

# CAR PRICE PREDICTION

**Presented by:**

**PRATAP LAVATE**

# Problem Statement

* 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. This project contains two phase:
* 1.Data Collection Phase
* 2.Model Building Phase

**1.Data Collection Phase:**

You have to scrape at least 5000 used cars data. You can scrape more data as well, it’s up to you. more the data better the model.

In this section You need to scrape the data of used cars from websites (Olx, cardekho, Cars24 etc.) You need web scraping for this. You have to fetch data for different locations. The number of columns for data doesn’t have limit, it’s up to you and your creativity. Generally, these columns are Brand, model, variant, manufacturing year, driven kilometers, fuel, number of owners, location and at last target variable Price of the car. This data is to give you a hint about important variables in used car model. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data.

Try to include all types of cars in your data for example- SUV, Sedans, Coupe, minivan, Hatchback.

**2.Model Building Phase:**

After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

Follow the complete life cycle of data science. Include all the steps like. 1. Data Cleaning

1. Exploratory Data Analysis
2. Data Pre-processing
3. Model Building
4. Model Evaluation
5. Selecting the best model.

# EDA(Exploratory Data Analysis)

## Data Description

The dataset contains 18865 records (rows) and 9 features (columns). Here, we will provide a brief description of dataset features. Since the number of features is 9, we will attach the data description i.e., 'Model', 'Brand', 'Variant', 'Manufacturing\_year', 'Driven\_km’, 'Fuel\_type', 'Transmission', 'Selling\_Price', 'location'.

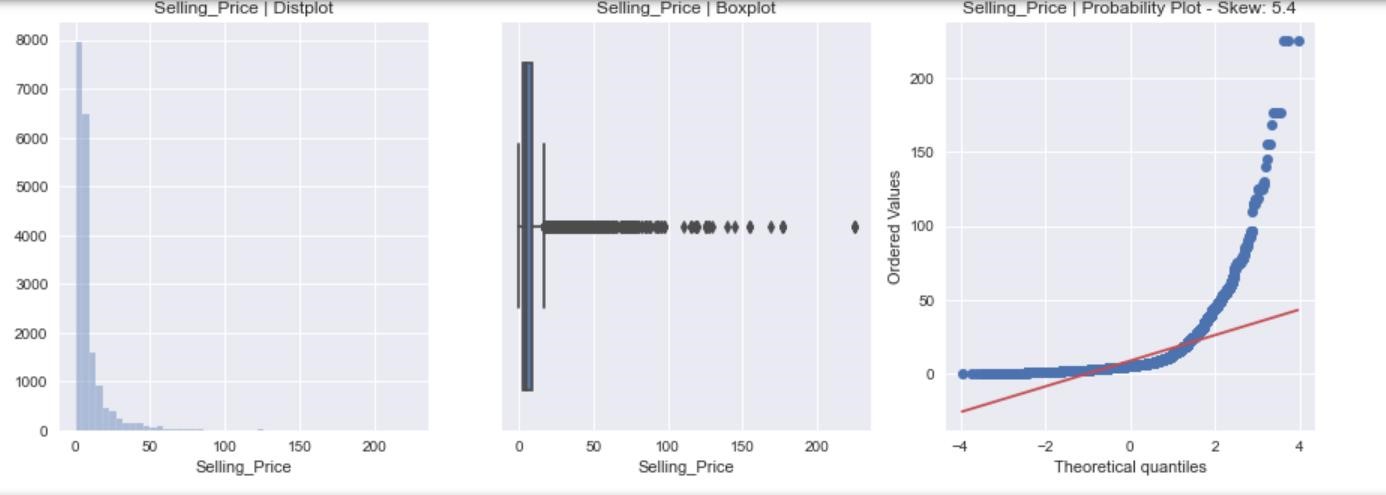
## Target Variable

• Price(Selling Price): It’s continuous type of data, so the model approach is carried out for Regression analysis.

**Regression:**

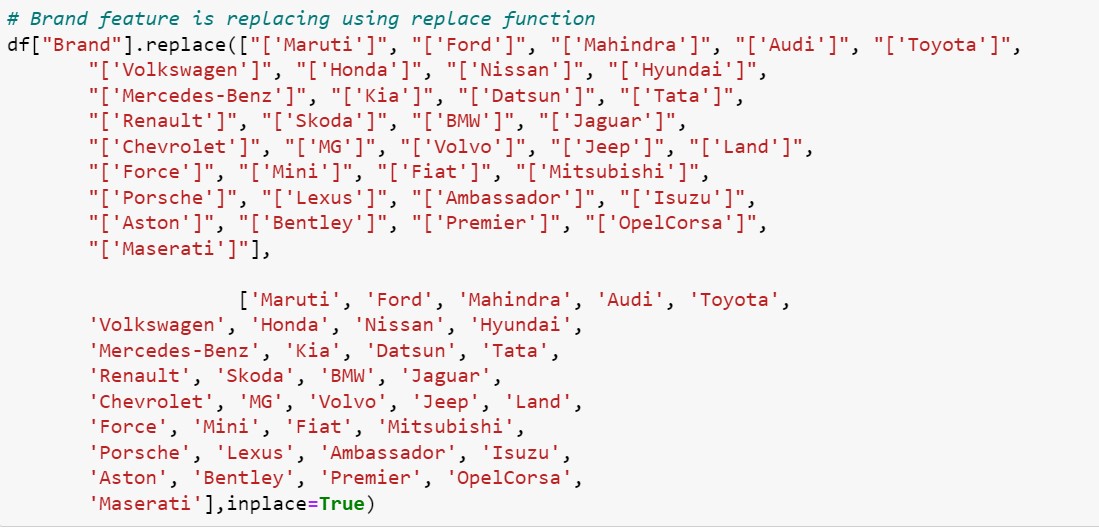
It’s an analysis is used when you want to predict a continuous dependent variable from a number of independent variables.

Independent variables with more than two levels can also be used in regression analysis.



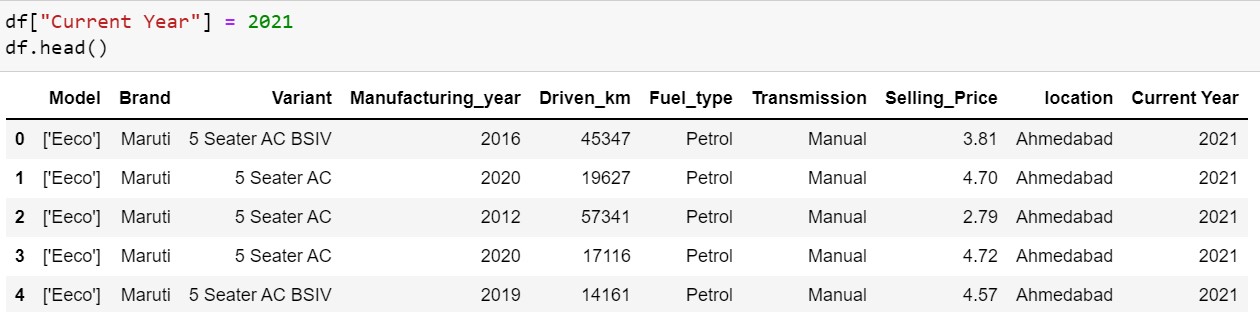
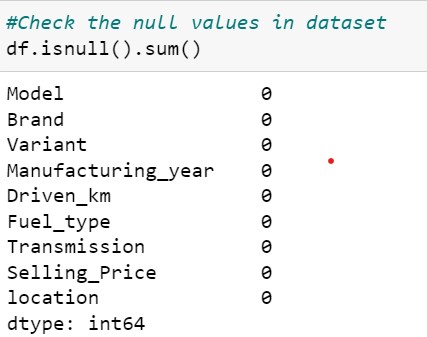
## Data Pre-processing

**Unique function for dataset**



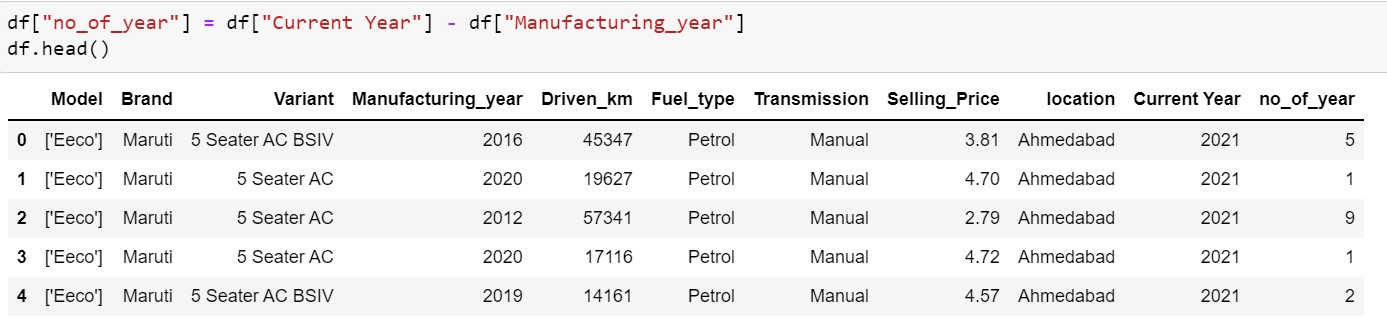


**No Null values in Dataset Adding Features in Datasets**

  Adding Current year in data frame

Created number of year by subtracting current year and

**Dropped Features** manufacturing year

1. Manufacturing Year
2. Current Year
3. Model
4. Variants

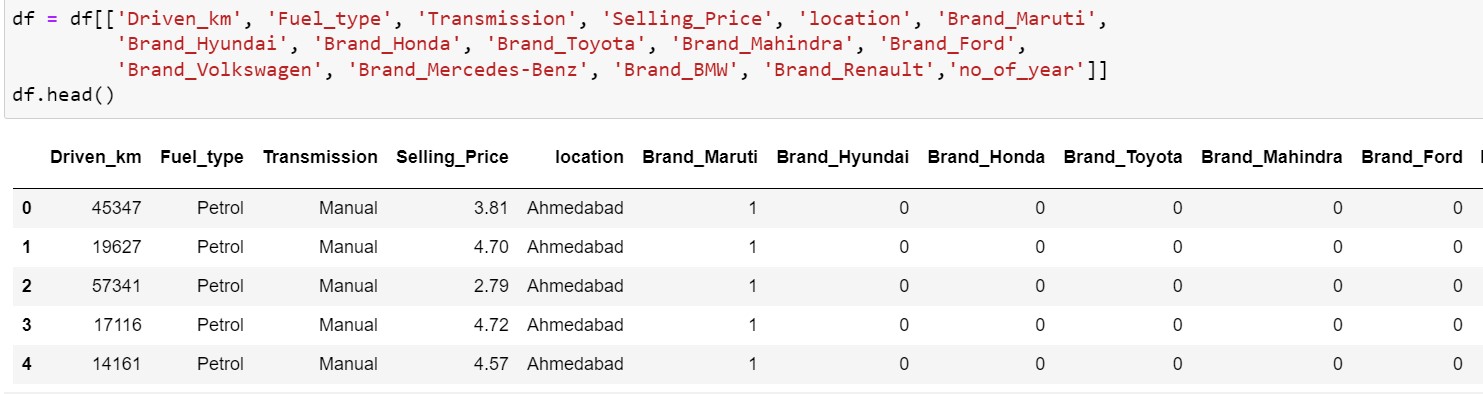
# Data Cleaning

**Encoding of Data Frame:**

The Encoding Technique is used for this problem:

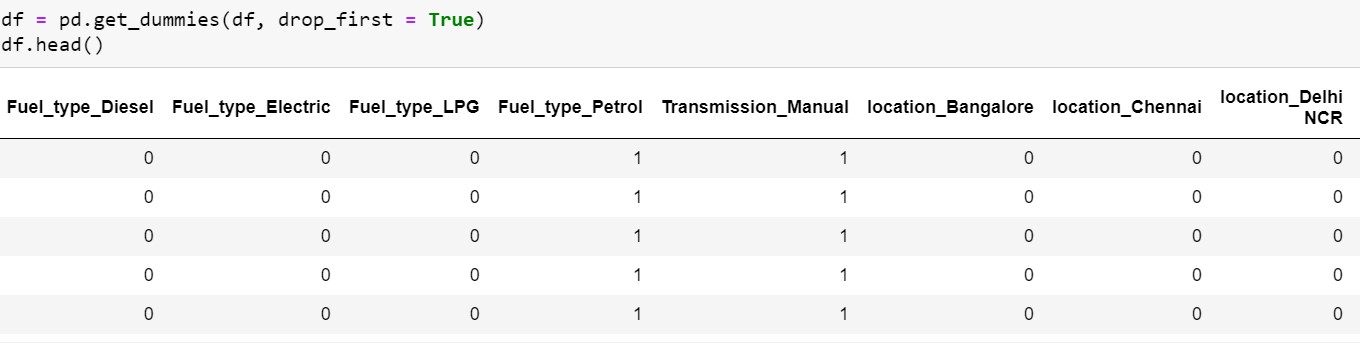
1. One hot encoding technique with multiple variables.
2. One hot encoding technique.

Firstly, proceed with One hot encoding technique with multiple variables for particular features i.e., Brand



The new data frame is created using one hot encoding technique with multiple variables.

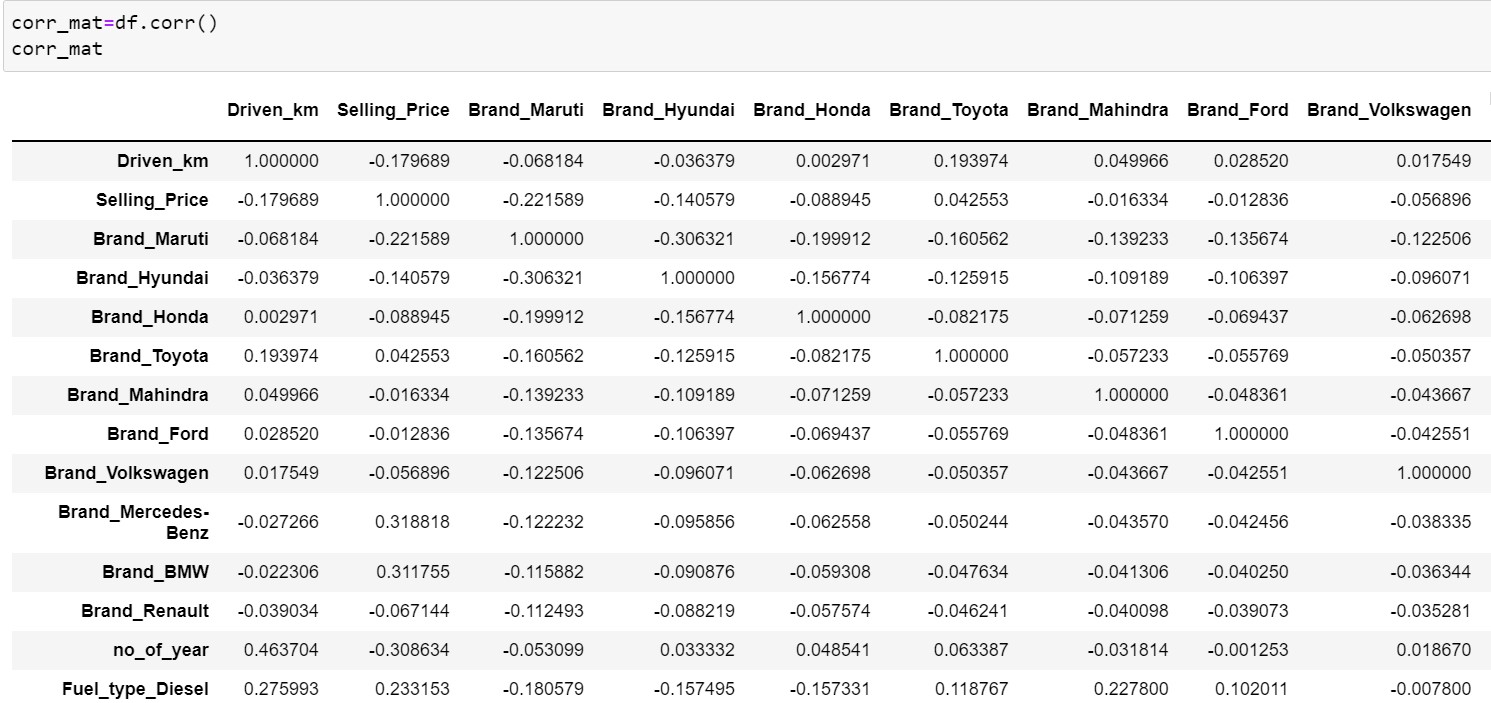
Secondly, proceed with One hot encoding technique i.e., transmission, location and fuel types.



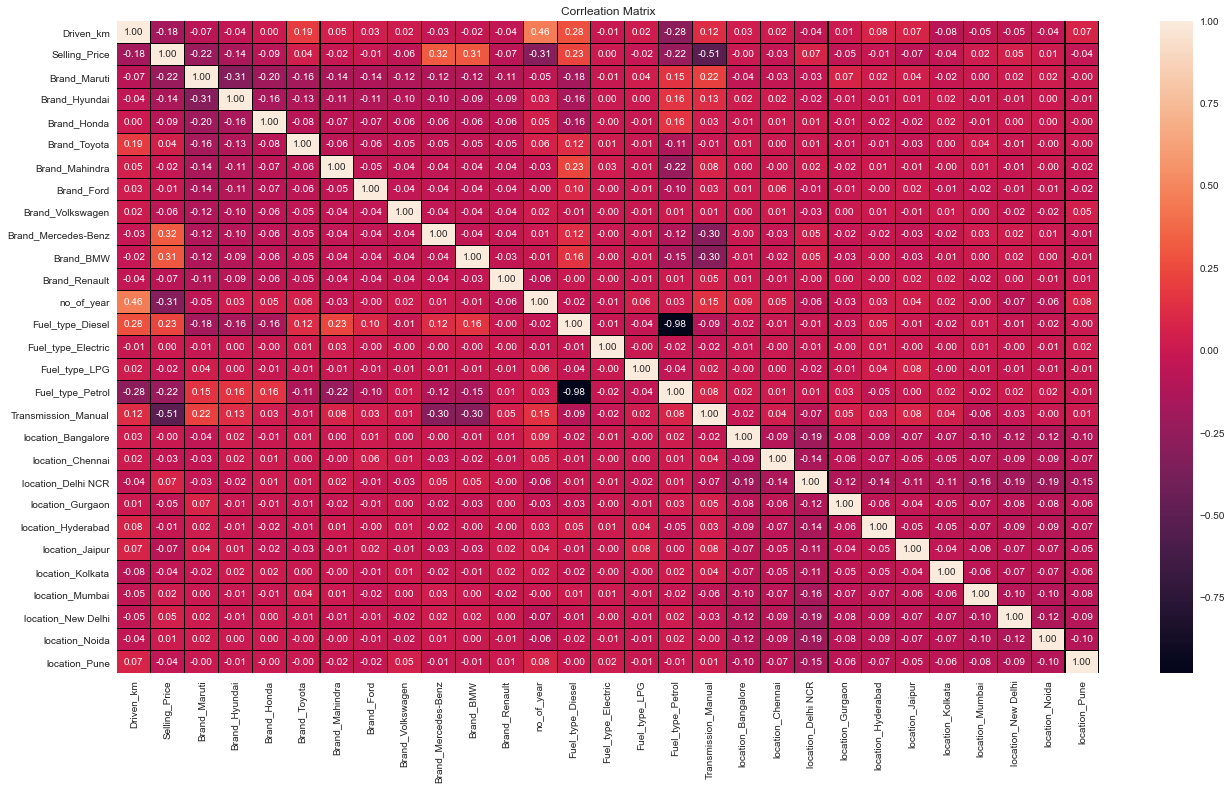
Now, let’s we can see all features is converted into numerical one after proceeding with encoding technique.

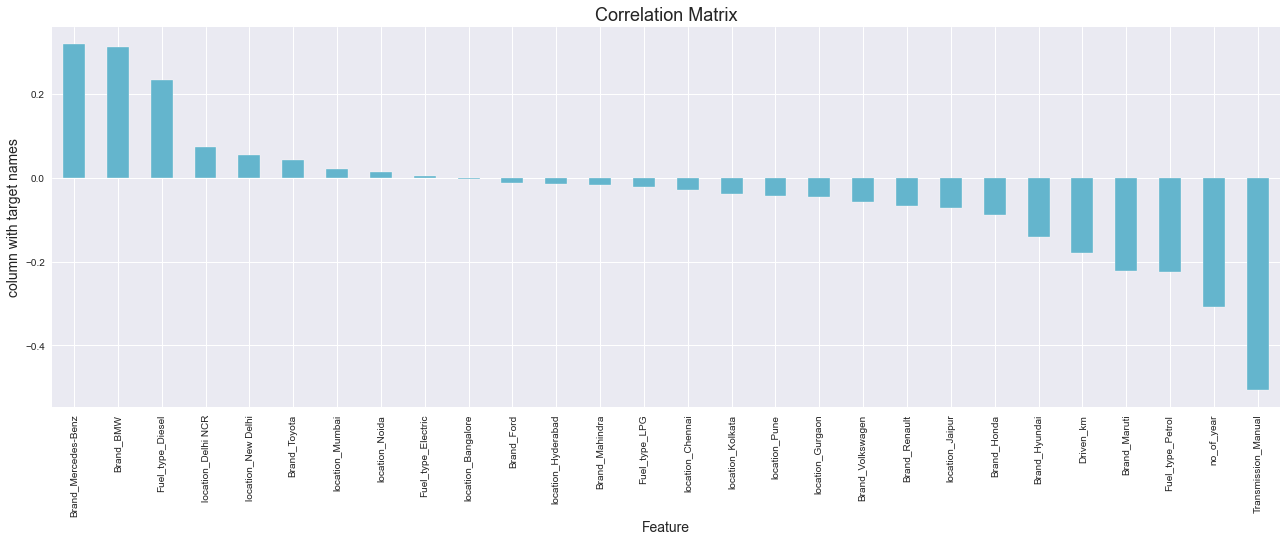
## Statistical Summary

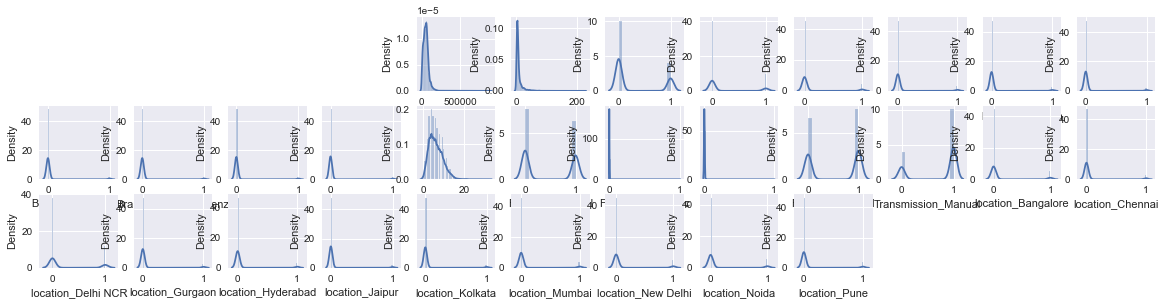
To see statistical information about the non-numerical columns in our dataset:

**Correlation matrix:**

A correlation matrix is simply a table which displays the correlation. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a heatmap.

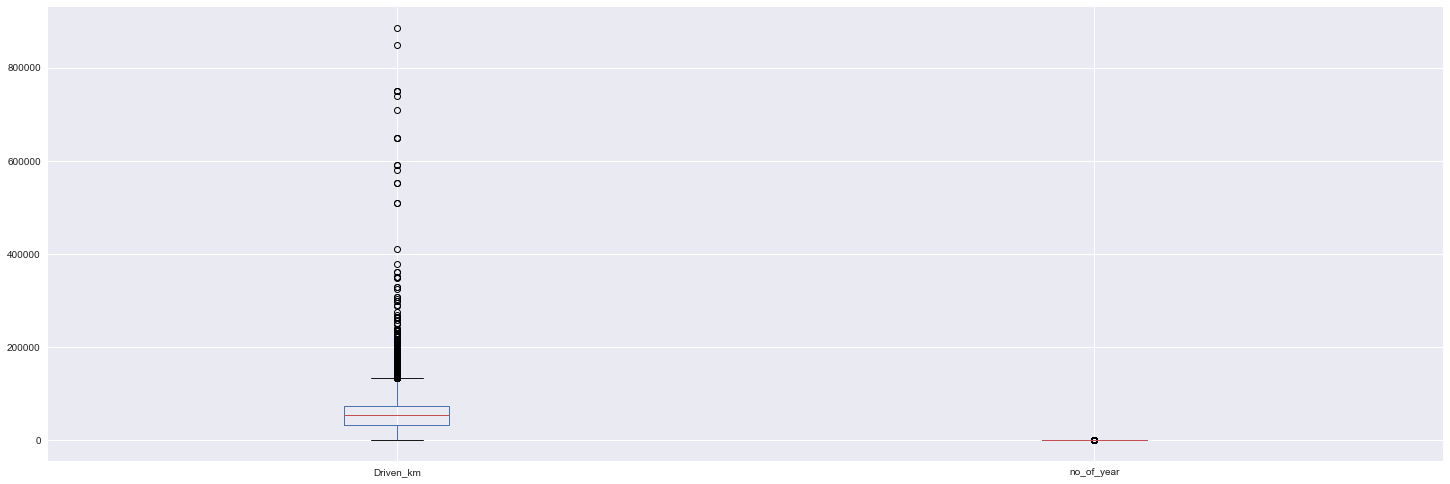






**Outliers Check:**

In this dataset, we applied one hot encoding method to categorical features. so, we check outlies for nominal features i.e., Diven\_Km, no\_of\_years and Selling Price. Only Driven\_km and no\_of\_years is considered because Selling Price is our target variable.



We can see outliers in Driven km due to various kilometers driven for different cars. so, we proceed with further steps.

**Checking Skewness:**

Now here, we are going to use Power transform function to handle skewness in dataset



## Model Building and Evaluation

These are modelling approach made to build an model :

* Linear
* k-nearest neighbors (KNN)
* Random Forest
* Decision Tree
* XGBoost

## Performance Metric

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Building** | **R2 score** | **MAE** | **MSE** | **RMSE** |
| **Linear** | 53.63 | 4.21 | 56.70 | 7.53 |
| **KNeighbors** | 65.71 | 2.67 | 41.92 | 6.47 |
| **Random** | 77.45 | 1.97 | 27.56 | 5.25 |
| **Decision** | 61.21 | 1.98 | 47.38 | 6.88 |
| **XGBoost** | 72.02 | 2.35 | 34.22 | 5.85 |

According to performance metric, the random forest has higher R2 score, So this is our best model.

**Comparison:**

|  |  |
| --- | --- |
| **Performance Metric** | **Cross -Validation Score** |
| **53.63** | -4.16 |
| **65.71** | 53.89 |
| **77.45** | 69.19 |
| **61.21** | 54.18 |
| **72.02** | 65.62 |

Comparing the performance model and cross-validation score the minimum difference is for xgboost. so finally, this is our best model.

# Hyper Parameter Tuning

**The Hyper parameter tuning is carried out for XGBoost Regressor model.**

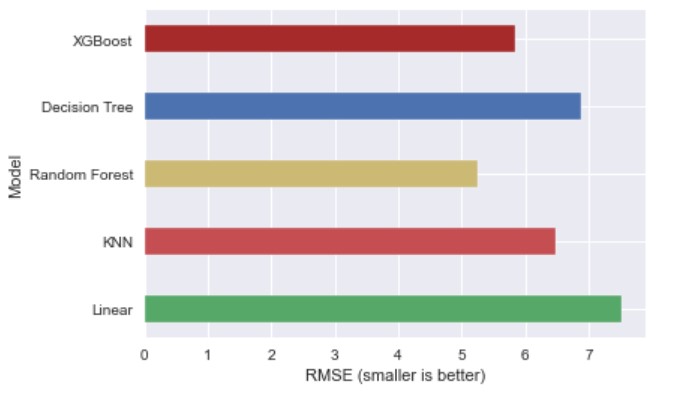
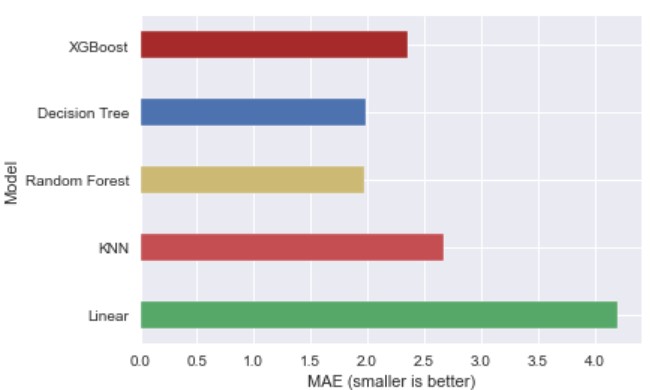
**Best Model**

Hyper parameter Tuning performance is carried out for XGBoost Regressor:

Hyper parameter Tuning i.e.,R2 score and Cross validation score = 78.18% and 69.29% respectively. Finally, XGBoost is best model for these dataset.

**Performance Interpretation:**

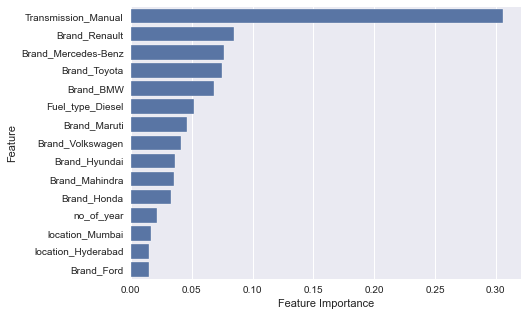
**MAE (Mean Absolute Error) RMSE (Root Mean Squared Error)**



**Feature Importance’s:**

Some of the models we used provide the ability to see the importance of each feature in the dataset after fitting the model. We will look at the

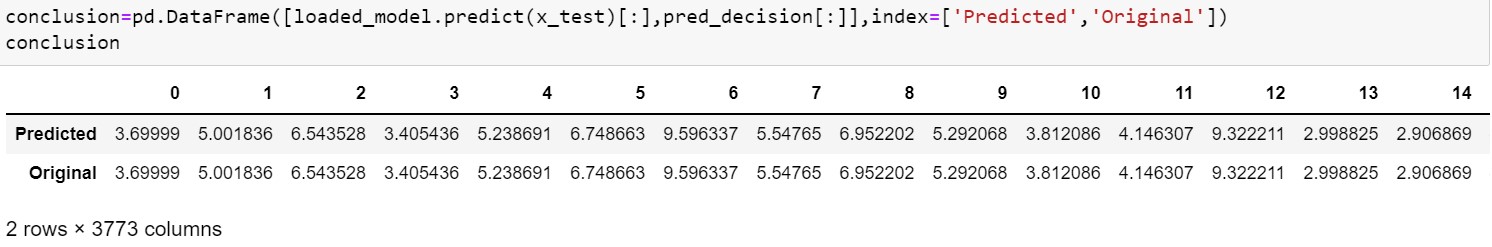
feature importance’s provided by XGBoost models. We have 29 features in our data which is a big number, so we will take a look at the 15 most important features.



Notice here in feature importance of XGBoost, the transmission manual feature plays a prominent role for target variable.

**Conclusion:**

* In this paper, we built several regression models to predict the selling price of cars by given some of the cars features. We evaluated and compared each model to determine the one with highest performance. We also looked at how some models rank the features according to their importance. In this paper, we followed the data science process starting with getting the data, then cleaning and pre-processing the data, followed by exploring the data and building models, then evaluating the results.
* As a recommendation, we advise to use this model (or a version of it trained with more recent data) by car market who want to get an idea about car price. The model can be used also with datasets that covered areas provided that they contain the same features. We also suggest that people take into consideration the features that were deemed as most important as seen in the previous section; this might help them estimate the car price is better.



## Thank you