

Intermediate Data Science

Writing in Data Science

Joanna Bieri DATA201

Today

- No new code, and no new homework!
- Today is about **writing**: how to explain your data science so other people understand it and trust it.
- At the end you will use these ideas on your HW 6 write-up.

Today

DISCUSS

You read two blog posts about data science writing.

What is one piece of advice that stuck with you?

Why Writing Matters

- AI tools can load data, make a plot, and fit a model in seconds.
- They cannot stand in front of your boss and explain what you did and why to trust it.
- Every analysis should answer three questions:
 - 1 **What** did you do? What data, what method, what model?
 - 2 **How** did you do it?
 - 3 **Why** did you make those choices?

Why Writing Matters

DISCUSS

AI can write your pandas code in seconds.

So what is left for you to do?

Explain Your Choices

Every choice changes your results, so every choice needs a sentence: what you did and why.

- Dropping rows or columns
- Filling in missing values (and with what)
- Cutoffs and bins (who counts as a “young adult”?)
- Which plot, and what the reader should see in it
- Which model or method, and why that one

One or two plain sentences is enough. The goal: nobody has to guess.

Explain Your Choices

WHICH IS BETTER?

A

I removed the outliers and made a histogram of sale amounts.

B

Four sales were over \$500. All four were catering orders, which are a different kind of customer, so I removed them and made a histogram of the regular sales. With them in, the histogram was one tall bar and a long empty tail.

Which is better? Why?

Explain Your Choices

ANSWER

- B says **how many, what they were, why** they were removed, and what the plot looked like before.
- You might disagree with removing them. You can only disagree because B told you!
- A makes the reader guess. What counts as an outlier? How many?

Explain Your Choices

FIX IT

I filled in the missing values with 0.

You can't see this data. What would you need to find out before you could rewrite this sentence? How would you find it out?

Explain Your Choices

ANSWER

- **How many** are missing? `df.isna().sum()`
- **Which rows** are they? `df[df['items'].isna()]`, then look at the dates. All Sundays? One week in December?
- **Why** are they missing? Check where the data came from, or ask the person who collected it.

Then write down what you found. For example: *23 of the 500 days were missing the number of items sold. All of them were days the shop was closed, so 0 is the right value.*

Describe Your Data

Before anyone can trust your results, they need to know where your data came from.

- 1 **Where** did it come from? (source and link)
- 2 **Who** collected it, when, and why?
- 3 What is **one row**?
- 4 What is **missing**, and who is NOT in the data at all?
- 5 Are you **allowed** to use it? (license, privacy)

From *Datasheets for Datasets* (Gebru et al.), shortened.

Describe Your Data: a real example

- 2020, the middle of the COVID pandemic. *The Lancet* publishes a big study: hydroxychloroquine linked to a higher risk of death.
- The data came from a private company, Surgisphere. Other scientists noticed the numbers looked strange.
- Independent reviewers asked for the full data. The company would not give it to them.
- Three of the four authors said they could no longer vouch for the data. Retracted about two weeks after it came out.

Describe Your Data

WHICH IS BETTER?

A

The data is from Kaggle.

B

Five years of daily sales from one coffee shop, shared on Kaggle by the owner [link]. One row is one day. 14 days are missing, all in December 2021, when the shop was closed. This is one shop, so I am not making claims about coffee shops in general.

Which is better? Which of the five questions does B skip?

Describe Your Data

ANSWER

- B answers four of the five questions in four sentences. It skips “are you allowed to use it?”
- B is honest about what the data cannot tell you. Readers trust that more.
- A gives the reader nothing to check.

Ethics: Who Is In Your Data?

- 2018, Reuters: Amazon stopped using a model it built to rank job applicants.
- It was trained on ten years of resumes sent to Amazon. Most came from men.
- It gave lower scores to resumes with the word “women’s”, as in “women’s chess club captain.”
- Amazon tried to fix it, could not be sure it was fair, and dropped it.

Ethics: Who Is In Your Data?

DISCUSS

Nobody wrote code that said "prefer men."

Where did the bias come from? How could someone
have caught it?

Reproducibility

Your work is **reproducible** if someone else can run your notebook on your data and get the same results.

In this class, that someone is your teammate, pulling your notebook from the team repo.

Reproducibility: five countries left out

- 2010: Reinhart and Rogoff (Harvard) publish a paper: countries with debt over 90% of GDP tend to shrink. Politicians quote it in arguments about cutting spending.
- 2013: Thomas Herndon, a grad student at UMass Amherst, cannot reproduce it for a class assignment. He asks for their spreadsheet.
- An Excel average stopped at row 44 instead of row 49. Five countries left out.
- After fixing that and a few other problems: about +2.2% growth instead of -0.1%.

Reproducibility

DISCUSS

The paper was out for three years.

Why did it take a student with a homework assignment
to find the mistake?

Reproducibility: Excel changes your data

- Type the gene name SEPT2 into Excel and it becomes the date September 2.
- 2016: researchers scanned genetics papers. About **1 in 5** papers with Excel gene lists had gene names turned into dates or numbers.
- When your work is in code, every change to the data is written down where someone can see it.

Reproducibility

WHICH IS BETTER?

A

```
df = pd.read_csv('/Users/maria/Desktop/HW6/insurance.csv')
```

B

```
df = pd.read_csv('data/insurance.csv')
```

Maria pushes her notebook. Which line works on her teammate's laptop?

Reproducibility

ANSWER

- B, a **relative path**. It works anywhere the repo is, as long as the data folder comes along.
- A only works on Maria's computer. Her teammate does not have a `/Users/maria`.
- Even better: download it in code, like we do with `kagglehub`.

Reproducibility

DISCUSS

Your notebook works fine for you.

Your teammate runs it and gets `NameError: name 'df_clean' is not defined`.

What could have happened?

Reproducibility

ANSWER

- Most likely: the cells were run out of order, or the cell that made `df_clean` was deleted. Your kernel still remembers it. A fresh kernel does not.
- Fix: **Restart the kernel and run all** before every commit.
- This is the “runs without errors from start to finish” line on every homework checklist.

Reproducibility: let your code write your numbers

If you type a number into your text by hand, it is wrong the next time the data changes.

```
n_rows, n_cols = df.shape  
print(f'The data has {n_rows} rows and {n_cols} columns.')
```

```
ave_tip = df['tip'].mean()  
print(f'The average tip is ${ave_tip:.2f}.')
```

An **f-string** puts the values right into the sentence. The notes have the full example.

Cite Your Sources

- **Data:** source and link.
- **Books and articles:** name the idea you used.
- **Code from the web:** link in a comment right above it.
- **AI**, from our syllabus: a quick debugging question or syntax lookup needs no citation. If you copy and paste code an AI wrote, say so: name the tool and what it helped with.

Cite Your Sources

WHICH IS BETTER?

A

I used AI.

B

I used ChatGPT to write the code for the grouped bar chart in Part 3, and to fix a KeyError. The interpretation is my own.

Which AI note is better? Why?

Cite Your Sources

ANSWER

- B names the tool, what it did, and where. A reader knows exactly what to check.
- One or two honest sentences is enough.
- A real one, for today: *I used Claude (Anthropic) to help draft these notes and slides, find the real-world examples, and write the made-up example paragraphs. I chose the topics, and I reviewed and edited everything.*

Your Own Voice

Syllabus: “The written discussion, interpretation, and reflection you submit must always be your own thinking, in your own words.”

- **Say what you see.** Where are the points, what goes up or down, what is weird?
- **It is okay to be unsure.** “I am not sure why there is a dip in 2015” is a great sentence.
- **Use plain words.** Write like you are explaining it to a friend in another major.
- **Paraphrase.** Close the tab and write it the way you would say it.

Your Own Voice

WHICH IS BETTER?

A

This analysis provides valuable insights into the complex dynamics of coffee sales, revealing significant trends that underscore the importance of seasonality.

B

Sales are highest in December and lowest in July. I expected summer to be busy because of tourists, so I am not sure why July is so slow. Maybe iced drinks are cheaper?

Which one looked at the plot?

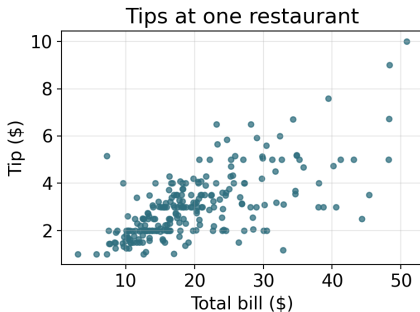
Your Own Voice

ANSWER

- A could be written without ever seeing the data. It does not name a single month or number.
- B says what it saw, what surprised them, and a guess to go check.
- “I am not sure why” is honest, and someone can follow up on it.

Your Own Voice

FIX IT

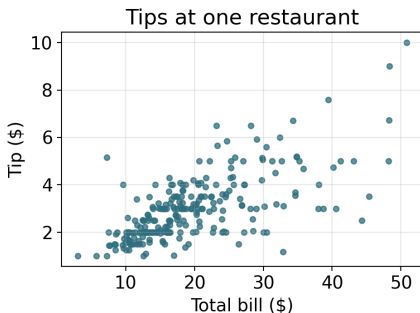


The plot shows a strong correlation between the variables.

Describe this plot in two or three sentences, in your own words.

Your Own Voice

ANSWER



One good version

Bigger bills get bigger tips, but the points spread out more as the bill gets bigger. One bill of about \$33 left a tip of about \$1.

Name the variables, say which way they go, and say what is odd.

The Shape of a Write-Up

- 1 **Question:** what are you trying to find out?
- 2 **Data:** where from, what is one row, what is missing?
- 3 **Methods:** what did you do and why?
- 4 **Results:** plots and numbers, each with a few sentences.
- 5 **Conclusion:** what did you learn, what next, what are you unsure about?
- 6 **Sources:** data, articles, code, AI.

Markdown headings for each part. Restart and run all before you open your Pull Request!

You Try

- 1 Open your HW 6 write-up notebook.
- 2 Pick one section or one plot and rewrite it using today's ideas.
- 3 Trade with a teammate: one thing that is clear, one place you had to guess.
- 4 Add a Data section and a Sources section if you do not have them yet.

Commit it to your branch like normal. It is part of HW 6, nothing extra to turn in!

Data Science Interview Practice

SAY IT OUT LOUD

Explain to a partner as if you were in an interview:

- Tell me about a time you explained a technical result to someone who does not code. How did you do it?
- How do you make sure someone else can reproduce your analysis?
- You find a mistake in your analysis after you already shared it. What do you do?

Questions?

Questions?