**PROJECT REPORT**

**ON**

**Secondhand Cars Price Prediction**



**Submitted by: Gyan Prakash Tripathi**

**ACKNOWLEDGMENT**

Approximately 40 million used vehicles are sold each year. Effective pricing strategies can help any company to efficiently sell its products in a competitive market and make a profit this blog,

The purpose of this project is to conduct a mini-project that encompasses the bulk of the data science process — from data collection (web-scraping: Beautiful Soup, Python), data cleaning, exploratory data analysis, model training, and testing stage.

The source of data comes from [cars24,](https://www.sgcarmart.com/main/index.php) and [www.cardekho.com](http://www.cardekho.com) , which are online car sales portals in India.

I will be **analysing the Old Cars Selling Price prediction using a Machine Learning dataset** using essential exploratory data analysis techniques and also, and I will be performing some data visualizations to better understand our data.

By doing data preprocessing, data analysis, feature selection, and many other techniques we built our cool and fancy machine learning model. And at the end, we applied many ml algorithms to get the very good accuracy of our model.

**Many thanks to Fliprobo Technology for providing me with this project to understand the Real-Time Field work present in Data Science Industry.**

I am very thankful to my friends and family who helped me through this study. So without any further due.

**ABSTRACT**

The prices of new cars in the industry are fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used car sales are on a global increase (Pal, Arora and Palakurthy, 2018). There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offer this service, their prediction method may not be the best. Besides, different models and systems may contribute to predicting the power of a used car’s actual market value. It is important to know their actual market value while both buying and selling.

**TAKEAWAYS FROM THE BLOG**

In this article, we do prediction using machine learning which leads to the below takeaways:

1. **Web Scraping:** Scraping data from websites like cars24, cardekho.com and olx.
2. **EDA:** Learn the complete process of EDA
3. **Data analysis:** Learn to withdraw some insights from the dataset both mathematically and visualize it.
4. **Data visualization:** Visualizing the data to get better insight from it.
5. **Feature engineering:** We will also see what kind of stuff we can do in the feature engineering part.
6. Eliminating features that had an insignificant effect on the response variable by evaluating the p-values and R² value of the mode

**PROBLEM STATEMENT**:

With the covid 19 impact in the market, we have seen a 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 the clients works with small traders, who sell used cars. With the change in the market due to covid 19 impact, THE 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.

**Model Building Phase** After collecting/scraping the data, we have around 5000 rows and 9 columns. We need to build a machine learning model. Before model building, we will be doing data pre-processing steps. We will try different models with different hyper parameters and select the best model.

**ABOUT THE DATASET**

**About the data:**

1. Number of features in dataset: 9

2. Number of data points in dataset: 5058

We have scraped price of 5058 old cars from Cars24 and cardekho.com This problem involves predicting the Selling prices of the old cars which are continuous and real-valued outputs. Thus, this is a **Regression Problem.**

**Features:**

**Here's a brief version of what features is in the data description file:**

1. **Brand Name:** Model and brand name of the car.

2. **Year**: Year of manufacturing/ model launch year

3. **Car\_variant**: Model type

4. **Selling Price:** The price on which the car is available on the website

5. **Kilometers\_Driven:** How many km the car is driven till now

6. **Fuel\_Type:** if the model run by petrol or diesel

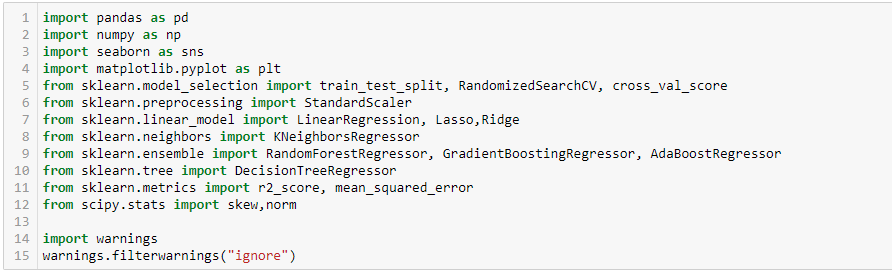
7. **Transmission:** If the model is run manually or automatic

8**. Owner\_Type**: How many owners are there or what source of car

9. **Location:** Which city of India the car is available importing Important Libraries:

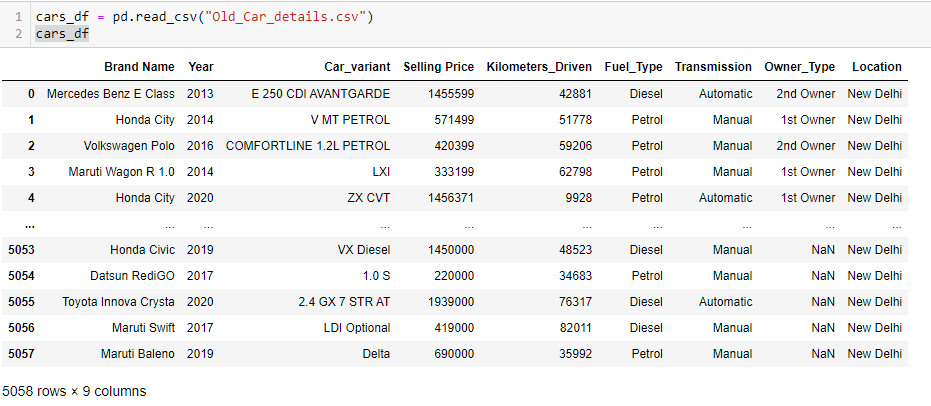
**Importing Important libraries:**

We need some libraries to be imported to work upon the dataset, we would import the dataset by using pandas’ read\_csv method.



**Loading Data Set into a variable:**

**Here I am loading the dataset into the variable cars\_df.**

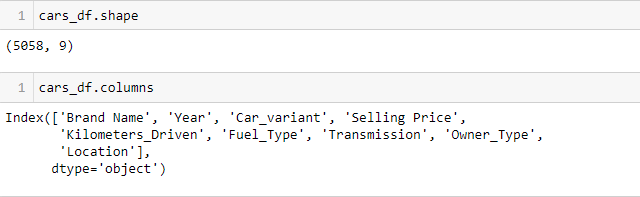


Selling Price is the target variable.

**Exploratory Data Analysis:**

E**xploratory Data Analysis** refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. We performed some bi-variate analysis on the data to get a better overview of the data and to find outliers in our data-set. Outliers can occur due to some kind of errors while collecting the data and need to be removed so that it doesn’t affect the performance of our model.

**Checking the shape of the dataset:**

****

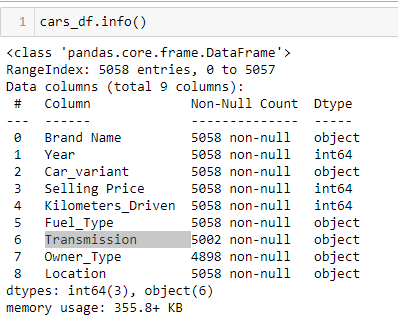
We can see that the dataset has 5058 rows and 9 columns..

**Getting detailed information about both the datasets:**

We have two datasets.

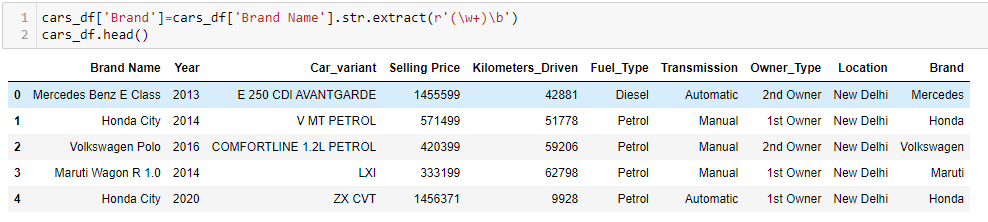
The dataset has 5058 observations and 9 columns including the target variable. The target variable is Selling price which is of integer data type.

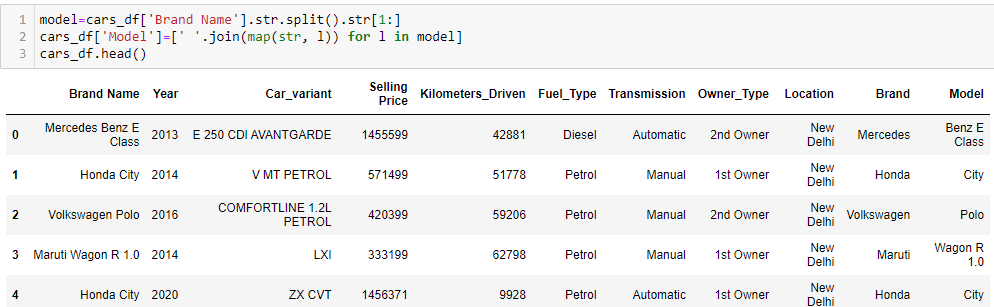
After applying the info() function to the datasets we get to know the data types of all the features and missing values of both the datasets.



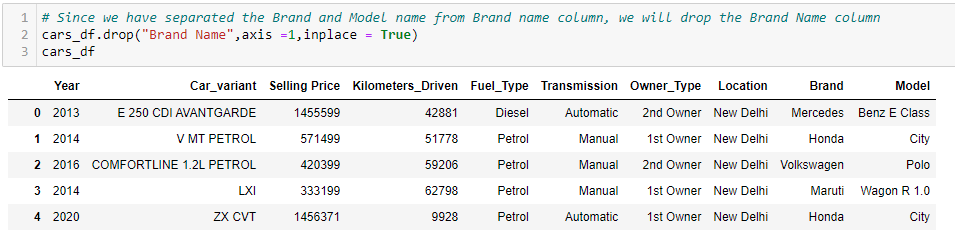
We concluded that we have 3 numerical and rest 6 categorical data types features in the datasets. Also, there are so many missing values in the Transmission and Owner type column.

**Separating Brand and Model name From Brand name column:**

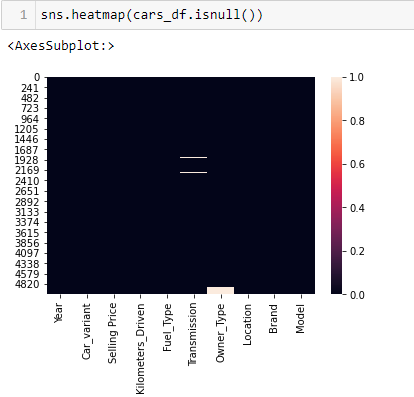
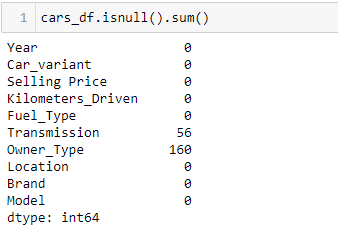
****

****

Now we can drop the brand name column as we have separate columns for brand and model

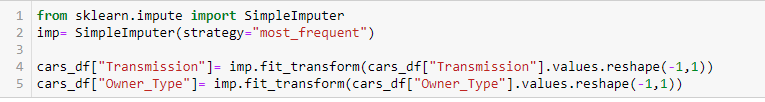


**Checking the Null Values:**

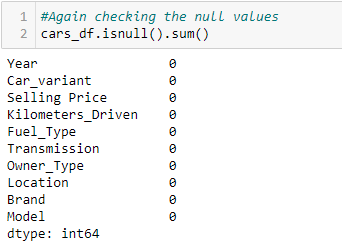


**Treating the Missing values in the dataset:**

We have null values in Transmission and Owner Type and both are categorical in nature. So we will apply Simple Imputer (most\_frequent) strategy



Checking the null values again

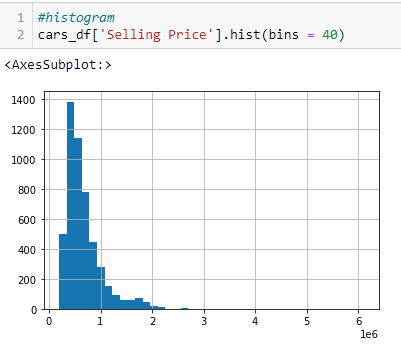
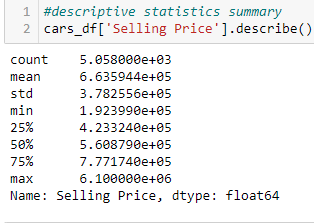
****

Now there is no null values in both the datasets.

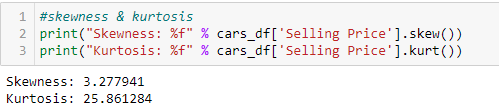
**Univariate Analysis:**

Uni means one, so in other words, the data has only one variable. Univariate data requires analysing each variable separately.  It doesn't deal with causes or relationships (unlike regression) and its major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

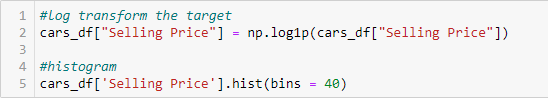
**Analyzing Target variable:**

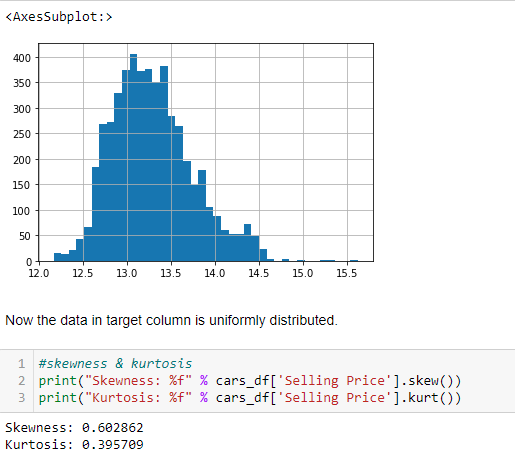
 

We can clearly see that the target variable has a normal distribution that is skewed towards the right. Now let's calculate the Skewness and Kurtosis:



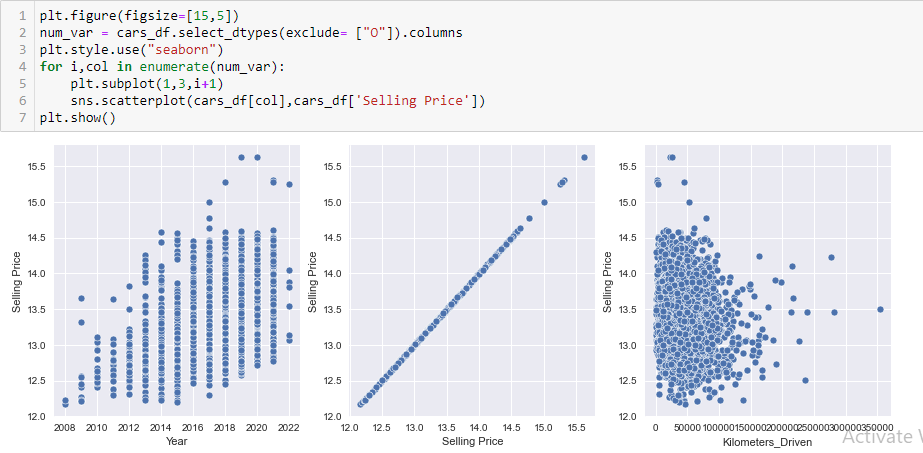
As we saw before, the target variable "Selling Price" is not uniformly distributed and it's skewed towards the right. Therefore**, we will try to use the log transformation to remove the skewness**.

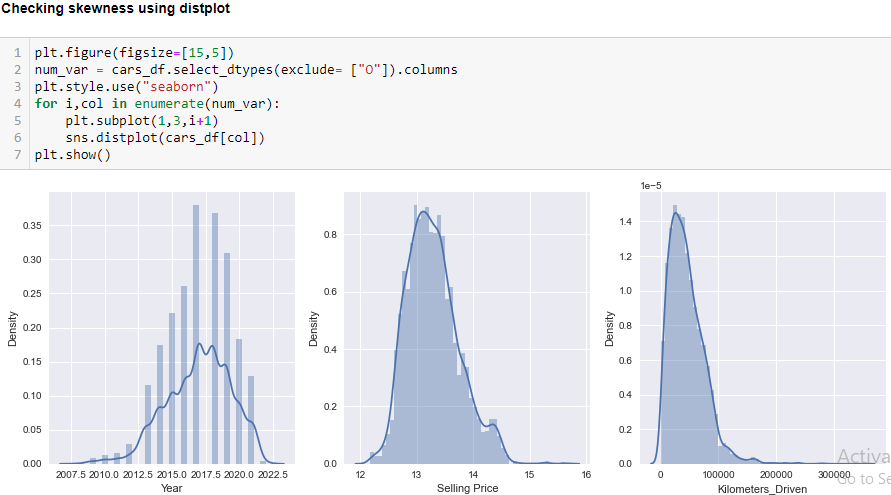




Now both the skewness and kurtosis are removed in the target variable.This looks almost normal distribution with a **Skew of 0.602862.**

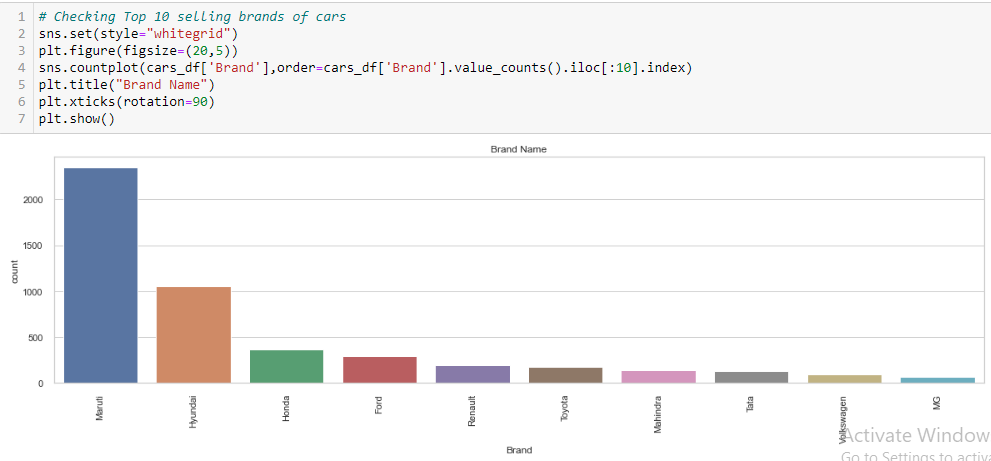
**Analysing Numerical columns:**

****

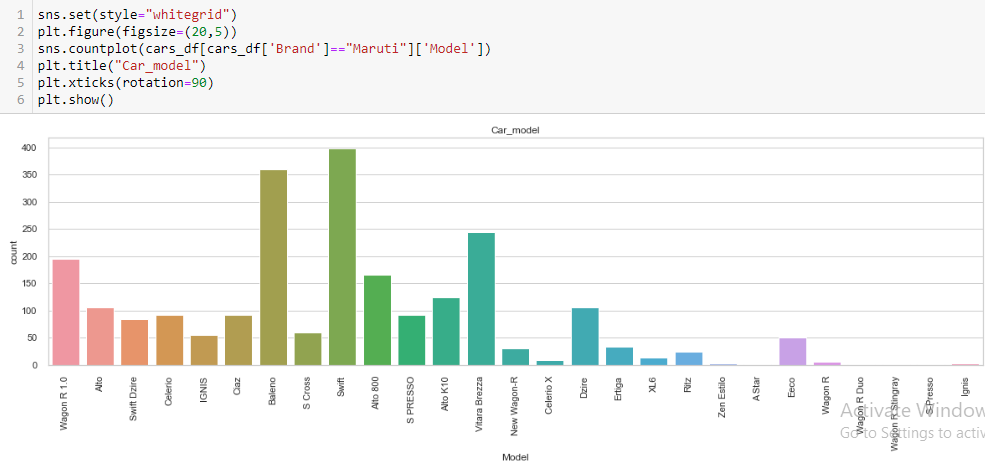
****

Here in all the numerical columns we can see through scatter plot that many numerical features are uniformly distributed.

**Analysing Categorical Columns:**

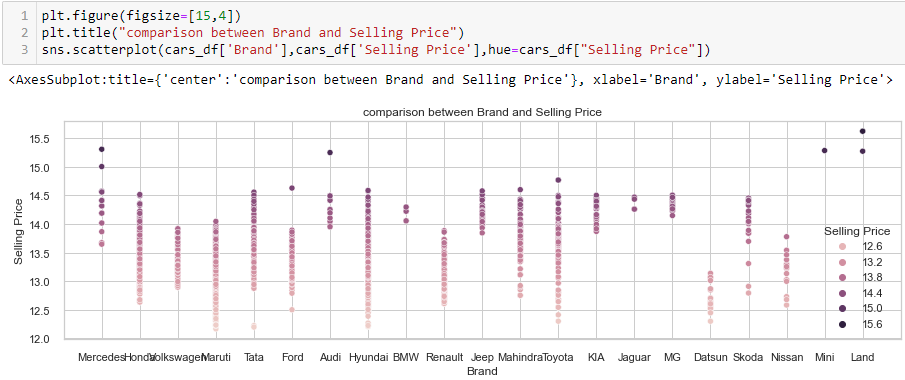


We can see from above count plot, Maruti is the highest selling brand followed by Hyundai. This may be because of their genuine price with genuine features.

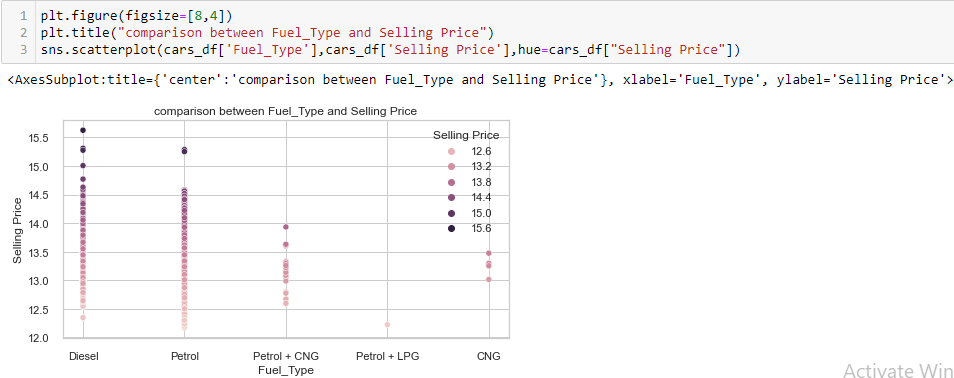
****From the above count plot we can see Maruti Swift model is highest selling model, folled by Baleno And Brezza.

**Bivariate Analysis:**

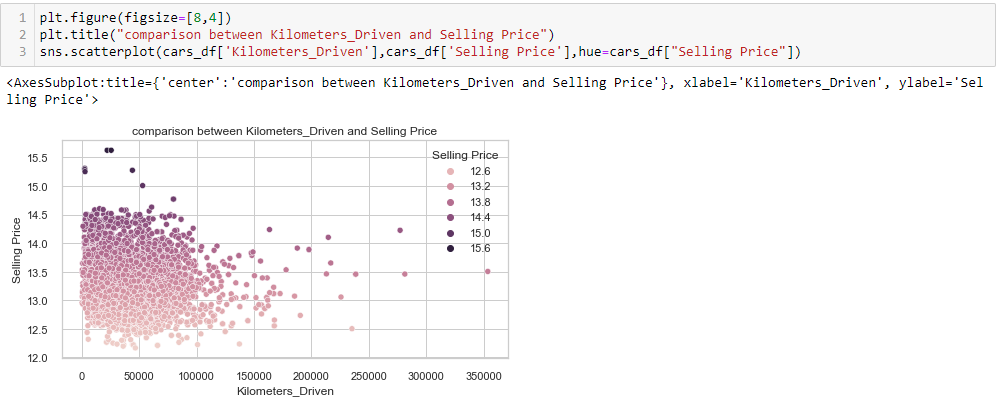
Bivariate analysis is finding some kind of empirical relationship between two variables. Specifically, the dependent vs independent Variables



Here we can Land Rover, Mercedes, Audi have highest selling Price as they are top brands .

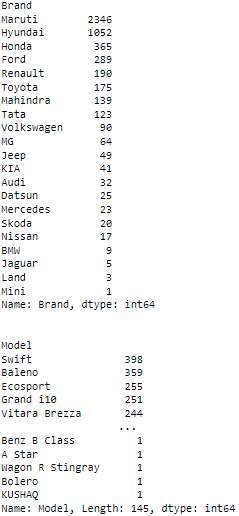
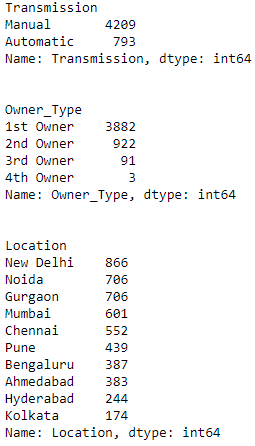
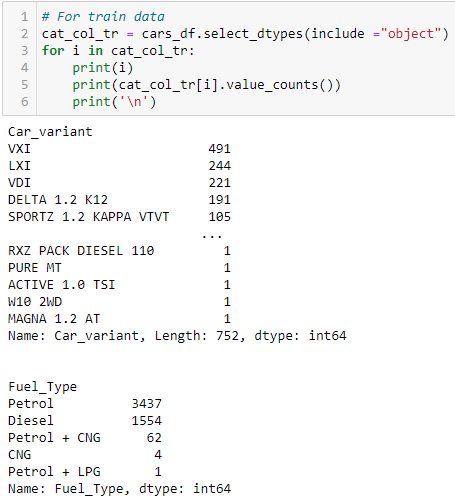


We can see from the above plot that Diesel cars are costly .They have highest selling price, followed by petrol cars. Petrol+LPG have lowest selling price.



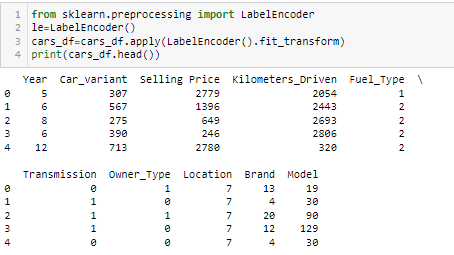
Selling Price does not depend much on Kilometer\_driven. It is mainly affected by Model of the cars and the fuel type together with Kilometer Driven.

**Categorical Columns – Value Counts**



### Converting categorical columns to numerical columns

### We will convert categorical columns into numerical columns using label encoder for further analysis.

****

**Checking for Correlation with Output Features:**

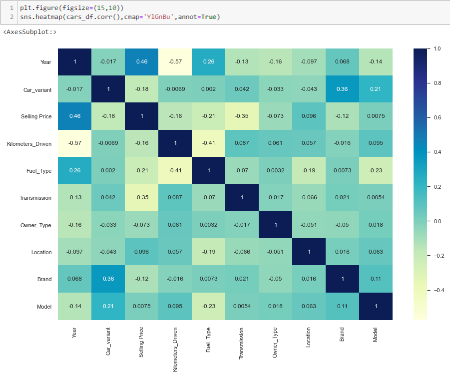
After this, we found the most important features relative to the target by building a correlation matrix. **A correlation matrix** is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. The correlation coefficient has values **between -1 to 1.**

**Correlations** are very useful in many applications, especially when conducting regression analysis.

**Multicollinearity:**

Multicollinearity is **a statistical concept where several independent variables in a model are correlated**. Two variables are considered to be perfectly collinear if their correlation coefficient is +/- 1.0. Multicollinearity among independent variables will result in less reliable statistical inferences.

Let’s check the correlation and multicollinearity through correlation heat map.



**Observation:**

1. Year of manufacturing, Location, Brand, and Model is positively correlated with the selling price. It means more latest is the model and more recently the model was manufactured, more is the price.
2. Other features are negatively correlated with the target feature.
3. There is not much multicollinearity among the independent features.

**Separating Independent and Dependent (target) features from Train Data:**



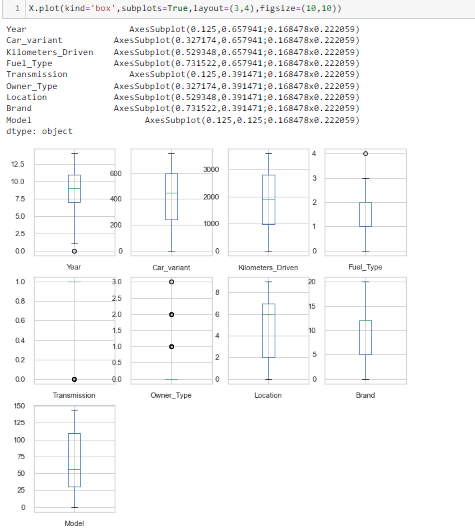
**Check for Outliers:**

An outlier is **a data point that is noticeably different from the rest**. They represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data. **A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data**.

#### Box Plot

#### This is the visual representation of the depicting groups of numerical data through their quartiles. **Boxplot** is also used for detecting the outlier in the data set.

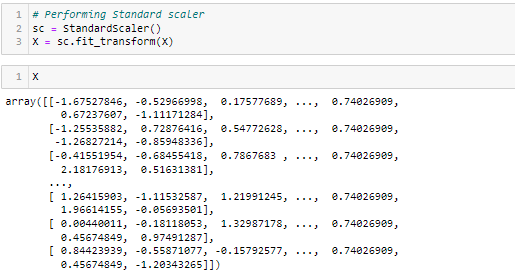
*I used a box plot in this dataset because It captures the summary of the data efficiently with a simple box and whiskers and allows me to compare easily across groups.*

****

**There are not many outliers in the columns. So we don’t need for outliers treatment.**

**Features Scaling / Standard Scaler:**

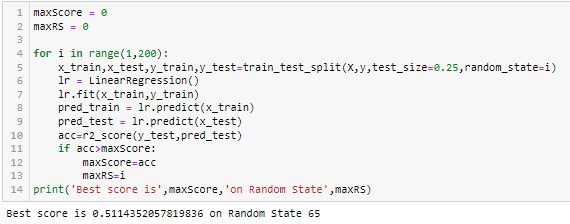
Feature Scaling is **a technique to standardize the independent features present in the data in a fixed range**. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.



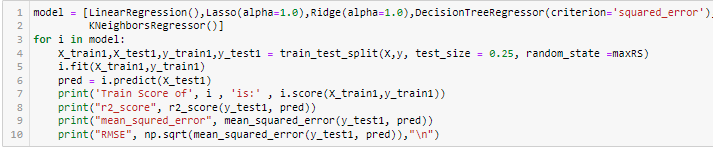
#### By using a standard scaler, I have scaled the data in one range.

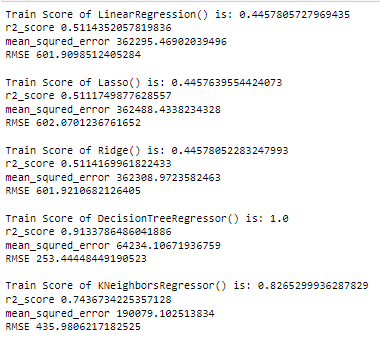
**Building Machine Learning Models:**

First, I will find the best random state on which I will get the maximum score.



**Applying train-test split with Best Random State and applying ML on Different Algorithms:**

****

****

**Conclusions:**

## Have checked Multiple Model and their score also. I have found that Decision tree regressor model is overfitting. Other models are working well. But KNeighborsRegressor is having less train and test score difference with least mean square error and least RMSE. Now i will check with the ensemble method to boost up score

## Using Ensemble Technique to boost up score:

### RandomForestRegressor:

### 

### AdaBoostRegressor:

### 

### GradientBoostingRegressor:

### 

## Conclusion:

## Here we can see the least difference between train score and test score is coming in GradientBoostingRegressor.So model is working well with both train model and test model.

## For AdaBoostRegressor, the difference is high, So the model is overfitting.

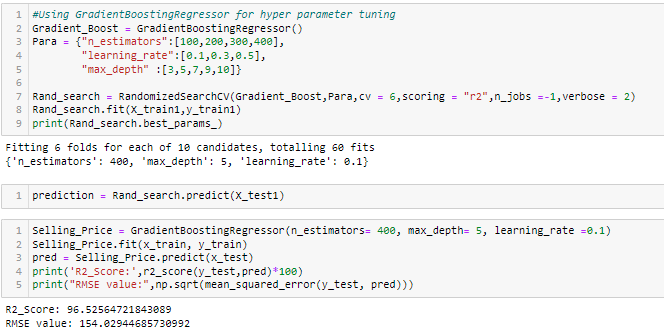
## For RandomForestRegressor, the difference is also more as compared to GradientBoostingRegressor.

## So selecting GradientBoostingRegressor as final model

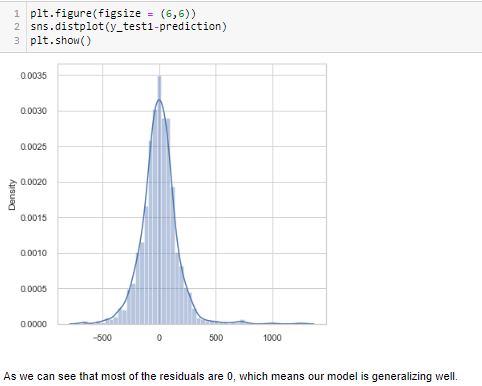
## Hyper Parameter Tuning:

[**Hyperparameter tuning**](https://towardsdatascience.com/hyperparameter-tuning-c5619e7e6624) (or hyperparameter optimization) is the process of determining the right combination of hyperparameters that maximizes the model performance. It works by running multiple trials in a single training process.

We are using Randomsearchcv method for hyperparameter tuning to find best parameters for GradientBoostingRegressor.

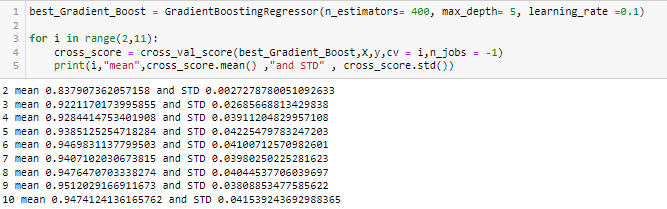


**The predicted y value is having a normalized curve which is good.**

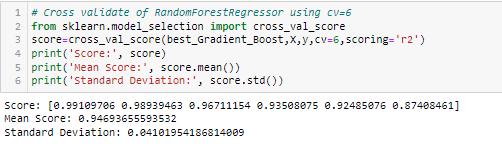
****

**Cross Validation:**

Cross-validation is **a resampling method that uses different portions of the data to test and train a model on different iterations**. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.



**Applying Cross validation Score=6**



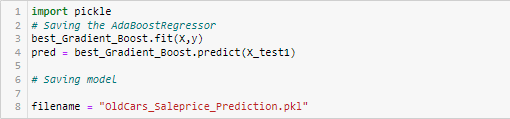
**Plotting y\_test1 vs predictions:**

* Simply plotting our predictions vs the true values.
* Ideally, it should be a straight line.

## 

**Saving the Model:**

We are saving the model by using python’s pickle library. It will be used further for the prediction.



Here we have predicted the target of test dataset and checked the shape of test dataset and target of test dataset to join them.

**CONCLUSION:**

* After Scraping old car prices for cities like Delhi, Noida, Gurgaon, Mumbai, Pune, Hyderabad, Bangalore, Ahmedabad, Chennai, Kolkata from different websites like Cars24 and CArdekho.com I have prepared an excel sheet and loaded the dataset for further EDA process.
* So, as we saw that we have done a complete EDA process, getting data insights, feature engineering, and data visualization as well so after all these steps one can go for the prediction using machine learning model-making steps.
* We have both numerical and categorical data types features in the datasets and the dependent variable of train data i.e. the **Selling price** is the numerical data type. So, I applied the regression method for prediction.
* Once data has been cleaned and missing value is replaced, Label encoding is applied to them to convert them into Numerical ones. I trained the model on five different algorithms but for most of the models, train and test data was having a variance, and the model was overfitting.
* Only Gradient Boost regressor worked well out of all the models, as there was less difference between train score and test score and RMSE was also low hence I used it as the final model and have done further processing.
* After applying hyperparameter tuning I got an accuracy(r2\_score) of 96% from the GradientBoostRegressor model after hyper parameter tuning which is a good score.

Then I saved the model.

**Limitations and Scope:**

* This study used different models in order to predict used car prices. However, there was a relatively small dataset for making a strong inference because number of observations was only 5058. Gathering more data can yield more robust predictions.
* Secondly, there could be more features that can be good predictors. For example, here are some variables that might improve the model: number of doors, gas/mile (per gallon), color, mechanical and cosmetic reconditioning time, used-to-new ratio, appraisal-to-trade ratio.
* Another point that has room to improve is that the data cleaning process can be done more rigorously with the help of more technical information. For example, instead of using ‘fill’ method, there might be indicators that helps to fill missing values more meaningfully.

As a suggestion for further studies, while pre-processing data, instead of using a label encoder, one hot encoder method can be used. Thus, all non-numeric features can be converted to nominal data instead of ordinal data (Raschka & Mirjalili, 2017).

**I hope this article helped you to understand Data Analysis, Data Preparation, and Model building approaches in a much simpler way.**

**Thank you**