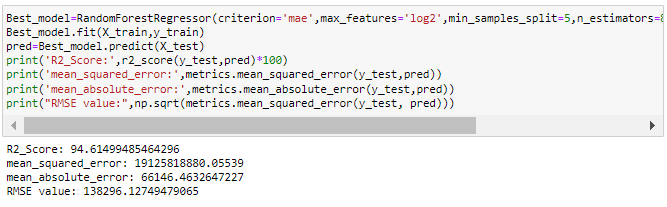
* **Run and Evaluate selected models**

Tuning the model using GCV.

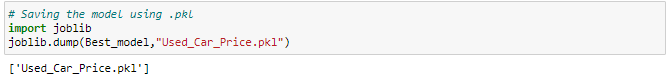
Obtaining the best parameters

****

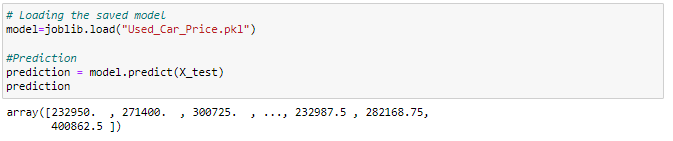
Got the best parameters for RandomForestRegressor.

****

**This is the final model with 94.61% as r2\_score after tuning which is good.**

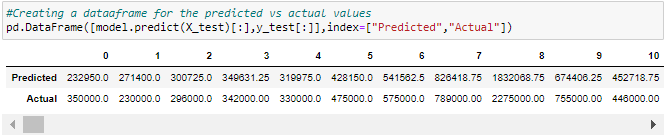
****

# Saving the model:

****

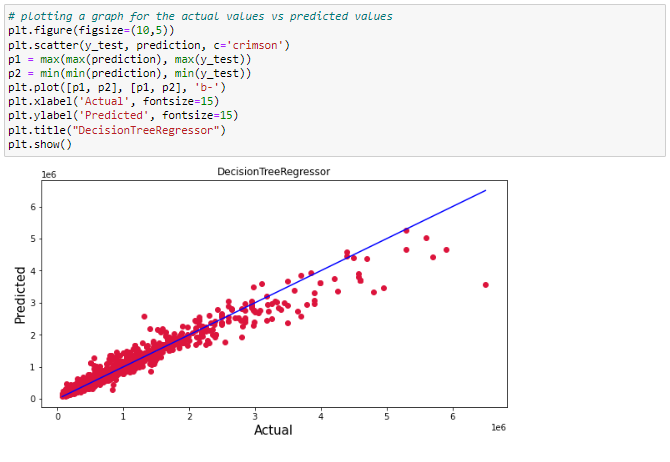
We have saved the model as Used\_Car\_Price Using .pkl

# Predictions:

****

Above are the predicted values and the actual values, which look similarwith a few exceptions

Plotting a graph for the actual values vs predicted values

****

## Observation

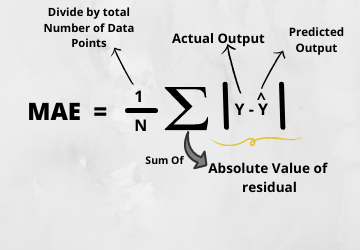
We have plotted the Actual vs Predicted. To get better insights - Bule line is the actual line and red dots are the predicted values

We can observe that the predicted values are nearby the actual values which says the model built works well.

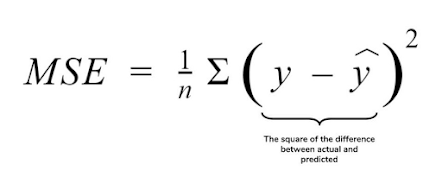
* **Key Metrics for success in solving problem under consideration**

The essential step in any machine learning model is to evaluate the accuracy and determine the metrics error of the model. I have used Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R2 Score metrics for my model evaluation:

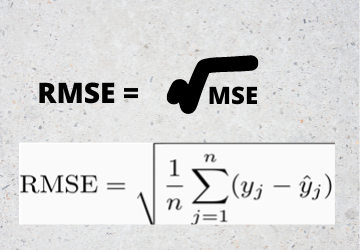
* **Mean Absolute Error (MAE):** MAE is a popular error metric for regression problems which gives magnitude of absolute difference between actual and predicted values. The MAE can be calculated as follows:



## **Mean Squared Error (MSE):** MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value. We perform squared to avoid the cancellation of negative terms and it is the benefit of MSE.



## **Root Mean Squared Error (RMSE):** RMSE is an extension of the mean squared error. The square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.



* **R2 Score:** I have used R2 score which gives the accurate value for the models used. On the basis of R2 score I have created final model.
* **Interpretation of the Results**

The dataset should be cleaned and scaled to build the ML models to get good predictions. I have performed few processing steps which I have already mentioned in the pre-processing steps where all the important features are present in the dataset and ready for model building.

In univariate analysis I have used count plots and pie plots to visualize the counts in categorical variables and distribution plot to visualize the numerical variables. In bivariate analysis I have used reg plots, bar plots, strip plots, line plots and violin plot to check the relation between label and the features. Used pair plot to check the pairwise relation between the features. The heat map and bar plot helped me to understand the correlation between dependent and independent features. Also, heat map helped to detect the multicollinearity problem and feature importance. Detected outliers and skewness with the help of box plots and distribution plots respectively. And I found some of the features skewed to right as well as to left. I got to know the count of each column using bar plots.

Aftercleaning and processing data, I performed train test split to build the model. I have built multiple regression models to get the accurate R2 score, and evaluation metrics like MAE, MSE and RMSE. I got Extreme Gradient Boosting Regressor (XGB) as the best model which gives 96.27% R2 score. I checked the cross-validation score ensuring there will be no overfitting and underfitting. After tuning the best model, the R2 score of Extreme Gradient Boosting Regressor has been increased to 96.90% and also got low evaluation metrics. Finally, I saved my final model and got the good predictions results for price of used cars.

**Conclusion**

* **Key Findings and Conclusions of the Study**

The case study aims to give an idea of applying Machine Learning algorithms to predict the sale price of the used cars. After the completion of this project, we got an insight of how to collect data, pre-processing the data, analyzing the data, cleaning the data and building a model.

In this study, we have used multiple machine learning models to predict the sale price of the used cars. We have gone through the data analysis by performing feature engineering, finding the relation between features and label through visualizations. And got the important feature and we used these features to predict the car price by building ML models. After training the model we checked CV score to overcome with the overfitting issue. Performed hyper parameter tuning on the best model and the best model’s R2 score increased and was giving R2 score as 96.90%. We have also got good prediction results of car price.

From the whole study we got to know that the continuous numerical variables having some strong positive linear relation with the label "Car\_Price". By comparing car price and categorical variables we got to know that the cars having automatic gear transmission, cars from the city Bangalore, cars using petrol and diesel as fuels, cars having the brands Benz and BMW and cars with 5-7 seating capacity have high sale price. While comparing continuous numerical variables and Car\_Price we found that cars which are having good mileage, engine displacement, less running in kms have good linear relation with the price that is the cars with this kind of qualities have high selling prices. We found outliers and removed them and further removed skewness. Looking at the heat map, I could see there were some features which were correlated with each other, so I used VIF method to remove the feature causing multicollinearity and scaled the data to overcome with the data biasness.

* **Learning Outcomes of the Study in respect of Data Science**

While working on this project I learned many things about the features of cars and about the car selling web platforms and got the idea that how the machine learning models have helped to predict the price of used cars which provides greater understanding into the many causes and benefits of selling old cars. I found that the project was quite interesting as the dataset contains several types of data. I used several types of plotting to visualize the relation between target and features. This graphical representation helped me to understand which features are important and how these features describe selling price of the old cars. Data cleaning was one of the important and crucial things in this project where I dealt with features having string values, features extraction and selection. Finally got XGBoosting Regressor as best model.

The challenges I faced while working on this project was when I was scrapping the real time data from cardekho website, it took so much time to gather data. The data was quite difficult to handle and cleaning part was challenging for me but fixed it well and it was unable to remove skewness in Seating\_cap column so moved further keeping it as it is.

Finally, our aim was achieved by predicting the sale price of used cars and built car price evaluation model that could help the clients to understand the future price of used cars.

* **Limitations of this work and Scope for Future Work**

The main limitation of this study is the low number of records that have been used.In the dataset our data is not properly distributed in some of the columns many of the values in the columns are “-“and some values which are not realistic. Because I have seen the column running in kms showing 0 kms and some of the cars having age as 0 years which are not possible in case of used cars. So, because of that data our models may not make the right patterns and the performance of the model also reduces. So that issues need to be taken care.

**Future work:** As future work, we intend to collect more data and to use more advanced techniques like artificial neural networks and genetic algorithms to predict car prices. In future this machine learning model may bind with various website which can provide real time data for price prediction. Also, we may add large historical data of car price which can help to improve accuracy of the machine learning model.