Kilger

MKT6971 2 credit course

Practicum I second semester

**Note: There is more sample code on blackboard under exercise #6**

Exercise #6

Name: **Krischelle Joyner**

This sixth exercise is to give you practice at interpreting your descriptor variable cluster means across the clusters. These descriptor variables maybe binary in nature – e.g. drink coke =yes or no. Remember that you are looking for as much separation in means across clusters for each descriptor variable as possible. Also remember that no cluster solution is perfect and that some variables will have means across clusters that are close to each other. As long as there a number of variables that have decent separation across clusters (say .3 or so different) then that’s going to be a decent solution.

Use your k means code as a starting spot from exercise #4. Find the k=# cluster solution that you thought worked best for exercise #4. Then use the following code to output the cluster number for each case, **substituting your driver variables for the ones in the sample code**. Use the maxcluster=# of clusters you chose as best in exercise #4. Example if the best solution was 4 clusters then:

*All your previous exercise 4 code here then…*

proc fastclus data=clusready out=myclust maxclusters=4;

var

healthy

ecofriend

import\_attract\_opp\_sex\_scale

spend\_time\_family\_scale;

run;

Note the out=myclust which creates a temporary SAS data set called myclust. In that data set is all of your original data plus special variable called **CLUSTER**. That variable contains the cluster number (in this case a number from 1 to 4) that indicates which cluster the case or person belongs to.

Now you want to get the means for your descriptor variables by cluster. To do that first we need to sort the data by cluster number so that we can use the BY statement in PROC MEANS. Do this by placing code like this below after the fastclus code above. This will sort your data set by cluster number and output a new temporary data set mysort.

Proc sort data= myclust out=mysort;

By cluster;

Run;

And then you can produce means for your descriptor variables like follows:

Proc means data=mysort;

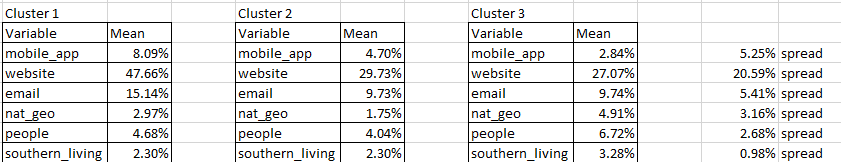
By cluster;

Var classic\_coke kfc\_chicken espn\_sports;

Run;

The BY statement tells SAS to group the means by cluster. Note that the means for binary variables such as classic\_coke can simply be interpreted as a proportion**.**

**Once you have obtained the descriptor variable means by cluster then comment on well or not so well the clustering solution discriminates on that descriptor variable. Are the descriptor variable means close to together? Far apart from each other? Remember that farther apart is better. Tell me what you see.**



**Looking at the means for each variable for each cluster this is what I see:**

The website variable has the largest mean spread between clusters at 20.59%.

Then email has the next largest mean spread at 5.41%.

Then mobile\_app has the next largest mean spread at 5.25%.

Then nat\_geo has the next largest mean spread at 3.16%.

Then people has the next largest mean spread at 2.68%.

Then southern\_living has the next largest mean spread at 0.98%.

The mobile\_app variable has the highest percentage 8.09% of people ordering a product through a mobile app in cluster 1. Then the percentage drops to 4.7% in cluster 2, and then drops to 2.84% in cluster 3.

The website variable has the highest percentage 47.66% of people ordering a product through a website in cluster 1. Then the percentage drops to 29.73% in cluster 2, and then drops to 27.07% in cluster 3.

The email variable has the highest percentage 15.14% of people ordering a product through an email in cluster 1. Then the percentage drops to 9.73% in cluster 2, and then rises to 9.74% in cluster 3.

The nat\_geo variable has a low percentage of 2.97% of people reading the National Geographic magazine in the last 6 months for cluster 1. Then the percentage drops to 1.75% in cluster 2, and then the highest percentage is 4.91% in cluster 3.

The people variable has a low percentage of 4.68% of people reading the People magazine in the last 6 months for cluster 1. Then the percentage drops to 4.04% in cluster 2, and then the highest percentage is 6.72% in cluster 3.

The southern\_living variable has a low percentage of 2.30% of people reading the southern\_living magazine in the last 6 months for cluster 1. Then the percentage stays the same at 2.30% in cluster 2, and then rises to 3.28% in cluster 3.

Based off these observations it seems that the following variables are the best in each cluster:

Cluster 1

Mobile\_app

Website

Email

Cluster 3

Nat\_geo

People

Southern\_living

Audible could advertise in magazines using cluster 3, but this would reduce the percentage of the audience audible would reach for each technology type (mobile\_app, email, website). I think audible would want at least 3 clusters from my analysis so I would present this analysis to them.