FLIP ROBO

CAR PRICE PREDICTION

Submitted by:

Rajashri Sadafule Darveshi

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to the sources Medium, Analytics Vidhya, TowardsDataScience, Data Training Projects and youtube channel’s which helped me to for this project. I would like to thank FlipRobo for giving me this opportunity to develop and accomplish this project.

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. With the change in market due to covid 19 impact, the previous ML models are not performing well.

- Conceptual Background of the Domain Problem

  A good knowledge of after sales market of cars is necessary. What makes a car valuable will be key.

- Review of Literature

  Not a lot of research is available on car prices after covid-19 impact.

  https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/how-consumers-behavior-in-car-buying-and-mobility-changes-amid-covid-19

  https://www.counterpointresearch.com/weekly-updates-covid-19-impact-global-automotive-industry/

- Motivation for the Problem Undertaken

  Due to covid-19 the car market has changed a lot, some cars have shot up in popularity and some gone down in price.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

Inbuilt function such as standardising and log will be used in tackling this problem.

R-square is a comparison of residual sum of squares (SSres) with total sum of squares(SStot). Total sum of squares is calculated by summation of squares of perpendicular distance between data points and the average line.

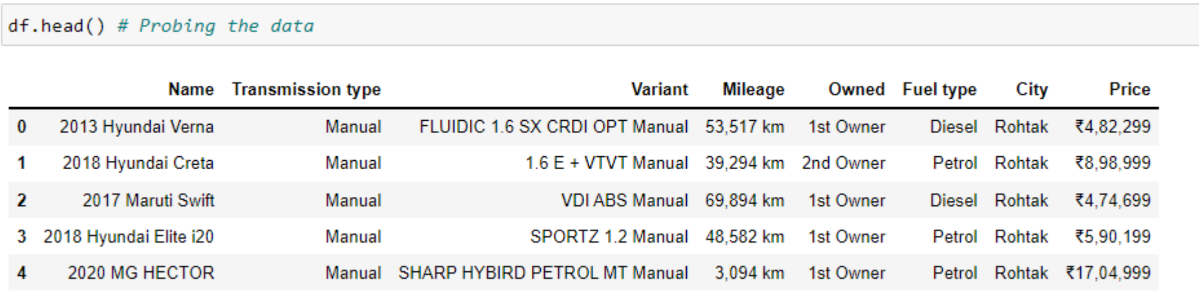
Where SSres is the residual sum of squares and SStot is the total sum of squares.

R-square is the main metric which I will use in this regression analysis.

Concordance index was also used. The concordance index or c-index is a metric to evaluate the predictions made by an algorithm. It is defined as the proportion of concordant pairs divided by the total number of possible evaluation pairs.

- Data Sources and their formats

The data was scraped from cars24 website; data was scraped for more than 20 cities where prices differ.



```
df.head() # Probing the data
```

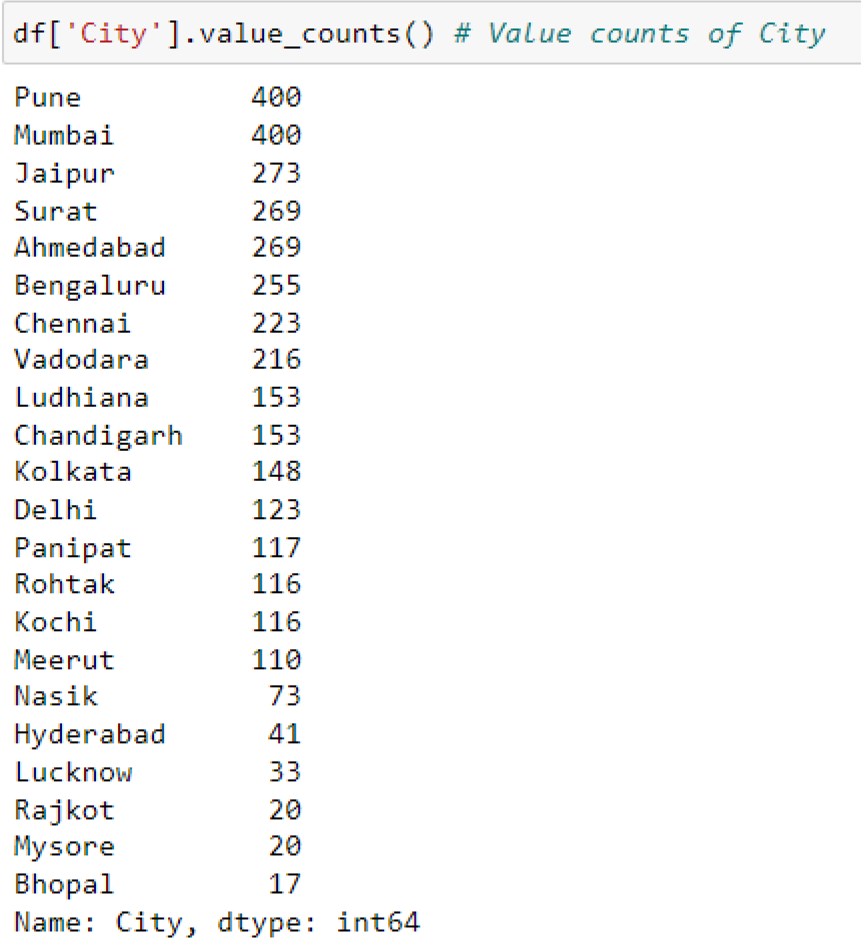
	Name	Transmission type	Variant	Mileage	Owned	Fuel type	City	Price
0	2013 Hyundai Verna	Manual	FLUIDIC 1.6 SX CRDI OPT Manual	53,517 km	1st Owner	Diesel	Rohtak	₹4,82,299
1	2018 Hyundai Creta	Manual	1.6 E + VTVT Manual	39,294 km	2nd Owner	Petrol	Rohtak	₹8,98,999
2	2017 Maruti Swift	Manual	VDI ABS Manual	69,894 km	1st Owner	Diesel	Rohtak	₹4,74,699
3	2018 Hyundai Elite i20	Manual	SPORTZ 1.2 Manual	48,582 km	1st Owner	Petrol	Rohtak	₹5,90,199
4	2020 MG HECTOR	Manual	SHARP HYBIRD PETROL MT Manual	3,094 km	1st Owner	Petrol	Rohtak	₹17,04,999

- Data Preprocessing Done

The years were extracted from the name of the car which contained lot of information.

Numerical variables were converted to integer type (form string) so I could perform deeper analysis on them.

Engine variants were classified under ranges; for example, engines were classified as 1.0 – 1.5 litre capacity; they were many more such ranges.



```
df['City'].value_counts() # Value counts of City
```

```
Pune          400
Mumbai        400
Jaipur        273
Surat         269
Ahmedabad     269
Bengaluru     255
Chennai       223
Vadodara      216
Ludhiana      153
Chandigarh    153
Kolkata       148
Delhi         123
Panipat       117
Rohtak        116
Kochi         116
Meerut        110
Nasik          73
Hyderabad      41
Lucknow        33
Rajkot         20
Mysore         20
Bhopal         17
Name: City, dtype: int64
```

- State the set of assumptions (if any) related to the problem under consideration

The main assumption is that there is no selection bias in the data which we have.

This is because we have cars from varying years and varying city; each city doesn’t have equal amount of data.

Here we can see the count of data per city.

- Hardware and Software Requirements and Tools Used

Pandas, Seaborn, ploty and sickit libraries were used throughout the project.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Regression and co relation.

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables

- Testing of Identified Approaches (Algorithms)

Decision tree regression

Random forest regression

Support vector regression

Natural log, and min-max scaling and finally hyper parameter tuning

- Run and Evaluate selected models

  All R-square values are from cross-validation of 4 samples

  Random forest regression: r-square = 0.93

  Decision tree regression: r-square = 0.90

  Support vector regression: r-square = 0.85

  And as per this data random forest was chosen as the best model; further hyper parameter tuning was performed.

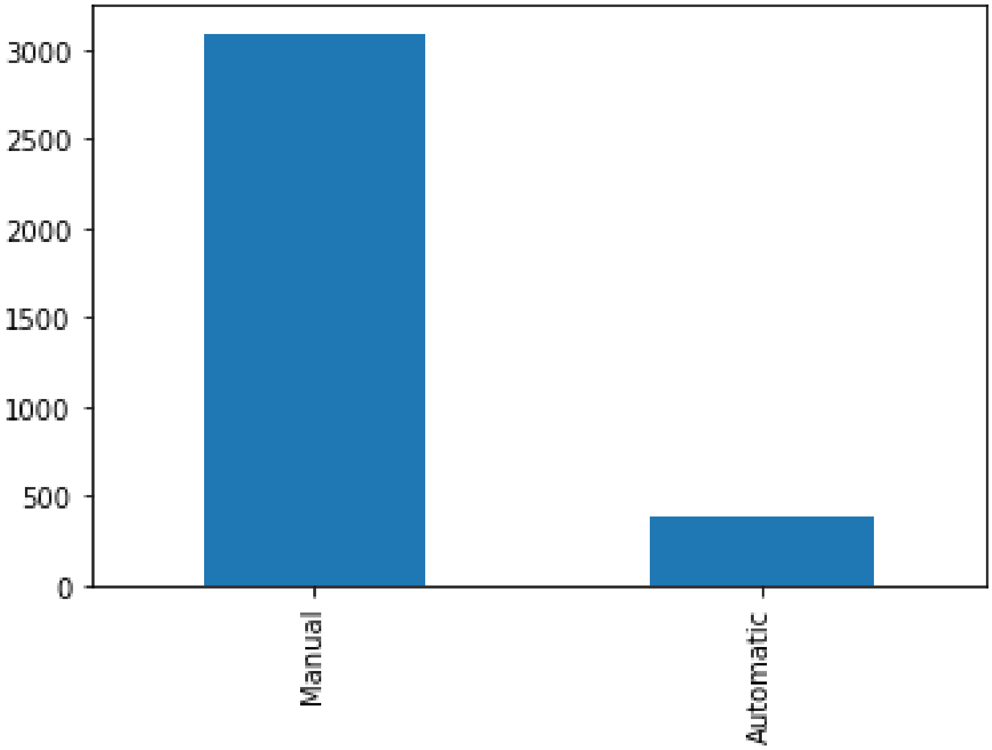
  Final model: r-squared = 0.94254259734771

  Which is an improvement from all previous attempts

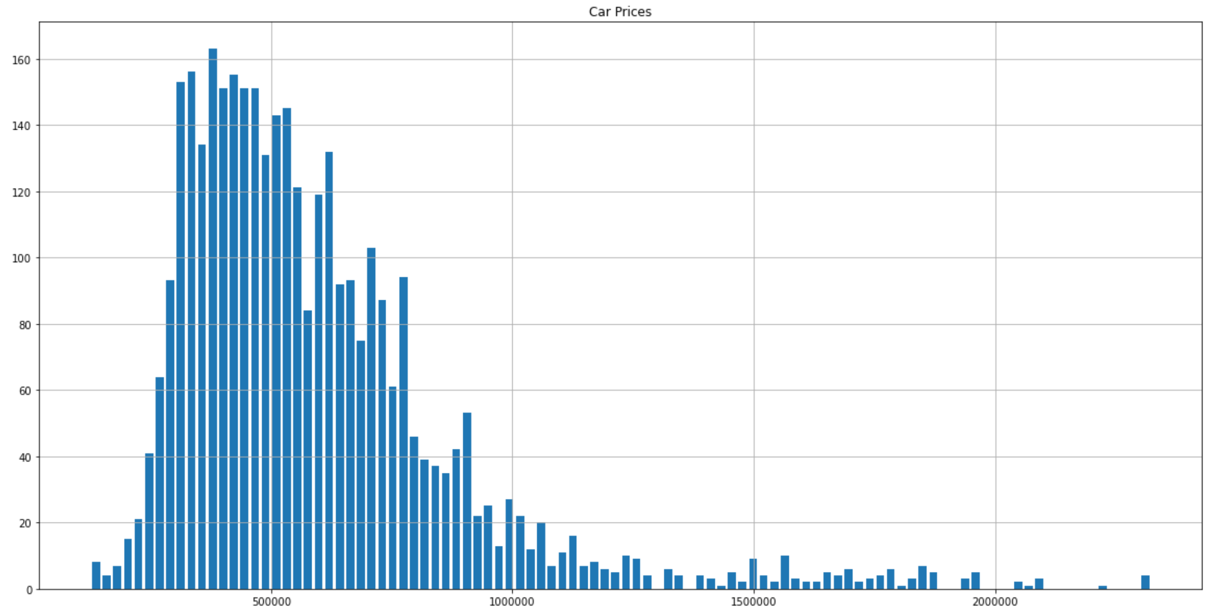
- Key Metrics for success in solving problem under consideration

  R-square was used to determine the success if an algorithm performed well or not.

- Visualizations

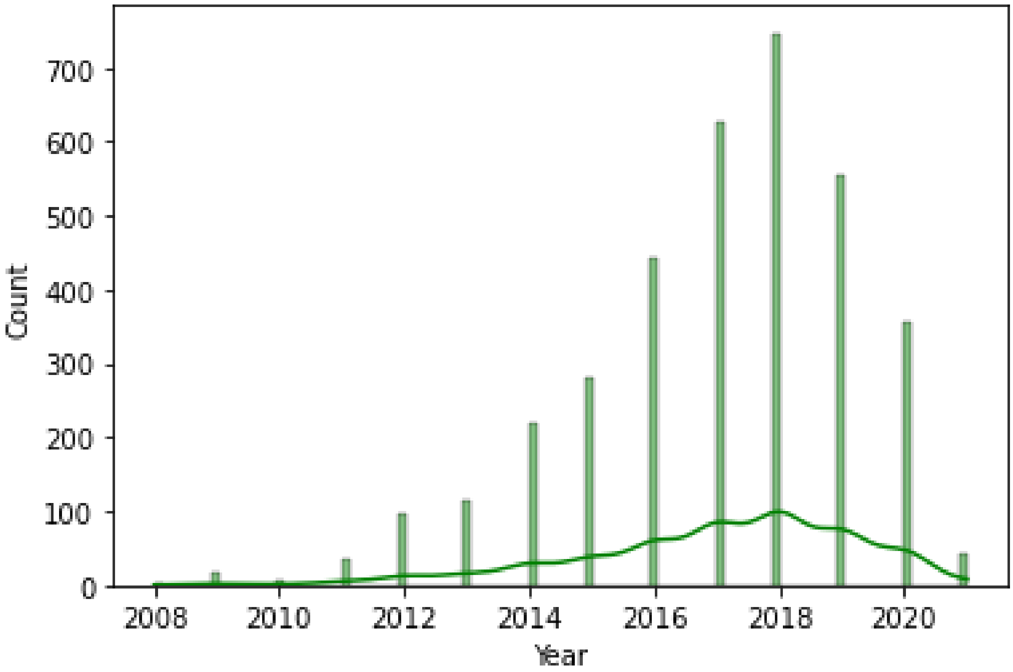


Plot of count of transmission type, automatic vs manual; we can observe majority of cars are Manual.

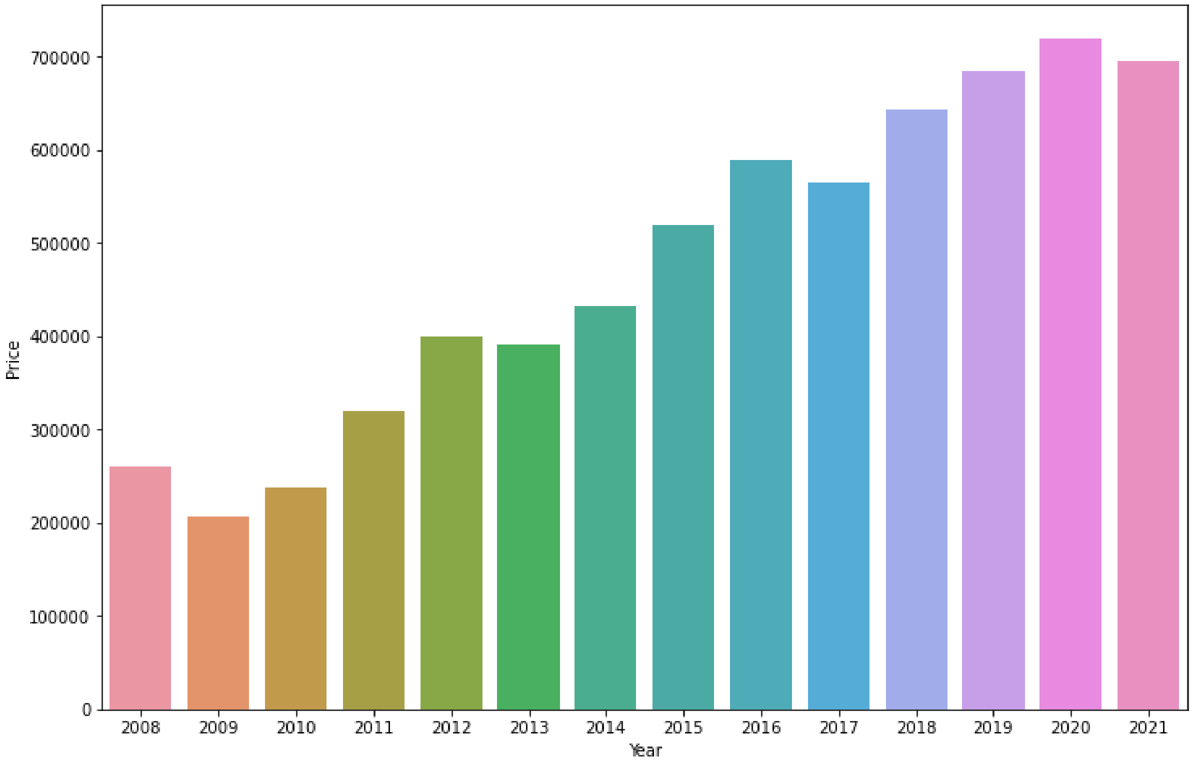


This is a histogram plot of price distribution of the cars.

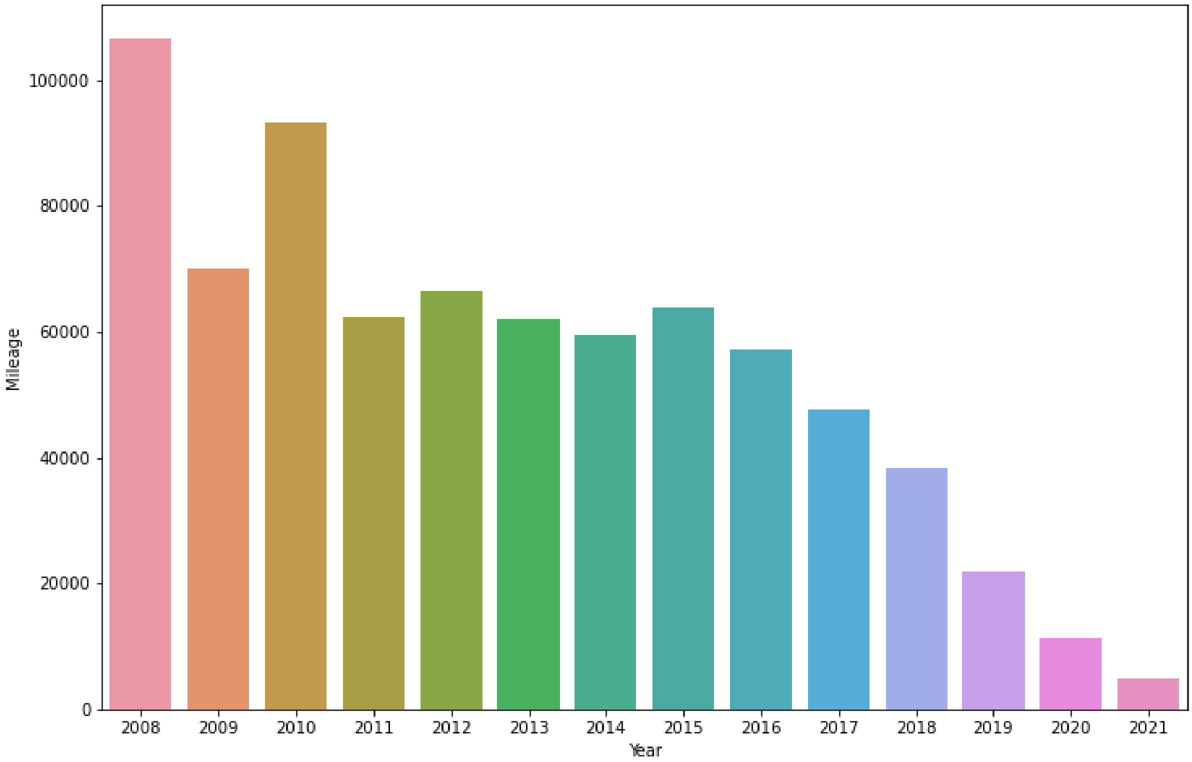
We can see that most of the cars are prices below Rs. 30,000,000. The average price of resale car is around Rs.5,000,000



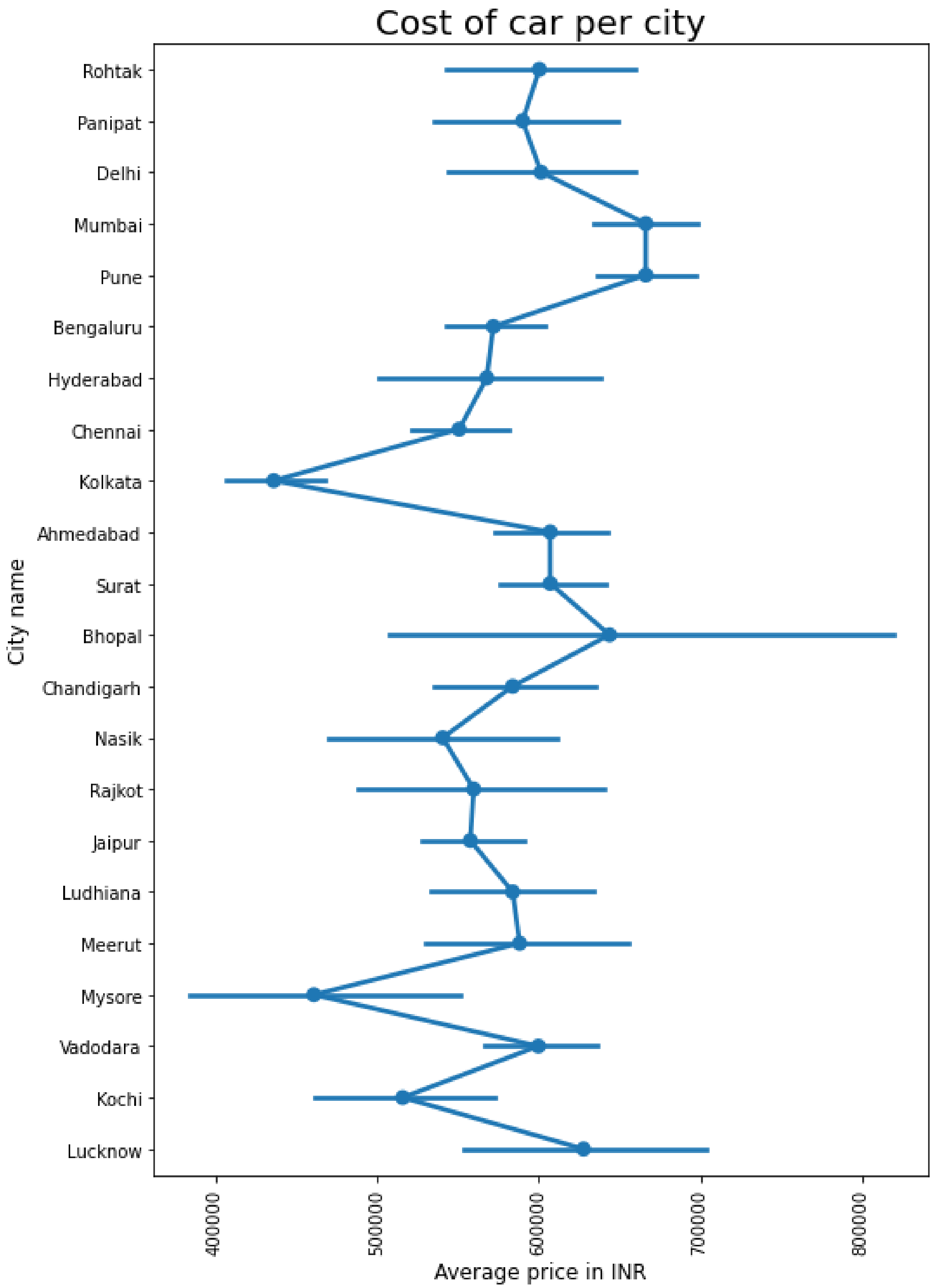
This is the plot of number of cars based on which year the car was manufactured. We can observe that bulk of cars is sold from 2015-2020. So cars ~6-2 years old are sold the most.



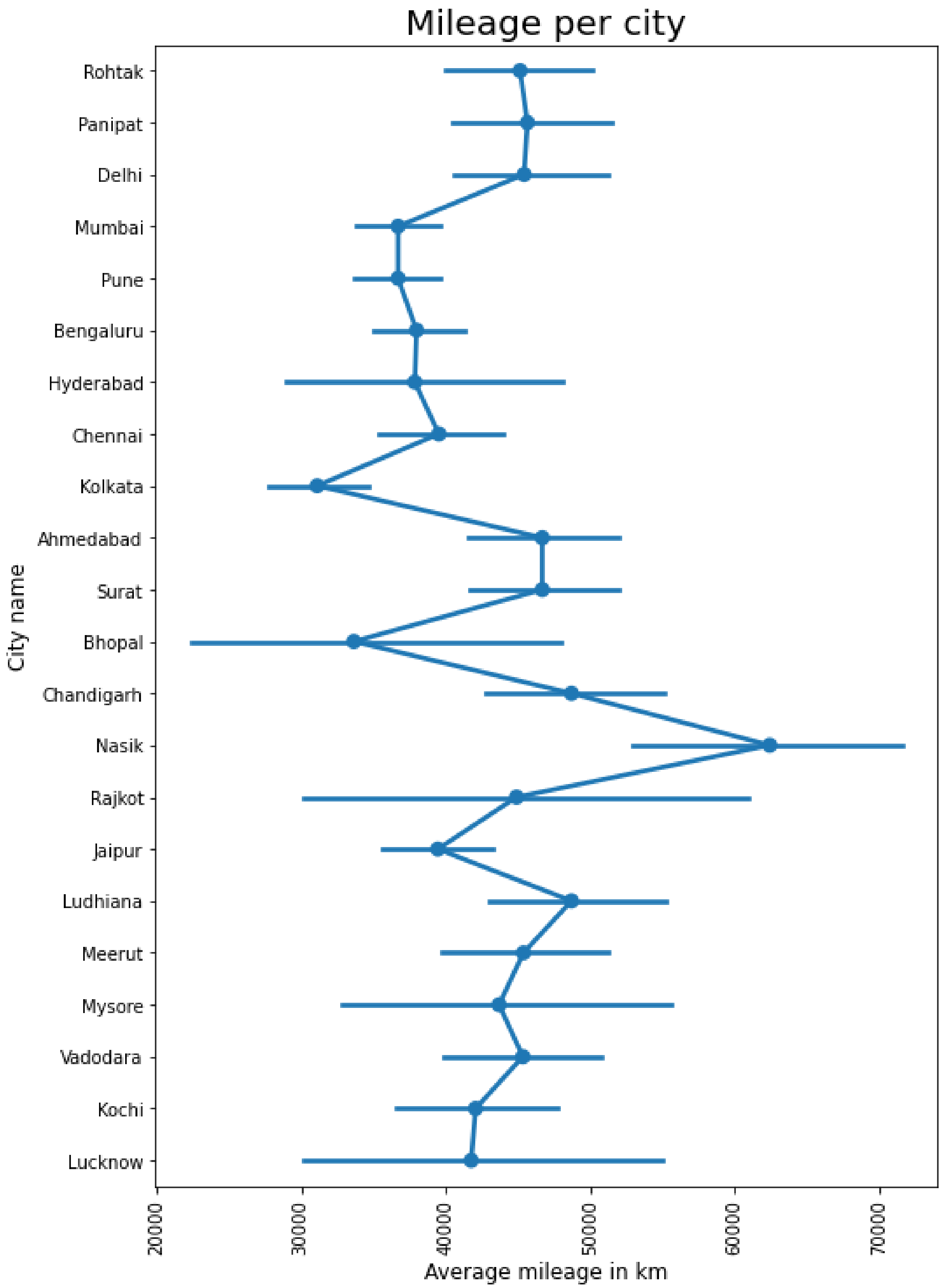
We can observe that newer cars are priced much higher on average as expected. The older the car is, the lower the resale value is.



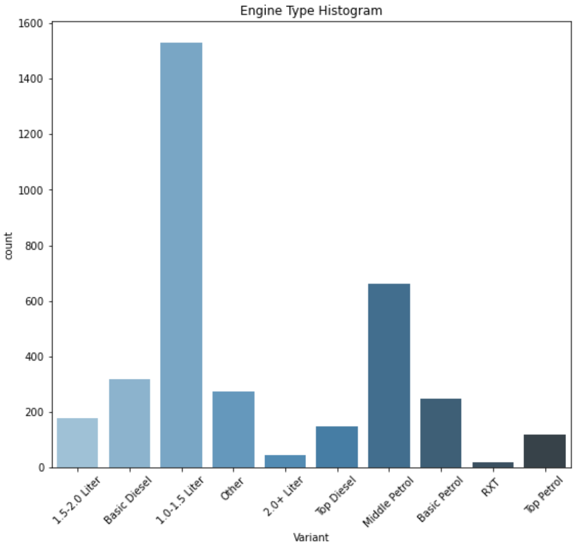
This plot shows the average mileage on a car (in km), based on which year the car was manufactured. We can observer that the older the car, the more mileage it has.

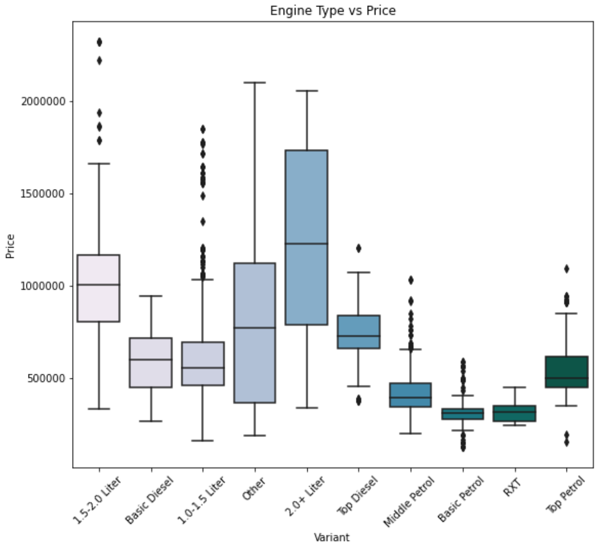


Average price of car based on city, Kolkata has the cheapest cars on average and Bhopal the highest price.



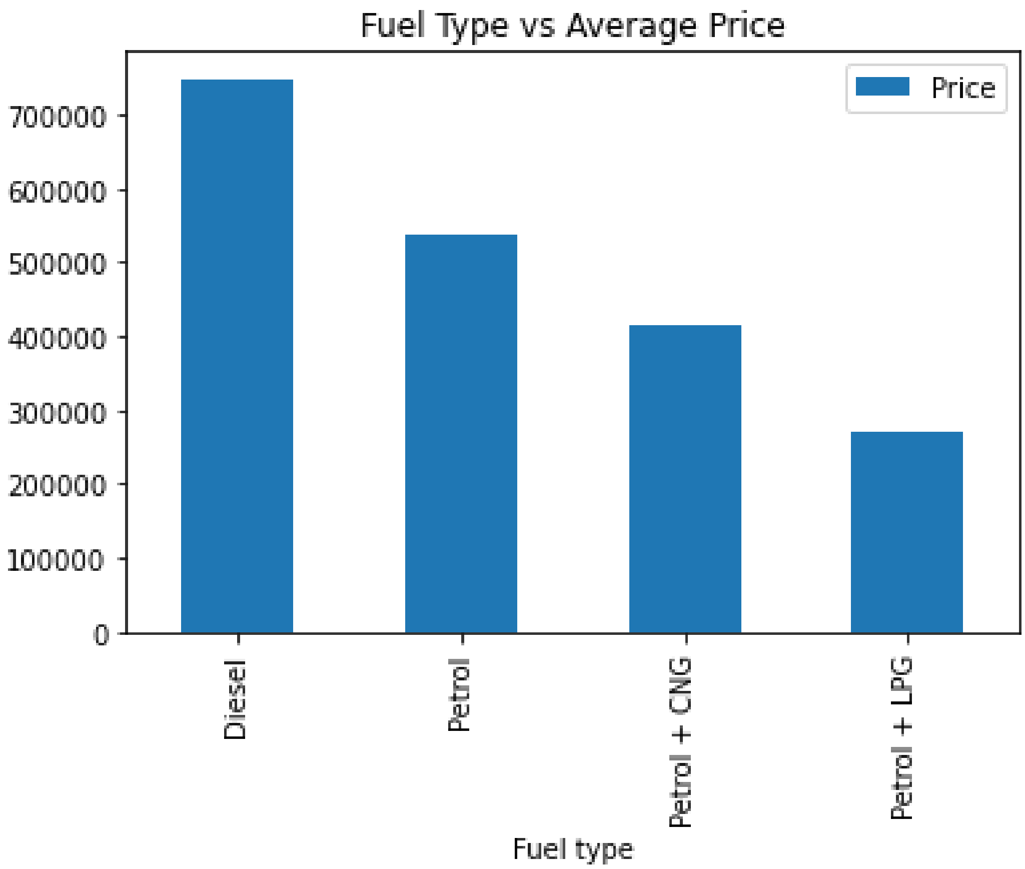
This plot shows the average mileage on car based on which city the car is listed in.



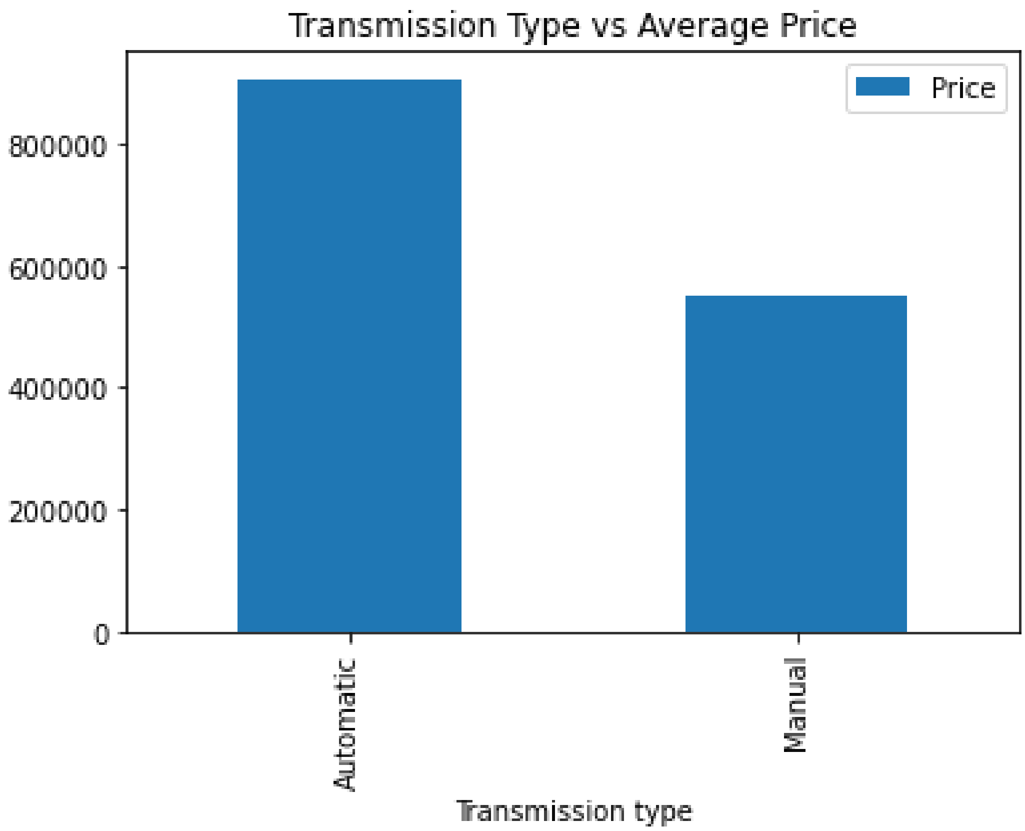


The above plot shows the count of engines found in cars, the plot to the right shows the average price of all those categories.

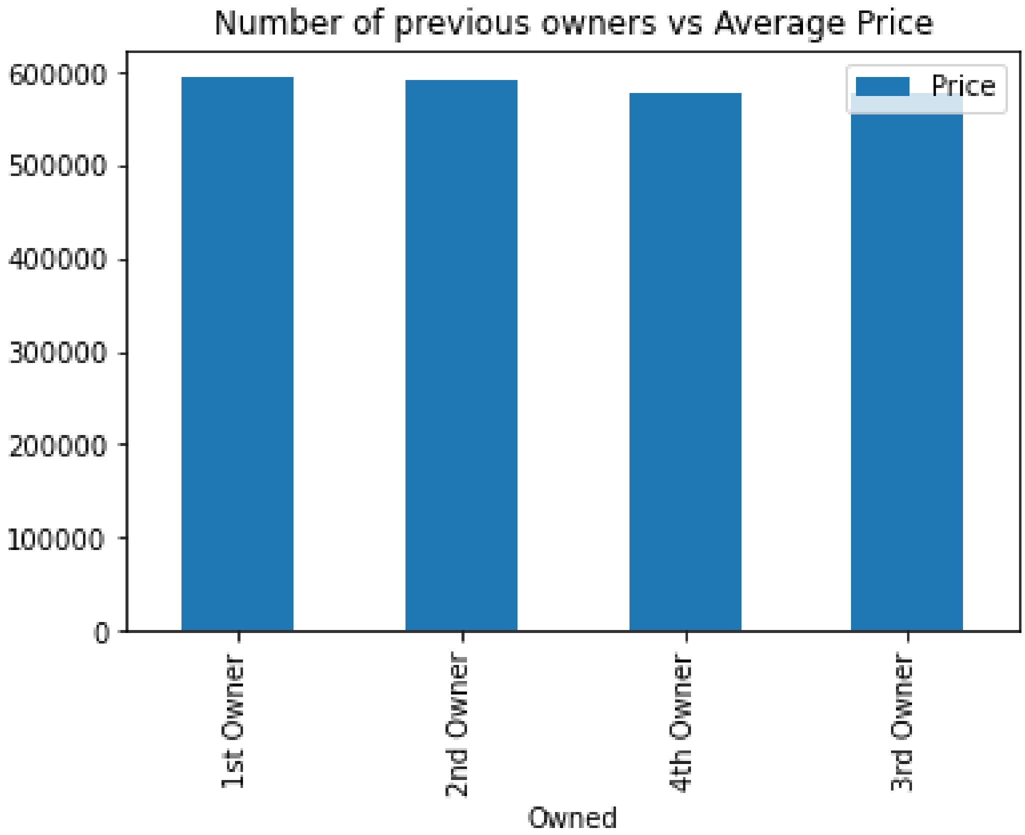
We can observe that 1.0 – 1.5 litre is the most common engine found in cars. As it is the cheapest engine. The most expensive engines are 2.0+ litres as expected heavier engines mean bigger cars which means much higher price.



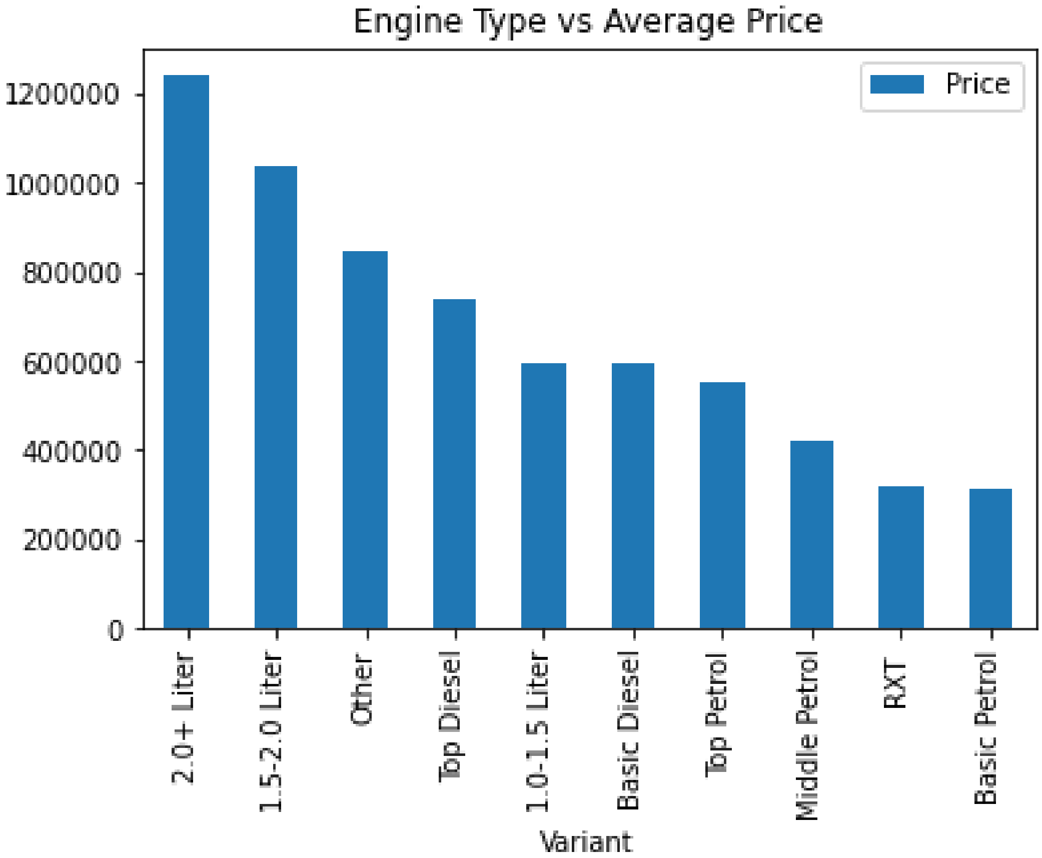
Diesel cars are resold for the most value, on average Rs. 7 lakh, petrol cars are price just above Rs. 5 lakh on average. CNG and LPG cars are the cheapest.



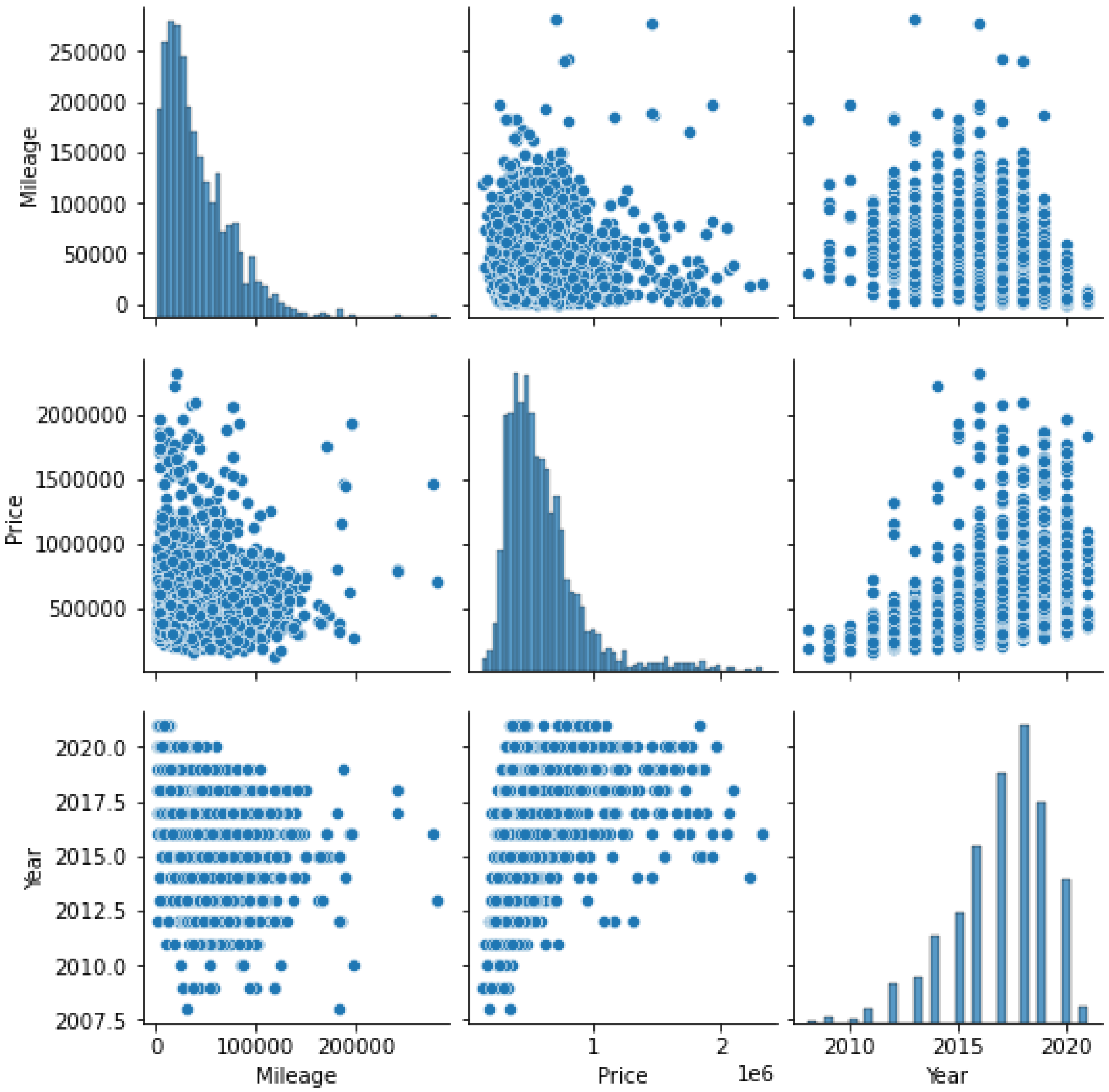
Average price of car according to type of transmission, as we can observe here that automatic cars are on average price above Rs 8 lakh and manual around 5 lakh.



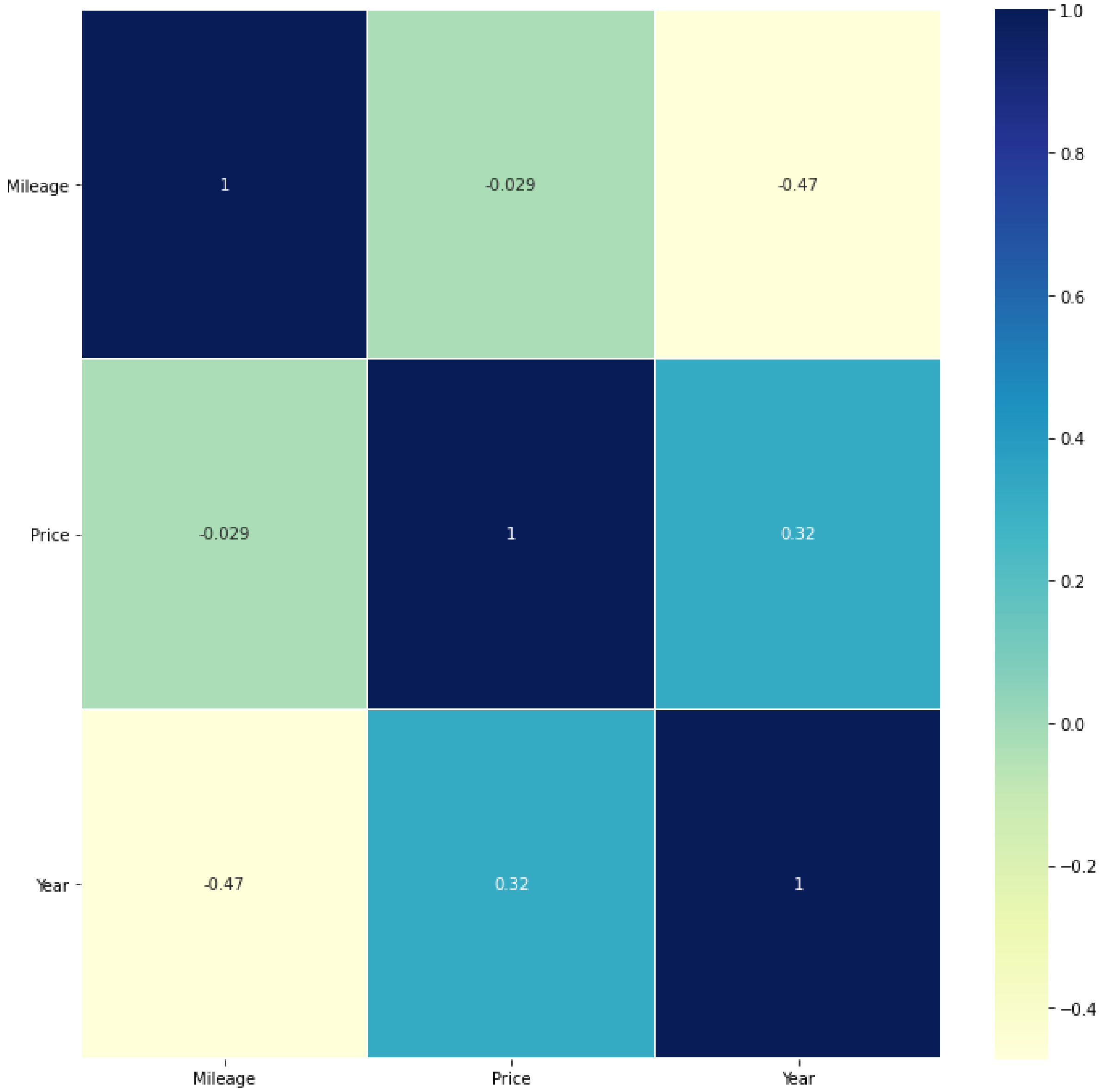
Average price of car based on how many previous owners it has had. We can only see a noticeable bump when there are more than 3 number of owners of the car. 1/2/3 owners doesn’t affect the price as much.



2.0+ litre engine is the most expensive



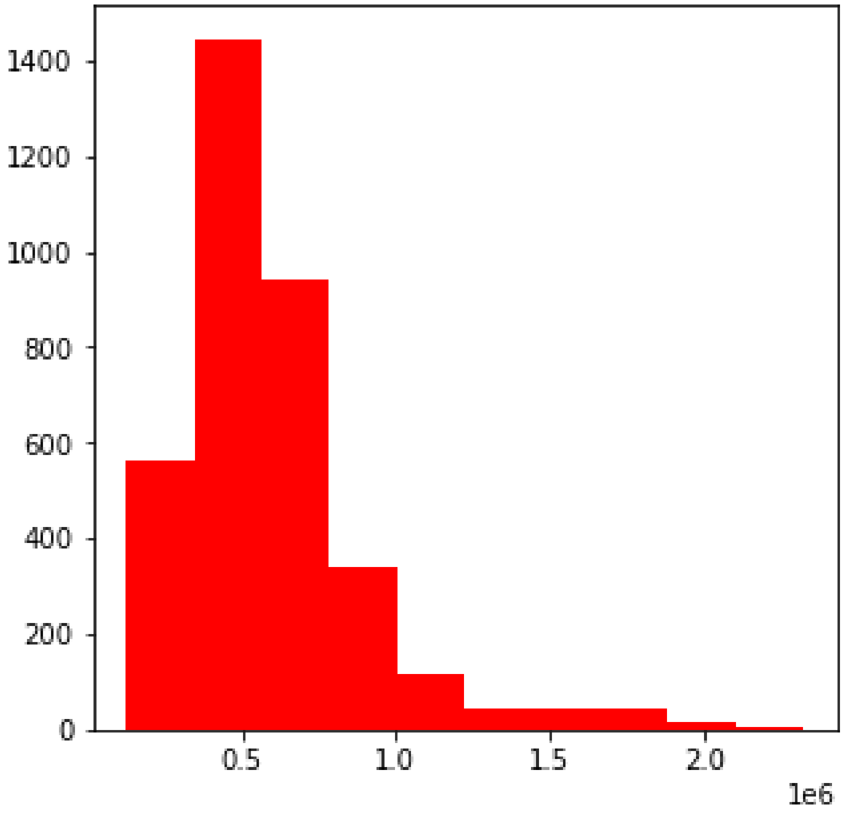
Multi variate analysis between the various numerical variables. Year, Mileage and Price.



Coorelation heatmap of the variables.

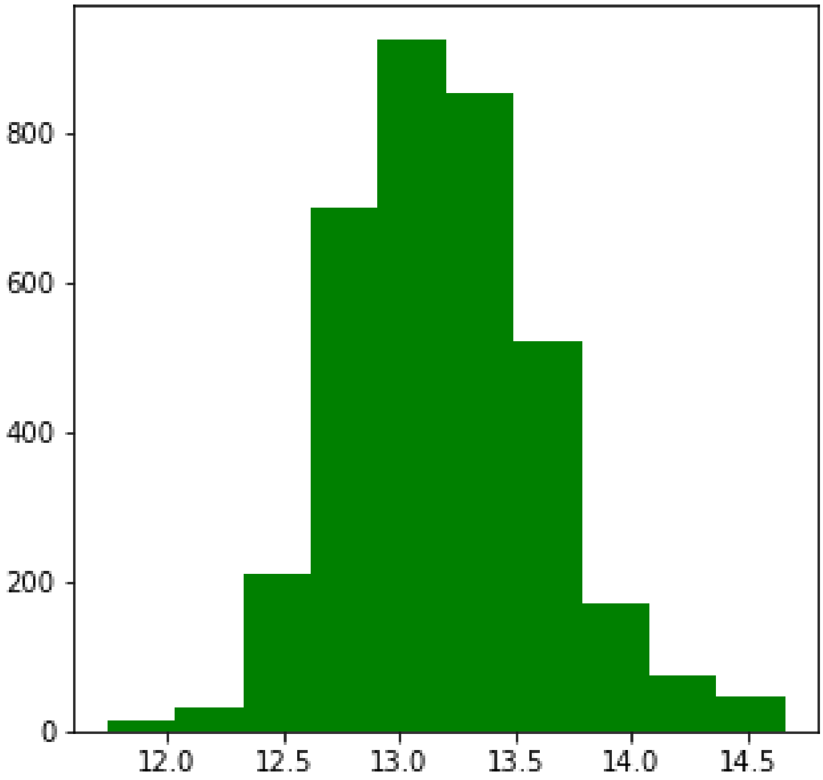
We can see that price is negatively affected by mileage i.e. the more mileage a car has the lower its price.

And the price is positively co-related with the year, as newer the more expensive it is.



This is the skew of Price.

Skew of Pirce: 2.0398565390925087



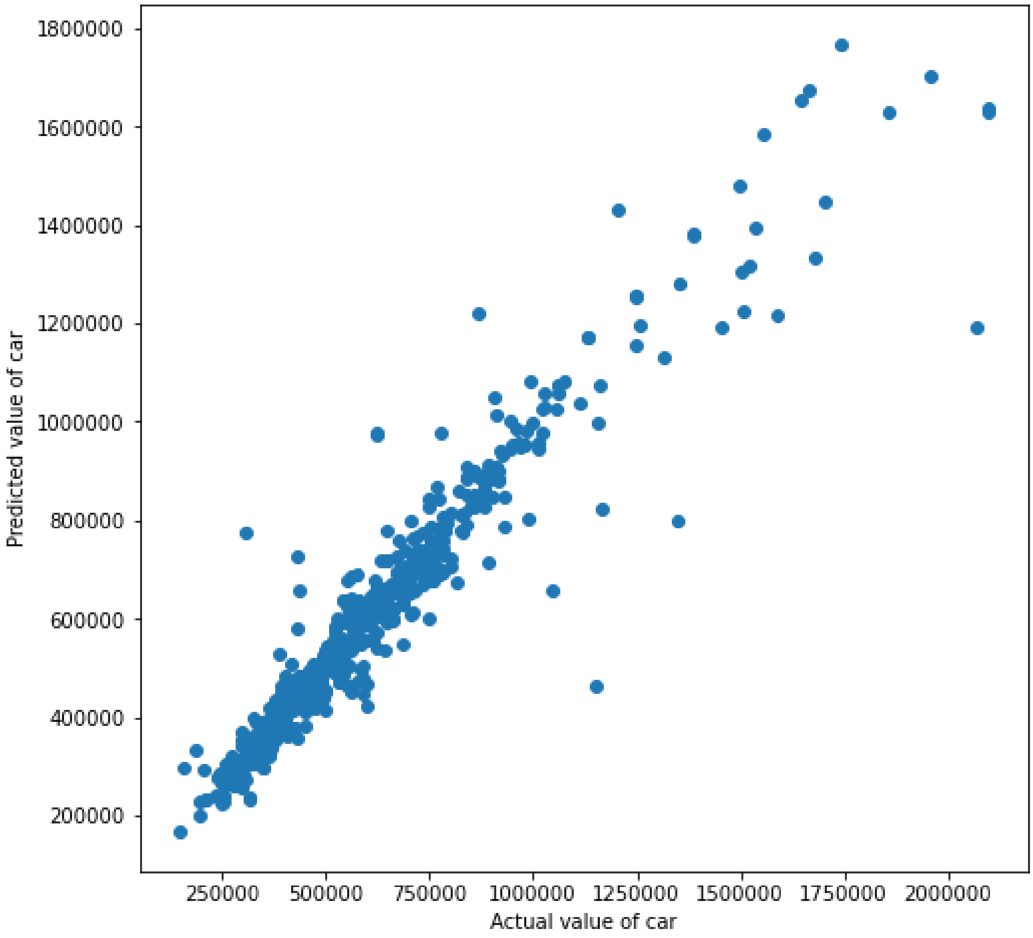
This is the skew of price after taking natural log.

Skew of Log-Transformed Price: 0.3518219716476838

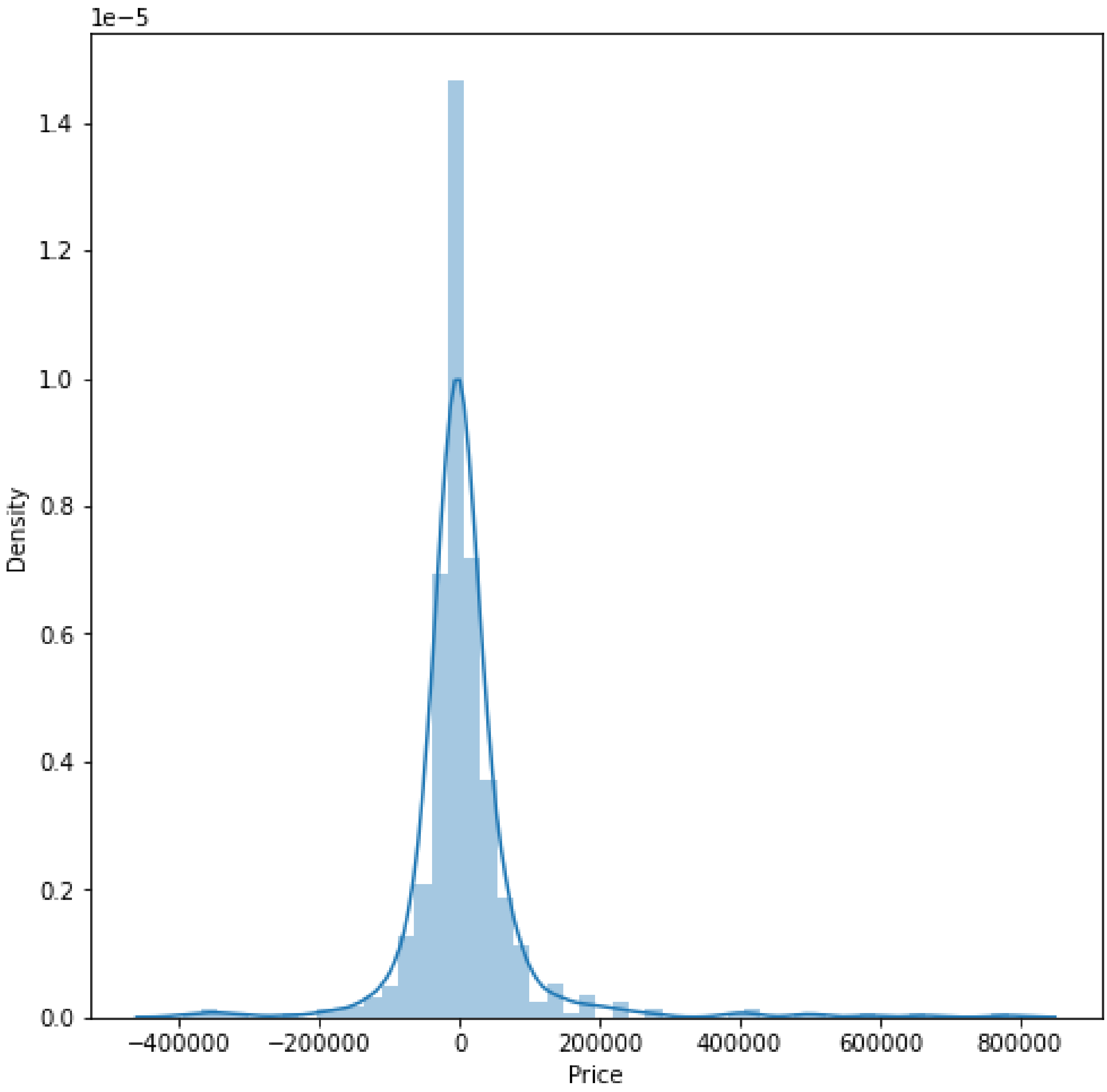
We can see that now we have a much better bell curve shape.

- Interpretation of the Results

Results were concluded from scatter plot of the predictions vs actual values and the mean absolute error between the two.



Final graph of the model predicting the values, we can see that the model is very accurate in determining the price.



Density plot of predicted price – actual price. We can see that most of the guesses are very near to the actual value. Which is good.

CONCLUSION

- Key Findings and Conclusions of the Study
  - The main component on which the price of a car depends is the engine size, the year which car was bought; the mileage on the car etc.

- The price also depends on which city the car was registered, as some cities have different tax rates and restrictions. Eg Delhi NCR has 10-year limit on diesel cars and 15 year on petrol cars, but no other city has such restrictions.

- Learning Outcomes of the Study in respect of Data Science
  - Random forest regression works best for this particular data set, hyper parameter tuning was performed and optimal parameters were found.

  - EDA is very powerful in understanding the data and pre-processing it before feeding it to the algorithm. Statistical methods work the best.

- Limitations of this work and Scope for Future Work
  - Post covid-19 car market is still evolving, and it will keep evolving for the foreseeable future. The algorithms will need to keep changing to keep up with the evolution.