Week 06

Boxplots & Audience

INFO 3402: Information Exposition

Brian Keegan, Ph.D.
Assistant Professor, Information Science
brian.keegan@colorado.edu

University of Colorado Boulder

Course Overview

Module	Week	Dates	Computational skill	Communication skill
Shaping	1	Jan 11, Jan 13	Loading	Documentation
	2	Jan 18, Jan 20	Aggregation	Summarization
	3	Jan 25, Jan 27	Joining	Validation
	4	Feb 1, Feb 3	Tidying	Tables
Distribution	5	Feb 8, Feb 10	Histograms	Perception
	6	Feb 15, Feb 17	Box plots	Audience
Comparison	7	Feb 22, Feb 24	Cat plots	Objectives
	8	Mar 1, Mar 3	Faceted plots	Simplicity
Trend	9	Mar 8, Mar 10	Line plots	Trust
	10	Mar 15, Mar 17	Stacked plots	Annotation
	11	Mar 22, Mar 24	Spring Break	
Relationship	12	Mar 29, Mar 31	Scatter plots	Fallacies
	13	Apr 5, Apr 7	Heatmaps	Persuasion
Spatial	14	Apr 12, Apr 14	Choropleths	Conventions
	15	Apr 19, Apr 21	Point plots	Design
Projects	16	Apr 26, Apr 28	Projects	

Shift in course structure

- We've been focusing on the computational side and slighting the communication side
  - Module Assignment 01 helped us calibrate where class needs are

- This week will have a stronger focus on the communication side (with some computation)
  - Weekly assignment will also support framing, analyzing, and writing up your MA02

- The first week of a module will be more computational, the second week more communication

- No changes in pace of weekly assignments, weekly quizzes, or module assignments

Weekly overview

- Notebook exercises (ungraded) – Assigned Tuesdays and reviewed Thursdays

- Add a Visualization Critique (ungraded) – Discussed in class on Thursdays

- Complete "Thursday Questions" form! (ungraded/optional) – Review in class on Thursdays

- Weekly Assignment (graded, 2%) – Assigned Tuesdays and due Sundays by midnight

- Weekly Quiz (graded, 1%) – End of class on Thursdays (open 12:00pm – 12:30pm)

- Module Assignment 02 (graded, 5%) – Due Tuesday, February 22 by 11am*

* We'll discuss due dates in a few slides

Readings

- Questions for Friday's Weekly Quiz 06 will be drawn from these readings
- It's important to be exposed to many examples of data visualization and storytelling!

- Dykes, B. (2019). "Chapter 4: The Anatomy of a Data Story."
  - Link goes to a CU Libraries page → click "Full Text Online" → Bulleted icon in upper left → Chapter 4
  - Data storytelling continuum, six elements, storytelling responsibilities, audience considerations
- Drucker, et al. (2018). "Chapter 9: Communicating Data to an Audience."
  - Audience knowledge of data, jargon, and visualization? What does audience know *vs*. want?
- Bach, et al. (2018). "Chapter 5: Narrative Design Patterns for Data-Driven Storytelling."
  - Types and applications of narrative patterns (argument, flow, framing emotion, engagement)

Weekly Assignment 06

○ **Skills**: Object-oriented style, histograms, annotation + thinking about audience for MA02

○ **Data**: U.S. House Statements of Disbursements

Weekly Assignment 04

WA04Q03 bug

- "AttributeError: 'numpy.ndarray' object has no attribute 'get_figure'"

  - **What's going on?** We're trying to make a figure with multiple subplot axes and getting an error. I haven't been able to reproduce this error, but the fact that at least half the class has it makes it very real.

  - **Why is this happening?** Some earlier versions of pandas and/or matplotlib incorrectly handle the axes objects within the "numpy.ndarray" objects

  - **How can we fix it?** Updating to pandas ≥ 1.4.0 and matplotlib ≥ 3.5.0 fixes it for some people. These are the versions I have as well and I don't get the error.
    ```
    conda update --all
    ```

    - Confusingly, I don't get the error when I go to earlier versions either…

    - Some students can't update: Try `conda install anaconda=2021.11` and then `conda update -all`

- This is severe enough to warrant filing an issue with the pandas developers. If you're interested in learning about open source development, email me and we can get this reproduced, documented, reported, and fixed (someday!).

Module Assignment 02

Module Assignment 01 feedback

- Write for a general audience, not to me and Samantha

  - Open with a question, provocation, case …anything to engage the reader

  - Do not assume the reader knows anything about data analysis and visualization, programming, *etc*.

  - "Like you said in lecture, I did X"

- There is a place for tutorial-style writing on Medium, but that's not what we're focused on

  - "I used X_df and the Y method to turn this into Z which looks like this"

  - We don't want a step-by-step narration of your analysis; we want narratives with polished results

- Quality matters

  - Use high-quality saved figures (docs) rather than screenshots whenever possible

  - No glaring spelling or grammatical problems; professional tone and style

Module Assignment 02

- Use the House expenditure data to tell a **<u>story</u>** about a US House disbursement data
  - You're welcome to bring in other data if it helps with your story, but disbursement data at a minimum
  - Emphasis on **<u>story</u>**! Engage the reader, identify characters/issues, build and resolve tension
  - Use WA06 to think about the audience you want to write for

- 700 – 1000 words with at least one visualization (extra credit for visualizations with annotation!)

- Module Assignment 02 will be due on Tuesday, February 22 by 11am
  - Submit URL of your Medium post to Canvas or save and submit as an HTML file
  - Tag your post on Medium with "INFO3402S22A1" and whatever other tags you'd like

U.S. House Disbursements

U.S. House Statements of Disbursements

- The U.S. House (but not the Senate!) publishes reports of all the office spendings of its members
  - Members have offices in D.C. and home district; staff; travel; office services
- Raw data from House Chief Administrative Officer available here: https://www.house.gov/the-house-explained/open-government/statement-of-disbursements/archive
  - PDF format before 2016 😖
- Cleaned data and documentation from Sunlight Foundation/ProPublica available here: https://projects.propublica.org/represent/expenditures
  - CSVs to 2009! 😃

- Expenditures also available by committee
  - Some committees are large and very active; others are smaller and less active

Exploratory questions

- Shifting personal or campaign expenses into office expenditures is unethical to criminal
  - What could be fingerprints of this kind of behavior?
- How is spending changing over time?
- Which representatives are the greatest outliers in their spending (high or low)?
- How do expenses differ by party? Geography? Seniority? Election competitiveness? Building?
- Also check out Clerk's member information (historical) and election data
  - District codes and members' bioguide IDs are your friend when it comes to keys for merging!

Exploratory questions need to be refined to more specific analyses, audiences, motivations!

Boxplots & Annotations

Notebook time!

○ Download "Week 06 – Lecture.ipynb"

    ○ We'll be using data from previous classes

Audiences

Four different articles using same dataset

○ Even months = `{February, April, June, August, October, December}`

○ Even years = `{1992, 1994, 1996, 1998, 2000, 2002}`

○ All these articles are about how politicians spend tax dollars on "themselves"

○ If you were born in an **even month** and an **even year** read this:
https://www.propublica.org/article/turnover-in-the-house-who-keeps-and-who-loses-the-most-staff

○ If you were born in an **even month** and an **odd year** read this:
https://www.buzzfeednews.com/article/johnstanton/taxpayers-foot-congressmans-hotel-bills

○ If you were born in an **odd month** and an **even year** read this: https://rollcall.com/2015/03/17/who-polices-how-members-spend-office-budgets/

○ If you were born in an **odd month** and an **odd year** read this:
https://www.businessinsider.com/congressional-staff-student-loan-debt-special-payment-tools-2021-12

Partner questions – what is

- What were your main takeaways from the article? What is at stake here?

- What kinds of analyses and visualizations did the article use… if any?

- Were you the audience for this article? Why? What turned you off or never engaged you?

- Who is the audience for this article? Why? What are cues from the language, framing, *etc*.?

Knaflic, C. (2020). Chapter 1. *Storytelling with Data: Let's Practice!*

Partner questions – what could be

- What could the author(s) have done to better understand the needs of an audience like you?

- What are issues that you care about related to national politics and political spending?

- What could have made you the audience for this story? Framing, characters, evidence, format…

- What are some missed potential audiences for this story?
  - List out some personas, what they care about, and why

- What are some kinds of analyses or visualizations that could have improved the story?
  - Sketch out some examples!

- If we reversed the framing, how would that clarify the audience or analysis strategy?

Knaflic's "Big Idea" Worksheet

- Who is your audience?
  - Who are the people or groups with whom you're communicating?
  - How could that group be narrowed?
  - What does your audience care about?
  - What action does your audience need to take?
- What is at stake?
  - What are benefits and risks of audience acting how you want?
- Form your big idea
  - Articulate a point of view
  - Convey what is at stake
  - What is a concise and complete statement of the idea?

Knaflic, C. (2020). Chapter 1. *Storytelling with Data: Let's Practice!* & Big Idea Worksheet.

Next class

Next Class

- Review concepts and exercises from last class
  - Complete "Thursday Questions" form! https://forms.gle/Kdu5Coynu4qinJJZ6 (ungraded/optional)

- Critique a data narrative or visualization
  - Share a link & a few sentences about a data story or visualization (ungraded/optional)

- Time to work on Weekly Assignment
  - Weekly Assignment due on Sunday by midnight by submitting HTML notebook to Canvas

- Weekly quiz at the end of class (12:00–12:30)