



Overview of k-Means

Dr. Goutam Chakraborty



Outline

- Differences between hierarchical and nonhierarchical clustering methods.
- Advantages and disadvantages of both types of methods.
- Mechanics of k-Means clustering



Hierarchical versus Nonhierarchical Clustering Methods

Hierarchical

- Involves a *tree-like* construction process where clusters at any level of the tree are a combination of clusters below that level.
- After an observation has joined another observation in a step, successive steps keep them together.

Nonhierarchical

- Assigns objects into *prespecified* number of clusters using a distance/similarity metric. No tree-like structure exists.
- Assignment of object to cluster is *not fixed* through the iteration process.
- Iterate to minimize or maximize a criterion such as separation between clusters or, within-cluster similarity.



Advantages and Disadvantages of Hierarchical Methods

- Advantages include:
 - **Ability to capture non-spherical clusters.**
 - No order effect, that is, the ordering of observations has no impact on cluster solutions.
 - No need to make an initial guess at number of clusters in the data.
- Disadvantages include:
 - **Does not scale well for large/complex data.**
 - Early combinations (even if it is a mistake) persist throughout the process.
 - Susceptible to outliers (depends on method).
 - In many segmentation studies, there is little theoretical reason to expect a hierarchical structure.
 - Too many choices of methods.



Advantages and Disadvantages of Nonhierarchical Methods

- Advantages include:
 - **Scale up well with large/complex data.**
 - Generally easy to understand.
- Disadvantages include:
 - **Makes assumptions about shape of clusters.**
 - Number of clusters need to be specified in advance.
 - Results might be influenced by the choice of initial seeds or, order of reading of seeds.
 - Susceptible to outliers.



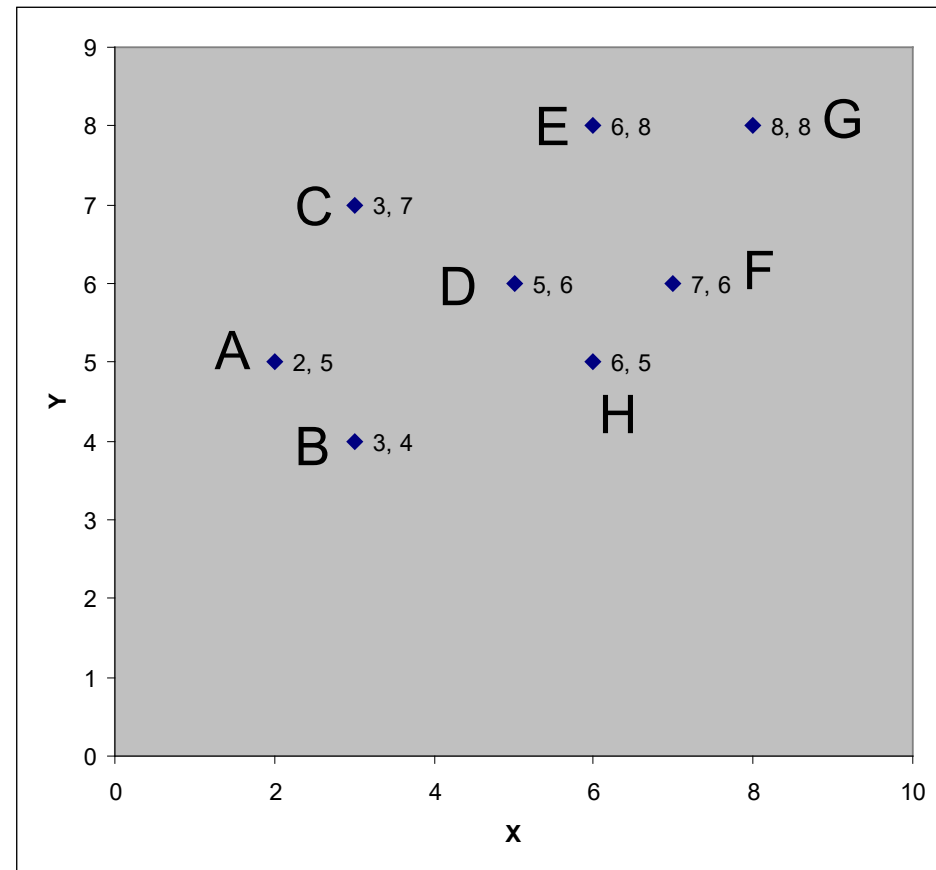
k -Means Procedure (Mechanics)

1. Select k cluster centers.
2. Assign cases to closest center.
3. Update cluster centers.
4. Reassign cases.
5. Repeat steps 3 and 4 until convergence.

A Numerical Example of k -Means Clustering

- Data from eight subjects (A, B, C, D, E, F, G, H) on two variables, X and Y.

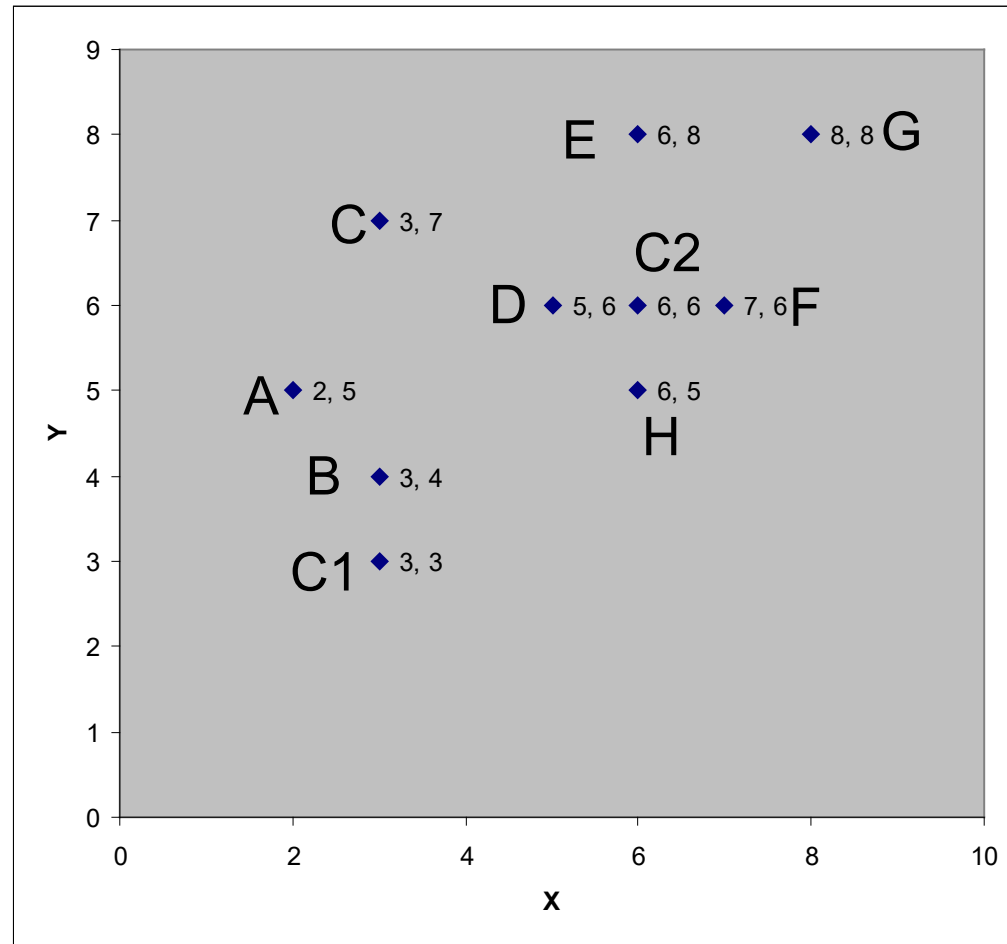
ID	X	Y
A	2	5
B	3	4
C	3	7
D	5	6
E	6	8
F	7	6
G	8	8
H	6	5



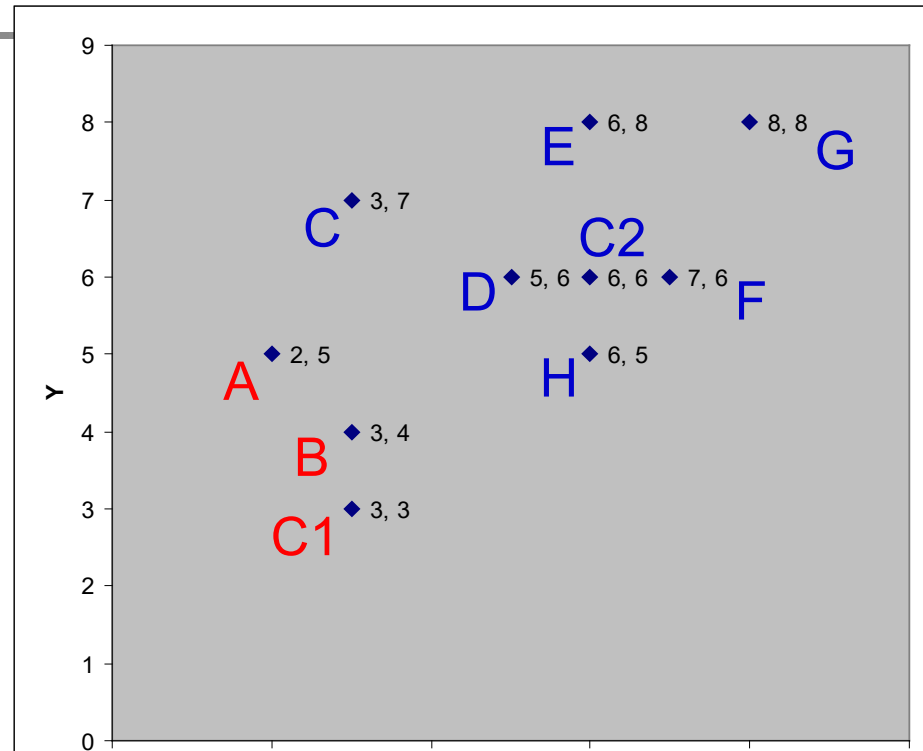
A Numerical Example of k -Means Clustering

- Center 1, C1(3,3) and Center 2, C2 (6,6) chosen at random.

ID	X	Y
A	2	5
B	3	4
C	3	7
D	5	6
E	6	8
F	7	6
G	8	8
H	6	5
Center1	3	3
Center2	6	6



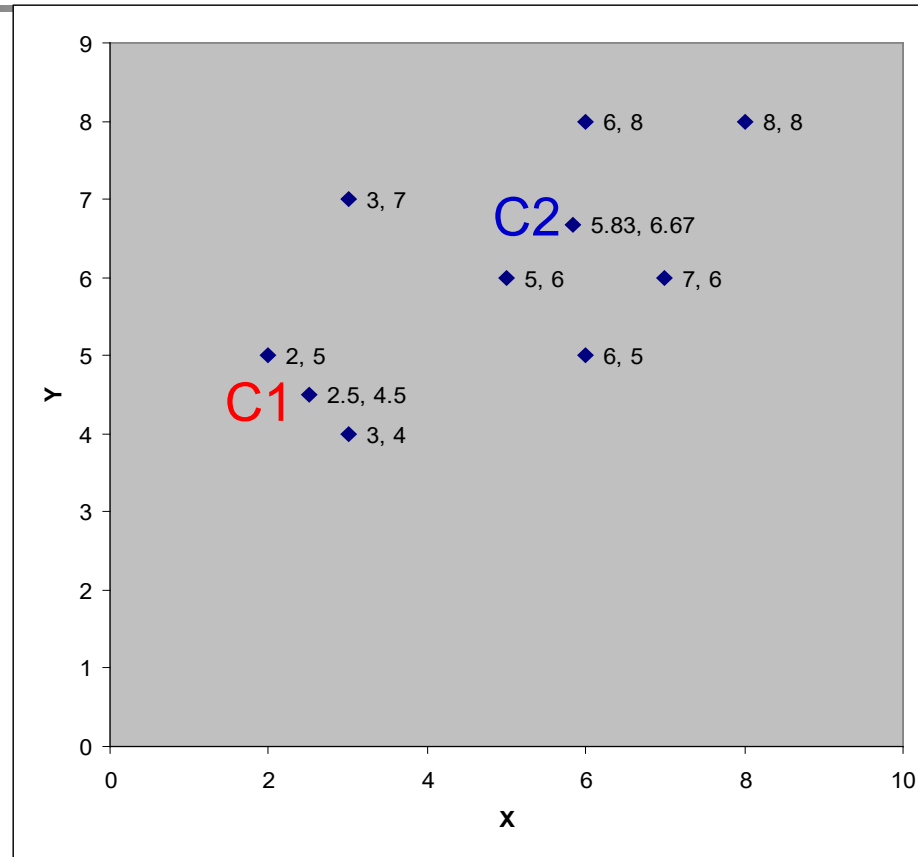
k-Means Numerical Example (Iteration 1)



ID	A	B	C	D	E	F	G	H	Center1	Center2
A	0.00000
B	1.41421	0.00000
C	2.23607	3.00000	0.00000
D	3.16228	2.82843	2.23607	0.00000
E	5.00000	5.00000	3.16228	2.23607	0.00000
F	5.09902	4.47214	4.12311	2.00000	2.23607	0.00000
G	6.70820	6.40312	5.09902	3.60555	2.00000	2.23607	0.00000	.	.	.
H	4.00000	3.16228	3.60555	1.41421	3.00000	1.41421	3.60555	0.00000	.	.
Center1	2.23607	1.00000	4.00000	3.60555	5.83095	5.00000	7.07107	3.60555	0.00000	.
Center2	4.12311	3.60555	3.16228	1.00000	2.00000	1.00000	2.82843	1.00000	4.24264	0

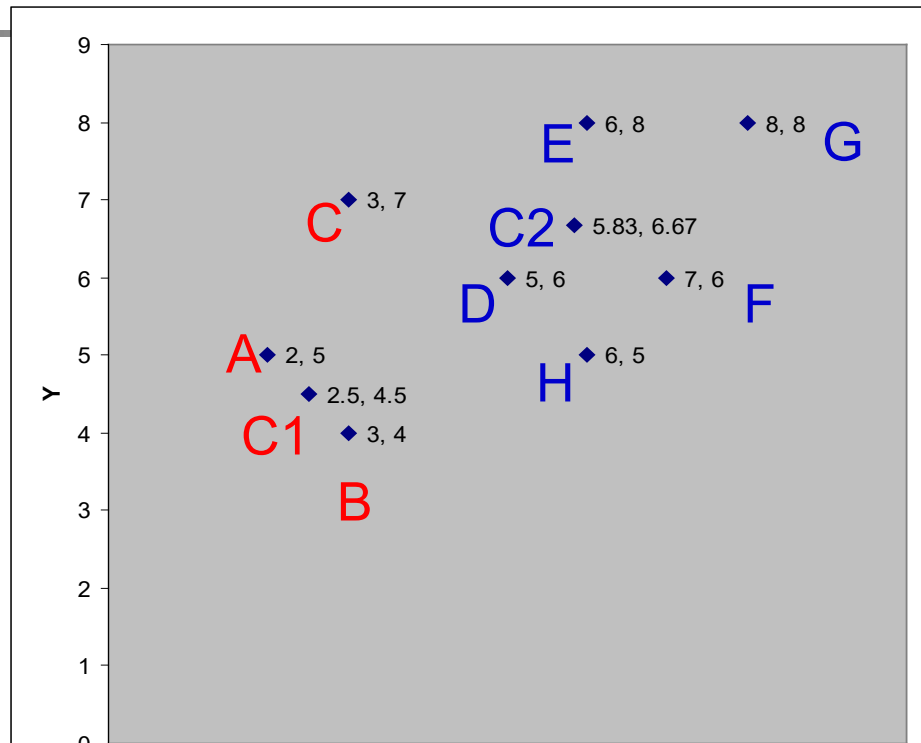
continued...

k -Means Numerical Example (Iteration 1)



- Update cluster centers (C1 and C2).
- New centers are C1(2.5,4.5) and C2(5.83,6.67).

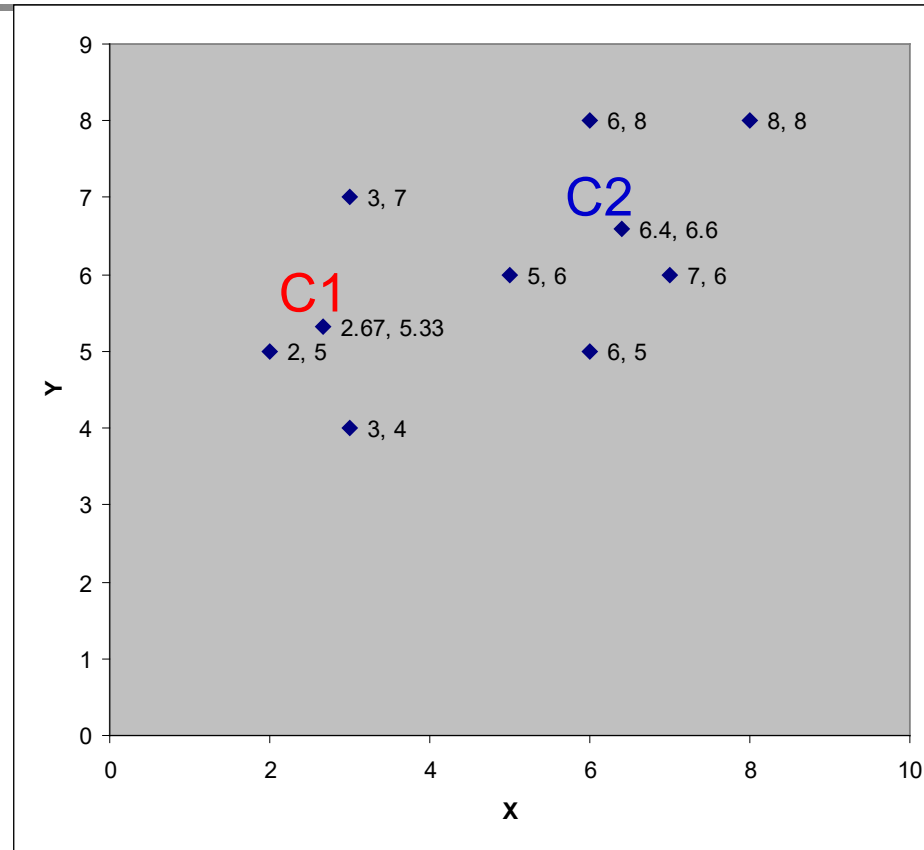
k-Means Numerical Example (Iteration 2)



ID	A	B	C	D	E	F	G	H	Center1	Center2
A	0.00000
B	1.41421	0.00000
C	2.23607	3.00000	0.00000
D	3.16228	2.82843	2.23607	0.00000
E	5.00000	5.00000	3.16228	2.23607	0.00000
F	5.09902	4.47214	4.12311	2.00000	2.23607	0.00000
G	6.70820	6.40312	5.09902	3.60555	2.00000	2.23607	0.00000	.	.	.
H	4.00000	3.16228	3.60555	1.41421	3.00000	1.41421	3.60555	0.00000	.	.
Center1	0.70711	0.70711	2.54951	2.91548	4.94975	4.74342	6.51920	3.53553	0.00000	.
Center2	4.17825	3.89073	2.84918	1.06668	1.34082	1.34826	2.54515	1.67863	3.97464	0

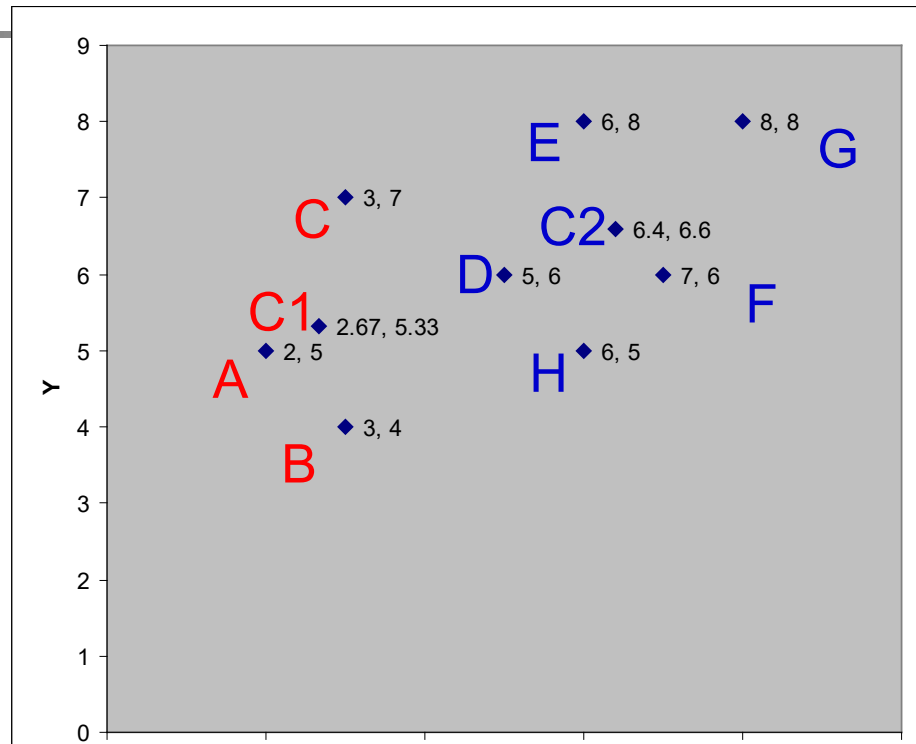
continued...

k -Means Numerical Example (Iteration 2)



- Update cluster centers (C1 and C2).
- New centers are C1(2.67,5.33) and C2(6.4,6.6)

k -Means Numerical Example (Iteration 3)



ID	A	B	C	D	E	F	G	H	Center1	Center2
A	0.00000
B	1.41421	0.00000
C	2.23607	3.00000	0.00000
D	3.16228	2.82843	2.23607	0.00000
E	5.00000	5.00000	3.16228	2.23607	0.00000
F	5.09902	4.47214	4.12311	2.00000	2.23607	0.00000
G	6.70820	6.40312	5.09902	3.60555	2.00000	2.23607	0.00000	.	.	.
H	4.00000	3.16228	3.60555	1.41421	3.00000	1.41421	3.60555	0.00000	.	.
Center1	0.74686	1.37033	1.70229	2.42442	4.26823	4.38153	5.96136	3.34631	0.00000	.
Center2	4.68188	4.28019	3.42345	1.52315	1.45602	0.84853	2.12603	1.64924	3.94028	0



SAS EM Interface Tour

Dr. Goutam Chakraborty



Outline

- Describe the basic navigation of SAS Enterprise Miner.
- Creating project, library (for data access) and diagram (for analysis) in SAS EM

SAS Enterprise Miner

The screenshot displays the SAS Enterprise Miner interface with a project named "My Project". The left sidebar shows a tree view with "Data Sources" (ORGANICS, PVA97NK), "Diagrams" (Organics, Predictive Analysis), and "Model Packages". The "Predictive Analysis" diagram is selected, showing a workflow: PVA97NK → Replacement → Data Partition → Impute → Regression. A "Decision Tree" node is also connected to the "Data Partition" node. The bottom left pane shows the "Property" and "Value" table for the selected diagram.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

Diagram Predictive Analysis opened

sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)

SAS Enterprise Miner – Interface Tour

Menu bar and shortcut buttons

The screenshot displays the SAS Enterprise Miner interface. The menu bar at the top includes File, Edit, View, Actions, Options, Window, and Help. Below the menu bar is a row of shortcut buttons. The main workspace shows a predictive analysis workflow diagram with the following steps: PVA97NK (Data Source) → Replacement → Data Partition → Impute → Regression. A branch from Data Partition leads to Decision Tree. The left pane shows a project tree with Data Sources (ORGANICS, PVA97NK), Diagrams (Organics, Predictive Analysis), and Model Packages. The bottom pane shows the Properties window for the selected node, with tabs for General, Train, and Class Targets. The General tab is active, showing properties like Node ID (Reg), Imported Data, Exported Data, Notes, Variables, Equation (Main Effects: Yes, Two-Factor Interactions: No, Polynomial Terms: No, Polynomial Degree: 2, User Terms: No), Term Editor, Class Targets (Regression Type: Logistic Regression), and General (Diagram Predictive Analysis opened). The status bar at the bottom shows the user sasdemo@SASBAP as sasdemo and the connection to SASApp - Logical Workspace Server (sasbap.demo.sas.com).

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
General	
Diagram Predictive Analysis opened	

SAS Enterprise Miner – Interface Tour

The screenshot displays the SAS Enterprise Miner interface. The **Project Panel** on the left, highlighted with a red box, shows a tree structure under 'My Project' with nodes for 'Data Sources' (ORGANICS, PVA97NK), 'Diagrams' (Organics, Predictive Analysis), and 'Model Packages'. The main workspace shows a workflow diagram for 'Predictive Analysis' with steps: PVA97NK → Replacement → Data Partition → Impute → Regression. A 'Decision Tree' node is also shown, connected to the 'Data Partition' step. The bottom-left pane shows the 'Property' window for the 'Predictive Analysis' diagram, with tabs for 'General', 'Train', and 'Class Targets'. The 'General' tab is active, showing properties like 'Node ID' (Reg), 'Imported Data', 'Exported Data', 'Notes', 'Variables', 'Equation' (Main Effects: Yes, Two-Factor Interactions: No, Polynomial Terms: No, Polynomial Degree: 2, User Terms: No), and 'Class Targets' (Regression Type: Logistic Regression). The status bar at the bottom indicates 'Diagram Predictive Analysis opened' and 'Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)'.

Project Panel

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

Diagram Predictive Analysis opened

sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)

SAS Enterprise Miner – Interface Tour

The screenshot displays the SAS Enterprise Miner interface. On the left, a tree view shows the project structure: My Project, Data Sources (ORGANICS, PVA97NK), Diagrams (Organics, Predictive Analysis), and Model Packages. The main workspace shows a workflow diagram for Predictive Analysis with nodes: PVA97NK, Replacement, Data Partition, Impute, Regression, and Decision Tree. A red circle highlights the Properties Panel on the left, which contains a table of properties for the selected node.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interaction	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit

Properties Panel

Diagram Predictive Analysis opened

sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)

SAS Enterprise Miner – Interface Tour

The screenshot displays the SAS Enterprise Miner software interface. The main window is titled "Enterprise Miner - My Project". The left sidebar shows a project tree with "My Project" expanded, containing "Data Sources" (ORGANICS, PVA97NK), "Diagrams" (Organics, Predictive Analysis), and "Model Packages". The "Predictive Analysis" diagram is selected, showing a workflow: PVA97NK → Replacement → Data Partition → Impute → Regression. A "Decision Tree" node is also connected to the "Data Partition" node. The bottom-left panel shows the "General" properties for the selected node, with a red box highlighting the "General" tab and the "General Properties" section.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No

General
General Properties

Help Panel

Diagram Predictive Analysis opened

sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)

SAS Enterprise Miner – Interface Tour

The screenshot displays the SAS Enterprise Miner interface. The main window is titled "Enterprise Miner - My Project" and contains a menu bar (File, Edit, View, Actions, Options, Window, Help) and a toolbar. On the left, a project tree shows the hierarchy: My Project > Data Sources > PVA97NK, and My Project > Diagrams > Predictive Analysis. Below the tree is a property grid for the selected "Predictive Analysis" diagram.

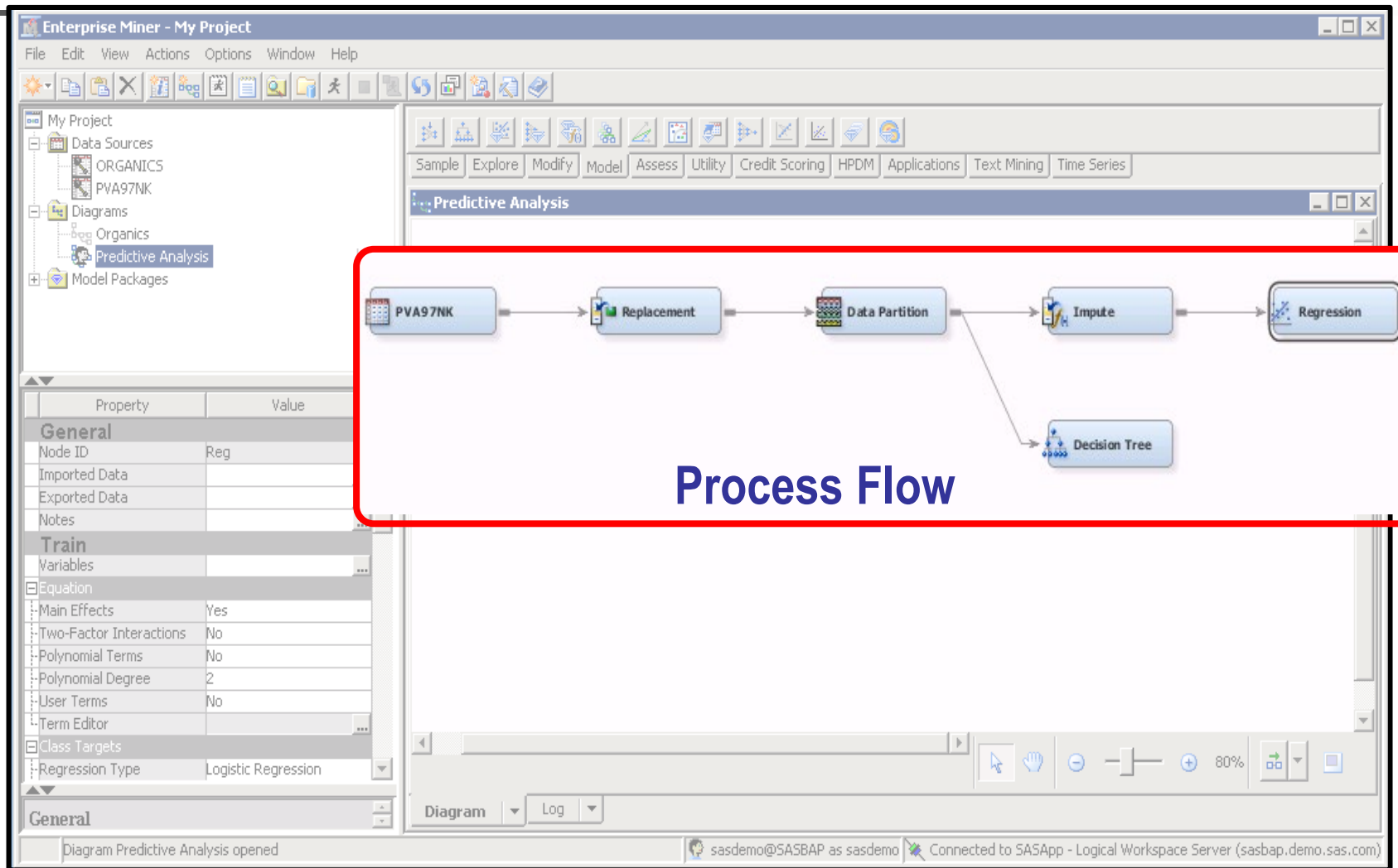
Property	Value
General	
Node ID	Reg
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression

The central "Predictive Analysis" diagram workspace is highlighted with a red border. It contains a flowchart with the following nodes and connections:

- PVA97NK** (Data Source) connects to **Replacement** (Transformation).
- Replacement** connects to **Data Partition** (Transformation).
- Data Partition** branches into two paths:
 - One path leads to **Impute** (Transformation), which then connects to **Regression** (Model).
 - Another path leads to **Decision Tree** (Model).

The text "Diagram Workspace" is overlaid in the center of this workspace. At the bottom of the interface, a status bar shows the user "sasdemo@SASBAP as sasdemo" and the connection "Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)".

SAS Enterprise Miner – Interface Tour



SAS Enterprise Miner – Interface Tour

The screenshot displays the SAS Enterprise Miner software interface. The main window is titled "Enterprise Miner - My Project". The left sidebar shows a project tree with "My Project" expanded, containing "Data Sources" (ORGANICS, PVA97NK), "Diagrams" (Organics, Predictive Analysis), and "Model Packages". The "Predictive Analysis" diagram is selected, showing a workflow: PVA97NK → Replacement → Data Partition → Impute → Regression. The "Data Partition" node is highlighted with a red box and labeled "Node". Below the workflow, a "Decision Tree" node is visible. The bottom left pane shows the "Property" table for the selected node.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

The bottom status bar shows "Diagram Predictive Analysis opened", the user "sasdemo@SASBAP as sasdemo", and the connection "Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)".

SAS Enterprise Miner – Interface Tour

The screenshot displays the SAS Enterprise Miner interface. The top menu bar includes File, Edit, View, Actions, Options, Window, and Help. Below the menu is a toolbar with various icons. A red box highlights the **SEMMA Tools Palette**, which contains the following tools: Sample, Explore, Modify, Model, Assess, Utility, Credit Scoring, HPDM, Applications, Text Mining, and Time Series.

The main workspace shows a workflow diagram with the following steps: PVA97NK → Replacement → Data Partition → Impute → Regression. A branch from Data Partition leads to Decision Tree.

On the left, the **My Project** tree shows the following structure:

- My Project
 - Data Sources
 - ORGANICS
 - PVA97NK
 - Diagrams
 - Organics
 - Predictive Analysis
 - Model Packages

Below the tree is the **Property** table:

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

At the bottom, the status bar shows: Diagram Predictive Analysis opened, sasdemo@SASBAP as sasdemo, and Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com).

SEMMA – Sample Tab

The screenshot displays the Enterprise Miner software interface. On the left, a project tree shows 'My Project' with sub-items: 'Data Sources' (ORGANICS, PVA97NK), 'Diagrams' (Organics, Predictive Analysis), and 'Model Packages'. Below this is a 'Property' table with columns 'Property' and 'Value'. The 'General' section includes 'Node ID' (Reg), 'Imported Data', 'Exported Data', and 'Notes'. The 'Train' section includes 'Variables'. The 'Equation' section includes 'Main Effects' (Yes), 'Two-Factor Interactions' (No), 'Polynomial Terms' (No), 'Polynomial Degree' (2), and 'User Terms' (No). The 'Class Targets' section includes 'Regression Type' (Logistic Regression). The main workspace shows a workflow diagram with nodes: 'PVA97NK', 'Replacement', 'Data Partition', 'Impute', 'Regression', and 'Decision Tree'. A red box highlights the 'Sample' tab in the top toolbar, which also includes 'Explore', 'Modify', 'Model', 'Assess', 'Utility', 'Credit Scoring', 'HPDM', 'Applications', 'Text Mining', and 'Time Series'. The bottom status bar shows 'Diagram Predictive Analysis opened' and 'sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)'.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

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

SEMMA – Explore Tab

The screenshot shows the SAS Enterprise Miner interface with the 'Explore' tab selected. The top toolbar contains the following tabs: Sample, Explore, Modify, Model, Assess, Utility, Credit Scoring, HPDM, Applications, Text Mining, and Time Series. The central workspace displays a flow diagram for 'PVA97NK' with nodes: Replacement, Data Partition, Impute, Regression, and Decision Tree. The bottom left panel shows the 'General' properties for the selected node, including Node ID, Imported Data, Exported Data, Notes, Variables, Equation, and Class Targets.

- Association
- Cluster
- DMDB
- Graph Explore
- Link Analysis
- Market Basket
- Multiplot
- Path Analysis
- SOM/Kohonen
- StatExplore
- Variable Clustering
- Variable Selection

SEMMA – Modify Tab

The screenshot displays the SAS Enterprise Miner interface. The 'Modify' tab is highlighted in the top toolbar, which is enclosed in a red rectangle. The main workspace shows a process flow diagram with nodes: PVA97NK, Replacement, Data Partition, Impute, Regression, and Decision Tree. The left sidebar shows the project structure with 'Predictive Analysis' selected. The bottom-left pane shows the 'Property' and 'Value' table for the selected node.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

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

Diagram Predictive Analysis opened

sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)

SEMMA – Model Tab

The screenshot displays the SAS Enterprise Miner interface. The top menu bar includes File, Edit, View, Actions, Options, Window, and Help. Below the menu is a toolbar with various icons. A red rectangle highlights the 'Model' tab, which is part of the SEMMA workflow. The 'Model' tab contains several sub-tabs: Sample, Explore, Modify, Model, Assess, Utility, Credit Scoring, HPDM, Applications, Text Mining, and Time Series. The main workspace shows a workflow diagram with nodes: PVA97NK, Replacement, Data Partition, Impute, Regression, and Decision Tree. The left sidebar shows a tree view of the project structure, including Data Sources, Diagrams, and Models. The bottom status bar indicates the user is logged in as sasdmo@SASBAP and is connected to the SASApp - Logical Workspace Server.

Model Tab

- AutoNeural
- Decision Tree
- DMine Regression
- DMNeural
- Ensemble
- Gradient Boosting
- Least Angle Regression
- MBR
- Model Import
- Neural Network
- Partial Least Squares
- Regression
- Rule Induction
- Two Stage

SEMMA – Assess Tab

Enterprise Miner - My Project

File Edit View Actions Options Window Help

My Project

- Data Sources
 - ORGANICS
 - PVA97NK
- Diagrams
 - Organics
 - Predictive Analysis
- Model Packages

Sample Explore Modify Model **Assess** Utility Credit Scoring HPDM Applications Text Mining Time Series

PVA97NK → Replacement → Data Partition → Impute → Regression

Decision Tree

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

Diagram Predictive Analysis opened

sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)

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

Beyond SEMMA – Utility Tab

The screenshot displays the Enterprise Miner software interface. The top menu bar includes File, Edit, View, Actions, Options, Window, and Help. Below the menu is a toolbar with various icons. A red rectangle highlights a secondary toolbar containing the following tabs: Sample, Explore, Modify, Model, Assess, Utility, Credit Scoring, HPDM, Applications, Text Mining, and Time Series. The left sidebar shows a project tree with 'My Project' expanded, containing 'Data Sources' (ORGANICS, PVA97NK), 'Diagrams' (Organics, Predictive Analysis), and 'Model Packages'. The bottom-left pane shows a property table for the 'Predictive Analysis' diagram.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

The main workspace displays a workflow diagram with nodes: PVA97NK, Replacement, Impute, and Regression. A 'Decision Tree' node is also visible. The bottom status bar shows 'Diagram Predictive Analysis opened' and 'sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)'.

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

Beyond SEMMA – HPDM Tab

The screenshot displays the SAS Enterprise Miner software interface. The main window is titled "Enterprise Miner - My Project". The toolbar at the top contains various icons and tabs. The "HPDM" tab is highlighted with a red rectangle. The workflow diagram in the center shows a sequence of steps: "PVA97NK" (Data Source) → "Replacement" (Transformation) → "Data Partition" (Transformation) → "Impute" (Transformation) → "Regression" (Model). A branch from "Data Partition" also leads to "Decision Tree" (Model). The left sidebar shows a tree view of the project structure, including "Data Sources", "Transformations", and "Models". The bottom status bar indicates "Diagram Predictive Analysis opened" and "Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)".

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

Beyond SEMMA – Applications Tab

The screenshot displays the SAS Enterprise Miner software interface. The title bar reads "Enterprise Miner - My Project". The menu bar includes File, Edit, View, Actions, Options, Window, and Help. The toolbar contains various icons for project management and analysis. The left pane shows a project tree with "My Project" expanded, containing "Data Sources" (ORGANICS, PVA97NK), "Diagrams" (Organics, Predictive Analysis), and "Model Packages". The bottom-left pane shows the "Property" table for the "Predictive Analysis" diagram.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

The main workspace shows a flow diagram with nodes: PVA97NK, Replacement, Data Partition, Input, Regression, and Decision Tree. A red box highlights the "Applications" tab in the top toolbar, which is part of a group including Sample, Explore, Modify, Model, Assess, Utility, Credit Scoring, HPDM, Applications, Text Mining, and Time Series. The "Applications" tab is currently selected. The bottom status bar shows "Diagram Predictive Analysis opened" and "sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)".

■ Incremental Response
■ Survival

Beyond SEMMA – Time Series Tab

The screenshot displays the Enterprise Miner - My Project window. The top menu bar includes File, Edit, View, Actions, Options, Window, and Help. Below the menu bar is a toolbar with various icons. A red rectangle highlights the 'Time Series' tab in the top menu bar, which is part of a larger set of tabs including Sample, Explore, Modify, Model, Assess, Utility, Credit Scoring, HPDM, Applications, Text Mining, and Time Series. The main workspace shows a diagram with nodes: PVA97NK, Replacement, Data Partition, and a Decision Tree. The left sidebar shows a tree view with 'My Project' expanded, containing 'Data Sources' (ORGANICS, PVA97NK), 'Diagrams' (Organics, Predictive Analysis), and 'Model Packages'. The bottom left pane shows the 'Property' and 'Value' table for the 'Predictive Analysis' diagram.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

Diagram Predictive Analysis opened

sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)

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

Credit Scoring Tab (Optional)

The screenshot displays the SAS Enterprise Miner interface. The main window is titled "Enterprise Miner - My Project". The left sidebar shows a project tree with "Data Sources" (ORGANICS, PVA97NK), "Diagrams" (Organics, Predictive Analysis), and "Model Packages". The bottom-left pane shows the "Property" table for the "Predictive Analysis" diagram.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

The main workspace shows a flow diagram with nodes: "PVA97NK" → "Replacement" → "Data Mining" → "Regression". The "Credit Scoring" tab is highlighted in the top toolbar. A red box highlights the toolbar area. A list of features is overlaid on the right side of the workspace:

- Credit Exchange
- Interactive Grouping
- Reject Inference
- Scorecard

The bottom status bar shows "Diagram Predictive Analysis opened" and "sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)".

Text Mining Tab (Optional)

The screenshot displays the SAS Enterprise Miner interface. The 'Text Mining' tab is highlighted in the top menu bar, which also includes 'Sample', 'Explore', 'Modify', 'Model', 'Assess', 'Utility', 'Credit Scoring', 'HPDM', 'Applications', and 'Time Series'. The main workspace shows a workflow diagram with nodes: 'PVA97NK', 'Replacement', 'Data Partition', 'Input', 'Regression', and 'Decision Tree'. A list of text mining tasks is overlaid on the right side of the diagram.

- Text Cluster
- Text Filter
- Text Import
- Text Parsing
- Text Profile
- Text Rule Builder
- Text Topic

The left sidebar shows a project tree with 'My Project' containing 'Data Sources' (ORGANICS, PVA97NK), 'Diagrams' (Organics, Predictive Analysis), and 'Model Packages'. The bottom left pane shows the 'Property' and 'Value' table for the 'Predictive Analysis' diagram.

Property	Value
General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression

The bottom status bar indicates 'Diagram Predictive Analysis opened' and 'sasdemo@SASBAP as sasdemo Connected to SASApp - Logical Workspace Server (sasbap.demo.sas.com)'.

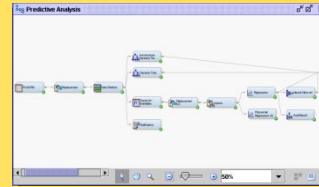
Analysis Element Organization



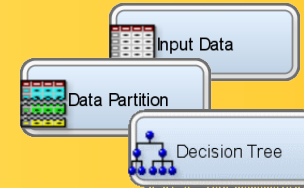
Projects



**Libraries
and
Diagrams**



**Process
Flows**



Nodes

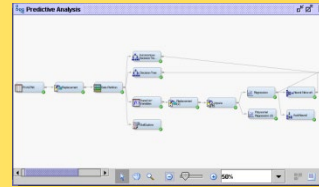
Analysis Element Organization



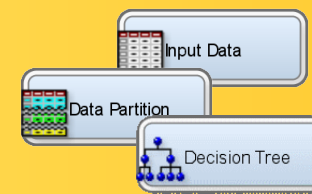
Projects



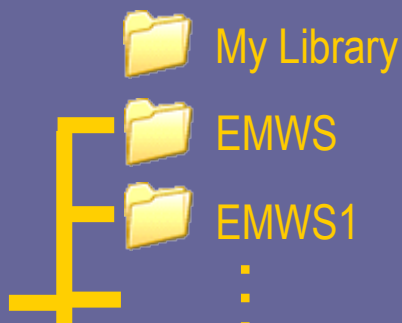
**Libraries
and
Diagrams**



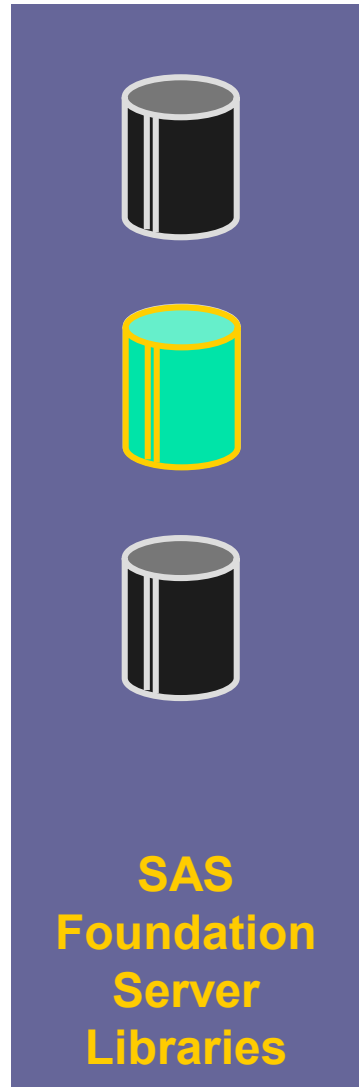
**Process
Flows**



Nodes



Defining a Data Source



- Select table
- Define variable roles
- Define measurement levels
- Define table role



SAS EM Demo

- Create a new project (save where you can access it and have enough space)
- Create a library (I will name it as *course*)
- Open the data set *kmeans_demotr* through the library
 - It is possible to set *specific* roles/levels of all variables in the data step creation process.
 - **Best practice:** use default selections for roles/levels of variables in data step creation process. Then, use a **Metadata** node to set *specific* roles/levels of all variables.
- Understand nature of your data via multiple methods:
 - Under data sources, right-click data table and select explore
 - In the diagram, right-click data table and select edit variables > then select variables and explore
 - Using nodes in **Explore tab** such as DMDB, Graph Explore, Multiplot and StatExplore



SAS EM Demo

Dr. Goutam Chakraborty



Outline

- Creating project, library (for data access) and diagram (for analysis) in SAS EM



SAS EM Demo

- Create a new project (save where you can access it and have enough space)
- Create a library (I will name it as *course*)
- Open the data set *kmeans_demotr* through the library
 - It is possible to set *specific* roles/levels of all variables in the data step creation process.
 - **Best practice**: use default selections for roles/levels of variables in data step creation process. Then, use a **Metadata** node to set *specific* roles/levels of all variables.
- Understand nature of your data via multiple methods:
 - Under data sources, right-click data table and select explore
 - In the diagram, right-click data table and select edit variables > then select variables and explore
 - Using nodes in **Explore tab** such as DMDB, Graph Explore, Multiplot and StatExplore



Transformations Before Clustering

Dr. Goutam Chakraborty



Outline

- Understand the variables and the business issues in the catalog company data set.
- Use SAS Enterprise Miner for checking distributions of variables (*bases*) and applying appropriate transformations.



A Few Cautions Before Beginning

- Most of the caveats mentioned in discussing hierarchical clustering also apply to k -means clustering. These include the following:
 - Selection of relevant clustering variables
 - Preprocessing of data to handle skewed distributions, outliers, missing values, and different measurement scales
 - Interpreting cluster profiles first using *bases* and then using descriptors



Business Problem and Data Description

- ABC is a supplier of identification products serving 90,000+ customers in the U.S.
- ABC wants to segment their customers based on their past and future expected transaction patterns with ABC, as well as selected firmographic variables.
 - ABC wants to consider between *2-10 segments*.
- ABC wants to profile and understand the segments using the *bases*.
- ABC also wants to profile and validate the segments using the *descriptors*.

SAS data set name: kmeans_demoTR

Base Variables

- **Lt_st_sales**: Total sales revenue from a customer.
- **Tele_rank**: ABC's internal estimate of ranking of customers based on future sales (smaller number is better).
- **Grow_dec**: ABC's internal estimate of which deciles customer falls in based on future growth potential.
- **RFM_group**: Seven categories (0-6) recency, frequency, and monetary grouping based on past year's transactions (higher number is better).
- **Hdcnt_last**: Number of employees in customer's location.
- **Industry**: Type of industry (based on two-digit SIC code) customer belongs in ten categories such as manufacturing, construction, and so on.

Descriptor and other managerially important variables

- **Lt_st_orders**: Total number of orders from a customer.
- **Divisions**: How many divisions within ABC a customer is buying from.
- **Acct_recency**: Time in months since last purchase.
- **Type_customer**: Four categories (platinum, gold, growable, and unspecified) of customers.
- **Reseller**: Whether the account is a reseller of ABC's products.
- **Zone**: Five categories of customer's primary location in the US (Western, Central, North, and NE, South, and SE, other).
- **Credit_risk**: ABC's internal estimate of customer's credit risk (five categories).



Plan of Analysis

1. Explore this data set using SAS Enterprise Miner. In particular, look at the distributions of *base* variables.
2. Use transformation as appropriate on *base* variables.
3. Run *k*-means using SAS Enterprise Miner.
4. Interpret results from *k*-means.



Why Do Base Variable Transformation before Clustering?

- To give *equal importance* to each variable in influencing cluster results
- To reduce *Skewness and Kurtosis* to a manageable number
- Ideally, we would like variable distribution to be close to Normal (if that's not possible, at least)



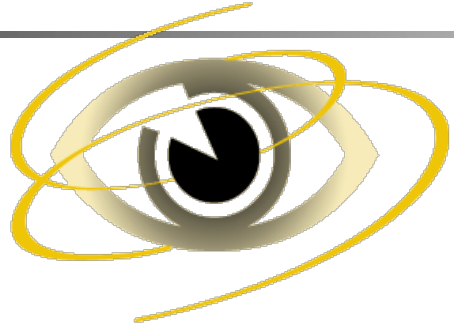
Types of Base Variable Transformation

For Numeric Variables:

- Scale transformation
 - Range or Centering transformation that does not change shape of the distribution
- Shape transformation
 - Power series and other transformations (such as double-standardization) that change both scale and shape
 - Examples are square, square root, inverse, log, and so on
- *Numeric to Categorical* transformation
 - Quantile, Bucket, Optimal Binning, and so on

For Categorical Variables:

- Combine very rare classes into “other” class
- Convert to numeric via WOE method



Checking Distributions and Handling Transformations

- This demonstration illustrates using SAS Enterprise Miner to get a feel for data, checking distributions, and handling transformations.



Summary of Checking Distributions and Handling Transformations

- Of the six base variables, there are three numeric variables and three categorical variables.
- The three categorical variables do not seem to have *very rare* classes.
- Of the three numeric variables, **HDCNT_LAST** and **lt_st_sales** show *large, right skew*.
 - Max. Normal method indicated log transformation for these two variables.
- Max. Normal method indicated square root transformation for the variable **tele_rank**.





Demo of k-Means

Dr. Goutam Chakraborty



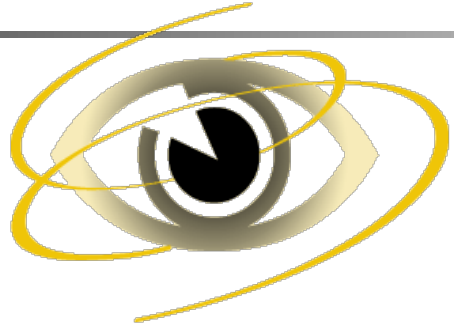
Outline

- Running k-Means and interpreting results



Plan of Analysis

1. Explore this data set using SAS Enterprise Miner. In particular, look at the distributions of *base* variables.
2. Use transformation as appropriate on *base* variables.
3. Run *k*-means using SAS Enterprise Miner.
4. Interpret results from *k*-means.



Applying k -Means

- This demonstration illustrates how to run k -means clustering and interpret the results.





Profiling k-Means Clusters

Dr. Goutam Chakraborty



Outline

- Profile kMeans clusters with base variables
 - Instead of using transformed variables, use the raw (untransformed) base variables for ease of business understanding.
- Profile kMeans clusters with descriptor variables



Recap Profiling Clusters : The Big Questions

Several types of questions are often asked in profiling:

- How is the average member *of one cluster* different from an average member *of a different cluster*?
- How is the average member *of any cluster* different from the average member *of the entire data*?
- How does the *distribution* of a variable *within a cluster* compare to the *distribution* of the same variable in the *entire data*?
- Which variables are *most important predictors* for **each** cluster?



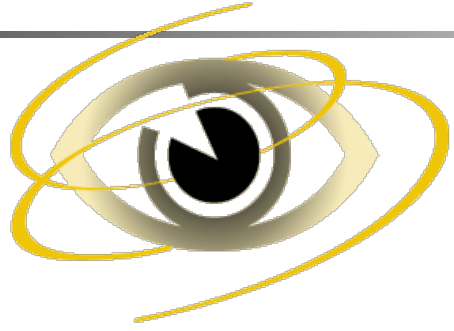
Recap Profiling Clusters with Bases

- Profiling involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters.
 - For **numeric** variables, this involves
 - comparing the mean of each variable across clusters
 - comparing the mean of each variable in a cluster with the mean for the same variable for the entire data
 - comparing the distribution (histogram) of each variable in a cluster with the distribution of the same variable for the entire data
 - For **categorical** variables, this involves comparing % members in each category within a cluster with the % members in the same category for the entire data



Profiling Clusters using Base Variables

- ❑ Save/export data from SAS EM using SAS code and then Use SAS EG and ANOVA on the saved data.
 - Not demonstrated but you should try on your own
- ❑ Use **Segment Profile** node in SAS EM
 - Set roles of untransformed base variables from “rejected to input”
 - Use results from segment profile node along with means/frequencies from SAS code (see below) to tell a story about each cluster
- ❑ Use SAS Code in SAS EM to get the *means by clusters* for interval variables and *cross-tab by clusters* for categorical variables
 - Create index and report index instead of raw mean/frequencies – not demonstrated here but you should try on your own



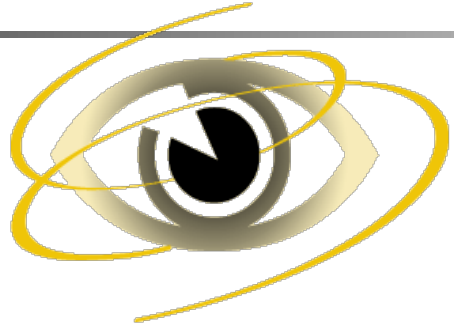
Profiling k -Means Clusters

- This demonstration illustrates how to profile k -mean cluster using bases.



Summary of Cluster Profiles (Bases)

Bases	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6
Life Time Sales	\$2,217	\$9,409	\$1,280	\$1,084	\$8,269	\$3,595
Grow Dec	More of d10	More of d01 – d03, d05	More of d08- d10	More of d08- d10	More of d04	More of d0-d03
Tele Rank	55,989	26,529	64,252	60,470	28,234	41,981
RFM Group	Mostly Groups 2 and 1	Group 6	Group 4	Group 0	Mostly Groups 3 and 6	Group 5
Employee Count	117	208	100	97	196	144
Industry	Slightly More Education	Slightly More Manufacturing	Slightly More Services	Somewhat More Unclassified	More Manufacturing	Slightly more public admin



Profiling k -Means Clusters Using Descriptors (**Self Study**)

- This demonstration illustrates profiling k -means clusters using descriptors.
- Do it on your own. **Note:** set the value of minimum worth in segment profile node to 0.001 (it was 0.01 for base variables)



Summary of Cluster Profiles (Descriptors)

Descriptors	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6
Type Customer		More gold, growable and platinum			More growable	More growable
Acct Recency	8.8	6.1	8.2	4.7	6.6	7.6
Divisions	1.6	2.0	1.5	1.5	2.0	1.8
Lifetime Orders	7.9	29.0	4.6	3.9	27.1	12.5
Credit Risk	Slightly more New and 1001	More 1003	More New	More New	More 1003	More 1003
Zone		Slightly more North NE			Slightly more Central	



Next Steps (for **You** to Do on your own)

- Sort the data differently and rerun cluster analysis to check for order effect.
 - This is one way to force the algorithm to use a very different set of starting seeds.
- Use different transformations on the base variables.
- Trim (or, Winsorize) outliers/atypical observations.
- Use a different method (you used Average) such as Ward's or Centroid method for the first stage in the clustering algorithm.
- Force a different number of cluster solutions (by switching from automatic to user specify in SAS Enterprise Miner and then specifying the number of clusters) and evaluate those solutions.

