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.

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.

Setup

To run this example, complete the following steps:

1 Open the BBall example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **BBall** and click **OK**.

2 Specify the K-Means Clustering procedure options

- Find and open the **K-Means Clustering** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Cluster Variables ..... **Height,FgPct,Points,Rebounds**
Label Variable ..... **Player**
Maximum Clusters **4**
Random Seed ..... **5025549** (for reproducibility)

Reports Tab

All Reports ..... **Checked**

Plots Tab

Bivariate Plots ..... **Checked**
Show Row Numbers ..... **Unchecked**
Show Row Labels ..... **Checked**

Scatter Plot Format (*Click the Button*)

Labels (Data Point Labels) ..... **Checked**

Report Options (*in the Toolbar*)

Variable Labels ..... **Column Names**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Minimum Iterations Summary

Minimum Iterations Summary

Iteration	Number of Clusters	Percent of Variation	
		Value	Bar
3	2	65.54%	||||||||||||||||||
4	3	46.48%	||||||||||||
8	4	29.17%	||||||||

This report is provided to help you determine the optimal number of clusters.

Iteration

The iteration number from the Iteration Detail report.

Number of Clusters

The number of clusters reported on.

Percent of Variation: Value

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.

Percent of Variation: Bar

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

Iteration Detail

Iteration Detail

Iteration	Number of Clusters	Percent of Variation	
		Value	Bar
1	2	72.03%	||||||||||||||||||||
2	2	72.03%	||||||||||||||||||||
3	2	65.54%	||||||||||||||||||
4	3	46.48%	||||||||||||
5	3	46.48%	||||||||||||
6	3	46.48%	||||||||||||
7	4	31.81%	||||||||
8	4	29.17%	||||||||
9	4	32.96%	||||||||

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

The iteration number reported on this line.

Number of Clusters

The number of clusters in this configuration.

Percent of Variation: Value

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.

Percent of Variation: Bar

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

Cluster Means

Cluster Means

	Cluster Mean		
Variables	**1**	**2**	**3**
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
Item 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

	Cluster Standard Deviation		
Variables	**1**	**2**	**3**
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
Item 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.

Analysis of Variance (ANOVA) Table

Analysis of Variance (ANOVA) Table

	Degrees of Freedom		Mean Square			
Variables	**DF1**	**DF2**	**Between**	**Within**	**F-Ratio**	**P-Value**
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.

Distances to Cluster Centers

Distances to Cluster Centers

		Distance to Cluster Center		
Row Label	**Cluster**	**1**	**2**	**3**
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.

Distances to Cluster Centers by Cluster

Distances to Cluster Centers for Items in Cluster 1

		Distance to Cluster Center		
Row Label	**Cluster**	**1**	**2**	**3**
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

Number of Items in This Cluster = 8

Distances to Cluster Centers for Items in Cluster 2

		Distance to Cluster Center		
Row Label	**Cluster**	**1**	**2**	**3**
1 Jabbar K.A.	2	2.4609	1.1263	4.0315
5 Chamberlain W	2	3.2456	1.1263	4.4712

Number of Items in This Cluster = 2

Distances to Cluster Centers for Items in Cluster 3

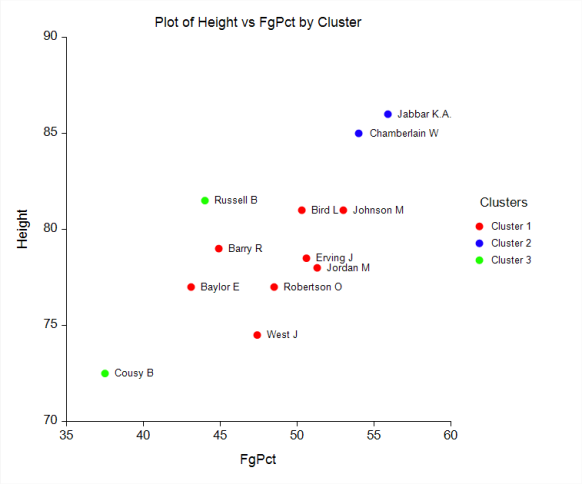
		Distance to Cluster Center		
Row Label	**Cluster**	**1**	**2**	**3**
6 Cousy B	3	2.9971	5.3790	1.9512
11 Russell B	3	3.3878	3.5197	1.9512

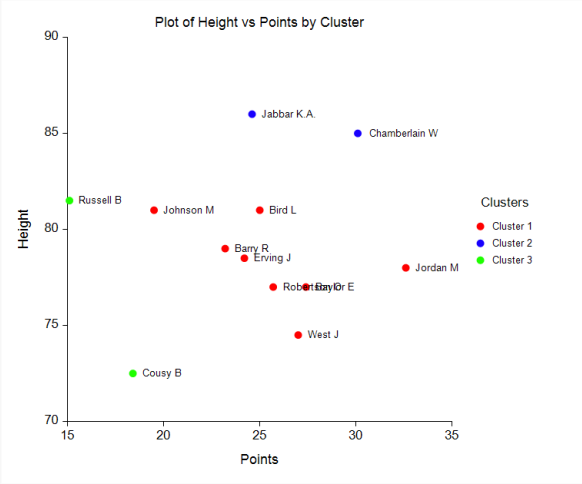
Number of Items in This Cluster = 2

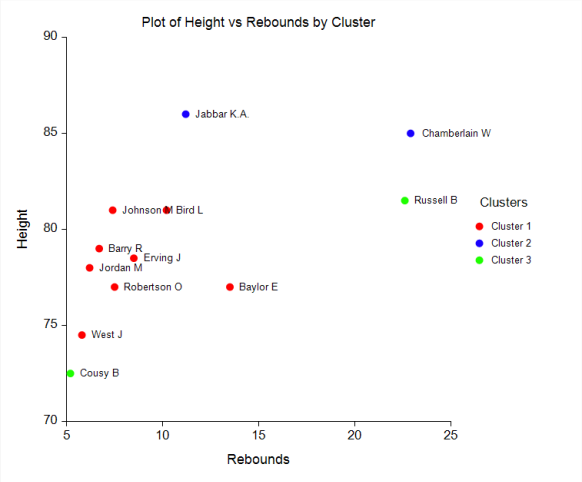
These sections show the same distances as in the previous distance report, except that the rows from only one cluster at a time are displayed. This makes it easier to see which items fell into each cluster.

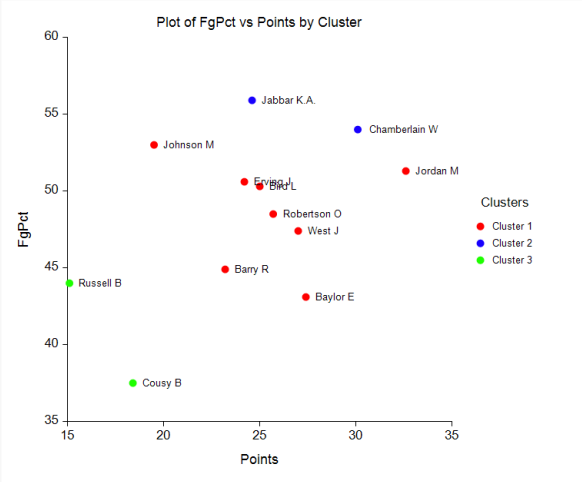
Bivariate Plots

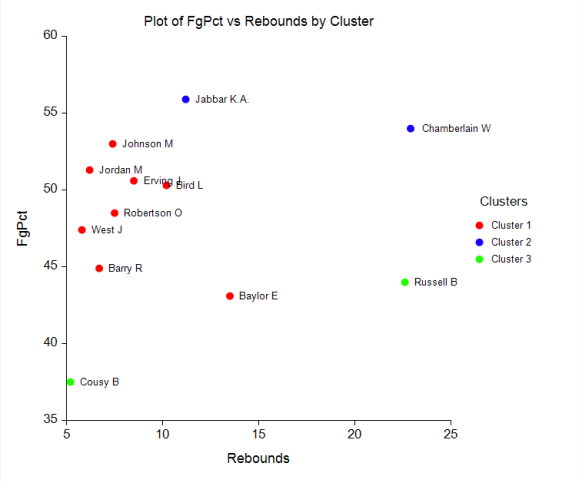
Bivariate Plots

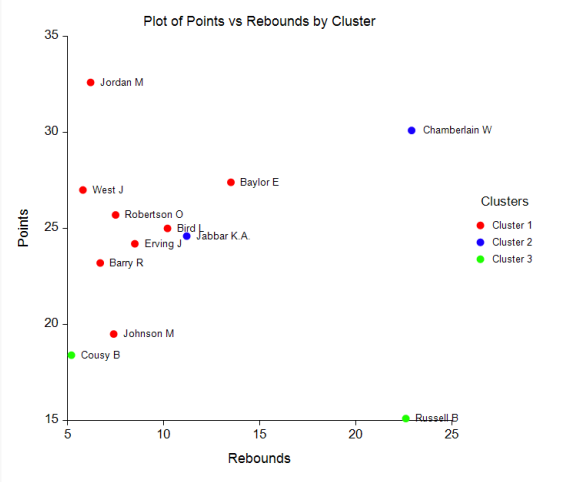












This series of scatter plots shows the data for each pair of variables with different clusters shown with different plotting symbols. The row numbers or row labels may be displayed at the side of plot symbols to help identify problem observations.

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.