Unit -1

1.) Data and information:

* Data is a collection of different facts, figures, object, symbols and events.

* Data can be also defined as collection of observations

* Data can be of two types they are Qualitative data and Quantative data.

Qualitative data:

* Qualitative data is a descriptive information.

* for example, the most common names india, etc..

Quantative data:

* Quantative data is a numerical information.

* for example : height, weight, customers in shop, leaves on tree and other measorment.

* Quantitative data can be of two types they are continuous and Discrete.

Information:

* Information is defined as classifed or organized data that has some meaningful value for the User.

* Information is used to make decisions and take actions.

* A Data is said to be informative if it has the following significant.

* Accuracy

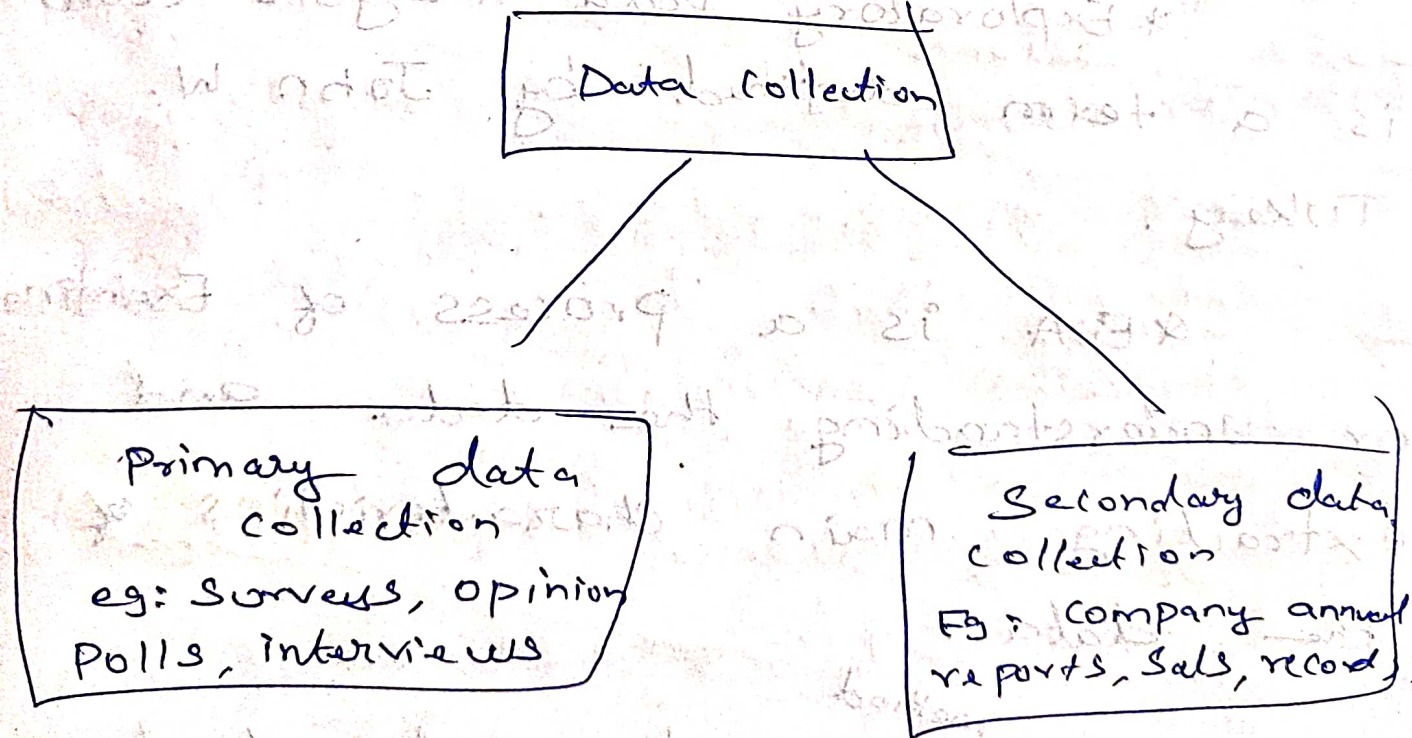
* Completeness

* Timeliness

Data collections:

* Data collection is defined as a method of collecting, analyzing the data for the purpouse of validation and ~~Don~~ research using some techniques.

* Data collection is classified into two types they are primary and secondary.



Problems in data:

* Inconsistent data
* Ambiguous data - errors in data.
* Duplicate data -
* Too much data:
* Inaccurate data,
* Hidden data.

2) EDA [Exploratory Data Analysis]

* Exploratory Data Analysis (EDA) is a term defined by John W. Tukey.

* EDA is a process of Examine or Understanding the data and extracting main characteristics of the data.

* EDA method is classified into two types they are graphical analysis and non-graphical analysis.

* EDA is the first critical step in analyzing the data.

* It is also known as visual analytics or descriptive statics

* It is the prattice of inspecting and exploring the data, before fitting predictors and other inferntial goals.

Statistical graphics:

* A statistical graph or chart is a pictorial representation of statistical data in graphical form.

eg: Histograms, box plots, scatter plots

Non-graphical analysis:

It uses statistical techniques to explore a single variable

Reasons to use EDA:

1) Detection of mistakes
2) Examine the data distribution
3) Handling Missing values.
4) Handling outliers
5) Removing duplicating
6) Normalizing and scaling
7) Determining relationships

3) Data Science:

* Data science is the domain of study that deals with data in large volume.

* Data science is combination of math and statistics.

life cycle of data science.

Data Science's are classified into five stages:

1) Capture - data Collection, Data acquistion, Data extraction

2) maintain :- Data warehousing, data cleansing, data processing.

3) process - Data mining, Clustering modeling.

4) Analyze - Predictive Analysis, regression, text mining.

5) Communicate. - Data reporting, Data visualization, Business intelligence, Decision making

Data Science tools:

1) Data Analysis:

SAS, Jupter, MATLAB, Excel.

2) Data Warehousing:

Informatica, AWS Redshift

3) Data Visualization:

Jupyter, Tableau, Cognos Raw, PowerBI

4) Machine Learning:

Spark, MKLib, Azure ML studio, mahout.

Application of Data Science:

* Healthcare
* Gaming - to next level.
* Image recognition
* Recommendation
* Logistics
* Fraud detection
* Internet search
* Speech recogonition
* Finance.

3) Types of EDA:

* There are four primary types of EDA, Namely

* Univariate non-graphical
* Univariate graphical
* Multivariate non-graphical
* Multivariate graphical.

* Univariate non-graphical:

* This is the simplest form of data Analysis,
* The data consists of one variable. Since it is an single variable.
* It does not deals with relationships.
* The main purpose of Univariate analysis is to describe the data and finding patterns.

Univariate graphical:

* Univariate non-graphical methods do not provide a full picture of the data.

* Hence graphical methods are required. Shape of distribution, frequency, mean, median, mode etc...

* Steam-leaf plots, which shows all data values, and the shap of distribution.

* Histograms, Bar plots: represents the frequency.

* Box Plots: mean, median, mode.

* Multivariate non graphical:

* Multivariate data arises from more than one variable

* Multivariate non-graphical EDA techniques generally shows

the more relationship between two or variables.

* Multivariate non-graphical represents data through cross-tabulation or statics.

* Multivariate graphical:

* Multivariate data uses graphics to display the relationships between two or more sets of the data.

* The most used graphic is bar plot or bar plot.

* other common types of multivariate graphics include

* Scatter plot
* Multivariate chart
* Run chart
* Bubble chart
* Heat map.

4) Sense of Data:

* In today's digital era data is very powerful tool.

* Just having possession of data is not sufficient.

One should be able to analyze the data in proper manner to get insight of data.

* making sense of the data means finding meanings from data generating inference from data.

* Data analaysis the process of making data meaning.

* Consider the process of creating something out of a set of raw materials.

for example: given a bolt of fabric and some notions.

Different pepople make different kinds of creation. one person might make a baby dress, another one make dance coustume and a another one make diwali lamp. But what they come up with is raw materials. It would be awfully difficult to make a working of car using bolt of fabric.

*Basic steps to make data sense:

1) managing data.
2) Getting familar with data.
3) Making sense of data.

*Managing data:

* It is an Ential process of analyzing qualitative data.

* First task is arrange or organize data. So it will make sense of it

* Separating different types of data. This will help to access the data easily.

* Keep the data in chronological order. makes data access fast.

* organizing data on topic wise or document wise.

Getting familar with data:

* Based on data collection and its size, it might take time to understand about data. Getting familar with data plays a vital role in output or result prediction.

Making Sense of data:

* Arrange the data in certain presentable or viewable format.

* Data is best seen using tables, charts, graphs and patterns.

6) Software for EDA:

1. R -

* It is open Source programming language.
* It is graphics supported
* R is foundation for Statistical computing.
* The R language is widely used for Statistical observations and data Analysis

2. Python:

* An interpreted, object oriented programming language.
* It is high level and has built-in data structures.
* It is rapidly used for application development and as well as for scripting.

3. Excel /spreadsheet:

* Excel tool is considered to indispensable part of analytics industry.

* It supports all the features like summarizing, visualizing data, data wrangling etc.

* It is managed by microsoft

* Microsoft excel is paid.

4. Weka:

* weka is an machine Learning tool,

* Easy to learn

* having an intuitive interface.

* It provides option for data-preprocessing, classification, clustering, association rule, and visualization.

* It is built on Java.

5. Tableau :

* Tableau is data visualizati Software.

* It is a fast visualizatio which allows exploring of data.

7). Visual Aids for EDA:

Univariate plots:

* It shows the frequency or the distribution shape of variable.

i) Histograms:

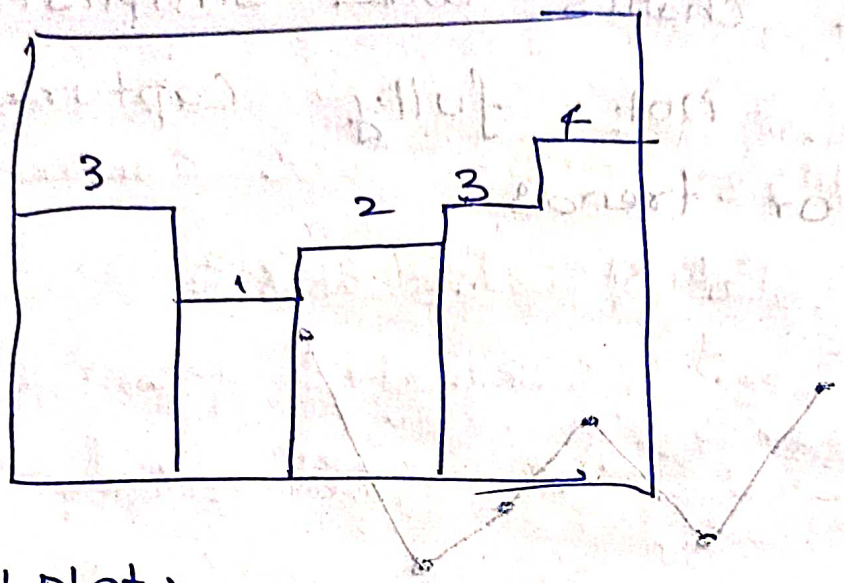
* Histograms are two-dimensional plots which contains x-axis and y-axis.

* x-axis shows time intervals

* y-axis show the frequency values.

* It is similar to Bar graph, but histogram does not contains gaps between bars.

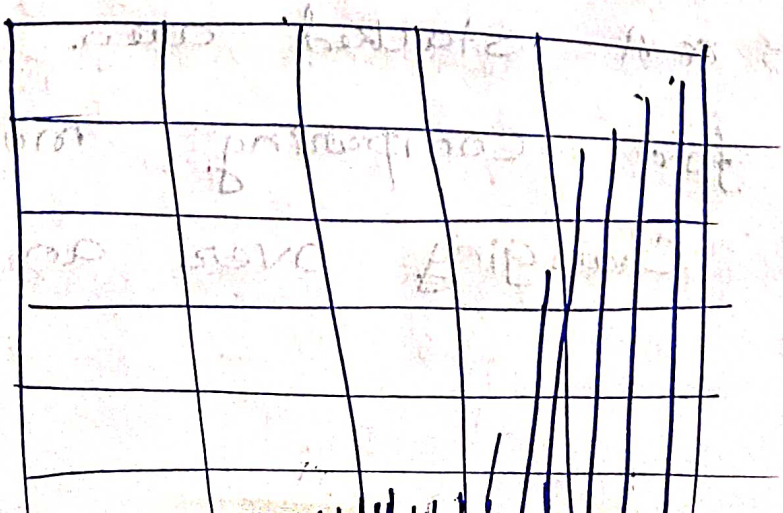
eg:



ii) Distplot:

* Dist plot are also known as second histogram because it is a slight improved version of the histogram.

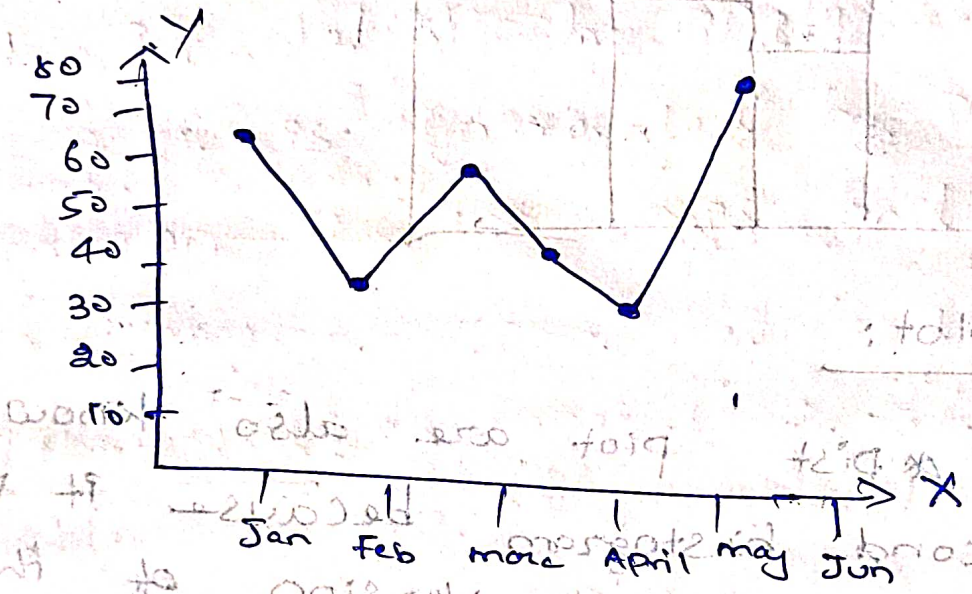
* Distplot gives us as KDE (Kernel Density Estimation).



iii) Line Chart:

* A line chart is a graphical representation of an asset's historical price.

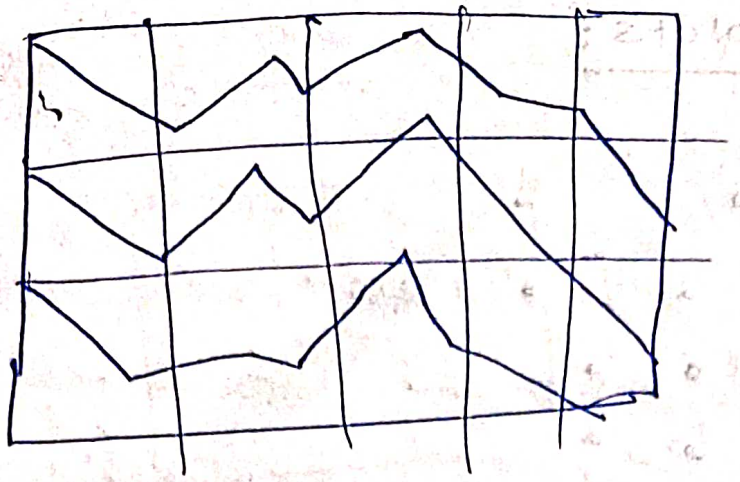
* Line charts are simplistic and may not fully capture patterns or trends.



iv) Stacked area plot/chart:

* A Chart combines the line chart and bar chart.

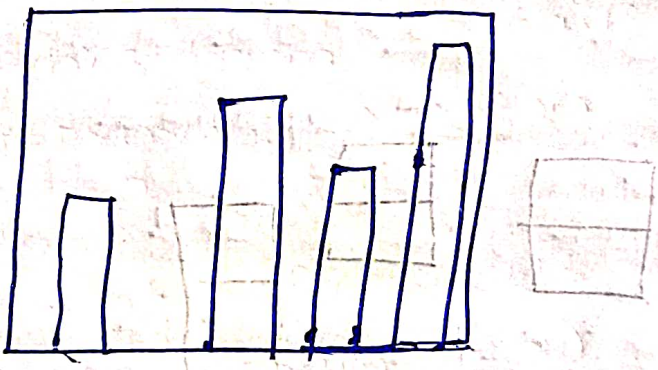
* A stacked area graph is useful for comparing multiple variables changing over an interval.



2. Bivariate plots:

Bivariate plots display the relationship between two variables on exploratory data analysis.

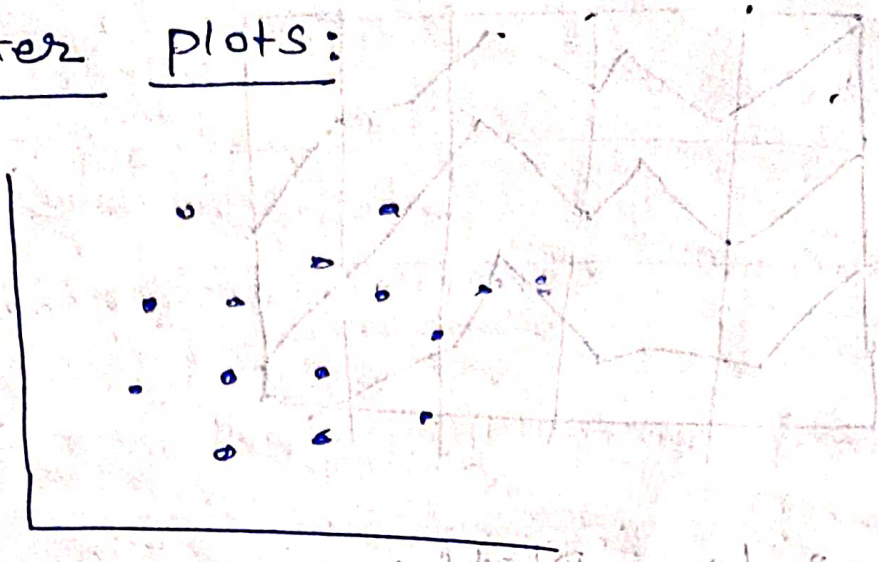
i) Bar graphs:



* Bar charts can be used to compare nominal or ordinal data.

* They are helpful for recognizing trends.

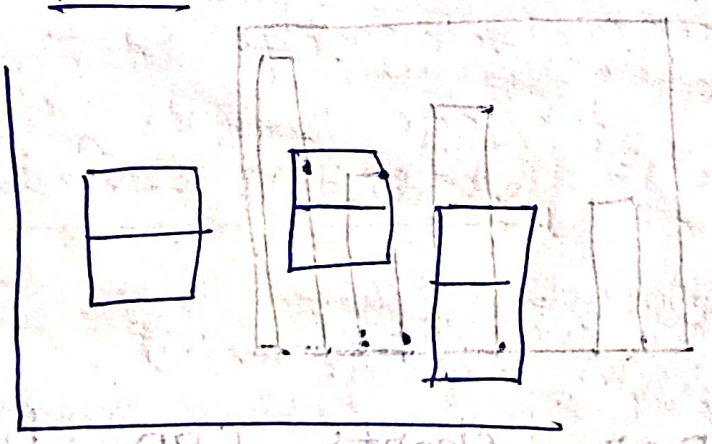
ii) Scatter plots:



* Scatter plots are commonly used in statistical analysis.

* It is correlated by plotting them on the x axis and y-axis.

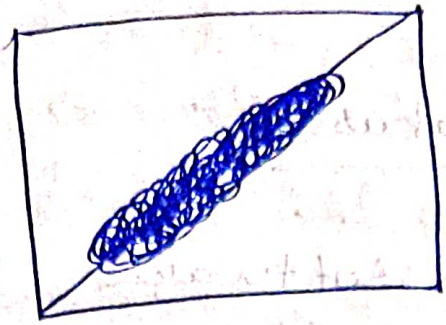
iii) Box plots:



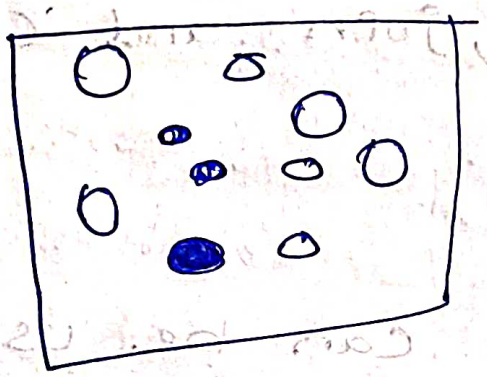
* Box plots are graphically boxical representation of two relationships.

* Box plots are suitable for identifying outliers.

iv) Correlation plots:



v) Cluster Map:



* It is also known as a dendrogram.

* It is also known as a Heat map.

3) Multivariate plots:

* It deals with more than 3 variables.

* It is also known as three dimensional plots.

* It is used in very large dimensional.

Text data:

Image data.

8) Merging Database:

*pandas Software in Python programming language for manipulation and analysis

*merge, join and concatenate.

Pandas merge:

*merge() can be used to a database's join operation. It is the most flexible operation that can be applied to data.

*It is used to combine data using one or more keys.

*Both many-to-one and many-to-many joins with merge().

*Merge() function is used for merging datasets.