Practice Set 1

Ziji Zhou

Due by 10pm ET on Friday

Practice Set Information

During the week, you will get further practice with the material by working through the Practice Set, a set of problems designed to give you practice beyond the examples produced in the text.

You may work through these problems with peers, but all work must be completed by you (see the Honor Code in the syllabus) and you must indicate who you worked with below.

Even then, the best approach here is to try the problems on your own before discussing them with peers, and then write your final solutions yourself.

GitHub Workflow

1. Before editing this file, verify you are working on the copy saved in *your* repo for the course (check the filepath and the project name in the top right corner).
2. Before editing this file, make an initial commit of the file to your repo to add your copy of the problem set.
3. Change your name at the top of the file and get started!
4. You should *save, knit, and commit* the .Rmd file each time you've finished a question, if not more often. You should also *push* your commits back onto GitHub occasionally (you can do this after each commit).
5. When you think you are done with the assignment, save the pdf as "*Name_thisfilename_date*.pdf" before committing and pushing (this is generally good practice but also helps me in those times where I need to download all student homework files).

Gradescope Upload

For each question (e.g., 3.1), allocate all pages associated with the specific question. If your work for a question runs onto a page that you did not select, you may not get credit for the work. If you do not allocate *any* pages when you upload your pdf, you may get a zero for the assignment.

You can resubmit your work as many times as you want before the deadline, so you should not wait until the last minute to submit some version of your work. Unexpected delays/crises that occur on the day the assignment is due do not warrant extensions (please submit whatever you have done to receive partial credit).

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (notes, textbook, etc) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

-

Problem 1 **MDSR Exercise 2.5 (modified)** Consider the data graphic for Career Paths at Williams College. Focus on the graphic under the "Major-Career" tab.

1.1 What story does the data graphic tell? What is the main message that you take away from it?

The graph tells the popularity of each major and industry as well as the distribution of each major into each industry and vice versa. There are obvious distributions of majors and industries (ie chem/bio into health profession) but many of the less career focused majors branch out into a large variety of professions. For example English/Literature had a very even spread among literally every industry.

1.2 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.

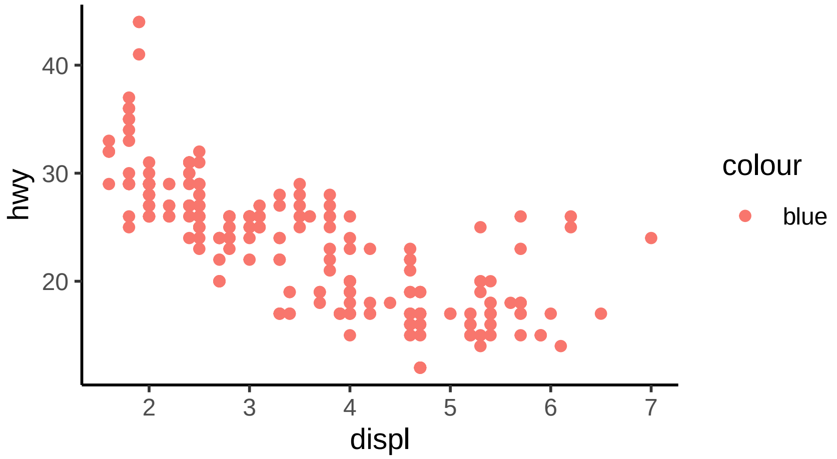
While i do not believe the taxonomy of this chapter covers the exact nature of the visualization given to us. The closest example would be of 3.2.4 Networks though it doesn't go in depth with the concept. The main visual cues and variables in this graph are the two variables of major and industry. The visual cues utilized through color for majors, size for popularity of major and industry, and line thickness for proportion of the correlation between each major and profession. The visualization of correlation can be described through the taxonomy in this chapter, though very inefficiently due to the qualitative nature of both major and industry. However, for example a faceted bar bar graph for each major and distribution of industry could help display the same data. The circular layout of the graph isn't covered in this chapter, as well as a plot fitting for the amount of qualitative information in this dataset.

1.3 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.

The graphic is a very efficient and clean cut presentation of the data. It is however extremely cluttered in the compilation graph with it very difficult to discern the individual lines and too messy to easily see a pattern of data. However, the interactivity of the graph helps with the cluster of the original graphic, helping narrow down and clarify the data presented. I found this graphic very interesting in mostly showing the wide variety of industries that the humanities feed off into. I thought it was very thought provoking and leads to many more questions about the connections of industries and majors that are seemingly disjointed, for example the few art majors that went into law or medicine. The colors in this graph could be more aesthetic, with the compilation leaving not the prettiest of pictures. I also would've perhaps created a lower cap for the compilation in order to declutter the graph and more easily see the patterns of data with everything present.

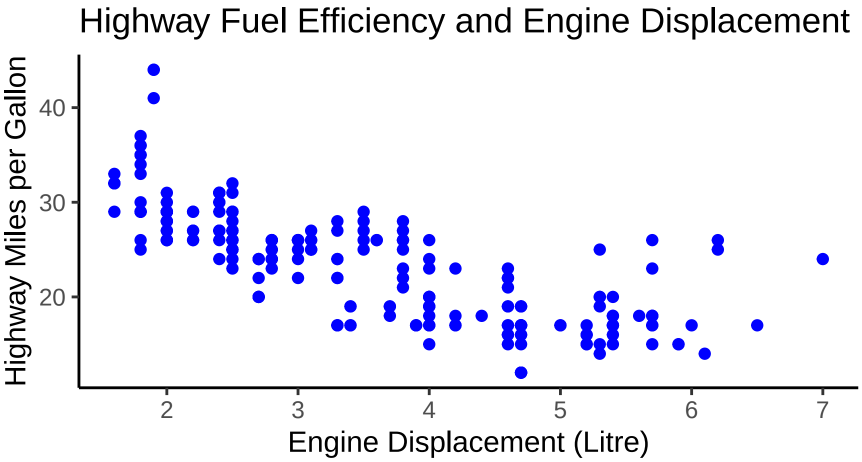
Problem 2 **Spot the Error** Explain why the following command does not color the data points blue, then write down (in a new code chunk) the command that will turn the points blue. Use the help file for the dataset to additionally update the graphic with informative axis labels and a title.

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



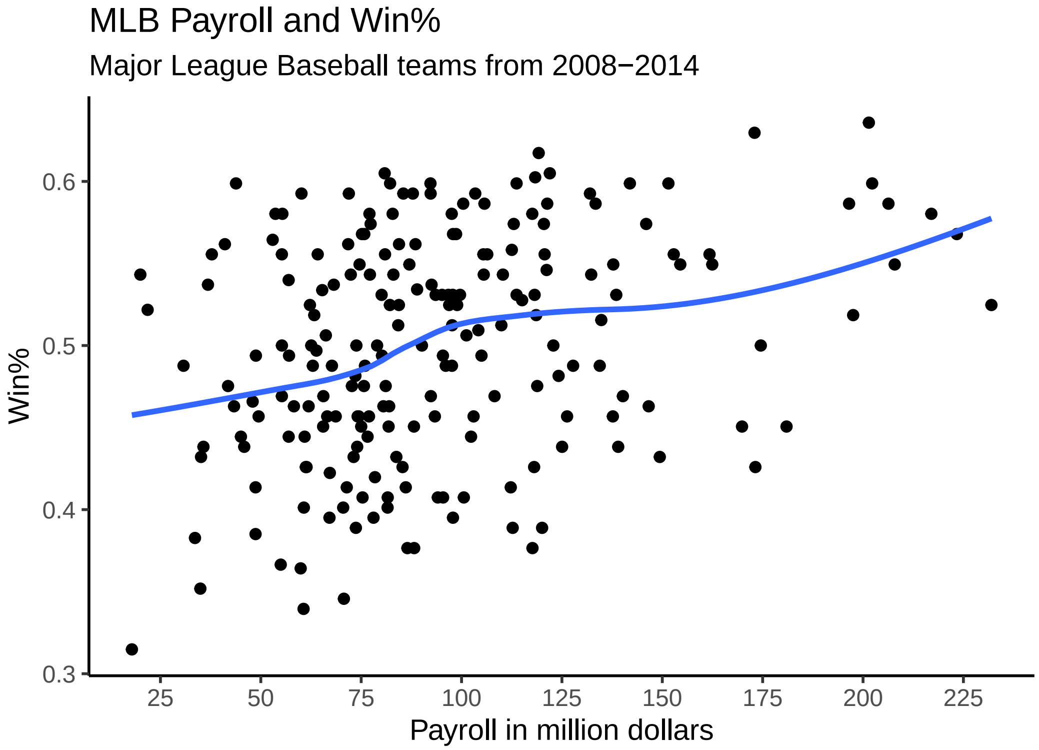
The command placed the `color = "blue"` argument inside the `aes()` function, leading to the program attempting to pair up the color of the dot with the `"blue"` data inside mpg, which doesn't exist, leading to the graph grouping all the points as `"blue"` and coloring it with a default red.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), color = "blue") +
  labs(title = "Highway Fuel Efficiency and Engine Displacement",
       x = "Engine Displacement (Litre)",
       y = "Highway Miles per Gallon")
```



Problem 3 **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?

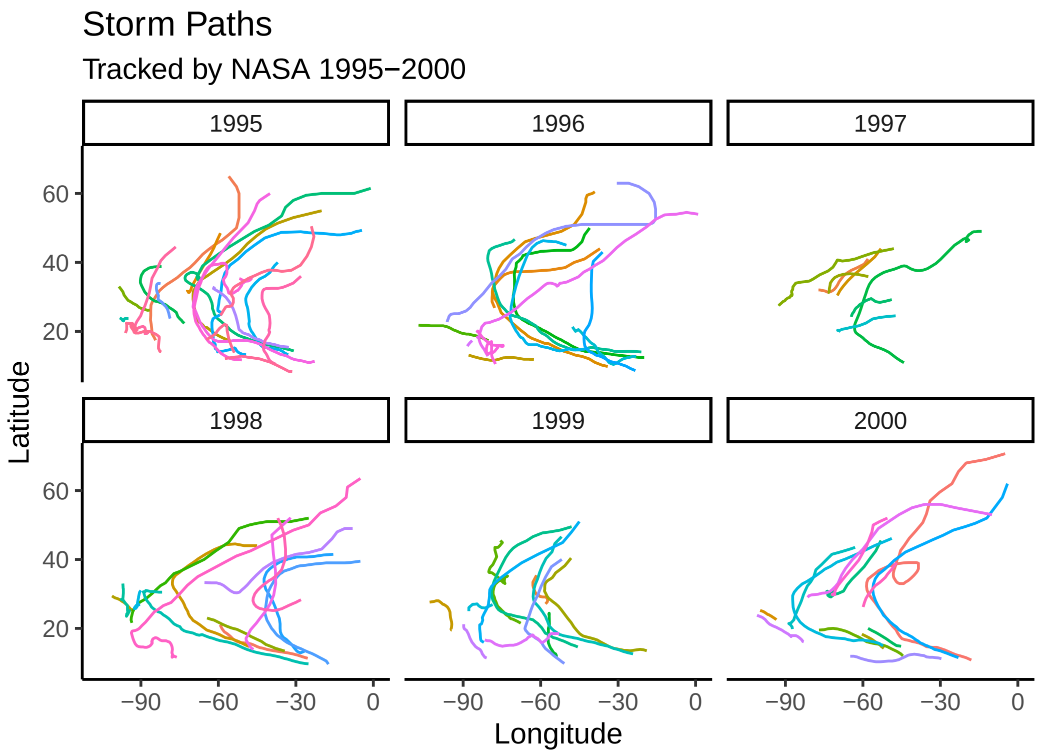
```
baseballdata = MLB_teams

ggplot(data = baseballdata, mapping = aes(x = payroll/1000000, y = W/(W+L))) +
  geom_point() +
  geom_smooth(method = "loess" , se = FALSE) +
  scale_x_continuous(n.breaks = 10) +
  labs(title = "MLB Payroll and Win%",
       subtitle = "Major League Baseball teams from 2008-2014",
       x = "Payroll in million dollars",
       y = "Win%")
```



The graph shows an obvious correlation of payroll and win percentage. However the data around the median payroll and win% sees much less correlation. Only when the payroll becomes exceeding below or above average does it reflect heavily on results.

Problem 4 **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")`.

```r
ggplot(data = nasaweather::storms, mapping = aes(x = long, y = lat)) +
        geom_path(aes(color = name)) +
        scale_color_discrete(guide = "none") +
        facet_wrap(~year, nrow = 2) +
        labs(title = "Storm Paths",
             subtitle = "Tracked by NASA 1995-2000",
             x = "Longitude",
             y = "Latitude")
```


Storm Paths
Tracked by NASA 1995–2000
Latitude
Longitude