AQA

A Level

A Level Mathematics

Cleaning Data

Name:

M M E

Mathsmadeeasy.co.uk

Total Marks:

L4- Cleaning Data- Questions

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]

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]

iii) Sketch the distribution of waiting times. [2]

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]

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]

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. [5]

ii) State on the tree diagram which possibility is a Type I error and which is a Type II error. [2]

iii) In words, and using the aforementioned context, describe what a Type II error is. [1]

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. [7]

ii) Explain your course of action for dealing with the outliers when creating a model for pickle eaten. [1]