CSE 416: Introduction to Machine Learning
Assignment #7
Instructor: Pemi Nguyen
due: Friday, May 20, 2022, 11:59 pm.

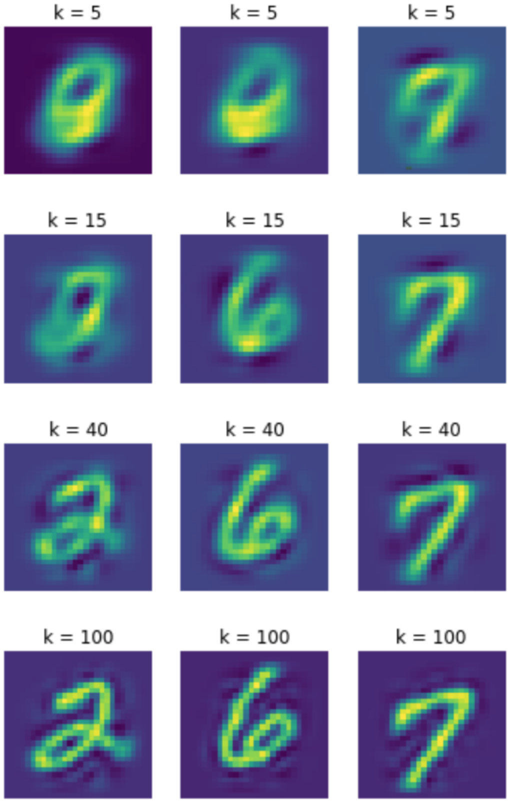
Instructions:

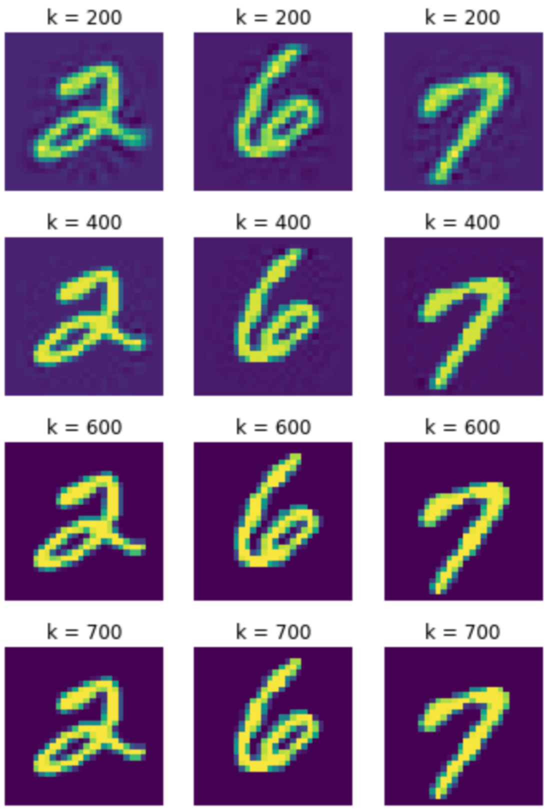
- **Answers:** Please provide detailed yet concise explanations for your work.

- **Turn-in:** Do not write your name on your pages (your Gradescope account will identify you to us) and do not include a copy of the exercise's question in what you turn in. You must use Gradescope to upload your written homework solutions. You will submit a single PDF file containing your solutions to all the exercises in the homework. You must follow the Gradescope prompts that have you link exercise numbers to your pages. **We do not accept handwritten solutions**, so please typeset them, such as Microsoft Word or LaTeX editor.

1. We're going to implement PCA reconstruction on the MNIST dataset. Here are the reconstructions of a few digits with different values of k, the number of principal components used.

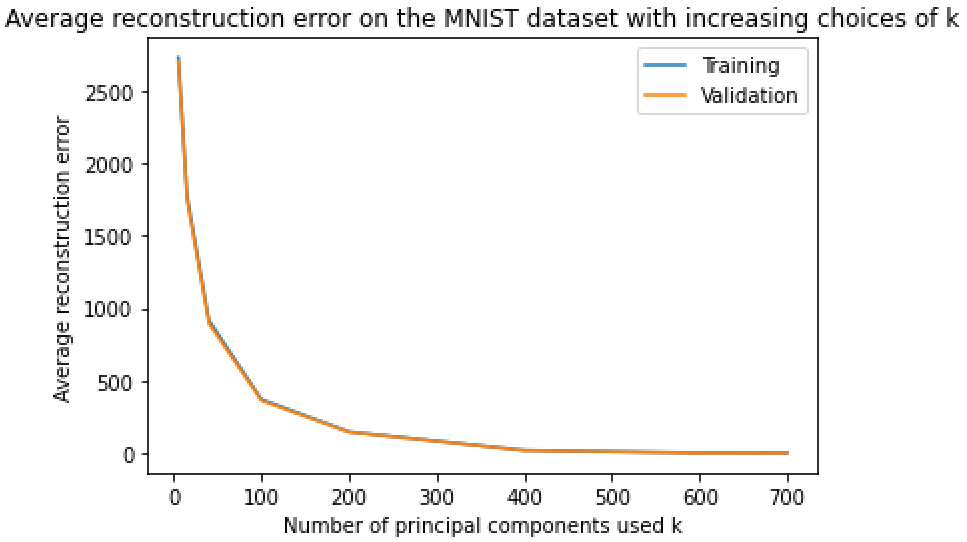
   Here is the Colab link of the source code if you're interested.





(a) *(2 points)* In general, what are the main goals of dimensionality reduction for PCA?

(b) *(2 points)* What can you say about the quality of the reconstructed images as we increase k? How is it related to the average reconstruction errors over increasing choices of k from the below plot:



(c) *(2 points)* Relating to the purposes of dimensionality reduction for PCA, which choice of k seem to be the best if we want to reconstruct images to improve the performance of an image classification model? (Note: There might not be a wrong answer, as long as you're able to defend your choice).

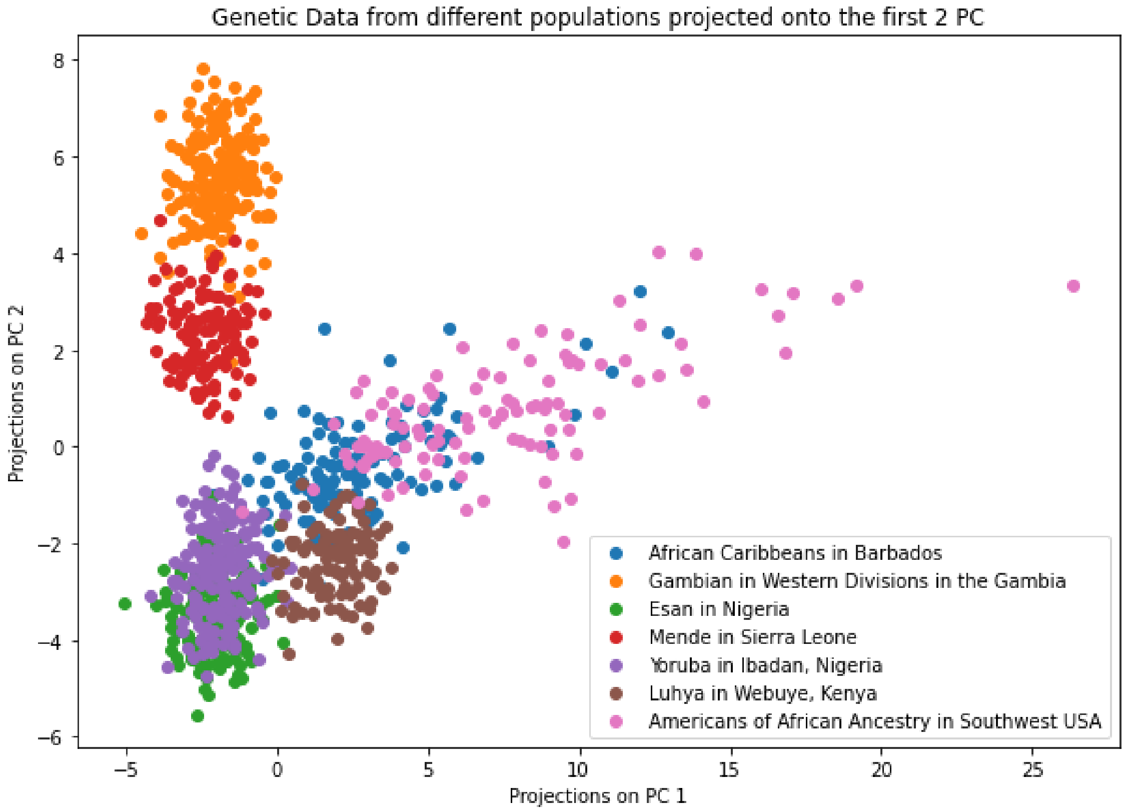
2. One interesting property of PCA is that the first few principal components tend to have interpretable meanings.

   The file pca-data.txt contains data from the 100 genomes project. Each of the lines in the file represents an individual. The first three columns contain: an individual's unique identifiers, his or her sex (1=male, 2=female), and the population to which they belong. (See the encodings here.) The subsequent columns of each line are a sample of the nucleobases from that individual's genome.

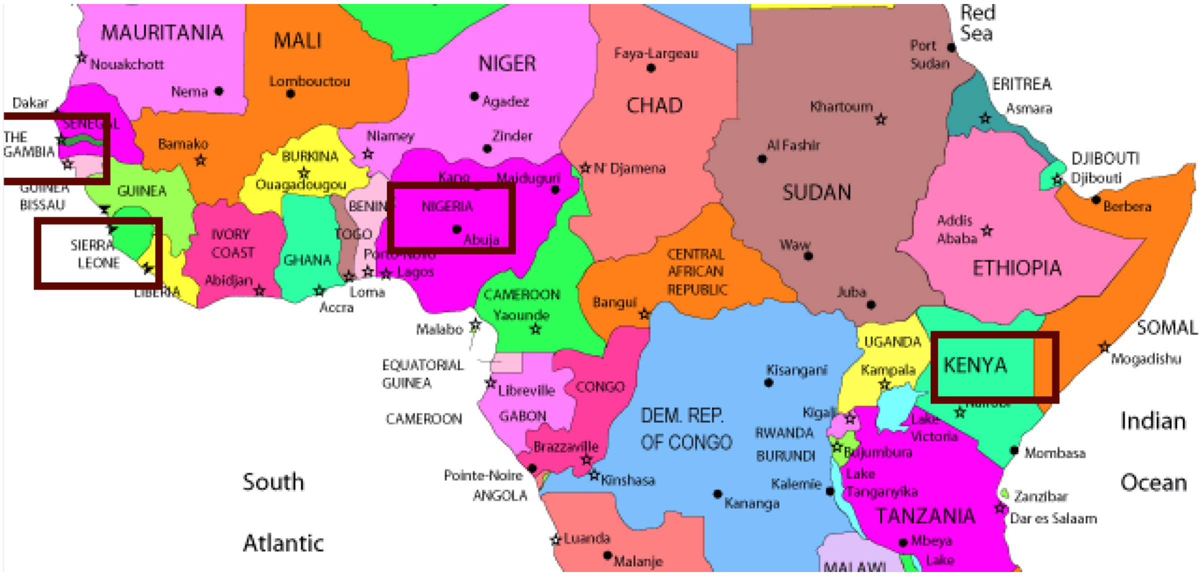
   For each column of nuclobases (A,C,G,T), compute the nuclobase that occurs most frequently. Call that the column-j mode. Convert the genetic data into a binary matrix X such that $X_{ij} = 0$ if person i has the column-j mode in position j and $X_{ij} = 1$ otherwise. The 1 entries correspond to mutations.

   (a) *(3 points)* Let's examine the first 2 principal components of X. These components contain lots of information about our data set. We create a scatter plot with each of the 995 rows of X projected onto the first two principal components, with each color representing a certain population group.

   From the plot below, suppose the possible interpretations of the first two principal components are that they refer to **the latitude and longitude of the origins of genetic data**. In your opinion, which principal component seems to refer to which meaning, and why?

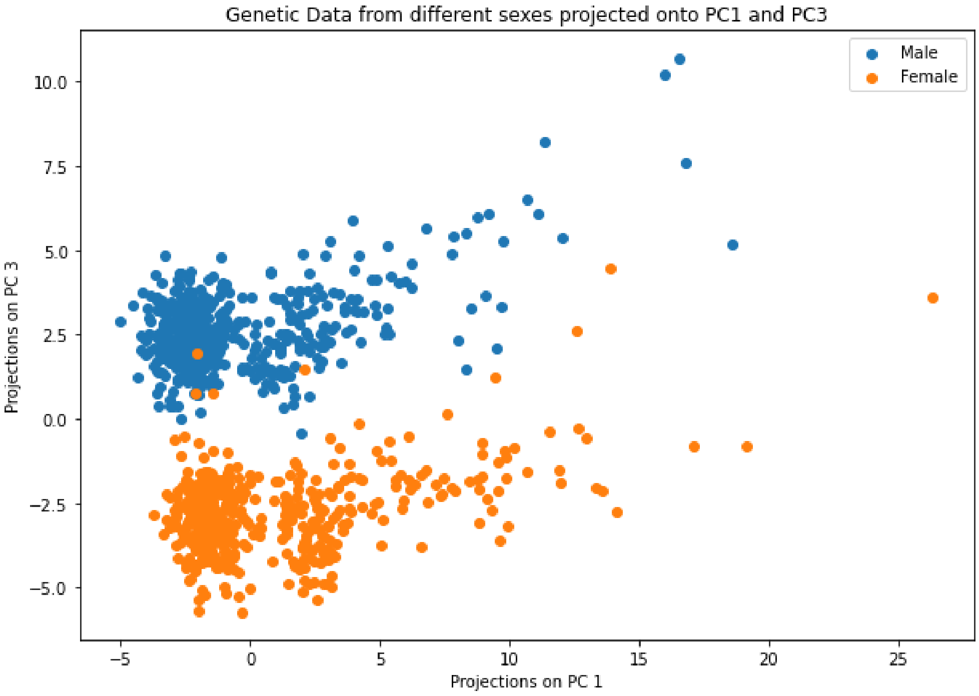


To allow you to visualize the locations of African countries on a map, here is a map where some African countries from the above populations are highlighted.



(b) *(1 points)* It is supposed that the third principal component seems to indicate the **sex**

of an individual. Based on the below plot with each individual projected onto the subspace spanned by the first and third principal components, do you agree? Why?

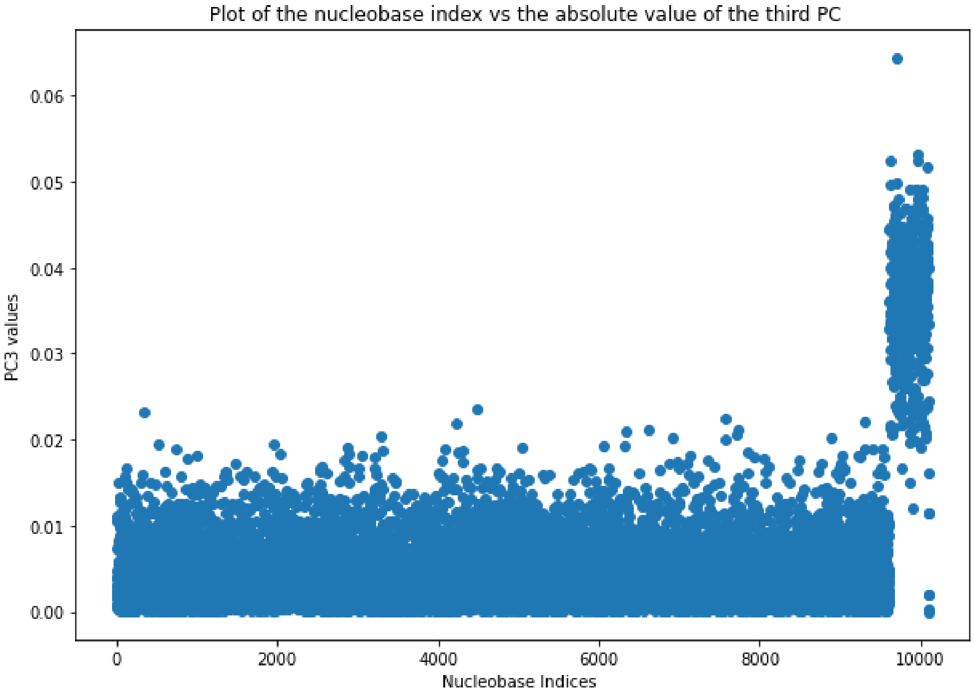


(c) **(Extra credit)** *(2 points)*

From the plot in part a, why are some clusters more concentrated than others? Give an interpretation based on history of human migration.

(d) **(Extra credit)** *(2 points)*

Based on the interpretation of the third principal component in part b, plot the nucleobase index vs the absolute value of the third principal component. What do you notice? What's a possible explanation? Hint: Think about chromosomes and biology.


Plot of the nucleobase index vs the absolute value of the third PC
PC3 values
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
2000
4000
6000
8000
10000
Nucleobase Indices