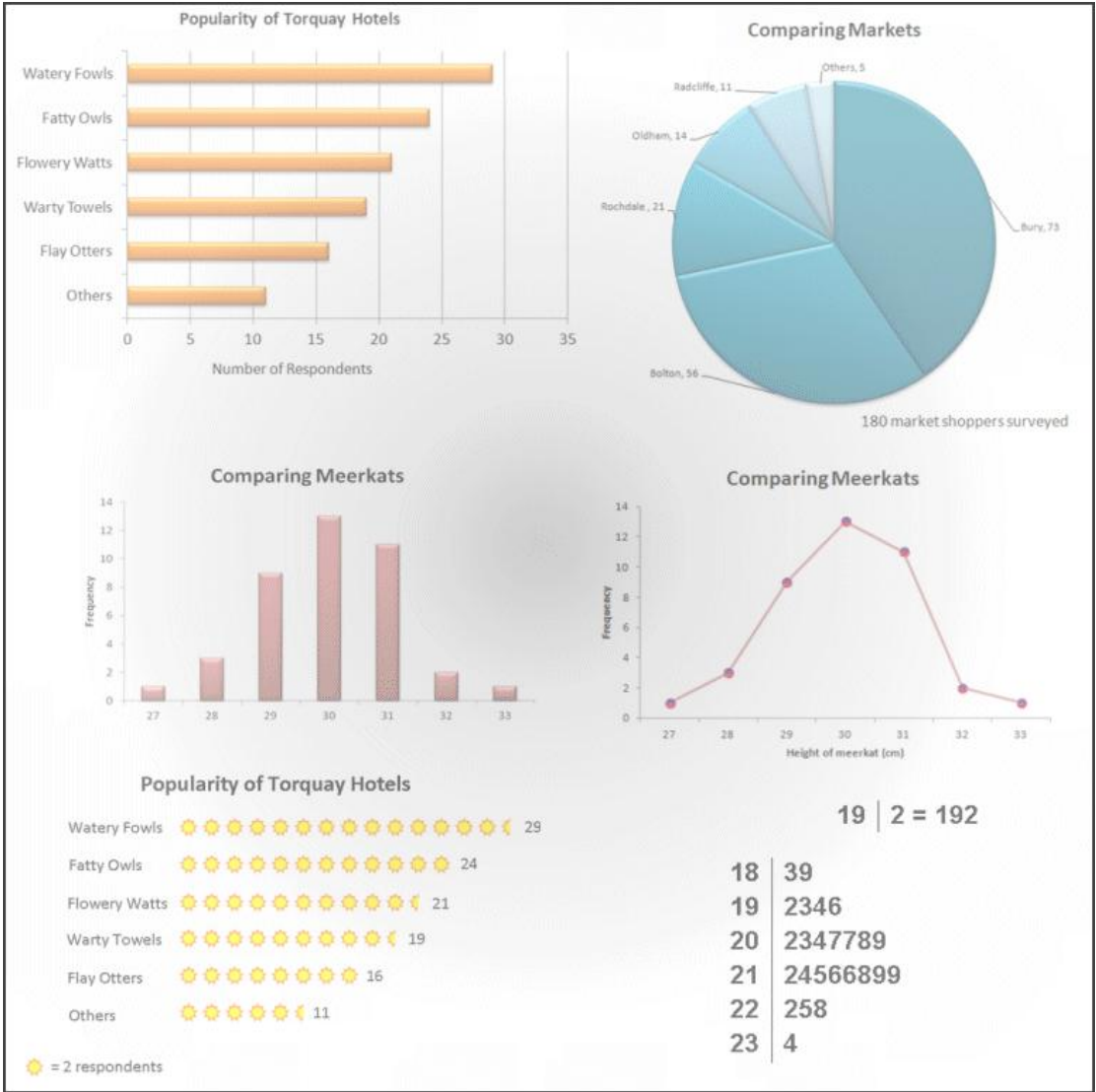


M.K. HOME TUITION

Mathematics Revision Guides

Level: GCSE Foundation Tier

REPRESENTING DATA



## REPRESENTING DATA.

### Discrete Data.

A set of data is said to be discrete if the values / observations belonging to it are distinct and separate, i.e. they can be counted (1,2,3,...).

Discrete data can be numeric, e.g. the number of goals scored by a football team, or non-numeric, such as blood groups or car colours. It can also be time-related, i.e. monthly sales figures, a temperature chart taken hourly.

Examples might include the number of puppies in a litter; the hourly number of patients in a doctor's surgery; the number of flaws in one metre of cloth; gender (male, female); blood group (O, A, B, AB).

### Continuous Data.

A set of data is said to be continuous if the values / observations belonging to it may take on any value within a finite or infinite interval. Examples of continuous data include height, weight, temperature and air pressure over time.

Continuous data can be converted into discrete data by rounding, e.g. heights of humans to the nearest cm, lifetime of a lightbulb to the nearest 100 hours, daily maximum temperatures over a month, to name a few.

### Representing non-numeric data.

Examples would include a sample of eye colours, methods of transport to work, blood groups and similar other examples.

Discrete non-numeric data can be illustrated by a bar chart or a pie chart. A line chart (frequency polygon) would be meaningless here, as the example overleaf will show.

### Collecting Data.

The collection, set or group of objects being studied is termed the **population**.  
Examples would include:

- All the pupils in Year 11 in a particular school
- All the people on the electoral register
- All people who watch soaps on television
- All the cars produced by a factory in one week
- All the hotels owned by a certain holiday chain

A **sample** is a small part of a population and can be selected using various ways.

A sample is said to be **biased** if there are any underlying factors distorting the data in such a way that it will not give a representative picture of the population.

There are several ways in which data can be collected directly.  
Data can either be ‘primary’ or ‘secondary’.

**Primary data** is information that is obtained directly from first-hand sources by means of surveys, observations or experiments.

Two common methods of collecting primary data are the tally chart and the questionnaire.

### The Tally Chart.

This is typically used to record data which can be easily observed and counted, usually over a time interval, although it can also be used for static data.

A few examples:

- The number of items paid for at a “10 items or less” supermarket checkout over one day
- The number of cyclists using a stretch of road over 15-minute intervals
- The number of passengers on a bus during various stages of a journey
- The number of cars, by year of registration, using a stretch of road over one hour

**Example (1):** A group of Year 11 pupils have been tallying the number of cyclists using a section of the A56 into Manchester during the morning rush-hour on a Tuesday morning.

They grouped the times into 15-minute sections beginning at 0700, 0715, 0730 .... up to the quarter-hour beginning at 0915.

The completed tally chart looks like this, with the ‘gate’ graphics :



This tally chart can then be re-displayed as a bar chart or a frequency polygon (see later).

### The Questionnaire.

Another way of obtaining information is by designing a questionnaire.

An ideal questionnaire should be clear, concise and not confusing to the user. It must also not be biased or vague in its wording.

Here are a few examples of data that can be obtained through a questionnaire :

- Choice of mode of personal transport to and from work
- The number of hours people spend doing physical exercise
- Choice of national daily newspaper read by a household
- The number of hours people spend per week surfing the Internet

All questionnaires must include a response section, typically a box for ticking. It is most usual to allow only one response choice, although the third case above can have multiple responses.

**Example (2):** This questionnaire concerns the popularity of TV soap operas.

**Speaking as a true Salfordian, I think Coronation Street is a better soap than EastEnders.**

**Do you agree ?**

☐ **Yes**                      ☐ **No**

This example is deliberately silly, but it illustrates an important point : the question must not be a persuasive or leading one in any way. There also is no “Don’t know / Not interested” response box.

Correct would be “Which soap do you prefer to watch – Coronation Street or EastEnders ?” , followed by three options, namely “Coronation Street” , “EastEnders” and “Don’t know / Not interested”.

**Example (3):** This questionnaire is about eating habits, specifically take-away food.

How often do you eat takeaway food ?

<input type="checkbox"/> Never	<input type="checkbox"/> Rarely
<input type="checkbox"/> Fairly often	<input type="checkbox"/> Very often

This time, the issue is one of vagueness.

There is no mention of time scale, and there is a lack of clarity as to what is meant by, “rarely”, “fairly often” or “very often”.

To a health-conscious eater, “fairly often” can mean twice a month; to a person on a poorer diet, “rarely” could mean twice a week.

How many times a week do you eat takeaway food ?

<input type="checkbox"/> Never	<input type="checkbox"/> 1 - 2
<input type="checkbox"/> 2 - 4	<input type="checkbox"/> More than 4

Most of the faults have been corrected here – there is now a well-defined time scale, and the vague words have been replaced by actual numeric quantities.

There is still one problem – which box would be ticked by a person who eats takeaway food twice a week ?

How many times a week do you eat takeaway food ?

<input type="checkbox"/> Never	<input type="checkbox"/> 1 - 2
<input type="checkbox"/> 3 - 4	<input type="checkbox"/> More than 4

This is the final form of the questionnaire, without any overlap among the categories in the tick boxes.

### **Biases in primary data collecting.**

We have seen one example of bias in the case of the badly-written questionnaire about TV viewers' favourite soap opera in Example (2). There, the question was crudely loaded in an attempt to influence the 'Yes/No' answer in favour of a 'Yes'.

Other biases are usually centred about location or time, and examples are easy to find or suggest.

**Example (4) :** A group of Year 11 pupils distributed a questionnaire outside the leisure centre about how much time a week people spent on physical exercise. A disproportionately large number of respondents ticked the "Over 5 hours per week" box.

(Users of the leisure centre were more likely to engage in physical exercise than non-users.)

**Example (5) :** The results of a questionnaire about whether the widening of the M60 motorway was a good idea to improve traffic flow. The response was a very strong 'No'.

(The people questioned all lived within 200 metres of the proposed motorway.)

**Example (6) :** The queues of daytime traffic into Mill Gate car park in Bury, as counted on three successive Mondays and Tuesdays. The apparent results showed that long queuing (over 20 cars waiting) was not an issue, yet many shoppers complained on Wednesdays and Fridays about difficulty in finding parking spaces.

(Wednesdays and Fridays are full market days in Bury; Mondays and Tuesdays are not.)

### **Avoiding bias in primary data collecting.**

Each of the questionnaires in the last three examples could have produced improved results in various ways. Here are some suggested corrections.

**Example (4) Correction :** Instead of distributing the questionnaire solely outside the leisure centre, the pupils should have chosen about four or five different town centre locations.

**Example (5) Correction :** Extend the area for distributing the questionnaire to include properties up to, say, 500 metres of the proposed motorway, rather than just the 'inner zone' of 200 metres.

**Example (6) Correction :** Count the traffic entering the car park for three complete weeks, Sundays to Saturdays, and not just Mondays and Tuesdays.

**Secondary data** is data collected by someone other than the user.

A few examples are :

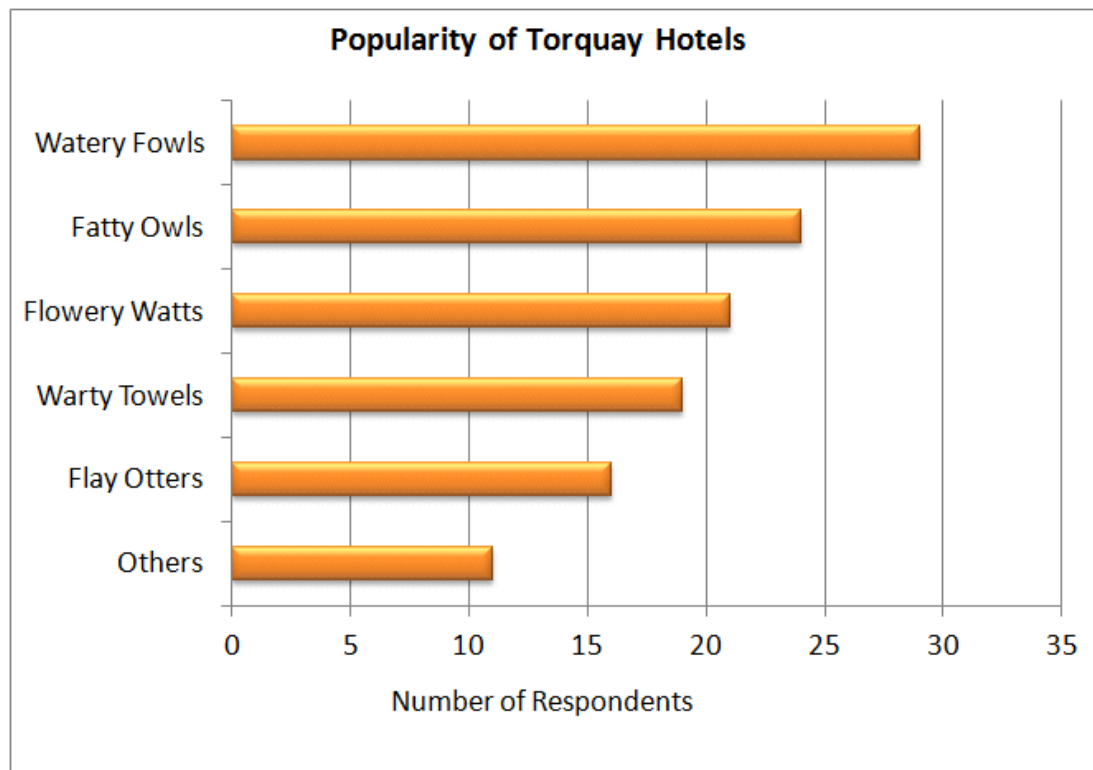
- Using data from the Met Office website to compare Manchester's rainfall statistics for January 2013 and January 2014.
- Using data from the Department of Transport and Greater Manchester Police to analyse road accident statistics.
- Looking up the database of Olympic Games statistics to measure trends in performance for particular athletic events.

### Representing Data – the Bar Chart.

**Example (7):** The West of England Tourist Board contacted a sample of 120 holidaymakers to find out their favourite hotel in Torquay.

The results were as follows: Watery Fowls was voted top by 29 people, Fatty Owls by 24, Flowery Watts by 21, Warty Towels by 19, Flay Otters by 16, whilst 11 voted for ‘others’.

The results can be shown in a bar chart (which could be vertical or horizontal) , where the lengths of the bars are proportional to the numbers of the respondents.



The hotel names here are the **group labels**: the numbers of respondents are the **frequencies**.

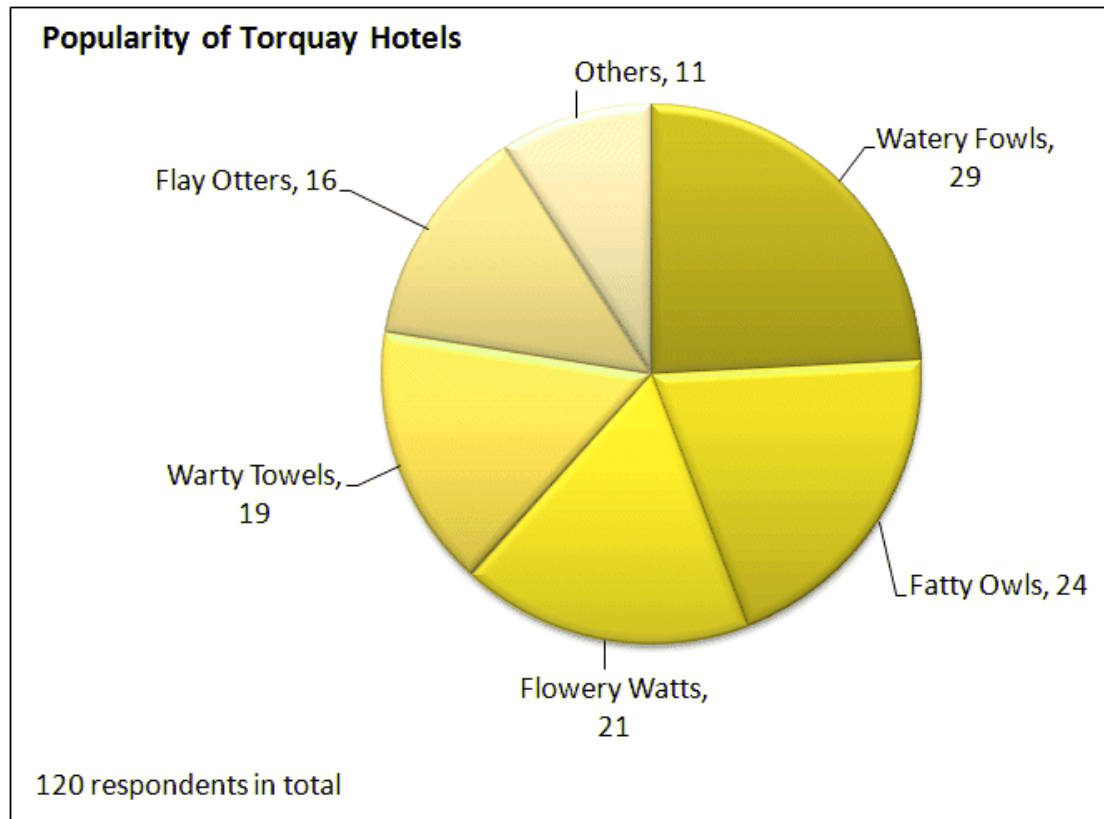


### Representing Data – the Pie Chart.

Another suitable representation would be a pie chart, which is a circle divided up proportionally like segments of a pie.

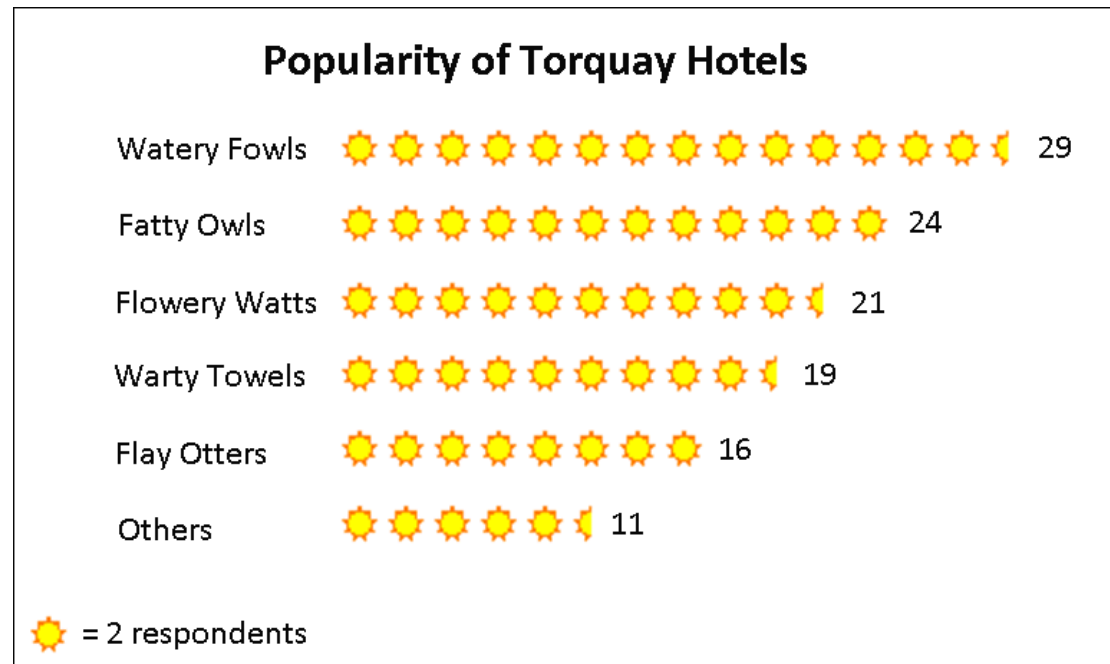
Here the sample size is 120, and so each respondent in the sample would be represented by a  $\frac{360}{120}^\circ$  slice of the pie, or  $3^\circ$ .

Thus, “Watery Fowls” with 29 respondents would occupy a  $29 \times 3^\circ$ , or an  $87^\circ$  slice, of the pie.  
“Fatty Owls” would take up  $24 \times 3^\circ$ , or  $72^\circ$ , and so forth.



### Representing Data – the Pictogram.

A **pictogram** is yet another way of representing discrete data. Here, a graphic symbol represents a numeric quantity.



The sun symbol here represents 2 respondents, with a special ‘half’ symbol to stand for one person. Pictograms can be entered fairly easily into a computer using one of the graphic fonts available, but are generally too tedious to be used in handwritten work.

Also, when one symbol represents (say) 10 people, it might be difficult to think up of many ‘partial’ symbols to represent 1, 2, 3... people.

**Example (8):** A survey was carried out among 180 market shoppers in the Bolton/Bury area to compare the popularity of different local markets.

Out of the 180 shoppers surveyed, 73 thought that Bury was the best market, with 56 for Bolton, 21 for Rochdale, 14 for Tommyfield in Oldham and 11 for Openshaw Market in Radcliffe. The remainder voted for 'others'.

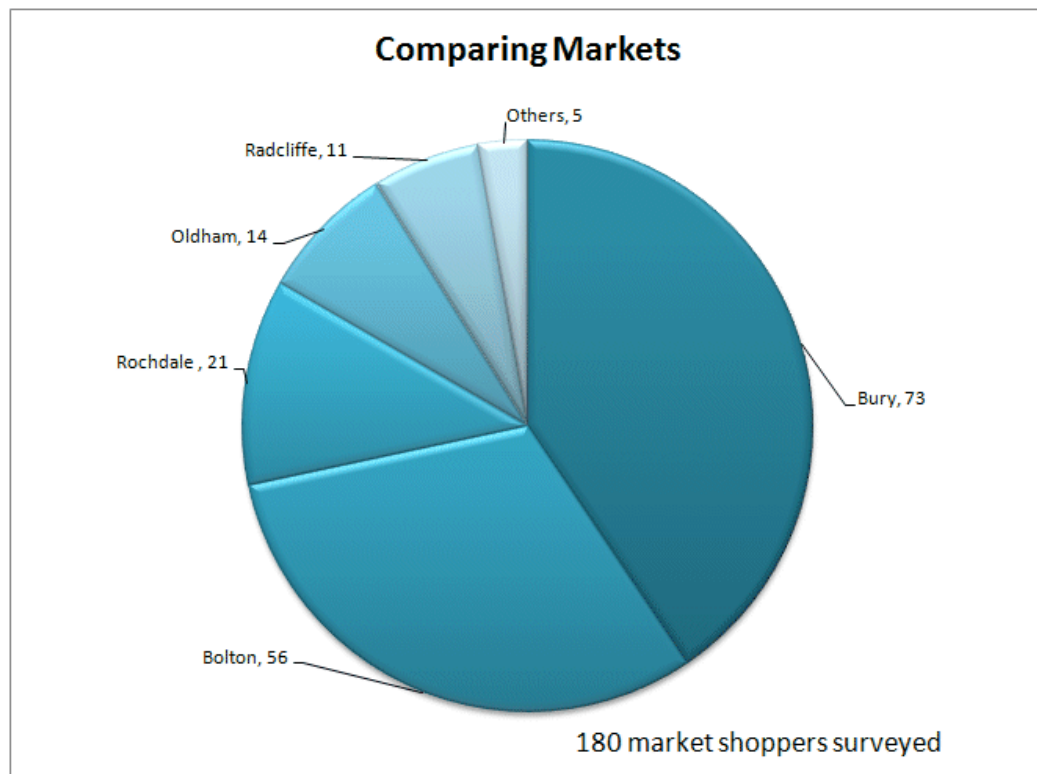
Produce a pie chart using the same data, including the 'others' option. Include labels and counts, as well as the total, on the chart.

The table of values is shown below, with 'others' found by subtracting the known total from 180:

Market	No. of shoppers	Pie chart angle
Bury	73	$146^\circ$
Bolton	56	$112^\circ$
Rochdale	21	$42^\circ$
Oldham	14	$28^\circ$
Radcliffe	11	$22^\circ$
Others	5	$10^\circ$
<b>TOTAL</b>	<b>180</b>	<b><math>360^\circ</math></b>

There are 180 shoppers and  $360^\circ$  in the pie chart, so 1 shopper is denoted by a  $\frac{360}{180}^\circ$  or  $2^\circ$  pie-slice.

The 'Bury' portion of the pie will thus be  $(73 \times 2)^\circ$  or  $146^\circ$ , Bolton  $(56 \times 2)^\circ$  or  $112^\circ$ , Rochdale  $42^\circ$ , Tommyfield (Oldham)  $28^\circ$ , Openshaw (Radcliffe)  $22^\circ$ , and 'others'  $10^\circ$ .



The data could have also been shown satisfactorily on a bar chart, but because the data is not numeric, a frequency polygon would have been unsuitable.

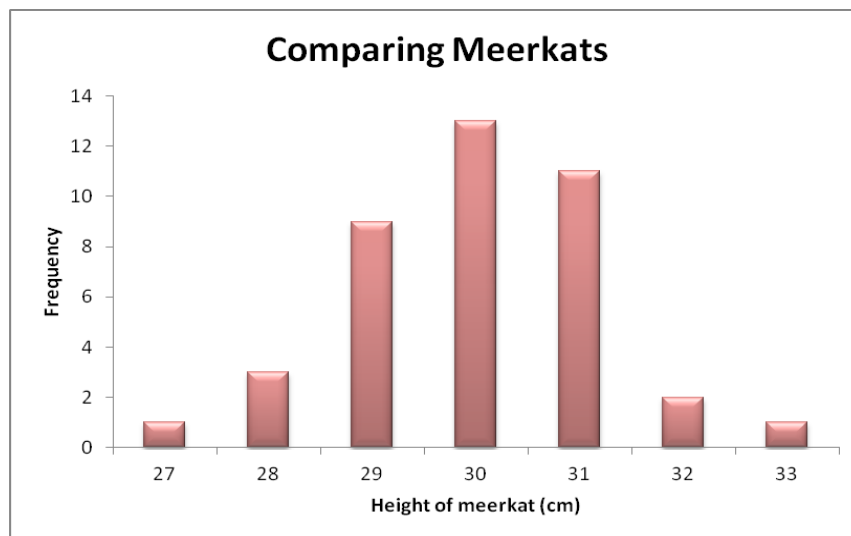
**Example (9):** The heights of 40 adult meerkats (to the nearest cm) on a nature reserve were recorded as follows:

Height (cm)	27	28	29	30	31	32	33
Frequency	1	3	9	13	11	2	1

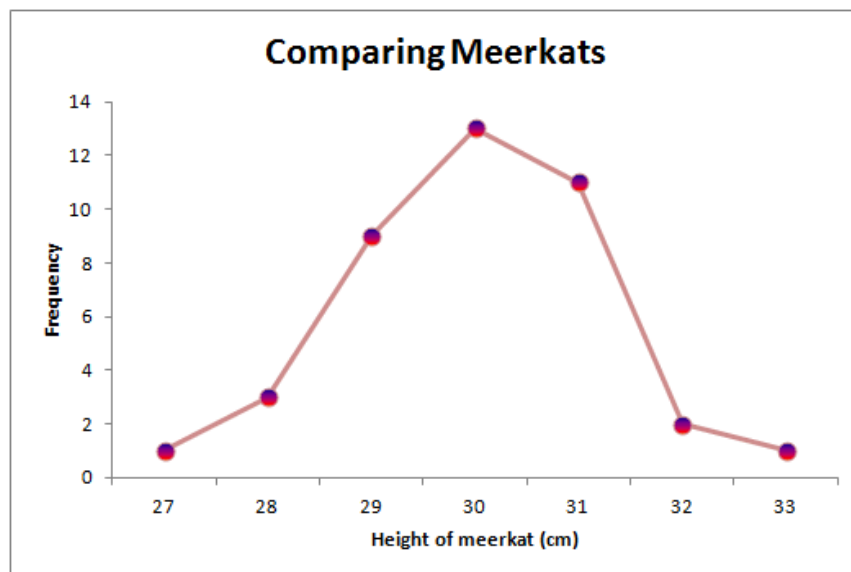
This time, the heights of the meerkats are the group labels.

The group labels (data items) are numeric and ordered this time, so the data can be illustrated by either a bar chart or a line chart (“frequency polygon”).

Bar chart:



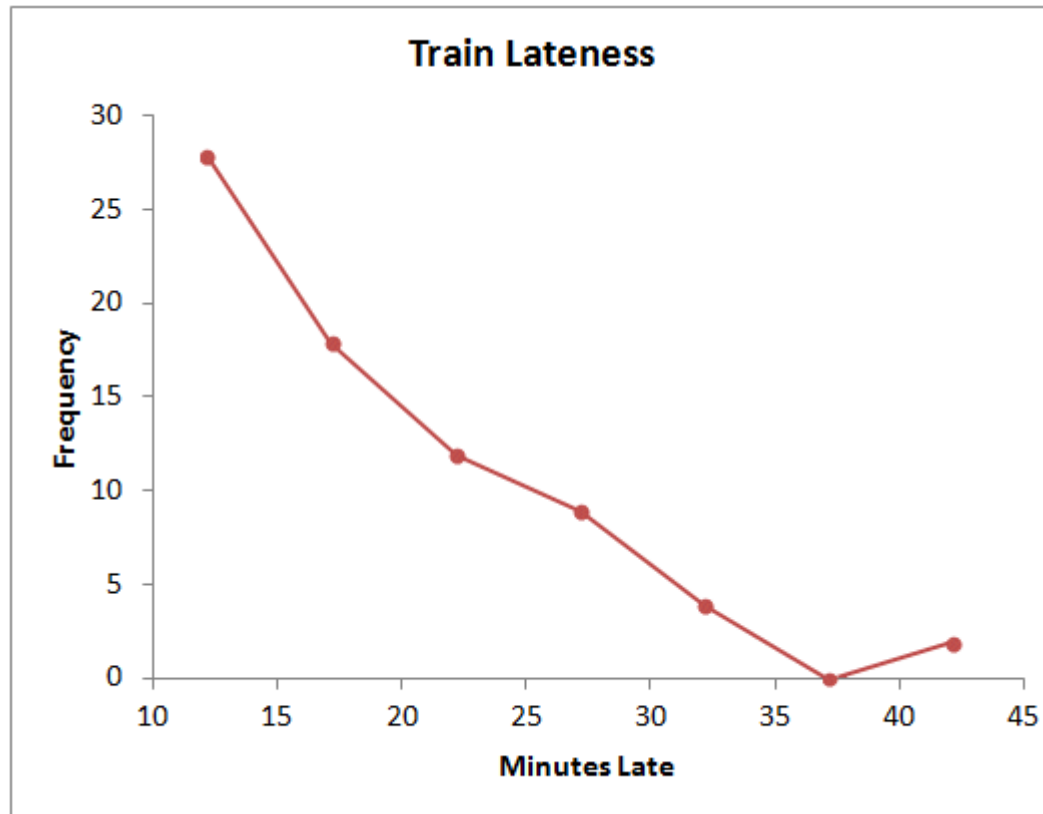
Frequency polygon:



**Example (10):** Here are the lateness statistics for trains arriving at Manchester Piccadilly railway station on a certain day. (Trains arriving within 10 minutes of their expected arrival time were not considered late.)

Lateness (minutes)	10-14	15-19	20-24	25- 29	30-34	35-39	40-44
Frequency	28	18	12	9	4	0	2

Express these statistics in a frequency polygon.



This time, the data is numeric and grouped, so we have to line up the points on the frequency polygon against the *midpoint* of each class interval, and not at either end.

The first data point, corresponding to the '10-14 minutes late' category, is therefore vertically above the '12 minutes' point on the horizontal axis, since 12 is halfway between 10 and 14.

**Example (11):** The table below shows the winners of the football World Cup from 1930 to 2014, along with the number of times each winning team had lifted the trophy.

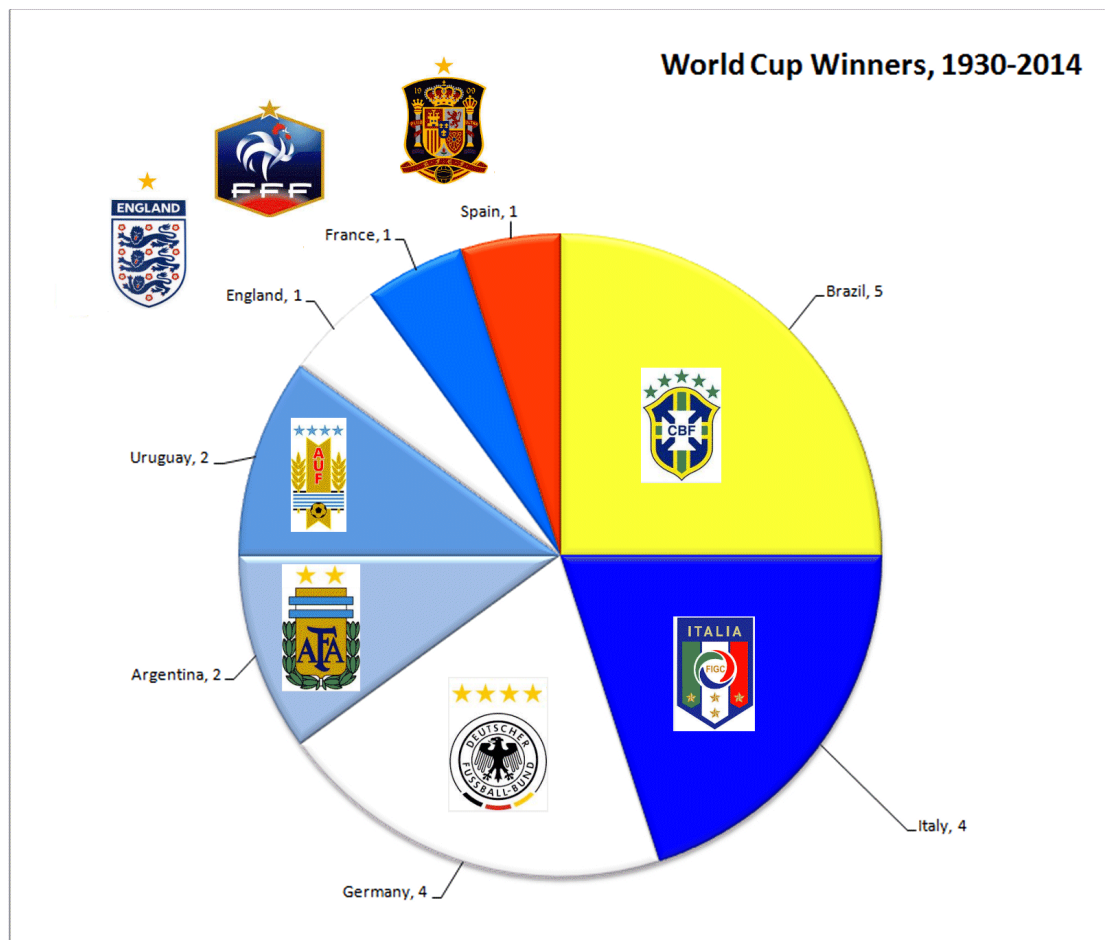
Winners	No. of wins
Brazil	5
Italy, Germany	4 each
Argentina, Uruguay	2 each
England, France, Spain	1 each

Illustrate this data on a pie chart.

We have 20 World Cups in total, so each single win equates to a  $\frac{360}{20}^\circ$ , or an  $18^\circ$  slice of the pie.

For example Brazil's slice, with 5 World Cup victories, takes up  $5 \times 18^\circ$ , or  $90^\circ$ .

Winners	No. of wins	Pie angle
Brazil	5	$90^\circ$
Italy, Germany	4 each	$72^\circ$ each
Argentina, Uruguay	2 each	$36^\circ$ each
England, France, Spain	1 each	$18^\circ$ each



(The question did not ask for the team badges, but it gives the chart a more professional effect !)

### Clustered and Stacked Bar Charts.

**Example (12)a:** Dave has been monitoring his mean monthly spending on gas, electricity and water over a four-year period from 2016 to 2019. He had made two changes to his lifestyle in 2019, firstly, he had replaced his gas cooker with an electric one, and secondly, he had bought an electric car.

#### Mean spending per month, £

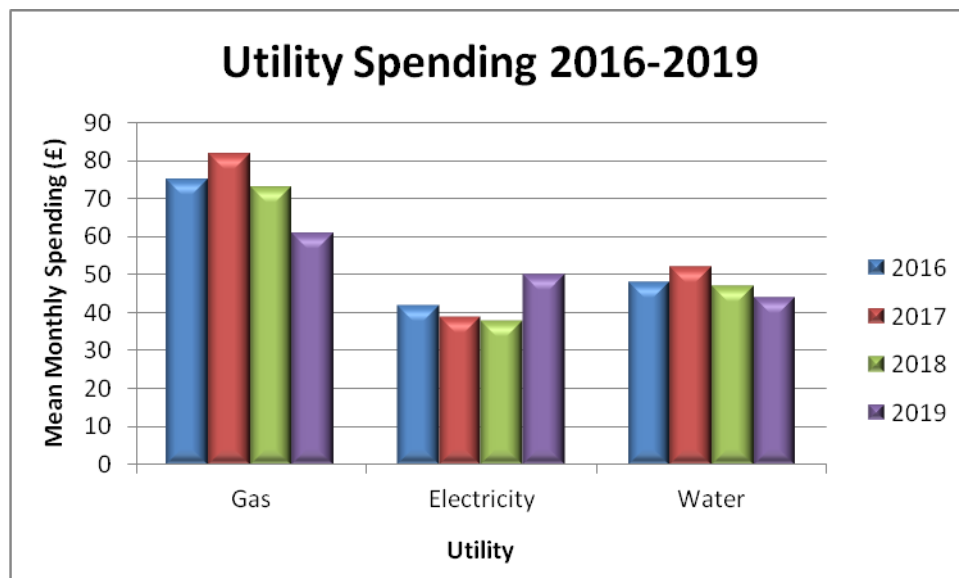
	2016	2017	2018	2019
Gas	75	82	73	61
Electricity	42	39	38	50
Water	48	52	47	44

He wants to use a bar chart to show how his spending on each utility over the years had changed over that four-year period, so he decides to cluster them into three groups of four, one for each utility

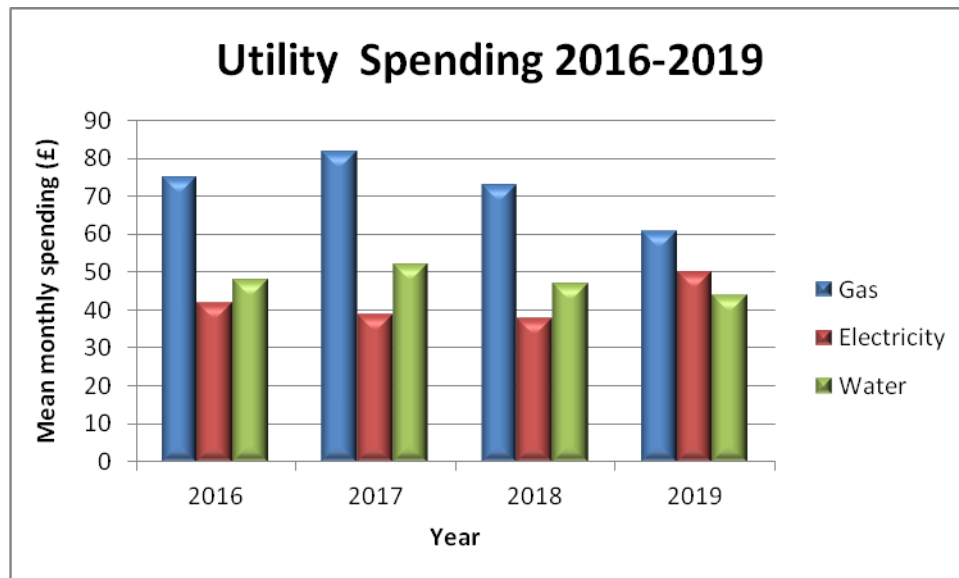
Thus the first cluster of four bars refers to his monthly spending on gas, from one year to the next. The monthly spending figures for electricity and water are grouped in a similar fashion.

It can be seen that Dave's spending on both gas and water peaked in 2017 before declining in 2018 and 2019. His spending on electricity, however, rose in 2019 after three years of slow decline.

In short, the comparison is between the *years* (the horizontal data label) , with each *utility* represented by a grouped bar.



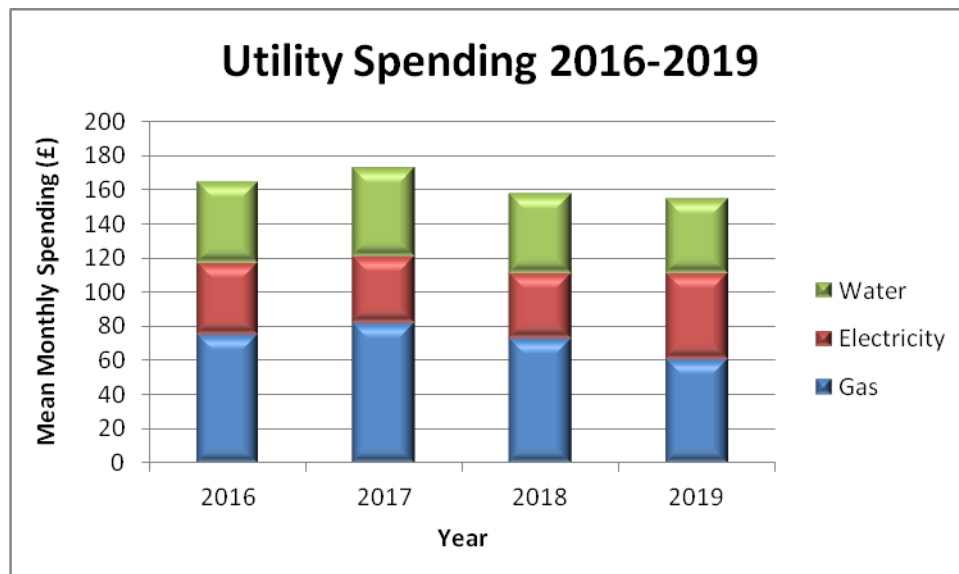
**Example (12)b:** Dave then wants to compare his *total* monthly spending on utilities, so he produces the bar chart below, where the comparison is between the *utilities* (the horizontal data label), with each *year* represented by a grouped bar. In other words, he groups the bars in four sets of three, one for each year.



This graph is not particularly useful for showing annual totals, because Dave would still have to add the values per utility.

**Example 12(c):** Dave wants to see how his *total* mean monthly spending on utilities had been trending over the last four years. .Show the results on a stacked bar chart.

Instead of grouping the columns by year, we can *stack* the utility columns for each year. The *totals* are clearer to see, even though the spending on *individual* utilities is less so.





### **Finding averages – the mode, median and mean.**

There are three different averages involving **numeric** data that feature heavily in statistics – the mode, the median and the mean.

The **mode** is simply the data item with the highest frequency, i.e. the one which occurs most often. The mode of the set of numbers 1, 2, 5, 3, 4, 2, 5, 4, 3, 4 is thus 4, since it occurs three times in the list.

The mode is particularly suited when ordering items for a shop, for instance - think mode = fashion.

Alone of the three averages, the mode can apply to non-numeric data as well.

As an example, a sports shop's replica football shirt sales by size on a certain week were 12 small, 15 medium, 19 large and 16 extra large. The modal size in terms of sales that week was 'large'.

The **median** can be worked out by sorting the data items in order, and finding the middle number.

The median of the (sorted) set of numbers 10, 12, 15, 19, 22, 26, 28 is therefore 19, because there are 7 numbers in all, and 19 is the fourth one, having exactly as many below it as above it.

Note: If the sample size is even, there will be two 'middle numbers'. The median is then found by adding the two middle numbers and halving the result.

Thus, the median of the set 12, 15, 19, 27, 32, 35, 41, 47 is taken as half of  $(27 + 32)$  or  $29\frac{1}{2}$ .

The **mean** can be worked out by adding the data items and dividing the result by the size of the sample. Thus, the mean of the set 12, 15, 19, 27, 32, 35, 41, 47 can be found by summing the numbers. The sum is 228 and there are 8 numbers in all, so the mean is equal to  $228 \div 8 = 28\frac{1}{2}$ .

This last set of numbers illustrates a disadvantage of the mode and median.

Since each number occurs only once, there is no mode. Also, the median of  $29\frac{1}{2}$  does not correspond to any actual number from the group, though that is less of an issue, as neither does the mean.

One more feature of a set of data is the **range** – this is simply the largest number in the group minus the smallest one.

**Example (12):** Taye received the following scores over an 11-week period on *Strict Dance*:

29, 31, 36, 29, 33, 39, 38, 36, 38, 38, 40

Find her modal, median and mean scores, as well as the range.

#### **Mode.**

The mode is simply the data item with the highest frequency, and Taye had received a score of 38 more often than any other. Her modal score is therefore 38.

#### **Median.**

We place the scores in numerical order as 29, 29, 31, 33, 36, **36**, 38, 38, 38, 39, 40.

The middle number is the 6<sup>th</sup> one in the list, having exactly as many below it as above it.

Hence Taye's median score is 36.

#### **Mean.**

The total of all the 11 scores is 387, so Taye's mean score is  $387 \div 11 = 35.2$ .

#### **Range.**

The highest score was 40, and the lowest was 29, so the range of Taye's scores is  $40 - 29 = 11$ .

**Example (13):** A survey was carried out to find out how many portions of fruit and vegetables each pupil in a class of 32 was eating per day. The results were given in the table below.

Portions	0	1	2	3	4	5	6	7	8
No. of pupils	2	3	6	10	5	2	1	0	3

Find the mode, median and mean of the values in the table.

**Mode.**

The mode is simply the data item with the highest frequency. More pupils (namely 10) have 3 portions of fruit and vegetables a day than any number of portions, so 3 is the mode.

**Median.**

To find the median, we could write out the data in full from the table as a list of 2 '0's, 3 '1's, 6 '2's and so forth:

0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 8, 8, 8

There are 32 numbers in all, so there are two middle numbers – the 16<sup>th</sup> and 17<sup>th</sup> in the list. Each is a 3, so 3 is the median.

This method was a little long-winded given the fact that the data was already grouped and sorted for us.

We could reason as follows;

Pupils 1-2 (2 of them) had no fruit and vegetables

Pupils 3-5 (3 of them) had one portion of fruit and vegetables

Pupils 6-11 (6 of them) had two portions

Pupils 12-21 (10 of them) had three portions, and so on..

The 16<sup>th</sup> and 17<sup>th</sup> pupils both fall into the '3 portions' category, so the median is 3.

**Mean.**

To find the mean, we need an extra row for subtotals and an extra column for totals as shown below.

Portions (Number) $x$	0	1	2	3	4	5	6	7	8	<b>TOTAL</b>
Pupils (Frequency) $f$	2	3	6	10	5	2	1	0	3	<b>32</b>
<b>Portions <math>\times</math> Pupils, <math>xf</math></b>	<b>0</b>	<b>3</b>	<b>12</b>	<b>30</b>	<b>20</b>	<b>10</b>	<b>6</b>	<b>0</b>	<b>24</b>	<b>105</b>

(Note: some books have the tables of data shown with columns and rows interchanged. The preference for the row format is purely arbitrary.)

We have multiplied the number of portions ( $x$ ) by the number of pupils ( $f$ ) to work out the total number of portions of fruit and vegetables eaten by the class.

As an example, there are 5 pupils who eat 4 portions of fruit and vegetables. They therefore contribute  $5 \times 4$  or 20 portions towards the total.

The total number of portions eaten by the class is 105, and so the mean is  $\frac{105}{32}$  or 3.3.

The range of the number of portions of fruit and vegetables eaten is  $8 - 0$  or 8.

**Example (14):** A football club has a squad of 23 players, and each player is currently on one of five different weekly wage structures.

Wage in £K	11	12	14	17	21
No. of players	4	4	6	5	4

Find the mode, median, mean and range of the players' wages.

**Mode:** The modal wage is £14 K, as more players are on that wage than any other.

**Median:** Eight players earn under £14 K, and the 9<sup>th</sup> to the 14<sup>th</sup> players earn £14 K.

There are 23 players in all, so the middle player's position is  $\frac{1}{2}(23+1)$  or 12. This 12<sup>th</sup> player in the list is in the £14 K wage band, and so the median wage is also £14 K.

**Mean:** We need an extra row and extra column for subtotals and totals as shown below.

Wage in £K, $x$	11	12	14	17	21	<b>TOTAL</b>
No. of players, $f$	4	4	6	5	4	<b>23</b>
<b>Sub-total, <math>xf</math></b>	<b>44</b>	<b>48</b>	<b>84</b>	<b>85</b>	<b>84</b>	<b>345</b>

There are 23 players earning a total weekly wage of £345 K, so the mean wage =  $\frac{345}{23} = 15$  or £15 K.

Finally, the range of the footballers' wages is £21 K - £11 K, or £10 K.

**Example (15):** The football club in Example (14) has lost one of their players on a wage of £17 K through retirement. To replace him, they have signed a Brazilian, Rondinho, and put him on a £155 K weekly wage. How will Rondinho's signing affect each average wage, and the range?

The table of players' wages will now look like this:

Wage in £K	11	12	14	17	21	155
No. of players	4	4	6	4	4	1

**Mode:** There are still more players on £14 K than on any other wage, so the mode is unaffected. (Had one of the players on a weekly wage of £14 K, there would have been **two** modal weekly wages, £14 K and £17 K.)

**Median:** The 9<sup>th</sup> to the 14<sup>th</sup> players still earn £14 K, and so the median is still unaffected at £14 K.

Wage in £K, $x$	11	12	14	17	21	155	<b>TOTAL</b>
No. of players, $f$	4	4	6	4	4	1	<b>23</b>
<b>Sub-total, <math>xf</math></b>	<b>44</b>	<b>48</b>	<b>84</b>	<b>68</b>	<b>84</b>	<b>155</b>	<b>483</b>

The 23 players are now earning a total weekly wage of £483 K, so the mean wage =  $\frac{483}{23}$  or £21 K.

This is a considerable difference from the earlier value of £15 K.

The range of the footballers' wages is now £155 K - £11 K, or £144 K.

Example (15) showed how different averages were differently affected by a value outside the 'normal' range, or an **outlier** – here it was the Brazilian footballer's wage.

The mode is not generally affected by outliers, and neither is the median. The mean and the range, however, are more sensitive, which is why care needs to be taken in selecting the most appropriate average.

**Example (16):** The heights of 25 sunflower plants grown in the same garden from the same seed packet were measured (in centimetres, rounded to the nearest cm) on one day and sorted in order.

186, 189, 192, 193, 194, 196, 202, 203, 204, 207, 207, 208, 209,  
212, 214, 215, 216, 216, 218, 219, 219, 222, 225, 228, 234.

Find the mode, median, mean and range.

**Mode:** There are three heights vying for the mode here – 207, 216 and 219 cm. These values each occur twice whilst all the others occur only once.

**Median:** The middle (13<sup>th</sup>) number out of the sorted 25 is 209, so the median height is 209 cm.

**Mean:** The sum of the heights (in cm) is 5228, so the mean is  $\frac{5228}{25}$  cm or 209.12 cm.

**Range:** The range is 234 – 186 cm, or 48cm.

This time, it seems that the mode is unsuitable because the range of values is large, the sample fairly small, and there is no one highly popular value. The median and mean are both suitable, because there are no outlying values and the distribution is uniform.

### Grouped Data.

When we have a large set of data, it is often more convenient to arrange it into groups.

The last example featured the individual heights of 25 sunflower plants, but grouping them helps illustrate the distribution more plainly. The following example has the same data, but grouped by **class intervals** of 10cm height.

**Example (17):** Find the modal and median classes from the grouped heights of sunflower plants below:

Height Range (cm)	180-189	190-199	200-209	210-219	220-229	230-239
Frequency	2	4	7	8	3	1

We can no longer find the actual mode or the median, but we can still find the class intervals in which they lie.

The **modal class** is 210-219 cm, since it has the highest frequency.

To find the **median class**, we now look for the class interval where the 13<sup>th</sup> data item occurs. The 13<sup>th</sup> data item occurs in the 200-209 cm class, so 200-209 cm is the median class

### Estimating the mean from grouped data.

Estimating a mean from grouped data needs a little more work. Because we no longer have actual data, the best we can do is assume that the data items are distributed evenly within their class intervals.

We therefore assume that all the data items take a class midpoint value, taken to be the mean of the end values of the class interval.

More often than not, due to the nature of the class groupings, the midpoint value would not seem 'exact', and have a messy ".5" in it. This is especially true for discrete or rounded data, where there are gaps between class intervals.

Thus, if a class interval of rounded data is quoted as '30-39', the effective class interval is 29.5 to 39.5, and the midpoint would be 34.5 and not 35.

Similarly a range quoted as '61-80' would have a midpoint of 70.5, not 70.

**Example (17a):** Estimate the mean from the grouped heights of sunflower plants below.  
 (Values have been rounded to the nearest cm.)

Height Range (cm) (Class Interval)	180- 189	190- 199	200- 209	210- 219	220- 229	230- 239
Frequency	2	4	7	8	3	1

Notice the gaps between the class intervals here at 189 to 190, 199 to 200 and so forth – indicative of discrete data.

Firstly we work out the class interval midpoints and place them in another row:

Height Range (cm) (Class Interval)	180- 189	190- 199	200- 209	210- 219	220- 229	230- 239
Frequency	2	4	7	8	3	1
<b>Class Interval Midpoint</b>	<b>184.5</b>	<b>194.5</b>	<b>204.5</b>	<b>214.5</b>	<b>224.5</b>	<b>234.5</b>

Finally we add a subtotals row and a totals column:

Height Range (cm) (Class Interval)	180- 189	190- 199	200- 209	210- 219	220- 229	230- 239	<b>TOTAL</b>
Frequency, <i>f</i>	2	4	7	8	3	1	<b>25</b>
Class Interval Midpoint, <i>m</i>	184.5	194.5	204.5	214.5	224.5	234.5	
Midpoint × Frequency, <i>mf</i>	<b>369</b>	<b>778</b>	<b>1431.5</b>	<b>1716</b>	<b>673.5</b>	<b>234.5</b>	<b>5202.5</b>

We then sum the "Midpoint × Frequency" values and divide them by the sample size.

The estimated mean is  $\frac{5202.5}{25}$  or 208.1 cm.

Note how this is slightly different from the actual mean of 209.25 found in Example (7).

**Example(18):** The lifetimes (in hours, rounded to the nearest hour) of 50 Alko batteries were tested under controlled conditions in a simulation lab.

The results were as follows:

Life (hrs)	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59
Frequency	2	3	4	7	13	11	6	4

Note that this is discrete data, hence the gaps between 24 and 25, 29 and 30 and so on.

Find the modal and median classes, and estimate the mean.

The **modal class** is 40-44 hours lifetime, as it has the highest frequency.

The **median class** is the one containing the 25<sup>th</sup> and 26<sup>th</sup> items.  
 This is the 40-44 hours class, actually containing the 17<sup>th</sup>-29<sup>th</sup> items.

To estimate the **mean**, we firstly find the class midpoints:

Life (hrs), $x$	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59
Frequency, $f$	2	3	4	7	13	11	6	4
Class midpoint, $m$	<b>22</b>	<b>27</b>	<b>32</b>	<b>37</b>	<b>42</b>	<b>47</b>	<b>52</b>	<b>57</b>

Next, we add the subtotals row and totals columns:

Life (hrs), $x$	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	<b>TOTAL</b>
Frequency, $f$	2	3	4	7	13	11	6	4	<b>50</b>
Class midpoint, $m$	22	27	32	37	42	47	52	57	
<b>Mid. <math>\times</math> Freq. , <math>mf</math></b>	<b>44</b>	<b>81</b>	<b>128</b>	<b>259</b>	<b>546</b>	<b>517</b>	<b>312</b>	<b>228</b>	<b>2115</b>

The estimated mean is  $\frac{2115}{50}$  or 42.3 hours.

If the source data is taken to be continuous, then the class interval midpoints are easier to find. The vast majority of examination questions assume continuous data, as in the example below.

Note how the class intervals are delimited by inequality symbols.

**Example (19):** An electrical retailer has analysed the money spent by customers in the first 500 transactions in the January sales, excluding purchases greater than £1000. The results were tabulated as follows:

Spending, $x$ , in £	$x < 50$	$50 \leq x < 100$	$100 \leq x < 200$	$200 \leq x < 400$	$400 \leq x < 600$	$600 \leq x \leq 1000$
Frequency, $f$	70	110	100	80	60	80

Estimate the mean amount of money spent per customer transaction by completing the table below:

Spending, $x$ , in £	$x < 50$	$50 \leq x < 100$	$100 \leq x < 200$	$200 \leq x < 400$	$400 \leq x < 600$	$600 \leq x \leq 1000$
Frequency, $f$	70	110	100	80	60	80
Class midpoint, $m$	25	75	150			
Mid. $\times$ Freq., $mf$						

The completed data table looks like this:

Spending, $x$ , in £	$x < 50$	$50 \leq x < 100$	$100 \leq x < 200$	$200 \leq x < 400$	$400 \leq x < 600$	$600 \leq x \leq 1000$	<b>TOTAL</b>
Frequency, $f$	70	110	100	80	60	80	<b>500</b>
Class midpoint, $m$	<b>25</b>	<b>75</b>	<b>150</b>	<b>300</b>	<b>500</b>	<b>800</b>	
Mid. $\times$ Freq., $mf$	<b>1750</b>	<b>8250</b>	<b>15000</b>	<b>24000</b>	<b>30000</b>	<b>64000</b>	<b>143000</b>

The estimated mean spending is therefore £  $\frac{143000}{500}$ , or £286 per customer.



**Example (20):** Sam is a keen gardener and his two most productive apple trees are a Cox and a Gala. The details of his harvest last autumn are as below.

<b>Cox</b>	
Weight of apple, $w$ , (g)	Frequency, $f$
$100 \leq w < 110$	25
$110 \leq w < 120$	60
$120 \leq w < 130$	90
$130 \leq w < 140$	180
$140 \leq w < 150$	140
$150 \leq w < 160$	80
$160 \leq w < 170$	25
<b>TOTAL</b>	<b>600</b>

<b>Gala</b>	
Weight of apple, $w$ , (g)	Frequency, $f$
$120 \leq w < 130$	15
$130 \leq w < 140$	30
$140 \leq w < 150$	140
$150 \leq w < 160$	170
$160 \leq w < 170$	120
$170 \leq w < 180$	25
<b>TOTAL</b>	<b>500</b>

- Write down the modal weight class for each variety of apple, and also the weight class in which the median lies.
- Estimate the total yield of apples (in kg), and hence the mean weight of an apple, from each tree.
- Sam reckons that the Cox tree had produced a heavier crop than the Gala, even though the individual apples from the Cox tree were of lower weight.

Comment on Sam's statement, giving supporting evidence or otherwise.

i) The modal weight class for the Cox apples is the  $130 \leq w < 140$  class, i.e. apples weighing between 130g and 140g each. The corresponding class for the Gala apples is the  $150 \leq w < 160$  class, or apples in the 150g-160g weight range.

ii) The total number of Cox apples is 600, so the median class interval is the one containing the 300<sup>th</sup> and 301<sup>st</sup> apples. This is the " $130 \leq w < 140$ " weight interval, as it contains the 176<sup>th</sup> to 355<sup>th</sup> apples.

By similar reckoning, the total number of Gala apples is 500, so the median class includes the 250<sup>th</sup> and 251<sup>st</sup> apples. This is the " $150 \leq w < 160$ " weight interval, containing the 186<sup>th</sup> to 355<sup>th</sup> apples.

Note how the modal and median classes are the same for both varieties of apple.

iii) We need to add class midpoint and sub-total columns in order to estimate the total crop weight of the apples from each tree, and thus the mean weight of a single apple.

<b>Cox</b>			
Weight of apple, $w$ , (g)	Frequency, $f$	Class midpoint, $m$	Mid. $\times$ Freq. , $mf$
$100 \leq w < 110$	25	<b>105</b>	<b>2625</b>
$110 \leq w < 120$	60	<b>115</b>	<b>6900</b>
$120 \leq w < 130$	90	<b>125</b>	<b>11250</b>
$130 \leq w < 140$	180	<b>135</b>	<b>24300</b>
$140 \leq w < 150$	140	<b>145</b>	<b>20300</b>
$150 \leq w < 160$	80	<b>155</b>	<b>12400</b>
$160 \leq w < 170$	25	<b>165</b>	<b>4125</b>
<b>TOTAL</b>	<b>600</b>		<b>81900</b>

The estimated total yield of Cox apples is **81.9 kg**, and as 600 apples were harvested in all, the

estimated mean weight of a single Cox apple is  $\frac{81900}{600}$  g or **136.5 g**.

<b>Gala</b>			
Weight of apple, $w$ , (g)	Frequency, $f$	Class midpoint, $m$	Mid. $\times$ Freq. , $mf$
$120 \leq w < 130$	15	<b>125</b>	<b>1875</b>
$130 \leq w < 140$	30	<b>135</b>	<b>4050</b>
$140 \leq w < 150$	140	<b>145</b>	<b>20300</b>
$150 \leq w < 160$	170	<b>155</b>	<b>26350</b>
$160 \leq w < 170$	120	<b>165</b>	<b>19800</b>
$170 \leq w < 180$	25	<b>175</b>	<b>4375</b>
<b>TOTAL</b>	<b>500</b>		<b>76750</b>

The estimated total yield of Gala apples is **76.75 kg**, and as 500 apples were harvested in all, the

estimated mean weight of a single Gala apple is  $\frac{76750}{500}$  g or **153.5 g**.

iii) From the results, the individual mean weight of a Gala apple is 17g more than that of a Cox, but as Sam harvested more Cox apples, the total estimated yield of Coxes was about 5 kg more than that of the Galas. Sam's assertions were therefore correct.

### Stem and Leaf Diagrams.

Another method of representing data is by the **stem and leaf** diagram.  
 It is a kind of bar chart where the numbers in the distribution act as the bars.

**Example (21):** The heights of 25 sunflower plants grown in the same garden from the same seed packet were measured (in centimetres, rounded to nearest cm ) on one day.

The data is the same as in Example (16).

207, 196, 209, 193, 214, 218, 234, 212, 216, 228, 189, 215, 192, 225, 203, 219, 194, 222, 186, 216, 204, 207, 208, 219, 202

Draw a stem and leaf diagram to illustrate the data.

First, we see that the smallest value is 186 and the largest is 234.  
 Each number can be divided into tens and units, and the 'tens' go in the left-hand columns of a table. This area to the left is the 'stem'.

Next, we gradually add values to the right of the column to give the 'leaves'.  
 These are the 'units' digits of the data.

**18**  
**19**  
**20**  
**21**  
**22**  
**23**

(We could have used sorted data as in Example 6, but this example will show how to create an **ordered** stem and leaf diagram from **unsorted** data)

The first data item is 207, so we add a 'leaf' with a 7 to the right of the 20 in the stem.

<b>18</b>	<b>18</b>	<b>18</b>	<b>18</b>	<b>18</b>	<b>18</b>	<b>18</b>
<b>19</b>	<b>19</b>	<b>19</b>	<b>6</b>	<b>19</b>	<b>6</b>	<b>19</b>
<b>20</b>	<b>20</b>	<b>7</b>	<b>20</b>	<b>7</b>	<b>20</b>	<b>7 9</b>
<b>21</b>	<b>21</b>	<b>21</b>	<b>21</b>	<b>21</b>	<b>21</b>	<b>21</b>
<b>22</b>	<b>22</b>	<b>22</b>	<b>22</b>	<b>22</b>	<b>22</b>	<b>22</b>
<b>23</b>	<b>23</b>	<b>23</b>	<b>23</b>	<b>23</b>	<b>23</b>	<b>23</b>

Next, we place 196 by adding a leaf with a 6 to the right of the 19 in the stem.  
 Then we put 209 to the right of the 7 in the '20' stem.

The first five value inputs are shown below.

After all 25 items have been entered, the stem and leaf diagram looks like the diagram below left. The data on the 'leaf' side is still not sorted, so we arrange the 'leaf' data in order as per the finished stem and leaf diagram on the right. A 'key' has also been included.

**19 | 2 = 192**

<b>18</b>	<b>96</b>	<b>18</b>	<b>69</b>
<b>19</b>	<b>6324</b>	<b>19</b>	<b>2346</b>
<b>20</b>	<b>7934782</b>	<b>20</b>	<b>2347789</b>
<b>21</b>	<b>48265969</b>	<b>21</b>	<b>24566899</b>
<b>22</b>	<b>852</b>	<b>22</b>	<b>258</b>
<b>23</b>	<b>4</b>	<b>23</b>	<b>4</b>

We can see here that the modal class is the 210-219, and that the median is 209 by counting 13 numbers down the **ordered** stem and leaf diagram.

### Scatter Diagrams.

Scatter diagrams, or scatter plots, are used to investigate connections between two features or variables, such as maths exam results against science exam results, or height against weight.

Scatter plots are two-dimensional, and are used to infer whether two sets of data are connected - in other words, correlated.

If there is strong correlation, then the data will tend to fall in a linear pattern; if there is little or no correlation, the data will generally fall in a random pattern.

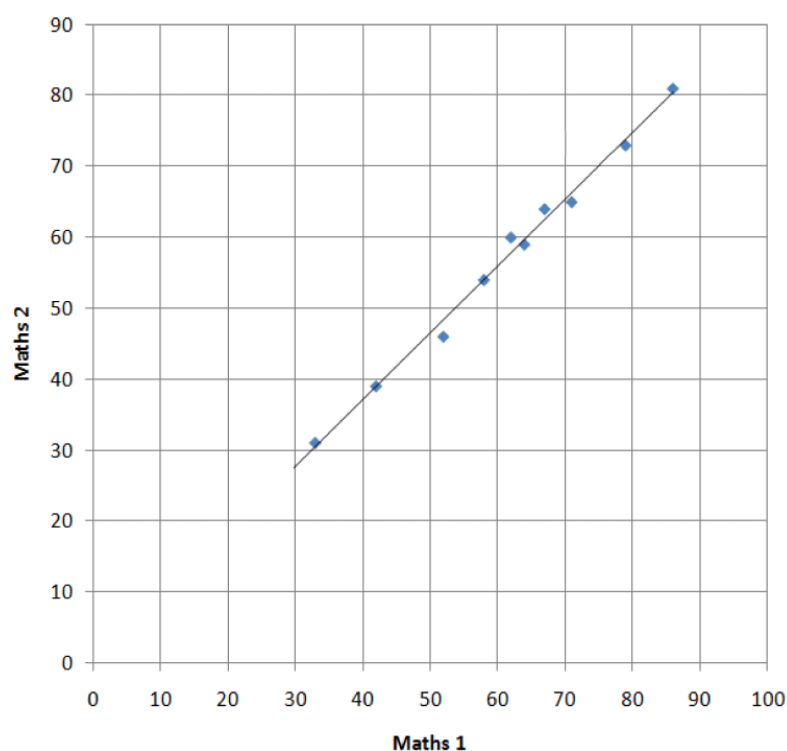
**Example (22):** Ten pupils were chosen at random from Year 11 and the percentage scores for the mock Maths 1 and Maths 2 exams were obtained as follows (Maths 1 quoted first).

64, 59	79, 73	86, 81	42, 39	58, 54
67, 64	33, 31	52, 46	71, 65	62, 60

Plot the pairs of results on a scatter diagram and attempt to draw a best-fit line. Is there a correlation between the result sets ?

i) There is a strong positive correlation between the results of the two maths exams, and hence it is easy to plot a line of best fit.

Scatter plots can also be used for the purposes of estimation, especially if the correlation between sets of data is strong.



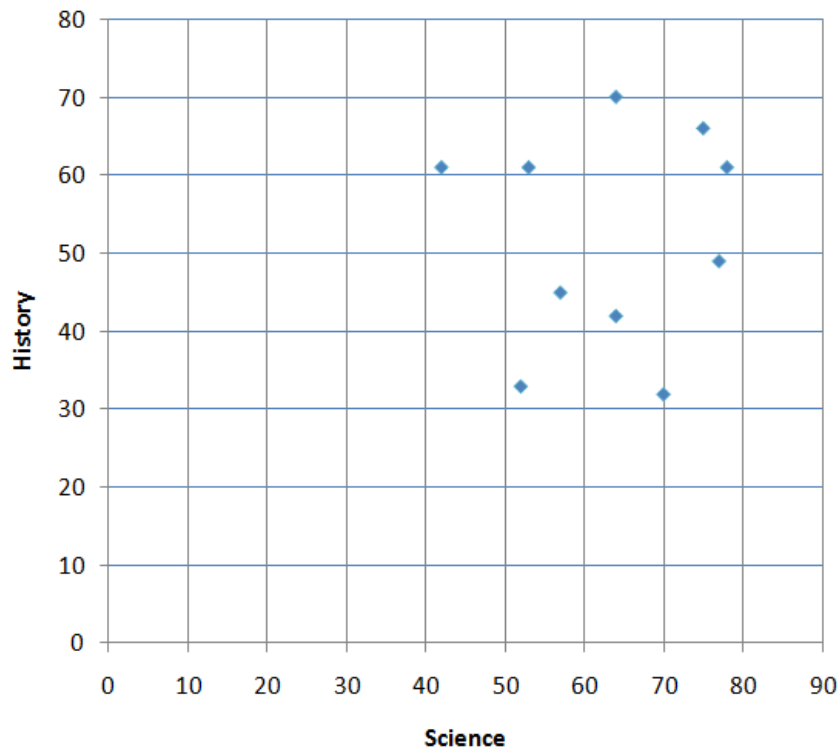
(Exam questions will usually have most of the points already plotted, with the pupil only expected to plot a few of them.)

**Example (23) :**

Eleven pupils were chosen at random from Year 11 and the results for the mock Science and History exams were obtained as follows (Science quoted first).

Plot the resulting scatter diagram. Is there any correlation at all between the data sets ?

78,61	77,49	75,66	70,32	64,42
64,70	57,45	53,61	52,33	42,61
35,51				



The results appear to be randomly scattered here, with no correlation at all apparent.

### The Two-Way Table.

A two-way table is used to illustrate two different sets of information about a sample of individual people or objects.

**Example (24):** A survey was carried out on the timekeeping of inter-city and suburban trains arriving at Manchester Piccadilly station over a 3-hour period on a certain day.  
Trains were recorded as either arriving on time or being late.

There were 114 train times surveyed in all, and of those 114, 42 were inter-city services.  
Of the suburban services, 63 arrived on time, and of the inter-city ones, three arrived late.

Complete the two-way table.

This partially-completed two-way table shows all the data as initially given:

	On time	Late	<b>Total</b>
Inter-city		3	<b>42</b>
Suburban	63		
<b>Total</b>			<b>114</b>

From there, it is a matter of simple arithmetic to complete the table.

There were 42 inter-city arrivals in total and three were late, so 39 were on time.

There were 114 arrivals in total and 42 were inter-city, so 72 were suburban.

There were 72 suburban arrivals and 63 were on time, so nine were late.

Finally, we can total up the columns to have 102 trains on time and twelve late in total.

	On time	Late	<b>Total</b>
Inter-city	39	3	<b>42</b>
Suburban	63	9	<b>72</b>
<b>Total</b>	<b>102</b>	<b>12</b>	<b>114</b>

We can also check the totals in case of any errors !

## Misrepresenting Data.

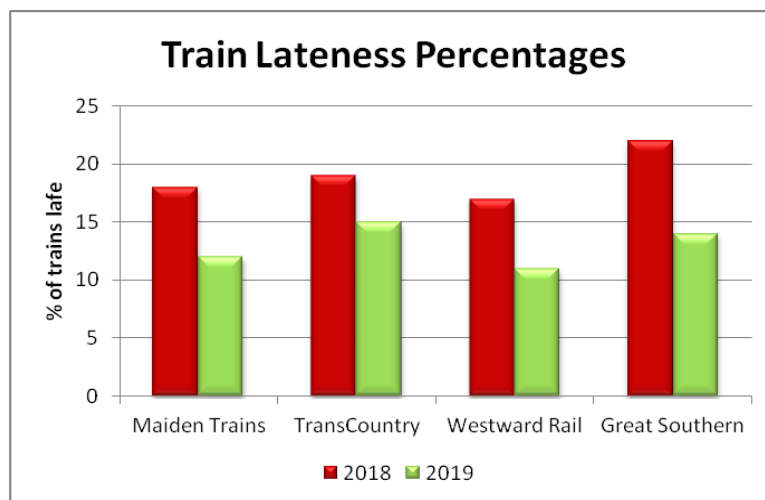
*“There are lies, damned lies, and statistics”.* (Attrib: Charles Wentworth Dilke.)

Statistical data can be distorted in a great many ways – here are some of the more common methods of ‘spin’ used by companies to sell products or by organisations to bolster a particular series of opinions.

### Non-zero vertical scales.

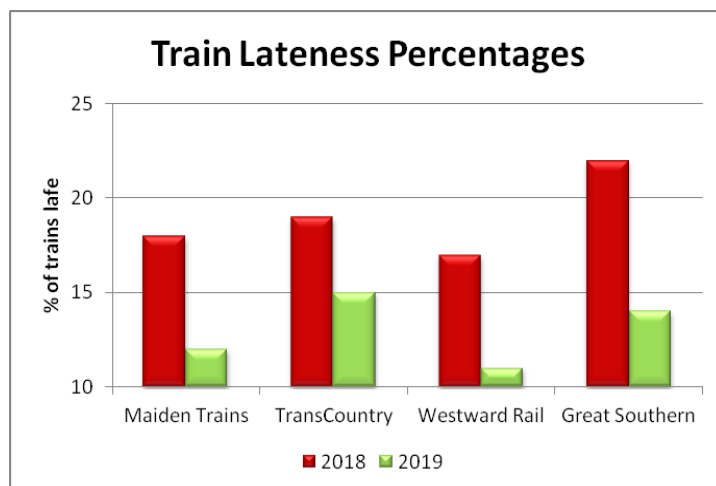
**Example (25):** The bar charts below show the percentages of late-running trains for four railway operators over two years.

The first bar chart has the vertical axis starting on zero and clearly shows how all four train companies had reduced lateness between 2018 and 2019. Thus Westward Rail had 17% of its trains running late in 2018, but that went down to 11% in 2019. By comparison, 12% of Maiden Trains services were late in 2019. All of the bars are of the correct relative height, so we can make valid inferences here.



The second bar chart has been altered by Westward Rail by having the vertical scale starting at 10% rather than 0%, thus making the heights of the bars no longer proportional to the actual percentages and distorting the true values.

Westward Rail’s percentage lateness rate now appears to be half that of its nearest rival, Maiden Trains, when 11% is not half of 12%. Also, the change from the previous year was a decrease from 17 to 11 percentage points, or about a third, yet on this second chart, the decrease in lateness looks enormous !



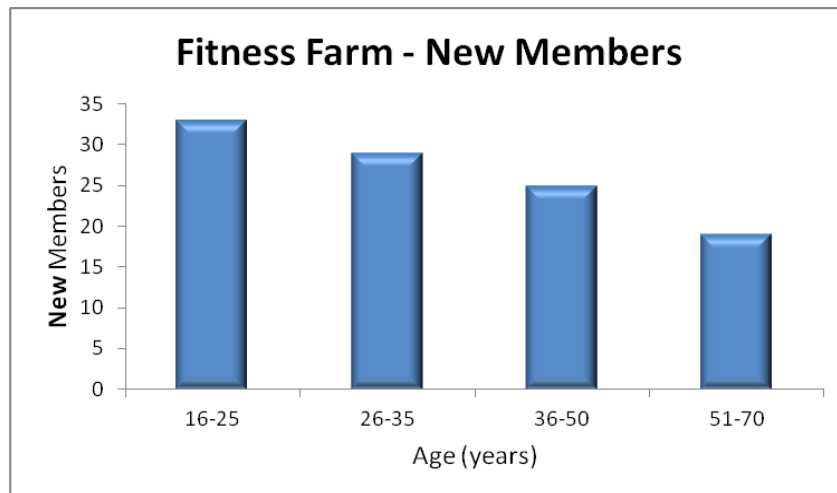
### Misleading grouping of data.

The next example shows another way of ‘fiddling’ a bar chart.

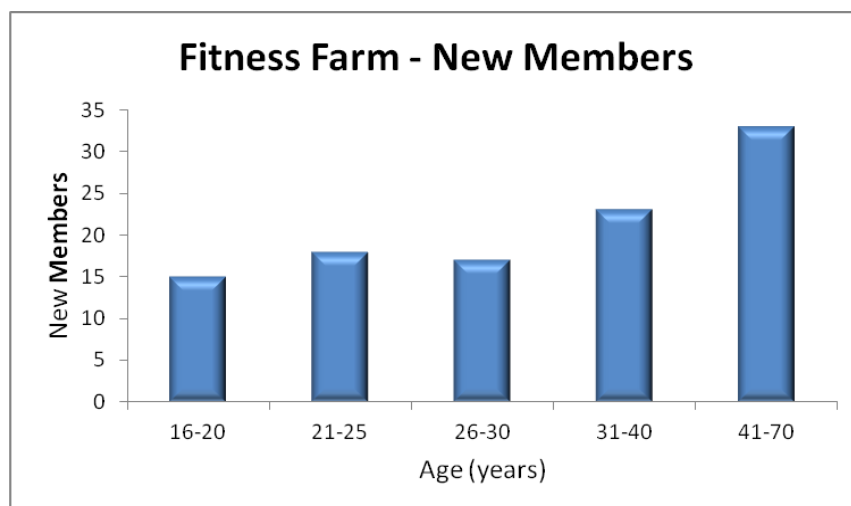
**Example (26):** Fitness Farm gym have published their new membership statistics for the latest year, by members’ ages :

Age range	16-20	21-25	26-30	31-35	36-40	41-50	51-70
New memberships	15	18	17	12	11	14	19

The gym had decided to publish this data as a bar chart in *Youth and Fitness* magazine.



A few weeks later, the following bar chart appeared in *Age and Health* magazine.



Both bar charts represent the same data, but the way in which the age groups are combined produces what appear to be two different bar charts.

In the first chart, it seems as if the gym is more popular with younger people; the second chart appears to lead to the opposite conclusion.

This is because the class intervals (here age ranges) vary greatly in width in each case, but the bar charts tend to hide the fact.

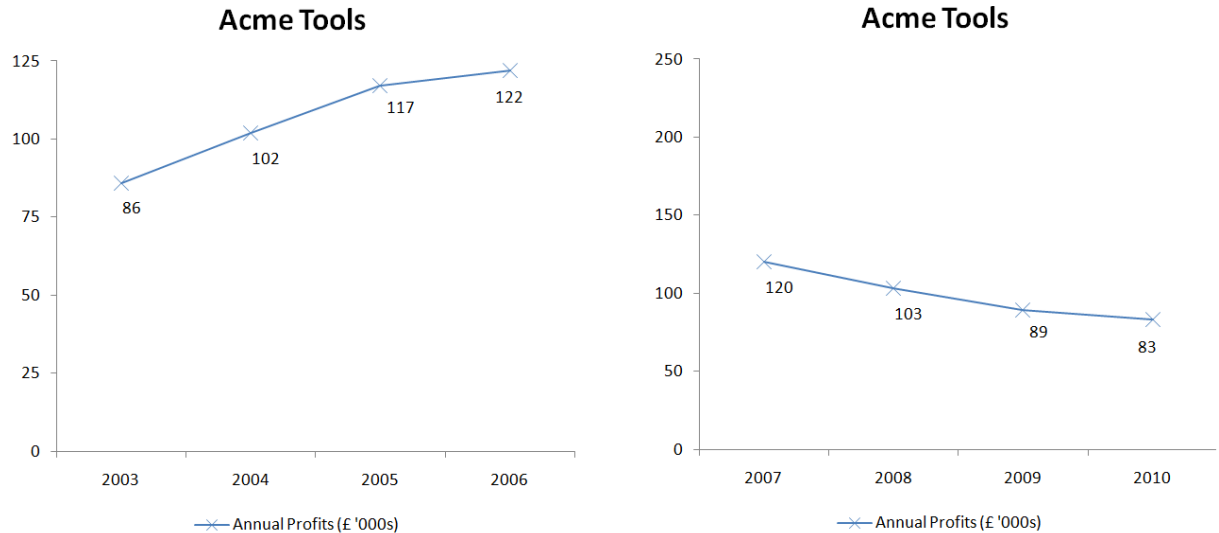
To illustrate the true pattern, it is better to use a histogram or a cumulative frequency diagram.



**Differing vertical scales.**

**Example (27):** Acme Tools showed four years of rising annual profits from 2003 to 2006, followed by four years of falling profits from 2007 to 2010.

These graphs appeared in a company leaflet.

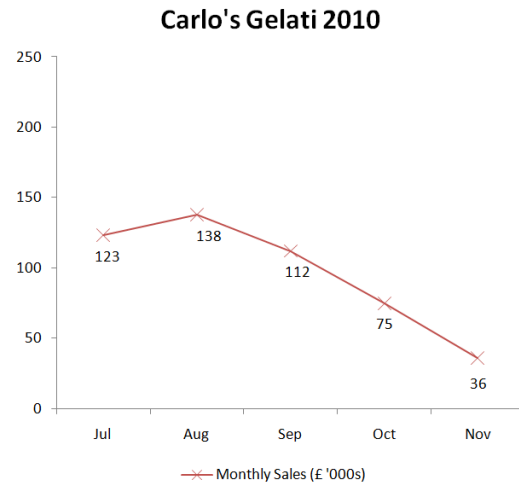
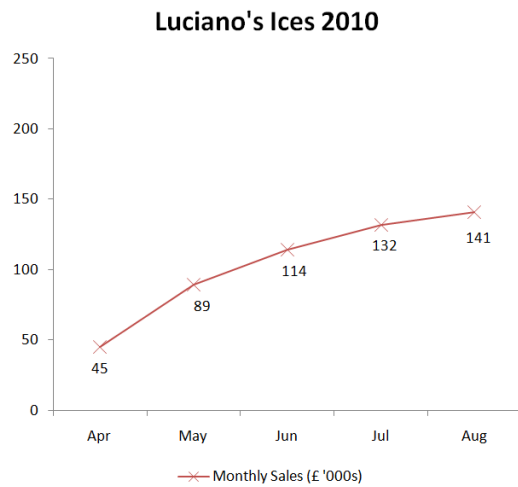


This time, the vertical scales begin at zero for both graphs, but they are still misleading because they use differing scales. While Acme Tools was showing improving profits, the graph exaggerated the fact by using a scale of £25,000 per vertical division. When the profits decreased from 2007 on, the graph used a scale of £50,000 per vertical division.

In short, the graphs have been 'fiddled' to make Acme Tools' good years look better by steepening upward graphs, and the bad years look less bad by flattening downward graphs.

**Differing seasonal timelines.**

**Example (28):** The next two graphs show the sales figures for two local ice-cream makers in 2010.

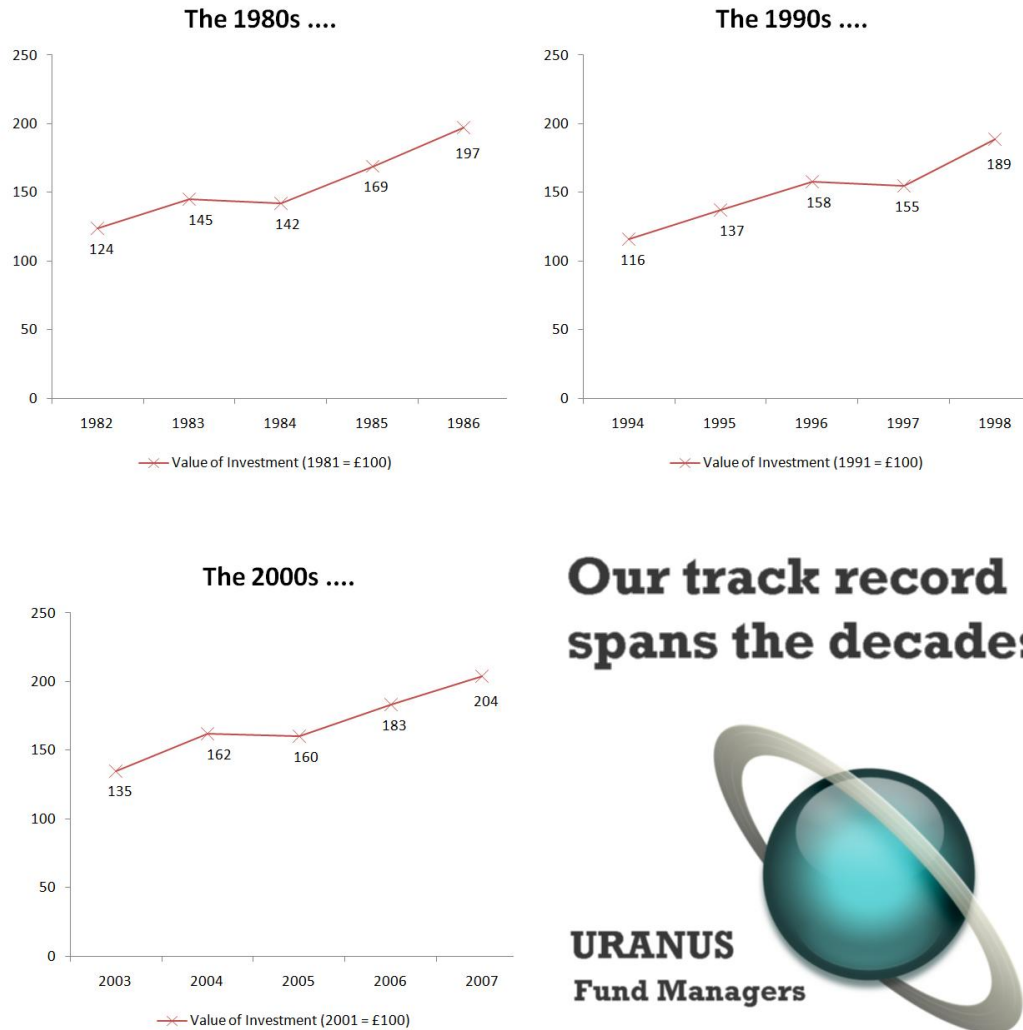


Luciano's story looks to be a successful one, whereas Carlo's looks very disappointing !  
Closer inspection of the time axis however shows that the comparison is unfair.

Luciano's graph covers spring and summer - natural times for ice cream sales to rise.  
Carlo's graph covers mainly autumn - natural times for ice cream sales to fall.

**Intentional omissions.**

**Example (29):** Uranus Fund Managers have published this advertisement in the national press.

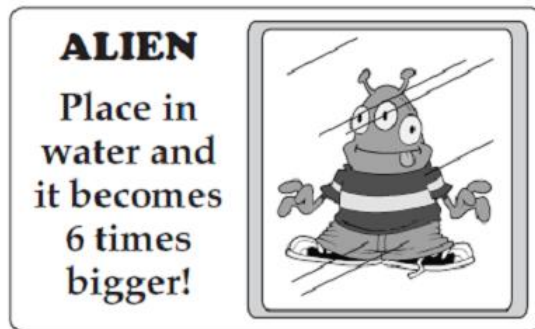


On the face of it, the fund managers seemed to have done very well over the last thirty years according to the graphs.

Closer inspection however reveals gaps in all of the timelines, such as 1987-1989 and 1999-2002 to give two examples. The fund managers might be hiding the fact that the value of their investments went *down* during those 'missing' years !

### Dimensional confusion.

#### Example (30):



The toy was 11 cm tall before placing in water, and a customer complained when the toy ended up being only 20 cm tall, yet the customer said it should have become 66 cm tall. .


Nevertheless, the trader was able to justify the claim on the packet. How ?

The ratio of the *heights* of the ‘aliens’ might have been in the ratio 11:20, but when this is converted into a ratio between *volumes* we have the ratio  $11^3:20^3$  or 1331:8000 – or almost exactly 1:6.

The alien toy does indeed become 6 times bigger – but in volume, not length !

### Using large numbers.

#### Example (31):

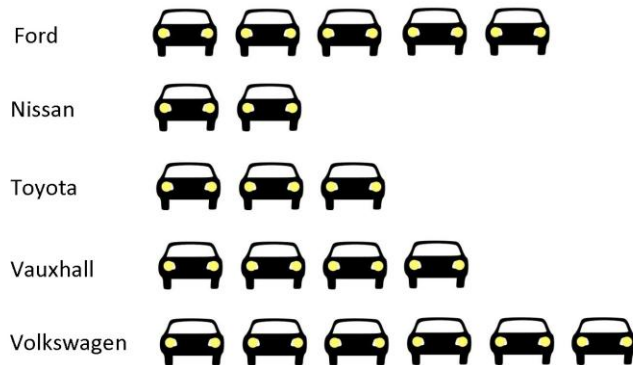
<p><b>The Daily Informer</b></p> <p><b>NHS staffing levels to be reduced by 3%</b></p> <p>The Government is announcing plans to reduce the workforce in the National Health Service by 3% from its current level of 1.3 million staff. In all, up to 40,000 jobs will be shed in the next four years.</p> 	<p><b>The Gleaner</b></p>  <p><b>40,000 HEALTH JOBS TO BE CUT</b></p> <p>The Government is planning to make huge cuts in the staffing levels of the NHS. It is expected that tens of thousands of jobs will be lost, out of a workforce of 1.3 million.</p>
---	--

Both newspaper headlines refer to the same proposal, but the eye’s attention is focused on the larger number, as large numbers have a more ‘sensational’ effect.

**Non-uniform pictograms.**

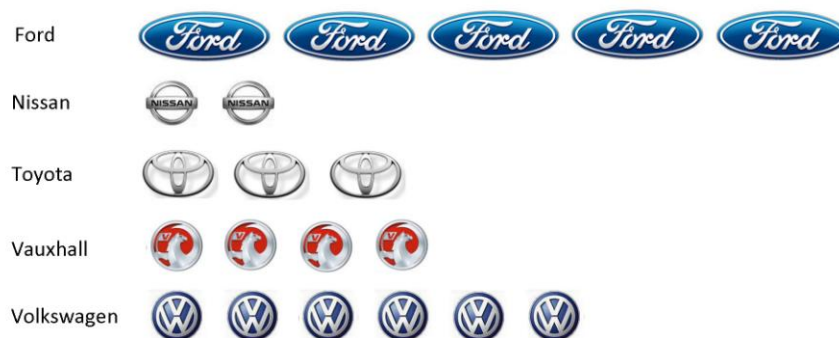
**Example (32):** A local car dealer had to send his regional manager a copy of a breakdown of new car sales for a certain week, and so he decided to represent the data on a pictogram.

**A. Boyce & Son**                      **New Car Sales w/e 28-02-2014**



He then thought the pictogram was a little too boring, so he decided to replace the car pictures with actual logos.

**A. Boyce & Son**                      **New Car Sales w/e 28-02-2014**



(All logos are copyrighted to their owners).

Considering that the dealership was owned by Ford, did the salesman have a different motive for changing the style of the pictogram, rather than “because it was a little too boring” ?

According to the first pictogram, it is clear that more Volkswagens were sold than Fords. On the second pictogram, a first glance suggests the exact opposite, because of the difference between the widths of the Volkswagen and Ford logos.

The dealer was trying to soft-soap his sales manager.