Data engineering Assignment - Term1

S&P 500 Stock Growth and ESG Score Analysis

1. Project Overview

In this university project, I aim to establish a strong data engineering foundation to enable valuable insights from our dataset. By focusing on the relationship between stock performance and Environmental, Social, and Governance (ESG) scores of S&P 500 companies, we plan to organize and prepare the data effectively. This involves calculating stock growth over a specified period and examining potential correlations with ESG risk scores, categorized by industry and levels of controversy. By building this data infrastructure, we make it possible to uncover how ESG factors might influence stock performance across various sectors and risk categories. This groundwork is essential for facilitating meaningful analysis and understanding the real-world impact of ESG considerations on financial outcomes.

2. Data sources:

S&P 500 Stocks (daily updated)

The Standard and Poor's 500 or **S&P 500 is the most famous financial benchmark in the world**. This stock market index tracks the performance of 500 large companies listed on stock exchanges in the United States. As of December 31, 2020, more than **$5.4 trillion** was invested in assets tied to the performance of this index. I used only the opservations between 2022-01-03 and 2024-10-16. Because the index includes multiple classes of stock of some constituent companies—for example, Alphabet's Class A (GOOGL) and Class C (GOOG)—there are actually 505 stocks in the gauge
https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks

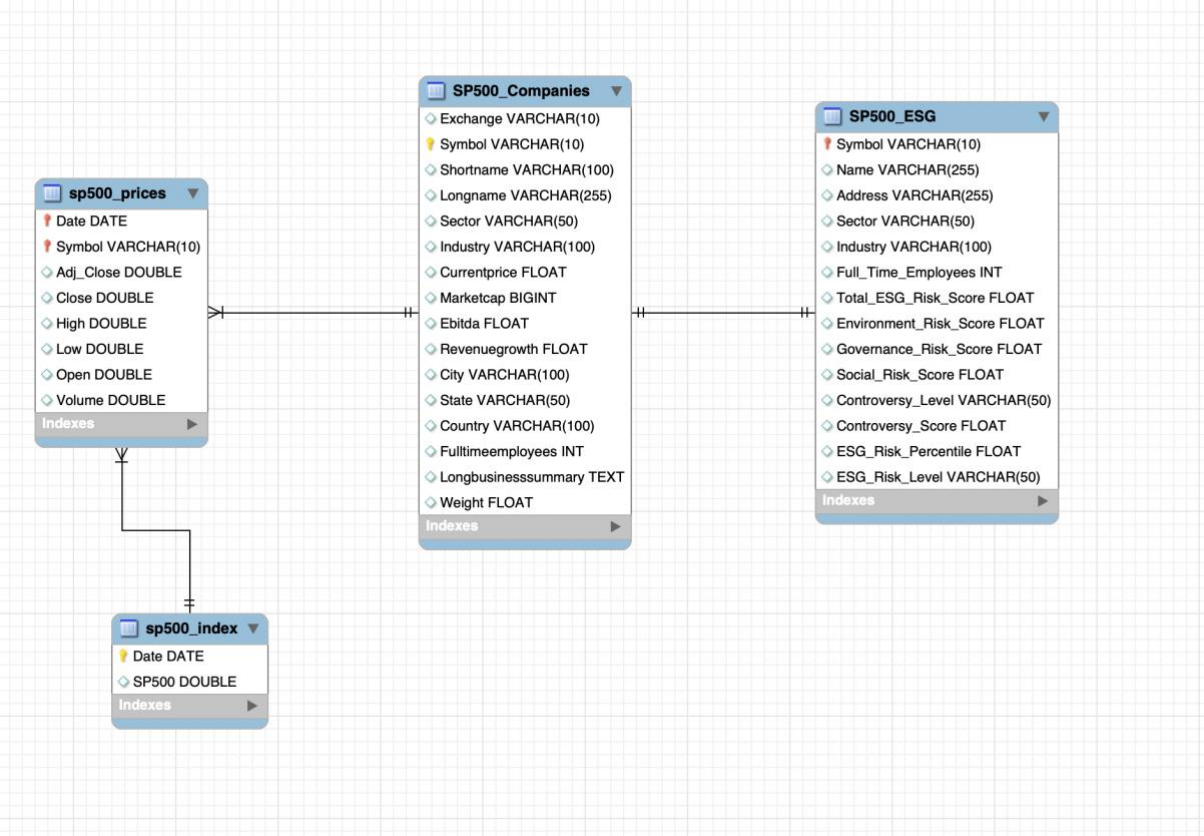
S&P 500 ESG Risk Ratings

This dataset exclusively showcases companies from the S&P 500 index. Researchers, investors, analysts, and policy-makers can utilize this dataset to gain insights into the ESG performance and risk profiles of these major corporations. Whether exploring trends, conducting ESG assessments, or making informed investment decisions, this dataset serves as a valuable resource for comprehending the sustainability and governance practices of S&P 500 companies.
https://www.kaggle.com/datasets/pritish509/s-and-p-500-esg-risk-ratings

3. Database Schema

The project relies on four key tables:

- **SP500_Companies**: Contains general information about each S&P 500 company, including financial metrics like Marketcap, Revenuegrowth, and company details like Sector and Industry.
- **SP500_ESG**: Stores ESG-related information, including Total_ESG_Risk_Score, Environment_Risk_Score, Governance_Risk_Score, Social_Risk_Score, and Controversy_Level for each company.
- **sp500_index**: Holds daily values of the S&P 500 index.
- **sp500_prices**: Contains daily stock price data for each company, including Adj_Close, Close, High, Low, Open, and Volume.



4. Analytical Plan

The project investigates the following primary business questions:

- **Sector-Level Analysis**: How does stock performance vary by sector, and what is the average ESG risk score for each sector?
- **Controversy-Level Analysis**: Is there a correlation between stock growth and controversy level (or other ESG sub-scores)?
- **ESG Scores and Stock Performance**: Does a higher ESG risk score correlate with lower or higher stock growth between 2022-01-03 and 2024-10-16.
- **Stock Price Volatility Analysis:** How does the stock price volatility of S&P 500 companies compare over time, and are there sectors or industries with notably higher volatility trends?
- **Environmental Risk Score vs. Revenue Growth**: Is there a relationship between the environmental risk score of companies and their revenue growth?

5. SQL Queries for Analysis

Calculate Stock Growth Over a Defined Period
The following query calculates the percentage growth in stock price for each company over a specified period:

```
SELECT Symbol, (MAX(Adj_Close) - MIN(Adj_Close)) / MIN(Adj_Close) * 100 AS Percent_Growth
FROM sp500_prices
WHERE Date BETWEEN '2022-01-03' AND '2024-10-16'
GROUP BY Symbol
```

Join Stock Growth with ESG Data by Sector
This query aggregates the stock growth data by sector, calculating the average stock growth and average ESG scores for each sector:

```
SELECT c.Sector, AVG(g.Percent_Growth) AS Avg_Percent_Growth, AVG(e.Total_ESG_Risk_Score)
AS Avg_Total_ESG_Risk_Score, AVG(e.Environment_Risk_Score) AS
Avg_Environment_Risk_Score, AVG(e.Governance_Risk_Score) AS Avg_Governance_Risk_Score,
AVG(e.Social_Risk_Score) AS Avg_Social_Risk_Score
FROM Stock_Growth AS g
JOIN SP500_ESG AS e ON g.Symbol = e.Symbol
JOIN SP500_Companies AS c ON g.Symbol = c.Symbol
GROUP BY c.Sector
ORDER BY Avg_Percent_Growth DESC;
```

Group by Controversy Level for Comparative Analysis
This query examines stock growth and ESG scores grouped by controversy level:

```
SELECT e.Controversy_Level, AVG(g.Percent_Growth) AS Avg_Percent_Growth,
AVG(e.Total_ESG_Risk_Score) AS Avg_Total_ESG_Risk_Score, AVG(e.Environment_Risk_Score) AS
Avg_Environment_Risk_Score, AVG(e.Governance_Risk_Score) AS Avg_Governance_Risk_Score,
AVG(e.Social_Risk_Score) AS Avg_Social_Risk_Score
FROM Stock_Growth AS g
```

```
JOIN SP500_ESG AS e ON g.Symbol = e.Symbol
GROUP BY e.Controversy_Level
ORDER BY Avg_Percent_Growth DESC;
```

Stock Price Volatility Analysis
Calculate Daily Volatility: Compute daily volatility as (High - Low) / Open for each stock in the specified period.

```
CREATE TEMPORARY TABLE Daily_Volatility AS
SELECT Symbol, Date, (High - Low) / Open AS Daily_Volatility
FROM sp500_prices
WHERE Date BETWEEN '2022-01-03' AND '2024-10-16';
```

Aggregate Volatility: Calculate the average volatility over the period for each company and further aggregate by sector or industry.

```
SELECT c.Sector, AVG(v.Daily_Volatility) AS Avg_Volatility, COUNT(DISTINCT v.Symbol) AS
Company_Count
FROM Daily_Volatility AS v
JOIN SP500_Companies AS c ON v.Symbol = c.Symbol
GROUP BY c.Sector
ORDER BY Avg_Volatility DESC;
```

Environmental Risk Score vs. Revenue Growth
Join ESG and Financial Data: Use Environmental_Risk_Score and Revenuegrowth from the SP500_Companies and SP500_ESG tables.Aggregate by Sector (optional): Group by Sector to observe sector-specific trends.

```
SELECT e.Environment_Risk_Score, c.Revenuegrowth, c.Sector
FROM SP500_Companies AS c
JOIN SP500_ESG AS e ON c.Symbol = e.Symbol
ORDER BY e.Environment_Risk_Score;
```

6. Closing Remarks

For a deeper analysis and more refined insights, exporting the aggregated data to specialized tools such as **Python**, **Power BI** or **Tableau** and **Excel**. These tools can provide advanced capabilities like time-series forecasting, regression analysis, and correlation matrices, enabling a more comprehensive understanding of the relationships between stock performance metrics, ESG scores, and sector-based trends.