



F01-10min Data Cleaning and Outliers (Foundation): 10-minute Pack

EDEXCEL 1ST0 · Calculator allowed · about 10 minutes

Total Marks

Name: _____ Date: ____ / ____ / ____

Answer ALL questions. Show all your working.

- 1** A leisure centre's membership sign-up log was entered by two different receptionists on the same day.

Row	Name entered	Date of birth
1	Mohammed Iqbal	03/05/2011
2	M. Iqbal	03/05/2011
3	Sara Owusu	21/11/2010
4	Leo Fischer	09/02/2012

- (a) Write down the two row numbers that most likely refer to the same person. (1)

- (b) Give two reasons why rows 1 and 2 are likely to refer to the same person, even though the names were not entered identically. (2)

(Total for Question 1 is 3 marks)

- 2** A list of ages recorded for a school trip includes the value 130 years. Write down why this value must be a data-entry error.

(Total for Question 2 is 1 mark)

- 3** A survey records a respondent's height as -160 cm. State why this value cannot be correct.

(Total for Question 3 is 1 mark)

- 4** A cafe records the volume of coffee sold in take-away cups one morning. Some volumes were entered in millilitres and some in litres by mistake.

Cup	A	B	C	D	E
Volume recorded	350	0.4	250	0.3	400

- (a) Explain why the volumes recorded for cups B and D are likely to have been entered in the wrong unit. (1)
- (b) Convert the volumes for cups B and D into millilitres, so that all five volumes use consistent units. (2)

.....
(Total for Question 4 is 3 marks)

- 5** For each statement about data cleaning, write down whether it is True or False.

- (i) An outlier is always caused by a mistake when the data was recorded. (1)
-
- (ii) Removing a genuine extreme value from a data set just because it looks unusual can make the data set less representative of what really happened. (1)
-
- (iii) If a data set contains duplicate entries for the same person, the true sample size may be smaller than the number of rows recorded. (1)
-

(Total for Question 5 is 3 marks)

Mark scheme · F01-10min Data Cleaning and Outliers (Foundation): 10-minute Pack

Question 1

- (a) **B1** 1 and 2 are correct
- (a) **Answer: Rows 1 and 2**
- (b) **B1** both rows have the same date of birth (03/05/2011)
- (b) **B1** 'M. Iqbal' is an abbreviated/shortened version of the name 'Mohammed Iqbal' or
- (b) **Answer: Both rows share the same date of birth, and 'M. Iqbal' is just a shortened version of the full name 'Mohammed Iqbal'.**

Question 2

- **B1** 130 years is an impossible age for a person, so it must be an error or
- **Answer: 130 years is an impossible age for a person (nobody lives that long), so it must be a data-entry error.**

Question 3

- **B1** a height cannot be negative or
- **Answer: A height cannot be negative, so -160 cm is impossible and must be an error.**

Question 4

- (a) **B1** 0.4 and 0.3 millilitres would be an impossibly small amount of coffee, so these must have been recorded in litres instead or
- (a) **Answer: 0.4 ml and 0.3 ml would be far too small for a cup of coffee, so these values must actually be in litres, not millilitres.**
- (b) **M1** multiplies each value by 1000 or
- (b) **A1** 400 ml and 300 ml, both correct
- (b) **Answer: Cup B = 400 ml, Cup D = 300 ml**

Question 5

- (i) **B1** False or
- (i) **Answer: False**
- (ii) **B1** True or
- (ii) **Answer: True**
- (iii) **B1** True or
- (iii) **Answer: True**