STAT 231: Problem Set 1B

Jack Dove

due by 5 PM on Friday, September 4

Series B homework assignments are designed to help you further ingest and practice the material covered in class over the past week(s). You are encouraged to work with other students, but all code must be written by you and you must indicate below who you discussed the assignment with (if anyone).

Steps to proceed:

1. In RStudio, go to File > Open Project, navigate to the folder with the course-content repo, select the course-content project (course-content.Rproj), and click "Open"
2. Pull the course-content repo (e.g. using the blue-ish down arrow in the Git tab in upper right window)
3. Copy ps1B.Rmd from the course repo to your repo (see page 6 of the GitHub Classroom Guide for Stat231 if needed)
4. Close the course-content repo project in RStudio
5. Open YOUR repo project in RStudio
6. In the ps1B.Rmd file in YOUR repo, replace "YOUR NAME HERE" with your name
7. Add in your responses, committing and pushing to YOUR repo in appropriate places along the way
8. Run "Knit PDF"
9. Upload the pdf to Gradescope. Don't forget to select which of your pages are associated with each problem. *You will not get credit for work on unassigned pages (e.g., if you only selected the first page but your solution spans two pages, you would lose points for any part on the second page that the grader can't see).*

If you discussed this assignment with any of your peers, please list who here:

ANSWER:

MDSR Exercise 2.5 (modified)

Consider the data graphic for Career Paths at Williams College at: https://web.williams.edu/Mathematics/devadoss/careerpath.html. Focus on the graphic under the "Major-Career" tab.

a. What story does the data graphic tell? What is the main message that you take away from it?

ANSWER: The data graphic shows the distribution of careers by major at Williams College. The main message that I took away from it is that, while Williams College is dominated by a few majors, most majors display a wide range of career outcomes.

b. Can the data graphic be described in terms of the taxonomy presented in this chapter? If so, list the visual cues, coordinate system, and scale(s). If not, describe the feature of this data graphic that lies outside of that taxonomy.

ANSWER: The designer uses color to differentiate each major. Additionally, he uses curve thickness (area) as his scale, displayed by a noticeably wide curve of economics majors choosing jobs at banking and financial firms. Each profession and major take up a different length of the circle's circumference, showing their relative sizes as well. There isn't much of a coordinate system, but the main story is told by each facet (major and career) and their corresponding connecting curves.

c. Critique and/or praise the visualization choices made by the designer. Do they work? Are they misleading? Thought-provoking? Brilliant? Are there things that you would have done differently? Justify your response.

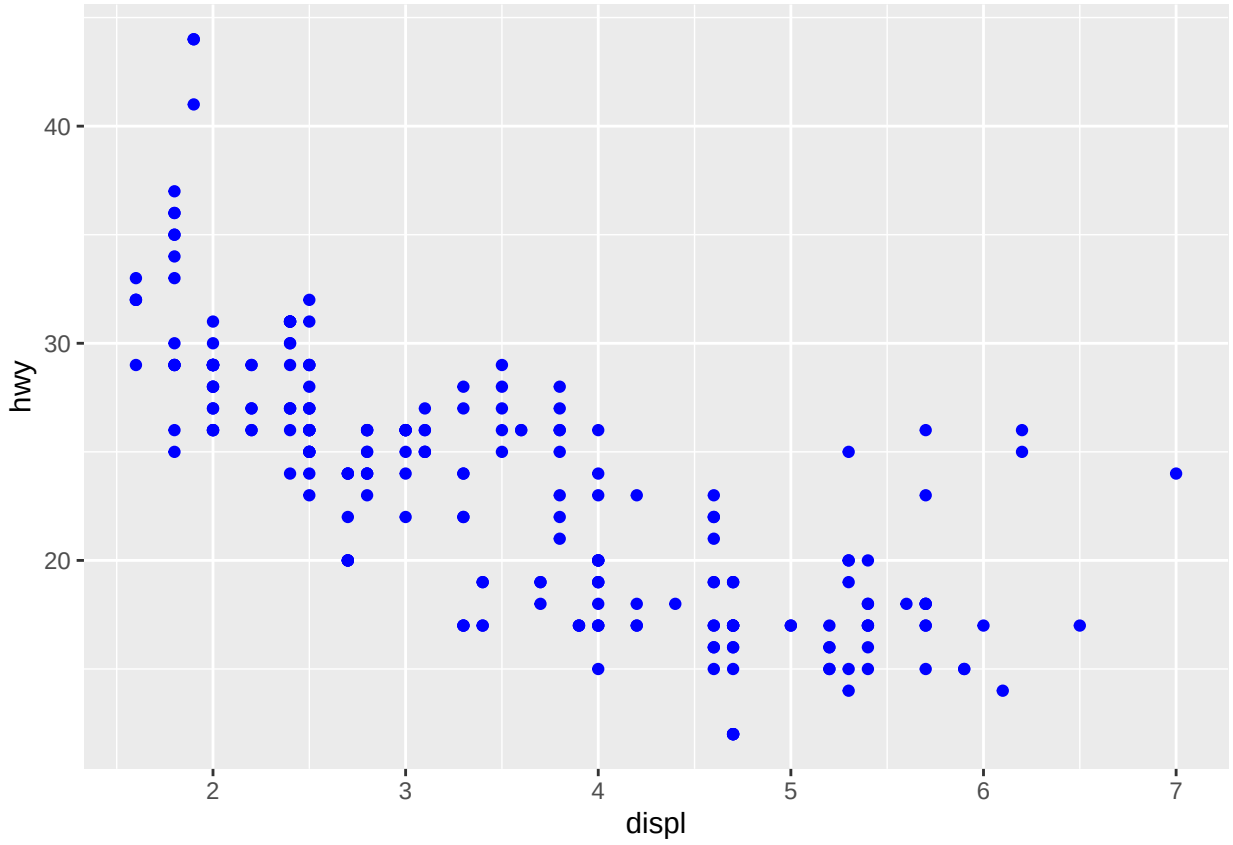
ANSWER: I think the weight of each connecting line and length of each major's piece of the circle is extremely well done; I could immediately tell what most kids major in, work in, and see how my current major performs. I'd brighten the color scale a bit and make it categorical, not sequential. Additionally, I'd try to put the most popular career choice for each major across from it when possible, as then angle could play a factor (the larger the angles of the lines, the greater career spread/divergence there would be for that particular major). Overall, the graphic is very informative yet simple, and the animated hovering makes its use intuitive.

Spot the Error (non-textbook problem)

Explain why the following command does not color the data points blue, then write down the command that will turn the points blue.

> ANSWER: Blue was included in the aes() component, which helps map each variable; however, the color does not help with mapping, but is still a viable input for geom_point, so I added it next to the mapping component, as shown in the corrsponding chart with blue points.

```
library(ggplot2)
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy),  color = "blue")
```

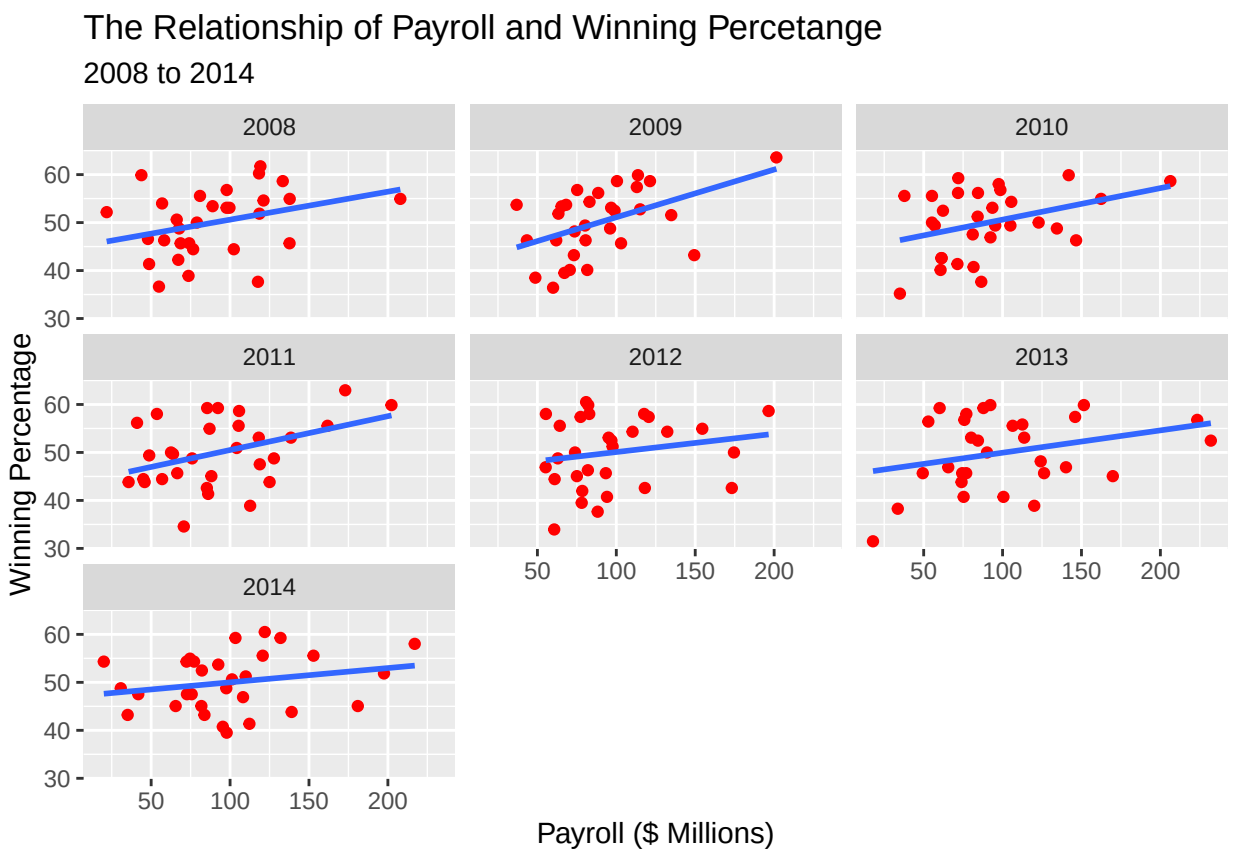


MDSR Exercise 3.6 (modified)

Use the `MLB_teams` data in the `mdsr` package to create an informative data graphic that illustrates the relationship between winning percentage and payroll in context. What story does your graph tell?

> ANSWER: My data graphic shows a consistent link between payroll and winning percentage over the course of seven seasons. All seven facets display positively-sloped trend lines. I faceted years because, since most baseball fans know the best teams of each year, they might find it valuable to see the overall season-by-season relationships between payroll and winning percentage.

```
data(MLB_teams)
ggplot(data=MLB_teams, mapping=aes(x=payroll/1000000, y=WPct*100)) +
  geom_point(color="red") +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    x="Payroll ($ Millions)",
    y="Winning Percentage",
    title= "The Relationship of Payroll and Winning Percetange",
    subtitle = "2008 to 2014"
  ) + facet_wrap(~yearID)
```



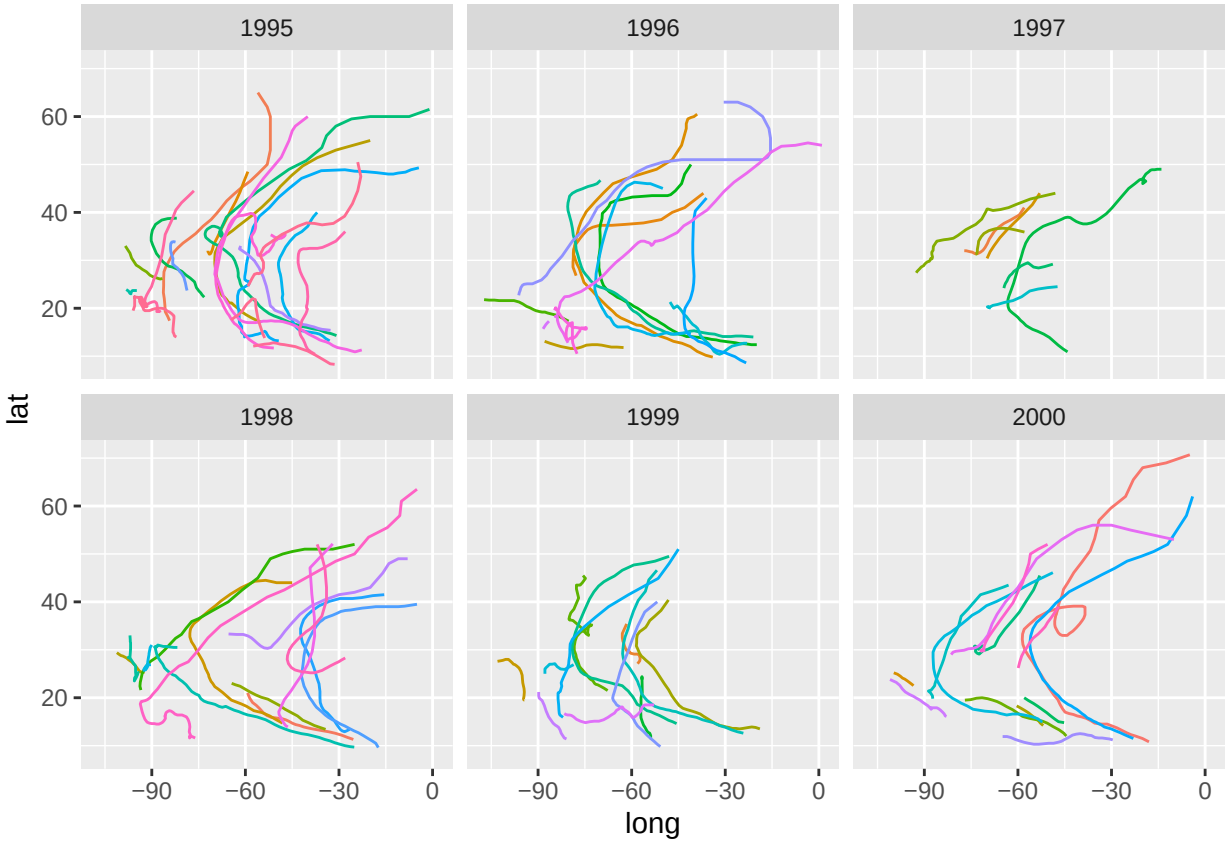
MDSR Exercise 3.10 (modified)

Using data from the `nasaweather` package, use the `geom_path()` function to plot the path of each tropical storm in the `storms` data table (use variables `lat` (y-axis!) and `long` (x-axis!)). Use color to distinguish the storms from one another, and use facetting to plot each `year` in its own panel. Remove the legend of storm names/colors by adding `scale_color_discrete(guide="none")`.

Note: be sure you load the `nasaweather` package and use the `storms` dataset from that package!

```
library(nasaweather)

ggplot(data=storms) +
  geom_path(mapping = aes(x=long, y=lat, color=name)) +
  facet_wrap(~year) + scale_color_discrete(guide="none")
```



Calendar assignment check-in

For the calendar assignment:

- Identify what questions you are planning to focus on
- Describe two visualizations (type of plot, coordinates, visual cues, etc.) you imagine creating that help address your questions of interest
- Describe one table (what will the rows be? what will the columns be?) you imagine creating that helps address your questions of interest

Note that you are not wed to the ideas you record here. The visualizations and table can change before your final submission. But, I want to make sure your plan aligns with your questions and that you're on the right track.

> ANSWER: Question 1: How much time did I spend athletically last fall vs. this fall? Question 2: How many more office hours did I attend this fall vs. last fall? Visualization 1: Stacked bar chart by weeks 1-12 of fall, two categories are 2019 and 2020, tracking athletic time spent in hours. Visualization 2: Campus chloropleth displaying time spent (2 charts: 2019 and 2020). Table: Office hours. Columns: 2019 and 2020. Rows: Statistics, Economics, Other, Total.