Used Cars Project
By Naeem Sufi

Content

- Brief description of data provided
- Cleaning of data
- Graphs showing the factors most heavily impacting the target attribute
- Insights from the graphs showing the factors most heavily impacting the target attribute
- Overview of ML model and its parameters
- Summary of most important factors used by the ML model for prediction
- Summary of key performance metrics for training and test data in tabular format for comparison

Dataset Description

Variable	Description
Name	Name of the car
Location	Location of the car
Year	Year of Model
Kilometers_Driven	Kilometers driven by cars
Fuel_Type	Type of Fuel on which car runs
Transmission	Type of Transmission on which car runs
Owner_Type	Type of Owner car has
Mileage	Mileage provided by car
Engine	Type of Engine in car
Power	Horsepower of car
Seats	Number of Seats in a car
New_Price	New Prices of car at the moment
Price	Price of car at the moment of release

Snapshot of Dataset

	S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.60	998.0	58.16	5.0	NaN	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	NaN	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.20	1199.0	88.70	5.0	8.61	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	NaN	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968.0	140.80	5.0	NaN	17.74

Data Description

	Year	Kilometers_Driven	Mileage	Engine	Power	Seats	New_Price	Price
count	7253.000000	7.253000e+03	7251.000000	7207.000000	7078.000000	7200.000000	1006.000000	6019.000000
mean	2013.365366	5.869906e+04	18.141580	1616.573470	112.765214	5.279722	22.624205	9.479468
std	3.254421	8.442772e+04	4.562197	595.285137	53.493553	0.811660	27.437394	11.187917
min	1996.000000	1.710000e+02	0.000000	72.000000	34.200000	0.000000	1.580000	0.440000
25%	2011.000000	3.400000e+04	15.170000	1198.000000	75.000000	5.000000	7.880000	3.500000
50%	2014.000000	5.341600e+04	18.160000	1493.000000	94.000000	5.000000	11.540000	5.640000
75%	2016.000000	7.300000e+04	21.100000	1968.000000	138.100000	5.000000	25.697500	9.950000
max	2019.000000	6.500000e+06	33.540000	5998.000000	616.000000	10.000000	375.000000	160.000000

Type of columns in dataset

```
Location             object
Year                  int64
Kilometers_Driven     int64
Fuel_Type            object
Transmission         object
Owner_Type           object
Mileage             float64
Engine              float64
Power               float64
Seats               float64
New_Price           float64
Price               float64
dtype: object
```

Missing values

```
Location              0
Year                  0
Kilometers_Driven     0
Fuel_Type             0
Transmission          0
Owner_Type            0
Mileage               2
Engine               46
Power               175
Seats                53
New_Price          6247
Price              1234
dtype: int64
```

Cleaning the Dataset

The dataset is cleaned for any missing data.

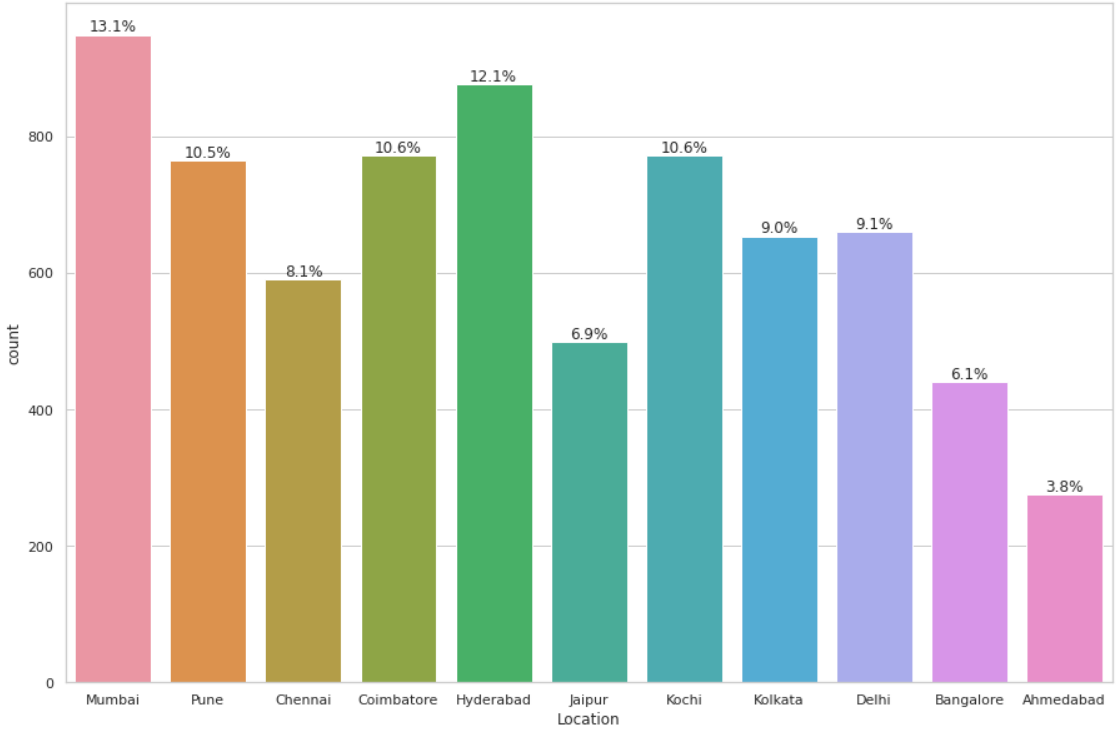
Zeros in the figure represents there are no missing values.

```
Year                          0
Kilometers_Driven             0
Mileage                       0
Engine                        0
Power                         0
Seats                         0
New_Price                     0
Price                         0
Fuel_Type_CNG                 0
Fuel_Type_Diesel              0
Fuel_Type_Electric            0
Fuel_Type_LPG                 0
Fuel_Type_Petrol              0
Transmission_Automatic        0
Transmission_Manual           0
Owner_Type_First              0
Owner_Type_Fourth & Above     0
Owner_Type_Second             0
Owner_Type_Third              0
Location_Ahmedabad            0
Location_Bangalore            0
Location_Chennai              0
Location_Coimbatore           0
Location_Delhi                0
Location_Hyderabad            0
Location_Jaipur               0
Location_Kochi                0
Location_Kolkata              0
Location_Mumbai               0
Location_Pune                 0
dtype: int64
```

Exploratory Analysis

Location

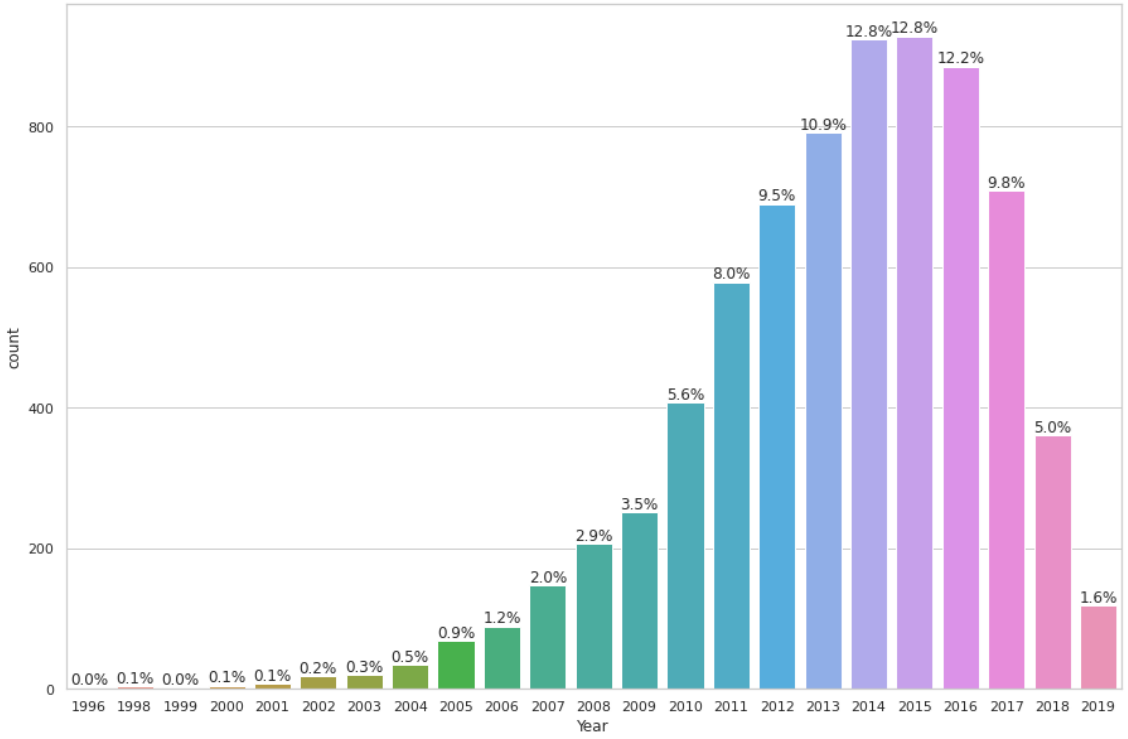
- Highest number of cars belong to Mumbai region with 13.1%
- Least number of cars belong to Ahmedabad.



Exploratory Analysis

Year

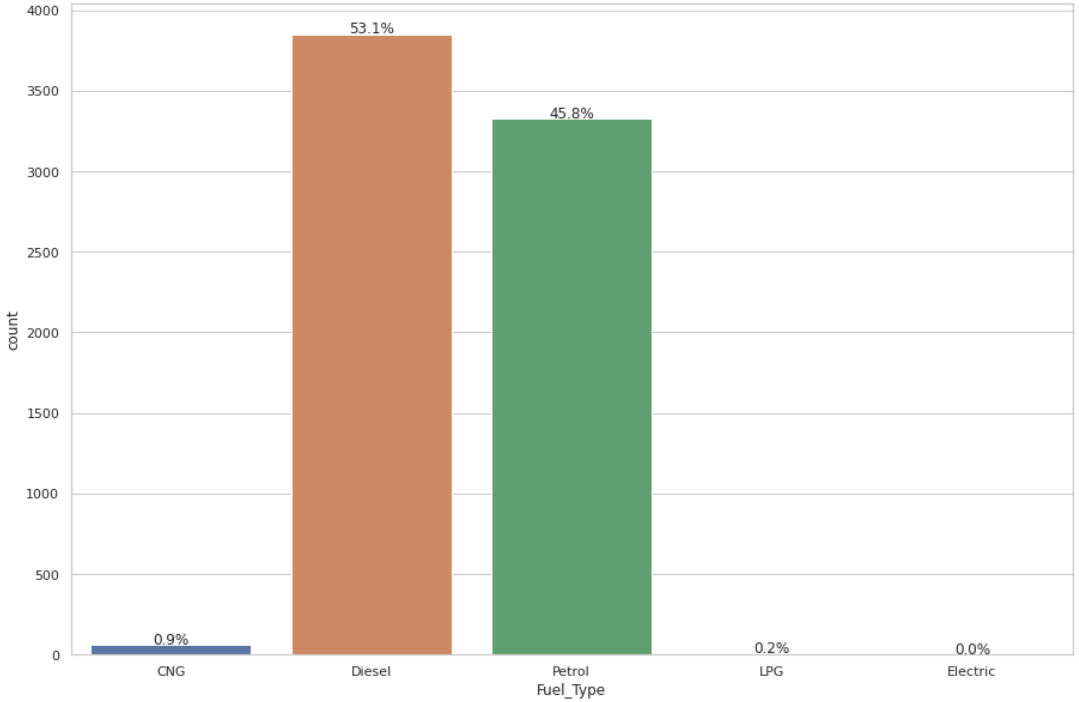
- Most cars are from the year 2014 and 2015 with same % i.e 12.8
- As we go in past number of cars decrease i.e we have few old cars
- Most cars available are from 2010 onwards



Exploratory Analysis

Fuel Type

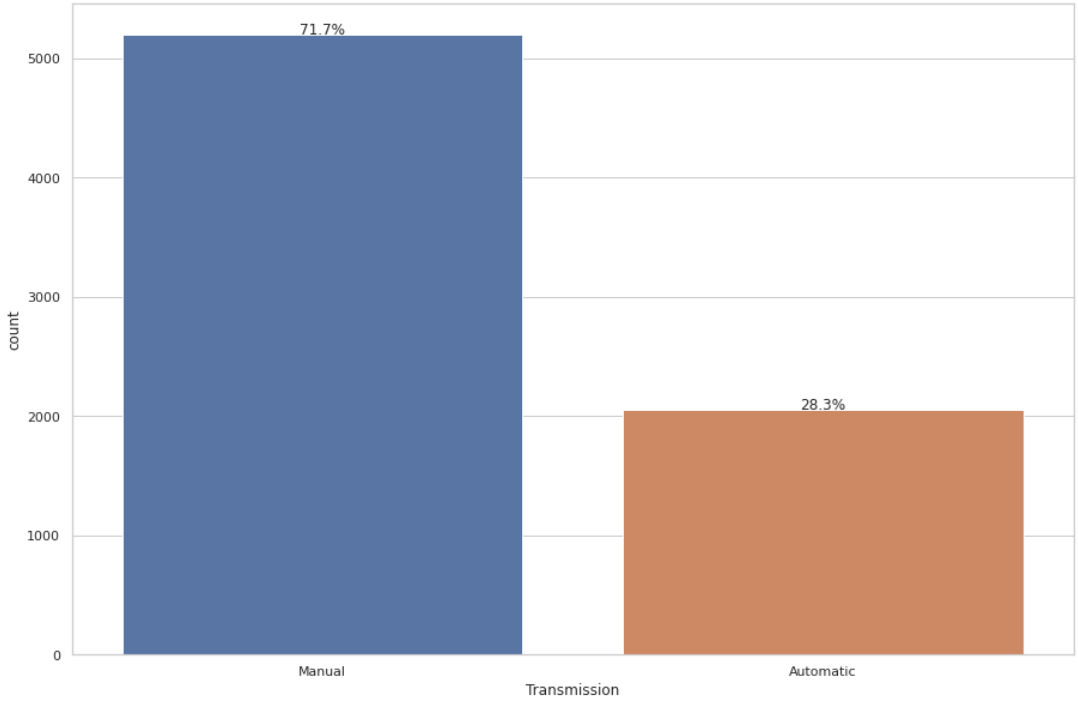
- Most common type of Fuel used in car is Diesel Followed by Petrol
- There is least number of cars on Electricity
- CNG LPG is also not very common



Exploratory Analysis

Transmission

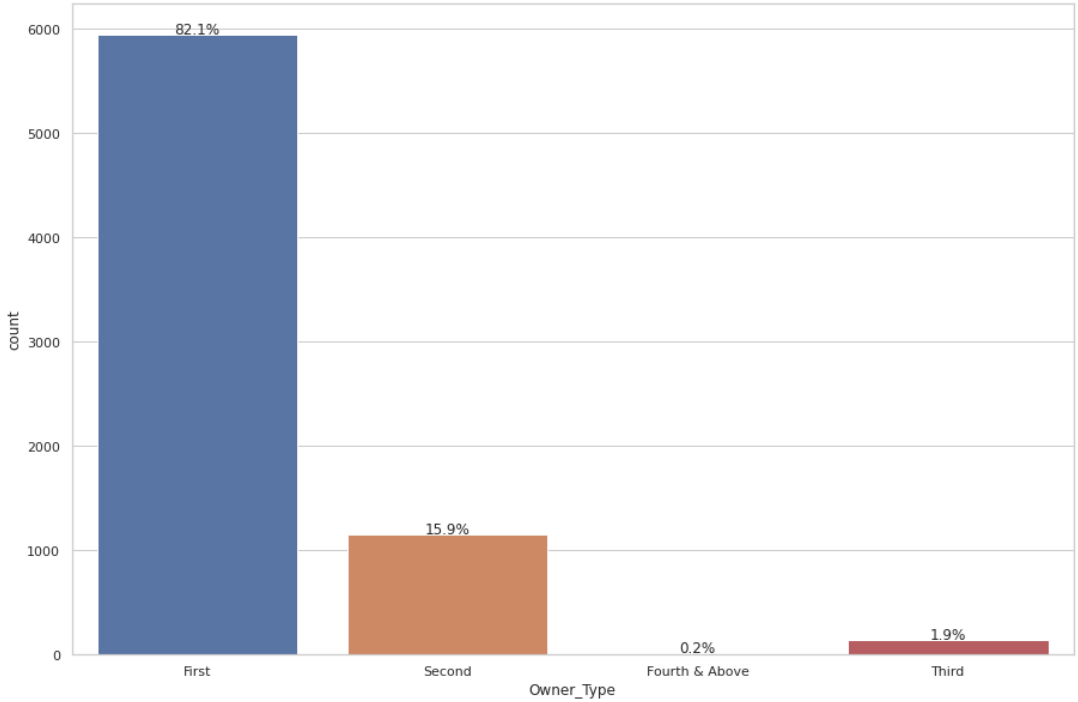
- Most of the cars sold are manual
- There are fewer cars that are are automatic



Exploratory Analysis

Owner Type

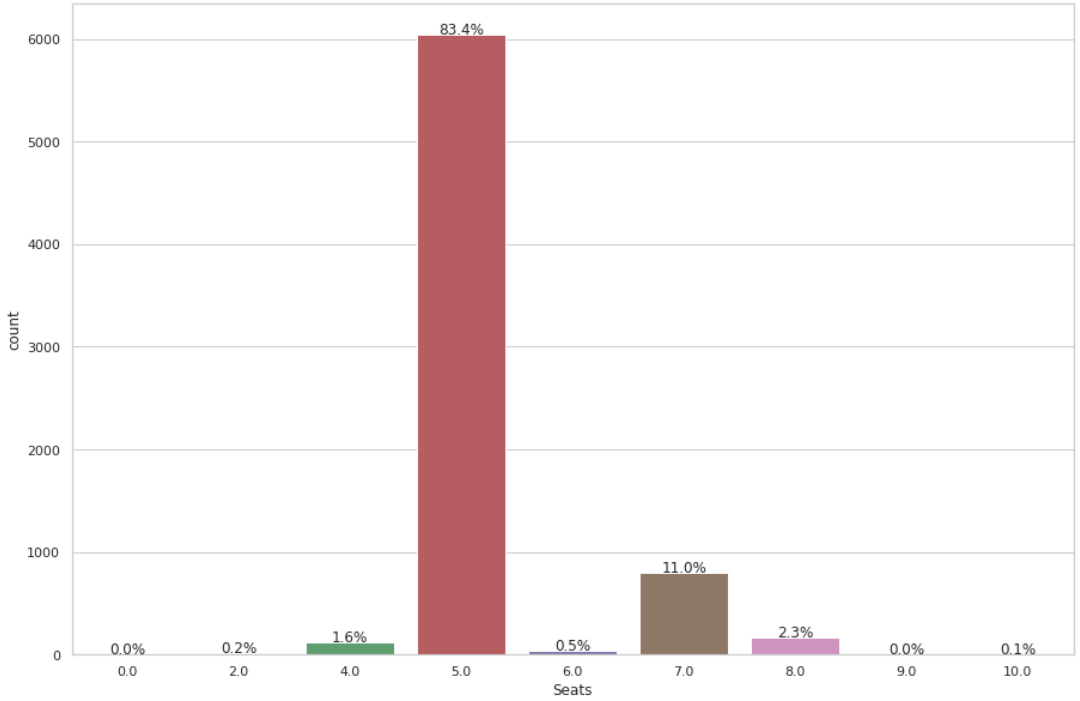
- Most of the cars have first owner
- Least number of cars have 4th owner



Exploratory Analysis

Seats

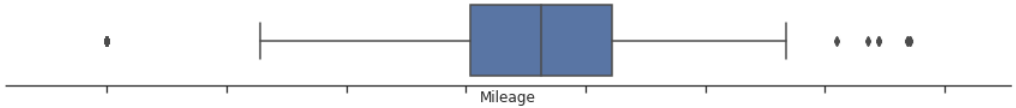
- Most Common number of seats in the cars are 5 and there is no car with 0 seats which makes sense
- There are very few cars with 9 and 10 seats
- Although there are cars with 7 seats which are second most common car

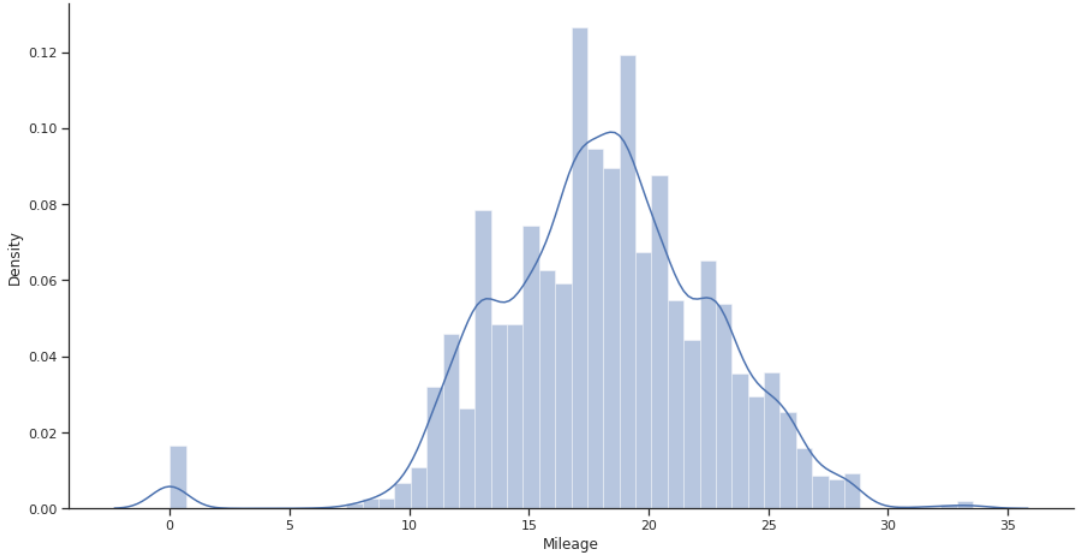


Exploratory Analysis

Mileage

- Outlier in mileage of car are towards far end of both sides

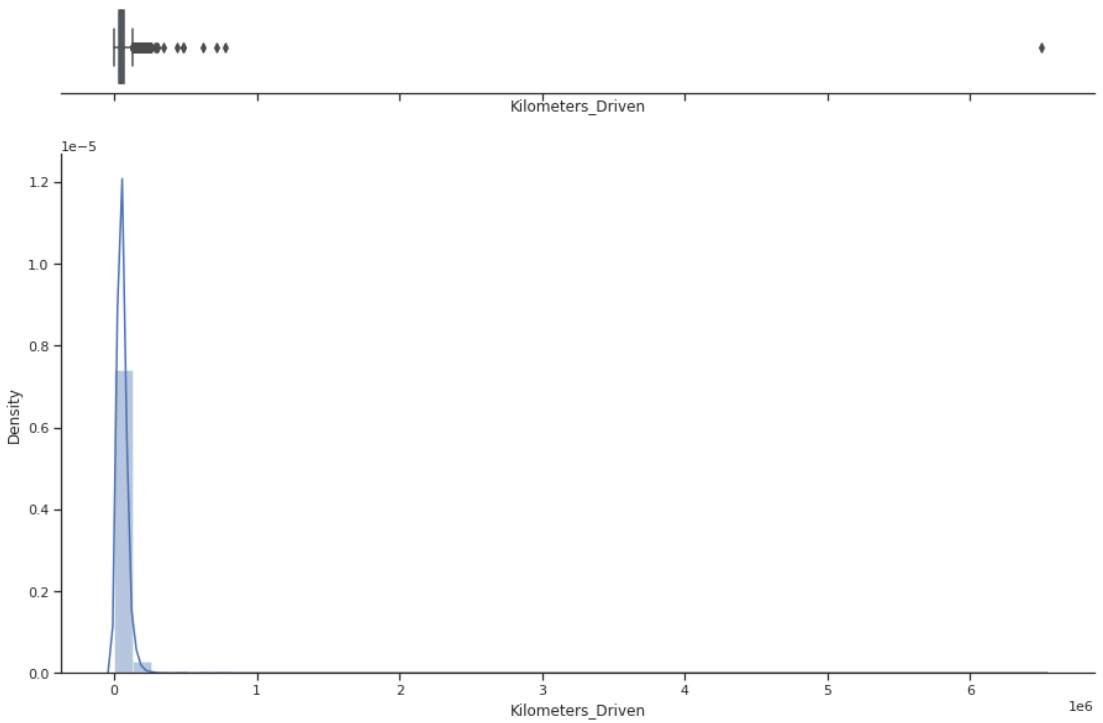




Exploratory Analysis

Kilometer Driven

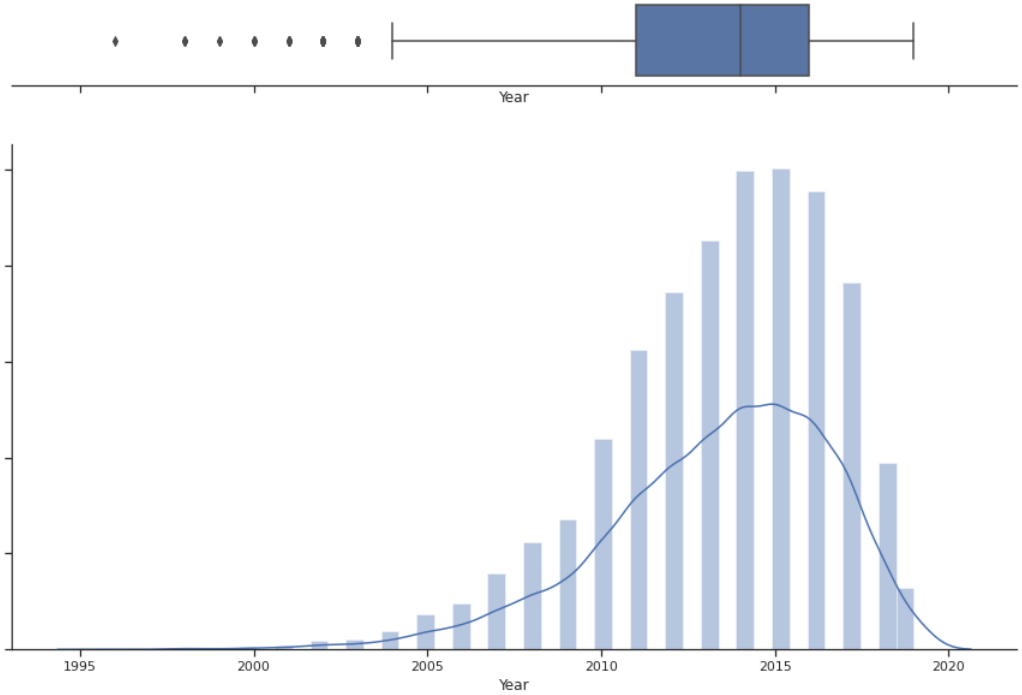
- Kilometer driven is right skewed



Exploratory Analysis

Year

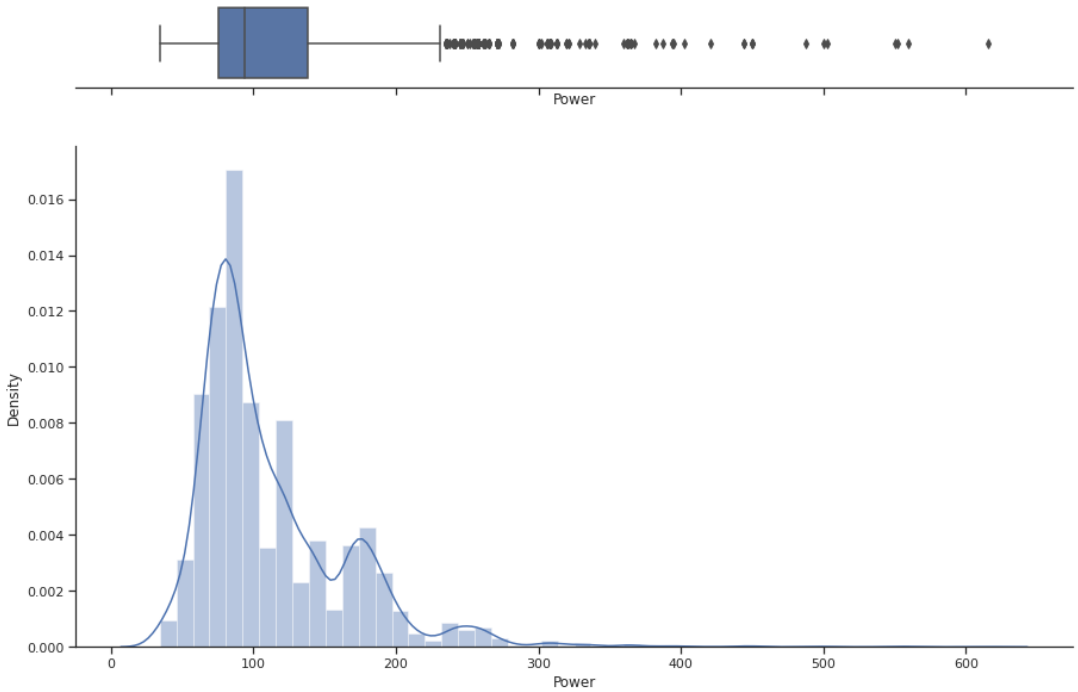
- There are very few old cars
- But the outliers are towards years in past 2010
- 2014 and 2015 are most common years for car models



Exploratory Analysis

Power

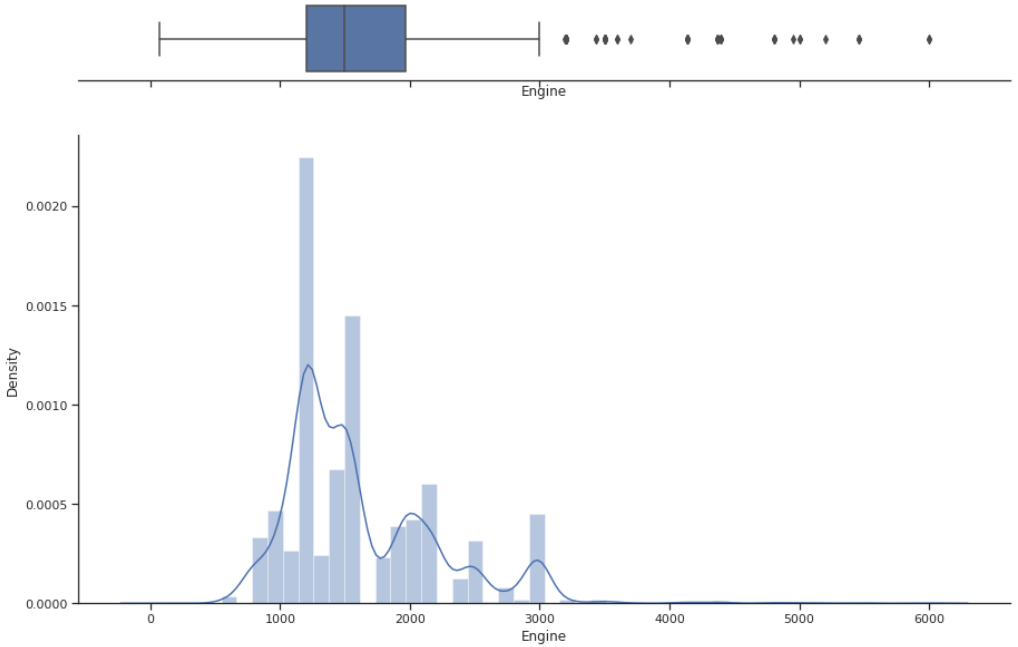
- Outliers in Power are towards far end on the right side
- 100 is most power we have
- There are cars with higher power than 100 but are considered outliers



Exploratory Analysis

Engine

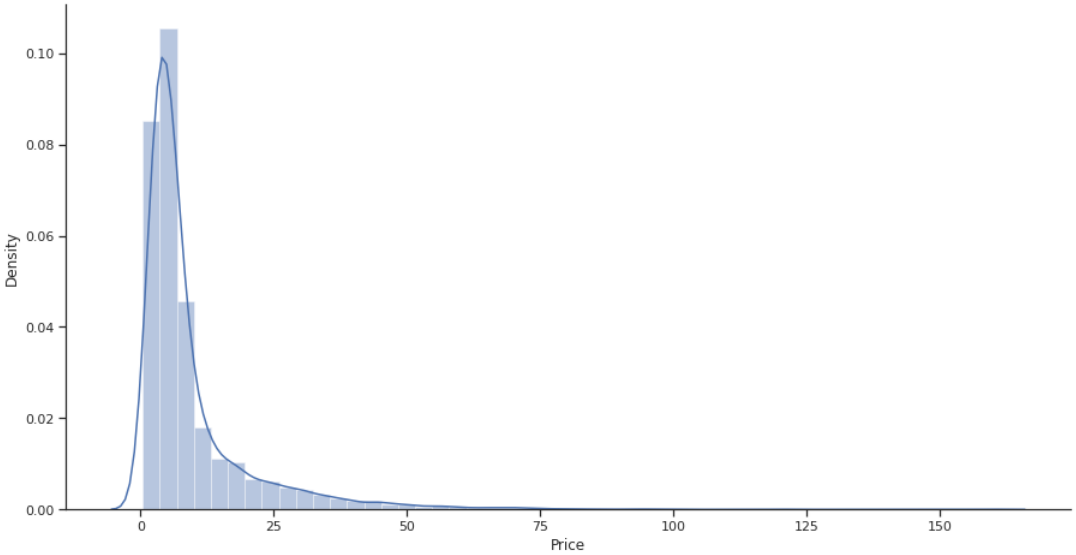
- Engine column is right skewed
- It has outlier on far right i.e car with more Engine value in CC are not common
- Most Common Range for car is with values between 1000CC to 1500CC



Exploratory Analysis

Price

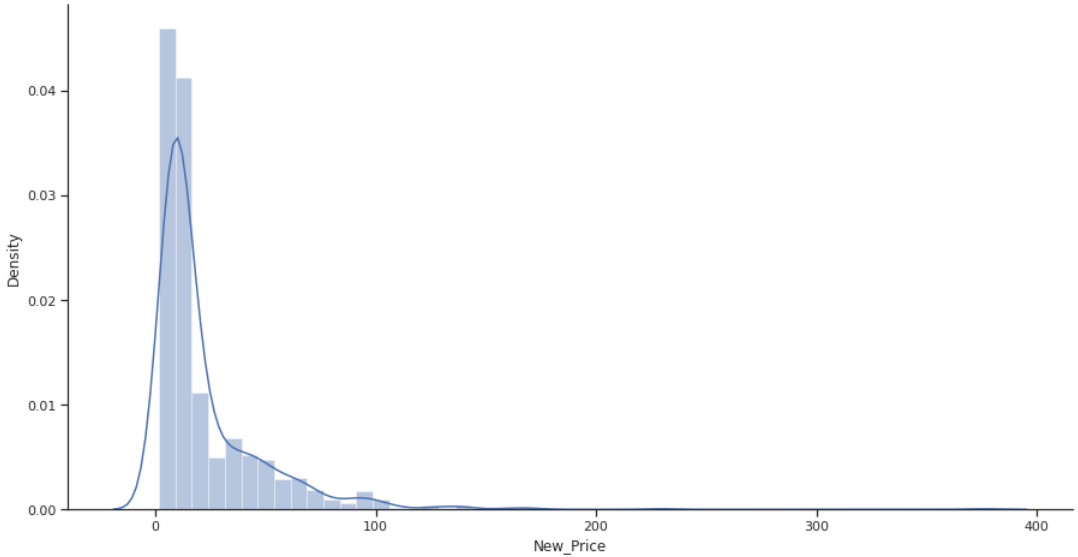
- Common Price of car lies in 5 to 20 lacs
- There very few cars with very high price considered outliers on far end of right side
- Most of the cars are below 20 lacs



Exploratory Analysis

New Price

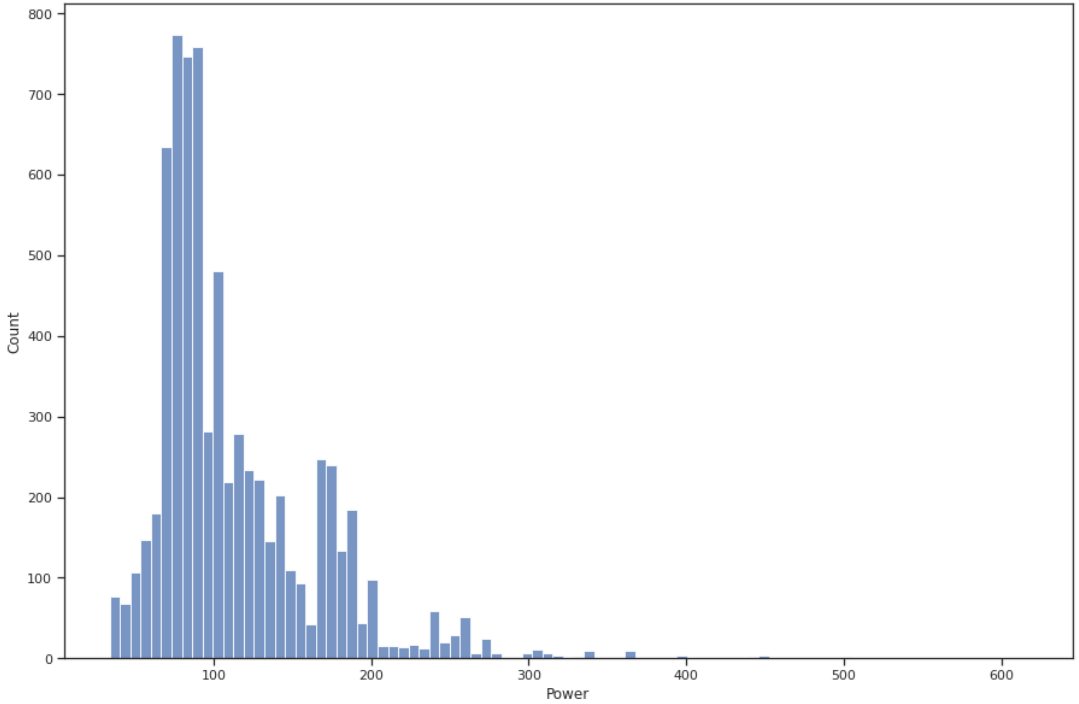
- Common New Price of car lies in 20 to 50 lacs
- There are very few cars with very high price considered outliers on far end of right side



Exploratory Analysis

Power

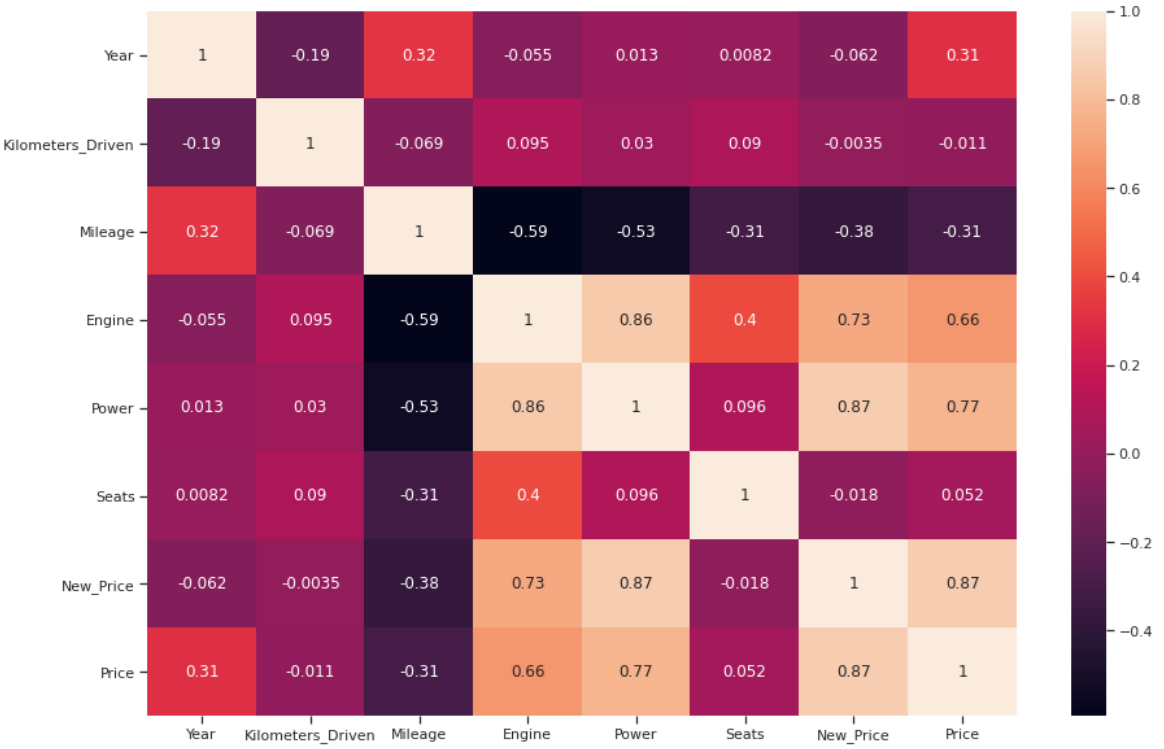
- Power is right skewed ie



Exploratory Analysis

Correlation Matrix

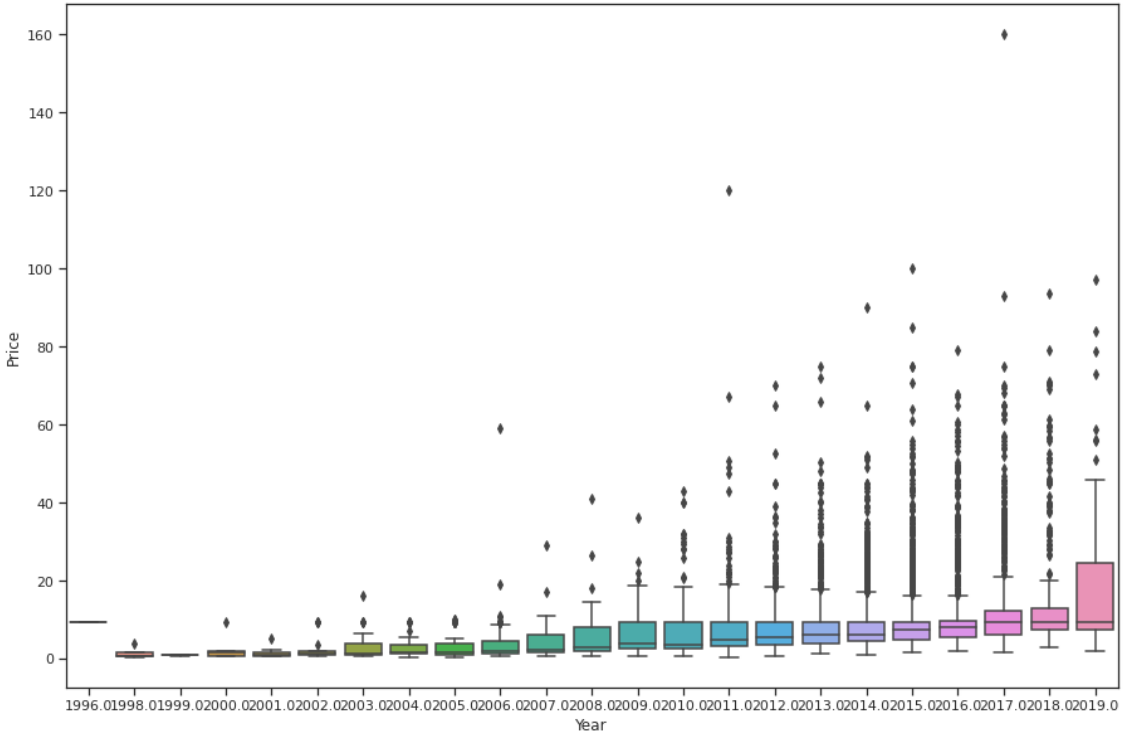
- Price is highly positively correlated to engine and power which means that with increase in engine and power price increases
- Price is negative related to Mileag but not strongly which means increase in mileage decrease prices
- Mileage is also negative correlated to engine and power which means increase in mileage decreases them



Exploratory Analysis

Price Vs Year

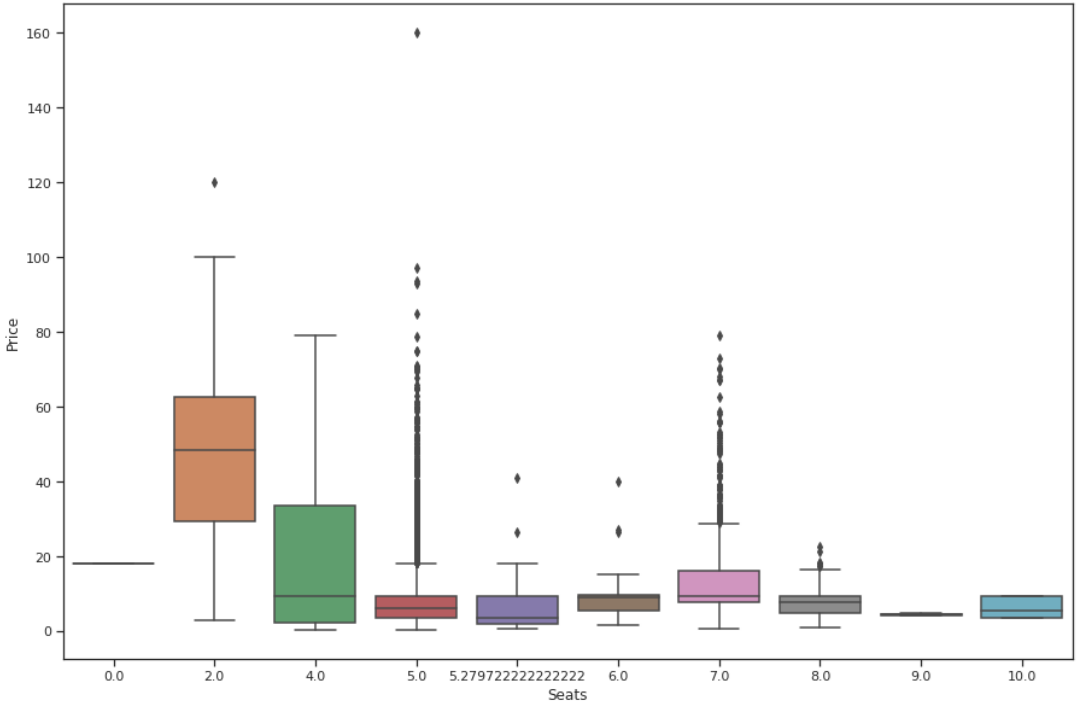
- Price of the car increases with year
- With Increase in price we can see some outliers as well



Exploratory Analysis

Price Vs Seat

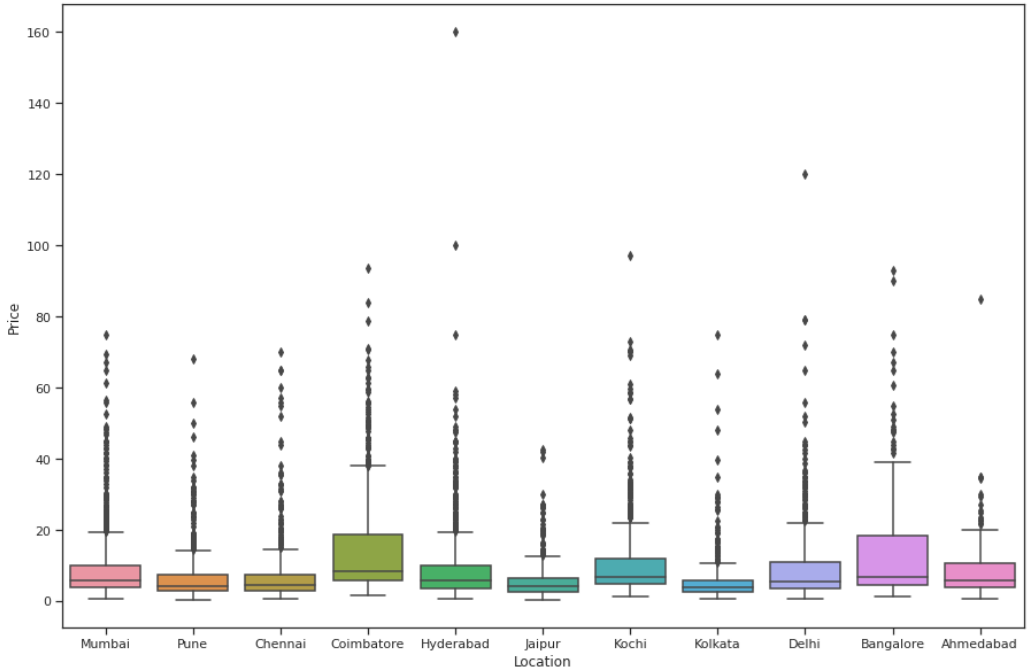
- With increase in seats outliers increase up side
- There is particular relationship that can be inferred between price and seats
- 2 seater cars are more expensive



Exploratory Analysis

Price Vs Location

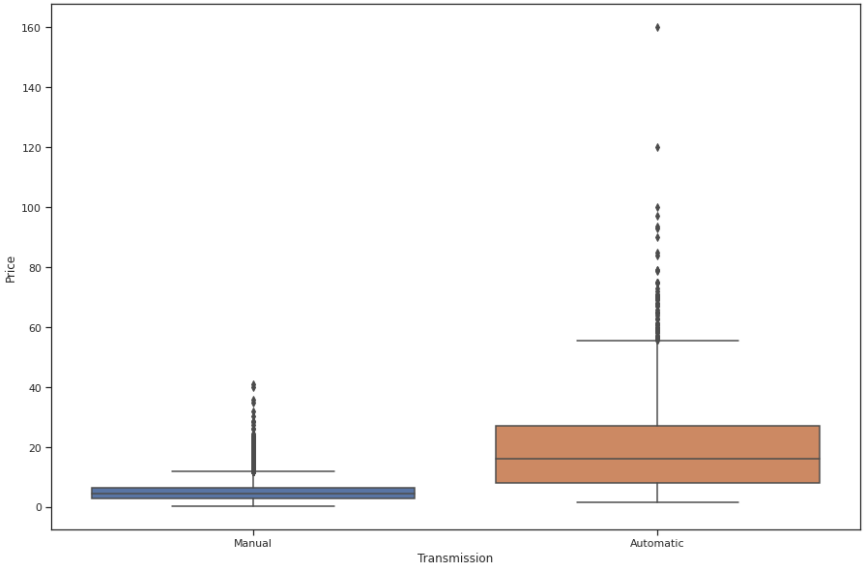
- Bangalore has the highest priced cars followed by Coimbatore
- Hyderabad has the most outliers in terms of price at upper end
- Kolkata hasthe lowest range when it comes to car prices followed by Jaipur and Chennai



Exploratory Analysis

Price Vs Transmission

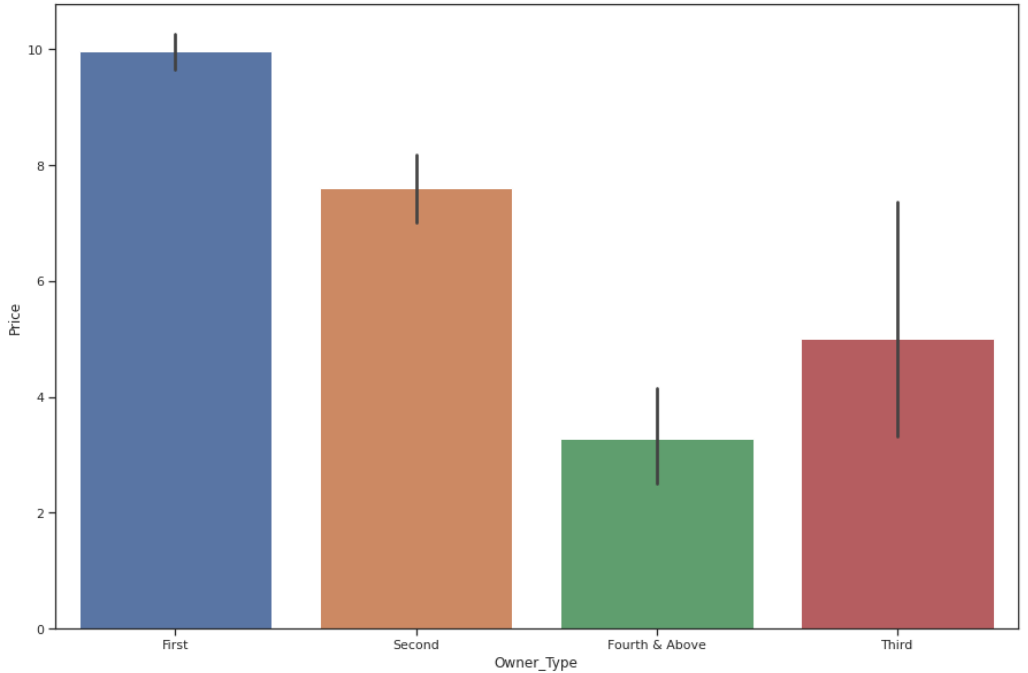
- Automatic cars are more Expensive.



Exploratory Analysis

Price Vs Owner type

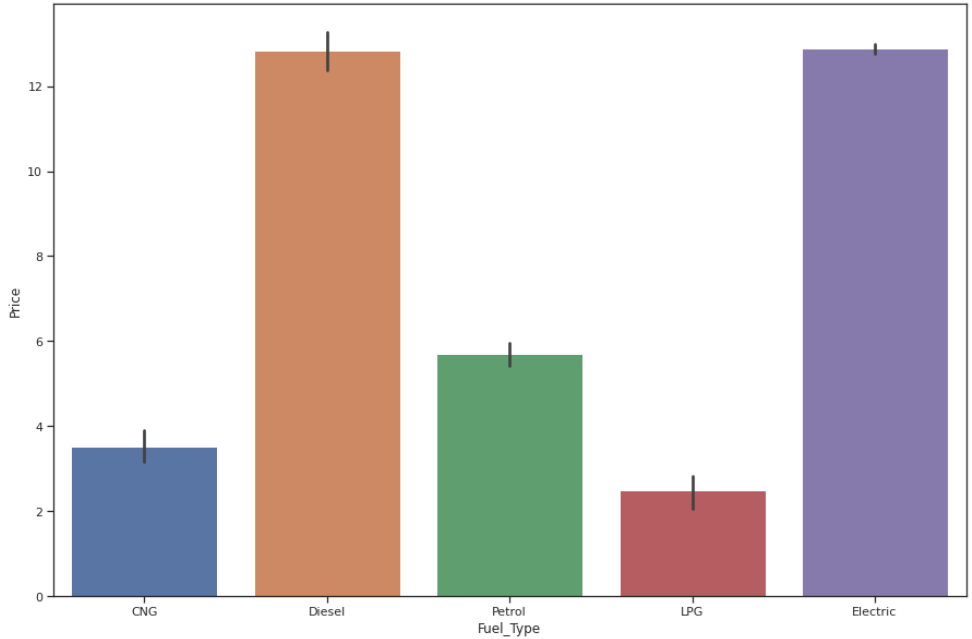
- First Hand cars are more expensive



Exploratory Analysis

Price Vs Fuel type

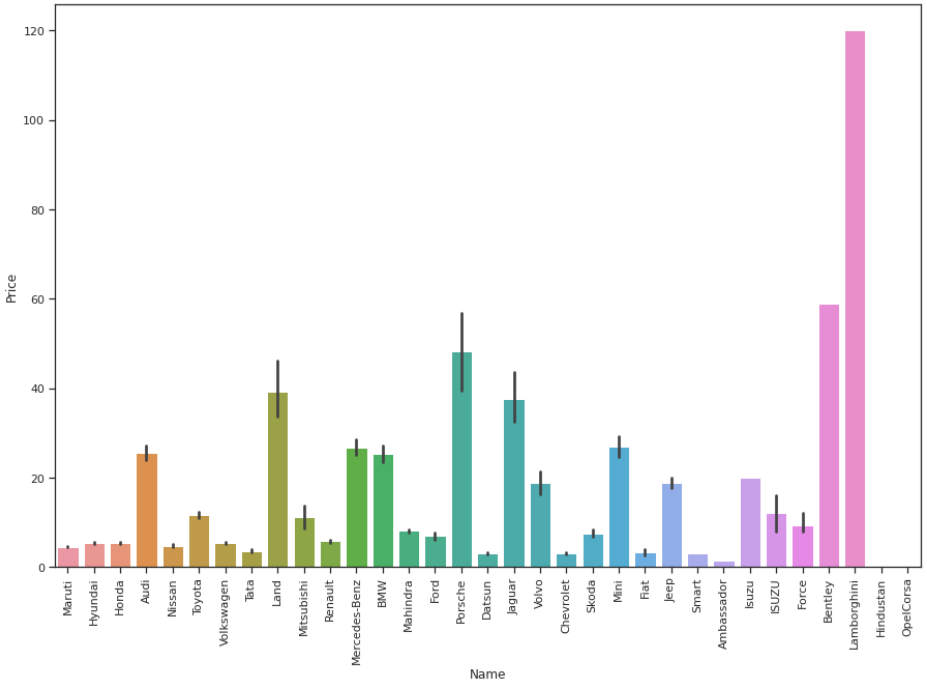
- Electric cars are expensive
- LPG cars are the cheapest



Exploratory Analysis

Price Vs Owner type

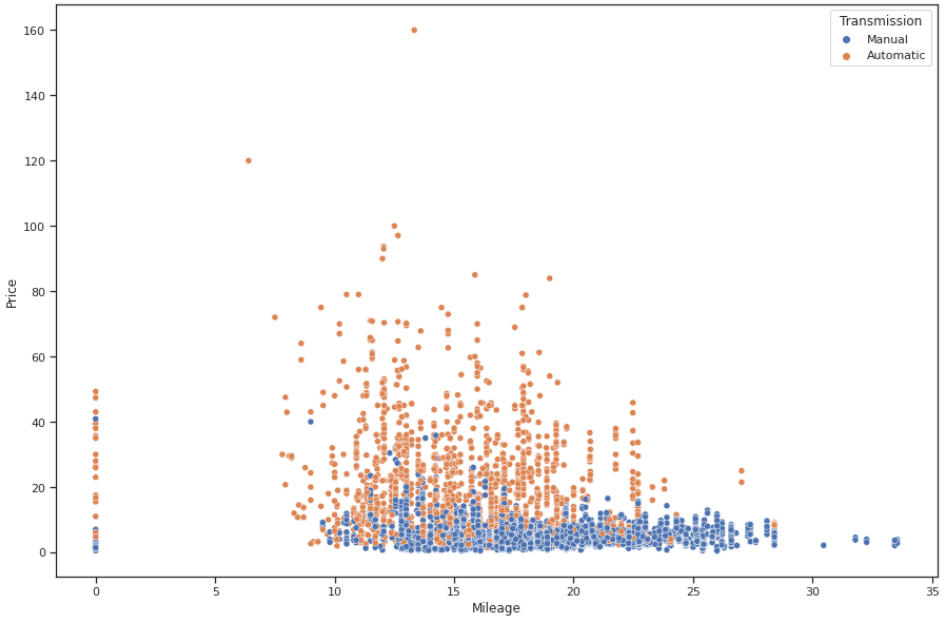
- Lamborghini is the most expensive car
- Ambassador is the cheapest car



Exploratory Analysis

Price Vs Mileage vs Transmission type

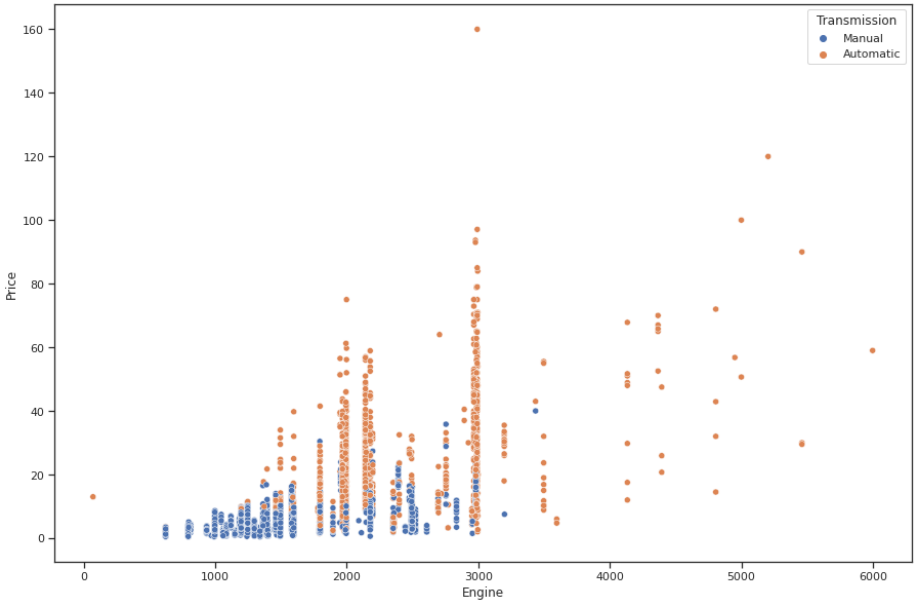
- For both transmission type with increase in mileage price decrease as both are negatively correlated.
- There are few outlier with 0 mileage



Exploratory Analysis

Price Vs Mileage vs Transmission type

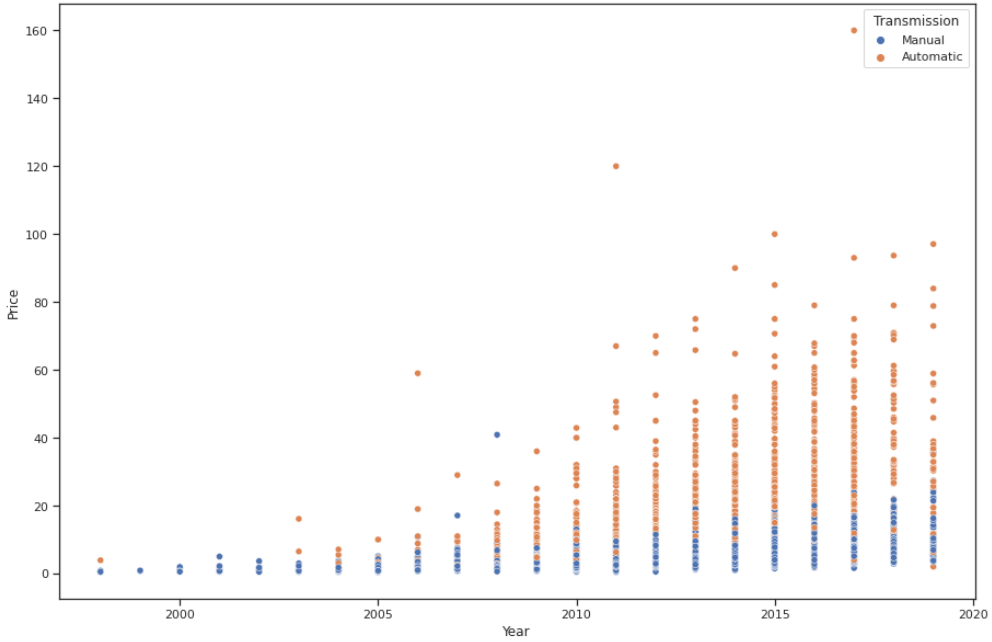
- For both transmission type with increase in Engine price Increase as both are positively correlated.



Exploratory Analysis

Price Vs Year vs Transmission type

- New cars are more expensive
- New automatic cars are more expensive than new manual cars



Fitting Model

- Dataset is splitted in 70/30 ratio 30% of testing 70% for training
- Dataset is fitted to Linear Regression Model
- Linear Regression Model is trained on 70 of training data
- 30% of data will be used in testing model performance

Test data 5 rows

	Year	Kilometers_Driven	Mileage	Engine	Power	Seats	New_Price	Price	Fuel_Type_CNG	Fuel_Type_Diesel	...	Location_Bangalore	Location_Che
0	2010	72000	26.60	998.0	58.16	5.0	NaN	1.75	1	0	...	0	
1	2015	41000	19.67	1582.0	126.20	5.0	NaN	12.50	0	1	...	0	
2	2011	46000	18.20	1199.0	88.70	5.0	8.61	4.50	0	0	...	0	
3	2012	87000	20.77	1248.0	88.76	7.0	NaN	6.00	0	1	...	0	
4	2013	40670	15.20	1968.0	140.80	5.0	NaN	17.74	0	1	...	0	

Predictions 5 rows

```
array([[ 9.96687472],
       [ 4.81140318],
       [ 7.85077468],
       [25.83046767],
       [ 8.50442979]])
```

Fitting Model

Coefficient of Columns

- The coefficients describes the mathematical relationship between each independent variable and the dependent variable.
- The sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable and the dependent variable. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.

```
The coefficient for Year is 0.7428800383982129
The coefficient for Kilometers_Driven is -2.2769705363408702e-05
The coefficient for Mileage is -0.15789869236275253
The coefficient for Engine is 0.0014322465241594397
The coefficient for Power is 0.09461940245125952
The coefficient for Seats is -0.7204056019984076
The coefficient for Fuel_Type_CNG is 0.7101056934618717
The coefficient for Fuel_Type_Diesel is 0.4758226447189013
The coefficient for Fuel_Type_Electric is -5.773159728050814e-14
The coefficient for Fuel_Type_LPG is 0.5437787550477492
The coefficient for Fuel_Type_Petrol is -1.7297070932002478
The coefficient for Transmission_Automatic is 1.053967646241683
The coefficient for Transmission_Manual is -1.0539676462464087
The coefficient for Owner_Type_First is -0.29745927315911347
The coefficient for Owner_Type_Fourth & Above is 0.45708515174929815
The coefficient for Owner_Type_Second is -0.5189963359248759
The coefficient for Owner_Type_Third is 0.35937045733947864
The coefficient for Location_Ahmedabad is -0.6364034600610686
The coefficient for Location_Bangalore is 0.8413990645822691
The coefficient for Location_Chennai is 0.5725719484810128
The coefficient for Location_Coimbatore is 1.401173685984825
The coefficient for Location_Delhi is -0.7034449492117574
The coefficient for Location_Hyderabad is 0.9035813286885747
The coefficient for Location_Jaipur is 0.35476255043029636
The coefficient for Location_Kochi is -0.3472298390862458
The coefficient for Location_Kolkata is -1.463703282651786
The coefficient for Location_Mumbai is -0.9628901174822942
The coefficient for Location_Pune is 0.04018307033495721
```

Fitting Model

Coefficient of Columns

- Column with higher values tends to have more impact predictor
- KM driven, transmission and location are important factor in prediction

```
The coefficient for Year is 0.7428800383982129
The coefficient for Kilometers_Driven is -2.2769705363408702e-05
The coefficient for Mileage is -0.15789869236275253
The coefficient for Engine is 0.0014322465241594397
The coefficient for Power is 0.09461940245125952
The coefficient for Seats is -0.7204056019984076
The coefficient for Fuel_Type_CNG is 0.7101056934618717
The coefficient for Fuel_Type_Diesel is 0.4758226447189013
The coefficient for Fuel_Type_Electric is -5.773159728050814e-14
The coefficient for Fuel_Type_LPG is 0.5437787550477492
The coefficient for Fuel_Type_Petrol is -1.7297070932002478
The coefficient for Transmission_Automatic is 1.053967646241683
The coefficient for Transmission_Manual is -1.0539676462464087
The coefficient for Owner_Type_First is -0.29745927315911347
The coefficient for Owner_Type_Fourth & Above is 0.45708515174929815
The coefficient for Owner_Type_Second is -0.5189963359248759
The coefficient for Owner_Type_Third is 0.35937045733947864
The coefficient for Location_Ahmedabad is -0.6364034600610686
The coefficient for Location_Bangalore is 0.8413990645822691
The coefficient for Location_Chennai is 0.5725719484810128
The coefficient for Location_Coimbatore is 1.401173685984825
The coefficient for Location_Delhi is -0.7034449492117574
The coefficient for Location_Hyderabad is 0.9035813286885747
The coefficient for Location_Jaipur is 0.35476255043029636
The coefficient for Location_Kochi is -0.3472298390862458
The coefficient for Location_Kolkata is -1.463703282651786
The coefficient for Location_Mumbai is -0.9628901174822942
The coefficient for Location_Pune is 0.04018307033495721
```

Fitting Model

Intercept of Model

- In machine learning intercept is called the bias, because it is added to offset all predictions that we make.
- Intercept or bias for the model is negative which means the bias is downward and -1489.9865 needs to subtracted anytime we are performing the linear modeling for this data

```
The intercept for our model is -1489.9865854690045
```

Fitting Model

R^2 for the Model

- This is R^2 values for the model
- It represent model fits with 48.6%

0.4868017479432566

Fitting Model

R^2 for the Model for training data after adding interaction terms

- This is R^2 values for the model with interaction terms
- It represent model fits with 74.6% which is better fit then model without interaction terms.

```
0.7468664922895023
```

Evaluation of Model on test data

Mean absolute error, Median absolute error and R squared

- These are different evaluation parameter.
- Lesser these values better the model
- Our model has Mean absolute error, Median absolute error and R squared value of 4.09, 2.38 and 0.48

MAE	Median AE	R^2
4.059	2.38	0.48

Conclusion

- Highest number of cars belong to Mumbai region with 13.1% and Least number of cars belong to Ahmedabad.
- Most cars are from the year 2014 and 2015 with same % i.e 12.8 and there are very few old cars
- Most common type of Fuel used in car is Diesel Followed by Petrol
- There are least number of cars on Electricity. CNG LPG is also not very common
- Most of the cars here are manual. Less cars are automatic
- Most of the cars have first owner least number of cars have 4th owner
- Outlier in mileage of car are towards far end of both sides
- Outliers in Power and towards far end on the right side. 100 is most power we have
- There are cars with higher power then 100 but are considered outliers
- Engine column has outlier on far right i.e car with more engine value in CC are less in quantity or are very rare
- Most Common Range of for cars are with values between 1000CC to 1500CC
- Commonly Price of cars lies in 5 to 20 lacs. There very few cars with very high price considered outliers on far end of right side
- Commonly New Price of car lies in 20 to 50 lacs. There very few cars with very high price considered outliers on far end of right side
- Price highly positively correlated to engine and power which means that with increase in engine and power price increase
- Price is negative related to Mileage but not strongly related which means increase in mileage will result in decrease prices
- Mileage is also negatively correlated to engine and power which means increase in mileage decreases them
- First Hand cars are more expensive
- Electric Cars are expensive

Conclusion

- LPG cars are the cheapest
- Lamborghini car is the most expensive car
- Ambassador car is the cheapest
- Kilometer driven is right skewed
- Bangalore and Coimbatore has highest priced cars
- Hyderabad has most outliers in terms of price at upper end
- Kolkata Jaipur Chennai has low range when it comes to car prices
- Automatic cars are more pricey
- For both transmission type with increase in mileage price decrease as both are negatively correlated.
- There are few outlier with 0 mileage
- KM driven, transmission and location are important factor in prediction
- Mode with interaction parameter is more well fitted
- For both transmission type with increase in Engine price Increase as both are positively correlated.
- New cars are more expensive
- New automatic care are more expensive than new manual cars
- Our model has Mean absolute error, Median absolute error and R squared value of 4.09, 2.38 and 0.48