Homework 5

For this homework you will create a github repo, set up github pages, clone the repo to your computer as an R project, create a `.qmd` file, and push those changes back to github to create a webpage! You'll submit the link to your github pages site (the one that looks like a nice website) and the github repo site.

The steps for setting things up exist in the first homework assignment and are not repeated here.

- Create a new `.qmd` document that outputs to HTML. You can give this a title of your choosing. Save the file in the main repo folder.
- In this document, answer the questions below. **Use tidyverse functions and manipulations for most aspects of this homework.**

This homework is meant to give you a chance to do some structured practice with the EDA material.

Summarizing Student Data

For this part, we'll use data that comes from the UCI machine learning repository. The data is about secondary education in two Portuguese schools. Information about the variables in the dataset can be found here. To download the data, head to this site and click the big blue button in the top right. When downloading the data I had to do two unzips of files... they actually provide code for reading in and merging both data files there in a `.R` file. We'll use this shortly.

Create an R project and place the data files into that directory.

You should read up on the variables. The dataset is generally about math and Portuguese scores (G1, G2, G3) for students from two different schools. They also measure a bunch of things about the students' home life.

Task 1: Read in the Data and Modify

We'll read in the data in two ways:

- First, modify the code provided with the download to read in the data from a local file source (your downloaded `.csv` files) and combine the two data frames. Use local paths as they do in their code.
- Second, read in and combine the data using functions from the `tidyverse`. Use an `inner_join()` on the variables they used in their code. Do you notice any issues? Make a note of the issue.
- Use an `inner_join()` on all variables other than `G1`, `G2`, `G3`, `paid`, and `absences`. Use this form of the combined data in further exercises.
- Next, for the math data, Portuguese, and combined data, choose four categorical variables you are interested in and convert those into `factor` variables in each tibble (use the same four variables in each). Use the `mutate()` function to accomplish this.

Task 2: Summarize the Data (Very Basic EDA)

We've talked about the general process of conducting an EDA. You try to understand how your data is stored, what is missing, and you try to summarize the variables both numerically and visually to understand relationships within the data.

Do the rest of these items on the **combined data**:

- Look at how the data is stored and see if everything makes sense.
- Document the missing values in the data.

Categorical variables

- Create a one-way contingency table, a two-way contingency table, and a three-way contingency table for some of the factor variables you created previously. Use `table()` to accomplish this.
  - Interpret a number from each resulting table (that is, pick out a value produced and explain what that value means.)
- Create a conditional two-way table using `table()`. That is, condition on one variable's setting and create a two-way table. Do this using two different methods:
  - Once, by subsetting the data (say with `filter()`) and then creating the two-way table
  - Once, by creating a three-way table and subsetting it
- Create a two-way contingency table using `group_by()` and `summarize()` from `dplyr`. Then use `pivot_wider()` to make the result look more like the output from `table()`.
- Create a stacked bar graph and a side-by-side bar graph. Give relevant `x` and `y` labels, and a title for the plots.

Numeric variables (and across groups)

The numeric variables are age, absences, and the three test grades variables (G1, G2, and G3) from each data set (math and Portuguese).

- Find measures of center and spread for three of these variables (including at least one G3 variable)
  - Repeat while subsetting the data in a meaningful way.
- Find measures of center and spread across a single grouping variable for three of these variables (including a G3 variable as one of them)
- Find measures of center and spread across two grouping variables for three of these variables (including a G3 variable as one of them)
- Create a correlation matrix between all of the numeric variables
- Create a histogram, kernel density plot, and boxplot for two of the numeric variables across one of the categorical variables (that is, create graphs that can compare the distributions across the groups on the same plot (no *faceting* here)). Add appropriate labels and titles.
- Create two scatterplots relating a G3 variable to other numeric variables (put G3 on the y-axis). You should jitter the points if they sit on top of each other. Color the points by a categorical variable in each. Add appropriate labels and titles.
- Repeat the scatter plot step but use faceting to obtain graphs at each setting of another categorical variable.
- Repeat the scatter plot step but use faceting to obtain graphs at each combination of two categorical variables.

After each summary or graph, you should discuss what is interesting about it or what it tells you!