UNIT: DATA ANALYSIS

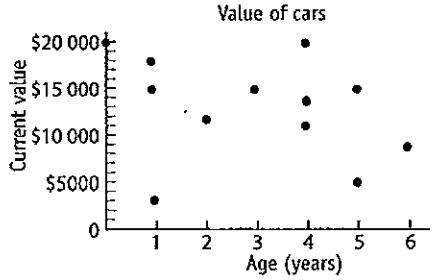
(iii) Correlation

Scatterplot

A scatterplot can be used to show the **relationship** between two separate sets of numerical data. One set of data is shown on the **horizontal** axis and the other on the **vertical** axis.

When drawing a scatterplot make sure that both axes are clearly labelled and that the scale is shown correctly. (All gaps between values on the axes must be the same.)

Example 1:
A scatterplot has been drawn showing the age and current value of certain cars.



(a) How many vehicles are shown on the scatterplot?
(b) What is the current value of the car that is 2 years old?
(c) What ages are the cars that are valued at $15 000?
(d) What is the difference in value between the two cars that are 5 years old?
(e) One vehicle has been extensively damaged in an accident. Which vehicle do you think this might be? Justify your answer.

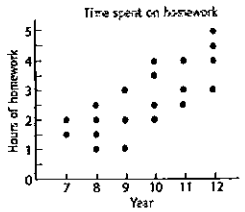
ANS
(a) 12 (b) $12 000
(c) There are 3 cars valued at $15 000. One is 1 year old, one is 3 years old and the other is 5 years old.
(d) Of the 5-year-old cars, one is valued at $15 000 and one at $5000. Difference = $15 000 - $5000 = $10 000
(e)The 1-year-old car valued at $3000. It is worth a lot less than you would expect for a car that is almost new.

Example 2:
The table shows the results of a survey of 20 high school students who were asked what year they were in at school and the number of hours they spent on homework each night.

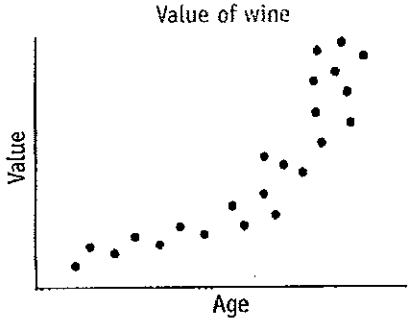
Year	Hours	Year	Hours
7	2	12	5
9	1	10	2½
10	3½	9	3
11	3	11	2½
8	2	8	1½
12	4	9	2
7	1½	12	4½
10	2	11	4
8	2½	12	3
8	1	10	4

Show the results on a scatterplot, placing the year on the horizontal axis and the hours on the vertical axis.

ANS:



Example 3:
A scatterplot has been drawn showing the value of certain bottles of wine as they age.



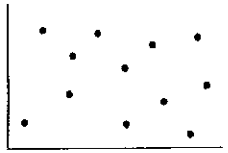
Describe any pattern.

ANS: The dots follow a pattern. The value increases slowly with age at first, then much more quickly. The pattern resembles an exponential curve.

A pattern is said to be linear if it forms a straight line. (By this we don't mean that every point falls on a straight line as if it had been drawn with a ruler, but rather that we can see a line formed by the points.)

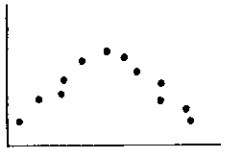
Example 4:
Determine whether or not there is a pattern and if so whether it is linear.

(a)

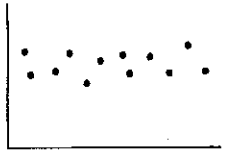


(b)

(c)



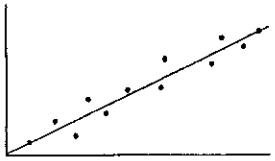
(d)



ANS: (a) no pattern (b) linear pattern (c) pattern, but not linear (d) linear pattern

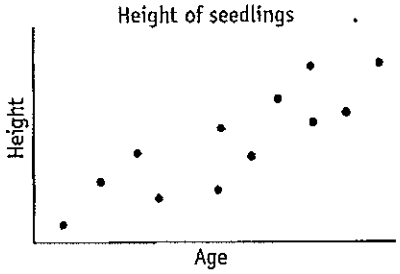
Line of fit (best fit)

A line of fit is a straight line that can be drawn to approximately show where most points lie.



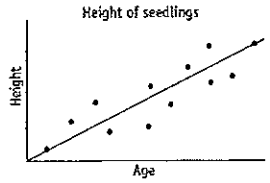
Note: There is no exact answer for a line of fit. Try to draw the line so that it goes roughly through the middle of the set of points and if possible passes through 2 points.

Example 5:
A scatterplot has been drawn showing the heights of certain seedlings after different lengths of time.



Draw a line of fit on the graph.

ANS:

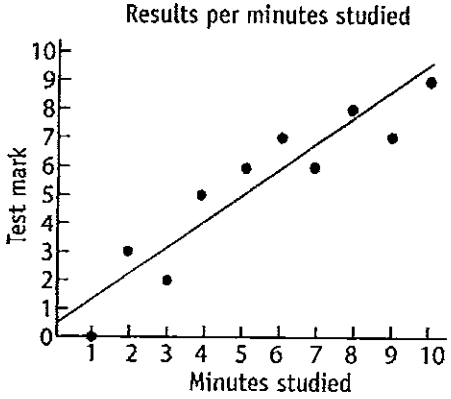


Example 6:
As an exercise in proving the value of study, a teacher gave ten students a sheet of strange facts. She gave them each a different number of minutes, from one to ten, to study the facts. She then gave a quick quiz. The table below shows the results.

Minutes studied	1	2	3	4	5	6	7	8	9	10
Test Mark	0	3	2	5	6	7	6	8	7	9

Draw a scatterplot of the results and draw in a line of fit.

ANS:

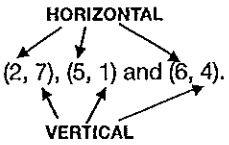


Median of points

On a scatterplot, the **median of a set of points** is the point for which the **horizontal part** is the **median** of the **horizontal values** and the **vertical** part is the **median** of the **vertical values.**

Example 7:

Find the median point of the points (2, 7), (5, 1) and (6, 4).



ANS:

Horizontal values are 2, 5 and 6. ∴ The median is 5.

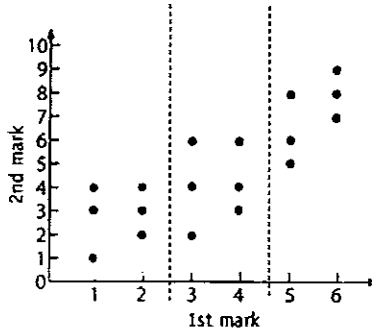
Vertical values are 1, 4 and 7. [In order] ∴ The median is 4. ∴ Median point is at (5, 4).

∴

**Median regression line

A median regression line is a more accurate line of fit. To draw a median regression Line follow these steps.

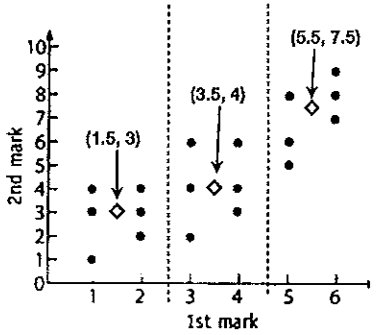
Step 1: divide the sets of points into three groups (with two vertical lines). There should be the **same** number of points in each group if possible; if not, the number of points in each group should be as close as possible, with the **same** number of points in the **1st** and **3rd** sections.



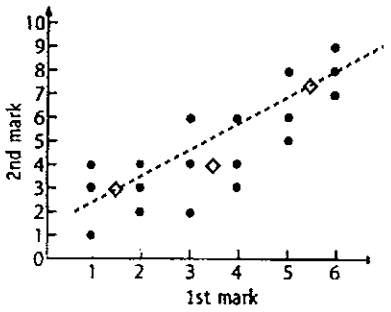
Step 2: The median point is then found **for each group** of points, and **marked** on the scatterplot.

1st group:
Horizontal values are 1, 1, 1, 2, 2, 2 ∴ Median is 1.5.
Vertical values are 1, 2, 3, 3, 4, 4. ∴ Median is 3. ∴ Median point is (1.5, 3)

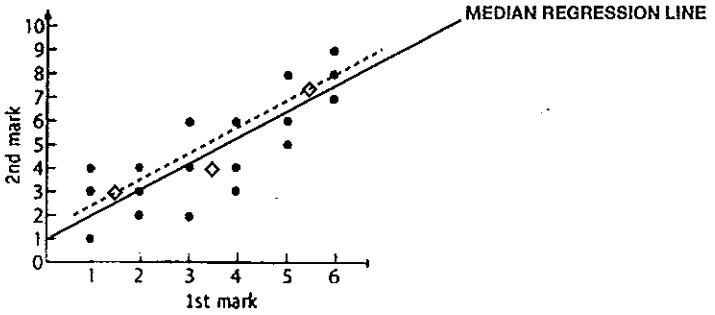
2nd group: Median point is (3.5, 4).
3rd group: Median point is (5.5, 7.5).



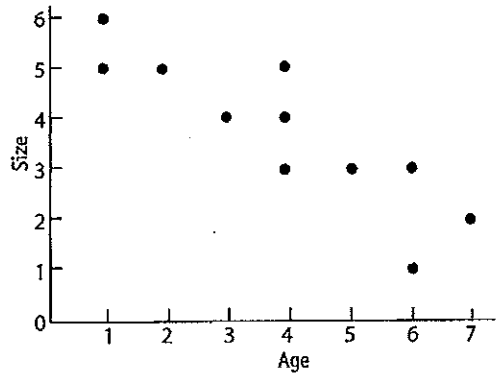
Step 3: Join the **medians** of the **1st** and **3rd** groups. This gives the slope of the required line.



Step 4: Slide this line **one third** of the way towards the **middle median point.** This gives the correct position for the **median regression line.**



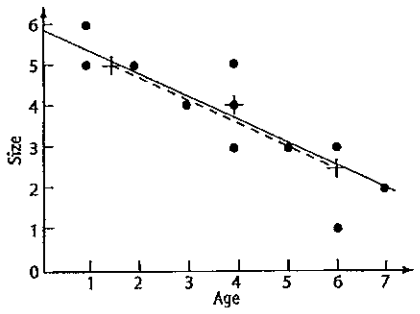
Example 8:
Consider the scatterplot below.



(a) Find the median of the first four points (moving from left to right across the scatterplot).
(b) Find the median of the middle three points.
(c) Find the median of the last four points.
(d) Draw in the median regression line.

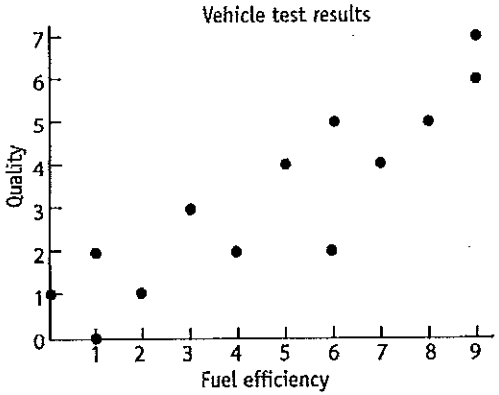
ANS:

(a) Horizontal values: 1, 1, 2, 3	**Median = 1.5**	**Vertical values: 4, 5, 5, 6**	**Median = 5**	**Median point is (1.5, 5).**
(b) Horizontal values: 4, 4, 4	**Median = 4**	**Vertical values: 3, 4, 5**	**Median = 4**	**Median point is (4, 4).**
(c) Horizontal values: 5, 6, 6, 7	**Median is 6.**	**Vertical values: 1, 2, 3, 3**	**Median is 2.5.**	**Median point is (6, 2.5).**



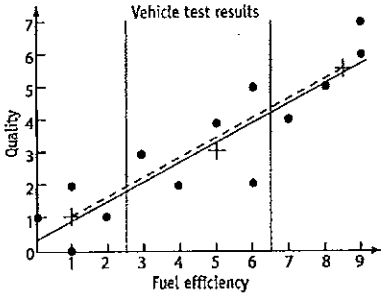
Example 9:

This scatterplot shows the results when a group of cars were tested for quality and fuel efficiency, and given a score out of ten in each category. Draw in the median regression line for the scatterplot.



ANS:

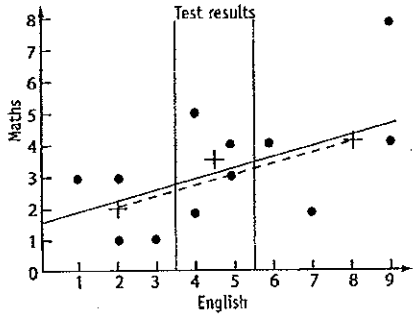
median 1st group = (1, 1) **median 2nd group (5, 3)** **median 3rd group (8.5, 5.5)**



Example 10:
Twelve students sat for tests in both English and Maths and their results, given as deciles, are shown in the table below. Show the points on a scatterplot and draw in the median regression line.

ENGLISH	MATHS	ENGLISH	MATHS
1	3	5	3
4	2	2	3
9	8	4	5
2	1	6	4
7	2	3	1
5	4	9	4

ANS:

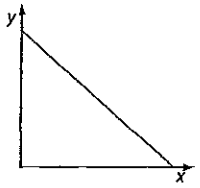


The Gradient of a line of fit

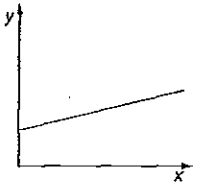
The **gradient** of a line is its **slope**. The gradient is **positive** if the line leans to the **right** (/) and is **negative** if the line leans to the **left** (\ ).

Example 11:
State whether the gradient of the line is positive or negative.

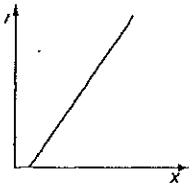
(a)



(b)



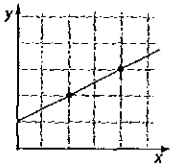
(c)



ANS: **(a) negative** **(b) positive** **(c) positive**

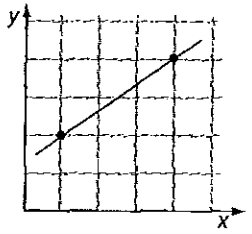
The symbol m is used for **gradient**.
The gradient of a straight line is found by dividing the **vertical** change in position by the **horizontal** change in position.

Gradient = $m = \frac{rise}{run}$



For example, the line in the diagram rises up one unit as it moves across two units, so the vertical change is 1 and the horizontal change is 2. The gradient is $m = \frac{1}{2}$

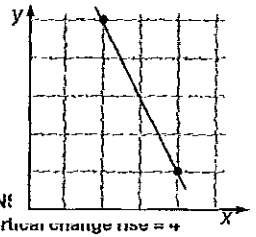
Example 12:
Find the gradient of the line joining the two given points:
(a)



ANS:
vertical change rise =2
horizontal change run = 3

$m = \frac{rise}{run} = +\frac{2}{3}$

(b)



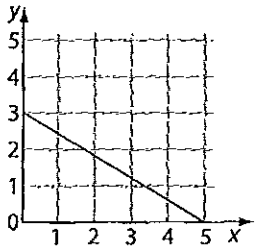
ANS:
vertical change rise = +
horizontal change run = 2

$m = \frac{rise}{run} = \frac{4}{2} = -2$

The Vertical Intercept (*y*-intercept)

The vertical intercept is where a line meets the vertical axis of a graph or plot.

Example 13:
Find the vertical intercept and gradient for this line.



ANS: *y* Int = -3 gradient = $-m = \frac{\text{rise}}{\text{run}} = -\frac{3}{5}$

The equation of a line

If the **gradient** of a straight line is ***m*** and if the **vertical intercept** is ***b***, then the equation of the line will be given by:

$$y = mx + b$$

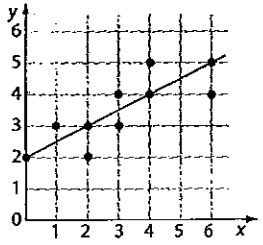
Example 14:
A line has gradient 2 and vertical intercept 3. What is the equation of the line?

ANS: $y = 2x + 3$

Example 15:
The equation of a line is given by y = 3x + 2.
(a) What is the gradient of the line?
(b) What is the vertical intercept?

ANS: (a) $m = 3$ (b) $b = 2$

Example 16:
A line of fit has been drawn for the points shown on the scatterplot as shown in the diagram.
(a) What is the gradient?
(b) What is the vertical intercept?
(c) What is the equation of the line of fit?



ANS: (a) $m = \frac{3}{6} = \frac{1}{2}$ (b) 2 (c) $y = \frac{1}{2}x + 2$

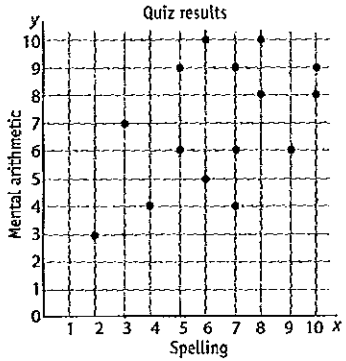
Example 17:

Pupils at a school did quizzes in spelling and mental arithmetic. The results in both quizzes for 15 of the pupils are given in the table.

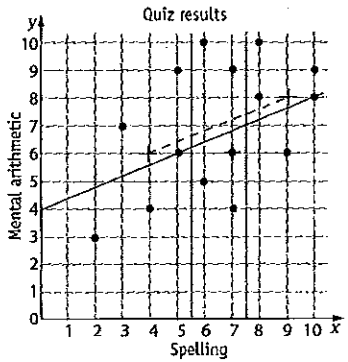
Spelling	Mental Arithmetic	Spelling	Mental Arithmetic
3	7	2	3
9	6	8	10
10	9	10	8
6	10	7	9
4	4	5	9
6	5	5	6
8	8	7	6
7	4		

(a) Show the results on a scatterplot. Place spelling on the horizontal axis and mental arithmetic on the vertical axis.
(b) Draw in the median regression line.
(c) Find the equation of the median regression line.

ANS:



Median of first five points is (4, 6).
Median of middle five points is (6.5, 6).
Median of last five points is (9, 8).



c $m = \dfrac{\text{vertical change in position}}{\text{horizontal change in position}}$

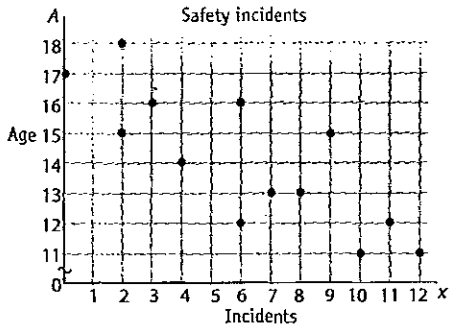
$= \dfrac{2}{5}$

Vertical intercept = 4

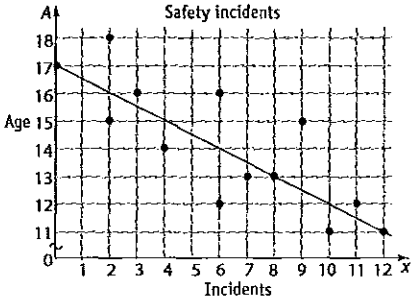
Equation is $y = \dfrac{2}{5}x + 4$

Example 18:

A survey was undertaken about the number of safety incidents in which students of different ages had been involved over a certain time period. The results are shown in the scatterplot below.



(a) Draw in a line of fit and find its equation.
(b) If a child was involved in 16 incidents, what is the likely age?
(c) If a student who is 5 were surveyed, how many incidents would you predict that he or she would be involved in?



$$m = \frac{\text{vertical change in position}}{\text{horizontal change in position}}$$
$$= -\frac{6}{12}$$
$$= -\frac{1}{2}$$

Vertical intercept = 17

Equation is $A = -\frac{1}{2}x + 17$

b When $x = 16$,

$$A = -\frac{1}{2} \times 16 + 17$$
$$= 9$$

The student is likely to be 9 years of age.

c When $A = 5$,

$$5 = -\frac{1}{2}x + 17$$
$$-12 = -\frac{1}{2}x$$
$$24 = x$$

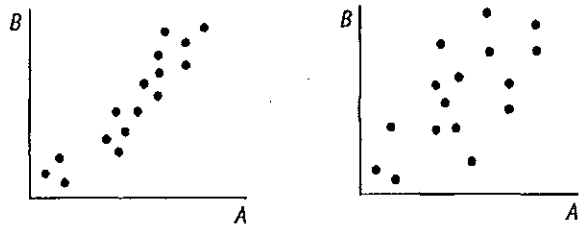
There are likely to be 24 incidents.

Correlation

Correlation refers to the **relationship between two variables.**
On a scatterplot, the **closer** the points are to **forming a straight line,** the **stronger** the correlation.

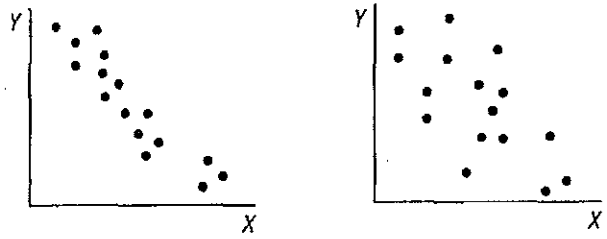
Positive Correlation
If there is a **positive correlation** between A and B say, then in general when A is small B is small, and when A is large B is large.

For example, there is a positive correlation between a person's height and the length of their feet. The taller a person, the more likely they are to have longer feet. The left diagram shows strong positive correlation and the right diagram shows weak positive correlation.

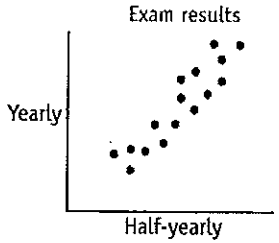


Negative Correlation
If there is a **negative correlation** between X and Y say, then in general when X is small Y is large, and when X is large Y is small.

For example, there is a negative correlation between the age of machinery and its reliability. The older a machine, the less reliable it is likely to be. The left diagram shows strong negative correlation and the right diagram shows very weak negative correlation.



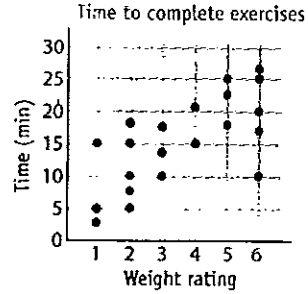
Example 19:
A scatterplot has been drawn showing the relationship between marks in the half-yearly exams and marks in the yearly exams.



Briefly describe any correlation between the two marks.

ANS: **There is a strong positive correlation between the two sets of marks.**

Example 20:
A scatterplot was drawn showing the results of an experiment where those taking part were tested on the time taken to complete an exercise. The participants were also given a weight rating. Those with a rating of 1 were under their ideal weight for their height. Those with a rating of 2 were very close to their ideal weight and those with ratings of 3 to 6 were (increasingly) overweight.



(a) Briefly describe the correlation.
(b) Saskia stated that everyone who is significantly overweight took longer to complete the exercise. Is she correct? justify your answer.

(a) There is a moderately-strong positive correlation between the time taken and the weight rating.
(b) Saskia is not correct. The strong positive correlation tells us that in general those who are overweight take more time to complete the exercise, but it is not true for every participant. For example, one person with the highest weight rating completed the task in 10 minutes, which is better than the majority of participants

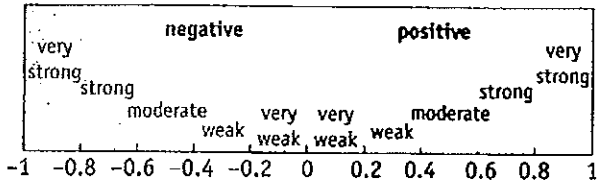
The correlation coefficient

The correlation coefficient is a number that gives us an idea of the **strength** of the **association** between two data sets.

If there is **perfect positive correlation** the **coefficient** will be **1**, and if there is **perfect negative correlation** the **coefficient** will be **-1**.

Otherwise the correlation coefficient will always be between **-1 and 1.**

If the **correlation coefficient is zero, there is no relationship** between the two data sets.



[You will not be asked to calculate the value of a correlation coefficient]

Example 21:

Students were asked to complete two different tasks and the time taken to complete each task was noted. The correlation coefficient between the time taken to perform the tasks was found to be 0.9. Briefly describe the relationship between the time taken for the tasks.

ANS: **There is a very strong positive correlation between the times taken to complete the two tasks.**

Example 22:

The correlation coefficient between the age and value of certain objects was found to be -0.3. What does this mean?

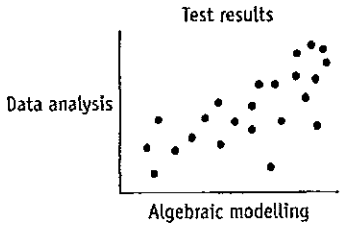
ANS: **There is a weak negative correlation between the age and value of the objects.**

**Causality

If there is a correlation between two variables, **it does not mean that one causes the other.** For example, there might be a strong correlation between the swimming ability and athletic ability of a group of students, but the fact that a child is a good swimmer doesn't cause him or her to be a good athlete.

Example 23:

The scatterplot shows the results of two different maths tests taken by students in year 12.



David commented that to do well in data analysis you must be good at algebraic modelling. Is this correct? Justify your answer.

ANS:
It is not necessarily correct. All those who did well in data analysis also did well in algebraic modelling, but there is no proof that being good at algebraic modelling causes someone to do well at data analysis.

Example 24:

Jacob notices that there is a strong positive correlation between marks achieved by students studying chemistry and those studying physics. Jacob is studying both of these subjects and decides to devote all of his time to chemistry believing that if he does well in chemistry he must also do well in physics. Briefly explain why this is a bad idea.

ANS:
Not every student who does well in chemistry does well in physics. It is not the ability in chemistry that causes the ability in physics.