Cluster Analysis

Summary

The **Cluster Analysis** procedure is designed to group observations or variables into clusters based upon similarities between them. The raw data for the procedure may be in either of two forms:

1. *n* rows or cases, each containing the values of *p* quantitative variables.
2. *n* rows and *n* columns if clustering observations or *p* rows and *p* columns if clustering variables, containing a measure of the "distance" between all pairs of items.

If raw data is input, the procedure will calculate distances between the observations or variables.

A number of different algorithms are provided for generating clusters. Some of the algorithms are *agglomerative*, beginning with separate clusters for each observation or variable and then joining clusters together based upon their similarity. Other methods begin with a set of seeds and match other cases or variables to those seeds.

The results of the analysis are displayed in several ways, including a dendogram, a membership table, and an icicle plot.

Sample StatFolio: *cluster.sgp*

Sample Data:

The file *cities.sgd* contains information of *n* = 10 large U.S. cities, obtained from www.city-data.com. The data consist of demographic, economic, and environmental variables. The table below shows a partial list of the data in that file:

City	*Population*	*Percent female*	*Land area*	*Median age*	*Median income*	*Highest temperature*
New York	8008278	52.6	303.3	34.2	38293	76.9
Boston	589141	51.9	48.4	31.1	39629	72.9
Chicago	2896016	51.5	227.1	31.5	38625	74.7
Washington	572059	52.9	61.4	34.6	40127	78.2
Atlanta	416474	50.4	131.7	31.9	34770	79.7
Los Angeles	3694820	50.2	469.1	31.6	36687	72.0
San Francisco	776733	49.2	46.7	36.5	55221	63.7
Miami	362470	50.3	35.7	37.7	23483	84.3
Houston	1953631	50.1	579.4	30.9	36616	53.0
Phoenix	1321045	49.1	474.9	30.7	41207	90.7

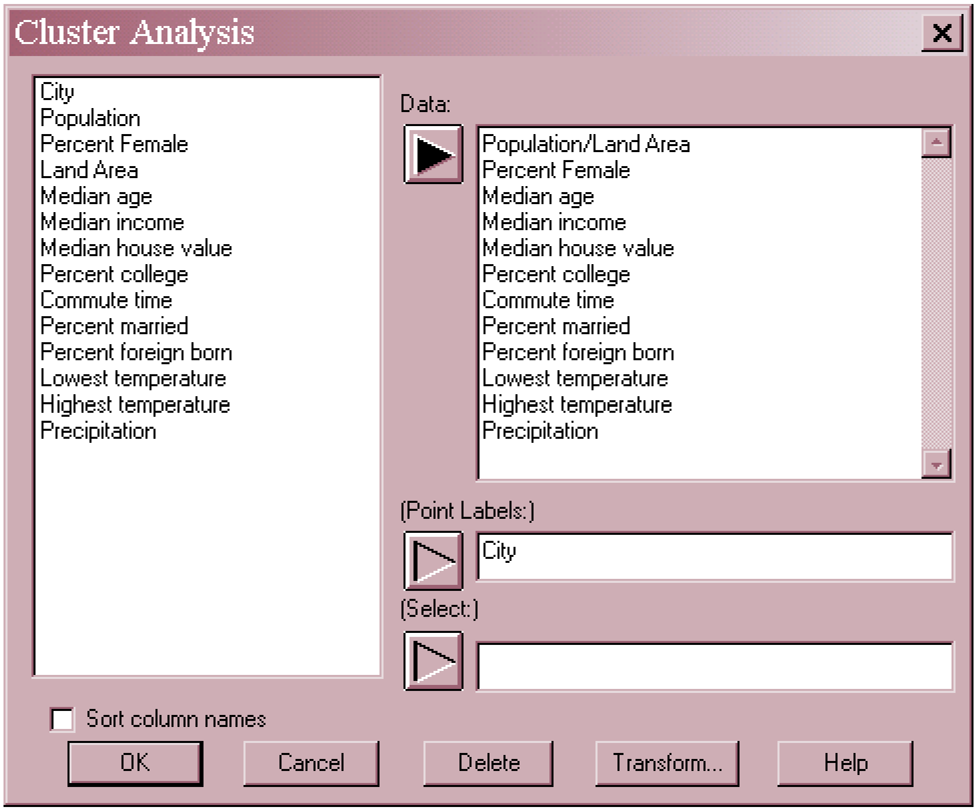
The cities will be clustered based upon the following *p* = 12 variables:

Population/Land Area
Percent Female
Median age
Median income
Median house value

Percent college
Commute time
Percent married
Percent foreign born
Lowest temperature
Highest temperature
Precipitation

Data Input

The data input dialog box requests the names of the columns containing the input data:



Data:
Population/Land Area
Percent Female
Median age
Median income
Median house value
Percent college
Commute time
Percent married
Percent foreign born
Lowest temperature
Highest temperature
Precipitation

(Point Labels:)
City

(Select:)

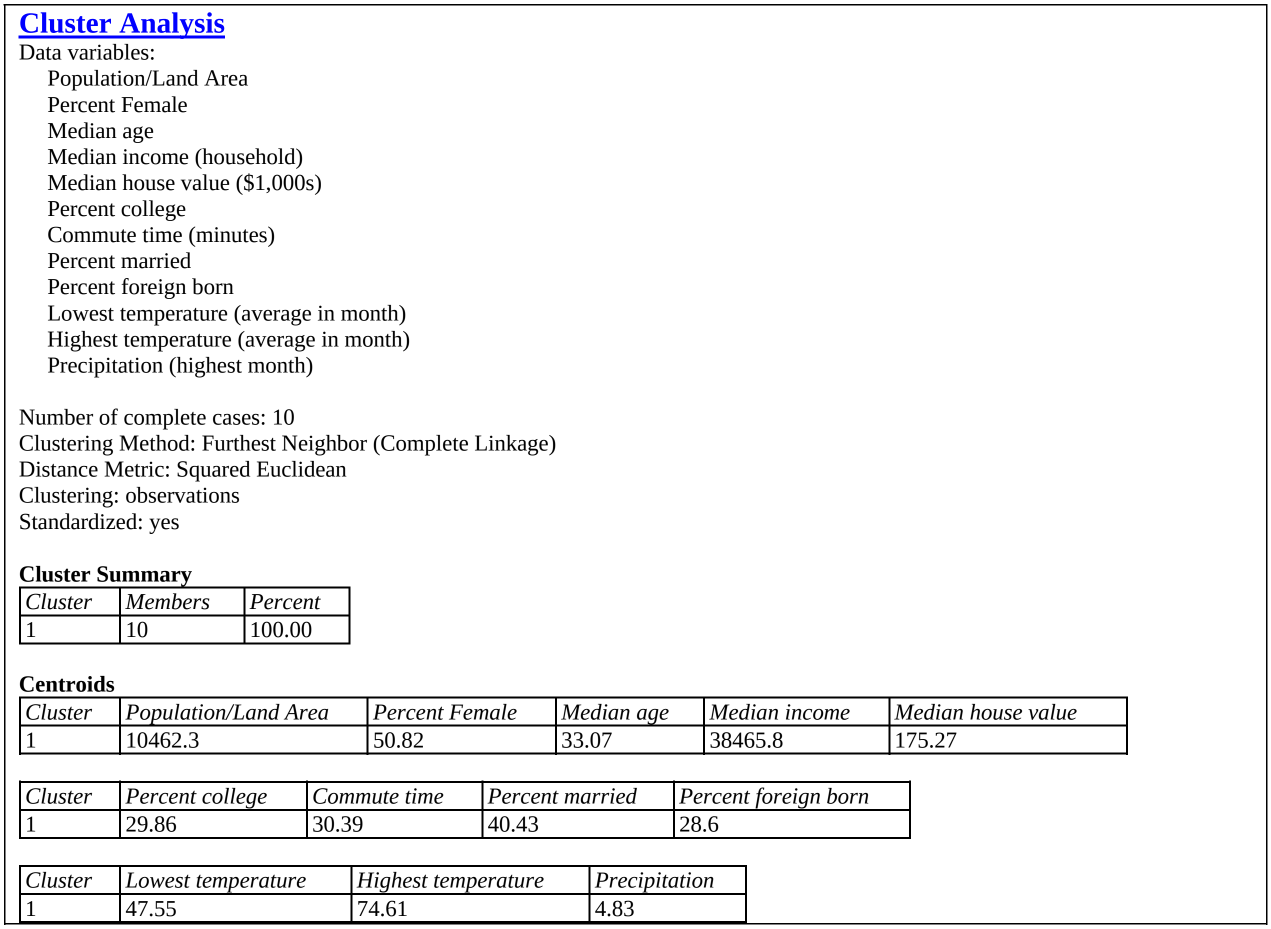
- **Data:** If observations are to be clustered, the names of *p* input variables containing the values for *n* cases, or an *n* by *n* matrix containing the distances between each case. If variables are to be clustered, the names of *p* input variables containing the values for *n* cases, or a *p* by *p* matrix containing the distance between each pair of variables.

- **Point labels:** optional labels for each row in the datasheet.

- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* summarizes the results of the clustering.



Cluster Analysis

Data variables:
Population/Land Area
Percent Female
Median age
Median income (household)
Median house value ($1,000s)
Percent college
Commute time (minutes)
Percent married
Percent foreign born
Lowest temperature (average in month)
Highest temperature (average in month)
Precipitation (highest month)

Number of complete cases: 10
Clustering Method: Furthest Neighbor (Complete Linkage)
Distance Metric: Squared Euclidean
Clustering: observations
Standardized: yes

Cluster Summary

Cluster	*Members*	*Percent*
1	10	100.00

Centroids

Cluster	*Population/Land Area*	*Percent Female*	*Median age*	*Median income*	*Median house value*
1	10462.3	50.82	33.07	38465.8	175.27

Cluster	*Percent college*	*Commute time*	*Percent married*	*Percent foreign born*
1	29.86	30.39	40.43	28.6

Cluster	*Lowest temperature*	*Highest temperature*	*Precipitation*
1	47.55	74.61	4.83

Included in the table are:

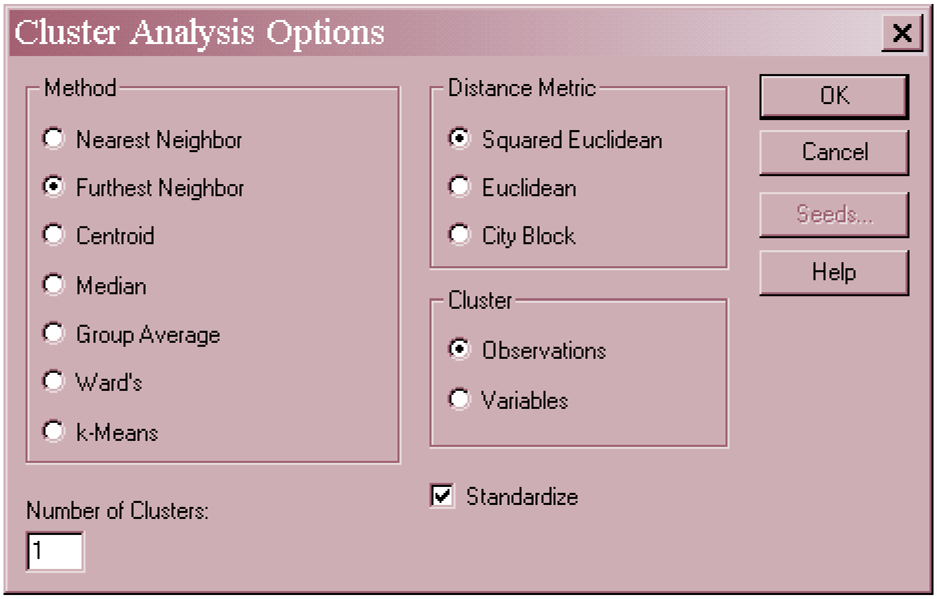
- **Input variables**: identification of the input variables.
- **Number of complete cases**: the number of cases *n* with information on all of the input variables. Any rows in the datasheet with missing values for any variable are excluded from the analysis.
- **Clustering Method**: the method used to derive the clusters (see the discussion below).
- **Distance Metric**: If the data consist of observations, the metric used to measure the distance between clusters. If a distance matrix has been entered, *User Matrix* will be indicated.
- **Clustering**: either *observations* or *variables*, depending on which items are being clustered.

- **Standardized**: whether the data have been standardized before the distances were calculated.

- **Cluster Summary**: the number of clusters created and the percentage of observations or variables placed into each cluster.

- **Centroids**: the average value for each variable in each cluster (if *observations* have been clustered).

Analysis Options



Cluster Analysis Options

Method
○ Nearest Neighbor
● Furthest Neighbor
○ Centroid
○ Median
○ Group Average
○ Ward's
○ k-Means

Distance Metric
● Squared Euclidean
○ Euclidean
○ City Block

Cluster
● Observations
○ Variables

☑ Standardize

Number of Clusters: 1

- **Method**: method used to create the clusters.

- **Number of Clusters**: the final number of clusters desired.

- **Distance Metric**: the metric used to measure distance between cases.

- **Cluster**: whether to generate clusters for observations or variables.

- **Standardize**: If checked, the variables will be standardized before doing the clustering. If clustering observations, each variable is standardized by subtracting its sample mean and then dividing by its sample standard deviation. If clustering variables, clustering is based on the sample correlation matrix rather than the sample covariance matrix.

- **Seeds Button**: when using the *k-means* method, displays a dialog box to input the k seeds.

Statistical Methodology

In order to create clusters of observations or variables, it is important to have a measure of "closeness" or "similarity" so that like items may be joined together. When observations are to be clustered, the closeness is typically measured by the distance between observations in the *p*-dimensional space of the variables. The *Cluster Analysis* procedure provides three different metrics for measuring the distance between two items, represented by *x* and *y*:

1. Squared Euclidian distance: $d(x,y) = \sum_{i=1}^{p} \left(x_i - y_i\right)^2$ (1)

2. Euclidian distance: $d(x,y) = \sqrt{\sum_{i=1}^{p} \left(x_i - y_i\right)^2}$ (2)

3. City Block distance: $d(x,y) = \sum_{i=1}^{p} \left|x_i - y_i\right|$ (3)

When clustering variables, distance is defined similarly except that *x* and *y* represent the location of two variables in the *n*-dimensional space of the observations, and the summation is over the observations instead of over the variables.

If some other distance metric is preferred, the user may input the distance matrix directly rather than inputting the original observations.

There are two basic type of methods provided to cluster items:

1. *Agglomerative hierarchical methods*: Agglomerative hierarchical clustering methods begin by placing each observation into a separate cluster. Clusters are then joined, two at a time, until the number of clusters is reduced to the desired target. At each stage, the clusters joined are the pair that are closest together.

2. *k-Means method*: This method begins by identifying *k* items as initial *seeds* for each cluster. Items are matched to the closest cluster.

Agglomerative Methods

The agglomerative methods begins by placing each item in a separate cluster and then combining clusters based on their distance from each other. The process continues until the desired number of clusters is formed. Where the methods differ is in how they define the distance between two clusters when one or both of the clusters contain more than one member:

1. *Nearest neighbor (single linkage)*: defines the distance between 2 clusters as the minimum distance between any member of one cluster and a member of the other.

2. *Furthest neighbor (complete linkage)*: defines the distance between 2 clusters as the maximum distance between any member of one cluster and a member of the other.

3. *Centroid*: defines the distance between 2 clusters as the distance between the centroids of each cluster, where the centroid is located at the average values of each variable over all members of the cluster.

4. *Median*: defines the distance between 2 clusters as the distance between the medians of each cluster, where the median is located at the median values of each variable over all members of the cluster.

5. *Group average (average linkage)*: defines the distance between 2 clusters as the average distance between all members of one cluster and all members of the other.

6. *Ward's method*: defines the distance between 2 clusters in terms of the increase in the sum of squared deviations around the cluster means that would occur if the two clusters were joined.

Method of k-Means

The method of *k-Means* works as follows:

1. *k* items are selected to be the initial *seeds* for the *k* desired clusters.

2. The remaining items are assigned to the cluster whose seed is closest to that item.

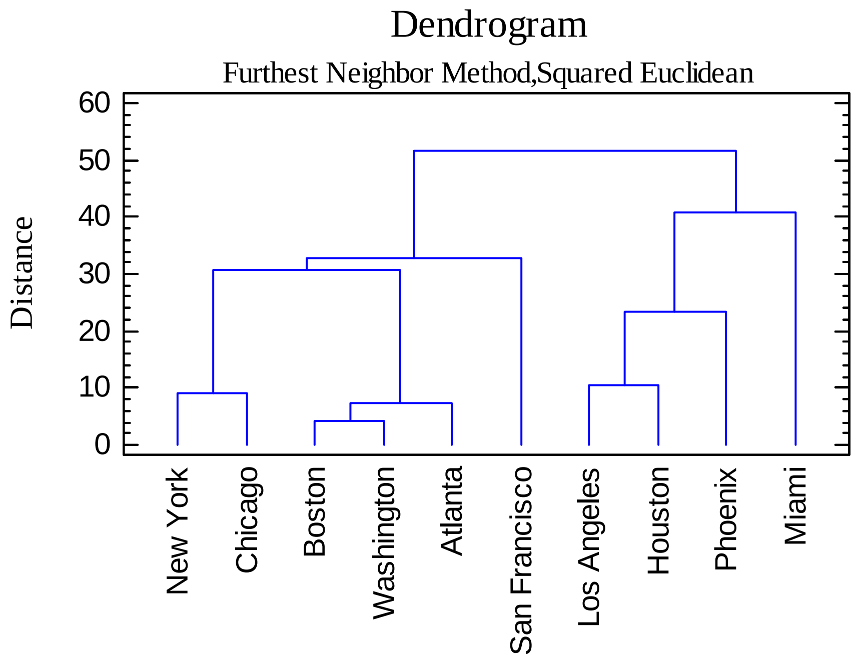
3. The centroids of each cluster are computed.

4. Each item is checked to determine whether it is closer to the centroid of another cluster than to the centroid of the cluster it is currently assigned to. If so, it is assigned to the other cluster and both centroids are recalculated.

5. Step 4 is repeated until no further changes take place.

Dendogram

The best way to view the output of a cluster analysis is usually by looking at the *Dendogram*:



Working from the bottom up, the dendogram shows the sequence of joins that were made between clusters. Lines are drawn connecting the clustered that are joined at each step, while the vertical axis displays the distance between the clusters when they were joined.

For example, the dendogram above shows the result of clustering the n = 10 cities in the example data file using the *furthest neighbor* method and *squared Euclidian distance*. At the start, each of the 10 cities forms a separate cluster. The first clusters joined were those containing *Boston* and *Washington*, at a distance of approximately 4. Next, *Atlanta* was joined to the cluster containing *Boston* and *Washington*. At the third step, *New York* and *Chicago* were joined into a single cluster, and then *Los Angeles* and *Houston* were joined. The procedure continued until a single cluster was formed.

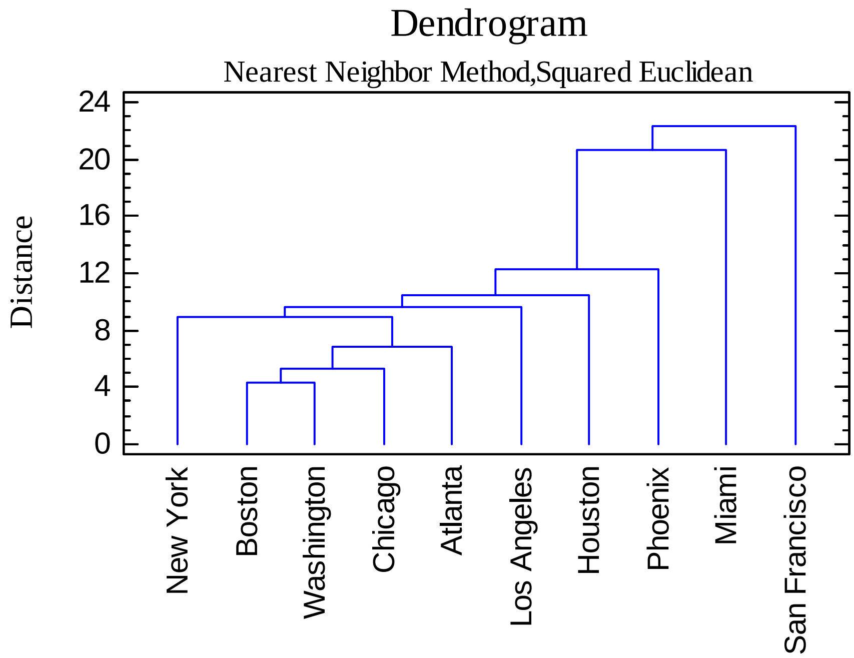
The general shape of the dendogram suggests grouping the cities into two groups:

Group #1: New York, Chicago, Boston, Washington, Atlanta and San Francisco.

Group #2: Los Angeles, Houston, Phoenix, and Miami.

Since Group #2 contains cities that tend to be located in warmer areas, it appears that climate plays an important role in grouping the cities when the *furthest neighbor* method is used.

A somewhat different grouping is obtained using the *nearest neighbor* method:



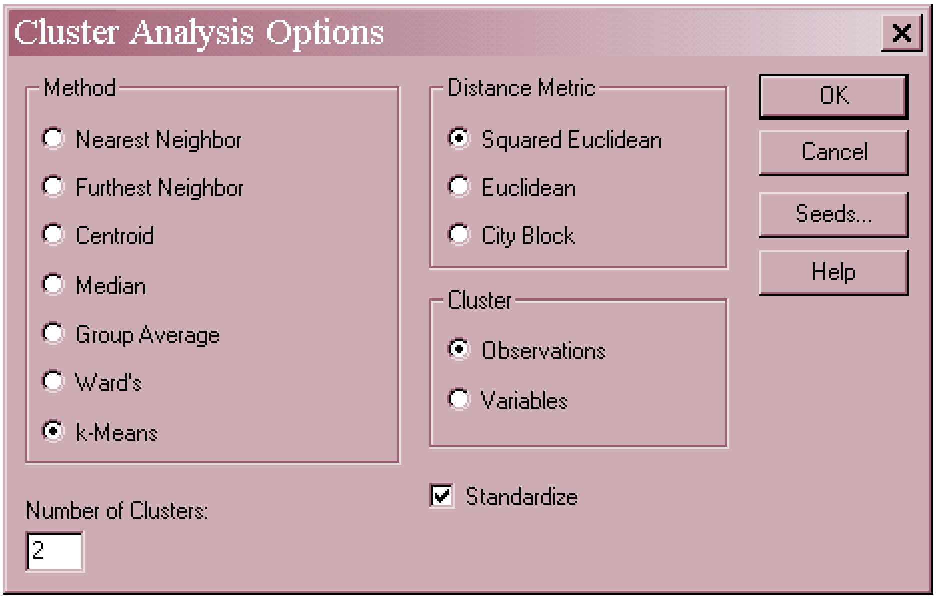
Particularly striking is the switch in location between Los Angeles and San Francisco. Los Angeles appears to join the other "big" cities sooner than with the previous method.

Membership Table

The *Membership Table* shows which observations or variables have been assigned to which clusters. Its use will be illustrated in the following example.

Example – Method of k-Means

Both of the methods used above indicate that *New York* and *Miami* are very different from each other. It is interesting to see what grouping would occur if one asked to create two clusters, using those two cities as seeds. To do so, access *Analysis Options*:



Method
- ○ Nearest Neighbor
- ○ Furthest Neighbor
- ○ Centroid
- ○ Median
- ○ Group Average
- ○ Ward's
- ● k-Means

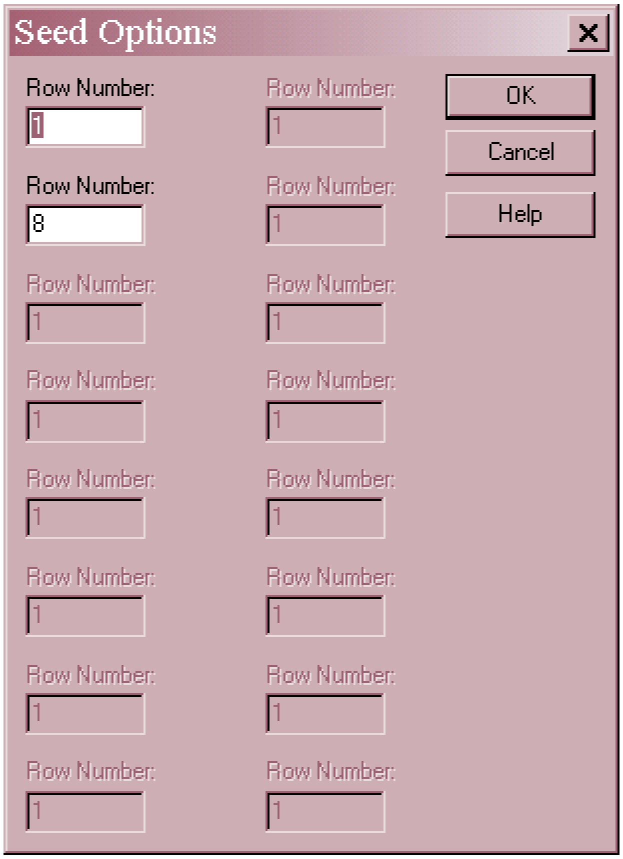
Distance Metric
- ● Squared Euclidean
- ○ Euclidean
- ○ City Block

Cluster
- ● Observations
- ○ Variables

☑ Standardize

Number of Clusters: 2

Select *k-Means* and set the *Number of Clusters* to 2. Then press the *Seeds* button and enter the row numbers of *New York* and *Miami*:



Row Number: 1	Row Number: 1
Row Number: 8	Row Number: 1
Row Number: 1	Row Number: 1
Row Number: 1	Row Number: 1
Row Number: 1	Row Number: 1
Row Number: 1	Row Number: 1
Row Number: 1	Row Number: 1
Row Number: 1	Row Number: 1

Press *OK* twice to generate the analysis. Although the dendogram is not available when using the *k-Means* method (since the clustering is not hierarchical), the *Membership Table* shows the final assignment of clusters:

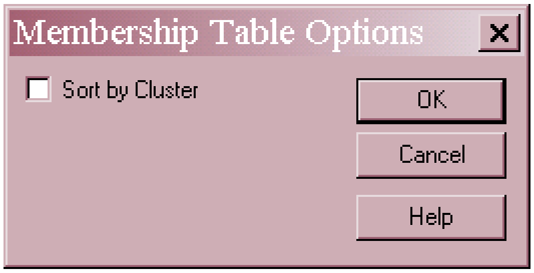
Membership Table
Clustering Method: k-Means

Distance Metric: Squared Euclidean

Row	*Label*	*Cluster*
1	New York	1
2	Boston	1
3	Chicago	1
4	Washington	1
5	Atlanta	1
6	Los Angeles	1
7	San Francisco	1
8	Miami	2
9	Houston	2
10	Phoenix	1

The only city that is placed with Miami is Houston. All of the others fall in the cluster with New York.

Pane Options



□ Sort by Cluster

Select *Sort By Cluster* to sort the items by cluster number.

Icicle Plot

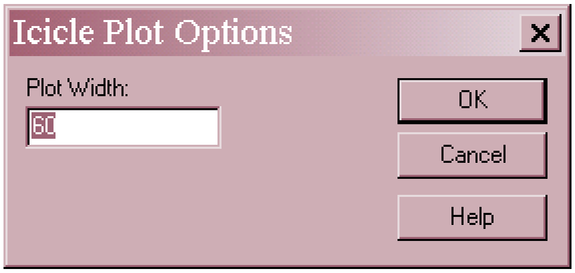
The *Icicle Plot* provides an additional way to illustrate the clustering that has occurred. It is most useful when the number of items is small:

Icicle Plot
Clustering Method: Furthest Neighbor (Complete Linkage)
Distance Metric: Squared Euclidean

```
                  Number of Clusters
                           1
Label      Row    1234567890
---------- ------ ----------
New York   1      XXXXXXX
                  XXXXXXX
Chicago    3      XXXXXXX
                  XXXX
Boston     2      XXXXXXXXX
                  XXXXXXXXX
Washington 4      XXXXXXXXX
                  XXXXXXXX
Atlanta    5      XXXXXXXX
                  XXX
San Franci 7      XXX
                  X
Los Angele 6      XXXXXX
                  XXXXXX
Houston    9      XXXXXX
                  XXXXX
Phoenix    10     XXXXX
                  XX
Miami      8      XX
```

Under each *Number of Clusters* is a row of X’s. Any items connected by contiguous X’s are contained in the same cluster. For example, the row beneath the “2” illustrates that when the cities were split into 2 clusters, the clusters consisted of the first 6 cities and the last 4 cities.

Pane Options



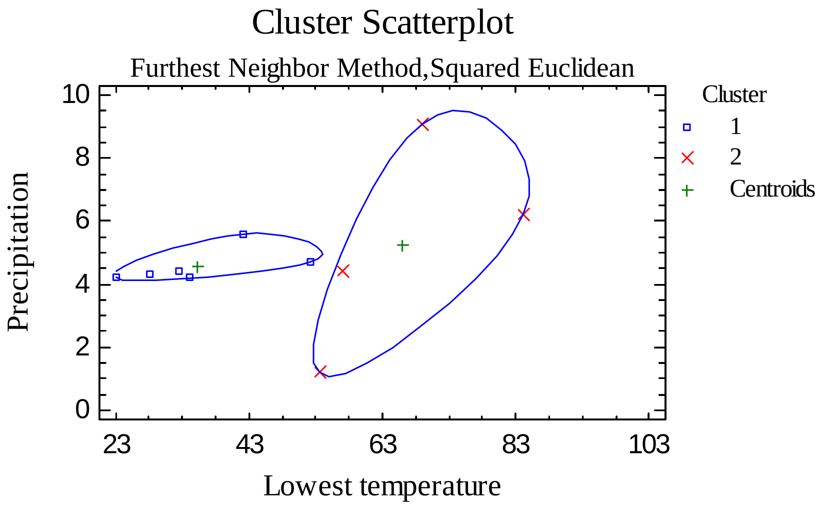
Icicle Plot Options

Plot Width: 60

- **Plot Width**: the maximum number of characters to be displayed across a single page.

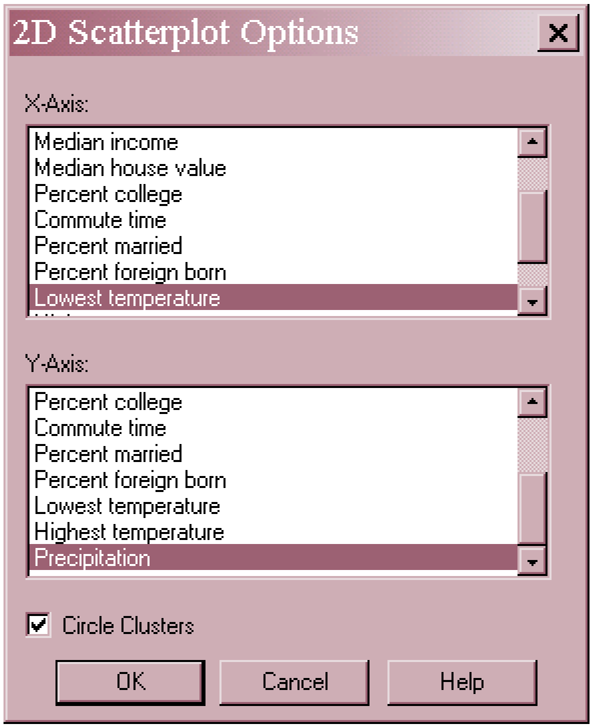
2D Scatterplot

The *2D Scatterplot* shows the clustering with respect to any two of the input variables:



Each observation in the datasheet is plotted, together with the centroids of the clusters. If desired, a spline may be used to connect observations on the borders of each cluster. In the sample data, the clusters are fairly well separated in the space of *Lowest temperature* and *Precipitation*.

Pane Options



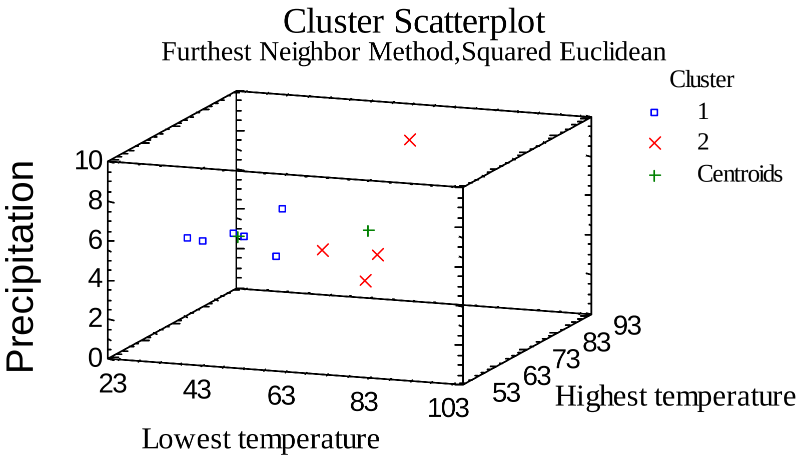
2D Scatterplot Options

☑ Circle Clusters

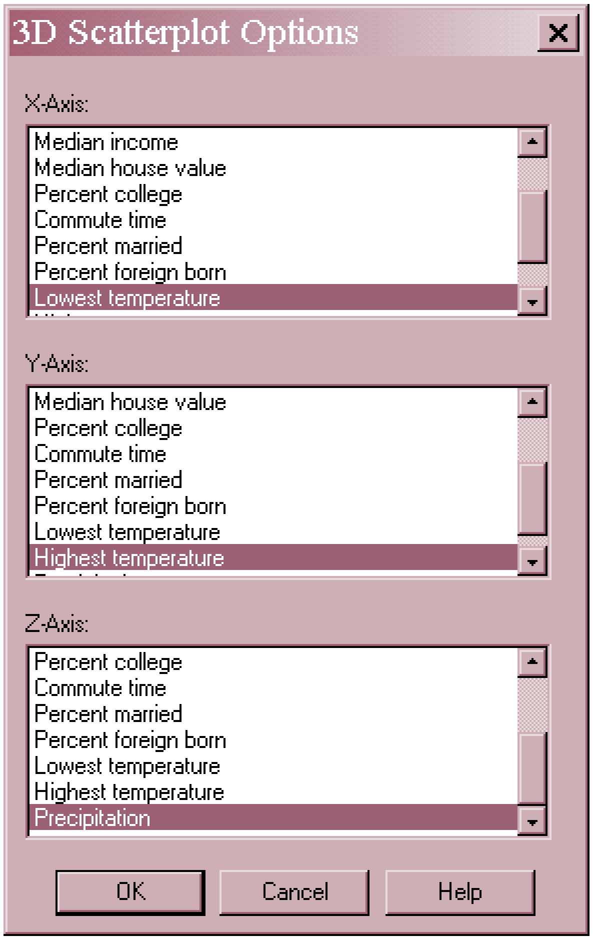
- **X and Y Axes:** the variables to be plotted on the horizontal and vertical axes.
- **Circle Clusters:** if checked, a spline will be used to connect the observations around the border of each cluster.

3D Scatterplot

The *3D Scatterplot* shows the clustering with respect to any three of the input variables:



Pane Options



3D Scatterplot Options

X-Axis:

Median income
Median house value
Percent college
Commute time
Percent married
Percent foreign born
Lowest temperature

Y-Axis:

Median house value
Percent college
Commute time
Percent married
Percent foreign born
Lowest temperature
Highest temperature

Z-Axis:

Percent college
Commute time
Percent married
Percent foreign born
Lowest temperature
Highest temperature
Precipitation

- **X, Y and Z Axes:** the variables to be plotted on the three axes.

Agglomeration Schedule

The *Agglomeration Schedule* provides a summary of each step in an agglomerative clustering algorithm:

Agglomeration Schedule
Clustering Method: Furthest Neighbor (Complete Linkage)
Distance Metric: Squared Euclidean

	Combined	*Combined*		*Previous Stage*	*Previous Stage*	*Next*
Stage	*Cluster 1*	*Cluster 2*	*Distance*	*Cluster 1*	*Cluster 2*	*Stage*
1	2	4	4.33537	0	0	2
2	2	5	7.24389	1	0	6
3	1	3	8.94147	0	0	6
4	6	9	10.4417	0	0	5
5	6	10	23.4536	4	0	8
6	1	2	30.6609	3	2	7
7	1	7	32.8948	6	0	0
8	6	8	40.9814	5	0	0

Cluster	*Smallest*
Number	*Row*
1	1
2	6

The top section of the output shows:

- **Stage**: the step number in the algorithm.

- **Combined Clusters**: the numbers of the observations or variables combined at each stage. For example, at stage 1, cities #2 and #4 were combined to form a single cluster. That cluster retains the smaller of the two numbers for the combined clusters (i.e., “2”). At the second stage, the cities in cluster 2 were combined with city #5.

- **Distance**: the distance between the clusters when they were joined.

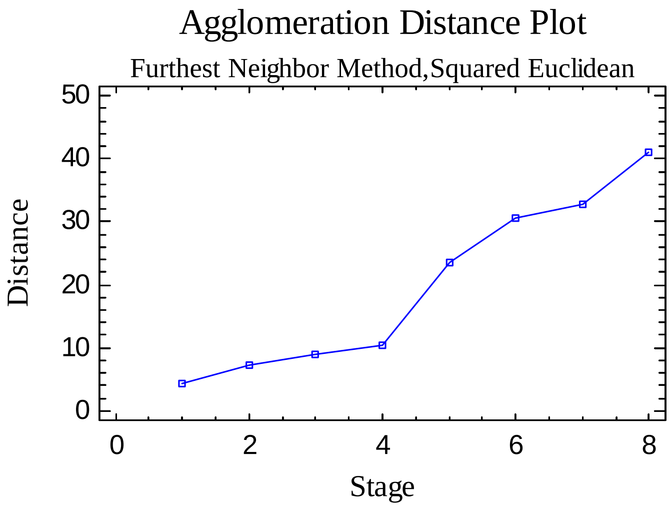
- **Previous Stage**: the stage number in which each cluster last appeared, or 0 if it has not been joined to any cluster in an earlier stage.

- **Next Stage**: the next stage in which the newly created cluster appears.

The bottom section of the output displays the smallest row number in the datasheet among the members of each cluster.

Agglomeration Distance Plot

The *Agglomeration Distance Plot* shows the minimum distance between clusters when they were combined:



Notice in the sample data that the distances through stage 4 are all quite small. The first four joins evidently happen between cities that are very similar to each other:

Stages 1 and 2: Boston, Washington and Atlanta
Stage 3: New York and Chicago
Stage 4: Los Angeles and Houston

After that, the clusters combined are at considerably further distances from one another.

The agglomeration distance plot can be helpful in determining how many natural clusters exist in the data.

Save Results

The following results may be saved to the datasheet:

1. *Cluster numbers*: the cluster numbers assigned to the data in each row of the input variables.

2. *Distance matrix*: the derived distance matrix between items being clustered.