Homework 5

For this homework you will create an R Markdown document that outputs to a PDF and you will upload both the .Rmd and .pdf files to wolfware. Be sure to put your name in the title of the document.

The purpose of this homework is to get practice reading in raw data.

To Do - Write a markdown file that goes through each step below. For problems requiring an explanation, just answer with text. Some data files are available on moodle in the assignment link.

General Questions

1. What is a delimiter? What is the most common delimiter that statisticians use?
2. What is an R package? What is the difference between `library()` and `require()`?

Deaths Data

1) Install and load the `readxl` package.
2) The `readxl` package has a few datasets built-in to practice reading in `xls` and `xlsx` data. Use the `readxl_example` function to print out the names of these datasets.

   The full path for each file can be found by using the filename inside the `readxl_example` function. We are interested in the `deaths.xlsx` dataset, which lists some information about famous historical deaths.

   Locate the dataset and open it in an Excel editor. How many tabs does it have? What problems do you see with it?
3) Using the full file path string, read in the first sheet of the famous deaths dataset without any additional arguments. What does this data look like?
4) Add appropriate arguments to your `read_excel` function to account for telling it which line to start reading and how many rows to read in. Then read in the first sheet again and store it as `deaths1`.
5) Read in the second sheet of the deaths data, adjusting the necessary arguments, and call it `deaths2`. Use `rbind` to combine the data to both sheets into an object called `all_deaths`.

Education Data

1) Download the `censusEd.xlsx` dataset from the assignment folder. This data is related to education enrollment and attendance across the United States. View it in an excel editor, and read in the first sheet.
2) The rows of edData that have Area_name equal to a state name (e.g. ALABAMA) represent the totals for that state. Using the `toupper` function, the built-in object `state.name`, and the `%in%` operator, filter the rows of edData to the totals for each state.
3) The column names that end in "D" represent yearly aggregates of enrollment in primary and secondary schools. Reduce edData by selecting only the state names and the columns that end in "D", call this `edDataD`.

4) This data now contains enrollment statistics for 1987-1996. The years are encoded in the column names - EDU0101**87**D = 19**87**. Select only the Area_name and 1996 columns, and arrange this by decreasing enrollment in 1996. Print the result to the console; what are the top 3 states in terms of number of students? Does this seem reasonable?

Scores Data

1) Download the scoresFull csv and put it in your `data` folder. Then read it in using `read_csv`.

Religion Data

1) Download the 'relig_income.txt' and 'relig_income.csv' files and put them in the `raw_data` folder. Open the .txt file; what kind of delimiter does it appear to be using?

2) Read in the .txt file using the appropriate function from the `tidyverse`. Store it as `relig` and print it to the console. Does it seem correct to you? If not, why?

3) Now read in the .csv file as `relig2` and print it to the console. Does this seem correctly formatted?

4) Use the `rename` function to change the variable `<$10k` to `$0-10k`.

5) This data is currently formatted in 'wide' format, which lists observations as rows and columns as variables. Use the `gather` function from the `tidyr` package to convert from wide to long format. Write code to print this data to the console.

Finished!