Assignment 4

Submission format:

* Submit your code as a .sas file. The filename MUST have the following format:

LastnameFirstname\_Username\_AsssignmentNumber.sas.

In this file the problems must be listed in order. Each problem must also have a comment line that precedes it. For example, for question 2, use:

\*\*\* q2 \*\*\*;

Each *part* of a problem must also have a comment line that precedes it. For example for question 2b, use:

\*\*\* q2b \*\*\*;

This ensures that I and the graders can follow your work.

* Include your problem #’s in your solution (as a comment).
* The output of the code and SAS LOG is submitted in a Word document. You have to specify clearly which problem does the output of the code and the SAS LOG belong to.The filename MUST have the same format ( but with .doc extension). In the Word document, you should have problem #’s too.

One way of obtaining the LOG and the ouput:

Run your code then go to RESULTS tab. Click on Download results as a RTF file.

Then you can click on the LOG tab and copy the LOG of the run.

Save both the log and the output in a DOC file and name it :

LastnameFirstname\_Username\_AssignmentNumber\_output.doc

* The code should be written so that, except for file-directory locations, the code can be directly run on any computer with SAS or SAS University ed.
* Any text that is not code ( such as responses to a question) will be included in the .sas file as comment lines. (IMPORTANT). The word document is strictly for SAS output. Everything else should be in the .sas file.
* You must submit the HW to the correct myCourses dropbox.
* Duedate: 5pm Monday, Oct 7th, 2019.
* Make sure you have both files in the submission.

This is based on the Rochester snowfall data from the previous homework. The SAS data set *snow* created there is in now stored as a permanent SAS data set in the *dirdata* directory so please use that file (data set). However, that data set includes Season ("1884-85", "1885-86", …) , but not Year (1884, 1885, …) so you will need to create the variable Year.

1. In SAS, and based on the work we did in Week 5 lecture code, create a graph of the snowfalls for November, December, January, February, and March vs. Year for the years 1950 through 2001. Connect the points by lines (like a time series) and overlay all the data on one graph. So this will be a total of 5 lines. Use different colors for each month—you can decide on the colors. Do not plot the points (you can try it, but it will make the plot look too busy). Include a legend and an appropriate title. (If you can, decide on an appropriate size and y-to-x aspect ratio for the graph.)

(Hint: lineattrs option when you make the plot, where you obtained the colors from R. See SASHelp\_MarkersAndColors.doc)

1. In SAS, let’s make more snowfall plots.

* Create plots of the snowfalls in January, February, and March vs. Year for *all* of the years in your data set. For each plot, plot only the points. (3 plots.)
* For each plot, also plot the best-fit regression line. (You may need to look up SAS help here.)
* Also plot the snowfalls in Feb vs. Jan, Mar vs. Jan, and Mar vs. Feb. Again, for each, plot the points and also plot the best-fit regression line. (3 plots.)
* *Arrange all 6 plots as 6 panel s in one graph*. (I do not want the individual plots, only this combined graph.) Use an appropriate arrangement of rows and columns.
* For all 6 plots, try to use the same scaling for the snowfall amounts. (You may need to look up SAS help here. And you may be surprised by what happens!)

1. (6pts) The new data frame we will use here is snow1 and looks like this (Year is integer, Total is gone):
2. Year Sep Oct Nov Dec Jan Feb Mar Apr May
3. 1 1884 0 0.0 1.0 27.1 22.2 17.0 3.5 19.5 0
4. 2 1885 0 1.7 8.2 8.4 16.9 16.0 6.5 7.0 0
5. 3 1886 0 0.0 22.2 12.5 12.0 18.4 6.3 1.2 0

The corresponding SAS data set is in the file *snow1.sas7bdat*. Note that Year is really the full snow season's earlier year. So year 1884 is really the 1884-1885 season.

* 1. (1) As efficiently as possible, reshape snow1 to snow1a, a SAS data set that should contain 3 columns that should be named Year, Month, and Snowfall.[[1]](#footnote-1)

Submit your code and a PROC PRINT of the first 10 records of your new data set.

For the remaining questions in Problem 3, please use snow1a as the basis for the input data.

* 1. (2) As efficiently as possible, make a table (for listing purposes—you can decide how to do this[[2]](#footnote-2)) of snowfall for only the last 10 years of data. Put months in rows and years in columns. Then do this again, but now with months in columns and years in rows.

Submit your code and your two tables.

* 1. (1) As efficiently as possible, find the total snowfall for each year. Save this into snow1Year. (This data set should, of course, contain two columns: Year and Snowfall.)

Submit your code and a PROC PRINT of the first 10 records of your new data set.

* 1. (2) As efficiently as possible, make boxplots of the snowfall for each month. (So, there should be 9 boxplots on your graph.) Each boxplot should be based on all of the years, from 1884 to 2001. All of the boxplots should be on one graph, with Month as the x-axis variable. The x-axis values should start in Sep and end in May.

Submit your code and the graph.

* 1. Bonus (1) As efficiently as possible, find the average annual snowfall for each full decade in the data set. (A decade is 10 years, starting in the “0” year. For 1930-1939, for example, call this the 1930 decade—this is actually called “the 1930’s” but we want to keep the variable as a number). Name these variables as Decade and Snowfall\_Yr. Then plot the snowfall vs. the decade. Use good axis labels (if possible) and a title in your graph. Please do not ask how to do this or post the answer to how to do this in the Discussion module of myCourses.

Submit your code, a PROC PRINT of the first 10 records of your new data set, and the plot.

* 1. Bonus (1). For each year, find the month that had the most snowfall. (For this problem, let’s not worry about ties or what your code does with ties.) Your output file should contain Year, Month, and Snowfall. Do this as efficiently as possible. Then make a table of the number of times each month had the maximum snowfall. Your answer should have months listed in the order Sep, Oct, …, May. Please do not ask how to do this or post the answer to how to do this in the Discussion module of myCourses.

Submit your code, a PROC PRINT of the first 10 records of your new data set,

and the table.

The second problem is based on the tb (tuberculosis) data set that was mentioned in the *TidyDataAndReshaping* pdf file, and is based on the example in *tidy-data.pdf* .

In the tb data set, as you know, are columns with names such as “f1524” – that means female, ages 15-24. You will be asked to recode the age-range part of this as shown in the listing below. Note: the tb data set is actually a subset of a much larger data set. Only the values that are highlighted need to be considered for the data set in this HW problem:

04 0-4

514 5-14

014 0-14

1524 15-24

2534 25-34

3544 35-44

4554 45-54

5564 55-64

65 65+

u NA

1. (3) In SAS, tb2000 data set is available.
   1. (1) “Melt” the data set in the appropriate way. Call the measured variable as cases. Call this new data set tb2000m1.

Submit your code and a PROC PRINT of the first 10 records of your new data set.

* 1. (1) Split values such as “f1524” into two columns, Sex (here “f”) and oldage (here “1524”). Use the name tb2000m2 for this new data set.

Submit your code and a PROC PRINT of the first 10 records of your new data set.

* 1. (1) Use a method of your choice to convert oldage (such as “1524”) into a new variable, age (such as “15-24”, as shown above). Then sort this data set by iso2, year, sex, and age. Then keep only these variables (in that order) and cases (as the last variable) in your final data set. Name this data set tb2000melt.

Submit your code and a PROC PRINT of the first 10 records of your new data set.

1. (1) (The intent of this problem is to see what additional work you might need to do if you decide not to reshape your data.) This is like problem 1, with snow1; however, now you need to use the SAS data set snow1 directly, that is *without reshaping it at all*.
   1. (1) (Like 1.c) In as few lines of code as you can, find the total snowall for each year. Save this into snow1AYear. (This data set should, of course, contain two columns: Year and Snowfall)

Submit your code and a PROC PRINT of the first 10 records of your new data set.

* 1. Bonus (1) (Like 1.d) In as few lines of code as possible, make boxplots of the snowfall for each month. (So, there should be 9 boxplots on your graph.) Each boxplot should be based on all of the years, from 1884 to 2001. All of the boxplots should be on one graph, with Month as the x-axis variable starting in Sep and ending in May. Please do not ask how to do this or post the answer to how to do this in the Discussion module of myCourses.

Submit your code and the graph.

* 1. Bonus (1). (Like 1.f) For each year, find the month that had the most snowfall. (For this problem, let’s not worry about ties or what your code does with ties.) Your output file should contain Year, Month, and Snowfall. Then make a table of the number of times each month had the maximum snowfall.

Submit your code, a PROC PRINT of the first 10 records of your new data set, and the table.

­­

1. You may want to think about what you want to do about month. See question 1.d, for example. [↑](#footnote-ref-1)
2. For listing purposes—don’t think that the word “table” means you should do this in PROC TABULATE. [↑](#footnote-ref-2)