FE582 – Assignment 2

Problem 1

Write an R code to scrape the website:

https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

Retrieve the content of the S&P 500 component stocks table (Symbol, Security, GISC Sector, GICS Sub-Industry, Headquarters Location, Date added, CIK, Founded). Create an R dataframe and perform exploratory data analysis and report summary statistics.

Problem 2

The data provided in the files contains several quantitative and categorical variables associate with each ticker. Please select a subset of 100 tickers from each file and use data for a specific year (ex: 2013). Use a small number of quantitative variables (10 or 12) out of ~76 columns available (example: After Tax ROE, Cash Ratio, Current Ratio, Operating Margin, Pre-Tax Margin, Pre-Tax ROE, Profit Margin, Quick Ratio, Total Assets, Total Liabilities, Earnings Per Share, etc…). The categorical variables available are GICS Sector, GICS Sub Industry, and possibly HQ Address (although this is sparse data for the 100 tickers subset selected).

Next, caclulate several distance and similarity functions to find the extreme values for distance and similarities between the subset of tickers that you chose. For each of the following cases, please define the function that allows you to calculate the quantity required, calculate the values for all ticker pairs, and rank the pairs by calculated value of distance or similarity, and report the top and bottom 10 values along with the ticker pairs for the following cases:

a) L_p-norm for $p = 1$
b) L_p-norm for $p = 2$
c) L_p-norm for $p = 3$
d) L_p-norm for $p = 10$
e) Minkovski distance (assign different weights for the feature components in the $\mathrm{L_p}$-norm based on your assessment on the importance of the features)
f) Match-Based Similarity Computation (use a small number of equi-depth buckets, ex: 3)
g) Mahalanobis distance
h) Similarity: overlap measure
i) Similarity: inverse frequency
j) Similarity: Goodall
k) Overall similarity between tickers by using mixed type data (choose a λ value for calculation)
l) Overall normalized similarity between tickers by using mixed type data (choose a λ value for calculation)