Context:

The used cars market contains uncertainties in pricing and supply. The sellers are in a difficult position to predict the actual price of the car as several factors such as mileage, brand, model, year can influence it. This issue is important to solve as it ensures the sellers are receiving accurate and fair benefits of selling their vehicle. It also protects the market growth.

The Objective:

Goal is to come up with an effective pricing model that can accurately predict the price of used cars and can help the business in devising profitable strategies using differential pricing

The key questions:

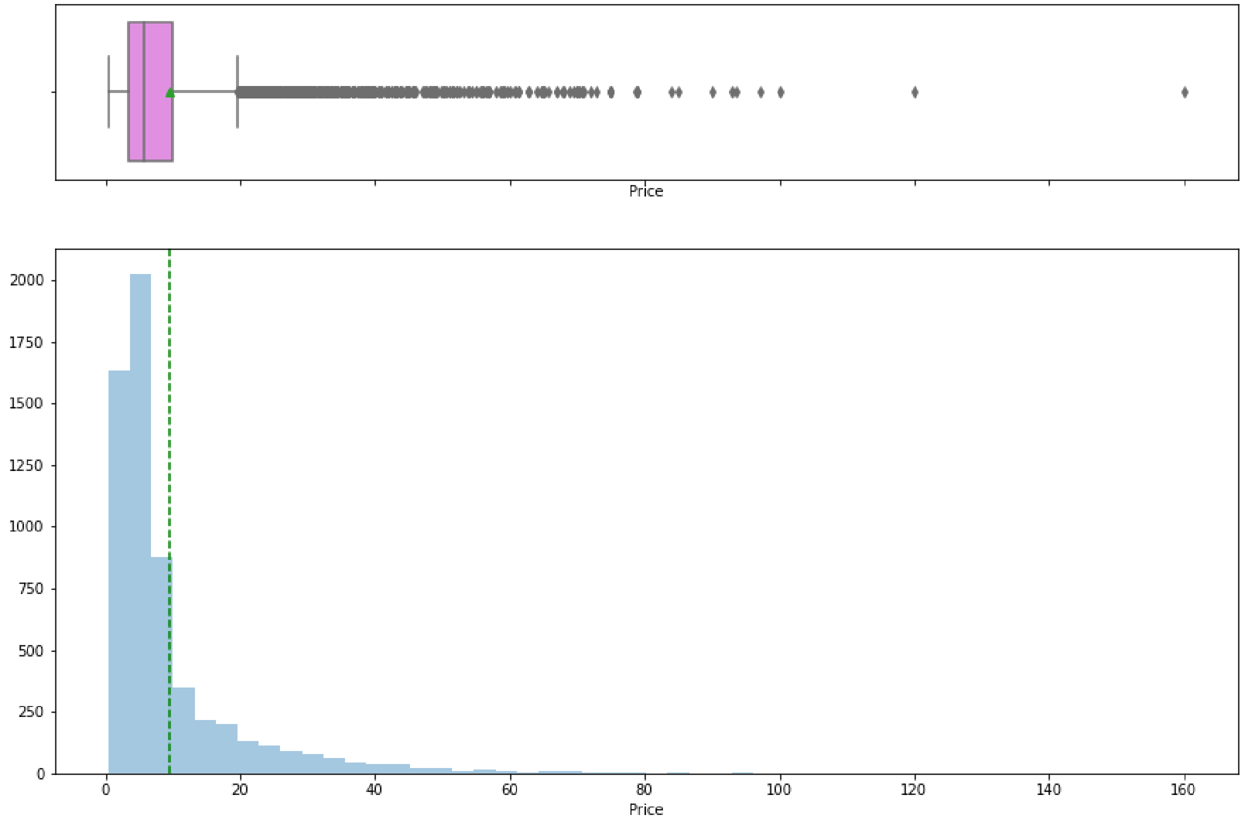
- The price of new vehicles is missing for many entries in the dataset. How to handle these missing values?
- Do we go with the simple linear regression approach or a random forest?

The problem formulation:¶

- We must build a regression model along with basic data preprocessing, Feature engineering, treating the missing values, Univariate analysis and model tuning in order to select the optimal model for calculating the price of any available used car.
- To develop this model, we need to identify the most important features in the dataset.

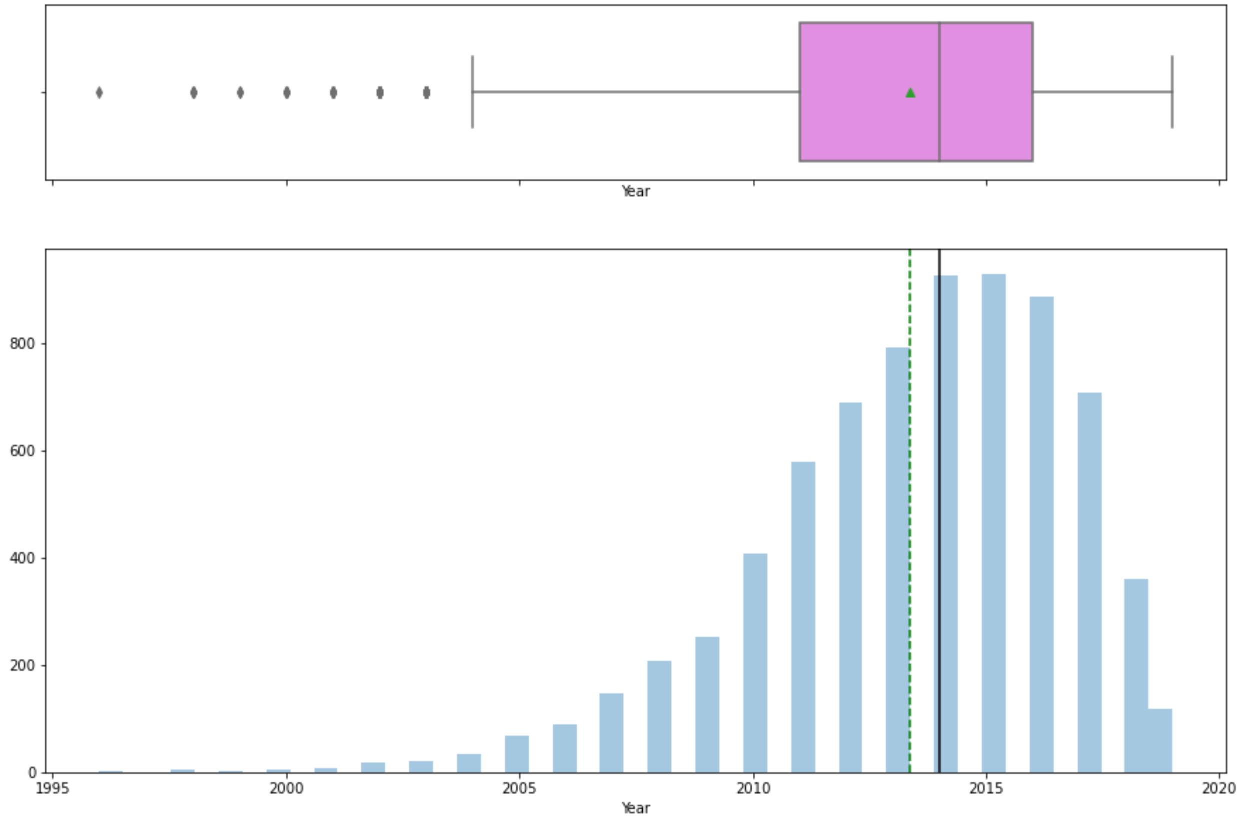
Observations & Insights

- There are some discrepancies and missing values in the dataset for ‘Kilometers_Driven’ and ‘Mileage’.
- Some of the attributes like Kilometers_Driven are highly skewed in the data



- Cars manufactured during 2011-2016 are more in the used cars market
- First hand owners are selling their cars more. This means second and third hand cars cannot be priced higher for sale and would not benefit the customer selling it
- Most people are selling the Manual transmission vehicles
- Missing values in seats column can be filled by using the Brand name and Model of the car
- Engine and Power have some manual entry errors
- The standard deviation of Kilometers_Driven and Engine volume are high that indicates these values are more spread out.
- More used cars present in the metropolitan and cosmopolitan cities
- Mahindra XUVs are more sold in Mumbai in used car market
- Similar type of transmission and owner_type are more frequent

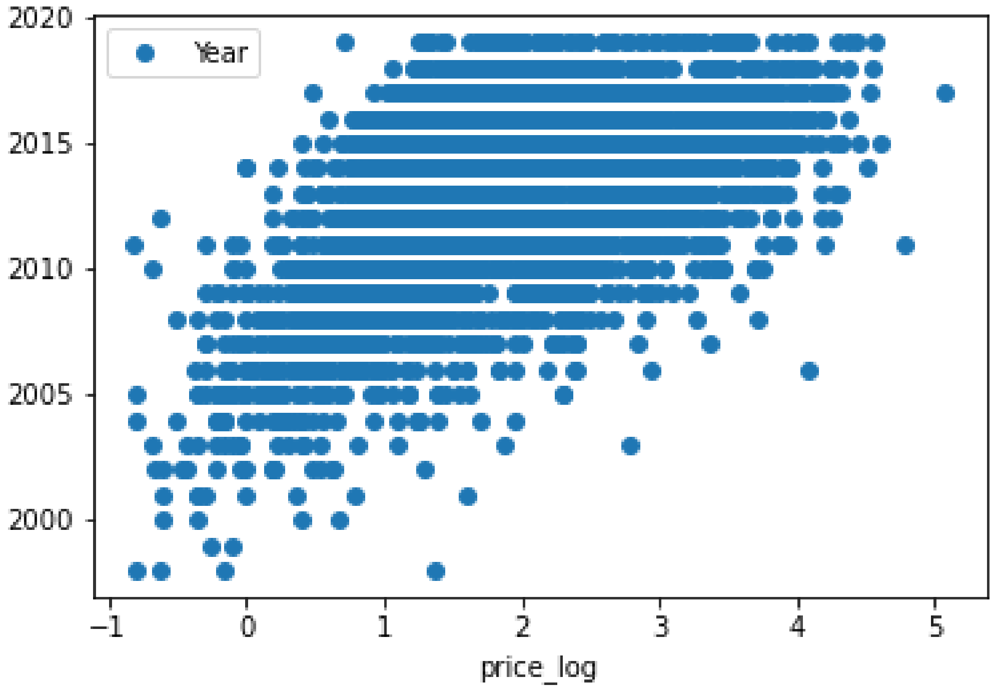
- Diesel vehicles tends to be more driven and they are used by a single customer for a long time before they sell it. And they have the highest engine power
- Enginer power is also highly rightly skewed so we can apply a log transformation on power.
- There are few outliers in Year which doesn't have any significance, especially entries before year 2007 would not make any significance. So we can ignore them.



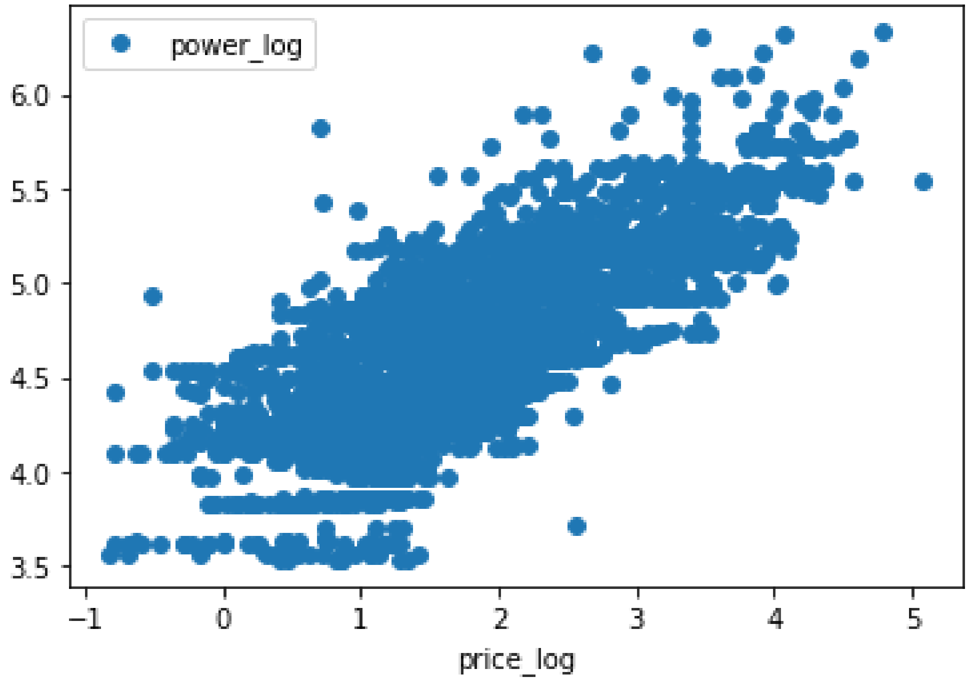
- Number of seats is not uniformly distributed with most of the data around 5 seats.
- Mumbai has the highest percentage of used cars for sale in market
- The price is higher for more recent cars. Meaning old cars are less priced comparitively with some outliers
- There doesn't seem to be a direct relationship between Kilometers driven and the price. Most of the cars are driven more than 80K

kilometers but their price seems to be centered around 100,000 to 300,000

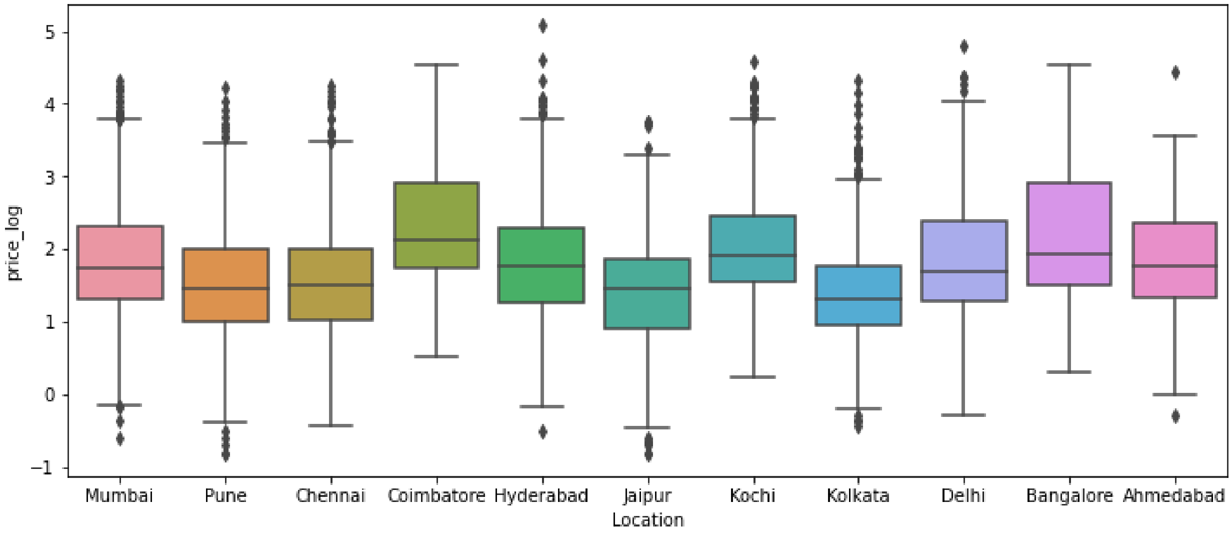
- Cars with more engine power tends to be expensive
- Engine volume doesnt seem to impact the price of the car
- Some Luxurious cars with limited seats are priced higher which can be ignored
- Year and Price are positively correlated. Mileage and Year also have positive correlation.



- Mileage and Power, Mileage and Engine Volume have a significant negative correlation
- Price and Power, Price and engine volume, Price and New car price have the most significant positive correlation



- Vehicle price is higher in cities like Banglore and Coimbatore. Bangalore being a silicon valley of India with more population needing cars for their daily commute. Coimbatore is developing as an IT Hub and more educated people are migrating there.



- Diesel and Automatic transmission vehicles are expensive.
- First and Second hand vehicles are priced more than others.