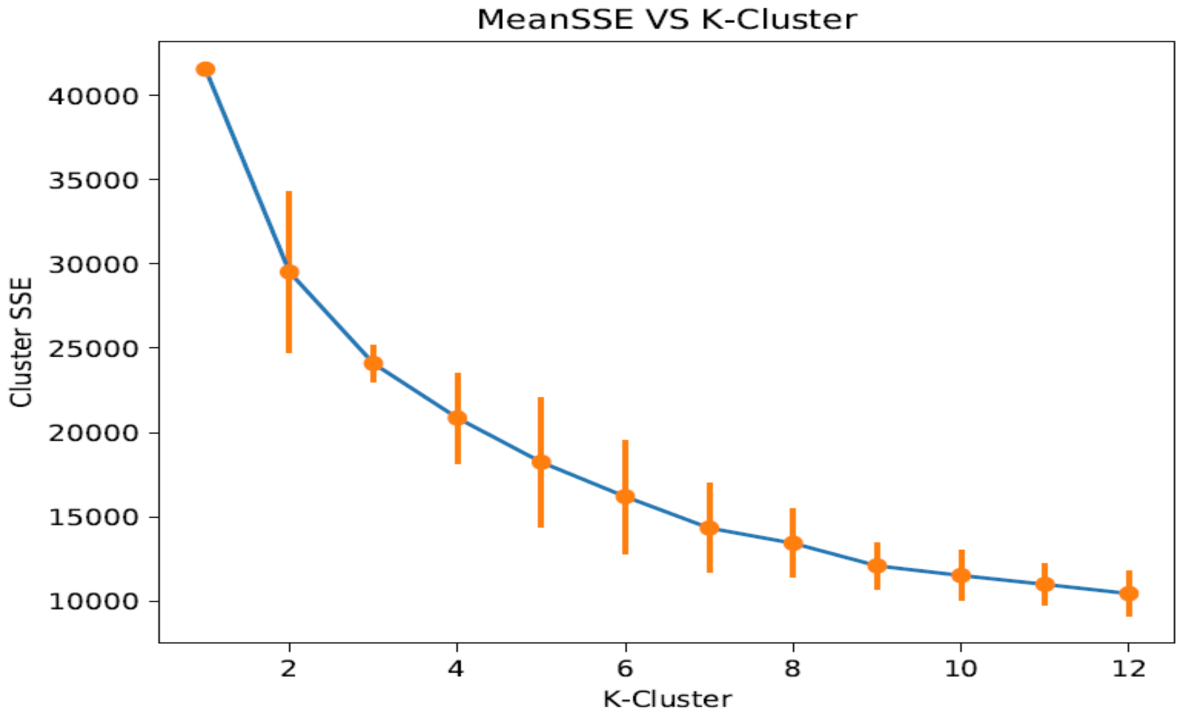
Question 1.

a) Below is the line plot of the mean SSE as a function of K, with error bars indicating 95% confidence interval. (The same can be generated through code, Please see the readme file for details)



b) Below is the table containing the 4 columns: k, μk, μk - 2σk and μk + 2σk for each of the values of k = 1, 2, ... , 12

K Value	μ_k	$\mu_k - 2\sigma_k$	$\mu_k + 2\sigma_k$
1	41562	41562	41562
2	29523.8836	24716.44274	34331.32446
3	24101.584	23001.42938	25201.73862
4	20865.6788	18188.37363	23542.98397
5	18240.3704	14410.3202	22070.4206
6	16197.9576	12832.61216	19563.30305
7	14333.8252	11686.78863	16980.86177
8	13429.1008	11390.54098	15467.66062
9	12090.8588	10738.246	13443.4716
10	11516.5804	10035.46508	12997.69572
11	10989.3824	9777.682488	12201.08231
12	10443.2524	9132.335575	11754.16923

c) When "K" increases and approaches to the total number of example N, SSE would keep decreasing and approaches to "0" when K reach to number of example "N", since the Euclidian distance from the cluster centroid will keep decreasing and eventually becomes 0 when there are clusters equal to total number of example (Since in this case centroid and data point would be the same).

If we use SSE to choose the optimal "K", it may give is more number of cluster than the actual (according to the domain knowledge of data). Since Best SSE will be when number of cluster approaches towards data sample "N", which is practically impossible. Which is why we use elbow method to find the optimal "K" value.

d) Another measure of cluster compactness and separation can be Euclidian distance between the data points. We can find the distance between the farthest points with-in the cluster to validate the cluster compactness and the distance between the closest points among the cluster to find the cluster separation.
Lower distance between the farthest points with-in the cluster signifies the better compactness of the cluster.
Higher distance between the closest points among the cluster signifies the better separation of the cluster.

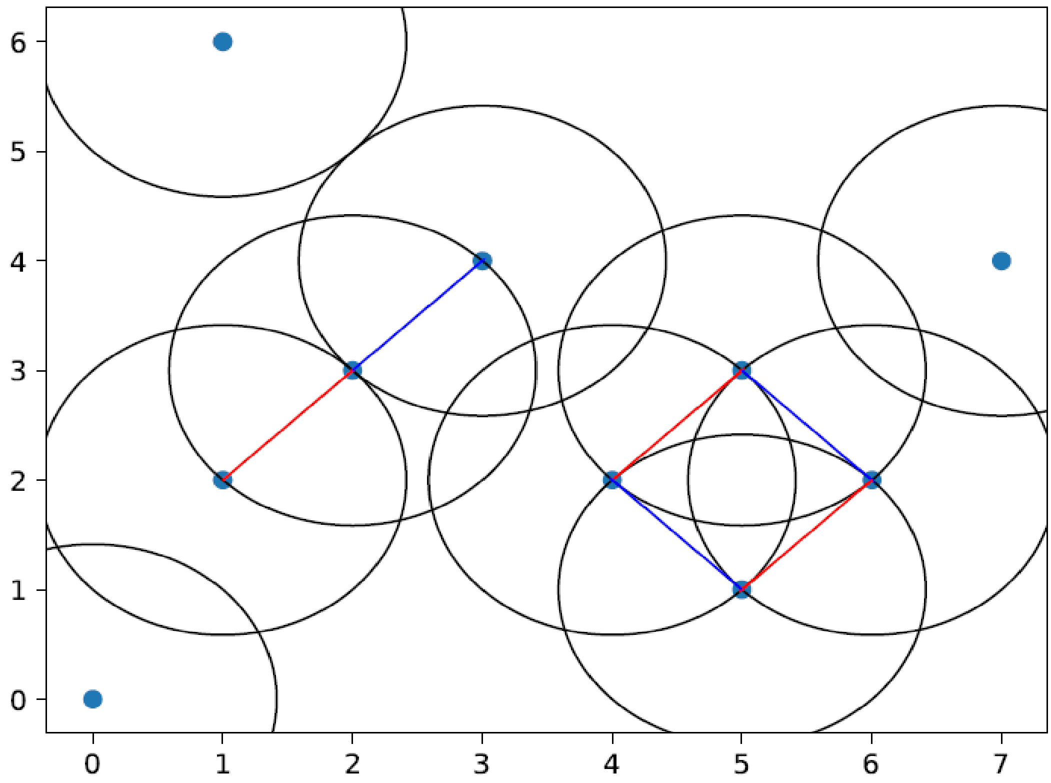
Another way to measure the cluster compactness could be the standard deviation with-in the cluster. Lower standard deviation will signifies the better compactness of the cluster since data points will be close to the cluster centroid (or average)

Question 2.

Handout sheet is attached with the submission with the detailed calculation (Question-2-Handout). Please see the readme file for more details.

Question 3.

Below is the plot which describe the density based cluster and to answer the questions.



a) Following are the two cluster and their points.

Cluster-1 :- { (1,2), (2,3), (3,4) }
Cluster-2 :- { (4,2), (5,1), (6,2), (5,3) }

b) Following are the two set of density connected points.

{ (1,2), (2,3), (3,4) }
{ (4,2), (5,1), (6,2), (5,3) }

c) Following are the points DBSCAN considered as noise.

{ (0,0), (1,6), (7,4) }

Question 4.

Confusion Matrix :-

	Truth	
Prediction	Positive	Negative
Positive	40 %	20 %
Negative	10 %	30 %

Accuracy :- **0.7**

Precision :- **0.67**

Recall :- **0.8**

F1 Score :- **0.727**

Specificity :- **0.6**