STAT 231: Problem Set 1B

Kim Zhou

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: Jamie Dailey

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 describes the industry alumnae have careers in compared to their major at Williams. My main takeaway was that within one career industry, there's a large variety of different, many unrelated majors. So, it doesn't necessarily matter what one majors in.

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:

This data graphic does have some of the visual cues listed in chapter 2. For example, color coordinates to whether the major is STEM, Humanities, or Social Sciences. Additionally, the area of the arc plays a role as double majors will have a two arcs, each with half thickness. The graphic doesn't appear to have a normal coordinate system and scale is hard to determine since there are no references/legends. What would be the coordinate system in this graphic is instead a circle of the different majors/career industries spread out around the circle. This is outside of our textbook taxonomy.

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:

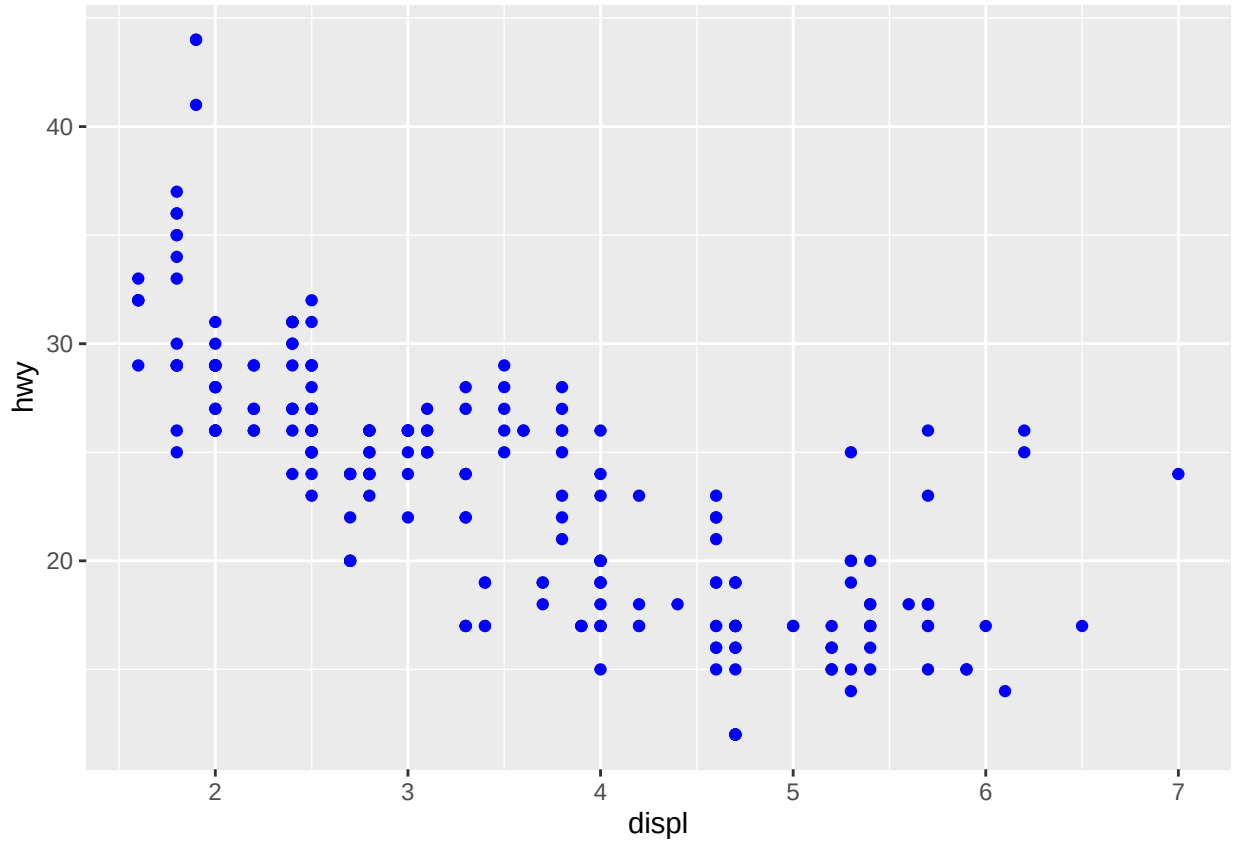
The visualization choice was definitely thought-provoking because it really appears that majors don't play a huge role in where you end up. The lack of any scale is really misleading however. Is there a reason different majors/careers are different sized on the outside? How many people does the super fat arc represent compared to the skinny one? We can only infer these answers and the actual numbers could be more or less than what we expect. I would have provided some quantification to these problems ie every 1/2 inch thickness represents _____ number of people. Additionally, I would have specified what each color represented. It would also be interesting to see how the authors determined which industries certain jobs go in, because often times careers are interdisciplinary.

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:

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



The following command did not turn the points blue because the color specification was in the aes function in geom_point which means that it was being mapped to a variable. Instead, we put the color specification for points in geom_point but left it outside of the aes() so to color all points blue.

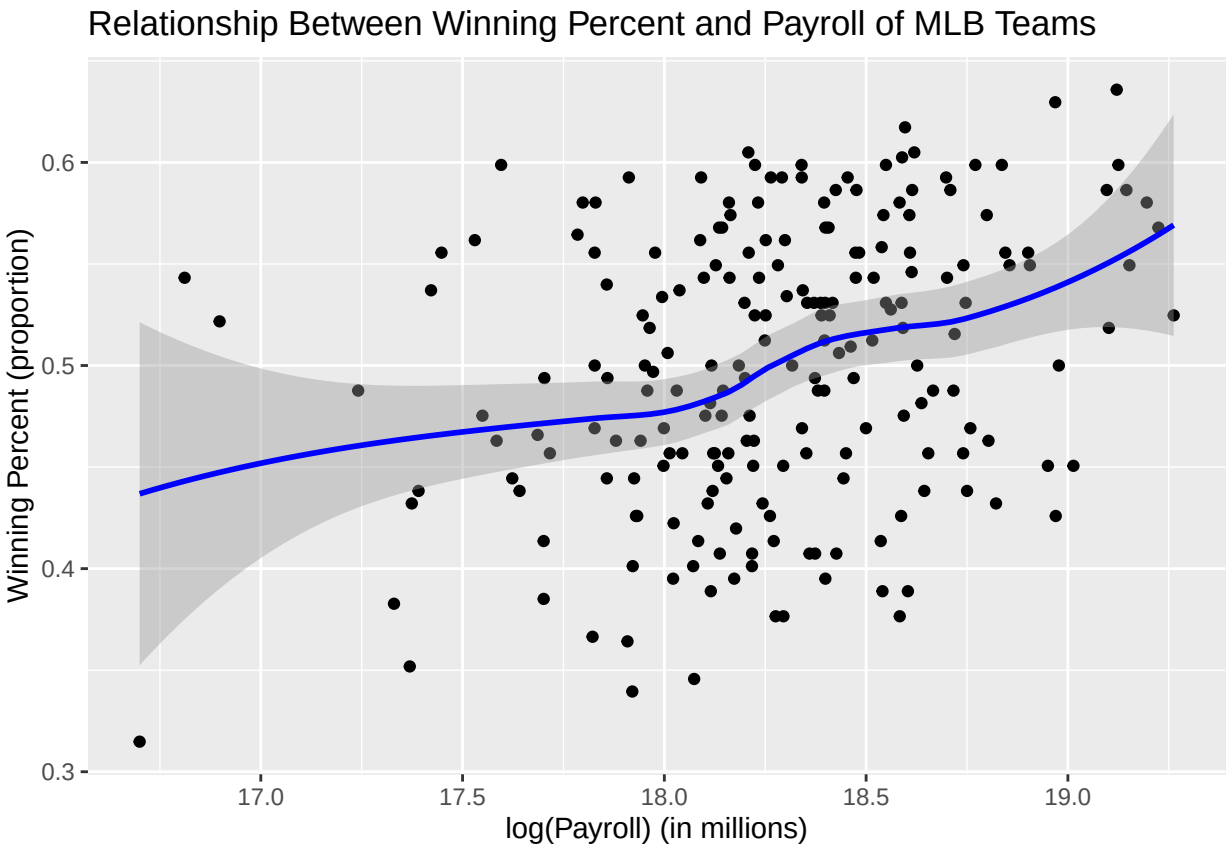
MDSR Exercise 3.6 (modified)

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:

```
library(mdsr)
mlb_teams <- MLB_teams %>%
  mutate(
    log_payroll = log(payroll)
  ) %>%
  select(name, log_payroll, WPct)


ggplot(data = mlb_teams, aes(x= log_payroll, y= WPct)) +
  geom_point() +
  labs(
    title = "Relationship Between Winning Percent and Payroll of MLB Teams",
      y = "Winning Percent (proportion)",
      x = "log(Payroll) (in millions)"
    ) +
  geom_smooth(color = "blue")
```



The graphic I created demonstrates that there is a weak, positive correlation between the payroll and winning percentage in the MLB. That is, a team with a higher payroll in the MLB, will generally have a higher winning percentage, ie, the more games they won/total games (per season).

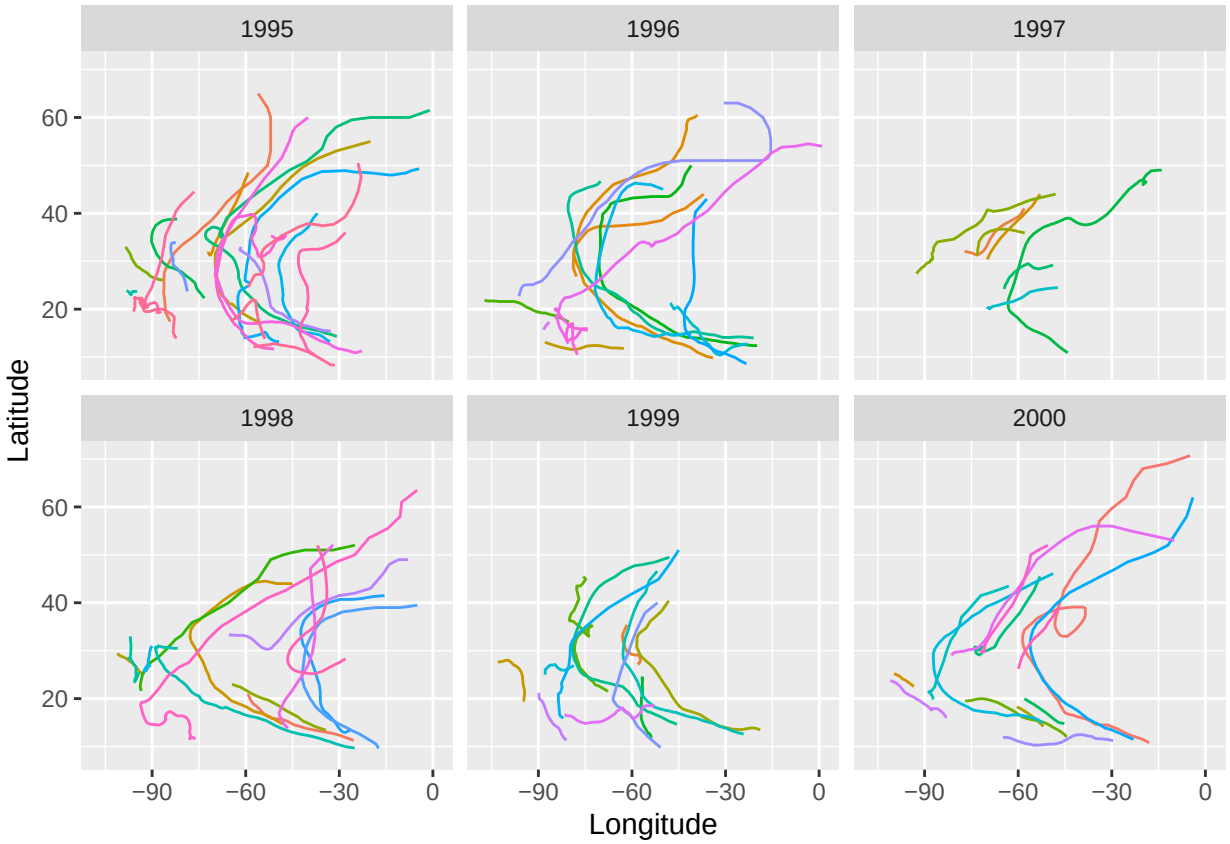
MDSR Exercise 3.10 (modified)

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)

nasa_storm <- nasaweather::storms
ggplot(data = nasa_storm) +
  geom_path(aes(x = long, y = lat, color = name)) +
  facet_wrap(~year) +
  scale_color_discrete(guide="none") +
  labs(
    x = "Longitude",
    y = "Latitude"
  )
```



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 want to focus on are how much time I spend listening to music and then compare that to how much time I spend in class. I also want to track how much time I actively work out and see how that compares to class and music.
- I think the easiest way to depict this data would be in a bar graph with x-axis as "activity" and y axis as total time over 14 days in hours most likely. The different activities would be in different colors to provide distinction. Another visualization could be comparing the time I spend doing an activity on different days. In this case, I would have to average the hours–this might be more effective if done in minutes–per day over the 2-week period and then on the x-axis I would have day of the week and on the y-axis is average time. Then plot and connect points for each activity with activities in different colors for distinction.
- The table for the bar graph would have rows that are each activity and columns labeled as dates for the days of data collection. Then the inputs into the table will be time spent doing said activity, either in minutes or hours.