How to Use PROC CLUSTER in SAS (With Example)

Authored by

stats writer

November 19, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Use PROC CLUSTER in SAS (With Example)*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=96630

ARABPSYCHOLOGY.COM

Cluster analysis is a cornerstone technique in machine learning and statistics, invaluable for uncovering inherent structures and patterns within complex datasets. This article serves as an expert guide on implementing hierarchical clustering using PROC CLUSTER in the SAS statistical software suite. We will meticulously review the fundamental concepts, necessary parameters, and execution steps for utilizing this powerful procedure. A comprehensive, practical example will illustrate how to apply PROC CLUSTER effectively to real-world data, followed by discussions on interpreting the resulting output, including the critical use of the dendrogram. By the conclusion of this tutorial, you will possess the requisite knowledge to perform sophisticated hierarchical clustering in SAS, enabling data-driven insights and supporting robust decision-making processes.

Fundamentals of Cluster Analysis

Clustering is an unsupervised learning technique employed to organize heterogeneous observations into homogenous groups, or clusters. The primary objective of this method is to partition a dataset such that observations within a single cluster are significantly more similar to one another than they are to observations belonging to different clusters. This concept of maximizing intra-cluster similarity while minimizing inter-cluster similarity is fundamental to effective data segmentation.

Successful cluster analysis requires careful consideration of both the distance metric used to measure similarity (e.g., Euclidean distance) and the linkage method employed to define cluster proximity. By identifying these latent structures, analysts can gain profound insights into underlying population segments or behavioral patterns that might otherwise remain obscured in the raw data.

Introducing PROC CLUSTER in SAS

In the SAS programming environment, the most straightforward and versatile way to execute hierarchical clustering is through the PROC CLUSTER procedure. This procedure is designed to perform various types of hierarchical clustering, allowing users to specify different methods for measuring distance and determining how clusters are merged or split.

The typical syntax for using PROC CLUSTER involves specifying the input dataset and the crucial parameters, most notably the **METHOD=** option, which dictates the specific hierarchical linkage strategy. Understanding these parameters is essential for producing meaningful and statistically sound cluster results. The following sections will guide you through a practical application of this powerful SAS procedure.

ARABPSYCHOLOGY.COM

Practical Example: Setting Up the Basketball Data

To demonstrate the practical application of PROC CLUSTER, let us consider a dataset compiled from 20 different basketball players. This dataset includes three key performance indicators: **points** scored, **assists** recorded, and **rebounds** collected. Our primary goal is to use clustering techniques to segment these players into distinct groups based on the similarity of their statistical profiles.

The initial step involves creating and reviewing the input dataset within the SAS environment. The code below illustrates the data creation process using the **DATALINES** statement, followed by the **PROC PRINT** statement, which allows us to verify the successful loading of the variables.

```
/*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;
```

ARABPSYCHOLOGY.COM

```
/*view dataset*/
proc print data=my_data;
```

The resulting dataset, which contains 20 observations corresponding to 20 players, is visualized below, confirming the structure of our input data prior to the clustering procedure.

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

ARABPSYCHOLOGY.COM

Executing the PROC CLUSTER Statement

With the data successfully loaded, we can now execute the core clustering routine. Our objective remains to identify inherent clusters of players who exhibit comparable statistical characteristics across the selected variables. The following SAS code demonstrates the necessary syntax for invoking **PROC CLUSTER**.

In this specific example, we utilize the **METHOD=AVERAGE** option. This average linkage method calculates the distance between two clusters as the average distance between all pairs of observations where one observation is in one cluster and the other is in the second cluster. This choice influences how the hierarchy is constructed and is a common technique for general-purpose clustering. The **VAR** statement explicitly tells the procedure which variables to use for calculating the similarity distance.

```
/*perform clustering using points, assists and rebounds variables*/
proc cluster data=my_data method=average;
var points assists rebounds;
run;
```

Interpreting the Initial Output Tables

Upon execution, **PROC CLUSTER** generates several output tables detailing the step-by-step process of the hierarchical grouping. These tables provide crucial information regarding the sequence in which observations and groups were combined, and the distance or measure of similarity at each merger stage. This output is critical for validating the clustering process and understanding the structure of the data aggregation.

Specifically, the initial tables detail the history of cluster mergers. Each line item typically shows which two clusters or observations were joined, the resulting cluster number, and the distance measure (or criterion value) associated with that merger. A smaller distance measure generally indicates a higher degree of similarity between the merged entities, confirming that the procedure is grouping the most similar entities first.

The image below illustrates the beginning of the hierarchical clustering process output, confirming the linkage method used and the variables included in the distance calculations.

ARABPSYCHOLOGY.COM

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	

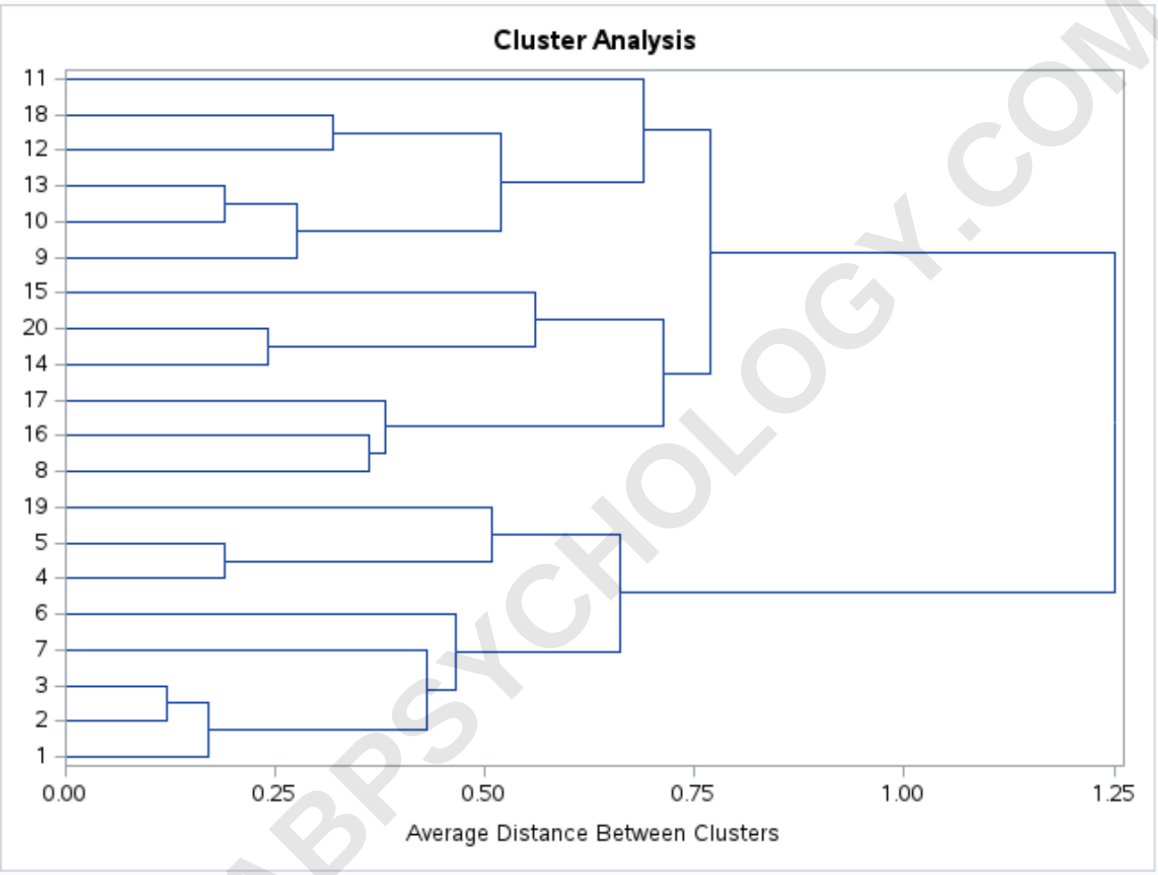
ARABPSYCHOLOGY.COM

Visualizing Hierarchical Structure with the Dendrogram

A key graphical output from **PROC CLUSTER** is the dendrogram, which provides a visual representation of the entire hierarchical clustering process. The dendrogram is invaluable for subjectively determining the optimal number of clusters for the dataset, as it graphically displays

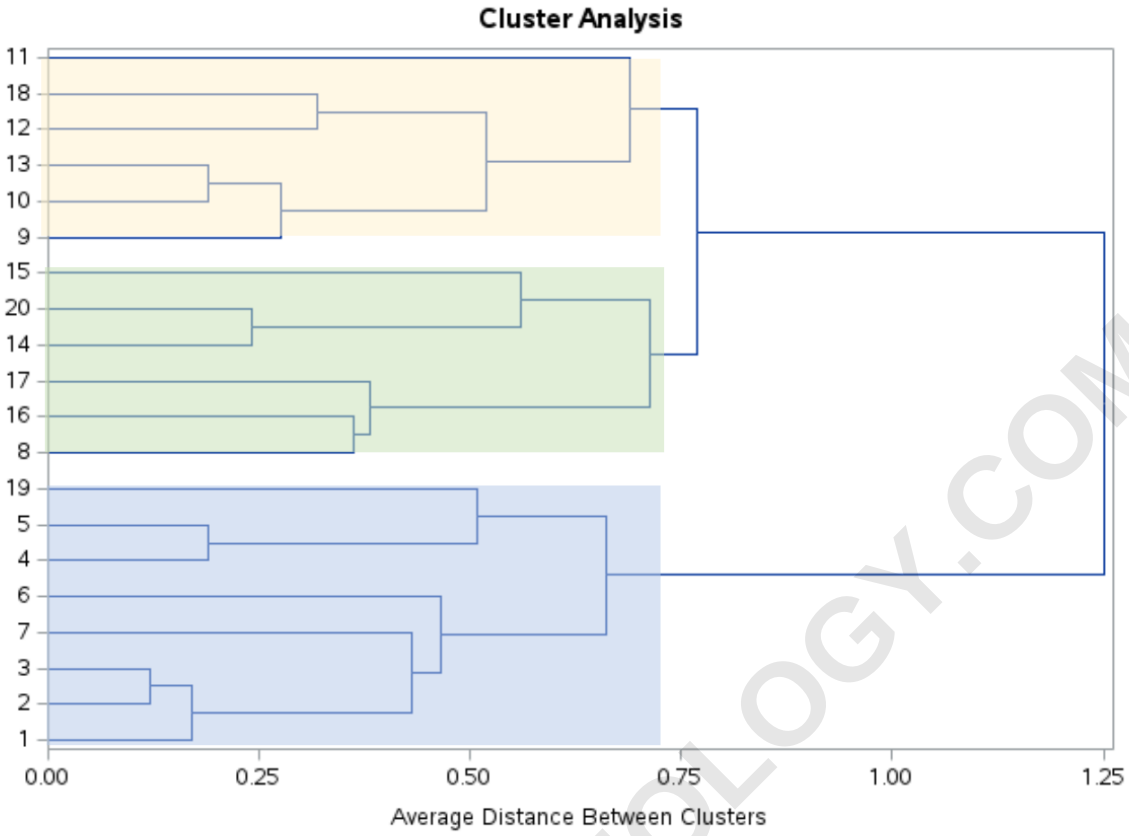
the distance at which clusters were fused.

The structure of the dendrogram is interpreted as follows: the Y-axis represents the individual observations (players, in this case), while the X-axis represents the average distance (or dissimilarity) between the clusters when they are merged. Taller vertical lines indicate that the clusters being joined are more distant from each other, suggesting a less natural grouping.



ARABPSYCHOLOGY.COM

By examining the structure of the branches, we look for natural cuts or large gaps in the fusion distances. Based on the visual inspection of this specific dendrogram, it appears that the 20 observations naturally aggregate into three distinct clusters before large distance jumps occur. This determination of three clusters (ncl=3) will guide the next step in our analysis.



Assigning Observations Using PROC TREE

Once the optimal number of clusters (in our case, three) has been identified using the dendrogram, we must formally assign each original observation to its corresponding cluster group. This crucial step is accomplished using the **PROC TREE** statement in SAS, which utilizes the cluster history generated by **PROC CLUSTER**.

The **NCL=3** option within **PROC TREE** explicitly instructs SAS to cut the dendrogram at the three-cluster level and assign cluster membership accordingly. Furthermore, we use the **OUT=clusts** option to create a new output dataset containing the original variables plus a new cluster assignment variable. The **COPY** statement ensures the original variables are included, and the **ID** statement is used to identify the players uniquely.

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

ARABPSYCHOLOGY.COM

```
proc sort;
by cluster;
run;

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

The final output dataset, shown below, now displays each basketball player's original statistics alongside their newly assigned cluster number (1, 2, or 3). This output is the foundation for further descriptive analysis of the clusters.

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

ARABPSYCHOLOGY.COM

For instance, the results clearly demonstrate that players identified by ID numbers 2, 3, 1, 4, 5, 7, 6, and 19 were all grouped into **Cluster 1**. The practical implication is that these eight individuals

share fundamentally similar patterns across their points, assists, and rebounds variables, suggesting a comparable role or performance tier within the dataset. Further analysis (like calculating cluster means) would reveal the specific characteristics defining this cluster.

Considerations for Linkage Methods

It is imperative to recognize that the choice of **linkage method** significantly influences the outcome and shape of the resulting clusters. In this demonstration, we employed the **average linkage method**. However, **PROC CLUSTER** supports a wide array of alternative hierarchical methods, each with unique properties and suitability for different data structures.

Common alternatives include:

Single Linkage (Nearest Neighbor): Tends to produce long, chain-like clusters.
Complete Linkage (Farthest Neighbor): Tends to produce compact, globular clusters.
Ward's Method: Minimizes the within-cluster variance, generally resulting in robust, spherical clusters.

Analyst must carefully select the appropriate linkage method based on prior domain knowledge and the theoretical shape of the expected clusters. Consulting the official SAS documentation is recommended for a complete listing and detailed explanation of all available options for the **METHOD=** parameter.

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

ARABPSYCHOLOGY.COM