Data Management and Visualization

Follow thecanadiandataguy

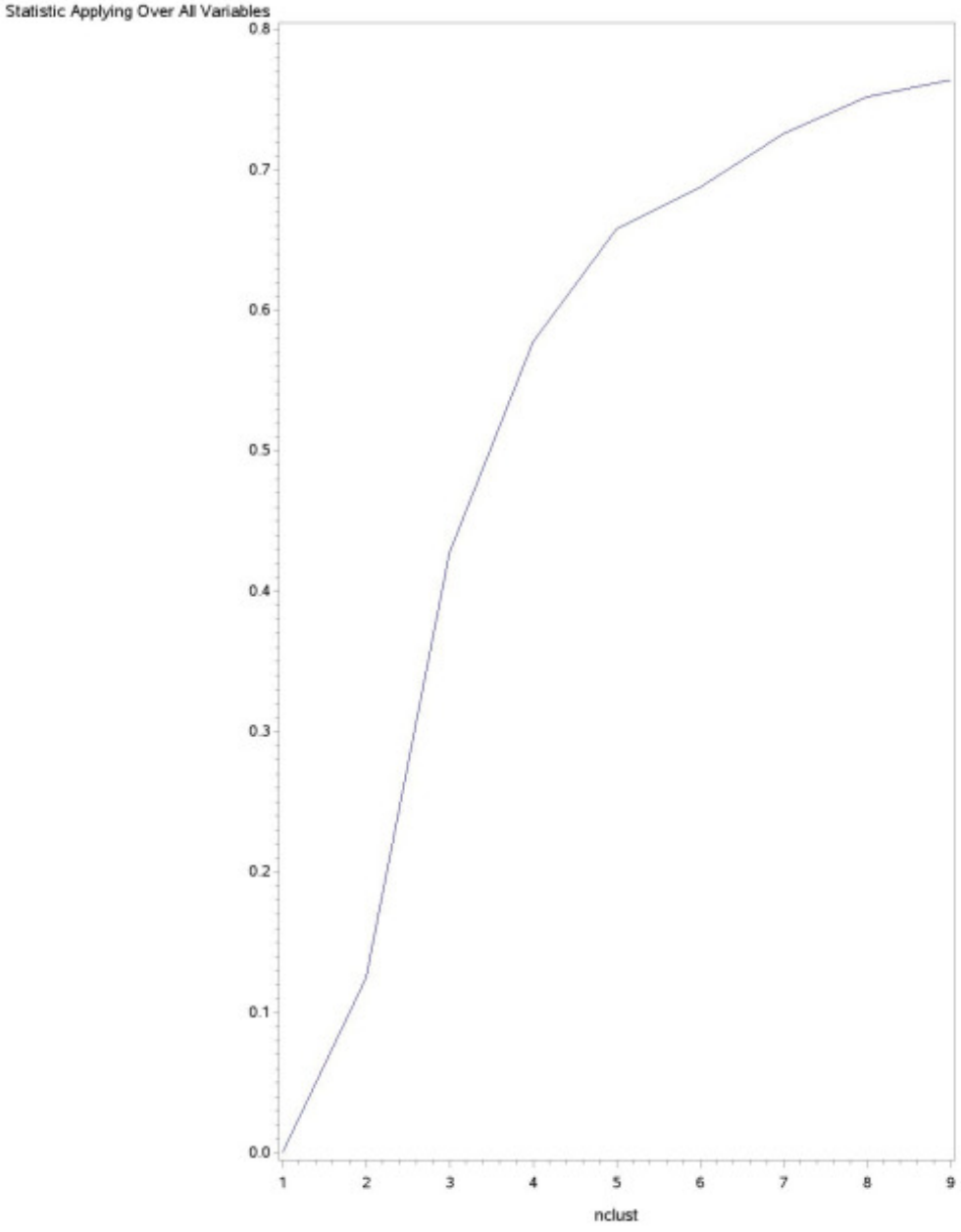
rss
archive

K-means

For the final assignment in machine learning, a k-means cluster analysis was conducted using variables used in previous assignments. The quantitative variables included in the k-means cluster analysis were: ***internetuserate*** (Internet net use per 100 people) ***incomeperperson*** (annual income) co***2emissions*** (annual) ***femaleemployrate*** (percentage females in the work place) ***alcconsumption***(annual alcohol consumption in liters) ***lifeexpectancy*** (average life expectancy) ***employrate*** (employment rate) ***urbanrate*** (percentage of land for urban use). The response variable that I have been looking at is ***breastcancerper100th*** (breast cancer cases per 100000). Data came from the gapminder data set provided through the course. All clustering variables were standardized to have a mean of 0 and a standard deviation of 1.

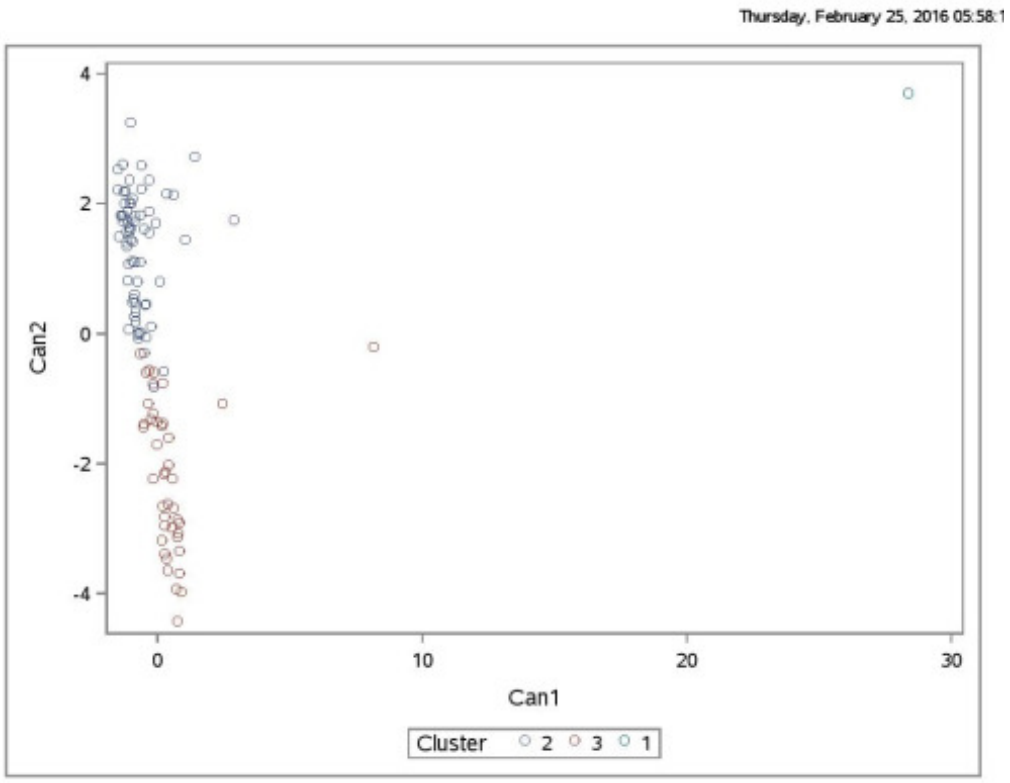
SAS code was included to split the data into a training set representing 70% of the data, 110 countries, and a test set representing 30% or 47 countries. k-mean cluster analysis was created for cluster sizes between 0 and 9, using Euclidean distance . The variance in the clustering variables that was accounted for by the clusters (r-square) was plotted for each of the nine cluster solutions in an elbow curve.

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

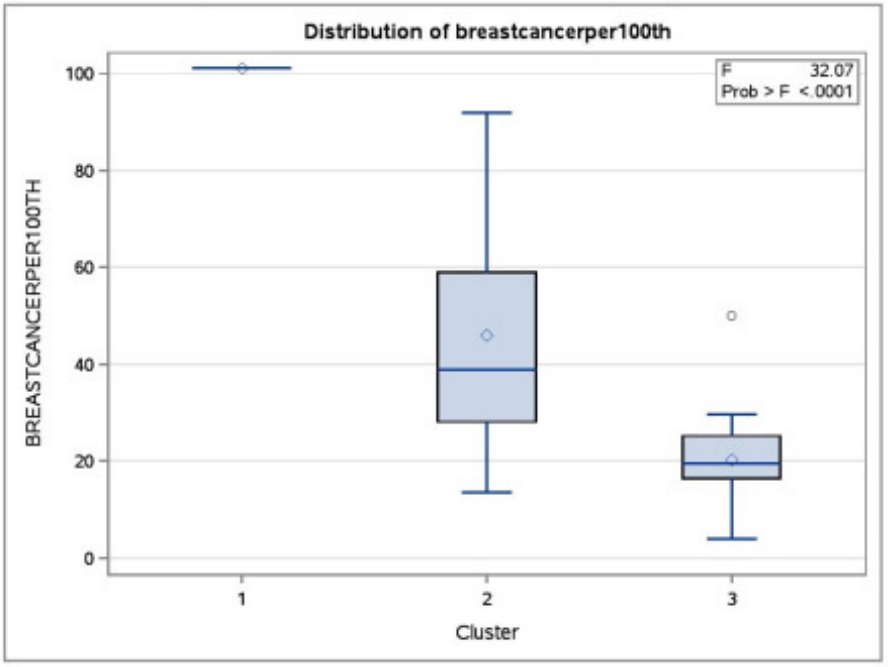


I decide to investigate the analysis with 5 clusters in further detail, although other points above 3 clusters could

be of interest.
Below is the plot of canonical discriminant analyses performed:



and the boxplot produced by the anova procedure:



Cluster 1 was 1 unique country, and as such showed little value for further analysis, but there was a good distinction between cluster 2 and 3 in both the scatterplot (red and blue points) and the boxplot showing little overlap between the regions.

Cluster Means

Cluster	internetuserate	incomeperperson	co2emissions	femaleemployrate	alcconsumption	lifeexpectancy	employrate	urbanrate
1	1.423671795	2.746781071	9.730348022	0.605146671	0.622063282	0.911754404	0.340993082	1.198952553
2	0.530518504	0.324339911	-0.079150669	-0.330096213	0.371281941	0.594906331	-0.383776390	0.577582846
3	-0.880200275	-0.582798932	-0.105410791	0.512173800	-0.607094126	-0.970725681	0.604095834	-0.949928649

Cluster 2 shows higher levels of internet use, higher income per person , alcohol consumption, and urbanization than cluster 3.
The anova analysis including tukey post hoc analysis was performed showing that there was a significant difference between means for breast cancer cases between both cluster 2 and 3 at a p value below 0.05. Cluster 3 showed the lowest incidence of breast cancer.

Comparisons significant at the 0.05 level are indicated by ***.				
CLUSTER Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
1 - 2	55.184	11.636	98.731	***
1 - 3	80.857	37.119	124.595	***
2 - 1	-55.184	-98.731	-11.636	***
2 - 3	25.674	17.166	34.181	***
3 - 1	-80.857	-124.595	-37.119	***
3 - 2	-25.674	-34.181	-17.166	***

Code used:

```
libname mydata “/courses/d1406ae5ba27fe300” access=readonly;

DATA clust;
set mydata.gapminder;
* create a unique identifier to merge cluster assignment variable with
the main data set;
idnum=_n_;
*/variables to keep from for cluster analysis*/
keep idnum Internetuserate incomeperperson co2emissions femaleemployrate alcconsumption lifeexpectancy
employrate urbanrate breastcancerper100th;
/*remove rows with missing data*/
IF MISSING(breastcancerper100th) THEN DELETE;
IF MISSING(Internetuserate) THEN DELETE;
IF MISSING(incomeperperson) THEN DELETE;
IF MISSING(co2emissions) THEN DELETE;
IF MISSING(femaleemployrate) THEN DELETE;
IF MISSING(alcconsumption) THEN DELETE;
IF MISSING(lifeexpectancy) THEN DELETE;
IF MISSING(employrate) THEN DELETE;
IF MISSING(urbanrate) THEN DELETE;
ods graphics on;
/* Split data to test and training data*/
proc surveyselect data=clust out=traintest seed = 123
samprate=0.7 method=srs outall;
run;
data clus_train;
set traintest;
if selected=1;
run;

data clus_test;
set traintest;
if selected=0;
run;
/*standardize*/
proc standard data=clus_train out=clustvar mean=0 std=1;
var Internetuserate incomeperperson co2emissions femaleemployrate alcconsumption lifeexpectancy
```

```
employrate urbanrate;
run;
%macro kmean(K);
proc fastclus data=clustvar out=outdata&K. outstat=cluststat&K. maxclusters= &K. maxiter=300;
var Internetuserate incomeperperson co2emissions femaleemployrate alcconsumption lifeexpectancy
employrate urbanrate;
run;
%mend;
%kmean(1);
%kmean(2);
%kmean(3);
%kmean(4);
%kmean(5);
%kmean(6);
%kmean(7);
%kmean(8);
%kmean(9);
/*extract r-square values */
data clus1;
set cluststat1;
nclust=1;
if _type_='RSQ';
keep nclust over_all;
run;
data clus2;
set cluststat2;
nclust=2;
if _type_='RSQ';
keep nclust over_all;
run;
data clus3;
set cluststat3;
nclust=3;
if _type_='RSQ';
keep nclust over_all;
run;
data clus4;
set cluststat4;
nclust=4;
if _type_='RSQ';
keep nclust over_all;
run;
data clus5;
set cluststat5;
nclust=5;
if _type_='RSQ';
keep nclust over_all;
run;
data clus6;
set cluststat6;
nclust=6;
if _type_='RSQ';
keep nclust over_all;
run;
data clus7;
set cluststat7;
nclust=7;
if _type_='RSQ';
keep nclust over_all;
run;
data clus8;
set cluststat8;
nclust=8;
if _type_='RSQ';
```

```
keep nclust over_all;
run;
data clus9;
set cluststat9;
nclust=9;
if _type_='RSQ';
keep nclust over_all;
run;
data clusrsquare;
set clus1 clus2 clus3 clus4 clus5 clus6 clus7 clus8 clus9;
run;
* plot elbow curve using r-square values;
symbol1 color=blue interpol=join;
proc gplot data=clusrsquare;
plot over_all*nclust;
run;
*********************************************************************************
further examine cluster solution for the number of clusters suggested by the elbow curve
*********************************************************************************
/*plot clusters for 3 cluster solution*/
proc candisc data=outdata3 out=clustcan;
class cluster;
var Internetuserate incomeperperson co2emissions femaleemployrate alcconsumption lifeexpectancy
employrate urbanrate;
run;
proc sgplot data=clustcan;
scatter y=can2 x=can1 / group=cluster;
run;
/* merge clustering variable and variables with breastcancerper100th var*/
data bc_data;
set clus_train;
keep idnum breastcancerper100th;
run;
proc sort data=outdata3;
by idnum;
run;
proc sort data=bc_data;
by idnum;
run;
data merged;
merge outdata5 bc_data;
by idnum;
run;
proc sort data=merged;
by cluster;
run;
proc means data=merged;
var breastcancerper100th;
by cluster;
run;
proc anova data=merged;
class cluster;
model breastcancerper100th = cluster;
means cluster/tukey;
run;
```

February 25, 2016 (6:24 pm)