Student: Balakrishna Mupparaju

Assignment: Project Milestone 1

Table of Contents

[Project Subject Area: 1](#_Toc194770370)

[Data Sources: 2](#_Toc194770371)

[Flat File: 2](#_Toc194770372)

[Website: 2](#_Toc194770373)

[API: 2](#_Toc194770374)

[Relationships: 3](#_Toc194770375)

[Project Approach/Plan: 3](#_Toc194770376)

[Concerns/Challenges: 4](#_Toc194770377)

[Ethical Implications: 5](#_Toc194770378)

# 

# Project Subject Area:

The project will focus on data wrangling and integration by combining three data sources—Kaggle dataset, Wikipedia, and Alpha Vantage API—to analyze S&P 500 companies. The goal is to explore the relationships between company financial data, market performance, and company metadata to gain insights into how these companies perform in the stock market and their broader industry.

# Data Sources:

## Flat File:

* **Description of Data Source**:

The Kaggle dataset contains structured data about S&P 500 companies, including financial metrics (e.g., revenue, market cap, P/E ratio, debt-to-equity ratio) and company details (e.g., sector, industry).

* **Link or Flat File Uploaded**:

[Kaggle S&P 500 Companies Dataset](https://www.kaggle.com/datasets/rohanrao/s-and-p-500-companies)

## Website:

* **Description of Data Source**:

The Wikipedia page lists the S&P 500 companies, providing metadata such as the company’s name, ticker symbol, sector, industry, headquarters, and CEO name. This data complements the Kaggle dataset by adding company-specific identifiers and additional company attributes.

* **Link**:

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

## API:

* **Description of Data Source**:

The Alpha Vantage API provides real-time and historical stock market data, including stock prices, trading volume, and various financial indicators for individual companies in the S&P 500.

* **Link**:

[Alpha Vantage API Documentation](https://www.alphavantage.co/documentation/)

## Relationships between data sets:

The three data sources are connected through the **Ticker Symbol**, which serves as the unique identifier for each company across all three datasets. The Kaggle and Wikipedia datasets include **Ticker Symbol** and company-specific data, while the Alpha Vantage API provides financial performance data (e.g., stock prices, earnings). By merging these datasets on the **Ticker Symbol**, we can combine the historical economic data from Alpha Vantage with the company-specific information from Kaggle and Wikipedia.

* **Kaggle Dataset**: Provides static financial metrics (e.g., market cap, revenue, P/E ratio) and company identifiers such as the **Ticker Symbol** and **Sector**.
* **Wikipedia Dataset**: Adds company metadata such as **CEO**, **Headquarters**, and **Date Added to S&P 500**, which can be merged with the Kaggle data.
* **Alpha Vantage API**: Provides real-time stock market data (e.g., **Open Price**, **Close Price**, **Volume**, **PE Ratio**), which is dynamic and can be added to the merged dataset to analyze market performance.

The merging of these datasets will allow for an enriched company analysis, combining historical financial data and current stock market performance.

# Project Approach/Plan:

1. **Data Collection**:
   * Download the Kaggle dataset for S&P 500 companies.
   * Use the Wikipedia API or web scraping techniques to extract company metadata (e.g., CEO, headquarters, sector).
   * Fetch real-time stock data from Alpha Vantage using the API for each company in the dataset.
2. **Data Cleaning**:
   * Clean missing values, handle duplicate rows and ensure consistency in the **Ticker Symbol** across all datasets.
   * Standardize date formats, especially for time-series data from the Alpha Vantage API.
   * Handle discrepancies between datasets, such as missing company names or sector details.
3. **Data Integration**:
   * Merge the three datasets using the **Ticker Symbol** to create a unified dataset with company financial metrics, metadata, and real-time stock performance data.
   * Perform data transformation (e.g., aggregating stock data, creating new features such as average stock price over a period).
4. **Analysis and Insights**:
   * Perform exploratory data analysis (EDA) to understand the relationships between stock performance, financial metrics, and company metadata.
   * Conduct correlation analysis and visualize key metrics (e.g., market cap vs. stock performance, P/E ratio vs. earnings).
5. **Presentation**:
   * Prepare a comprehensive report or presentation highlighting insights, potential trends, and recommendations based on the analysis.

# Concerns/Challenges:

1. **Data Inconsistencies**:

The primary challenge will be handling inconsistencies between the three data sources. For example, missing or inconsistent entries for company names or ticker symbols could make merging the datasets challenging. Data formatting issues could also arise, especially when working with stock data from the Alpha Vantage API.

1. **API Limitations**:

The Alpha Vantage API has rate limits (e.g., limited number of requests per minute/hour). This could affect the ability to fetch real-time data for many companies simultaneously. We may need to implement delays or batch requests to overcome this limitation.

1. **Data Quality**:

Some companies in the S&P 500 list may not have complete data in one of the datasets (e.g., missing revenue data in the Kaggle dataset or outdated CEO information in Wikipedia). Cleaning this data while ensuring its accuracy will be time-consuming.

1. **Data Synchronization**:

The Kaggle dataset provides historical financial metrics, Wikipedia offers company metadata, and the Alpha Vantage API provides real-time stock data. Aligning these datasets for analysis, especially time-sensitive data like stock prices, may require careful handling.

# Ethical Implications:

1. **Data Privacy and Security**:

While the data used for this project is publicly available, ensuring that any data obtained via APIs is handled responsibly is essential. Users must not share sensitive or proprietary data that could breach terms of service agreements.

1. **Bias in Data**:

The project is based on publicly available data, but it’s essential to recognize potential biases in the data. For example, the Kaggle dataset and Wikipedia entries might be skewed by data availability for large, publicly recognized companies, while smaller or newer companies may not be as well represented.

1. **Financial Implications**:

Since the project involves financial data (stock prices, market performance), there could be implications for users relying on this analysis for investment decisions. The ethical responsibility to ensure the accuracy and clarity of the analysis is paramount to avoid misleading conclusions that could affect financial decisions.

1. **Transparency and Documentation**:

Documenting how the data was collected, cleaned, and analyzed is essential, ensuring transparency in the methodology used. Any assumptions made during the data wrangling or analysis process should be communicated to avoid misinterpreting results.

By addressing these concerns and ethical considerations, the project can contribute meaningful insights into S&P 500 companies while adhering to data handling best practices.