STATISTICS FOR DATA SCIENCE

Usecase : ① Bank → HDFC

Q/ HDFC have already two ATM's at location Ⓐ & Ⓑ. They are planning whether they will set up ATM at location Ⓒ or not ?

Ⓐ ATM Ⓑ ATM Ⓒ ATM

This task is basically for data analyst or data scientist

Usecase ② : Find the avg. size of the shark throughout the world ?

Amazon Interview Quest^n for DA

③ Amazon Big Billion Day Sale
{asked in Intuit (product based) company}
* which month should you select for Big Billion Day Sale ?

Statistics [Life cycle of DATA SCIENCE PROJECT]

①

Requirement gathering → Business Analyst will discuss with Product Manager & write all the requirement

↑↑ Responsible for this Product Manager or Project Manager

② → DA OR DS (Sent to this team)

(i) Data Analyst
(ii) Data Scientist
(iii) Big Data Engi
(iv) Cloud Engineer

Service Based Company

↓

e.g. T.C.S., HCL

giving service to client.

Product Based Company

↓

creating on their own product

e.g. Apple, Google, IBM

Google

↓

Sales (department)

↓

Domain Expertise

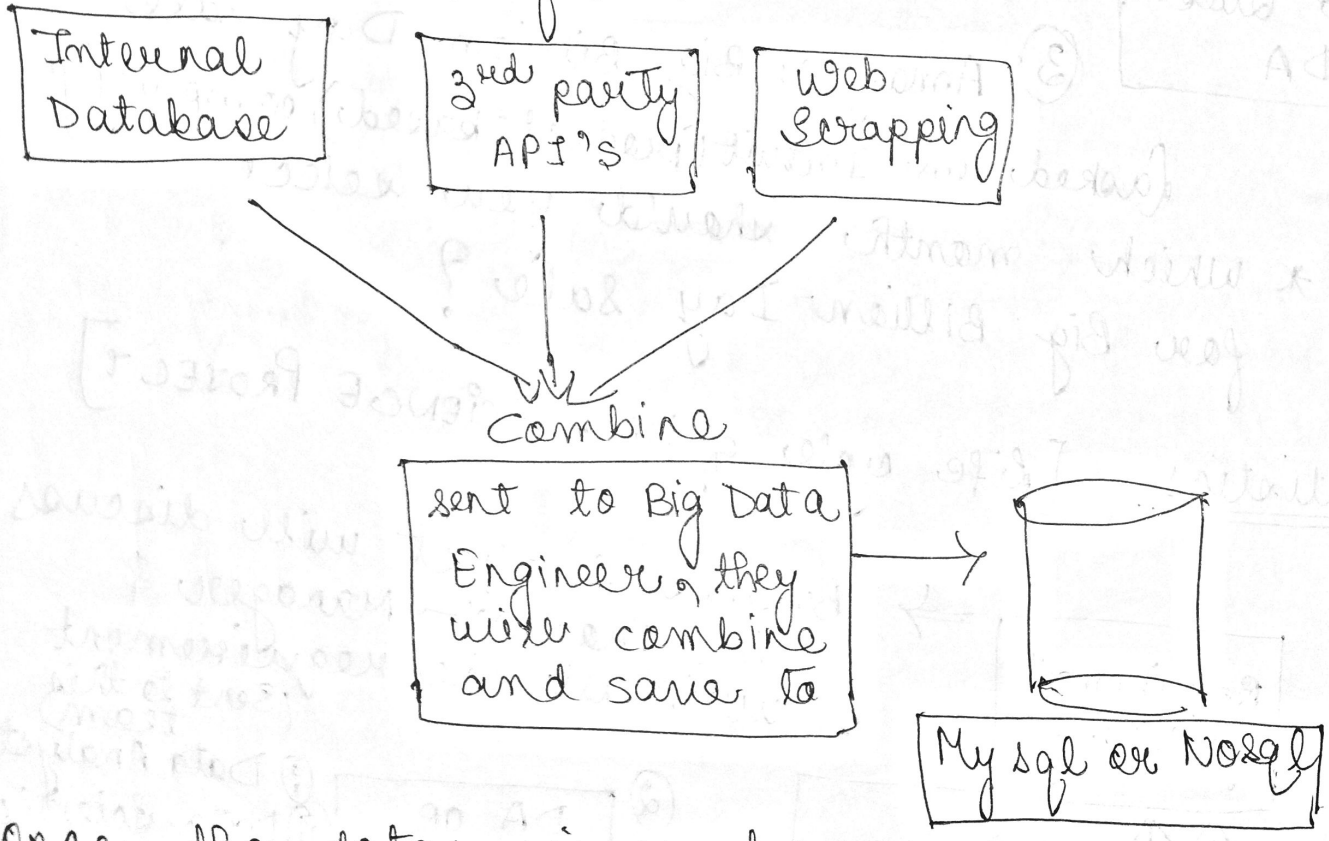
↓

role: Product Manager

② ~~DA OR DS skill req~~

Product Manager and BA will require here to discuss where the data came from with DA or DS.

DATA CAN GET FROM



Once the data is ready life cycle of Data Scientist Project start

EDA → FE → FS → Model Training

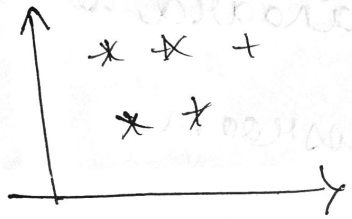
(exploratory Data Analysis) (feature engineering) (feature selection) (training with ML, DL)

Hyperparameter tuning

{Improve the performance of model}

* In all the steps Statistics will be used!!

Analysis of Data



{Summarisation of Data}

⇓

Descriptive Statistics

Pie Chart

⇒ Descriptive Stats

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

Avg. age {is a part of Descriptive Stats} is a measure of Central Tendency

Interview Question How Statistics is used in Machine Learning!!! ***

Definition

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

Data : "facts or pieces of information"

E.g. (1) Ages of students in a classroom

{24, 25, 32, 29, 18}

with the help of this data we can analyse this data by finding mean, median, mode, standard deviation

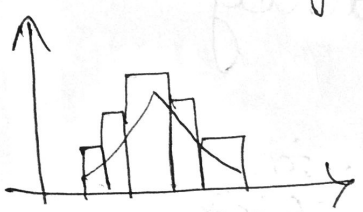
eg. (2) weight of students in a classroom.

TYPE OF STATISTICS

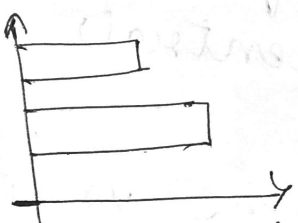
Descriptive Stats [EDA+FE] (used in)

(1) It consist of organising & summarising the data

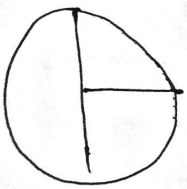
e.g.



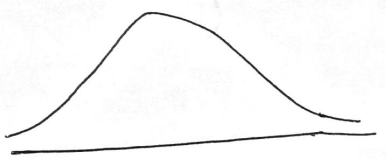
Histogram,



Bar chart

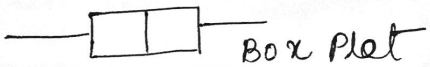


pie,



distribution,

Candles stick



Box Plot

Inferential Stats

(1) It consists of collecting sample data and making conclusions about population data using some experiments.

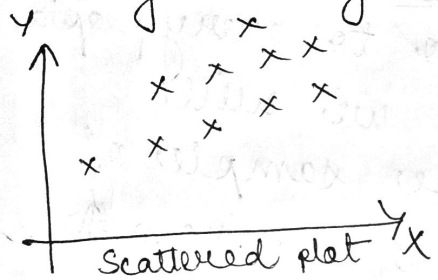
** making conclusions can be done by Hypothesis testing

e.g. of Descriptive Stats

Let, X & Y features

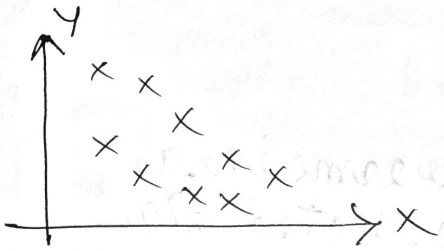
↓ ↓

height weight



⇒ when x ↑, y ↑

⇒ when x ↓, y ↓



⇒ when x ↓, y ↑

⇒ when x ↑, y ↓

e.g. of Inferential Stat

Let We have a University of 500 people, take a 60 people of

class A

↓

Sample data ⇒ Age

↓

Avg. age of entire university

↓

Hypothesis Testing

C.I ⇒ Confidence Interval

p-value

① z-test

② t-test

③ chisquare test

④ F test

We are taking

(Sample data)

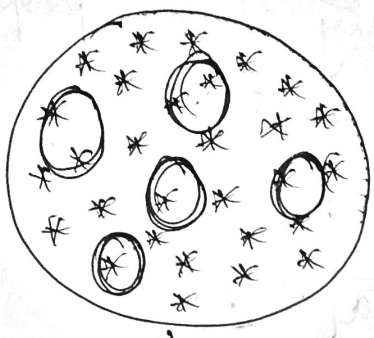
& making conclusions

about (Population data)

through Hypothesis testing.

Sample data Vs Population Data

News will conduct some exit poll



Punjab State
(10cr. popn)

Population Data

we can't go to every people to ask, so we will select some sample of size 1000 and ask from them

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

Ages {21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22}
Weight {: - - - - - - - - - - -}

According to descriptive stats :- we get quest's like

(i) what is the avg. age of students in the classroom

(ii) Relationship betⁿ age and weight.

According to Inferential stats :- we get questions like

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

e.g. ② 1000 students in University

class A → 50 girls 50 boys

avg. marks 95% 92%

Ans:- girls performing well than boys.

* Popn (N) sample (n)

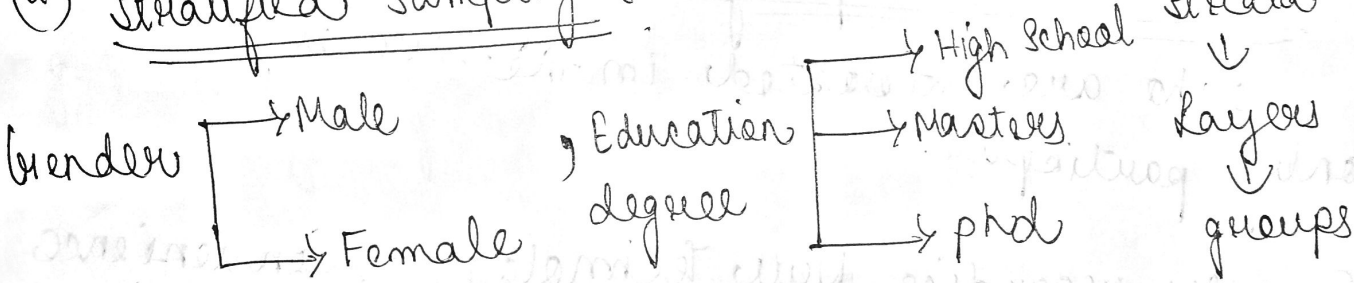
SAMPLING TECHNIQUES

(choosing of sample techniques)

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

e.g. exit poll, movie review, lottery

(2) Stratified Sampling :



Blood group

In Exist Poll, we first apply stratified sampling for neglecting the age <18 & choosing the age >18 [then for that popn we will apply random sampling]

Population [Exit Poll]

↓ Stratified Sampling ↓

<18 year >/=18 year [voting]

↓

Apply Random Sampling

③ Sistematic Sampling

Select every nth individual out of popn (N)

At {Airport}

↓

{Credit card}

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

① Survey regarding New Technology ⇒ Convenience Sampling

② RBI → conduct survey ⇒ for [Women] → Married, Unmarried

we will use (i) Stratified Women B then (ii) Random sampling

③ Credit Card :- Stratified + Random Sampling

caller

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

e.g:- age = 14 , 25 , 1000.

Variables

Ages = [24, 25, 26, 27] ⇒ collection

Two different types of Variable

(1) Quantitative Variable :- Measured Numerically (Mathematical Operators)

e.g. Age , weight , rainfall (cm) , temp^{re} , distr

(2) Qualitative Variables :- Categorical Variables (Based on some characteris-tics)

e.g. ; yender , Types of flowers , types of movies.

Quantitative

Discrete Variable

e.g. ; whole numble values

① No. of Bant Accounts person can have {1, 2, 3, 4, 5}

② No. of children in family

Continous Variable

e.g. ; continuous

① Height , weight , ages , Rainfall , speed

Assessment ① What kind of variable is ① Marital Status
② Yanga River Length
③ Movie duration ④ Pincode ⑤ IQ ⑥ Gender

Ans:- ① Categorical Variable
② Continuous ③ Continuous
④ discrete ⑤ ~~Continuous~~ discrete ⑥ Categorical

~~HISTOGRAMS~~