CS 638: Data Science in Wisconsin

Website: https://wisc-ds-projects.github.io/f20/

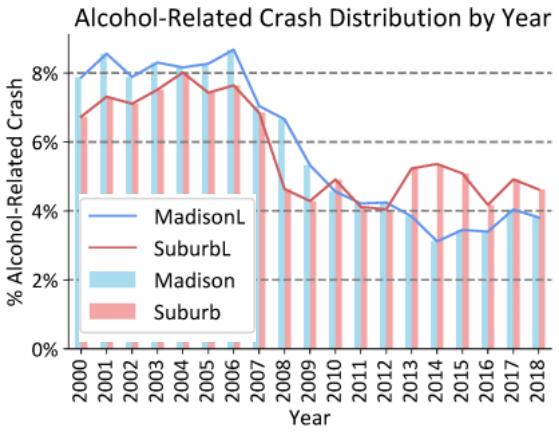
Email: compsci638-2-f20@g-groups.wisc.edu (~20 people -- have discussions+share resources)

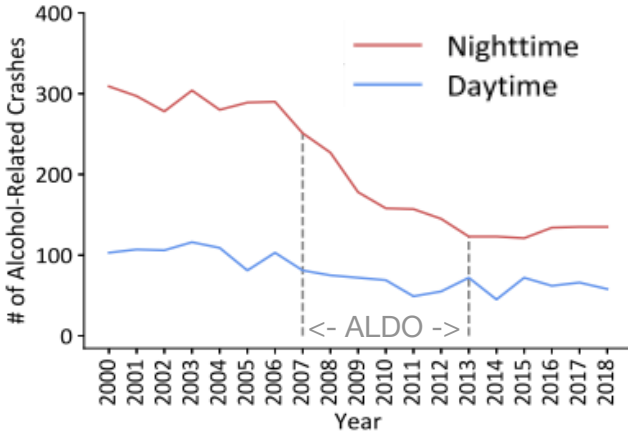
Optional text: *The Visual Display of Quantitative Information* by Edward R. Tufte

Note: 2-credits, letter grades, doesn't count for CS elective

Logistics:

- I won't "teach" much per se, but I'll give detailed feedback on your work
- Form teams of 2-4, choose topic (due by end of the weekend!)
- Every week, each person updates team's draft report (in markdown, HTML, or Latex)
- Draft should be ready before Monday meeting; be prepared to present to class
- Plots should have explicit questions, observations, conclusions





What makes a good plot? (above examples adapted from student plots)

- Obvious stuff: legible font, labeled axes, correct data (use common sense!)
- Lots of **context** (annotations, interesting subcategorizations); never show a single line or set of bars. Rule of thumb: maximize how many numbers are represented (https://trailsofwind.figures.cc/)
- Audience is **intelligent, busy, not STEM.** Explain, minimize mental load; don't "dumb down"
- Minimize non-data ink: remove gridlines, top/right borders, box around legend (be **minimalist**)
- **Design before code**: think about how it should look, then write necessary code to make it. Don't start with the things that are easy to do with matplotlib and force your data to conform
- Interesting data: you don't know until you plot it. You should make MANY plots each week, then **only show us the best of the best**. Be ready to verbally summarize other results.
- Communicate about the plot verbally and with text. What are you asking? What can you conclude? What new questions arise? What is your **methodology**; what are threats to validity?
- Narrative: the plots should be sequenced in a meaningful way to **tell a story**
- **Printer friendly**: colors are OK, but only if they work if turned to gray

Real data is not like "toy" data used in classes

- Missing, different people collect different ways, practice change over time, incorrect entries
- Is dirty data better than no data? Data quality can itself be an object of study (e.g., heartbeat data).
- Lacking data should never stop you (conduct a survey, https://www.reddit.com/r/madisonwi.json)

Evaluation

- I'll "grade" each student's weekly set of plots on a 0-5 scale, roughly as follows. **0:** no plot; **1:** plot hard to read or has mistakes; **2:** data is "thin" and lacks context; **3:** well-designed data-dense plot; **4:** the plot is used to make a convincing argument; **5:** the plot is so compelling I'll want to share it with others. Other factors include the methodology, written+verbal communication, how will it fits with other plots of the team, etc.
- **Quality over quantity**: can you produce a visualization that everybody in Madison will want to see and talk about? https://paulbutler.org/2010/visualizing-facebook-friends/
- If your final presentation is amazing, I'll grade you solely on that, regardless of the plot grades throughout the semester.
- A good final presentation conveys most interesting findings from the report, has actionable suggestions tailored to the audience, and motivates people to read the original document.

Formats

Example formats: https://github.com/wisc-ds-projects/wisc-ds-projects.github.io/tree/master/f20/example

Have one private repo per team. Add instructor to it. I'll give feedback following each Monday class.
https://git-scm.com/download/win or https://git-scm.com/download/mac

Markdown (.md)

- simplest
- can embed images with <img src="...">
- section headers: **# header**
- links: **[link name](link URL)**

HTML (.html)

- more control over styling

Latex (.tex => .pdf)

- get it here: https://www.latex-project.org/get/
- can run **pdflatex** to compile .tex files to a PDF
- very elegant typesetting
- embed PDF images using **\includegraphics[width=\columnwidth]{filepath}**
- reference figures with **\label{name}** and **\ref{name}**