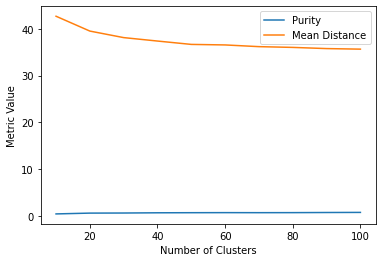
|  |  |  |
| --- | --- | --- |
| **Name (netid):** Your Name (Your Netid)  **CS 441 - HW 4: Dealing with Data**  Complete the claimed points and sections below. |  | |
| **Total Points Claimed** | **[** | **] / 142** |
| 1. Clustering and Fast Retrieval    1. Test Kmeans Purity & Centroids | [ | ] / 15 |
| b. Questions | [ | ] / 10 |
| c. Fast 1-NN Retrieval | [ | ] / 15 |
| 1. Estimating PDFs    1. Histograms | [ | ] / 10 |
| b. Clustering | [ | ] / 10 |
| c. Gaussian Mixture Model | [ | ] / 15 |
| 1. PCA and Data Compression    1. Display Principal Components | [ | ] / 5 |
| b. Scatter Plot | [ | ] / 5 |
| c. Plot cumulative explained variance | [ | ] / 5 |
| d. Time & Accuracy   1. Stretch Goals    1. Rotate Using PCA and comparison To original approach | [  [ | ] / 10  ] / 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 |

## Clustering & Fast Retrieval

* 1. Test KMeans Purity [15]

K Vs Mean\_Distance Plot

K Vs Purity Plot



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

* 1. Questions [10]
     1. As you increase K, do you expect the purity to increase? Why or why not?

**Increasing K in k-means clustering may or may not increase the purity of the clusters. It depends on the data and the structure of the clusters. If the clusters are well-separated and distinct, increasing K may improve the purity as it allows the algorithm to capture more subtle variations in the data. However, if the clusters are overlapping or poorly separated, increasing K may lead to fragmentation of the clusters and decrease the purity.**

* + 1. 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?

**No, the average distance of a sample to the centroid is not guaranteed to monotonically decrease with each iteration, as k-means is sensitive to the initial centroid positions and can converge to local optima.**

* + 1. 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?

**No, Kmeans is not guaranteed to give the optimal clustering that minimizes the sum of distances between each sample and its center, even after enough iterations, because it can get stuck in local optima. The algorithm's final result depends on the initial choice of centroids and can converge to a suboptimal solution.**

* + 1. Does improving the Kmeans objective (i.e. achieving lower mean squared error) necessarily improve expected purity ? Why or why not?

**Improving the Kmeans objective does not necessarily improve the expected purity, because the objective only focuses on minimizing the sum of distances between each sample and its center, without taking into account the distribution of the samples across the clusters. Therefore, it is possible to have a low objective value, but with clusters that have a high degree of overlap and low purity.**

* 1. Fast Retrievals
     1. Brute Force: [5]

|  |  |  |
| --- | --- | --- |
| Test Error | Time to Add | Time to Search |
| 0.0089 | 0.1147sZ | 4.2405s |

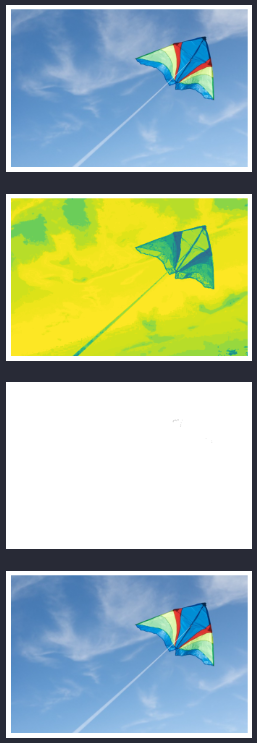
* + 1. LSH: [5]

|  |  |  |  |
| --- | --- | --- | --- |
| Test Error | Time to Add | Time to Search | Nbits parameter |
| 0.3943 | 0.1554s | 0.5845s | 8 |

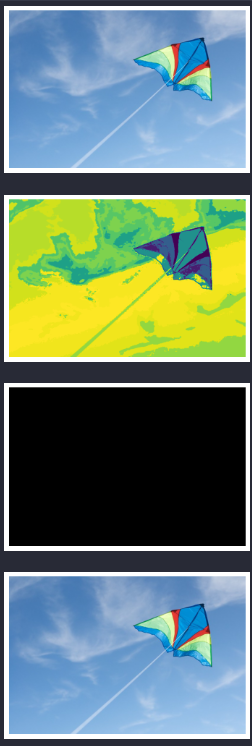
## Estimating PDFs

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

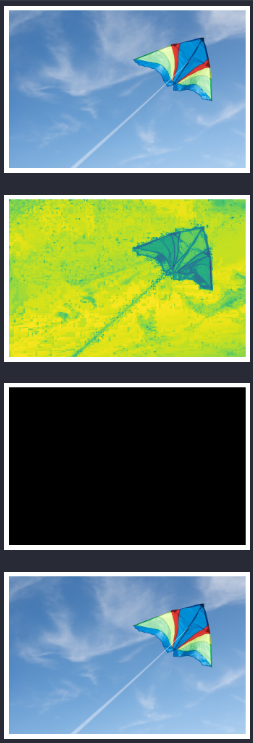
* 1. Histogram: [10]



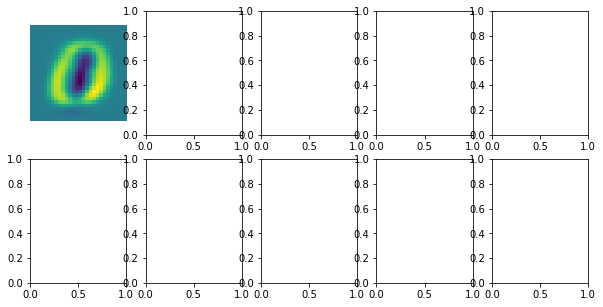
* 1. Clustering: [10]



* 1. Gaussian mixture Model: [15]

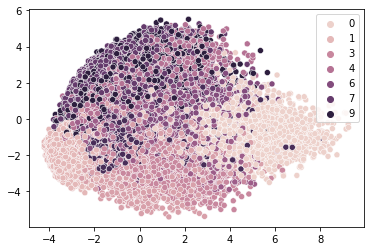


1. **PCA and Data Compression**
2. First 10 principal components [5]



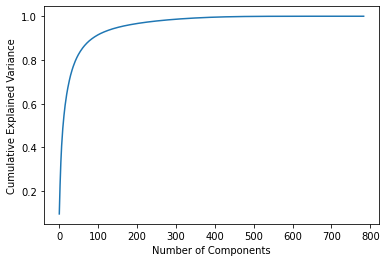
Visualization

1. Scatterplot [5]



PLOT

1. Cumulative explained Variance [5]



PLOT

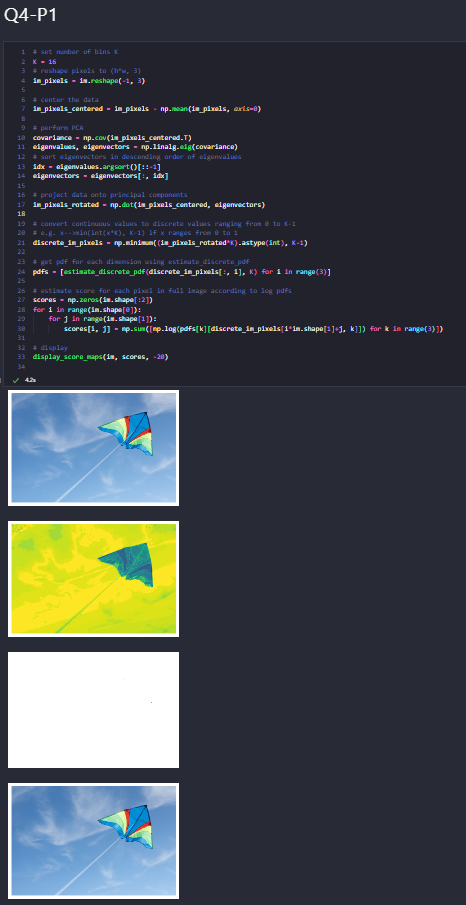
1. Faiss [10]

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total Time | Test Error | Dimensions |
| Brute Force (PCA) | 20.55 | 0.0331 | 784 |
| Brute Force | 0.1147 | 0.0089 | 784 |
| LSH | 0.1554 | 0.3943 | 784 |

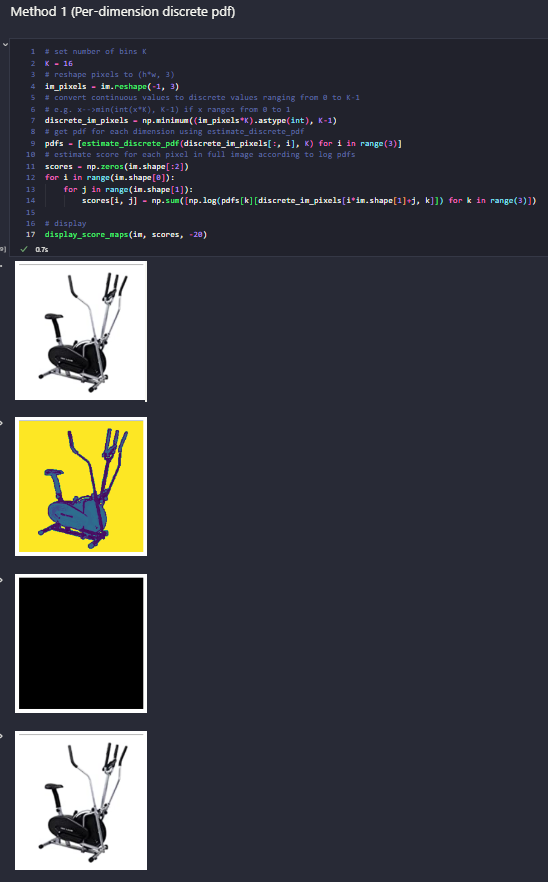
Note: the last two rows are copied from 1.c for reference.

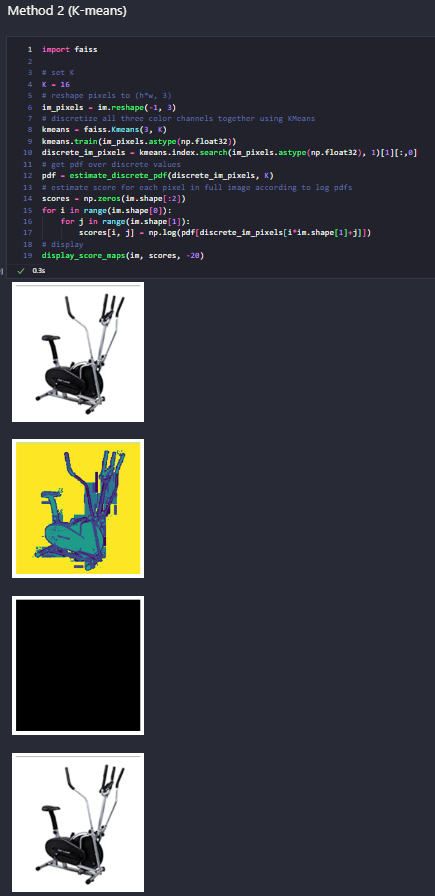
# Stretch Goals

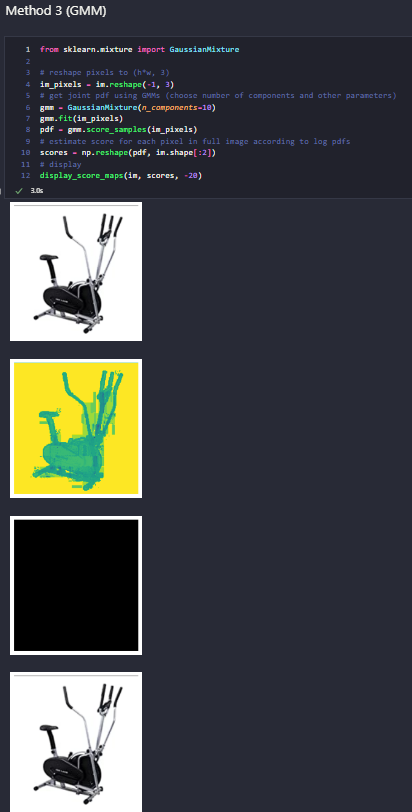
* 1. PDFS after using PCA to rotate your data [15]



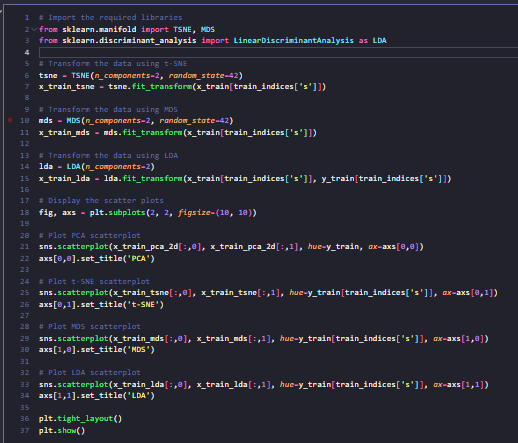
* 1. Apply Part 2 to your own choice of image, with the same deliverables [10]

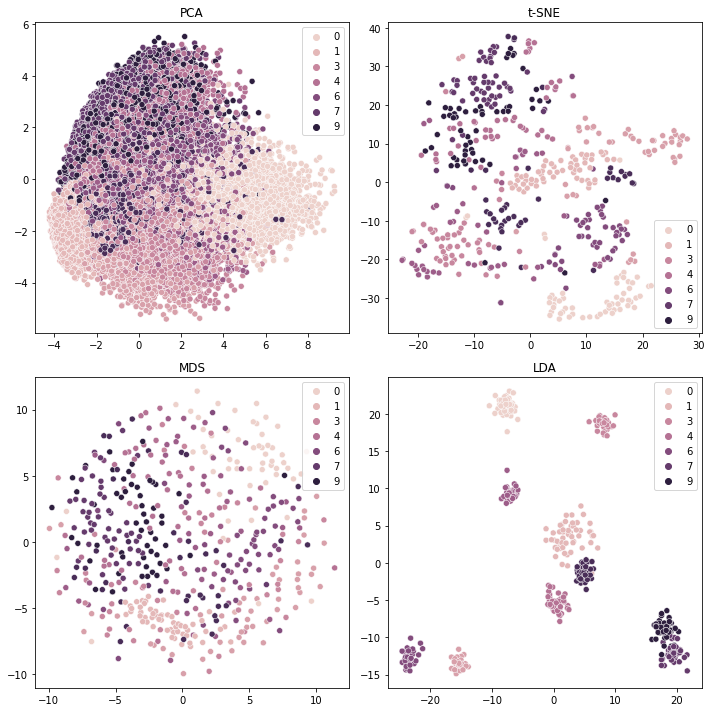
­





* 1. Scatterplots using at least two of t-SNE, MDS, and Linear Discriminant Analysis [15]





# Acknowledgments / Attribution

“None”.