PHPM 672/677 Assignment #4: Reshaping & Combining Tables (Small)
Due date: Submit on E-Campus by 11:59pm Monday 3/3
Mid point check due date: Email by 11:59pm Monday 2/25 (next week)

Submission. Submit on E-campus by 11:59pm the day before the class they are due.

1. Lab (2 points. GRADED. **Due with MIDPOINT**)
   - Fill in the excel sheet and submit
2. Midpoint email (1 point. SEPARATE from Lab)
   - Describe in one sentence, what each of the tables are (there is a total of 8).
   - What is the unit (row) of each table?
   - For each table that does not have the required unit of analysis as "county year", explain how you will convert the given table into the required "county year" table. If not applicable write NA.
   - When linking up all the tables to have all the variables in one table,
     - Which tables link up as 1-to-1 matching ? What are the matching variables?
     - Which tables link up as 1-to-N matching ? What are the matching variables?
3. Assignment (9 points):
   - Commented code (lnameN.sas; where N indicates the assignment number and lname is your last name)
   - Output from your code (lnameN.log & (lnameN.lst or lnameN.html))
   - Answers to questions: Readme.txt

Late Assignments. Each student will be allowed one late assignment, due 7 days from the due date. NO other late assignments or make up will be accepted.

Plagiarism: If you consult any outside sources when doing your work, you are expect to further document these sources. Give credit where credit is due. Plagiarism will not be tolerated.

All handed in homework should state at the top any assistance with debugging and programming, as well as citations of any program segments copied from a website.

Required & recommended readings for this assignment

1. UCLA link Optional 3: Combining Tables (Lec 6-7) on course website
2. The little SAS book (online book available from the library): Chapters 1 & 6

Guideline for assignment grading (Total of 12)

- Assignment (Total 6)
  - 1, 2 or 3: Submitted code that does not run.
  - 4: Mostly running but incorrect.
  - 5: Correct and meets requirements (i.e use of arrays and loops)
  - 6: Correct & Elegant. Comments.
- Answers to questions on the assignment (Total 3)
- Midpoint Check email (Total 1)
- Lab (Total 2) – DON'T forget to submit this with this assignment if you have not made an early submission the week before.

Assignment 4: Reshaping and Combining Tables (Small)

By the end of this assignment, you should be able to

- Append multiple tables (e.g stack tables on top of each other to increase the number of rows) using `set`

- Link up multiple tables using a shared key (e.g align the rows using the shared key, and link multiple tables to increase the number of variables in the tables) using `merge`.
- Know which table the data came from using `in`
- Aggregate/Combine multiple rows into one row by group processing `proc summary`
- Reshape tables to flip rows & columns using `proc transpose`
    - Also transpose (flip rows & columns) by groups of rows

[STEP 1: Download] Setting Up: Get data from website into the appropriate folder: The links are provided only as a reference for documentation of the data. All data are provided on the class website under Data as a zipped file. Look at the data using proc contents/proc print (use the label option) to understand what variables you have. Do NOT open these tables using point & click. If you need more information, look at information on the webpages.

1. Texas Discharge data (www.dshs.state.tx.us/thcic/hospitals/Inpatientpudf.shtm).
2. Texas Physician Supply data (http://www.dshs.state.tx.us/chs/hprc/PHYS-lnk.shtm)
3. Texas Death data (http://soupfin.tdh.state.tx.us/deathdoc.htm)
4. Texas Health Regions Look-up Table (http://www.dshs.state.tx.us/chs/brfss/pages/counties.shtm)

Big Picture: The goal is to combine all the data above into an analytic file to answer the questions at the bottom in the readme section. Given that this is your first time, we are first going to do this on "tiny" datasets, so you can see the full flow of data. This is assignment 4. That means the answers to your questions below will not be correct because the data is not the full dataset. Then in assignment 5, you will adapt your program in assignment 4 to process the full data and get the correct answers. In addition, you will expand your analysis to other units of analysis.

Folder Structure: Good file organization is essential to keeping track of your code. For assignments 4 and 5, I recommend the following folder structure

assign4/ : all sas programs for assignment 4
    src/ : all source data for assignment 4 (**you should unpack the data files you downloaded here**)
    data/ : all data for assignment 4
assign5/ : all sas programs for assignment 5
    data/ : all data for assignment 5 (you should copy the data files you need for assignment 5 into here)

[STEP 2: Subset] Generating tiny datasets for Assignment 4: Copy the following code into a sas program and run. This will subset all tables to only those relevant to 4 counties. Remember to fill in the correct PATH, and repeat for all 8 datasets. <u>DO NOT submit this code. This is not part of the assignment. You should not be doing anything else in this code.</u>

```
libname src 'PATH/assign4/src/';
libname data 'PATH/assign4/data/';

data data.phy2010;
    set src.phy2010;
    where county in ('Brazos', 'Brewster', 'Austin', 'Childress');

data data.phy2011;
    set src.phy2011;
    where county in ('Brazos', 'Brewster', 'Austin', 'Childress');

* repeat the above code for all 8 datasets;
* This will generate a tiny dataset with information on only 4 counties;
```

For assignment 4, you will be working with the TINY datasets so that you can see the full dataset as you build your process. Also, I have left out the problem data rows. You should be using proc print data=fn (obs=100) in between data steps and proc steps to see what your code does to the tables in each step.

[STEP 3: Combine all the tables]

- Keep only the rows and columns you need for the analysis. This analysis covers years 2010-2012 and 4 counties `'Brazos', 'Brewster', 'Austin', 'Childress'`.

- The initial unit of analysis for this assignment is "county year". Thus, you will have to have 1 observation per "county year" in each table unless it is a county look up table. The county look up table has 1 observation per county which should link to all years.

- Which tables need to be changed to have 1 row per "county year"? How? Change these tables to the right unit of analysis. Following are things to consider.
  - If you have multiple tables of the same kind of data (one table per year), append them as appropriate. Hint: rename variables as needed. Convert variable type to be the same if needed. Add year variable as needed during the appending process using `in` option.
  - If the table is wide (years are going across columns), then convert to long (so that the years are going down rows)
  - If you have multiple rows per "county year", then aggregate the rows into one row using the correct statistics (sum, mean etc)

- Now link all the tables together to have all the variables in one table (counts of discharge, physician supply, number of deaths, border county, etc).
  - What are the variables you must line up (matching vars), to get the tables to link up properly? (Hint: 1-to-1 matching and 1-to-N matching have to be done separately)

Readme file: Answer the following questions in a text file called readme.txt
You must write code to answer them. The answer can come from the log or the lst file.

<County Year Analysis>

1. How many total observations do you have when the merge is complete?
2. How many observations are missing patient data?
3. How many counties in your dataset are missing patient data for AT LEAST 1 year? (notice, we want the number of counties, not the number of observations that are missing patient data) ?
4. What is the average size (in terms of the population variable) of all observations (county/years) that have missing patient data?
5. Which observations (which county/years) have more deaths than inpatient discharges? How many observations have more deaths than physicians?
6. Compare the Ratio of DPC per 100,000 Population between the border and non-border counties. Which has a higher average Ratio of DPC per 100,000 Population? Also give the actual values (DPC=direct primary care physicians)