DA6823

Exercise #6

Name: Austin Thrash - FFK221

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.**

Based on my work in Exercise #4, cluster 5 was the best choice. However, when doing the previous exercise, I had discovered that my cluster may not have been very good. After completing this exercise, I believe that is the case. Across each of the 5 clusters, each descriptor variable mean exhibits a difference between 0.005 and 0.015. The only descriptor variable with a noticeable difference in means across clusters is “more likely to use a brand that character uses” with 1.5 difference on a 5 point scale.

2. **Finally, write a short one paragraph description of each cluster using the means from the driver and descriptor variables.**

2.1. Respondents in the first cluster read less Outdoor life and Field and Stream magazines when compared to the other clusters. However more respondents in this cluster watch more Fox News and Animal planet when compared to the average of the other clusters. This cluster also contains the second most amount of female respondents. All driver variables for this first cluster have the smallest difference between corresponding driver variables in other clusters.

2.2 Cluster 2 has similar demographics when it comes to Outdoor Life and Field and Stream magazine, as well as almost exactly matching the demographics for Fox News and Animal Planet in Cluster 5. With this cluster however, there seems to be more of a balance between male and female respondents. Another difference with this variable is that is has the largest difference between corresponding driver variable means in other clusters.

2.3 With Cluster 3, we see an increase in respondents who have read Outdoor Life and Field and Stream magazines. This cluster also contains the largest amount of respondents who watch Fox News when compared to the other clusters. A noticeable difference can also be seen in the number of hispanic respondents in this cluster. This cluster contains a 0.20 difference when compared to other clusters as they are all fairly balanced between non-hispanic and hispanics.

2.4 In Cluster 4 we see the largest amount of respondents who read Outdoor Life and Field and Stream magazines. However we see that we have close to the average mean of respondents who watch Fox News and Animal Planet; not too big of a difference can be seen when comparing this variables means to other clusters. This cluster contains the largest difference between male and female respondents with a difference of about 0.1, thus making it the cluster with the most amount of female respondents.

2.5 The fifth cluster is mostly similar to all the previous clusters, however contains more respondents who read Field and Stream magazine. However when comparing this clusters driver variable means to other clusters we notice that there are no close comparisons except for cluster 2, but the similarities are not close enough to label them as significant.

Although these differences can be summarized through careful examination of the results, they still are not strong enough to indicate that good clustering has been achieved. I am going to continue to swap out/remove variables to see what will lead to better clustering results and larger differences between means.