Ahmed Wael Fahmy

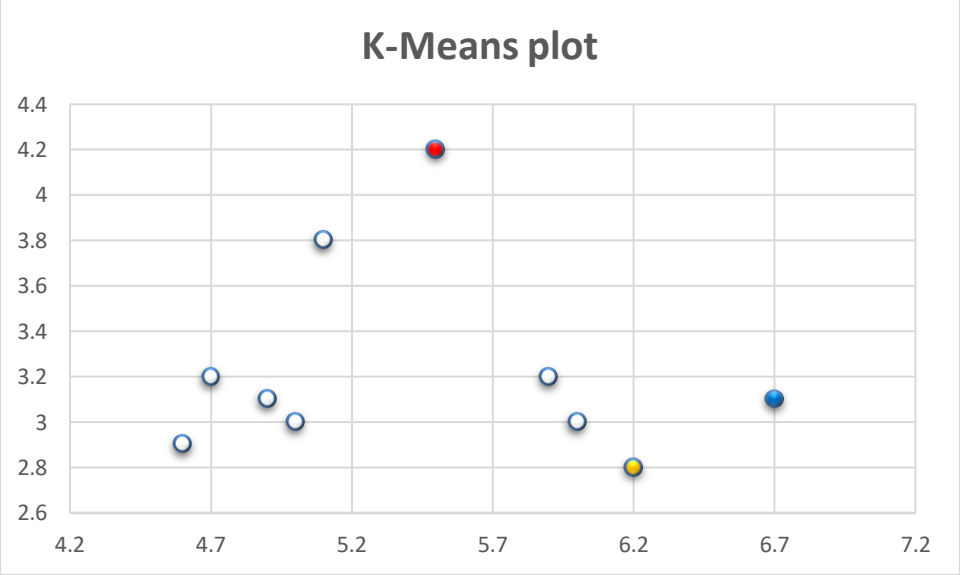
900160127

**Machine Learning**

**Assignment # 3**

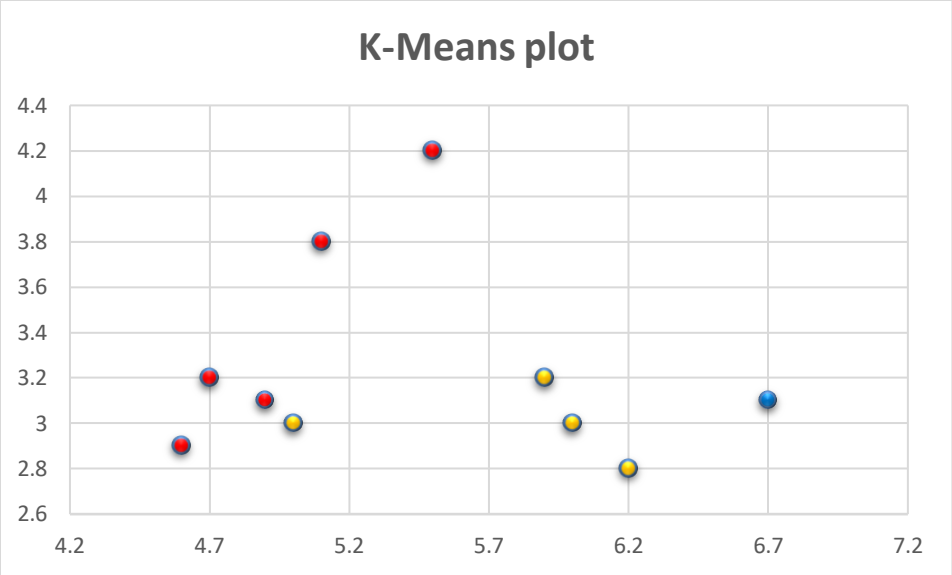
Q1)

Initial Plot



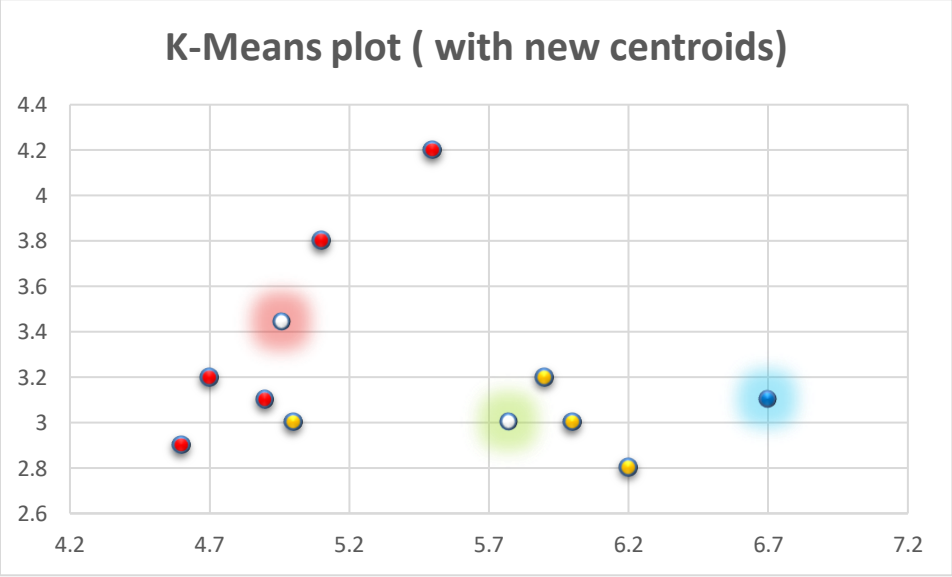

Iteration 1:

| x | y | distance from red | distance from blue | distance from yellow | minimum | centroid |
|---|---|---|---|---|---|---|
| 5.5 | 4.2 | 0 | 1.62788206 | 1.565247584 | 0 | red |
| 5.1 | 3.8 | 0.565685425 | 1.74642492 | 1.486606875 | 0.5656854 | red |
| 4.7 | 3.2 | 1.280624847 | 2.002498439 | 1.55241747 | 1.2806248 | red |
| 5.9 | 3.2 | 1.077032961 | 0.806225775 | 0.5 | 0.5 | yellow |
| 4.9 | 3.1 | 1.252996409 | 1.8 | 1.334166406 | 1.2529964 | red |
| 6.7 | 3.1 | 1.62788206 | 0 | 0.583095189 | 0 | blue |
| 5 | 3 | 1.3 | 1.702938637 | 1.216552506 | 1.2165525 | yellow |
| 6 | 3 | 1.3 | 0.707106781 | 0.282842712 | 0.2828427 | yellow |
| 4.6 | 2.9 | 1.58113883 | 2.109502311 | 1.603121954 | 1.5811388 | red |
| 6.2 | 2.8 | 1.565247584 | 0.583095189 | 0 | 0 | yellow |



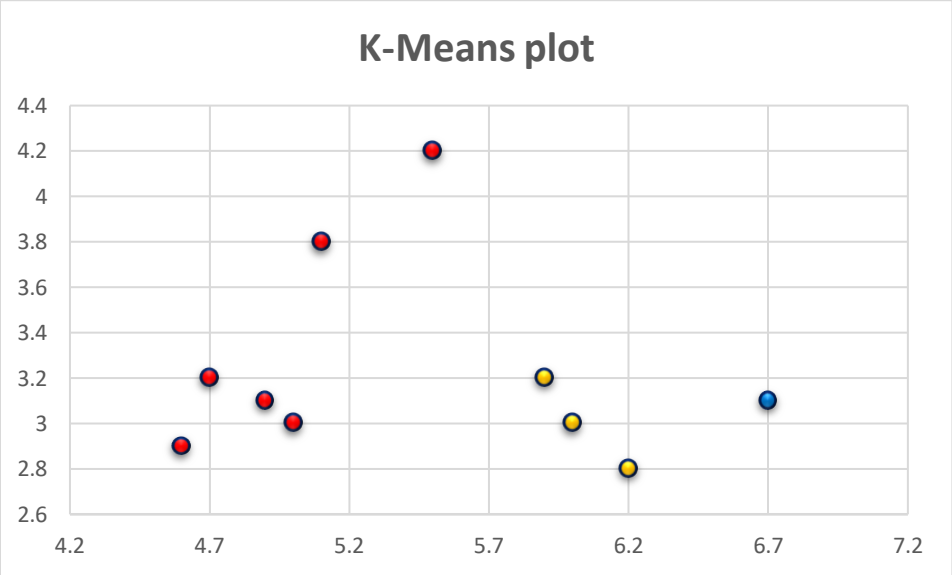

| | New centroids | |
|---|---|---|
| | x | y |
| red | 4.96 | 3.44 |
| blue | 6.7 | 3.1 |
| yellow | 5.775 | 3 |

Glowing points are centroids

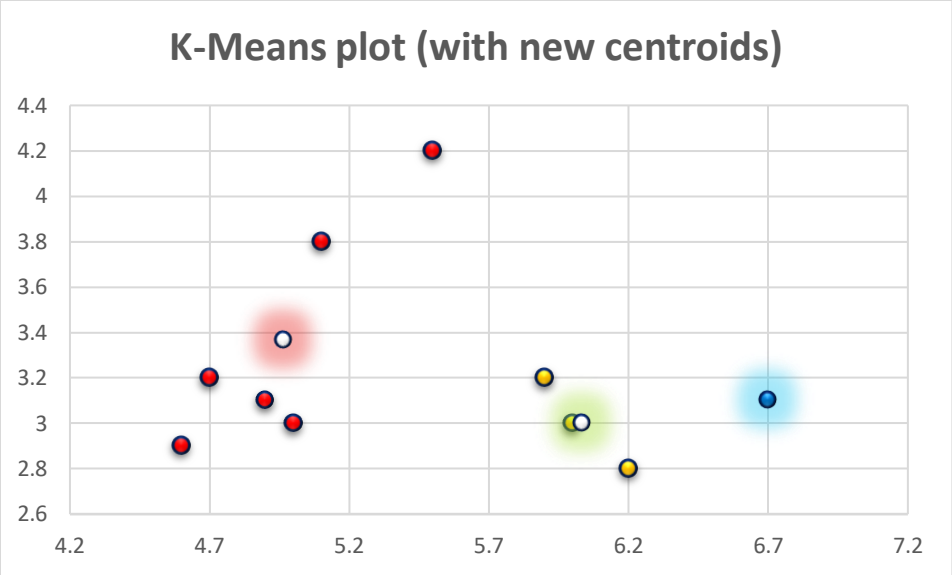


Iteration 2:

| x | y | distance from red | distance from blue | distance from yellow | minimum | centroid |
|---|---|---|---|---|---|---|
| 5.5 | 4.2 | 0.932308962 | 1.62788206 | 1.231107225 | 0.932309 | red |
| 5.1 | 3.8 | 0.386264158 | 1.74642492 | 1.046721071 | 0.3862642 | red |
| 4.7 | 3.2 | 0.35383612 | 2.002498439 | 1.093446386 | 0.3538361 | red |
| 5.9 | 3.2 | 0.970154627 | 0.806225775 | 0.235849528 | 0.2358495 | yellow |
| 4.9 | 3.1 | 0.34525353 | 1.8 | 0.880695748 | 0.3452535 | red |
| 6.7 | 3.1 | 1.772907217 | 0 | 0.930389703 | 0 | blue |
| 5 | 3 | 0.441814441 | 1.702938637 | 0.775 | 0.4418144 | red |
| 6 | 3 | 1.129247537 | 0.707106781 | 0.225 | 0.225 | yellow |
| 4.6 | 2.9 | 0.64899923 | 2.109502311 | 1.179247642 | 0.6489992 | red |
| 6.2 | 2.8 | 1.395421083 | 0.583095189 | 0.469707356 | 0.4697074 | yellow |




| | New centroids | |
|---|---|---|
| | x | y |
| red | 4.966666667 | 3.366666667 |
| blue | 6.7 | 3.1 |
| yellow | 6.033333333 | 3 |

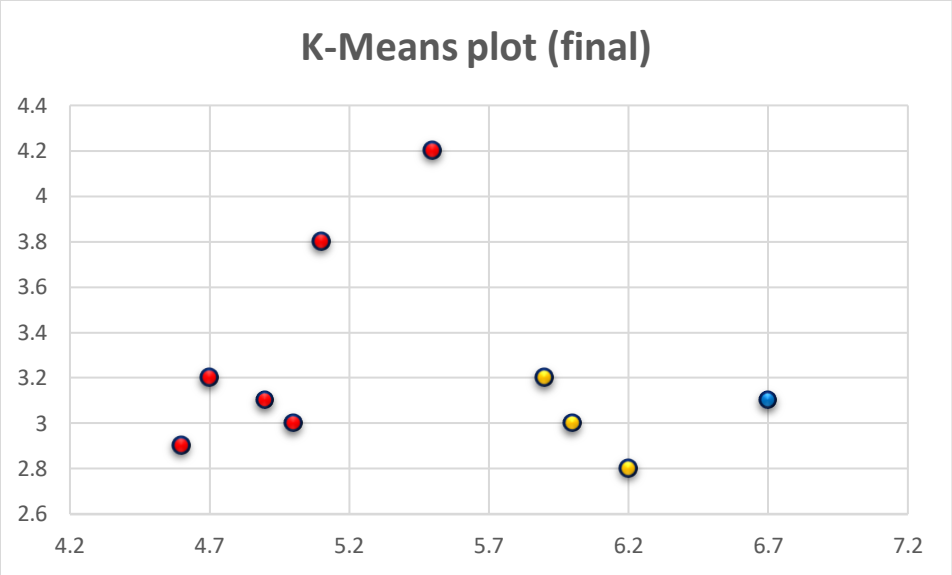



Glowing points are centroids

Iteration 3:

| x | y | distance from red | distance from blue | distance from yellow | minimum | centroid |
|---|---|---|---|---|---|---|
| 5.5 | 4.2 | 0.989388139 | 1.62788206 | 1.31318104 | 0.9893881 | red |
| 5.1 | 3.8 | 0.45338235 | 1.74642492 | 1.229272594 | 0.4533824 | red |
| 4.7 | 3.2 | 0.314466038 | 2.002498439 | 1.348249894 | 0.314466 | red |
| 5.9 | 3.2 | 0.94809751 | 0.806225775 | 0.240370085 | 0.2403701 | yellow |
| 4.9 | 3.1 | 0.274873708 | 1.8 | 1.137736544 | 0.2748737 | red |
| 6.7 | 3.1 | 1.753726192 | 0 | 0.674124947 | 0 | blue |
| 5 | 3 | 0.368178701 | 1.702938637 | 1.033333333 | 0.3681787 | red |
| 6 | 3 | 1.096458947 | 0.707106781 | 0.033333333 | 0.0333333 | yellow |
| 4.6 | 2.9 | 0.593483127 | 2.109502311 | 1.436817471 | 0.5934831 | red |
| 6.2 | 2.8 | 1.357284871 | 0.583095189 | 0.260341656 | 0.2603417 | yellow |

As there is no change in the clustering, this is the final clustering obtained:


K-Means plot (final)
4.4
4.2
4
3.8
3.6
3.4
3.2
3
2.8
2.6
4.2
4.7
5.2
5.7
6.2
6.7
7.2

Q2)

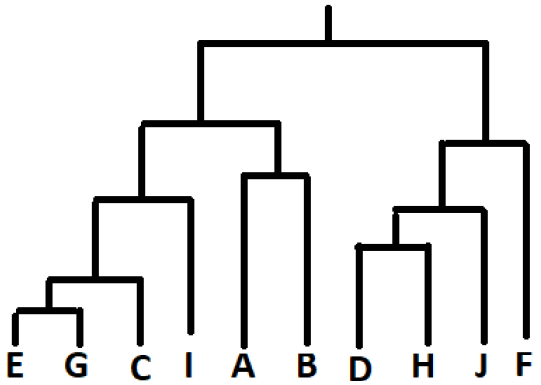
Letters were assigned to the coordinates to be able to refer to them.

Distances between every point and every other point:

| | x | y | a | b | c | d | e | f | g | h | i | j |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | 5.5 | 4.2 | | 0.5657 | 1.2806 | 1.077 | 1.253 | 1.6279 | 1.3 | 1.3 | 1.5811 | 1.5652 |
| b | 5.1 | 3.8 | 0.5657 | | 0.7211 | 1 | 0.728 | 1.7464 | 0.8062 | 1.2042 | 1.0296 | 1.4866 |
| c | 4.7 | 3.2 | 1.2806 | 0.7211 | | 1.2 | 0.2236 | 2.0025 | 0.3606 | 1.3153 | 0.3162 | 1.5524 |
| d | 5.9 | 3.2 | 1.077 | 1 | 1.2 | | 1.005 | 0.8062 | 0.922 | 0.2236 | 1.3342 | 0.5 |
| e | 4.9 | 3.1 | 1.253 | 0.728 | 0.2236 | 1.005 | | 1.8 | 0.1414 | 1.1045 | 0.3606 | 1.3342 |
| f | 6.7 | 3.1 | 1.6279 | 1.7464 | 2.0025 | 0.8062 | 1.8 | | 1.7029 | 0.7071 | 2.1095 | 0.5831 |
| g | 5 | 3 | 1.3 | 0.8062 | 0.3606 | 0.922 | 0.1414 | 1.7029 | | 1 | 0.4123 | 1.2166 |
| h | 6 | 3 | 1.3 | 1.2042 | 1.3153 | 0.2236 | 1.1045 | 0.7071 | 1 | | 1.4036 | 0.2828 |
| i | 4.6 | 2.9 | 1.5811 | 1.0296 | 0.3162 | 1.3342 | 0.3606 | 2.1095 | 0.4123 | 1.4036 | | 1.6031 |
| j | 6.2 | 2.8 | 1.5652 | 1.4866 | 1.5524 | 0.5 | 1.3342 | 0.5831 | 1.2166 | 0.2828 | 1.6031 | |

**Single linkage:**

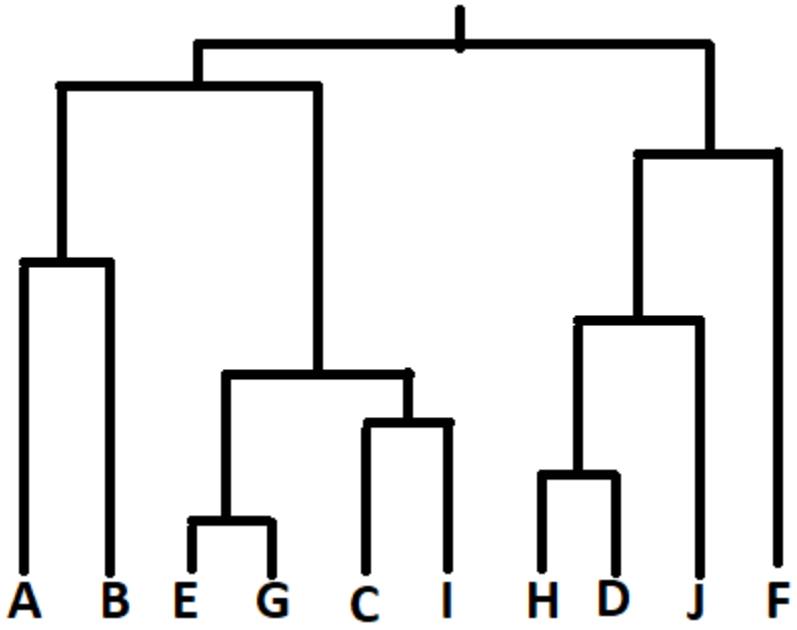
- Closest two are E and G (0.141421356) and are now clustered together
- Closest two are C and E within cluster 1 (0.223606798). C, E, G Now clustered
- Closest two are D and H (0.223606798) and are clustered together
- Closest two are J and H within cluster 2 (0.282842712). J D H Now clustered
- Closest two are I and C within cluster 1 (0.316227766). C E G I Now clustered
- Closest two are A and B (0.565685425) and are now clustered together
- Closest two are F and J within cluster containing (DHJ) (0.583095189). DHJF Formed
- Closest two are B within cluster (A, B) and C within cluster containing (EGCI) (0.721110255). EGCIAB formed
- 2 remaining clusters are joined
- **Dendrogram**:



**Complete linkage:**

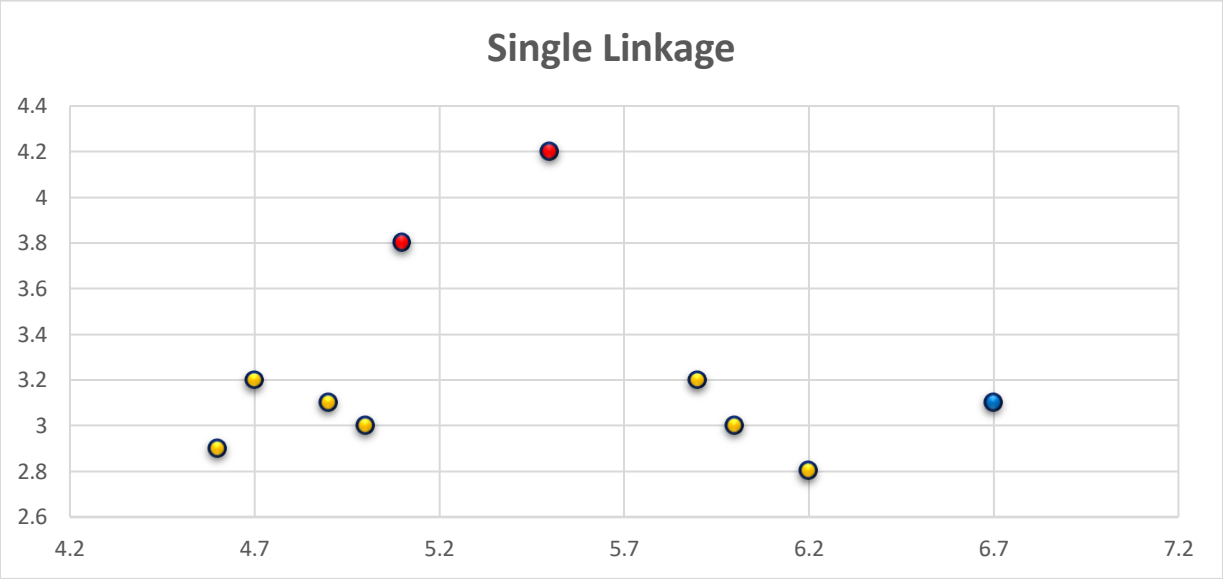
- Closest two are E and G (0.141421356) and are now clustered together
- Closest two are D and H (0.223606798) and are clustered together
- Closest two are C and I (0.316227766) and are clustered together
- Closest two are cluster EG and cluster CI (0.412310563) and are clustered together forming cluster (E, G, C, I)
- Closest two are cluster HD and J (0.5) and are clustered together forming (H, D, J)
- Closest two are A and B (0.565685425) and are now clustered together
- Closest two are cluster HDJ and F (0.806225775) and are clustered together forming (H, D, J, F)
- Closest two are clusters AB and EGCI (1.58113883) they form cluster (A, B, E, G, C, I)
- The remaining clusters are then merged together. All now clustered

**Dendrogram:**

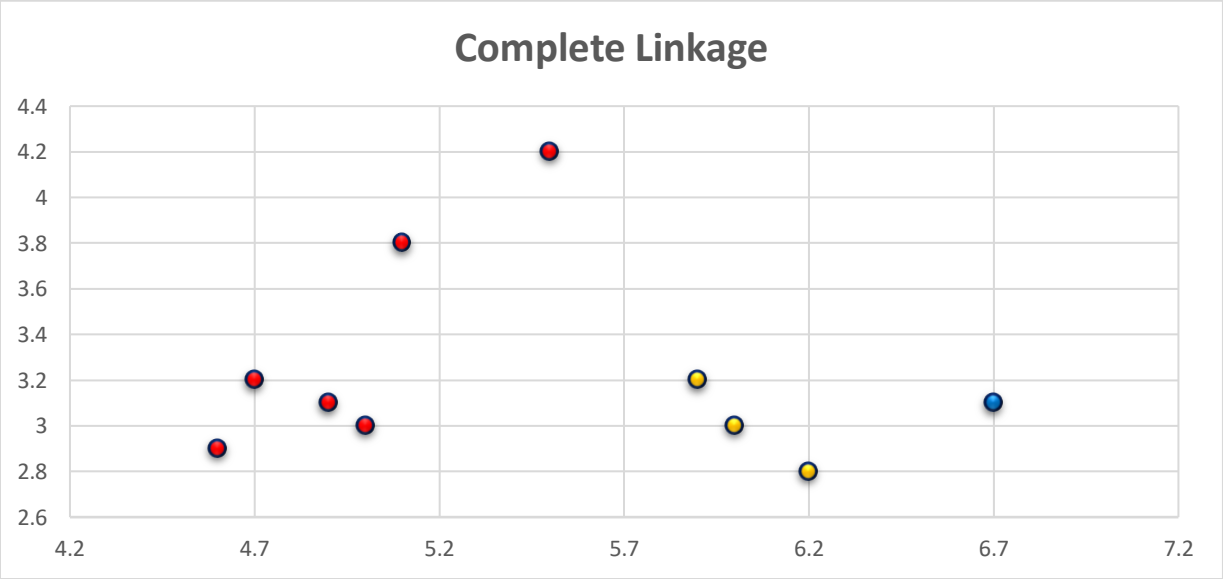


**Q3)**

When partitioned into 3 clusters, the single linkage is made up of **(AB), (F), (EGCDHJI)** Which looks as follows:




The complete linkage is partitioned into **(ABEGCI), (F), (HDJ)** Which looks as follows:




As can be seen, the results for complete linkage and K-means were the same, while Single linkage was different. Visually, single means produced worse results.

As for intra and inter cluster distances:

K-Means:

| inter cluster distance | | |
|---|---|---|
| red | yellow | blue |
| 2.993772063 | 0.534045074 | 0 |
| Average = | 1.175939046 | |
| Intra Cluster distance | | |
| 3.555779427 | | |

Complete Linkage:

| inter cluster distance | | |
|---|---|---|
| red | yellow | blue |
| 2.993772063 | 0.534045074 | 0 |
| Average = | 1.175939046 | |
| Intra Cluster distance | | |
| 3.555779427 | | |

Single Linkage:

Centroids:

| | | |
|---|---|---|
| red | 5.3 | 4 |
| blue | 6.7 | 3.1 |
| yellow | 5.328571429 | 3.028571429 |

| inter cluster distance | | |
|---|---|---|
| red | yellow | blue |
| 0.565685425 | 4.325183886 | 0 |
| Average = | 1.63028977 | |
| Intra Cluster distance | | |
| 4.009467777 | | |

As can be seen, the K-Means and Complete Linkage methods outperform the Single linkage by having smaller Intra and Inter Cluster distances. The yellow cluster from single linkage was too elongated, which is an issue with single linkage.

Q4)

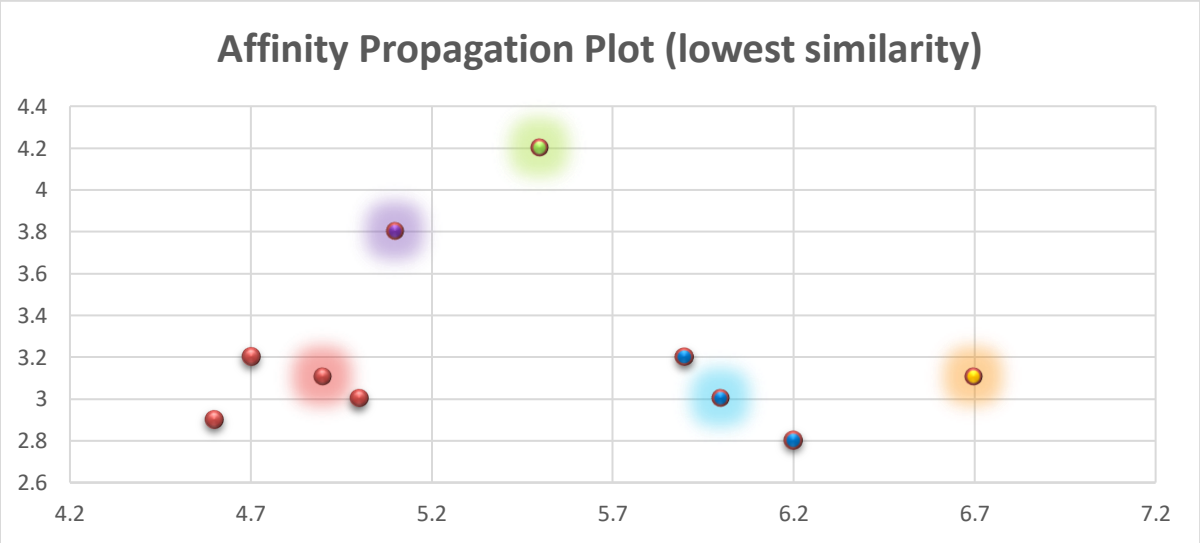
When Self-Prefrance value is the lowest similarity (-0. 141421356) the output from the program is :

**[1, 2, 5, 8, 5, 6, 5, 8, 5, 8]**

**Which is:**

| letter | number | x | y | exemplar number | exemplar |
|---|---|---|---|---|---|
| a | 1 | 5.5 | 4.2 | 1 | a |
| b | 2 | 5.1 | 3.8 | 2 | b |
| c | 3 | 4.7 | 3.2 | 5 | e |
| d | 4 | 5.9 | 3.2 | 8 | h |
| e | 5 | 4.9 | 3.1 | 5 | e |
| f | 6 | 6.7 | 3.1 | 6 | f |
| g | 7 | 5 | 3 | 5 | e |
| h | 8 | 6 | 3 | 8 | h |
| i | 9 | 4.6 | 2.9 | 5 | e |
| j | 10 | 6.2 | 2.8 | 8 | h |

Glowing points are exemplars



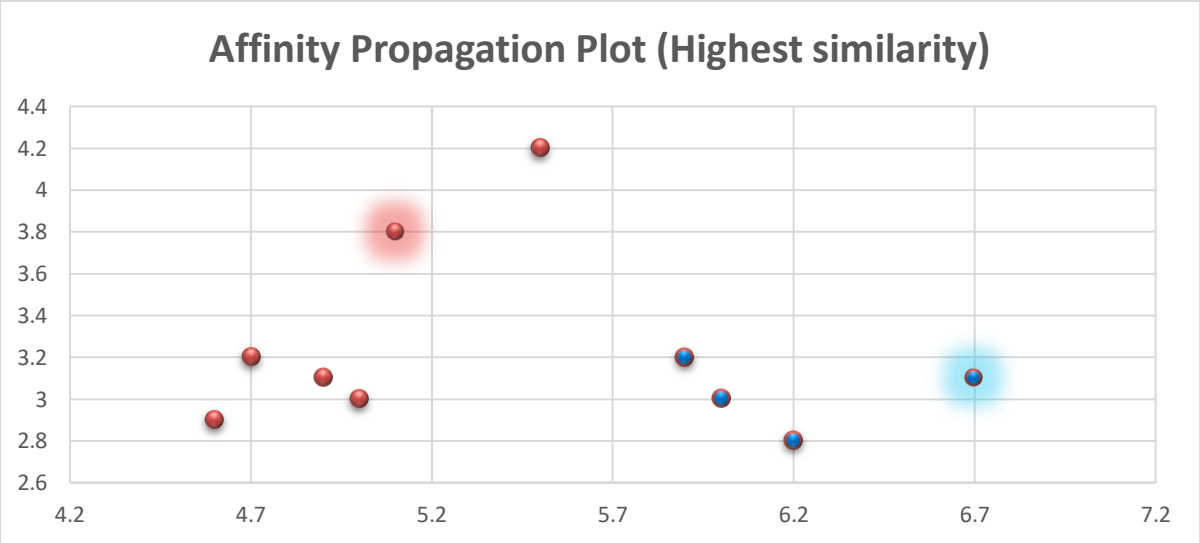
When Self-Prefrance value is the Highest similarity (-2.109502311) the output from the program is :

**[2, 2, 2, 6, 2, 6, 2, 6, 2, 6]**

Which is:

| letter | number | x | y | exemplar number | exemplar |
|---|---|---|---|---|---|
| a | 1 | 5.5 | 4.2 | 2 | b |
| b | 2 | 5.1 | 3.8 | 2 | b |
| c | 3 | 4.7 | 3.2 | 2 | b |
| d | 4 | 5.9 | 3.2 | 6 | f |
| e | 5 | 4.9 | 3.1 | 2 | b |
| f | 6 | 6.7 | 3.1 | 6 | f |
| g | 7 | 5 | 3 | 2 | b |
| h | 8 | 6 | 3 | 6 | f |
| i | 9 | 4.6 | 2.9 | 2 | b |
| j | 10 | 6.2 | 2.8 | 6 | f |

Glowing points are exemplars




As we can See, the highest similarity setting produces fewer, large clusters while the lowest similarity produces fewer, smaller clusters. Both intra and inter cluster similarity is better for the lowest similarity than the highest similarity setting as the clusters are fewer and closer, and so are the point within them.

The highest similarity setting produced similar results to the k-means and complete linkage but with one less cluser, as the number of clusters in not fixed in the AP algorithm. The lowest similarity method produced better metrics but a bit too much clusters.

Perhaps using the average or mean similarity as a self similarity will produce better overall clustering.

For the lowest similarity setting:

| inter cluster distance | | |
|---|---|---|
| cluster 5 | cluster 8 | cluster 1,2,6 |
| 0.748236244 | 0.534045074 | 0 |
| Average = | 0.256456264 | |

For Highest similarity the inter cluster similarity:

| cluster 2 | cluster 6 |
|---|---|
| 1.434 | 1.252 |
| Average = | 1.343 |

Which reinforces the points made above.

Comparing intra-cluster similarities for different numbers of clusters would not make sense as a larger number of clusters makes the similarity higher, as clusters are more frequent (average will be smaller as we divide by a larger number) and clusters are closer to eachother.