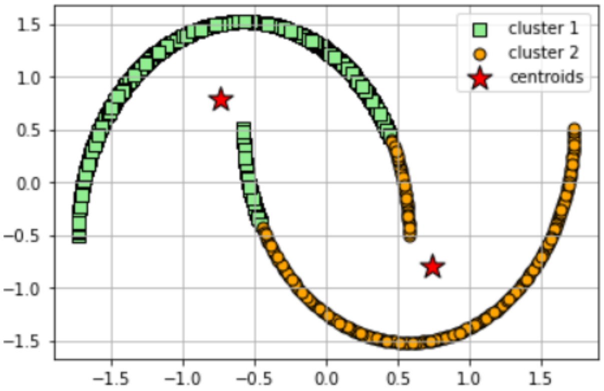
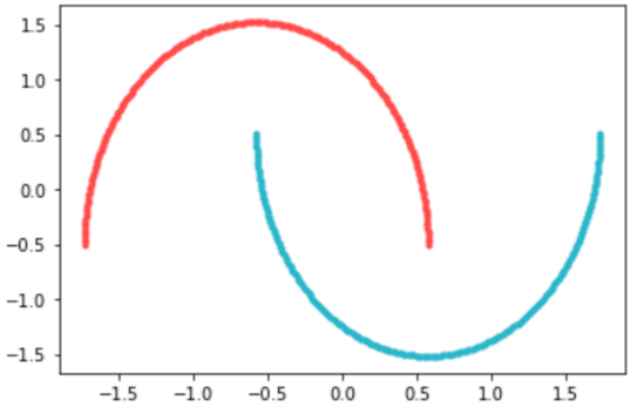
Assignment 5 Report

1a.



We plotted the points and coloured them according to the labels in the preceding plot.
The K-Means Algorithm predicts the outcome. The centroids have been highlighted in red.
Clusters form around the central points. This is to be expected, given that K-Means clustering is used which works by employing a distance measure to categorise the points.
We can't remark on the classification's nature because the ground truth isn't available.
However, by plotting the points in 2D, we can see that they form two semi-circles, which may be the case.
The two semi-circles are two separate clusters, in this case.

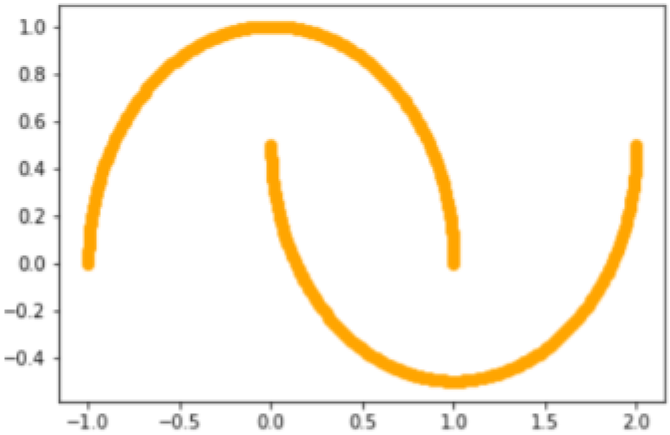
1b.



In the above image, the points have been plotted and coloured based on the labels predicted by the DBSCAN Algorithm.
Because clustering is based on the distance between two locations and the fact that a certain number of minimum samples are necessary to designate a point a core point, The points that are closest to the same cluster should be expected.

We can observe from the graph that the spots that are near together are grouped together. We also don't see any noise points in the graph.
We can't remark on the classification's nature because the ground truth isn't available. However, when we plot the points in 2D, we can see that they form two semi-circles, and it's possible that the two semi-circles are two different clusters. As a result of this assumption, we can deduce that the true character of the points has been captured.

1c.



When we plot the points in 2D, we can see that they form two semi-circles. We can't choose the optimum classification technique for the dataset since ground truth isn't provided.

1. K_Means: In the plot the clustered around the centers of the two cluster. This is expected since the K-means works by clustering the set of points in R radius as a cluster.

2. DBSCAN: In this plot the clusters can be seen as two groups each forming a cluster.

This happens because the DBSCAN algorithm model uses the relative distance between two points and assigns a set of ts (Core point) as a representative of the cluster and does not consider a single global representative of a cluster unlike the K-Means. The set of representative points for a cluster in DBSCAN is set by considering the neighboring points in a fixed radius from the point. And the core points are actual points from the dataset. Whereas in DBSCAN, the cluster representative point is outside the dataset.

1d.

DBSCAN
Pros: This returns the clusters formed and shows the noise data.
Cons: Need to optimize the parameters to get results. While elbow method is available for KMeans to find the K- cluster value.

K Means
Pros: The number of clusters existing in the dataset can be found by identifying the elbow from the graph using the ELBOW method.
Cons: It does not recognize the noise. It also classifies the noise as a cluster

2a.

- The simplest guideline for the number of iterations required for tSNE to converge is that the more iterations, the better. This, however, is not possible in the case of large data sets, as it may take a lot of time to complete ex :10,000 iterations . In contrast, if you use too few rounds, the clusters may not be apparent, and you may not get the results you want. Instead you will observe similar to the plot, find a large clump of data points in the centre of your tSNE plot when the number of iterations is equal to 250.

- If we look closely at the tSNE plots, we can see that the largest distance-gap between data points is about 100. This basic rule of thumb shows that the algorithm has arrived at its destination.Convergence and increasing the number of iterations will only make a slight difference. As a result, trials with 1000 and 2000 iterations achieve the same results.

2b.

- The t-SNE algorithm begins by computing the probability of point similarity in high-dimensional space and the probability of point similarity in the corresponding low-dimensional space. The conditional probability that a point A would choose point B as its neighbour if neighbours were chosen in proportion to their probability density under a Gaussian (normal) distribution centred at A is used to compute point similarity. For a perfect representation of data points in lower-dimensional space, it then tries to minimise the difference between these conditional probabilities (or similarities) in higher-dimensional and lower-dimensional space. t-SNE uses a gradient descent approach to minimise the sum of Kullback-Leibler divergence of overall data points to quantify the minimization of the sum of difference of conditional probability.

- A stochastic model is the t-SNE. Stochastic models are used to evaluate the probability of various outcomes by allowing for random fluctuation in inputs. A set of equations with input that expresses uncertainty across time is used to create stochastic models. As a result, when stochastic models are performed at different times, they will produce different outcomes.