Homework 2

Weeks 4 and 5

```
library(tidyverse)

# Load up the mpg dataset (from the ggplot2 package)
data("mpg")
```

1. Create a basic plot comparing the miles per gallon by different driving location (`cty` or `hwy`). To do this, you'll need to transform those two columns to be in a key-value pair using `gather`. What do you see in this plot/what should be the main takeaways?
2. Take the plot you created in 1 and make it publication ready however you see fit (scale, labels, color, theme, etc.).
3. Create another plot on your own using the `mpg` data set. Explain why you chose to create the plot that you did, why you chose the variables you did, and why you think it is an important relationship to look at. Explain what you see in your plot.
4. Calculate each of the following and tell me what we can take away from each statistic:

- Count of `drv`
- Quartiles of `hwy`
- Mean and median of `cty`

Weeks 6 and 7

5. Take a look at the `mpg` data set. If we were to predict `hwy` using a linear regression model, what do you think would be good to use as predictors? Use any pre-analysis steps or general knowledge of the data set to support your ideas.
6. Using the `mpg` data set, build a model using `displ`, `drv`, and `class` to predict `hwy`. Explain your output from a **practical** perspective.
7. Check regression assumptions; explain why each assumption is met or not.

```
# Load up the geyser data set in the MASS package
library(MASS)
data("geyser")
```

8. Perform some pre-analysis on the geyser data set; explain what you see.
9. Build a simple linear regression model predicting `duration` by `waiting`. Explain output and assumptions.
10. Build a polynomial model predicting the same thing as 9. You choose the degree. Explain your choices and output. Compare this model with the one you built in 9. Which do you think is better?