Things to Remember

I - What is statistics

- By Statistics, we mean methods specially adapted to the elucidation of quantitative data affected to a marked extent by a multiplicity of causes".

  Yule and Kendal

- Difference between descriptive and inferential statistics

Basis of comparison	Descriptive Statistics	Inferential Statistics
Meaning	Descriptive statistics seeks to describe the data, but do not attempt to make inferences from the sample to the whole population	Inferential statistics deals with making inferences about a population from a sample
What it does ?	Summarize, organize and present the data in a meaningful way	Conclusion and prediction of data

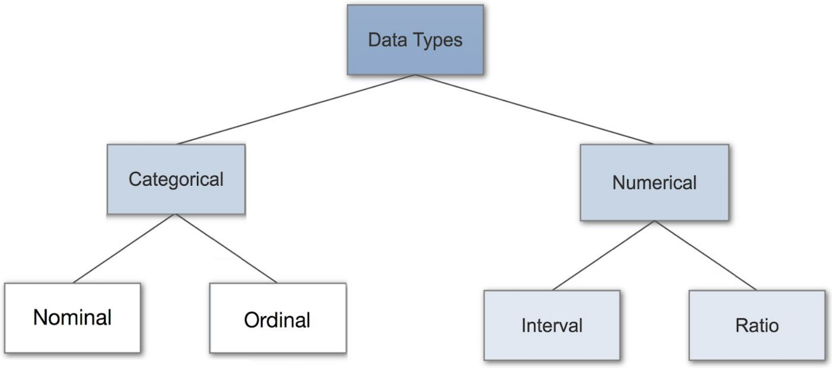
II - Data

Data Vs Information - When analysts are bewildered by plethora of data, which do not make any sense on the surface of it, they are looking for methods to classify data that would convey meaning. The idea here is to help them draw the right conclusion. Data needs to be arranged into information.

Raw Data - Raw Data represent numbers and facts in the original format in which the data have been collected. We need to convert the raw data into information for decision making.

Types of Data:

It is very important to have a good understanding of the different data types, also called measurement scales, is a crucial prerequisite for doing Exploratory Data Analysis (EDA)



Types of Data		
Categorical Data (represents characteristics, also called as qualitative data)	Nominal Data	Ordinal Data
	This data represents discrete units and use to label variables that have no quantitative value . Nominal data has no order	Ordinal data represent discrete and ordered units.Order is important in case of this data
Example	What is your gender ● Male ● Female What programming languages you know : ● Python ● R ● SAS	Winners in Hackathon ● First ● Second ● Third Proficiency in programming ● High ● Medium ● Low
Visualization methods	**Bar chart and Pie chart**	
Numerical Data	Discrete Data Data which can't be measured but can be counted. Data can take on only certain values	Continuous Data Data which can be measured but can't be counted Two types of Continuous data ● Interval Data - Ordered units have the same difference. But it has no true zero points. ● Ratio Data -same as

		interval values, with the difference that they do have an absolute zero
Example	Team members in a cricket team- It can be 11 but not 11.5	Interval Data -Temperature of a particular place ● -10 ● -5 ● 0 ● 5 ● 10 (here 0 has no true meaning) Ratio Data - Equal difference ● 0 ● 5 ● 10 ● 15
Visualization Technique	Boxplots and Histogram	

III - Measures of Central Tendency

Measures of Central Tendency			
	Mean	**Median**	**Mode**
Meaning	The mean is simply the average and considered the most reliable measure of central tendency. The mean is computed by the sum of all values, divided by the number of values.	The median is the “middle” value or midpoint in your data	The mode is the value or category that occurs most often within the data
Example	Uber Rating - After every ride, you give a rating for your experience and final rating which comes for the driver is calculated	With 10,000 people, the mean salary might be $45,000, but the range is $20,000 to $3,000,000 with a mean of $100,000. Mean is	Which is the most popular video on youtube? How will you find out? - Ans - The one which has the maximum likes

	using mean	affected by extreme values. In order to get a real figure in cases where we have outliers in data median is calculated	

III - Measures of Dispersion

Meaning - refers to the idea of variability within your data. It answers unambiguously the question "What is the magnitude of departure from the average value for different groups having identical averages?".

Different types of measures of dispersion

1) **Range** is the simplest of all measures of dispersion. It is calculated as the difference between the maximum and minimum value in the data set.

Range =Largest Value − Lowest Value

The range is also the most affected by outliers as it uses only the extreme values.It is advisable to use range only for very small distributions with no outliers

2) **Interquartile Range** is the distance between the lower and upper quartiles of a data.

IQR = Q3 - Q1

IQR is considered a good measure of variation in skewed datasets as it is resistant to outliers.

3) **Standard deviation i**s a measure of how much data values deviate away from the mean. Larger the standard deviation, the greater the amount of variation.

SD = $\sqrt{\ }$ Σ (Data value - arithmetic mean)^2 / Total number of values in the dataset

S**tandard deviation** is a good measure of variability for normal distributions or distributions that aren't extremely skewed

4) **Coefficient of variation** is equal to the standard deviation divided by the mean. It is a useful measure for comparing the variability between two different datasets. ***For eg. if we need to compare the sales of Apple mobile phones between India and the US, the***

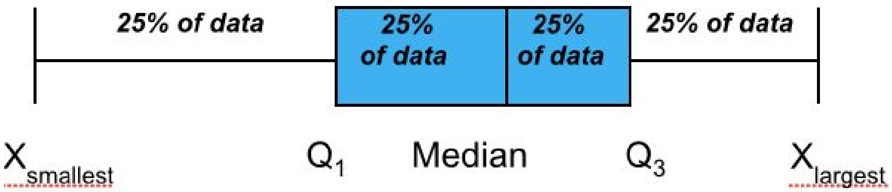
coefficient of variation would be used as it's a relative measure free of units of measurement.

Standard deviation will not be useful as sales in India would be given in INR and for US in dollars and won’t give any meaningful result,therefore coefficient of variation is used and is also called as relative standard deviation

IV - Boxplot

Boxplot is five numbers that help describe the centre, spread and shape of data are:

- ^{x}smallest
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- ^{x}largest



v**) Skewness -** It refers to a lack of symmetry. Skewness results in inequality in the values of mean, median and mode and lower and upper quartiles are not situated at equal distance from median.

- Skewness may be positive or negative
- In case of positive skewness for a distribution
  - Mean > Median > Mode
  - (Q_3 - Median) > (Median - Q_1)

- In case of negative skewness for a distribution
  - Mean < Median < Mode
  - (Q_3 - median) < (median - Q_1)

Relationships among the five-number summary and distribution shape

Left-Skewed	***Symmetric***	***Right-Skewed***

$Median – X_{smallest}$ $>$ $X_{largest} – Median$	$Median – X_{smallest}$ $\approx$ $X_{largest} – Median$	$Median – X_{smallest}$ $<$ $X_{largest} – Median$
${}^{Q}1\ {}^{-X}smallest$ $>$ ${}^{X}largest\ {}^{-Q}3$	${}^{Q}1\ {}^{-X}smallest$ $\approx$ ${}^{X}largest\ {}^{-Q}3$	${}^{Q}1\ {}^{-X}smallest$ $<$ ${}^{X}largest\ {}^{-Q}3$
$Median – Q_1$ $>$ $Q_3 – Median$	$Median – Q_1$ $\approx$ $Q_3 – Median$	$Median – Q_1$ $<$ $Q_3 – Median$