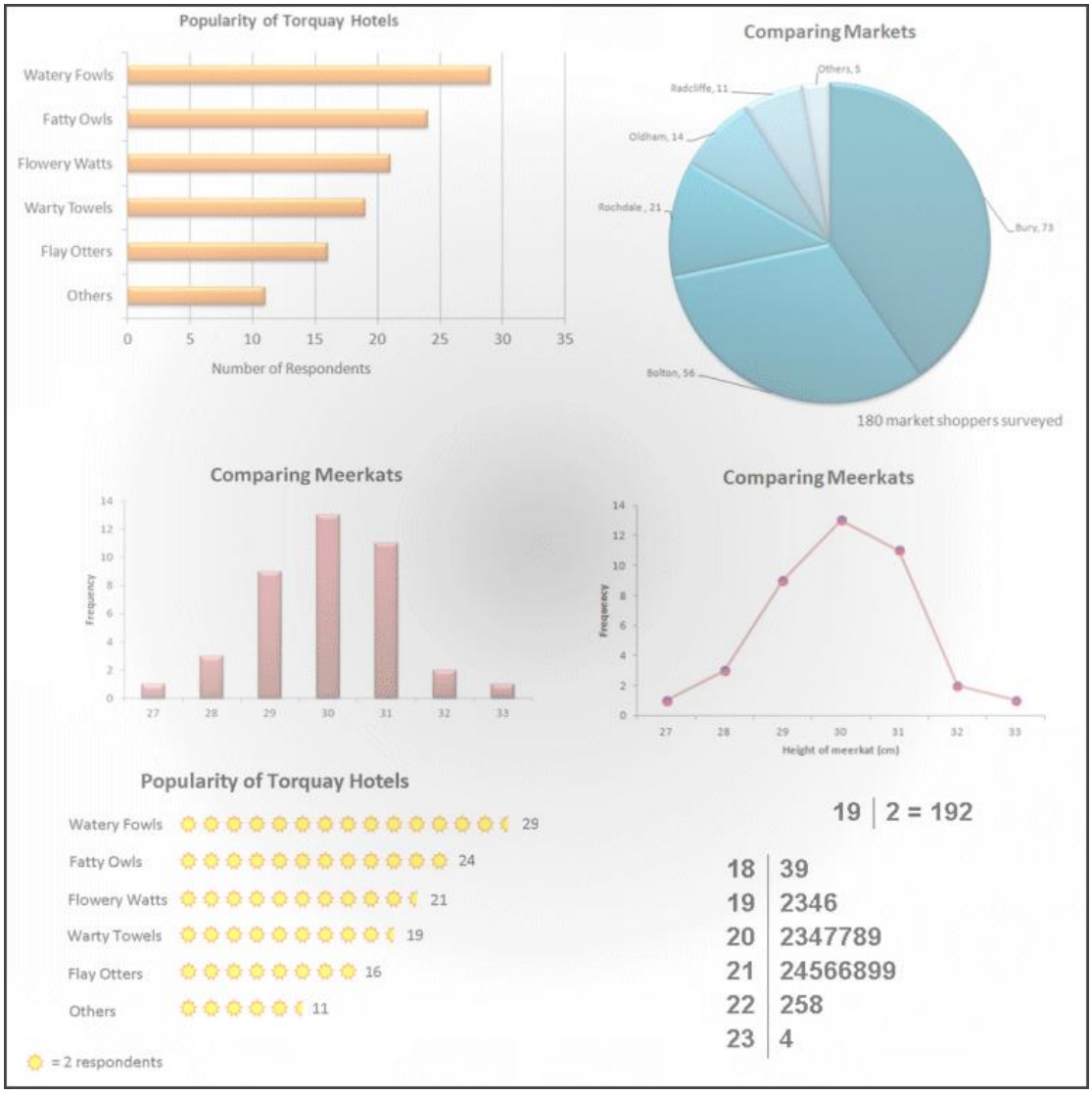
M.K. HOME TUITION

Mathematics Revision Guides
Level: GCSE Higher Tier

REPRESENTING DATA



Version: 4.3 Date: 11-11-2019

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 (O, A, B, AB). It can also be time-related, i.e. monthly sales figures, a temperature chart taken hourly.

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 discrete non-numeric data.

Examples would include a sample of eye colours, methods of transport to work, species of birds in a garden and similar other examples.

Discrete non-numeric data can be illustrated by a bar chart, a pie chart or a pictogram.

Two different sets of data can be represented by a double bar chart, two-way stem-and-leaf diagram, but a more versatile method is the scatter diagram, discussed in its own section.

If the data can be subdivided into categories in two different ways, we can display the results in a two-way table. An example would be a fleet of company cars, counted by engine size and by manufacturer.

Representing discrete numeric data.

Examples would include shoe sizes sold in a week, daily maximum temperatures over a month, and numeric grades scored in an examination.

Discrete numeric data can be illustrated by a bar chart, a line chart (frequency polygon), or a stem-and-leaf diagram. The pie chart and pictogram are less suited to this kind of data.

Two different sets of data can be represented by a two-way stem-and-leaf diagram, but a more versatile method is the scatter diagram, discussed in its own section.

Other important representations are the histogram and cumulative frequency diagram, again discussed in their own sections.

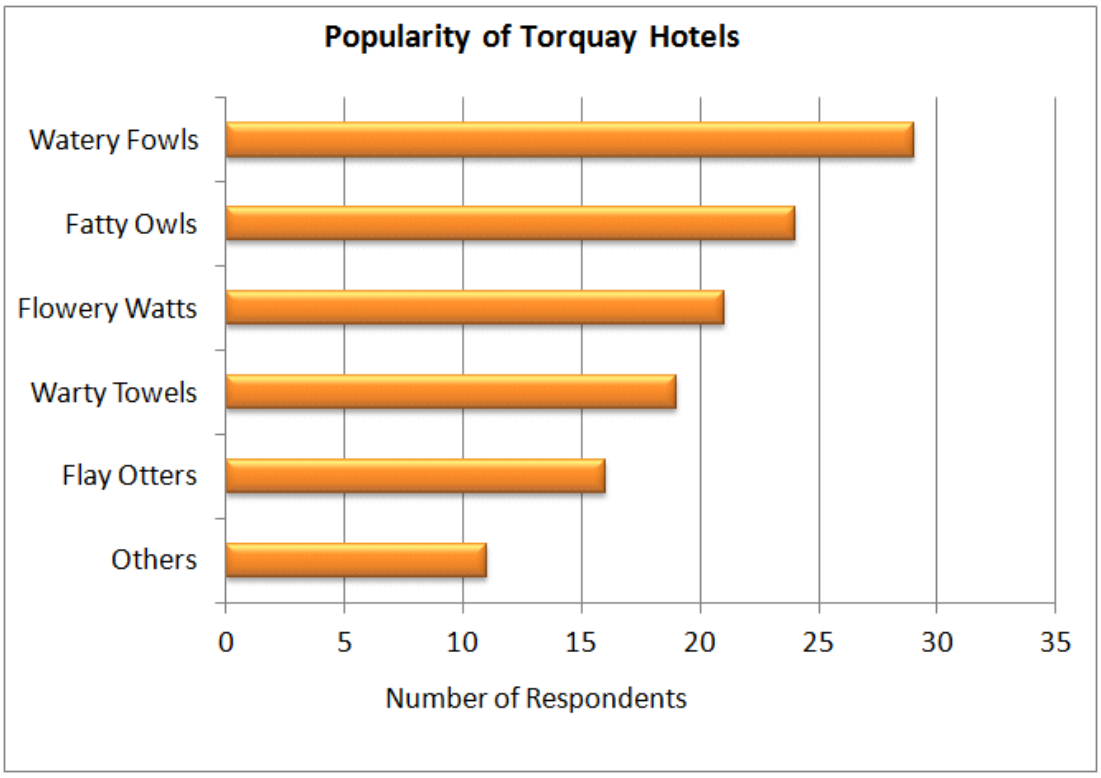
Finally we have the time series to investigate trends in data, as well as moving averages. This also includes year-on-year comparison charts.

Example (1): 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.

Bar Chart:

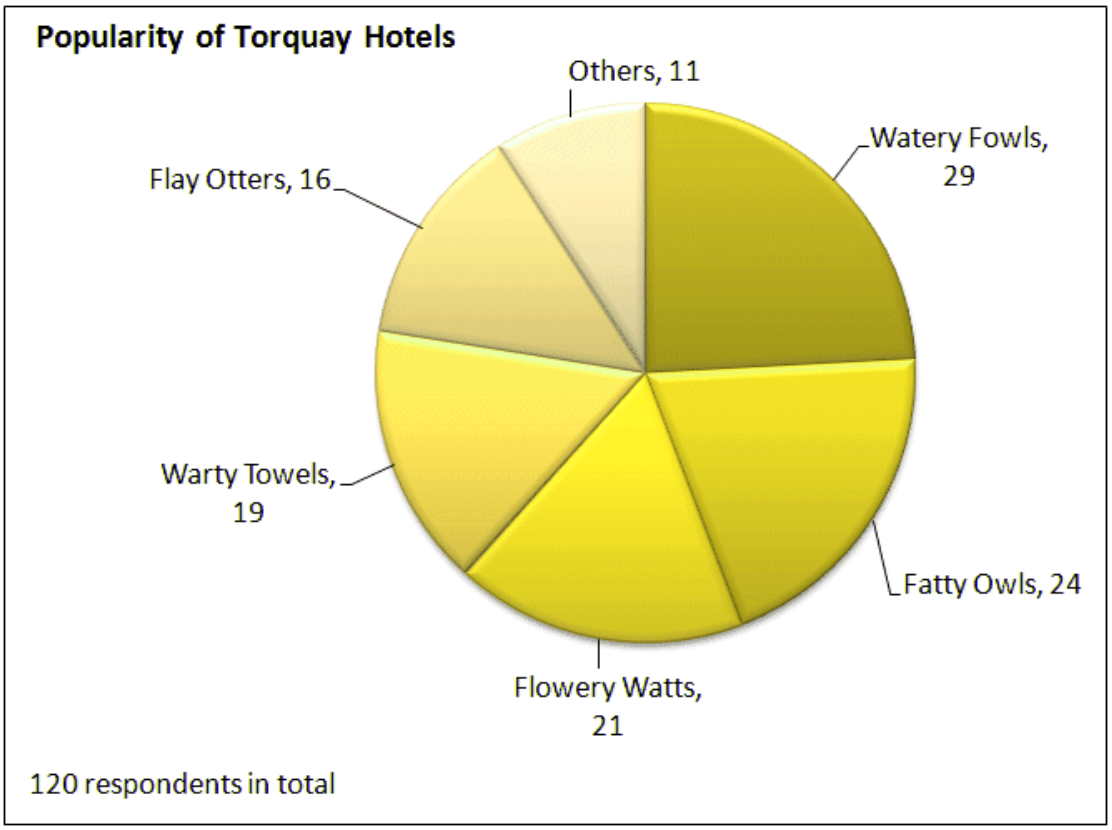


The hotel names here are the **group labels**: the numbers of respondents are the **frequencies.**

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 $\frac{360}{120}°$ slice of the pie, or 3°.

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

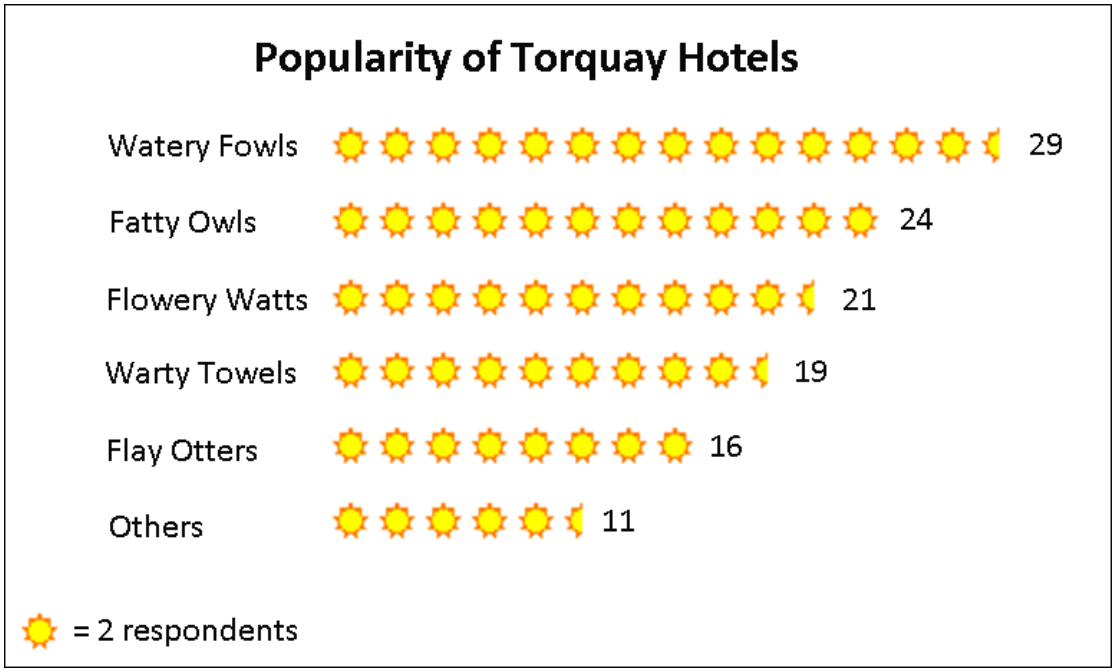
Pie Chart:



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 (2): 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'.

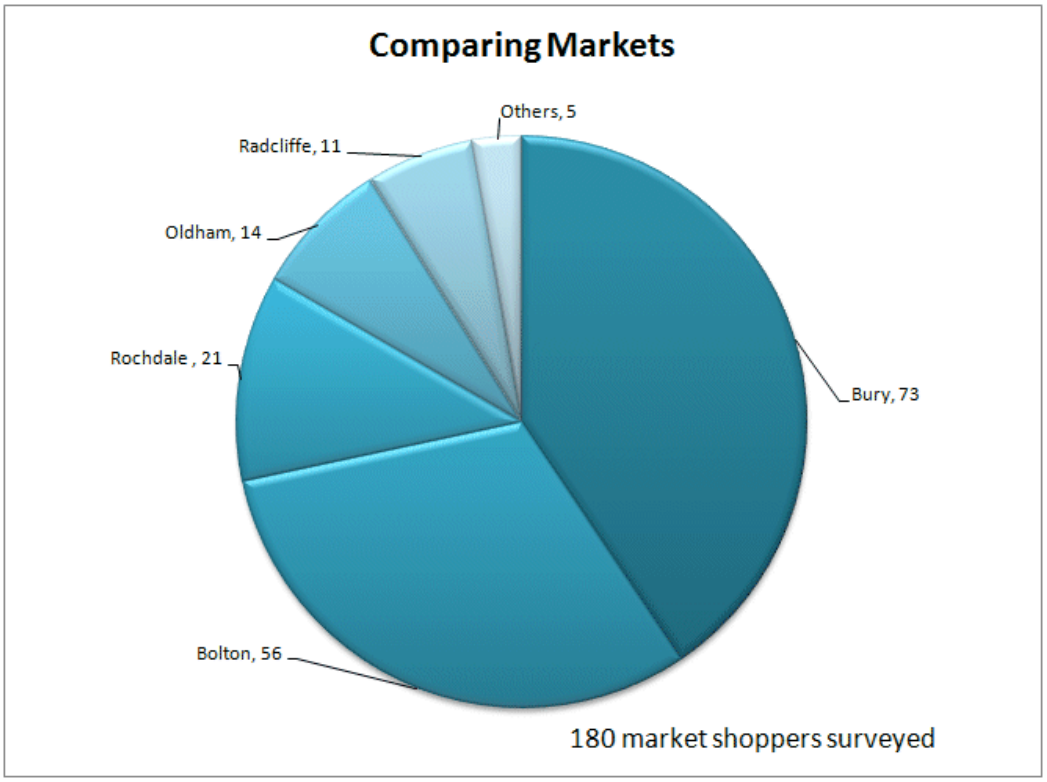
Produce a pie chart using the same data, including the 'others' option. Include labels and counts, as well as the total, on the chart.

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°
Bolton	56	112°
Rochdale	21	42°
Oldham	14	28°
Radcliffe	11	22°
Others	5	10°
TOTAL	**180**	**360°**

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

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



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.

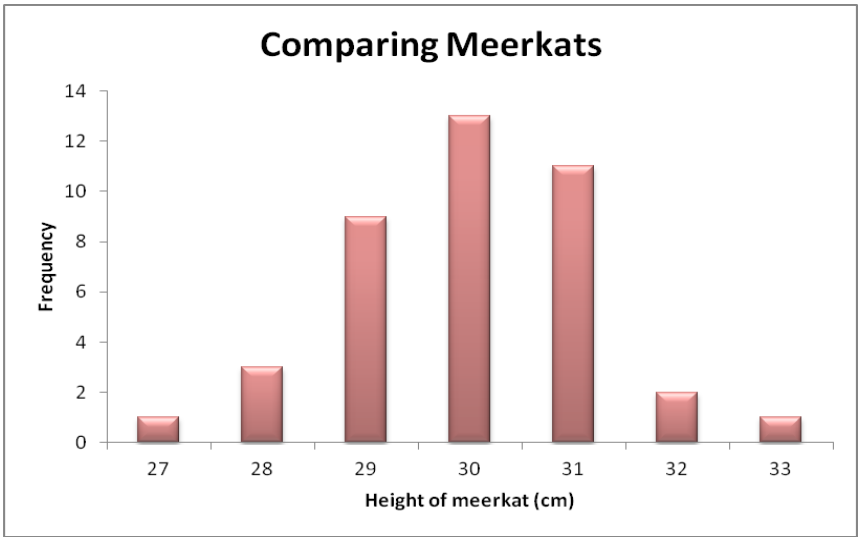
Example (3): The heights of 40 adult meerkats (to the nearest cm) on a nature reserve were recorded as follows:

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

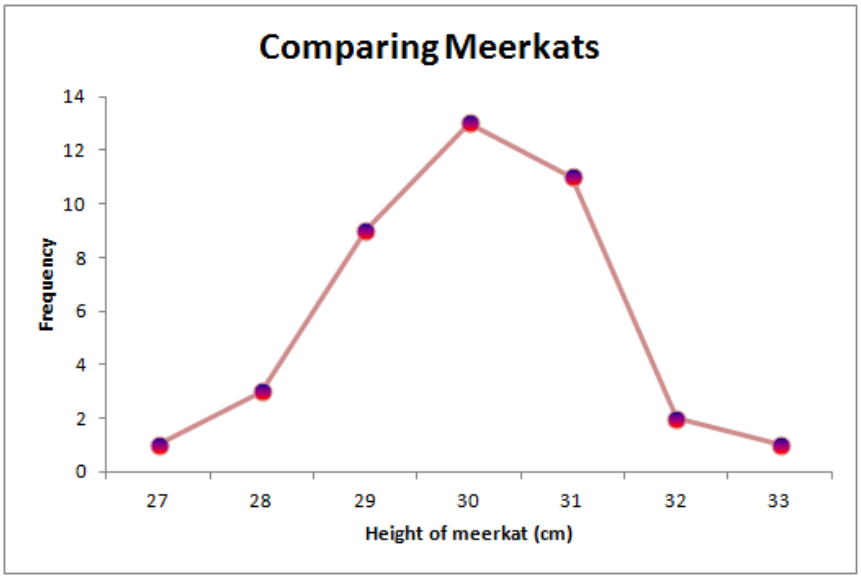
This time, the heights of the meerkats are the group labels.

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"). The frequency polygon can help visualise the distribution of the data, with most of the meerkats' heights clustered around 30 cm.

Bar chart:



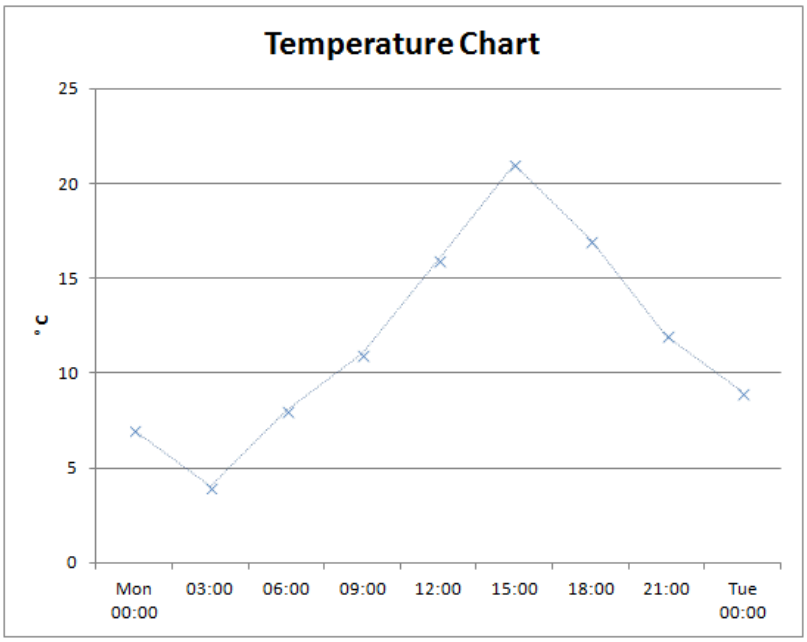
Frequency polygon:



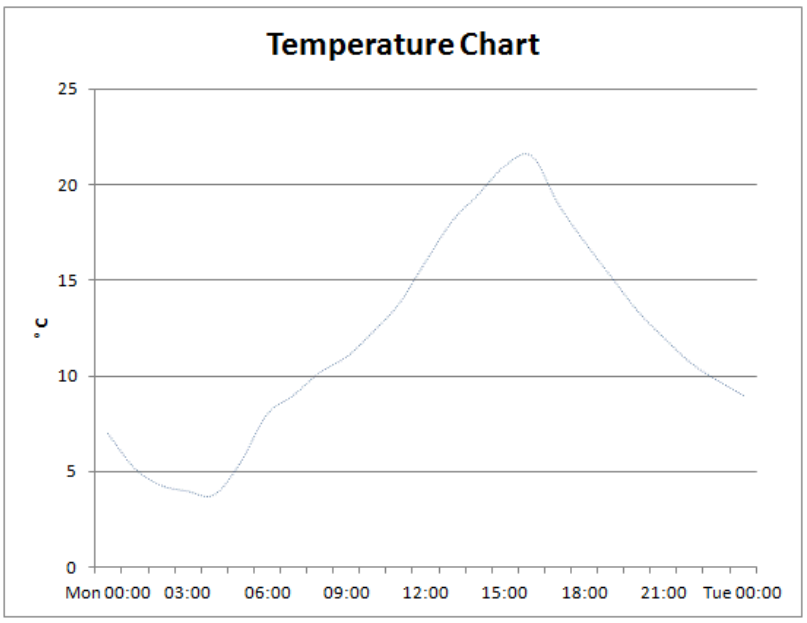
Line charts are also useful when plotting data against time, as per the next example.

Example (4): Here are some temperature readings taken at discrete 3-hour intervals over 24 hours. Plot the results on a frequency polygon.

Time	Mon 00:00	03:00	06:00	09:00	12:00	15:00	18:00	21:00	Tue 00:00
Temp (°C)	7	4	8	11	16	21	17	12	9



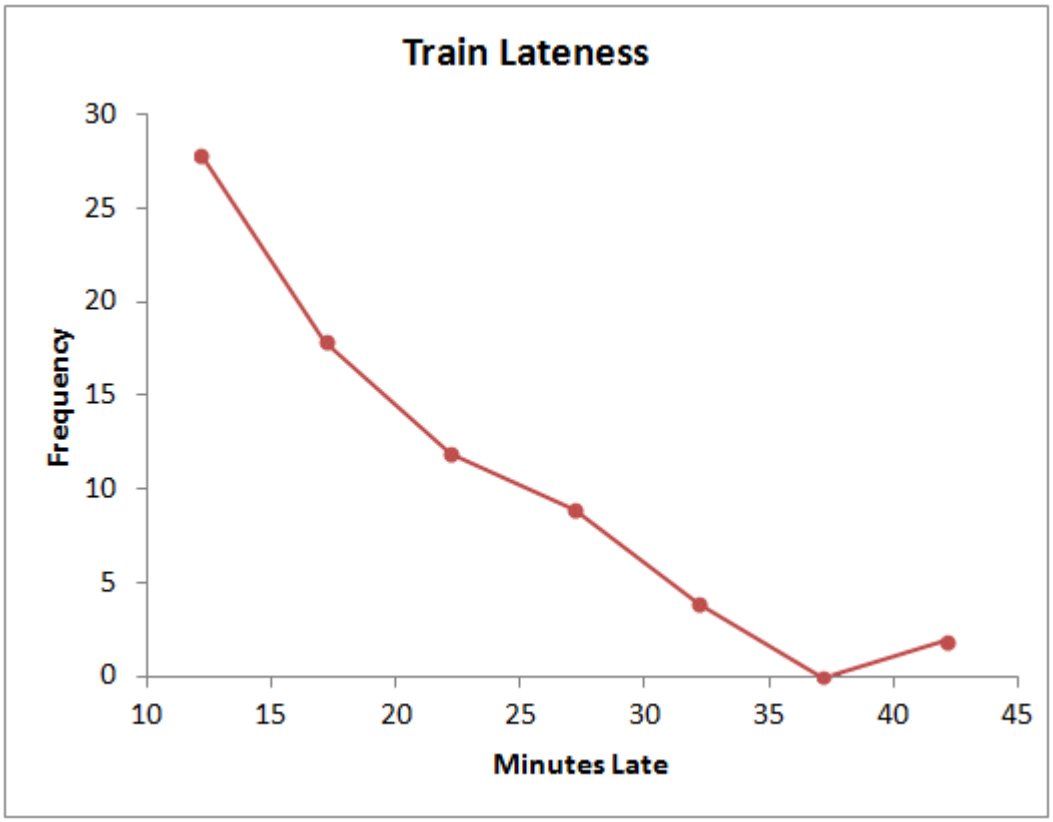
The temperature readings were selected at 3-hour intervals, and so the readings had been joined by straight lines. A weather station however monitors the temperature continuously, resulting in a smooth curve, as shown below.



Example (5): 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.

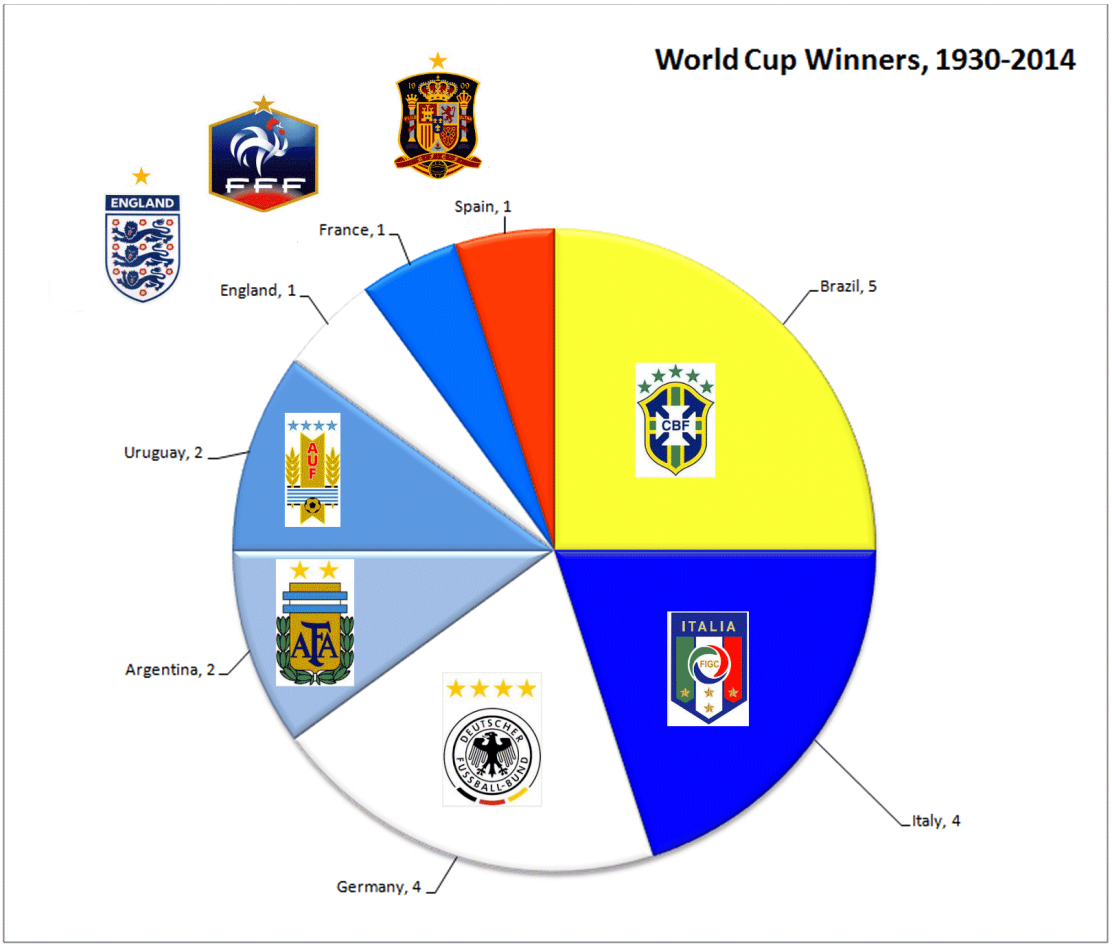
Example (6): The table below shows the winners of the football World Cup from 1930 to 2014, along with the number of times each winning team had lifted the trophy.

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 $\frac{360}{20}$°, or an 18° slice of the pie.
For example Brazil's slice, with 5 World Cup victories, takes up $5 \times 18°$, or $90°$.

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



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

Comparative Data.

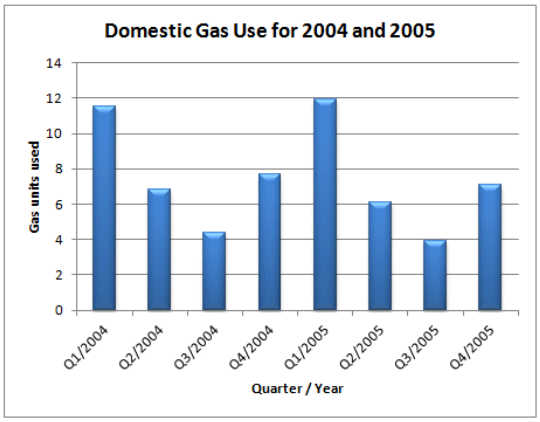
Comparative data is commonly used where the horizontal data label is *time*, i.e. for year-on-year comparisons, as in the next example.

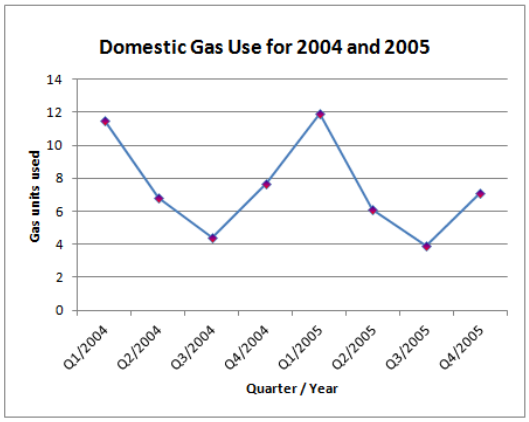
Example (7): A household is monitoring its quarterly gas consumption over two years.

Year	2004				2005			
Qtr.	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Units	11.5	6.8	4.4	7.7	11.9	6.1	3.9	7.1

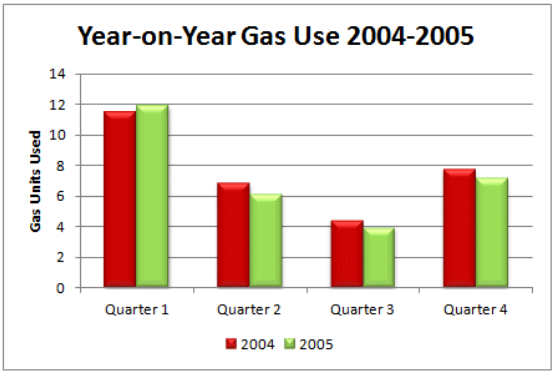
Here, the quarter of the year is the horizontal data label, and the units of gas used are the frequency (even though it is no longer a whole number). This also happens to be a time-related series. Again, the most suitable forms of representation are the bar and line charts.

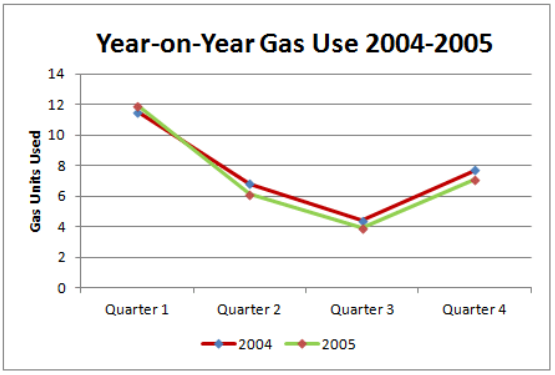
We can display the gas consumption as a straight time series as below:





It might however be more useful to display year-on-year charts, as they can reveal trends:





The chart on the left is an example of a clustered bar chart, with additional examples shown later.

Based on the 2004 consumption of gas, the household was able to reduce gas usage in the second, third and fourth quarters in 2005, after a slight increase in the first quarter.

Care must be taken with comparative pie charts, as the next example will show.

Example (8): There are two further education colleges in a town.
The Community College has 400 students, of whom 120 study Arts subjects, 150 Humanities subjects and 130 study Science subjects.
The Technical College has 900 students, of whom 225 study Arts, 150 study Humanities and 525 study Sciences.

Display the results on two comparative pie charts.

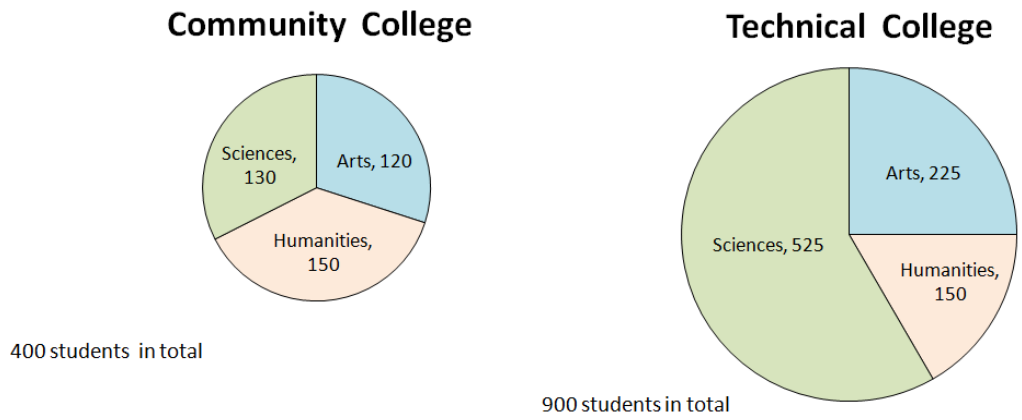
There are 400 students at the Community College, so 120 students (for instance) are represented by a $\frac{120}{400}\times 360°$ or a 108° slice of the pie, i.e. 1 student = $\frac{1}{400}\times 360°$ or 0.9°.

Subject	No. of students	Pie chart angle
Arts	120	108°
Humanities	150	135°
Sciences	130	117°
TOTAL	**400**	**360°**

There are 900 students at the Technical College, so 150 students (for instance) are represented by a $\frac{150}{900}\times 360°$ or a 60° slice of the pie, i.e. 1 student = $\frac{1}{900}\times 360°$ or 0.4°.

Subject	No. of students	Pie chart angle
Arts	225	90°
Humanities	150	60°
Sciences	525	210°
TOTAL	**900**	**360°**

The comparative pie charts are shown below:



Notice how the sizes of the pie charts are related here. The numbers of students at the college are in the ratio 400:900 or 4:9, and therefore the *areas* of the pie charts are related in the same way. Thus, the two slices for Humanities have equal areas, as they both represent 150 students, even though their angles are different.

We therefore need to take square roots to find the ratio between the *radii* of the charts: here the ratio is $\sqrt{4}:\sqrt{9}$, or 2:3. The radius of the pie for the Technical College is thus 1½ times that for the Community College.

Clustered and Stacked Bar Charts.

Example (9)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, £

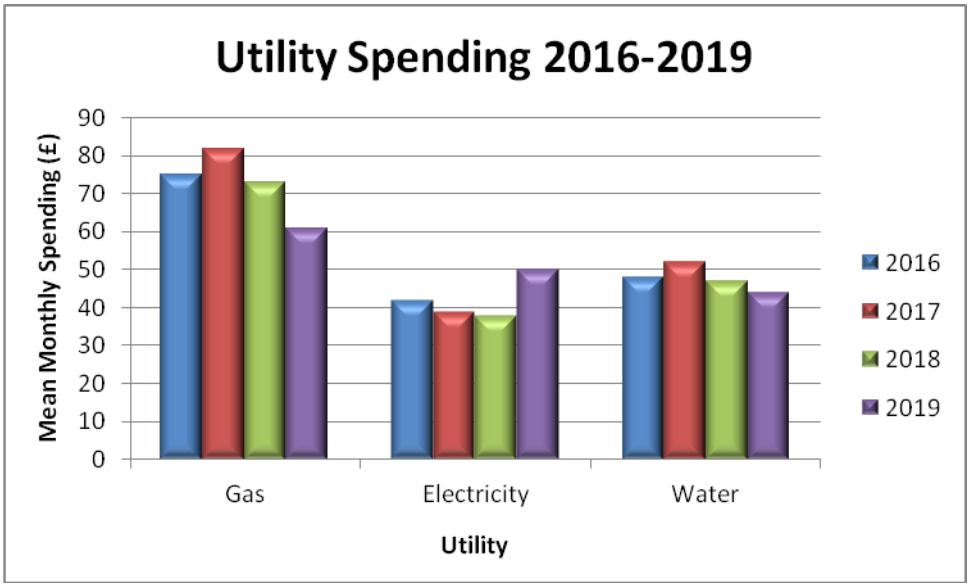
	2016	2017	2018	2019
Gas	75	82	73	61
Electricity	42	39	38	50
Water	48	52	47	44

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

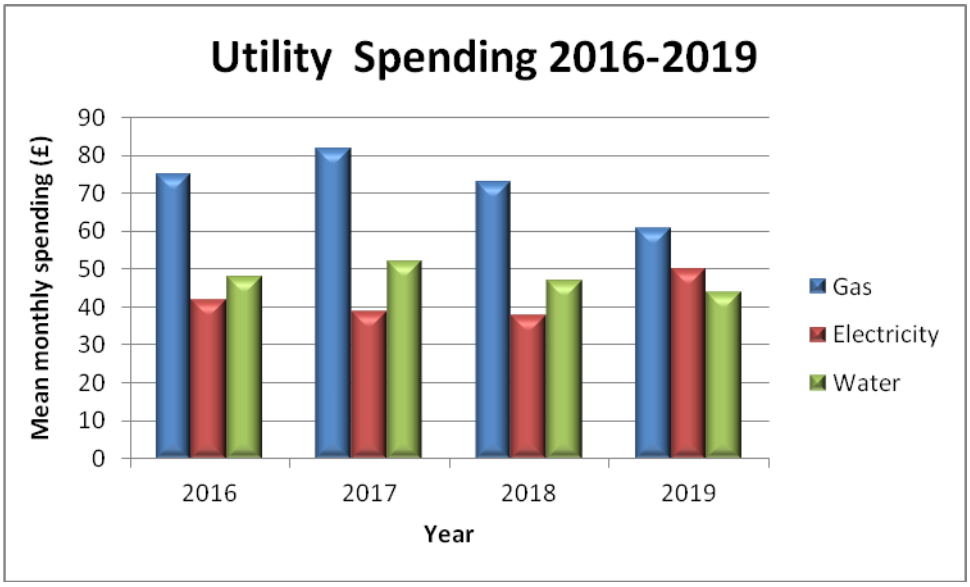
Thus the first cluster of four bars refers to his monthly spending on gas, from one year to the next. The monthly spending figures for electricity and water are grouped in a similar fashion.

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.



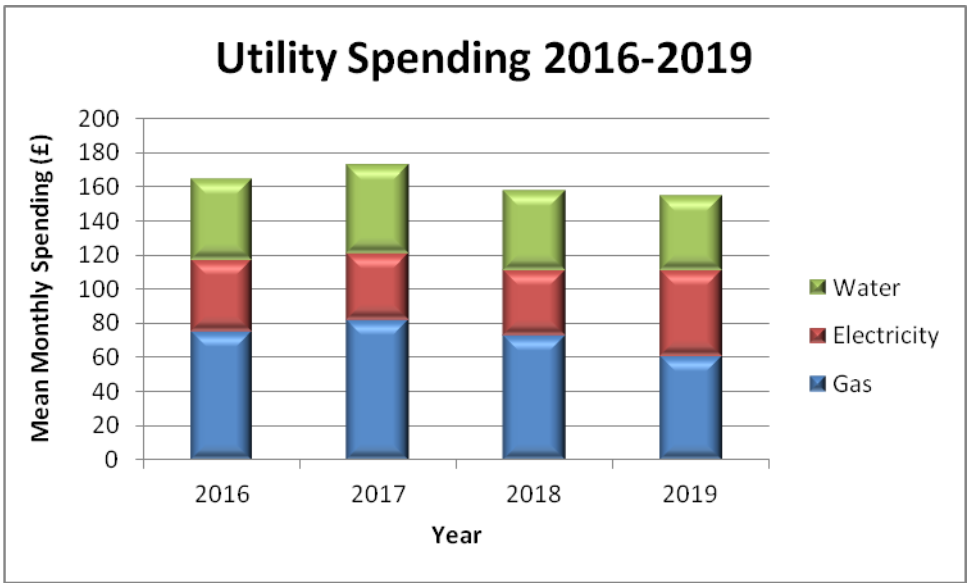
Example (9)b: Dave then wants to compare his *total* monthly spending on utilities, so he produces the bar chart below, where the comparison is between the *utilities* (the horizontal data label), with each *year* represented by a grouped bar. In other words, he groups the bars in four sets of three, 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 9(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.



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 (10): 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	**Total**
Inter-city		3	**42**
Suburban	63		
Total			**114**

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 and 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	**Total**
Inter-city	39	3	**42**
Suburban	63	9	**72**
Total	**102**	**12**	**114**

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

Two-way tables are concise, but they lack the visual impact of most of the other representations.

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 (11): 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

207, 196, 209, 193, 214, 218, 234, 212, 216, 228, 189, 215, 192, 225, 203,219, 194, 222, 183, 216, 204, 207, 208, 219, 202

Draw a stem and leaf diagram to illustrate the data.

First, we see that the smallest value is 183 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'.

18
19
20
21
22
23

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

(We could have used sorted data as in Example 3, 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.
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.

18		18		18		18		18		18	
19		19		19	6	19	6	19	6 3	19	6 3
20		20	7	20	7	20	7 9	20	7 9	20	7 9
21		21		21		21		21		21	4
22		22		22		22		22		22	
23		23		23		23		23		23	

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.

18	93
19	6324
20	7934782
21	48265969
22	852
23	4

19 | 2 = 192

18	39
19	2346
20	2347789
21	24566899
22	258
23	4

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.

Misrepresenting Data.

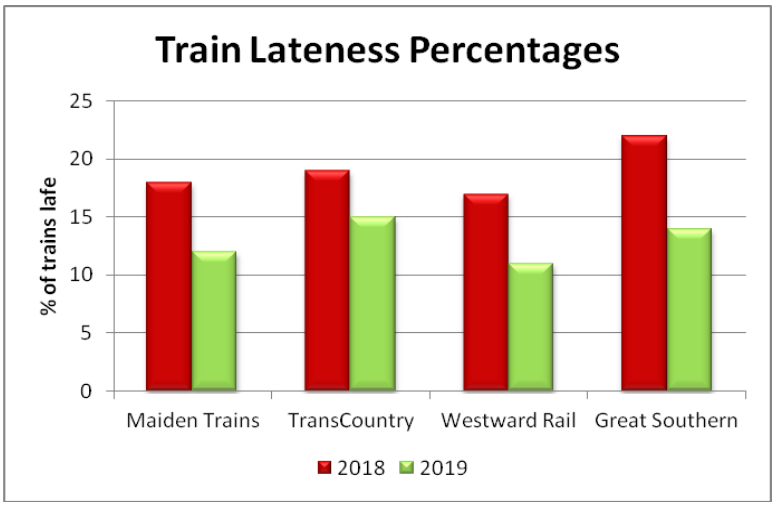
"There are lies, damned lies, and statistics". (Attrib: Charles Wentworth Dilke.)

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.

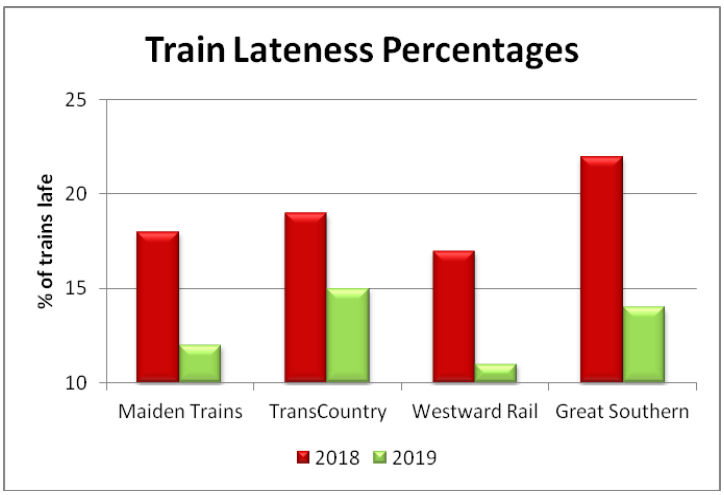
Example(12): The bar charts below show the percentages of late-running trains for four railway operators over two years.

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 !



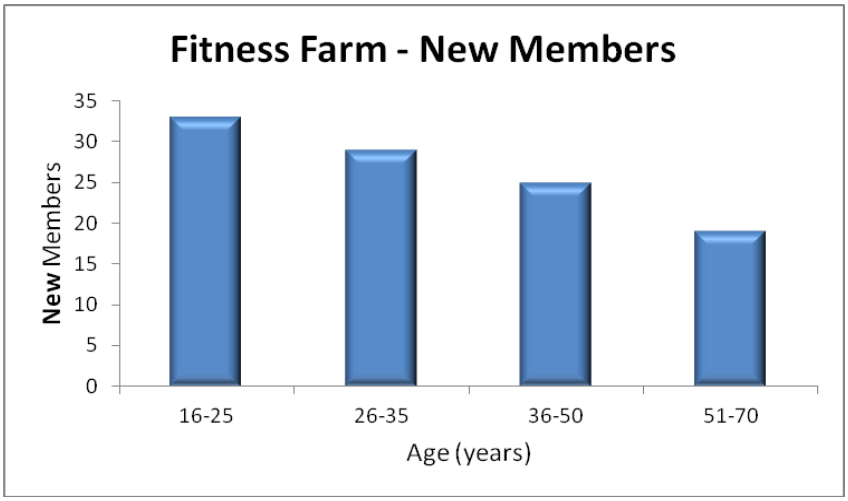
Misleading grouping of data.

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

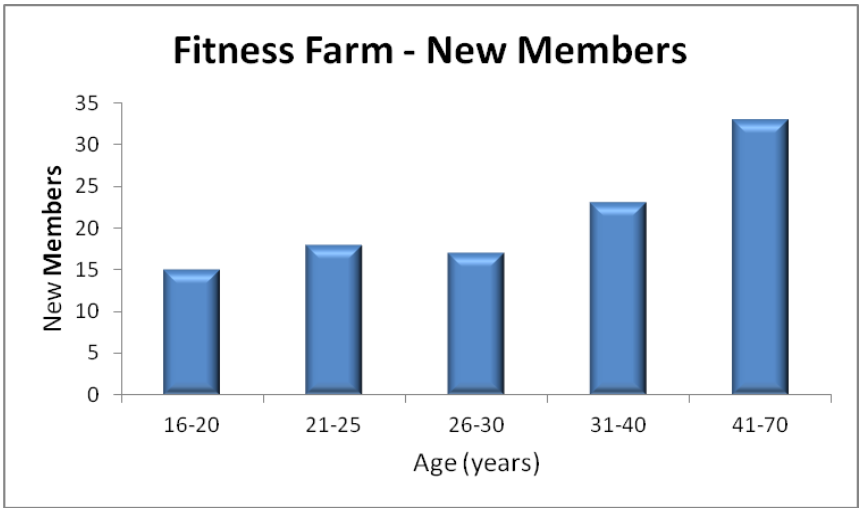
Example (13): Fitness Farm gym have published their new membership statistics for the latest year, by members' ages :

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 whch 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.

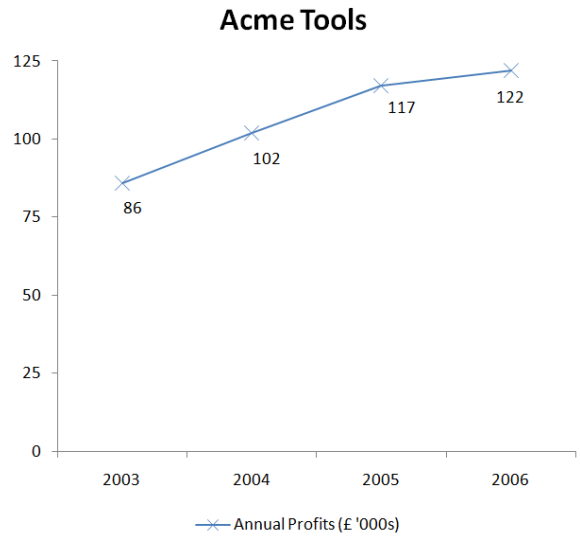
This is because the class intervals (here age ranges) vary greatly in width in each case, but the bar charts tend to hide the fact.

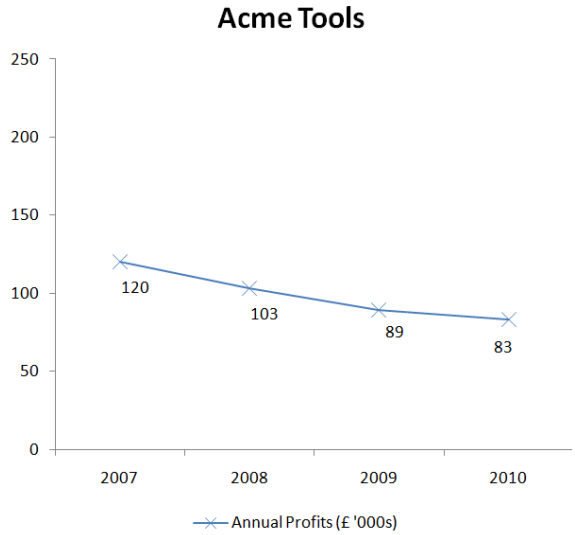
To illustrate the true pattern, it is better to use a histogram or a cumulative frequency diagram.

Differing vertical scales.

Example (14): 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.



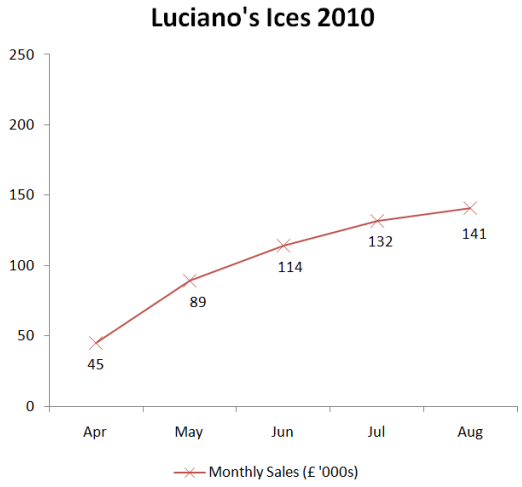


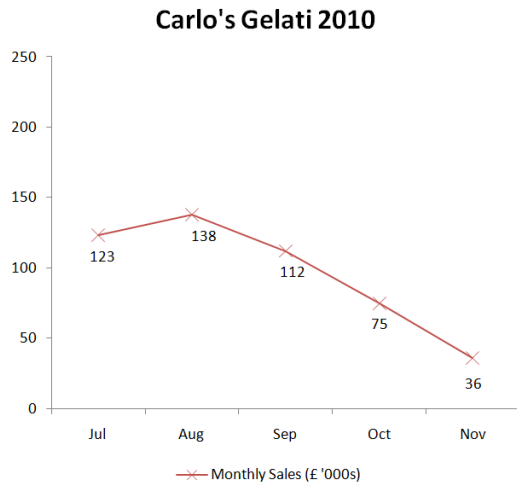
This time, the vertical scales begin at zero for both graphs, but they are still misleading because they use differing scales. While Acme Tools was showing improving profits, the graph exaggerated the fact by using a scale of £25,000 per vertical division. When the profits decreased from 2007 on, the 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 (15): The next two graphs show the sales figures for two local ice-cream makers in 2010.



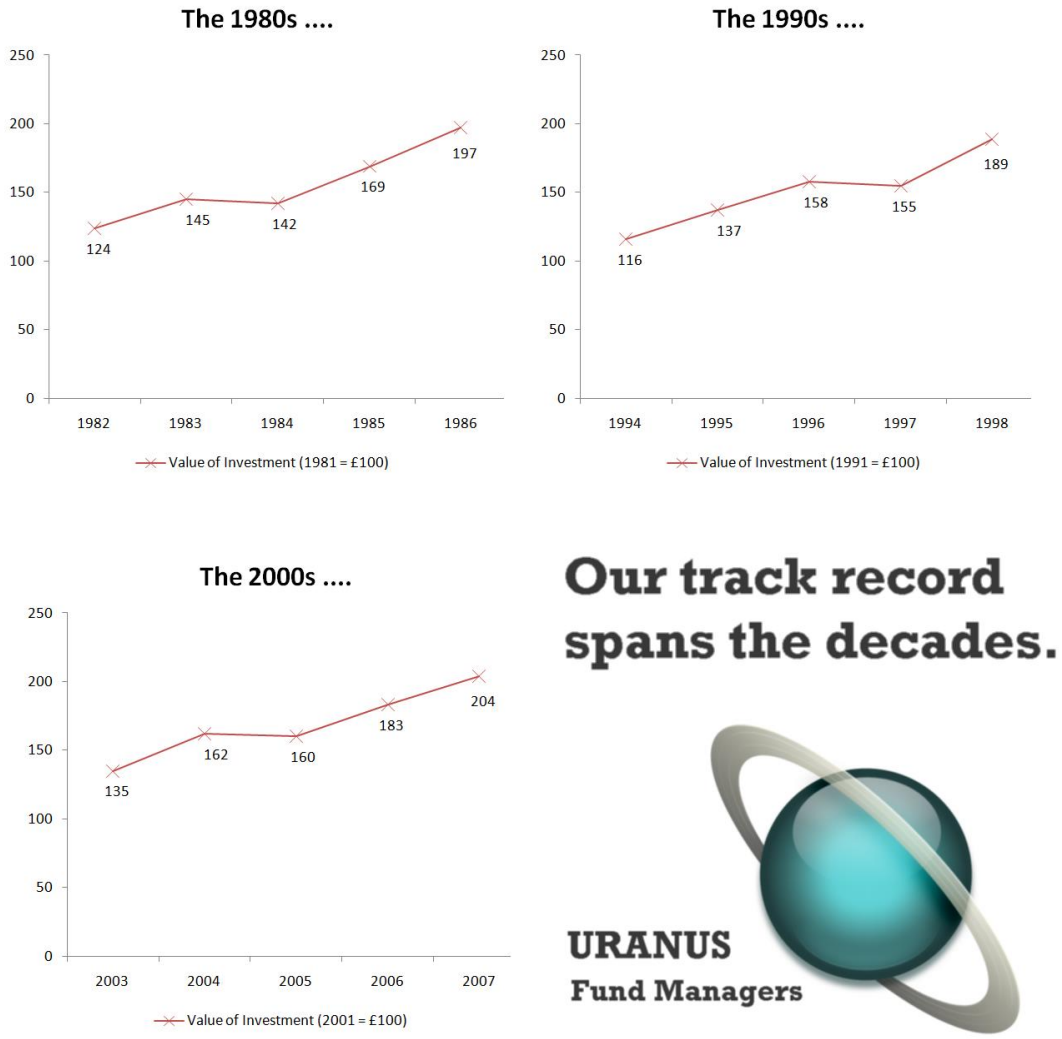


Luciano's story looks to be a successful one, whereas Carlo's looks very disappointing !
Closer inspection of the time axis however shows that the comparison is unfair.

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 (16): Uranus Fund Managers have published this advertisement in the national press.

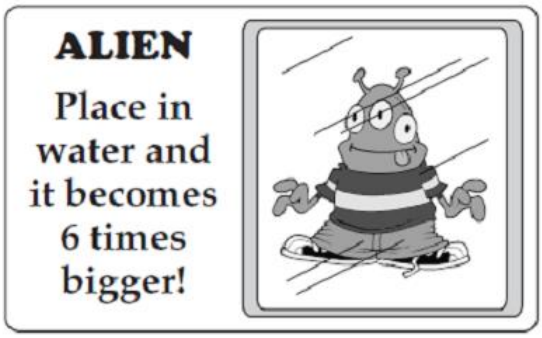


On the face of it, the fund managers seemed to have done very well over the last thirty years according to the graphs.

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 (17):



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 (18):


The Daily Informer

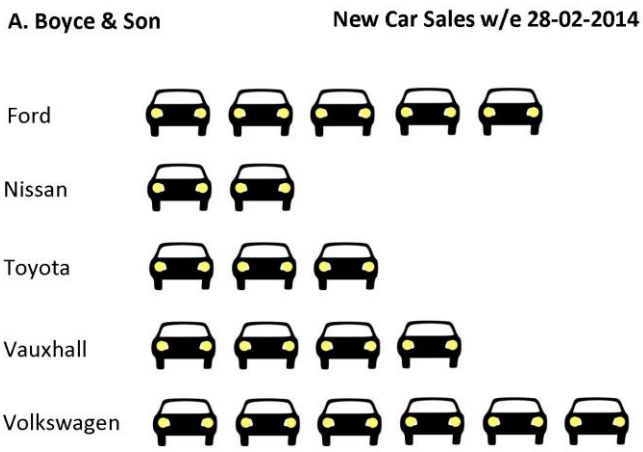
NHS staffing levels to be reduced by 3%

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.

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 (19): 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.



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

(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.