



SAS® FORUM  
UNITED KINGDOM 2015

# Data Exploration and Visualisation in SAS Enterprise Miner

Dr Iain Brown, Senior Analytics Specialist Consultant,  
SAS UK & Ireland

# Agenda

- **SAS Presents – Thursday 11<sup>th</sup> June 2015 – 14:30**
- **Data Exploration and Visualisation in SAS Enterprise Miner**
- *The session looks at:*
  - *Data Visualisation and Sampling*
  - *Variable Selection*
  - *Missing Value Imputation*
  - *Outlier Detection*

# The Analytics Lifecycle

## BUSINESS MANAGER

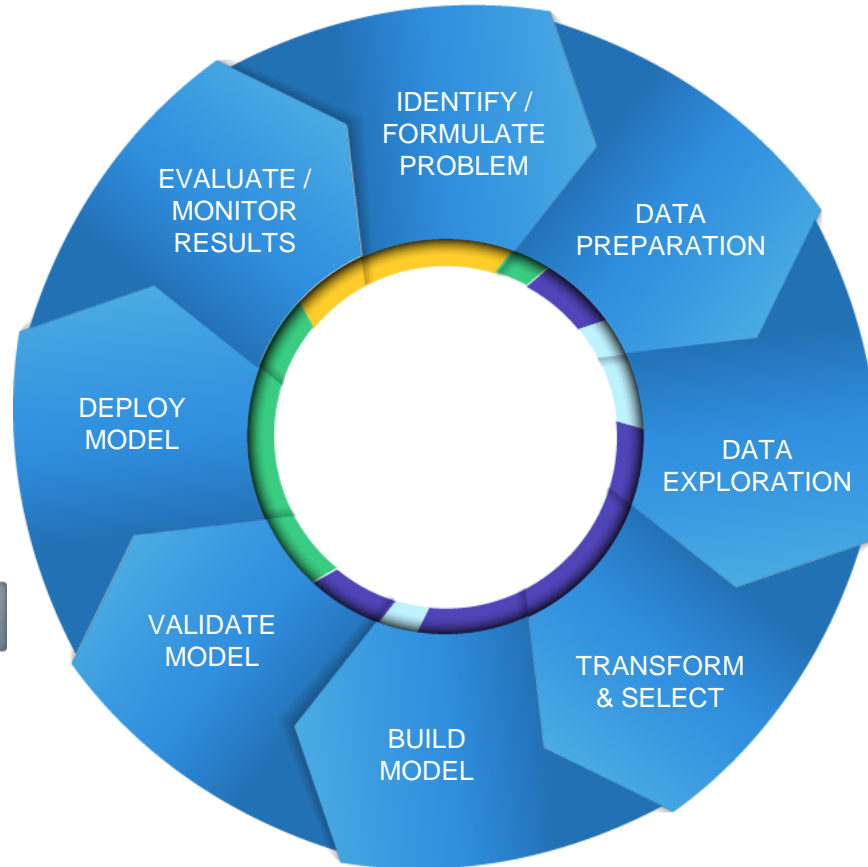


Domain Expert  
Makes Decisions  
Evaluates Processes and ROI

## IT SYSTEMS / MANAGEMENT



Model Validation  
Model Deployment  
Model Monitoring  
Data Preparation



## BUSINESS ANALYST



Data Exploration  
Data Visualization  
Report Creation

## DATA MINER / STATISTICIAN



Exploratory Analysis  
Descriptive Segmentation  
Predictive Modeling

# The Analytics Lifecycle

## BUSINESS MANAGER

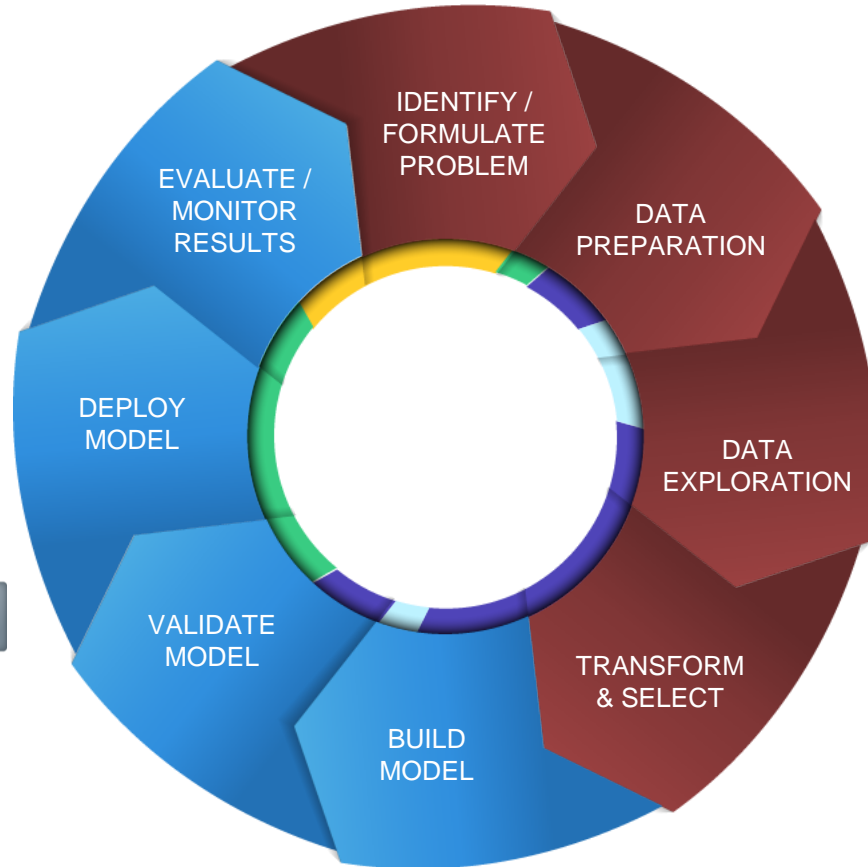


Domain Expert  
Makes Decisions  
Evaluates Processes and ROI

## IT SYSTEMS / MANAGEMENT



Model Validation  
Model Deployment  
Model Monitoring  
Data Preparation



## BUSINESS ANALYST



Data Exploration  
Data Visualization  
Report Creation

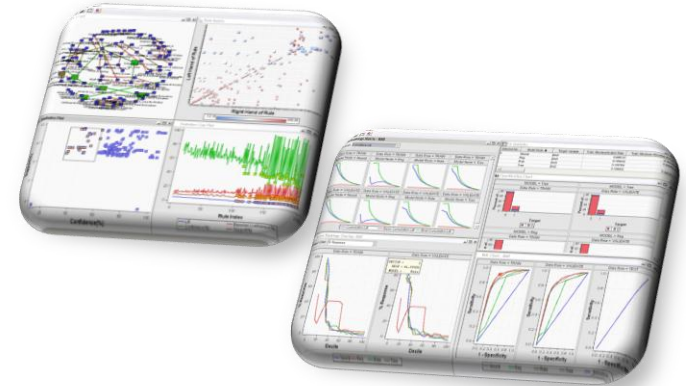
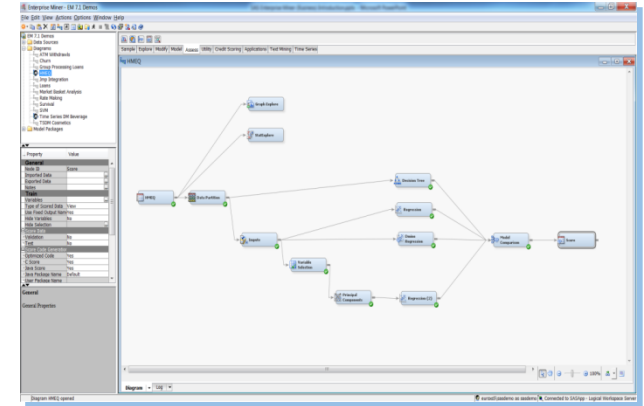
## DATA MINER / STATISTICIAN



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



# Model Development Process

## S<sub>ample</sub>

- Input Data
- File Import
- Sample
- Data Partition
- Merge
- Filter
- Append

## E<sub>xplore</sub>

- Association
- Cluster
- Variable Selection
- Market Basket
- StatExplore
- Variable Clustering
- MultiPlot
- Path Analysis

## M<sub>odify</sub>

- DMDB
- SOM/Kohonen
- Graph Explore

- Transform Variables
- Impute
- Replacement
- Interactive Binning
- Rules Builder
- Drop
- Principal Components

## M<sub>odel</sub>

- Decision Tree
- AutoNeural
- Dmine Regression
- DMNeural
- Ensemble
- Gradient Boosting
- LARS
- MBR

- Neural Network
- SVM
- Partial Least Squares
- Regression
- Rule Induction
- TwoStage
- Model Import

## A<sub>ssess</sub>

- Model Comparison
- Score
- Segment Profile
- Decisions
- Cutoff

# 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

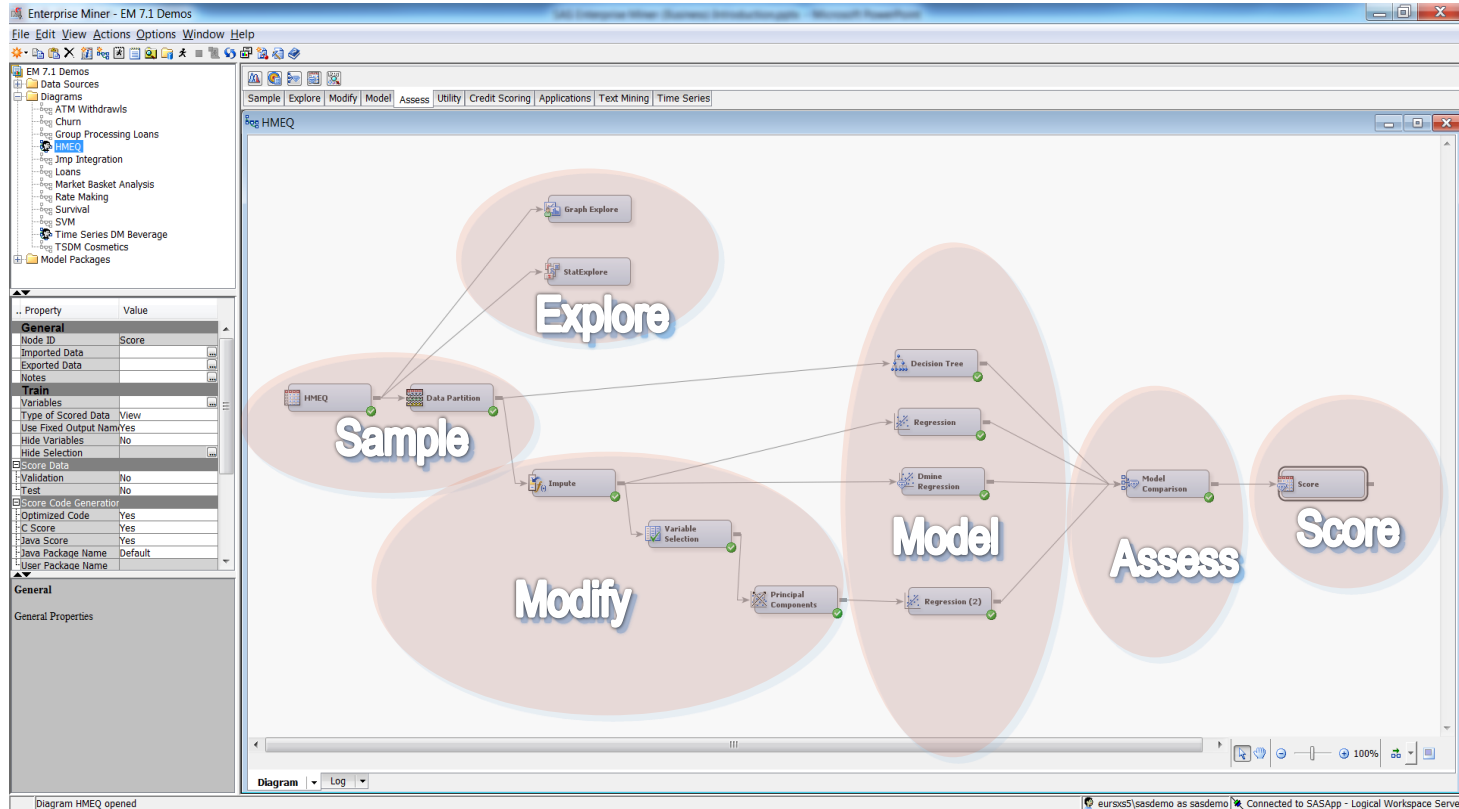
## HPDM

- 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

# SEMMA in Action – Repeatable Process





# Data Visualisation And Sampling



THE  
POWER  
TO KNOW.

# Visualisation

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



# SEMMA Process

## S<sub>ample</sub>



Input Data



File Import



Sample



Data Partition



Merge



Filter



Append



Time Series

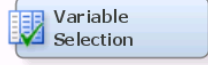
## E<sub>xplore</sub>



Association



Cluster



Variable Selection



Market Basket



StatExplore



Variable Clustering



MultiPlot



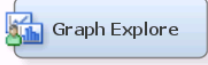
Path Analysis



DMDB

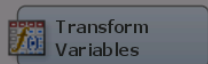


SOM/Kohonen

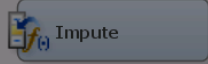


Graph Explore

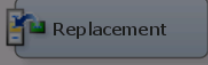
## M<sub>odify</sub>



Transform Variables



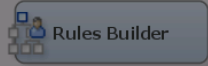
Impute



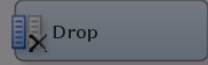
Replacement



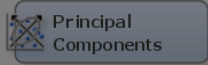
Interactive Binning



Rules Builder



Drop



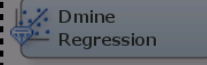
Principal Components



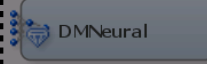
Decision Tree



AutoNeural



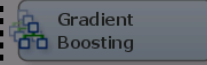
Dmine Regression



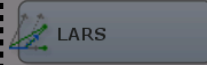
DMNeural



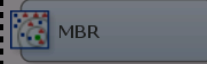
Ensemble



Gradient Boosting



LARS



MBR

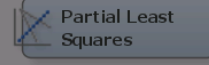
## M<sub>odel</sub>



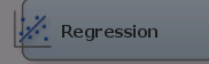
Neural Network



SVM



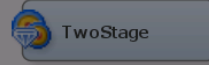
Partial Least Squares



Regression



Rule Induction

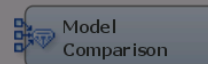


TwoStage

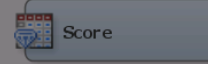


Model Import

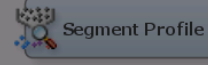
## A<sub>ssess</sub>



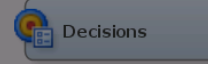
Model Comparison



Score



Segment Profile



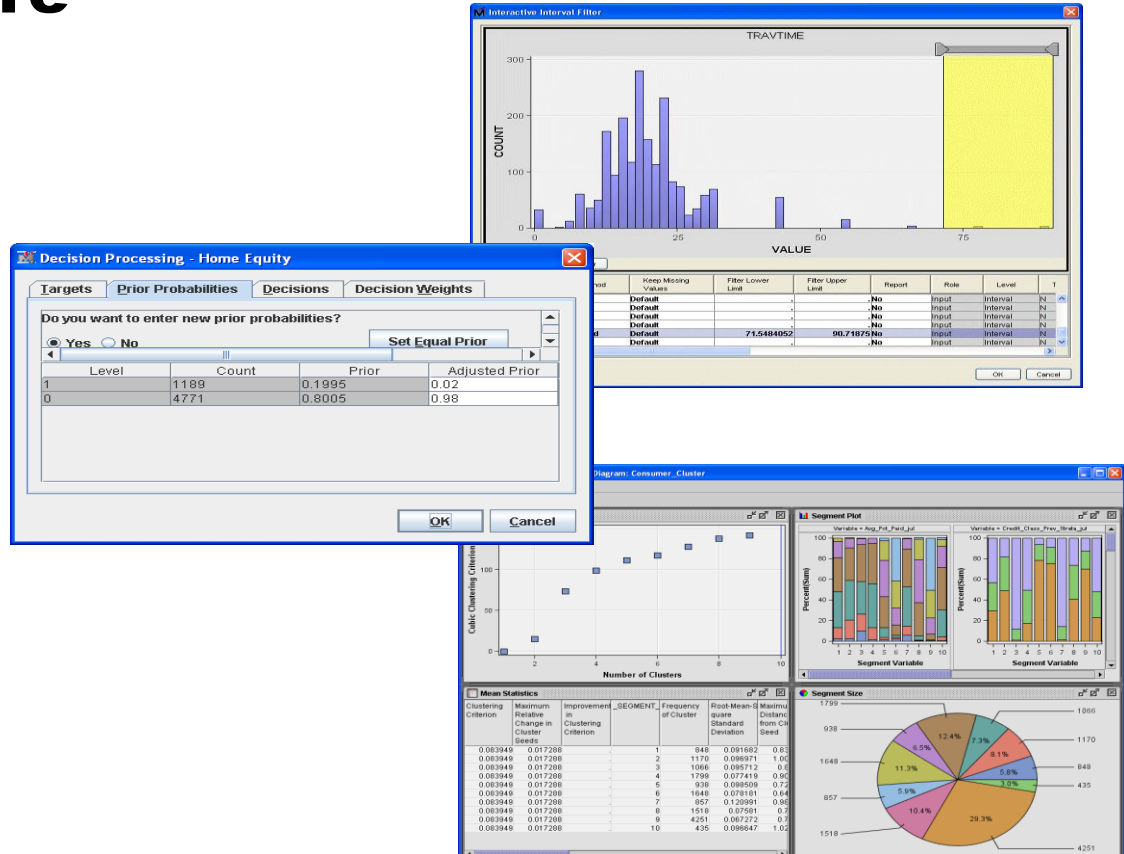
Decisions



Cutoff

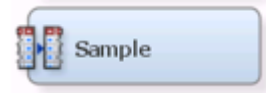
# Sample and Explore

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



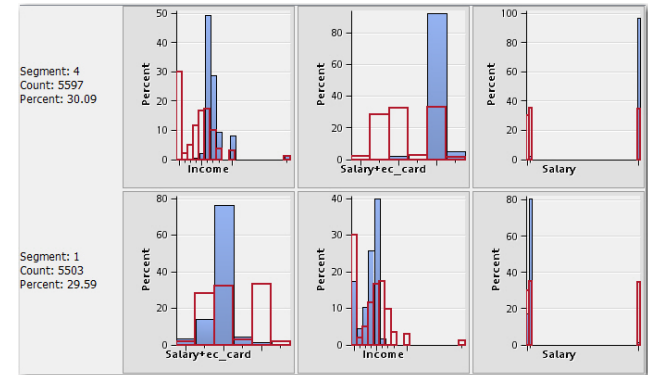
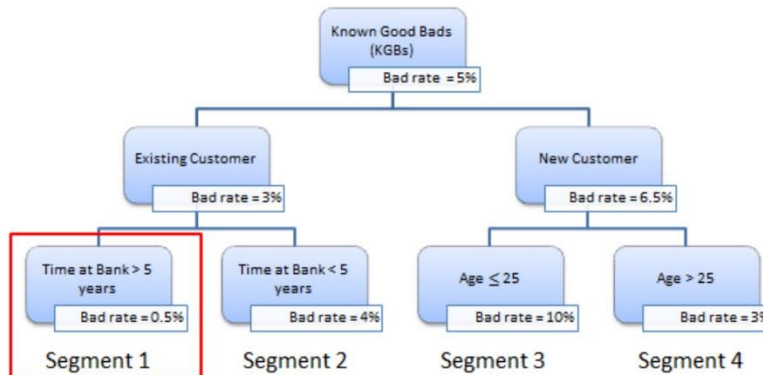
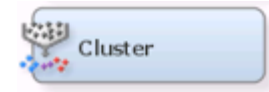
# Sampling

- 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



# Sampling

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



# Variable Selection



THE  
POWER  
TO KNOW.

# Variable Selection Routines

- Variable Selection
  - **Variable Selection Node**
    - Relationship of independent variables to dependent target
    - R-Square of 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





# Interactive Grouping Node Example Results

- Automatic and interactive variable grouping
- Computes Weights of Evidence



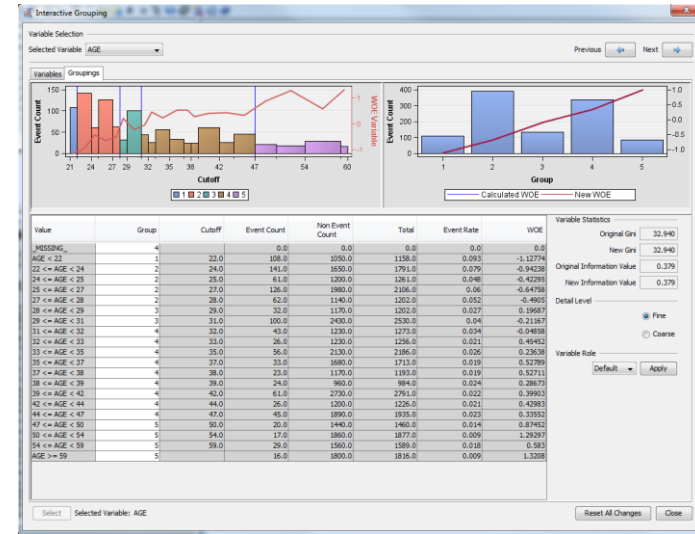
Interactive Grouping

Variable Selection  
Selected variable: AGE

Previous Next

| Variable | Label                 | Pre-Defined Grouping | Level    | Calculated Role | New Role | Original Gini | Original Information Value | Gini Statistic | Information Value |
|----------|-----------------------|----------------------|----------|-----------------|----------|---------------|----------------------------|----------------|-------------------|
| AGE      | Age                   |                      | INTERVAL | Input           | Default  | 32.94         | 0.376                      | 32.94          | 0.376             |
| INC1     | Salary rec_card       |                      | INTERVAL | Input           | Default  | 25.166        | 0.225                      | 25.166         | 0.225             |
| INC0B1   | Time at Job           |                      | INTERVAL | Input           | Default  | 23.129        | 0.209                      | 23.129         | 0.209             |
| INCOME   | Income                |                      | INTERVAL | Input           | Default  | 24.631        | 0.207                      | 24.631         | 0.207             |
| INC      | Salary                |                      | INTERVAL | Input           | Default  | 23.909        | 0.204                      | 23.909         | 0.204             |
| STATUS   | Status                |                      | NOMINAL  | Input           | Default  | 22.487        | 0.203                      | 22.487         | 0.203             |
| CARDS    | Credit Cards          |                      | NOMINAL  | Input           | Default  | 13.41         | 0.155                      | 13.41          | 0.155             |
| EC_CARD  | EC_card holders       |                      | INTERVAL | Input           | Default  | 15.697        | 0.133                      | 15.697         | 0.133             |
| PERF_H   | Num in household      |                      | INTERVAL | Input           | Default  | 18.434        | 0.132                      | 18.434         | 0.132             |
| TEL      | Telephone             |                      | INTERVAL | Rejected        | Default  | 9.164         | 0.088                      | 9.164          | 0.088             |
| PROF     | Profession            |                      | NOMINAL  | Rejected        | Default  | 9.659         | 0.056                      | 9.659          | 0.056             |
| CAR      | Type of vehicle       |                      | NOMINAL  | Rejected        | Default  | 9.455         | 0.051                      | 9.455          | 0.051             |
| IMACD    | Time at Address       |                      | INTERVAL | Rejected        | Default  | 10.933        | 0.044                      | 10.933         | 0.044             |
| NRBLDAN  | Num Mynbk Loans       |                      | INTERVAL | Rejected        | Default  | 9.026         | 0.041                      | 9.026          | 0.041             |
| CHILDREN | Num of Children       |                      | INTERVAL | Rejected        | Default  | 18.46         | 0.04                       | 18.46          | 0.04              |
| REGN     | Region                |                      | INTERVAL | Rejected        | Default  | 8.503         | 0.027                      | 8.503          | 0.027             |
| CASH     | Requested cash        |                      | INTERVAL | Rejected        | Default  | 7.922         | 0.025                      | 7.922          | 0.025             |
| PRODUCT  | Type of Business      |                      | NOMINAL  | Rejected        | Default  | 7.471         | 0.022                      | 7.471          | 0.022             |
| FINL0AN  | Num frshd Loans       |                      | INTERVAL | Rejected        | Default  | 6.434         | 0.017                      | 6.434          | 0.017             |
| LOANS    | Num of running lo...  |                      | INTERVAL | Rejected        | Default  | 5.35          | 0.015                      | 5.35           | 0.015             |
| COV      | Large region          |                      | INTERVAL | Rejected        | Default  | 5.025         | 0.013                      | 5.025          | 0.013             |
| BUREAU   | Credit Bureau Res...  |                      | INTERVAL | Rejected        | Default  | 4.351         | 0.009                      | 4.351          | 0.009             |
| NAT      | Nationality           |                      | NOMINAL  | Rejected        | Default  | 2.302         | 0.004                      | 2.302          | 0.004             |
| TITLE    | Title                 |                      | NOMINAL  | Rejected        | Default  | 2.028         | 0.004                      | 2.028          | 0.004             |
| RESID    | Residence Type        |                      | NOMINAL  | Rejected        | Default  | 0             | 0                          | 0              | 0                 |
| LOCATION | Location of Credit... |                      | INTERVAL | Rejected        | Default  | 0             | 0                          | 0              | 0                 |

Select Selected variable: AGE Reset All Changes Close



Auto-updating IV and Gini

Fine/Coarse Detail

# Missing Value Imputation



THE  
POWER  
TO KNOW.

# Impute

- Missing Value Imputation

- **Impute** Node

- Complete case required for models such as Regression
    - Multiple imputation techniques e.g. Tree, Distribution, Mean, Mode



| <b>Class (categorical) variables</b> | <b>Interval (numeric) variables</b> |
|--------------------------------------|-------------------------------------|
| <i>Input/Target</i>                  | <i>Input/Target</i>                 |
| 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                    |

# Outlier Detection



THE  
POWER  
TO KNOW.

# Detect and Treat

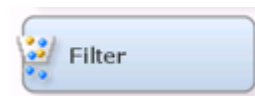
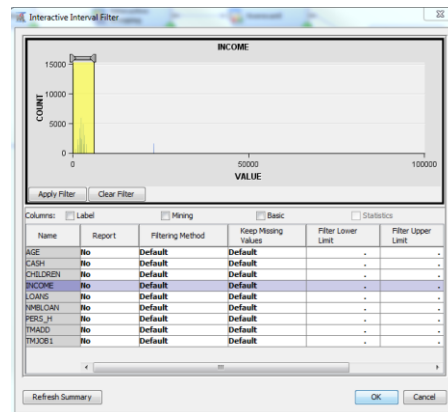
- Outlier Detection

- Filter Node

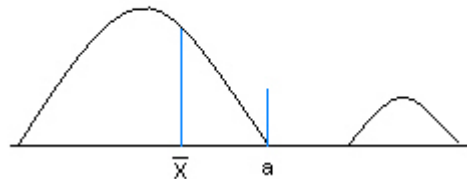
- Automated and Interactive filtering
    - Identify and exclude extreme outliers

- 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.



# Summary



THE  
POWER  
TO KNOW.

# 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)



SAS® FORUM  
UNITED KINGDOM 2015

## Questions and Answers

[Iain.Brown@sas.com](mailto:Iain.Brown@sas.com)



THE  
POWER  
TO KNOW.

[www.SAS.com](http://www.SAS.com)