Cars Price Prediction Analysis

By:

LAKSHITA KAIN

DATA SCIENTIST(INTERN)

FLIPROBO TECHNOLOGIES

Introduction

- 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. Due to rising fuel prices, fuel economy is also of prime importance.

PROBLEM STATEMENT

- Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and it's fuel type.
- This study focuses on predicting a used car's selling price all over India using various supervised machine learning models based on car features.

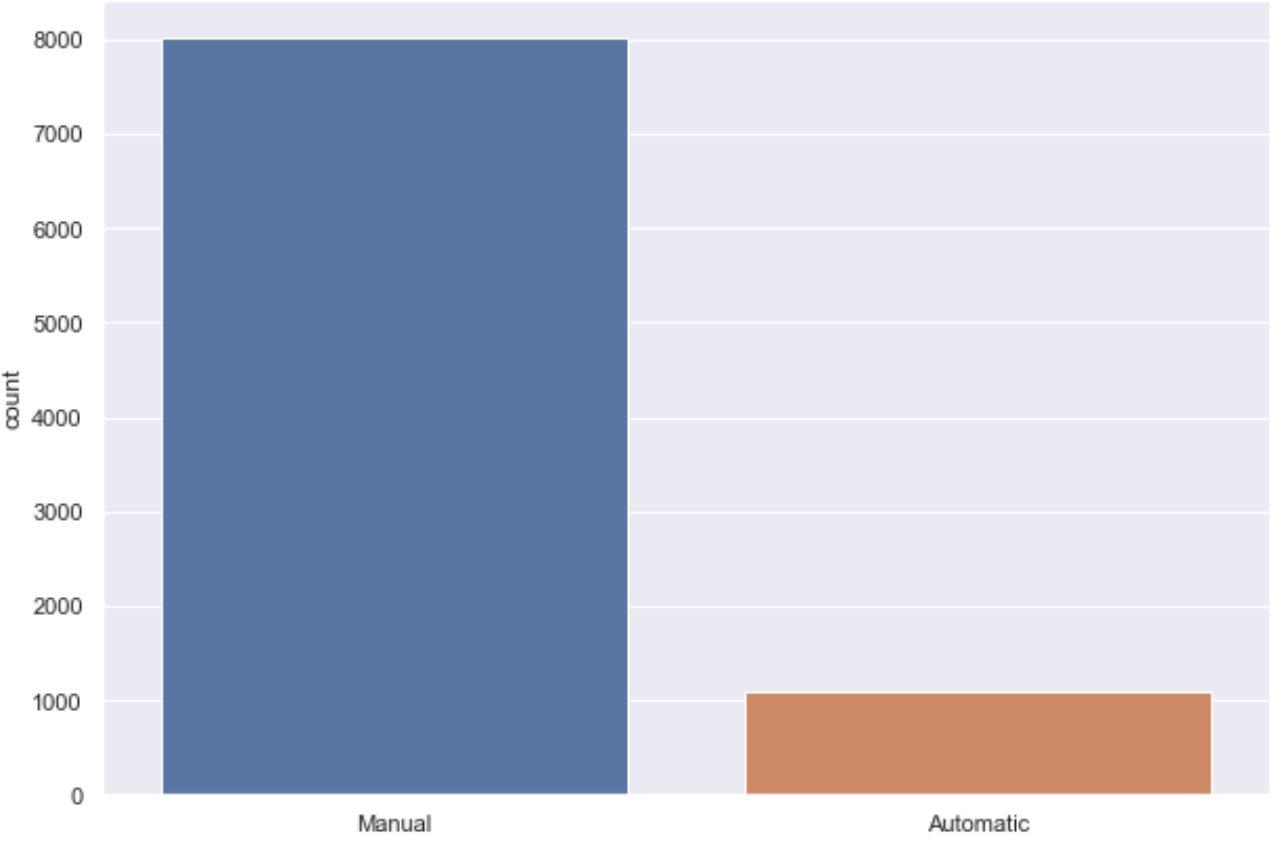
DATA COLLECTION

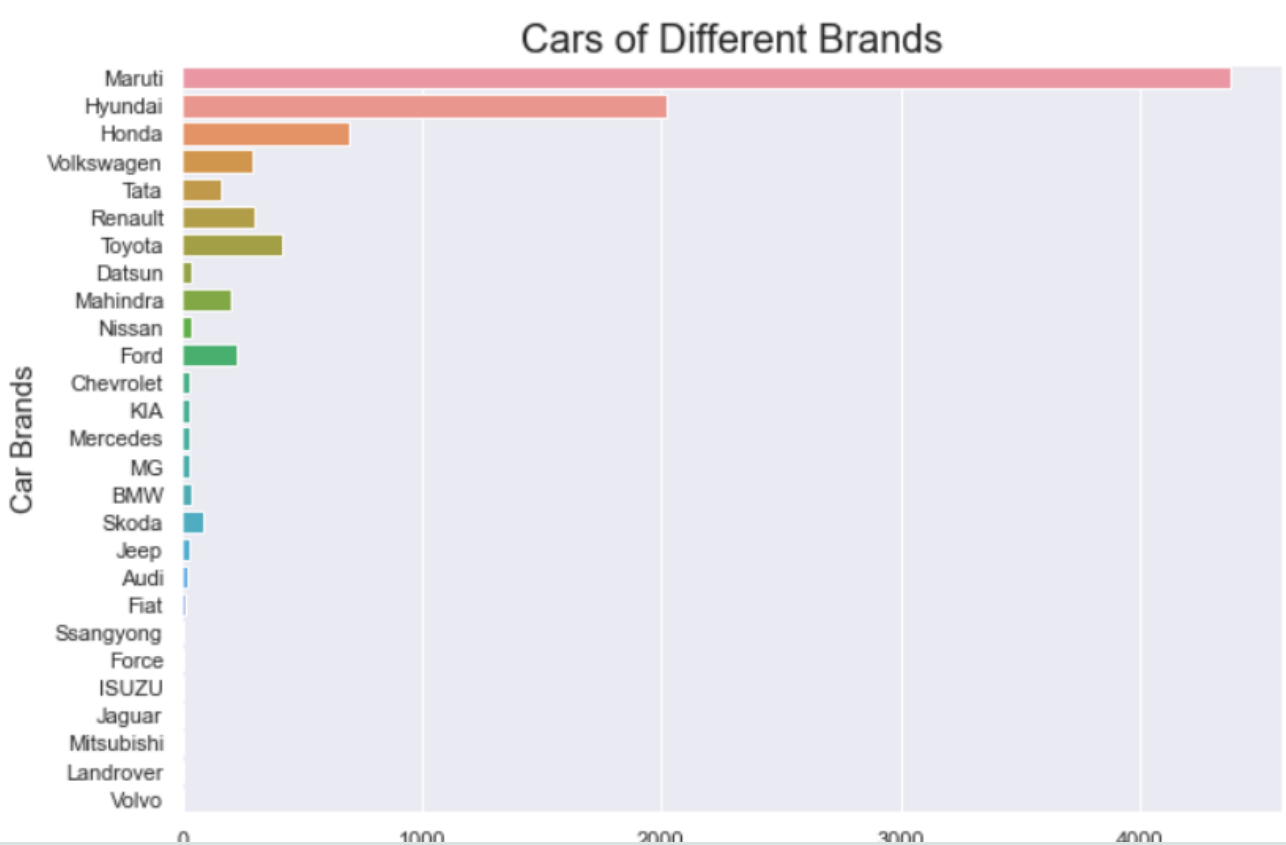
Data has been scrapped using Selenium webdriver with python. Selenium is a powerful tool for controlling web browsers through programs and performing browser automation.

Our data has been retrieved from https://www.car24.com which consists of used car information such as model, brand, location, kilometers driven, fueltype, transmission, engine, Price, Monthly EMI price and direct URL. The data collected is from Delhi – NCR, Chennai, Maharashtra, Pune, Hyderabad, Gujarat, Jaipur Chandigarh and Kolkata.

EXPLORATORY DATA ANALYSIS

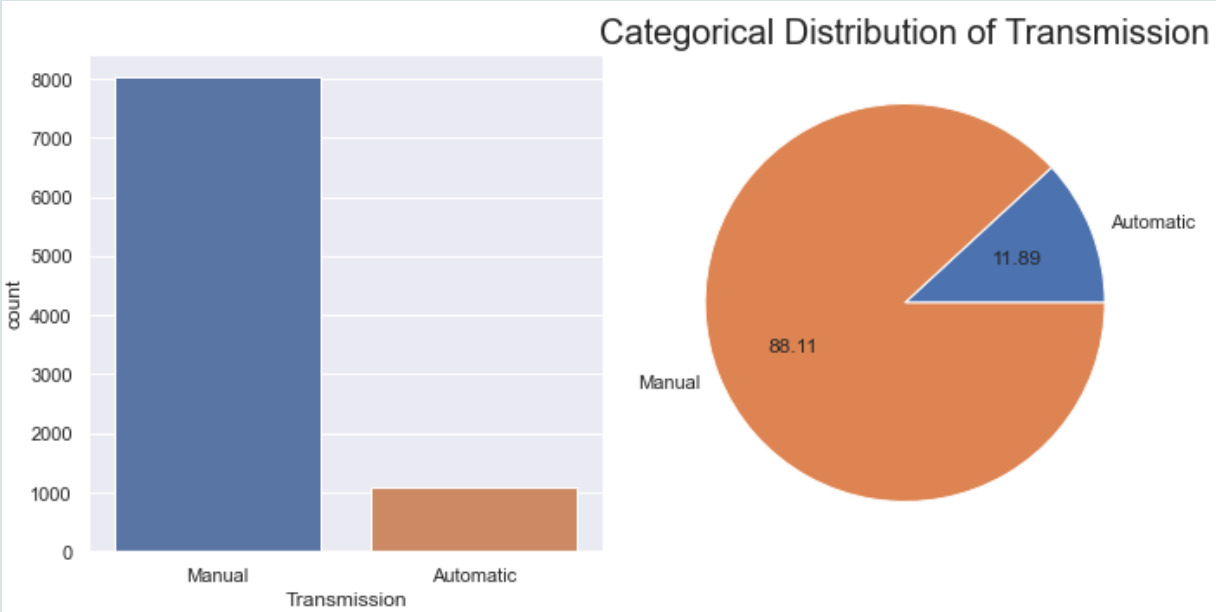
- Shape of our dataset : `records -9100, features - 13`
- There are no null values
- There are no duplicated fields or attributes.
- "Unnamed : 0" is an necessary column for our analysis therefore we drop it.
- There are 10 object datatype columns and int datatype attributes.
- "Kilometers", "Price" and "Monthly_EMI" have object datatype which needs to be changed for further analysis.

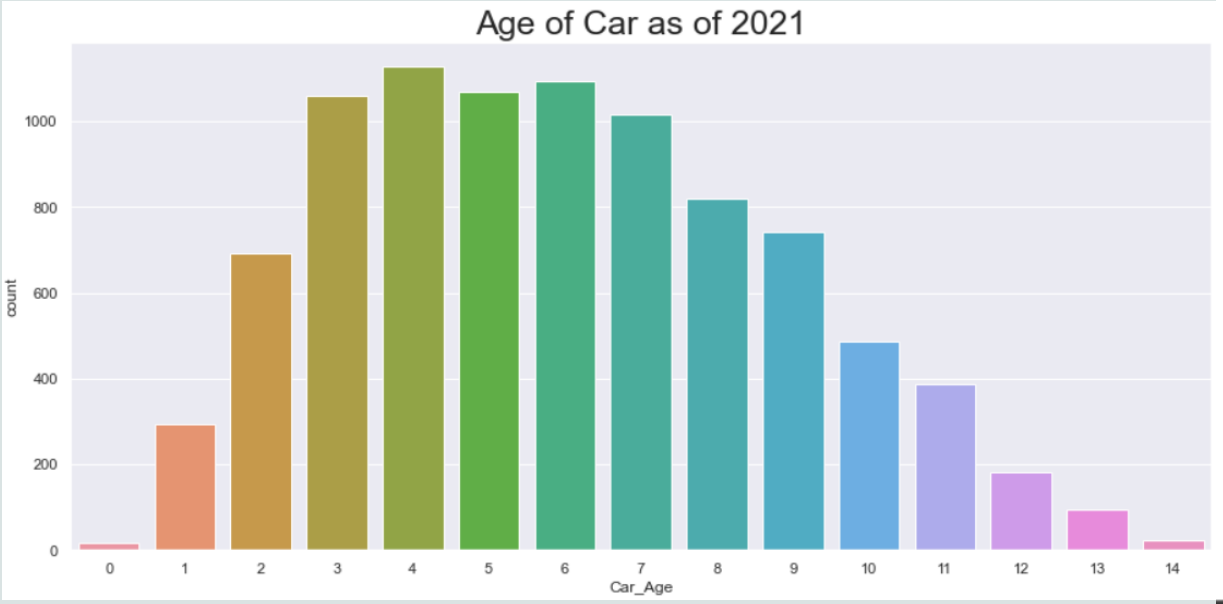


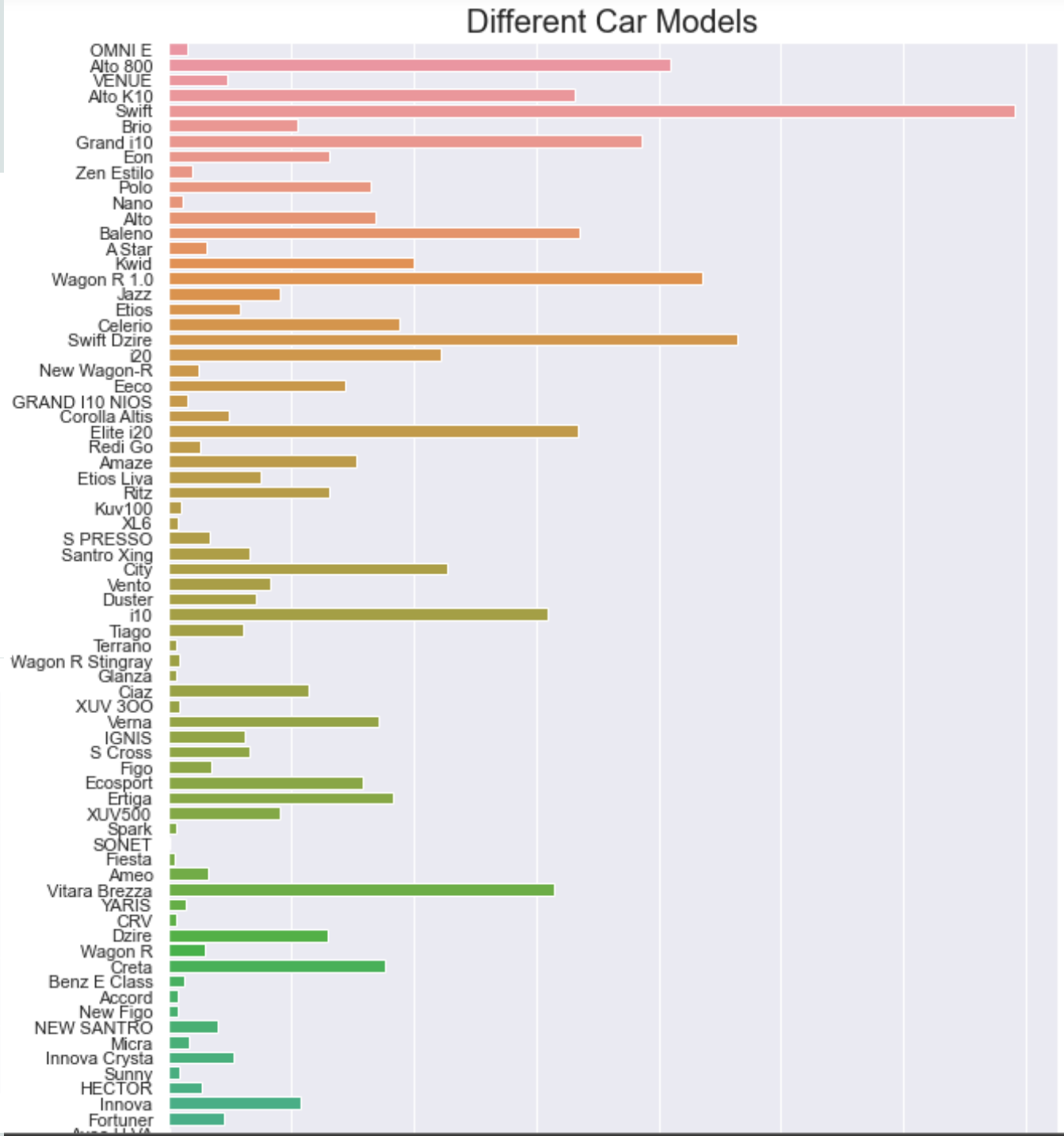


UNIVARIATE ANALYSIS OF FEATURES

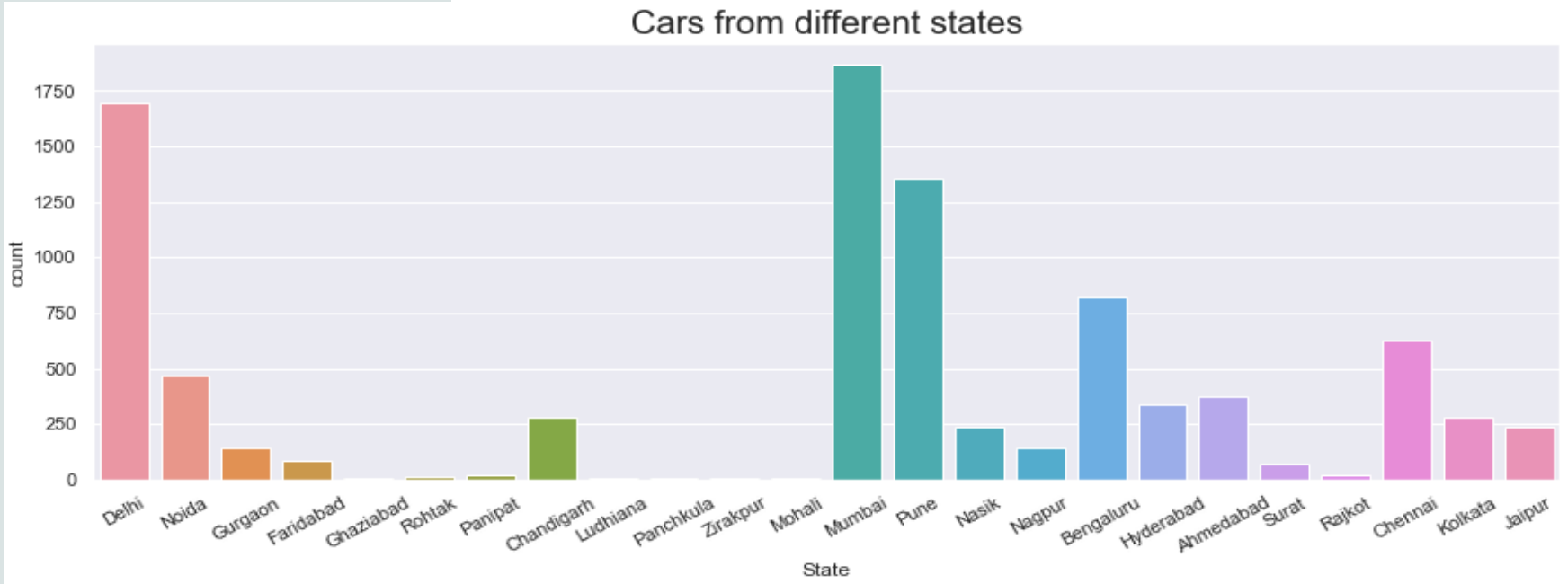
UNIVARIATE ANALYSIS OF FEATURES

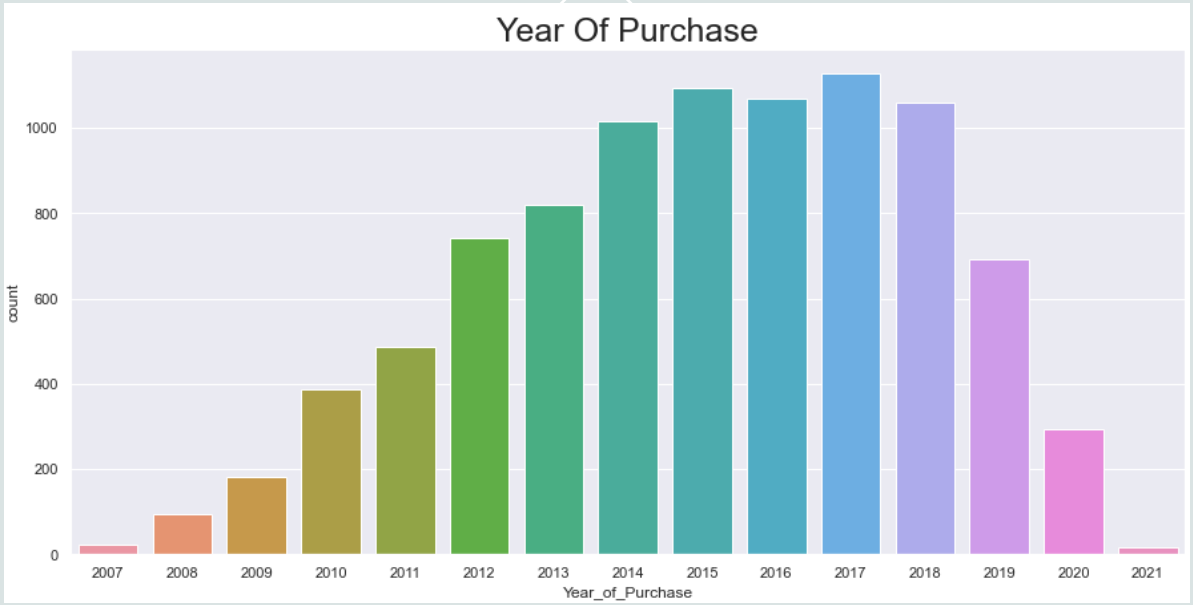


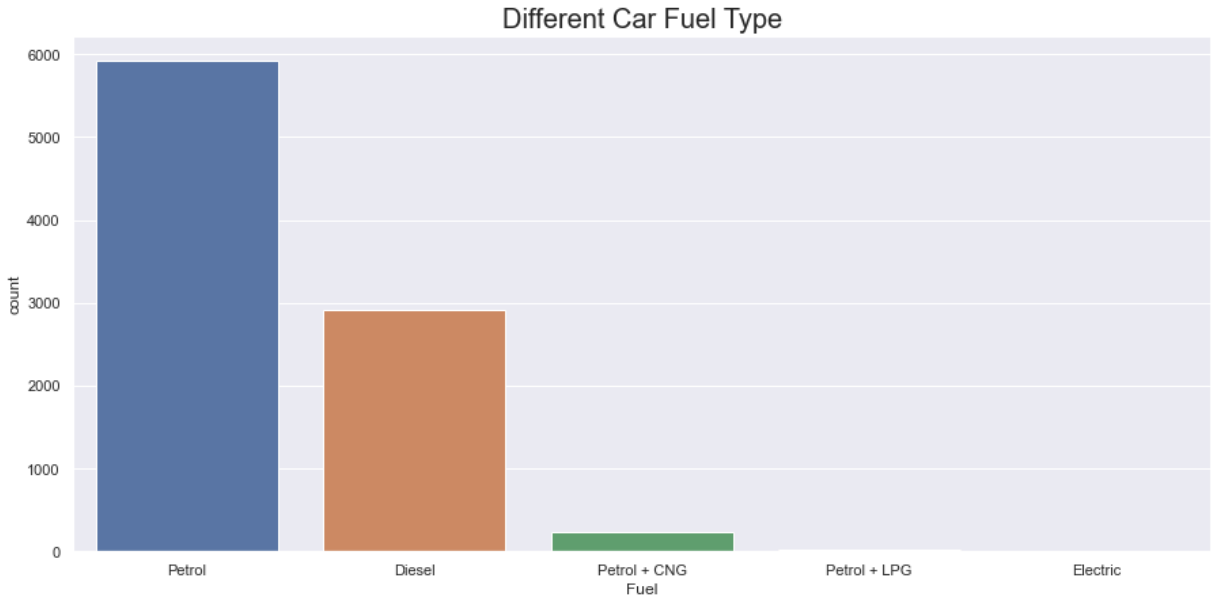




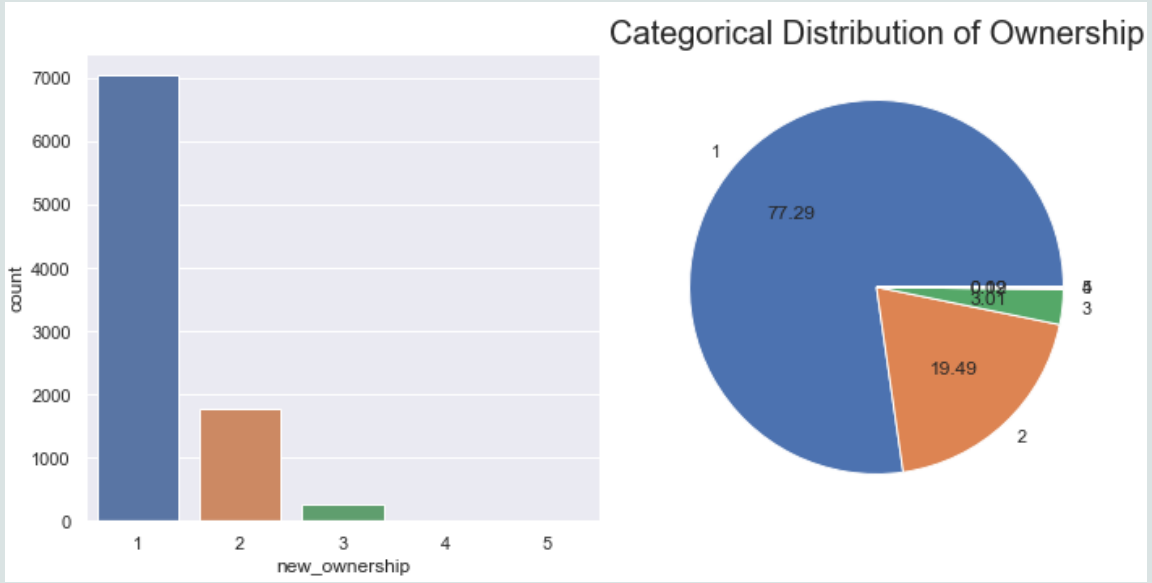
UNIVARIATE ANALYSIS OF FEATURES

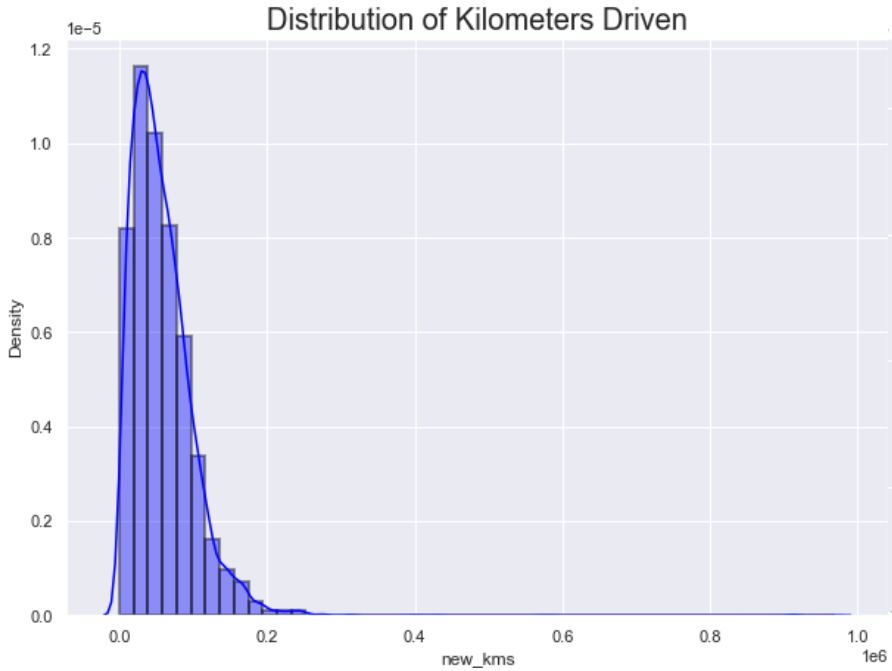


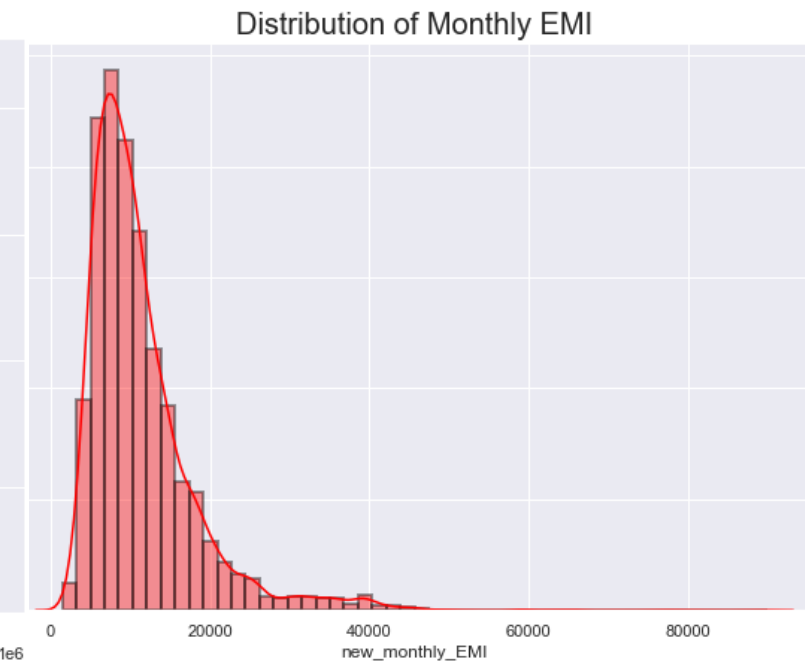




UNIVARIATE ANALYSIS OF FEATURES



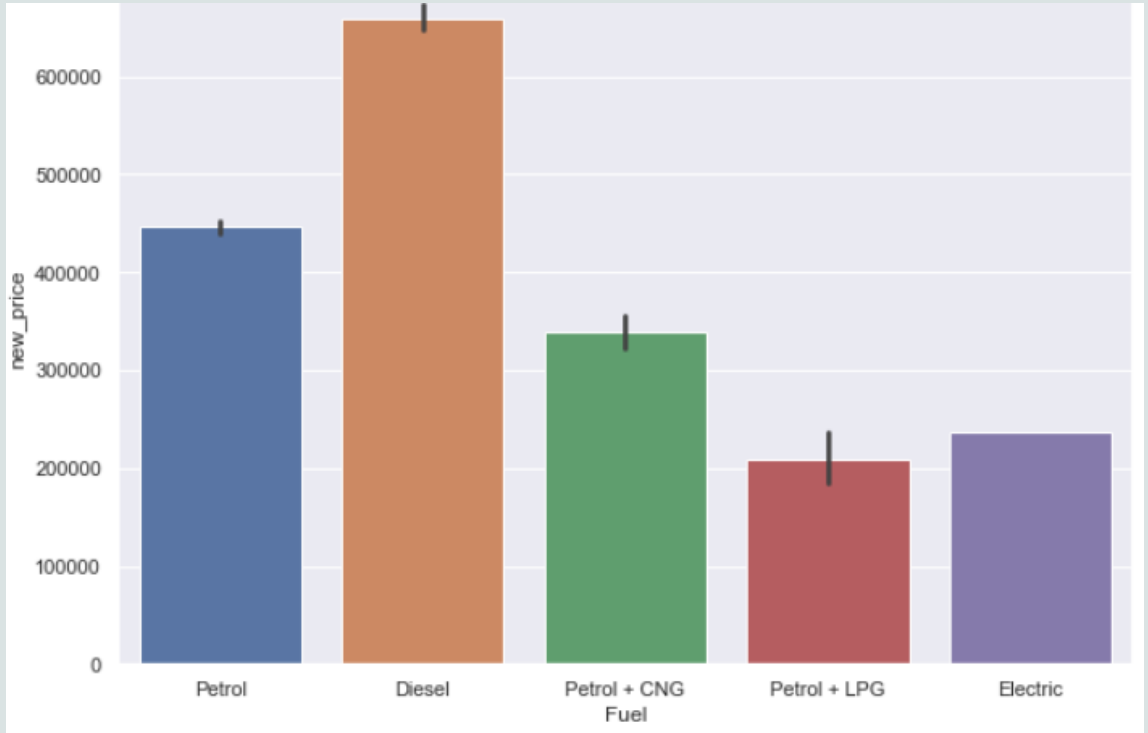




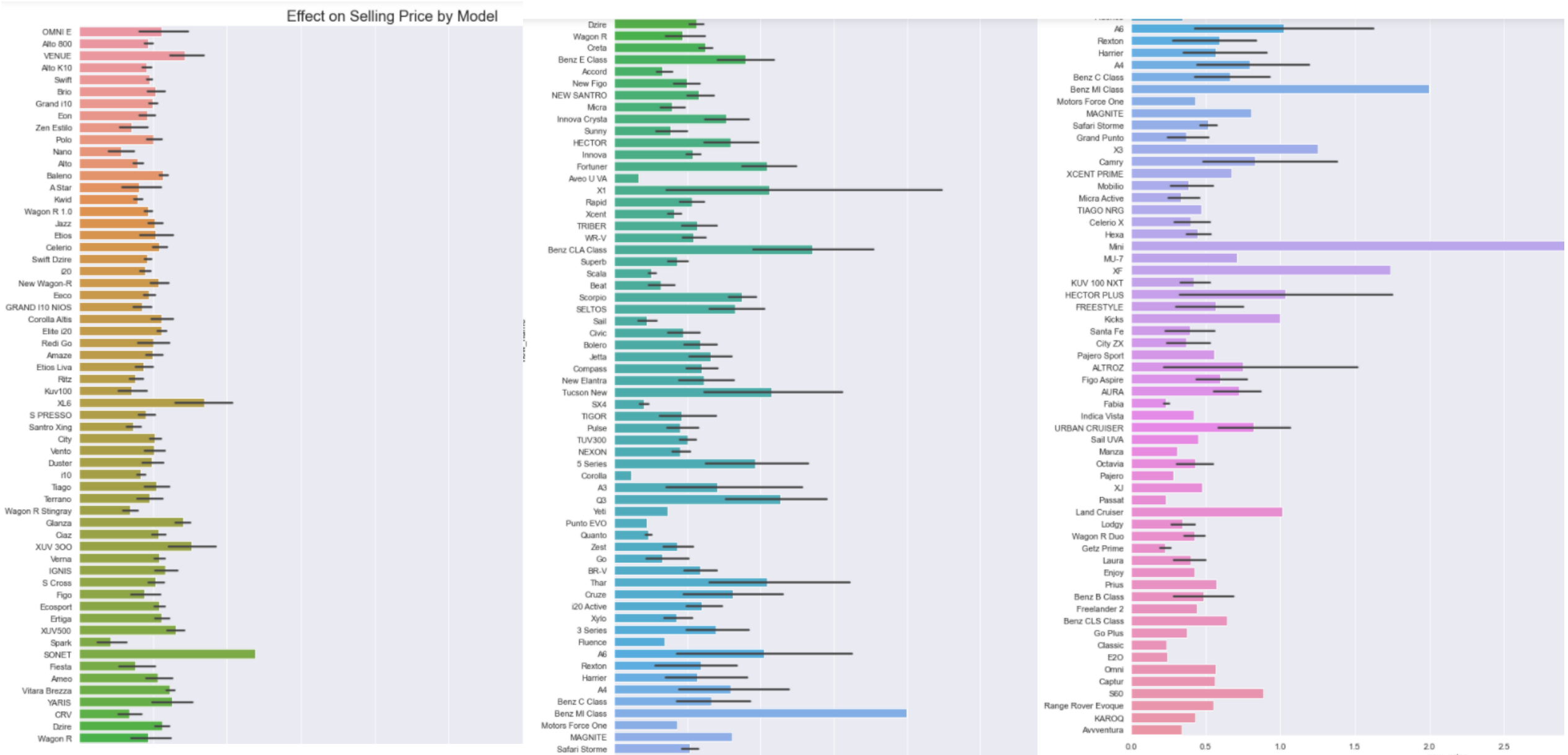
INSIGHTS

- Car from Maruti are high in number.
- Frequency of Cars with Manual transmission are 8x then Automatic cars in our dataset.
- Most of the cars are sold from Mumbai followed by Delhi.
- Most cars sold are from the first owners.
- Our Data is right – skewed for features kilometers driven, Price, and Monthly EMI.
- Petrol Cars advertised are High number.

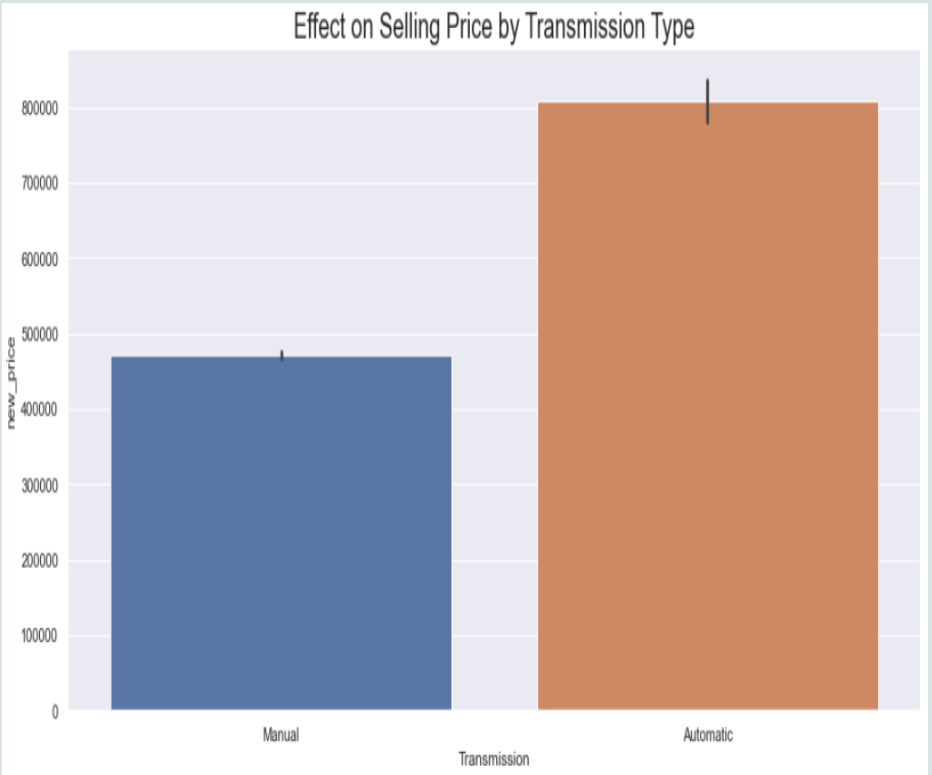
BIVARIATE ANALYSIS OF FEATURES



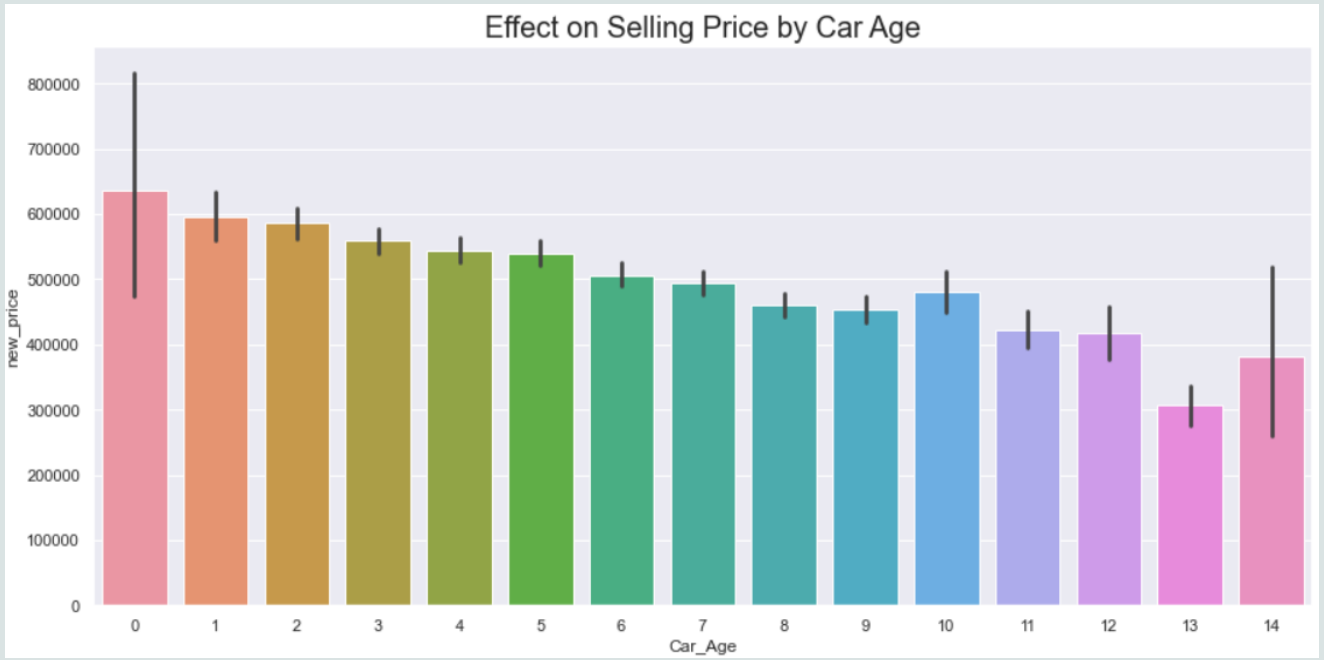
BIVARIATE ANALYSIS OF FEATURES

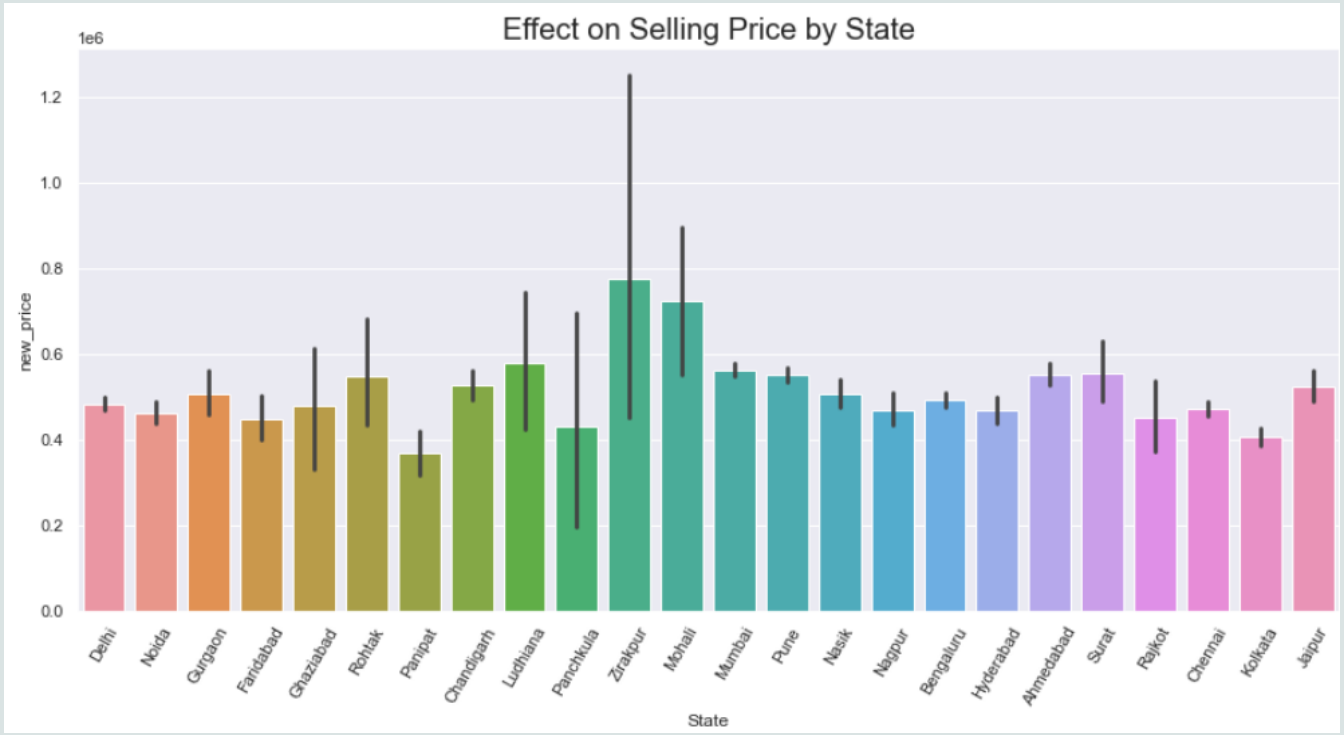


BIVARIATE ANALYSIS OF FEATURES



BIVARIATE ANALYSIS OF FEATURES

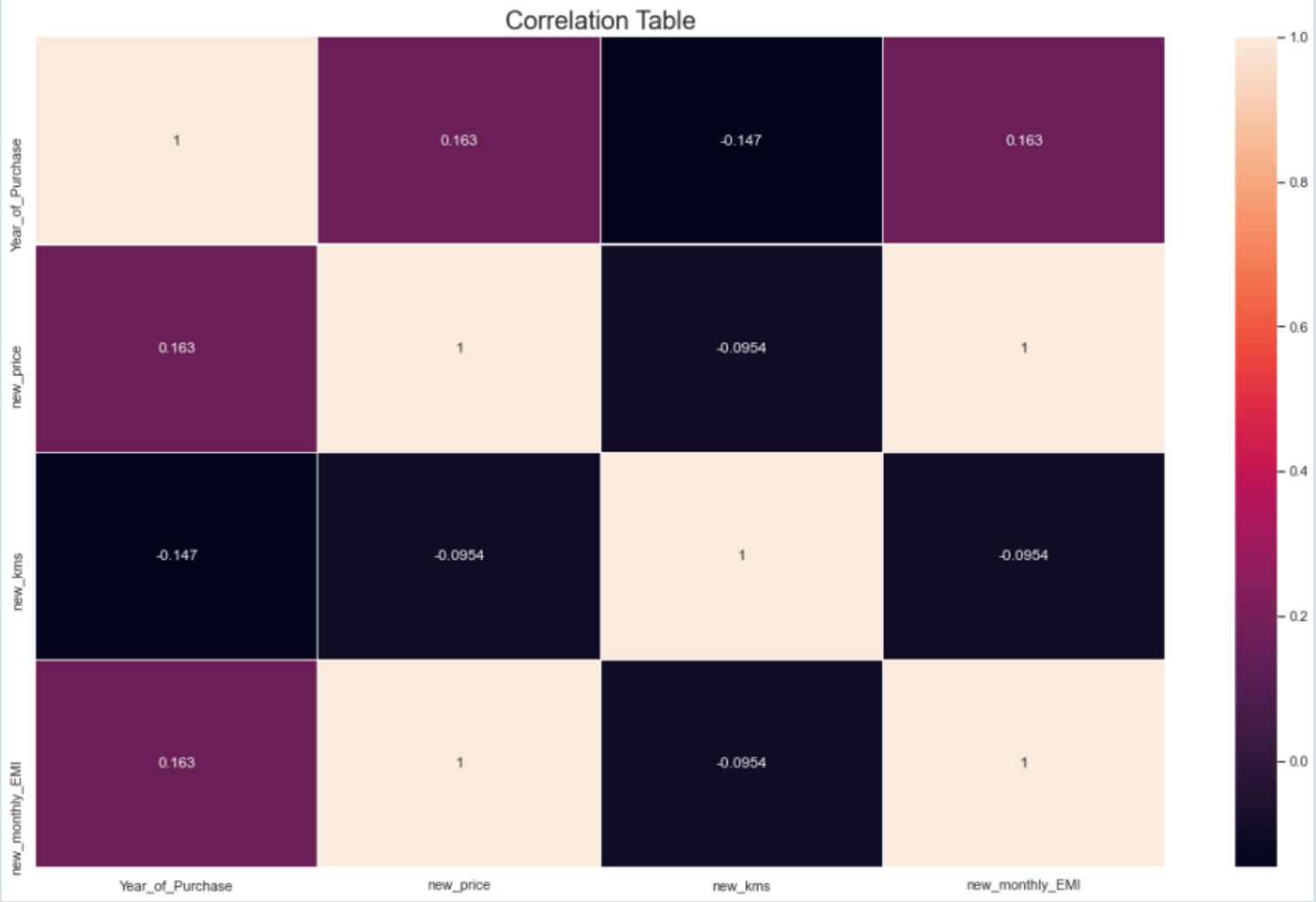


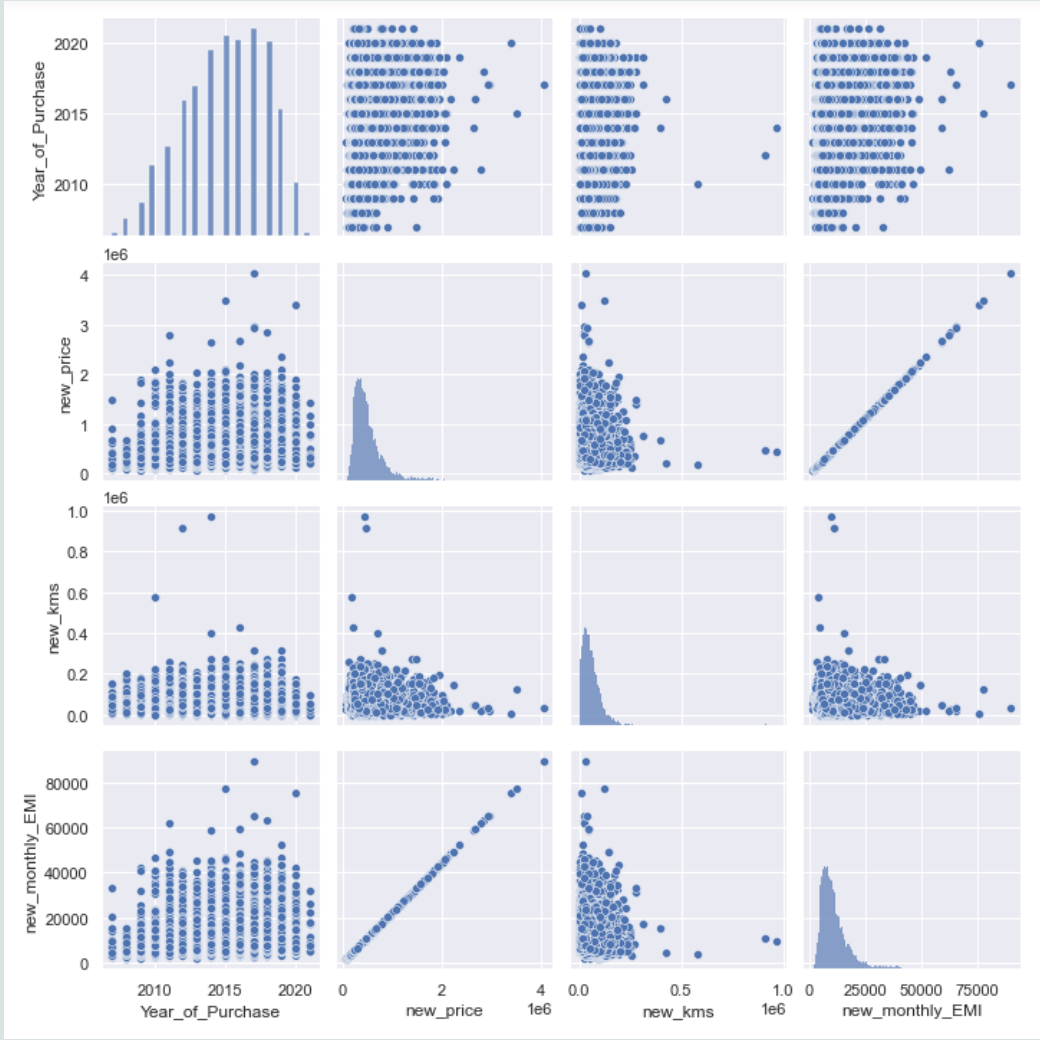


INSIGHTS

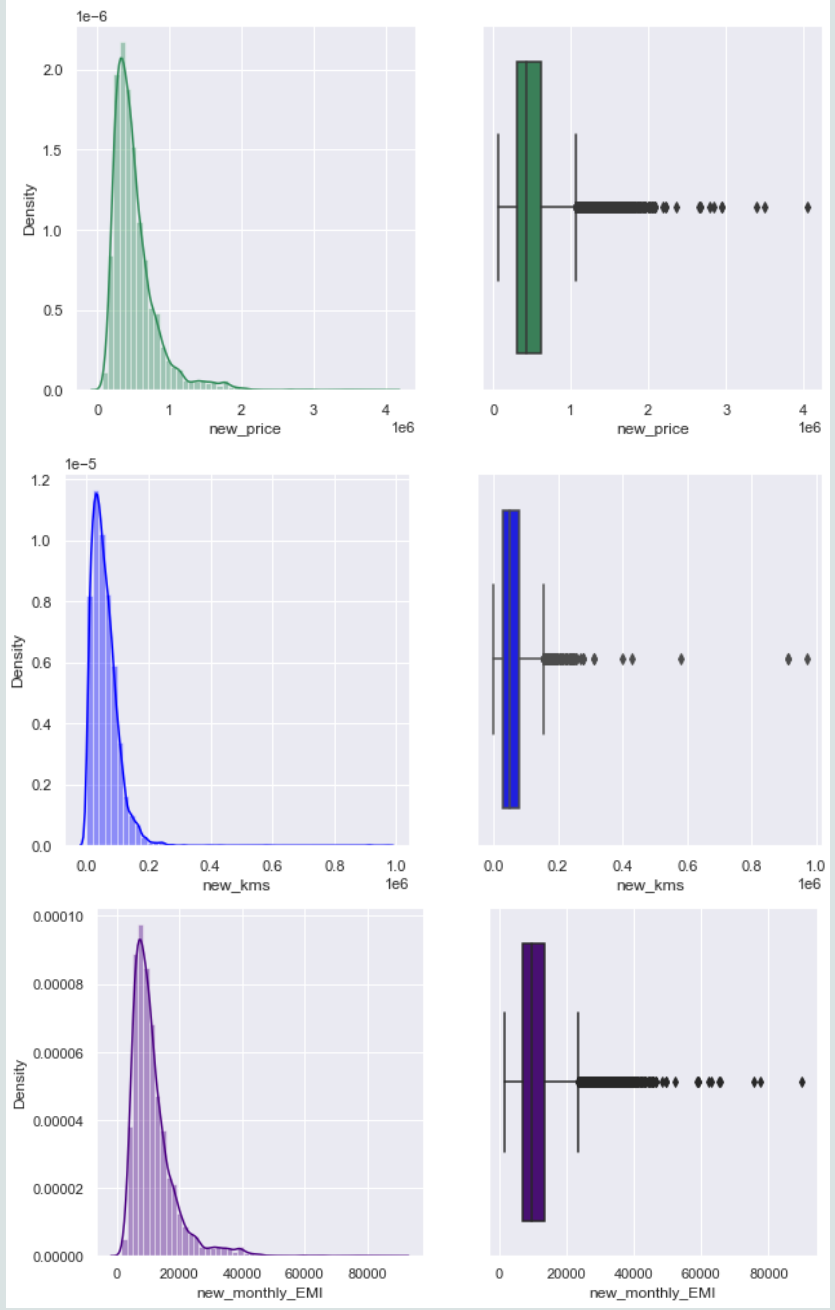
- Car from "Jaguar" costs the highest whereas cars from "Chevrolet" costs less.
- Cars with Diesel as Fueltype have higher selling price followed by petrol. However cars with Petrol+LPG have lowest selling price.
- Mini Cooper is the most expensive car in the lot.
- Car value decreases as the number of owners increases.
- Automatic Transmission Cars costs more than the Manual transmission cars.
- As the age of car increases it's selling price decreases. Older the car lower the value.
- Cars from Zirakpur costs highest and cars from panipat costs lowest.
- Price of a car is positively correlated with the Kilometers driven.

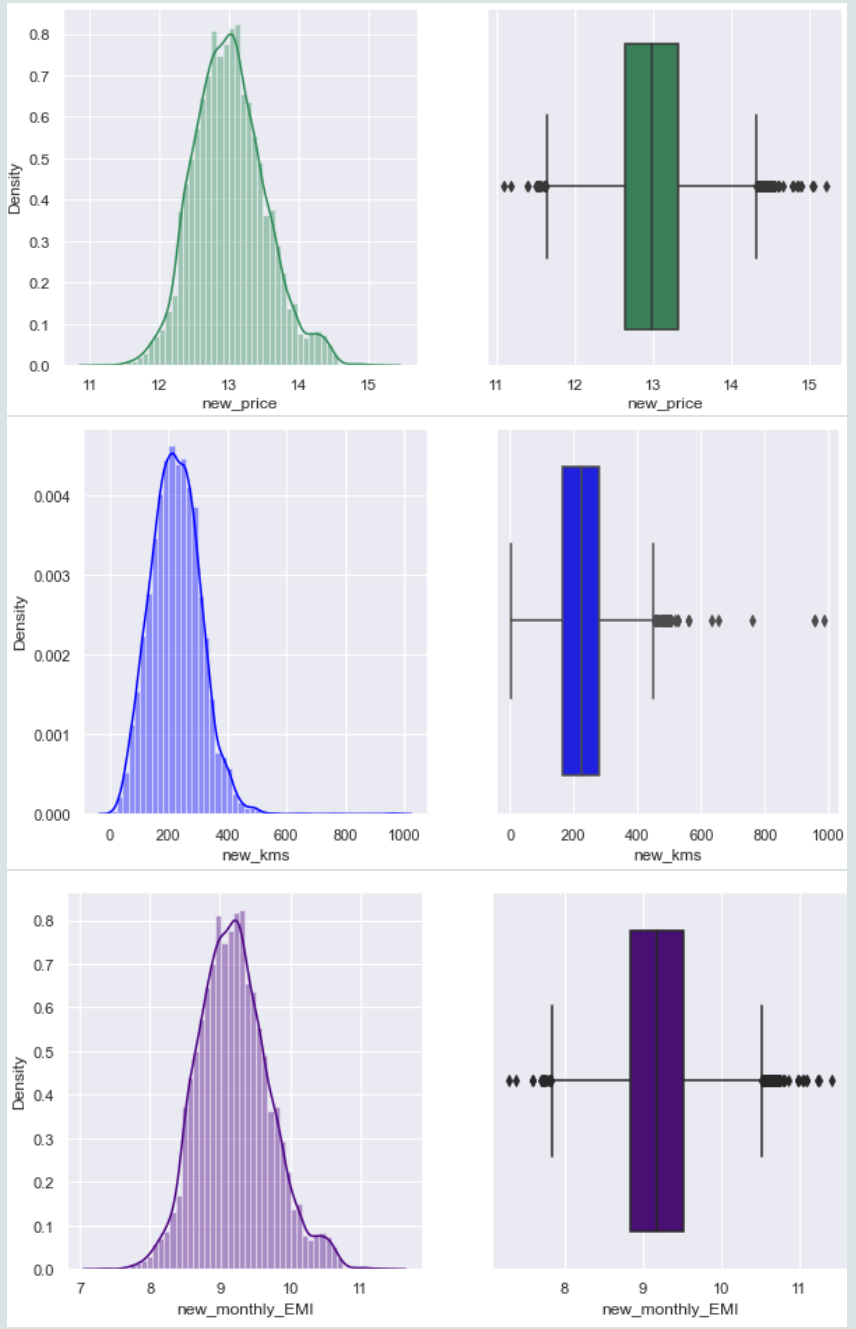
MULTIVARIATE ANALYSIS OF FEATURES





SKEWNESS REMOVAL USING LOG AND SQUARE ROOT TRANSFORMATION





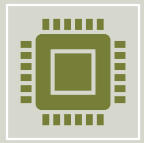
PREPARING DATASET FOR MODELLING

- Categorical features of our dataset were encoded using LabelEncoder.
- Dataset was split into two parts 25% for testing and 75% for training.

Models Used and their Evaluation Metrics

	r2 Score (%)	Mean Absolute Error(e-05)	Mean Squared Error (e-09)	Root Mean Squared Error(e-05)
Linear Regression	99.99	2.8340	1.3370	3.6571
Lasso Regularization	99.99	4.8300	3.5660	5.9721
Decision Tree Regressor	99.99	0.0007	8.0760	0.0089
Random Forest Regressor	99.96	0.0006	8.8880	0.0094
Ridge Regularization	99.99	2.8344	1.3370	3.6578
K-Nearest Neighbors	7.08	0.3788	0.2430	0.4932
Gradient Boost Regressor	99.96	0.0030	9.0090	0.0094
Adaptive Boosting Regressor	98.88	0.0410	0.0029	0.0540
Bagging Regressor	99.99	0.0007	8.7580	0.0093

Results



Almost all of our algorithms performed well except the K – Nearest Neighbor Regressor.

Linear Regression model yielded high accuracy apart from it, Linear Regression is the fastest algorithm amongst all the other ensemble algorithms.

Hence we move forward with the hyperparameter tuning of Linear Regression model with L1(Lasso Regularization) to analyze if our model is overfitted or not.

HYPERPARAMETER TUNING

Randomized Search CV was used to tune our Linear Regression model with L1(Lasso) using the following parameters.

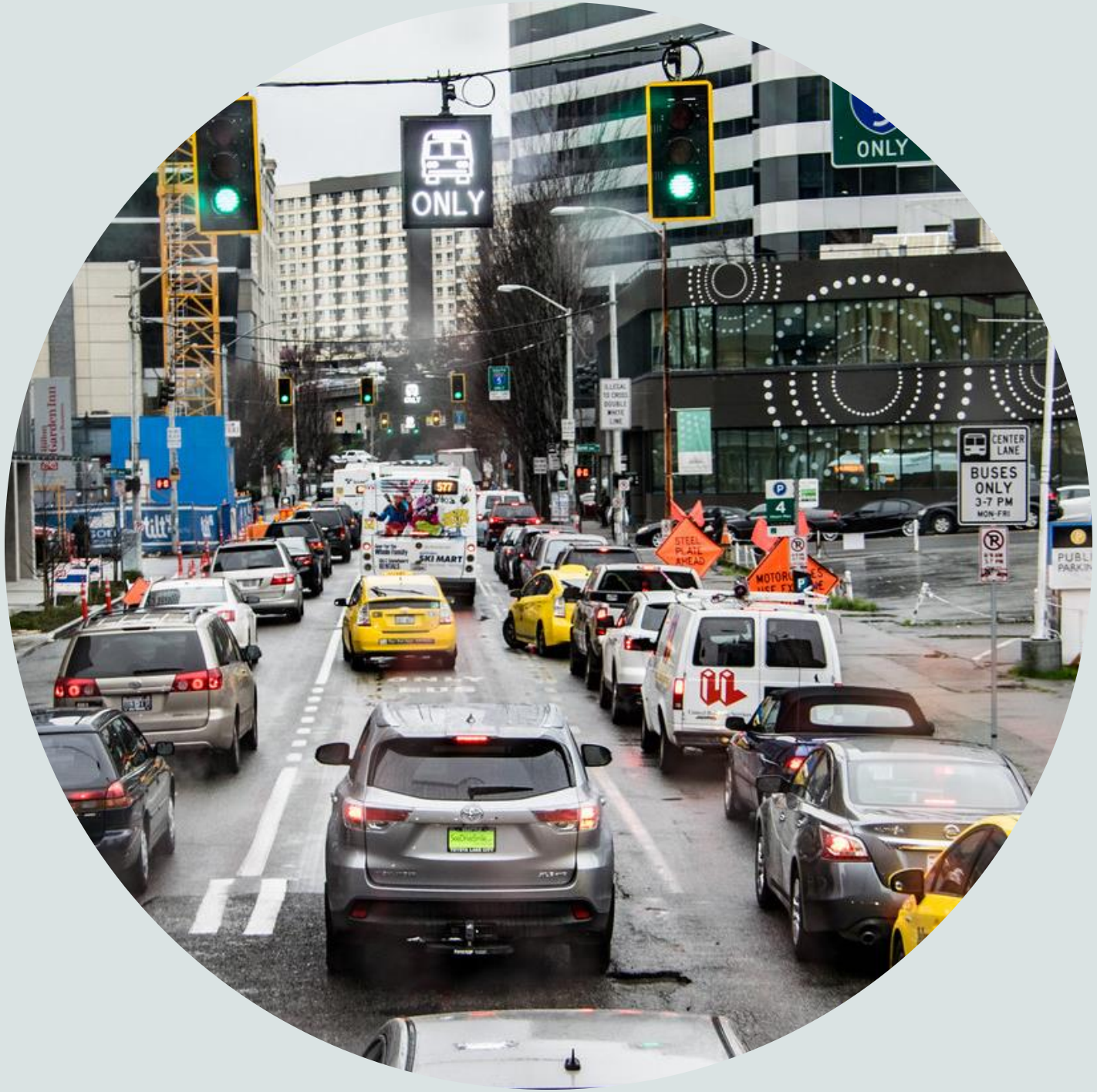
```
param_grid = {'fit_intercept' : [True,False],
              'normalize' : [True, False],
              'precompute' : ['auto',True,False],
              'selection' : ['cyclic','random'],
              'max_iter' : [1000,1500,2000,2500]}

rand_search = RandomizedSearchCV(estimator=LassoCV(),
                                 param_distributions=param_grid,
                                      n_iter=100)
```

Final Tuned Model

R2 Score for Tuned Lasso Regression Model: 0.9999989905373903
Mean Absolute Error for our Tuned Lasso Regression Model: 0.0003993377218317804
Mean Squared Error for our Tuned Lasso Regression Model: 2.578524406311133e-07
Root Mean Squared Error for our Lasso Linear Regression Model: 0.0005077917295812461

Thank You!!!


ONLY
ONLY
CENTER LANE
BUSES ONLY
3-7 PM
MON-FRI
STEEL PLATE AHEAD
SKI MART