STATS/CSE 780 - Homework Assignment 3

Instructor: Pratheepa Jeganathan

2025-10-31

Instruction

- Due before 10:00 PM on Friday, November 14, 2025.
  - Submissions will not be accepted after this due date.

Assignment Standards

Your assignment must comply with the following standards:

- Writing Tools: Use Quarto-Jupyter Notebook for your report and supplemental materials.
- Title Page: Include your name and student number on the title page. Assignments without a title page will not be graded.
- Formatting:
  - Use an 11-point font (Times New Roman or a similar typeface).
  - Apply 1.5 line spacing and ensure 1-inch margins on all sides.
  - This document used these formats.
- Submission Components:
  - Submit a PDF copy of the report (2–3 pages) and supplemental material (less than 10 pages) to Avenue to Learn.
- Collaboration: While you may discuss problems with peers, the final assignment must be your own work.

- Ensure the writing and referencing are appropriate for graduate-level work.
- Various tools may be used to verify the originality of submitted work, including publicly available internet tools.

Report:

- The report must not exceed 3 pages, including tables and figures.
- Choose figures and tables judiciously for the report; additional figures and tables may be included in the supplemental material and referenced in the report.
- You may include one page for the `bibliography` and one page for the `title page`.
- `Python code are not` allowed in the report.

Supplemental Material:

- Include only Python code that generated the results discussed in the report.
- Ensure that the PDF does not display chunk messages, warnings, or extended data frames.
- `Screenshots of Python code are not` allowed in the supplemental material.

Generative AI Policy

- Generative AI is not permitted in this assignment, except for the use of `GitHub Copilot as an assistant for coding`.
- Clearly indicate in the code comments where GitHub Copilot was used as a coding assistant.
- In alignment with McMaster academic integrity policy, it "shall be an offence knowingly to submit academic work for assessment that was purchased or acquired from another source". This includes work created by generative AI tools. Also state in the policy is the following, `Contract Cheating is the act of`outsourcing of student work to third parties" with or without payment." Using Generative AI tools is a form of contract cheating. Charges of academic dishonesty will be brought forward to the Office of Academic Integrity.

Question

Find a dataset that is suitable for the cluster analysis using the methods covered in class. Some sites for dataset search are

1) Google Dataset Search, or
2) Kaggle Datasets, or
3) UCI Machine Learning Repository.

Requirement:

Do not use datasets that have been used in class or collected for your research (not publicly available) or in the textbooks used in this course or R or Python package data.

The dataset must have **at least five variables**.

1. Briefly describe your chosen dataset and clearly explain where it was sourced.
2. Carry out a thorough cluster analysis of your chosen data set using:
    a. agglomerative or divisive hierarchical clustering;
    b. k-means clustering; and
    c. k-means or hierarchical clustering after principal component analysis.
3. Your report must include comparison of the clustering results obtained using these methods.
4. Apply uniform manifold approximation and projection (UMAP) to your chosen dataset and visualize the results. Discuss whether UMAP provides any clustering results that are consistent with the clustering results obtained using the methods in question 2.

Grading scheme

Grading scheme for all the questions is given below.

1.		Source of the dataset [1] describe your dataset (data types, summaries, outliers, missing value analysis, etc.) [3] state the problem to be addressed or explain why the dataset is fit to clustering [2] Any data/statistical transformation or any pre-processing for cluster analysis and principal component analysis [2]
2.	a.	apply hierarchical clustering, describe choosing the number of clusters, evaluate the hierarchical clustering [3]
	b.	apply k-means clustering, describe choosing the number of clusters, evaluate the k-means clustering [3]
	c.	apply PCA, choose the number of PCs (and say why), apply k-means or hierarchical clustering on PCs, describe choosing the number of clusters, evaluate the clustering results [5]
3.		at least two comparisons of the clustering results obtained using these methods [2]
4.		apply UMAP, visualize the results, discuss whether UMAP provides any clustering results that are consistent with the clustering results obtained using the methods in question 2 [3]
References		Reference list starts on a new page, references are appropriate and list out in the **report** [2]
Supplementary material		Supplementary material starts on a new page, code readability, all codes are within the margins, the R codes and the outputs for the questions are presented [3]

The maximum point for this assignment is 29. We will convert this to 100%.

Some datasets that may be used for this assignment

1. Ship Performance Clustering Dataset: https://www.kaggle.com/datasets/jeleeladekunlefijabi/ship-performance-clustering-dataset.
2. Glass Identification: https://archive.ics.uci.edu/dataset/42/glass+identification
3. Heart Failure Clinical Records: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records