**EXPLORATORY DATA ANALYSIS**

Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore data, and possibly formulate hypotheses that might cause new data collection and experiments.

EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset

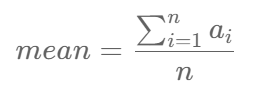
## TYPES OF EDA

1. Univariate Non graphical
2. Multivariate Non graphical
3. Univariate graphical
4. Multivariate graphical

**UNIVARIATE NON*GRAPHICAL:***

*Non Graphical exploratory data analysis is the first step when beginning to analyze your data as part of the general data analysis approach.This preliminary data analysis step focuses on four points, i.e. for mechanisms that you will want to examine. These include:*

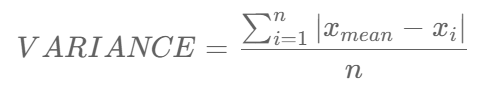
* *measures of central tendency, i.e. the mean, the media and mode,*
* *measures of spread, i.e. variability, variants and standard deviation,*
* *Skewness and kurtosis: Two more useful univariates descriptors are the skewness and kurtosis of the distribution. Skewness is that the measure of asymmetry and kurtosis may be a more subtle measure of peakedness compared to a normal distribution*
* ***more about central tendency...*** *Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. Along with the variability (dispersion) of a dataset, central tendency is a branch of descriptive statistics. The central tendency is one of the most quintessential concepts in statistics.*
* *Mean- it is the average of the numbers i.e. sum of the values divided the total number of values*

**

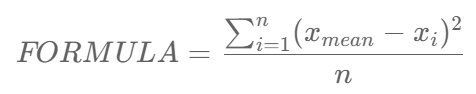
* *Median- it is the middle value of a dataset whrn the data is arranged in ascending order. If the total number of values in the dataset is even then mean of the two middle values is taken.*
* *Mode- it is the most frequently occured value in the dataset. In some cases, dataset may contain multiple modes, while some datasets may not have any mode at all.  
  The selection of a central tendency measures 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.*
* ***more about spread...*** *A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population. It is usually used in conjunction with a measure of central tendency, such as the mean or median, to provide an overall description of a set of data.*
* *RANGE: it is the difference between the maximum and the minimum value of the dataset. $$*

**

* *QUARTILE AND INTERQUARTILE RANGE: Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half. We then find the median for all 4 quartiles and then find the difference of the largest and the smallest quartile value which is known as Interquartile range. semi interquartile range is the half of interquartile range.*
* *The absolute and mean absolute deviation show the amount of deviation (variation) that occurs around the mean score. To find the total variability in our group of data, we simply add up the deviation of each score from the mean. The average deviation of a score can then be calculated by dividing this total by the number of scores.*
* *ABSOLUTE AND MEAN ABSOLUTE DEVIATION: it is the sum of absolute difference of mean and each value then then dividing this sum by total number of values. $$*

**

* *VARIANCE- it is the sum of squares of difference of mean and each value then dividing it total number of values.$$*

**

*If the scores in our group of data are spread out, the variance will be a large number. Conversely, if the scores are spread closely around the mean, the variance will be a smaller number. However, there are two potential problems with the variance. First, because the deviations of scores from the mean are 'squared', this gives more weight to extreme scores. If our data contains outliers (in other words, one or a small number of scores that are particularly far away from the mean and perhaps do not represent well our data as a whole), this can give undo weight to these scores. Secondly, the variance is not in the same units as the scores in our data set: variance is measured in the units squared. This means we cannot place it on our frequency distribution and cannot directly relate its value to the values in our data set.Calculating the standard deviation rather than the variance rectifies this problem.*

* *STANDARD DEVIATION: The standard deviation is a measure of the spread of scores within a set of data. Usually, we are interested in the standard deviation of a population. We are normally interested in knowing the population standard deviation because our population contains all the values we are interested in. Therefore, you would normally calculate the population standard deviation if: (1) you have the entire population or (2) you have a sample of a larger population, but you are only interested in this sample and do not wish to generalize your findings to the population.However, in statistics, we are usually presented with a sample from which we wish to estimate (generalize to) a population, and the standard deviation is no exception to this. Therefore, if all you have is a sample, but you wish to make a statement about the population standard deviation from which the sample is drawn, you need to use the sample standard deviation. Confusion can often arise as to which standard deviation to use due to the name "sample" standard deviation incorrectly being interpreted as meaning the standard deviation of the sample itself and not the estimate of the population standard deviation based on the sample.$$*

**

***MULTIVARIATE NON-GRAPHICAL:***

*These techniques generally show the relationship between two or more variables in the form of either cross-tabulation or statistics. For each combination of categorical variable (usually explanatory) and one quantitative variable (usually outcome), we can create a statistic for a quantitative variables separately for each level of the categorical variable, and then compare the statistics across levels of the categorical variable.*

* *CROSS TABULATION: Cross tabulation is usually performed on categorical data — data that can be divided into mutually exclusive groups.An example of categorical data is the region of sales for a product. Typically, region can be divided into categories such as geographic area (North, South, Northeast, West, etc) or state (Andhra Pradesh, Rajasthan, Bihar, etc). The important thing to remember about categorical data is that a categorical data point cannot belong to more than one category.Cross tabulations are used to examine relationships within data that may not be readily apparent. Cross tabulation is especially useful for studying market research or survey responses. Cross tabulation of categorical data can be done with through tools such as SPSS, SAS, and Microsoft Excel.*
* *benefits of cross tabulation:*
* ***Eliminates confusion while interpreting data:*** *Raw data can be difficult to interpret. Even for small data sets, it is all too easy to derive wrong results by just looking at the data. Cross tabulation offers a simple method of grouping variables, which minimizes the potential for confusion or error by providing clear results.*
* ***Helps in deriving innumerable insights:*** *As we observed in our example, cross tabulation can help us derive great insights from raw data. These insights are not easy to see when the raw data is formatted as a table. Since cross tabulation clearly maps out relations between categorical variables, researchers can gain better and deeper insights — insights that otherwise would have been overlooked or would have taken a lot of time to decode from more complicated forms of statistical analysis.*
* **Offers data points to chart out a course of action:** Cross tabulation makes it easier to interpret data, which is beneficial for researchers who have limited knowledge of statistical analysis. With cross tabulation, people do not need statistical programming to correlate categorical variables. The clarity offered by cross tabulation helps professionals evaluate their current work and chart out future strategies.

**DATA VISUALIZATION:**

It is the process of visualizing data through graphs which makes it easier to understand data patterns.

- **BOXPLOT:** Box plots are used to visualize summary statistics of a dataset, displaying attributes of the distribution like the data’s range and distribution.

- **HISTOGRAM:** it consists of a number of adjacent, equal-width vertical columns, drawn so that there is no space between the columns. The columns correspond to “bins” that together span the range of the data. The data are divided among these bins, with the height of each bar corresponding to the number of data points falling into each bin. The taller the bar, the more data points fall into the range of that bin.Histograms are useful exploratory data visualizations for spotting outliers, skew, bimodality, and other shape features in the distribution as well as for comparing subgroups in the data. The presence of strong skewness or outliers should lead researchers to investigate the use of median and IQR as summary statistics and nonparametric hypothesis tests instead of traditional parametric tests.

- **SCATTER PLOT:** A Scatter Chart (also called a Scatter Plot, Scatter Graph, or Scatter Diagram) is a visualization design that uses Cartesian coordinates to display values in dotsYou can use Scatter Plot examples to determine relationships or associations between key data points.The actual analysis comes in when you discern the type of relationship existing between key metrics you’re tracking closely. You can use Scatter Diagrams to uncover hidden “cause-and-effect” relationships between two key variables in your data.

- **CORRELATIONAL MATRIX:** A correlation matrix is a common tool used to compare the coefficients of correlation between different features (or attributes) in a dataset. It allows us to visualize how much (or how little) correlation exists between different variables. This is an important step in pre-processing machine learning pipelines. Since the correlation matrix allows us to identify variables that have high degrees of correlation, they allow us to reduce the amount of features we may have in a dataset. This is often referred to as dimensionality reduction and can be used to improve the runtime and effectiveness of our models.

# **Univariate graphical :**

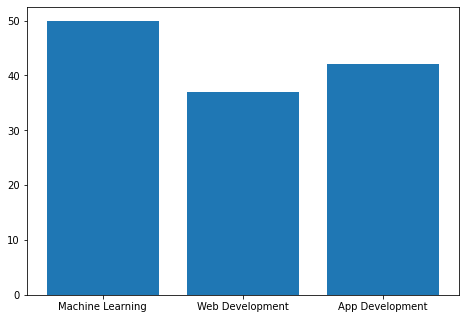
Non-graphical methods don’t provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:

* Stem-and-leaf plots, which show all data values and the shape of the distribution.
* Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
* Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.  
  Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range , variance), Quartiles (interquartile range), and Standard deviation.

Univariate data can be described through:

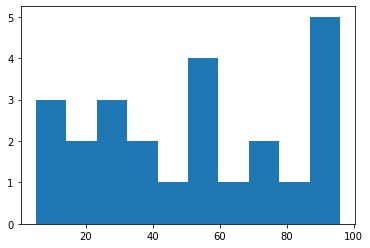
## Bar Charts

The bar graph is very convenient while comparing categories of data or different groups of data. It helps to track changes over time. It is best for visualizing discrete data.



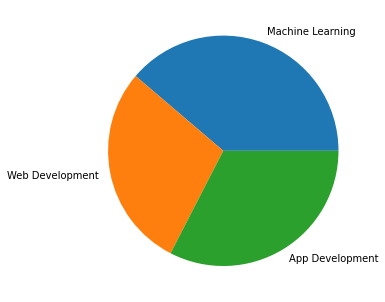
## Histograms

Histograms are similar to bar charts and display the same categorical variables against the category of data. Histograms display these categories as bins which indicate the number of data points in a range. It is best for visualizing continuous data.



## Pie Charts

Pie charts are mainly used to comprehend how a group is broken down into smaller pieces. The whole pie represents 100 percent, and the slices denote the relative size of that particular category.



Similar to histograms, a frequency polygon is used for comparing datasets or displaying the cumulative frequency distribution.

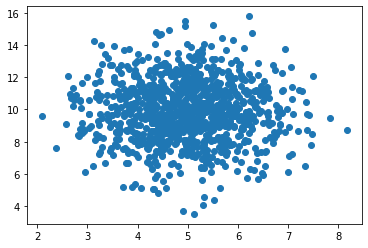
# **Multivariate graphical:**

Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

Other common types of multivariate graphics include:

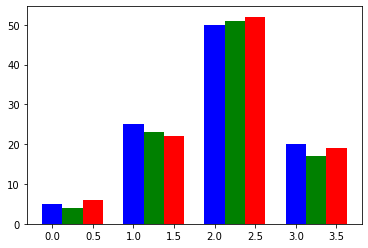
### Scatter plot :

Scatter Plot is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.



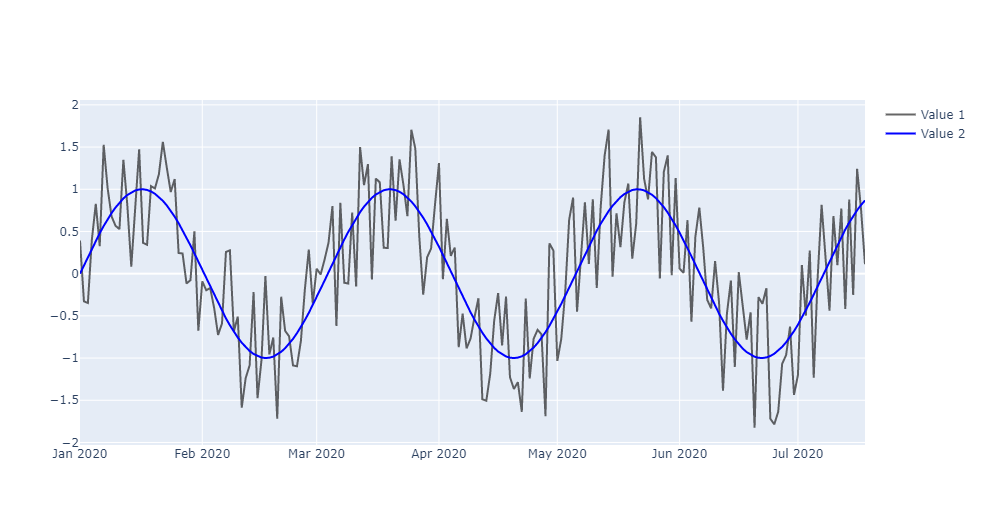
### Multivariate chart :

It is a graphical representation of the relationships between factors and a response.



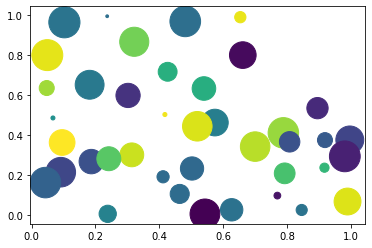
Run chart :

It is a line graph of data plotted over time.



Bubble chart :

Which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.

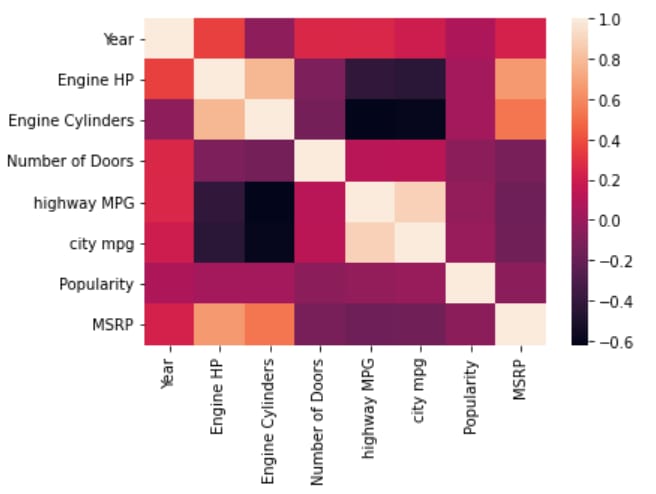


### 

### 

### Heat map:

It’s a graphical representation of data where values are depicted by color.



**Authors:**

1. Vibhansh Bhatia

2. Aditya Yadav

3. Kumar Kanishk Singh

4. Priyanshu Meena

**Sources:**

1.<https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20is,statistical%20summary%20and%20graphical%20representations>.

2. <https://www.geeksforgeeks.org/exploratory-data-analysis-eda-types-and-tools/>

3. <https://statistics.laerd.com/statistical-guides/measures-of-spread-standard-deviation.php>

4. <https://corporatefinanceinstitute.com/resources/knowledge/other/central-tendency/>

5. <https://humansofdata.atlan.com/2016/01/cross-tabulation-how-why/>

# 6. <https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20is,statistical%20summary%20and%20graphical%20representations>.

7. [GeeksforGeeks](https://www.geeksforgeeks.org/exploratory-data-analysis-eda-types-and-tools/)

8. [TowardsDataScience](https://www.geeksforgeeks.org/exploratory-data-analysis-eda-types-and-tools/)

9. [IBM](https://www.ibm.com/in-en/cloud/learn/exploratory-data-analysis)