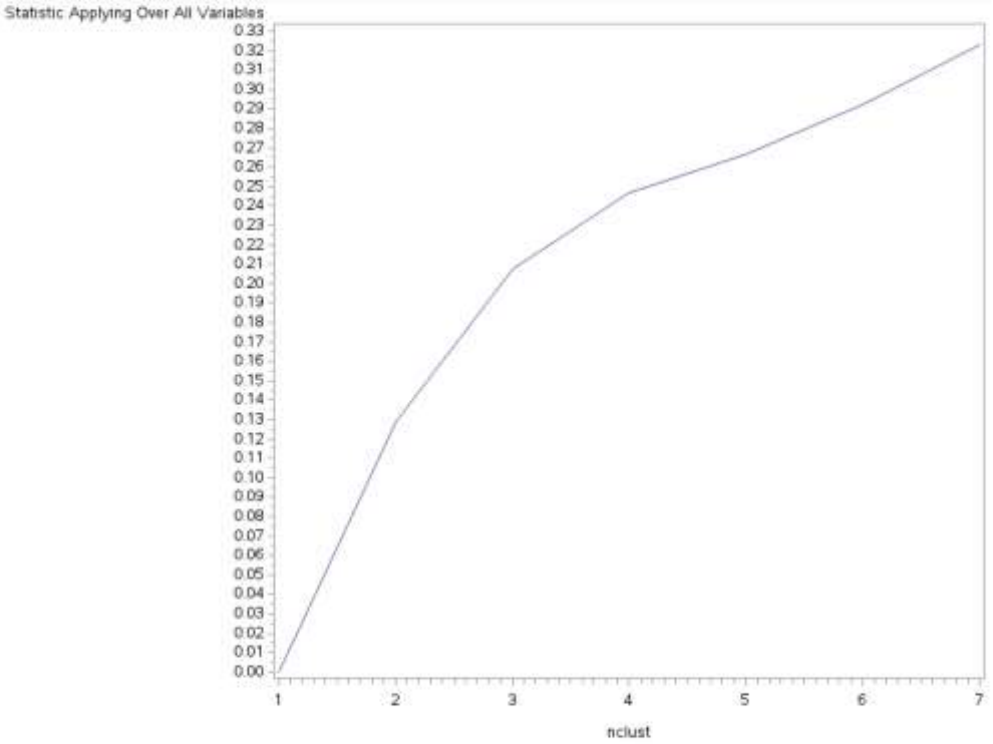
Machine learning for data analysis

Week 4 assignment

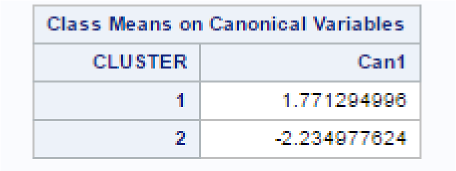
Running a k-means cluster analysis

I chose addhealth as my dataset. 12 variables (H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3) were analyzed. BIO_SEX is a binary variable. H1GH59 and H1GH60 are quantitative variables. Others are ordered categorical variables. Each of them has more than 3 levels, so that they can be treated as quantitative variables.

K=7. There is a bend at nclust equal to 2, 3 and 4. So I further evaluated these cluster solutions.

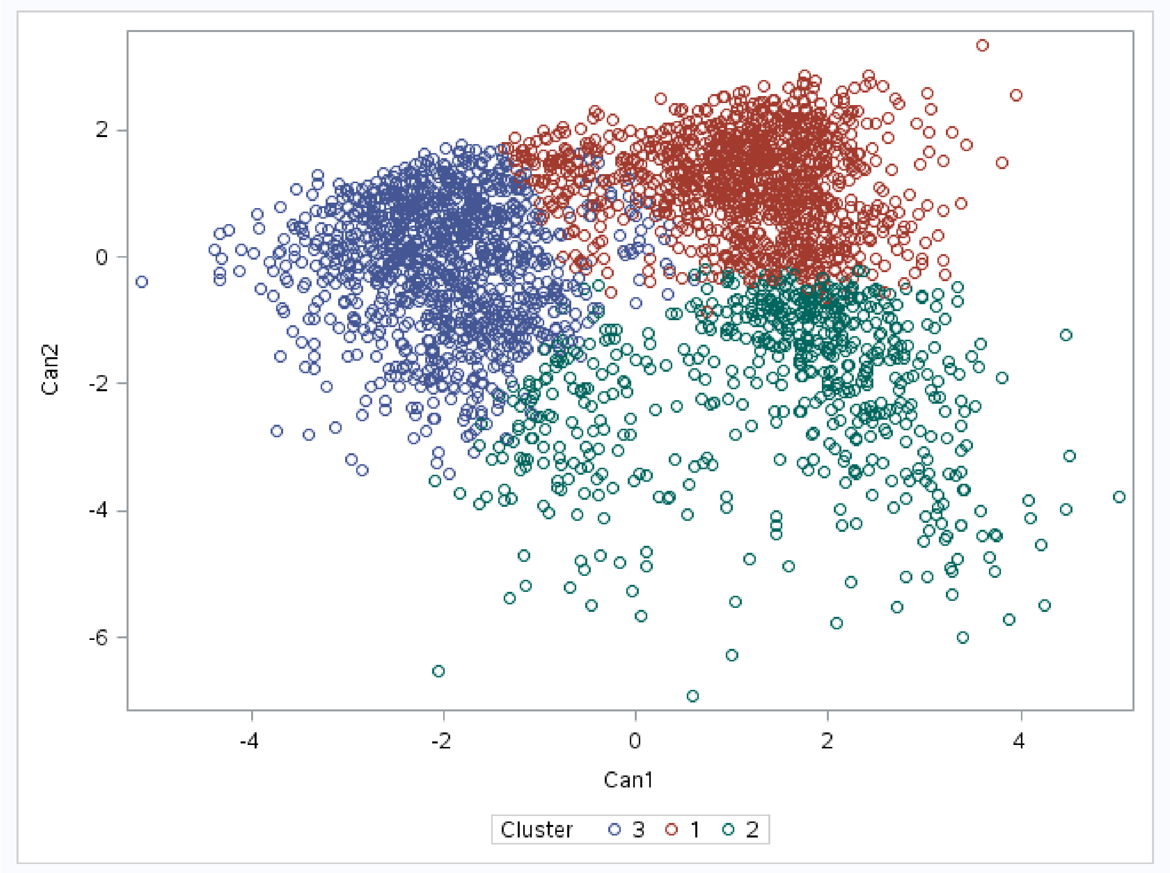


For nclust=2, can2 is invalid.

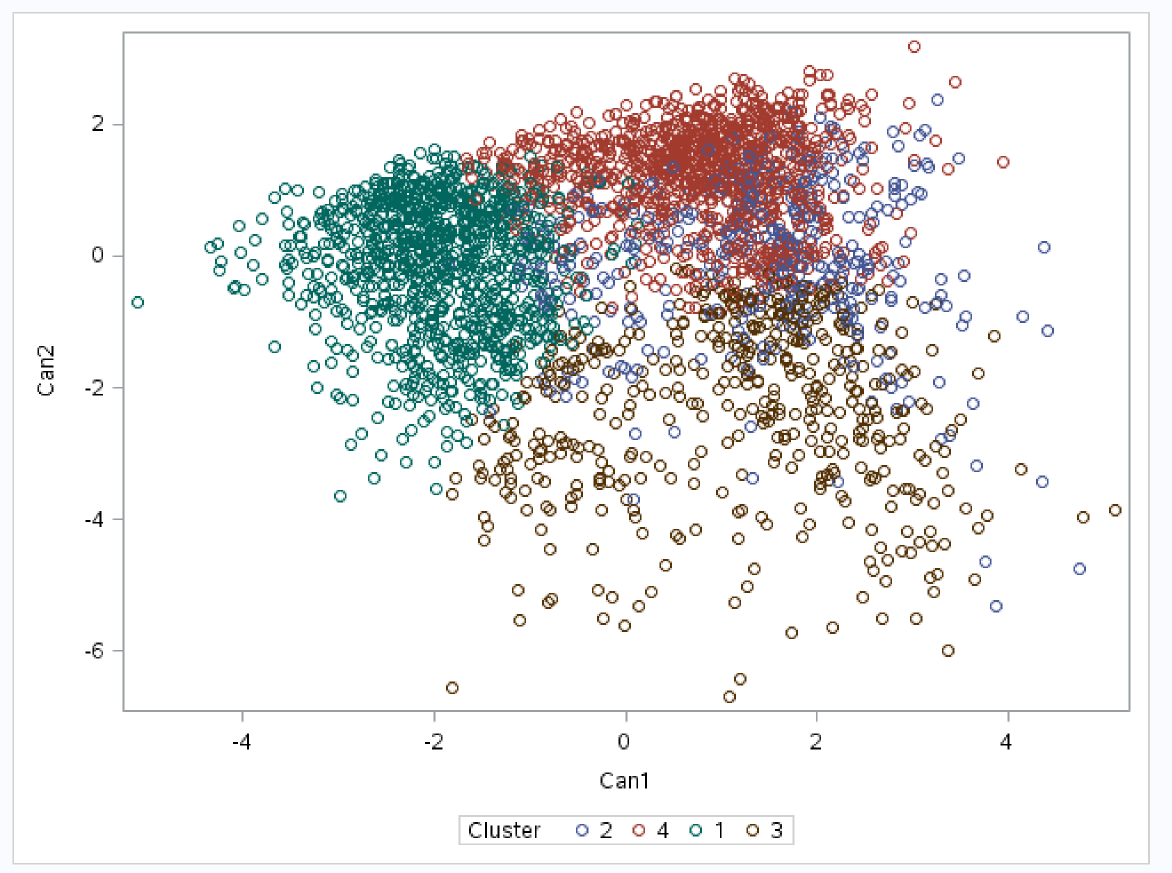


Class Means on Canonical Variables	
CLUSTER	Can1
1	1.771294996
2	-2.234977624

For nclust=3, the cluster 2 is more spread out than the other two clusters, meaning is variance within cluster is much higher. There is almost no overlap between each of the clusters.

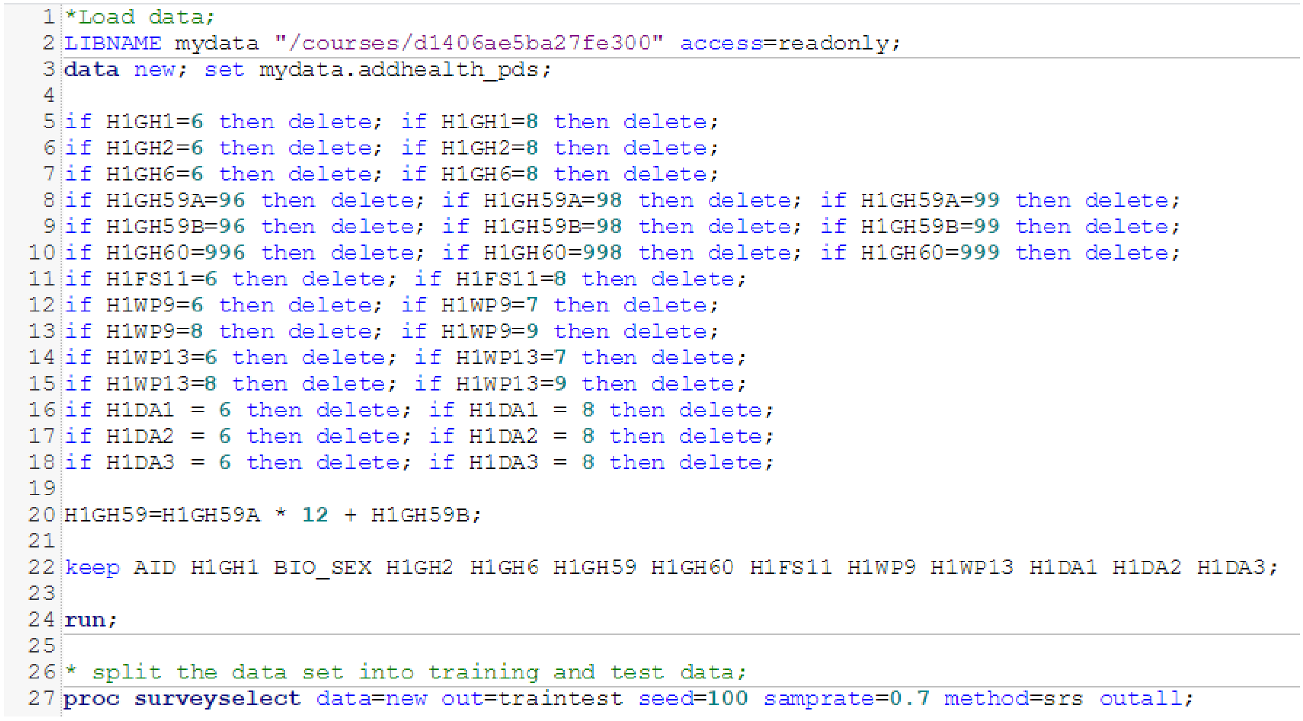


For nclust=4, the most part of cluster 2, cluster 4 and cluster 3 are overlapped, indicating the variance between cluster is not high enough. This should not be the best cluster solution.

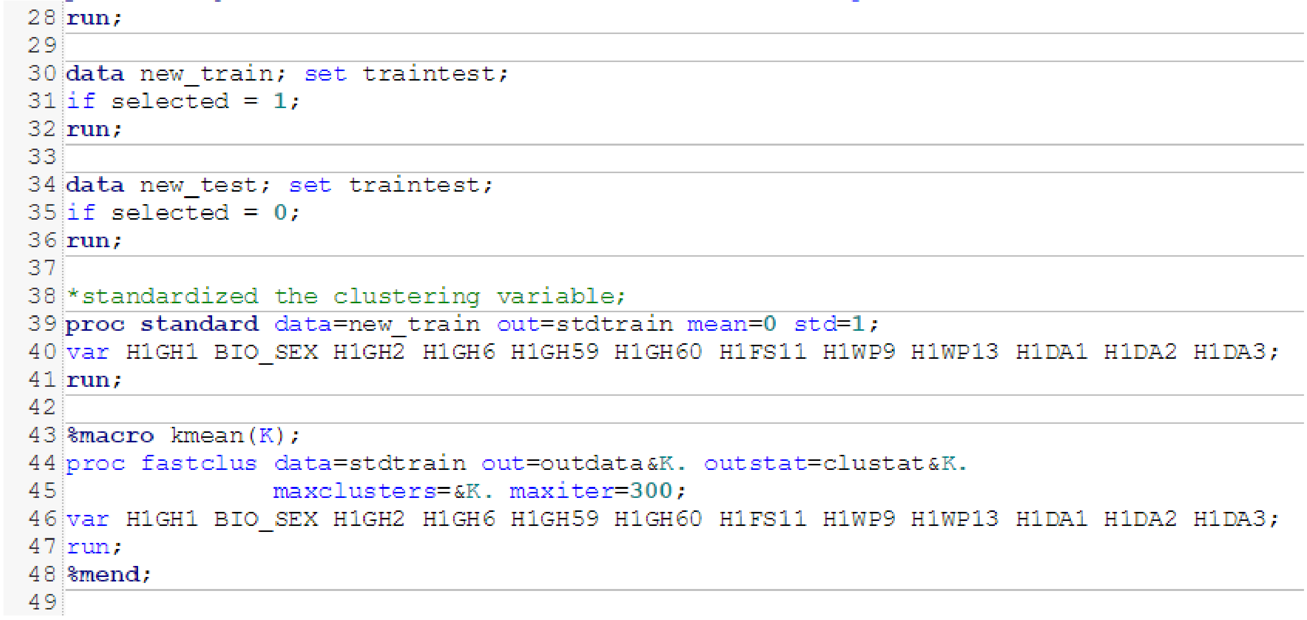


Overall, nclust=3 may be the best cluster solution.

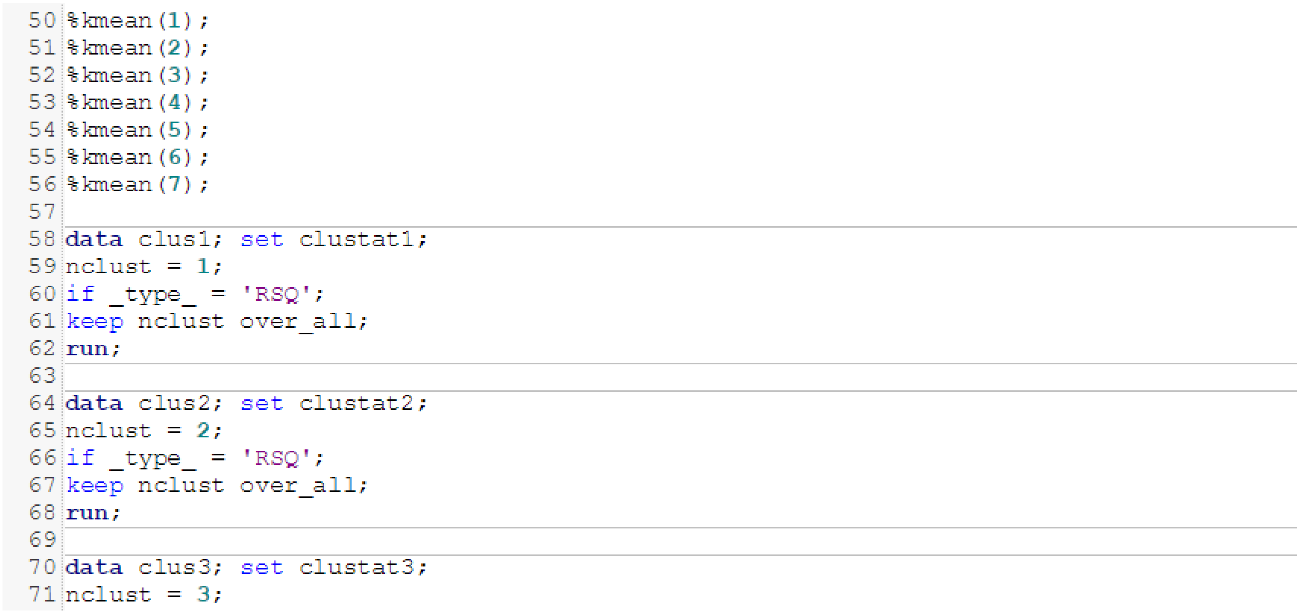
My code:



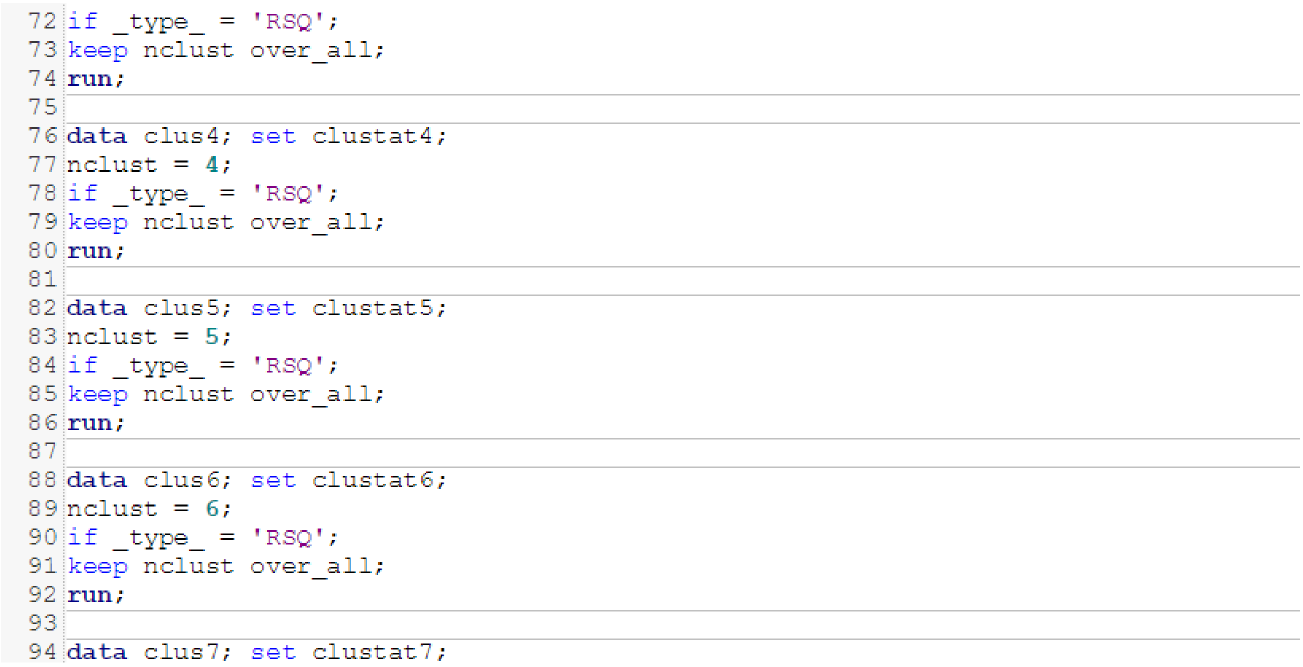
```
*Load data;
LIBNAME mydata "/courses/d1406ae5ba27fe300" access=readonly;
data new; set mydata.addhealth_pds;

if H1GH1=6 then delete; if H1GH1=8 then delete;
if H1GH2=6 then delete; if H1GH2=8 then delete;
if H1GH6=6 then delete; if H1GH6=8 then delete;
if H1GH59A=96 then delete; if H1GH59A=98 then delete; if H1GH59A=99 then delete;
if H1GH59B=96 then delete; if H1GH59B=98 then delete; if H1GH59B=99 then delete;
if H1GH60=996 then delete; if H1GH60=998 then delete; if H1GH60=999 then delete;
if H1FS11=6 then delete; if H1FS11=8 then delete;
if H1WP9=6 then delete; if H1WP9=7 then delete;
if H1WP9=8 then delete; if H1WP9=9 then delete;
if H1WP13=6 then delete; if H1WP13=7 then delete;
if H1WP13=8 then delete; if H1WP13=9 then delete;
if H1DA1 = 6 then delete; if H1DA1 = 8 then delete;
if H1DA2 = 6 then delete; if H1DA2 = 8 then delete;
if H1DA3 = 6 then delete; if H1DA3 = 8 then delete;

H1GH59=H1GH59A * 12 + H1GH59B;

keep AID H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;

run;

* split the data set into training and test data;
proc surveyselect data=new out=traintest seed=100 samprate=0.7 method=srs outall;
```



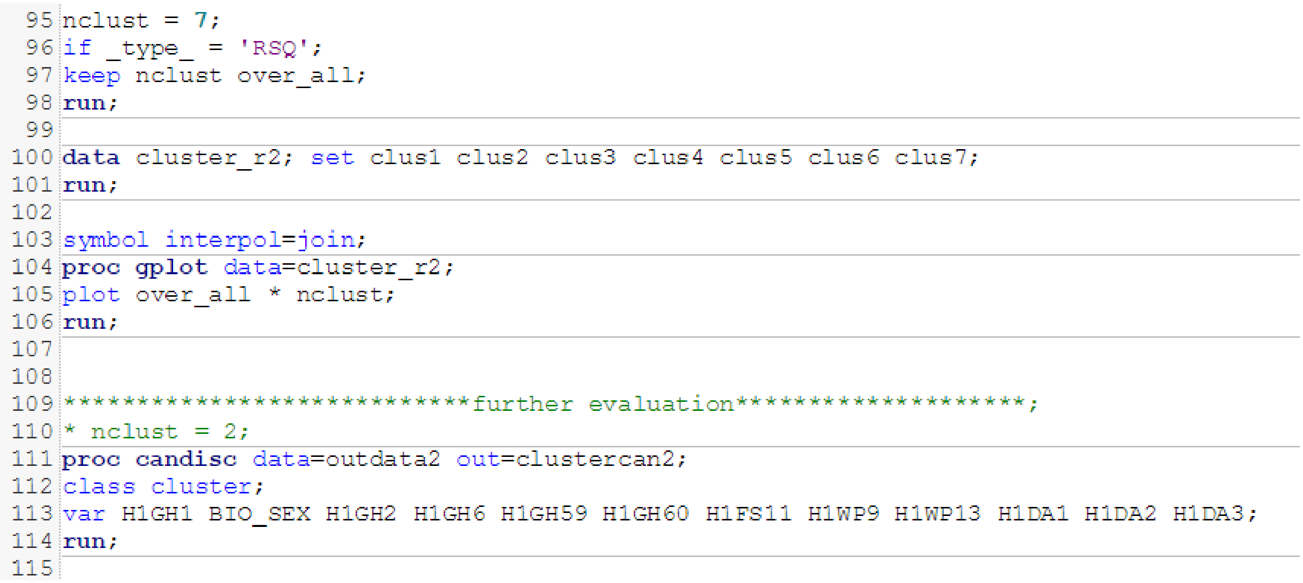
```
run;

data new_train; set traintest;
if selected = 1;
run;

data new_test; set traintest;
if selected = 0;
run;

*standardized the clustering variable;
proc standard data=new_train out=stdtrain mean=0 std=1;
var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
run;

%macro kmean(K);
proc fastclus data=stdtrain out=outdata&K. outstat=clustat&K.
              maxclusters=&K. maxiter=300;
var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
run;
%mend;

```



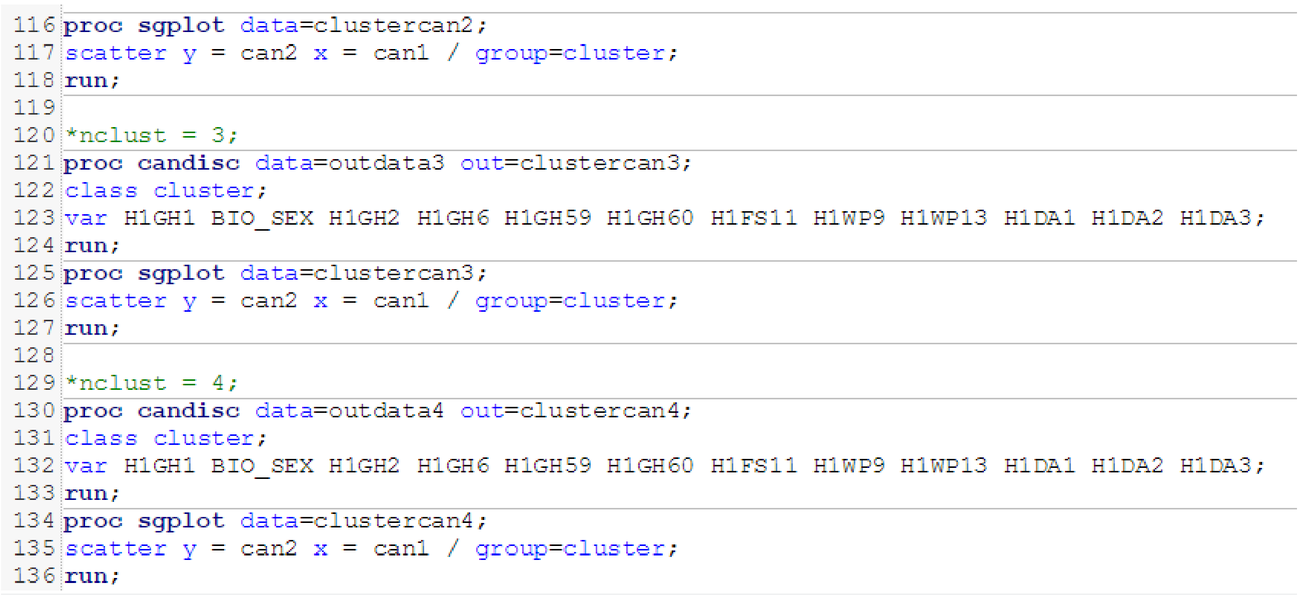
```
%kmean(1);
%kmean(2);
%kmean(3);
%kmean(4);
%kmean(5);
%kmean(6);
%kmean(7);

data clus1; set clustat1;
nclust = 1;
if _type_ = 'RSQ';
keep nclust over_all;
run;

data clus2; set clustat2;
nclust = 2;
if _type_ = 'RSQ';
keep nclust over_all;
run;

data clus3; set clustat3;
nclust = 3;
```



```
if _type_ = 'RSQ';
keep nclust over_all;
run;

data clus4; set clustat4;
nclust = 4;
if _type_ = 'RSQ';
keep nclust over_all;
run;

data clus5; set clustat5;
nclust = 5;
if _type_ = 'RSQ';
keep nclust over_all;
run;

data clus6; set clustat6;
nclust = 6;
if _type_ = 'RSQ';
keep nclust over_all;
run;

data clus7; set clustat7;
```



```
nclust = 7;
if _type_ = 'RSQ';
keep nclust over_all;
run;

data cluster_r2; set clus1 clus2 clus3 clus4 clus5 clus6 clus7;
run;

symbol interpol=join;
proc gplot data=cluster_r2;
plot over_all * nclust;
run;


****************************further evaluation********************;
* nclust = 2;
proc candisc data=outdata2 out=clustercan2;
class cluster;
var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
run;

```



```
proc sgplot data=clustercan2;
scatter y = can2 x = can1 / group=cluster;
run;

*nclust = 3;
proc candisc data=outdata3 out=clustercan3;
class cluster;
var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
run;
proc sgplot data=clustercan3;
scatter y = can2 x = can1 / group=cluster;
run;

*nclust = 4;
proc candisc data=outdata4 out=clustercan4;
class cluster;
var H1GH1 BIO_SEX H1GH2 H1GH6 H1GH59 H1GH60 H1FS11 H1WP9 H1WP13 H1DA1 H1DA2 H1DA3;
run;
proc sgplot data=clustercan4;
scatter y = can2 x = can1 / group=cluster;
run;
```