STAT 231: Problem Set 1B

Joshua Kim

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:

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 graph shows the relationship between William's College student's choice of major and the fields of their careers. We see that every field/industry attracts talent from each field of majors. The main message that I take away from this graph is that although we may believe that our major has a significant impact on our careers, we can see that despite their choice of major, there were still many people who chose to go into different industries. Thus, as a senior, this encourages me to broaden my job search range, and try different industries. However, we do see some strong correlation – for instance, many political science majors go into law, while many biology majors enter the health/medicine industry.

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 graph is a polar graph, using a cartesian system. The scale is categorical, being split in half of the circle between majors and field of career. The visual cues in this graph are the width of the arc representing the strength of the relationship of major and career field, length of the major/field representing the share of the major/field, and color representing different majors.

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.

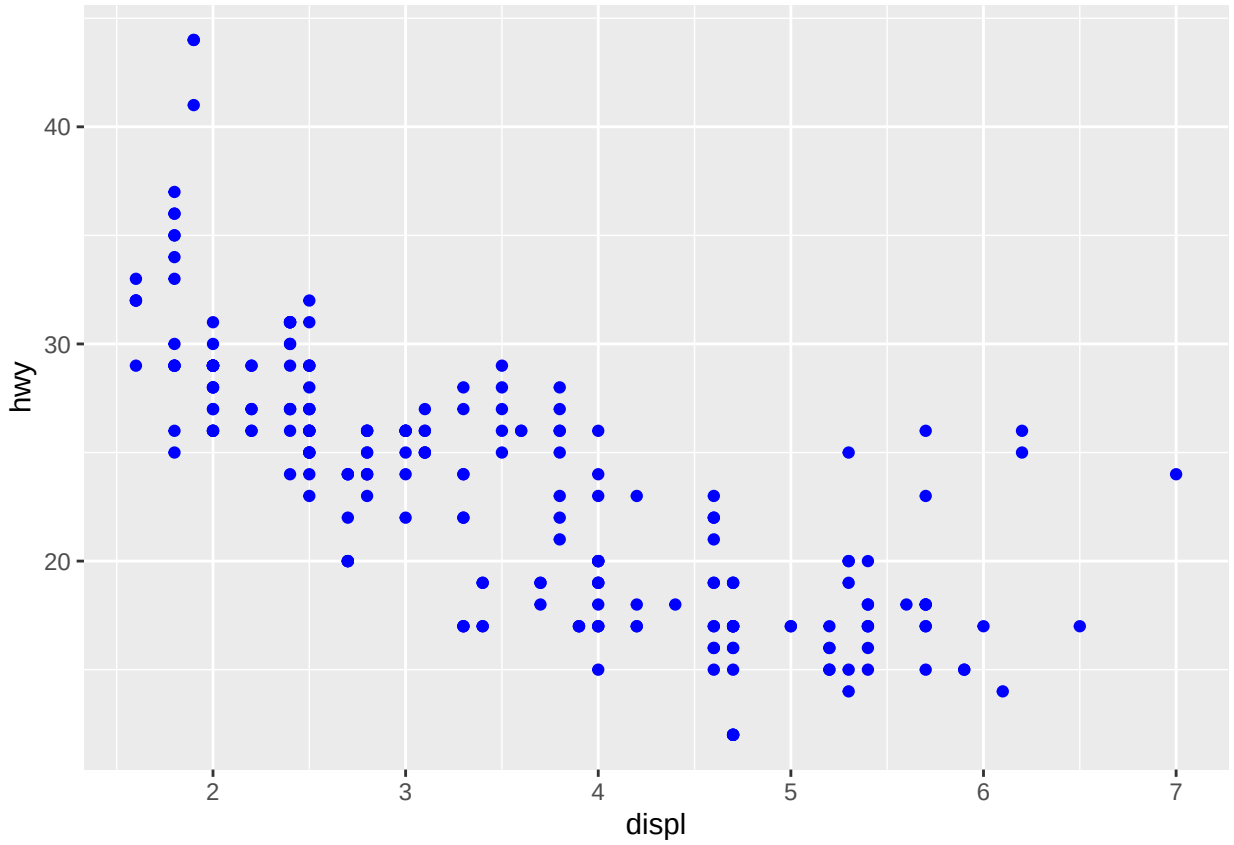
   ANSWER:I really enjoy the graph as I think it tells the main point of that all fields recruit from different majors. Although it would be nice to know what percentage each major represents, I think that this would be too busy for the graph and the message it is trying to tell. I really think that this graph is brilliant and even aesthetically pleasing because of the good contrasts of blue, red, and green. It is a busy graph but I also think it portrays the complex nature of jobs and how there is a chance for everyone. I found this graph very encouraging and even a bit motivational so that I could find jobs in different fields even in a tight job market due to the pandemic.

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 function color ="blue" was within the aesthetics, which does not refer to the geompoint function. Thus by changing it outside of the aesthtics command, we are able to change the points into blue.

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



MDSR Exercise 3.6 (modified)

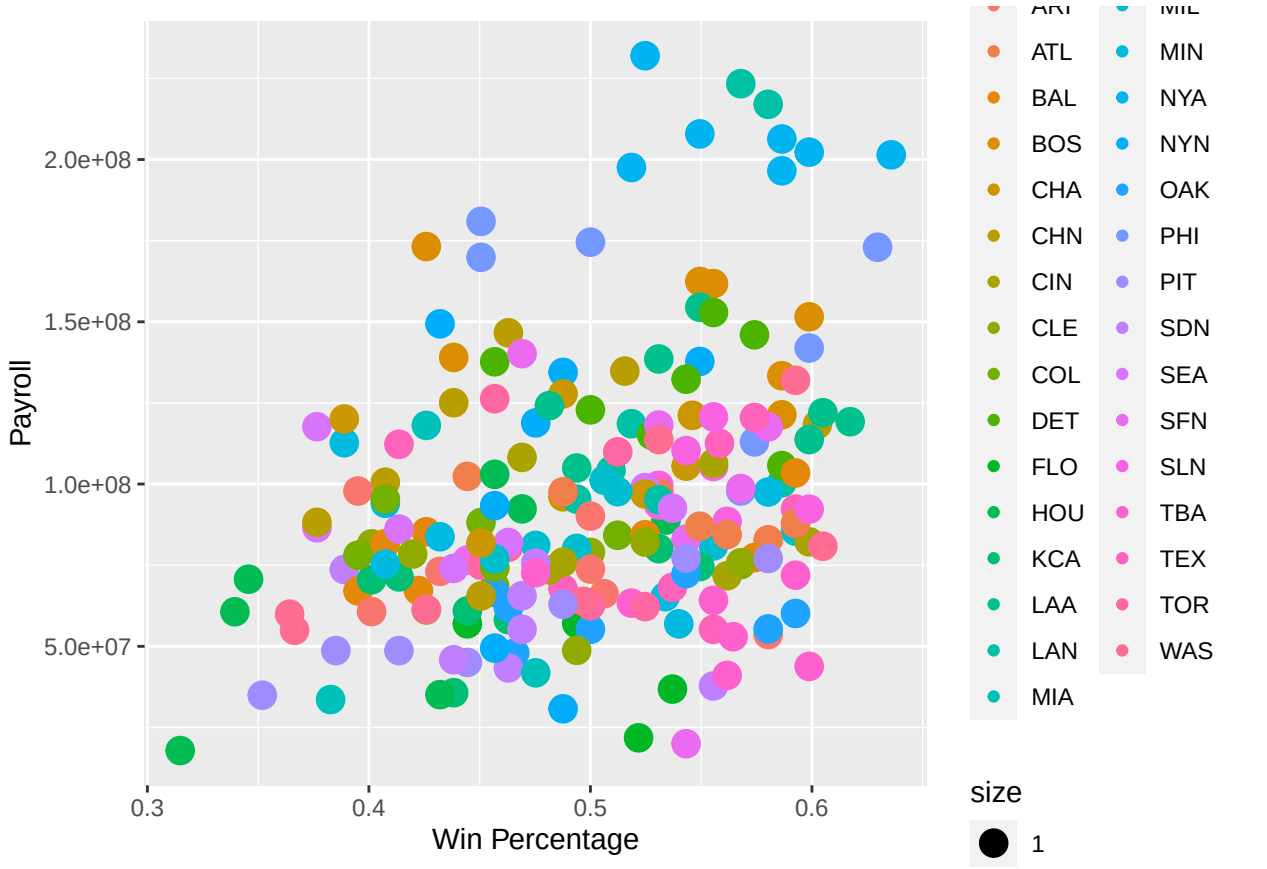
Use the `MLB_teams` data in the `mdsr` package to create an informative data graphic that illustrates the relationship between winning percentage and payroll in context. What story does your graph tell?

> ANSWER: This graph shows the relationship between payroll and win percentage with each point referring to a unique year and team. We can see that there is a positive relationship between payroll and win percentage, but we also observe that there are specific teams that do even better with more payroll.

```
library(mdsr)
library(ggplot2)
library(tidyverse)
library(dplyr)

MLB_teamsWPPR <- ggplot(MLB_teams, aes(x = WPct, y = payroll)) +
  xlab("Win Percentage") +
  ylab("Payroll") +
  geom_point (aes(size=1, color = teamID))

MLB_teamsWPPR
```



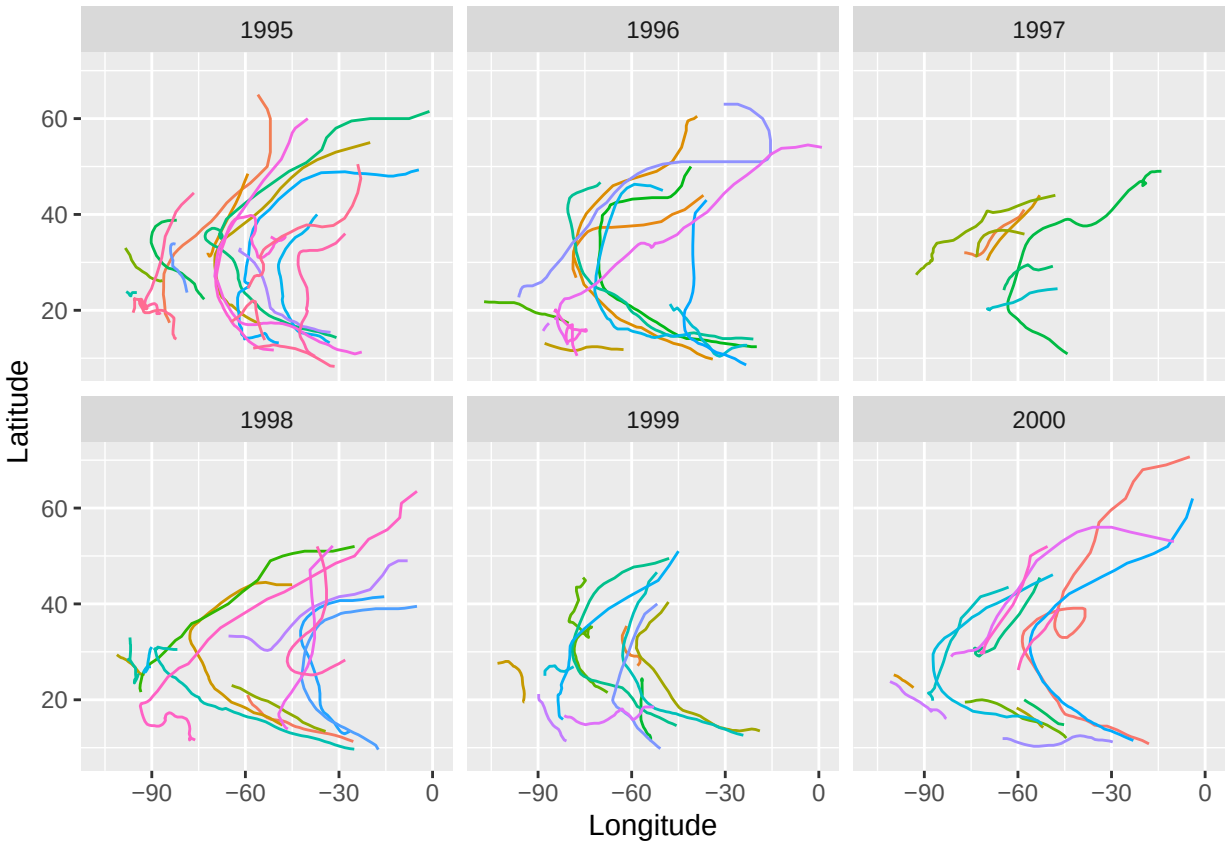
MDSR Exercise 3.10 (modified)

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(nasaweather::storms, aes(x = long, y = lat)) +
  geom_path(aes(x = long, y = lat, colour = name)) +
  facet_wrap(~year) +
  scale_color_discrete(guide="none") +
  xlab("Longitude") +
  ylab("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 answer 'How do I spend my time on my business vs work vs school vs leisure on each day?' I would also like to further segment that data, and learn how I am spending my time doing different types of work I do for my E-commerce store so that I can better learn to budget my time and maximize effeciency of my time spent. I believe that I can create a segmented bar graph faceted by day on a cartesian plane. This way I can visulize how I spent each day of the week. Visual cues can be color for category of time spent(business vs work vs school vs leisure), length of the bar will indicate time spent. Meanwhile a second visualization could be a line graph on a cartesian plane. With time spent on the y axis and days on the x axis, I can visualize how I spent my time on my business each day without needing to facet. Visual cues include color for each category and angle for direction of time spent. One table idea that I have is to have the columns be business vs work vs school vs leisure while the rows to be individual days. This will help me understand the exact time that I spent each day.