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)

2.

A sixth form college has 84 students in Year 12 and 56 students in Year 13

The head teacher selects a stratified sample of 40 students, stratified by year group.

(a) Describe how this sample could be taken.

(3)

The head teacher is investigating the relationship between the amount of sleep, s hours, that each student had the night before they took an aptitude test and their performance in the test, p marks.
For the sample of 40 students, he finds the equation of the regression line of p on s to be

$$p = 26.1 + 5.60s$$

(b) With reference to this equation, describe the effect that an extra 0.5 hours of sleep may have, on average, on a student's performance in the aptitude test.

(1)

(c) Describe one limitation of this regression model.

(1)

3. The relationship between two variables p and t is modelled by the regression line with equation

$$p = 22 - 1.1\,t$$

The model is based on observations of the independent variable, t, between 1 and 10

(a) Describe the correlation between p and t implied by this model.

(1)

Given that p is measured in centimetres and t is measured in days,

(b) state the units of the gradient of the regression line.

(1)

Using the model,

(c) calculate the change in p over a 3-day period.

(2)

Tisam uses this model to estimate the value of p when $t = 19$

(d) Comment, giving a reason, on the reliability of this estimate.

(1)

4. Helen is studying one of the qualitative variables from the large data set for Heathrow from 2015.

She started with the data from 3rd May and then took every 10th reading.

There were only 3 different outcomes with the following frequencies

Outcome	*A*	*B*	*C*
Frequency	16	2	1

(a) State the sampling technique Helen used.

(1)

(b) From your knowledge of the large data set

(i) suggest which variable was being studied,

(ii) state the name of outcome *A*.

(2)

George is also studying the same variable from the large data set for Heathrow from 2015. He started with the data from 5th May and then took every 10th reading and obtained the following

Outcome	*A*	*B*	*C*
Frequency	16	1	1

Helen and George decided they should examine all of the data for this variable for Heathrow from 2015 and obtained the following

Outcome	*A*	*B*	*C*
Frequency	155	26	3

(c) State what inference Helen and George could reliably make from their original samples about the outcomes of this variable at Heathrow, for the period covered by the large data set in 2015.

(1)

5. Fred and Nadine are investigating whether there is a linear relationship between Daily Mean Pressure, p hPa, and Daily Mean Air Temperature, t °C, in Beijing using the 2015 data from the large data set.

Fred randomly selects one month from the data set and draws the scatter diagram in Figure 1 using the data from that month.

The scale has been left off the horizontal axis.

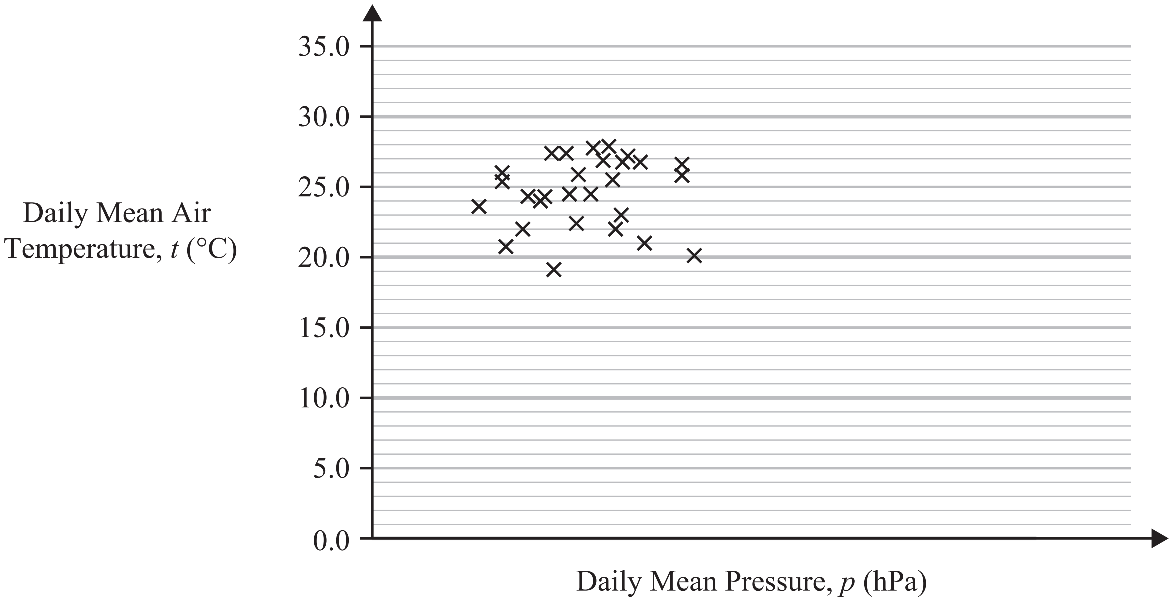


Figure 1

(a) Describe the correlation shown in Figure 1.

(1)

Nadine chooses to use all of the data for Beijing from 2015 and draws the scatter diagram in Figure 2.

She uses the same scales as Fred.

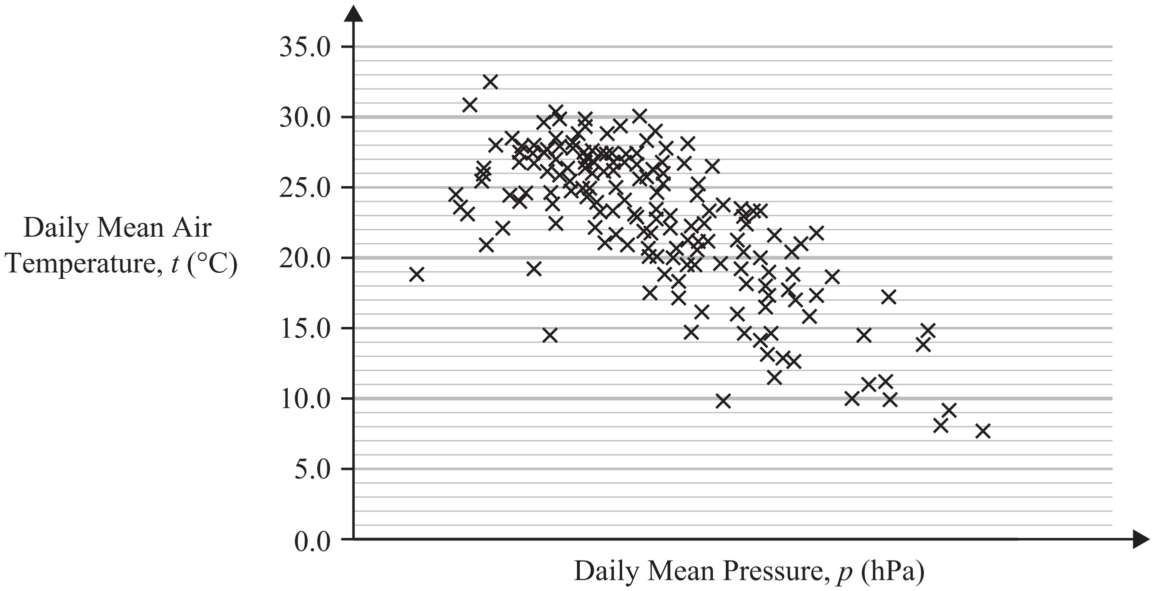


Figure 2

Question 5 continued

(b) Explain, in context, what Nadine can infer about the relationship between p and t using the information shown in Figure 2.

(1)

(c) Using your knowledge of the large data set, state a value of p for which interpolation can be used with Figure 2 to predict a value of t.

(1)

(d) Using your knowledge of the large data set, explain why it is not meaningful to look for a linear relationship between Daily Mean Wind Speed (Beaufort Conversion) and Daily Mean Air Temperature in Beijing in 2015.

(1)

6. Jerry is studying visibility for Camborne using the large data set June 1987.

The table below contains two extracts from the large data set.

It shows the daily maximum relative humidity and the daily mean visibility.

Date	**Daily Maximum Relative Humidity**	**Daily Mean Visibility**
Units	**%**	
10/06/1987	90	5300
28/06/1987	100	0

(The units for Daily Mean Visibility are deliberately omitted.)

Given that daily mean visibility is given to the nearest 100,

(a) write down the range of distances in metres that corresponds to the recorded value 0 for the daily mean visibility.

(1)

Jerry drew the following scatter diagram, Figure 2, and calculated some statistics using the June 1987 data for Camborne from the large data set.

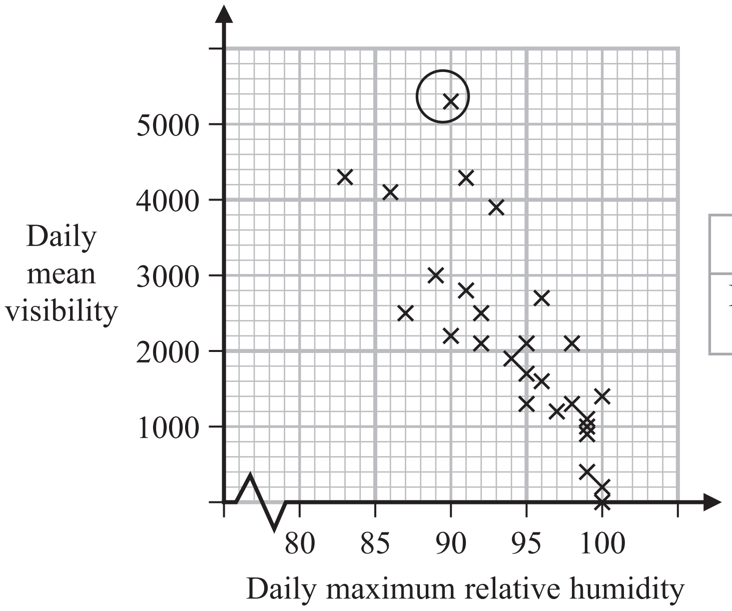


Figure 2

	Q_1	IQR
Daily mean visibility	1100	1600
Daily maximum relative humidity (%)	92	8

Jerry defines an outlier as a value that is more than 1.5 times the interquartile range above Q_3 or more than 1.5 times the interquartile range below Q_1.

(b) Show that the point circled on the scatter diagram is an outlier for visibility.

(2)

(c) Interpret the correlation between the daily mean visibility and the daily maximum relative humidity.

(1)

Jerry drew the following scatter diagram, Figure 3, using the June 1987 data for Camborne from the large data set, but forgot to label the *x*–axis.

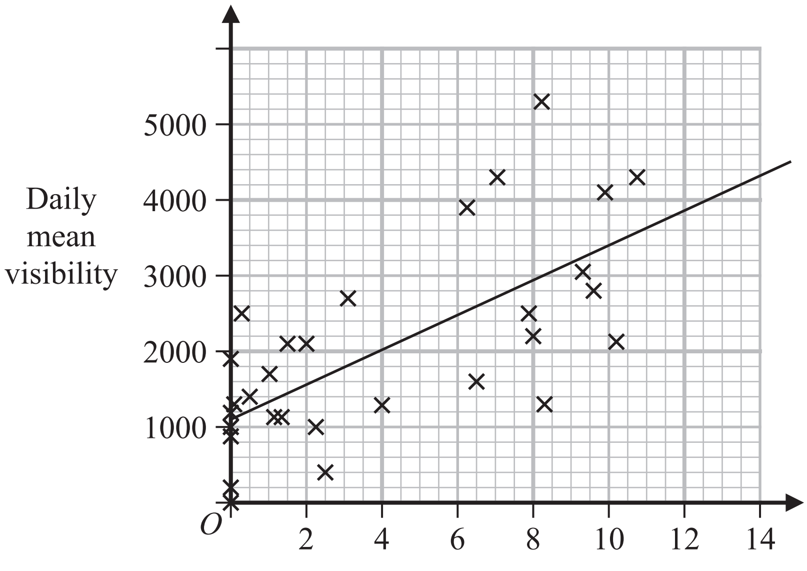


Figure 3

(d) Using your knowledge of the large data set, suggest which variable the *x*-axis on this scatter diagram represents.

(1)

7. 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)

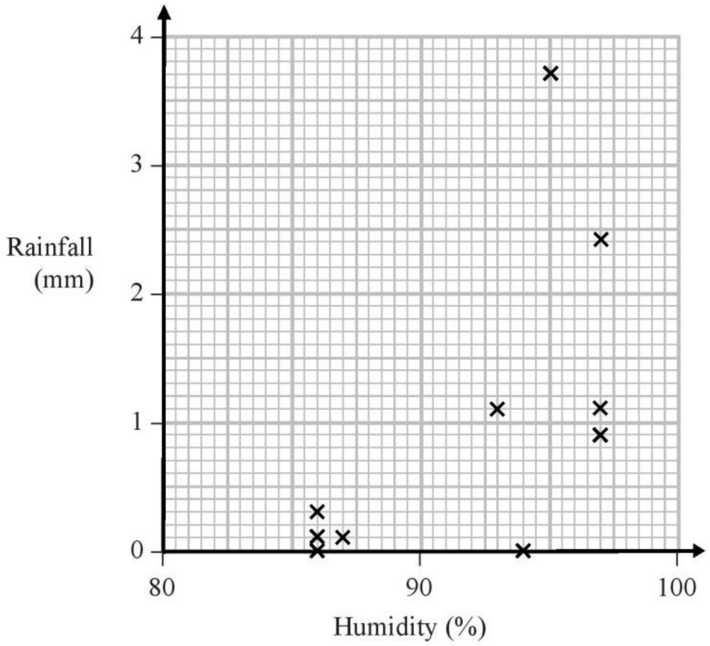
(b) Give a reason why Sara might (i) include

(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)

Question 7 continued

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)