FOUNDATION

F01-30min Data Cleaning and Outliers (Foundation): 30-minute Pack

Total Marks

EDEXCEL 1ST0 · Calculator allowed · about 30 minutes

Name: _____ Date: ____ / ____ / _____

Answer ALL questions. Show all your working.

1 Two rows in a database both list 'Age: 12, Name: Tom Lee, School: Elm Road'. Write down what type of data problem this is.

(Total for Question 1 is 1 mark)

2 A column of exam marks records one entry as 'ninety-two' while every other entry is a number such as 85. Write down the data-cleaning problem shown here.

(Total for Question 2 is 1 mark)

3 A vet's spreadsheet has two entries for the same day: 'R. Ahmed, 14:00, Consultation' and 'Raj Ahmed, 14:00, Consultation'. Give two reasons why these are likely to be duplicate entries for the same appointment.

(Total for Question 3 is 2 marks)

4 Jasmine is investigating whether pupils in her year group spend more time on homework on weekdays or at weekends. She plans a survey, collects responses using a paper questionnaire, then enters the data into a spreadsheet before working out any averages. (The statistical enquiry cycle has four stages: plan, collect, process, interpret.)

(a) Jasmine notices that some pupils have written their homework time in minutes and others in hours. At which stage of the enquiry cycle should Jasmine correct this problem, before she calculates any averages? **(1)**

(b) Explain why it is important to clean the data during the process stage before moving on to the interpret stage. **(1)**

(c) Jasmine also finds that 3 questionnaires have been returned completely blank. Give one appropriate action she could take with these blank responses at the process stage. **(1)**

(d) State one reason it would NOT be sensible for Jasmine to guess plausible homework times to fill in the 3 blank questionnaires. **(1)**

(Total for Question 4 is 4 marks)

5 A clinic logs patient appointment dates. Two different members of staff entered the dates in different formats.

Patient	Date entered
P	04/07/2026
Q	2026-07-11
R	09/07/2026
S	2026-07-03

(a) Write down which format has been used for patient Q's appointment date. **(1)**

.....

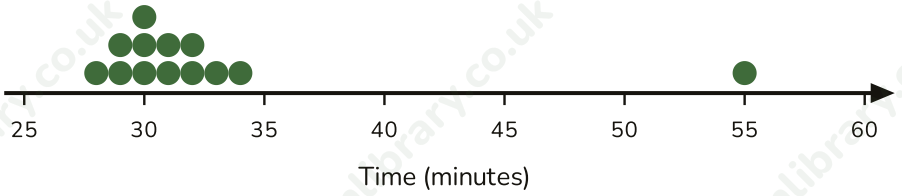
(b) Explain why mixing date formats like this in one column could cause a problem when the appointments are sorted into date order. **(1)**

(c) Rewrite patient Q's date using the same day/month/year format as the other patients. **(1)**

.....

(Total for Question 5 is 3 marks)

6 The dot plot shows the time, in minutes, taken by 13 runners to complete a 5 km fun run.



(a) Write down the modal time. **(1)**

.....

(b) Write down the outlier in this data set. **(1)**

.....

(c) Give a reason why 55 minutes is considered an outlier. **(1)**

(Total for Question 6 is 3 marks)

---

7 A gym's sign-in sheet lists 8 rows for one class, taken automatically from members scanning their cards. The Member IDs recorded are, in order: M501, M502, M501, M503, M504, M502, M505, M501. Work out how many unique members attended, and write down the Member ID that was recorded most often, stating how many times it appears.

(Total for Question 7 is 3 marks)

8 The weekly pocket money, in £, of 9 pupils in a class is recorded below.

£5, £6, £5, £7, £6, £5, £6, £7, £45

(a) Work out the mean and the median weekly pocket money for all 9 pupils. **(3)**

(b) One pupil's pocket money, £45, is much higher than the rest. Excluding the £45, work out the new mean and new median for the remaining 8 pupils. **(3)**

(c) Comment on the effect that removing the £45 had on the mean compared with its effect on the median. **(1)**

(Total for Question 8 is 7 marks)

9 The weekly earnings, in pounds, of 9 employees at a small cafe are: 310, 320, 305, 315, 300, 325, 318, 312, 980. One employee is the manager, who genuinely earns much more than the others (this is not an error). Work out the median weekly earnings and the mean weekly earnings for all 9 employees, and state which average better represents the typical earnings, giving a reason.

(Total for Question 9 is 3 marks)

10 The number of emails received by 11 employees in one day, given in ascending order, is: 8, 9, 9, 10, 11, 11, 11, 12, 13, 14, 40. Find the lower quartile (Q1) and upper quartile (Q3) of the data, work out the interquartile range (IQR), and then use the rule that a value is an outlier if it is higher than Q3 + 1.5 x IQR to decide whether 40 is an outlier.

(Total for Question 10 is 4 marks)

Mark scheme · F01-30min Data Cleaning and Outliers (Foundation): 30-minute Pack

Question 1

- **B1** duplicate entry / the same record entered twice oe
- **Answer: A duplicate entry: the same record has been entered twice.**

Question 2

- **B1** the value is recorded as text/words instead of a number, so it is in an inconsistent format oe
- **Answer: The value is written as words instead of a number, so it is in an inconsistent format and will not be read as numeric data.**

Question 3

- **B1** both entries share the same time, 14:00 oe
- **B1** 'R. Ahmed' is a shortened form of 'Raj Ahmed' oe
- **Answer: Both entries share the same time (14:00), and 'R. Ahmed' is simply a shortened version of the name 'Raj Ahmed'.**

Question 4

- (a) **B1** process cao
- **(a) Answer: The process stage**
- (b) **B1** if inconsistent or incorrect data is analysed, the averages/conclusions calculated would be wrong or misleading, so cleaning first ensures the interpretation is based on accurate, comparable data oe
- **(b) Answer: If Jasmine analyses uncleaned data, her averages and conclusions could be wrong or misleading, so the data must be made accurate and consistent before it is interpreted.**
- (c) **B1** remove/exclude them from the data set, since they contain no usable information oe
- **(c) Answer: Remove the 3 blank questionnaires from the data set, since they contain no usable information.**
- (d) **B1** guessed values are not genuine data and could bias/distort the results, as they may not reflect what those pupils actually do oe
- **(d) Answer: Guessed values would not be genuine data and could bias the results, since they may not reflect what those pupils actually do.**

Question 5

- (a) **B1** year-month-day (ISO) format oe
- **(a) Answer: Year-month-day format (2026-07-11)**
- (b) **B1** a computer or reader could misread which part is the day and which is the month, so the appointments could be sorted into the wrong order oe
- **(b) Answer: A date such as 04/07/2026 could be misread as the 4th of July or as April 7th depending on the format expected, so mixed formats could sort the appointments in the wrong order.**
- (c) **B1** 11/07/2026 cao
- **(c) Answer: 11/07/2026**

Question 6

- (a) **B1** 30 cao
- **(a) Answer: 30 minutes**
- (b) **B1** 55 cao
- **(b) Answer: 55 minutes**
- (c) **B1** 55 is much greater than (isolated from) the rest of the data, which is clustered between 28 and 34 minutes oe
- **(c) Answer: 55 minutes is far higher than every other runner's time, which are all clustered between 28 and 34 minutes.**

Question 7

- **M1** groups the 8 rows by Member ID oe
- **A1** 5 unique members cao
- **B1** M501, recorded 3 times, cao
- **Answer: 5 unique members; M501 was recorded most often, 3 times.**

Question 8

- (a) **M1** 5+6+5+7+6+5+6+7+45 (= 92), oe
- (a) **A1** mean = £10.22 (awrt 10.22) cao
- (a) **B1** median = £6 cao
- **(a) Answer: Mean = £10.22 (awrt), median = £6**
- (b) **M1** 5+6+5+7+6+5+6+7 (= 47), oe
- (b) **A1** mean = £5.875 (= £5.88 to 2 dp) cao
- (b) **B1** median = £6 cao
- **(b) Answer: Mean = £5.875 (= £5.88 to 2 dp), median = £6**
- (c) **B1** the mean changed a lot (from £10.22 to £5.88 approx) but the median stayed the same (£6), showing the median is much less affected by (more resistant to) an outlier than the mean oe
- **(c) Answer: The mean fell a lot (from about £10.22 to about £5.88) but the median stayed exactly the same (£6), showing the median is far less affected by an outlier than the mean.**

Question 9

- **B1** median = 315 cao
- **B1** mean = 387.22 (awrt) cao
- **B1** median is more representative, since the manager's high (genuine) salary pulls the mean up oe
- **Answer: Median = £315, mean = £387.22 (awrt); the median better represents typical earnings because the manager's high salary pulls the mean up.**

Question 10

- **B1** Q1 = 9 cao
- **B1** Q3 = 13 cao
- **B1** IQR = 4 (ft from Q3 - Q1)
- **B1** boundary = 19 (ft), and 40 correctly identified as an outlier since 40 > 19
- **Answer: Q1 = 9, Q3 = 13, IQR = 4, upper boundary = 19; since 40 is greater than 19, it is an outlier.**