The Proposal of final Project

Student: Zheng Fang and Pengfei Yang
DA 210-02 / CS 181-02: Data Systems
Spring 2023
Instructor: Dr. Tanya Amert

The Central Question:

We want to know the influences of pandemic of COVID-19 in United States on three representative stocks in each of the 11 sectors of the US stock market based on GICS (Global Industry Classification Standard).

Data Sets:

We will be using two types of databases.

(1) Trends in Number of COVID-19 Cases and Deaths in the US Reported to CDC

(a) This data set obtained from the CDC website, which contains the weekly number of new COVID-19 cases in the United States from January 29th, 2020 to March 22nd, 2023 (present).

(b) Link: https://covid.cdc.gov/covid-data-tracker/#trends_weeklycases_select_00

(c) This database consists of 166 rows, with each row representing a week.

(d) The database contains 4 columns.

Variable Name	Variable Introduction	Variable Type
Geography	Region name	String
Date	Start date of the week (including month, day, and year)	String
Weekly Cases	Newly reported COVID-19 cases for the week	Int
New Historic Cases	Confirmed cases not from current week but reported during this week	Int

(e) In this file, the Independent value is: Geography and Date

(f) The Dependent value is: Weekly Cases and New Historic Cases

(2) US Stock Dataset

a) The second dataset we will be using is based on the Global Industry Classification Standard (GICS), which is a classification system for industries developed jointly by Morgan Stanley Capital International (MSCI) and Standard & Poor's (S&P) in 1999.

b) This system divides all companies into 11 different industries, and for each industry, we will analyze the stock information of **three** top companies.

c) Link: https://www.kaggle.com/datasets/footballjoe789/us-stock-dataset

d) As the data package is too large (8GB) and contains daily US stock market information

for each publicly listed company, we will only need to analyze the US stock market performance of the 33 selected companies.

e) This database consists of 33 CSV files, each containing information on a company's US stock market performance for each trading day from its initial public offering to March 24th, 2023. Each CSV file contains 8 columns.

Variable Name	Variable Introduction	Variable Type
Data	The date of the trading day	String
Open	The opening price of the stock on the trading day	Float
High	The highest price of the stock on the trading day	Float
Low	The lowest price of the stock on the trading day	Float
Close	The last price at which a stock trades during a regular trading session	Float
Volume	The number of shares traded in a particular stock, index, or other investment over the trading day	Float
Dividends	A payment to shareholders that consists of additional shares rather than cash	Float
Stock Splits	The number of shares of that company increases	Float

f) In these files, the Independent value is: Date

g) The Dependent value is: Open, High, Low, Close, Volume, Dividends, Stock Splits

h) Due to different listing times for each company, the number of rows in the CSV files varies. Below is an introduction to the CSV file data for the companies we have selected in each industry.

*** Information Technology**

* AAPL.csv
    - This CSV file contains the stock information of Apple Inc.
    - The CSV file has 10661 rows and 8 columns.
* MSFT.csv
    - This CSV file contains the stock information of Microsoft Corp.
    - The CSV file has 9335 rows and 8 columns.
* AMZN.csv
    - This CSV file contains the stock information of Amazon.com, Inc.
    - The CSV file has 6509 rows and 8 columns.

*** Financials**

* V.csv
    - This CSV file contains the stock information of Visa

    - The CSV file has 3782 rows and 8 columns.
  * JPM.csv
    - This CSV file contains the stock information of JPMorgan Chase
    - The CSV file has 10849 rows and 8 columns.
  * C.csv
    - This CSV file contains the stock information of Citigroup Inc
    - The CSV file has 11658 rows and 8 columns.

*** Healthcare**

  * JNJ.csv
    - This CSV file contains the stock information of Johnson & Johnson
    - The CSV file has 15414 rows and 8 columns.
  * MRNA.csv
    - This CSV file contains the stock information of Moderna
    - The CSV file has 1082 rows and 8 columns.
  * BNTX.csv
    - This CSV file contains the stock information of BioNTech SE – ADR
    - The CSV file has 871 rows and 8 columns.

*** Consumer Discretionary**

  * QSR.csv
    - This CSV file contains the stock information of Restaurant Brands International Inc
    - The CSV file has 2086 rows and 8 columns.
  * MAR.csv
    - This CSV file contains the stock information of Marriott
    - The CSV file has 6295 rows and 8 columns.
  * NKE.csv
    - This CSV file contains the stock information of Nike Inc
    - The CSV file has 10669 rows and 8 columns.

*** Industries**

  * UPS.csv
    - This CSV file contains the stock information of UPS
    - The CSV file has 5881 rows and 8 columns.
  * UAL.csv
    - This CSV file contains the stock information of United Airlines
    - The CSV file has 4314 rows and 8 columns.
  * CAT.csv
    - This CSV file contains the stock information of Caterpillar Inc.
    - The CSV file has 15414 rows and 8 columns.

*** Communication Services**

  * NFLX.csv
    - This CSV file contains the stock information of Netflix
    - The CSV file has 5247 rows and 8 columns.
  * ZM.csv
    - This CSV file contains the stock information of Zoom
    - The CSV file has 992 rows and 8 columns.

    * T.csv
        - This CSV file contains the stock information of AT&T Inc.
        - The CSV file has 9917 rows and 8 columns.

*** Consumer Staples**

    * KO.csv
        - This CSV file contains the stock information of The Coca-Cola Company
        - The CSV file has 15414 rows and 8 columns.
    * PG.csv
        - This CSV file contains the stock information of Procter & Gamble Co
        - The CSV file has 15414 rows and 8 columns.
    * UL.csv
        - This CSV file contains the stock information of Unilever plc
        - The CSV file has 10843 rows and 8 columns.

*** Energy**

    * XOM.csv
        - This CSV file contains the stock information of Exxon Mobil Corporation
        - The CSV file has 15414 rows and 8 columns.
    * CVX.csv
        - This CSV file contains the stock information of Chevron Corporation
        - The CSV file has 15414 rows and 8 columns.
    * SHEL.csv
        - This CSV file contains the stock information of Shell plc
        - The CSV file has 7151 rows and 8 columns.

*** Real Estate**

    * AMT.csv
        - This CSV file contains the stock information of American Tower Corporation
        - The CSV file has 6311 rows and 8 columns.
    * PLD.csv
        - This CSV file contains the stock information of Prologis, Inc
        - The CSV file has 6376 rows and 8 columns.
    * CCI.csv
        - This CSV file contains the stock information of Crown Castle International Corp.
        - The CSV file has 6192 rows and 8 columns.

*** Materials**

    * DD.csv
        - This CSV file contains the stock information of DuPont de Nemours, Inc.
        - This CSV file has 12815 rows and 8 columns.
    * BHP.csv
        - This CSV file contains the stock information of BHP Group Limited
        - This CSV file has 10849 rows and 8 columns.
    * SHW.csv
        - This CSV file contains the stock information of The Sherwin-Williams Company
        - The CSV file has 10849 rows and 8 columns.

*** Utilities**

* SO.csv
    - This CSV file contains the stock information of The Southern Company
    - The CSV file has 10396 rows and 8 columns.
* ED.csv
    - This CSV file contains the stock information of Consolidated Edison, Inc.
    - The CSV file has 15414 rows and 8 columns.
* AEP.csv
    - This CSV file contains the stock information of American Electric Power Company, Inc.
    - The CSV file has 15414 rows and 8 columns.

Outline and Functions of the Project:

1. We want to first store information from CDC (Centers for Disease Control and Prevention) web pages locally in HTML format. The Data Table for Weekly Case Trends - The United States on this page provides the weekly number of new COVID-19 cases in the United States from January 29, 2020, to the present. Therefore, our first step is downloading the information we need to our computer.
2. Download the US stock market information we found on Kaggle. This dataset contains the daily change information of most of our US stocks since their establishment. We need to select the three representative stocks in each of the 11 sectors of the US stock market and have the same working directory as the HTML file in Step 1.
3. Next, we need to use xpath to convert the HTML table about the COVID-19 epidemic in the United States into a Tidy Data frame and remove the unrelated columns.
4. Similarly, the CSVs of the 33 stocks we selected were converted into Tidy Data Frame using Pandas, and unrelated columns were removed.
5. For the 33 stocks from 11 sectors, we need to unify their daily change information with the weekly change information of new cases. Therefore, we need to calculate the average value of the stock information every seven days and then form a new Tidy Data Frame, which records the weekly stock change information.
6. The weekly change rate of new epidemic cases was produced using the data frame on the COVID-19 epidemic in the United States. Store the weekly change rate of new topics in a new column of this data frame.
7. Using the 33 data frames that provide stock information for the 11 sectors, create their weekly rate of change in stock prices. Store the weekly change rate of each stock's stock price in a new column in their corresponding data frame.
8. Compare the weekly change rate of each stock with the weekly change rate of new cases of the US epidemic. If both values increase or decrease simultaneously, mark this week as a positive proportion; Conversely, if one of the two values is an increase and the other is a decrease, then keep this week inversely proportional.
9. By calculating the proportion of positive and negative weeks in total weeks, we can obtain the time proportion of each stock's stock price increase and decrease caused by the US epidemic.
10. Based on the proportion in Step 9, summarize whether the overall impact of the US pandemic on the stock prices of these 33 stocks is positive or negative.