STATS/CSE 780
Homework Assignment 3

Pratheepa Jeganathan

13 March, 2023

Instruction

- **Due before 10:00 PM on Friday, March 24, 2023.**
- **Submit a copy of the PDF with your report (2-3 pages) and technical supplemental material (less than 10 pages) to Avenue to Learn using the link that was emailed to you.**
  - Technical supplemental material can only include R/Python codes for the results reported.
- **Late penalty for assignments: 15% will be deducted from assignments each day after the due date (rounding up).**
- **Assignments won't be accepted 48 hours after the due date.**

Assignment Standards

Your assignment must conform to the Assignment Standards listed below.

- RMarkdown or LaTeX is strongly recommended to write the report, and RMarkdown must be used to write the supplemental material.
- Report is about the results by applying data science methods and how you interpret or discuss the results. Don't show in the report how you do the analysis using R/Python.
- Technical supplemental material is how you produce the report results using R/Python. Don't print chunk messages, warnings, or extended data frames in the PDF.
- Write your name and student number on the title page. We will not grade assignments without the title page.

- Eleven-point font (times or similar) must be used with 1.5 line spacing and margins of at least 1~inch all around (The RMarkdown for this assignment used these formats).
- Your report may not exceed **three pages**, including tables and figures. It would help if you chose the tables and figures for the report. You can also keep other tables and figures in the supplementary material (**less than 10 pages**) and refer to the report. In addition, you may use one page for the bibliography and one page for the title page.
- You may discuss homework problems with other students, but you must prepare the written assignments yourself.
- No screenshots are accepted for any reason.
- The writing and referencing should be appropriate to the graduate level.
- The instructor may use various tools, including publicly available internet tools, to check the originality of the submitted work.

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.

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.

Provide a clear and concise description of the results. Clearly state what conclusions can be drawn from your analysis in the context of your chosen dataset.

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 preprocessing 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]
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 26. We will convert this to 100%.