Eric Gagliano (egagli)
Inferring Structure from Complex Systems

Lab 8

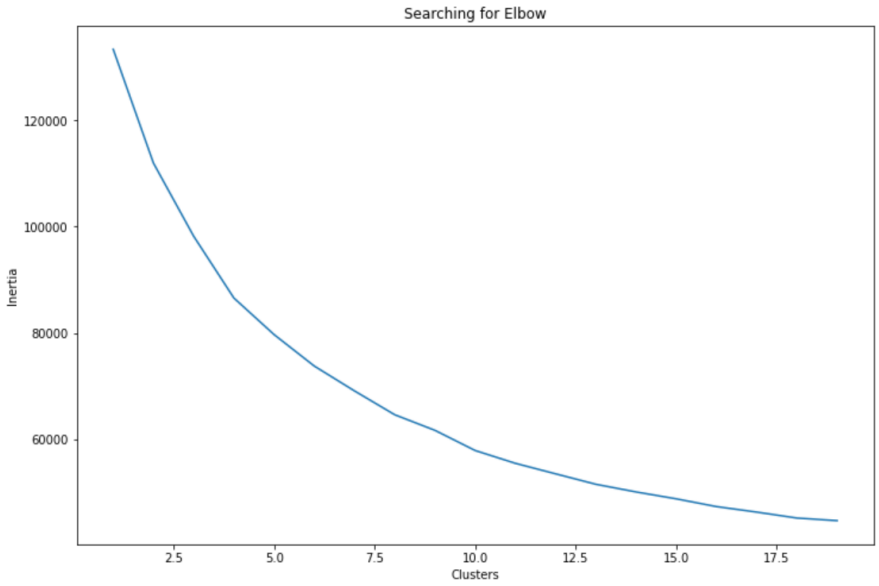
Github link:
https://github.com/egagli/amath563/blob/main/labs/8/Lab_8_AMATH_Decomposition.ipynb

1. Which combination of variables do you expect to give the best separation? Why?

I would think purchases and purchases frequency because these should be highly correlated. The more things you buy, the more money spent on purchases. Of course different items will cost different amounts of money, but this general rule seems to hold.

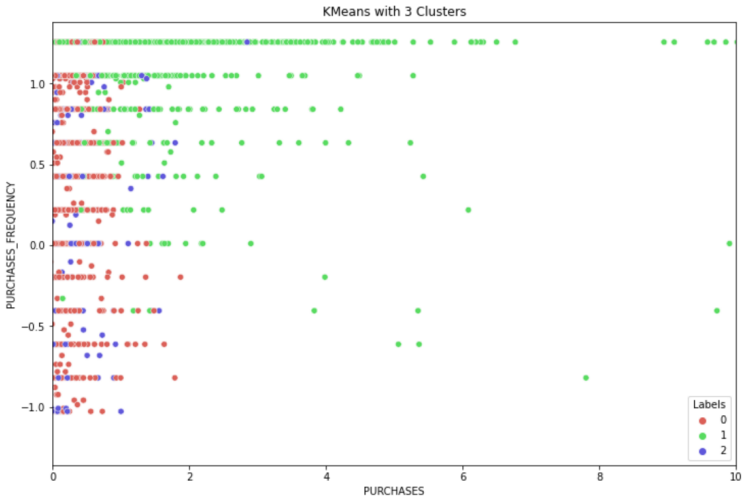
2. What value of k did you choose for your k-means clustering?

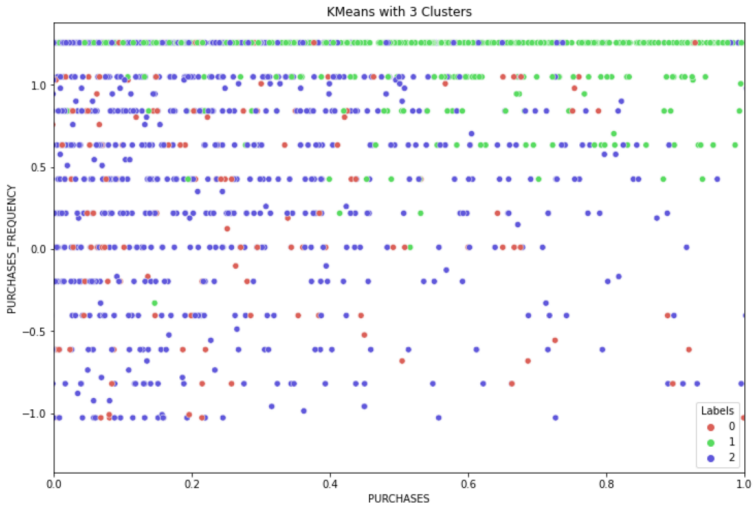
I chose a value of k=3 based off of the inertia vs k plot shown here.



3. Give a brief description of the types of customers grouped into each cluster.

It looks like in the 3 clusters we have green which are high spenders and high frequency spenders. Red is medium and highly variable frequency spenders. Blue is low spenders, though they still have medium purchasing frequency.





4. When you took the PCA and SVD, what value of r did you choose for the dimension of your projection?

I chose a value of r=5 for my projection. Increasing r would increase the performance of these methods relative to clustering.

5. According to your confusion matrices, which clusters had the most and least agreement between the original coordinates and the SVD and PCA spaces?

It looks like the PCA has almost perfect agreement with the predictions using k-means clustering, only getting 7 out of 864 wrong, translating to an accuracy of 99.2%. It looks like cluster 0 was most poorly falsely predicted, and cluster 2 was most accurately predicted. SVD was weaker, misclassifying 86, translating to an accuracy of 90.0%. Similar to PCA, cluster 0 was most poorly falsely predicted, and cluster 2 was most accurately predicted. Keep in mind, label 1 has a much larger number of members, so in both PCA and SVD, when adjusted for relative number of members, the misclassifications are pretty uniform. Using a higher dimension r would lead to much closer agreement for both methods when compared to k-means clustering.

6. Please upload images of your two confusion matrices.