Data Science Project report

Title: Variation of prices of the sold cars according to their features in India.

Group members:

1. Name : Tejas DR , SRN : PES1UG19CS537
2. Name : Vittal Gudadinni , SRN: PES1UG19CS580
3. Name : Y Nikhil Bharadwaj , SRN: PES1UG19CS586
4. Name : Yashas LS , SRN : PES1UG19CS590

Abstract:

The data set chosen was "prices and details of sold cars in India". To start it will include the information about the sold cars ex : the kilometers it was driven until the survey time , mileage , company, power, fuel type etc. , We are going to perform various statistical operations ex : data cleaning, plotting suitable graphs pertaining to the data chosen, normalizing the data ,checking for normal plots, conducting hypothesis testing , finding the correlation between variables etc.,

Introduction:

The problem statement stated is “Variation of prices of the sold cars according to their features in India”. How the prices of different cars are going to vary based on their features like company, mileage, power, engine capacity etc. Statistical data analysis is a procedure of performing various statistical operations. It is a kind of quantitative research, which seeks to quantify the data, and typically, applies some form of statistical analysis. Depending upon the number of variables, the researcher performs different statistical techniques. We took the data set and performed statistical operations on it to analyze and get useful information from it.

Also there are many other statistical ways to determine the results as mentioned above. By doing this we can easily analyze the data and draw useful information from the obtained results ex: if a person wants to buy a car he can get clear picture by observing. We use python for performing operations throughout our dataset.

Data Set:

We have taken this data set from the net link:
https://github.com/tdr652/hidude/blob/main/Prices%20and%20details%20of%20Used%20Cars%20in%20India.csv
This data set contains main information about the sold cars, the main features of this data set are: Kilometers driven, Fuel type, company of the vehicle, transmission, Mileage, Engine(CC), Power, Seats, Price, Owner type, Fuel type, Year of purchase.

It contains 12 columns and 661 rows. There are 7 categorical columns and 5 numerical columns. Columns: company name, Location, Year, Fuel type, transmission, owner type, and seats belong to the categorical set and columns: Kilometers driven, Mileage (kmpl), Engine (CC), Power (bph) and Price (L) belongs to numerical set. In numerical columns there are two discrete

data columns and three continuous data columns.

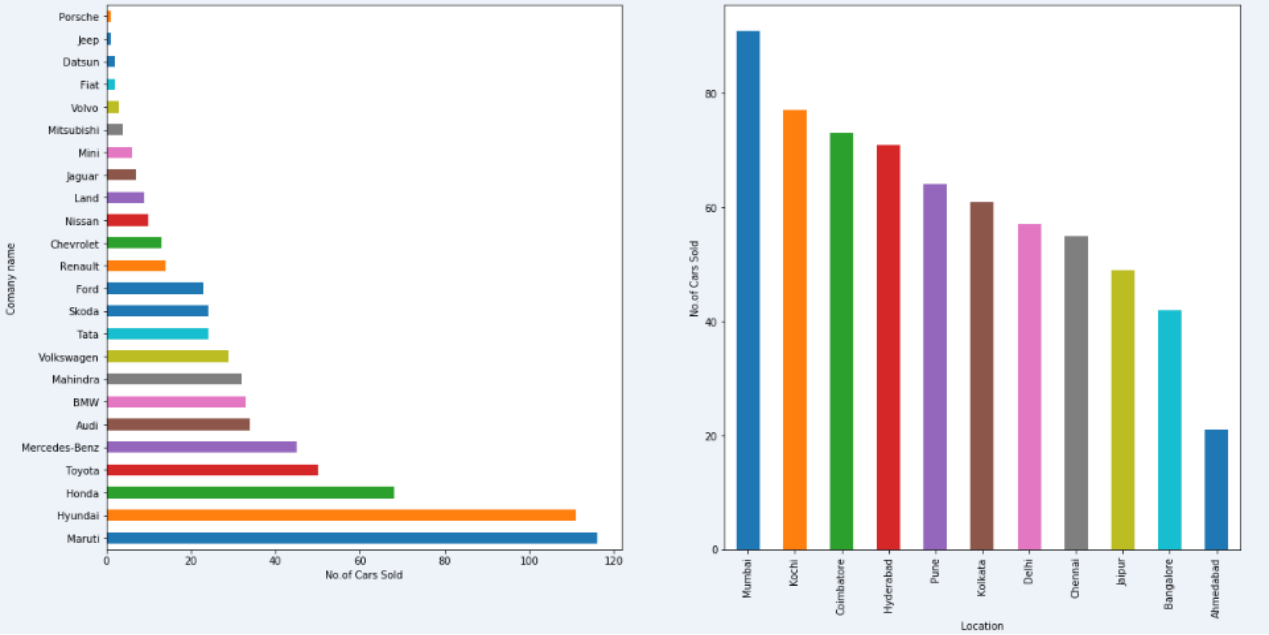
Preprocessing or Data cleaning:

About 3.2% of data contains NAN values. These NAN values can't be just ignored so it has to be filled with some values so for :
Categorical : mode
Numerical : mean or median
Is calculated with respect to those columns and then filled. Outlier is which a data point which is located outside the whiskers of box plot. The outliers were found with the help of box plot graph. We used IQR technique to filter the outliers. The rule of this technique is that anything not in the range of (Q1 – 1.5 IQR) and (Q3 + 1.5 IQR) is an outlier, and can be removed.

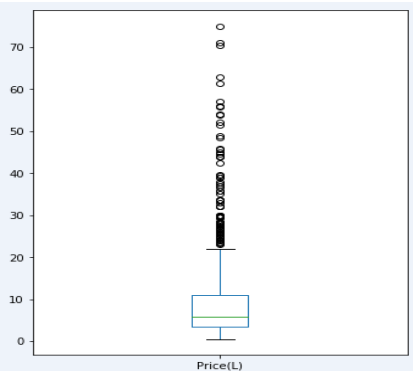
Data cleaning is very important because they do not cause experimental errors in the observation, deviation among the data is very less and there will be less variability.

Exploratory Data Analysis:

Graphical representation of the data is better as it can easily be understood by looking at it.
We usually use bar charts for categorical data and histograms or boxplots for numerical data.

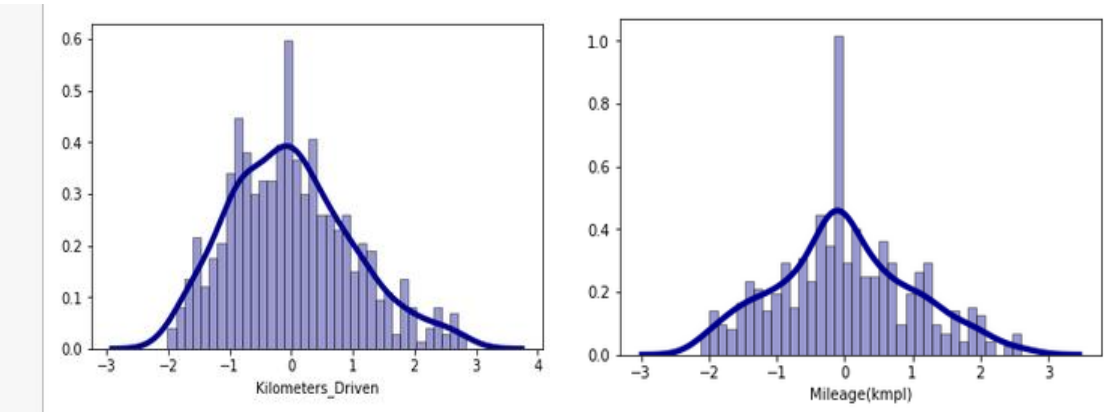


As we can see the cars of Maruti Company were bought by most people and cars belonging to Volvo Company were bought by the least people. Similarly on location basis Mumbai sold the most cars and Ahmedabad sold the least cars. Similarly the plots were plotted for the rest of the columns and insights were made.



This is graph pertaining to the Price (L) column which is numerical and as you can see the mean price lies between 6 to 7 lakhs and as you can see there are a lot of outliers which are removed further by the method mentioned above in the data cleaning process. Similarly plots were plotted for other numerical columns and insights were made. After removing outliers it was observed that the rows were decreased from 661 to 546.

After normalizing and plotting the histograms for the numerical columns it was seen that columns Kilometers_Driven and Mileage (kmpl) followed approximately normal plot.



Hypothesis testing:

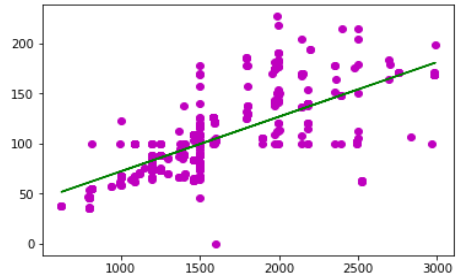
Hypothesis Testing is basically an assumption that we make about the population parameter. Hypothesis testing is an essential procedure in statistics. A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. We used the z-value approach for obtaining the result.
We considered the column Kilometers_Driven and Mileage (kmpl).

For Kilometers_Driven column:
The following hypothesis was chosen
Ho: $\mu = 57000$
Ha: μ != 57000
we draw a sample of size 60
mean: 55885.32
actual z-value: 1.9599639845400545
hypo z-value: 0.315433661947
Since the hypo z-value is less than actual z-value we fail to reject the NULL hypothesis.

For Mileage (kmpl) column:
The following hypothesis was chosen
Ho: $\mu >= 20$
Ha: $\mu < 20$
we draw a sample of size 25
mean: 17.64
actual z-value: 1.6448536269514729
hypo z-value: 3.18059299191
Since the hypo z-value is less than actual z-value we reject the NULL hypothesis.

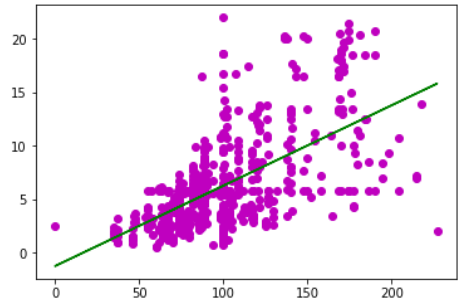
Correlation:

When a scatter plot was plotted between Engine and Power column we see a positive relation between them. The following insights were observed:
the correlation coefficient is: 0.7252088020109353



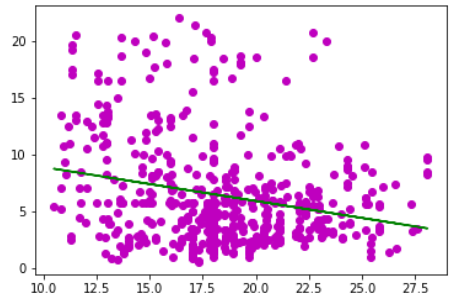
Coefficients: [0.05452268]
RMSE: 628.39

When a scatter plot was plotted between Power and Price column we see a positive relation between them. The following insights were observed:
the correlation coefficient is: 0.604957972663688



Coefficients: [0.07511686]
RMSE: 12.96

When a scatter plot was plotted between Mileage and Price column we see a slight negative relation between them. The following insights were observed:
the correlation coefficient is: -0.24514088476365808



Coefficients: [-0.29839877]
RMSE: 19.21
As you can see the line is almost horizontal

Results and Conclusions:

We can conclude that as the Engine capacity increases the power increases and in turn price increases with power and engine as seen by the statistical tests performed, hence cars with high Engine capacity and power has high price. Price of the cars slightly decreases with increase in mileage, but we can say almost there is no relation between them.