MATHSAURUS

AS Mathematics Exam Questions by Topic
Chapter 15c: Scatter diagrams, correlation and regression

These questions are taken from the Specimen Exam materials and the real 2018 papers for the new syllabus AS and A-level mathematics courses and arranged by chapter of the textbooks by Goldie et al (available here: `https://amzn.to/39umfr5` and `https://amzn.to/3hE8kBL`). There are a mixture of questions from OCR A, OCR B (MEI), Edexcel and AQA. Although the style of questions varies a little across the exam boards the content of the syllabus is almost identical so these are suitable for students preparing for any exam board.

Free problem sets for all other chapters, as well as video solutions, full past papers and other content for GCSE and A-level maths can be found at:

`https://mathsaurus.com/`

Edexcel AS 2018 Paper 2 Question 1:

1. A company is introducing a job evaluation scheme. Points (x) will be awarded to each job based on the qualifications and skills needed and the level of responsibility. Pay (£y) will then be allocated to each job according to the number of points awarded.

Before the scheme is introduced, a random sample of 8 employees was taken and the linear regression equation of pay on points was $y = 4.5x - 47$

(a) Describe the correlation between points and pay.

(1)

(b) Give an interpretation of the gradient of this regression line.

(1)

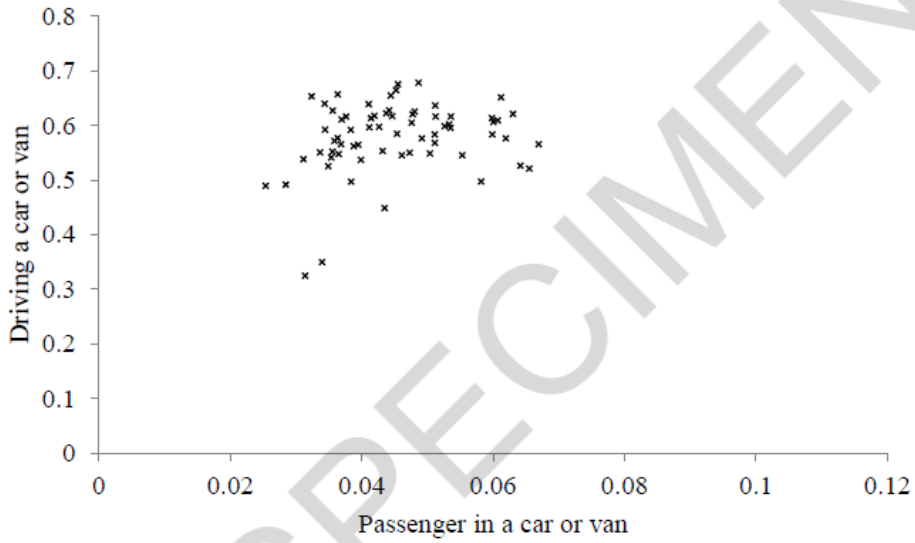
(c) Explain why this model might not be appropriate for all jobs in the company.

(1)

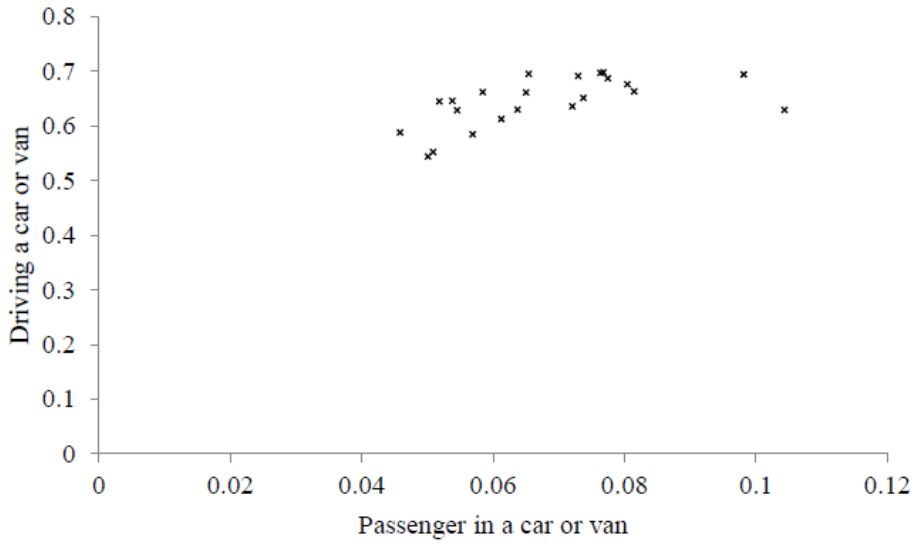
13 The table and the four scatter diagrams below show data taken from the 2011 UK census for four regions. On the scatter diagrams the names have been replaced by letters.
The table shows, for each region, the mean and standard deviation of the proportion of workers in each Local Authority who travel to work by *driving* a car or van and the proportion of workers in each Local Authority who travel to work as a *passenger* in a car or van.
Each scatter diagram shows, for each of the Local Authorities in a particular region, the proportion of workers who travel to work by *driving* a car or van and the proportion of workers who travel to work as a *passenger* in a car or van.

	Driving a car or van		Passenger in a car or van	
	Mean	Standard deviation	Mean	Standard deviation
London	0.257	0.133	0.017	0.008
South East	0.578	0.064	0.045	0.010
South West	0.580	0.084	0.049	0.007
Wales	0.644	0.045	0.068	0.015

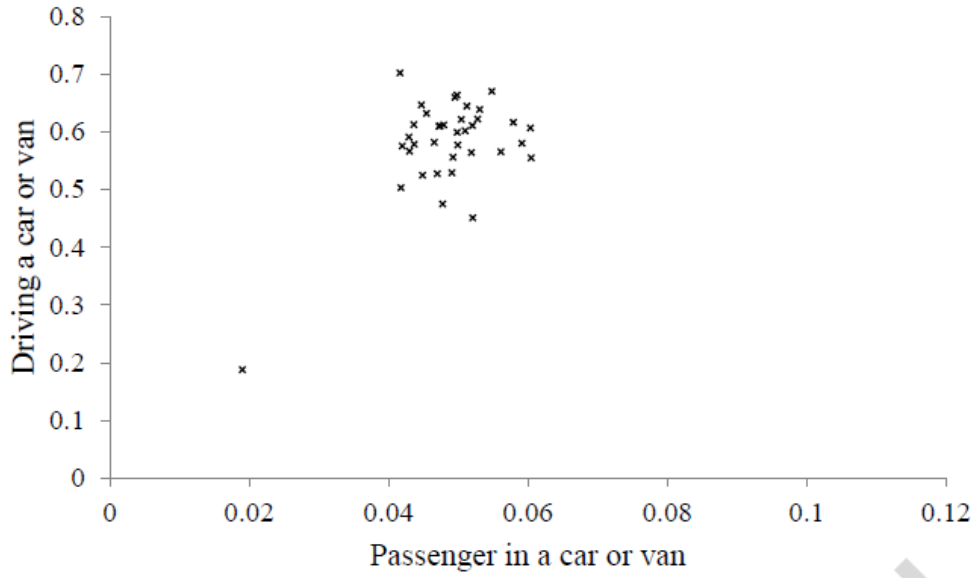
Region A



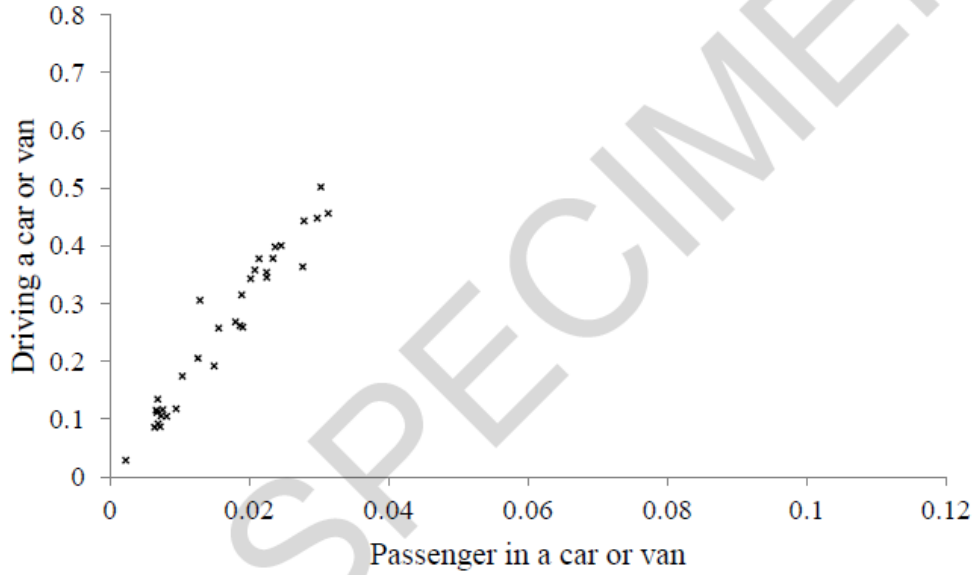
Region B



Region C



Region D



(i) Using the values given in the table, match each region to its corresponding scatter diagram, explaining your reasoning. **[3]**

(ii) Steven claims that the outlier in the scatter diagram for Region C consists of a group of small islands. Explain whether or not the data given above support his claim. **[1]**

(iii) One of the Local Authorities in Region B consists of a single large island. Explain whether or not you would expect this Local Authority to appear as an outlier in the scatter diagram for Region B. **[1]**

OCR B MEI AS 2018 Paper 2 Question 11:

11 The pre-release material contains data concerning the death rate per thousand people and the birth rate per thousand people in all the countries of the world. The diagram in Fig. 11.1 was generated using a spreadsheet and summarises the birth rates for all the countries in Africa.

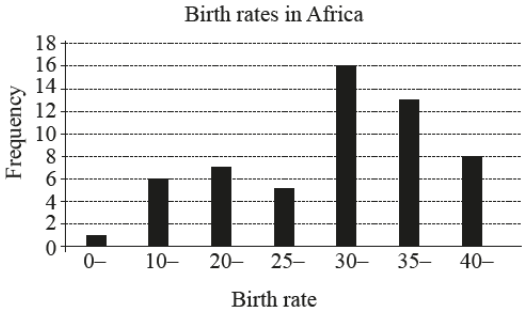


Fig. 11.1

(i) Identify **two** respects in which the presentation of the data is incorrect. **[2]**

Fig. 11.2 shows a scatter diagram of death rate, y, against birth rate, x, for a sample of 55 countries, all of which are in Africa. A line of best fit has also been drawn.

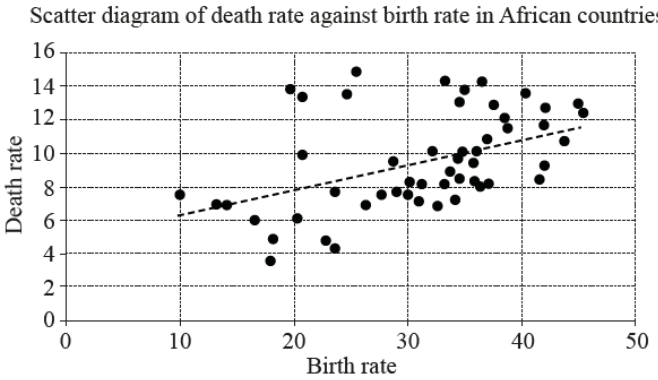


Fig. 11.2

The equation of the line of best fit is $y = 0.15x + 4.72$.

(ii) (*A*) What does the diagram suggest about the relationship between death rate and birth rate? **[1]**

(*B*) The birth rate in Togo is recorded as 34.13 per thousand, but the data on death rate has been lost. Use the equation of the line of best fit to estimate the death rate in Togo. **[1]**

(*C*) Explain why it would not be sensible to use the equation of the line of best fit to estimate the death rate in a country where the birth rate is 5.5 per thousand. **[1]**

(*D*) Explain why it would not be sensible to use the equation of the line of best fit to estimate the death rate in a Caribbean country where the birth rate is known. **[1]**

(*E*) Explain why it is unlikely that the sample is random. **[1]**

Including Togo there were 56 items available for selection.

(iii) Describe how a sample of size 14 from this data could be generated for further analysis using systematic sampling. **[2]**

9 The box and whisker diagrams in Fig. 9.1 summarise the birth rates per 1000 people for all the countries in three of the regions as given in the pre-release data set. They were drawn as part of an investigation comparing birth rates in different regions of the world.

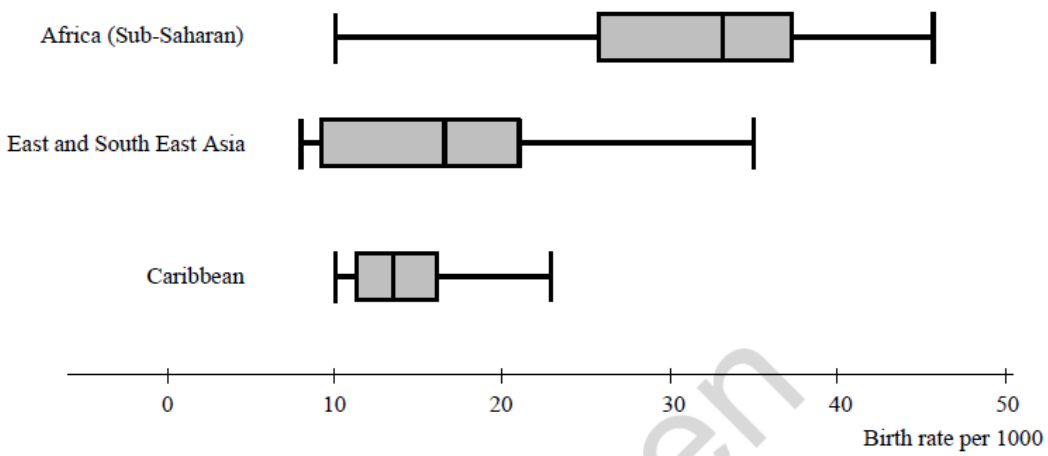


Fig. 9.1

(i) Discuss the distributions of birth rates in these regions of the world. Make three different statements. You should refer to **both** information from the box and whisker diagrams **and** your knowledge of the large data set. **[3]**

(ii) The birth rate for all the countries in Australasia are shown below.

Country	Birth rate per 1000
Australia	12.19
New Zealand	13.4
Papua New Guinea	24.89

(*A*) Explain why the calculation below is not a correct method for finding the birth rate per 1000 for Australasia as a whole.

$$\frac{12.19+13.4+24.89}{3} \approx 16.83$$

[1]

(*B*) Without doing any calculations, explain whether the birth rate per 1000 for Australasia as a whole is higher or lower than 16.83. **[1]**

The scatter diagram in Fig. 9.2 shows birth rate per 1000 and physicians/1000 population for all the countries in the pre-release data set.

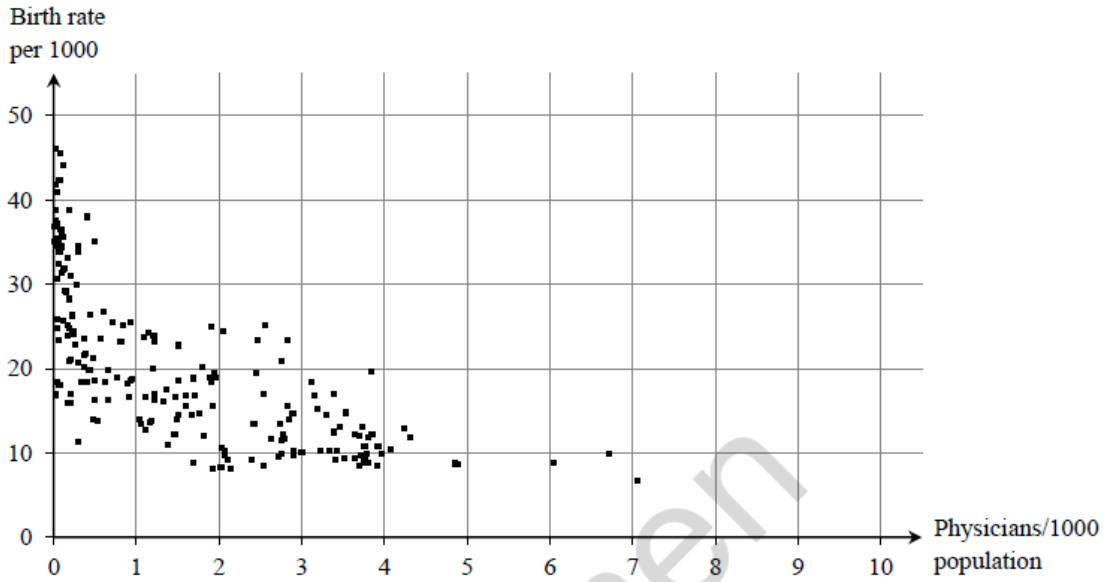


Fig. 9.2

(iii) Describe the correlation in the scatter diagram. **[1]**

(iv) Discuss briefly whether the scatter diagram shows that high birth rates would be reduced by increasing the number of physicians in a country. **[1]**

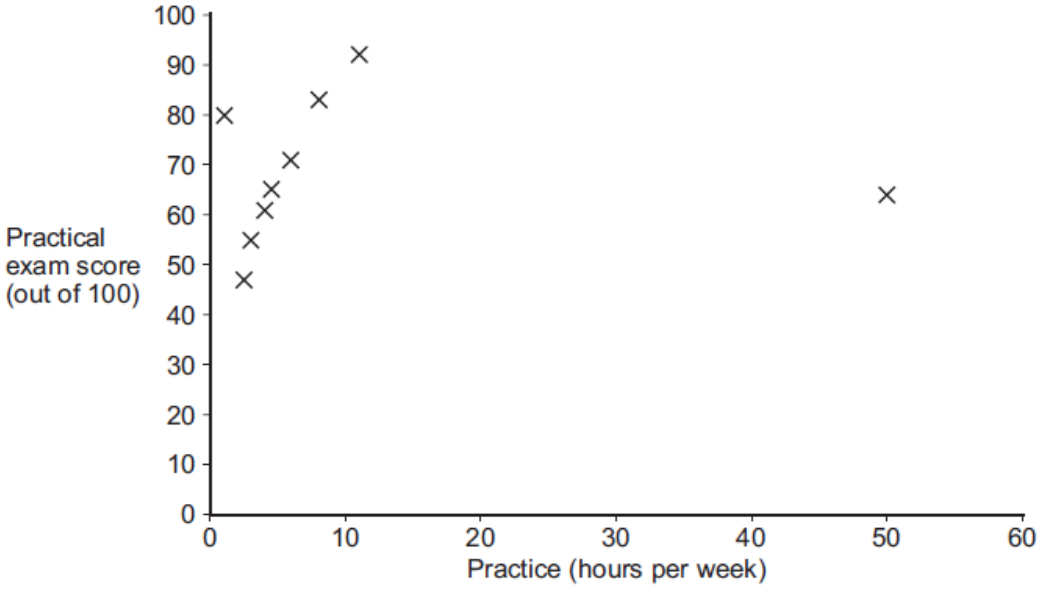
AQA AS 2018 Paper 2 Question 18:

18 Jennie is a piano teacher who teaches nine pupils.

She records how many hours per week they practice the piano along with their most recent practical exam score.

Student	Practice (hours per week)	Practical exam score (out of 100)
Donovan	50	64
Vazquez	6	71
Higgins	3	55
Begum	2.5	47
Collins	1	80
Coldbridge	4	61
Nedbalek	4.5	65
Carter	8	83
White	11	92

She plots a scatter diagram of this data, as shown below.



18 (a) Identify two possible outliers by name, giving a possible explanation for the position on the scatter diagram of each outlier.

[4 marks]

First outlier _____

Possible reason _____

_____

Second outlier _____

Possible reason _____

_____

18 (b) Jennie discards the two outliers.

18 (b) (i) Describe the correlation shown by the scatter diagram for the remaining points.

[1 mark]

_____

_____

_____

18 (b) (ii) Interpret this correlation in the context of the question.

[1 mark]

_____

_____

_____

Edexcel AS Sample Paper 2 Question 4:

4. Sara was studying the relationship between rainfall, r mm, and humidity, h%, in the UK.
She takes a random sample of 11 days from May 1987 for Leuchars from the large data set.

She obtained the following results.

h	93	86	95	97	86	94	97	97	87	97	86
r	1.1	0.3	3.7	20.6	0	0	2.4	1.1	0.1	0.9	0.1

Sara examined the rainfall figures and found

$$Q_1 = 0.1 \qquad Q_2 = 0.9 \qquad Q_3 = 2.4$$

A value that is more than 1.5 times the interquartile range (IQR) above Q_3 is called an outlier.

(a) Show that $r = 20.6$ is an outlier.

(1)

(b) Give a reason why Sara might:

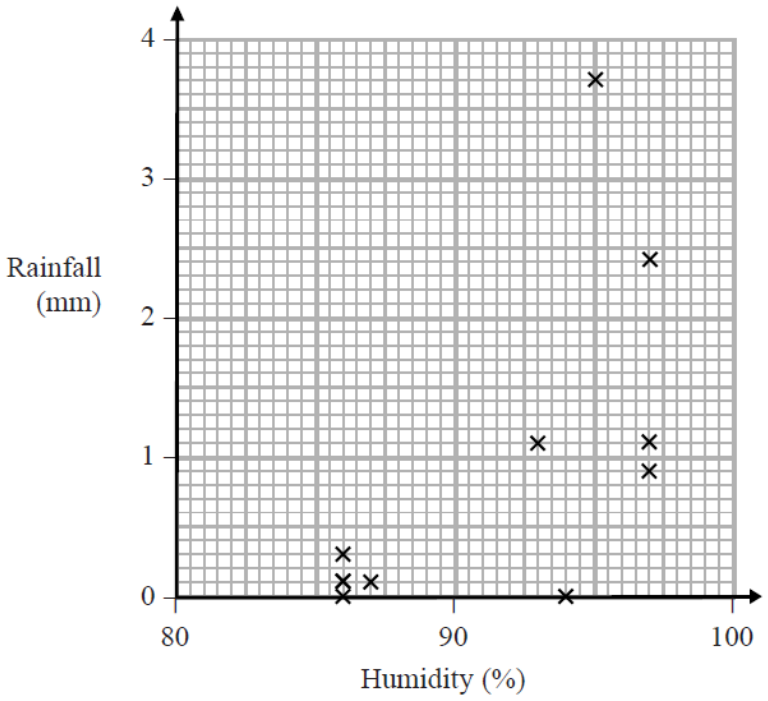
(i) include

(ii) exclude

this day's reading.

(2)

Sara decided to exclude this day's reading and drew the following scatter diagram for the remaining 10 days' values of r and h.



(c) Give an interpretation of the correlation between rainfall and humidity.

(1)

The equation of the regression line of r on h for these 10 days is $r = -12.8 + 0.15h$

(d) Give an interpretation of the gradient of this regression line.

(1)

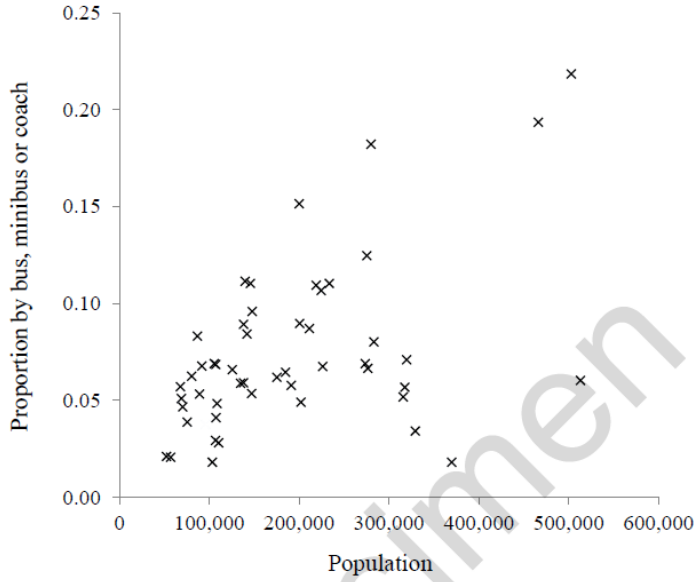
(e) (i) Comment on the suitability of Sara's sampling method for this study.

(ii) Suggest how Sara could make better use of the large data set for her study.

(2)

OCR A AS Sample Paper 1 Question 11:

11 The scatter diagram below shows data taken from the 2011 UK census. The scatter diagram shows, for each of the Local Authorities in the North East and North West regions, the total population of the Local Authority and the proportion of its workforce that travel to work by bus, minibus or coach.



(i) Samuel suggests that, with a few exceptions, the data points in the diagram show that Local Authorities with larger populations generally have higher proportions of workers travelling by bus, minibus or coach. On the diagram in the Printed Answer Booklet draw a ring around each of the data points that Samuel might regard as an exception. **[1]**

(ii) Jasper suggests that it is possible to separate these Local Authorities into more than one group with different relationships between population and proportion travelling to work by bus, minibus or coach. Discuss Jasper's suggestion, referring to the data and to how differences between the Local Authorities could explain the patterns seen in the diagram. **[3]**

---

11 Christa used Pearson's product-moment correlation coefficient, r, to compare the use of public transport with the use of private vehicles for travel to work in the UK.

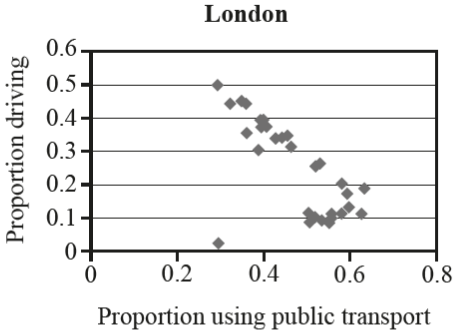
(i) Using the pre-release data set for all 348 UK Local Authorities, she considered the following four variables.

Number of employees using public transport	x
Number of employees using private vehicles	y
Proportion of employees using public transport	a
Proportion of employees using private vehicles	b

(a) Explain, in context, why you would expect strong, positive correlation between x and y. **[1]**

(b) Explain, in context, what kind of correlation you would expect between a and b. **[2]**

(ii) Christa also considered the data for the 33 London boroughs alone and she generated the following scatter diagram.



One London Borough is represented by an outlier in the diagram.

(a) Suggest what effect this outlier is likely to have on the value of r for the 32 London Boroughs. **[1]**

(b) Suggest what effect this outlier is likely to have on the value of r for the whole country. **[1]**

(c) What can you deduce about the area of the London Borough represented by the outlier? Explain your answer. **[1]**