Data Science Project
Sub-Project – VII
Clustering

Student's Name: Anmol Bishnoi **Mobile No: 7042845211**

Roll Number: B19069 **Branch: CSE**

1 a.

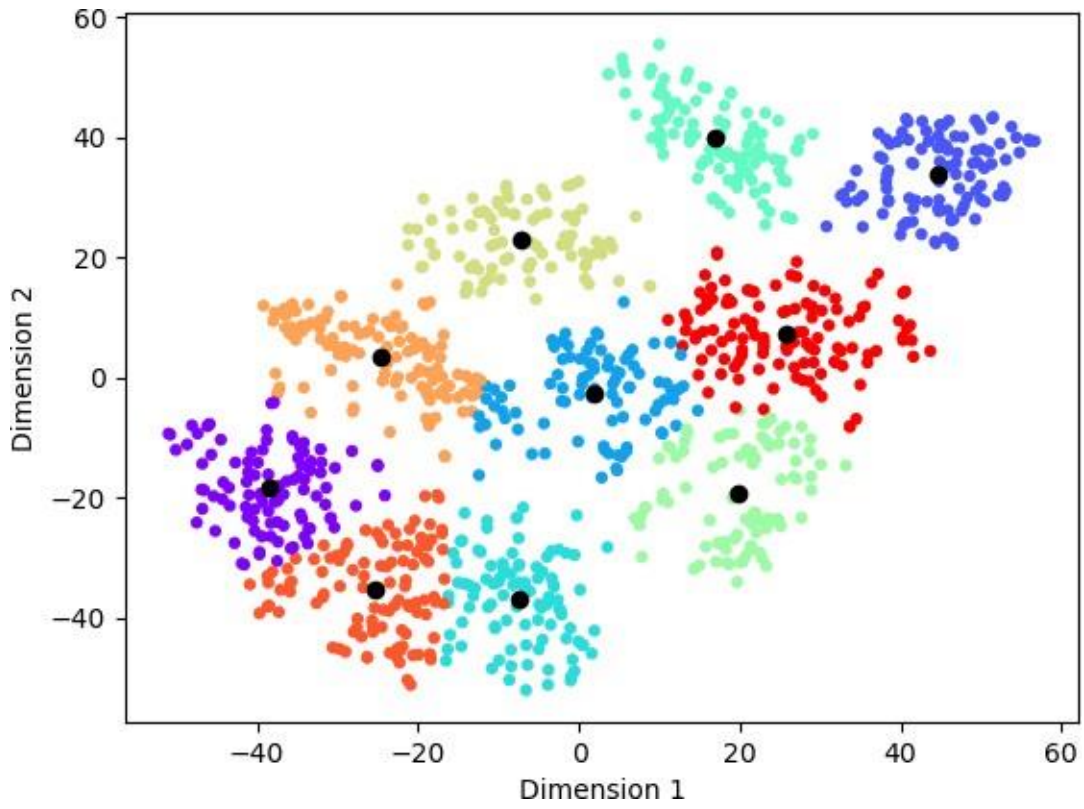


Figure 1 K-means (K=10) clustering on the mnist tsne training data

Inferences:

1. K-Means is a clustering algorithm of unsupervised learning. Looking at the above graphs, it has formed circular clusters. The grouping looks accurate though the purity score is 0.69. Being an unsupervised algorithm, K-Means can provide well-formed clusters and is a good clustering algorithm.
2. Yes, K-Means forms circular boundaries. Though some cluster boundaries look like oval shaped but majority have circular boundaries.

b.

The purity score after training examples are assigned to the clusters is **0.69**

c.

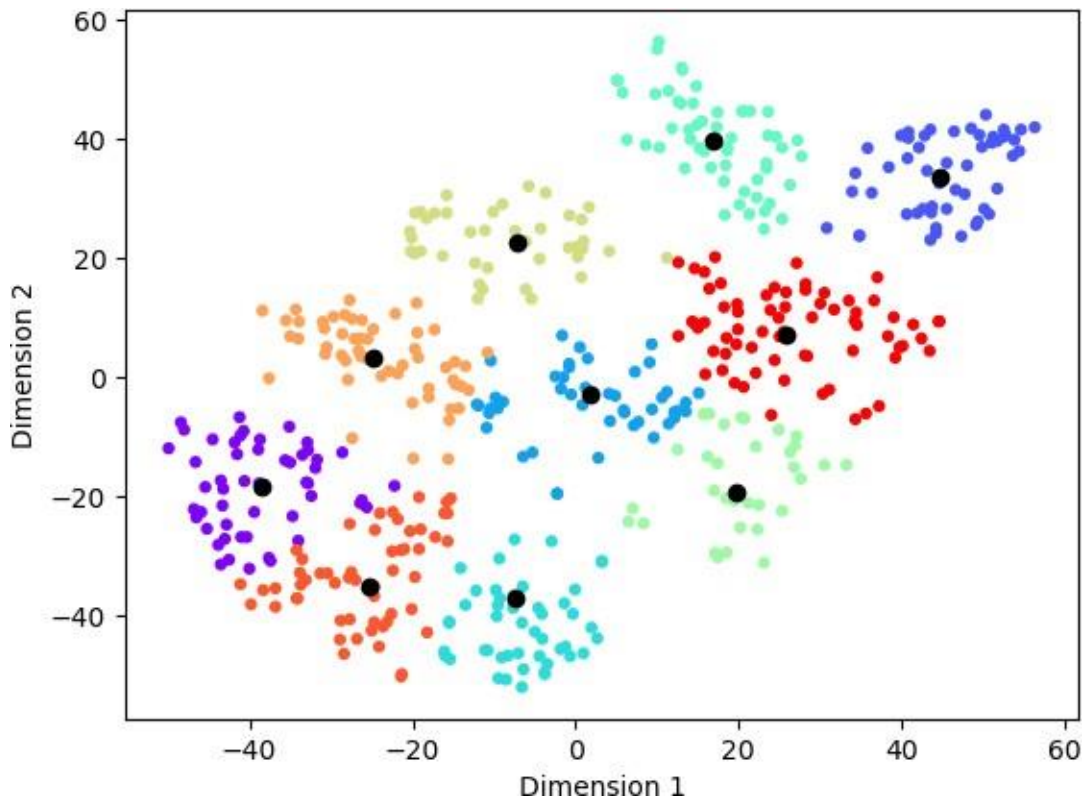


Figure 2 K-means (K=10) clustering on the mnist tsne test data

Inferences:

1. Both the graphs are almost similar, the clusters and the cluster centers. The purity scores are also similar.

d.

The purity score after test examples are assigned to the clusters is 0.**676**

Inferences:

1. Train purity is slightly higher than test purity. This may be because the model is fit on train data and hence gives more accurate results on the train data.
2. The value of K is to be chosen after experimenting. Small K and high values of K does not give accurate results. Elbow method can be used for calculating the optimum K. K-Means also does not perform well when the data consists of several outliers.

2 a.

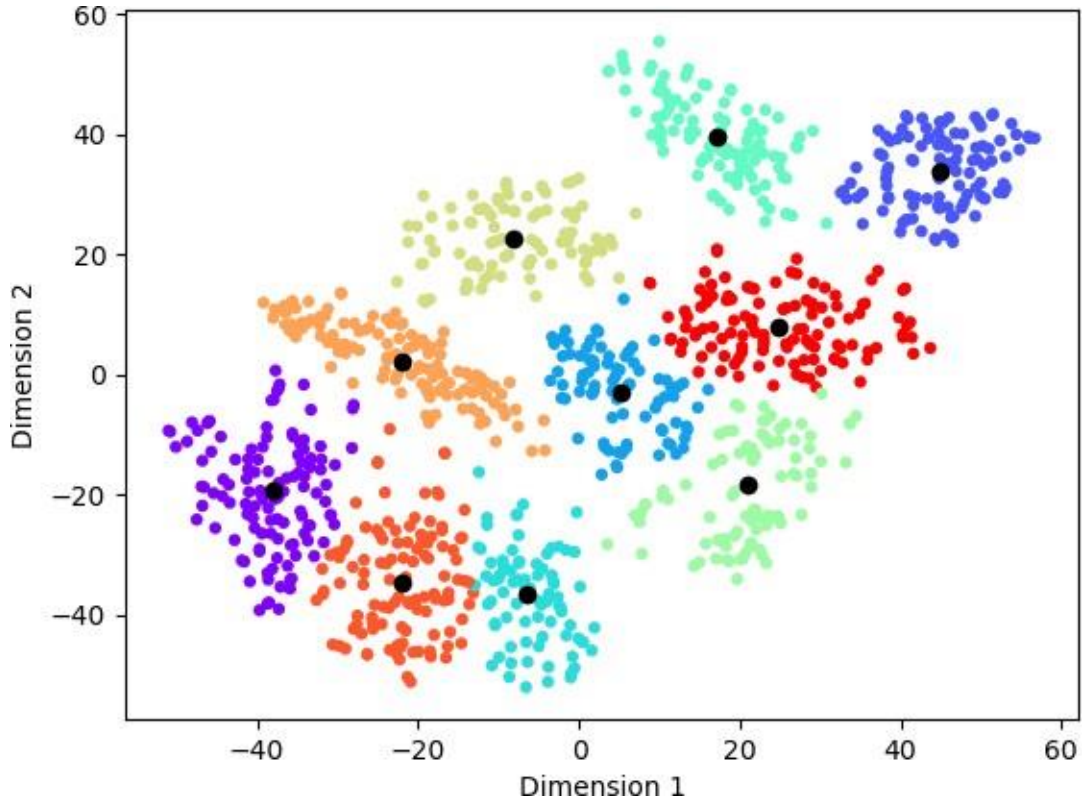


Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. GMM is an algorithm of unsupervised learning. Looking at the above graphs, it has formed elliptical clusters. The grouping looks accurate though the purity score is 0.708. Being an unsupervised algorithm, GMM can provide well-formed clusters and is a good clustering algorithm.
2. Yes, GMM forms elliptical boundaries. Though some cluster boundaries look like circular shaped but majority have elliptical boundaries.
3. Surprisingly, both the graphs don't differ much but the boundaries are more circular in K-means and elliptical in GMM.

b.

The purity score after training examples are assigned to the clusters is **0.708**

c.

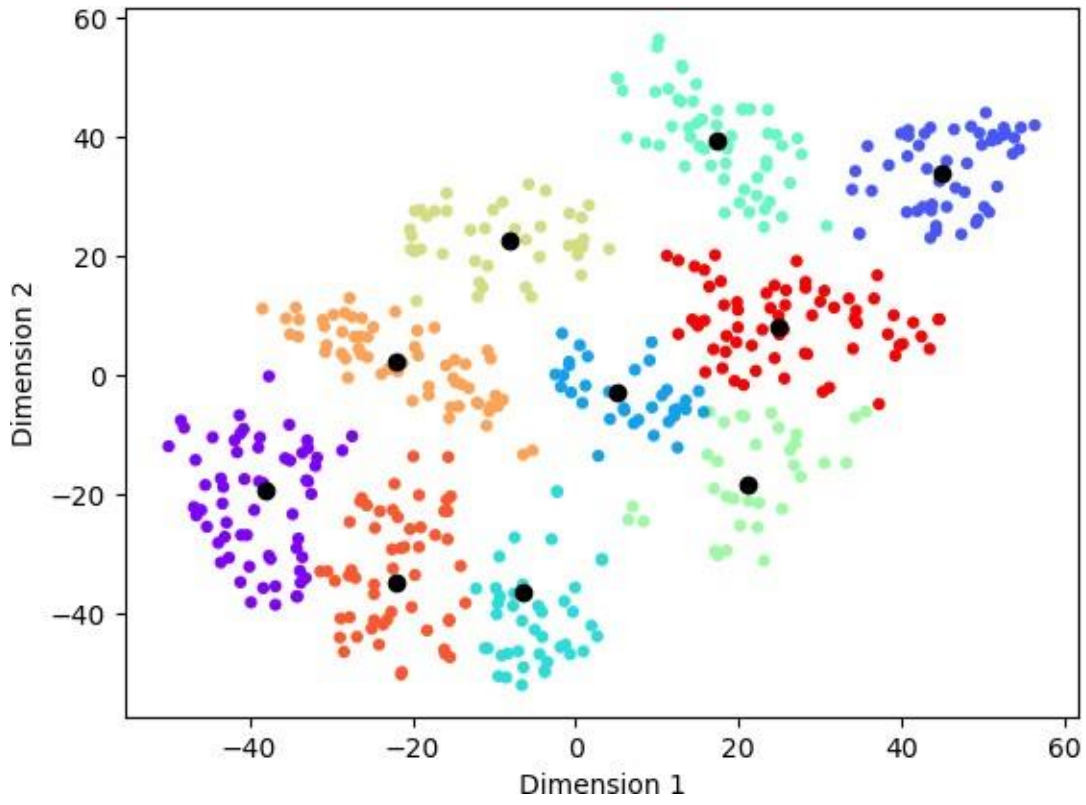


Figure 4 GMM clustering on the mnist tsne test data

Inferences:

1. Both the graphs are almost similar, the clusters and the cluster centers. The purity scores are also similar.

d.

The purity score after test examples are assigned to the clusters is 0.**704**

Inferences:

1. Train purity is slightly higher than test purity. This may be because the model is fit on train data and hence gives more accurate results on the train data.
2. The main limitation of the GMM algorithm is that, for computational reasons, it can fail to work if the dimensionality of the problem is too high.

3 a.

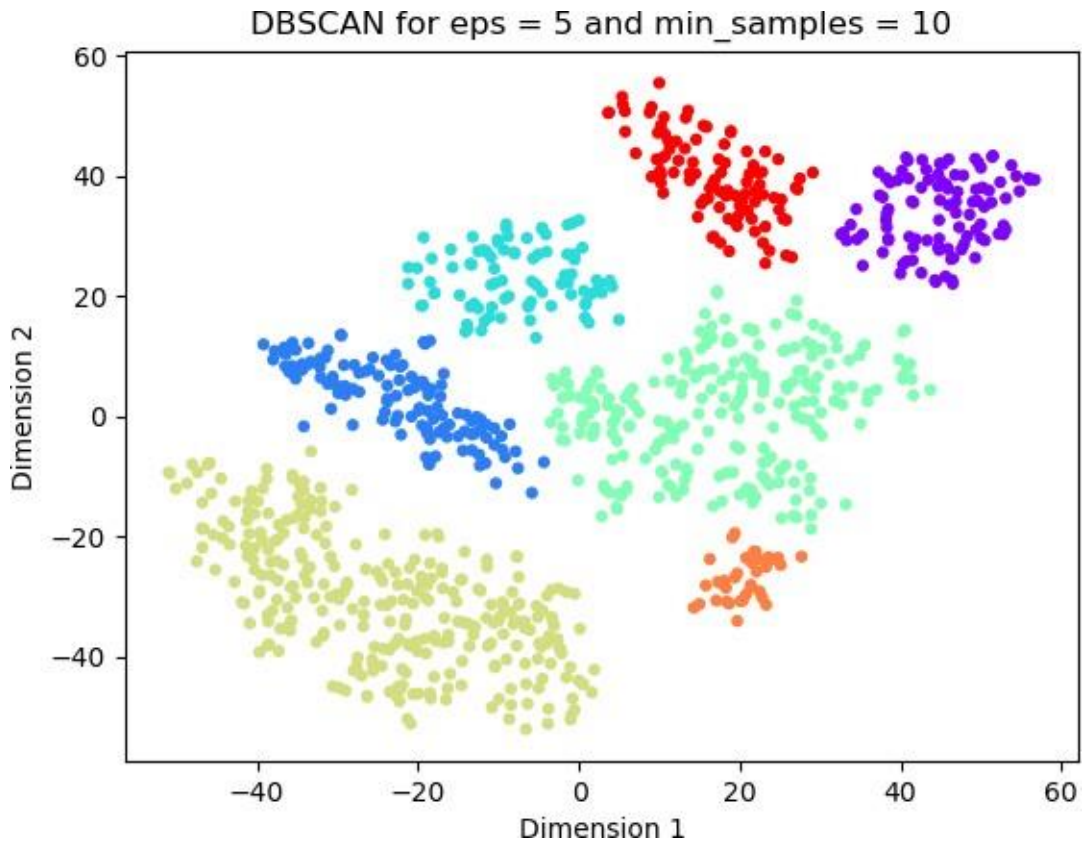


Figure 5 DBSCAN clustering on the mnist tsne training data

Inferences:

1. DBSCAN ignore outliers (-1 cluster). This might be both positive and negative. Purity score is less and the reason can be outliers.
2. Yes, here are differences between the number and location of clusters formed with K-Means and GMM and DBSCAN. Also the cluster boundaries are different as DBSCAN has no definite boundary shape.

b.

The purity score after training examples are assigned to the clusters is **0.602**

c.

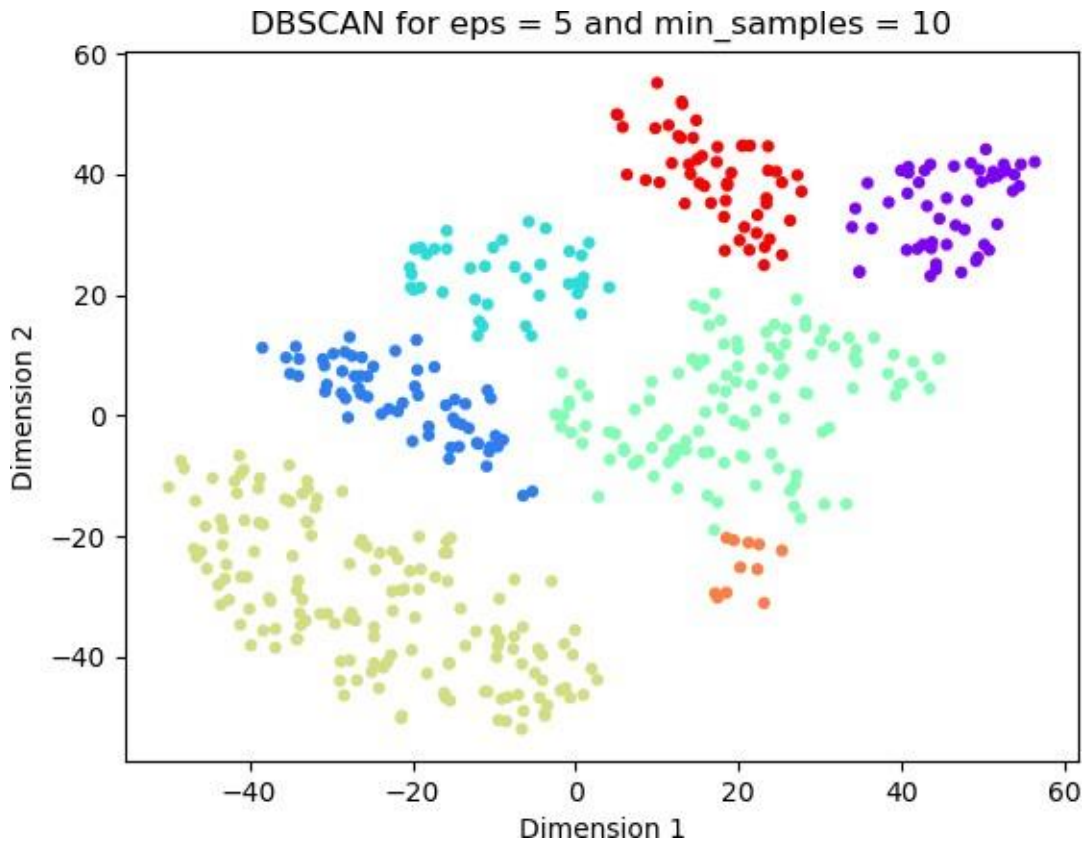


Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. Both the graphs are almost similar, the clusters and the cluster centers. The purity scores are also similar.

d.

The purity score after test examples are assigned to the clusters is **0.598.**

Inferences:

1. Train purity is slightly higher than test purity. This may be because the model is fit on train data and hence gives more accurate results on the train data.
2. DBSCAN struggles with clusters of similar density.

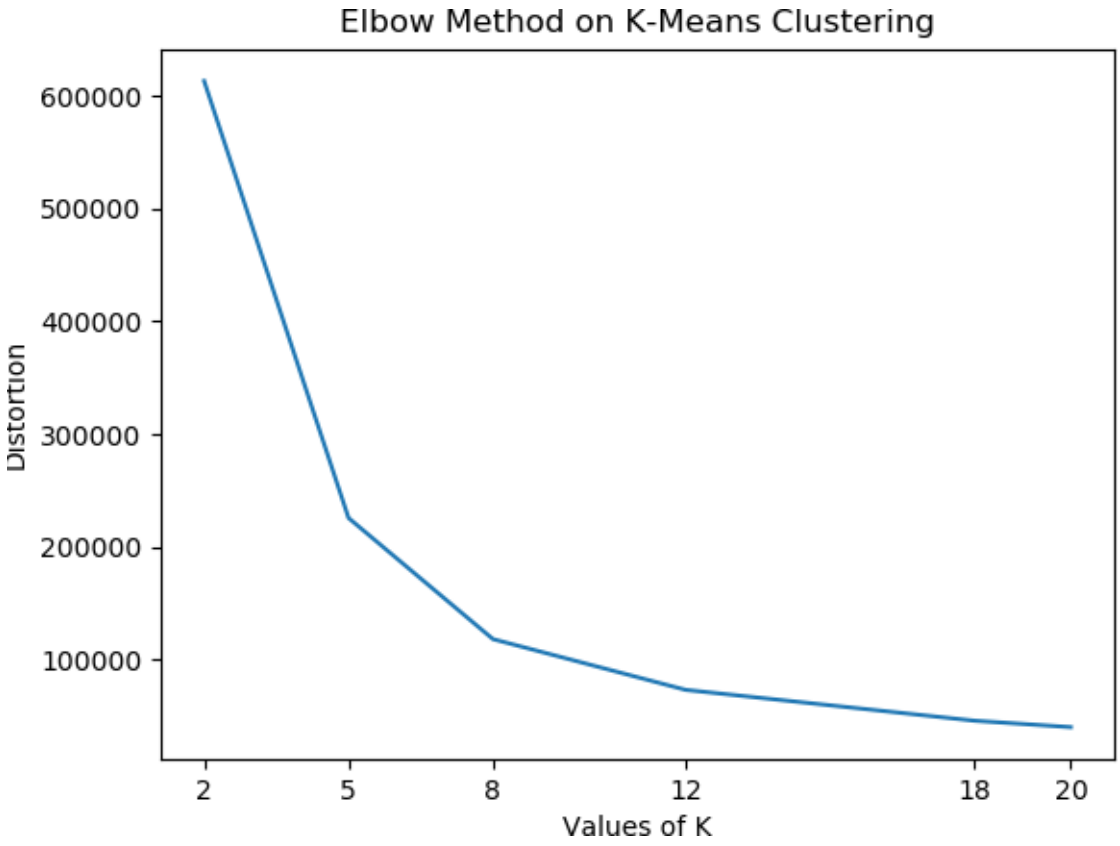
Bonus Questions

Question A

Note: (Plotted 24 graphs, 12 of K-Means and 12 of GMM, 2 each for every value of K in the code but not pasted in the report.)

K-Means:

K	Distortion Score
2	613349.83
5	225666.88
8	118198.67
12	73211.55
18	45828.03
20	40145.30

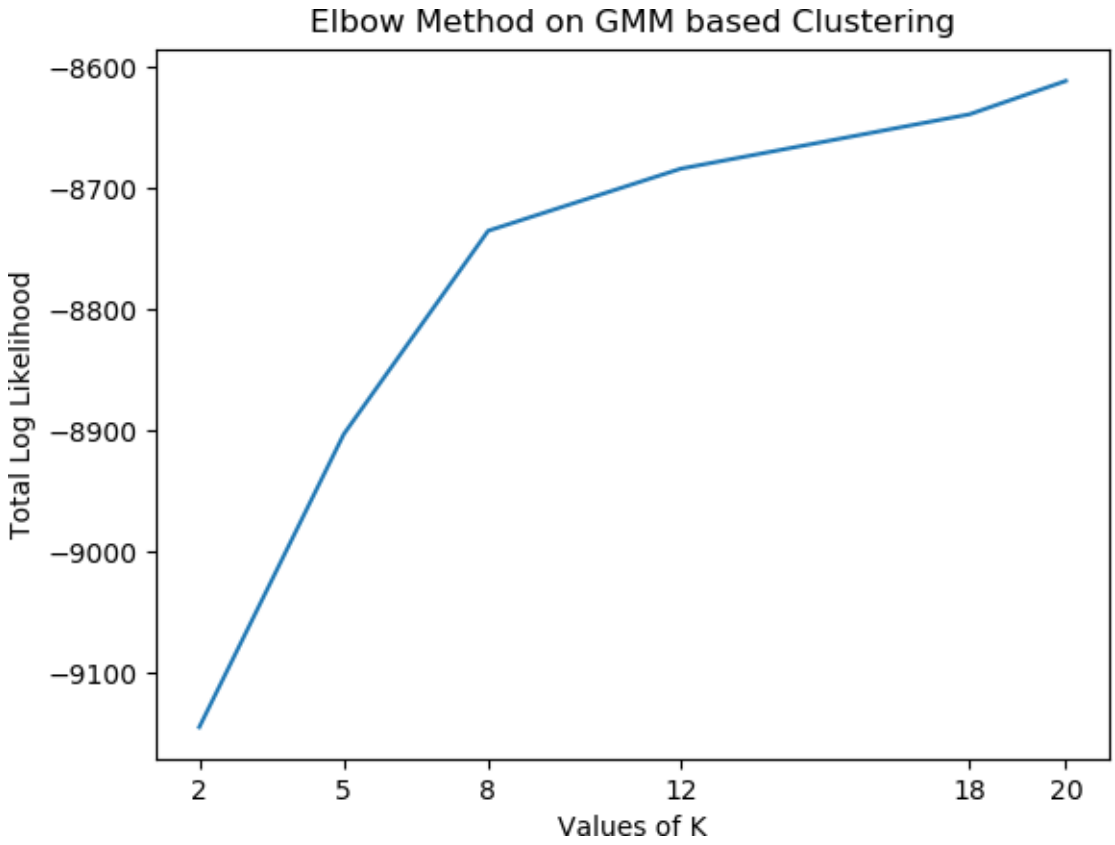


Optimal value of K using K-Means = 8

The most optimal value for K = 8. The distortion decreases almost exponentially till 8 after which the slope is almost linear.

GMM:

K	Total Log Likelihood
2	-9144.70
5	-8903.15
8	-8735.33
12	-8684.24
18	-8639.24
20	-8611.72



Optimal value of K using GMM = 8

The most optimal value for K = 8. The total log likelihood increases almost exponentially till 8 after which the slope is almost linear.

Question B

Note: (Plotted 12 graphs)

Eps	MinPoints	Purity Score	Cluster Formed
1	1	0.036	716
	10	NA	0
	30	NA	0
	50	NA	0
5	1	0.288	8
	10	0.602	7
	30	0.951	2
	50	NA	0
10	1	0.1	1
	10	0.1	1
	30	0.1	1
	50	0.488	5

The max purity scores is for Eps = 5 and Min_Samples = 30