Unit 4 · Data Science

WORKSHEET

Full marks 33

Name _____ Class _____ Roll no. ____ Date _____ Marks _____ / 33

1 Key words [4]

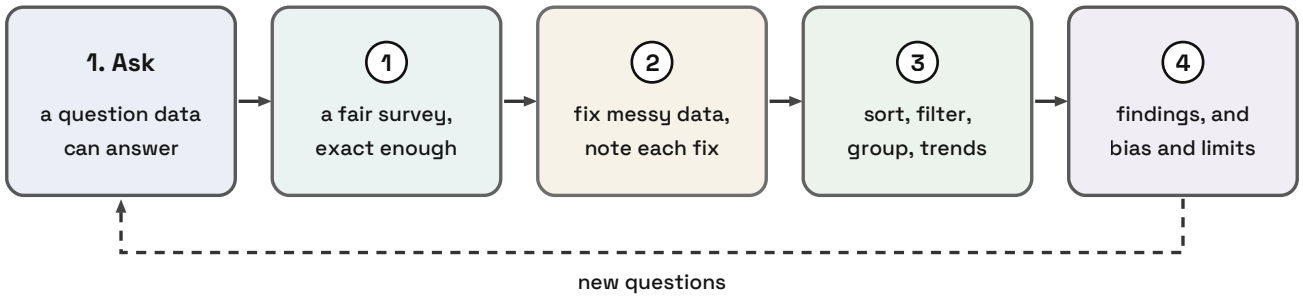
Fill in the blanks with words from the box. Two words in the box are not needed. Each blank carries 0.5 mark.

DataFrame	**mean**	**median**	**merge**	**multivariate**	**outlier**	**parameters**	**quartile**	**Series**	**simulation**

In pandas, a table with named columns is a _____, and each of its columns is a _____. A value far from the rest is an _____: it pulls the _____ towards it, but hardly moves the _____. A chart that shows three or more variables at once is a _____ chart. A program that acts out a real system with made-up data is a _____, and the numbers that set how its data is made, such as 25 people an hour, are its _____.

2 Label the diagram [2]

The diagram shows the steps of a data investigation. Write the name of each missing step in place of 1 to 4. Each carries 0.5 mark.



1. _____ 2. _____

3. _____ 4. _____

3 Multiple choice [5]

Choose the correct option for each of the following questions.

i. `districts[districts["height"] > 4000]` gives what? [1]

a. the number of districts higher than 4,000 m
b. a DataFrame with only the rows of the districts whose height is over 4,000 m
c. the heights, with 4,000 taken away from each
d. True if any district is higher than 4,000 m

ii. A column of house rents in Bhaktapur has a mean of Rs 32,000 and a median of Rs 18,000. What does that tell you? **[1]**

a. Most rents are around Rs 32,000.

b. A few very high rents pull the mean up; half the rents are Rs 18,000 or less.

c. The data must have a mistake, because the mean and median should be equal.

d. Half the rents are over Rs 32,000.

iii. A program reads a file and prints `people["marks"].sum()`, and the answer is `"657285"`, in quotes. What has happened? **[1]**

a. The class scored 657,285 marks.

b. The marks column was read as text, so `sum()` joined the strings `"65"`, `"72"` and `"85"` end to end.

c. pandas can't add marks.

d. The file has three rows missing.

iv. Why does a simulation program often start with `random.seed(…)`? **[1]**

a. to make the numbers more random

b. to make the program run faster

c. so that it makes the same random numbers every run, and a result can be checked or a bug found again

d. because random numbers can't be made without it

v. A chart shows the rain of every district (across) against its height (up), with the dots colored by province. How many variables does it show? **[1]**

a. 1 **b.** 2

c. 3 **d.** 77

4 What does it print? [3]

Write exactly what the program prints, in the box under it.

```
1  import pandas as pd
2
3  shops = pd.DataFrame({
4      "shop": ["Asha", "Bikash", "Chiya Ghar", "Dhaba", "Everest"],
5      "town": ["Thimi", "Banepa", "Thimi", "Banepa", "Thimi"],
6      "cups": [120, 85, 300, 95, 30]})
7  print(shops.shape)
8  busy = shops[shops["cups"] > 100]
9  print(len(busy), busy["shop"].tolist())
10 print(shops.groupby("town")["cups"].sum().to_dict())
11 print(shops["cups"].mean(), shops["cups"].median())
```

OUTPUT

5 Trace the program [4]

The function `way()` from the unit puts each answer to "How do you come to school?" into a group. Fill in a row for each answer, then answer the question under the table.

```
1  def way(text):
2      text = text.strip().lower()
3      if text == "":
4          return "no answer"
5      for word in ["walk", "foot"]:
6          if word in text:
7              return "walk"
8      for word in ["bus", "micro"]:
9          if word in text:
10             return "bus or micro"
11     for word in ["cycle", "bicycle"]:
12         if word in text:
13             return "bicycle"
14     return "other"
```

Answer typed	`text` after line 2	What `way` gives back
`"Micro "`		
`"BIKE"`		
`two spaces only`		
`"bus, then walk"`		
`"Motorcycle"`		
`"On Foot"`		

Which answer is put in a group that is clearly wrong? Why does it happen?

6 Turned round by a third variable [4]

Read, then answer.

Two schools' end-of-year exam results, with made-up numbers:

	School A	School B
Grade 9	45 of 50 pass: ..... %	170 of 200 pass: ..... %
Grade 10	120 of 200 pass: ..... %	25 of 50 pass: ..... %
Both grades	165 of 250 pass: ..... %	195 of 250 pass: ..... %

i. Fill in the six percentages in the table.
ii. Which school does better in each grade, and which over both grades together? Explain how both can be true.
iii. A newspaper writes "School B is the better school". What should it write instead?

7 Find and fix the bug [3]

Read the program, then answer.

A student copied the unit's ticket counter simulation but changed one line. It now prints `Average wait: 0.0 minutes` on every morning, however many people come.

```
import random

def one_morning(per_hour, counters):
    free_at = [0] * counters
    waits = []
    now = 0
    while True:
        now = now + random.expovariate(per_hour / 60)
        if now > 180:
            break
        k = free_at.index(min(free_at))
        start = now
        waits.append(start - now)
        free_at[k] = start + random.uniform(1, 3)
    return waits

random.seed(5)
waits = one_morning(25, 1)
print("Average wait:", round(sum(waits) / len(waits), 1), "minutes")
```

i. Which line is wrong? Write it correctly.

ii. Why does the mistake make every wait 0?

8 Write a program [4]

Write the program in Python on the lines. Any program that gives the right result gets the marks.

The file **districts.csv** has one row for each of Nepal's 77 districts, with the columns `district`, `belt` (Mountain, Hill or Terai), `height`, `slope`, `rain` (mm a year) and `people` (per km^2). Using pandas, write a program that prints:

i. the three Hill districts with the least rain;

ii. how many Terai districts get more than 2,000 mm of rain;

iii. the median rain of each belt.

9 Short answer [2]

Write a short answer to the following question.

People per km^2 in Nepal's districts run from 2 (Manang) to 4,936 (Kathmandu). Why does a chart of this need a log scale, and what must the chart say on that axis?

10 Think [2]

Think, then answer.

A news report says: "Students who use phones more than 3 hours a day score 15% lower in exams. Phones are ruining results." Write two questions you would ask before believing it, and say which word in the report claims more than the data can show.

11 Activity: Clean, then join (not marked)

At a computer, open the unit's page (the QR code) and find **Clean, then join.** Switch on the cleaning steps in the order below, and write the readings after each.

Steps switched on	Districts matched	Population, Hill	Population, Mountain
3. Numbers from text			
3, then 1. Trim spaces, fix capitals			
3, 1, then 2. One spelling per district			
3, 1, 2, then 4. Remove repeated rows			
Which step changed the Hill total the most?			