ODSC Data Visualization Workshop (Notes)

FAQ on ggplot2:

- What is ggplot2?
  - a graphing system that accurately "describes the properties of a plotting system"
- Why use ggplot2?
  - ggplot2 makes it *"Easy to reason about how data drives visualization"*, *"Easy to iterate"*, and *"Easy to be consistent"*
- Why does ggplot2 use the `+` instead of the pipe (`%>%`)?
  - From Hadley's interview, *"I think I was reading about operator overloading and I thought "Oh maybe I could do this with '+' instead", and it kind of makes sense, you know, because you're adding layers to the plot"*
- What is the future of ggplot2?
  - *"the fundamental challenge"* of *"harness[ing] the best of static visualization and the best of interactive visualization"* - Hadley Wickham
    - *"I think the future of visualization in R is fundamentally and inextricably tied to visualization in JavaScript"*; *"one of the big problems with JavaScript visualization is how do you turn that into a publishable artifact"* - Hadley Wickham
- How has ggplot2 changed data visualizations?
    - *"My general thesis of visualization is that the quality of the best visualization has maybe improved 10% in the last 150 years. The best visualization you can make today is only slightly better than the best visualization someone could make 150 years ago. But the time it takes you to make them has probably decreased by three orders of magnitude."* - Hadley Wickham

---

(Rough) Notes on EDA:

1. Understand the dataset
   - [ ] Look at the 'rectangle'
2. Note the metadata
   - [ ] document the dimensions, size, etc.
     - Check out the inspectdf package
3. Get your summary stats:
   - [ ] document with `skimr::skim()`, `dplyr::group_by()`, `dplyr::summarize()`

- skimr package
- dplyr summary functions

4. What is the question you're trying to answer?
   - [ ] Write down any assumptions or expectations
     - What are your assumptions about the relationships/distributions for the variables in the dataset?
5. Start with univariate graphs:
   - [ ] tables, histograms, density/ridge-line plots, box-plots, violin plots, etc.
     - Check out janitor::tabyl()
   2. Are these what you expected?
      - [ ] yes ➜ confirm our assumptions & document findings
      - [ ] no ➜
        - [ ] adjust our previous expectations or
        - [ ] re-consider the validity of the data
          - [ ] Communicate with stakeholders about expectations/ assumptions
6. Bivariate EDA:
   - [ ] 2x2 tables, bar/column graphs, trend lines, scatter plots, box-plots with categorical variables
   2. Are these what you expected?
      - [ ] yes ➜ confirm our assumptions & document findings
      - [ ] no ➜
        - [ ] adjust our previous expectations or
        - [ ] re-consider the validity of the data
          - [ ] Communicate with stakeholders about expectations/ assumptions
7. Multivariate EDA:
   - [ ] Bubble charts, stacked bar/column graphs, facets, etc.
     - Check out raincloud plots
   2. Are these what you expected?
      - [ ] yes ➜ confirm our assumptions & document findings
      - [ ] no ➜
        - [ ] adjust our previous expectations or
        - [ ] re-consider the validity of the data
          - [ ] Communicate with stakeholders about expectations/ assumptions