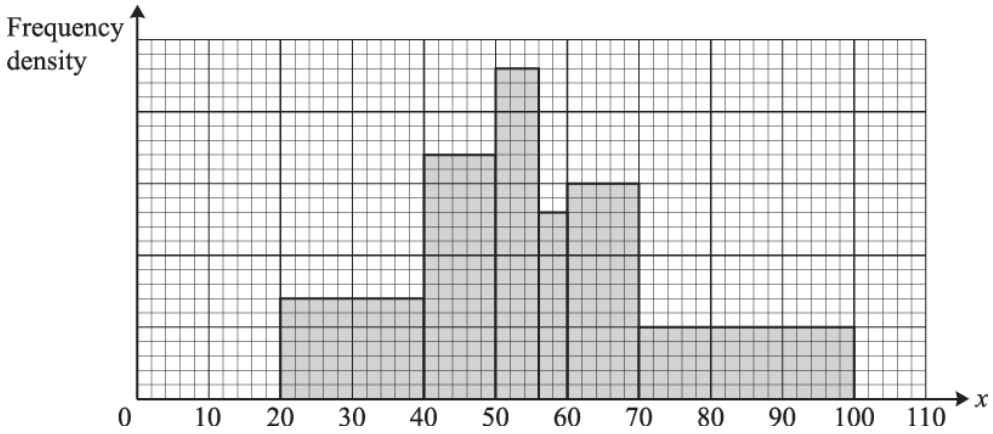
1. The masses, x grams, of 800 apples are summarised in the histogram.



i. On the frequency density axis, 1 cm represents a units. Find the value of a.

[3]

ii. Find an estimate of the median mass of the apples.

[4]

2. The lengths, in centimetres, of 18 snakes are given below.

24 62 20 65 27 67 69 32 40 53 55 47 33 45 55 56 49 58

i. Draw an ordered stem-and-leaf diagram for the data.

[3]

ii. Find the mean and median of the lengths of the snakes.

[2]

iii. It was found that one of the lengths had been measured incorrectly. After this length was corrected, the median increased by 1 cm. Give two possibilities for the incorrect length and give a corrected value in each case.

[2]

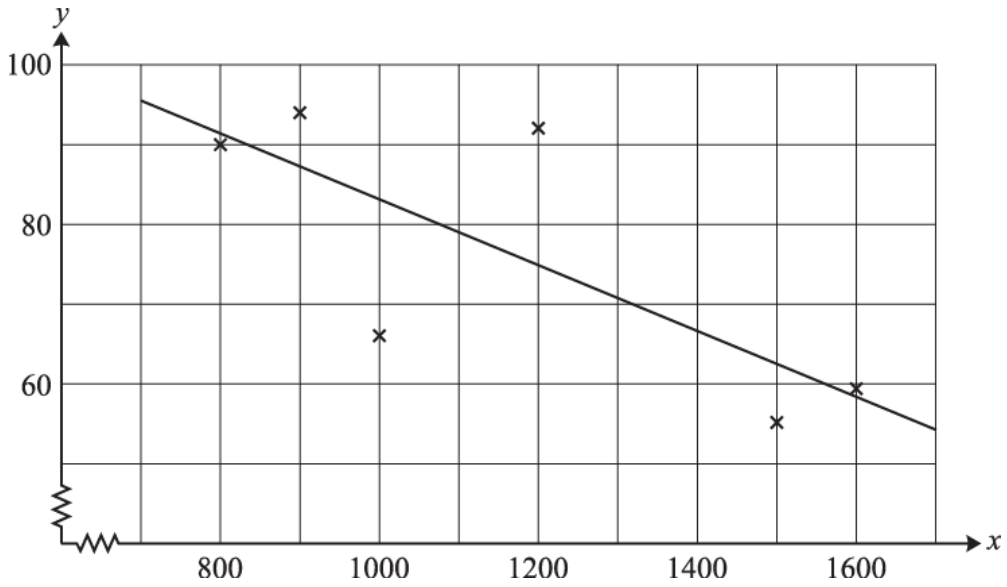
3. The Gross Domestic Product per Capita (GDP), x dollars, and the Infant Mortality Rate per thousand (IMR), y, of 6 African countries were recorded and summarised as follows.

$n = 6$ $\quad \Sigma x = 7000$ $\quad \Sigma x^2 = 8\ 700\ 000$ $\quad \Sigma y = 456$ $\quad \Sigma y^2 = 36\ 262$ $\quad \Sigma xy = 509\ 900$

i. Calculate the equation of the regression line of y on x for these 6 countries.

[4]

The original data were plotted on a scatter diagram and the regression line of y on x was drawn, as shown below.



ii. The GDP for another country, Tanzania, is 1300 dollars. Use the regression line in the diagram to estimate the IMR of Tanzania.

[1]

iii. The GDP for Nigeria is 2400 dollars. Give two reasons why the regression line is unlikely to give a reliable estimate for the IMR for Nigeria.

[2]

iv. The actual value of the IMR for Tanzania is 96. The data for Tanzania ($x = 1300$, $y = 96$) is now included with the original 6 countries. Calculate the value of the product moment correlation coefficient, r, for all 7 countries.

[4]

v. The IMR is now redefined as the infant mortality rate per hundred instead of per thousand, and the value of r is recalculated for all 7 countries. Without calculation state what effect, if any, this would have on the value of r found in part (iv).

[1]

4. At a stall in a fair, contestants have to estimate the mass of a cake. A group of 10 people made estimates, m kg, and for each person the value of $(m - 5)$ was recorded. The mean and standard deviation of $(m - 5)$ were found to be 0.74 and 0.13 respectively.

i. Write down the mean and standard deviation of m.

[2]

The mean and standard deviation of the estimates made by another group of 15 people were found to be 5.6 kg and 0.19 kg respectively.

ii. Calculate the mean of all 25 estimates.

[2]

iii. Fiona claims that if a group's estimates are more consistent, they are likely to be more accurate. Given that the true mass of the cake is 5.65 kg, comment on this claim.

[2]

5. The table shows information about the numbers of people per household in 280 900 households in the northwest of England in 2001.

Number of people	1	2	3	4	5 or more
Number of households	86900	92500	45000	37100	19400

i. Taking '5 or more' to mean '5 or 6', calculate estimates of the mean and standard deviation of the number of people per household.

[5]

ii. State the values of the median and upper quartile of the number of people per household.

[2]

6. The stem-and-leaf diagram shows the heights, in metres to the nearest 0.1 m, of a random sample of trees of species *A*.

5	
5	9
6	1 4
6	5 5 9
7	2 3 3 4
7	5 6 6 6 7 8
8	0 3 4
8	5

Key: 6 | 4 means 6.4 m

i. Find the median and interquartile range of the heights.

[3]

ii. The heights, in metres to the nearest 0.1 m, of a random sample of trees of species *B* are given below.

7.6 5.2 8.5 5.2 6.3 6.3 6.8 7.2 6.7 7.3 5.4 7.5 7.4 6.0 6.7

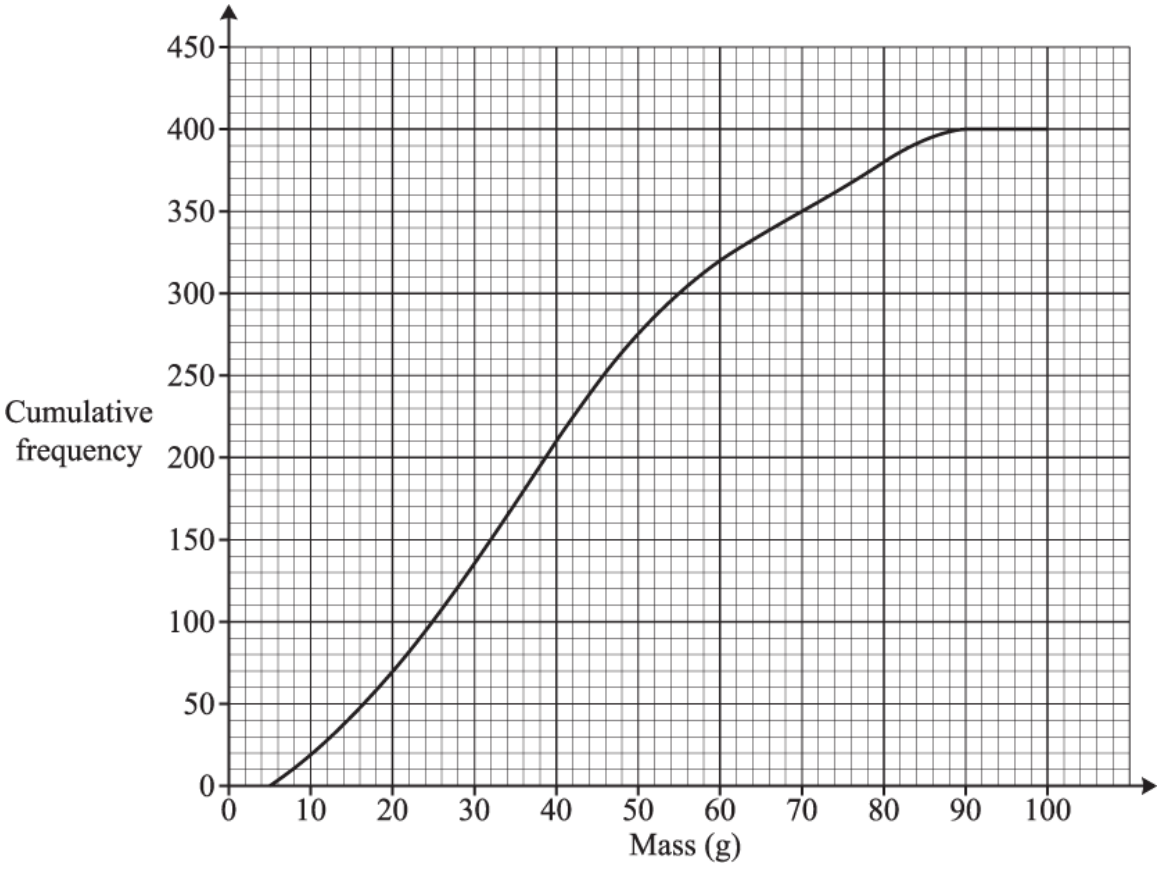
In the answer book, complete the back-to-back stem-and-leaf diagram.

[2]

iii. Make two comparisons between the heights of the two species of tree.

[2]

7. The masses, in grams, of 400 plums were recorded. The masses were then collected into class intervals of width 5 g and a cumulative frequency graph was drawn, as shown below.



i. Find the number of plums with masses in the interval 40 g to 45 g.

[1]

ii. Find the percentage of plums with masses greater than 70 g.

[2]

iii. Give estimates of the highest and lowest masses in the sample, explaining why their exact values cannot be read from the graph.

[2]

iv. On the graph paper in the answer book, draw a box-and-whisker plot to illustrate the masses of the plums in the sample.

[4]

v. Comment briefly on the shape of the distribution of masses.

[1]

8. The masses, m grams, of 52 apples of a certain variety were found and summarised as follows.

$n = 52$	$\sum(m - 150) =$	$\sum(m - 150)^2 =$
	-182	1768

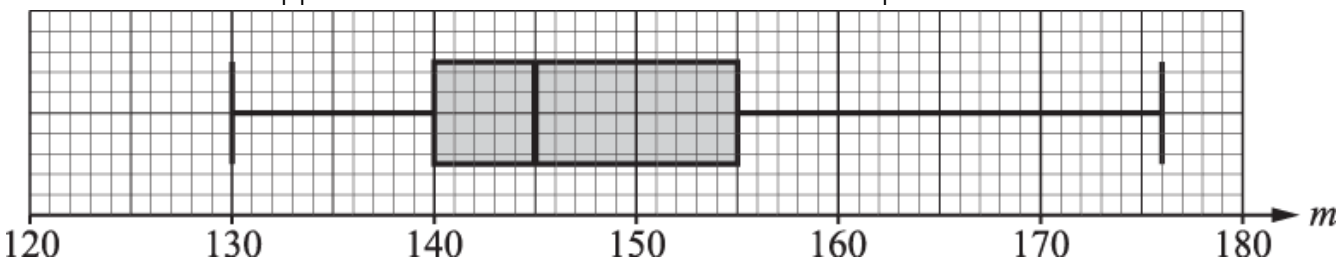
i. Find the mean and variance of the masses of these 52 apples.

[5]

ii. Use your answers from part **(i)** to find the exact value of $\sum m^2$.

[3]

The masses of the apples are illustrated in the box-and-whisker plot below.



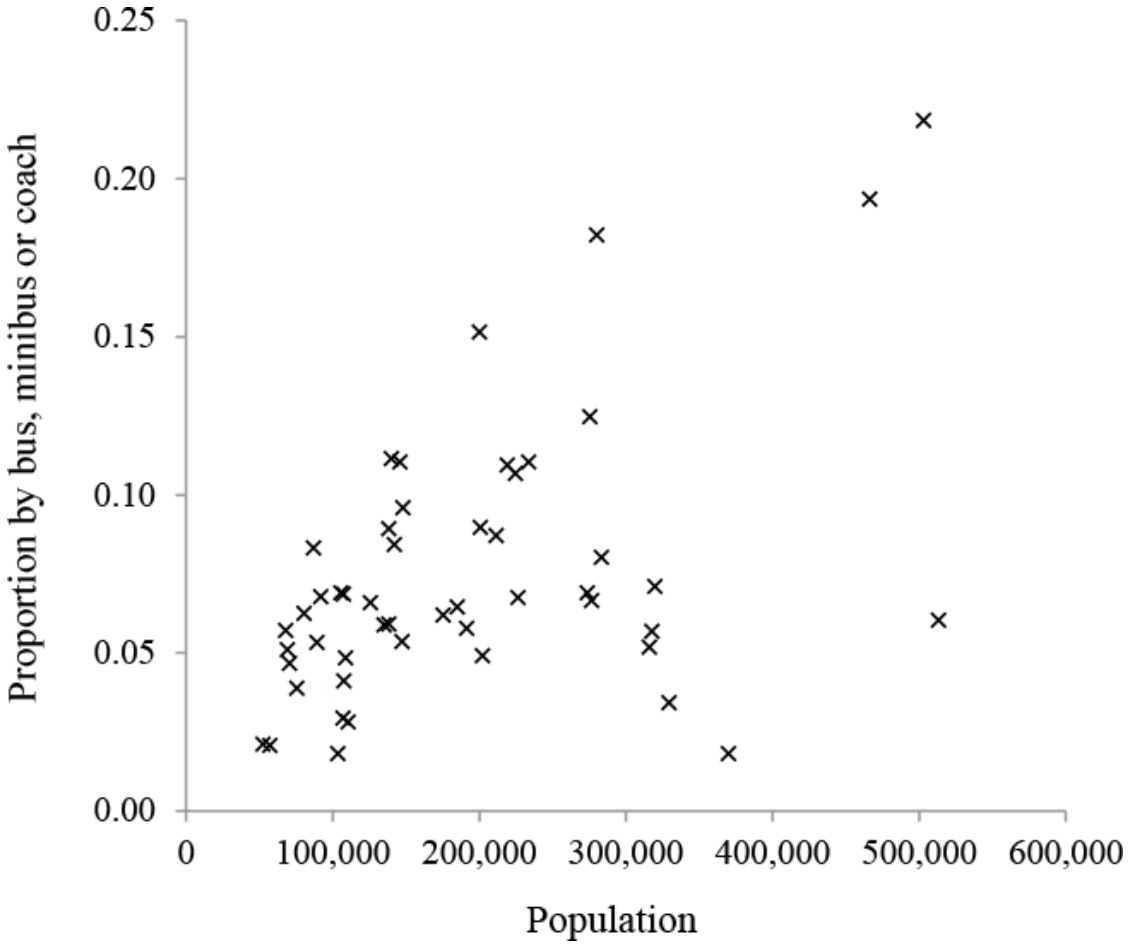
iii. How many apples have masses in the interval $130 \leqslant m < 140$?

[2]

iv. An 'outlier' is a data item that lies more than 1.5 times the interquartile range above the upper quartile, or more than 1.5 times the interquartile range below the lower quartile. Explain whether any of the masses of these apples are outliers.

[3]

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



(a) 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 above draw a ring around each of the data points that Samuel might regard as an exception. [1]

(b) 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]

10. Clara used some data from the 2011 UK census to summarise information on carbon emissions due to travel to work, in two Local Authorities. Her results are shown below.

	Method of travel to work	Individual motorised transport	Shared motorised transport	Public transport	No motorised transport	
	Carbon emissions category	High	Medium	Low	None	Total
Local Authority A	Number of workers	174 374	42 112	61 483	76 024	353 993
	Percentage of workers	49.3	11.9	17.4	21.5	100
Local Authority B	Number of workers	39 433	9944	4614	16 232	70 223
	Percentage of workers	56.2	14.2	6.6	23.1	100

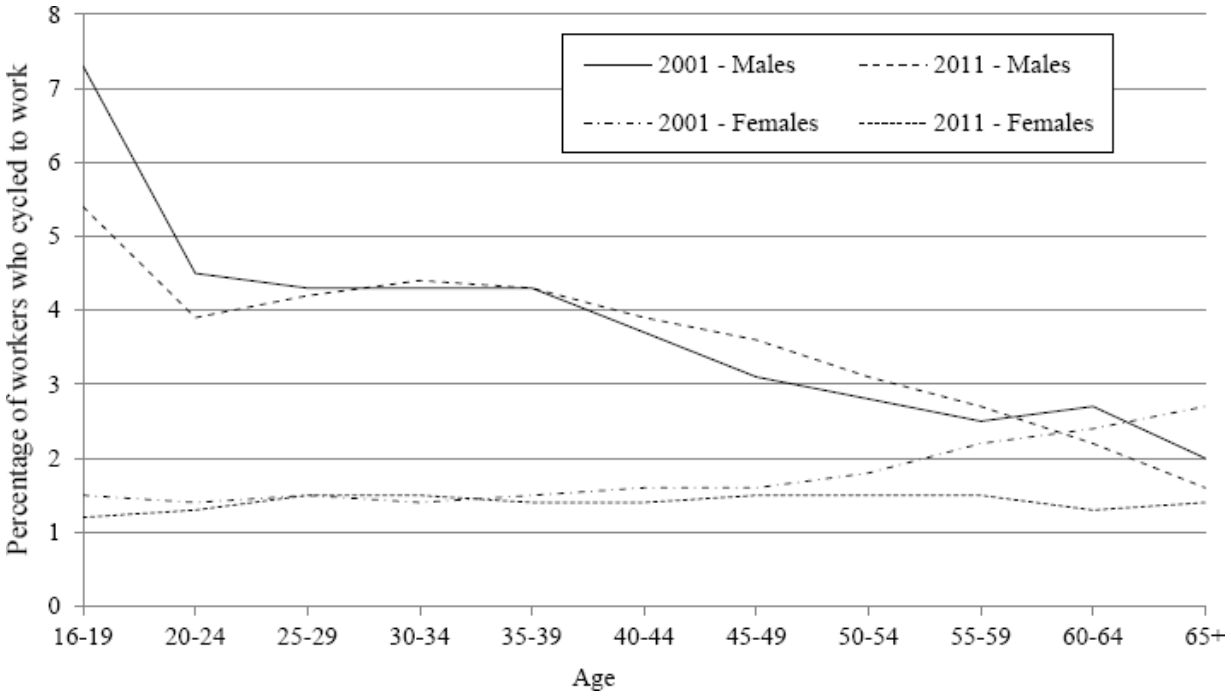
(a) Clara calculated the values for the column headed "shared motorised transport" by doubling the value in the "passenger in a car or van" column of the original data set. Explain what assumption she has made and what other adjustment would need to be made to the data to take account of this. [2]

(b) Clara suggests that the average carbon emissions per worker due to travelling to work is larger in region B than in region A.

(i) Use data from the table to support Clara's suggestion. [1]

(ii) Use data from the table to argue against Clara's suggestion. [1]

11. The diagram below shows some "Cycle to work" data taken from the 2001 and 2011 UK censuses. The diagram shows the percentages, by age group, of male and female workers in England and Wales, excluding London, who cycled to work in 2001 and 2011.



The following questions refer to the workers represented by the graphs in the diagram.

(a) A researcher is going to take a sample of men and a sample of women and ask them whether or not they cycle to work. Why would it be more important to stratify the sample of men? [1]

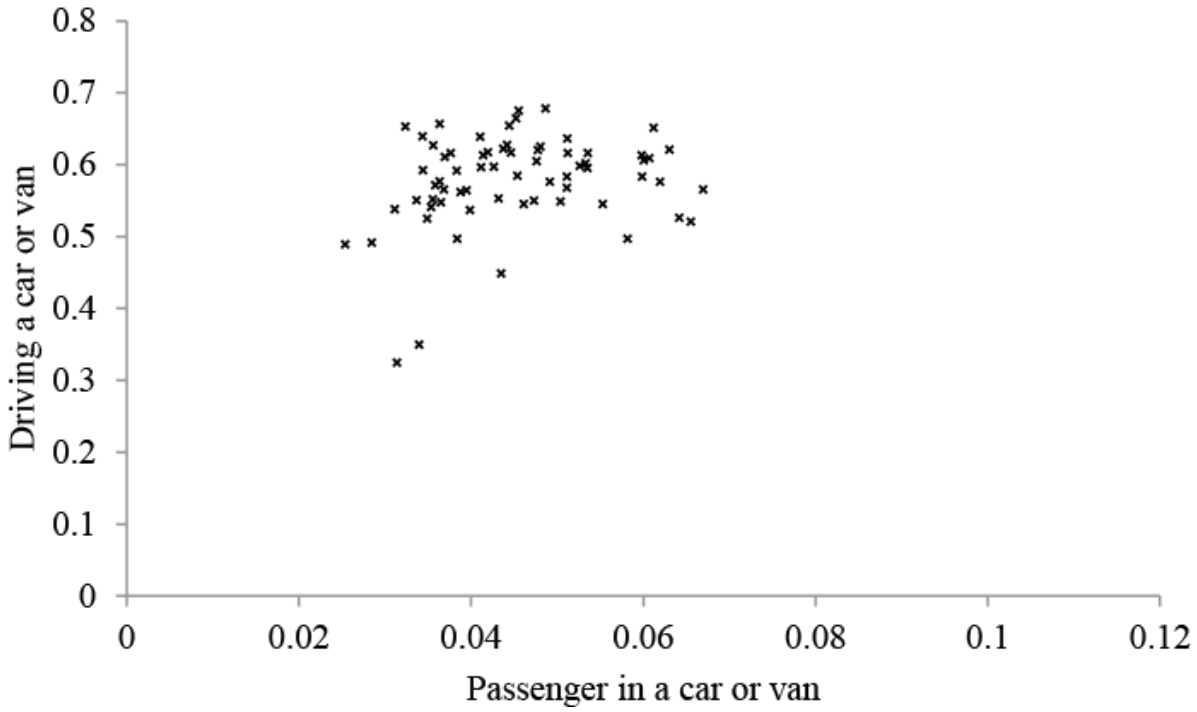
(b) A research project followed a randomly chosen large sample of the group of male workers who were aged 30-34 in 2001. Does the diagram suggest that the proportion of this group who cycled to work has increased or decreased from 2001 to 2011? Justify your answer. [2]

(c) Write down one assumption that you have to make about these workers in order to draw this conclusion. [1]

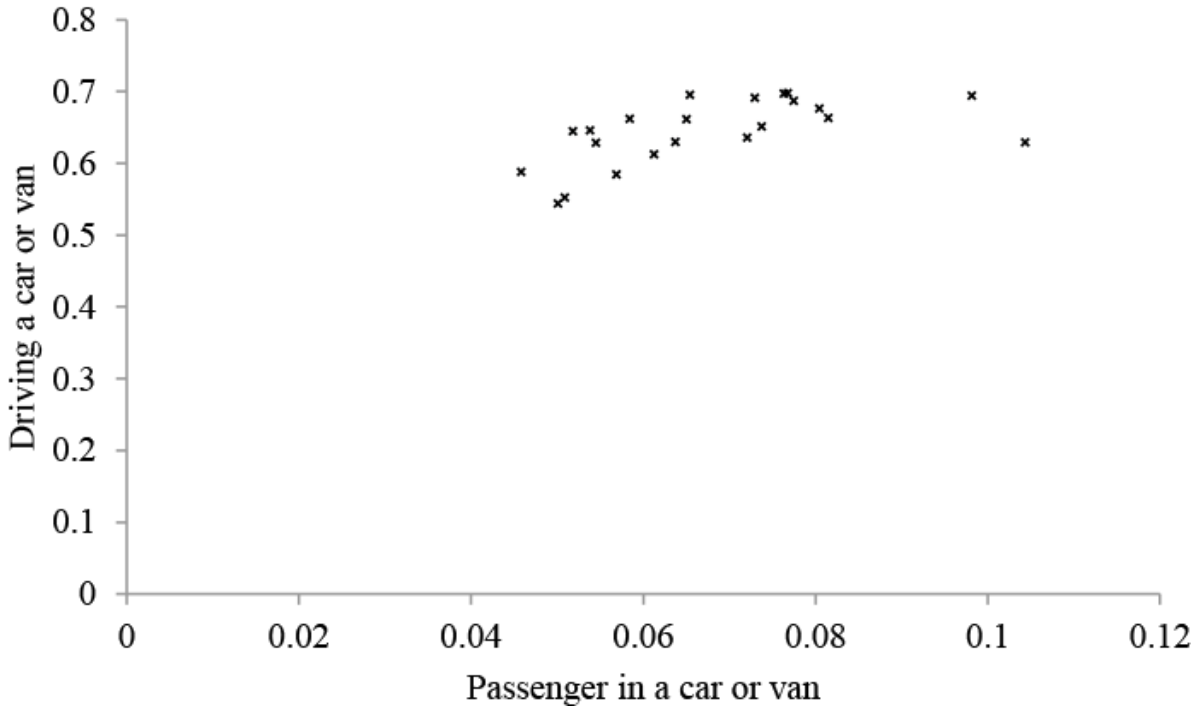
12. The table and the four scatter diagrams below show data taken from the 2011 UK census for four regions.
On the scatter diagrams the names have been replaced by letters.
The table shows, for each region, the mean and standard deviation of the proportion of workers in each Local Authority who travel 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

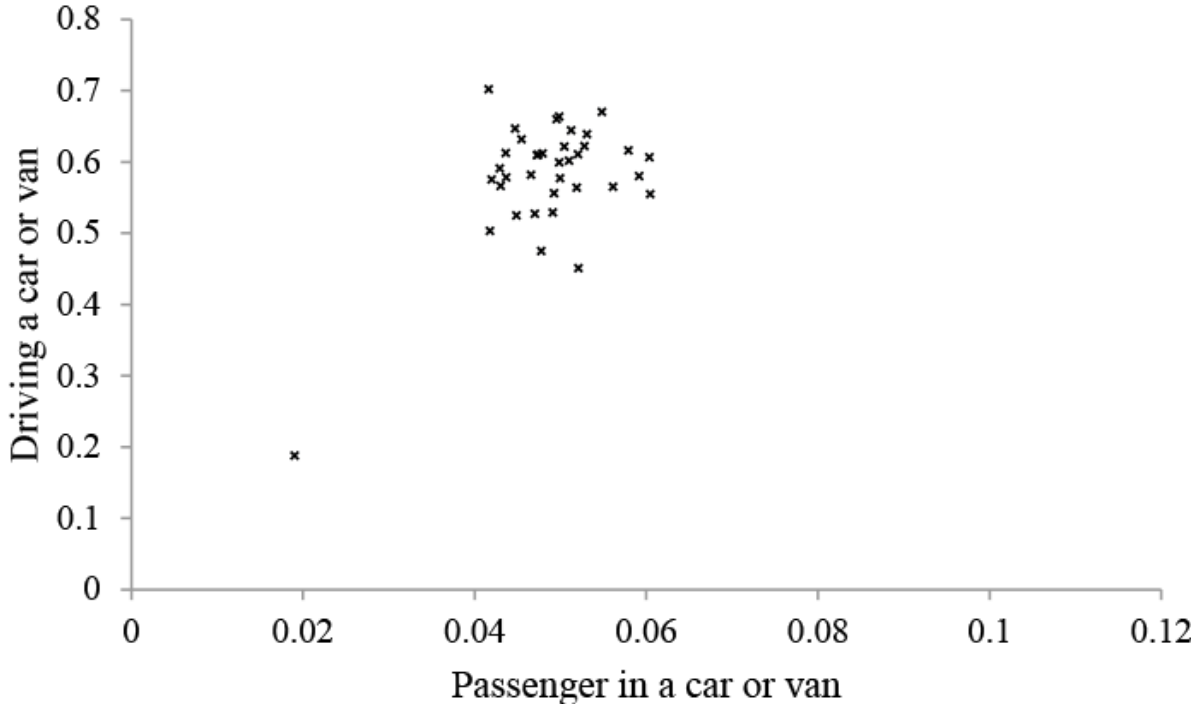
Region A



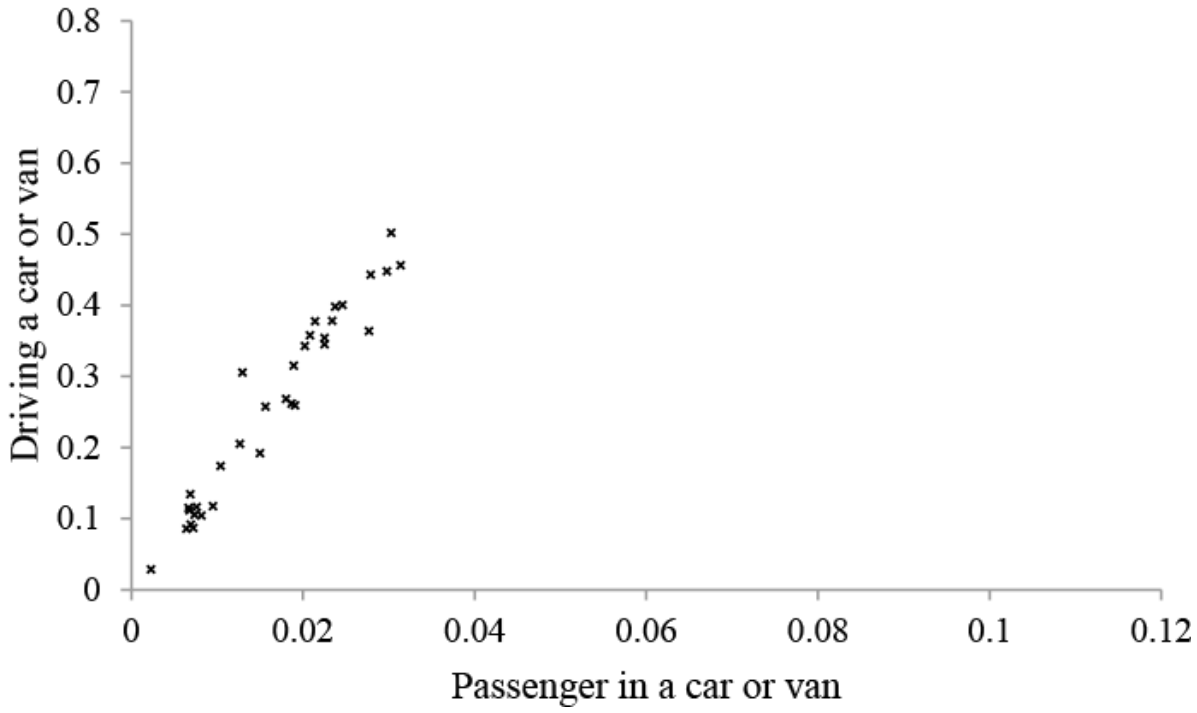
Region B



Region C



Region D



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

(b) 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]

(c) 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]

13. Frances used the pre-release data set to produce the following table which shows information about the residents of Norwich in 2011.

Age	0 to 15	16 to 24	25 to 44	45 to 64	65 and over	Total
Number of residents	21 707	22 921	40 894	27 645	19 345	132 512

(a) State the upper class boundary of the "25 to 44" class. [1]

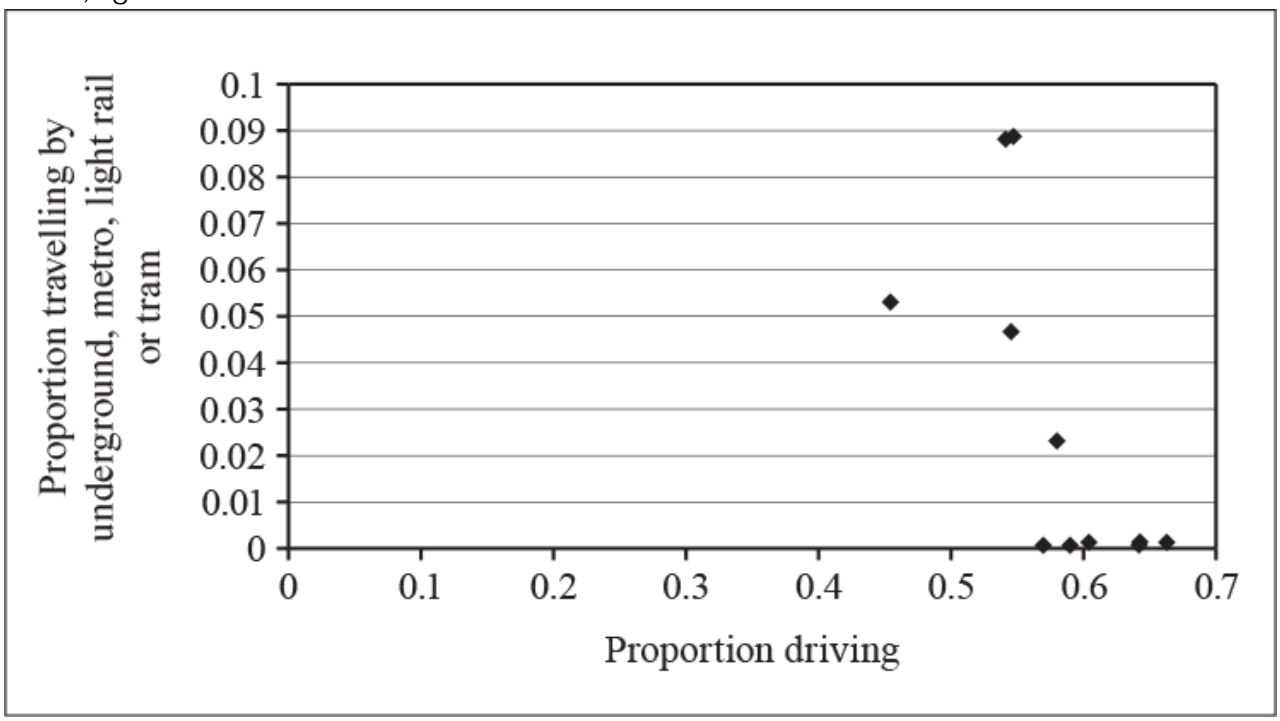
Frances used these data to calculate estimates of the mean and standard deviation of the ages of these residents. She assumed that the oldest resident was aged 105.

(b) Calculate these estimates. [3]

(c) Use these estimates to discuss whether there may be any outliers. [2]

(d) Jacob suggested that more reliable estimates could be obtained by using the value 80 for the last class, instead of the midpoint. Explain, with a reason, whether you think this suggestion is a good one. [1]

14. The scatter diagram shows data taken from the pre-release data set for several Local Authorities in a region of the UK. The diagram shows, for each Local Authority, the proportion of workers driving to work, and the proportion travelling to work by underground, metro, light rail or tram.



(a) On the diagram above, identify the points corresponding to two distinct sections of the population represented in the diagram. [1]

(b) Suggest a reason why there are two distinct sections of the population represented by the points in the diagram. [1]

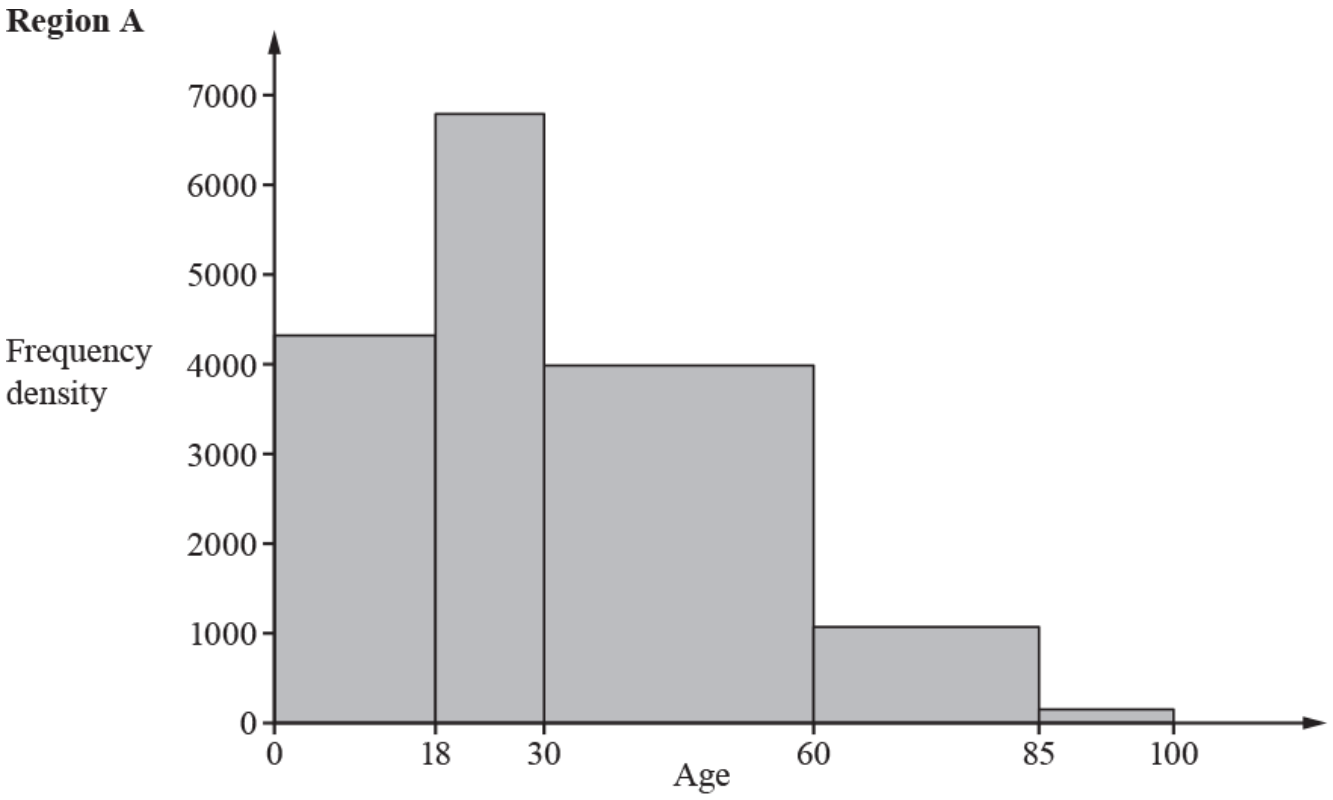
The data for another local authority in this region can be represented by the point (0.62, 0.004).

(c) (i) To which of the two distinct sections of the population does this Local Authority belong? Explain your answer. [1]

(ii) What can you deduce about this Local Authority? [1]

(d) A student suggests that the Local Authority represented by the point (0.55, 0.089) is a non-metropolitan district. Comment on this suggestion. [1]

15. John used data from the 2011 UK census to produce the following histogram for region A.



In the Census report, the age classes were given as follows.

0 to 4	5 to 7	8 to 9	10 to 14	15	16 to 17	18 to 19	20 to 24	25 to 29	30 to 44	45 to 59	60 to 64	65 to 74	75 to 84	85 to 89	90 and over

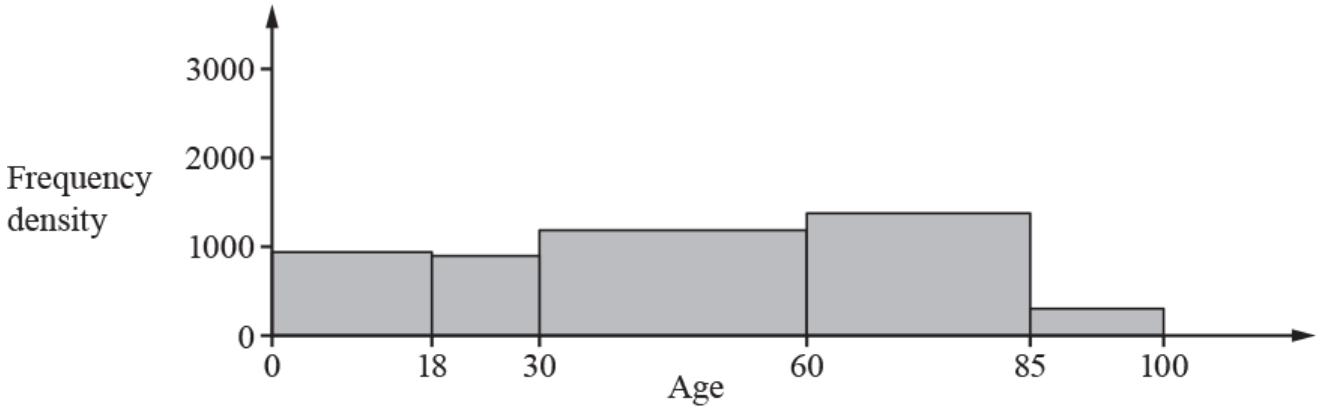
John combined classes to give the classes shown in the histogram.

(a) (i) Explain the reason for John's choice of upper class boundary for the first class. [1]

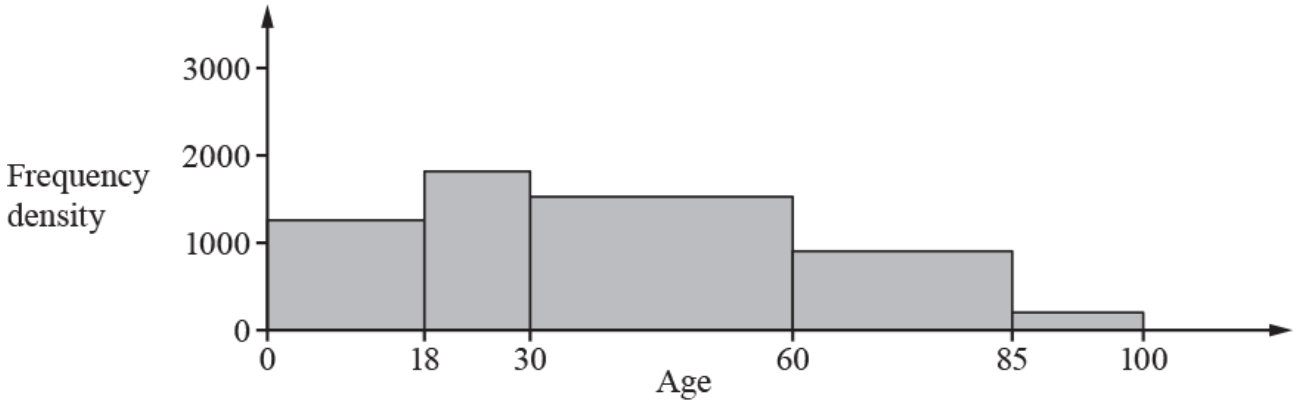
(ii) Suggest a reason for John's choice of upper class boundary for the last class. [1]

John also produced similar histograms for two other UK regions, B and C.

Region B



Region C



(b) Which of the three regions had the largest proportion of people aged 85 and over? Without detailed calculations, explain your answer. [3]

The mean ages, in years, of the populations in the three regions were 47.5, 39.5 and 31.5.

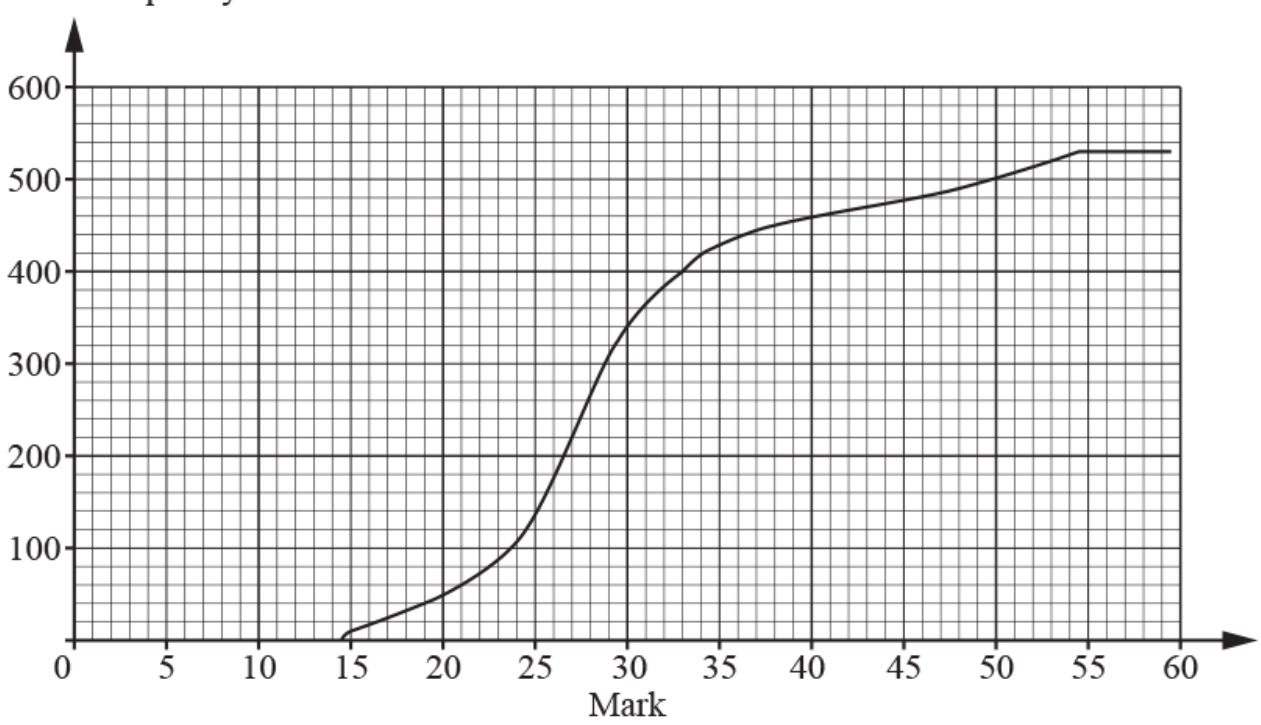
(c) For each of these means, state the region to which it corresponds. Justify your answers. [3]

John made the following claim.

> "The histograms show that a child living in region B in 2011 could expect to live longer than a child living in region A in 2011."

(d) Is this claim justified? Give a reason for your answer. [1]

16. The marks of some students in an examination were summarised in a grouped frequency distribution, using the following classes: 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, all inclusive. A cumulative frequency diagram was drawn, as shown below.



(i) How many students took the examination? [1]

(ii) 20% of students gained the top grade. Find the minimum mark for the top grade. [3]

(iii) A teacher said

"The cumulative frequency graph shows that the highest mark scored by any student was 54 or 55."

Explain why this statement is incorrect, and give an improved statement about the highest mark. [2]

(iii) State which class is the modal class, explaining how you know. [2]

17. The mean and standard deviation of the weights, w grams, of a sample of 75 stones were found to be 52.3 and 5.8 respectively.

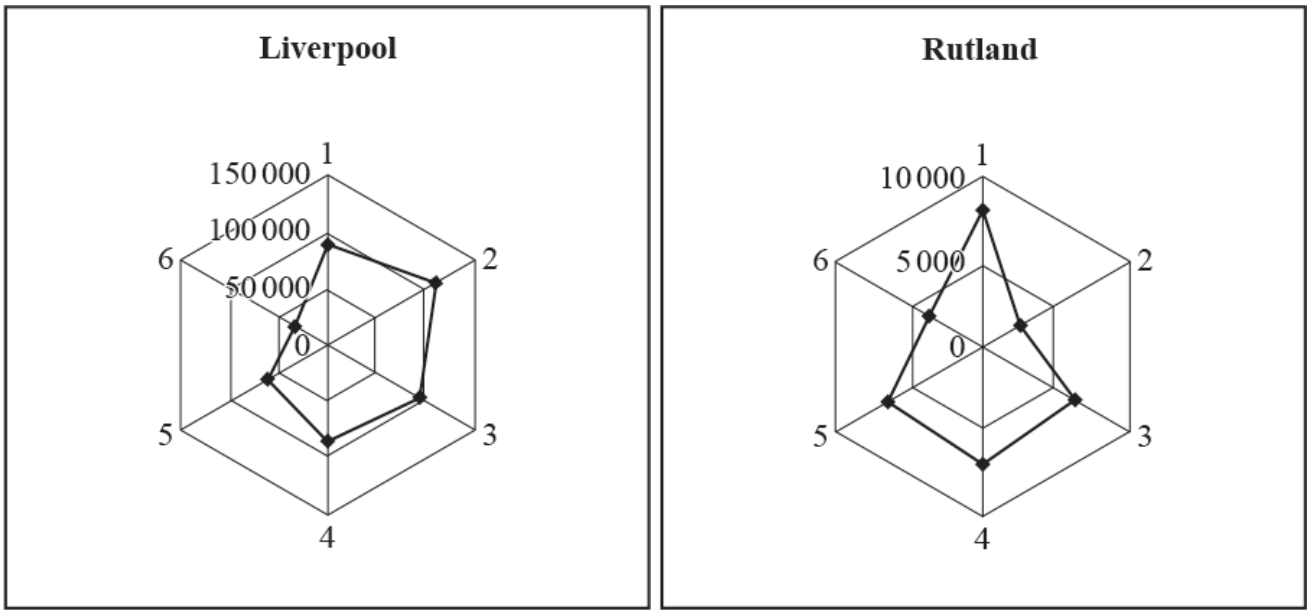
(i) Find the value of Σw^2. [2]

The weights, x grams, of another sample of 100 stones were found and were summarised as follows.

$$n = 100 \qquad \Sigma x = 5760 \qquad \Sigma x^2 = 335\,497$$

(ii) Calculate the mean and standard deviation of the weights of all 175 stones. [4]

18. The radar diagrams illustrate some population figures from the 2011 census results.



Each radius represents an age group, as follows:

Radius	1	2	3	4	5	6
Age group	0–17	18–29	30–44	45–59	60–74	75+

The distance of each dot from the centre represents the number of people in the relevant age group.

(a) The scales on the two diagrams are different. State an advantage and a disadvantage of using different scales in order to make comparisons between the ages of people in these two Local Authorities. [2]

(b) Approximately how many people aged 45 to 59 were there in Liverpool? [1]

(c) State the main two differences between the age profiles of the two Local Authorities. [2]

(d) James makes the following claim.

"Assuming that there are no significant movements of population either into or out of the two regions, the 2021 census results are likely to show an increase in the number of children in Liverpool and a decrease in the number of children in Rutland."

Use the radar diagrams to give a justification for this claim.

[2]

19. The table shows information, derived from the 2011 UK census, about the percentage of employees w of travel to work in four Local Authorities.

Local Authority	Underground, metro, light rail or tram	Train	Bus	Drive	Walk or cycle
A	0.3%	4.5%	17%	52.8%	11%
B	0.2%	1.7%	1.7%	63.4%	11%
C	35.2%	3.0%	12%	11.7%	16%
D	8.9%	1.4%	9%	54.7%	10%

One of the Local Authorities is a London borough and two are metropolitan boroughs, not in London.

(a) Which one of the Local Authorities is a London borough? Give a reason for your answer. [1]

(b) Which two of the Local Authorities are metropolitan boroughs outside London? In each case give a reason for your answer. [2]

(c) Describe one difference between the public transport available in the two metropolitan boroughs, as suggested by the table. [1]

(d) Comment on the availability of public transport in Local Authority B as suggested by the table. [2]

20. Using the 2001 UK census results and some software, Javid intended to calculate the mean number of people who travelled to work by underground, metro, light rail or tram (UMLT) for all 348 Local Authorities. However, Javid noticed that for one LA the entry in the UMLT column is a dash, rather than a 0. See the extract below.

Data extract for one LA in 2001			
Work mainly at or from home	UMLT	Train	Bus, minibus or coach
295	–	4	4

Javid felt that it was not clear how this LA was to be treated so he decided to omit it from his calculation.

(a) Explain how the omission of this LA affects Javid's calculation of the mean. **[1]**

The value of the mean that Javid obtained was 2046.3.

(b) Calculate the value of the mean when this LA is not removed. **[2]**

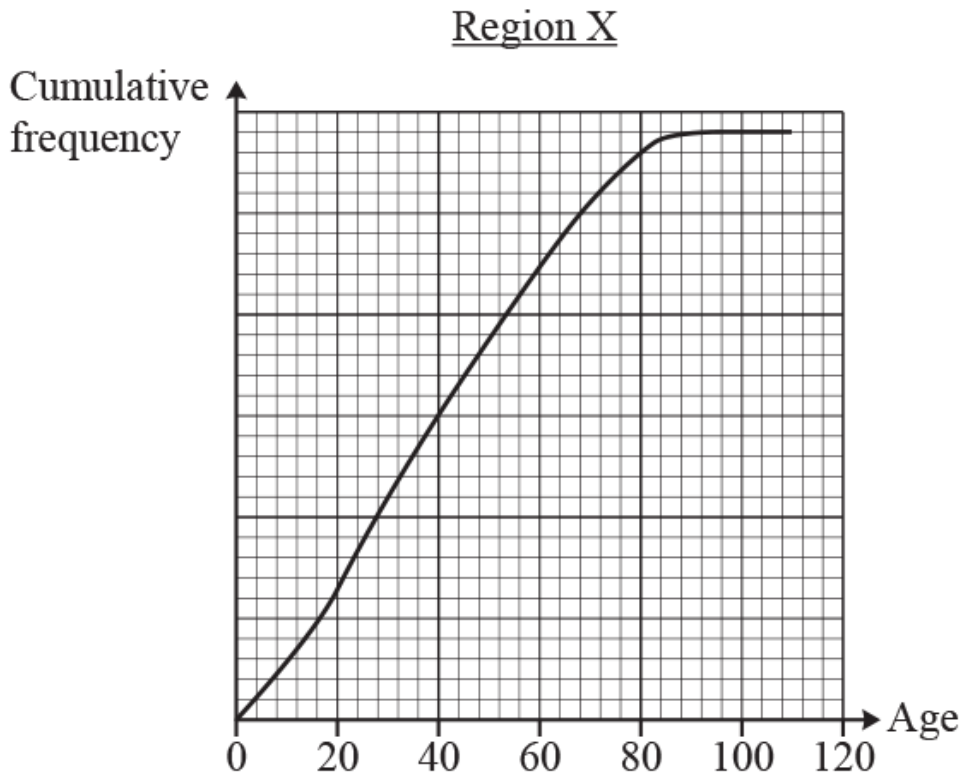
Javid finds that the corresponding mean for all Local Authorities for 2011 is 2860.8. In order to compare the means for the two years, Javid also finds the total number of employees in each of these years. His results are given below.

Year	2001	2011
Total number of employees	23 627 753	26 526 336

(c) Show that a higher proportion of employees used the metro to travel to work in 2011 than in 2001. **[2]**

(d) Suggest a reason for this increase. **[1]**

21. Paul drew a cumulative frequency graph showing information about the numbers of people in various age-groups in a certain region X. He forgot to include the scale on the cumulative frequency axis, as shown below.



(a) Find an estimate of the median age of the population of region X. [1]

(b) Find an estimate of the proportion of people aged over 60 in region X. [2]

Sonika drew similar cumulative graphs for another two regions, Y and Z, but she included the scales on the cumulative frequency axes, as shown below.

(c) Find an age group, of width 20 years, in which region Z has approximately 3 times as many people as region Y. [1]

(d) State one advantage and one disadvantage of using Sonika's two diagrams to compare the populations in Regions Y and Z. [2]

(e) Without calculation state, with a reason, which of regions Y or Z has the greater proportion of people aged under 40. [1]

22. The marks of 24 students in a test had mean m and standard deviation $\sqrt{6}$. Two new students took the same test. Their marks were $m - 4$ and $m + 4$.

Show that the standard deviation of the marks of all 26 students is 2.60, correct to 3 significant figures. [3]

END OF QUESTION paper

Question			Answer/Indicative content	Marks	Part marks and guidance	
1		i	Attempt find total area, (even if includes a^2) eg $20 \times 1.4a + 10 \times 3.4a + 6 \times 4.6a + 4 \times 2.6a + 10 \times 3a + 30a$ or 28a + 34a + 27.6a + 10.4a + 30a + 30a or $20 \times 1.4 + 10 \times 3.4 + 6 \times 4.6 + 4 \times 2.6 + 10 \times 3 + 30$ or $28 + 34 + 27.6 + 10.4 + 30 + 30$ or $7 \times 20 + 17 \times 10 + 23 \times 6 +$**.....** or $160a$ or 160 or 16 or $16a$ (if area, not ht)	M1	eg tot <u>area</u> = $16cm^2$ or $16a$ M1 '800/16 (= 50) M1 $a \times 10 = 50$ $a = 5$ A1 eg tot area = 400 (sqs) M1 800/400 (= 2) M1 $1.4a$ x20 = 70 × 2 $a = 5$ A1	Trial methods, eg: $a = 5$ gives $7 \times 20 + 17 \times 10 + 23 \times 6 + \ldots. = 800$ M1 But no of apples = 800 M1 Hence $a = 5$ A1
		i	800 ÷ their total (must involve area, not ht) eg $160a = 800$, 800÷	M1dep	Correct ans with nothing incorrect seen: M1M1A1 But where the correct answer clearly results from incorrect working, eg $a = 800/167 = 4.8$ rounded to $a = 5$, then max M1M1A0 **Examiner's Comments**	$a = 10$ gives $14 \times 20 + 34 \times 10 + 46 \times 6 +. = 1600$ M1 But no of apples = 800 M1 Hence $a = 5$ A1 NOT "1cm = 5" (because may just come from counting squares) <u>NB total ht = 16cm so if 16 seen, must clearly be area eg 800/16 may score 0 or 2</u>
		i	$a = 5$ "Box" ⇒ area. "Square" possibly ⇒ area	A1	This question gave rise to many different approaches, only some of which were valid. Many candidates appreciated the need to find the total area, but many did not appear to understand what units they were using. Some used cm^2 or small squares but others took the scale from the x-axis and assumed a scale for the y-axis. In many cases it appeared that candidates were not aware that they were actually choosing a scale in their calculation. Marking was generous and many candidates scored two marks for attempting to find the total area (in any units) and relating this correctly to the total frequency of 800. However, because of the muddle	

					over scales, few knew how to take the final step and find a. Some candidates considered only one block without considering the total area of the histogram. Others only considered the heights of the blocks rather than their areas. Many candidates used the range (80) and found 800 ÷ 80, which led to an incorrect method in almost every case.	
		ii	$\frac{1}{2}$ total area or $\frac{1}{2}$ total no. apples ft their 6(i)	Blf		Examples of correct methods: $400-(7\times20+17\times10)\quad(=90)$ $50+\frac{"90"}{23\times6}\times6\ =54$ $200-(70+85)\ (=45)$ $50+\frac{"45"}{69}\times6\quad=54$
		ii	Median is in 50 – 56 class stated or implied	M1		
		ii	Calculate (approx) $\frac{2}{3}$ of way along class or $\frac{1}{3}$ of way from top of class	M1		$400.5-(7\times20+17\times10)\quad(=90.5$ $50+\frac{"90.5"}{23\times6}\times6\ =54$
					Correct ans with nothing incorrect seen: M1M1A1 But where the correct answer clearly results from incorrect working, eg a = 800/167 = 4.8 rounded to a = 5, then max M1M1A0	
		ii	Median = 53.9 to 54 Not eg 54.2	A1	**Examiner's Comments** Few fully correct answers were seen. Most candidates found half the total area or frequency. Many identified the correct class (50–56) but some of these just gave the midpoint of this class as the median. Others tried to find exactly where in the class the median was situated, but only some of these could	Use of LB = 49.5: eg median $=49.5+\text{appr}\frac{2}{3}\times6=5$ B1M1A1

					handle the necessary proportion calculation. A few candidates found the mid-point of the range, giving an answer of 60. Examiner's Comments Most candidates answered this part well. A few omitted one or two digits, but the most common error was misalignment. Many candidates did not appear to appreciate that the shape of a stem-andleaf diagram is important. The lengths of the leaves show, at a glance, the general shape of the distribution of the data. Hence the alignment of the figures is important. For example, many candidates showed the 3rd digit in the first row clearly aligned with the 4th digit in one or more of the other rows. Also when candidates made an error and crossed out a digit, this usually resulted in misalignment when the correct digit was inserted. A few omitted the key. A small number drew a box-and-whisker diagram instead of a stem-and-leaf diagram.	Data Presentation and Interpretation
		Total		7		
2	i	2 | 0 4 7 3 | 2 3 4 | 0 5 7 9 5 | 3 5 5 6 8 6 | 2 5 7 9		B1	B1 for stem correct AND (3 branches correct OR 5 branches correct nos but incorrectly ordered)	Ignore "0" and / or "1" in stem, without leaves Allow incorrect alignment. Allow space instead of line. Allow left-facing diag
	i			B1	B1 for all correct	If all digits are in correct rows and orders, award this mark unless: 4th row is not the longest OR eg a 3rd digit in one row is clearly aligned with a 4th digit in another
	i	2 | 4 means 24 or similar		B1		

		ii	47.6 (3 sf) or $\frac{857}{18}$ or $47\frac{11}{18}$ (cm) oe	B1	cao ft wrong diag Examiner's Comments	eg 857 ÷ 18 = 41.6 B0 but $\frac{857}{18} = 41.6$ ISW B1
		ii	51 (cm)	B1ft	Most candidates answered this part correctly, although arithmetical errors in finding the mean were not uncommon. Also, for the median, a few candidates found the 9th value instead of the mean of the 9th and 10th. Follow-though from an incorrect diagram was allowed for the median.	
		iii	49 (or 9th no.) becomes 51	B1	No marks for identifying 49 & 53 alone or 51 & 55 alone Examiner's Comments	NB NO ft from wrong diag NOT eg '51 or higher'
		iii	or 53 (or 10th no.) becomes 55	B1	Many candidates correctly identified the 49 and 53 as possibilities for the incorrect value, but some gave incorrect replacements, most commonly 50 and 54. A few gave answers that suggested that they did not understand what a median is. Some candidates understood the instruction "Give two possibilities for the incorrect length …" to mean "Give two possible explanations for incorrect measuring". This gave rise to answers such as "The snakes moved while being measured."	Allow embedded answer eg 53 identified as incorrect and state (55 + 49) ÷ 2 = 52 scores 2nd B1
			Total	7		

3		i	$S_{xx} = 8700000 - \frac{7000^2}{6}$ $(= 533333)$ $S_{xy} = 509900 - \frac{7000\times456}{6}$ $(= -22100)$	M1	Correct subst'n in any correct *S* formula	
		i	$b = -\frac{"22100"}{"533333"}$ or $-\frac{663}{16000}$ $(= -0.0414)$	M1	Correct subst'n in any correct *b* formula from two correct *S* formulae	
		i	$y - \frac{456}{6} = "-0.0414"(x - \frac{7000}{6})$	M1	ft their *b* except if using *r*	or $a = \frac{456}{6} - ("-0.0414")\times\frac{7000}{6}$ oe ft "
					or $y = -\frac{663}{16000}x + \frac{3979}{32}$ or $y = -0.041x + 124$	
					Examiner's Comments	
		i	y = –0.0414*x* + 124 (3 sf)	A1	This part was answered very well on the whole. A few candidates made a sign error when substituting *b* (which is negative) in order to find *a*. Some simply lost the minus sign in *b*. Some found *b* and stopped. Others found *r* instead of what was asked. A few found r and then used this as their value of *b*.	Allow *y* = –0.04*x* + 124 if –0.041… seen
		ii	70 to 72	B1	or 71 per thousand, NOT 71000 **Examiner's Comments** Many candidates ignored the instruction to "use the regression line in the diagram" and used their equation from (i). A few candidates misinterpreted the situation, giving an answer such as 71 000.	No ft from **(i)** Ignore method
		iii	Extrapolation oe	B1	Allow "2400 is beyond graph" } "Not shown on the graph" or } "Line drops low, or below 0" } "Outlier" }	"Line only allows for countries poorer than Nigeria" 1st B1 Allow "Value for Nigeria is –ve 1st B1

		iii	Corr'n not high or small sample	B1	Poor corr'n oe, or pts not close to line oe 2nd B1 **Examiner's Comments** Most candidates gave one correct answer, using the word "extrapolation" or some equivalent wording. However, many either gave no second reason or gave one that was, in effect, equivalent to their first reason (eg "The IMR will become negative"). Some candidates gave the valid second reason, namely that the diagram does not show good linear correlation.	NOT "Other factors may apply" oe Ignore all else
		iv	$S_{xx} = 8700000 + 1300^2 - \frac{(7000+1300)^2}{7}$		or $10390000 - \frac{(8300)^2}{7} = \frac{3840000}{7}$	
		iv	$S_{yy} = 36262 + 96^2 - \frac{(456+96)^2}{7}$	M1	or $45478 - \frac{552^2}{7} = \frac{13642}{7}$ o	Correct sub in any correct S formula M1 Correct value of any S seen or implied by r A1
		iv	$S_{xy} = 509900 + 1300\times96 - \frac{8300\times552}{7}$	A1	or $634700 - \frac{8300\times552}{7} = -\frac{138700}{7}$ o	
		iv	$r = \frac{"-19814.3"}{\sqrt{"548571"\times"1948.86"}}$	M1	Correct subst'n in any correct r formula from 3 correct subs in 3 correct S formulae, ie all correct method **Examiner's Comments** This part was answered well by most candidates. A few used the original totals, just changing the value of n from 6 to 7. Others made the opposite error, finding new totals, but with n = 6. Sensibly, most wrote down their new totals such as $\sum xy$, but some were incorrect and, without any indication of method, these lost a method mark.	SC If n = 6, but otherwise correct allow M1A0M1A0 (ans r = –0.574, must see wking)
		iv	= –0.606 (3 sf)	A1		
		v	No effect oe	B1	Stay the same oe Allow just "No"	Ignore all else

					Examiner's Comments A few candidates thought that r would decrease because the values used in the formula would decrease, but most stated correctly that r would be unchanged.	Data Presentation and Interpretation
			Total	**12**		
4		i	574	B1	NB 0.13 seen within working; B0 **Examiner's Comments** Many candidates gave the correct mean, but many gave 5.13 for the standard deviation (or even 0.74 for the mean). Some divided 0.74 by 10 before adding it to 5. Others confused this question with questions involving finding the mean and standard deviation of two groups combined. These candidates tried to find Σx^2 by working backwards from σ = 0.13 and from there they tried to find the new standard deviation.	
		i	0.13 or 'the same'	B1		eg $\frac{\Sigma x^2}{10} - (\text{their mean})^2 = 0.13^2$ sco
		ii	(10 × '5.74' + 15 × 5.6) ÷ 25oe all correct	M1	eg $5.74 \times \frac{2}{5} + 5.6 \times \frac{3}{5}$ ft their 5.74	NB (5.74 + 5.6) ÷ 2 = 5.67 M0A0
		ii	= 5.656 = 5.66 (3 sf)	A1ft	**Examiner's Comments** Some candidates found the unweighted mean of the two means or simply added the two means.	NB 5.7 with no wking: M0A0 even if already penalised elsewhere for over-rounding
		iii	1st gp (or one gp) is more consistent (or less spread oe)	B1ft	2nd gp (or one gp) more accurate or etc but less consistent or etc	1st gp (or one gp) more consistent or etc 2nd gp (or the other gp) more accurate or etc

		iii	but less accurate	B1ft		Ignore all other, eg ignore 'Claim false' or 'Claim true' etc even if it contradicts other statements Reference to mean of all 25 does not score
		iii	(or mean further from true mean oe)		If neither B1 scored, but state 'consistency does not imply accuracy' or similar: SC B1 Equiv answers accepted, but no others	
		iii			<u>Examiner's Comments</u> There was some confusion here. Some candidates considered only the mean of all 25 people, stating that it was very close to the true mean and was therefore "consistent" in some sense. Some gave general answers such as "It is untrue because they are just guessing." Others only compared the means of the two groups, correctly noting that one was nearer to the true mass than the other. A few appreciated that the standard deviation represents the consistency of a group's guesses, then compared the standard deviations of the groups and gave a fully correct answer, although a few thought that a higher standard deviation means greater consistency.	Follow through their values for 1st gp: eg if 1st gp sd = 5.13: 1st gp less accurate and less consistent oe B1B1 Similar for other ft.
			Total	**6**		
5		i	$\frac{\Sigma fx}{\Sigma f}$ attempted $\left(= \frac{662000}{280900}\right)$	M1	3 terms of Σfx correct.. and $\div \Sigma f$ Allow incorrect Σf NOT Σx	Use of 5 or 6 instead of 5.5 for last value of x: all M-marks can be scored,but no A-marks. (ans: 5 gives 2.32 and 1.23; 6 gives 2.39 and 1.40) Use of 5 and 6 instead of 5.5 (probably with freqs 19400/2) could lead to correct mean M1A1, but possibly M1M1A0 for sd. $\div$ 5 or $\div$ 6 M0A0

	i	= 2.36 (3 sf)	A1			
	i	$\frac{\Sigma fx^2}{\Sigma f}$ attempted $\left(=\frac{2042350}{280900}=\right.$ 7.270737)	M1	3 terms of Σfx^2 correct and $\div \Sigma f$ Allow incorrect Σf NOT Σx	$\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}$ 3 terms of num correct and $\div \Sigma f$ M2 $(86900 \times 1.36^2 + 92500 \times 0.36^2 + 45000 \times 0.64^2 + 371001.64^2 + 194003.1^2)$, $\left(\frac{482210.64}{280900}\right)$ 2 terms of num correct and $\div \Sigma f$ M1 Allow incorrect Σf but NOT if $\Sigma = \Sigma x$	
	i	$-$ ″2.36″^{2} (= 1.70 to 1.72, 3 sf)	M1	dep +ve result $\div$ 5 or $\div$ 6 M0M0A0 allow 1.3	NB √ not requ'd for M1M1	
	i	s.d. = 1.31 or 1.30 (3 sf)	A1	**Examiner's Comments** Despite this question being in principle straightforward, candidates found many false paths down which to travel. Some simply made arithmetical errors. Others divided by 5 or 6 or by Σx, instead of Σf. Some used Σx instead of Σfx, and Σx^2 instead of Σfx^2. Some found the mean correctly, but used $\sqrt{(\Sigma fx^2 - \bar{x}^2)}$ for the standard deviation. A few found $\Sigma(fx)^2$ or $(\Sigma fx)^2$. The last class caused difficulty for some candidates. The instruction was to treat "5 or more" as "5 or 6", This led some candidates to find two values of fx for this class, one for 5 and one for 6. Some of these halved the class frequency, giving a plausible method. Others did not. A few failed to take the square root at the end. Candidates who attempted to find the standard deviation using $\Sigma(x - \bar{x}^2)$ became lost in the arithmetic. Probably the safest method, both for achieving correct answers and for enabling examiners to understand	Correct answer(s) without working score full marks

					candidates' working, is to complete a table showing the values of x, f, fx and fx^2, and the totals for the last three columns.	Data Presentation and Interpretation
		ii	2	B1	allow IQR = 3 – 1 = 2, ie UQ = 3 implied	Ignore working for both, even if
		ii	3	B1	**Examiner's Comments** Some candidates gave values taken from the second row of the table instead of the top row. Others tried to interpolate, giving answers such as 2.3 and 3.6.	Incorrect NB 3, 2 B0B0 unless labelled correctly
			Total	**7**		
6		i	Median = 7.45 (m)	B1	cao	
		i	IQR = 7.75 – 6.7	M1	allow 7.775 – 6.6 or 77.5 – 67 or 77.75 – 66 or 7.8 – 6.5 even though this is an incorrect method or 78 – 65	These pairs of values only, and subtract, for M1 eg

		i	= 1.05 (m) allow 1.175 or 1.18 NOT 1.3	A1	allow 10.5 or 11.75 or 11.8 but only if med = 74.5 **Examiner's Comments** Candidates used a variety of methods, many of them incorrect. In particular, the use of $\frac{n}{4}$ and $\frac{3n}{4}$ nstead of $\frac{n+1}{4}$ and $\frac{3(n+1)}{4}$ n finding the quartiles was a common source of error. The latter method, although correct, requires interpolation, which some candidates failed to do correctly. The most successful candidates used the simplest method for the quartiles. This method takes the lower quartile to be the median of the lower half of the values (excluding the middle value if there is one). The IQR, using this method, is found by 7.75 " 6.7 = 1.05. Some candidates misread the key and gave answers which were 10 times the correct ones.	7.45, 7.75 – 6.7 = 1.05 B1M 7.45, 7.775 – 6.6 = 1.175 B1M 7.45, 7.8 – 6.5 = 1.3 B1M 7.45, 7.7 – 6.5 = 1.2 B1M 7.45, 77.5 – 67 = 10.5 B1M 74.5, 77.5 – 67 = 10.5 B0M 74.5, 7.75 – 6.7 = 10.5 B0M 74.5, 77.75 – 66 = 11.75 B0M 7.45, 78 – 65 = 13 B1M 74.5, 78 – 65 = 13 B0M 74.5, 77 – 65 = 12 B0M
		ii	4 2 2 | 5 | 5 3 3 0 | 6 8 7 7 | 6 4 3 2 | 7 6 5 | 7 | 8 5 | 8	B1*	correct digits in correct leaves, ignore order, allow one omitted or extra or misplaced or incorrect digit	Allow a separate diag with leaves to left of stem. If only a separate diag is drawn, with leaves to right of stem: all correct including order, alignment and key: B1
		ii	Complete correct diag including order and key and alignment	B1 dep	key: eg 8|6|4 means 6.8 (*B*) and 6.4 (*A*) allow just 8 | 6 means 6.8 NOT 8 | 6 means 8.6 Allow 8 | 6 means 68, if consistent with (i)	If all digits are in correct rows and orders, & correct key, award this mark unless EITHER: 1. eg a 2nd digit in one row is clearly aligned with a 3rd digit in another OR 2. 1st, 3rd, 4th & 5th rows are very different lengths, eg because of crossing out and replacement

					Examiner's Comments Most candidates ordered the digits correctly, but many failed to align them properly. These candidates seemed unaware that one of the points of a stem-and-leaf diagram is to illustrate the general "shape" of the distribution, which depends upon the leaves being of the correct lengths, i.e. the digits being aligned correctly. In some cases, misalignment was caused by crossing out incorrect work and replacing it with correct digits, but in the wrong place. A few candidates, faced with the dilemma of how to align properly after crossing out, started a new diagram. This was acceptable, so long as the leaves for B were on the left of the stem. Wise candidates firstly drew a rough diagram on the left hand side of the answer space, and then gave their final version in the place expected. Many candidates gave the key incorrectly, not appreciating that for the digits in the left hand half of the diagram, 2 | 5 means 5.2, not 2.5. Others simply omitted the key. The digits 4, 2, 2 were often seen in the second row instead if the first.	
		iii	A higher overall A has more taller trees or fewer shorter A has higher median (mean, ave, medium)	B1	B shorter overall B has fewer taller trees or more shorter B has lower median (mean, ave, medium)	One correct comment on size: B1. One correct comment on spread or shape: B1. The following are examples only. Ignore any working; mark the statements only. Allow "First set" or "Right" for A, "Second set" or "Left" for B. NOT A higher than B NOT B has shorter trees than A Allow just quoting the two medians, even if wrong, so long as med of A is gter than med of B. Similarly if quote IQRs
		iii	B more evenly spread or distributed B more spread out B has larger range or IQR or sd Ranges of both are similar	B1	A less evenly spread or distributed A less spread out A has smaller range or IQR or sd Allow A's heights are more consistent	NOT any reference to outliers NOT any reference to sample size NOT any reference to indiv trees NOT two comments on size

			A is nearer to normal A is negatively skewed A has a (unique) mode, or modal class or peak; (B doesn't)		Not other comment about skew Ignore any other reference to mode or most common Ignore all else even if incorrect **Examiner's Comments** To be sure of gaining both marks in questions of this type, candidates should follow the following guidelines: 1 Always refer to the context. 2. Give answers that refer to the groups as a whole, rather than to individual values. 3. Give one answer about size and one about spread. Many candidates fell down on one or more of these criteria. While it is not absolutely impossible to gain the marks without adhering to these guidelines, it is extremely difficult to give a convincing answer that does not do so.	NOT two comments on spread eg highest on both is 8.5 B0
			Total	**7**		
7		i	35	B1	Allow 30 to 40 inclusive **Examiner's Comments** A generous range was allowed here (30 to 40), but a few candidates made errors in reading the graph and gave answers outside this range. The fact that the scales on the two axes are different may have confused some candidates, A few candidates just read off the cumulative frequency for a mass of 45 g. Others found the average of the two cumulative frequencies for 40 g and 45 g.	
		ii ii	$\frac{50\pm2}{400} \times 100$ oe = 12% to 13%	M1 A1	NOT $\frac{50\pm2}{450} \times 100$ **Examiner's Comments** Almost all candidates answered this question correctly. A few	NOT $\frac{100\pm2}{400 \text{or} 450} \times 100$ NOT $\frac{350\pm2}{400} \times 100$ (unless sub from

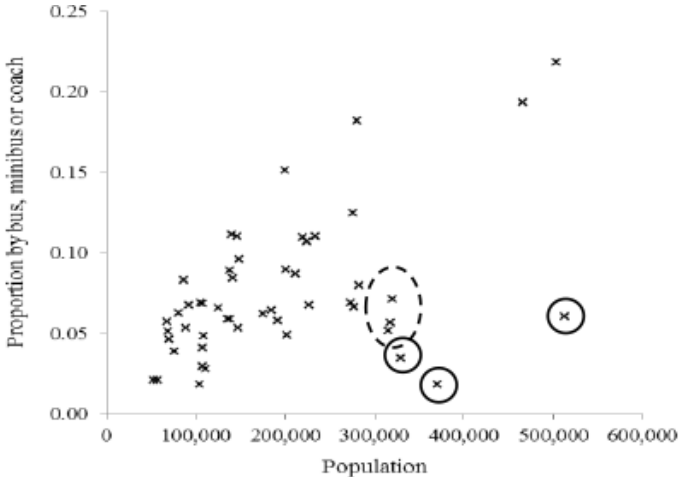
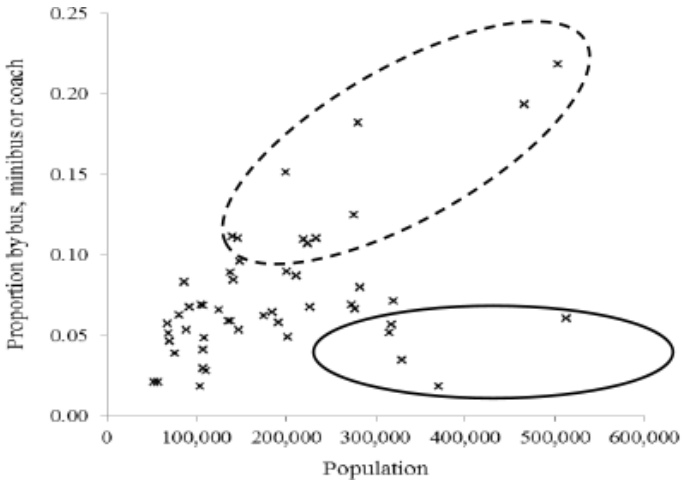
					omitted to subtract the cumulative frequency from the total (400). Others thought the total frequency was 450. Some found the correct decimal (0.125) but not the percentage.	Data Presentation and Interpretation
		iii	eg 7.5, 87.5 or 5, 90 or 5–10, 85–90	B1	or any values in intervals 5 – 10 & 85 – 90 No raw data given. Not have each data value Exact values not given or can't be read off oe Ignore all else for 2nd B1, **not 1st B1** **Examiner's Comments**	NOT “Because it's cumulative frequency”
		iii	“Classes” or “intervals” or “groups” or “mid-points” or “bounds” seen Data lost oe	B1	Many candidates thought that the highest mass was 100 g. Many incorrect answers were given for the reason why the exact values cannot be read off. Some simply restated the question (‘Exact values cannot be read off the graph.’). Other reasons included the following ‘Because it is a curve’; ‘Because it is cumulative frequency’;, ‘Masses start from 5 and level off after 80’; ‘Cumulative frequency graph does not show the range or the frequency;, ‘The resolution is too low’; ‘The scale is too small; and ‘Because we don't know the full range of masses’.	NOT “Because it's a line of best fit” NOT “Because graph is difficult to read” NOT “Because graph is a curve” NOT “Cont data has no exact data pts”
		iv	Median = 39 ± 1 drawn	B1	or stated	Mark diagram even if contradicts
		iv	Quartiles = 25 ± 1, 55 ± 1 drawn	B1	or stated	statements of values in (iv) or (iii)
		iv	Ends in ranges 5 – 10 & 85 – 901 drawn	B1f	or ft (iii) or ft (iii) mark intention (allow unruled lines)	If no diagram, award max B1B1B1 for
		iv	Correct B&W plot ± 11 drawn	B1f	**Examiner's Comments** The diagram was often correct, although large minority of	statements of med, quartiles & ends

					candidates thought that the median was above 40, perhaps because they assumed that the total was 450. Some candidates reduced their chances of marks by drawing free-hand, thus making it unclear at which values their lines were drawn. A few drew the maximum line at 100, even though the value for the highest mass, given in their answer to part (ii), was, e.g., 90 g or 88 g.	
		v	Stretched out at top end oe Not symmetrical More concentrated towards lower end More values (or data) in lower half of range Median closer to lowest value Average towards lower end More plums have lower masses Majority of distribution towards lower end More below 50 (or 45) Upper whisker longer than lower whisker	B1	Positive skew, Skewed to right (or to higher values) Larger skewness at top Larger plums more spread than smaller ones Ignore all else No need for context **Examiner's Comments** The requirement here was to note either the longer whisker at the top or the fact that there were more masses in the lower half of the range than in the upper part. Many candidates gave inadequate answers such as 'The spread is fairly even' or 'There is wide spread of masses' or 'The IQR is nearer the lower end' or 'The majority are between 25 and 55'. The concept of 'skew' is not in the specification, but candidates could gain the mark by stating that the data had positive skew. Many stated (wrongly) that there was negative skew. 'Positive correlation' was not infrequently seen.	**NOT** any of below: more large extremes than small extremes IQR is towards the lower end skewed to the left (or to lower values) majority below 39 distribution towards lower end
			Total	**10**		
8		i	$\frac{-182}{52}$ or –3.5 seen or implied	B1	NB in (i) and (ii) $1768 + 150^2 \times 52 = 1171768$ is incorrect and scores no marks in either part, except possible ft in (ii).	$\sum m = 150 \times 52 - 182$ or 7618 B1

		i	Mean = 150 – "3.5"	M1	or $\frac{-182+150\times52}{52}$ or $\frac{7800-182}{52}$ B1M1	Data Presentation and Interpretation "7618" ÷ 52 M1 = 146.5 A1
		i	= 146.5 or 147	A1		
		i	$\frac{1768}{52}-("-3.5")^2$ alone, eg not if + 150	M1	Allow within √ sign	or $\frac{'1117168'}{52}-'146.5'^2$ fully correct method M1 = 21.75 A1 Examiner's Comments Most candidates found the mean of m correctly, although a large minority only found the mean of (m – 150). The variance caused problems for many candidates. Only a minority appreciated that they only needed to find the variance of
		i	= 21.75 or 21.8	A1	Not ISW, eg $\sqrt{21.75}$ (or 4.66) M1A0 ans 4.66, no working, M1A0 NB $\frac{1768}{52}-"146.5"^2$ or 1768" –('–3.5^2')M0A0	(m – 150). Many tried to "uncode" the data, for example by finding $\frac{1768}{52}+150$ (or even $\frac{1768}{52}+150^2$) before subtracting either the (mean of m)^2 or (the mean of (m – 150))^2. Some candidates mixed coded and uncoded values, giving $\frac{1768}{52}-"146.5"^2$ or 1768 –('–3.5^2'). Some attempted firstly to find $\sum m^2$. But most of these candidates appeared not to understand the meaning

						of the "$\sum$" sign and made little or no progress. Some candidates gave $\sum(m-150)^2 = 1768$, which is correct, but continued with working such as $\sum m^2 - 300m + 150^2 = 1768$. A strange, but not uncommon, error in the variance calculation was $1768 - \frac{(-3.5)^2}{52}$
		ii	$\frac{\Sigma m^2}{52} - "146.5"^2 = "21.75"$ or $\sum m^2 = ('21.75' + '146.5^2') \times 52$ ft their mean & +ve var from (i) for M2	M2	Allow M1 for $\frac{\Sigma m^2}{52} - "3.5"^2 = "21.75"$ or $\sum m^2 = ('21.75' + '3.52') \times 52$	$\sum(m-150)^2 = 1768$ $\sum m^2 - 300\sum m + 150^2 \times 52 = 1768 \geq 2$ terms correct M1 $\sum m^2 = 1768 + 300 \times "7618" - 150^2 \times 52$ correct method M1 $= 1117168$ A1 Correct ans, no wking M1M1A1 If incorrect ans given with no wking, possibly M1M1 for (ii) may be obtained by correct method seen in (i), However M1M0 or M0M0 is more likely.
		ii	$\sum m^2 = 1117168$ ISW	A1	Exact; no ft from (i) eg 147 or 21.8	**Examiner's Comments** Only a few candidates made any progress in this part. Some tried to find $\sum m^2$ from $\sum(m-150)^2$, but most did not know how to handle the $\sum$ sign.
		iii	(52 + 1) ÷ 4 = 13.25 or (26 + 1) ÷ 2 = 13.5 (⇒ 13th apple has mass < 140)	M1	The correct method is in the 1st column. However, most candidates will give the allowed method in the middle column and score both marks. NB 3rd column Allow 52 ÷ 4 or 26 ÷ 2 (= 13) M1	Allow 52 ÷ 4 or 26 ÷ 2 (= 13) M1 (⇒ 13th apple has mass 140) ⇒ (no. below 140 =) 12 A0
		iii	⇒ (no. below 140 =) 13	A1	⇒ (no. below 140 =) 13 A1	**Examiner's Comments** Many candidates saw the point and recognised that 140 is the lower quartile. From that point, most candidates just divided 52 by 4 to give an answer of 13. These candidates

						gained full marks. However, the correct method involves finding $\frac{52+1}{4}$ or $\frac{26+1}{2}$, which leads to the conclusion that the lower quartile lies between the 13th and 14th values and hence there are 13 values below the lower quartile. Many candidates used a wholly incorrect method using "scaling", such as $\frac{52}{46} \times 10 = 11$
		iv	IQR = 15 seen or implied	B1	or 22.5 seen or implied	
		iv	155 + 1.5 × 15 = 177.5 (or > 176) or 140 – 1.5 × 15 = 117.5 (or < 130)	B1	176 – 155 = 21 (or < 22.5) or 140 – 130 = 10 (or < 22.5)	$\frac{176-155}{15} = 1.4$ (or <1.5) or $\frac{140-130}{15} = \frac{2}{3}$ (or <1.5) Equivalent correct methods may be seen For 2nd B1 allow 14 ≤ IQR ≤ 16
		iv	No outliers	B1	Ignore method	**Examiner's Comments** Many candidates answered this question correctly. However, some misread the description of outliers given in the question and found, for example, 1.5 × the upper quartile. Many other candidates just gave a verbal answer with little or no calculation to support it. These generally gained no marks or possibly just one mark. A few candidates quoted the convention that outliers are indicated by dots on a box-and-whisker plot. The wording of this question meant that these candidates could not score more than one mark.
			Total	**13**		

9		a		B1 (AO2.2b) [1]	At least the three with solid rings. No extras other than those in the dashed ring.	
		b	 e.g. the dotted ringed group are “metropolitan districts” which have good infrastructure, so they have high proportions of travelling by bus. The solid ringed group are probably large “unitary authorities” which are not urban, so they don’t have good bus services. The unringed points are a mix of small “unitary authorities” and “non-metropolitan districts” which are difficult to tell apart with these data.	B1 (AO2.2b) E1 (AO1.2) E1 (AO2.3)	For identifying (not necessarily using the diagram) the two subpopulations shown as being one in which there is a positive correlation between the two variables, and one in which larger populations do not appear to lead to increases in the proportion	Identifying *some* points of those ringed as being in different sub-populations

				[3]	travelling by bus. For identifying two distinct subpopulations in terms of the structure of the large data set For explaining why it might be difficult to tell the others apart.	
			Total	4		
10		a	She has assumed that any car has exactly two people in it: one passenger and the driver. Subtract the value in "Passenger in a car or van" from the value in "Driving a car or van" to get the number of people driving alone.	B1 (AO2.2b) B1 (AO2.2a) [2]	Must refer to "Driving a car or van", or equivalent	
		b	i) The proportion using individual motorised transport in region B (56.2) is greater than region A (49.3)	B1 (AO2.3)	Or other valid reason taken	

				[1]	from data	
		b	ii) The proportion using no motorised transport in region B (23.1) is greater than region A (21.5)	B1 (AO2.3) [1]	Or other valid reason taken from data	
			Total	**4**		
11		a	e.g. From the data given, the proportions of men who cycle to work show much more variability than women, with greater proportions of younger men cycling than older men.	E1(AO2.4) [1]		
		b	The proportion decreased e.g. These workers were in the 40-44 group in 2011, which is a smaller proportion of the population than the 30-34 group in 2001.	B1(AO2.2a) B1(AO2.2b) [2]		
		c	e.g. The age group is still approximately the same size in 2011 Very few (or no) males in this age group join the workforce between 2001 and 2011 Very few (or no) males in this age group leave the workforce between 2001 and 2011 The overall size of the workforce in this age group has not changed much The sample is representative of the whole population	B1(AO2.2b) [1]	For any relevant assumption	
			Total	**4**		

12		a	E.g. The only region with very low location on both variables is Region D which is therefore London. E.g. The region with the lowest standard deviation is Region B, so this is Wales E.g. The only value where the other two differ much is sd of *driving*; the wider spread on Region C including the outlier suggests that this is the Southwest, so Region A is the South East.	E1(AO2.2a) E1(AO2.2a) E1(AO2.2b) [3]	Or any other valid reason to connect Region D with London Or any other valid reason to connect Region B with Wales Careful argument involving mean and / or standard deviation	OR E1 for one region correct with good reasoning OR E2 for two regions correct with good reasoning	Data Presentation and Interpretation
		b	E.g. No the data only shows that this LA has low proportions of car use for travelling to work. E.g. No, many LAs in Region D (London) have similar proportions and they are not small islands.	E1(AO2.2b) [1]	Or any other valid explanation of why the data given is insufficient to draw this conclusion	Identifying the LA as the Scilly Isles is not relevant; this requires information that is not in the supplied data.	
		c	E.g. On a large island, methods of travel to work are unlikely to be different to any other LA; people will still be travelling to work on the roads, and provision of public transport will be similar to any other LA.	E1(AO2.2b) [1]	Or any other valid explanation of how large islands are likely to have similar patterns of method of travel to other LAs	Candidates may, but need not, identify the LA as Anglesey, but this is not sufficient to award the mark	

			Total	5	Data Presentation and Interpretation		
13		a	45	B1(AO3.1b) [1]	Allow 44 years 364 days or similar		
		b	$\bar{x} = 39.6$ (3 sf) s = 24.3 (3 sf)	B1(AO1.1) M1(AO1.1) A1(AO1.1) [3]	(UCB = 106 because oldest is 105) (If seen) sub in correct formula using any x's within classes Allow 24.2 (3 sf)	If use 105 as UCB: mean = 39.5 (3 sf) B0 sd = 24.2 (3 sf) M1A1	
		c	39.6 + 2×24.3 = 88.2 39.6 – 2×24.3 = –9 Hence <u>may be</u> outliers at top, but not at bottom	M1(AO1.1a) A1f(AO2.2b) [2]	Allow just 39.6 + 2×24.3 = 88.2 for M1 A1 for both limits and full conclusion	ft their $\bar{x}$ and s ft their $\bar{x}$ and s	
		d	No, most in 65+ class will be nearer the lower end oe	E1(AO3.2b) [1]	or imply class is weighted towards left		

			Total	**7**	Data Presentation and Interpretation	
14		a	Identify points on (or close to) *x*-axis, & those not	**B1(AO1.1)** **[1]**		
		b	Some (or 5 or 6) areas have no metro etc	**E1(AO2.2a)** **[1]**	In some (or 5 or 6) areas, none use metro / Any equivalent description, in context	
		c	**i) Either** few use metro so "no metro" group **Or** some use metro so in group with metro	**E1(AO2.2a)** **[1]**	or no metro in area so “no metro” group	
		c	ii) Probably travel to diff area to get metro	**E1(AO2.4)** **[1]**		
		d	Unlikely. Large prop use metro etc	**E1(AO2.2b)** **[1]**	Unlikely. There is metro etc.	
			Total	**5**		
15		a	(i) "0 to 17" means 0 ≤ age < 18	**E1(AO 1.2)** **[1]**	Allow "17" means up to 17 yrs, 364 days / or any correct	

		a	(ii) Original class had no UCB but for histogram an UCB was needed. Few people live > 100	E1(AO 1.1) [1]	or other sensible reason		
		b	B has more aged 85 - 100 than A, which has larger total B & C have similar totals, and there are more aged 80 - 100 in B B	E1(AO3.1a) E1(AO2.4) B1(AO2.2a) [3]	dep on E1E1 earned		
		c	A: 31.5; B: 47.5; C: 39.5 A has greatest proportion of area towards left B has greatest proportion of area towards right	E1(AO 2.2a) E1(AO 3.1a) E1(AO 2.4) [3]	or similar correct explanation		
		d	Not justified B has higher mean than A But could be caused by older people moving away from A, or moving into B		or B has higher proportion of older than A		

				B1 [1]	or eg people move to B to retire	or other sensible
			Total	9		
16		i	530 (± 5)	B1 [1]	**Examiner's Comments** In this question some tolerance was allowed in reading the graph, but a few candidates lost marks through misreading the scale on either or both axes. A few candidates gave the answer 600.	
		ii	$\frac{20}{100} \times$ their 530 (= 106) Read graph at cf = their 530 – their 106 Min mk = 34 (± 1)	M1 M1 A1 [3]	May be implied by ans or mark on graph seen on graph or implied by correct ans cao If ans in range ignore wking, M1M1A1	0.8 × their 530 (= 424) Read graph at cf their 424 ± 10 Not nec'y integer If ans not in range and 1st M1 scored, 2nd M1 can be scored only by mark drawn on

	graph from their 424 ± 10

Examiner's Comments

Most candidates answered this question correctly. A few read the graph from 106 instead of from 424.

iii

Type 1 answer		Type 2 answer	
Individual marks unknown or Data is in classes or groups or ranges or Upper bounds used 'Classes' or 'groups' may be implied eg by "between"		No incr in freq above *a* Curve not incr above *a* Curve stops incr at *a* Curve stops incr at *a* Horiz or level or stnry or plateaus from *a* Line horiz before *a* Curve does not reach *a*	where $54 \leq a \leq 55$
Hiest in class 50 – 54 or between 50& 54	B1		eg Hiest mk between 54 and 59 B1B0
Allow 50 – 55 or 49.5 – 54.5	B1	Highest mk is ≤ 54 Allow ≤ 55	eg Hiest mk is in class 55–59 B1B0 Ignore all else
	[2]		

					The two B-marks are independent	

Examiner's Comments

Many candidates showed that they did not really understand the nature of a cumulative frequency graph. Some stated that, since the graph goes up to 60, the highest mark was 60. Some recognised that, because the data is grouped, the highest mark cannot be precisely identified, but some of these went on to say that the highest mark was in the 55 - 59 class. Some candidates recognised that the fact that the curve became flat at about 54.5 meant that no marks were higher than this, but many of these went on to say that the highest mark was 54, (or 54.5, or 55). A few candidates said that the highest mark was between 50 and 59.
Some said that there was only one highest mark, so the teacher could not be correct in saying that it was 54 or 55. Most of these candidates went on to say that the highest marks was 54.5.

		i v	Steepest part of graph oe or Slope most verticalor similar 25 – 29	B1 B1	or Greatest increase in cf or Increases by largest amount or Greatest frequency oe (dep on 25–29) Allow 25 – 30	NOT Greatest cum freq NOT Most students are in this class Ignore all else The two B-

				[2]	marks are independent **Examiner's Comments** Most students gave the correct answer of 25 – 29, generally with a correct reason such as that the graph is steepest here or that the increase in cumulative frequency is greatest here. A few gave an incorrect reason, such as "The cumulative frequency is greatest in this class". Some candidates thought that the mode was where the mark (rather than the frequency) was highest, and so gave the answer 55 – 59. A few candidates found the class which contained the median, rather than finding the modal class.	
			Total	8		
17		i	$5.8^2 = \frac{\Sigma w^2}{75} - 52.3^2$ $\Sigma w^2 = 207669.75$ or $\frac{830679}{4}$ oe	M1 A1 [2]	or $5.8 = \sqrt{(\frac{\Sigma w^2}{75} - 52.3^2)}$ Allow 208000 with correct working, no errors seen NOT other ans that rounds to 208000 **Examiner's Comments** A common error was to omit to square 5.8 or 52.3 or both. A	

					few candidates confused Σ with E and thought that Var(X) = $\Sigma x^2 - (\Sigma x)^2$. Many others thought that $\Sigma w^2 = (\Sigma w)^2 = (75 \times 52.3)^2$.	
		ii	$\text{mean} = \frac{75\times52.3+5760}{75+100}$ = 55.3 (3 sf) $\text{var} = \frac{"207\,669.75"+335\,497}{75+100} - "55.329"^2$ (= 42.5…..) sd = 6.52 (3 sf)	M1 A1 M1 A1 [4]	or $\frac{3922.5+5760}{175}$ or $\frac{9682.5}{175}$ or $\frac{543166.75}{175} - "55.329"^2$ Allow 6.51art 6.52 or 6.51	$\frac{\text{Their(i)}+335\,497}{75+100} - (\text{their mean of } 175)^2$ NB ans 6.76 prob'y from mean = 55.3 M1A1M1A0 but check wking NB May see 55.3 used in sd calc'n, but correct sd given (6.52). This gets full marks on the assumption

					that although candidate wrote “55.3” she used more sig figs in the calc’n	

Examiner’s Comments

Many candidates found the mean and standard deviation of the second sample of stones alone. Some found the unweighted mean of the two means.

			Total	**6**		
18		a	In all parts, once mark gained, ignore all else **Advantage: Type 1 answers:** **State or imply compare proportions (or distributions or structure or profile or pattern)** Examples: Can comp proportions (or distributions or structure or profile) Allow can see props Can compare areas' age groups relative to size of area Easier to see age groupp distributions **Disadvantage: Type 1 answers: State or imply pop sizes not <u>easy</u> to compare** Examples: Diag does not show relative sizes of the authorities	E1 (AO1.1) E1 (AO1.1)	Allow eg "Group 1" for 0-17s etc. **Advantage: Type 2 answers: State or imply with same scale, sizes of diags wd be very different** Examples: Prevents diag from becoming too big or too small to use effectively	Allow "children" for 0-17s **NOT** e.g: Easy to compare large area with small Easier to see results Easy to compare populations Because L is	

			R'd appears to have more in 0-17, but actually L'l has more in this group Hard to compare because diff nos rep by same size on diags Can't compare numbers (or results or pops or sizes) easily Can't compare numbers (or results or pops or sizes) without calculation	[2]	If one set of values is a lot lower than the other, it will be hard to compare them on the same scale. **Disadvantage:** **Type 2 answers: State or imply mismatch between diag size and pop size** Examples: Confusing because same size diag for diff size populations Looks as if same no. of people in each Might miss the fact that scales are diff, looks as bigger than R Can compare age in small & large areas **NOT eg** Can't compare results Can't compare numbers Easy to be mistaken when comparing	

					if more 0-17s in R'd than L'1	

Examiner's Comments

Many candidates gave correct answers here, although frequently they used far more words than were required. Some gave inadequate answers such as (for the advantage) "It's easier to see the results", "It's easier to compare the populations" or "Can compare ages in small and large populations". For the disadvantage some inadequate answers were "Can't compare results"; "Easy to make a mistake".

Many candidates wrote long essays which did not necessarily gain any marks.

Examples of brief, acceptable answers are:

Advantage: The different scales make it easy to compare the age group proportions in Liverpool with those in Rutland.

Disadvantage: The different scales make it hard to compare the age group numbers in Liverpool with those in Rutland.

		b	90000. Allow between 75000 & 95000 incl.	B1 (AO2.2b) [1]	Allow reasonable ans given as range eg "Much more than 50000 but < 100000"	

Examiner's Comments

					Most candidates answered this question correctly. A few wrote 100 000 even though the dot is clearly below this value.	Data Presentation and Interpretation
		c	"L" = Liverpool. "R" = Rutland NB: Must be about 60-74s and/or 18-29s and/or 0-17s Answer type 1 Compare <u>proportions</u> in two age groups. Examples: Any two of eg: L has smaller prop of 60-74 (than R'd) L has smaller prop of 0-17s (than R'd) L has larger prop of 18-29s (than R) eg, L prop of 18-29s is 4 × R prop 18-29s R has smaller prop of 18-29s R has hier prop of 0-17s	E1 (AO2.2b)	Answer type 2 Compare gps with largest (or smallest) props. Allow "number" instead of prop only for this type of answer Examples: L's hiest no. (or mode) is 18-29s AND R's hiest no. (or mode) is 0-17s E1 only L'ls smallest is	Answer type 3 Comp <u>props</u> in same age gps Examples: L has high prop 18-29s AND R has low prop 18-29s E1 only	

				E1 (AO2.2b) [2]	75+ AND R's smallest is 18-29 E1 only (75+ allowed in this case only) NOT "number" except in ans about modes or smallest. Ignore all else.	R has high prop 60-74s AND L has low prop 60-74s E1 only **NOT** eg L has **more** 18-29s than R
					<u>Examiner's Comments</u> Again, many answers were inadequate, such as "Rutland has a higher proportion of middle aged people". Answers that compared numbers (as opposed to proportions) of people in the two areas were not accepted. For example, "Liverpool has more people in the 18-29 age group than Rutland.	
		d	Must state gp who are likely to have babies ie 18-29s or 30-44s or 18-44s. (Allow 0-29s or "young") This gp is large in L, AND is small in R	E1 ind (AO2.4) E1dep (AO2.4) [2]	Inadequate ans: L high prop of young, who will have babies E1 R high prop of old E0	Allow "number" instead of "proportion" NOT just This gp is larger in L

					<u>Examiner's Comments</u> Most candidates recognised the key point here, which was that children are being born during the 10 years from 2011. However, some only stated that Liverpool had a large proportion of people in the potential child-bearing group, and failed to state that Rutland has a small proportion in this group. Some candidates just argued about the sizes of the groups changing as people get older, without considering new births at all. An example of an acceptable answer is: People in age group 18-29 (and/or 30-44) are likely to give birth to children during the next 10 years. There is a high proportion of this group in Liverpool, but a low proportion in Rutland. Acceptable answers to Question 13 parts (a), (c) and (d) can be found in the published mark scheme. Throughout this question, a common error was to write about "numbers", rather than "proportions", of people in the two areas.	
			Total	**7**		
19		a	E.g. C is a London borough because it has a (very) large percentage of metro	**E1 (AO2.2b)** **[1]**		
		b	E.g. D is a metropolitan borough because it has a large percentage of metro E.g. A is a metropolitan borough because it has a large percentage of metro train	**E1 (AO2.2b)** **E1 (AO2.2b)** **[2]**	or bus	

					or bus	
		c	E.g. D has some sort of metro, but A does not	E1FT (AO2.2b) [1]	or A has more trains (or buses) than D.	FT their two LAs from (ii)
		d	E.g. It looks like there is very little public transport available because a very small percentage of people use bus or train.	E2 (AO2.2b) (AO2.2b) [2]	E1 for very little available E1 for reference to supporting evidence	
			Total	**6**		
20		a	Wrong total	E1 (AO 3.2b) [1]		
		b	2046.3 × 347 ÷ 348 = 2040.4	M1 (AO 1.1a) A1 (AO 1.1) [2]		
		c	<table><tr><th>2011</th><th>2001</th></tr><tr><td>2860.8 × 348 ÷ 26526336</td><td>2040.4×348÷ 23627753</td></tr><tr><td>= 0.0375</td><td>= 0.0301</td></tr></table> Hence 2011 higher	M1 (AO 2.4) A1 (AO 2.2a) [2]	M1 both calculations A1 both	or without "× 348" <table><tr><td>= 0.0000789</td><td>= 0.000086803</td></tr></table>

					results and conclusion	Data Presentation and Interpretation
		d	More metros built.	E1 (AO 2.2b) [1]	or sensible alternative	
			Total	**6**		
2 1		a	38 to 39	B1 (AO 3.1a) [1]		
		b	eg $\frac{1.25}{5.75}$ or $\frac{7}{29}$ = 0.2 to 0.24	M1 (AO 3.1a) A1 (AO 1.1) [2]	Use heights, any units, eg cm or squares	
		c	eg 40 to 60	B1 (AO 1.1) [1]	Any correct range	
		d	Can easily compare proportions or age profile Cannot easily compare numbers in age groups	E1 (AO 2.2b) E1 (AO 2.2b) [2]		
		e	Z. Graph steeper for below 40	E1 (AO 2.2b) [1]		
			Total	**7**		

2 2			$\frac{\Sigma(x-m)^2}{24} = (\sqrt{6})^2$ $\Sigma(x-m)^2 = 144$ Mean of all 26 is m $S^2(26) = \frac{144+2\times4^2}{26}$ $= \frac{176}{26}$ or $\frac{88}{13}$ $S = \sqrt{\frac{88}{13}}$ (= 2.60 (3 sf) **AG**)	M1 (AO 3.1a) M1 (AO 1.1a) A1 (AO 1.1) [3]	$\frac{\Sigma x^2}{24} - m^2 = (\sqrt{6})^2$ $\Sigma x^2 = 144 + 24m^2$ $S^2(26) = \frac{144+24m^2+(m-4)^2+(m+4)^2}{26} - m^2$ $= \frac{176}{26}$ or $\frac{88}{13}$ Must see $\sqrt{\frac{88}{13}}$ or equivalent and ans 2.60	
			Total	3		