SPSS Workshop ---Techniques for Clustering

Data

In this lesson we use the dataset ***churn.txt***, containing information on 1,477 customers of a telecommunication firm who has at some time purchased a mobile phone. The customers fall into one of three groups: current customers, involuntary leavers, and voluntary leavers. The file contains information on the customer account including length of time spent on local, long distance and international calls, the type of billing scheme, and a variety of basic demographics, such as age and gender. The data are typical of what is often referred to as a churn example (hence the file name).

Introduction

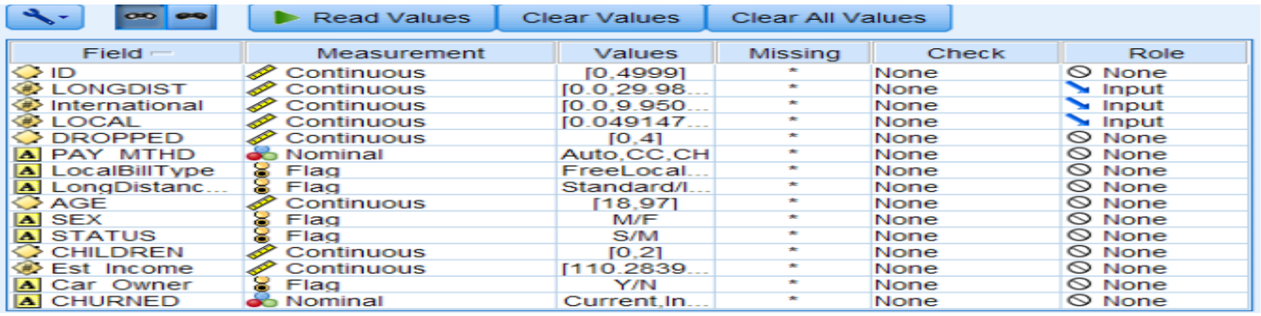
We will use the data file *churn.txt*. We will attempt to find natural segments, or clusters of customers to see whether they can be targeted for different promotions and to explore if cluster differences relate to customer status.

Steps:

1. Click **File…New Stream** → Place a **Var.File** Node on the Canvas → Double-click the node to input the data of **Churn.txt** from **Blackboard** → then Click **OK.**

2. Place a Table node to the above of Churn.txt and connect them → Examine the data and close the **Table** window

3. Place a **Type** node to the right of Churn.txt and connect them → Double click Type node to edit it →Define the role of three fields (**LONGDIST, International, LOCAL**) as **INPUT** corresponding to the amount of time spent on long distance, international, and local telephone calls, in minutes.(As shown below)



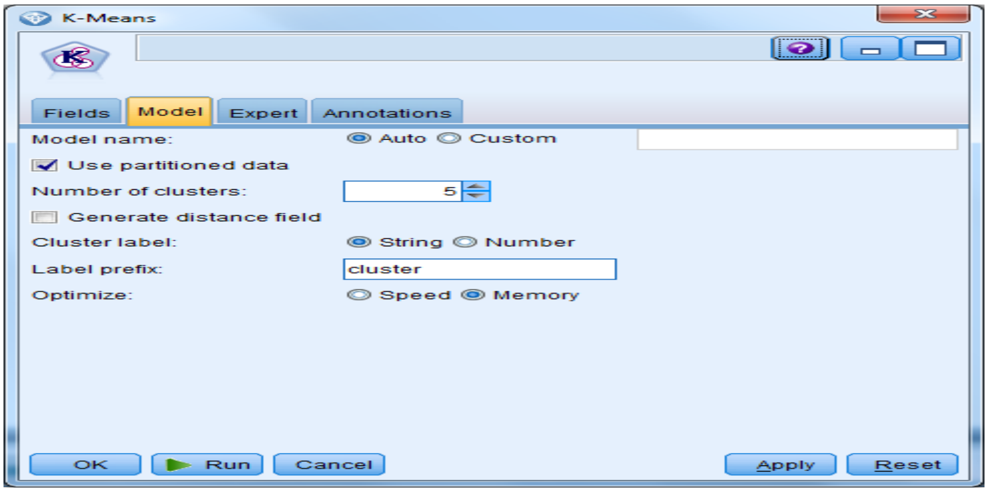
Field	Measurement	Values	Missing	Check	Role
ID	Continuous	[0,4999]	*	None	None
LONGDIST	Continuous	[0.0,29.98…	*	None	Input
International	Continuous	[0.0,9.950…	*	None	Input
LOCAL	Continuous	[0.049147…	*	None	Input
DROPPED	Continuous	[0,4]	*	None	None
PAY_MTHD	Nominal	Auto,CC,CH	*	None	None
LocalBillType	Flag	FreeLocal…	*	None	None
LongDistanc…	Flag	Standard/I…	*	None	None
AGE	Continuous	[18,97]	*	None	None
SEX	Flag	M/F	*	None	None
STATUS	Flag	S/M	*	None	None
CHILDREN	Continuous	[0,2]	*	None	None
Est_Income	Continuous	[110.2839…	*	None	None
Car_Owner	Flag	Y/N	*	None	None
CHURNED	Nominal	Current,In…	*	None	None

Figure 1

Notes: **Selection of Fields for Clustering**

In this example we will be using fields that have the same scale, i.e., minutes of time. This certainly isn't required for a clustering solution, but the selection of the input fields is an important decision. Unfortunately, very little advice is given about this in most references on clustering, where it is often assumed that the fields to be included will be **obvious**.

4. Click OK to close Type node→ Place a **K-Means** node K-Means from Modeling palette → Connect to the Type node → Double-click to edit the K-Means node.



K-Means

Fields | Model | Expert | Annotations

Model name: ● Auto ○ Custom

☑ Use partitioned data

Number of clusters: 5

☐ Generate distance field

Cluster label: ● String ○ Number

Label prefix: cluster

Optimize: ○ Speed ● Memory

Figure 2

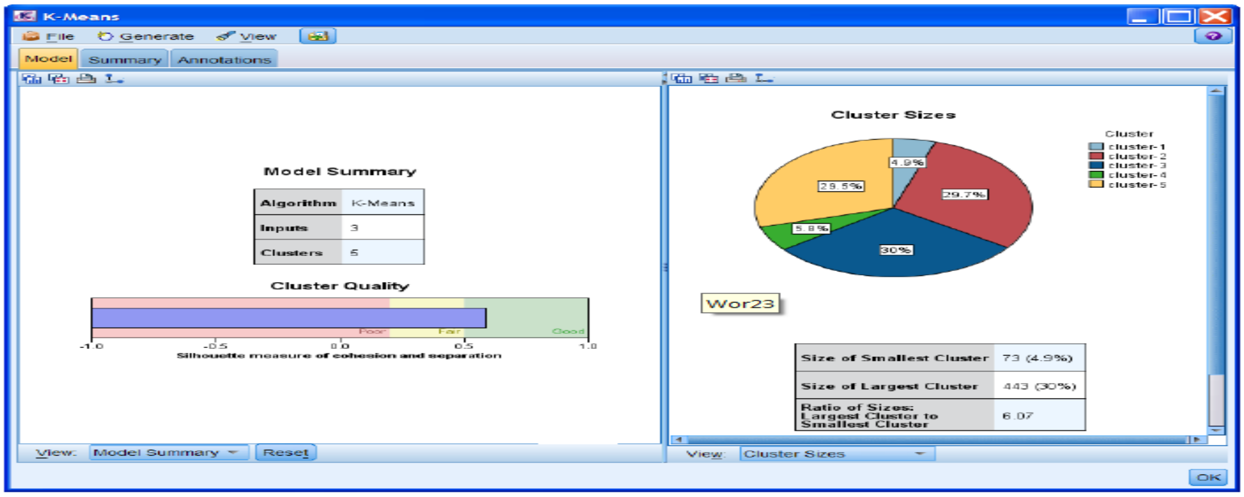
Notes: The **default Number of Clusters** is set to **5**, which means that the K-means model is going to return a 5-cluster solution. This number can be increased or decreased accordingly per the discussion above. There is no good sense for the number of clusters in the churn data, so we will retain the default value.

5. Run the **K-Means** node

Browsing the K-Means Results

6. Once the model has run, double-click the **K-Means model** node K-Means in the Models manager → Click the **Summary** tab (not shown)

7. Click **Model** tab



Model Summary

Algorithm	K-Means
Inputs	3
Clusters	5

Size of Smallest Cluster	73 (4.9%)
Size of Largest Cluster	443 (30%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	6.07

Figure 3

In the left (main) panel we see our five clusters are induced by three input variables, as set in the Type node, and the cluster quality falls in the "Good" range.

In the right panel we see a pie chart that illustrates the relative size of each cluster. The smallest cluster has 73 records, or about 5% of the file, which is verging on being too small to be useful. The largest cluster has 443 records. The ratio of sizes of the largest cluster to the smallest cluster is 6.07.

Exploring the Cluster Profiles

More options are provided to explore the cluster profiles in a visual way.

8. Click **Clusters** on the View drop-down list (in the left panel)

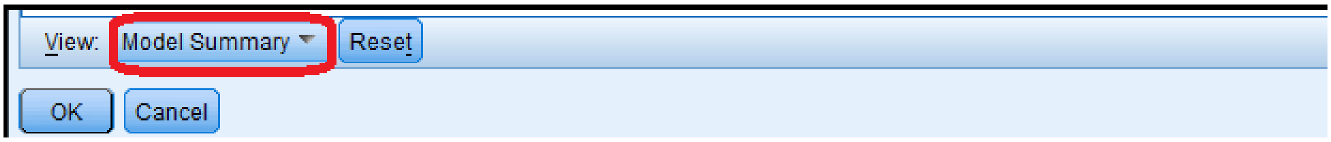
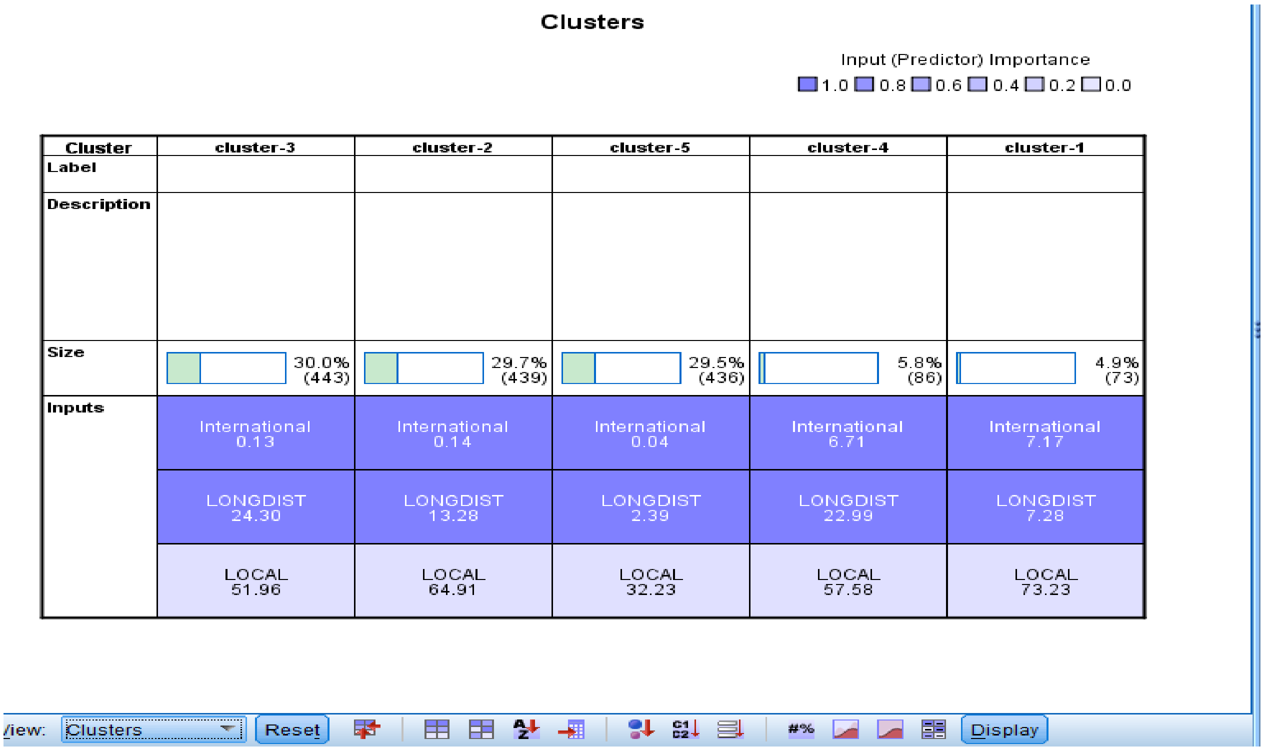


Figure 4



Clusters

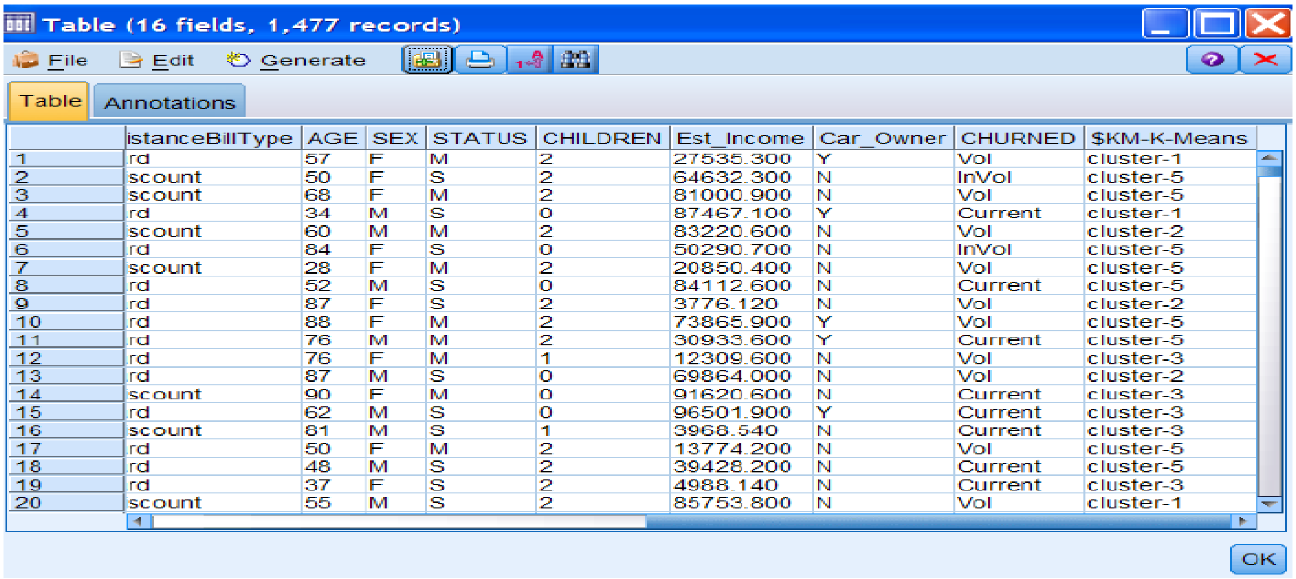
Cluster	cluster-3	cluster-2	cluster-5	cluster-4	cluster-1
Label					
Description					
Size	30.0% (443)	29.7% (439)	29.5% (436)	5.8% (86)	4.9% (73)
Inputs	International 0.13	International 0.14	International 0.04	International 6.71	International 7.17
	LONGDIST 24.30	LONGDIST 13.28	LONGDIST 2.39	LONGDIST 22.99	LONGDIST 7.28
	LOCAL 51.96	LOCAL 64.91	LOCAL 32.23	LOCAL 57.58	LOCAL 73.23

Figure 5

We can now easily compare the clusters. We see that three of the clusters (2, 3, and 5) have the majority of cases. Cluster 5 includes customers who use their phone very little, as they have the smallest mean minutes of usage for all three input fields. Customers in cluster 3 use more long distance minutes, while those in cluster 2 are similar but use fewer long distance minutes and more local minutes.

It is too early to tell whether this cluster solution is useful, as it needs to be examined in terms of the goals of the data-mining project as well as business and organizational knowledge about mobile phone customers.

9. Close the K-Means generated model window →Place another **Table** node in the Stream canvas to the right of the **K-Means generated model** node →**Connect** the **K-Means generated model** node to the **Table** node →Run the **Table** node →Scroll to the **right** in the Table window (shown below)

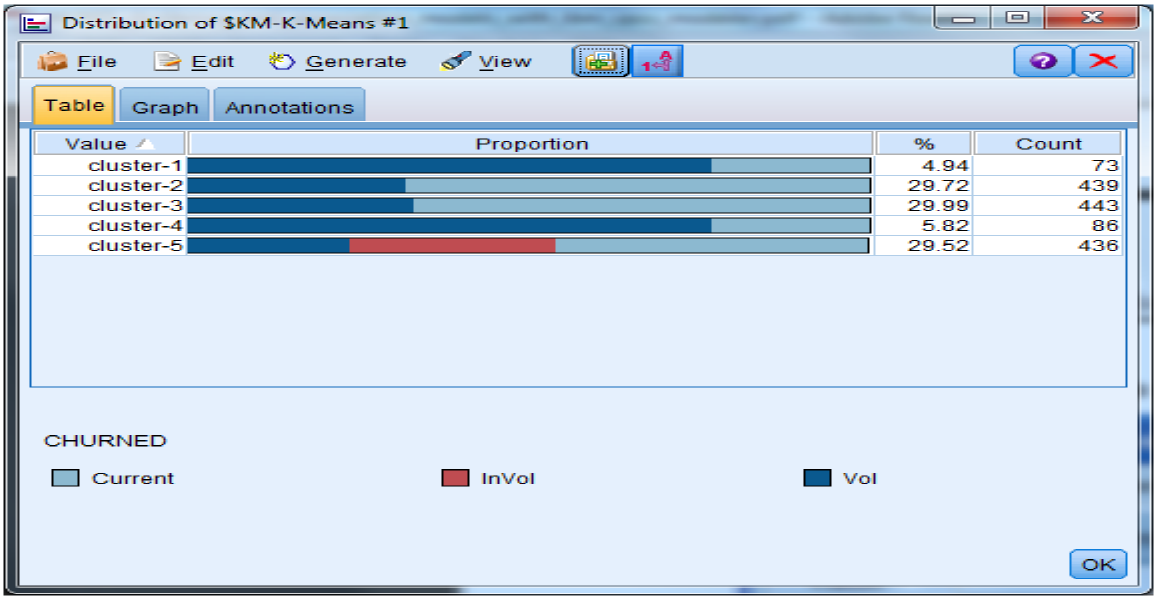


	istanceBillType	AGE	SEX	STATUS	CHILDREN	Est_Income	Car_Owner	CHURNED	$KM-K-Means
1	rd	57	F	M	2	27535.300	Y	Vol	cluster-1
2	scount	50	F	S	2	64632.300	N	InVol	cluster-5
3	scount	68	F	M	2	81000.900	N	Vol	cluster-5
4	rd	34	M	S	0	87467.100	Y	Current	cluster-1
5	scount	60	M	M	2	83220.600	N	Vol	cluster-2
6	rd	84	F	S	0	50290.700	N	InVol	cluster-5
7	scount	28	F	M	2	20850.400	N	Vol	cluster-5
8	rd	52	M	S	0	84112.600	N	Current	cluster-5
9	rd	87	F	S	2	3776.120	N	Vol	cluster-2
10	rd	88	F	M	2	73865.900	Y	Vol	cluster-5
11	rd	76	M	M	2	30933.600	Y	Current	cluster-5
12	rd	76	F	M	1	12309.600	N	Vol	cluster-3
13	rd	87	M	S	0	69864.000	N	Vol	cluster-2
14	scount	90	F	M	0	91620.600	N	Current	cluster-3
15	rd	62	M	S	0	96501.900	Y	Current	cluster-3
16	scount	81	M	S	1	3968.540	N	Current	cluster-3
17	rd	50	F	M	2	13774.200	N	Vol	cluster-5
18	rd	48	M	S	2	39428.200	N	Current	cluster-5
19	rd	37	F	S	2	4988.140	N	Current	cluster-3
20	scount	55	M	S	2	85753.800	N	Vol	cluster-1

Figure 6

Each record has a value for the new field $KM-K-Means, which records cluster membership.

10. Close the **Table** window → Place a **Distribution** node from **Graphs Palette** in the Stream canvas near the **K-Means** generated model node and connect them → Edit the **Distribution** node (not shown) → Select **$KM-K-Means** in the **Field** box and **CHURNED** in the **Overlay Color** field box → Select the **Normalize by color** checkbox →Click **Run.**



Value	Proportion	%	Count
cluster-1		4.94	73
cluster-2		29.72	439
cluster-3		29.99	443
cluster-4		5.82	86
cluster-5		29.52	436

CHURNED

▨ Current 🟥 InVol 🟦 Vol

Figure 7

We see that, interestingly, all involuntary churners—those customers dropped by the company—are in cluster 5. This cluster was composed of customers with generally low usage. That all of them fall into one cluster may be a tangible indicator that this cluster solution is useful. Thus, we may find that predicting which customers will need to be dropped (because of late bill payment or other problems) will be more accurate by including cluster membership in a model.

Voluntary churners, often the most critical customer group, tend to concentrate in clusters 1 and 4, which are also the smallest clusters. This may not be satisfactory and thus may require, despite the argument above, a different cluster solution. Involuntary churners are exclusively associated with Cluster 5, the set of customers who use the phone very little and therefore are apparently dropped by the company.

11. When there are only a few fields used in clustering, and they are continuous measurement, a scatterplot can be useful to visualize the clustering solution.

Close the **Distribution graph** window → Place a **Plot** node in the Stream canvas near the **K-Means** model node and connect them →Edit the **Plot** node →Select **LONGDIST** as the **X field**, **International** as the **Y field**, and **$KM-K-Means** as the **Overlay Color** field → Run the **Plot** node

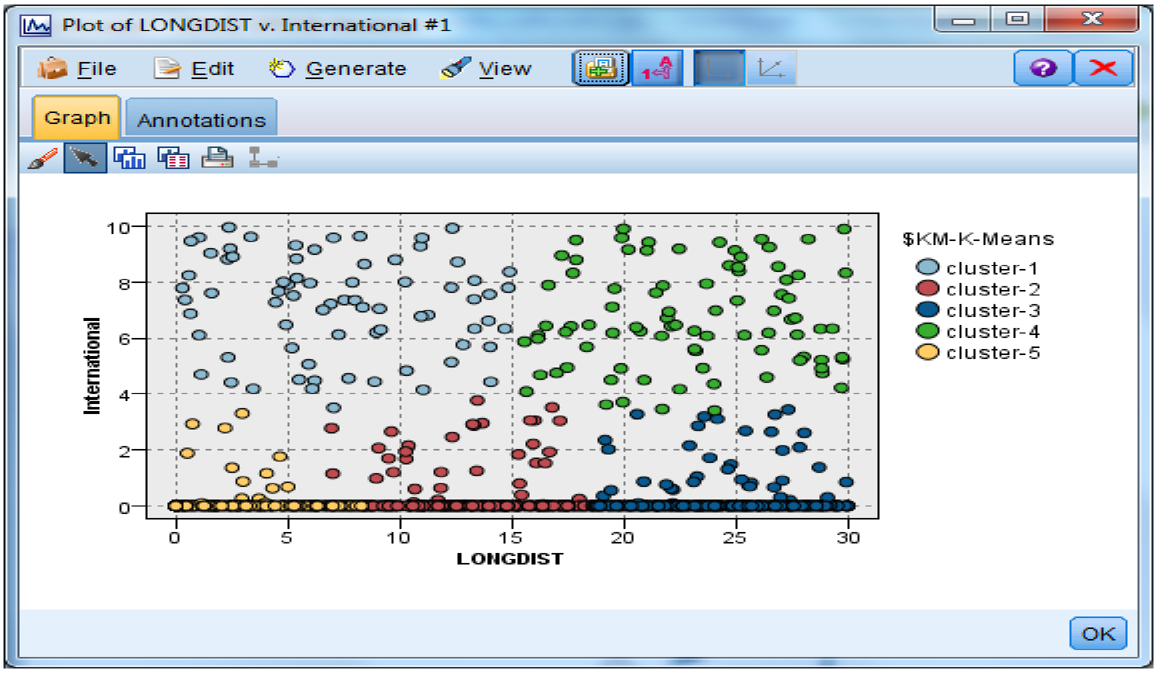


Figure 8

We see graphically that Cluster 5 is composed of customers who have relatively low usage on both fields, while cluster 4 contains customers who have high usage on both types of calls. If you refer back to Figure 2.6, you can see why cluster 1's distance, or proximity, was greatest to cluster 3.

We can also observe how tightly, or loosely, the clusters are grouped, i.e., how restricted their range is in the two dimensions. Clusters 2 and 5 are especially tightly grouped (in the bottom left half of the plot), as compared to clusters 1 and 4. Note how the data are spread throughout the two-dimensional space, with no one cluster truly well-separated from any other. Remember that we got a five-cluster solution because that is what we requested, not because the data naturally have five separate groupings. This is why other cluster solutions should be requested when running K-Means.

12. Close the **Plot** window

13. The Clustering stream

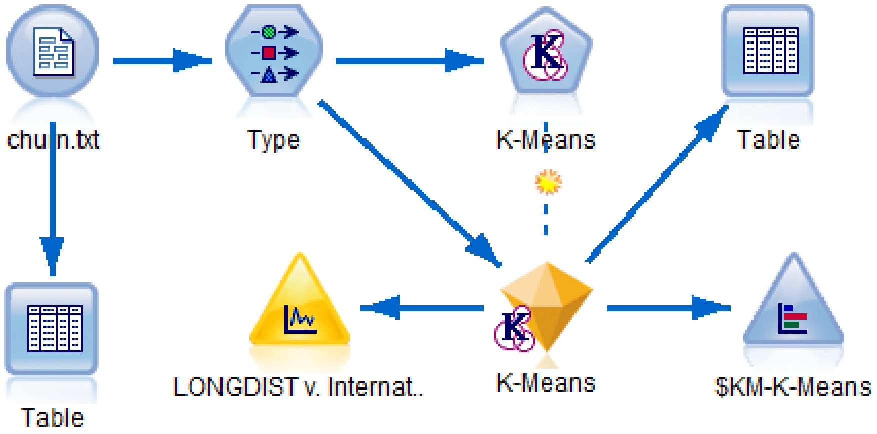


Figure 9

Notes:

What to Look For When Clustering

There are some standard principles that can be applied to any clustering solution. We mention the most critical in this section.

Number of Records per Cluster

The goal of cluster analysis is to create groupings of cases that reflect natural groupings in the data. But to be useful, clusters should not be too small in size and contain only a few cases. While it is certainly possible for a few outlying cases to form their own cluster, 5 or 10 cases in a dataset of 1,000 records are too tiny a cluster to be practically useful in the great majority of circumstances.

There is no hard and fast rule concerning the minimum number of cases in a cluster, but 5-10% of the file size is a reasonable limit to use upon first examination. If clusters are found that are too small, you have several options:

- Rerun the analysis requesting a solution with fewer clusters.
- Combine the small clusters with others that are nearby (in distance), although this is most easy to do with K-means where distances between clusters can be readily output.
- Drop the cluster from the final solution. This is reasonable when the percentage of cases in a small cluster is truly a tiny fraction of the file, or the cases are uninteresting for other reasons.

Note that points 3 suggest that the final clustering solution will not apply to all records (also known as incomplete coverage). This strategy is fairly common in data mining, where the overall goal is not necessarily to model all the records but instead to find the most promising relationships.

Number of Clusters

The intent of cluster analysis is to reduce the dimensionality of the data so that a relatively small number of clusters can represent hundreds or thousands of cases and several variables. However, if a data file is large, with tens of thousands of records, and many variables are used in clustering, then even a clustering solution with 100 clusters will be a significant reduction in dimensionality.

Nonetheless, a solution with so many clusters is rarely helpful or practically useful. This is because, as we discuss in the next section, cluster solutions cannot be accepted on face value but must instead be validated to determine their adequacy. Trying to understand the characteristics of many clusters is time-consuming. Thus we recommend limiting the number of clusters to at most a dozen or so, unless you have strong reasons to believe otherwise.

Validation

A clustering solution should not be accepted until it has been validated. There are at least three approaches to validation.

First, differences between the clusters should be investigated, using both the fields included in the clustering and then other important fields. If there is an outcome field that you hope to predict it should typically not be included among the fields used to create the clusters, but you can observe how it varies across the clusters. In general you examine a profile of each cluster so that the characteristics of each cluster can be described both numerically and in words. With this information in hand, you can then determine whether or not the clusters are truly different in ways that you find substantively important (because the clustering solution does not guarantee this will be true).

Often in modeling we use training and test datasets, and although this is less commonly done in clustering, there is no reason why this technique can't be applied. A test dataset can be created (perhaps the same one to be used later in modeling), and a second cluster analysis done, attempting to replicate the cluster solution found in the training data. If the same basic clusters are found in the test data (the clusters will never be identical), then you have more confidence that the clustering solution will apply to future data.