Clustering

Instructions to run the code

- **Filename = elkans_analysis_birch.py (Run this code, choose k to replicate the Elkan's experiment on the birch dataset from the Elkan approach paper)**

- **Filename = K-Means_Project.py**
  - **I have run this code in local Jupyter Notebook**

  - After some research, I found out that unique () function from Numpy is not available in all versions of python. I have used this line "np.split(df1[:, 3], np.cumsum(np.unique(df1[:, 0], return_counts=True)[1])[:-1])" in my code to a process that calculates the rand index in order to determine the cluster quality.

  - Therefore, please run this on **python 2.7** and with **numpy verison 1.7.1** (I have verified and got the desired outputs in jupyter notebook as well)

  - Please change the pathway directory of the files used in the analysis (I have added comments on the code on where this can be done)

  - Rest of the code I have designed it to be user interactive, hence options are displayed with comments on the screen.

  - In the code, when you press 7 in the options given it will exit (it would prompt you to enter some k value so don't mind that bug.)

  - The accuracy on student's dataset would be a bit less because I reduced the dimensionality of it for this analysis to avoid running into higher runtime issues. (need to do PCA or some other measure to choose more relevant features maybe)