STAT 231: Problem Set 1B

Kevin Ma

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: Office Hours On Friday

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 story the data graphic tells is about the career choices of different majors that graduated from Williams College. The main message that I take away from it is that some majors concentrate their alumni to specific industries while others are more broad. For example, computer science majors are concentrated to the technology sector while psychology majors are more evenly divided.

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: While there is no identifiable coordinate system or scale, there are various visuals cues utilized by the graphic. The graphic first uses length to demonstrate how many students belong to each group of majors. The graphic also uses color to identify the different major groups. In addition, to demonstrate the "flow" of students to their respective career choice, the graphic uses area and shade the highlight how many students of a particular group of majors goes to a certain industry. Area would encompass the different thickness of each line. A feature that lies outside of the taxonomy is that double-majors have their lines split in two, which does not really fall in any one item (possible area?)

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.

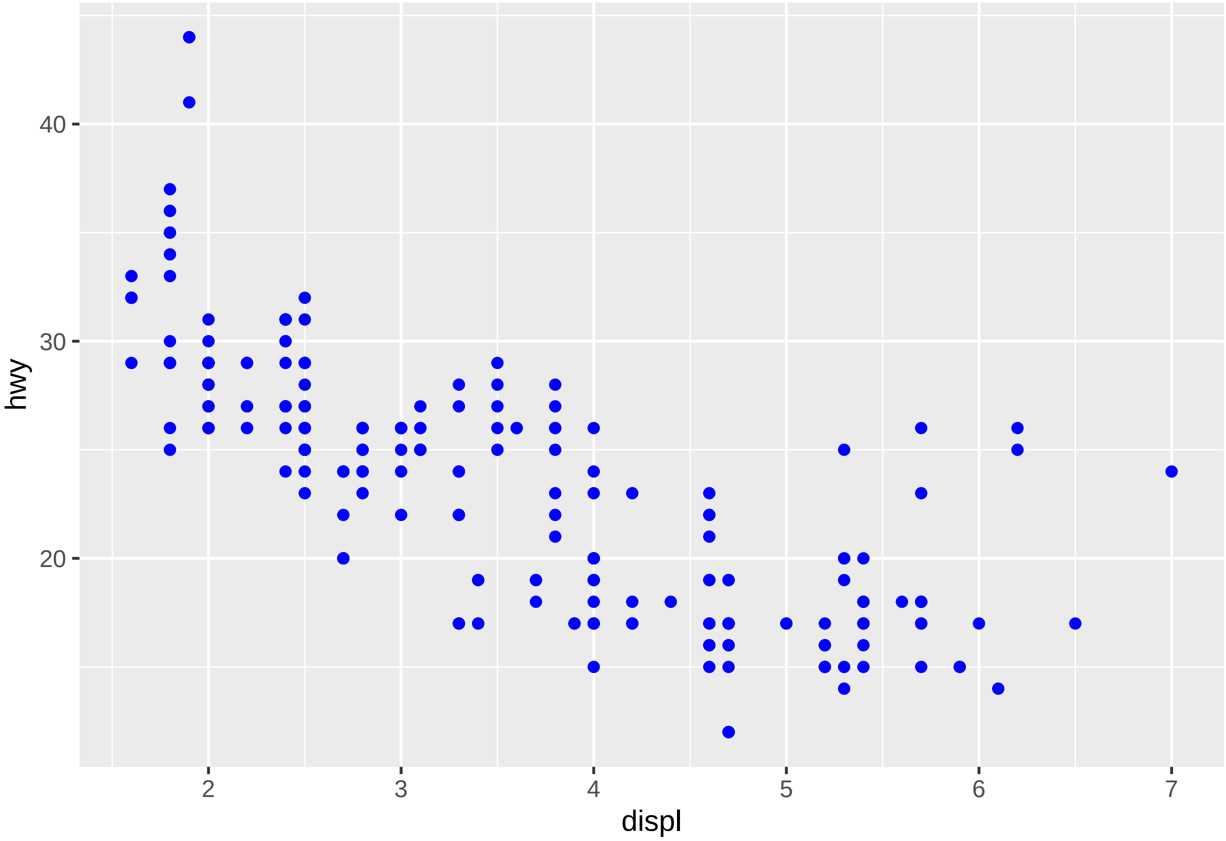
   ANSWER: I think that the visualization choice made by the designer is very clever because there are a lot of different majors and a lot of different industries. Thus, the wheel approach makes the messaging more intuitive and interpretable. A bar chart would be too complicated for the major group "English" as the alumni are spread across too many industries. However, it is hard to tell in this wheel format how many alumni belong to each major group and industry, so I think adding some additional labelling would have made the graphic more clear.

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 command does not turn the data points blue the code “color =”blue"" is within the aes function which is part of mapping. As such, it is treated as an aesthetic, which means “blue” is treated as a variable rather than as an instruction. Because there is no variable “blue”, all of the points are the default color. The command that will turn the points blue is: ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy), color = “blue”)

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

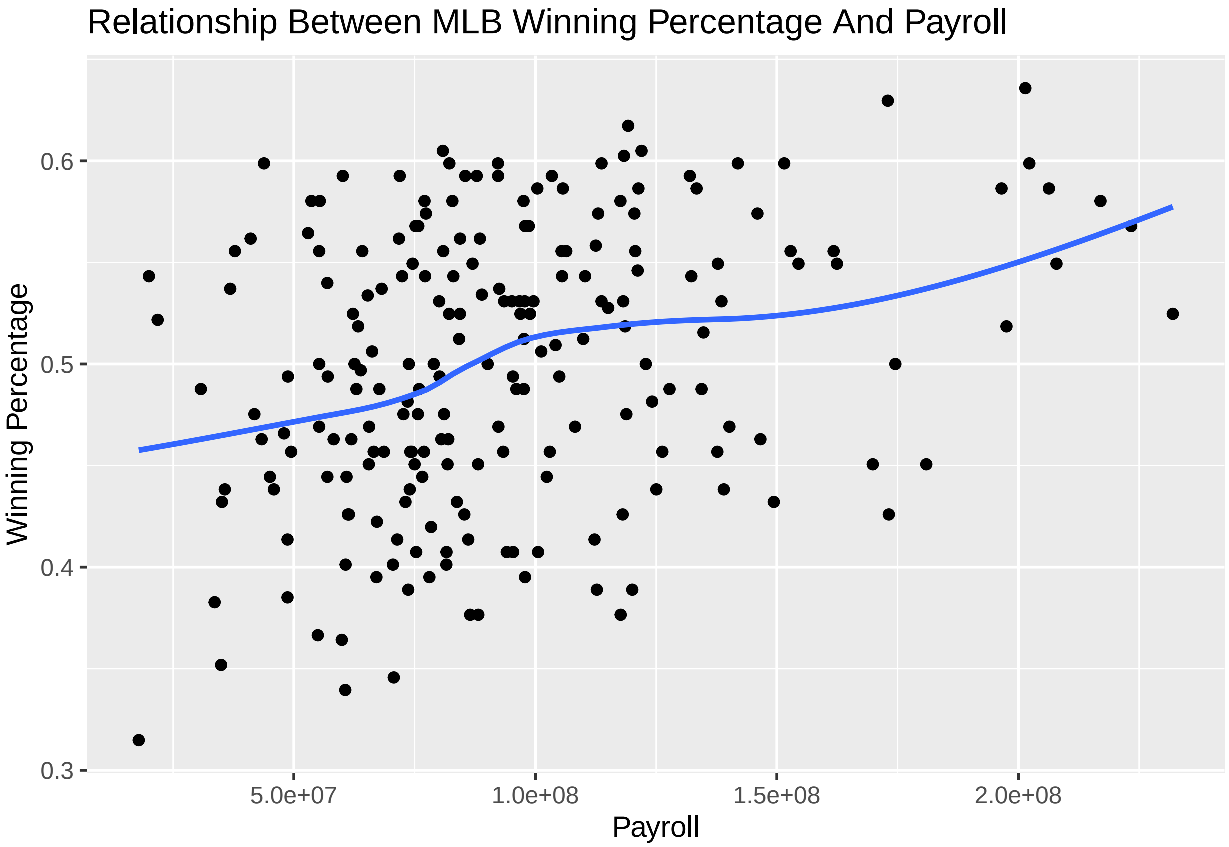


MDSR Exercise 3.6 (modified)

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: My graph shows that from 2008 to 2014 that there was a positive correlation between a MLB team's payroll and their winning percentage during the season. What I plotted is the relationship between the dependent variable Winning Percentage and the independent variable Team Payroll.

```
library(ggplot2)
ggplot(data = MLB_teams, mapping = aes(x = payroll, y = WPct)) +
  geom_point()+
  geom_smooth(method = "loess", se = FALSE) +
  labs(
    title = "Relationship Between MLB Winning Percentage And Payroll",
    y = "Winning Percentage",
    x = "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")`.

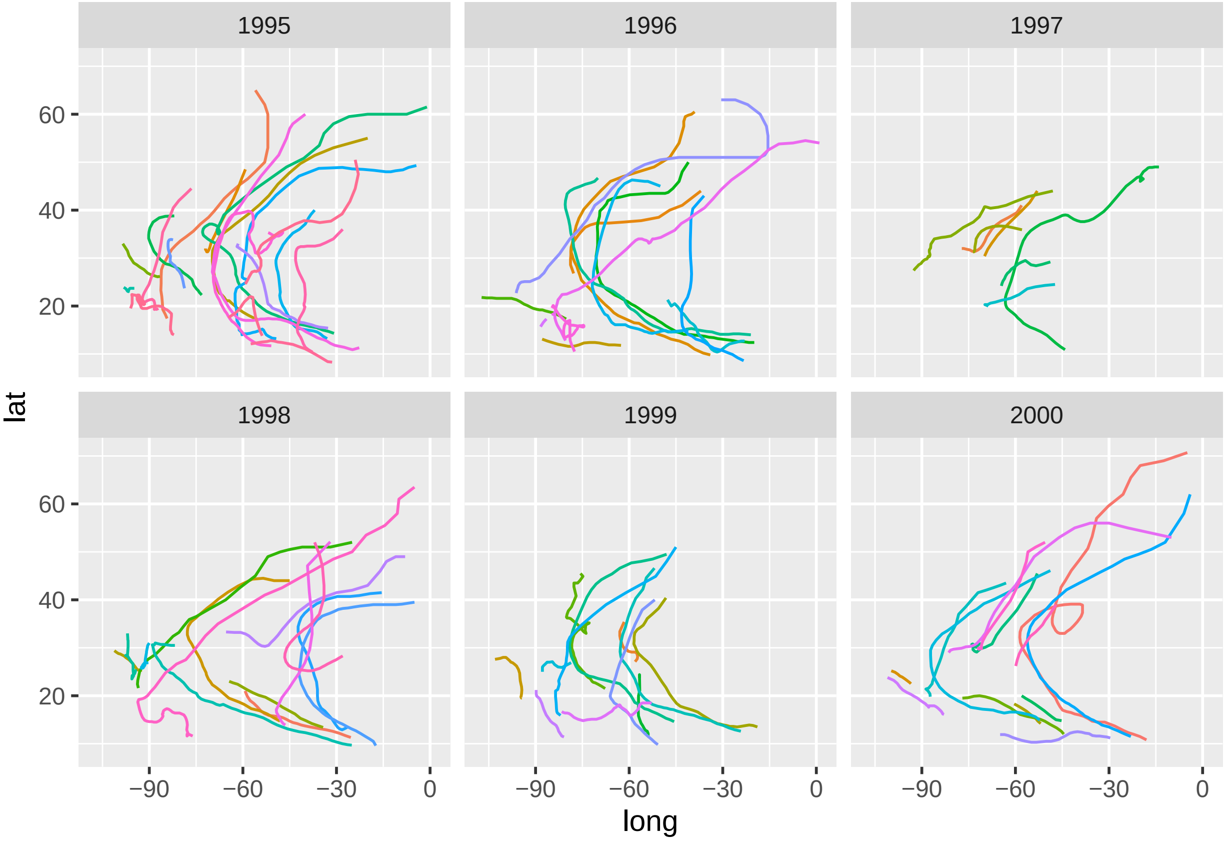
Note: be sure you load the `nasaweather` package and use the `storms` dataset from that package!

```
install.packages("nasaweather", repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/72/g78dd9gd7jz4cs0ty13xcw580000gn/T//RtmpjH0OVj/downloaded_packages
```

```
library(nasaweather)
```

```
library(ggplot2)
ggplot(data = storms) +
  geom_path(data = storms, aes(x = long, y = lat, color = name))+
  facet_wrap(~year, ncol = 3)+
  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: I have two main questions that I am planning to focus on. The first is how much time do I spend on Amherst class-related activities? This includes classes, homework, thesis work, and preparation for exams. The second is how much time I spend on a particular class relative to other classes? The first visualization can be a line graph where I chart the amount of time I spend on class work per day over the course of two weeks. On the y axis will be the amount of time spent on class work while the x-axis will be dates in chronological order. The second vizualization will be an animated pie chart showing the fraction of time spent on each class over a given period of time. The pie chart will use color to show different classes. If I track data for two weeks, I think it will be interesting to have the pie chart reflect class time spent on each section on the average of a given day. For example, the pie chart will first show how much time I spend on each class on Sunday, then Monday and so on for the rest of the week. The data table I will create will have to contain various sources of data. I image in that each row will be a date. For each row/observation there will be seven columns. The first three columns will store the amount of time spent during on that class during the day. The next three columns will store the fraction of overall class time spent that a single class had during that day. The last column will have the total amount of time spent on class work during the day, so it will be the sum of the first three columns.