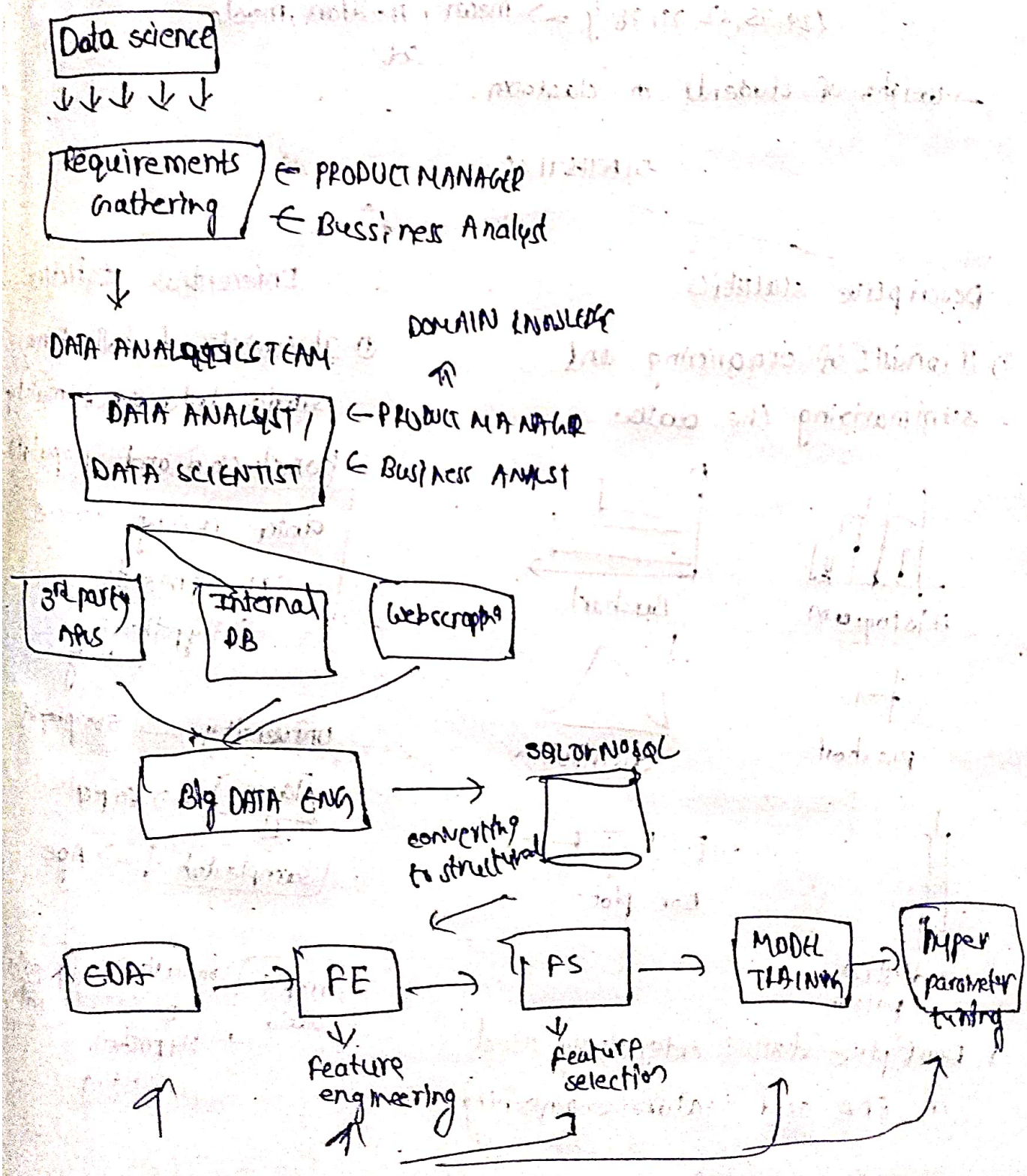
STATISTICS

* Find the avg size of shark throughout the world?
* Amazon Big Billion day sale best date? → which month should you select?

LIFE CYCLE OF DATA SCIENCE PROJECT



statistics will be used → Analysis of data and summarizing of data.

STATISTICS:-

statistics is the science of collecting, organizing and analyzing the data.

Data:-

→ facts or pieces of information.

Eg: Ages of students in classroom.

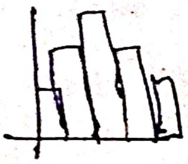
{24, 25, 32, 29, 28} ⟹ mean, median, mode, sd

→ weights of students in classroom

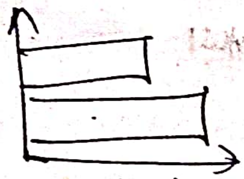
STATISTICS

Descriptive statistics

① it consists of organizing and summarizing the data.



Histogram

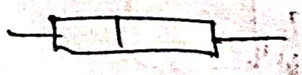


Barchart.

pie chart

distribution

candlestick pattern



Box plot.

* Descriptive stats is extensively used in EDA and feature engineering

Inferential statistics.

① It consists of collecting sample data and making conclusion and population data using some experiments. → Hypothesis testing.

University → 500 people

Class → 60 people

⇓

Sample data ⟹ Age.

Sample data → Conclusions → population Data

⇓

Hypothesis Testing.

SAMPLE DATA vs population data

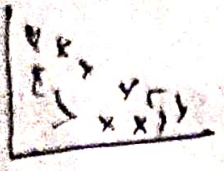
scatter plot

Punjab

10CR

Population DATA

Sample size=1000



Population data (N)

Sample data (n)

eg: Lets 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:

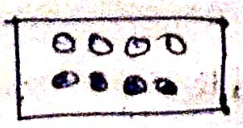
1) what is the avg age of students in the classroom?
2) Relationship b/w age & weight?

Inferential stats:

1) Are the avg age of students in the classroom less than avg. age of students in the university?

Sampling Techniques:

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



N=8 n=4

Picking each marble probability is same.

2) stratified sampling:

Sample will be selected based on groups, Layers and clusters.

Gender: MALE, FEMALE

education degree: High school, Master, PHD.

3) systematic sampling : → Airport   nth person

{credit card}

every 8th person → every 9th person.

→ select every nth individual out of population (N).

4) convinience sampling:

only those who are interested in the survey will only be participated.

{AI survey → General AI survey}

Variable

A variable is a property that can take any values.

ex: age = 14
age = 25
age = 100

→ Two types of variables:

1) Quantitative variable → measured Numerically
{Mathematical operation}
ex: age, weight, rainfall, temperature.

2) qualitative variable → categorical variables {Based on some characteristics}
ex: Gender, Types of movies.

Quantitative variable

↓ Discrete variable →
ex:
→ No. of children → whole number → limited
→ No. of bank accounts.

↓ Continuous variable:
→ decimal; ex: age, height,
→ unlimited speed, rainfall.

Assessment :

- what kind of variable is Marital status ← Categorical variable
- " " " " " Ganga river length ← continuous
- " " " " Movie Duration ← continuous
- " " " Pincode ← Discrete
- " " " " IQ ! ← Continuous.

Day-2

AGENDA → Histograms
Measure of CT
MEASURE OF DISPERSION
PERCENTILES & QUARTILES
5 NUMBER SUMMARY (BOX PLOT)

Histogram :

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 56, 70, 90, 95, 100}

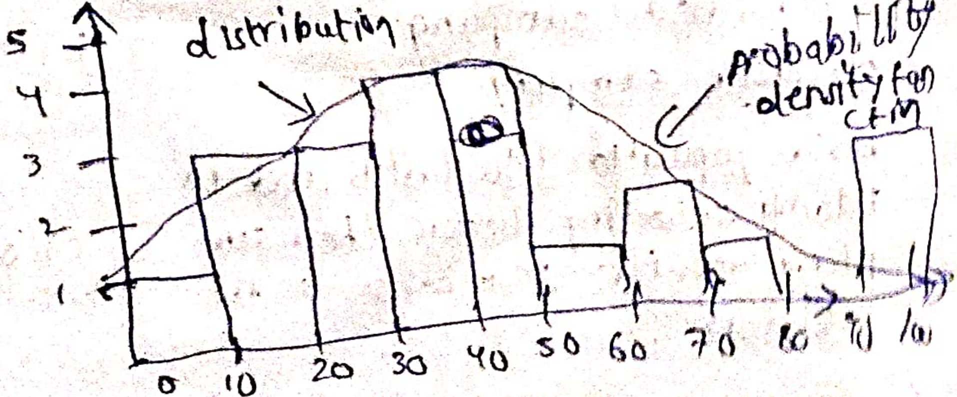
steps to create histogram

1. sort the numbers
2. Bins → Number of groups
3. Bin size → size of Bins.

Number of Bins = $\frac{\text{Max value}}{\text{size of Bin}}$ $= \frac{100-0}{10} = 10$

suppose we want 10 binsize then Number of groups = 10

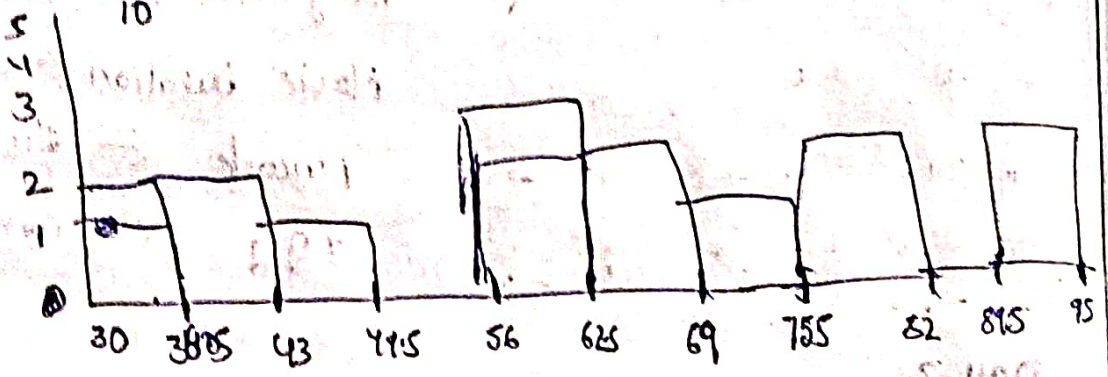
suppose we want 20 bins then, binsize = $\frac{100-0}{20} = 5$



weight = {30, 35, 35, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 98, 98}

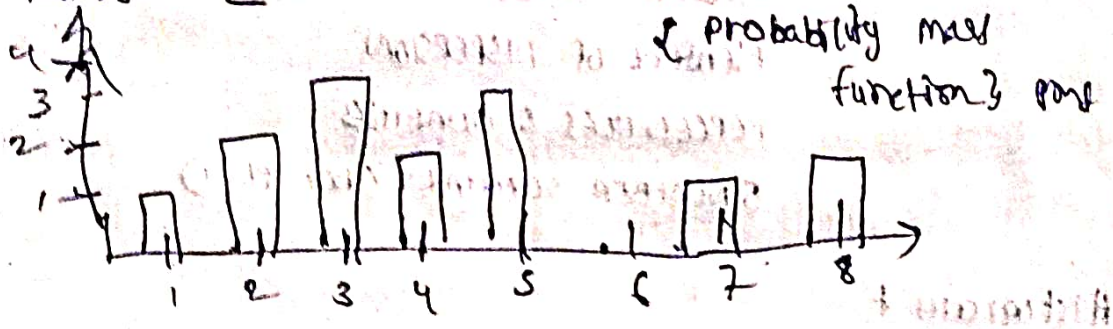
bins = 10

binsize = $\frac{95-30}{10} = 6.5$



Discrete Bank account

No. of Banks = [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5]



{ probability mass function } pmf

→ Pdf for continous variable

→ Pmf for discrete variable.

Sampling

probability sampling

a) simple random sampling

b) systematic sampling

c) stratified sampling

d) cluster sampling

Divide population into clusters and randomly selecting clusters, then sampling all the members in the clusters.

Non probability sampling.

a) convinience sampling

selecting individual who are easy to reach

b) Judgemental sampling

select individual based on researcher's judgement.

c) snowball sampling

Existing study subjects recruit future subjects

multi-stage sampling.

combining multiple sampling method. Involves selecting clusters and randomly sampling

among their aquantances

4) Quota sampling

Age, Group, Gender, caste.

→ selecting sampling depending on usecases.

Qualitative data

Nominal

↓ No Rank

Eg: M/F → Gender

Blood Group

Pincode

→ categorical values

Ordinal

- Categorical data we can assign rank

Good Bad better

1 3 2

ext customer feedback.

scales of measurement:

→ scales of measurement describe the nature of information within the values assigned to variable

+ primary scales of measurement

1) Nominal side
2) ordinal side
3) Interval
4) Ratio.

this scale classifies data into distinct categories that do not have an intrinsic order.

characteristics:

i) Data is categorized based on names, labels, or qualities

ii) Categories are mutually exclusive.

iii) No logical order among categories

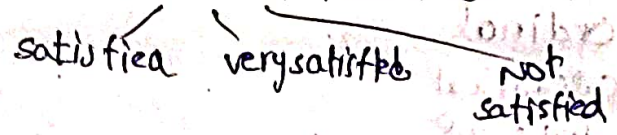
ordinal scale:

This scale classifies data into categories that can be ranked or ordered.

characteristics:

i) Data is categorized and ranked in specific order
ii) The interval between ranks are not necessarily equal.

eg: customer feedback



Interval scale:

The Interval scale not only categorizes and orders but also specify the exact difference between intervals. It lacks a true zero point.

characteristics:

1) Data is ordered with consistent interval between values.
2) Allows for meaningful comparison of values.
3) No true zero point.

eg: Temp in fahrenheit — 10°F, 20°F, 30°F
calendar years — 2024, 2016, 2020

Ratio scale:

The order matters. The differences are measurable. It contains a zero starting point.

eg: student marks in a class

ASSIGNMENT

① length of different rivers in world → Interval scale.
② favorite food based on Gender? → ordinal scale.
③ Marital status? → Nominal scale.
④ IQ measurement → Interval scale