Data Wrangling

This lab is a bit different from the others. We'll be working through the exercises together as a class, but be sure to record them in your lab report like normal. You can find these exercises in the slides, which are posted here. The On Your Own component will be as usual, on your own (or in pairs). To be sure you have all the necessary packages loaded, be sure to add the following chunk at the top of your lab report.

```
knitr::opts_chunk$set(message = FALSE, warning = FALSE)
library(dplyr)
library(ggplot2)
library(oilabs)
library(pnwflights14)
```

On your own

For the following questions, it may be helpful to sketch out what you want the final data set to look like (what are the rows? what are the columns?), then work backwards to figure out how to create that dataset from the original data.

- The plot below displays the relationship between the mean arrival delay and the mean distance travelled by every plane in the data set. It also shows the total number of flights made by each plane by the size of the plotted circle. Please form the chain necessary to create the dataset that lies underneath this plot. You will also want to exclude the edge cases from your analysis, so focus on the planes that have logged more than 20 flights and flown an average distance of less than 2000 miles.

- Once you have that dataset, you can check your answer by using the following code to generate the plot. What does the relationship appear to be between mean distance that planes fly and their mean arrival delay?

```
ggplot(delay, aes(dist, delay)) +
 geom_point(aes(size = count), alpha = 1/2) +
 geom_smooth() +
 scale_size_area()
```

- Say you live halfway between Portland and Seattle. Based on the 2014 data, which airport should you fly out of to minimize overall delays? (however you define them)
- Say you fly very often between PDX airport in Portland and JFK airport in New York. Which day of the week would you recommend that I buy a ticket for if I want to minimize delays? Any tips for the best airlines to go through?

Challenge: Remake two plots that show up in the slides - the points plot of average delays by carrier and the bar chart of total count of flights by carrier out of PDX - but display the carriers in descending order.