STAT 231: Problem Set 1B

Sean Wei

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: Zach Ostrow

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: This data graphic is conveying what major(s) Williams College alums graduated with and then career choice they ended up choosing. You can also look at it from the other perspective - what majors did certain career choices have? The main message that I took away from this was that career choices have many different backgrounds and that your major doesn't always 100% dictate your career path.

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: Yes, the data graphic can be described in terms of the taxonomy presented in this chapter. Visual cues include: color (refers to majors) and arc length (shows how many people are in the majors). Coordinate system: seems polar because coordinates are placed on a radius and angle. Scale: Categorical (the different majors and career choices).

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 thought that this graphic was very interesting and unique. From my understanding, it conveys clearly the relative amount of different majors at Williams College along with what Williams College alums choose as their career path. One thing I would have considered doing differently would be having more drastic color differences between the majors because the compilation of everything looks a little confusing. I also would have appreciated some quantitative values to go along with the visual.

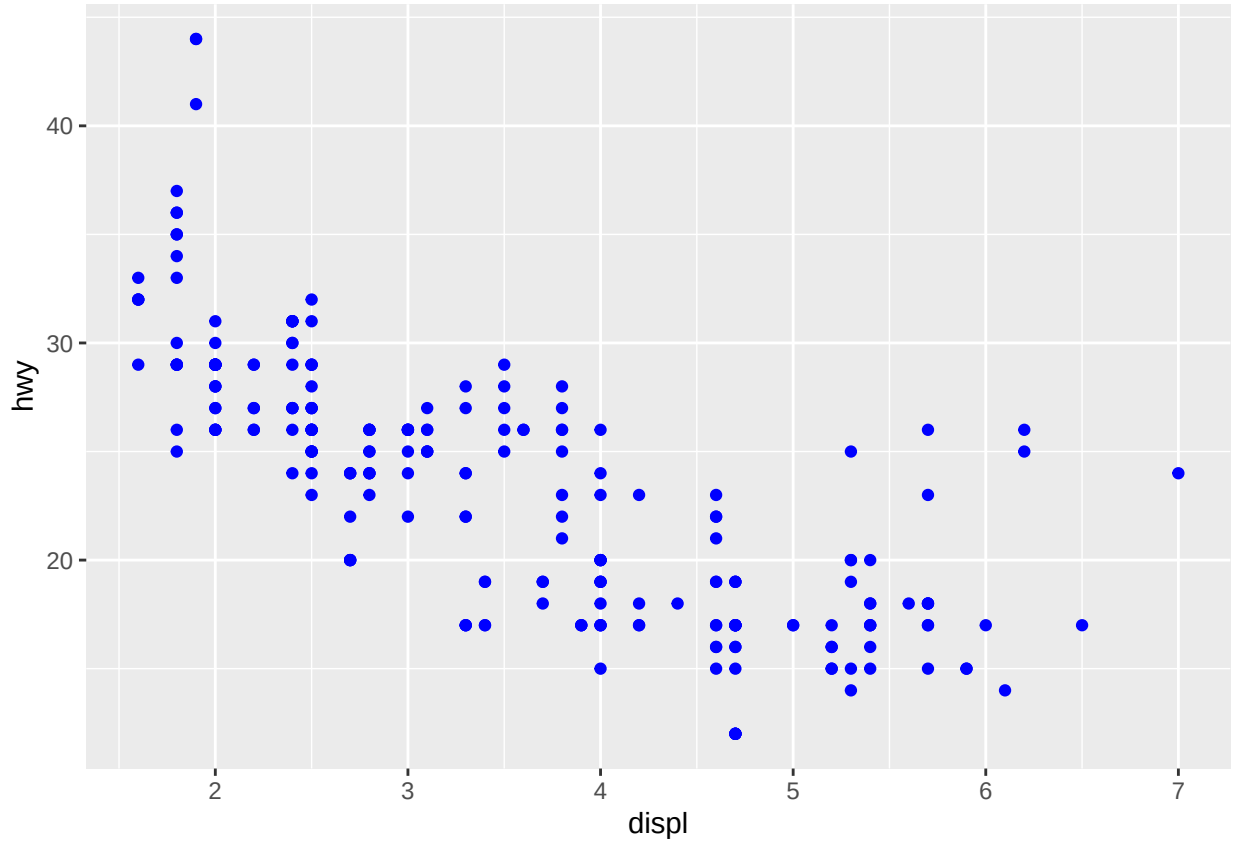
Spot the Error (non-textbook problem)

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

> ANSWER: The following command does not color the data points blue because you need to specify the color outside aes(). The aesthetic values in aes() are data which are properly fit to a scale (which explains why it is interpreted with a legend). In the first command, since color is inside aes(), essentially a new factor variable and legend are generated.

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

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

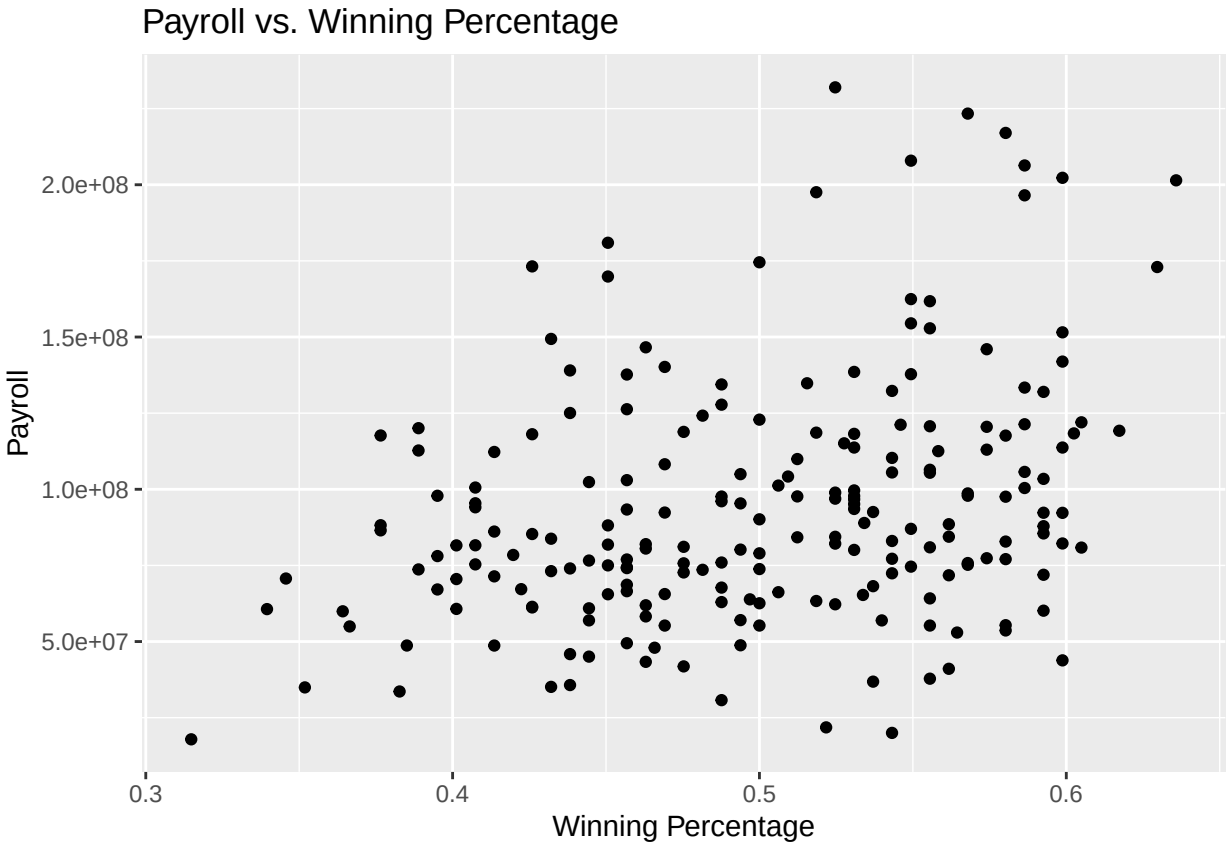


MDSR Exercise 3.6 (modified)

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: From this graph, it is evident that there is a medium to strong positive linear association between winning percentage and payroll. As such, as winning percentage increases we expect payroll to as well.

```
#glimpse(MLB_teams)
ggplot(MLB_teams, aes(x = WPct, payroll)) +
  geom_point() +
  labs(x = "Winning Percentage", y = "Payroll") +
  ggtitle("Payroll vs. Winning Percentage")
```

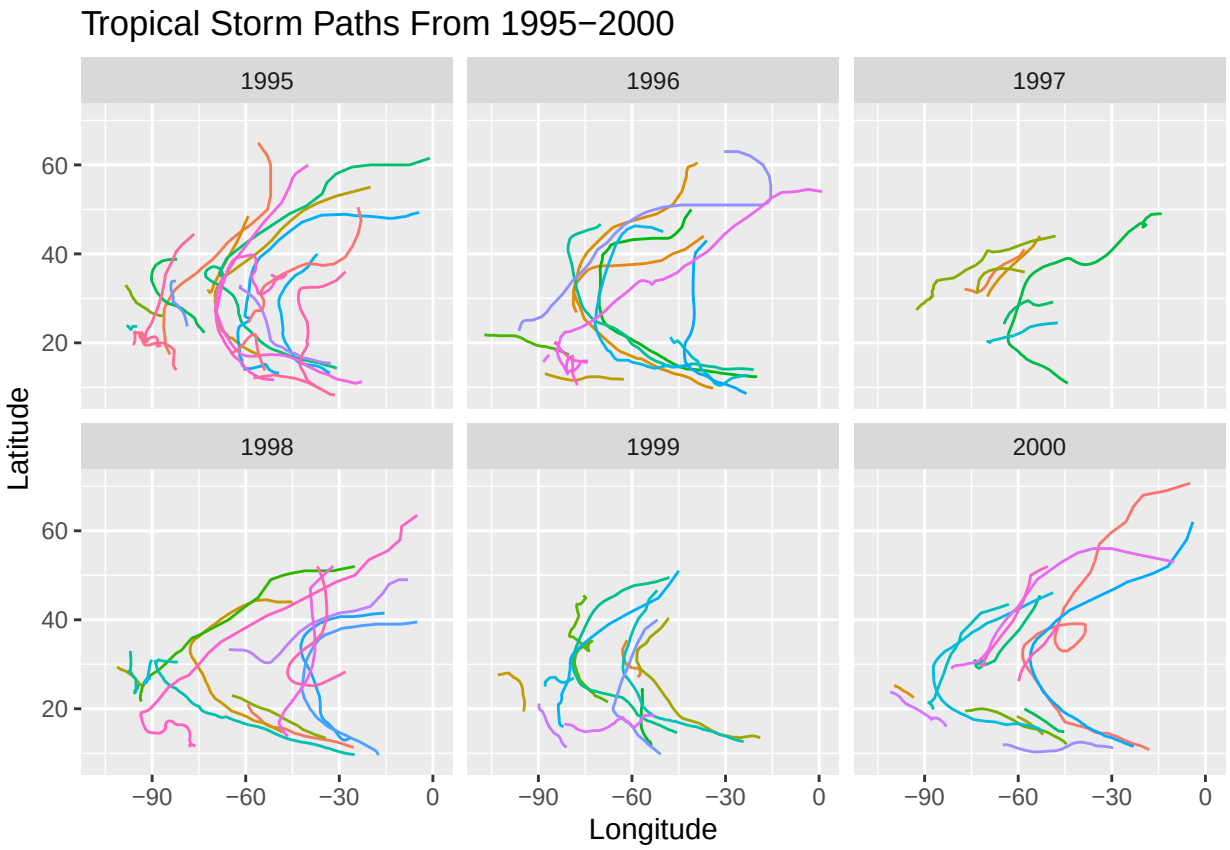


MDSR Exercise 3.10 (modified)

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)
#glimpse(storms)
ggplot(storms, aes(x = long, y = lat)) +
  geom_path(aes(color = name)) +
  facet_wrap(~year) +
  scale_color_discrete(guide="none") +
  labs(x = "Longitude", y = "Latitude") +
  ggtitle("Tropical Storm Paths From 1995-2000")
```



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: I think the questions that I am planning to focus on are exercise time, how much time studying (productivity), and how much sleep I'm getting versus how much I should be getting. One plot could be a time plot where there's a horizontal line signifying 8 hours and I'll overlay that with the amount of sleep I actually got over time. Another visualization could just be how am I spending all of my time, with different colors representing the amount of time I spend on exercise, sleep, work, and leisure each day. For a table, the columns could be the different things I'm tracking (i.e work, sleep, etc.) as well as possibly one where it is a calculation of how far off/how many more hours I was from the expected amount of hours. In that case, each row would be it's own day.