CARS 4U

Data Analysis

Objective

- Come up with a pricing model that can effectively predict the price of used cars.
- Help business in devising profitable strategies using differential prices.
- To extract actionable insights from the available data by leveraging customer information .
- Explore and generate value out of company's data.
- Perform uni-variate and multi-variate analysis.

Data information

Variables	Description
S.No	Serial Number
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 Citie
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

Variables	Description
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.
Price	The price of the used car in INR Lakhs

Observations	Variables
94289	13

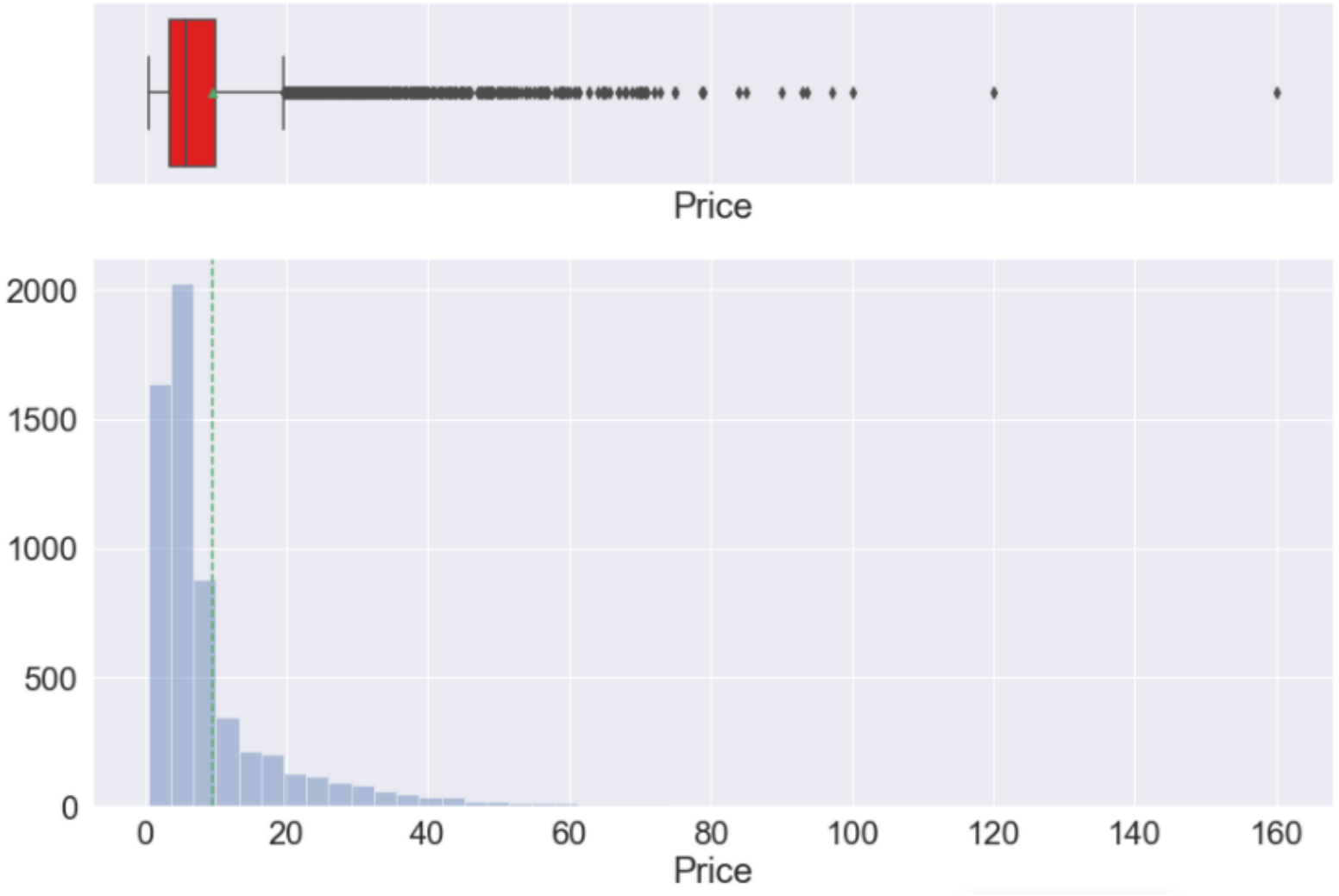
Notes:
1 duplicated observation was found on the data

Missing Values and Outliers

- It was found a representative number of Outliers, 7481 datapoints and the following assumptions were made to treat them.
  - **Variable 'Price':** it was found 1234 missing values. As this variable would be target variable, in other words Price would be the goal we want to predict, models were created excluding missing values of this variable and in a later stage they were filled out by using the model to predict their value.
  - **Variable 'New Price':** it was found 6247 missing values. This variable was not taken in consideration due to its big number of missing values. Some attempts to better understand this variable by looking each entry was done but the amount of variability within the variable made it hard to consider to impute values based on the available data.

- A great number of outliers were found within numerical variables
  - It was used a technique called flooring and capping where all the extreme values were brought to their lower whisker, which represents 1.5 times the lower quantile (.25) and in the upper side values were brought to upper whisker which represents 1.5 times the upper quantile (.75) in order to reduce the spread of the data and increase the accuracy of the models.

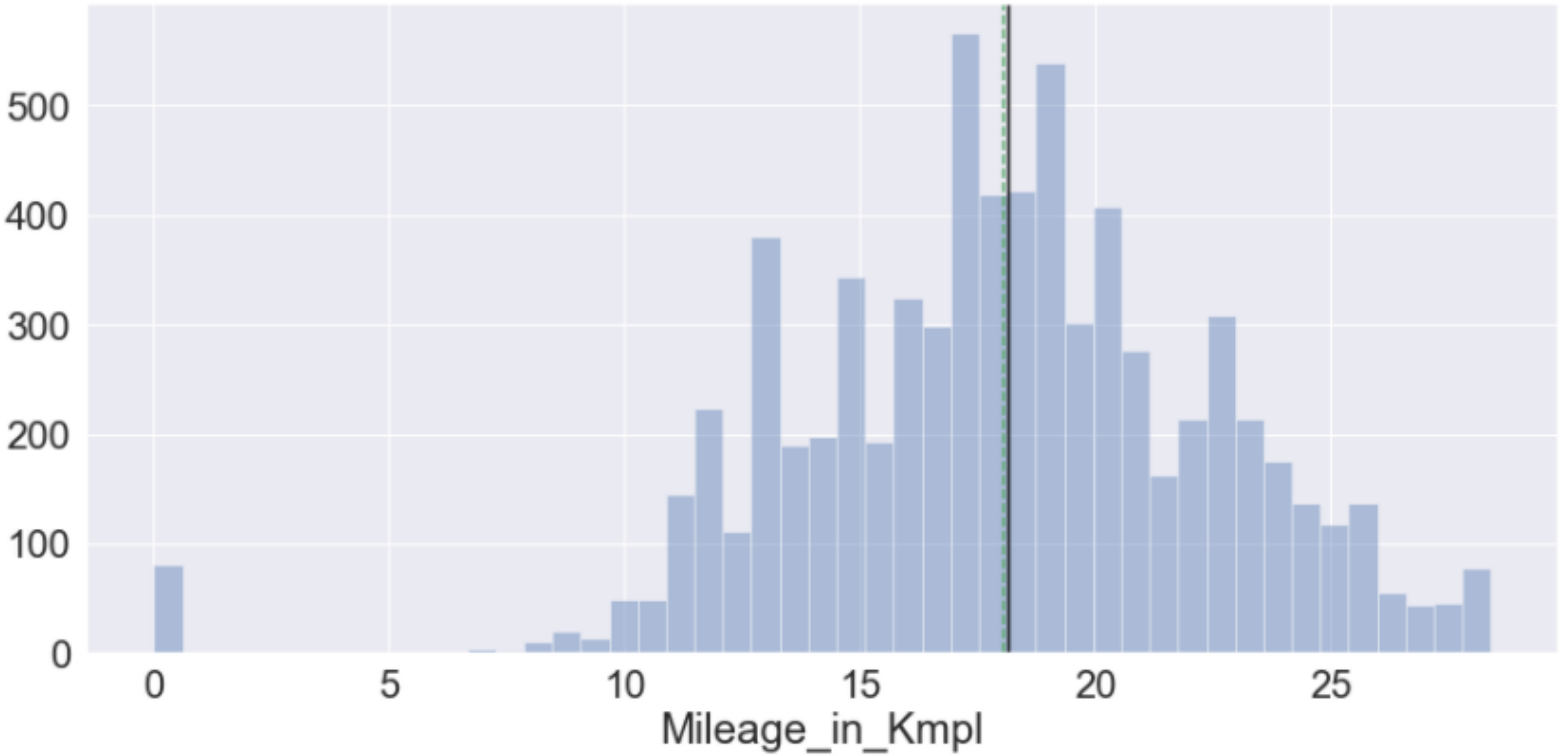
Mean (– – –)
Median (———)



EDA – Univariate Analysis

- Price is right skewed distributed which means a big concentration of cars are below mean, 10 Lakh.
- A big number of outliers could be found in this variable.
- Its is evident that outliers are highly impacting the mean.

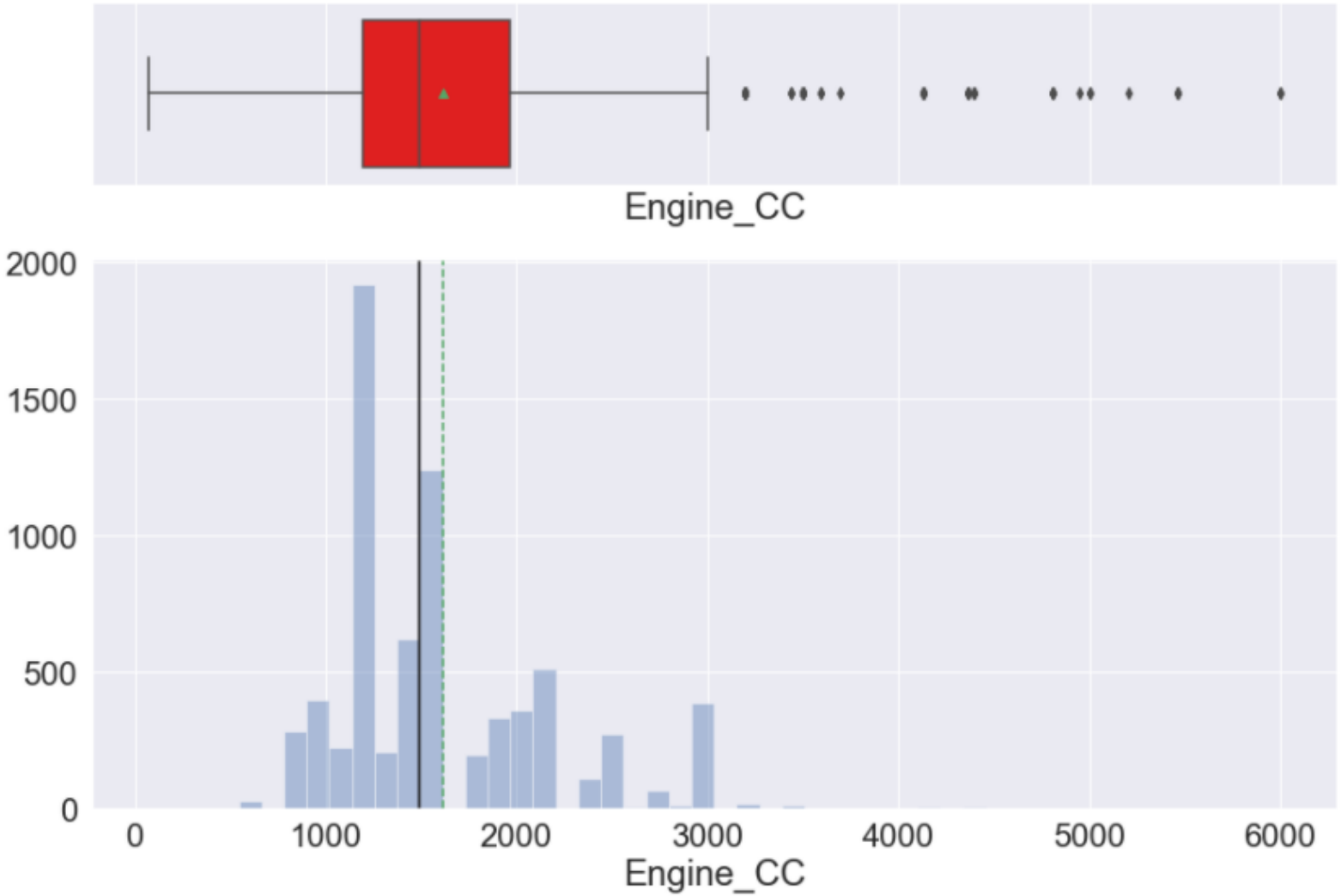
Mean (– – –)
Median (———)



EDA – Univariate Analysis

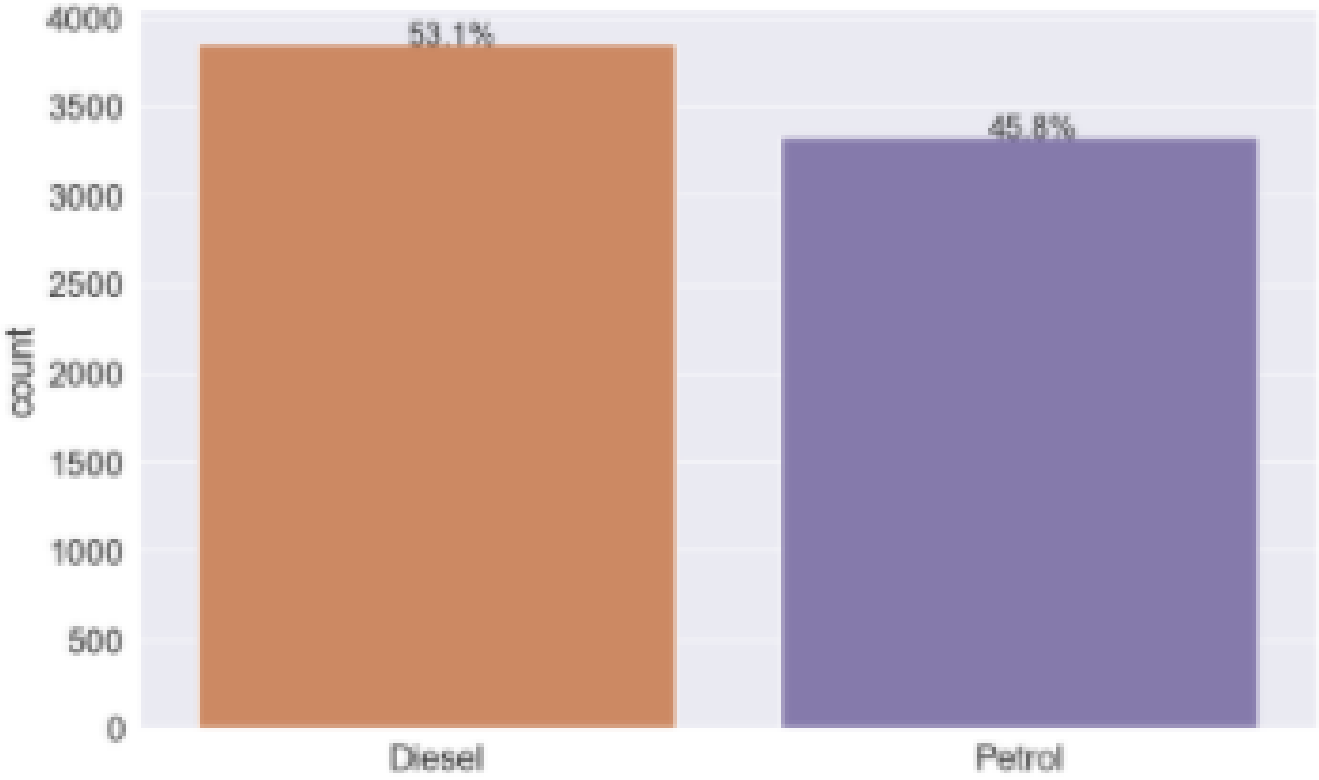
- Mileage is normally distributed, but it can be observed that there are Outliers on the lower scale.

Mean (– – –)
Median (———)



EDA – Univariate Analysis

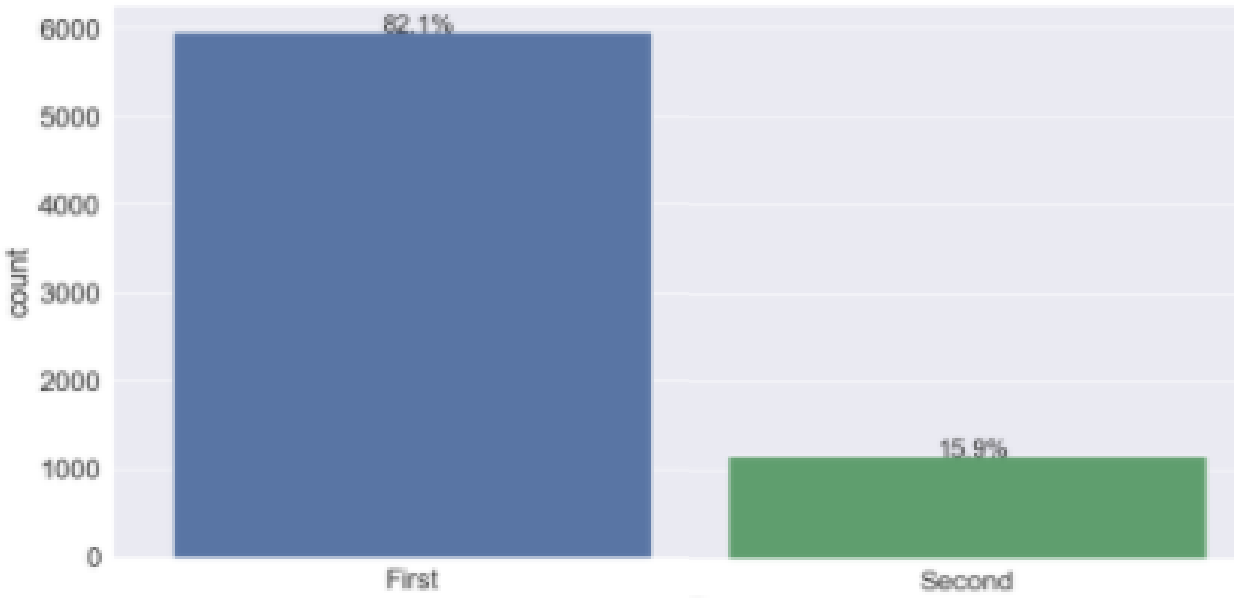
- Engine CC behaves more as categories. For instance, from 800 to 1500, from 1500 to 2300 and so one.



Fuel Type

EDA – Univariate Analysis

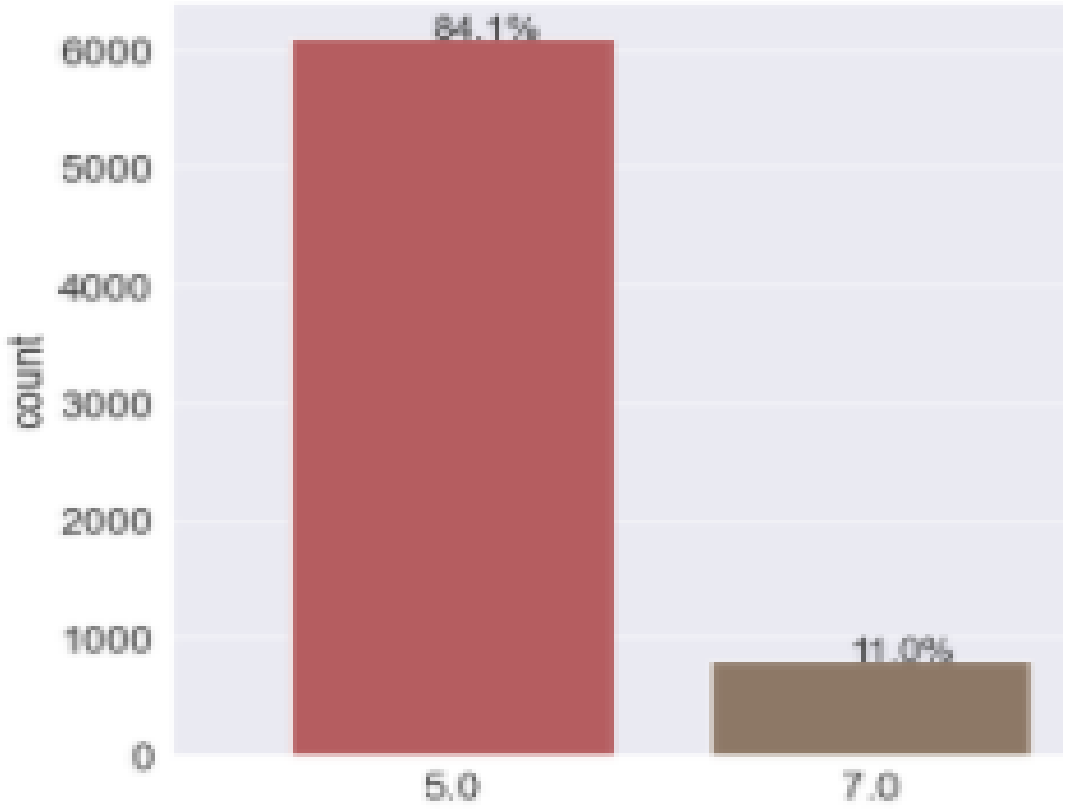
- 98.9% of the cars from the data have their fuel type as either Diesel or Petrol with diesel having the bigger share 53.1%.
- Remaning 1.1% are grouped between Eletric, LPG and CNG.



Owner

EDA – Univariate Analysis

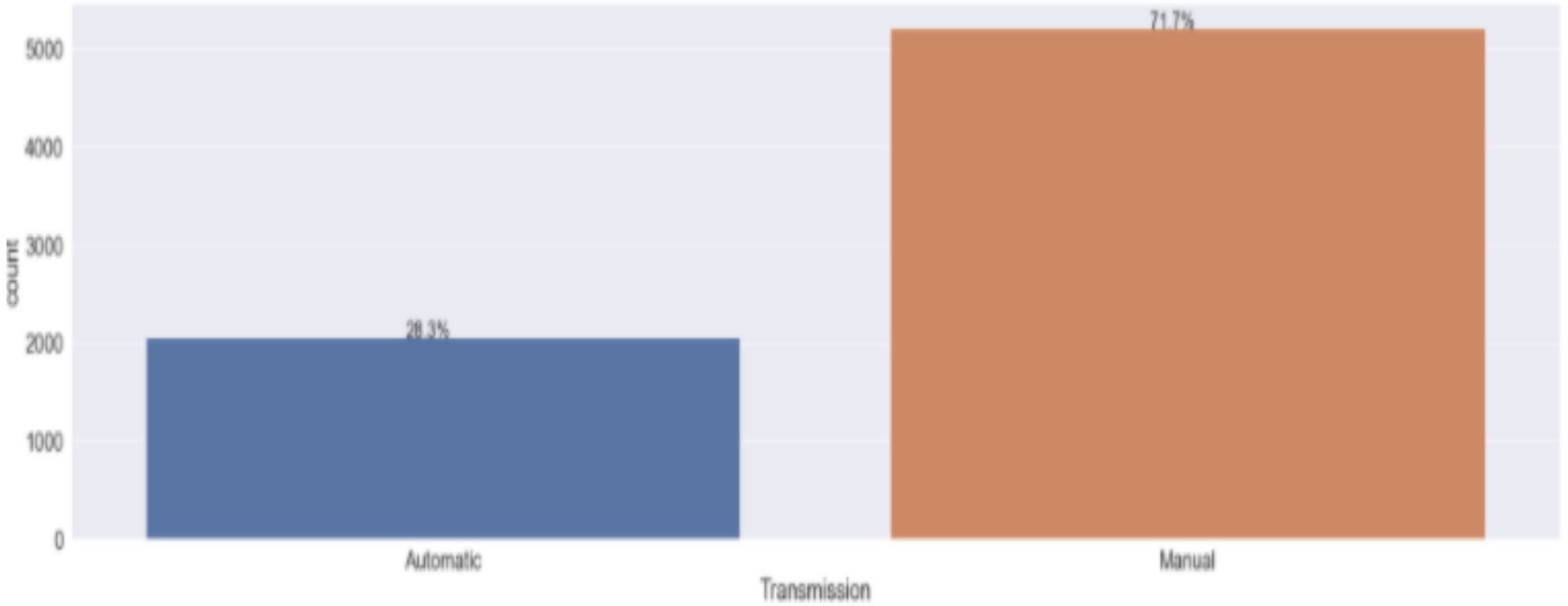
- 82.1% of cars were first owners, 15.9% second owners and the remaining 2% spread between third and forth owner.



Quantity of seats

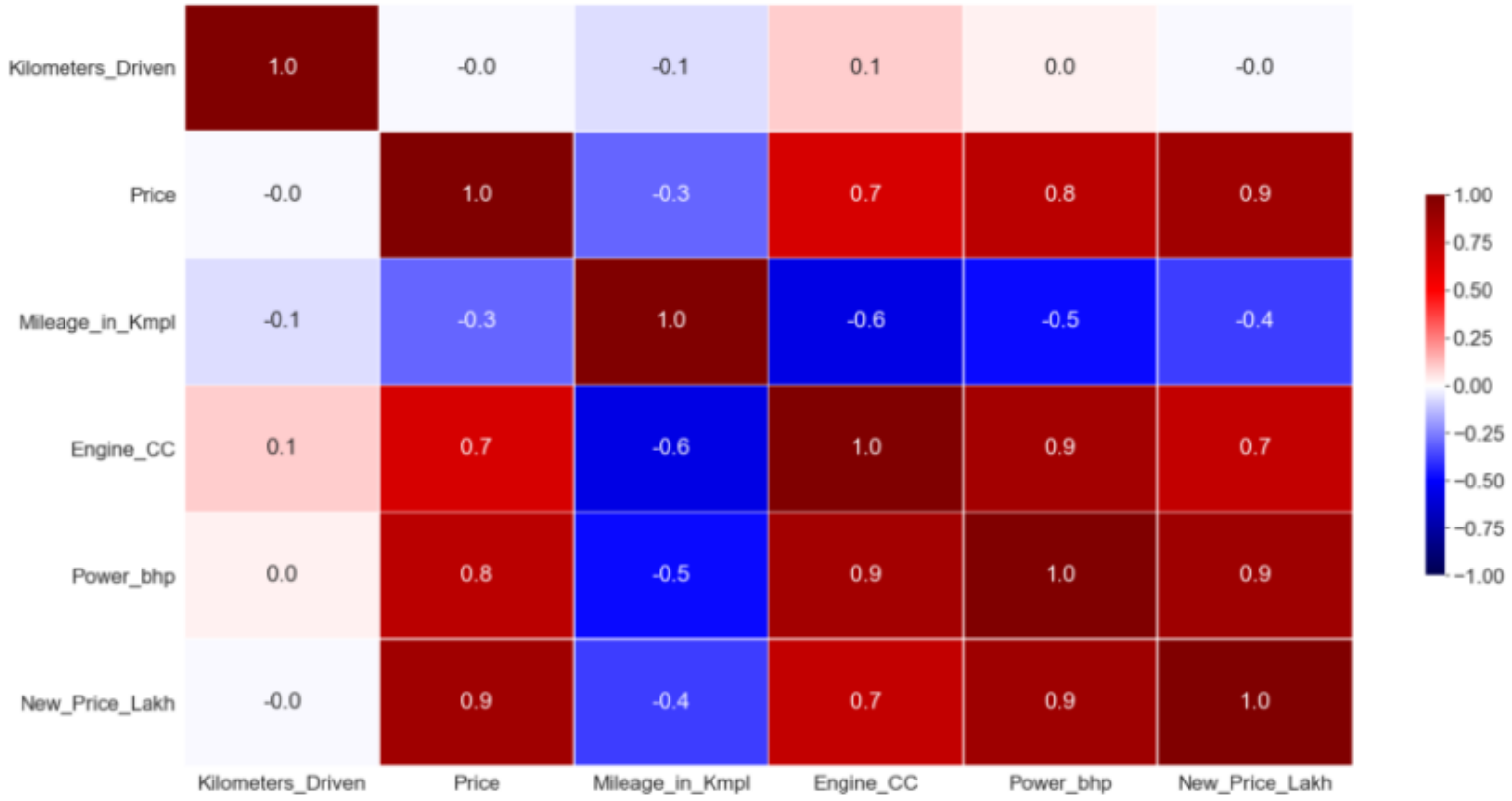
EDA – Univariate Analysis

- 95.1% of the cars were 5 or 7 seat cars. With 5 seats representing 84.1%.
- Remaining 4.9% were divided in 8 seats with 2.3% and the remaining 2.6% spread among 2, 4, 10 seats.



EDA – Univariate Analysis

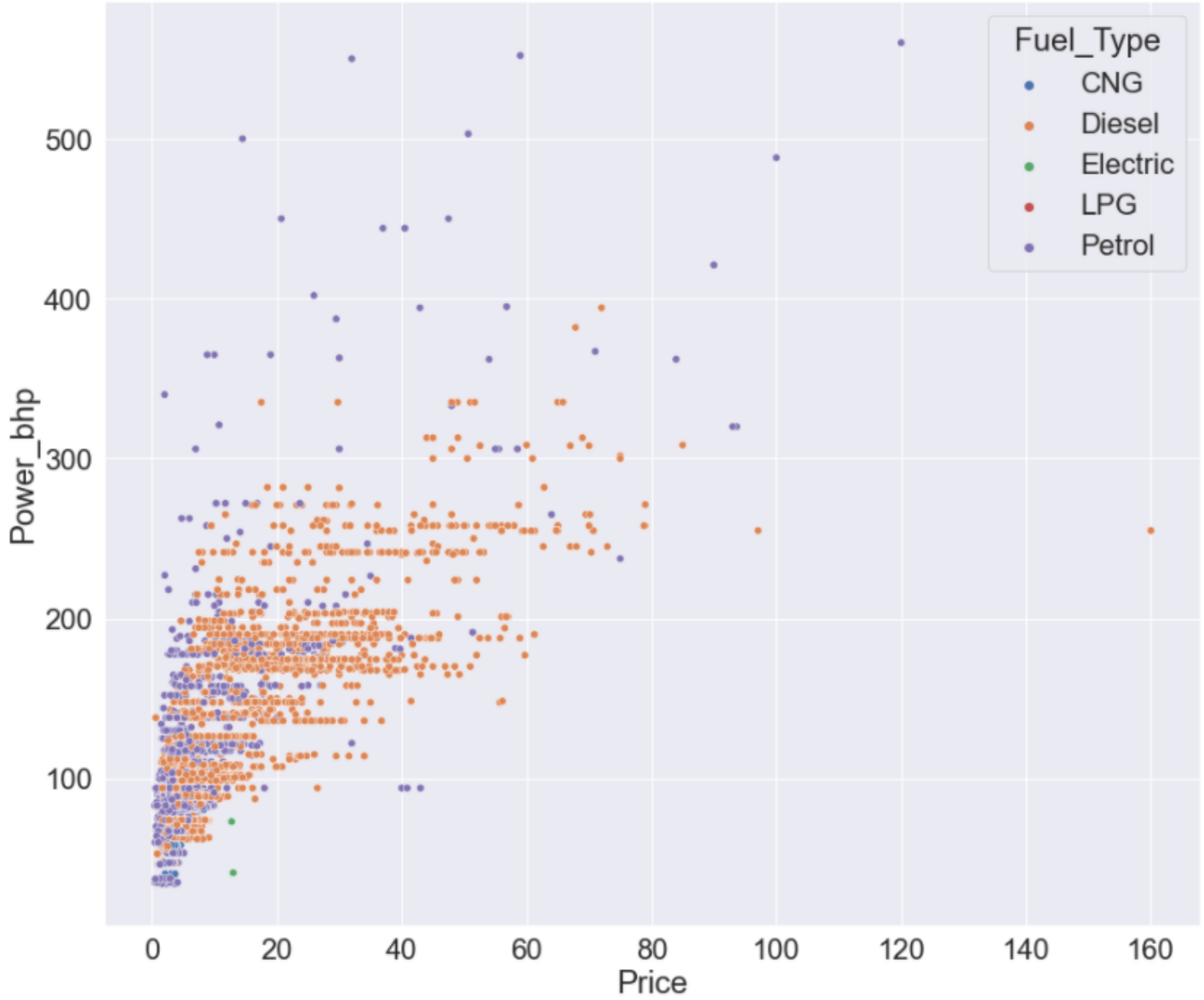
- 71.7% of the cars have Manual transmission .



EDA – Correlation Matrix

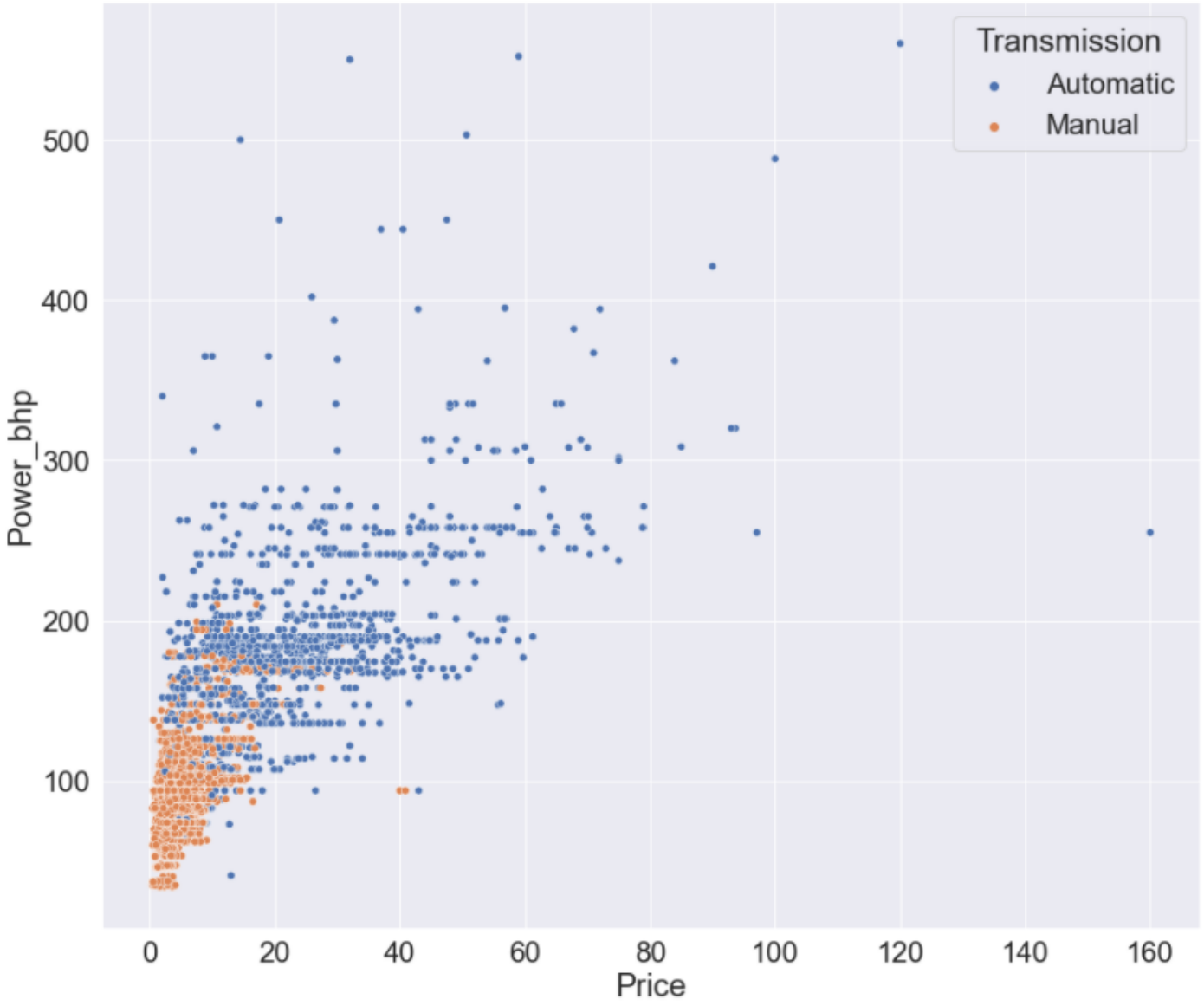
- Positive correlation were found in Variable Price by variables Engine_CC, Power_bhp, and New_price_Lakh meaning that when Price goes up those variables follow the same behavior.

- Negative correlation were found in Variable 'Miles in Kmpl' (Fuel economy) by variables Engine_CC, Power_bhp, New_price and Price, meaning when this variable increase makes other variables decrease.



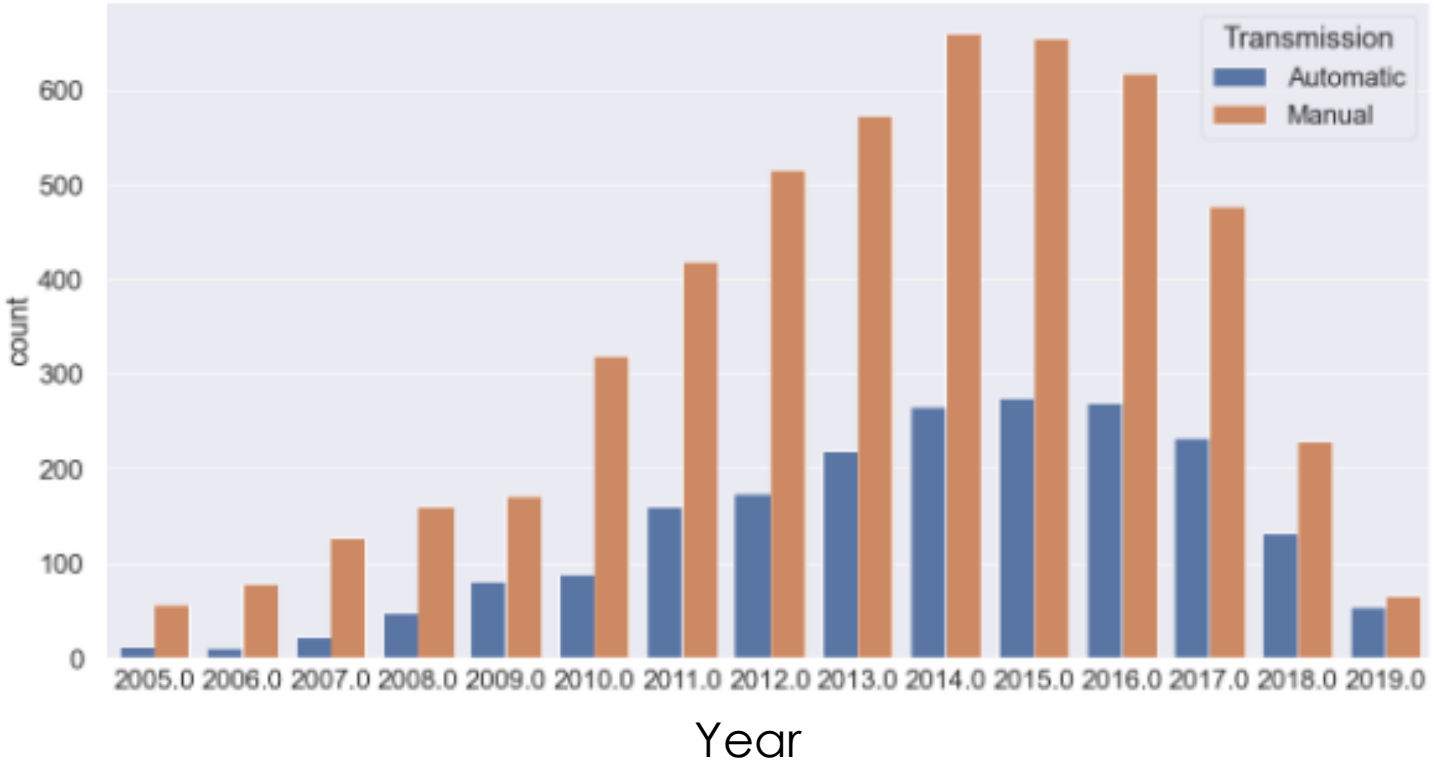
EDA – Multivariate Analysis

- It can be observed that cars that use Diesel and Petrol do have almost same Power up to 400 bhp, after that Petrol becomes the only choice for customers who are looking for cars with 400 bhp or higher.
- Below 80 bhp most expensive cars are electric powered.
- It can be observed that most of Diesel cars are more expensive than other vehicles with different fuel type.



EDA – Multivariate Analysis

- It can be observed that cars with automatic transmission are more expensive than manual ones and they are found in most cases above 80bhp.
- An interesting factor is that cars above 220 bhp are almost not found with manual transmission.
- Lower income customers suitable cars up to 150 bhp.



EDA – Multivariate Analysis

- It can be observed that the gap between cars with manual transmission to automatic one have been reduced from 2017 on which may indicate a trend towards automatic transmission.
- It can also be observed that most of cars in the market are between 2011-2017 (manufacturing year). As of now cars on the market are 4 to 10 years old.

Conclusions

1. 98.9% of the cars from the data have their fuel type as either Diesel or Petrol with diesel having the bigger share 53.1%.
   - For cars with power 400 bhp or above Petrol is the only choice.
   - Diesel cars are more expensive than Petrol ca
2. 82.1% of cars were first owners, 15.9% second owners
3. 95.1% of the cars were 5 or 7 seat cars. With 5 seats representing 84.1%.
4. It can be observed that cars with automatic transmission are more expensive than manual ones but represents only 28.3% of the cars however the gap between manual transmission and automatic one is decreasing which may indicate a trend towards automatic transmissions.
   - Cars above 220 bhp are found only with automatic transmission
5. Positive correlations related to variable Price are: Power, Engine and New Price meaning when those variable increase, Price also increase.
6. Negative correlations related to variable Price are: Km per liters (in US MPG), which is a good insight we it tells when Price increase, cars become less economics.

Recommendations

1. As of now there are a trend for electrification. The percentages of Petrol and Diesel should be followed closely since it can decrease over the years.
2. Diesel cars are mostly more expensive than Petrol, any reverse change on this direction should be flagged.
3. There is a trend towards automatic transmission, cars in this category might be easier to be sold, incentives and discounts for sellers of this category may increase the amount of customer awareness about brand Cars4U.
4. Company should define strategy over market segmentation where for customers with high income, cars with Petrol fuel, Diesel, Electric or above 200 bhp can be more suitable.
5. Customers with lower income, strategy should be more around cars with manual transmission, petrol and below 150 bhp.
6. Some locations are highly sensitive to price changes where other are not, next slide locations are listed.

Prediction Model Information

1. It was used a Linear Regression Model to predict "Prices" of old cars based on their characteristics. The following variables were taken in consideration in order the model predicts with the accuracy informed below:
   - Manufacturing year , Kilometers Driven, Mileage in Kmpl, Engine capacity, Power, Location, Fuel Type, Transmission Type, Type of ownership, Quantity of Seats
2. Regression model was initially used to predict missing values found in variable 'Price' and then a second regression model was created containing the whole data.
   - The model was able to explain 86.46% of the variance in the testing data, which can be considered as a good result to move forward.
   - Model was consistent and considered good when testing data was compared to training data where it was found a difference of 0.00351 or 0.3% between comparison of R^2.
   - Technics were applied to attempt to reduce the absolute error and the Mean Absolute Error found was 1.4108 where 0 represents no errors.
3. Regression coefficients also brought some insights:
   - Variables with positive relationship to Price (meaning Increase in Price cause an Increase in the variable)
     - Year, Engine, Power, Locations: Ahmedabad, Bangalore, Chennai, Coimbatore, Hyderabad and Jaipur and variables Automatic Transmission, Owner Type First / second /third. Car qty seats: 2 / 4.
   - Variables with negative relationship to Price (meaning Increase in Price cause a decrease in the variable)
     - Kilometers Driven, Mileage, Locations: Delhi, Kochi, Kolkata, Mumbai and variables: Fuel type, Transmission Manual and seats 5 to 9.