STAT 440 – Homework 4

Students are encouraged to work together on homework. However, sharing or copying any part of the homework is an infraction of the University's rules on Academic Integrity.

Final submissions must be uploaded to our Compass 2g site on the Homework page. No email, hardcopy, or late submissions will be accepted.

The HW Report should include the output generated from the following exercises:
1-bde, 2-abdfhj, 3-efg

Getting the program file ready

a. Create a folder on the hard drive with the following pathname – C:\440\hw4. Save all data files accompanying this assignment in that folder. If you cannot create the folder because you are working on a university computer and don't have permission, create the …\440\hw4 folder elsewhere.
b. Assign the library reference **hw4** to the folder 'C:\440\hw4'. Use this library as your permanent library for this assignment. If you could not create the folder, assign the library reference **hw4** to your …\440\hw4 folder.
   Note: If you are using a folder other than 'C:\440\hw4', you must change any pathname references in your program file to 'C:\440\hw4' before submitting your homework.

Submitting your work to Compass 2g

You are to submit two (and only two) files for your homework submission.

1. Your SAS program file which should be saved as **HW*n*_*YourNetID*.sas**. For example, my file for the HW4 assignment would be HW4_dunger.sas. All program statements and code should be included in one program file.

2. Your Report including all relevant output to address the exercises. For this homework, use ODS to send your results to a Rich Text Format (RTF) file called ***YourNetID*_HW*n*.rtf**. Only include your final set of output. Do not include output for every execution of your SAS program. Use the template file **hw3 template.sas** as your guide.

   Once the results have been sent to the .rtf file, you may open it in Word and include your own responses in the relevant areas (as directed in the exercises).

You have an unlimited number of submissions, but only the last one will be viewed and graded. Homework submissions must always come as a pair of files, as described above.

1. Product data

You will be working with the SAS data files **inventory** (which contains the model ID and price of various products) and **purchase** (which contains the model ID, quantity purchased, and customer who purchased the product).

a. Merge the **inventory** and **purchase** data sets to create a new, temporary SAS data set called **purchase_price_*NetID*** based on the Model number.
   - Add the Price value found in the **inventory** data set to each observation in the **purchase** data set.
   - There are some models in the **inventory** data set that were not purchased (and, therefore, are not in the **purchase** data set). Do not include these product models in the new data set.
   - Compute a new variable called TotalCost that calculates the total invoice cost for each Model purchased.

b. Print the data portion of **purchase_price_*NetID*** including all variables, excluding observation numbers. (Include results in the HW Report.)

c. Using a separate DATA step, create a list of all Models (and the Price) that were not purchased in a permanent SAS data set called **not_purchased_*NetID***.

d. Print the data portion of **not_purchased_*NetID***, excluding observation numbers. (Include results in the HW Report.)

e. Repeat parts (a) and (c), but this time create both new data sets in a single DATA step. (Include the log entries and any resulting notes from the lines of this DATA step in the HW Report. Not the entire session log!)

2. Consumer Expenditure Survey

The Consumer Expenditure Survey (CE) is conducted by the U.S. Department of Labor, Bureau of Labor Statistics to provide data on the buying habits of American consumers. The survey collects data at each quarter of the year at both the consumer unit (i.e., family) and member (i.e., individual person) level. Thus, each consumer unit (CU) may be composed of multiple members (i.e., a family could have 1, 2, 3,… members). A CU may or may not participate in all the interviews (e.g., respond to 1st and 4th quarters, but skip 2nd and 3rd).

You will use the following SAS data sets.

- **fmli071** – a subset of the family-level interview data for the 1st Quarter of 2007
- **fmli072** – a subset of the family-level interview data for the 2nd Quarter of 2007
- **fmli073** – a subset of the family-level interview data for the 3rd Quarter of 2007
- **fmli074** – a subset of the family-level interview data for the 4th Quarter of 2007

- **memi071** – a subset of the member-level interview data for the 1st Quarter of 2007
- **memi072** – a subset of the member-level interview data for the 2nd Quarter of 2007
- **memi073** – a subset of the member-level interview data for the 3rd Quarter of 2007
- **memi074** – a subset of the member-level interview data for the 4th Quarter of 2007

Description:

- The specifications of each variable in each data file can be found in the file **Interview Data Dictionary.pdf**. It contains information on every one of the hundreds of variables from the original survey, but only a subset of those variables are used in the data sets provided.
- There are two additional variables in this data set which I extracted from the values of the NEWID variable in the original BLS data.

Variable Name	Description
CU_ID	Consumer Unit sequence number – This uniquely identifies the CU and can be any number from 1 through 9999999.
INT_NUM	Interview Number – Value of 2It is possible for a CU to skip an interview. For example, a CU could have a 2nd, 3rd and 5th interview, but no 4th interview.

- Every data set is sorted in ascending order by CU_ID.
- In the **fmli** data sets, CU_ID is unique to each observation. That is, a valid CU_ID occurs at most once in each of the four **fmli** data sets.
- In the **memi** data sets, CU_ID may occur more than once if the CU (i.e. household) has more than one member. For example, a family of four would share the same CU_ID and so those four observations in a **memi** data set would all have the same CU_ID.

Source: U.S. Department of Labor, Bureau of Labor Statistics, Consumer Expenditure Survey, Interview Survey, 2007. http://www.icpsr.umich.edu/cocoon/ICPSR/STUDY/25623.xml

a. Use PROC CONTENTS to view the descriptor portion of each of the eight data sets. Construct a table that lists the number of observations and number of variables in each data set. This table can be made in Word and does not have to be compiled using SAS. (Include the table in the HW Report. Do not include the output from PROC CONTENTS.)
b. Which, if any, of the eight data sets are alike in structure? Which, if any, of the eight data sets are unlike in structure? These can be answered directly in Word. (Include your response in the HW Report.)

c. Concatenate (but do not interleave) the four family-level data sets. Also create a new variable called QTR that uniquely identifies during which quarter of 2007 the interview took place. Name the resulting temporary data set **fmli07*_NetID***.
d. Print the descriptor portion of the new data set. (Include your results in the HW Report.)

e. Concatenate (but do not interleave) the four member-level data sets. Also create a new variable called QTR that uniquely identifies during which quarter of 2007 the interview took place. Name the resulting temporary data set **memi07*_NetID***.
f. Print the descriptor portion of the new data set. (Include your results in the HW Report.)

g. Merge the data sets **fmli07*_NetID*** and **memi07*_NetID*** into a new permanent data set called **ce07*_NetID***. This should be a merge that matches a consumer unit with all corresponding family member interview responses.
h. Print the descriptor portion of the new data set. (Include your results in the HW Report.)

i. How many consumer units participated in all four quarterly interviews in 2007? Create a temporary SAS data set called **all_four*_NetID*** containing only the CU_ID of those who fit this description.
j. Print the descriptor portion of the new data set. (Include your results in the HW Report.)

3. Hockey data

In this exercise, you will continue to work with the hockey data. The SAS data set **skaters** is the same as the one used in HW3 which created from a raw data set in HW2. Recall that it contains career data for every one of 6498 players in the history of the National Hockey League to play a position other than goaltender. Data is current as of February 24, 2015.

The raw data set **hockey goalies 24FEB15.dat** contains career data for all 716 players in the history of the National Hockey League to play goaltender. Data is current as of February 24, 2015.

Field	Name	Description
1	Player	NHL Player
2	First	First year of NHL career
3	Last	Last year of NHL career
4	GP	Games Played
5	GS	Games Started
6	W	Wins
7	L	Losses
8	TOL	Ties/Overtime/Shootout Losses
9	GA	Goals Against
10	SA	Shots Against
11	SV	Saves
12	SV_PCT	Save Percentage
13	GAA	Goals Against Average
14	SO	Shutouts
15	MIN	Minutes
16-20	v16-v20	[Read in, but drop before they get to the SAS data set **goalies**.]
21	G	Goals (that the player recorded, not opponents)
22	A	Assists (that the player recorded, not opponents)
23	PTS	Points (that the player recorded, not opponents)
24	PIM	Penalties in Minutes
25-28		[Do not read these in.]

a. Write a DATA step to read the values of **hockey goalies 24FEB15.dat** into SAS. The output data set is to be a temporary SAS data file called **goalies_*YourNetID***. Choose appropriate names, labels, and other attributes as needed for the variables.
Hint: Remember to look back at HW2 when you did something similar for **skaters**.

b. Concatenate **skaters** and **goalies_*YourNetID*** into a new data set called **all_*YourNetID***.
    - The variables Player, First, Last, GP, G, A, PTS, PIM appear in both input data sets, and they each mean the same thing in both data sets.
    - TOI (**skaters**) and MIN (**goalies**) mean the same thing. Be sure that all values appear only in the variable TOI in the **all** data set.
    - Be sure that all goaltenders (that is, players who come from the **goalies** data set) each have the value "G" for the variable Pos.

The raw data set **hockey HOF.csv** contains a list of all 263 individuals inducted into the Hockey Hall of Fame as a player. Data is current as of February 24, 2015.

Field	Name	Description
1	Player	Player
2	Year	Year that the Player was inducted into HOF
3	First	First year of NHL career
4	Last	Last year of NHL career
5	GP1	Games Played as Skater
6	G	Goals
7	A	Assists
8	PTS	Points
9	PM	Plus/Minus
10	PIM	Penalties in Minutes
11	GP2	Games Played as Goalie
12	W	Wins
13	L	Losses
14	TOL	Ties/Overtime/Shootout Losses
15	SV_PCT	Save Percentage
16	GAA	Goals Against Average

c. Write a DATA step to read the values of **hockey HOF.csv** into SAS. The output data set is to be a temporary SAS data file called **hof_*YourNetID***. Choose appropriate names, labels, and other attributes as needed for the variables.

d. Merge **all_*YourNetID*** and **hof_*YourNetID*** into the following two new data sets in a single DATA step using Player as the reference variable:
    1. **NHLhof_*YourNetID*** containing those observations appearing in both data sets.
    2. **otherhof_*YourNetID*** containing those observations appearing only in **hof**.

    - **all** contains only NHL players, but **hof** contains all players in the Hall of Fame, some of whom did not play in the NHL, and so those players will have no recorded stats.
    - Most variables appear in both input data sets, and they each mean the same thing in both data sets.
    - It's okay to leave GP, GP1, and GP2 in both output data sets.
    - The source of **skaters** and **goalies** put a star (*) after the name of each player in the Hall of Fame. I augmented the values in **hof** to also include a star in the same way so that matching would be much easier. This fact shouldn't affect your program, yet I want you to be aware of the stars floating around.

e. Print the descriptor portion of both new data sets. Include only the first table containing the number of observations and variables for each. (Include results in the HW Report.)

f. Print a list of all NHL goalies in the Hall of Fame who have recorded more than 15,000 saves. Include only the Player's name, First, Last, GP, SV, and SV_PCT for each player. (Include results in the HW Report.)

g. Print a list of all NHL players in the Hall of Fame who have recorded more than 30 short-handed goals in their career. Include only the Player's name, First, Last, GP, G, and SH for each player. (Include results in the HW Report.)

Personal Challenge: Not for any points or bonus and not to be included in your HW4 submission. Just for your own development.

Refer back to Exercise 2. Can the **ce07_*NetID*** data set created in parts (c), (e), and (g) over several steps be created in a single DATA step?

- If so, try to write a program that merges the original eight data sets as directed in the three parts above, but in a single DATA step.
- If not, try to explain why such a single DATA step could not perform the desired merge for these data.