DANA 4840 Project details

1. Before applying any clustering method on your data, it is important to evaluate if the data sets contain meaningful clusters. If the answer is yes, then a natural question to ask is how many clusters are there. This process is described under the chapter on **assessing clustering tendency**: https://www.datanovia.com/en/lessons/assessing-clustering-tendency/

2. You should perform three clustering analyses (k-means, PAM and hierarchical agglomerative clustering) for this project:

(a) identify an on-line data source and extract the required data to perform an application of partitioning clustering (eg, k-means clustering and PAM/CLARA). You should compare both k-means and PAM results and it make more sense to compare the results of both algorithms using the same dataset. Students must use R to complete the clustering analysis.

(b) identify another (different) data set to perform an application of hierarchical clustering. Students must explore various linkage methods and compare their results.

Ideally, data used for the project for at least one of the clustering analyses should be fairly large (at least 150 observations). Note that agglomerative hierarchical clustering is typically performed on smaller datasets.

3. It is important to explain/formulate the purpose of the clustering analyses for the chosen datasets. You should work on a project that has practical and meaningful uses, rather than work on any dataset just for the sake of doing clustering.

4. Marks are allocated for the creation of beautiful and meaningful graphs and plots as covered in the textbook.

5. You must include measures of cluster validation and analysis on the optimal number of clusters as covered by the following chapters in your textbook:

https://www.datanovia.com/en/lessons/assessing-clustering-tendency/
https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/
https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/
https://www.datanovia.com/en/lessons/choosing-the-best-clustering-algorithms/

6. There is a separate research component to this project whereby each team will answer two questions/topics of their choice related to cluster analysis that goes beyond what was discussed in class.

Example of questions are:

- What is the difference between k-means and kmeans++?
- Is there such a thing as k-modes clustering and if so, how does it work and when do we use it?
- What are the ways to identify outliers in clustering?
- What's the difference between MacQueen's kmeans and Hartigan-Wong's kmeans and Lloyd's kmeans?

There is no need to present complicated questions/topics as the intention is for each team to explore clustering beyond what is covered in the classroom. You will be assessed how well you present and explain your answers in the powerpoint (audio included) file as well as in the typed report (with R code). For this portion (two research questions), you are allowed to use R or Python. Please get approval early on the research questions since different teams should not share the same research questions.

7. For each team, please submit a proposal by July 04 (uploaded to Brightspace) identifying the datasets (at least one dataset for partitional clustering, and one dataset for hierarchical clustering). Although no marks are allocated to the proposal (the purpose is to make sure you have a lead on this project by identifying the data sets and have thought of the research questions), you will lose 10% of the project marks for not handing in a project proposal on time. You should include the 2 research questions with the proposal but you are allowed to change the research questions if you get instructor's approval of the modification before the project deadline. You are advised to work on this research portion early in case you run into surprises and need to change your research questions.

8. The final project deliverables (to be uploaded to Brightspace) are due Tuesday 30 July:

(a) a powerpoint file with audio (voices from all team members) that is no longer than 30 minutes. It is expected that most teams will produce powerpoint file that run for between 8 to 15 minutes. Each student must present their work.

(b) a typed report (pdf, generated from R markdown file is best) with explanation that captures all the code and output from R/Python, including all the graphs and plots used in the project.

(c) All data and programs (data files, R and Python files, R markdown .Rmd files etc) are to be uploaded to Brightspace separately so that they can easily be executed if there are questionable portions in your project. Please organize your work so that it is easy for the instructor to grade your work.

Grading Rubric:

Powerpoint with audio (clear explanation, attractive and useful content included)	20
Report writing and organization, free of glaring grammatical errors	20
R code – readability, proficiency, documentation and comments	10
Content (purpose of applying clustering method to chosen data, measures for cluster validation, analysis on the optimal number of clusters, results)	20
Research questions component – how well did the team answer the questions of interest	20
Ease to grade entire project – organization and other factors	10