STAT 231: Problem Set 1B

Xander Schwartz

due by 5 PM on Friday, February 7

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 out of the course repo project in RStudio
5. Open up 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

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:The data shows the relationships between what students studied at Williams and what career path they took. The data shows that in almost every career there are students from every discipline (and vice versa).

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: Thee graphic uses a somewwhat confusing coordinate system that is probably closer to polar (as it is a circle) than anything else. In this chart, the data is made of of strands connecting to other parts of the other semi-circle (there are two distinct semi-circles of careers and studies, and the strands do not connect within the same semi-circle). There are some relatively arbitrary color scales in use showing the discipline (green for STEM, blue for social studies, etc.) and the thickness of the strands represents the amount of connenctions (presumably).

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.

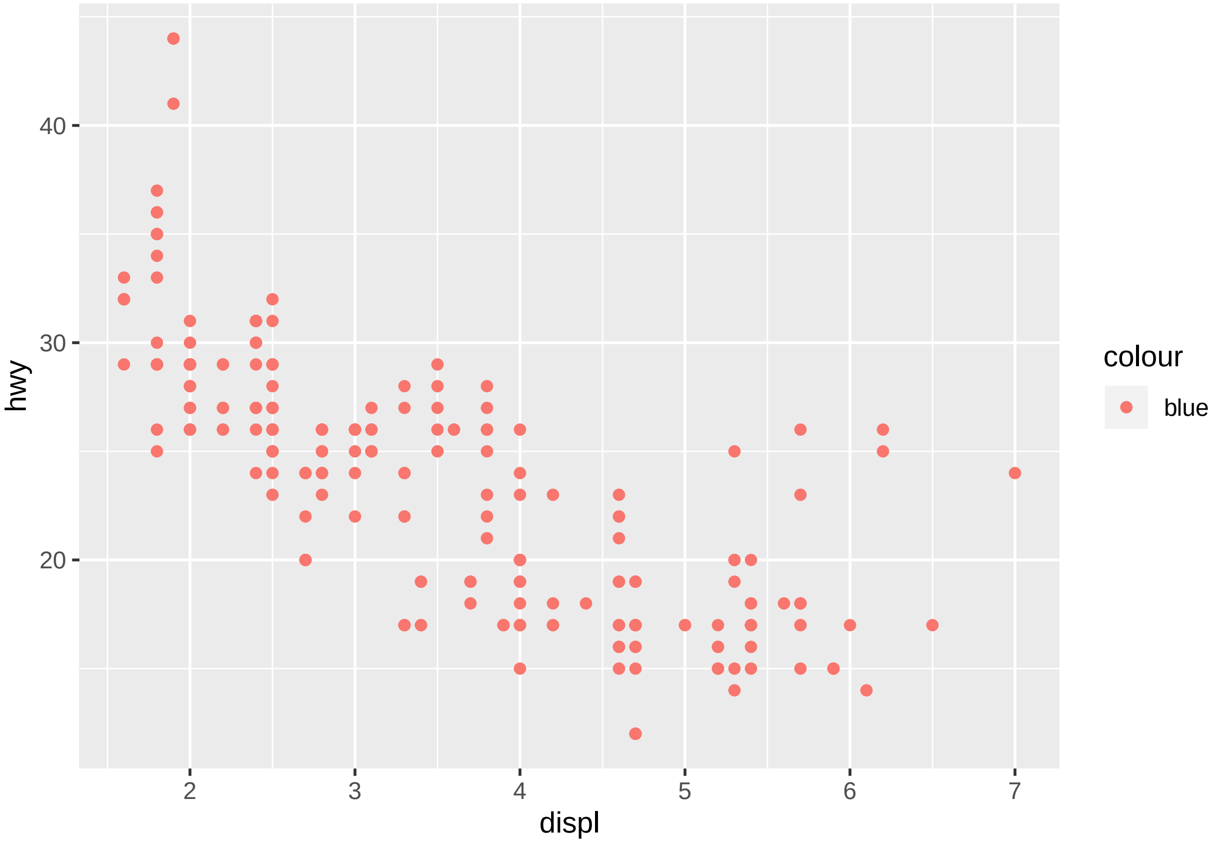
   ANSWER: In my mind, there is too much going on to fully understand everything. While it is very concise in packing a lot of informtion in one chart, it might be easier to have 14 seperate scaled pie charts of each discipline showing career paths than this mess. It feels like the designer prioritized artsiness over clarity.

Spot the Error (non-textbook problem)

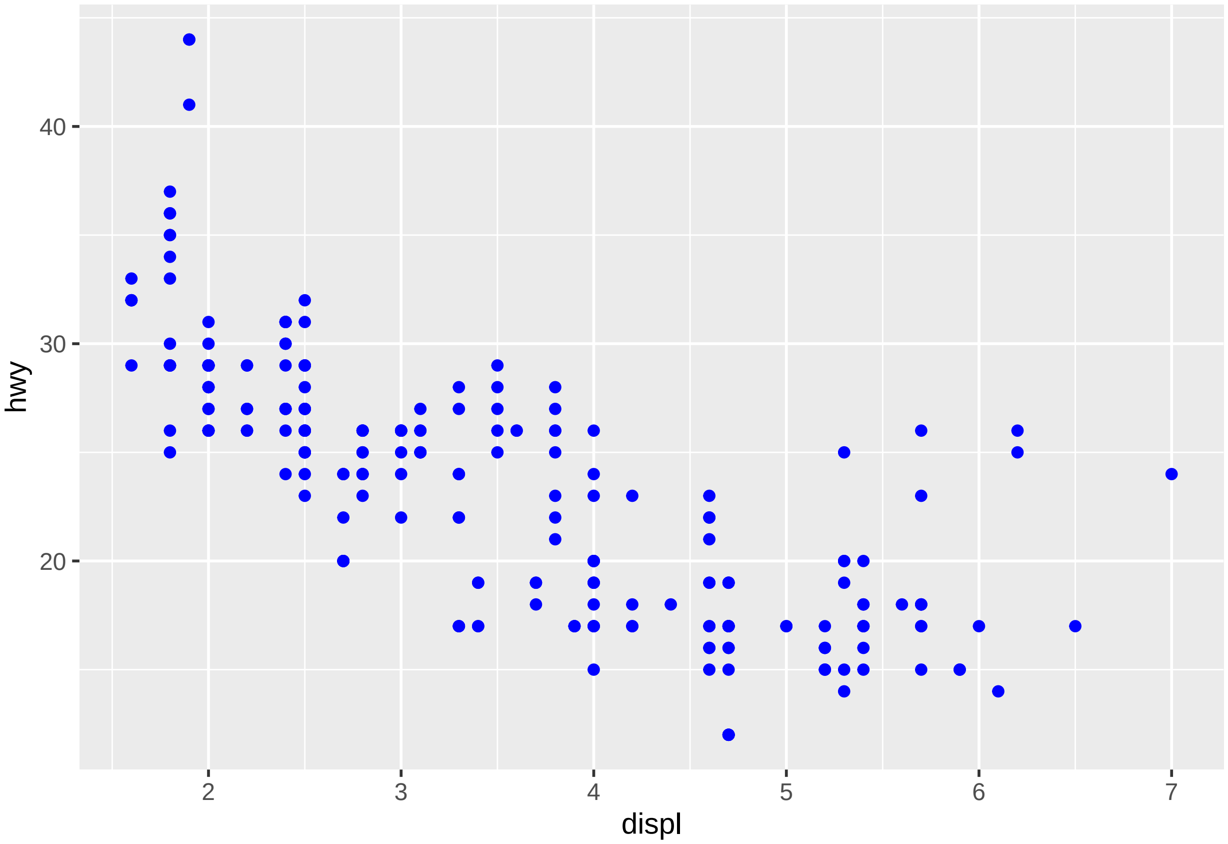
Explain why the following command does not color the data points blue, then write down the command that will turn the points blue.

ANSWER: In this exaample, mapping was setting the legend to say "blue" rather than the points.

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



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

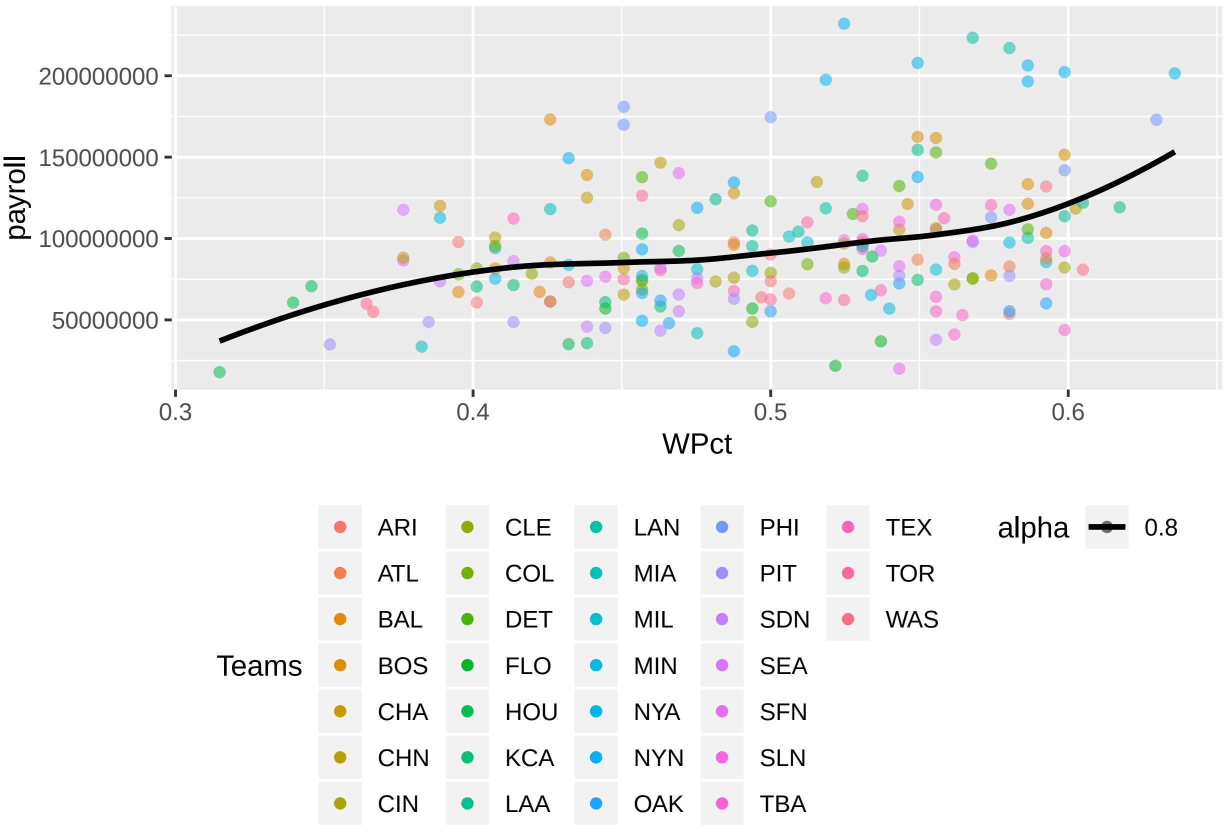

40
30
20
hwy
2
3
4
5
6
7
displ

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 this graph tell?

> ANSWER:The chart shows a positive but not overly strong correlation between winning perecentage and payroll. The slope seems to show more signifigance at higher ends of the spectrum than in the middle. This indicates that payroll is a decent predictor of on field success, although correlataion and causation are likely not the same in this scenario, as many baseball teamas are not spending money because their teams are not good in the first place, as opposed to vice versa (you might as well save money on the payroll while the team is bad).

```
Teams <- MLB_teams$teamID
ggplot(data = MLB_teams, aes(x=WPct,y=payroll,alpha=.8))+
  geom_point(aes(color=Teams)) +geom_smooth(se=F,color="Black")+
  theme(legend.position="bottom")
```

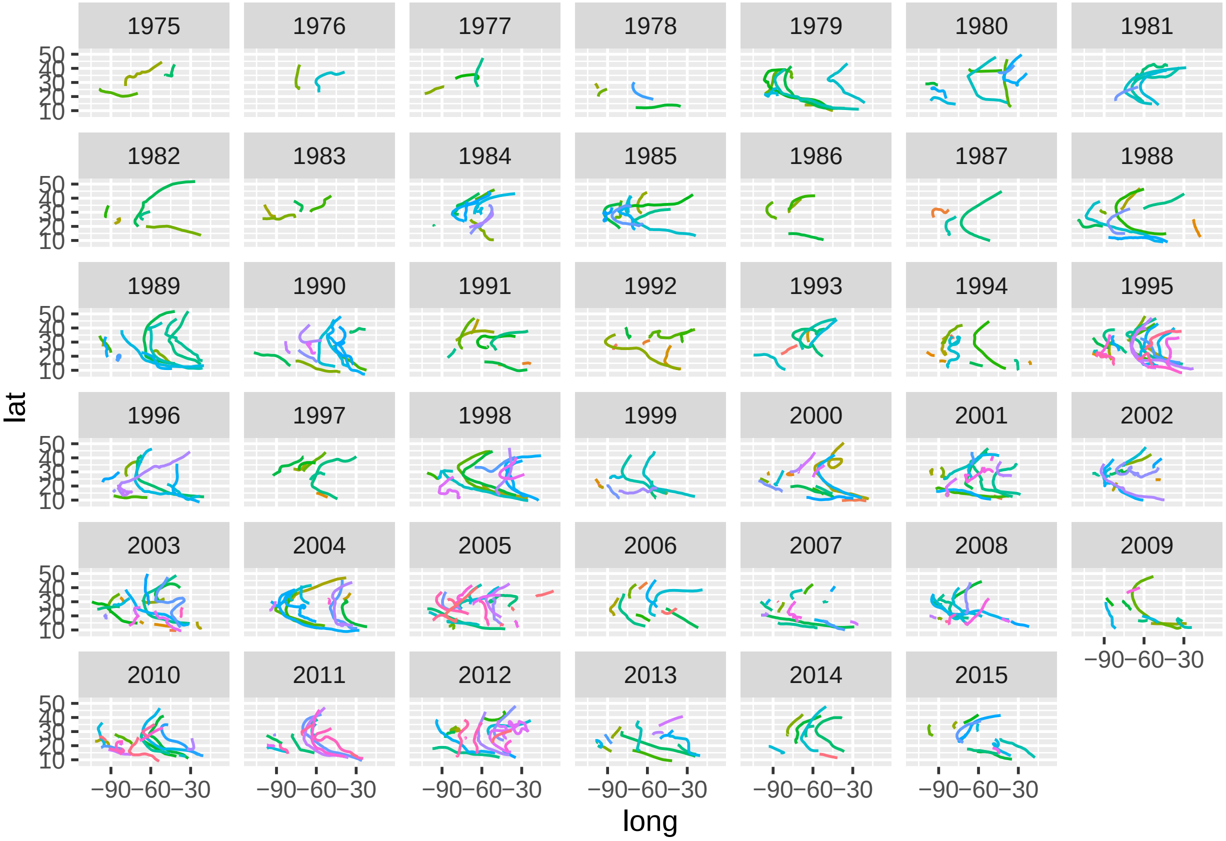


MDSR Exercise 3.10 (modified)

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!

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


lat
long