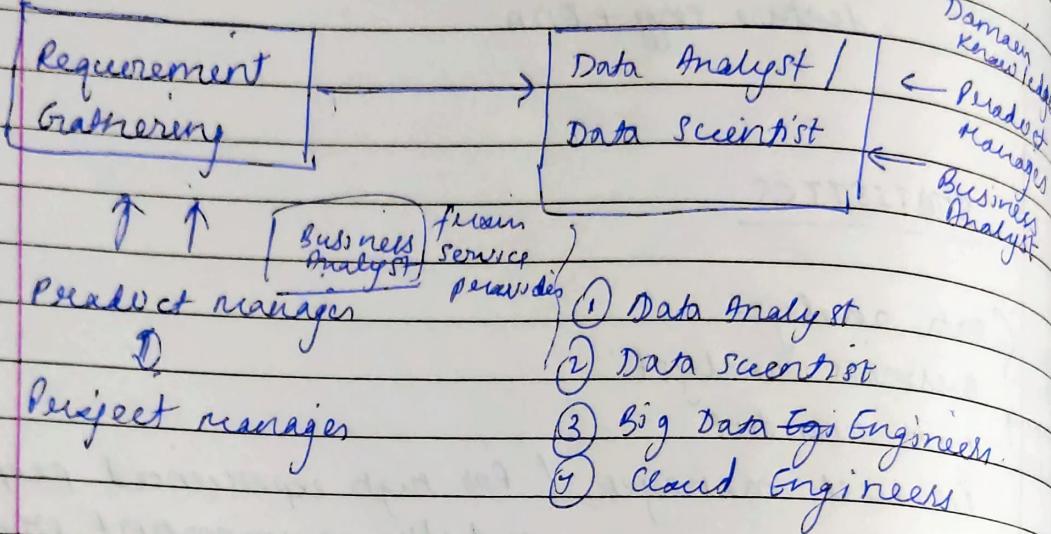
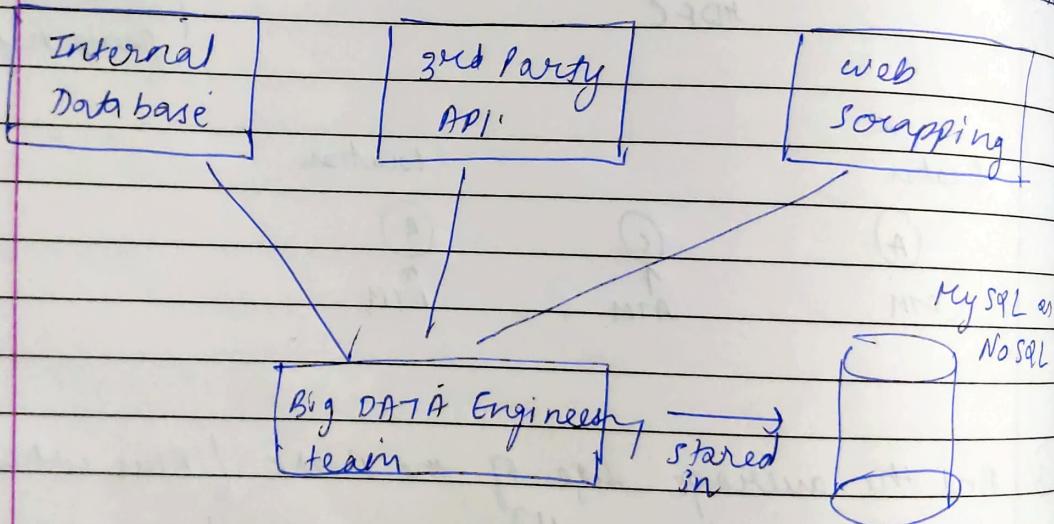


Life cycle of data science project

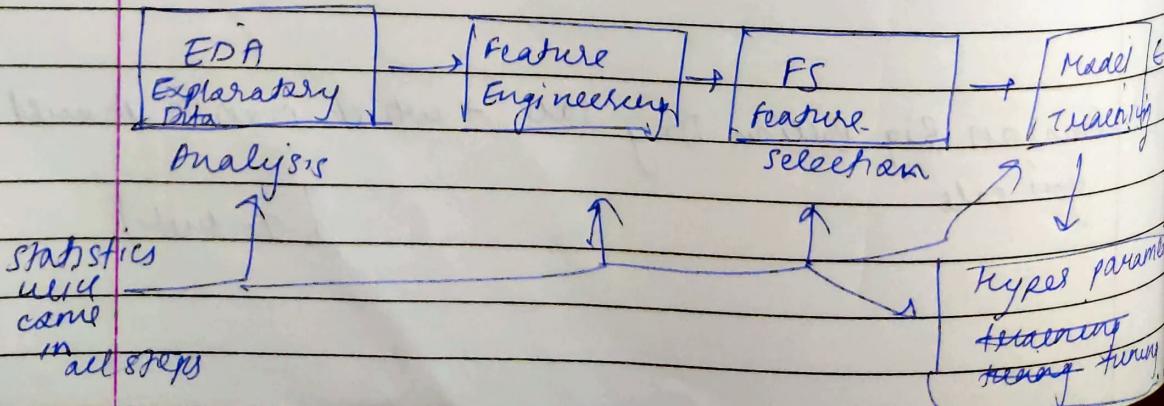
Data Analytics team



Data will be collected from different data sources



Life cycle of DS Project



statistics is used for analysis of data

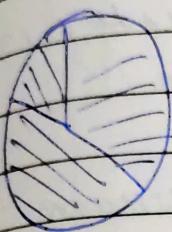
4

descriptive stats



* * *

⇒ summarizing the data



⇒ descriptive stats

Age Age = {12, 13, 14, 18, 20} → finding average age
is also kind of stats
where we are finding central tendency

Descriptive stats

ML
Machine learning

To improve
performance of
models

Statistics

Defn : Statistics is the science of collecting, organizing and analyzing the data.

Data - facts or pieces of information

Ex - ① Ages of students in classroom

{ 24, 25, 37, 29, 28 } we can find mean, median, mode, standard deviation

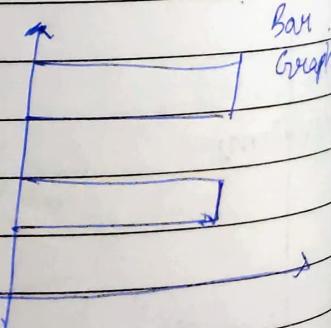
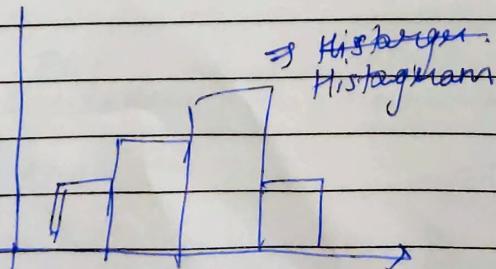
② Weights of students in classroom

statistics

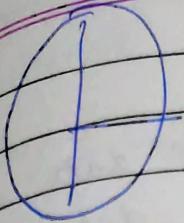
Descriptive (used in stats) (EDA + FE)

Inferential stats
Inferential

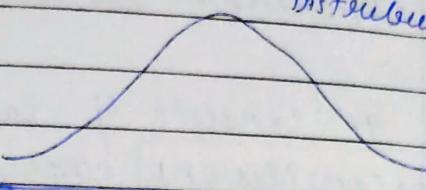
(i) It consists of organizing and summarizing the data using below visualizations



Pie



Distribution



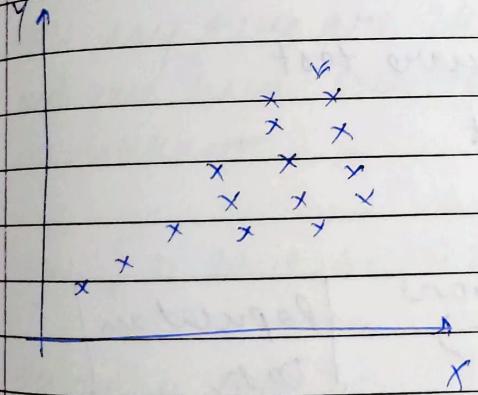
candle stick



Box plot



Example :-



$X \uparrow Y \uparrow$
 $X \downarrow Y \downarrow$

height weight



$X \uparrow Y \downarrow$

Inferential statistics

→ It consists of collecting sample data and making conclusions about population data using some experiments.

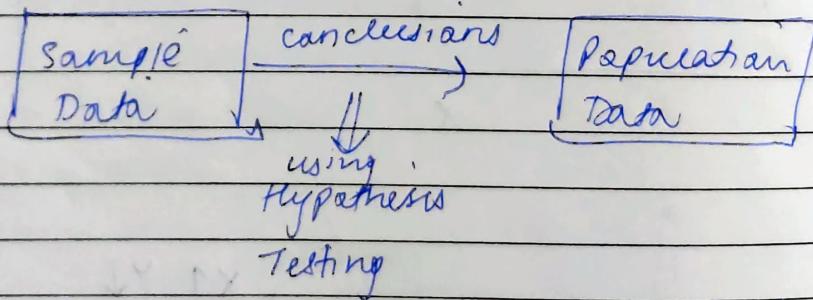
Hypothesis Testing

Example: -

University of 500 people
Class A — 60 people

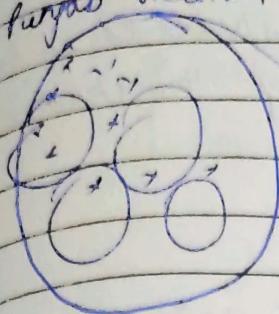
↓
[sample data] — Average age of the university can be found using

- ①
- ② t test
- ③ chi square test
- ④ F test



Sample Data vs Population Data

Punjab Election



sample size = 1000

Total population - 10 crore



Population Data

→ Exit poll results are given using hypothesis testing

Example :-

Let's say there are 20 classrooms in a university and you have collected the age of students in one classroom.

$$\text{Age} = \{21, 20, 18, 34, 17, 22, 19, 34, 25, 26, 33\}$$

Descriptive stats :-

- What is the average age of students in the classroom?
- Relationship b/w age and weight

Population $\rightarrow N$
Sample $\rightarrow n$

Page No.:
Date:

Inferential statistics

Is Due the average age of the students in the class less than the average age of the students in the university?

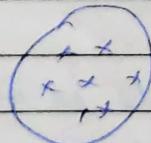
1000 students

Class A =	50 girls	50 Boys
	↓	↓
	95%	92%

The above data can we do a conclusion that girls have performed better than boys
we need to do hypothesis testing

Sampling Techniques

- ① Simple Random Sampling : - Every member of the population (N) has an equal chance of being selected for your sample.



Exit poll - Random sampling
General survey - 4
Census - 4
Movie review - 4

1. Stratified sampling

Gender $\left\{ \begin{array}{l} \text{Male} \\ \text{Female} \end{array} \right.$

2. Education $\left\{ \begin{array}{l} \text{High school} \\ \text{Degree} \quad \left[\begin{array}{l} \rightarrow \text{Master} \\ \rightarrow \text{PhD} \end{array} \right. \end{array} \right.$

In exits polls, first we apply stratified sampling
on

age < 18 & ≥ 18 years, then apply
random sampling

systematic sampling \rightarrow

selecting every n^{th} individual out of Population (N)

In a mall, one credit card seller thinks that
they will approach every

5th person $\left\{ \begin{array}{l} \text{Person} \xleftarrow{f} 0 \\ \text{Person} \xrightarrow{f} 0 \end{array} \right. \rightarrow$ 5th person

(4) Convenience sampling — only those who are interested in the survey will only participate

- Survey regarding new technology - Convenience Sampling
- RBI survey - only women - stratified + Random Sampling
→ Married women
- Credit card calls → stratified + Random Sampling

Variable :-

A variable is a property that can take any values

variable

age = 14

age = 25

variables

Ages = [24, 25, 16] → collection

Two types :-

(1) Quantitative Variable - measured numerically
{ mathematical operations can be done }

Ex - Age, weight, height, rainfall (con)

Qualitative variables → Categorical variables

(Based on some characteristic they are grouped together)

Ex - Grades, types of flowers, types of movies.

Quantitative variable

1

↓
discrete variables

↓
continuous variables

Eg - whole number

Eg - No. of bank accounts
1, 2, 3, 5 2.5x

No. of children - whole numbers

Eg - continuous

Eg - Height, weight,
ages,
rain

Rainfall

Example -

Marital status - Categorical

Ganga river length - continuous

Horse deviation - continuous

Pineapple - Discrete

IQ - continuous Discrete