STAT 506: Homework 2

For these problems you will need to access the data in the PG1/data folder. Use the `libname` statement we learned to load this each time you work on your assignments. You should call it '`pg1`' to be consistent with the SAS materials.

I tried to *italicize* the parts where I expect you to actually show me something in your homework solutions if it is not obvious.

1. Initial Exploration of the National Parks Summary Data with Various Procedures

a. Write a PROC PRINT step to display only the first 12 observations in **pg1.np_summary**. *Show your code and the output.*

b. Add a VAR statement to the PROC PRINT step (from part a) to include only the variables **Reg**, **ParkName**, and **Type** (in that order). Notice that **Type** is based on **ParkName**. Do you observe any possible inconsistencies in the **Type** abbreviations used for the different types of parks? *Show your code and output, and answer the question.*

c. Now using all the observations in **pg1.np_summary**, write a PROC FREQ step that uses a TABLES statement to produce separate frequency tables for **Reg** and **Type**. Which codes/values appear only once each in **pg1.np_summary** for these variables? *Show your code and output, and answer the question.*

d. Write a PROC MEANS step for all the observations in **pg1.np_summary**. Calculate summary statistics for just the **DayVisits** and **TentCampers** columns. What are the minimum values for the number of recreational day visitors and for the number of tent campers? *Show your code and output, and answer the questions.*

e. Write a PROC UNIVARIATE step for all the observations in **pg1.np_summary**. Calculate summary statistics for just the **DayVisits** variable. What are the two lowest values and two highest values of **DayVisits**? *Show your code and the relevant part of the output, and answer the questions.*

f. Write a PROC PRINT step and use a WHERE statement to display only the row/observation that had the maximum number of **DayVisits**. (It's OK to just hardcode in a value here.) *Show your code and output.*

2. Further exploring the National Parks Summary Data

a. Open the program **p103p04.sas** (from the "practices" folder). Add a WHERE statement to print only the rows where **ParkName** includes the word "Preserve" anywhere in the name of the park using wildcards. What codes (in **Type**) are currently being used to denote Preserves? *Show your code and output, and answer the question.*

b. Edit the VAR statement to additionally include the **DayVisits** variable. Add a second WHERE statement (below the previous one) to include only observations that had between 3,000 and 300,000 (inclusive) Recreational Day Visitors. Run the code to see if you get the expected results. *Show your code and the corresponding Log notes.*

c. Combine the two previous WHERE statements into one WHERE statement that uses both conditions (Preserves with between 3,000 and 300,000 visitors) for subsetting. *Show your code and output.*

3. Using a Macro Variable

a. Create a macro variable named **regcode** and use it to store the text "MW". *Show your code.*

b. Write a PROC MEANS step to calculate summary statistics for the variable **ACRES** in **pg1.np_summary**. Use a WHERE statement to only include observations with the variable **REG** equal to your macro variable **regcode**. If done properly, this should be 18 observations. *Show your code, corresponding log notes, and output.*

c. Change the value stored in the **regcode** macro variable to "IM". Rerun that statement and rerun the same PROC MEANS step as before. This time, there should be 52 observations included. *Show your code, corresponding log notes, and output.*

d. Remove the WHERE statement from the PROC MEANS step and replace it with the statement: `BY reg;` Run the edited step and observe the output. *Show just your code and corresponding log notes.*

4. **Using Formats**

   a. Write a step to examine the descriptor portion of the **pg1.np_westweather** table. Which format is currently being used to display the **DATE** variable? *Show your code and answer the question.*

   b. Write a PROC PRINT step to display the first 6 observations of **pg1.np_westweather**. Use the **DATE9.** format to display **DATE**, and use the **4.1** format to display both **SNOW** and **SNOWDEPTH**. *Show your code and output.*

5. **Sorting the National Parks Summary Data**

   a. Write a PROC SORT step to read **pg1.np_summary** and create a temporary sorted table named **np_sorted**. Include a BY statement to order the data by first by **Reg** and then by descending **DayVisits**. Add a WHERE statement to select **Type** equal to either "NP" or "NS". *Show your code and the corresponding log notes.*

   b. Write a PROC PRINT step to display only the first 16 observations from **np_sorted** and the only the variables **Reg**, **Type**, **DayVisits**, and **ParkName** (in that order). *Show your code and output.*

6. **Using PROC SORT to Subset a Table**

   Write a PROC SORT step which will split the **pg1.np_westweather** table into two new temporary tables named **newyearsdays** and **others**. The table **newyearsdays** should include just the first recorded observation for each unique occurrence of the variables **NAME** and **YEAR**. For example, the first observation in **newyearsdays** should be for Death Valley on Jan. 1, 2015. The second observation should be for Death Valley on Jan. 1, 2016, etc. The table **others** should include all the other observations from the original table.

   [Note that **pg1.np_westweather** is helpfully already sorted by **date**. Use the `nodupkey` option and other corresponding syntax in your PROC SORT. The table **newyearsdays** should contain 12 observations.]

   Show your code and corresponding log output.