Data Analytics Project 2

Problem 1

Question 1)

Sl No	User	Jaws	Star Wars	Exorcist	Omen	Cluster ID
0	Kevin	4	2	4	5	0
1	George	4	3	1	2	1
2	Adele	5	4	2	3	1
3	James	2	3	1	1	1
4	William	3	4	2	4	1
5	Matt	3	3	3	3	1
6	Keith	5	4	5	4	1
7	Arnie	1	2	1	2	0
8	Sally	3	2	5	4	0
9	Sam	5	3	5	4	1

Question 2)
Looking at the SSE curve, we can see that the curve takes a sharp change at k = 2.
Therefore, k=2 is the best value according to the Elbow Method.

Question 3)

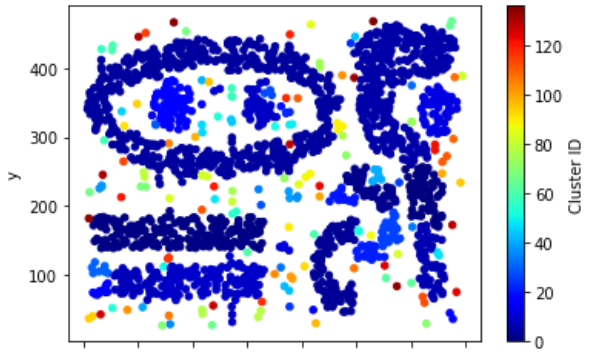
Clustering Algorithm	Cophenetic Correlation Coefficient
Group Average	0.4886522573
Complete Link	0.6063706366
Single Link	0.3558041132

The Complete Link (MAX) Algorithm has the highest Cophenetic Correlation Coefficient.
Therefore, the Complete Link (MAX) algorithm shows best match with the class labels.

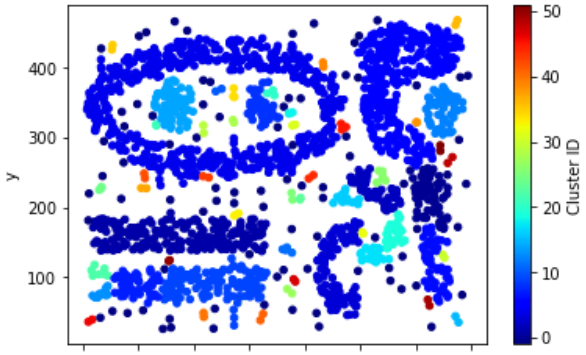
Question 4)

Eps	Min Samples	Number of Clusters
12.5	1	137
12.5	2	52
12.5	3	24
12.5	4	24
12.5	5	28

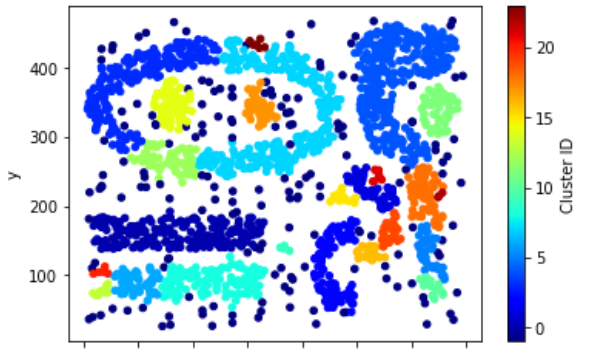
Min Samples = 1, eps = 12.5



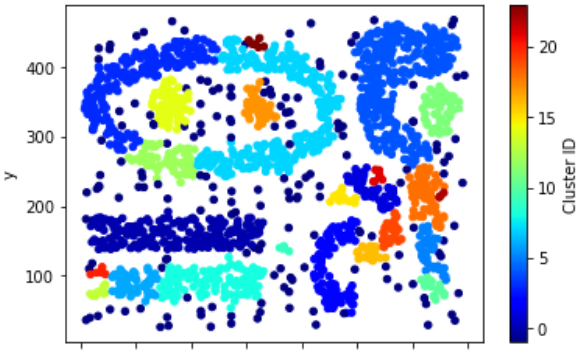
Min Samples = 2, eps = 12.5



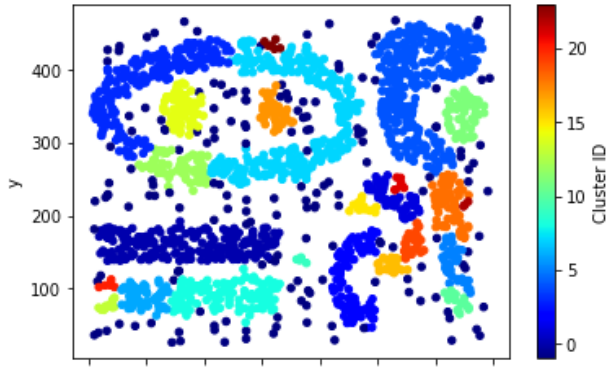
Min Samples = 3, eps = 12.5



Min Samples = 4, eps = 12.5

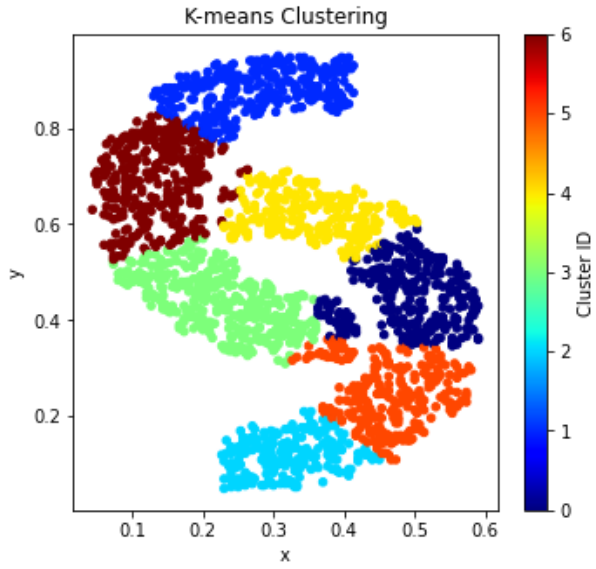


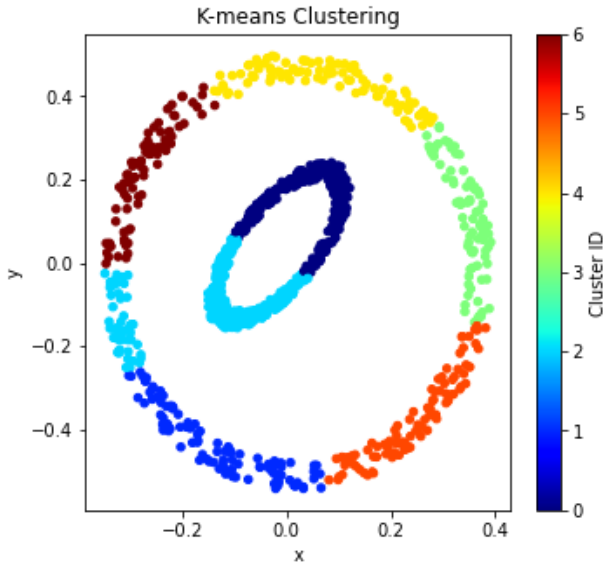
Min Samples = 5, eps = 12.5



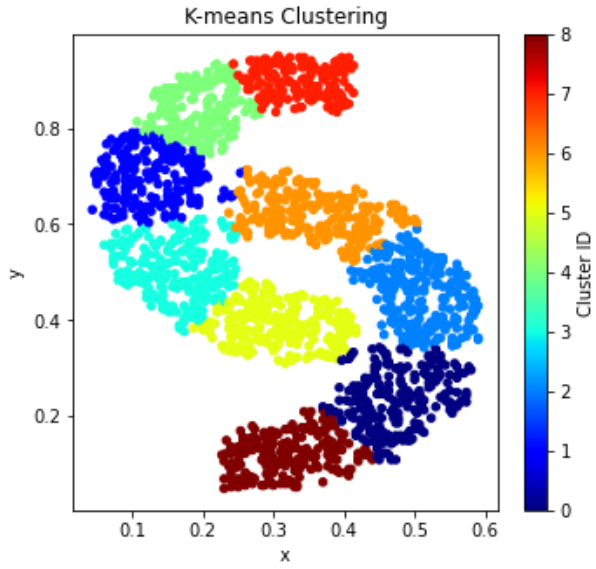
Question 5)

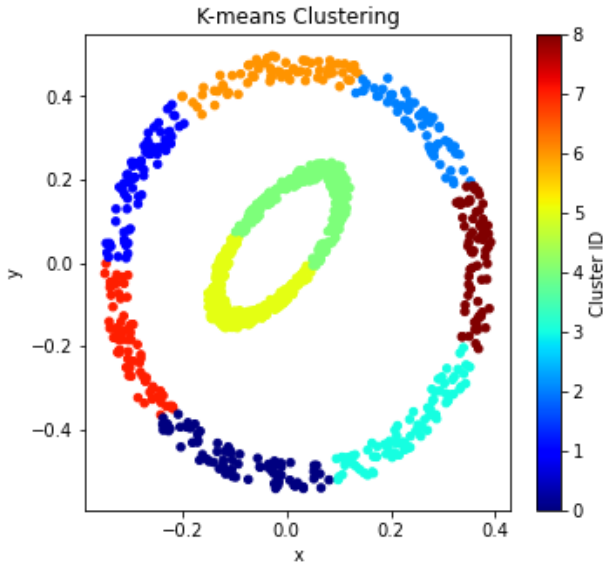
k = 7





k = 9





Problem 2

Question 1)

Cluster 1 :
Purity = $\frac{890}{1927} = 0.462$
Cluster 2 :
Purity = $\frac{1344}{2631} = 0.510$
Cluster 3 :
Purity = $\frac{1241}{2470} = 0.502$
Cluster 4 :
Purity = $\frac{562}{1343} = 0.418$
Cluster 5 :
Purity = $\frac{196}{279} = 0.702$
Cluster 6 :
Purity = $\frac{1558}{1649} = 0.945$
The maximum purity metric is 0.945

Question 2)
The maximum purity metric amongst all the 10 clusters is 0.958 of cluster 4.
The value has increased as there are more clusters than required.