CS506 HW1

1. Working with the Algorithms
   a) Explain your data representation and how you determined certain parameters.
      a. K-Means++ Clustering
         - In K-Means++ Clustering.ipynb file, I added comments for each step and also showed what variables it accepts and what results it returns for each function.
         - There is a popular method known as elbow method which is used to determine the optimal value of k to perform the K-Means++ Clustering Algorithm. I firstly used cost function to calculate the sum of squared distances of samples to their closest cluster center. Then I made a plot which plots the various values of cost with changing k. From the plot, we can see that the elbow is clearly forming at K=5, so I think the optimal value of k in K-Means++ Clustering will be 3.
         - There is a dictionary representing each data point belongs to its corresponding cluster. There is also a plot visualizing the clusters in 3D space.

         Note: since the time taken is long if running on the entire dataset, I just run the algorithm on the top 10000 rows.

      b. Hierarchical Clustering
         - In Hierarchical Clustering.ipynb file, I added comments for each step.
         - Hierarchical Clustering does not require to determine the optimal value of k manually as it can be determined by dendrogram. After deleting listings that contain missing values and scaling the dataset, I created a dendrogram to know the clusters that we want the data to be split to. Once one big cluster is formed, the longest vertical distance without any horizontal line passing through it is selected and a horizontal line is drawn through it. The number of vertical lines this newly created horizontal line passes is equal to number of clusters, therefore the number of clusters will be 2 or 3.
         - There is a dictionary representing each data point belongs to its corresponding cluster. There is also a plot visualizing the clusters in 3D space.

         Note: since the time taken is long if running on the entire dataset, I just run the algorithm on the top 8000 rows.

      c. GMM Clustering
         - In GMM Clustering.ipynb file, I added comments for each step.
         - AIC, BIC can be used to determine the optimal value of k to perform the GMM Clustering Algorithm. The optimal number of clusters is the value that minimizes the AIC or BIC. Therefore, I made a plot which plots the various AIC

and BIC with changing k. From the plot, we can see that AIC and BIC starts to increase again at k = 12 or k = 13, so I think the optimal value of k would be around 15.

- There is a dictionary representing each data point belongs to its corresponding cluster. There is also a plot visualizing the clusters in 3D space.

Note: for this algorithm, I run the algorithm on the entire dataset.

b) List the pros and cons of the various clustering algorithms.

a. K-Means++ Clustering

Pros:

- Relatively simple to implement
- Scales to large datasets
- Guarantees convergence

Cons:

- Being dependent on initial values
- Sensitivity to different sizes and density
- Sensitivity to noise and outliers

b. Hierarchical Clustering

Pros:

- Easy to implement
- No prior information about the number of clusters required
- Outputs a structure that is more informative

Cons:

- Algorithm can never undo what was done previously
- Not suitable for large datasets
- Sensitivity to noise and outliers

c. GMM Clustering

Pros:

- More flexible in terms of cluster covariance
- Will not bias the means towards zero, or bias the cluster sizes

Cons:

- Long computation time
- Need to decide the number of clusters

2. Data Visualization

a) This heatmap is useful but not enough to draw conclusions about the expressiveness of areas within NYC because the dark and light color area could represent that the expensiveness within that area but it is too specific and could not have a broad perspective to see which region is more expensive.

b, c) Code in Data Visualization.ipynb

e) The findings are in agreement with what I have in mind about the cost of living for neighborhoods in NYC because by looking at the plot and the average price within certain clusters we can see that the most expensive area is Manhattan, and the least expensive area is Queens.

3. Image Manipulation
   Code in Image Manipulation.ipynb