Practice Set 1

Kauila Tengan

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)

- Mahathi Athreya and Caroline Tilton - problem 1.3; Mahathi - problem 4 code

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

- Textbook section 2.2 for question 1

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 Major-Career graphic tells us a few things. On one hand, it shows what all of the alumns majored in. It also shows how many students majored in those respective majors. Next, the graph details what career fields the Williams College students entered. To take it one step further, each career also showed which and how many majors entered that field. On the left side, it shows us that economics, english, history, poli sci, psych, art and biology have always been very popular majors (many students majored in those fields). It also shows that law, banking, health, and higher education were popular career paths for students. If we hover over the labels on the right, the thickness of the line indicates how many students from each respective field entered that career.

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.

This is not a cartesian plane coordinate system. However, the graphic has clear labels of each of the majors and occupations. The color corresponds to the categorical variable `major`. The size (perhaps a combination of area and volume) both shows the amount of total majors and occupations (the length of the cutout of the circumfrence of the circle); but also, the size of area of the corresponding lines also shows how many of the major went to each occupation. The direction simply connects the major to the career.

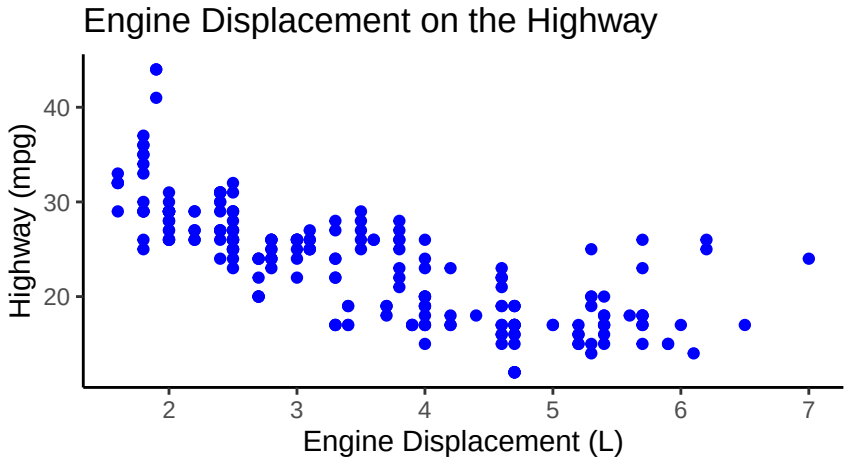
1.3 Critique and/or praise the visualization choices made by the designer. Do they work? Are they misleading? Thought-provoking? Brilliant? Are there things that you would have done differently? Justify your response.

Hovering over the jobs individually, I think this visualization does a good job of simply showing the audience the distribution of each college major at Williams entering a job. It does not tell us a specific number, but considering there are over 15,000 observations, I like the simplicity. Rather, we can gage how many of what major have entered these jobs. However, things do get confusing when looking at the compilation visualization. The shades of red are fairly similar so it can be confusing to look at first glance. The other critique I have is how they mapped out the double majors. The caption says they have two strands (one coming from each major). Perhaps this could be misleading. We do not know exactly how many double majors there were nor do we know which majors they doubled in. Perhaps this could skew the thickness of the lines if these individuals are being counted twice (once for each major). Overall, I think this was a solid attempt at conveying A LOT of information in a significant manner. If this were done on a cartesian plot, we would be looking at 15 different bar charts of each major which may be overwhelming.

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"))

ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), color = "blue") +
  labs(title = "Engine Displacement on the Highway",
       y = "Highway (mpg)",
       x = "Engine Displacement (L)",
       color = "Blue")
```

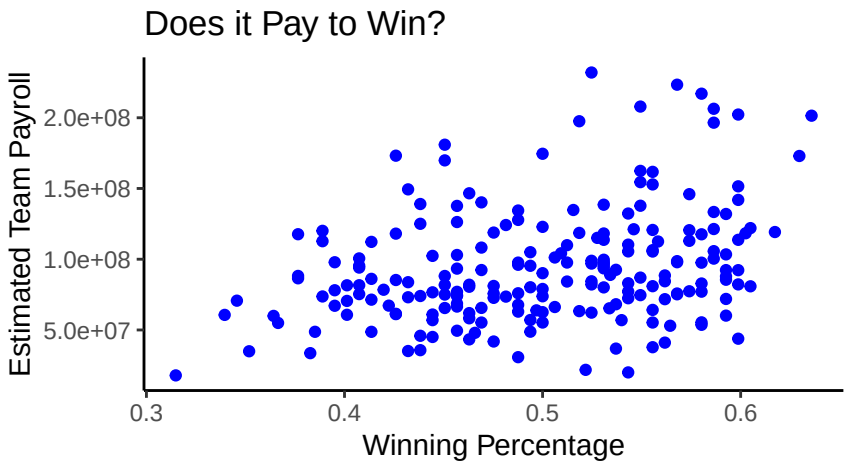


Problem 3 **MDSR Exercise 3.6 (modified)** Use the `MLB_teams` data in the **mdsr** package to create an informative data graphic that illustrates the relationship between winning percentage and payroll in context. What story does your graph tell?

Overall, the graph details a positive, linear relationship between the estimated sum of players' salaries on a team and the team's win percentage. I have some hesitancy with this graphic as it is not a uniform distribution and there is a noticable spread between the concentration of points on the lower half of the graph and the dispersed points in the top right. The points fan out instead of being tight from the lower left to the upper right. However, a majority of the highest payroll does fall between a win percentage of 50 and 60%. And the lowest payroll is associated with the lowest win percentage.

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

ggplot(data = MLB_teams) +
  geom_point(mapping = aes(x = WPct, y = payroll), color = "blue") +
  labs(title = "Does it Pay to Win?",
       y = "Estimated Team Payroll",
       x = "Winning Percentage")
```



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")`.

```
g <- filter(storms, type == "Tropical Storm")
ggplot(data = g) +
  geom_path(mapping = aes(x = long, y = lat, colour = name)) +
  labs(title = "Path of Tropical Storms",
       x = "Longitude",
       y = "Latitude") +
  facet_wrap(~year, nrow = 5) +
  scale_color_discrete(guide = "none")
```