Lesson 9 Statistics!: Correlations and simple linear regression

Goals for Today

- Today we're going to be talking about correlations and linear regression

Readings for Today

Required

- OpenIntro Statistics, **All of Chapters 8**
- DataCamp, **Regression in R**
- Peng, **Art of Data Science - Chapter 5**

Optional

- If you're a little uncomfortable with math (and that's okay!), some of these YouTube videos could help as refreshers:
  - Khan Academy - Correlation and Causation
  - Khan Academy - Regression Example
  - Linear regression in R - with Examples

Important Links and Files

- The file we worked on in class is available here

Written Assignment #2

Details

- **Due Date**:
  - 2021-10-15 (**this Friday!**) prior to the start of class (i.e., needs to be submitted before 1PM EST).
- **Format**:
  - You'll submit this via Google Classroom. Again, it can be in whatever format you like - so long as it uploads! Take care if you're trying to use a Google Doc, however, as the copy/paste function of code seems to be a little funky.
- **Working Style**
  - You can do this individually or as a group - it is entirely up to you! If you do work in groups, please make note in the document that you did so and list everyone's name.

Data

Everyone complains about bad drivers - but which state has the *worst* bad drivers? Let's explore that question and see if we can find any interesting relationships.

The data and data dictionary/context are available here. You can download the data directly with these commands:

```
library(tidyverse)
data <- read_csv(url("https://raw.githubusercontent.com/fivethirtyeight/data/master
```

Note: for full credit, you *must* provide the functions you used to obtain your answers!

Questions

Answer these as thoroughly as you can and please provide the code that you've used to generate your answer.

1. Which state(s) has the highest number of drivers involved in fatal collisions per billion miles?

2. What is the state-level average percentage of drivers involved in fatal collisions who were alcohol-impaired? (Note: remember which of these data elements are US states and which ones are not!)

3. Identify if a given state is above or below the median car insurance premiums from this data set. Now test, using a two-sided t-test, if the average number of drivers involved in fatal collisions per billion miles is different between these two groups. Use a confidence threshold of 90%. Answer the question by referring back to the hypothesis you're testing.

4. Compare the average percentage of drivers involved in fatal collisions while speeding between states on the West Coast (including Alaska and Hawaii) to every state on the East Coast (Florida up through Maine) using a t-test. Test if the East Coast drivers have a *lower* average percentage than West Coast (hint: use a one-sided test!) using a 95% confidence threshold.

5. Is there evidence of a statistically significant correlation between the percentage of drivers involved in fatal collisions who were alcohol impaired and the number of drivers involved in fatal collisions per billion miles? Use a confidence threshold of 90%

6. Examine all of the possible correlations between the numeric variables and reported which variables, if any, are statistically correlated. Make sure you report the correlation coefficient for each pair of variable and that you're using a confidence threshold of 95% each time.