CS 322 Assignment 3 PCA

There is only one coding problem in this assignment. For a coding problem, you must submit an actual executable file as usual.

Academic honesty: You must complete the work on your own. You may discuss questions with classmates or visit office hours. However, if you and a classmate submit identical solutions, it will be treated as cheating and will receive zero credit—unless you can prove the similarity is an extraordinary coincidence. For more complex assignments or projects, collaboration may be allowed and even encouraged.

Folder submission: Place all files in one folder. Name the folder exactly as your name appears in NEIUPORT. **For example, if your name is Jessica Fatima, submit the folder as Jessica_Fatima.**

Due date: Thursday (11/6).

Grace period: You may submit up to three days late without a penalty of 20%.

Late policy: Submissions more than one week past the due date will not be accepted or graded.

Q1: Read the file named "PCA_disk.py". Finish your work based on it. Overwrite the original file and save your work there.

Perform PCA: Implement the PCA process using the steps below (50 points in total):

- De-mean the dataset by subtracting the mean from each dimension. (10 points)
- Compute the covariance matrix of the mean-centered data. (20 points)
- Calculate the eigenvalues and eigenvectors of the covariance matrix, then sort them in descending order. (20 points)

Q2: Variance Analysis (20 points) Determine and print the explained variance ratio for each principal component. Based on the how the first two principal component explain the variability in the data, would you say the data spreads out like a basketball or a disk?

Q3: Data Projection and Table Creation (20 points) Project the mean-centered 3D data onto the first two principal components derived from PCA. Create a Pandas DataFrame to display the original data alongside their projections onto these two principal components. Print the first 5 rows of this DataFrame to verify the results.

Q4: Visualization (10 points) Create a scatter plot of the projected data points based on the first two principal components obtained from PCA. Ensure that the plot clearly represents how the principal components relate to the projected data.