

Clustering big datasets with XLSTAT - Using k-means clustering followed by an AHC

[demoCluster3.xls](#)

Dataset to cluster

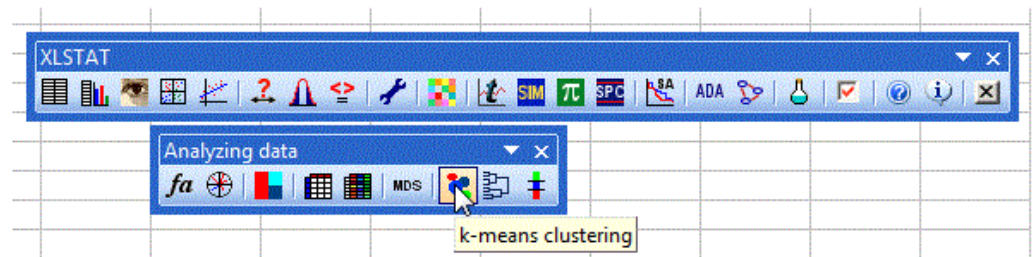
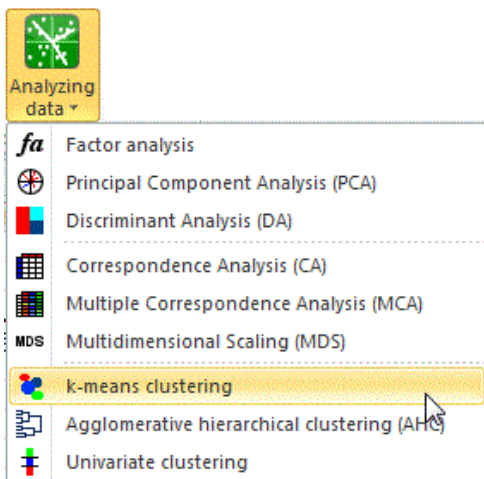
An Excel sheet containing both the data and the results for use in this tutorial can be downloaded by clicking [here](#).

The data are from the US Census Bureau and describe the changes in the population of 51 states between 2000 and 2001. The initial dataset has been transformed to rates per 1000 inhabitants, with the data for 2001 serving as the focus for the analysis. Our aim is to create homogeneous clusters of states based on the demographic data we have available. This dataset is not very big but it will illustrate how to deal with much bigger dataset.

Note: if you try to re-run the same analysis as described below on the same data, as the k-means method starts from randomly selected clusters, you may obtain different results from those listed hereunder. To fix the seed, go to the XLSTAT Options, Advanced tab, then check the "fix the seed" option.

Setting up the k-means clustering

Once XLSTAT is activated, select the **XLSTAT / Analyzing data / k-means clustering** command, or click on the corresponding button of the **Analyzing data** toolbar (see below).



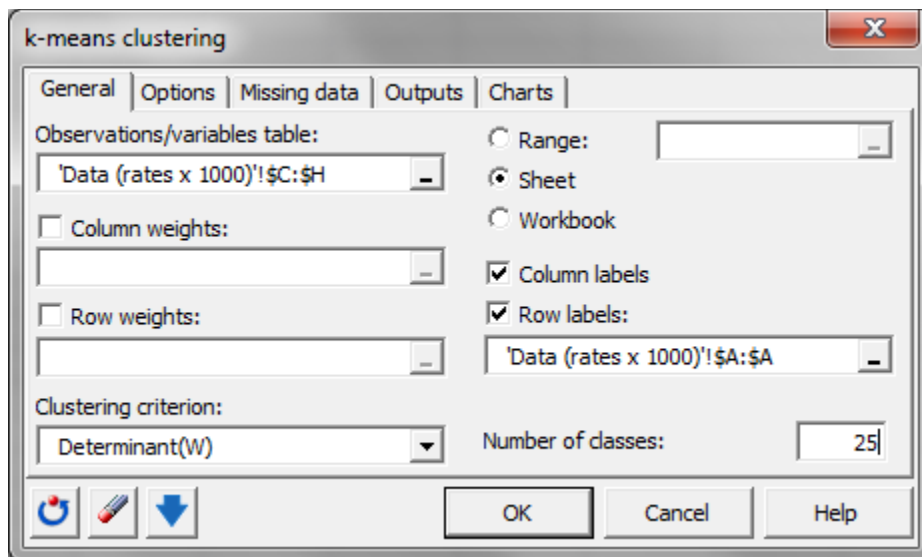
Once you've clicked the button, the k-means clustering dialog box appears.

Select the data on the Excel sheet with the mouse. (Note: There are several ways of selecting data with XLSTAT - for further information, please check the tutorial on [selecting data](#).) In this example, the data start from the first row, so it is quicker and easier to use the "column selection" mode. This explains why the letters corresponding to the columns are displayed in the selection boxes (C to H).

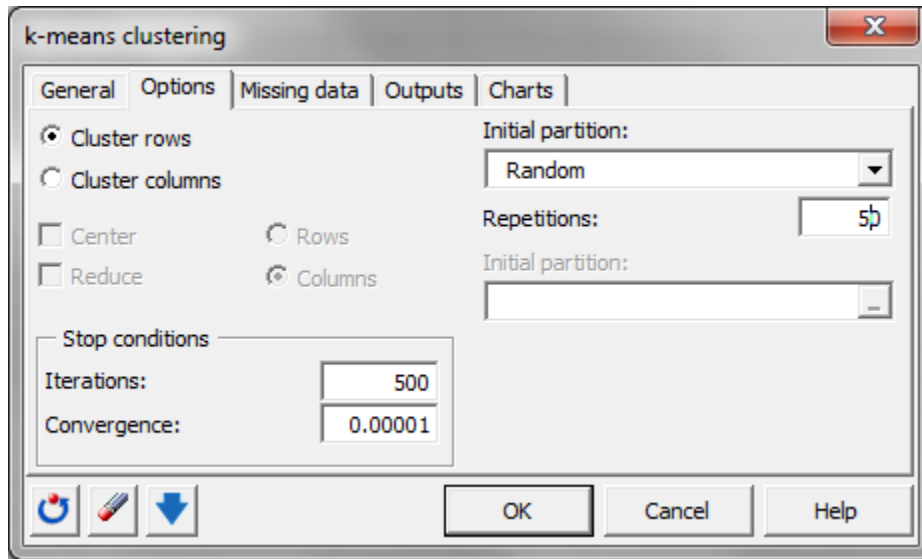
The Total population variable was not selected, as we are interested mainly in the demographic dynamics. The last column was not selected as it is fully correlated with the column preceding it. The **observations labels** were selected as they are available.

We set the **number of groups** to be created to 25. In the case of much bigger dataset you may use a bigger number.

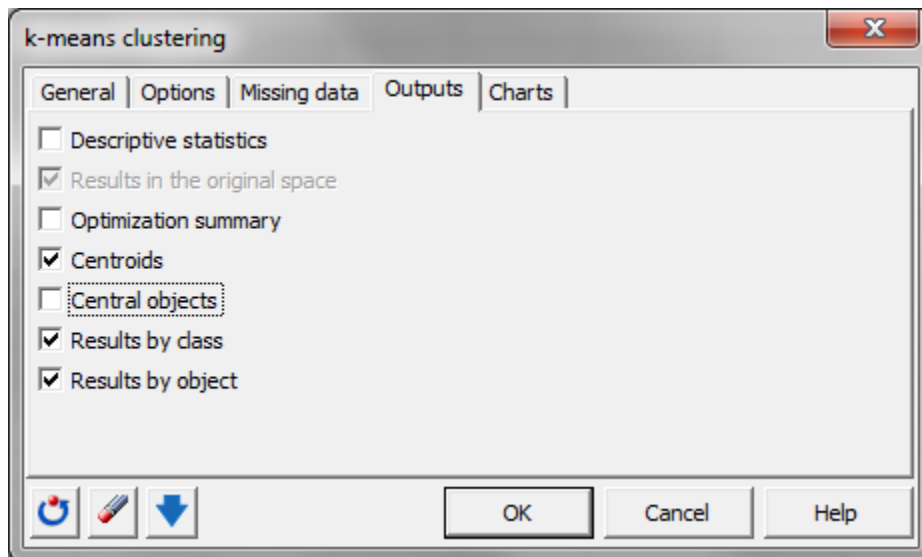
The selected criterion is "**Determinant(W)**" as it allows to remove the scale effects of the variables.



In the options tab we increased the number of **repetitions** to 50 in order to increase the quality and the stability of the results.



In the **Outputs** tab we select only the **Centroids** which we will use in the AHC, the **results by class** as it will give us the samples within each class and the **results by object** to get the table of the sample with an attribution variable.



Once you clicked on **OK** the results of the k-means clustering will appear in a new sheet.

Agglomerative Hierarchical Clustering on the results of the k-means clustering

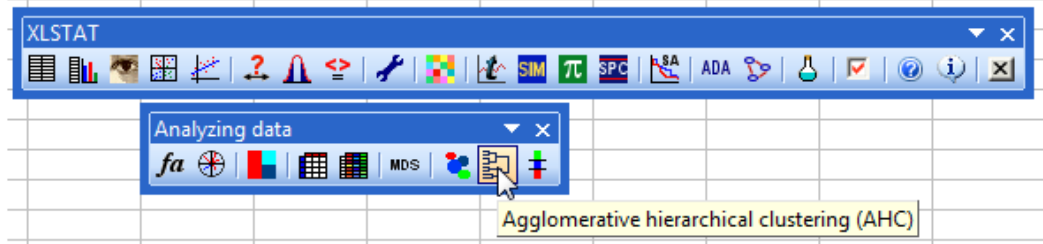
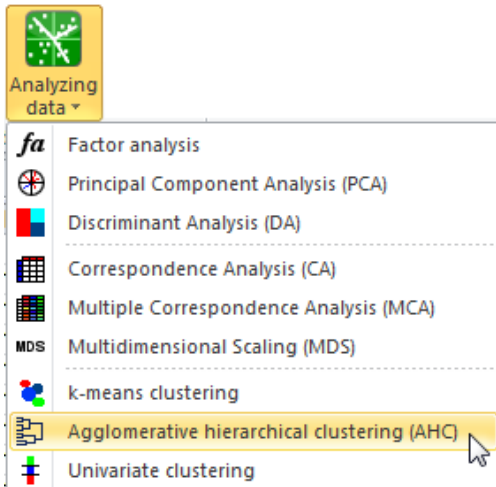
We are going to work on the table **Class centroids**.

Class centroids:										
Class	Net Domestic Mig.	Federal/Civilian move from abroad	Net Int. Migration	Period Births	Period Deaths	< 65 Pop. Est.	Sum of weights	Within-class variance		
1	-2,896	-0,029	1,029	14,082	10,065	868,153	3,000	2,193		
2	-1,718	-0,239	2,095	15,952	4,645	941,949	1,000	0,000		
3	14,251	-0,034	4,295	15,883	7,766	869,535	1,000	0,000		
4	1,798	-0,024	1,286	12,495	9,849	857,772	4,000	28,529		
5	-2,014	-0,044	7,879	15,372	6,716	894,029	1,000	0,000		
6	6,443	-0,058	3,759	15,860	7,028	902,848	3,000	17,486		
7	-1,844	-0,019	2,070	12,647	9,246	864,853	4,000	7,139		
8	3,471	-0,012	1,892	12,502	8,681	871,038	4,000	11,389		
9	-7,523	-0,045	5,242	14,601	9,744	880,508	2,000	2,943		
10	12,521	-0,035	5,763	12,544	10,128	826,278	1,000	0,000		
11	-2,499	-0,293	4,325	15,438	6,874	866,178	1,000	0,000		
12	3,700	-0,081	3,272	14,130	7,768	887,389	4,000	5,204		
13	-2,348	-0,001	1,582	13,852	9,036	877,198	2,000	1,003		
14	-9,582	-0,039	0,965	12,407	9,543	852,059	2,000	31,005		
15	-6,363	-0,052	2,280	14,535	9,346	868,324	1,000	0,000		
16	0,828	-0,038	0,904	13,761	9,833	875,695	2,000	1,073		
17	-6,727	-0,043	1,654	15,244	8,435	882,803	2,000	10,842		
18	-1,737	-0,029	1,375	12,880	7,874	881,138	2,000	15,234		
19	-3,115	-0,048	0,523	15,984	10,207	879,314	1,000	0,000		
20	-6,558	-0,037	1,790	14,184	8,903	864,922	1,000	0,000		
21	27,349	-0,031	6,450	14,221	7,168	888,644	1,000	0,000		
22	6,750	-0,061	1,423	13,424	8,679	879,395	3,000	18,917		
23	-7,207	-0,009	6,214	13,387	8,492	870,056	2,000	29,591		
24	-2,193	-0,001	0,881	11,554	11,341	845,889	2,000	4,412		
25	-5,729	-0,018	3,271	20,406	5,519	913,764	1,000	0,000		

Another important table is the table containing the information about which states are clustered together.

Results by class:										
Class	1	2	3	4	5	6	7	8	9	
Objects	3	1	1	4	1	3	4	4	2	
Sum of weights	3	1	1	4	1	3	4	4	2	
Within-class variance	2,193	0,000	0,000	28,529	0,000	17,486	7,139	11,389	2,943	
Minimum distance to centroid	0,696	0,000	0,000	2,990	0,000	2,100	1,894	2,676	1,213	
Average distance to centroid	1,149	0,000	0,000	4,507	0,000	3,280	2,286	2,913	1,213	
Maximum distance to centroid	1,621	0,000	0,000	5,529	0,000	4,423	2,699	3,284	1,213	
	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	District of Columbia	
	Ohio			Maine		Georgia	Massachusetts	Oregon	Illinois	
	Oklahoma			Rhode Island		Texas	Missouri	Vermont		
				South Dakota			Montana	Wisconsin		

Select now the option **XLSTAT / Analyzing data / Agglomerative Hierarchical Clustering** command, or click on the corresponding button of the "Analyzing data" toolbar (see below).



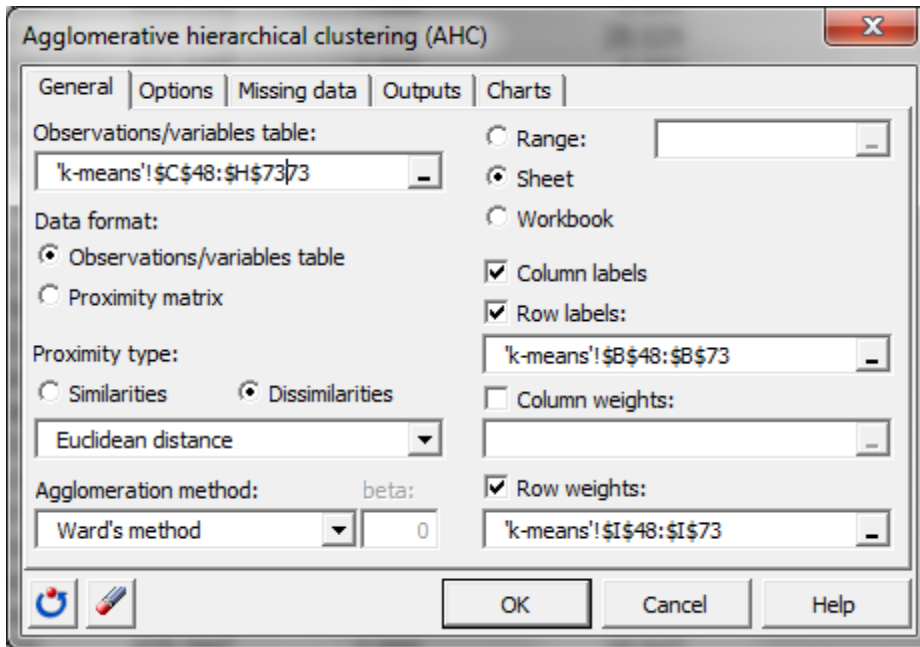
In the **General** tab you need to select the data to cluster. Select the original variables describing the 25 classes in the table **Class centroids** from the variable Net domestic Migration to the variable <65 Pop. Est.

We will use the **Proximity type: Dissimilarities** and the **Euclidian distance**, as well as the **Ward's method** as the agglomeration method.

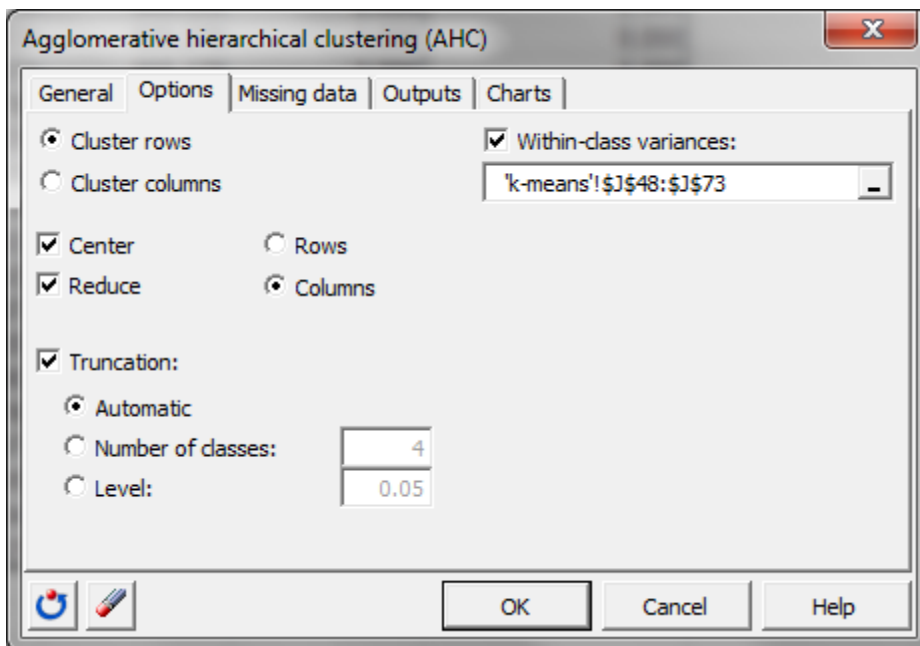
We have the name of the variables included in the selection so we tick the option **Column labels** and select the **Row labels** that are the cluster number (1-25).

We will use the **Row weights** option and select the column **Sum of weights** of the same table **Class centroids**.

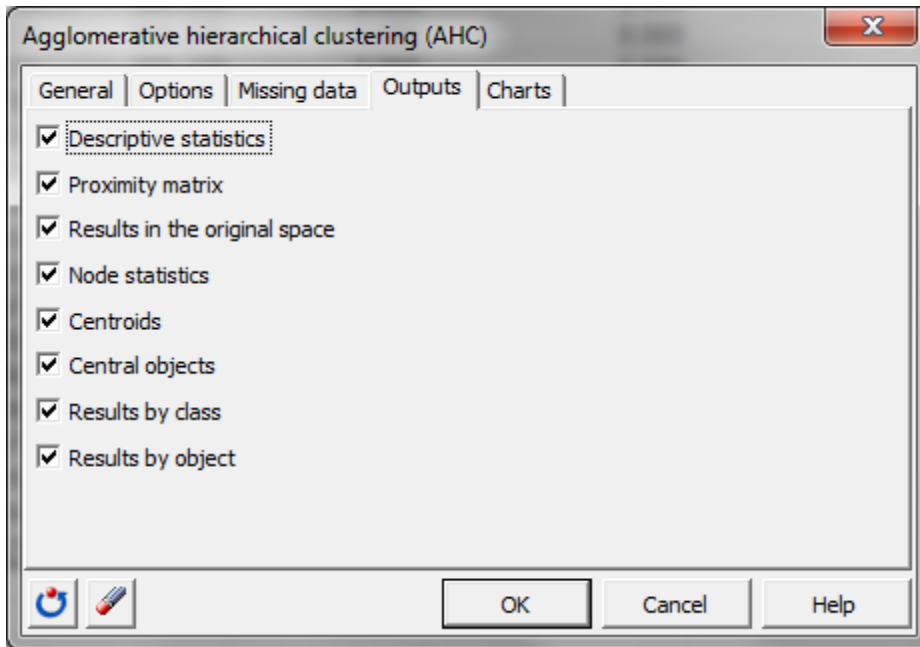
We can leave the selection of a new **sheet** to display the results.



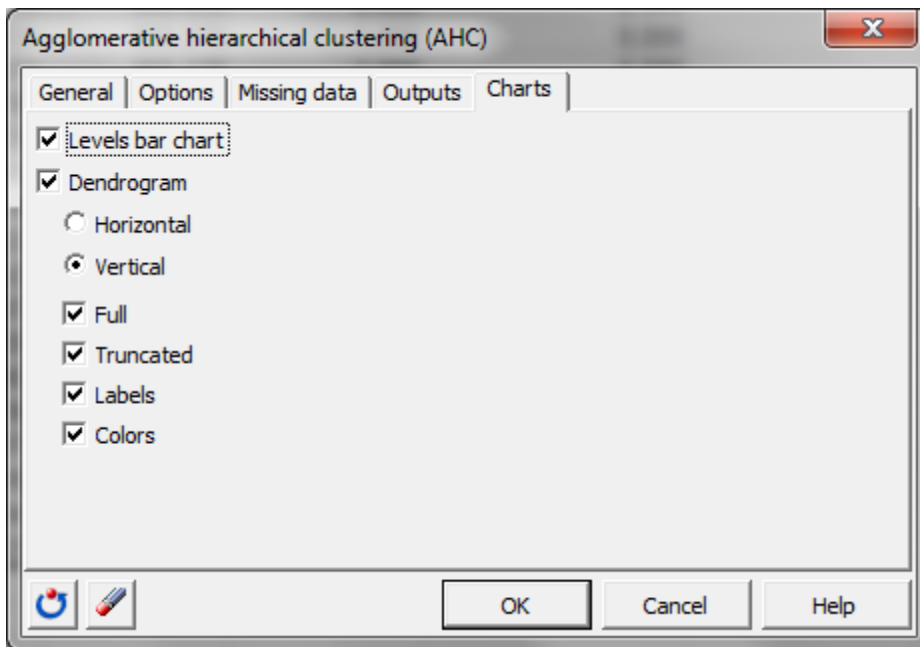
In the **Options** tab we are still clustering the rows as the classes are in rows but in this type of clustering (AHC after k-means) you need to include the **Within-class variances**. You will find this information in the same table as before: Class centroids, in the last column **Within class variance**.



We can select all the **Outputs** for this analysis.

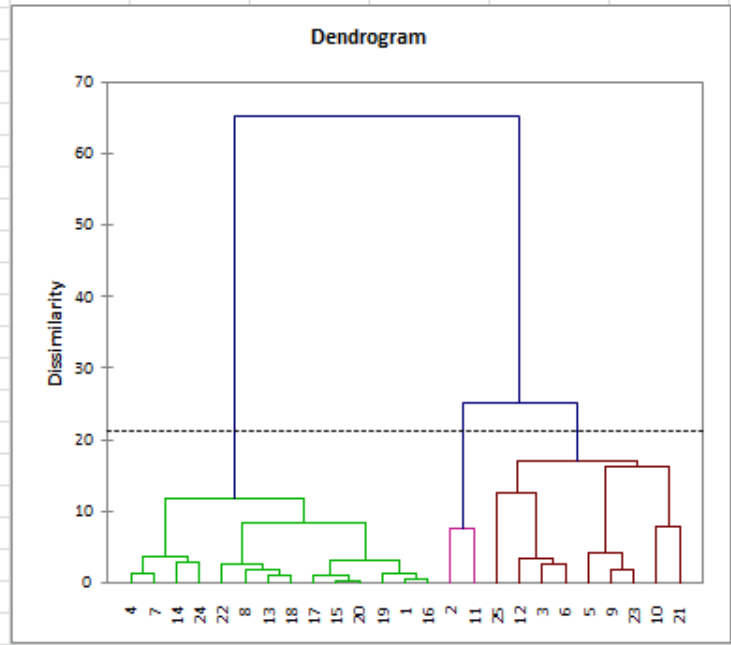
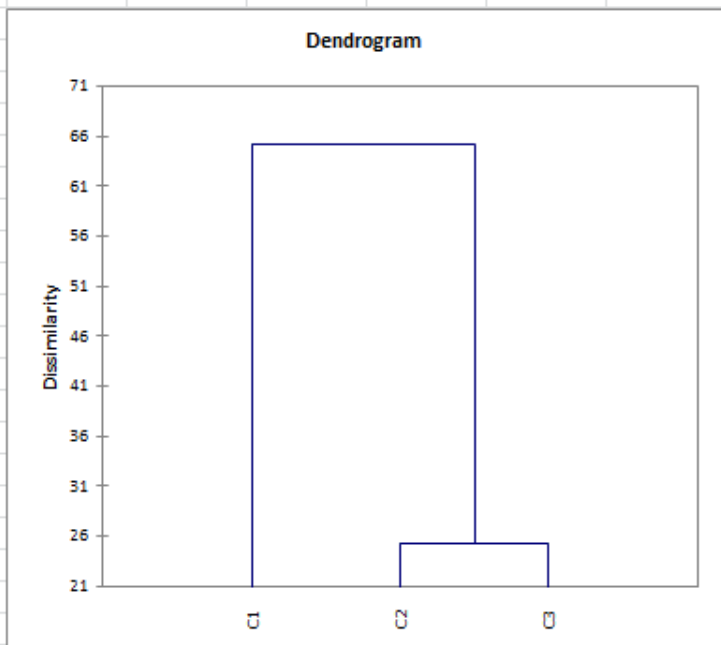


Finally in the tab **Charts**, select all the charts. Pay special attention the dendrogram type and select the option **Vertical**.



Results of the Agglomerative Hierarchical Clustering

In the results of the AHC, look at the two dendrograms that gives you the composition of the 3 clusters. You can see how the 25 clusters are again clustering in three final clusters.



Also you can look into the decomposition of the variance within and between classes.

Variance decomposition for the optimal classification:

	Absolute	Percent
Within-class	666,280	81,64%
Between-classes	149,793	18,36%
Total	816,072	100,00%

You can finally use the table obtained in the AHC to recode the table obtained in the k-means clustering so as to have the final results. Go to **XLSTAT / Preparing data / Coding**.

You need to select the column Class from the classification table obtained in the k-means clustering as the variable to recode. Select the table Result by object from the AHC including the name of the columns as the coding table. You then select the option **Column labels**.

To append the new column to the first table select the option Range and the first cell next to the table. Also uncheck the option **Display report header** so as to not have anything else displayed.

	A	B	C	D	E	F	G	H	I	J
3	Observation	Class	Distance to centroid			Observation	Class			
4	Alabama	1	1,621				1	1		
5	Alaska	2	0,000				2	2		
6	Arizona	3	0,000				3	3		
7	Arkansas	4	4,100				4	1		
8	California	5	0,000				5	3		
9	Colorado	6	3,317				6	3		
10	Connecticut	7	2,699				7	1		
11	Delaware	8	2,930				8	1		
12	District of Columbia	9	1,213				9	3		
13	Florida	10	0,000				10	3		
14	Georgia	6	2,100				11	2		
15	Hawaii	11	0,000							
16	Idaho	12	3,059							
17	Illinois	9	1,213							
18	Indiana	13	0,708							
19	Iowa	14	3,937							
20	Kansas	15	0,000							
21	Kentucky	16	0,733							
22	Louisiana	17	2,328							
23	Maine	4	5,529							
24	Maryland	12	1,897							
25	Massachusetts	7	1,894							
26	Michigan	13	0,708							
27	Minnesota	18	2,760							
28	Mississippi	19	0,000							
29	Missouri	7	2,584				24	1		
30	Montana	7	1,968				25	3		
31	Nebraska	20	0,000							
32	Nevada	21	0,000							
33	New Hampshire	22	4,804							
34	New Jersey	23	3,846							
35	New Mexico	17	2,328							
36	New York	23	3,846							
37	North Carolina	22	2,087							

Coding ✖

Data: Range: Sheet

Coding table: Workbook

Column labels Display the report header

Finally you obtained the results of the classification for the all the states.

Results by object:				Results by object:	
Observation	Class	Distance to centroid	Class	Observation	Class
Alabama	1	1,621	1	1	1
Alaska	2	0,000	2	2	2
Arizona	3	0,000	3	3	3
Arkansas	4	4,100	1	4	1
California	5	0,000	3	5	3
Colorado	6	3,317	3	6	3
Connecticut	7	2,699	1	7	1
Delaware	8	2,930	1	8	1
District of Columbia	9	1,213	3	9	3
Florida	10	0,000	3	10	3
Georgia	6	2,100	3	11	2
Hawaii	11	0,000	2	12	3
Idaho	12	3,059	3	13	1
Illinois	9	1,213	3	14	1
Indiana	13	0,708	1	15	1
Iowa	14	3,937	1	16	1
Kansas	15	0,000	1	17	1
Kentucky	16	0,733	1	18	1
Louisiana	17	2,328	1	19	1
Maine	4	5,529	1	20	1
Maryland	12	1,897	3	21	3
Massachusetts	7	1,894	1	22	1
Michigan	13	0,708	1	23	3
Minnesota	18	2,760	1	24	1
Mississippi	19	0,000	1	25	3
Missouri	7	2,584	1		
Montana	7	1,968	1		
Nebraska	20	0,000	1		
Nevada	21	0,000	3		

The following video shows how to do this tutorial. (This video does not have sound.)

http://www.youtube.com/watch?feature=player_embedded&v=GyL-lZnLT-4