**Workshop Exercise 3**

**Exercise III: Data Management tasks**

**Datasets = thai1.sas7bdat, thai2.sas7bdat**

**Thai1.sas7bdat**

Information on n=8582 sixth grade students, including the following variables:

SCHOOLID: a character variable identifying the school the student came from

MALE: a dummy variable with a value of 1 for males and 0 for females

PPED: a dummy variable with a value of 1 for those students that had pre-primary education and a 0 for those who did not.

REP1: a dummy variable with a value of 1 for students who had repeated a grade in 1st through 5th grades.

**Thai2.sas7bdat**

Information on n=357 schools, including the following variables:

SCHOOLID: a character variable identifying the school

MSESC: the school mean SES

1. Using Proc Contents, get information on both data sets. How many observations are in each?
2. Sort both datasets by Schoolid (be sure they are not open in the viewtable browser before sorting)
3. Merge the two datasets, thai1 and thai2 by SchoolID.
4. Check the log to see how many observations are in your merged dataset. Why do you think you have this number?
5. Re-merge the two datasets, but this time only include cases that are in both datasets. How many observations do you have in your merged dataset now?

**Stacking Data:**

**Datasets = wave1.sas7bdat, wave2.sas7bdat, wave3.sas7bdat**

In this part of the exercise, we stack three waves (wave1.sas7bdat, wave2.sas7bdat, wave3.sas7bdat) of data from the Tecumseh Community Health Study to make a single long dataset containing information on all three waves. You can find information about the Tecumseh Study in the Appendix.

1. Using Proc Contents, get information about the wave1, wave2, and wave3 datasets.
	* 1. How many observations are in each wave of data?
		2. How many variables are in each wave of data?
2. Use a data step to stack all three datasets together, and make a **permanent** SAS dataset, called Allwaves.sas7bdat,
3. Get descriptive statistics for all numeric variables in your new dataset for each Round (variable = ROUND).
	1. What is the sample size in each round? Does it match the dataset sizes for the three waves of data?
4. How many and what percentage of participants smoked cigarettes in each round? Consider carrying out a crosstab using Proc Freq.
	1. How many observations are in each Round (variable = ROUND) of data?
	2. What is the percent of smokers (variable=CIG) in each Round? Be sure you report the correct percentages.
	3. How many missing values are there for the crosstab of Round by Cig?