CARS4U PROJECT

2021

OBJECTIVES

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

BUSINESS PROBLEM OVERVIEW

BACKGROUND

- Cars4U is a budding Tech Start that seeks to gain market share in the industry
- As a senior data scientist at Cars4U, my task is to come up with a pricing model that can effectively predict the price of used cars and can help the business in devising profitable strategies using differential pricing.

SOLUTION APPROACH (MACHINE LEARNING)

- Define the problem and perform an Exploratory Data Analysis
- Illustrate the insights based on EDA
- Data pre-processing
- Model building - Linear Regression
- Test assumptions of linear regression model
- Model performance evaluation
- Actionable Insights & Recommendations

FINANCIAL IMPLICATIONS

- Increase Sales and Profit Margin
- Sets Price to outwit Competition
- Increase Revenue
- Effect Allocation and redistribution of resources especially for Advert campaigns
- Targeted Marketing and Media Campaigns
- Curbing costs

DATA MANIPULATION

- Identification of Missing Values
- Fixing Columns (New_Price and S. No) were dropped. New_Price because of too many missing data
- Conversion of Data types
- Treating Missing Values and Missingness
- Prior to modelling, Year was dropped in place of a new variable, Age created for a cleaner prediction

DATA INFORMATION

Variable	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 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	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)

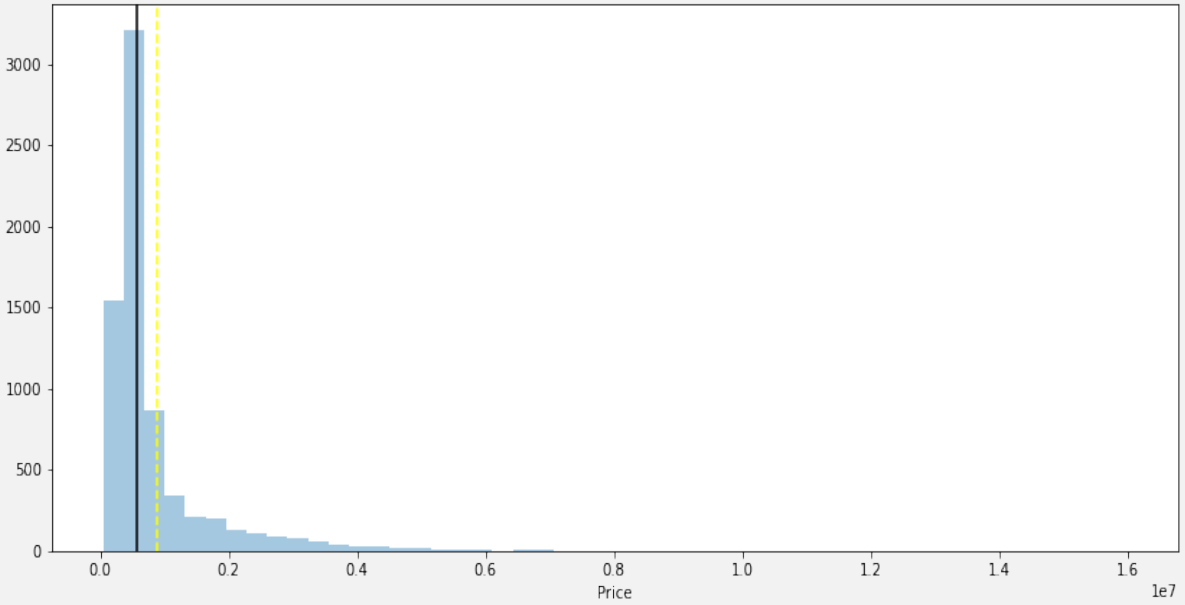
Observations	Variables
7253	14

Numerical	Object
5 S.No Seats Price Year Kilometers_ Driven	9 Name Location Fuel_Type Transmission Owner_Type Mileage Engine Power New_Price

Missing Values in Data
New_Price 6247
Price 1234
Seats 53
Power 46
Engine 46

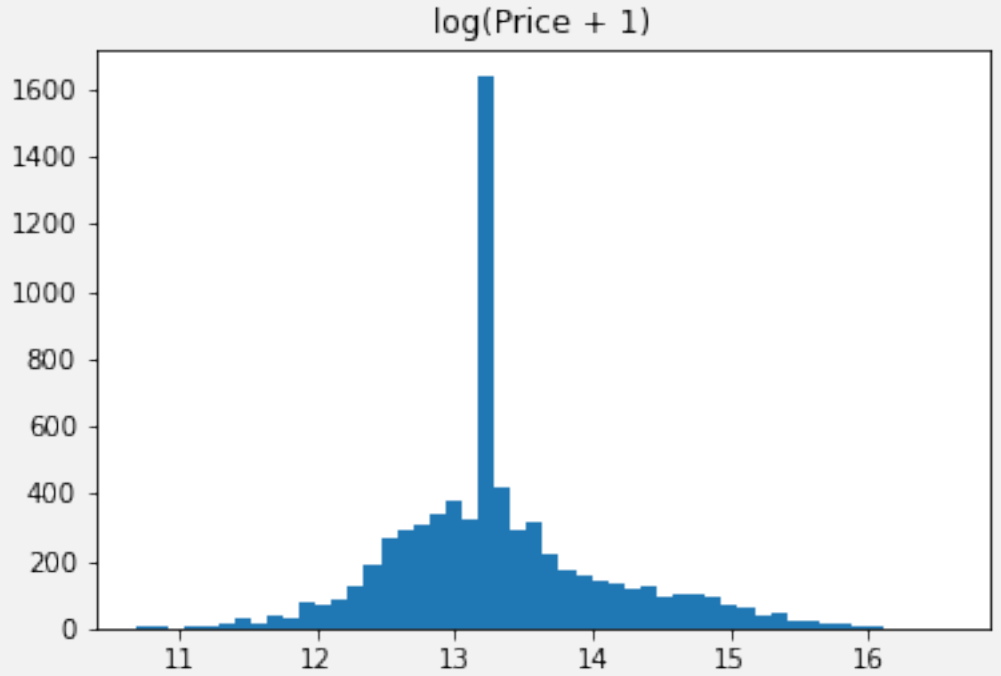
EXPLORATORY DATA ANALYSIS

PRICE



- There are massive outliers. This may be accounted for by erroneous data collection or variability in methods
- Price distribution appear normal but heavily rightly skewed
- 50% of potential sales topped 564,000 and below while 75% of potential revenue was below 850,000
- We will further treat the skewness using Log Transformation

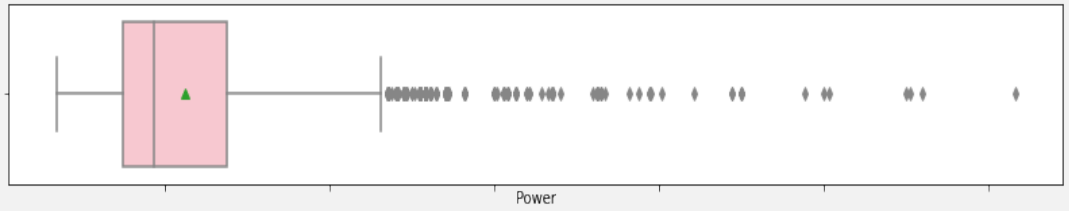
Log Transformation-Price

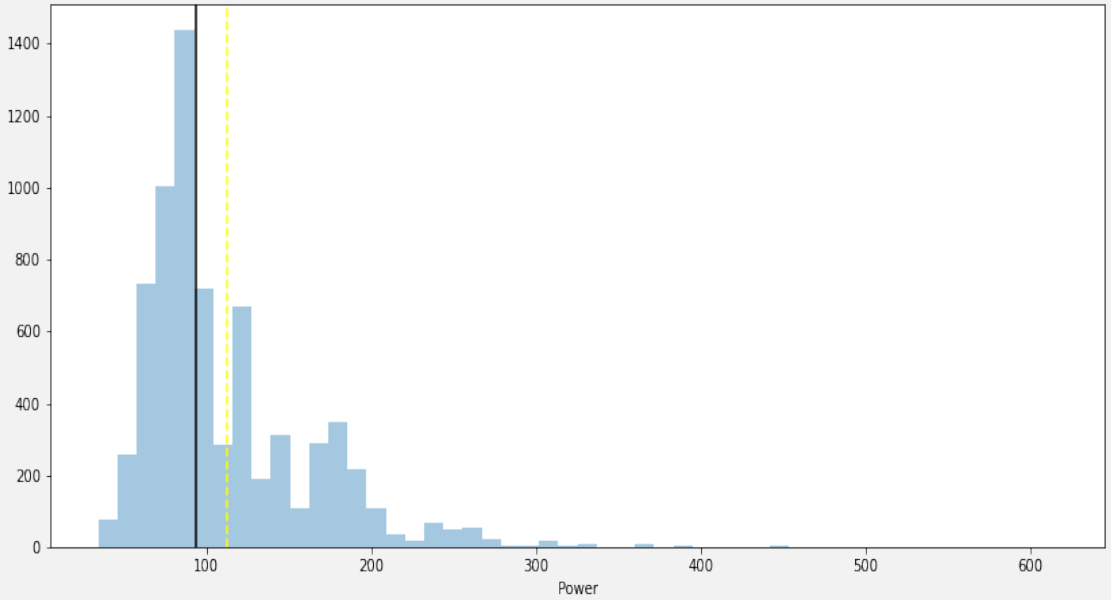


- Note that the skewness has greatly been eliminated to affirm a Normal Distribution
- Price definitely behaves better on a Log scale

EXPLORATORY DATA ANALYSIS

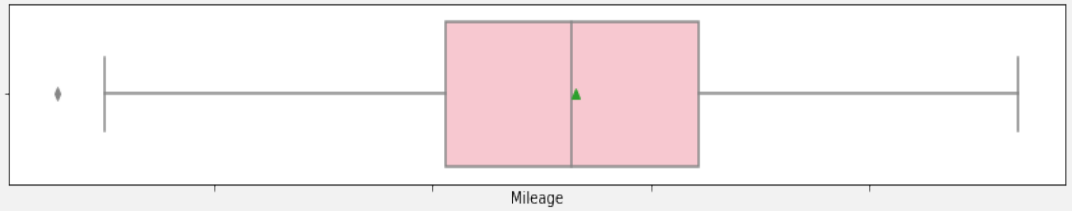
POWER

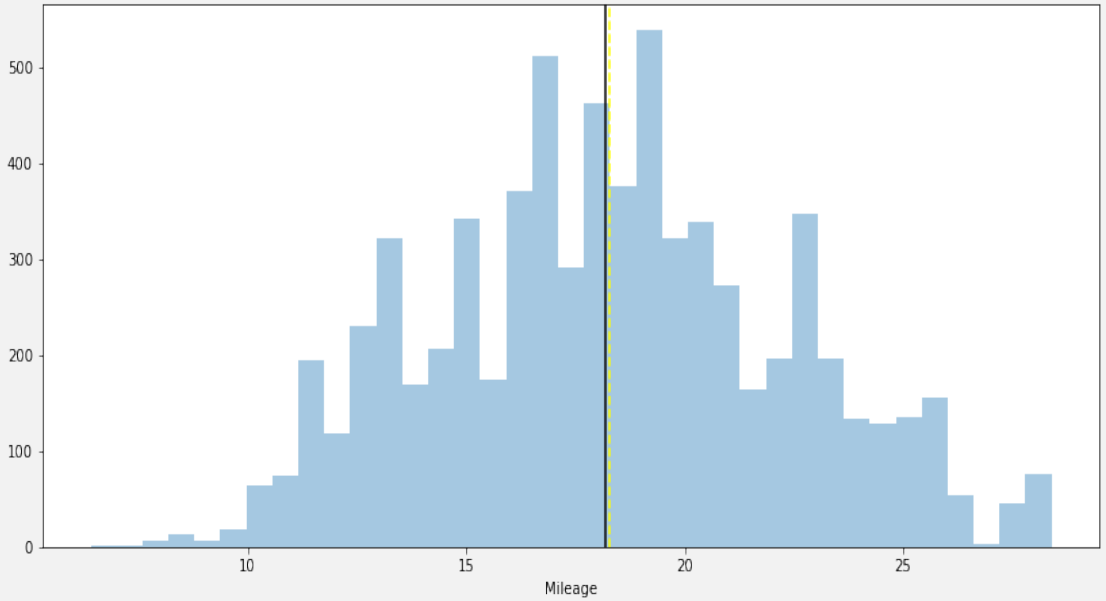




- There are significant outliers hence Right skewness.
- Power is a normal distribution. Mean (112.56) is approx. equal to Median (93.7)
- There is a presence of outliers which may indicate variability in data collection
- Outliers will be treated for a perfect modeling

MILEAGE

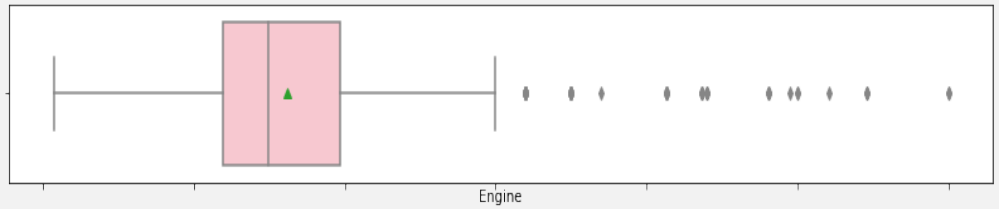


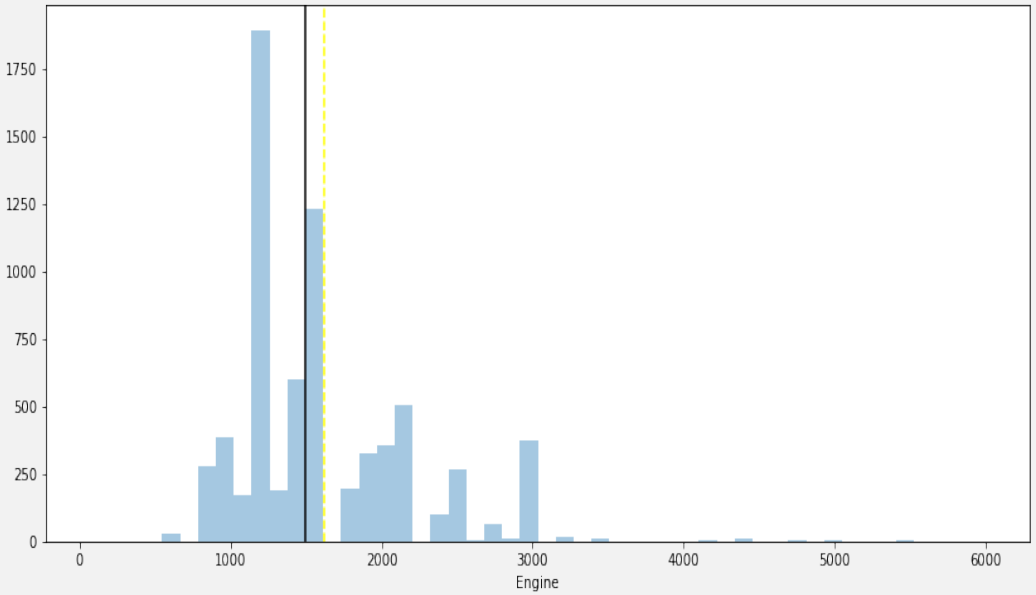


- The Mileage distribution is approximately near perfectly normal
- The Mean(18.214) > Median (18.19) by a little margin.
- The presence of a few outliers but with little or no impact on skewness

EXPLORATORY DATA ANALYSIS

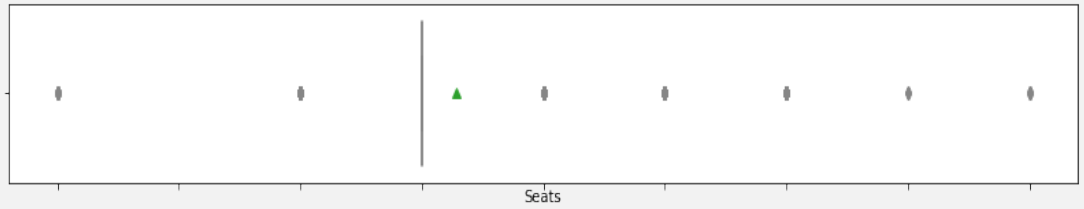
ENGINE

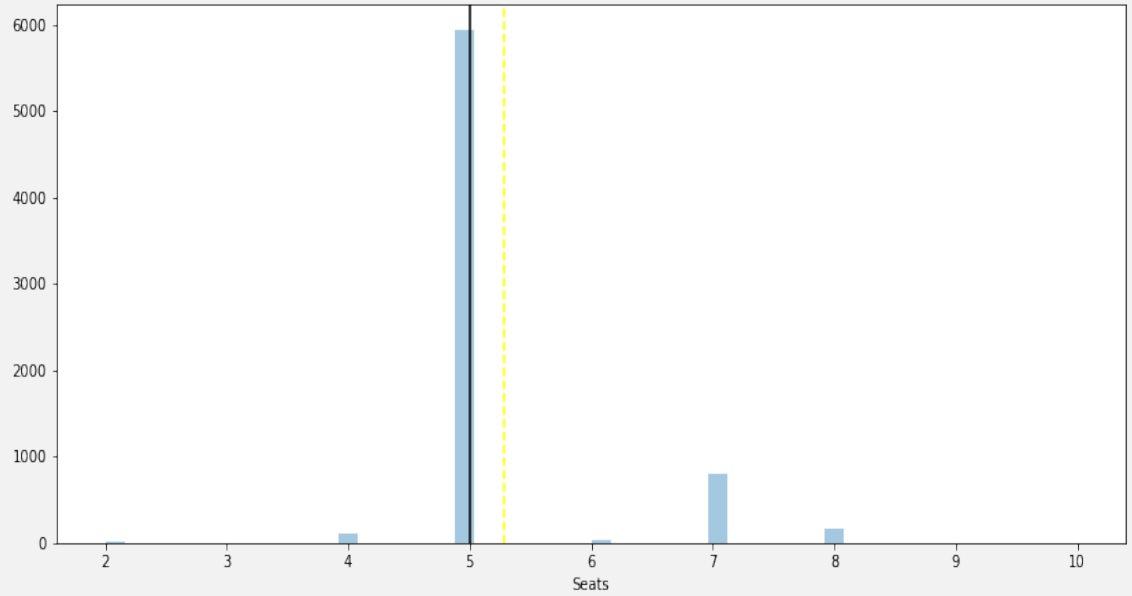




- The mean > median, hence the skewness to the right
- Presence of a lot of outliers
- Below 75% of the used cars on sale had a cumulative 1998 bhp and the upper quartile aggregated 5998 bhp
- There is an obvious possibility of error in the data gathering owing to the amount of outliers present

SEATS

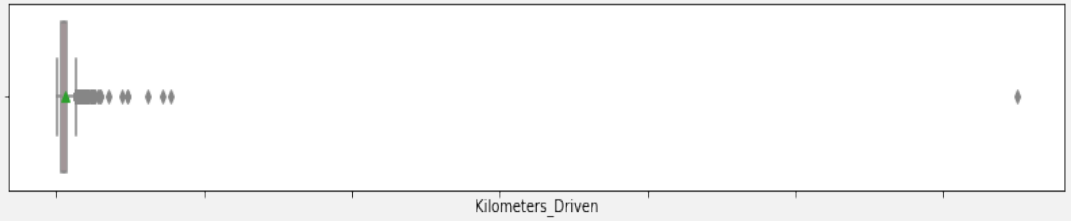


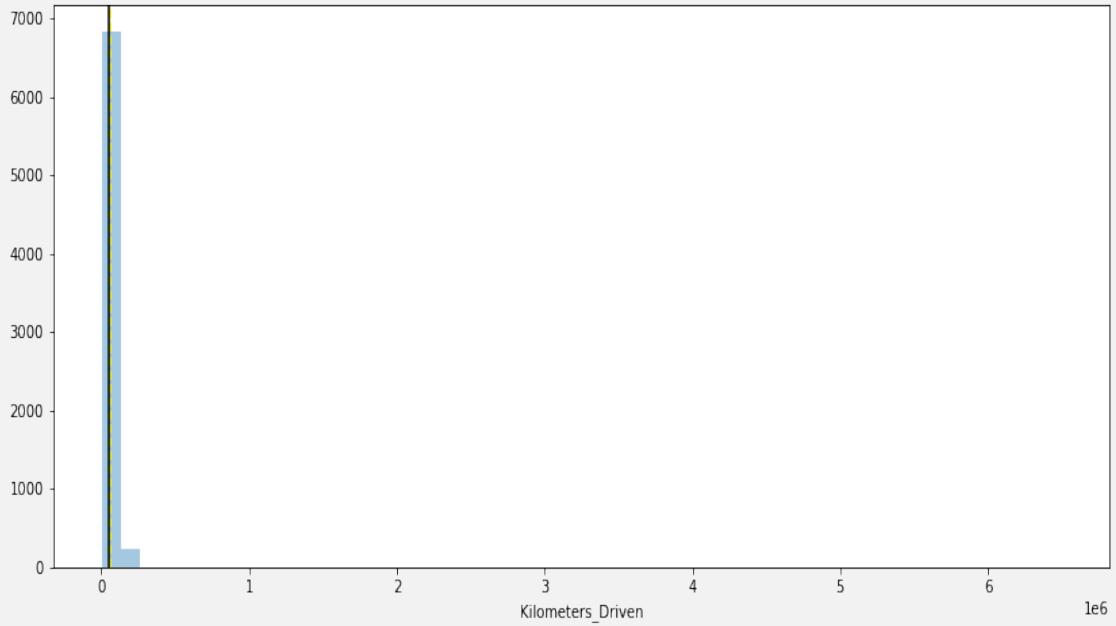


- Seats is an obvious uniform distribution as seen from the Histogram plot

EXPLORATORY DATA ANALYSIS

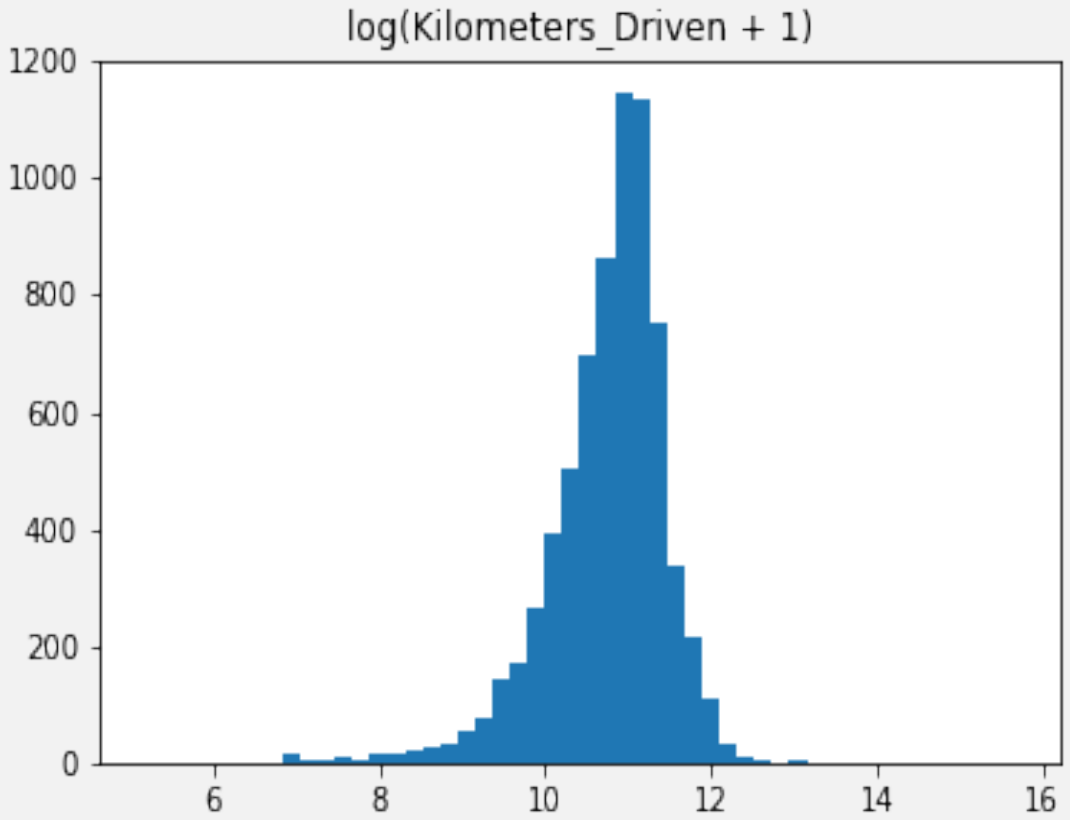
KILOMETERS DRIVEN





- There are massive outliers. Kilometer_Driven Distribution suggests a normal one but will be investigated with a LogTransformation
- This is heavily rightly skewed as seen from the range(6499829) and standard deviation(85231)

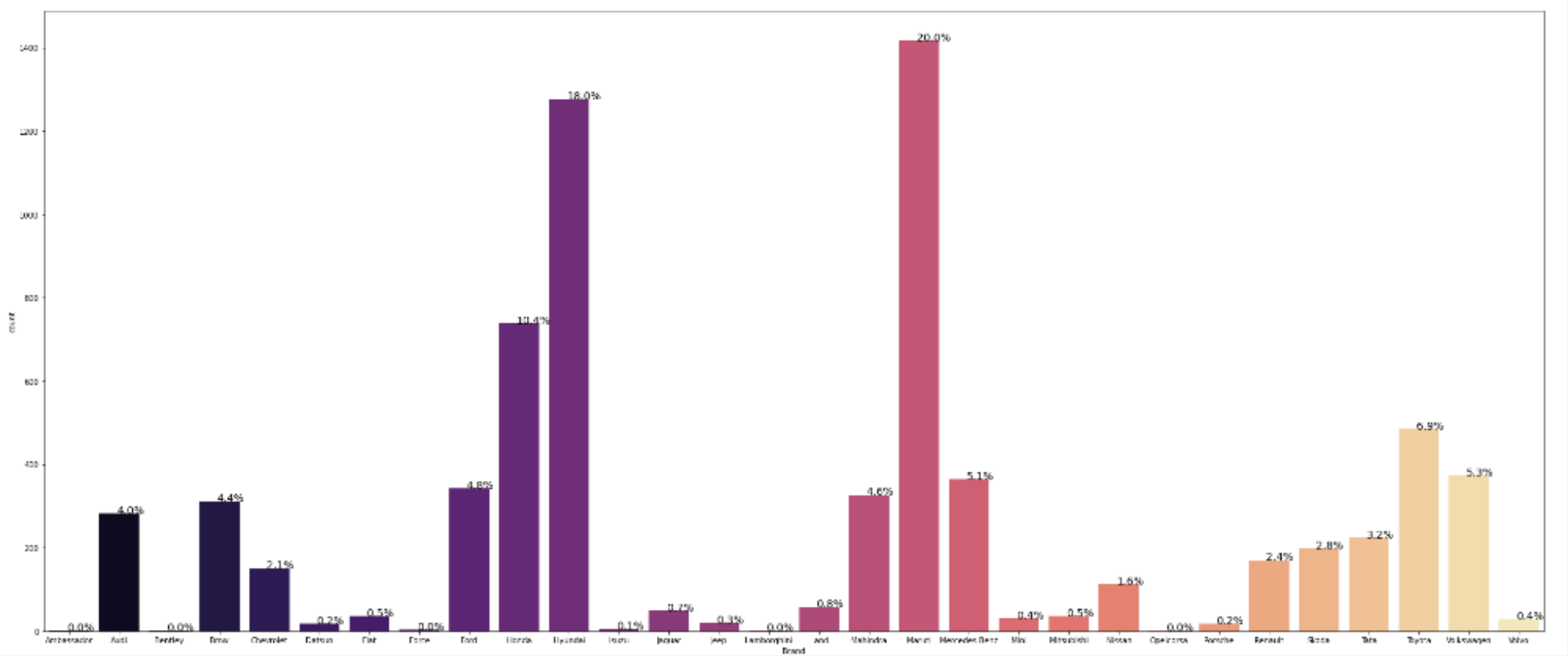
LOG TRANSFORMATION



- The Power of Log Transformation and its effect on kurtosis. We can see an apparent Normal distribution

EXPLORATORY DATA ANALYSIS

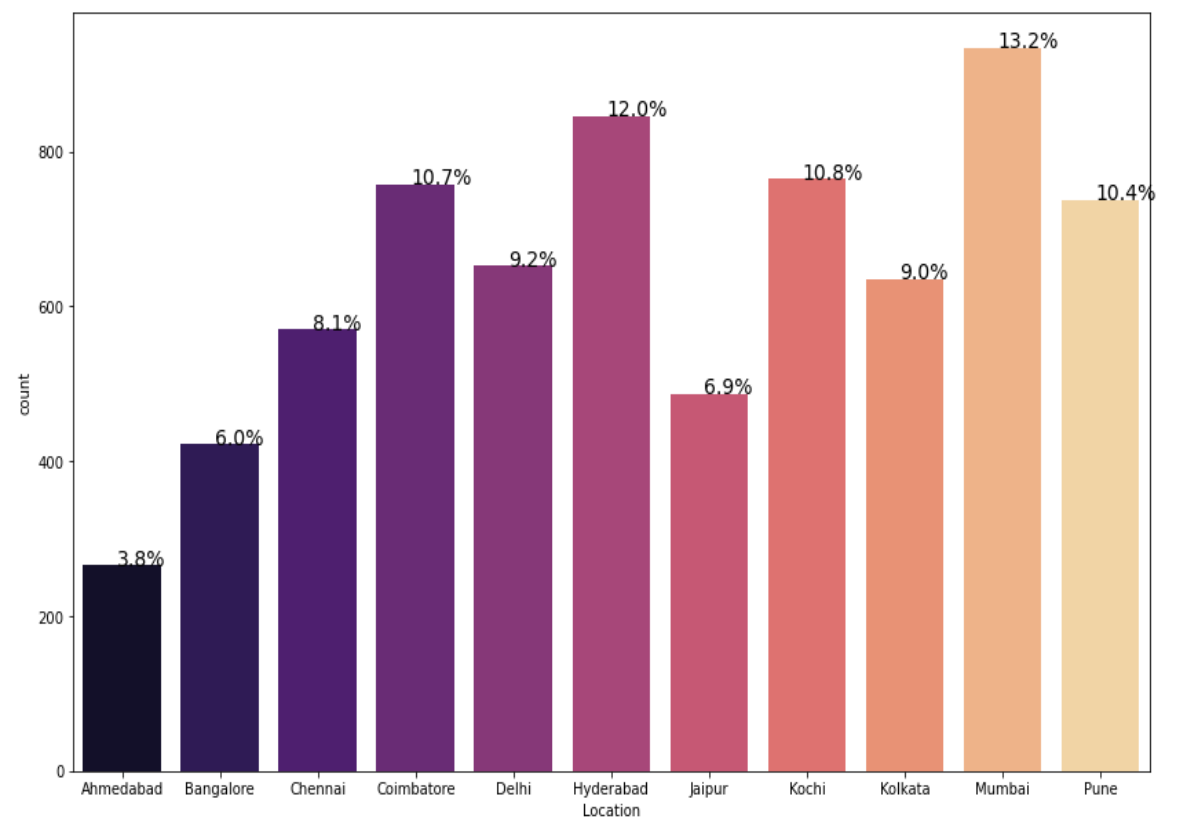
BRAND



- Maruti (20%), Hyundai(18%) and Honda (10.4%) respectively account for availability compared to other brands in the distribution
- Ford Brand came a distant 4th for used cars on sale at 4.8%

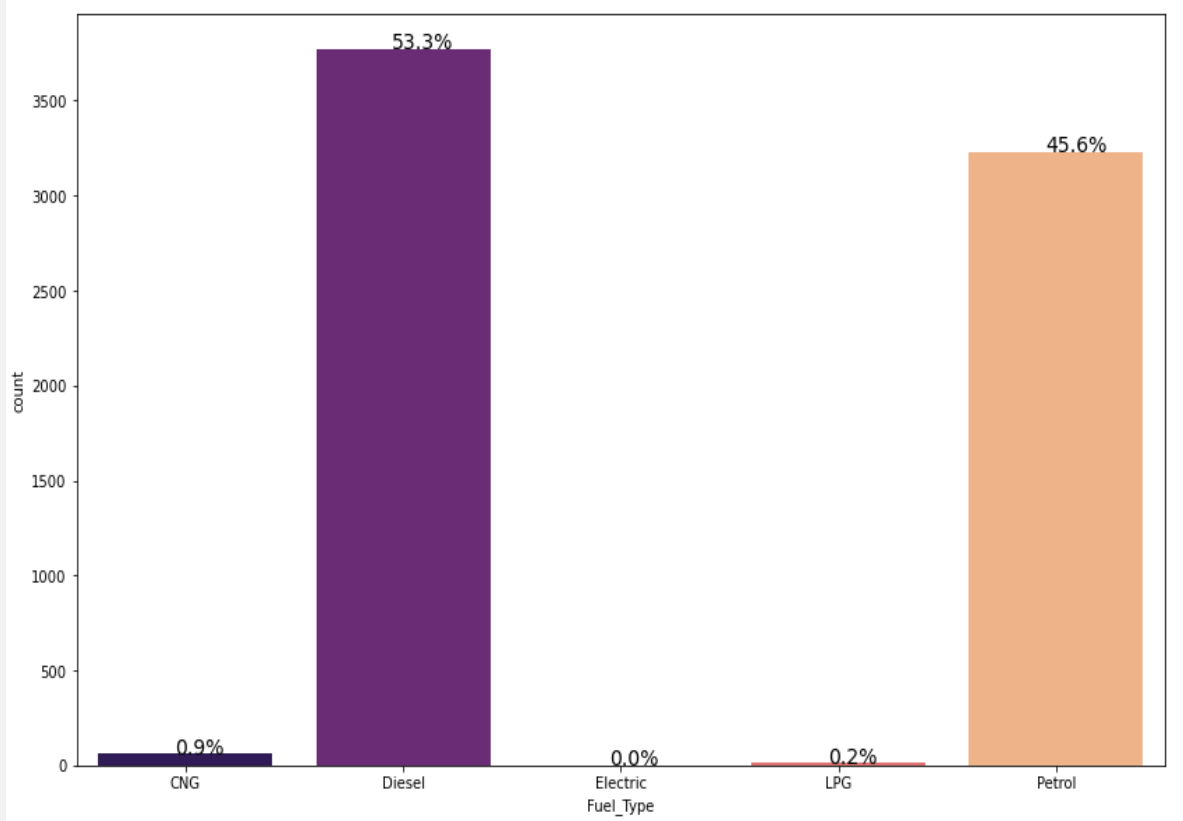
EXPLORATORY DATA ANALYSIS

LOCATION



- Mumbai (13.2%), Hyderabad(12,0%) and Kochi(10.8%) respectively account for over one-third of units on sale across the 11 different locations
- The least in terms of units on sale is Ahmedabad

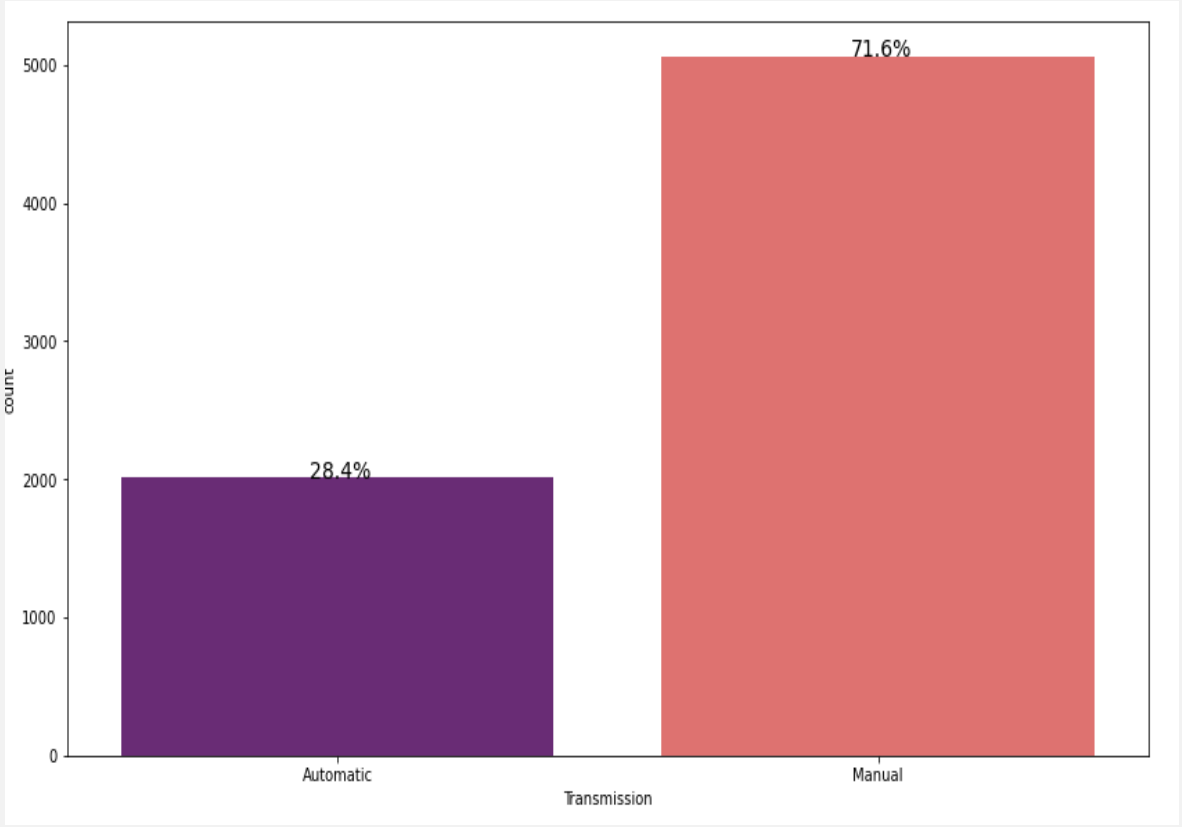
FUEL_TYPE



- Diesel used cars topped potential sales at 53.3% followed by Petrol
- Clearly, stock for used cars of CNG, LPG and Electric fuel types are at a distant low

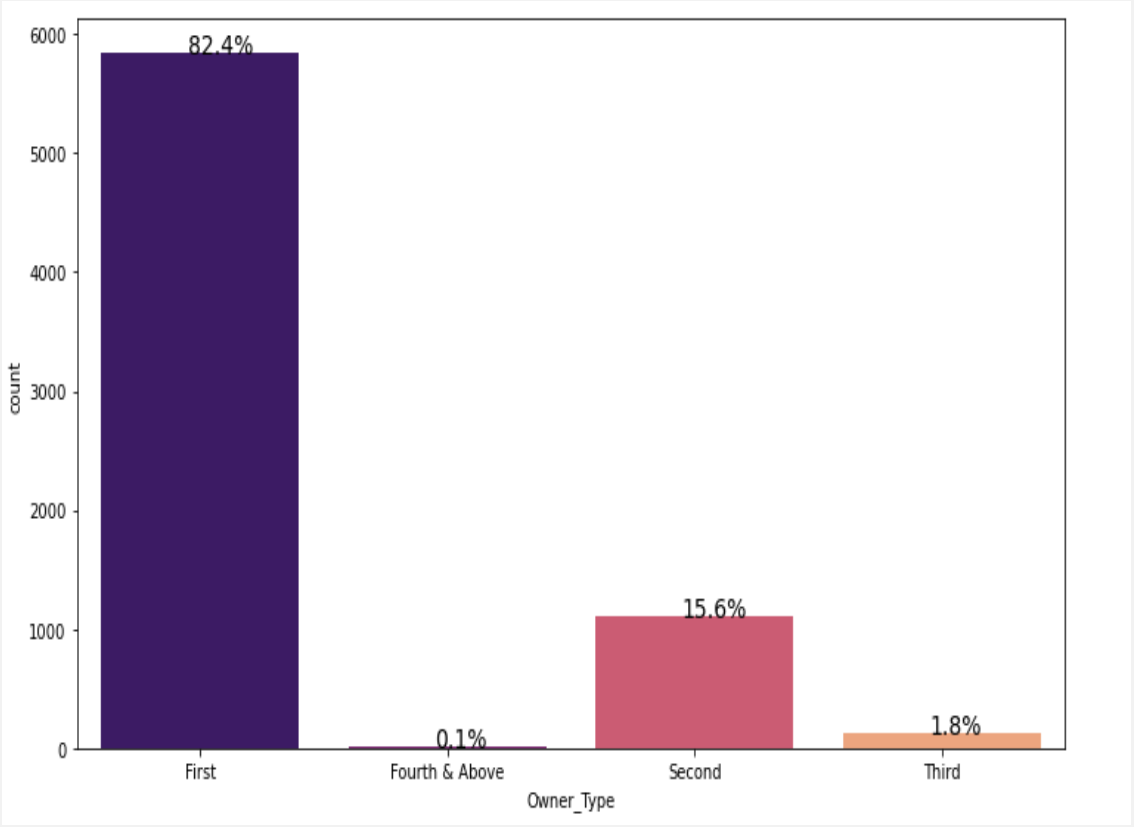
EXPLORATORY DATA ANALYSIS

TRANSMISSION



- Used cars with Manual transmission is the most predominant amongst car dealers with a 71.6% inventory position
- Automatic transmission has less attraction in the used car market with a 28.4% share

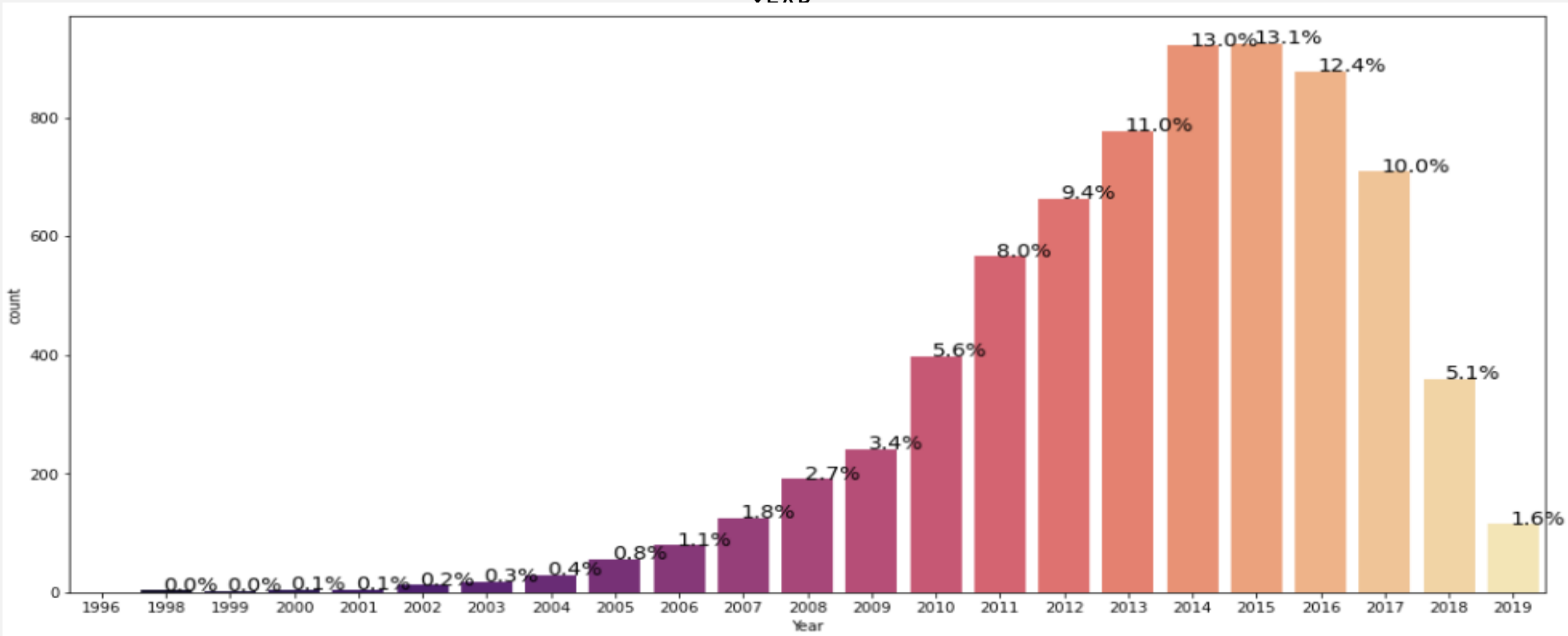
OWNER TYPE



- Dealers preference for used cars by First owners trumped the rest by 82.4% followed by a distant 15.6% of Second Owner types.
- There is a very low affinity for used cars by Third and fourth owners.

EXPLORATORY DATA ANALYSIS

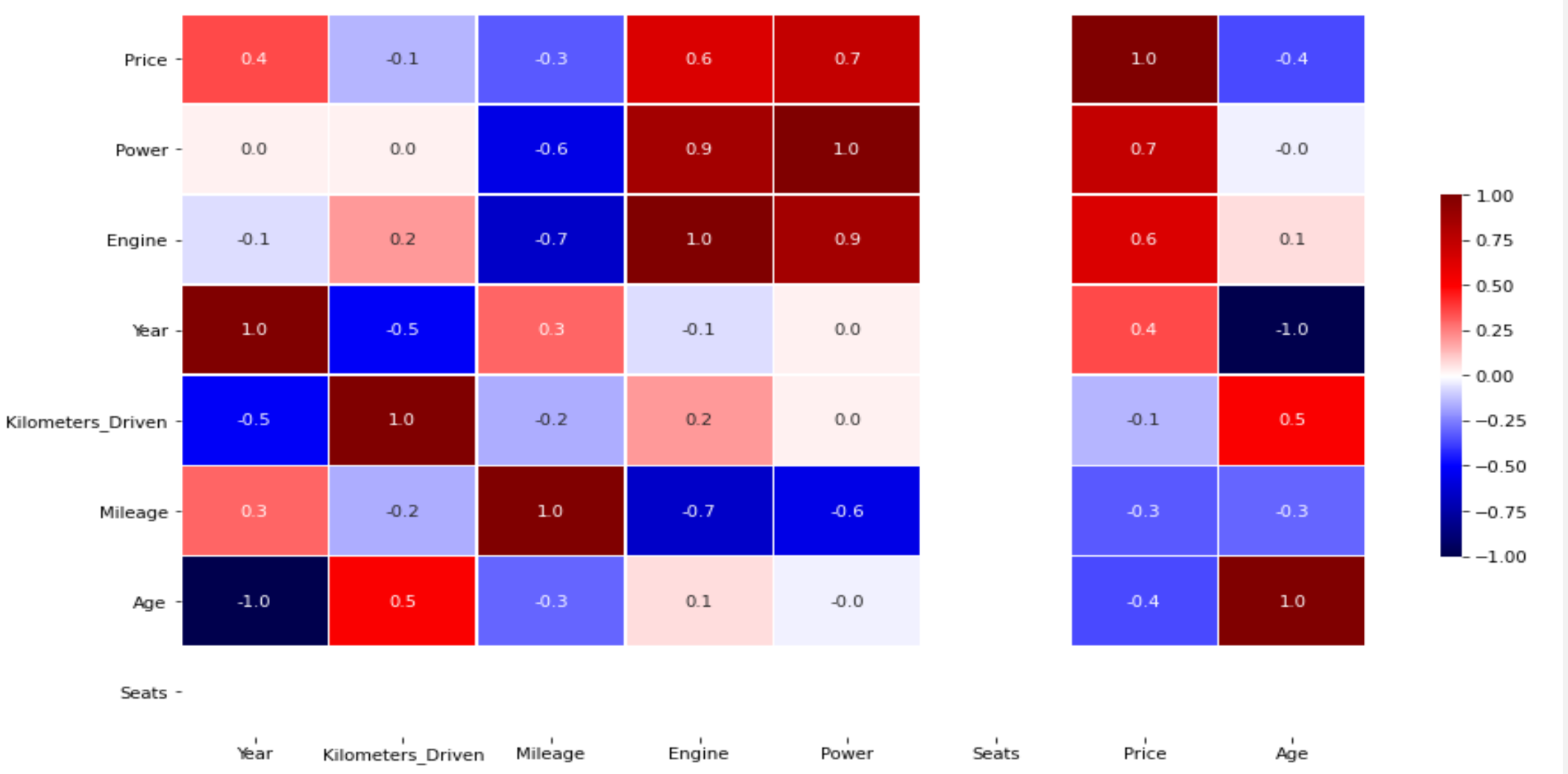
YEAR



- 2015 (13.1%), 2014(13%), 2016(12.4%) and 2013(11.0%) models respectly account for close to 50% of potential sales indicating indicating preference for fairly used cars
- Older models between 1998 to 2007 are less attractions

EXPLORATORY DATA ANALYSIS

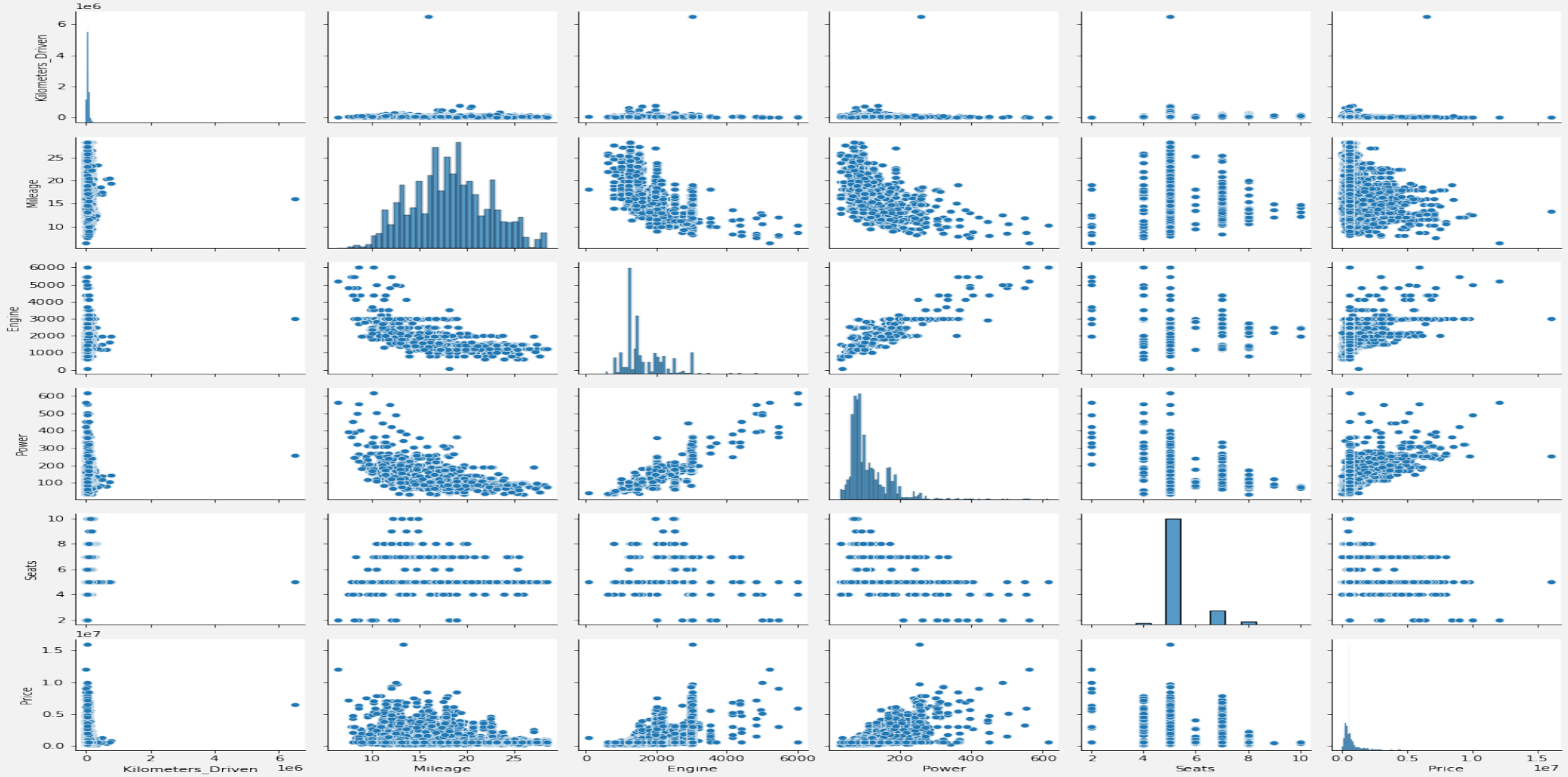
Heat Map



- As earlier observed , Price is highly correlated with Engine and Power
- Engine is highly correlated with Power as well but inversely proportional to Mileage
- There is very little or no correlation between Price the rest.

EXPLORATORY DATA ANALYSIS

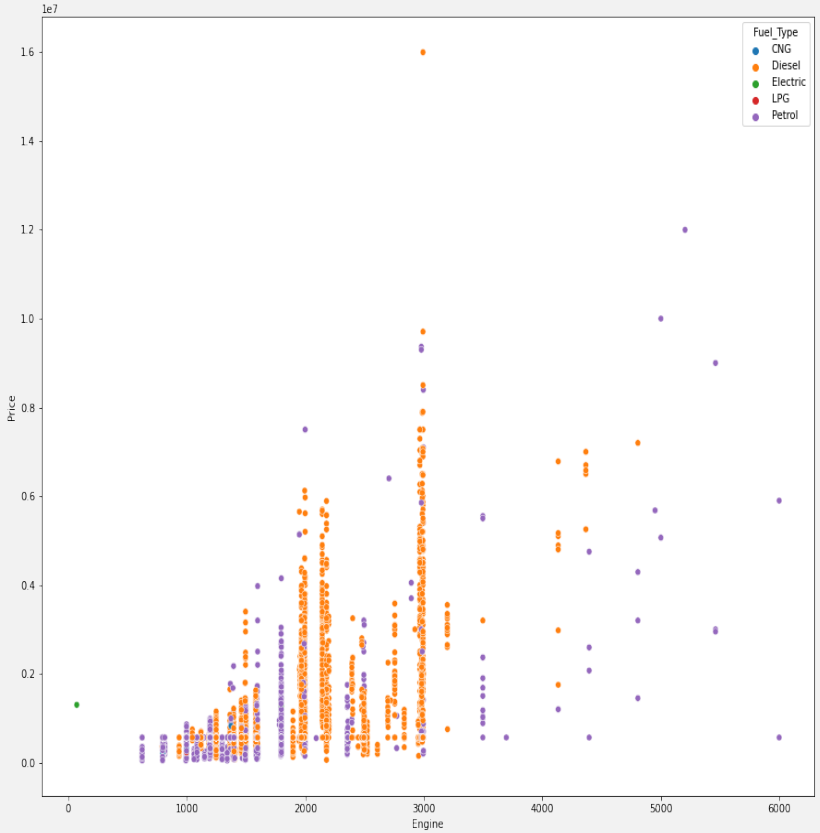
PAIR PLOT



There is quite a significant amount of correlation between Price and other variables (Engine, Power, Mileage, Kilometer Driven)

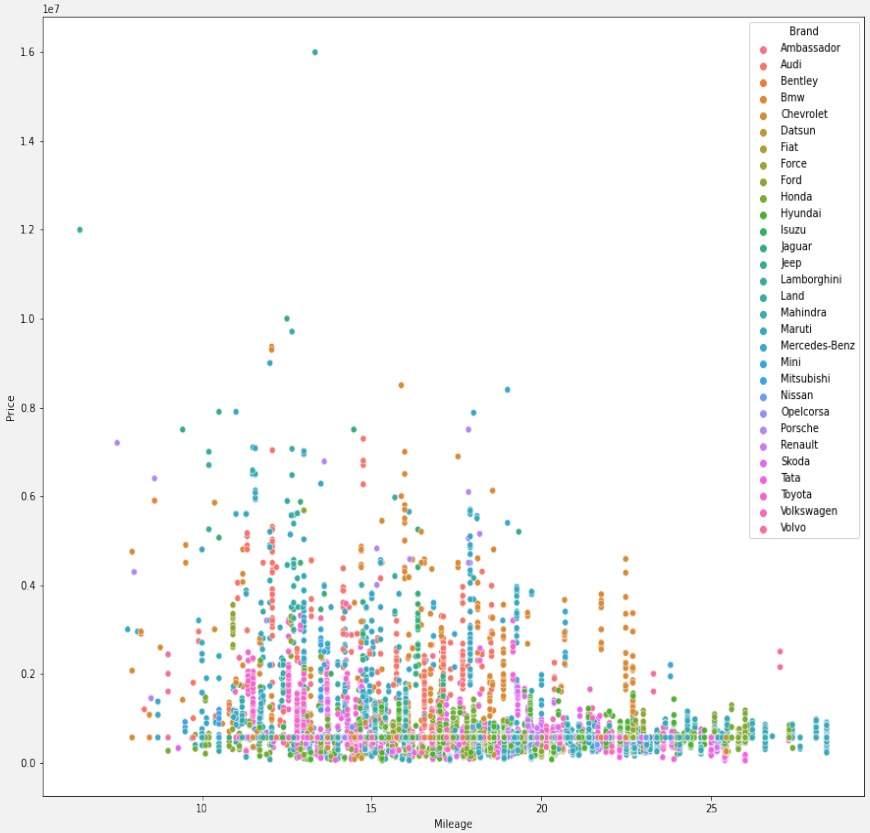
EXPLORATORY DATA ANALYSIS

PRICE-ENGINE-FUEL TYPE



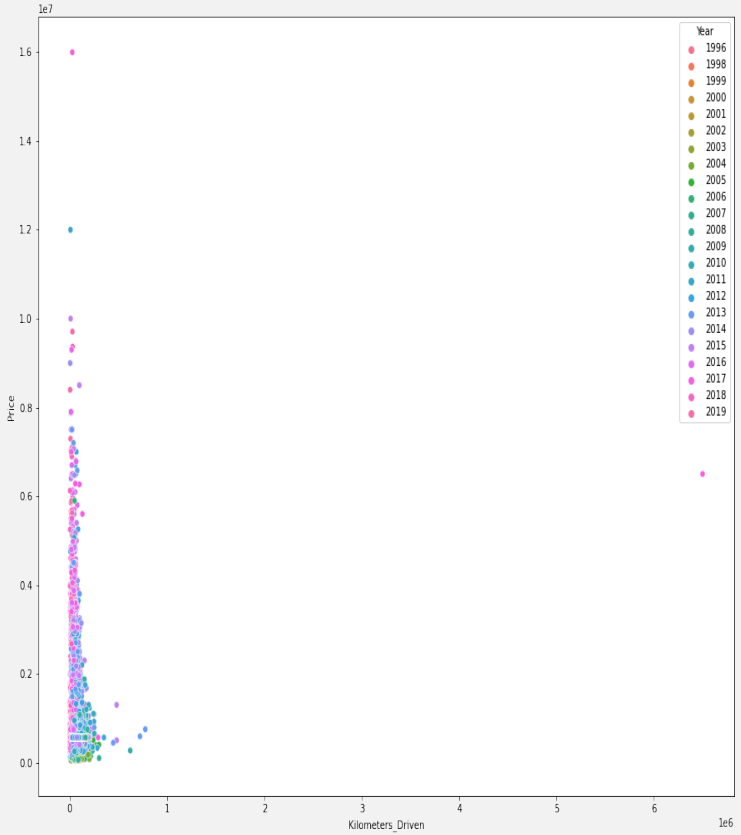
- The higher the Engine capacity,the more robust the fuel type (Diesel) and thus the higher the Price
- The point dispersion especially for the Petrol accounts for a chunk of the outliers seen

PRICE-MILEAGE-BRAND



- Maruti tops the data with lower mileage and equally priced higher followed by Hyundai.
- We can see that cars with higher mileages were priced quite low

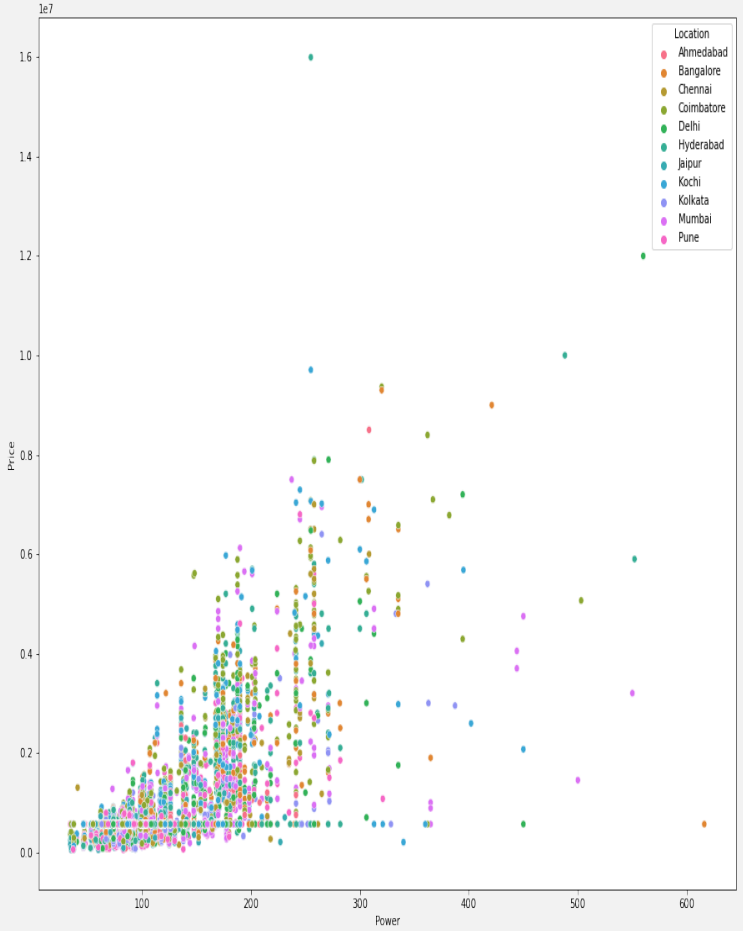
PRICE-KILOMETER DRIVEN-YEAR



- Kilometers_Driven has an adverse impact on Price.
- Used car models of year 2015 with less Kilometers covered prior to sale had the most preference

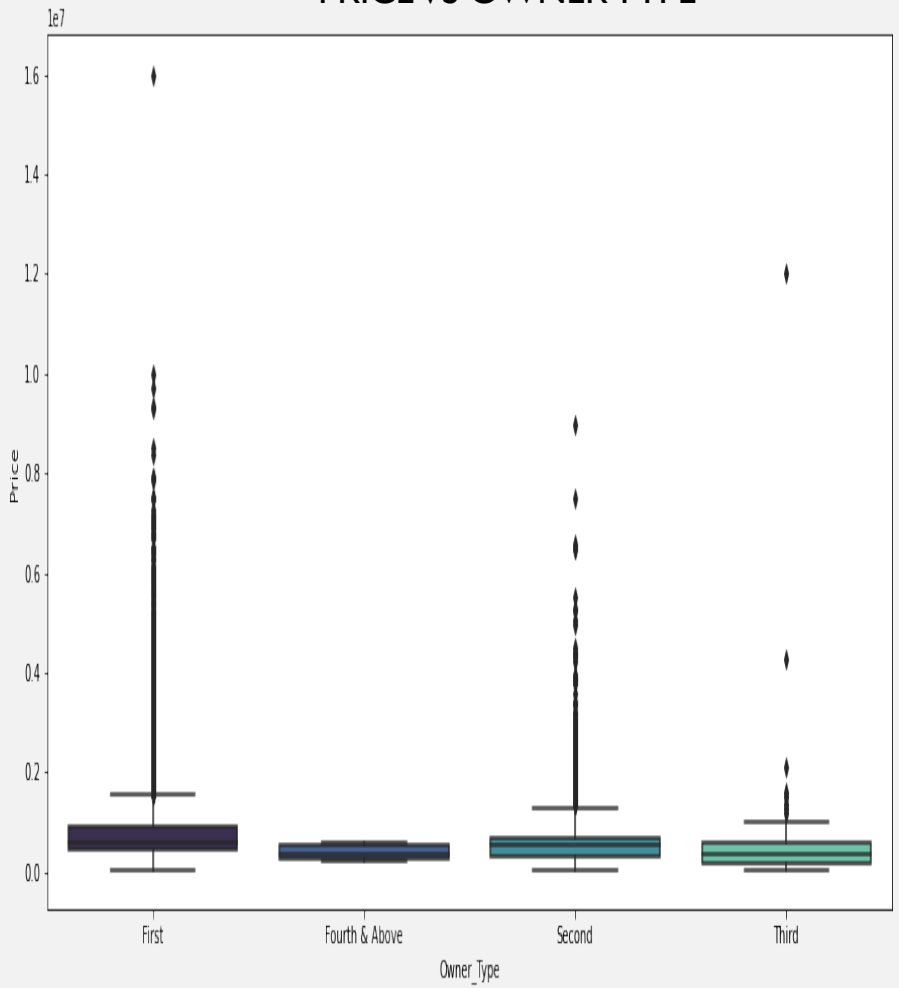
EXPLORATORY DATA ANALYSIS

PRICE-POWER-LOCATION



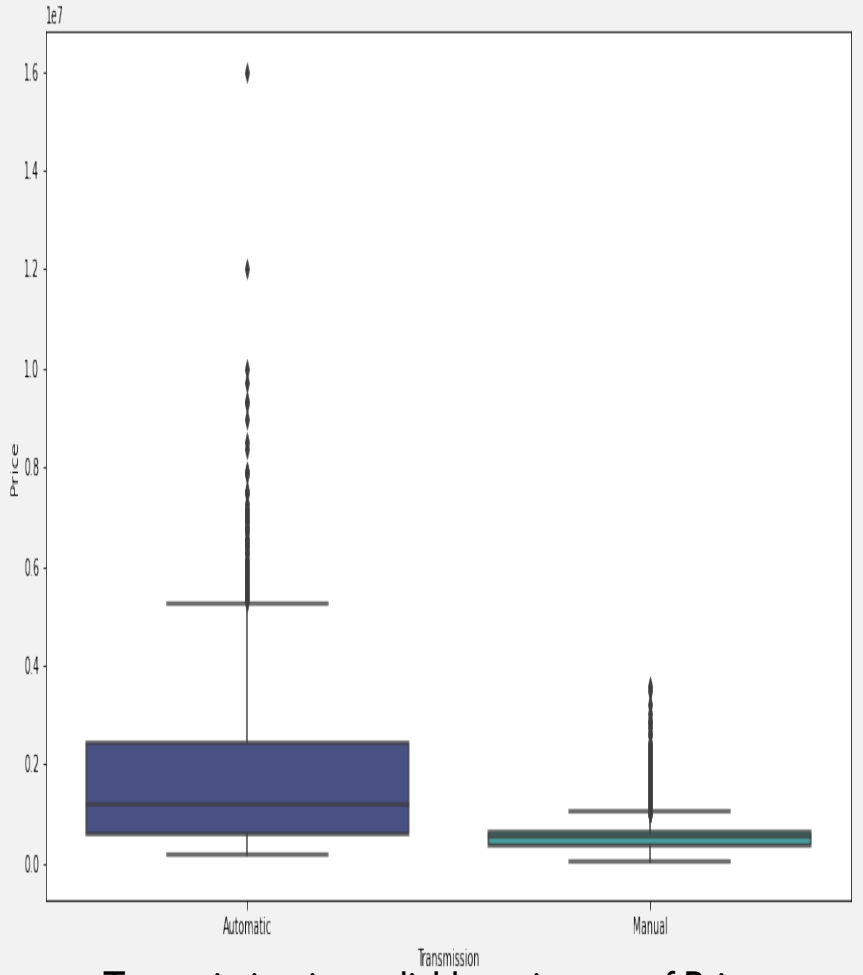
- The obvious linearity between Price and Power is unquestionable
- The frequency of data points representing Mumbai reflects an unmatched inventory of used car units on sale

PRICE VS OWNER TYPE



- We can see the Owner_Type determines the Pricing as the First Owner_Types are most sought after followed by the second
- Pressure from a lot of Outliers

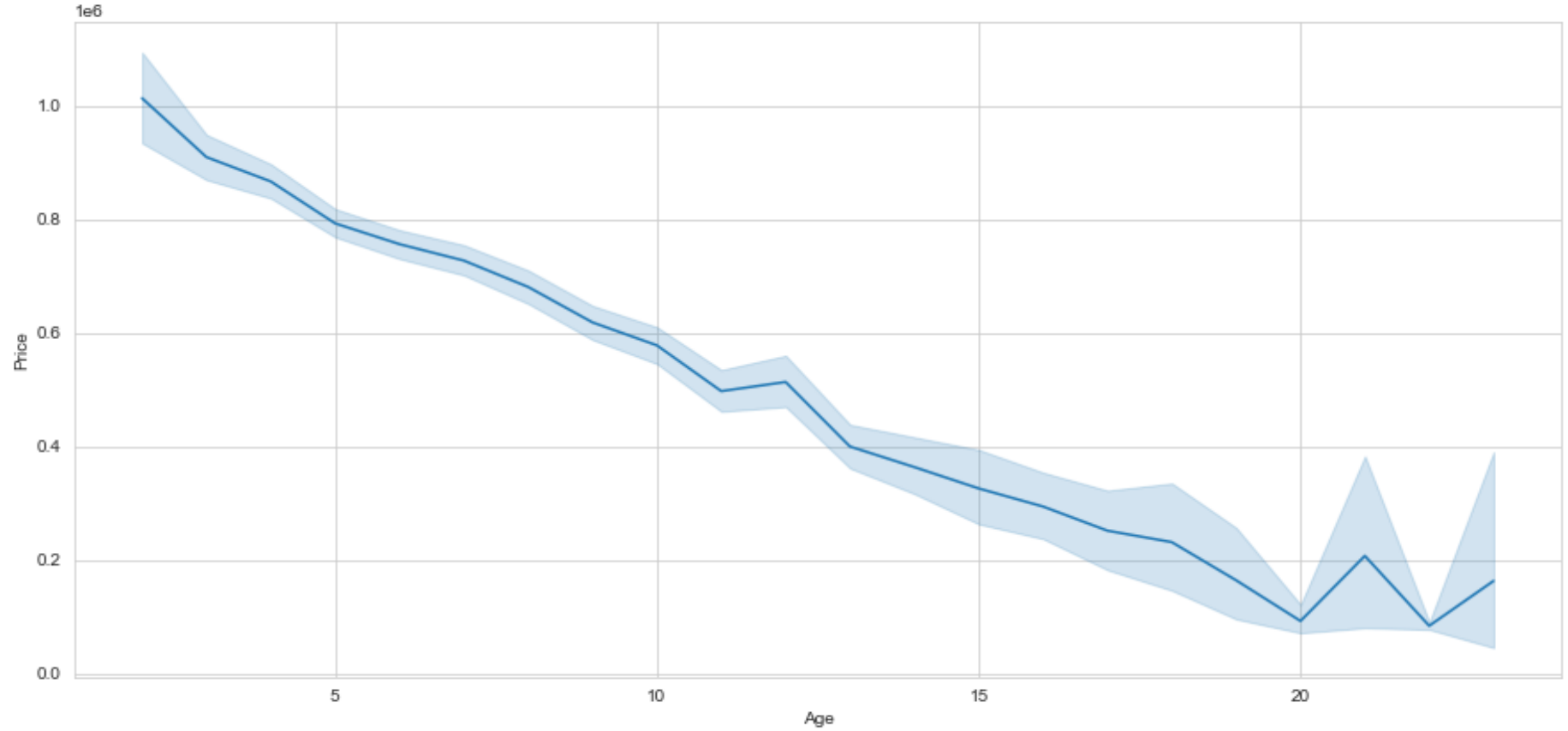
PRICE VS TRANSMISSION



- Transmission is a reliable estimator of Price as can be seen from the plot
- Automatic Transmission is higher priced than the Manual

EXPLORATORY DATA ANALYSIS

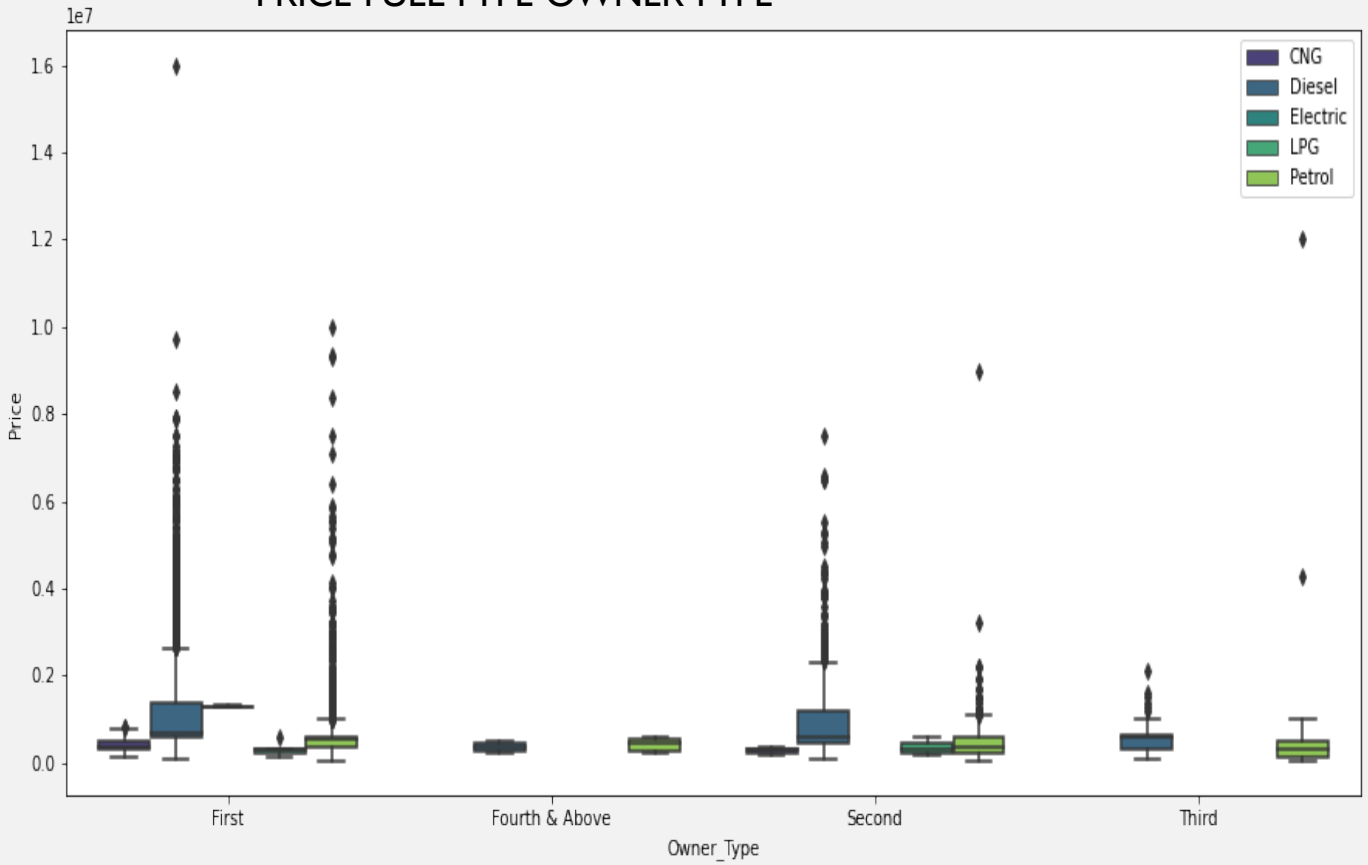
PRICE-AGE



- We can clearly infer from the plot that there is an inverse relation between Price and Age
- This implies that a unit increase in the Age of a used vehicle leads to a decrease in the Price of that vehicle.

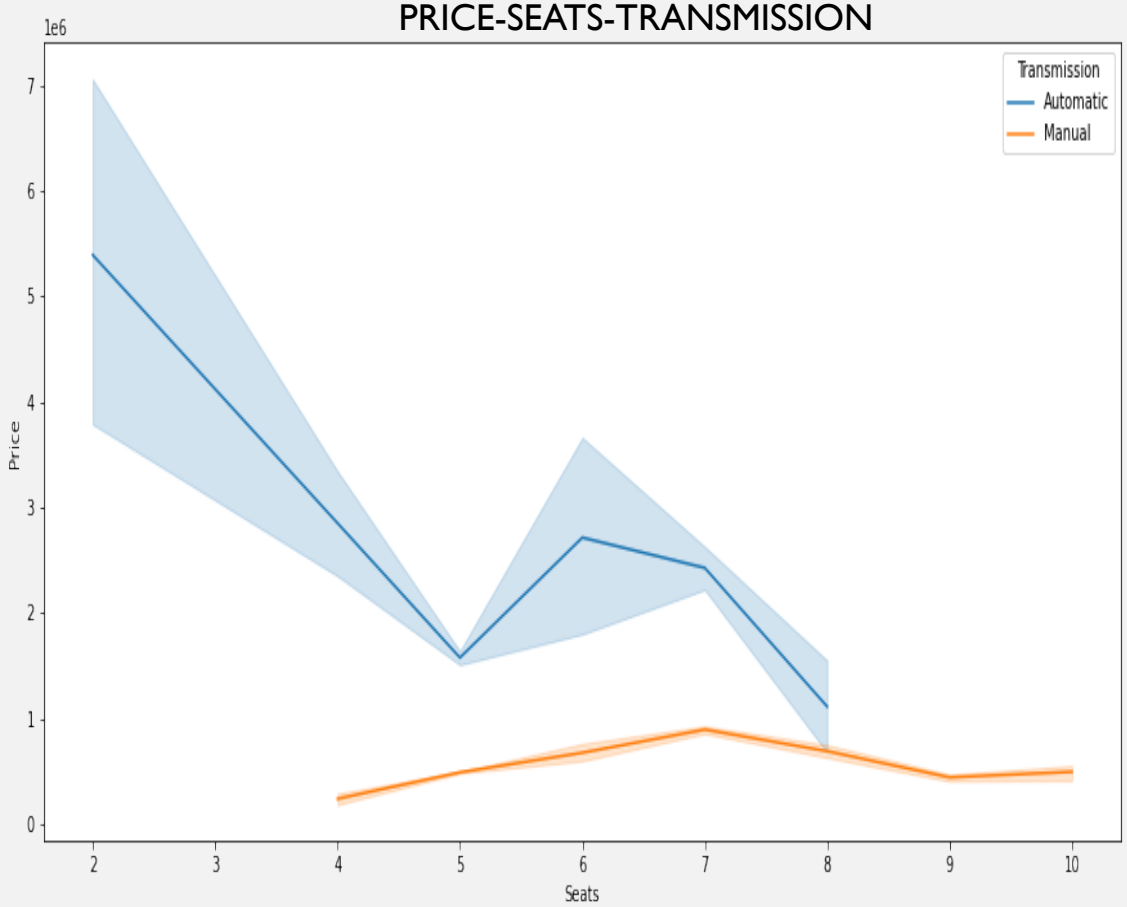
EXPLORATORY DATA ANALYSIS

PRICE-FUEL TYPE-OWNER TYPE



- Right away, it is apparent used cars driven by first owners with Diesel fuel types are higher priced followed by second owner types
- Also amongst first owner types, diesel fuel types are preferred compared to Petrol hence are priced higher
- Interestingly, petrol is prefered by buyers of fourth owner types and are priced almost at par with Petrol of second owner types

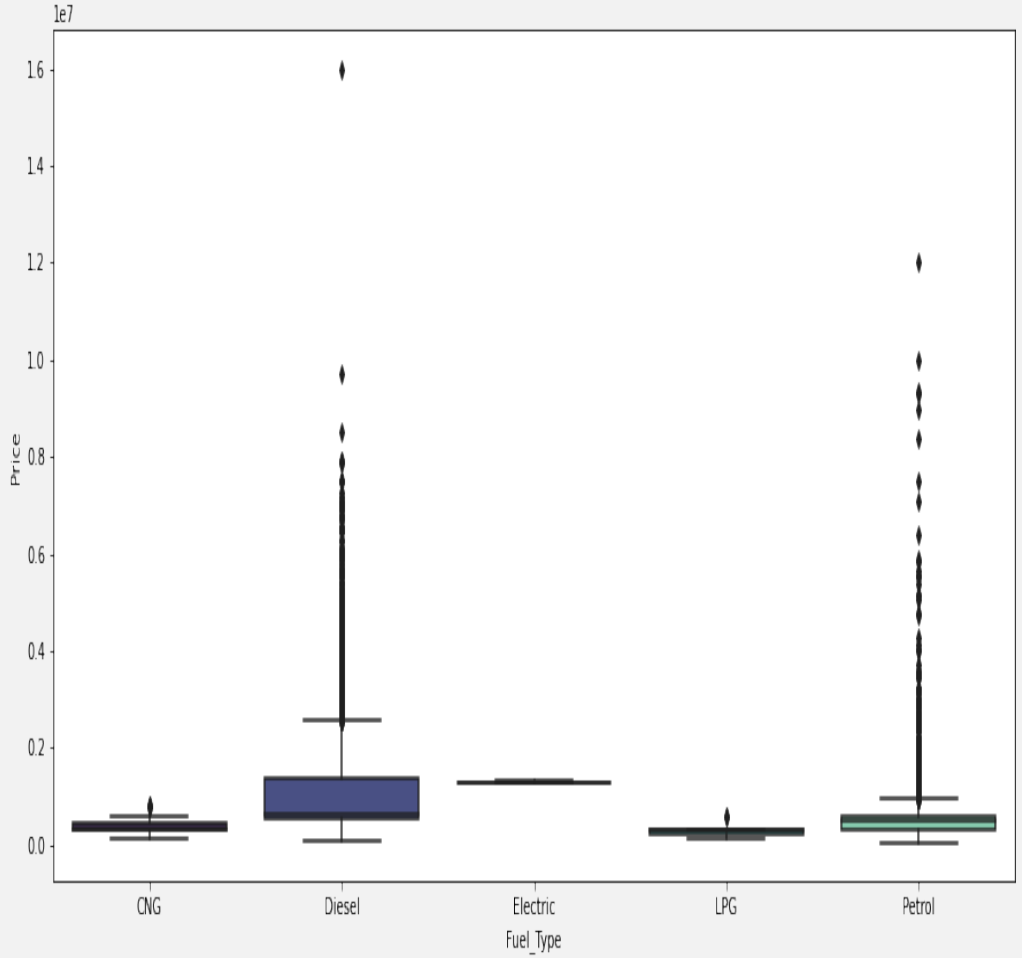
PRICE-SEATS-TRANSMISSION



- Used cars with Automatics transmission decreased in price with an increase in the number of seats
- This pattern clearly speaks to luxury models with automatic transmission like Coupes and SuVs
- Manual used cars are priced quite low as the number of seats increase

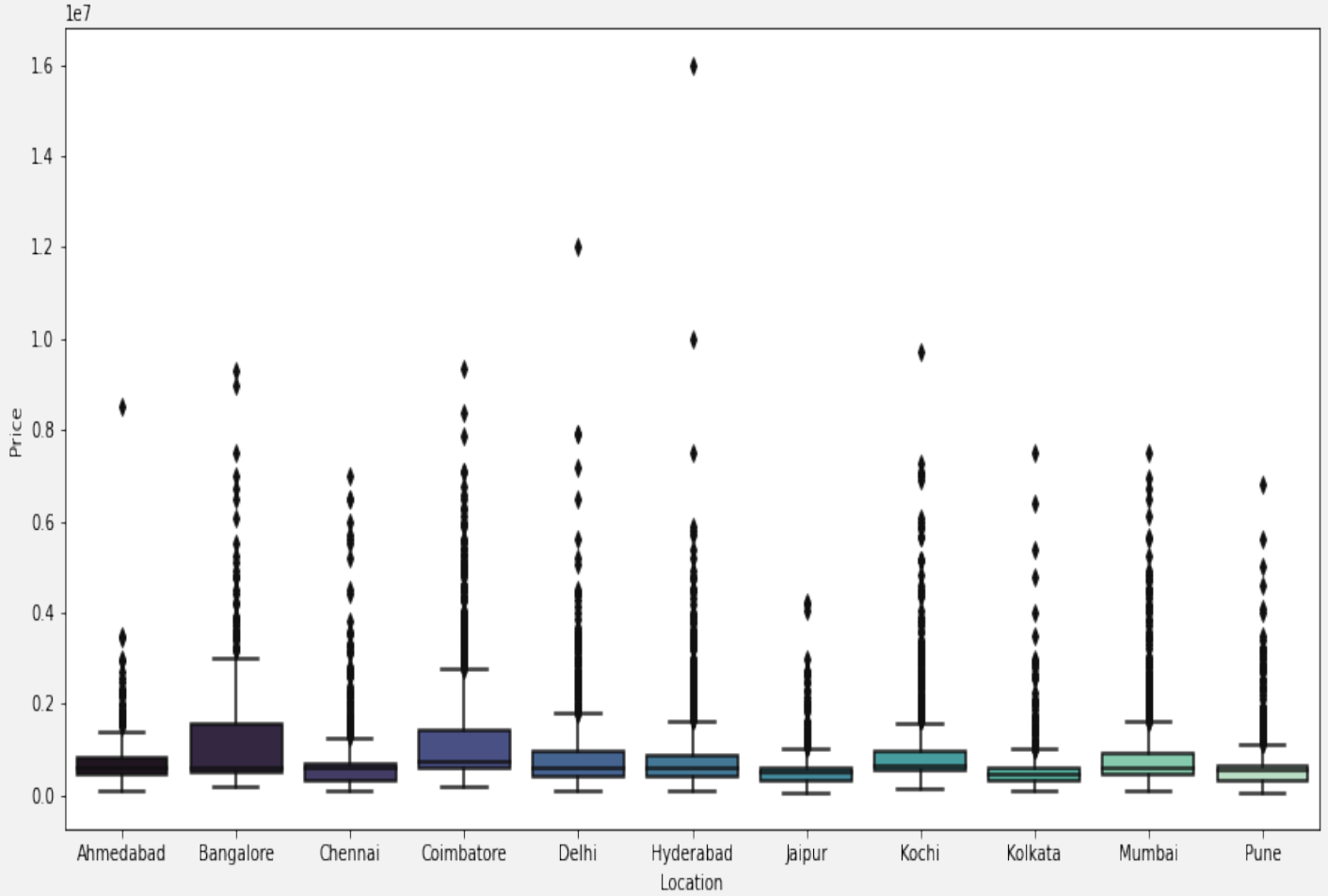
EXPLORATORY DATA ANALYSIS

PRICE VS FUEL TYPE



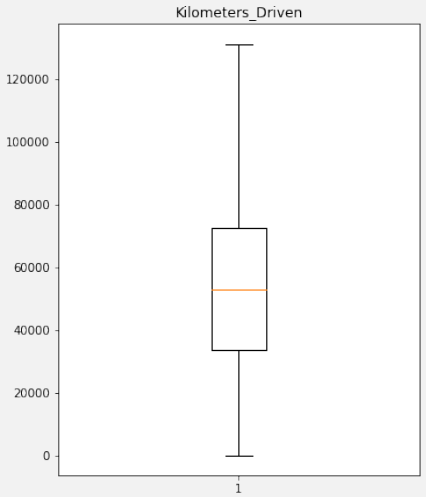
- Diesel is obviously higher priced followed by Petrol.
- Same can be of the preference by Buyers.
- We can see that these two fuel types are the reason for the outliers
- These will be treated

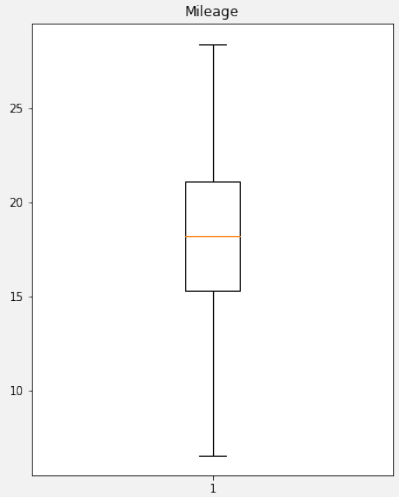
PRICE VS LOCATION

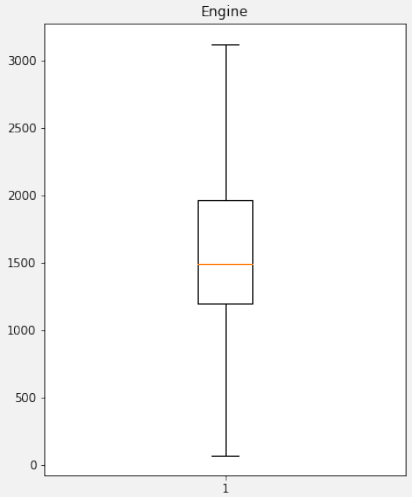


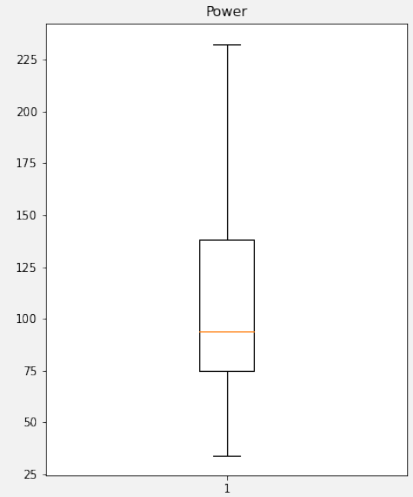
- Bangalore accounts for the most priced cars in India followed by Coimbatore
- Interestingly, Mumbai with the highest car inventory, has averagely priced cars on sale
- Delhi & Kochi and Hyderabad return regular price regimes for used cars in stock

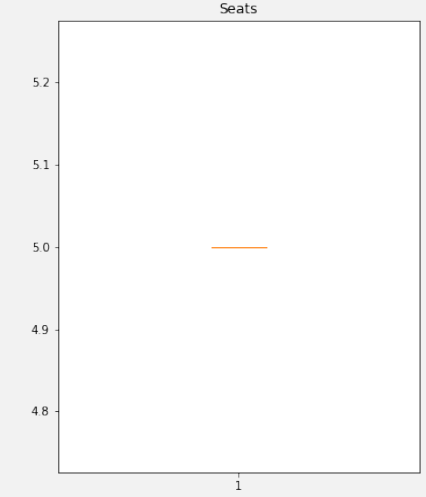
OUTLIER TREATMENT

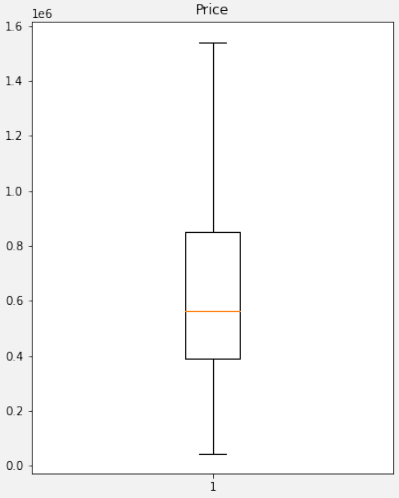












As seen. Outliers Treated

MODEL PERFORMANCE SUMMARY

OVERVIEW OF ML MODELS AND PARAMETERS

- Prior to Modeling, Outliers were treated and further Data-Preprocessing was done to identify independent variables viable for the prediction process
- Next step was to create dummy variables using a function that automates On-Hot encoding for a more surgical approach
- Finally, we imported the Linear Regression function from the Sklearn library in Python to model, Train and evaluate independent variables (X) against a target or dependent variable(Y=Price) in this case.
- The following parameters were generated: Intercept of the Linear Equation: 6.95 and a Matrix of Co- efficients
- Mean Absolute Error (MAE) on Test : 1.69 : Effectively, MAE describes the typical magnitude of the residuals.
- Root Mean Square Error (RMSE) on Test : 2.33
- R-Squared score on Testing : 0.69 ; R-Squared score on Training : 0.697 ; Adjusted R-Squared score on Modeling: 0.915

*R*2: (coefficient of determination) is a regression metric which tells us the amount of variance explained. Best possible score is 1.0, and it can be negative because the model can be arbitrarily worse.

A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

***R*2 value is 0.69, which means that in this model's independent variables are able to explain 69% of the variance in the dependent variable**

FACTORS EMPLOYED FOR ML PREDICTION

- **Adjusted. R-squared**: It reflects the fit of the model.
  - R-squared values range from 0 to 1, where a higher value generally indicates a better fit, assuming certain conditions are met.
  - In our case, the value for Adj. R-squared is **0.915**, which is excellent!
- **Constant coefficient** is the Y-intercept.
  - It means that if all the dependent variables (features: like Country, status, Adult mortality and so on..) coefficients are zero, then the expected output (i.e., the Y) would be equal to the const coefficient. In our case: 6.95
- **Schooling coeff**: It represents the change in the output Y due to a change of one unit in the Schooling (everything else held constant).
- **std err**: It reflects the level of accuracy of the coefficients.
  - The lower it is, the higher is the level of accuracy.
- **P >|t|**: It is p-value.
  - Ho : Independent feature is not significant
  - Ha : Independent feature is that it is significant
- Pr(>|t|) gives P-value for each independent feature to check that null hypothesis. we are considering 0.05 (5%) as significance level
- A p-value of less than 0.05 is considered to be statistically significant
- **ASSUMPTIONS:**
  1. Little or No Multicollinearity. (Passed)
  2. Mean of residuals should be 0 (Passed)
  3. No Heteroscedacity (Passed) .
  4. Linearity of variables (Passed)
  5. Normality of error terms (Passed)

CONCLUSIONS-KEY INSIGHTS FROM MODELING

- Kilometer Driven and Age rates come out to be very significant, as expected. As they increase, the Price decreases, as visible in the negative co-efficient
- A 1 unit increase in Manual Transmission Type, leads to a decrease in Price by 1.0162
- Equally, any increase in a unit of the second owner type and third owner type respectively leads to a decrease in Price
- Engine and Power have a very strong effect on Price as an increase in the Brake Horse power and the Engine Capacity apparently increase the Price of the vehicle
- An increase in any unit of the Fuel Types leads to an automatic increase in the Price

CONCLUSIONS-MAJOR KEY INSIGHTS

- Bangalore accounts for the most priced cars in India followed by Coimbatore
- Interestingly, Mumbai with the highest car inventory, has averagely priced cars on sale
- Mumbai (13.2%), Hyderabad(12,0%) and Kochi(10.8%) respectively account for over one-third of units on sale across the 11 different locations
- Ahmedabad has the lowest inventory of use cars on sale
- Diesel used cars topped potential sales at 53.3% followed by Petrol
- Stock for used cars of CNG, LPG and Electric fuel types are at a distant low
- Dealers preference for used cars by First owners trumped the rest by 82.4% followed by a distant 15.6% of Second Owner types.
- There is a very low affinity for used cars by Third and fourth owners.
- Used cars with Manual transmission is the most predominant amongst car dealers with a 71.6% inventory position
- Automatic transmission is higher priced than its manual counterpart in the used car market
- Maruti (20%), Hyundai(18%) and Honda (10.4%) respectively account for availability compared to other brands in the distribution
- Ford Brand came a distant 4th for used cars on sale at 4.8%
- The higher the Engine capacity and brake horse power, the more robust the fuel type (Diesel) and thus the higher the Price
- First owners with Diesel fuel types are higher priced followed by second owner types
- The older a used car vehicle, the lesser the Price

BUSINESS RECOMMENDATIONS

Based on the key insights regarding the Pricing of Used Car vehicles, generated, the following are ideal recommendations to the board;

- These insights can be inferred to form the basis of a formidable marketing vis-à-vis advert campaigns to gain a competitive edge, more market share and ultimately increased revenue
- Highly Priced vehicular inventory in Bangalore should have premium resources channeled to cater to high net worth in view of opportunities to grow the group's bottom line
- As evidenced in the analysis, an extensive and aggressing marketing initiative ought to be considered in the lower priced inventory locations to encourage sales
- Restocking of Inventories should be purely based on factors that favor pricing to improve customer satisfaction, patronage and ultimately revenue
- Price Estimation should be peculiar to each location as the standard of living across cities varies which ultimately is reflective of the purchasing power of the potential buyers
- It is quite apparent the training data is not exhaustive. More effort should be geared toward leveraging on a wider observation sample to draw more inclusive, extensive and impactful insights
- Following from the statement, dealerships should incline toward stocking newer used car models (lower aged) but equally strive to strike a balance between both extremes as a function of the standard of living per location. Same applies to Fuel types.
- Models vis-à-vis ownership types should follow the aforementioned recommendations to endear patronage to used cars based on these core determinants to design their campaigns per location while allocating resources more effectively toward increasing their bottom-lines

THANK YOU