HW 4, STAT 450
Due: Monday, November 8

Reading: Sections 13.1 - 13.4 from *R for Data Science*

```
library(tidyverse)
library(nycflights13)
```

Exercise 1. Consider the following data frames:

```
facts <- tibble(
  state_abbr = c("CA", "CA", "OR", "OR", "TX"),
  year = c(2019, 2010, 2019, 2010, 2019),
  population = c(39.5, 37.2, 4.2, 3.8, 29.0)
)
state <- tibble(
  state_abbr = c("CA", "OR"),
  state_name = c("California", "Oregon")
)
```

(a) What is the *key* that relates the two data frames?

(b) Try to predict the output of the following code. Then run the code in R to check if your prediction was correct.

```
inner_join(facts, state, by = "state_abbr")
left_join(facts, state, by = "state_abbr")
```

Exercise 2. Use `group_by()` and `summarise()` to compute the mean arrival delay for each flight destination. Then join that data frame with the grouped summaries with the `airports` data frame, which contains information about each airport. This is what the resulting data frame should look like after the join:

```
## # A tibble: 105 x 10
##    dest  count arr_delay_mean name           lat    lon   alt    tz dst   tzone 
##    <chr> <int>          <dbl> <chr>        <dbl>  <dbl> <dbl> <dbl> <chr> <chr> 
##  1 ABQ     254           4.38 Albuquerqu~   35.0 -107.   5355    -7 A     Americ~
##  2 ACK     265           4.85 Nantucket ~   41.3  -70.1    48    -5 A     Americ~
##  3 ALB     439          14.4  Albany Intl   42.7  -73.8   285    -5 A     Americ~
##  4 ANC       8          -2.5  Ted Steven~   61.2 -150.    152    -9 A     Americ~
##  5 ATL   17215          11.3  Hartsfield~   33.6  -84.4  1026    -5 A     Americ~
##  6 AUS    2439           6.02 Austin Ber~   30.2  -97.7   542    -6 A     Americ~
##  7 AVL     275           8.00 Asheville ~   35.4  -82.5  2165    -5 A     Americ~
##  8 BDL     443           7.05 Bradley In~   41.9  -72.7   173    -5 A     Americ~
##  9 BGR     375           8.03 Bangor Intl   44.8  -68.8   192    -5 A     Americ~
## 10 BHM     297          16.9  Birmingham~   33.6  -86.8   644    -6 A     Americ~
## # ... with 95 more rows
```

Bonus: Use the data frame from Exercise 2 to visualize the spatial distribution of arrival delays. Here's some code to draw a map of the United States:

```r
library(maps)
library(mapproj)
states <- map_data("state")
ggplot() +
  geom_polygon(data = states, aes(x = long, y = lat, group = group),
               fill = "white", color = "black") +
  coord_map()
```

On this map, plot the coordinates (longitude and latitude) of each destination airport. The use the `color` of the points to display the average delay time for each airport.[1] You might also what to use `filter()` to remove the airports located in Alaska and Hawaii.

[1] I recommend using the `viridis` color scale: https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html