AQA

A Level

A Level Mathematics

Cleaning Data (Answers)

Name:

M M E

Mathsmadeeasy.co.uk

Total Marks:

L4- Cleaning Data- Answers

AQA

1) **You have a dataset containing one million individual customer records. You are concerned with the average time, in minutes, a customer must wait to have their call answered. A snapshot of the data is recorded a spreadsheet, shown below.**

ID	Name	Age	Postcode	Time Rang	Time Answered
1	B Smith		S12 3AW	10:15	10:17
2	J Haq		N1 3JW	14:22	14:22
3	C Brook		B4 9LP	12:45	13:01
…	…	…	…	…	…
1,000,000	A Tandem		NG16 1AL	09:02	09:07

i) **Write a method for obtaining the average time a customer waited.**

[1 mark for any of the methods correctly stated- 1 max]

The average can be given by some variation on the following.

The mean, which is

$$\mu = \frac{\Sigma\,(Time\ Answered - Time\ Rang)}{1{,}000{,}000}$$

or the median, involving finding the middle data point

$$M = \frac{1{,}000{,}000 + 1}{2}$$

of the matrix

$$[Time\ Answered - Time\ Rang]_M$$

or the mode using a frequency table and finding the maximum value

Minutes	Frequency
0	
1	
…	
n	
Maximum (F)	

The range for waiting time is 49 minutes, the median is 2 minutes and the mean is 11 minutes.

ii) State the longest time a customer had to wait.

[1 mark]

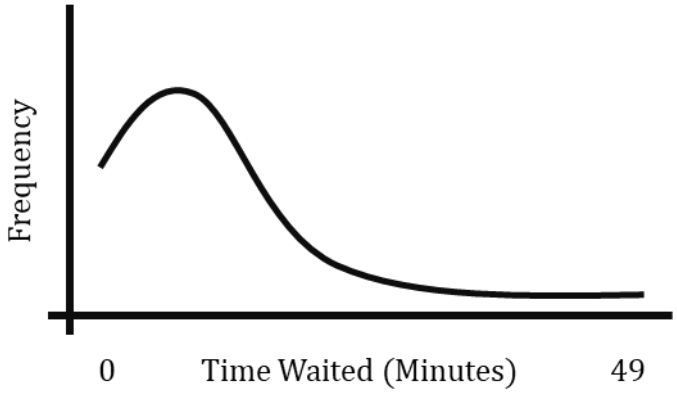
The range is (longest time – shortest time). We know that shortest time was 0 minutes, as it is shown in the table above. Making the longest time 49 minutes.

iii) Sketch the distribution of waiting times.

[1 mark for correct x-axis]

[1 mark for positive skew]

This graph must have positive skew to reflect the fact that the mean is greater than the median. The limit of the x-axis must be from 0 to 49 minutes.



Owing to a virus, some of the values between Time Rang and Time Answered might have been switched.

iv) Suggest a method of identifying these records that does not involve looking at every row.

[1 mark for any method which involves identifying negative time]

The switched data essentially means that the phone was answered before it rang, giving negative time. Any plot or sort would identify these.

v) A scatter graph has been produced (x-axis is Time Rang and y-axis is Time Answered). How could you use this to identify the erroneous data points?

[1 mark]

Any point below the line y=x would indicate that the phone was answered before it rang.

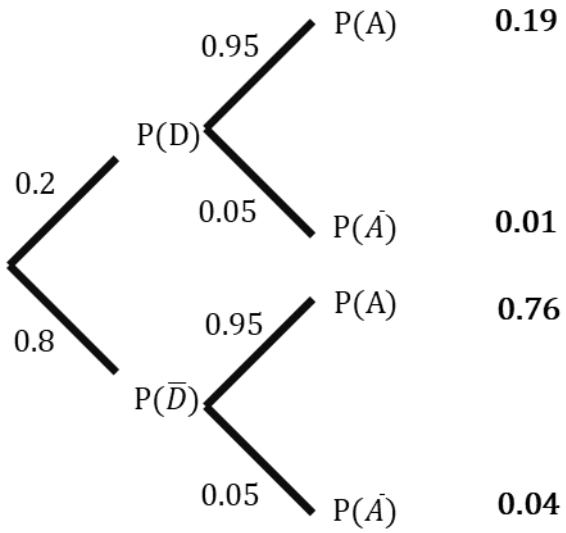
2) **A test for a disease in blood is 95% accurate (A), regardless of whether is the result is positive or negative. Only one in five people have the disease (D).**

i) **Draw a tree diagram showing the four possibilities and calculate probabilities of them occurring.**

[1 mark for each correct final probability – 4 max]

[1 mark for diagram showing correct method]

ii)



iii) **State on the tree diagram which possibility is a Type I error and which is a Type II error.**

Define the null hypothesis:

The person is disease free $\overline{D}$.

[1 mark]

A Type I error is rejecting the null hypothesis when it is true. In this instance, it is given along the branch that is $P(\bar{A}|\overline{D}) = 0.04$.

[1 mark]

A Type II error is accepting the null hypothesis when it is true. In this instance, it is given along the branch that is $P(\bar{A}|D) = 0.01$.

iv) **In words, and using the aforementioned context, describe what a Type II error is.**

[1 mark]

A Type II error is the hospital saying the individual is disease free, when they have the disease, because the test returned the wrong result.

3) The large dataset contains information about the amount(g) of pickle eaten per week per person in households in the South East and South West. This subset of data is shown ordered below along with the calculated quartiles one (25%), two (50%) and three (75%).

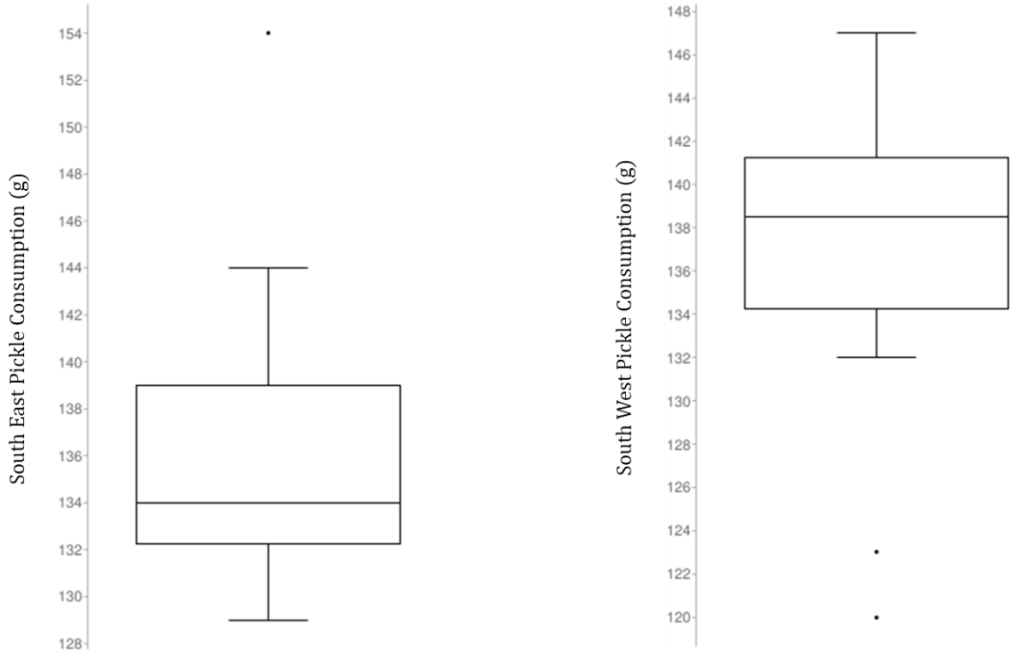
	South East	South West
	129	120
	130	123
	130	132
	133	135
	134	136
	134	138
	134	138
	134	139
	136	139
	138	139
	139	141
	139	142
	144	146
	154	147
Q1	132.3174605	134.4123184
Q2	134.1068108	138.5526584
Q3	138.8031286	141.4838487

i) **Draw a box plot, with outliers (if necessary) for each region.**

[1 mark for correct box drawn on each diagram- 2 max)

[1 mark for correct whiskers drawn on each diagram- 2 max]

[1 mark for each correctly identified outlier – 3 max]



Here, the maximum and minimum values (whiskers) are obtained from the table. The quartiles are already given. The outliers are indicated with asterisks and are defined as an outlier if their value is 1.5 times the interquartile range above or below the third or first quartile respectively.

ii) **Explain your course of action for dealing with the outliers when creating a model for pickle eaten.**

[1 mark]

In this instance, there is no justification for removing the outliers. It is not beyond the bounds of reason to entertain that people in the South-East might have eaten ~150g a week or those in the South-West eating ~120g. The modeller should accept that these are possible values and they should be included when predicted the amount of pickle people will eat in the future.