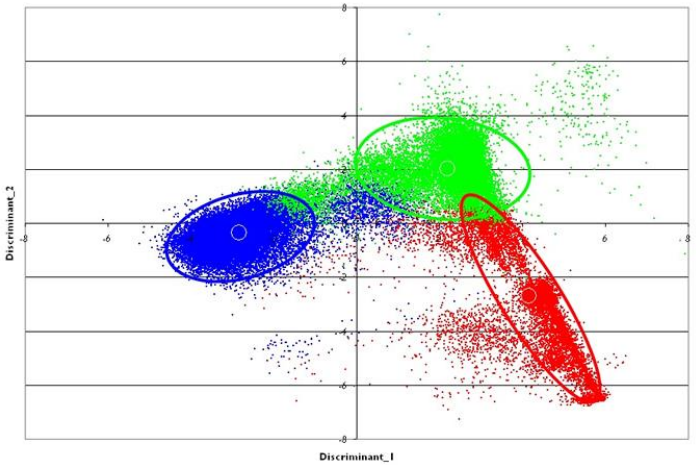
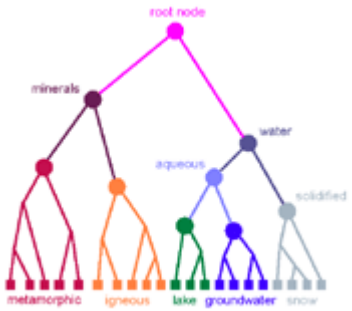
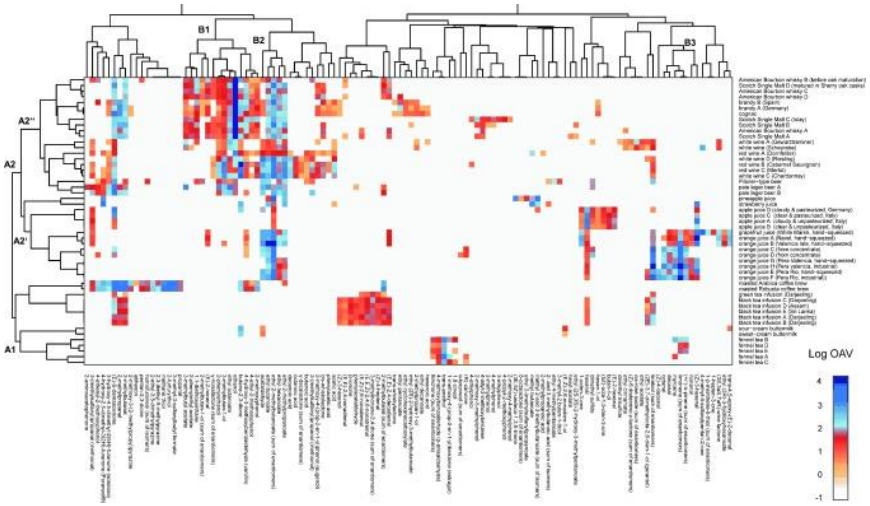
DISCOVERING PATTERNS



Module 7 : Cluster Analysis





Dr Rita Chakravarti
Institute of Systems Science
National University of Singapore
Email: rita@nus.edu.sg

© 2018 NUS. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.

© 2018 NUS. All rights reserved.

CLUSTER ANALYSIS

© 2018 NUS. All rights reserved.

Agenda

At the end of this module, you will be able to:

- Understand what is Cluster Analysis
- Application of Cluster Analysis
- Types of Clustering Methods
- How to Profile Clusters
- How to Validate the Created Clusters
- Understand how to perform these techniques using JMP, R & SPSS
- Limitations of Cluster Analysis

© 2018 NUS. All rights reserved.

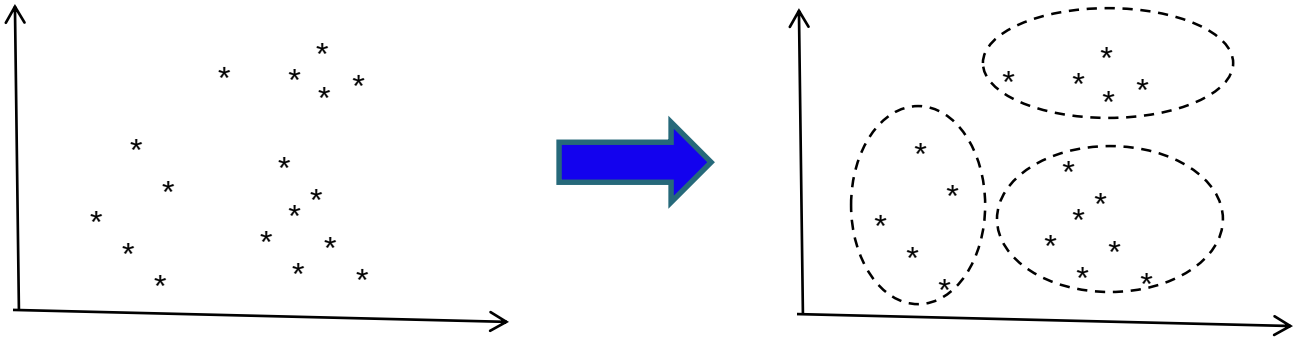
What is Cluster Analysis?

Goal of Clustering

- To segment the data into a set of ***homogeneous*** clusters (i.e.: members within same clusters are similar enough) of observations for the purpose of generating insight to ***formulate cluster strategy***

- i.e. each object in a cluster should have more similarity within the same group such as characteristic, lifestyle, purchase preferences, product holdings etc. **but**

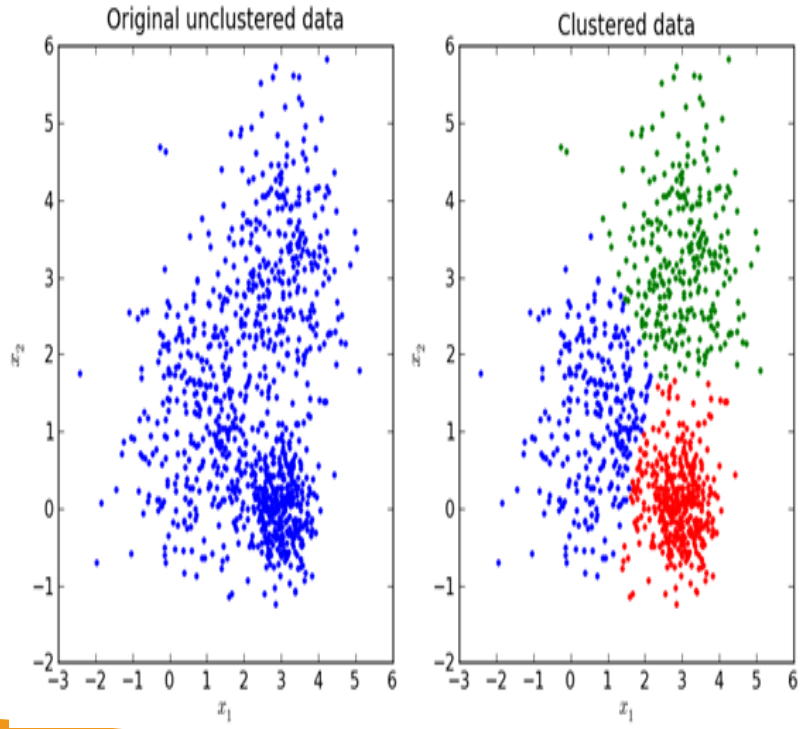
- Less similarity between groups

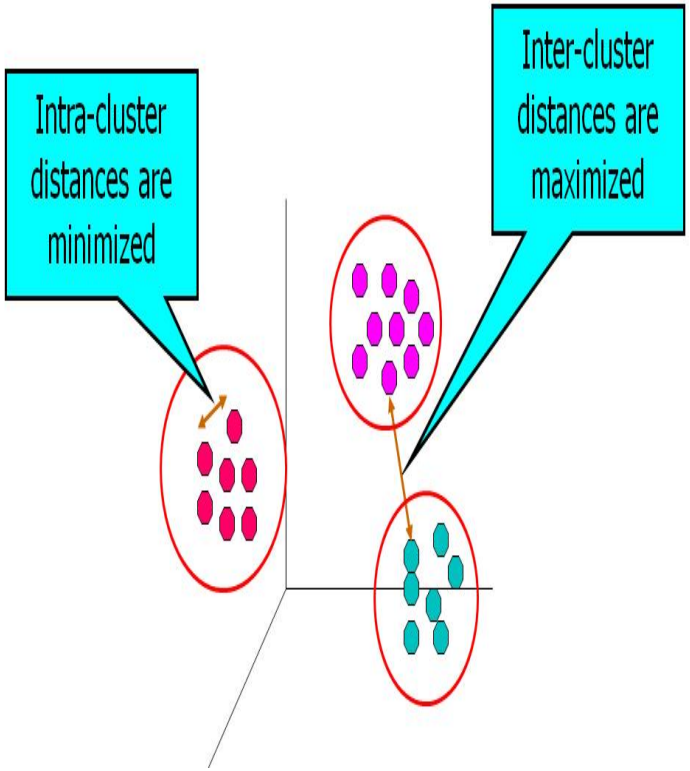


© 2018 NUS. All rights reserved.

Characteristics of Cluster Analysis

Cluster Analysis **is an unsupervised learning - there are no pre-defined classes**





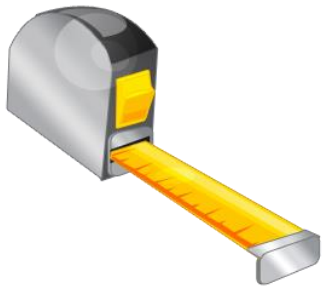
© 2018 NUS. All rights reserved.

Technicalities of Cluster Analysis

- **Need to begin by determining a distance measure** which will represent how far apart two data points are

- If each data point is represented by straightforward **numerical variables** then it is relatively easy;

- Measures of Distance
  - ***Euclidean distance.***
  - Squared (or absolute) Euclidean distance.
  - City-block (Manhattan) distance.
  - Chebychev distance.
  - Mahalanobis distance (D2)
  - ……



- In most cases we use Euclidean distance

- However it is important that the measures of distance are ***scaled or normalized*** so that one component of distance does not dominate
  - One measure of scale is the Z score
  - Transform the data so that the distances have mean 0, and variance of 1

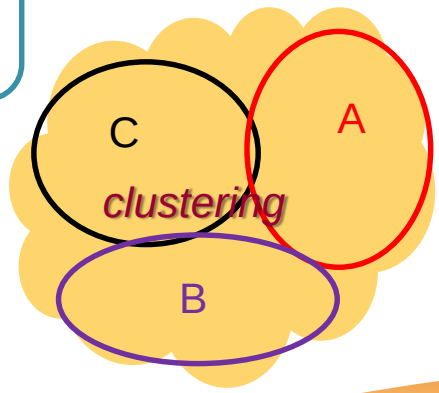
© 2018 NUS. All rights reserved.

Utility of Cluster Analysis

Decide action

- Specific action on particular group, e.g.:
  - Targeting on group B (customers with potential interest in some new service)
  - Increasing influence to C (customers with possibility in developing some interest in new service)
  - Identify similar cities for test marketing
  - Identify group of firms that are prime target for takeover

Description of each group is achieved through **profiling**



© 2018 NUS. All rights reserved.

Cluster Analysis : Be Mindful About Data Type

With categorical* variables it is more difficult to define distances

- If the categories can be related to levels of some measurement then we can define distance/association measures based on this
- Example: Your measurement scale may be (Likert Scale)
  - Strongly like
  - Like
  - Neutral
  - Dislike
  - Strongly dislike

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
Scale Week is a worthwhile feature on The Research Bunker Blog.	○	○	○	●	○
I would like to read more posts about survey rating scales.	○	○	○	○	●
Vance Marriner is, without a doubt, the most insightful contributor to The Research Bunker Blog.	●	○	○	○	○

- If one object scores “Like” and another scores “Neutral” then the distance could be 1
- If one object scores “Like” and another scores “Dislike” then the distance could be 2

*https://pdfs.semanticscholar.org/aa3c/6df25998ed83e005bb5ea13acdcf5d58b500.pdf

© 2018 NUS. All rights reserved.

Distance Measure Used For Clustering

Euclidean Distance is the most popular distance measure

- Given two cases *i*, and *j* (in the *p*-dimensional space) the distance is defined by

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2}$$

- Euclidean distance is highly scale dependent.
- Normalization on continuous measurements is recommended before computing the Euclidean distance, to convert all measurements to the same scale

Other measures

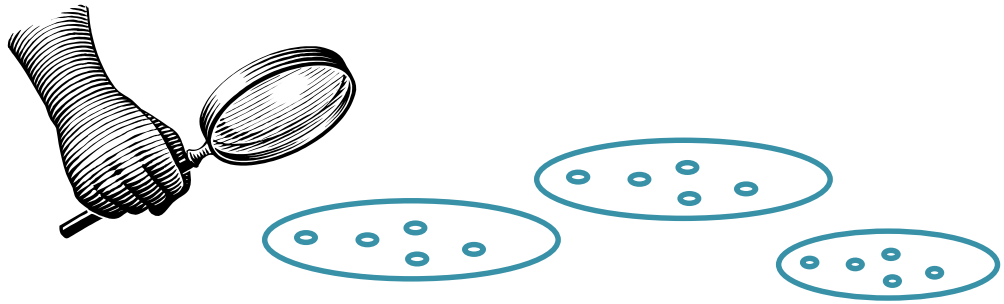
- Manhattan distance for numerical data
- There are also measures for categorical* data and mixed data

*Reference: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.140.8831&rep=rep1&type=pdf

© 2018 NUS. All rights reserved.

Cluster Interpretation For Strategy Formulation

Interpreting Clusters



Can the clustering be explained in practical terms?

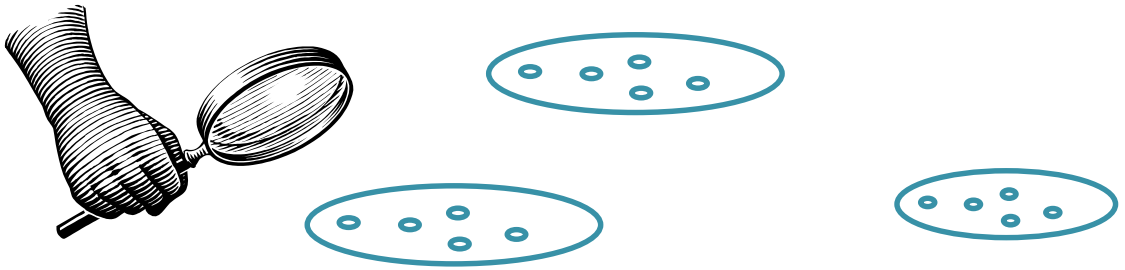
Examine the value of each variable of the cluster centroid, (the cluster profile) of each cluster.

- Look at the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters
- Cluster solutions failing to show substantial variation between clusters indicate that this may be a spurious clustering
  - Other cluster solutions should be examined.

© 2018 NUS. All rights reserved.

Interpretation May Require Research!!

Interpreting Clusters



The clusters should also be assessed against the analyst's

- Prior expectations based on theory
- Or practical experience in previous work

If no practical interpretation can be offered, more original research may be needed into the background of the data

May use case studies, specific measures, or pre-existing labels

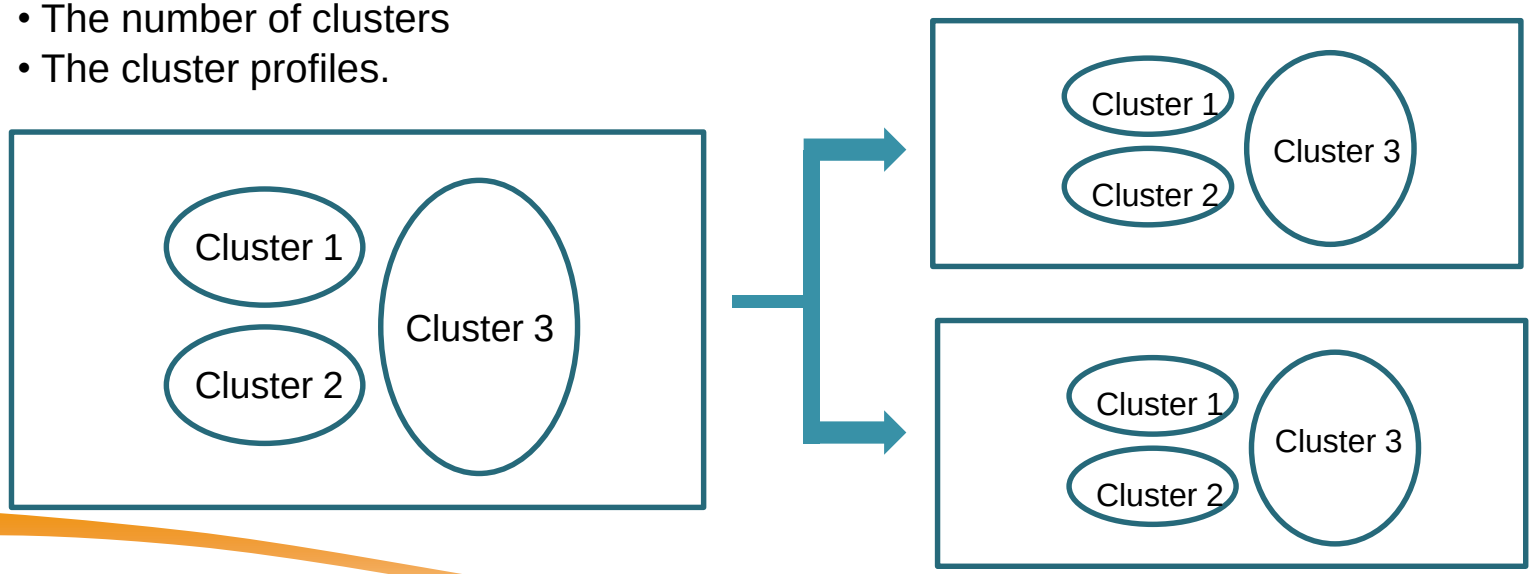
© 2018 NUS. All rights reserved.

Validation of Clusters

Validation is essential in cluster analysis since the clusters are descriptive of structure and require additional support for their relevance:

One method is Cross-validation

- Cross-validation empirically validates a cluster solution by
  - Creating two sub groups from the original group of data points by randomly splitting the group
  - Perform cluster analysis on **each** group
  - Comparing the cluster solutions from each subgroup for consistency with respect to the
    - The number of clusters
    - The cluster profiles.



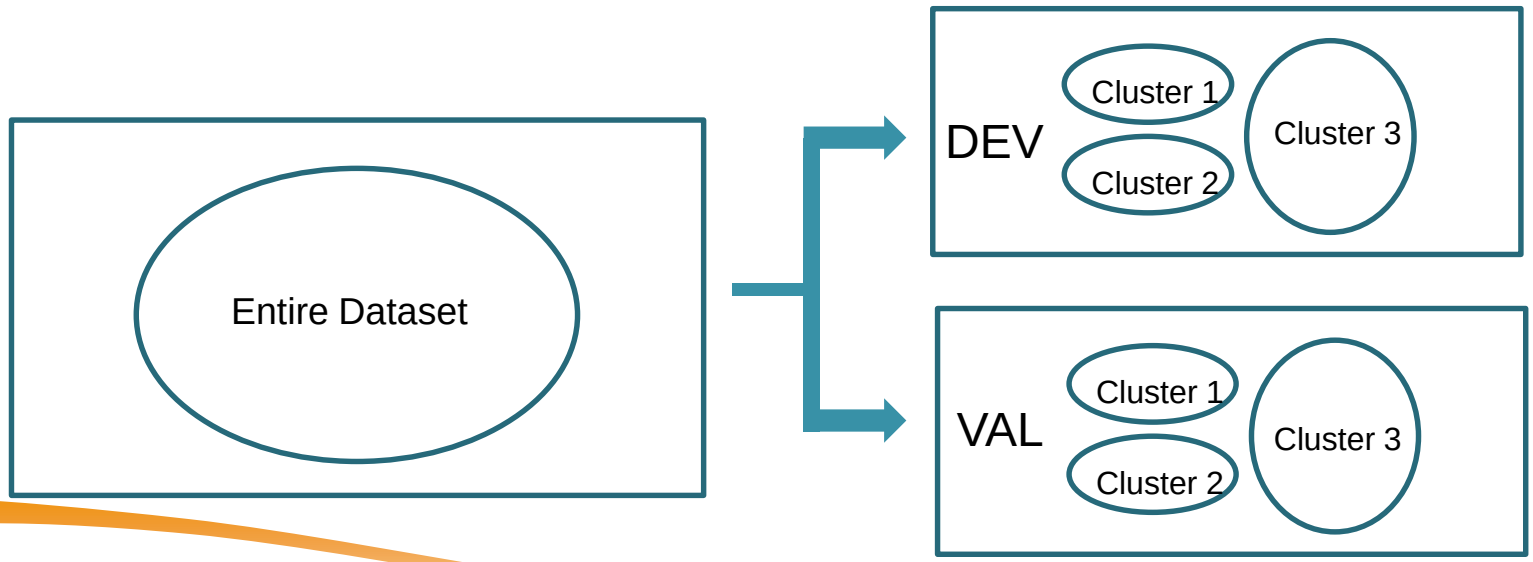
© 2018 NUS. All rights reserved.

Validation of Clusters

Second method is to split randomly the clustering data set into two parts :

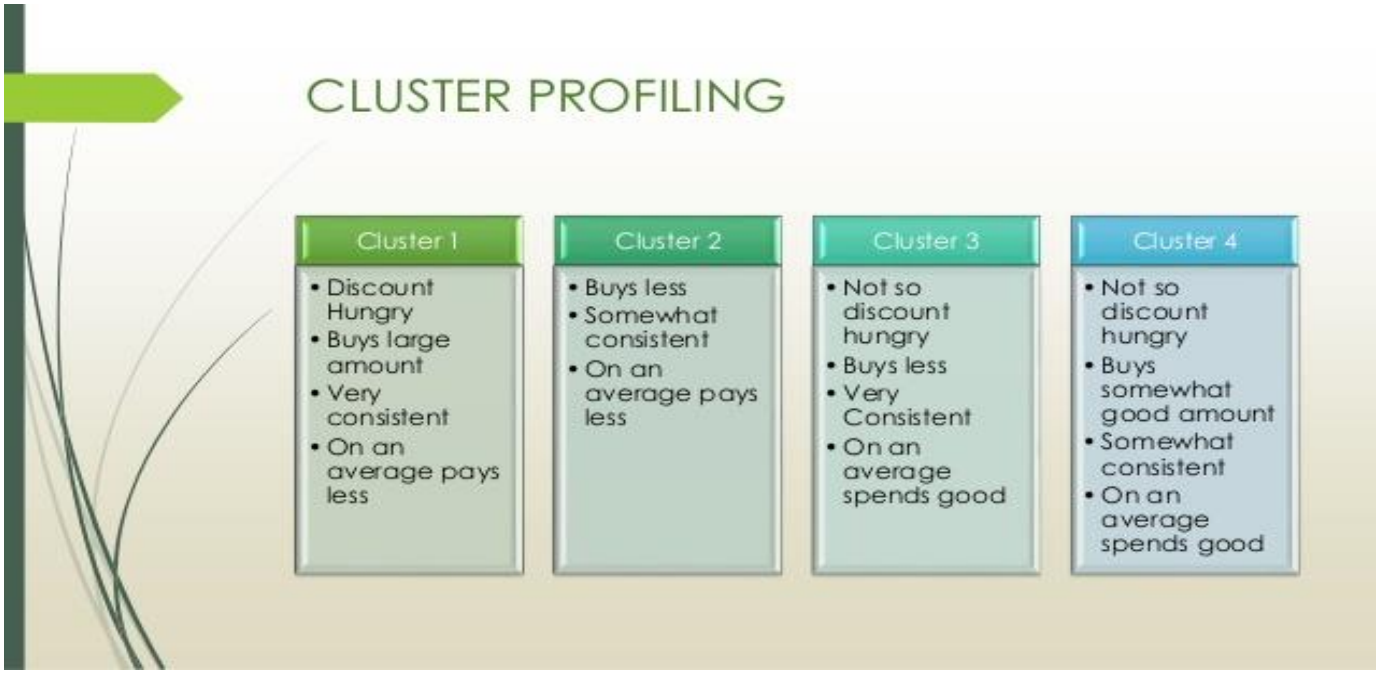
Development and Validation (50:50)

- Perform cluster analysis on development sample **only**
  - Create cluster profiles
  - Validate the results on the validation sample using development sample clustering algorithm (which becomes the implementation code)



© 2018 NUS. All rights reserved.

Profiling of Clusters



Profiling of Clusters should be done with those variables where their pattern or distribution is different within each cluster

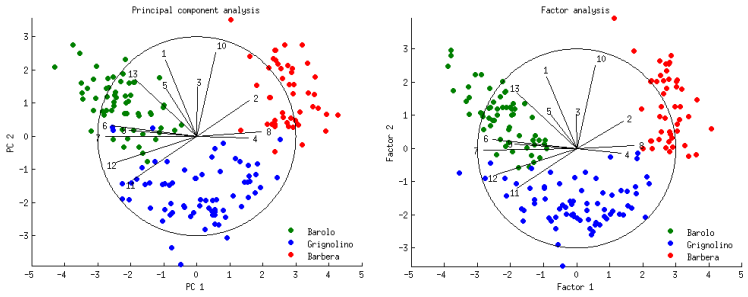
e.g. If age is a clustering variable and the age distribution is same in the parent (development) data set and within all clusters then age is not a discriminating variable and **should not** be used for profiling

© 2018 NUS. All rights reserved.

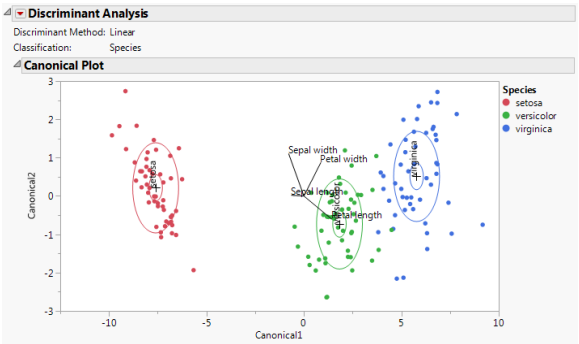
Efficient Method of Cluster Analysis

Data Exploration → Dimension Reduction → Clustering → Profiling

- **Factor analysis** reduces the dimensions (thus the number of variables), and minimizes multicollinearity effects.
- **Cluster analysis** identifies the grouping (thus structures within the data)
- **Discriminant analysis** checks the goodness of fit of the model and profiles the clusters.



Factor analysis



Discriminant analysis

© 2018 NUS. All rights reserved.

How does Cluster Analysis help us?

Determine if different clusters exist and if there is a “natural” grouping for members within the data set.

Identify those variables (or ranges of variables) that most strongly define the clusters.

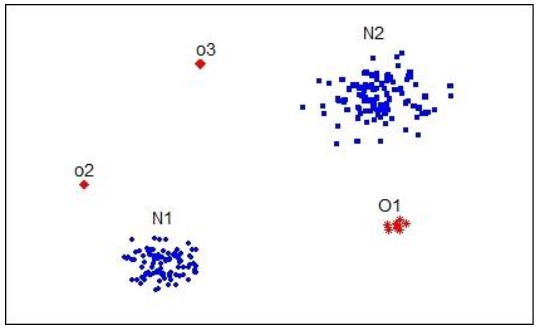
Determine the rationale underlying the clusters. The results of the clusters should make sense from a business point of view & should generate business opportunities.

Identify anomalies or outliers; i.e. members who do not fit into any cluster

Suggests a classifying scheme for new data members

© 2018 NUS. All rights reserved.

Dealing with Outliers



Outliers are data points that do not naturally fit into any cluster

- They only join groups at the end of the clustering process; i.e. when everything fits into one cluster

They may distort your clustering by changing your similarity measures

Usually they have some unusually high or low variable value

It may be necessary to remove these from your analysis to get a more well-defined clustering

But it may be good to investigate where the outlier came from

- Was it faulty data collection or transcription?
- Or was the outlier a representative of some different, new population?

© 2018 NUS. All rights reserved.

Application of Cluster Analysis

Cluster analysis has been largely used in research such as:

- **Healthcare/Medical Research**
  - e.g. classifying patients on the basis of clinical and/or laboratory type observations.
- **Urban Planning**
  - e.g. Identifying groups of houses according to their house type, value, and geographical location

© 2018 NUS. All rights reserved.

Application of Cluster Analysis

Marketing

- Clustering of neighborhoods using U.S. postal zip codes has been used successfully to group neighborhoods by lifestyle
- Group neighborhoods into 40 clusters using various measures of consumer expenditure and demographics.
- Knowledge of lifestyles can be used to estimate the potential demand for products (e.g. sports utility vehicles) and services (e.g. pleasure cruises)

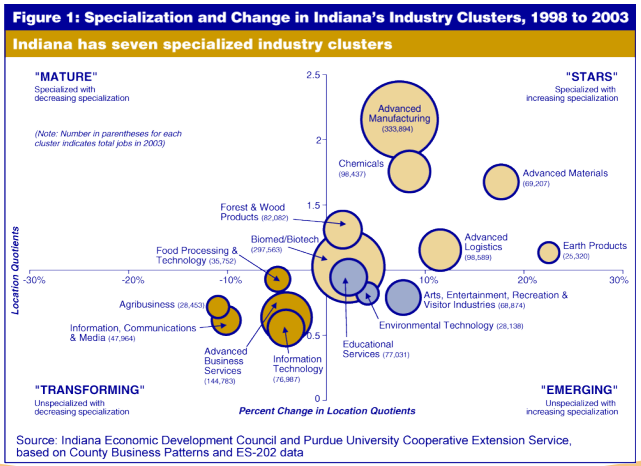
https://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/lecture-notes/lec11.pdf

© 2018 NUS. All rights reserved.

Application of Cluster Analysis

Industry analysis

- For a given industry, we are interested in finding groups of similar firms based on measures such as growth rate, profitability, market size, product range, and presence in various international markets.
- These groups can then be analyzed in order to understand industry structure and to determine, for instance, who is a competitor, and how can we capture new opportunity

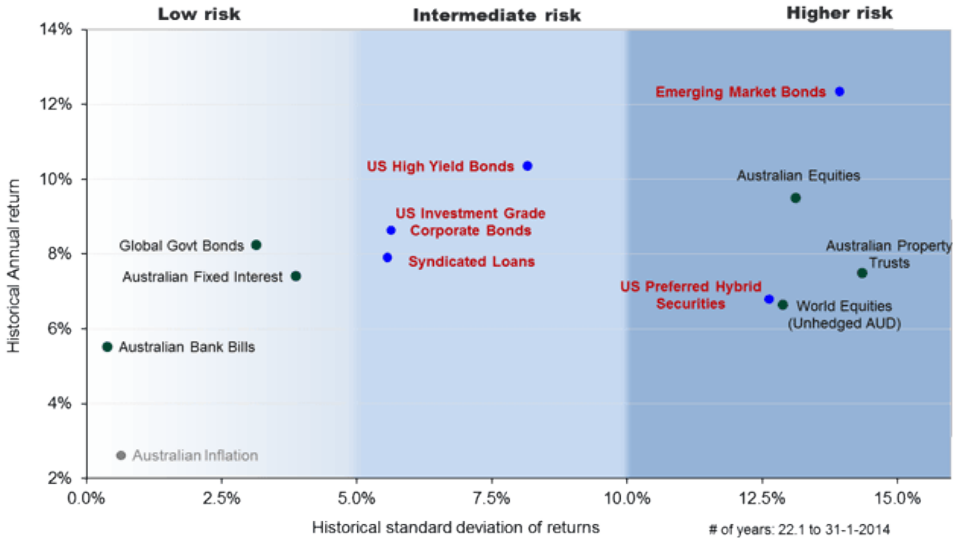


© 2018 NUS. All rights reserved.

Application of Cluster Analysis

Finance: Balancing Portfolios

- Given data on a variety of investment opportunities (e.g., stocks), one may find clusters based on performance variables such as return, volatility, and other characteristics, such as industry and market capitalization.
- Selecting securities from different clusters can help create a balanced portfolio (for a better risk management)



© 2018 NUS. All rights reserved.

Application of Cluster Analysis

Apparel Industry: Uniforms for women in U.S. Army

- The design of a new set of sizes for army uniforms for women
- The cluster analysis study came up with a new clothing size system with only 20 sizes to fit different body types
- 20 sizes are combinations of five measures: chest, neck, shoulder circumference, sleeve outseam, and neck-to-buttock length

A completely new insightful view can be gained by examining clusters of records

(Berry & Lincoff 1997)

© 2018 NUS. All rights reserved.

Data Sources for Clustering

Big Data

- Model systems may store and process very big data (e.g. weblogs) Example, Google's MapReduce Framework use Map Functions to distribute the computation across different machines.

Multimedia Data

- Clustering Video / Audio Clips
- Temporal ordering of records represents its meaning (Flickr, YouTube)

Biological Data

- Clustering Images (Position of Pixel represent its context)
- Gene/Protein Sequences

Social Network Data

- Clustering different typed nodes / links together (e.g. NetClus)

Customer Behavioural Data (Bank, Insurance, Telco….)

- Clustering different types of customers

Manufacturing Data

- Clustering different types of machines based on performance

© 2018 NUS. All rights reserved.

TYPES OF CLUSTERING METHODS

© 2018 NUS. All rights reserved.

Types of Cluster Analysis

Hierarchical Clustering

- Agglomerate Method
- Divisive Method

K-Means Method

Kohonen's Clustering Method (Self Organizing Maps SOM)

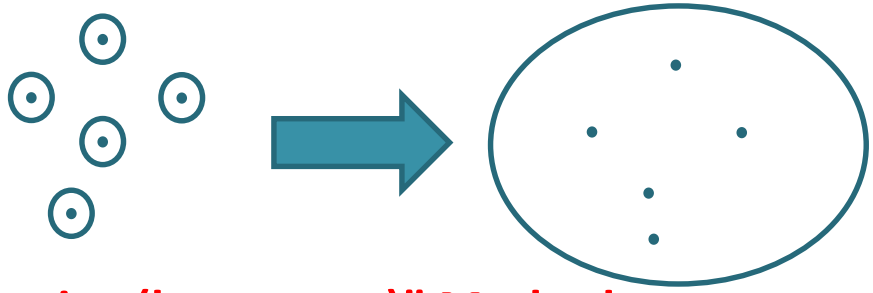
K Nearest Neighbour (KNN)

Two Step Clustering

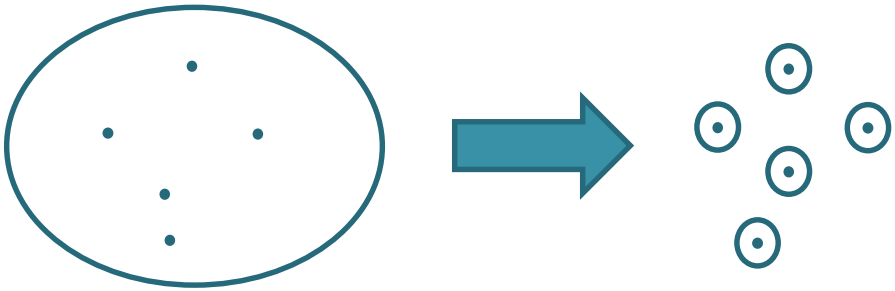
Source: http://www.stat.columbia.edu/~madigan/W2025/notes/clustering.pdf

© 2018 NUS. All rights reserved.

Two Approaches to Hierarchical Clustering



- **"Agglomerative (bottom up)" Methods**
  - We build up clusters from a group of disparate individuals



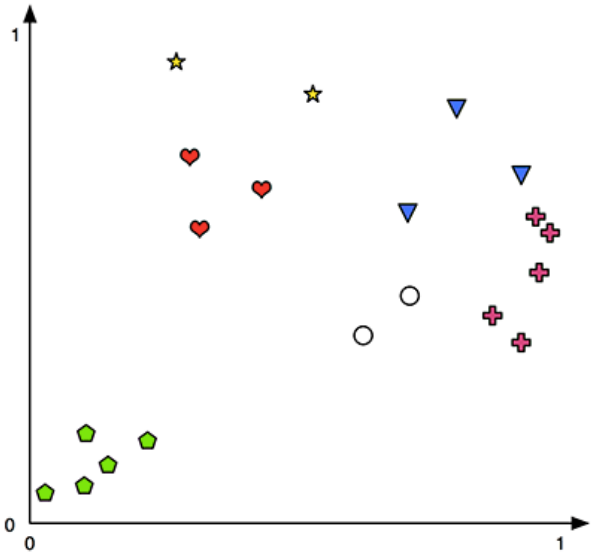
- **"Divisive(top down)" Methods (Partitional)**
  - Breakdown a single cluster (of all individuals) into several different clusters

© 2018 NUS. All rights reserved.

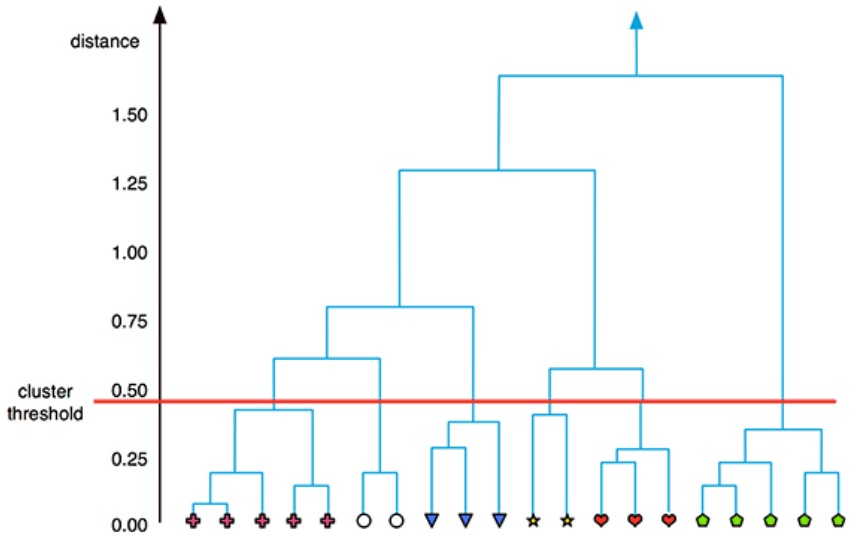
Agglomerate or Hierarchical Clustering

Agglomerate or Hierarchical Clustering

- Start with each cluster comprising exactly one record and then progressively agglomerating (combining) the two ***nearest*** clusters until there is just one cluster left at the end, which consists of all the records.



sample dataset



hierachical clustering of the sample dataset

© 2018 NUS. All rights reserved.

Agglomerate or Hierarchical Clustering

Quite outdated nowadays

It is called hierarchical or agglomerate because it starts with a solution where each record comprises a cluster & gradually groups records up to the point where all of them fall into one supercluster.

In each step it calculates the distances between all pairs of records & groups the most similar ones

The analyst should consult this information, & identify the point where the algorithm starts to group disjoint cases, & then decide on the number of clusters to retain.

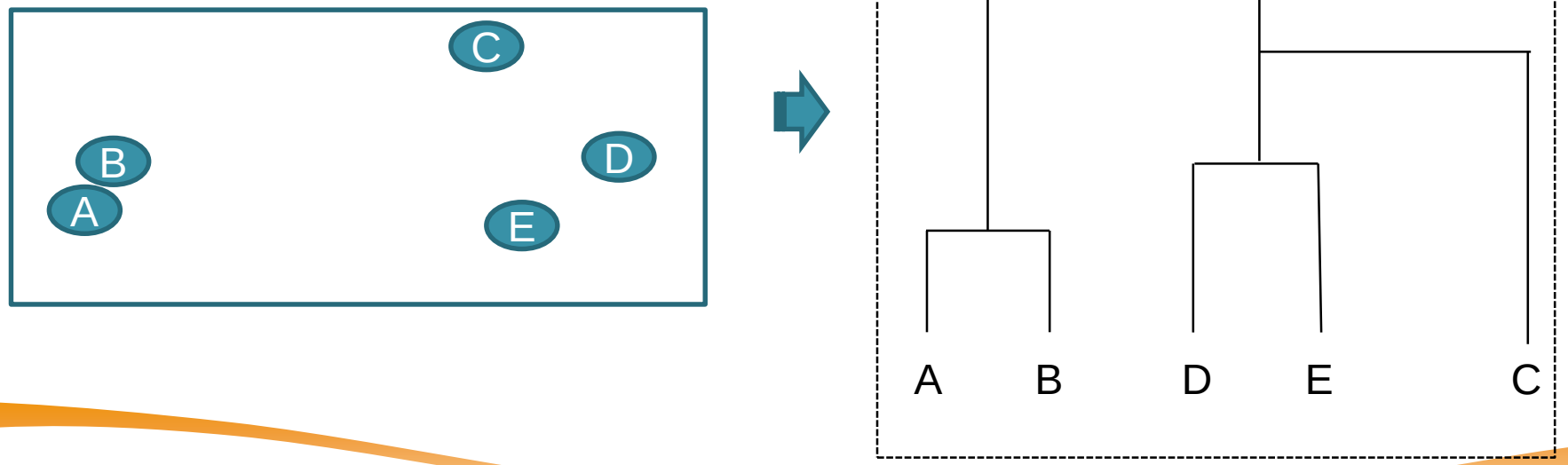
This algorithm cannot effectively handle more than a few thousand cases. Thus, it cannot be directly applied in most business clustering tasks.

© 2018 NUS. All rights reserved.

Agglomerate or Hierarchical Clustering

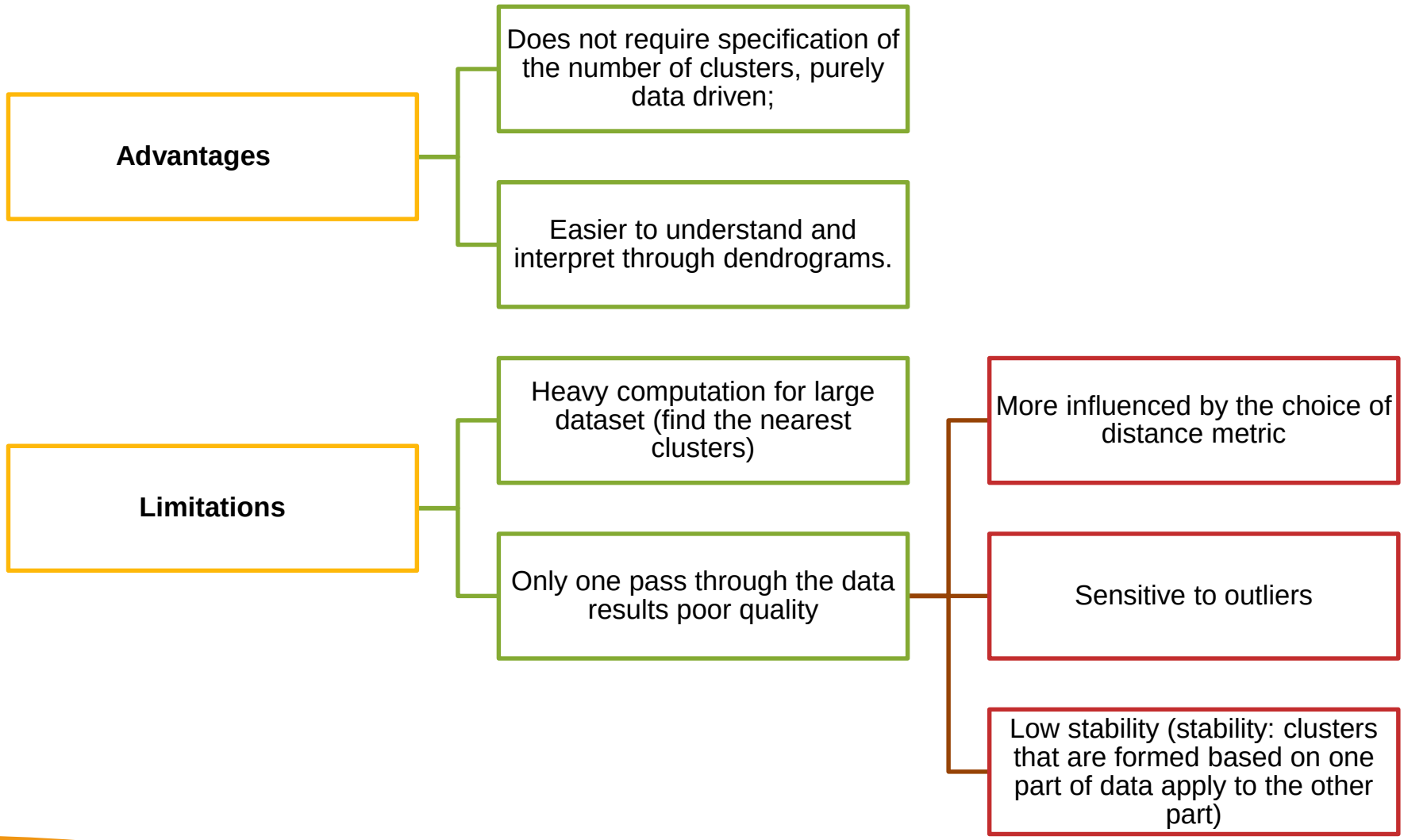
A dendrogram is a tree diagram used to illustrate the arrangement of the clusters produced by hierarchical clustering.

Example, suppose the below data is to be clustered using Euclidean distance as the distance metric.



© 2018 NUS. All rights reserved.

Agglomerate or Hierarchical Clustering



© 2018 NUS. All rights reserved.

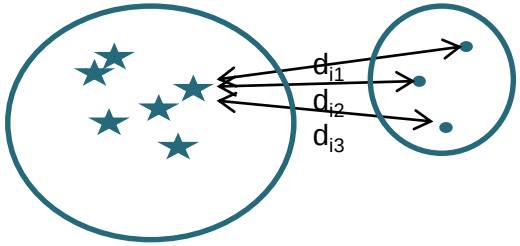
Agglomerate or Hierarchical Clustering

1. Start with all the data points

- Determine the distances between every data point
- Work out which two data points are closest together
- Combine these two data points together to form your first cluster

2. Then recalculate the distances

- Between every data point
- Also between every data point and every cluster
- Also between every cluster



3. Identify the two clusters with the smallest average distance between them

4. Combine these two clusters together to form a cluster

5. Repeat steps 2 to 4 until all data points are in a single cluster

© 2018 NUS. All rights reserved.

Divisive Clustering Methods

A cluster hierarchy can also be generated top-down

This variant of hierarchical clustering is known as divisive clustering

- One starts with all objects in one cluster
- Then the cluster is split using a flat* cluster algorithm (typically used in document clustering)
- The procedure is applied recursively until each object is in its own singleton clusters
- Top-down clustering is conceptually more complex than bottom-up clustering since a second, flat clustering algorithm as a ``subroutine'' is required
- It has the advantage of being more efficient if one does not generate a complete hierarchy all the way down to individual objects
- k-means is perhaps the most widely used flat clustering algorithm due to its simplicity and efficiency

*http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab6-ClusteringWithKMeans.pdf

© 2018 NUS. All rights reserved.

Nonhierarchical Clustering

***k*-Means** is a typical nonhierarchical clustering algorithm

- The goal is to divide the sample into *predetermined* number *k* of non-overlapping clusters so that clusters are as homogeneous as possible with respect to the measurements used

Kohonen clustering (sometimes called *k*-Means with a twist)

- Using Kohonen network, which is an important unsupervised learning neural network performing *Self Organizing Map*
- Is more data driven, without the requirement of prespecified number of clusters

k-Nearest Neighbors algorithm (KNN):

- It is a non-parametric method used for classification and regression.
- In both cases, the input consists of the *k* closest training examples in the feature space.
- The output depends on whether *k*-NN is used for classification or regression

© 2018 NUS. All rights reserved.

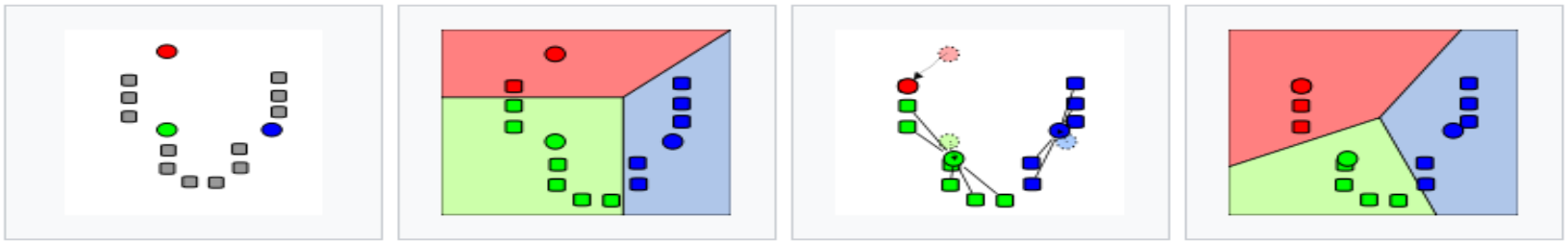
K – Means Clustering

This is an efficient & perhaps the fastest clustering algorithm that can handle both long (many records) and wide datasets (many input fields)

It is distance based and unlike the hierarchical algorithms, it does not need to calculate the distance between all pairs of records

The number of clusters to be formed is predetermined and specified in advance

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

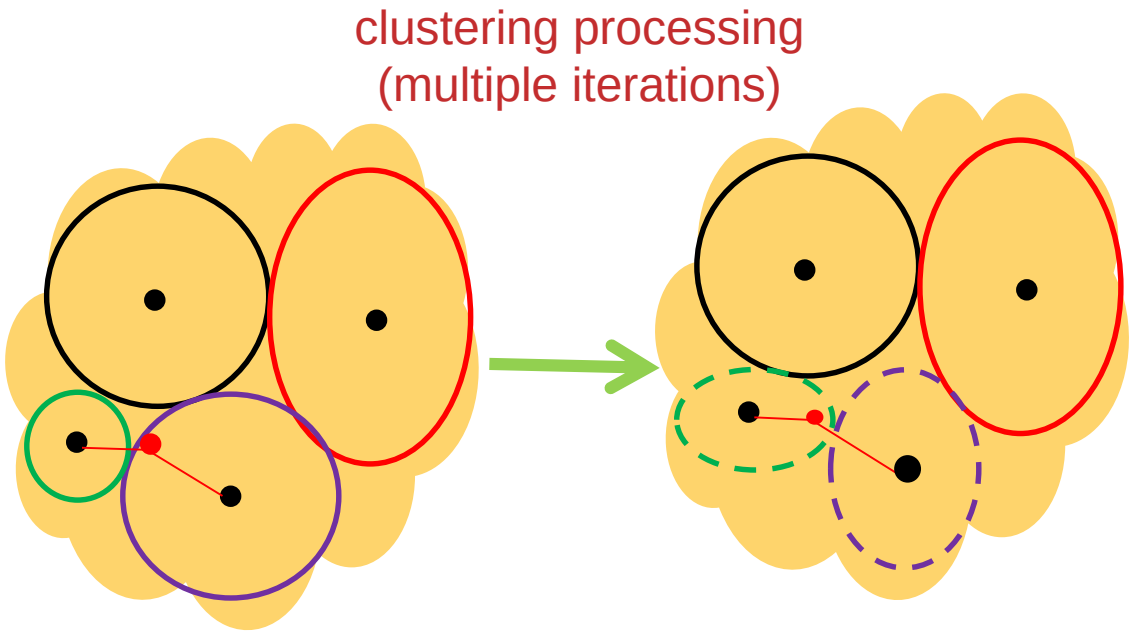
Reference : Applied Multivariate Statistical Analysis : Johnson & Wichern

© 2018 NUS. All rights reserved.

K – Means Clustering

- Usually a number of different solutions should be tried & evaluated before approving the most appropriate.

- K-Means Clustering is best for handling **continuous** clustering fields



© 2018 NUS. All rights reserved.

K – Means Clustering

Step 1 : Start with *k* initial clusters (user choose *k*)

Step 2 : At every step, each data record is reassigned to the cluster with “closest” centroid

Step 3 : Re-compute the centroids of clusters that lost or gained a record, and

- repeat the previous step as the next iteration

Step 4 : Stop when moving any more records between clusters increases cluster dispersion

© 2018 NUS. All rights reserved.

K – Means Clustering

The choice of the number of clusters can

- Either be driven by external consideration (e.g.: previous knowledge, practical constraints, etc)
- Or we can try a few different values for *k* and compare the resulting clusters

Initial partition into *k* clusters

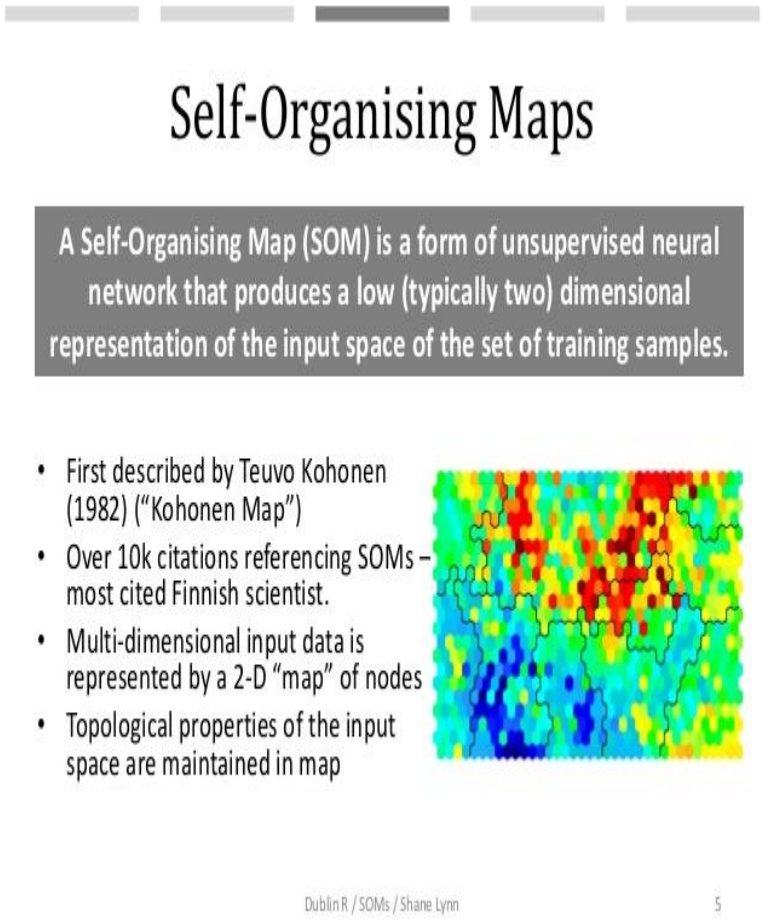
- Any available external information suggesting a certain partitioning, or the centroids of the *k* clusters, should be used
- In case no information is available for the initial partition, the algorithm can rerun with different randomly generated starting partitions to reduce the chance of resulting poor solution

© 2018 NUS. All rights reserved.

Kohonen Network (Self Organizing Maps)

Kohonen Networks are based on Neural Networks & typically a two-dimensional grid or map of the clusters, hence the name self organising maps.

Kohonen Networks usually take a longer time to train than k-Means & Two Step algorithms, & provide a different view on clustering that is worth trying.

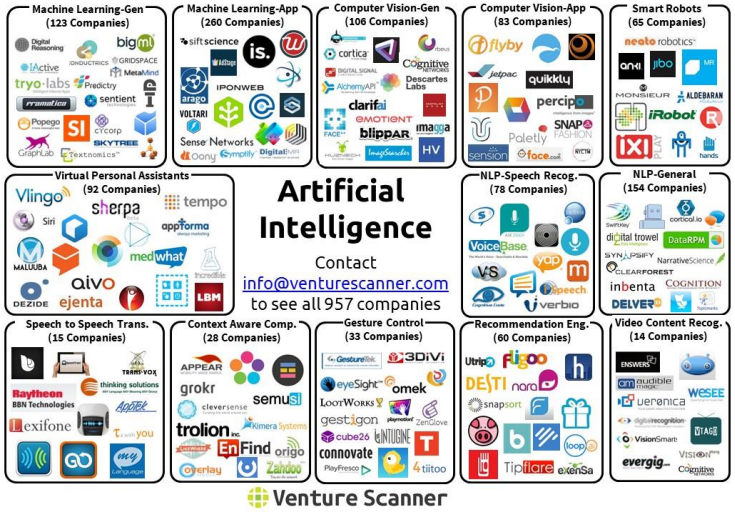


© 2018 NUS. All rights reserved.

Segmentation – A Strategic Tool – Application of Clustering Technique

One size does not fit all

Segmentation is the practice of dividing a customer base/market/softwares/tools/machines into groups that are similar in specific ways so that the description/profile of the segment can be used for strategic purposes



© 2018 NUS. All rights reserved.

Bank Case Study

Behaviour Segmentation For Salaried Customers

© 2018 NUS. All rights reserved.

Behavior Segmentation - Approach

"Segmentation"

Aim: *To identify actionable customer clusters based on the their past behavior*

Step 1: Group the customers into clusters based on a set of criteria

Step 2: Compare & identify the clusters

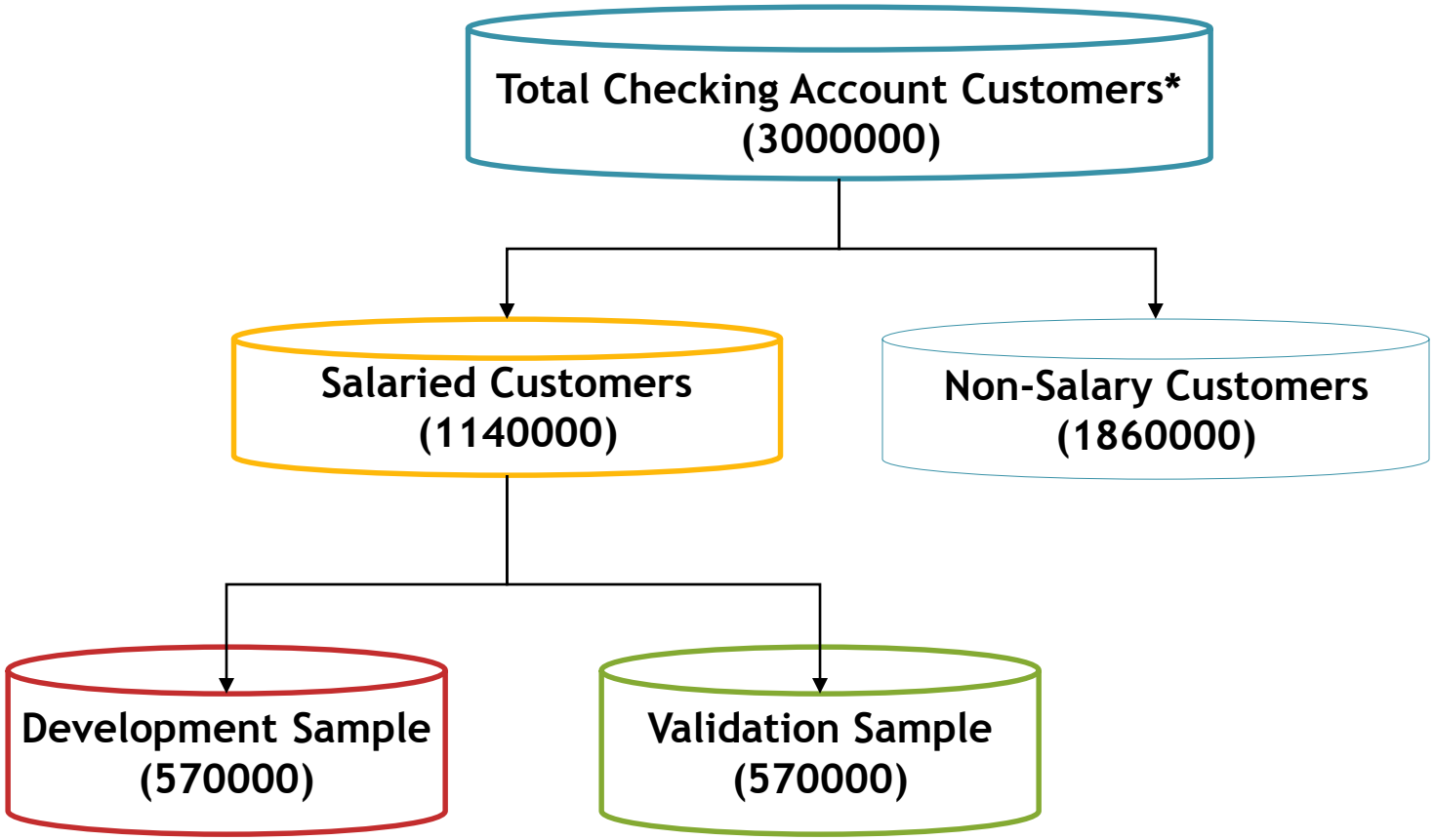
Step 3: Profile each cluster to know our customer better

Step 4: Develop strategy for each cluster

Step 5: Develop suitable actionable insights for the clusters

© 2018 NUS. All rights reserved.

Segmentation was performed on a sample of 570K customers

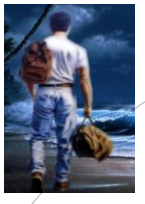


*As of October 04

© 2018 NUS. All rights reserved.

7 Key Segments were identified in the Salaried Base

Cluster 1:
Priority Customers



Cluster 7:
Dead Wood

Cluster 2:
High Asset Customers

A universe of
1.1 Million
customers

Cluster 6:
De-growing Segment

Cluster 3:
Growth Segment

Cluster 5:
Depositors

Cluster 4:
Super Active Customers

© 2018 NUS. All rights reserved.

Description of each cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	Priority Customers	**High Asset Customers**	**Growth Segment**	**Super Active Customers**
Revenue	**8200**	**5100**	**2500**	**2900**
Key Description	***Most profitable customers*** ***Highest*** *Asset Product Holding and high deposit balances* *All are Local Customers*	***2nd Most profitable customers*** ***Low*** deposits & ***High Asset*** Product Holding (amount) All are Local Customers	***Medium Profitable Customers*** *These customers are* ***growing*** *their deposit balances* *Average Asset Product Holding* *Comparatively young foreigner customers*	***Medium Profitable Segment*** ***Highest*** # of debit transactions ***Low Deposit and Average Asset*** Product Holding *Comparatively young foreigner customers*

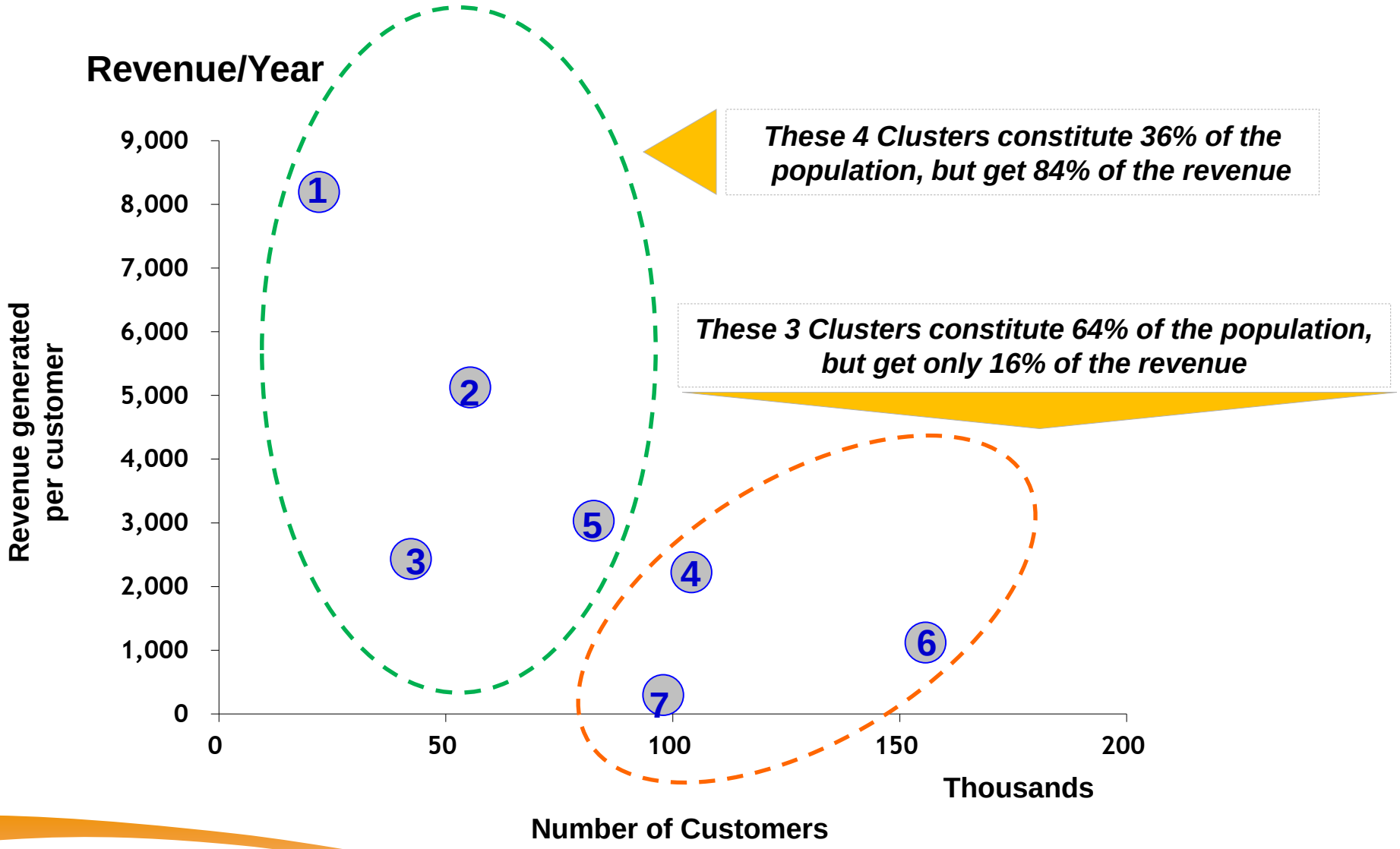
© 2018 NUS. All rights reserved.

Description of each cluster

	Cluster 5	Cluster 6	Cluster 7
	Depositors	De-growing Segment	Dead-Wood
Revenue	3000	1200	300
Key Description	***Medium Profitable customers*** ***Highest*** deposit balances ***Low Asset*** Product Holding and thus do not contribute significantly to revenue Comparatively older local customers	***2nd Lowest Profitable Customers*** Have been ***reducing*** their balances Very Low deposit balances Low Asset product holding	***Lowest Profitable Customers*** Very low balances Very low Asset product holding

© 2018 NUS. All rights reserved.

Top 4 clusters share as high as 84% of revenue with 36% of the population



© 2018 NUS. All rights reserved.

Profile of 7 clusters based on the behavior across products…

	Cluster 1 Priority Customers	Cluster 2 High Asset Customers	Cluster 3 Growth Segment	Cluster 4 Super Active Customers	Population Average
Size	**4%**	**10%**	**8%**	**19%**	**100%**
Avg. Revenue per customer	**8200**	**5100**	**2500**	**2900**	**2223**
Deposit Avg. Month-End balance	15382	3951	3259	6879	9261
Deposit Avg. bal growth	**0.5**	0.5	0.8	0.4	0.45
Liabilities - # of Debit Transactions/ Month	271	167	136	284	144
Tenure – Checking Account (months)	145	116	87	95	96
Loan holding	100%	100%	93%	73%	52%
Overdraft holding	95%	96%	89%	69%	49%
Card Holding	45%	36%	24%	21%	17%
Total Loan – Disbursed Amount	290833	195599	89579	72422	66680
Overdraft – Disbursed Amount	264132	176187	82650	66869	60728
Other Asset products – Disbursed Amt.	26701	19412	6928	5554	5952
% Local	100%	100%	98%	98%	90%
Age (yrs)	39	35	31	31	35

© 2018 NUS. All rights reserved.

Profile of 7 clusters based on the behavior across products…

	Cluster 5 Depositors	Cluster 6 De-growing Segment	Cluster 7 Dead-Wood	Population Average
Size	**15%**	**28%**	**17%**	**100%**
Avg. Revenue per customer	**3000**	**1200**	**300**	**2223**
Deposit Avg. Month-End balance	39290	1794	2519	9261
Deposit Avg. bal growth	0.5	0.2	0.7	0.45
Liabilities - # of Debit Transactions/ Month	123	90	66	144
Tenure – Checking Account (months)	118	86	80	96
Loan holding	23%	45%	11%	52%
Overdraft holding	21%	42%	10%	49%
Card Holding	11%	12%	4%	17%
Loan – Disbursed Amount	26990	40576	5920	66680
Overdraft – Disbursed Amount	9,176	36881	5363	60728
Other Asset products – Disbursed Amt.	2379	3696	557	5952
% Local	89%	87%	77%	90%
Age (yrs)	39	35	37	35

© 2018 NUS. All rights reserved.

Recommendations for each Segment

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	Priority Customers	**High Asset Customers**	**Growth Segment**	**Super Active Customers**
Strategy	***Maintain/ Enhance***	***Enhance***	***Grow***	***Grow***
Actionable Insights	***Most Profitable customers*** Engage them with personal care & keep them happy Further enhance the relationship by offering investment & insurance products	***2nd Most profitable customers*** Need to engage with these customers to keep them happy Focus on growing their deposit balances with the bank	***A Growing Segment*** As they are growing their balances, we can proactively target them with Up-sell/Cross-sell offers to increase the penetration of asset products	***Highly active in terms of debit transactions*** Can encourage them to maintain their balances Increase penetration of Asset products Can be good target for Cards

© 2018 NUS. All rights reserved.

Recommendations for each Segment

	Cluster 5	Cluster 6	Cluster 7
	Depositors	**De-growing Segment**	**Dead-Wood**
Strategy	***Grow***	***Maintain/Grow***	***Maintain/Grow***
Actionable Insights	***Highest Deposit Balances*** Focus on increasing penetration of Asset products like Auto & Mortgage, as well as Credit Cards	***Customers which are reducing their balances*** Focus on re-growing their balances Target them with X-sell offers of Asset products Keep a check on the risk	***Lowest profitable Customers with very low balances & asset product holding*** Carry on BAU campaigns without spending much on personalized marketing

© 2018 NUS. All rights reserved.

Cluster Analysis for Customer Segmentation

Advantages

- Clustering of customers according to their attribute preferences
- Cluster customer with similar behaviours/characteristics
- Clusters of similar brands/products can help identifying competitors / market opportunities

Disadvantages

- Choice of cluster-forming variables can be random
- Determine the right number of cluster often time-consuming
- Highly dependent on the analyst's interpretation

© 2018 NUS. All rights reserved.

HANDS ON EXERCISE

Statistical Tools- JMP, R, SPSS Modeler

© 2018 NUS. All rights reserved.

Cluster Analysis Using JMP Pro 13

Data Set: European Countries

© 2018 NUS. All rights reserved.

K Means Cluster by JMP

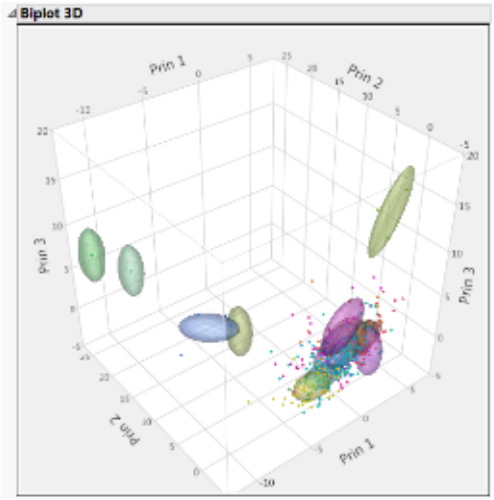
K MEANS CLUSTER

Group Observations Using Distances

Use the K Means Cluster platform to group observations that share similar values across a number of variables. Use the *k*-means method with larger data tables, ranging from approximately 200 to 100,000 observations.

The K Means Cluster platform constructs a specified number of clusters using an iterative algorithm that partitions the observations. The method, called *k-means*, partitions observations into clusters so as to minimize distances to cluster centroids. You must specify the number of clusters, *k*, in advance. However, you can compare the results of different values of *k* to select an optimal number of clusters for your data.

3D Biplot



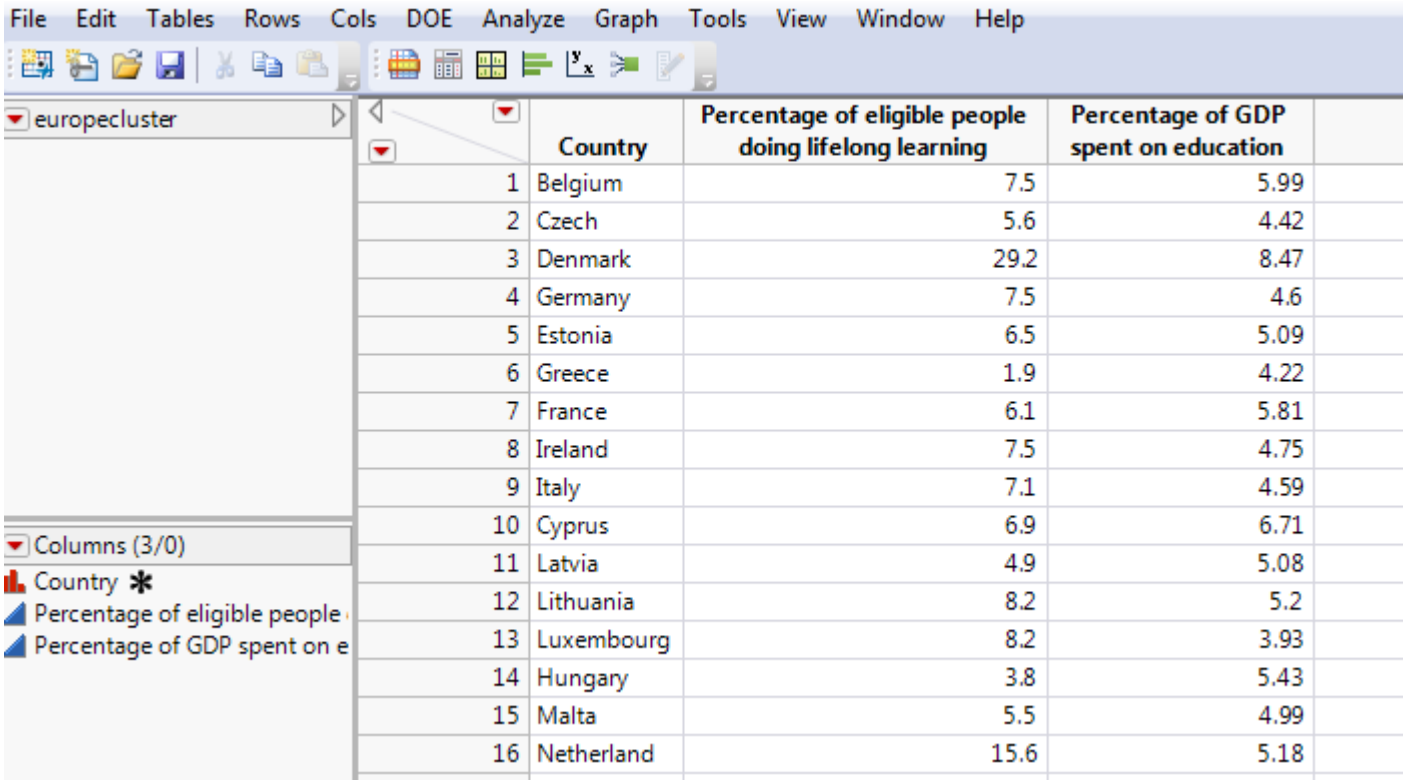
Summary of Clustering Methods

Method	Data Type or Modeling Type	Data Table Size	Specify Number of Clusters
Hierarchical Cluster	Any	With Fast Ward, up to 200,000 rows With other methods, up to 5,000 rows	No
K Means Cluster	Numeric	Up to millions of rows	Yes
Normal Mixtures	Numeric	Any size	Yes
Latent Class Analysis	Nominal or Ordinal	Any size	Yes

http://www.jmp.com/support/help/13/

© 2018 NUS. All rights reserved.

Select Appropriate Data Set



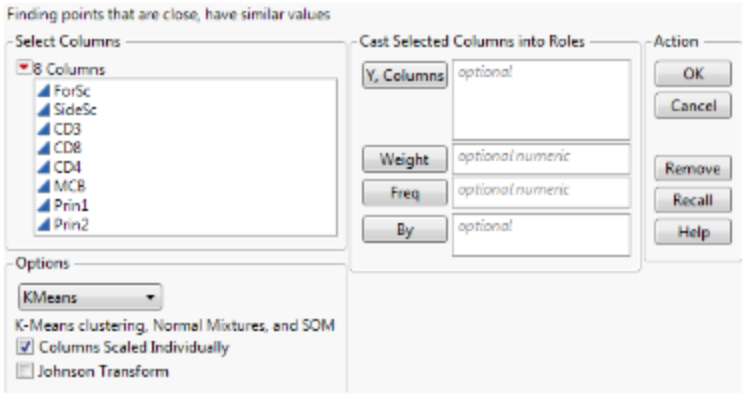
	Country	Percentage of eligible people doing lifelong learning	Percentage of GDP spent on education
1	Belgium	7.5	5.99
2	Czech	5.6	4.42
3	Denmark	29.2	8.47
4	Germany	7.5	4.6
5	Estonia	6.5	5.09
6	Greece	1.9	4.22
7	France	6.1	5.81
8	Ireland	7.5	4.75
9	Italy	7.1	4.59
10	Cyprus	6.9	6.71
11	Latvia	4.9	5.08
12	Lithuania	8.2	5.2
13	Luxembourg	8.2	3.93
14	Hungary	3.8	5.43
15	Malta	5.5	4.99
16	Netherland	15.6	5.18

Data File: European-Countries

© 2018 NUS. All rights reserved.

K Means Cluster - Launch

K Means Cluster Launch Window



Y, Columns

The variables used for clustering observations.

Note: *K*-Means clustering supports only numeric columns.

Weight

A column whose numeric values assign a weight to each row in the analysis.

Freq

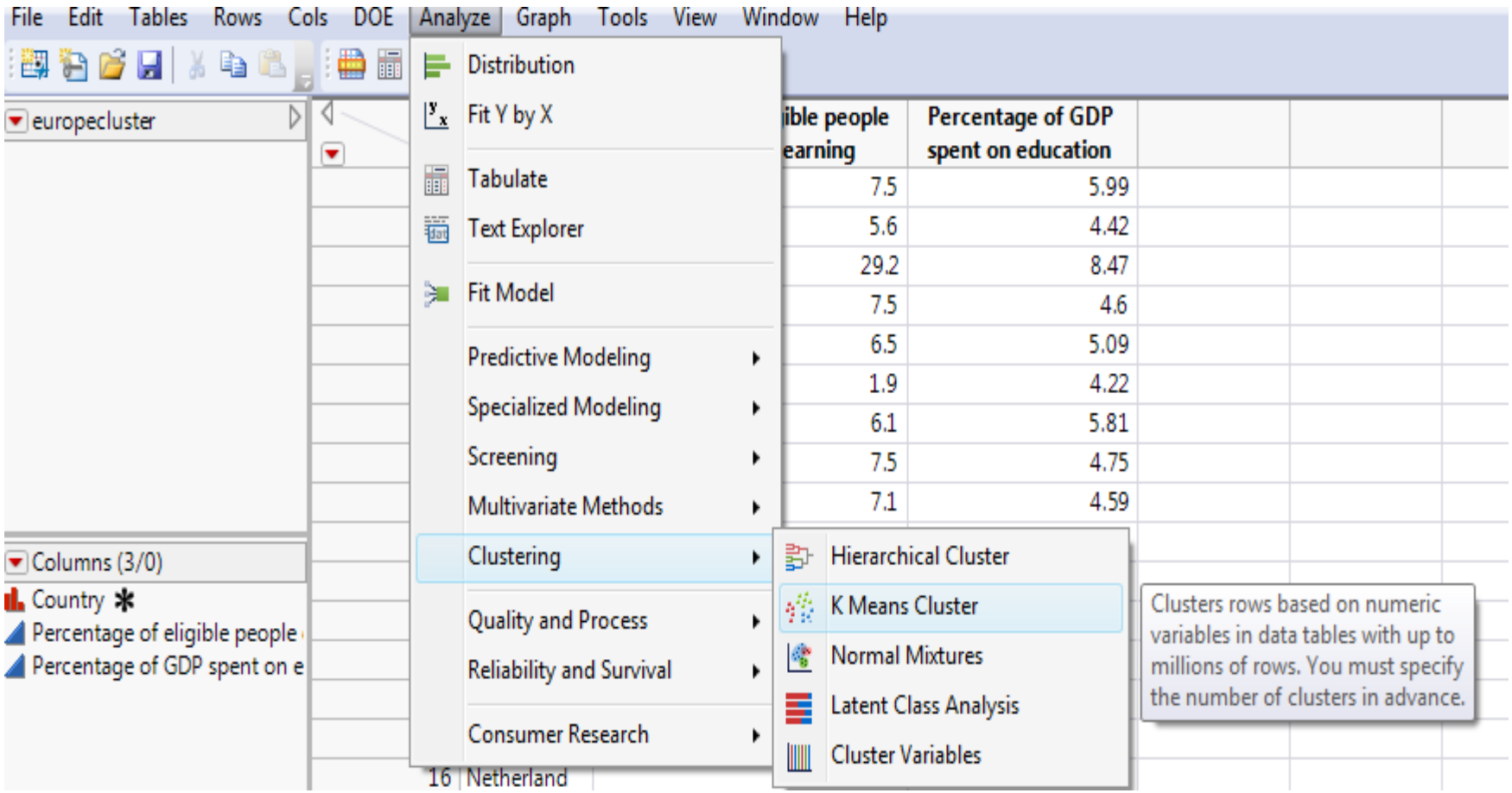
A column whose numeric values assign a frequency to each row in the analysis.

By

A column whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed. The results are presented in separate reports. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

© 2018 NUS. All rights reserved.

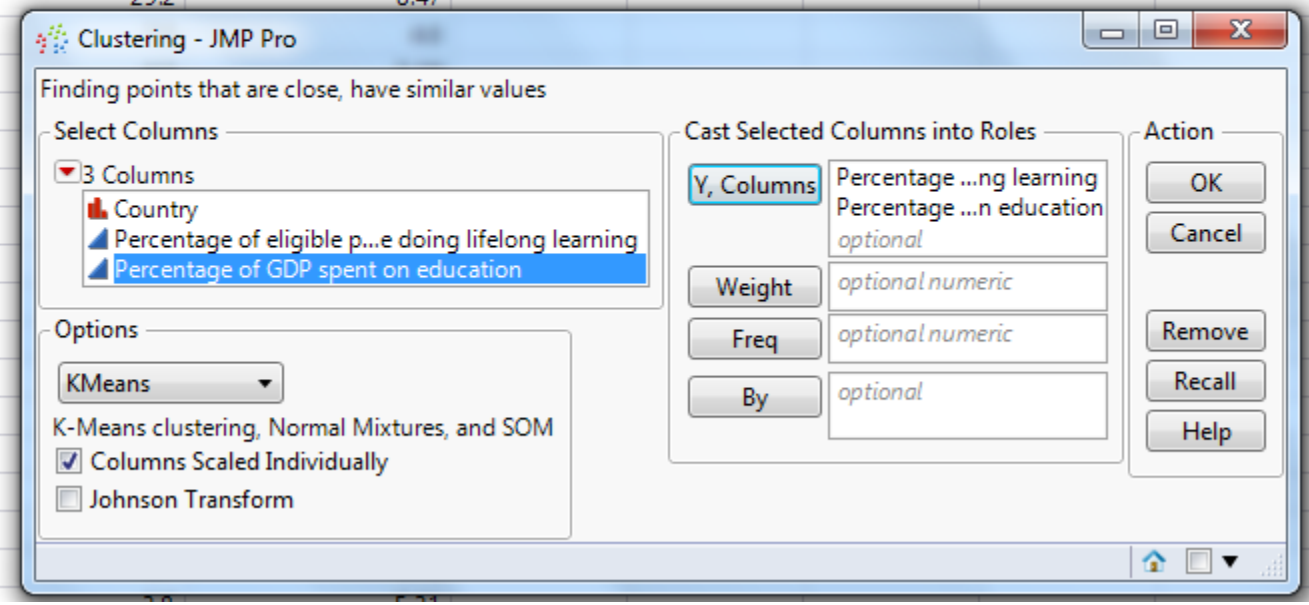
Select Clustering (K Means) Method



© 2018 NUS. All rights reserved.

Select Variables (Numeric) To be Used For Clustering

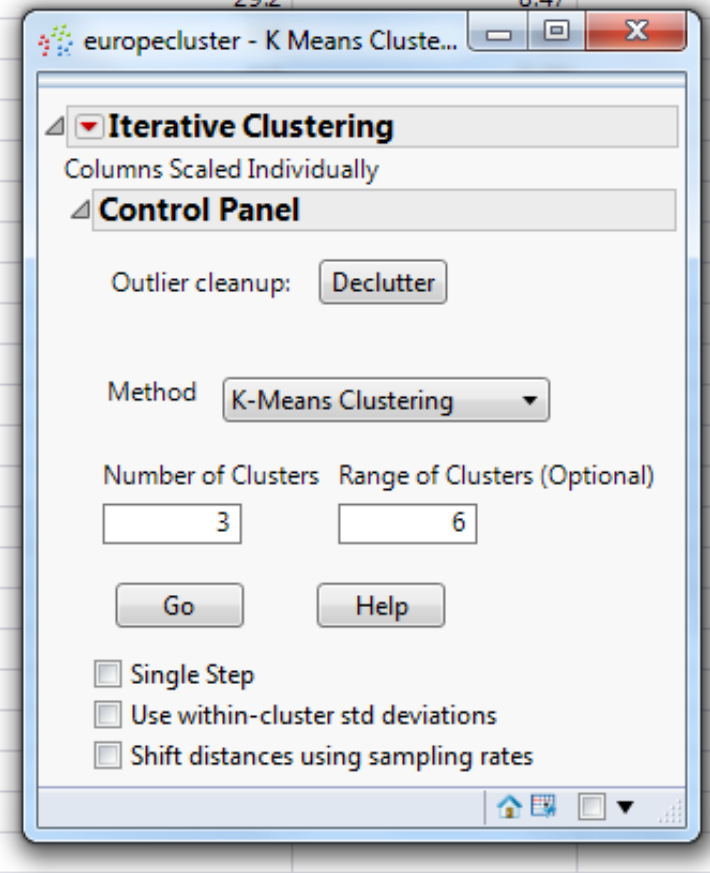
Percentage of eligible people doing lifelong learning	Percentage of GDP spent on education
7.5	5.99
5.6	4.42
29.2	8.47



© 2018 NUS. All rights reserved.

Select Number of Clusters (Individual or Range)

	Country	Percentage of eligible people doing lifelong learning	Percentage of GDP spent on education
1	Belgium	7.5	5.99
2	Czech	5.6	4.42
3	Denmark	29.2	8.47
4	Germany		
5	Estonia		
6	Greece		
7	France		
8	Ireland		
9	Italy		
10	Cyprus		
11	Latvia		
12	Lithuania		
13	Luxembourg		
14	Hungary		
15	Malta		
16	Netherland		
17	Austria		
18	Poland		
19	Portugal		
20	UK		



© 2018 NUS. All rights reserved.

Cluster Output

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	3	-1.6313	
K-Means Clustering	4	-0.8029	
K-Means Clustering	5	-0.4991	
K-Means Clustering	6	0.55953	Optimal CCC

Columns Scaled Individually

Control Panel

K Means NCluster=3

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	17	2	0
2	2		
3	1		

Cluster Means

Cluster	Percentage of eligible people doing lifelong learning	Percentage of GDP spent on education
1	6.4	5.11647059
2	21.1	5.235
3	29.2	8.47

Cluster Standard Deviations

K Means NCluster=4

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	4	2	0
2	13		
3	1		

K Means NCluster=4

Cluster Summary

Cluster	Count
4	2

Cluster Means

Cluster	Percentage of eligible people doing lifelong learning	Percentage of GDP spent on education
1	8.4	5.99
2	5.78461538	4.84769231
3	29.2	8.47
4	21.1	5.235

Cluster Standard Deviations

K Means NCluster=5

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	9	2	0
2	6		
3	1		
4	3		
5	1		

Cluster Means

Cluster	Percentage of eligible people doing lifelong learning	Percentage of GDP spent on education
1	6.07777778	4.63
2	5.46666667	5.77666667
3	29.2	8.47
4	12.3	5.27666667
5	26.6	5.29

K Means NCluster=6

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	2	3	0
2	6		
3	1		
4	1		
5	9		
6	1		

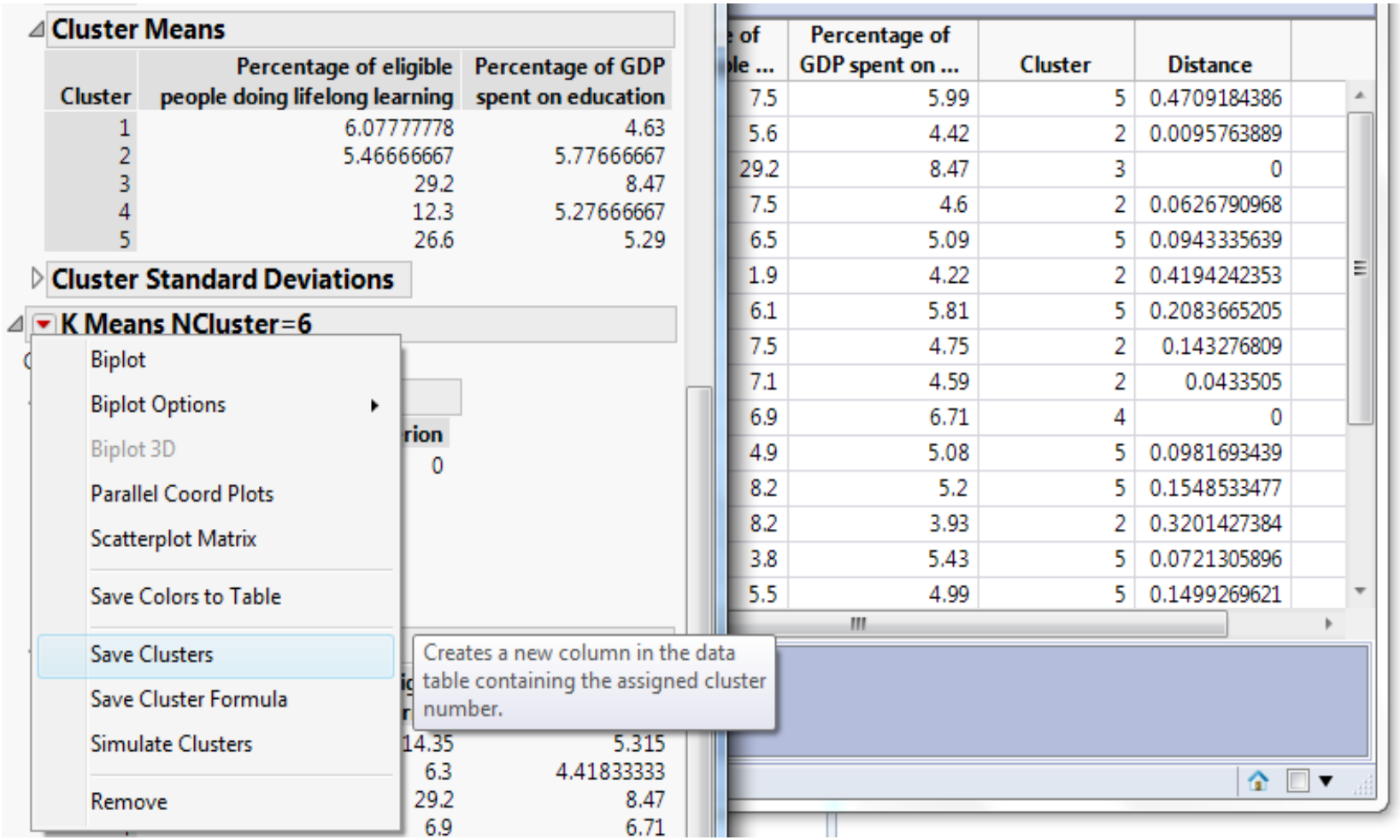
Cluster Means

Cluster	Percentage of eligible people doing lifelong learning	Percentage of GDP spent on education
1	14.35	5.315
2	6.3	4.41833333
3	29.2	8.47
4	6.9	6.71
5	5.66666667	5.36777778
6	26.6	5.29

Cluster Standard Deviations

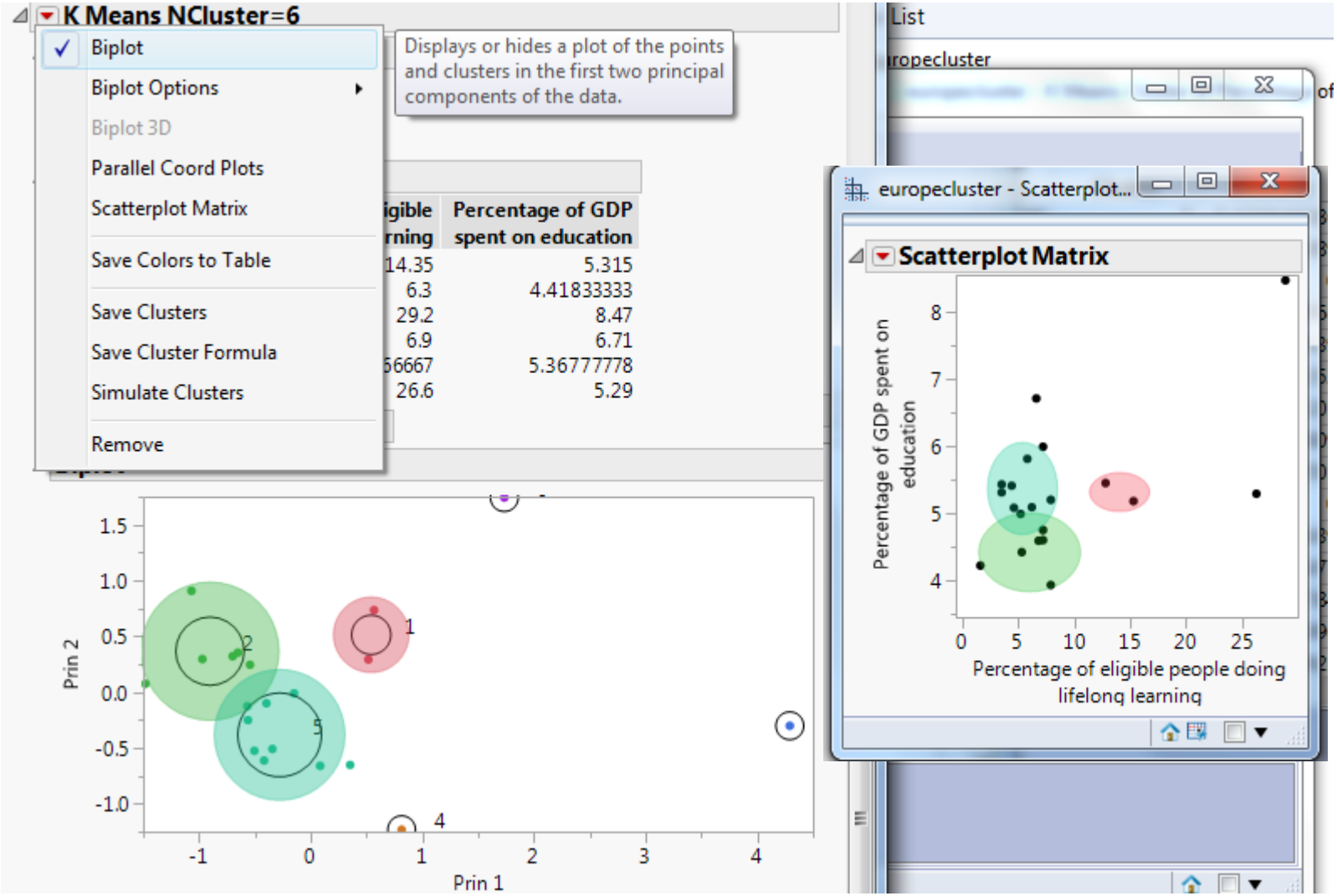
© 2018 NUS. All rights reserved.

Assign Cluster



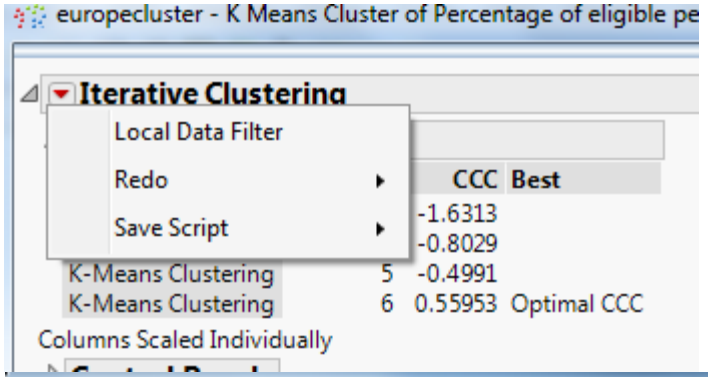
© 2018 NUS. All rights reserved.

Cluster Visualization



© 2018 NUS. All rights reserved.

Cluster Script



Script Window - JMP Pro

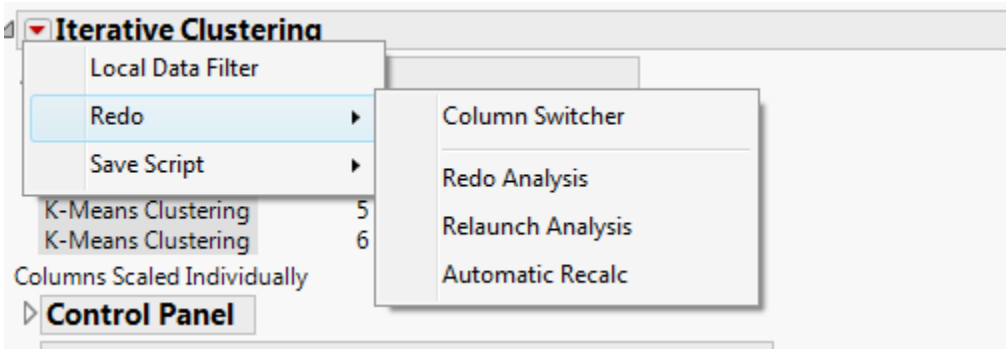
File Edit Tables DOE Analyze Graph Tools View Window Help

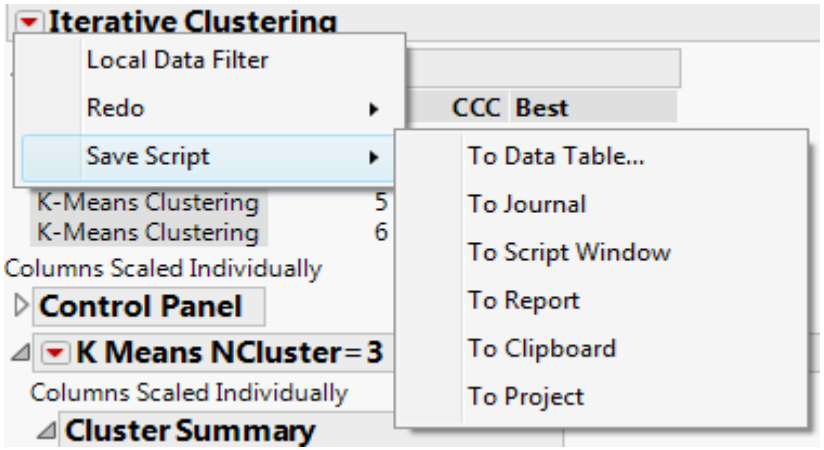
europecluster

```
K Means Cluster(
    Y(
        :Percentage of eligible people doing lifelong learning,
        :Percentage of GDP spent on education
    ),
    {Number of Clusters( 3 ), Name( "K-Means Clustering" ), Go},
    {Number of Clusters( 4 ), Name( "K-Means Clustering" ), Go},
    {Number of Clusters( 5 ), Name( "K-Means Clustering" ), Go},
    {Number of Clusters( 6 ), Name( "K-Means Clustering" ),
    Go( Parallel Coord Plots, Biplot( 1 ) )},
    SendToReport( Dispatch( {}, "Control Panel", OutlineBox, {Close( 1 )} ) )
);
```


© 2018 NUS. All rights reserved.

Additional Necessary Functionality





© 2018 NUS. All rights reserved.

Cluster Analysis Using R

Data Set: cluster_children

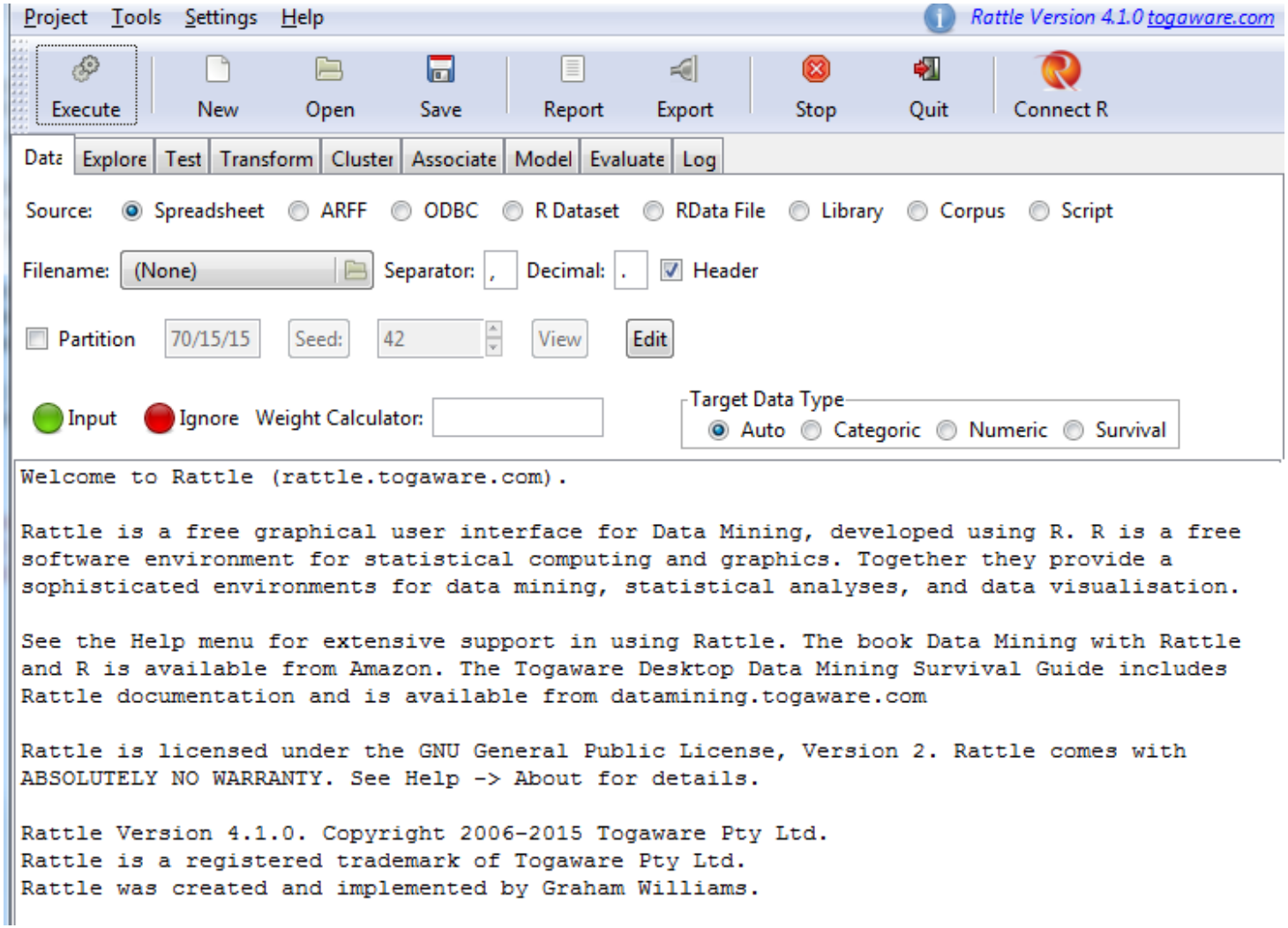
© 2018 NUS. All rights reserved.

Launch R Studio

```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(rattle)
Rattle: A free graphical interface for data mining with R.
Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
>
```

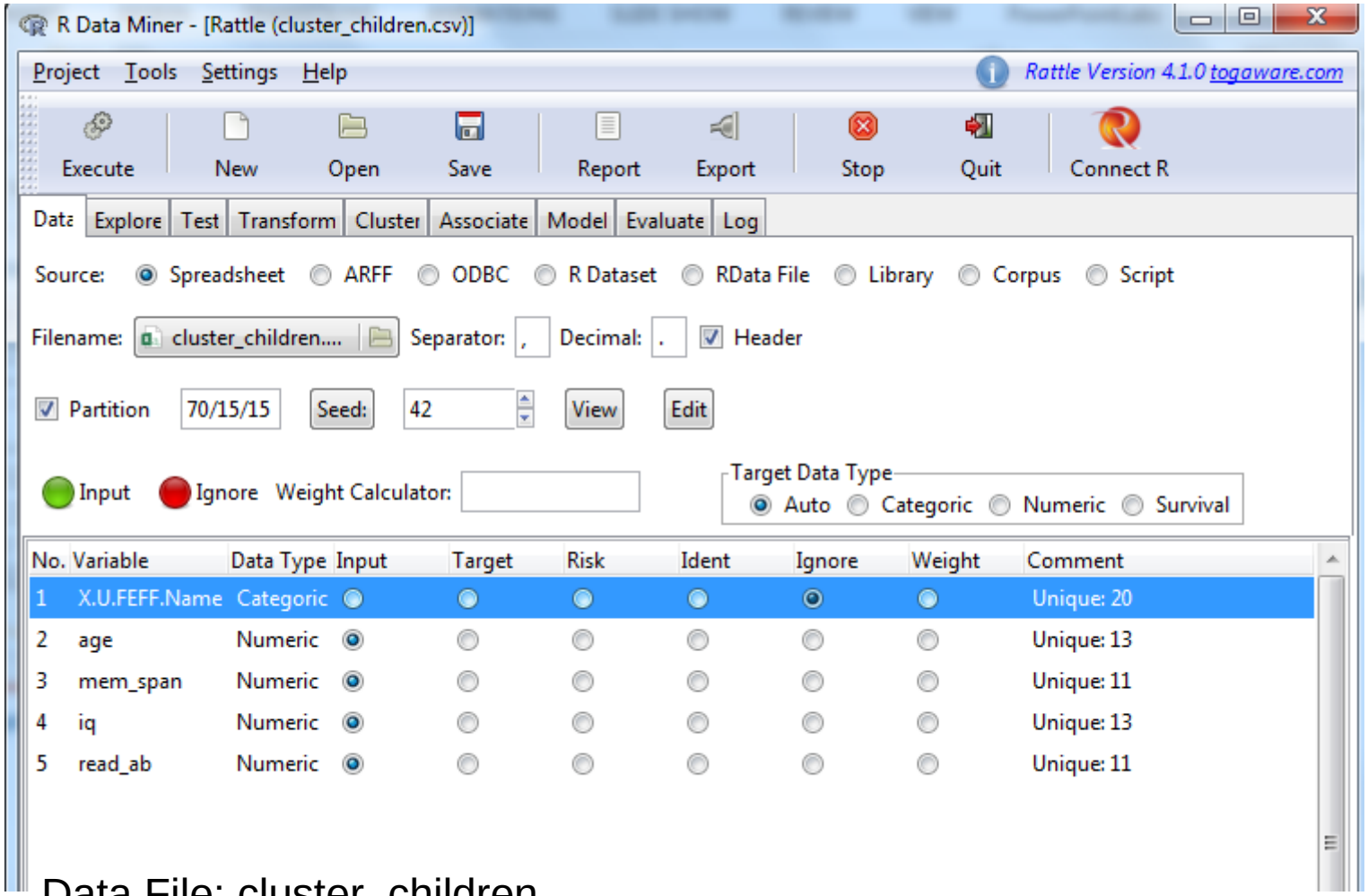

© 2018 NUS. All rights reserved.

Launch Rattle



© 2018 NUS. All rights reserved.

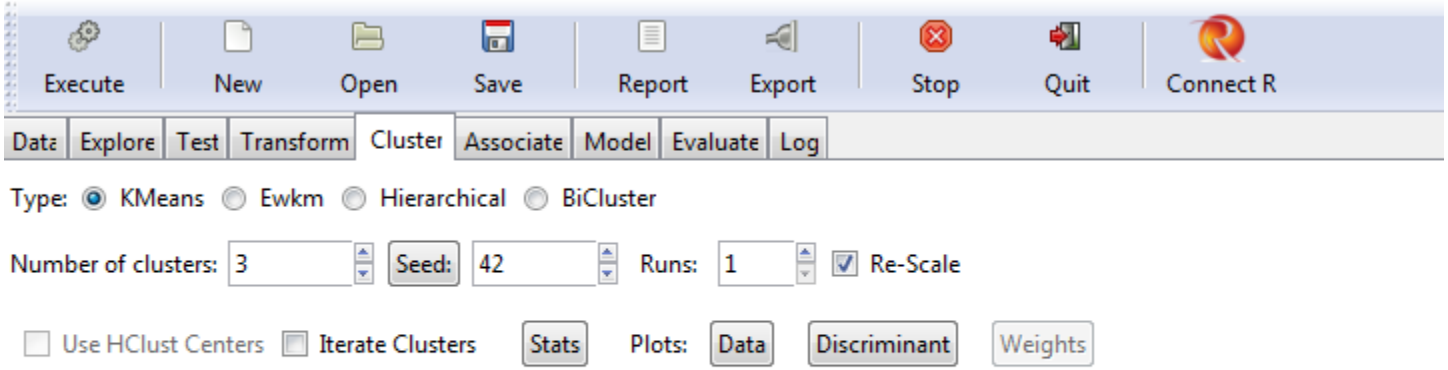
Select Appropriate Data Set & Variable & Execute



Data File: cluster_children

© 2018 NUS. All rights reserved.

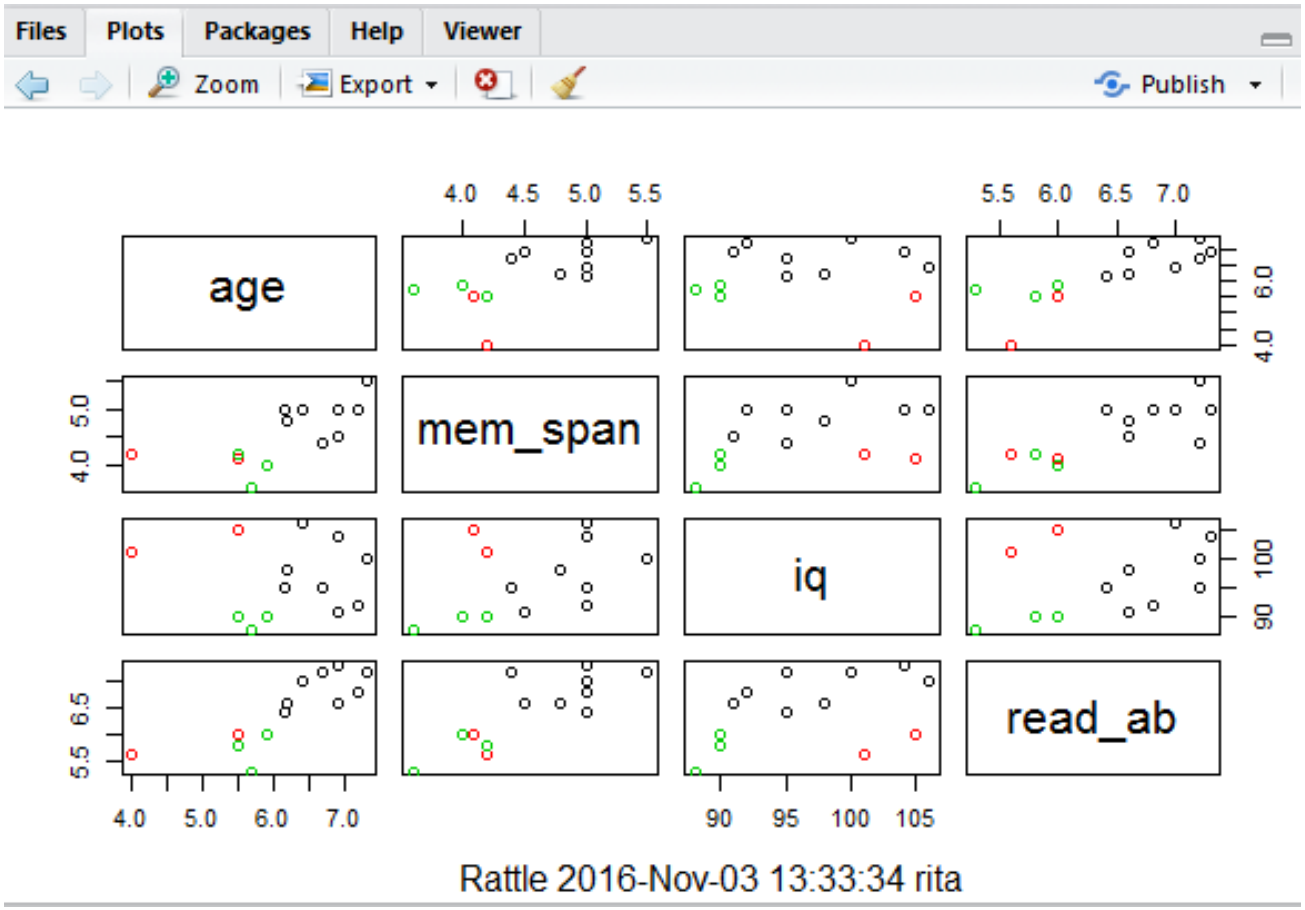
K Means Cluster – Rattle Output




```
Cluster sizes:

[1] "9 2 3"

Data means:

      age  mem_span        iq   read_ab
0.6850649 0.5413534 0.4880952 0.6000000

Cluster centers:

        age  mem_span         iq   read_ab
1 0.8434343 0.7192982 0.54938272 0.8111111
2 0.2272727 0.2894737 0.83333333 0.2500000
3 0.5151515 0.1754386 0.07407407 0.2000000

Within cluster sum of squares:

[1] 1.3352486 0.1493822 0.1322849

Time taken: 0.00 secs
```

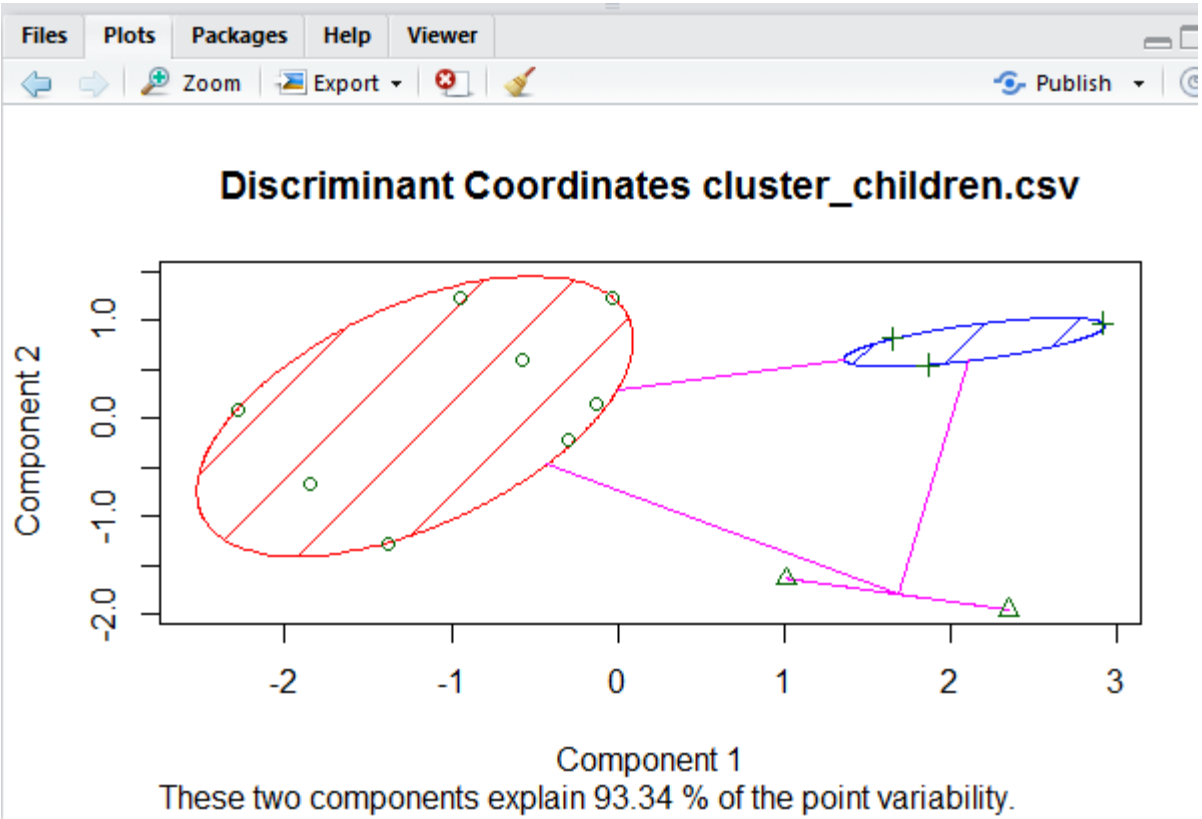

© 2018 NUS. All rights reserved.

K Means Cluster – Plots



© 2018 NUS. All rights reserved.

K Means Cluster – Plots



© 2018 NUS. All rights reserved.

Cluster Analysis Using SPSS Modeler

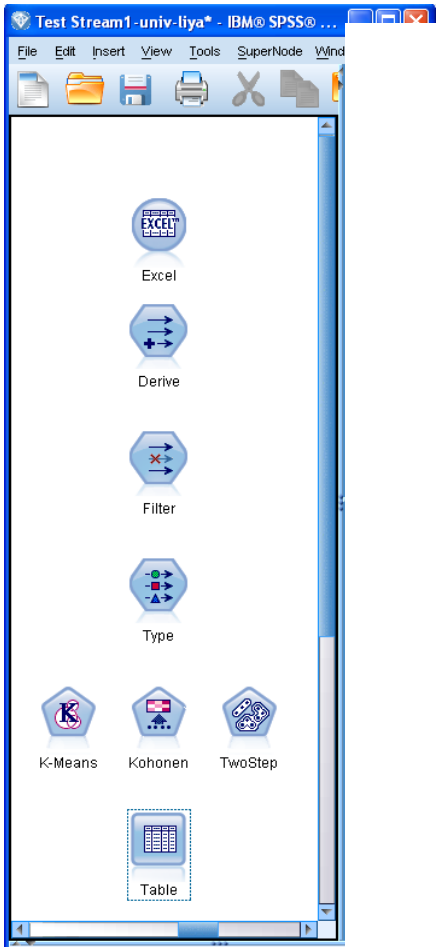
Data Set: UScolleges

© 2018 NUS. All rights reserved.

Clustering using SPSS Modeler

Nodes involved have following operations:

- **Source**: Access dataset e.g. Excel
- **Derive**: Generate field based on existing ones
- **Filter**: Rename or remove field for next step
- **Type**: Specify the type (continuous/ nominal/ ordinal) of fields
- **Modeling** – Clustering algorithms : K-means, Kohonen, TwoStep etc.
- **Table**: Final result capture & validation



© 2018 NUS. All rights reserved.

Data File: US Colleges

- Dataset: American College and University
  - 1302 records with 23 fields:
    - College name, State, Public/private,
      Math SAT, Verbal SAT, Tuition fee,
      No. of applicants received,
      No. of applicants accepted, …

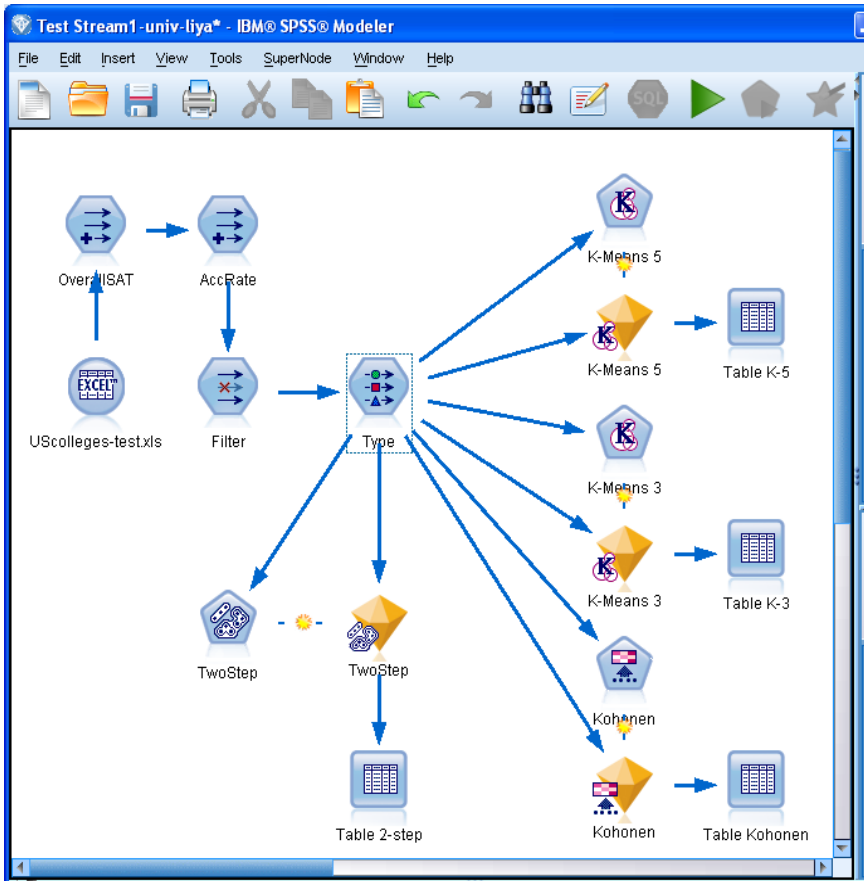
(source: G. Shmueli , N. R. Patel, P. C. Bruce, "Data Mining for Business Intelligence", Wiley, 2010)

- Application
  - An education agent to recommend US universities to applicants (outside of US) given his/her SAT scores and financial capability
  - The company needs an optimal overall plan for resource allocation to cope with potential demands to different groups of colleges, such as private, public, top school, …

© 2018 NUS. All rights reserved.

Case Study: US Colleges (cont.)

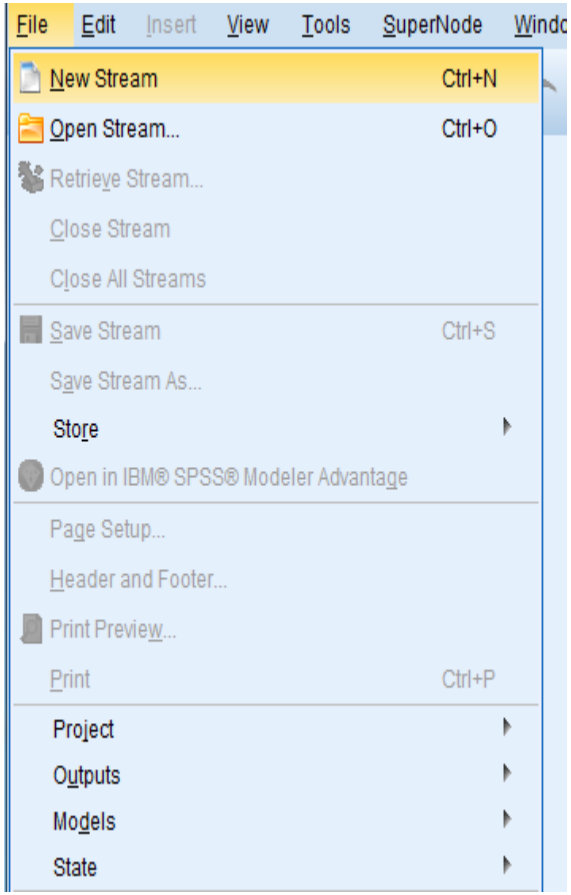
Demo Stream :
One needs to create a stream to do any analysis

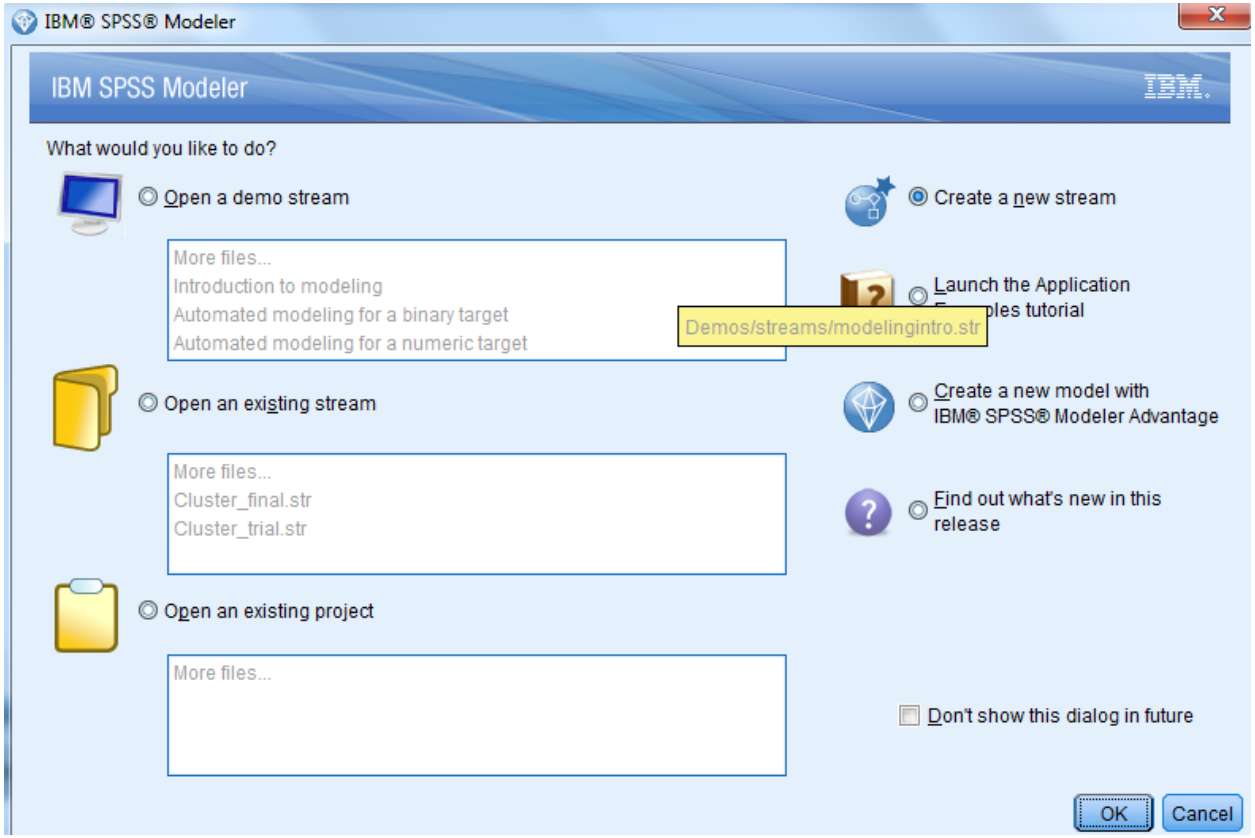


© 2018 NUS. All rights reserved.

Creation of Stream

Launch SPSS Modeler – start a new stream

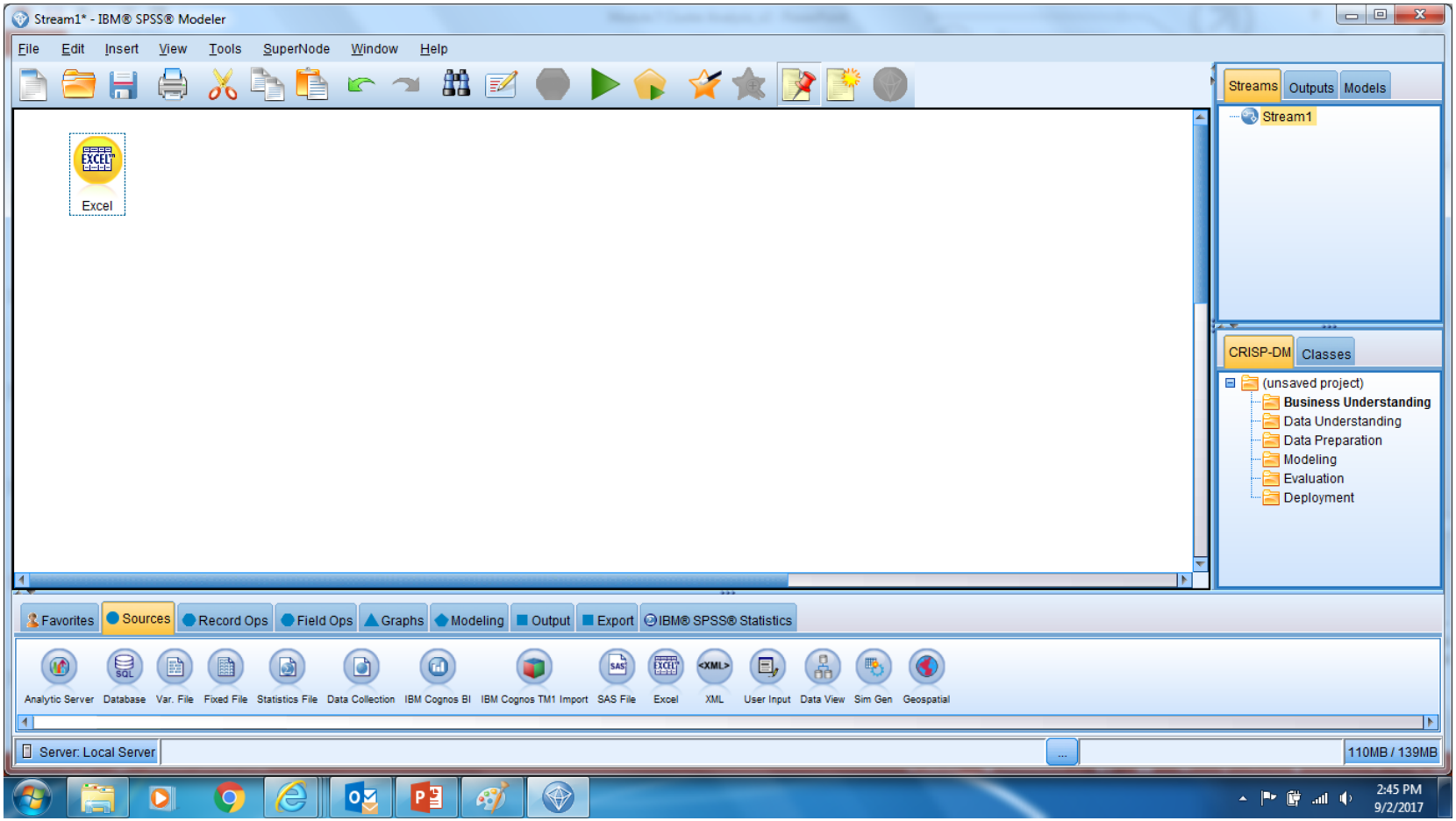




© 2018 NUS. All rights reserved.

Creation of Stream

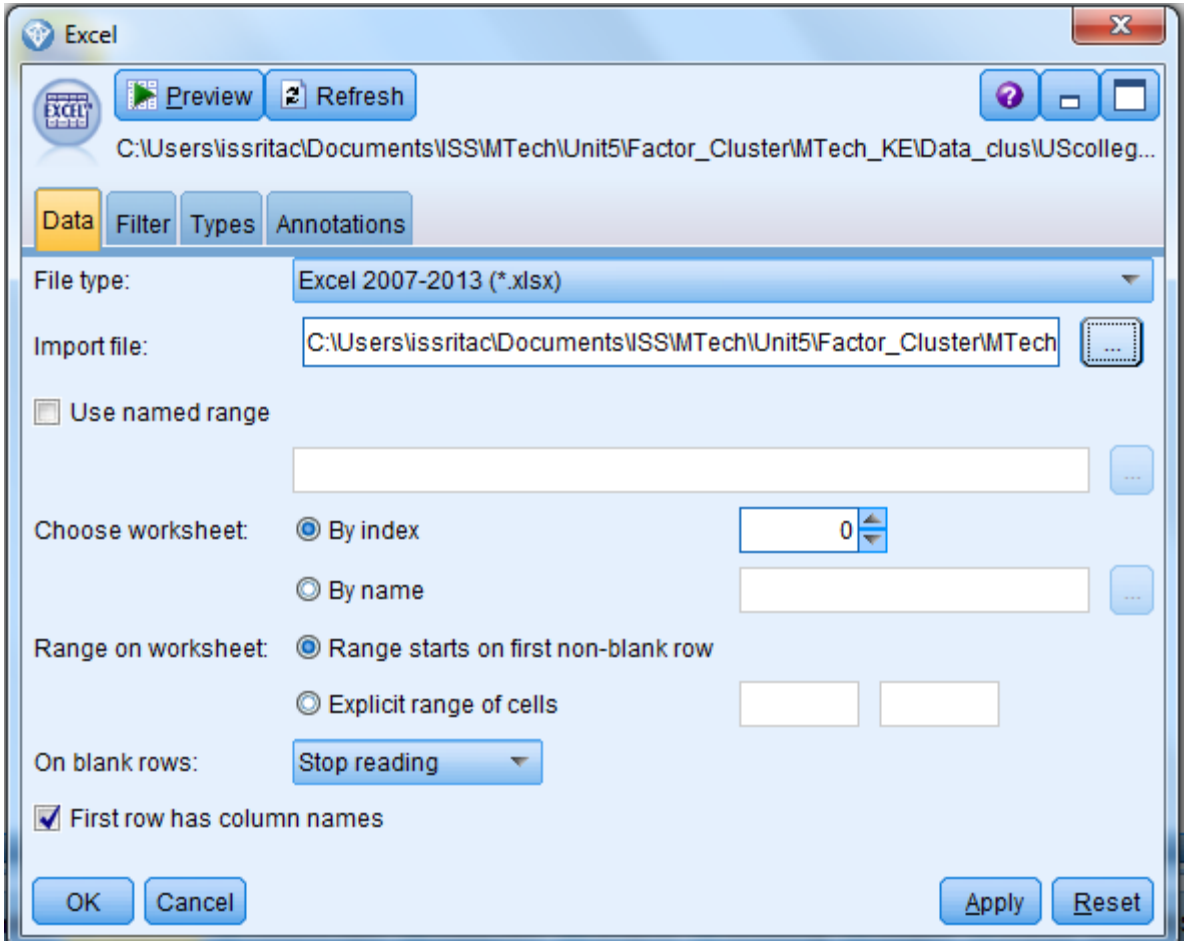
Drag & drop necessary icons on the Stream screen and link them up



© 2018 NUS. All rights reserved.

Creation of Stream

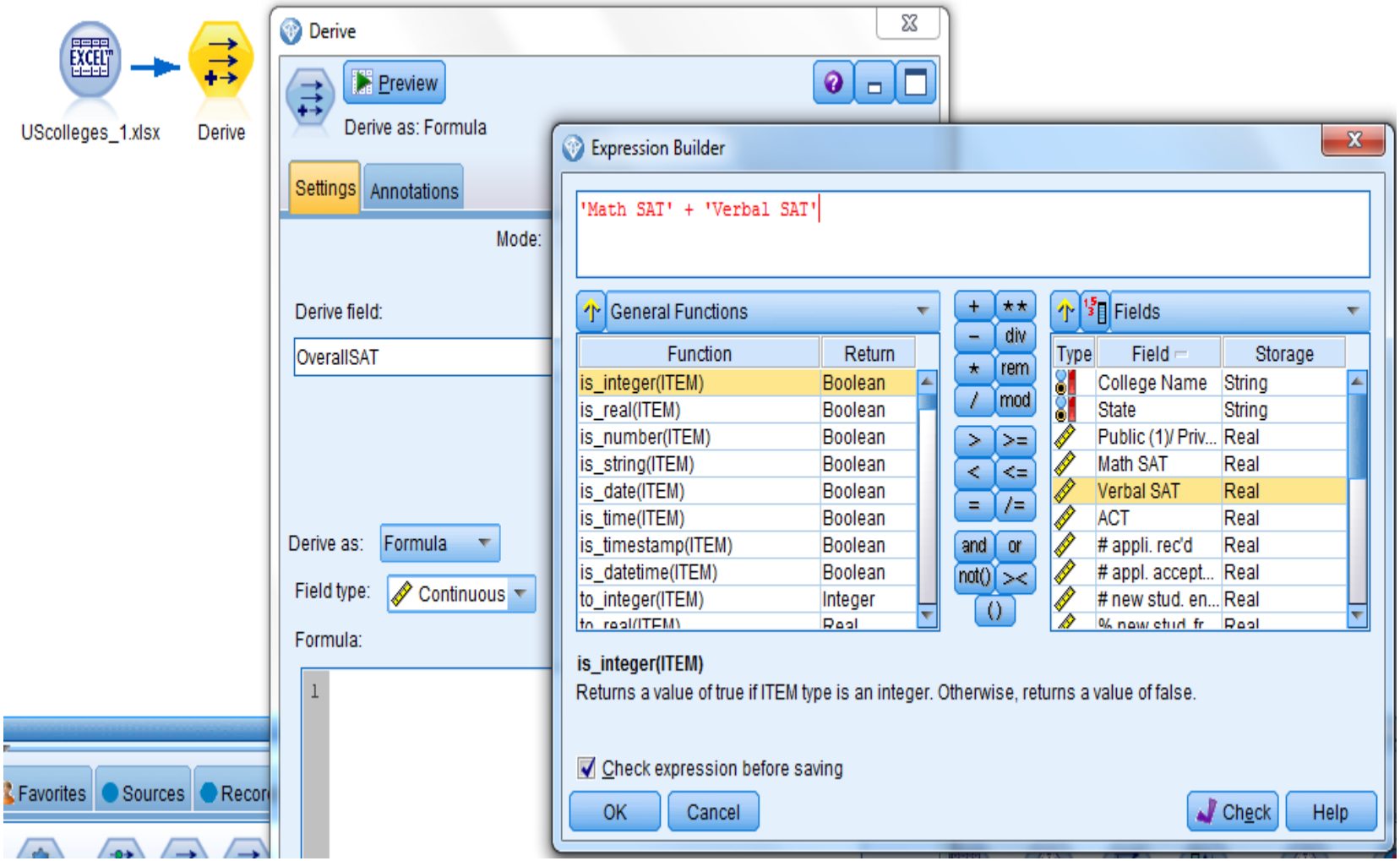
Import file and start various operations



© 2018 NUS. All rights reserved.

US Colleges: *Data Preparation*

New variable creation



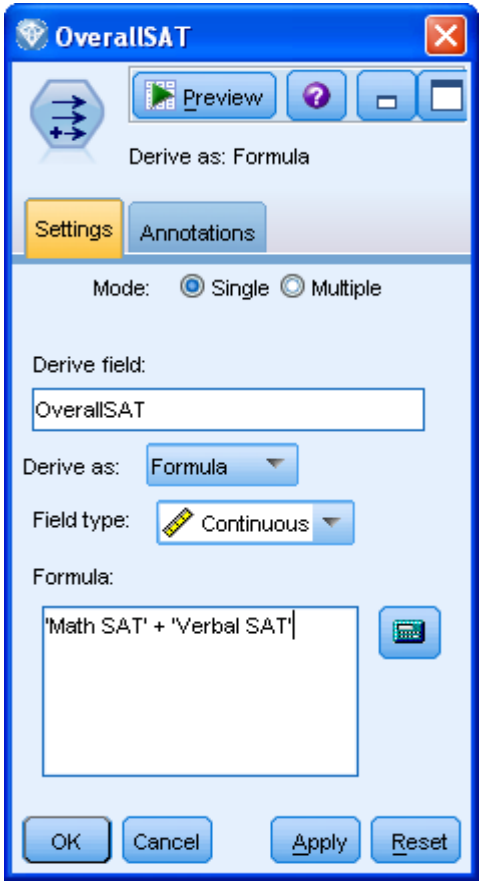
© 2018 NUS. All rights reserved.

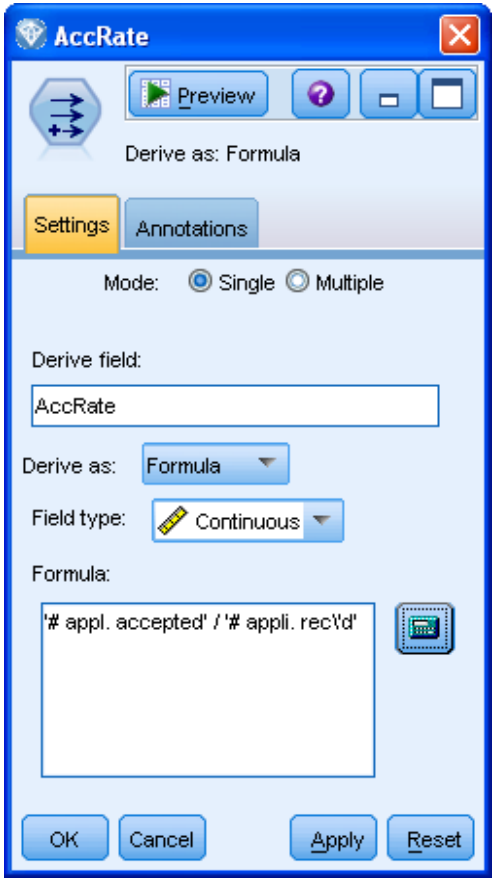
US Colleges: *Data Preparation*

Data source: Uscolleges_1.xls

Derived fields (using derive nodes)

- Overall SAT = Math SAT + Verbal SAT
- AccRate = appl. accepted / appl. rec\'d





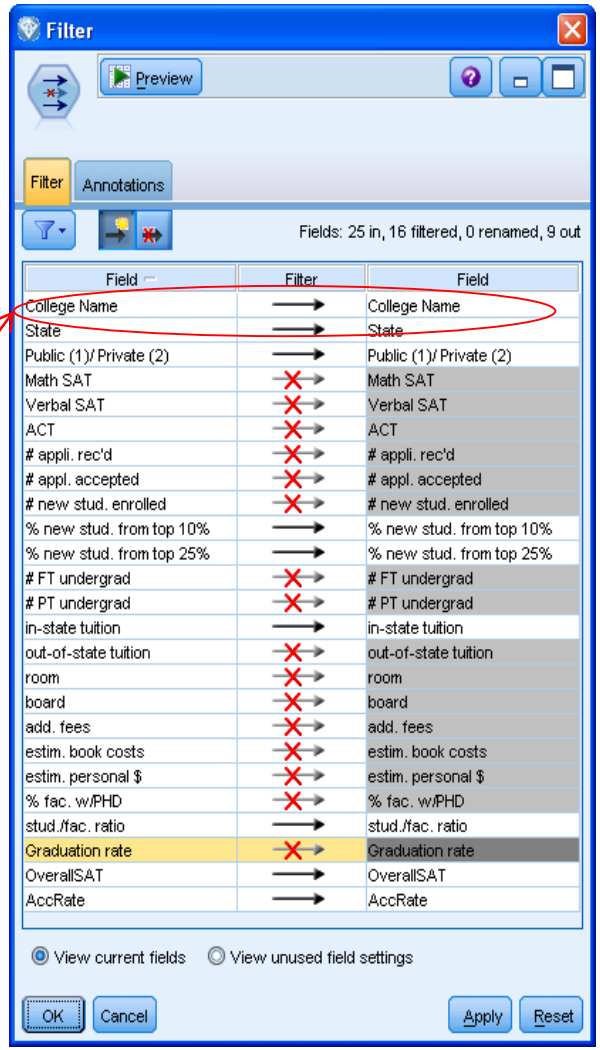
© 2018 NUS. All rights reserved.

US Colleges: *Data Preparation* (cont.)

Fields selected for clustering (using filter node)

- Public/private
- % new stud. from top 10%
- % new stud. from top 25%
- In-state tuition
- Stud./fac. ratio
- **OverallSAT**
- **AccRate**

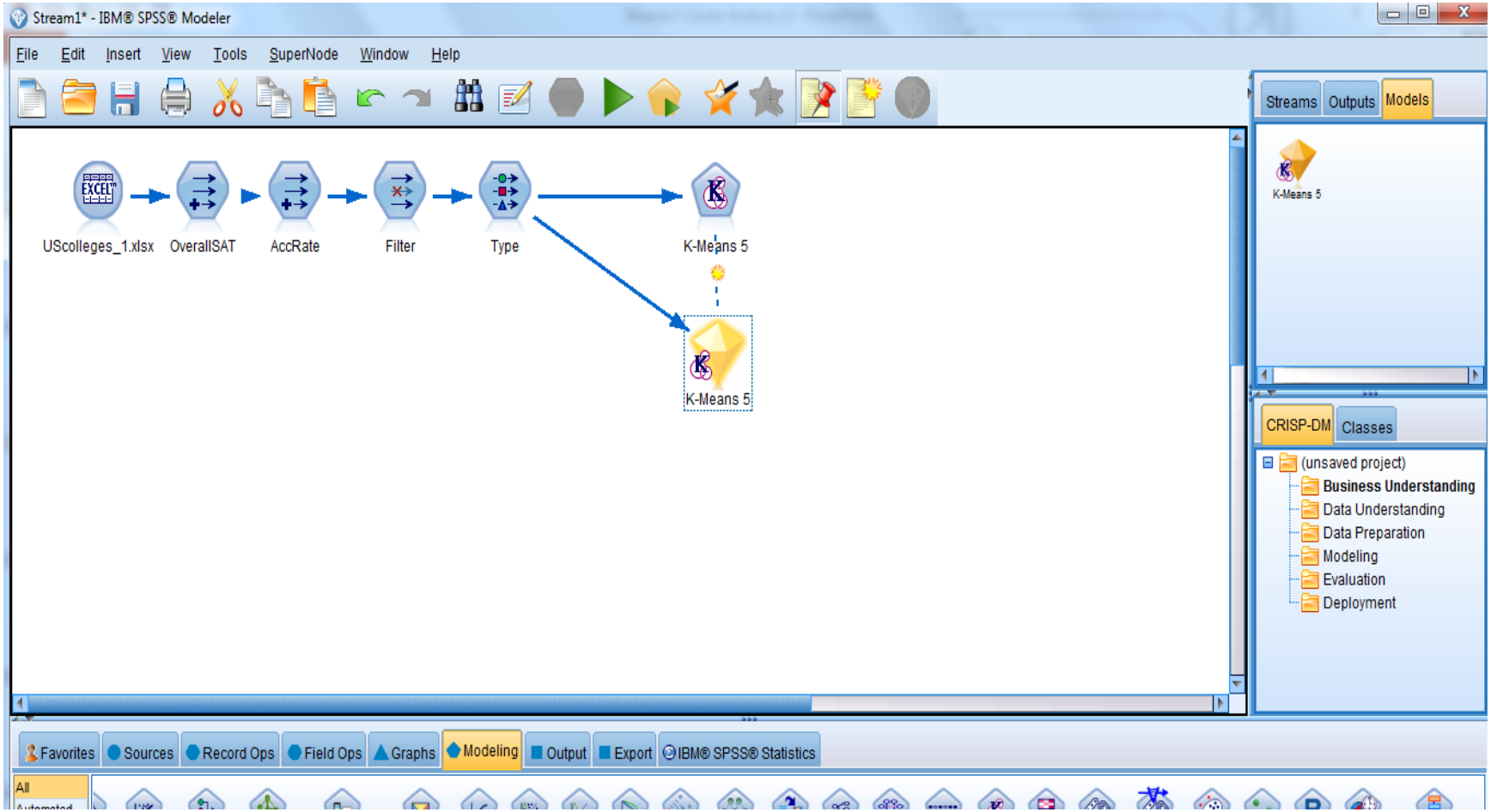
Not for clustering, but keep for result evaluation



© 2018 NUS. All rights reserved.

US Colleges: *Clustering Method Selection*

Select Modeling Method



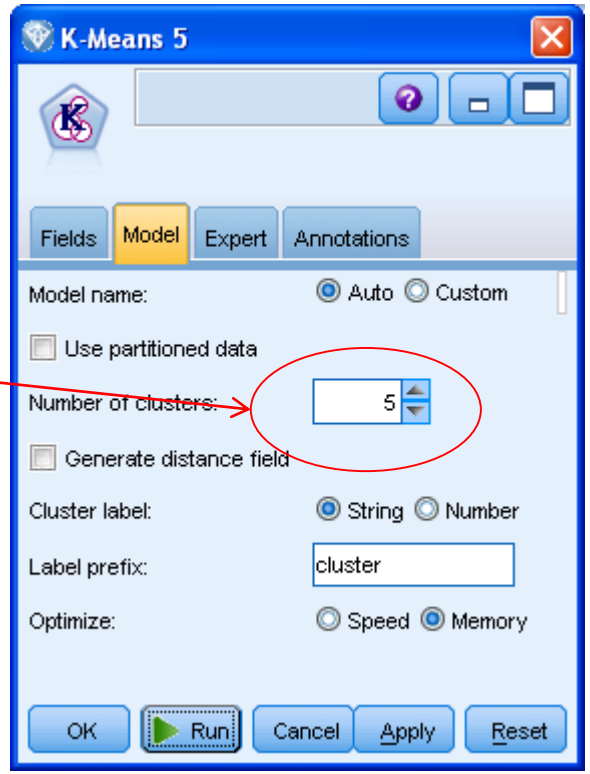
© 2018 NUS. All rights reserved.

US Colleges: *k-Means*

For demo purpose, we have two *k*-Means nodes

- k = 5
- k = 3

We first run with k = 5



© 2018 NUS. All rights reserved.

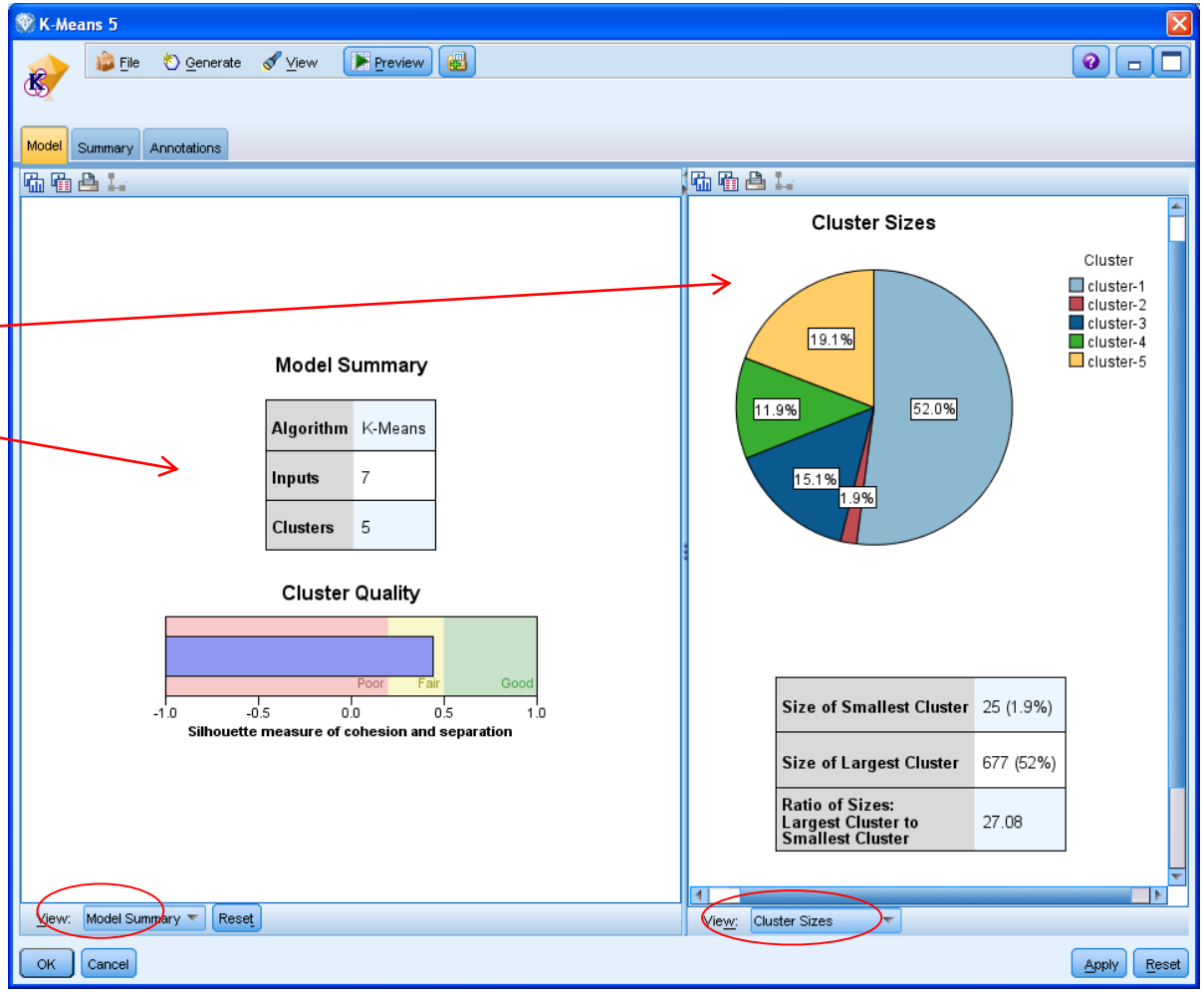
US Colleges: *k-Means* (k = 5)

Double click the nugget

- "K-means 5"

Cluster size

Model summary



© 2018 NUS. All rights reserved.

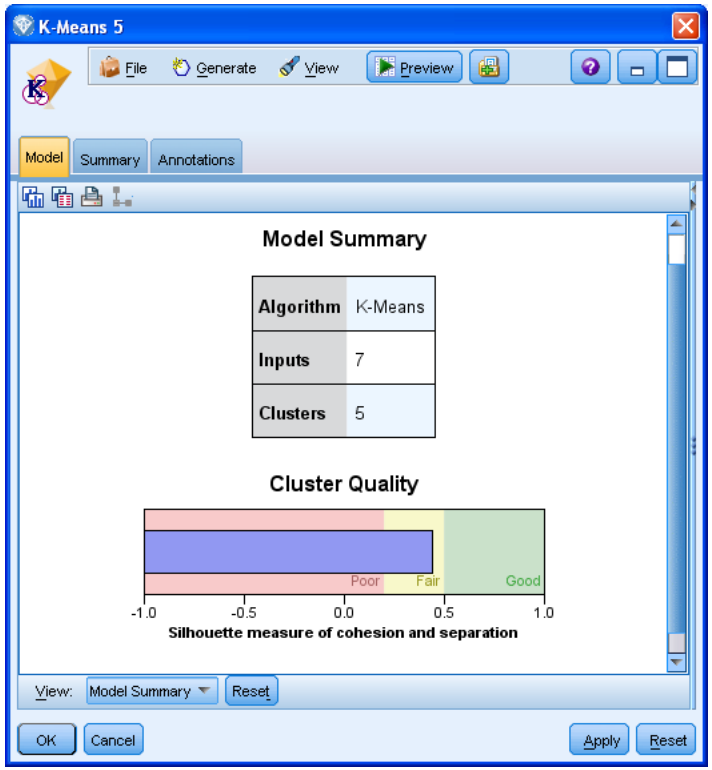
Cluster Quality

Silhouette coefficient is a measure of both cohesion and separation. For each element in a cluster, calculate :

- the average distance to all other elements in its cluster, $\boldsymbol{d}_{ave}$ &
- the average distance to all elements in each of the other clusters, $\boldsymbol{D}_{ave}$

Silhouette =

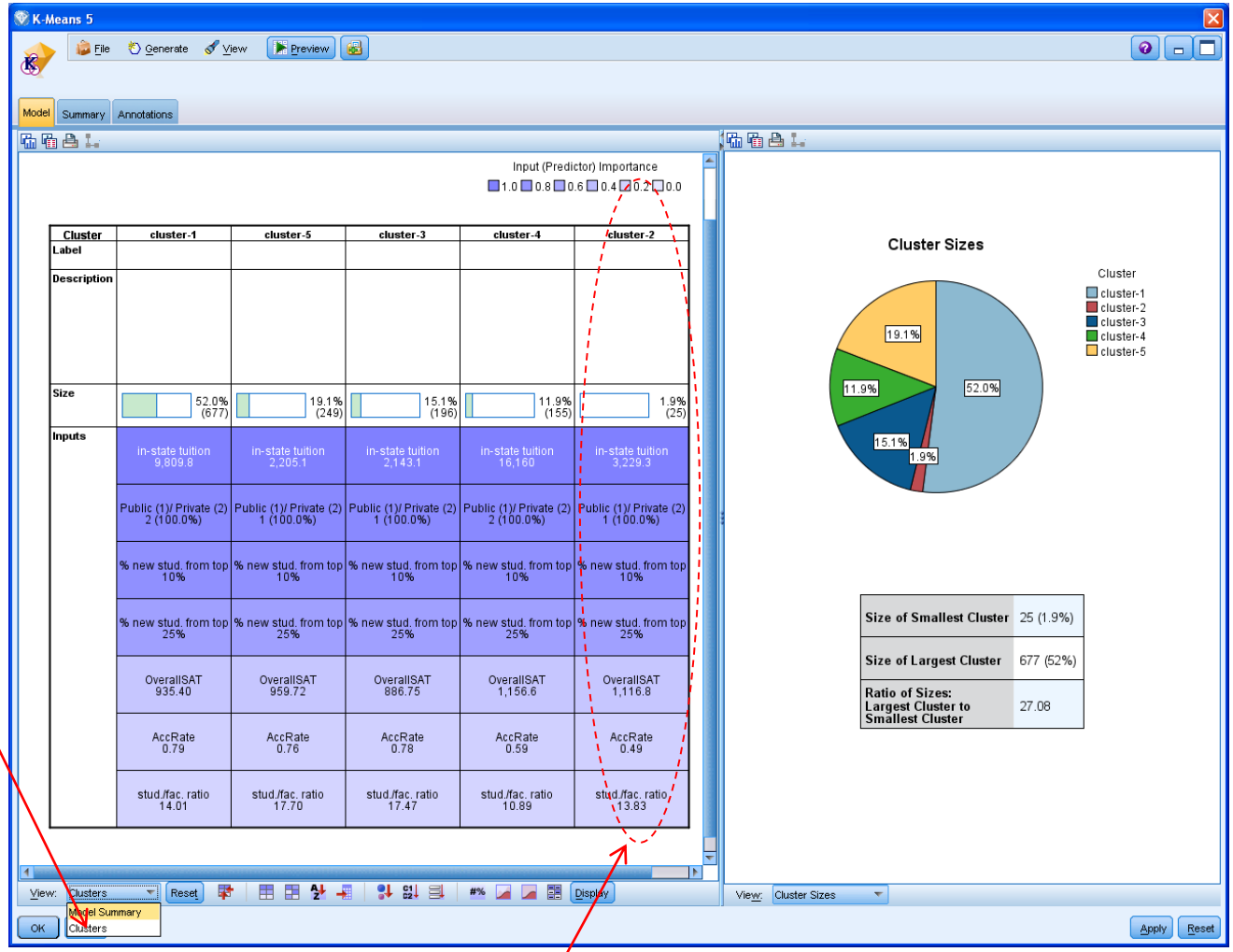
- $(\boldsymbol{D}_{ave} - \boldsymbol{d}_{ave}) / \text{Max}(\boldsymbol{D}_{ave}, \boldsymbol{d}_{ave})$



© 2018 NUS. All rights reserved.

US Colleges: *k-Means* (k = 5) (cont.)

Choose "cluster"

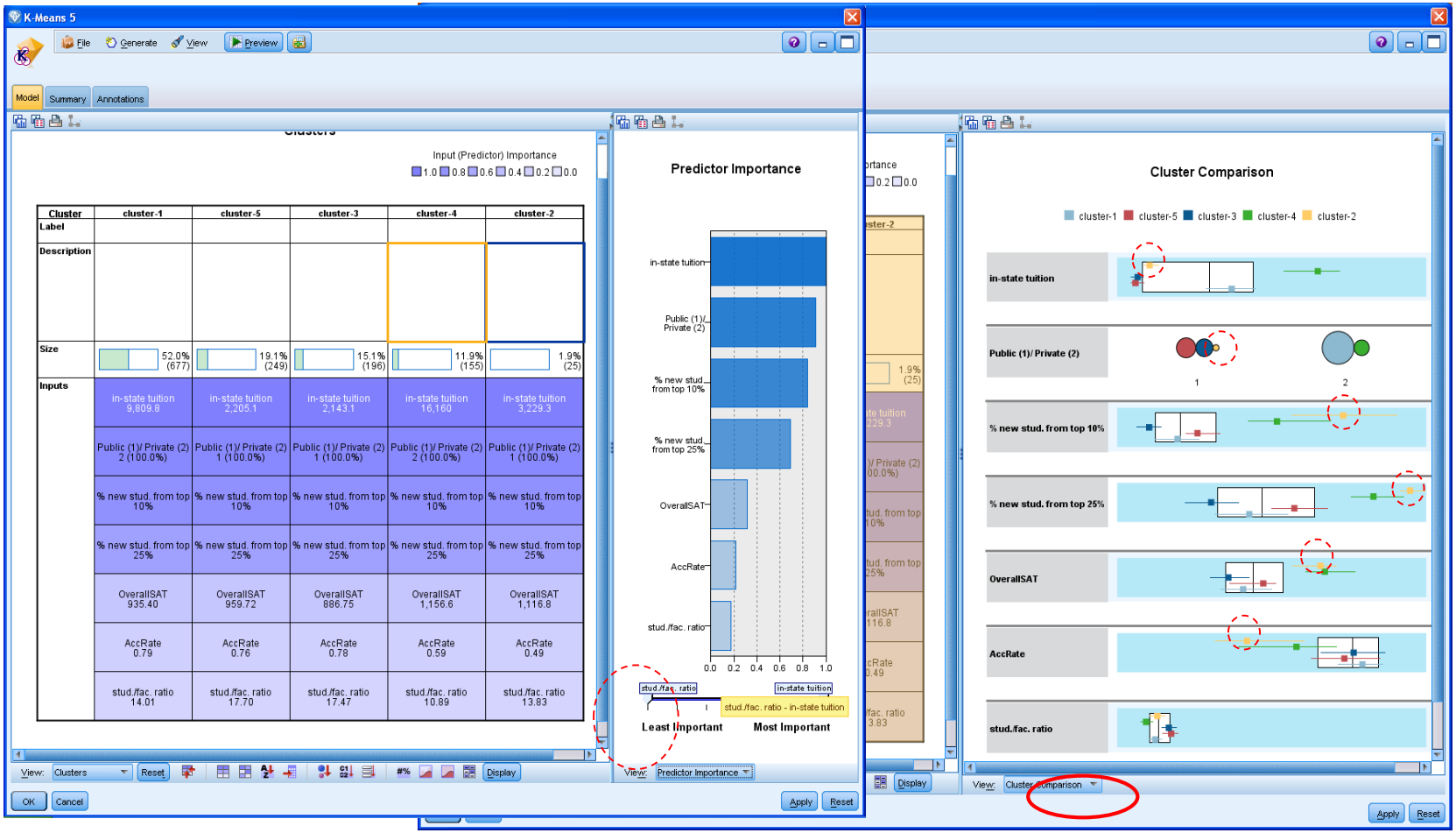


The smallest cluster

© 2018 NUS. All rights reserved.

US Colleges: *k-Means* ($k = 5$) (cont.)

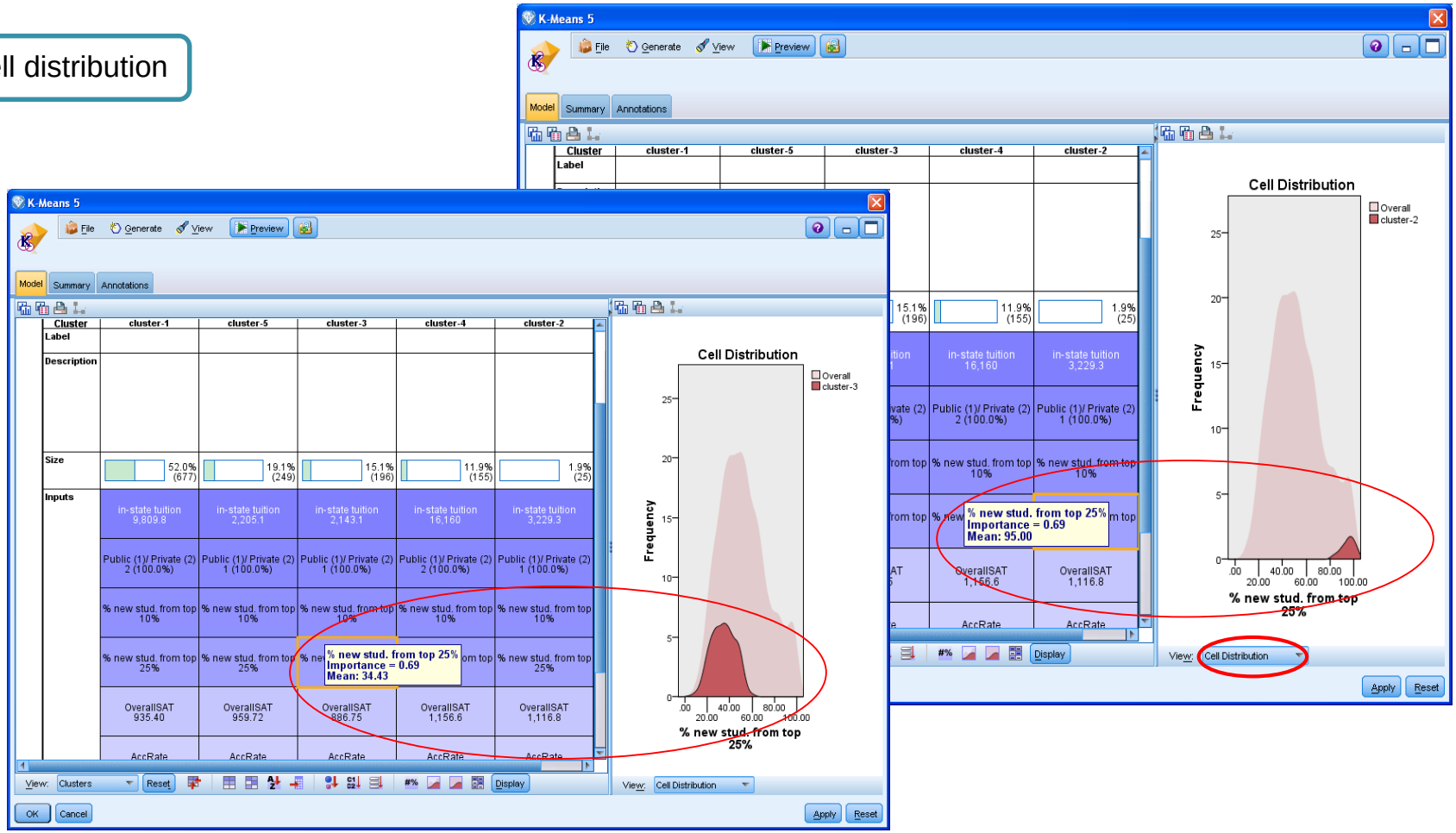
Cluster comparison (cluster-2, the smallest cluster)



© 2018 NUS. All rights reserved.

US Colleges: *k-Means* (k = 5) (cont.)

Cell distribution

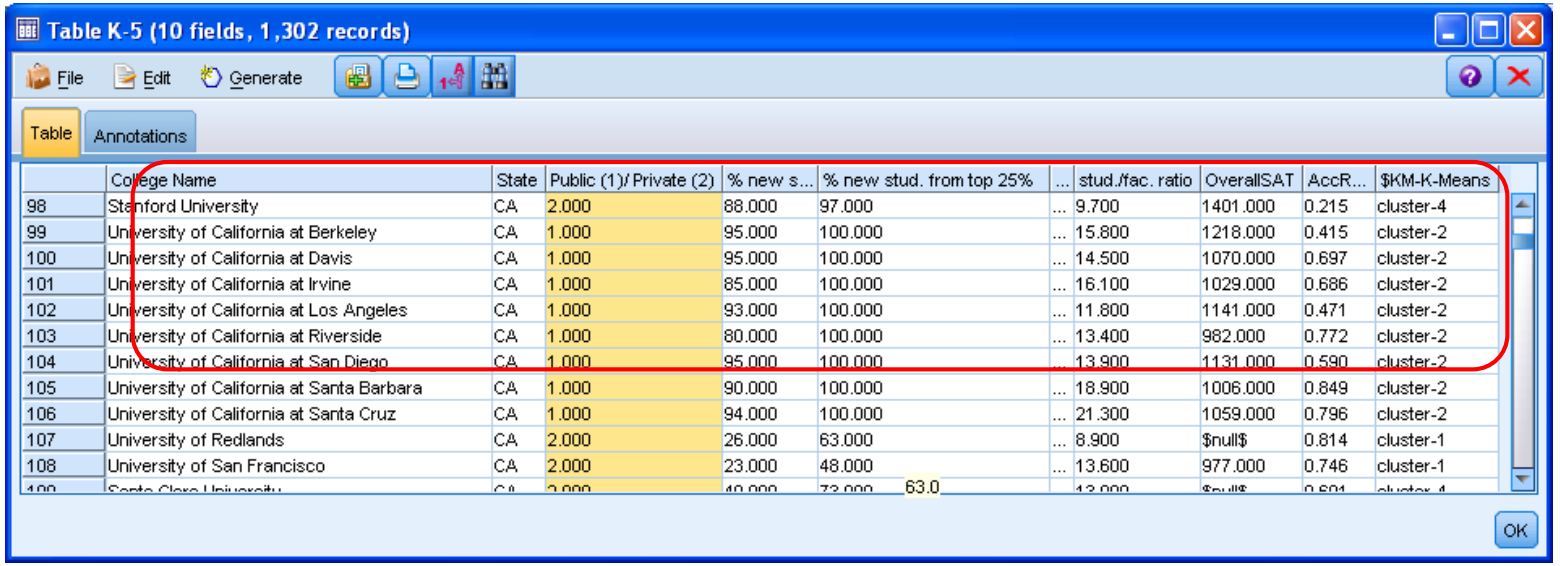


© 2018 NUS. All rights reserved.

US Colleges: *k-Means* ($k = 5$) (cont.)

Run the Table K-5

- In the cluster-2
  - We found some famous top public universities



	College Name	State	Public (1)/ Private (2)	% new s...	% new stud. from top 25%	...	stud./fac. ratio	OverallSAT	AccR...	$KM-K-Means
98	Stanford University	CA	2.000	88.000	97.000	...	9.700	1401.000	0.215	cluster-4
99	University of California at Berkeley	CA	1.000	95.000	100.000	...	15.800	1218.000	0.415	cluster-2
100	University of California at Davis	CA	1.000	95.000	100.000	...	14.500	1070.000	0.697	cluster-2
101	University of California at Irvine	CA	1.000	85.000	100.000	...	16.100	1029.000	0.686	cluster-2
102	University of California at Los Angeles	CA	1.000	93.000	100.000	...	11.800	1141.000	0.471	cluster-2
103	University of California at Riverside	CA	1.000	80.000	100.000	...	13.400	982.000	0.772	cluster-2
104	University of California at San Diego	CA	1.000	95.000	100.000	...	13.900	1131.000	0.590	cluster-2
105	University of California at Santa Barbara	CA	1.000	90.000	100.000	...	18.900	1006.000	0.849	cluster-2
106	University of California at Santa Cruz	CA	1.000	94.000	100.000	...	21.300	1059.000	0.796	cluster-2
107	University of Redlands	CA	2.000	26.000	63.000	...	8.900	$null$	0.814	cluster-1
108	University of San Francisco	CA	2.000	23.000	48.000	...	13.600	977.000	0.746	cluster-1
109	Santa Clara University	CA	2.000	40.000	73.000		13.000	$null$	0.691	cluster-4

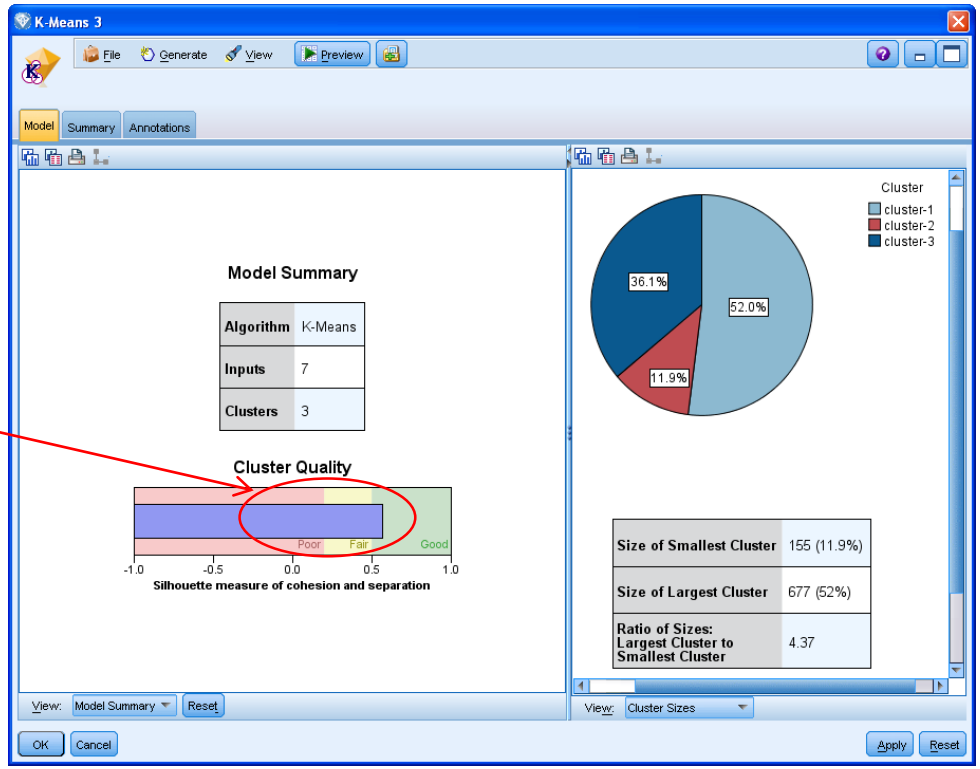
© 2018 NUS. All rights reserved.

US Colleges: *k-Means* ($k = 3$)

Now run *k*-means with $k = 3$

Double click the nugget "K-means 3"

- The Silhouette measure is higher than that of $k = 5$
- Is the clustering more meaningful than that of $k = 5$?



© 2018 NUS. All rights reserved.

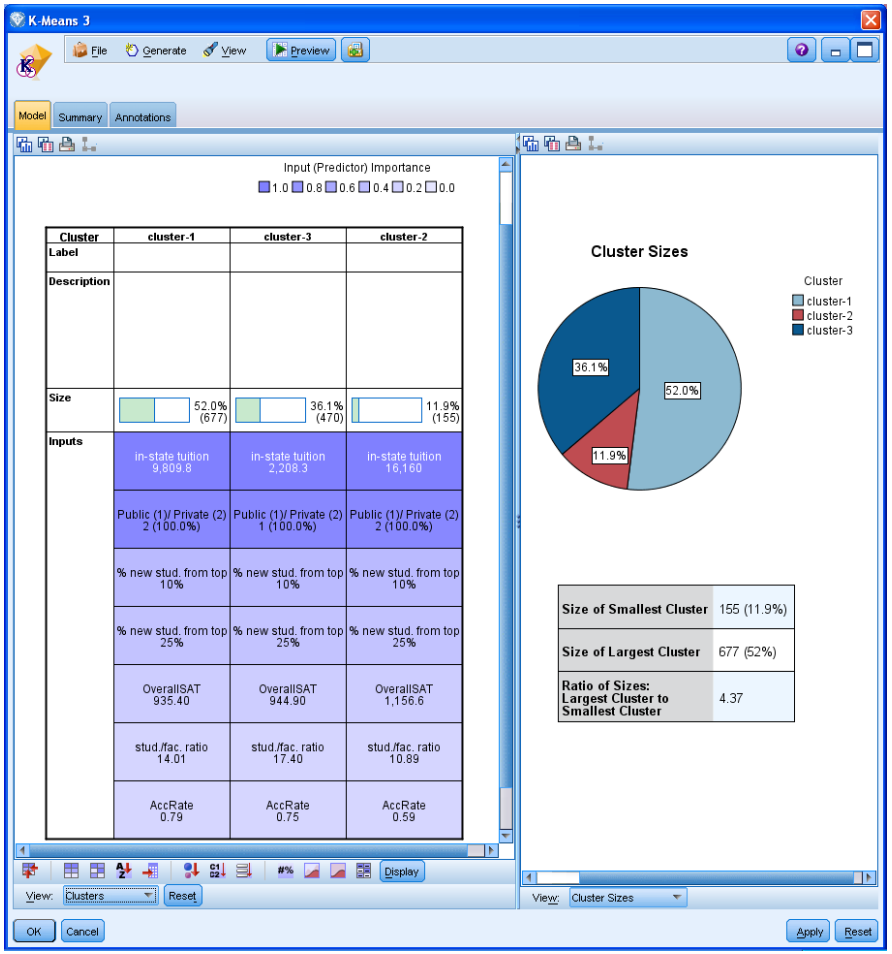
US Colleges: *k-Means* (k = 3) (cont.)

Result analysis

- Cluster 1: private schools
  - lowest overall SAT & highest acceptance rate
- Cluster 2: private schools
  - highest overall SAT & lowest acceptance rate
- Cluster 3:
  - includes all public schools

Compare with k = 5

- Three previous clusters of
  - public schools are now in
  - one cluster
- **Which one is more useful?**



© 2018 NUS. All rights reserved.

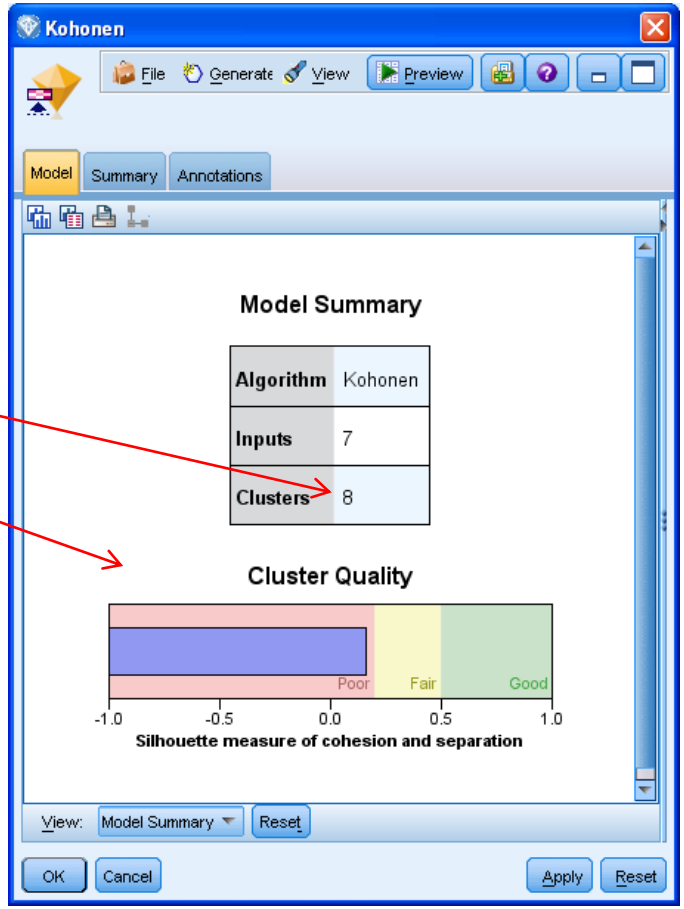
US Colleges: *Kohonen clustering*

Run Kohonen clustering

Open the nugget "Kohonen"

- 8 clusters formed
- The cluster quality seems poorer than that of *k*-means ?

Further analysis is needed



© 2018 NUS. All rights reserved.

TwoStep Clustering Mechanism

Step 1:

- Pre-clustering to make little clusters

Step 2:

- Clustering of pre-clusters

Twostep clustering algorithm offers function to exclude outliers

© 2018 NUS. All rights reserved.

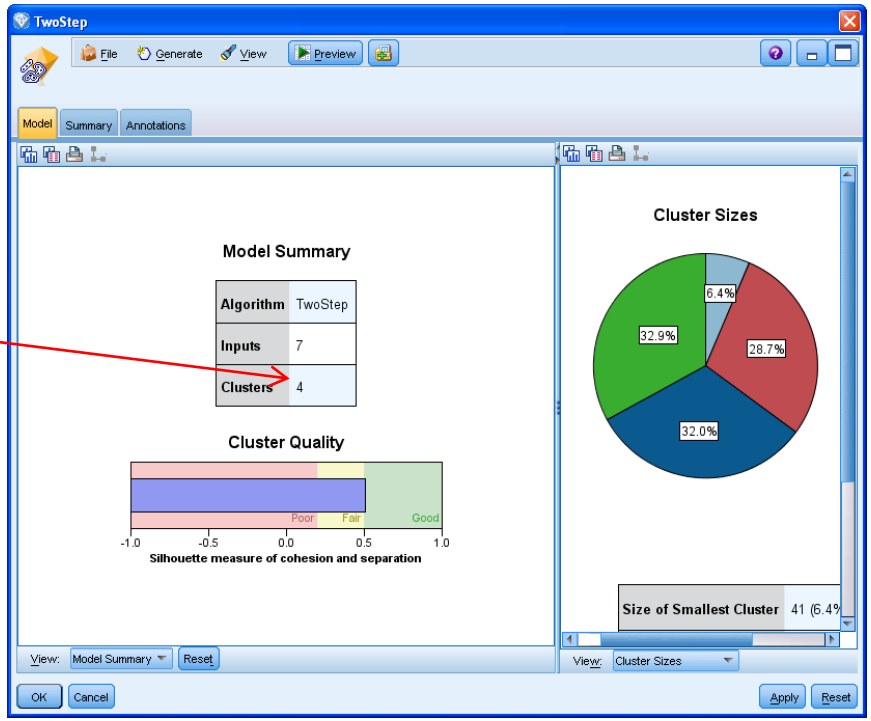
US Colleges: *TwoStep clustering*

Run TwoStep clustering

Open the nugget "TwoStep"

- 4 clusters formed(using Euclidean distance)

Further analysis is needed

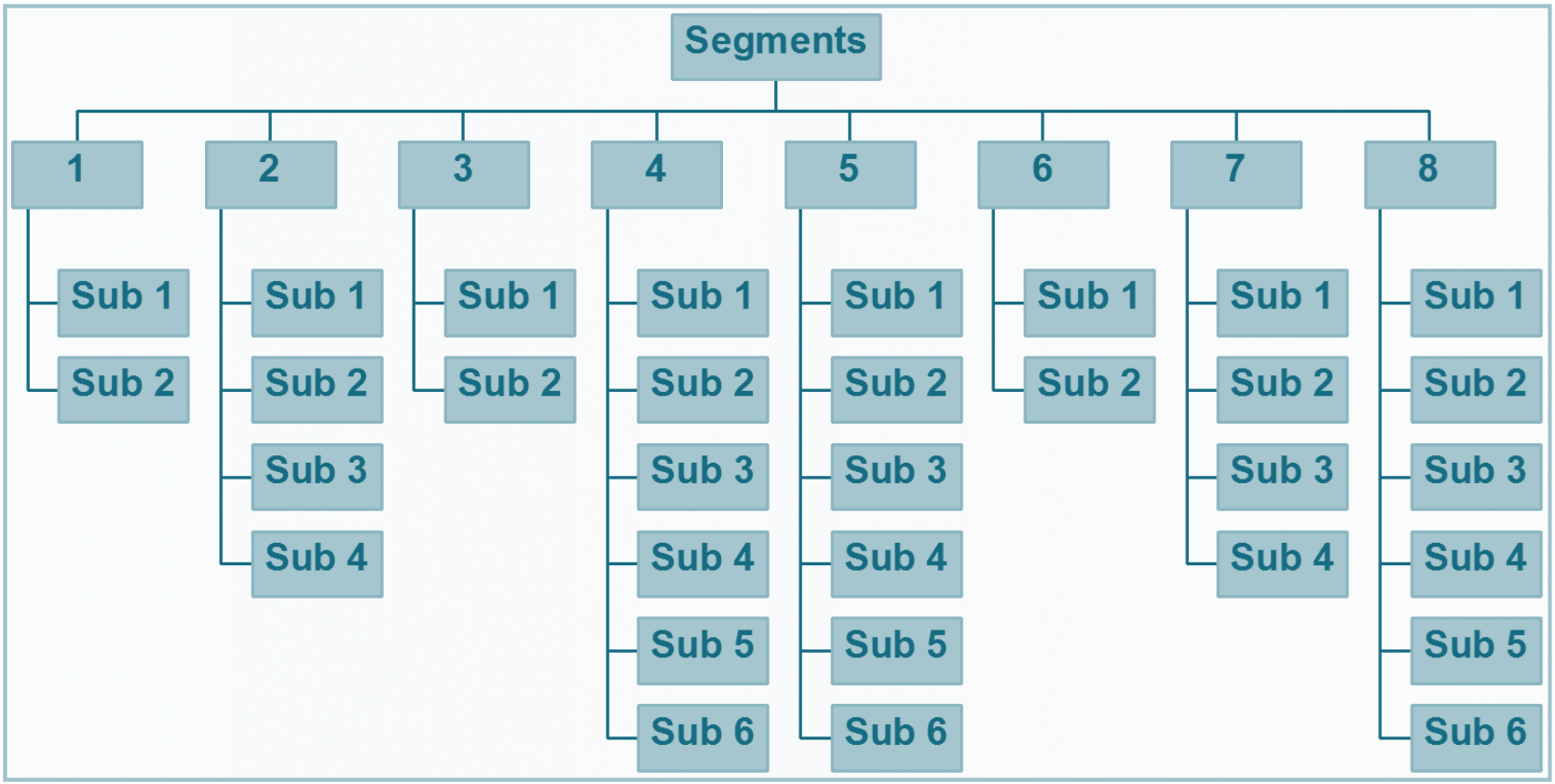


© 2018 NUS. All rights reserved.

REAL LIFE EXAMPLE OF TWO STEP CLUSTERING

First Level Segmentation Variables : Revenue, Utilization, Credit Risk, Attrition Risk

Second Level Segmentation Variables : Please see slide 96



© 2018 NUS. All rights reserved.

REAL LIFE EXAMPLE OF TWO STEP CLUSTERING

Note how categorical variables are treated

Variable	Definition
Gender	Female vs. others.
Marital status	Single vs. others.
Age as of latest month	Older (>45 years) vs. others.
Home ownership	Buying or owning vs. others.
Occupation: "White-collar, educated jobs" vs. others.	White-collar educated jobs include: Snr Govt Officer, Bus Mngr./Exec., Cler/Sales Mngr., Health Prof., Build Prof, Bus. Prof., Social Prof., Tech/Skill, Misc Prof, Nurse/Med Tech, Build Tech, Para-Prof.
Spend	Monthly average of customer sales amount in the last 12 months.
Overseas transactions as % of total transactions. The following spending categories are used:	Proportion of average transactions in a particular merchant type as % of average number of total transactions in the last 12 months.
Travel	Airlines, auto rental, hotel, travel services, and other transportation.
Restaurant & Entertainment	Restaurant, bar, amusement entertainment
Shopping	Department store and clothing/apparel stores. retail stores
Day to Day	Telecommunications, utilities, gas station, grocery, insurance payments.
Cash	Cash advance from manual disbursement and ATM.
Home	Home improvement, electronics appliances, furniture

© 2018 NUS. All rights reserved.

Validating Clusters

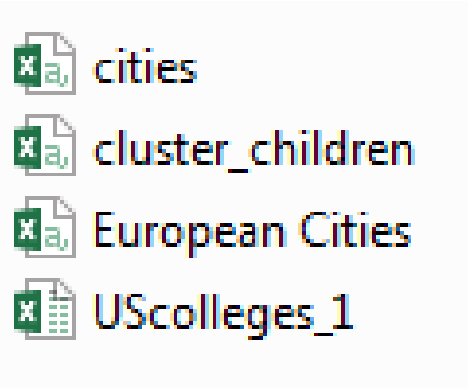
As described the method in slide 12, in order to apply clustering result for decision, validating is necessary.

It may be reviewed from

- Cluster interpretability (through human examination)
  - Is the interpretation of the resulting clusters reasonable?
    - E.g.: in the US College case study, k-means with k=5 and k=3
- Cluster stability (through experimental test)
  - Partition the data and see how well clusters that are formed based on one part apply to the other part
- Cluster separation (through statistical test)
  - Statistical tests to examine the ratio of between-cluster variation to within-cluster variation (but their usefulness is somewhat controversial)

© 2018 NUS. All rights reserved.

Data Sets For Experimentation



Manufacturing Industry Example

http://www.jmp.com/en_us/events/ondemand/mastering-jmp/pinpointing-and-reducing-defects.html

© 2018 NUS. All rights reserved.

Thank You!

© 2018 NUS. All rights reserved.