Willy Esquivel-lopez – G01127937

Miner2 username: Gopher123 (Best Public Score: Phase 1 = .79, Phase 2 = .70)

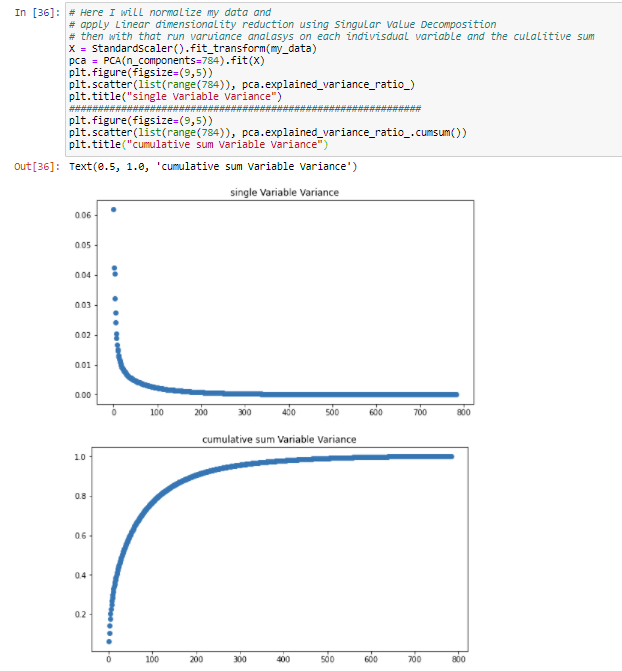
**HW4 Report**

**My approach:**

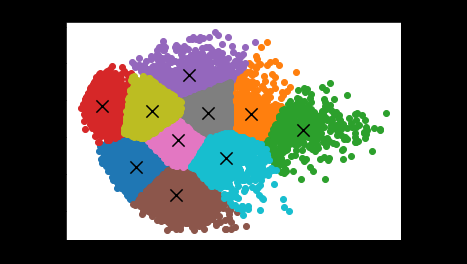
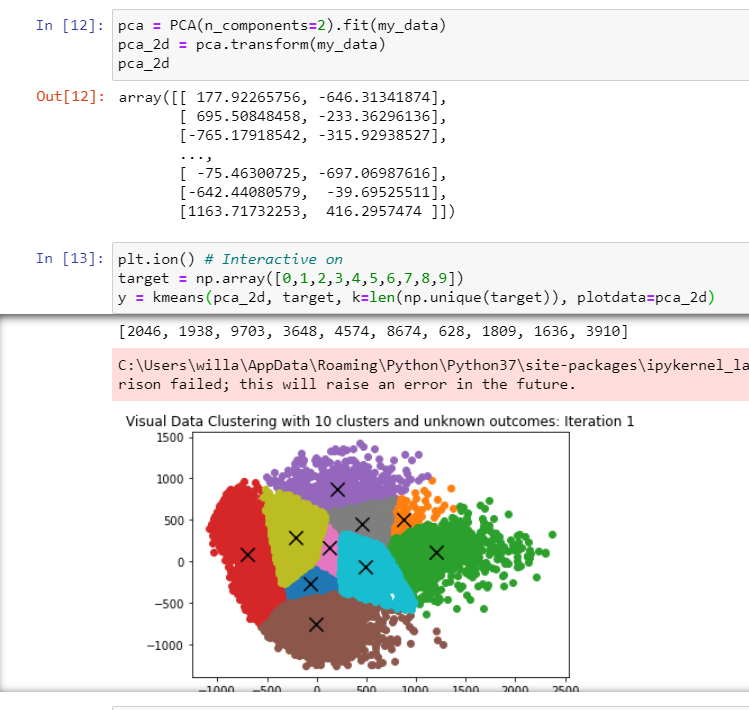
When seeing that the small dataset for Iris and the Visual Data set for Phase 2 was unlabeled and that the target value was not given, I could no longer use the traditional train test split among x and y axis. upon closer examination of the data, I could still not find any correlation at first glance for either of them. I plotted the visual data using a 28x28 grid to visually understand the data and any trends.



I tried plotting it to make sense and began estimating some numbers such as mean, std div, and generating some random centroids given those values. But given all that information I could not generate any useful correlation or knowledge. Thus, I looked for other means of plotting the data, I began trying to solve for the variance and discovered a library “sklearn.decomposition.PCA” which I could use to reduce the dimensionality of my data by using Singular Value Decomposition. The Lower Dimensional space, which would now allow me to analyze variance on single or the cumulative sum of my variables.

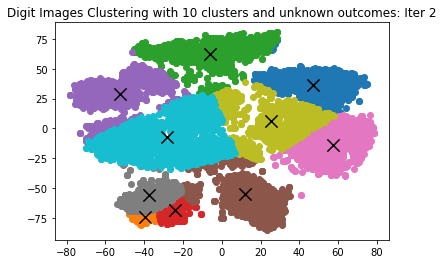
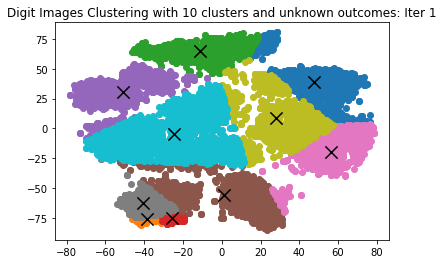
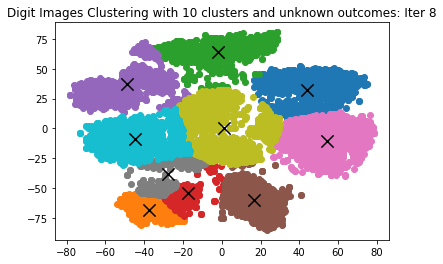


This was especially useful as it gave me a better insight as to what the variance was for each component and could determine the estimation of number of clusters in my data and each component variation. Now with this given I began to Implement My K-Means algorithm in a fashion where I would be able to see the plotting of each iteration and the K-Means progression in clustering the data. Thus, I applied PCA and ran my K-Means algorithm.

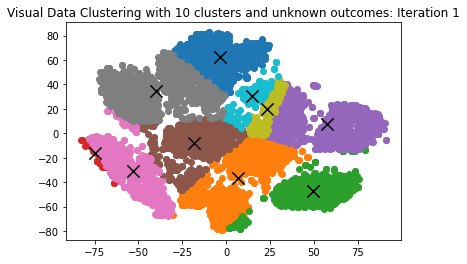


Unfortionately the best scores I could get with this sapproach was around mid 60%. So I had to think of another way to cluster my datapoints. I began to do reaserch on other means of representing my clustering of my data.

Best = [7641, 5051, 5572, 5908, 5520, 3758, 9406, 2209, 1548, 8090]

**…**8 iterations with .70 V-measure on miner

Best = [875, 1530, 667, 5213, 1167, 2649, 1058, 4626, 774, 5728]

Iter =8 with .72 V-measure

