# Project Title: - Used Car Price Prediction

**Introduction: -**

The Prices of new cars increasing globally. Due to the increased price of new cars , The customers of used cars sales are on global increase.

I will be trying to develop the model to predict the best price to the vehicle.

By studying the variety of features, effectively we can determine the worthiness of used car

**Problem Statement: -**

Predicting the price of used cars given by the features

**Flow Chart: -**

1. Data Collection
2. Data Cleaning
3. Data Visualisation
4. Train and Test Machine Learning Models
5. Compare the accuracy of the Machine Learning Model and Select the best one

**Data Collection: -**

The Dataset that is used is from Kaggle.

For accurate and real time analysis, Data is Prepared from Scratch. The Data is Collected from QuikrCar Website.

The Data set Having 9 Columns like

```
Car_Name, Year, Selling_Price, Present_Price, Kms_Driven,
       Fuel_Type, Seller_Type, Transmission, Owner.
```

Dataset:

| | Car_Name | Year | Selling_Price | Present_Price | Kms_Driven | Fuel_Type | Seller_Type | Transmission | Owner |
|---|---|---|---|---|---|---|---|---|---|
| **0** | ritz | 2014 | 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual | 0 |
| **1** | sx4 | 2013 | 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual | 0 |
| **2** | ciaz | 2017 | 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual | 0 |
| **3** | wagon r | 2011 | 2.85 | 4.15 | 5200 | Petrol | Dealer | Manual | 0 |
| **4** | swift | 2014 | 4.60 | 6.87 | 42450 | Diesel | Dealer | Manual | 0 |

**Data Cleaning: -**

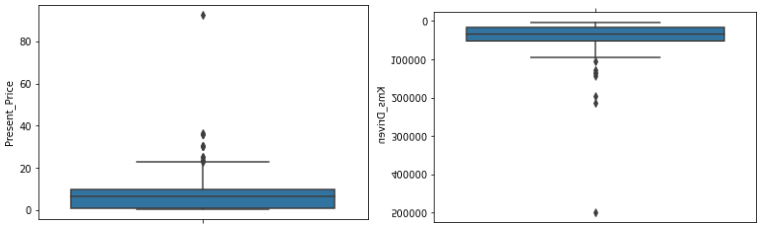
1. Loading the Data Set
2. After Loading the Data set the shape of the Data set is (301,9)
   301 rows and 9 columns
3. We have to Check for the Duplicates, NULL Values and the
   Data Types
4. Check all the columns are unique

Here I found all the Datatypes, Columns, and also no Null values and Duplicates

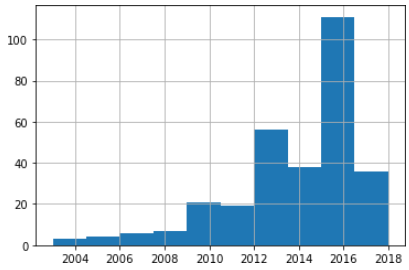
**Exploratory Data Analysis:**

**Data Visualisations**

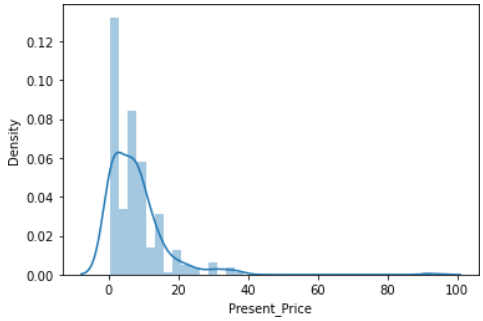
1. Check for the outliers, Here I found there are no outliers in the data



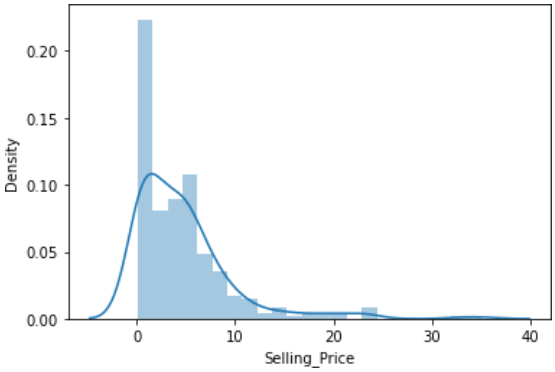

2. After clear analysis of data, it is observed that we have
   - Cars from 98 different brands
   - Most Cars are manufactured in the year 2015 to 2017 and Least Cars are manufactured in the year 2003 to 2005




   - Observed the occurrences of present price and selling price

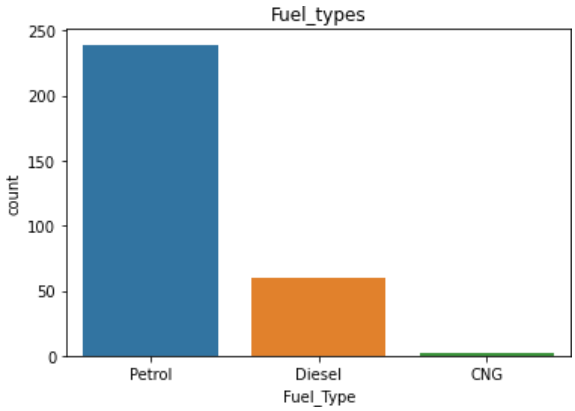



The Occurences of Present Price is high at 0 to 20

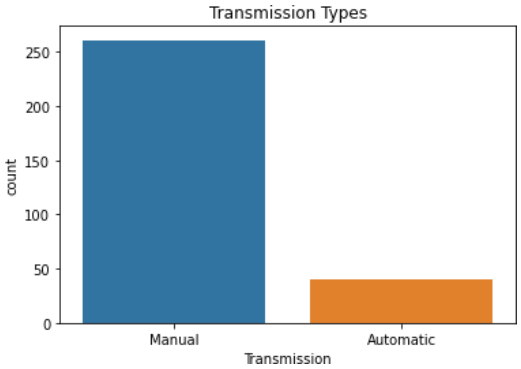



The Occurences of Selling Price is high at 0 to 10

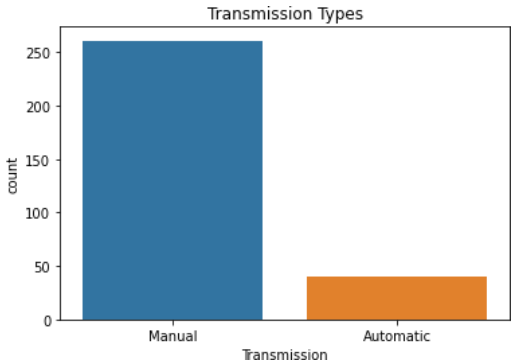
- Here, Petrol cars are more than Deisel and CNG



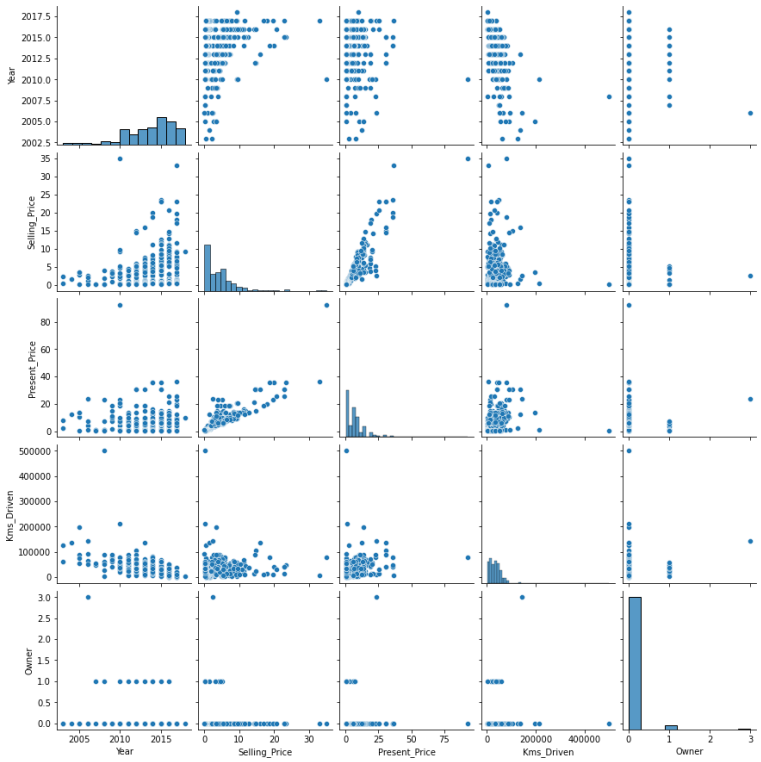

- Manual transmission are compared to Automatic

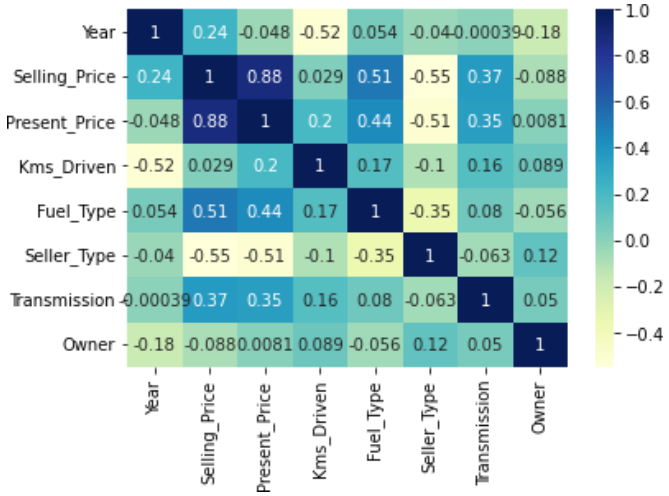



- Dealers seller types are more than Individual




- There is Strong relationhip between (selling_price ,present_price), (Kms_driven,selling price) and (fuel_type ,Selling_price)
  Weak relationship between (selling price,owner)




**Model Building :**

1. Encoding the Categorical Data

| | Year | Present_Price | Kms_Driven | Fuel_Type | Seller_Type | Transmission | Owner |
|---|---|---|---|---|---|---|---|
| **0** | 2014 | 5.59 | 27000 | 0 | 0 | 0 | 0 |
| **1** | 2013 | 9.54 | 43000 | 1 | 0 | 0 | 0 |
| **2** | 2017 | 9.85 | 6900 | 0 | 0 | 0 | 0 |
| **3** | 2011 | 4.15 | 5200 | 0 | 0 | 0 | 0 |
| **4** | 2014 | 6.87 | 42450 | 1 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **296** | 2016 | 11.60 | 33988 | 1 | 0 | 0 | 0 |
| **297** | 2015 | 5.90 | 60000 | 0 | 0 | 0 | 0 |
| **298** | 2009 | 11.00 | 87934 | 0 | 0 | 0 | 0 |
| **299** | 2017 | 12.50 | 9000 | 1 | 0 | 0 | 0 |
| **300** | 2016 | 5.90 | 5464 | 0 | 0 | 0 | 0 |

301 rows × 7 columns

2. Dropping the Irrelevant column Car_Type for building a model

3. The test set is 20% of overall dataset
4. The Performance metrics after running the model with Linear Regression Method

```
from sklearn.linear_model import LinearRegression
```

```
lr=LinearRegression()
lr.fit(x_train,y_train)
```

```
LinearRegression()
```

```
predictions = lr.predict(x_test)
```

```
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

```
mse=mean_squared_error(y_test,predictions)
np.sqrt(mse)
```

```
1.7134244742422997
```

```
mae=mean_absolute_error(y_test,predictions)
mae
```

```
1.2671998472916832
```

```
r2_score(y_test, predictions)
```

```
0.8401532365377773
```

5. The Performance metrics after running the model with Random Forest Regressor Method

```
from sklearn.ensemble import RandomForestRegressor
```

```
Rr=RandomForestRegressor()
Rr.fit(x_train,y_train)
```

```
RandomForestRegressor()
```

```
data_pred = Rr.predict(x_test)
```

```
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

```
mse=mean_squared_error(y_test,data_pred)
np.sqrt(mse)
```

```
0.8777990259734857
```

```
mae=mean_absolute_error(y_test,data_pred)
mae
```

```
0.4891491803278689
```

```
r2_score(y_test, data_pred)
```

```
0.9580468954421506
```

6. The Performance metrics r2_score increased in Random Forest regressor method when com pared to Linear Regression method

**Conclusion:**

According to this Used Car Prediction Dataset Random Forest Regressor is the Best fit Model