# Machine Learning for Data Analysis

**Assignment – Week 4**

**Running a k-means Cluster Analysis**

**By** [martcsusa@gmail.com](mailto:martcsusa@gmail.com)

This week's assignment deals with **k-means Cluster Analysis** and interpretation of results.

Cluster analysis is an unsupervised machine learning method that partitions the observations in a data set into a smaller set of clusters where each observation belongs to only one cluster. The goal of cluster analysis is to group, or cluster, observations into subsets based on their similarity of responses on multiple variables. Clustering variables should be primarily quantitative variables, but binary variables may also be included.

**About My research**

For research purposes of Machine Learning course we are advised to use ADDHEALTH modified dataset and I’m interested in how substances like **alcohol, marijuana, smoking** etc. Affecting adolescent life and what we can predict and up to what level. I will be using ADDHEALTH for the first time, for my previous courses I was using GAPMINDER, so, this is bit new for me.

**The purpose of current research is to determine what ratings are affecting the most GPA score.**

Variable **GPA1 (grade point average up to 4.0)** will be external validation/response variable.

**Sample**

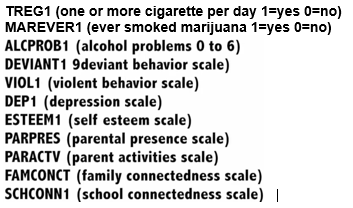
ADDHEALTH - **The sample to** be used represent adolescent various data collected in 2004.

**Procedure**

Data were during 2004.

**Measures (current study)**

Following **target**, **explanatory binary categorical** and e**xplanatory quantitative** variables used in **k-means Cluster Analysis:**



**Program code**

Program code was slightly modified to be executed in 2 MACROs supplying corresponding parameters and can be found in **appendix A**.

**Interpretation of results:**

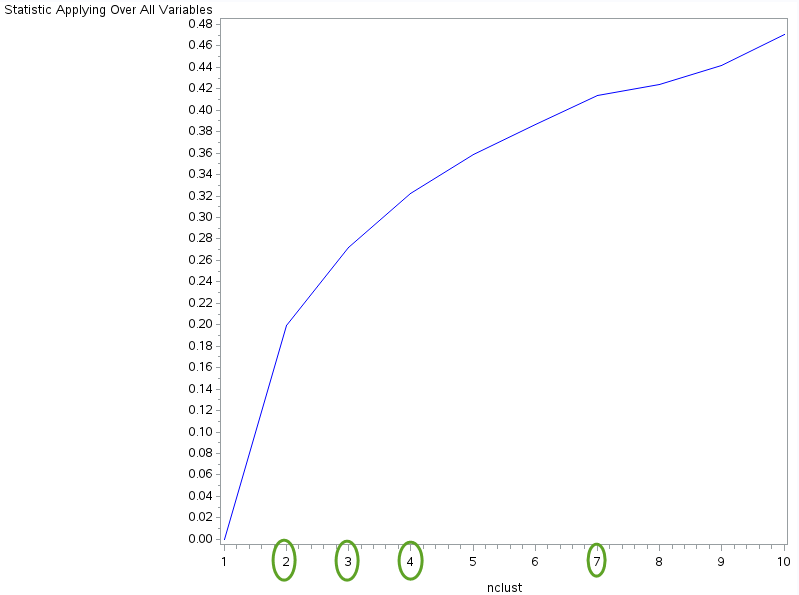
For the purpose of this study a **k-means cluster analysis was conducted to identify underlying subgroups of adolescents based on their similarity of responses on 11 variables that represent characteristics that could have an impact on school achievement**. Clustering variables included two binary variables measuring whether adolescent is regular smoker (**TREG1**) , if adolescent had ever used marijuana (**MAREVER1**) , and quantitative variables measuring alcohol problems (**ALCPROB1**), a scale measuring engaging in deviant behaviors (**DEVIANT1**) (such as vandalism, other property damage, lying, stealing, running away, driving without permission, selling drugs, and skipping school), and scales measuring violence (**VIOL1**), depression (**DEP1**), self-esteem (**ESTEEM1**) family connectedness (**FAMCONCT**), and school connectedness (**SCCONN1**), parental presence (**PARPRES**), parental activities (**PARACTV**).

The training and test sets created with **70% in training (N=3203)** and **30% in test (N=1372)**. Observations with missing values removed prior creation of both sets.

The k-means cluster analyses were applied on the training set to create K=1-10 clusters using Euclidean distance.

All clustering variables were standardized using STANDARD procedure to have a mean of 0 and a standard deviation of 1

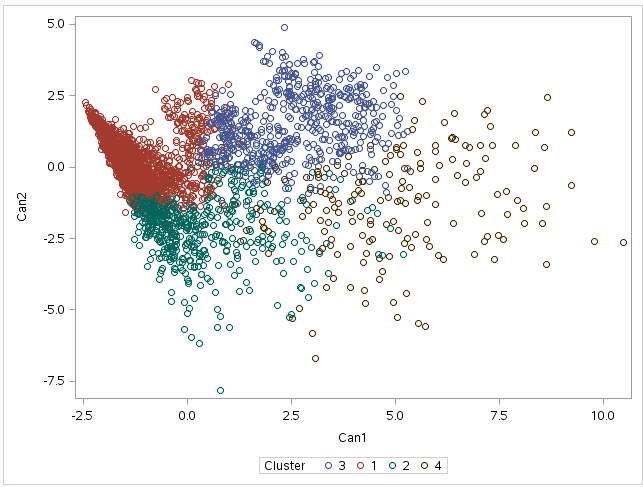
The output of GPLOT: variance in cluster variables (R-square) or elbow curve give us general understanding how significant variable variance changed for each cluster solution.

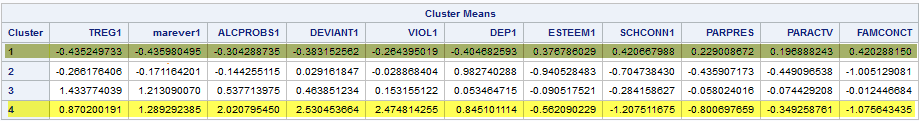


Cluster 2 has the most significant change in variance in cluster variables (20%), cluster 3 has increase about 10% and 4 and 7 Are about 4 % respectively. Based on visual observation of elbow curve solution with 4 and 3 cluster data were executed.

**Four (4) Cluster solution**

Canonical discriminant analyses (PROC CANDISC) was used to reduce the 11 clustering variable down 4 canonical variables representing most of the variance in the clustering variables. Subsequent plot produced graph based on which we can conclude that cluster 1(red) has the most density, but some part of values is very close to cluster 3, which doesn’t has strong concentration and separated into 2 parts. The second the most concentrated cluster is cluster 2, which very close to cluster 1 and has little spread. The most troubled cluster is #4, its values interacting with all clusters. Based on created plot we may try to reduce cluster solution to may be 3 to check how clusters 1, 2 and 3 are separate/if separated from each other.



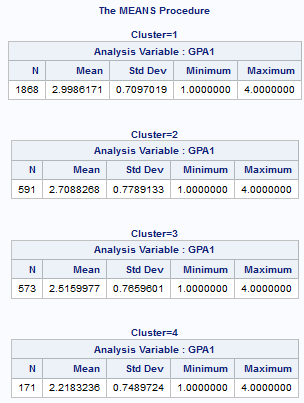


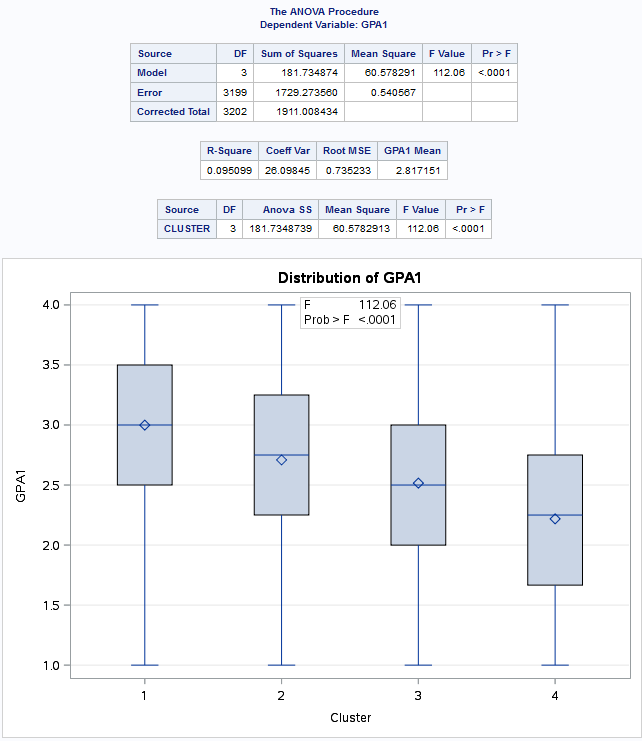
The MEANS table created for 4(four) cluster solution also shows us that **cluster 4 is the most troubled cluster**, where adolescents have very high r-square mean values if they are regular smokers, mostly tried marijuana, having alcohol problems And having deviant and violent behavior. At the same time they probably can have significant level of depression, low self-esteem very low school and family connectedness, and probably having moderately low parent activity. In a contrary, **cluster 1 shows** that adolescents who never tried marijuana and non-smokers have low level of depression, moderate to high self-esteem, moderate parent activity, parent present and high school family consecutiveness.

Interesting results shown for **cluster 2**, where level of depression is very high and at the same time self-esteem is the lowest among all 4 clusters, adolescents showing low school and family consecutiveness, lowest parental activity, moderate deviant behavior and with enough confidence we can say that adolescents in cluster 2 are not regular smokers and probably didn’t try marihuana.

**Cluster 3** is very close to cluster 4 with lower values in all indicators.

To check significant difference in GPA between 4 clusters, the ANOVA test were conducted. A turkey test was used for post hoc comparisons between the clusters. Results indicated significant differences between the clusters on **GPA (F=3, with error=3199) and F value = 112.06, p<.0001**). The **turkey** 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 1 (as expected)** **have the highest GPA** (**MEAN= 2.998, SD=0.71**), and **cluster 4 had the lowest GPA (MEAN=2.21, SD=0.74).**





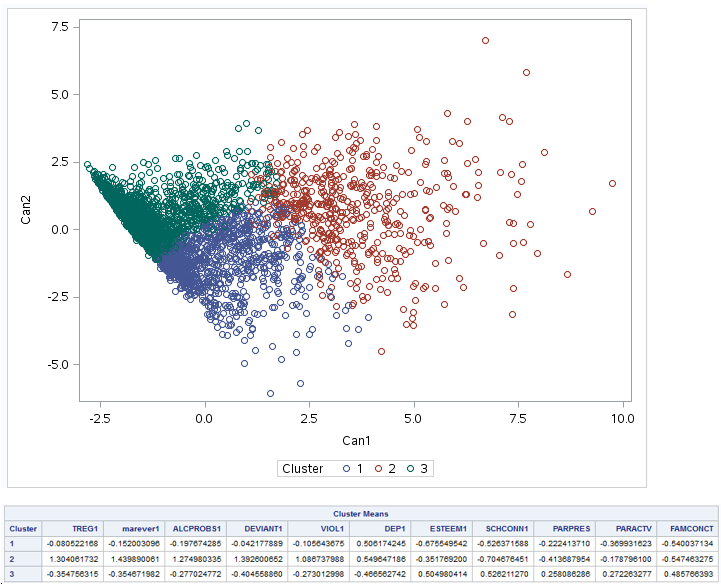
**Three (3) Cluster solution**

Canonical discriminant analyses (PROC CANDISC) was used to reduce the 11 clustering variable down 3 canonical variables representing most of the variance in the clustering variables. Subsequent plot produced graph based on which we can conclude that cluster 3(grey) has the most density, but some part of values is very close to cluster 2, which doesn’t has strong concentration. The second the most concentrated cluster is cluster 1, which very close to cluster 1 and has little spread. The most troubled cluster is #2, its values interacting with all clusters. Based on created plot we may conclude that 3(three) cluster solution looks better than 4(four) clusters solution, clusters are better concentrated.

The MEANS table created for 4(four) cluster solution also shows us that **cluster 2 is the most troubled cluster**, where adolescents have very high r-square mean values if they are regular smokers, mostly tried marijuana, having alcohol problems And having deviant and violent behavior. At the same time they probably can have significant level of depression, low self-esteem very low school and family connectedness, and probably having moderately low parent activity. In a contrary, **cluster 3 shows** that adolescents who never tried marijuana and non-smokers have low level of depression, moderate to high self-esteem, moderate parent activity, parent present and high school family consecutiveness.

Interesting results shown for **cluster 1**, where level of depression is very high and at the same time self-esteem is the lowest among all 4 clusters, adolescents showing low school and family consecutiveness, lowest parental activity, moderate deviant behavior and with enough confidence we can say that adolescents in cluster 2 are not regular smokers and probably didn’t try marihuana.

To check significant difference in GPA between 4 clusters, the ANOVA test were conducted. A turkey test was used for post hoc comparisons between the clusters. Results indicated significant differences between the clusters on **GPA (F=2, with error=3200) and F value = 161.50, p<.0001**). The **turkey** 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 3 (as expected)** **have the highest GPA** (**MEAN= 3.01, SD=0.70**), and **cluster 2 had the lowest GPA (MEAN=2.37, SD=0.78).**

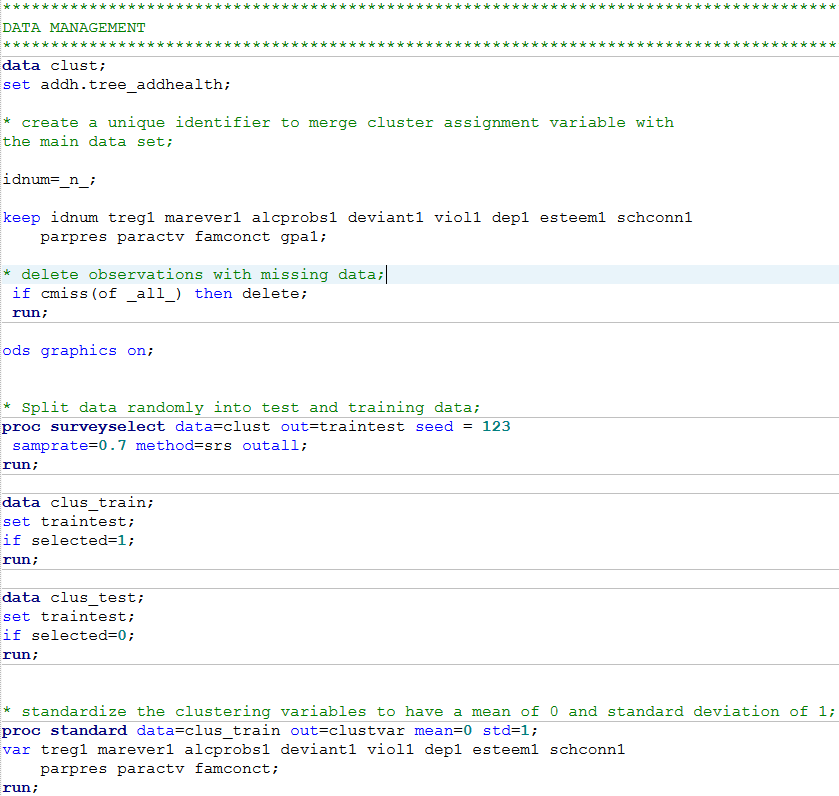


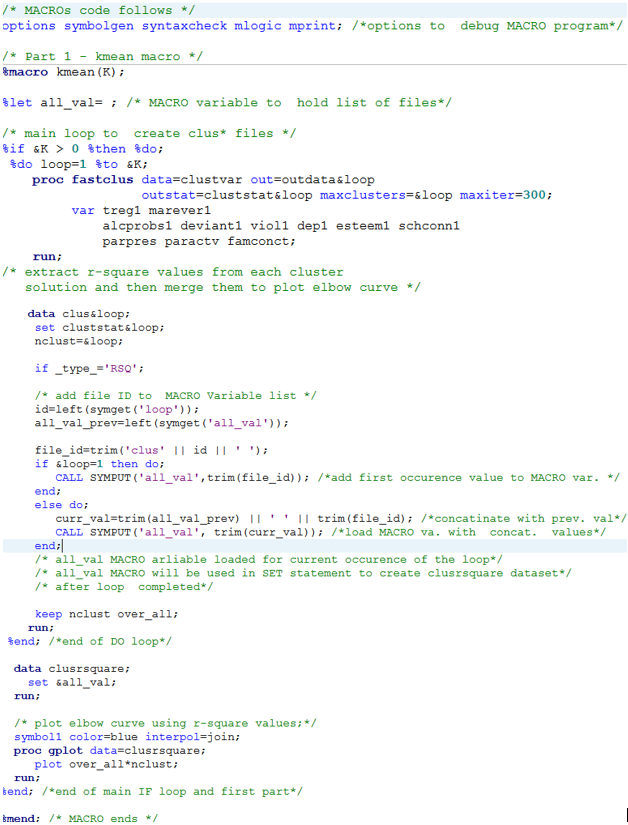


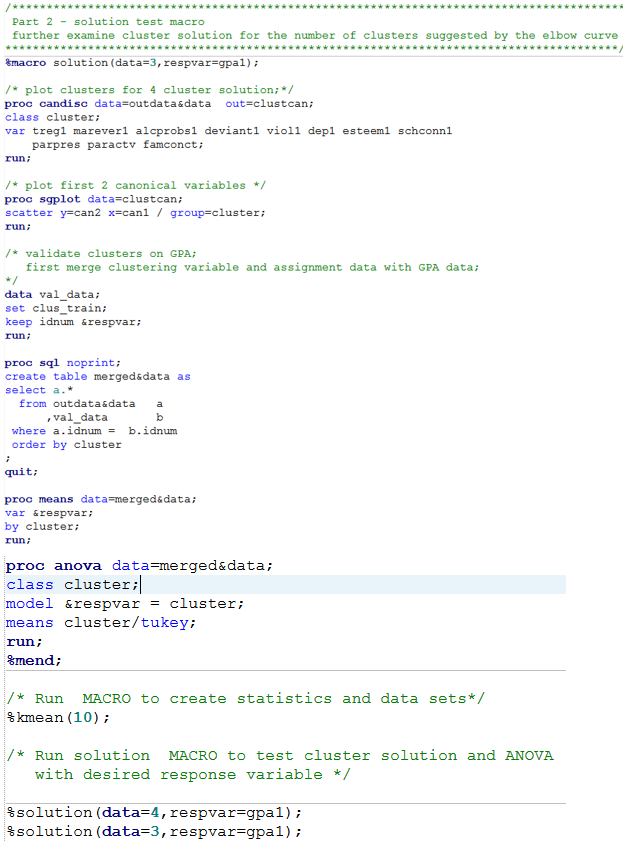
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* end of week 4 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Appendix A.**

Program code snapshots.







http://mapolarbear-da.blogspot.com/2017/01/ml-assignment-week3-lasso.html

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*