Homework 2

Instructions

1. Put all your answers and results in a **single Microsoft Word document** with your team number, and save all the R codes used for this assignment in a **single R Script file** with your team number. **Submit your Word file and R file into eLearning**. Grading will be based on the answers and results provided in your **Word document**. We check you R codes only when we see any potential problems (e.g., something suspicious). Correct R codes without good answers in the Word file receive no credits.
2. **Late submissions are not acceptable** and will be rejected by eLearning.
3. A professional quality report is expected–messy or hard-to-read reports will be penalized.
4. Explain your answers. **Be as clear as possible**. Vague answers–even if they are long–will not receive full credit. Information in excess of what the question warrants is acceptable as long as it is relevant and correct. Incorrect information, even if unwarranted, will be penalized. Therefore, **proofread your report to tidy it up before submission**.

Questions

1. Clustering Stores: The DUNGAREE data set shows the number of pairs of four different types of dungarees sold at stores over a specific time period. Each row represents an individual store. There are six columns in the data set. One column is the store identification number, and the remaining columns contain the number of pairs of each type of jeans sold. (**5 points**)

Name	Model Role	Data Type	Description
STOREID	Ident	Numeric	Identification number of the store
FASHION	Input	Numeric	Number of pairs of fashion jeans sold at the store
LEISURE	Input	Numeric	Number of pairs of leisure jeans sold at the store
STRETCH	Input	Numeric	Number of pairs of stretch jeans sold at the store
ORIGINAL	Input	Numeric	Number of pairs of original jeans sold at the store
SALESTOT	Ignore	Numeric	Total number of pairs of jeans sold (the sum of FASHION, LEISURE, STRETCH, and ORIGINAL)

Use R to run k-mean clustering (based on the code shown in class):

(a) Import the data to R and remove the column(s) that you are not going to use. Copy the R code used below.

(b) Examine the input variables: Are there any unusual data values? Are there missing values that should be replaced?

(c) Normalize the data. Copy the R code used below. What would happen if you did not standardize/normalize your inputs?

(d) Run k-means clustering using a seed = 42, and choose k = 20. Copy the R code used below.

(e) Based on the results, does k=20 clusters seem appropriate? Why or why not?

(f) In the next run, specify a maximum of six clusters, and run the k-means clustering algorithm again. Copy the R code used below.

(g) Plot profile plot of centroids for the six clusters generated in (f). Copy the code used and the result below.

(h) Using the profile plot of centroids, interpret the characteristics of each cluster as it relates to types of jeans sold at stores. Describe these clusters, and their similarities and differences in words.

2. Clustering Pharmaceutical Firms: 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. (**5 points**)

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.

Name	Model Role	Data Type	Description
Symbol	Ignore	Categoric	Company stock symbol
Name	Ignore	Categoric	Company name
Market_Cap	Input	Numeric	Market capitalization (in billions of dollars)
Beta	Input	Numeric	Beta
PE_Ratio	Input	Numeric	Price to earnings ratio
ROE	Input	Numeric	Return on equity
ROA	Input	Numeric	Return on investment
Asset_Turnover	Input	Numeric	Asset turnover
Leverage	Input	Numeric	Leverage
Rev_Growth	Input	Numeric	Estimated revenue growth
Net_Profit	Input	Numeric	Net profit margin
Median_ Recommendation	Ignore	Categoric	Median recommendations (across major brokerages)
Location	Ignore	Categoric	Location of company headquarters
Exchange	Ignore	Categoric	Stock exchange on which the firm is listed

Use R to run hierarchical clustering (based on the code shown in class):

(a) Import the data to R, set row names to the “Symbol” column, and remove all the columns that you are not going to use for clustering. Copy the R code used below.

(b) Normalize the data. Copy the R code used below.

(c) Based on single linkage, run hierarchical clustering to generate Dendrogram. Copy the code used and the result below.

(d) If we are interested in 6 clusters based on Dendrogram in (c), what are the members of each cluster? Copy the code used and the result below.

(e) Based on complete linkage, run hierarchical clustering to generate Dendrogram. Copy the code used and the result below.

(f) If we are interested in 6 clusters based on Dendrogram in (e), what are the members of each cluster? Copy the code used and the result below.

(g) Do (d) and (f) lead to the same six clusters? Explain why.