DA6823

Exercise #6

Name: Moneeb Abu-Esba KVG805

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. You will have specified them in market segmentation exercise #1 as descriptor variables. Remember that you are looking for as much separation in means across clusters for each descriptor variable as possible – use the criteria discussed in class to evaluate the solution for the descriptor variables. . Also remember that no cluster solution is perfect and that some variables will have means across clusters that are close to each other.

Use your k means code as a starting spot from exercise #4 k means clustering. Find the k=# cluster solution that you thought worked best in that exercise. 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**.**

1. **One 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.** most of them do...
   The clusters fluctuate between close and far, but they all have a 0.1 difference in mean.
2. **Finally, write a short one paragraph description of each cluster using the means from the driver and descriptor variables.**
   The drivers are very far apart with most if not all having at least a 1.0 difference. It shows a lot of discrimination. It is not so the same for the descriptor variables as they have a difference around 0.1. I believe this is due to the descriptor variables being vary broad as some of them are gender and if they are legal guardians. With the descriptor variable the clusters stopped at K=3 as it is very small not seen as good and helpful.

what does this last sentence mean? Does not make any sense... -3

where are your descriptions of each cluster? this is the most important part of the assignment! -7