USED CAR PRICE PREDICTION

Final Project – 31 Oct 2020

Group 1

Adi Lukmanto

Muhammad Naradi Karim

Yenny Ligninasari

Tutor

Abdullah Ghifari

Overview

In this final project, we will be predicting the price of used cars given the data collected from various sources and distributed across various locations in India.

- Data description
- Data preprocessing
- Exploratory data analysis
- Feature Engineering
- Model
- Conclusion

Price Prediction User

To be able to predict used cars market value can help both buyers and sellers.

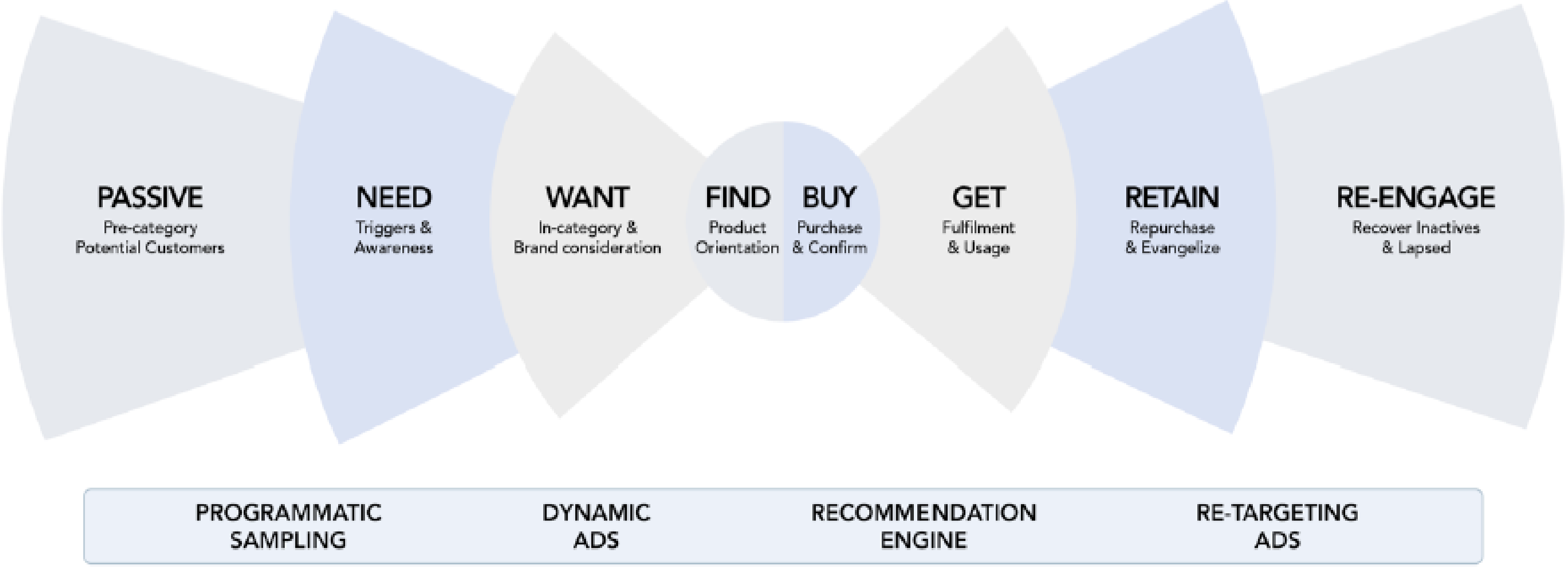
- Used car buyers (dealers)
- Online pricing services
- Individuals (sellers)

Business Understanding

With a nearly endless amount of data — constantly-evolving market trends and consumer demand, to name a few — it's hard to parse what used car dealers should pay attention to and what they shouldn't.

Machine learning contributes to give the fittest feature which influences the customer in purchasing the car which indirectly gives the company or the research market a result in predicting the future sales for cars and boost sales.

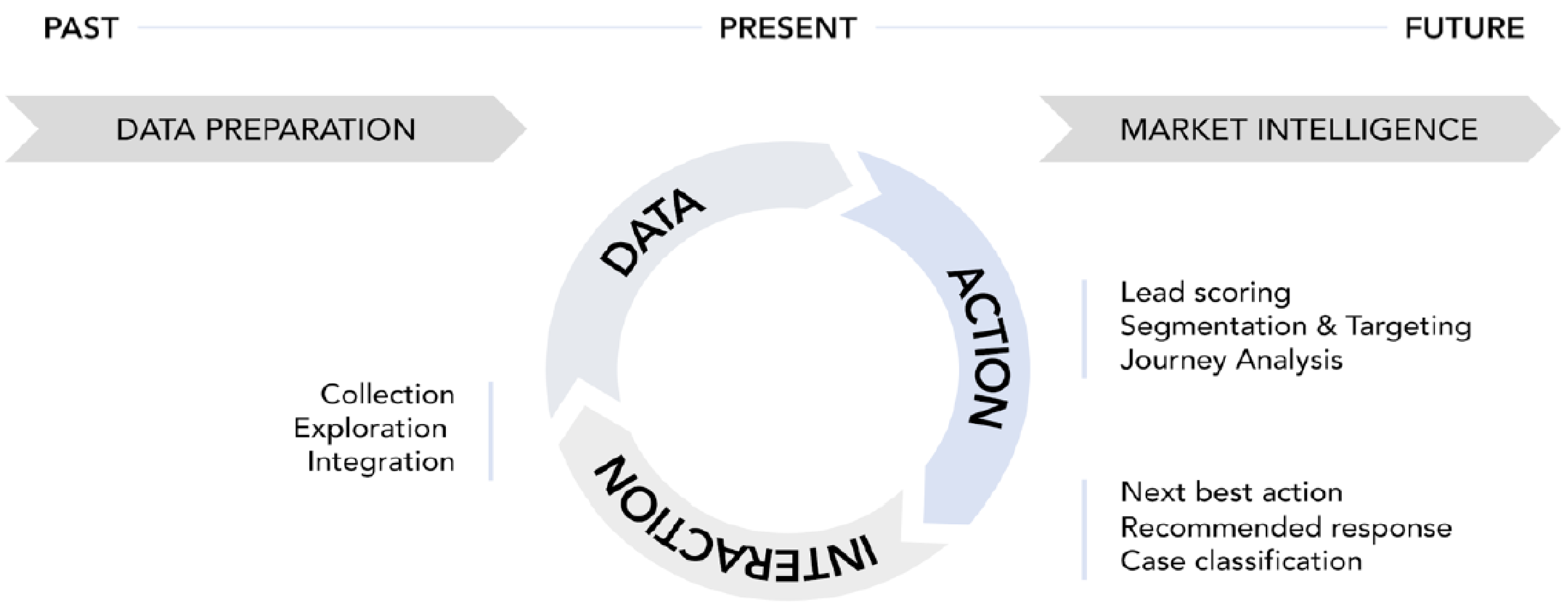
Goal of Machine Learning for Business



Source : Machine Learning consumer journey of Amazon.com
(adapted from Hackermoon, 2018)

Process of Machine Learning for Business

Three stages of the understand-deliver-measure cycle



PAST

PRESENT

FUTURE

DATA PREPARATION

MARKET INTELLIGENCE

Collection
Exploration
Integration

Lead scoring
Segmentation & Targeting
Journey Analysis

Next best action
Recommended response
Case classification

Source : Mari & Rohner 2016

Data Description

Dataset: used_car_data.csv

- Source : https://drive.google.com/folderview?id=1cOxWoIfsFRIMYIdCbKvGp9u0mgsbnkch
- Dataset has 6019 rows and 12 columns
  - 6019 listings
  - 12 columns, with attributes describing different characteristics of the car listings

Features Description

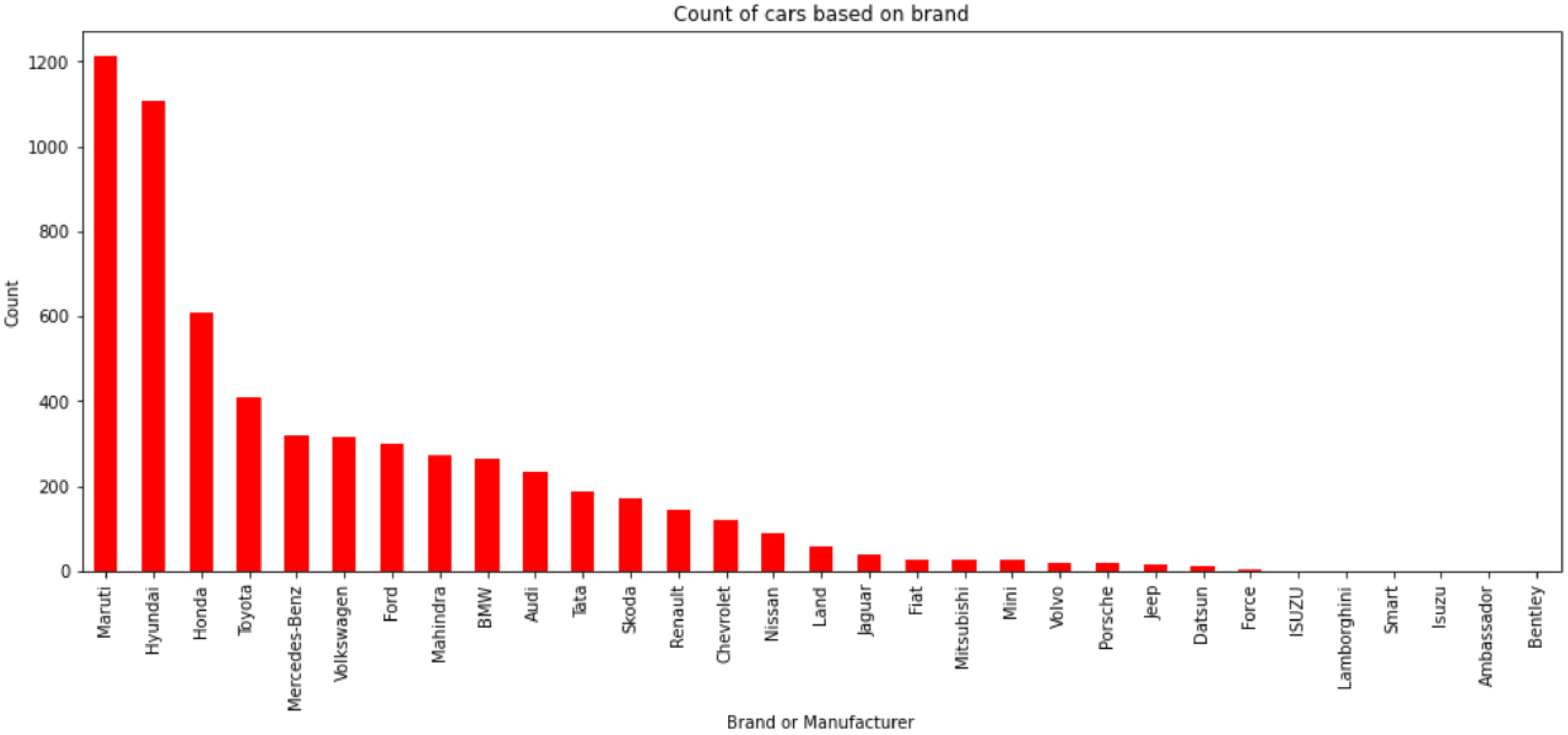
Following features are given in dataset to make the prediction

1. **Name:** The brand and model of the car.
2. **Location:** The location in which the car is being sold or is available for purchase.
3. **Year:** The year or edition of the model.
4. **Kilometers_Driven:** The total kilometres driven in the car by the previous owner(s) in KM.
5. **Fuel_Type:** The type of fuel used by the car.
6. **Transmission:** The type of transmission used by the car.
7. **Owner_Type:** Whether the ownership is Firsthand, Second hand or other.
8. **Mileage:** The standard mileage offered by the car company in kmpl or km/kg
9. **Engine:** The displacement volume of the engine in cc.
10. **Power:** The maximum power of the engine in bhp.
11. **Seats:** The number of seats in the car.
12. **Price:** The price of the used car in INR Lakhs (INR 100,000)

Data Preprocessing

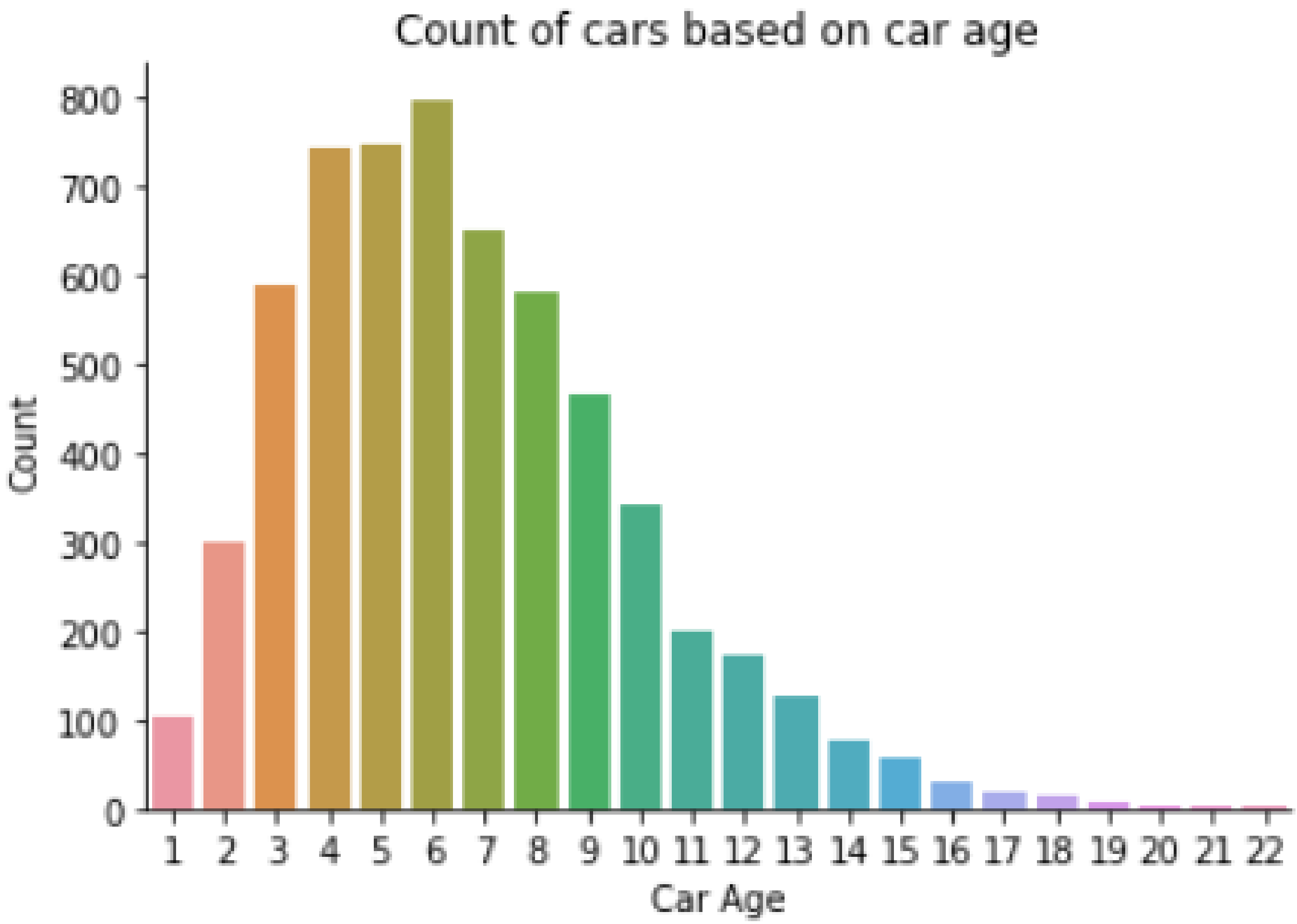
1. Separate ‘Car Brand’ and ‘Model’ names in two separate columns
2. Change ‘Year’ feature to ‘Car Age’
3. Fill missing and null values for features ‘Mileage’, ‘Engine’, ‘Power’, and ‘Seats’
4. Convert string to numeric: feature ‘Mileage’, ‘Engine’, and ‘Power’

Top Car Brand



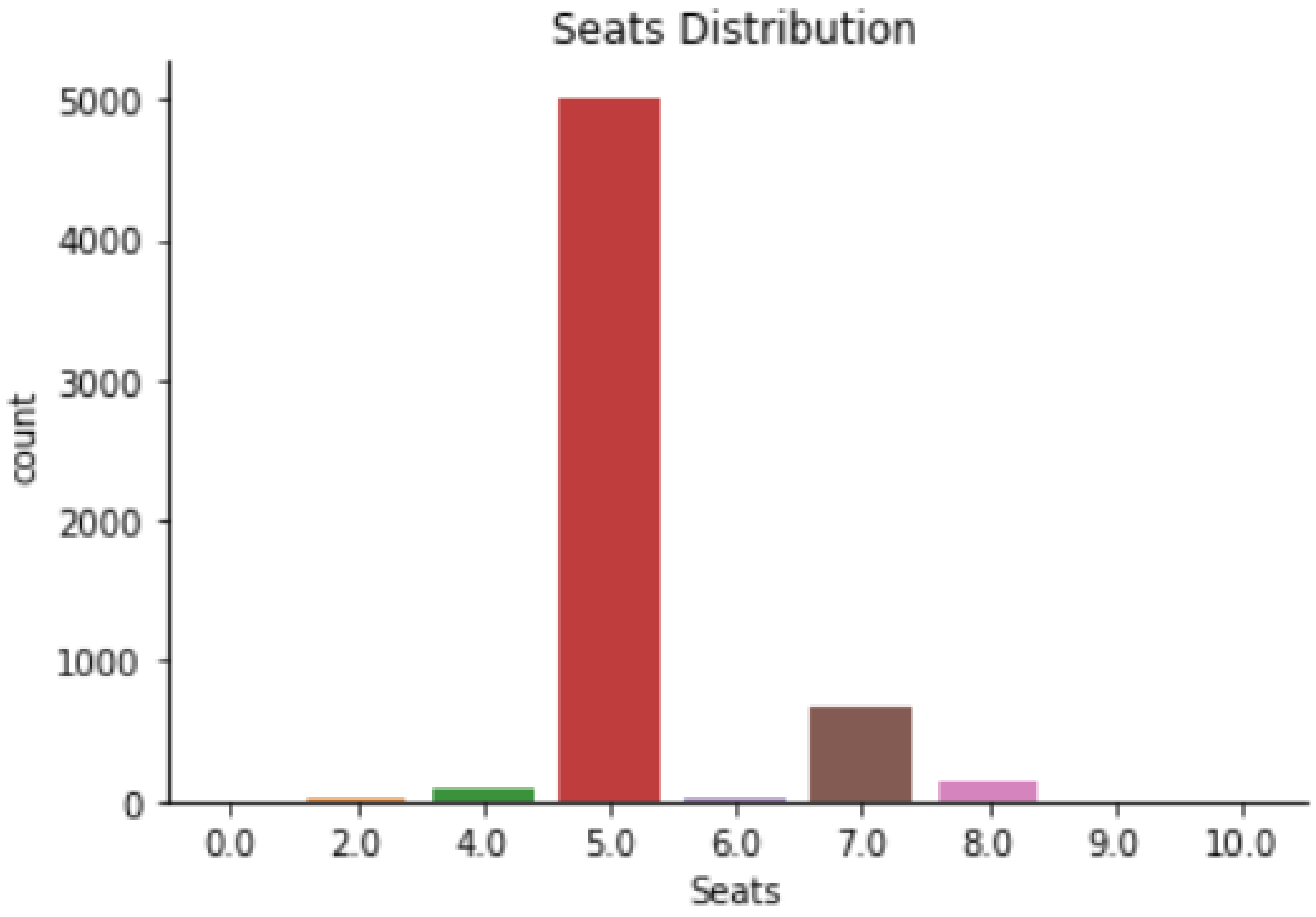
- Maruti is the leading car brand, followed by Hyundai.

Car Age



- Most number of cars in the dataset are built between 2010 to 2017 (age 3-10)

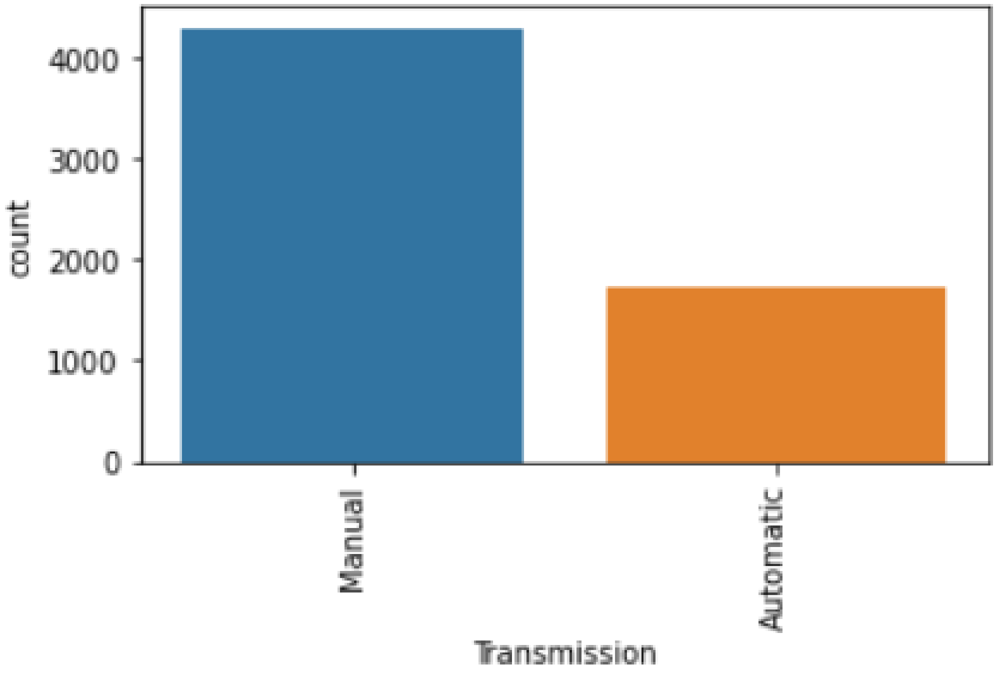
Car Seats

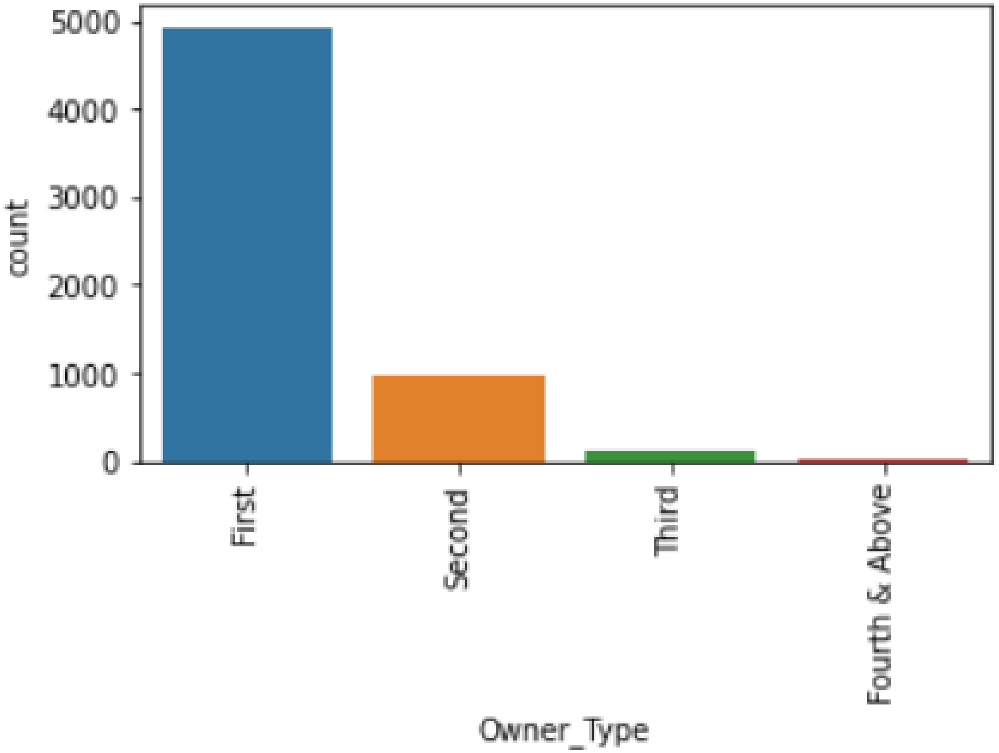


- Most of cars in the listing have 5 Seats

Other features with reference to number of cars

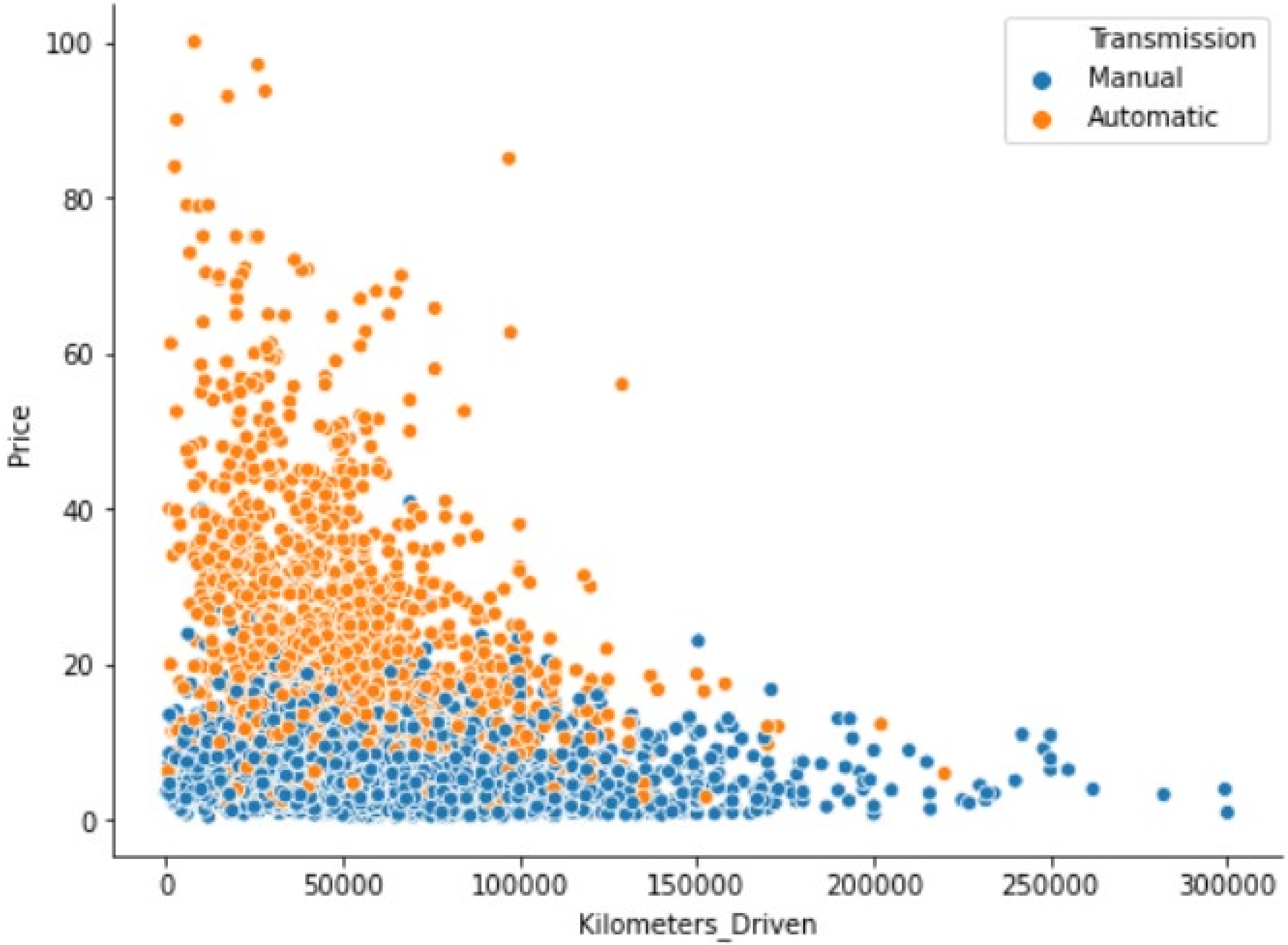
- **Manual cars are listed more than Automatic cars.**
- **Most of the listed cars are from first hand owners.**



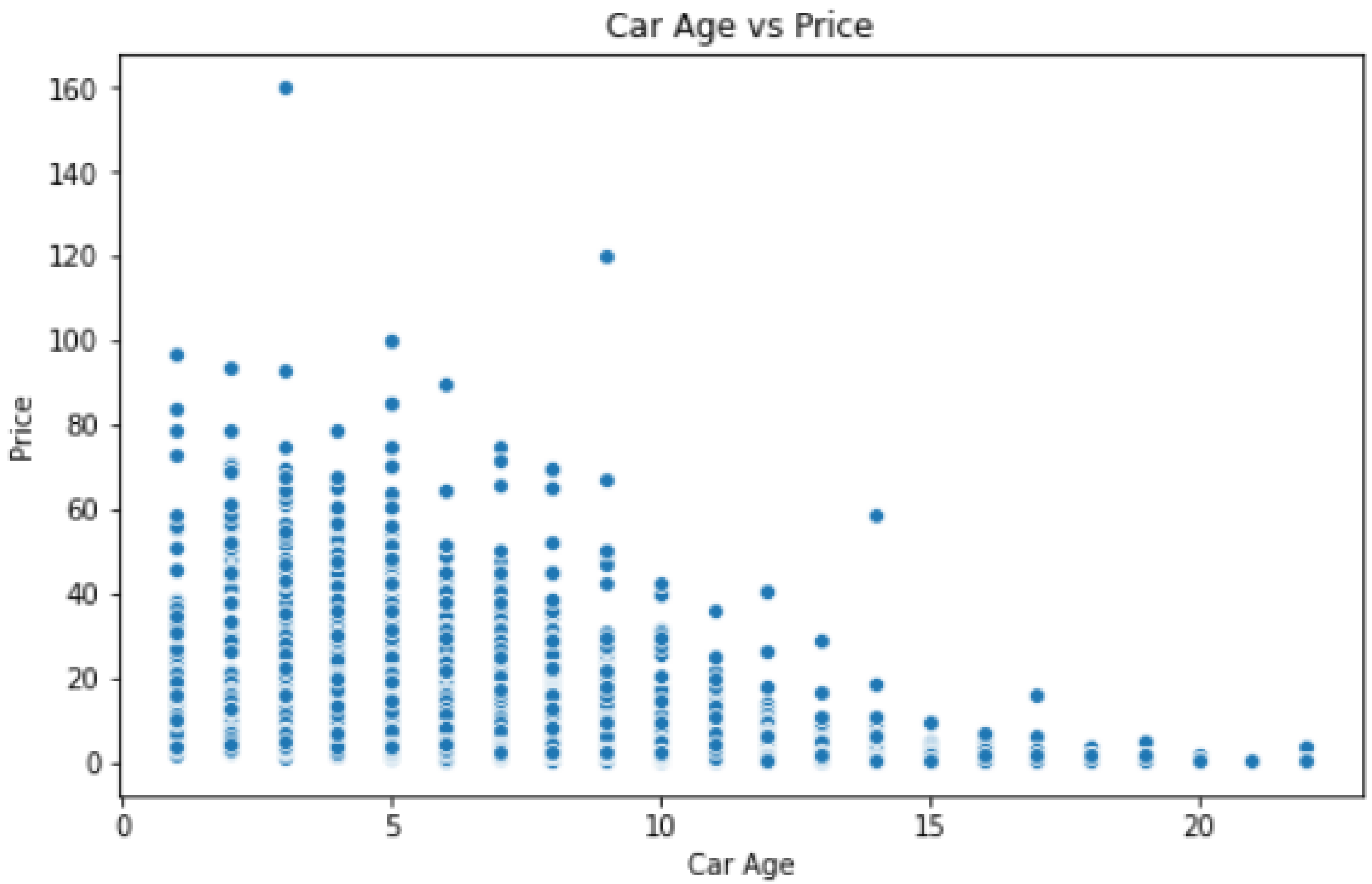


Kilometers_Driven vs Price

- Automatic cars are more expensive than manual cars and cars with less Kilometers_Driven also cost more.

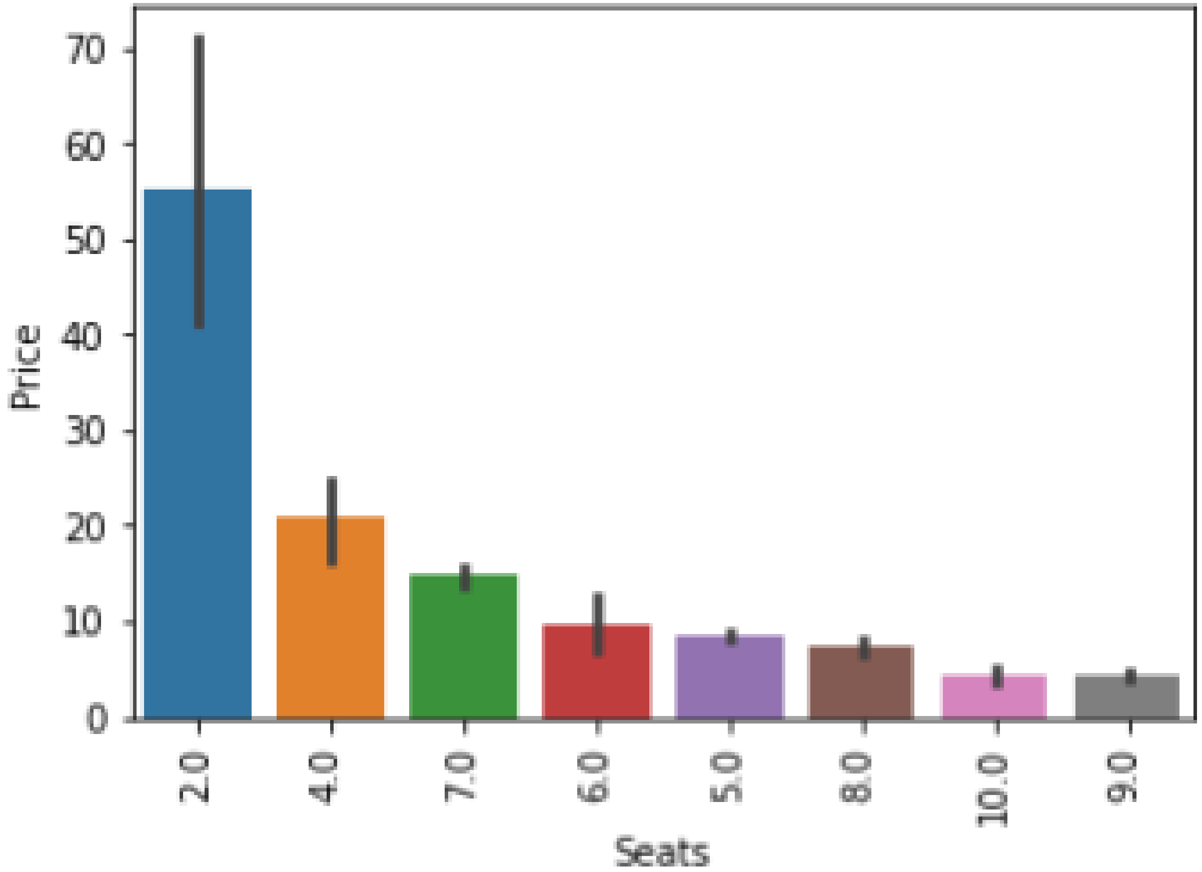


Car Age vs Price



- Cars ranging between the years 2012 to 2019 (age 1 – 8) cost more.

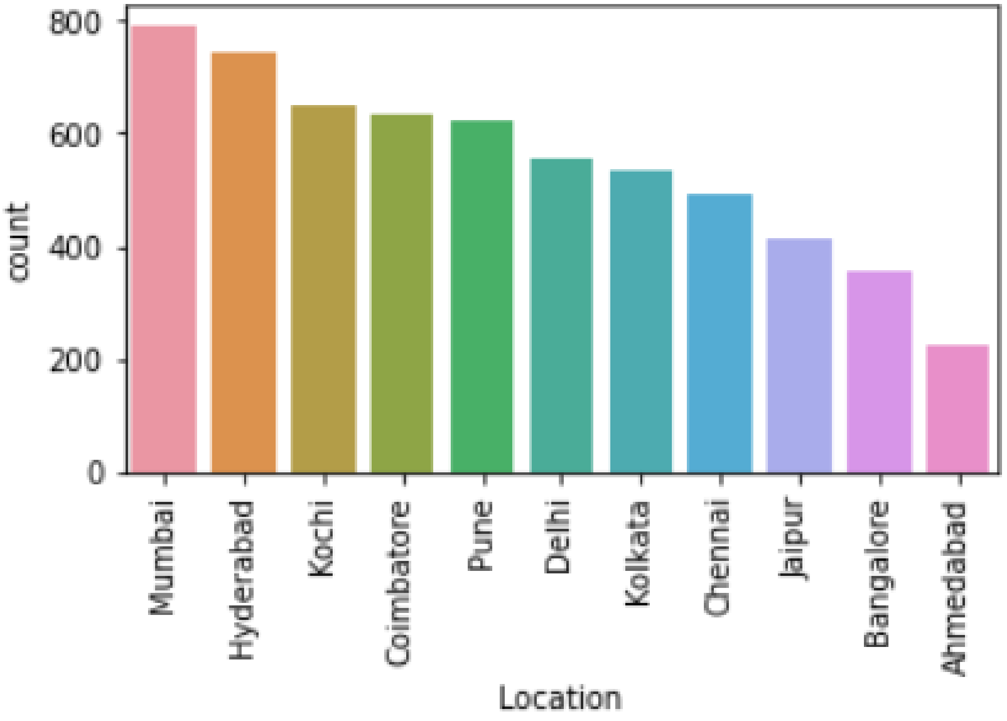
Seats vs Price

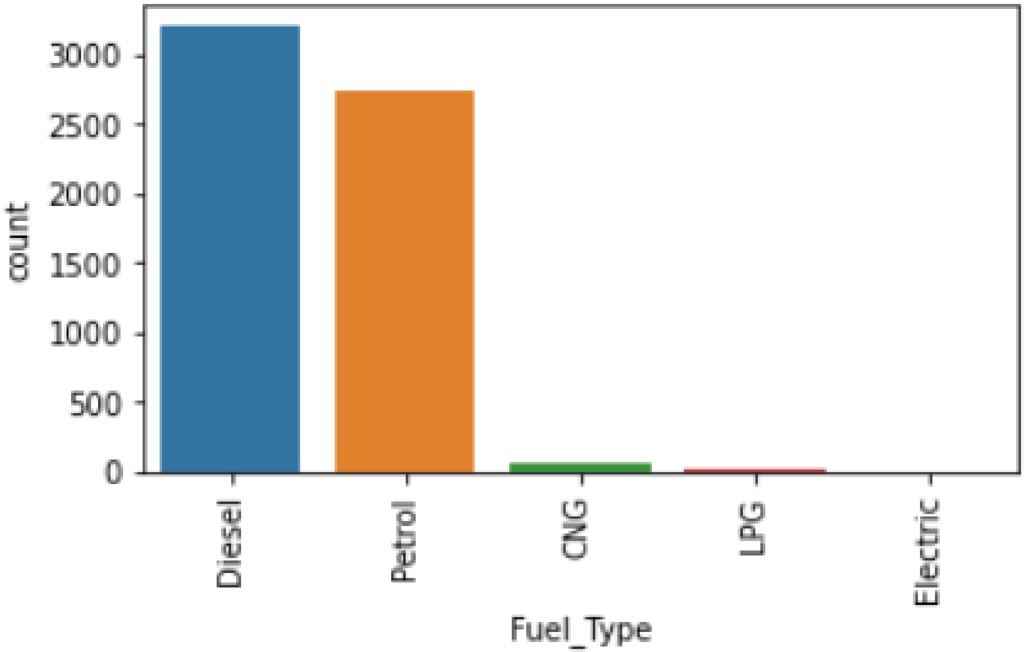


- Two-seater cars are the most expensive in the listing.

Other features with reference to number of cars (2)

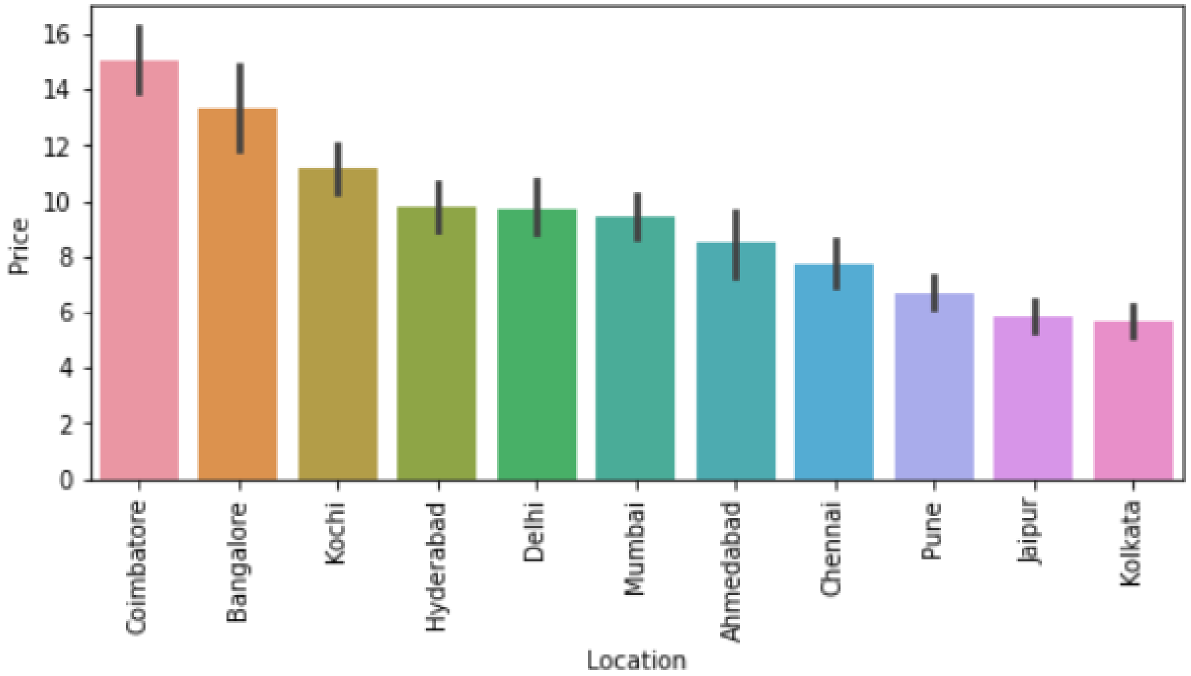
- **Most of the used cars in the listings are in Mumbai, Hyderabad and Kochi**
- **Diesel and petrol are the most listed fuel types**

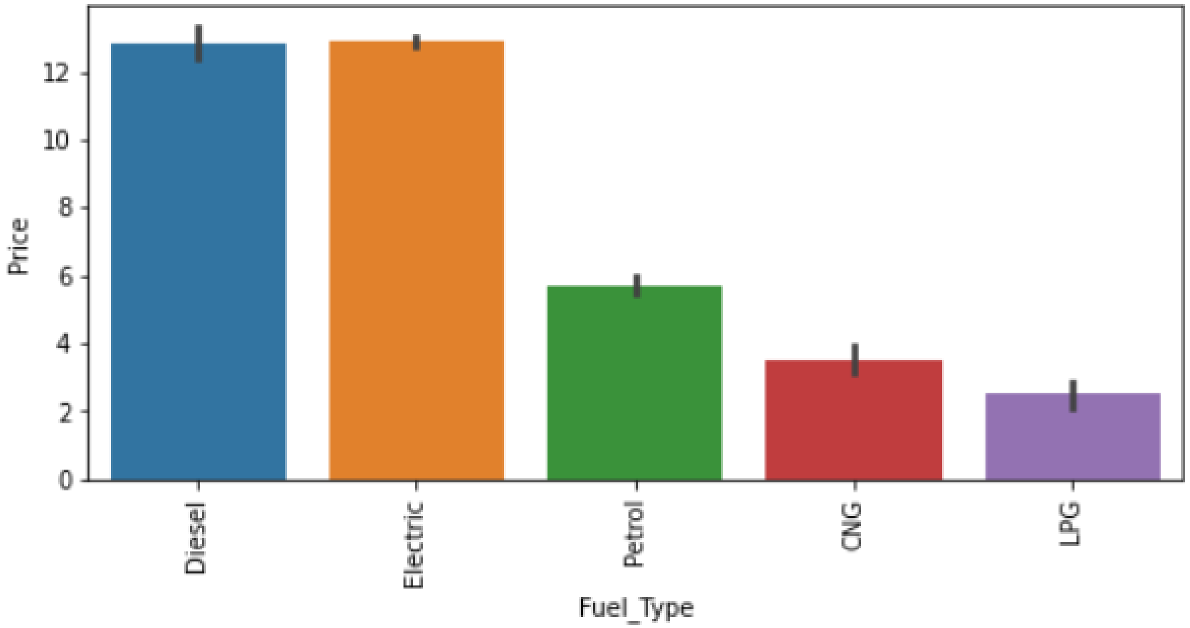




Other features with reference to price

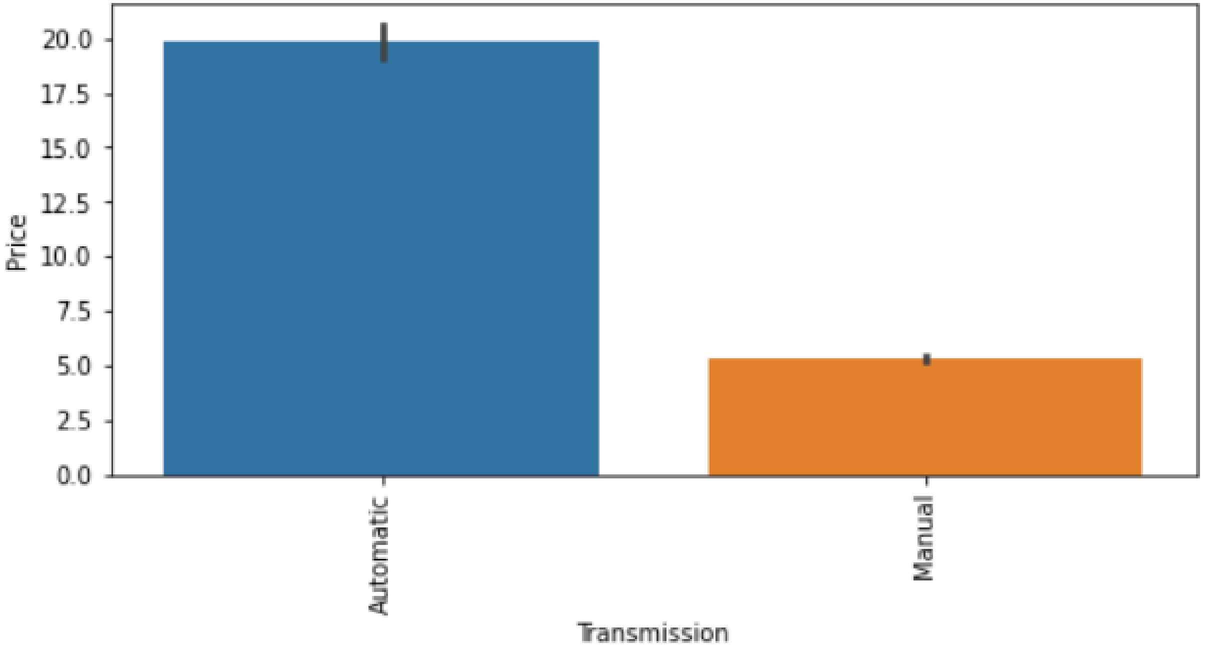
- **Used cars from Coimbatore have higher price than other cars origin in dataset**
- **Diesel and electric cars are more costly.**

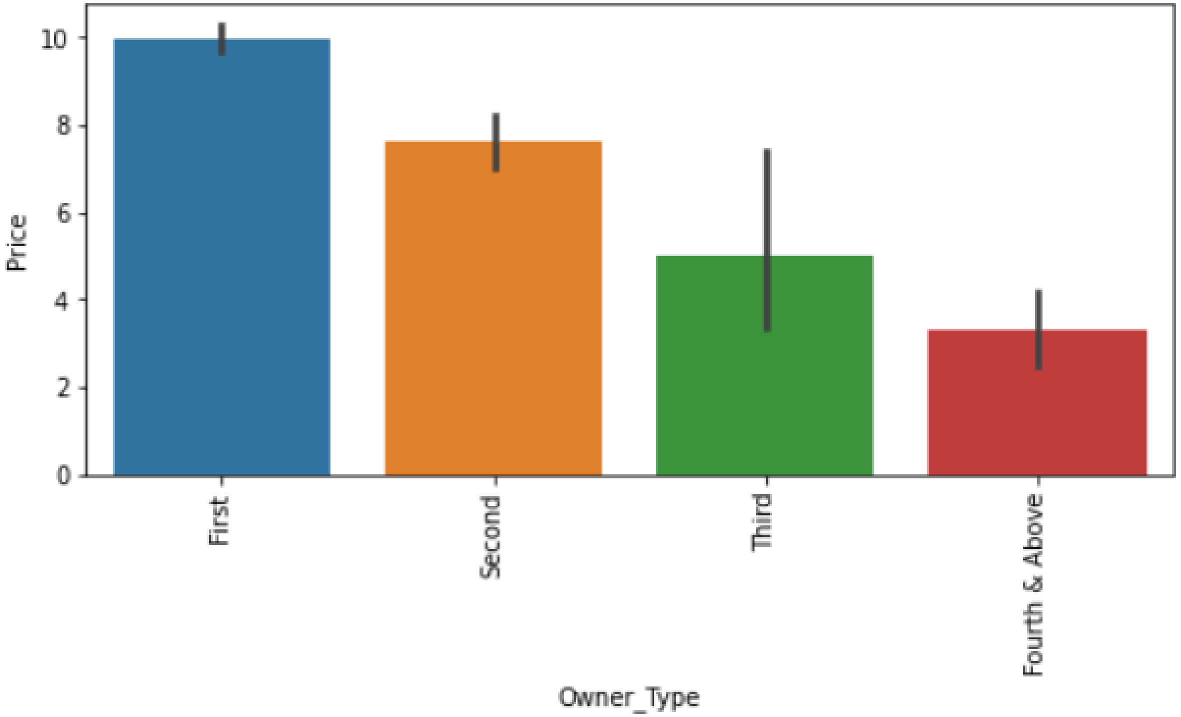




Other features with reference to price (2)

- **Automatic cars more expensive than manual cars.**
- **First-hand cars are the most costly and followed by second-hand cars.**

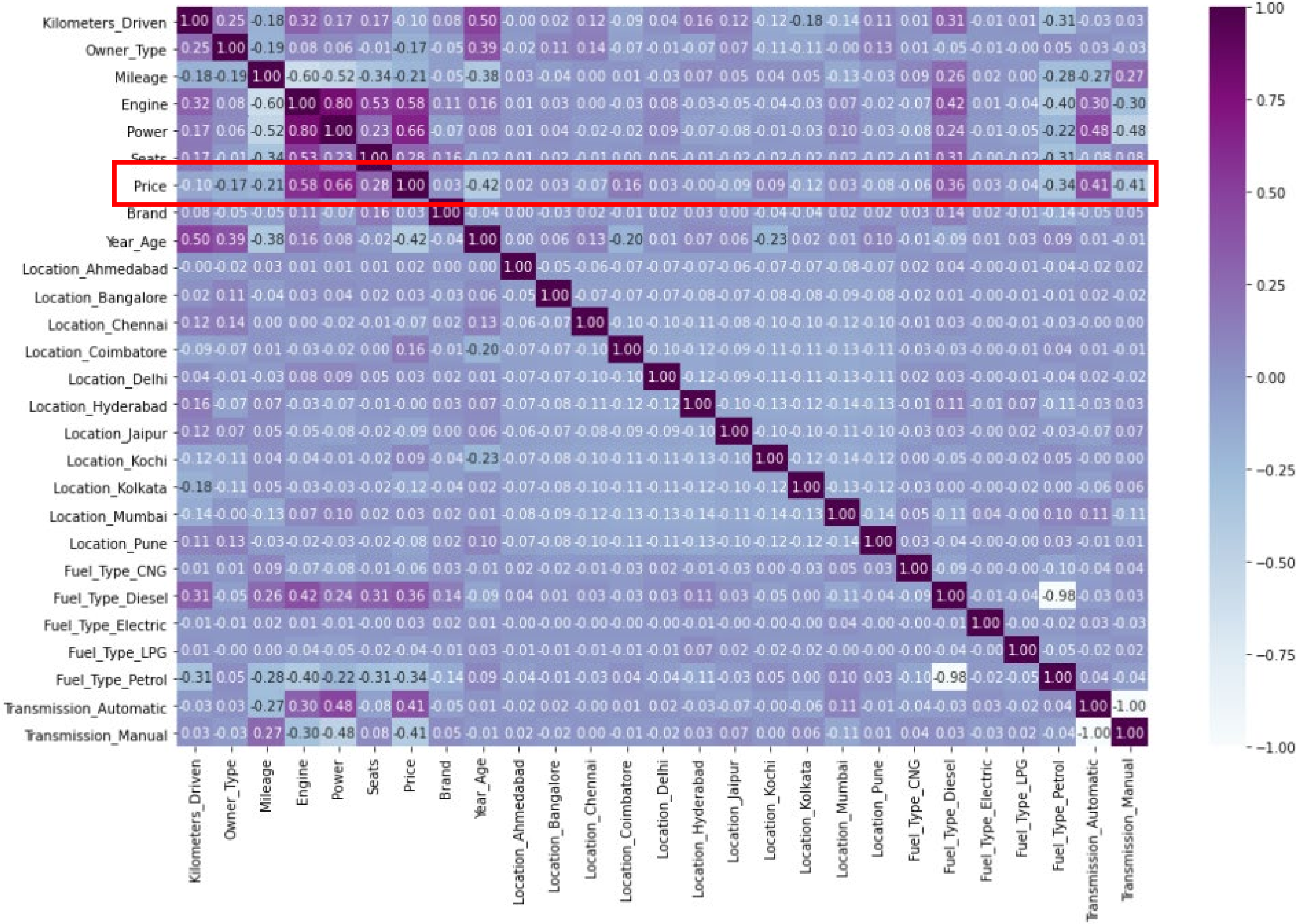




Feature Engineering

1. Remove outliers for features 'Kilometers_Driven','Mileage','Engine','Power’, and 'Price’
2. Delete unnecessary column ‘Model’
3. Transform feature values using StandardScaler package
4. Log transform target ‘Price’
5. Split dataset (70% train and 30% test)
6. Modelling: using scikit-learn tools

Heatmap: Correlation between features and price

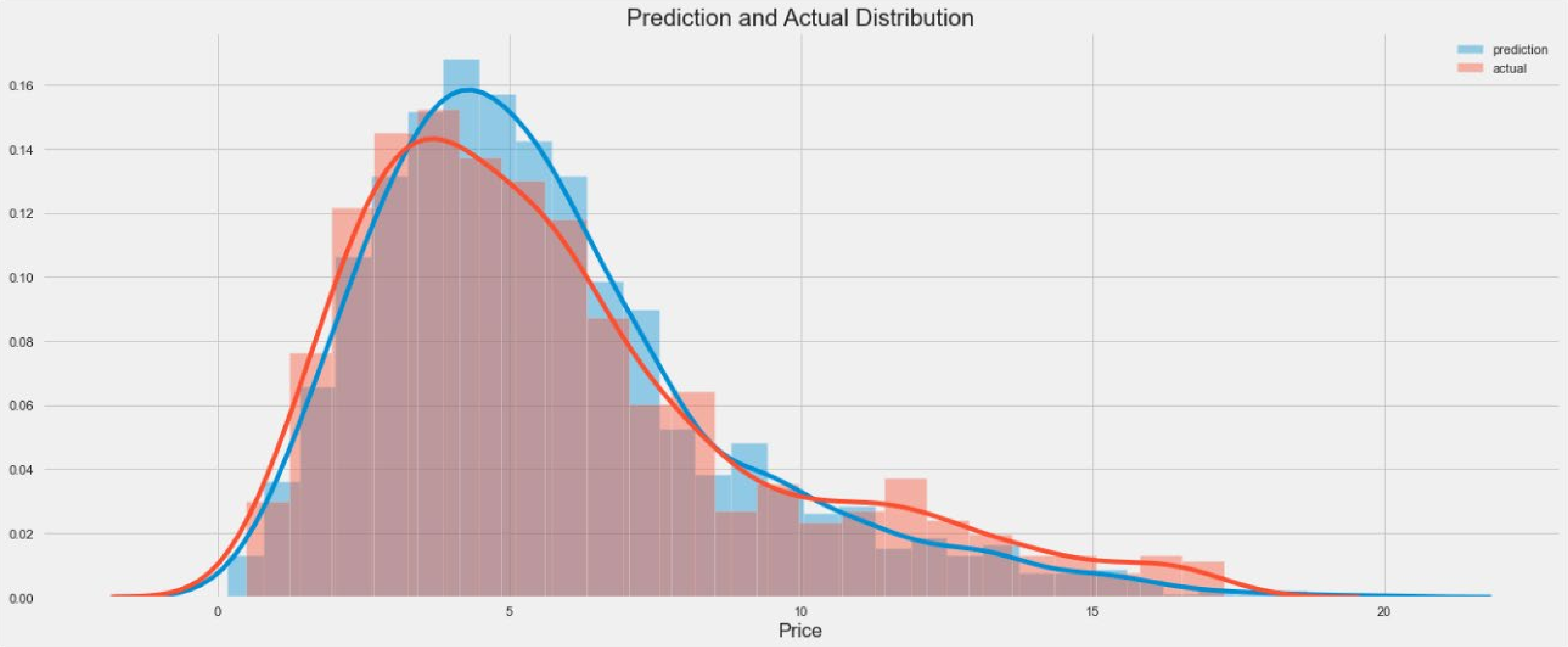


Modelling Summary

Model	MAE	RMSE
Linear Regression	1.082	1.622
Decision Tree Regression	1.05	1.684
Support vector regression	0.845	1.34
Random Forest Regression	0.775	1.191

Note: unit is in INR Lakhs (INR 100,000)

Linear Regression - Prediction on test data



MAE = 1.082 and RMSE = 1.622

Random Forest Regression

Hyperparameter / Random Search Cross Validation

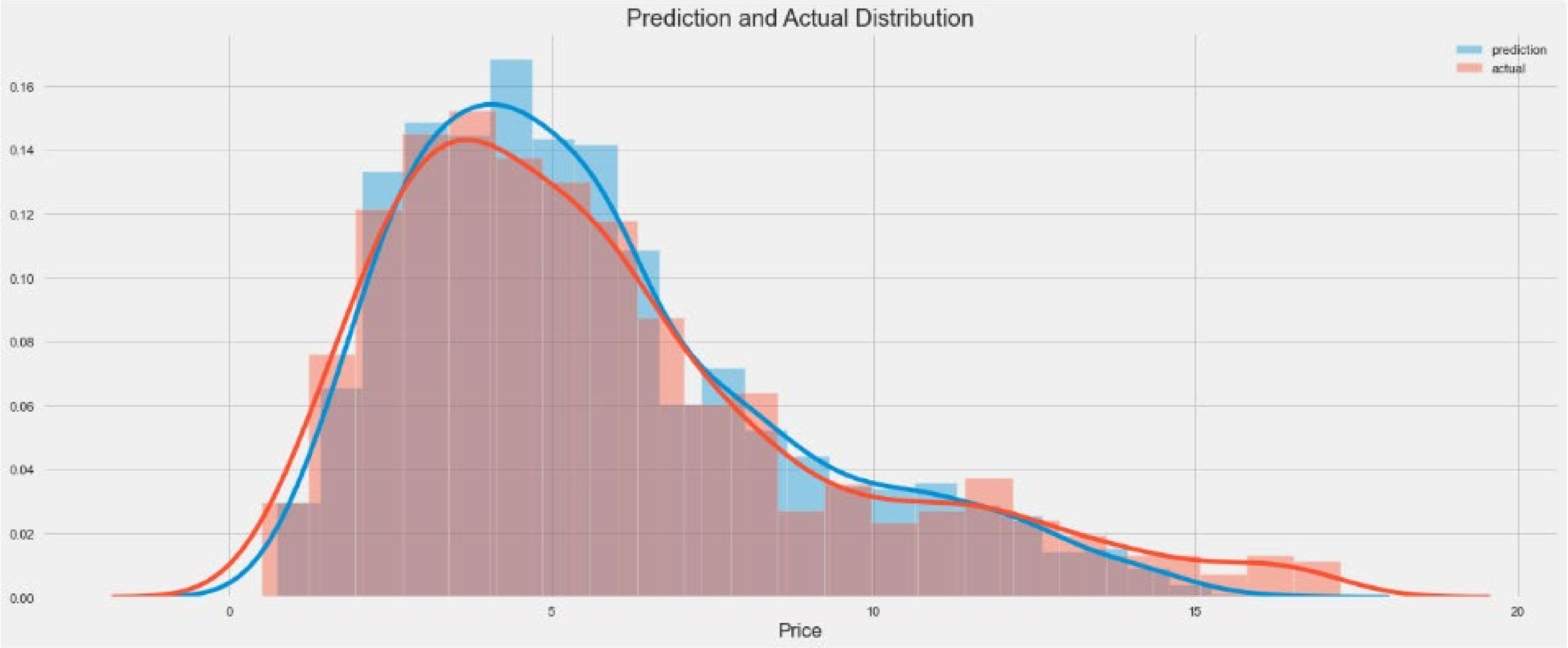
- Using RandomizedSearchCV
- Random Hyperparameter Grid

```
random_grid={'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [20, 40, 50, 100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]}
```

- Random search of parameters, using 5 fold cross validation, search across 100 different combinations, and use all available cores
- Fit the random search model
- Best parameters from fitting the random search:

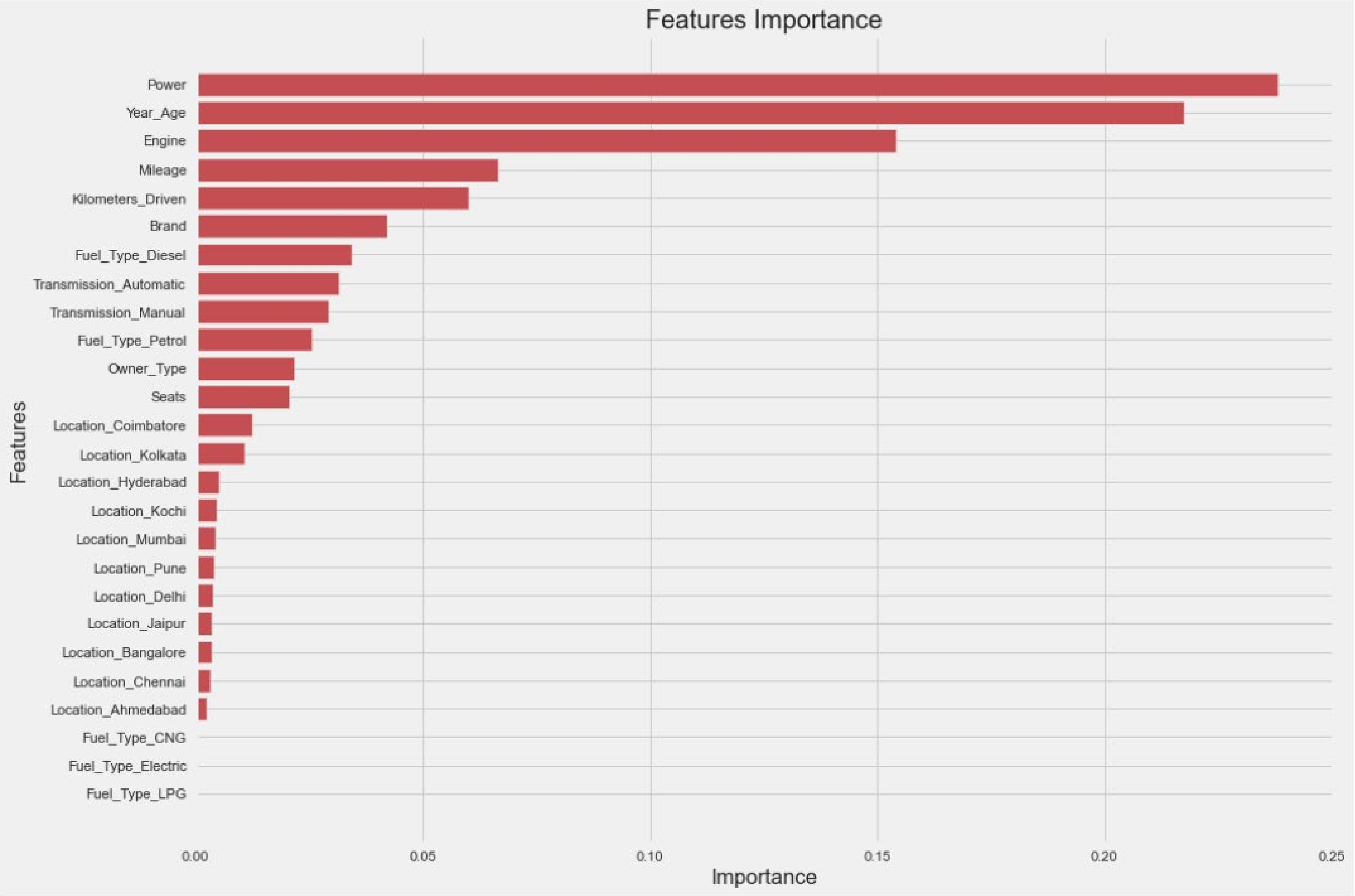
```
{'n_estimators': 800,
 'min_samples_split': 2,
 'min_samples_leaf': 1,
 'max_features': 'sqrt',
   'max_depth': 80}
```

Random Forest - Prediction on test data



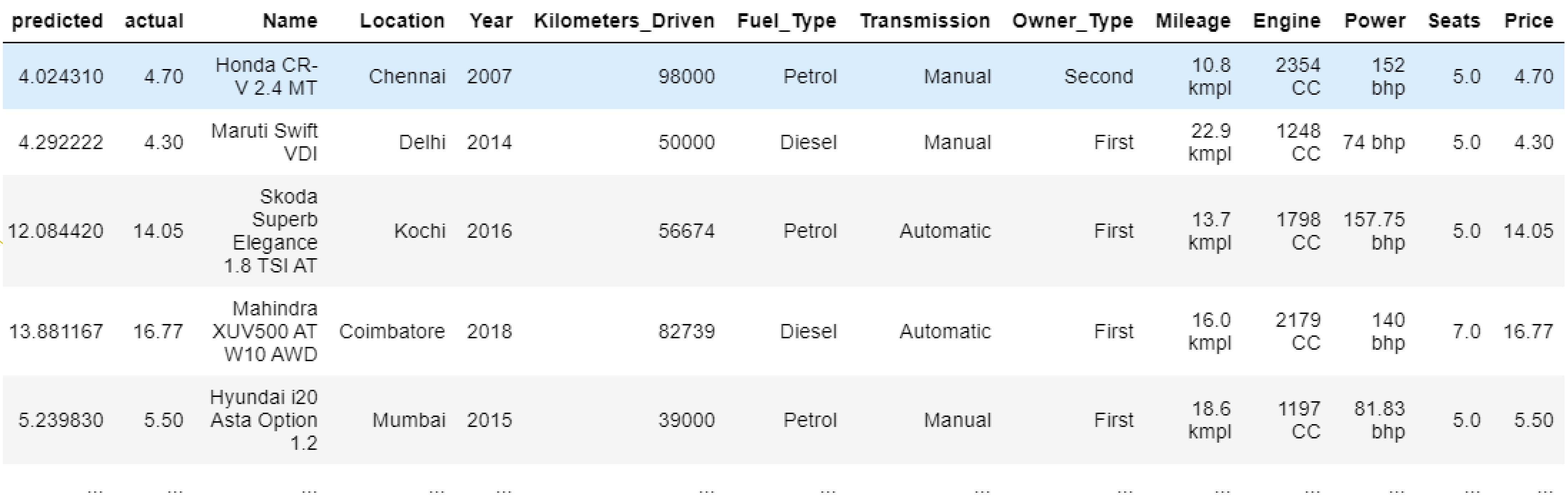
MAE = 0.787 and RMSE = 1.228

Random Forest - Features Importance



Business Insight

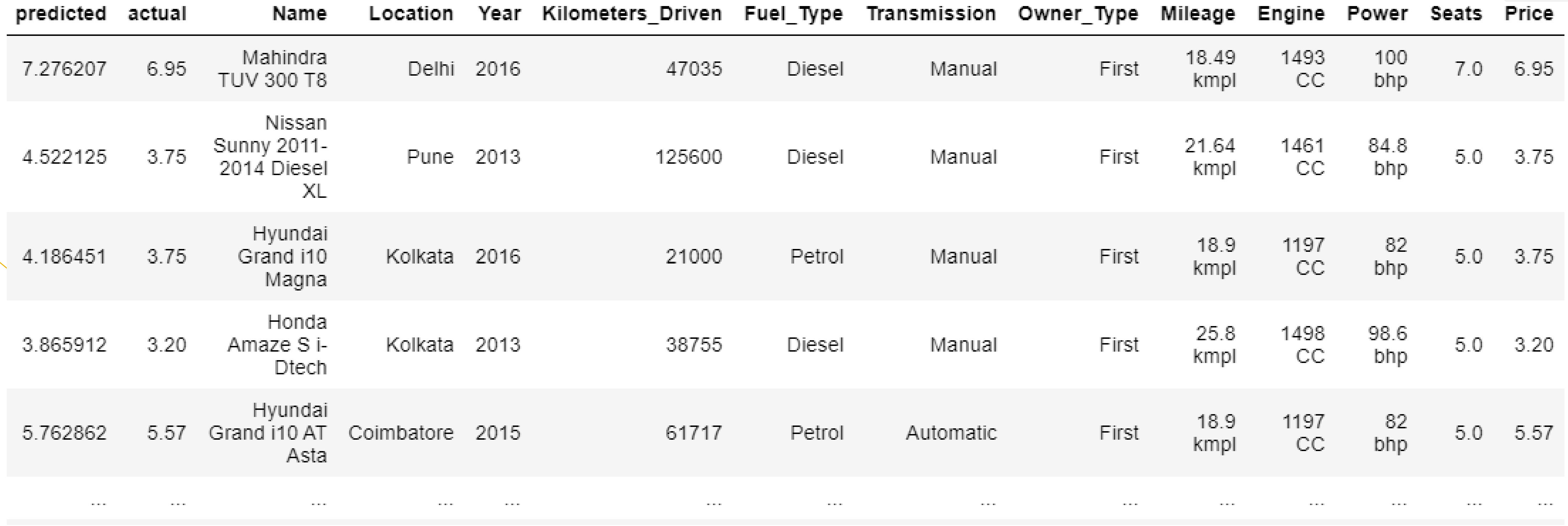
- Examples of overpriced used car listings, based on our price prediction



predicted	actual	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price
4.024310	4.70	Honda CR-V 2.4 MT	Chennai	2007	98000	Petrol	Manual	Second	10.8 kmpl	2354 CC	152 bhp	5.0	4.70
4.292222	4.30	Maruti Swift VDI	Delhi	2014	50000	Diesel	Manual	First	22.9 kmpl	1248 CC	74 bhp	5.0	4.30
12.084420	14.05	Skoda Superb Elegance 1.8 TSI AT	Kochi	2016	56674	Petrol	Automatic	First	13.7 kmpl	1798 CC	157.75 bhp	5.0	14.05
13.881167	16.77	Mahindra XUV500 AT W10 AWD	Coimbatore	2018	82739	Diesel	Automatic	First	16.0 kmpl	2179 CC	140 bhp	7.0	16.77
5.239830	5.50	Hyundai i20 Asta Option 1.2	Mumbai	2015	39000	Petrol	Manual	First	18.6 kmpl	1197 CC	81.83 bhp	5.0	5.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Business Insight (2)

- Examples of underpriced used car listings, based on our price prediction



predicted	actual	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price
7.276207	6.95	Mahindra TUV 300 T8	Delhi	2016	47035	Diesel	Manual	First	18.49 kmpl	1493 CC	100 bhp	7.0	6.95
4.522125	3.75	Nissan Sunny 2011-2014 Diesel XL	Pune	2013	125600	Diesel	Manual	First	21.64 kmpl	1461 CC	84.8 bhp	5.0	3.75
4.186451	3.75	Hyundai Grand i10 Magna	Kolkata	2016	21000	Petrol	Manual	First	18.9 kmpl	1197 CC	82 bhp	5.0	3.75
3.865912	3.20	Honda Amaze S i-Dtech	Kolkata	2013	38755	Diesel	Manual	First	25.8 kmpl	1498 CC	98.6 bhp	5.0	3.20
5.762862	5.57	Hyundai Grand i10 AT Asta	Coimbatore	2015	61717	Petrol	Automatic	First	18.9 kmpl	1197 CC	82 bhp	5.0	5.57
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Conclusion

- Five top factors that predict Price of used cars are :
  - Power
  - Car Age
  - Engine
  - Mileage (fuel consumption)
  - Kilometers Driven

- Limitation of dataset features:
  - In future research, we can collect data to explore other factors that influence the sales/sales period of used vehicles. For example sales days, level of discount from the original price, etc. Incorporating these factors in the analysis can improve to choose non-overage vehicles and have a positive impact on profit.

Thank You

Used Car Price Prediction