

QQQ – Statistics Year 1 – Chapter 4 – Correlation (Version 2)

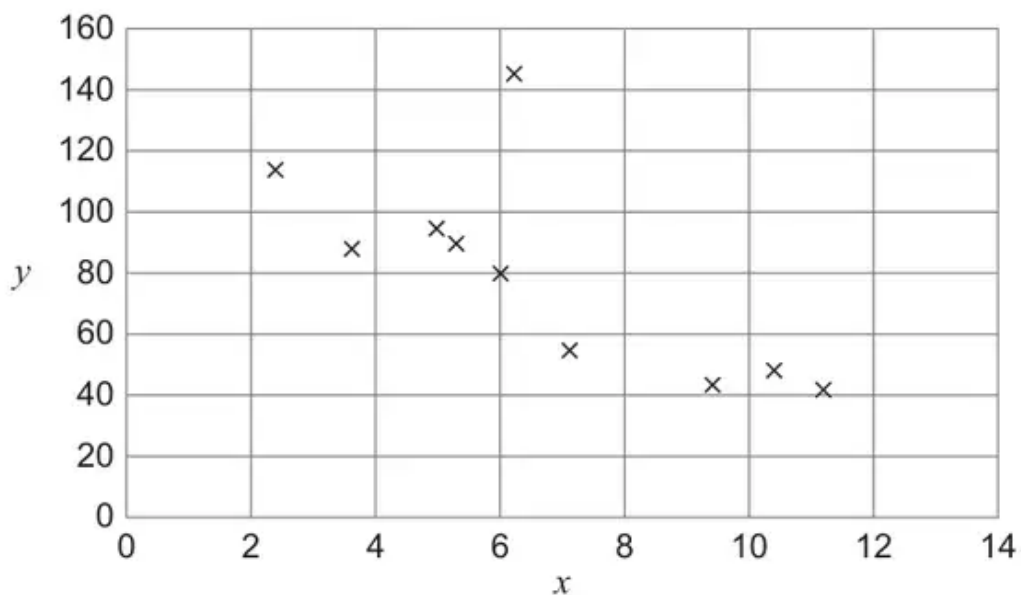
Total Marks: 22

(22 = Platinum, 20 = Gold, 18 = Silver, 15 = Bronze)

1. The table below shows the daily salt intake, x grams, and the daily Vitamin C intake, y milligrams, for a group of 10 adults.

Adult	A	B	C	D	E	F	G	H	I	J
x	5.3	6.2	3.6	10.4	2.4	9.4	6	5	7.1	11.2
y	90	145	88	48	114	44	80	95	55	41

A scatter diagram of the data is shown below.



One of the adults is an outlier. Identify the letter of the adult that is the outlier.

Circle your answer below.

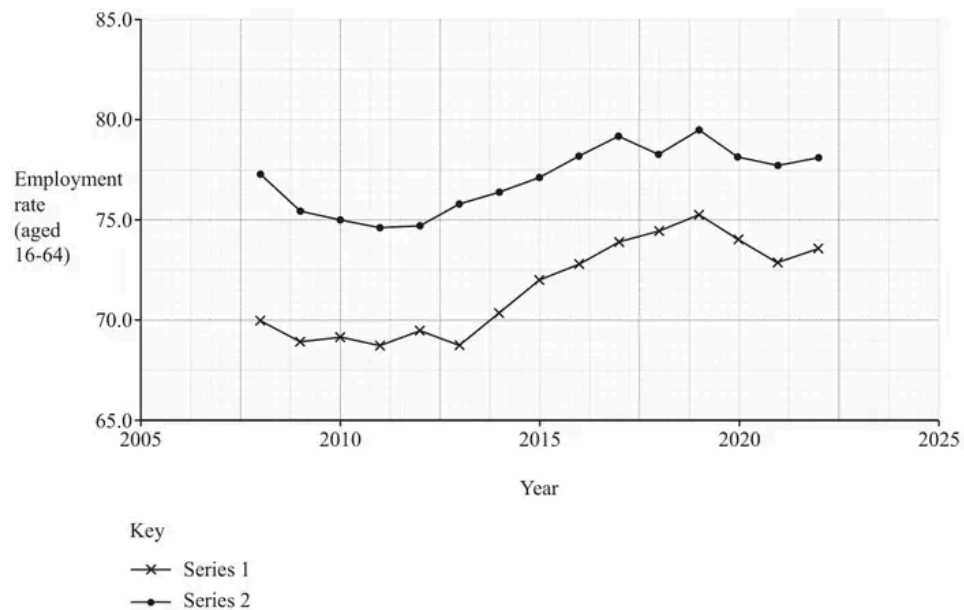
[1 mark]

- A
- B
- E
- J

(Total 1 marks)

2. The pre-release material contains data for percentage employment rates in different areas of England.

The diagram shows percentage employment rates between 2008 and 2022 for the North West of England and the South East of England.



- (a)** Use your knowledge of the pre-release material to explain whether Series 1 (the lower time series) or Series 2 (the upper time series) represents the data for the South East of England. [1]

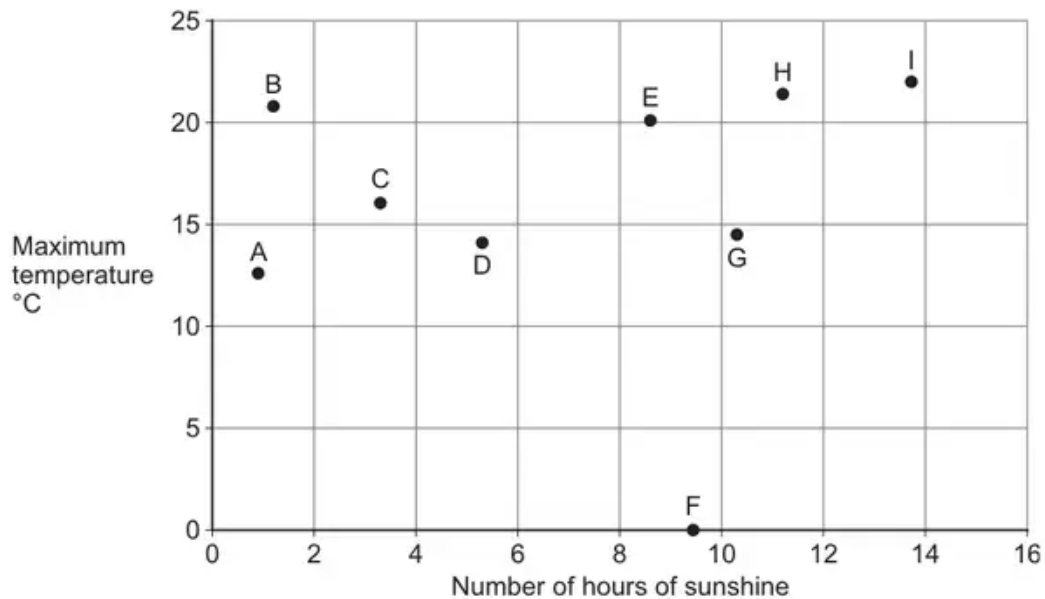
A student uses the LINEST function to model the data for Series 2.

- (b)** Assuming that a straight line model is reasonable, use your knowledge of the pre-release material to explain whether it would be appropriate to use this model to calculate an estimate of the percentage employment rate in Cornwall, in the South West of England, in 2011. [1]

(Total 2 marks)

3. The number of hours of sunshine and the daily maximum temperature were recorded over a 9-day period in June at an English seaside town.

A scatter diagram representing the recorded data is shown below.



One of the points on the scatter diagram is an error.

- (a) (i) Write down the letter that identifies this point. [1 mark]
- (ii) Suggest one possible action that could be taken to deal with this error. [1 mark]
- (b) It is claimed that the scatter diagram proves that longer hours of sunshine cause higher maximum daily temperatures.

Comment on the validity of this claim.

[1 mark]

(Total 3 marks)

4. The pre-release material contains information for countries in the world concerning real GDP per capita in US\$ and mobile phone subscribers per 100 population. In an investigation into the relationship between these two variables, a student takes a sample of 20 countries in Africa. The student draws a scatter diagram for the data, which is shown in **Fig. 5.1**.

Africa 1st sample

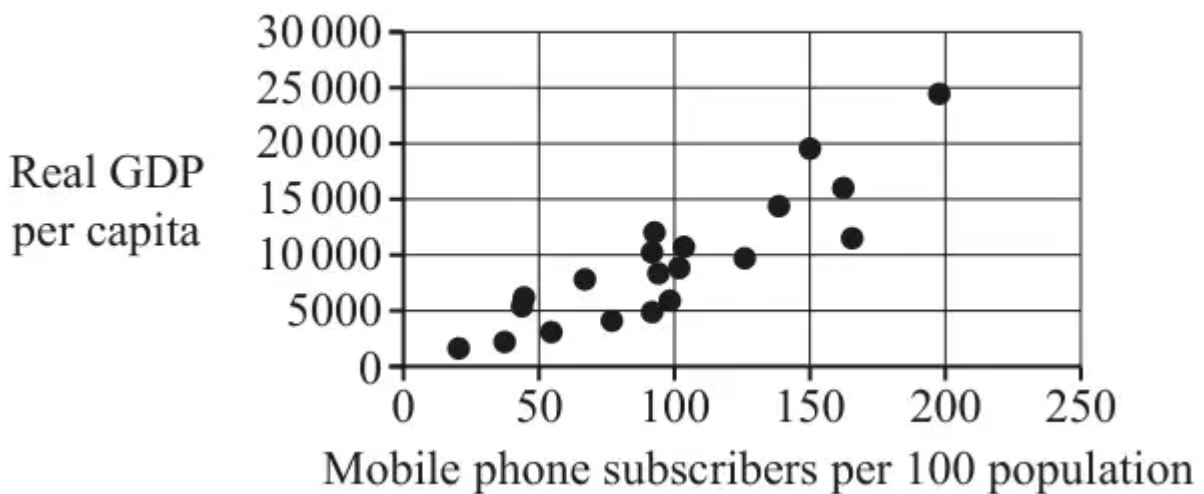


Fig. 5.1

- (a) What does **Fig. 5.1** suggest about the relationship between real GDP per capita and the number of mobile phone subscribers per 100 population? [1]

Another student collects a **different** sample of 20 countries from Africa, and draws a scatter diagram for the data, which is shown in **Fig. 5.2**.

Africa 2nd sample

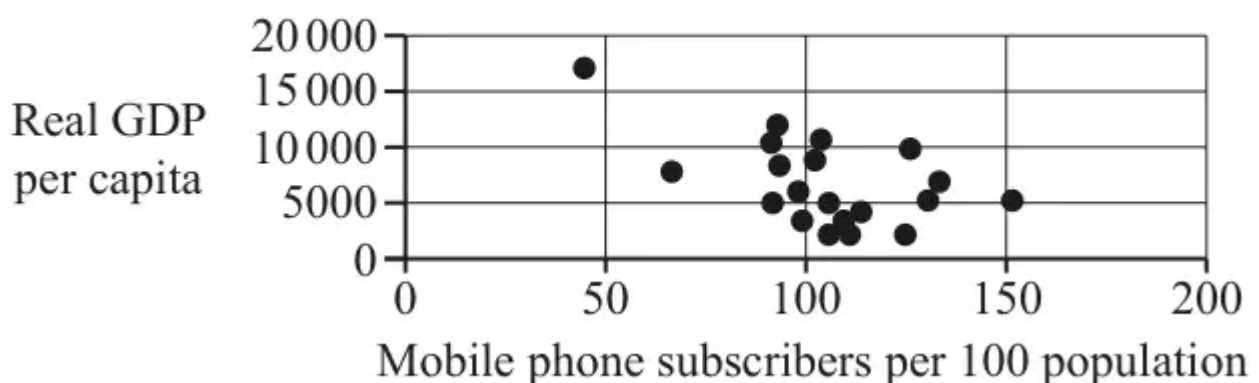


Fig. 5.2

- (b) What does **Fig. 5.2** suggest about the relationship between real GDP per capita and the number of mobile phone subscribers per 100 population? [1]
- (c) Explain whether either of the two scatter diagrams is likely to be representative of the true relationship between real GDP per capita and the number of mobile phone subscribers per

100 population, for countries in Africa.

[1]

(Total 3 marks)

OCR MEI June 2024 Paper 2 Q5

5. This question deals with information about the populations of Local Authorities (LAs) in the North of England, taken from the 2011 census.

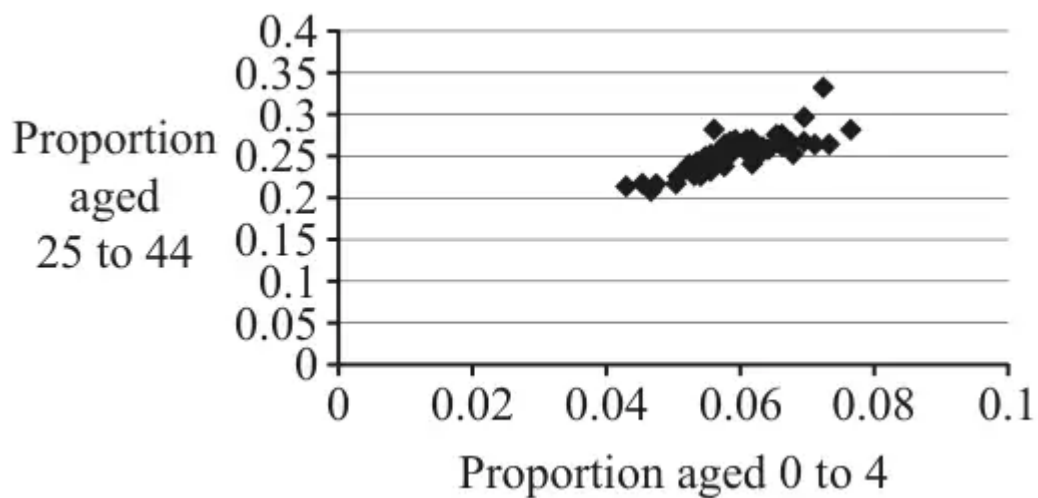


Fig. 1

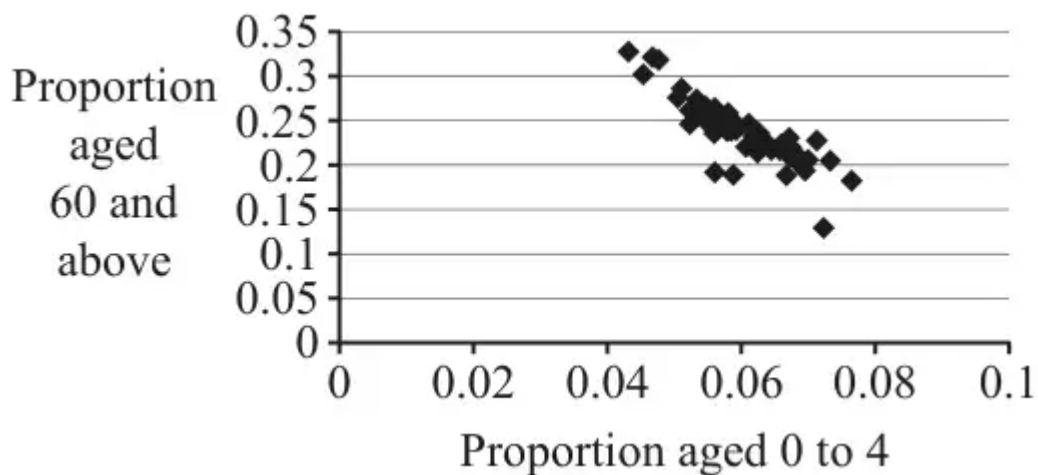


Fig. 2

Fig. 1 and **Fig. 2** both show strong correlation, but of two different kinds.

- (a) For each diagram, use a single word to describe the kind of correlation shown. [1]
- (b) For each diagram, suggest a reason, in context, why the correlation is of the particular kind described in part (a). [2]

Fig. 3 is the same as **Fig. 2** but with the point *A* marked.

Fig. 4 shows information about the same LAs as **Fig. 2** and **Fig. 3**.

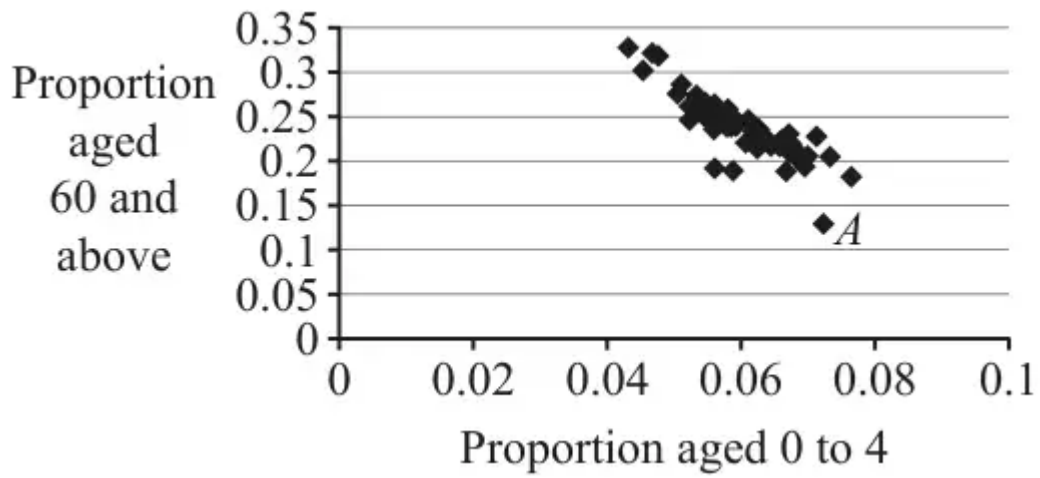


Fig. 3

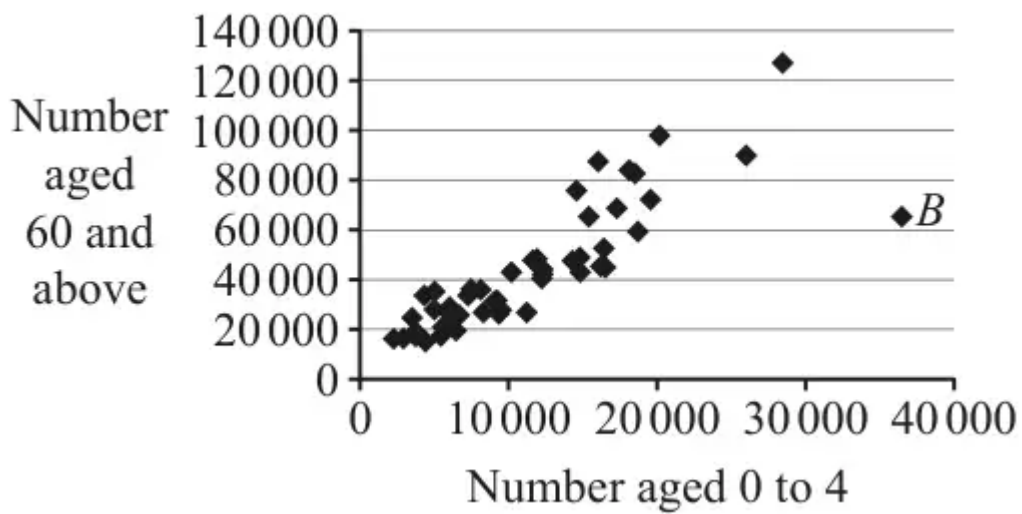


Fig. 4

- (c) Point *A* in Fig. 3 and point *B* in Fig. 4 represent the same LA.

Explain how you can tell that this LA has a large population.

[1]

(Total 4 marks)

6. The pre-release material contains information on Pulse Rate and Body Mass Index (BMI). A student is investigating whether there is a relationship between pulse rate and BMI. A **section** of the available data is shown in the table.

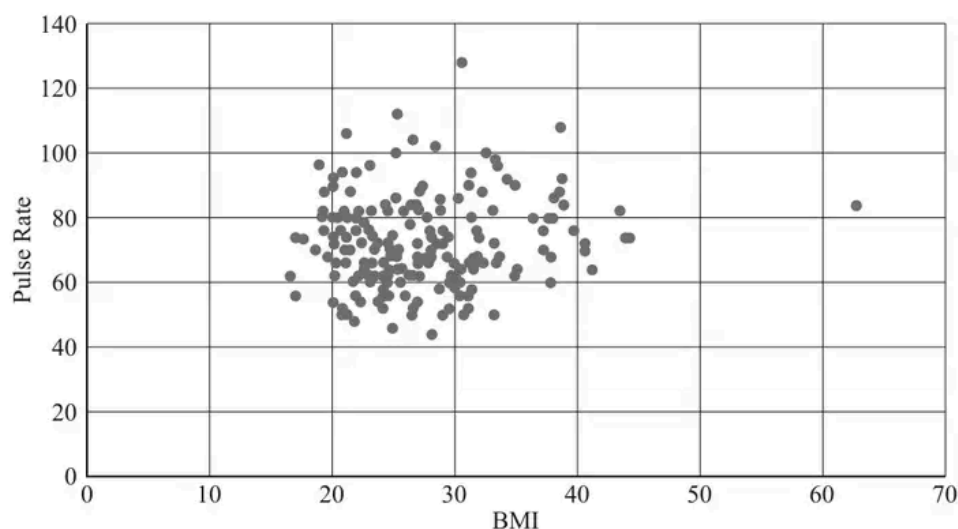
Sex	Age	BMI	Pulse
Male	62	29.54	60
Female	20	23.68	#N/A
Male	17	26.97	72
Male	35	24.7	64
Male	17	20.09	54
Male	85	23.86	54
Female	81	24.04	#N/A

The student decides to draw a scatter diagram.

- (a) With reference to the table, explain which data should be cleaned before any analysis takes place. [1]

The student cleans the data for BMI and Pulse Rate in the pre-release material and draws a scatter diagram.

Scatter diagram of Pulse Rate against BMI



The student identifies **one** outlier.

- (b) On the copy of the scatter diagram in the Printed Answer Booklet, circle this outlier. [1]

The student decides to remove this outlier from the data. They then use the LINEST function in the spreadsheet to obtain the following formula for the line of best fit.

$$P = 0.29Q + 64.2,$$

where P = Pulse Rate and Q = BMI.

They use this to estimate the Pulse Rate of a person with BMI 23.68.

They obtain a value of 71 correct to the nearest whole number.

- (c) With reference to the scatter diagram, explain whether it is appropriate to use the formula for the line of best fit. [1]

It is suggested that all pairs of values where the pulse rate is above 100 should also be cleaned from the data, as they must be incorrect.

- (d) Use your knowledge of the pre-release material to explain whether or not all pairs of values with a pulse rate of more than 100 should be cleaned from the data. [1]

(Total 4 marks)

7. Giovanni believes there is a relationship between the mass of a car, m kg, and its fuel consumption, c miles per gallon (mpg).

For a sample of 25 cars, he obtains the following summary statistic.

$$\sum (c - \bar{c})^2 = 394$$

(a) Find the standard deviation of the fuel consumption of the 25 cars. (1)

Using m as the explanatory variable, Giovanni creates the linear regression model

$$c = a + bm$$

where a and b are constants.

Using his model, he concludes that, on average

- the fuel consumption is 3.5 mpg lower for each additional 500 kg of mass
- the fuel consumption of a car with a mass of 1700 kg is 20 mpg

(b) (i) Find the value of b (2)

(ii) Find the value of a (2)

(Total 5 marks)

Solutions

1.

Scheme	Marks	AO
Circles second answer	B1	2.2a
	(1 mark)	

Typical solution

B

2.

Mark scheme (a)

Scheme	Marks	AO
Series 2 represents SE England because employment rates are generally higher in this region (than in the North West)	B1	2.4
	[1]	

Notes

Accept: Series 2 represents SE England because employment rates are generally lower in the North West

Mark scheme (b)

Scheme	Marks	AO
Not appropriate because (the pattern of) employment rates are different in different areas of England	B1	3.5a
	[1]	

3.

Mark scheme (a)

Scheme	Marks	AO
(i) Deduces the correct letter	B1	2.2a
	(1)	

(ii) States an appropriate action to deal with the error	E1	2.4
	(1)	

Typical solution

(i)

F

(ii)

Remove the point from the set of data

Mark scheme (b)

Scheme	Marks	AO
Indicates somewhere in their response that the scatter diagram does not imply causality (ISW) Or Identifies any counter example from the data and concludes	E1	2.4
	(1)	
	(3 marks)	

Typical solution

The claim is invalid because correlation does not imply causality.

4.

Mark scheme (a)

Scheme	Marks	AO
positive association or positive correlation	B1	2.2b
	[1]	

Notes

May be a contextual interpretation e.g. 'as the mobile phone subscribers per 100 population increases, real GDP per capita also increases.'

Ignore references to 'strong/weak' etc. 'Linear' not required. Contradictory statements lose the mark e.g. 'positive correlation' followed by incorrect interpretation.

Mark scheme (b)

Scheme	Marks	AO
negative association or negative correlation	B1	2.2b
	[1]	

Notes

Accept 'as the mobile phone subscribers per 100 population increases, real GDP per capita decreases' o.e.

Ignore references to 'strong/weak' etc. 'Linear' not required. Contradictory statements lose the mark e.g. 'negative correlation' followed by incorrect interpretation.

Mark scheme (c)

Scheme	Marks	AO
Neither/None of them – samples are same size and different samples can lead to different conclusions.	B1	2.4
Or		
'Neither as the patterns of the data contrast/are different'.		
	[1]	

Notes

Allow e.g. neither since combining results suggests no correlation.

SC: IF they interpret (b) as 'little/no correlation' then they can score for a correct FT reason e.g. superimposing (a) onto (b) implies no real relationship – so 2nd diagram more likely to represent the true situation.

SC: From pre-release material/large data set accept 'the first scatter diagram is more representative of the true relationship as the PMCC is positive for this data when all countries in Africa are considered'

Any reference to sample size being too small is B0 as comparing these two scatter diagrams of equal size. e.g.

'Neither as sample size is too small' is B0

'Neither as total in both samples $\neq$ total in Africa' is B0

'Neither as the patterns of the data contrast/are different' is B1

Mark scheme (a)

Scheme	Marks	AO
Fig 1: Positive. Fig 2: Negative	B1	1.2
	[1]	

Notes

Ignore all else

Mark scheme (b)

Scheme	Marks	AO
Fig. 1: 25-44s have children so proportion of 0-4 is directly related to proportion of 25-44.	B1	2.4
Fig. 2: High proportion of older people means lower proportion of families with young children	B1	2.4
	[2]	

Notes

B1 (Fig. 1): oe e.g. “25-44s more likely to be parents”

Must be in context so do not accept generic descriptions of the correlation.

B1 (Fig. 2): Oe e.g. “>60s less likely to live with young children”

Must be in context so do not accept generic descriptions of the correlation.

In both parts:

- Need not refer to proportion but may not refer to amount (i.e. do not accept answers including references to ‘amount’ or ‘number’)
- SC Max [1/2] for correct description in context but referencing ‘amount’ or ‘number’ or ‘population’
- Ignore irrelevant statements

Condone ‘children’ as shorthand for the 0-4 age group in this question (and ‘elderly’ as shorthand for the >60 age group)

Mark scheme (c)

Scheme	Marks	AO
This LA has a similar number of >60s to many other LAs, but these represent by far the lowest proportion of its population.	B1	2.4

OR		
This LA has a large number of 0-4s, but its proportion of 0-4s is similar to several other LAs. (Hence must have large number in total.)		
	[1]	

Notes

oe. Must make a correct connection between number and proportion and compare to other LAs to show that this is one of the largest

Do not accept statements referring to 25-44 (from Fig. 1)

6.

Mark scheme (a)

Scheme	Marks	AO
Remove any data where #N/A is in the column, as there is no data available	B1	2.4
	[1]	

Notes

LDS advantage

Comments such as exclude 20/81 year old female are B0 as this is only an extract of the full set of data

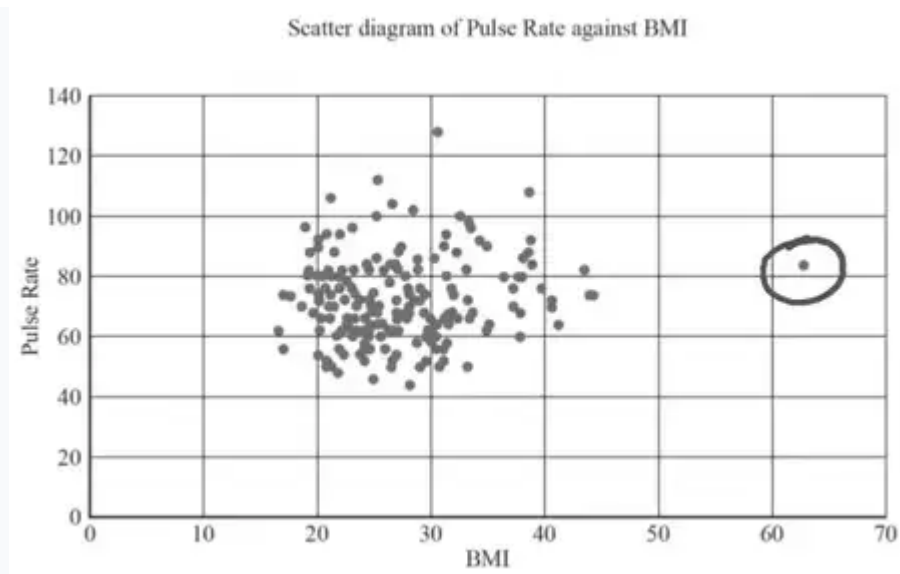
Must refer to N/A/missing data but accept 'not available/non-existent'

Remove data without a pulse reading is B0 as we could have missing BMI data too

Mark scheme (b)

Scheme	Marks	AO
(62.77, 84) ringed <u>and no others</u>	B1	1.1
	[1]	

Notes



Mark scheme (c)

Scheme	Marks	AO
No evidence of a linear relationship oe so unlikely to be reliable	B1	2.2b
	[1]	

Notes

Any comment relating to interpolation or extrapolation is B0 as we want a comment on the appropriateness of the model considering the scatter diagram

Condone 'there appears to be little correlation/weak positive correlation' etc

'No/Zero correlation' is B0 e.g the PMCC could be 0.1 etc

Mark scheme (d)

Scheme	Marks	AO
None of the pulse rates are that unusual so should not be removed	B1	2.2b
	[1]	

Notes

LDS advantage

Need 'no/keep' and reason

Condone 'No as they are not outliers'

Accept 'Higher pulse rates are not uncommon'

7.

Mark scheme (a)

Scheme	Marks	AO
$\sigma = \left[\sqrt{\frac{S_{cc}}{n}} = \sqrt{\frac{394}{25}} \right] = 3.9698 \dots \text{ awrt } \mathbf{3.97}$	B1	1.1b
	(1)	

Notes

B1: awrt 3.97 (allow $s =$ awrt 4.05)

Mark scheme (b)(i)

Scheme	Marks	AO
$b = \frac{-3.5}{500} \text{ or } 16.5 = a + 2200b$	M1	2.2a
$= \underline{-0.007}$	A1	1.1b
	(2)	

Notes

Mark (b)(i) and (b)(ii) together

M1: deducing that the first bullet point can be used to:

Either set up a correct numerical expression for b (allow $\pm$) implied by ± 0.007 or $\pm \frac{3.5}{500}$

or write a correct equation from the first bullet point e.g. $16.5 = a + 2200b$

Allow equivalent use of the first bullet point e.g. $(20 - 3.5k) = a + (1700 + 500k)b$

NB: $(\pm)3.5 = a + 500b$ is M0

A1: -0.007 oe (allow standard form) correct answer scores M1A1.

Mark scheme (b)(ii)

Scheme	Marks	AO
$20 = a + \text{“} -0.007 \text{”} (1700) \text{ or } 20 = a + 1700b$	M1	2.2a
$= \underline{31.9}$	A1	1.1b
	(2)	

	(5 marks)	
--	-----------	--

Notes

M1: using the second bullet point with their value of b to find a

Either substituting “their b ” into $20 = a + 1700b$ oe e.g. $20 - \text{“their } b\text{”} \times 1700$

or writing the correct equation $20 = a + 1700b$ oe

A1: 31.9 oe correct answer scores M1A1.