STAT 231: Problem Set 1B

Brandon Kwon

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: Alastair Poole

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 displays the relationship between major and possible career paths through the utilization of arcs that connect parts of the circle to other parts of the circle. As a result, I believe the story the data graphic is trying to convey is one that emphasizes the fact that the type of major one pursues in college is very much related to the career pathway that he/she chooses in the future, but that there is no one main path that a student can take in order to pursue a specific career. One can major in practically anything and still arrive at the same destination as another student who majors in something completely different.

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 can partially be described in terms of the taxonomy presented in this chapter. In terms of visual cues, I see that position, length, direction, and color contribute to this representation greatly. Position is significant in that each part (whether major or career) is placed in a piece of the circle in a way that allows the viewer to make connections easily. Length is also important because each arc has different lengths as to allow the viewer to make better comparisons in terms connecting a major to a specific career path. Direction plays a role not necessarily as an indicator of "slope" but rather as an indicator of which part connects to what as a part of the circle. (Does the arc rise upward? Does the arc aim downward to a career path?) I do understand that direction, however, does not contribute as much to the graphic as it would for numerical variables. Lastly, the color is a very significant indicator of differentiating each major besides the labels themselves. A coordinate system is not utilized in this data graphic due to it being a circle with arcs with no designated numerical positions. Scale is utilized in the sense that each variable represented in this data graphic is categorical. Type of major is categorical, as well as type of career. There is no sense of ordering to these variables.

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.

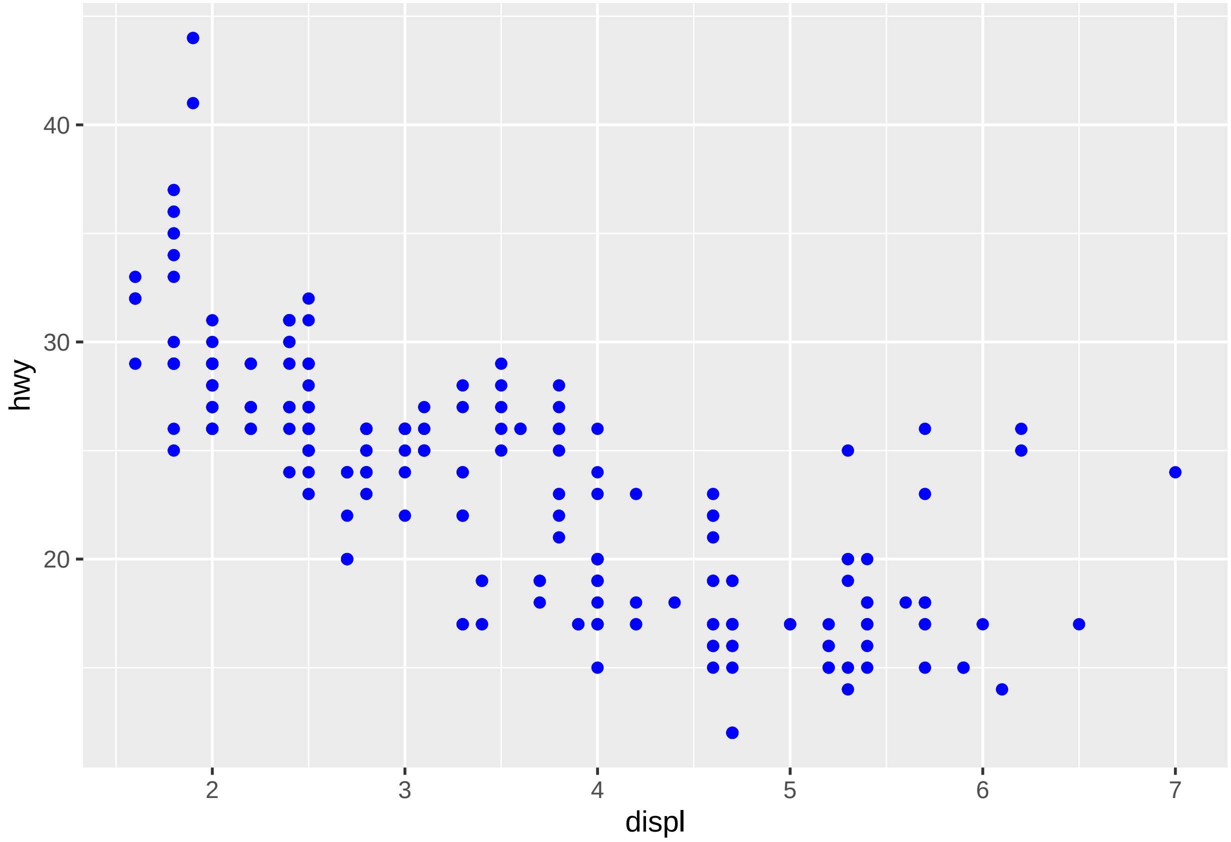
   ANSWER: I believe the designer did an excellent job conveying the relationship between major type and career type. The length/direction of the arcs as well as the color of each major allowed me to make distinct connections. The fact that I can view each career path distinctly, while also being able to see the compilation of arcs, is helpful in terms of what my purpose would be: if I wanted to view a specific career pathway and its connection with different majors, I can utilize my ability to view that career only. If wanted a broad picture of how majors relate to career pathways in general, I have the option to view the compilation of arcs in its entirety. One aspect I find puzzling, however, is that for some arcs, they "thin" out towards the middle. I do not understand why the designer wanted to convey this feature, but I am interested in learning why this feature was displayed in the graphic. If I were to change one characteristic of this graph, I would also consider utilizing different shades in allowing each career pathway to be represented more distinctly in eyes of the viewer. I would also provide numerical values to the thicknesses of each arc more explicitly to allow for better reference.

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 the "color" aspect is inside the aes() function, thus implying that color is connected to the variables themselves. However, if the "color" aspect is outside the aes() function, this aspect is applied to the dataset in its entirety, thus enabling all of the data points to become blue.

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

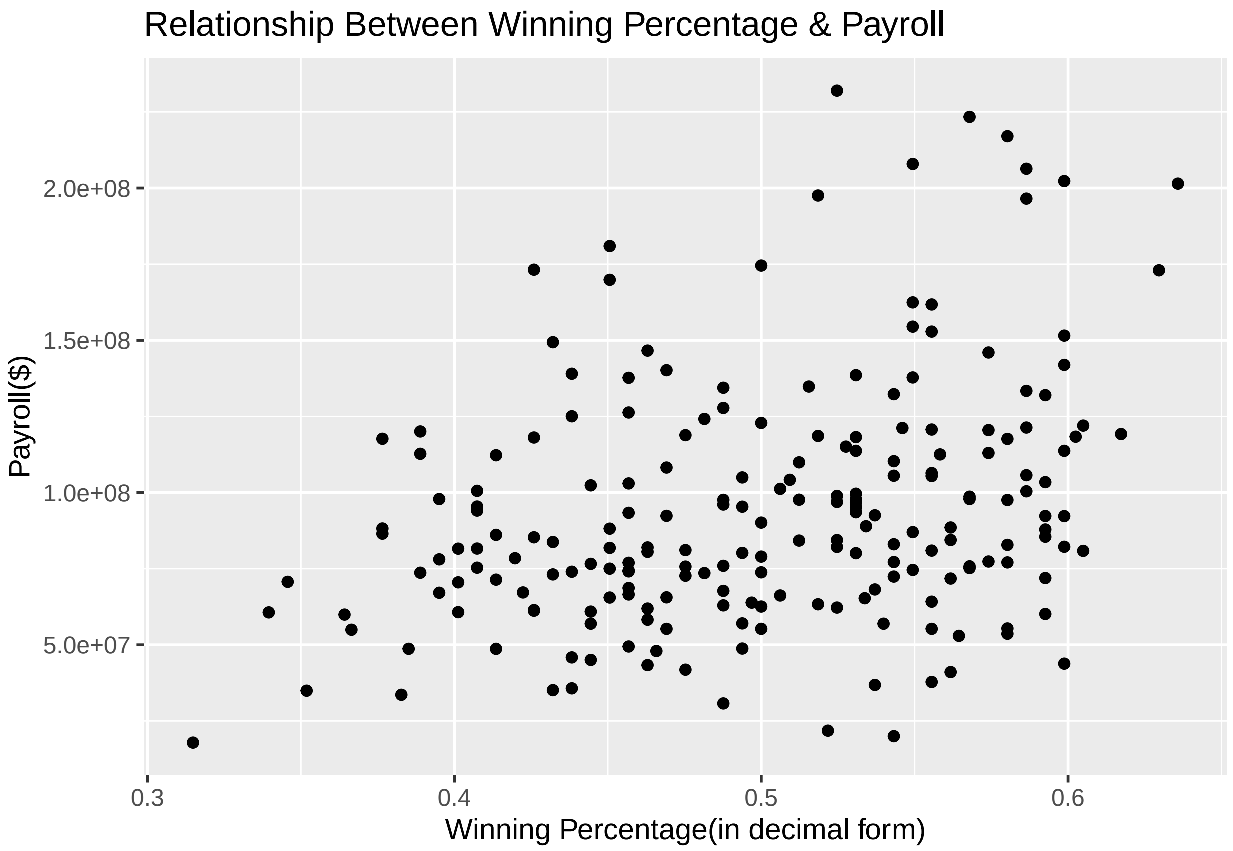


MDSR Exercise 3.6 (modified)

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: We see that there seems to be a positive correlation between payroll and winning percentage in the MLB. This correlation is not strong, but we can somewhat conclude that as the winning percentage improves for a team, the payroll also increases for the players. This definitely makes sense because the story the graph tells is that winning games and thus having a better winning percentage garners positive results for the team such as more cap space (due to more viewership, sponsorship, and more recognition in general). Thus, this will allow for better salaries for the players logistically.

```
ggplot(data = MLB_teams) +
  geom_point(aes(x = WPct, y = payroll)) +
  xlab("Winning Percentage(in decimal form)") +
  ylab("Payroll($)") +
  ggtitle("Relationship Between Winning Percentage & Payroll")
```

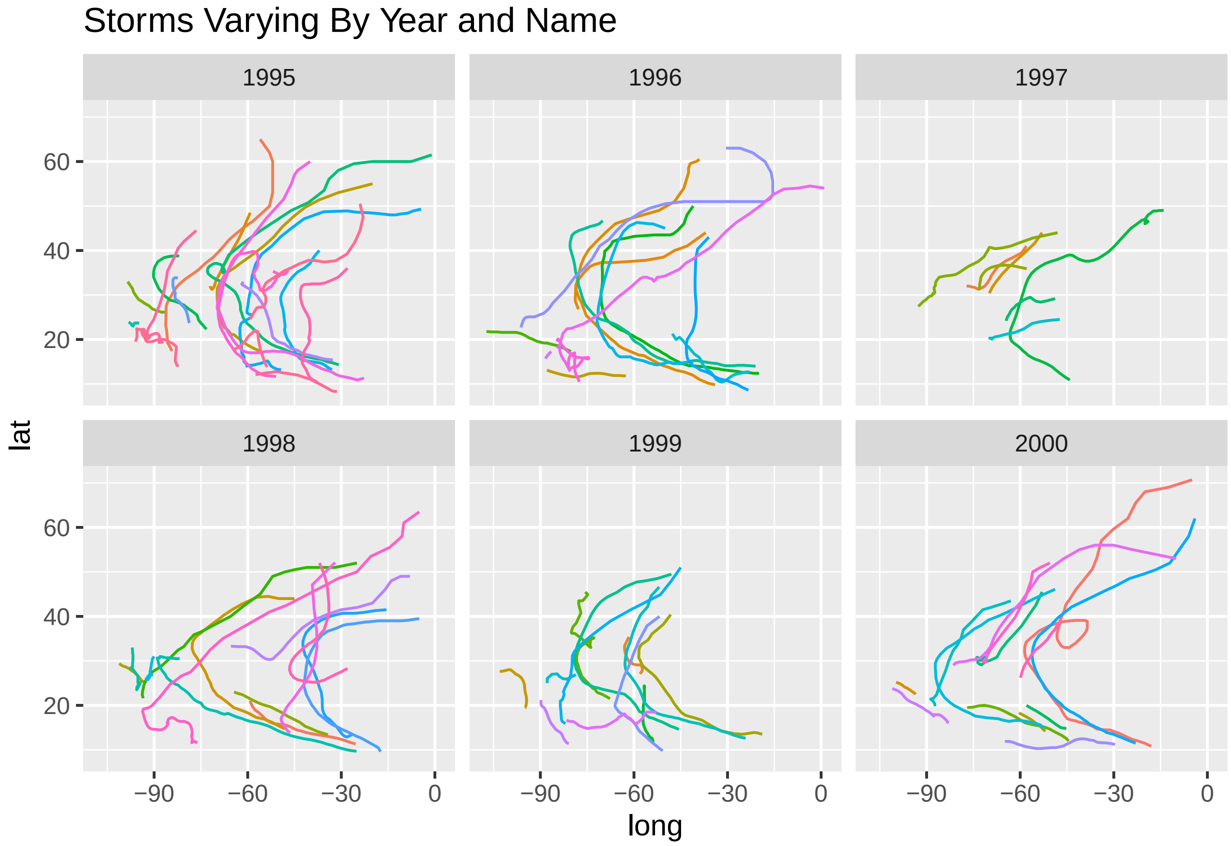


MDSR Exercise 3.10 (modified)

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!

```r
library(nasaweather)
## view(storms)
ggplot(data = storms) +
  geom_path(aes(x = long, y = lat, color = name)) +
  facet_wrap(~year) +
  scale_color_discrete(guide="none") +
  ggtitle("Storms Varying By Year and Name")
```



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: Questions: How does my time focusing on school work compare to my time taking naps/sleeping? What relationship is present among studying for school, doing leisurely activities, and working at the hospital? How much do I get distracted from my school work due to social media, Youtube, and Netflix?
>
> Visualizations: One type of visualization I plan to consider is to use a bar graph that represents the time I spend studying for school, doing leisurely activities, and working at the hospital. On the horizontal axis, I plan to use a time scale that represents each day I will record (Monday-Sunday, Monday-Sunday). Each bar that represents each day will then be split (by color) by the category that I listed above. The splits will be determined by percentage of time that I spend on each activity per day as compared to the other activities that I carry out (with the exception of sleeping). Another type of visualization I plan to consider is to create a scatterplot that conveys the relationship between the time I spend on social media, Youtube, and Netflix and the time I spend studying/working on homework. Each point that I plot will represent a day of the week. I plan to use the facetting function (by time on social media, Youtube, and Netflix) to create better distinction on the relationship between the two variables.
>
> Table: The rows I plan to use can be labelled by day. The columns would then be separated by time (in hours and minutes) spent studying for school, time spent doing leisurely activities, and time working at the hospital.