NOVA IMS Information Management School

Data Mining

(Practical Classes)

MAA 2015/2016

Enterprise Miner Practical Classes

Frederico Jesus
fjesus@novaims.unl.pt

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Agenda

i. The Data Mining Project: What are we going to do?

ii. The SEMMA Approach

iii. Introduction to SAS Enterprise Miner
   i. Setting up a new SAS Miner Project
   ii. Data sources
   iii. SEMMA Process

iv. Segmentation Problem

Agenda

i. The Data Mining Project: What are we going to do?

ii. The SEMMA Approach

iii. Introduction to SAS Enterprise Miner
   i. Setting up a new SAS Miner Project
   ii. Data sources
   iii. SEMMA Process

iv. Segmentation Problem

Data Mining Project

Tugas:

- Tugas is a Portuguese e-tailer offering an assortment of goods within 5 major categories: Clothes, housekeeping, kitchen, small appliances and toys;
- Tugas started a loyalty program 2 years ago. Among other objectives, the program's aim is to gather Customer information to better drive the marketing efforts;
- There is enough historical information to start producing sound knowledge about their Customers. IT department provided an ABT.

Variable	Description
Custid	Customer ID
Age	Customer Birthday Year
Income	Customer Income
Frq	# Purchases last 18 months
Rcn	Months since last visit
Mnt	Amount spent last 18 months
Clothes	% spent on clothes
Kitchen	% spent on kitchen products
SmallAppliances	% spent on small appliances

Variable	Description
Toys	% spent on toys
HouseKeeping	% spent on house keeping
Education	Degree of Education
Marital_Status	Marital Status
Gender	Customer Gender
Dependents	1 Customer has dependents / 0 if not
PerNetPurch	% Purchases made through the Web
PerCatPurch	% Purchases made through Catalog
Recomendation	Customer Recommendation

Data Mining Project

Segmentation / Cluster Analysis:

- Process of creating k homogeneous groups in which individuals within each group are more similar to each other and less similar to others in other groups;
- The main objective is to address customers in a different manner, according to its characteristics, without having the need to relate individually with each customer (which would be impossible in any event for the majority of the cases);
- Segmentations should have specific proprieties and objectives;

Agenda

i. The Data Mining Project: What are we going to do?

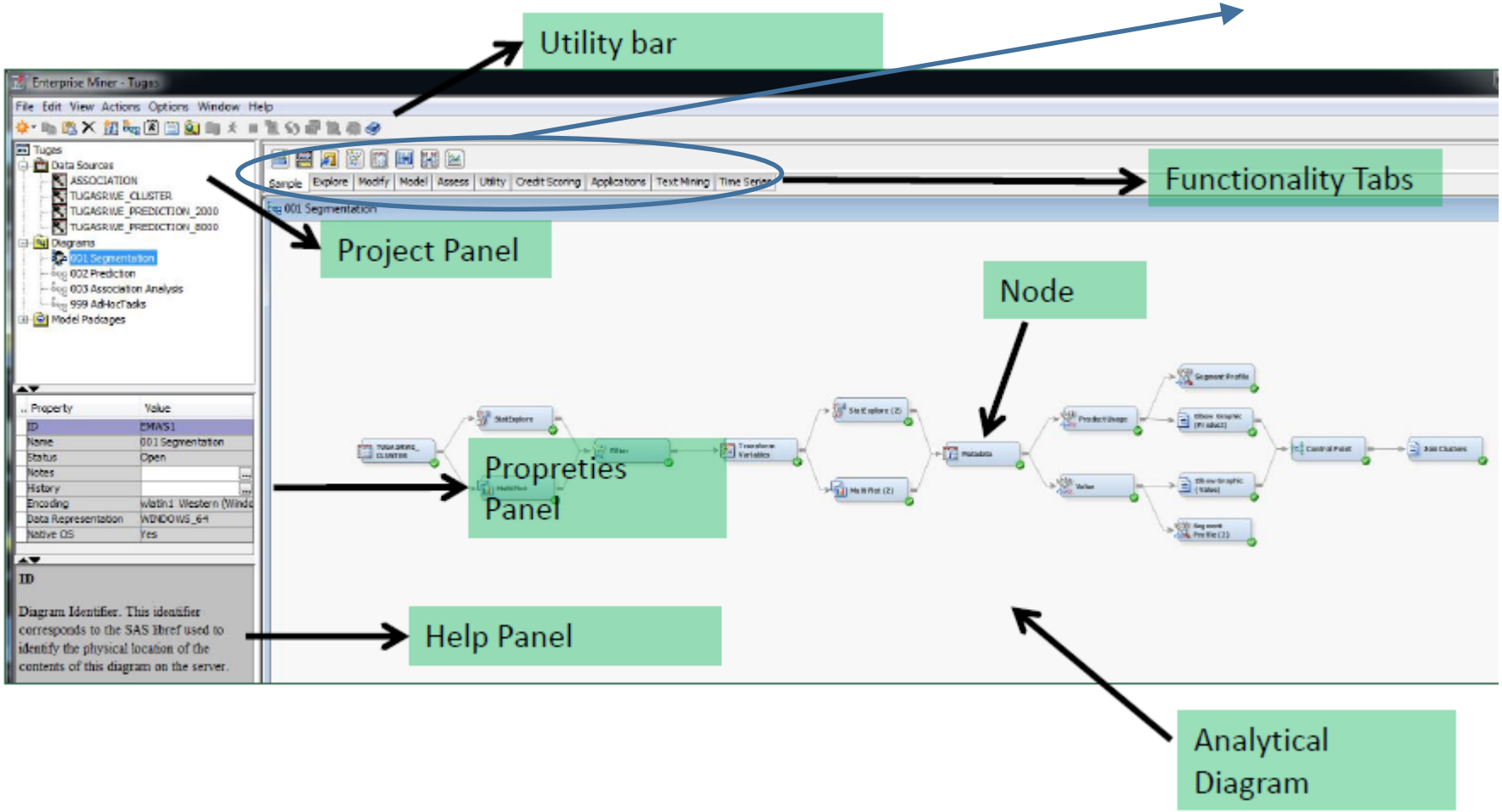
ii. The SEMMA Approach

iii. Introduction to SAS Enterprise Miner
   i. Setting up a new SAS Miner Project
   ii. Data sources
   iii. SEMMA Process

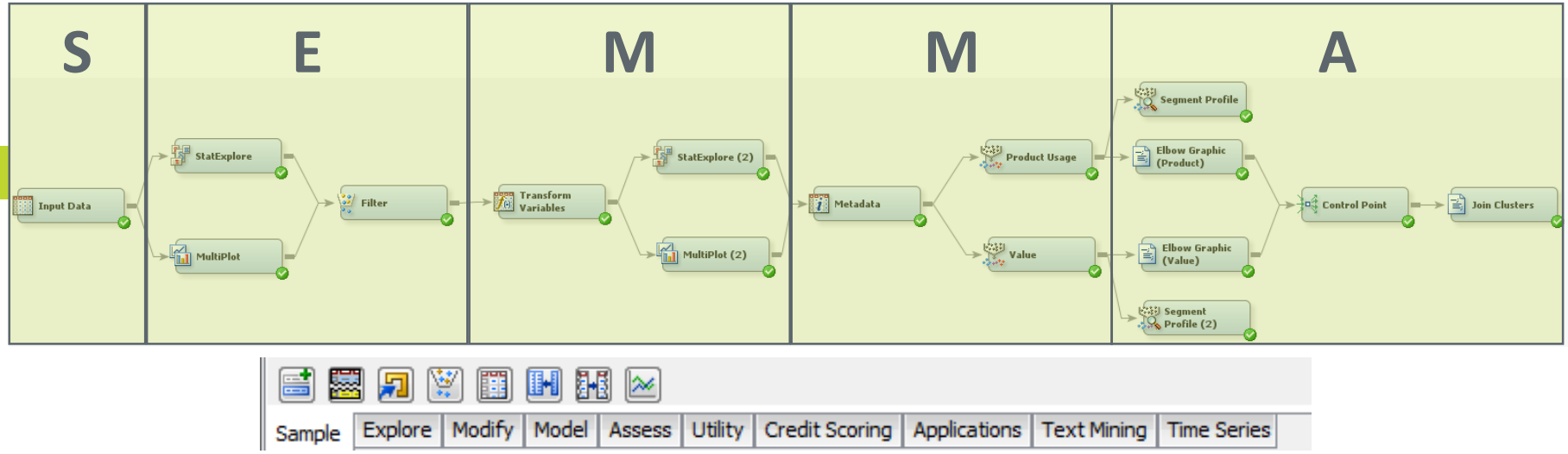
iv. Segmentation Problem

Environment Overview

SEMMA
(SAS approach for DM)



Introduction to SAS EM (SEMMA)



SEMMA is an acronym for the following:

- 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;
- 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;
- Modify: Modify the data by creating, selecting, and transforming the variables to focus the model selection process;
- Model: You model the data by using the analytic techniques to search for a combination of the data that reliably predicts what is intended;
- Assess: Assessment of the alternatives of the model/segmentation in the project, with objective of choosing the one that better serves the purposes.

Agenda

i. The Data Mining Project: What are we going to do?

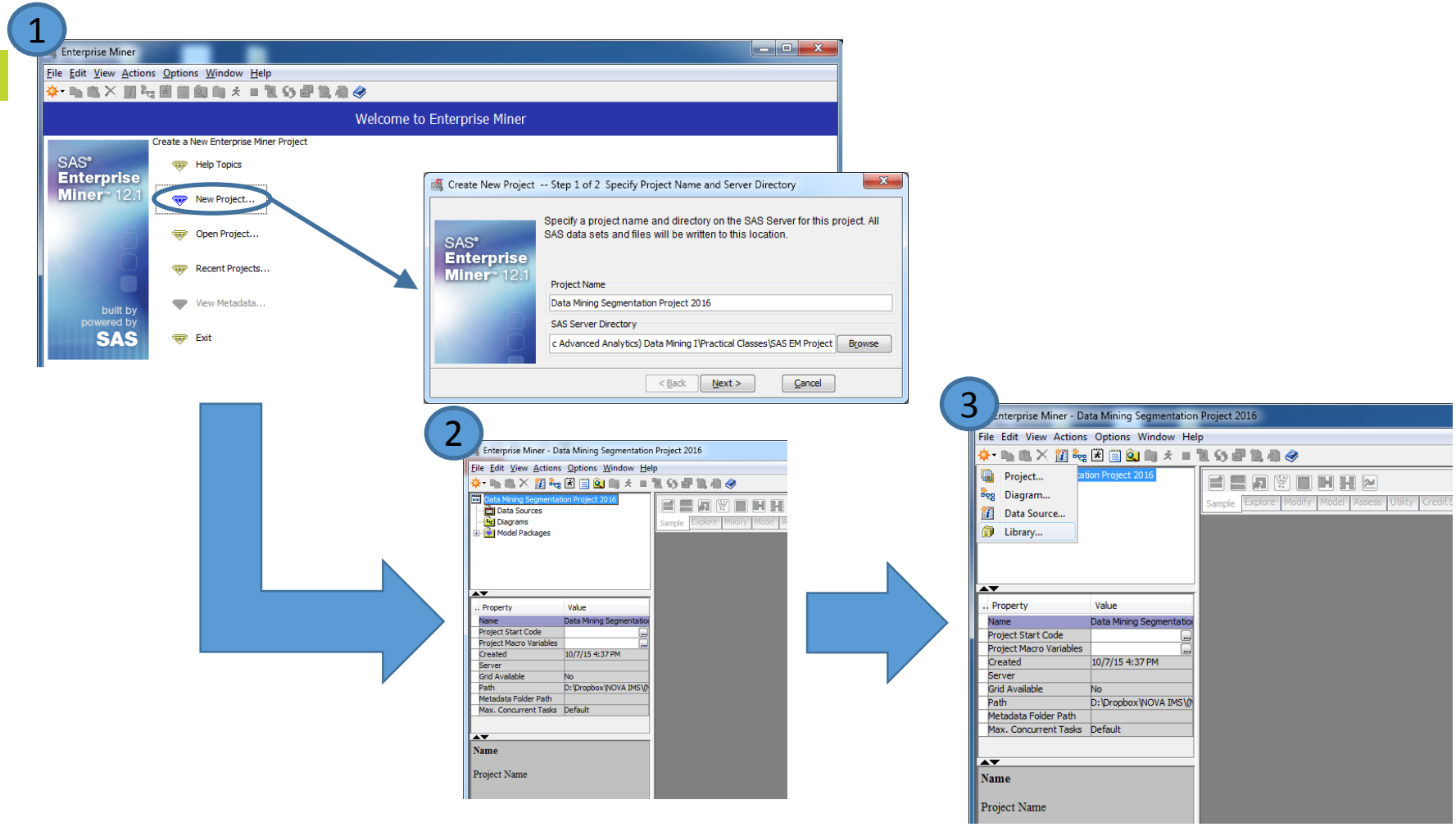
ii. The SEMMA Approach

iii. Introduction to SAS Enterprise Miner
   i. Setting up a new SAS Miner Project
   ii. Data sources
   iii. SEMMA Process

iv. Segmentation Problem

Setting up a new Project

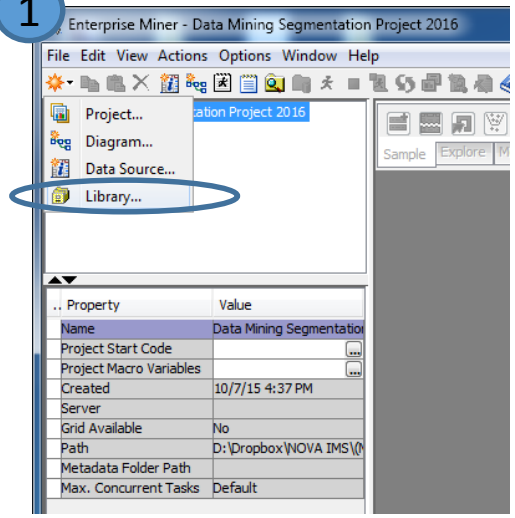
Open SAS Enterprise Miner Workstation 12.1….



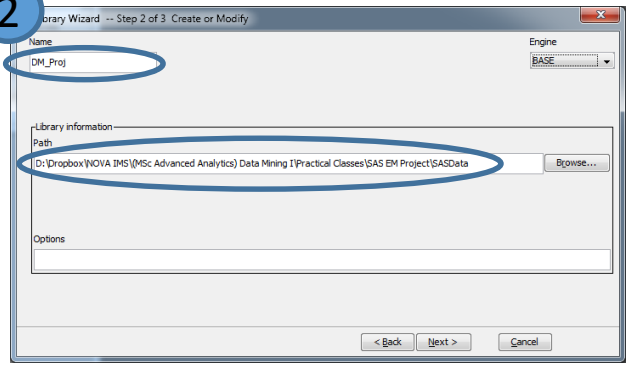
Setting up a new Project – SAS Library

Assign the libraries

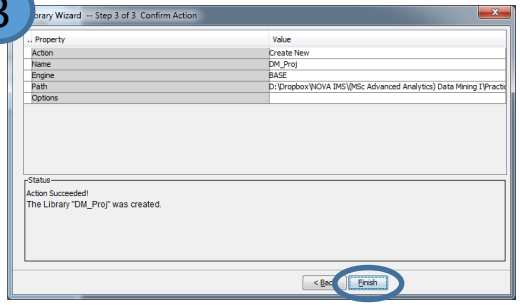
1



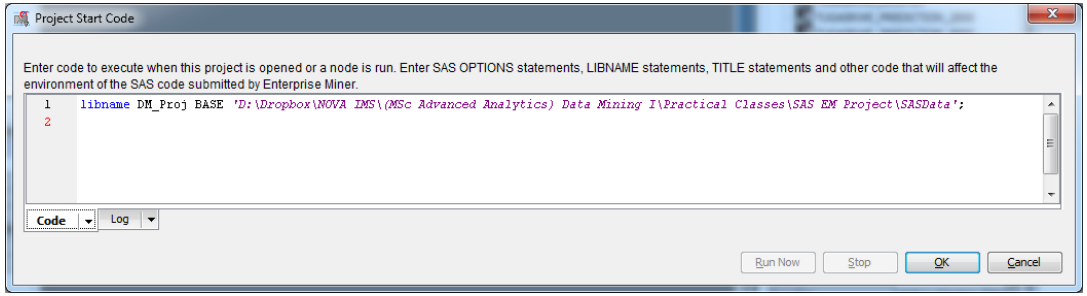
2



3

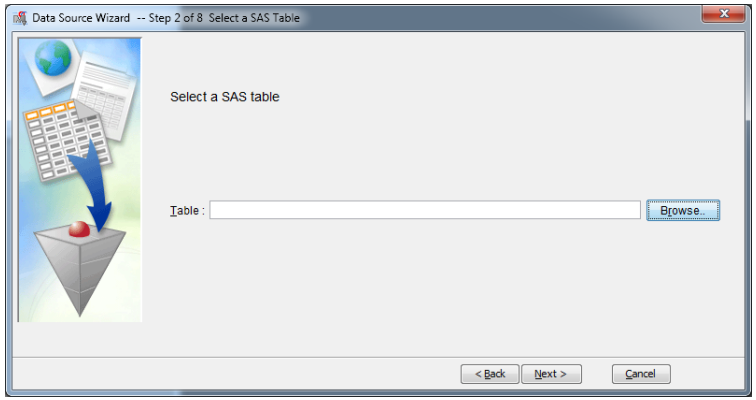


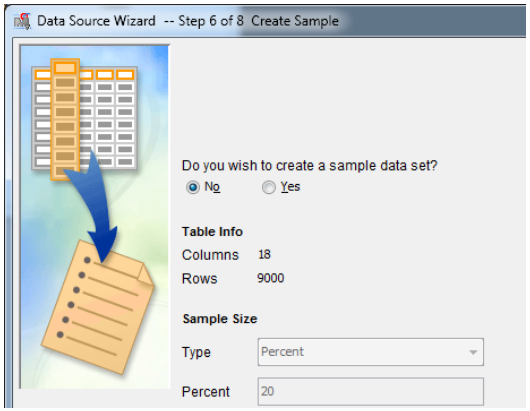
Or…

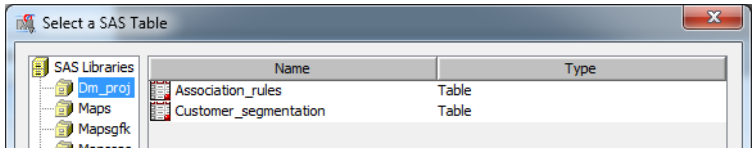


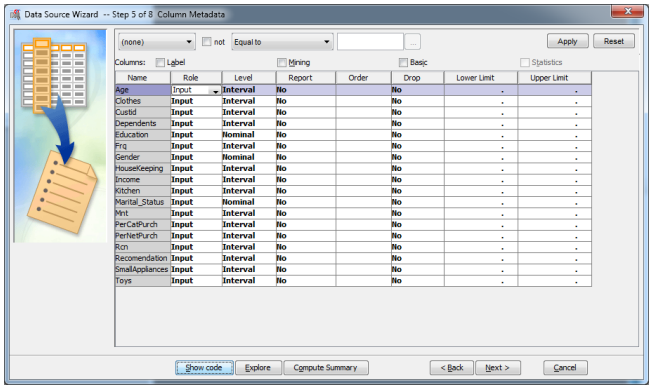
Setting up a new Project – Import Data

Contrarily to SAS EG, we need to specify which datasets we want to use in the DM process.



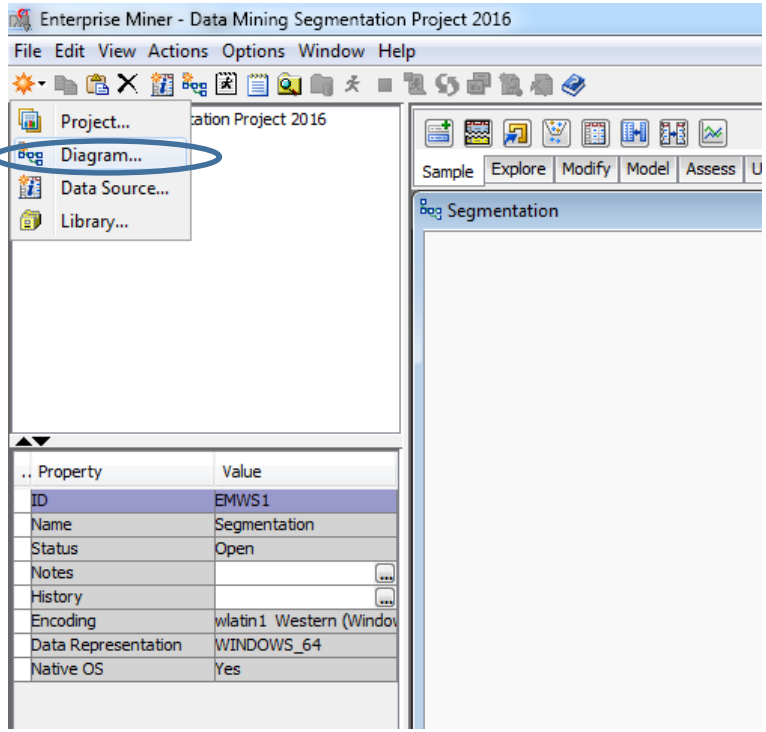


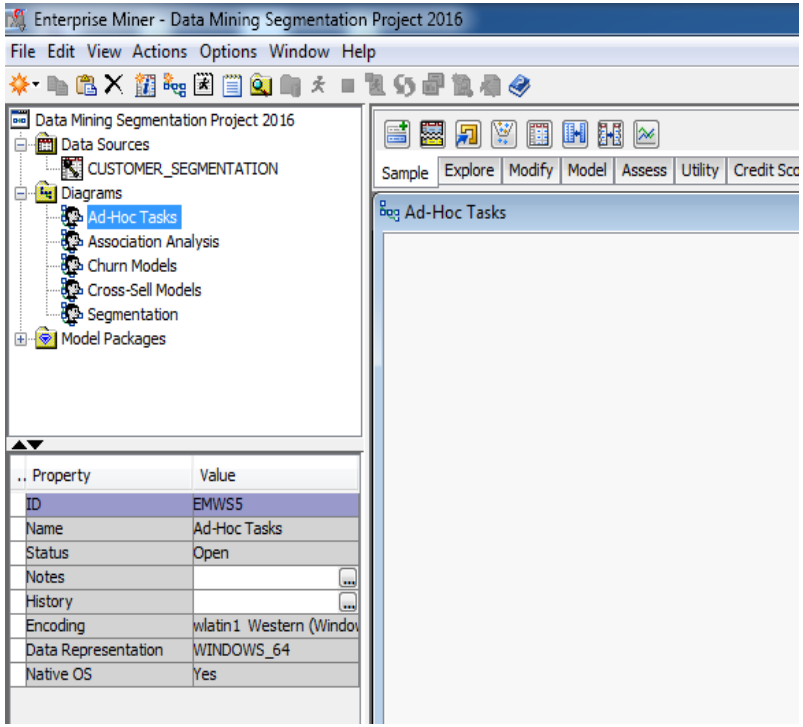




Setting up a new Project – Diagram

Create diagrams





Several Diagrams can be stored within a single “EM Project”.
Diagrams should be in respect to independent tasks/projects.

Agenda

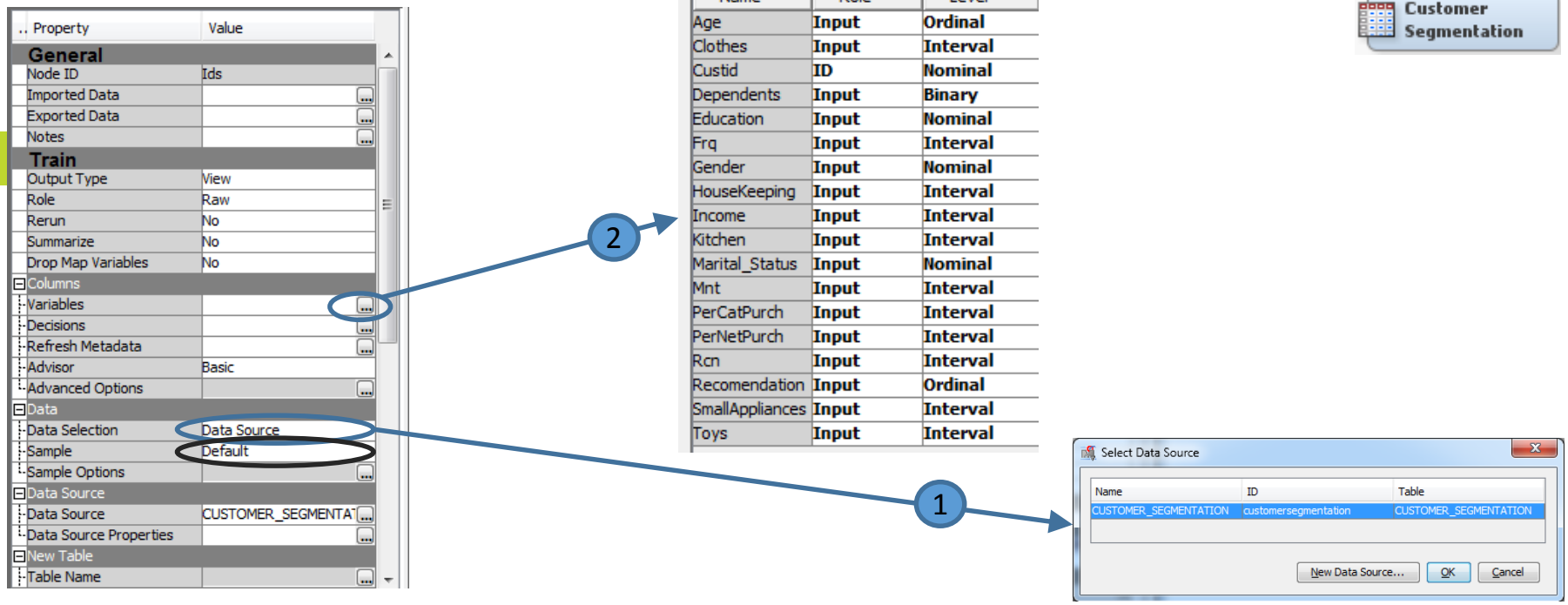
i. The Data Mining Project: What are we going to do?

ii. The SEMMA Approach

iii. Introduction to SAS Enterprise Miner
   i. Setting up a new SAS Miner Project
   ii. Data sources
   iii. SEMMA Process

iv. Segmentation Problem

Input Data Source



1- The "Data Source" option allow us to:

- Evoke the SAS dataset(s) to be used in the project as well as define its role on the modelling/segmentation processes;

2 - The "Variables" option allow us to:

- Review of the variables within the dataset;
- Role definition (ID, Input, Target, Rejected, Segment, etc)
- Level (Nominal, Binary, Ordinal, Interval) definition of each variable;
- In some cases it might be useful to conduct a random sample of data.

The Results pane allow us to identify ouliers;

Input Data Source (Cont.)

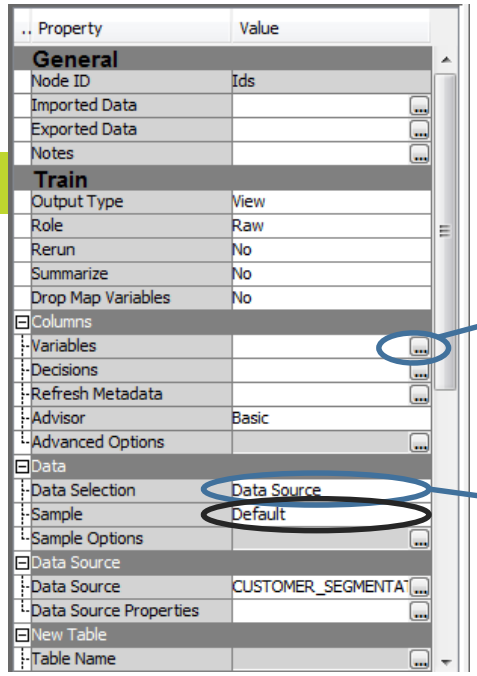
Variable Roles:

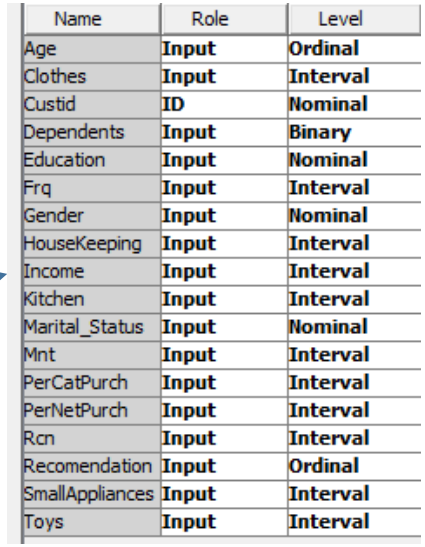
- ID: Identify the ID variable in the dataset. It must present unique values per observation/record and it is not used in any calculations;
- Input: Variables classified as “input” will be the ones used for segmentation/modelling tasks;
- Target: Dependent variable of the problem, i.e., the one we are going to try understand or predict. Naturally, is mandatory for predictive tasks;
- Rejected: These variables are excluded from subsequent analysis/tasks;
- Cost: Useful for calculating profits and take decisions which maximizes it;
- (…)

Variable levels:

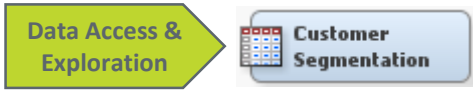
- Binary;
- Nominal;
- Ordinal;
- Interval.

Project Development – Input Data Source

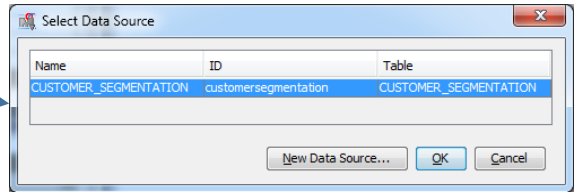




Name	Role	Level
Age	Input	Ordinal
Clothes	Input	Interval
Custid	ID	Nominal
Dependents	Input	Binary
Education	Input	Nominal
Frq	Input	Interval
Gender	Input	Nominal
HouseKeeping	Input	Interval
Income	Input	Interval
Kitchen	Input	Interval
Marital_Status	Input	Nominal
Mnt	Input	Interval
PerCatPurch	Input	Interval
PerNetPurch	Input	Interval
Rcn	Input	Interval
Recomendation	Input	Ordinal
SmallAppliances	Input	Interval
Toys	Input	Interval



1. Select the data source to be used;
2. Variables´ definitions.



1- The "Data Source" option allow us to:

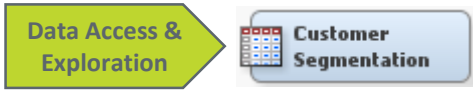
- Evoke the SAS dataset(s) to be used in the project as well as define its role on the modelling/segmentation processes;

2 - The "Variables" option allow us to:

- Review of the variables within the dataset;
- **Role** definition (ID, Input, Target, Rejected, Segment, etc)
- **Level** (Nominal, Binary, Ordinal, Interval) definition of each variable;
- In some cases it might be useful to conduct a random sample of data.

The Results pane allow us to identify *outliers*;

Project Development – Input Data Source



Variable Roles:

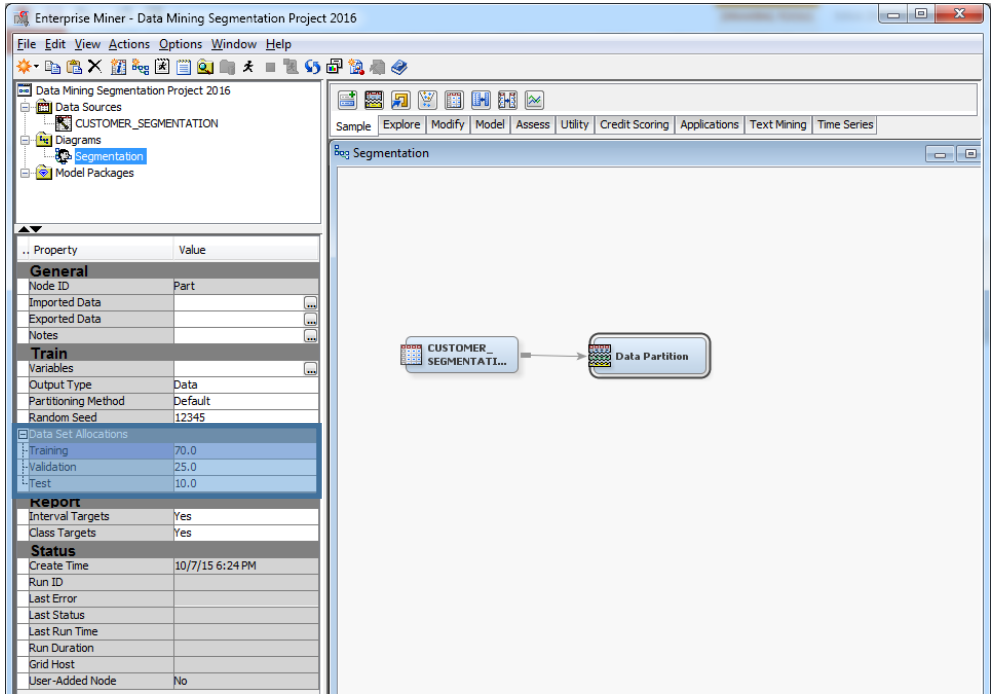
- ID: Identify the ID variable in the dataset. It must present unique values per observation/record and it is not used in any calculations;
- Input: Variables classified as “input” will be the ones used for segmentation/modelling tasks;
- Target: Dependent variable of the problem, i.e., the one we are going to try understand or predict. Naturally, is mandatory for predictive tasks;
- Rejected: These variables are excluded from subsequent analysis/tasks;
- Cost: Useful for calculating profits and take decisions which maximizes it;

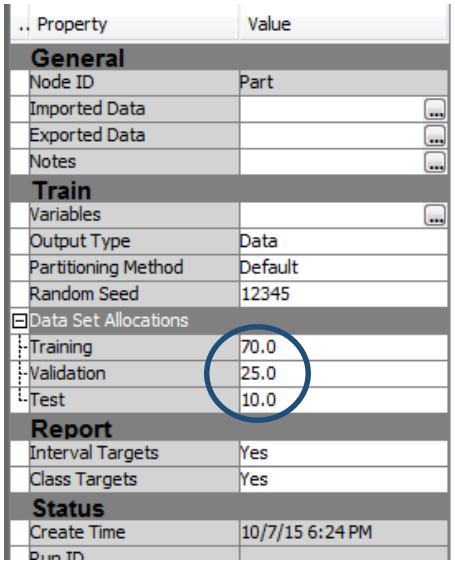
(…)

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”.

Project Development – Data Partition





The Data Partition allow to split the data into three subsets, in a randomly manner:

- Training Set (%) – Individuals in which the model will be trained;
- Validation Set (%) – Individuals in which the model will stop the training process (immediately before the error starts to increase);
- Test Set (%)– Individuals in which the model will be tested in order to give a non-biased error estimate.

Project Development – StatExplore

Property	Value
General	
Node ID	Stat2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Data	
Number of Observations	100000
Validation	No
Test	No
Standard Reports	
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No
Cross-Tabulation	
Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Variab	1000
Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	No
Number of Bins	5
Correlation Statistics	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No
Status	
Create Time	10/7/15 9:35 PM
Run ID	3942128e-9eed-4912-8be
Last Error	
Last Status	Complete
Last Run Time	10/8/15 12:51 AM
Run Duration	0 Hr. 0 Min. 2.44 Sec.
Grid Host	
User-Added Node	No

1

2

Cross Tabulations:

- Education * Dependents;
- Dependents * Gender;

Chi Square and Correlation Statistics:

- Useful in presence of target variables

Project Development – StatExplore

Data Access & Exploration → StatExplore

Class Variables

Data Role	Variable Name	Level	CODE	Frequency Count	Type	Percent	Level Index	Role	Label	Plot
TRAIN	Education	Graduation	0	4430	C	49.22222	4	INPUT	Education	1
TRAIN	Education	2nd Cycle	5	1496	C	16.62222	3	INPUT	Education	1
TRAIN	Education	Master	1	1304	C	14.48889	5	INPUT	Education	1
TRAIN	Education	1st Cycle	3	1104	C	12.26667	2	INPUT	Education	1
TRAIN	Education	PhD	4	593	C	6.588889	7	INPUT	Education	1
TRAIN	Education		6	47	C	0.522222	1	INPUT	Education	1
TRAIN	Education	OldSchool	2	26	C	0.288889	6	INPUT	Education	1
TRAIN	Gender	M	0	5784	C	64.26667	3	INPUT	Gender	1
TRAIN	Gender	F	1	3214	C	35.71111	2	INPUT	Gender	1
TRAIN	Gender	?	2	2	C	0.022222	1	INPUT	Gender	1
TRAIN	Marital_Status	Married	2	3273	C	36.36667	3	INPUT	Marital_Status	1
TRAIN	Marital_Status	Single	3	2294	C	25.48889	4	INPUT	Marital_Status	1

Class Variables:

- Missing values´ analysis;
- First approach to odd (and possibly) error values.

Interval Variables

Ordered Inputs	Data Role	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
1	TRAIN	HouseKee...	4	0	9000	0	77	6.929667	7.881981	2.229297	6.887084
2	TRAIN	Toys	4	0	9000	0	62	7.037	7.924137	2.095795	5.643974
3	TRAIN	Rcn	53	0	9000	0	549	62.46233	69.7566	4.174275	21.10057
4	TRAIN	Kitchen	4	0	9000	0	75	7.039556	7.847558	2.049489	5.620054
5	TRAIN	Mnt	383	0	9000	6	3052	622.1437	646.7193	0.981047	-0.05771
6	TRAIN	Dependents	1	282	8718	0	1	0.707272	0.455041	-0.91121	-1.16996
7	TRAIN	Frq	17	0	9000	3	59	19.84811	10.90239	0.697817	-0.41357
8	TRAIN	Custid	6004	0	9000	1001	11000	6000.566	2887.654	-0.00373	-1.20329
9	TRAIN	Clothes	51	0	9000	1	99	50.447	23.42034	-0.07825	-0.91836
10	TRAIN	SmallAppli	28	0	9000	1	74	28.524	12.58512	0.314701	-0.42248
11	TRAIN	PerNetPurch	45	0	9000	4	88	42.42989	18.49656	-0.2663	-1.0349
12	TRAIN	Income	70021	46	8954	10000	140628	69969.01	27592.16	0.008695	-0.92939
13	TRAIN	PerCatPurch	55	0	9000	12	96	57.57011	18.49656	0.266305	-1.0349
14	TRAIN	Recomend...	4	0	9000	1	5	3.433	1.017876	-0.13016	-1.05877
15	TRAIN	Age	1966	0	9000	1936	1996	1966.062	17.29591	0.00776	-1.19594

Interval Variables:

- Missing values´ analysis;
- 57% of purchases are thought catalogs; 43% using the Tugas´ website;
- Some 80% of sales are clothes and small appliances.

Project Development – MultiPlot

Data Access & Exploration → MultiPlot

General	
Node ID	Plot
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Type of Charts	Bar Charts
Bar Chart Options	
Graph Orientation	Vertical
Include Missing Values	Yes
Interval Target Charts	Mean
Show Values	Yes
Statistic	Freq
Numeric Threshold	20
Scatter Options	
Confidence Interval	Yes
Regression Equation	No
Regression Type	Linear
Status	
Create Time	10/7/15 9:39 PM
Run ID	773466fc-acac-4af3-aa12
Last Error	
Last Status	Complete
Last Run Time	10/7/15 9:40 PM
Run Duration	0 Hr. 0 Min. 22.40 Sec.
Grid Host	
User-Added Node	No

Applications:

- Provides useful insights about the data (e.g., visualization of empirical distributions);
- Identifies possible data quality/integrity problems;
- If a target variable is specified, provides additional graphics, useful for explanatory variables´ selection.

Techniques:

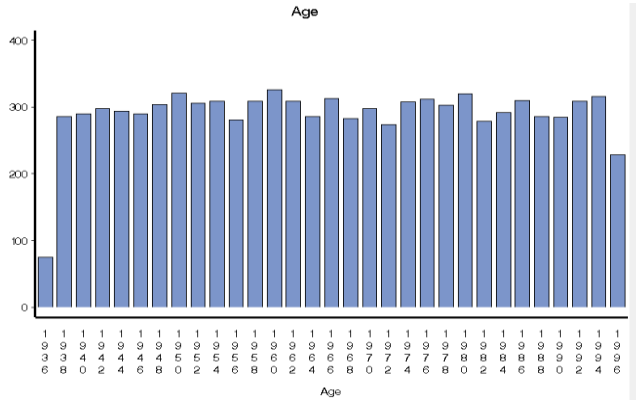
- Barcharts / Histograms (enhanced in presence of target/dependent variables);
- Scatterplots (enhanced in presence of target/dependent variables);

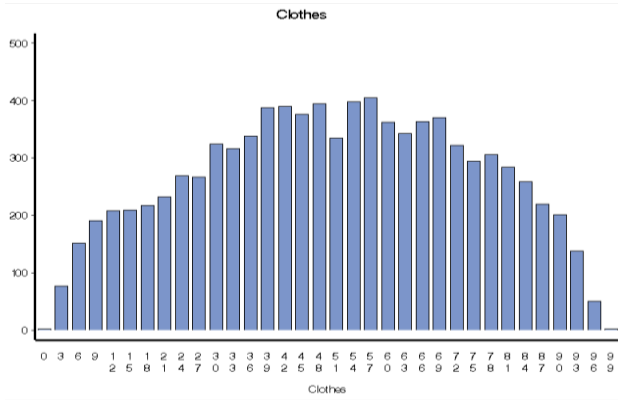
Project Development – MultiPlot

Histograms / Frequency diagrams allow us to:

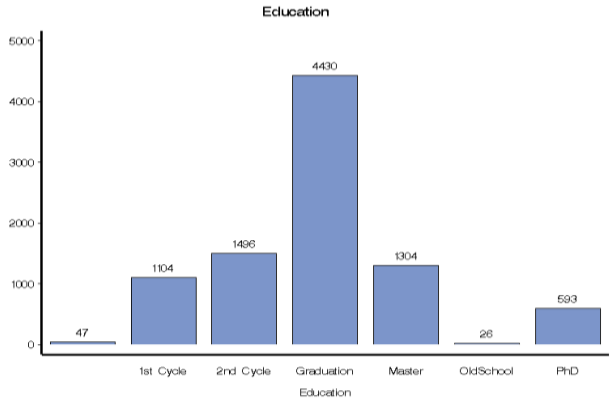
Data Access & Exploration → MultiPlot

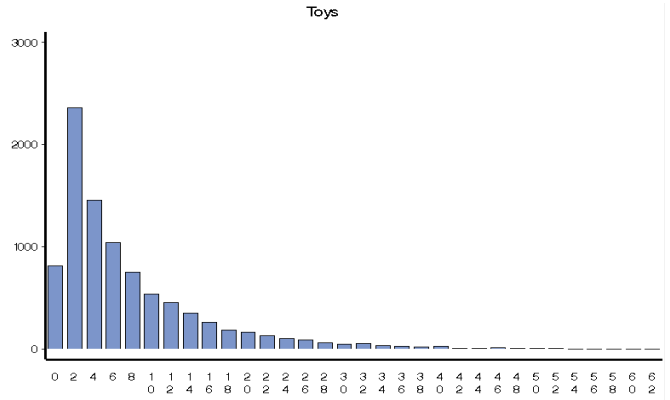
Assess data´s distribution





Identify inconsistencies and potential outliers in our data



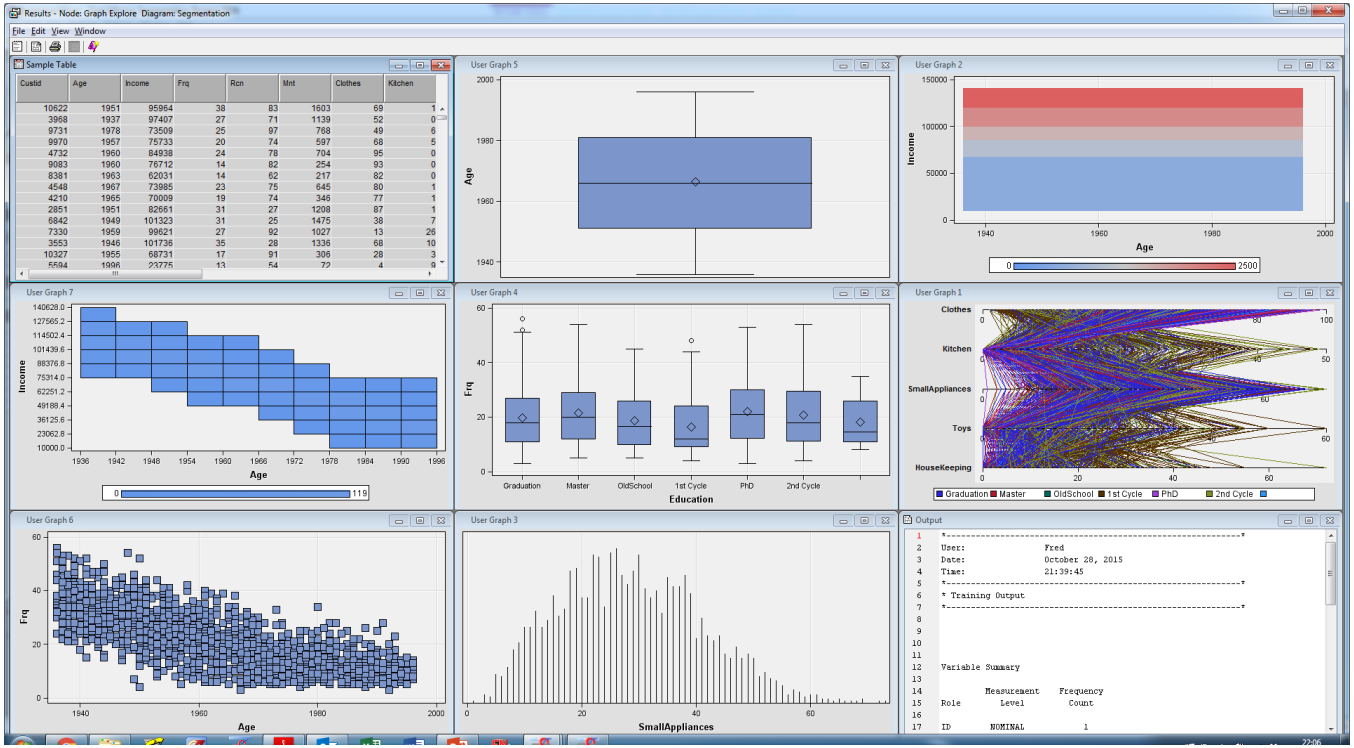


Project Development – GraphExplore

Data Access & Exploration → Graph Explore

Applications:

- An advanced visualization tool for interactive data exploration;
- Analysis of univariate distributions, investigate multivariate distributions, create scatter and box plots, constellation and 3-D charts, and so on.
- Allows users to custom-made their plots…



Review

Variable	Description
Custid	Customer ID
Age	Customer Birthday Year
Income	Customer Income
Frq	# Purchases last 18 months
Rcn	Months since last visit
Mnt	Amount spent last 18 months
Clothes	% spent on clothes
Kitchen	% spent on kitchen products
SmallAppliances	% spent on small appliances

Variable	Description
Toys	% spent on toys
HouseKeeping	% spent on house keeping
Education	Degree of Education
Marital_Status	Marital Status
Gender	Customer Gender
Dependents	1 Customer has dependents / 0 if not
PerNetPurch	% Purchases made through the Web
PerCatPurch	% Purchases made through Catalog
Recomendation	Customer Recommendation

Business Needs → Data Integration → Data Access & Exploration → Data Preparation → Modelling / Segmentation → Assessment

- Segmentation framework;

- Input Data;
- Stat Explore;
- Multiplot;
- Graph Explore;

- Filter;
- Impute;
- Replacement;
- Transform;
- Metadata;

- À Priori;
- Hierarchical;
- Non-hierarchical;

Project Development – Filter

Applications:

- The Filter node allow the creation and application of filters to the input data;
- Filters can exclude observations, such as extreme outliers and errant data;
- Missing values can also be addressed.

Outliers:

- Class variables: Automatic and manual settings for exclusion;
- Interval variables: Percentiles´ criteria, # (2?) standard deviations from the mean (!); manual settings.

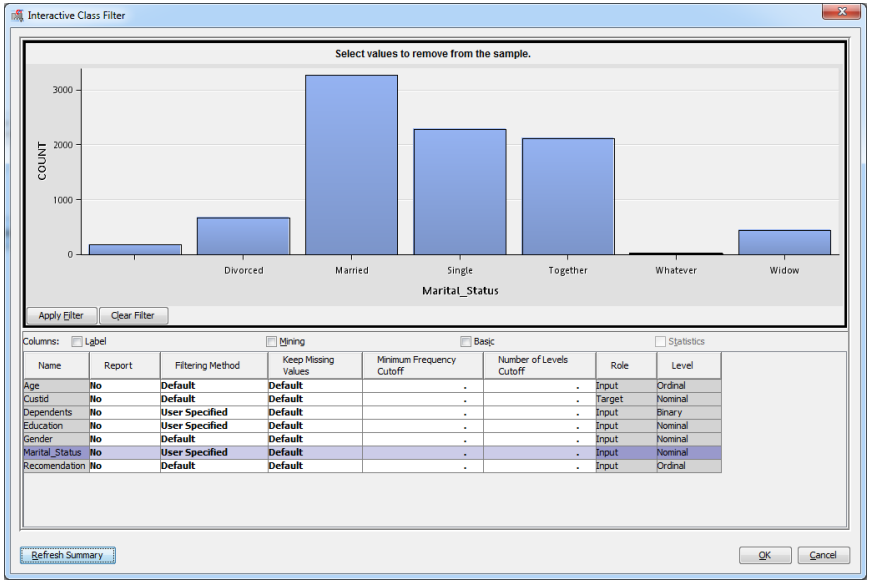
Missing Values:

- Keep Missing Values: Yes (addressed next)

Project Development – Filter

.. Property	Value
General	
Node ID	Filter
Imported Data	
Exported Data	
Notes	
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
Class Variables	
Class Variables	
Default Filtering Method	None
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percentage	0.01
Maximum Number of Levels Cutoff	25
Interval Variables	
Interval Variables	
Default Filtering Method	None
Keep Missing Values	Yes
Tuning Parameters	
Score	
Create score code	Yes
Update Measurement Level	No
Status	
Create Time	10/7/15 9:51 PM
Run ID	8ff6fa71-2e99-44b7-9851-41beae2
Last Error	
Last Status	Complete
Last Run Time	10/29/15 7:40 AM
Run Duration	0 Hr. 0 Min. 16.56 Sec.
Grid Host	
User-Added Node	No

1



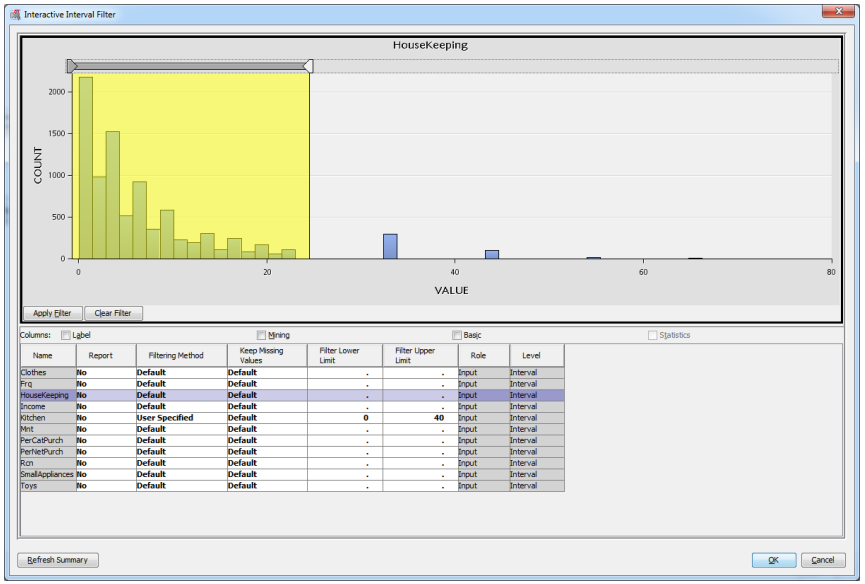
Class Variables:

- Manual selection of values to exclude.

Project Development – Filter

.. Property	Value
General	
Node ID	Filter
Imported Data	
Exported Data	
Notes	
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
Class Variables	
Class Variables	
Default Filtering Method	None
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percentage	0.01
Maximum Number of Levels Cutoff	25
Interval Variables	
Interval Variables	
Default Filtering Method	None
Keep Missing Values	Yes
Tuning Parameters	
Score	
Create score code	Yes
Update Measurement Level	No
Status	
Create Time	10/7/15 9:51 PM
Run ID	8ff6fa71-2e99-44b7-9851-41beae2
Last Error	
Last Status	Complete
Last Run Time	10/29/15 7:40 AM
Run Duration	0 Hr. 0 Min. 16.56 Sec.
Grid Host	
User-Added Node	No

2



Interval Variables:

- Manual selection of values to exclude;
- Automatic criteria.

Project Development – Filter

Results:

Filter Limits for Interval Variables
(maximum 500 observations printed)

Variable	Role	Minimum	Maximum	Filter Method	Keep Missing Values	Label
Kitchen	INPUT	0	40	MANUAL	Y	Kitchen

Excluded Class Values
(maximum 500 observations printed)

Variable	Role	Level	Train Count	Train Percent	Label	Filter Method
Dependents	INPUT	.	282	3.13333		MANUAL
Education	INPUT	OldSchool	26	0.28889		MANUAL
Marital_Status	INPUT	Whatever	15	0.16667		MANUAL

Number Of Observations

Data Role	Filtered	Excluded	DATA
TRAIN	8644	356	9000

Project Development – Replacement

Filter Node:

- Keep Missing Values: Yes

General	
Node ID	Repl2
Imported Data	
Exported Data	
Notes	
Train	
Interval Variables	
Replacement Editor	
Default Limits Method	None
Cutoff Values	
Class Variables	
Replacement Editor	
Unknown Levels	Ignore
Score	
Replacement Values	Computed
Hide	No
Report	
Replacement Report	Yes
Status	
Create Time	10/28/15 11:32 PM
Run ID	b61d8b8f-f492-42a1-b330-daf3f1d
Last Error	
Last Status	Complete
Last Run Time	10/29/15 7:30 AM
Run Duration	0 Hr. 0 Min. 11.88 Sec.
Grid Host	
User-Added Node	No

1

2

Interactive Replacement Interval Filter

Columns: Label Mining Basic Statistics

Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	Replace Method	Lower Replacement Value	Upper Replacement Value	Role	Level
Clothes	Default	Default	.	.	Default	.	.	Input	Interval
Frq	Default	Default	.	.	Default	.	.	Input	Interval
HouseKeeping	Default	Default	.	.	Default	.	.	Input	Interval
Income	Default	Default	.	.	Default	.	.	Input	Interval
Kitchen	Default	Default	.	.	Default	.	.	Input	Interval
Mnt	Default	Default	.	.	Default	.	.	Input	Interval
PerCatPurch	Default	Default	.	.	Default	.	.	Input	Interval
PerNetPurch	Default	Default	.	.	Default	.	.	Input	Interval
Rcn	Default	Default	.	.	Default	.	.	Input	Interval
SmallAppliances	Default	Default	.	.	Default	.	.	Input	Interval
Toys	Default	Default	.	.	Default	.	.	Input	Interval

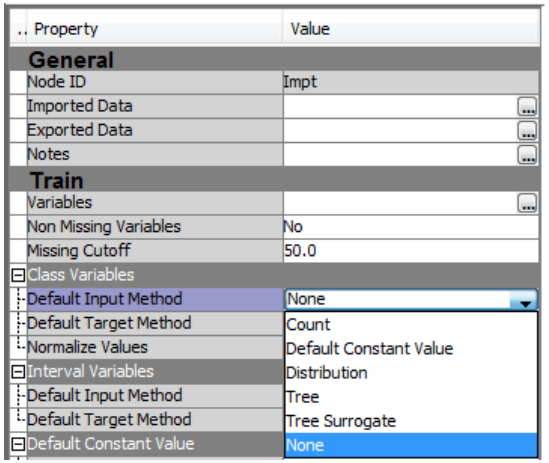
Generate Summary OK Cancel

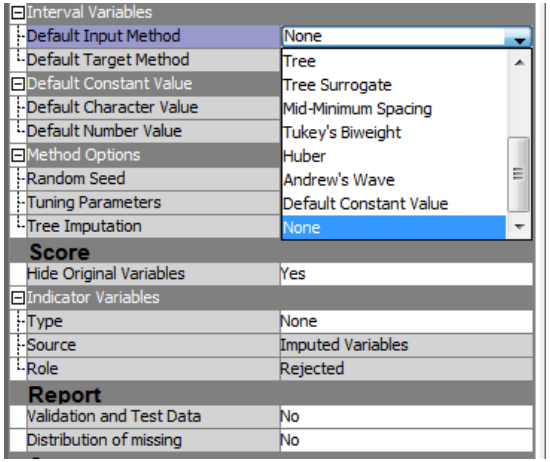
Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
Age	1979		166	N		1979
Age	1951		163	N		1951
Age	1961		163	N		1961
Age	1974		162	N		1974
Age	1976		162	N		1976
Age	1992		162	N		1992
Age	1960		158	N		1960
Age	1949		157	N		1949
Age	1950		154	N		1950
Age	1959		153	N		1959
Age	1966		153	N		1966
Age	1958		152	N		1958
Age	1978		152	N		1978
Age	1983		151	N		1983
Age	1940		150	N		1940
Age	1942		150	N		1942
Age	1989		150	N		1989
Age	1953		149	N		1953
Age	1986		149	N		1986

OK Cancel

Project Development – Inpute





Applications:

- The Impute node enables missing values´ replacement in the data set;
- This step may be, depending upon the context and techniques, critical as some methods ignore records with missing values.
- The impute node exports new values by creating new variables that contain replacements for missing values. In other words, it does not replace variables in the data set (with prefix IMP_).

Missing Values:

- Class variables: Surrogate Trees;
- Interval variables: Statistics (usually median, which is less sensitive to outliers)

Project Development – Transform Variables

Property	Value
General	
Node ID	Trans
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Formulas	
Interactions	
SAS Code	
Default Methods	
Interval Inputs	None
Interval Targets	None
Class Inputs	None
Class Targets	None
Treat Missing as Level	No
Sample Properties	
Method	First N
Size	Default
Random Seed	12345
Optimal Binning	
Number of Bins	4
Missing Values	Use in Search
Grouping Method	
Cutoff Value	0.1
Group Missing	No
Number of Bins	Variables
Add Minimum Value to Offset Val	Yes
Offset Value	1

Applications:

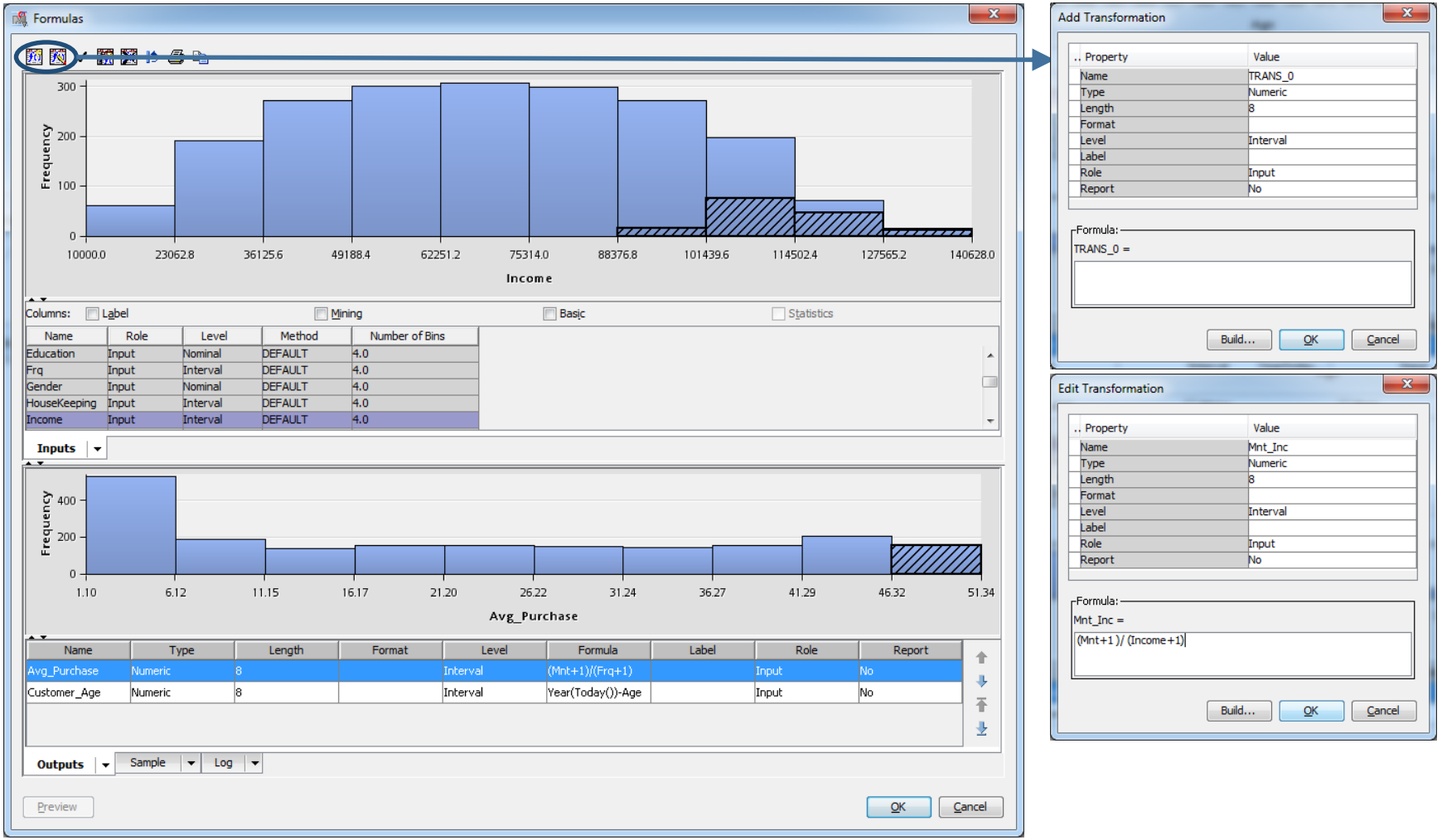
- Creation of new variables (business, variations, absolute vs ratios, etc);
- Transformation of existing variables (optimized in presence of target variables);
- Both types may improve (predictive) model´s performance or improved pattern´s findings (in case of descriptive tasks);
- Transformed variables are added to the data set with specific prefixes, and original variables are updated with “rejected” role;
- Typical objectives associated with variables´ transformation are normalizations (variance stabilization), non-linearity removal and counter non-normality. Transformations can be a function of one or more variables.

Common transformations:

- Standardization; Optimal binning; quantiles…

Project Development – Transform Variables

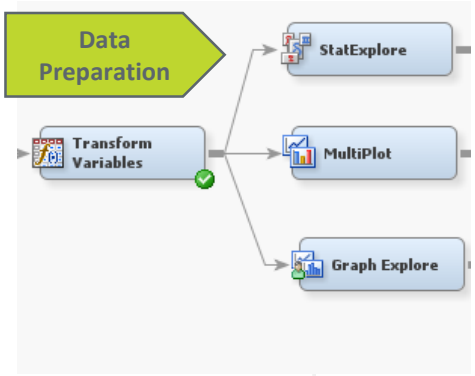
Business Needs → Data Preparation → Transform Variables

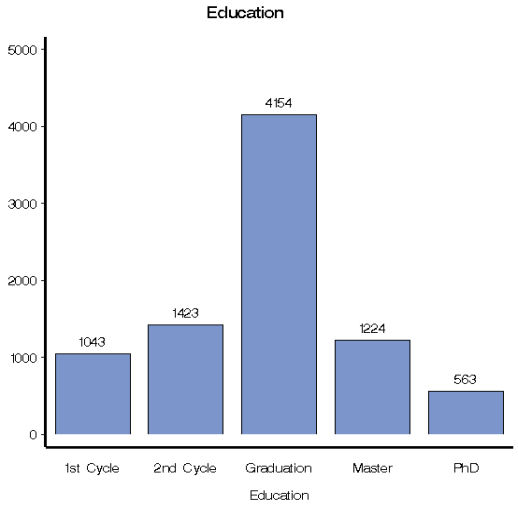


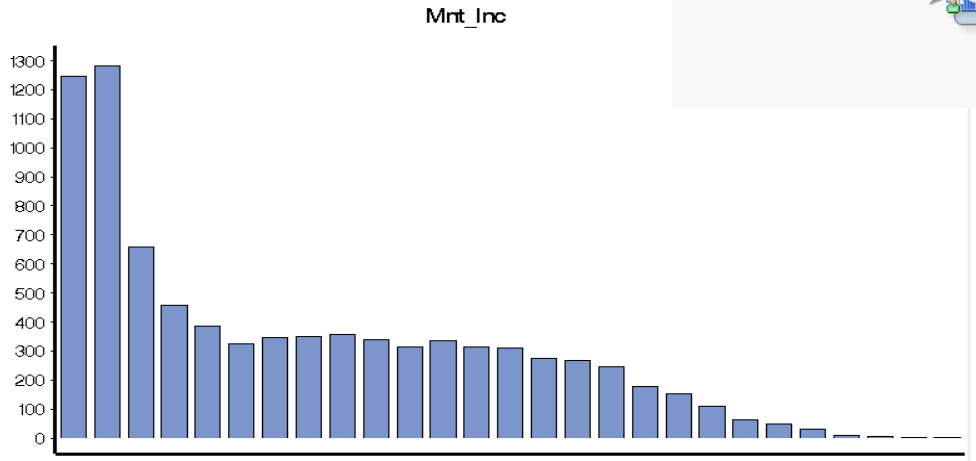
Project Development – Reassess variables

Applications:

- Reassess variables;
- Do a second exploratory analysis confirming the desired effect of previous changes and transformations;

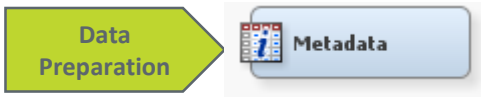






Ordered Inputs	Data Role	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Abs C.V.	Coefficient of Variation
1	TRAIN	HouseKeeping	4	0	8409	0	59	6.906172	7.826853	2.144048	5.920686	INPUT	HouseKee...	1.133313	1.133313
2	TRAIN	Toys	4	0	8409	0	60	7.021406	7.897146	2.063998	5.341356	INPUT	Toys	1.124724	1.124724
3	TRAIN	Rcn	53	0	8409	0	549	62.35688	69.38733	4.195415	21.40605	INPUT	Rcn	1.112745	1.112745
4	TRAIN	Kitchen	4	0	8409	0	59	7.008087	7.756412	1.927079	4.419611	INPUT	Kitchen	1.10678	1.10678
5	TRAIN	Mnt	383	0	8409	6	3052	622.4121	647.5513	0.987711	-0.03996	INPUT	Mnt	1.04039	1.04039
6	TRAIN	Mnt_Inc	0.005557	0	8409	.0001972	0.026832	0.007071	0.005941	0.641137	-0.75198	INPUT	Mnt_Inc	0.840108	0.840108
7	TRAIN	Frq	17	0	8409	3	59	19.86193	10.91261	0.700695	-0.40756	INPUT	Frq	0.549423	0.549423
8	TRAIN	Clothes	51	0	8409	1	99	50.54929	23.40166	-0.08411	-0.91928	INPUT	Clothes	0.462947	0.462947
9	TRAIN	SmallAppliances	28	0	8409	1	74	28.49352	12.56658	0.321573	-0.4154	INPUT	SmallAppli...	0.441033	0.441033
10	TRAIN	PerNetPurch	45	0	8409	4	88	42.47009	18.45873	-0.26939	-1.03042	INPUT	PerNetPurch	0.434629	0.434629
11	TRAIN	Income	70087	0	8409	10000	140628	70022.72	27605.8	0.010554	-0.92904	INPUT	Income	0.394241	0.394241
12	TRAIN	AGE_Years	49	0	8409	19	79	48.93329	17.27888	-0.00702	-1.19328	INPUT	AGE_Years	0.353111	0.353111
13	TRAIN	PerCatPurch	55	0	8409	12	96	57.52991	18.45873	0.269389	-1.03042	INPUT	PerCatPurch	0.320855	0.320855

Project Development – Metadata



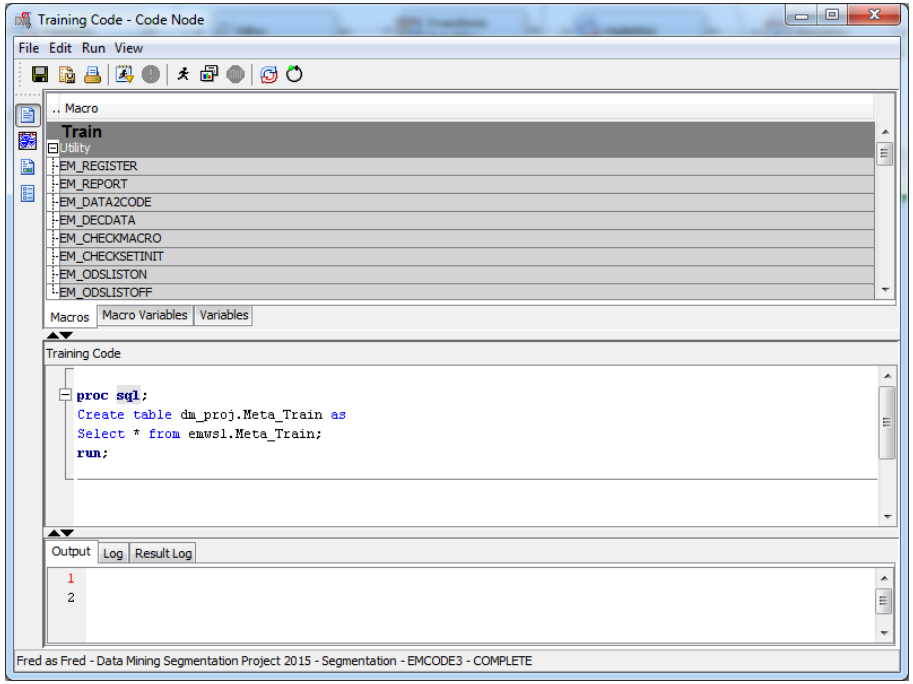
Application:

- Change variables characteristics (roles, measurements, presence) in the middle of the analytical process.

.. Property	Value
General	
Node ID	Meta
Imported Data	
Exported Data	
Notes	
Train	
Import Selection	
Summarize	No
Advanced Advisor	No
Rejected Variables	
Hide Rejected Variables	No
Combine Rule	None
Variables	
Train	
Transaction	
Validate	
Test	
Score	
Status	
Create Time	10/29/15 9:05 AM
Run ID	77294109-683f-4c1c-a3d6-ae74f9
Last Error	
Last Status	Complete
Last Run Time	10/29/15 9:05 AM
Run Duration	0 Hr. 0 Min. 5.96 Sec.
Grid Host	
User-Added Node	No

Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report
Age	N	Default	Input	Default	Ordinal	Default	Default	Default
Clothes	N	Default	Input	Default	Interval	Default	Default	Default
Custid	N	Default	ID	Default	Nominal	Default	Default	Default
Dependents	N	Default	Input	Default	Binary	Default	Default	Default
Education	N	Default	Input	Default	Nominal	Default	Default	Default
Frq	N	Default	Input	Default	Interval	Default	Default	Default
Gender	N	Default	Input	Default	Nominal	Default	Default	Default
HouseKeeping	N	Default	Input	Default	Interval	Default	Default	Default
Income	N	Default	Input	Default	Interval	Default	Default	Default
Kitchen	N	Default	Input	Default	Interval	Default	Default	Default
Marital_Status	N	Default	Input	Default	Nominal	Default	Default	Default
Mnt	N	Default	Input	Default	Interval	Default	Default	Default
PerCatPurch	N	Default	Input	Default	Interval	Default	Default	Default
PerNetPurch	N	Default	Input	Default	Interval	Default	Default	Default
Rcn	N	Default	Input	Default	Interval	Default	Default	Default
Recomendation	N	Default	Input	Default	Ordinal	Default	Default	Default
SmallAppliances	N	Default	Input	Default	Interval	Default	Default	Default
Toys	N	Default	Input	Default	Interval	Default	Default	Default

Project Development – Cluster Analysis



- Exporting Metadata´s output table to SAS Enterprise Guide allow us to use have more flexibility in Cluster Analysis (see provided SAS Code at NOVAIMS Online)

Project Development – Cluster Analysis

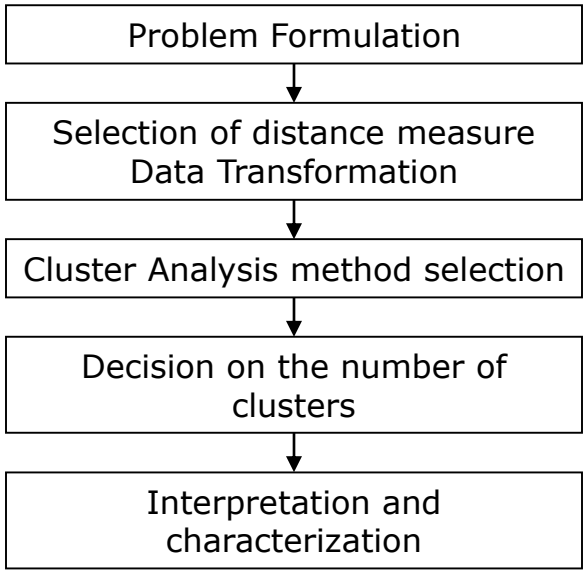
Modelling / Segmentation → Cluster

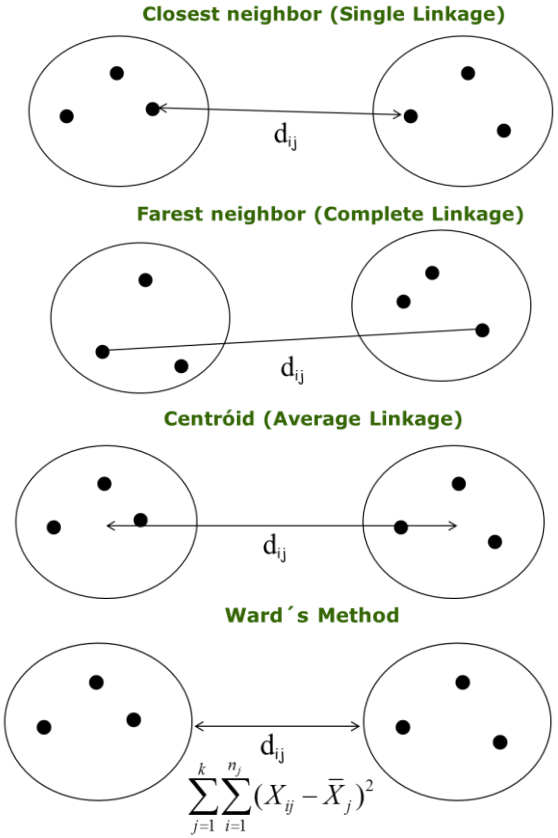
Application:

- Conduct Cluster Analysis.
- Hierarchical and non-hierarchical algorithms available;
- Identify the “natural” clusters within our data.

Algorithms:

- Hierarchical Methods: Ward, Complete and Centroid;
- Non-hierarchical: k-Means.





Project Development – Cluster Analysis

Modelling / Segmentation → Cluster

General	
Node ID	Clus
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Cluster Variable Role	Segment
Internal Standardization	Standardization
Number of Clusters	
Specification Method	Automatic
Maximum Number of Clusters	10
Selection Criterion	
Clustering Method	Ward
Preliminary Maximum	50
Minimum	2
Final Maximum	20
CCC Cutoff	3
Encoding of Class Variable	
Ordinal Encoding	Rank
Nominal Encoding	GLM
Initial Cluster Seeds	
Seed Initialization Method	Default
Minimum Radius	0.0
Drift During Training	No
Training Options	
Use Defaults	Yes
Settings	
Missing Values	
Interval Variables	Default
Nominal Variables	Default
Ordinal Variables	Default
Scoring Imputation Method	None
Score	
Cluster Variable Role	Segment
Hide Original Variables	Yes
Cluster Label Editor	
Report	
Cluster Graphs	Yes
Tree Profile	Yes
Distance Plot and Table	Yes

- Standardization Definition;
- Cluster´s number manual definition;
- Class Variables´ encoding;
- Missing values´ handling;
- Cluster´s labelling.

Project Development – Cluster Analysis

Modelling / Segmentation

General	
Node ID	Clus
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Cluster Variable Role	Segment
Internal Standardization	Standardization
Number of Clusters	
Specification Method	User Specify
Maximum Number of Clusters	3
Selection Criterion	
Clustering Method	Ward
Preliminary Maximum	50
Minimum	2
Final Maximum	3
CCC Cutoff	3
Encoding of Class Variables	
Ordinal Encoding	Rank
Nominal Encoding	GLM
Initial Cluster Seeds	
Seed Initialization Method	Princomp
Minimum Radius	0.0
Drift During Training	No
Training Options	
Use Defaults	Yes
Settings	
Missing Values	
Interval Variables	Default
Nominal Variables	Default
Ordinal Variables	Default
Scoring Imputation Method	None
Score	
Cluster Variable Role	Segment
Hide Original Variables	Yes
Cluster Label Editor	
Report	
Cluster Graphs	Yes
Tree Profile	Yes
Distance Plot and Table	Yes

Product Segmentation

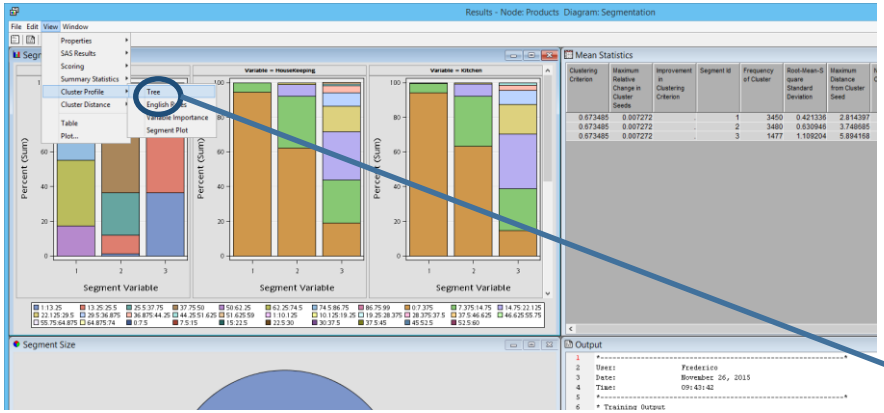
Name	Use	Report	Role	Level
SmallAppliances	Yes	No	Input	Interval
HouseKeeping	Yes	No	Input	Interval
Toys	Yes	No	Input	Interval
Kitchen	Yes	No	Input	Interval
Clothes	Yes	No	Input	Interval
Rcn	No	No	Input	Interval
Frq	No	No	Input	Interval
AGE_Years	No	No	Input	Interval
Education	No	No	Input	Nominal
Gender	No	No	Input	Nominal
Income	No	No	Input	Interval
PerNetPurch	No	No	Input	Interval
Mnt_Inc	No	No	Input	Interval
Recomendation	No	No	Input	Ordinal
Marital_Status	No	No	Input	Nominal
Custid	No	No	ID	Nominal
Dependents	No	No	Input	Binary
PerCatPurch	No	No	Input	Interval
Mnt	No	No	Input	Interval

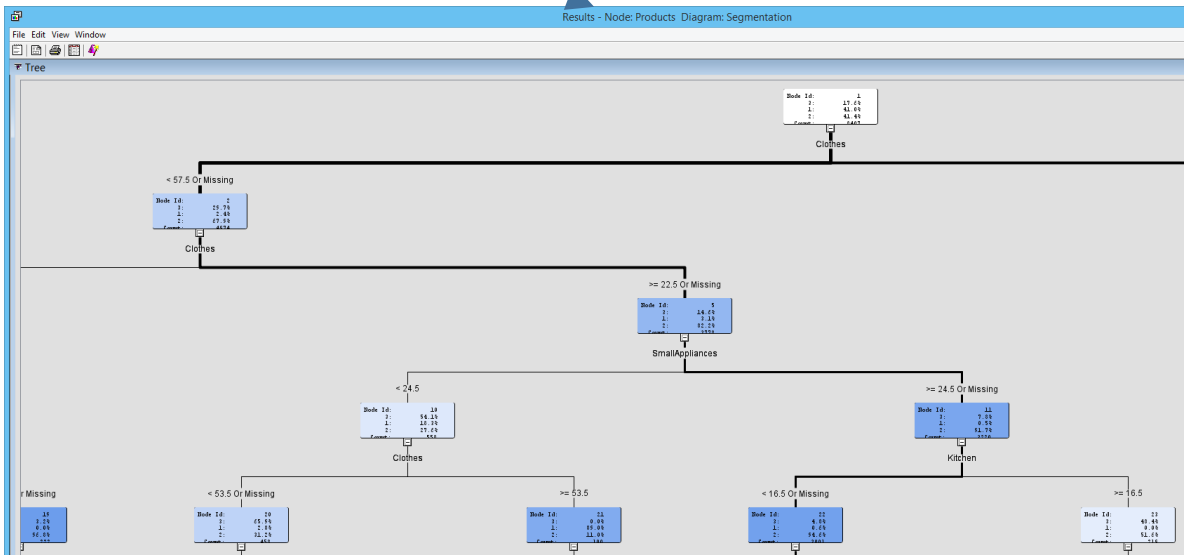
Value Segmentation

Name	Use	Report	Role	Level
Recomendation	Yes	No	Input	Ordinal
Income	Yes	No	Input	Interval
Rcn	Yes	No	Input	Interval
AGE_Years	Yes	No	Input	Interval
PerCatPurch	Yes	No	Input	Interval
Mnt_Inc	Yes	No	Input	Interval
Mnt	Yes	No	Input	Interval
Frq	Yes	No	Input	Interval
Custid	No	No	ID	Nominal
Clothes	No	No	Input	Interval
HouseKeeping	No	No	Input	Interval
Gender	No	No	Input	Nominal
SmallAppliances	No	No	Input	Interval
Kitchen	No	No	Input	Interval
Marital_Status	No	No	Input	Nominal
Dependents	No	No	Input	Binary
Education	No	No	Input	Nominal
Toys	No	No	Input	Interval
PerNetPurch	No	No	Input	Interval

Project Development – Results

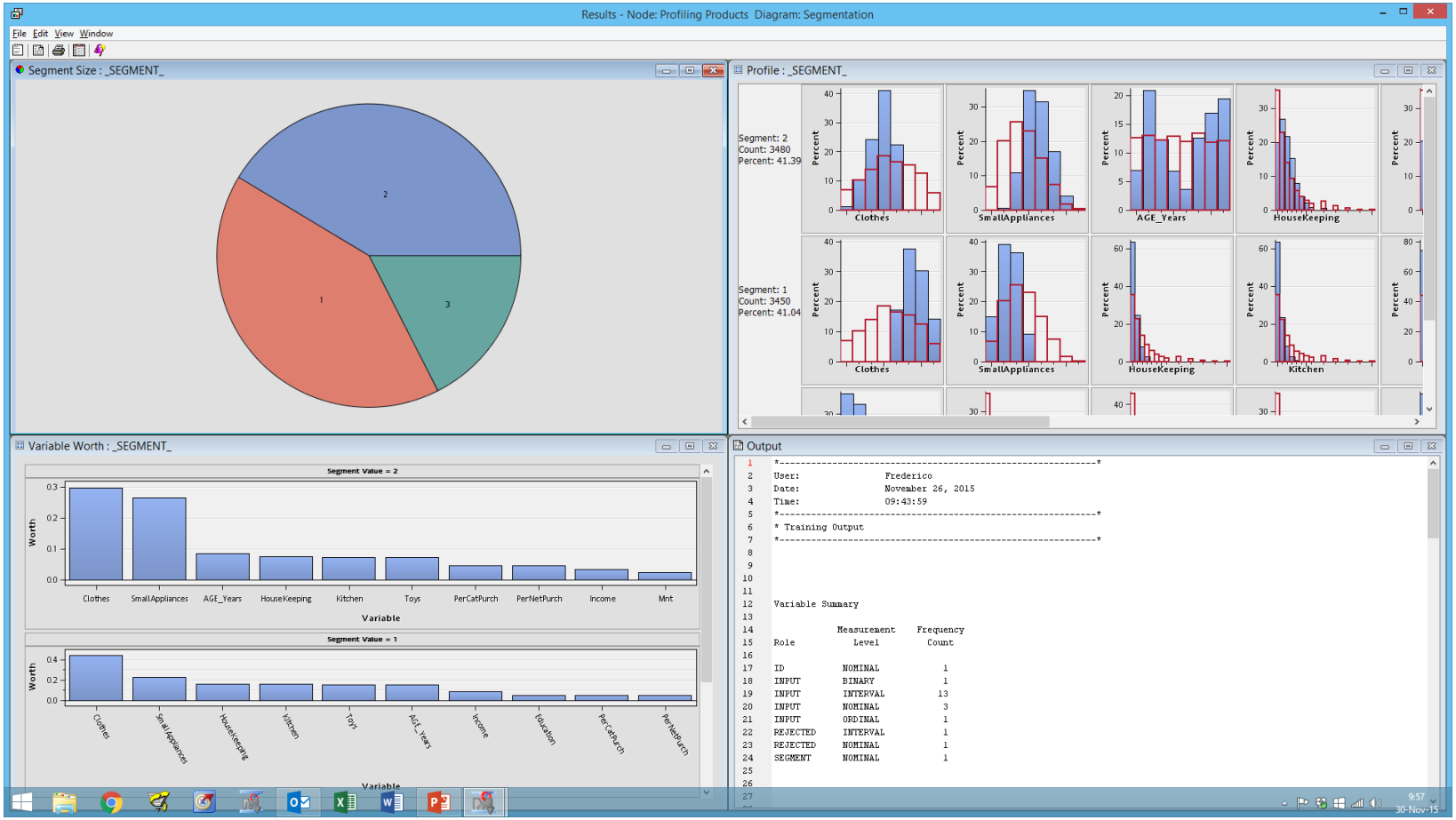
Assessment → Profiling Products





Project Development – Profiling

Assessment → Cluster



Project Development – Self Organizing Maps

Modelling / Segmentation → SOM/Kohonen

Property	Value
General	
Node ID	SOM
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Method	Batch SOM
Internal Standardization	Standardization
Segment	
Row	5
Column	5
Seed Options	
Initial Method	Principal Components
Radius	0.0
Batch SOM Training	
Use Defaults	Yes
Local-Linear Smoothing	Yes
Nadaraya-Watson Smoothing	Yes
Local-Linear Options	
Convergence Criterion	1.0E-4
Max Iterations	10
Nadaraya-Watson Options	
Convergence Criterion	1.0E-4
Max Iterations	10
KohonenVQ	
Maximum Number of Clusters	10
Kohonen	
Batch Training	No
Use Defaults	Yes
Kohonen Options	
Neighborhood Options	
Use Defaults	Yes
Neighborhood Options	
Encoding of Class Variables	
Ordinal Encoding	Default
Nominal Encoding	Default
Missing Values	
Interval Variables	Default
Nominal Variables	Default
Ordinal Variables	Default
Scoring Imputation Method	None
Score	
Segment Role	Segment
Exported Variables	All
Hide Original Variables	Yes

- SOM/Kohonen available methods and standardization of variables
- Cluster´s number definition (# cells in the U Matrix);
- Missing values´ handling;

Project Development – Self Organizing Maps

Assessment → SOM/Kohonen