L6 Your Data Is Already Messy

Data Science · Cycle 1 · Guided notes

Name: _____

Date: _____ Period: _____

Spiral warm-up

1 From L5: bring column B's largest value to row 2.

Select the whole block from _____, header row included.

Data > Sort range > Advanced range sorting options.

Tick _____.

Sort by column B, _____ (largest first).

Sort column B alone and every row goes _____. Nothing turns _____.

2 From L4: a sheet started Wed 9 Sep, one row per night, the header in row 1.

Tue 22 Sep, morning bell. Rows of data: _____

They sit in rows _____ to _____.

So =MIN on column B reads _____

A missed night keeps its row: blank cells, and _____ in Notes.

The idea

RAW is the record: nothing already in RAW is _____. Every fix happens on _____, one CLEANING LOG row per fix.

	Three words (VDOE DS.7)
Valid	one _____ within a _____
Complete	no _____ value missing
Unique	no _____, no extra entries

	Four defects, in this order
1	missing _____
2	inconsistent _____
3	a _____ that drifted
4	values that _____

Objectives

- Keep RAW untouched and clean only the CLEAN tab.
- Find the four defects and name the word each breaks: valid, complete or unique.
- Check a column with MIN, MAX, COUNT and COUNTA, and log every change with its why and how.

Key words

drift: a definition or unit changing partway

outlier: could be right, far from the rest

data validation: checking data before you change it

COUNT and COUNTA: cells with numbers and cells not empty

Examples

Example 1

Data validation: the range check on a trip log. *Constructed example.*

	A	B	C	D
1	**Date**	**Travel time (min)**	**Mode**	**Notes**
2	3/2/2026	18	Bus	
3	3/3/2026	21	bus	
4	3/4/2026	12	walk	
5	3/5/2026	0	Bus	forgot to time
6	3/6/2026	14	Walking	
7	3/9/2026	19	Bus	
8	3/10/2026	15 min	bus	
9	3/11/2026	13	walk	
10	3/12/2026	20	Bus	
11	3/13/2026	11	Walking	
12	3/13/2026	11	Walking	

	=MIN	=MAX	=COUNT	=COUNTA	Rows of data
B2:B12					

MIN flags _____. COUNT below COUNTA: text in _____.

COUNTA equals the rows, yet one trip was never timed: the _____ hid it.

Example 2

Column C. Find walk, replace with Walking, box unticked, then ticked.

Cell	Before	Box unticked	Box ticked
C4	walk		
C6	Walking	Walkinging	Walking
C9	walk		
C11	Walking		

Spellings: _____

Real modes: _____

Box: **Match entire cell contents**

Sheets ignores _____ unless **Match case** is ticked.

Search is set to _____, so RAW is untouched.

Then find bus, replace with Bus, box ticked. Cells that change: _____. Two Replace alls, _____ log rows.

Example 3

B5 says 0 and its note says forgot to time. Blank or zero? Then the log.

Did I forget, or did nothing happen? _____

So B5 becomes _____. A 0 means you measured _____.

AVERAGE _____ a blank and _____ a 0.

Row / cell	B5	row 12
Date logged	3/16/2026	3/16/2026
Column	Travel time (min)	
What it said	0	
What I changed it to		
Why		
How		

Example 4

AVERAGE(B2:B12) before and after the fixes, with B8 typed as 15.

Before: sum _____ ÷ count _____ = _____ minutes. The 0 is counted; 15 min is skipped.

After: sum _____ ÷ count _____ = _____ minutes.

One fix alone	Sum ÷ count	Mean
Clear the 0 in B5		
Type 15 in B8		
Delete row 12		

Moved the mean most: _____

The first Homework 5 sentence: Clearing the 0 on 3/5 moved my mean travel time from _____ to _____ minutes.

Practice

Practice 1 to 6 are Quiz 1 review and stay with you. Practice 7 goes with Homework 5.

1 Write valid, complete or unique.

(a) the same Tuesday typed twice _____

(b) a mood of 7 in a column rated 1 to 5 _____

(c) a night with no value and nothing in Notes _____

(d) yes, Yes and Y in one column _____

(e) a test row typed while setting up the sheet _____

2 A night logger's column on Tue 22 Sep should hold 13 values, B2:B14. =COUNT(B2:B14) returns 11. =COUNTA(B2:B14) returns 12.

Blank cells: _____ Cells holding text: _____

What can COUNT and COUNTA not tell you?

3 Find swim, replace with Swimming, **Match case** unticked. Fill in each cell with **Match entire cell contents** unticked, then ticked.

Cell before	Box unticked	Box ticked
swim		
Swim		
Swimming		
swimmer		

4 Real data: one VDOT counter on I-95 southbound at the Capital Beltway, vehicles per day (annual average). *VDOT annual average daily traffic files, segment 190002.*

Year	Route	Vehicles per day
2019	I-95 SB	60,000
2020	I-95 SB	56,000
2021	I-95 SB	64,000
2022	I-95	217,386
2023	I-95 SB	64,000

=MAX flags _____ in _____.

Outlier or drift? _____

217,386 ÷ 64,000 = _____, about _____ times its neighbors.

The clue in the Route column: _____

What drifted: _____ Word: _____

What the 2022 row most likely counts:

A fix a log can defend:

5 A student's reading log. *Constructed example.*

	A	B	C
1	**Date**	**Minutes read**	**Book**
2	9/1/2026	25	Dune
3	9/2/2026	30	Dune
4	9/3/2026		Dune
5	9/4/2026	1500	Dune
6	9/5/2026	20	dune
7	9/5/2026	20	dune

Find four problems. For each, write which row it is in (row numbers as Sheets shows them, header in row 1) and label it **valid**, **complete**, or **unique**.

Row	Problem	Label

6 Write the stage of the data cycle, exact name.

(a) Typing tonight's minutes into RAW

(b) Replacing dune with Dune on CLEAN, with a log row

(c) Computing the mean minutes read per night

7 Your own CLEAN tab, one number column. Write the possible range before you look.

Column	Smallest possible	Largest possible

=MIN	=MAX	=COUNT	=COUNTA	Rows so far

Anything to investigate? _____

One log row of my own:

Row / cell	What it said	What I changed it to

Why: _____

How (the Sheets step, so someone else could repeat it):

Homework 5, first draft: Changing _____

moved my _____ from _____

to _____.

Why it moved: _____

Check: the range is written before MIN and MAX are read; the log row has a Why and a How.