STAT 231: Problem Set 1B

Alastair Poole

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: Brandon Kwon

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 main takeaway that I see is that Williams College has a diverse network of alums, and each major at Williams College has lead some alums into each sector of today's diverse job market. Essentially, the message is that no matter which major you pursue at college, you'll be able to attain a career in any number of different areas, though some majors lean more heavily to certain career paths than others.

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 can partially be explained by the taxanomy presented in this chapter. The visual cues are as follows: Firstly, color is used, with green representing STEM, orange representing the "soft sciences," and blue representing the pure humanities majors. Size/area is also used, as thicker lines indicate more majors gravitated towards a particular career path. Position is also used, simply because all the majors are grouped onto the left side of the data graphic, and all career paths are grouped together on the right side. This ensures all lines are moving roughly "east-west" across the circle. This data graphic does not contain any coordiante system, nor would it make sense to do so. Scale is only used because each section of the data graphic is categorical.

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.

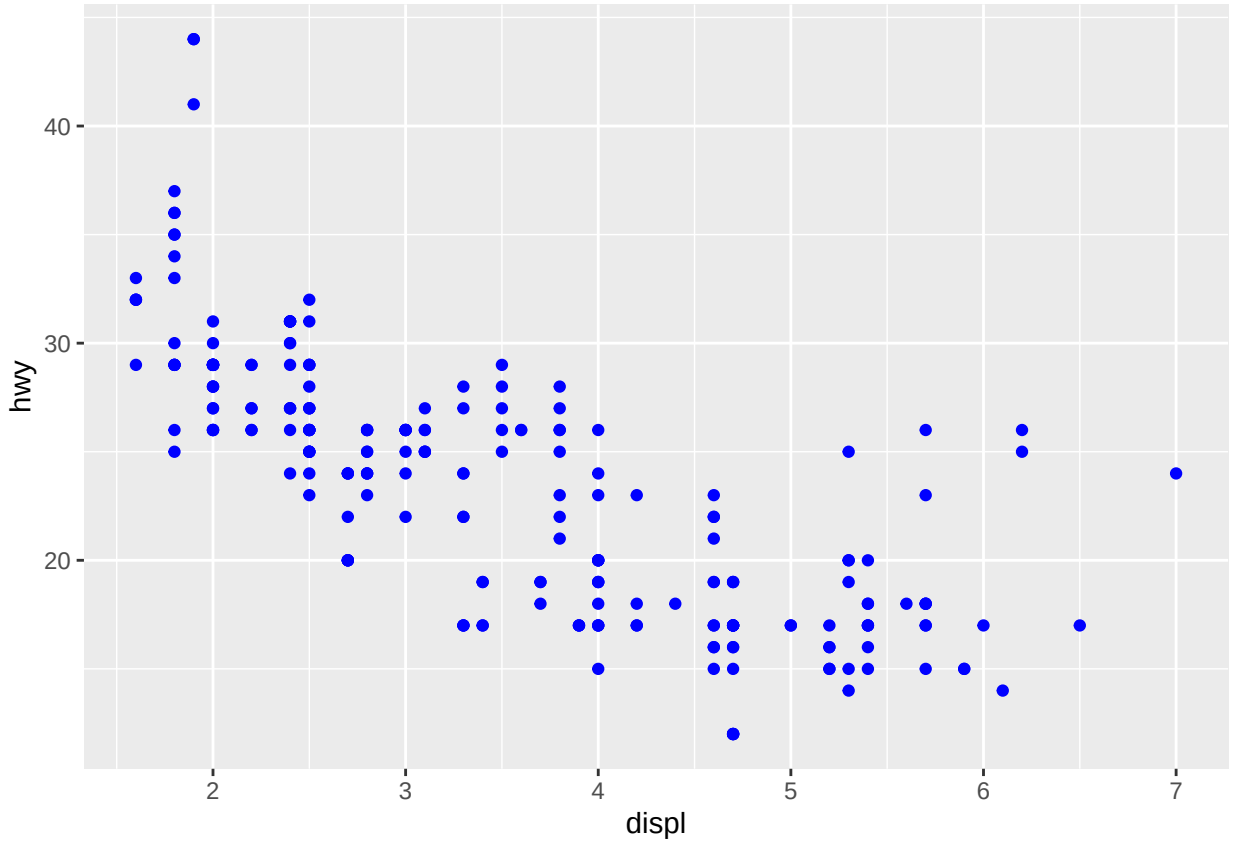
ANSWER: This data graphic is effective in relaying the message that no matter the major, a student can pursue any career path. However, seeing as this data graphic is comprised of 15,600 data points, the thickness of the lines makes it very difficult to distinguish exactly how many alums from a particular major went on to pursue a particular career. I would edit this graphic, perhaps making the number of alums from one major a categorical variable, and then creating a key where there would be clearly labeled line thicknesses representing different relative numbers of majors going into careers. I think the decision to use color to separate the major categories was helpful, and makes the data graphic more visually appealing.

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 when 'colot = "blue"' is inside the aes function, the function interprets "color" as being a third variable in addition to x and y. That is why in the original code, a key is created that lists each color as blue instead of actually coloring in the data points blue.

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

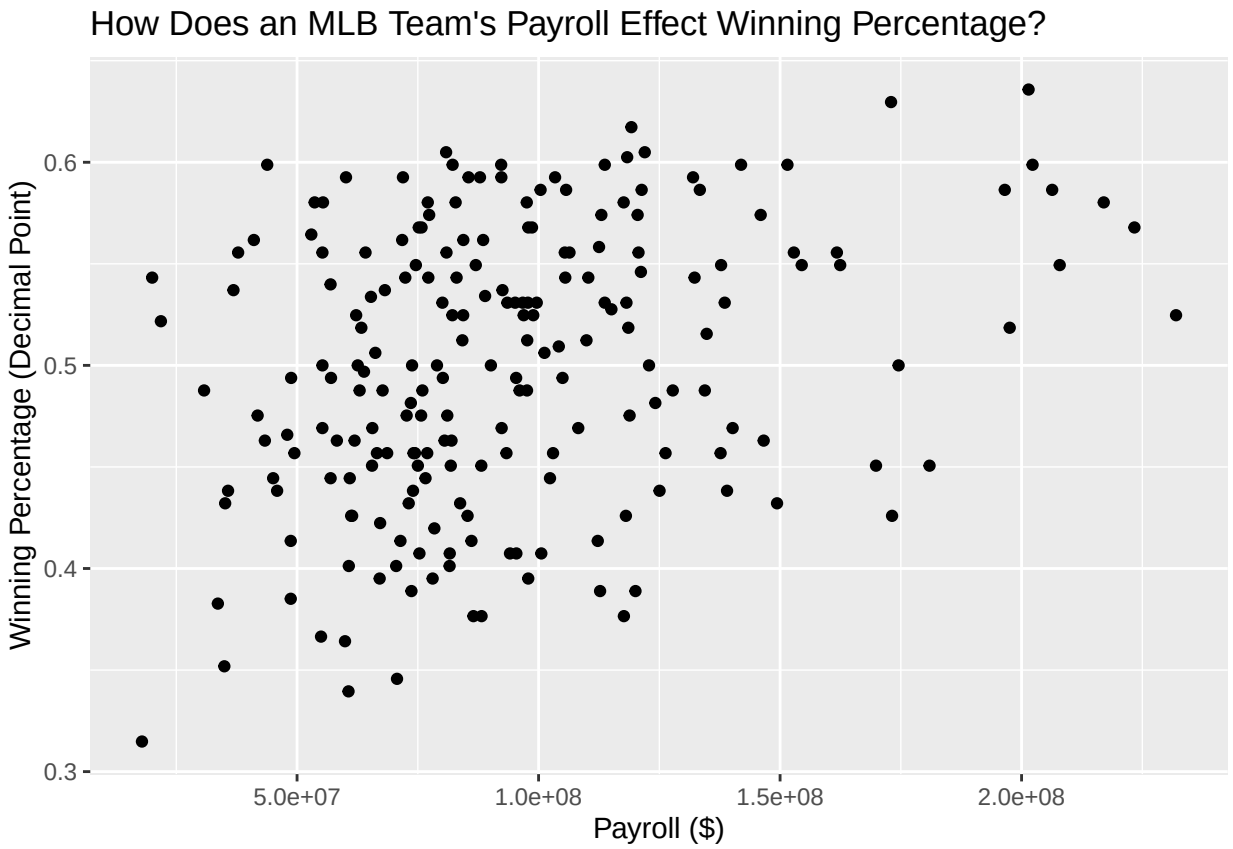


MDSR Exercise 3.6 (modified)

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 general story is that teams that roster more expensive players end up winning more games. This makes sense because the best players in the MLB will demand the most money, and these best players are the ones responsible for earning wins. This graphic illustrates that there is a positive association between payroll and winning percentage, as the payroll increases, so does the winning percentage. This is particularly true for the most expensive MLB rosters, as all seven rosters with a payroll above $200,000,000 have won over 50% of their games, whereas winning percentage varies much more on MLB teams that have payrolls less than $100,000,000.

```
ggplot(data = MLB_teams) +
  geom_point(mapping = aes(x = payroll, y = WPct)) +
  xlab("Payroll ($)") +
  ylab("Winning Percentage (Decimal Point)") +
  ggtitle("How Does an MLB Team's Payroll Effect Winning Percentage?")
```

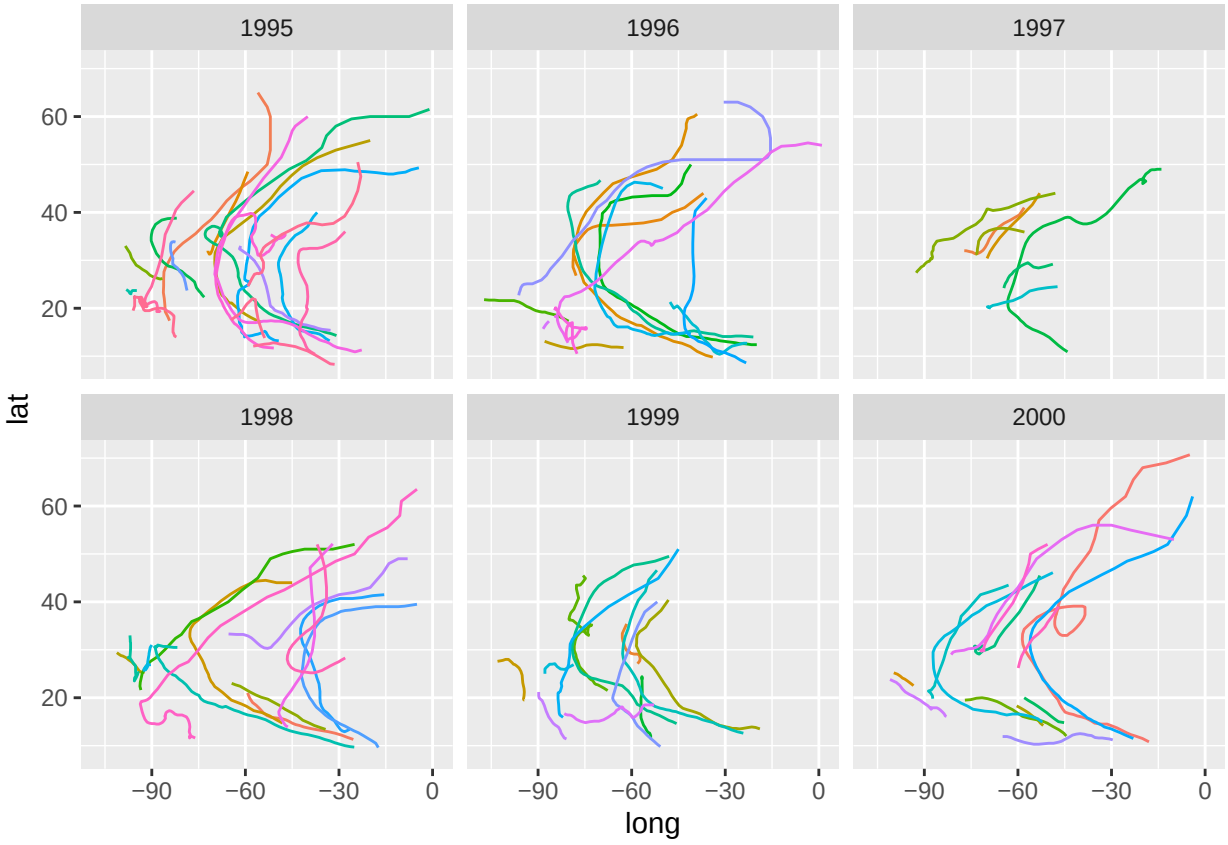


MDSR Exercise 3.10 (modified)

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)
ggplot(data = storms) +
  geom_path(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: The questions I plan on focusing on is how much time do I spend being productive (working on homework, studying for exams, searching and applying for jobs) vs. how much time do I spend being unproductive (lying around, watching TV, playing on my phone)? This is especially important because as a remote student, I do not have to worry about walking to and from class, I simply open up my computer. This leads me to my next question, which I am curious both because of my remote environment and because of Covid regulations in MA right now: How much time do I spend outside my bedroom (working, at the gym, eating) vs. how much time do I spend in my bedroom (classes, homework, sleeping)? As for visualizations to help me answer these question, I was thinking about creating two bar charts (one for productivity and one for non-productivity; one for outside my bedroom and one for inside my bedroom), where each chart has number of hours spent on the y axis, and each axis includes categorical variables which would be the specific activites I engage in. One data table will include the rows being number of hours planned and number of hours spent, while the columns may be the various activities I engage in as well as the locations I go to. This is a rough idea so I plan on discussing with either or both of Andrea and Professor Correia.