Student Name

Mathematics Stage 5 Core

DATA ANALYSIS B

Book 3 Interpret data involving 2 numerical variables using graphical representations

Version: 260207

Report mistakes & suggest improvements:
https://MrDingMaths.com

CONTENTS

OVERVIEW

SYLLABUS CONTENT

MA5-DAT-C-02 displays and interprets datasets involving bivariate data

Interpret data involving 2 numerical variables, using graphical representations

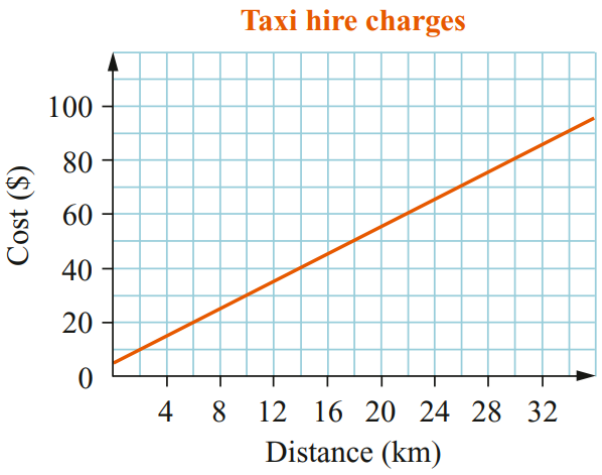
- Describe informally the association between 2 numerical variables and apply terminology about form (linear), strength (strong, moderate or weak) and direction (positive or negative)
- Use the line of best fit, by eye, to make predictions between known data values (interpolation) and what might happen beyond known data values (extrapolation)
- Explain the limitations of the model when making predictions

REVIEW OF PRIOR KNOWLEDGE

1 Identify the gradient and y-intercept in these equations.

a $y = 3x + 4$	**b** $y = 8x - 2$	**c** $y = 0.2x + 5$	**d** $y = 9 + 3x$
$m =$	$m =$	$m =$	$m =$ .....
$c =$	$c =$	$c =$	$c =$ .....
e $y = -8 + 3x$	**f** $y = 7 - 2x$	**g** $y = x - 5.52$	**h** $y = -55.68x$
$m =$	$m =$	$m =$	$m =$ .....
$c =$	$c =$	$c =$	$c =$ .....

2



a From the graph, find the cost of travelling 22 km.

.....

b How far can a passenger travel for $30?

.....

c Find the y-intercept.
What is its meaning in this situation?

.....

d Identify the gradient. What is its meaning in this situation?

.....

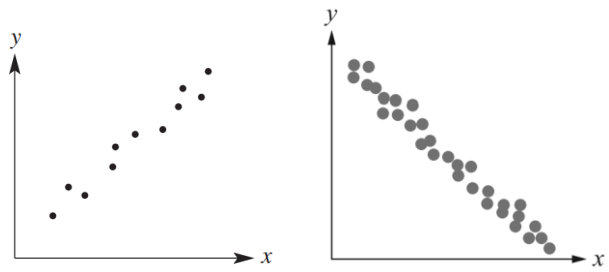
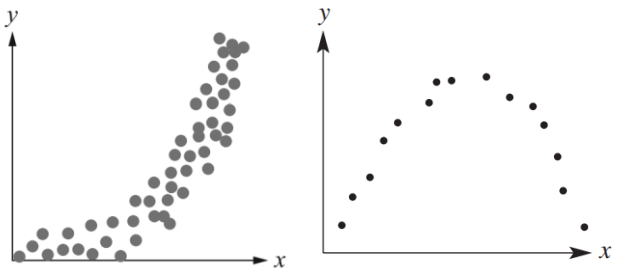
e Write and equation relating cost C and distance d .....

CORRELATION

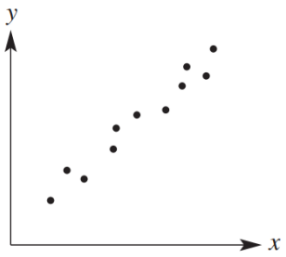
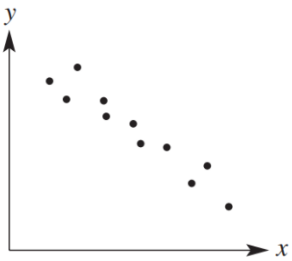
ASSOCIATION

- An **association** is a relationship between two variables.

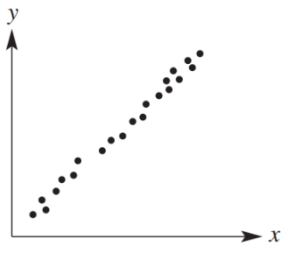
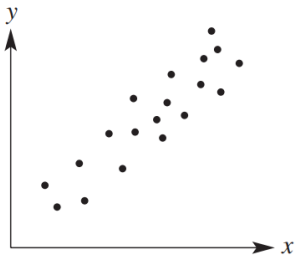
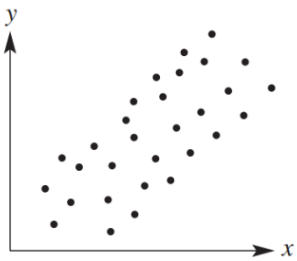
FORM OF AN ASSOCIATION

Linear (Correlation)	Non-Linear
The data points loosely form a straight line. 	The data points don't form a straight line 

DIRECTION OF A CORRELATION

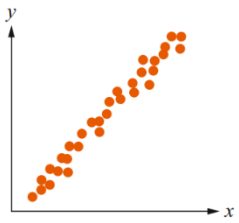
Positive	Negative
Positive gradient for a line of best fit. 	Negative gradient for a line of best fit. 

STRENGTH OF A CORRELATION

Strong	Moderate	Weak
Data points form a clear line with not much random scatter. 		Data points are significantly scattered and do not form a neat line. 

Example Identify direction and strength of a correlation.

a

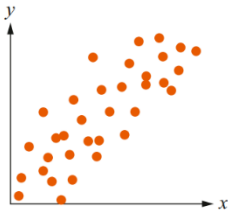


Direction:

Strength:

Strong positive

b

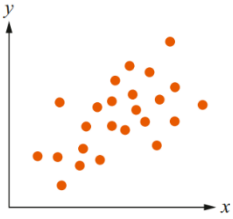


Direction:

Strength:

Moderate positive

c

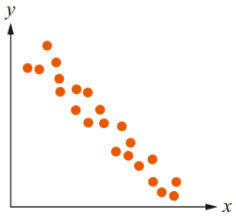


Direction:

Strength:

Weak positive

d

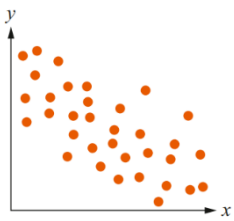


Direction:

Strength:

Strong negative

e

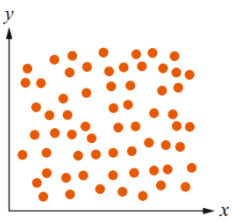


Direction:

Strength:

Moderate negative

f



Direction:

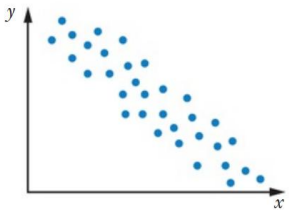
Strength:

No association

Key ideas

1. The direction of an association can be positive or negative, based on the …… of the line of best fit.
2. The strength of an association describes how …… the dots are from the line of best fit.
3. If the data points of scattered completely randomly, then there is …… association.

1



a As the value of x increases, the value of y ……

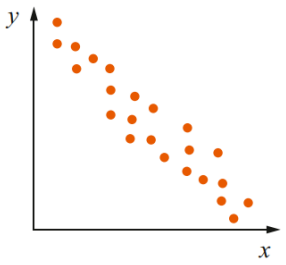
decreases

b What association is shown on the scatterplot? Positive or negative?

negative

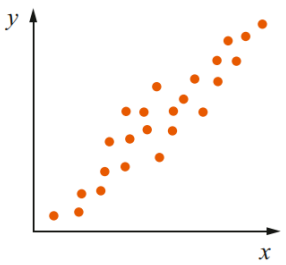
2 For each scatterplot, state whether the correlation is positive, negative, or zero.

a



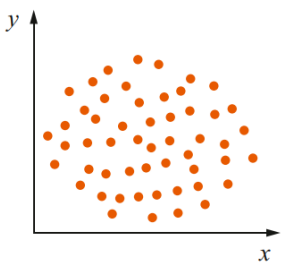
negative

b



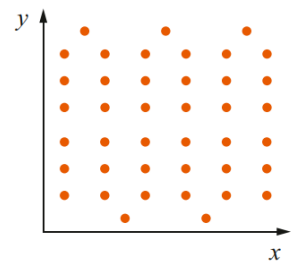
positive

c



zero

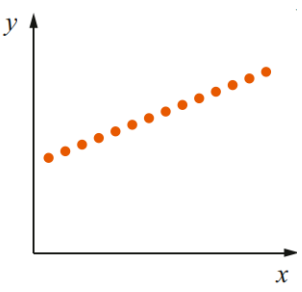
d



zero

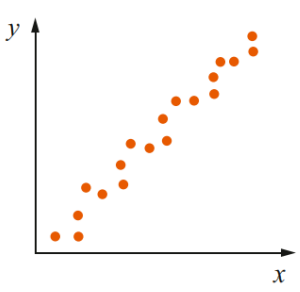
3 For each scatterplot, state whether the correlation is perfect, strong, or moderate, weak, or no correlation.

a



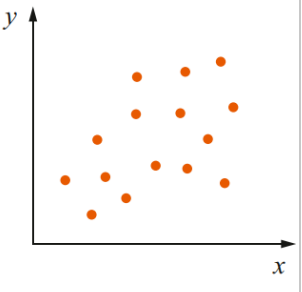
perfect

b



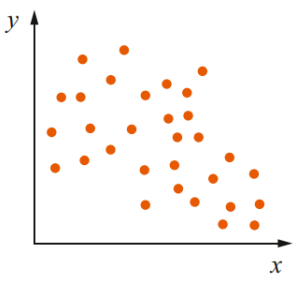
strong

c



weak

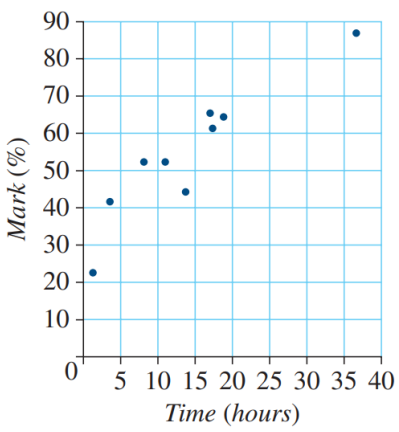
d



No correlation

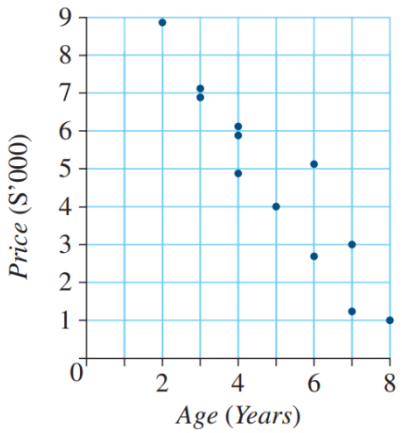
4 State whether the correlation in these scatterplots is positive/negative or strong/moderate/weak.

a



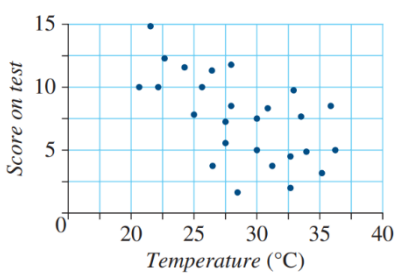
moderate

b



strong

c



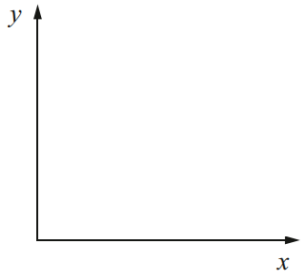
weak

5 If two variables have a strong correlation, will the data points be closer to or further from the line of best fit than with a weak correlation?

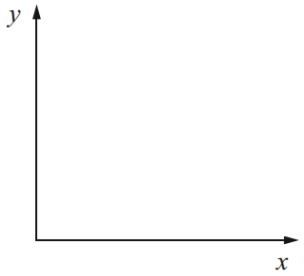
closer

6 Draw a scatterplot with these correlations.

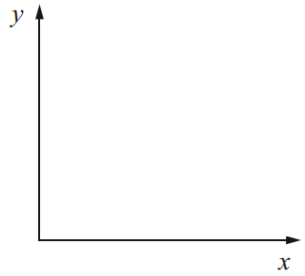
a strong positive



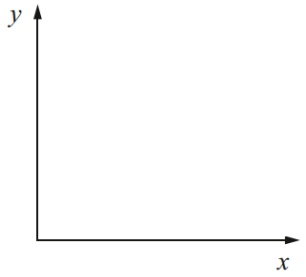
b weak negative



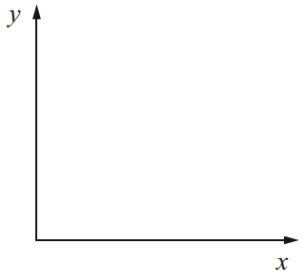
c perfect negative



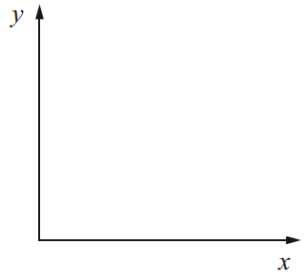
d no correlation



e perfect positive

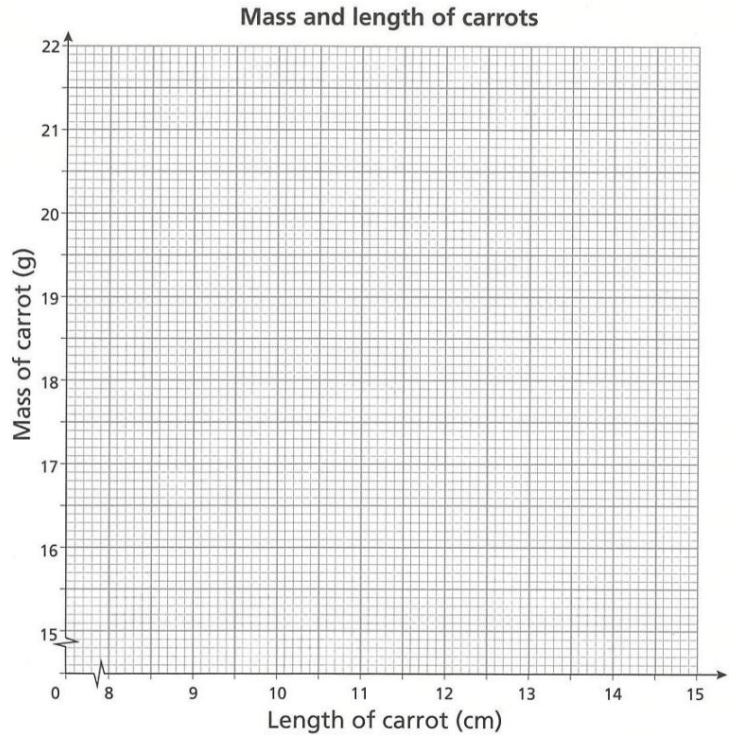


f moderate negative



7 Show the table as a scatterplot.

Length of a carrot (cm)	8	12	14.5	11	13	13.5	10	10.5	14	15
Mass of a carrot (g)	15.5	19	21	17.5	19	20	17	16.5	20.5	19



a Describe the relationship between the length of a carrot and the mass of a carrot.

The longer the carrot, the greater its mass.

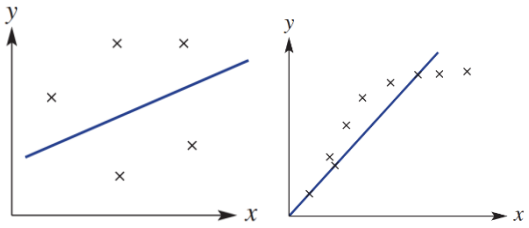
b What is the direction of the correlation shown by the scatterplot?

positive

c Why might a longer carrot have a lower mass than a shorter carrot?

A longer carrot might be thinner.

8 Explain why the shown line of best fits would **not** be appropriate for the data.



The form of the association is not linear

9 Predict whether each pair of variables have a positive, negative or zero correlation.

a number of alcoholic drinks consumed and reaction time. .....

negative

b number of cigarettes smoked and probability of lung cancer. .....

positive

c price of a packet of biscuits and the change from $10. .....

negative

d number of road accidents and number of ambulances on the road. .....

positive

e size of back muscles and swimming speed. .....

positive

f length of femur (thigh bone) and length of humerus (arm bone). .....

positive

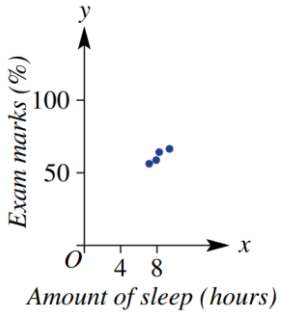
g tennis shot hit rate and time spent practising tennis shots. .....

positive

h time spent awake and time spent asleep. .....

negative

10 A person presents you with this scatter plot and suggests a strong correlation between the amount of sleep and exam marks.
What might be a problem with the person's graph and conclusions?



The axis is not appropriate, the distances between points is likely larger.

There are only 4 data points. It is not enough to make a conclusion.

11 **2012 HSC Standard 2 Band 4**

Which of the following relationships would most likely show a negative correlation?

(A) The population of a town and the number of hospitals in that town.

(B) The hours spent training for a race and the time taken to complete the race.

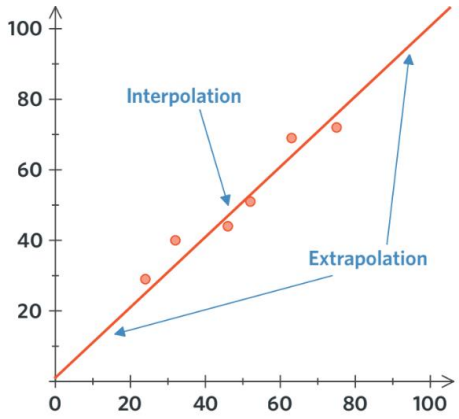
(C) The price per litre of petrol and the number of people riding bicycles to work.

(D) The number of pets per household and the number of computers per household.

B

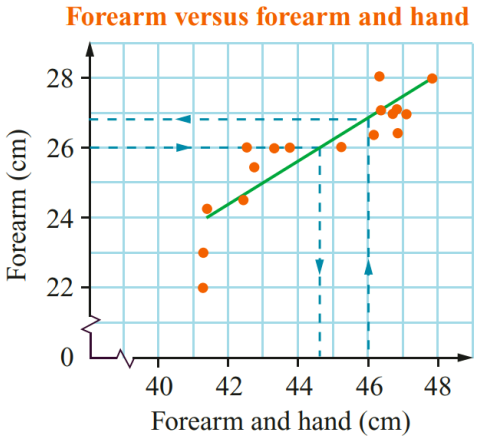
INTERPOLATION AND EXTRAPOLATION

- We use a line of best fit to **predict** data values.
- **Interpolation** predicts values within existing data.
- **Extrapolation** predicts values outside existing data.
- Interpolation is **more accurate** than extrapolation; the relationship may change beyond known data.



Example Make predictions using the line of best fit.

The scatterplot shows the length of the forearm-and-hand against the length of the forearm. Use the given line of best fit to predict:



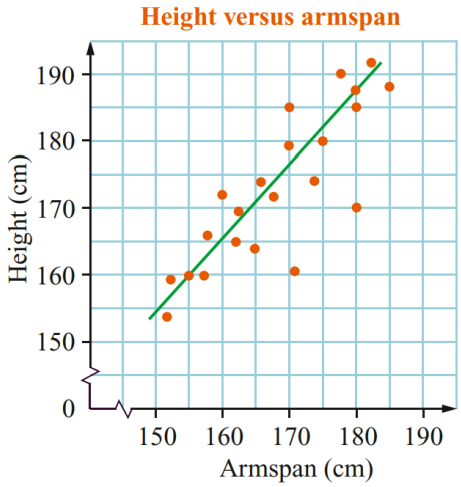
The forearm measurement for a student with a forearm-and-hand measurement of 46 cm.

Around 26.8 cm

The forearm-and-hand measurement for a student with a forearm measurement of 26 cm.

Around 44.6 cm

This scatterplot shows the heights and armspans of a group of students.



Use the given line of best fit to predict:

The height of a student with 175 cm armspan.

Around 182 cm

The armspan of a student with 185 cm height.

Around 178 cm

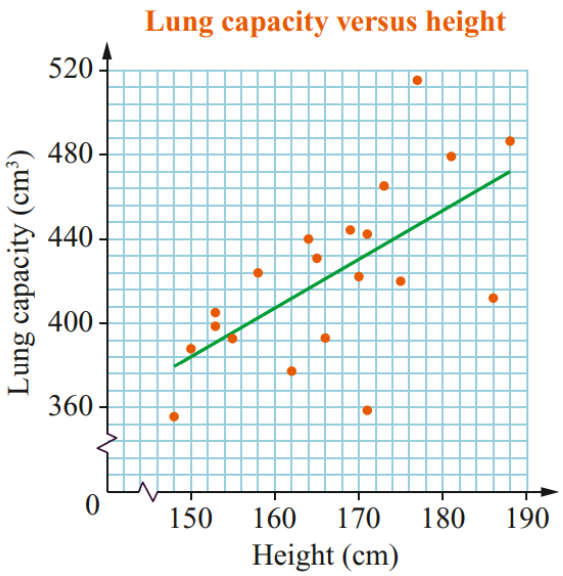
Why would it not make sense to extend the line?

There are no data points beyond the current line of best fit.

Key ideas

1. Interpolation and extrapolation uses a …… …… …… …… to make predictions based on the value of one given variable.
2. Interpolation is making predictions …… the range of known data points.
3. Extrapolating is making predictions …… the range of known data points.

1 This scatterplot shows the height and lung capacity of a group of students.
A line of best fit has been drawn for these points.
Use the line of best fit to predict:



a The lung capacity of a student with a height of:

i 160 cm ..... 408 cm³

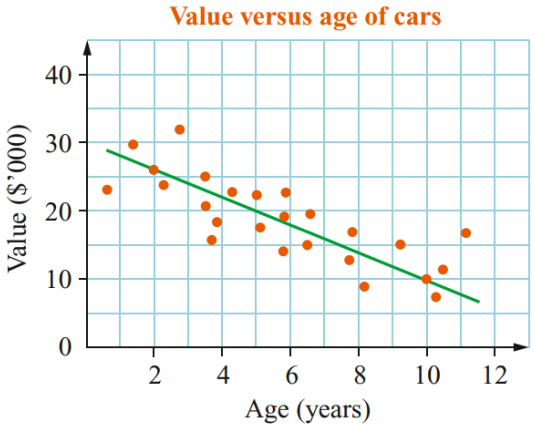
ii 178 cm ..... 448 cm³

b The height of a student with a lung capacity of

i 400 cm³ ..... 157 cm

ii 420 cm³ ..... 166 cm

2 This scatterplot shows the age and value for a sample of cars of a particular model.
A line of best fit has been drawn for these points. Use the line to predict:



a the value of a car of this model of age

i 10 years $10 000

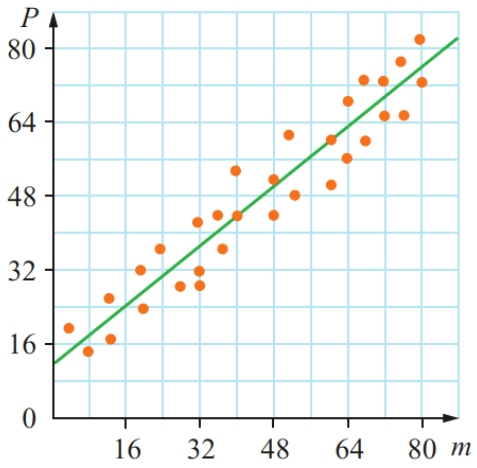
ii 5 years $20 000

b the age of a car of this model with a value of

i $15 000 ..... 7.5 years

ii $25 000 ..... 2.5 years

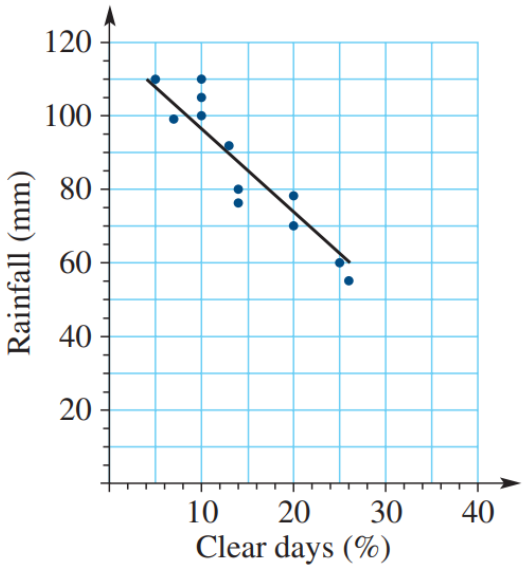
3 Use the graph to estimate:



a the value of P when $m = 80$. 75

b the value of m when $P = 50$. 48

4 The scatterplot opposite shows the rainfall (in mm) and the percentage of clear days for each month in a year.



a How many months had 10% of clear days?

3

b What was the percentage of clear days when the rainfall was 70mm?

20%

c Describe the strength of the relationship as strong, moderate, or weak.

strong

d Predict the rainfall in the month given the following percentage of clear days:

i 15% .....

85%

ii 22% .....

70%

e Predict the percentage of clear days in the month given the following rainfall:

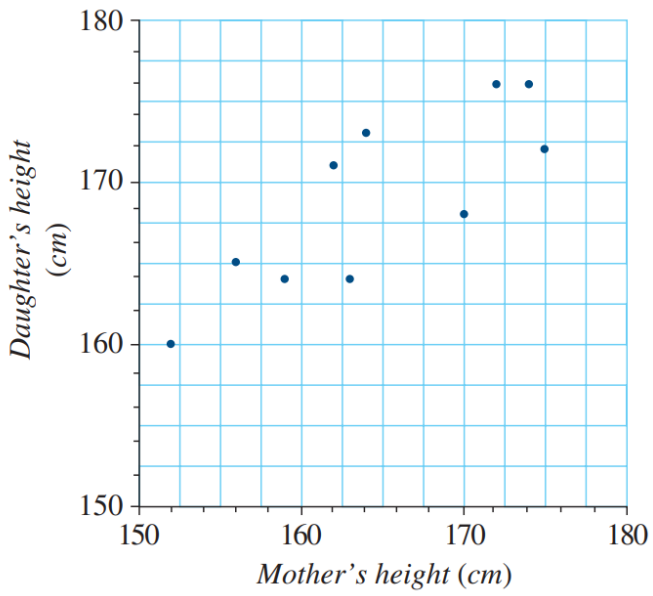
i 80 mm .....

17%

ii 90 mm .....

12.5%

5 The scatterplot shows a mother's height and her daughter's height.



a Draw a line of best fit.

b Describe the strength of the relationship as strong, moderate, or weak.

Moderate

c Estimate the daughter's height if the mother's height is 170 cm.

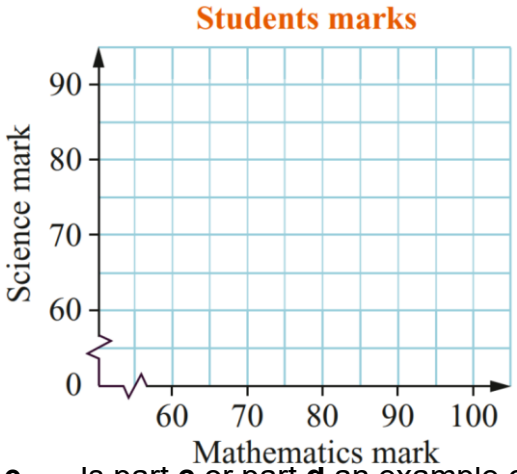
172 cm

d Estimate the mother's height if the daughter's height is 162 cm.

154 cm

6 The results of a group of students on Mathematics and Science tests are compared.

Student	1	2	3	4	5	6	7	8	9	10
Maths test (M)	64	67	69	70	73	74	77	82	84	85
Science test (S)	68	73	68	75	78	73	77	84	86	89



a Show these results as a scatterplot.

b Describe the strength of the relationship as strong, moderate, or weak.

strong

c Use a line of best fit to estimate the score in Science of a student who scores 80 in Mathematics.

82

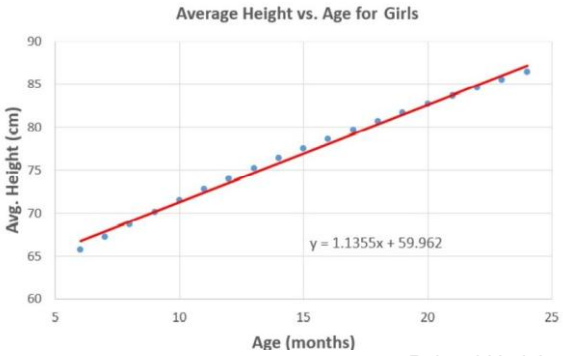
d Use the line of best fit to estimate the score in Mathematics of a student who scores 60 in Science.

56

e Is part **c** or part **d** an example of extrapolation?

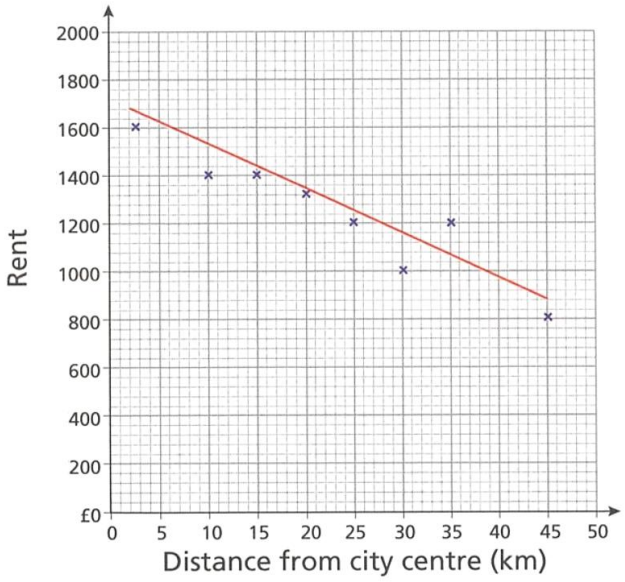
d as it is outside the range of existing data.

7 April collects data on the average height vs age for baby girls.
Explain why this line of best fit is not useful for predicting the age of 30-year-old women.



Baby girl height growth might be linear, but this trend does not continue to a woman when she is 30.

8 The scatter graph shows the cost of rent per month for office space and the distance from the city centre.



a What is the direction of correlation shown in the scatter graph?

negative

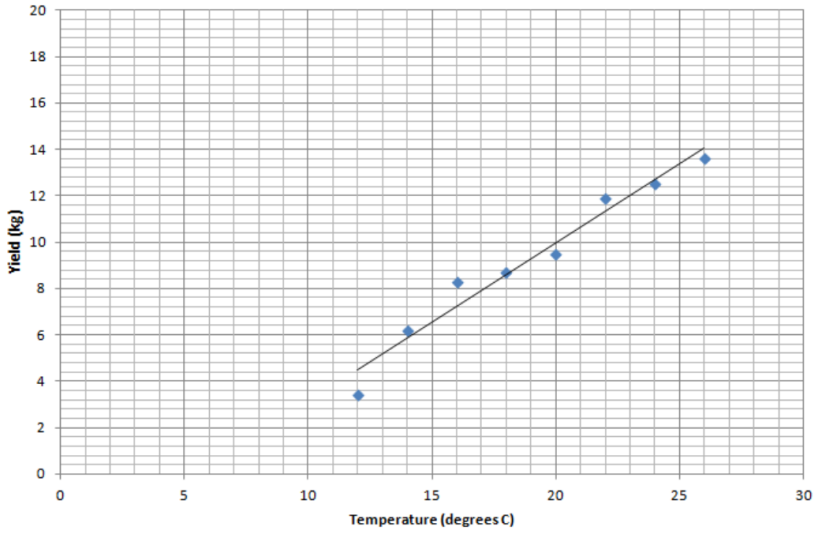
b Estimate the distance from the city centre for an office space that costs $1000 per month.

38 km

c If the graph was extended to include a distance of up to 100 km from the city centre, do you think the trend shown by the line of best fit would continue? Explain your answer.

No. The trend is unlikely to continue so far outside the existing range of data.

9 Experiments were run in a greenhouse to determine the effect of temperature on berry production. Sections of the greenhouse were kept at different temperatures and the mass of berries produced was then measured (the yield). The gathered data is shown in the scatter plot, along with the line of best fit.



a Use the line of best fit to estimate the yield for plants kept at 19°C.

9.2 kg

b At what temperature would you expect a yield of 16 kg?

28.5°C

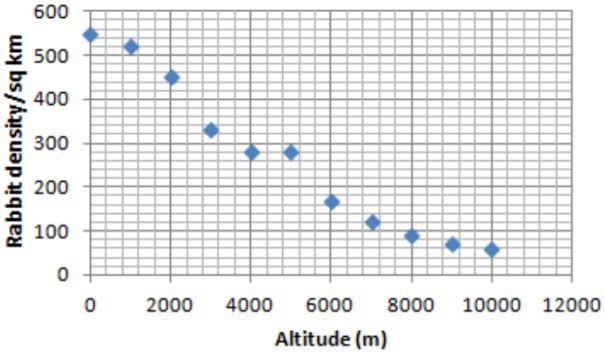
c Extrapolate to determine the approximate temperature at which the bushes would produce no berries. How might the lowest data point affect your answer?

5.5°C. The lowest datapoint could suggest that the relationship may not be linear at very low temperatures, affecting the accuracy.

d Would it be sensible to extrapolate to determine the yield for a temperature of 50°? Why?

No. Predictions beyond the existing range may not follow the same trend. Heat stress could cause the yield to drop rapidly.

10 The plot shows the density of the rabbit population per square kilometre mapped against altitude in a region.



a An altitude of 0 m indicates sea level. What does your line of best fit imply will happen to the density of the rabbit population if the land were to dip below sea level?

The line of best fit would suggest that the rabbit population would be high, which is not accurate.

b Where does your line of best fit intersect the horizontal axis and what does this imply?

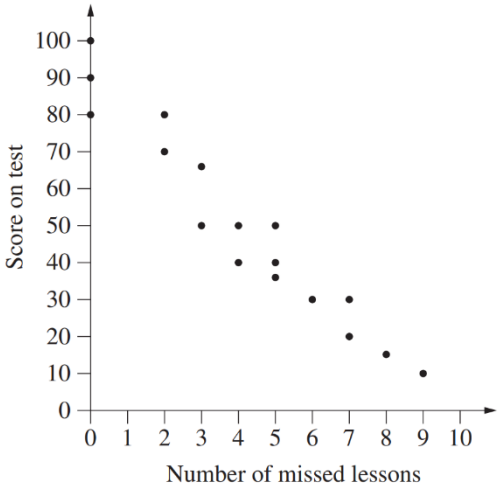
The population should be zero at around 10 000 m, which is not true.

c By observing the flow of the data points, what can you suggest is the true trend at higher altitudes?

The association is likely non-linear, there won't be many rabbits, but they could still survive at higher altitudes.

11 2020 HSC Standard 1

A group of students sat a test at the end of term. The number of lessons each student missed during the term and their score on the test are shown on the scatterplot.



Describe the strength and direction of the linear association observed in this dataset.

Strong negative

Calculate the range of the test scores for the students who missed no lessons.

100 – 80 = 20

Draw a line of best fit on the scatterplot.

Meg did not sit the test. She missed five lessons. Use the line of best fit drawn to estimate Meg's score on this test.

Between 41 and 46 (depends on your line of best fit)

John also did not sit the test and he missed 16 lessons. Is it appropriate to use the line of best fit to estimate his score on the test? Briefly explain your answer.

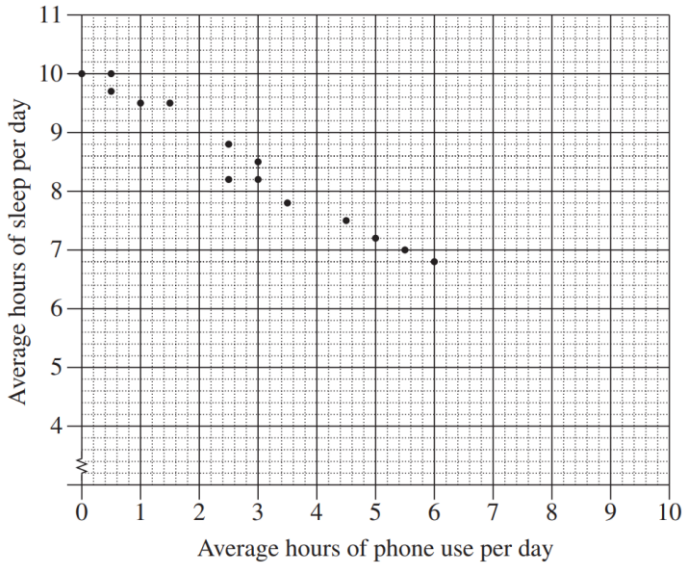
No. Extrapolation beyond the range of data is usually inappropriate.

12 2022 HSC Standard 1

A teacher surveyed the students in her Year 8 class to investigate the relationship between the average number of hours of phone use per day and the average number of hours of sleep per day. The results are shown on the scatterplot below.

The data for two new students, Alinta and Birrani, are shown in the table below.
Plot their results on the scatterplot.

	Average hours of phone use per day	*Average hours of sleep per day*
Alinta	4	8
Birrani	0	10.5



By first fitting the line of best fit by eye on the scatterplot, estimate the average number of hours of sleep per day for a student who uses the phone for an average of 2 hours per day.

Around 9 hours.

EQUATION OF A LINE OF BEST FIT

Equation of Line of Best Fit

Dependent variable Independent variable

$$y = mx + c$$

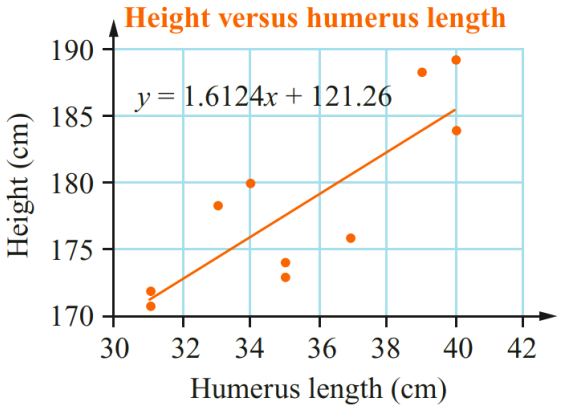
Gradient or rate of change; m per x

y-intercept y value at $x = 0$

- **Gradient** m: change in y per unit increase in x.
- **y-intercept** c: value of y when x is 0.
- The c value often lacks real meaning as the line of best fit rarely extends to zero.

Example Interpret the intercept and gradient of the regression line.

A spreadsheet calculated the equation of the line of best fit.



Interpret the meaning of the gradient.

Height increases by 1.6124 cm

per 1 cm increase in humerus length.

Interpret the meaning of the y-intercept.

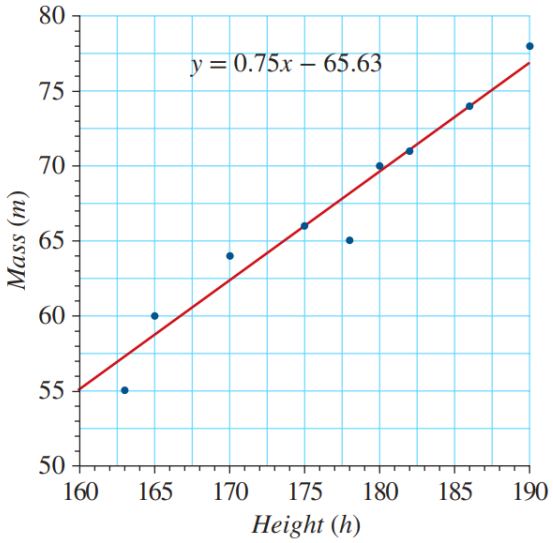
Does it make sense in this scenario?

When humerus length is 0,

height is likely 121.26 cm.

This does not make sense as the length of a humerus will never be 0.

The scatterplot shows the heights and masses of 9 people.



Interpret the meaning of the gradient.

..... increases by .....

per ..... increase in .....

Interpret the meaning of the y-intercept.

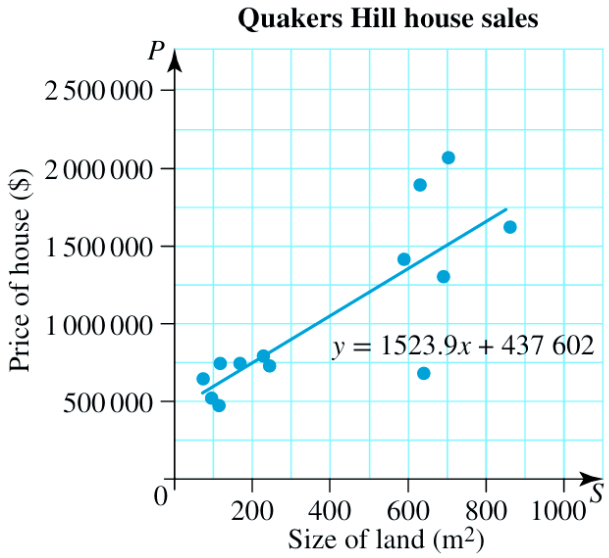
When ..... is 0,

..... would be -65.63 kg.

Does the y-intercept make sense in this scenario?

.....

1



a What is the independent variable?

Size of land (m^2)

b What is the dependent variable?

Price of house ($)

c What is the gradient?

1523.9

d What does it represent?

..... increases by .....

per ..... increase in .....

House price increases by $1523.90 per 1 m^2 increase in land.

e What is the y-intercept?

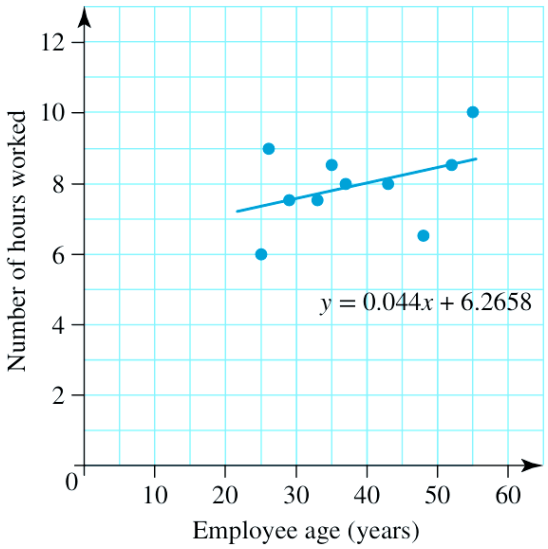
437 602

f What does it represent?

When ..... is 0, ... would be .....

When land size is 0, house price would be $437 602

2



a What is the gradient?

0.044

b What does it represent?

Hours worked increases by 0.044 hours per 1 year increase in employee age.

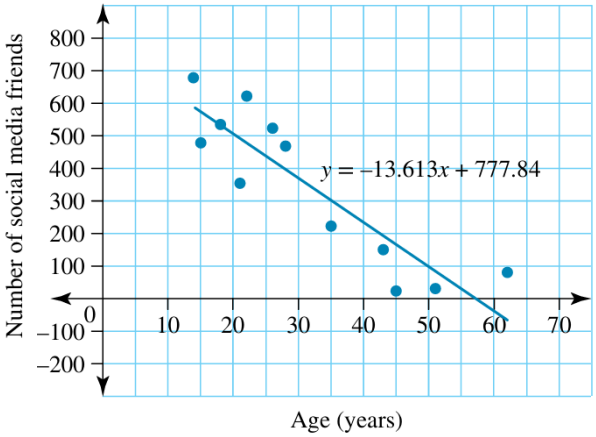
c What is the y-intercept?

6.2658

d What does it represent?

An employee who is 0 years old would work 6.2658 hours.

3



a Interpret the meaning of the gradient.

Number of social media friends decreases by 13.613 per year of age.

b Interpret the meaning of the y-intercept.

A person who is 0 years old would have 777.84 social media friends.

INTERPOLATING AND EXTRAPOLATING USING THE EQUATION

1. Identify the given variable: independent (x) or dependent (y).
2. Substitute the known value into the equation.
3. Solve to find the unknown variable.

Example Make predictions using the linear regression equation.

The equation of a line of best fit is given by $y = 0.6x - 0.7$.

The independent variable x is forearm-and-hand length.

The dependent variable y is forearm length.

Predict the forearm length of a student with forearm-and-hand length of 45 cm

1. Let $x = 45$
2. $y = 0.6(45) - 0.7$
3. $= 26.3$ cm

Why did we use $x = 45$ in this question?

The question gives us ..... length, which is the ..... variable, so x.

Predict the forearm-and-hand length of a student with a forearm length of 25 cm

1. Let $y = 25$
2. $25 = 0.6x - 0.7$
3. $25 + 0.7 = 0.6x$

$\frac{25+0.7}{0.6} = x$

$42.8 \text{ cm} = x$

Why did we use $y = 45$ in this question?

The question gives us ..... length, which is the ..... variable, so y.

The equation of a line of best fit is given by $y = 1.2x - 36$,
where the independent variable is armspan and the dependent variable is height.

Calculate the height of a student with armspan of length 178 cm.

Let ..... $= 178$

177.6 cm

Calculate the armspan length of someone with height 163 cm.

Let ..... $= 163$

165.83 cm

1 A brand of medication for babies bases the dosage on the age (in months) of the child.
The regression equation for this situation is $M = 0.312A + 0.157$, where M is the amount of medication in mL and A is the age in months.

a Identify the independent variable and dependent variable.

Independent: age in months | Dependent: amount of medication in mL

b Calculate the amount of medication required for a child aged 6 months.

Let …… $= 6$

……

……

2.029 mL

c Determine the age of a child who requires 2.5 mL of the medication.
Give your answer correct to 1 decimal place.

Let …… $= 2.5$

……

……

7.5 months

2 A survey of the nightly room rate for Sydney hotels and their proximity to the Sydney Harbour Bridge produced the regression equation $C = -50.471d + 281.92$, where C is the cost of a room per night in dollars and d is the distance to the bridge in kilometres.

a Calculate the cost of a hotel room 2.5 km from the bridge.
Give your answer correct to the nearest cent.

……

……

……

$155.74

b Determine the distance of a hotel room from the bridge if the cost of the room was $115.
Give your answer correct to 2 decimal places

……

……

……

3.31 km

3 The equation of the line of best fit connecting hip measurement (H), in cm, and waist measurement (W), in cm, is $W = 0.7H - 2.1$. Use the equation to predict:

a The waist of a person with a hip measurement of 96 cm.

65.1 cm

b The hip measurement of a person with a waist of 65 cm.

95.86 cm

4 The equation of the line of best fit connecting grape yield (G), in tonnes, of a vineyard and the number of frosts (n) during the growing season is $G = -0.14n + 5.6$. Predict:

a The yield when there are 5 frosts.

4.9 t

b The yield when there are 12 frosts.

3.92 t

c The most number of frosts if you want the yield to be at least 3.5 t.

15 frosts

5 Talula was investigating the relationship between the number of weeks (n) people have been on a new diet and their overall weight loss in pounds (w).
She determines the line of best fit for the relationship as $w = 3.5 + 1.1n$.

a Talula says the gradient is 3.5 and y-intercept is 1.1. Explain why she is incorrect.

The gradient is 1.1 and y-intercept is 3.5.

b Interpret the gradient of the line of best fit in the context of the data.

1.1 pound weight loss per week.

3.5 pound weight loss at 0 weeks!

c Does it make sense to interpret the vertical intercept in the context of the data?

No, you won't lose weight just by signing up.

6 Answer the following questions for the equation $I = 0.43 + 1.1s$
where I is the number of insects caught and s is the area of a spider's web in cm^2.

a Interpret the meaning of the gradient.

1.1 increase in insects caught per 1 cm^2 increase in web area.

b Interpret the meaning of the y-intercept.

0.43 insects caught when web area is 0 cm^2

c Does the y-intercept make sense?

No, a web area of 0 won't catch any insects.

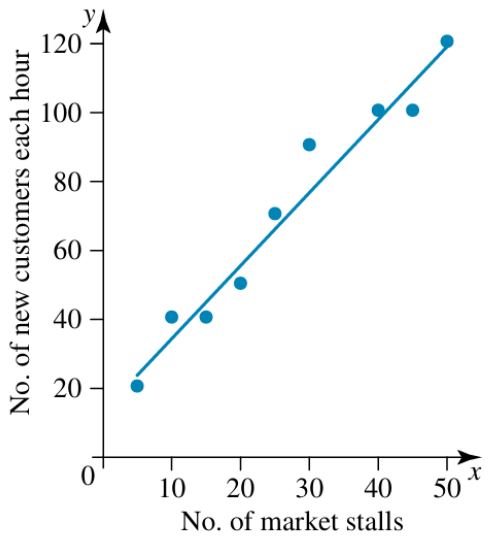
d Is the trendline of the data positive or negative?

Positive.

e Determine how many insects are likely to be caught if the area of the web is 60 cm^2.
Give your answer correct to the nearest whole number.

54 insects

7 In the scatterplot shown, the line of best fit has the equation of $c = 13.33 + 2.097m$, where c is the number of new customers each hour and m is the number of market stalls.



a Using the line of best fit, find the number of new customers expected if there are 30 market stalls.

75 new customers per hour.

b Use the equation to find the number of market stalls required to expect 150 new customers.

65 stalls

c Explain why part **a** is interpolation, while part **b** is extrapolation.

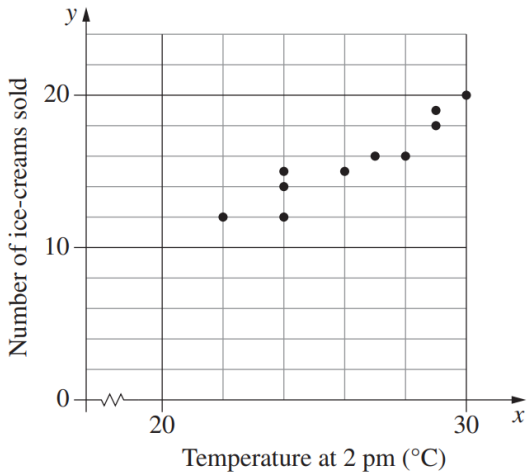
30 market stalls (x) is within the existing range of data. 150 new customers (y) is not.

8 **2023 HSC Standard 1**

The scatterplot shows the number of ice-creams sold, y, at a shop over a ten-day period, and the temperature recorded at 2 pm on each of these days.

The data are modelled by the equation of the line of best fit given below:

$y = 0.936x - 8.929$, where x is the temperature.



Sam used a particular temperature with this equation and predicted that 23 ice-creams would be sold.
What was the temperature used by Sam, to the nearest degree?

34 degrees.

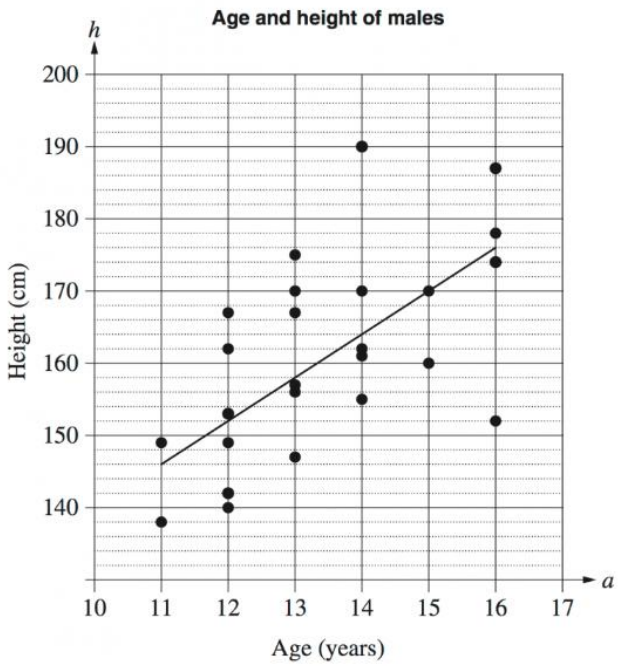
In using the equation to make the prediction in part (a), was Sam interpolating or extrapolating?
Justify your answer.

Extrapolating, as 23 ice creams is not within the existing range of data.

9 2013 HSC Standard 2 Band 5

Ahmed collected data on the age (a) and height (h) of males aged 11 to 16 years.

He created a scatterplot of the data and constructed a line of best fit to model the relationship between the age and height of males.



Determine the gradient of the line of best fit.

6

Explain the meaning of the gradient of the line of best fit.

Males should grow 6 cm per year between ages of 11 to 16

Determine the equation of the line of best fit.

$h = 6a + 80$

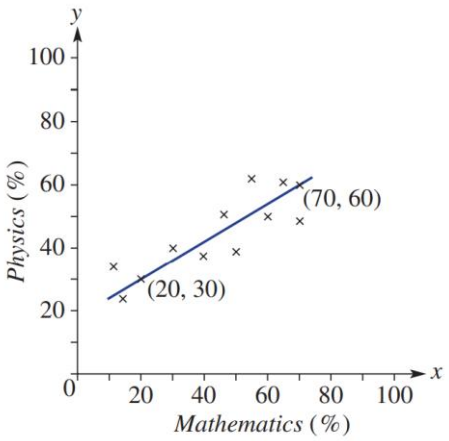
Use the equation to predict the height of a 17-year-old.

182 cm

Why would this model not be useful for predicting the height of a 45-year-old male?

People eventually stop growing when they become adults. The height vs age trend is not always linear.

10 This scatter plot shows a linear relationship between Mathematics marks and Physics marks in a small class of students. A trend line passes through (20, 30) and (70, 60).



a Find the equation of the line if the y-intercept is 18.

$y = 0.6x + 18$

b Use your equation to find the Physics score if the Mathematics score is:

i 40

42

ii 90

72

c Use your equation to find the Mathematics score if the Physics score is:

i 36

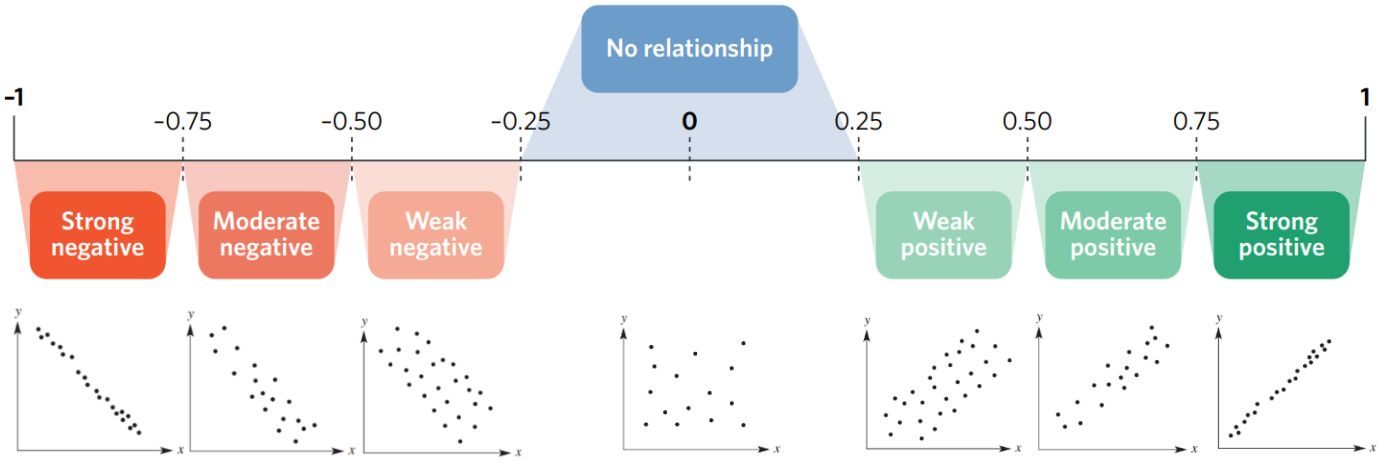
30

ii 78

100

PEARSON'S r VALUE

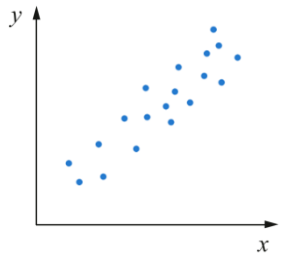
- **Pearson's correlation coefficient** (r) precisely measures association **strength**.
- r ranges from -1 to 1.
- You'll learn how to calculate r in a future lesson.
- In stronger relationships, data points are closer to the line of best fit, yielding more accurate predictions.



Example Interpret Pearson's r.

The members of a sporting club are training to increase thigh circumference.
Each fortnight, the thigh circumference of each member is measured.
At the conclusion of the training period, the correlation coefficient comparing thigh circumference with time is 0.73.

a Draw what the scatterplot might look like.



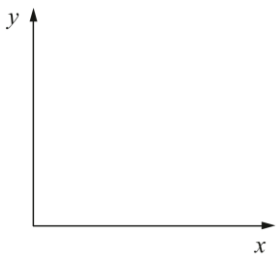
b What can you conclude about the relationship between thigh circumference and time?

There is a moderate, positive correlation between weight training and thigh circumference.

Scientists are trying to find a method of reducing knee pain.
The correlation between quadricep strength and knee pain is -0.83.

Draw what the scatterplot might look like.

What can you conclude about the relationship between quadricep strength and knee pain?



Stronger quadriceps generally lead to less knee pain.

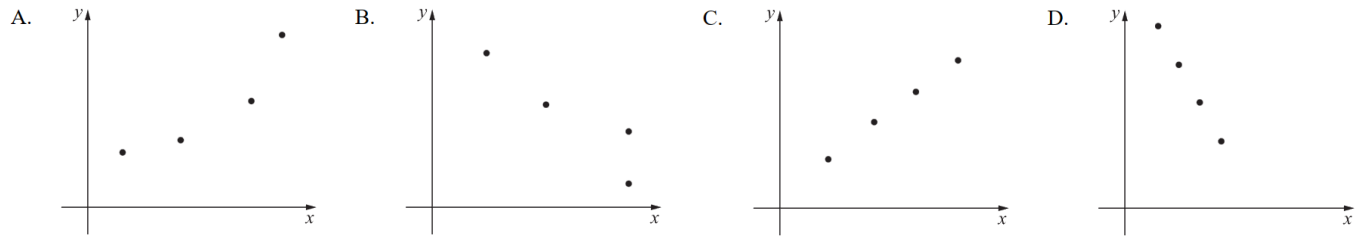
Key ideas

1. Pearson's r value precisely states the …… of a linear association.
2. The …… the magnitude of r, the more …… the predictions of a line of best fit.

1 2020 HSC Standard 2 Band 4

For a set of bivariate data, Pearson's correlation coefficient is –1.

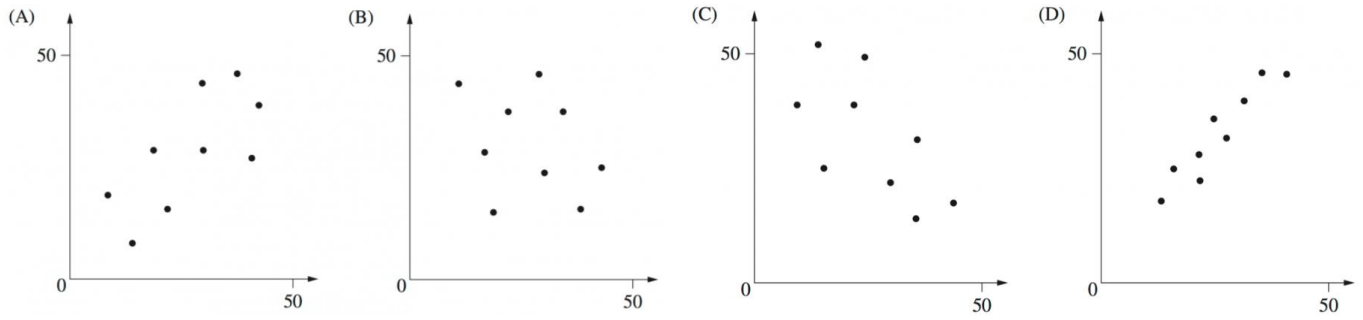
Which graph could best represent this set of bivariate data?



D

2 2013 HSC Standard 2 Band 4

Which graph best shows data with a correlation closest to 0.3?



A

3 2016 HSC Standard 2 Band 4

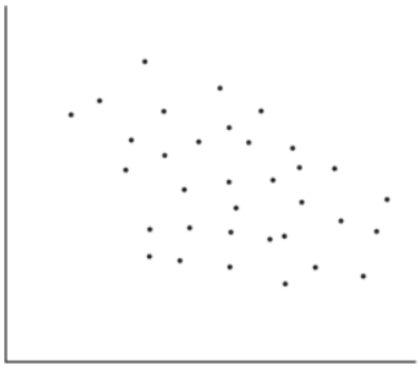
The graph shows a scatterplot for a set of data.
Which of the following is the best approximation for the correlation coefficient of this set of data?

(A) -1

(B) -0.3

(C) 0.3

(D) 1



B

4 2008 HSC Standard 2 Band 4

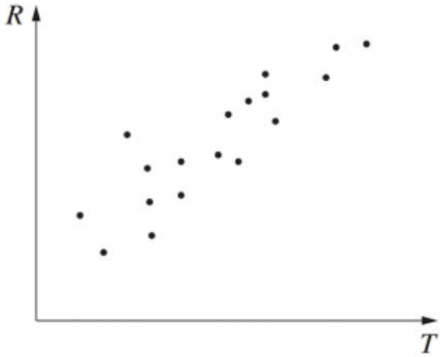
A scatterplot is shown. Which of the following best describes the correlation between R and T?

(A) Positive

(B) Negative

(C) Positively skewed

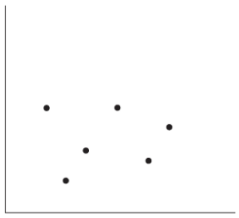
(D) Negatively skewed

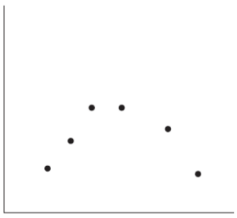


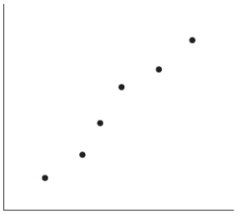
A

5 **2011 HSC Standard 2 Band 4**

In which graph would the data have a correlation coefficient closest to –0.9?

(A) 

(B) 

(C) 

(D) 

D

6 A gym is using a new training program to increase bicep circumference.
Each fortnight the bicep circumference of each person is measured.
At the end of the program, the bicep measurements are compared with time.
The correlation coefficient is 0.64. What does this mean?

……

……

……

The relationship between time in the training program and bicep circumference is moderate and positive.

7 Ten people are on a 6-week diet to lose weight. Their weight is recorded every week.

The correlation coefficient of the data is −0.77. Does this show that the diet works?

……

……

……

Yes. There is a strong negative correlation between time in the diet and weight (weight decreases with time)

8 The line of best fit comparing salary, in thousands, at age 35 (S) and number of years of study (n) is given by $S = 3.5n + 9.8$.

a Predict the salary of someone who has completed:

i 10 years of study (left school year 10)

$44 800

ii 12 years of study (no university)

$51 800

iii 17 years of study (double degree)

$69 300

b The value of Pearson's correlation coefficient is $r = 0.6$.
How accurate do you expect these predictions to be?

……

Moderately accurate.

9 HSC Sample Question Band 3

A student claimed that as time spent swimming training increases, the time to run a 1 kilometre time trial decreases.

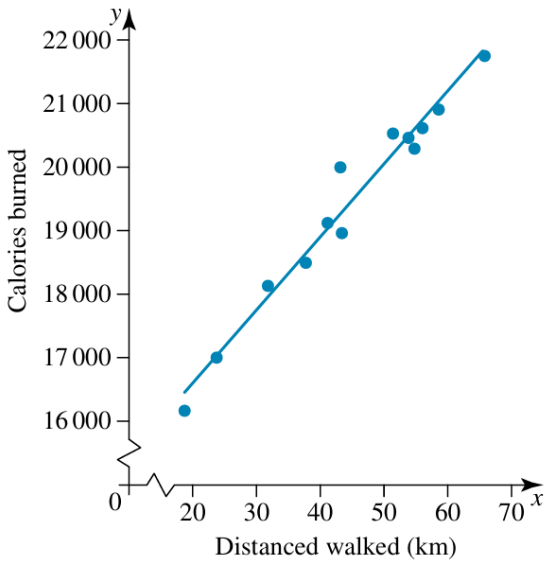
After analysing the data, the student found the correlation coefficient, r to be – 0.73.

What does this correlation indicate about the relationship between the time a student spends swimming training and their 1 kilometre run time trial times.

As time spent training swimming increases, 1 km run time decreases.

There is a moderate negative correlation.

10 For three months, Cameron has been wearing an exercise tracking wristband that records the distance he walks and the number of calories he burns. The graph shows his weekly totals. The regression line equation for this data is $y = 14301 + 115.02x$.



a Using the equation for the regression line, determine the number of calories burned if a person walked 50 km in a week.

20052 calories

b Is this interpolation or extrapolation?

Interpolation

c Due to an injury, in one week Cameron only walked 10 km and he forgot to wear his tracker. Determine the number of calories this distance would burn.

15451.2 calories

d Is this interpolation or extrapolation?

Extrapolation

e Pearson's correlation coefficient for this data is 0.9678.
What does this mean for the direction and strength of the correlation?

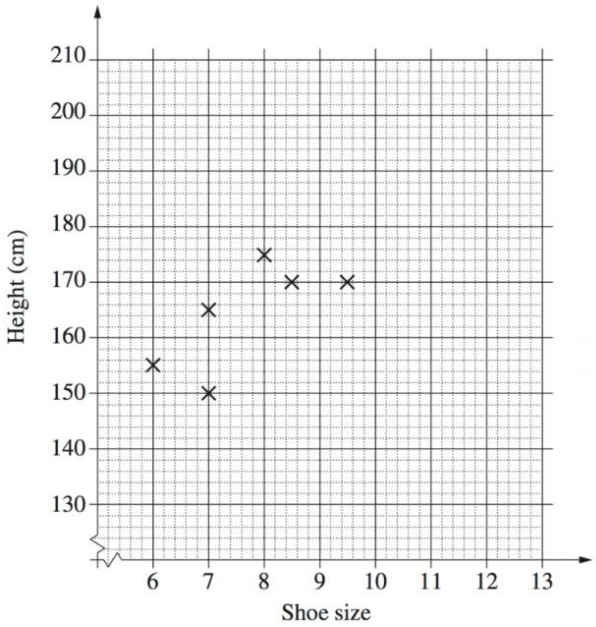
Strong positive

11 2015 HSC Standard 2 Band 5

The shoe size and height of ten students were recorded.

Shoe size	6	7	7	8	8.5	9.5	10	11	12	12
Height	155	150	165	175	170	170	190	185	200	195

Complete the scatter plot and draw a line of fit by eye.



Use the line of fit to estimate the height difference between a student who wears a size 7.5 shoe and one who wears a size 9 shoe.

Around 13 cm

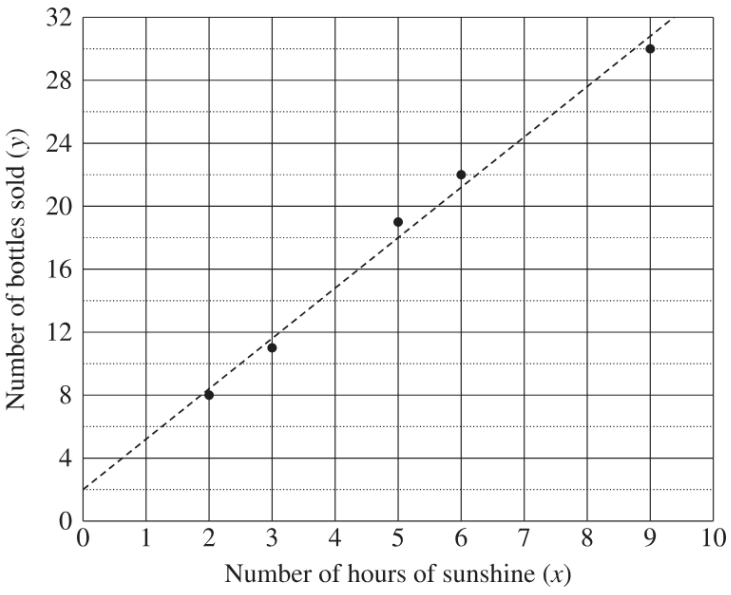
A student calculated the correlation coefficient to be 1 for this set of data. Explain why this cannot be correct.

A correlation coefficient of 1 means all data points would lie perfectly on the line of best fit, which is not the case.

12 2021 HSC Standard 2 Band 5

A salesperson is interested in the relationship between the number of bottles of lemonade sold per day and the number of hours of sunshine on the day.

The diagram shows the dataset used in the investigation and the least-squares regression line.



What is the equation of the least-squares regression line relating to the dataset?

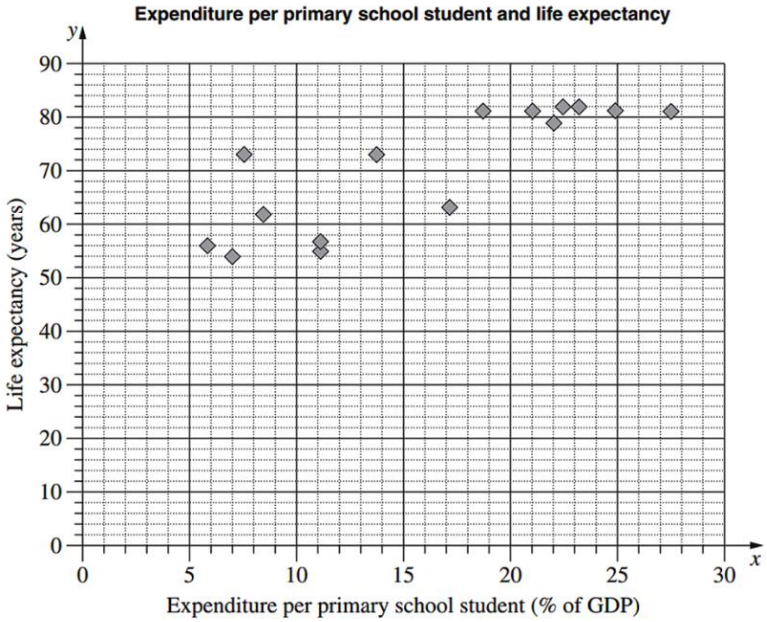
$y = 3.2x + 2$

Suppose a sixth dataset was collected on a day that had 10 hours of sunshine and 45 bottles of lemonade sold. What would happen to the gradient of the line?

The gradient would increase (steepen)

13 2014 HSC Standard 2 Band 6

The scatterplot shows the relationship between expenditure per primary school student, as a percentage of a country's Gross Domestic Product (GDP), and the life expectancy in years for 15 countries.



For the given data, the correlation coefficient, r, is 0.83. What does this indicate about the relationship between expenditure per primary school student and life expectancy for the 15 countries?

There is a strong positive correlation between expenditure per primary student and life expectancy.

For the data representing expenditure per primary school student, $Q_L = 8.4$ and $Q_U = 22.5$. Calculate the interquartile range.

14.1

Another country has an expenditure per primary school student of 47.6% of its GDP.
Would this country be an outlier for this set of data? Justify your answer with calculations.

Yes, the upper limit for being an outlier is 43.65%

On the scatterplot, draw the least-squares line of best fit $y = 1.29x + 49.9$

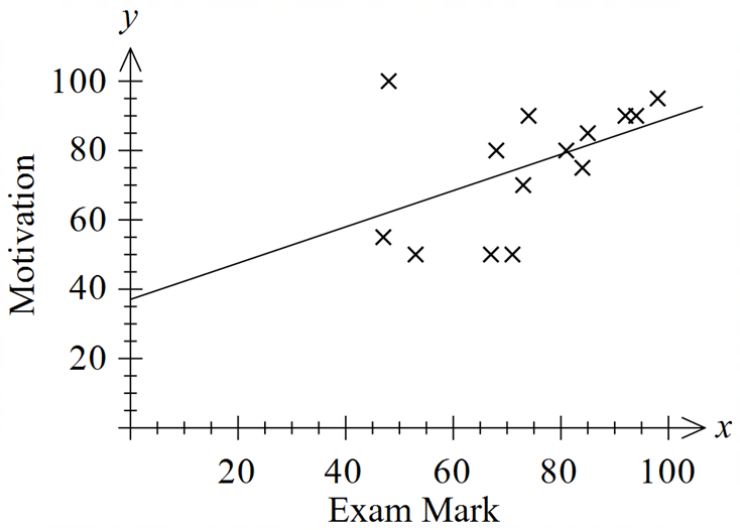
Using this line, estimate the life expectancy in a country which has an expenditure per primary school student of 18% of its GDP.

Around 73 years.

Why is this line NOT useful for predicting life expectancy in a country which has expenditure per primary school student of 60% of its GDP?

60% GDP is far outside the existing range of data.

14 The scatterplot compares examination mark to a student's self-reported motivation score.



a Identify which student is most likely to be considered the outlier. Circle the data point.

(45, 100)

b Pearson's correlation coefficient r is $r = 0.48$.
Describe the strength of the relationship between exam mark and motivation.

Weak positive.

c The equation of the line of best fit is $y = 0.52x + 37.1$.
Predict the motivation score of a student who scored 86.

81.82

d Would the linear regression model be useful for predicting the motivation of a student who scored 20 in the exam?

No. It is far outside the existing range of data.

e With the outlier removed, the new equation of the line of best fit is $y = mx + 4.1$.
Describe how m would change.

m would become more positive as the gradient of the line of best fit would become more positive, steeper.

f Would removing the outlier increase or decrease Pearson's correlation coefficient r?
Give a brief explanation for your answer.

Increase. On average, the datapoints would now be closer to the line of best fit,

BIVARIATE DATA ANALYSIS USING A CALCULATOR

Equation of Line of Best Fit (on a calculator)

$$y = a + bx$$

y-intercept
y value at $x = 0$

Gradient
rate of change;
m per x

CASIO FX-82 AU

1. Set calculator *Statistics Mode 2-var*: MODE 2 2
2. Enter your data into the table using numbers, arrows and equals button.
3. Press AC to save the data.
4. Calculate regression statistics: SHIFT 1 5 Note: regression line is $y = A + Bx$

   r 3 Pearson's r

   A 1 y-intercept

   B 2 Gradient

   Then press =

CASIO FX-8200 AU

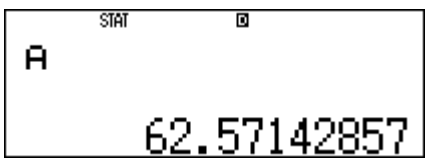
1. Set the calculator to Statistics Mode using $\text{Home} \rightarrow \text{Statistics} \rightarrow 2 - \text{Variable}$
2. Enter your data into the frequency distribution table.
3. Press $\text{EXE} \rightarrow \text{Reg Results}$

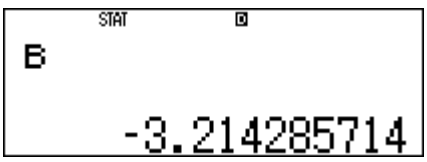
Example Calculate regression statistics using a calculator.

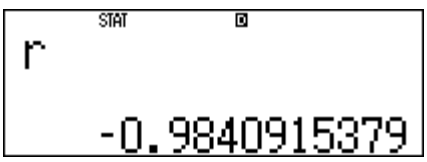
Find the equation of the line of best fit and Pearson's r value for this set of data.
Round values to 2 d.p.

Time (minutes)	2	4	6	8	10	12	14
Weight that can be held (kg)	55	52	46	33	28	25	19

y-intercept: gradient: r:







The equation is $y = 62.57 - 3.21x$ with $r = -0.98$

1 Find Pearson's r and equation of the line of best fit $y = a + bx$ for the tables below.

x	8	14	17	17	22	27	30	33
y	11	14	20	16	22	29	28	35

a Pearson's correlation coefficient r

0.98

b Strength and direction

Strong positive

c Equation of the line of best fit $y = a + bx$

$y = 2.06 + 0.94x$

x	10	14	20	22	28	35	38	43	47
y	9	15	16	13	24	20	29	22	27

a Pearson's correlation coefficient r

0.86

b Strength and direction

Strong positive

c Equation of the line of best fit $y = a + bx$

$y = 6.82 + 0.44x$

2 The time t it takes a number of students to swim 25 m freestyle is recorded in the table.

Age (n)	18	15	22	17	11	26	14	19	15
Time (s)	24	28	25	21	33	30	29	21	26

a Pearson's correlation coefficient r

$= -0.26$

b The equation of the line of best fit $y = a + bx$

$y = 30.35 - 0.23x$

Identify the strength and direction of the association.

Negative weak

Interpret the gradient and y-intercept

Time taken to swim 25 m decreases by 0.23 of a second each year of age

At age 0 years, it would take 30.35 seconds to swim 25 m

3 The results of group of students on Mathematics and Science tests are compared.

Maths test (M)	64	67	69	70	73	74	77	82	84	85
Science test (S)	68	73	68	75	78	73	77	84	86	89

a Find the equation of the regression line.

$S = 8 + 0.93x$

b Predict the score in Science of a student who scores 80 in Mathematics.

82.4

c Predict the score in Mathematics of a student who scores 60 in Science.

55.9

d Is part **b** or **c** an example of extrapolation?

c, as 60 in science is outside the existing range of data.

BIVARIATE DATA ANALYSIS USING A SPREADSHEET

The table below shows the humerus length (elbow to shoulder) compared to height.

Humerus length (cm)	37	35	40	31	35	33	31	40	34	39
Height (cm)	176	174	184	172	173	178	171	189	180	188

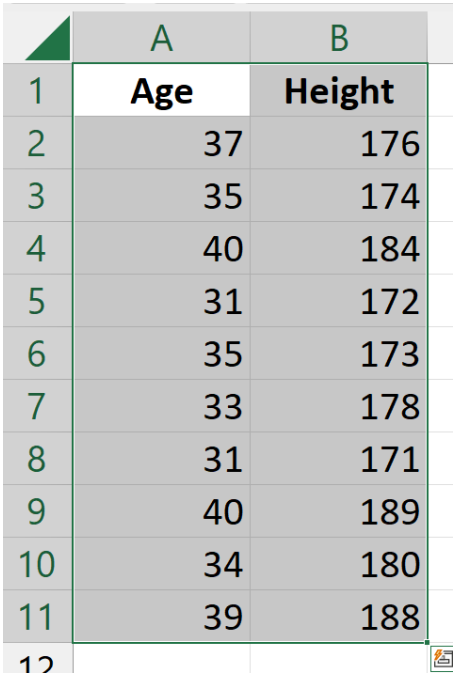
We will use a spreadsheet to analyse this data.

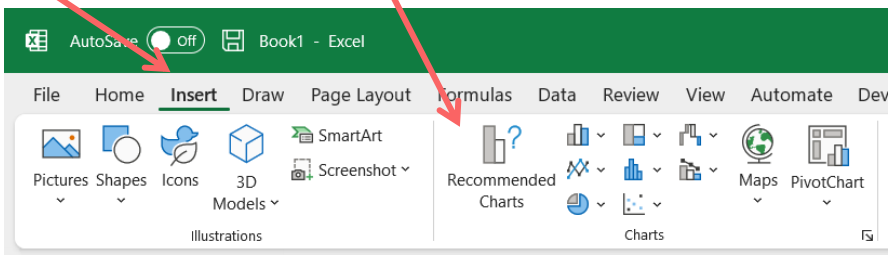
1. Enter data onto a spreadsheet.
   NOTE: It is good practice to have the tables vertical.
   (easier to scroll down than across)

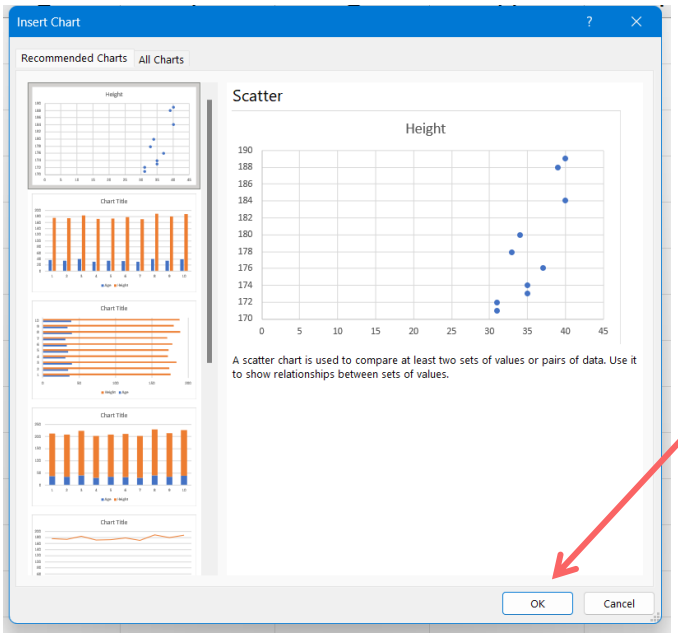
	A	B
1	**Age**	**Height**
2	37	176
3	35	174
4	40	184
5	31	172
6	35	173
7	33	178
8	31	171
9	40	189
10	34	180
11	39	188

GRAPHING THE DATA

2. **Highlight** the entire table, including the headers.
3. Go to the **INSERT** tab and click **recommended charts.**
4. Click **Scatter** and **OK**







CALCULATING PEARSON'S COEFFICIENT

5. In an empty cell, type $= CORREL($ and highlight the first column. add a comma , then highlight the second column.

	A	B
1	**Age**	**Height**
2	37	176
3	35	174
4	40	184
5	31	172
6	35	173
7	33	178
8	31	171
9	40	189
10	34	180
11	39	188
12		
13		
14	=CORREL(A2:A11	
15	CORREL(**array1**, array2)	

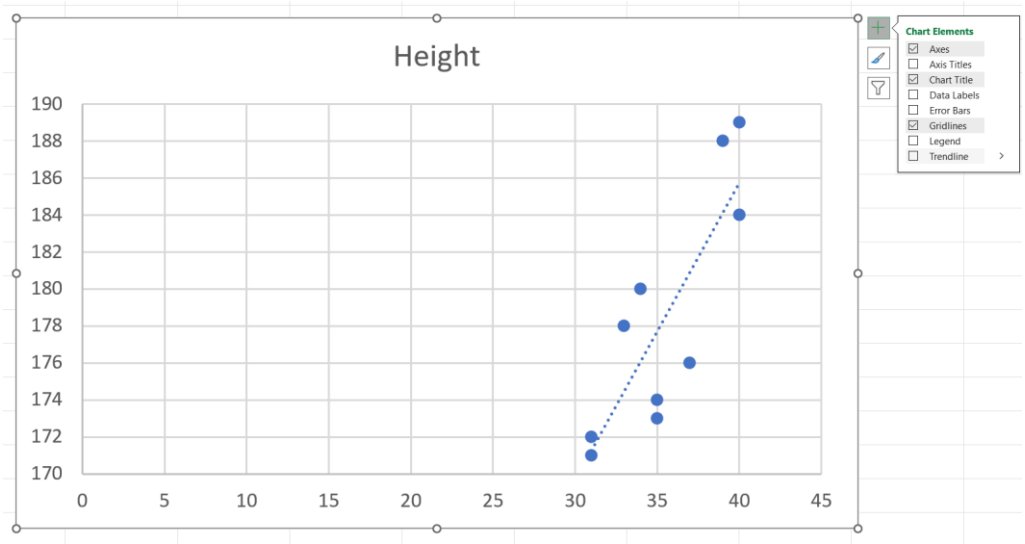
	A	B
1	**Age**	**Height**
2	37	176
3	35	174
4	40	184
5	31	172
6	35	173
7	33	178
8	31	171
9	40	189
10	34	180
11	39	188
12		
13		
14	=CORREL(A2:A11,B2:B11	
15	CORREL(array1, **array2**)	

6. Press ENTER and the spreadsheet will calculate Pearson's R

14	0.83627
15	

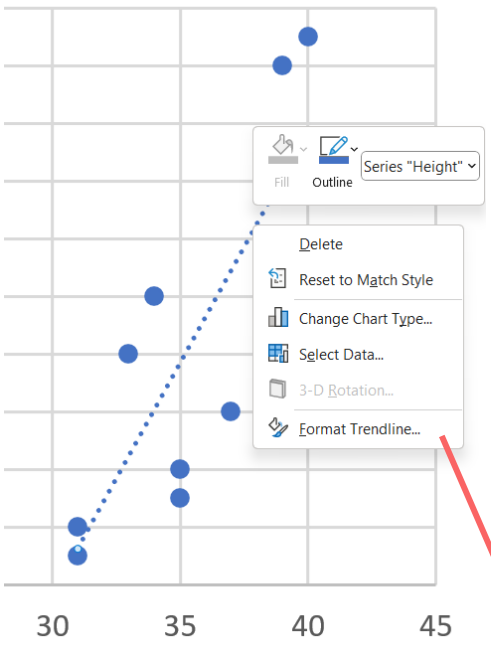
ADDING A LINE OF BEST FIT

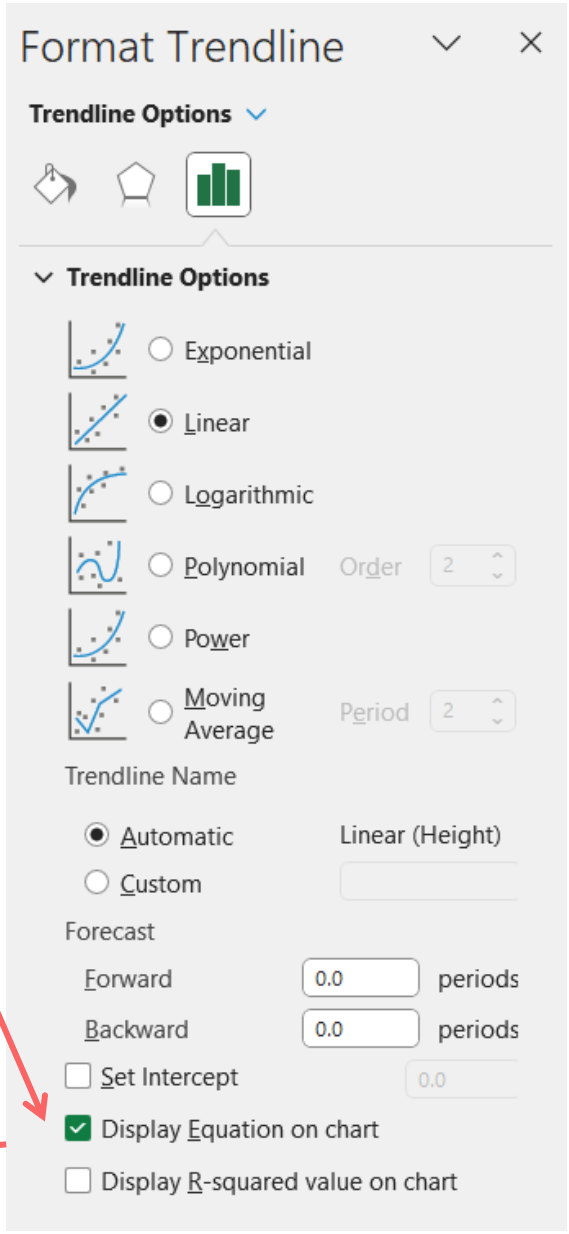
7. Click on your **graph**.
8. Click on the **plus** + **sign** to the top right of your graph.
9. Click on **trendline**. This line is called the **least-squares regression line**. It is a precise line of best fit that is calculated so that the distances of all the points to the line is as small as possible.

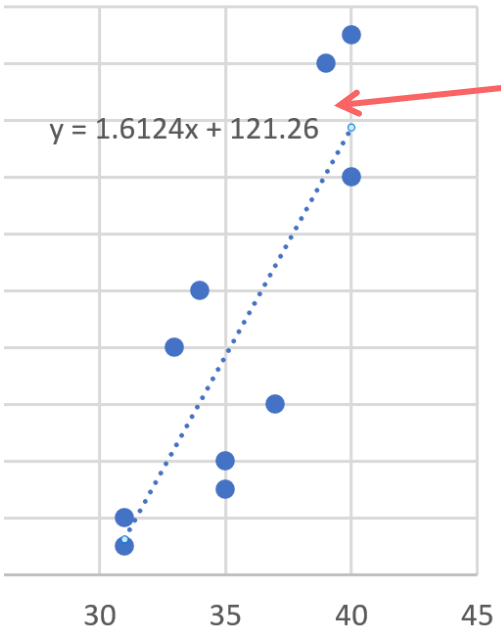


CALCULATING EQUATION OF THE LINE

10. **Right click** on the trendline.
11. Click **format trendline.** A sidebar will pop up.
12. Click **display equation on chart**.
13. You now have the equation of your trendline.







1 The data shows the average height of 17-year-old girls from 1850 to 1990.

Year	1850	1870	1890	1910	1930	1950	1970	1990
Height (cm)	147	149	155	158	160	161	164	170

a Pearson's correlation coefficient r

0.98264

b Strength and direction

Strong positive

c Equation of the line of best fit $y = mx + c$

$y = 0.1524x - 134.57$

d Interpret the gradient and y-intercept

Average height of 17-year-old girls increases by 0.1524 cm each year

At the year 0 AD, the average height of a 17-year-old girl is -134.57 cm

2 The data shows the time students spent studying and their final mark on their exam.

Time studying (h)	7	9	14	12	8	3	11	15	5	9
Mark	64	67	80	92	73	88	77	89	59	75

a Pearson's correlation coefficient r

0.44715

b Strength and direction

Moderate positive

c Equation of the line of best fit $y = mx + c$

$y = 1.3051x + 64.262$

d Interpret the gradient and y-intercept

Mark increases by 1.3051 for each hour spent studying

If you spend 0 hours studying, you will get a mark of 64.262

3 The data shows the population of Australia between 1931 and 2011.

Year	1931	1941	1951	1961	1971	1981	1991	2001	2011
Population (millions)	6.53	7.11	8.42	10.55	13.07	14.92	17.28	19.27	22.34

a Pearson's correlation coefficient r

0.9918

b Strength and direction

Strong positive

c Equation of the line of best fit $y = mx + c$

$y = 0.203x - 386.87$

d Interpret the gradient and y-intercept

The population increases by 0.203 million each year

The population at 0 AD would be -386.87 million

4 May is investigating whether there is a link between the weight and systolic blood pressure of several patients at the hospital she works at. She collects the following information.

Weight (kg)	75	83	62	66	98	57	88	111	79	60
Blood pressure (mmHg)	99	115	108	86	127	90	122	143	103	105

a Plot this data as a scatterplot on a spreadsheet.

b Find the equation of the regression line.

$y = 0.8712x + 41.93$

c Use the equation of the regression line to predict the systolic blood pressure of patients with the following weights. Round to the nearest whole.

i 115 kg

142 mmHg

ii 140 kg

164 mmHg

d Which of the predictions from part **c** is more reliable? Explain your answer.

115 kg is only slightly outside the range of existing data so is more reliable.

5 June is looking to see whether there is a relationship between the number of goals and behinds in AFL football games. The table below is a sample of the data she collects.

Goals (g)	14	5	13	15	12	18	17	7	11	25
Behinds (b)	8	15	13	7	8	9	16	9	12	13

a Plot this data on a scatterplot, with g on the horizontal axis and b on the vertical axis.

b Calculate the value of Pearson's correlation coefficient for the data.

0.02448

c Considering your answers to **b**, will the regression line be useful in making accurate predictions for this dataset? Explain your answer.

No. $r = 0.02448$ means there is no meaningful relationship between the two variables.

Investigation Height and armspan.

Measure the height and armspan of the students in your class
(separate results for males and females) and record the information in a table
(or directly into a spreadsheet).

a Draw a scatterplot of the data using a spreadsheet.

b Calculate the value of r. Describe the correlation between these two variables in terms of direction (positive/negative) and strength (perfect/strong/moderate/weak).

……

……

c Is one variable a good predictor of the other? ……

d Use technology to add the least-squares regression line of best fit to the data.

e Interpret the gradient and y intercept of your trendline.

……

……

f Mr Ding has a height of 181 cm. Predict his armspan.

……

g Compare your prediction to his actual handspan.
Is his handspan shorter or longer than expected?

……

h Comment on the validity of your model.

……

……

A familiar folk tale is that storks deliver babies, used by parents since ancient times to avoid answering the awkward question “where do babies from from?”

We will investigate whether there is a correlation between storks and birth rate.

Stork population data is available for 17 European countries, shown in the table below.

We will compare this to human birth rate.

a Transcribe the data to Excel and create a scatterplot of Storks (x) vs Birthrate (y).

b Calculate Pearson’s correlation coefficient between Storks and Birthrate

……

0.29876

c Comment on the strength and direction of association.

……

Moderate positive

Country	Storks (breeding pairs)	Human birthrate (10^3/yr)
Albania	100	83
Austria	300	87
Belgium	200	118
Bulgaria	5000	117
Denmark	5	59
France	140	374
Germany	3300	301
Greece	2500	106
Holland	200	188
Hungary	5000	124
Italy	150	551
Poland	30 000	610
Portugal	1500	120
Romania	5000	367
Spain	8000	439
Switzerland	150	82
Turkey	25 000	610

From this, you might say:

“The data shows that it’s there is a relationship between storks and babies, maybe storks actually deliver babies!”

We will try to disprove this causal relationship.

A **confounding** variable is a third variable that influences the two variables of interest.

The area of each country is shown.

Create a scatterplot and calculate Pearson's r of area vs storks and area vs birthrate.

Country	Area (km^2)
Albania	28 750
Austria	83 860
Belgium	30 520
Bulgaria	111 000
Denmark	43 100
France	544 000
Germany	357 000
Greece	132 000
Holland	41 900
Hungary	93 000
Italy	301 280
Poland	312 680
Portugal	92 390
Romania	237 500
Spain	504 750
Switzerland	41 290
Turkey	779 450

d Area vs Storks r: ……

0.40201

e Area vs Birthrate r: ……

0.82742

f Complete the sentences

This study demonstrates that …… does not necessarily imply ……

correlation, causation

While there appeared to be a correlation between stork populations and birth rates, further investigation revealed a …… factor: the …… of each country.

confounding, area

Statistics, Storks, and Babies – Numberphile https://www.youtube.com/watch?v=-n-d1ApBTFw

In the late 1800s, researchers noticed something odd: places with more storks also had higher birth rates.

This wasn't because storks were delivering babies. The ***confounding variable*** was area; countries with larger area have more rural areas – there are more storks in rural areas and households prefer large families in rural places. This is a classic example of a "spurious correlation" - when two things appear related but have no real cause-and-effect link.

This stork-baby story teaches a crucial lesson about interpreting data: correlation doesn't always mean causation. Just because two things happen together doesn't mean one causes the other.

A modern example many people believe is that violent video games cause aggressive behaviour. While there's a correlation between playing these games and aggression, it doesn't prove causation. Other factors, like a person's temperament or home environment, might explain both the game choice and the behaviour.

In the 1950s, the correlation-causation confusion took a darker turn. When scientists first linked smoking to lung cancer, tobacco companies used the stork-baby logic to defend their product. They argued, "Just as storks don't really bring babies, cigarettes might not cause cancer. It could be another spurious correlation."

This tactic muddied the waters of public understanding for years, with deadly consequences. Millions continued smoking, unaware of the real risks, because of this misuse of statistical concepts.

But how do we know for sure that smoking causes lung cancer?

Scientists used multiple approaches:

- Large-scale studies showed that smokers had much higher rates of lung cancer than non-smokers.
- Laboratory experiments demonstrated that cigarette smoke damages DNA and causes mutations in lung cells.
- When people quit smoking, their lung cancer risk decreased over time.
- Animals exposed to cigarette smoke developed lung cancer.
- A clear *dose-response relationship* exists: more smoking leads to higher cancer risk.

This combination of evidence from different angles establishes a causal relationship.

Whether in science, policy-making, or everyday life, understanding the difference between correlation and causation is crucial for making informed decisions and spotting misleading claims.

https://www.tylervigen.com/spurious-correlations