BAN 5743 Exercise 2 (10 points) Solution

Exercise Description:

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 their loyalty program participants and collected data that includes whether or not these customers purchased any of the organic products.

The **ORGANICS** data set has 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 above, for now, this exercise will focus on the binary variable **TARGETBUY**.

1. Initial Data Exploration

For all exercises, assignments, projects related to this class, please do this first. Click Options > Preferences in the top menu and make sure you **set sampling method to random and sample size to Max** in Interactive Sampling.

a. Create a new diagram named **Organics**.

b. Create a new library for the data set **Organics** using library wizard.

c. Define the data set **ORGANICS** as a data source for the project. Use basic option in step 4 of the metadata advisor options.

d. Go through and finish the data step creation. Then add the data set to the diagram.

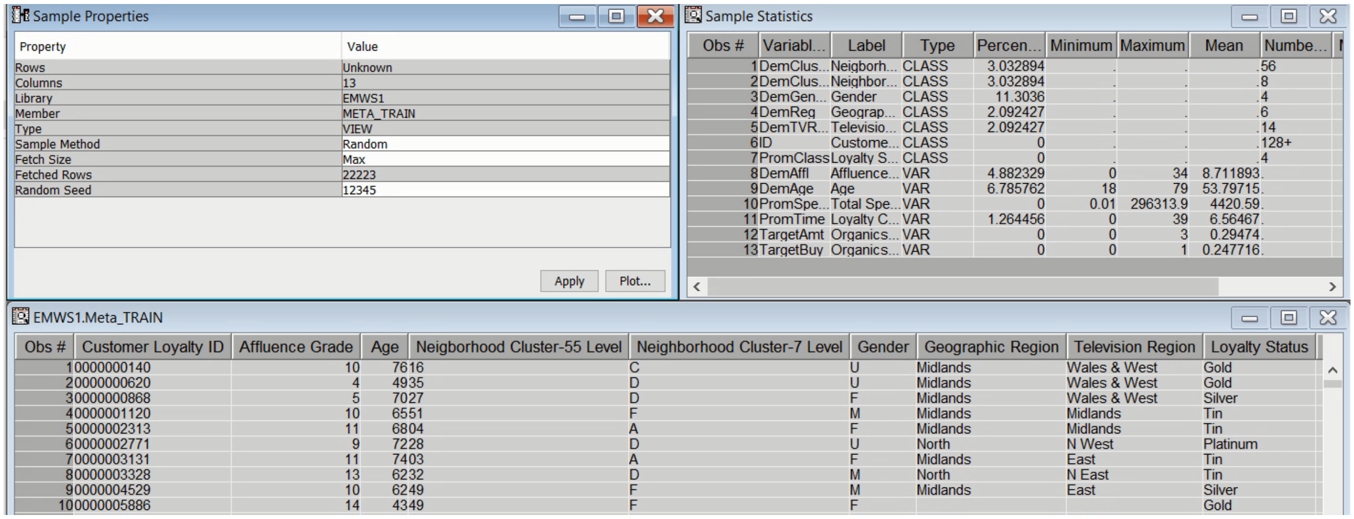
e. Attach a Metadata node and connect it with the data.

f. Check the model role and measurement level for each variable using the Metadata node with the model roles and measurement levels of the table printed above. Fix as needed (by clicking on role or level) any mismatch in the roles or measurement levels of variables in the above step.

g. Right-click on the data source in the diagram.

h. Choose edit variables and explore all variables. In the 'Sample Properties' window make sure you set sampling method to random and sample size to Max**. Include a screenshot of first 10**

observations in your report. **(0.5 Points)**



Sample Properties

Property	Value
Rows	Unknown
Columns	13
Library	EMWS1
Member	META_TRAIN
Type	VIEW
Sample Method	Random
Fetch Size	Max
Fetched Rows	22223
Random Seed	12345

Sample Statistics

Obs #	Variabl...	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...
1	DemClus...	Neigborh...	CLASS	3.032894	.	.	.	56
2	DemClus...	Neighbor...	CLASS	3.032894	.	.	.	8
3	DemGen...	Gender	CLASS	11.3036	.	.	.	4
4	DemReg	Geograp...	CLASS	2.092427	.	.	.	6
5	DemTVR...	Televisio...	CLASS	2.092427	.	.	.	14
6	ID	Custome...	CLASS	0	.	.	.	128+
7	PromClass	Loyalty S...	CLASS	0	.	.	.	4
8	DemAffl	Affluence...	VAR	4.882329	0	34	8.711893	.
9	DemAge	Age	VAR	6.785762	18	79	53.79715	.
10	PromSpe...	Total Spe...	VAR	0	0.01	296313.9	4420.59	.
11	PromTime	Loyalty C...	VAR	1.264456	0	39	6.56467	.
12	TargetAmt	Organics...	VAR	0	0	3	0.29474	.
13	TargetBuy	Organics...	VAR	0	0	1	0.247716	.

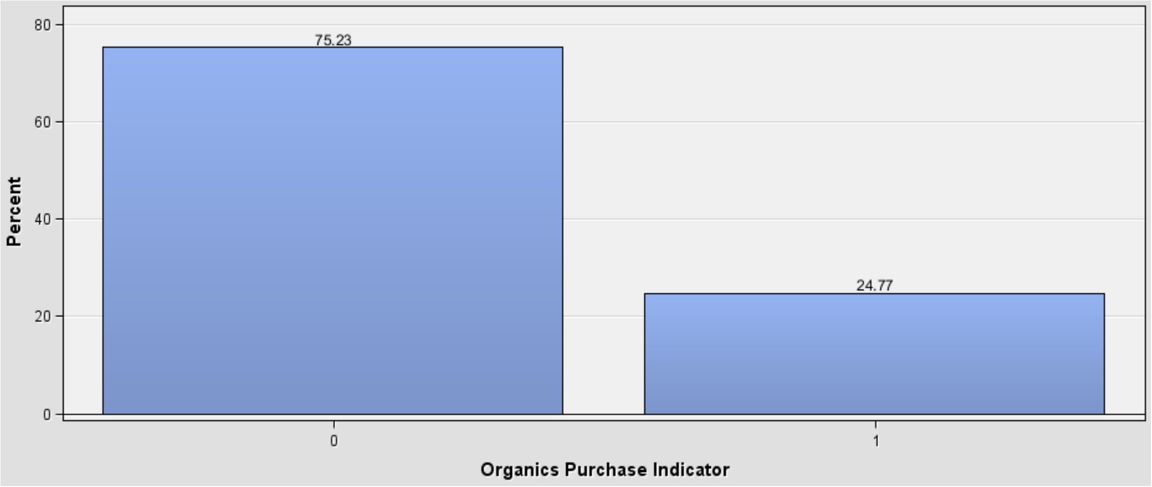
EMWS1.Meta_TRAIN

Obs #	Customer Loyalty ID	Affluence Grade	Age	Neigborhood Cluster-55 Level	Neighborhood Cluster-7 Level	Gender	Geographic Region	Television Region	Loyalty Status
1	0000000140	10	76	16	C	U	Midlands	Wales & West	Gold
2	0000000620	4	49	35	D	U	Midlands	Wales & West	Gold
3	0000000868	5	70	27	D	F	Midlands	Wales & West	Silver
4	0000001120	10	65	51	F	M	Midlands	Midlands	Tin
5	0000002313	11	68	04	A	F	Midlands	Midlands	Tin
6	0000002771	9	72	28	D	U	North	N West	Platinum
7	0000003131	11	74	03	A	F	Midlands	East	Tin
8	0000003328	13	62	32	D	M	North	N East	Tin
9	0000004529	10	62	49	F	M	Midlands	East	Silver
10	0000005886	14	43	49	F	F			Gold

- **i.** In the pop-up box, select the designated variable below and then click on the Explore button.
  - **2)** Select TargetBuy.
    - **a.** Create a frequency histogram for the variable TargetBuy.
    - **b.** Make sure the vertical axis is percentages and you display the percentage values in the histogram (hint: right-click on the graph…).
    - **c.** **Turn-in a copy of the histogram as a part of your deliverable.** **(1 Point)**

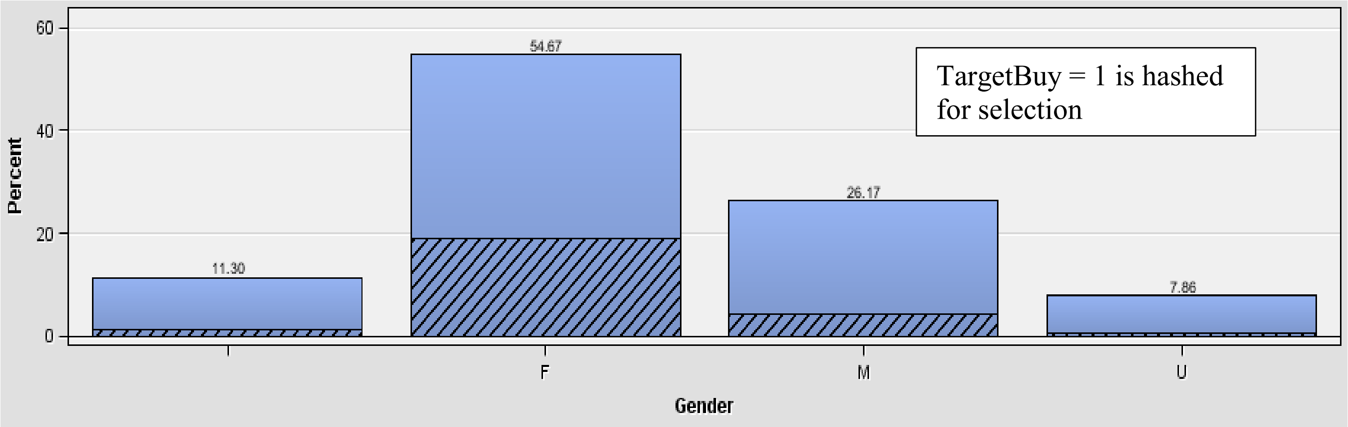
Solution: To change the frequency to percentage in the graph, right click on the histogram -

> select data options -> change response static to percentage from frequency. The percentages are skewed with Organic buyers accounting for 3 times fewer in this data. This may create some problems in model building because models tend to work best when the percentages are close to each other.



- **3)** Select DemGender.
  - **a.** Create a frequency bar chart for the variable DemGender.
  - **b.** **Comment on what you see in this histogram.**
  - **c.** **Turn-in a copy of the histogram as deliverable.** **(1 Point)**

Solution: There are more women in this data set (55% compared to 26% males). However, there is a difference in organic purchases when you look within genders. You can also see that there is a 3rd value of unknown gender and several with missing gender.



TargetBuy = 1 is hashed for selection

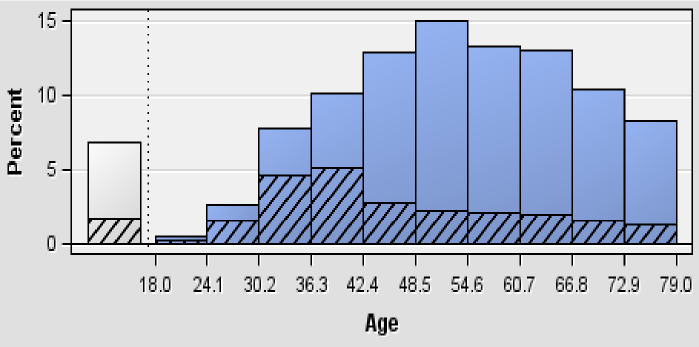
4) Select DemAge.

a. Create a frequency histogram for the variable DemAge.

b. Comment on what you see in this histogram.

c. Turn-in a copy of the histogram as deliverable. **(1 Point)**

Solution: The variable is left skewed. The distribution has a long left-tail. There are missing values (grey bar) in DemAge. When viewing targetbuy = 1 by DemAge, the highest percentge of organic purchasers moves from the 48.5-54.6 bin to the 36.3-42.4 bin. Those who purchase organic products are younger than the group as a whole.



5) Select PromSpend.

a. Create a frequency histogram for the variable PromSpend.

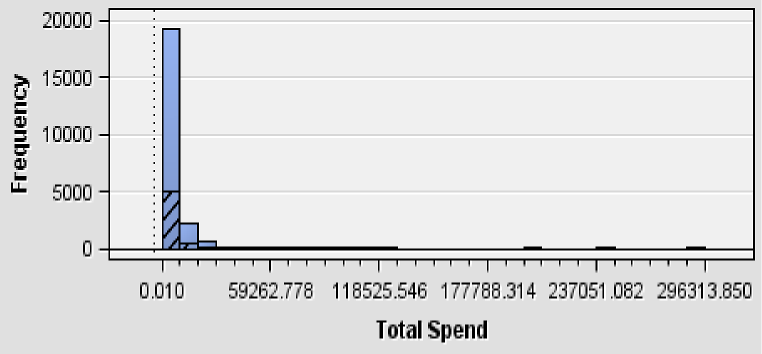
b. Change the number of bins to 30.

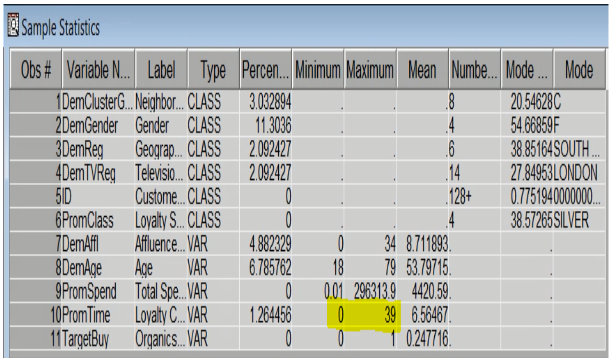
c. Is this variable right or left skewed?

d. Does that make sense based on the variable's description?

e. Turn-in a copy of the histogram as deliverable. **(1 Point)**

*Solution***:** Promotional Spend is right skewed with a very long right tail. This makes sense that more people would spend small amounts of money while only a few will spend large amounts. Further, the largest optimum number of bins would be **39 – 0 = 39** (As 39 is the highest value and 0 is the lowest value for thisvariable)





Obs #	Variable N...	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	Mode ...	Mode
1	DemClusterG...	Neighbor...	CLASS	3.032894	.	.	.	8	20.54628	C
2	DemGender	Gender	CLASS	11.3036	.	.	.	4	54.66859	F
3	DemReg	Geograp...	CLASS	2.092427	.	.	.	6	38.85164	SOUTH ...
4	DemTVReg	Televisio...	CLASS	2.092427	.	.	.	14	27.84953	LONDON
5	ID	Custome...	CLASS	0	.	.	.	128+	0.775194	0000000...
6	PromClass	Loyalty S...	CLASS	0	.	.	.	4	38.57265	SILVER
7	DemAffl	Affluence...	VAR	4.882329	0	34	8.711893	.	.	
8	DemAge	Age	VAR	6.785762	18	79	53.79715	.	.	
9	PromSpend	Total Spe...	VAR	0	0.01	296313.9	4420.59	.	.	
10	PromTime	Loyalty C...	VAR	1.264456	0	39	6.56467	.	.	
11	TargetBuy	Organics...	VAR	0	0	1	0.247716	.	.	

2. **Summarize your findings for the supermarket manager in a report that addresses each of the items below including screenshots of supporting information. Consider these when writing your report (6 Points)**

   a. How would you efficiently display or examine the variables in the data set?

   b. Are there any variables that have missing data? How would this affect any future predictive analysis?

   c. Are the variables normally distributed? How do you deal with these data to prepare them for predictive modeling?

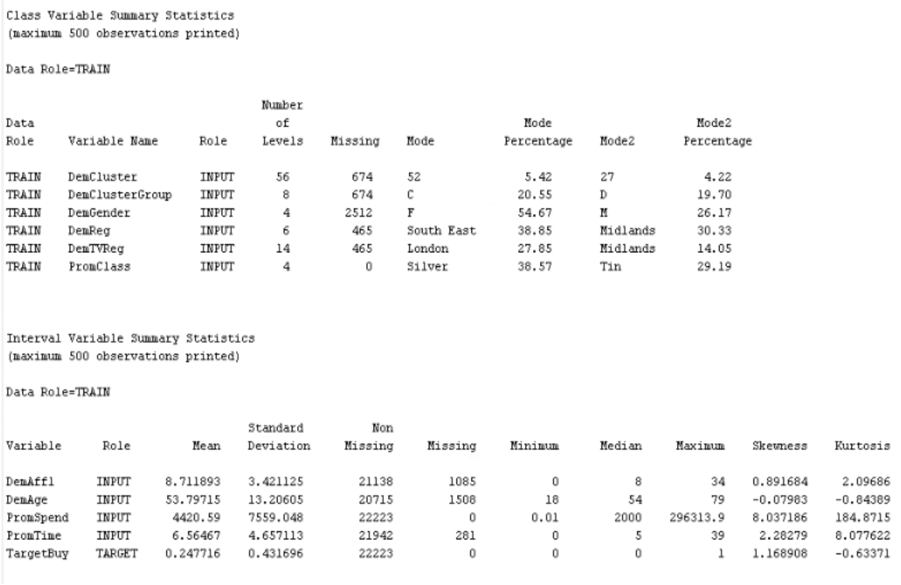
   d. Do all the values seem like they are realistic or do you need to make adjustments to any particular variable? Use appropriate techniques in Enterprise Miner to fix any issues and re-run the analysis.

   e. What can you say from this analysis about those who buy organics and those who do not?

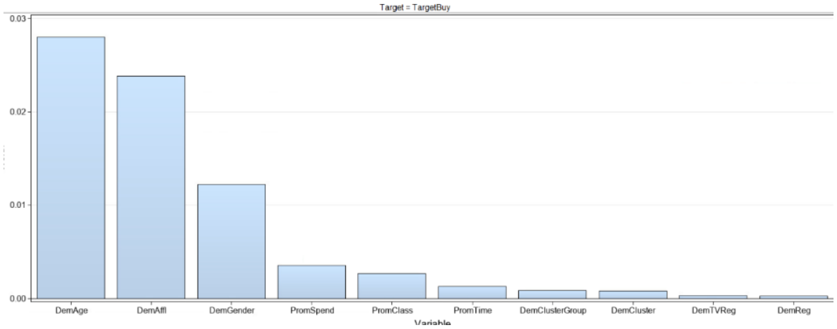
   f. How can the manager use the information to help design a marketing plan?

Solution:

To further examine the variables and display their core information more efficiently, we investigated summary statistics. In these statistics we look for outliers with extreme minimums and maximums, missing data, and observe for variables that may need to be scaled or transformed for further analysis. We also continue to look at this data in histograms as this displays skewness as well as allows us to quickly know information about the data. The below screenshot of information shows summary statistics. We observe again that gender and age are missing quite a few values. We also see that the region most represented in the data is the South East and the prominent loyalty class is Silver. Using stat explorer, we also observe the higher correlations exists between the affluence of the customer and the age of the customer between each and the target variable. The higher the affluence, the more likely they are to purchase organic product. Whereas, as the age of the customer goes up the less likely they are to purchase organic products. The bar chart below shows the importance of the variables in relationship to the target variable.

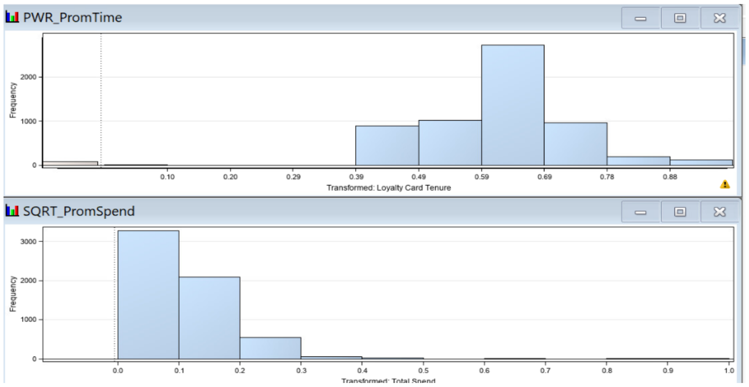


```
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                   Number
Data                                 of                            Mode                     Mode2
Role     Variable Name     Role    Levels   Missing   Mode        Percentage   Mode2       Percentage

TRAIN    DemCluster        INPUT     56        674    52              5.42     27             4.22
TRAIN    DemClusterGroup   INPUT      8        674    C              20.55     D             19.70
TRAIN    DemGender         INPUT      4       2512    F              54.67     M             26.17
TRAIN    DemReg            INPUT      6        465    South East     38.85     Midlands      30.33
TRAIN    DemTVReg          INPUT     14        465    London         27.85     Midlands      14.05
TRAIN    PromClass         INPUT      4          0    Silver         38.57     Tin           29.19


Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                  Standard      Non
Variable     Role        Mean    Deviation   Missing    Missing    Minimum    Median     Maximum    Skewness    Kurtosis

DemAffl      INPUT    8.711893    3.421125     21138       1085          0         8          34    0.891684     2.09686
DemAge       INPUT    53.79715    13.20605     20715       1508         18        54          79    -0.07983    -0.84389
PromSpend    INPUT     4420.59    7559.048     22223          0       0.01      2000    296313.9    8.037186    184.8715
PromTime     INPUT     6.56467    4.657113     21942        281          0         5          39     2.28279    8.077622
TargetBuy    TARGET   0.247716    0.431696     22223          0          0         0           1    1.168908    -0.63371
```



Exploring in multiplot we again observed that PromSpend (how much a customer spent) is skewed, as well as observed that PromTime (How much time the customer has been in the loyalty program) is skewed. We decided that transformation to be performed on these two variables. We also decided to use the impute node in SAS EM and impute the age category with the mean. After rerunning the analysis with the time variable is more evenly distributed and the spend variable still appears skewed, but the value is now closer to 0. We suggest that management use this information to influence decision making on who to expect to buy organic products and who does not. For example, running a decision tree in SAS EM with these variables will give us an importance and inform how to design a marketing plan. Below are the variables by importance in this analysis.

Managers should utilize age and affluence to design marketing strategy. For example, using the information from the analysis above, if a promotion were given to customers younger than the median age of 45 and whom have a higher affluence, management could expect more response to the campaign. Considering the gender distribution of the loyalty customers, management could also expect a response to a campaign to their women customers based on the makeup of the amount of the loyalty customers.



3. This part is a stand-alone exercise using file import node.
   a. First, use the file import node to import the Excel data file **Smalldata**.
   b. Run the file import node and then answer the following questions.

6) Does any variable have a role of rejected? (0.5 Points)

Solution: One variable was rejected after the file import node was run on the Excel file. Because there are more than 50 levels in Zip_Code, the variable was rejected.

7) If yes, can you guess why it is rejected? (0.5 Points)

Solution: Because there are more than 50 levels in Zip_Code, the variable was rejected. This could be changed using a metadata node should the variable be needed for anlaysis.

   c. There are several variables that SAS EM may have assigned a measurement level of **<u>Interval</u>**. But, these should really be **<u>binary</u>**.

8) Which are these variables and why should they have binary measurement level? (**1 Point**)

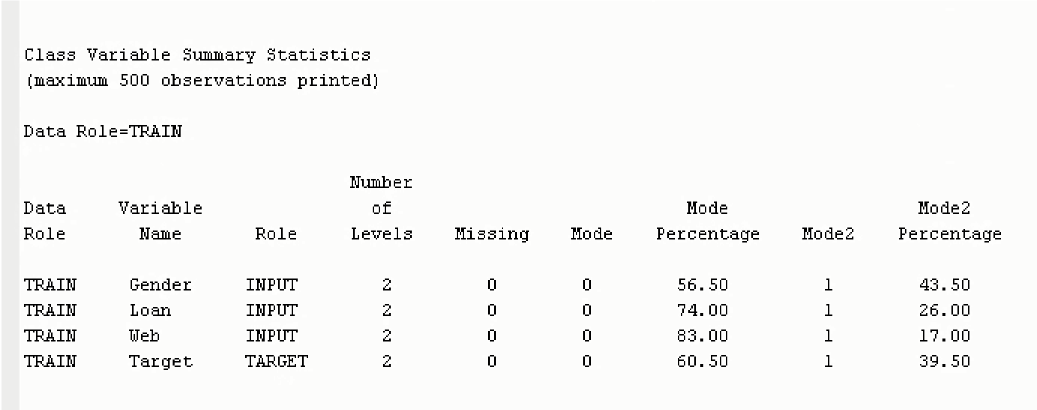
Solution: Gender, Target, Web, and Loan should all be set to binary as indicated by their minimum and maximum values and the histograms during variable exploration.

   d. Fix their levels to binary before going forward.
   e. Attach a StatExplore node to the File Import node and run the Stat Explore node.

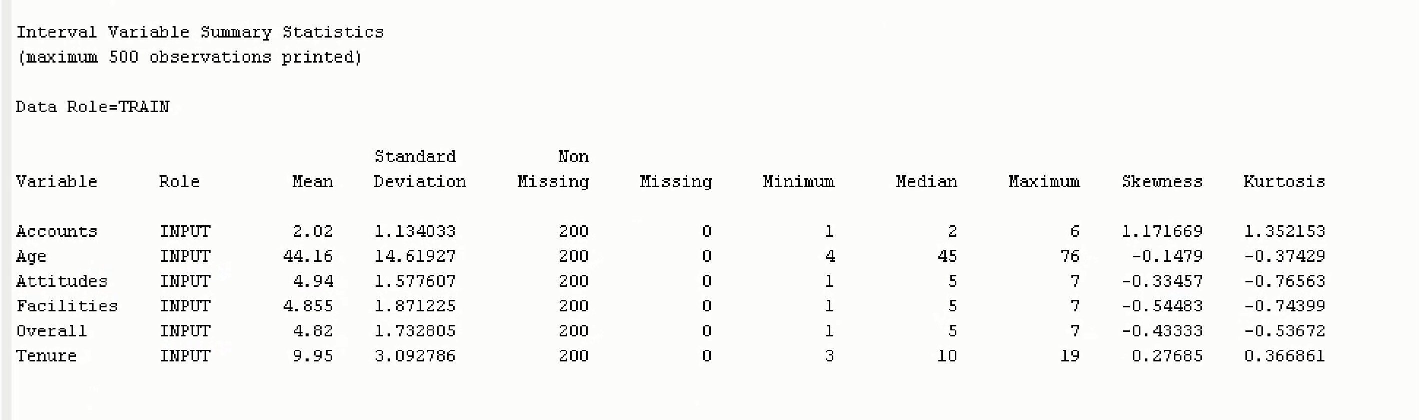
9) Report the class variable summary statistics and the Interval variable summary statistics

from the output window of the StatExplore node. (1 Point)

Solution:



```
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                               Number
Data      Variable               of                                 Mode                   Mode2
Role        Name       Role    Levels    Missing    Mode    Percentage    Mode2    Percentage

TRAIN     Gender     INPUT        2         0         0        56.50         1         43.50
TRAIN     Loan       INPUT        2         0         0        74.00         1         26.00
TRAIN     Web        INPUT        2         0         0        83.00         1         17.00
TRAIN     Target     TARGET       2         0         0        60.50         1         39.50
```



```
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                     Standard        Non
Variable      Role          Mean    Deviation    Missing    Missing    Minimum    Median    Maximum    Skewness    Kurtosis

Accounts      INPUT         2.02     1.134033        200          0          1         2          6    1.171669    1.352153
Age           INPUT        44.16     14.61927        200          0          4        45         76     -0.1479    -0.37429
Attitudes     INPUT         4.94     1.577607        200          0          1         5          7    -0.33457    -0.76563
Facilities    INPUT        4.855     1.871225        200          0          1         5          7    -0.54483    -0.74399
Overall       INPUT         4.82     1.732805        200          0          1         5          7    -0.43333    -0.53672
Tenure        INPUT         9.95     3.092786        200          0          3        10         19     0.27685    0.366861
```

Deliverables:

✓ *As you complete the exercise, create a report in Microsoft Word and in this report answer the questions in the exercise description.*

✓ *Make sure you comment or explain and not just provide snapshots of data.*

✓ *Limit your report to no more than 7* ***pages*** *including tables and diagrams.*

✓ *Copy and paste supporting tables/diagrams as needed to justify any of your answer. You may need to shrink your table/ diagrams but please ensure they are readable.*

✓ *Make sure you print your name, student ID#, student email on the cover page of the report and turn-in the report as communicated by your instructor.*

✓ *Please also put a running header/footer with your name, on each page of your exercise solution report.*

Failure to follow these instructions will result in deduction of points