**Name:**

**Assignment 7 - SAS #1 – Introduction to Working with SAS**

**MAKE SURE YOU HAVE COMPLETED THE IN-CLASS EXERCISE FIRST BEFORE YOU DO THIS ASSIGNMENT. AND MAKE SURE YOU’VE CHANGED THE SAMPLING DEFAULTS AS SPECIFIED ON PAGE 30 OF THE IN-CLASS EXERCISE! OTHERWISE, YOU WILL GET THE WRONG ANSWERS!**

1. **Initial Data Exploration**

A supermarket is offering a new line of organic products. The supermarket's management wants   
to determine which customers are likely to purchase these products.

The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of the loyalty program participants and collected data that includes whether these customers purchased any of the organic products.

The **ORGANICS** data set contains 13 variables and over 22,000 observations. The variables in the data set are shown below with the appropriate roles and levels:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Measurement Level** | **Description** |
| **ID** | ID | Nominal | Customer loyalty identification number |
| **DemAffl** | Input | Interval | Affluence grade on a scale from 1 to 30 |
| **DemAge** | Input | Interval | Age, in years |
| **DemCluster** | Rejected | Nominal | Type of residential neighborhood |
| **DemClusterGroup** | Input | Nominal | Neighborhood group |
| **DemGender** | Input | Nominal | M = male, F = female, U = unknown |
| **DemRegion** | Input | Nominal | Geographic region |
| **DemTVReg** | Input | Nominal | Television region |
| **PromClass** | Input | Nominal | Loyalty status: tin, silver, gold, or platinum |
| **PromSpend** | Input | Interval | Total amount spent |
| **PromTime** | Input | Interval | Time as loyalty card member |
| **TargetBuy** | Target | Binary | Organics purchased? 1 = Yes, 0 = No |
| **TargetAmt** | Rejected | Interval | Number of organic products purchased |

**🖉** Although two target variables are listed, this exercise concentrates on the binary variable **TargetBuy**.

* 1. Create a new diagram named **Organics**.
  2. Define the data set **AAEM.ORGANICS** as a data source for the project.
     1. Set the roles for the analysis variables as shown in the table on the first page.

(You can go back and modify variable roles even after you complete the wizard by right-clicking on the **Organics** data source and selecting **Edit Variables…**)  
  
The variable **DemClusterGroup** contains collapsed levels of the variable **DemCluster**. Presume that, based on previous experience you believe that **DemClusterGroup** is sufficient for this type of modeling effort. Set the model role for **DemCluster,** along with **TargetAmt,** to Rejected.

**Include a screen shot showing DemCluster and TargetAmt is rejected (as on page 11 or 13 of the in-class exercise).**

* + 1. Finish the **Organics** data source definition.
    2. Examine the distribution of the target variable **TargetBuy**. You can do this by right-clicking on the data source and selecting **Edit Variables…** Then right-click on **TargetBuy** and then click the **Explore** button.  
         
       What is the proportion of individuals who purchased organic products (hint: take a look at the Means column in the “Sample Statistics” window)?   
         
       **ANSWER:**
  1. Add the **AAEM.ORGANICS** data source to the **Organics** diagram workspace.
  2. Explore the **Organics** data source and answer the following questions:  
       
     Are there more males or more females in the sample? **ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
     (Hint - plot **DemGender** using a bar chart)  
       
     **Include a screen shot displaying the bar chart.**  
     What is maximum age of the people in the sample? **ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
       
     Plot **DemAffl** using a histogram with 30 “X” bins. Based on that, is it more common to have a high income or a low income (note: wealthier households have greater affluence scores)?   
      **ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
       
     Looking only at the **DemAffl** histogram you just created, people in this sample most frequently receive an affluence grade between   
      **ANSWER:** \_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_