

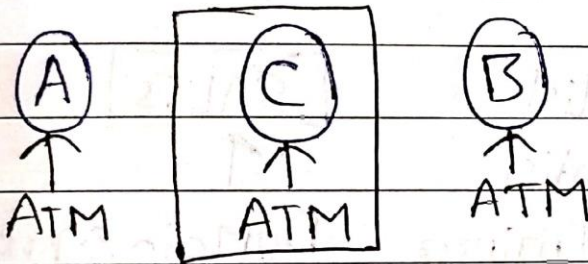
:- Hare Krishna:-

## Statistics:-

Use case:-

HDFC

30 KMS



Statistician

↳ 5 years

↳ Data Analyst

↳ Business Analyst

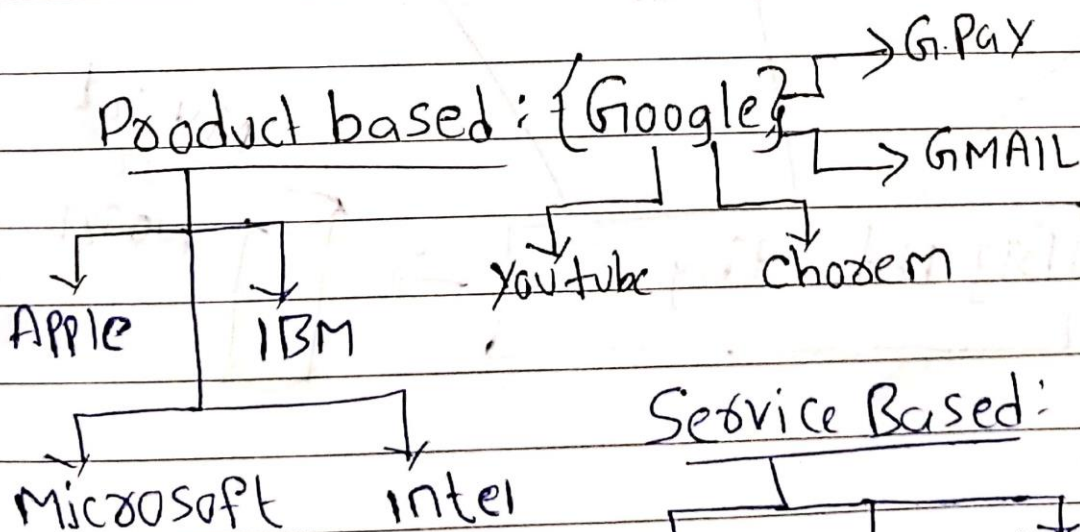
↳ Data Scientist

↳ Product MANAGERS

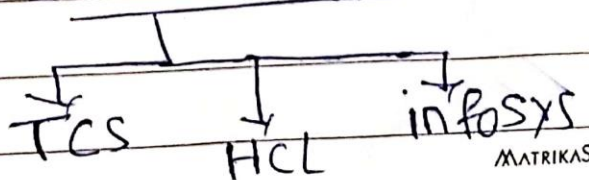
① DATA ANALYST

② DATA SCIENTIST

Product Based company  
&  
Service Based company



Service Based:





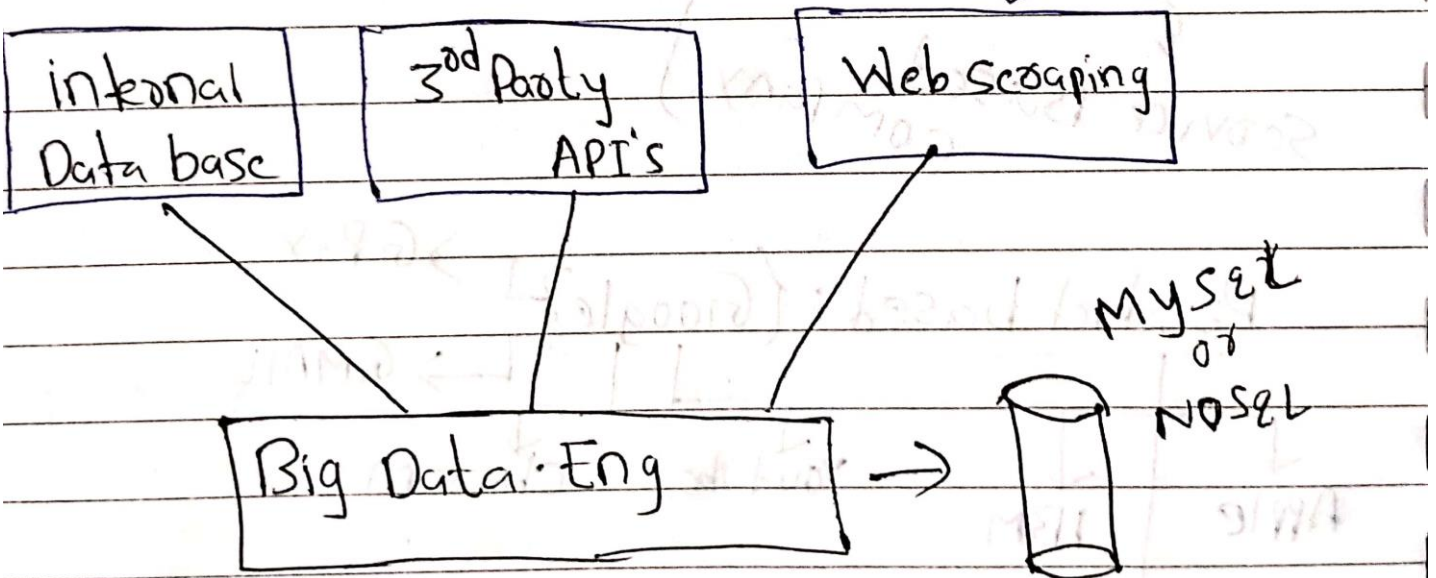
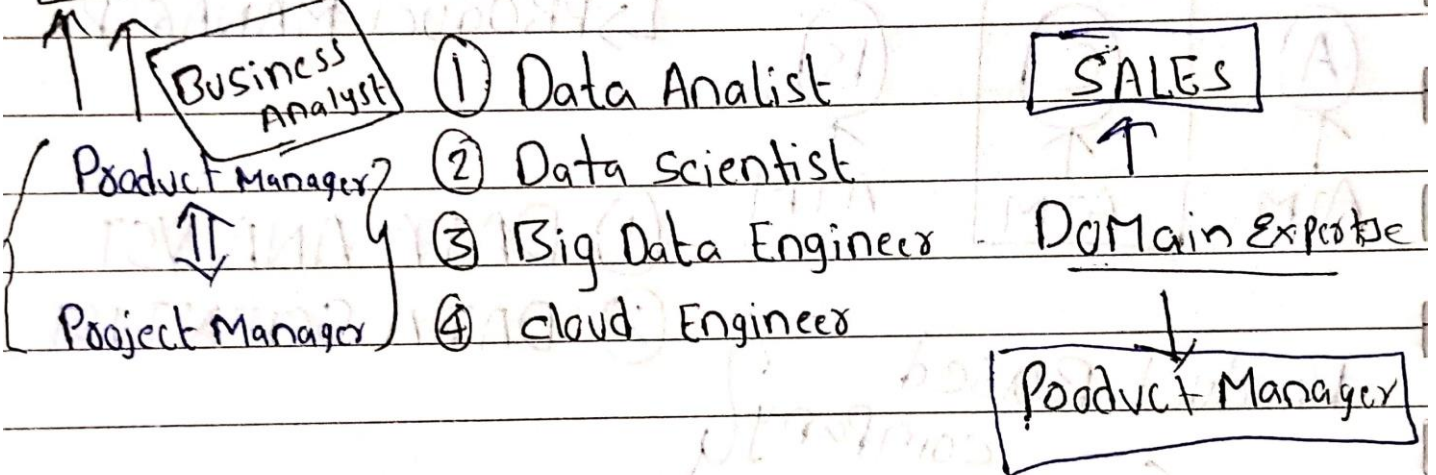
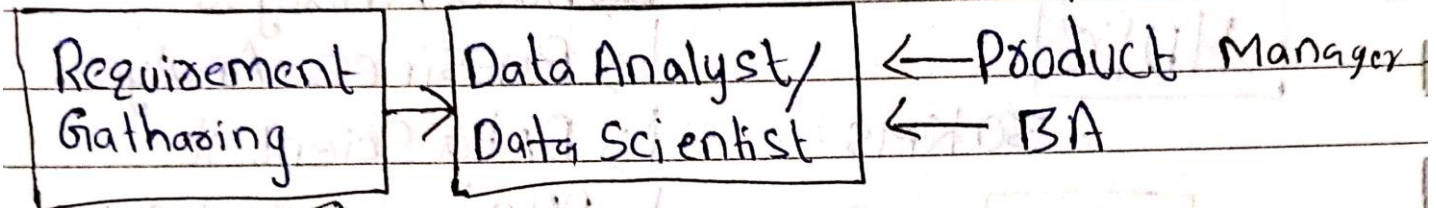
Date: / /

# Statistics:-

{ life cycle of data science projects }

DATA SCIENCE

Domain Knowledge

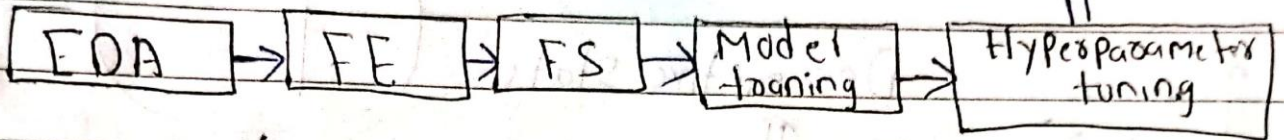




Date / /

## Life cycle of DS Projects.

improve the program model



EDA :- Exploratory Data Analysis

FE :- Feature Engineering

FS :- Feature Selection

Statistics are used in

→ EDA

→ FA

→ FS

Training with ML Algorithms

↔ Model training

→ Hyperparameter tuning.

improve the Project model

Product base Companies:- focuses on making the highest quality product for its customers

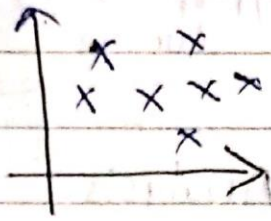
Ex:- Microsoft, Google, Cisco, Adobe, Apple...etc

Service based companies:- Provide required service to the another company products

Ex:-

TCS, Infosys, IBM, Wipro, Accenture  
.. etc.

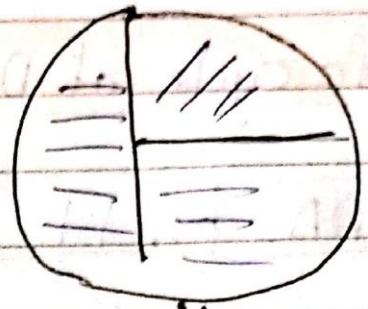


Analysis of Data:

Descriptive Stats



Summarizing the data



Descriptive Stats

Age = {12, 13, 14, 18, 20, 25}  $\Rightarrow$  Average Age

descriptive stats

Measures the central  
Tendency

Definition of Statistics:-

Statistics is the science of collecting, organising and analysing the data.

Data :- "Parts or Pieces of information"

eg: Ages of students in classroom

{24, 25, 32, 29, 28}  $\Rightarrow$  Mean, Median,

Standard

deviation

Mode



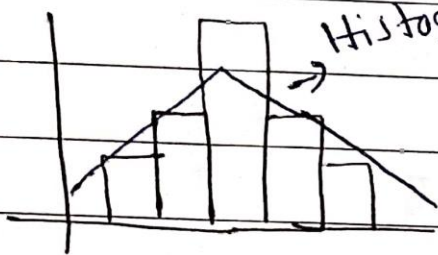
Date: / /

# Statistics

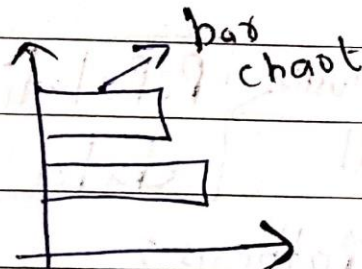
## Descriptive stats

[EDA + FE]

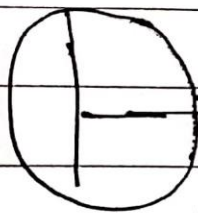
① it consists of organising and summarizing the data



Histogram



bar chart



pie



candlestick



Box plot.

## inferential stats

① it consists of collecting sample data and making conclusion about population data using some experiment

Hypothesis testing

university  $\rightarrow$  500 people

class A  $\rightarrow$  60 people

Sample data  $\Rightarrow$  Age

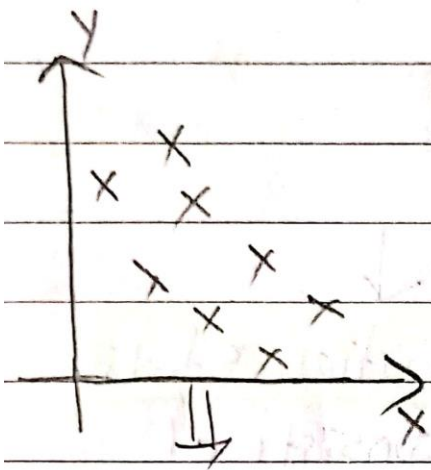
Average age of entire university

Hypothesis testing

C.I  $\Rightarrow$

Confidence interval

Date: / /



(height) (weight)

x

y

-

-

-

-

-

-

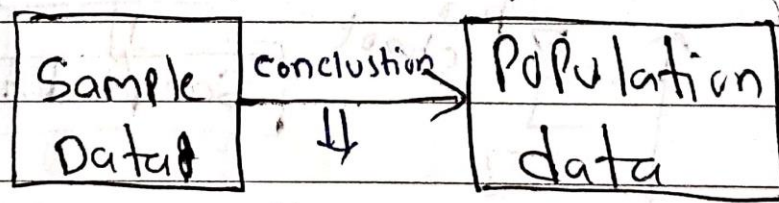
x ↑	y ↑
x ↓	y ↓

Scatter plot

P-Value

x ↑	y ↓
x ↓	y ↑

- ① Z-test
- ② t-test
- ③ Chi square test
- ④ F-test



Hypothesis testing





## Sample data & Population data:

Sample  
Size = 100



Punjab

1008

Population data

Exit Pole:

Party A will win

Party B will loose

Eg: let say that there are 20 classrooms in a university and you have collected the age of students in one classroom.

Ages = { 21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22 }

Weight = { - - - - - }

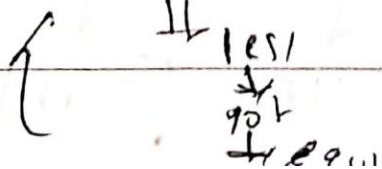
## Descriptive stats:

what is the average age of students in the classroom?

Relationship b/w Age & gender?

## Inferential stats:

Are the average age of the students in the classroom less than the average age of the student in the



University

MATRIKAS



1000 students  $\Rightarrow$  hypothesis testing

Date: / /

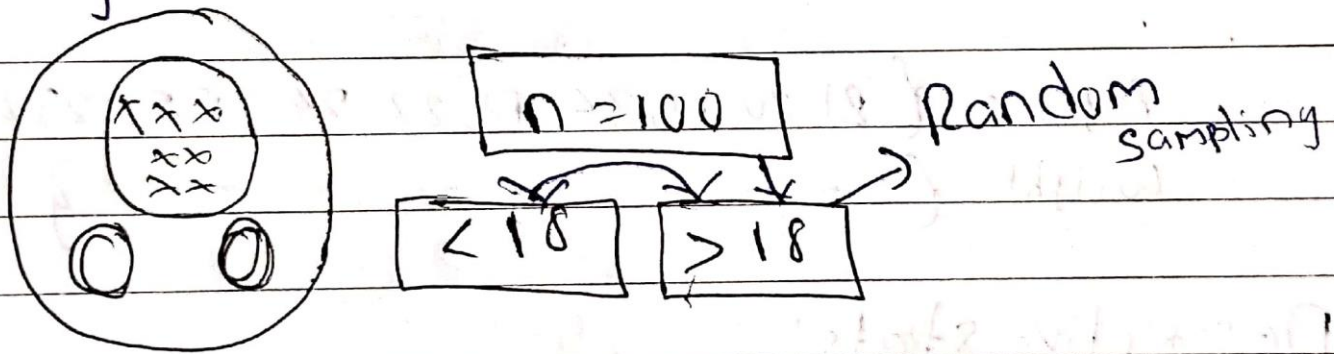
class A  $\Rightarrow$  50 girls 50 boys  
 $\Downarrow$   $\Downarrow$   
 95% 92%

choose a sample  $\Rightarrow$

Sampling Techniques: Population (N)  
 Sample (n)

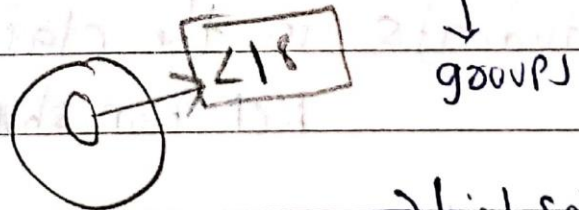
## ① Simple Random Sampling:

Every member of the Population (N) has an equal chance of being selected for your sample (n)



Strata  $\rightarrow$  layers  $\rightarrow$  clusters

## ② Stratified Sampling:



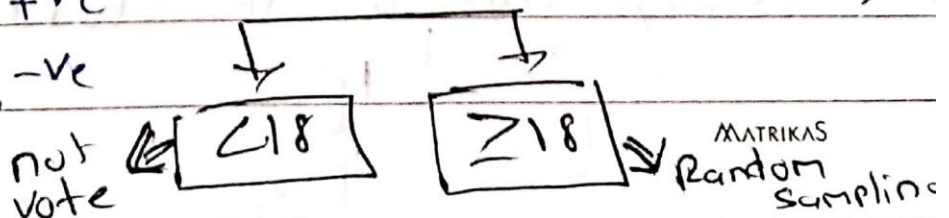
Gender.  $\rightarrow$  male  
 $\rightarrow$  female

Education  $\rightarrow$  high school  
 Degree  $\rightarrow$  master  
 $\rightarrow$  phd

blood group  $\rightarrow$  O +ve  
 $\rightarrow$  B +ve  
 $\rightarrow$  B -ve

Population

{exit policy}





Date: / /

### ③ Systematic Sampling:

select any  $n^{\text{th}}$  person

out of Population (N)

5<sup>th</sup> Person

→ { Air Pool }

⇒ selling

{ Credit Card }

select every  $n^{\text{th}}$  individual

out of Population

card

### \* Convenience

#### Sampling:

only those who are interested in the survey will only

{ Data Science Survey → General AI Survey } Participate

### ① Survey Regarding New Technology:

⇒ Convenience

### ② RBT Survey ⇒ Women

⇒ married Sampling

Stratified

+

Random Sampling

### ③ Credit Card:

Stratified +

Random Sampling

MATRIKAS



Variable: A variable is a property that can take any values

Eg: age = 14

Variable

age = 25

Age = [24, 25, 26, 27, 28]

age = 100

collection

Two different type of Variable

① Quantitative Variable

Measured numerically (mathematical operation)

Ex: age, weight, height,

dist, rainfall, temp,

② Qualitative Variable

Categorical Variables

(based on some characteristics they are grouped together)

Eg: Gender, types of flowers,

type of movies

Quantitative Variable

Discrete Variable

Continuous Variable



Date: / /

## Discrete Variable:

eg: Whole number → fixed

eg:- N.O of Bank accounts

{1, 2, 3, 4} 2.5 ✗

eg:- N.O of children:- whole n.O

## Continuous Variable:

eg:- continuous → decimal values

eg:- height, weight, age, speed

qualitative  $\left\{ \begin{array}{l} \rightarrow \text{named} \\ \rightarrow \text{not named} \end{array} \right\}$

Categorical variable

## Quantitative Variable

Discrete Variable

Continuous Variable

Whole number

Bank account = {1, 2, 3, 4}

pin code = { - - - - }