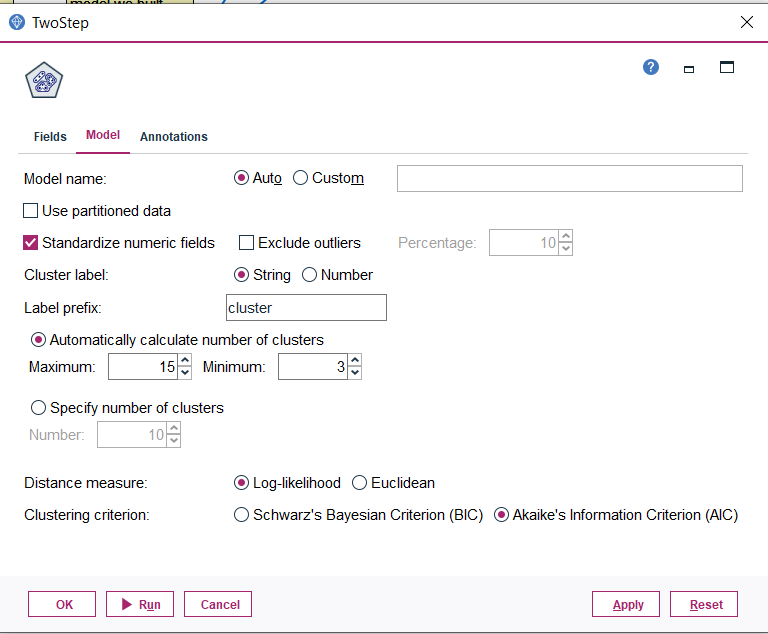
I believe we were also talking about AIC for cluster selection.

<https://www.ibm.com/support/pages/choosing-number-clusters-k-means-cluster-analysis>

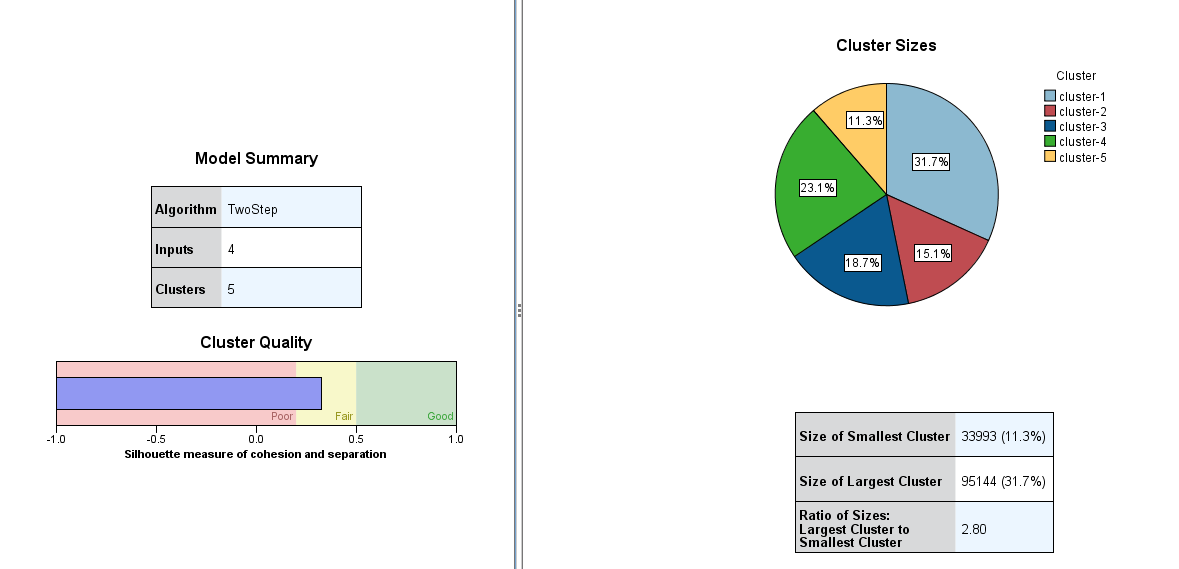
Per the IBM documentation (it is the same for SPSS and modeler), they do not have a way to generate an AIC curve to select the optimal number of clusters as we did with the R class for example. It is built into the two-step node.

If you look about halfway down, “Automatically calculate number of clusters” is selected. For the criterion, you can select BIC or AIC but I believe you were asking about AIC. Unfortunately, they do not have a good graph of that at least to what I know or could figure out in trying to produce one. I did do a hack to pull it out of SPSS and run some R code to generate the graph, but that is another matter.

Default for number of cluster minimum is 2. I would consider between 3 and 5 to be normal expectations. Two seems to be a bit too few, and once you get beyond five you normally start getting these tiny slices.

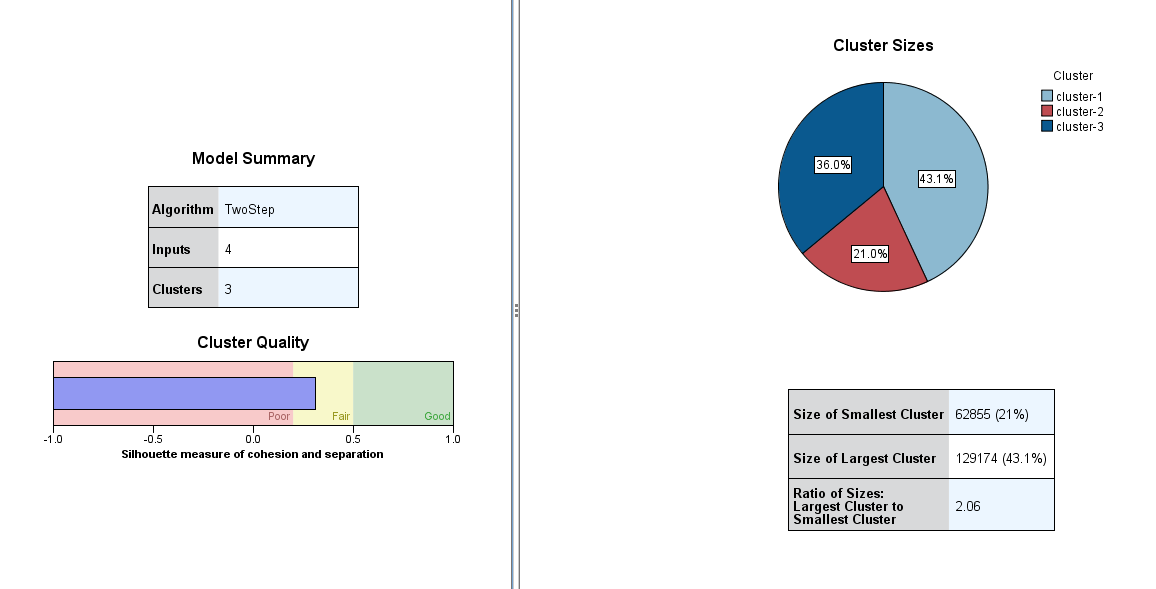


The model it produced looks something like this:



1. Settings included no removal of outliers, minimum number of clusters set to three, and standardized numeric fields. This will take some of the variance out of the data.
2. Silhouette is fair. That means the cohesion (internal distances within a cluster) and separation (distances between clusters) is fair. In this case a 0.3. By comparison, the k-means was producing 0.5 which is right on the line for good.
3. It produced five clusters and if you look at the spread these are just about ideal.
4. The ratio of large to small cluster is 2.80 which is also good.

Let us change it up a bit. Set the model to exclude outliers and this is the result:



1. Silhouette remained unchanged at 0.3.
2. We got fewer clusters, but the distribution looks to have improved.
3. Ratio for large to small has gone down to 2.06 which is even better.

For the two-step using the AIC and auto cluster selection, I would say eliminating the outliers would produce a slightly better model. Both are a bit less in terms of quality than the K-Means that we produced. That is what would lead me to believe a K-means with some outlier removal would be the best for this situation.