ASSIGNMENT – 2

Harry Preet Singh – 301368450   
CMPT 459, SPRING 2022

TASKS:

1. This task has been completed in kmeans.py file provided during submission
2. For the K-Means clustering using Random Initialization, I had high disparity in the silhouette scores every time I ran the fitting function, hence I averaged the silhouette scores from 5 runs for each value of k and have reported that here.  
     
   **Best K =4**  
   Silhouette Scores Values for k ranging from 2 – 9:  
   Text

   Description automatically generated  
     
   Plotting the Silhouette Scores:  
   Chart, line chart

   Description automatically generated  
     
   **Here I found k = 4 as best and with the highest silhouette score of 0.204**. The elbow point occurs after k = 5 hence due to random fluctuations, future runs might have k = 5 as best k
3. Same as before, I averaged the silhouette scores from 5 runs for each k for K-means ++ and have reported that here.   
     
   **Best K = 2**  
   Silhouette Scores Values for k ranging from 2 – 9:  
   Text

   Description automatically generated  
     
   Plotting the Silhouette Scores:  
   Chart, line chart

   Description automatically generated  
   **Here I found k = 2 as best and with the highest silhouette score of 0.28**. The elbow point occurs after k = 3.  
     
   **Conclusion:**
   * The silhouette score for K-means ++ is higher for best k than in the random initialization, hence we can conclude that the K-means ++ does seem optimal.
   * Further, I found that the running time for k means ++ clustering after initialization was way faster than the random initialization and hence proving that k means ++ converges faster
   * For future analysis, maybe we can calculate the error for both the clustering algorithms to access further insight into these conclusions.
   * Lastly, though k means ++ does perform better, the silhouette score itself isn’t that impressive as it is less than 0.4 letting me to conclude that the clustering seems not so reasonable.
4. **Scatter Plot Visualizing the clusters with the best k = 4 from task 2, the PCAs are dimensionally reduced from 100 to 2 to simplify plotting.**Chart, scatter chart

   Description automatically generated
   * Here we can see the 4 different clusters, with the red cluster being having a definite boundary.
   * Green and Blue clusters seem to share their boundary with each other, and the clustering seems mixed
   * Yellow cluster seems to be spread on both green and blue clusters and further only has few values compared to rest, implying it might consist of outliers.