CARS4U

Contents

- ✓ Huge demand for used cars in the Indian Market
- ✓ Sales of new cars have slowed down in the recent past
- ✓ Demand is shifting towards the pre-owned market.
- ✓ Used cars are very different beasts with huge uncertainty in both pricing and supply.
- ✓ Pricing scheme of these used cars becomes important in order to grow in the market.

Objective

- Explore and visualize the dataset.
- Build a linear regression model to predict the prices of used cars.
- Generate a set of insights and recommendations that will help the business.

Business Problem Overview and Solution Approach

- Cars4U is a budding tech start-up that aims to find footholes in this market
- Linear Regression Model we shall use , to predict the price of used cars
- Help the business in devising profitable strategies using differential pricing
- Liner Regression Model will help to understand the independent variables required to predict the price.

Data Overview

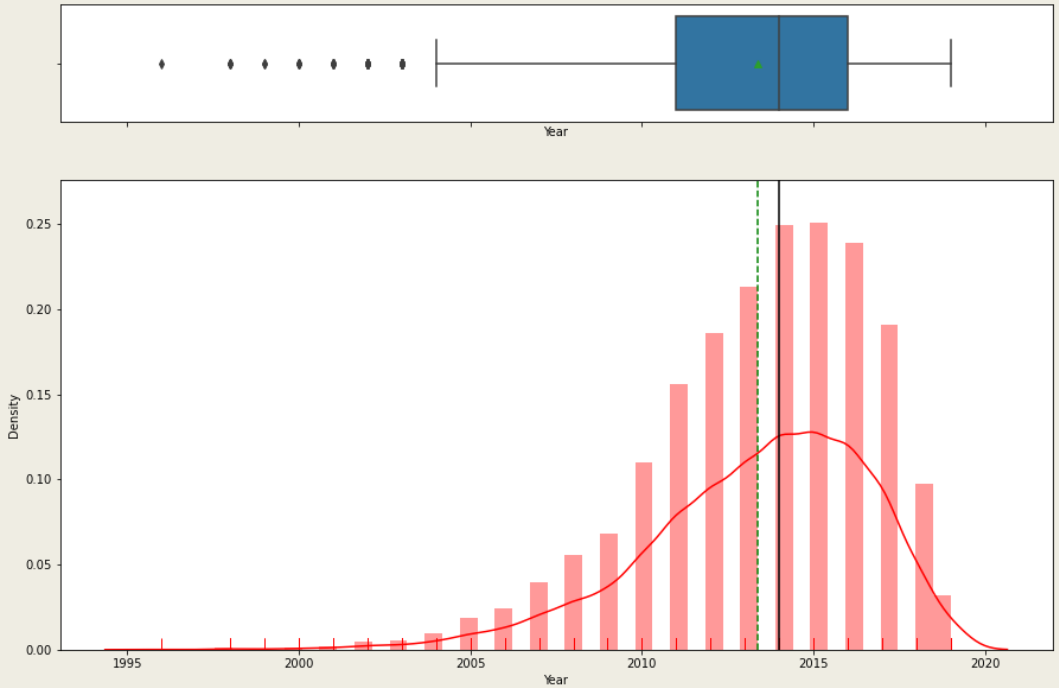
- - S. No. : Serial Number
- - Name : Name of the car which includes Brand name and Model name
- - Location : The location in which the car is being sold or is available for purchase Cities
- - Year : Manufacturing year of the car
- - Kilometers driven : The total kilometers driven in the car by the previous owner(s) in KM.
- - Fuel Type : The type of fuel used by the car. (Petrol, Diesel, Electric, CNG, LPG)
- - Transmission : The type of transmission used by the car. (Automatic / Manual)
- - Owner : Type of ownership
- - Mileage : The standard mileage offered by the car company in kmpl or km/kg
- - Engine : The displacement volume of the engine in CC.
- - Power : The maximum power of the engine in bhp.
- - Seats : The number of seats in the car.
- - New Price : The price of a new car of the same model in INR Lakhs.(1 Lakh = 100, 000)
- - Price : The price of the used car in INR Lakhs (1 Lakh = 100, 000)

Key Points –

- In the data set we have 7253 rows and 13 columns
- In the given dataset we do not have duplicate values
- We do have null values in the dataset
- Data Preprocessing is required for couple of variables.
- Out layers are supposed to be processed.

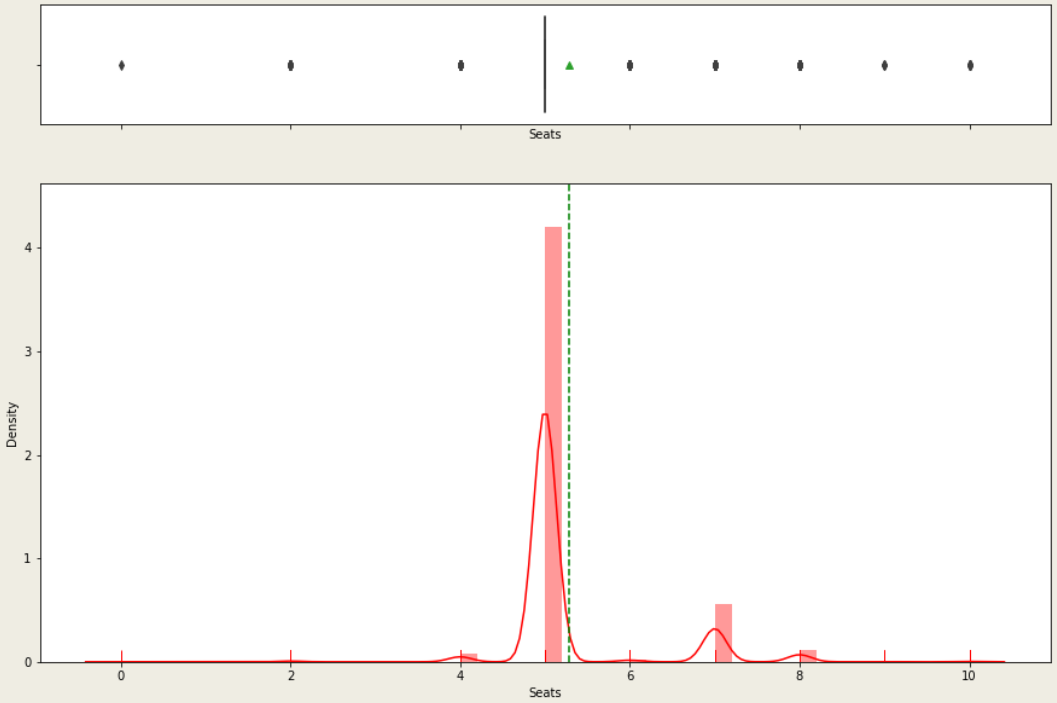
Exploratory Data Analysis

Year



- Majority of the cars are from 2014 and 2015
- Year has left skew , lot of data points are towards the left side.
- Data transformation is required

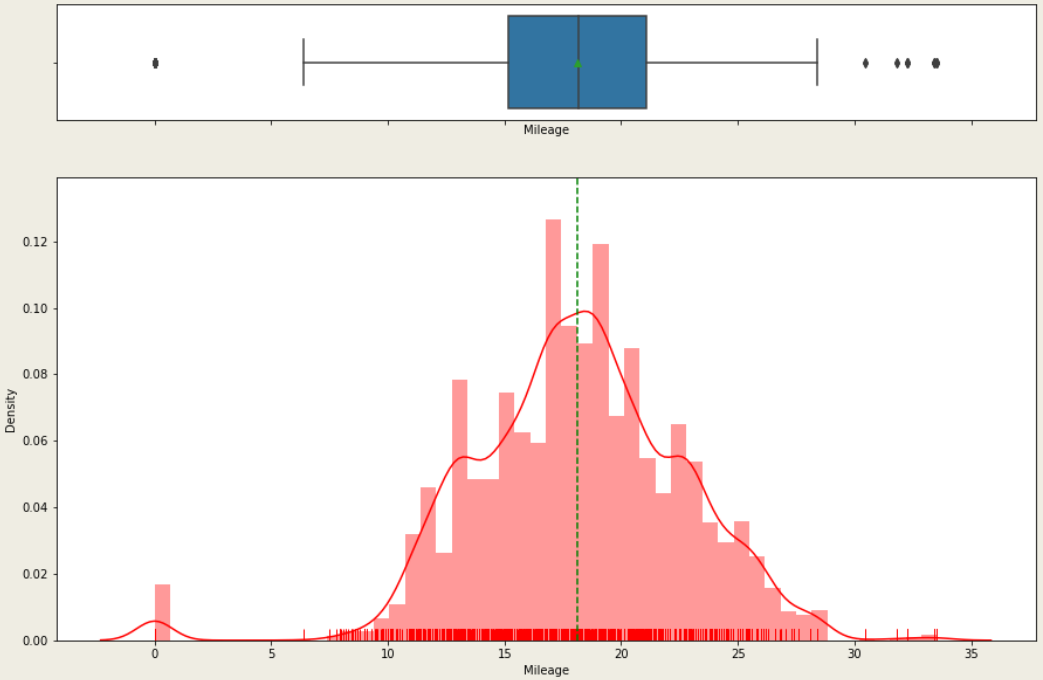
Seats



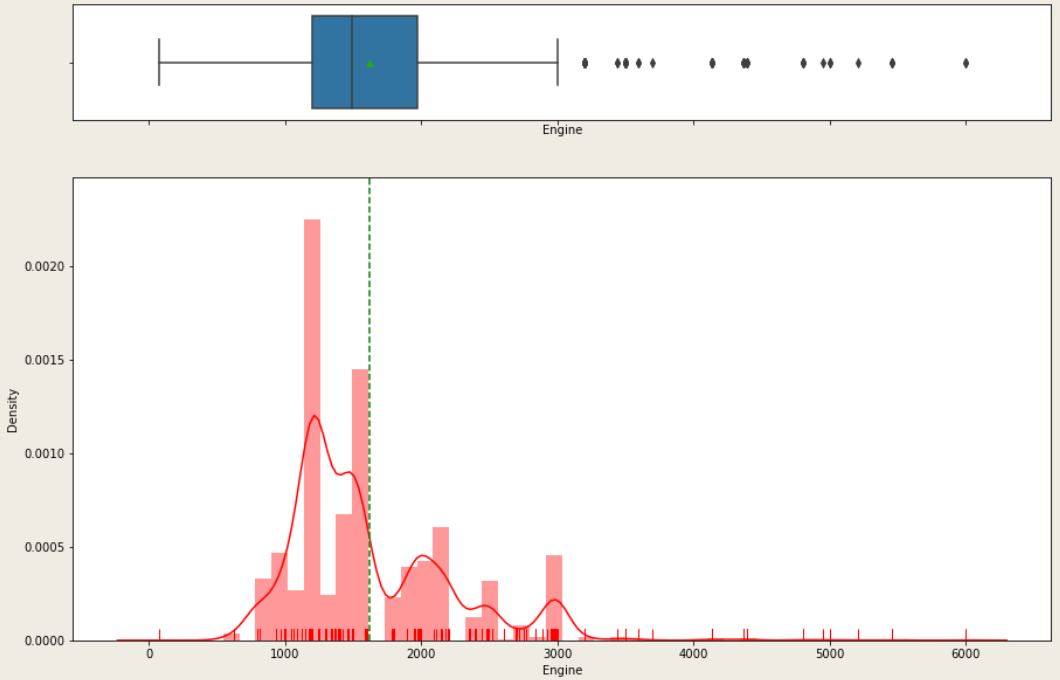
- Distribution of the data looks very normal.
- Majority of the cars has 5 seats
- It looks like we do have 2 seat cars , some may be are sports cars

Exploratory Data Analysis

Mileage



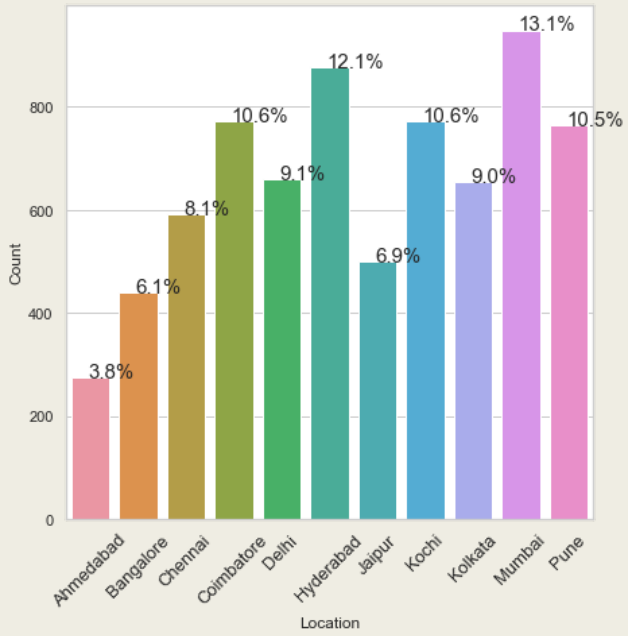
Engine



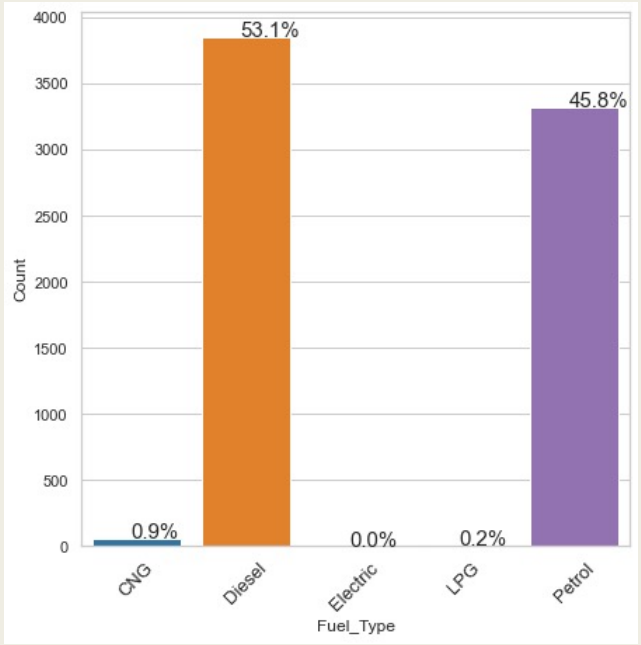
- Mileage looks normal distribution.
- Engine do have right skew data

Exploratory Data Analysis

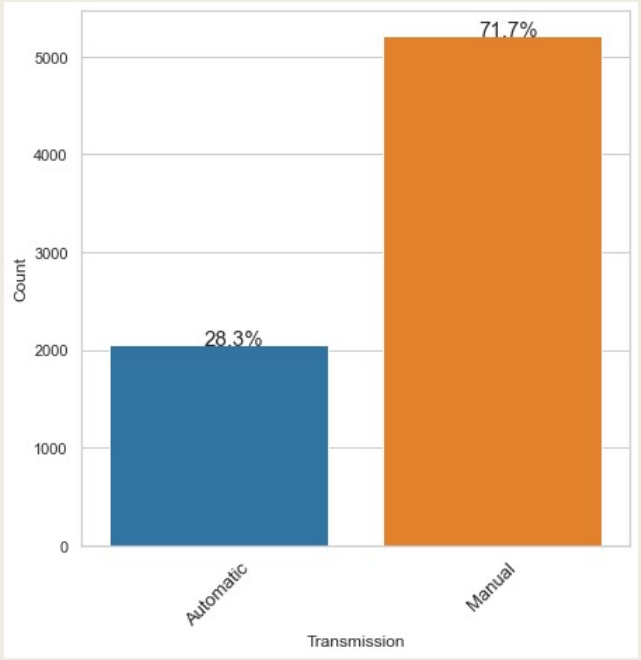
Location



Fuel Type

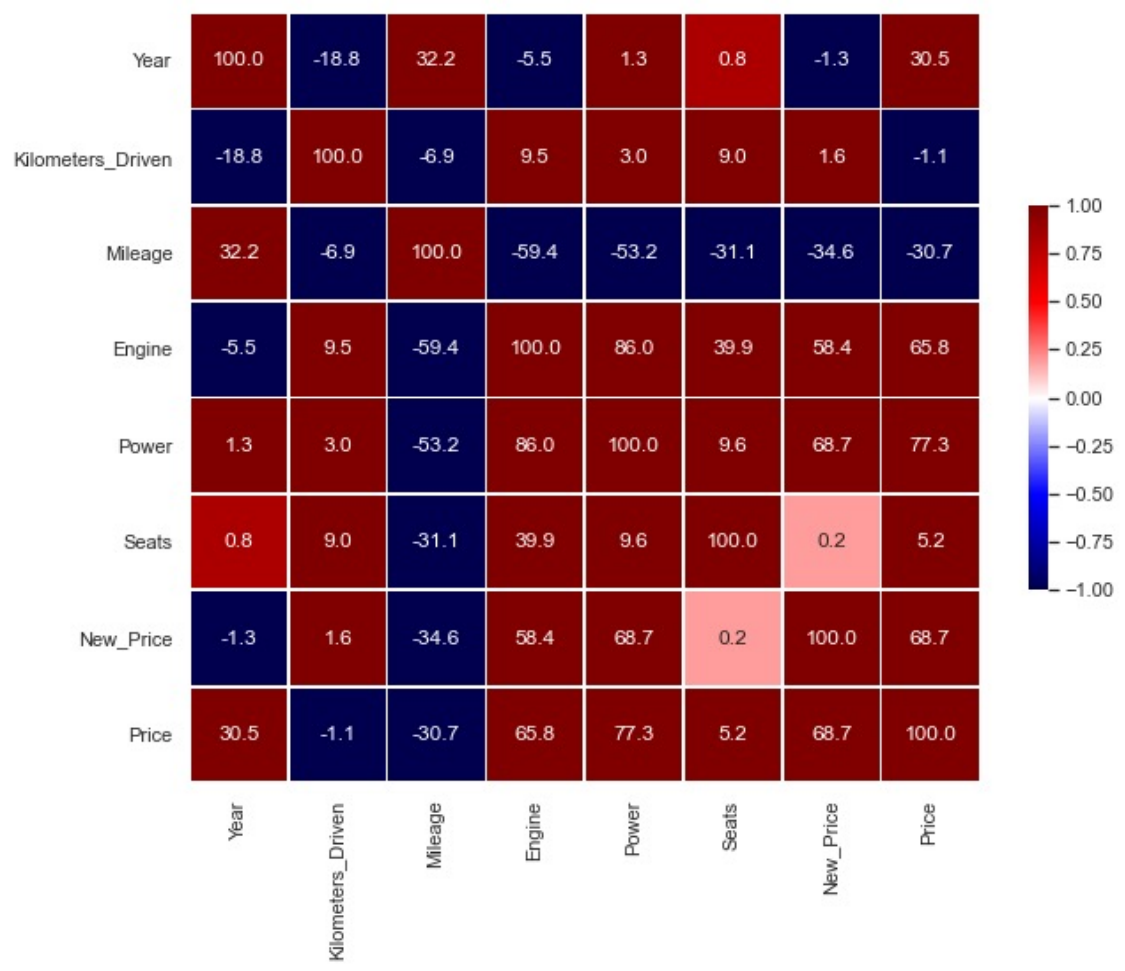


Transmission



- Mumbai has highest number of cars and Ahmedabad has lowest
- Majority of the cars belongs to Diesel and Petrol.
- Manuals cars are double when compared to Automatic cars

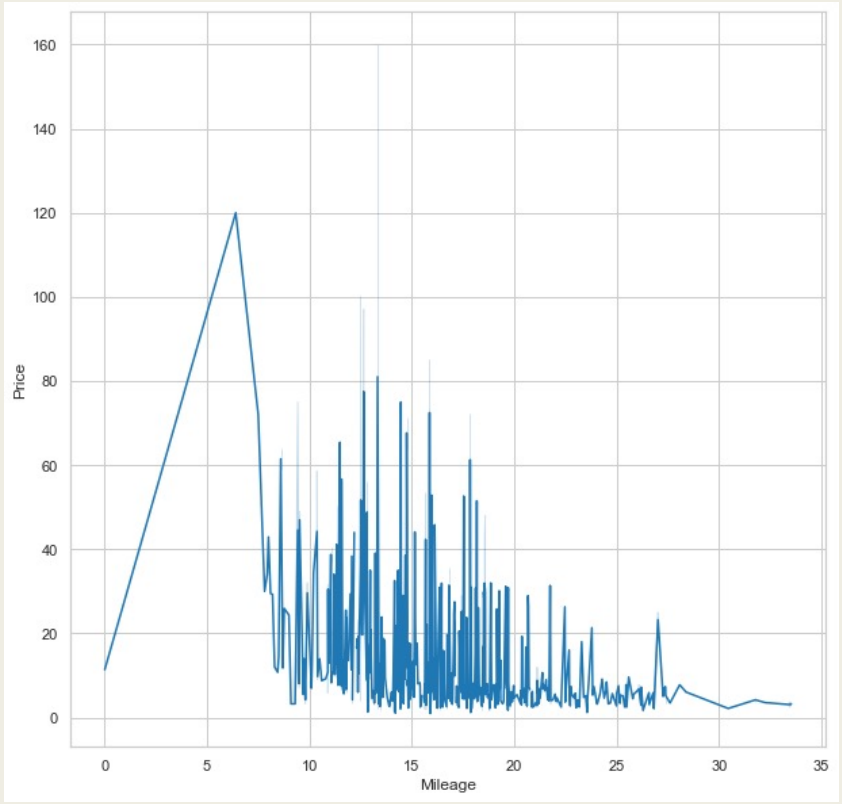
Exploratory Data Analysis - Correlation



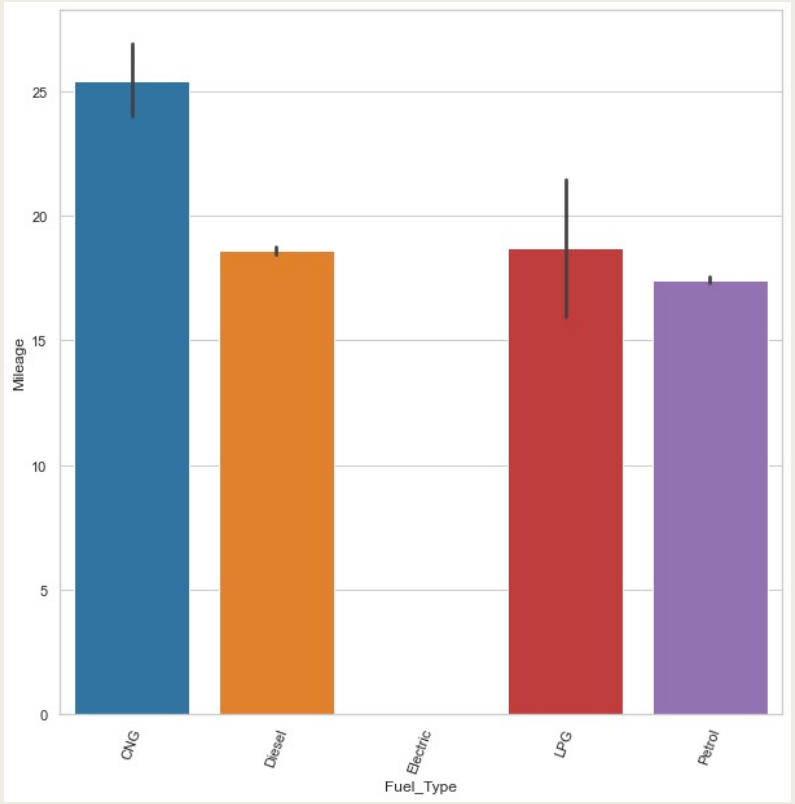
- ✓ - Majority of the variables are not correlated.
- ✓ - Power and New Price are correlated. Maybe one variable can be removed.
- ✓ - Power and Engine has correlation value of 86% , which means to say they are correlated.

Exploratory Data Analysis

Price vs Mileage



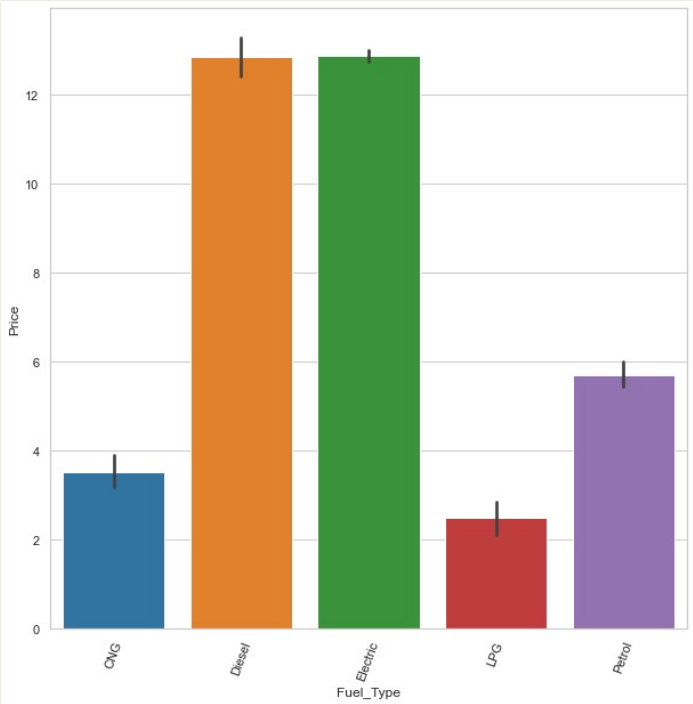
Fuel Type vs Mileage



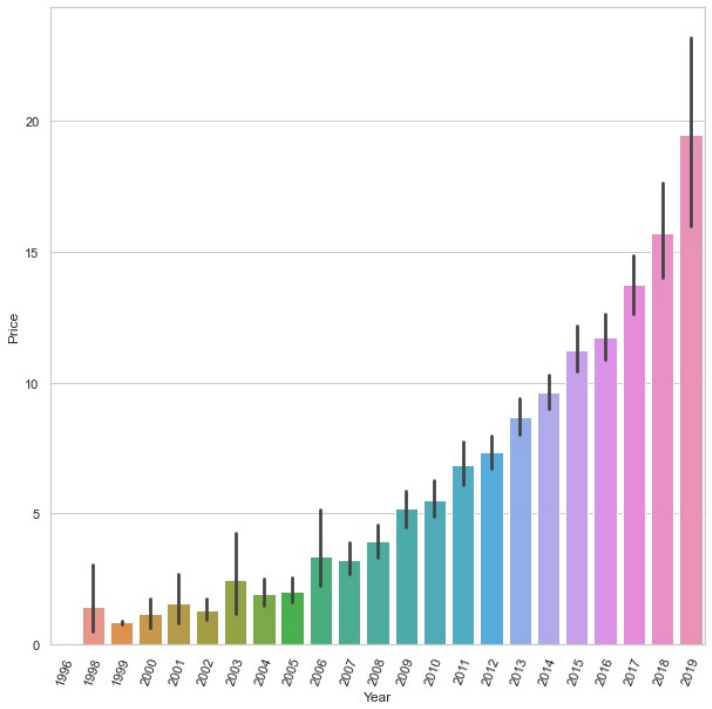
- Less the mileage of the car , more price
- For majority of the cars , mileage is from 10 to 30.
- Only for couple of car's mileage is more than 30
- We do have incomplete values in the dataset , due to that couple of cars has 0 mileage
- CNG cars have more mileage when compared to all others.
- LPG and Diesel has almost same mileage
- Petrol has less mileage when compared to other fuel types

Exploratory Data Analysis

Fuel Type vs Price



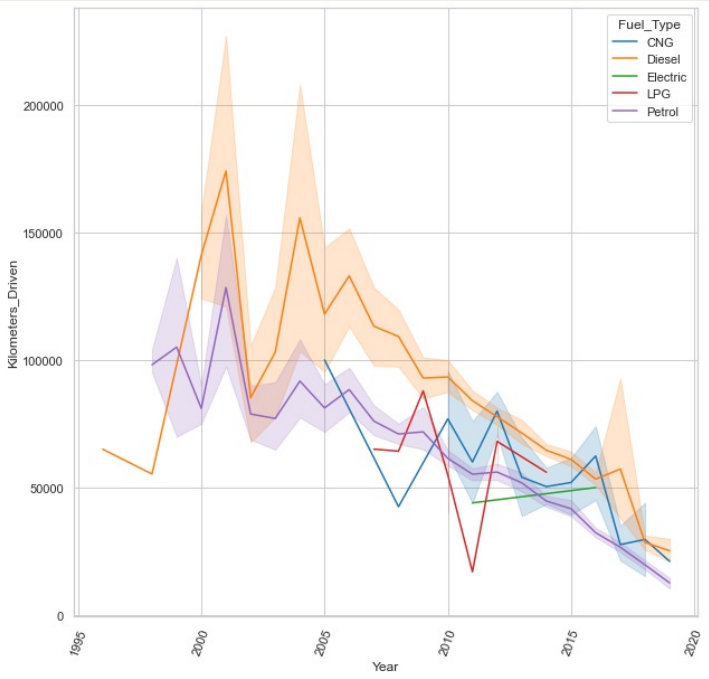
Year vs Price



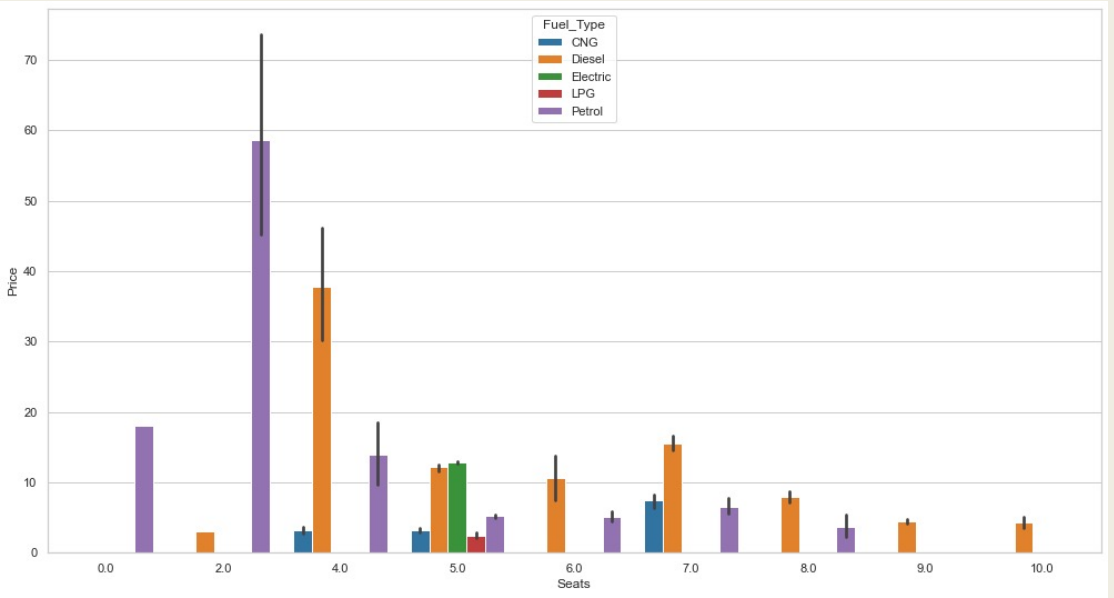
- ✓ - Diesel and Electric cars are bit costly when compared to other cars.
- ✓ - LPG is less when compared to all other.
- ✓ - High the year , more the price of the car.
- ✓ - 2019 cars has highest price in the given dataset.

Exploratory Data Analysis

Year vs Kilometer driven



Price vs Seats



- Cars with 2015 and greater , drove less miles when compared to others.
- Diesel cars drove more miles when compared to all other cars
- Petrol cars with 2 seats are costly when compared to all others , maybe these are sports cars.
- Diesel cars with 4 seats are next in line.

Conclusion

- Car Price come out to be very significant, as expected. variables - New Price and Name of the car is not required to predict the price of the car.
- Higher the mileage , Price of the car is less.
- Year 2015 and high has less number of kilometers when compared to others as expected.
- Less the mileage of the car , more price
- For majority of the cars , mileage is from 10 to 30 just for few cars mileage is more than 30
- CNG cars have more mileage when compared to all others , after that LPG and Diesel has almost same mileage when compare to all other Petrol has less mileage.
- Diesel and Electric cars are a bit costly, LPG is less when compared to all others.
- Cars with 2015 and greater year, drove less miles when compared to others.
- Diesel cars drove more miles when compared to all other cars
- Petrol cars with 2 seats are costly when compared to all others , maybe these are sports cars. Diesel cars with 4 seats are next in line.
- Power and New Price are correlated. Maybe one variable can be removed.
- Power and Engine has correlation value of 86% , which means to say they also correlated.

- Less number of Years , less mileage cars has more value.

Model Performance Summary

- **Linear regression** attempts to model the relationship between two variables by fitting a **linear** equation to observed data. Regression Model was performed on the data provided with assumptions of Linear Model
- Missing values were replaced with Median.
- Out layers were treated to get better accuracy
- Data was split into 70 and 30 percent ratio.
- Root Mean Square error of the model is 0.11 with Absolute mean value of 0.08
- Highest accuracy of the model so far is 74% , Using R^2 accuracy was evaluated.
  1. *Mean of residuals is ZERO*
  2. *Linearity of variables*
  3. *Normality of error terms*
  4. *No Heteroscedacity*

Accuracy on Training Data	71%
Accuracy on Test Data	74%

Business Insights and Recommendations

- Using model we can predict car price for almost 74% accuracy.
- Rather then Linear Regression , random forest model might have provided some more high accuracy.
- Coefficients of the equation are: [6.12289782e-04 4.80148904e+02 -1.36196508e-01 -1.22280577e-01 -3.09944308e-02 -1.21222387e-02 9.28298060e-01 1.16250943e+00 -2.40227134e-04 -1.21375435e-01]