STAT 231: Problem Set 1B

Jett Knight

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: It shows the majors of over 15,000 Williams College alumni, and shows what career field they went into after graduating. The main message I can see is first, what majors at Williams College tend to be more popular than others, and second that the major you graduate with actually has a small impact on the career you choose to go into afterwards.

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: Yes it can. The visual cues used are color, area and position. A polar coordinate system is used. It is on a categorical scale.

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.

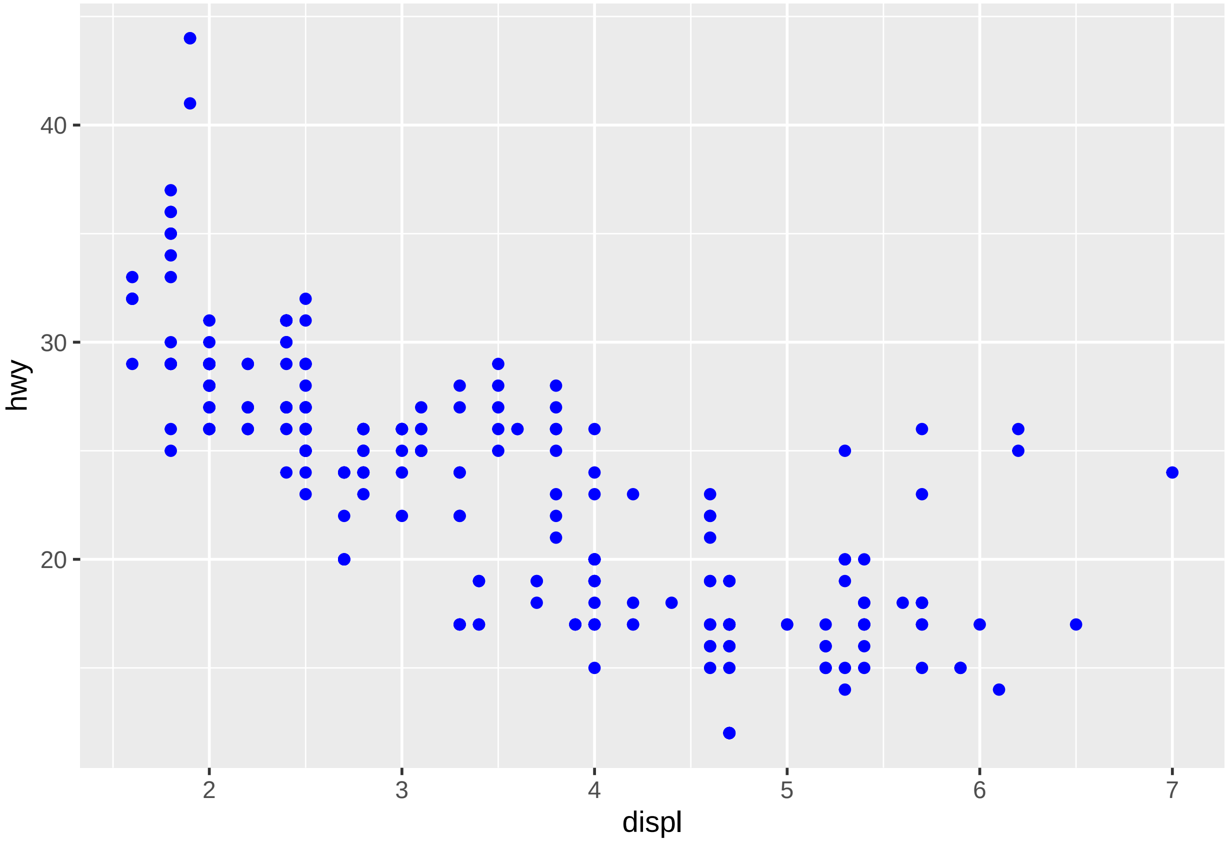
   ANSWER: Even though people are usually better at seeing differences in position than angles, this graphic really works, in my opinion. Color is used well both to differentiate the halves of the graphic and the different majors from each other. The graphic is also not overcrowded with unnecessary text.

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: Color is chosen in the aesthetics command. When it is done here, it only changes what the color is labeled as in the legend, not the actual color itself.

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

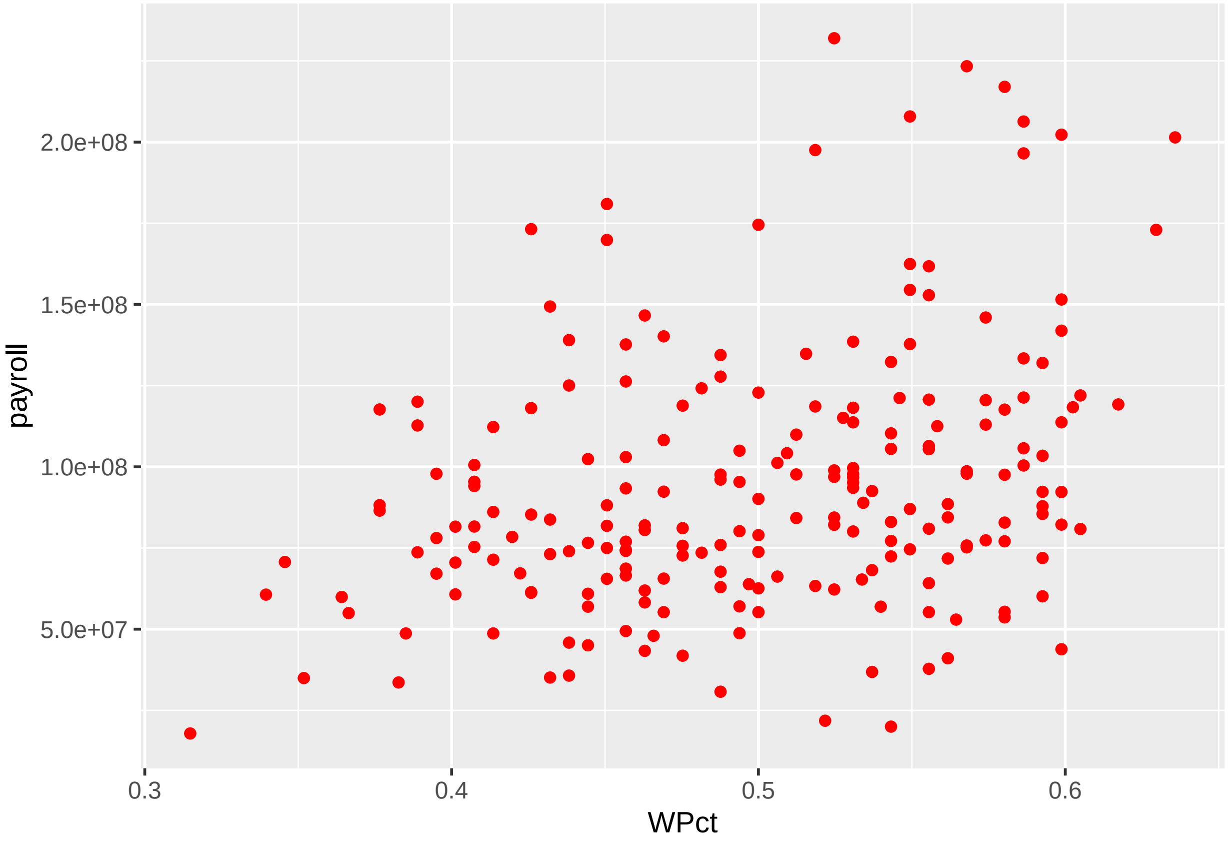


MDSR Exercise 3.6 (modified)

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: There is not a particularly strong relationship between winning percentage and payroll. While we don't see any high payrolls for low win percentages, we see plenty of high win percentages associated with low payrolls.

```r
library(mdsr)
ggplot(data = MLB_teams, aes(WPct, payroll)) + geom_point(color = "red")
```

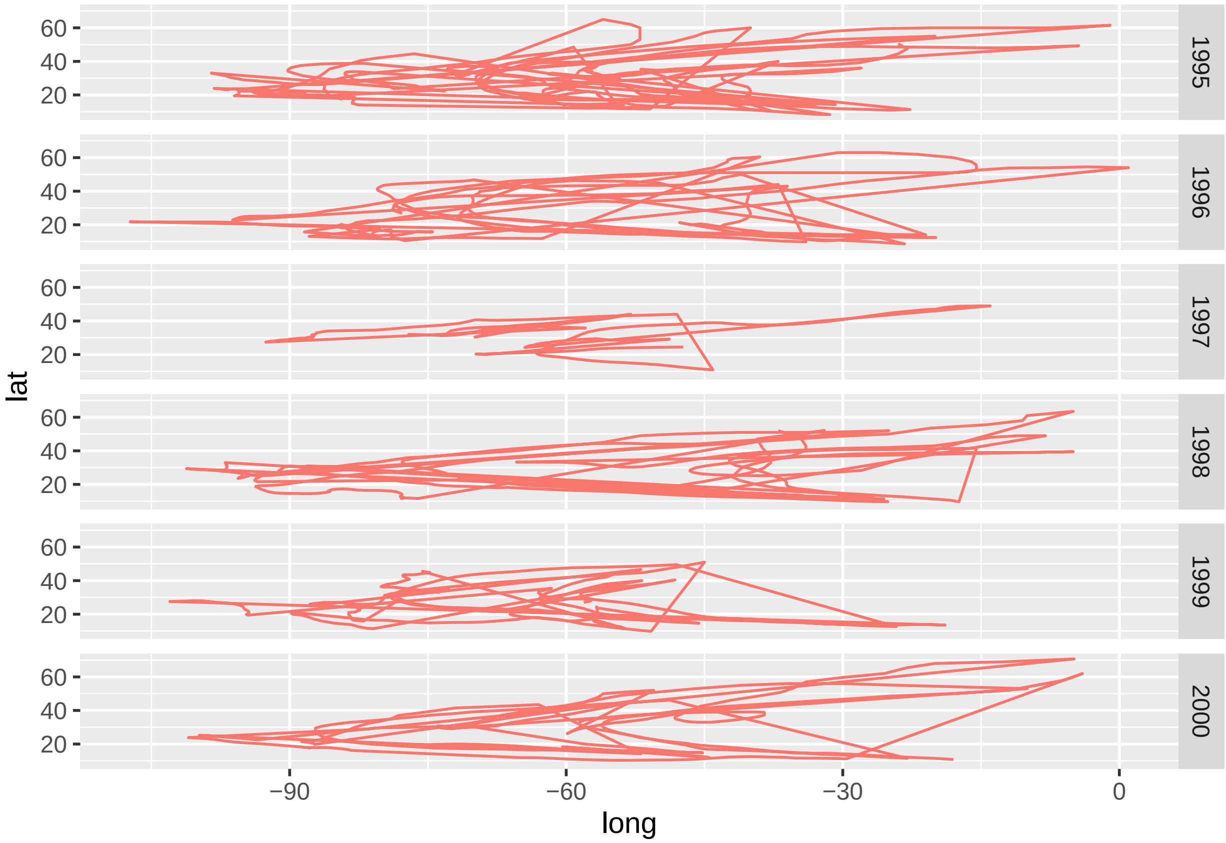


MDSR Exercise 3.10 (modified)

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(data = storms, aes(long, lat)) + geom_path(aes(color = "type")) + facet_grid("year") + scale_colo
```



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'm going to focus on how much time I spend on each course and how much time I spend watching TV. I could use scatterplots and bar charts to show these. One table could be comparing time spent in data science vs time spent doing computer science.