

Assignment Instructions: Assignment 4

Purpose

The purpose of this assignment is to use k-Means for clustering.

Directions

An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data gathered on 21 firms in the pharmaceutical industry are available in the file *Pharmaceuticals.csv*. For each firm, the following variables are recorded:

1. Market capitalization (in billions of dollars)
2. Beta
3. Price/earnings ratio
4. Return on equity
5. Return on assets
6. Asset turnover
7. Leverage
8. Estimated revenue growth
9. Net profit margin
10. Median recommendation (across major brokerages)
11. Location of firm's headquarters
12. Stock exchange on which the firm is listed

Use cluster analysis to explore and analyze the given dataset as follows:

- a. Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.
- b. Interpret the clusters with respect to the numerical variables used in forming the clusters.
- c. Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)
- d. Provide an appropriate name for each cluster using any or all of the variables in the dataset.

File Attached: *Pharmaceuticals.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_4** 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_4 folder on GitHub

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