

WJEC/Eduqas Geography A-Level

Independent Investigation Non-Exam Assessment

3 and 4- Data Presentation and Analysis
Notes



Introduction

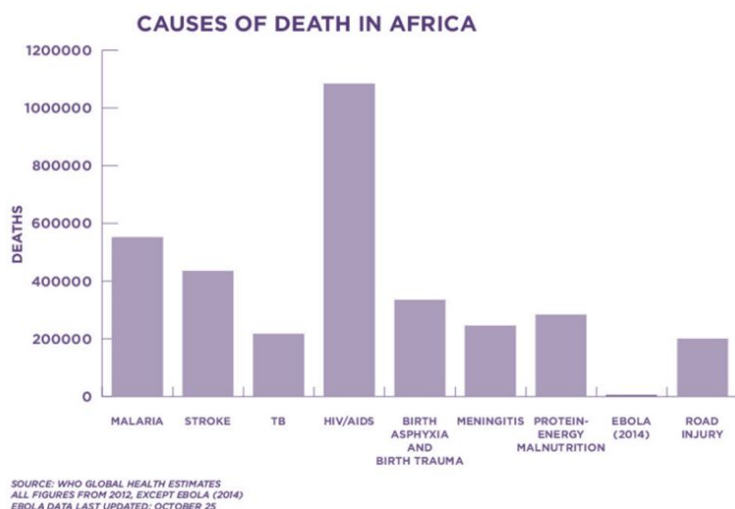
Once you have collected your data, you must collate and present it through a range of **presentational techniques**. The graphs, charts, maps and other sources you create undergo statistical analysis and/or a written qualitative analysis. This analysis is used with regards to your **hypotheses/ sub-questions**, which should eventually allow for your **main hypothesis/statement/question** to be answered or proved.

(N.B. It should be noted that some data within this guide has been created and manipulated to show the data presentation methods, and is not entirely accurate. Unless the graph has been taken from an external source (as referenced) it should be assumed that the data within the figure is false. This data has been manipulated to show clear data presentation methods that will serve as an educational resource for data presentation, rather than resources you can use as valid sources within your investigation)

Data Presentation Methods: Graphs/Charts

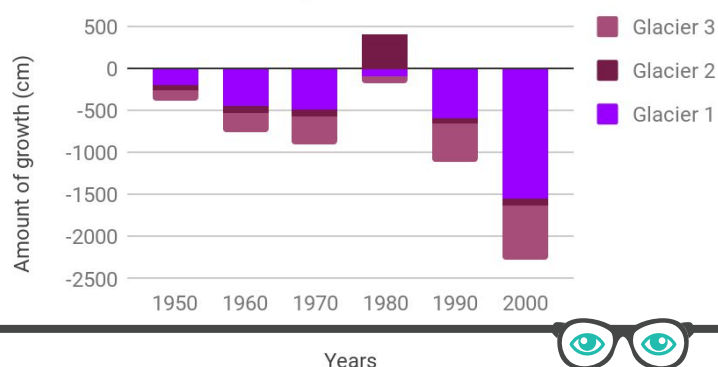
Bar Charts

Bar charts are useful when **tracking a change** (normally over time), or when **comparing factors** across different groups. The horizontal axis (x) usually contains the **independent variable**, which could be time, or the groups that will be compared.



A **simple bar chart** has the **independent variable** on the **horizontal axis** and the **dependent variables** on the **vertical axis**. This is useful to identify **relationships or correlations** between a **subject** (e.g. number of deaths) and a **factor** (e.g. types of deaths). If the changes in your data are gradual and your data is **categorical**, you should consider whether a line graph would be more suitable to determine trends.

Amount of Glacier growth from 1950 - 2000



For multiple subjects, a **stacked bar chart** could be more suited, which uses a **colour scheme** to separate the subjects. Ensure the colour scheme is obvious, has a clear **key**, and each subject can be defined in **greyscale**.

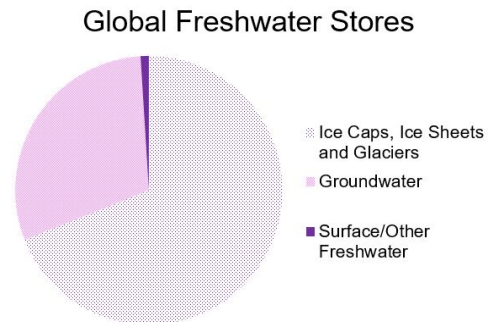


Bar charts can be simplistic and, although useful, higher level candidates should also use more technical data presentation.

Pie Charts

Pie charts are a useful way of presenting a wide range of data, especially that which is from **questionnaires** and **foot count/traffic surveys** (although useful, **make sure not to overuse them**). Sometimes just writing the **numeric figures** is sufficient, or using a **compound bar graph**, which could both be used instead of a pie chart. Pie charts allow **easy interpretation of data** by the reader, but **can also be misread**. When creating a pie chart it is recommended that:

- It is **2D**
- The **data is not labelled**
- The **segments have no gaps** between them
- The **colours are clear** and the different segments could be identified if the document was printed in greyscale (patterns are useful)
- The **key is explicit and easy to understand**
- There are not **too many segments**

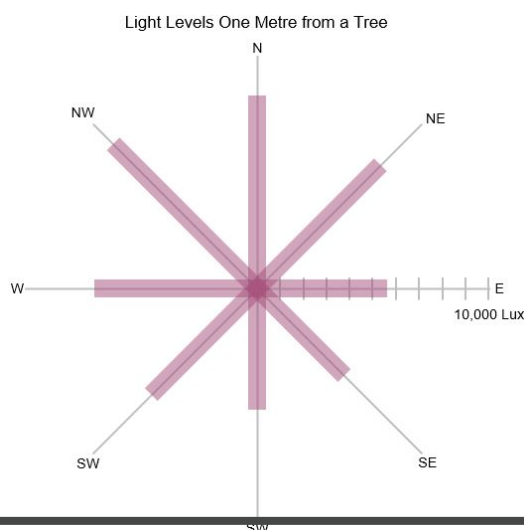
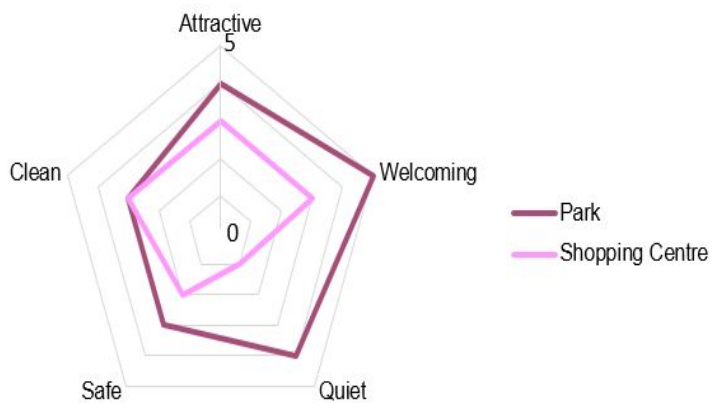


Radar Graphs

Radar graphs are most effective at displaying data from **environmental quality surveys**, or data about different locations. Data from various locations can be overlaid or compared on different charts.

It is important that all of **the scales are in the same direction**. All of the positive, highest scores should be in the same area of the graph (either all in the middle or all surrounding the outside). For example, rather than having quiet, welcoming, and **unsafe** all on the outside, it should be quiet, welcoming, and **safe** as shown in the radar graph. There is no limit to the amount of data sets that you can use, but using **too many sets may make the graph confusing**.

Radar Graph Showing Environmental Quality in Jakarta



Similar to a radar graphs, **rose graphs** use **multi-directional axes** to represent data, but with bars instead of lines. Rose graphs use **compass directions** for the axes directions, and you should define how far from a central point you are measuring when collecting the data. They could be useful for assessing forest cover (light levels), noise levels or wind speed, though there are many other possibilities. If you were investigating noise in a city centre area, you could use a rose graph over a **wider area** (10 metres in each direction).

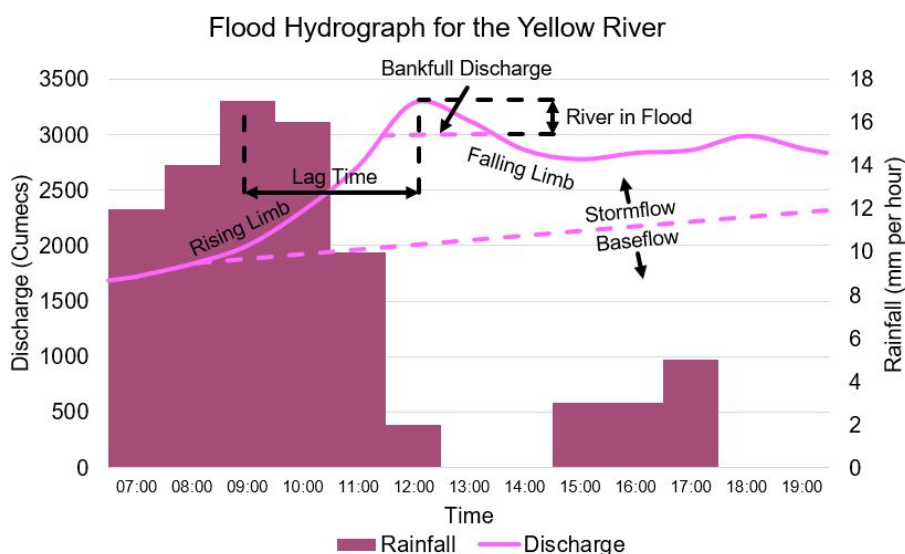
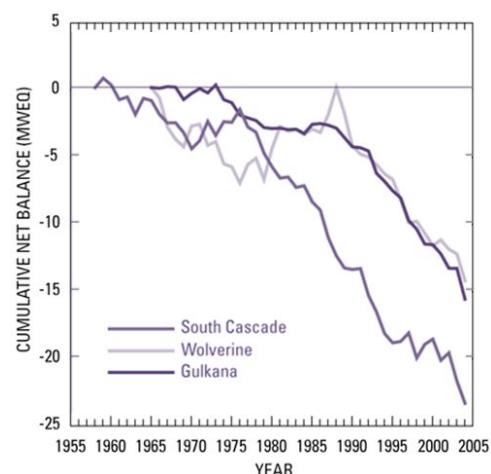


Line Graphs

Line graphs are useful for **tracking a change, usually over time**. In line graphs, the change that is being tracked will usually be a **gradual change** so that every point can be joined up in one line. A **key** could be used to track how **several factors change** over the same period.

Source: <https://pubs.usgs.gov/fs/2009/3046/>

Line graphs may be simplistic as stand alone graphs, but can also be used in **combination graphs**. For example, Flood Hydrographs use bar charts (precipitation) and line graphs (river discharge). You can create combination graphs by selecting 'Combo Chart' in data formatting programmes.

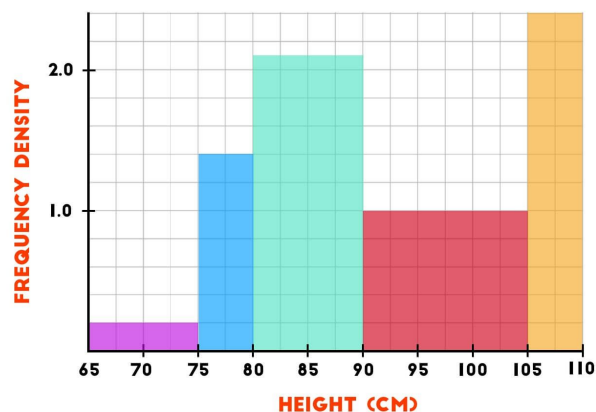


Histograms

Histograms are simply **bar charts of varying thicknesses**; for data with **different class widths**, a histogram is most appropriate. The **area** of a histogram's bar (the **frequency density** multiplied by the **class width**) is the frequency of your reading.

For example, in the figure, the red box represents a frequency of 15 people between the height of 90cm and 105cm.

There are histogram generators online, or use Excel - type all your data into a table, then highlight the table and click insert statistical chart.



Source: www.pythagorasandthat.co.uk

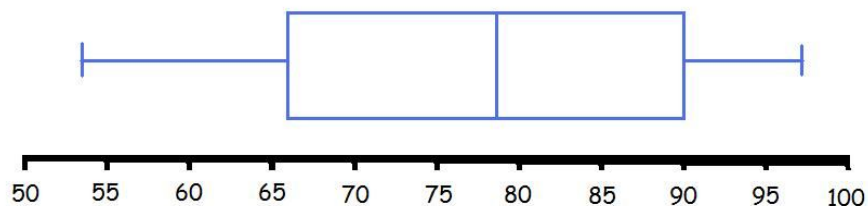


Box Plot

Box plots (sometimes called a Box and Whisker graph) are a pictographic way to represent the **median, range and interquartile range**. They are used to compare the spread of results and can be used to compare multiple sets of **continuous data**.

A box plot is easy to draw:

1. Draw an **appropriate scale horizontally** - Make sure your scale includes your maximum and minimum results, and should be for the variable you measured (e.g. the height of waves, time taken to erode, etc)
2. Draw a **small vertical line** where the median occurs. Repeat this for the maximum, minimum, upper and lower quartile (see later on how to calculate these values).
3. Join up the median, upper and lower quartile to form a box. Finally, draw a horizontal line connecting the maximum and minimum to the central box. Your diagram should look similar to the figure.

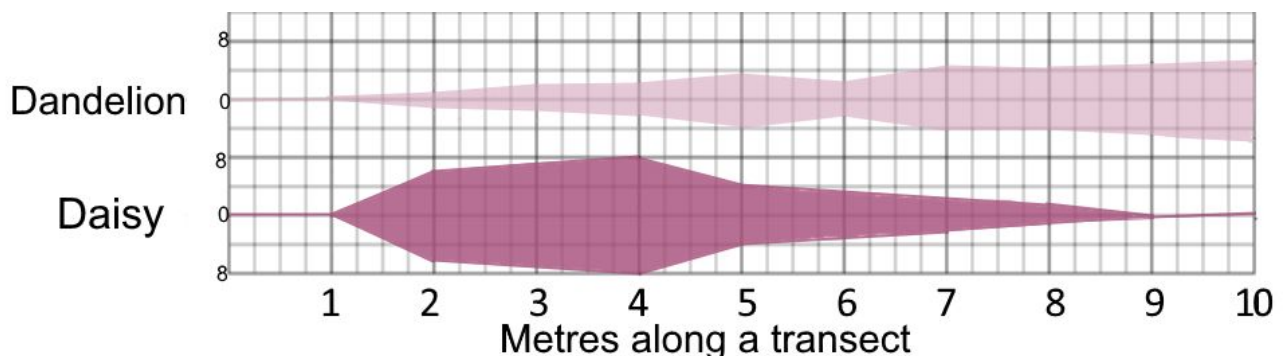


Kite Diagram

Kite diagrams show the **changes in frequency** of a factor over a **measured distance**, usually along a **transect**. **Multiple factors** being counted along the same transect can be shown in kite diagrams, which make them useful for **comparing spatial distribution** - especially of **plants and animals**.

Drawing kite diagrams:

Distance along transect (metres)		1	2	3	4	5	6	7	8	9	10
Plant	Daisy	0	12	14	16	8	6	4	2	0	0
	Dandelion	0	2	3	4	7	5	8	8	9	10



How a kite diagram works:


















- Y axis** - the y axis works like a **mirror**. In each **section** (e.g. the daisy section), the y axis should be as **wide** as your **largest piece of data**. In the **middle** of your section is **zero**, which is the **line of symmetry/ mirror line**. Each side of the mirror line goes up to **half of the largest piece of data**. In this example, the largest number is **16**, so each side of the mirror line goes up to **8** (because 8 is **half** of 16).
- Plotting points** - to plot points, your **value** should be **halved** and each half should be plotted on **either side of the mirror line**. This will create a **symmetrical shape** when all the points are plotted and joined up.
- Labelling** - all of your sections should be **labelled** with the **factor** you are measuring and the **distance** of the transect. On your y axis, you should also label **numbers**. Make sure to label the **zero line (mirror line) and the maximum value** (half the highest value). All of the sections (daisy, dandelion etc.) should be the **same size** so that you can **compare**. **Do not change the size of the sections** on the same graph. Always use the **biggest number** in the **entire** set of data to work out how wide your section should be.

Pictograms

Pictograms use **icons** or **pictures** to display sets of **discrete data** (data that has a **finite** count, i.e. **cannot have a decimal point**). Each icon represents a number, so that a completed pictogram will show the **frequency of a factor in different sets of data**. The icon usually resembles what is being counted. Here is an example of a completed pictogram that has been created from a building use survey.

A pictogram showing the different types of buildings on Main Street:

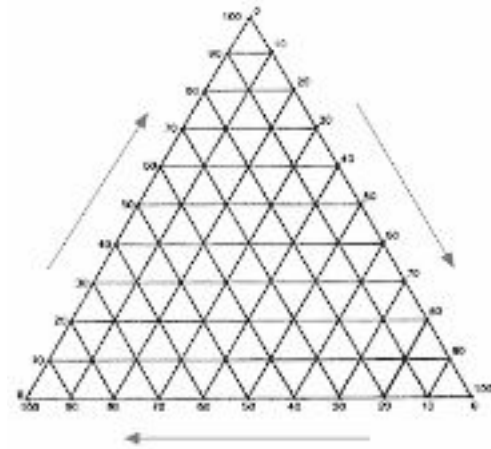
Residential 		18	Key: 1 icon = 1 building E.g.  = 1 house Overall - 41 buildings on Main Street
Industrial 		1	
Commercial 		8	
Entertainment 		5	
Public Building 		3	
Transport 		2	
Services 		4	

Pictograms are useful for presenting **simple counts** in interesting and understandable ways. However, pictograms can become **confusing** when there are **many numbers involved** because it



would require **counting many icons**, which is **unclear**. To present larger counts, a key could be used which **condenses down** the counts into **ratios** (e.g. 1 icon = 10 buildings) but in doing this, you must ensure your counts all have the same **highest common factor**. For example, if you want to use a key of 1 icon = 10 buildings, your different building counts should all divide by 10. Using **half icons or quarter icons** to represent a smaller number becomes **messy and confusing**, so generally it is best to stay away from pictograms when many numbers are involved.

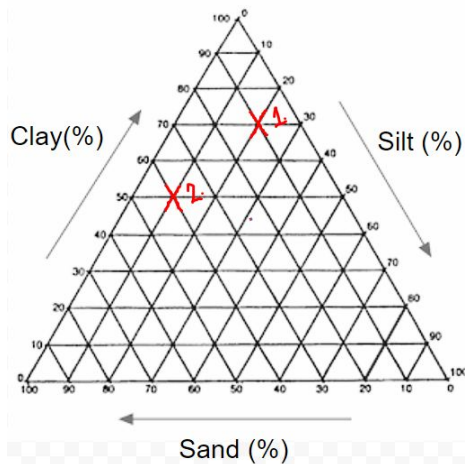
Triangular Graphs



Triangular graphs **compare the composition** of different data sets by using **3 variables that add to 100%**. Each data set is **plotted** as a point on the triangular axis shown.

Each axis (side of the triangle) goes up to **100%**. The data you plot must be composed of **3 variables** that **collectively** and **exclusively** add to make 100%.

For example, a triangular graph could be used to show **soil content** in 10 different locations (the **data sets**). In these 10 different areas, the soil content is mainly a mixture of clay, silt and sand. Here is how some areas would be plotted on a triangular graph.



Area 1: Clay - 70%, Silt - 20%, Sand - 10%.

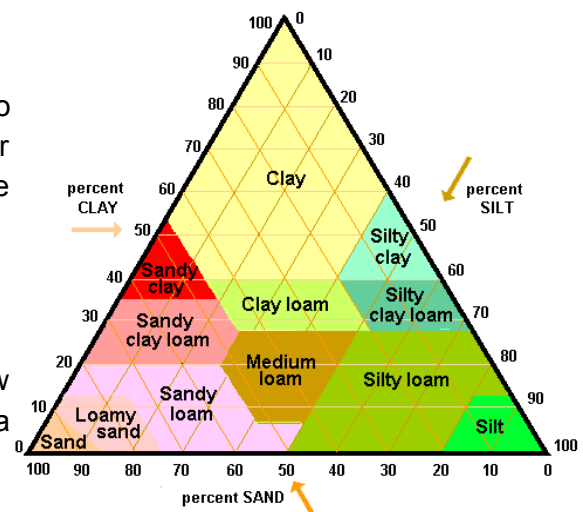
Area 2: Clay - 50%, Silt - 10%, Sand - 40%.

The easiest way to **plot points** is to read **horizontally from the left axis**, then **diagonally downwards from the right axis**, then **diagonally upwards from the bottom axis**. Ensure you label each point plotted as well as each axis.

Furthermore, you could split your graph up into **predetermined definable areas**, so that wherever your points lie indicates what your point could be defined as. Here is an example:

Source: <http://oneplan.org/Water/soil-triangle.asp>

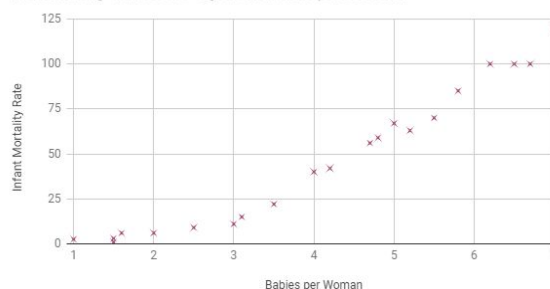
When your points are plotted, the graph easily show what the composition of your data set means in a **geographical context**.



Scatter Graphs

Scatter graphs are used to show the **relationship** or **correlation** between an **independent variable** and a **dependent variable**. Every piece of data is plotted like a coordinate on an axis: the **x axis** is the **independent variable**, i.e. the **cause**. The **y axis** is the **dependent variable**, i.e. the **effect**. Scatter graphs are useful in proving that a variable has a definite effect on a factor that is being observed in your investigation. If you are considering using a scatter graph, ensure that you have enough data collected so that a **clear** correlation can be identified.

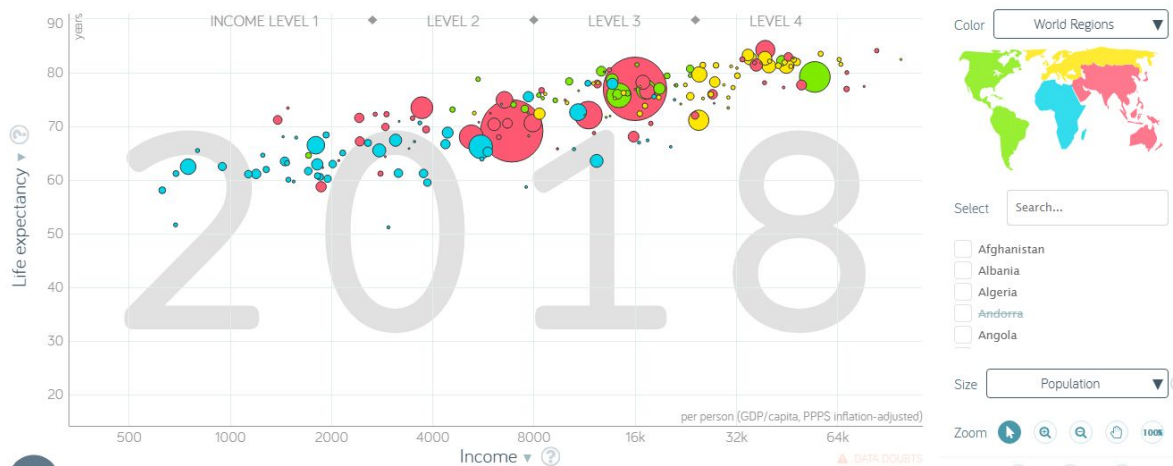
Babies per Woman vs. Infant Mortality Rate (Globally)
 Infant Mortality - number of 0 - 5 year olds that die per 1000 born.



If you want to **compare** how **multiple factors** are affected by the **same independent variable**, the different factors could be colour coded so your graph is **clear**. Ensure that you include a **key** so that each factor's colour code is obvious.

A **line of best fit** can be drawn on a scatter graph, which shows the **average trend** of your scatter graph. A line of best fit can be added on some softwares such as Microsoft Excel, or it can be drawn by you. A good way to check if a line of best fit is **accurate** is if there are approximately an **equal number of points above and below your line, excluding anomalies by circling them**.

Bubble charts are a type of scatter graph where the size of the plotted point also shows another **variable**.



Source: [https://www.gapminder.org/tools/#\\$chart-type=bubbles](https://www.gapminder.org/tools/#$chart-type=bubbles)

This bubble chart from Gapminder uses the size of the dots to show the country's population size. Using this type of graph allows 3 separate sets of data to be presented on one graph. However, a bubble chart should only be used if the information is **relevant** to your investigation. Ensure you indicate within your key what the size of your bubbles refer to.



Logarithmic Scales

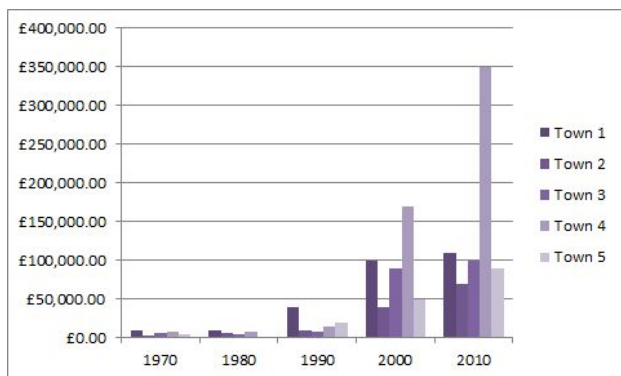
A **logarithmic scale** is a **non-linear** scale where numbers within a **large range** are **condensed** down into a **smaller, easier to understand scale**.

The **mathematical function** 'log' has a **base number**, which indicates how your raw data is condensed. The maths can be complicated, but log can be applied to your data by the **'log' button on your calculator or on a spreadsheet**, meaning all you have to do is **input your data**. Different base numbers can be used, but \log_{10} is most commonly used (10 is the base number here). In instances where \log_{10} is used, every time your **raw data increases by 1^{10}** , your **logarithmic scale goes up by 1**. This table shows the logarithmic scale that would be created when inputting certain numbers, and it also shows what would appear on your calculator when typing your raw data in.

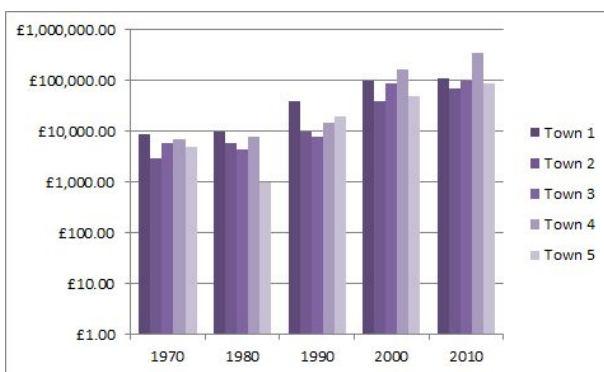
Raw number	$\log_{10}(x)$ format	Number that is used in your scale
10	$\log_{10}(10)$	1
100	$\log_{10}(100)$	2
1000	$\log_{10}(1000)$	3
10000	$\log_{10}(10000)$	4

The **Richter scale** is an example of a logarithmic scale, where a **4.0 magnitude** earthquake is **10 times stronger** than a **3.0**, **100 times stronger** than a **2.0**, and **1000 times stronger** than a **1.0**.

This scale is very useful when your raw data has a **very large range**, because it ensures smaller numbers are still **clear**. For example, this scale would be useful if some of your figures were in triple digits, and other figures were in six digits. In these two graphs, house price data in 5 towns is being compared. Note how the comparison is much clearer to see on the logarithmic scaled graph.



Without logarithmic scale.



With logarithmic scale.

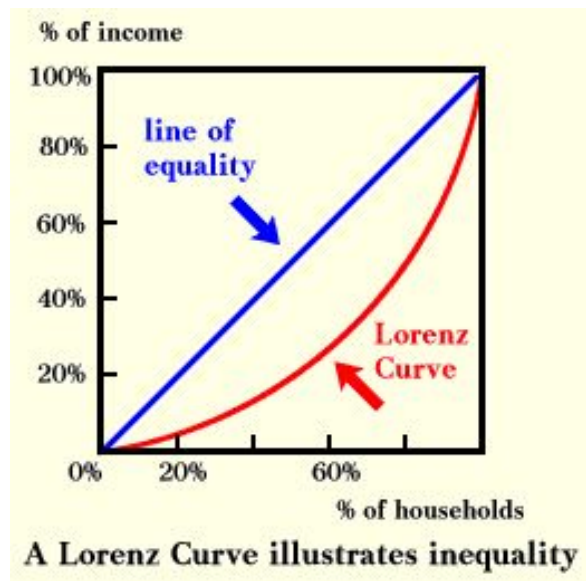
If logarithmic scales are used, make sure that you indicate this to avoid confusion.



Lorenz Curves

A Lorenz curve is a **graphical representation** of the **distribution** or **equality** of something. A **straight, diagonal line from the origin** represents **perfect equality**. The **further away** the Lorenz curve is from this, the **more diverse the sample is** and the more unevenly the values are spread out. From a Lorenz curve, the Gini Coefficient can be calculated.

1. **Rank** the data
2. **Order** the data by Rank
3. Calculate the **proportion** (percentage) of each data from the total data
4. Calculate the **cumulative proportion** by increasing rank (calculate the running total from adding the percentage of each line in turn to the line before it)
5. **Graph** the ranks on the **x axis** against the cumulative proportion on the y axis.
6. Draw the **perfect equality** diagonal line.



Data Presentation Methods: Cartographical

GIS

A GIS (**geographic information system**) is a form of data analysis and presentation. Any **digital presentation of data** in comparison to its **location or spatial distribution** can be classed as GIS. For colleges with GIS software, this is often simple to create (so see your teacher for more specific guidance). For those without specific GIS software, any of the cartographic data presentation methods may be considered GIS, if you create them **digitally** rather than drawn by hand.

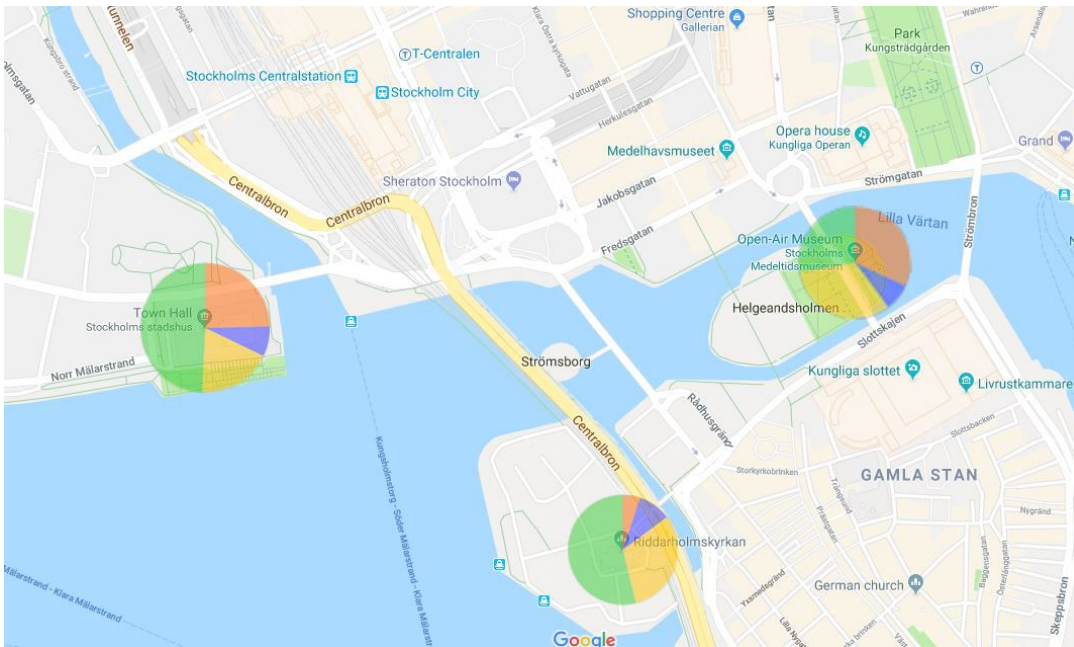
GIS is more **accurate** than hand drawing, as it minimises the **risk of human error** in drawing graphs or pinpointing locations. It can enable you to **display a large data set**, or multiple variables/information. For example, for each location sampled, a bar chart could display the average scores of several questions on an environmental survey or differing demographic characteristics along the rural-urban continuum. GIS is a key addition to create a high-level fieldwork report that looks **scientific** and **reliable**.

Map Overlays

Overlaying graphs or other data such as pie charts on to maps can be a powerful tool to easily display data. This map uses **proportional pie charts** and Google Maps, to display **footcount survey data**. The larger the pie chart, the greater the number of people that were surveyed in the 10 minute data collection period at the: Stockholm Town Hall, Riddarholmen Church and Stockholm Middle-Ages Museum. The various colours represent the **different proportions of people by age** (years) in each survey location.



Key: 0-15 16-30 31-45 46+



Map Credit (Google, 2018)

You could also use graphs such as bar charts as an overlay on the map, or even **qualitative data** such as quotes from questionnaire respondents. These maps are **high level additions** and should **not be overused**. The graphs which you overlay should **not be too large**, which would obscure the map, **or too small** so that the segments are difficult to see. Always ensure the data presentation method can be easily understood. To overlay graphs onto a map, you can create the graphs digitally, remove any data labels/axis etc. and save them as an image. Then remove the background using the tools on a programme such as PowerPoint and move the image onto a map. You can then adjust the size of the image to make them proportional, but make sure that the image is cropped right to the edges of the pie charts, otherwise this would be inaccurate.

Image Overlays

Similar to map overlays, image overlays may be used to display both a set of values and their location (often on a **smaller scale**). Using an image as a background may allow the use of **3D graphs** - such as 3D bar charts - to display data. For example, you could overlay the rates of erosion over an image of the coastline studied, hence showing localised variations in erosion rates.

3D graphs may look impressive, however caution must be taken to ensure that 3D graphs don't **overcomplicate** the image or **disguise trends** in data. If it's difficult to tell the start of the bar (and so the location of the data taken) or the height of the bars (and so

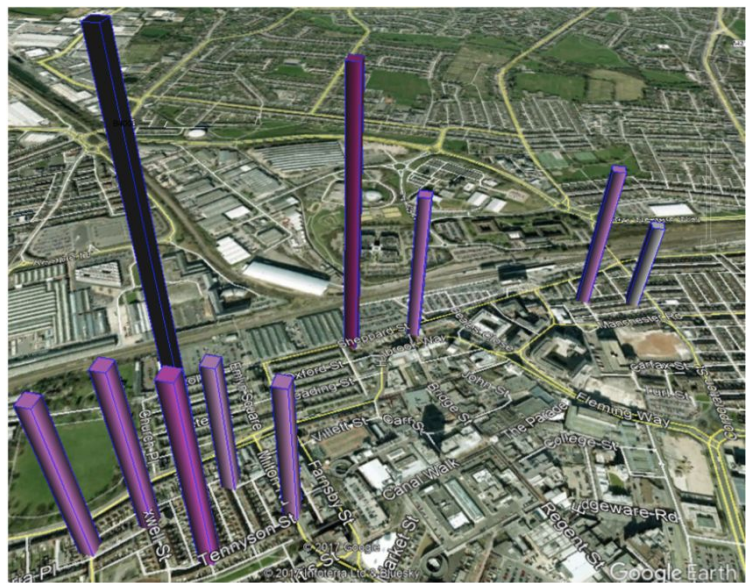


Figure 12 Overall Relative EQS Results in Central



the value of your dependant variable) - use 2D graphical presentation methods instead.

Isoline Maps

Isoline maps use lines to present areas in which points are of an **equal value**. Isolines are drawn using **geospatial** data (data that is specific to a location), and overall they show how the **value of a factor changes spatially**.

For example, an isoline map could be drawn to show **pedestrian density** in a town centre:

Raw data from several pedestrian counts.

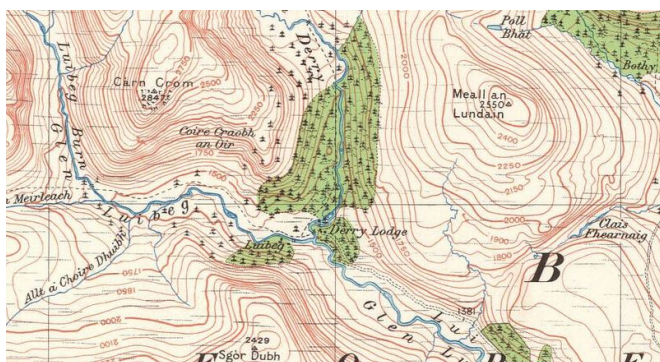


Isoline map for pedestrian density.



The isolines shown on the map are **10, 20, 30, 40, and 50**. Each of these lines show the **estimated** point at which a pedestrian count would show this number. Note that the lines show **estimations** where there is a **lack of data**, which makes isoline maps useful. However, it is also important to consider as estimating values can lead to **inaccuracies** and may disregard trends.

The **10 line** separates numbers **10 and below** from higher numbers. The **20 line** separates all values **below or equal to 20 but more than 10**. The **30 line** separates all values **below or equal to 30 but more than 20**, and so on. To make the isoline map even clearer, a **colour code** could be added to emphasise the different sections. When using isoline maps, ensure your lines' values increase by the same amount (e.g. all of the isolines increase by 10 in this example).



Source: <https://cairngormwanderer.wordpress.com/page/18/>

Isoline maps can be used for anything that shows many **different geospatial values**, and they are often used in weather or topography (the contour lines on maps are isolines). Take this map of the Cairngorms in Scotland for example. Each line represents elevation above sea level, and each isoline increases by 500m. The isolines get closer together when the elevation increases rapidly.



Dot Map

A dot map is a cartographical data presentation technique that uses small dots on a map to show the **distribution** or **density** of an **observation**. Wherever you have a recorded observation, a dot is marked on that **exact location on the map**. If you have two observations in the same exact location, the dot can be placed slightly to the side so that it can still be seen. Eventually, the map will have enough dots that a clear pattern is shown, which indicates both **spatial distribution** of your observation and the **density** of your observation within certain areas.

What to consider when using a dot map:

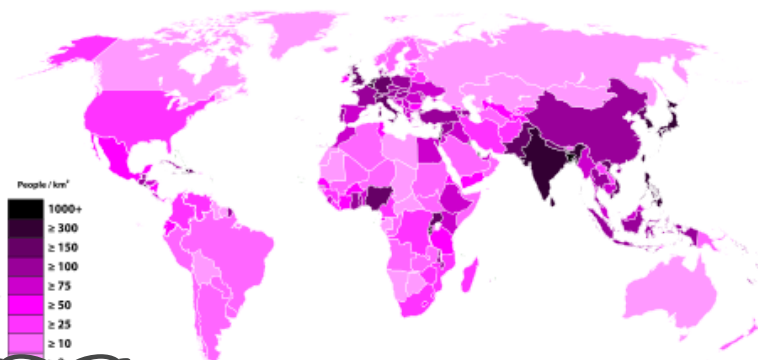
- Collect a **large amount of data**. Dot maps are only useful when you have a lot of data so that a trend in **density** can be spotted. If you have a small amount of dots, your trends will be unclear. Make sure to choose something you are able to make **hundreds of observations** in.
- Use **appropriate dot sizes**. If your dots are too large, they may overlap meaning observations are lost. Ideally, your dot size should be large enough to create **density patterns** but small enough so that (if you were to zoom in) every dot could be seen. Compare the two dot maps (Source: <http://learnGIS.uk/creating-statistical-dot-density-map-qgis/>). The first dot map clearly uses dots that are too large, as all of the map is concealed and there is no density pattern. The second map uses dots that are appropriate for the dot map.



- What do your dots **represent**? **1 dot** could represent **1 observation (1:1)**, or it could represent **more than 1 observation (1:many)**. If you have a **large amount of data**, it may be useful to include a **key** where **1 dot represents more than 1 observation**, e.g. 1:10 ratio, where 1 dot actually represents 10 observations. You may also wish to include a **colour coordinated key**, where the colours of the dots show different observations.
- Is a dot map the most **appropriate** method for your data? If you are recording something that has a **yes/no** or **existent/non-existent** observation, dot maps can be useful. However, if you are recording something that **can be recorded on a continuum or scale**, a **choropleth map** may be more suited.

Choropleth Map

A choropleth map uses **colouring** or **shading in predetermined areas** to show the **average prevalence** of a **phenomenon**. This factor can be



Countries by Population Density in 2015



recorded on a **scale** or **continuum**, and each colour indicates a different section (called **data classes**) of your scale.







Source: https://simple.wikipedia.org/wiki/File:Countries_by_Population_Density_in_2015.svg

Choropleth maps are useful when analysing the **intensity/prevalence/frequency** of a phenomenon in different areas. **Patterns** can be spotted within your map, which makes choropleth maps useful for analysing factors that vary **spatially**.







Things to consider:

- **Borders** - your sectioned **sub-areas** should be small enough so that your average value does not **ignore** obvious **variances** within your data. Your borders should also be large enough so that you can collect a sufficient amount of data within your **timeframe**. If you have many small areas that show the same observation, it may be worth **condensing** these areas into one larger area.
- **Colours and keys** - use colours that can be seen in **greyscale** and try to choose colours that can be distinguished from each other clearly. Your key should use quoted numbers rather than vague interpretations.

Clear numbers.

	<10%
	10-20%
	21 - 30%
	31 - 40%
	41 - 50%
	>50%

Vague description.

	Very Low
	Low
	Slightly Low
	Average
	Slightly High
	High

- **Number of Data Classes** - There should be a **compromise** in the number of data classes used in the key. If you have **too few** data classes, there may be a **generalisation** of your data. **Too many**, and the colours in your choropleth map could be **indistinguishable**.

Flow Lines

Flow lines can **represent movement** of people, animals etc., to or from a specific place. They are best **not confused with desire lines**, which can have a different meaning.

The example below shows the areas of Barcelona which students travel from, to reach St. Peter's School Barcelona. The **width of the arrow** is used to represent how many students travel from a particular area to the school. It is only a generic indication, but can be a useful tool, which could also represent migration, people travelling to an event etc. and is therefore most likely to be used in an Urban Environments or Changing Places fieldwork investigation. They are useful for analysis when **considered with additional sources** of data.

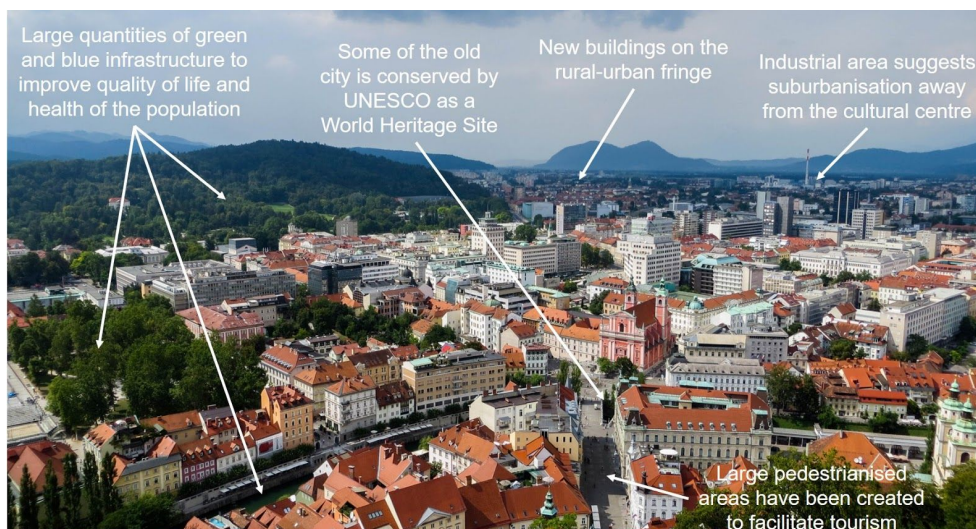




Data Presentation Methods: Qualitative Data

Photographs

If you are using photographs as a method of data presentation then they should be **clearly annotated**, with information relating to your hypotheses. You may use **place names, geographic theories or observations** that you made when taking the photos to give them greater meaning and relate them to your investigation.





NE Ljubljana as seen from the Ljubljana Castle - It is appropriate to **state which direction the photograph is taken from** if known.

Map Credit: (Google, 2018)

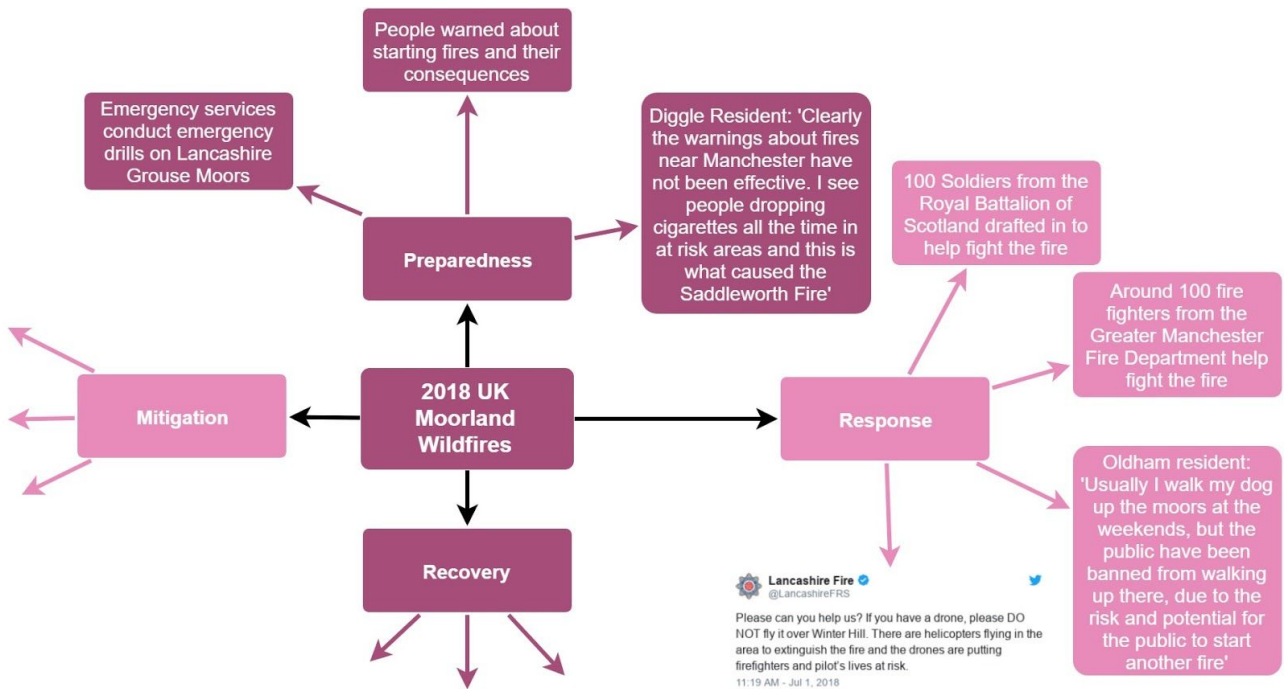
Photographs are a great way to enrich your report and make it more stimulating to read. You may take pictures of large area, but analysing smaller areas in detail is just as important. You may want to take images of your survey sites. When taking photos consider:

- The **angle** which you are taking the photograph at - what does it show and what does it not show? Is it objective?
- Be careful not to photograph individual **people** as you do not have their permission.
- What is the area like that you are photographing? Is it **safe**? If not, it may be advisable to not take pictures if there is a risk, for example risk of your camera/phone being stolen.
- Can you take pictures in the **same location** with the **same angle** at **different times** and then compare these later?
- Could you find images **online** and then photograph the same place yourself? What are the differences if there are any?

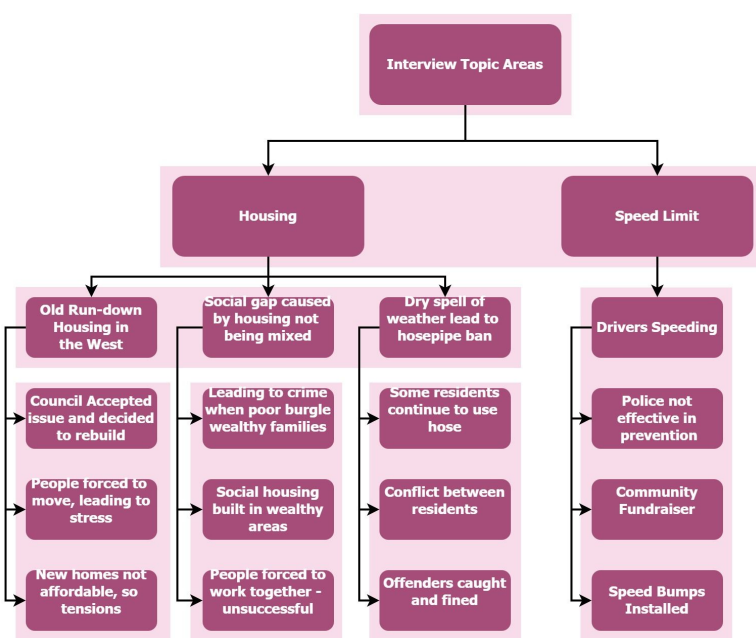
You can also compare pictures that you take to maps that may be online. You can use Bing Maps to find Ordnance Survey maps, and Google maps to show terrain and different places. This creates an insightful source of data presentation.



Quotes/Mind Maps



Mind maps are another useful tool used to present a wide range of **qualitative** (and also quantitative) data. It may be useful when creating a mindmap to structure it around a part of the **core theory** - in the example above, the Hazard Management Cycle. This gives **clear structure** to the mind map and demonstrates your **knowledge of the geographical context** in your investigation. It is important to ensure that you do not put **too much information** on the mind map, which would reduce its readability and effectiveness as a data presentation technique. You could also sort the data based on your hypotheses or the sub areas which you are investigating in your fieldwork.



A mind map could also take the form shown on the left, which codes and splits different areas from an **interview** into different topic sections, with each quote linking to the next. This can be useful not only for data presentation, but helping you to sort through the **relevant** sections of an interview transcript, which is likely to contain a lot of data raw data that will be **hard to analyse** otherwise.



Word Clouds

Word clouds are very useful for presenting data from qualitative sources such as **questionnaires**. It would otherwise be difficult to present this data in an easily viewable form. Websites such as wordclouds.com allow you to **paste text**, such as questionnaire and interview transcripts. The **most common words** will appear larger, so you can easily assess the **themes** present.

You can also compare word clouds between different locations. It may be useful to remove words such as 'the', 'and', 'I' etc. as these are frequent, but offer no benefit when included. You should also be careful as phrases such as 'not welcoming' will be split into 'not' and 'welcoming' which may incorrectly suggest trends which are not present. You could hyphenate these phrases to overcome the problem. It may be best to include adjectives which provide the greatest insight in this qualitative format, but using only adjectives may omit other relevant data.



Data Analysis Methods

Measures of Central Tendency

The term 'measures of central tendency' refers to a group of **statistical tests**. These statistical tests describe data distribution in relation to the '**middle**' value to indicate the **concentration** of the values in the **central part** of the **distribution of frequencies of the whole data**.

The numbers below will be used as an example for each measure of central tendency.

13 25 79 82 1 45 49 45 67 45 1

Mean:

The mean is calculated by **adding up all the data** and dividing by the **number of data items**.

For example, using the numbers above, the sum would be 452 and there 11 numbers, so the mean would equal 41.1 to 3sf.

Mode:

The **most appearing number**. In the example above, the mode is 45.

Median:

The median is the **midpoint** value. The data needs to be ranked first from lowest to highest value.

1 1 13 25 45 45 45 49 67 79 82

- When there is an **odd number** of data items, the **median is a whole number**. As in the example above, there are 11 data items, so the median is 45.
- When there is an **even number** of data items, the median lies across the **two items** at the midpoint. The median is therefore an **average** (mean) of the **two middle items**.

Measures of Dispersion

The term 'measures of dispersion' refers to a group of statistical tests. These statistical tests describe data distribution.

Range:

The range describes the **spread of the data**. Simply, **subtract the highest number from the lowest number**. In the example above, the range would be: $82 - 1 = 81$

Interquartile range:

The interquartile range shows where the middle 50% of the data lie. Anomalies should be ignored in this calculation.

- Find the **median** using the method above. (45)
- Find the **lower quartile** by calculating the **median of the lower half of the data**. (13)
- Find the **upper quartile** by calculating the **median of the upper half of the data**. (67)



- The **difference** between the lower and upper quartiles is the interquartile range. (67-13 = 54)

Standard deviation

This shows by how much most piece of data vary from the mean.

1. Find the **mean** of the data.
2. Calculate, in a separate column, how **each piece of data differs** from the mean.
3. **Square** this value.
4. Use this equation:

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

where S = the standard deviation of a sample,
 Σ means "sum of,"
 X = each value in the data set,
 \bar{X} = mean of all values in the data set,
 N = number of values in the data set.

Using the example above: **Mean = 41.1** (3sf)

Data value	Variance from the mean ($x - \bar{x}$)	Variance from the mean squared ($x - \bar{x}$) ²
13	28.1	789.61
25	16.1	259.21
79	37.9	1436.41
82	40.9	1672.81
1	40.1	1608.01
45	3.9	15.21
49	7.9	62.41
45	3.9	15.21
67	25.9	670.81
45	3.9	15.21
1	40.1	1608.01
Sum:		8152.91



$$\text{Standard deviation} = \sqrt{\frac{8152.91}{11}} = 27.2 \text{ (3sf)}$$

Variance:

This shows how far each piece of data varies from the average. It is simply equal to the **square of the standard deviation**.

Data Analysis Methods: Statistical Testing

Spearman's Rank

Spearman's Rank tests the **relationship (correlation) between two sets of data**. For example, it could test the correlation between age of respondents and the score for their perception of the city centre, or the sediment size along a coast and the rate of erosion there. Completing Spearman's Rank is best in a **table**, as shown below with a series of steps to follow:

1st set of data	r_1 - ranks for 1st set of data	2nd set of data	r_2 - ranks for 2nd set of data	d - difference between ranks	d^2 - difference squared

Method

1. List a set of data (e.g age of respondent) in the first column. Then **rank** each piece of data relative to each other in the second column - for example, the youngest person will rank 1, the second youngest is 2, etc.
2. List your second set of data and rank each piece (similar to your first set) in the following columns. If there is more than one respondent with the same answer (for example the same score or same age) then you may rank them consecutively in any order. Ensure that you do not skip any rank; as a check, ensure that your lowest/ worst rank number is the same as your sample size (e.g. 20th is the last rank, and there are 20 people in your sample)
3. Calculate the **difference between the two ranks** - along one row, take the second rank from the first rank ($r_2 - r_1$).
4. **Square** this difference and record the value.
5. Repeat steps 3 and 4 for each row. Add up all values in the final column.
6. Complete these two word equations with your own values, remembering to calculate the brackets first:

$$(6 \times \text{the sum of the final column}) \div (n \times n \times (n - 1))$$

$$\text{Spearman's Rank} = 1 - (\text{the value you calculated above!})$$

As a check, your value must be between 1 and -1. Alternatively (if you're more maths-y!) the actual equation is :

$$R = 1 - \frac{6 \sum d^2}{n(n-1)}$$



NB: This equation may come up in your exam, so be familiar with it. However, the word equations above are exactly the same steps.

7. To finish, you must **describe the correlation** between your data.
 - If the value (ignoring the sign) you calculated is greater than 0.5, then your data has a strong correlation. Or if the value you calculated is smaller than 0.5, then your data has a weak correlation.
 - If your Spearman's Rank is positive, then your correlation is positive. A negative correlation will cause a negative Spearman's Rank.

Chi-Squared Test

The Chi-Squared test looks at the **relationship** between a set of data of interest (such as that that you have collected or observed from your fieldwork) and a **theoretical/expected set** of data to decide whether the difference between the two is **significantly different**. It is used to see how closely the data from the research fits with the widely accepted findings or what you expected to find. This test only checks to see if there is an association between two sets of data, **not** what the nature of the relationship might be between those sets, nor the strength of any relationship.

It can be used on data which has the following characteristics:

- The data must be in the form of **frequencies counted** in a number of groups (% cannot be used).
- The total number of observations must be **> 20**.
- The observations must be **independent** (i.e. one observation must not influence another).
- The expected frequency in any one category must not normally be **> 5**.

Method

1. State the hypothesis being tested – there is a significant difference between sample groups. It is convention to give a null hypothesis – no significant difference between the samples.
2. Tabulate the data as shown in the example below. The data being tested for significance is the 'observed' frequency and the column headed 'O'
3. Calculate the 'expected' number of frequencies that you would expect to find in the column headed 'E'.
4. Calculate the statistic using the formula

$$\Sigma \frac{(O - E)^2}{E}$$

5. Calculate the degrees of freedom.
Degrees of freedom = number of rows - 1
6. Compare the calculated figure with the critical values in the significance tables using the appropriate degrees of freedom. Read off the probability that the data frequencies you are testing could have occurred by chance.



If the calculated value **exceeds** the tabulated critical value for the correct number of degrees of freedom at the given confidence level (usually 95%), then **reject the null hypothesis**. This means that it can be stated with 99% confidence that there is a statistically significant difference in the data sets, and this difference is not down to chance. If the calculated Chi-Squared value is smaller than the critical value, **accept the null hypothesis**.

An easy way of remembering this: **MRSA**

More than
Reject
Smaller than
Accept

Exemplar: (modified from the WJEC/Eduqas Teaching Guidance)

Investigating the size of pebbles along a beach to determine whether the variations in pebble size are **significant or random**. If there is no difference in the sizes of pebbles, the sites should all have approximately the same frequency of pebbles > 5cm.

Null hypothesis: There is no significant difference in the sizes of pebbles sampled along the beach.

Alternative hypothesis: there is a significant difference in the sizes of pebbles sampled along the beach.

Beach site	O Number of pebbles > 5 cm long	E Mean number of pebbles > 5 cm long	(O-E)	(O-E) ²	(O-E) ² /E
1	40	18	22	484	20.89
2	15	18	3	9	0.5
3	5	18	13	169	9.39
4	12	18	6	36	2
					Σ 38.78

Degrees of freedom: $4 - 1 = 3$

Degrees of freedom	Probability of error 0.10	Probability of error 0.05
1	2.706	3.841
2	4.605	5.991
3	6.251	7.815
4	7.779	9.488
5	9.236	11.070
6	10.645	12.592
7	12.017	14.067
8	13.362	15.507
9	14.684	16.919
10	15.987	18.307

0.05 (95% confidence level) = 7.815



“As the calculated value of 38.78 exceeds the tabulated figure at 3 degrees of freedom at the 95% confidence (7.814), it can be stated with 95% confidence that there is a statistically significant difference in pebble size along this stretch of beach.”

For a further example involving two samples, look to the [WJEC/Eduqas teaching guidance](#) document page 72.

T-test

The Student's t-test looks at the **means of two sets of data** and decides whether there is a significant difference between the two. It looks at the degree of overlap between the two samples. It applies to data that is measured on an **interval** or ratio scale and for data that is **normally** distributed around the mean.

The null hypothesis is that the two data sets are **the same** (there is no significant difference between them). The alternate hypothesis is that there is a significant difference between the two data sets.

Method

1. Calculate the **mean** and **standard deviation** for the two sets
2. Plug the values into this formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

\bar{x}_1 = Mean of sample 1

\bar{x}_2 = Mean of sample 2

S_1 = Standard deviation of sample 1

S_2 = Standard deviation of sample 2

N_1 = Number of subjects in sample 1

N_2 = Number of subjects in sample 2

To see if your t value is significant you will need to calculate the degrees of freedom and compare your calculated t value to the appropriate critical value.

These critical values give 95% confidence. This means that if your calculated t value is the same or higher than the critical value, you can be **95% confident** that you have a **significant difference** between your two sets of data.

$$\text{Degrees of freedom} = n_1 + n_2 - 2$$

If calculated $t \geq$ critical t you **reject** your null hypothesis and accept your alternative hypothesis.

If calculated $t <$ critical t you **accept** your null hypothesis and reject your alternative hypothesis.



An easy way of remembering this: **MRSA**

More than
Reject
Smaller than
Accept

Mann-Whitney U

Mann-Whitney U looks at the **medians of two sets of data** and decides whether there is a **significant difference** between the two.

It can be used on data that has the following characteristics:

- The 2 samples are **independent**
- The data is ordinal- it can be **ranked**
- There are at least **6** pairs of data
- It does not require a normal distribution
- It does not require there to be the same number of data sets

Method

1. Label one data set 'sample A' and the other 'sample B' and find the n values

Sample A: 22, 18, 25, 33, 31, 28, 19, 24, 29 n_a (number of data points in A) = 9

Sample B: 26, 18, 30, 16, 35, 21, 31, 17, 18, 27 n_b (number of data points in B) = 10

2. **Rank** all of the data points in sample A and Sample B all together as one set (order the data in each sample for ease)

A	18	19	22	24	25	28	29	31	33		
Rank(R_a)	4	6	8	9	10	13	14	16.5	18		$\Sigma R_a = 98.5$
B	16	17	18	18	21	26	27	30	31	35	
Rank(R_b)	1	2	4	4	7	11	12	15	16.5	19	$\Sigma R_b = 91.5$

Where ranks are tied, add up the corresponding ranks, divide by the number of tied ranks and give this rank to all the tied ranks.

E.g. 18, 18 and 18 are tied across ranks 3, 4 and 5

$(3+4+5) \div 3 = 4$ so all the 18s get a rank of 4. The next number in the ranking (19) gets a rank of 6 as 3, 4 and 5 have been used by the 18s.

E.g. 31 and 31 are tied across ranks 16 and 17.

$(16 + 17) \div 2 = 16.5$ so both 31s get a rank of 16.5.

3. **Sum up** the **ranks** of sample A and sample B

4. Calculate the **U values** using the formula:



$$U_a = n_a n_b + \frac{n_a(n_a + 1)}{2} - \sum R_a$$

and

$$U_b = n_a n_b + \frac{n_b(n_b + 1)}{2} - \sum R_b$$

$$U_a = (9 \times 10) + \frac{9(9+1)}{2} - 98.5$$

$$U_b = (9 \times 10) + \frac{10(10+1)}{2} - 91.5$$

$$U_a = 36.5$$

$$U_b = 53.5$$

5. Select the **smaller of the two U values**

Smaller U value is $U_a = 36.5$

6. Look up the critical values in the table at the given level of significance.

Level of significance: 5% ($P = 0.05$)

		Size of the largest sample (n_2)																				
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22			
Size of the smallest sample (n_1)	3	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9			
	4	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14	15	16			
	5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20	22	23			
	6		5	6	8	10	11	13	14	16	17	19	21	22	24	25	27	29	30			
	7			8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38			
	8				13	15	17	19	22	24	26	29	31	34	36	38	41	43	45			
	9					17	20	23	26	28	31	34	37	39	42	45	48	50	53			
	10						23	26	29	33	36	39	42	45	48	52	55	58	61			
	11							30	33	37	40	44	47	51	55	58	62	65	69			
	12								37	41	45	49	53	57	61	65	69	73	77			
	13									45	50	54	59	63	67	72	76	80	85			
	14										55	59	64	67	74	78	83	88	93			
	15											64	70	75	80	85	90	96	101			
	16												75	81	86	92	98	103	109			
	17													87	93	99	105	111	117			
	18														99	106	112	119	125			
	19															113	119	126	133			
	20																127	134	141			

Mann-Whitney U is an **exception to the 'MRSA' rule** for these statistical tests.

If the smaller U is smaller than or equal to the critical value, reject the null hypothesis. There is a significant difference between the two data sets.

If the smaller U value is greater than the critical value, accept the null hypothesis. There is no significant difference between the two data sets.

E.g. As 36.5 is greater than 20, the null hypothesis is accepted. There is no significant difference between the two data sets.



Location quotient

The location quotient (LQ) is used to determine the **spatial distribution** (the extent of clustering/dispersal) of a phenomenon in a **subset** of data compared to the **total** data, for instance the concentration of an industry in a **region** compared to the **nation**. They are often used in demography, economics and any type of locational analysis.

1. Find the **proportion** of subset and the **total** with the phenomenon observed
2. **Divide** the **proportion** of the **subset** by the **proportion** of the **total**

E.g. Ethnic diversity- Proportion of people who are White British in England's regions

Region	White British Population	Total population	<i>Proportion</i>
South West	510800	536000	<i>95.3%</i>
England	42279236	53010000	<i>79.8%</i>

$$\frac{95.3}{79.8} = 1.19 \quad LQ = 1.19$$

3. Interpretation of location quotient results:

If the LQ is **greater than 1**, this indicates a **high spatial concentration** for that subset compared to the total set.

If LQ = 1, the share of the total is **equal** for the subset and the total set

If the LQ is **less than 1**, this indicates a **low spatial concentration** for that subset compared to the total set.

As the LQ of the proportion of people of people who are White British in the South West is greater than 1, this indicates a higher concentration than the average for England. The South West has low ethnic diversity.



Critically Analysing Data

There is no set method for analysing data, but it is important that within your analysis you include:

- How the data shown links to your **hypotheses** or **sub-questions**.
- **Thorough** analysis. Comment on what your data actually **shows** about the subject you are investigating, such as patterns or frequent opinions.
- **Quoted** numerical data and qualitative data rather than **only** discussing overall trends.
- **Interrogation** of data. Ensure that **trends** shown within the data presentation have been discussed **thoroughly** and **clearly**. Do not leave any **gaps** in your analysis, e.g. **do not ignore anomalies or points that disprove your hypotheses**.
- Comments on the **accuracy** of your data. Comment on how **precise** your data is as this will make your conclusions more **believable** and **confident**. For example, comment on the **degree of accuracy** of your graph (e.g to 4 significant figures) so that you can prove your data is not missing harder to spot trends.
- Comments on the extent to which your data is **representative**. If you have used a lot of **investigation sites** - for example - then comment on this, because it shows your data represents your locational context realistically and wholly.
- Links to the **theory** behind your data. Give reasons as to why **data patterns** have arose using geographical theory. The purpose of the investigation is to **extend your geographical understanding** so show that it has been extended.

Using Context

It is important that, within your investigation, your **deductions** from your data are **supported** with **geographical theory** and **locational context**.

Rather than just **describing** trends, you should **explain why** these trends occur. This may include **geographical context** from your **exam specification**, or it could be **wider knowledge** that you have **researched** (and referenced). Geographical theory is important as it proves that your **conclusions have a valid reasoning** behind them. For example, if you were investigating why there are more wildfires in an area of Manchester than in an area of northern Scotland, explain your data using theory of how climate affects wildfire prevalence.

Locational context is just as important to include in your analysis, as your location's **external and internal factors** will impact upon your data. For example, if your location is close to a coal burning factory, you could explain that the poor air quality in your Environmental Quality Survey (EQS) is most likely due to this. Locational context could also be useful for **explaining trends that do not match your hypothesis**. An external factor could cause the **geographical theory to not correlate with your data trends**, so rather than ignoring this, you could explain how a locational factor could cause these trends.



How to Write an Analysis

It is recommended that you analyse in **hypotheses order** rather than **presentational technique order**. Using this structure, **one** hypothesis is analysed first using all of your different sets of data, then **another** hypothesis is analysed. This may lead to **repeats** in your **figures** if the data overlaps into different hypotheses, but you can always **reference** the figures later on in your analysis even if you have included it in a prior paragraph (e.g. see Figure 9).

Example analysis

In this example enquiry, the student is investigating **how deforestation in the Carlisle area could have been a contributing factor to 2018 flooding in Carlisle**. Here is a **brief** example of an acceptable way to set out analysis of data. For clarity, only one figure is analysed.

Hypothesis 2: Deforestation in Carlisle causes a surplus of water in the drainage basin.

Clear link to hypothesis.

General comment on relationship.

Explaining trend using data from figure and manipulating data.

Reasons behind anomalies using geographical theory.

In order to conclude **whether there was a surplus of water in the drainage basin**, the saturation of the soil was measured in 11 different sites, and the trees in a 100m² were counted. This soil saturation reading was taken from an **average of 10 separate readings**, meaning the data is an **accurate representation** of soil water content. **Figure 1 shows the saturation content has a clear relationship with the number of trees**. In **general**, as the number of trees increases, the saturation content of the soil decreases, **showing that the trees decrease the saturation content in the soil**, therefore possibly showing that **deforestation increases saturation content** (which may cause flooding). For example, where there were **2 trees** the saturation content was high at **70%**, however where **23 trees** were present, the saturation content was only **19%**, showing a **51% decrease** in soil saturation from 2 trees to 23 trees. There is a **disruption to the pattern** at 8 trees, because the soil saturation content is only 32% which is unusually low. However, this may be because the **soil content is less permeable clay and the relief is around 17° steeper than the other sites (secondary source author, date)**, perhaps giving reason as to why there is less infiltration in this area. **Overall**, Figure 1's negative correlation between the number of trees and soil moisture content proves that in areas of Carlisle, the lack of dense forestry causes water surplus.

Accuracy comment.

Graphical trend.

Identifying anomalies.

Summary.



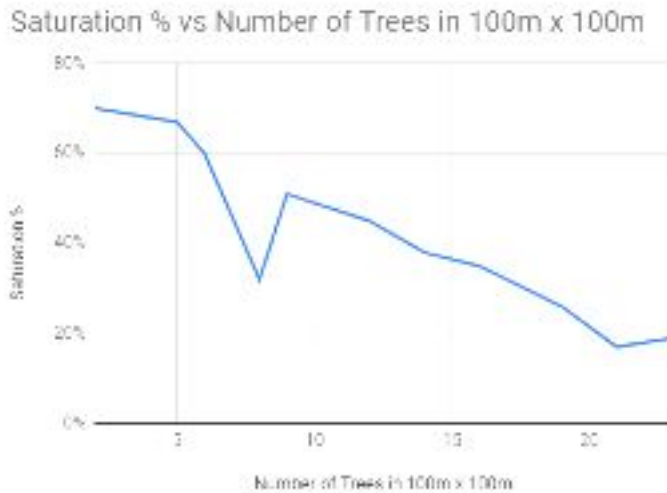


Figure 1 - A graph showing saturation content of soil vs number of trees in 11 sites in Carlisle.

You would be required to talk about **more than one figure**, and higher level students could talk about the figures **interchangeably** to spot more trends. For example, a candidate could have performed a tree count and referenced areas of deforestation within their analysis of Figure 1 to prove the areas of water surplus were in areas of deforestation.

