Assignment Instructions: Assignment 5

Purpose

The purpose of this assignment is to use Hierarchical Clustering

Directions

The dataset Cereals.csv includes nutritional information, store display, and consumer ratings for 77 breakfast cereals.

Data Preprocessing. Remove all cereals with missing values.

- Apply hierarchical clustering to the data using Euclidean distance to the normalized measurements. Use Agnes to compare the clustering from single linkage, complete linkage, average linkage, and Ward. Choose the best method.
- How many clusters would you choose?
- Comment on the structure of the clusters and on their stability. Hint: To check stability, partition the data and see how well clusters formed based on one part apply to the other part. To do this:
    - Cluster partition A
    - Use the cluster centroids from A to assign each record in partition B (each record is assigned to the cluster with the closest centroid).
    - Assess how consistent the cluster assignments are compared to the assignments based on all the data.
- The elementary public schools would like to choose a set of cereals to include in their daily cafeterias. Every day a different cereal is offered, but all cereals should support a healthy diet. For this goal, you are requested to find a cluster of "healthy cereals." Should the data be normalized? If not, how should they be used in the cluster analysis?

File Attached: Cereals.csv

Learning Outcomes

The assignment will help you with the following course outcomes:

1. Think critically about how to use machine learning algorithms to solve a given business problem.
2. Know how to formulate business problems and identify relevant data to use in modeling frameworks.
3. Know how to evaluate the appropriateness and estimate the performance of using k-Means for a given task.
4. Know how to use software tools (such as R) effectively to implement k-Means.
5. Foster the communication and presentation of statistical results and inferences.

Requirements

All due dates are included in the Assignment Schedule.

General Submission Instructions

All work must be your own. Copying other people's work or from the Internet is a form of plagiarism and will be prosecuted as such.

1. Create a new folder called **Assignment_5** in your previously created GitHub repository.
2. If you are using R, then upload the R Markdown file, the knitted pdf/html file, and any other data file you might have used for the assignment.
3. If you using Python, then share the Jupyter/Google Colab notebook in our Assignment_5 folder on GitHub

Provide the link to your git repository in Canvas for the assignment.