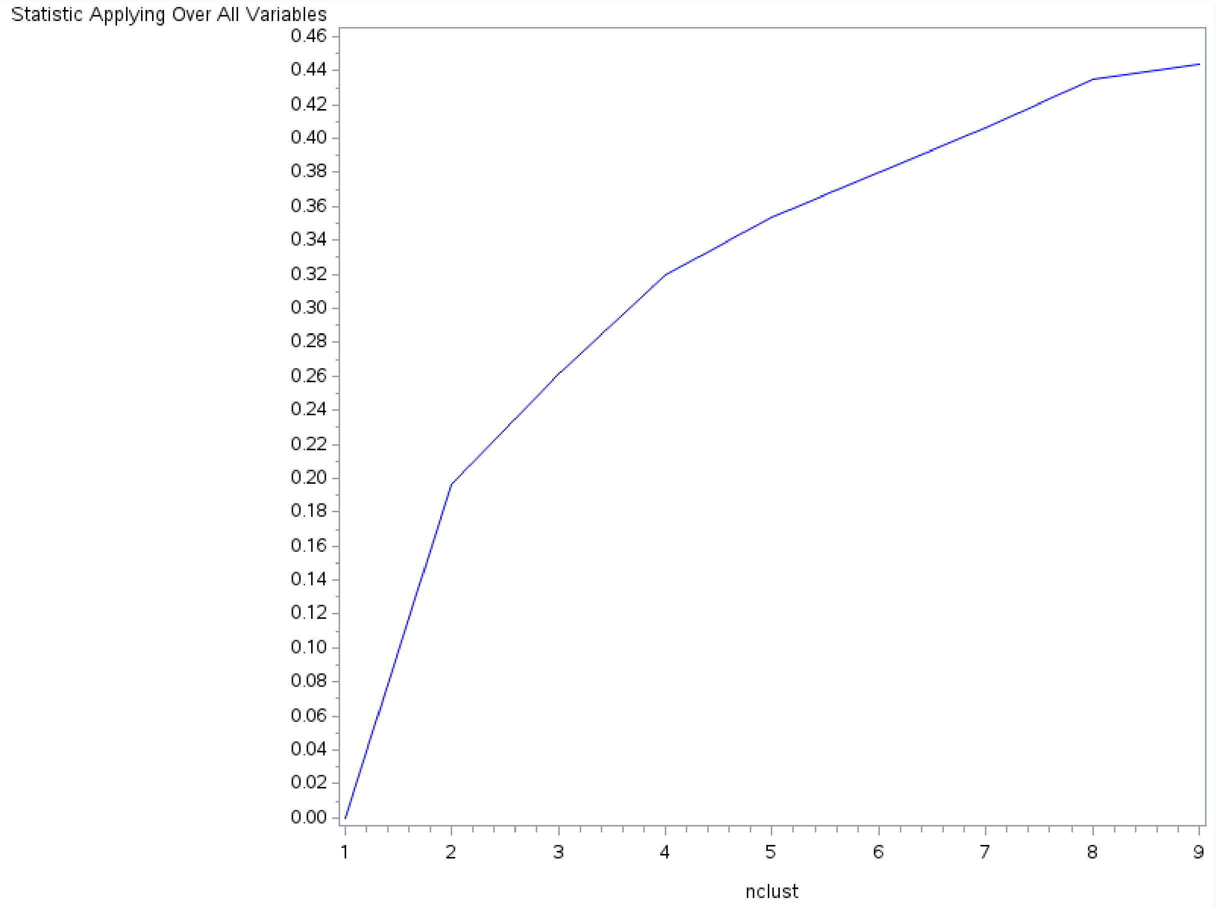
Example of how to write results for a k-means cluster:

A k-means cluster analysis was conducted to identify underlying subgroups of adolescents based on their similarity of responses on 11 variables that represent characteristics that could have an impact on school achievement. Clustering variables included two binary variables measuring whether or not the adolescent had ever used alcohol or marijuana, as well as quantitative variables measuring alcohol problems, a scale measuring engaging in deviant behaviors (such as vandalism, other property damage, lying, stealing, running away, driving without permission, selling drugs, and skipping school), and scales measuring violence, depression, self-esteem, parental presence, parental activities, family connectedness, and school connectedness. All clustering variables were standardized to have a mean of 0 and a standard deviation of 1.

Data were randomly split into a training set that included 70% of the observations (N=3201) and a test set that included 30% of the observations (N=1701). A series of k-means cluster analyses were conducted on the training data specifying k=1-9 clusters, using Euclidean distance. The variance in the clustering variables that was accounted for by the clusters (r-square) was plotted for each of the nine cluster solutions in an elbow curve to provide guidance for choosing the number of clusters to interpret.

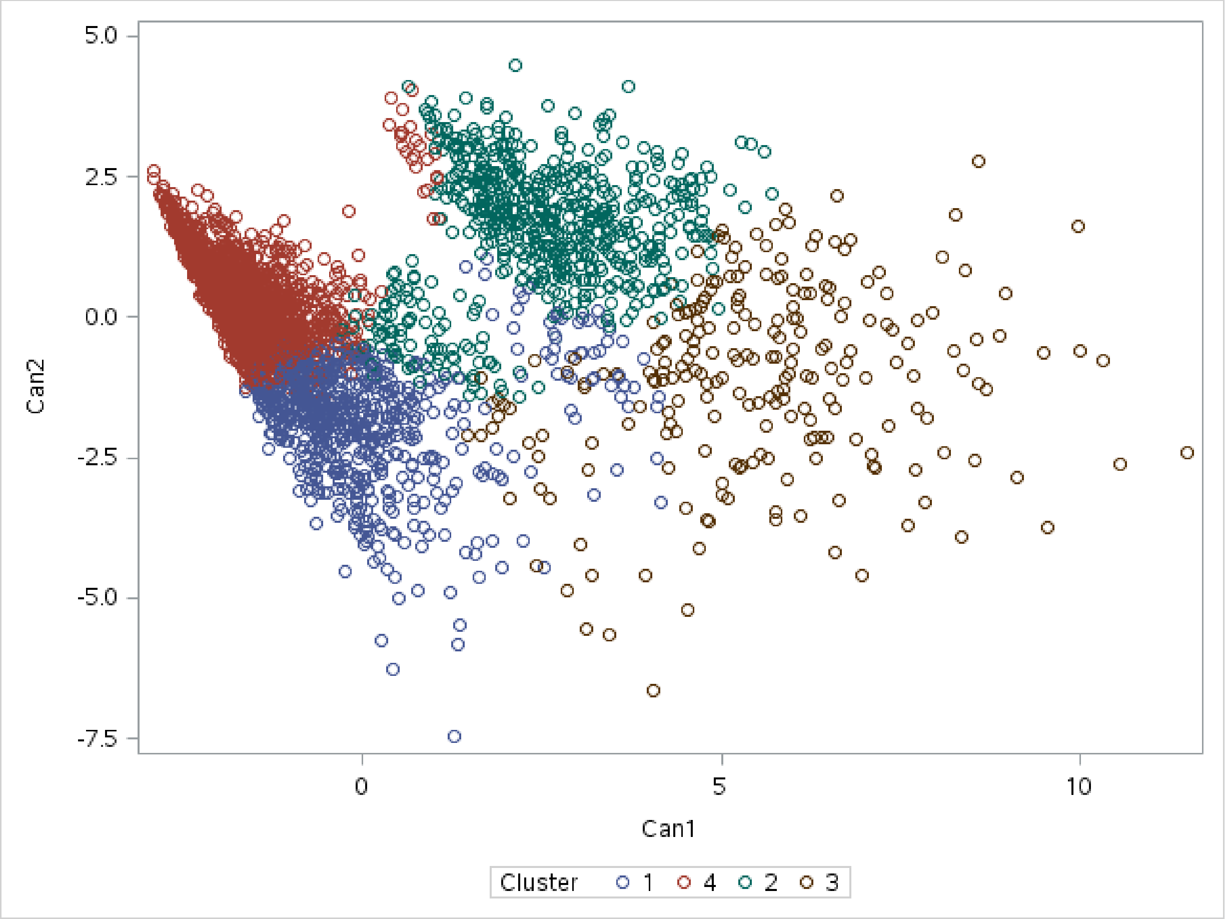
Figure 1. Elbow curve of r-square values for the nine cluster solutions



The elbow curve was inconclusive, suggesting that the 2, 4 and 8-cluster solutions might be interpreted. The results below are for an interpretation of the 4-cluster solution.

Canonical discriminant analyses was used to reduce the 11 clustering variable down a few variables that accounted for most of the variance in the clustering variables. A scatterplot of the first two canonical variables by cluster (Figure 2 shown below) indicated that the observations in clusters 1 and 4 were densely packed with relatively low within cluster variance, and did not overlap very much with the other clusters. Cluster 2 was generally distinct, but the observations had greater spread suggesting higher within cluster variance. Observations in cluster 3 were spread out more than the other clusters, showing high within cluster variance. The results of this plot suggest that the best cluster solution may have fewer than 4 clusters, so it will be especially important to also evaluate the cluster solutions with fewer than 4 clusters.

Figure 2. Plot of the first two canonical variables for the clustering variables by cluster.



The means on the clustering variables showed that, compared to the other clusters, adolescents in cluster 1 had moderate levels on the clustering variables. They had a relatively low likelihood of using alcohol or marijuana, but moderate levels of depression and self-esteem. They also appeared to have fairly low levels of school connectedness parental presence, parental involvement in activities and family connectedness. With the exception of having a high likelihood of having used alcohol or marijuana, cluster 2 had higher levels on the clustering variables compared to cluster 1, but moderate compared to clusters 3 and 4. On the other hand, cluster 3 clearly included the most troubled adolescents. Adolescents in cluster three had the highest likelihood of having used alcohol, a very high likelihood of having used marijuana, more alcohol problems, and more engagement in deviant and violent behaviors compared to the other clusters. They also had higher levels of depression, lower self-esteem, and the lowest levels of school connectedness, parental presence, involvement of parents in activities, and family connectedness. Cluster 4 appeared to include the least troubled adolescents. Compared to adolescents in the other clusters, they were least likely to have used alcohol and marijuana, and had the lowest number of alcohol problems, and deviant and violent behavior. They also had the lowest levels of depression, and higher self-esteem, school connectedness, parental presence, parental involvement in activities and family connectedness.

In order to externally validate the clusters, an Analysis of Variance (ANOVA) was conducting to test for significant differences between the clusters on grade point average (GPA). A tukey test was used for post hoc comparisons between the clusters. Results indicated significant differences between the clusters on GPA (F(3, 3197)=82.28, p<.0001). The tukey post hoc comparisons showed significant differences between clusters on GPA, with the exception that clusters 1 and 2 were not significantly different from each other. Adolescents in cluster 4 had the highest GPA (mean=2.99, sd=0.71), and cluster 3 had the lowest GPA (mean=2.36, sd=0.78).

👍 👎 ⚑