STAT 231: Problem Set 1B

Lauren Pelosi

due by 5 PM on Friday, September 4

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:N/A

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 graphic shows the volume of alums of particular majors that filtered into particular jobs. Within particular majors there seem to be dominant career paths; for example, history majors are very likely to go into law, economics majors are very likely to enter the banking/financial industry, and biology majors frequently enter in health and medicine.

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:There are certain visual cues used: color indicates major, and area indicates volume of alumi. However, the coordinate system and scale cannot be described with the chapter's taxonomy.

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.

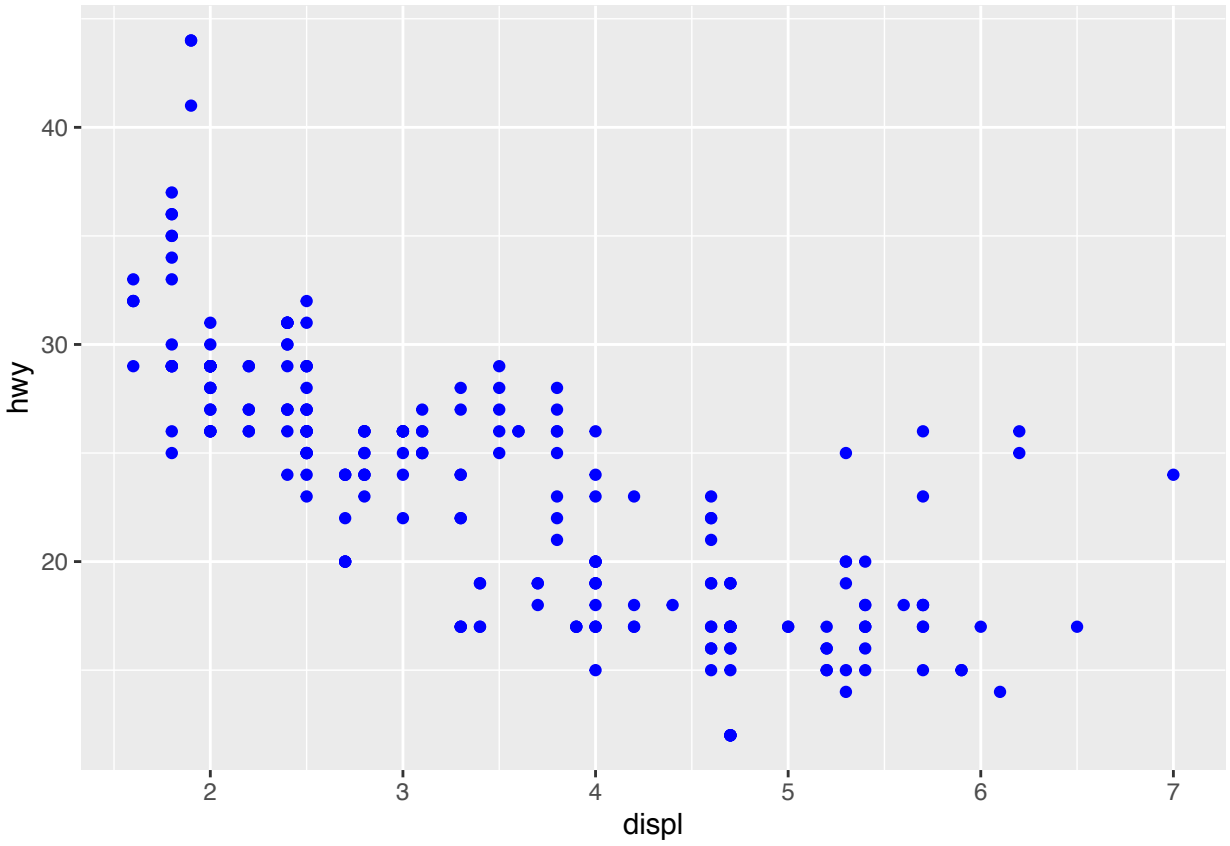
   ANSWER: I like the use of color and area as visual cues. However, I find that the double major component complicates what we're seeing and makes the graphic a bit misleading. I might have left double majors out of this graphic, and created another graphic showing major crossovers. That way, no individual's path will be represented twice.

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:When placed inside aes(), color = "blue" does not have meaning; inside aes(), you can equate color to a variable to use color as an indicator for different levels within that variable. However, you need color outside aes() to change the scatterplot in a way that is not tied to the data. Below, I've changed the code in that way.

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

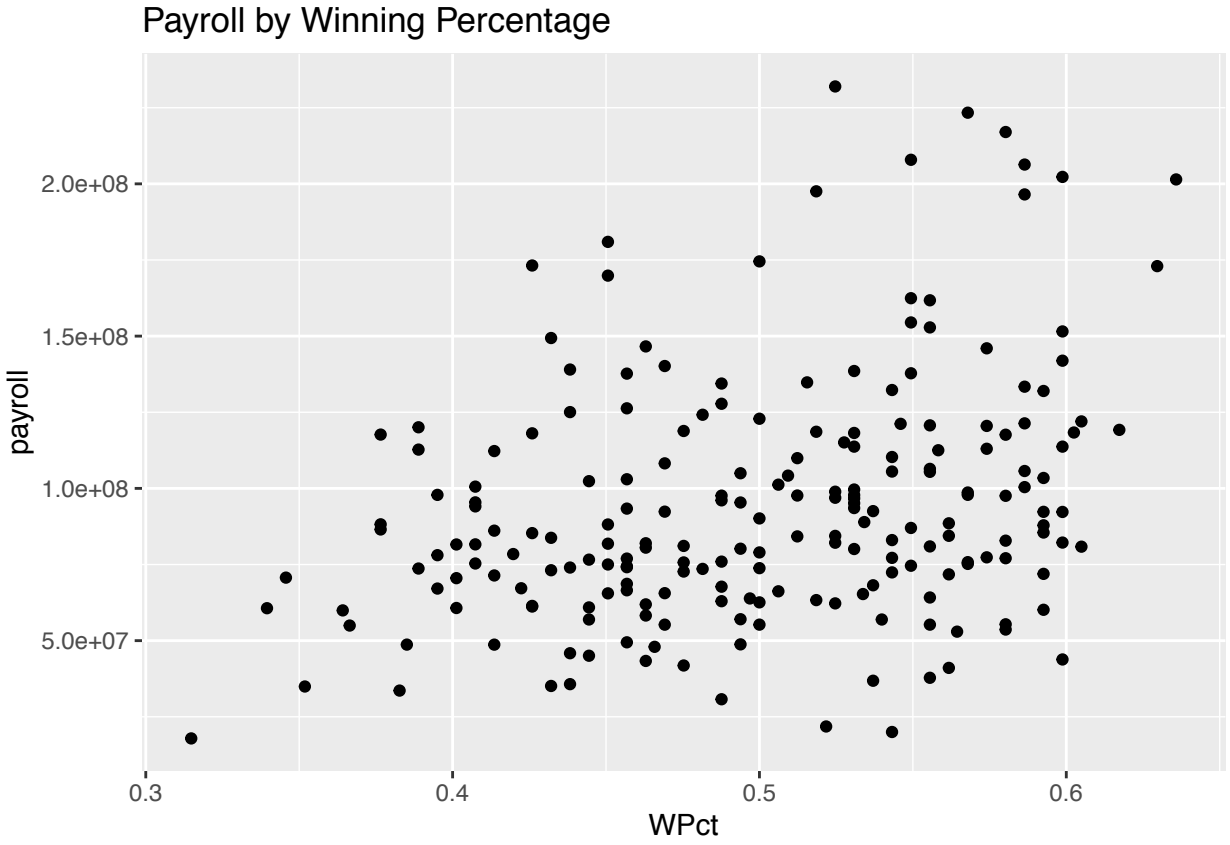


MDSR Exercise 3.6 (modified)

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 seems to indicate that as winning percentage increases, payroll increases on average as well. We could run a linear regression to test this.

```
MLB_teams %>%
  ggplot + geom_point(aes(x=WPct, y=payroll)) +
  labs(title = "Payroll by Winning Percentage")
```



MDSR Exercise 3.10 (modified)

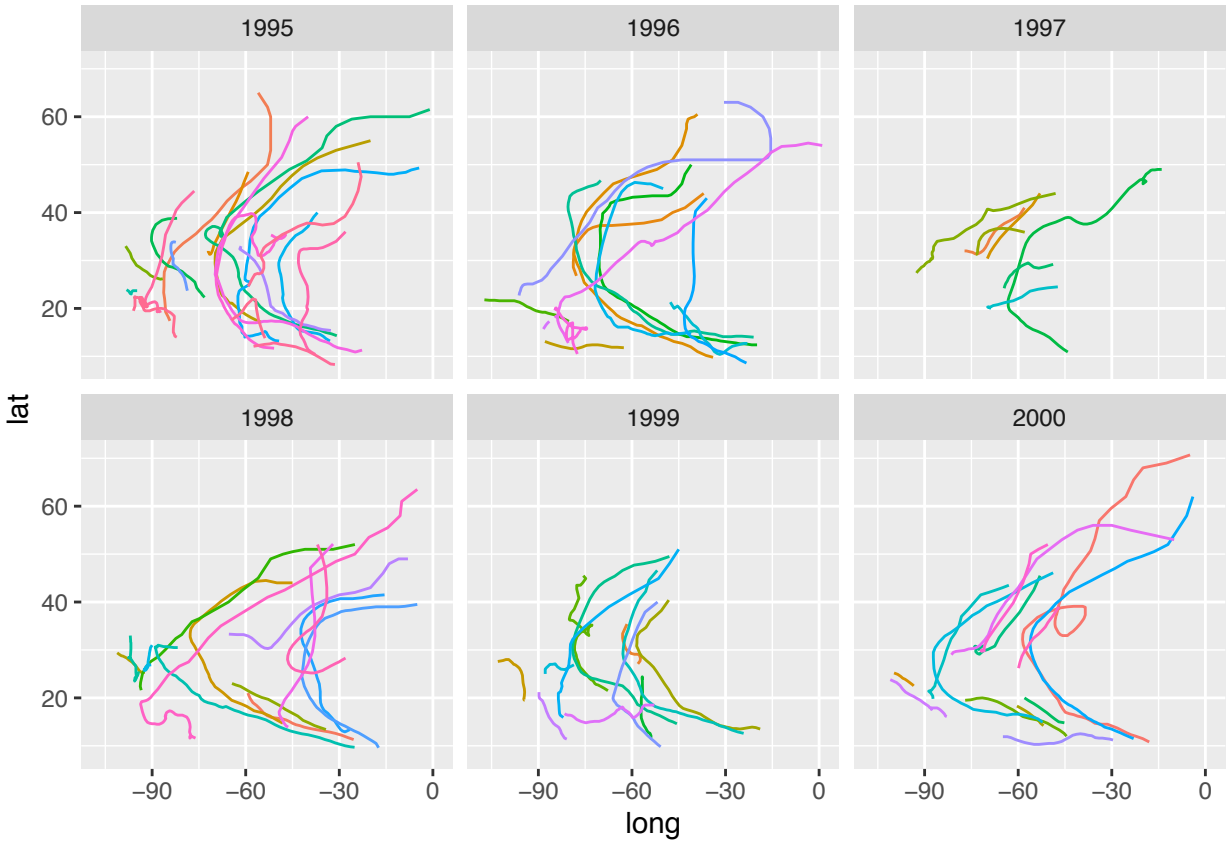
Using data from the `nasaweather` package, use the `geom_path()` function to plot the path of each tropical storm in the `storms` data table (use variables `lat` (y-axis!) and `long` (x-axis!)). Use color to distinguish the storms from one another, and use facetting to plot each `year` in its own panel.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

```
## # A tibble: 2,747 x 11
##    name    year month   day  hour   lat  long pressure  wind type        seasday
##    <chr>  <int> <int> <int> <int> <dbl> <dbl>    <int> <int> <chr>         <int>
##  1 Allis~  1995     6     3     0  17.4 -84.3     1005    30 Tropical D~       3
##  2 Allis~  1995     6     3     6  18.3 -84.9     1004    30 Tropical D~       3
##  3 Allis~  1995     6     3    12  19.3 -85.7     1003    35 Tropical S~       3
##  4 Allis~  1995     6     3    18  20.6 -85.8     1001    40 Tropical S~       3
##  5 Allis~  1995     6     4     0  22   -86        997    50 Tropical S~       4
##  6 Allis~  1995     6     4     6  23.3 -86.3      995    60 Tropical S~       4
##  7 Allis~  1995     6     4    12  24.7 -86.2      987    65 Hurricane         4
##  8 Allis~  1995     6     4    18  26.2 -86.2      988    65 Hurricane         4
##  9 Allis~  1995     6     5     0  27.6 -86.1      988    65 Hurricane         5
## 10 Allis~  1995     6     5     6  28.5 -85.6      990    60 Tropical S~       5
## # ... with 2,737 more rows
```

```
storms %>%
  ggplot + geom_path(aes(y=lat, x = long, color =name)) +
                       facet_wrap(~year) +
  scale_color_discrete(guide="none")
```


1995
1996
1997
1998
1999
2000
lat
long
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 would like to explore the breakdown of the time I'm allocating to each class. I hypothesize that I am spending significantly more time on one class than the others; however, I'd like to look into this because it could be that I enjoy that class less, and so it only seems like it is taking up a disproportionate amount of my time. Another question I would like to explore is if I let my schoolwork dictate the time I spend exercising. I try to exercise for a consistent amount of time daily despite schoolwork, and I would like to see if I'm successful.
>
> A bar chart (with a bar for each class on the x-axis and weekly hours on the y axis) with error bars will be useful to represent the number of hours I spend on each class and to help determine if there is a statistically significant difference. A scatterplot of hours spent on schoolwork in a day (x-axis) and hours spent exercising on that day (y-axis), arranged in descending order along the x-axis can help me determine if as schoolwork increases excersise decreases or stays the same. (A linear regression would help with this.)
>
> A data table including the columns Date, Hours Spent on each of my four classes, Total Hours Spent on Schoolwork, and Hours Spent Exercising will help me answer these questions. Each row will correspond to a day.