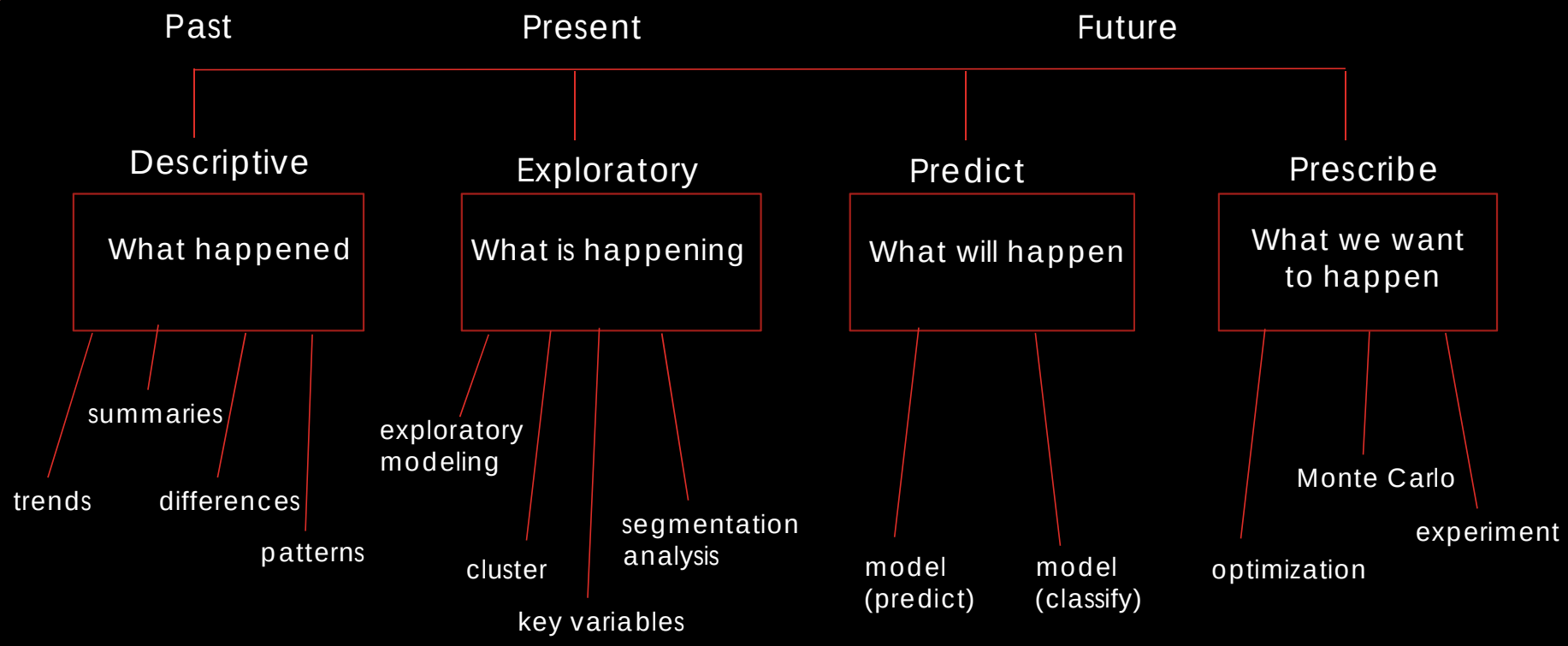
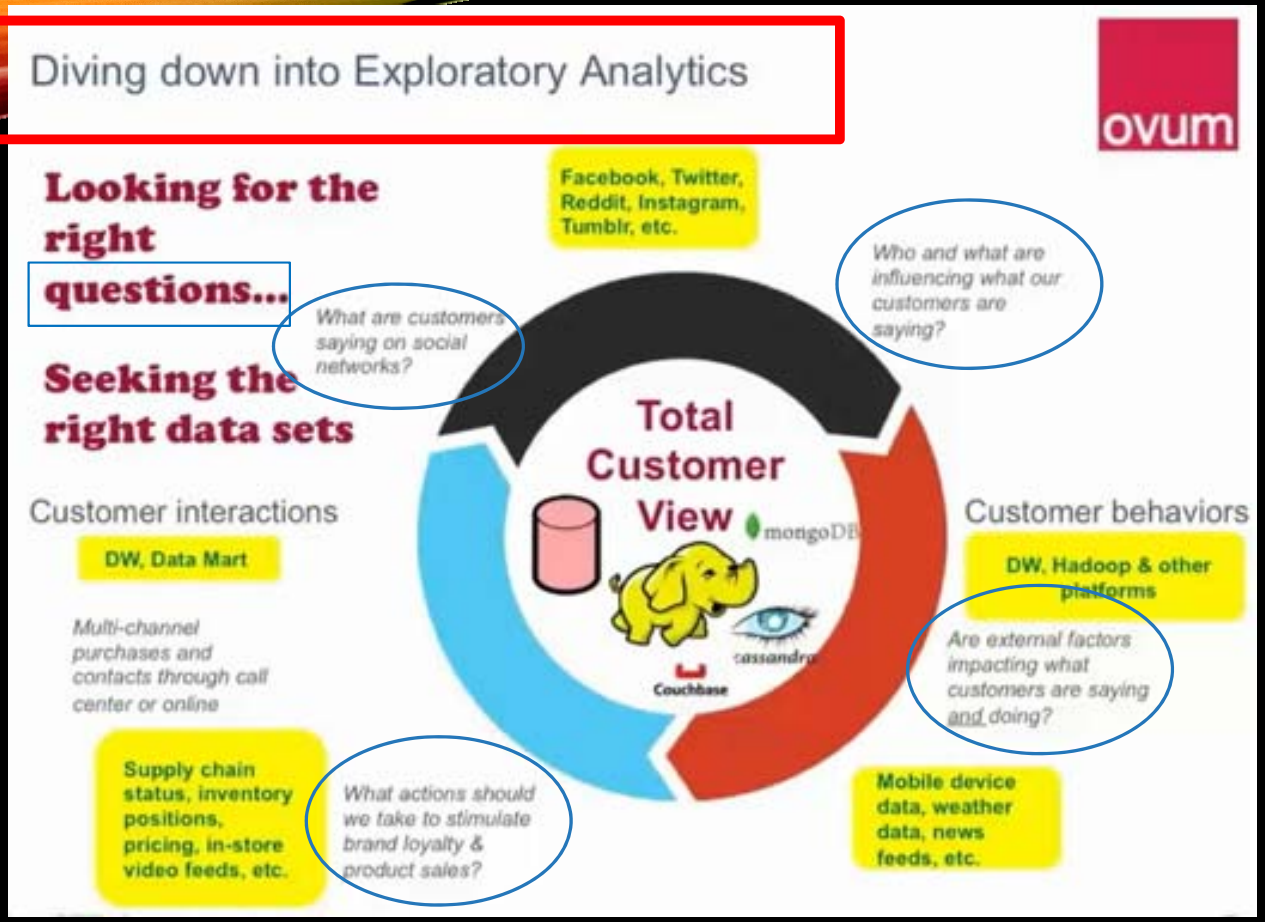
EXPLORATORY ANALYSIS



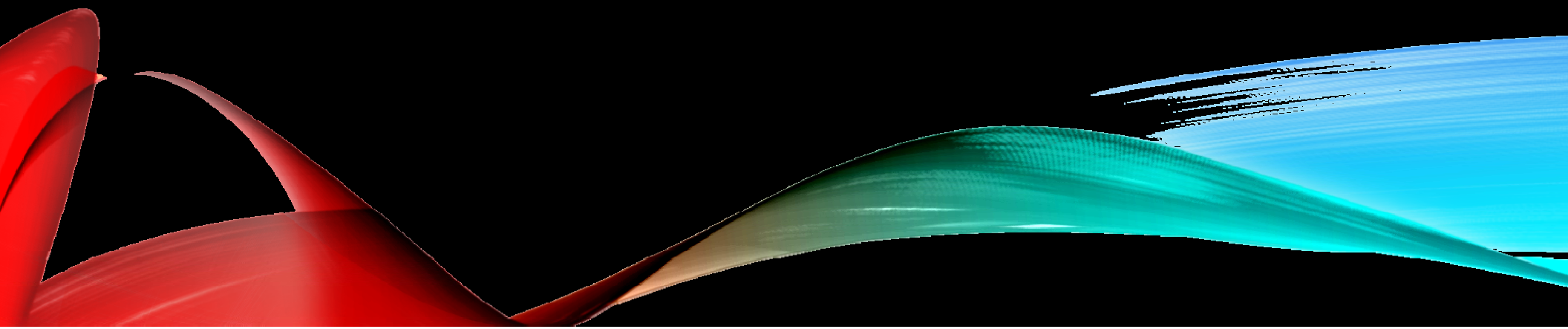
Source: Jim Grayson, Ph.D. and Mia Stephens



Data Wrangling for Agile Analytics | BrightTalk | 1-21-2015
Tony Baer, Principal Analyst, Ovum Sean Kandel, Co-Founder & CTO, Trifacta

- Cluster Analysis
- Segmentation Analysis
- Partition

CLUSTER ANALYSIS



Clustering is the technique of grouping rows together that share similar values across a number of variables. It is a wonderful exploratory technique to help you understand the clumping structure of your data. JMP provides three different clustering methods:

- Hierarchical clustering is appropriate for small tables, up to several thousand rows. It combines rows in a hierarchical sequence portrayed as a tree. In JMP, the tree, also called a dendrogram, is a dynamic, responding graph. You can choose the number of clusters that you like after the tree is built.
- *K*-means clustering is appropriate for larger tables, up to hundreds of thousands of rows. It makes a fairly good guess at cluster seed points. It then starts an iteration of alternately assigning points to clusters and recalculating cluster centers. You have to specify the number of clusters before you start the process.
- Normal mixtures are appropriate when data is assumed to come from a mixture of multivariate normal distributions that overlap. Maximum likelihood is used to estimate the mixture proportions and the means, standard deviations, and correlations jointly. This approach is particularly good at estimating the total counts in each group. However, each point, rather than being classified into one group, is assigned a probability of being in each group. The EM algorithm is used to obtain estimates.

After the clustering process is complete, you can save the cluster assignments to the data table or use them to set colors and markers for the rows.

JMP Book: Multivariate Methods

Clustering is a multivariate technique of grouping rows together that share similar values. It can use any number of variables. The variables must be numeric variables for which numerical differences make sense. The common situation is that data are not scattered evenly through *n*-dimensional space, but rather they form clumps, locally dense areas, modes, or clusters. The identification of these clusters goes a long way toward characterizing the distribution of values.

JMP provides two approaches to clustering:

- *hierarchical clustering* for small tables, up to several thousand rows
- *k-means* and *normal mixtures* clustering for large tables, up to hundreds of thousands of rows.

JMP Book: Multivariate Methods

Hierarchical clustering is also called *agglomerative clustering* because it is a combining process. The method starts with each point (row) as its own cluster. At each step the clustering process calculates the distance between each cluster, and combines the two clusters that are closest together. This combining continues until all the points are in one final cluster. The user then chooses the number of clusters that seems right and cuts the clustering tree at that point. The combining record is portrayed as a tree, called a *dendrogram*. The single points are leaves, the final single cluster of all points are the trunk, and the intermediate cluster combinations are branches. Since the process starts with $n(n + 1)/2$ distances for n points, this method becomes too expensive in memory and time when n is large.

Hierarchical clustering also supports character columns. If the column is ordinal, then the data value used for clustering is just the index of the ordered category, treated as if it were continuous data. If the column is nominal, then the categories must match to contribute a distance of zero. They contribute a distance of 1 otherwise.

JMP offers five rules for defining distances between clusters: Average, Centroid, Ward, Single, and Complete. Each rule can generate a different sequence of clusters.

JMP Book: Multivariate Methods

Clustering

Use clustering to automatically group rows having similar characteristics.

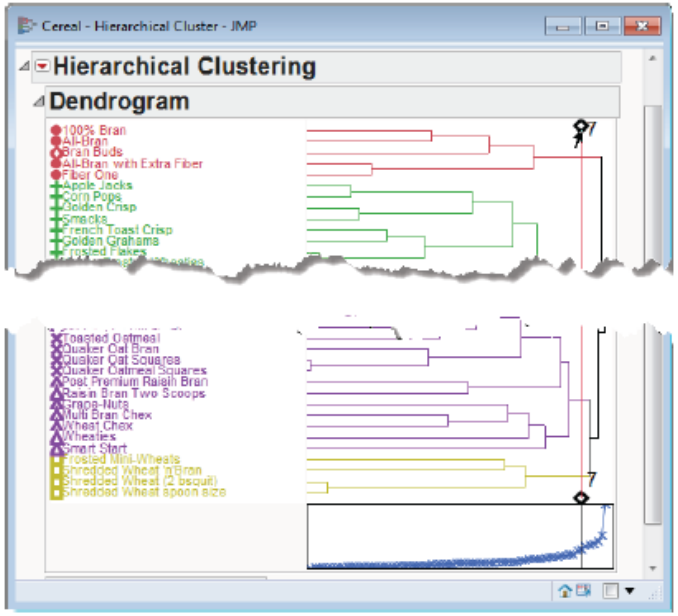
Hierarchical Clustering

1. From an open JMP® data table, select **Analyze > Multivariate Methods > Cluster.**
2. Select one or more variables from **Select Columns** and click **Y, Columns.**
3. If available, select a **Label** variable.
4. Select the desired **method** (bottom left corner) and click **OK.**

JMP will generate:

- A **dendrogram**, showing the clusters formed at each step.
- A **scree plot**, showing the distance bridged each step.
- The **clustering history**, giving cluster statistics for each step.

Example: Cereal.jmp (Help > Sample Data)

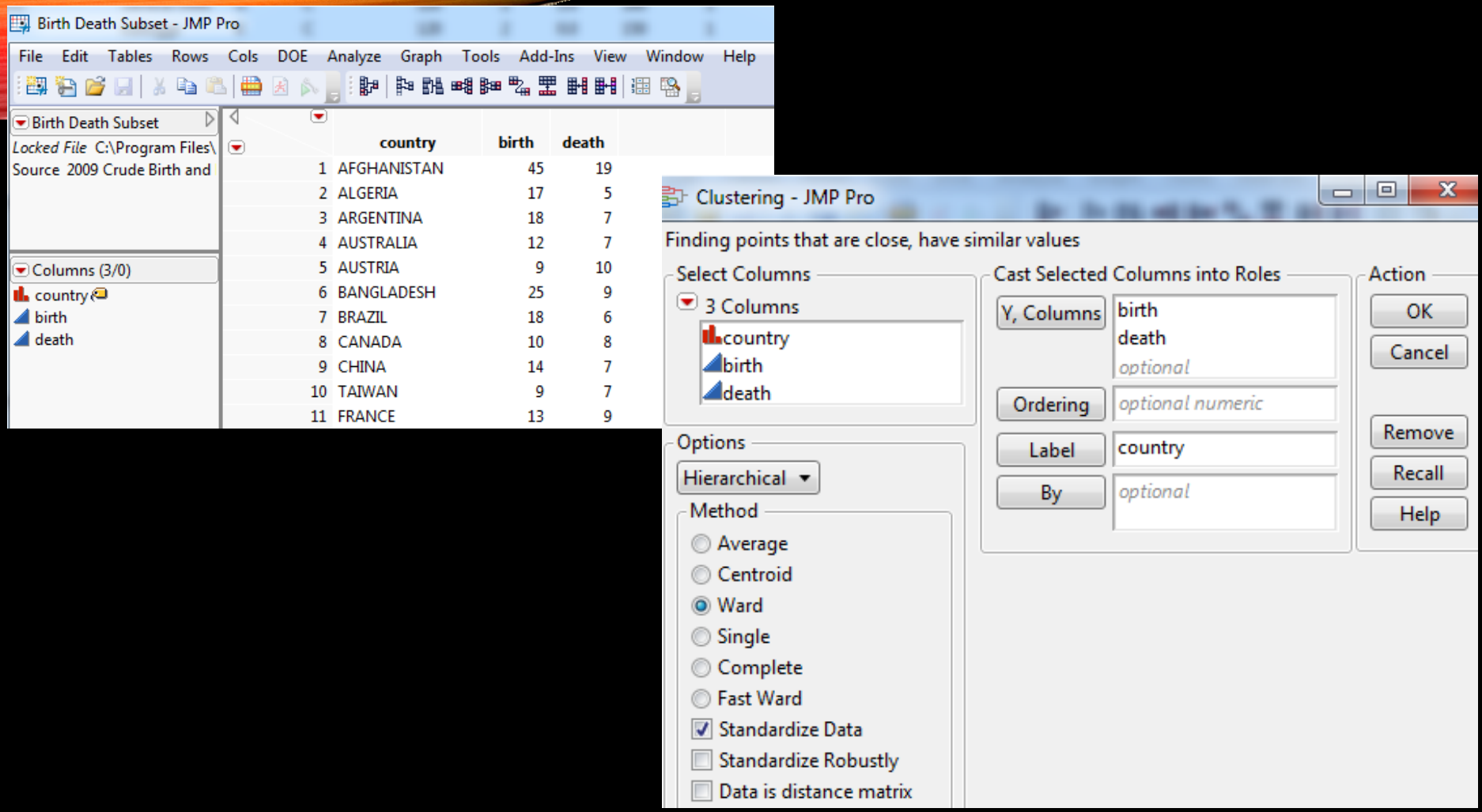


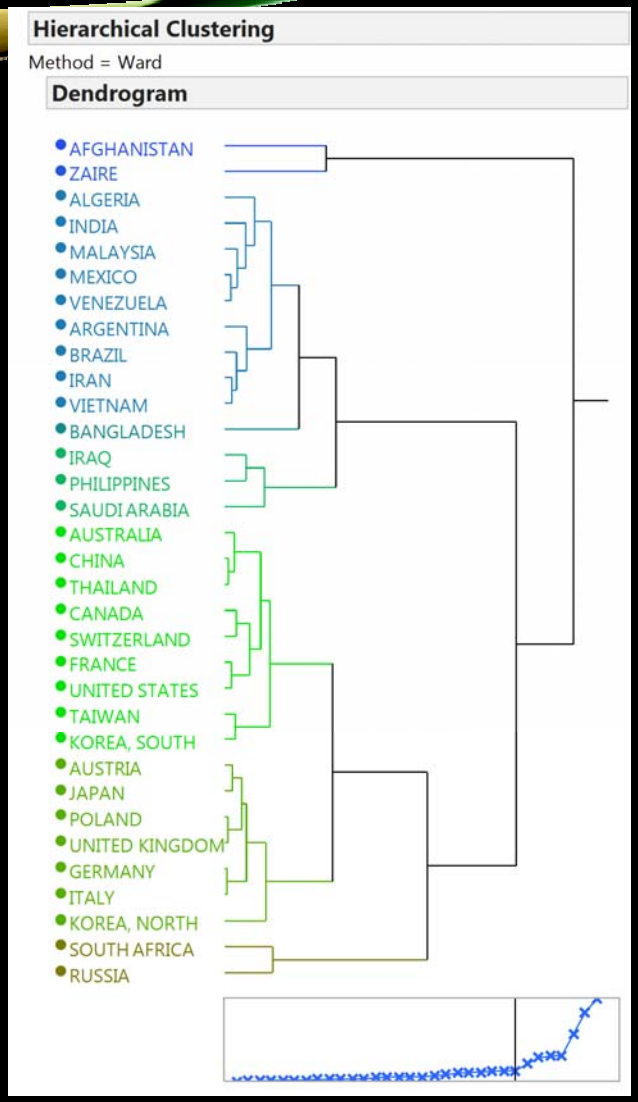
Clustering

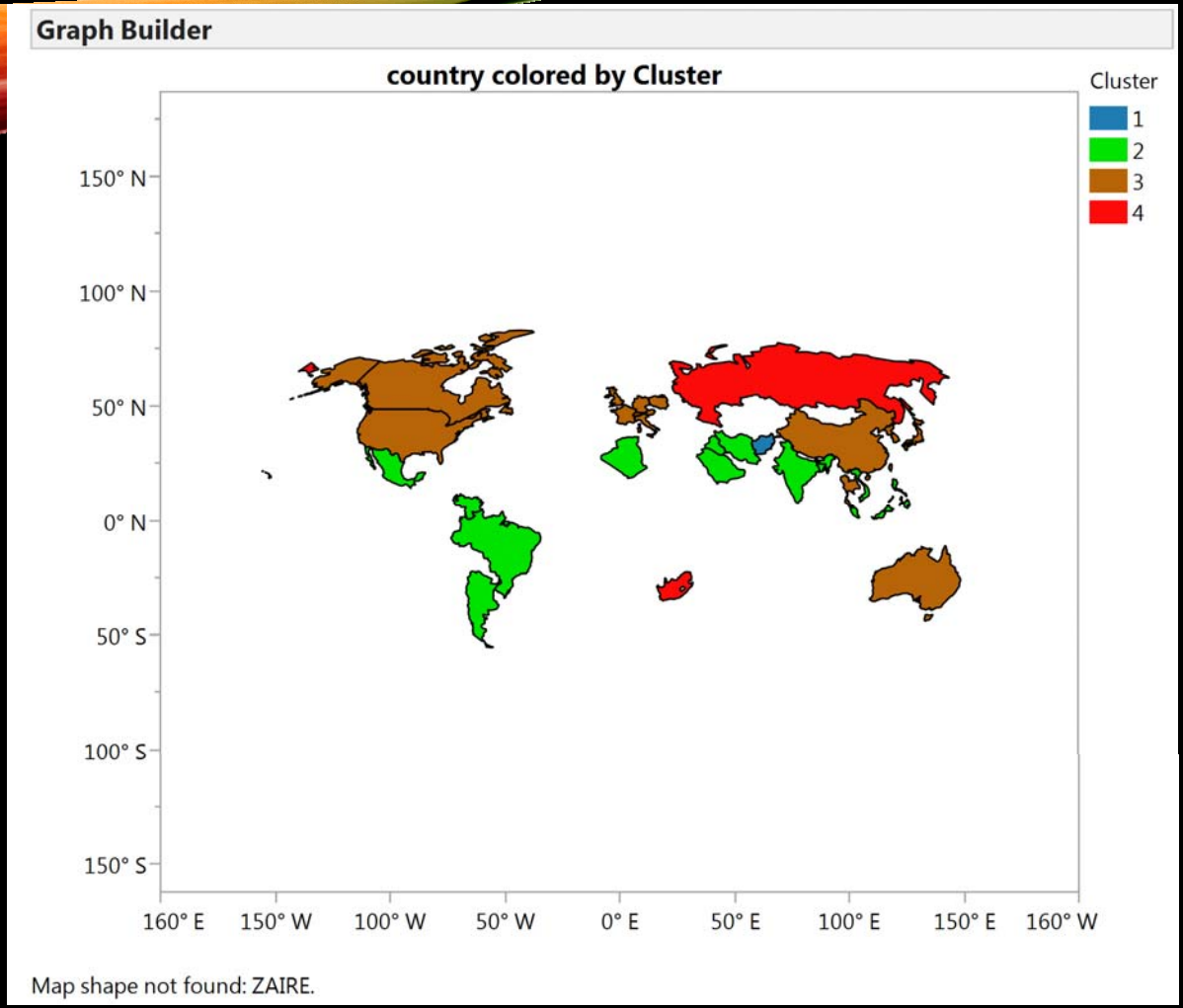
Tips:

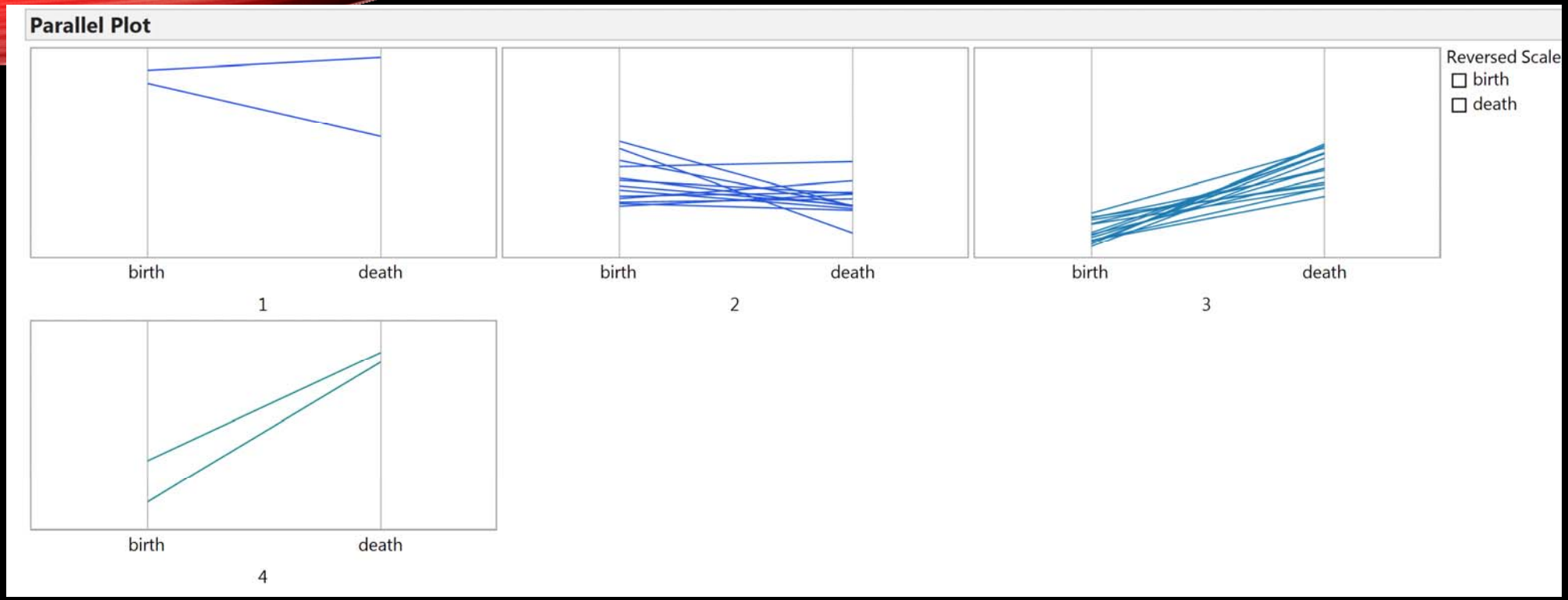
- To **color clusters**, to **mark or save clusters**, or to **request other options**, click the **top red triangle**.
- To dynamically change the number of clusters, click and drag one of the **black diamonds** left or right.

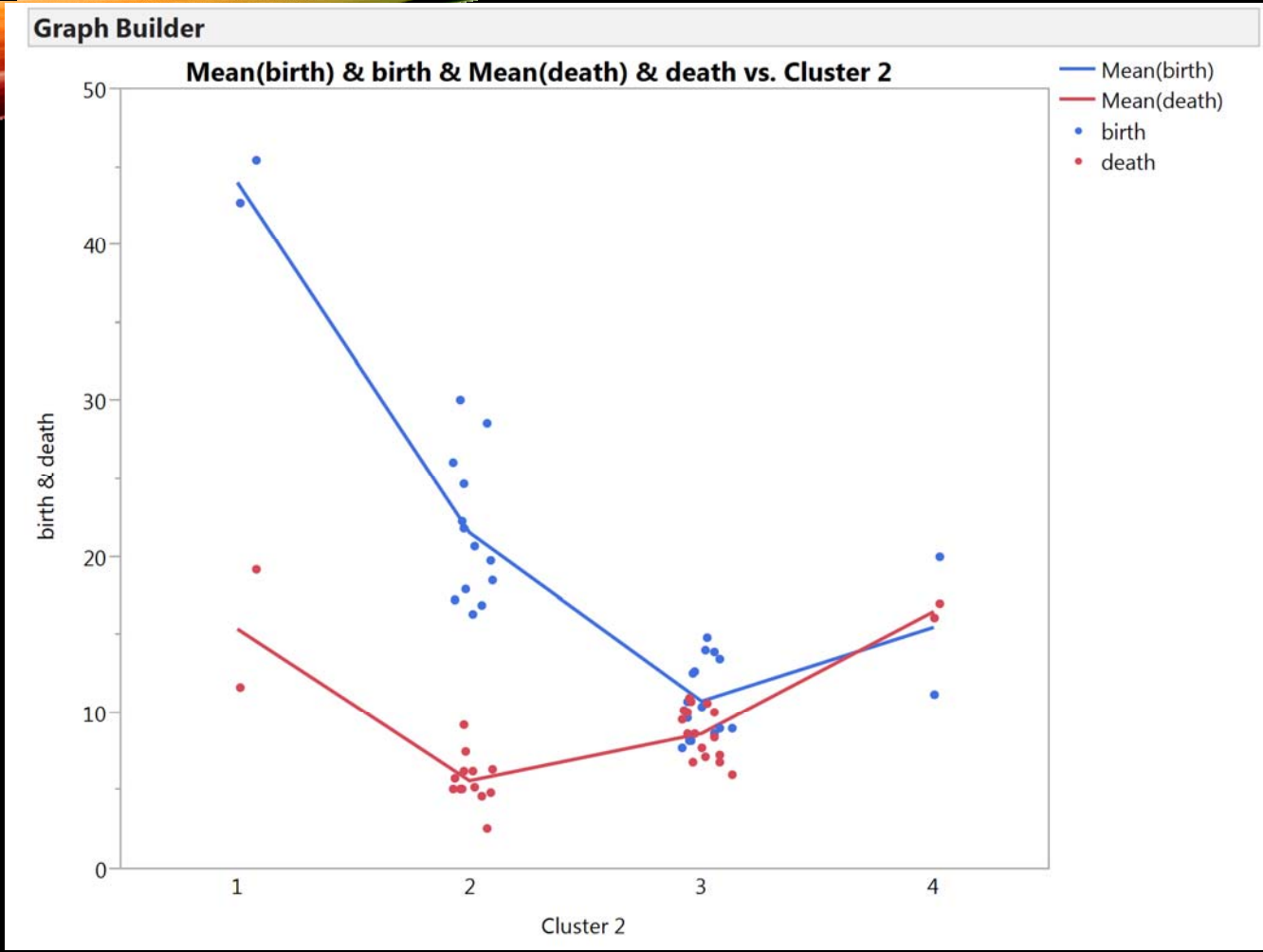
BIRTH DEATH SUBSET DATA JMP HELP FILE

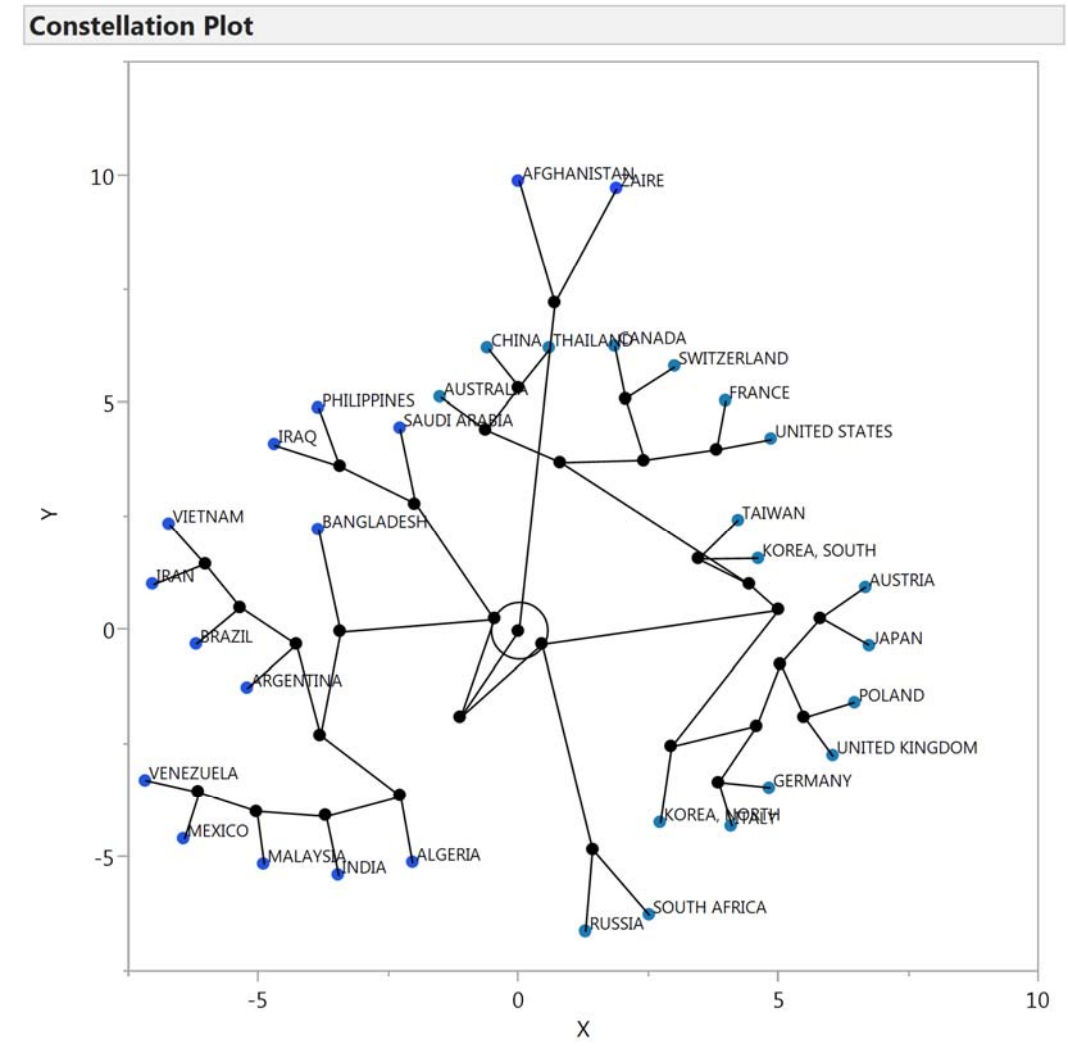

Birth Death Subset - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help
Birth Death Subset
Locked File C:\Program Files\
Source 2009 Crude Birth and
Columns (3/0)
country
birth
death
country birth death
1 AFGHANISTAN 45 19
2 ALGERIA 17 5
3 ARGENTINA 18 7
4 AUSTRALIA 12 7
5 AUSTRIA 9 10
6 BANGLADESH 25 9
7 BRAZIL 18 6
8 CANADA 10 8
9 CHINA 14 7
10 TAIWAN 9 7
11 FRANCE 13 9
Clustering - JMP Pro
Finding points that are close, have similar values
Select Columns
3 Columns
country
birth
death
Cast Selected Columns into Roles
Y, Columns
birth
death
optional
Ordering
optional numeric
Label
country
By
optional
Action
OK
Cancel
Remove
Recall
Help
Options
Hierarchical
Method
Average
Centroid
Ward
Single
Complete
Fast Ward
Standardize Data
Standardize Robustly
Data is distance matrix


Hierarchical Clustering
Method = Ward
Dendrogram
AFGHANISTAN
ZAIRE
ALGERIA
INDIA
MALAYSIA
MEXICO
VENEZUELA
ARGENTINA
BRAZIL
IRAN
VIETNAM
BANGLADESH
IRAQ
PHILIPPINES
SAUDI ARABIA
AUSTRALIA
CHINA
THAILAND
CANADA
SWITZERLAND
FRANCE
UNITED STATES
TAIWAN
KOREA, SOUTH
AUSTRIA
JAPAN
POLAND
UNITED KINGDOM
GERMANY
ITALY
KOREA, NORTH
SOUTH AFRICA
RUSSIA


Graph Builder
country colored by Cluster
Cluster
1
2
3
4
150° N
100° N
50° N
0° N
50° S
100° S
150° S
160° E
150° W
100° W
50° W
0° E
50° E
100° E
150° E
160° W
Map shape not found: ZAIRE.


Parallel Plot
Reversed Scale
birth
death
birth
death
1
birth
death
2
birth
death
3
birth
death
4

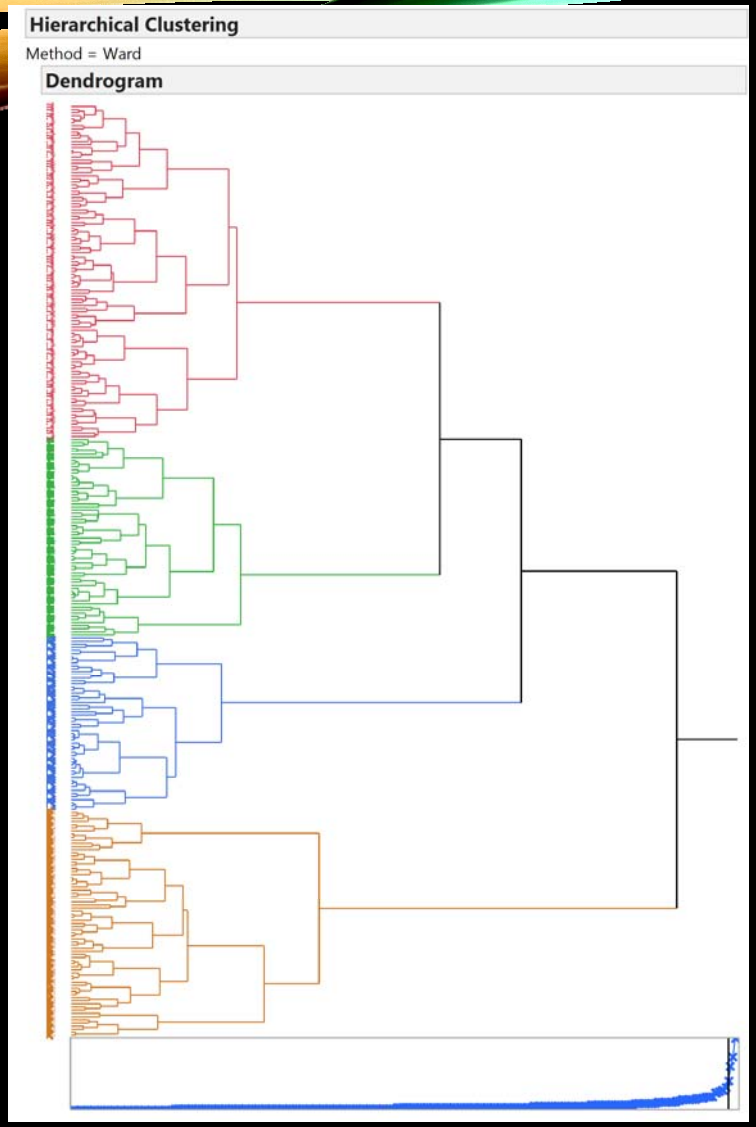

Graph Builder
Mean(birth) & birth & Mean(death) & death vs. Cluster 2
Mean(birth)
Mean(death)
birth
death
birth & death
50
40
30
20
10
0
1
2
3
4
Cluster 2

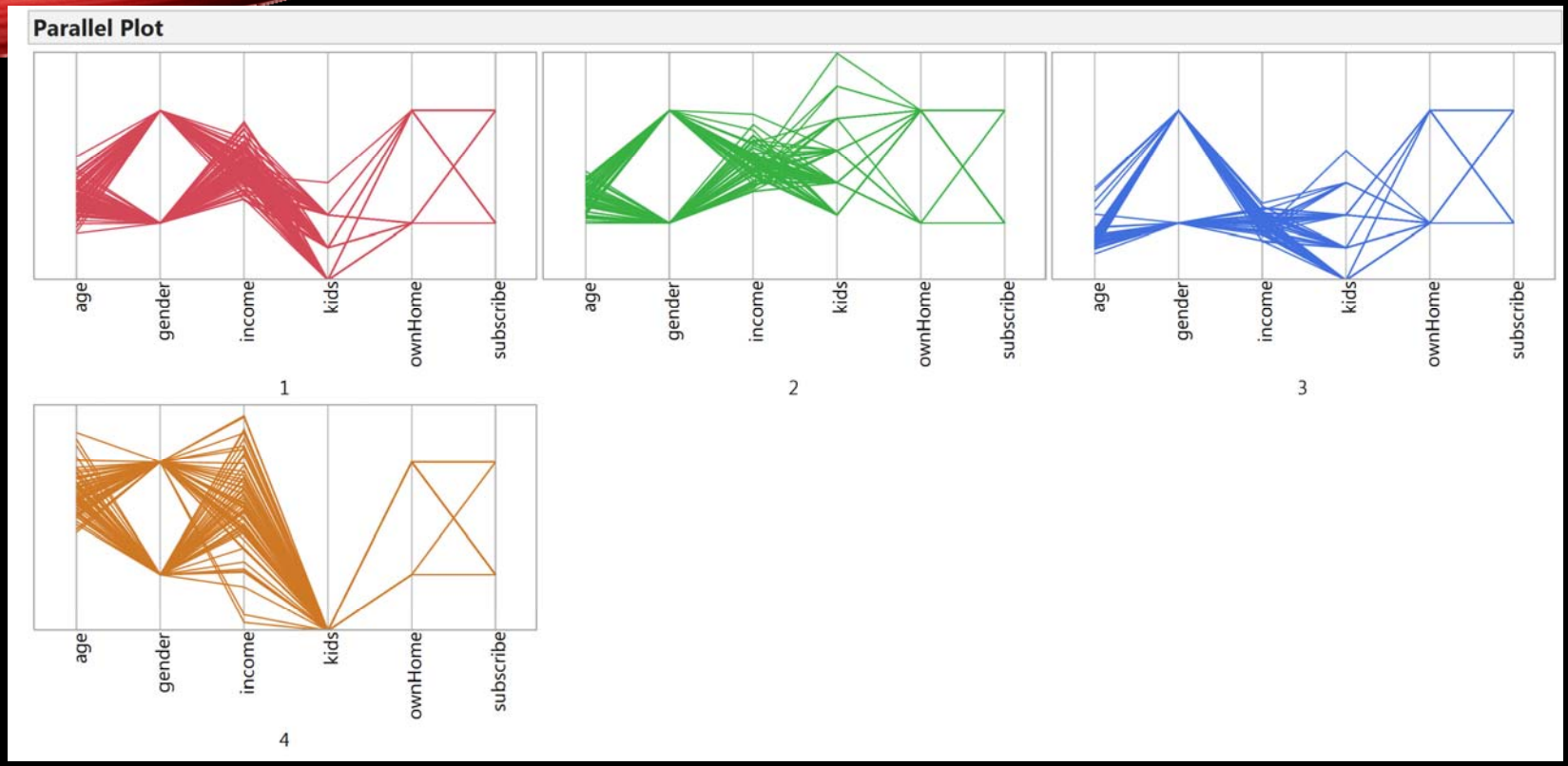

Constellation Plot
AFGHANISTAN
ZAIRE
CHINA
THAILAND
CANADA
SWITZERLAND
FRANCE
AUSTRALIA
PHILIPPINES
SAUDI ARABIA
IRAQ
UNITED STATES
VIETNAM
BANGLADESH
TAIWAN
KOREA, SOUTH
IRAN
AUSTRIA
BRAZIL
JAPAN
ARGENTINA
POLAND
UNITED KINGDOM
VENEZUELA
GERMANY
MEXICO
MALAYSIA
INDIA
ALGERIA
SOUTH AFRICA
RUSSIA
Y
X
10
5
0
-5
-5
0
5
10

Data Set: rintro-chapter5.csv

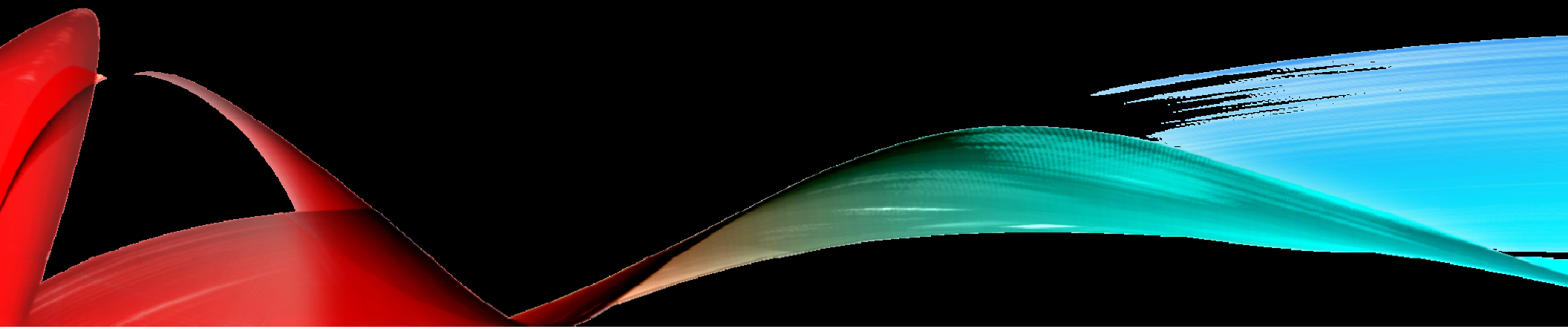
R for Marketing Research and Analytics by Chris Chapman and Elea McDonnell Feit (Springer)

The data set is a simulated data set for a consumer segmentation project. The scenario is for a subscription based service. Data collected are 300 data points for respondents on age, gender, income, number of children, whether they own (or rent) their homes (y/n) and whether they subscribe to the service (yes/no).

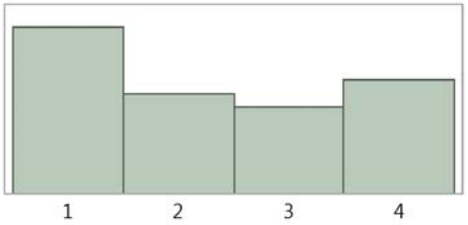

Hierarchical Clustering
Method = Ward
Dendrogram


Parallel Plot
age
gender
income
kids
ownHome
subscribe
1
2
3
4

SEGMENTATION ANALYSIS



Cluster



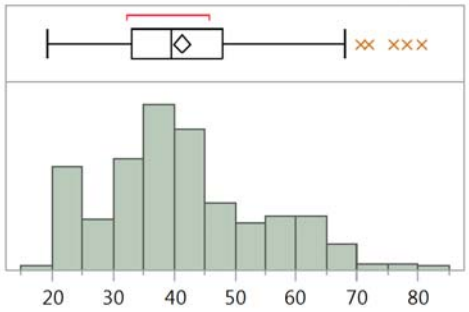
Frequencies

Level	Count	Prob
1	107	0.35667
2	64	0.21333
3	56	0.18667
4	73	0.24333
Total	300	1.00000

N Missing 0

4 Levels

age



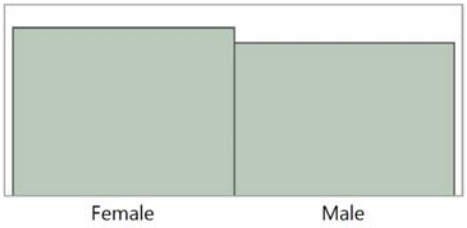
Quantiles

100.0%	maximum	80.4862
99.5%		79.3303
97.5%		68.0547
90.0%		60.9415
75.0%	quartile	47.9262
50.0%	median	39.4877
25.0%	quartile	32.9722
10.0%		24.3254
2.5%		21.9673
0.5%		19.9935
0.0%	minimum	19.2599

Summary Statistics

Mean	41.19965
Std Dev	12.707427
Std Err Mean	0.7336636
Upper 95% Mea	42.643448
Lower 95% Mean	39.755851
N	300

gender



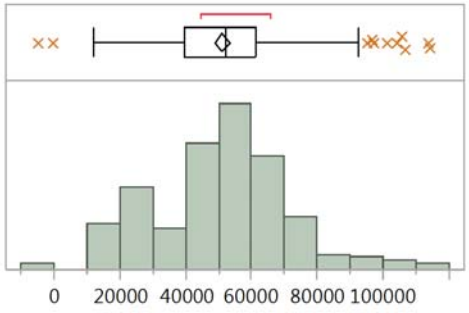
Frequencies

Level	Count	Prob
Female	157	0.52333
Male	143	0.47667
Total	300	1.00000

N Missing 0

2 Levels

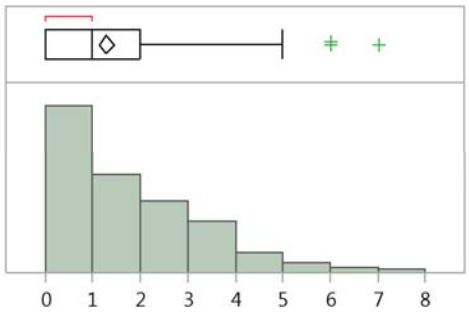
income



Quantiles		
100.0%	maximum	114278
99.5%		113863
97.5%		96695.8
90.0%		73748.2
75.0%	quartile	61436.5
50.0%	median	52014.4
25.0%	quartile	39610.9
10.0%		22201.5
2.5%		15712.4
0.5%		-2916.2
0.0%	minimum	-5183.4

Summary Statistics	
Mean	50936.536
Std Dev	20137.549
Std Err Mean	1162.642
Upper 95% Mea	53224.534
Lower 95% Mean	48648.539
N	300

kids



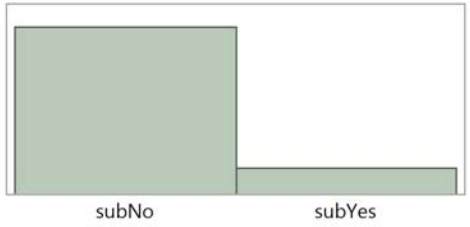
Quantiles		
100.0%	maximum	7
99.5%		6.495
97.5%		5
90.0%		3
75.0%	quartile	2
50.0%	median	1
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	1.27
Std Dev	1.4084432
Std Err Mean	0.0813165
Upper 95% Mea	1.4300252
Lower 95% Mean	1.1099748
N	300

ownHome

Frequencies		
Level	**Count**	**Prob**
ownNo	159	0.53000
ownYes	141	0.47000
Total	300	1.00000
N Missing	0	
2 Levels		

subscribe



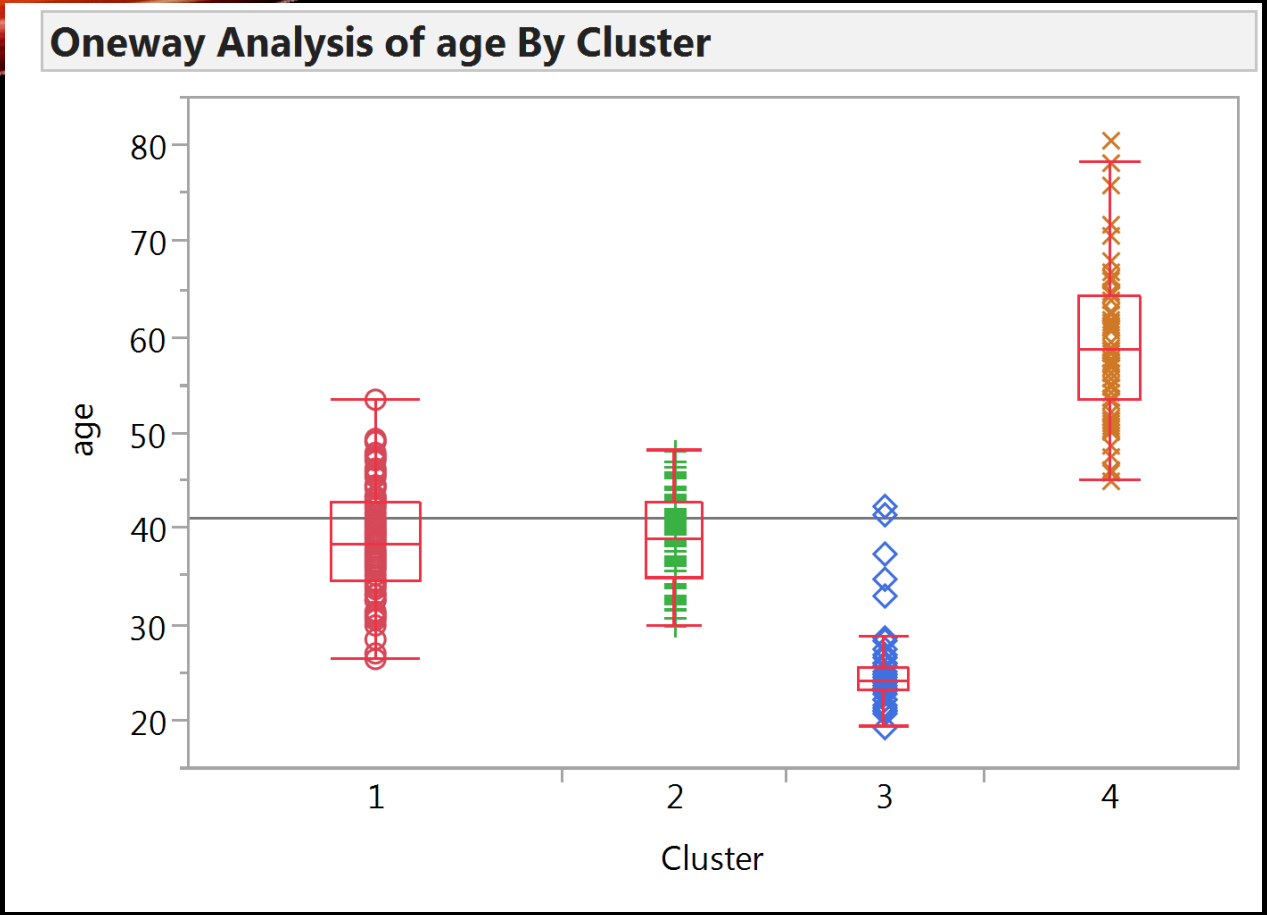
Frequencies		
Level	**Count**	**Prob**
subNo	260	0.86667
subYes	40	0.13333
Total	300	1.00000
N Missing	0	
2 Levels		

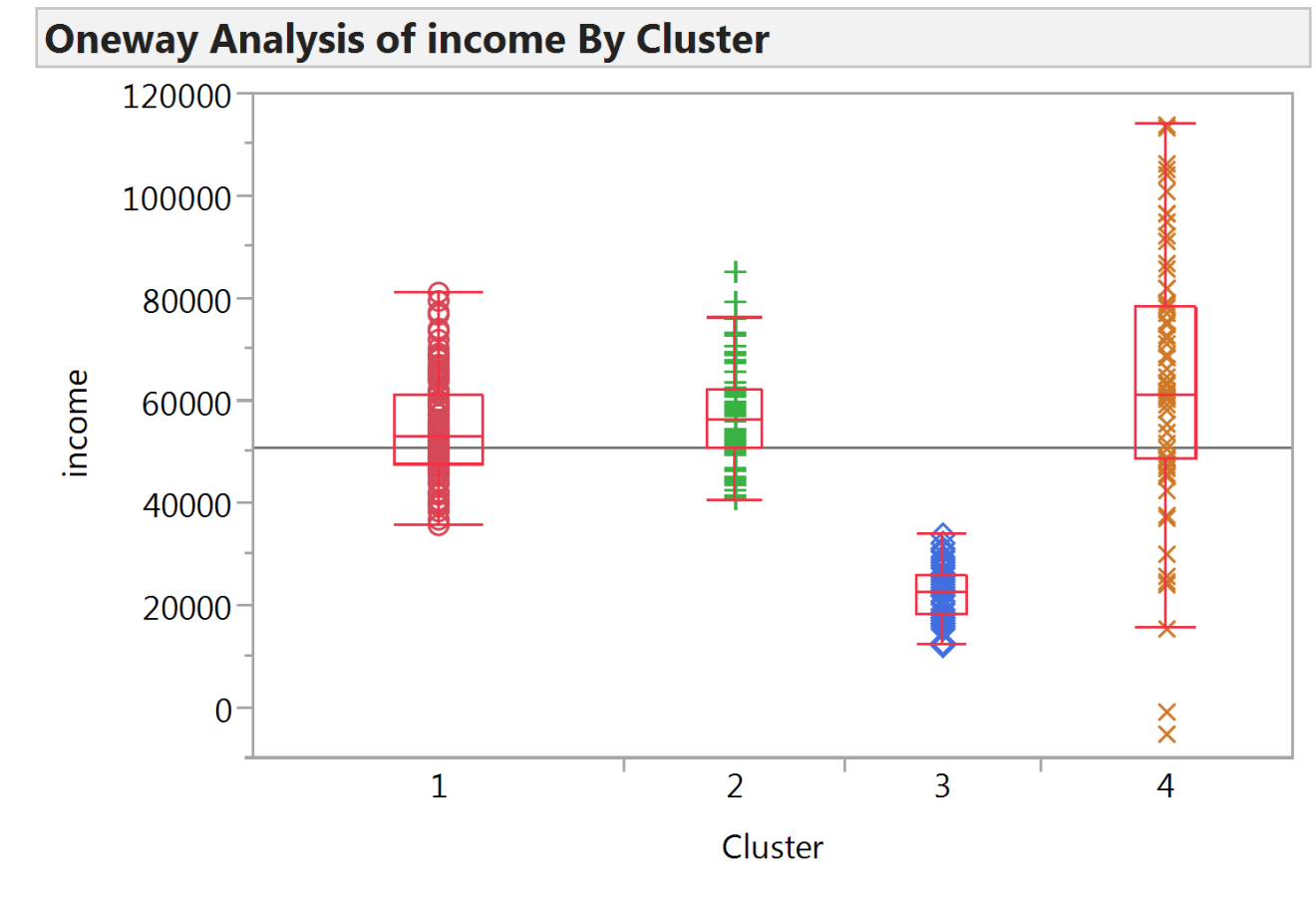
Tabulate

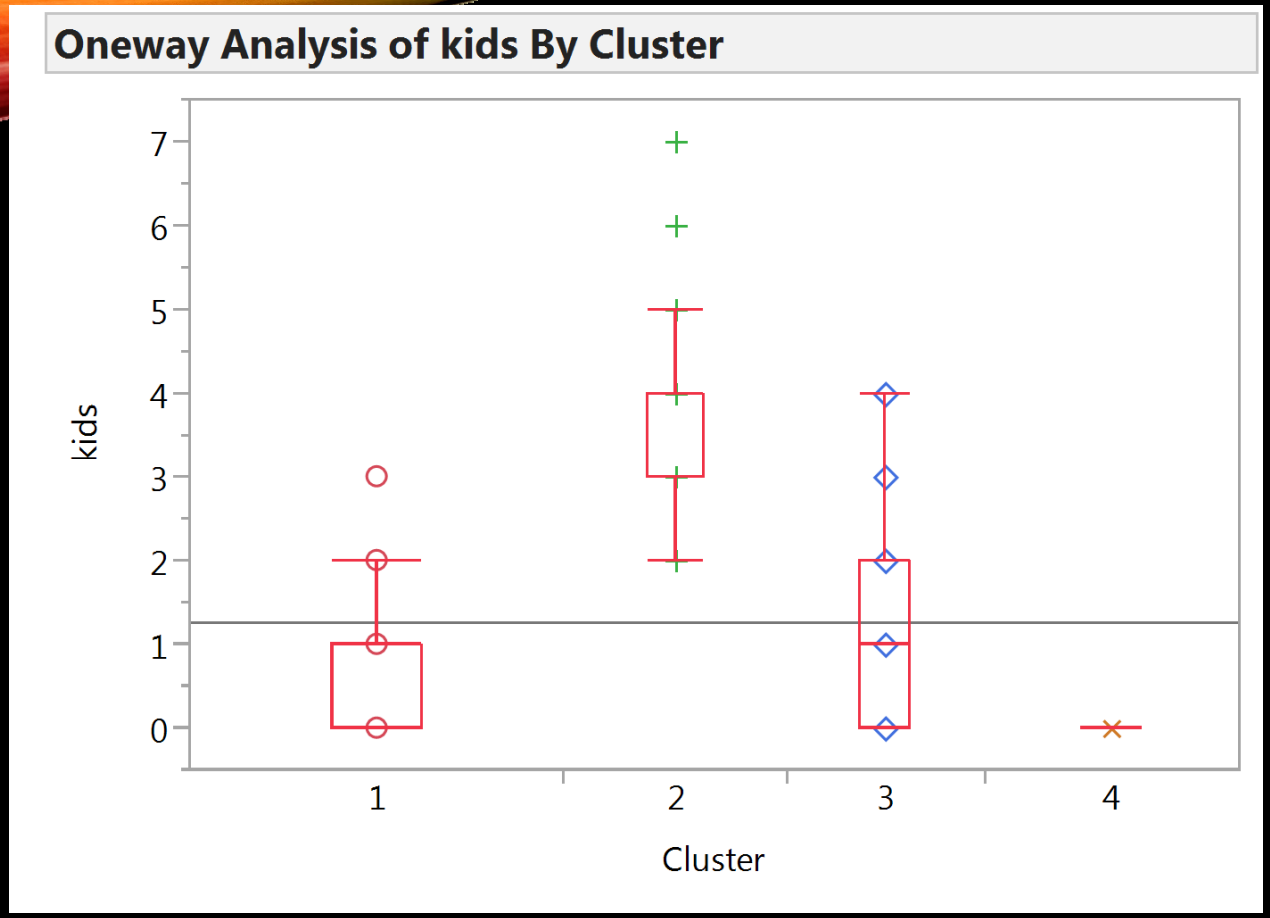
	age		income		kids	
Cluster	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
1	38.69	5.32	54301.39	10115.89	0.95	0.74
2	38.87	4.62	56886.86	9736.62	3.31	1.11
3	25.22	4.54	22305.68	5115.16	1.20	1.05
4	59.17	7.36	62751.17	24557.24	0.00	0.00

Tabulate

	gender		ownHome		subscribe	
	Female	Male	ownNo	ownYes	subNo	subYes
Cluster	Row %	Row %	Row %	Row %	Row %	Row %
1	57.01%	42.99%	67.29%	32.71%	85.05%	14.95%
2	56.25%	43.75%	40.63%	59.38%	92.19%	7.81%
3	39.29%	60.71%	78.57%	21.43%	82.14%	17.86%
4	52.05%	47.95%	23.29%	76.71%	87.67%	12.33%


Oneway Analysis of age By Cluster
age
80
70
60
50
40
30
20
1
2
3
4
Cluster


Oneway Analysis of income By Cluster
120000
100000
80000
60000
40000
20000
0
income
1
2
3
4
Cluster

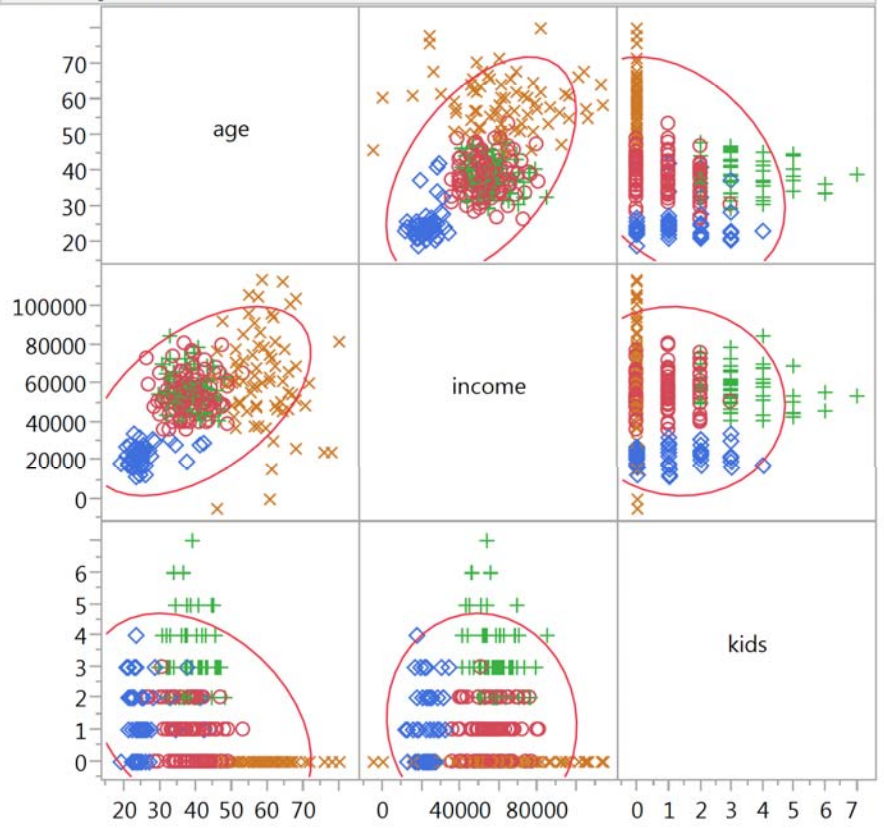

Oneway Analysis of kids By Cluster
kids
7
6
5
4
3
2
1
0
1
2
3
4
Cluster

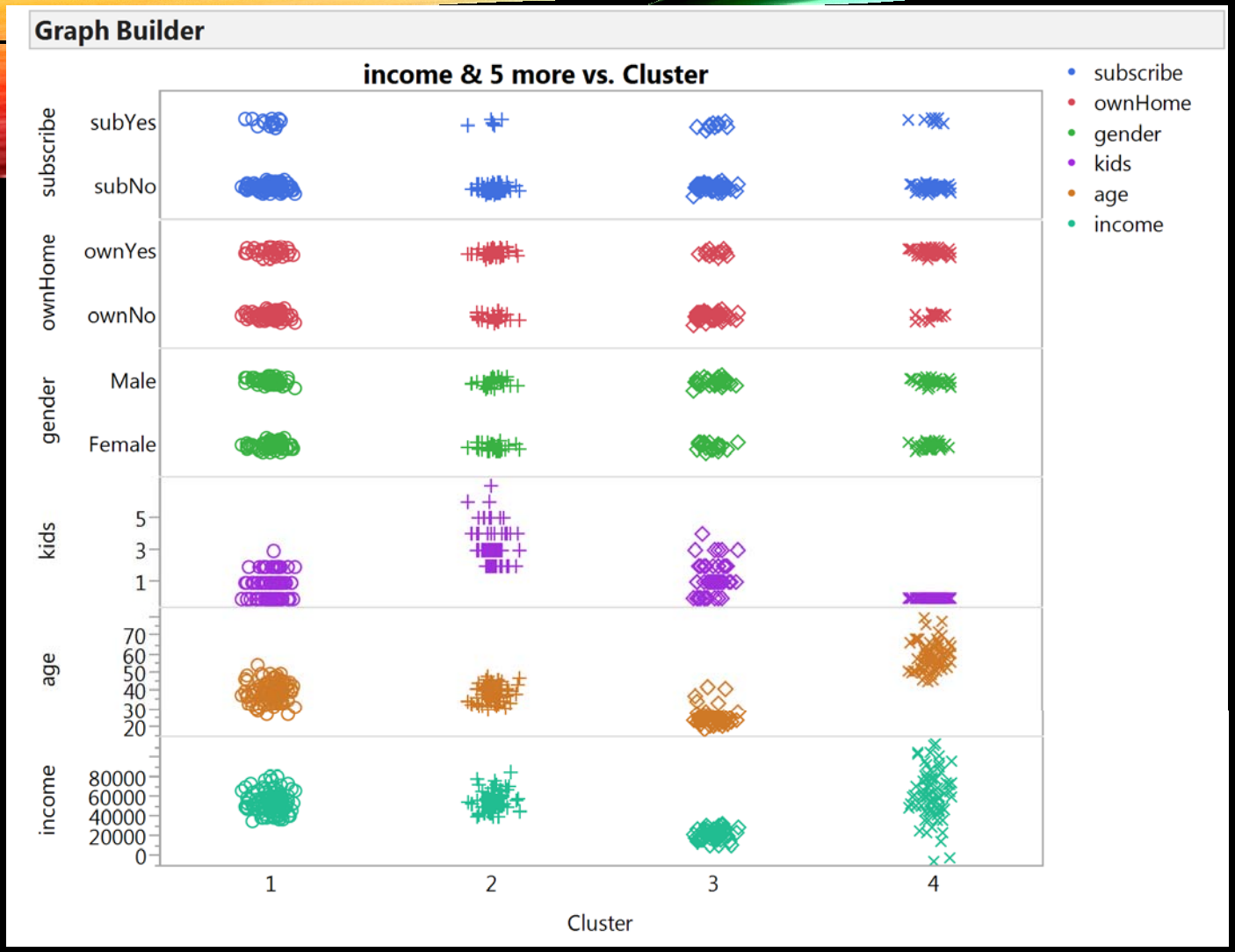
Multivariate

Correlations

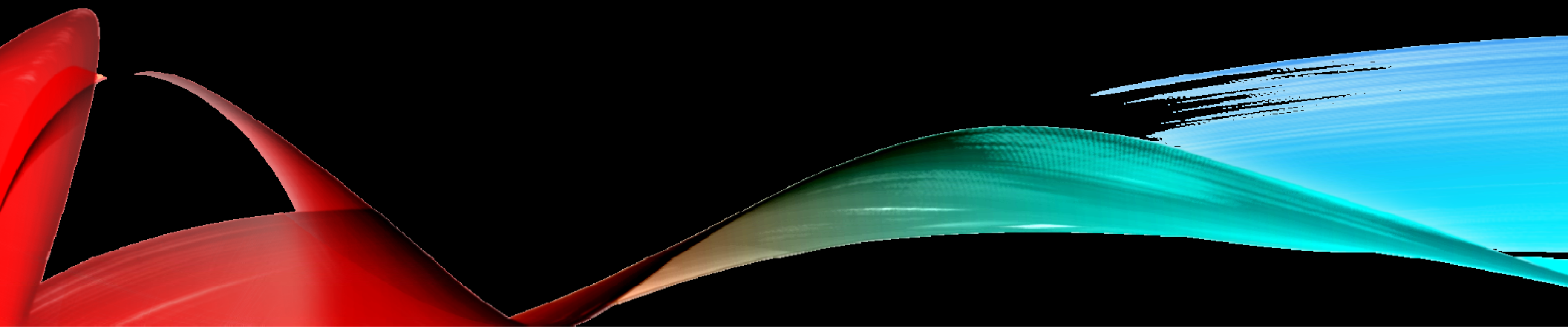
	age	income	kids
age	1.0000	0.5126	-0.3601
income	0.5126	1.0000	-0.0403
kids	-0.3601	-0.0403	1.0000

Scatterplot Matrix



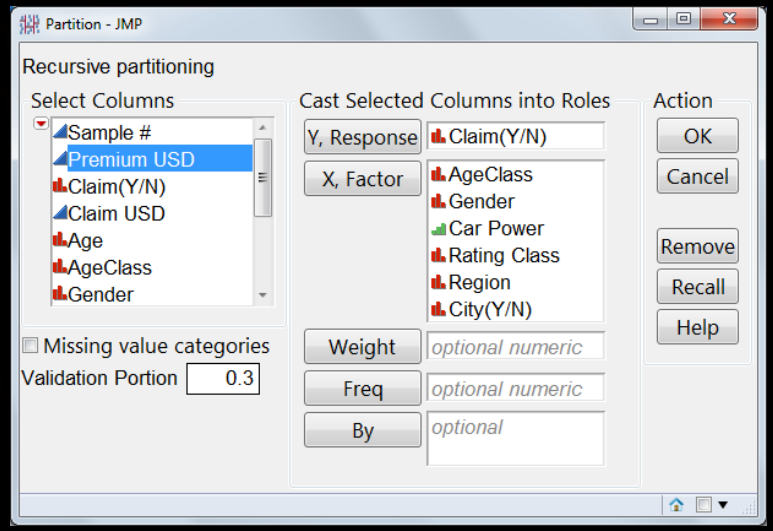

Graph Builder
income & 5 more vs. Cluster
subscribe
subYes
subNo
ownHome
ownYes
ownNo
gender
Male
Female
kids
5
3
1
age
70
60
50
40
30
20
income
80000
60000
40000
20000
0
1
2
3
4
Cluster
subscribe
ownHome
gender
kids
age
income

PARTITION



INTRODUCTION TO PARTITION

- Partition (**Analyze > Modeling > Partition**) is a data mining tool, used for data exploration and for building predictive models.
- We will introduce the tool here, and will revisit the topics again later.
- Example: **Auto Raw Data.jmp** in the Sample Data Directory.

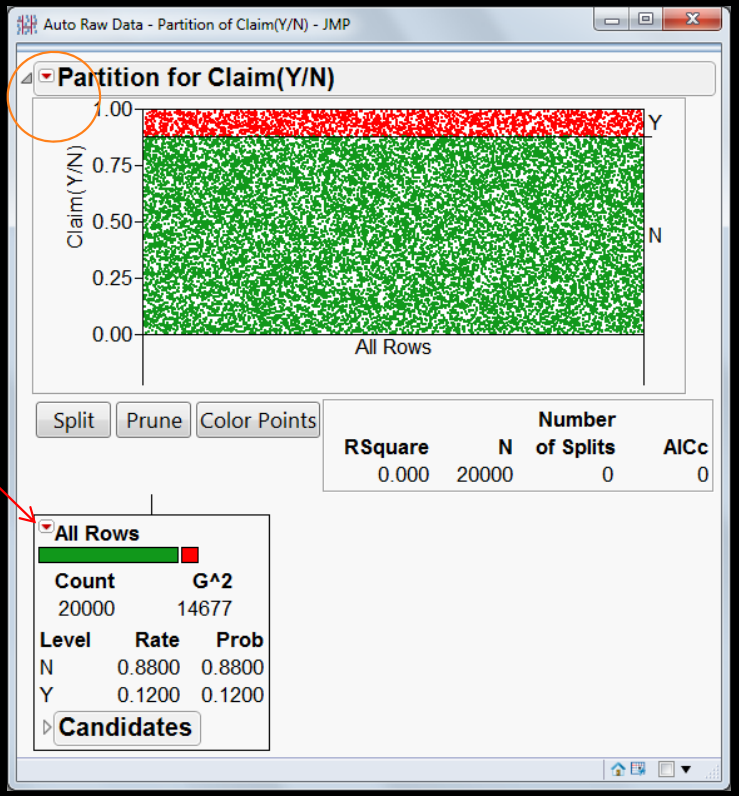


Courtesy of Mia Stephens: DSI 2014 Presentation

INTRODUCTION TO PARTITION

- JMP displays a graph with a line drawn at the overall response rate.

Click on the top red triangle and select **Display Options > Show Split Prob** to display the overall response rate.



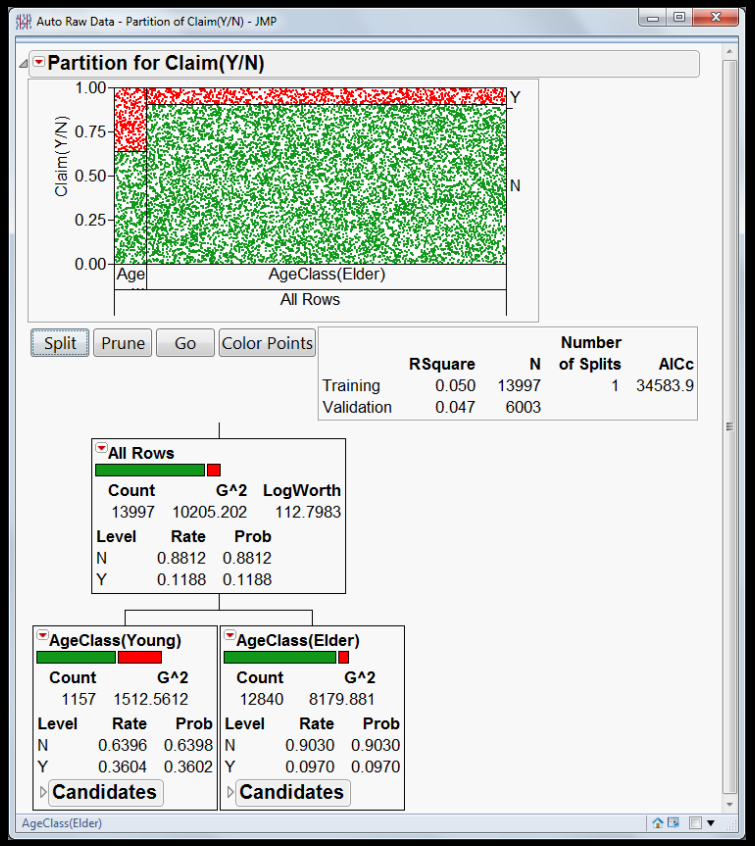
Courtesy of Mia Stephens: DSI 2014 Presentation

INTRODUCTION TO PARTITION

- Click the **Split** button. The original observations will be split into two nodes, or leaves.

Interpretation:

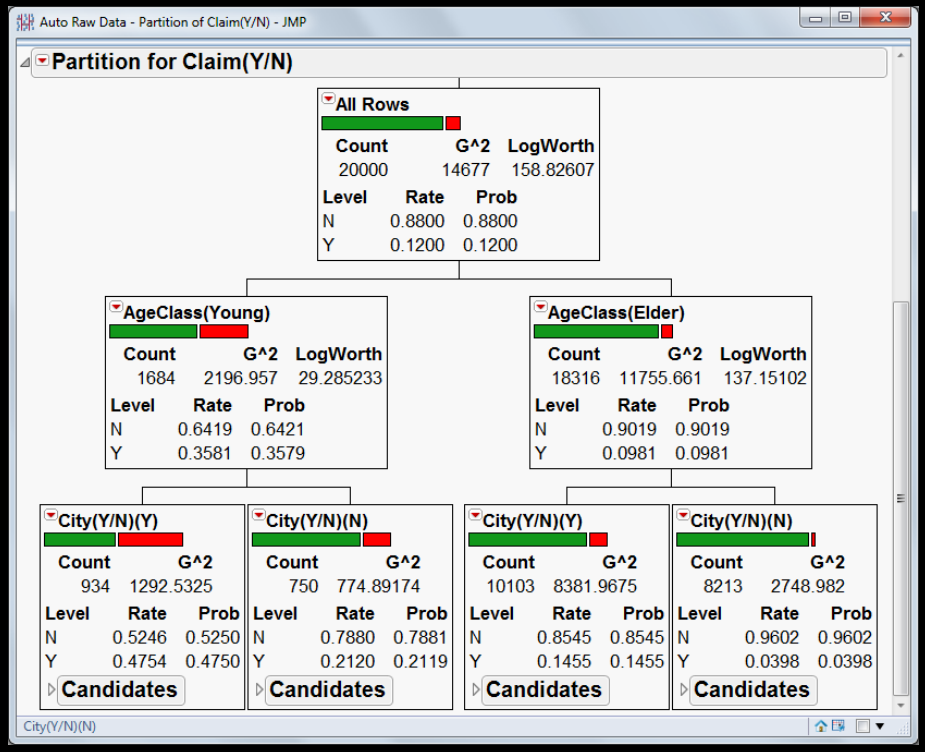
- In the left leaf, corresponding to AgeClass = Young, the probability that there is a claim is 0.3604.
- In the right leaf, corresponding to AgeClass = Elder, the probability that there is a claim is 0.097.



Courtesy of Mia Stephens: DSI 2014 Presentation

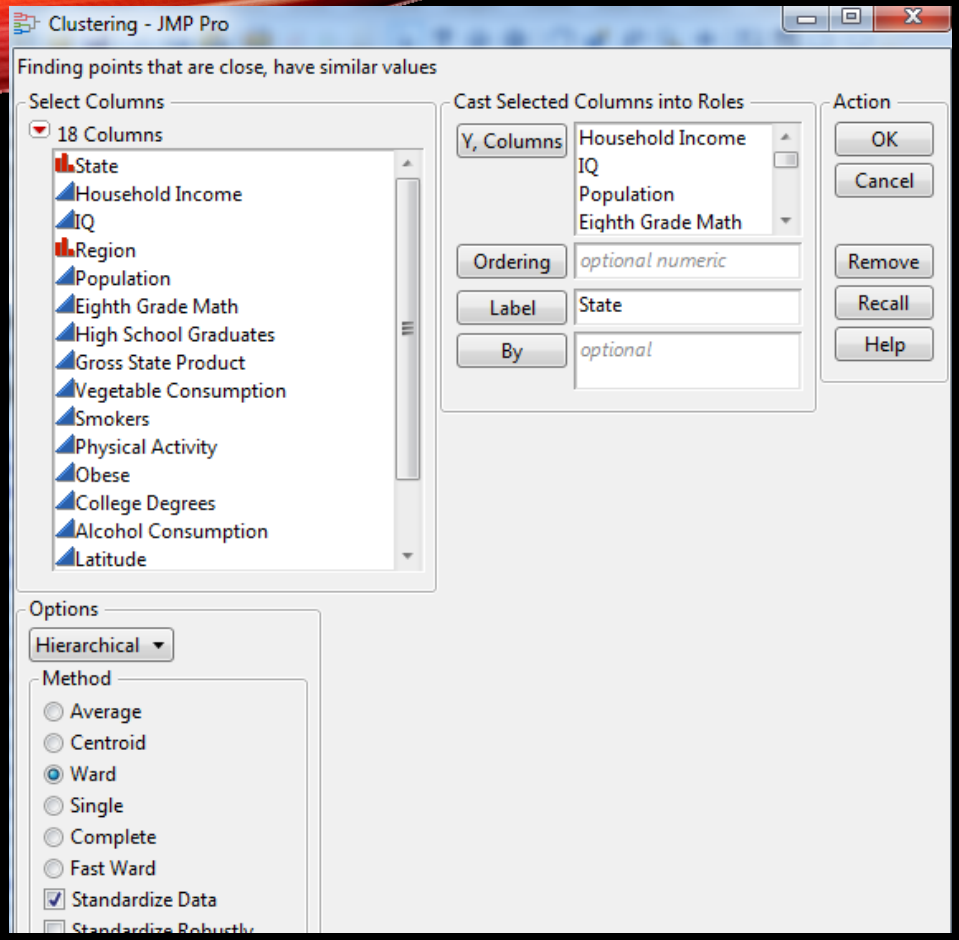
INTRODUCTION TO PARTITION

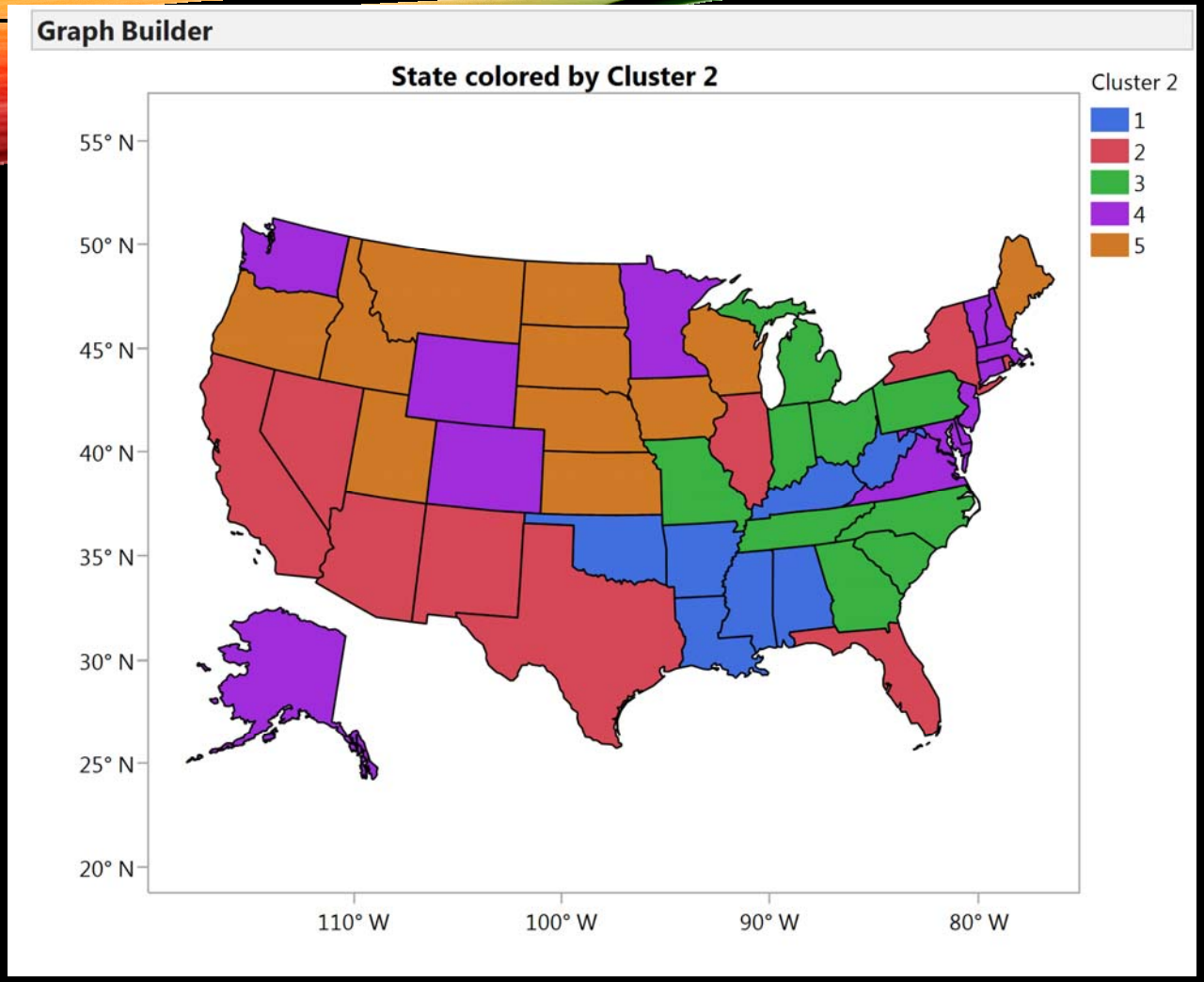
- Click the **Split** button again, and then a third time. The original observations will be split into two nodes, or leaves.

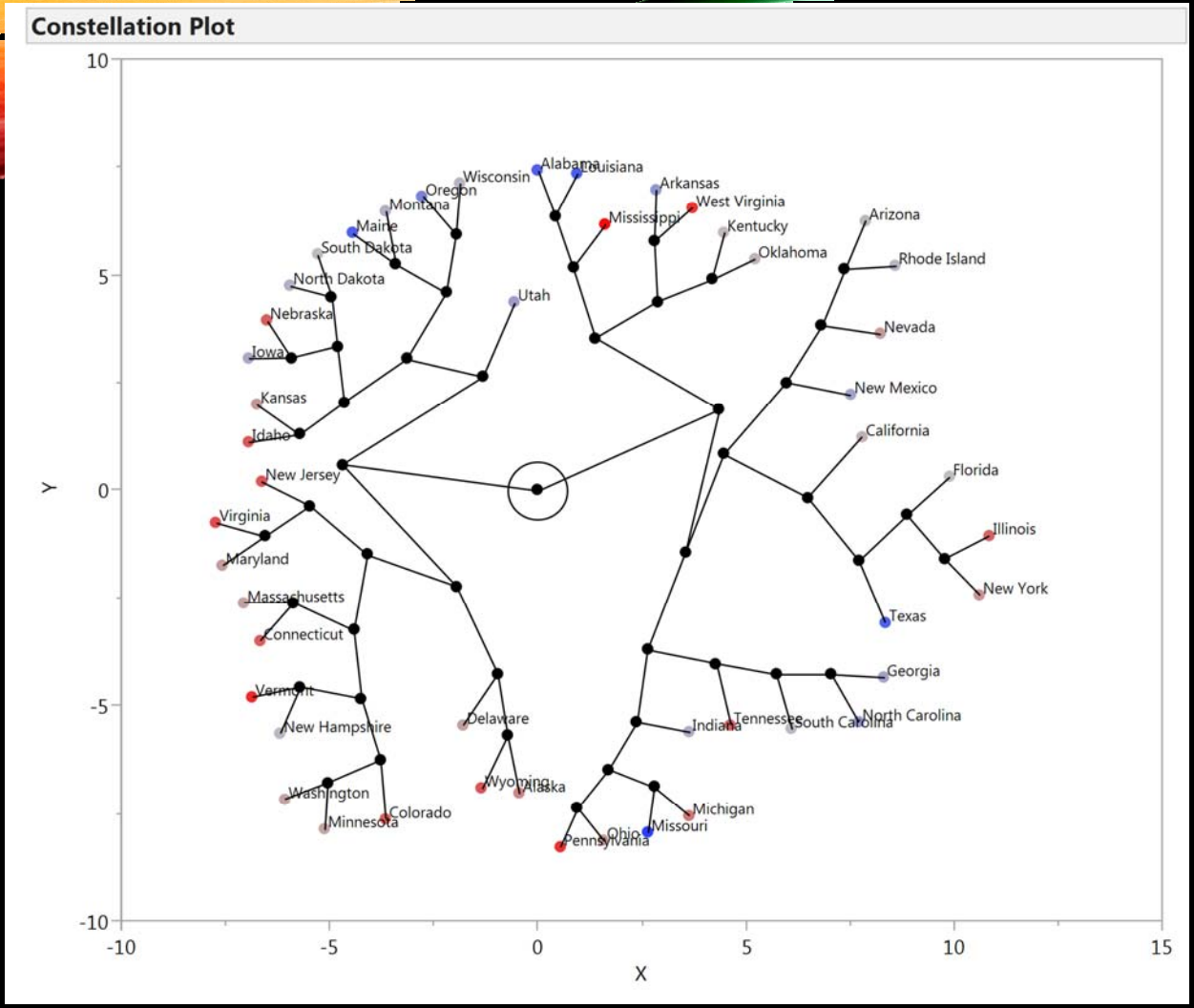


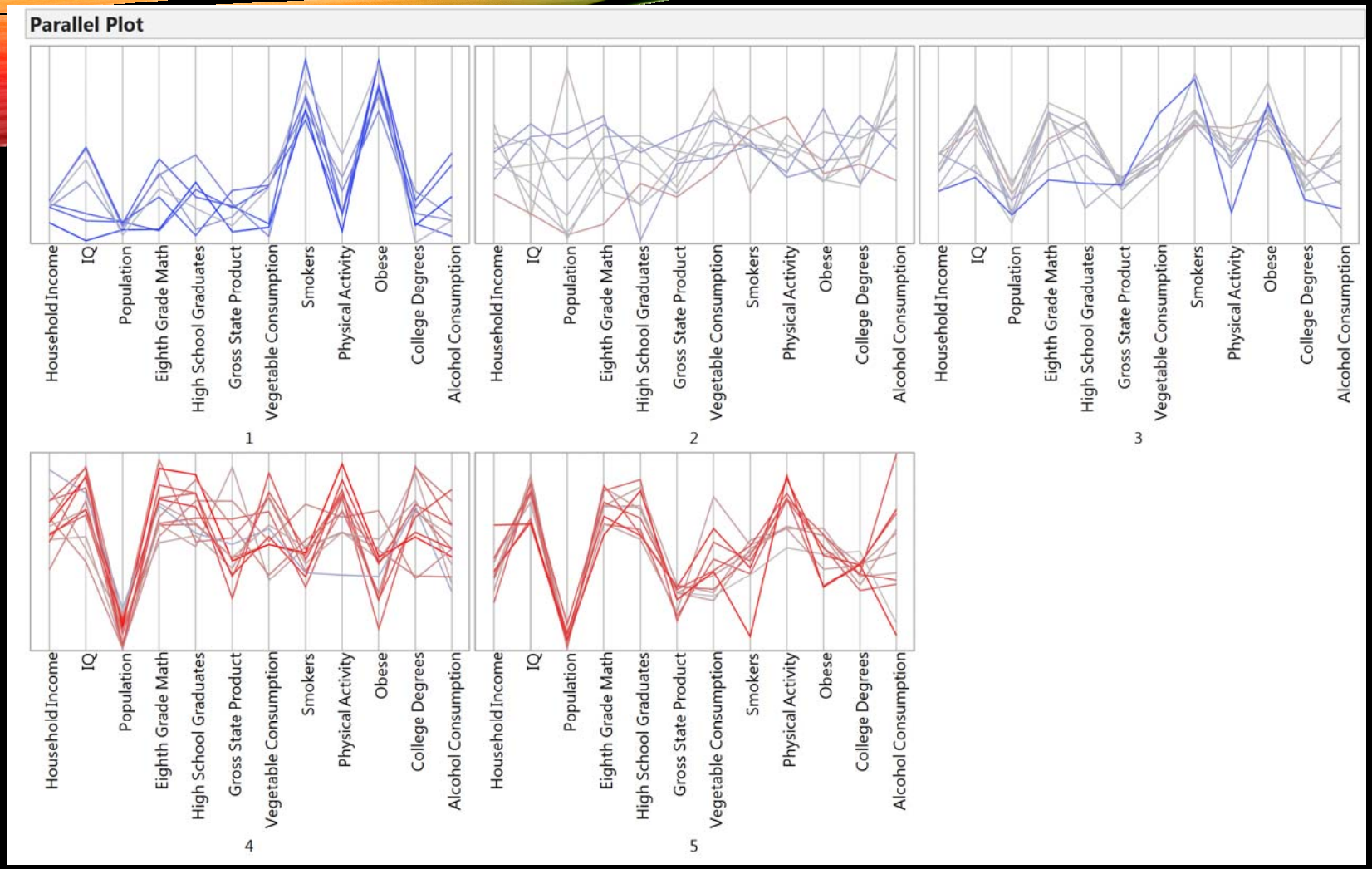
Courtesy of Mia Stephens: DSI 2014 Presentation

US DEMOGRAPHICS JMP HELP FILE

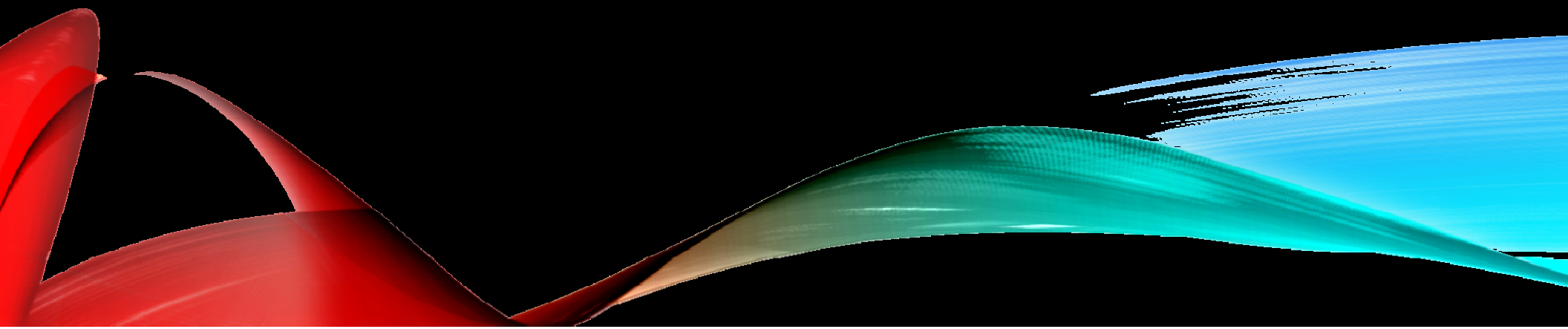

Clustering - JMP Pro
Finding points that are close, have similar values
Select Columns
18 Columns
State
Household Income
IQ
Region
Population
Eighth Grade Math
High School Graduates
Gross State Product
Vegetable Consumption
Smokers
Physical Activity
Obese
College Degrees
Alcohol Consumption
Latitude
Cast Selected Columns into Roles
Y, Columns
Household Income
IQ
Population
Eighth Grade Math
Ordering
optional numeric
Label
State
By
optional
Action
OK
Cancel
Remove
Recall
Help
Options
Hierarchical
Method
Average
Centroid
Ward
Single
Complete
Fast Ward
Standardize Data


Graph Builder
State colored by Cluster 2
Cluster 2
1
2
3
4
5
55° N
50° N
45° N
40° N
35° N
30° N
25° N
20° N
110° W
100° W
90° W
80° W

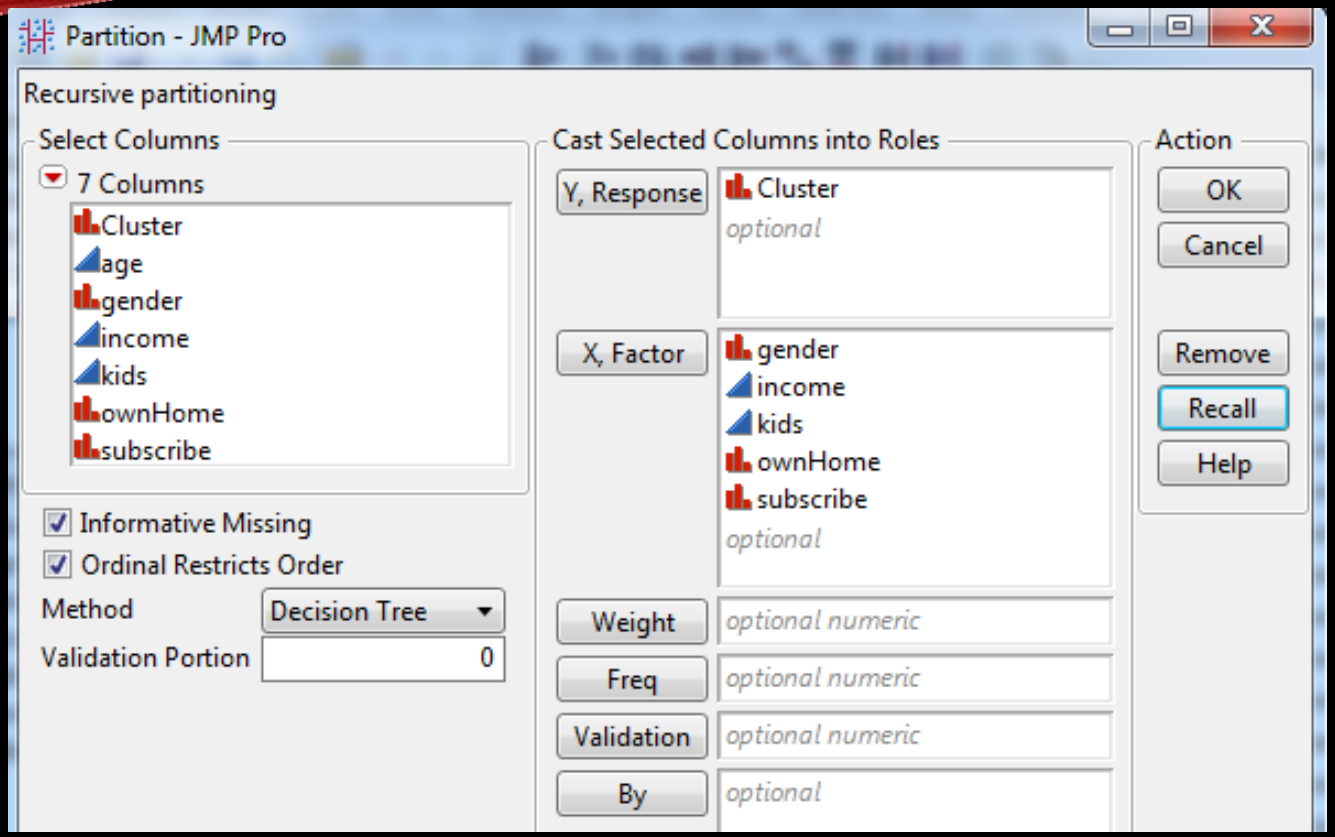

Constellation Plot
Y
X
10
5
0
-5
-10
-10
-5
0
5
10
15
Alabama
Louisiana
Arkansas
Wisconsin
Oregon
Montana
West Virginia
Mississippi
Maine
Kentucky
Arizona
South Dakota
Oklahoma
Rhode Island
North Dakota
Utah
Nebraska
Nevada
Iowa
New Mexico
Kansas
California
Idaho
Florida
New Jersey
Illinois
Virginia
New York
Maryland
Texas
Massachusetts
Connecticut
Georgia
Vermont
Delaware
Indiana
Tennessee
North Carolina
New Hampshire
South Carolina
Wyoming
Alaska
Washington
Michigan
Colorado
Minnesota
Ohio
Missouri
Pennsylvania


Parallel Plot
Household Income
IQ
Population
Eighth Grade Math
High School Graduates
Gross State Product
Vegetable Consumption
Smokers
Physical Activity
Obese
College Degrees
Alcohol Consumption
1
2
3
4
5

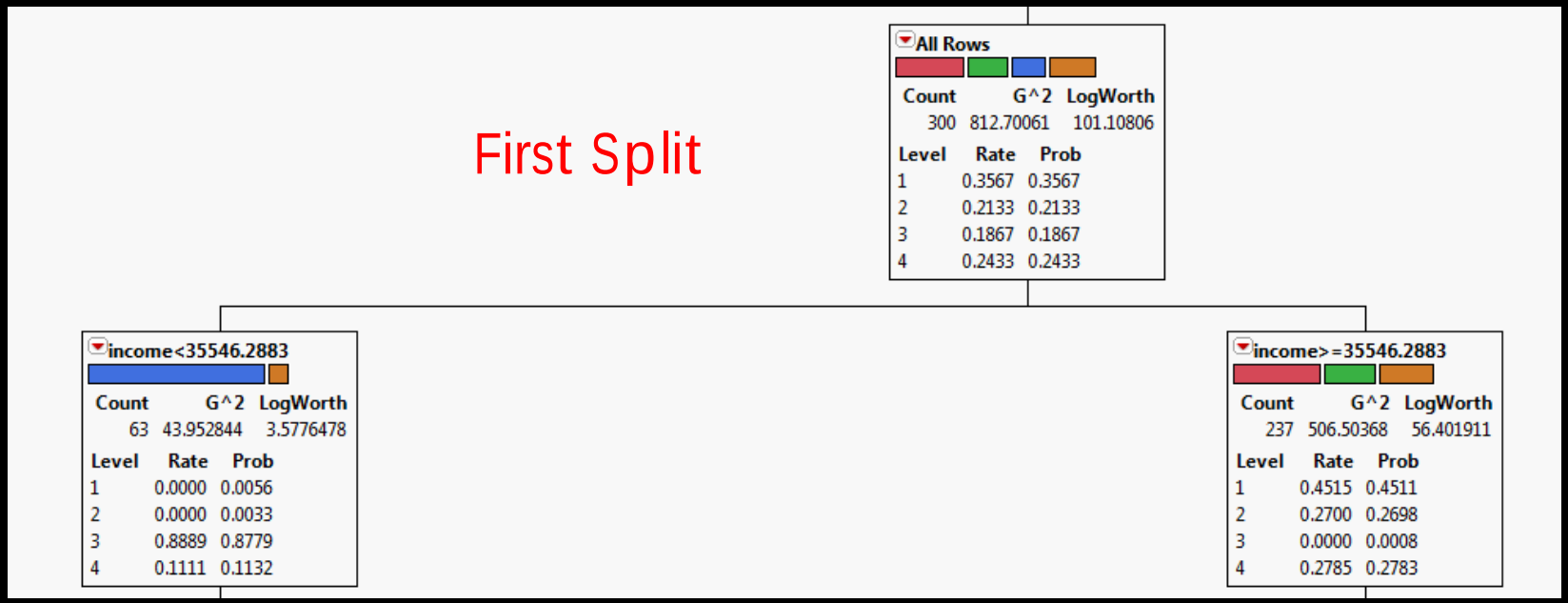
PARTITION USING EXAMPLE DATA SET

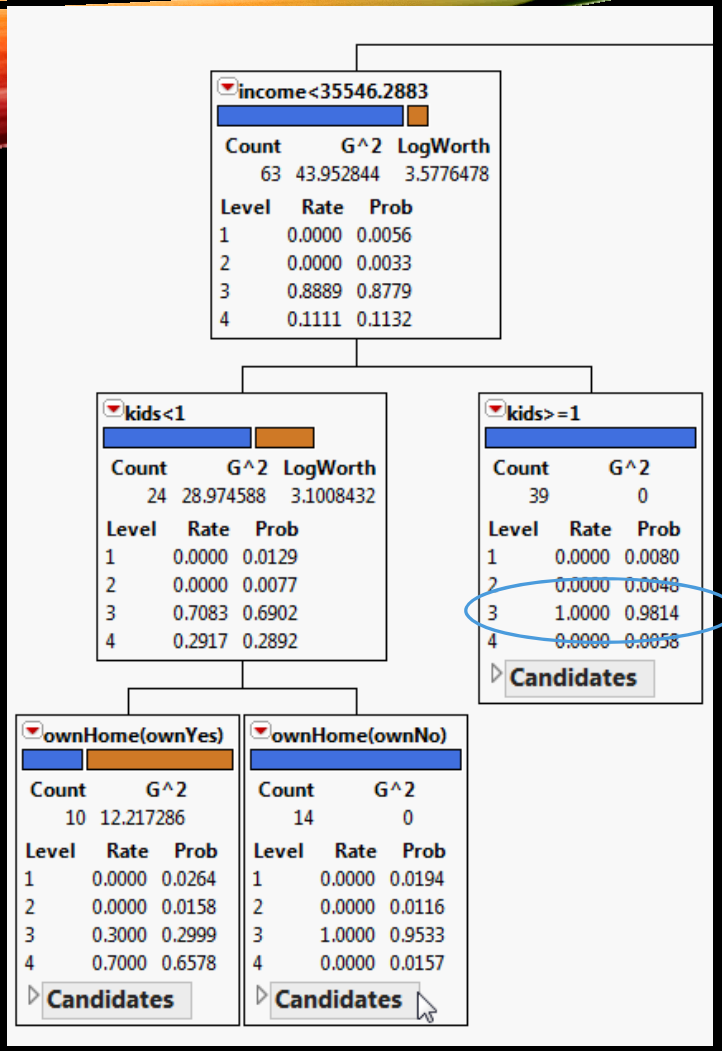


Analyze > Modeling > Partition



First Split


All Rows
Count G^2 LogWorth
300 812.70061 101.10806
Level Rate Prob
1 0.3567 0.3567
2 0.2133 0.2133
3 0.1867 0.1867
4 0.2433 0.2433
income<35546.2883
Count G^2 LogWorth
63 43.952844 3.5776478
Level Rate Prob
1 0.0000 0.0056
2 0.0000 0.0033
3 0.8889 0.8779
4 0.1111 0.1132
income>=35546.2883
Count G^2 LogWorth
237 506.50368 56.401911
Level Rate Prob
1 0.4515 0.4511
2 0.2700 0.2698
3 0.0000 0.0008
4 0.2785 0.2783

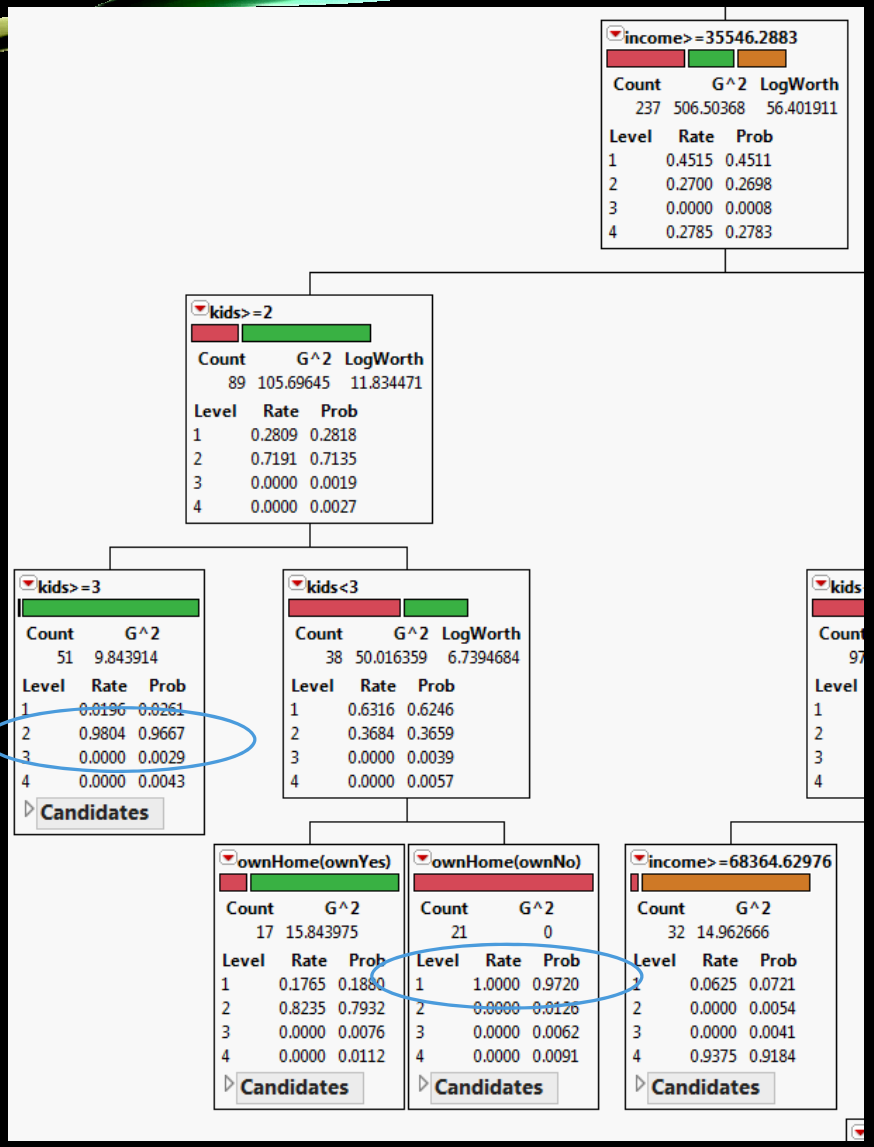


Down
Left
Branch

Cluster 3
Lower income
One or more children
Doesn't own home

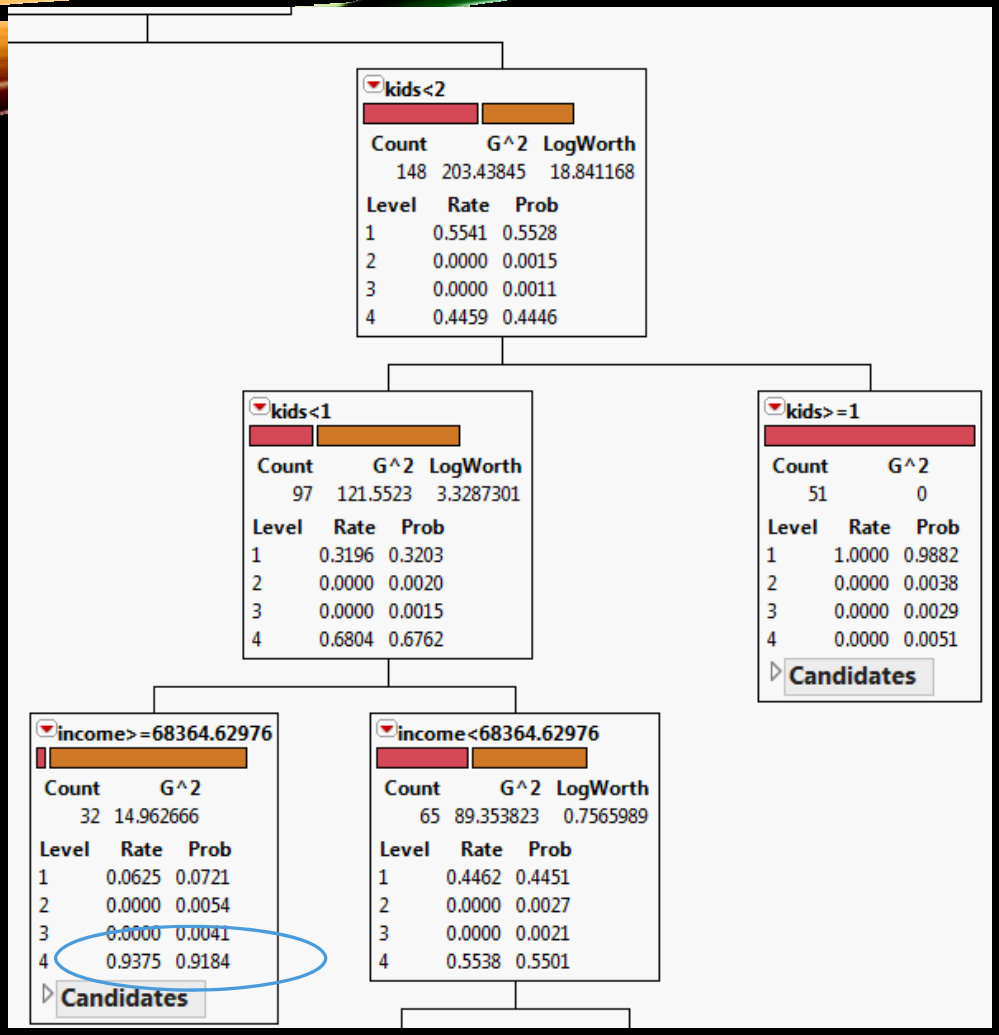
Cluster 1
Higher income
Less than 3 children
Own home

Cluster 2
Higher income
More than 3 children
Own home?



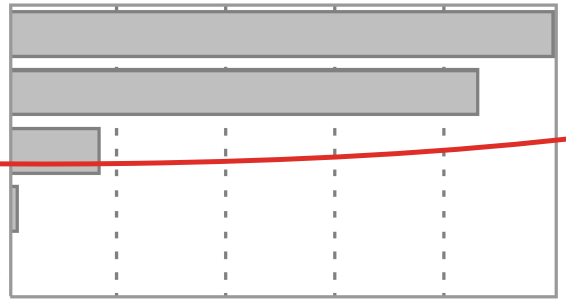
Down
Right
Branch

Cluster 4
Highest income
No children



Column Contributions

Term	Number of Splits	G^2	Portion
kids	4	340.069361	0.4891
income	4	294.468882	0.4235
ownHome	3	55.8843608	0.0804
subscribe	1	4.83988976	0.0070
gender	0	0	0.0000



Leaf Report

Response Prob

Leaf Label	1	2	3	4
income<35546.2883&kids<1&ownHome(ownYes)	0.0264	0.0158	0.2999	0.6578
income<35546.2883&kids<1&ownHome(ownNo)	0.0194	0.0116	0.9533	0.0157
income<35546.2883&kids>=1	0.0080	0.0048	0.9814	0.0058
income>=35546.2883&kids>=2&kids>=3&subscribe(subNo)	0.0069	0.9859	0.0029	0.0043
income>=35546.2883&kids>=2&kids>=3&subscribe(subYes)	0.2208	0.7230	0.0228	0.0334
income>=35546.2883&kids>=2&kids<3&ownHome(ownYes)	0.1880	0.7932	0.0076	0.0112
income>=35546.2883&kids>=2&kids<3&ownHome(ownNo)	0.9720	0.0126	0.0062	0.0091
income>=35546.2883&kids<2&kids<1&income>=68364.62976&income>=78019.71522	0.0174	0.0080	0.0062	0.9684
income>=35546.2883&kids<2&kids<1&income>=68364.62976&income<78019.71522	0.1677	0.0115	0.0088	0.8120
income>=35546.2883&kids<2&kids<1&income<68364.62976&ownHome(ownYes)&income>=60168.28251	0.0317	0.0120	0.0092	0.9470
income>=35546.2883&kids<2&kids<1&income<68364.62976&ownHome(ownYes)&income<60168.28251	0.4639	0.0047	0.0036	0.5279
income>=35546.2883&kids<2&kids<1&income<68364.62976&ownHome(ownNo)	0.6154	0.0064	0.0049	0.3733
income>=35546.2883&kids<2&kids>=1	0.9882	0.0038	0.0029	0.0051