CSE 255 Homework 2
Name: Yansong Gao
PID: A53100210

Tasks(PCA & Clustering):

1. The "reconstruction error" is 113183.433705

2. Using 3 PCA dimensions, use the middle value as the last two dimension. The "reconstruction error" is 11714.415511

3. The last two clusters' centroids are
(4.07021276596, 3.96063829787, 3.87021276596, 3.96808510638, 3.89574468085)
and
(2.66666666667, 2.65, 2.88333333333, 2.96666666667, 2.33333333333)
respectively.
And cluster one has 470 nodes, cluster two has 30 nodes.

4. If we compress the data by replacing each point by one of the two centroids of the two clusters found in question 3. The reconstruction error is 557.917198582

Tasks(Community Detection):

1. There are 3 connected components in the graph. And there are 40 nodes in the largest connected component.
2. The normalized cut cost is 0.422405876951
3. Starting from the 50% 50% two community from question2, using greedy algorithm, we can get the normalized cut cost 0.0981704596162. The two communities are :
   1. [697, 703, 708, 713, 719, 745, 747, 753, 769, 772, 774, 798, 800, 803, 805, 810, 811, 819, 828, 823, 830, 840, 880, 890, 869, 856]
   2. [825, 861, 863, 864, 876, 878, 882, 884, 886, 888, 889, 893, 729, 804]
4. Now if we have four communities, then the cost is 0.280598888492. The four communities are:
   A. [697, 703, 713, 719, 745, 747, 772, 774, 800, 803, 805, 810, 819, 823, 828, 830, 840, 856, 880],
   B. [708, 753, 769, 798, 811, 869, 890],
   C. [825, 864, 884, 893],
   D. [729, 804, 861, 863, 876, 878, 882, 886, 888, 889]]