Grade 10 Computing

Unit 4 · Data Science

ANSWER KEY · FOR TEACHERS

Full marks 33

Give the marks for the points in the marking scheme; an answer in other words that says the same is right.

1 Key words

1. DataFrame
2. Series
3. outlier
4. mean
5. median
6. multivariate
7. simulation
8. parameters

Each carries 0.5 mark. **K · 4**

2 Label the diagram

1. Collect
2. Clean
3. Analyze
4. Report

Each carries 0.5 mark. **K · 2**

3 Multiple choice

i. **b.** a DataFrame with only the rows of the districts whose height is over 4,000 m **U · 1**

The comparison inside gives True or False for each row, and the outer brackets keep the rows that are True: a smaller table, with all its columns. `len(...)` of it would give the number of districts.

ii. **b.** A few very high rents pull the mean up; half the rents are Rs 18,000 or less. **U · 1**

The median is the middle rent: half are below it. A mean far above the median means a few very large values (outliers) are pulling it up, so the mean is not a typical rent.

iii. **b.** The marks column was read as text, so `sum()` joined the strings `"65"`, `"72"` and `"85"` end to end. **A · 1**

A column with even one value that isn't a number is read as text, and the sum of text is the strings joined together. Check `dtypes`, find the values that aren't numbers, and convert the column on purpose.

iv. **c.** so that it makes the same random numbers every run, and a result can be checked or a bug found again **K · 1**

Without a seed, every run is different, which is what you want for exploring. With a seed, the same "random" numbers come every time, so you and anyone checking your work get exactly the same result.

v. **c.** 3 **U · 1**

Rain, height and province: two numbers by position and a category by color. The 77 are rows (districts), not variables.

4 What does it print?

MARKING SCHEME **U · 3**

- ✓ `(5, 3)`, and `2 ['Asha', 'Chiya Ghar']`: the filter keeps only the rows with more than 100 cups.
- ✓ `{'Banepa': 180, 'Thimi': 450}`: one total per town (85 + 95; 120 + 300 + 30).
- ✓ `126.0 95.0`: the mean is 630 ÷ 5; the median is the middle of 30, 85, 95, 120, 300. Chiya Ghar's 300 pulls the mean above four of the five shops.

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

5 Trace the program

ANSWER **A · 4**

Answer typed	`text` after line 2	What `way` gives back
”Micro ”	”micro”	bus or micro
”BIKE”	”bike”	other
two spaces only	””	no answer
”bus, then walk”	”bus, then walk”	walk
”Motorcycle”	”motorcycle”	bicycle
”On Foot”	”on foot”	walk

“Motorcycle” is counted as a **bicycle:** `"cycle" in "motorcycle"` is True, because `in` finds the letters anywhere in the text, even inside another word.

MARKING SCHEME

- ✓ The `text` column: spaces off both ends and small letters (`"micro"`, `""` for the spaces only).
- ✓ “Micro ” bus or micro, “BIKE” other, the spaces no answer.
- ✓ “bus, then walk” is walk, because the walk rule is checked first; “On Foot” walk.
- ✓ “Motorcycle” as bicycle, because “cycle” is found inside “motorcycle”.

6 Turned round by a third variable

ANSWER **A · 4**

i. School A: 90%, 60%, 66%. School B: 85%, 50%, 78%.

ii. School A does better in both grades (90% against 85%, 60% against 50%), but School B does better over both grades (78% against 66%). In both schools Grade 10 passes less often than Grade 9, and School A has most of its students in Grade 10 (200 of 250), School B in Grade 9 (200 of 250). So School A’s overall rate is mostly its Grade 10 rate, pulled down by **which grade its students are in**, not by how well they do. This is Simpson’s paradox.

iii. Compare grade by grade: School A’s pass rate is higher in both Grade 9 and Grade 10; its overall rate is lower only because most of its students are in Grade 10, which passes less often at both schools.

MARKING SCHEME

- ✓ The six percentages: 90, 60, 66 and 85, 50, 78.
- ✓ A better in each grade, B better overall.
- ✓ The reason: most of A’s students are in the grade that passes less often, which pulls its overall rate down (Simpson’s paradox).
- ✓ The report compares like with like, grade by grade, and gives the reason for the overall rates.

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

7 Find and fix the bug

ANSWER A·3

i. Line 12. It should be `start = max(now, free_at[k])`.

ii. With `start = now`, every person is served the moment they arrive, even when the counter is still busy with someone else, so `start - now` is always 0. A person must start at whichever is later: when they arrive, or when the counter is free.

Fixed, it prints:

```
OUTPUT
Average wait: 5.9 minutes
```

MARKING SCHEME

✓ Line 12.

✓ `start = max(now, free_at[k])`.

✓ The program ignores when the counter is free, so nobody ever waits: `start - now` is always 0.

8 Write a program

ANSWER A·4

```
import pandas as pd

districts = pd.read_csv("districts.csv")
hill = districts[districts["belt"] == "Hill"]
print(hill.sort_values("rain").head(3))
wet = districts[(districts["belt"] == "Terai") & (districts["rain"] > 2000)]
print(len(wet))
print(districts.groupby("belt")["rain"].median())
```

```
OUTPUT
     district  belt  height  slope  rain  people
76   Udayapur  Hill     430     16  1691     148
75  Terhathum  Hill    1395     22  1713     132
9   Bhaktapur  Hill    1378     10  1745    3510
6
belt
Hill        2195.5
Mountain    2163.5
Terai       1900.0
Name: rain, dtype: float64
```

MARKING SCHEME

✓ `import pandas as pd` and `pd.read_csv("districts.csv")`.

✓ i: the Hill rows filtered, sorted by rain (smallest first) and the first three taken (`head(3)`, or `nsmallest(3, "rain")`).

✓ ii: two conditions joined with `&`, each in brackets, and `len()` of the result.

✓ iii: `groupby("belt")["rain"].median()`.

9 Short answer

MARKING SCHEME U·2

✓ On an ordinary scale Kathmandu sits at the top and the other districts are squashed together at the bottom; on a log scale each step is ten times the one before (1, 10, 100, 1,000), so small values get as much room as big ones.

✓ The axis must be named as a log scale, with its ticks labeled, because equal gaps no longer mean equal differences.

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

10 Think

ANSWER HA · 2

Questions (any two): How many students, from where, and how were they chosen? How was phone use measured: by the students' own guesses, or by the phones? Could a third thing explain both, such as students with less help at home or who work in the evenings, and were the groups compared like with like? Is a 15% difference bigger than the usual spread of marks?
"Ruining" claims that phones **cause** the lower marks; the data only shows that the two go together.

MARKING SCHEME

- ✓ Two sensible questions (the sample, how phone use was measured, a third variable, the size of the difference against the spread).
- ✓ "Ruining" claims a cause, when the data shows only that phone use and lower marks go together.

11 Activity: Clean, then join

(not marked)

WHAT STUDENTS SHOULD FIND

- Step 3 only: 7 of 13 matched; Hill 1,597,983 (with Bhaktapur twice); Mountain 172,767.
- Steps 3 and 1: 10 of 13; Hill 4,517,280; Mountain 172,767. Kathmandu, Lalitpur and Dhading come back, over 2.9 million people: the biggest change.
- Steps 3, 1 and 2: 13 of 13; Hill 5,347,392; Mountain 435,391.
- All four: 13 of 13; Hill 4,915,260; Mountain 435,391 (without Rasuwa, whose population is missing); Terai 719,859 throughout.

Marks by level	Knowledge	Understanding	Application	Higher ability	**Total**
Marks	7	8	16	2	**33**

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking