STAT 231: Problem Set 1B

LORRAINE OLOO

due by 5 PM on Friday, February 26

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: Teddy Baraza, Lovemore Nyaumwe

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 graphic data shows: (15600 Williams College Alums) 1. The distribution of majors and the careers they pursue
2. The distribution of majors taken over time since 1930
3. The distribution of the majors and double- majors they took.
My main take away is the distribution of majors at Williams over time, since 1930, and how it is constantly changing.

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 data graphic uses a radial coordinate system and the visual cue of color to distinguish between the 15 majors and the 15 career paths choosen by the alumnis. The scale used is categorical; that is the 15 majors and the 15 career paths. The graphic also have clear labelling to provide the context. The inclusion of the arcs in the graphic lies outside the taxonomy discussed in the chapter.

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.

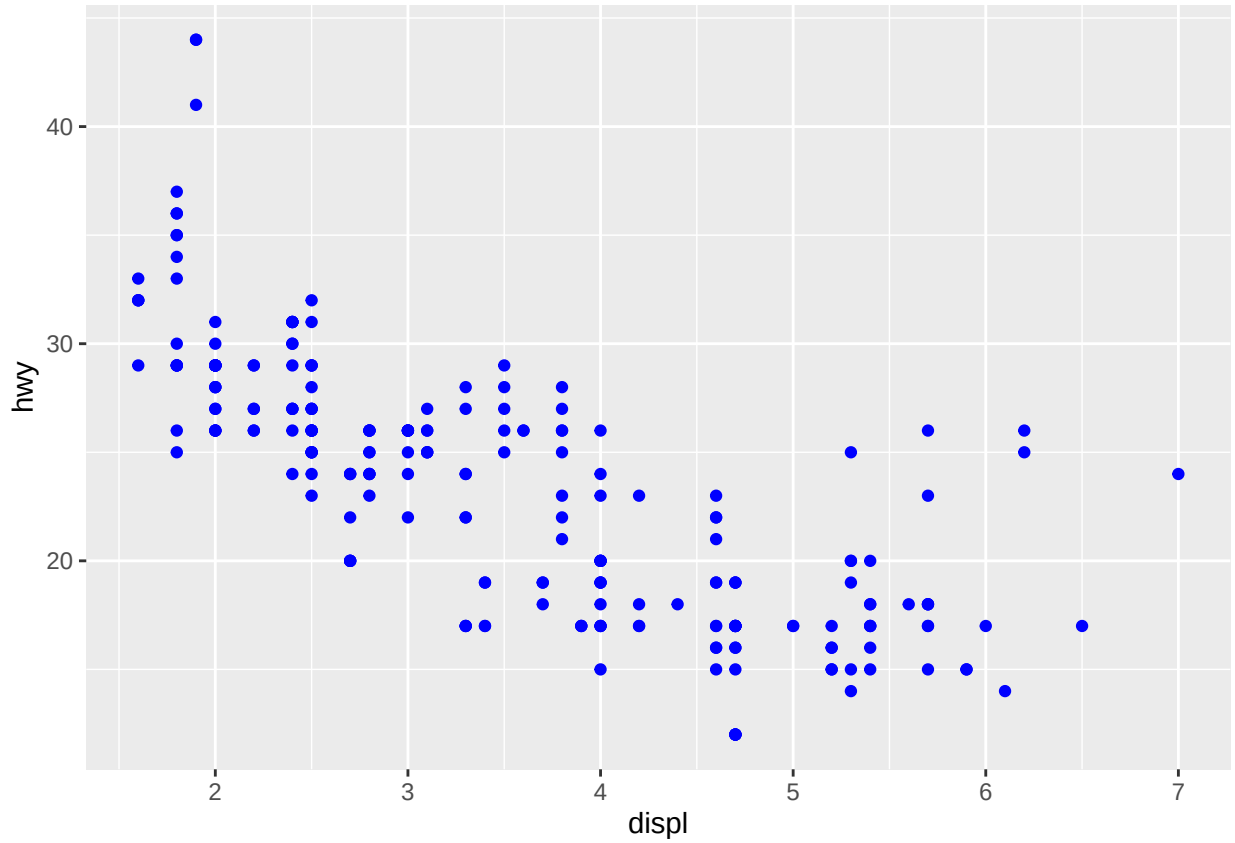
ANSWER: Trying to group all the maojor ideas in one graph was a smart idea. A reader only needed to look through one graph. I found trying to include all the information about the different career paths based on major and double majoring in one pie chart made it quite challenging to understand what was going on. Separating the majors into different charts would have made it easier to understand. A bar chart would have been a better idea and more clear to see. However, this would mean that a reader would have to go through more graphs which is time consuming.

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 color is presented as a label because it is inside the aesthetic function. We need to move it outside.

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



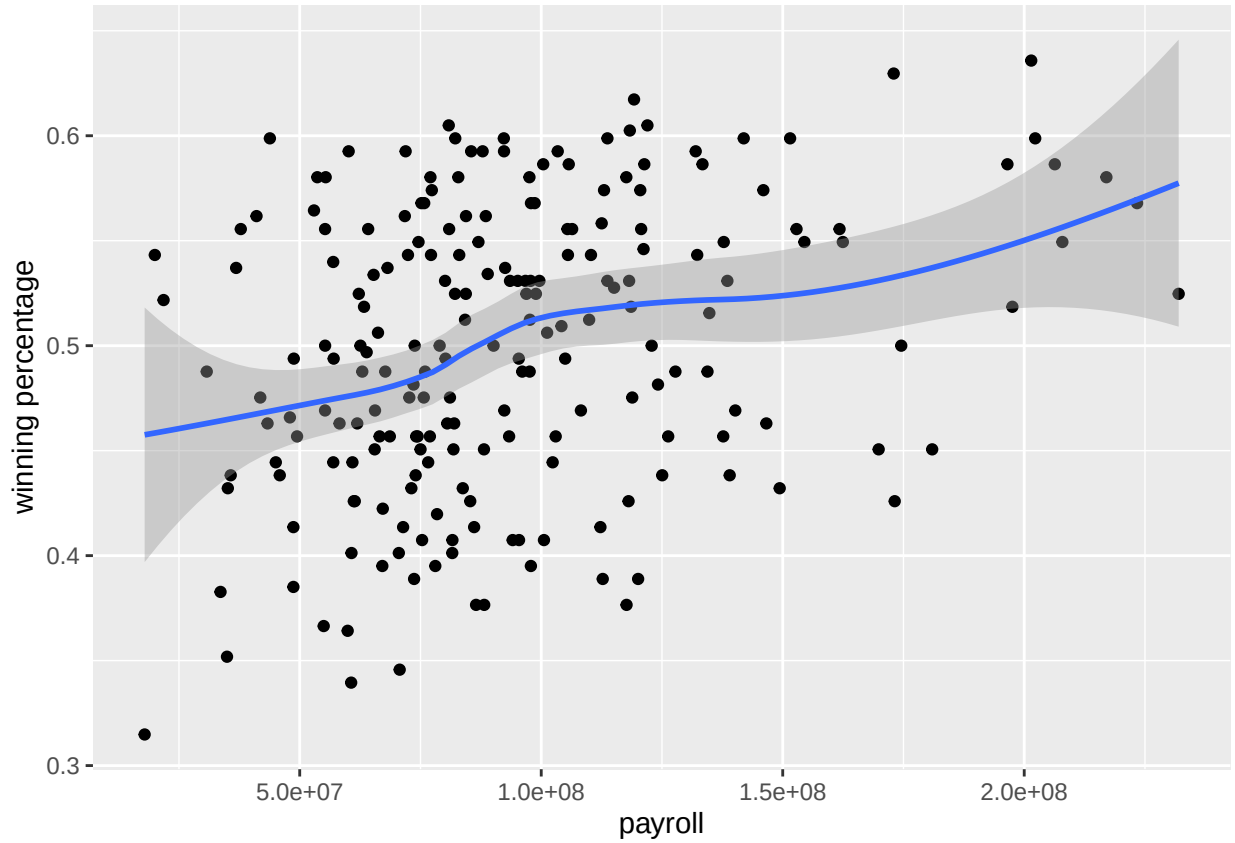
MDSR Exercise 3.6 (modified)

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: The graph shows a moderate positive correlation between the the sum of the salaries of the players on each team (payroll) and the winning percentage. The higher the payroll, the higher the winning percentage was.

```
library(mdsr)


ggplot(
  data = MLB_teams,
  aes(x = payroll, y = WPct)
) +
  geom_point() +
  geom_smooth() +
  xlab("payroll") +
  ylab("winning percentage")
```



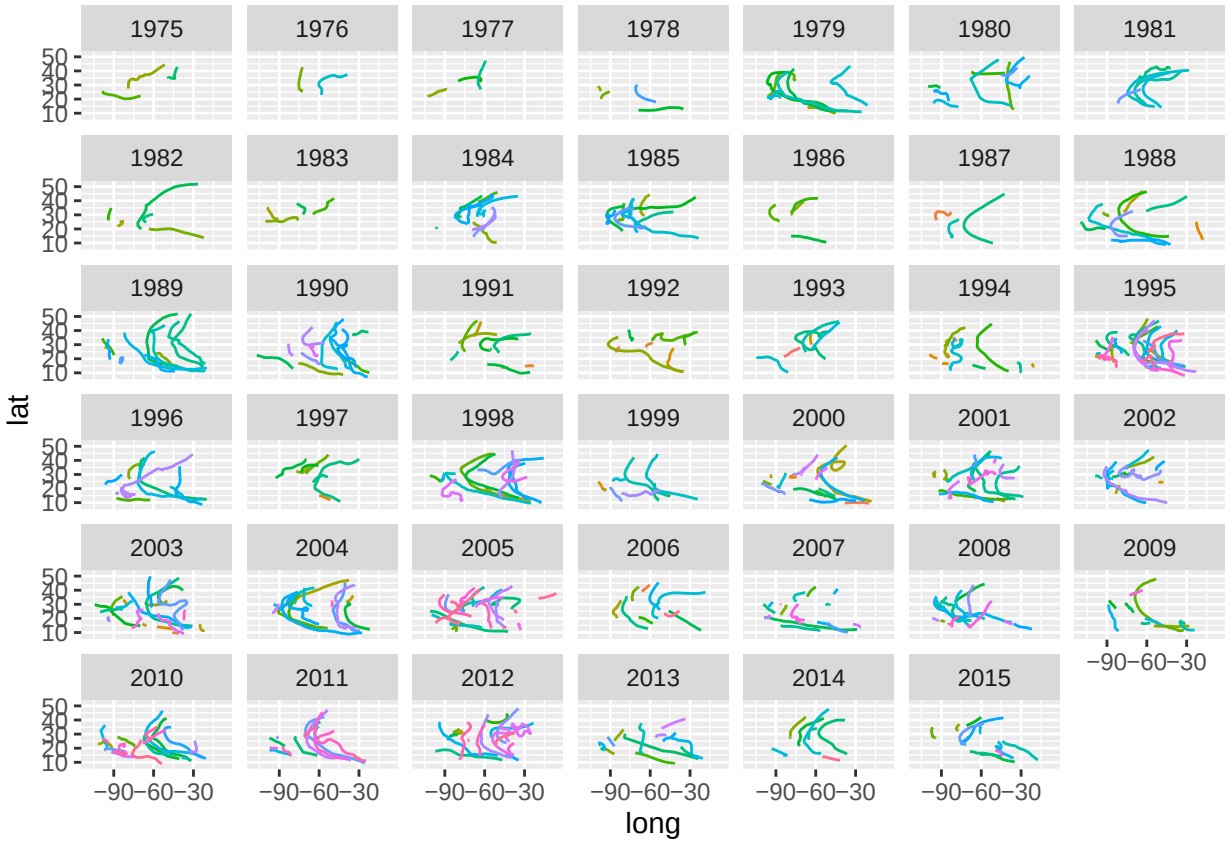
data

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!

```
ggplot(data = storms, aes(x = long, y = lat, color = name))+geom_path() + 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: - Identify what questions you are planning to focus on: 1. How much time do I spend on social media during a typical weekday and a typical weekend?

2. How much time do I spend doing my course assignments vs reading/preparing for the course classes?

- Describe two visualizations you imagine creating that help address your questions of interest
  I will be using bar graphs to show the times. I will use hours to measure the amount of time spent on the activities. To show the difference, I will use color.

-Describe one table
For the first question, the columns will be:
Day of the week, total hours, total minutes, weekday/weekend (I will get this data straight from my phone rather than a google calender and create a spreedsheet that I will make sure to be updating after a week)

For the second question, the columns will be:
Course, reading hours, reading minutes, assignement hours, assignment minutes

The reason for also including minutes is incase the hours are to low to be meaningful