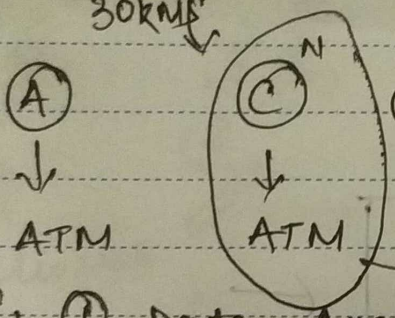
Neuron : Statistics 03/09/2022

↳ DATA ANALYST
↳ BUSINESS ANALSYT
↳ DATA SCIENTIST
↳ PRODUCT MANAGERS

Use case

HDFC

30KMS



↳ ① Data Analyst
② DATA scientist.

① Asked in Product Based company.

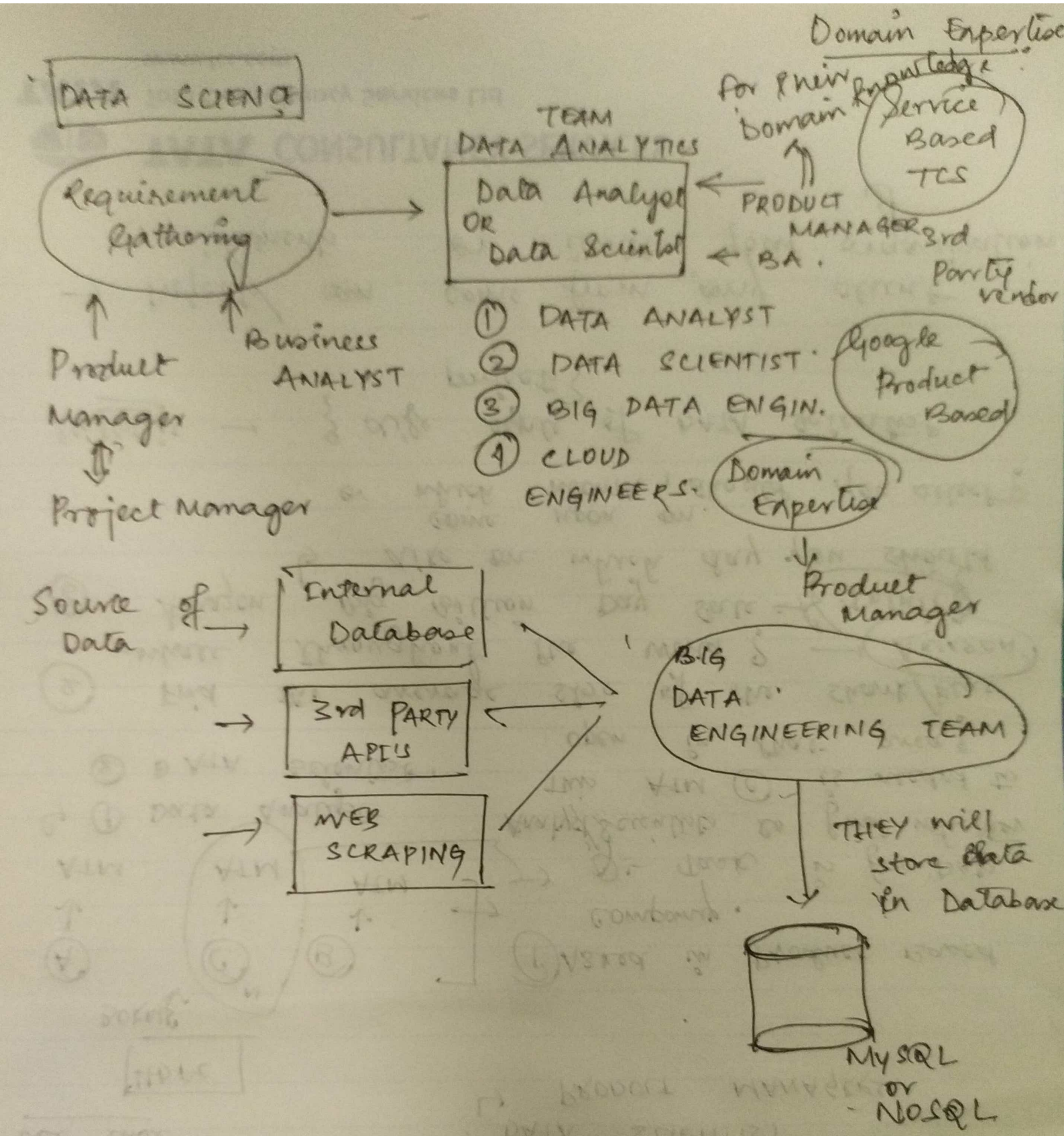
Q:- Task is for Data Analyst/Scientists to find whether this ATM (C) is needed to open in that area?

② Find the average size of the shark/Blue whale throughout the world? → AMAZON

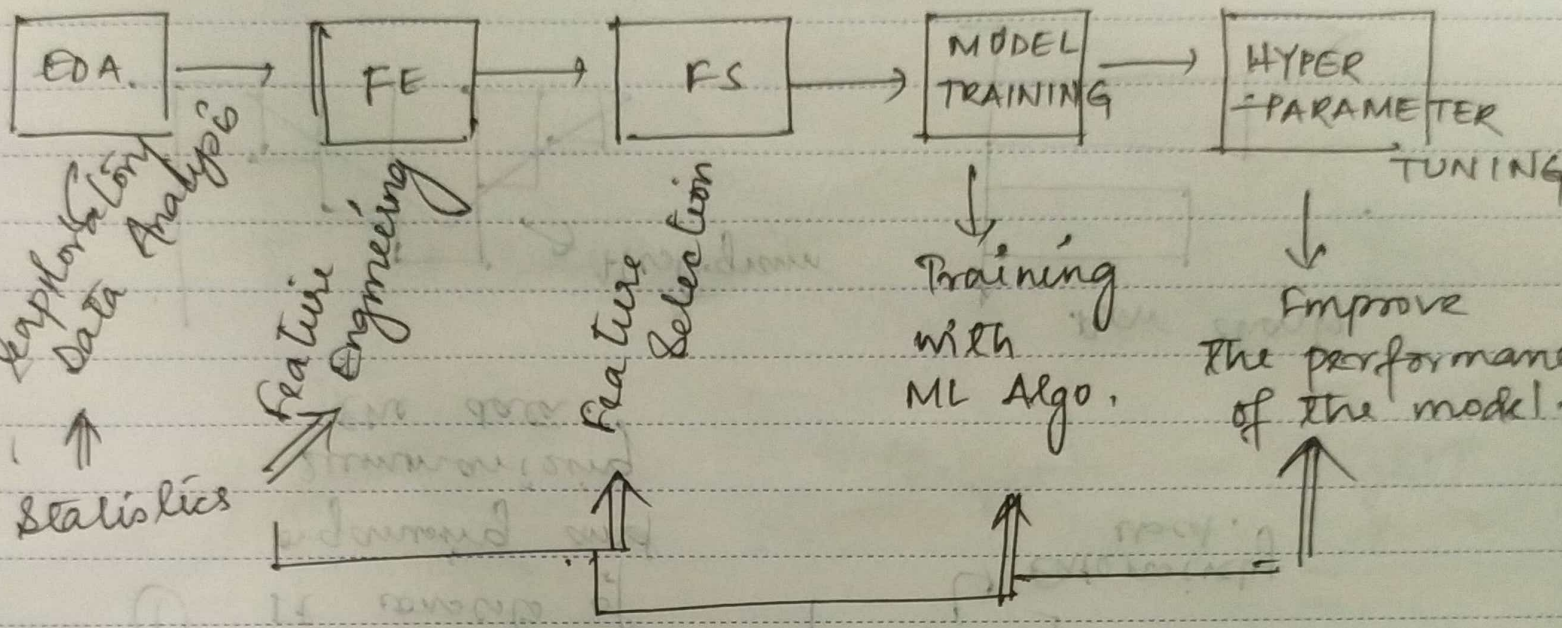
③ Amazon Big Billion Day Sale → Intuit
↳ like on which day you should come upon on.
or which month, should you select?

Statistics → { life cycle of DATA scientist projects }

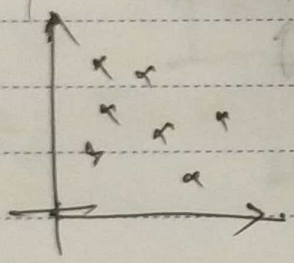
→ Projects/Requirements can come from any client or inside your organizatio


DATA SCIENCE
Requirement Gathering
Product Manager
Business ANALYST
Project Manager
TEAM
DATA ANALYTICS
Data Analyst
OR
Data Scientist
BA.
PRODUCT MANAGER
for their domain knowledge
Domain Expertise
Service Based TCS
3rd Party vendor
1 DATA ANALYST
2 DATA SCIENTIST.
3 BIG DATA ENGIN.
4 CLOUD ENGINEERS.
Google Product Based
Domain Expertise
Product Manager
Source of Data
Internal Database
3rd PARTY APIs
WEB SCRAPING
BIG DATA ENGINEERING TEAM
THEY will store data in Database
MySQL
or
NoSQL

Life CYCLE OF DATA SCIENCE PROJECT

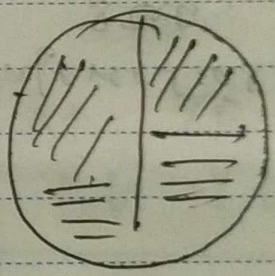


↳ Analysis of DATA



⇒ Summarying the Data

descriptive Stats



pie charts part of descriptive statistics.

Age = { 12, 13, 14, 18, 20, 25 }

⇒ Average Age?

Also known as Measures of central Tendency.

] Part of descriptive statics

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

Data → facts or pieces of information.

Eg: Ages of students in classrooms.

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

① Standard deviation

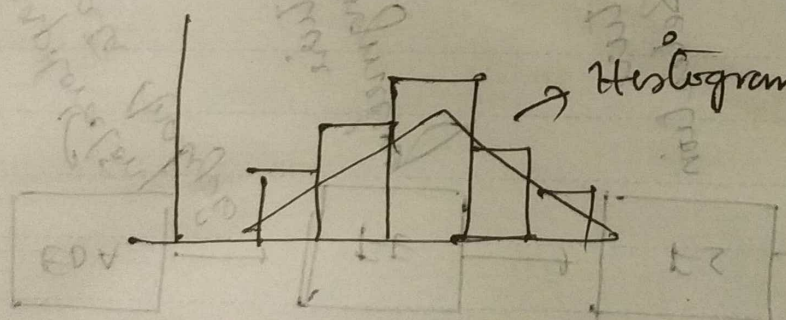
② Weights of students is classrooms.

STATISTICS

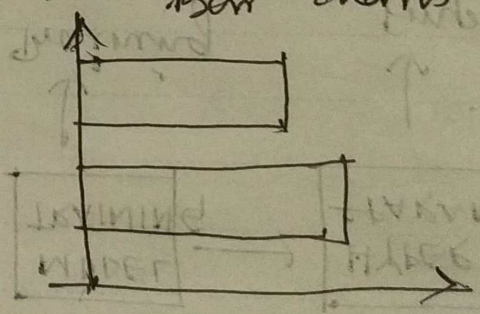
Descriptive Stats → [EDA + FE] ← Inferential Stats.

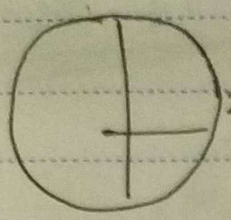
Extensively used.

① It consists of organising and summarizing the data.

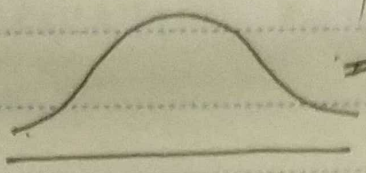


Bar charts

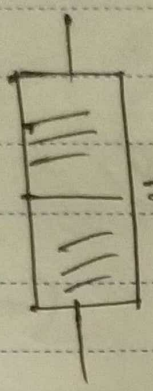




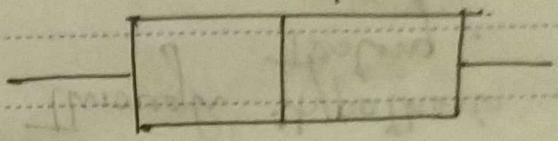
Pie



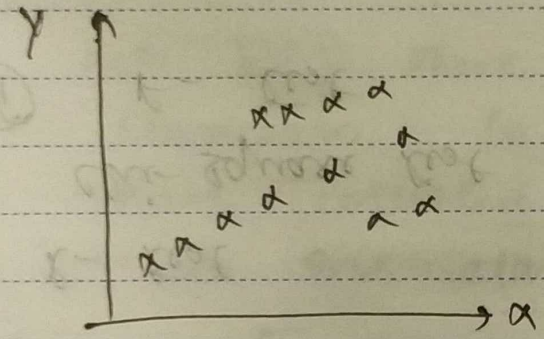
Gaussian distribution



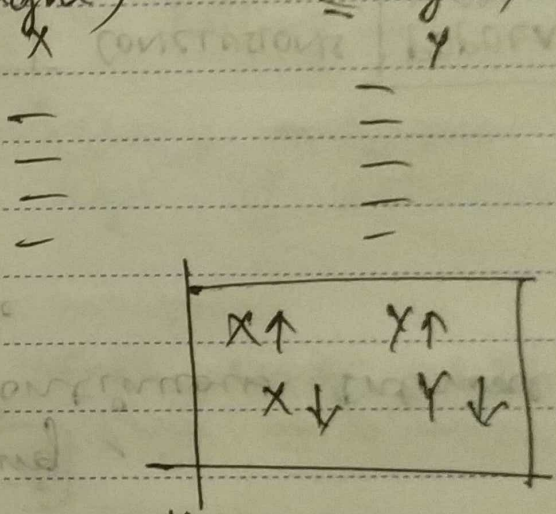
candelstick



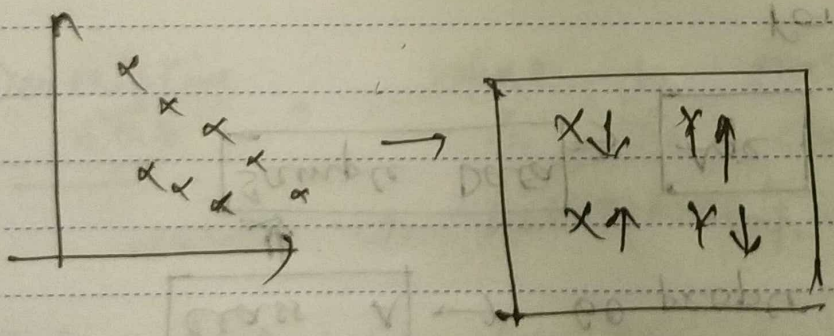
Berr plot



scatter plot



Here we are able to summarize the data. from the graph.



Inferential Stats.

① 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 ⇒ Avg. age of Entire university

For this we will use Hypothesis Testing,

CF = Continuous Interval

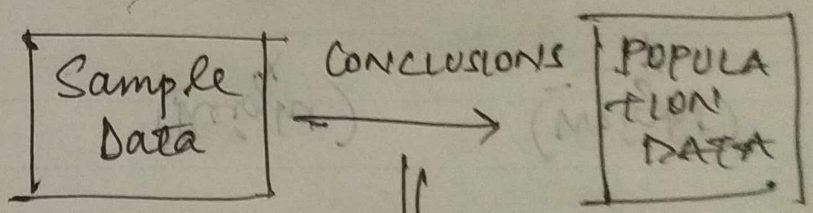
P-value

① Z- test

② t- test

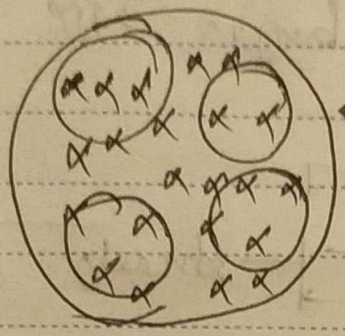
③ Chi- square test

④ F- test



Through Hypothesis Testing.

SAMPLE DATA Vs POPULATION DATA.



Sample Size = 1000 (For various Region)

Punjab → 10 Cr → Population Data

EXIT POLL

Suppose : PARTY A will win / PARTY B will Loose

↳ Risk of Wrong Hypothesis Testing.

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

Ages : [21, 20, 18, 34, 17, 22, 24, 25, 26, 22, 23]

weight : [: _____]

Descriptive Stats : → What is the avg. of students in the classrooms.?

→ Relationship b/w age & gender?

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

1000 Students

Class A ⇒ 50 girls 50 boys

95% ↓ 92%

POPULATION → (N)

SAMPLE → (n)

Sampling Techniques

(1) Simple Random Sampling :-

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

Ex :- Exit Poll.

Lottery

Picking up a marbel from a beg full of marbels.

(2) Sorafied Sampling.

Strata → Layers → Clusters ⇒ Groups.

Gender → Male, Female

Eductation Degree → High School, Master, Phd.

Blood Groups → A+, A-, —

<18 | >18

Like voting & Non voting groups ← Sorafied sampling

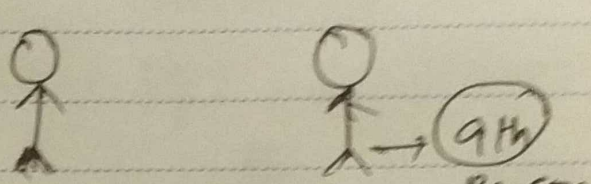
Random Sampling

Groups >18 <18

(3) ~~Systematic Sampling~~ :- ← {AIRPORT}

Systematic Sampling {CREDIT CARD}

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



→ 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}

Ex):

(1) Survey Regarding New Technology :→ Convience Sampling

(2) RBI Survey ⇒ women → Stratified Sampling + Random Sampling → Married women.

(3) Credit Card :

(4) Stratified + Random Sampling

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

Eg :- 19 age = 100
age = 25

Variables : Ages = [24, 25, 26 ...] ⇒

→ Two different types of variables:

① Quantitative variable → Measured Numerically { Mathematical operations }
Eg :- Age, weight, height, rainfall (cm), temp, distance.

② Qualitative variables → categorical variables
Eg :- Gender, Types of flowers, Types of Movies.

{ Based on some characteristics they are grouped together }.

Quantitative Variables

Discrete variable

Eg: whole number

Eg: No. of Bank accounts
{1, 2, 3, 4, 5} {2.5} ✗

Eg: No. of children :
(Always will focus on whole numbers)

Continuous variable

Eg: Continuous

Eg: Height, weight, Ages, Rainfall, Speed.

ASSESSMENT :-

(1) What kind of variable is : Marital Status
→ Categorial variable

(2) Ganga River length → Continuous

(3) Movie duration → "

(4) Periods → Discrete

(5) IQ → ~~Continuous~~ Discrete