CISC 7204: Data Science and Visualization
Fall 2024

Assignment 2: Python Data Visualization and Dimensionality Reduction

Due: 11:59 pm on Friday, October 4

Overview

The purpose of this assignment is to give you further hands-on practice with Python data visualization and PCA. Unlike other assignments, you **must** use Python and the provided Google Colab notebook for completing this assignment.

The assignment builds on the in-class activity for September 25. You will start by working through the Colab notebook in class, which provides a guided overview of matplotlib for visualization and scikit-learn for PCA (building on the external tutorials you completed on September 11). After completing this work, you will then use the end of the notebook to write code for a full PCA of an unpublished bioinformatics dataset.

At the end of the assignment, you have the option of doing a bonus activity involving UMAP. Successful completion of the bonus activity will earn you an additional 15 points. The bonus activity is optional, and you can still earn 100% on Assignment 2 without attempting it.

Detailed instructions are provided in the Colab notebook. You should follow those instructions when working on the in-class activity and the assignment.

Dataset

For part of the assignment, you will work with a real bioinformatics dataset about gene expression in endometrial cancer. This dataset is unpublished, and you will not find reference to it anywhere online. It contains measurements of 50 genes related to fatty acid metabolism in 210 control or endometrial cancer samples (the exact gene names have been redacted for confidentiality reasons). Types of endometrial cancer represented include G1/G2/G3 endometriod carcinoma, serous endometrial carcinoma, and clear cell endometrial carcinoma.

Materials to Submit

You should submit the following materials on UMMoodle:

1) A copy of your modified Colab notebook with all of the “EXERCISES” completed and the full code for your PCA of the endometrial cancer dataset (and UMAP, if doing the optional bonus).

2) Your final PCA scatterplot for the penguins dataset in png format.

3) Your final PCA scatterplot for the endometrial cancer dataset in png format.

4) [OPTIONAL] Your final UMAP scatterplot in png format.

You will lose points if you do not follow the file format requirements!

Grades will be based on a combination of your performance on the basic exercises (30 points) and on the PCA with the endometrial cancer dataset (70 points).