1. For this problem, you will use the data in the `HW11-ClusteringData.csv` file for clustering. The file contains 1600 data points in $\mathbb{R}^2$ (i.e., 2-dimensional data points), and each row corresponds to a data point. The first two columns of the file contains the values of data points, and the third column contains the true labels of data points for your reference.

   The data points are generated using a Gaussian mixture model with 4 components. The mean vectors of the components are given by

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 4 \\ 6 \end{bmatrix}, \ \boldsymbol{\mu}_2 = \begin{bmatrix} -3 \\ 3 \end{bmatrix}, \ \boldsymbol{\mu}_3 = \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \ \text{and} \ \boldsymbol{\mu}_4 = \begin{bmatrix} -1 \\ -7 \end{bmatrix}.$$

   Also, $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4) = (0.1875, 0.25, 0.3438, 0.2188)$, where π_k is the probability that a data point belongs to component k.

   (a) Plot the data points using different colors for each label.
   Ans: See Figure 1.

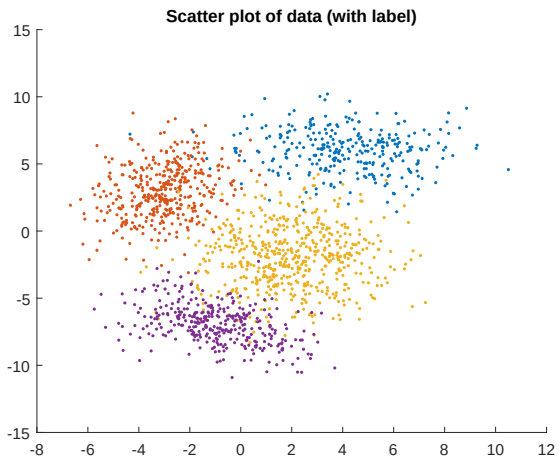

Figure 1: Plot of data points.

   (b) Cluster the data points using the k-means clustering for $k \in \{2, 3, \ldots, 7\}$. Use the silhouette coefficient or silhouette score to evaluate the quality of clustering for different values of k. The silhouette coefficient of a data point captures both intra-cluster cohesion (the average distance to other data points in the same cluster) and inter-cluster separation (the average distance to data points in the *nearest* cluster). For instance, if data point $\mathbf{x}_i$ belongs to the l-th cluster, then its silhouette coefficient $s(i)$ is given by

$$s(i) = \frac{\beta(i) - \alpha(i)}{\max(\alpha(i), \beta(i))},$$

   where

$$\alpha(i) = \frac{\sum_{\mathbf{x}_j \in C_l \setminus \{\mathbf{x}_i\}} dist(\mathbf{x}_i, \mathbf{x}_j)}{|C_l| - 1}, \ \beta(i) = \min_{\ell \in \{1,\ldots,k\} \setminus \{l\}} \frac{\sum_{\mathbf{x}_j \in C_\ell} dist(\mathbf{x}_i, \mathbf{x}_j)}{|C_\ell|}$$

C_ℓ is the ℓ-th cluster, and $|C_\ell|$ is the number of data points in C_ℓ. The silhouette coefficient lies in [-1, 1], and the larger the average silhouette coefficient of the data points it is, the better the clustering is.

For $k \in \{2, \ldots, 7\}$, plot the average silhouette coefficient of the data points and determine the optimal value of k that maximizes the average silhouette coefficient using (a) Euclidean distance and (b) Manhattan distance

Ans: See Figures 2 and 3. You can see that optimal value of k for both cases is 4.

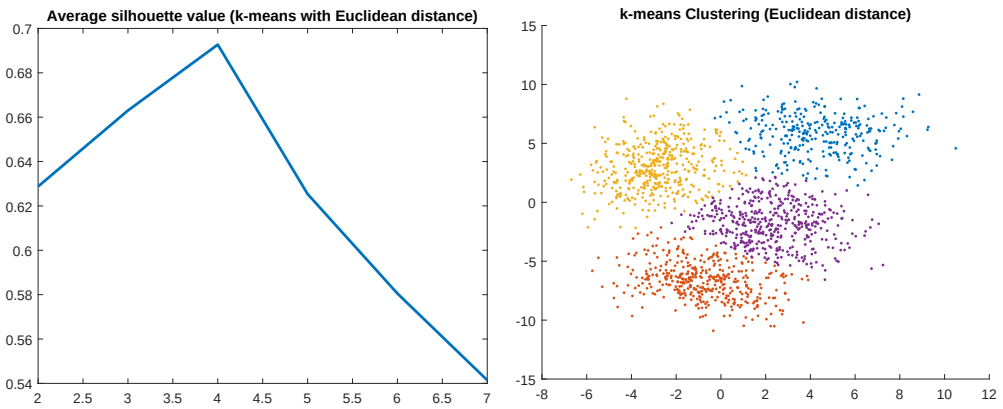

Figure 2: Plot of silhouette coefficient as a function of k and k-means clustering for $k = 4$ (Euclidean distance).

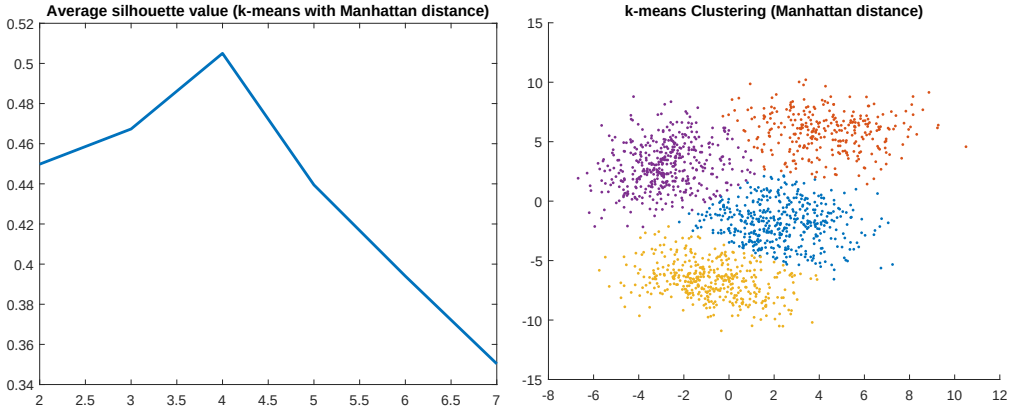

Figure 3: Plot of silhouette coefficient as a function of k and k-means clustering for $k = 4$ (Manhattan distance).

(c) Use the Expectation-Maximization (EM) algorithm to estimate π_k and $\boldsymbol{\mu}_k$, $k \in \{1, 2, 3, 4\}$. Compare the estimated values to the true values provided above. Also, plot the distribution of Gaussian mixture model using the estimated parameters. You can plot either the probability density function or the contour of the density function.

Ans: The contour of the Gaussian mixture distribution along with the average silhouette coefficient (as a function of the number of component distributions) are shown in Figure 4.

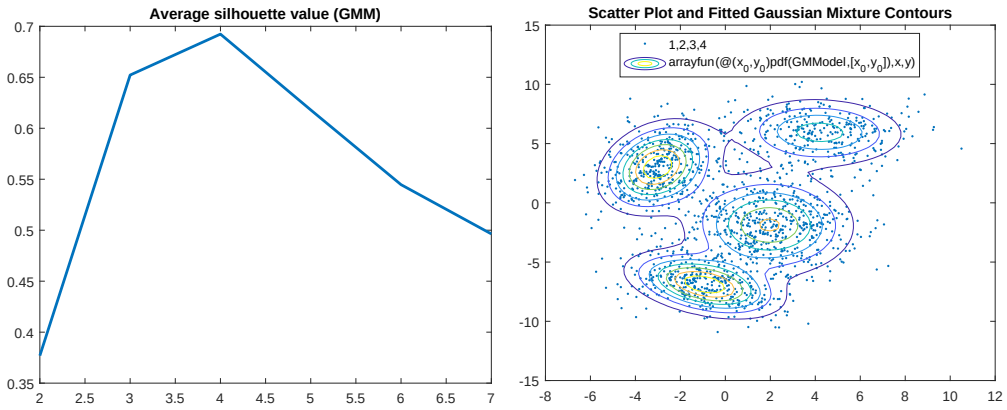

Figure 4: Plot of silhouette coefficient as a function of k and the contour of Gaussian mixture distribution.