UNIT-1

① Intro to BA : Role of Analytics for Data Driven Decision Making

⇒ Business Analytics means ⇒ "means utilizing tools, techniques and procedure to analyze past business data in order to gain insight and help in decision making & problem solving"

- Data is Everything, Analytics is necessary for survival. Analytics can be used for process improvement (eg. using it to reduce time to deliver order), problem solving (predicting customer cancellation & frauds), decision making (switch to a new market etc.),

② Introduction to concept of Big data Analytics.

Big Data ⇒ refers to extremely large/complex sets of data that cannot be easily managed, procured or analysed with traditional data processing tools.

⇒ The goal of big data analytics is to extract valuable insights, patterns and knowledge.

⇒ Big data is characterized by 3 Vs :-

① Volume (large amount of data), ② Velocity (High speed at which data is generated & processed) ③ Variety (diff. type of data like structured / unstructured)

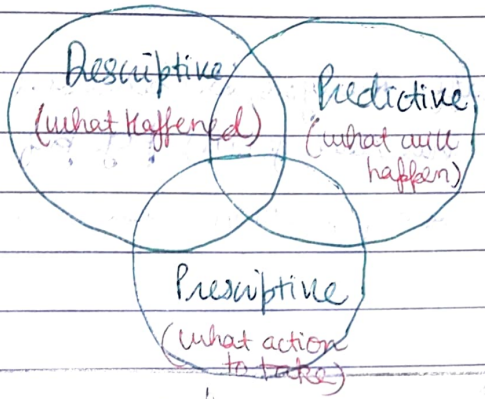
Sources of big data
- Social media data
- Satellite data
- Banking, financial, healthcare data
- Entertainment data
- Research data. Data from sensors, meters etc

① Structured data :- Has clear structure and follows a regular sequence. Organized formated in a way which is easily readable by both Humans & machines. It generally follows tabular structure. eg ① Relational Database entries :- Employee data in company database with columns Name, ID, Department, Salary ② Spreadsheet :- Data in Excel or Similar software like financial data with date, expense etc.

(Relational Database is used to handle complex, large datasets while Spreadsheets are simpler, used for basic data manipulation & smaller datasets and calculations)

② Unstructured data :- lacks formal structure. Doesn't conform to fixed schema and is often human generated.
Example word docs, pdfs, video clips, images, audio recordings etc

③ Types :- Descriptive Analytics, Predictive Analytics and Prescriptive Analytics



Descriptive is Numerical based on - (covered in upcoming pages (one shot))

Descriptive

⇓

Summarizes historical data to understand what happened in the past.

[Tells you what happened in the past]

Eg. You run a ecommerce business, it will involve looking at the past data (sales) and answer ques. like :-

① what were our best selling products last year?

② which months had the highest sales?

Predictive

⇓

use historical data & statistical algorithms to make prediction about future outcomes.

[Take educated guess of what will happen in the future]

Eg. Continuing with ecommerce one, it would forecast future sales or specific trends like sale of certain clothes ↑ during certain seasons.

Prescriptive

⇓

goes a step further and recommends action to optimize or improve outcomes. gives advice/action to achieve desired result

[Specific action to achieve best outcome]

Eg. Predictive analytics told us sale would go up during this period I so prescriptive analytics might recommend ↑ing inventory of those products, running targeted ads or discounts.

Social Media Analytics

⇒ Both play crucial role in digital marketing and online presence strategies of business and organizations. By leveraging insights from these analytics, entities can make informed decision to optimize their online performance & user engagement.

◉ WEB Analytics

⇒ It provides insights into performance of website ⇒ that how users are interacting with their online content. Tools like Google Analytics is used for this.

Key Aspect of this are :- Traffic Analysis, User behavior (Time they spend), Conversion tracking (Sale/Purchase/Leads genrate), Page performance (load time), A/B testing (making two page eg. one page with more images and seeing which perform better).

◉ Social Media Analytics

⇒ It involves collection and analysis of data from social media platforms.

Its Key Aspects include :- Audience insights, Engagement metrics, Sentiment analysis (+ve, -ve, neutral), influencer impact, campaign performance etc.

⇒ Tools used are Google analytics, Facebook/Insta/Twitter/Pinterest analytics, etc.

(5) Overview Machine learning Algorithms and Statistical Software Packages

◉ Machine learning Algorithms

They are set of rules / statistical model that enable computer to learn and make predictions without being explicitly programmed to do so. They learn from data and improve overtime.

Eg. A easy commonly used example of Machine learning Algo is "Spam email filter", which often uses a technique called Naive Bayes classification.

- Example :- Spam Email filter
- Problem :- You recieve lots of spam emails. You want to automatically filter out rather than check each manually.
- Machine Based Algorithm :-
  - ① Training :- give dataset of email labeled as spam or not spam. It learns patterns / probabilities of certain words being used in spam email like offer, free, discount, urgent and calculate probability of it being spam.
  - → This is how machine learning can automate tasks.

◉ How Machine learning Operates :-

⇒ By training model on a dataset, allowing algorithm to identify patterns and rel^{n} b/w data. The trained model can then be used to make decision when presented with new / unseen data.

◉ Statistical Software / Software Packages

⇒ They are the type of computer program designed to perform statistical analysis on data. Helps researchers, scientists, analyst to make sense of data by providing various tools

- Benifits :-
  - ① Helps in Data Analysis :- Analyse trend / pattern / rel^{n} in large
  - ② Visualizations :- Help creating visual representation for more easy intipration of large data.
  - ③ Modelling :- Helps in the creation of statistical models.

Popular Softwares :- SPSS (statistical package for Social Sciences), R Python language, SAS (Statistical Analysis software), MATLAB (Matrix laboratory

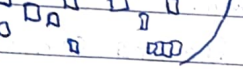
⊙ Descriptive Statistics

1. Data

Based on Data Structure

Structured Data
(Data in a matrix labelled as rows & columns)

Unstructured Data
(Data not in matrix form, no fixed structure)



Based on Scale of Measurement (Type of Data)

→ **Cross sectional data**
- Taking picture at one moment of time (screenshot)

Eg: Survey 100 people on street today and ask their fav. colour. It gives you picture of people's preference at that particular moment.

→ **Panel Data** (aka longitudinal data)
- Taking picture of the same grp over a period, allowing to observe changes within that grp.

Eg: Tracking salaries made by your friend in their 20s, 30s, 40s, 50s, 60s etc.

(Panel :- over a period
Time series :- regular interval)

→ **Time Series Data** :- Taking pics of same thing at regular time intervals like every 10 days to see how it evolves/ fluctuate Eg: tracking stock price of a company for every 15 days.

2. ★ Types of Data Measurement Scales ★

Nominal
- In it data refers to names which can be given a numerical value to represent them (no necessary means rank/order)

Eg: Single – 1
Married – 2
Children (with) – 3

(1, 2, 3 are not ranking/ order just for easy interpretation)

Ordinal
- In it data has clear order or ranking but interval b/w categories may not be consistent.

Eg: Student performance
1st } while there is an
2nd } order that 1st is
3rd } better than 2nd
but the gap b/w 1st and 2nd & 2nd and 3rd may differ a lot.

Interval
- In it data has clear order and consistent interval b/w them.

Eg: Difference in temp b/w 20°C and 30°C will be same as difference b/w 30°C and 40°C.

★ But here in interval scale there is no true zero point. Eg: 0°C doesn't mean there is no temp. (= it mean less temp.)

Ratio
- It has clear order, consistent interval and a true zero point

Eg: A person can have 0 income too.

3. Population and Sample

⊙ **Population** :- The entire grp you are interested in studying.

⊙ **Sample** :- Smaller grp, selected from population to represent & make inferences about the larger grp.

Eg: You want to find avg. height in college. You cannot ask 2000 students ⇒ You select 10 students from each course and infer about from it about all.

4. Measures of Central Tendency

★ Mean (Average)

⇒ It is the arithmetical average value of data

$$\text{Mean} = \bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

If, with frequency;

$$\text{mean} = \sum_{i=1}^{n} \frac{f_i x_i}{n}$$

★ ⇒ Imp. property of mean is that summation of deviation from observation from mean is zero.

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

★ "Median" is the thing in descriptive stats which is not calculated based on the entire data.

★ Median

⇒ Divides data into two equal halves. Portion of observation below & above median will be 50%.

Individual Series

⇒ Arrange the data, find middle most value.

Discrete Series :-

Odd no. of observations,

$$\left(\frac{n+1}{2}\right)^{th} \text{ term} = \text{Median}$$

Even no. of observation,

$$\frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2} = \text{will give a value} = \text{Medi}$$

Continous Series

⇒

$$\text{Median} = l + \left(\frac{\frac{N}{2} - C.F_0}{f}\right) \times h$$

◉ Calculate CF.

- ★ l = lower limit of median class
- ★ f = frequency of median class
- ★ CF = CF of class above median class
- ★ h = class width.

⇒ Median class is the class which has the highest frequency.

★ Mode

⇒ Mode is the frequently occuring value in dataset.

★ ⇒ Mode is the only measure of central tendency which is valid for Nominal (qualitative data) since, the mean and median for nominal data is meaningless.

⇒ Eg. There is a data of marital status of customer namely (a) Married (b) Single (c) with children (d) without children. Mean, Median are meaningless here, but mode can tell what is the most occuring customer type.

Individual / Discrete Series

⇒ Series with highest frequency.

Continous series

⇒

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

★ Modal class :- class interval with highest frequency.

l = lower limit modal class
f_0 = frequency of class above modal class
f_1 = frequency of modal class
f_2 = frequency of class below modal class.
h = class width

Mode = 3 Median − 2 Mean

(5.) Percentile, Quartile and Decile

⇒ Percentile, Quartile and Decile are used to tell the position of the observation (and its value) in the dataset.

and from it we can tell value.

⊙ Percentile score can be used in identifying pos^n of a student in class and in asset management.

⇒ P_x = ◯ ← gives value of data at which x% of data is below that value.

⊙ Eg P_{10} = value below which 10% of data lies.

$P_{x\%} = \left[\frac{x(n+1)}{100}\right]^{th}$ obs here, n = no. of observation in data.

Percentiles

- Quartiles
  - ⊙ Special type of percentile which divide data in 4 equal parts.
  - Eg. $Q_1 = P_{25}$ = Contains first 25% of data
  - ★ $Q_2 = P_{50}$ = " " 50% " " = Median
  - $Q_3 = P_{75}$ = " " 75% " "
  - $Q_4 = P_{100}$ = Contains 100% of data
- Deciles
  - ⊙ Special type of percentile which divide into 10 equal parts.
  - Eg. P_{10} = first decile = P_{10} = contains 10% of data
  - P_{20} = 2nd " = P_{20} = contains 20% of data etc.

Time of the failures of wire cut (in hrs)

2	22	39	56	76
3	24	44	66	77
5	24	45	67	78
9	26	45	75	78
21	31	46	75	78

Eg. ⇒ Suppose this a a data.
(1) Company likes to know by what time 10% of the and 90% of the wire are cut.

Ans. $n = 25$ here, $P_{10} = \frac{10(25+1)}{100} = 2.6^{th}$ observation

⇓

Value at 2nd pos^n = 3 hr

so, value at approx $(2.6)^{th}$ ~~will be~~ obs will be :-

~~5 hrs~~ = value at 2nd + 0.6th obs.
= 3hr + 0.6 (value at 3rd pos − value at 2nd)
= 3 + 0.6 (5−3)
= 3 + 1.2 = 4.2hr Ans

while, $P_{90} = \frac{90(25+1)}{100} = \frac{26\times 9}{10} = \frac{234}{10} = 23.4^{th}$ obs

value at 23.4^{th} obs = 23rd obs + 0.4^{th} obs
= 78 + 0.4th (value at 24th obs − 23rd obs)
= 78 + 0.4 (78−78) = 78 hr. Ans

(6.) Measures of Variation

- Range
- IQD
- Variance
- Standard Deviation

⊙ Range

⇒ Difference b/w highest and lowest value in the set. It tells how much data is spread.

Eg. [70, 85, 92, 60, 75] (Score of students)

• Highest = 92, • lowest = 60 ⇒ Range = 92 − 60 = 32

This mean the score spread across a range of 32 poi

⊙ Inter-Quartile Distance (IQD) [also k/a IQR (inter Quartile Range)]

⇒ It is the measure of distance b/w Q_1 and Q_3.

Eg. Dataset = [10, 15, 20, 25, 30, 35, 40]

Q_2 = 25 (median)

⇒ lower half of data = [10, 15, 20] = Q_1 = 15

upper half of data = [30, 35, 40] = Q_3 = 35

IQD = IQR = 35 − 15 = 20

↑ it means middle 50% of data has a range of 20 units.

⇒ IQR is used to identify outliners in data.

↓

means, those values in dataset which differ significantly and their presence can cause incorrect conclusion.

⇒ Formula to identify outliner using IQR :-

- lower Bound = $Q_1 - 1.5 \times IQR$ (Here −)
- Upper Bound = $Q_3 + 1.5 \times IQR$ (Here +)

Eg. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 90]

$Q_1 = 3$, $Q_3 = 9$ ⇒ IQR = 9 − 3 = 6

outliner value = $Q_1 - 1.5$ IQR = 3 − 1.5(6) = −6

outliner value = $Q_3 + 1.5$ IQR = 9 + 1.5(6) = 9 + 9.0 = 18

Here, in our dataset there is a value which is higher than '18' (upper bound) outliner = "90" so, 90 is a outliner.

V. Imp. FORMULA Trick ★★

Variance = σ^2 =

⊙ Variance and Standard deviation

Note, $\bar{x}$ = Sample mean, μ = Population mean

⊙ Variance = $\sigma^2 = \sum_{i=1}^{n} \frac{\Sigma (x-\mu)^2}{N}$ [S.D. = $\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x-\mu)^2}{N}}$]

(↑ In case of population variance)

⊙ Sample Variance = $s^2 = \sum_{i=1}^{n} \frac{\Sigma (x-\bar{x})^2}{N-1}$ [SD for sample $s = \sqrt{\sum_{i=1}^{n} \frac{(x-\bar{x})^2}{N-1}}$]

(↑ k/a sample variance, we divide by n−1 in case we are calculating variance for sample. This is k/a "Bassels correction". It is denoted by s^2.)

[If, we divide for calculating sample by "n" only then, it will give less value which is incorrect. ⇒ this is also k/a downward bias. To save from it we divide by n−1.]

⊙ Degree of freedom

Eg. You want to create a sample of 4 numbers with mean = 10

⊙ 1st choice (free choice) = 8
2nd choice (, , , 4) = 12
3rd choice (4, , ,) = 9
} Then sum = 29

⇒ Now, for mean to be 10, sum of 4 no. should be 40 then only $\frac{40}{4} = 10$. So, 4th no. should be 40 − 29 = 11 ✓

⇒ So, we didn't have the freedom to choose this number. So,

Degree of freedom = N − 1

7. Chebyshev's Theorem

⇒ Chebyshev Theorem / Inequality tells that how much of your data lies b/w certain interval defined by mean & standard deviation.

⇒ Probability of finding a random selected value in an interval defined by $\mu \pm k\sigma$ is $1 - \frac{1}{k^2}$ i.e.,

$$P(\mu - k\sigma \leq x \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

(P = probability, μ = mean, X = value in data set, σ = S.D., K = no. of SD from mean)

Q. Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate portion of customers who are spending b/w 8000 and 16000.

Ans. Given, mean = μ = 12000

S.D. = 2000

Now, $\mu \pm k\sigma \Rightarrow 12000 \pm k(2000)$

$= 8000$ → from here $k = 2$

$= 16000$ → from here $k = 2$

So, Now,

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P(8000 \leq x < 16000) = 1 - \frac{1}{2^2} = 0.75$$

→ That is the portion of customers spending b/w 8000 and 16000 is at least 0.75 or 75%.

8. Skewness and Kurtosis

★ Skewness

⇒ Skewness is the measure of asymmetry of probability distribution of a real valued random variable. In other words skewness tells us, the portion of data b/w μ and $\mu - k\sigma$ is same as μ and $\mu + k\sigma$ (k = some +ve constant).

It tells us the extent to which distribution leans to the left or right of mean.

$$g_1 = \frac{\sum_{i=1}^{n} (x_i - \mu)^3}{n} \Big/ \sigma^3$$

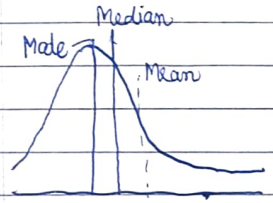
Here, x_i = each value in dataset, μ = mean of dataset, n = no. of terms, σ = S.D.

If,

g_1 = +ve = +ve skewness = tail longer in right side

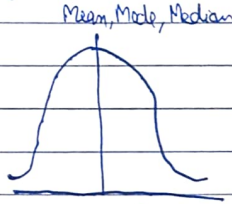
g_1 = -ve = -ve " = tail " " left side

g_1 = close to 0 = symmetrical distribution.



+ve Skewness

(Mean > Median > Mode)



Symmetrical

(Mean = Median = Mode)

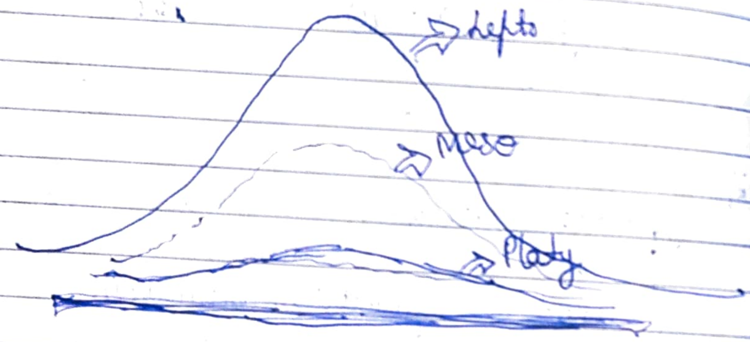
(Mean < Median < Mode)

★ Kurtosis

Kurtosis tells us the peakness where as skewness told where the tail is झुका हुआ.

$$(K)\ \text{Kurtosis} = \frac{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n}}{\sigma^4}$$

If,
- $K < 3 \Rightarrow$ Platykurtic
- $K = 3 \Rightarrow$ Mesokurtic
- $K > 3 \Rightarrow$ Leptokurtic



⊙ Graphs (Data Visualisation)

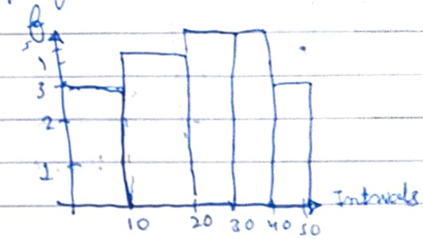
① Histogram ("Frequency polygon and Frequency curve")

⇒ If you are not given data in frequency form, make in frequency.

Eg. 1, 3, 27, 32, 5, 63, 49, 50, 20, 30, 46, 50, 27, 29, 30, 32.

Now draw a frequency:-

Interval/Bucket/Bin	f
0-10	3
10-20	4
20-30	5
30-40	5
40-50	3

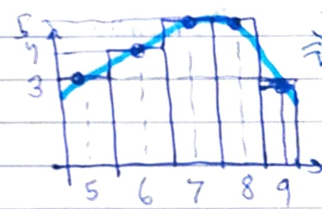


⇓

If we take mid values

0-10	5	3
10-20	6	4
20-30	7	5
30-40	8	5
40-50	9	3

⇒



⇒ draw with free hand
↳ K/a Frequency curve

⇒ draw st lines through mid points of pillars
↳ K/a " Frequency poly

[Make sure to extend freq. polygon to make it closed since polygons are closed]

★ Ogive Curve ★

⇒ curve made/represent cummulative frequency (C.F.)

⇒ It can be less than or more than type

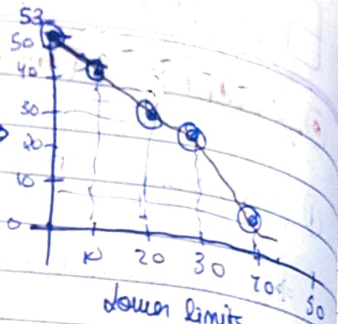
⇒ When they both meet, the intersection point is median. Also we can calculate median from each one separately also.

Eg

	f
0-10	3
10-20	27
20-30	15
30-40	3
40-50	5
	53

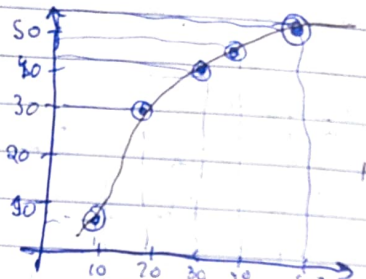
more than

more than	CF
more than 0	53
more than 10	48
" " 20	45
" " 30	30
" " 40	3



less than

less than	CF.
less than 10	3
20	30
30	45
40	48
50	53



⇒ Now, $N = \Sigma f = 53$

Median $= \frac{53+1}{2}^{th} = \frac{54}{2} = 27^{th}$ observation

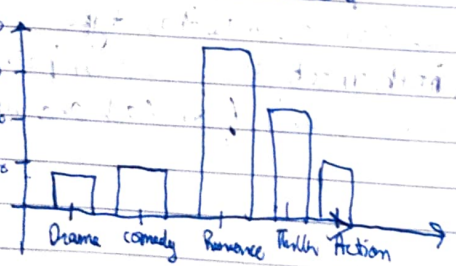
On Y-axis locate 27th value and corresponding value of X will be the value of 27th observation.

② Bar Graphs

↳ used in case of Nominal data where names are associated with Numerical value.

Eg.

Category	People
Drama	20
Comedy	30
Romance	100
Thriller	80
Action	40



③ Pie Chart

Eg

Subject	Frequency	Degrees given in chart
Science	6	72°
P.E.	4	48°
Maths	9	108°
English	7	84°
History	4	48°
Total =	30	360° ✓

×12°

Now, $\frac{360}{30} = 12°$

① Draw a circle

②  draw st. line from centre. Put '0' on it and start marking angles using the outer portion of 0.

③  Ans.

④ 5 Point Summary & Box-Plot (or Box & whisker Plot)

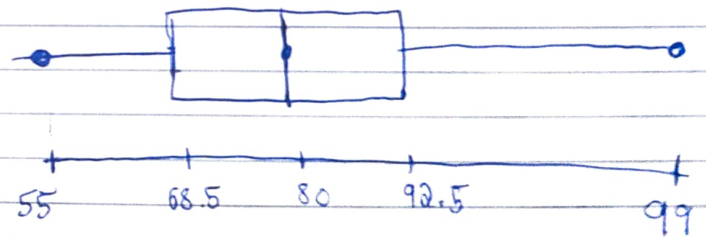
Concise description of dataset which gives 5 things:-

① Minimum Value
② Q_1
③ Q_2 or median
④ Q_3
⑤ Maximum Value

⇒ graphical representation of 5 point summary is known as Box plot.

Eg Data = [55, 66, 71, 75, 80, 85, 90, 95, 99]

- Min = 55
- $Q_1 = P_{25} = \frac{25(n+1)}{100} = \frac{9}{4}(9.5)^{th} = 2^{th} + 0.5(3^{rd} - 2^{nd}) = 66 + 0.5(71-66) = 66 + 2.5 = 68.5$
- $Q_2 = 80$
- $Q_3 = P_{75} = \frac{75(9+1)}{100} = \frac{3}{4}(7.5)^{th} = 7^{th} + 0.5^{th} = 90 + 0.5(95-90) = 92.5$
- Max = 99



★ Box Range from Q_1 to Q_3. Lines from max & min and median is marked inside box. [In case of outliers, they are displayed as point beyond]