Name (netid): Your Name (Your Netid)
CS 441 - HW 4: Dealing with Data

Complete the claimed points and sections below.

Total Points Claimed **[ ] / 142**

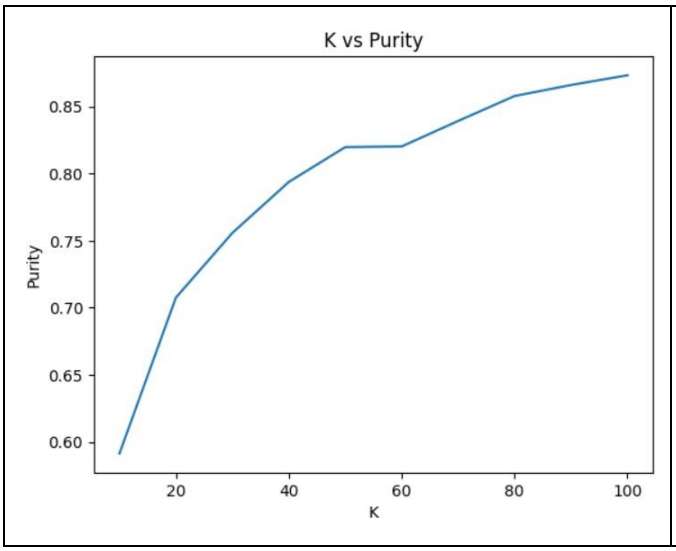
1. Clustering and Fast Retrieval
    a. Test Kmeans Purity & Centroids [ ] / 15
    b. Questions [ ] / 10
    c. Fast 1-NN Retrieval [ ] / 15
2. Estimating PDFs
    a. Histograms [ ] / 10
    b. Clustering [ ] / 10
    c. Gaussian Mixture Model [ ] / 15
3. PCA and Data Compression
    a. Display Principal Components [ ] / 5
    b. Scatter Plot [ ] / 5
    c. Plot cumulative explained variance [ ] / 5
    d. Time & Accuracy [ ] / 10
4. Stretch Goals
    a. Rotate Using PCA and comparison To original approach [ ] / 15
    b. Try Part 2 with your own images [ ] / 10
    c. Plot Using t-SNE and MDS [ ] / 15
    d. Completed HW3 survey by DATE [ ] / 2

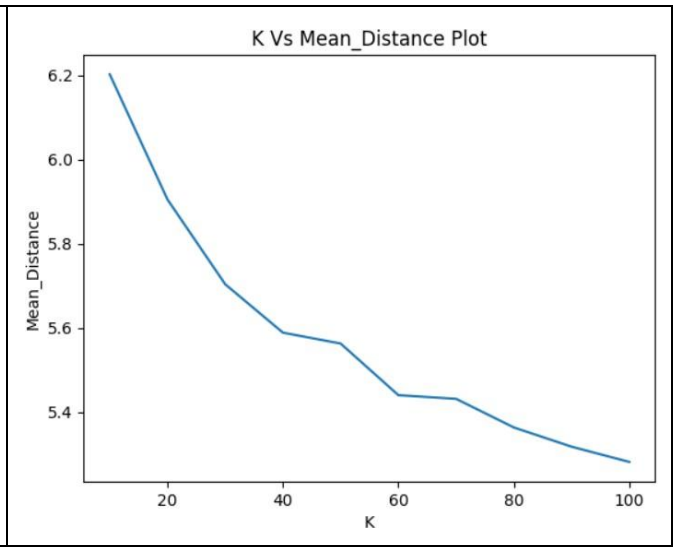
1. Clustering & Fast Retrieval

a. Test KMeans Purity

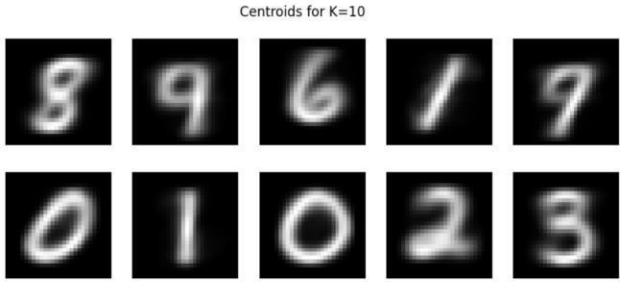
[15]

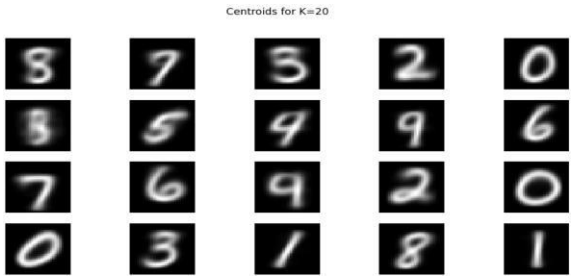
K Vs Purity Plot	K Vs Mean_Distance Plot

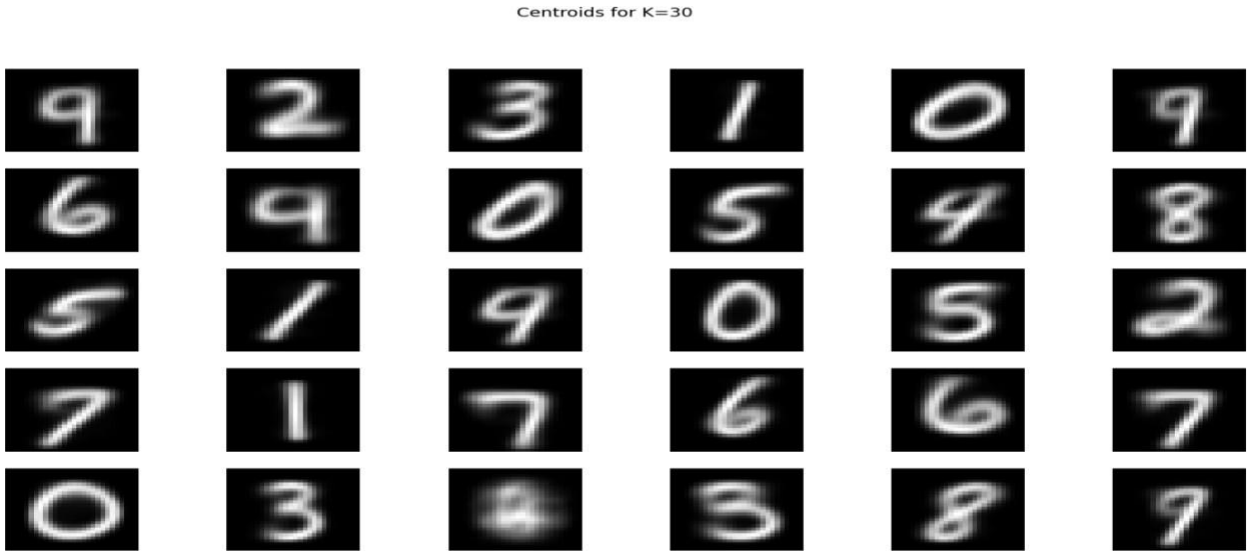




Paste images of centroids for K = 10, K = 20, and K=30 below (three rows).







b. **Questions** [10]

i. As you increase K, do you expect the purity to increase? Why or why not?
ii. In a given run, is the average distance of a sample to centroid guaranteed to monotonically decrease with each iteration (i.e., cannot increase)? Why or why not?
iii. If you do enough iterations, is Kmeans guaranteed to give you the optimal clustering that minimizes the sum of distances between each sample and its center? Why or why not?
iv. Does improving the Kmeans objective (i.e., achieving lower mean squared error) necessarily improve expected purity? Why or why not?

c. Solution

i. As you increase the K, the purity may or may not increase, depending on the data and the structure of the clusters. Increasing K means increasing the number of clusters and if the data has distinct and well separated clusters, then increasing k may lead to high purity as the cluster become more specialized. However, if the data has overlapping clusters or does not have distinct clusters, increasing k may lead to overfitting and lower purity.
ii. No, the given run, the average distance of a sample to centroid is not guaranteed to monotonically decrease with each iteration because the k means algorithm is sensitive to initialization and can coverage to a suboptimal solution depending on the initial position of the centroids.
iii. No, k means is not guaranteed to give optimal clustering that minimize the sum of distances between each sample and its center, even if you do enough iterations. K means is an iterative algorithm that uses random initialization of centroids, so it can sometimes stick in local optima that may not be a global optimum. In other words, we can say the final clustering depends on the random seed, and different initialization may produce different clustering results. Therefore, it is possible that k means may converge to a suboptimal solution instead of the global optimum.
iv. No, it does not necessarily improve expected purity because the kmeans algorithm optimizes for compact and well-separated clusters, but it does not explicitly optimize for cluster purity. In other words, two clusters may be well separated and compact but still contain a mix of different class labels, resulting in low clustering purity. Therefore, while minimizing the sum of squared distances can improve the clustering quality in terms of compactness and separation, it does not guarantee high purity.

d. Fast Retrievals
i. Brute Force: [5]

Test Error	Time to Add	Time to Search
0.0059	0.4727 seconds	16.5441 seconds

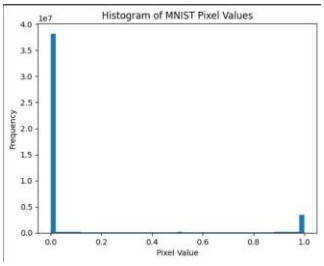
ii. LSH: [5]

Test Error	Time to Add	Time to Search	Nbits parameter
0.6079	0.24113512	2.48537611	256

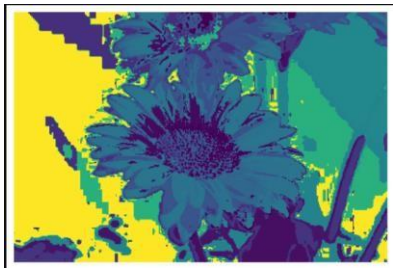
2. Estimating PDFs

Include the generated images (score map, thresholded score map, thresholded RGB) from the display code.

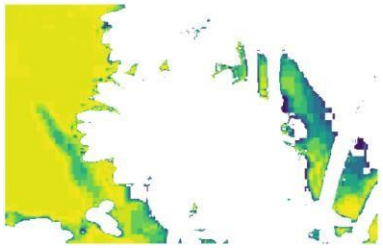
a. Histogram: [10]



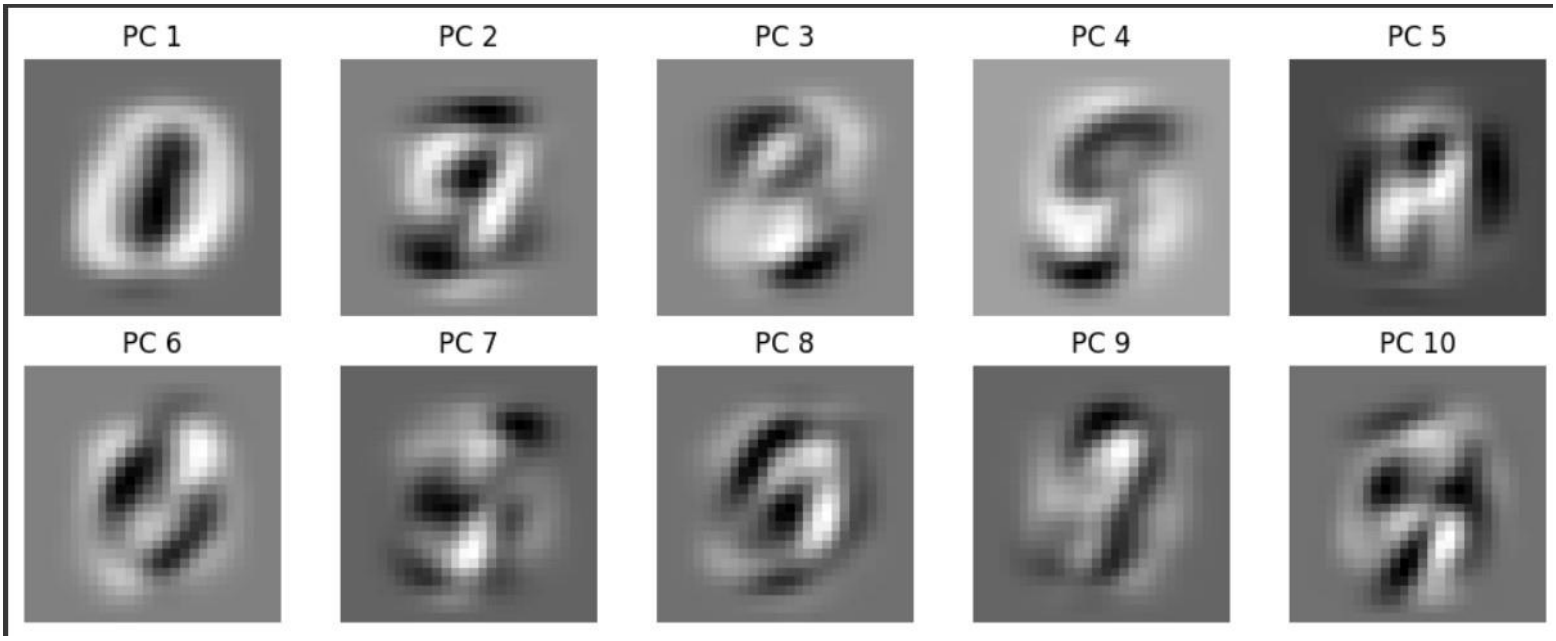
b. Clustering:



c. Gaussian mixture Model: [15]



Visualization



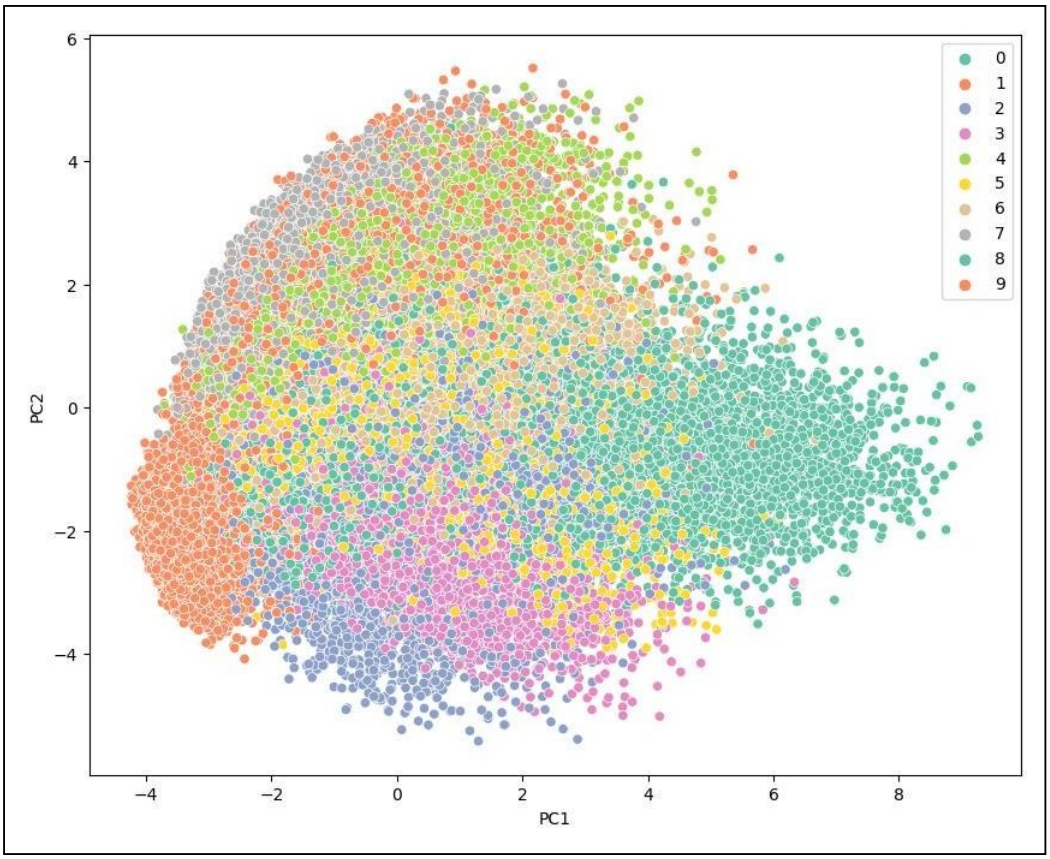
3. PCA and Data Compression

1. First 10 principal components [5]

   Plotted above in Visualization table--

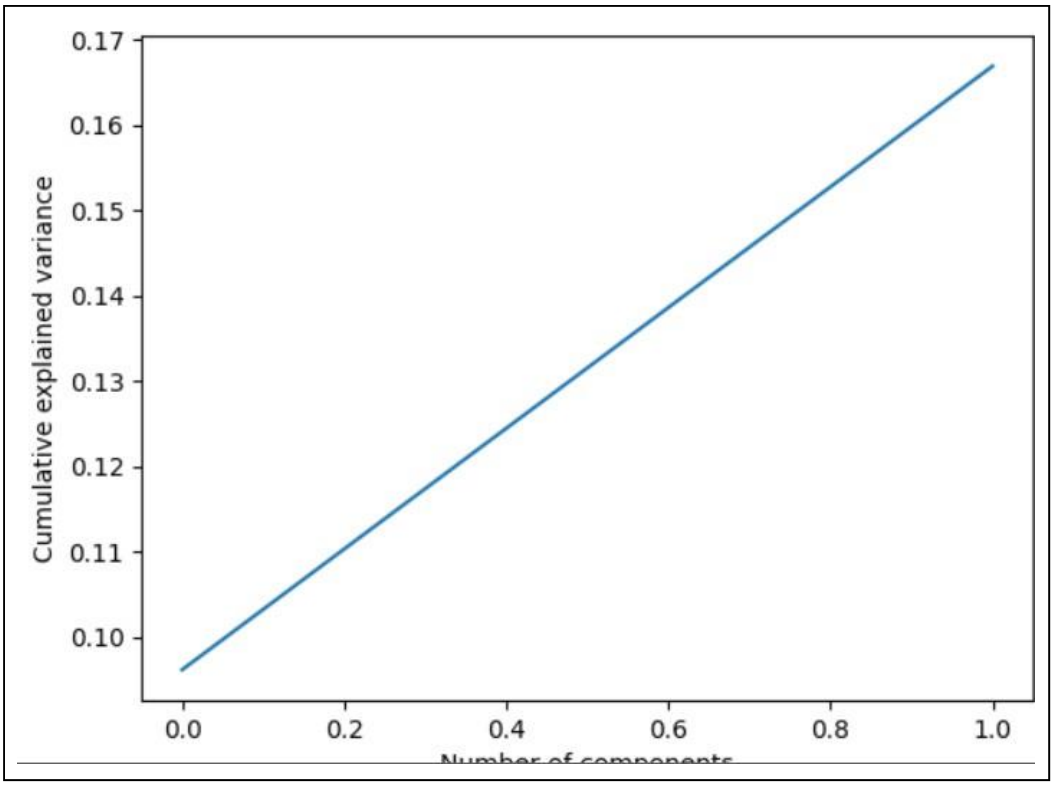
2. Scatterplot [5]

PLOT



3. Cumulative explained Variance [5]

PLOT



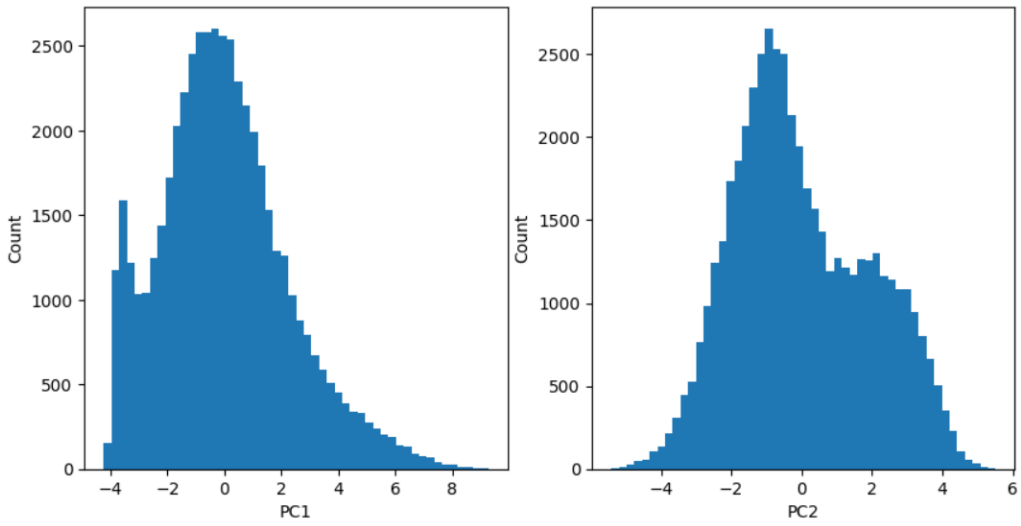
4. Faiss [10]

	Total Time	Test Error	Dimensions
Brute Force (PCA)	0.298798	16.63%	768
Brute Force	17.45316 sec	18.48%	768
LSH	2.54050326	69.29%	768

Note: the last two rows are copied from 1.c for reference.

4. Stretch Goals

a. PDFS after using PCA to rotate your data [15]



b. Apply Part 2 to your own choice of image, with the same deliverables [10]

Applied using Flower image –

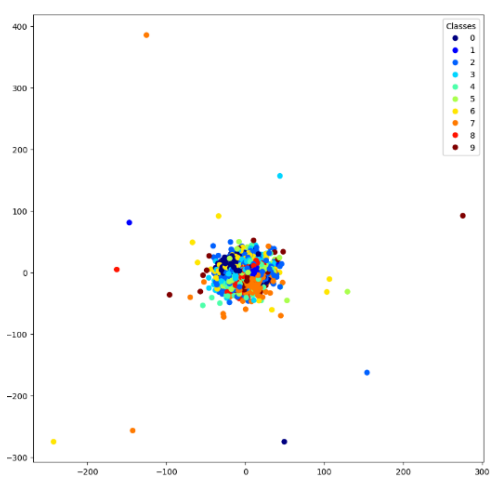
Fig1:

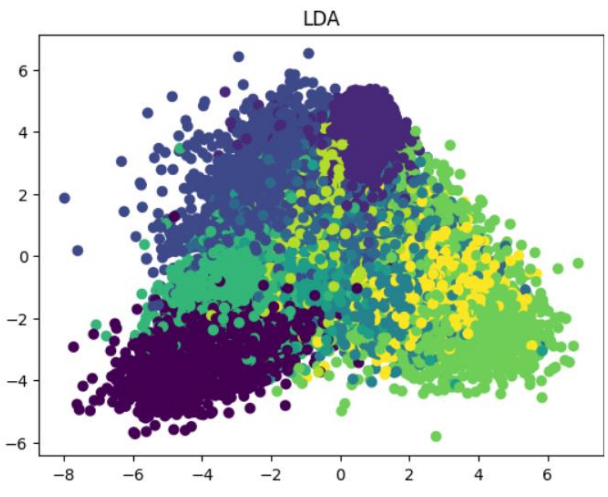
https://drive.google.com/file/d/1HId1aDP5Hg3hleSJH3RJobs5pp8gHDTL/view?usp=sharing [Cropped image]

Fig 2:

https://drive.google.com/file/d/1_mAdoAzaUuA5o_K2FYh5InU3hYZ_3crQ/view?usp=sharing [Original Image]

c. Scatterplots using at least two of t-SNE, MDS, and Linear Discriminant Analysis [15]
(Plotted using MDS and Linear Discriminant Analysis)





Acknowledgments / Attribution

The code and understanding of MDS algorithm and statistical analysis guide taken from--

Multi-Dimensional Scaling--

“Multidimensional scaling by optimizing goodness of fit to a numeric hypothesis”Kruskal, J. Psychometrika, 29, (1964).

sklearn.manifold.MDS — scikit-learn 1.2.2 documentation

Linear Discriminant Analysis---

sklearn. discriminant_analysis.LinearDiscriminantAnalysis — scikit-learn 1.2.2 documentation

The Statistical analysis was guided by the textbook “Introduction to Statistically Learning” by Gareth James et al. (Springer, 2013).