

**NOVA**

**IMS**

Information  
Management  
School

# Data Mining

(Practical Classes)

MAA 2015/2016

## Enterprise Miner Practical Classes

Frederico Jesus  
[fjesus@novaaims.unl.pt](mailto:fjesus@novaaims.unl.pt)

# 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

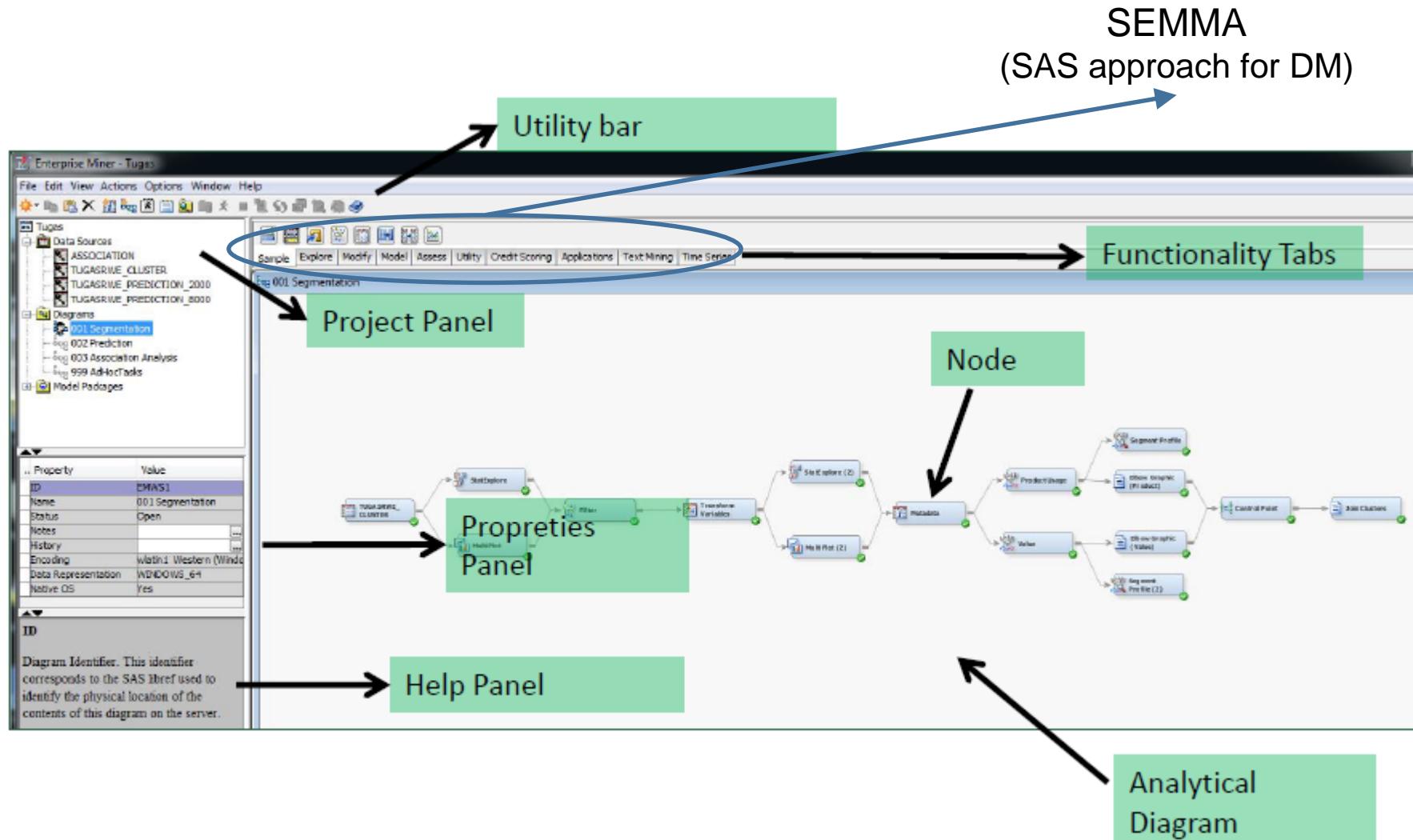
## Segmentation / Cluster Analysis:

- 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 properties and objectives;

# Agenda

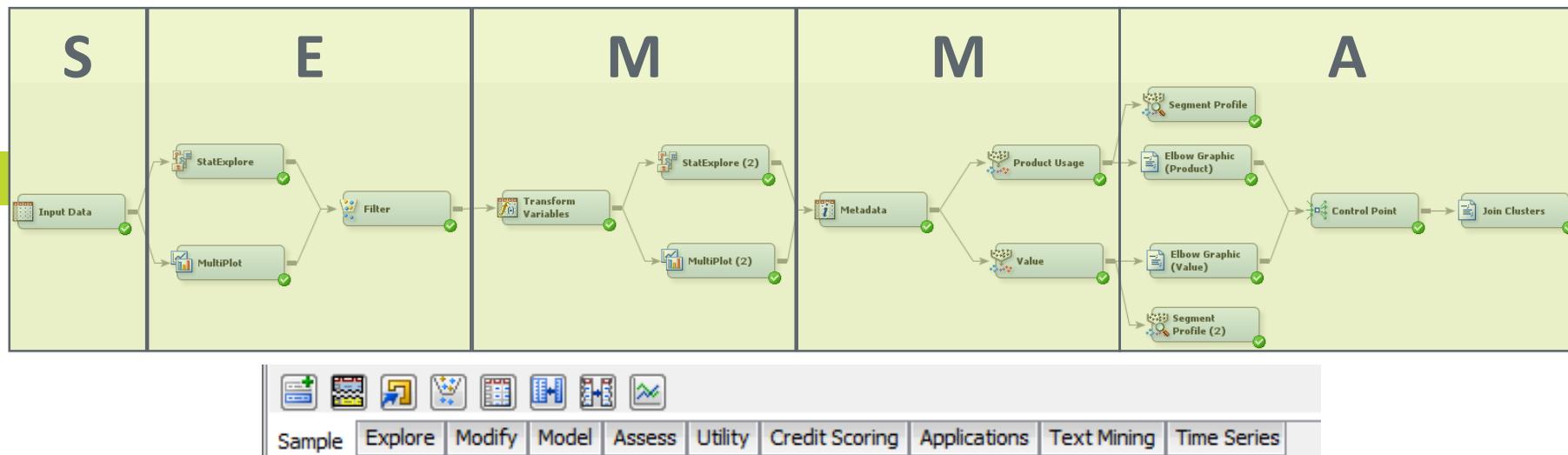
- i. The Data Mining Project: What are we going to do?
- ii. The SEMMA Approach
- iii. Introduction to SAS Enterprise Miner
  - i. Setting up a new SAS Miner Project
  - ii. Data sources
  - iii. SEMMA Process
- iv. Segmentation Problem

# Environment Overview



**SEMMA**  
(SAS approach for DM)

# Introduction to SAS EM (SEMMA)



SEMMA is an acronym for the following:

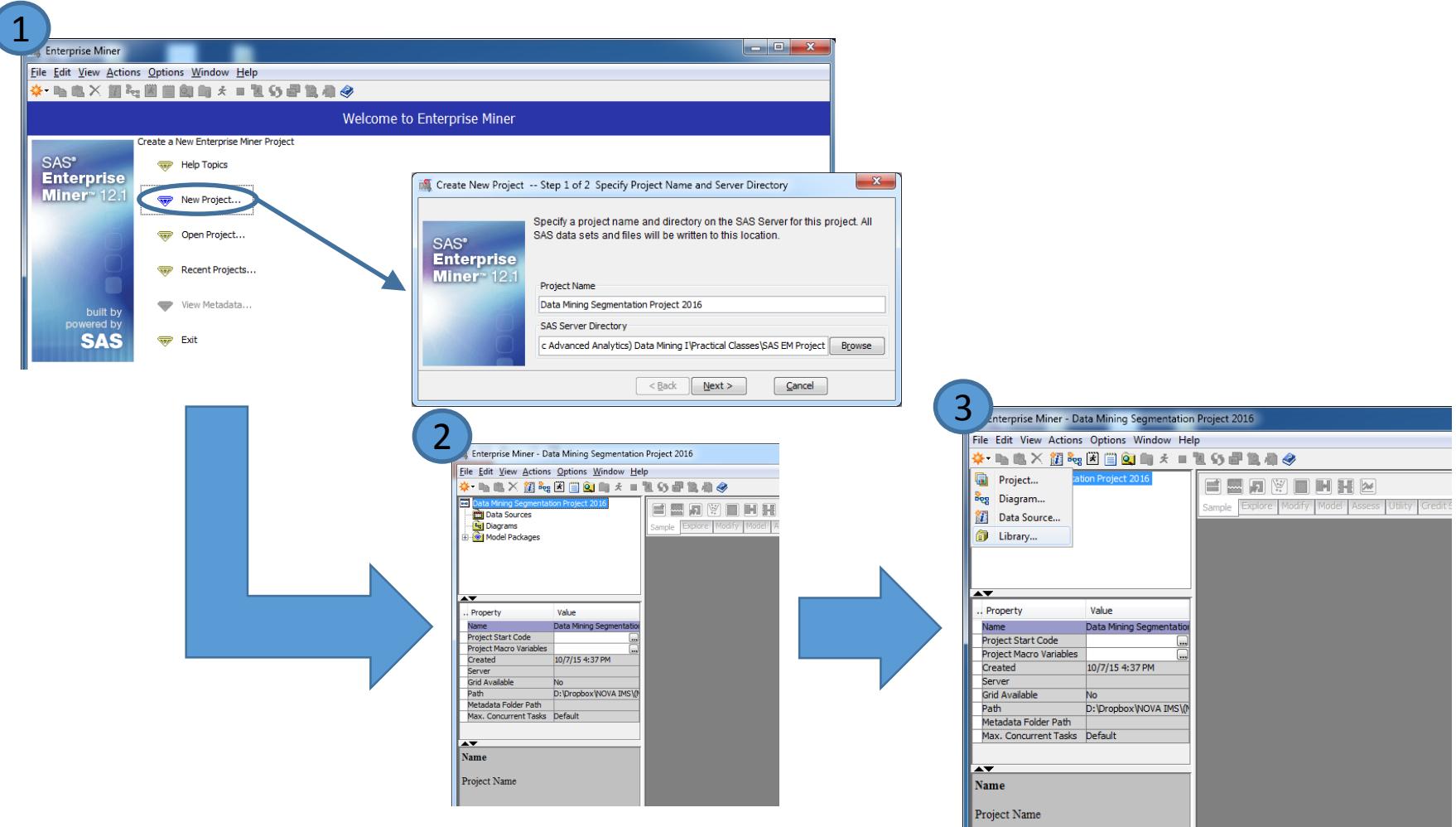
- Sample: Sample the data by creating one or more datasets. Datasets should be large enough to contain the significant information, yet small enough to process, especially where computing resources are (more) limited;
- Explore: Explore the data by searching for anticipated (hypothesized) relationships, 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

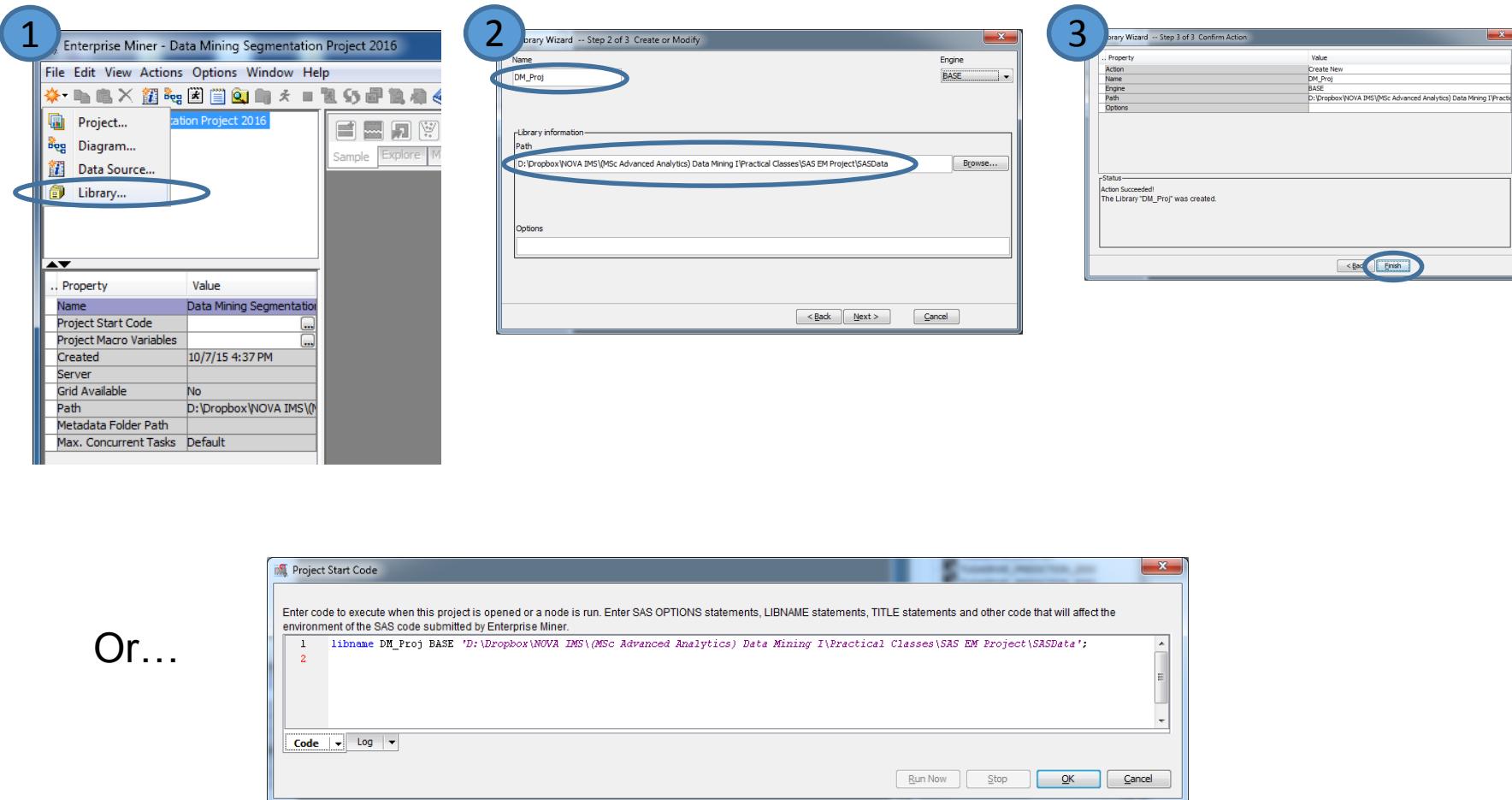
# Setting up a new Project

Open SAS Enterprise Miner Workstation 12.1....



# Setting up a new Project – SAS Library

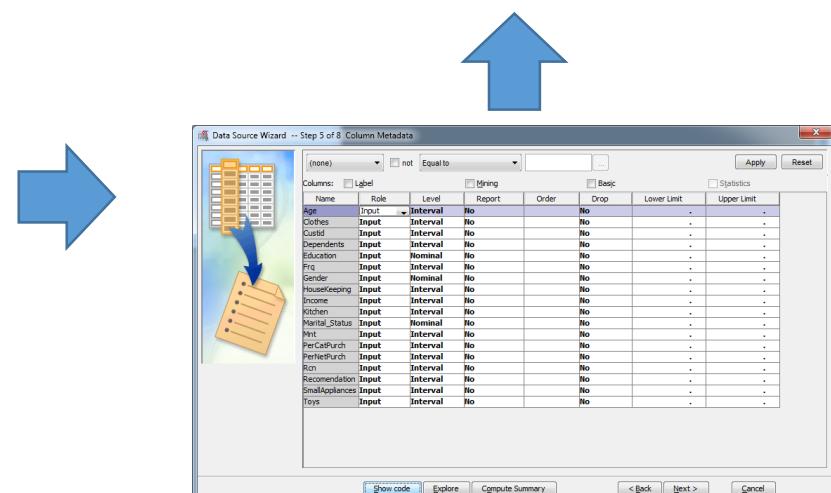
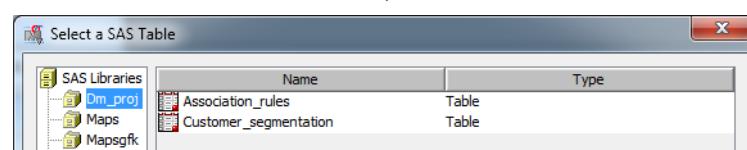
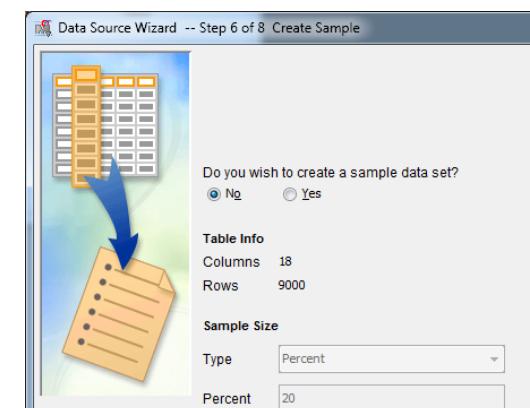
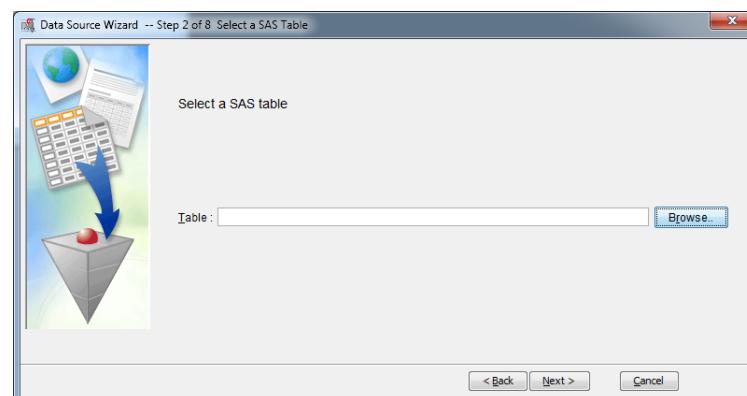
## Assign the libraries



Or...

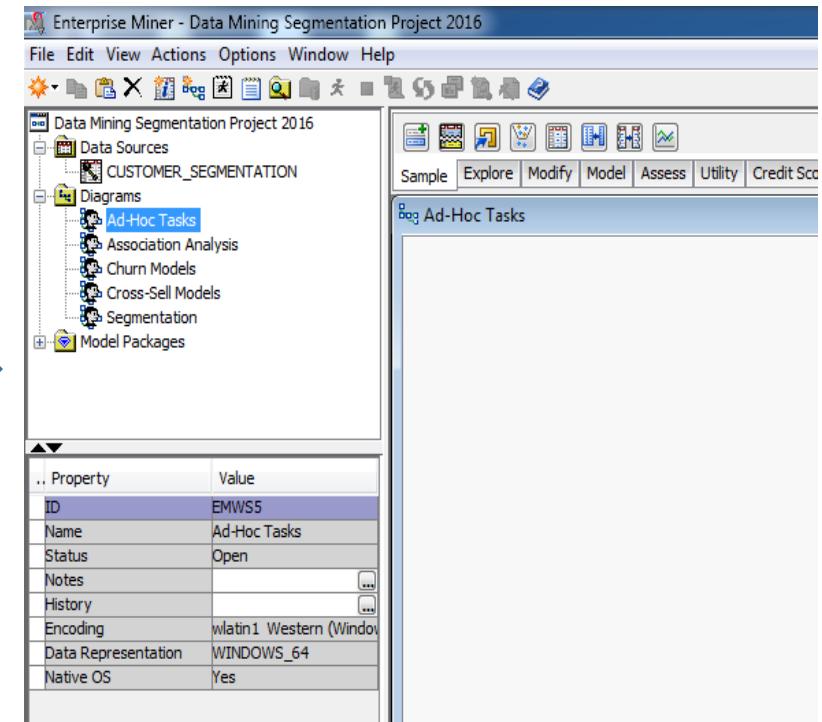
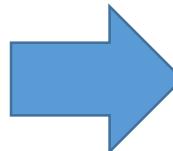
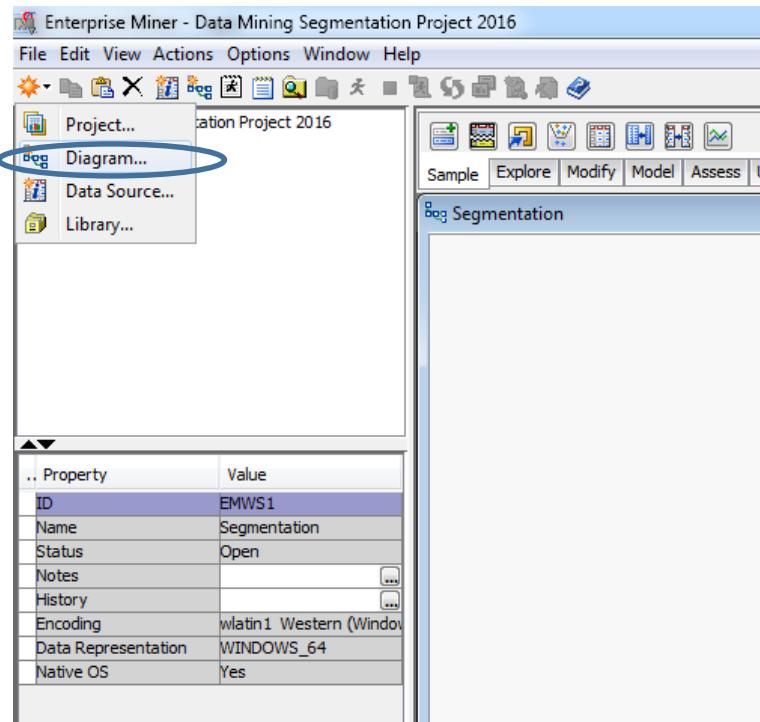
# Setting up a new Project – Import Data

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

.. Property Value

**General**

- Node ID: Ids
- Imported Data
- Exported Data
- Notes

**Train**

- Output Type: View
- Role: Raw
- Rerun: No
- Summarize: No
- Drop Map Variables: No

**Columns**

- Variables
- Decisions
- Refresh Metadata
- Advisor: Basic
- Advanced Options

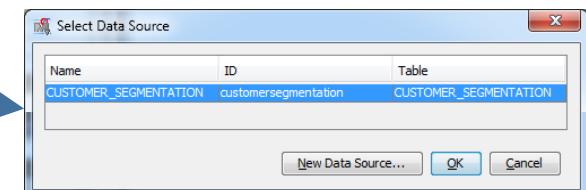
**Data**

- Data Selection: Data Source (circled)
- Sample: Default (circled)
- Sample Options

**Data Source**

- Data Source: CUSTOMER\_SEGMENTATION (circled)
- Data Source Properties
- New Table
- Table Name

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- 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;

# 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

.. Property Value

**General**

- Node ID: Ids
- Imported Data
- Exported Data
- Notes

**Train**

- Output Type: View
- Role: Raw
- Rerun: No
- Summarize: No
- Drop Map Variables: No

**Columns**

- Variables
- Decisions
- Refresh Metadata
- Advisor: Basic

**Data**

- Data Selection: Data Source
- Sample: Default
- Sample Options

**Data Source**

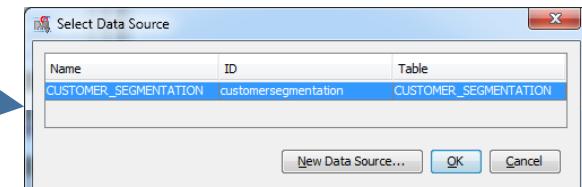
- Data Source: CUSTOMER\_SEGMENTATION
- Data Source Properties
- New Table
- Table Name

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

Data Access & Exploration

Customer Segmentation

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

Data Access &amp; Exploration

Customer  
Segmentation

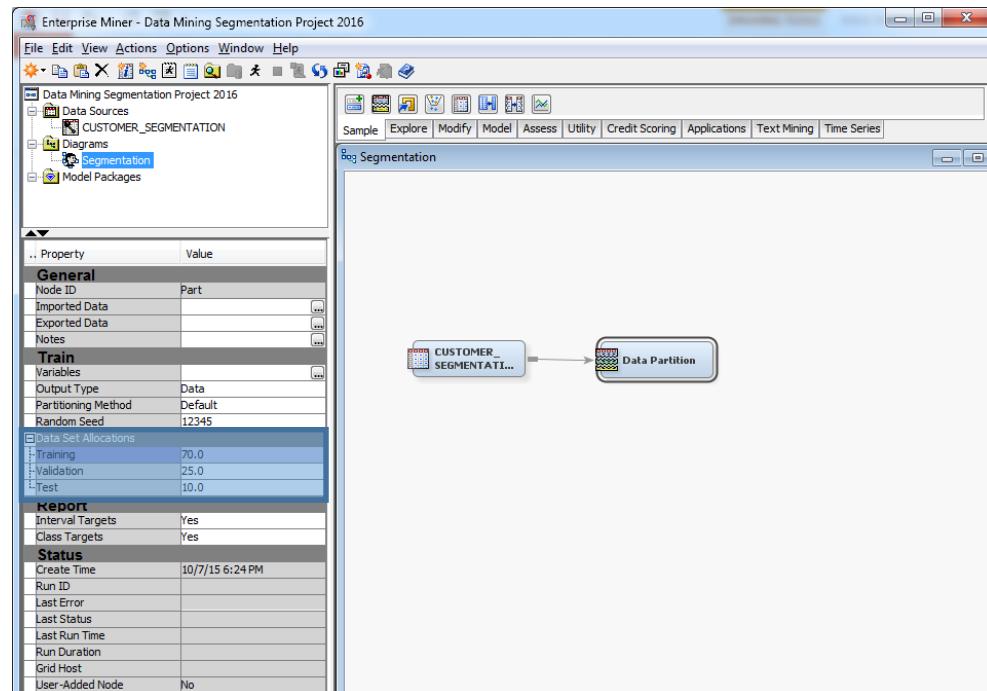
## Variable Roles:

- 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 have “ratios”.

# Project Development – Data Partition



Data Access & Exploration

Data Partition

.. Property	Value
<b>General</b>	
Node ID	Part
Imported Data	
Exported Data	
Notes	
<b>Train</b>	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
<b>Data Set Allocations</b>	
Training	70.0
Validation	25.0
Test	10.0
<b>Report</b>	
Interval Targets	Yes
Class Targets	Yes
<b>Status</b>	
Create Time	10/7/15 6:24 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

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

Data Access & Exploration



Property	Value
<b>General</b>	
Node ID	Stat2
Imported Data	[...]
Exported Data	[...]
Notes	[...]
<b>Train</b>	
Variables	[...]
<b>Data</b>	
Number of Observations	100000
Validation	No
Test	No
<b>Standard Reports</b>	
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No
Cross-Tabulation	[...]
<b>Variable Selection</b>	
Hide Rejected Variables	Yes
Number of Selected Variables	1000
<b>Chi-Square Statistics</b>	
Chi-Square	Yes
Interval Variables	No
Number of Bins	5
<b>Correlation Statistics</b>	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No
<b>Status</b>	
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



## Cross Tabulations:

- Education \* Dependents;
- Dependents \* Gender;

## Chi Square and Correlation Statistics:

- Useful in presence of target variables

# Project Development – StatExplore

Data Access & Exploration



Class Variables										
Data Role	Variable Name	Level	CODE	Frequency Count	Type	Percent	Level Index	Role	Label	Plot
TRAIN	Education	Graduation		0	4430C	49.22222	4INPUT	Education	1	
TRAIN	Education	2nd Cycle		5	1496C	16.62222	3INPUT	Education	1	
TRAIN	Education	Master		1	1304C	14.48889	5INPUT	Education	1	
TRAIN	Education	1st Cycle		3	1104C	12.26667	2INPUT	Education	1	
TRAIN	Education	PhD		4	593C	6.588889	7INPUT	Education	1	
TRAIN	Education			6	47C	0.522222	1INPUT	Education	1	
TRAIN	Education	OldSchool		2	26C	0.288889	6INPUT	Education	1	
TRAIN	Gender	M		0	5784C	64.26667	3INPUT	Gender	1	
TRAIN	Gender	F		1	3214C	35.71111	2INPUT	Gender	1	
TRAIN	Gender	?		2	2C	0.022222	1INPUT	Gender	1	
TRAIN	Marital_Status	Married		2	3273C	36.36667	3INPUT	Marital_Status	1	
TRAIN	Marital_Status	Single		3	2294C	25.48889	4INPUT	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
1TRAIN	HouseKee...		4	0	9000	0	77	6.929667	7.881981	2.229297	6.887084
2TRAIN	Toys		4	0	9000	0	62	7.037	7.924137	2.095795	5.643974
3TRAIN	Rcn		53	0	9000	0	549	62.46233	69.7566	4.174275	21.10057
4TRAIN	Kitchen		4	0	9000	0	75	7.039556	7.847558	2.049489	5.620054
5TRAIN	Mnt		383	0	9000	6	3052	622.1437	646.7193	0.981047	-0.05771
6TRAIN	Dependents		1	282	8718	0	1	0.707272	0.455041	-0.91121	-1.16996
7TRAIN	Frq		17	0	9000	3	59	19.84811	10.90239	0.697817	-0.41357
8TRAIN	Custid		6004	0	9000	1001	11000	6000.566	2887.654	-0.00373	-1.20329
9TRAIN	Clothes		51	0	9000	1	99	50.447	23.42034	-0.07825	-0.91836
10TRAIN	SmallAppli		28	0	9000	1	74	28.524	12.58512	0.314701	-0.42248
11TRAIN	PerNetPurch		45	0	9000	4	88	42.42989	18.49656	-0.2663	-1.0349
12TRAIN	Income		70021	46	8954	10000	140628	60960.04	27592.16	0.008695	-0.9239
13TRAIN	PerCatPurch		55	0	9000	12	96	57.57011	18.49656	0.266305	-1.0349
14TRAIN	Recomend...		4	0	9000	1	5	3.433	1.017876	-0.13016	-1.05877
15TRAIN	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



<b>General</b>	
Node ID	Plot
Imported Data	<input type="button" value="..."/>
Exported Data	<input type="button" value="..."/>
Notes	<input type="button" value="..."/>
<b>Train</b>	
Variables	<input type="button" value="..."/>
Type of Charts	Bar Charts
<input checked="" type="checkbox"/> Bar Chart Options	
Graph Orientation	Vertical
Include Missing Values	Yes
Interval Target Charts	Mean
Show Values	Yes
Statistic	Freq
Numeric Threshold	20
<input checked="" type="checkbox"/> Scatter Options	
Confidence Interval	Yes
Regression Equation	No
Regression Type	Linear
<b>Status</b>	
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);

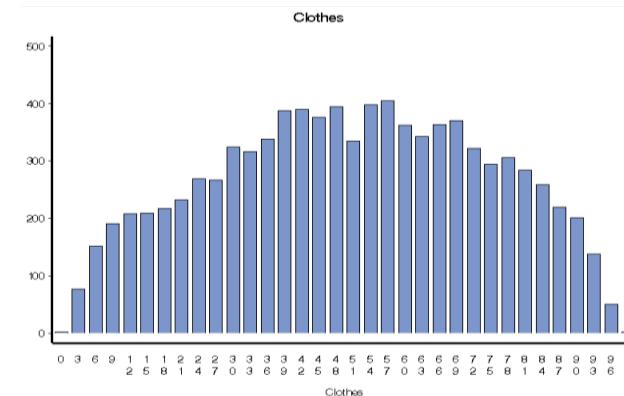
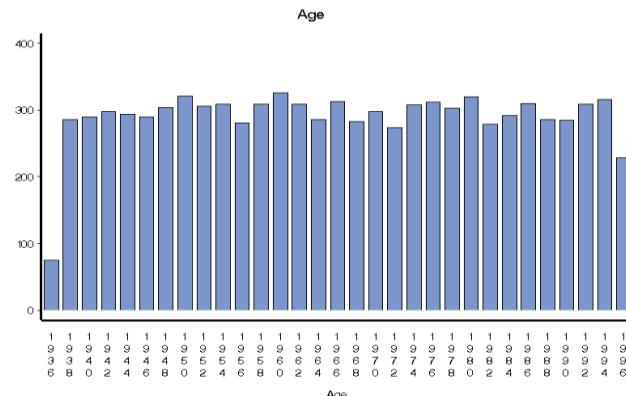
# Project Development – MultiPlot

Histograms / Frequency diagrams allow us to:

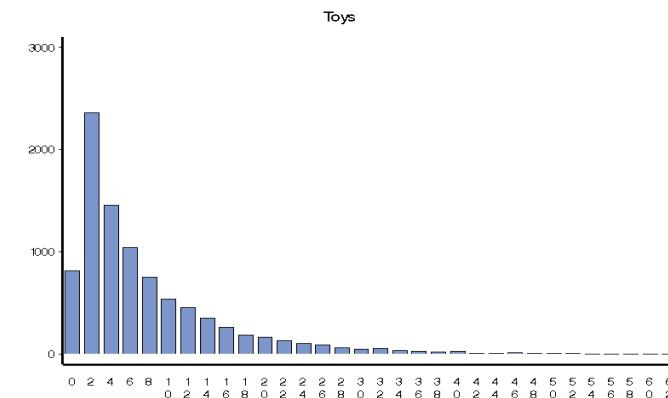
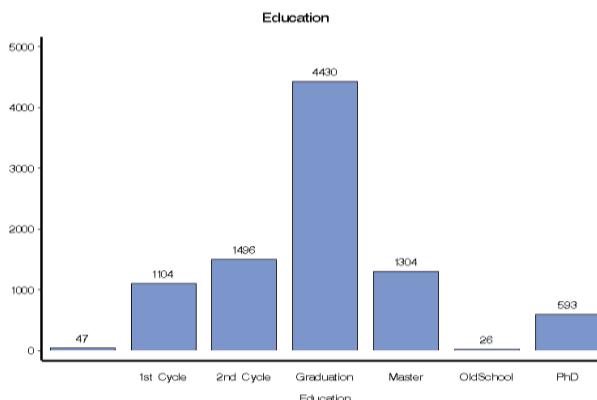
Data Access & Exploration



Assess data's distribution



Identify inconsistencies and potential outliers in our data



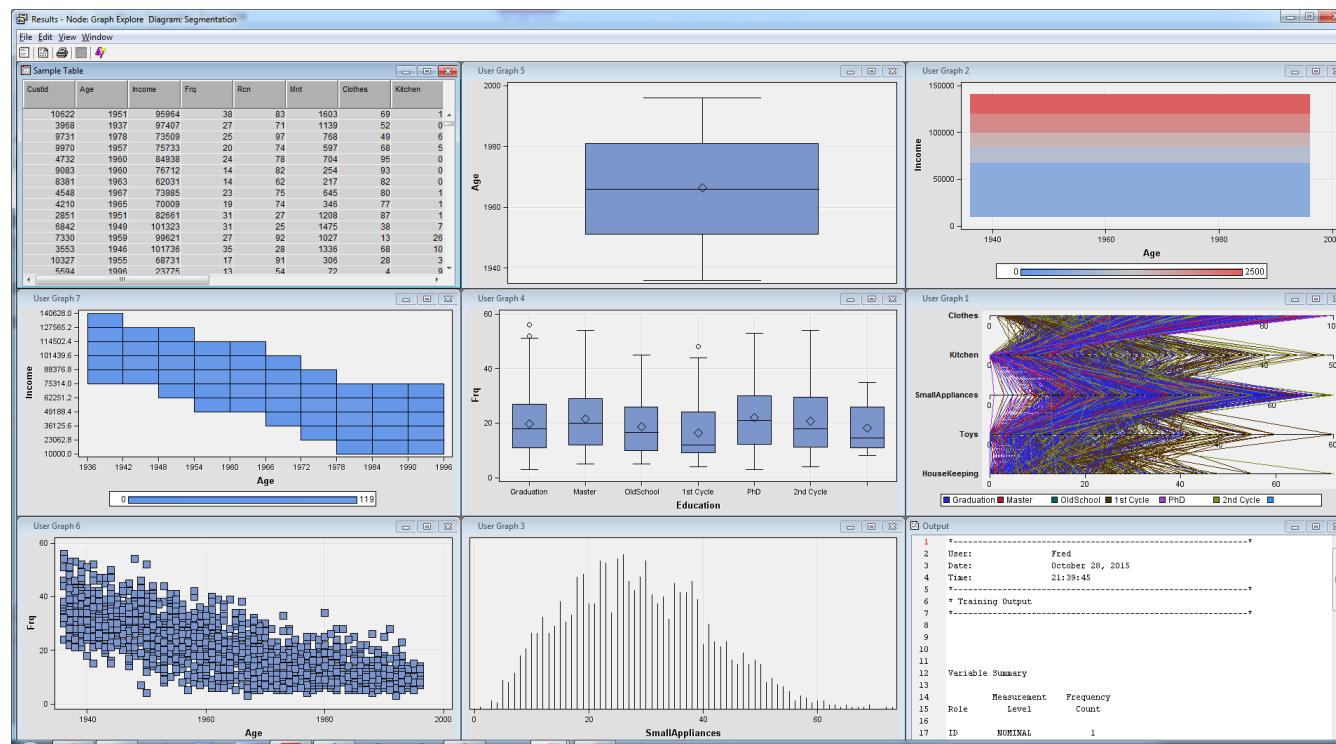
# Project Development – GraphExplore

Data Access & Exploration



## 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



- 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

Data Preparation

Filter

.. Property	Value
<b>General</b>	
Node ID	Filter
Imported Data	
Exported Data	
Notes	
<b>Train</b>	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
<input type="checkbox"/> Class Variables	
<input type="checkbox"/> 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
<input type="checkbox"/> Interval Variables	
<input type="checkbox"/> Interval Variables	
Default Filtering Method	None
Keep Missing Values	Yes
Tuning Parameters	
<b>Score</b>	
Create score code	Yes
Update Measurement Level	No
<b>Status</b>	
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



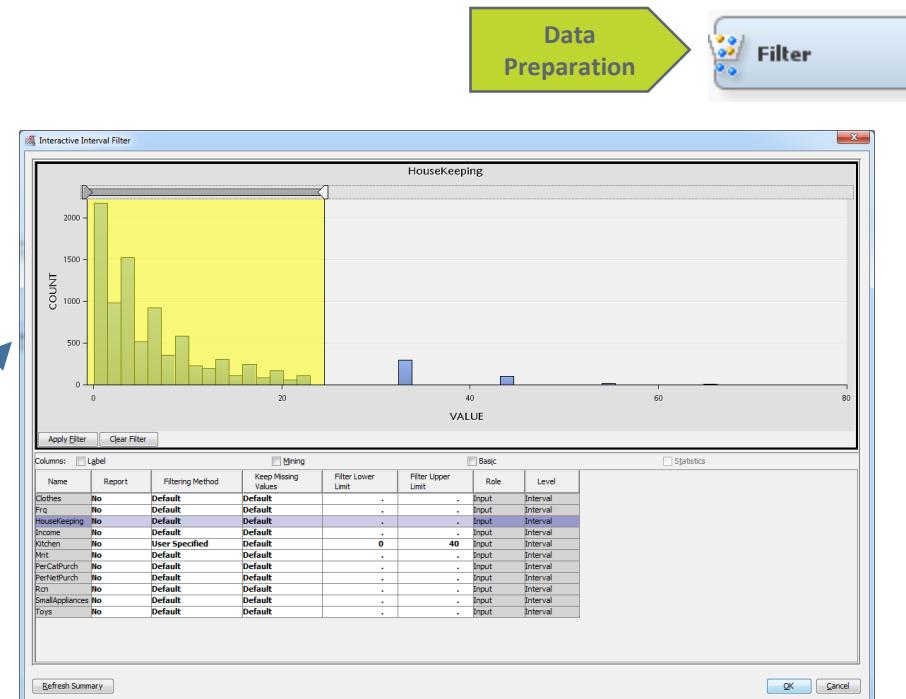
## Class Variables:

- Manual selection of values to exclude.

# Project Development – Filter

.. Property	Value
<b>General</b>	
Node ID	Filter
Imported Data	<input type="button" value="..."/>
Exported Data	<input type="button" value="..."/>
Notes	<input type="button" value="..."/>
<b>Train</b>	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
<input type="checkbox"/> Class Variables	
Class Variables	<input type="button" value="..."/>
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
<input type="checkbox"/> Interval Variables	
Interval Variables	<input type="button" value="..."/>
Default Filtering Method	None
Keep Missing Values	Yes
Tuning Parameters	<input type="button" value="..."/>
<b>Score</b>	
Create score code	Yes
Update Measurement Level	No
<b>Status</b>	
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

Data  
Preparation

 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	[...]
<b>Train</b>	
Interval Variables	[...]
Replacement Editor	[...] (circled)
Default Limits Method	None
Cutoff Values	[...]
Class Variables	[...]
Replacement Editor	[...]
Unknown Levels	Ignore
<b>Score</b>	
Replacement Values	Computed
Hide	No
<b>Report</b>	
Replacement Report	Yes
<b>Status</b>	
Create Time	10/28/15 11:32 PM
Run ID	b61d0b8f-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

The diagram illustrates the process of using the Filter Node. A blue arrow labeled '1' points from the 'Replacement Editor' button in the General tab of the Filter Node to the 'Interactive Replacement Interval Filter' dialog. Another blue arrow labeled '2' points from the 'Replacement Editor' button in the 'Score' tab of the Filter Node to the 'Replacement Editor-WORK.OUTPUT' dialog.

**Interactive Replacement Interval Filter Dialog:**

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
Freq	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

**Replacement Editor-WORK.OUTPUT Dialog:**

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

# Project Development – Inpute

.. Property	Value
<b>General</b>	
Node ID	Impt
Imported Data	[...]
Exported Data	[...]
Notes	[...]
<b>Train</b>	
Variables	[...]
Non Missing Variables	No
Missing Cutoff	50.0
<b>Class Variables</b>	
Default Input Method	None
Default Target Method	Count
Normalize Values	Default Constant Value
<b>Interval Variables</b>	
Default Input Method	Distribution
Default Target Method	Tree
Default Constant Value	Tree Surrogate
	None
<b>Score</b>	
Default Input Method	None
Default Target Method	Tree
Default Constant Value	Tree Surrogate
Default Character Value	Mid-Minimum Spacing
Default Number Value	Tukey's Biweight
<b>Method Options</b>	
Random Seed	Huber
Tuning Parameters	Andrew's Wave
Tree Imputation	Default Constant Value
	None
<b>Score</b>	
Hide Original Variables	Yes
<b>Indicator Variables</b>	
Type	None
Source	Imputed Variables
Role	Rejected
<b>Report</b>	
Validation and Test Data	No
Distribution of missing	No



## 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
<b>General</b>	
Node ID	Trans
Imported Data	<input type="button" value="..."/>
Exported Data	<input type="button" value="..."/>
Notes	<input type="button" value="..."/>
<b>Train</b>	
Variables	<input type="button" value="..."/>
Formulas	<input type="button" value="..."/>
Interactions	<input type="button" value="..."/>
SAS Code	<input type="button" value="..."/>
<input type="checkbox"/> Default Methods	
Interval Inputs	None
Interval Targets	None
Class Inputs	None
Class Targets	None
Treat Missing as Level	No
<input type="checkbox"/> Sample Properties	
Method	First N
Size	Default
Random Seed	12345
<input type="checkbox"/> Optimal Binning	
Number of Bins	4
Missing Values	Use in Search
<input type="checkbox"/> Grouping Method	
Cutoff Value	0.1
Group Missing	No
Number of Bins	Variables
Add Minimum Value to Offset Value	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:

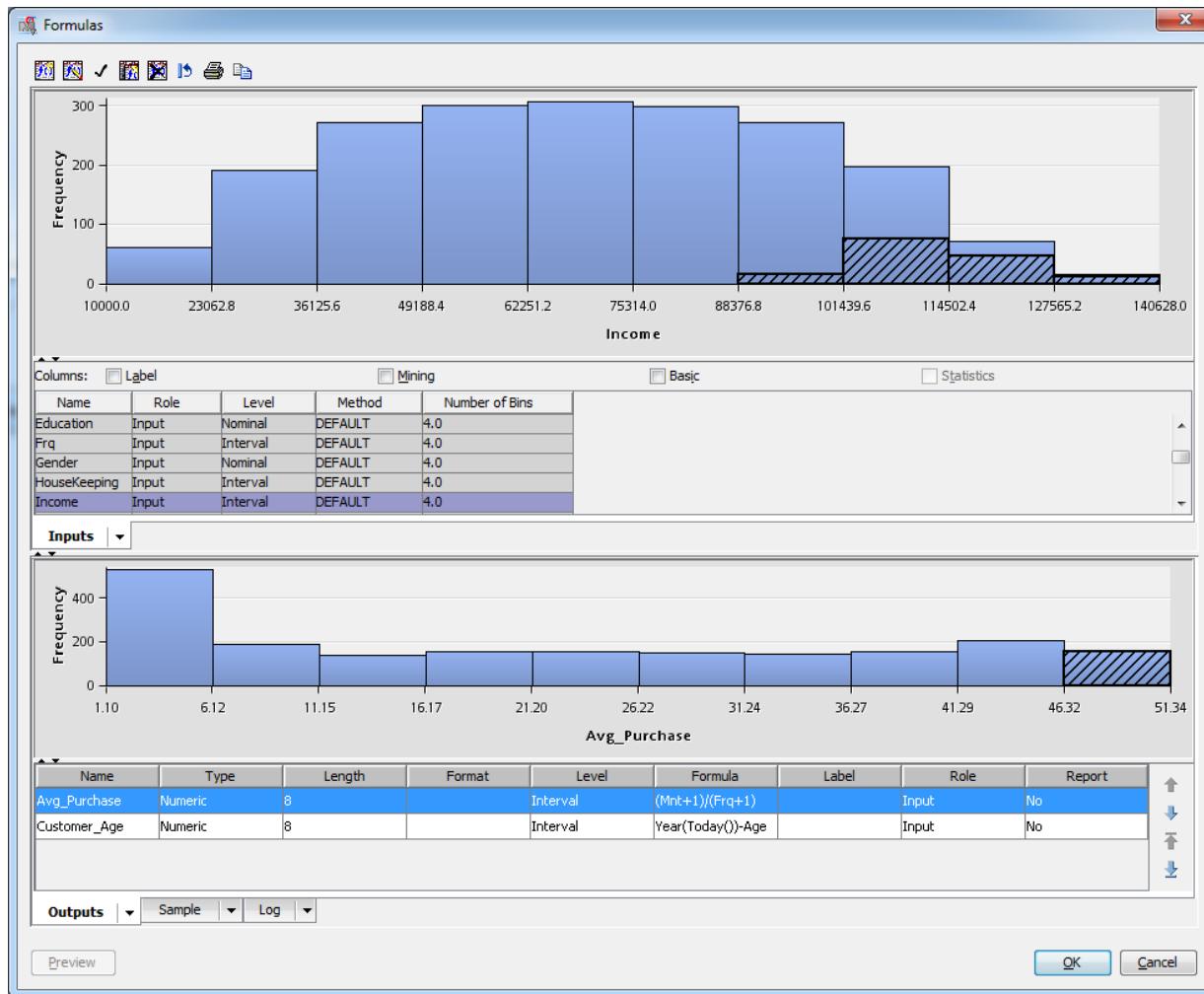
- Standardization; Optimal binning; quantiles...

# Project Development – Transform Variables

Business  
Needs

Data  
Preparation

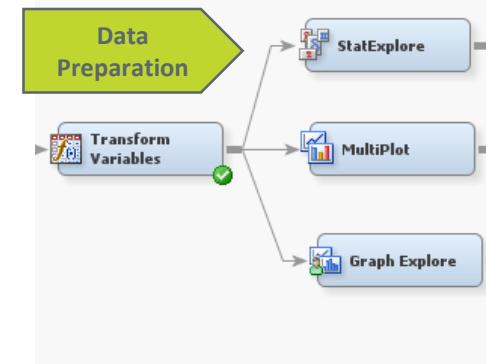
Transform  
Variables



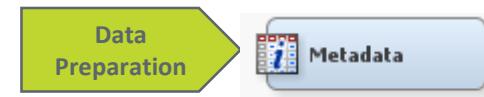
# Project Development – Metadata

## Applications:

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



# Project Development – Transform Variables



## Application:

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

.. Property	Value
<b>General</b>	
Node ID	Meta
Imported Data	...
Exported Data	...
Notes	...
<b>Train</b>	
Import Selection	...
Summarize	No
Advanced Advisor	No
<input type="checkbox"/> Rejected Variables	
<input type="checkbox"/> Hide Rejected Variables	No
<input type="checkbox"/> Combine Rule	None
<input type="checkbox"/> Variables	
<input type="checkbox"/> Train	...
<input type="checkbox"/> Transaction	...
<input type="checkbox"/> Validate	...
<input type="checkbox"/> Test	...
<input type="checkbox"/> Score	...
<b>Status</b>	
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	<b>Default</b>	Input	<b>Default</b>	Ordinal	<b>Default</b>	<b>Default</b>	<b>Default</b>
Clothes	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
Custid	N	<b>Default</b>	ID	<b>Default</b>	Nominal	<b>Default</b>	<b>Default</b>	<b>Default</b>
Dependents	N	<b>Default</b>	Input	<b>Default</b>	Binary	<b>Default</b>	<b>Default</b>	<b>Default</b>
Education	N	<b>Default</b>	Input	<b>Default</b>	Nominal	<b>Default</b>	<b>Default</b>	<b>Default</b>
Frq	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
Gender	N	<b>Default</b>	Input	<b>Default</b>	Nominal	<b>Default</b>	<b>Default</b>	<b>Default</b>
HouseKeeping	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
Income	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
Kitchen	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
Marital_Status	N	<b>Default</b>	Input	<b>Default</b>	Nominal	<b>Default</b>	<b>Default</b>	<b>Default</b>
Mnt	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
PerCatPurch	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
PerNetPurch	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
Rcn	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
Recomendation	N	<b>Default</b>	Input	<b>Default</b>	Ordinal	<b>Default</b>	<b>Default</b>	<b>Default</b>
SmallAppliances	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>
Toys	N	<b>Default</b>	Input	<b>Default</b>	Interval	<b>Default</b>	<b>Default</b>	<b>Default</b>