Stat 604
Assignment 11 - SAS

You should have all of the information you need to complete this assignment by viewing Lessons SAS01 through SAS08.
Perform each of the exercises listed below. To the extent you have been taught to control it, your output should match that in the PDF file posted on eCampus. Download the file **zip_codes.sas7bdat** from the **Assignment Data** section of eCampus to your personal SAS homework library folder for use in this assignment. Be sure to include "access=readonly" in any libref that you assign to the homework folder.

1. Begin your program with the required header, filename, and libname statements. As always, your program must include comments in the appropriate places.
2. The zip_codes data set was created from a raw text file. Consequently, all of the variables are text even though the original data may have been numeric. Get familiar with the data before starting to write your program code. You are to create in your personal SAS library folder a new permanent data set that is a "cleaned up" version of this data set. Efficiency must be a consideration in this data step as you decide when to drop or keep variables and in the order of any conditional processing. Assign labels to the variables as shown in the sample output posted on eCampus. The output data set must contain the same variables as shown in the sample data. Decommissioned zip codes are to be removed. Please review the input data and output data very carefully before asking a lot of questions about the requirements of this assignment. The cleaning process will update and transform some of the variables as described below:
    a. Some attributes like length and data type cannot be changed once they are in the descriptor portion of the data set. You are to use a method similar to explicit conversion to reduce the length of the county variable to 31 in the output data set. You will also need to transform some of the values as follows: Most of the **county** names end with the word County, Parish, or Borough. This is redundant and takes up extra space. Use one or more text manipulation functions to remove these words from each value. Make sure you do not inadvertently remove the word Borough if it occurs somewhere besides the end of the county name. If the county name does not end in one of these words it must remain unchanged. The name of the variable containing the transformed county names must remain **county** in the output data set.
    b. Use an explicit conversion method to convert the **estimated_population** variable from character to numeric.
    c. There are 5 distinct values of **timezone** that have names separated by an underscore. For those records, use one or more conditional statements and the left substr function to replace the _ with a blank space in the **timezone** variable.
3. This step draws heavily from the concepts covered in the chapter on Summarizing Data (Prog2-Ch3). We want to use the "clean" data set created in the previous step to create a temporary data set in which the information is consolidated by city within each state. You are to collapse the data from the original data set in a new temporary data set with one row per city much like you would summarize data by department. (TIP: You may find it helpful to wait until you have your summaries working before you put in code to remove observations and variables. That way you are better able to see what is happening from observation to observation.)

   a. Use a sort procedure to sort the clean data set in place as needed to summarize the data in the subsequent steps.
   b. Since the original variables of **zip** and **estimated_population** would only reflect the data in the last row of each city they are not needed in the resulting data set.
   c. Create a new variable labeled Est. City Population that contains the total of all the estimated population values for each city. The method used to create this value must be tolerant of any missing values. The values in this variable must be displayed with the comma separator as shown in the output on eCampus.
   d. Create a new variable labeled Zip Codes that contains a list of all the zips for that city separated by commas. You can create this value in a manner similar to the way you would manually create a numeric summary. When you create a manual summary, you take the value of the current record and add it to the summary value that is carried over from the previous records. Since you cannot literally "add" character values together you will concatenate instead of adding. You will concatenate the current zip code text to the list of zip codes that may have already been accumulated from previous observations in your zip code list variable. This list could be up to 1,700 characters.
   e. Output only those cities in which the estimated city population is greater than 0.
4. Open the PDF destination and choose one of the options so that bookmarks are not created for this output.
5. Print the descriptor portion and a subset of the data portion from both data sets created in this assignment. Subset the data by printing out only the observations for the cities Albany, Center, Reno, Rome, Paris, San Juan, Juneau, and Washington. Be sure labels are shown on your printed output.
6. Convert the program and log to PDF files and submit them to WebAssign along with your SAS output.