**CLASS EXERCISE 29/1/2020**

On the USDA NASS Database, I accessed the Quick Stats tool to mine the database by:

* Selecting county as the geographical level for the state of NC for years 2010-2018.
* Getting a census for crops, dairy and energy for different Program, Sector, Group, Category until the status of number of records turns from red to green
* Downloading the selected data by clicking on “GetData/Spreadsheet”.
* Renaming the downloaded file to USDARawData.csv.

The data contained 22181 observations and 21 variables. The name of the file is usda\_data.csv.

**DATA CLEANING & EXPLORATION STEPS**

* Using S.A.S Enterprise Miner, I specified a library name with a proper column header format using “options validvarname=V7” i.e without spaces.
* I imported the data file and I saved it in a library called “USDA\_RAW”.
* I performed an exploratory analysis on the data with “PROC CONTENTS”, the output showed that the datatypes of all the columns were characters and some needed to be numeric.
* I removed the observations with the “D” and “Z” shown in the PROC CONTENTS output.
* I converted the “Value” character column to numeric type with the input function..
* I tried to query the data to derive a table for fertilizers used for crops across different counties in North Carolina for years 2012 to 2015 with a value greater than $200,000.
* I’ll also try formatting the Value variable to dollar
* I used the “proc mean” statement to get the mean of the Value variable in the table.
* There were 21155 observations with a mean of 3116631.00