STAT 231: Problem Set 1B

Matthew Perkins

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: The graphic shows groupings of majors at Williams College and the career paths followed by those who choose a particular major. The main message I take away from this graph is that major choice is somewhat predictive of ultimate career path, although there are still many alternative paths students take.

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 data can be described fairly well using the taxonomy from the chapter. The graphic uses primarily color and area to distinguish major/career groups and the amount of people in each group. The data appears in a circle most like a polar coordinate system, and the scale used by the graphic is categorical.

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.

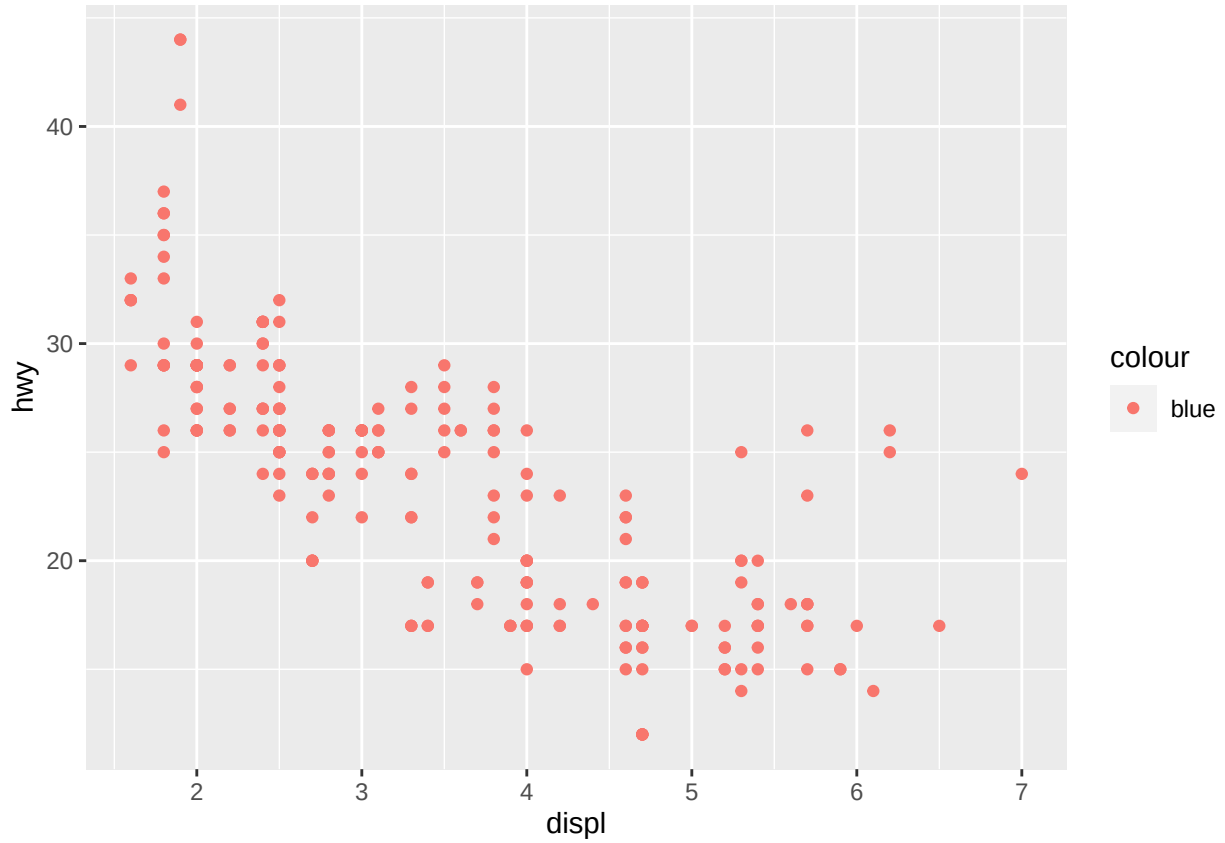
   ANSWER: I may have used more colors in the graphic beyond blue/brown/green to distinguish more between majors like economics and culture studies which are quite different, but otherwise I think the graphic is designed very well, highlighting differences in paths between groups, and the interactivity is crucial for making everything much more clear.

Spot the Error (non-textbook problem)

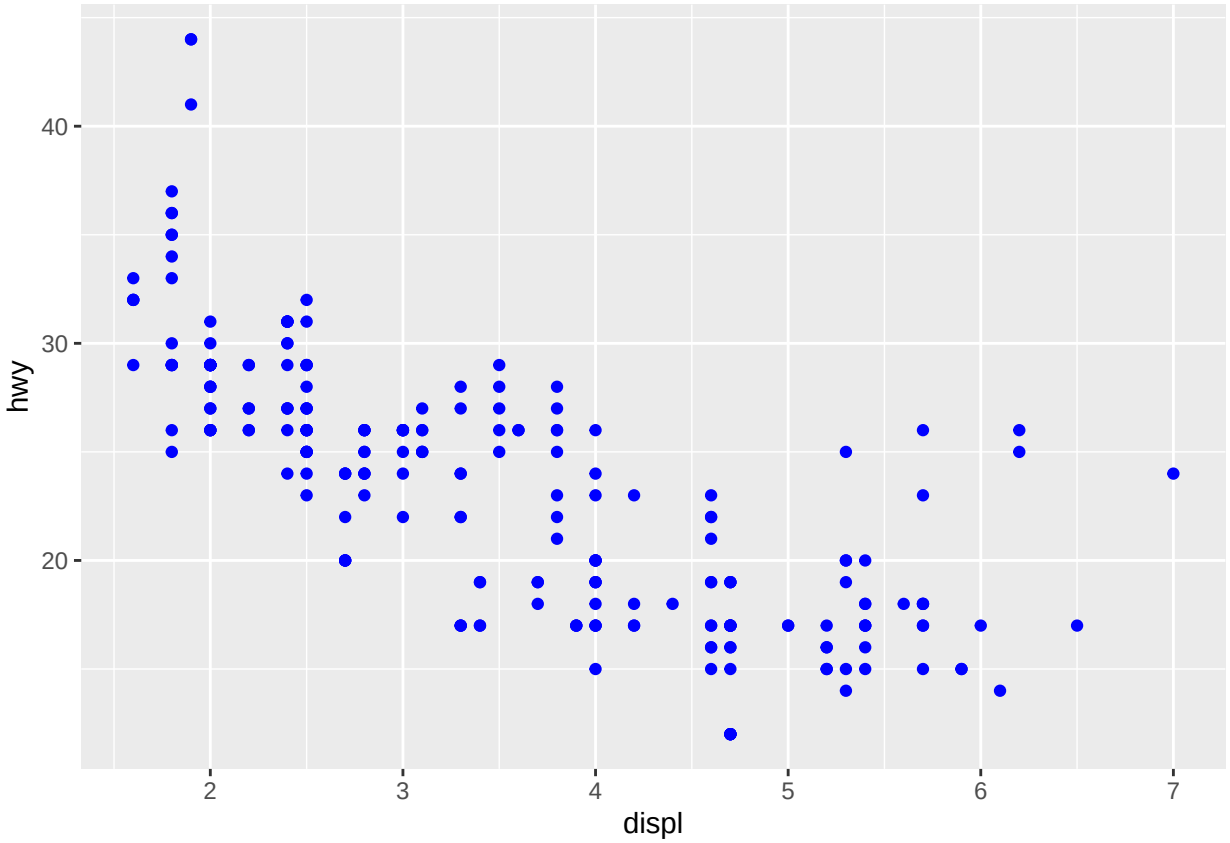
Explain why the following command does not color the data points blue, then write down the command that will turn the points blue.

> ANSWER: This command does not color the data points blue because it should not go into aes() since it is not an explicit mapping for a particular variable. Putting it outside the aesthetic fixes the error.

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




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


hwy
40
30
20
2
3
4
5
6
7
displ

MDSR Exercise 3.6 (modified)

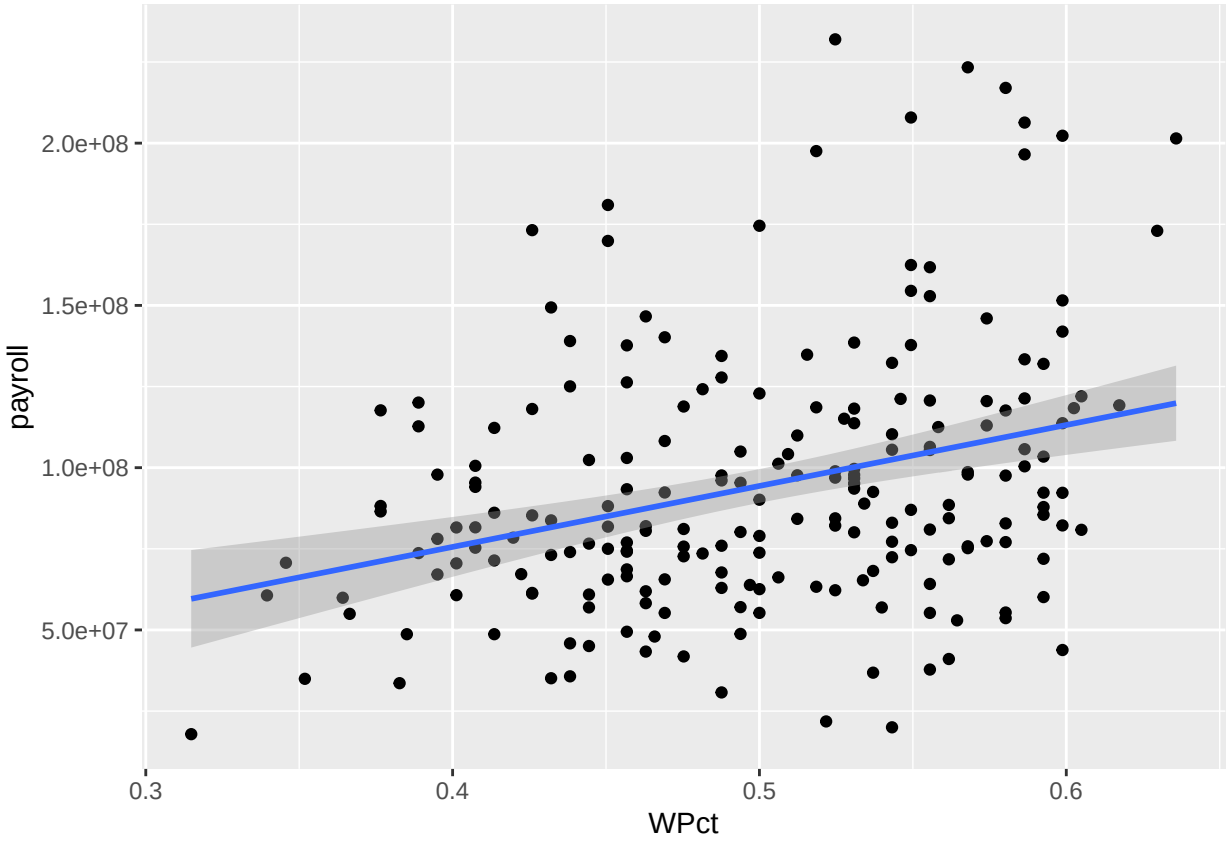
Use the `MLB_teams` data in the `mdsr` package to create an informative data graphic that illustrates the relationship between winning percentage and payroll in context. What story does your graph tell?

> ANSWER: In general, the higher the winning percentage the higher the payroll tends to be, but the relationship is not incredibly strong.

```
MLB_teams
```

```
## # A tibble: 210 x 11
##    yearID teamID lgID      W     L  WPct attendance normAttend payroll metroPop
##     <int> <chr>  <fct> <int> <int> <dbl>      <int>      <dbl>   <int>    <dbl>
##  1   2008 ARI    NL       82    80 0.506    2509924      0.584  6.62e7  4489109
##  2   2008 ATL    NL       72    90 0.444    2532834      0.589  1.02e8  5614323
##  3   2008 BAL    AL       68    93 0.422    1950075      0.454  6.72e7  2785874
##  4   2008 BOS    AL       95    67 0.586    3048250      0.709  1.33e8  4732161
##  5   2008 CHA    AL       89    74 0.546    2500648      0.582  1.21e8  9554598
##  6   2008 CHN    NL       97    64 0.602    3300200      0.768  1.18e8  9554598
##  7   2008 CIN    NL       74    88 0.457    2058632      0.479  7.41e7  2149449
##  8   2008 CLE    AL       81    81 0.5      2169760      0.505  7.90e7  2063598
##  9   2008 COL    NL       74    88 0.457    2650218      0.617  6.87e7  2754258
## 10   2008 DET    AL       74    88 0.457    3202645      0.745  1.38e8  4296611
## # ... with 200 more rows, and 1 more variable: name <chr>
```

```
ggplot(data = MLB_teams) +
  geom_point(mapping = aes(x = WPct, y = payroll)) +
  geom_smooth(aes(x = WPct, y = payroll), method = "lm")
```


2.0e+08
1.5e+08
1.0e+08
5.0e+07
payroll
0.3
0.4
0.5
0.6
WPct

MDSR Exercise 3.10 (modified)

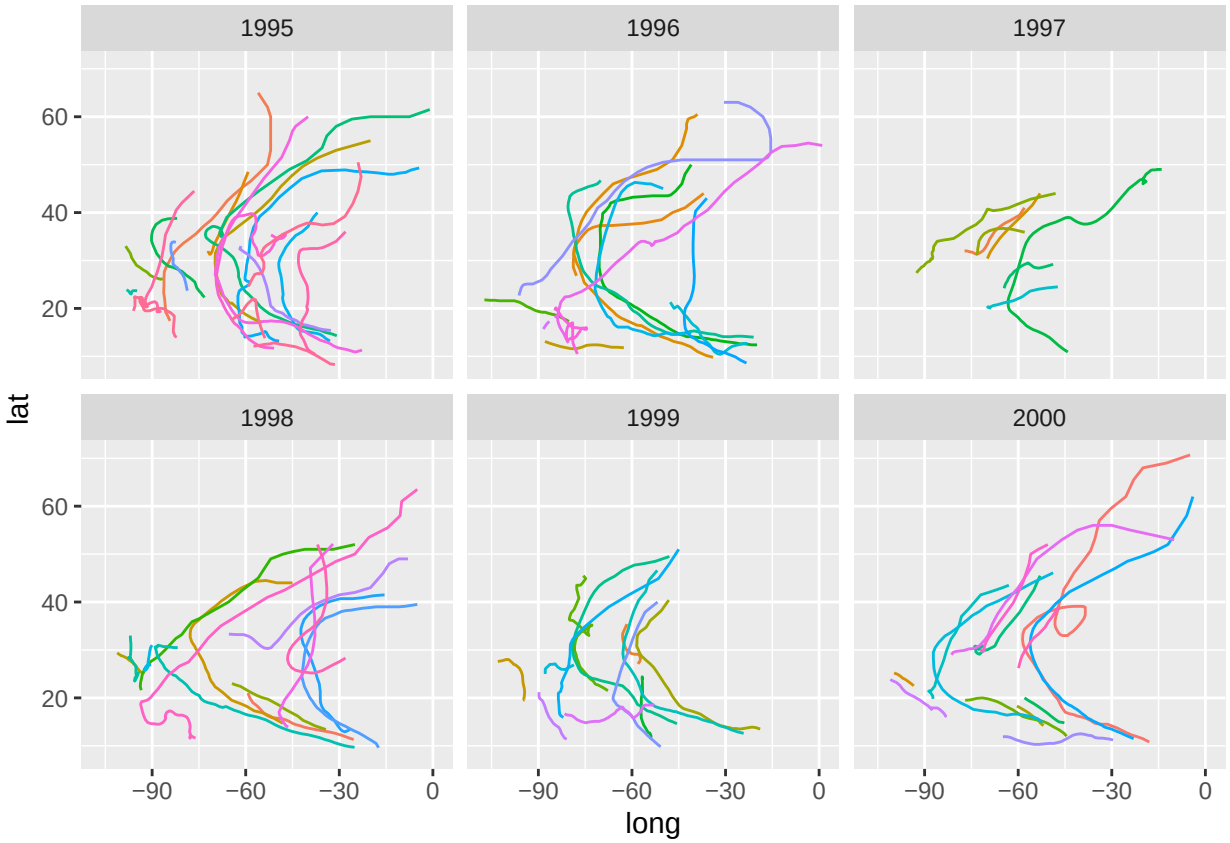
Using data from the `nasaweather` package, use the `geom_path()` function to plot the path of each tropical storm in the `storms` data table (use variables `lat` (y-axis!) and `long` (x-axis!)). Use color to distinguish the storms from one another, and use facetting to plot each `year` in its own panel.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

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


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 plan to focus on understanding which activities I spend most of my time on, and how the amount of work vs. leisure changes as the semester progrsses. The first visualization will be a bar graph comparing the total cumulative time spent on each activity, and the second visualization will be a scatter plot showing the percentage of time spent on work as time progresses. The table will show each day as a row and each activity as a column and will count how much time was spent on a given activity on a given day.