

Term Project

Saravanan Janarthanan

DSC540 Data Preparation, Bellevue University

Professor: Catherine Williams

Week 4 – Assignment due date: 4/7/2024

Project: **S&P 500 Component Stock performance**

The project objective is to analyze the S&P 500 stock's performance data by using historical information, stock company information and live query of price and changes through online API calls

Data Sources

Flat File: SNP_500_Performance_data.csv will be used, it has 503 rows and 30 columns

| Column name | Details |
|-------------------------|--|
| Symbol | Stock Ticker Symbol |
| Name | Name of the company |
| Market Cap | Market Capitalization |
| P/E | Price Earnings ratio |
| EPS | Earnings per share |
| Net Income | Net income |
| Beta | Coefficient quantifies the volatility of a stock's returns concerning the market or S&P 500, determined through a 60-month historical regression analysis of the stock's return in comparison to the return of the S&P 500 |
| Dividend | Dividend per share |
| Div Yield | Dividend yield |
| Extended Hr. Last Price | Extended hours last price |
| Extended Hr. Change | Extended Hours price change compared to previous day close price |
| %Chg (Ext) | Extended hour price change % compared to previous day close price |
| Time | Date of the snapshot |
| Last Price | Normal trading hours last price |
| Change | Normal trading hours last price change compared to previous day close price |
| %Chg | Normal trading hours last price change % compared to previous day close price |
| Opinion | Buy or Sell option summary |
| 20D Rel Str | 20 Day relative strength, 70% states over bought and 30% and below states over sold |
| 20D His Vol | 20-day historical volatility, the average deviation from average price over last 20 days |

| | |
|-------------|---|
| 20D Avg Vol | 20-day average volume traded |
| 52W Low | 52-week low price |
| 52W High | 52-week high price |
| Wtd Alpha | a measure of how much this stock has risen or fallen |
| YTD %Chg | Price difference between end of last calendar year close price to current price |
| 1M %Chg | Price difference between last one-month close price to current price |
| 3M %Chg | Price difference between last 3-month close price to current price |
| 52W %Chg | price difference between previous 52 week closing price and current price |
| Sector | Business Sector of the Stock |
| ISIN | International Securities Identification Number |
| Volume | Volume traded on the snap shot date |

API

URL: <https://polygon.io/>

Polygon.io API will be used to query the stock price details online through REST calls. Python polygon module will be used to fetch details or can use regular “urllib.request” module to make raw requests. API key needs to be added to the request for authorization

Sample URL that returns JSON or CSV

<https://api.polygon.io/v2/aggs/grouped/locale/us/market/stocks/2023-01-09?adjusted=true&apiKey=.....>

| Response label | Details |
|----------------|--|
| T | The exchange / Ticker symbol that this item is traded under. |
| c | The close price for the symbol in the given time period. |
| h | The highest price for the symbol in the given time period. |
| l | The lowest price for the symbol in the given time period. |
| n | The number of transactions in the aggregate window. |
| o | The open price for the symbol in the given time period. |
| t | The Unix Msec timestamp for the end of the aggregate window. |
| v | The trading volume of the symbol in the given time period. |
| vw | The volume weighted average price. |

Term Project – Identify Datasets

4

Web site:

URL: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies#S&P_500_component_stocks

[List of S&P 500 companies - Wikipedia](https://en.wikipedia.org/wiki/List_of_S%26P_500_companies#S&P_500_component_stocks)

This site details the S&P 500 company details and will be used for web scraping the below details. It will have 503 rows and 8 columns

| Column name | Details |
|-----------------------|--|
| Symbol | Ticker Symbol or Stock Symbol |
| Security | Company name |
| GICS Sector | Industry taxonomy - Global Industry classification Standard Sector - that classifies the company sector it is grouped |
| GICS Sub-Industry | Industry taxonomy - Global Industry classification Standard sub-industry - that classifies the company sub section of the sector it is grouped |
| Headquarters Location | Company Headquarters city and State |
| Date added | Date the stock was added to S&P 500 index list |
| CIK | Central Index Key assigned to the company |
| Founded | Company founded year |

Relationship

Each dataset is linked through the **Ticker or stock symbol**, which is universally utilized across stock markets. This symbol serves as a means to correlate information from these three sources

Project Approach / Plan

The approach involves key activities such as gathering and comprehending the data, amalgamating the datasets utilizing a shared identifier (ticker / stock symbol), addressing data cleanliness by handling missing values and outliers, and crafting visualizations to discern the interrelationships among data attributes.

Data Collection and Understanding:

- Access the datasets from their respective financial sources through diverse methods, this activity is completed except the API calls
- Gain insight into the organization, format, and contents of each dataset, comprehensively understanding its structure and layout
- Identify the common stock ticker symbol field across all datasets, especially API calls as it will return over and above S&P 500 stocks

Data Integration:

- Merge or join the datasets based on the common stock ticker symbol field.
- Choose an appropriate method for integration, such as concatenation, merging, or joining, depending on the structure and relationships between datasets.

Data Cleaning and Preprocessing:

- Handle missing values, outliers, and inconsistencies in the datasets.
- Standardize column names, data types, and formats across datasets to ensure compatibility.
- Perform any necessary transformations or preprocessing steps to prepare the data for analysis.

Data Analysis and Modeling:

- Define the objectives and questions to be addressed with the combined dataset.

Visualization:

- Create visualizations, dashboards, or reports to communicate the results effectively.

The exploratory data analysis (EDA) and modeling approach will build upon the preceding data preparation activities.

Concerns or Challenges

Expertise in the sector or specific stocks is required to effectively manage outliers and missing values. Additionally, any modifications to the S&P 500 index may necessitate adjustments to the list of stocks employed

Ethical implications

Ensuring the accuracy and integrity of the stock information utilized for analysis is paramount. Although verifying the reliability of data sources can be challenging, efforts are made to minimize errors or inaccuracies. It's important to note that since these data are sourced from openly available channels, there are no concerns regarding data privacy violations, fair usage, regulatory compliance, or conflicts of interest with any stock company stakeholders. Moreover, it's emphasized that this information is solely used for educational purposes.