Problemset 2, Econ 980w, Spring 2019:
Estimating top income shares
using the Survey of Consumer Finances ("SCF")

The purpose of this exercise is to give you some feel for the basic tasks involved in analyzing actual data. This includes in particular

- downloading the data from the internet,
- getting rid of all the unnecessary data in these datasets,
- converting the data to an appropriate file format
- reading them into the statistical software used,
- and generating some descriptive statistics.

Once these tasks are completed, we can proceed to ask statistical questions that can be answered using methods you learned in this class, such as

- Can we conclude from these data that inequality has increased?
- Are there significant differences in poverty rates between different demographic groups?
- Is the distribution of wealth more unequal than the distribution of incomes?

We will analyze data from the Survey of Consumer Finances, or SCF. As described on the homepage of the federal reserve,
http://www.federalreserve.gov/econresdata/scf/scfindex.htm:

"The Survey of Consumer Finances (SCF) is a triennial survey of the balance sheet, pension, income, and other demographic characteristics of U.S. families. The survey also gathers information on the use of financial institutions."

The data we will be using are available in various formats; Excel format will be the easiest to use. Download the datasets ("Summary Extract Public Data") for 1989 and for 2016 from the above webpage.

These data sets contain a lot of variables. Different columns correspond to different variables, different rows correspond to different observations (or entries).

The variables we will be using are the following:
- WGT (sample weight)
- ASSET (total value of assets held)
- DEBT (total value of debt)
- NETWORTH (difference between assets and debt)
- INCOME (total income)
- WAGEINC (wage and salary income)
- EDUC, (education of the household head)
- AGE (age of the household head)
- RACE (of the respondent)

Furthermore, for some technical reasons, these data sets contain five entries for every household. We will only keep the first entry, and delete the remaining four entries.

For figuring out the required R commands, the following two cheatsheets will be helpful:
https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf
https://github.com/rstudio/cheatsheets/raw/master/data-import.pdf
A comprehensive overview of working with data in R is available here:
http://r4ds.had.co.nz/

Steps of your assignment:

1. Pick one of the survey years.
2. Download the corresponding data from the internet.
3. Create a new R Script.
4. Write library("readxl") and library("tidyverse") on the first two lines, to load the necessary packages.
5. On the third line, import your .dta file using "read_xlsx()"
6. Delete all variables except for the ones listed on the previous page, using "select()"
7. Save the R script, and then execute it clicking "Run" in the top right corner of editor.
8. Look at the data set by clicking on it in the top right window.
9. Close the data viewer.
10. The dataset has five lines for each household. We need to delete all but one of them. Do so using "slice()".
11. Continue working on your script. Calculate the poverty line (0.6 times the median income), the .9 quantile of the income distribution, and the .99 quantile Throughout use weights WGT, for instance using "weighted.mean" in the "summarise()" function.
12. Generate indicator variables for the following:
    - whether a household is below the poverty line.
    - whether a household is among the top 10 percent (top 1 percent) of income earners.
    - whether the survey respondent is non-Hispanic white (race=1), and
    - whether the respondent is black (race=2).
13. Calculate the poverty rate for the full population, for the subpopulation of non-Hispanic white households, and for the subpopulation of black households.
14. Calculate a variable incsharetop10 containing household income divided by mean income, times an indicator for whether the household is among the top 10 percent. The mean of this variable will be the share of incomes going to the top 10 percent. Calculate this mean.
15. Do the same for the top 1% income share
16. Replicate the entire analysis for a different survey year. To do so, you just have to change the data set you import in your R script!