

Tutorial of the steps followed in the first practical class with SAS Miner Guide

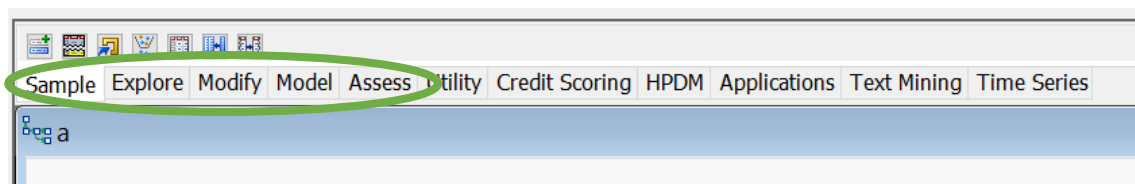
What is SAS Miner?

"SAS Enterprise Miner streamlines the data mining process to create highly accurate predictive and descriptive models based on analysis of vast amounts of data from across an enterprise. Data mining is applicable in a variety of industries and provides methodologies for such diverse business problems as fraud detection, householding, customer retention and attrition, database marketing, market segmentation, risk analysis, affinity analysis, customer satisfaction, bankruptcy prediction, and portfolio analysis."



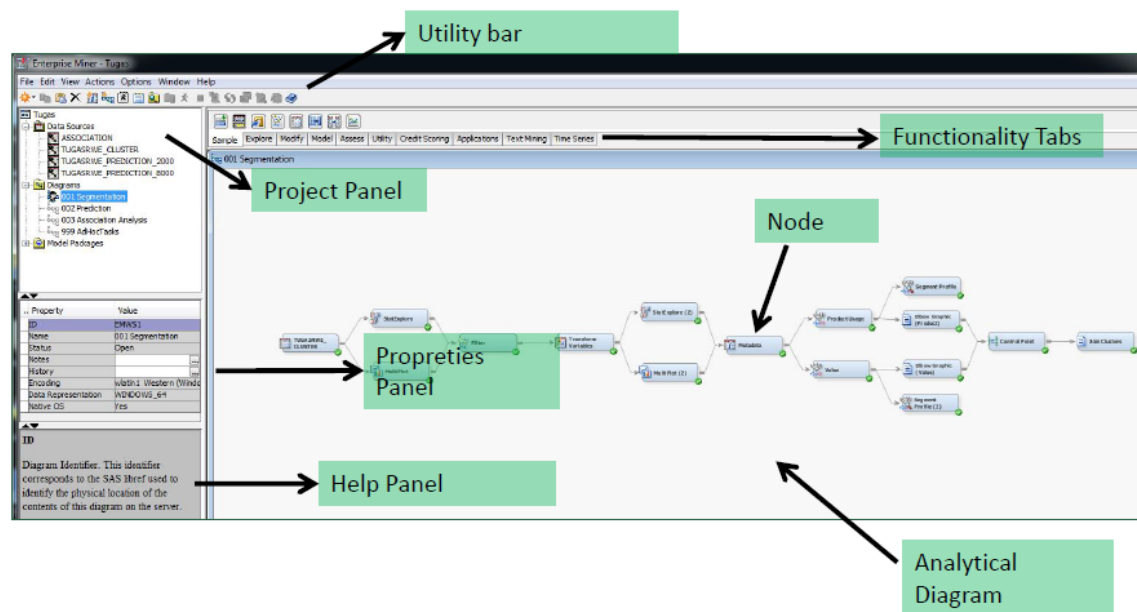
The SEMMA Approach

In SAS Enterprise Miner, the data mining process follows the SEMMA Approach.

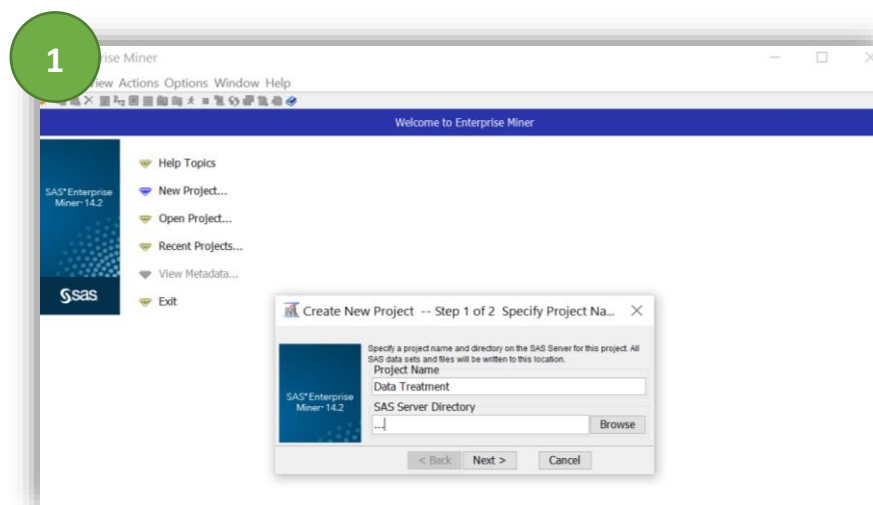


- **Sample:** Sample the data by creating one or more datasets. Datasets should be large enough to contain the significant information, yet small enough to process, especially where computing resources are (more) limited. This step includes the use of data preparation tools for data import, merge, append, and filter, as well as statistical sampling techniques;
- **Explore:** Explore the data by searching for anticipated (hypothesized) relationships, unanticipated patterns, and anomalies in order to improve awareness and improved understanding about the subject under analysis. This step includes the use of tools for statistical reporting and graphical exploration, variable selection methods, and variable clustering;
- **Modify:** Modify the data by creating, selecting, and transforming the variables to focus the model selection process. This step includes the use of tools for defining transformations, missing value handling, value recoding, and interactive binning;
- **Model:** You model the data by using the analytic techniques to search for a combination of the data that reliably predicts what is intended. This step includes the use of techniques such as linear and logistic regression, decision trees, neural networks, partial least squares, LARS and LASSO, nearest neighbor, and importing models defined by other users or even outside SAS Enterprise Miner;
- **Assess:** Assessment of the alternatives of the model/segmentation in the project, with objective of choosing the one that better serves the purposes; This step includes the use of tools for comparing models and computing new fit statistics, cutoff analysis, decision support, report generation, and score code management.

Environment Overview



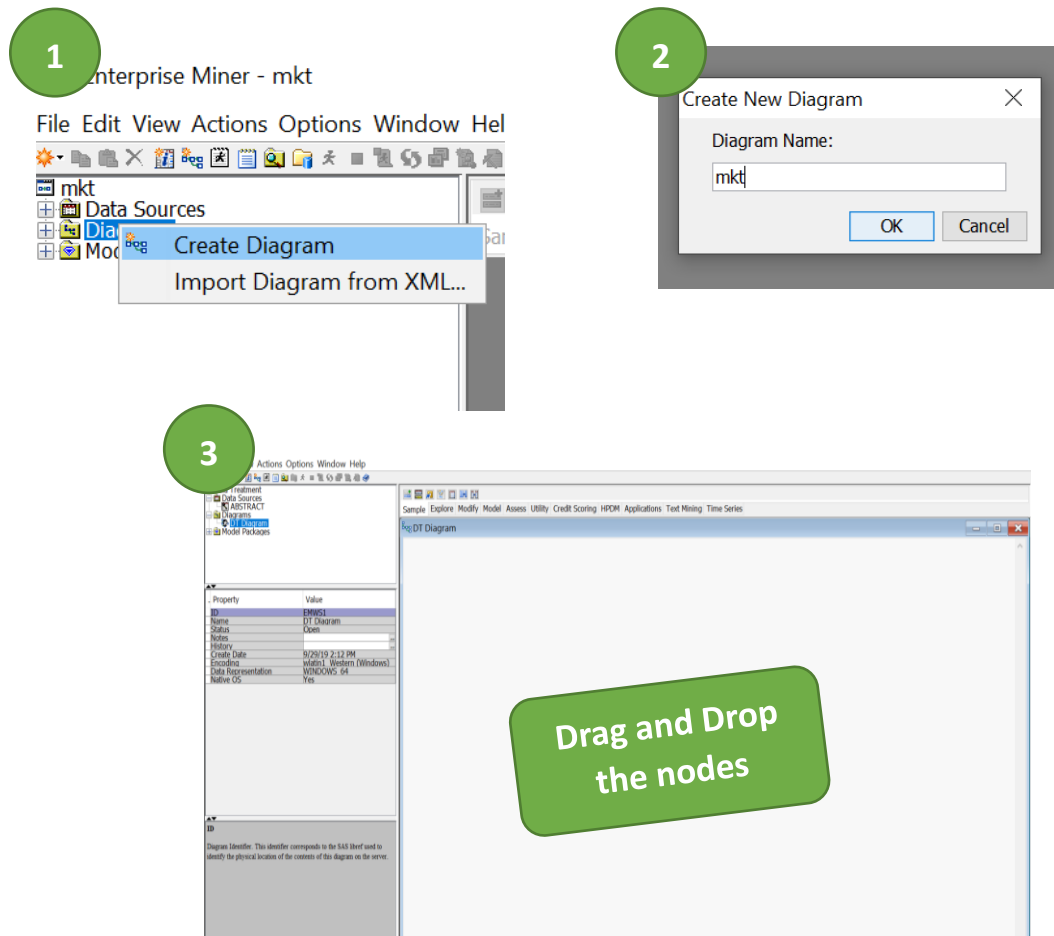
Create a new SAS Enterprise Miner Project



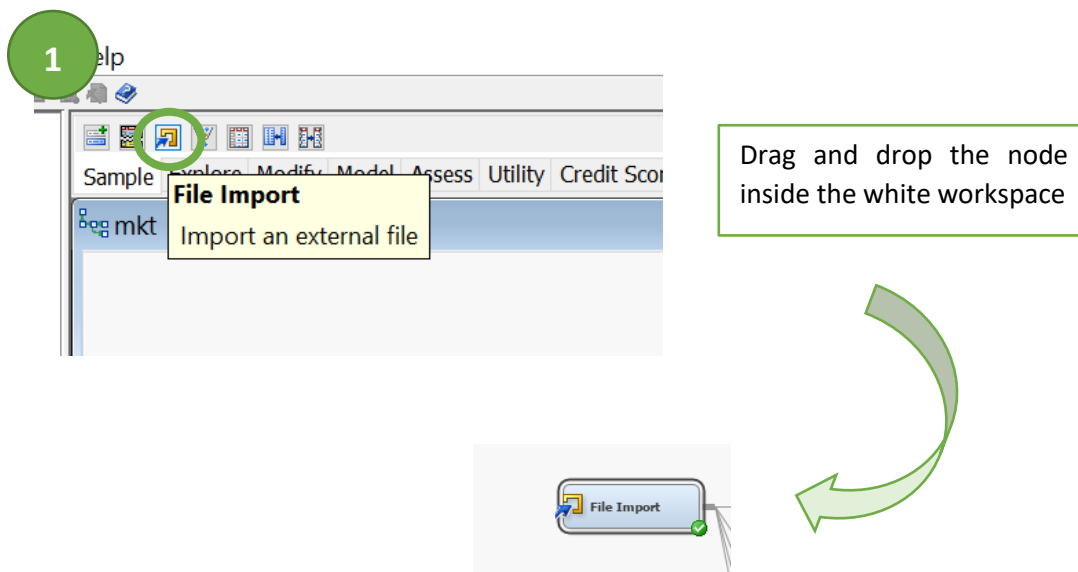
This creates a folder in the local you have defined as SAS Server Directory. Every action that is done in SAS Miner is recorded in that folder

- DataSourcees
- HPDM
- Meta
- Reports
- System
- Workspaces
- emproject.properties
- project.emp

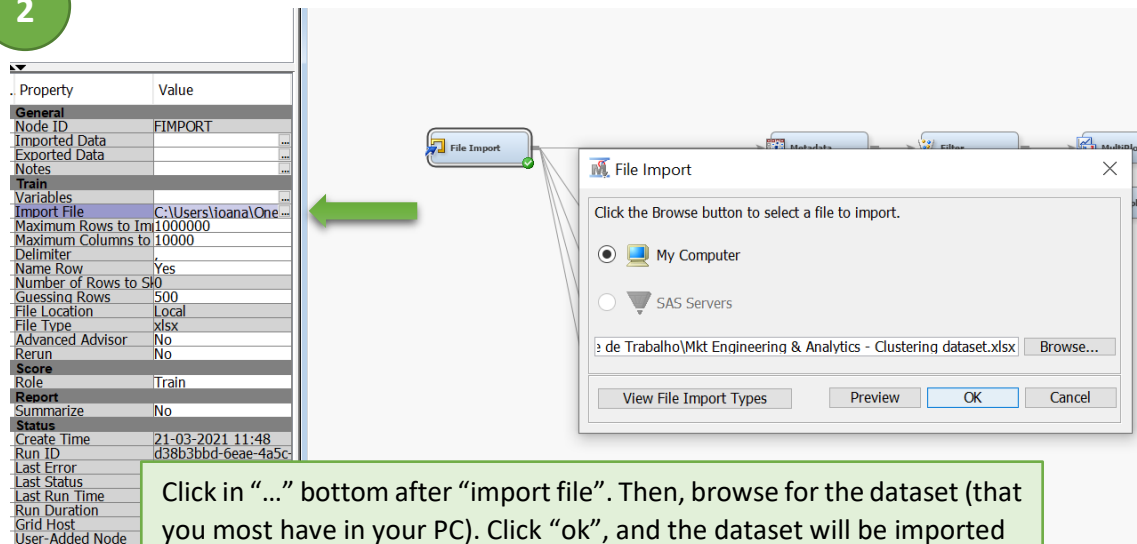
Create a new SAS Enterprise Miner process flow diagram



1- File import (to import the dataset in the diagram)

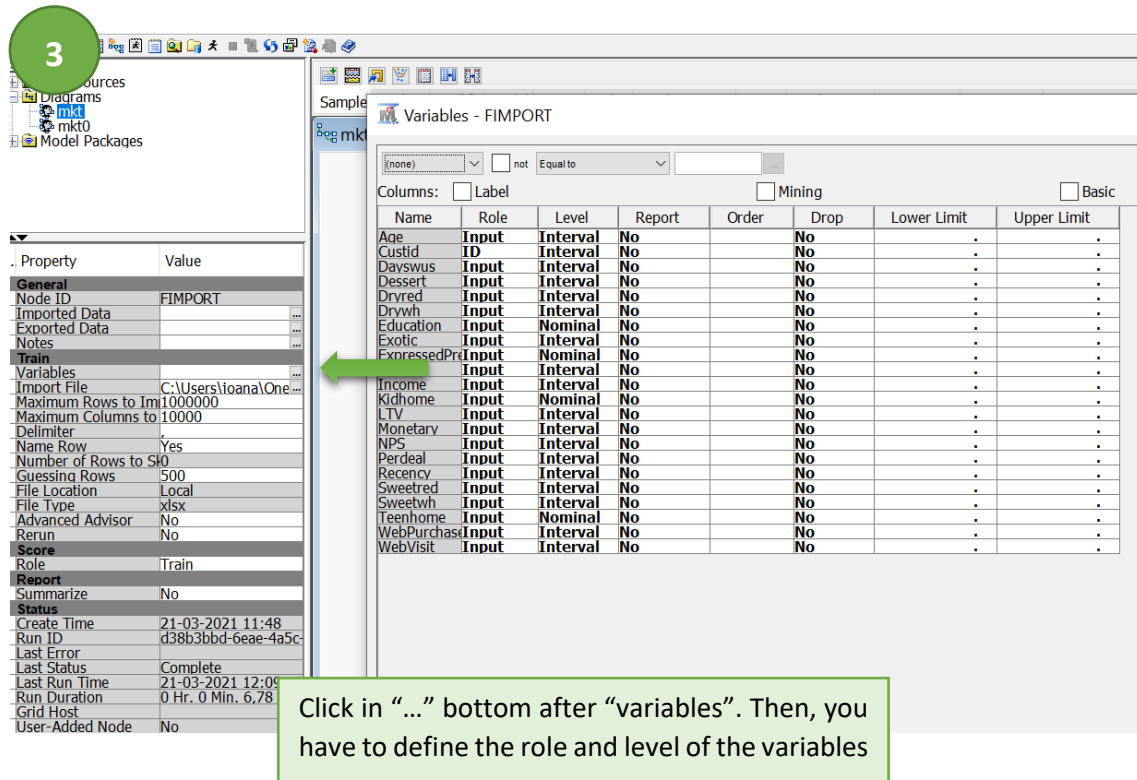


2



Click in "... " bottom after "import file". Then, browse for the dataset (that you most have in your PC). Click "ok", and the dataset will be imported

3



Click in "... " bottom after "variables". Then, you have to define the role and level of the variables

Variable Roles:

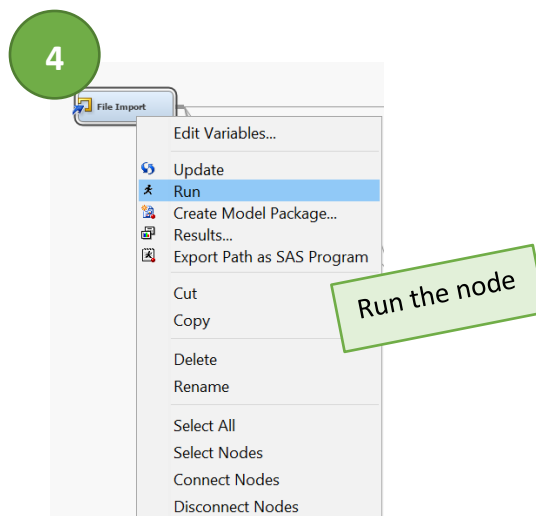
- ID: Identify the ID variable in the dataset. It must present unique values per observation/record and it is not used in any calculations;
- Input: Variables classified as "input" will be the ones used for segmentation/modelling tasks;
- Target: Dependent variable of the problem, i.e., the one we are going to try understand or predict. Naturally, is mandatory for predictive tasks;

- Rejected: These variables are excluded from subsequent analysis/tasks;

(...)

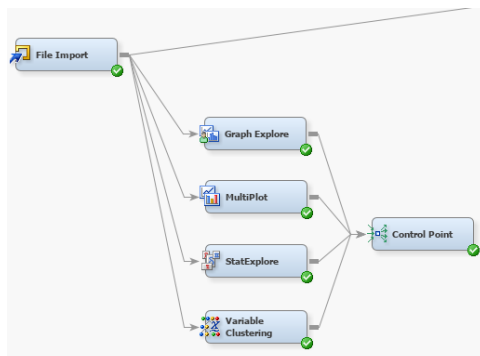
Variable levels:

- Binary: Variables that assume only two values, numeric or character (e.g., "Yes" or "No" / 1 or 0);
- Nominal: Numeric or character as a means of separating properties or elements into non-sortable different classes or categories (e.g., eye colors);
- Ordinal: where only comparisons such as "greater", "less", or "equal" between measurements are possible. Can be both numeric or string (e.g., year of birth);
- Interval: Numeric variables used when it is possible to meaningful have "ratios".



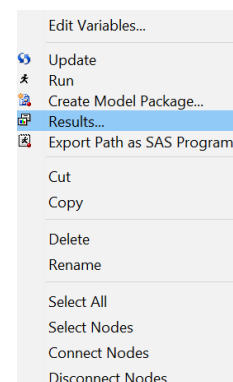
2- Explore the data

Typically, we use the stat explore node, multiplot node, graph explore node and variable clustering node to explore the data (descriptive statistics, graphics, ...). You can find these nodes on the explore tab. Then, just drag and drop them.



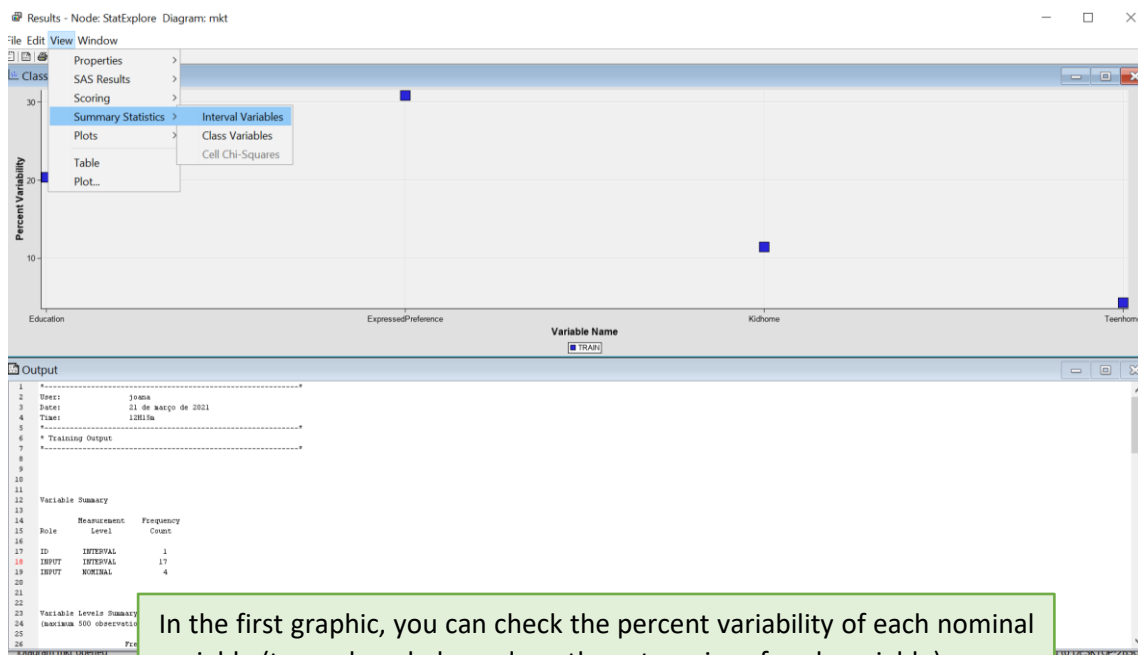
You can use a control point to run all nodes once. Just go to the utility tab, drop the control point node and run it.

The nodes are descriptive, so you don't need to change nothing in the properties box.



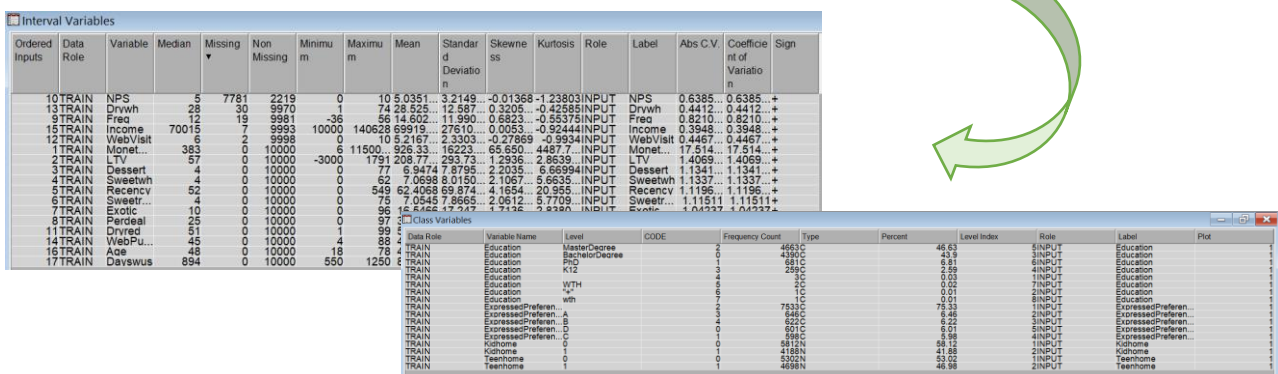
Just click in the "results" in each node to check the results of them

a) Stat explore

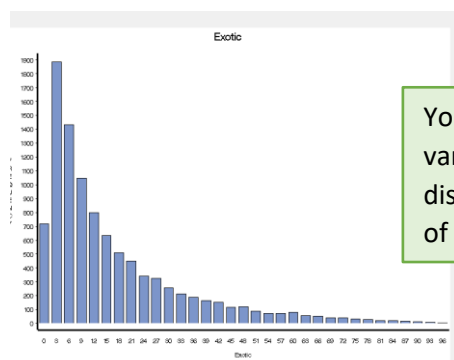


In the first graphic, you can check the percent variability of each nominal variable (to see how balanced are the categories of each variable).

You can also go to “view”, click “summary statistics” and then choose “interval variables” or “class variables” to check some statistics of the variables

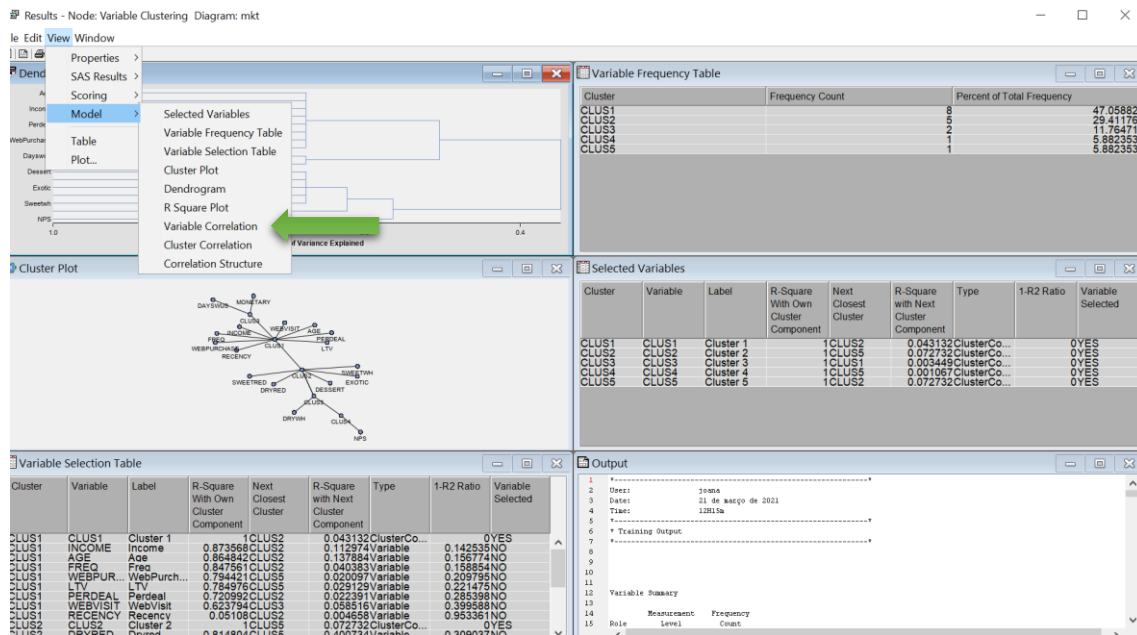


b) Multiplot

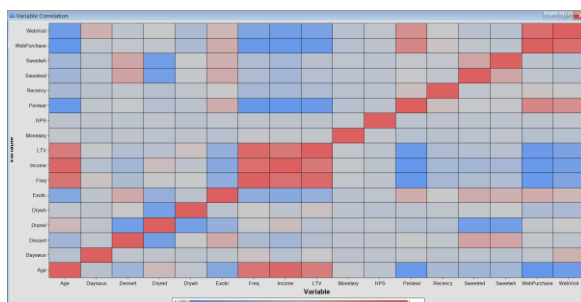


You can check the histogram of each variable. Looking for it you can examine the distribution, the missing values and outliers of each variable

c) Variable clustering



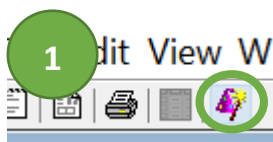
This node is mainly used as a way to reduce the dimensionality of the data (creating clusters of variables). However, in this phase, we will use it just to check the correlation between variables. For that, just click “view”, “model”, and “variable correlation”.



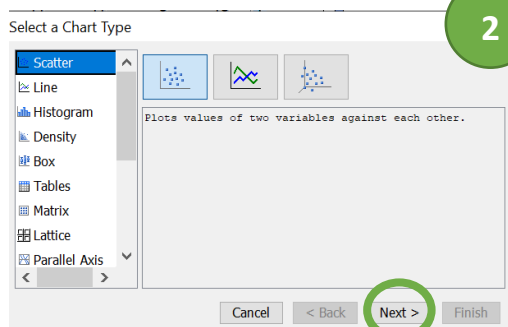
“very” red: close to 1

“very” blue: close to -1

d) Graph explore



We will use this node to create other graphics that might be useful to understand our data.



One interesting graphic is the scatter plot that allow you to check for multidimensional outliers

3

Chart Roles

Missing required roles: Y.

Use default assignments

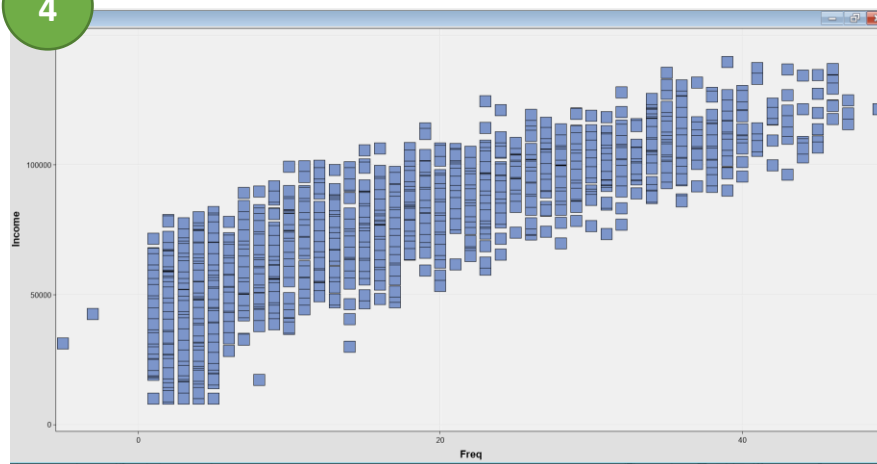
Variable	Role	Type	Description	Format
Age		Numeric	Age	BEST.
Custid		Numeric	Custid	BEST.
Dayswus		Numeric	Dayswus	BEST.
Dessert		Numeric	Dessert	BEST.
Dryred		Numeric	Dryred	BEST.
Drywh		Numeric	Drywh	BEST.
Education		Character	Education	\$14.
Exotic		Numeric	Exotic	BEST.
ExpressedPreference		Character	ExpressedPreference	\$1.
Freq	X	Numeric	Freq	BEST.
Income	None	Numeric	Income	BEST.
Kidhome	None	Numeric	Kidhome	BEST.
LTV	X	Numeric	LTV	BEST.

☐ Allow multiple role
 Group
 Group Index
 Color
 Tip

Cancel < Back Next > Finish

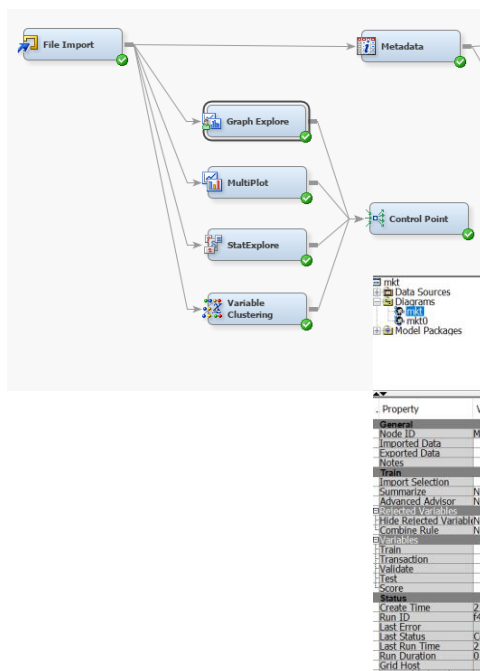
After selecting the graphic, define which variable will appear in the X and Y axis. Then, click "finish"

4



3- Metadata node

Sometimes, in the middle of the project you may want to change the role of some variables (e.g. to reject some of your analysis). As such, you may simple go to the utility tab and drag and drop the metada node and make the changes.

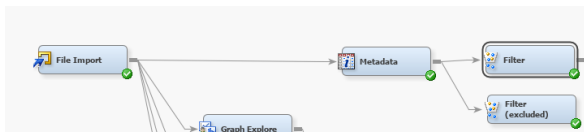


Just click in "..." bottom after "train" and then put the new role and level.

Sample Variables - Meta

Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report
Age	N	Default	Input	Default	Interval	Default	Default	Default
Custid	N	Default	ID	Default	Interval	Default	Default	Default
Dayswus	N	Default	Input	Default	Interval	Default	Default	Default
Dessert	N	Default	Input	Default	Interval	Default	Default	Default
Dryred	N	Default	Input	Default	Interval	Default	Default	Default
Drywh	N	Default	Input	Default	Interval	Default	Default	Default
Education	N	Default	Input	Default	Interval	Default	Default	Default
Exotic	N	Default	Input	Default	Interval	Default	Default	Default
ExpressedPref	N	Default	Input	Default	Interval	Default	Default	Default
Freq	N	Default	Input	Default	Interval	Default	Default	Default
Income	N	Default	Input	Default	Interval	Default	Default	Default
Kidhome	N	Default	Input	Default	Interval	Default	Default	Default
LTV	N	Default	Input	Default	Interval	Default	Default	Default
Monetary	N	Default	Input	Default	Interval	Default	Default	Default
NPS	N	Default	Input	Default	Interval	Default	Default	Default
Referred	N	Default	Input	Default	Interval	Default	Default	Default
Sweetwh	N	Default	Input	Default	Interval	Default	Default	Default
Teenhome	N	Default	Input	Default	Interval	Default	Default	Default
WebPurchase	N	Default	Input	Default	Interval	Default	Default	Default
WebVisit	N	Default	Input	Default	Interval	Default	Default	Default

4- Filter node (to filter the outliers)



Just go to the sample tab and drag and drop a filter node. Connect always to the previous nodes, to assure that the data running in the node is the same that you have changed before

In the properties box you can find a lot of definitions.

Indicate whether you want to export a data set containing filtered or excluded observations.

Excluded: Export the excluded observations.

Filtered: (default setting) Export the included observations.

All: Export all observations.

Indicate whether you want to filter the training data set or all imported data sets.

Applies a specific filter (you choose) in a variable-by variable basis (instead of applying a default filter to all class variables).

No if you want to filter out observations that contain missing values for class variables. The default setting for the Keep Missing Values property is Yes.

Property	Value
General	
Node ID	Filter
Imported Data	...
Exported Data	...
Notes	...
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data	Yes
Class Variables	
Class Variables	
Default Filtering	None
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency	1
Minimum Cutoff for	0.01
Maximum Number	25
Interval Variables	
Interval Variables	...
Default Filtering	User-Specified Limits
Keep Missing Values	Yes
Tuning Parameter	...
Score	
Create Score Code	Yes
Update Measurement	No
Status	
Create Time	12/6/19 10:58 AM
Run ID	e7b58bf6-0a93-4e...
Last Error	
Last Status	Complete
Last Run Time	12/6/19 12:45 PM
Run Duration	0 Hr. 0 Min. 4.89
Grid Host	
General	

Specify the method that you want to use to filter class variables.

Rare Values (Count): Drop rare levels that have a count less than the level that you specify in the Minimum Frequency Cutoff property.

Rare Values (Percentage): (default setting) Drop rare levels that occur in proportions lower than the percentage that you specify in the Minimum Cutoff for Percentage property.

None: No class variable filtering is performed.

Specify the method that you want to use to filter interval variables. The Default Filtering Method applies only to input variables.

Mean Absolute Deviation (MAD): eliminates values that are more than n deviations from the median. You specify the threshold value for the number of deviations, n, in the Cutoff for MAD property.

User-Specified Limits: specifies a filter for observations that is based on the interval values that are displayed in the Filter Lower Limit and Filter Upper Limit columns of your data table. You specify these limits in the Variables table.

Metadata Limits: are the lower and upper limit attributes that you can specify when you create a data source or when you are modifying the Variables table of an Input Data node on the diagram workspace.

Extreme Percentiles: filters values that are in the top and bottom pth percentiles of an interval variable's distribution. You specify the upper and lower threshold value for p in the Cutoff Percentiles for Extreme Percentiles property.

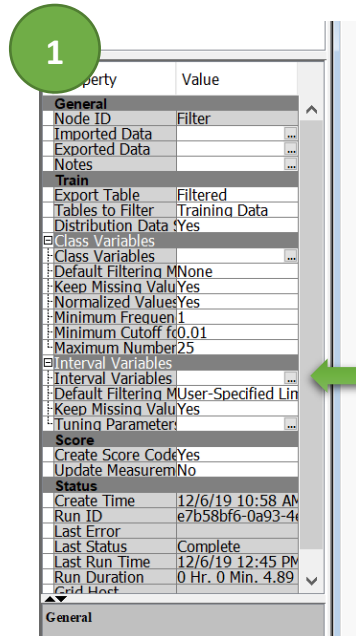
Modal Center: eliminates values that are more than n spacings from the modal center. You specify the threshold value for the number of spacings, n, in the Cutoff Percentiles for Modal Center property.

Standard Deviations from the Mean: (default setting) filters values that are greater than or equal to n standard deviations from the mean. You must use the Cutoff Percentiles for Standard Deviations property to specify the threshold value that you want to use for n.

None: do not filter interval variables.

The most important steps are:

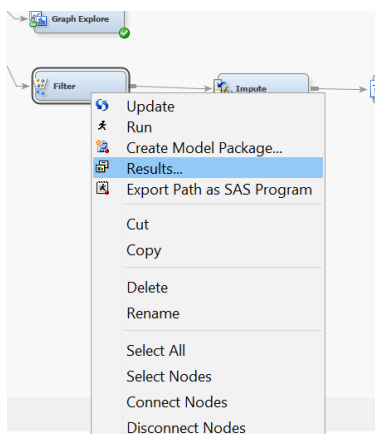
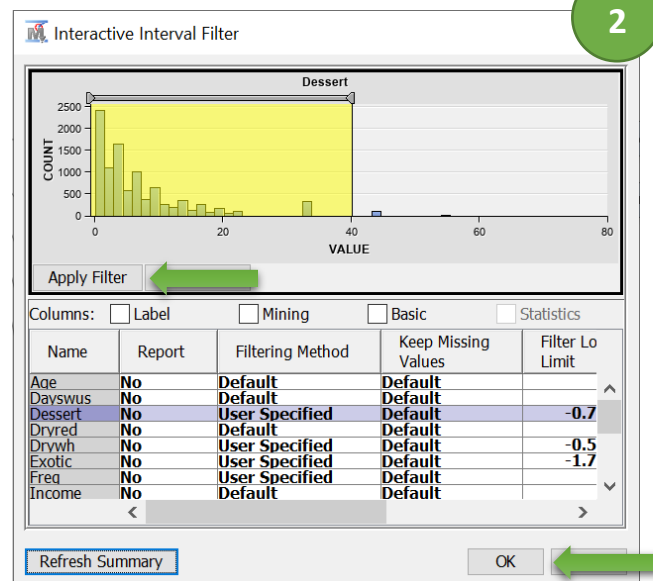
- Guarantee that you specify “none” and “user specified limits” in the default filtering method of categorical and interval variables, respectively.
- In the “export table”, select the option filtered



Click in the “...” bottom after interval variables and you can filter the data.

The yellow area is the one that we will maintain. What is outside that is considered an outlier and will be excluded.

- 1- define the yellow region
- 2- click “apply”
- 3- do the same for all variables
- 4-click “ok”



Then you must check how many observation you excluded

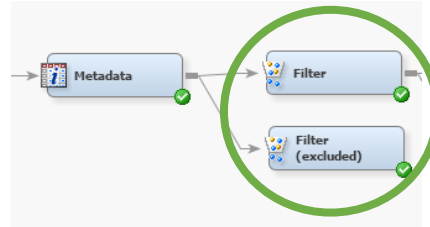
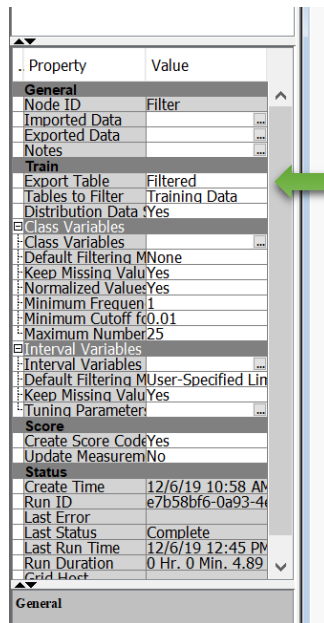
Data Role	Filtered	Excluded	DATA
TRAIN	9565	435	10000

Statistics for Original and FILTERED Data
(maximum 500 observations printed)

Data Role=TRAIN Variable=Circuits

Statistics	Original	Filtered

It is important to analyze the outliers. Thus, you can extract a dataset with the excluded observations. For that, just copy and paste the filter node that you had created (ctrl C, ctrl V) and change the “export table” property to excluded. The, connect it to the source.



- 5- After change the data is always good to see some descriptive statistics and graphics. So, you can add again more descriptive nodes (explore tab).

