July 2020

Politics, Statistics & Me

Alexandra Anderson
Todd Hartman

Q-Step, University of Sheffield

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE

Nuffield Foundation

The University Of Sheffield.

Why Statistics?

The world around us is governed by politics and affects nearly every aspect of our lives. Parliament is responsible for making and changing laws, the Government decides how the country is run, sets taxes, and chooses how to spend money and deliver public services, such as the NHS or the police. Local governments are responsible for functions such as social care, schools, and housing.

Often, statistics are used to inform these different bodies on how best to govern, and they are also used to inform us, the public, about what is going on in politics. Statistics are therefore powerful tools that help us understand patterns of behaviour, explain past events, or predict possible futures.

Overview

This series of activities uses statistical concepts to explore political issues, both past and present, helping students develop their vital critical thinking and analytical skills. In six thematic chapters, this course will cover key statistical concepts such as descriptive statistics, visualizing data in graphs, charts, and maps, explore how surveys are conducted and the problem of bias.

These activities are best conducted in order, as the 5th chapter will inform the activities in the 6th.

With these activities, the aim is to support and empower students to become active, informed, and engaged citizens in their knowledge of statistical knowledge and their application to everyday politics.

Table of Contents

Here is are the main themes will be exploring and where to find them in this document:

Key Statistical Terms ....p. 04

1. Electoral Turnout: How Many People Decide to Vote? (Descriptive Statistics) ....p. 08
2. Women MPs in the UK Parliament (Graphs and Charts)..................p. 22
3. The UK's Main Political Parties (Scatter plots and line graphs)........p. 34
4. Mapping Politics (Spatial Data Analysis)...p. 46
5. What is the Public Thinking? (Polling and Surveying).........................p. 56
6. Media, Bias and Politics (Bias)...p. 65

Key Statistical Terms

A **bar chart** uses rectangle bars with different heights, or lengths, to represent and display categories of data with rectangular bars where the heights are proportional to the values that they
represent.

In statistics, **bias** occurs when the results differ from the true quantitative parameter being estimated.

Cartogram is a type of map which distorts the area (typically shown in a Mercator or Gall Peter's projection) in order to demonstrate how the data is equally represented.

Causation is what we know as a 'cause and effect' relationship between two variables. It is the change of one variable that we can define as changing another variable.

The **Choropleth map** will distinguish predefined areas by using different types of shading, colouring or symbols that represent the average values of a particular quantity in those areas.

Compound questions (or double-barrelled questions) is when a question is asking about more than one topic or theme - so it is asking two distinct things but only allows for a single answer.

A **convenience sample** is when participants are chosen for accessibility and availability.

Correlation informs us how strongly a pair of variables are linearly related and change together. It only tells us that a relationship exists, but does not tell us the how or why.

"Correlation does not imply causation" means that essentially, just because two things appear to be related does not mean that one causes the other.

Exclusion bias happens when specific individuals are excluded from the population sample of a study.

A h**exagon** map presents individual areas with a hexagon for a boundary. Hexagram maps can offer clarity in the way they standardize geographic spaces, due to the ability for the hexagon to tessellate well.

A **histogram** is similar to a bar chart, but instead of showing categorical data, its columns represent a continuous quantitative variable (e.g. age). It uses vertical columns to show frequencies (how many times each score occurs) and it will not have spaces between the columns because the data is continuous.

A **leading question** is when a question steers a participant to a desired response.

A **line chart** is a type of graph that displays data as a series of data points called 'markers' which are connected by straight line segments. The line shows how something changes in value (often this change is time).

The **line of best fit** is a line drawn through the dots in a scatter plot that best expresses the relationship between the points.

A **loaded question** makes presumptions about a participant's experience, forcing a response in a predetermined way.

The **mean** is the average number in a data set. It can be found by adding all the data points together and then dividing by the number of data points.

The **median** is the middle number in a data set and it can be found by ordering all the data points and finding the one number in the middle. Sometimes there are two middle numbers, and the median would be the mean of those two numbers.

The **Mode** is the most common number in a data set. This is the number that occurs the highest number of times.

Non-structured questions - also often called open-ended questions - are questions where the participant is asked to write/say their response to a question, and they are not given a list of choices.

Non-probability sample is people who were selected based on non-random criteria.

Observer bias arises when the researcher subconsciously influences research due to their own expectations, which can alter how a survey is carried out, or how the results are recorded.

A **partially structured question** uses a list of choices with the option to include an unstructured answer.

Pie charts are used to compare parts of a whole, not the difference between groups. By using pie slices in a circle we can demonstrate the relative sizes of data and how they add up to the whole. Often, this data takes the form of a percentage.

Polls normally only ask one or two questions with limited analysis of data, and are often employed in political races to gather immediate feedback and predict an outcome.

The **population** is the whole, and would consist of every member in a group. A population may refer to an entire group of people but it can also refer to a group of events, objects, measurements.

A **population sample** is the group of individuals who are selected to participate in a study.

Probability is the likelihood that an event will occur.

The **range** is the difference between the lowest and highest values.

A **ranking question** uses a list to asks participants to demonstrate how they feel about something by comparing it to others.

A **rating question** asks participants to demonstrate their opinion about something by using a scale.

Recall Bias occurs when participants are unable to remember past experiences or events accurately, and memories can be influenced by subsequent experiences and events.

Reporting bias is when only certain results are chosen to be included in an analysis, leaving out relevant evidence.

Response bias (also known as survey bias) is a general term that describes the different tendencies in participants to answer untruthfully or inaccurately. For example, participants may be reluctant to give truthful answers if they are not deemed socially acceptable. This type of bias frequently happens when participants are asked to self-report (like in structured interviews or surveys) and it can also be the result of poor survey design.

Scaling and axis manipulation is used to alter the way that a scale looks for a graph.

A **scatter plot** displays points in order to present the relationship between two sets of data. The position of each dot on the horizontal (X) and vertical (Y) is a useful way of presenting data because the dots show the value of individual points, in addition to patterns in the data as a whole.

Selection bias involves choosing individuals selectively over others in a study, and therefore proper randomisation is not achieved. This means that the population sample obtained is not representative of the population intended to be analysed.

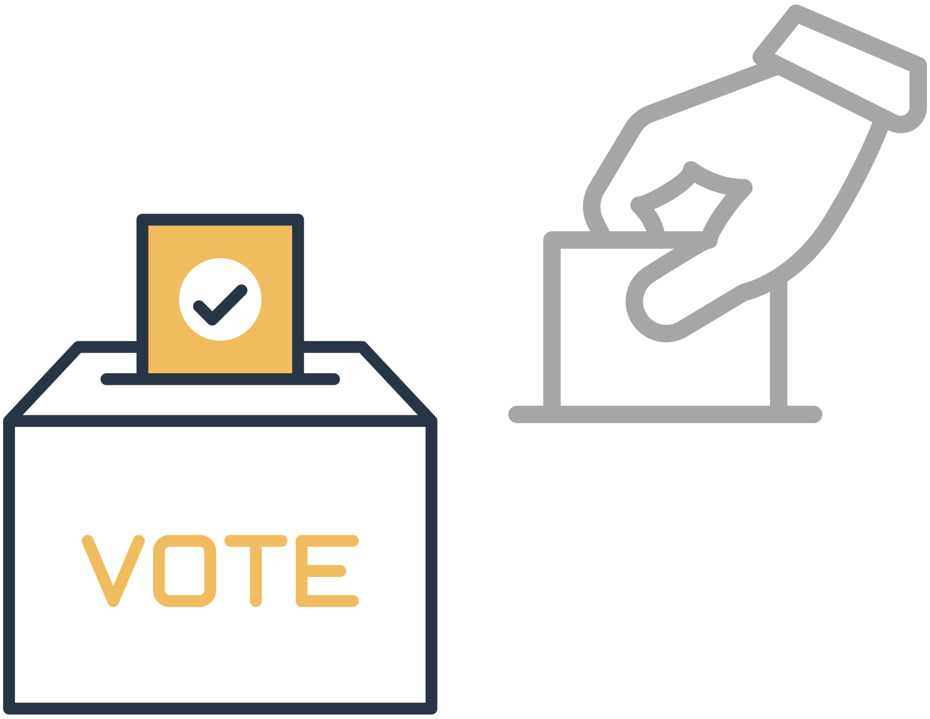
Surveys use multiple questions to gather data facts, behaviours, attitudes and preferences from a targeted population.

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE
Nuffield Foundation
The University Of Sheffield.

1. Electoral Turnout: How Many People Decide to Vote? (Descriptive Statistics)

In this chapter, we are going to use ***descriptive statistics*** *to analyse electoral turnout.*

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.



1918-2019: Who was allowed to vote?

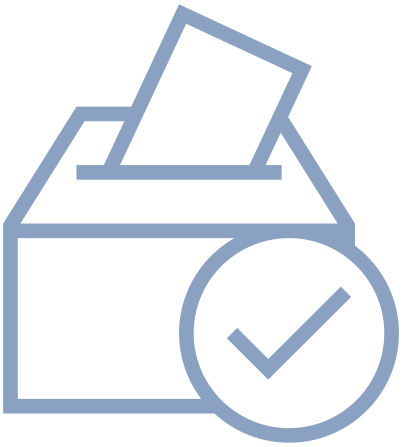
In 2018, the UK marked the 100th anniversary of The Representation of the People Act of 1918 that allowed some women, and all men, to vote for the first time.

To understand how many people decide to vote in elections in the UK, let's look at who has the right to vote:

Before 1918, women did not have the right to vote, and male voters had to be taxpaying householders aged over 21. In 1928, women aged 21 and over were granted the right to vote. In 1969, the voting age was lowered on an equal basis to 18 for both women and men.

Let's keep in mind that now (in 2020), to vote in a general election you must:

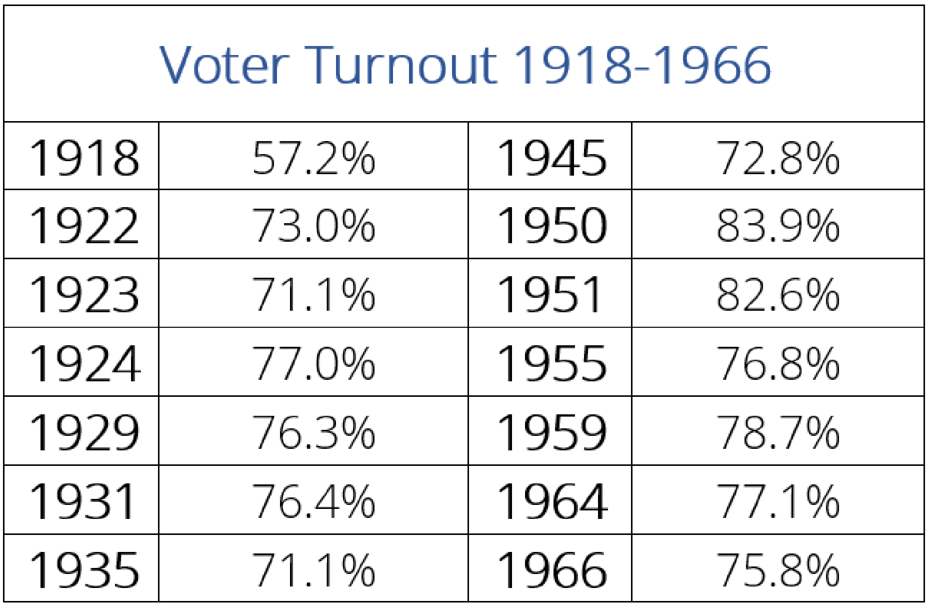
- be registered to vote
- be 18 or over on the day of the election ('polling day')
- be a British, Irish or qualifying Commonwealth citizen
- be resident at an address in the UK (or a British citizen living abroad who has been registered to vote in the UK in the last 15 years)
- not be legally excluded from voting



Activity 1: Descriptive Statistics

From 1918 to 2019, there were 28 elections. How many people decided to exercise their right to vote?

We are going to start our workshop by looking at the first fifty years, or fourteen elections, to understand how many people (who were eligible to vote) participated in each election.



Voter Turnout 1918-1966

1918	57.2%	1945	72.8%
1922	73.0%	1950	83.9%
1923	71.1%	1951	82.6%
1924	77.0%	1955	76.8%
1929	76.3%	1959	78.7%
1931	76.4%	1964	77.1%
1935	71.1%	1966	75.8%

1A: Mean, Median and Mode

Mean, median, and mode are different measures of the center in a numerical data set. They each try to summarize a dataset with a single number to represent a "typical" data point from the dataset.

The **Mean** is the average number in a dataset. It can be found by adding all the data points together and then dividing by the number of data points.

Let's do this with the above numbers. This is what it would look like:

First step - add up all the numbers:

57.2 + 73.0 + 71.1 + 77.0 + 76.3 + 76.4 + 71.1 + 72.8 + 83.9 + 82.6 + 76.8 + 78.7 + 77.1 + 75.8 = 1049.8

We then divide this number by the number of data points. for this set, there are 14 data points.

1049.8 ÷ 14 = **74.99**

So the Mean is 74.99%

The **Median** is the middle number in a data set and it can be found by ordering all the data points and finding the one number in the middle. Sometimes there are two middle numbers, and the median would be the mean of those two numbers.

Let's use the above numbers again, starting with putting them in numerical order:

57.2 / 71.1 / 71.1 / 72.8 / 73.0 / 75.8 / 76.3 / 76.4 / 76.8 / 77.0 / 77.1 / 78.7 / 82.6 / 83.9

In this set of numbers, there are two in middle. So we say that the median is the mean (average) of the two:

(76.3 + 76.4) / 2 = **76.35**

The **Mode** is the most common number in a data set. This is the number that occurs the highest number of times.

Have a look at the numbers again - are there any that occur more than once?

71.1 % occurs twice, so that is the mode.

When to use mean, median and mode?

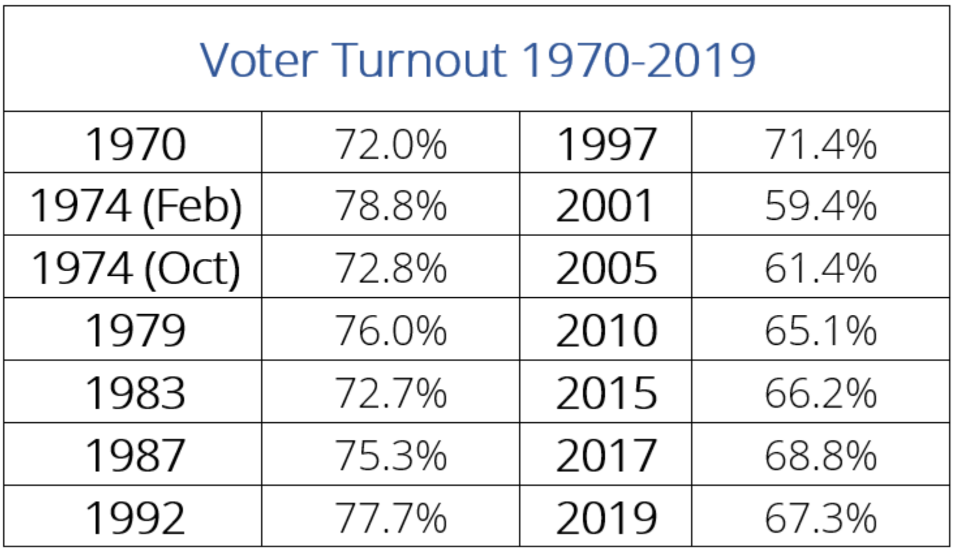
There are different ways to summarise the data that are beneficial to a specific data set.

For example, the mean is more sensitive to extreme outliers and can be pulled in either direction. A good example of this is to think about at house prices that are used to describe a neighbourhood:

> *If Mark Zuckerberg lived in your neighbourhood, and enlarged the house and expanded into the property around him, then the mean and median home values will look very different because of that house's affect on the overall property values. Which dispersive statistic do you think would give you a more accurate reflection of the neighbourhood?*

1B: Let's try it ourselves

Let’s calculate the mean, mode and median again, this time with the election turnout from the past 50 years:



Voter Turnout 1970-2019

1970	72.0%	1997	71.4%
1974 (Feb)	78.8%	2001	59.4%
1974 (Oct)	72.8%	2005	61.4%
1979	76.0%	2010	65.1%
1983	72.7%	2015	66.2%
1987	75.3%	2017	68.8%
1992	77.7%	2019	67.3%

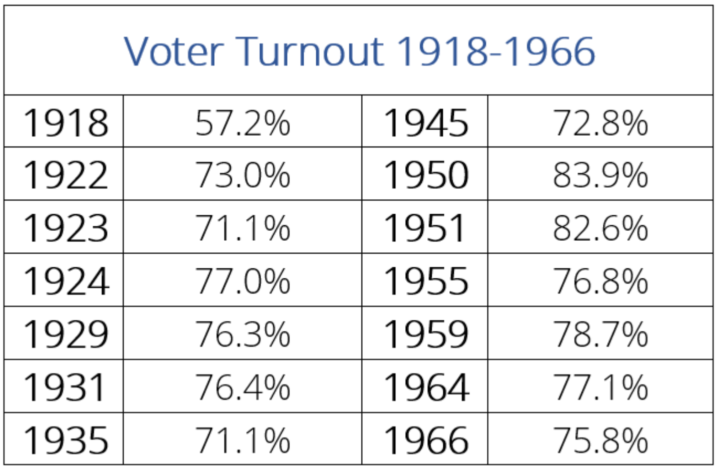
What is the **mean**?

What is the **median?**

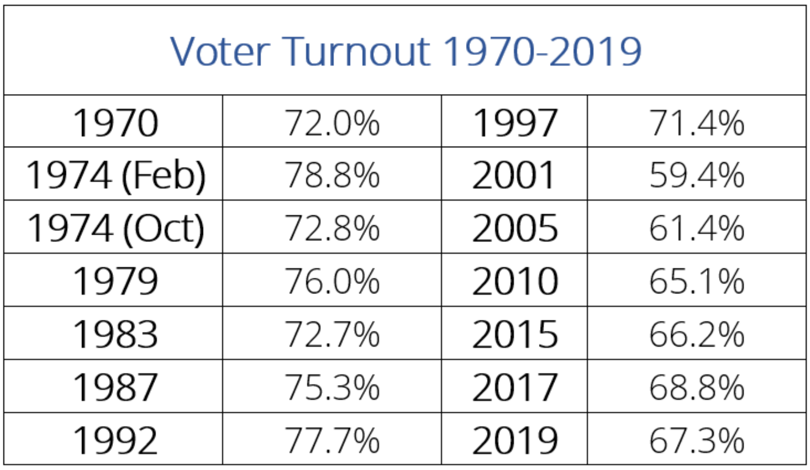
What is the **mode**?

1C: Let's Analyse the full 100 years

Let’s calculate the mean, mode and median for the combined datasets from 1918-2020.



Voter Turnout 1918-1966			
1918	57.2%	1945	72.8%
1922	73.0%	1950	83.9%
1923	71.1%	1951	82.6%
1924	77.0%	1955	76.8%
1929	76.3%	1959	78.7%
1931	76.4%	1964	77.1%
1935	71.1%	1966	75.8%



Voter Turnout 1970-2019			
1970	72.0%	1997	71.4%
1974 (Feb)	78.8%	2001	59.4%
1974 (Oct)	72.8%	2005	61.4%
1979	76.0%	2010	65.1%
1983	72.7%	2015	66.2%
1987	75.3%	2017	68.8%
1992	77.7%	2019	67.3%

What is the **mean**?

What is the **median**?

What is the **mode**?

Part 2: When was voter turnout the lowest, and when was it the highest?

*How can we check that the mean, mode and median that we found above really represent the data? Here we are going to measure the **Range** to check how spread out our numbers are.*

Use the Voter Turnout Tables from previous pages for this activity.

The **Range** is the difference between the lowest and highest values.

So for the percentages from 1918 to 1969, the lowest number is 57.2% and the highest number is 83.9%. The difference between the two is 26.7%.

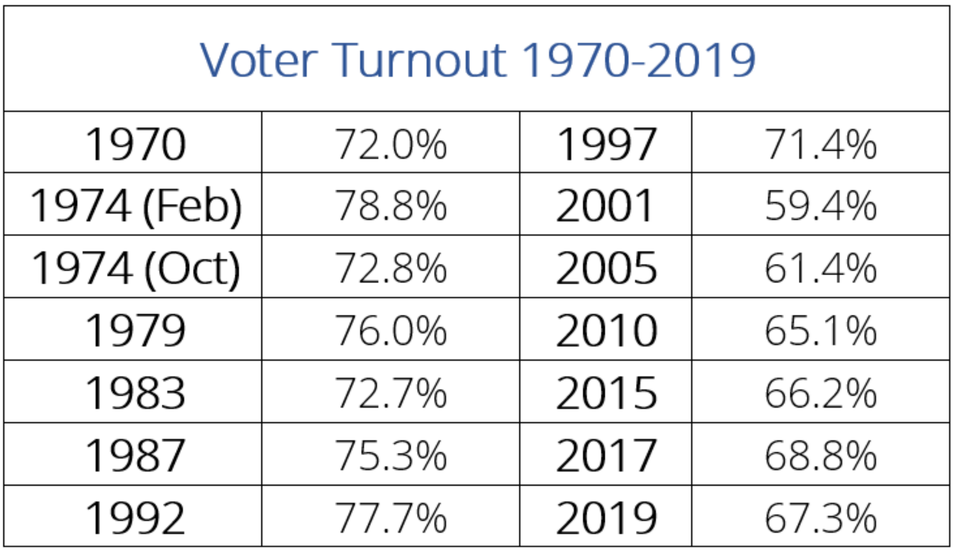
So the range is 26.7%

1) What is the **range** from 1970 to 2019?

2) And what is the **range** for the total numbers from 1918 to 2019?

1) What is the **range** from 1970 to 2019?

2) And what is the **range** for the total numbers from 1918 to 2019?



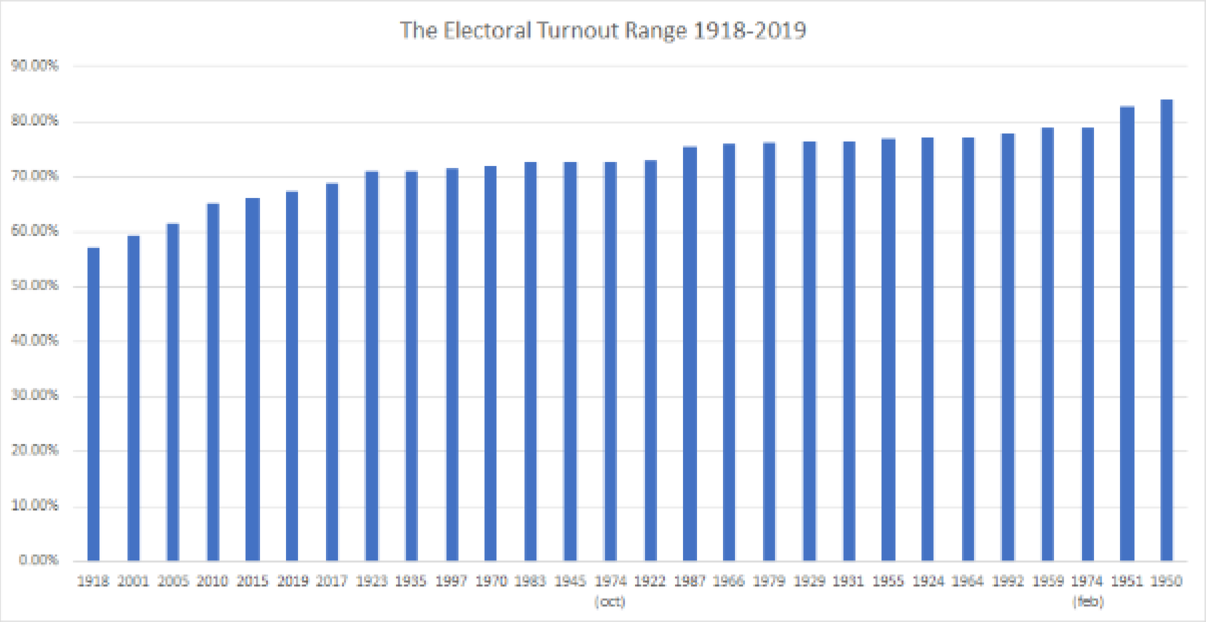
Voter Turnout 1970-2019

1970	72.0%	1997	71.4%
1974 (Feb)	78.8%	2001	59.4%
1974 (Oct)	72.8%	2005	61.4%
1979	76.0%	2010	65.1%
1983	72.7%	2015	66.2%
1987	75.3%	2017	68.8%
1992	77.7%	2019	67.3%

What do you think the range is telling us about electoral turnout for each set of numbers?

Let’s think about how the range is useful - for example, what if the electoral turnout in 1966 was 5%, and in 1924 it was 95%? The range tells us what the mean, median, and mode don’t: essentially, how spread out our numbers are. Why do you think this is important?

Here is a visual representation of the range from 1918 to 2019 to help you visualise these numbers:



Part 3: Final Discussion

After looking at the descriptive statistics (the mean, mode and the median), and a dispersive statistic (the range) here are some questions to contemplate:

3A: What do these descriptive tell us? And what do they not tell us?

1 What do you think are the positives and negatives of each method?

2 For this data set about electoral turnout over the past 100 years, which statistic (mean, mode, median) do you think worked best?

3 How did these statistics help us understand electoral turnout over the past hundred years?

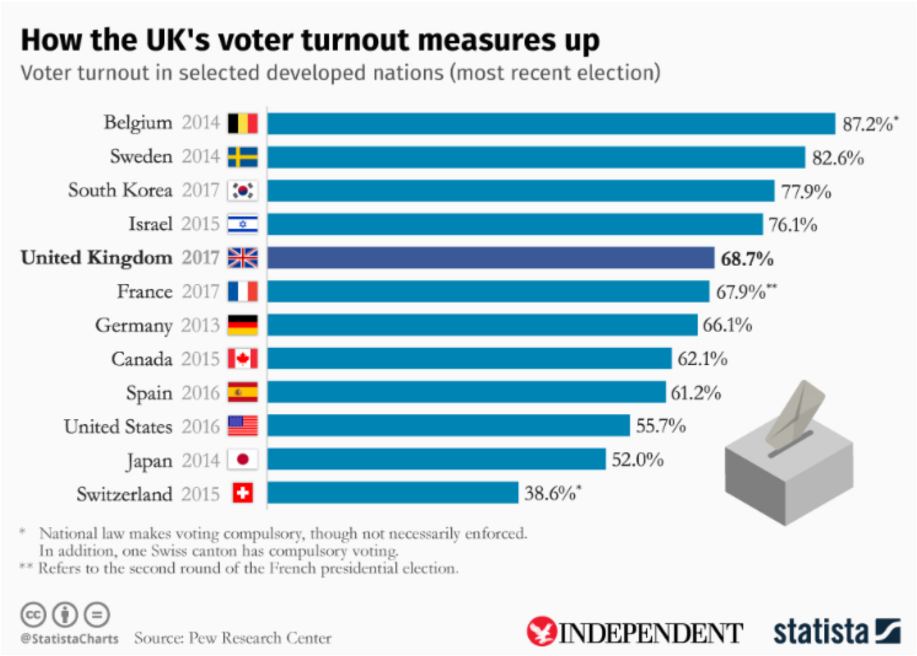
3B: Voter Turnout - How does the UK compare?

In the past, many people have fought for your right to vote (when you turn 18). Many campaigners, such as the women's suffrage movement or those who took part in the 1819 Peterloo protests and subsequent massacre, gave their lives to win the right to vote.

Through voting for our elected representatives, we get to have a voice in how our country is run. In the UK, the MPs we elect make decisions about taxes, education, housing, the NHS as well as the environment.

Not everyone who is eligible to vote actually does. There are multiple reasons for this: sometimes they are interested in politics, or do not like any of the candidates, or are worried they are not well enough informed. Some countries have compulsory voting, like Australia and Belgium. And even though voting is mandatory, does not infringe on their rights as they are allowed to submit a blank ballot at a polling station rather than vote for a candidate.

So how does the UK compare? The bar graph (a statistical concept we will explore in the next chapter) below will provide some useful information.



* National law makes voting compulsory, though not necessarily enforced. In addition, one Swiss canton has compulsory voting.
** Refers to the second round of the French presidential election.

Source: Pew Research Center

Source: Statista

3C: Here are some more specific questions for discussion:

1 Why do you think people should vote, and are enough people voting?

2 Do you think the criteria for eligibility to vote affect who goes out to vote? What kind of extra data would be helpful in understanding this?

3 How does the UK's voter turnout compare to other countries around the world? Use the graph on the previous page to compare. Many countries that do not have a compulsory voting system still get strong turnouts. For example, Sweden had a voter turnout of 82.6% in 2014, South Korea had 77.9% in 2017, Israel had 76.1% in 2015 and New Zealand had 75.7% in 2017.

4 Given the percentage of people who have decided to vote in the last hundred years, do you think there should be compulsory voting in the UK (like in Australia or Belgium)?

5 If the voting age was lowered to 16 years old, do think it would affect the percentage of people who participate in elections? Why or why not?

2: Women MPs in the UK Parliament (Graphs and Charts)

*In this chapter we are going to start our exploration of **Data Visualisation** by exploring who we vote for in the UK, both currently and in the past.*

Data Visualisation is how we represent information (data) in a visual context. This can include bar charts, histograms or pie charts and today we will explore all three. By using data visualisations, we make it easier to detect trends, patterns, and outliers, making it easier for our brains to process both big and small data.

Graphs are used to present data, or information, as clearly and understandably as possible. Some types of graphs are more useful for some types of data - and today we are going to understand why.

This chapter consists of 3 activities where we will explore and create different types of graphs and charts, using data about female representation in UK and EU elections.

Since 1918 who have we been voting for at General Elections?

From 1918, if you were at least 21 years old, both men and women could be elected to Parliament.

The first female to sit in the House of Commons was Nancy Astor in the 1919 by-election. Before Astor, Constance Markievicz was the first female MP elected, but along with other Sinn Fein MPs, did not take her seat.

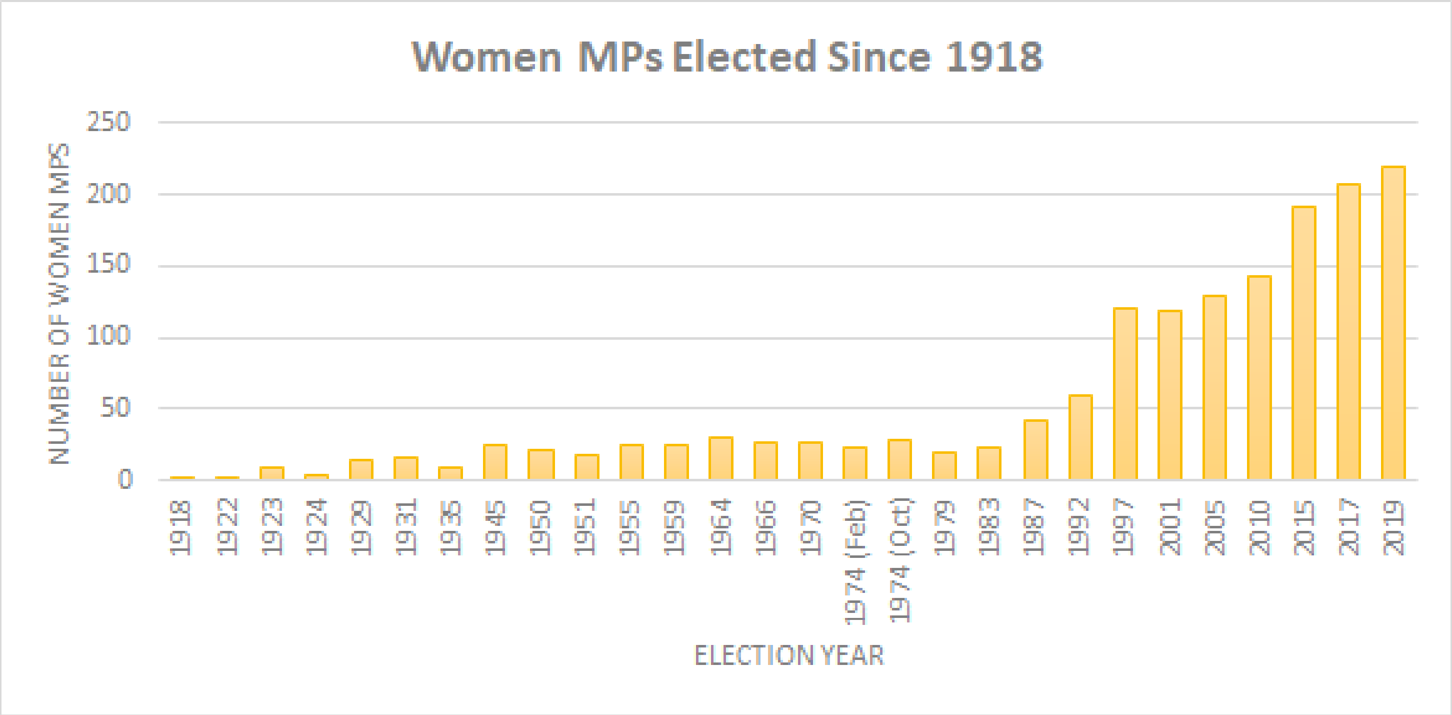
Women never held more than 10% of seats until 1997. What does female representation look like now? We will explore all of this further in the following pages.

To see how the number of female MPs has increased over time, we are going to use some data visuals.

Activity 1: The Bar Chart

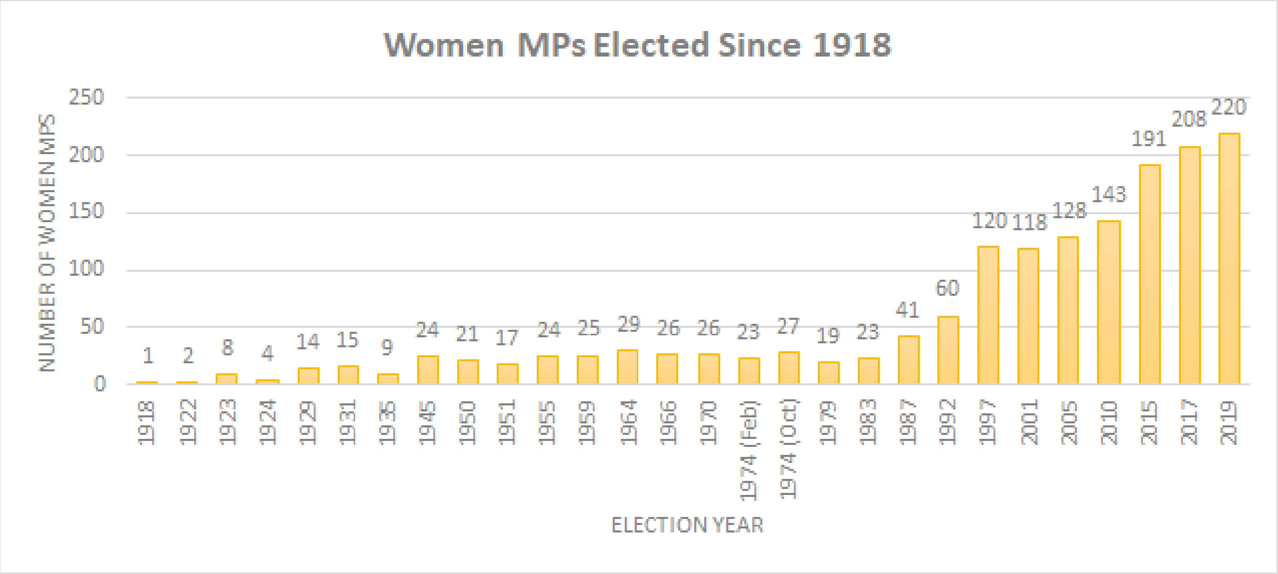
First, let's start with a bar chart. **A bar chart** uses rectangle bars with different heights, or lengths, to represent and display categories of data.

A bar chart has a vertical axis (also known as the Y axis, and is the line going top to bottom) with numbers on it, and a horizontal axis (also known as the X axis, and is the line going side to side) that show the categorical data with rectangular bars where the heights are proportional to the values that they represent. In the example below, the categorical data is the election year, and the vertical axis represents the amount of women MPs elected:



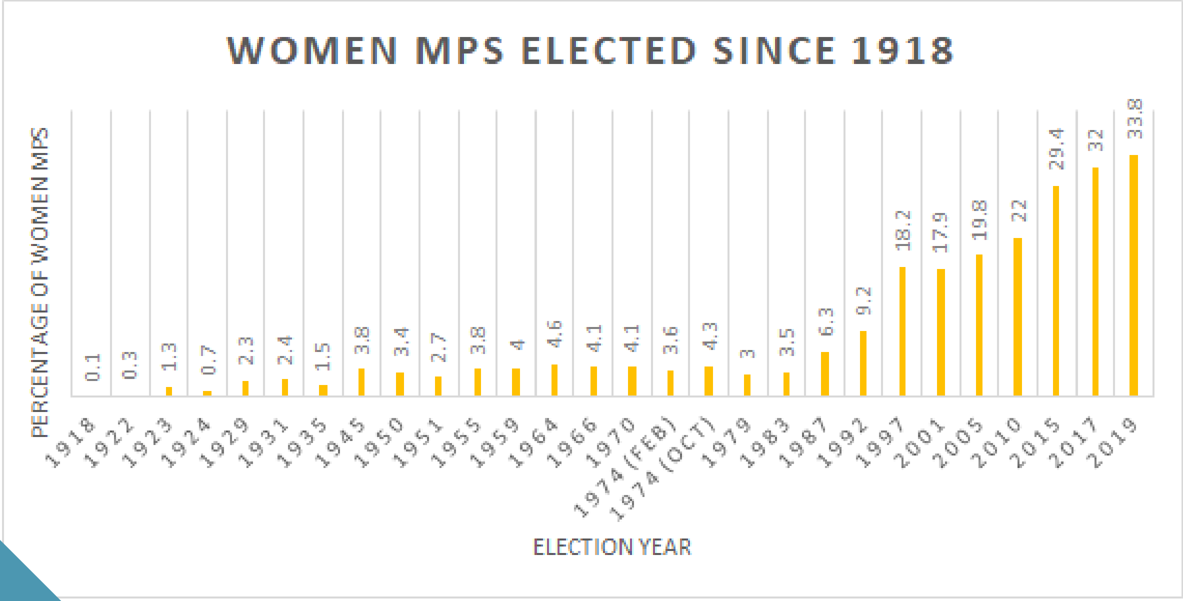
Looking at this graph - what information is missing that you think would be useful.

For example, what if we added the exact data for each bar?

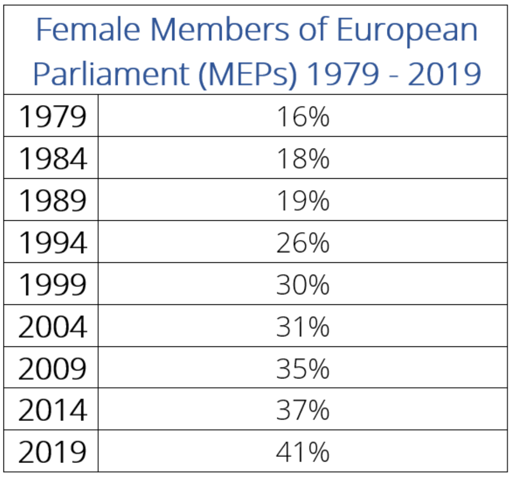


Also - how would knowing the total amount of MPs for each year affect how you view the graph? Perhaps it would be better to have a percentage?

For example, even though the 2001 General Elections saw the election of 118 female MPs, it is hard to tell what this means if we do not know what this number is out of the total. Here is how a graph showing the percentage of female MPs, out of total MPs, would be displayed to display this type of data:



Here is the data for Members of European Parliament (MEPs) gender balance since 1979.



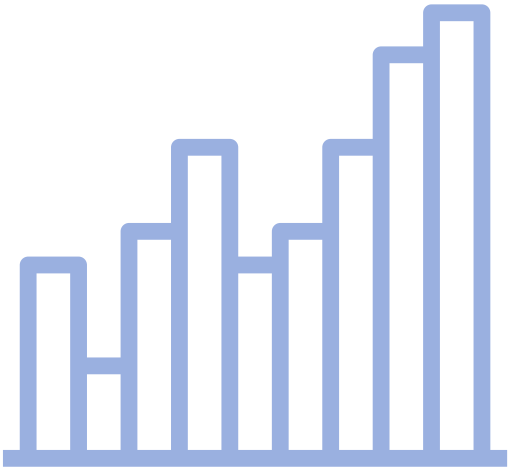
Female Members of European Parliament (MEPs) 1979 - 2019	
1979	16%
1984	18%
1989	19%
1994	26%
1999	30%
2004	31%
2009	35%
2014	37%
2019	41%

Create a bar chart to demonstrate the data *(feel free to draw this by hand)*:

Now compare your bar chart to the previous charts. How does the UK compare to the EU Parliament for its gender diversity?

What is missing from the data in the graph that would be helpful for your comparison?

Activity 2: The Histogram

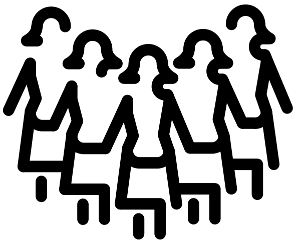


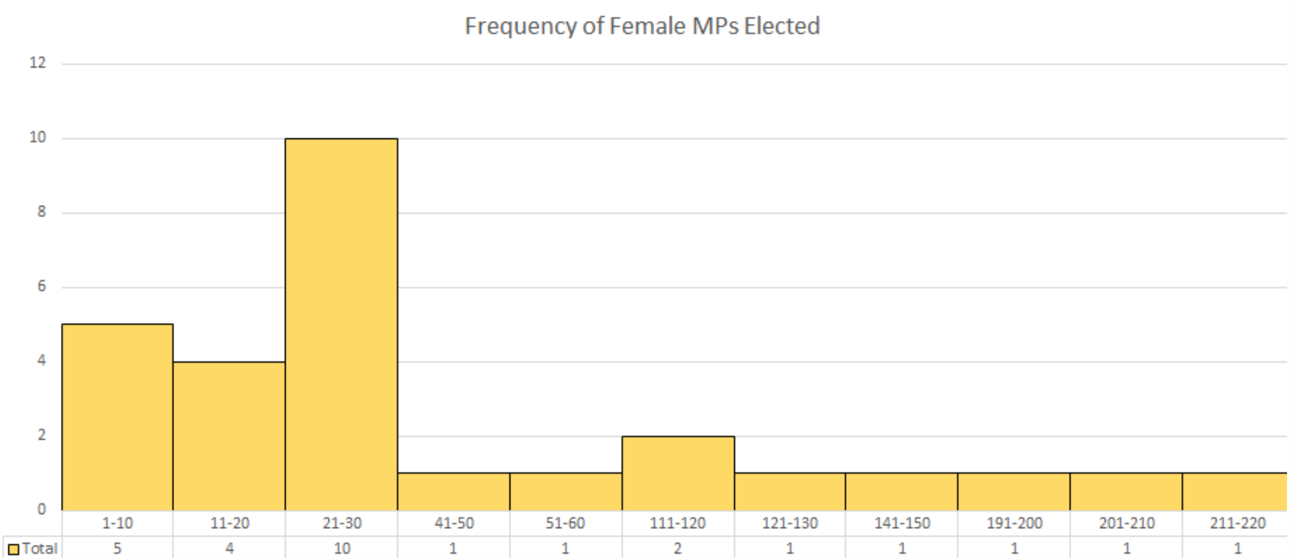
In this activity, we are going to look at the **histogram**. A histogram is similar to a bar chart, but instead of showing categorical data, its columns represent a continuous quantitative variable (e.g. age).

A histogram uses vertical columns to show frequencies (how many times each score occurs) and it will not have spaces between the columns because the data is continuous.

Let’s use the same data as the activity above to create a histogram: we are going to look at the frequency in which multiples of female MPs were elected.

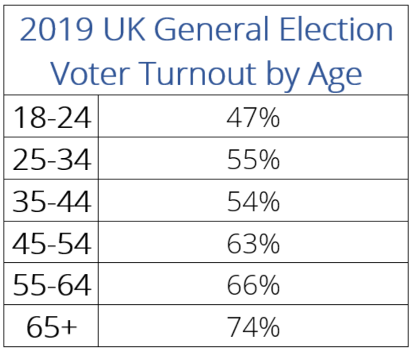
This will be a useful way for you see how data can be presented in multiple ways.





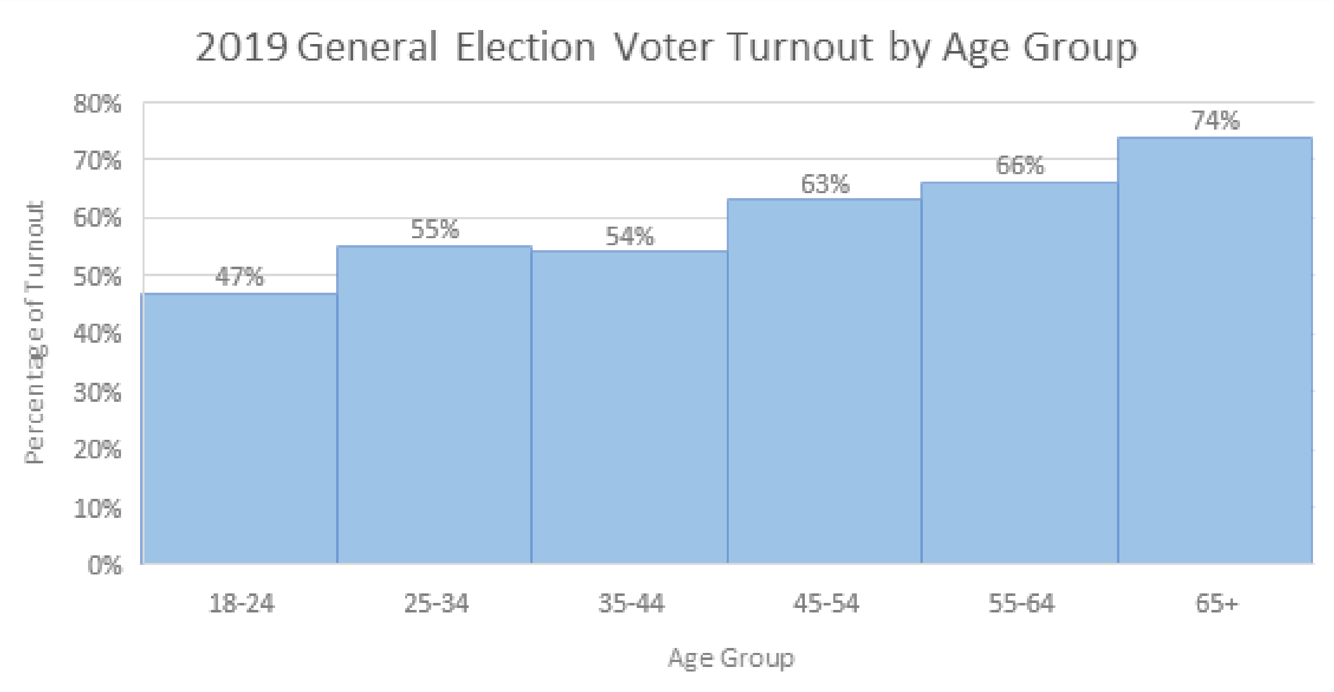
In this histogram, the frequency of women MPs elected is measured in groups that tell us how many times **1-10** Women MPs were elected, how many times **11-20** women MPs were elected, how many times **21-20** women MPs were elected and the graph measures up to how many times **211-220** women MPs were elected. From this histogram, we can extrapolate that often, over the past 100 years, low numbers of female MPs were elected. *How is this histogram a useful alternative to the previous bar charts? Notice there is a jump in the graph from* ***51-60*** *and* ***110-120*** *- why do think that is? Are there any other jumps in numbers?*

Here is another example of a histogram, demonstrating the turnout for the 2019 General Elections by age group. The turnout is the frequency being examined, and the age groups are the categorical data. Here is the data we used:

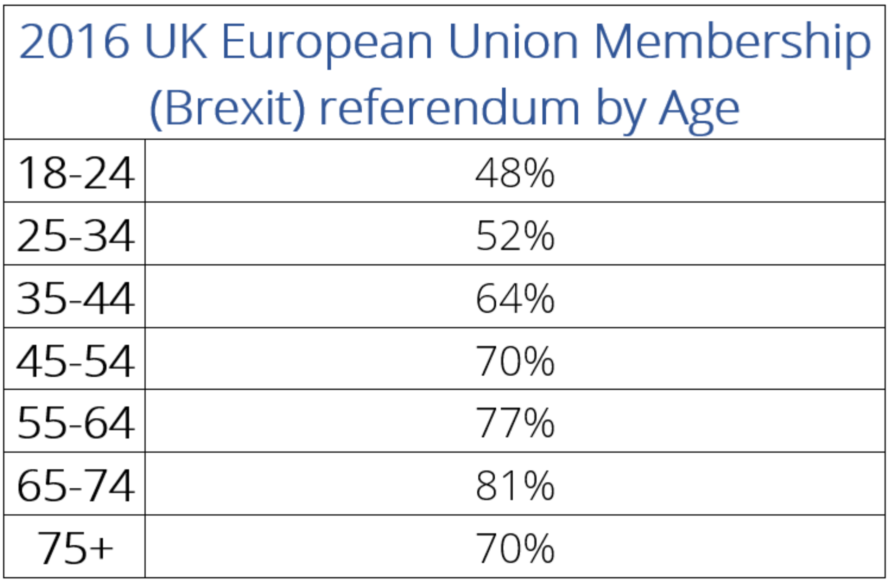


2019 UK General Election Voter Turnout by Age	
18-24	47%
25-34	55%
35-44	54%
45-54	63%
55-64	66%
65+	74%

And here is the histogram representing that data:



Now it is your turn to draw a histogram. Here is the data for the election turnout for the 2016 UK European Union membership (Brexit) referendum:



2016 UK European Union Membership (Brexit) referendum by Age	
18-24	48%
25-34	52%
35-44	64%
45-54	70%
55-64	77%
65-74	81%
75+	70%

Now create the histogram, using the example above for inspiration and guidance *(feel to draw the histogram by hand)* :

1. How does the turnout between the two events compare? How does it affect our ability to understand voter turnout for each event?

2. Why do you think the extra age bracket (age 65-74) was included in the EU Referendum data?

3. Why do you think there are larger percentage differences in some age groups and not others?

4. Why do you think some age groups appear to be more likely to vote than others?

5. Why do you think knowing the voter turnout for separate age groups is useful? For both the General Election and Brexit Referendum?

Activity 3: The Pie Chart

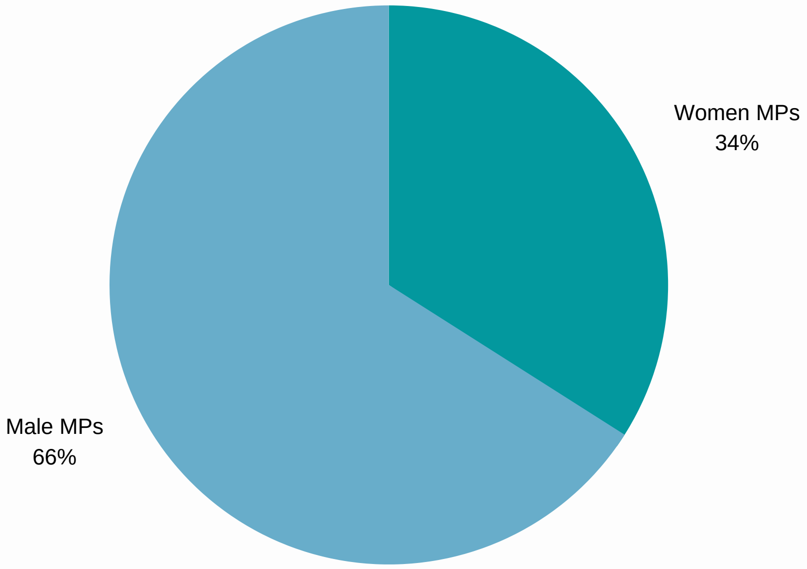
Now let’s look at pie charts.

By using pie slices in a circle we can demonstrate the relative sizes of data and how they add up to the whole. Often, this data takes the form of a percentage.

Pie charts are a useful way of showing how a total amount is divided up.

Pie charts are used to compare parts of a whole, not the difference between groups.

In the 2019 General Election, 220 female MPs were elected. This consisted of 34% of all MPs and was more than any previous election. What does this look like?



How does this compare to the General Elections in 1979, when only 11 Women MPs were elected, which was only 3% of the total MPs?

Draw a pie chart to help you visualise the difference!

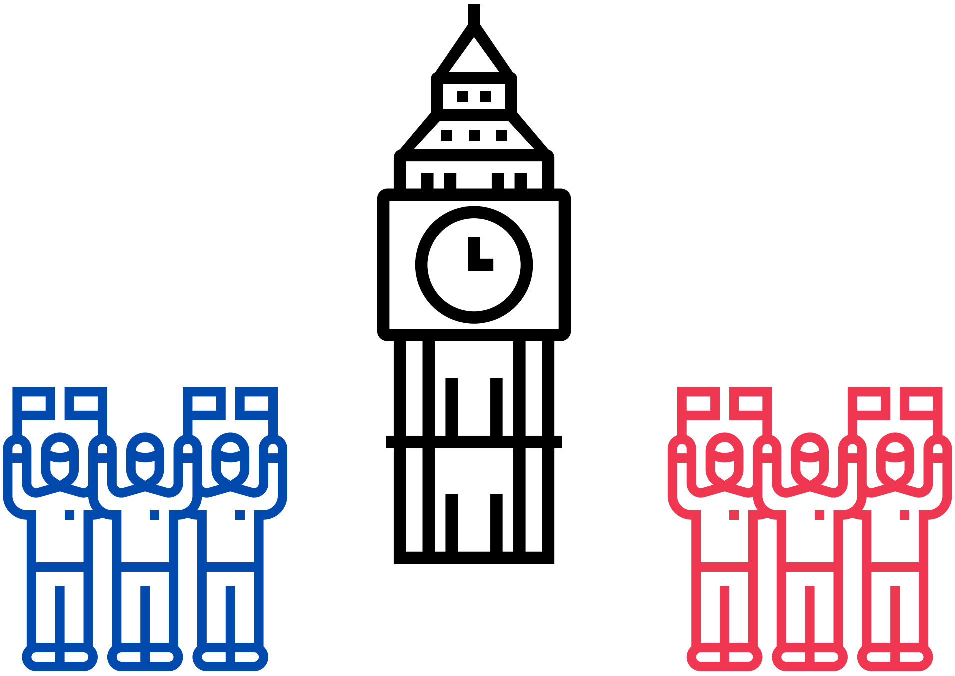
But how does the UK also compare to the current gender composition of the EU Parliament, where in 1979 Women MEPs consist of 16% of the total MEPs? And currently, where Women MEPs consist of 41% of the total MEPs?

Draw the two pie charts to help you compare:

3. The UK’s Main Political Parties (Scatterplots and line graphs)

In this chapter we are going to look at a couple more types of data visualisations: scatter plots and line graphs.

We are going to look at the UK's main political parties to understand relationships between variables.



Our Political Parties

The UK political system consists of multiple parties. Since the 1920s there have been essentially two main political parties in the UK who take up a number of seats in the House of Commons: the Conservative Party and the Labour Party.

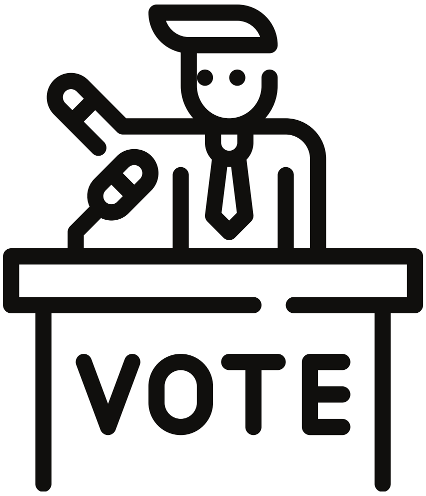
The first-past-the-post electoral system used for general elections tends to maintain the dominance of these two parties. The first-past-the-post system is the way in which we elect MPs.

How this system works is that voters in a constituency mark their choice with an X - and the candidate with the largest number of votes wins.

In the 2019 elections, the Scottish National Party, the Liberal Democrats, the Democratic Unionist Party, Sinn Fein, Plaid Cymru, Social Democratic and Labour Party, the Alliance Party of Northern Ireland and the Green Party were also elected.

The Conservative and Labour Parties

First organized in the 1830s, the modern day Conservative Party was actually an outgrowth of the Tory movement or party, which has been traced back to the end of the 17th Century. Members and supporters are often referred to as Tories and the party itself is also colloquially known as the Tory Party. The Conservative Party has been the governing party in the UK since 2010.



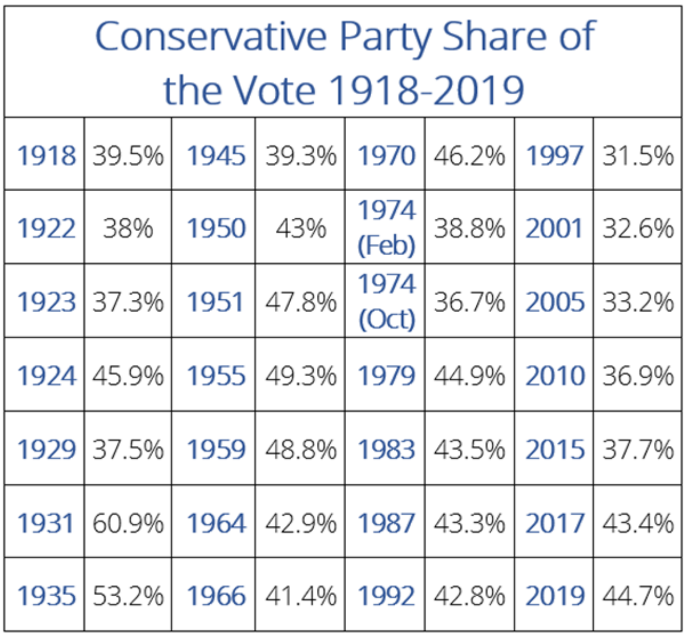
The Labour Party emerged at the turn of the 19th Century, emerging from the socialists and trade union movements of the period. It supplanted the Liberal Party which has previously been one of the two dominant parties (that included the Conservatives) in the UK. Currently, the Labour Party forms the Official Opposition to the Conservative government, as it won the second-largest number of seats in the 2019 general election.

Activity 1: Scatterplots and Line Charts

How has the UK supported these two parties since 1918? First, we are going to use scatterplots to examine voter support for the Conservative Party over the past hundred years.

A Scatter Plot displays points in order to present the relationship between two sets of data. The position of each dot on the horizontal (X) and vertical (Y) is a useful way of presenting data because the dots show the value of individual points, in addition to patterns in the data as a whole. In this example, the 'whole' is a century in time - the past 100 years.

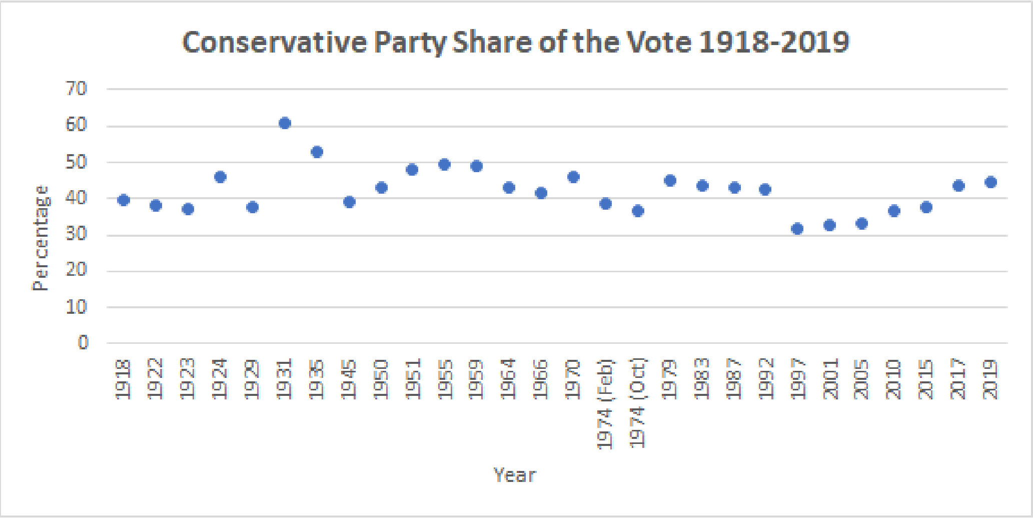
Here is the Data:



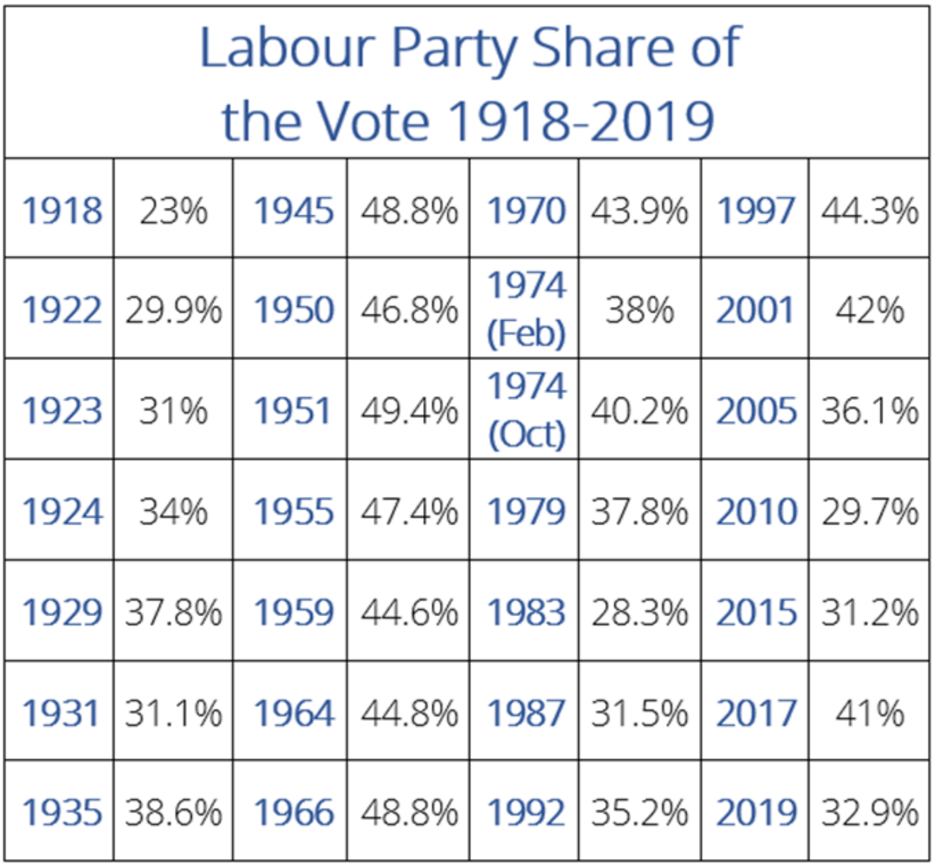
Conservative Party Share of the Vote 1918-2019

1918	39.5%	1945	39.3%	1970	46.2%	1997	31.5%
1922	38%	1950	43%	1974 (Feb)	38.8%	2001	32.6%
1923	37.3%	1951	47.8%	1974 (Oct)	36.7%	2005	33.2%
1924	45.9%	1955	49.3%	1979	44.9%	2010	36.9%
1929	37.5%	1959	48.8%	1983	43.5%	2015	37.7%
1931	60.9%	1964	42.9%	1987	43.3%	2017	43.4%
1935	53.2%	1966	41.4%	1992	42.8%	2019	44.7%

And then here is the scatter plot created from that data:



Let’s now draw a scatter plot of the voter support for Labour over the past 100 years. Use the table of data below and look at the previous scatterplot for inspiration and guidance.



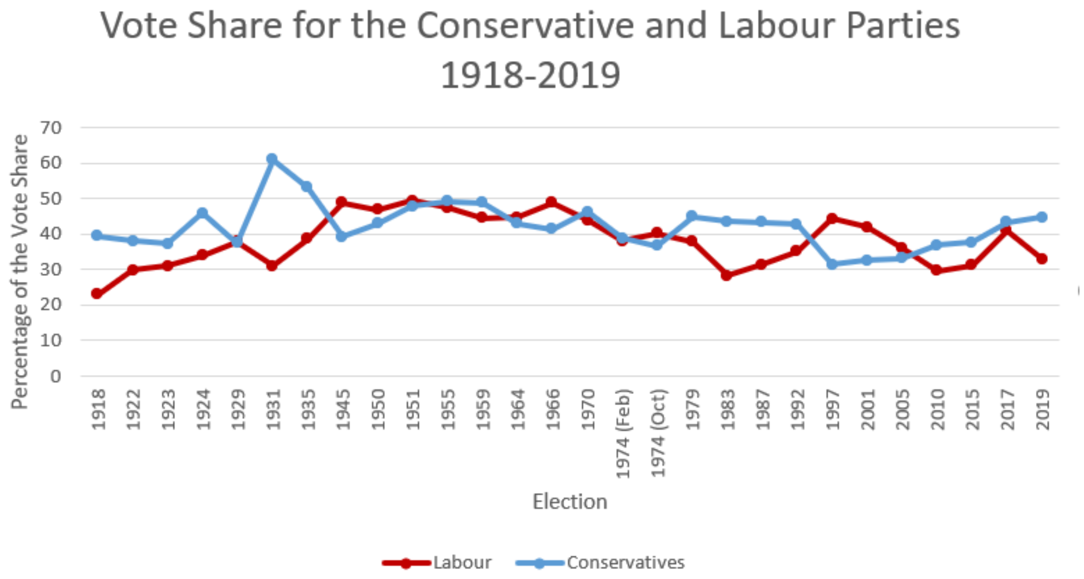
Labour Party Share of the Vote 1918-2019

1918	23%	1945	48.8%	1970	43.9%	1997	44.3%
1922	29.9%	1950	46.8%	1974 (Feb)	38%	2001	42%
1923	31%	1951	49.4%	1974 (Oct)	40.2%	2005	36.1%
1924	34%	1955	47.4%	1979	37.8%	2010	29.7%
1929	37.8%	1959	44.6%	1983	28.3%	2015	31.2%
1931	31.1%	1964	44.8%	1987	31.5%	2017	41%
1935	38.6%	1966	48.8%	1992	35.2%	2019	32.9%

What are the similar and different patterns between the two?

Using a pen, marker or pencil, connect the dots and create a line of data for both graphs above. This is how you make a line graph (or chart)! A line chart is a type of graph that displays data as a series of data points called 'markers' which are connected by straight line segments. The line shows how something changes in value (often this change is time).

Here is how they would look together on one graph, with the blue colour (Series 1) representing the Conservative share of the vote, and the red colour (Series 2) representing the share for Labour. Drawing the line helps us understand the past relationships between the data.



Why do you think having both sets of data on the same graph useful?

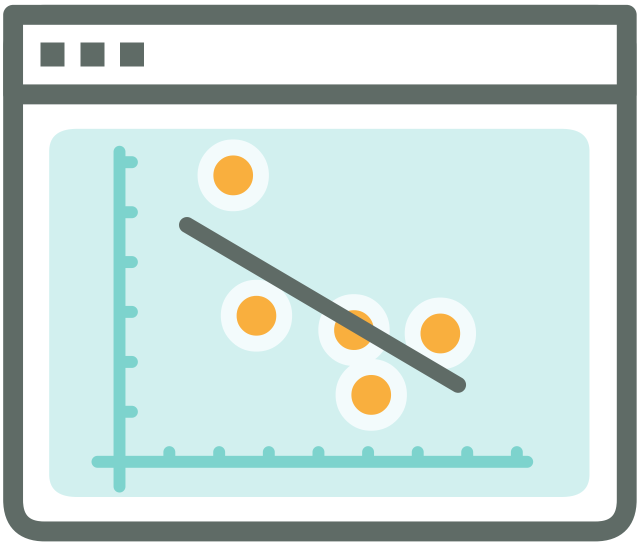
So why not present the data in a bar graph? In a bar graph, you use rectangular blocks to represent many different types of graphs. A line in a line graph will show one type of data, and is particularly useful for showing trends over a period of time.

Activity 2: What Can we do with these Points of Data?

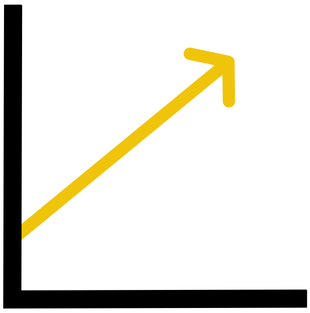
In order to identify the relationship between the points of data, instead of connecting the dots, drawing a '**line of best fit**' would be helpful.

The line of best fit is a line drawn through the dots in a scatterplot that best expresses the relationship between the points.

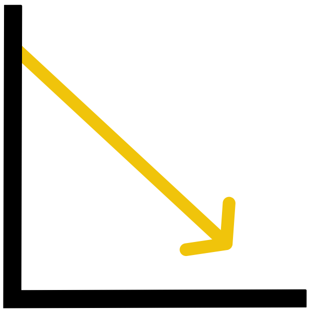
Social scientists will often use **linear regression analysis** to create a straight line that results from the relationship between the two variables. Was does this look like?



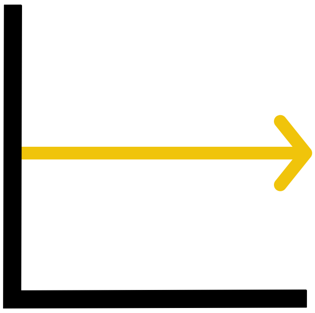
A positive relationship between X and Y is when the data shows an uphill pattern from left to right. As the X-value increases, so does the Y value. The data is moving right and up.



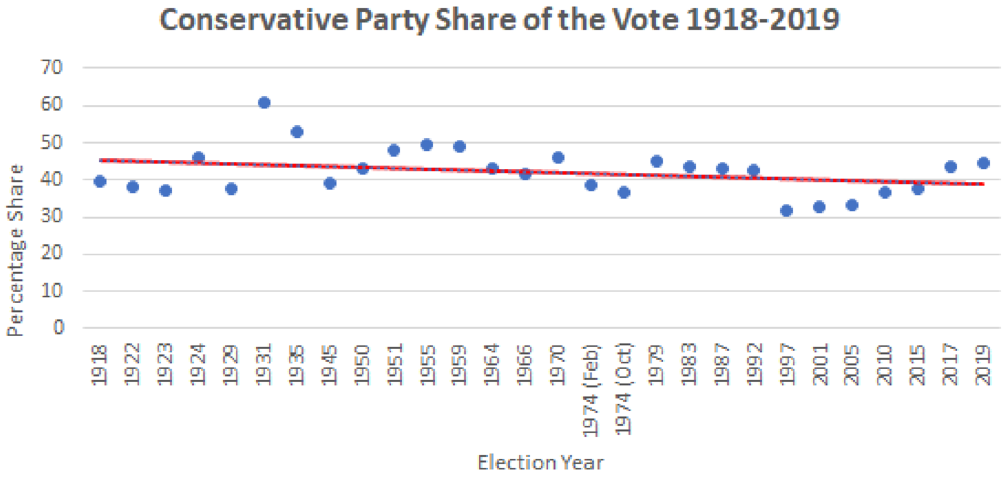
A negative relationship will show a downhill pattern from left to right, where Y value decreases and moves down as the x value increases.



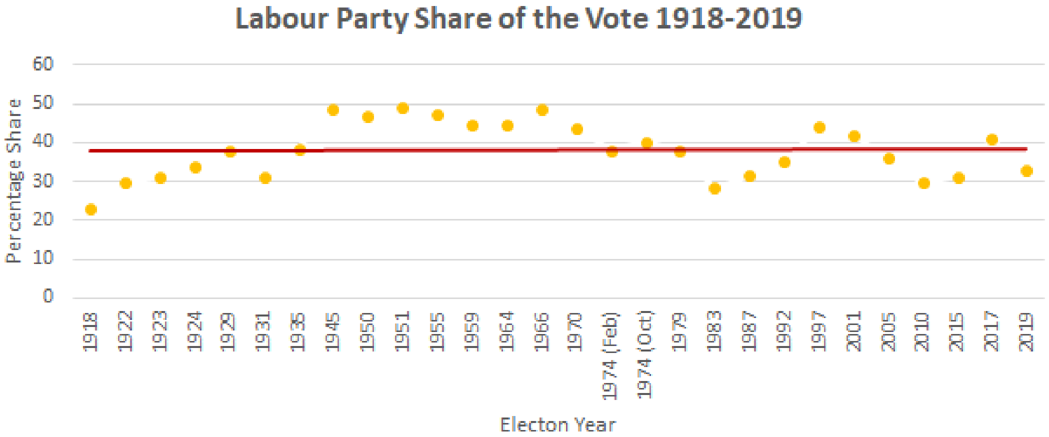
If there doesn't appear to be any pattern, and no clear line moving up or down, then that means no relationship exists between X and Y.



Here is the linear regression for the vote share for the Conservative Party from 1918 to 2019:



And here it is for the labour party:



For each graph, would you say that the 'line of best' fit' is positive, negative or that there is no relationship between the variables?

Activity 3: Correlation Does Not Equal Causation.

It is important to note that the scatterplot and line graphs in these activities only suggest a linear relationship between the two sets of values. They do not suggest that the passage of time is responsible for the support for Labour and Conservative to increase or decrease. Instead, we need to think about what the passage of time infers.

You may have heard of the expression "**correlation does not imply causation**". This reminds us to be careful when reading graphs, as we cannot always conclude that there exists a cause-and-effect relationship between two variables because we think we see an association (or correlation) between them.

Essentially, just because two things appear to be related does not mean that one causes the other.

Let's be a bit more specific in order to make things clearer:

Correlation informs us how strongly a pair of variables are linearly related and change together. It only tells us that a relationship exists, but does not tell us the how or why.

Causation, on the other hand, is what we know as a 'cause and effect' relationship between two variables. It is the change of one variable that we can define as changing another variable.

An example that is often used to demonstrate that correlation does not equal causation is the statistics that show that when sales of ice-cream increase, so do the sale of sunglasses. Rather than concluding that ice cream affects people's ability to handle sunlight or desire to wear sunglasses, we have to look beyond these two variables and think of a missing causal link: sunshine. When it is sunny outside, and it is often hot, and people are more likely to want to buy ice cream or sunglasses in this kind of weather.

So, for example, in the two graphs about the history of the vote share for the Labour and Conservative Parties (from earlier in this chapter), the two variables in the graph are time and percentage of the vote share. Would you argue that it is the passage of time that affects the vote share for each party? Or could there be other types of data that could be affecting their vote share?

What are some possible unmeasured variables that could explain the correlation? List three and explain why you think they would be helpful below.

1

2

3

4: Mapping Politics (Spatial Data Analysis)

*In this chapter, we are going to explore **spatial data** (Maps!) by using a variety of political topics to demonstrate the huge variety in which maps can be used to show data. There will be questions about what you see throughout the chapter in order to help you develop your understanding.*

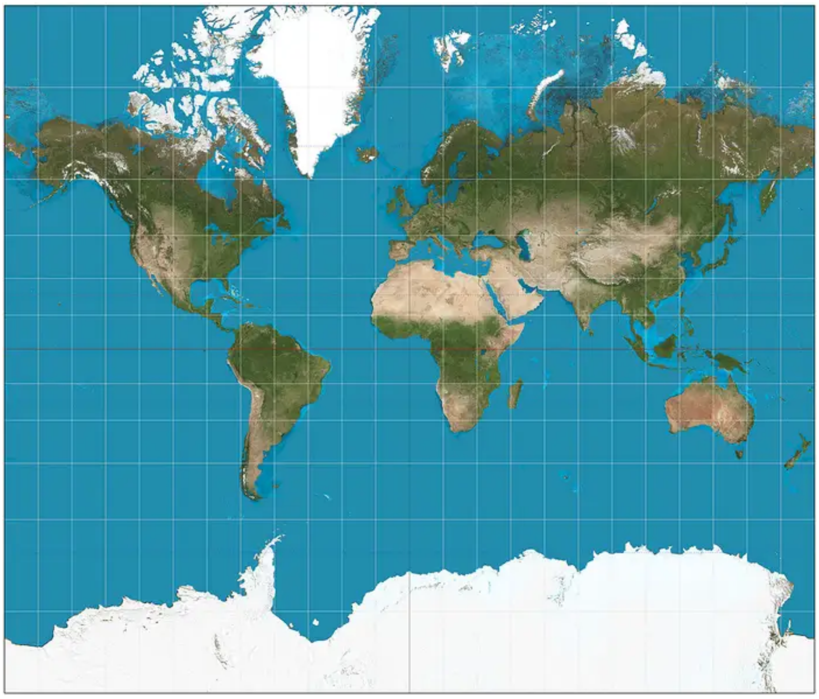
Maps are a really useful and interesting way of presenting information as they represent the real world on a much smaller scale. Maps can show how a value varies across geographic space - and the relationship between these spaces. They show us where specific things happen, and how densely.

We do need to be careful with spatial data, because maps are manipulated to reflect a specific view of the world, and we should always scrutinize the story that the mapmaker is trying to tell us.

How we see the World

Maps are not as straightforward as they seem...

This is the map we are normally presented within understanding our world:



The Mercator projection. Wikimedia Commons

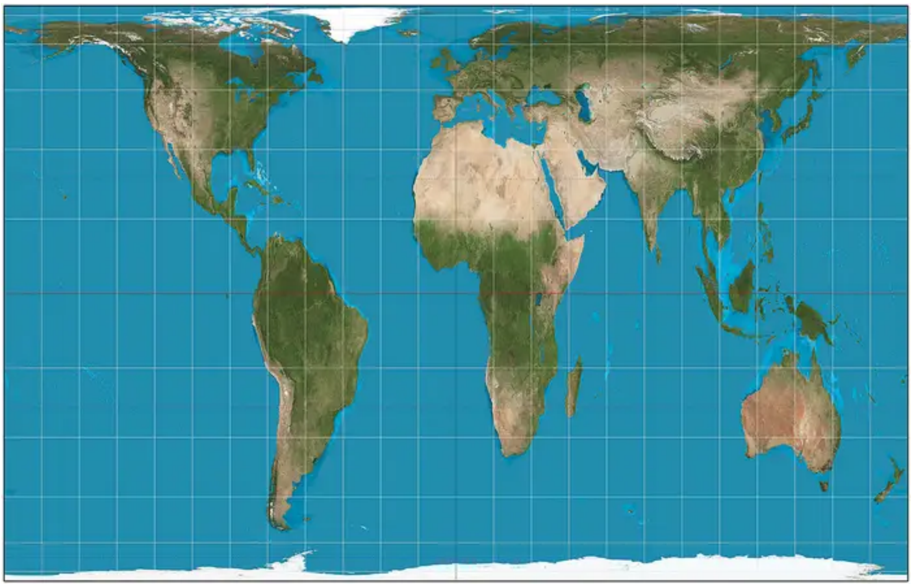
This map is actually quite distorted because the Earth is spherical, and any attempt to flatten this view ends up distorting the data. Look at Greenland and Africa for example, they look like they are close to the same size, but Africa is actually 14 times bigger.

Created in 1569 by a Flemish cartographer, Gerardus Mercator, this map is known as the **Mercator Map Projection**. We have been using this map widely since the 16th century - I'm sure you have seen it in google maps for example.

The map was initially created to preserve directional bearing, making it useful for navigation.

So are there any alternatives?

This map is known as the **Gall-Peters Projection**. In 1974 Arno Peters published this projection, and it was also drawn previously by a nineteenth-century Scottish mapmaker, James Gall. The projection is described as an "equal-area" map because it accurately scales surface areas.



The Gall-Peters projection. Wikimedia Commons

However, it is important to note that the map is still nonetheless distorted as it is a two-dimensional visualization of a three-dimensional earth.These different perspectives are important, as they highlight how we conceive what the world looks like.

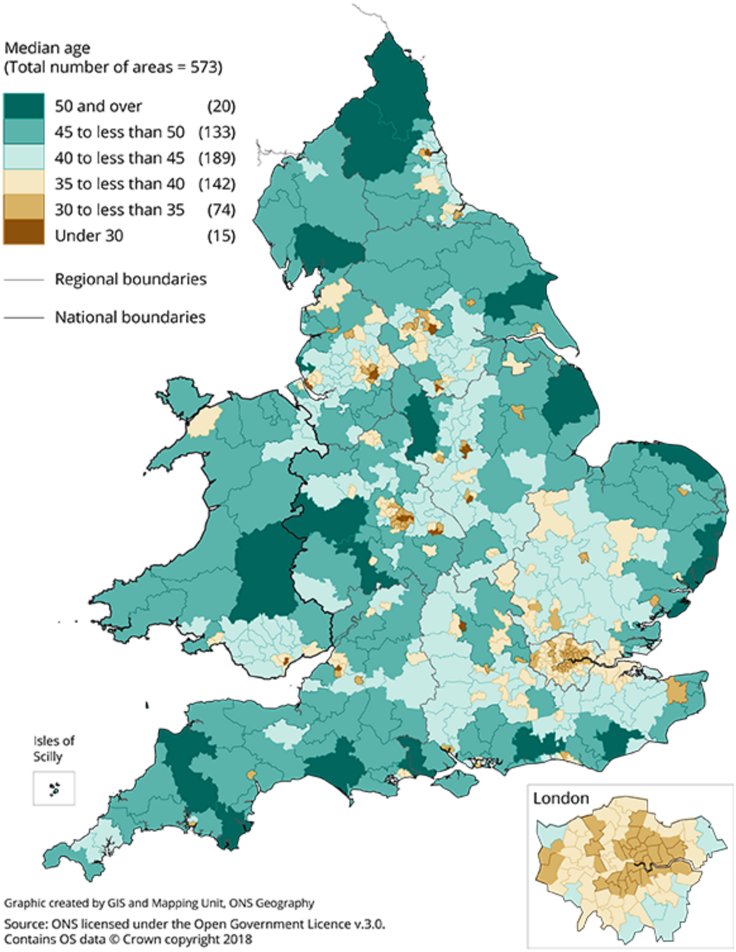
When we use spatial data (also known as geospatial data) we are presenting information and how it relates to a specific location on earth's surface.

Activity 1: The Choropleth Map

There are many different ways of presenting spatial data.

First, we are going to start with **Choropleth map**, a type of spatial data you have probably seen before and is used very commonly in the media and research.

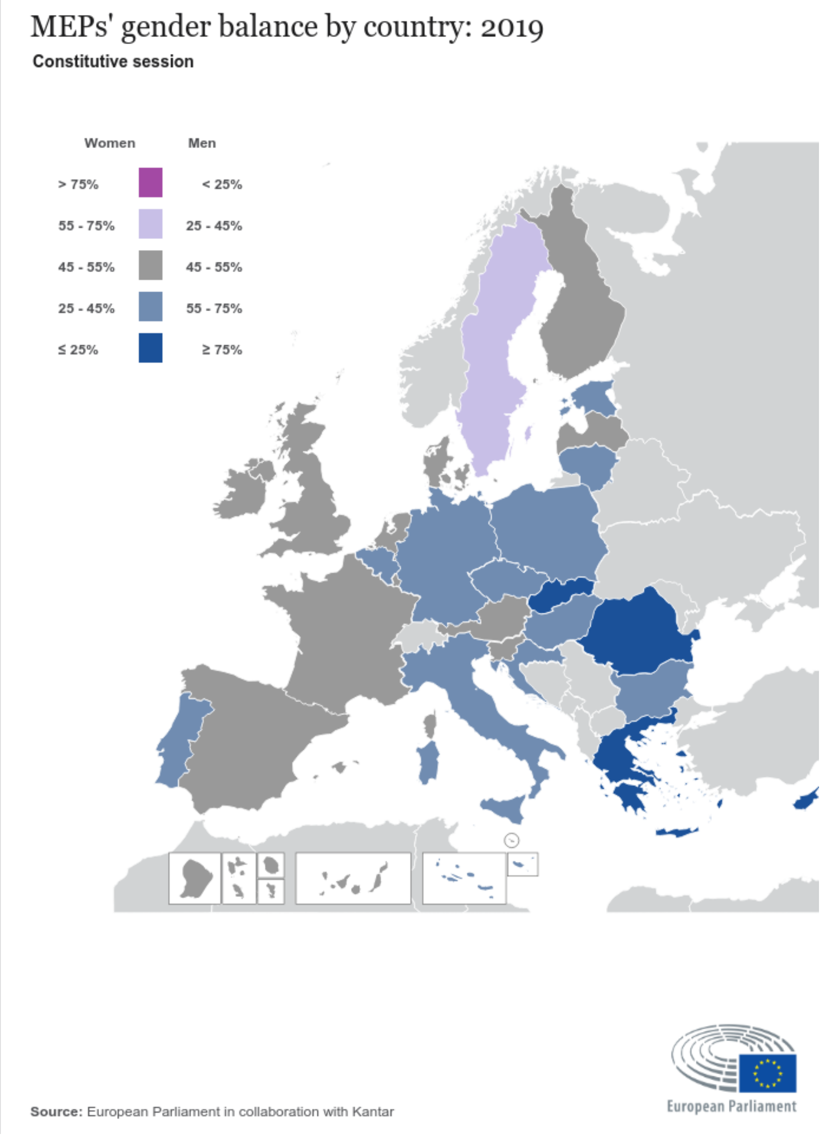
The Choropleth map will distinguish predefined areas by using different types of shading, colouring or symbols that represent the average values of a particular quantity in those areas. You have probably often seen this type of map to show characteristics of a population:



Graphic created by GIS and Mapping Unit, ONS Geography

Source: ONS licensed under the Open Government Licence v.3.0.
Contains OS data © Crown copyright 2018

Let’s start with a simple example. The map below relates to our last activity, and shows how many female MPs individual MPs were voting into the EU Parliament by country.



MEPs' gender balance by country: 2019

Constitutive session

Source: European Parliament in collaboration with Kantar

Why do you think this visual is useful?

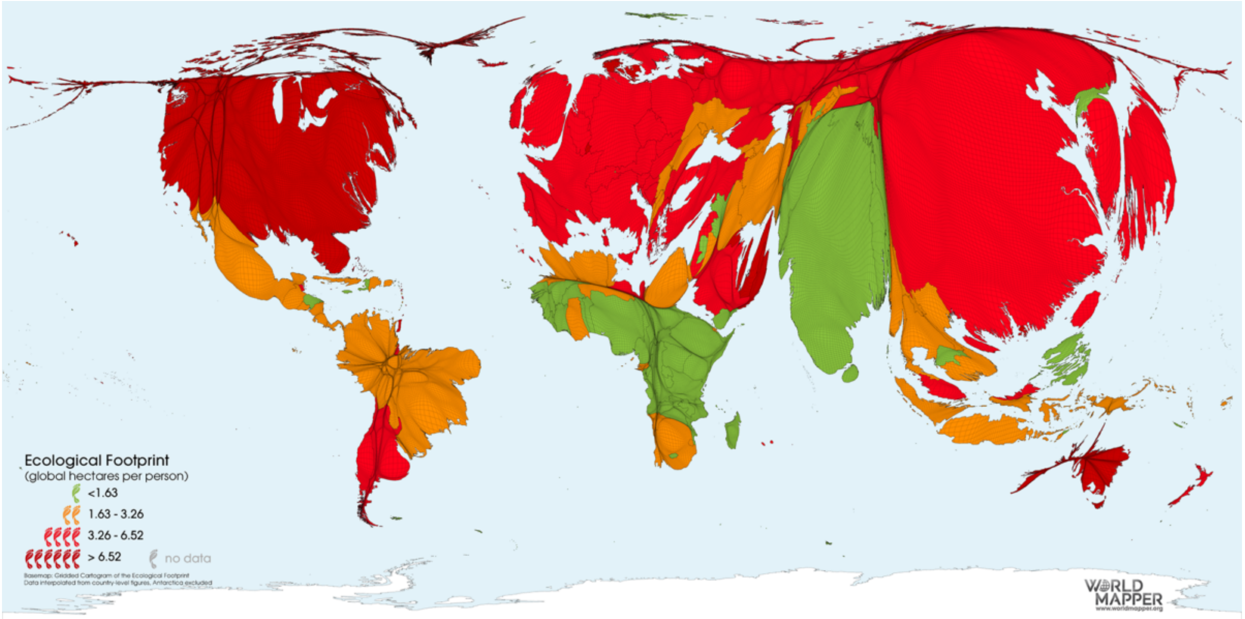
What are the colours telling us?

What information is missing from this map that you think would be helpful in understanding the number of female MPs elected in the EU?

Activity 2: Cartograms

Cartograms are a type of map which distorts the area (typically shown in a Mercator or Gall Peter's projection) in order to demonstrate how the data is equally represented.

Ecological Footprint of Consumption 2019


Ecological Footprint
(global hectares per person)
<1.63
1.63 - 3.26
3.26 - 6.52
> 6.52
no data
Basemap: Gridded Cartogram of the Ecological Footprint
Data interpolated from country-level figures, Antarctica excluded
WORLD MAPPER
www.worldmapper.org

In this map, the areas are altered to demonstrate the total ecological footprint of individuals in each country, in order to quantify humanity's impact on the natural environment. An ecological footprint describes the impact of an individual person or community on the environment, expressed as the amount of land required to sustain their use of natural resources.

Using this map as a guide, answer the following questions:

Which areas of the world have the best ecological footprint?

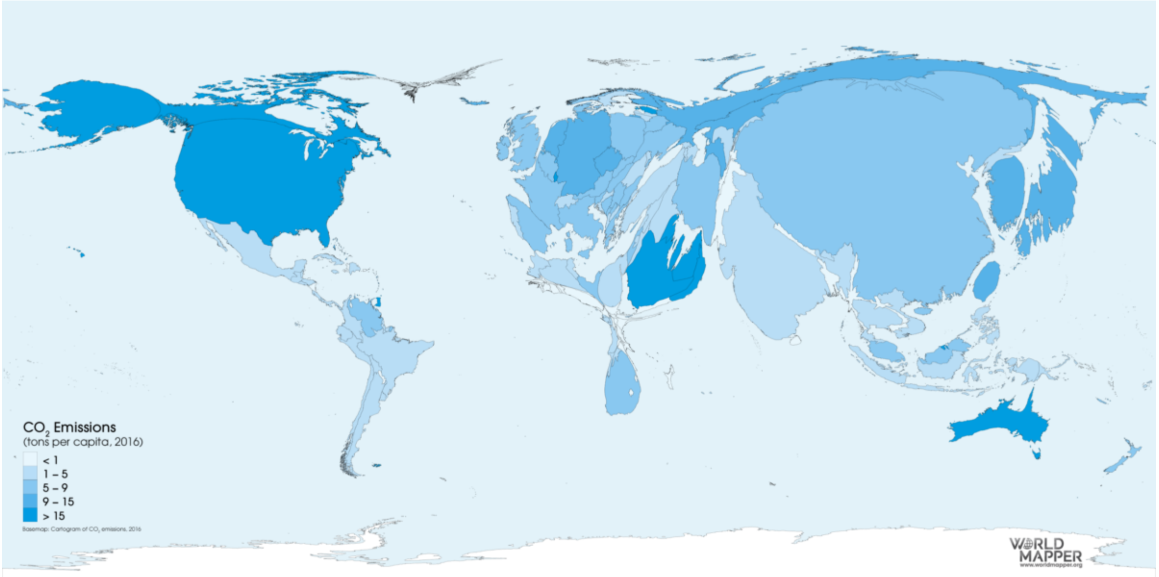
Which have the worst?

How does the UK compare to the rest of the world? And what about the EU?

Why do you think it is useful to see this type of data in this way? Do you find this type of map distortion helpful?

Here is another cartograph that demonstrates the CO2 emissions for each country in 2016.

CO_2 Emissions per capita 2016



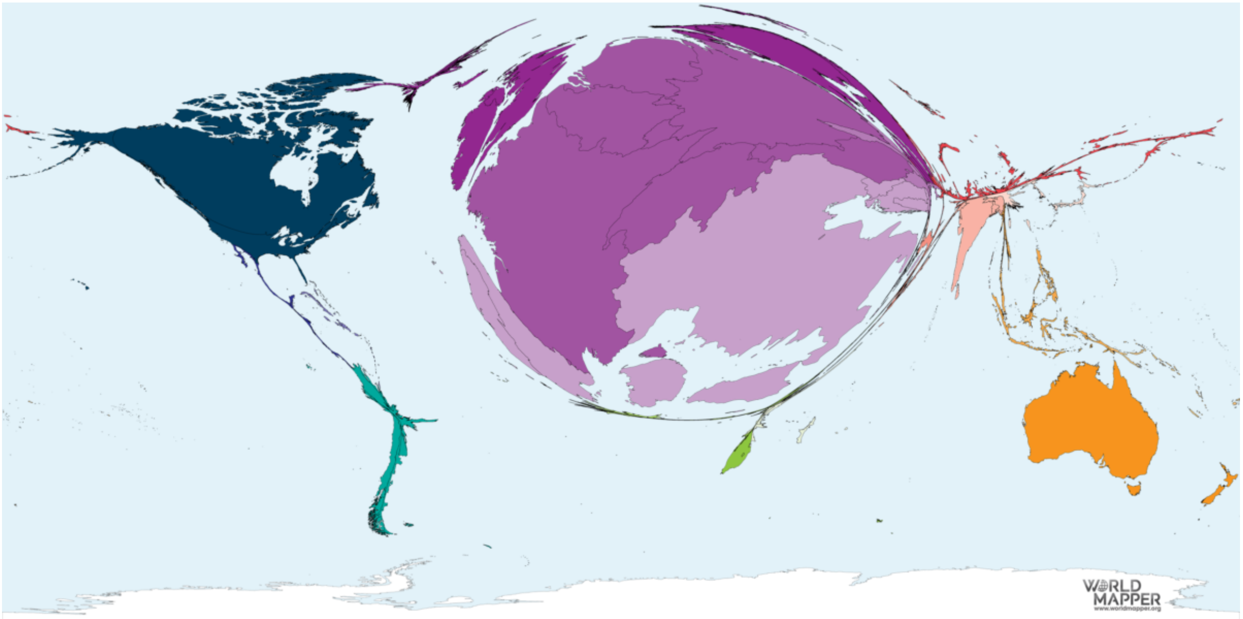
This map distorts the areas to show which countries have highest and lowest carbon emissions. *How does it correspond with the findings of the first cartograph map above?*

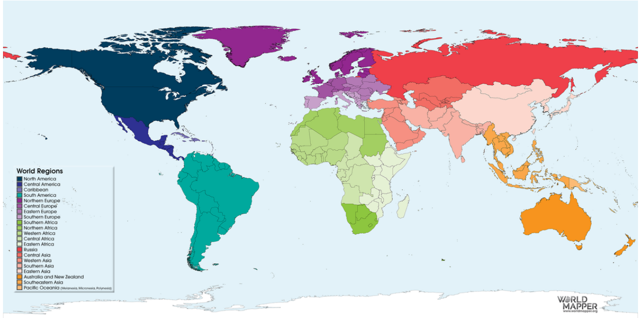
How does it differ?

Why do you think that is?

Lastly, let's look at this graph that looks an important day for Climate Crisis Protests. on Friday 15th March 2019, the largest *School Strike for the Climate* took place, where An estimated 1.4 million young people in 123 countries skipped school in order to demand action from political leaders to take action to prevent climate change.

#FridaysForFuture Climate Demonstrations (March 15th 2019)





Reference Map

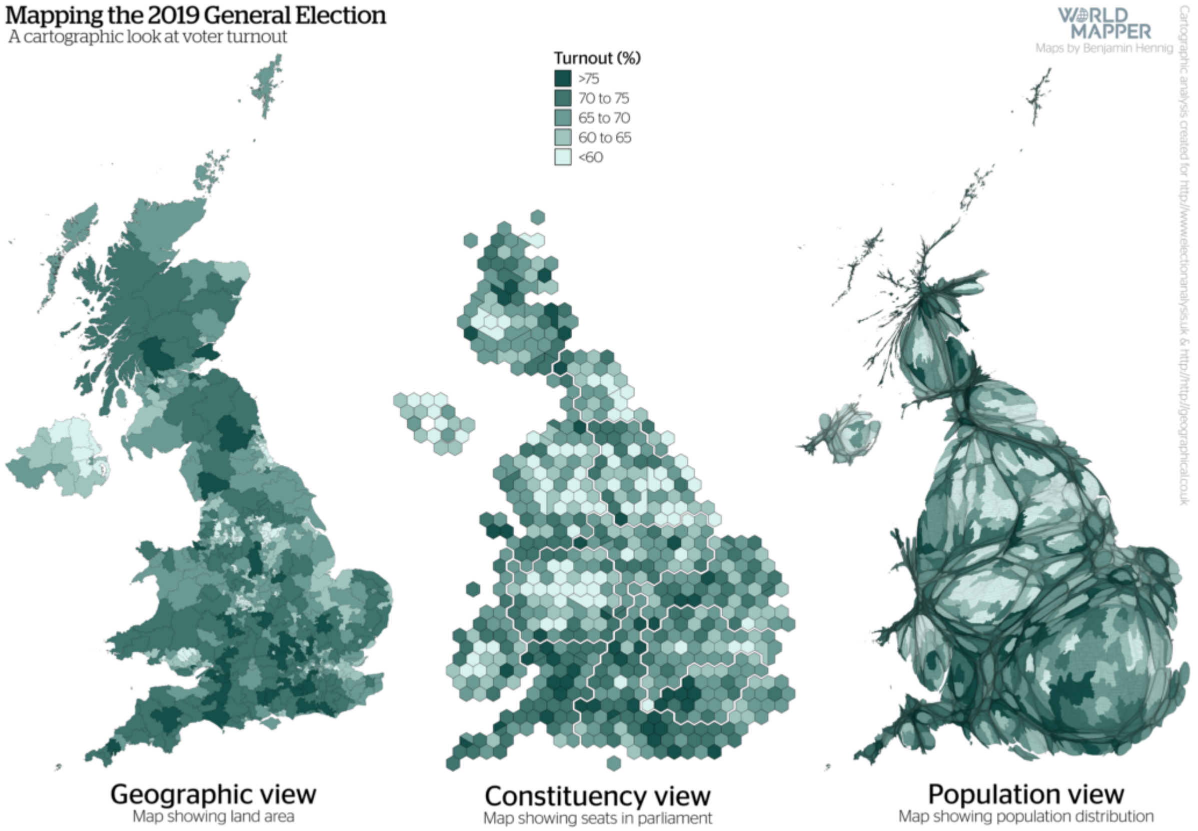
How is the reference map helpful compared to a legend in the corner (that we saw in the two previous maps)?

Are you surprised about the areas that had the most turnout for the climate demonstrations?

If you compare to the two graphs above, why do you think the northern hemisphere appears to have held more demonstrations?

Activity 3: Comparative Maps

In this section will be comparing maps. Here is an example of three different ways of mapping the voter turnout at the recent 2019 UK General Election:


Mapping the 2019 General Election
A cartographic look at voter turnout
Turnout (%)
>75
70 to 75
65 to 70
60 to 65
<60
WORLD MAPPER
Maps by Benjamin Hennig
Cartographic analysis created for http://www.electionanalysis.uk & http://http//geographical.co.uk
Geographic view
Map showing land area
Constituency view
Map showing seats in parliament
Population view
Map showing population distribution

As you can see, there is a geographic (choropleth) view, a constituency (hexagon) view, and a population (cartograph) view.

The geographic views show voter turnout in a conventional choropleth map, in the form of the map of the UK that we are used to seeing.

The population view shows the map altered by how each area is resized according to the number of people living in that area (cartograph). In all three of these maps, the different shades of green represent levels of voter turnout.

The constituency view instead uses its hexagons to present individual constituencies. Hexagon maps can offer clarity in the way they standardize geographic spaces, due to the ability for the hexagon to tessellate well. They are still able to produce recognizable geographic representations, as you can see in the map above showing seats in the UK parliament.

1 *From these maps, which area do you think has the best voter turnout? Which had the least?*

2 *How do the nations of Scotland, Wales, Northern Ireland and England compare?*

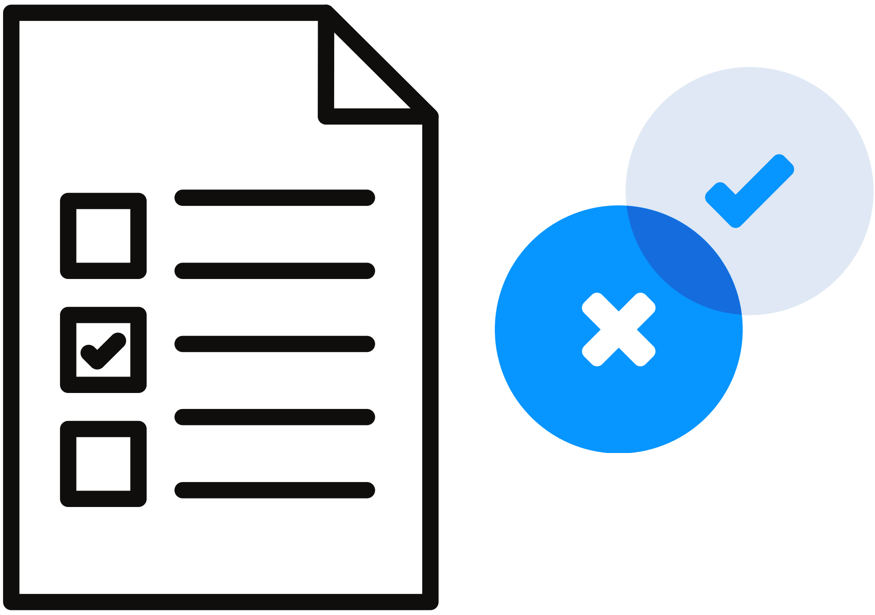
3 *What do you think is the advantage of each map?*

4 *Do you find one easier to understand than another?*

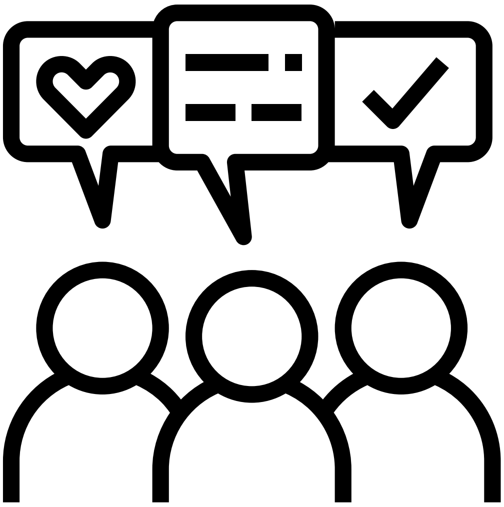
5. What is the Public Thinking? (Polling and Surveying)

In this chapter, we are going to explore what is going on behind the numbers in political statistics - the survey questionnaire.

Political Scientists, journalists, media outlets and researchers use polls and surveys as a means of understanding what people are thinking and feeling because we cannot get everything we want to know from administrative data collected from other purposes. The data collected from surveys is often what we see in graphs and maps.



Polls vs Surveys



Polls normally only ask one or two questions with limited analysis of data, and are often employed in political races to gather immediate feedback and predict an outcome.

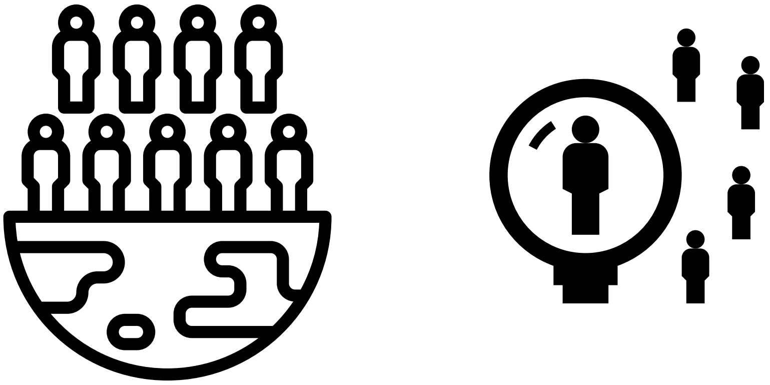
Surveys, on the other hand, use multiple questions to gather data facts, behaviours, attitudes, and preferences from a targeted population.

Polls and surveys help us understand what the public is thinking. They allow the public to express themselves and create a quantifiable way of demonstrating the opinion of large groups of people.

In expressing their opinion and judgments on special and important political issues, the results from polls and surveys help inform decisions made in both local and national government.

To who are we asking questions?

In statistics, the **population** is the whole group, and would consist of every member in a group. For example, everyone who lives in the UK would form a population. A population thus may refer to an entire group of people but it can also refer to a group of events, objects, measurements - like hospital visits or football game attendance.



A **population sample** is the group of individuals who are selected to participate in a study, or for the purpose of today's activities, a survey. While it can be possible to survey an entire population -for example, like in a census - often this is impractical so a sample is surveyed instead.

Activity 1: Types of Question

Our first activity is going to help you identify types of questions and how they are suitable for different subjects.

A **Rating Question** asks participants to demonstrate their opinion about something by using a scale. For example, in this question, the rating is out of 5:

Please describe how you felt about your breakfast this morning.
Unsatisfied (1)
Somewhat Satisfied (2)
Satisfied (3)
Very Satisfied (4)
Extremely Satisfied (5)

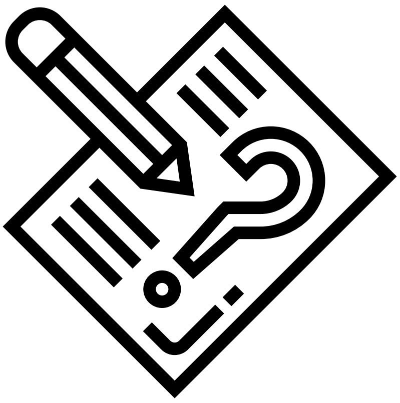
A **ranking question** uses a list to asks participants to demonstrate how they feel about something by comparing it to others. For example:

Please rank the following sporting activities in order of preference (starting with 1 for your favourite activity).
__Football
__Running
__Rugby
__Field Hockey

Which type of questions do you think is best for gathering a participant's opinion on something?

-A rating question asks participants to think about individual items.

-A ranking question asks participants to list their responses in order of preference but does not tell you if a person likes or dislikes an item on that list.



Non-structured questions - also often called open-ended questions - are questions where the participant is asked to write/say their response to a question, and they are not given a list of choices. These types of questions are useful for when you do not know what you expect from your participants, and would like to explore new ideas. For example:

Why did you decide to take certain GCSE subjects?

____

____

If you have an idea of what you would like to explore, but would like to give participants the opportunity to propose new ideas or give an answer you may not have thought of, a **Partially Structured Question** can be used. This is often done using a list of choices with the option to include an unstructured answer. For example:

Which GSCE subjects did you enjoy the most this academic year?

( ) Maths
( ) English Literature
( ) English Language
( ) Art and Design
( ) Modern languages
( ) History
( ) Geography
( ) Biology
( ) Chemistry
( ) Physics
( ) Other ____________________

It is possible to also ask questions in a way that can distort your participant's true answer.

Here are some examples of how this is done:

If you ask a **leading question**, you are steering your participant to a desired response. An example of a leading question would be "Do you like our current Prime Minister?", where the participant would not necessarily feel comfortable choosing 'no' as an absolute, so would feel forced to say 'yes'. Instead, a survey should ask something along the lines of "How would you rate the performance of our Prime Minister on a scale of 1 to 10?"

A **loaded question** makes presumptions about a participant's experience, forcing a response in a predetermined way. An example would be "Who did you vote for in the 2019 General Elections?", when it should be "Did you vote in the 2019 General Elections?"

You should also avoid **compound questions**, (or **double-barrelled questions**), which is when a question is asking about more than one topic or theme - so it is asking two distinct things but only allows for a single answer. An example of this would be: "How satisfied are you with your MP and Political Party on a scale of 1 to 10". Instead, there should be two questions: one about the MP, and one about the Political Party.

Let's Create a Survey

Formulate groups consisting of 3 to 5 students. If you are at home, you can do this activity on your own and simply ask these questions to classmates, friends, or family virtually.

Here are three political themes for you to think about:

1) The Government's role in affecting climate change.
2) The right to vote at 16.
3) Government influence over social media.

For each of these themes, think about an overarching research question - the purpose of your survey.

Now, choose a topic (it doesn't necessarily have to be one of the three above) and a corresponding research question you created - and come up with five to ten questions. Make sure you think about and discuss with your group what you think is the best wording, format and response options for each question.

Bonus: What would be a couple examples of poor or leading questions?

Activity 2: Let's Execute the Survey!

Choose what you think are the best or most interesting examples of questions that you brainstormed for the theme you chose in the last activity.

Now survey 10 people in your class (or whoever you can get in touch with, including friends and family). Write out the surveys on blank pieces of paper and hand them out (this can also be done in an email, Word Document etc).

If you are doing this activity in person, make sure that no one writes their name down on their survey so that all of the data is anonymous. *Why do you think is important?*

Keep in mind that this may not necessarily be the best population sample because you are limited in who you can approach and may not be able to ask diverse types of people (e.g. age, gender, race etc.) - this will be discussed further in the next chapter.

Activity 3: Analyse the Data

Now, analyse your data. Look for common responses. Are there any answers that are similar?

What is it telling you?

How would you visually present the data from your survey?

Feel free to flip back to the previous chapters for inspiration! Think about what you learned in chapter 2, 3 and 4.

For example, pie charts are useful for showing data about the whole of one category in time, while bar graphs are useful for showing differences between different categories of data.

Bonus: Draw it!

Make sure you keep the materials from this chapter for when you do the next and final chapter.

6. Media, Bias and Politics (Bias)

When you see data presented to you in the media, it is important to be able to critically examine the statistics for any bias.

Often, headlines will be made for attention-grabbing, and not for accuracy. We see numbers used more and more in the news, and they can be used to persuade you to conduct a certain way. For example, who to vote for in an election or referendum.

With the Covid-19 pandemic, there isn't a day that goes by in the news that we don't see numbers, graphs, and statements about statistics to explain how the UK, and the world, is handling the crisis.

But what do all these numbers really mean? And are they always used objectively?

Today's activities will help you scrutinize the statistics you see in the media.

What is bias?

Sometimes, research is biased when information is presented in a warped way in order to guide someone to reach a desired conclusion.

In statistics, bias occurs when the results differ from the true quantitative parameter being estimated. There are many different types of bias in statistical research.

Examples of different types of biases that we will be exploring today:

-**Selection bias** involves choosing individuals selectively over others in a study, and therefore proper randomisation is not achieved. This means that the population sample obtained is not representative of the population intended to be analysed.

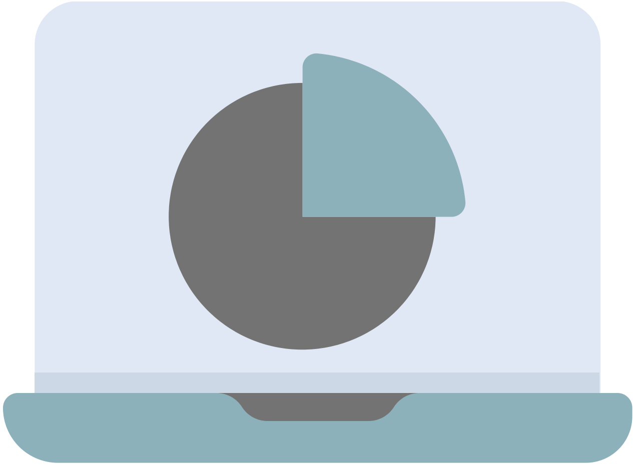
-**Exclusion bias** happens when specific individuals are excluded from the population sample of a study.

-**Observer bias** arises when the researcher subconsciously influences research due their own expectations, which can alter how a survey is carried out, or how the results are recorded.

-**Reporting bias** is when only certain results are chosen to be included in an analysis, leaving out relevant evidence.

-**Recall Bias** occurs when participants are unable to remember past experiences or events accurately, and memories can be influenced by subsequent experiences and events.

-**Response bias** (also known as **survey bias**) is a general term that describes the different tendencies in participants to answer untruthfully or inaccurately. For example, participants may be reluctant to give truthful answers if they are not deemed socially acceptable. This type of bias frequently happens when participants are asked to self-report (like in structured interviews or surveys) and it can also be the result of poor survey design.



Activity 1: Lying in Statistics

Our first activity is going to help you scrutinise the numbers, and statements about them, that you see in the media.

When looking at statistics you should be asking yourself these questions:

When was the data collected? Where did the data come from? What questions were asked and how were participants asked to answer? Is there any information that is missing that could affect the results? Can you think of a way to measure the data in a better way?

In this activity, we are asking you to examine 5 different statements about statistics.

For each statement, answer these three questions:

1) Is there is any bias evident here?

2) What type(s) of bias?

3) How would you improve the statement?

Researchers surveyed a population sample in Sheffield to find out how the UK would vote if there was a snap election.

Researchers surveyed a random sample of the UK population to find out why they voted the way they did in the 1997 elections.

Researchers collected data on a random sample of individuals from around the UK and found the mean age of Conservative supporters in the UK.

Activity 2: How did we do?

Now, let's scrutinise our own work!

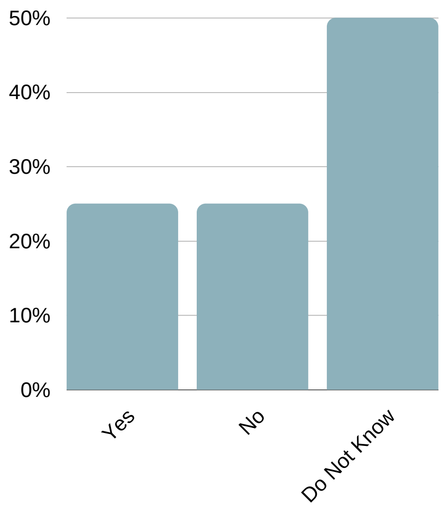
What may have been biased about your survey and its results in the last chapter?

How could you make a misleading headline about your research?

For example, I will give you some data with an accompanying headline, and you can tell me what is misleading about the statement and graph about this data (please keep in mind that this data is made up for the purpose of this experiment).

Graph Title: Results show that Only 25% percent of the US will vote for Trump.

Research Question: Will you vote for Donald Trump in the 2020 Election?



In the graph above we can see that half the population is undecided, which means that in the end there may be a lot more than 25% of people who vote for Donald Trump.

So now, brainstorm three headlines from your survey that would be misleading, but technically correct:

1.

2.

3.

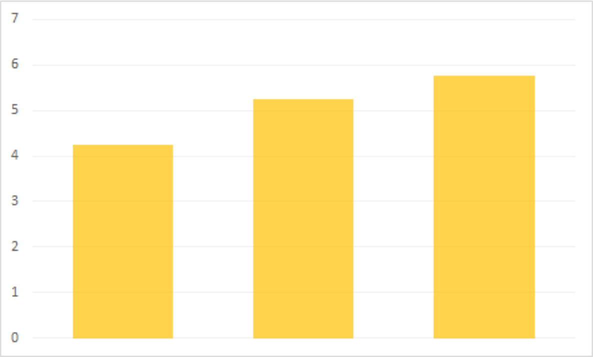
Next, form small groups of 4 to 5 of your classmates. Share your headlines and see who can work out what may be misleading about your statements. After, share you original data to check if they were right!

Activity 3: The Right Visual

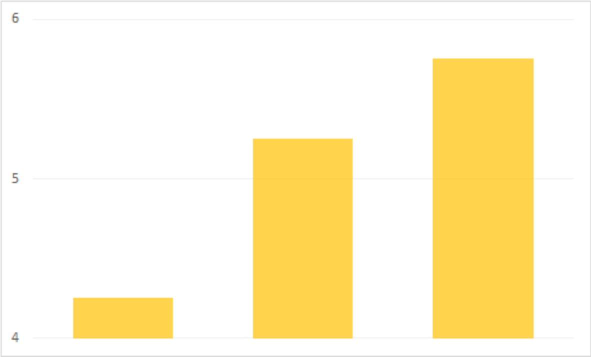
In this this activity we are going to explore how statistics can often be presented in skewed ways. Here are some ways that statistics are visually misrepresented:

Scaling and Axis manipulation is used to alter the way that a scale looks for a graph. This changes the impression of the magnitude of differences:

A.

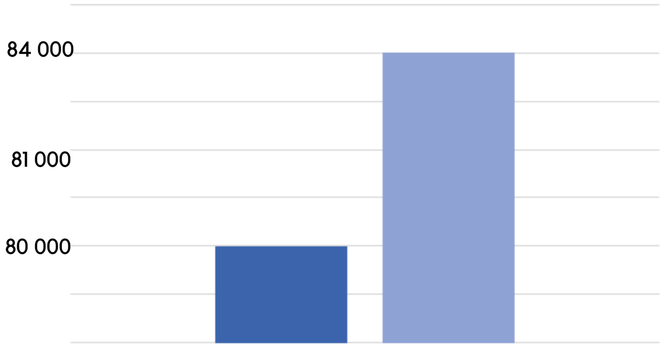


B.



Here is another example of Scaling and Axis manipulation.

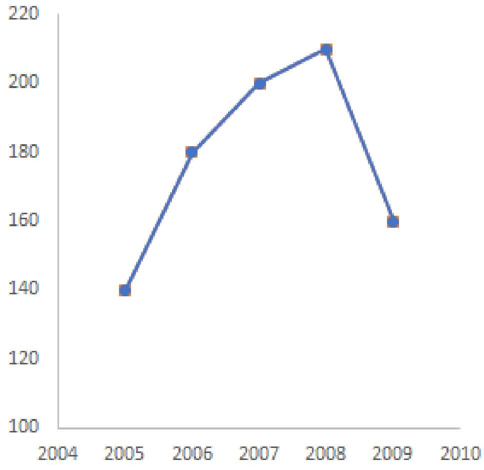
C.



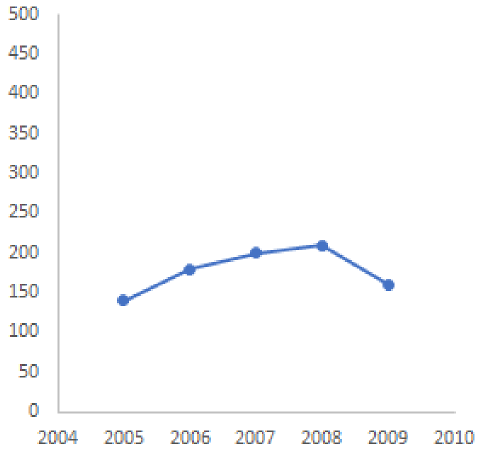
Have a close look at bar charts B and C - how are they misleading? How should the information be presented?

Here is an example of the data can be misrepresented in a line graph. Both graphs show the same data but with different scales.

D.



C.



Which one appears more truthful to you? Why?

Now, is there a way of visually demonstrating your data in a skewed way? Draw it!:

Activity 4: Can you *infer* with your data?

The reason why we want to have the right type of population sample is because we want to be able to **infer** our results on the whole population that the sample represents. To infer means 'to deduce', and in survey results this means that we want to be able to draw a conclusion from a sample, that reflects what the population is thinking as a whole.

In order to infer our survey data, our results therefore have to be probable.

Probability is the likelihood that an event will occur.

Probability in statistics is what allows us to use an observed sample of data to make inferences about an unobserved population.

An example of probability is a coin toss. When you toss a coin, there is a 50% probability that it will land heads up after you flip it. There is also a 50% probability that it will land tails up. In statistics, we often use decimals to represent probability. So for a coin toss, the probability would be 0.5

How does this apply to a population sample? When conducting a survey, you want to think carefully about who your participants should be. For example, just because your research topic is about the elderly, does not mean that you should solely interview people over the age of 65. The results would differ according to who we ask: male or female, young or old, Conservative or Labour supporter.

In order to demonstrate what the population is thinking, we have to survey a sample that is representative of the entire population. That is the only way that we can get the right probability to infer our results on an unobservable population.

Think of who you surveyed in Chapter 5 - can you infer statistically from your population sample?

In the survey activity in Chapter 5 we essentially showed the sample as an example of what GCSE students (or your friends and family) are thinking in the UK. This was a **non-probability sample**: which means that people were selected based on non-random criteria. A form of this is a **convenience sample**, which when participants are chosen for accessibility and availability - which is what we did!

Here are some questions to help you reflect on how your survey could, or could not, be inferred on the UK population as a whole:

1. Is your population sample biased in any way?

2. And how would you improve your population sample?

3. What about the questions you asked? Is there any type of bias that could arise from the wording or type of questions you asked?

4. In hindsight, is there anything you would change?

Q-Step
A step-change in quantitative social science skills
Funded by the Nuffield Foundation, ESRC and HEFCE
Nuffield Foundation
The University Of Sheffield.