STAT 440 – Homework 02

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\hw02. 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\hw02 folder elsewhere.

b. Assign the library reference **hw02** to the folder 'C:\440\hw02'. Use this library as your permanent library for this assignment. If you could not create the folder, assign the library reference **hw02** to your …\440\hw02 folder.
   Note: If you are using a folder other than 'C:\440\hw02', you must change any pathname references in your program file to 'C:\440\hw02' 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 HW02 assignment would be HW02_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 Portable Document Format (PDF) file called **HW*n*_YourNetID.pdf**. For example, my file for HW01 would be HW01_dunger.pdf. Only include your final set of output. Do not include output for every execution of your SAS program. Use the template file **hw01 template.sas** as your guide.

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. 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 (xxxi142) 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 xxxi142 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.
   fmli142
   - There is one record per CU.
   - Each CU is uniquely identified by NEWID.
   - 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.
   memi142
   - There are multiple records per CU.
   - There is one record per member.
   - Unique records are defined by the combination of NEWID 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, NEWID is unique to each observation. That is, a valid NEWID occurs at most once in each of the four **fmli** data sets.
   - In the **memi** data sets, NEWID 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 NEWID and so those four observations in a **memi** data set would all have the same NEWID.

   a. Use PROC CONTENTS to view the descriptor portion of each of the two data sets. (Include the table in the HW Report.)

b. Look at the Variable Attributes table for the **fmli142** and the **memi142** data sets. Using the **Interview Dictionary** file, create a user-defined format for each variable that would benefit from having a format to interpret its levels.
- Note that you might not need a unique format for each variable. A format can be applied to multiple variables.
- You should be able to do some copying-and-pasting from the **Interview Dictionary** file to save time on typing.

c. Create two new data sets, **fmli142_*NetID*** and **memi142_*NetID***, that assign your newly created formats to the appropriate variables. Also create appropriate labels for variables needing explanation.

d. Print the descriptor portion of both datasets. (Include your results in the HW Report.)

e. Print the first 10 observations in both datasets. (Include your results in the HW Report.)
- For **fmli142_*NetID***, print the variables NEWID, CUID, AGE_REF, BLS_URBN, MARITAL1, and FINCATAX.
- For **memi142_*NetID***, print the variables NEWID, CU_CODE, MARITAL, and SALARYX.

f. Create another new format for SALARYX, the variable containing the amount of wages/salary income received before any deductions.

- Those between $30,000 and $70,000 are designated as Middle Class.
- Those above $70,000 but less than $120,000 are designated as Upper Middle Class.
- Those less than $12,000 are designated as Impoverished.
- Those above $120,000 are designated as Upper Class.
- Other values are designated as Lower Class.
- Missing values should be represented with the word Missing.

g. Use a PROC DATASETS step to update the format of SALARYX in the **memi142_*NetID*** dataset.
- Read the documentation for PROC DATASETS for help, in particular the MODIFY statement.

h. Print the first 10 observations in the updated **memi142_*NetID*** dataset, but only the NEWID, EDUCA, and SALARYX variables. (Include your results in the HW Report.)