DA6823

Exercise #5

Name: Austin Thrash - FFK221

This fifth exercise is to give you practice at clustering market segments using GAP analysis

You can likely reuse a bunch of your code from exercise 4. So instead of doing a k means clustering, you are going to use PROC HPCLUS to do a gap analysis clustering. **Be sure to cut and paste the tables here that help you answer the questions below.**

1. Using the same drivers you did for your k means assignment, run your PROC HPCLUS using FIRSTPEAK as your criterion. How many clusters does it say is optimum? Cut and paste the ABC Statistics table as well as the FIRSTPEAK table below.

ABC Statistics

Number of Clusters	Logarithm of Within-Cluster SSE		Gap	Simulation Adjusted Standard Deviation	One Standard Error Adjusted Gap
	Input	Reference			
2	11.3760	12.4730	1.0970	0.0161	1.0809
3	11.1821	12.0231	0.8410	0.0125	0.8284
4	11.0141	11.7922	0.7781	0.0188	0.7594
5	10.8850	11.7379	0.8529	0.0109	0.8420
6	10.8020	11.4117	0.6096	0.0301	0.5795

Estimated Number of Clusters

Criterion	Number of Clusters
FIRSTPEAK	5

Based on the output of HPCLUS, the number of cluster that is optimum is 5. This can be seen in the 'Estimated number of clusters' table.

2. Repeat step #2 except use GLOBALPEAK as your criterion. How many clusters does it say is optimum? Is it the same as Step #1 above? Cut and paste the ABC Statistics table as well as the GLOBALPEAK table below

Estimated Number of Clusters

Criterion	Number of Clusters
GLOBALPEAK	5

We can see that using global peak also suggests that 5 clusters is the optimum.

3. What number of clusters did you pick on the previous K means exercise? Is it the same as the HPCLUS suggested number of clusters? How might you decide which to use?

In the previous K means exercise, the test results also suggested that 5 clusters was the optimum. Both test suggest the same number of clusters, it seems like 5 is the best.

4. Examine the cluster means for the drivers for the result either in Step #1 or Step #2 above. Do they look like there is decent discrimination among the clusters for the driver variables?

Within Cluster Statistics

Variable	Cluster	Mean	Standard Deviation
mag_entertainment	1	1.3098	6.2824
	2	3.5701	7.3683
	3	1.5667	5.3944
	4	2.4889	4.2333
	5	1.3112	1.8596
understand_nat	1	3.9243	13.7542
	2	4.3153	10.0532
	3	4.1543	9.0599
	4	2.7254	6.1196
	5	4.4159	4.8261
eco_friend	1	2.4855	12.4787
	2	4.1463	9.2179
	3	3.6085	8.3519
	4	2.8908	5.8147
	5	4.4391	4.9072
sport_exe	1	4.3979	14.2349
	2	4.1799	8.9283
	3	1.5114	5.6652
	4	2.8082	5.3838
	5	4.4311	5.2895

Upon examining the cluster means for my driver variables, I believe there is decent discrimination among the clusters for those variables. When comparing any of the driver variables means across the 5 clusters, we can see that the smallest difference is usually no smaller than 0.1, while our greatest difference can be between 1.5 - 2. This variation across clusters leads me to believe that the distinction between each cluster is significant.

5. Turn in your code, output and this report.