## horizontal line



**Predictive Analysis**

**Karan Ashok**

# Initial Data Exploration

A supermarket is beginning to offer a line of organic products. The supermarket’s management would like 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 have now collected data that includes whether these customers have purchased any of the organic products.

The ORGANICS data set contains over 22,000 observations and 18 variables. The variables in the data set are shown below with the appropriate roles and levels.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Measurement Level** | **Description** |
| CUSTID | ID | Nominal | Customer loyalty identification number |
| GENDER | Input | Nominal | M = male, F = female, U = unknown |
| DOB | Rejected | Interval | Date of birth |
| EDATE | Rejected | Unary | Date extracted from the daily sales data base |
| AGE | Input | Interval | Age, in years |
| AGEGRP1 | Input | Nominal | Age group 1 |
| AGEGRP2 | Input | Nominal | Age group 2 |
| TV\_REG | Input | Nominal | Television region |
| NGROUP | Input | Nominal | Neighborhood group |
| NEIGHBORHOOD | Input | Nominal | Type of residential neighborhood |
| LCDATE | Rejected | Interval | Loyalty card application date |
| LTIME | Input | Interval | Time as loyalty card member |
| ORGANICS | Target | Interval | Number of organic products purchased |
| BILL | Input | Interval | Total amount spent |
| REGION | Input | Nominal | Geographic region |
| CLASS | Input | Nominal | Customer loyalty status: tin, silver, gold, or platinum |
| ORGYN | Target | Binary | Organics purchased? 1 = Yes, 0 = No |
| AFFL | Input | Interval | Affluence grade on a scale from 1 to 30 |

# 

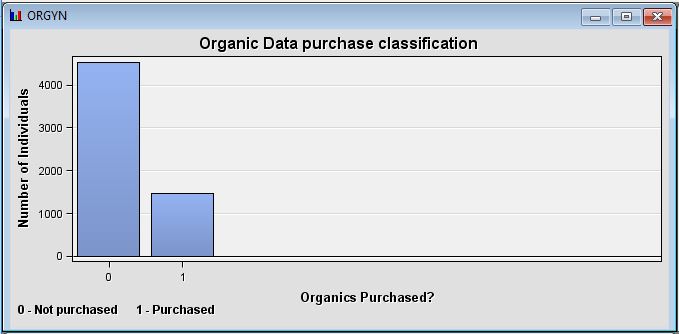
# 

# Questions

## What is the proportion of individuals who purchased organic products? Display the screen shot showing the bar chart for the target variable. Aside from the corrections identified below, do you have any concerns about any of the other variables in the data set? Are there any variables that should not be included as input variables in your analysis?

Proportion of individuals who purchased organic products = (Number of individuals who purchased organic products / (Total Number of Individuals) = 1460/6000 = 0.243

24.3% individuals purchased the organic products.

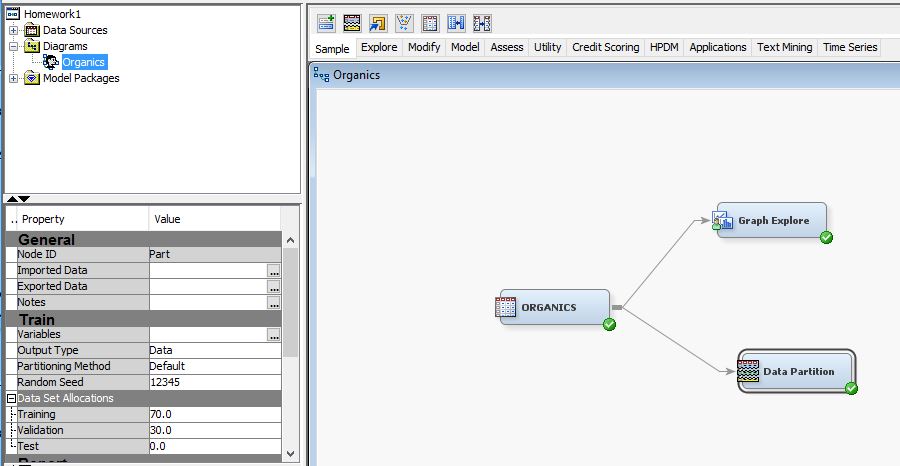


Variables such as GENDER, AGE, REGION, AFFL could have been used to add extra classification parameters to our analysis.

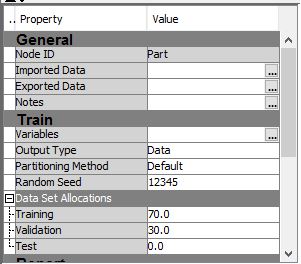
Variables such as BILL, CLASS, TV\_REG can be rejected.

## Partition your dataset into training and validation sets as follows: 70% of the data for training and 30% for validation. Display a screenshot of your diagram at this point and a screenshot showing the place where you made changes to the partitioning percentages.

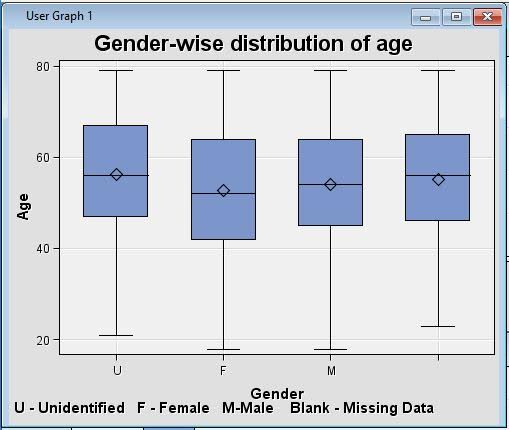
Please find below the screenshot our diagram at this point:



Please find below the screenshot showing the place where we made the partition changes:



## Display and explain the boxplot for the GENDER-wise distribution of AGE. Use the Graph Explore node to do this.



The **box plot** is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum. Lowest line represents minimum value, highest line represents maximum value, IQR is the difference between Third and First quartile which is represented by the box, Line inside the box represents median and the quadrilateral in the box represents the mean value. Below are the details for each of the box plot represented above:

**Gender - Unidentified**

Minimum age - 21

Maximum age - 79

Mean age - 56.21

First Quartile - 47

Median - 56

Third Quartile - 67

**Gender - Female**

Minimum age - 18

Maximum age - 79

Mean age - 52.78

First Quartile - 42

Median - 52

Third Quartile - 64

**Gender - Male**

Minimum age - 18

Maximum age - 79

Mean age - 54.14

First Quartile - 45

Median - 54

Third Quartile - 64

**Gender - Blank**

Minimum age - 23

Maximum age - 79

Mean age - 55.10

First Quartile - 46

Median - 56

Third Quartile - 65