How can PROC CLUSTER be used in SAS, and can you provide an example?

Authored by

stats writer

June 23, 2024

PROC CLUSTER is a SAS procedure used for clustering analysis, which is the process of grouping objects into clusters based on their similarities. This procedure allows users to identify patterns and relationships within a dataset, making it useful for data exploration and classification tasks.

One example of using PROC CLUSTER in SAS is to group customer data based on their purchasing behavior. This can help businesses identify different segments of customers and tailor their marketing strategies accordingly. The procedure can be used to cluster customers based on variables such as purchase frequency, amount spent, and types of products purchased. The resulting clusters can then be further analyzed to understand the characteristics and preferences of each group. This information can be used to develop targeted marketing campaigns and improve customer satisfaction.

Use PROC CLUSTER in SAS (With Example)

Clustering is a technique in machine learning that attempts to find clusters of observations within a dataset.

The goal is to find clusters such that the observations within each cluster are quite similar to each other, while observations in different clusters are quite different from each other.

The easiest way to perform clustering in SAS is to use PROC CLUSTER.

The following example shows how to use PROC CLUSTER in practice.

Example: How to Use PROC CLUSTER in SAS

Suppose we have the following dataset that contains information about points, assists and rebounds for 20 different basketball players:

```
/*create dataset*/
data my_data;
input points assists rebounds;
datalines;
18 3 15
20 3 14
19 4 14
14 5 10
14 4 8
15 7 14
20 8 13
28 7 9
30 6 5
31 9 4
35 12 11
33 14 6
29 9 5
25 9 5
25 4 3
```

```
27 3 8
29 4 12
30 12 7
19 5 6
23 11 5
;
run;

/*view dataset*/
proc printdata=my_data;
```

Obs	player_ID	points	assists	rebounds
1	1	18	3	15
2	2	20	3	14
3	3	19	4	14
4	4	14	5	10
5	5	14	4	8
6	6	15	7	14
7	7	20	8	13
8	8	28	7	9
9	9	30	6	5
10	10	31	9	4
11	11	35	12	11
12	12	33	14	6
13	13	29	9	5
14	14	25	9	5
15	15	25	4	3
16	16	27	3	8
17	17	29	4	12
18	18	30	12	7
19	19	19	5	6
20	20	23	11	5

Suppose we would like to perform clustering to attempt to identify "clusters" of players that have similar stats to each other.

The following code shows how to use PROC CLUSTER in SAS to perform clustering:

/*perform clustering using points, assists and rebounds variables*/

proc clusterdata=my_data method=average; var points assists rebounds;run;

The first tables in the output provide information about how the clustering was performed:

The CLUSTER Procedure
Average Linkage Cluster Analysis

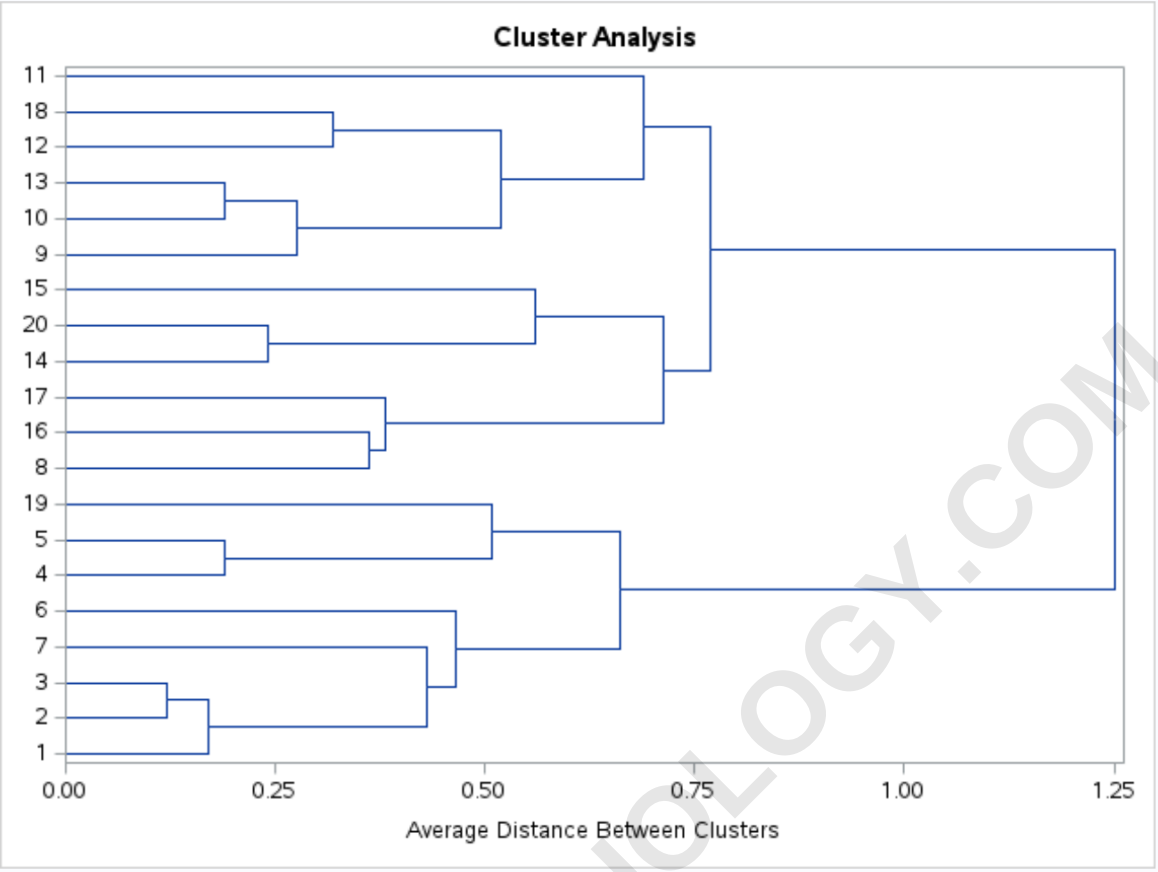
Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	51.0645891	40.0401485	0.7416	0.7416
2	11.0244406	4.2529440	0.1601	0.9017
3	6.7714966		0.0983	1.0000

Root-Mean-Square Total-Sample Standard Deviation	4.790982

Root-Mean-Square Distance Between Observations	11.73546

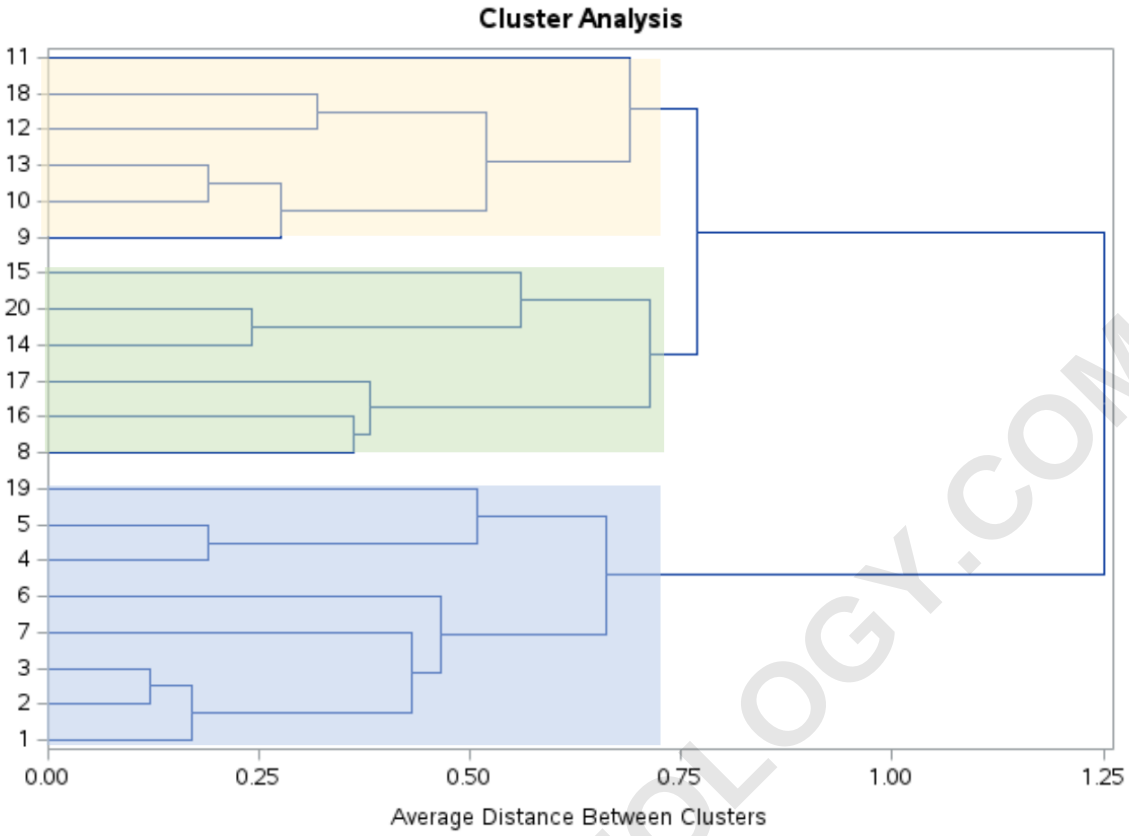
Cluster History					
Number of Clusters	Clusters Joined		Freq	Norm RMS Distance	Tie
19	OB2	OB3	2	0.1205	
18	OB1	CL19	3	0.1704	
17	OB4	OB5	2	0.1905	T
16	OB10	OB13	2	0.1905	
15	OB14	OB20	2	0.241	
14	OB9	CL16	3	0.2761	
13	OB12	OB18	2	0.3188	T
12	OB8	OB16	2	0.3615	
11	CL12	OB17	3	0.3811	
10	CL18	OB7	4	0.4317	
9	CL10	OB6	5	0.4648	
8	CL17	OB19	3	0.5077	
7	CL14	CL13	5	0.5183	T
6	CL15	OB15	3	0.5588	
5	CL9	CL8	8	0.6615	
4	CL7	OB11	6	0.6881	
3	CL11	CL6	6	0.7124	
2	CL3	CL4	12	0.7677	
1	CL5	CL2	20	1.251	

A dendrogram is also produced so that we can visually inspect the similarity between observations in the dataset:



The y-axis shows the individual observations and the x-axis shows the average distance between clusters.

From looking at this dendrogram, it appears that the observations naturally group themselves into three clusters:



We can then use the PROC TREE statement with ncl=3 to tell SAS to assign each observation in the original dataset to one of three clusters:

```
/*assign each observation to one of three clusters*/
proc treedata=clustd noprint ncl=3 out=clusts;
copy points assists rebounds;
id player_ID;
run;
proc sort;
by cluster;
```

```
run;

/*view cluster assignments*/
proc printdata=clusts;
id player_ID;
run;
```

The resulting dataset shows each of the original observations along with the cluster they belong to:

player_ID	points	assists	rebounds	CLUSTER	CLUSNAME
2	20	3	14	1	CL5
3	19	4	14	1	CL5
1	18	3	15	1	CL5
4	14	5	10	1	CL5
5	14	4	8	1	CL5
7	20	8	13	1	CL5
6	15	7	14	1	CL5
19	19	5	6	1	CL5
10	31	9	4	2	CL4
13	29	9	5	2	CL4
9	30	6	5	2	CL4
12	33	14	6	2	CL4
18	30	12	7	2	CL4
11	35	12	11	2	CL4
14	25	9	5	3	CL3
20	23	11	5	3	CL3
8	28	7	9	3	CL3
16	27	3	8	3	CL3
17	29	4	12	3	CL3
15	25	4	3	3	CL3

For example, we can see: that players with ID's 2, 3, 1, 4, 5, 7, 6 and 19 all belong to cluster 1.

This tells us that these eight players are "similar" across the points, assists and rebounds variables.

Note: For this example we chose to use average as the linkage method for clustering. Refer to the for a complete list of other linkage methods you can use.

The following tutorials explain how to perform other common tasks in SAS: