Assignment 1

# Overview

In this assignment you'll load some data into a python notebook, and use some basic functions to do some basic analysis.

Please present this in a notebook format. Make sure that it is organized and makes sense, you're attempting to illustrate and communicate to others, so please try to make it clear and readable. Take a look at this (<https://www.kaggle.com/firefliesqn/tuning-deepsort-helmet-mapping> ) for formatting/layout example – sections are labelled, there’s an explanation, it’s easy to read. Ignore the contents, that’s totally different.

# Tasks

* Age:
  + Make and plot a Hist and Pmf for age.
  + What fraction of the people in the data are 51? What fraction are older than 51?
  + What is the median age? Does the distribution of the sample data seem to mirror the working age population?
* Demographics:
  + Consider some of the demographic features: Education, Race, Hisp, MaritalStatus, and Nodeg. This data came from the United States, does it appear to be a representative sample of the US population? Demonstrate this in some way in your code.
* Earnings:
  + Make and plot a graph or graph of your choosing of the 3 earnings values, in order to answer the below question. Identify how the graph gave you your answer.
    - What is one conclusion could you draw from visualizing of the earnings in the different year? Please express it in plain language/non-statistics terms/as though you were explaining to one of your friends what happened to earnings between 1974 and 1978?
  + Which has the greatest effect size on 1978 earnings: Race, Hispanic, MaritalStatus, or Nodeg? What could you investigate further in an attempt to explain this?
  + Plot a histogram and PMF, and compute useful descriptive statistics (think: average...) for the 1978 earnings value. Use the "Cohorts" code from the quiz to break the data into cohorts, plotting each group (either on one chart, or separately, whichever makes the most sense for examining the data - state specifically why you chose 1 vs many charts.
  + What is the difference in median income between the groups? Theorize a reason for the difference between the groups that could be explained with more data. Briefly describe the data you'd need. This does not need to be something you have data for, or know how to solve right now - just one well founded hypothesis on what might explain the difference.
  + Are there outliers in the 1978 earnings data? Demonstrate this in some way with your code. What can you infer from the presence of outliers that may impact analysis of this data

Deliverables

* Your Jupyter notebook.