Experiment 02 - Exploratory Data Analysis and Visualization

| Roll No. |  |
| --- | --- |
| Name |  |
| Class | D15-A |
| Subject | DS using Python Lab |
| LO Mapped | LO1: Understand the concept of Data science process and associated terminologies to solve real-world problems  LO2: Analyze the data using different statistical techniques and visualize the outcome using different types of plots. |
|  |  |

**Aim**:

1. To perform Data Visualization for the selected data set using Matplotlib and Seaborn
2. To perform Exploratory Data Analysis for the selected data set

**Introduction**:

**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is the first step in your data analysis process developed by “John Tukey” in the 1970s. In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. By the name itself, we can get to know that it is a step in which we need to explore the data set.

When you are trying to build a machine learning model you need to be pretty sure whether your data is making sense or not. The main aim of exploratory data analysis is to obtain confidence in your data to an extent where you’re ready to engage a machine learning algorithm.

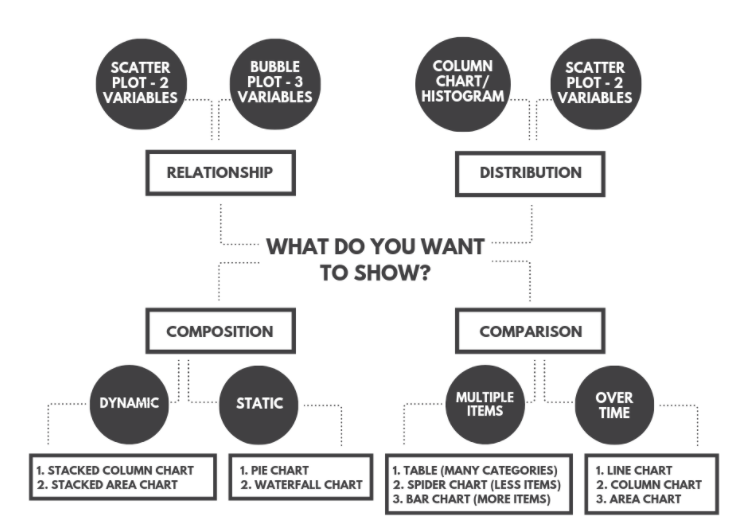
**Why do we do EDA?**

Exploratory Data Analysis is a crucial step before you jump to machine learning or modeling your data. By doing this you can get to know whether the selected features are good enough to model, are all the features required, are there any correlations based on which we can either go back to the Data Preprocessing step or move on to modeling.

Once EDA is complete and insights are drawn, its feature can be used for supervised and unsupervised machine learning modeling.

In every machine learning workflow, the last step is Reporting or Providing the insights to the Stakeholders and as a Data Scientist you can explain every bit of code but you need to keep in mind the audience. By completing the EDA you will have many plots,heat-maps, frequency distribution, graphs, correlation matrix along with the hypothesis by which any individual can understand what your data is all about and what insights you got from exploring your data set.

Data visualization is very critical to market research where both numerical and categorical data can be visualized, which helps in an increase in the impact of insights and also helps in reducing the risk of analysis paralysis



Advantages of Data visualization:

1. Better Agreement:

In business, for numerous periods, it happens that we need to look at the exhibitions of two components or two situations. A conventional methodology is to experience the massive information of both the circumstances and afterward examine it. This will clearly take a great deal of time.

1. A Superior Method:

It can tackle the difficulty of placing the information of both perspectives into the pictorial structure. This will unquestionably give a superior comprehension of the circumstances. For instance, Google patterns assist us with understanding information identified with top ventures or inquiries in pictorial or graphical structures.

1. Simple Sharing of Data:

With the representation of the information, organizations present another arrangement of correspondence. Rather than sharing the cumbersome information, sharing the visual data will draw in and pass on across the data which is more absorbable.

1. Deals Investigation:

With the assistance of information representation, a salesman can, without much of a stretch, comprehend the business chart of items. With information perception instruments like warmth maps, he will have the option to comprehend the causes that are pushing the business numbers up just as the reasons that are debasing the business numbers. Information representation helps in understanding the patterns and furthermore, different variables like sorts of clients keen on purchasing, rehashing clients, the impact of topography, and so forth.

1. Discovering Relations Between Occasions:

A business is influenced by a lot of elements. Finding a relationship between these elements or occasions encourages chiefs to comprehend the issues identified with their business. For instance, the online business market is anything but another thing today. Each time during certain happy seasons, like Christmas or Thanksgiving, the diagrams of online organizations go up. Along these lines, state if an online organization is doing a normal $1 million business in a specific quarter and the business ascends straightaway, at that point they can rapidly discover the occasions compared to it.

1. Investigating Openings and Patterns:

With the huge loads of information present, business chiefs can discover the profundity of information in regard to the patterns and openings around them. Utilizing information representation, the specialists can discover examples of the conduct of their clients, subsequently preparing for them to investigate patterns and open doors for business.

**Descriptive analysis - Central tendency**

A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. We can think of it as the tendency of data to cluster around a middle value. In statistics, the three most common measures of central tendency are the **mean**, **median**, and **mode**. Each of these measures calculates the location of the central point using a different method. Choosing the best measure of central tendency depends on the type of data we have

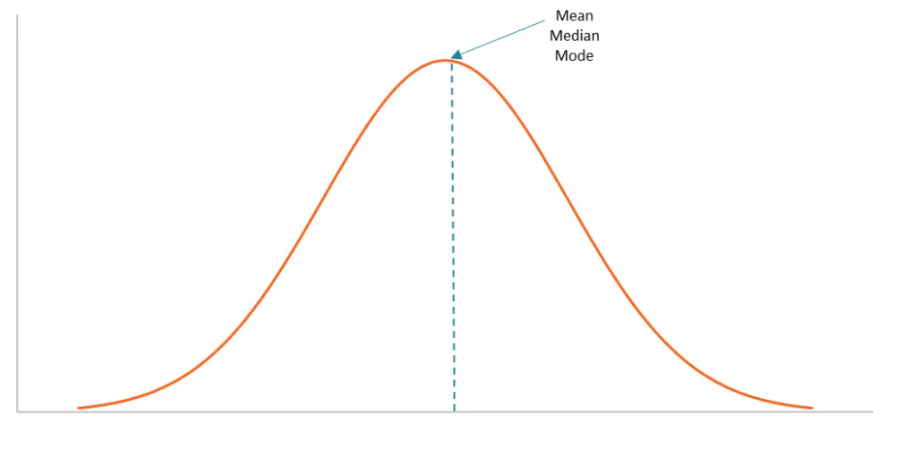
1. **Mean (Average):** Represents the sum of all values in a dataset divided by the total number of the values.
2. **Median:** The middle value in a dataset that is arranged in ascending order (from the smallest value to the largest value). If a dataset contains an even number of values, the median of the dataset is the mean of the two middle values.
3. **Mode:** Defines the most frequently occurring value in a dataset. In some cases, a dataset may contain multiple modes, while some datasets may not have any mode at all.

Even though the measures above are the most commonly used to define central tendency, there are some other measures, including, but not limited to, geometric mean, harmonic mean, midrange, and geometric median.

The selection of a central tendency measure depends on the properties of a dataset. For instance, the **mode** is the only central tendency measure for categorical data, while a **median** works best with ordinal data.

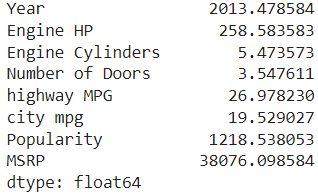
Although the **mean** is regarded as the best measure of central tendency for quantitative data, that is not always the case. For example, the mean may not work well with quantitative datasets that contain extremely large or extremely small values. The extreme values may distort the mean. Thus, you may consider other measures.

The measures of central tendency can be found using a formula or definition. Also, they can be identified using a frequency distribution graph. Note that for datasets that follow a normal distribution, the mean, median, and mode are located on the same spot on the graph.

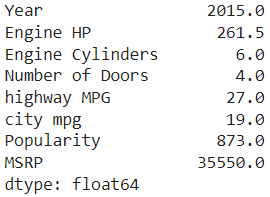


**Output:**

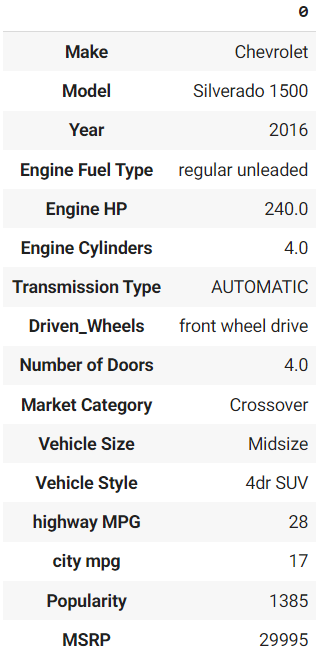
df.mean()



df.median()



df.mode().T

****

Inferences:

* Year, MSRP and MSRP have high differences between their mean and median. Hence, their data is a skewed distribution.
* Highway MPG and City MPG have very similar mean and median. Hence, their data is a symmetrical distribution.
* Most cars were made in 2016.
* Most cars have 4 doors.

**Descriptive analysis - Dispersion and Distribution**

A **distribution** tells you how likely certain events are, e.g For the normal distribution (continuous) you can talk about the probability that you get a number between 3 and 7, or for a discrete distribution like the Poisson, how likely are you to get 3?

A measure of **dispersion** tells you, if you see many events happen, how spread out are they going to be.

Dispersion:

Most measures of dispersion have the same units as the quantity being measured. In other words, if the measurements are in meters or seconds, so is the measure of dispersion.There are two main types of dispersion methods in statistics which are:

1. **Absolute Measure of Dispersion**

An absolute measure of dispersion contains the same unit as the original data set. Absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or mean deviations. It includes range, standard deviation, quartile deviation, etc. The types of absolute measures of dispersion are:

* 1. **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6
  2. **Variance:** Deduct the mean from each data in the set then squaring each of them and adding each square and finally dividing them by the total no of values in the data set is the variance. Variance (σ2)=∑(X−μ)2/N
  3. **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. = √σ.
  4. **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
  5. **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

1. **Relative Measure of Dispersion**

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

* 1. **Coefficient of Range**
  2. **Coefficient of Variation**
  3. **Coefficient of Standard Deviation**
  4. **Coefficient of Quartile Deviation**
  5. **Coefficient of Mean Deviation**

Coefficient of Dispersion

The coefficients of dispersion are calculated (along with the measure of dispersion) when two series are compared, that differ widely in their averages. The dispersion coefficient is also used when two series with different measurement units are compared. It is denoted as C.D.

The common coefficients of dispersion are:

**Range** => C.D. = (Xmax – Xmin) ⁄ (Xmax + Xmin)

**Quartile Deviation =>** C.D. = (Q3 – Q1) ⁄ (Q3 + Q1)

**Standard Deviation** => C.D. = S.D. ⁄ Mean

**Mean Deviation** => C.D. = Mean deviation/Average

Distribution:

When we use the term normal distribution in statistics, we usually mean a probability distribution. Good examples are the Normal distribution, the Binomial distribution, and the Uniform distribution.

A distribution in statistics is a function that shows the possible values for a variable and how often they occur.

Before we jump on to the explanation of distributions, let’s see what kind of data we can encounter. The data can be discrete or continuous.

1. **Discrete Data:** As the name suggests, can take only specified values. For example, when you roll a die, the possible outcomes are 1, 2, 3, 4, 5 or 6 and not 1.5 or 2.45.
2. **Continuous Data:** Itcan take any value within a given range. The range may be finite or infinite. For example, A girl’s weight or height, or the length of the road. The weight of a girl can be any value from 54 kgs, or 54.5 kgs, or 54.5436kgs.

Types of Distributions

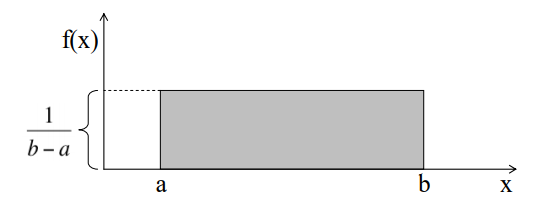
1. **Bernoulli Distribution:**
   1. A Bernoulli distribution has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.
   2. The probability mass function is given by: px(1-p)1-x where x € (0, 1).
   3. It can also be written as



1. **Uniform Distribution:**
   1. When we roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution.
   2. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.
   3. A variable X is said to be uniformly distributed if the density function is:



* 1. The graph of a uniform distribution curve looks like



1. **Binomial Distribution:**

The properties of a Binomial Distribution are

* 1. Each trial is independent.
  2. There are only two possible outcomes in a trial - either a success or a failure.
  3. A total number of n identical trials are conducted.
  4. The probability of success and failure is the same for all trials. (Trials are identical.)
  5. The mathematical representation of binomial distribution is given by:



1. **Normal Distribution:**

The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:

* 1. The mean, median and mode of the distribution coincide.
  2. The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
  3. The total area under the curve is 1.
  4. Exactly half of the values are to the left of the center and the other half to the right.

The PDF of a random variable X following a normal distribution is given by:



The mean and variance of a random variable X which is said to be normally distributed is given by:

Mean => E(X) = µ

Variance => Var(X) = σ^2

1. **Poisson Distribution:**

A distribution is called Poisson distribution when the following assumptions are valid:

* 1. Any successful event should not influence the outcome of another successful event.
  2. The probability of success over a short interval must equal the probability of success over a longer interval.
  3. The probability of success in an interval approaches zero as the interval becomes smaller.

1. **Exponential Distribution:**

Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result

Some examples of Exponential distributions are:

* 1. Length of time between metro arrivals,
  2. Length of time between arrivals at a gas station
  3. The life of an Air Conditioner

**Output:**

Q1 = df.quantile(0.25)

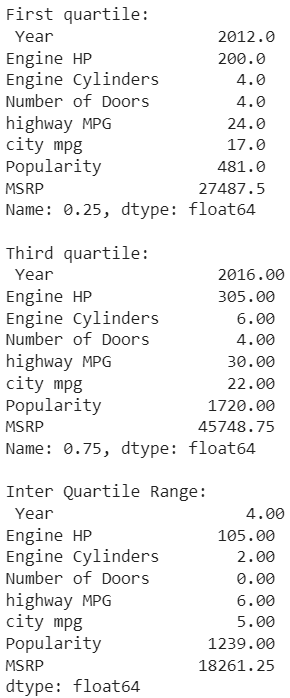
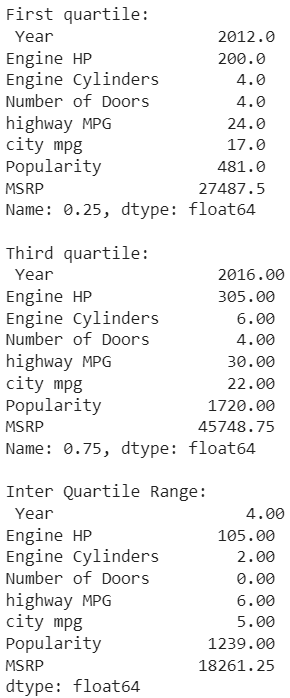
print("First quartile:\n",Q1)

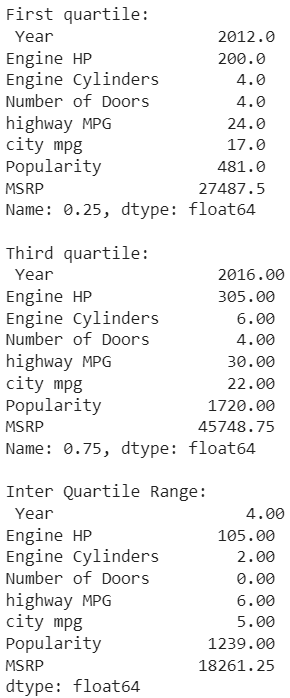
Q3 = df.quantile(0.75)

print("\nThird quartile:\n",Q3)

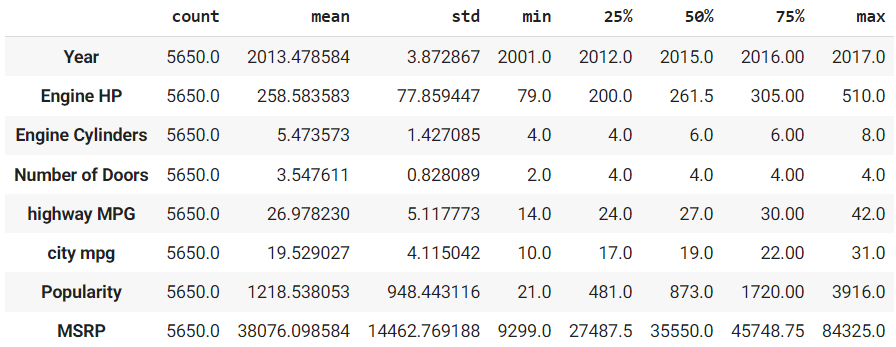
IQR = Q3 - Q1

print("\Inter Quartile Range:\n",IQR)



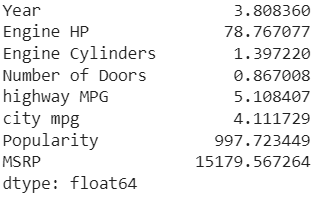


df.describe().T



#Standard deviation

df.std()



**Inferences:**

1. Majority of cars have Highway MPG between 24 and 30, as the first quartile is 24 and the third quartile is 30. The median is 27 which shows that this data has a symmetrical distribution.
2. Almost all of the cars manufactured have four doors, as both first and third quartile are equal to 4.
3. Manufactured cars give better Highway MPG as compared to city MPG.
4. The standard deviation values of the dataset show that Engine HP, Popularity and MSRP have the most spread out values among all cars manufactured.

**Correlation**

Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.

Correlation, statistical technique which determines how one variable moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It’s a bi-variate analysis measure which describes the association between different variables. In most of the business it’s useful to express one subject in terms of its relationship with others.

**Types of Correlation:**

Based on the degree of correlation:

**1. Positive Correlation:**

Two features (variables) can be positively correlated with each other. It means that when the value of one variable increases then the value of the other variable(s) also increases.

**2. Negative Correlation:**

Two features (variables) can be negatively correlated with each other. It means that when the value of one variable increases then the value of the other variable(s) decreases.

**3. No Correlation:**

Two features (variables) are not correlated with each other. It means that when the value of one variable increases or decreases then the value of the other variable(s) doesn’t increase or decrease.

Based on the change in proportion:

**1. Linear:** If the value of the amount of change in one variable tends to preserve a constant ratio to the amount of change in other variables, then the correlation is said to be linear. For Example, Whenever the price rises by 10%, then supply rises by 20%.

**2. Non-linear:** If the value of the amount of change in one variable does not preserve a constant ratio to the amount of change in the other variables, then it is said to be a Non-linear correlation. It is also known as the Curvilinear correlation. For Example, Whenever price rises by 10%, but supply rises sometimes by 20%, sometimes by 10%, and sometimes by 40%.

Based on the number of variables studied:

**1. Simple Correlation:** When we consider only two variables(Bivariate analysis) and check the correlation between only those variables, it is said to be a Simple Correlation. For example, Price and demand, Height and Weight, Income and consumption, etc.

**2. Multiple Correlation:** When we consider more than three or three variables for correlation simultaneously, it is termed Multiple Correlation. For example, When we study the relationship between the yield of rice per hectare and both the amount of rainfall along with the number of fertilizers are used to find the relationship with rice production.

**3. Partial Correlation:** When one or more variables are kept constant and the relationship is studied between the remaining variables, then it is termed Partial Corr. Study the relationship between 2 variables and assume other variables are constant. For example, Relationship between rainfall and rice yields under constant temperature.

**Correlation vs Causation**

Correlation:

It is a numerical measure of the direction and magnitude of the mutual relationship between the variables(X and Y).

Causation:

X is the cause of change in Y i.e, the change of Y is the effect of change in X.

NOTE:

– If X and Y are correlated then X and Y may or may not have a casual relationship.

– If X and Y have a causal relationship then X and Y must be correlated.

**Reasons Behind Correlation**

It may happen because of several reasons like:

1. Mutual dependence Between the variables:

Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect.

When two variables(X and Y) affect each other mutually, we cannot say X is the cause or Y is the cause.

For Example, The price of a commodity is affected by demand and supply.

2. Due to pure chance:

In a small sample, X and Y are highly correlated but in the universe X and Y are not correlated.

For Example, Correlation between income and weight of a person. This may be due to:

– Sampling fluctuations

– Bias of investigator in selecting the sample.

Such a relation is called a non-sense or spurious correlation.

3. Correlation due to any third common factor:

Both the correlated variables may be influenced by one or other variables.

– X and Y don’t have a direct correlation.

For Example, It is between the production of tea and rice per hectare. Here they are not directly correlated, instead the cause is the good rainfall well in time.

**Utility of Correlation**

1. It is very useful for Economists to study the relationships between variables.

2. It helps in measuring the degree of relationship between the variables.

3. We can also test the significance of the relationship.

4. Sampling error can also be calculated by knowing the correlation.

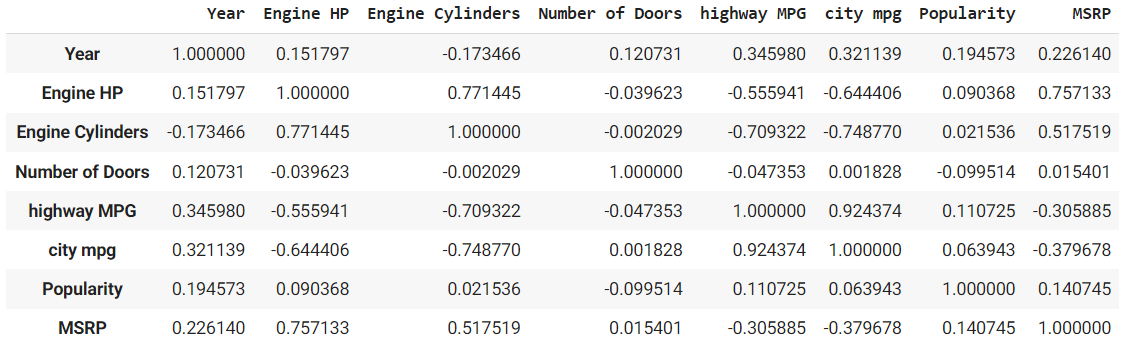
5. It is the basis for the study of regression.

6. Estimate the value of one variable based on the other variable.

7. It is used to determine the relationship between datasets in business.

**Output:**

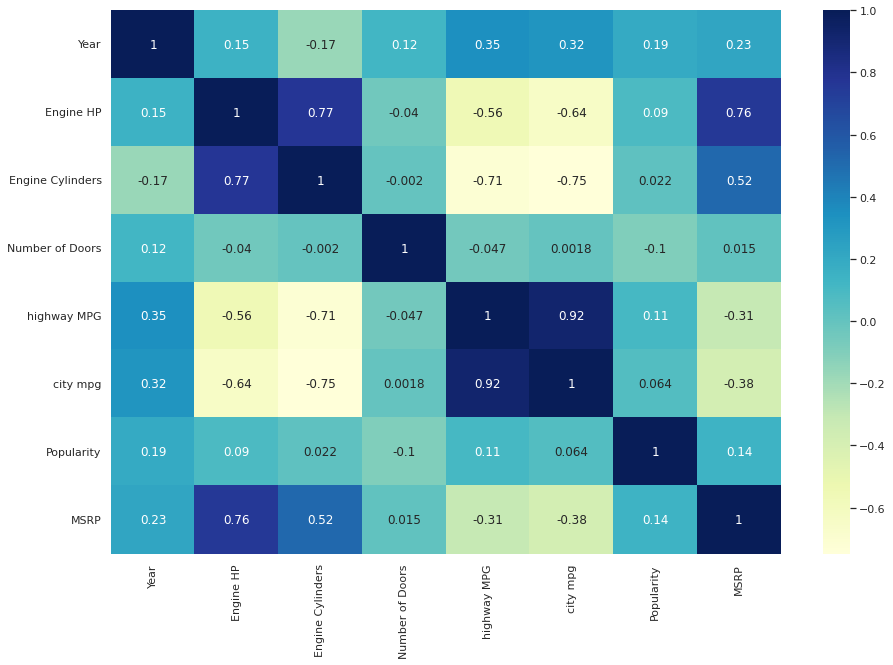
df.corr()



#Displaying a heatmap for this correlation matrix

plt.figure(figsize=(15,10))

sns.heatmap(df.corr(),cmap="YlGnBu",annot=True)

****

Inferences:

Positive correlation:

1. Cylinders and HP - higher the number of cylinders higher will be the horsepower.
2. Highway mpg and City mpg - higher the highway mpg higher will be the city mpg.

Negative correlation:

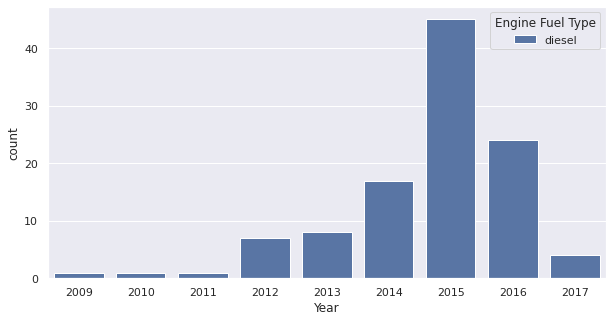
1. MPG and Cylinders - higher the number of cylinders lesser will be the MPG.
2. MPG and HP - higher the number of Power lesser will be the MPG.

No correlation:

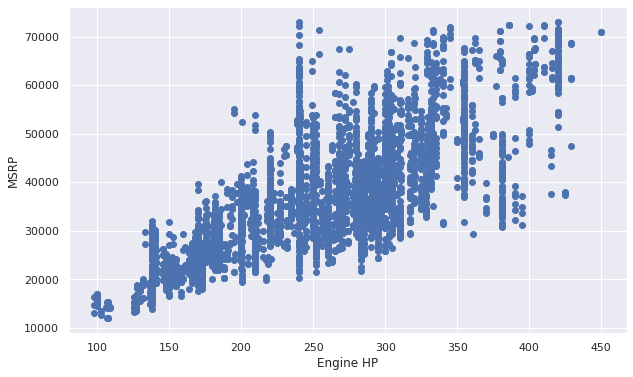
1. Engine HP and Popularity
2. City mpg and Popularity
3. Engine Cylinders and Popularity

**Formulate and Answer any 5 questions about the data**

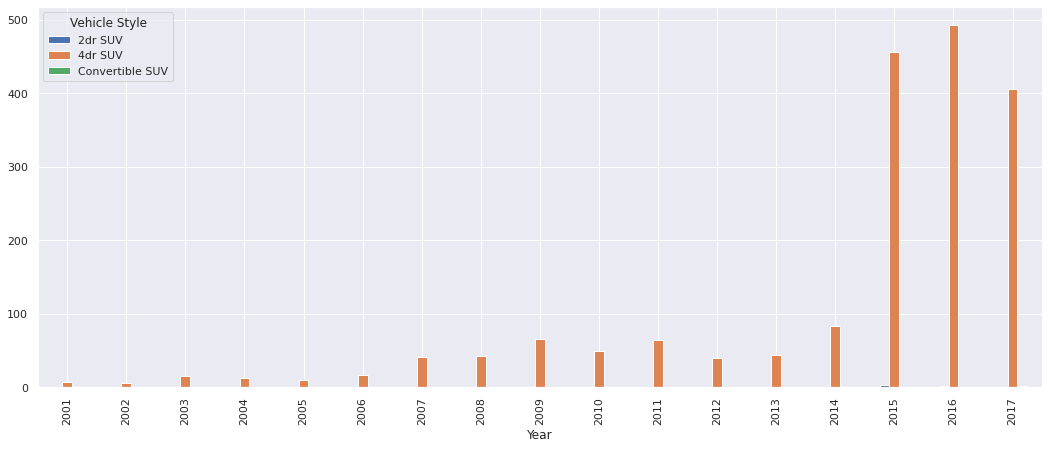
1. Has the use of diesel cars increased or decreased over the years?
2. Use of diesel cars increased steadily from 2010 to 2015. Thereafter, the usage has gone down drastically.



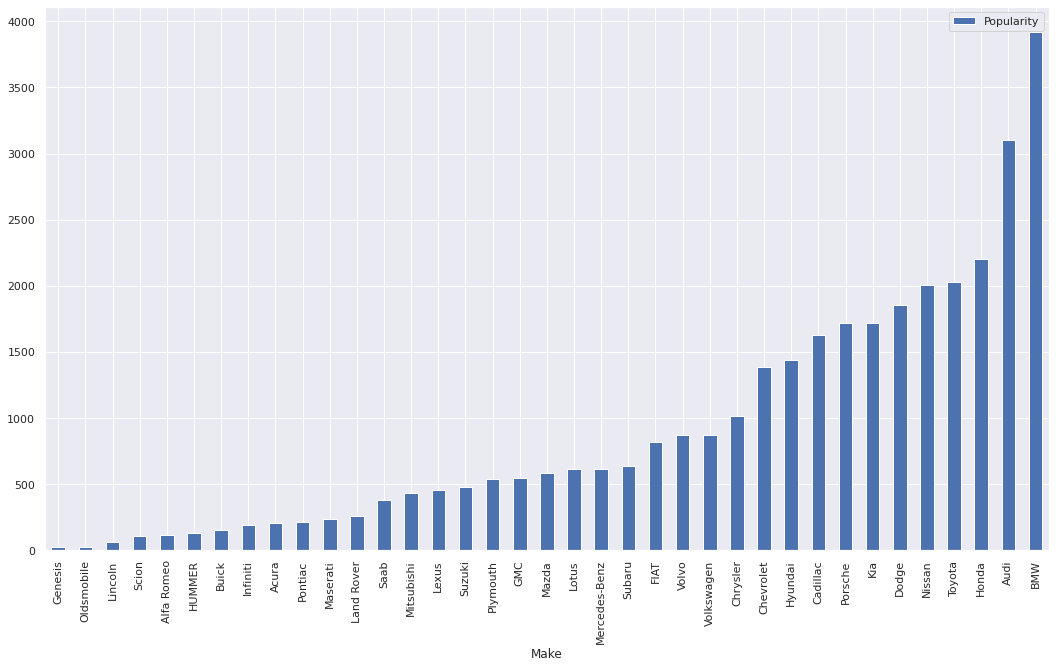
1. Do higher priced cars come with more powerful engines, that is, higher horsepower?
2. Scatterplot above suggests that, yes higher priced cars do provide higher engine horsepower.



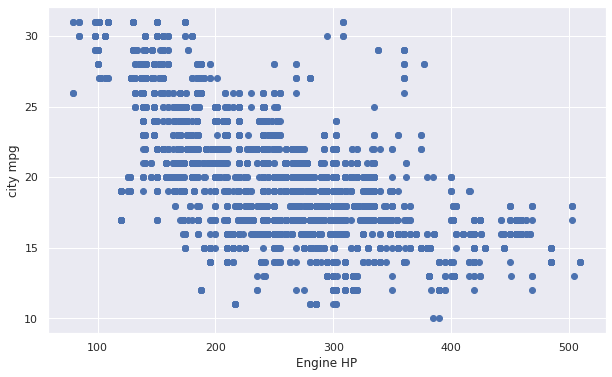
1. Has production of SUVs increased over the years?
2. From the graph we can clearly see that SUV production has gone up significantly as compared to previous decades.



1. Which brand is least popular and which one is most popular?
2. Genesis is least popular while BMW is the most popular brand.



1. Does higher miles per gallon (MPG) value provide higher horsepower?
2. Scatterplot above suggests that, no, higher MPG value does not provide higher power.



**Data Visualization**

Data visualization is defined as a graphical representation that contains the information and the data. By using visual elements like charts, graphs, and maps, data visualization techniques provide an accessible way to see and understand trends, outliers, and patterns in data.

In modern days we have a lot of data in our hands i.e, in the world of Big Data, data visualization tools, and technologies are crucial to analyze massive amounts of information and make data-driven decisions. It is used in many areas such as:

* To model complex events.
* Visualize phenomenons that cannot be observed directly, such as weather patterns, medical conditions, or mathematical relationships.

**Univariate Analysis Techniques for Data Visualization**

**1. Distribution Plot**

It is one of the best univariate plots to know about the distribution of data.

When we want to analyze the impact on the target variable(output) with respect to an independent variable(input), we use distribution plots a lot.

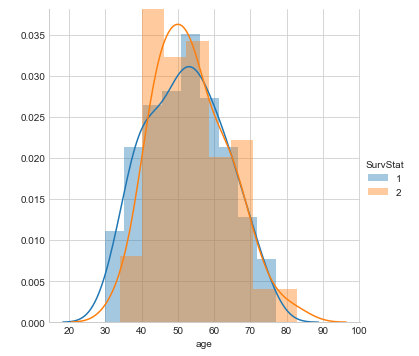
This plot gives us a combination of both probability density functions(pdf) and histogram in a single plot.

Implementation:

The distribution plot is present in the Seaborn package.

The code snippet is as follows:

sns.FacetGrid(hb,hue='SurvStat',size=5).map(sns.distplot,'age').add\_legend()



Some conclusions inferred from the above distribution plot:

* From the above distribution plot we can conclude the following observations:
* We have observed that we created a distribution plot on the feature ‘Age’(input variable) and we used different colors for the Survival status(output variable) as it is the class to be predicted.
* There is a huge overlapping area between the PDFs for different combinations.
* In this plot, the sharp block-like structures are called histograms, and the smoothed curve is known as the Probability density function(PDF).

NOTE:

The Probability density function(PDF) of a curve can help us to capture the underlying distribution of that feature which is one major takeaway from Data visualization or Exploratory Data Analysis(EDA).

**2. Box and Whisker Plot**

This plot can be used to obtain more statistical details about the data.

The straight lines at the maximum and minimum are also called whiskers.

Points that lie outside the whiskers will be considered as an outlier.

The box plot also gives us a description of the 25th, 50th,75th quartiles.

With the help of a box plot, we can also determine the Interquartile range(IQR) where maximum details of the data will be present. Therefore, it can also give us a clear idea about the outliers in the dataset.

Implementation:

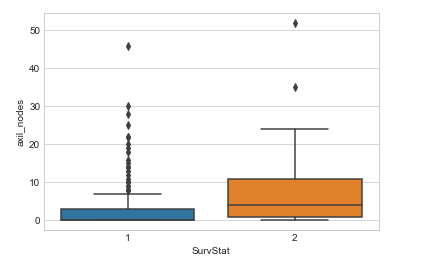
Boxplot is available in the Seaborn library.

Here x is considered as the dependent variable and y is considered as the independent variable. These box plots come under univariate analysis, which means that we are exploring data only with one variable.

Here we are trying to check the impact of a feature named “axil\_nodes” on the class named “Survival status” and not between any two independent features.

The code snippet is as follows:

sns.boxplot(x='SurvStat',y='axil\_nodes',data=hb)



Some conclusions inferred from the above box plot:

* From the above box and whisker plot we can conclude the following observations:
* How much data is present in the 1st quartile and how many points are outliers etc.
* For class 1, we can see that very little or no data is present between the median and the 1st quartile.
* There are more outliers for class 1 in the feature named axil\_nodes.

NOTE:

We can get details about outliers that will help us to well prepare the data before feeding it to a model since outliers influence a lot of Machine learning models.

**3. Violin Plot**

The violin plots can be considered as a combination of Box plot at the middle and distribution plots(Kernel Density Estimation) on both sides of the data.

This can give us the description of the distribution of the dataset like whether the distribution is multimodal, Skewness, etc.

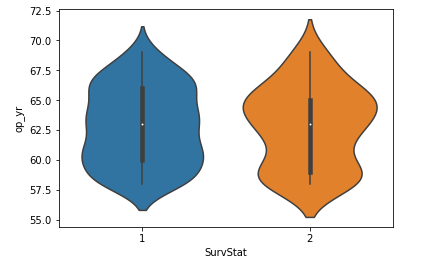
It also gives us useful information like a 95% confidence interval.

Implementation:

The Violin plot is present in the Seaborn package.

The code snippet is as follows:

sns.violinplot(x='SurvStat',y='op\_yr',data=hb,size=6)



Some conclusions inferred from the above violin plot:

* From the above violin plot we can conclude the following observations:
* The median of both classes is close to 63.
* The maximum number of persons with class 2 has an op\_yr value of 65 whereas, for persons in class1, the maximum value is around 60.
* Also, the 3rd quartile to median has a lesser number of data points than the median to the 1st quartile.

**4. Density Plots**

Another quick and easy technique for getting each attribute's distribution is Density plots. It is also like a histogram but having a smooth curve drawn through the top of each bin. We can call them as abstracted histograms.

Example

In the following example, Python script will generate Density Plots for the distribution of attributes of Pima Indian Diabetes dataset.

from matplotlib import pyplot

from pandas import read\_csv

path = r"C:\pima-indians-diabetes.csv"

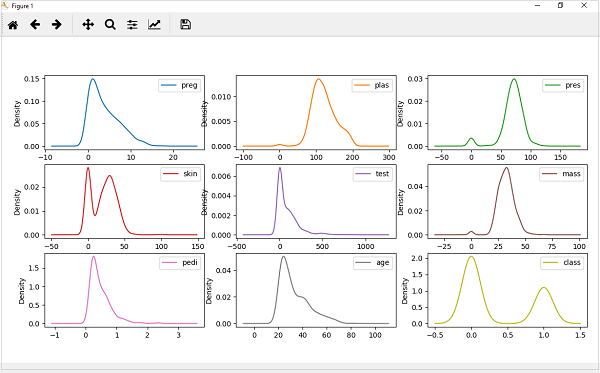
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']

data = read\_csv(path, names=names)

data.plot(kind='density', subplots=True, layout=(3,3), sharex=False)

pyplot.show()

Output



From the above output, the difference between Density plots and Histograms can be easily understood.

**Bivariate Analysis Techniques for Data Visualization**

**1. Line Plot**

This is the plot that you can see in the nook and corners of any sort of analysis between 2 variables.

The line plots are nothing but the values on a series of data points will be connected with straight lines.

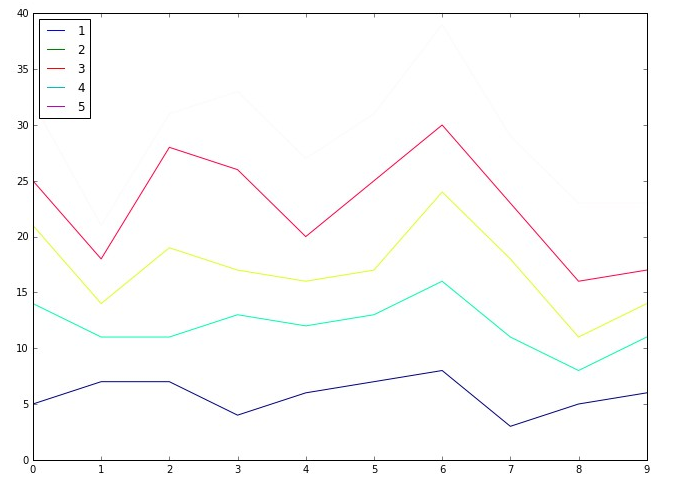
The plot may seem very simple but it has more applications not only in machine learning but in many other areas.

Implementation:

The line plot is present in the Matplotlib package.

The code snippet is as follows:

plt.plot(x,y)



Some conclusions inferred from the above line plot:

* From the above line plot we can conclude the following observations:
* These are used right from performing distribution Comparison using Q-Q plots to CV tuning using the elbow method.
* Used to analyze the performance of a model using the ROC- AUC curve

**2. Bar Plot**

This is one of the widely used plots, that we would have seen multiple times not just in data analysis, but we use this plot also wherever there is a trend analysis in many fields.

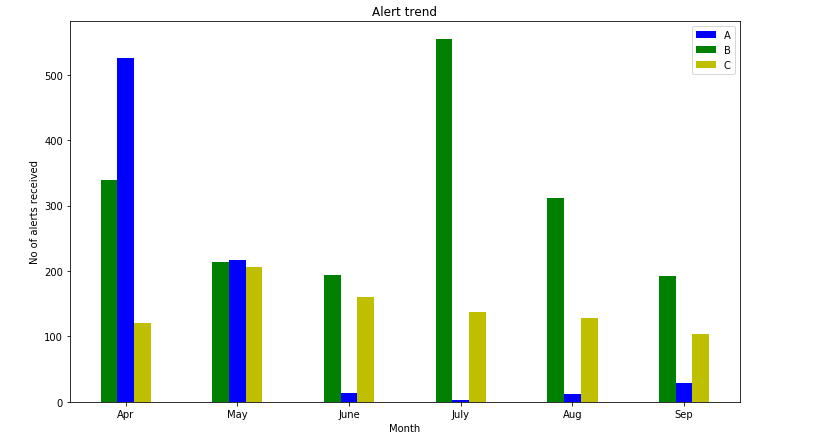
Though it may seem simple it is powerful in analyzing data like sales figures every week, revenue from a product, Number of visitors to a site on each day of a week, etc.

Implementation:

The bar plot is present in the Matplotlib package.

The code snippet is as follows:

plt.bar(x,y)



Some conclusions inferred from the above bar plot:

* From the above bar plot we can conclude the following observations:
* We can visualize the data in a cool plot and can convey the details straight forward to others.
* This plot may be simple and clear but it’s not much frequently used in Data science applications.

**3. Scatter Plot**

It is one of the most commonly used plots used for visualizing simple data in Machine learning and Data Science.

This plot describes us as a representation, where each point in the entire dataset is present with respect to any 2 to 3 features(Columns).

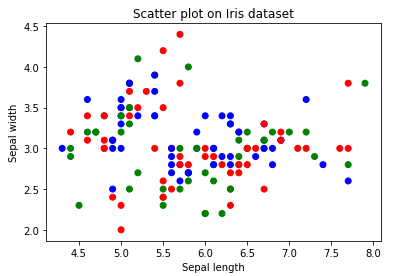
Scatter plots are available in both 2-D as well as in 3-D. The 2-D scatter plot is the common one, where we will primarily try to find the patterns, clusters, and separability of the data.

Implementation:

The scatter plot is present in the Matplotlib package.

The code snippet is as follows:

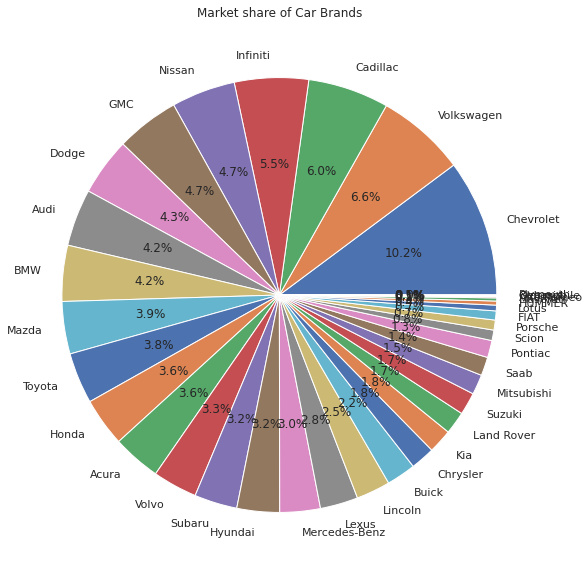
plt.scatter(x,y)



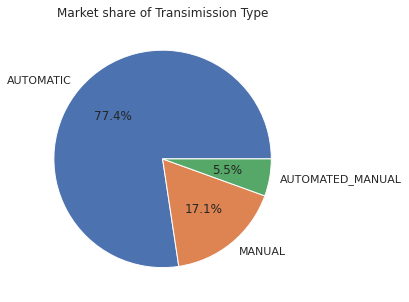
Some conclusions inferred from the above Scatter plot:

* From the above Scatter plot we can conclude the following observations:
* The colors are assigned to different data points based on how they were present in the dataset i.e, target column representation.
* We can color the data points as per their class label given in the dataset.

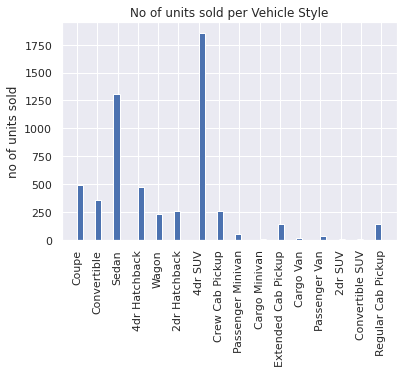
**Output:**



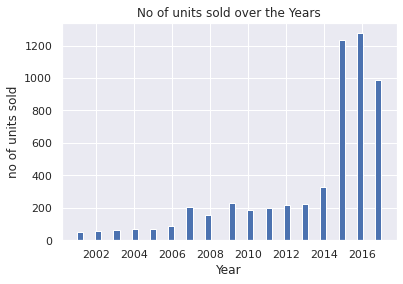
**Inference:** Chevrolet sold the most number of cars



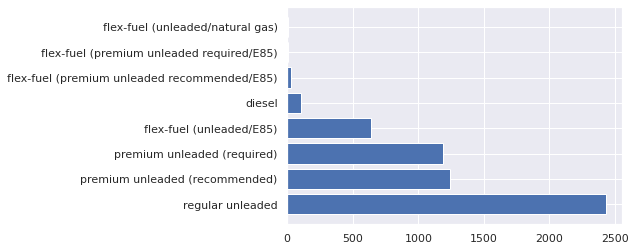
**Inference:** Most people prefer an automatic transmission car over other types



**Inference:** 4dr SUV was the most sold vehicle style.



**Inference:** In the year 2015 there was a spike in car sales



**Inference:** Regular unleaded is the most popular engine fuel type

**Conclusion**:

Hence we learned about exploratory data analysis and various types of statistical measures of data along with correlation. We also learnt about visualization and applied these concepts with hands-on experience on our chosen dataset.