Chapter 446

K-Means Clustering

Introduction

The k-means algorithm was developed by J.A. Hartigan and M.A. Wong of Yale University as a partitioning technique. It is most useful for forming a small number of clusters from a large number of observations. It requires variables that are continuous with no outliers. Discrete data can be included but may cause problems.

The objective of this technique is to divide *N* observations with *P* dimensions (variables) into *K* clusters so that the within-cluster sum of squares is minimized. Since the number of possible arrangements is enormous, it is not practical to expect the best solution. Rather, this algorithm finds a "local" optimum. This is a solution in which no movement of an observation from one cluster to another will reduce the within-cluster sum of squares. The algorithm may be repeated several times with different starting configurations. The optimum of these cluster solutions is then selected.

Technical Details

The k-means clustering algorithm is popular because it can be applied to relatively large sets of data. The user specifies the number of clusters to be found. The algorithm then separates the data into spherical clusters by finding a set of cluster centers, assigning each observation to a cluster, determining new cluster centers, and repeating this process.

Assume that you have *N* rows (observations), which are separated into *K* groups. The k^{th} cluster contains n_k observations. Each row consists of *P* variables. A missing value in the i^{th} variable of the j^{th} row of the k^{th} group is designated by δ_{ijk}.

The data are standardized by subtracting the variable mean and dividing by the standard deviation. The standardized data elements are referred to as z_{ij}.

Cluster Initialization

The method of initializing the clusters influences the final cluster solution. For each trial, **NCSS** randomly assigns each point to a cluster. This configuration is optimized using the k-means algorithm. Trying several random starting configurations will greatly increase the probability of finding the global optimum solution for a particular number of clusters.

Goodness-of-Fit Criterion

The goodness-of-fit criterion used to compare various cluster configurations is based on the within-cluster sum of squares, WSS_K, where

$$WSS_K = \left(\frac{NP}{NP - m}\right) \sum_{k=1}^{K} \sum_{i=1}^{P} \sum_{j=1}^{n_k} (1 - \delta_{ijk})(z_{ij} - c_{ik})^2$$

where c_{ik} is the average (center) value of the i^{th} variable in the k^{th} cluster.

The percent of variation is defined as

$$PV_K = 100\frac{WSS_K}{WSS_1}$$

Data Structure

The data given in the following table contain information on twelve of the most famous superstars in basketball. The stats are on a per game basis for games played through the 1989 season.

BBall dataset (subset)

Player	Height	FgPct	Points	Rebounds
Jabbar K.A.	86.0	55.9	24.6	11.2
Barry R	79.0	44.9	23.2	6.7
Baylor E	77.0	43.1	27.4	13.5
Bird L	81.0	50.3	25	10.2
Chamberlain W	85.0	54.0	30.1	22.9
Cousy B	72.5	37.5	18.4	5.2
Erving J	78.5	50.6	24.2	8.5
Johnson M	81.0	53.0	19.5	7.4

Missing Values

You control the fate of observations with missing values by setting a percent-missing parameter. Observations with more than the specified percentage of missing values are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Cluster Variables

Designates the variables to be clustered. Note that the k-means algorithm assumes that all of your variables are continuous with no outliers. If your data do not meet these requirements, use caution when applying this technique.

Label Variable

An optional variable containing row labels that you may want to use to document your output. You can use dates (like Jan-23-95) as labels. Here is how. First, enter your dates using the standard date format (like 06/20/93). In the Variable Info screen, change the format of the date variable to something like *mmm-dd-yyyy* or *mm-dd-yy*. The labels will be displayed as labels. Without changing the variable format, the dates will be displayed as long integer values.

Cluster Options

Minimum and Maximum Clusters

These options specify a minimum and maximum number of clusters to try. Although the k-means algorithm finds a cluster configuration for a fixed number of clusters, ***NCSS*** lets you specify a range of values to try for the number of clusters. Various goodness-of-fit tests help you determine the optimum number of clusters.

Often, values between two and five are used here, although your data might require more.

Reported Clusters

This is the number of clusters to use for reporting purposes. This is the so-called "optimum" number of clusters. Usually, you will have to make two passes through your data. On the first pass, you will determine the optimum number of clusters. On the second pass, you will obtain the information about the clusters.

Other Options

Random Starts

The first box specifies the number of random initial configurations to try for each value between the minimum and maximum cluster range. Since the k-means algorithm finds a local optimum, it is thought that trying several random, initial configurations will lead to the global optimum (or near optimum). Of course, as this value is increased, the program's running time also increases.

Max Iterations

This option specifies the maximum number of retries before the algorithm is aborted.

Percent Missing

An observation with missing values may be clustered by using only the non-missing data. This option specifies the percentage of missing values to allow in an observation before it is skipped. For example, an observation with five variables, with two values missing, would be 60% complete. If the value of this option were 50, this observation would be kept, while an observation with three missing values would be skipped.

Reports Tab

The following options control the format of the reports.

Select Reports

Minimum Iteration Report – Distance by Cluster Report

These options specify which reports are displayed.

Report Options

Precision

This allows you to specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only**.**

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Bivariate Plots Tab

These options control the attributes of the bivariate plots.

Bivariate Plot Format

Bivariate Plots

Specify whether to display the indicated plots. Click the plot format button to change the plot settings.

Show Row Numbers

Check to show the row numbers next to the points on the plots. Data point labels must also be checked on the plot format.

Show Row Labels

Check to show the row labels next to the points on the plots. Data point labels must also be checked on the plot format.

Storage Tab

These options let you specify where to store various row-wise statistics.

Storage Variable

Store Cluster Id in Variable

You can automatically store the cluster identification number of each row into the variable specified here. The configuration stored is for the number of clusters specified in the Reported Clusters option.

Warning: Any data already in this variable is replaced by the cluster number. Be careful not to specify variables that contain important data.

Example 1 – K-Means Clustering

This section presents an example of how to run a K-Means cluster analysis. The data used are shown above and found in the BBall dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the K-Means Clustering window.

1 Open the BBall dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **BBall.NCSS**.
- Click **Open**.

2 Open the K-Means Clustering window.

- Using the Analysis menu or the Procedure Navigator, find and select the **K-Means Clustering** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the K-Means Clustering window, select the **Variables tab**.
- Double-click in the **Cluster Variables** box. This will bring up the variable selection window.
- Select **Height, FgPct, Points, Rebounds** from the list of variables and then click **Ok**. "Height, FgPct, Points, Rebounds" will appear in the Cluster Variables box.
- Double-click in the **Label Variable** box. This will bring up the variable selection window.
- Select **Player** from the list of variables and then click **Ok**. "Player" will appear in the Label Variable box.
- Enter **4** for the **Maximum Clusters**.

4 Specify the report.

- On the K-Means Clustering window, select the **Reports tab**.
- All reports and plots should be selected.

5 Specify the plots.

- On the K-Means Clustering window, select the **Plots tab**.
- Check the **Bivariate Plots** checkbox.
- Click on the plot format button and check the **Labels** checkbox under **Data Point Labels**.
- Uncheck **Show Row Numbers**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Minimum Iteration Section (Results may vary)

Minimum Iteration Section

Iteration No.	No. of Clusters	Percent of Variation	Bar Chart of Percent
2	2	66.09	||||||||||||||||||
4	3	46.48	|||||||||||||
8	4	29.17	|||||||||

This report is to help you determine the optimum number of clusters. The results may vary slightly from those shown.

Iteration No.

The iteration number from the Iteration Report.

No. of Clusters

The number of clusters reported on.

Percent of Variation

This gives the within sum of squares for the number of clusters reported on in this line as a percentage of the within sum of squares with no clustering. As more and more clusters are added, this value should fall. Select as the optimum number of clusters the point where this percentage fails to decrease dramatically.

Bar Chart of Percent

This gives a visual display of the Percent of Variation values.

Iteration Section

Minimum Iteration Section

Iteration No.	No. of Clusters	Percent of Variation	Bar Chart of Percent
1	2	72.02716	|||||||||||||||||||
2	2	66.08894	|||||||||||||||||
3	2	70.51944	|||||||||||||||||||
4	3	46.47721	||||||||||||
5	3	47.91439	|||||||||||||
6	3	46.47721	||||||||||||
7	4	31.81219	||||||||
8	4	29.17248	||||||||
9	4	32.96159	|||||||||

This report is especially useful in helping you determine if you have selected enough random starting configurations. If you have specified enough starting configurations, two or three of them will be optimum (minimum percent variation) for each number of clusters. If this does not occur, you should increase the number of random starting configurations (Initial Configurations) and re-run the problem.

Iteration No.

The iteration number reported on this line.

No. of Clusters

The number of clusters in this configuration.

Percent of Variation

This gives the within sum of squares for the number of clusters reported on in this line as a percentage of the within sum of squares with no clustering. As more and more clusters are added, this value should fall. Select as the optimum number of clusters the point where this percentage fails to decrease dramatically.

Bar Chart of Percent

This gives a visual display of the Percent of Variation values.

Cluster Means

Cluster Means

Variables	Cluster1	Cluster2	Cluster3
Height	78.25	85.5	77
FGPct	48.6375	54.95	40.75
Points	25.575	27.35	16.75
Rebounds	8.225	17.05	13.9
Count	8	2	2

This report shows the means of each of the variables across each of the clusters. The last row shows the *count* or number of observations in the cluster.

Cluster Standard Deviations

Cluster Standard Deviations

Variables	Cluster1	Cluster2	Cluster3
Height	2.171241	0.7071068	6.363961
FGPct	3.357694	1.343503	4.596194
Points	3.770089	3.889087	2.333452
Rebounds	2.544321	8.273149	12.30366
Count	8	2	2

This report shows the standard deviations of each of the variables across each of the clusters. The last row shows the count (number of observations) in the cluster.

F-Ratio Section

F-Ratio Section

Variables	DF1	DF2	Between Mean Square	Within Mean Square	F-Ratio	Prob Level
Height	2	9	48.125	8.222222	5.85	0.023532
FGPct	2	9	101.6469	11.31653	8.98	0.007170
Points	2	9	72.7475	13.34056	5.45	0.028096
Rebounds	2	9	75.04459	29.46	2.55	0.132844

This report summarizes the results of performing a one-way ANOVA on each variable, using the currently defined clusters as the factor. This report helps you investigate the importance of each variable in the clustering process.

Caution should be used with this report since it ignores the correlation that exists among the variables. A better approach to reducing the number of variables would be to save the cluster configuration and run a Discriminant Analysis with variable selection, since this would account for the correlation among the variables.

Distance Section

Distance Section

Row	Cluster	Dist1	Dist2	Dist3
1 Jabbar K.A.	2	2.4609	1.1263	4.0315
2 Barry R	1	0.9139	3.1499	1.9940
3 Baylor E	1	1.4427	3.1724	2.2139
4 Bird L	1	0.8398	1.8867	2.7392
5 Chamberlain W	2	3.2456	1.1263	4.4712
6 Cousy B	3	2.9971	5.3790	1.9512
7 Erving J	1	0.4724	2.4891	2.5912
8 Johnson M	1	1.6497	2.5426	2.8064
9 Jordan M	1	1.5532	2.8939	4.0067
10 Robertson O	1	0.3409	2.9490	2.5629
11 Russell B	3	3.3878	3.5197	1.9512
12 West J	1	1.0971	3.6374	2.8439

This report displays the relative distance of each row to the cluster centers. It is provided to help determine how sharp the clustering has been. If the distance from each point to its designated center is much less than the distance from the point to the other centers, the cluster configuration does a good job of clustering. However, if the smallest distance is close in value to the distance to one of the other clusters, there is ambiguity as to which cluster the point belongs. Such a solution is not as desirable.

Individual Distance Section

Distance Section for Cluster 1

Row	Cluster	Dist1	Dist2	Dist3
2 Barry R	1	0.9139	3.1499	1.9940
3 Baylor E	1	1.4427	3.1724	2.2139
4 Bird L	1	0.8398	1.8867	2.7392
7 Erving J	1	0.4724	2.4891	2.5912
8 Johnson M	1	1.6497	2.5426	2.8064
9 Jordan M	1	1.5532	2.8939	4.0067
10 Robertson O	1	0.3409	2.9490	2.5629
12 West J	1	1.0971	3.6374	2.8439

Count = 8

Distance Section for Cluster 2

Row	Cluster	Dist1	Dist2	Dist3
1 Jabbar K.A.	2	2.4609	1.1263	4.0315
5 Chamberlain W	2	3.2456	1.1263	4.4712

Count = 2

Distance Section for Cluster 3

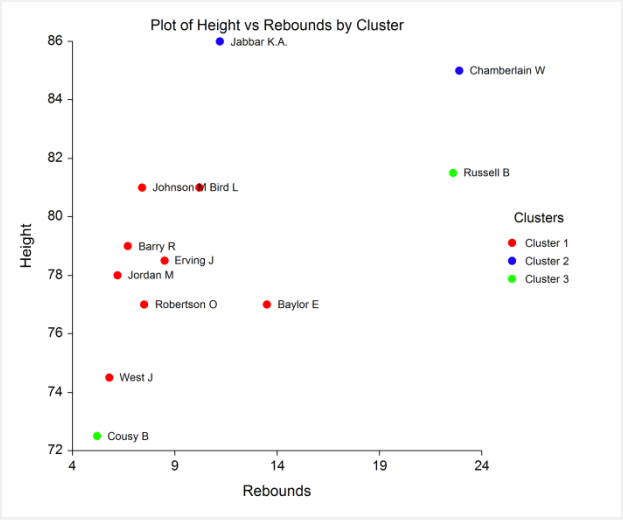
Row	Cluster	Dist1	Dist2	Dist3
6 Cousy B	3	2.9971	5.3790	1.9512
11 Russell B	3	3.3878	3.5197	1.9512

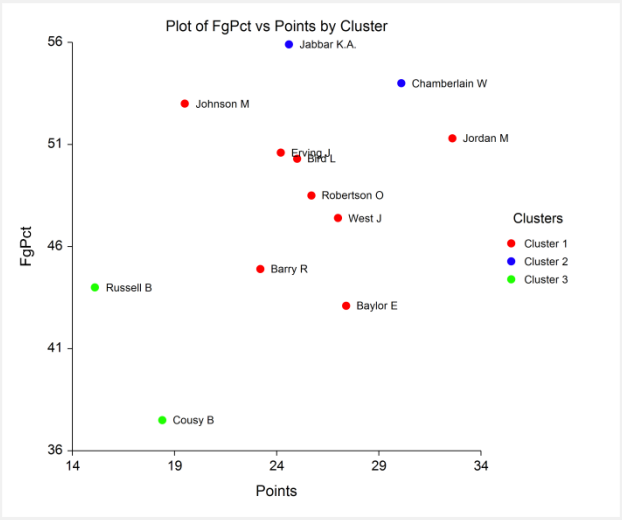
Count = 2

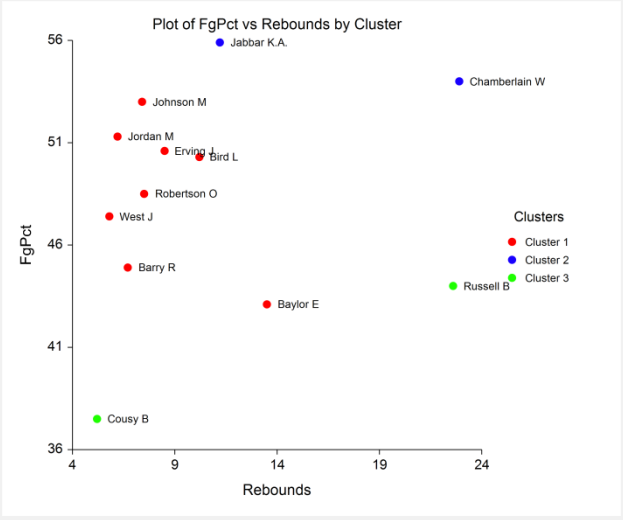
These sections show the same distances as in the previous distance report, except that the rows from only one cluster at a time are displayed. This makes it easier to see which items fell into each cluster.

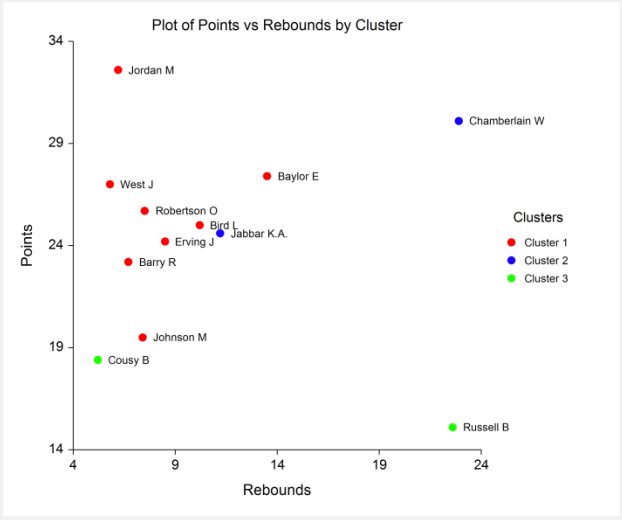
Bivariate Plots Section

K-Means Clustering









This series of scatter plots shows the data for each pair of variables with different clusters shown with different plotting symbols. The row numbers may be displayed at the side of plot symbols to help identify problem observations.

These plots will help you find outliers, anomalies, and various other problems. Note that because of the multivariate nature of the data, your cluster configuration may be good yet still show little pattern in these plots.