

CAR PRICE PREDICTION PROJECT

Submitted by:

B. JERISH

# ACKNOWLEDGMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to Flip Robo Technologies Bangalore for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I want to thank my SME Miss. Khushboo Garg for providing the project and helping us to solve the problem and addressing out our Query in right time.

I would like to express my gratitude towards my parents & members of Flip Robo for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to industry persons for giving me such attention and time.

# INTRODUCTION

## BUSINESS PROBLEM FRAMING

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.

## REVIEW OF LITERATURE

This project is more about exploration, feature engineering and classification 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.

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 increasingly digital world.

# ANALYTICAL PROBLEM FRAMING

## MATHEMATICAL/ANALYTICAL MODELLING OF THE PROBLEM

In our scrapped dataset, our target variable "**selling\_price**" is a continuous variable. Therefore, we will be handling this modelling problem as classification.

This project is done in two parts:

 Data Collection phase

 Model building phase

**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 kilometres, 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. Note – The data which you are collecting is important to us. Kindly don’t share it on any public platforms.

**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:

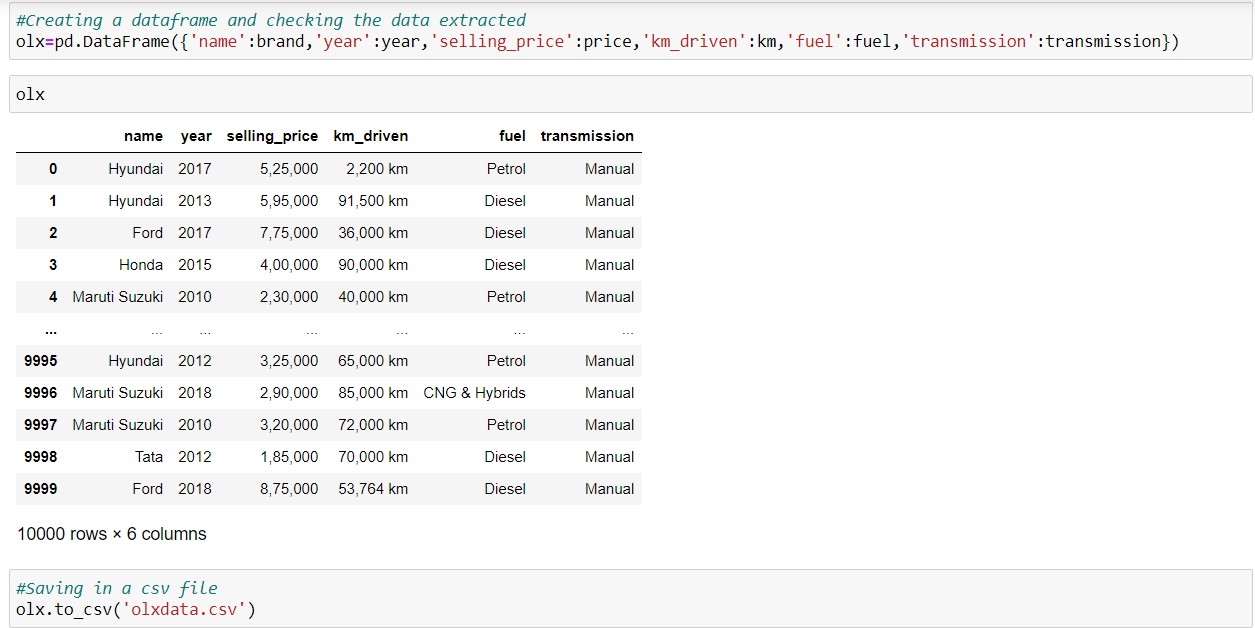
1. Data Cleaning
2. Exploratory Data Analysis
3. Data Pre-processing
4. Model Building
5. Model Evaluation
6. Selecting the best model

## DATA SOURCES AND THEIR FORMATS

We collected the data from difference websites like olx and cars24. The data is scrapped using Web scraping technique and the framework used is Selenium.

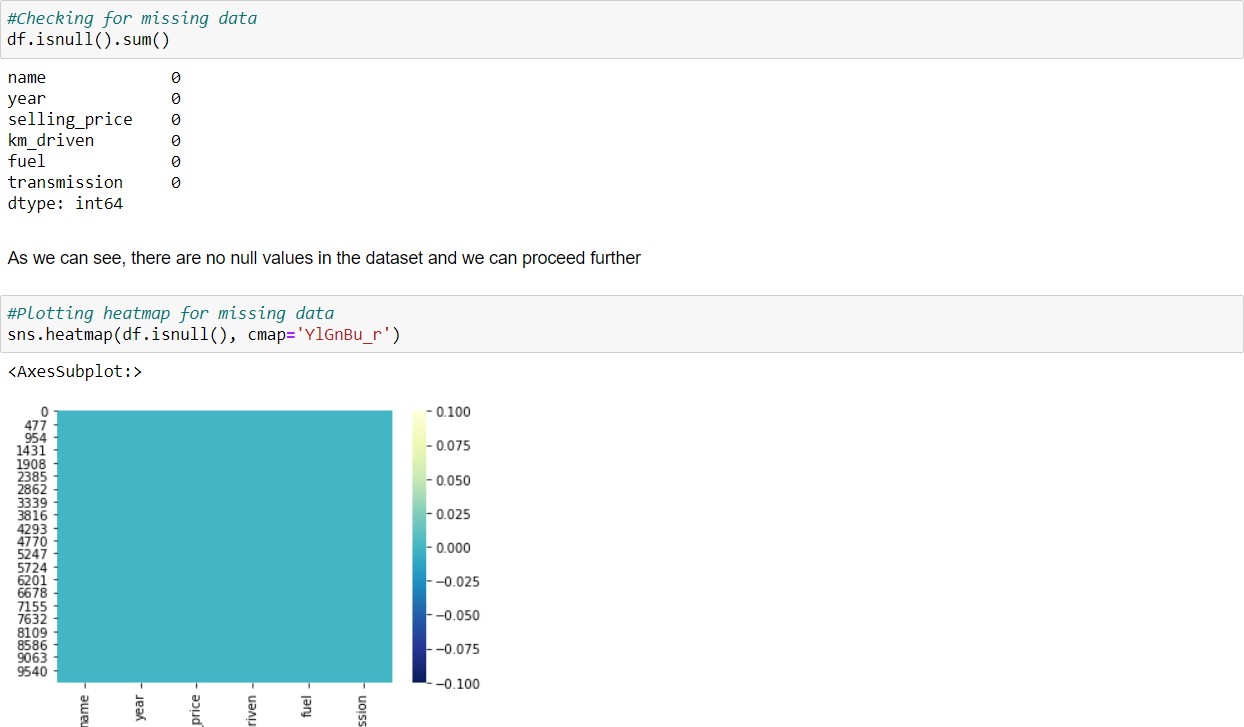
We scrapped nearly 10000 of the data and saved each data in a separate data frame

In the end, we combined all the data frames into a single data frame and it looks like as follows:



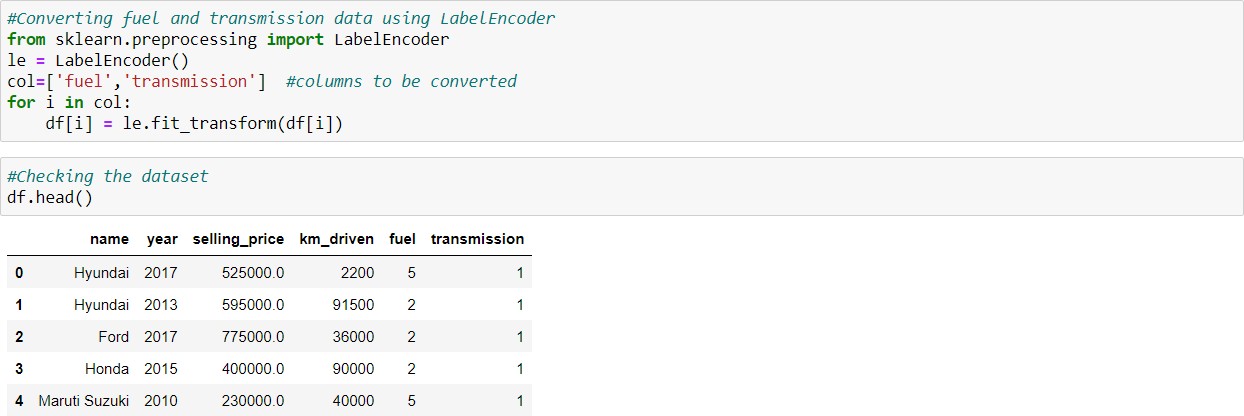
## DATA PRE-PROCESSING

**Checking for null-values**

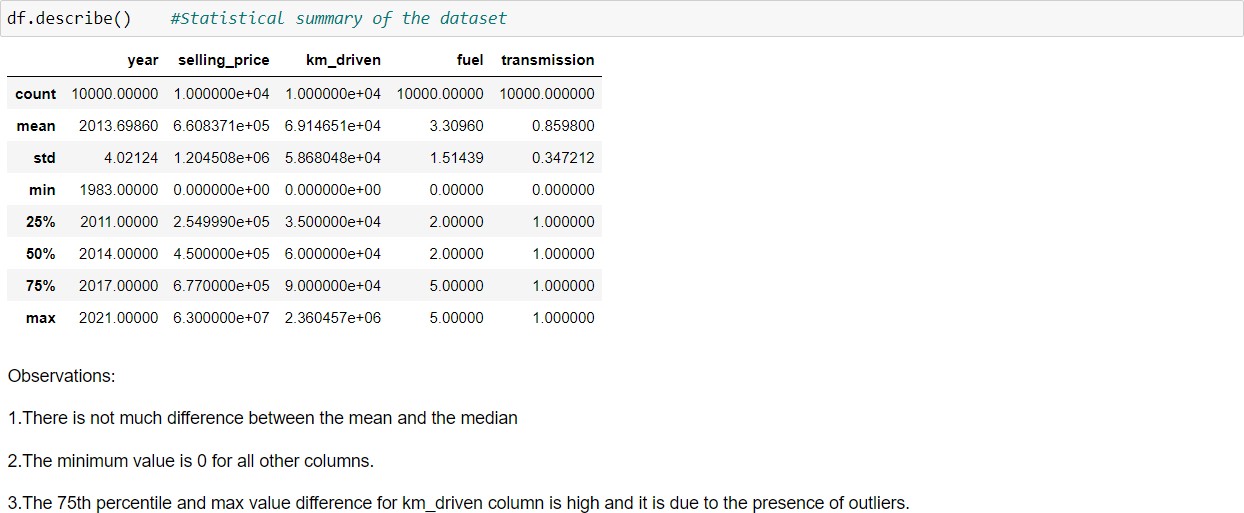


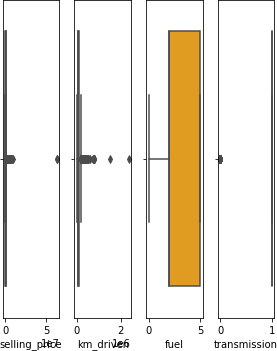
**Data Cleaning**

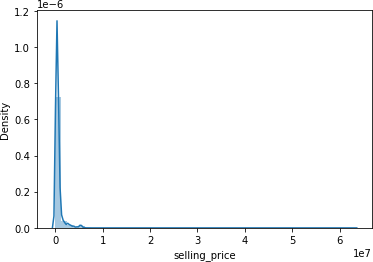
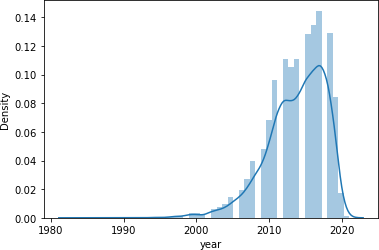
* Removing words from object data and converting them into int data type using lambda operation.
* Removing ‘,’, replacing missing data with highest weightage data.
* Removing unnecessary features

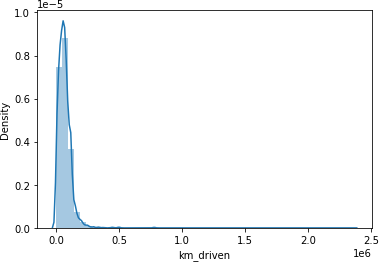
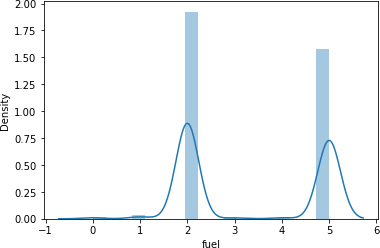
**Converting categorical data into numeric data**

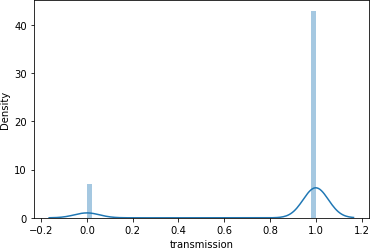
**Statistical summary of the dataset**



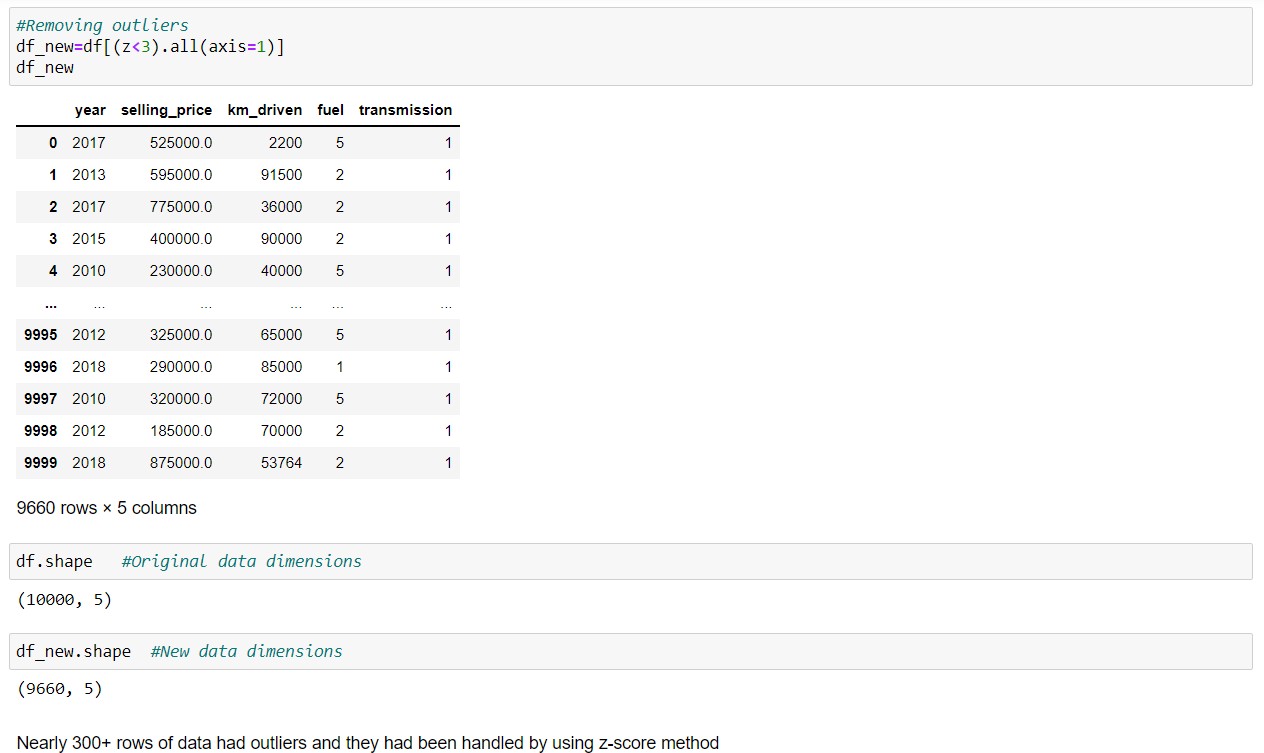
**Checking outliers**

**Checking skewness and data distribution**



**Handling outliers using z-score method**



# HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED

For doing this project, the hardware used is a laptop with high end specification and a stable internet connection. While coming to software part, I had used anaconda navigator and in that I have used **Jupyter notebook** to do my python programming and analysis.

For using an CSV file, Microsoft excel is needed. In Jupyter notebook, I had used lots of python libraries to carry out this project and I have mentioned below with proper justification:

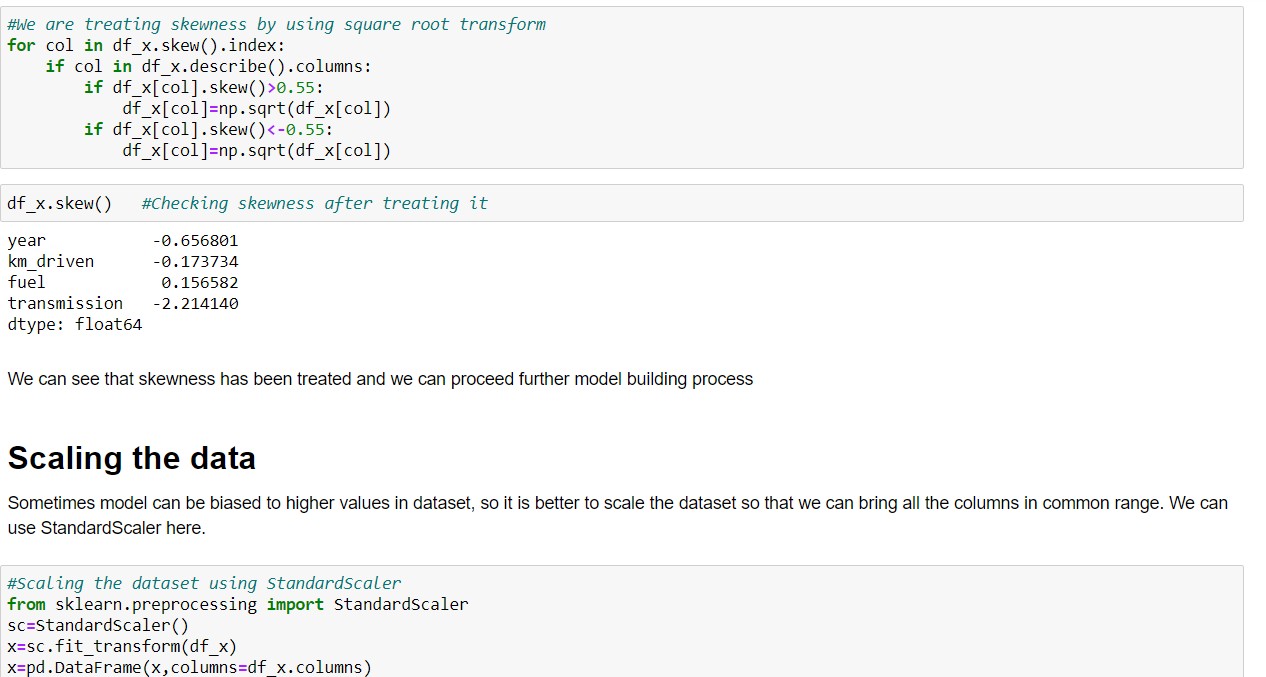
Libraries Used:

## MODEL/S DEVELOPMENT AND EVALUATION

**Preparing dataset for model training**

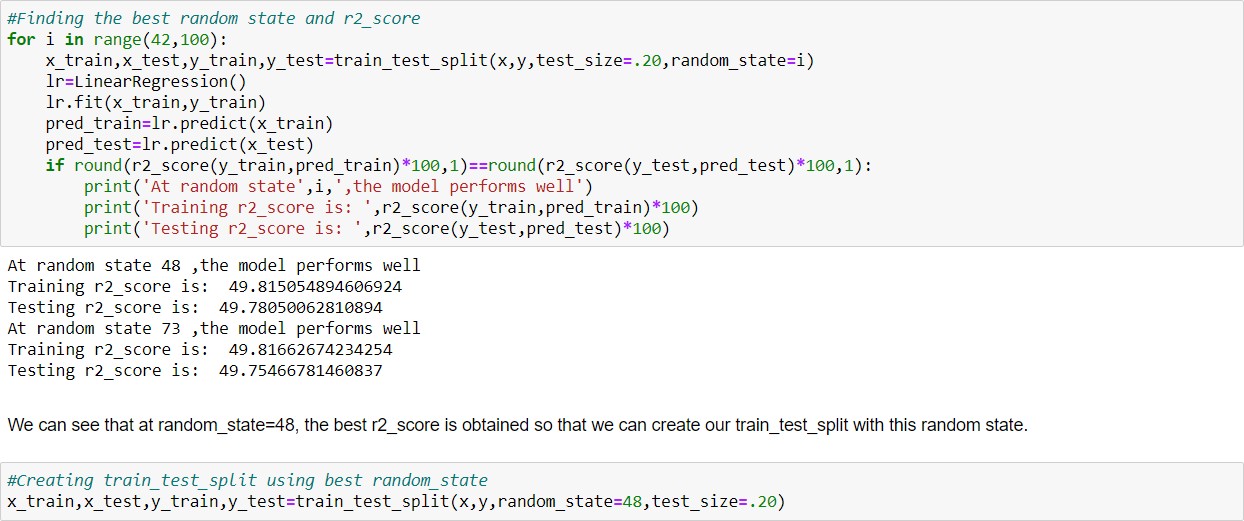


## Treating skewness and scaling the data

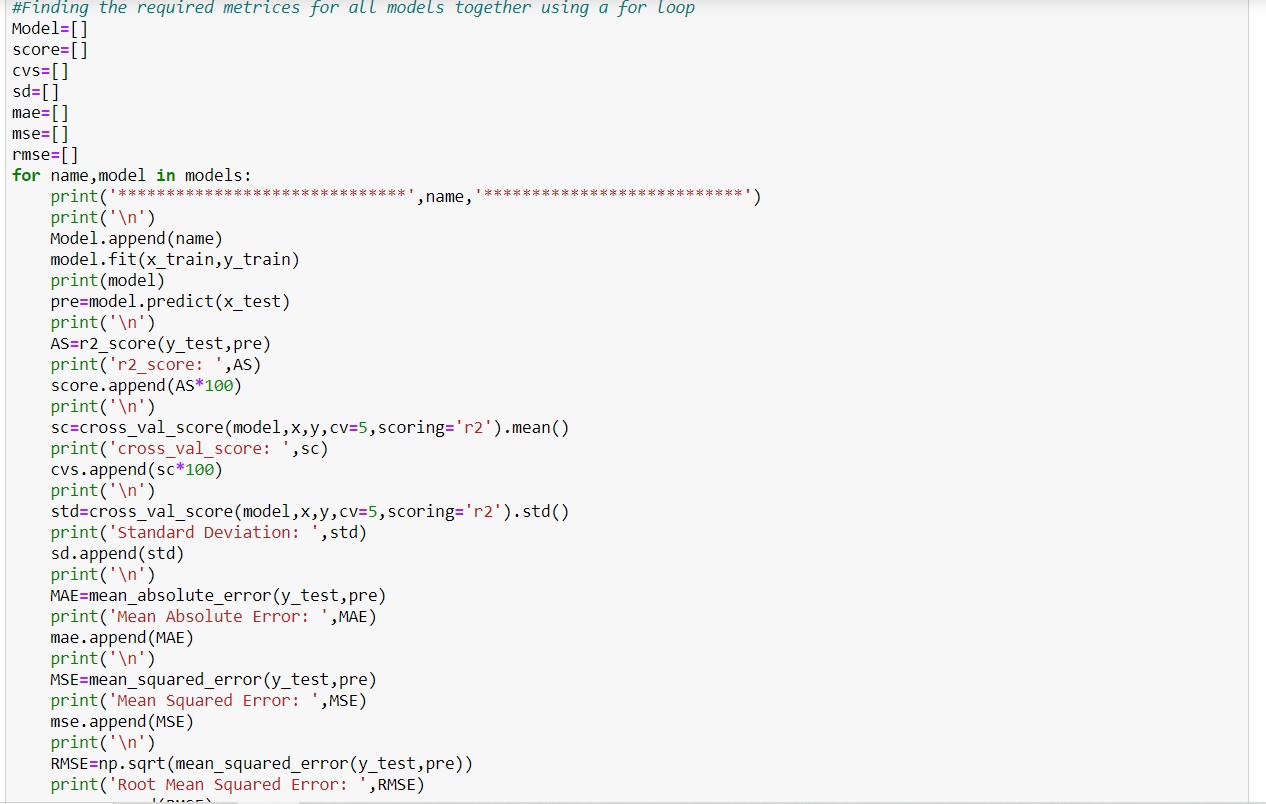


**Model Building**

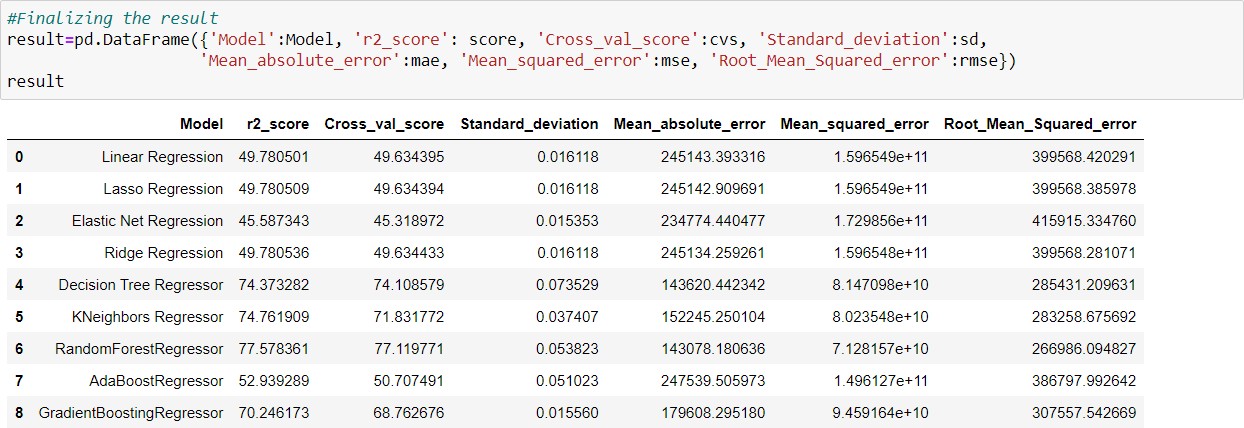
Finding best random state and r2\_score



After initializing the instances of the model, we will run a for loop with the required metrics and other regression related scores and we will append the obtained values to empty lists. The code will be as follows:



The final outputs of the models we used are stored in a dataframe and it is shown below:



We can see that Random Forest Regressor and KNeighbors Regressor are performing well compared to other algorithms. Now we will try Hyperparameter Tuning to find out the best parameters and try to increase the scores.

## Hyperparameter Tuning

**Random Forest Regressor**



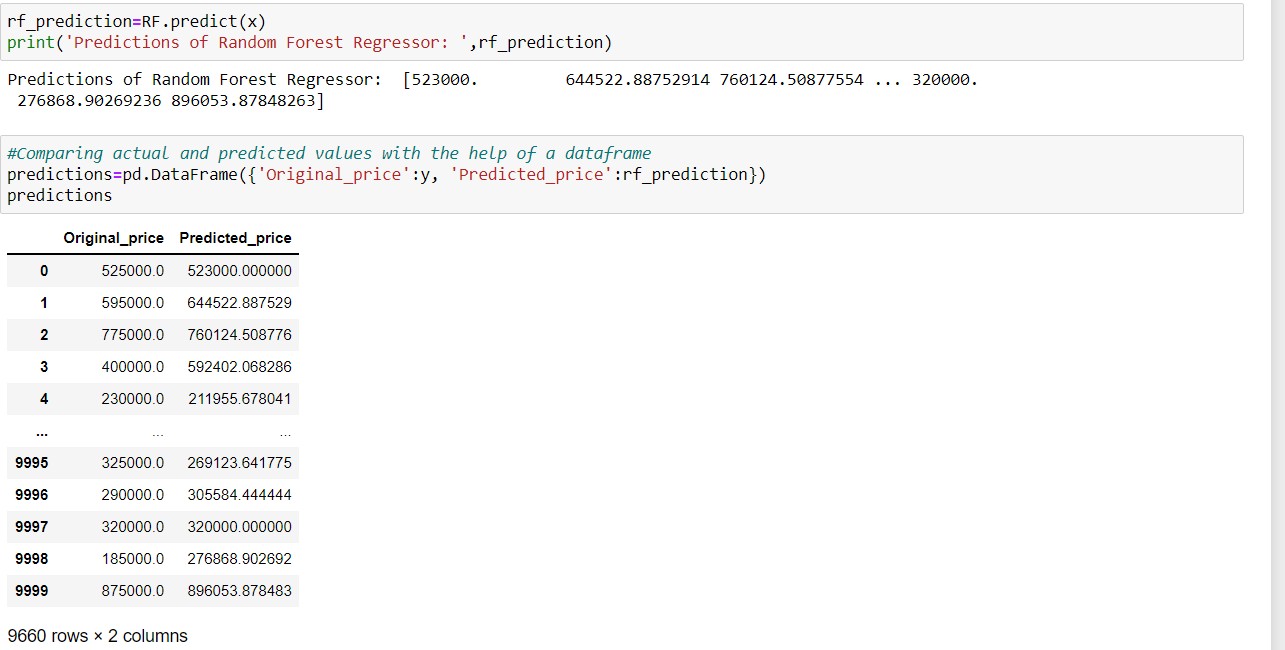
**KNeighbors Regressor**



After applying Hyperparameter Tuning, we can see that RandomForestRegressor is the best performing algorithm among all other algorithms as it is giving a r2\_score of 78.02 and cross validation score of 77.01. It has also the less amount of error values obtained. Lesser the RMSE score, the better the model. Now we will finalize the model. The scores are improved; r2\_score from 77.57 to 78.02 and cross\_val\_score from 77.11 to 77.01, which is good

# FINAL THE MODEL

**Comparing original and predicted price**

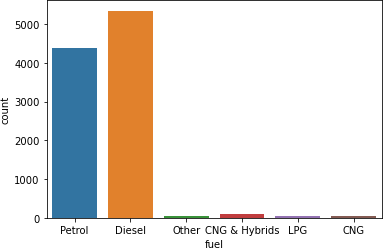


**Saving the model**

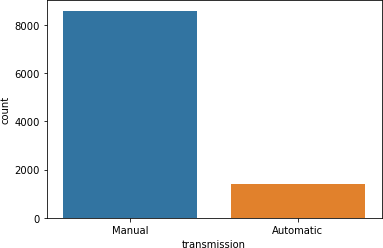


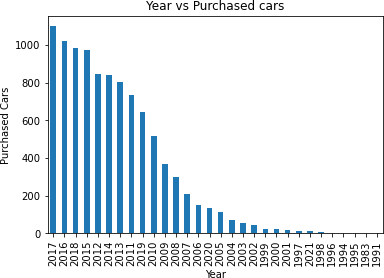
**Fuel count**

# DATA VISUALIZATION



**Transmission count**

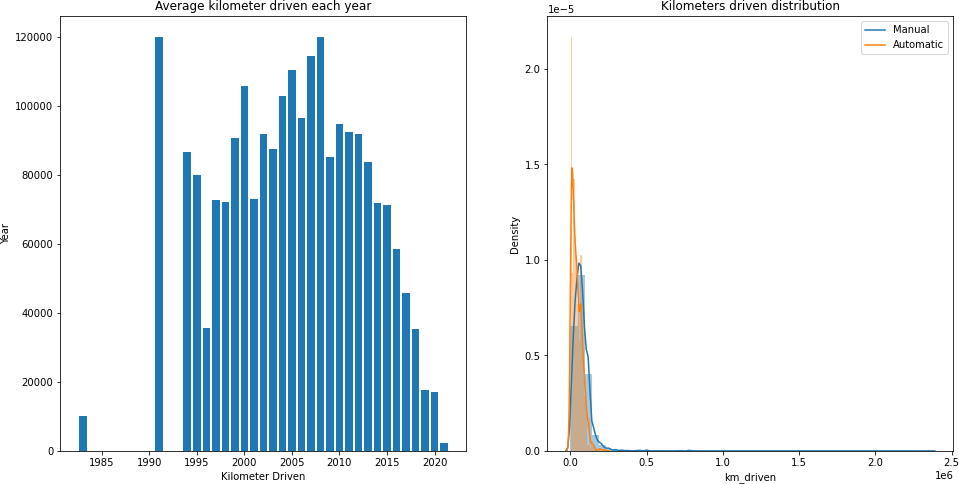


**Year vs Purchased cars**

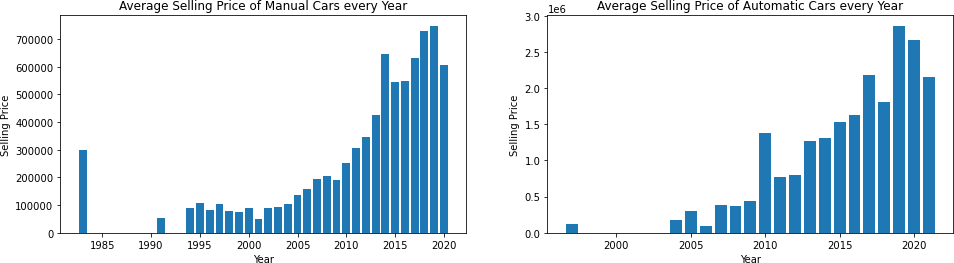
**Pairplot for transmission**



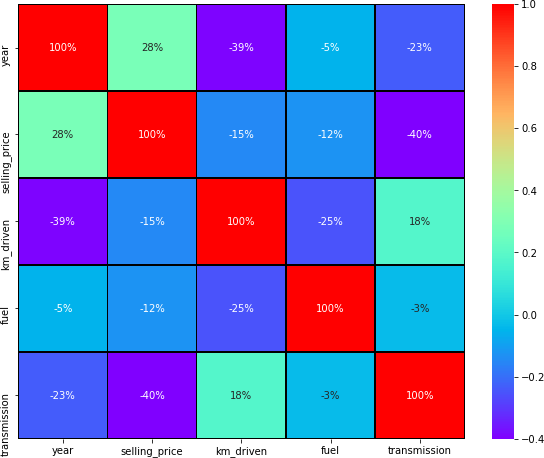
**Year vs km**



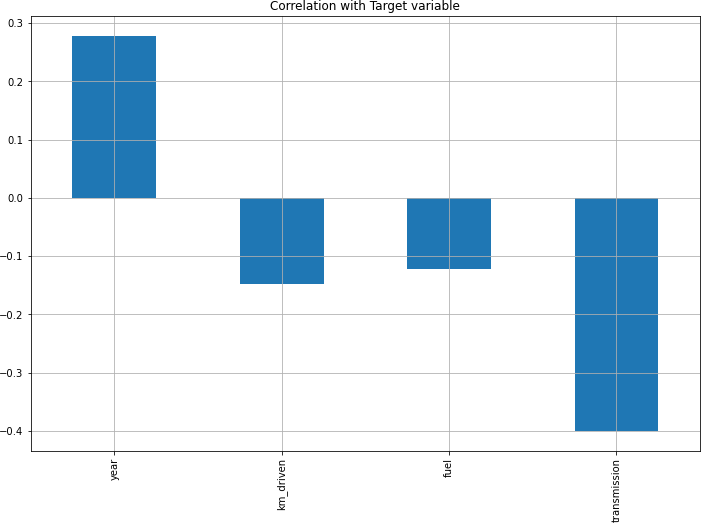
**Year vs Selling price**



**Correlation heatmap**



**Correlation with target variable**



# CONCLUSION

**Key Findings and Conclusions of the Study**

**->** After the completion of this project, we got an insight of how to collect data, pre-processing the data, analysing the data and building a model.

**->** First, we collected the used cars data from different websites like olx, cardekho and it was done by using Web scraping. The framework used for web scraping was Selenium, which has an advantage of automating our process of collecting data.

**->** We collected almost 10000 of data which contained the selling price and other related features.

**->** Then, the scrapped data was combined in a single data frame and saved in a csv file so that we can open it and analyse the data.

**->** We did data cleaning, data-preprocessing steps like finding and handling null values, removing words from numbers, converting object to int type, data visualization, handling outliers, etc.

**->** After separating our train and test data, we started running different machine learning classification algorithms to find out the best performing model.

**->** We found that RandomForest and KNeighbors Algorithms were performing well, according to their r2\_score and cross val scores.

**->** Then, we performed Hyperparameter Tuning techniques using GridSearchCV for getting the best parameters and improving the scores. In that, RandomForestRegressor performed well and we finalised that model.

**->** We saved the model in pkl format and then saved the predicted values in a csv format.

**The problems we faced during this project were:**

Website was poorly designed because the scrapping took a lot of time and there were many issues in accessing to next page.

More negative correlated data were present.

Many outliers were removed as some of the selling price and km values were out of range.

No information for handling these fast-paced websites were informed so that we were consuming more time in web scraping itself.