STAT 231: Problem Set 1B

Evan Daisy

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: Any major at Williams college could lead someone into any number of different fields.

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 creators of this graph use position on the left side of a polar coordinate system to indicate an alum's major and position on the right side to indicate the field in which they work. (Arc) Length is used to indicate the number of alums who majored in a particular field or work in a particular profession. Color is used to indicate major, and the creators employ layering/interactivity to allow the user to see the trajectories of every individual with a particular major or in a particular field. They use faceting to give separate charts depicting the different career paths charted by those who double-majored in a particular field, and they use an animation to show how the career choices of majors has changed over time.

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.

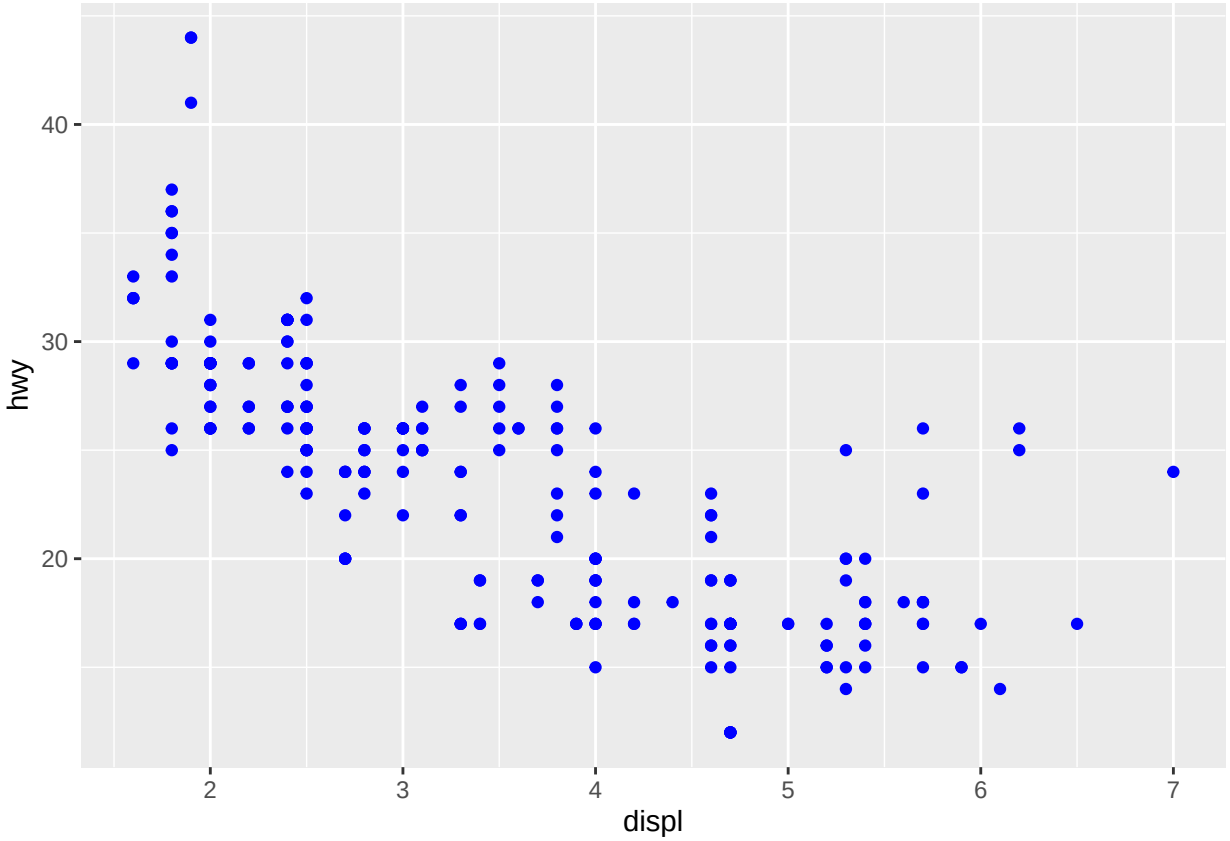
ANSWER: I would say that the first two tabs are great interactive ways in which to show how career choice varies by major over time (in the case of the second tab). The one question I might have is whether these graphics obey the area principle (is the area of major regions are truly proportional to the number of individuals in that major? Is the area of a stream entering a profession from a major proportional to the number of people who took that career path?) It's quite possible that they do, but the designers would need to be careful. The third tab I found a bit more confusing; I had to read the caption a few times to realize what was happening, and I found the title misleading, because the easiest things to see in these graphics are the career paths of single majors, not those of double-majors.

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: This command is claiming that mapping color to "blue" is an aesthetic, but it isn't because there are no variables involved. Color would instead simply be an attribute of the scatterplot and we would leave it outside of the aes() function, see fixed example below.

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

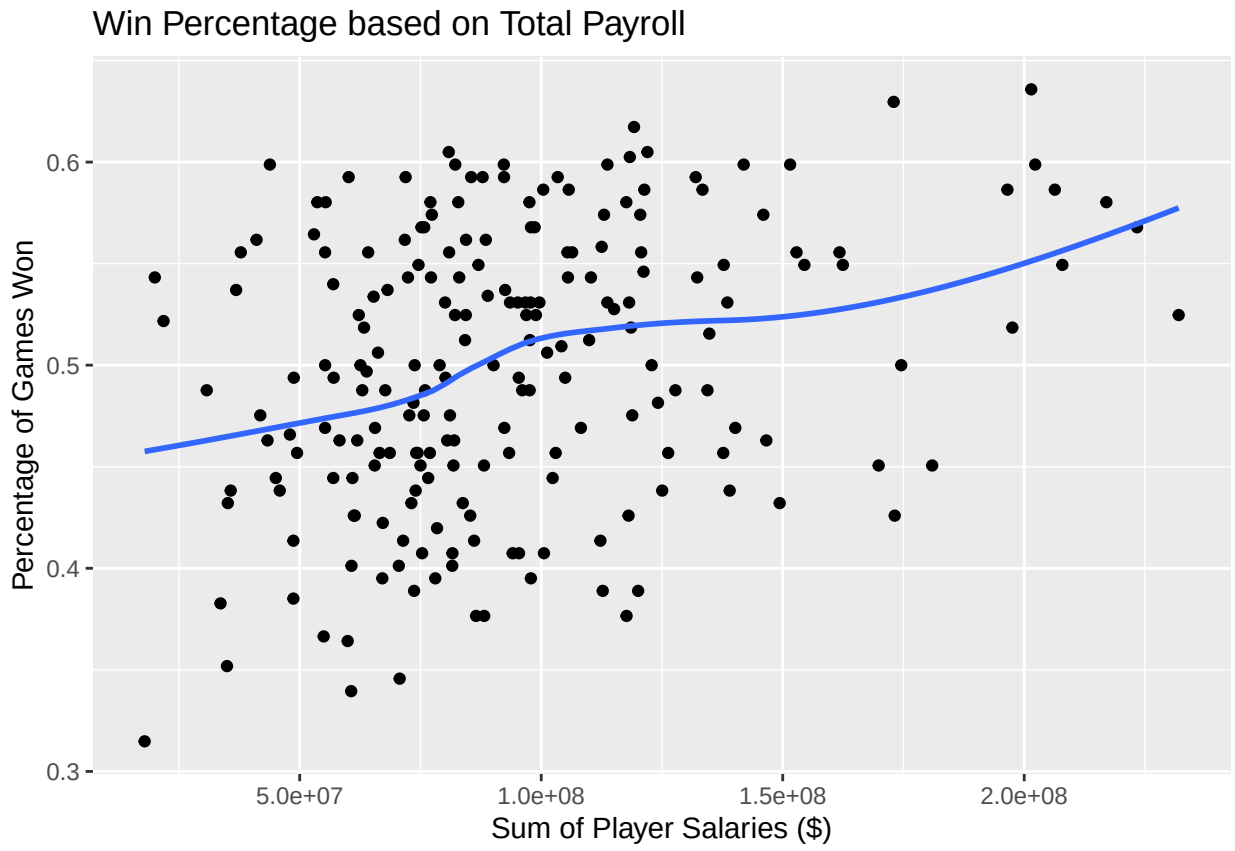


MDSR Exercise 3.6 (modified)

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 tells that there is a weak positive association between win percentage and total payroll of a team in a given year.

```
help(MLB_teams)
ggplot(data = MLB_teams) +
  geom_point(aes(x = payroll, y = WPct)) +
  geom_smooth(aes(x = payroll, y = WPct), method = "loess", se = FALSE) +
  labs(title = "Win Percentage based on Total Payroll") + xlab("Sum of Player Salaries ($)") +
  ylab("Percentage of Games Won")
```




```
MLB_teams %>%
  summarize(r = cor(payroll, WPct))
```

```
## # A tibble: 1 x 1
##       r
##   <dbl>
## 1 0.319
```

MDSR Exercise 3.10 (modified)

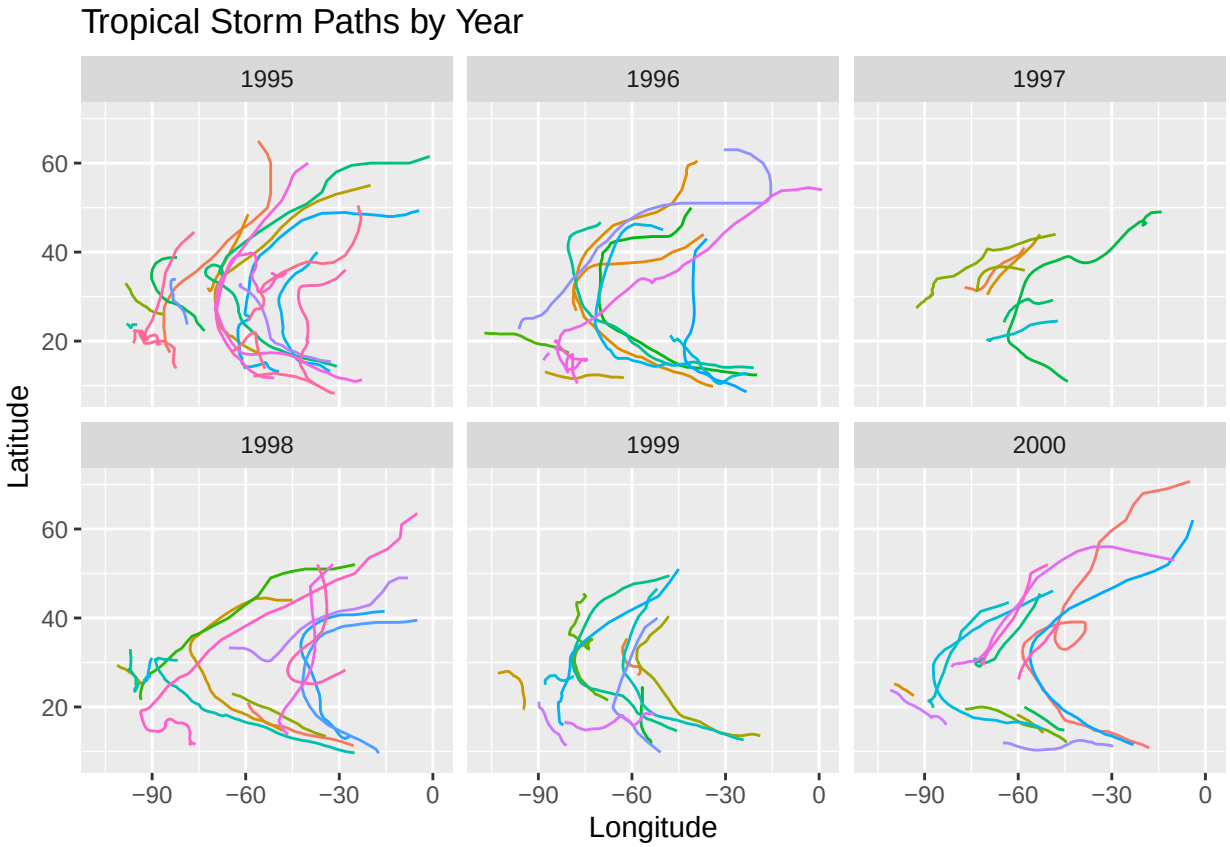
Using data from the `nasaweather` package, use the `geom_path()` function to plot the path of each tropical storm in the `storms` data table (use variables `lat` (y-axis!) and `long` (x-axis!)). Use color to distinguish the storms from one another, and use facetting to plot each `year` in its own panel. 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)
glimpse(storms)
```

```
## Rows: 2,747
## Columns: 11
## $ name     <chr> "Allison", "Allison", "Allison", "Allison", "Allison", "Allis~
## $ year     <int> 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1995, 1~
## $ month    <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6~
## $ day      <int> 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 8~
## $ hour     <int> 0, 6, 12, 18, 0, 6, 12, 18, 0, 6, 12, 18, 0, 6, 12, 18, 0, 6,~
## $ lat      <dbl> 17.4, 18.3, 19.3, 20.6, 22.0, 23.3, 24.7, 26.2, 27.6, 28.5, 2~
## $ long     <dbl> -84.3, -84.9, -85.7, -85.8, -86.0, -86.3, -86.2, -86.2, -86.1~
## $ pressure <int> 1005, 1004, 1003, 1001, 997, 995, 987, 988, 988, 990, 990, 99~
## $ wind     <int> 30, 30, 35, 40, 50, 60, 65, 65, 65, 60, 60, 45, 30, 35, 35, 4~
## $ type     <chr> "Tropical Depression", "Tropical Depression", "Tropical Storm~
## $ seasday  <int> 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 8~
```

```
ggplot(data = storms) + geom_path(aes(x = long, y = lat, color = name)) +
  scale_color_discrete(guide="none") + facet_wrap(~year, nrow = 2) +
  labs(title = "Tropical Storm Paths by Year") + xlab("Longitude") +
  ylab("Latitude")
```


Tropical Storm Paths by Year
1995
1996
1997
1998
1999
2000
Latitude
Longitude
60
40
20
−90
−60
−30
0

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 plan to focus on where I spend most of my time (my room, common room, science center, etc.), with whom I spend most of my time (alone, with friends), and on what classes I spend most of my time. I can imagine creating bar charts with each of these categorical variables on the x-axis and total amount of time on the y-axis to represent the overall distribution of how I spend my time. I would then facet by time of day (morning, afternoon, evening, night) to see how I spend time differently during different hours. For all of these charts I would map area of the bars to amount of time spent on a specific activity and color to each categorical variable. I can imagine creating a table with time of day as rows (with bins morning, afternoon, evening, night), one of my categorical variables (where, with whom, on what class) as columns, and total time spent as entries.