Cars4U Project

Background & Context

- There is a huge demand for used cars in the Indian Market today. As sales of new cars have slowed down in the recent past, the pre-owned car market has continued to grow over the past years and is larger than the new car market now. Cars4U is a budding tech start-up that aims to find footholes in this market.
- In 2018-19, while new car sales were recorded at 3.6 million units, around 4 million second-hand cars were bought and sold. There is a slowdown in new car sales and that could mean that the demand is shifting towards the pre-owned market. In fact, some car sellers replace their old cars with pre-owned cars instead of buying new ones. Unlike new cars, where price and supply are fairly deterministic and managed by OEMs (Original Equipment Manufacturer / except for dealership level discounts which come into play only in the last stage of the customer journey), used cars are very different beasts with huge uncertainty in both pricing and supply. Keeping this in mind, the pricing scheme of these used cars becomes important in order to grow in the market.

Objective

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

Data Dictionary

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)

Exploratory Data Analysis(EDA) and Data Preprocessing

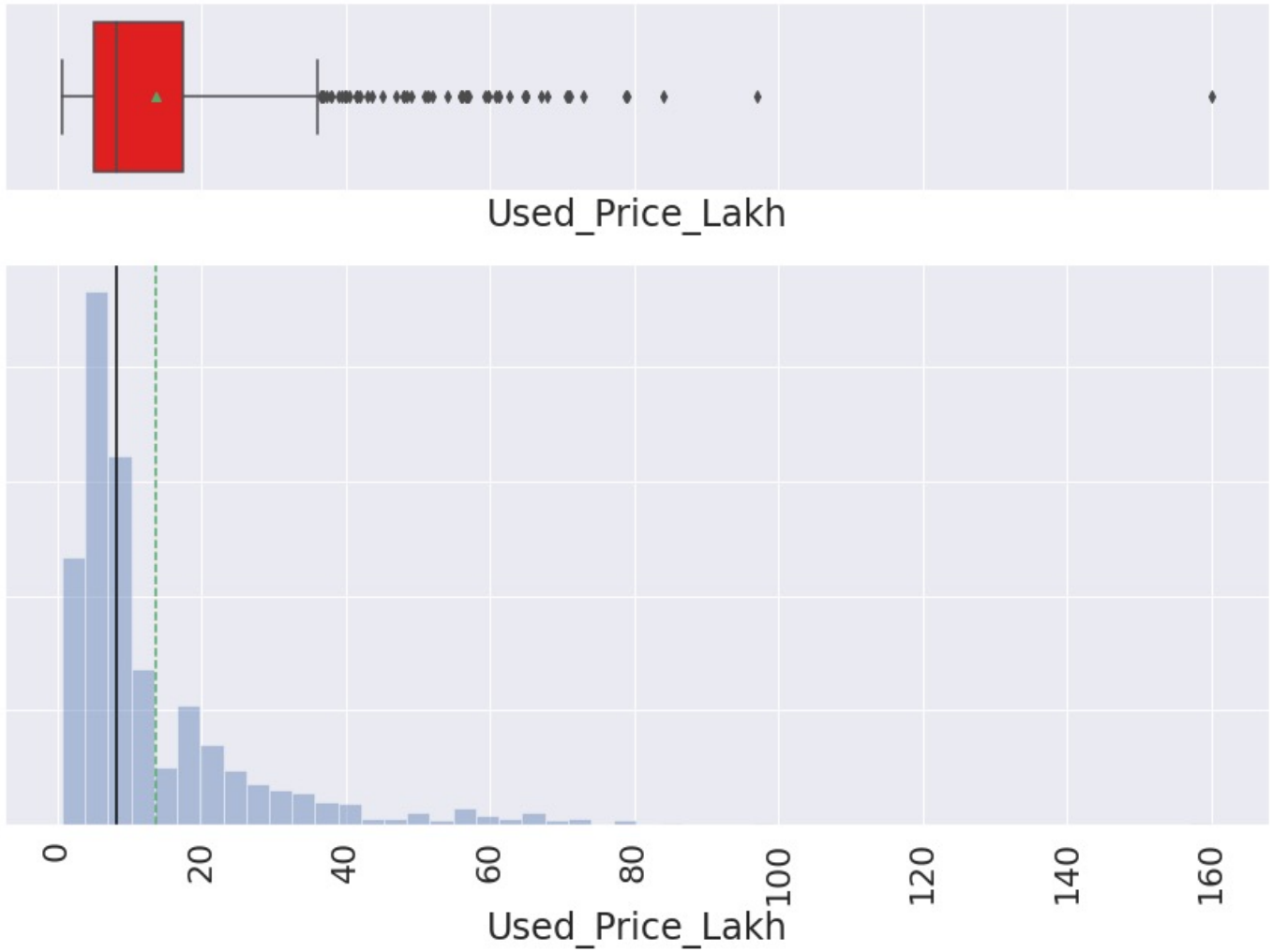
Univariate Analysis

- We will examine the numerical variables of each dataset

Univariate Analysis(Used Price Lakh)

- Used Price will act as our dependent variable
- The data is highly skewed to the left. This is due to the large amount of outliers.

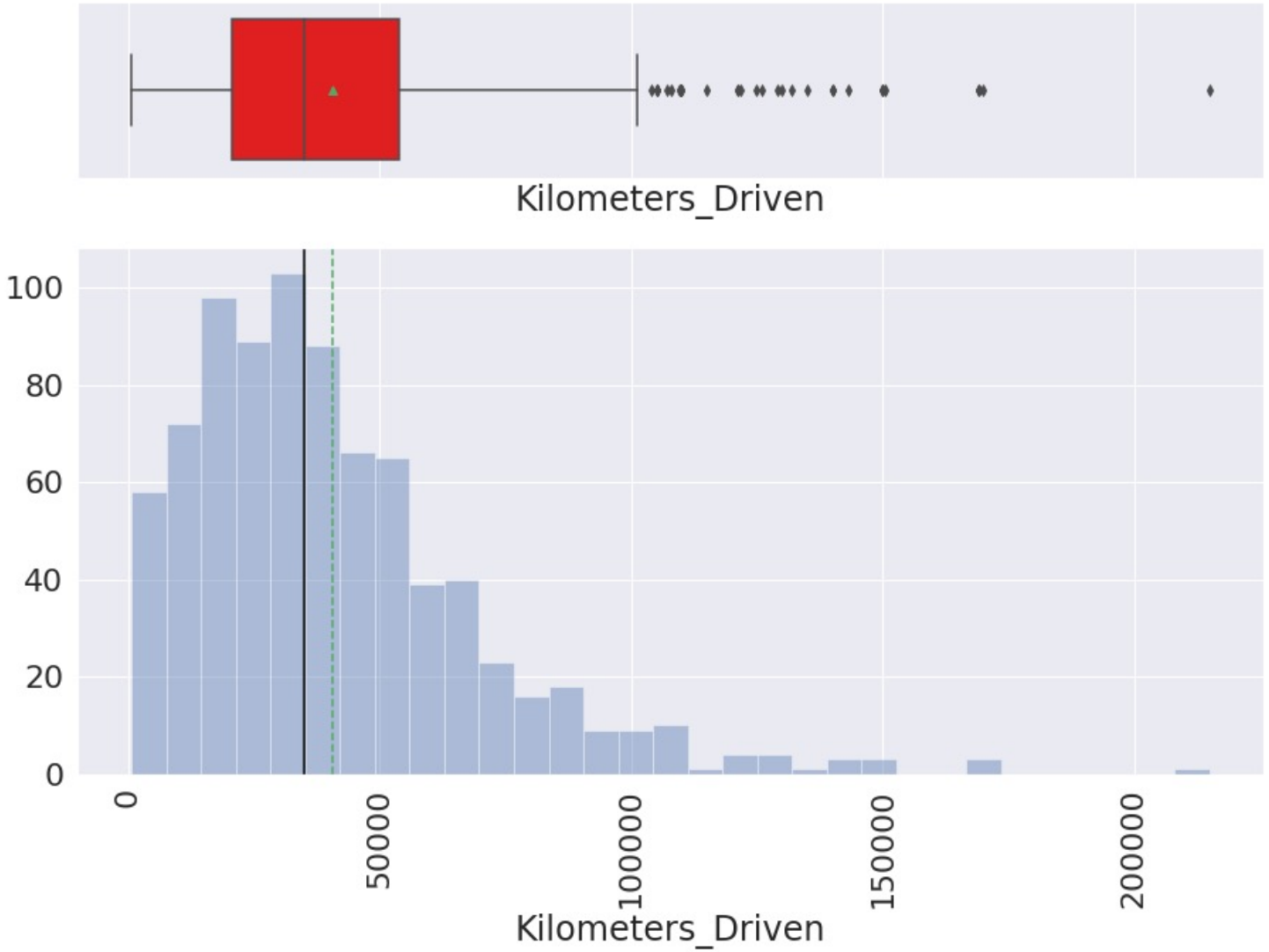
- We will treat these outliers because it could adversely affect our model



Univariate Analysis(Kilometers Driven)

- The data is highly skewed to the right.

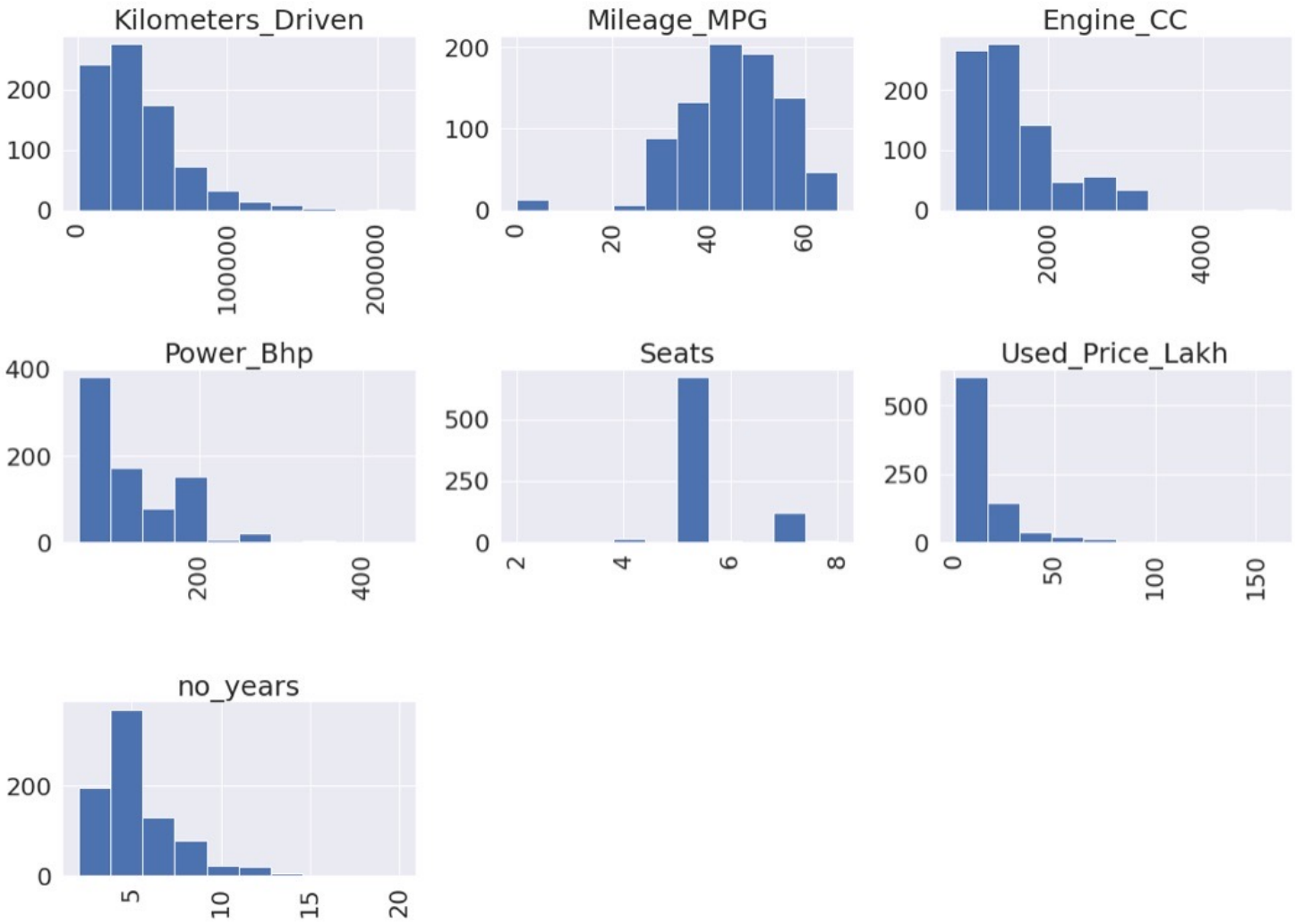
- There are a significant amount of outliers after 100000 km driven.
- The average kilometers driven is around 40498 and the median is 34895.



Univariate Analysis

Observation

- **Mileage and Seats** is somewhat normally distributed
- The other plots are skewed to the right
- Interpretation of left and right skewed:
  - **Engine data** is skewed to the right indicating the displacement volume of the engine is beyond 3000CC
  - **Power data** is skewed to the right indicating there are some observations where the maximum power of the engine is beyond 200Bhp
  - **No_years** is skewed to the right indicating there are some observations where the number of years owned is beyond 14



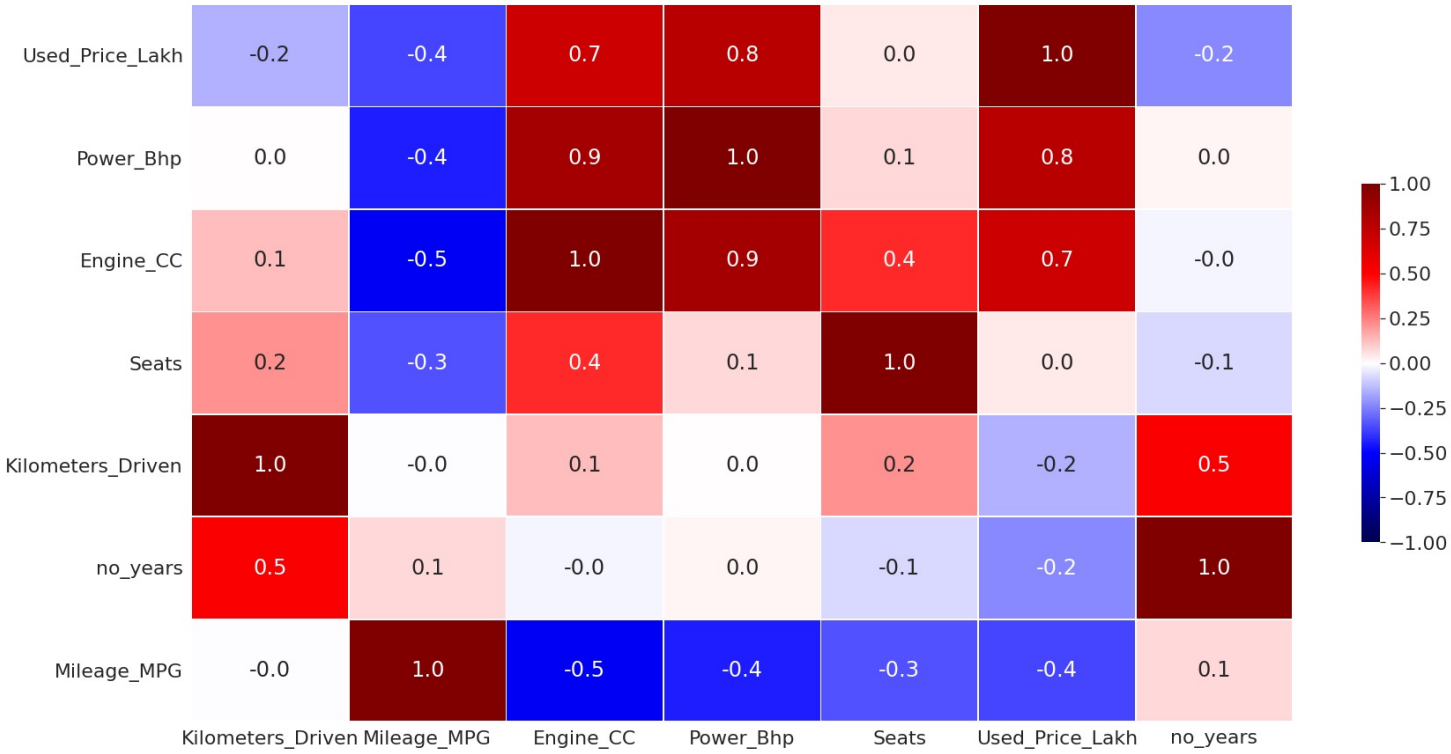
Exploratory Data Analysis(EDA)

Bivariate Analysis

Bivariate Analysis

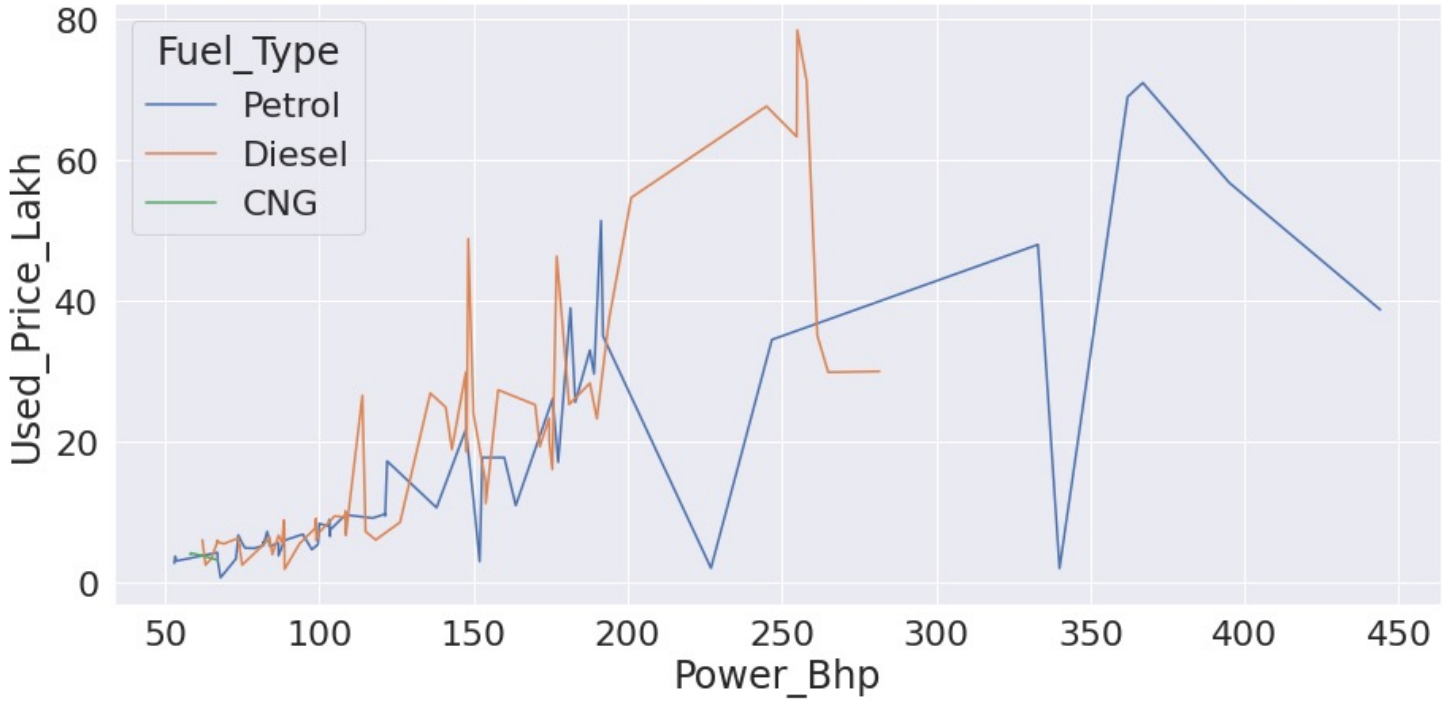
Observations

- The highest positive correlations are between the used price and engine, used price and power. This indicates that as maximum power and displacement volume of the vehicle's engine increase, so will the price.
- There is a low positive correlation between no of years and used prices.
- Mileage and kilometers driven are negatively correlated with used prices. That could be an indicator that the price decreases as the mileage and kilometers driven increase.
- Seats has a correlation of 0 with used price. This means we may have to drop it when fitting the linear regression model



Bivariate Analysis(Power vs Used Price vs Fuel Type)

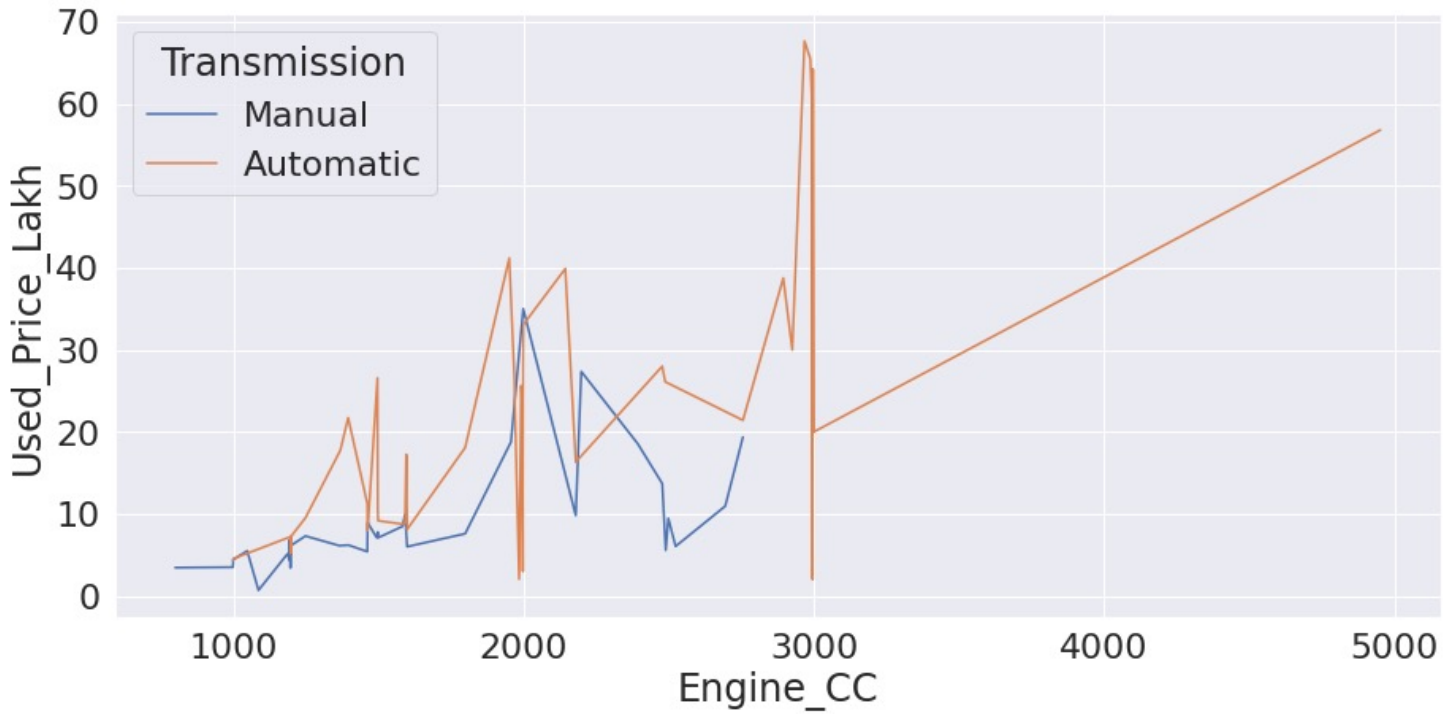
- **Observations**
- There is a positive correlation between the maximum power of the vehicle and its price.
- Diesel seems to have a stronger positive correlation with power and price compared to Petrol



Bivariate Analysis(Engine vs Used Price vs Transmission)

Observations

- The chart shows that the price increases along with the displacement of the engine's volume indicating a positive correlation.
- Automatic had the highest positive correlation when compared to Manual.



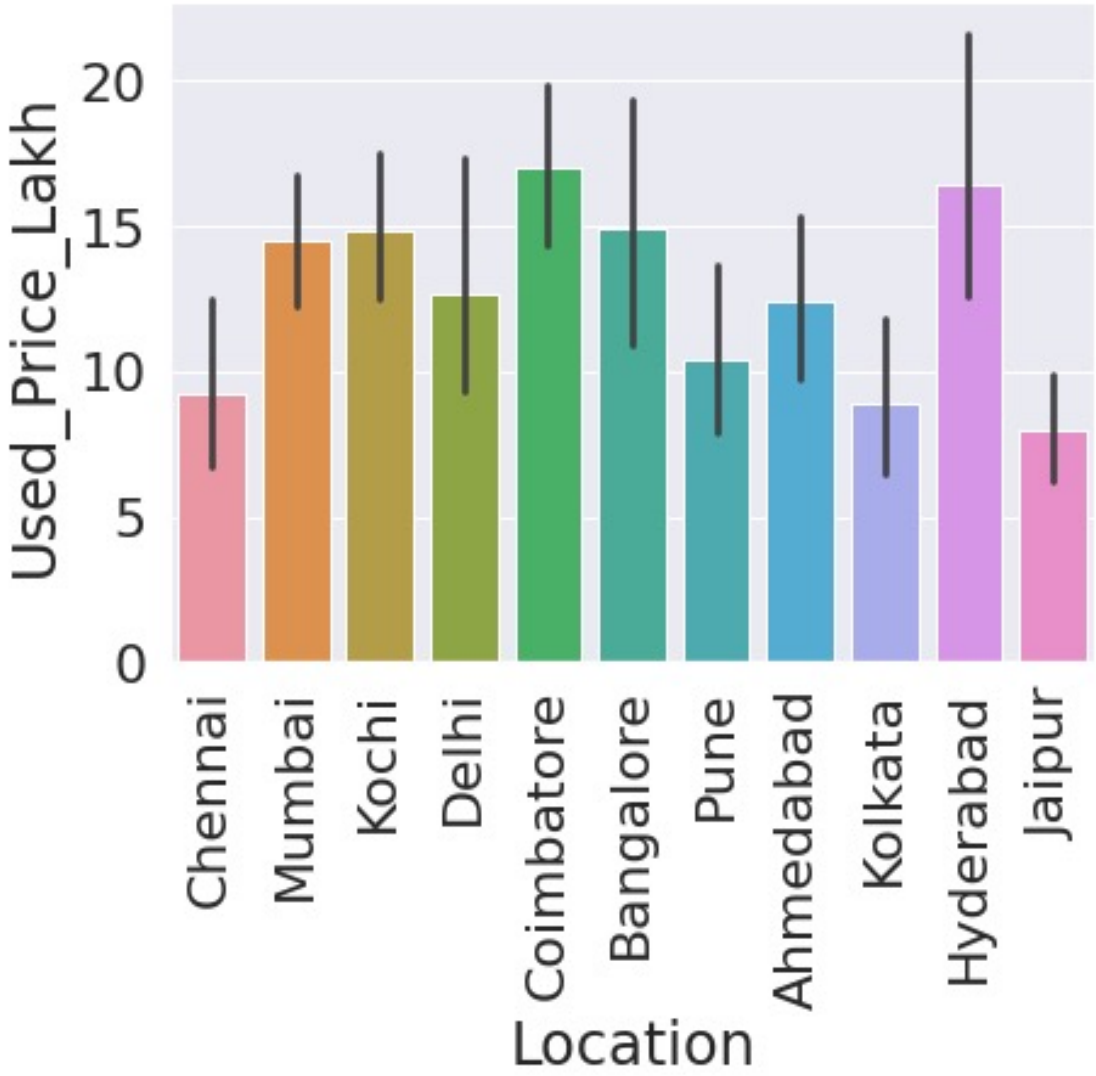
Exploratory Data Analysis(EDA)

Bivariate Analysis (Used Price vs Categorical variables)

Bivariate Analysis(Location vs Used Price)

Observations

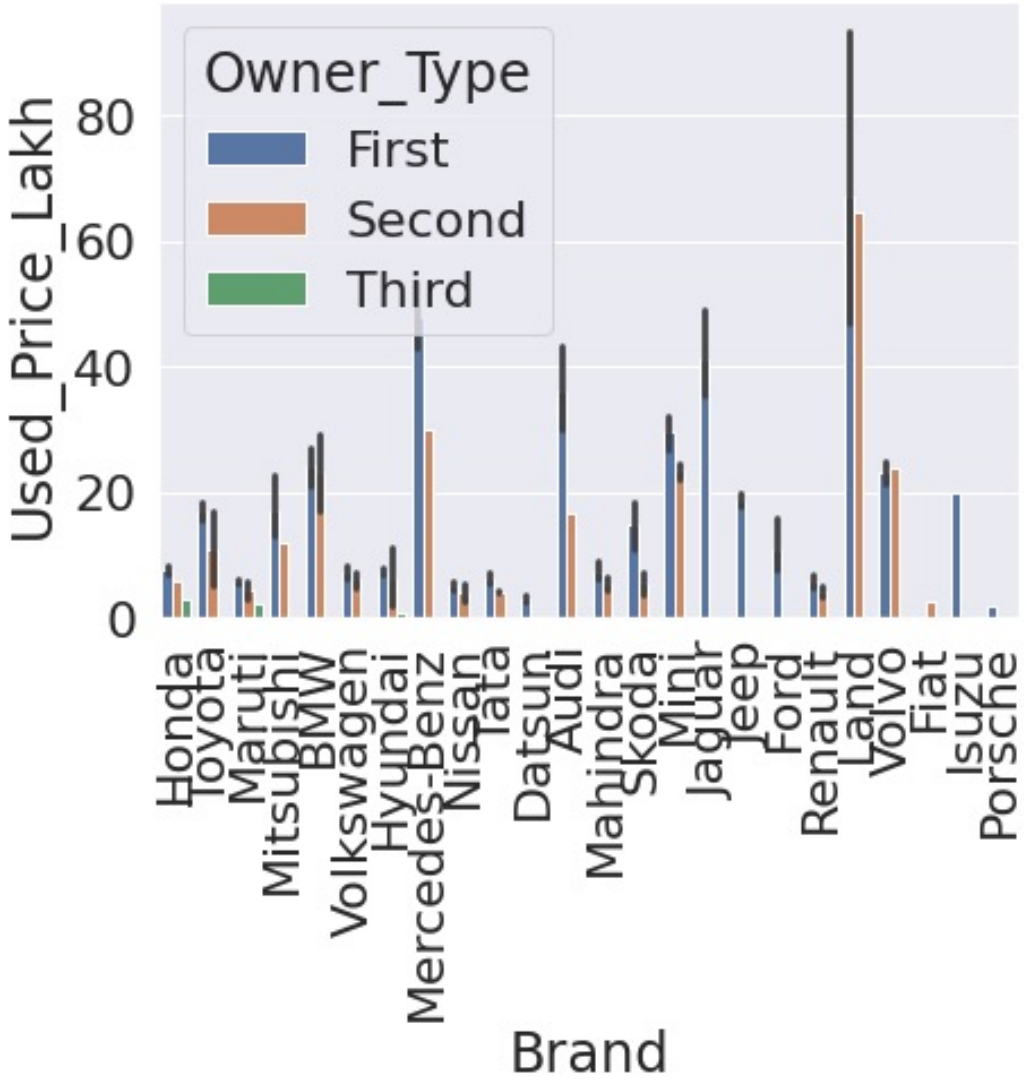
- The highest used car prices were competitive in Coimbatore, Bangalore, Ahmedabad and Delhi.
- Jalpur contains the lowest price.



Bivariate Analysis(Brand vs Used Price vs Owner Type)

Observations

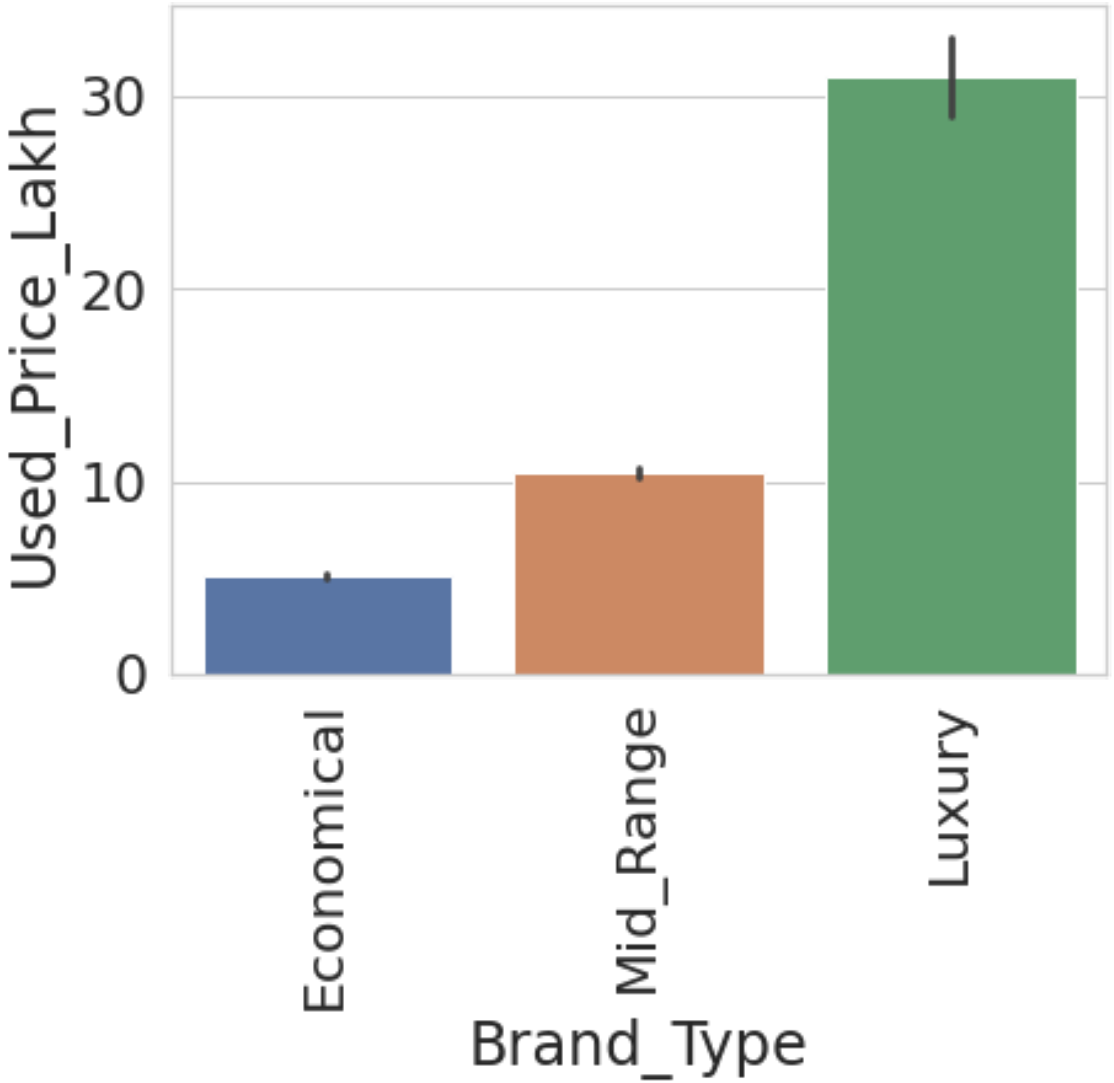
- Land had the highest prices among first and second car owners.
- Prices for first and second car owners for the BMW, Volvo and Toyota
- We can further classify the column brands based on their level of price using binning



Bivariate Analysis(Brand Type vs Used Price)

Observations

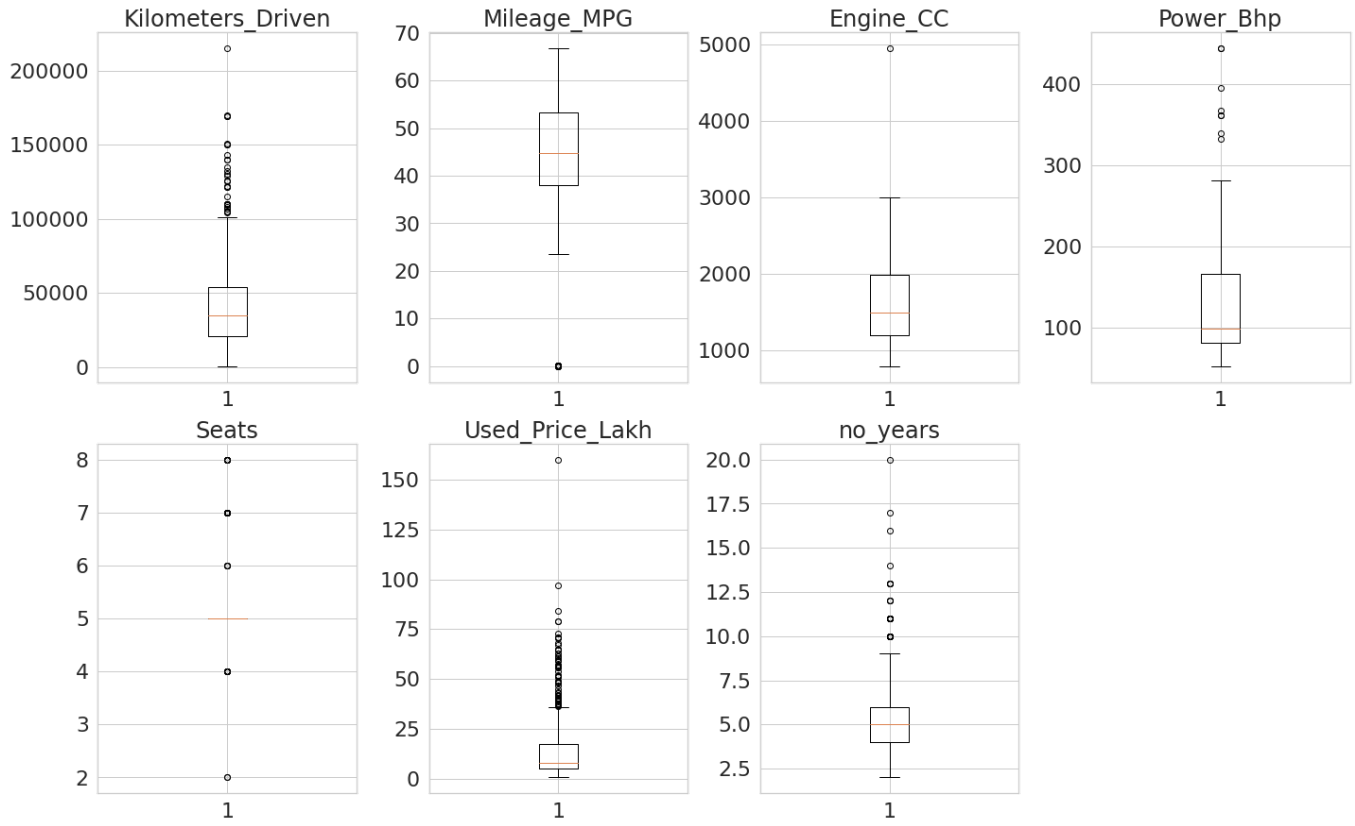
- Luxury vehicles obviously have the highest prices followed by the Mid Range and then Economical types



Examining Outliers

Observations

- The small circles in the upper and lower areas represent the outliers.
- Used Price, Lakh, Kilometers Driven, Power Bhp, No. years have the most significant outliers in the upper area.
- Engine has upper outliers as well, but they are much less significant.
- Seats has outliers in the upper and lower areas.
- Mileage is the only one with outliers in the lower area.
- Treating these outliers could create a more accurate model



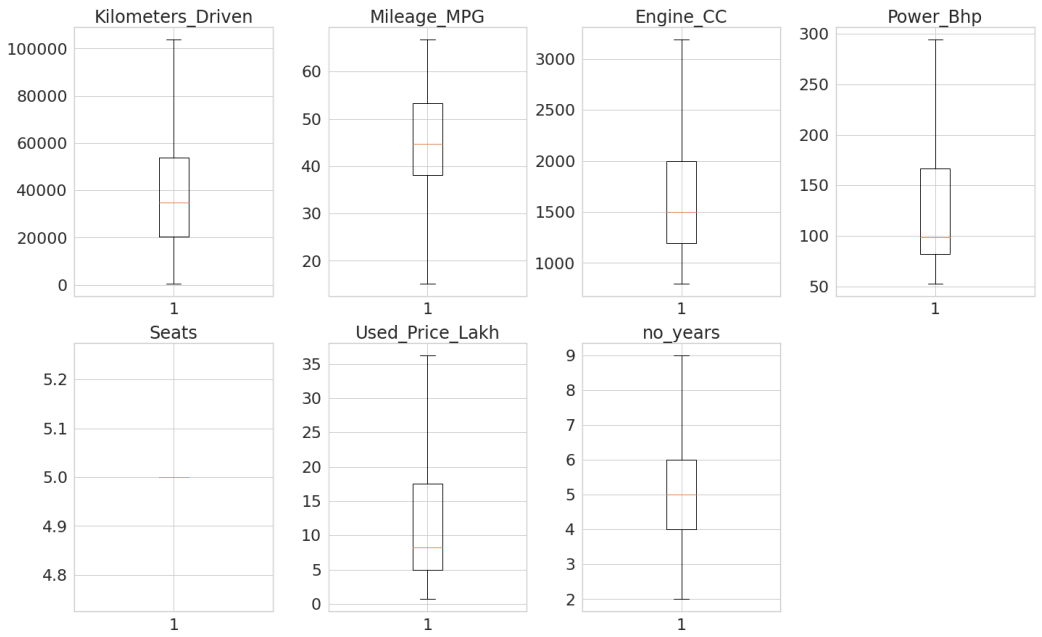
Outlier Treatment

We treated the outliers in several steps

1) We multiply 1.5 and IQR and subtract it from the 1st quantile
2) We multiply 1.5 and IQR and add it to the 3rd quantile
3) This will remove any outliers on the outsides of the lower and upper whiskers.

Observations

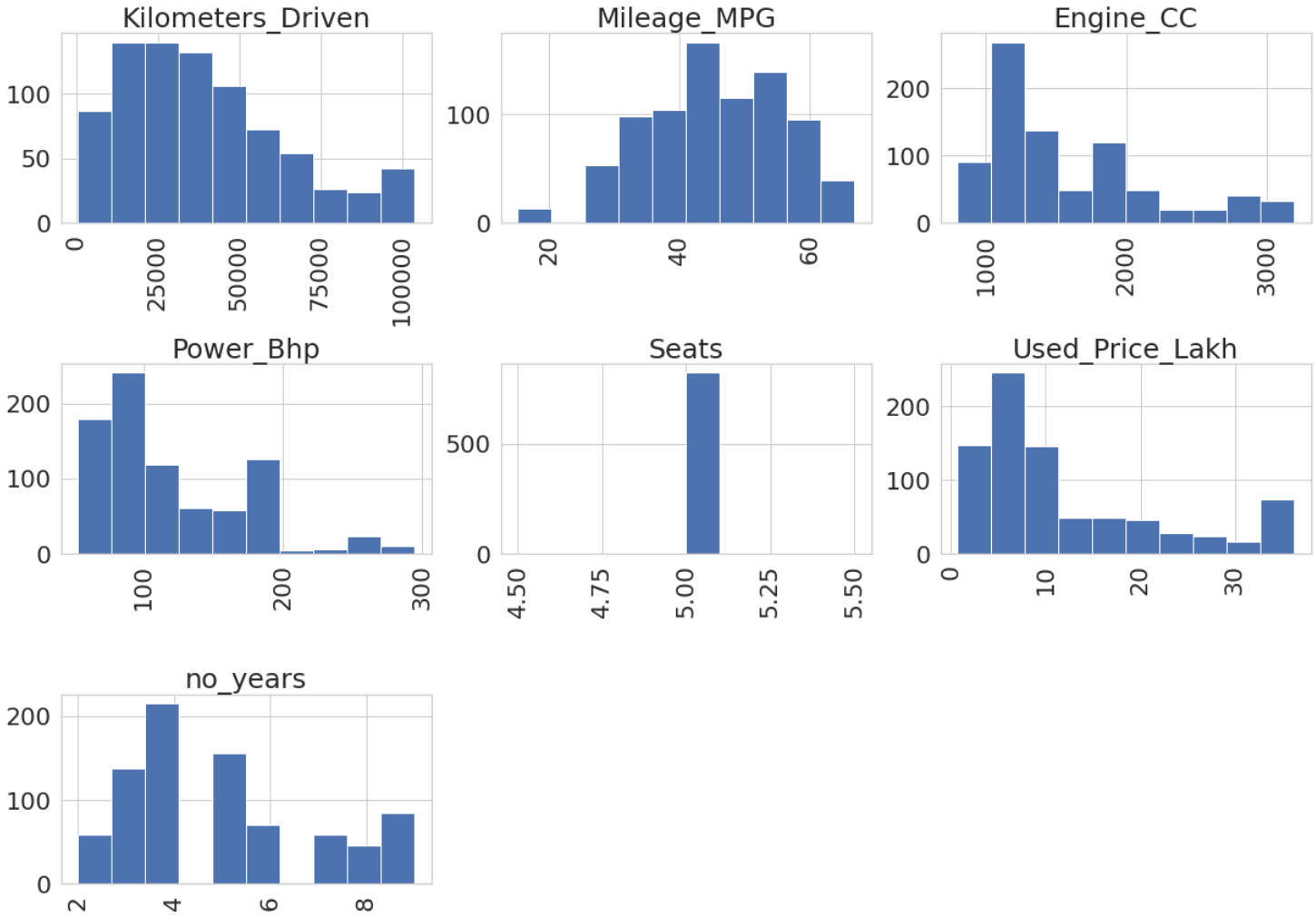
The small circles that represented the outliers are no longer visible



Log Transformations

Purpose of Log Transformations

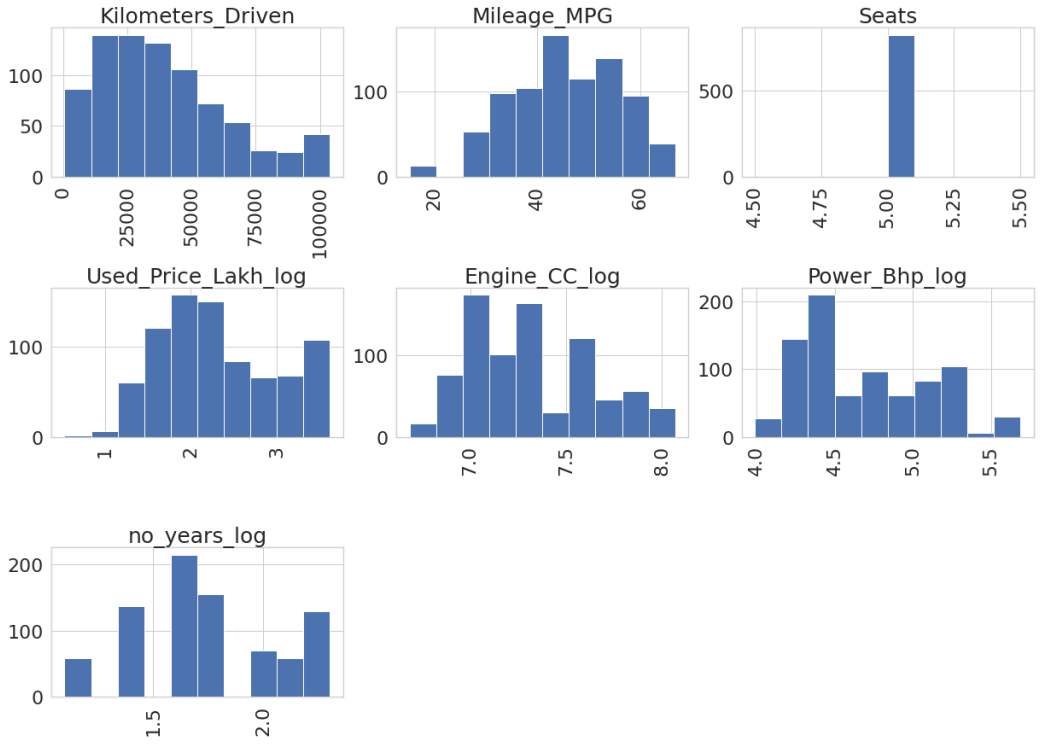
- **Log Transformation is used to make a column appear more normally distributed**
- **Observations**
- Some of the numerical columns are not normally distributed. Used Prices, Engine, Power Bhp and no of years are still skewed.
- We will use log transformation to make these columns appear more normally distributed.



Post Log Transformation

Observations

- Used Prices, Engine, Power Bhp and no of years appear to be more normally distributed.
- Now we can prepare the model.



Data Preparation

Model Performance

- First we will define the x and y variables
- For X variable we will drop Used Price, Brand and Seats.
- For y Used Price will be assigned
- We will also use pandas function get dummies to make the categorical variables