Data Mining Homework 10
D'yangelo Grullon
4/30/15

1)
   a) Team members:
      i) Joseph Ville - k-Means algorithm and stuff
      ii) Dyangelo Grullon - k-Means algorithm and stuff
      iii) Meridangela Gutierrez - data normalization, and data visualization
   b) Python
   c) Overall architecture: for k = 9 and 39, pick k seed points for the centers of the k clusters. Keep repeating until no data points change from one cluster to another. This is our stopping criteria.
      i) Interface used for development: Python Idle
      ii) Subroutines used: random.randint
      iii) We use a Cluster object

2) Aqua Flu cid: 0
   Crimean Cold cid: 1
   SSE: 238739944.25
   Normalized: False

K=9	Min ROW of any seat for this cluster	Max ROW for this cluster	Min AISLE for this cluster	Max AISLE for this cluster
Aqua Flu	1	51	1	10
Crimean Cold	8	51	1	10

   You cannot compare these clusters; the sum of squared error is huge. Therefore, these clusters are not cohesive.

3) Aqua Flu cid: 0
   Crimean Cold cid: 13
   SSE: 59129497.19
   Normalized: False

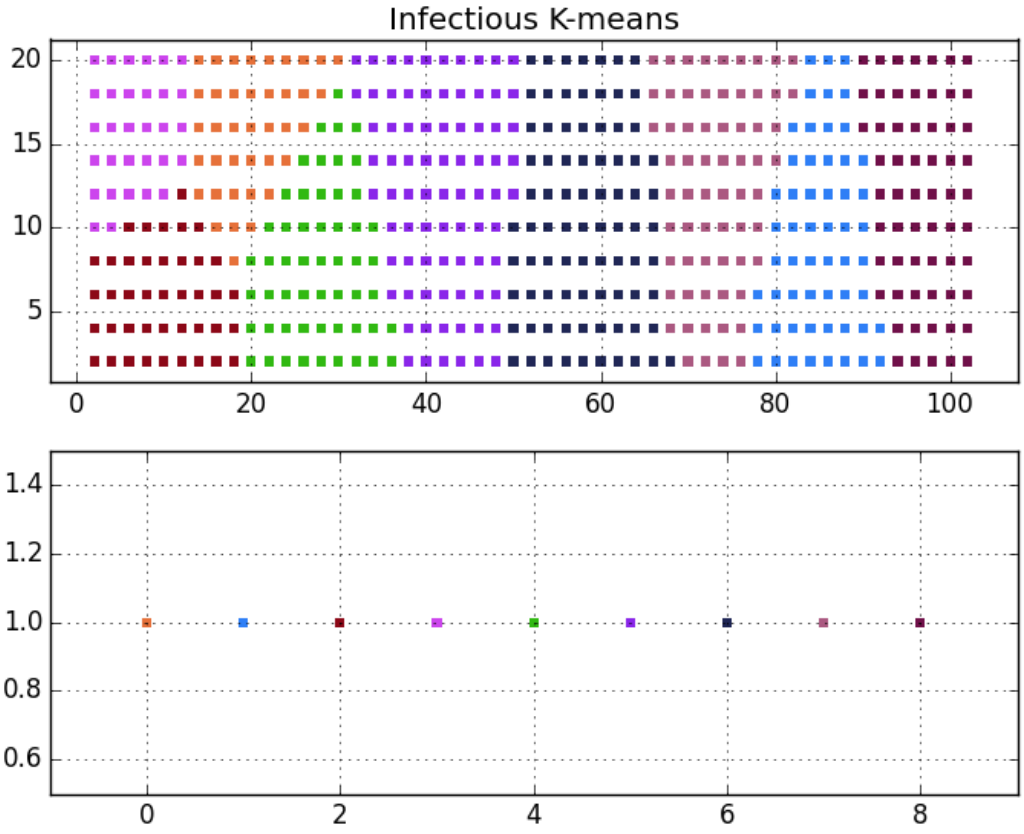
K=39	Min ROW of any seat for this cluster	Max ROW for this cluster	Min AISLE for this cluster	Max AISLE for this cluster
Aqua Flu	2	42	1	10
Crimean Cold	10	51	1	10

Better cohesiveness, than k = 9, so it would seem that increasing k would improve the clusters if they were widely separated before, but the measure of cohesiveness is still very high. In addition, the clusters are apparently/visually too spread out as well.

4) Aqua Flu cid: 2
   Crimean Cold cid: 1
   SSE: 2368.56
   Normalized: True

K=9	Min ROW of any seat for this cluster	Max ROW for this cluster	Min AISLE for this cluster	Max AISLE for this cluster
Aqua Flu	1	15	1	10
Crimean Cold	39	51	1	10

5) As the normalizing factor decreases, the cluster cohesiveness decreases as well. That is to say that the importance of rows and aisles in the distance metric begins to dominate the distance metric, the cluster cohesiveness suffers. From row 15 to 25, cluster 5 begins to hog all of the data points because proximity because more of a factor in the distance metric. Cluster 1 wobbles in size, but overall stays the same.
6)
   a) High blood pressure, high body temperature, nausea, cloudy urine, difficulty swallowing/Dry Mouth and high pulse rate.
   b) Individuals with aqua flu might have underlying symptoms that are not observable after passing through a certain stage, but individuals who have recently acquired aqua flu might exhibit certain symptoms as expressed by the center of mass of cluster 0: High Body Temp or a fever.
   c) Manhattan distance (diamond shape) because the virii probably spread due to proximity, and the closest passengers to an infected passenger are in a diamond shape relative to the host (front, back, left and right). As the virii continue to spread, the farthest edges of the cluster maintain this diamond shape relative to the group of hosts.
7) Unsupervised learning may provide plenty of insight into the hidden behavior or patterns in the data, but the attribute selection process, combined with their determining their influence to the distance metric (normalizing factor) are what really determine whether you find patterns in the data or not. Data lies, but we can chisel the truth from it.


Infectious K-means
20
15
10
5
0
20
40
60
80
100
1.4
1.2
1.0
0.8
0.6
0
2
4
6
8