Statistics

Usecase

HDFC

doc → A → ATM

30Kms

C → ATM

doc → B → ATM

PRODUCT BASED COMPANY

X Statistician ⟶ 5 years

- DATA ANALYST
- Business Analyst
- DATA SCIENTIST
- PRODUCT MANAGERS

① DATA ANALYST

② DATA SCIENTIST

② Find the average size of the shark throughout the world?

blue whale

Amazon

③ Amazon Big Billion Day Sale {Intuit} ⟶ Which month should you select?

Statistics {life cycle of DATA Science Project}

HCL

Google → Domain Expert

SERVICE {TCS}

④

DATA SCIENCE

Requirement Gathering

Business Analyst

PRODUCT MANAGER

PROJECT MANAGER

DATA ANALYTICS TEAM

DATA ANALYST / DATA SCIENTIST

← PRODUCT MANAGER

← BA

Apple

PRODUCT BASED

Domain Knowledge

{Google}

Youtube, GPAY, Google Ads, GMAIL

SALES

Domain Expertise

Product Manager

① DATA ANALST

② DATA SCIENTIST

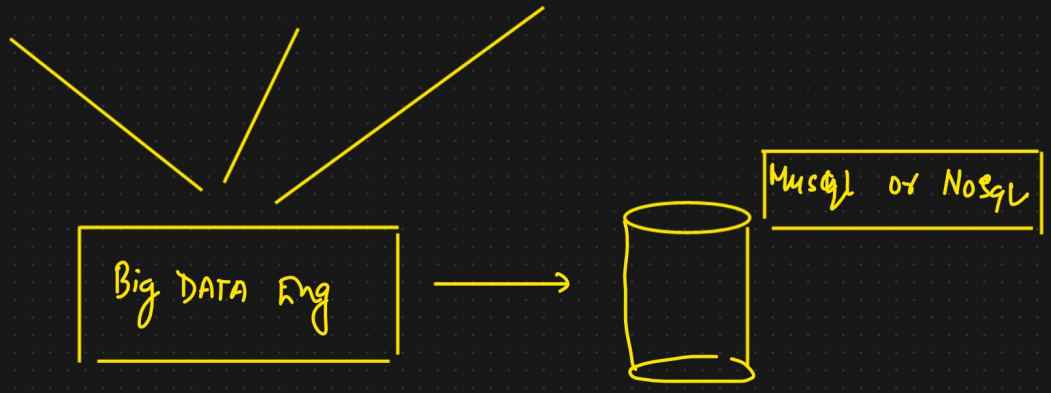
③ BIG DATA Engineers

④ Cloud Engineers

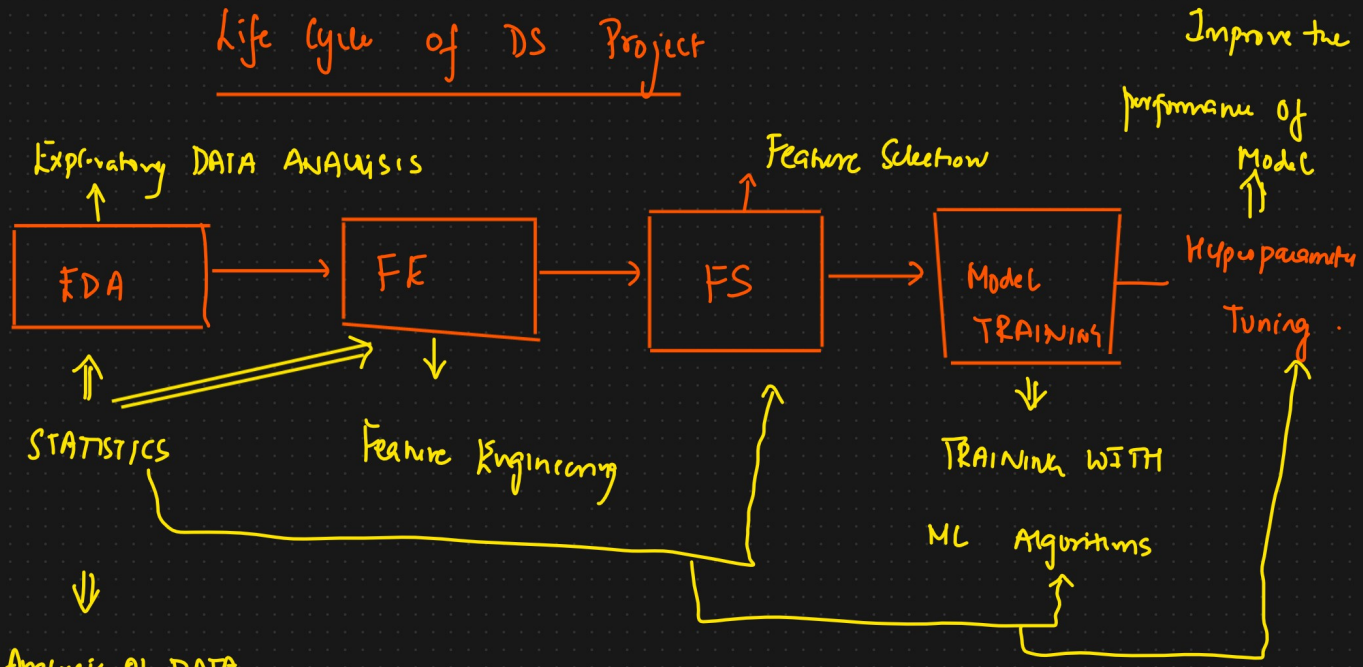
Internal DATABASE

3rd party API's

Web Scraping

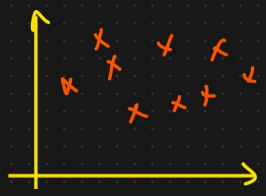


Life Cycle of DS Project



Analysis of DATA

DESCRIPTIVE STATS



⇒ Summarizing the DATA

⇒ DESCRIPTIVE STATS

Age = { 12, 13, 14, 18, 20, 25 } ⇒ Average Age ⇒ Measure of Central Tendency

⇓

DESCRIPTIVE STATS

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

Data : "facts or pieces of information"

Eg: Ages of students in classroom

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

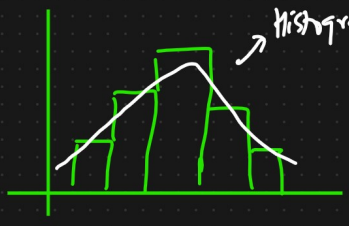
Standard deviation

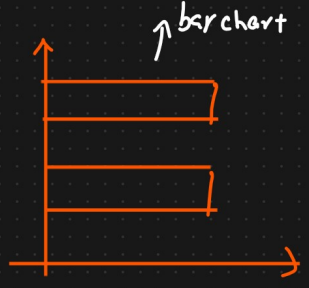
② Weights of students in classroom

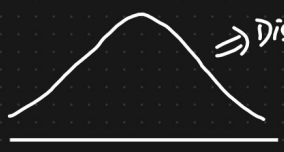
STATISTICS

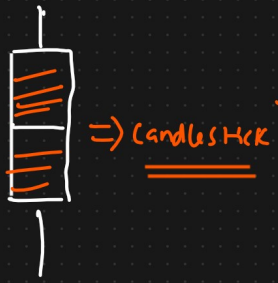
Descriptive Stats [EDA + FE]

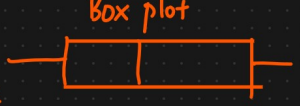
① It consists of organising and summarising the data.











Inferential Stats

② It consists of collecting sample data and making conclusion about population data using some experiments

Hypothesis Testing

University → 500 people

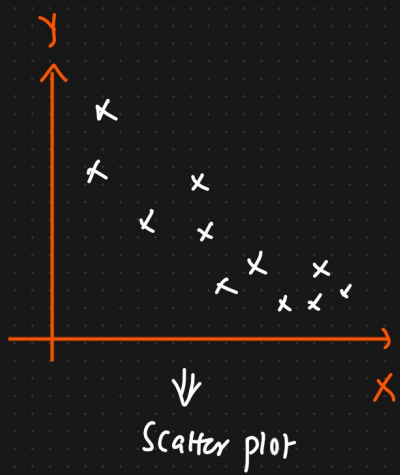
Class A → 60 people

Sample data ⇒ Age ⇒ Average age of the entire university

Yes

Hypothesis Testing

C.I ⇒ Confidence Interval



(height) x — (weight) y

x↑	y↑
x↓	y↓
x↑	y↓
x↓	y↑

p - value

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

Sample Data → Conclusions → Population Data

Conclusions ↓ Hypothesis Testing

Sample DATA Vs Population DATA

EXIT POLL ⇑

PARTY A will win

PARTY B will loose ✓

Eg: Let's say 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 between Age & Gender?

Inferential Stats : Are the average age of the students in the classroom less than the average age of the students in the university?

⇓

Greater
⇓
Equal

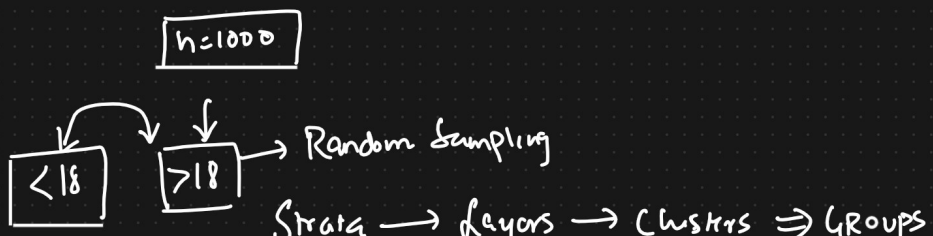
Hypothesis Testing
⇑
1000 Students

Class A ⇒ 50 girls → ⇓ 95%
50 boys → ⇓ 92%

Choose a sample
⇑
Sampling Techniques

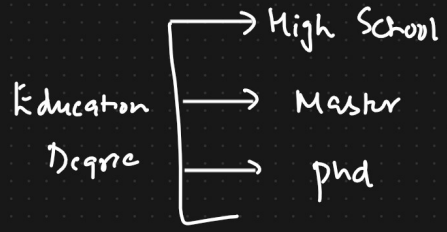
Population (N) Sample (n)

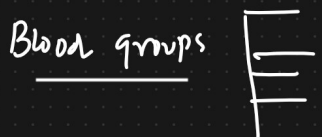
① Simple Random Sampling : Every member of the population (N) has an equal chance of being selected for your sample (n)

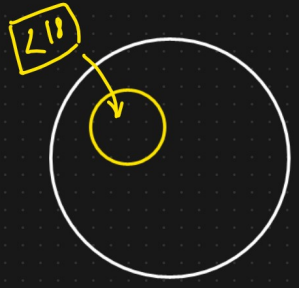


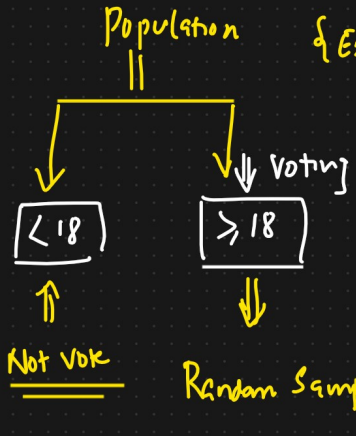
Strata → Layers → Clusters ⇒ GROUPS

② Stratified Sampling







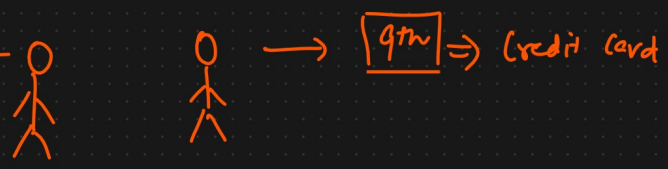


③ Systematic Sampling → {AIRPORT}

n^{th} person

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

{CREDIT CARD}



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

④ Convenience Sampling ÷ Only those who are interested in the survey will only participate

{DATA SCIENCE SURVEY → General AI Survey}

Interview job for a specific ⇓ {Fill the Form}

① Survey Regarding New Technology ⇒ Convenience Sampling

② RBI Survey ⇒ Women ⇒ Stratified + Random Sampling · Married Women

③ Credit Card ÷ Stratified + Random Sampling

① Variable : A variable is a property that can take any values

Eg: age = 14

age = 25

age = 100

Variables

Ages = [24, 25, 26, 27, 28, 29] ⇒ Collection

Two diffcunt type of variable

① Quantitative Variable ⟶ Measured Numerically {Mathematical Operation}.

Ex: Age, weight, height, rainfall (cm), temp, distance

② Qualitative Variables ⟶ Categorical Variables { Based on some characteristics they are grouped together }.

Eg: Gender, Types of flowers, Types of Movies

Quantitative variable

↓ Discrete Variable

↓ Continuous Variable.

Eg: Whole number → fixed

Eg: No of Bank Accounts

{1, 2, 3, 4, 5}

2.5 X

Eg: No of Children : Whole numbers

Pincode = fixed

Eg: Continuous → Decimal values

Eg: Height, weight, ages, Rainfall

Speed

Categorical variables.

Marital → Married

→ Not Married

Assessment

Gender ? Categorical

① What kind of variable is Marital Status ? Categorical variable

Ganga River length ? Continuous

Movie duration ? Continuous

Pincode ? Discrete

IQ ? Discrete

105.75, 90.5,

Pancard

AMLPA - -

↓

Categorical variables

Pincode

360059

360058

360057

Fixed

⇒ It is many ?

Categorical

↓

FE

does no. of Categories

Continuous

Discrete ←

Continuous

Whole number

Bank Account = { 2, 3, 4, 5 }

Pincode = { }

C

6

Categorical

	Cities	Gender	Pincode
5	—	M	—
	—	F	—
	—		—
	—		—
			—
			—
			—
			—

PAN ←

⇒ Categorical

Profile Building

1. LinkedIn
2. Github
3. Instagram
4. Resume
5. Mock Interviews

⇉ Jobs, opportunities

3 → Freelancing