Week 12

Scatterplots

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	Type	Skill
Shaping	1	Jan 11, Jan 13	Computation	Loading
	2	Jan 18, Jan 20	Computation	Aggregating
	3	Jan 25, Jan 27	Computation	Joining
	4	Feb 1, Feb 3	Computation	Tidying
Distribution	5	Feb 8, Feb 10	Computation	Histograms
	6	Feb 15, Feb 17	Communication	Audience
Comparison	7	Feb 22, Feb 24	Computation	Cat plots
	8	Mar 1, Mar 3	Communication	Persuasion
Trend	9	Mar 8, Mar 10	Computation	Time series
	10	Mar 15, Mar 17	Communication	Uncertainty
	11	Mar 22, Mar 24	Spring Break	
Relationship	12	Mar 29, Mar 31	Computation	Scatter plots
	13	Apr 5, Apr 7	Communication	Fallacies
Spatial	14	Apr 12, Apr 14	Computation	Choropleths
	15	Apr 19, Apr 21	Communication	Conventions
Projects	16	Apr 26, Apr 28	Projects	

Readings

- Questions for Friday's Weekly Quiz 12 will be drawn from these readings
  - Schwabish (2021). Chapter 8: Relationships.

  - Wilke (2019). Chapter 12: Visualizing Associations Among Two or More Quantitative Variables.

  - Yau (2011). Chapter 6: Visualizing Relationships.

- Skim through the tutorial for Scatterplots on Python Graph Gallery.

Weekly Assignment 12

- **Skills**: Review joining data, visualizing relationships, customizing scatterplots

- **Data**: U.S. county data on COVID, elections, and Census

Module Assignment 04

Module Assignment 04

- Use Wikipedia data to tell a story about the production and/or demand for information about some topic or event
  - Current event, VIP, popular culture. anniversary, front page article, trending articles, comparing things
    - Russia-Ukraine conflict if you absolutely want, but it's not terribly original
  - You're welcome to use and explore anything in wikifunctions!
  - Explore other wikis like Conservapedia, Fandom, TVTropes, *etc*.
    - Change the endpoint and most wikifunctions should still work! (get_pageviews won't work outside of Wikipedia)

- 700 – 1000 words with at least one visualization

- Module Assignment 04 will be due on **Wednesday, March 30 by 11:59pm**
  - Wednesday after Spring Break
  - Submit URL of your Medium post to Canvas or save and submit as an HTML file
  - Tag your post on Medium with "INFO3402S22A4" and whatever other tags you'd like

Outline

- Describe why you or anyone might be interested in the topic

  - What's the tension or hook?

- Some background on Wikipedia and why you're looking at this metric

  - Article size, number of contributors, pageviews, *etc*.

- A visualization of at least property of a Wikipedia article (or several) changing over time

  - Extra credit for including annotations!

- Some discussion of the interesting features of the time series

  - Trends, patterns, anomalies, *etc*.

- Some discussion about what these findings reveal

  - Information seeking, production, biases, upending some conventional wisdom, *etc*.

Module Assignment 2 shout-outs

○ Jessie Bart – House Food and Beverage Spending

  ○ Great setup and motivation, links to external stories, contrasting with prevalence of food insecurity

○ Bailey Gimpel – Congressional Expenditures on Equipment

  ○ Deeper dive into the Congressional offices spending more than median on equipment

○ Brian Lee – Big Spenders on Travel in the House

  ○ Breaking down different kinds of travel expenses and comparing member vs. officer spending

○ Michael Murdock – Congressional Spending across Parties

  ○ Differences in spending categories by parties and deeper dive into differences within travel

○ Estevan Sandoval – You've Got Mail, Just Not for Long

  ○ Background on franking as an expense and deep dive into top and bottom spenders on franking

○ Ken Vue – 2021Q3 Meal Expenditures

  ○ Looking at meal expenses per day and per receipt, finding some troublesome outliers

Module Assignment 05

Module Assignment 05

○ Use the U.S. county data to identify an *unusual* relationship between variables

  ○ Between "us_counties.csv", "analytic_data2021.csv", and "Unemployment.xlsx" there are close to 1,000 different county-level variables to explore

  ○ A relationship between poverty and income is **not** unusual: it's to be expected!

  ○ Other relationships might be trivial: employment from one dataset is likely similar with employment from another

  ○ The relationship should be strong-ish: if it's linear, a correlation above 0.2 or below -0.2

  ○ Identifying this relationship can be top-down (sorting, correlograms, pairplots, *etc.*) or bottom-up (exploring pairs)

  ○ Like WA12 Questions 2 & 3, explore whether this relationship shows up or disappears in related variables

  ○ Make a case for there being a causal rather than a random relationship between these variables

○ 700 – 1000 words with at least one visualization

○ Module Assignment 05 will be due on **Wednesday, April 13 by 11:59pm**

  ○ Submit URL of your Medium post to Canvas or save and submit as an HTML file

  ○ Tag your post on Medium with "INFO3402S22A5" and whatever other tags you'd like

Next class

Notebook

○ Download "us_counties.csv" and "Week 12 – Lecture.ipynb"

○ Use scatterplot functions in different visualization libraries

○ Customize scatterplot appearance

○ Explore different types of relationships and their correlations

○ Basic modeling of linear relationships

○ Visualizing relationships with pairplots, correlograms, and clustermaps

Next class

Next Class

○ Review concepts and exercises from last class

    ○ Complete "Thursday Questions" form! https://forms.gle/VjuzDawFkMaJmMfM7 (ungraded/optional)

○ Time to brainstorm and work on Weekly Assignment 12

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