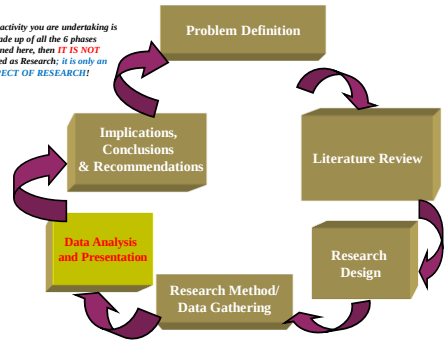
Data Analysis and Presentation

The Research Cycle/Phases

N/B: If the activity you are undertaking is not made up of all the 6 phases mentioned here, then IT IS NOT considered as Research; it is only an ASPECT OF RESEARCH!



A. Data Analysis

- Data analysis is the **categorizing, ordering, manipulating, and summarizing data** to obtain answers to research questions.
- **Purpose**: to obtain meaning from the collected data.

WHY DO WE ANALYZE DATA

- The purpose of analysing data is to obtain usable and useful information. The analysis, irrespective of whether the data is qualitative or quantitative, may:

1. describe and summarise the data
2. identify relationships between variables
3. compare variables
4. identify the difference between variables
5. forecast outcomes

B. Data Interpretation

- Is the process of attaching meaning to data…

Numbers do not speak for themselves.

For example, what does it mean that 55 youth reported a change in behavior. Or, 25% of participants rated the program a 5 and 75% rated it a 4. What do these numbers mean?

Interpretation is the process of attaching meaning to the data.

C. Data Presentation

- Data can be presented in various forms depending on the type of data collected.
- Some examples of Data Presentation Include:

1. Frequency Distribution.
2. Graphical Representations e.g. charts, graphs etc.
3. Descriptive Measures- e.g. mean, median, range, variance, standard deviation.

1. FREQUENCY DISTRIBUTION

- A frequency distribution is a table showing how often each value (or set of values) of the variable in question occurs in a data set.
- A frequency table is used to summarize categorical or numerical data. Frequencies are also presented as relative frequencies, that is, the percentage of the total number in the sample.

EXAMPLE: Frequency distribution of number of cups of coffee		
No.	Frequency	Percent
0-3	24	30.0
4-7	50	62.5
8-11	6	7.5
TOTAL	80	100

2. GRAPHICAL REPRESENTATIONS

- There are different types of graphs and charts that you can use to represent your research data.
- A good graph or chart can show as much as several paragraphs of words. But how do you choose which style of graph to use?
- The following slides show the ***common types of statistical graphs and charts (and their meanings/uses***) widely used in any research.

Remember to ALWAYS LABEL your graphs and charts with an appropriate:
TITLE, X-axis, Y-axis

a.)Line Graphs

- A line chart graphically displays data that changes continuously over time.
- Uses of line graphs:
  - i. When you want ***to show trends***. For example, how house prices have increased over time.
  - ii. When you want ***to make predictions based on a data histor*** over time.
  - iii. When ***comparing two or more different variables, situations, an*** ***information over a given period o time***.
- Line graphs have ***an x-axis and a y-axis***.
- In most cases, ***time is distributed on the horizontal axis.***

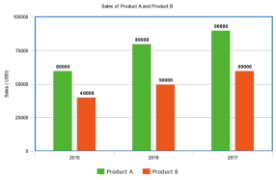
Example: The line graph below shows annual sales of a particular business company for the period of six consecutive years.

Remember to ALWAYS LABEL your graphs and charts with an appropriate: TITLE, X-axis, Y-axis

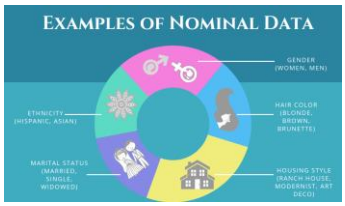
b.)Bar Chart

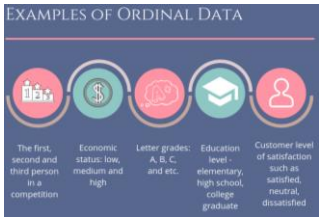
- Bar charts represent categorical data with rectangular bars.
- They are commonly used to compare several categories of data.
- Bar Charts Uses:
  - i. When you want to display ***data that are grouped into nominal or ordinal categories*** (see nominal vs ordinal data).
  - ii. To ***compare data among different categories***.
  - iii. Bar charts are ideal for ***visualizing the distribution of data when we have more than three categories***.

Example: The bar chart below represents the total sum of sales for Product A and Product B over three years.



Remember to ALWAYS LABEL your graphs and charts with an appropriate: TITLE, X-axis, Y-axis

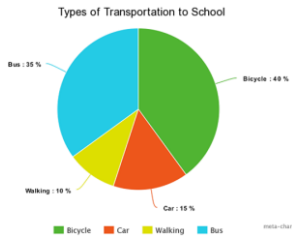




c.)Pie Chart

- Pie charts are good for illustrating and showing sample break down in an individual dimension.
- Pie Chart Uses:
  - i. When you want to ***create and represent the composition of something***.
  - ii. It is very useful for ***displaying nominal or ordinal categories of data***.
  - iii. To ***show percentage or proportional data.***
  - iv. When ***comparing areas of growth*** within a business such as profit
  - v. Pie charts work best for ***displaying data for 3 to 7 categories.***

Example: The pie chart below represents the proportion of types of transportation used by 1000 students to go to their school.

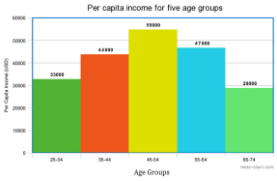


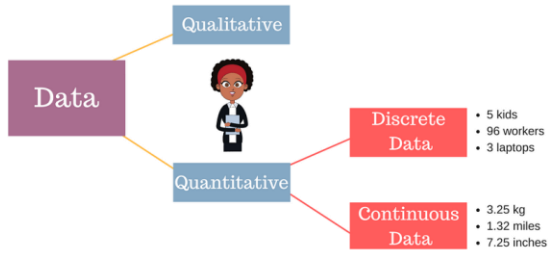
d.)Histogram

- A histogram shows **continuous data** in ordered rectangular columns. A histogram *displays a frequency distribution (shape) of a data set.*
- At first glance, histograms look alike to bar graphs. However, there is a key difference between them. ***Bar Charts represents categorical data and histogram represent continuous data.***
- Histogram Uses:

i. When the *data is continuous*.
ii. When you want to *represent the shape of the data's distribution.*
iii. When you want to see whether the outputs of two or more processes are different.
iv. To summarize large data sets graphically.
v. To communicate the data distribution quickly to others.

Example: The histogram below represents per capita income for five age groups.



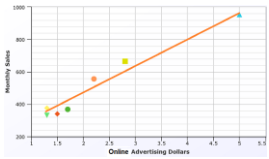


e.)Scatter plot

- A scatter plot shows a relationship between two variables.
- Scatter plots also help you predict the behavior of one variable (dependent) based on the measure of the other variable (independent).
- Scatter plot uses:

i. When trying *to find out whether there is a relationship between 2 variables*.
ii. To *predict the behavior of dependent variable* based on the measure of the independent variable.
iii. When having paired numerical data.
iv. When working with root cause analysis tools to identify the potential for problems.
v. When you just want to visualize the correlation between 2 large datasets without regard to time.

Example: The Scatter plot below presents data for 7 online stores, their monthly e-commerce sales, and online advertising costs for the last year.



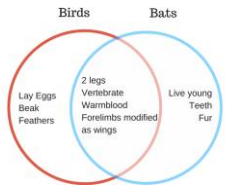
- The orange line you see in the plot is called **"line of best fit"** or a **"trend line"**.
- This line is used to help us make predictions that are based on past data.

f.)Venn Chart/Diagram

- Venn Diagrams use overlapping circles to visualize the logical relationships between two or more group of items.
- The items in the overlapping section have specific common characteristics. Items in the outer portions of the circles do not have common traits.
- Venn Chart Uses:

i. When you want to *compare and contrast groups of things.*
ii. To categorize or group items.
iii. To *illustrate logical relationships from various datasets*.
iv. To identify all the possible relationships between collections of datasets.

Example: The following science example of Venn diagram compares the features of birds and bats.

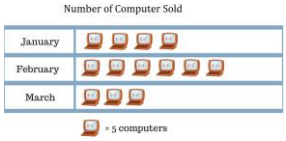


g.)Pictogram

- A pictogram displays numerical information with the use of icons or picture symbols to represent data sets.
- Pictogram Uses:

i. When your audience prefers and understands better displays that include icons and illustrations. Fun can promote learning.
ii. It's habitual for info graphics to use of a pictogram.
iii. When you want **to compare two points** in an emotionally powerful way.

Example: The following pictogram represents the number of computers sold by a business company for the period from January to March.



Summary of Graphs and Charts

- Graphs and Charts enable one to do the following: ***Inform, Compare, Show Change, Organise or Reveal Relationships*** between variables… (ICCOR)

Intended Goal…	Graph or Chart to Use
1.) Inform	• Pictogram • Donut Chart
2.) Compare (*you want to compare categories or show compositions*)	• Bar Chart (*to compare many categories*) • Pie Chart (*to show composition*)
3.) Change (*you want to show change over time or by location*)	• Line Chart • Timeline
4.) Organise (*you want to show groupings, rankings or processes*)	• Venn diagram • Ordered Bar Chart
5.) Reveal Relationships between variables…	• Histogram (*for distribution of one variable*) • Scatter Plot (*relationship between two continuous variables*)

Example Question….

When comparing the data for men and women, a researcher found that there was a difference in the proportion of friendly and aggressive social interactions. This is shown in Table 2 below.

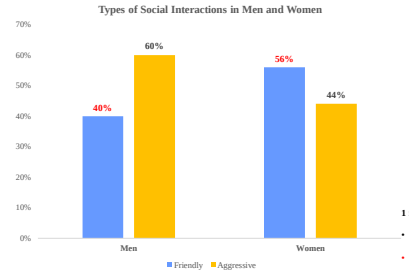
Table 2: Percentage of friendly and aggressive social interactions reported by men and women.

	Men	Women
Friendly	40%	56%
Aggressive	60%	44%

a) Draw a suitable graphical display to represent the data in Table 2. Label your graph appropriately. **(4 marks)**

Answer….

The goal is to show **COMPARISON**…therefore, the most suitable would be a **Bar Chart.**



1 mark for each of the following:

- appropriate **title** with **reference to the IV and the DV**
- **accurately plotted** data, **bars must not touch**
- accurate **y-axis label**, percentage/% is sufficient.
- accurate **x-axis label** with key, where necessary.

3. DESCRIPTIVE MEASURES

- Measures of **central tendency** and **dispersion** are common descriptive measures for summarising numerical data.

1. Measures of central tendency are measures of the location of the middle or the center of a distribution.

- The most frequently used measures of central tendency are the ***mean***, ***median*** *and* ***mode***.

2. A measure of dispersion is a numerical value describing the amount of variability present in a data set.

- The ***standard deviation (SD)*** is the most commonly used measure of dispersion.
- The ***range*** can also be used to describe the variability in a set of data and is defined as the difference between the ***maximum*** *and* ***minimum*** values.

DEFINITIONS:

- The ***mean is the average***; it is the sum of all the data entries divided by the number of entries.
- The ***median is the middle score***. If we have an even number of events we take the average of the two middles.
  - *N/B: The* ***median is a more representative form of a measure*** *of central tendency (average) when there is unusual data,. –* ***this is because any 'extreme' or 'unusual' scores that would otherwise artificially inflate or deflate the average if the mean was calculated are disregarded and do not feature in the calculation.***
- The ***mode of a set of data is the number with the highest frequency***.
- The ***Standard Deviation*** is a ***quantity expressing by how much the members of a group differ*** from the mean value for the group.

What do mean, median, mode and standard deviation values tell us about data??

1.) The Mean

- The mean is ***the average of the data***, which is the sum of all the observations divided by the number of observations.

Example: the wait times (in minutes) of five customers in a bank are: 3, 2, 4, 1, and 2. The mean waiting time is calculated as follows: $\frac{3+2+4+1+2}{5} = \frac{12}{5} = 2.4 \text{ min}$

- On average, a customer waits 2.4 minutes for service at the bank.
- **Interpretation:**

The mean is used to describe the data sample with **a single value that represents the center of all the data**.

Example question…

In a research study, a researcher obtained a volunteer sample of 10 students aged 17 years . The participants were asked to complete two puzzle tasks as quickly as possible.

- Task A was to find 10 differences in a *'spot the difference'* puzzle while *working in silence*.
- Task B was to find 10 differences in another 'spot the difference' puzzle while *listening to music* through headphones.

The time taken to complete each task was recorded for each student.

Participant	Task A (silence)	Task B (music)
1	67	82
2	45	70
3	58	60
4	43	59
5	72	77
6	90	105
7	101	90
8	37	59
9	54	83
10	63	89

a) Explain **one reason why the mean would be the most appropriate** measure to summarise the data above. **(2 marks).**

b) **Calculate the mean** values **for both Task A and Task B**. Show your workings.(2 marks)

Answers…

a) Explain **one reason why the mean would be the most appropriate** measure to summarise the data above. **(2 marks).**

The mean can be said to be representative of all the data collected as it is calculated using all the individual values in a set of data.

b) **Calculate the mean** values **for both Task A and Task B**. Show your workings.**(2 marks)**

TASK A: Mean= $\frac{67+45+58+43+72+90+101+37+54+63}{10}$ = **63 SECONDS**

TASK B: Mean= $\frac{82+70+60+59+77+105+90+59+83+89}{10}$ = **77.4 SECONDS**

2.) The Median

- The median is **the middle value**. It is the value that splits the dataset in half.
- To find the median, **order your data from smallest to la**rgest, and *then find the data point that has an equal amount of values above it and below it*.
- The method for locating the median varies slightly depending on whether your dataset has an even or odd number of values.

In this dataset with the odd number of observations, notice how the number 12 has six values above it and six below it. Therefore, **12 is the median of this dataset.**

Median Odd
23
21
18
16
15
13
12
10
9
7
6
5
2

Median Even	
	40
	38
	35
	33
	32
	30
28	29
	27
	26
	24
	23
	22
	19
	17

When there is an even number of values, you count in to the two innermost values and then take the average. The average of 27 and 29 is 28. Therefore, **28 is the median of this dataset.**

Example…

- The following descriptive statistics were calculated. The mean and median are in blue.
- Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
Sales	70	100.057144	143.93872	11.997446	1.4339691	98.5	77	73	150	92	107

- Both the mean of 100,057 and median of 98,500 ***indicate*** where the center of the data is located, and ***what the typical daily number of newspapers sold is.***
- Thus, the typical number of newspapers sold daily is about 100,000.

3.) The Mode

- The mode is **the value that occurs most frequently in a data set**.
- The mean and median require a calculation, but *the mode is determined by counting the number of times each value occurs in a data set*.

Interpretation:

- The mode can be used with mean and median *to provide an overall characterization of your data distribution*.
- The mode can also be used to identify problems in your data.

For example: a bank manager collects wait time data for customers who are cashing checks and for customers who are applying for home equity loans. Because these are two very different services, the wait time data included two modes. The data for each service should be collected and analyzed separately.

4.) The Standard Deviation

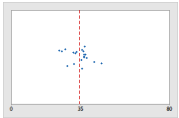
- The standard deviation is the most common measure of dispersion, or how spread out the data are about the mean.

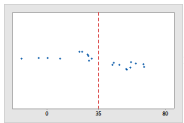
Interpretation:

- Use the standard deviation to determine how spread out the data are from the mean. A higher standard deviation value indicates greater spread in the data.
- The standard deviation can also be used to establish a benchmark for estimating the overall variation of a process.
- ***For example:*** **Hospital discharge times**

  Administrators track the discharge time for patients who are treated in the emergency departments of two hospitals. Although the average discharge times are about the same (35 minutes), the standard deviations are significantly different. The standard deviation for hospital 1 is about 6. On average, a patient's discharge time deviates from the mean (*dashed line*) by about 6 minutes. The standard deviation for hospital 2 is about 20. On average, a patient's discharge time deviates from the mean (dashed line) by about 20 minutes.

Hospital 1





Hospital 2

What calculations do I use??

Do you want to know how many individuals checked each answer?	Frequency
Do you want the proportion of people who answered in a certain way?	Percentage
Do you want the average number or average score?	Mean
Do you want the middle value in a range of values or scores?	Median
Do you want to show the range in answers or scores?	Range
Do you want to compare one group to another?	Cross tab
Do you want to report changes from pre to post?	Change score
Do you want to show the degree to which a response varies from the mean?	Standard deviation