DA6823

Exercise #4

Name: Austin Thrash - FFK221

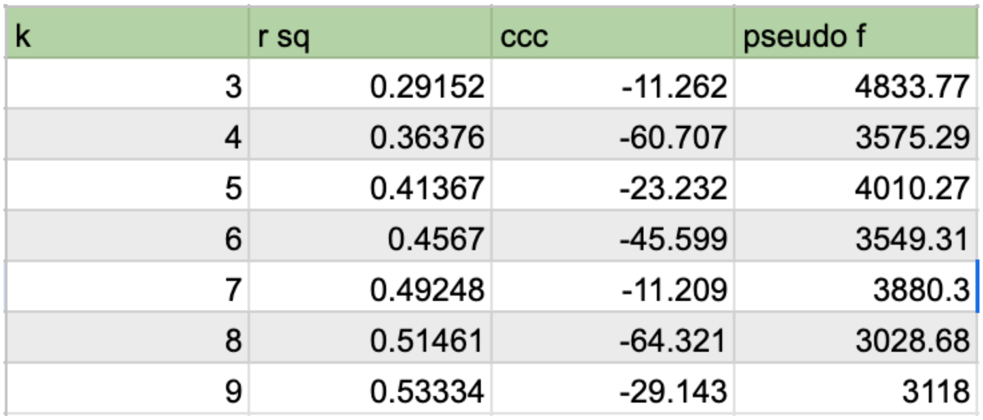
This forth exercise is to give you practice at clustering market segments using k means clustering. **Be sure that you cut and paste the tables and plots that are requested in the questions below!!!**

You can reuse the code from exercise 3. You will need to add some extra code in the PROC FACTOR procedure to create a temporary SAS file that creates a temporary SAS system file that contains all of your original variables plus the two PCA factor variables that were created in PROC FACTOR. **Watch the lecture carefully to see how to do this.**

1. Run your k means cluster analysis using PROC CLUSTER using the temporary file just created above. Use all of your single driver variables plus the two factor variables you created in the PROC FACTOR procedure. Run it each time for k = 3 to k= 9 (that is 6 runs). Be sure to include these runs in your submission.

   2. Build a table that contains the results for the k=3 to k=9 runs with values for the following statistics:

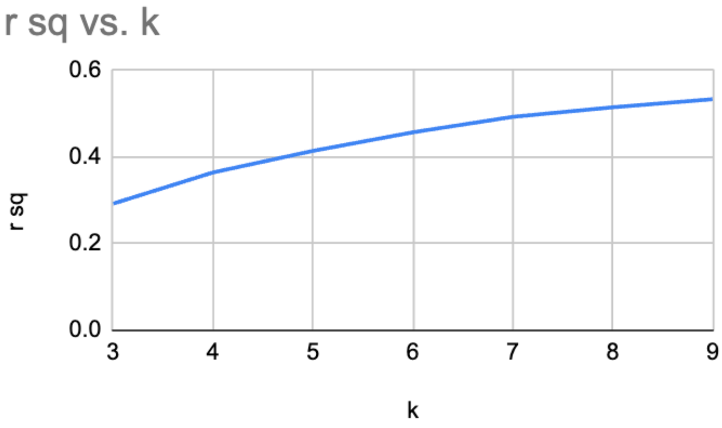
      a. Number of clusters
      b. R square
      c. CCC
      d. Pseudo F



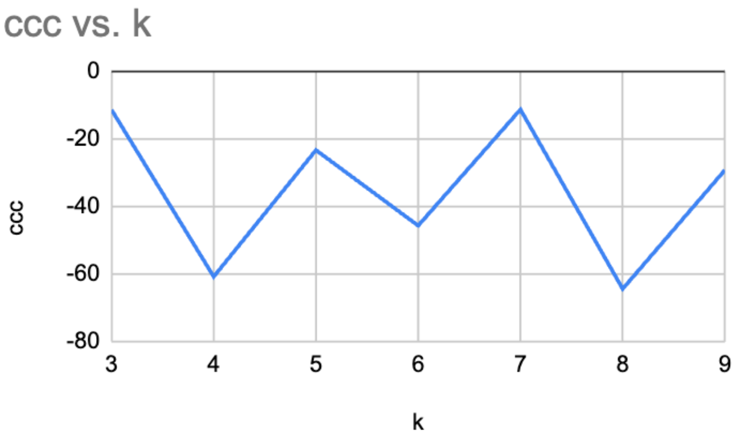
k	r sq	ccc	pseudo f
3	0.29152	-11.262	4833.77
4	0.36376	-60.707	3575.29
5	0.41367	-23.232	4010.27
6	0.4567	-45.599	3549.31
7	0.49248	-11.209	3880.3
8	0.51461	-64.321	3028.68
9	0.53334	-29.143	3118

Well, bad news bears, all my CCC values turned out the be negative. This could be because my clusters are not well-defined or that distinct from each other. This might stem from creating factors that have not been the abstractly described with the proper variables. My two factors were Brand Loyalty and Cost Consciousness. I will have to go back and try different variables again in order to get better clustering results.

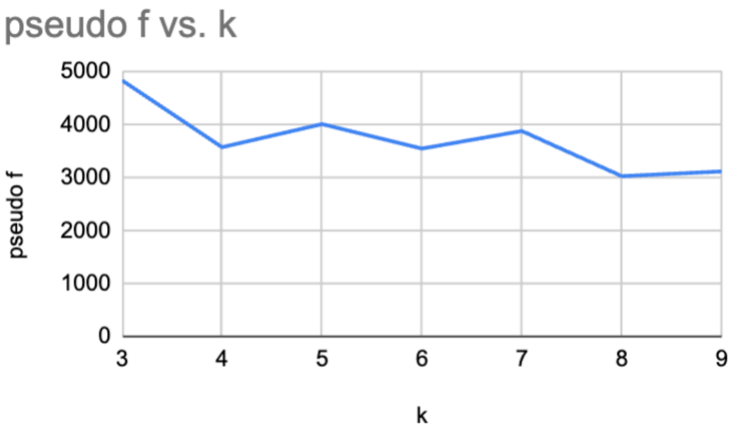
3. Use whatever plotting software you like to create three different plots that the following against the number of clusters
    a. R square



    b. CCC



    c. Pseudo F



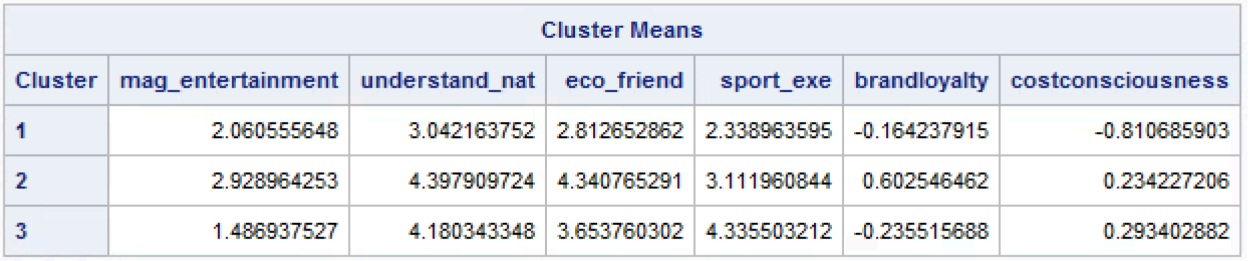
4. Examine the CCC plot and tell me what the rule is to suggest how many clusters to keep. How many does it suggest for this k means analysis. Note that it may not suggest any specific number of clusters. Tell me what you see!

Based on our CCC plot, our local maximum is 3 clusters, however this plot may not be a good indicator of how many clusters to keep due to all the CCC values being negative, meaning our clustering is not good to begin with.

5. Examine the Pseudo F plot and tell me what the rule is to suggest how many clusters to keep. How many does it suggest for this k means analysis. Note that it may not suggest any specific number of clusters. Tell me what you see!

Based on our Pseudo F plot, we will chose the 3 clusters as our Pseudo F value only decreases onward. Like stated, this is likely to be sub-optimal.

6. Examine the means of the driver variables for the solution that is suggested by the CCC plot. Do they look like this might be a good solution? Why or why not? Cut and paste the driver means table for the best solution. Note the criteria that I discuss in class for deciding if this is a decent solution or not.

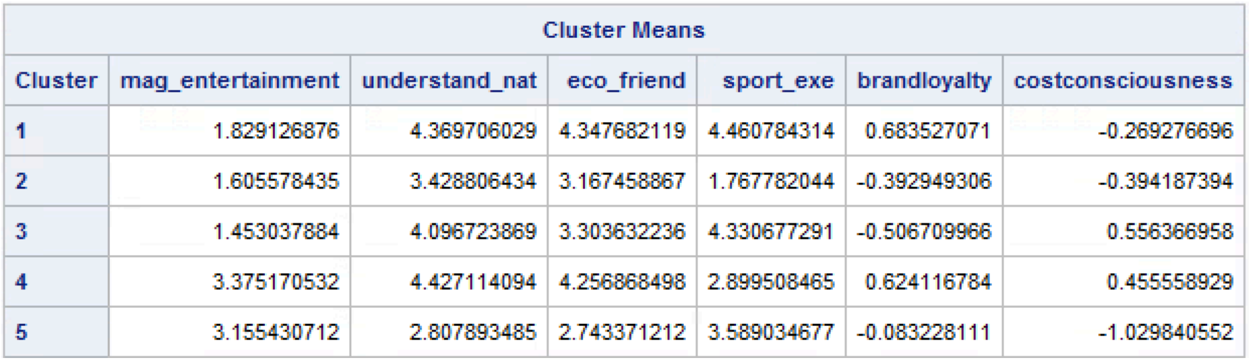


Cluster Means

Cluster	mag_entertainment	understand_nat	eco_friend	sport_exe	brandloyalty	costconsciousness
1	2.060555648	3.042163752	2.812652862	2.338963595	-0.164237915	-0.810685903
2	2.928964253	4.397909724	4.340765291	3.111960844	0.602546462	0.234227206
3	1.486937527	4.180343348	3.653760302	4.335503212	-0.235515688	0.293402882

Based on my CCC plot, it suggested using 3 clusters. However after calculating the number of ties, we can see that about 2 out of 3 comparisons turn up to be ties. This comes out to about 66%, which is not a good percentage as we are expecting to only have between 10% to 20% turn up to be ties.

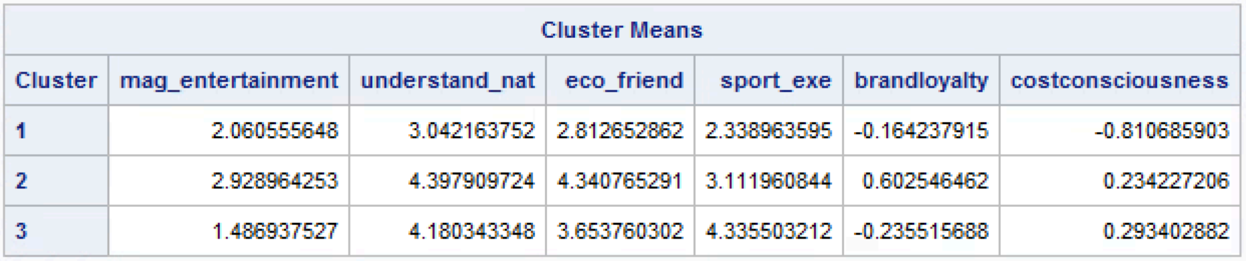
This value significantly changes when using the second local maximum CCC which is -23.232, which is 5 clusters. When using 5 clusters, we can observe that about 1 out of 5 comparisons, 20%, turn up to be ties. This percentage is better but not great. The table for 5 clusters can be seen below.



Cluster Means

Cluster	mag_entertainment	understand_nat	eco_friend	sport_exe	brandloyalty	costconsciousness
1	1.829126876	4.369706029	4.347682119	4.460784314	0.683527071	-0.269276696
2	1.605578435	3.428806434	3.167458867	1.767782044	-0.392949306	-0.394187394
3	1.453037884	4.096723869	3.303632236	4.330677291	-0.506709966	0.556366958
4	3.375170532	4.427114094	4.256868498	2.899508465	0.624116784	0.455558929
5	3.155430712	2.807893485	2.743371212	3.589034677	-0.083228111	-1.029840552

7. Examine the means of the driver variables for the solution that is suggested by the Pseudo F plot. Do they look like this might be a good solution? Why or why not? Cut and paste the driver means for the best solution. Note the criteria that I discuss in class for deciding if this is a decent solution or not.



Cluster Means

Cluster	mag_entertainment	understand_nat	eco_friend	sport_exe	brandloyalty	costconsciousness
1	2.060555648	3.042163752	2.812652862	2.338963595	-0.164237915	-0.810685903
2	2.928964253	4.397909724	4.340765291	3.111960844	0.602546462	0.234227206
3	1.486937527	4.180343348	3.653760302	4.335503212	-0.235515688	0.293402882

Similarly to the CCC plot, the Pseudo F plot suggest we use cluster 3 as it is the first local maximum. The number of ties is the same as the previous answer as it is the same cluster. Examining the Pseudo F plot, we can see that we have good values for cluster 3 and 5. Sense cluster 5 has better tie percentages and retains a good pseudo f value, it is likely a better choice.

8. **If your solution above satisfies the criteria for a good solution, you are set. If neither the CCC or the Pseudo F plot suggest a best solution, try removing some of the drivers or adding one or two new drivers until you get something that satisfies the criteria for a decent solution. When you have that solution, then cut and paste this new table with number of clusters, R square, CCC and pseudo F below. Also cut and paste the driver means for this new solution below.**

Based on our CCC plot and Pseudo F plot, cluster 7 is another suggest solution, so we will try that before adding or removing any drivers to see if we can improve our CCC values. After examining cluster 7, it is full of ties, this leads me to believe that cluster 5 is the best to use with my current setup, however this is not an optimal solution so I am going to continue to try new variables and see if I can obtain better results.