ENTERPRISE MINER:
1 – DATA EXPLORATION AND VISUALISATION

JOZEF MOFFAT, ANALYTICS & INNOVATION PRACTICE, SAS UK

10, MAY 2016

AGENDA

- **SAS Webinar – 10th May 2016 at 10:00 AM BST**
- **Enterprise Miner: Data Exploration and Visualisation**
- *The session looks at:*

  - Data Visualisation and Sampling

  - Variable Selection

  - Missing Value Imputation

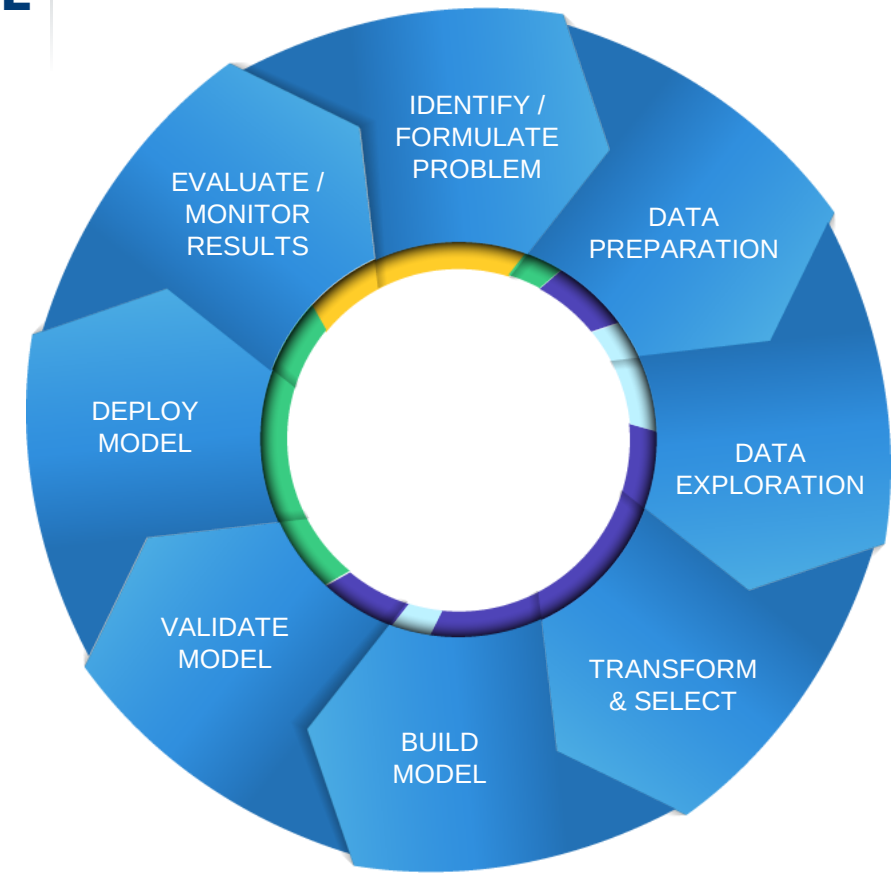
  - Outlier Detection

ROLE OF SAS ENTERPRISE MINER

THE ANALYTICS LIFECYCLE

PREDICTIVE ANALYTICS AND DATA MINING

Domain Expertise
Decision Making
Process and ROI Evaluation



Data Exploration
Data Visualization
Report Creation

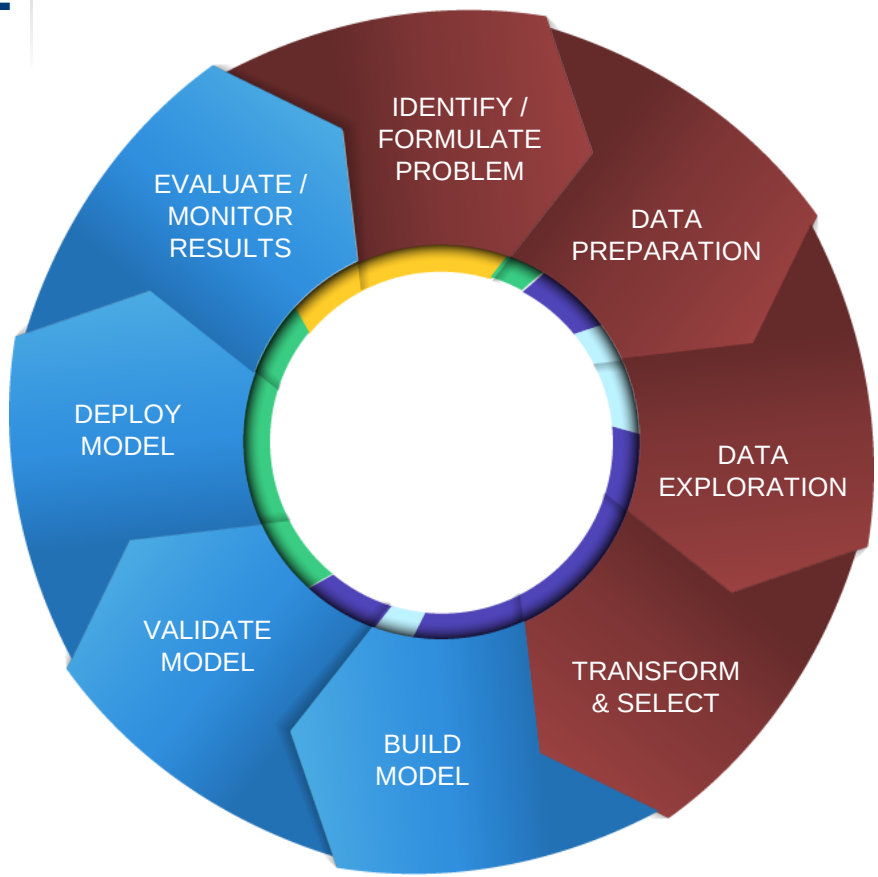
Model Validation
Model Deployment
Model Monitoring
Data Preparation

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

THE ANALYTICS LIFECYCLE

PREDICTIVE ANALYTICS AND DATA MINING

Domain Expertise
Decision Making
Process and ROI Evaluation



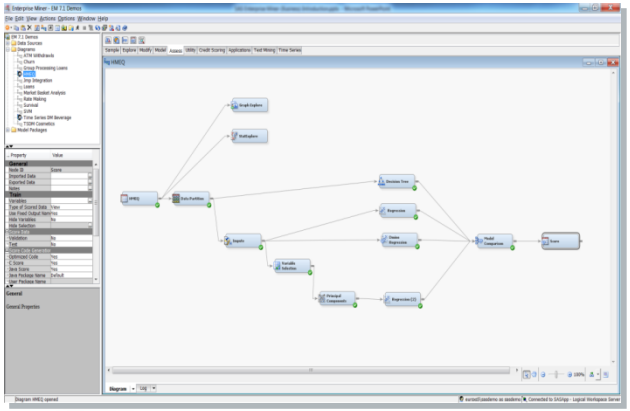
Data Exploration
Data Visualization
Report Creation

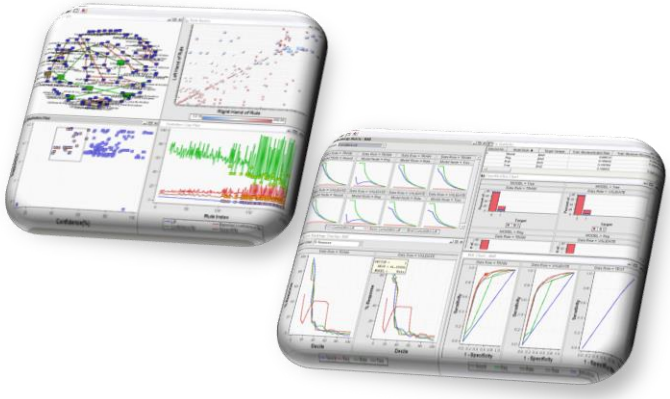
Model Validation
Model Deployment
Model Monitoring
Data Preparation

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

SAS® ENTERPRISE MINER™

- Modern, collaborative, easy-to-use data mining workbench
- Sophisticated set of data preparation and exploration tools
- Modern suite of modeling techniques and methods
- Interactive model comparison, testing and validation
- Automated scoring process delivers faster results
- Open, extensible design for ultimate flexibility

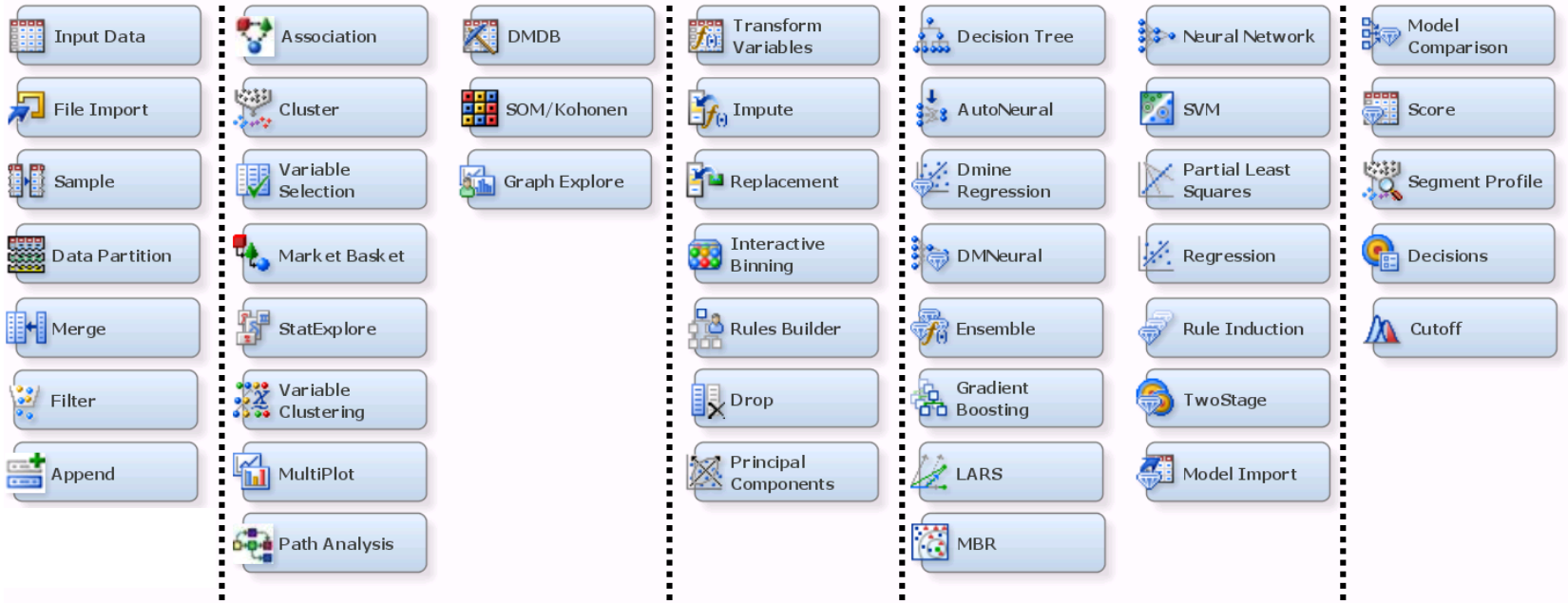




SAS® ENTERPRISE MINER™

SAS® ENTERPRISE MINER™
MODEL DEVELOPMENT PROCESS

Sample **E**xplore **M**odify **M**odel **A**ssess



THE POWER TO KNOW.

SAS® ENTERPRISE MINER™

SAS® ENTERPRISE MINER™

MODEL DEVELOPMENT PROCESS

Utility

- Metadata
- SAS Code
- Start Groups
- End Groups
- Control Point
- Reporter
- Score Code Export
- Ext Demo
- Open Source Integration
- Register Model

Apps.

- Incremental Response
- Survival

Time Series

- TS Correlation
- TS Data Preparation
- TS Decomposition
- TS Dimension Reduction
- TS Exponential Smoothing
- TS Similarity

HPDM

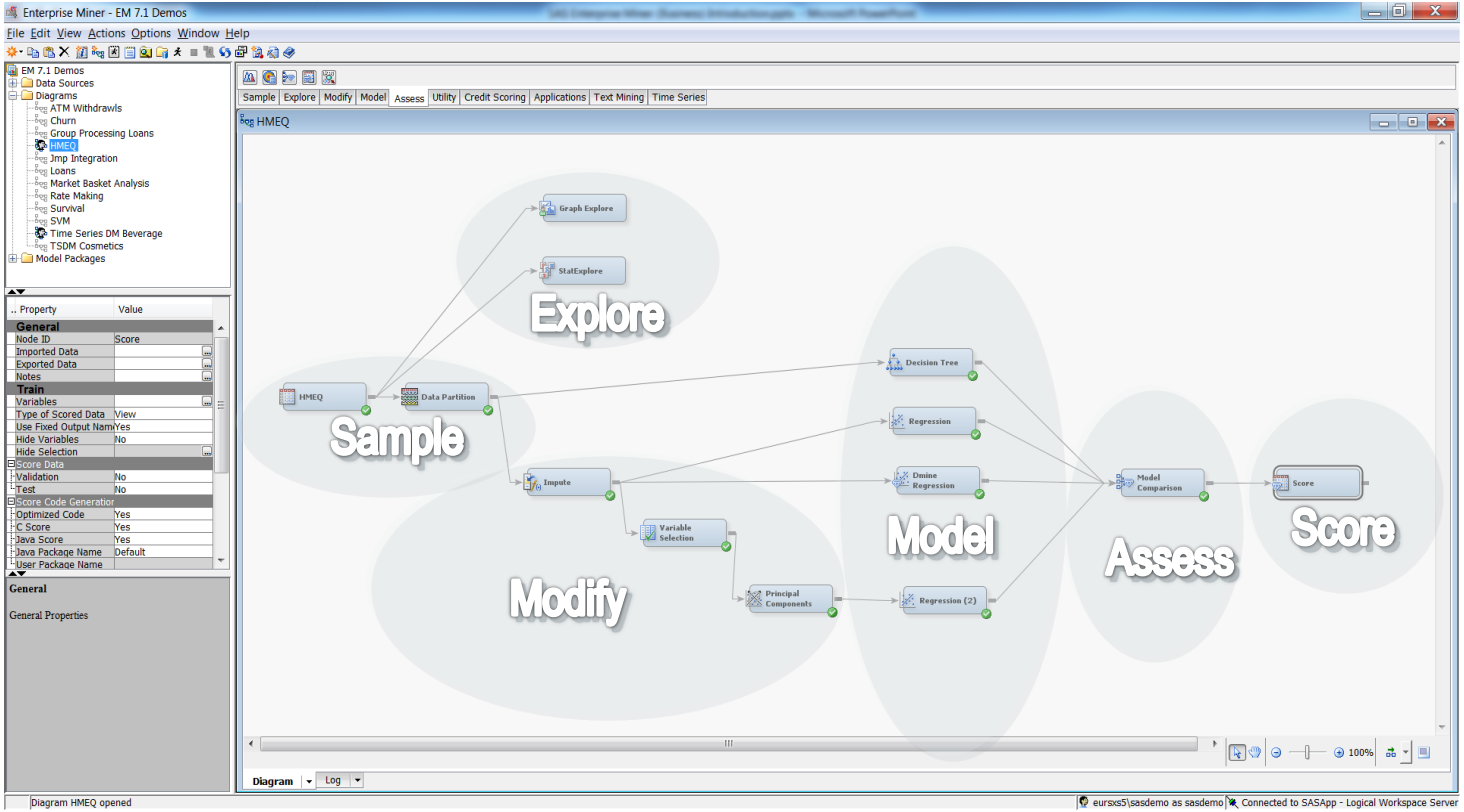
- HP Cluster
- HP Data Partition
- HP Explore
- HP Forest
- HP GLM
- HP Impute
- HP Neural
- HP Principal Components
- HP Regression
- HP SVM
- HP Text Miner
- HP Transform
- HP Tree
- HP Variable Selection

Credit Scoring

- Interactive Grouping
- Reject Inference
- Scorecard

SAS® ENTERPRISE MINER™

SEMMA IN ACTION – REPEATABLE PROCESS



DATA VISUALISATION AND SAMPLING

VISUALISATION

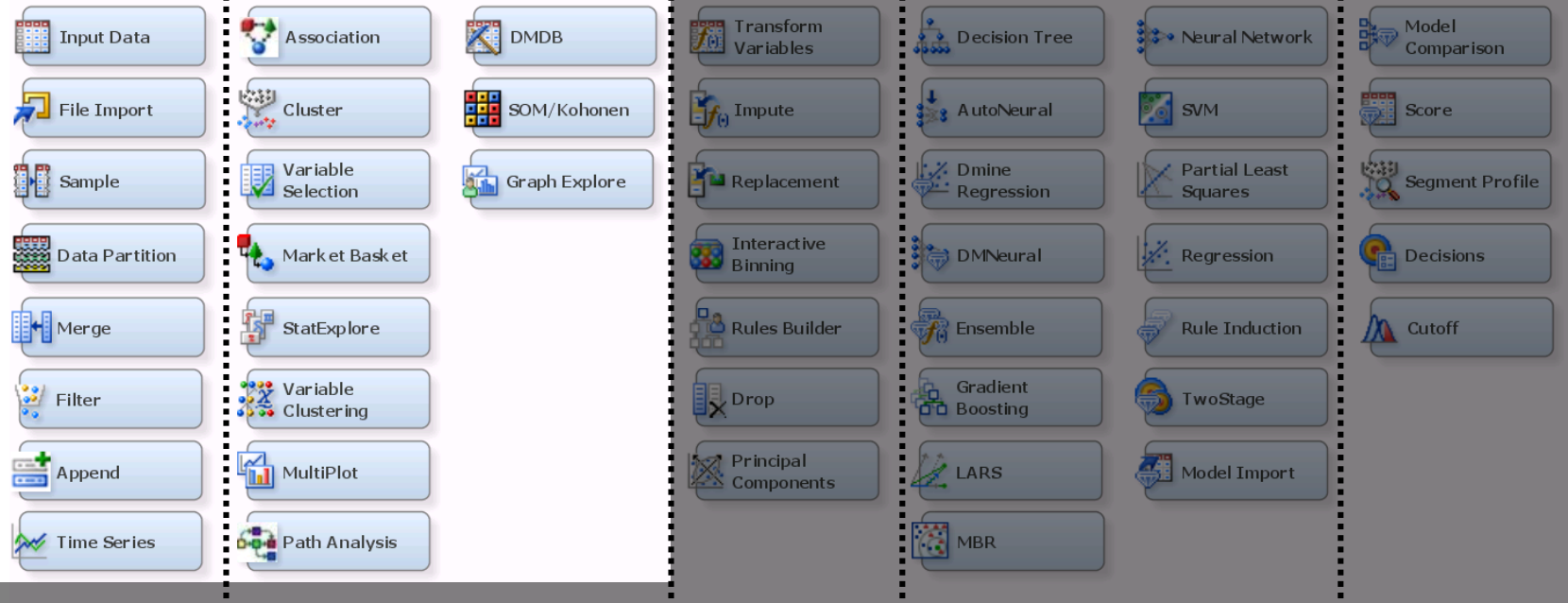
"Quickly find related patterns within a set of data via interactive pictures."

DATA VISUALISATION AND SAMPLING

SAS® ENTERPRISE MINER™

SEMMA PROCESS

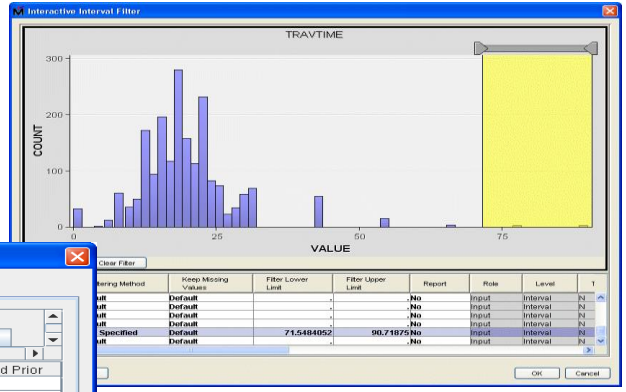
Sample **E**xplore **M**odify **M**odel **A**ssess

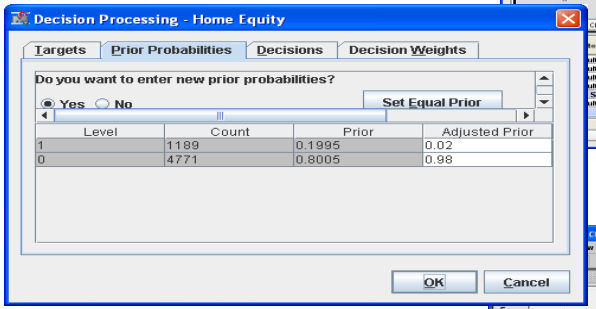


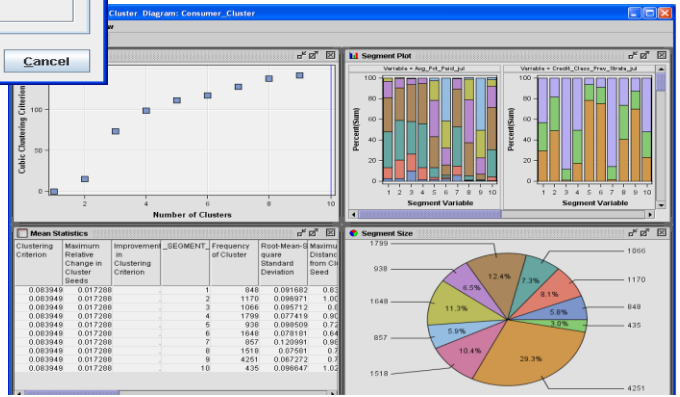
THE POWER TO KNOW.

SAMPLE AND EXPLORE

- Data selection
  - Required & excluded fields
  - Sample balancing
  - Data partitioning
- Data evaluation
  - Statistical measures
  - Visualization
  - Identifying outliers
  - Analytical segmentation
  - Variable creation & selection



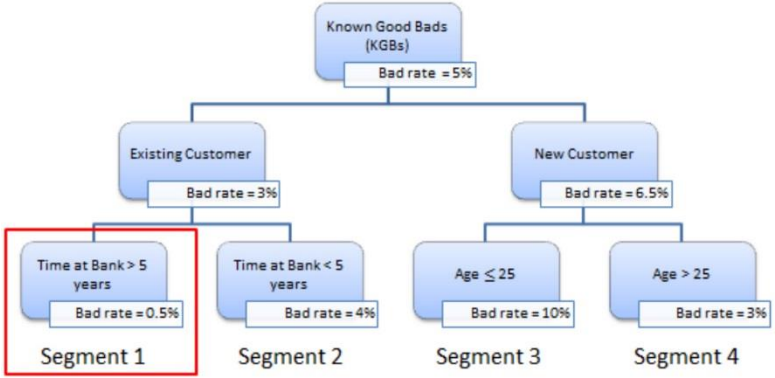


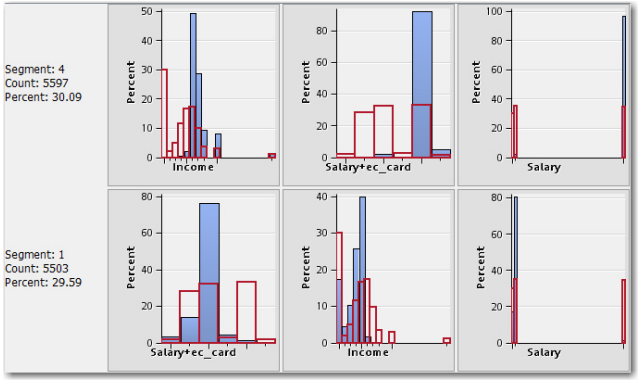


SAMPLING

- Sampling
  - **Sample** Node:
    - Stratified / Simple Random Sampling
    - Used for over/under sampling input data
  - **Data Partition** Node:
    - Random sampling into Training, Validation and Test sets
    - Prevent model over fitting
  - **Filter** Node:
    - Select time period of interest
    - Filter based on pre-defined flag

- Segmentation
  - **Cluster** Node
    - Unsupervised, k-means clustering algorithm
    - Data driven
    - Output tree based descriptions





DEMO

VARIABLE SELECTION

VARIABLE SELECTION ROUTINES

- Variable Selection
  - **Variable Selection** Node
    - Relationship of independent variables to dependent target
    - R-Square and Chi-square selection criteria
  - **Variable Clustering** Node
    - Identify correlations and covariance's between input variables
    - Select Best variable from cluster or Cluster Component
  - **Interactive Grouping** Node
    - Computes Weights of Evidence
    - GINI and Information Values for variable selection

DEMO

MISSING VALUE IMPUTATION

MISSING VALUE IMPUTATION

IMPUTE

- Missing Value Imputation
  - **Impute** Node
    - Complete case required for models such as Regression
    - Multiple imputation techniques e.g. Tree, Distribution, Mean, Mode

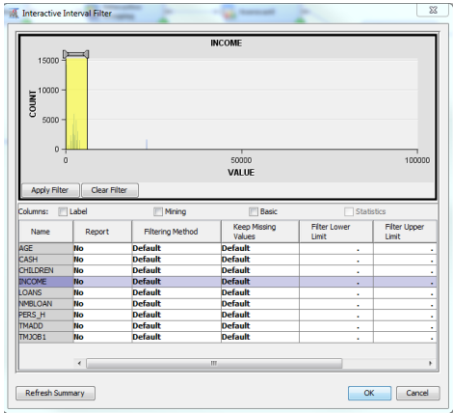
Class (categorical) variables	Interval (numeric) variables
Input/Target	*Input/Target*
Count	Mean
Default Constant Value	Median
Distribution	Midrange
Tree (only for inputs)	Distribution
Tree Surrogate (only for inputs)	Tree (only for inputs)
	Tree Surrogate (only for inputs)
	Mid-Minimum Spacing
	Tukey's Biweight
	Huber
	Andrew's Wave
	Default Constant

DEMO

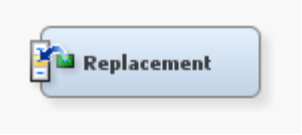
OUTLIER DETECTION

OUTLIER DETECTION

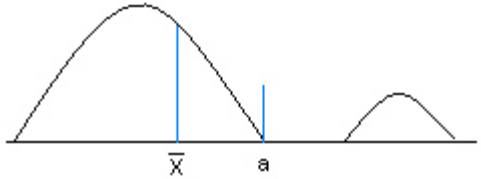
DETECT AND TREAT



- Outlier Detection
  - **Filter** Node
    - Automated and Interactive filtering
    - Identify and exclude extreme outliers
  - **Replacement** Node
    - Generates score code to process unknown levels when scoring
    - Interactively specify replacement values for class and interval levels



Replace any value > a with the mean.



DEMO

SUMMARY

SUMMARY

- Comprehensive data mining toolset
  - Variety of visualisation and sampling methodologies
  - Number of approaches to data and dimension reduction
  - Importance of enhancing data prior to model development
- Garbage in = Garbage out (GIGO)

QUESTIONS AND ANSWERS

JOZEF.MOFFAT@SAS.COM

THE POWER TO KNOW.

THE POWER TO KNOW.

www.SAS.com