"what they are"

* Datascience → Mining large amounts of data to identify patterns

↳ + PL, Stats, ML, algorithms

* Big Data → ✓ Huge Data Volumes

+ data Capture, Storage, Sharing Querying.

Data Analytics → process and perform Statistical analysis of data

✓ How data can be used to draw Conclusions, solve problems

① ⟶ "What they do"

Data Scientists → predicts future based on Past patterns

→ capture / examine data from multiple disconnected Sources

→ Develop New Analytical Methods, ML Models

② Big Data prof :- Analyse sys Bottlenecks

✓ large scale data processing Systems creation

✓ Architect scalable distributed Systems

Scanned with CamScanner

* Data Analysts — Acquire, process, summarize data
— package data for insights
— Design / create data reports using various reporting tools

⟶ where used?

* Data Science → Search Engines, Fin. services, E Com.
Big Data → Fin, Comm, Retail
Data Analytics — Healthcare, Travel, IT Industry.

⟶ what skills?

* DS → Prog Skills — Python, Stats, Math, ML, Hadoop, SQL, Story telling, visualization

Big Data → Java, Scala } PL, NoSQL — MongoDB, CassDB,
⟶ Apache Hadoop framework, distributed systems

⟶

Data Analytics → prog skills — R, Stats, Math, visualization skills.

$123K | $88 | $61 ⟶

✓ Tagging people and objects

✓ Google search, driving, ⇒ Reco systems, Self parking, 'like a website' & a song
ML is mandatory for products.

Scanned with CamScanner

(8) ML = using data / to answer questions

"Training" / "prediction / Inference"

Data → Training → Model → Serve prediction

↓

answer qn.

Key (X)

→

7 Steps of ML :-

Wine or beer?

Colour, % } features

Colour	%	Wine (or) Beer

Data preparation

✓ Training, Evaluating

✓ Internet of things (IoT)

✓ Cloud Computing

✓ Big Data

✓ Block chain

✓ Artificial Intelligence

5 technologies

Difference between.

Block Chain :-

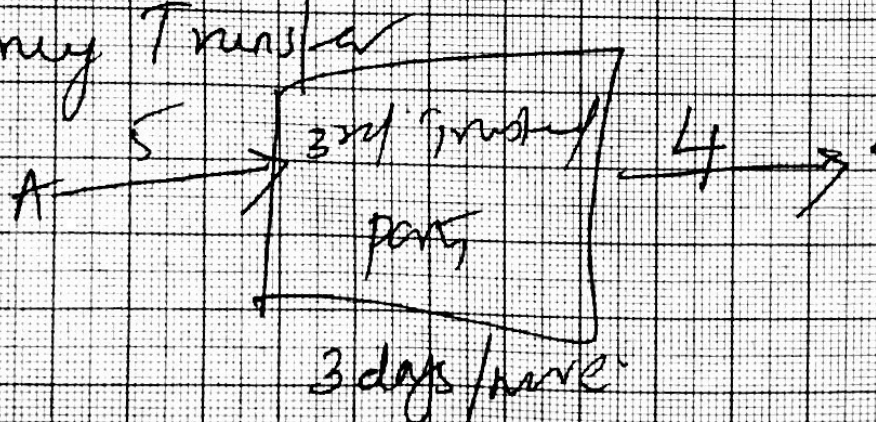
Bit coin ≠ Block Chain

↓

Big Money

enables movement of digital assets.

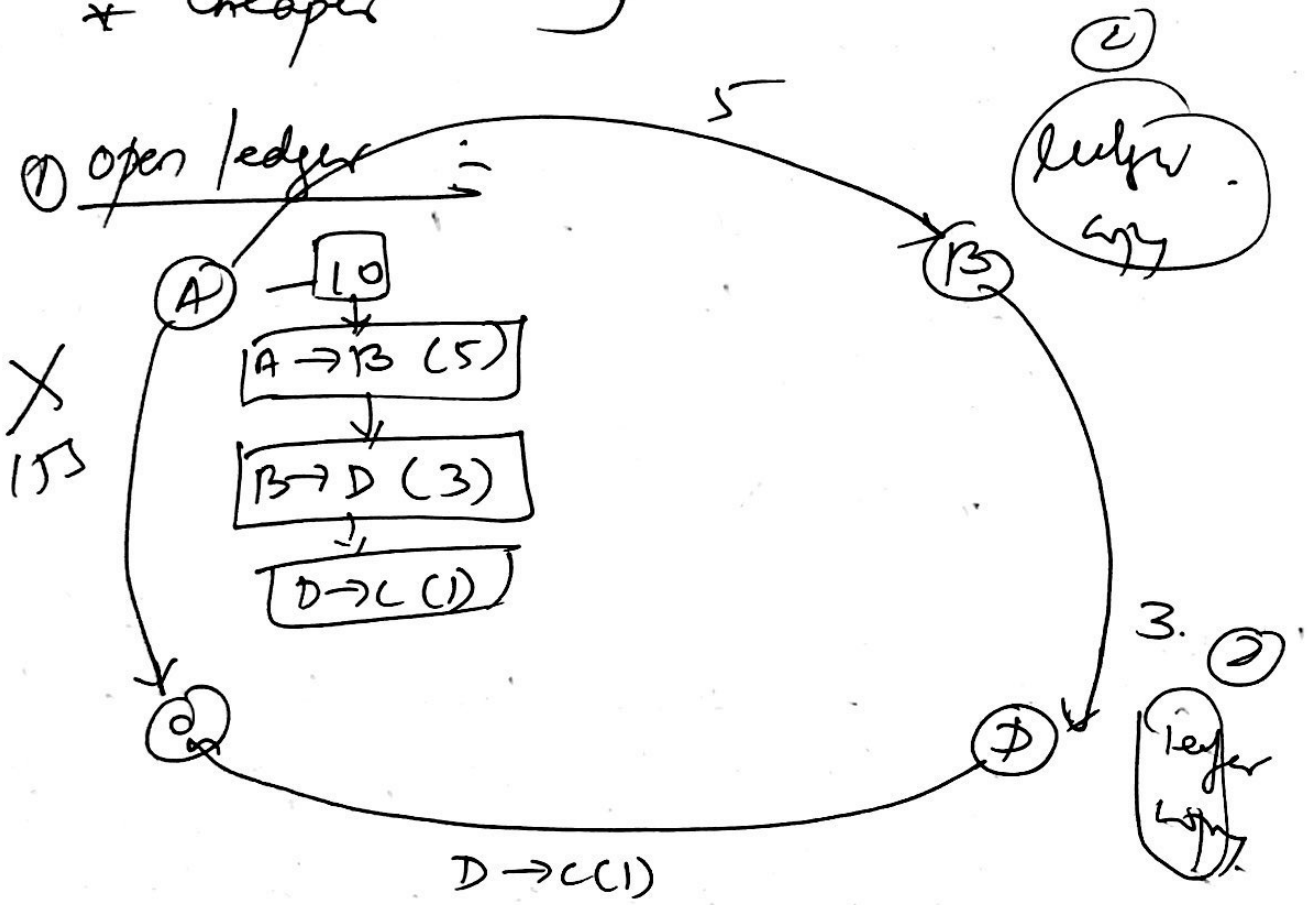
✓ Money Transfer



Scanned with CamScanner

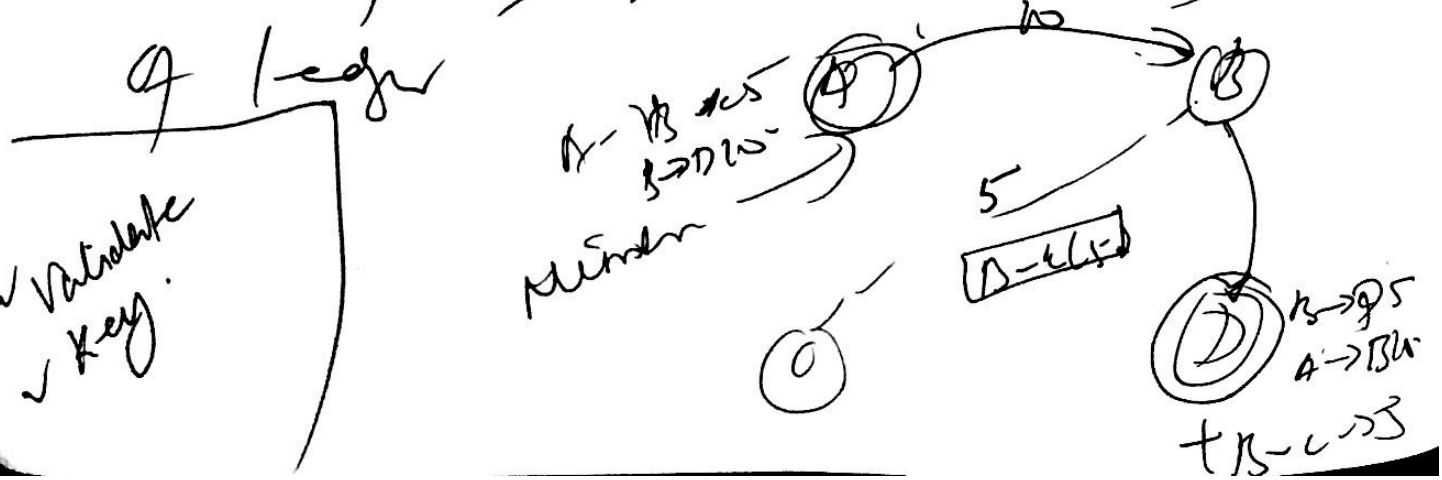
* No third party
* Instantly
* Cheaper

} Transfer → Blockchain



✓ Chain of txn's open to everyone

② Distributed ledger
✓ multiple ledger copies
✓ All parties have same version of ledger



Scanned with CamScanner

✓ Descriptive statistics, Exploratory Data Analysis
→ Data description
→ Data visualization techniques
→ Graph forms, Description of a variable, Relationship between two or variables graphically
→ Data Summarization. — || Descriptive ||

Descriptive → Central Tendency Measures
Mean, Median, Mode.
— Measures of Dispersion,

+ prob Dist^n for expressing data
— More than Mean / Median /
20% data < x; 30% data < y; ....

∴ Not only idea of dispersion centrality — entire characterization of the data set

✓ Inferential statistics — Data is a sample from a larger population

HT of IIITDM students — finite data - but may not be available completely
— Sampling of 50 students

→ Descriptive alone may not suffice

Use of Anova in Inferential Statistics.

Scanned with CamScanner

→ Something about the population

eg: avg ht of IIITDM students < 150, = 150

Sample observation → population prediction is Inferential statistics.

" Sample Observation → population Inference

finite / Infinite population

✓ Regression Analysis

- └ Inferencing, Descriptive, optimization
- └ Overlap with m/c learning
- └ SLR → Dep(o/p) variables relship with indep(i/p) variables
- └ Line fitting for Data
- └ prediction

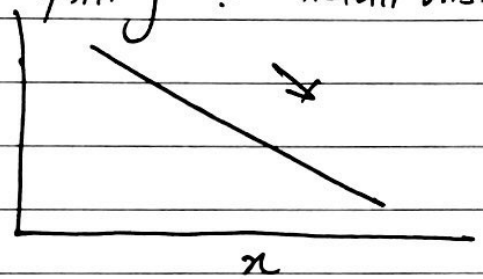
(Indep) examples.

- rainfall → crop yield (dep var)
- speed → Mileage
- exercise → weight
- ..

Reg. Analysis → Relationship b/w dep var (o/p) with 1/more indepen. variables (i/p)

x axis – Indep var – Rainfall
y " → crop yield – Dep.

Line fitting :- Relationship b/w two variables.



" Given Indep variable can I predict Dep var? "
↓ Regression Answers.

ML (essence of prediction) ? Accuracy of given,

" Historic Data " → ML

---

(22). Major part Next – Machine learning

Data Mining / Pattern Recognition / Statistical learning related topics, Not same.

ML → focus on algorithm that converts data to knowledge.

Scanned with CamScanner

(L2 Contd)

Regression for Continuous Variables.

Classifn → o/p variable is discretized

—— Categorical —— Nominal

" ADT of Supervised / Unsupervised learning

↓

Clustering

↓

Association Rule mining

clustering → task of grouping set of objects into clusters / groups based on similarity defined over common set of attributes / features.

Variety of Diverse Groups.

K Means, Hierarchical, etc.

ARM another USL :-

- relships between features over a set of objects.

Scanned with CamScanner

Machine learning → broadly classified as supervised/ unsupervised.

↳ SL → Create a function / relship from training Data : One explicit output Variable — atleast.
labelled data / training data

" Algorithmic mapping between I/p Variable and

✓ Unsupervised → Task of creating patterns from data which have no explicit features
Patterns from unlabelled data (No o/p Variables)

Sup. learning problems ← Classification
Regression problems.

SL → establish relationship between I/p and O/p.

Classification — o/p variable is Discrete Categorical and not a nominal Categorical Variable
(No explicit ordering of classes)

Eg. Male / female opposed to Amount of crop yield prediction
↳ Continuous Variable.

←

Scanned with CamScanner

Grouping of customers based on age, gender, nationality.

- Design of experiments - Data gen.
- Active learning
- Reinforcement learning

Active learning → (some data) → Start with

look @ it to understand type of data required — DoE.

+ Deep learning NN, etc.

Scanned with CamScanner

Lecture 03

Descriptive statistics :- Science of learning from Data

Data — Numbers / text / symbols — represent some information

— 'values' of Qualt / Quant. Variables.

— Num / Quant. (cont / Discrete)

Categ / Qualt (Always Discrete)

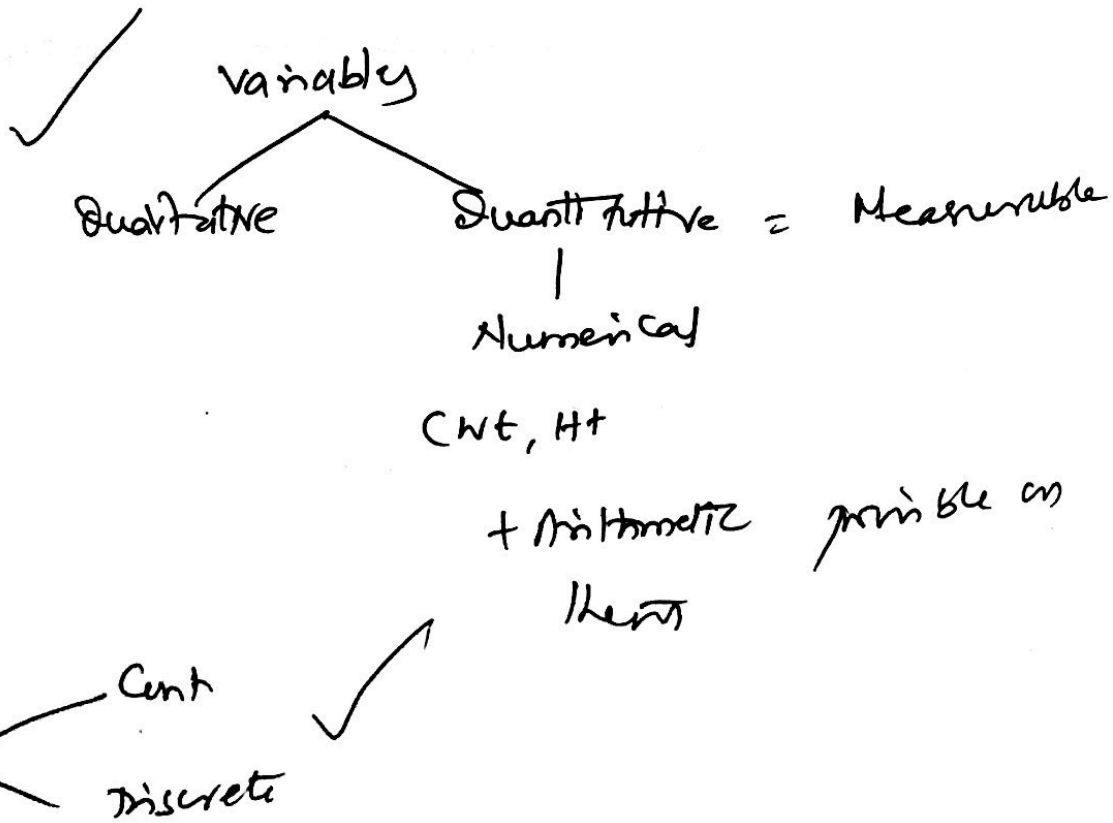
↳ Nominal , Ordinal.

Des. stats — varies accordg to data types

Data — No / text / symbols

↳ frequency.

Data = Values.



Scanned with CamScanner

CV → within Interval any value is possible
→ One Square within a certain interval

1000 students ; Ht (student/s)
(lowest — highest).
eg: (120 — 140 cm)
135 possible, 135.5 also possible.

→ No measuring scale for a part accuracy
"Nothing that prohibits 135.658 for height"

CV → any value between certain interval

No. of uniq people entering IIITDM / day :-
Multiple entries not ✗

No of uniq. ppl per day → variable of interest
for a year → 365 data points / one per day

0 (possible) lower bound ; up bound - eg. 10000
only 0, 1, 2, ...
1.5 not possible.
∴ Discrete variable

Discrete in sense that only integer values are possible ✗

Scanned with CamScanner

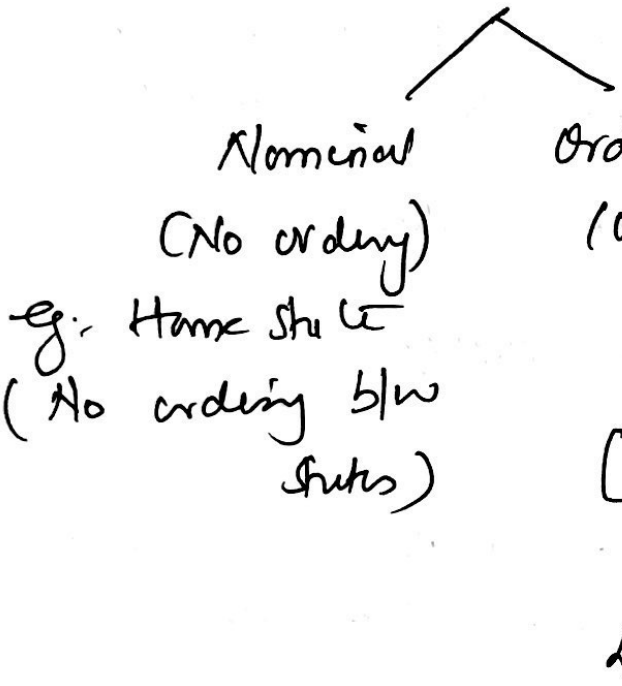
Qualitative :- represents characteristics that can be grouped / characterized

eg: Gender (male / female)

Home town → Any set of names

= diff states of India

Cat. variable (Inherently) discrete



∴ Cat Variables :- No arith operations possible on them

Greater / lesser / Comparison possible with ordinal Variables.

Descriptive Statistics :-

Quantitatively describe data

Means — visual → Graphs / tables

"Crunch the data"

Scanned with CamScanner

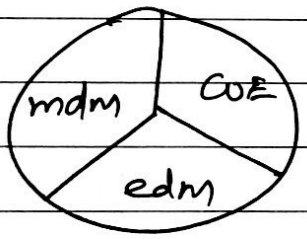
1|2|3|4 types/numbers that quickly summarise the data.

data about data / Metadata → Descriptive stats.

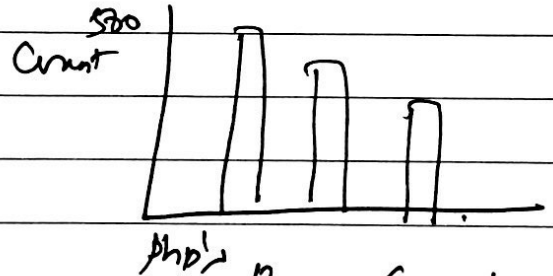
Descriptive → Cannot make conclusions b/yond the data one has
"No Generalisation / Prediction / Influence possible"

Lot of data → Concise Repsn. required for simpler interpretation — graphs, nos, etc.,

Des. Stats can be for multiple variables also.



Pie of Branch St. Demo.



Bar Graph

Describe each variable / also show inter-relationship between variables.

Graph repsn of Sngle Variables (above)

Highest education (cat. variable) 4 values -
(School / Bach / Masters / Doctorate)

Scanned with CamScanner

Highest education v/s Frequency of occurrence

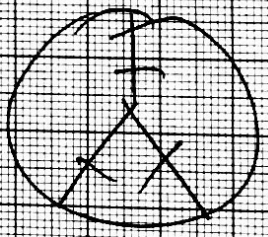
↳ Quick summary of max, least, average, etc.

another view <3

↳ High < Masters < PhD (years)

⇒ Ordinal ?

Categorical variables — another way is pie chart.



Coincidentes

✓ Not suited for ordinal / nominal variables

✓ Move departments — the most beaten

Quant :- Boxplot — Num. data

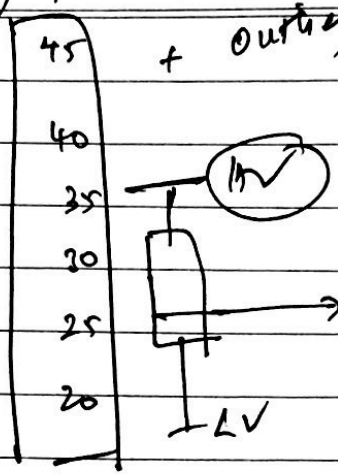
(avg, variance ...)

↳ Central tendency captured (median)

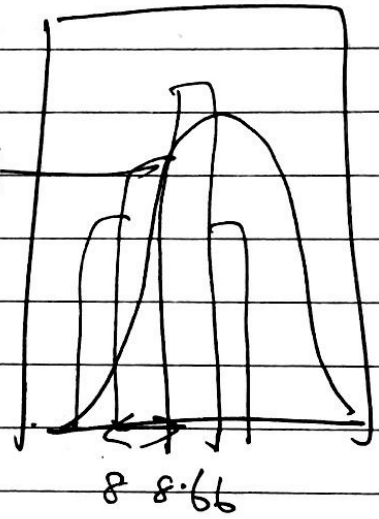
Scanned with CamScanner

This is Repeated Next page from slide

left (window based) content



range of (25 – 33)



6 points.

Histogram: Richest for num. data

x axis – diff values

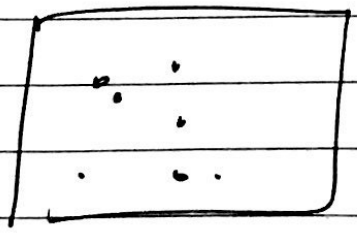
Column – Bins

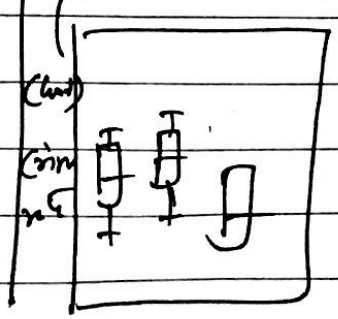
→ empirical Construction

Multiple Variables → Scatter plots (2 Quant) ✓ Both Numeric ✓ Relationship b/w x, y is Captured well.

Box plots, (1 Cat, 1 Quant)

Contingency plots (2 Cat)





MBox plots :-

Scanned with CamScanner

from this page :— Next lecture;

✓ Pie chart — Not suited for ordinal vars.
✓ for Nominal variables

✓ Left out depts / genes !! . . pie is not a good representation.

" pie gen shud represent all values "

2 ways for rep Quant variables :—
Box plot, Histogram

* might be interested in average, variance, summary statistic.

Box plot → Captures & Central Tendency
" line within the box " — median of dataset
← two hinges of dataset → [ ]

↓
" data variability
lower, upper quartile
25% 75% percentile.

✓ Histograms – pick of funcs for Numerical data

? How many data points within a range is answered by Histograms.

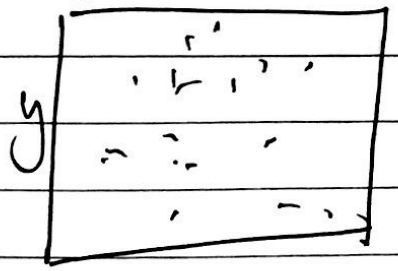
✓ Columns → Bins

Histograms are first step towards → "Empirical Construction"

Graphical Repn for Multiple Variables:-
Scatter / Box plots / Contingency tables.

Sp:- Rep 2 Quantitative Variables
Bp:- One Cat v/s One Quant. Variable
CT:- 2 Cat variables with freq of occurrence as Item.

Sp:- good to capture relationship b/w variables



x ↑ → y ↑.

* Bad @ Understanding each individual Variable.

Extended Box plots:- County v/s Crime Rate
(Cat) (Quant)

✓ Multiple Box plots on the same graph
" Country wide understanding of crime Rate

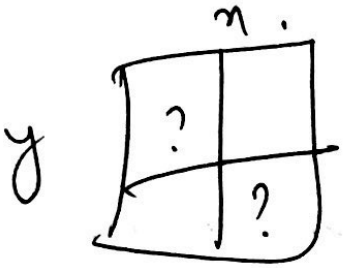
Scanned with CamScanner

⇒ Scatter

Contingency table :- 2 cat variables.

y < MBA-Y / MBA-N

x < Worked as Mgr / No.



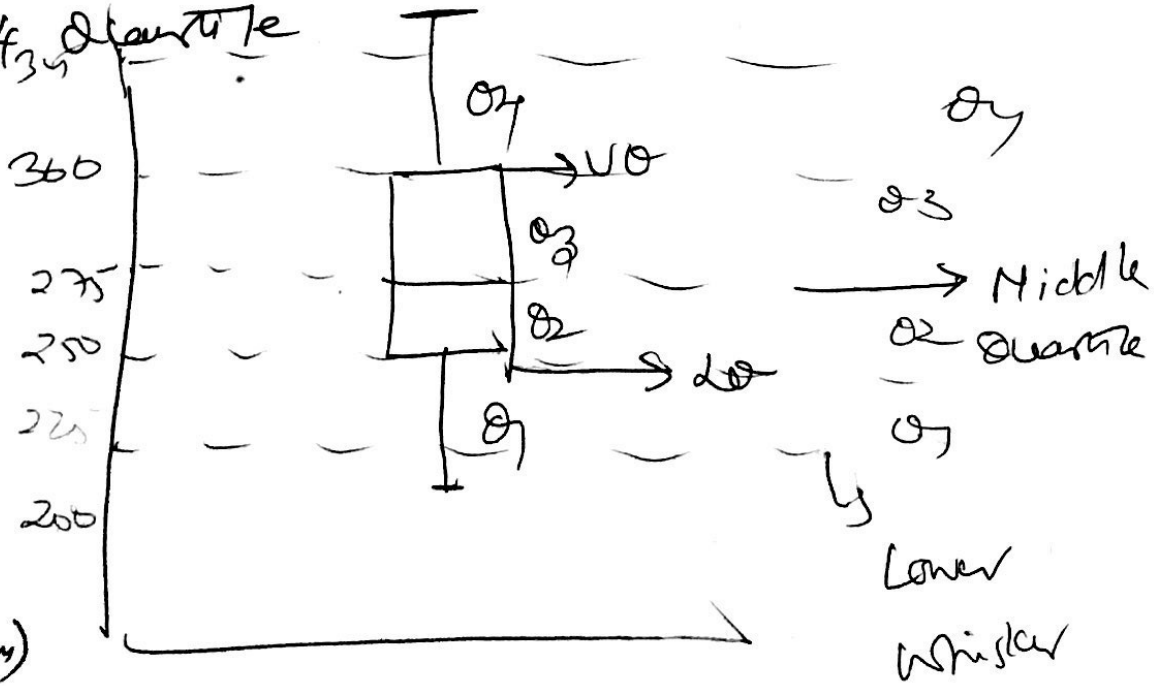
Box plots :- distributional characteristic of group of scores, levels of scores

✓ scores are sorted

✓ four equal sized groups are made from ordered scores.

⇒ 25% scores in each group

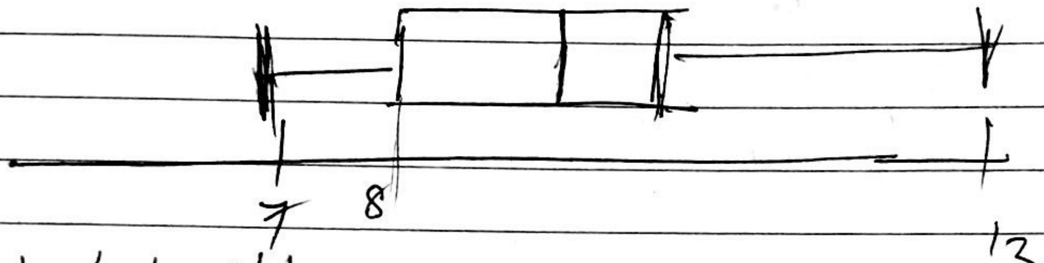
✓ 4 Quantile



Hist → | Bar Charts → Ordinal / Nominal also

Scanned with CamScanner

7 8 8 8 9 9 | 9 9 10 10 13



→ Stem + Leaf plot :

21, 22, 22, 24, 31, 34, 34, 35, 37, 43, | 44, 46, | 48, 49, 49, 52, 55, 55, 56, 67, 68, 68

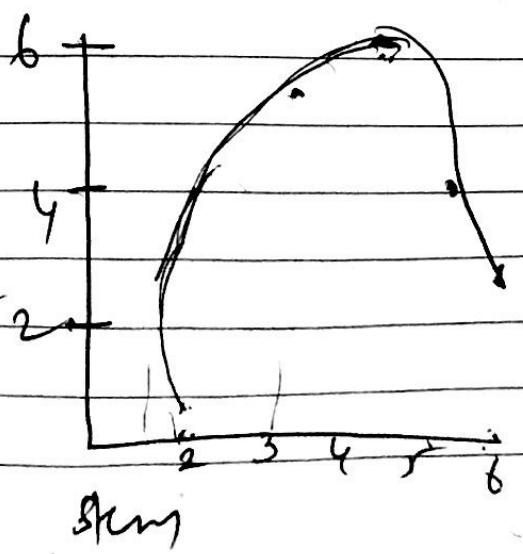
(22)

Stem	leaves
2	1224
3	14457
4	346889
5	2556
6	788

(* mode, median determining

large Datasets with small Data values

Range / spread - known.



→ θ'

All here

⇒ Lecture 04.

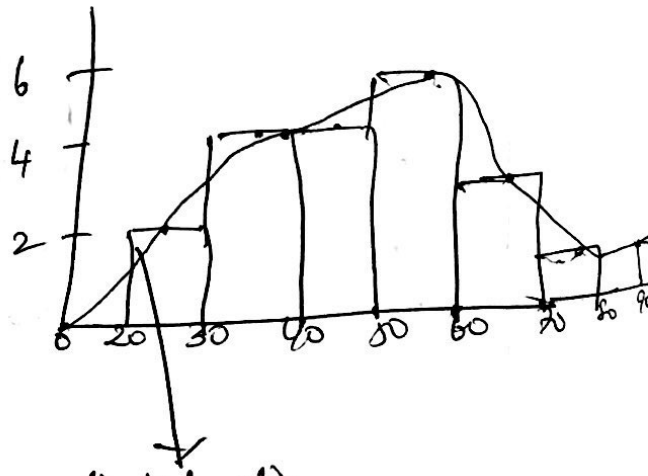
Descriptive statistics — Summary, Central Tendency

✓ Summarising data from numbers.
→ Measures of central tendency
→ Dispersion
→ Skew and kurtosis

✓ Histogram — rich representation; no of values within any range.

✓ Bins vs frequency counts

20-30 → 2	3
4	1
4	0
5	1



distribution

✓ Simpler than distribution?
✓ Central tendency

@ Centre → min → max — in b/w

? pt in histogram ⟶ Convergence to 50% of area

? in central

Scanned with CamScanner

✓ Balancing act ? → x axis is balance and bags are weights. ? fulcrum for balancing.

✓ Measure of central tendency – ? central value.

Measures of dispersion — ? data dispersed around central value. far / close from curve

→ skew | kurtosis

↳ shape of the dispersion distribution.

" distn leans one side / other "

Kurtosis → ? fat are tails of distribution.

Mean = Balancing in seesaw (Centre tendency).

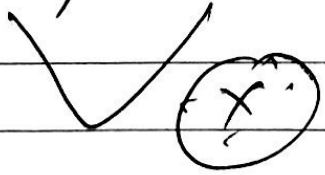
Median — Another central tendency measure.

↓ sorted order — middle / value.

Mode → not freq. occurring

" common range "

? to use Mean / Median / Mode.



outlier – outside majority

↳ reason – error data — Bad outlier.

↳ imp part (story) — Good outlier

Scanned with CamScanner

Mean is impacted by outliers more.

Median more for outliers

Bad outliers → Median is preferred

8, 8, 90.

∴ Median remains same

✓ Mean and really large salaries

→ → outliers.

99% days → low 1 rupee/day.

1% " → 10 crores.

Stock market trading → extreme ups

1000 days

Not confirming to → Bad outliers

Instead of 8 → 80

3, 4, 3, 12, 3, 4, 3, 6, 90 → 4, 90

old Mean = 4.58 →

Median :- 1, 2, 3, 3, 3, (4, 4) 5, 6, 7, 8, (60), 9

Still median unchanged

✓ Good outlier → part (skew) to tail

Lose 1 Rs every day on 99% days
108 (lo come) " 1% "
" outliers

10,000 days :-

Median = . −1, −1, −1, ... 10,000.

Median = −1

Mean = $\frac{-1 + -1 + \cdots + 10000}{1000}$ = >> +ve No.

Mean is a better measure here

Mode :- Most popular / frequent value
for distributions that are symmetric

↳ Can be used for Nominal variables
✓ useful in multimodal distributions
↓
many peaks to distribution

+ Cont. prob distn. with two or more modes

Scanned with CamScanner

0485

Khan Academy

Box plot for odd

(1) 6, 15, 19, 21, 24, 38, 47, 55, 56, 65, 78

$n = 11$ points

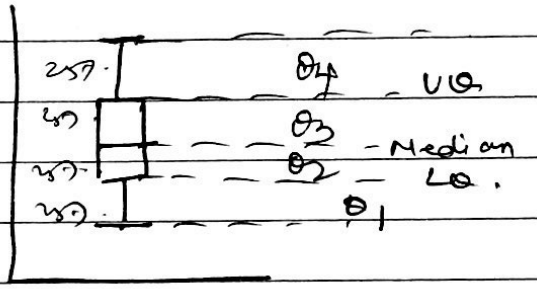
Q_1 ↑ Q_2 (Middle quartile) ↑ Q_3 (upper quartile)

Generic Box:

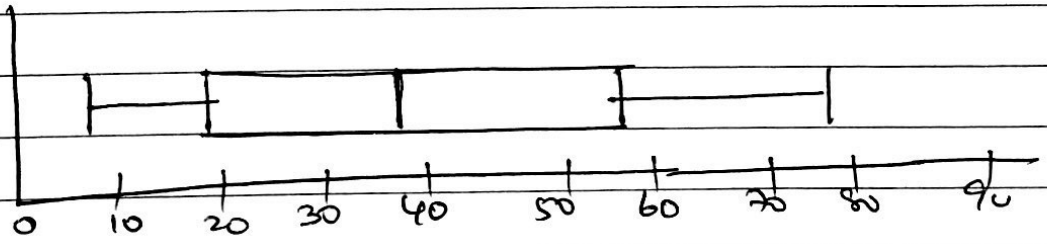
$IQR = Q_3 - Q_1$.

Outlier $< Q_1 - 1.5\,IQR$

$> Q_3 + 1.5\,IQR$

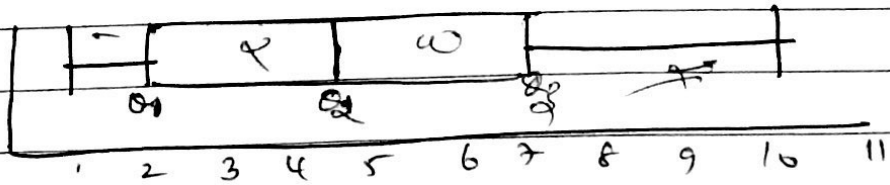


(2)



Ex:- 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 10 (14 pts)

Median = 4.5 (Q_2)

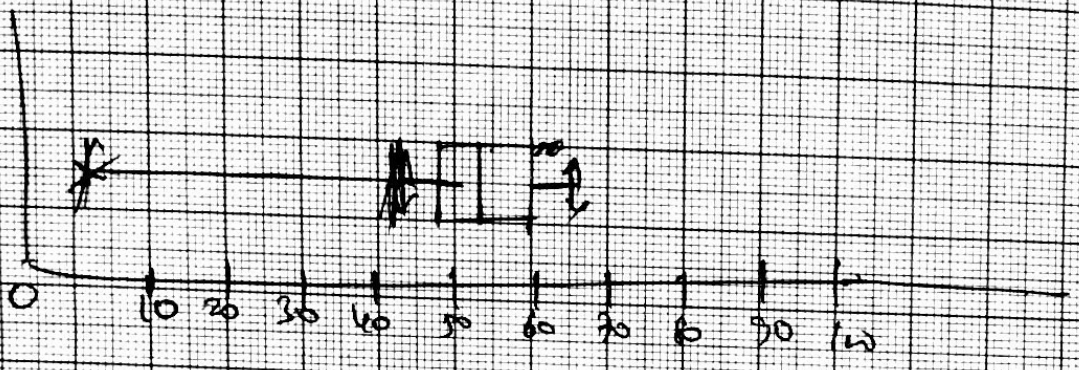


Scanned with CamScanner

collected for outlier

2, 43, 49, 50, 57, 57, 53, 54, 60, 62, 63

Q_1 Q_2 Q_3 (11)



(11)

$Q_1 - 1.5IQR = 49 - 1.5 \times 11$

$= 49 - 16.5$

$= 32.5$

∴ 2 is an outlier ✓

∴ Lower whisker =

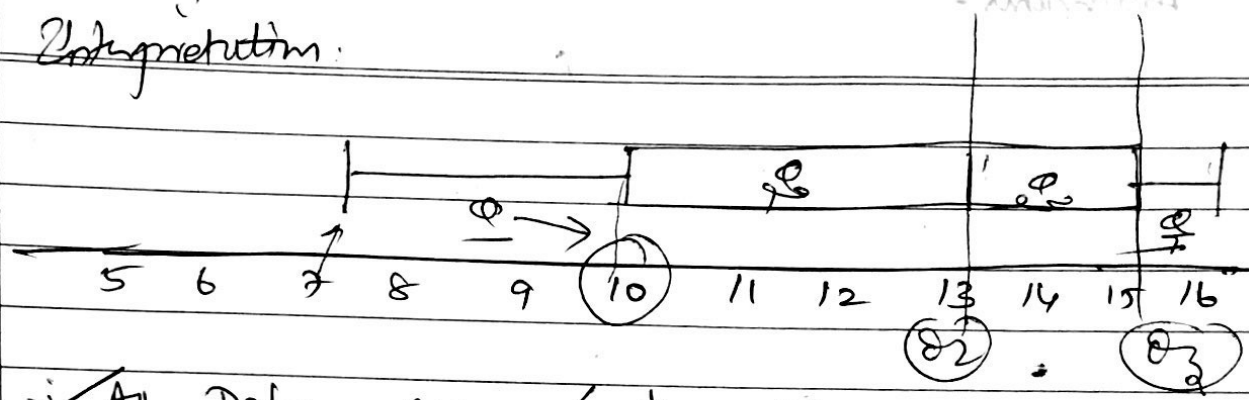
$Q_3 + 1.5 = 60 + 16.5 = 76.5$

Upper whisker is far

Scanned with CamScanner

Interpretation:

(4)



(i) All Data are < than 17

(ii) 75% " > 10 year old

eg: 7 [10 13 15 16] (i) (7 data)

6/7 is >= 10 ⇒ > 75%

eg: 7 9 11 12 14 15 15 16 (ii)

10 13 15

6/8 = 3/4 ≥ 75%

(iii) only 1 person is 17 or 18 year old.

"Can't say" | Multiple 17's / 18's can be there as well.

(iv) exactly 50% are > 13

"can't say" → True with (ii) but not (i).

Scanned with CamScanner

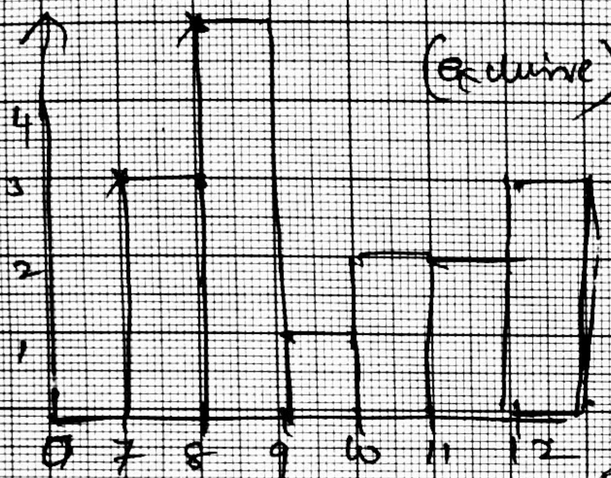
Histograms:-

Khan

① 7, 7, 7, 8, 8, 8, 8, 8, 9,
10, 10, 11, 11, 12, 12, 12

12 − 7 = 5/5 = 1

(Exclusive)



Age of students

5, 7, 5, 9, 7, 7, 6, 9, 9, 9, 10,
12, 12, (7)

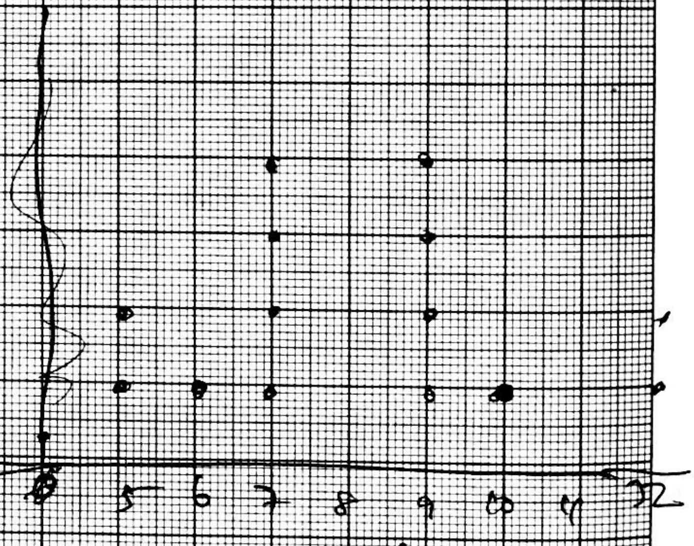
(11)

(i) freq. table.

Age	5	6	7	9	10	12
#	2	1	3	4	1	2

(4) (3)

(ii) Dot plot



②

0→5	5→10	10→20	20→40	40→50	50→60
7	9	12	8	14	89
1	9	6	2	7	3

Af = M.CS / A.CS × freq

= 5/5 × 7 = 7 → (1)

how many 7 9 (3)
(∵(1,1))

"look @ data

freq of highest stack

range of ages?
= max − min
= (17)

5/10 ×12

Ans 1:

1, 3, 27, 32, 5, 63, 26, 25, 28, 16, 4, 40,

29, 19, 22, 57, 58, 9, 42, 6 ✓

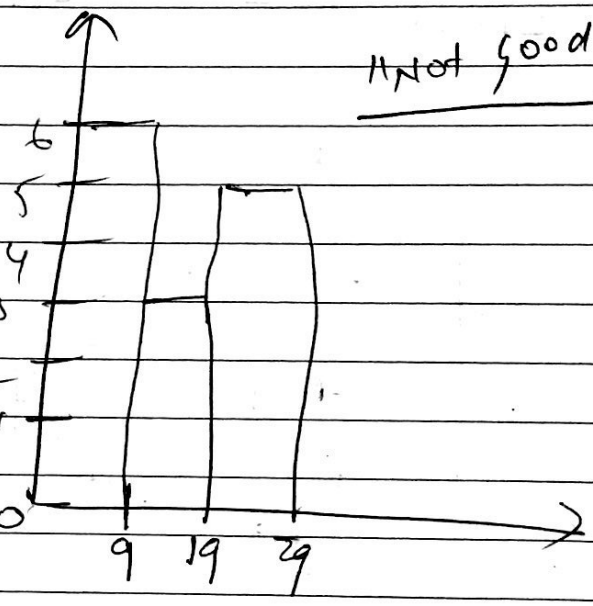
"Bucketing Data"

(Inclusive)

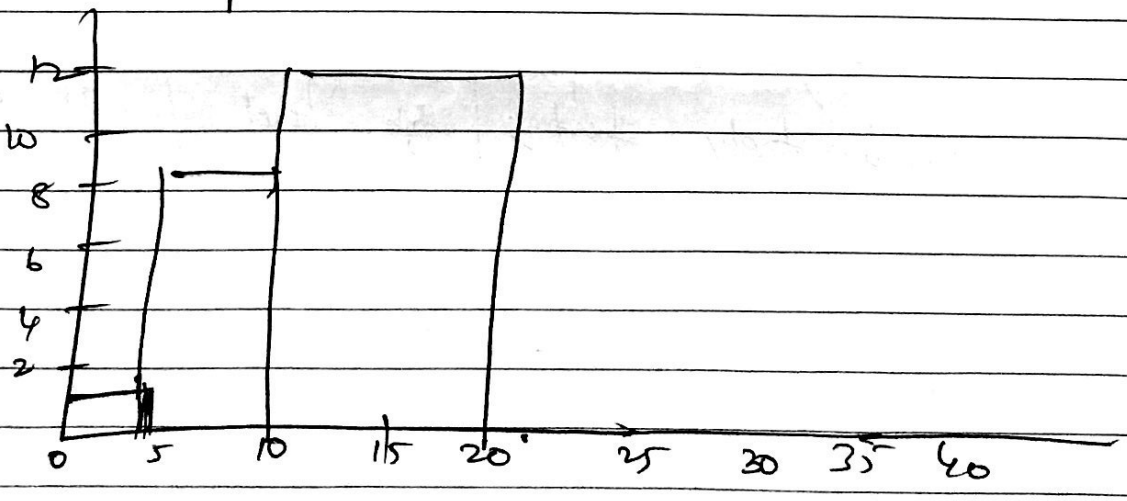
Bucket	Count (#)
0 - 9	6
10 - 19	3
20 - 29	5
30 - 39	1
40 - 49	2
50 - 59	2
60 - 69	1

Visualizing

"Not good"



(2)

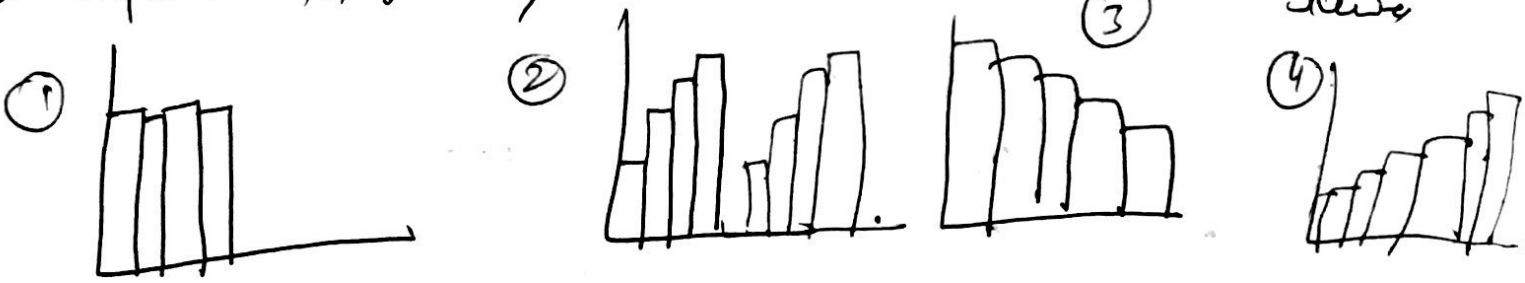


Make it larger byt.

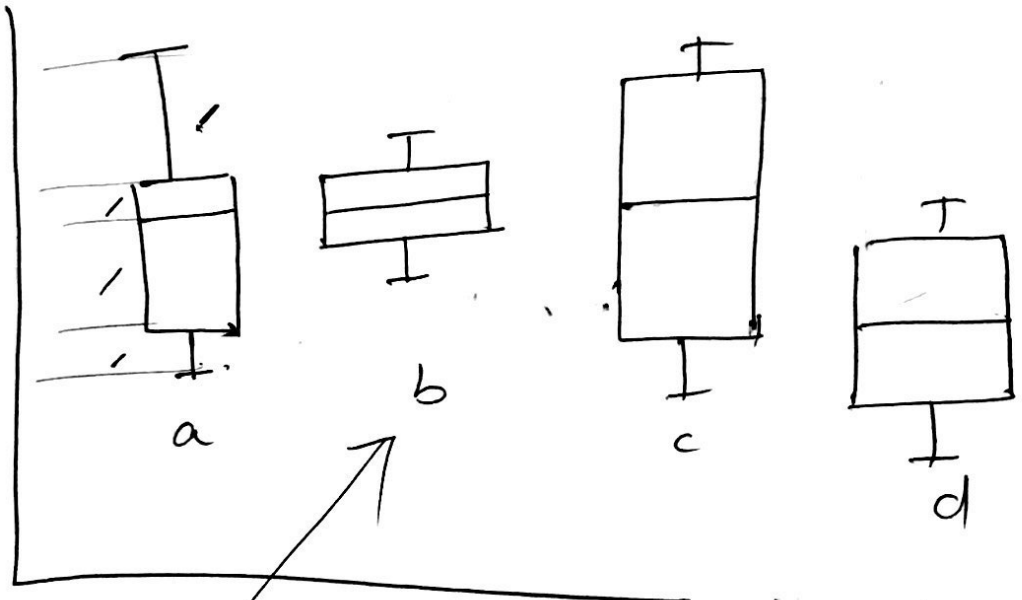
Scanned with CamScanner

✓ Histograms v/s Bar chart (Categorical Var on x axis)
(Numerical) (Discrete)

✓ Uniform Histogram, Bimodal, Right skewed, Left skewed



→ Box plot Interpretation



✓ (b) → Comparatively short; high level of agreement.
Spread out factor is small

(a,c) → Comp. tall; varied views

(c,d) → One Box plot is higher/lower than other

Scanned with CamScanner

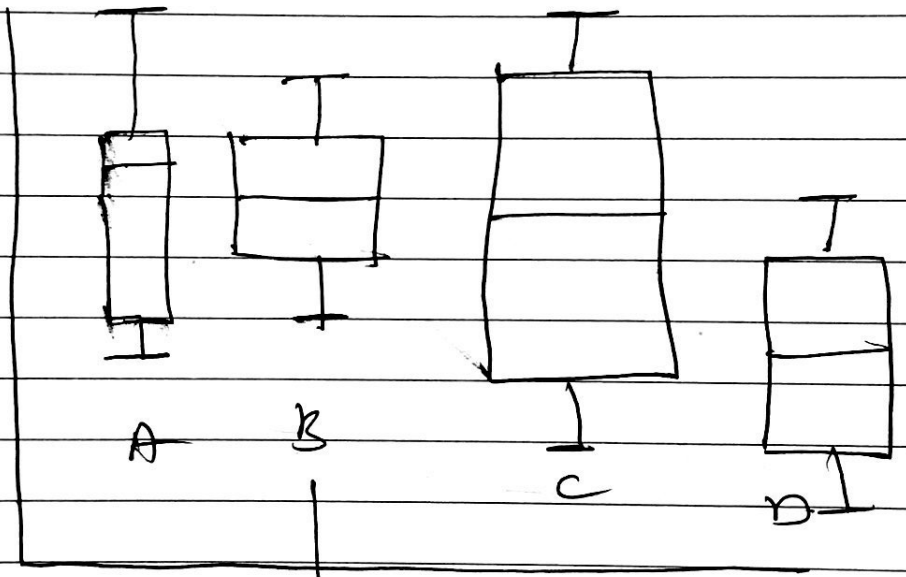
(*) 4 sectors are varied → (a)

→ longer upper - whisker → Maximum Variation in third quartile

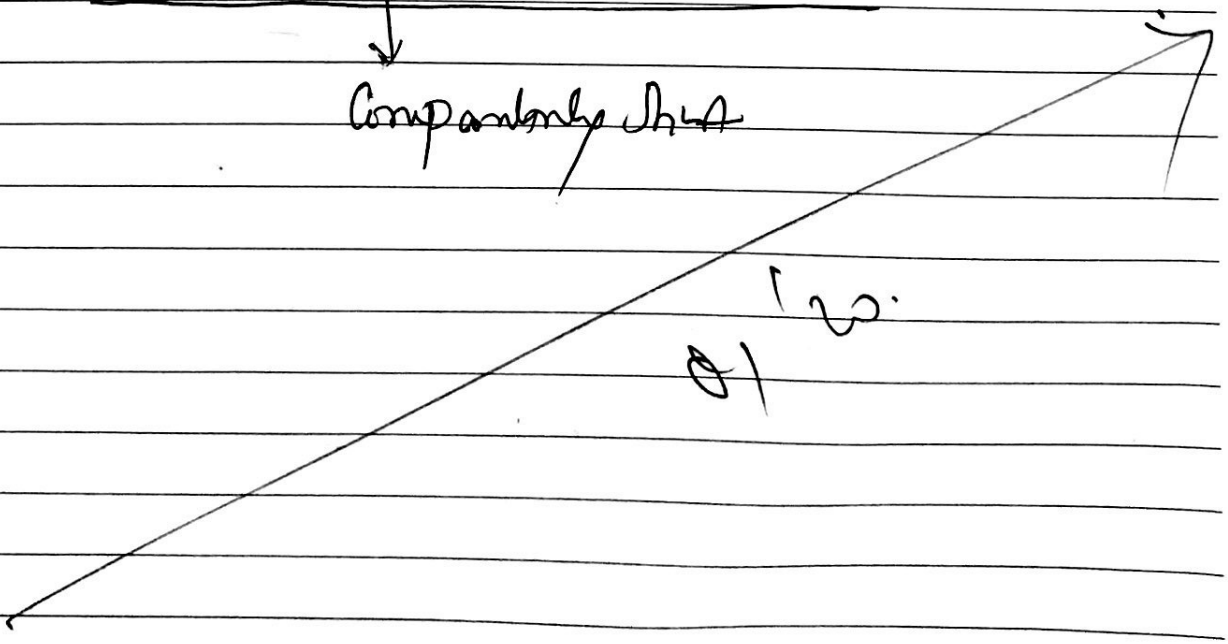
shorter whisker → similarity ↑

(*) Same Median, Different Distributions (a, b, c).

" Nothing significant"



Comparable Data



Scanned with CamScanner