Practice1

Sebatian Montesinos

Due by midnight, Friday, Feb. 18

Practice1

Reminder: Practice assignments may be completed working with other individuals.

For academic integrity, the second page of this assignment provides a place where you will list who you worked with (and for which problems) as well as outside resources that you used (meaning resources besides our textbook, course materials, and default R help for functions/packages.) The problems begin after that. (Next week, this page will be the cover page and the workflow review will be removed.)

Reading

The associated reading for the week is Chapter 2, Chapter 3, and Section 8.2.

Git Workflow Review

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.
5. You should occasionally *push* the updated version of the .Rmd file back onto GitHub. When you are ready to push, you can click on the Git pane and then click **Push**. You can also do this after each commit in RStudio by clicking **Push** in the top right of the *Commit* pop-up window.
6. When you think you are done with the assignment, save the pdf as "*YourFirstInitialYourLastName_thisfilename*.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). For example, I would save this file as AWagaman_Practice1.pdf.

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook, course materials in the repo, labs, R help menu) 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)

-

1 - Understanding the Taxonomy

Steer your web browsers to the John Hopkins University COVID Dashboard's US map. The map is the primary graphic on the page and is our focus for this problem.

> part a - What story does the data graphic tell? What is your main take away message from it?

Solution: This graphic displays the relative distribution of covid cases across the country. The primary takeaway is that covid is distributed everywhere but there are hotspots where the ceoncetration is particularly high in the south/mideast area of the US.

> part b - If possible, describe the data graphic in terms of Yau's taxonomy. That is, list the visual cues, coordinate system, scales, and context. If any features do not fit in one of these four categories, list them separately.

Solution:The coordinate system is geographic, meaning it displays the data using a MAP of the US divided into counties. The primary visual cue is color/shade, which draws one's attention to the relative amount of covid deaths in a location. The other visual cue is position/location, since you can be naturally drawn to areas in the US with less or more covid cases. The scale is numeric, in that it shows the proportion of covid cases per 100,000 people in a county. Finally, the context of the graphic includes the title, the detailed description under the graphic explaining the source and purpose of the graph, and the information on the right and left hand sides indicating the places in the US with the most covid.

> part c - Critique the display. Are there aspects of the visualization you would praise? Are there aspects you would change if given the chance? Is there anything misleading you would want to fix if this was your graphic? Justify your response.

Solution: The graph pretty clearly shows the general areas in the US struggling most with a high proportion of covid cases. The shading is fairly easy to follow and read. I think it is also helpful that the specific areas with the highest amount of covid cases are listed in a sidebar, since we cannot tell which counties are worse off just by looking at the shade corresponding to the highgest rate of covid cases. Two changes that could help one get an even better sense of the daat would be to have a dot for each state that scales in size for the proportion of cases per state, that would appear on that state. This could help one get an even better sense of the more general amount of covid in wider areas than counties.

2 - Enhancing a plot

The *mpg* data set is available in R (use the help file to learn more about it). We are interested in examining the relationship between the variables *hwy* and *displ* across the variable *drv*.

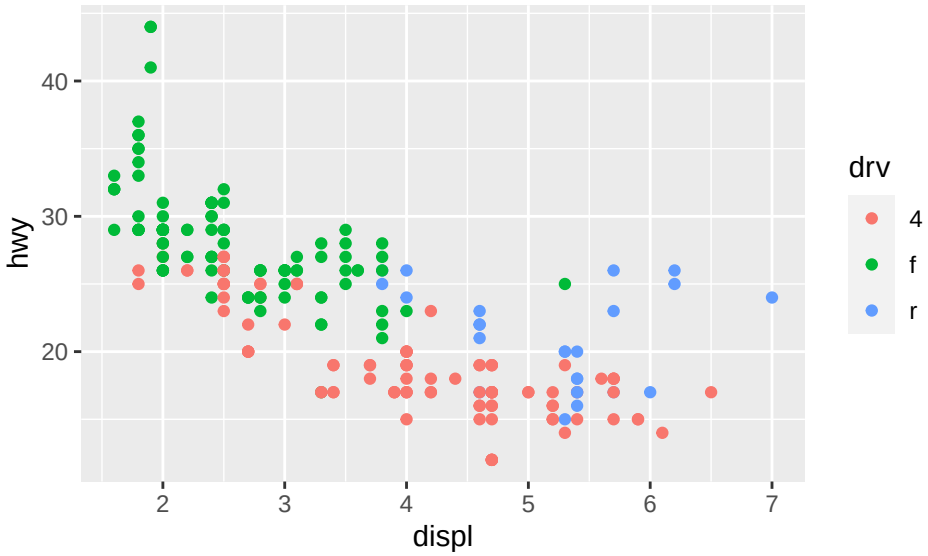
A preliminary plot is provided which needs enhanced. Add new code chunks to make new plots as described below.

```
# preliminary plot, eval = FALSE means plot will not show in pdf
g <- ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point()
g
```

part a - Add *drv* to the preliminary plot using color or size.

Solution:

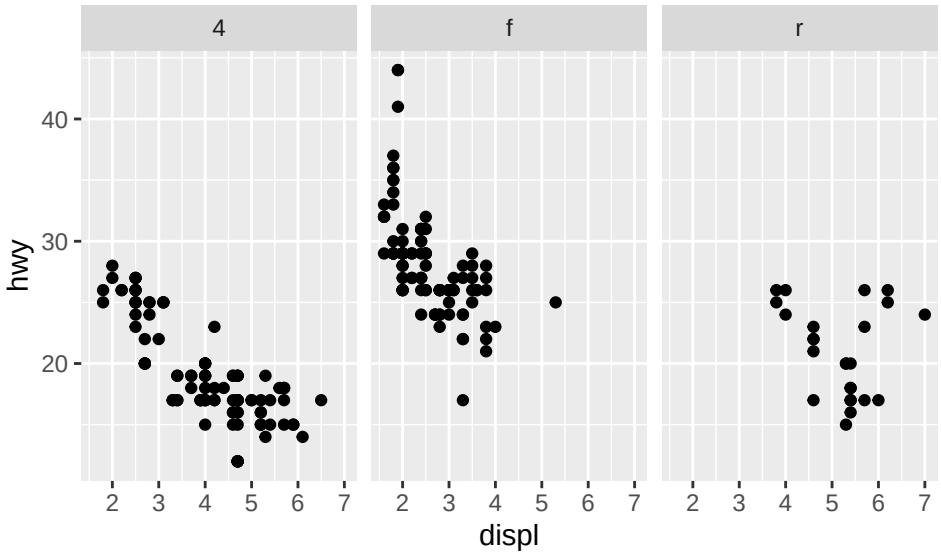
```
g <- ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point()

g
```



part b - Use facets to incorporate *drv* to the preliminary plot instead of using either color or size.

Solution:

```
g <- ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  facet_wrap(~drv, nrow = 1)
g
```



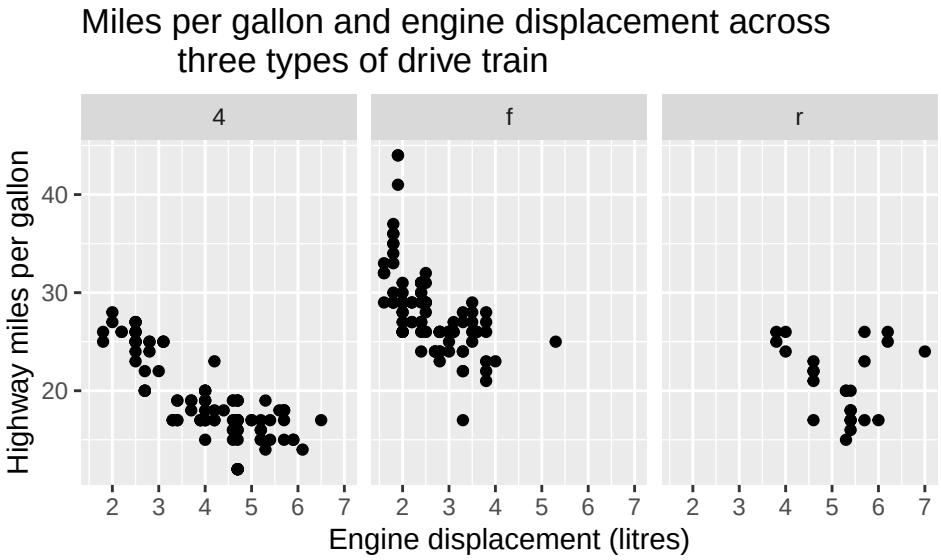
part c - Which graphic do you prefer - the one from part a or part b - for exploring the relationship between these 3 variables? Justify your response.

Solution: I prefer the facet solution because it allows me to more clearly see where the data for each DRV variable is grouped and how they differ. The colors help a little with this but I prefer to see them as independent groups.

part d - Improve your preferred plot by adding a title and making more appropriate axis labels.

Solution:

```
g <- ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  facet_wrap(~drv, nrow = 1) +
  ylab("Highway miles per gallon") +
  xlab("Engine displacement (litres)") +
  ggtitle("Miles per gallon and engine displacement across
          three types of drive train")

g
```


Miles per gallon and engine displacement across three types of drive train
4
f
r
Highway miles per gallon
40
30
20
2
3
4
5
6
7
Engine displacement (litres)

3 - Baseball (Based on MDSR 3.5)

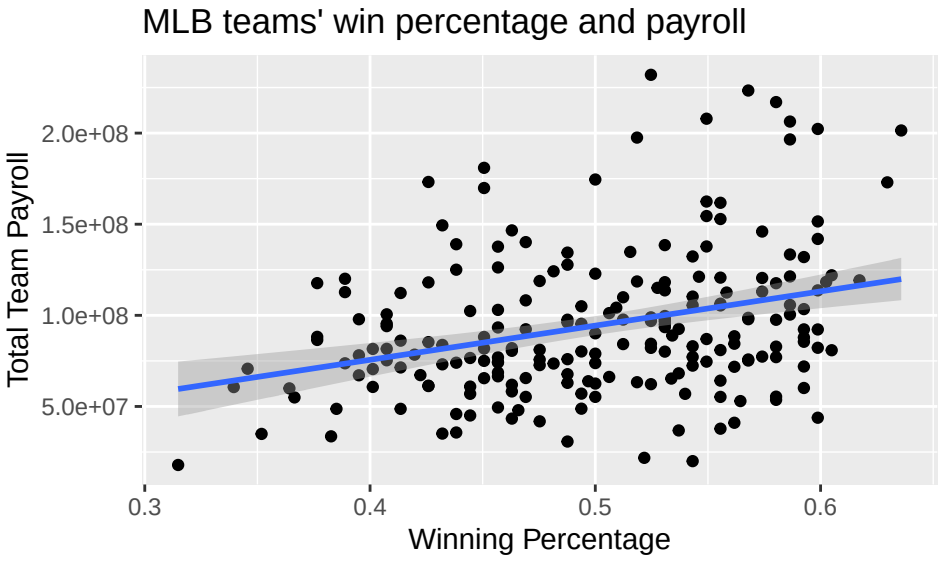
We want to explore the relationship between winning percentage and payroll in context using the *MLB_teams* data in the *mdsr* package.

> part a - Create an informative data graphic that illustrates the relationship between these 2 variables. Be sure your graphic has appropriate labels and a title (i.e. it has context).

Solution:

```
p <- ggplot(data = MLB_teams, mapping = aes(x = WPct, y = payroll)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ggtitle("MLB teams' win percentage and payroll") +
  ylab("Total Team Payroll") +
  xlab("Winning Percentage")
p
```

```
## `geom_smooth()` using formula 'y ~ x'
```



> part b - Now, add a third variable to your plot, making sure to update titles, labels, etc. as needed.

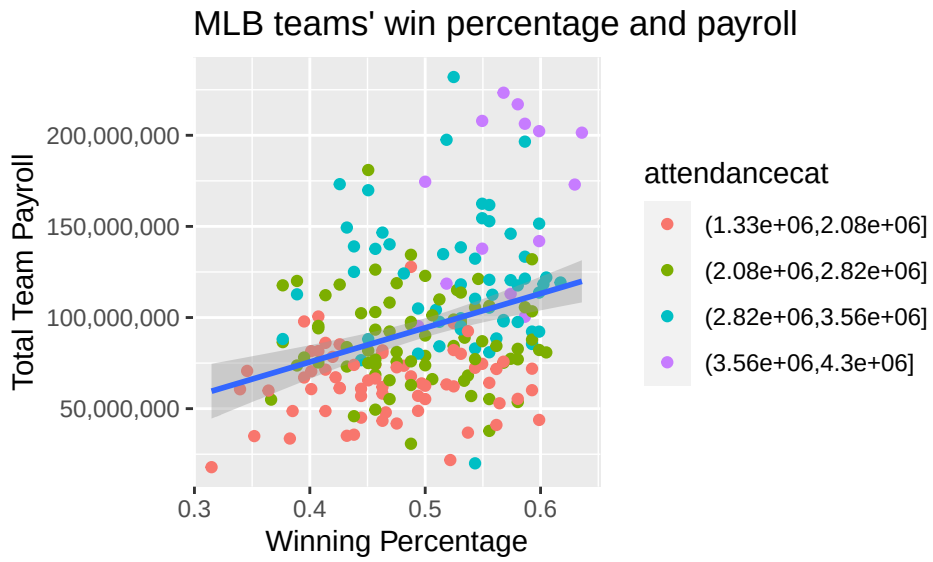
Use the help file to choose a variable that you think will be interesting to add. (You may play around with this, of course.)

Solution:

```
MLB_teams <- MLB_teams %>%
  mutate(attendancecat = factor(cut(attendance, 4)))
p <- ggplot(data = MLB_teams, mapping = aes(x = WPct, y = payroll))+
  geom_point(aes(color = attendancecat)) +
```

```
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::comma) +
  ggtitle("MLB teams' win percentage and payroll") +
  ylab("Total Team Payroll") +
  xlab("Winning Percentage")
p
```

```
## `geom_smooth()` using formula 'y ~ x'
```



part c - What story does your graph from part b tell?

Solution: Attendance appears to be a modulating factor in payroll in addition to winning percentage, since teams with lower attendance cluster at a low team payroll while teams with very high attendance cluster at a high team payroll. Thus, both winning percentage and attendance appear to be related to team payroll.

4 - Storms (Based on MDSR 3.8)

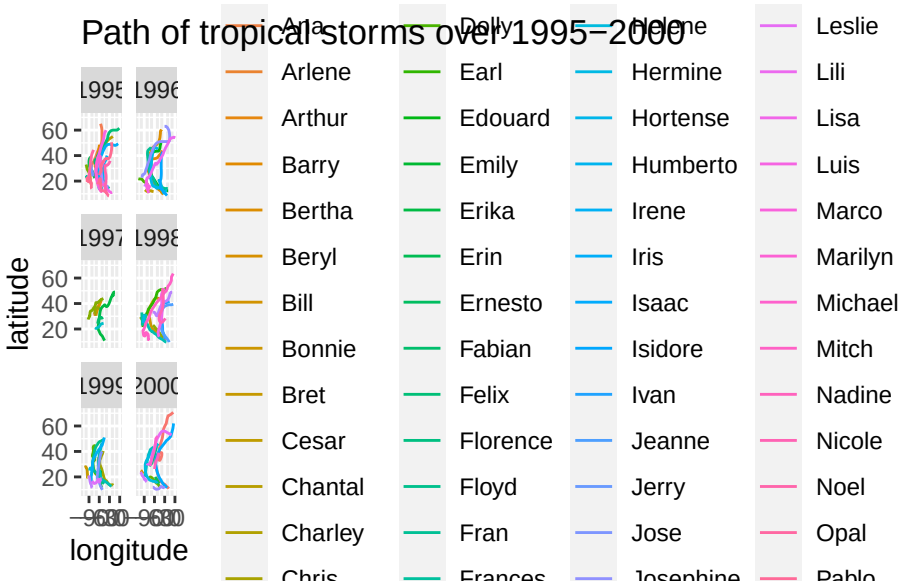
MDSR 3.8 reads: "Using data from the *nasaweather* package, use the *geom_path()* function to plot the path of each tropical storm in the *storms* data table . Use color to distinguish the storms from one another, and use faceting to plot each year in its own panel."

part a - Complete MDSR 3.8

Hint: latitude should be your y-axis and longitude should be your x-axis.

Solution:

```
v <- ggplot(data = storms, mapping = aes(x = long, y = lat, color = name)) +
  geom_path() +
  facet_wrap(~year, nrow = 5) +
  ggtitle("Path of tropical storms over 1995-2000") +
  ylab("latitude") +
  xlab("longitude")
v
```



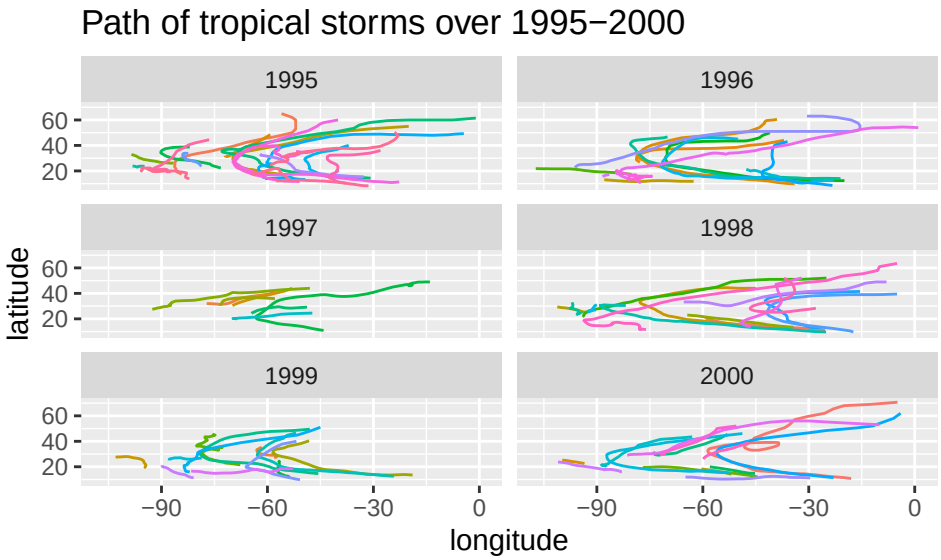
part b - How useful do you find the legend of storm names and colors? If your overall goal was to just look for common paths, would you need the names?

Solution: There are so many different storms names associated with so many different shades of color that it is not useful at all. It would take way too long and be too hard to associate a particular storm name with a specific color/spot on the graph. One can easily assess common paths without this convoluted legend.

part c - Remove the legend of storm names/colors by adding *scale_color_discrete(guide = "none")*. Be sure your final graph has appropriate labels and a title.

Solution:

```
v <- v + scale_color_discrete(guide = "none")
v
```



5 - Metabolic Rate

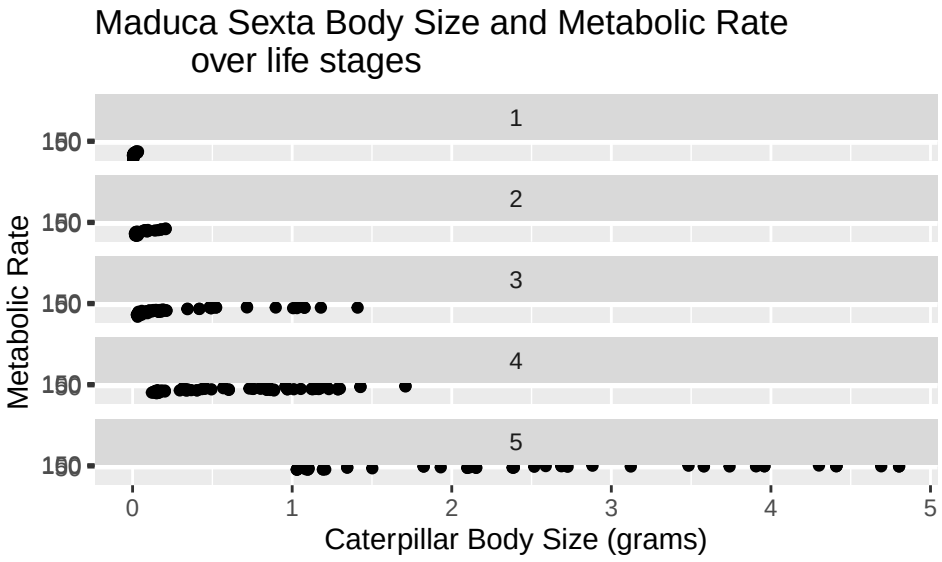
The *MetabolicRate* data set from the *Stat2Data* package contains measurements of body size and metabolic rates for Manduca Sexta caterpillars. We want to model the relationship between these variables using a regression (with body size as the explanatory variable). However, there is some concern that the relationship between the original variables is not linear. (You can investigate this yourself.)

Create an informative data graphic that shows the relationship between these two variables - *BodySize* and *Mrate*. Use appropriate changes to scale to identify a relationship that regression modeling seems appropriate for (you do not need to add the regression line). Then, add the variable *Instar* to the plot in some way. Be sure your final plot has a title, appropriate labels, and a legend (as appropriate). Finally, describe what your graphic reveals in a few sentences.

Hint: The help menu for the data set will describe what the variables are and may help you identify potential scales to use. However, do not change the variables in the plot - use the variables above.

Solution: There is an logarithmic relationship between body size and metabolic rate, wherein metabolic rate increases greatly with body size from 0 to 1 grams but slowly levels out after that. The life stage of the caterpillar also seems to effect body size, as later life stages are associated with greater body sizes and metabolic rates.

```
z <- ggplot(data = MetabolicRate, mapping = aes(x = BodySize, y = Mrate)) +
  geom_point() +
  coord_trans(y = "log10") +
  facet_wrap(~Instar, nrow = 6) +
  ggtitle("Maduca Sexta Body Size and Metabolic Rate
          over life stages") +
  ylab("Metabolic Rate") +
  xlab("Caterpillar Body Size (grams)")
z
```



Going forward, every plot created for our assignments should have clear context in terms of good labels and a title. Get in the practice of using labs().