STAT4996 Capstone Project Proposal

Group members: Zhenzhen Zhu (zz8vy), Xinru Cheng (xc9mb), Yimeng Xiao (yx6df), Lauren Stover (les2fd)

We are interested in working on a data analysis project about corporate finance. We obtained some data from Simfin.com (https://simfin.com/data/find/companies Note: Registration is needed for data access), which contains company names, ticker, industry code, the number of stock shares, and quarterly and annually financial statistics including Revenue, COGS (cost of goods sold), operating expenses, SG&A (selling, general & administrative expenses), EBITDA (earnings before interest, taxes, depreciation, and amortization), Net income, etc. since Q1 2008. Data in the data set comes from financial statements uploaded on the SEC server, and has been cleaned and organized already, though there are potentially some missing observations. One challenge is that we need to extract the data set from the virtual environment created by the data provider (We have followed the github tutorials about downloading the data from the server and viewing the data in the virtual environment, but we still need to figure out how to download it to our local computer).

Questions of interest include but not limit to:

1. What are the best predictors that can predict if a company is financially healthy? (Using regression tree and industry-specific knowledge)
2. Identify companies that are likely to earn the highest (predicted) net profit (using time series forecasting) based on past performance.
3. Which industry has the biggest potential for growth in terms of net profit? Are there any industries that are correlated in terms of growth trend?
4. Which companies are the industry leaders? What are the metrics that identify industry leaders?
5. Identify companies in a particular segment that appear to be under-evaluated through comparative analysis. Such companies have great potential for investment. We can build valuation metrics including company's relative growth rate, geographic and product line diversification, perceived management quality (may require information outside of the obtained data set), and balance sheet risk.

To analyze this time series data, we will use techniques generally like smoothing, SARIMA model, estimation,forecasting, clustering, decision trees, and etc.We might need to compensate with other data in order to make our analysis more thorough and dynamic. Starting from the next semester, we may want to first of all accomplish the following things:

1. Look closely at any missing observations, either find the missing value from other resources or eliminate that observation.
2. Consult domain experts and conduct exploratory data analysis to find out variables that can possibly predict a company's financial status, future net profit, and/or potential for growth.
3. Subset the data set in a few ways as appropriate:
    a. Subset/group by industry: retail, medical/drug, software/hardware, business...

b. Subset by financial statements: income statement, balance sheet, cash flow statement.