# Stat 610 Homework 3

Due Wednesday, September 25, 11:59pm

## Assignment

In this assignment, you will make some functions that download datasets from the census bureau, do some processing, and either store them as a variable in R or write them to the disk. As we discussed in class, please make a first attempt without using generative AI. If you are happy with the code you wrote on your own, you can turn in just what you wrote. If you felt like you needed to use chatGPT or something similar, compare the code you wrote to what chatGPT gives you in response to the questions. Note whether the code works and whether there are any stylistic or substantive differences with what you wrote.

You should do the following:

1. The datasets in question are available at addresses of the form `https://www2.census.gov/programs-surveys/cps/datasets/<yyyy>/basic/<month><yy>pub.csv` where `<yyyy>` is a four-digit description of a year (e.g. 2000, 1995), `<yy>` is a two-digit description of a year (e.g. 00, 95), and `<month>` is `jan`, `feb`, `mar`, `apr`, `may`, `jun`, `jul`, `aug`, `sep`, `oct`, `nov`, `dec`. (You can look at the files at https://www2.census.gov/programs-surveys/cps/datasets/2020/basic/ to see an example).

   Supposing that you have a variable `year` giving the four-digit year and a variable `month` giving the month, write a line of code that will construct the url where the relevant data can be accessed.

2. Note that the function `read.csv` can read files from a url. Write a function that will take as arguments a four-digit year and a month (as in the previous question), will download the data and return it as a data frame. Since the files are very large, please only download a subset of them using the `nrows` argument to `read.csv` (e.g. `read.csv(<file-name>, nrows = 1000)`).

   The function should look something like this:

   ```
   get_data <- function(year, month) {
     <do stuff>
     <return something>
   }
   ```

3. Modify your function so that it checks whether it was given a valid year and month. (Year should be 1994 or later and month should be one of the codes specified in the first question.) If the year or the month is not valid, print a message saying what the problem was and return `NA`.

4. The variable `prtage` gives the age of the individual surveyed. Modify your function so that it takes two additional arguments, maximum age and minimum age, and so that it subsets

the dataset to only include rows for which `prtage` is between the minimum and maximum age.

5. Modify your function so that there is an option to write the data to the disk instead of returning a data frame object. You will need to use the `write.csv` function, which also requires you to specify an output file name. Make sure that your output file name contains information about the year, month, and age range you used.

## Submission parameters

You should submit an Rmd file and the corresponding pdf or html on canvas. The files should contain both the code you ran and the answers to the problems.