STAT 231: Problem Set 1B

YOUR NAME HERE

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: You see the distribution of job outcomes for each major and the distribution of majors that went into each job outcome. The graph shows that there tend to be job sectors that majors tend to conglomerate to, but there are people of all majors in each sector 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: The visual cues of the data graphic can be described with the textbook's taxonomy to some extent, but the graphical, interconnected network and lack of numerical data shown well lie outside the textbook's normal taxonomy for coordinate systems and scales. Every major and job department is given its own color and patial circumference along a circle split into two halves for jobs/majors, and all jobs are connected to the majors that contribute to its workforce while all majors are connecte dot the job departments their graduates venture into. The width of the curved line connected is proportional to the portion of the total major/workforce combination that line represents. We can assume this means the scale is numeric and based on percentages to some extent, however it is not explicitly stated anywhere, and this can only be assumed since the sum of line widths seems to correspond to the width of the partial circumference a major/job occupies.

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.

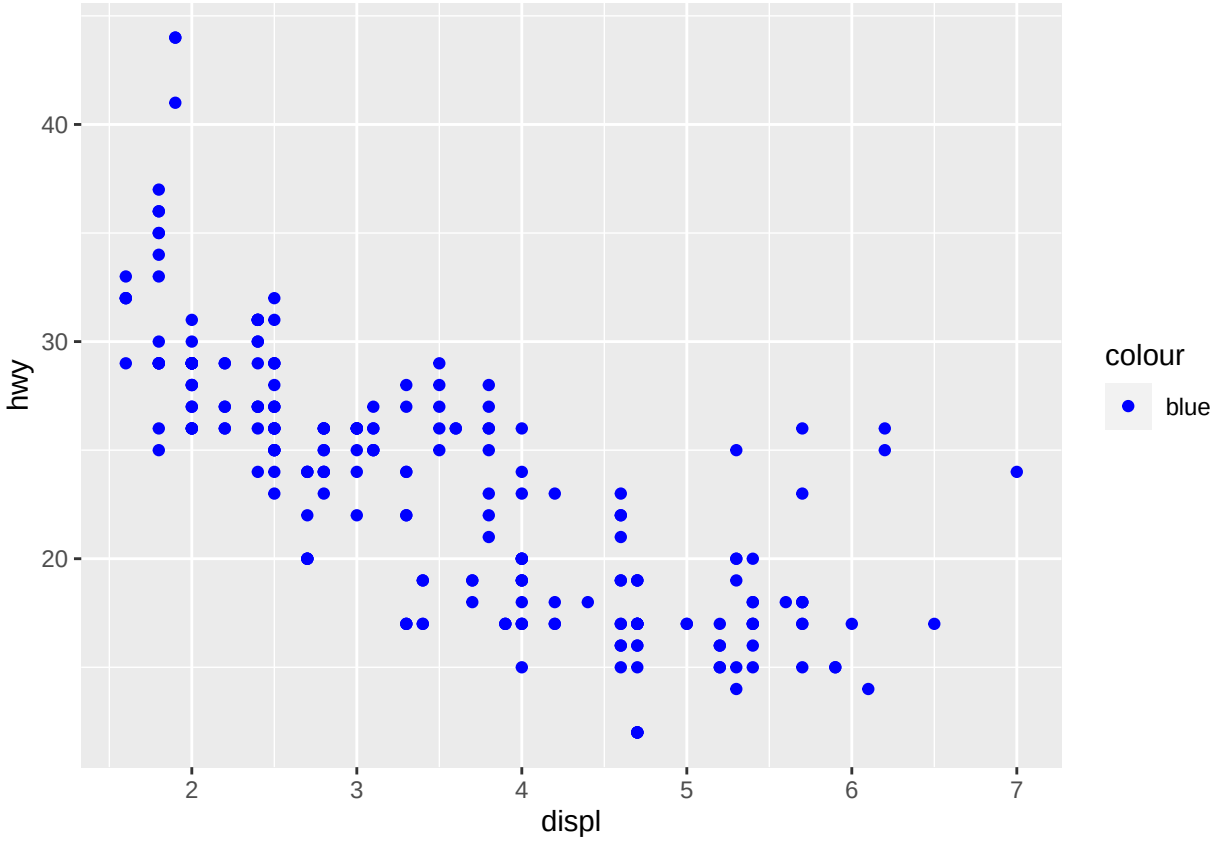
ANSWER: The visualiztion is thought-provoking since it shows how there is correlation between major and job, but the strength of the correlation caries greatly dpeending on what job or major you're looking at. Furthermore, basically everything is connected which shows that your major does not necessarily determine your job path, and everyone in a job department isn't necessarily from a particular major or group of majors. On the other hand, percentages should've been explicitly shown somehow. There was a clear lack of numerical scale of any kind, and the width of lines was hard to follow as they curved, looped, bended, etc. The differentiation of color for nearby jobs/majors was very helpful though in following along.

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:The geom_point funciton can affect how and where points are drawn, and the color option in the aes function only differentiates groups by color but isn't used to designate a particular color to all points.

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



MDSR Exercise 3.6 (modified)

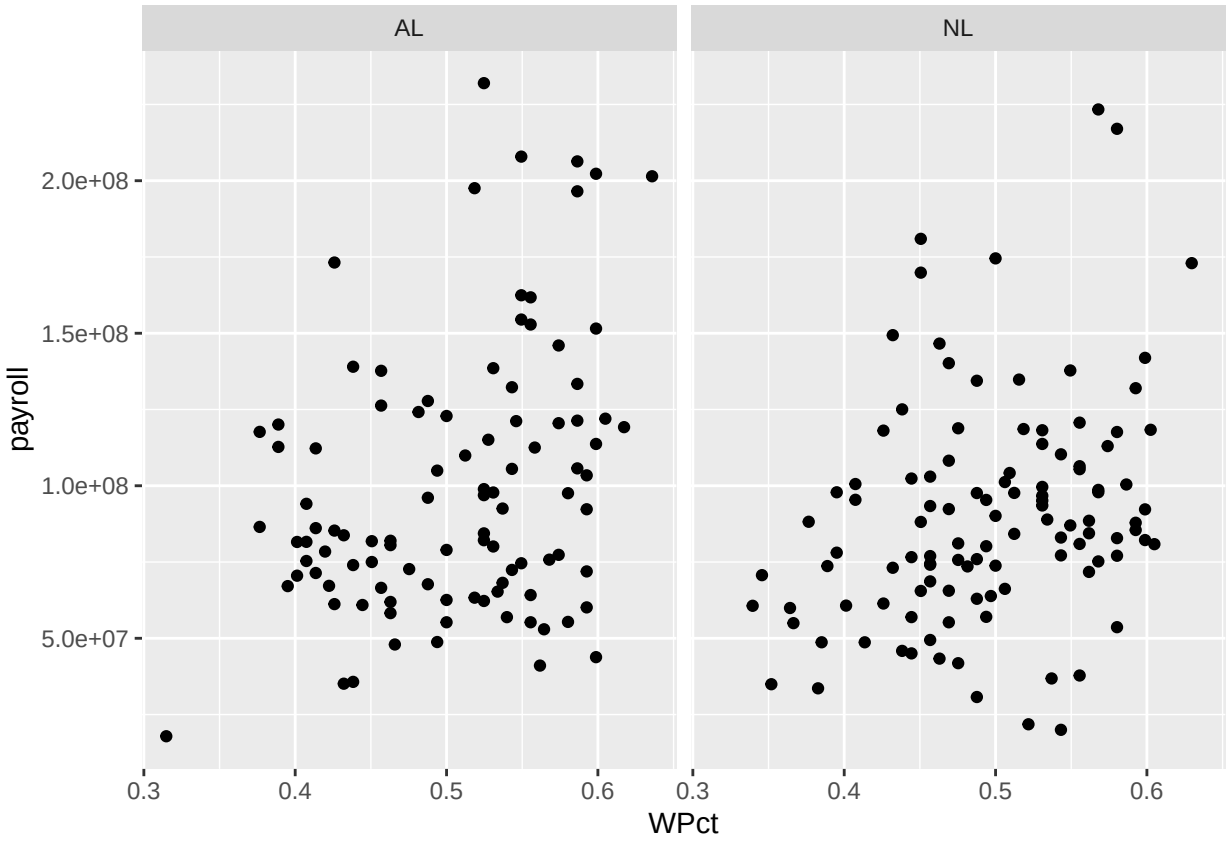
Use the `MLB_teams` data in the `mdsr` package to create an informative data graphic that illustrates the relationship between winning percentage and payroll in context. What story does your graph tell?

> ANSWER: There is a small correlation between winning percentage and payroll, but it's very weak. Even wehn accounting for year, league, and even the individual team, there isn't a clear, consistent pattern in the relationship between payroll and winning percentage.

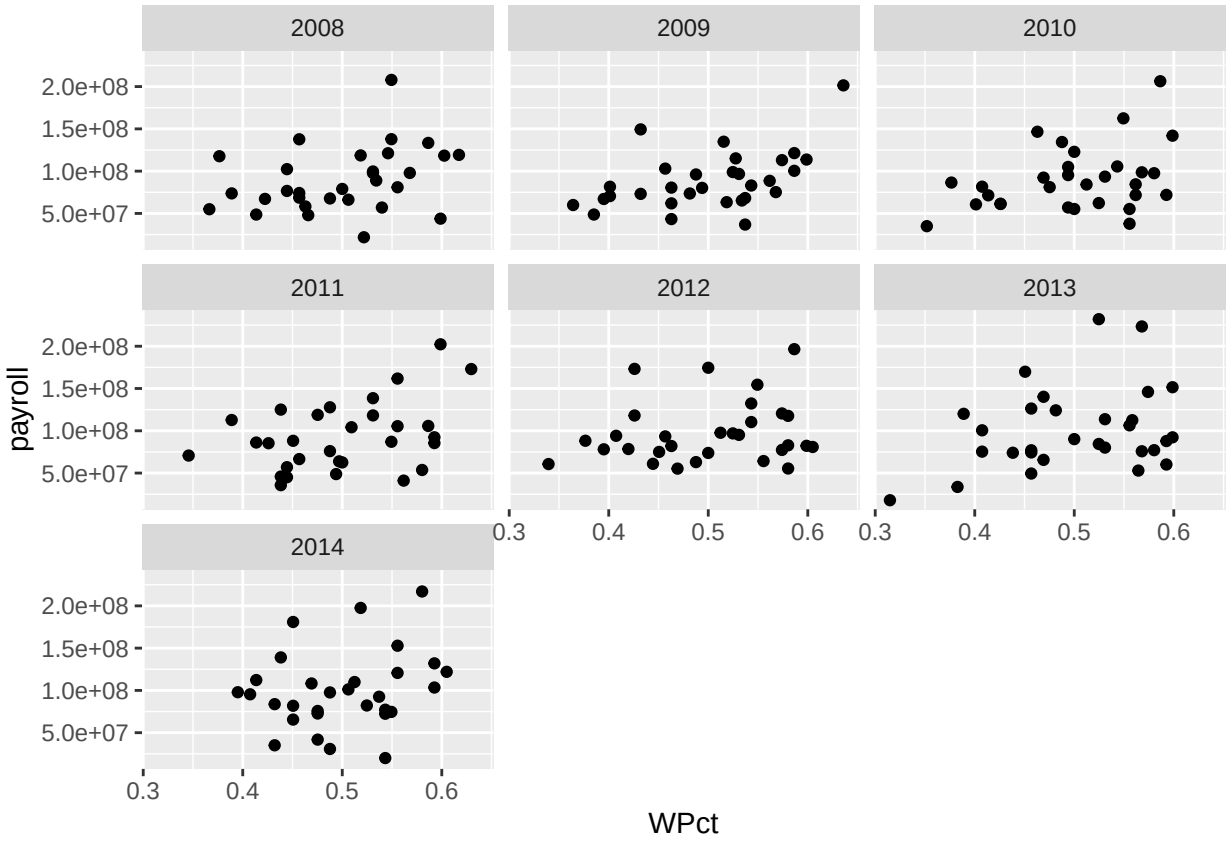
```
library(mdsr)
MLB_teams
```

```
## # 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>
```

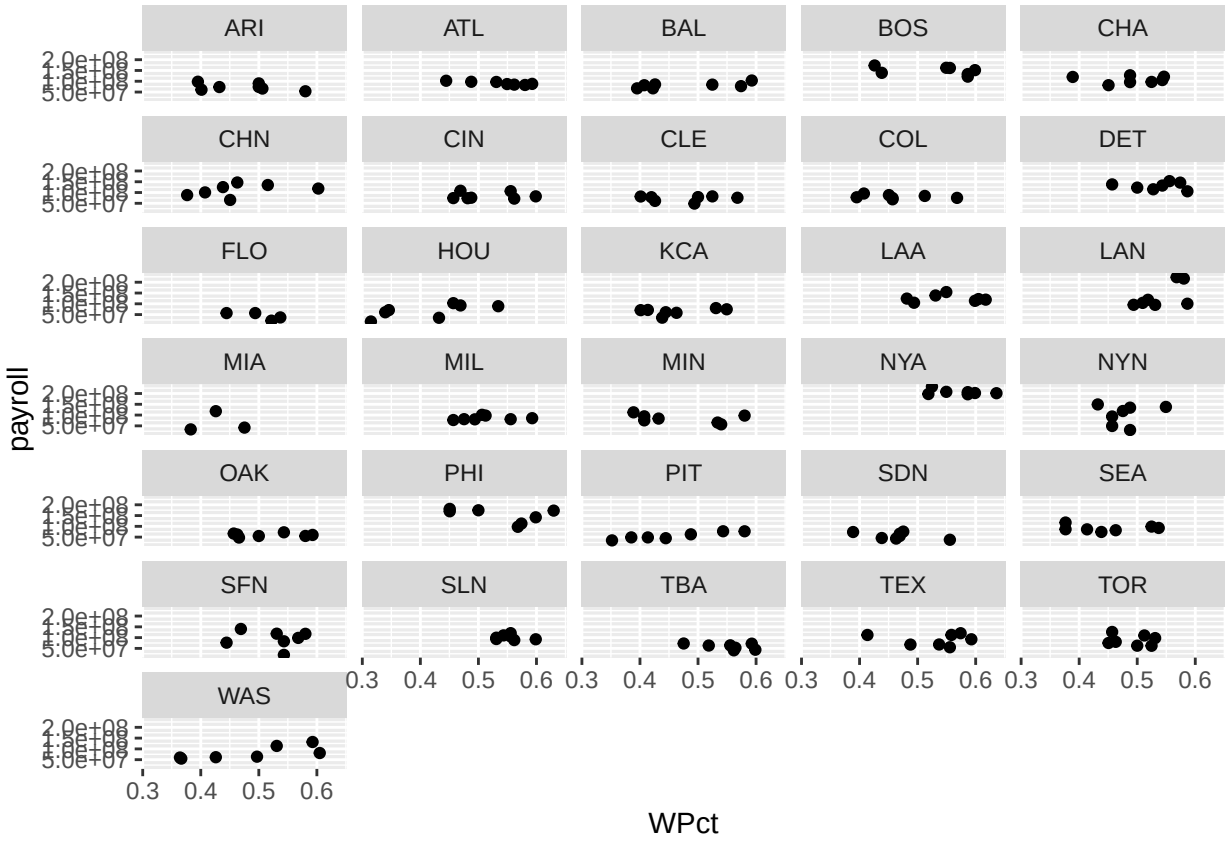
```
rel <- ggplot(data = MLB_teams) + geom_point(mapping = aes(x = WPct, y = payroll))

rel + facet_wrap(~lgID)
```




```
rel + facet_wrap(~yearID, nrow=3)
```




```
rel + facet_wrap(~teamID, nrow=7)
```

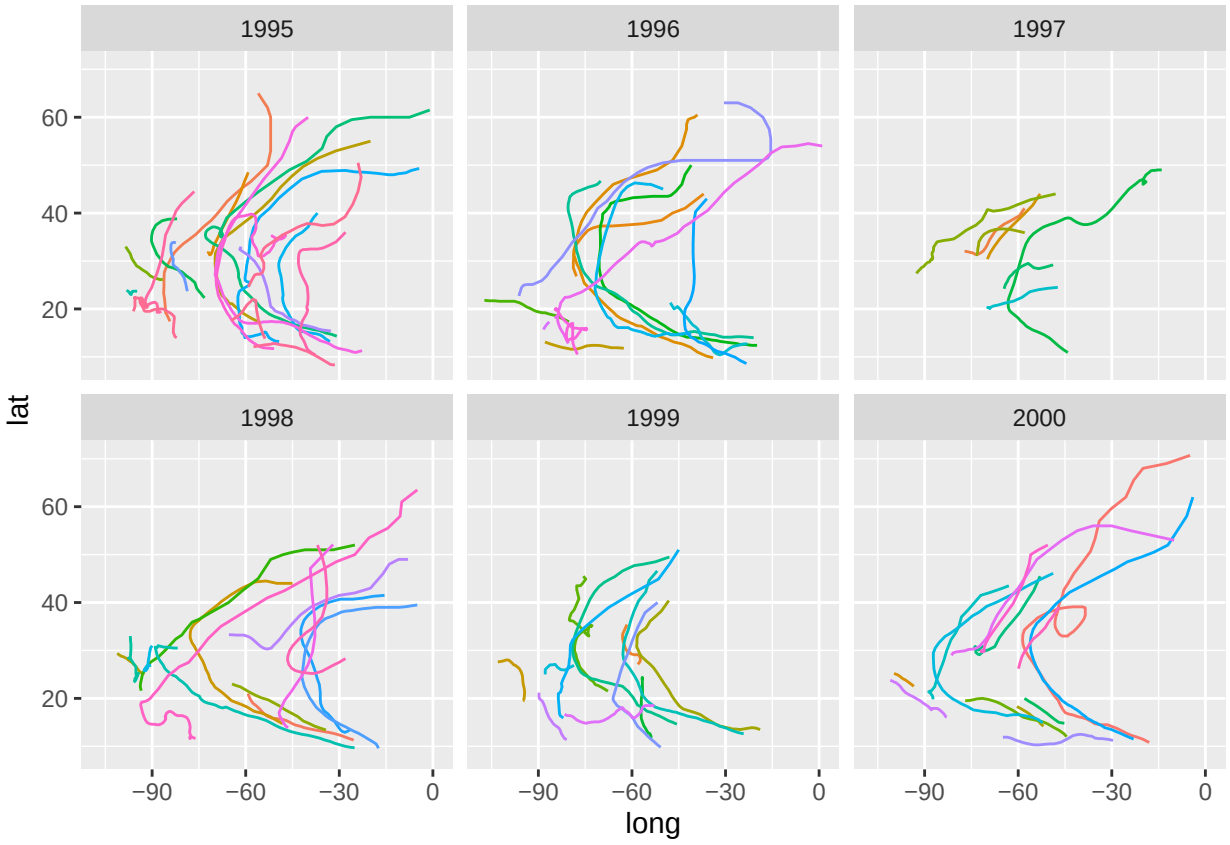

ARI
ATL
BAL
BOS
CHA
CHN
CIN
CLE
COL
DET
FLO
HOU
KCA
LAA
LAN
MIA
MIL
MIN
NYA
NYN
OAK
PHI
PIT
SDN
SEA
SFN
SLN
TBA
TEX
TOR
WAS
payroll
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.3
0.4
0.5
0.6
WPct

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)
ggplot(data = storms) + geom_path(mapping = aes(y=lat, x = long, color = name)) + scale_color_discrete(g
```



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: Does going to the gym make me more or less likely to follow the schedule I preemptively made for afterwards? Are there certain days of the week where I am more/less likely to do as I say I will, and are there certain days where I unexpectedly do things I wasn't planning to more often? I could use pie charts to compare what I intend to do each day in terms of activites vs. what I actually do. I could also make a split bar graph with percentages for how closely I follow my schedule on the y axis and on the x axis have time after gym reservation and time before gym reservation. A table could also show my distribution of time allocation to each activity across the week by having rows for each day and columns for each acitivity. Cells could store time spent, and two tables could be made for what I plan and what I actually do.