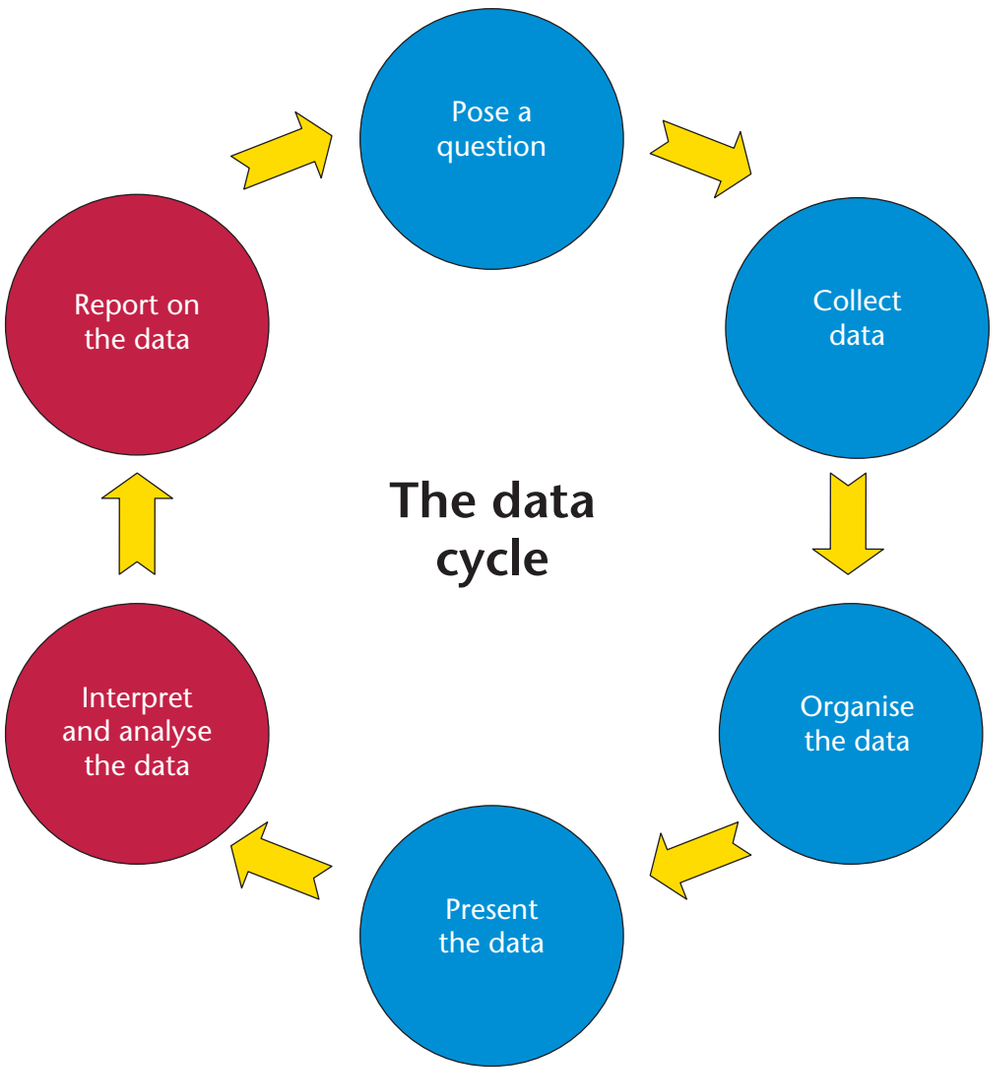
Chapter 10
Interpret, analyse and report on data

In this chapter, you will develop and practise some critical data analysis skills. This means looking at reported data and analysing the whole data handling cycle for this data. You need to decide which way of representing data is best in a given situation. In summarising data, some measures are more appropriate for different types of data. You also need to recognise some ways in which bias can appear in data, including methods of collecting, representing and summarising data.



10 Interpret, analyse and report on data

10.1 Which graph is best?

You have learnt that certain types of graphs are best for displaying certain kinds of information. The type of graph depends mostly on the type of data that needs to be represented. Here is a summary of the advantages of different types of graphs:

Tables show more information than graphs but the patterns are not as easy to see. They do not give a visual impression of particular trends.

Pie charts show a whole divided into parts. They show how the parts relate to each other and how the parts relate to a whole. They do not show the quantities involved.

Bar graphs show the amounts or quantities involved but do not show the relationship as effectively as pie charts. They are useful for showing **quantitative** data. Bar charts allow us to compare the quantities of different categories, for example, the sales of different items.

A **double-bar graph** is used to compare two or more things for each category. For example, we could use a double-bar graph to compare the differences between males and females.

Histograms are used to represent numerical data that is grouped into equal class intervals. Histograms are useful to show the way the data is spread out.

Broken-line graphs show trends or changes in quantities over time.

CHOOSE THE BEST REPRESENTATION

1. Which kind of graph is best to represent each of the following? Explain your answers.
   (a) Showing the value of the rand against the US dollar over several years

   .....

   (b) Comparing the monthly sales of six different makes of car in 2014 and 2015

   .....

   (c) The proportion of people of different age groups in a town

   .....

   .....

   (d) The quantities of different crops produced on a farm

   .....

   .....

   (e) The percentages of different goods sold to make up the total sales for a shop

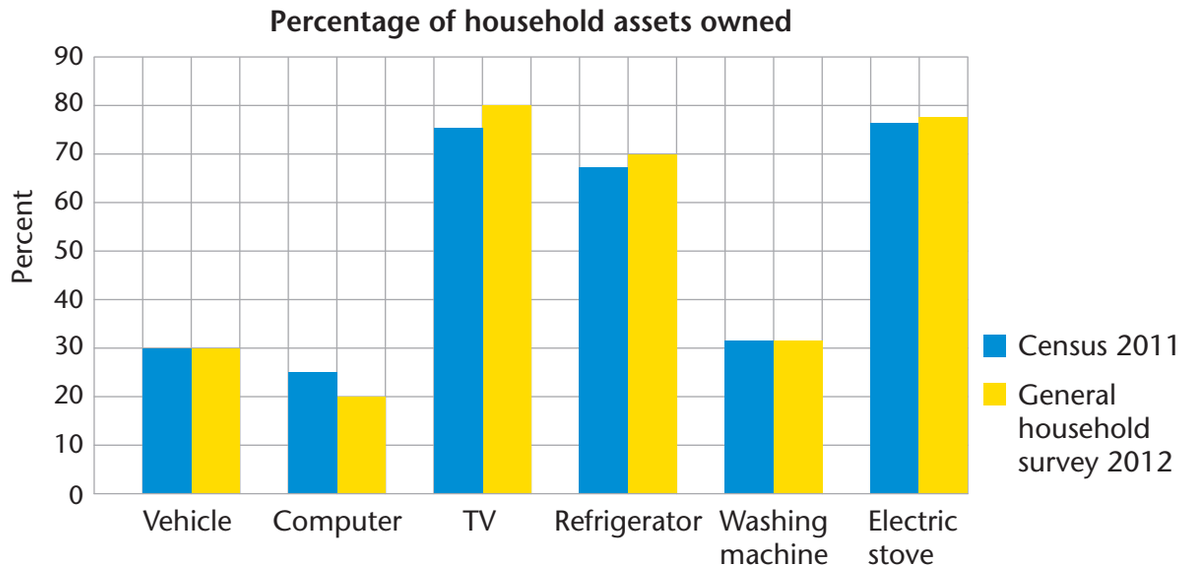
   .....

   .....

   (f) The change in HIV infection rates over time

   .....

2. This graph was published by Statistics South Africa to show the assets owned by South Africans. The blue bar shows the Census 2011 results and the yellow bar shows the General Household Survey 2012 results.



Give reasons for your answers to the questions below.

(a) Is it useful to show the differences in the results of Census 2011 and the General Household Survey 2012?

.....

.....

(b) Is it useful to collect data on assets that people own?

.....

.....

(c) Is it useful to show that lower percentages of people own certain assets?

.....

.....

.....

(d) The different coloured bars represent the two different surveys. Draw up a table to show the data in table form. (Read the percentages as accurately as you can from the graph and round off the data to the nearest whole number for the table.)

(e) Does the table show the data as effectively as the double bar chart? Give your own opinion.

.....

.....

3. The table below shows the employment status of people ages 15–64 years in South Africa. Discuss some ways of representing the data (e.g. graphs). Justify your answers.

	Jul–Sept 2012	Apr–June 2013	Jul–Sep 2013
	Number of people (thousands)		
Population 15–64 years old	33 017	33 352	33 464
Labour force	18 313	18 444	18 638
Employed	**13 645**	**13 720**	**14 028**
Formal sector (non-agricultural)	9 663	9 694	10 008
Informal sector (non-agricultural)	2 197	2 221	2 182
Agriculture	661	712	706
Private households	1 124	1 093	1 132
Unemployed	**4 668**	**4 723**	**4 609**
Not economically active	**14 705**	**14 908**	**14 826**
Discouraged work-seekers	2 170	2 365	2 240
Other (not economically active)	12 535	12 543	12 586
Unemployment rate (%)	25,5	25,6	24,7

(a) The percentages of the employed, unemployed, and not economically active people in July–September 2013.

.....

(b) The change in the employment **rates** over three time periods

.....

(c) The proportions of employed people who work in the formal sector, informal sector, agriculture and private households.

.....

(d) The numbers of the employed and unemployed over the three time periods.

.....

.....

10.2 The effects of summary statistics on how data is reported

Information articles often use averages to report information. The articles might not use the exact terms for average that you have learnt about: the mean, median and mode. Instead, they may use terms such as 'most'. However, it is important to be sure which kind of average a report refers to, because they give us different information.

- Remember that the **mean** is useful for describing a set of measurement values, but can also be used for other numerical data sets. The word 'average' usually refers to the 'mean' if it is not explained further. The mean is not reliable if a data set is too spread out.
- The **median** is the value in the middle of a data set when it is arranged in order. Half the values in the data set are lower than the median and half of them are higher than the median. The median is often the average used when data values are not uniformly distributed, because the mean is affected by extreme values in the data set, while the median is not. For example, house prices vary widely, so the median would be a better description of the data than the mean. When the median is given in a report, the writer should state that they are using the median or middle value.
- The **mode** is the number that occurs most often in a set of data. For example, if we collect data about people's favourite colours, the data set would be a list of colours, and the mode would be the colour that comes up most often. The mode can also be used for numbers. Not all data sets have a mode, because sometimes none of the numbers occurs more than once.

Example

The standard way of reporting house prices in South Africa and internationally is the median house price, which is used by economists in financial reports. The median is regarded as more useful than the mean house price because the sale of a few expensive houses would increase the mean, but would not affect the median.

If a bank gives bonds for eight houses to the value of R100 000, and for two houses to the value of R1 million, then the mean would be R280 000. This does not seem to be an accurate reflection of the value of the houses, because it is distorted by the higher values. The median house price would be R100 000, which is an accurate reflection of the prices.

Remember that the median is the middle point, and half of the values fall below the median, and half above. If the median is lower than the mean, this shows us that there are high values that are distorting the mean.

USING DIFFERENT SUMMARY STATISTICS

1. What kind of average is used in each of these statements?
   (a) The average family has 2,6 children. .....
   (b) Most families have 3 children. .....
   (c) Most people prefer red cars. .....
   (d) The average height for women is 1,62 m. .....
   (e) More people shop after work than at any other time during the day. .....

2. The mean monthly salary of all the staff at company ABC is R8 000 per month, but the median salary is R5 000.
   (a) Explain why the two summary statistics are so different.

   .....

   .....

   .....
   (b) Which summary statistic gives a better idea of the salaries at the company? Give reasons for your answer.

   .....

   .....

10.3 Misleading graphs

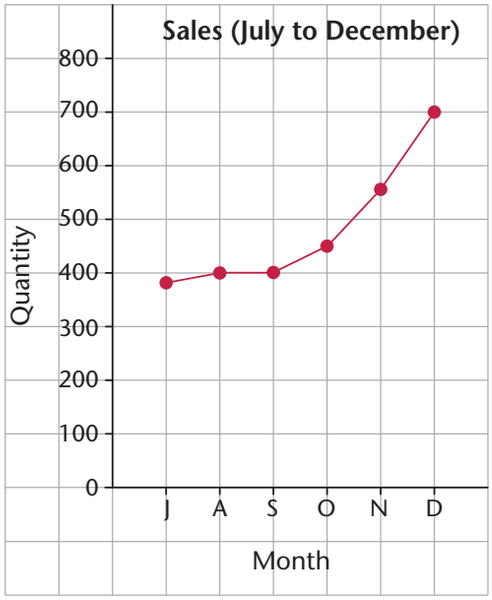
The media (newspapers, magazines, television), regularly use graphs to show information. Unfortunately, the information is often manipulated to emphasise a particular result. This may be because the writer simply wants to make his or her argument more obvious to the reader.

Changing the scale of the axis

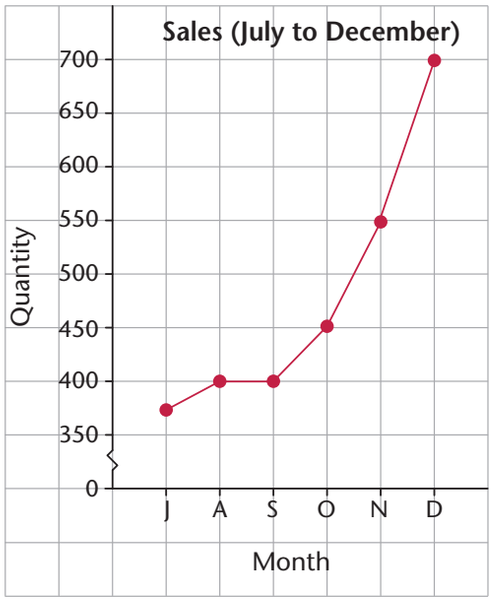
If you change the scale of the vertical axis on bar graphs and line graphs, you will change the way the graphs look. For a bar graph, the larger the spaces between the numbers on the vertical axis, the bigger the difference between the bars. The smaller the spaces between the numbers on the axis, the smaller the difference in the height of the bars. The same is true for a line graph which will either have sharp points or be much flatter depending on how you have changed the scale.

Example

The two broken-line graphs below show the same sales data for a business over a period of six months. Which graph gives the more accurate impression?



Graph A



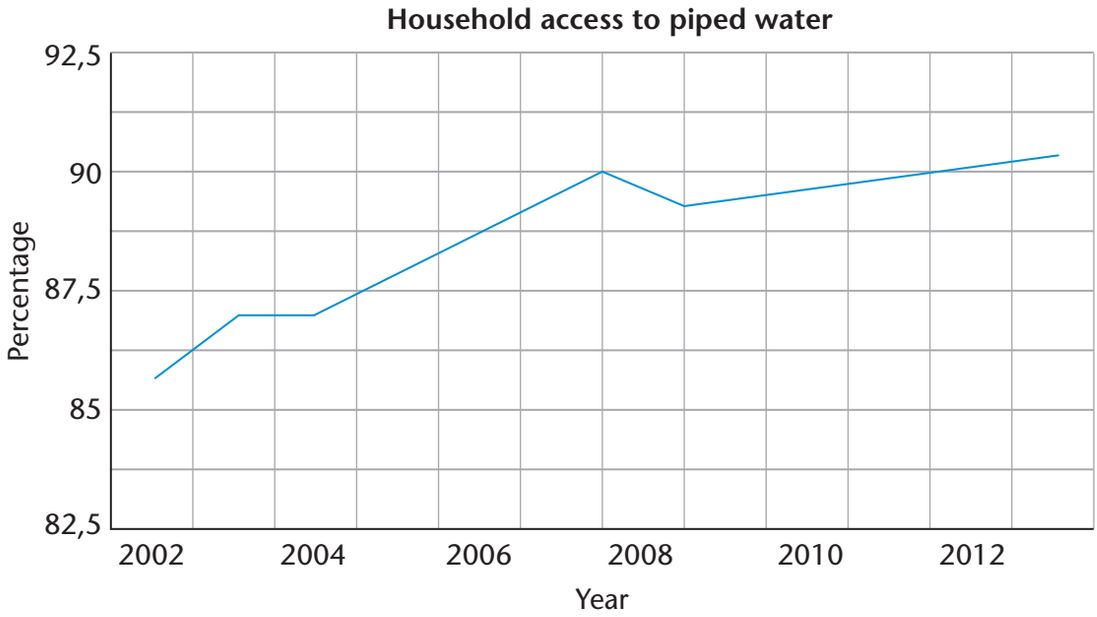
Graph B

Graph B has a different scale on the vertical axis. The vertical axis does not start at 0 and so **two** blocks on the vertical axis represent 100 items instead of only **one** block, as in Graph A. This makes it look as if the sales increased rapidly over the six months.

Note that it is not necessarily wrong to change the scale on the axes or not to start at 0. For example, graphs showing stock exchange fluctuations rarely show the origin on the graph and stockbrokers are taught to interpret the graphs in that form. Sometimes small changes in data values have important effects and in these cases, it may be valid to change the scale to show these.

ANALYSING GRAPHS

1. This graph from Statistics South Africa shows the increase in the percentage of households that had access to piped water over a ten-year period.



(a) Comment on the scale used on the vertical axis. Is this a misleading graph?

.....

.....

.....

.....

.....

.....

(b) How could you redraw the graph so that the differences on the graph are more noticeable?

.....

.....

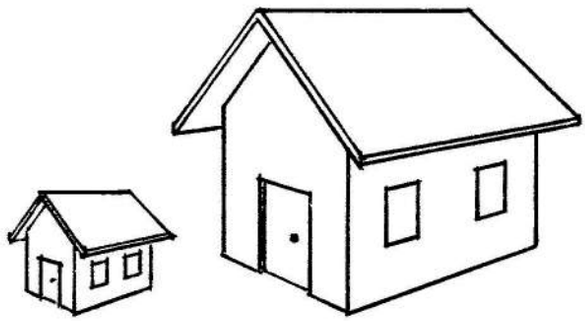
(c) How could you draw the graph so that the differences are less noticeable?

.....

.....

2. In this graph the height of the houses represents the number of sales.

Home Sales



2010 2014

Do you think that this graph is misleading? Give reason(s) for your answer.

.....

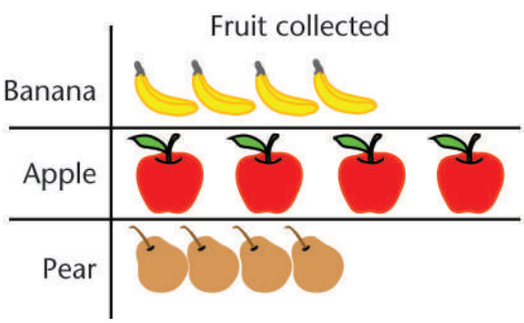
.....

.....

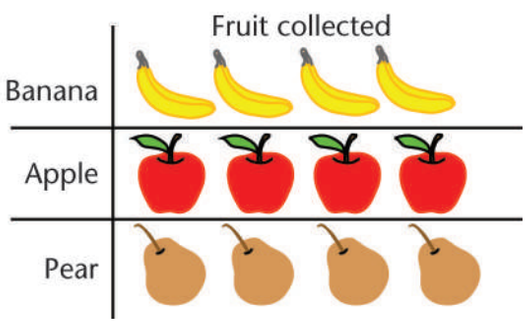
.....

3. Look at the two graphs below:

Graph A



Graph B



Which graph do you think is drawn correctly? Explain your answer.

.....

.....

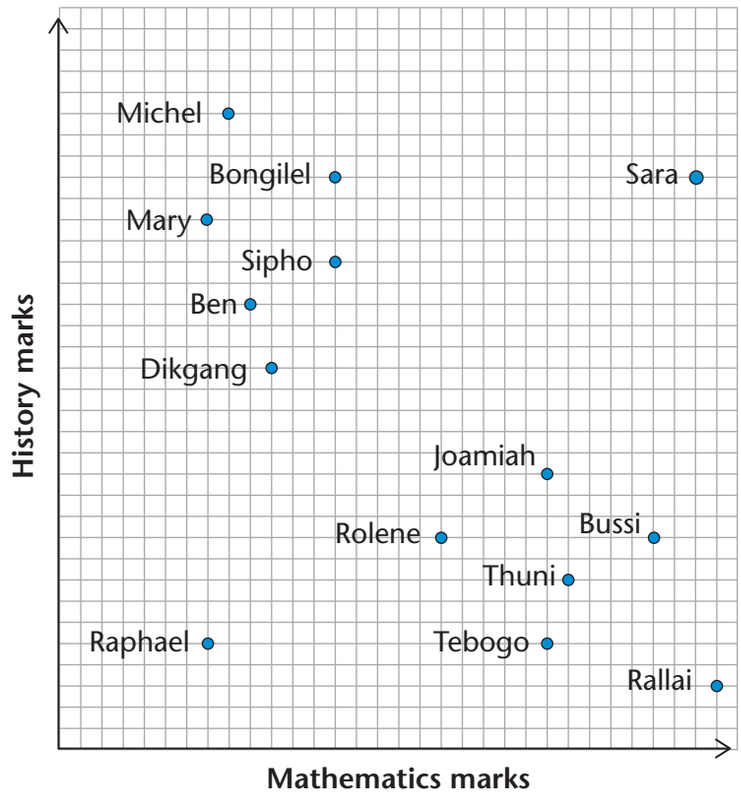
.....

.....

10.4 Analysing extreme values and outliers

A data item that is very different from all (or most) of the other items in a data set is called an **outlier**.

It is sometimes difficult to notice outliers in numerical data. However, outliers often become clearly noticeable when data is displayed with graphs.



1. The above scatter plot shows the performance of a group of learners in Mathematics and History. Which of the points on the scatter plot can be regarded as outliers? Give reasons for your answer.

.....

.....

.....

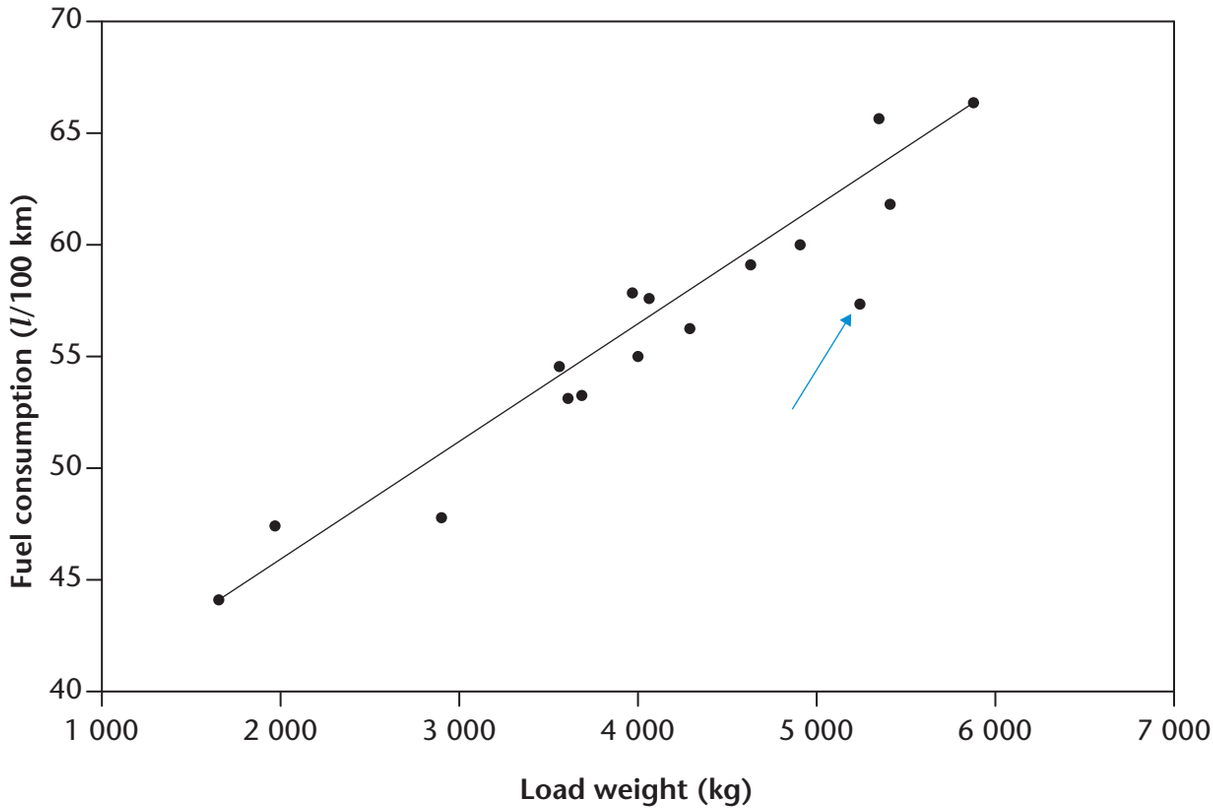
.....

.....

.....

Outliers in data sets can be very important. We need to decide whether there is a particular reason for the value being so different to the others. Sometimes it gives us important information. In some cases, the data collected for that point could be wrong.

The scatter plot below is for data collected by a transport company.



The company uses just one type of truck. Before each transport job, the company has to specify the price for the job. In order to specify a price before a job, the company needs to estimate how much their costs will be for doing the job. One of the main costs is the cost of fuel, and the main factor influencing the amount of fuel used is the distance. The load weight also plays a role: the greater the load weight, the higher the fuel consumption (litres/100 km).

The table on the next page gives information that was recorded for previous transport jobs. The jobs are numbered from 1 to 16 and for each job the values of the four variables *distance*, *load weight*, *amount of fuel used* and *fuel consumption rate* are given.

2. (a) Which of the four variables are represented on the scatter plot given above?

.....

(b) What are the values of these two variables for the point indicated by the blue arrow on the scatter plot?

.....

Job number	Distance (km)	Load weight (kg)	Fuel used (litres)	Fuel consumption (litres/100 km)
1	1 304	5 445	879	67.4
2	1 320	2 954	639	48.4
3	1 151	4 705	698	60.6
4	1 371	4 378	787	57.4
5	325	3 673	176	54.2
6	1 630	5 995	1 113	68.3
7	1 023	5 357	600	58.7
8	620	4 988	382	61.6
9	73	1 992	35	47.9
10	1 071	5 529	680	63.5
11	370	4 140	218	58.9
12	1 423	4 062	843	59.2
13	394	4 068	221	56.1
14	1 536	1 678	682	44.4
15	1 633	3 736	887	54.3
16	435	3 644	241	55.4

3. (a) Consider the scatter plot and the data set. What is the effect of load weight on fuel consumption?

.....

(b) Is job 7 an exception in this respect? Explain your answer.

.....

.....

4. Further investigations revealed that the driver for jobs 2 and 7 was the same person, and that he was not the driver for any other jobs. What may this indicate?

.....

.....

FIND OUTLIERS

Researchers collected data on the population of some African countries plus the Seychelles, the income per person, and the percentage of the income spent on health.

Country	Total population (in 1 000s)	Total annual national income per person (US$)	Percentage of income spent on health
Angola	18 498	4 830	4,6
Botswana	1 950	13 310	10,3
DRC	66 020	280	2,0
Lesotho	2 067	1 970	8,2
Malawi	15 263	810	6,2
Mauritius	1 288	12 580	5,7
Mozambique	22 894	770	5,7
Namibia	2 171	6 250	5,9
Seychelles	84	19 650	4,0
South Africa	50 110	9 790	8,5
Swaziland	1 185	5 000	6,3
Tanzania	43 739	1 260	5,1
Zambia	12 935	1 230	4,8

1. What are the three variables in this table?

   .....

   .....

   .....

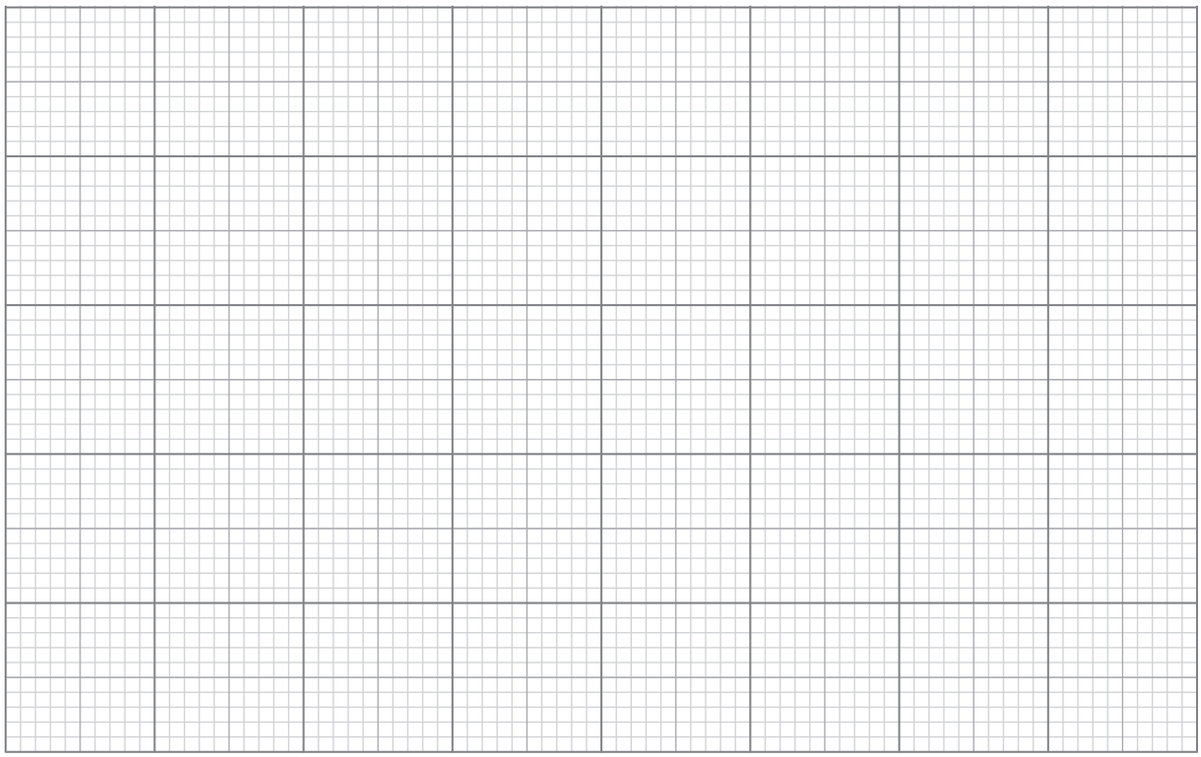
2. Why do you think it is important to look at income per person in this case, rather than the total income?

   .....

   .....

   .....

3. Plot the points for the national income per person and the percentage spent on health care for each country.



4. Write a short report on the data in the table and what the scatter plot shows you about the data. Comment on the general trend and any outliers.

.....

.....

.....

.....

.....

.....