# Homework 3

## Instructions

For this homework you will create a python notebook (`.ipynb` file) using Google Colab and submit a link to your notebook. This can be done by clicking on the 'share' icon in the top right. Please make sure to change the sharing settings so that anyone with the link can see the notebook. No edits should be made after the due date!

The purpose of this homework is to practice with `pandas` and doing some basic numerical summaries that would be part of a larger EDA. Most homework assignments will have a part that pushes you beyond what was in the lectures! Learning to search for the right questions and browsing stackoverflow are really life skills that we should hone :)

### Reading and Manipulating Data with `pandas`

1. This question doesn't involve any programming. Suppose you have a pandas `DataFrame`. In a markdown cell, describe what each of the following operators/methods allows you to subset from a data frame (that is, describe all of the types of subsetting you can do with the operator):

   - `[]`
   - `.iloc[]`
   - `.loc[]`

2. Read in the BreastCancer.dat data file available in the assignment link. (Open the file in a program such as notepad or wordpad to determine the delimiter - although a program like notepad++ is a better choice.) Upload this file to your Colab notebook and read it in using a relative file path (just the file name).

   a. Save the data as an object called `cancer_data`.
   b. Use the `.head()` method to look at the data.
   c. Return just the `grade` column using the column *attribute*
   d. Use the `.loc[]` method to print out all rows where the `size` is larger than 30.
   e. Use the `.loc[]` method to print out all rows where the `size` is greater than 30 and the `grade` is 3.
   f. Use `[]` to return just the `age`, `size`, and `grade` columns.
   g. Use `.loc[]` to return the rows where `meno` is equal to `premenopausal` along with the `age`, `size`, and `grade` columns.

3. There are two files about mosquitos available at:

   - https://www4.stat.ncsu.edu/~online/datasets/mosquito.txt
   - https://www4.stat.ncsu.edu/~online/datasets/mosquito2.txt

   a. Determine the delimiter and read in the mosquito.txt file as an object called `mosq_data`.
   b. Similarly, read in the mosquito2.txt file. Note this file doesn't contain column names! The columns are the same as the other file though. Use an attribute from `mosq_data` to assign the column names as you read in the data. Save this data as an object called `mosq_data2`.
   c. Combine the two datasets into one data frame using the `concat()` function from `pandas` (see https://pandas.pydata.org/docs/reference/api/pandas.concat.html).

## Summarizing Data Numerically

For this part, we'll use the StudentData.txt data that comes from the UCI machine learning repository. Information about the variables in the dataset can be found here. I want you to look at the math scores data set.

You should read up on the variables. The dataset is generally about math scores (G1, G2, G3) for students from two different schools. They also measure a bunch of things about the students' home life. Hopefully you can make some interesting connections!

### Task 1: Read in the data

- You can either read this dataset from the URL or download it and read it in locally. Check out the first few observations of the data.

### Task 2: Summarize the Data

This data has many categorical variables and a few numeric. You should do the following:

**Categorical variables**

- Create a one-way contingency table, a two-way contingency table, and a three-way contingency table
  - 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. 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 `.loc`) and then creating the two-way table
  - Once, by creating a three-way table and subsetting it

**Numeric variables (and across groups)**

The numeric variables are age, absences, and the three test grades variables (G1, G2, and G3).

- Find measures of center and spread for three of these variables (including G3 as one of them)
  - Repeat while subsetting the data by some grouping variable (say with `.loc`)
- Find measures of center and spread across a single grouping variable for three of these variables (including G3 as one of them)
- Find measures of center and spread across two grouping variables for three of these variables (including G3 as one of them)
- Create a correlation matrix between all of the numeric variables

## Submission

Now you can grab the 'share' link and submit that! Please make sure to change the sharing settings so that anyone with the link can see the notebook. Good luck and let us know if you run into issues.