UNIT-I

Descriptive statistics and methods for Data science

Data science :-

Data science is an inter-disciplinary field

→ which uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data

→ Apply knowledge and actionable insights from data to a broad range of application domains.

Data science plays a key role in data mining, machine learning and big data and many more.

→ All applications work by creating and operating a digital model of the real-like scenario that they are trying to automate. like speech recognition, Healthcare, sports, Gaming, Banking & Finance, Self-Driving cars, Security, Internet Search, Robots, Digital Advertisements, Fraud & risk-Detection, Delivery logistics, Image Processing, Internet of Things (IOT). etc. ..

Data science methodology:-

Data science applications require a particular methodology & skills.

1) Exploratory Data Analysis (EDA):

It is critical for a data science practitioner to understand the domain and how the data represents the domain. He needs to understand the basic properties of data (mean, variance, range, distribution etc.)

correlation between the variable that he is trying to predict and the input variables, as well as the relationship b/w the input variables themselves.

So, The datascience practitioner needs to be aware of techniques to programatically explore the data and draw insights that are meaningful for modelling.

2) Data visualization:-

To deal with larger volumes of data that analyzable across several dimensions. Data visualization techniques are vital to be able to abstract data and detect patterns.

3) Data manipulation:-

Data science often needs to merge - multiple datasets to create a common dataset.

The student needs to ~~merge~~ be able to summarize data at various levels & in general be very well versed with data merging, splitting and computing relevant measures.

4) Feature selection & Extraction:

We should use the insights gained from the EDA and visualization exercise to determine the correct features that have Predictive power.

5) Model Development:-

To determine the correct machine learning algorithm that can be used for a scenario. The student needs to familiar with common algorithms and the merits & demerits of those algorithms.

6) model Performance management:

once a model is built, its predictions needs to be checked for accuracy. and also be able to boost performance by using additional features (or) other techniques.

Importance of Data science:-

1. Data science helps brands to understand their customers in a much enhanced and empowered manner.

2. It allows brands to communicate their story in such a engaging and Powerful manner.

3. Big data is a new field that is constantly growing and evolving.

4. Its findings and results can be applied to almost any sector like travel, healthcare and education among others.

5. Data science is accessible to almost all sectors.

Statistics:- The word statistics is derived from italian word "Stato". (Political State) It is the discipline that concerns the collection, organization, analysis, interpretation and Presentation of data.

It deals with every aspect of data, including the planning of data collection interms of the design of surveys and experiments.

There are two types of statistics:

Descriptive statistics:

organizing & summarizing data using informal methods like graphing and using numbers is called descriptive stat.

Inferential statistics:-

organizing & summarizing data using formal methods is called inferential statistics.

Role of Statistics:-

→ Statistics is a language of data

→ Statistics provides a scientific way to extract and retrieve the information hidden inside the data.

→ Statistics cannot do miracles.

→ " " change the Process (or) Phenomenon.

→ " " do anything in no time. Tools can be developed but development needs time.

Statistical Tools:-

several tools & components are available

→ Graphical tools: Provide visualization from first hand information

→ Analytical tools:- Provide quantitative information

both approaches work together and are inseperable.

Graphical Tools

1. 2D & 3D Plots
2. Scatter Diagram
3. Pie Diagram
4. Histogram
5. Barchart
6. Stem & leaf Plot
7. Box Plot.

Analytical Tools

Central tendency of data: mean median, mode, G.m, H.m, Quartiles etc.

Dispersion of data: variance, S.D, standard error, meandeviation, absolute deviation, range etc.

Collection of Data:-

The basic Problem of Statistical enquiry is, to collect facts and figures relating to a Particular Phenomenon under study. collection of data is the Process of enumeration together with the Proper recording of results. The success of an enquiry depends on the Proper collection of data.

Statistical data may be classified as two types They are i) Primary data ii) Secondary data.

Primary Data:

Data originally collected by an investigator for the first time for any statistical investigation and it is the type of 'data' that is collected by researchers directly from main 'sources'

Sources of Primary Data:-

- → Direct Personal investigation
- → Questionnaire received through mail, e-mail, e-forms (google forms), online surveys etc.
- → Questionnaire sent through surveyors.

Secondary data:

Data which has already been collected by some Person (or) Agency for any statistical investigation.

Sources of secondary Data:-

- → Published sources: Ex: Books, Journals, articles, web Pages etc.
- → Data collected from survey agencies.

→ Public reports the data Ex: municipalities

→ blogs

→ Library, & Internet sources etc.

Examples:-

An example of Primary data is the national census data collected by the government while an example of secondary data is the data collected from online sources.

The secondary data collected from an online source could be the primary data collected by another researcher.

The Government, after succesfully completion of the national census, they share the results in newspapers, online magazines, Press releases etc. This is Primary data.

Another government agency that is trying to allocate the state budget for healthcare, education etc, may need to access the census results. Then the census data is secondary data for this government agency.

Advantages:-

Some common advantages of Primary data are its authenticity, Specific nature and up to date information while secondary data is very cheap and not time-consuming.

→ Primary data is very reliable because it is usually objective and collected directly from the original source. It also gives upto date information about a research topic compared to secondary data.

→ secondary data is not expensive making it easy for people to conduct secondary research. It does not take so much time and most of the secondary data sources can be accessed for free.

Similarities b/w Primary & secondary data:

1. contains same content:

Secondary data was once Primary data when it was newly collected by the first researcher. The content of the data collected does not change and therefore has the same content as Primary data.

It does not matter if it was further visualized in the secondary form, the content does not change. Ex:- definitions, Theorems, Postulates that were made years ago but still remain the same.

Primary data and secondary data both have applications in business and research. They may, however, differ from each other in the way in which they are collected, used and analyzed.

Differences between Primary & Secondary data:

→ Primary data is very expensive while secondary is economical. When working on a low budget, it is better for researchers to work with secondary data, then analyze it to uncover new trends.

→ Primary data is more accurate and reliable while secondary data is relatively less reliable and accurate. This is mainly because the secondary data sources are not regulated and are subjected to personal bias.

→ Primary data is available in crude form while secondary data is available in a refined form.

→ The secondary data is usually made available to the public in a simple form for a layman to understand while primary data are usually raw and will have to be simplified by the researcher.

Variables and Types of Variables:-

The values that are altering according to circumstances are referred to as variables. A variable can occurs in any form. such as trait, factor (or) a statement that will constantly be changing according to the changes in the applied environment.

Variables in statistics are broadly divided into four categories such as

i) Independent variables
ii) Dependent variables
iii) Categorical variables
iv) Continuous variables

Apart from, Quantitative and Qualitative variables hold data as nomial, ordinal, interval and ratio.

Independent variables:-

The independent variable is the one that is computed to view the impact of dependent variables. It is also called as resultant variables, predictor (or) experimental variables.

Ex:- ① A manager asks 100 employees to complete a project. He should know the capacity of the individual employee. He wants to know the reason behind smart guys & failure guys.

The 1st reason is that some will be working hard for day & night to complete the project with in the estimated time, and the other one is that some guys are born intelligent and smarter than others.

Their hardwork and IQ is independent variable.

Ex:2) A Tutor asks 100 students to complete a maths Test. The tutor wants to know why some students perform better than others. He thinks that

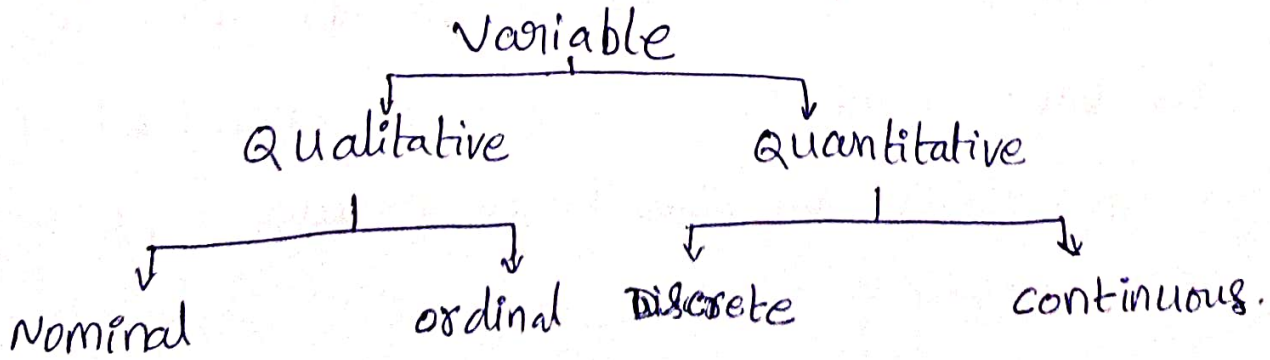
i) some students spend more time revising for their Test & ii) some students are naturally more intelligent than others.

Their Revising time & intelligence is independent-variables.

Dependent variables:-

The dependent variable is also called a criterion variable which is applied in non-experimental circumstances. The dependent variable has relied on the independent variable.

In the above examples, completion of Project is the dependent variables. & Getting better marks in the test is dependent variables.



Quantitative variables:-

Variable that reflects a notion of magnitude. i.e, If the values it can take are numbers. A Quantitative variable represents thus a measure & is numerical.

→ Quantitative variables are divided into two types: Discrete and continuous.

Discrete variables:-

Quantitative discrete variables are ~~visible~~ variables for which the values it can take are countable and have a finite number of possibilities.

Ex:- No. of children per family.

No. of students in a class.

Continuous variables:-

The variables that can take any value on an interval are called continuous variables.

Qualitative (or) Categorical variables:-

These are ~~also called as~~ not numerical and which values fits into categories.

Qualitative variable is a variable which takes as its values modalities, categories (or) even levels, in contrast to quantitative variables which measure a quantity on each individual.

These are divided into two types. They are

i) Nominal and ii) Ordinal.

Nominal variables:-

A Qualitative nominal variable is a Qualitative variable where no ordering is possible (or) implied in the levels.

Ex:- 1) The variable gender is nominal because there is no order in the levels male/female.

2) Eye colour is nominal because there is no order among blue, brown (or) green eyes.

This variables can have between two levels

Ex:- 1) Do you smoke? Yes/no

2) What is your gender? male/female

Ordinal variables:-

An ordinal variable is a qualitative variable with an order implied in the levels.

Ex:- ① The road accidents has been measured on a scale such as light, moderate and fatal accidents,

② Another example is health, which can take values such as poor, reasonable, good (or) excellent.

Data Visualization:-

Data visualization is the graphical representation of information and data.

The main goal of data visualization is to make it easier to identify patterns, trends and outliners in large data sets.

→ By using visual elements like charts, graphs and maps

→ Data visualization tools provide an accessible way to see and understand trends, outliers, & patterns in data.

→ In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Benefits of Data Visualization:-

→ The ability to absorb information quickly, improve insights and make faster decisions.

→ An increased understanding of the next steps that must be taken to improve the organization.

→ An improved ability to maintain the audience's interest with information they can understand.

→ An easy distribution of information that increases the opportunity to share insights with everyone involved.

→ Eliminate the need for data scientists since data is more accessible and understandable.

→ An increased ability to act on findings quickly and therefore achieve success with greater speed and less mistakes.

General types of data Visualization:-

Charts, Tables, Graphs, maps, info graphics, Dash boards

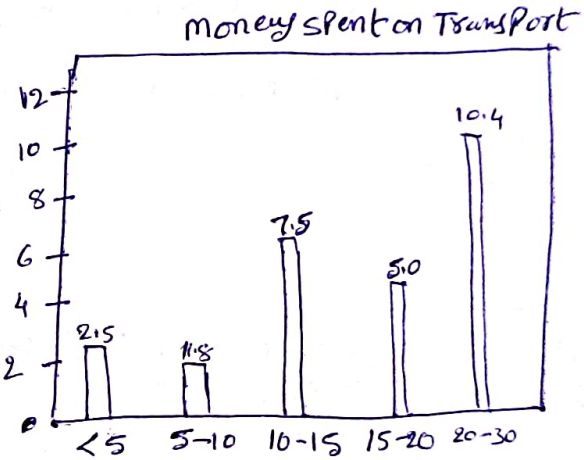
Ex:- Bar chart, Pie chart, Histogram, Scatter diagram etc..

Bar Graphs:-

→ Visualizes the relative (or) absolute frequencies of observed values of a variable.

→ consists of one bar for each category either horizontal (or) vertical.

→ Height of each bar is decided by the frequency of the respective category shown y-axis.

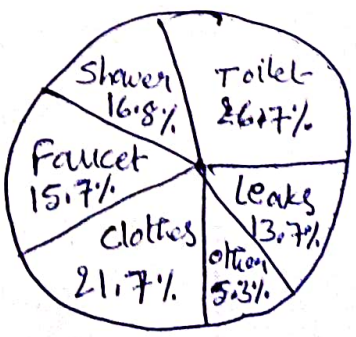


Pie diagrams:-

→ visualizes the absolute and relative frequency.

→ A circle Partitioned into segments where each of the segments represents a category.

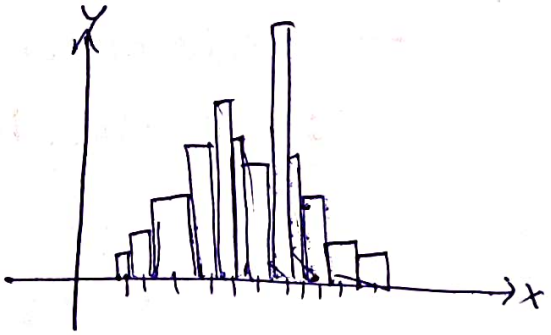
→ Size of each segment depends upon relative frequency and is determined by the angle. (ie relative frequency x 360°)



Histogram:-

It visualizes the distribution of data over a continuous interval (or) certain time period. Each bar in a histogram represents the tabulated frequency at each interval.

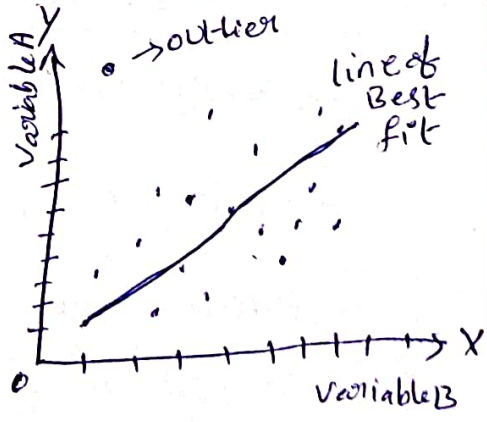
Histograms help give an estimate as to where values are concentrated, what the extremes are and whether there are any gaps (or) unusual values. They also useful for giving a rough view of the probability distribution.



Scatter diagram:-

It is also known as Scatter Graph, Dant Graph, X-Y Plot, Scatter Chart,

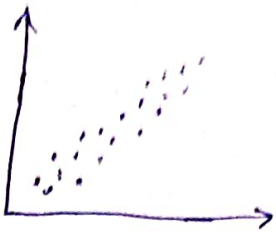
Scatter Plots use a collection of points placed using cartesian coordinates to display values from two variables. By displaying a variable in each axis, you can detect its relationship (or) correlation b/w the two variables exists.



Points that end up far outside the general cluster of points are known as outliers.

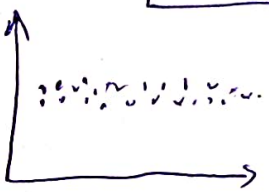
Various types of correlation can be interpreted through the patterns displayed on scatter plots.

They are
- Positive (values increase together),



- Negative (one variable increases as the other decreases)

- Null (No correlation)



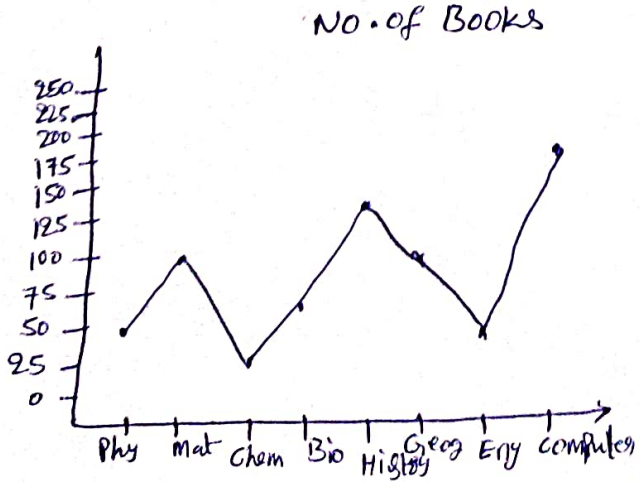
- Linear

- Exponential

- U-shaped

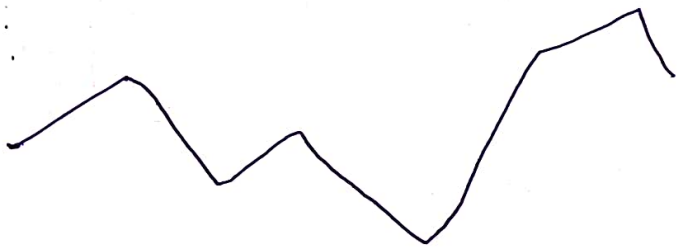
The strength of the correlation can be determined by how closely packed the points are to each other on the graph.

Line Graph:- It is a type of chart which displays information as a series of data points called markers connected by straight line segment.



Sparkline :-

A sparkline is a very small line chart drawn without axes or coordinates, It present the general shape of the variation in some measurement such as temperature (or) stock market price in a simple and highly condensed way.



Sparkline chart.

Population Vs Sample :-

→ It is a complete set of all possible observations of the type which is to be investigated.

i) Finite & Infinite Population ii) Hypothetical & Existant Population

→ Sample is a subset of population. and There is a procedure to draw a sample from population such a process is called Sampling.

i) Random sampling ii) Purposive sampling

iii) Simple sampling iv) Stratified sampling.

A study of the sample will give correct idea of the population.