Assignment 6: Due Tuesday October 23, 2018, 5pm

This assignment is worth 3 homework grades. (30 points)

You are working with a researcher who is interested in writing a grant and needs some preliminary data. The data can be extracted from several relational databases (similar to the Veterans Affairs Corporate Data Warehouse). This will require you to link several tables of data together. Therefore, it is important for you to understand that within this type of data, there are two identifiers for individuals. There is a unique identifier for each individual (uniqueID) and there is a site specific identifier. Individuals may visit more than one site for care within this healthcare system. At each of these sites, they will obtain an identifier unique to that site (siteID). Therefore, an individual may have multiple siteIDs (dependent on the number of sites they visit), but they will only have one uniqueID.

The data *description.xlsx* contains the variable names, a brief description and, where important, a range of values and the value key, for each of the data sets listed below:

The following are the data sets that you will be working with:

1. MainPatientFile.csv: Contains demographic information for the cohort. This is unique at the siteID level (i.e. there is a row of data for each).
2. CohortCrosswalk.csv: This contains the file that links the siteID with uniqueID. Unique at the uniqueID level.
3. OutpatientVisits.csv: Contains information about outpatient visits. Unique at the siteID level.
4. ODiagnosisCrosswalk.csv: Contain the linkage between the diagnosis code and the diagnosis for outpatient visits. Unique at the diagnosis code.

You are to create a data set called *demo* that contains one row of data per individual, which contains the following variables:

- UniqueID
- Sex
- Birth Date
- Race
- Marital Status
- Income
- Age (based on date record last updated)
- A1_dx: Indicator of whether an individual has chronic condition A1 prior to 2015: In order to designate an individual as having a chronic condition, he/she must have at least two independent outpatient diagnoses of the condition**.** Chronic condition A1, can be classified at multiple levels (A1.1 through A1.6), all levels should be included. NOTE: Use the diagnosis name and **not** the diagnosis code to determine if the individual has the chronic condition. Diagnosis codes can change over time, but the diagnosis name will remain the same.

Deliverables:

1. Order the day by UniqueID and print the first 10 observations of the demo data file.
2. Create plots of the distribution of chronic condition A1, by sex, by race, and by marital status

Some important information

Individual demographics should be defined based on the last site visited (most recently updated record) for an individual.

Make sure your code is well documented/commented. The main SAS program should be saved as LASTNAME_FIRSTNAME_HW6. Everyone **MUST** submit this file. Any formats that are written should be saved in a separate file LASTNAME_FIRSTNAME_HW6_FORMATS, and should be temporary. The main program should call this program. Similarly, any macros that are written should be saved in a separate file LASTNAME_FIRSTNAME_HW6_MACROS, and should be called by the main program. You may only submit a MAXIMUM of 3 SAS files. However, the only program that I will run is the main program (LASTNAME_FIRSTNAME_HW6). Make sure to use the %let statement to indicate the pathname in only relevant places. Assume that I will save all code in the same folder. **DO NOT** create any permanent SAS data sets or formats.

All output requested should be **CLEARLY** commented and titled within the SAS program. However, no additional output should be "submitted". I will run the code and look at the output created. The only output that should be provided is what is asked for.