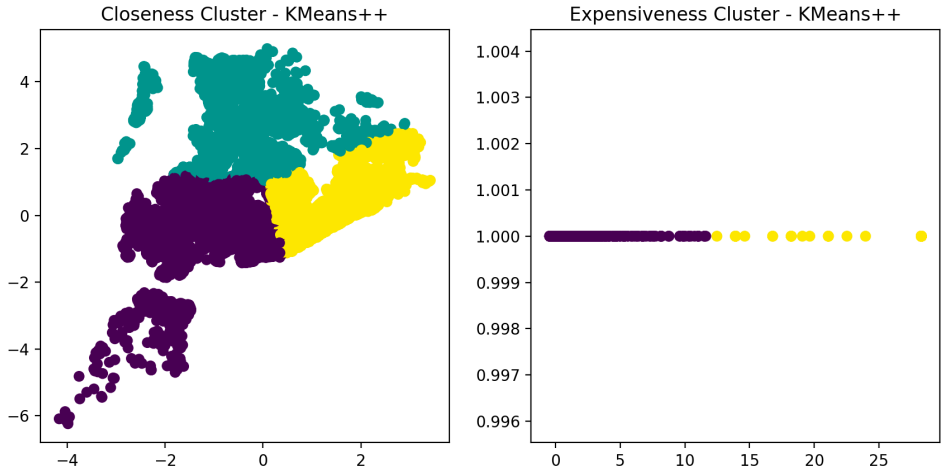
Homework 1 Solutions

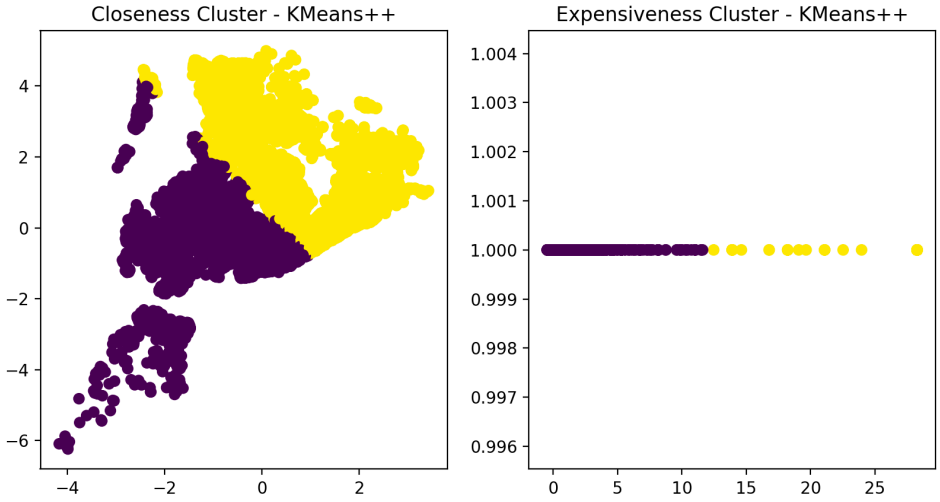
Problem 1:

a)

- **Kmeans++**
  - **Data Representation:**
    - **Set cluster number to be 3 for closeness and 2 for expensiveness**



    - **Set cluster number to be 2 for closeness and 2 for expensiveness**

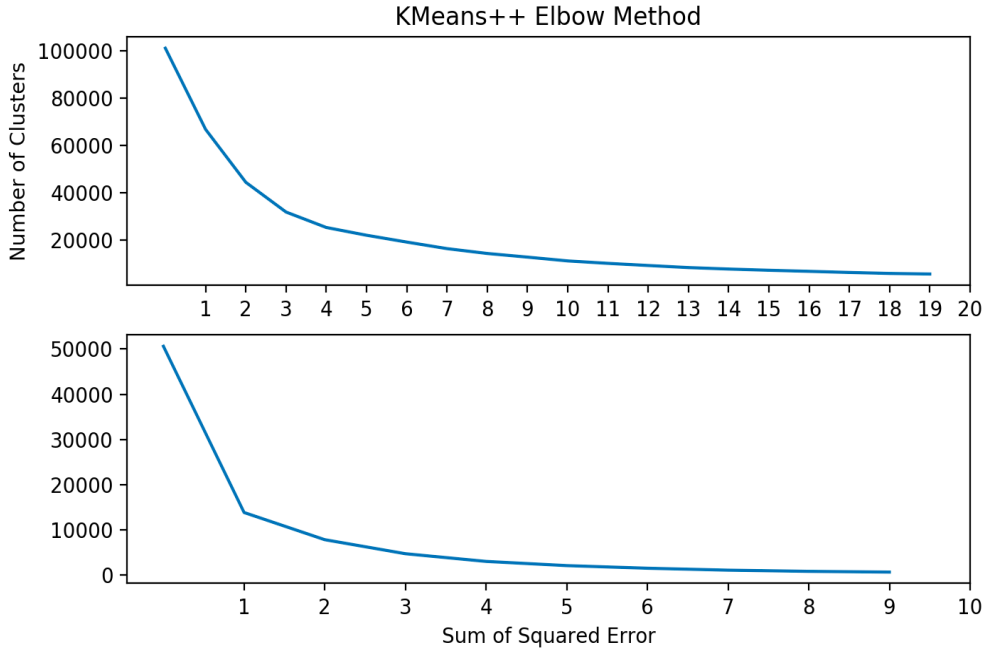


  - **Method of Deciding on the Number of Clusters:**

    Elbow Method

    Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.

    As shown in the figures as below, the approximately optimal choices of cluster number for this specific problem is 3 or 4 for closeness (figure 1) and 2 or 3 for expensiveness (figure 2). All possible parameters should be tried to see which achieves the best performance.



- **Pros & Cons:**

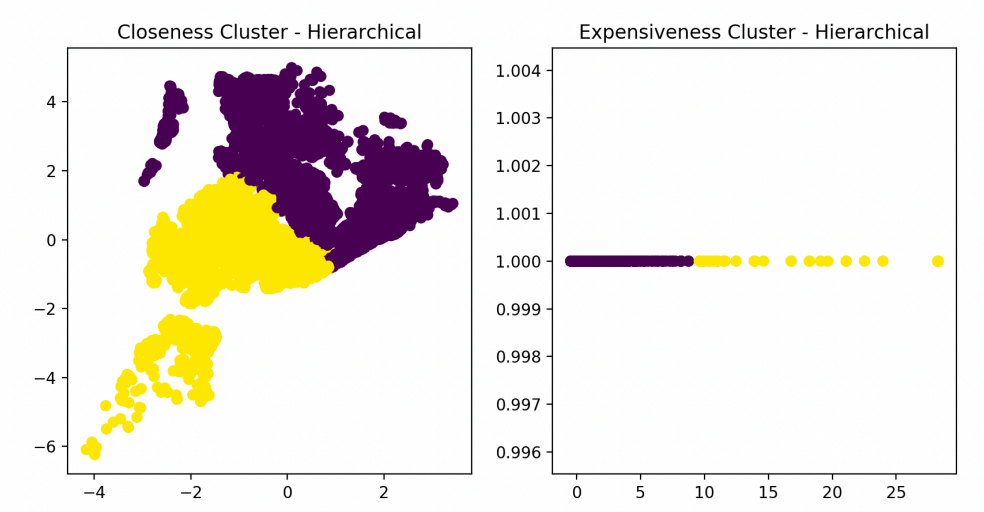
  Pros:

  - Relatively simple to implement.
  - Scales to large data sets.
  - Guarantees convergence.
  - Can warm-start the positions of centroids.
  - Easily adapts to new examples.

  Cons:

  - Needs to choose the number of clusters (k) manually.
  - Depends on initial values of k. Need to run KMeans for several times to test the performance of different values of k.
  - Has trouble in clustering data of varying sizes and density (this problem is a good example of this cons).
  - Has trouble in clustering outliers. Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored
  - Has Curse of Dimensionality with number of dimensions increasing.

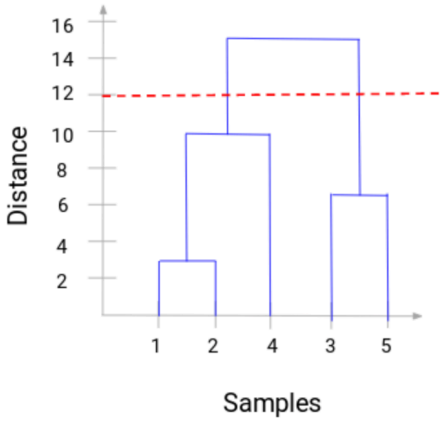
- **Hierarchical**
  - **Data Representation**
    - **Set cluster number to be 2 for closeness and 2 for expensiveness**



- **Method of Deciding on the Number of Clusters**

  Dendrogram

  A dendrogram is a tree-like diagram that records the sequences of merges or splits (See the example in the figure below). Whenever two clusters are merged, we will join them in the dendrogram and the height of the join will be the distance between these points. More the distance of the vertical lines in the dendrogram, more the distance between those clusters. Keep merging until there is only one cluster. We will get a tree-like diagram. The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the set threshold.



- **Pros & Cons**

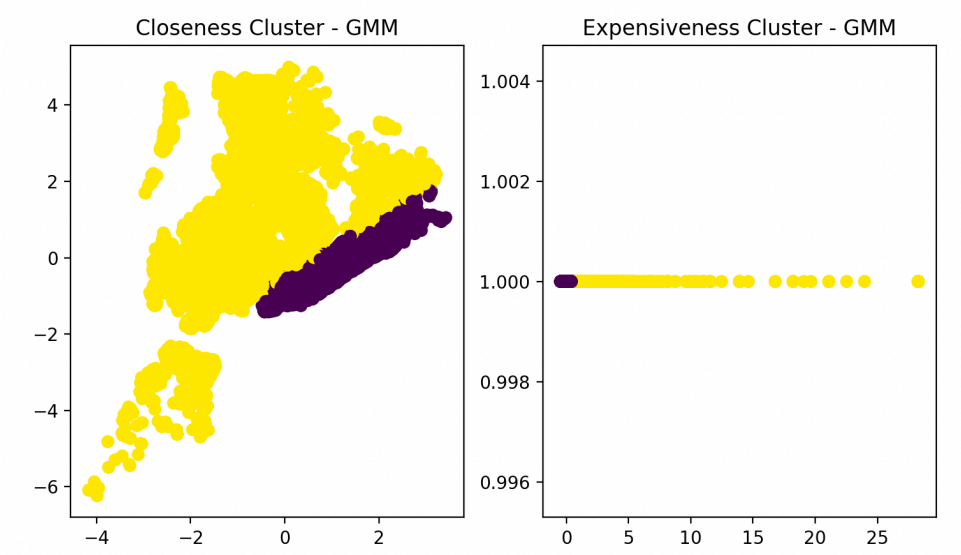
  Pros:

  - Do not have to assume any particular number of clusters. Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level
  - They may correspond to meaningful taxonomies, e.g., shopping websites—electronics (computer, camera, ..), furniture, groceries

  Cons:

  - Sensitivity to noise and outliers
  - Difficult handling different sized clusters and irregular shapes
  - Breaking large clusters

- **GMM**
  - **Data Representation**
    - **Set cluster number to be 2 for closeness and 2 for expensiveness**



  - **Method of Deciding on the Number of Clusters**

    Silhouette Coefficient

    The Silhouette Coefficient is defined for each sample and is composed of two scores:

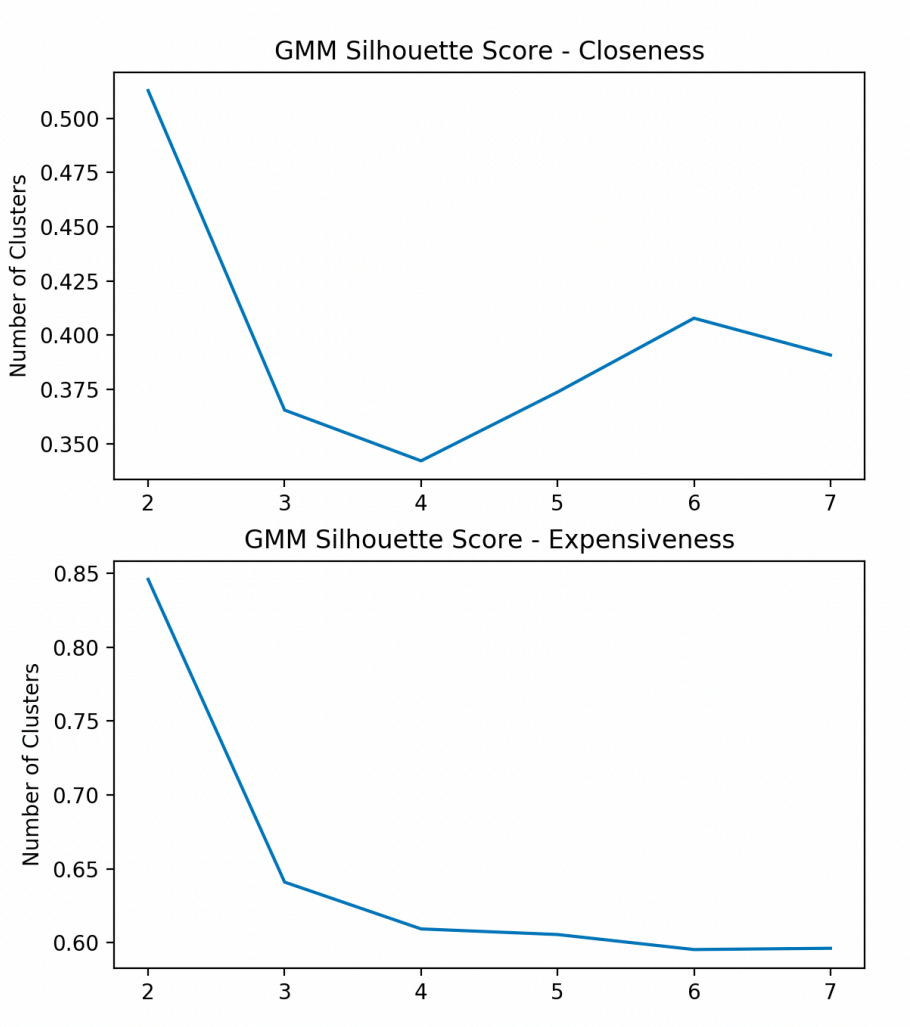
    - a: The mean distance between a sample and all other points in the same class.
    - b: The mean distance between a sample and all other points in the next nearest cluster.

    The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

    A higher Silhouette Coefficient score relates to a model with better defined clusters.

    In this case, 2 clusters performs best for both the cluster problems (See figure below).



- **Pros & Cons**

  Pros:

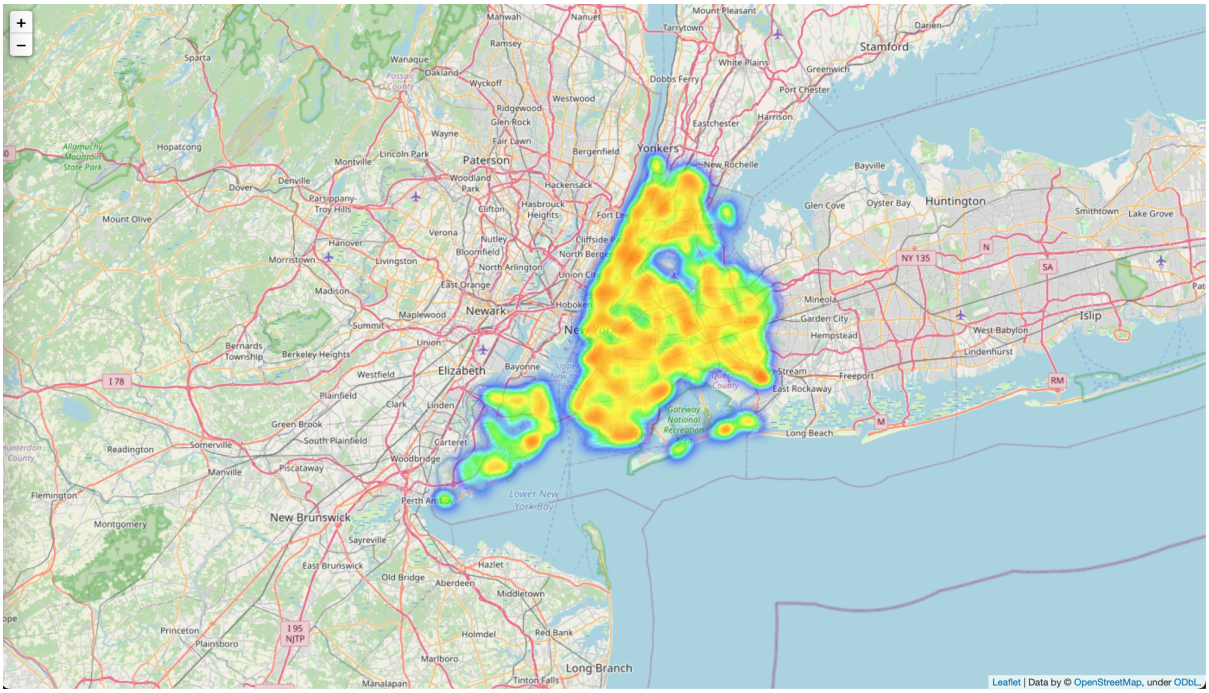
  - Speed: it is the fastest algorithm for learning mixture models.
  - Agnostic: as this algorithm maximizes only the likelihood, it will not bias the means towards zero, or bias the cluster sizes to have specific structures that might or might not apply.

  Cons:

  - Singularities: when one has insufficiently many points per mixture, estimating the covariance matrices becomes difficult, and the algorithm is known to diverge and find solutions with infinite likelihood unless one regularizes the covariances artificially.
  - Number of components: this algorithm will always use all the components it has access to, needing held-out data or information theoretical criteria to decide how many components to use in the absence of external cues.

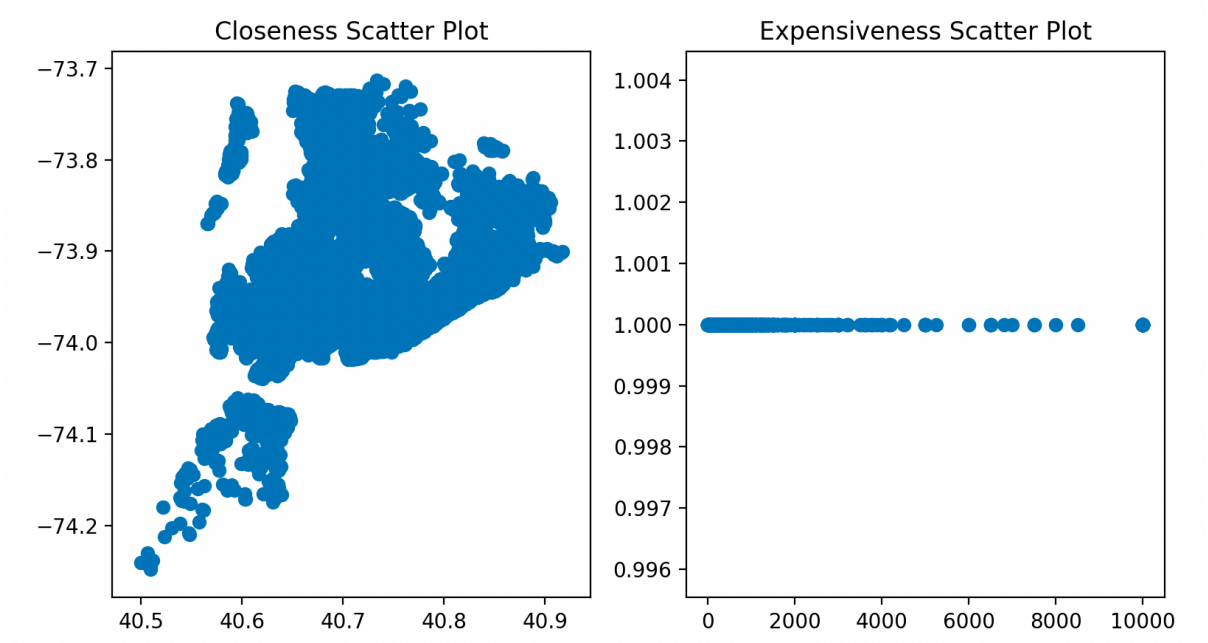
Problem 2:

a) e) Heatmap



Heapmap gives a better indication of the expensiveness of areas within NYC by coloring the areas from warm to cold colors with the corresponding price dropping down. This finding agrees with the actual cost of living in NYC as it's more expensive to live in the urban areas especially the centers of such areas.

b) Scatter Plot



Problem 3:

Original image:

Manipulated image: