Chapter 16: Scatter diagrams and correlation

Key words

- Correlation
- Bivariate data
- Scatter diagram
- Dependent variable
- Positive correlation
- Trend
- Negative correlation
- No correlation
- Line of best fit
- Extrapolation

In this chapter you will learn how to:

- draw a scatter diagram for bivariate data
- identify whether or not there is a positive or negative correlation between the two variables
- decide whether or not a correlation is strong or weak
- draw a line of best fit
- use a line of best fit to make predictions
- decide how reliable your predictions are
- recognise the common errors that are often made with scatter diagrams.

The term 'supply and demand' may be something that you have already heard about. Manufacturers are more likely to deliver efficient services if they fully understand the connections between the demands of customers and the quantities of goods that must be produced to make the best profit.

On a hot day it can be frustrating to go for an ice cream and find that the vendor has run out. Vendors know that there is a good link between the hours of sunshine and the number of ice creams they will need. A knowledge of how good the **correlation** is will help them ensure they have enough stock to keep everyone happy.

RECAP

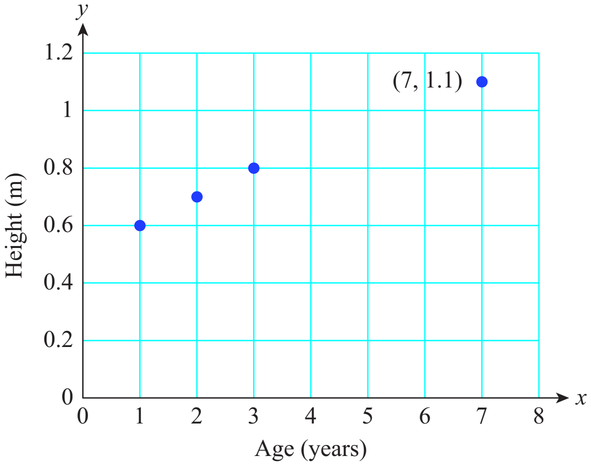
You should already be familiar with the following scatter diagram concepts:

Scatter diagrams

These graphs are used to compare two quantities (recorded in pairs).

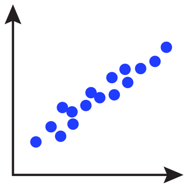
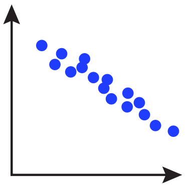
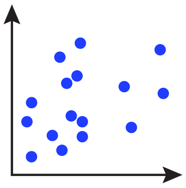
The diagram allows you to see whether the two sets of data are related (correlated) or not.

Age	1	7	3	2
Height	0.6	1.1	0.8	0.7



Correlation

The pattern of points on the scatter diagram shows whether there is a positive or negative correlation or no correlation between the variables.

Positive correlation	**Negative correlation**	**No correlation**
Points clustered around a 'line' sloping up to the right 	Points clustered around a 'line' sloping down to the right. 	Points are not in a line. 

16.1 Introduction to bivariate data

REWIND

You learned how to summarise data and draw conclusions on it in chapters 4 and 12. ◀

LINK

Correlation is used to establish relationships between variables in biology. For example, what is the relationship between the length of a particular bone and the height of a person?

So far you have seen how to summarise data and draw conclusions based on your calculations. In all cases the data has been a collection of single measurements or observations. Now think about the following problem.

An ice cream parlour sells its good throughout the year and the manager needs to look into how sales change as the daily temperature rises or falls. He chooses 10 days at random, records the temperature and records the total takings at the tills. The results are shown in the table:

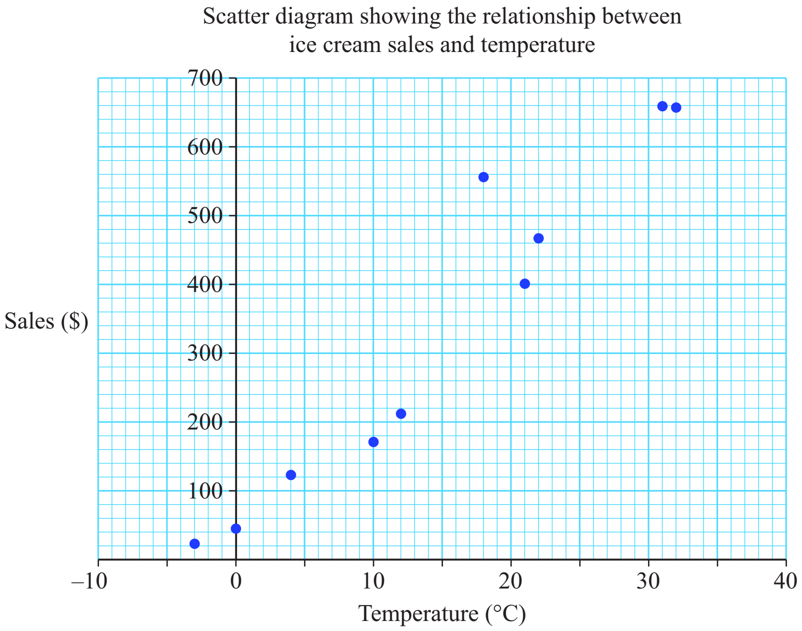
Day	A	B	C	D	E	F	G	H	I	J
Temperature (°C)	4	18	12	32	21	−3	0	10	22	31
Total takings (sales) ($)	123	556	212	657	401	23	45	171	467	659

Notice that *two measurements* are taken on each day and are recorded as *pairs*. This type of data is known as **bivariate data**. You can see this data much more clearly if you plot the values on a **scatter diagram**.

Drawing a scatter diagram

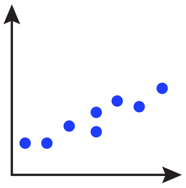
To draw a scatter diagram you first must decide which variable is the **dependent variable**. In other words, which variable depends on the other. In this case it seems sensible that the total takings will depend on the temperature because people are more likely to buy an ice cream if it is hot!

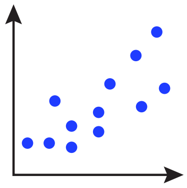
The scatter diagram will have a pair of axes, as shown below, with the dependent variable represented by the vertical axis. If the data in the table are treated as if they are co-ordinates, then the diagram begins to take shape:

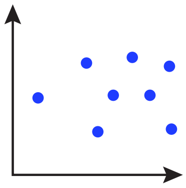


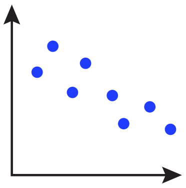
Notice that there seems to be a relationship between the ice cream sales and the temperature. In fact, the sales rise as the temperature rises. This is called a **positive correlation**. The **trend** seems to be that the points roughly run from the bottom left of the diagram to the top right. Had the points been placed from the top left to bottom right you would conclude that the sales decrease as the temperature increases. Under these circumstances you would have a **negative correlation**. If there is no obvious pattern then you have **no correlation**. The clearer the pattern, the stronger the correlation.

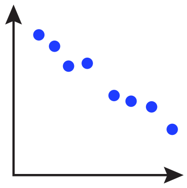
Examples of the 'strength' of the correlation:


Strong positive


Weak positive


No correlation

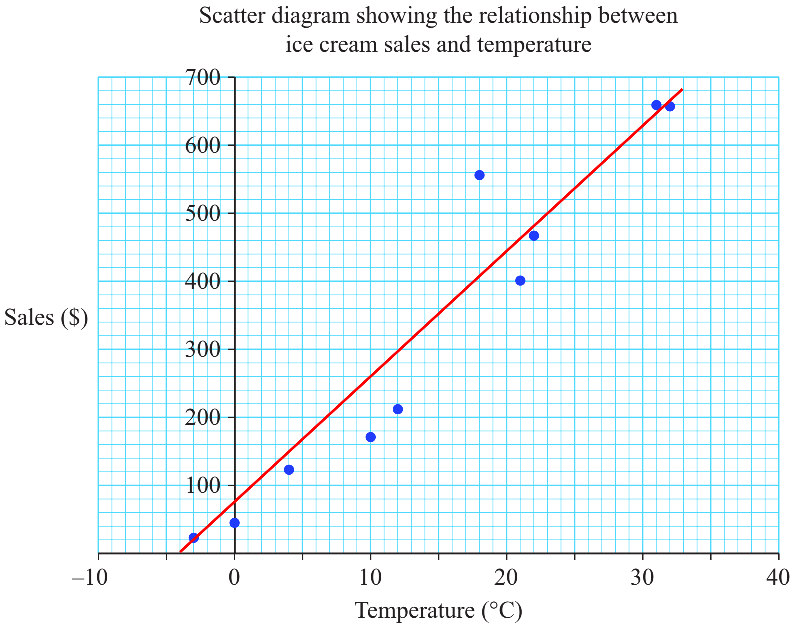

Weak negative


Strong negative

You should always be ready to state whether or not a correlation is positive, negative, strong or weak.

Notice on the graph of ice cream sales that one of the results seems to stand outside of the general pattern. Unusually high sales were recorded on one day. This may have been a special event or just an error. Any such points should be noted and investigated.

You can also show the general trend by drawing a **line of best fit**. In the diagram below a line has been drawn so that it passes as close to as many points as possible.

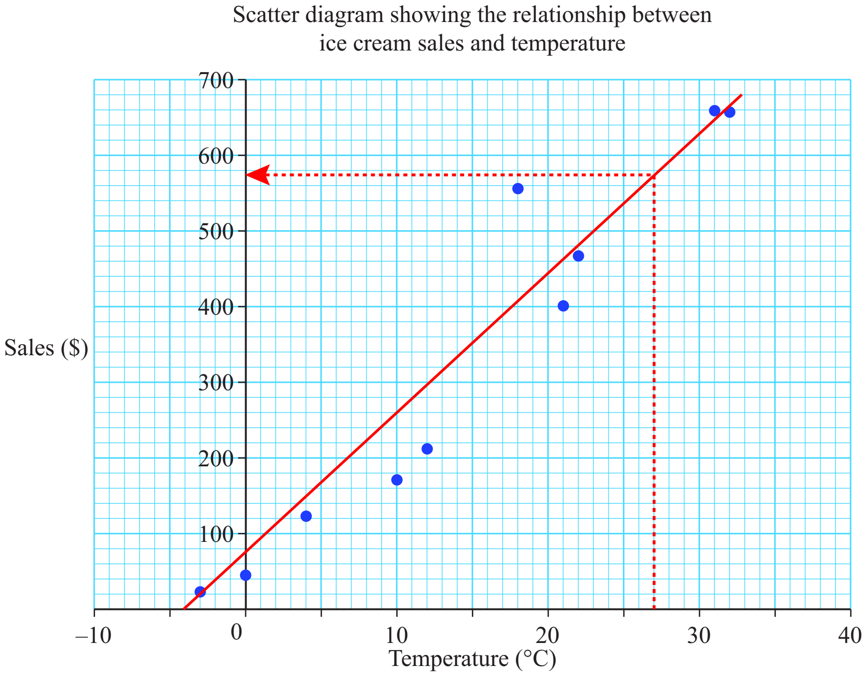


This is the line of best fit and can be used to make predictions based on the collected data.

For example, if you want to try to predict the ice cream sales on a day with an average temperature of 27°, you carry out the following steps:

1. Locate 27° on the temperature axis.
2. Draw a clear line vertically from this point to the line of best fit.
3. Draw a horizontal line to the sales axis from the appropriate point on the line of best fit.
4. Read the sales value from the graph.

The diagram now looks like this:



Here, the estimated value is approximately $575.

Worked example 1

Mr. Leatherfoot claims that a person's height, in cm, can give a very good idea of the length of their foot. To investigate this claim, Mr. Leatherfoot collects the heights and foot lengths of 10 people and records the results in the table below:

Person	A	B	C	D	E	F	G	H	I	J
Foot length (cm)	28.2	31.1	22.5	28.6	25.4	13.2	29.9	33.4	22.5	19.4
Height (cm)	156.2	182.4	165.3	155.1	165.2	122.9	176.3	183.4	163.0	143.1

a Draw a scatter diagram, with Height on the horizontal axis and Foot length on the vertical axis.

b State what type of correlation the diagram shows.

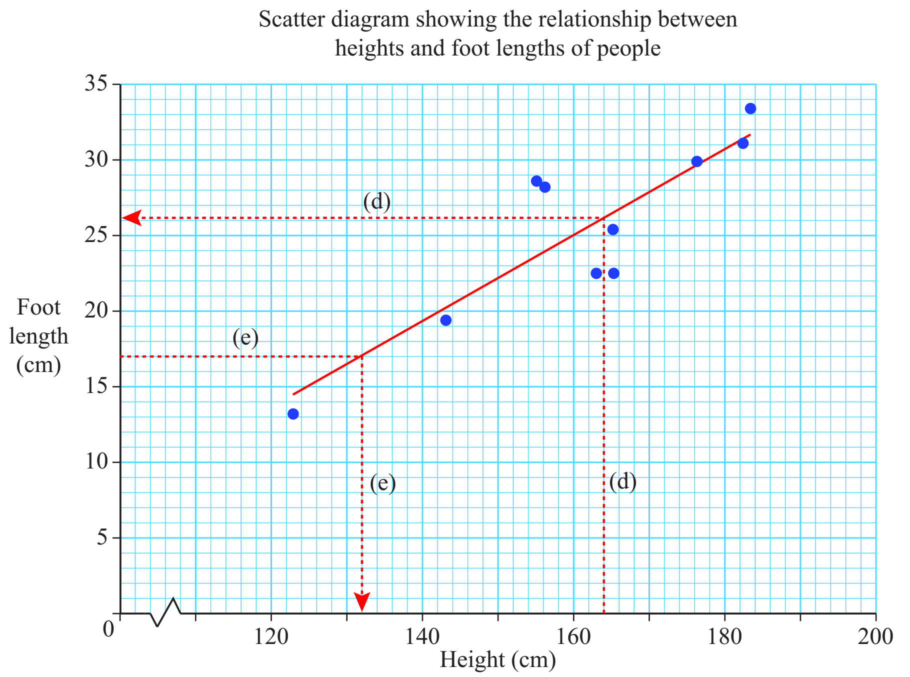
c Draw a line of best fit.

d Estimate the foot length of a person with height 164 cm.

e Estimate the height of a person with foot length 17 cm.

f Comment on the likely accuracy of your estimates in parts **(d)** and **(e)**.

a



When commenting on correlation, always make sure that you refer back to the original context of the question.

b This is a strong positive correlation because foot length generally increases with height.

c The line of best fit is drawn on the diagram.

d The appropriate lines are drawn on the diagram. A height of 164 cm corresponds to a foot length of approximately 26 cm.

e A foot length of 17 cm corresponds to a height of approximately 132 cm.

f Most points are reasonably close to the line, so the correlation is fairly strong. This means that the line of best fit will allow a good level of accuracy when estimates are made.

Golden rule

Before you try to draw and interpret some scatter diagrams for yourself you should be aware of an important rule:

- never use a diagram to make predictions *outside* of the range of the collected data.

For example, in the foot length/height diagram above, the data does not include any heights above 183.4 cm. The trend may not continue or may change 'shape' for greater heights, so you should not try to predict the foot length for a person of height, say, 195 cm without collecting more data.

The process of extending the line of best fit beyond the collected data is called **extrapolation.**

Prediction when correlation is weak

If you are asked to comment on a prediction that you have made, always keep in mind the strength of the correlation as shown in the diagram. If the correlation is weak you should say that your prediction may not be very reliable.

Stating answers in context

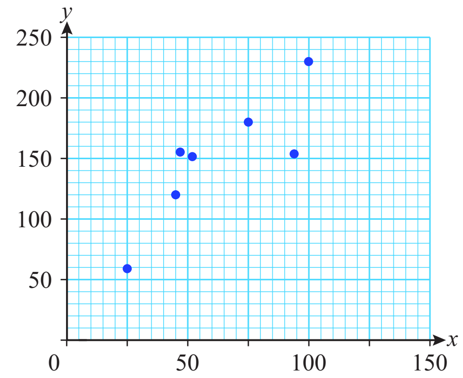
It is good to relate all conclusions back to the original problem. Don't just say 'strong positive correlation'. Instead you might say that 'it is possible to make good predictions of height from foot length' or 'good estimates of ice cream sales can be made from this data'.

Exercise 16.1

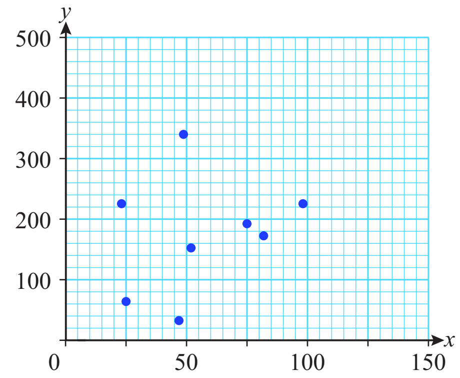
Applying your skills

1 What is the correlation shown by each of the following scatter diagrams? In each case you should comment on the strength of correlation.

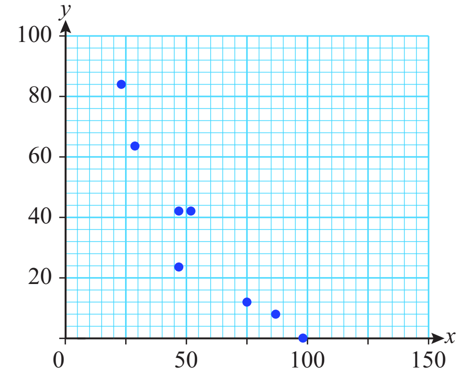
a



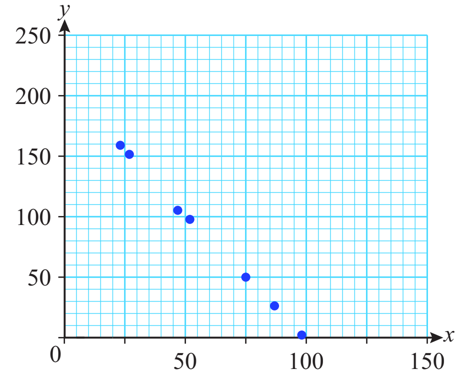
b



c



d



2 The widths and lengths of the leaves (both measured in cm) on a particular tree are recorded in the table below.

Width (cm)	14	25	67	56	26	78	33	35	14	36	13	36	25	62	25
Length (cm)	22	63	170	141	76	201	93	91	24	91	23	67	51	151	79

a Draw a scatter diagram for this data with the lengths of the leaves shown on the vertical axis.
b Comment on the strength of correlation.
c Draw a line of best fit for this data.
d Estimate the length of a leaf that has width 20 cm.

3 Emma is conducting a survey into the masses of dogs and the duration of their morning walk in minutes. She presents the results in the table below.

Duration of walk (min)	23	45	12	5	18	67	64	15	28	39
Mass (kg)	22	5	12	32	13	24	6	38	21	12

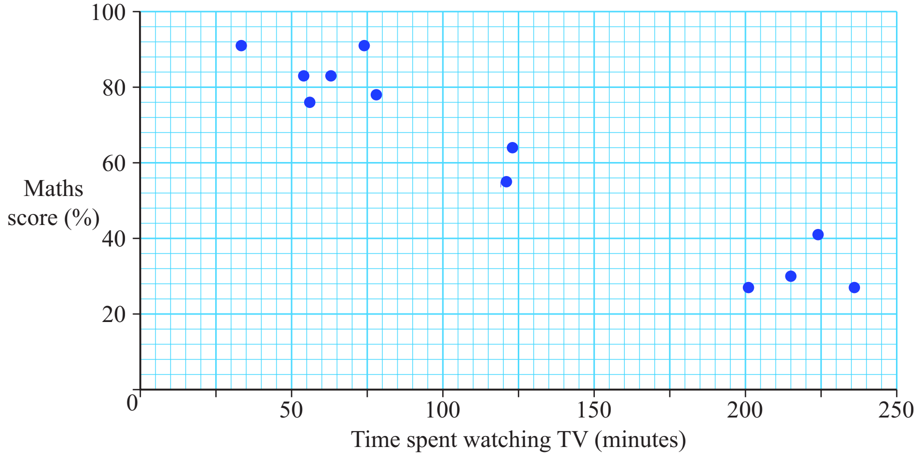
a Draw a scatter diagram to show the mass of each dog against the duration of the morning walk in minutes. (Plot the mass of the dog on the vertical axis.)
b How strong is the correlation between the masses of the dogs and the duration of their morning walks?
c Can you think of a reason for this conclusion?

4 Mr. Bobby is investigating the relationship between the number of sales assistants working in a department store and the length of time (in seconds) he spends waiting in a queue to be served. His results are shown in the table below.

Number of sales assistants	12	14	23	28	14	11	17	21	33	21	22	13	7
Waiting time (seconds)	183	179	154	150	224	236	221	198	28	87	77	244	266

a Draw a scatter diagram to show the length of time Mr. Bobby spends queuing and the number of sales assistants working in the store.
b Describe the correlation between the number of sales assistants and the time spent queuing.
c Draw a line of best fit for this data.
d Mr. Bobby visits a very large department store and counts 45 sales assistants. What happens when Mr Bobby tries to extend and use his scatter diagram to predict his queuing time at this store?

5 Eyal is investigating the relationship between the amount of time spent watching television during a week and the score on a maths test taken a week later. The results for 12 students are shown on the scatter diagram below.



Scatter diagram showing the relationship between time watching TV and maths score

The table shows some of Eyal's results, but it is incomplete.

TV watching (min)		34	215	54		78	224	236	121	74	63	
Maths score (%)	64		30	83	76	78	41		55	91	83	27

a Copy the table and use the scatter diagram to fill in the missing values.

b Comment on the correlation between the length of time spent watching television and the maths score.

c Copy the diagram and draw a line of best fit.

d Aneesh scores 67% on the maths test. Estimate the amount of time that Aneesh spent watching television.

e Comment on the likely accuracy of your estimate in part **(d)**.

Summary

Do you know the following?

- You can use a scatter diagram to assess the strength of any relationship between two variables.
- If one of the variables generally increases as the other variable increases, then you say that there is a positive correlation.
- If one of the variables generally decreases as the other variable increases, then you say that there is a negative correlation.
- The clearer the relationship, the stronger the correlation.
- You can draw a line of best fit if the points seem to lie close to a straight line.
- The line of best fit can be used to predict values of one variable from values of the other.
- You should only make predictions using a line of best fit that has been drawn within the range of the data.

Are you able to…..?

- draw a scatter diagram
- describe the relationship between the variables shown
- use a scatter diagram to make predictions.

Examination practice

Exam-style questions

1 The table below shows the sizes (in square metres) and prices (in UK pounds) of several paintings on display in a gallery.

Painting	A	B	C	D	E	F	G	H
Painting area (m^2)	1.4	2.3	0.8	0.1	0.7	2.2	3.4	2.6
Price ($)	2400	6565	1800	45	8670	4560	10 150	8950

Painting	I	J	K	L	M	N	O
Painting area (m^2)	1.1	1.3	3.7	1.5	0.4	1.9	0.6
Price ($)	3025	4560	11230	4050	1450	5420	1475

a Draw a scatter diagram for this data. The price should be represented by the vertical axis.
b Which painting is unusually expensive? Explain your answer clearly.
c *Assuming that the unusually expensive painting is not to be included* draw a line of best fit for this data.
d A new painting is introduced to the collection. The painting measures 1.5 m by 1.5 m. Use your graph to estimate the price of the painting.
e Another painting is introduced to the collection. The painting measures 2.1 m by 2.1 m. Explain why you should not try to use your scatter diagram to estimate the price of this painting.

2 A particular type of printing machine has been sold with a strong recommendation that regular maintenance takes place even when the machine appears to be working properly.

Several companies are asked to provide the machine manufacturer with two pieces of information: (x) the number of hours spent maintaining the machine in the first year and (y) the number of minutes required for repair in the second year. The results are shown in the table below.

Maintenance hours (x)	42	71	22	2	60	66	102
Repairs in second year (y) (minutes)	4040	2370	4280	4980	4000	3170	940

Maintenance hours (x)	78	33	39	111	45	12
Repairs in second year (y) (minutes)	1420	3790	3270	500	3380	4420

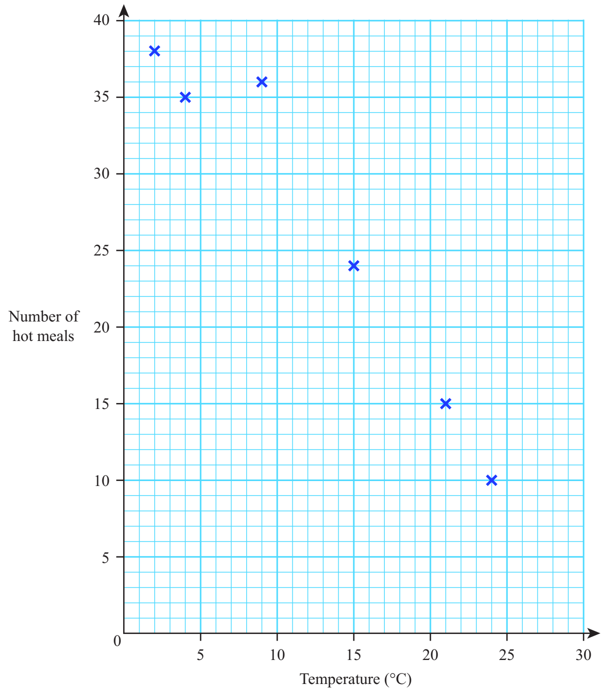
a Draw a scatter diagram to show this information. You should plot the second year repair times on the vertical axis.
b Describe the correlation between maintenance time in the first year and repair time needed in the second year.
c Draw a line of best fit on your scatter diagram.
d Another company schedules 90 hours of maintenance for the first year of using their machine. Use your graph to estimate the repair time necessary in the second year.
e Another company claims that they will schedule 160 hours of maintenance for the first year. Describe what happens when you try to predict the repair time for the second year of machine use.
f You are asked by a manager to work out the maintenance time that will reduce the repair time to zero. Use your graph to suggest such a maintenance level and comment on the reliability of your answer.

Past paper questions

1 On the first day of each month, a café owner records the midday temperature (°C) and the number of hot meals sold.

Month	J	F	M	A	M	J	J	A	S	O	N	D
Temperature (°C)	2	4	9	15	21	24	28	27	23	18	10	5
Number of hot meals	38	35	36	24	15	10	4	5	12	20	18	32

a Complete the scatter diagram.
The results for January to June have been plotted for you.



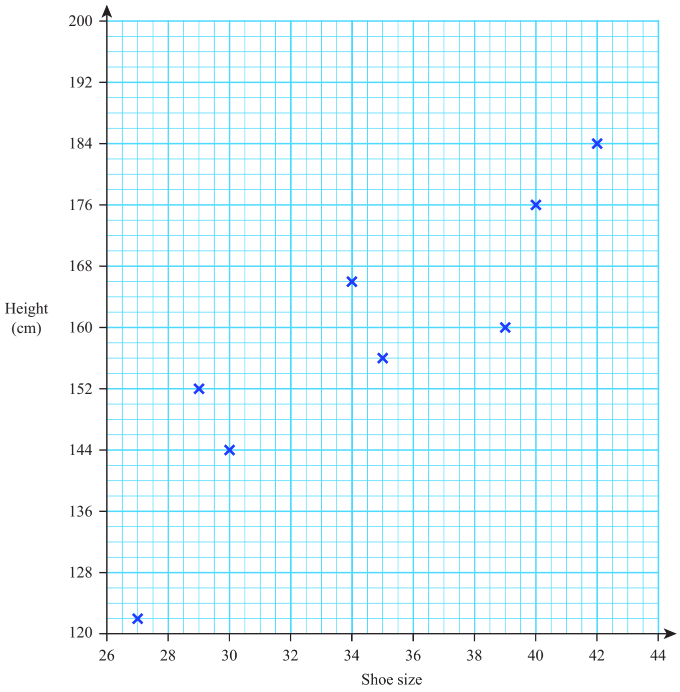
[2]

b On the grid, draw the line of best fit. [1]

c What type of correlation does this scatter diagram show? [1]

[Cambridge IGCSE Mathematics 0580 Paper 13 Q18 May/June 2013]

2 The scatter diagram shows the results of height plotted against shoe size for 8 people.



a Four more results are recorded.

Shoe size	28	31	38	43
Height (cm)	132	156	168	198

Plot these 4 results on the scatter diagram. [2]

b Draw a line of best fit on the scatter diagram. [1]

c What type of correlation is shown by the scatter diagram? [1]

[Cambridge IGCSE Mathematics 0580 Paper 12 Q17 October/November 2014]