



F01-test Data Cleaning and Outliers (Foundation): Test Pack

EDEXCEL 1ST0 · Calculator allowed · about 25 minutes

Total Marks

Name: _____ Date: ____ / ____ / ____

Answer ALL questions. Show all your working.

- 1** A school office assistant is entering sign-up details for a new coding club into a spreadsheet. The raw data entered is shown below.

Entry	Name	Age (years)	Height
1	Aliyah Khan	14	1.58 m
2	Ben Ward	(not recorded)	1.62 m
3	Chloe Bird	13	152 cm
4	Chloe Bird	13	152 cm

- (a) State the problem with entry 2. (1)
- (b) State the problem with entries 3 and 4. (1)
- (c) Entry 3's height is recorded in a different unit (cm) from the rest of the table (m). (1)
Explain why this inconsistency could cause a problem when the data is analysed.

(Total for Question 1 is 3 marks)

- 2** A list of 8 delivery times, in minutes, is: 18, 21, 19, 22, 200, 20, 19, 21. Write down the value that is a data-entry error, and state what its correct value was most likely intended to be.

(Total for Question 2 is 2 marks)

- 3** A vet records the weight, in kg, of 8 puppies at a check-up. One entry is missing.

Puppy	A	B	C	D	E	F	G	H
Weight (kg)	4.2	3.9	(missing)	4.5	4.0	3.8	4.1	4.3

- (a) Give one appropriate way the vet could deal with the missing weight for puppy C. (1)
- (b) Give one disadvantage of simply deleting puppy C's record from the data set instead. (1)
- (c) The vet decides to estimate the missing value using the mean weight of the other 7 puppies. Work out this estimate. (2)

(Total for Question 3 is 4 marks)

- 4** A spreadsheet records 5 delivery dates for the same week: 03/06/2027, 4/6/27, 05-06-2027, 2027-06-06, 07/06/2027. Give two reasons why these dates are difficult to compare directly.

(Total for Question 4 is 2 marks)

- 5** A researcher records the resting pulse rate, in beats per minute, of 8 volunteers: 66, 70, 68, 690, 71, 69, 67, 70. Identify the value that is almost certainly a data-entry error, explain why it cannot be a genuine pulse rate, and suggest the most likely correct value.

(Total for Question 5 is 3 marks)

- 6** A gym's sign-in sheet for one evening class lists 10 entries, taken automatically from members scanning their cards.

Row	Name	Member ID
1	Grace Oduya	M1042
2	K. Yilmaz	M1077
3	Grace Oduya	M1042
4	Daniel Cross	M1055
5	Kerem Yilmaz	M1077
6	Priti Shah	M1061
7	Daniel Cross	M1055
8	Amir Bello	M1089
9	Grace Oduya	M1042
10	Priti Shah	M1061

- (a) Work out how many unique members actually attended the class. (2)

- (b) Write down the Member ID that was recorded the greatest number of times, and state how many times it appears. (2)

(Total for Question 6 is 4 marks)

- 7** A clinic's records show the weight, in kg, of 10 newborn babies. One value is missing: 3.2, 3.4, 3.1, 3.5, 3.3, 3.6, 3.0, 3.4, (missing), 3.2. Give one appropriate way the clinic could estimate the missing weight, then use the mean of the other 9 weights to work out this estimate, and state one disadvantage of simply deleting this baby's record instead.

(Total for Question 7 is 4 marks)

Mark scheme · F01-test Data Cleaning and Outliers (Foundation): Test Pack

Question 1

- (a) **B1** identifies that the age has not been recorded / is missing oe
- (a) **Answer: Ben Ward's age is missing (not recorded).**
- (b) **B1** identifies that entries 3 and 4 are duplicate entries for the same person oe
- (b) **Answer: Entries 3 and 4 are duplicates (the same person, Chloe Bird, has been entered twice).**
- (c) **B1** explains that mixing units could lead to heights being compared or averaged incorrectly (e.g. 152 could be mistaken for 152 m) oe
- (c) **Answer: If the units are not made consistent, 152 could wrongly be treated as if it were in metres (152 m) when calculating an average or comparing heights, giving a meaningless result.**

Question 2

- **B1** 200 identified as the error cao
- **B1** 20 stated as the likely intended value oe
- **Answer: 200 is the error. The intended value was most likely 20 (an extra zero was typed by mistake).**

Question 3

- (a) **B1** any sensible strategy, e.g. re-weigh puppy C to get the true value, or estimate/impute using the mean of the other puppies' weights oe
- (a) **Answer: Re-weigh puppy C to find the true value, or estimate the missing weight using the mean of the other 7 puppies.**
- (b) **B1** loses information about puppy C / reduces the sample size, which could make the data set less representative oe
- (b) **Answer: Deleting the record loses all information about puppy C and reduces the sample size, which could make the results less representative.**
- (c) **M1** $(4.2+3.9+4.5+4.0+3.8+4.1+4.3) / 7$ oe
- (c) **A1** 4.11 kg (awrt 4.1) cao
- (c) **Answer: 4.11 kg (awrt 4.1 kg)**

Question 4

- **B1** different orderings are used, e.g. day/month/year and year-month-day oe
- **B1** inconsistent use of leading zeros and full/abbreviated years oe
- **Answer: The dates use different orderings (some day/month/year, one year-month-day) and inconsistent leading zeros and year lengths, so they cannot be compared or sorted reliably as they are.**

Question 5

- **B1** 690 identified as the error cao
- **B1** explains 690 bpm is far higher than any realistic resting pulse rate oe
- **B1** 69 stated as the most likely correct value cao
- **Answer: 690 is the error; a resting pulse of 690 bpm is impossible for a human, so an extra digit was probably typed. The correct value was most likely 69 bpm.**

Question 6

- (a) **M1** groups the 10 rows by Member ID oe
- (a) **A1** 5 cao
- (a) **Answer: 5 unique members**
- (b) **B1** M1042 cao
- (b) **B1** 3 (times) cao
- (b) **Answer: M1042, recorded 3 times**

Question 7

- **B1** sensible method, e.g. estimate using the mean of the other 9 weights, or check the hospital notes or
- **M1** $(3.2+3.4+3.1+3.5+3.3+3.6+3.0+3.4+3.2) / 9$ or
- **A1** 3.3 kg or
- **B1** disadvantage of deleting, e.g. loses information about that baby and reduces the sample size or
- **Answer:** Estimate the missing weight using the mean of the other 9 weights, which is 3.3 kg; deleting the record instead would lose information about that baby and reduce the sample size.