BUS 336: Data Analytics & Visualization

Instructor: Gohram Baloch

Assignment 1 – Clustering

Total Marks: 60

Due date: 11:59 PM Wednesday June 05, 2024

Instructions:

Please read the instructions carefully as your assignment will be graded based on the assumption that the instructions are followed.

1. This assignment is based on the data file "NasdaqReturns.csv," which has been uploaded to Canvas. The assignment focuses on clustering and should be completed individually.
2. You must use either R or Tableau while answering questions. If you use any other tool such as Excel, you won't receive any mark.
3. You will need to submit:
   a. A single PDF file on Crowdmark. While answering each question, you report must include lines of R code used and the resulting output (e.g., tables, graphs).
   b. The complete .R file and tableau file (.twb) along with csv file used, to be uploaded on Canvas.
4. Late submissions will not be accepted.

Please review the instructions carefully and make sure to follow the submission requirements.

Clustering Stock Returns

When building portfolios of stocks, investors seek to obtain good returns while limiting the variability in those returns over time. This can be achieved by selecting stocks that show different patterns of returns. In this problem, we will use clustering to identify clusters of stocks that have similar returns over time; an investor might select a diverse portfolio by selecting stocks from different clusters.

For this problem, we will use the dataset `NasdaqReturns.csv`, which contains monthly stock returns from the NASDAQ stock exchange during 2000-2009. The companies selected in this dataset are limited to those that were listed on the stock exchange for this entire time period and whose stock price never fell below $1. The NASDAQ is the second-largest stock exchange in the world, and it lists many technology companies. The variables in the dataset are described in the table below:

Table 1: Data Codebook

Variable	Description
`StockSymbol`	The symbol identifying the company of the stock
`Industry`	The industry the stock is classified under
`SubIndustry`	The sub-industry the stock is classified under.
`Ret2000.01 - Ret2009.02`	The return for the stock during the variable's indicated month. The variable names have the format "RetYYYY.MM," where YYYY is the year and MM is the month. For instance, variable Ret2005.02 refers to February 2005. The value stored is a proportional change in stock value during that month. For instance, a value of 0.05 means the stock increased in value 5% during the month, while a value of -0.02 means the stock decreased in value 2% during the month. There are 120 of these variables, for the 120 months in our dataset

Part I: Data Exploration (15 points)

a) How many companies are in this dataset? How many companies are there in each of the industries? **(5 points)**

b) In the aftermath of the dot-com bubble bursting in the early 2000s, the NASDAQ was quite tumultuous. In December 2000, how many stocks in this dataset saw their value increase by 10% or more? And how many saw decrease by 10%? **(5 points)**

c) Entering the Great Recession, most stocks lost significant value, but some sectors were hit harder than others. In October 2008, which three industries had the **worst average return** across all stocks in that industry? **(5 points)**

Part II: Cluster Analysis (45 points)

Let us now cluster the stocks according to the monthly returns. For the remainder of this problem, make sure that you are just clustering the observations based on the variables `Ret2000.01 - Ret2009.12` (`StockSymbol`, `Industry`, and `SubIndustry` should not be used to cluster the observations). You can do this by creating a new dataset only containing the variables `Ret2000.01 - Ret2009.12`.

a) In this analysis, we will not normalize our data prior to clustering. Why is this a valid approach for this problem and dataset? **(5 points)**

b) Cluster the data using Hierarchical clustering. Clearly indicate which distance metrics you used for distances. Plot the resulting dendrogram. What do you think are reasonable choices for the number of clusters to select, based on the dendrogram? Select a specific number of clusters to use for the rest of the problem and justify your choice. **(15 points)**

c) Extract cluster assignments from your hierarchical clustering model and add a column to the original data frame (nasdaq table), using the number of clusters you selected in the previous problem. Save your new data frame as csv file for visualization in Tableau. **(5 points)**

d) Describe each cluster by plotting a bar graph showing the number of stocks in each clusters. By looking at the bar plot, do you think your choice of the number of clusters in part (b) is justifiable? Explain your answer. **(5 points)**

e) Plot a heatmap to describe your clusters both by the number of stocks and industry of the companies in the cluster. Hint: x-axis as clusters, y-axis as Industry. **(5 points)**

f) For some months, we expect there to be significant differences between the returns of stocks in different clusters. Plot a barplot of average returns for February 2000 for clusters. Identify clusters with negative average returns and the clusters with positive average returns. **(5 points)**

g) In the introduction to this problem, we discussed the value of a diverse portfolio and how we might achieve this objective by selecting stocks from different clusters. Propose a diverse portfolio of stocks to minimize overall risk using the results of your clusters. Justify your selected portfolio. **(5 points)**