STAT 440 – Homework 06

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.

Getting the program file ready

a. Create a folder on the hard drive with the following pathname – C:\440\hw06. 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\hw06 folder elsewhere.

b. Assign the library reference **hw06** to the folder 'C:\440\hw06'. Use this library as your permanent library for this assignment. If you could not create the folder, assign the library reference **hw06** to your …\440\hw06 folder.
   Note: If you are using a folder other than 'C:\440\hw06', you must change any pathname references in your program file to 'C:\440\hw06' 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 HW06 assignment would be HW06_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.

   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. 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. This is an extension of the data you worked with back in HW2. Consider revisiting that assignment as it may help you with these exercises.

   The Consumer Expenditure Survey (CE) is conducted by the Bureau of Labor Statistics to provide data on the buying habits of American consumers. The Interview data you'll explore generally tracks consumer units' (CU) large expenditures, such as major appliances and cars. An Interview "quarter" refers to the calendar quarter in which the interview occurred. For example, any Consumer Unit interviewed in April, May, or June would have their data stored in the quarter 2 (2007Q2) datasets. During an interview, the CU is asked to report expenditures for a reference period of three months. So, for a CU interviewed in April, their expenditures in the YYQ2 file are for January, February, and March.

   The Interview survey collects data at each quarter of the year at both the consumer unit (i.e., family) level 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.

   fmli07*i*
   - There is one record per CU.
   - Each CU is uniquely identified by CU_ID.
   - It 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.
   - Variables include demographics for the reference person and spouse of reference person, income at the CU level, sample housing unit information, and summary level expenditures.

   memi07*i*
   - There are multiple records per CU.
   - There is one record per member.
   - Unique records are defined by the combination of CU_ID and MEMBNO.
   - Variables include demographics about CU members, member level income, and member relationship status to the reference person.

   Description:
   - The specifications of each variable in each data file can be found in the **Interview Dictionary** file. 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.
   - 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.

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.)
   - Note that all the **fmli07** data sets have the same number of variables, as do the four **memi07** data sets.

b. 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 **fmli2007_*NetID***.

c. Print the descriptor portion of the new data set. (Include your results in the HW Report.)
   - Use the results of part (a) to check the math of your concatenation, and comment on whether the results are as you would expect.

d. 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 **memi2007_*NetID***.

e. Print the descriptor portion of the new data set. (Include your results in the HW Report.)
   - Use the results of part (a) to check the math of your concatenation, and comment on whether the results are as you would expect.

f. Merge the data sets **fmli2007_*NetID*** and **memi2007_*NetID*** into a new permanent data set called **ce2007_*NetID***. This should be a merge that matches a consumer unit with all corresponding family member interview responses.

g. Print the descriptor portion of the new data set. (Include your results in the HW Report.)
   - Use the results of part (a) to check the math of your concatenation, and comment on whether the results are as you would expect.

h. How many consumer units participated in at least three of the four quarterly interviews in 2007? How many consumer units participated in all four quarterly interviews in 2007? Create SAS data sets called **atleast_three_*NetID*** and **all_four_*NetID*** containing only the CU_ID of those who fit these descriptions.

i. Print only the first table of the descriptor portion of the new data sets. (Include your results in the HW Report.)