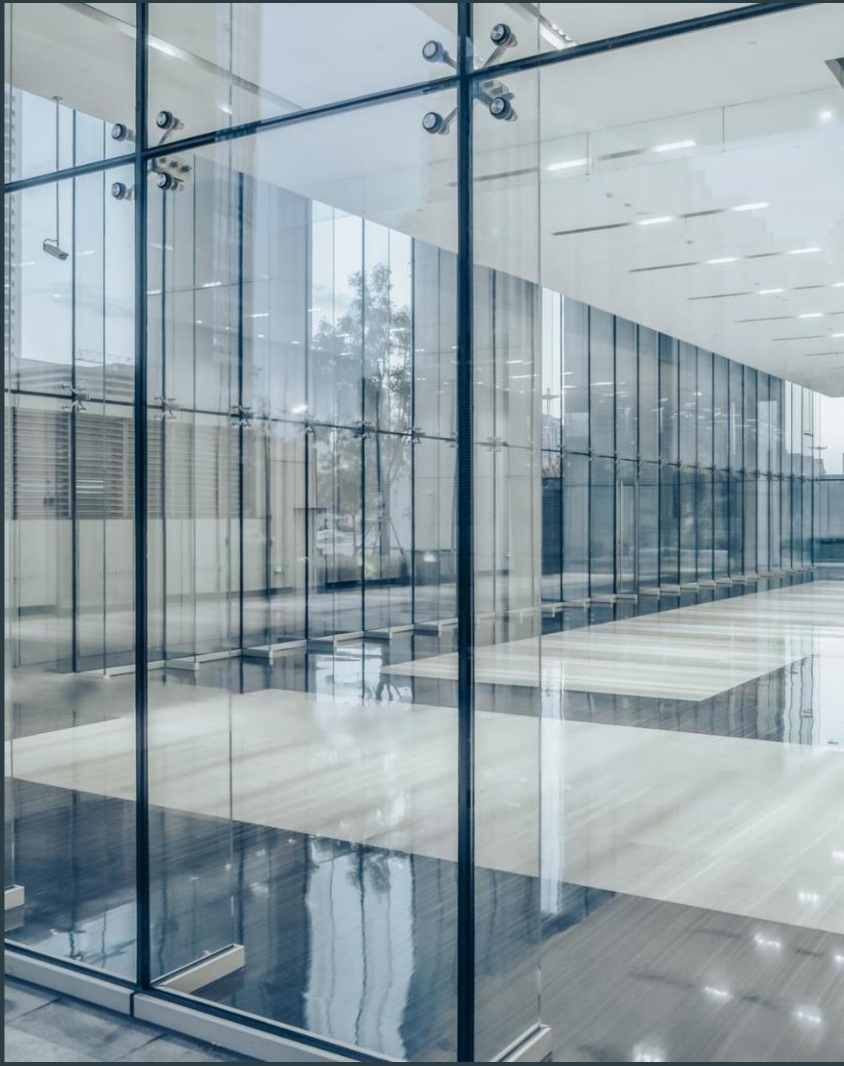
PYTHON / EDA FINAL PROJECT: S&P 500 INDEX ANALYSIS

- Student: Rina Rafalski
- Mentor: Alon Tam
- Program: DRA

NAYA College

Table of contents



Introduction

- The Standard and Poor's 500 or S&P 500 is the most famous financial benchmark in the world. It tracks the largest 500 publicly traded U.S. companies. Investors have long used the S&P 500 as a benchmark for their investments as it tends to signal overall market health. The index is a popular choice for long-term investors who wish to watch growth over the coming decades.
- In this project will be conducted comparative analysis of the S&P 500 index over the last 10 years (2014-2023) by using descriptive statistics and data visualization plots.
- The main goal for this analysis is to answer the Initial Research Questions that aimed to help long-term investors to make informed decisions.

Initial Research Questions

1. The Standard and Poor's 500 or S&P 500 is the most famous What are the trends in stock prices and trading volumes for S&P 500 companies over the last 10 years (2014-2023)?
2. What are the historical performance and growth trends of the S&P 500 index as a whole?
3. Which sectors have shown the most consistent growth?
4. How do specific industries within the S&P 500 perform compared to the overall index?
5. How does the market capitalization and revenue growth of S&P 500 companies correlate with long-term stock performance?
6. What are the top-performing companies in terms of revenue growth and market capitalization?
7. What are the historical volatility and risk profiles of top-performing companies?
8. How does a company’s weight in the S&P 500 influence its long-term returns? Do higher-weight companies tend to perform better?

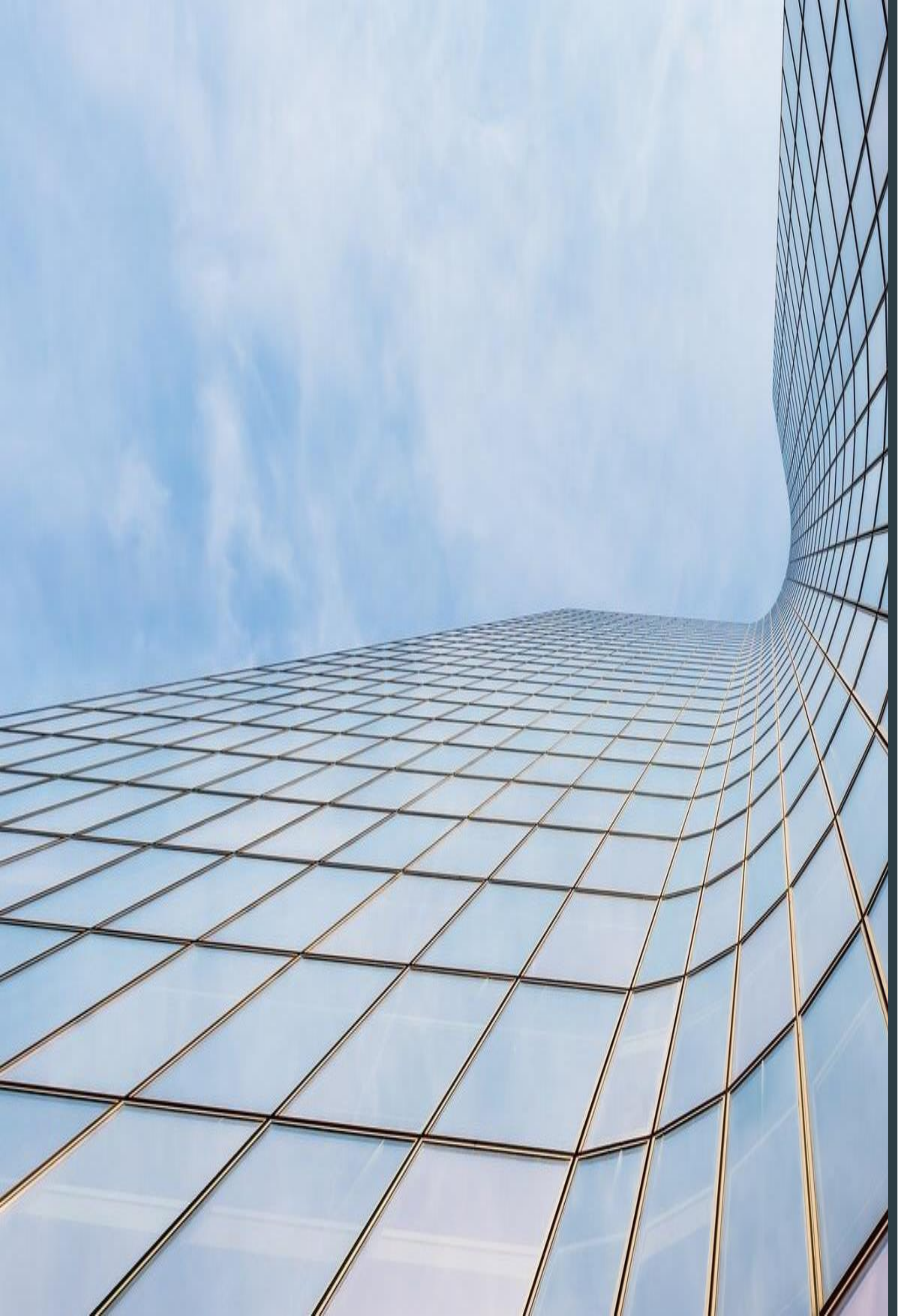
Data description

The source of the data is [S&P 500 Stocks (daily updated)]() dataset that was downloaded from online community platform for data scientists www.kaggle.com.

The dataset contains data from the time period 28/07/2014 – 26/07/2024

The dataset unites 3 subsets, each in separate csv files:

- 1. Companies subset: contains details about S&P 500 companies, including symbol, name, sector, industry, market cap, revenue growth, and other financial metrics
- 2. Stocks subset: contains daily stock prices for each company that included in S&P 500 index
- 3. Index subset: Contains daily S&P 500 index



Data preprocessing

Data preprocessing

Companies Subset

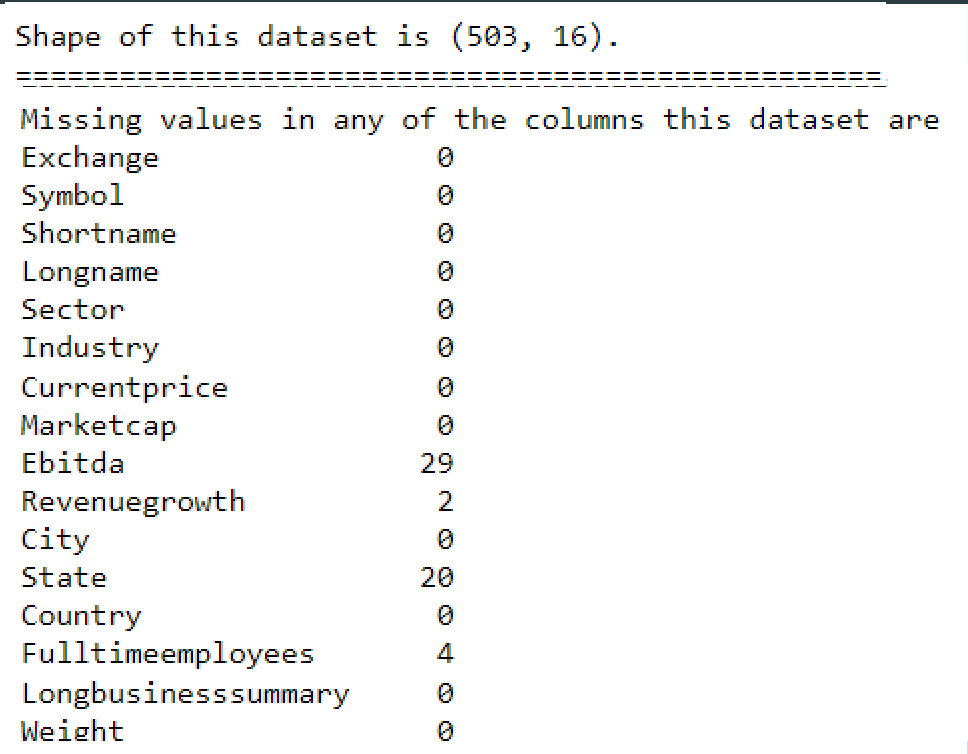
- The shape of this dataframe is 503 rows, 16 columns.
- Columns with null (missing) values
  - Ebitda: 29
  - Revenurgrowth: 2
  - State: 20
  - Fulltimeemplyees: 4

Handling Missing Values

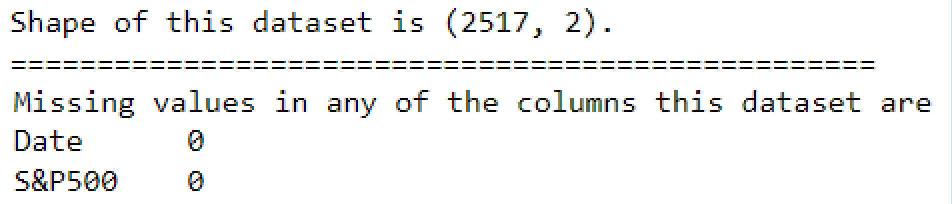
- State: the column has been removed as it was not used for EDA
- Revenurgrowth: 2 missing values were extracted from Yahoo Finance library
- Fulltimeemplyees: 4 missing values was inserted manually according to information from the available companies' annual reports.

SP500 index Subset

- The shape of this dataframe is 2,517 rows, 2 columns.
- The subset has no missing values



```
Shape of this dataset is (503, 16).
==================================================
Missing values in any of the columns this dataset are
Exchange                0
Symbol                  0
Shortname               0
Longname                0
Sector                  0
Industry                0
Currentprice            0
Marketcap               0
Ebitda                 29
Revenuegrowth           2
City                    0
State                  20
Country                 0
Fulltimeemployees       4
Longbusinesssummary     0
Weight                  0
```

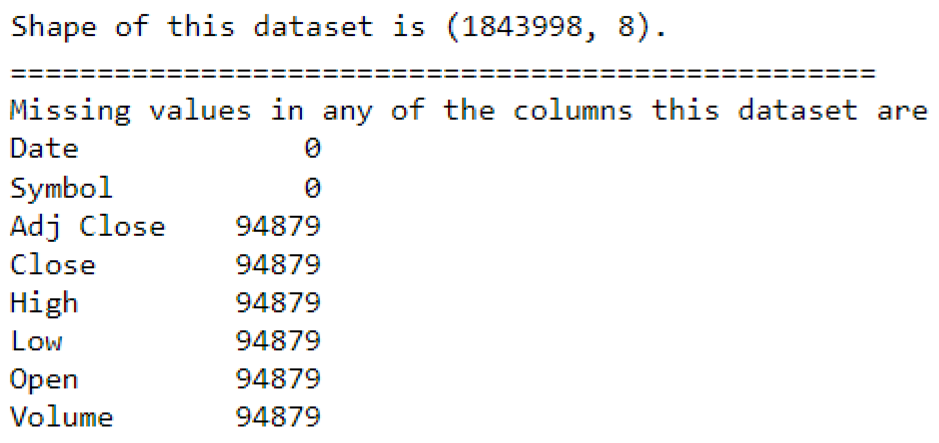


```
Shape of this dataset is (2517, 2).
==================================================
Missing values in any of the columns this dataset are
Date      0
S&P500    0
```

Data preprocessing

Stocks Subset

- The shape of this dataframe is 1,843,998 rows, 8 columns.
- Columns with null (missing) values
  - There are 94,879 rows in the data set that have only Date and Symbol.



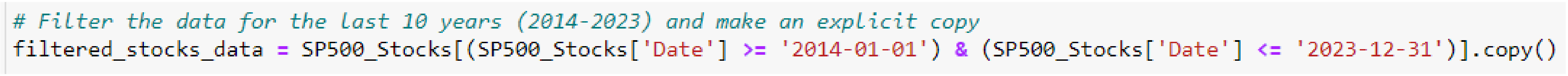
```
Shape of this dataset is (1843998, 8).
==================================================
Missing values in any of the columns this dataset are
Date              0
Symbol            0
Adj Close     94879
Close         94879
High          94879
Low           94879
Open          94879
Volume        94879
```

Handling Missing Values

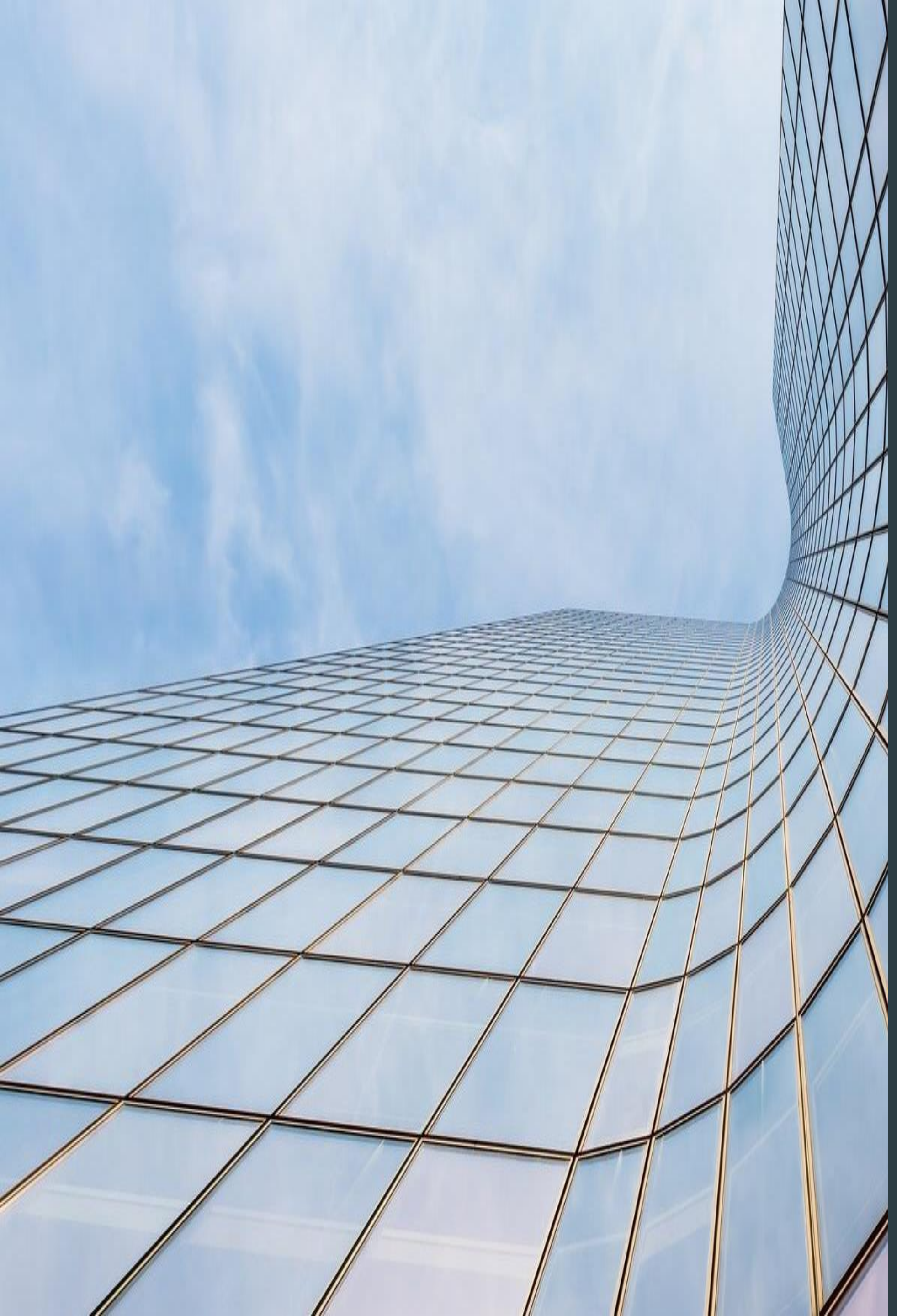
- Since the rows with missing values contain only the date and symbol, and no numerical data, dropping these rows would be an efficient method to maintain the integrity of the data for further analysis. These rows do not contribute any useful information for time series analysis and can be safely removed.

Filtering the data

- The data was filtered for the last decade (2014-2023) :



```
# Filter the data for the last 10 years (2014-2023) and make an explicit copy
filtered_stocks_data = SP500_Stocks[(SP500_Stocks['Date'] >= '2014-01-01') & (SP500_Stocks['Date'] <= '2023-12-31')].copy()
```



Exploratory Data Analysis

EDA - 1.Trends in stock prices and trading volumes

Q1. **What are the trends in stock prices and trading volumes for S&P 500 companies over the last 10 years?**

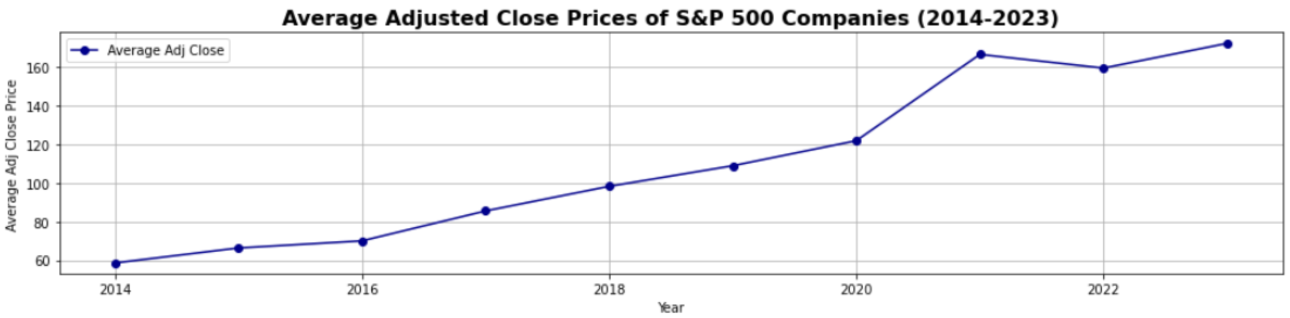
Insights

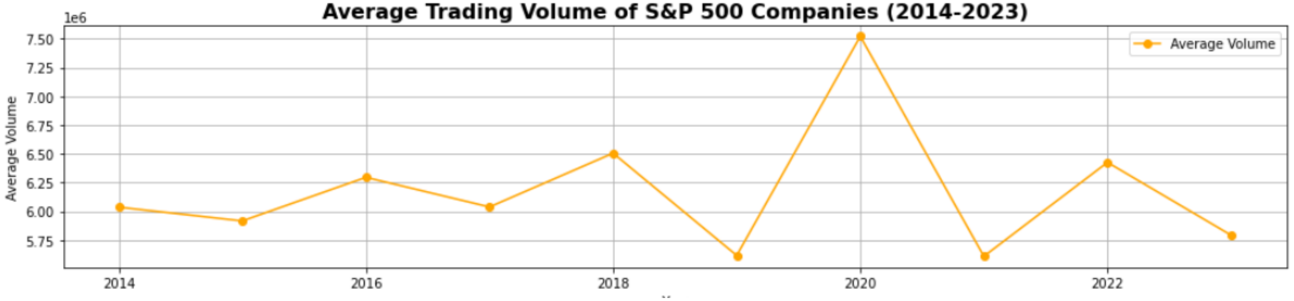
Average Adjusted Close Prices

- There is a clear upward trend in the average adjusted close prices of S&P 500 companies over the last decade.
- The rise is relatively consistent, with notable increases around 2016-2020 and a sharp increase in 2021.
- The data suggests strong growth and recovery, especially in the years following the initial impact of the COVID-19 pandemic in 2020.

Average Trading Volume

- The average trading volume shows more fluctuation rather than a consistent trend.
- The trading volume peaks significantly around 2020, which aligns with the onset of the COVID-19 pandemic.





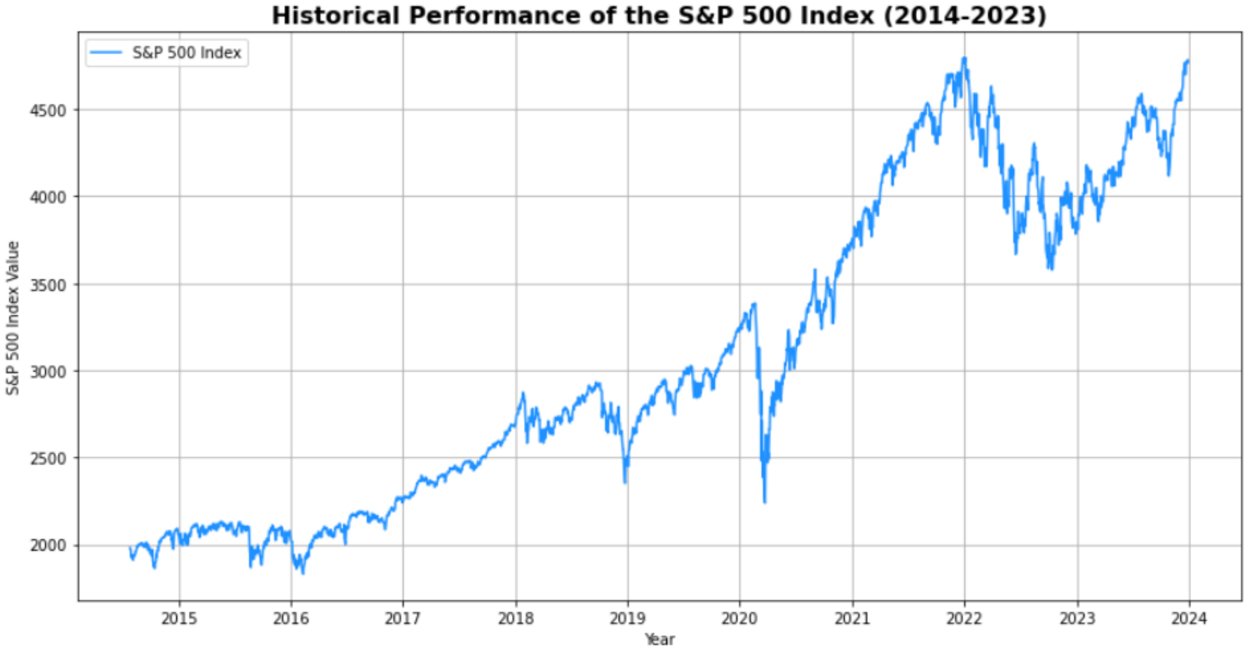
- After 2020, trading volumes seem to fluctuate, with a notable dip in 2021, followed by a peak in 2022, and another dip in 2023.
- This pattern indicates that while the prices were generally rising, the interest and activity in trading S&P 500 stocks varied possibly reflecting market responses to various events.

EDA - 2. Performance of the S&P 500 Index

Q2. What are the historical performance and growth trends of the S&P 500 index as a whole over the last 10 years ?

Insights

- Steady Growth: The S&P 500 index has shown a consistent upward trend over the past decade, indicating overall growth in the market. This growth reflects the increasing value of the companies within the S&P 500.
- **Volatility:** There are noticeable dips in the index, particularly around 2020 (due to the COVID-19 pandemic), and 2022. These dips correspond to periods of market volatility but are followed by recoveries, demonstrating the resilience of the market.
- **Significant Recovery:** The sharp recovery after the 2020 dip is particularly noteworthy, indicating strong market performance in the years following the initial impact of the pandemic.

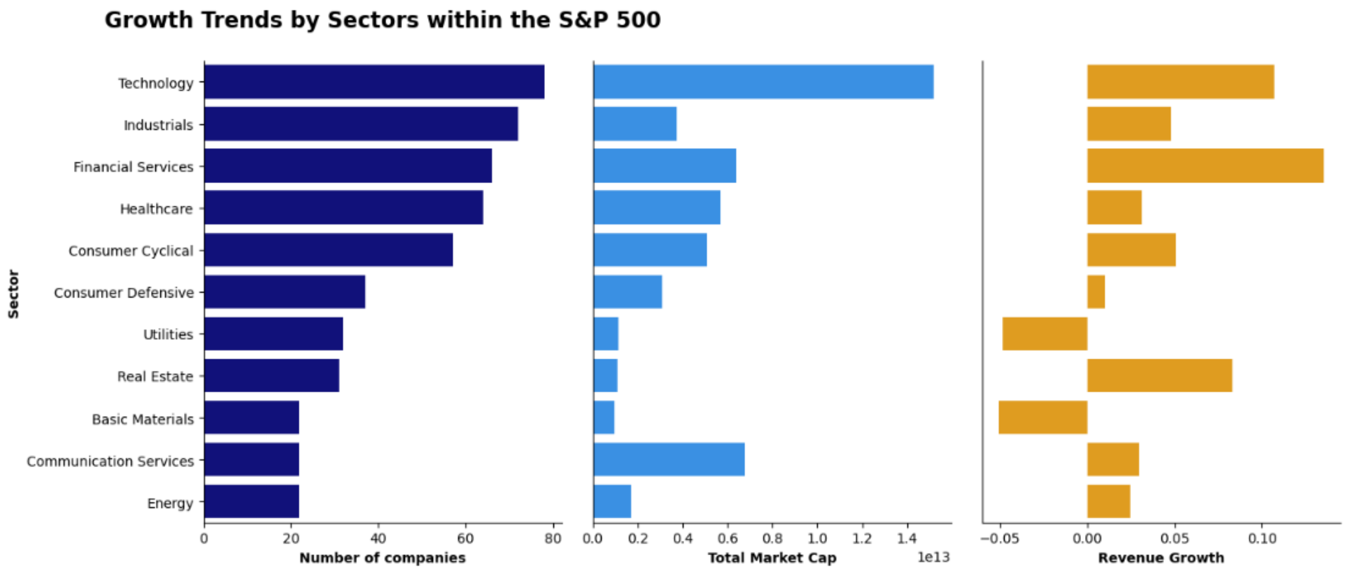


EDA - 3. Growth Trends by Sectors

Q3. **Which sectors have shown the most consistent growth?**

Insights

- Technology sector having the most companies, followed by Industrials and Financial Services.
- The Technology sector stands out with the highest market cap, indicating its significant contribution to the overall S&P 500 market value.
- Sectors like Technology and Industrials have higher revenue growth, while sectors like Utilities and Energy have lower or even negative growth.
- The visualization suggests that there is no correlation between the Number of companies and Market Cap: some sectors (like Technology) have both a large number of companies and a high market cap, while others (like Communication Services) have fewer companies but still maintain a significant market cap.

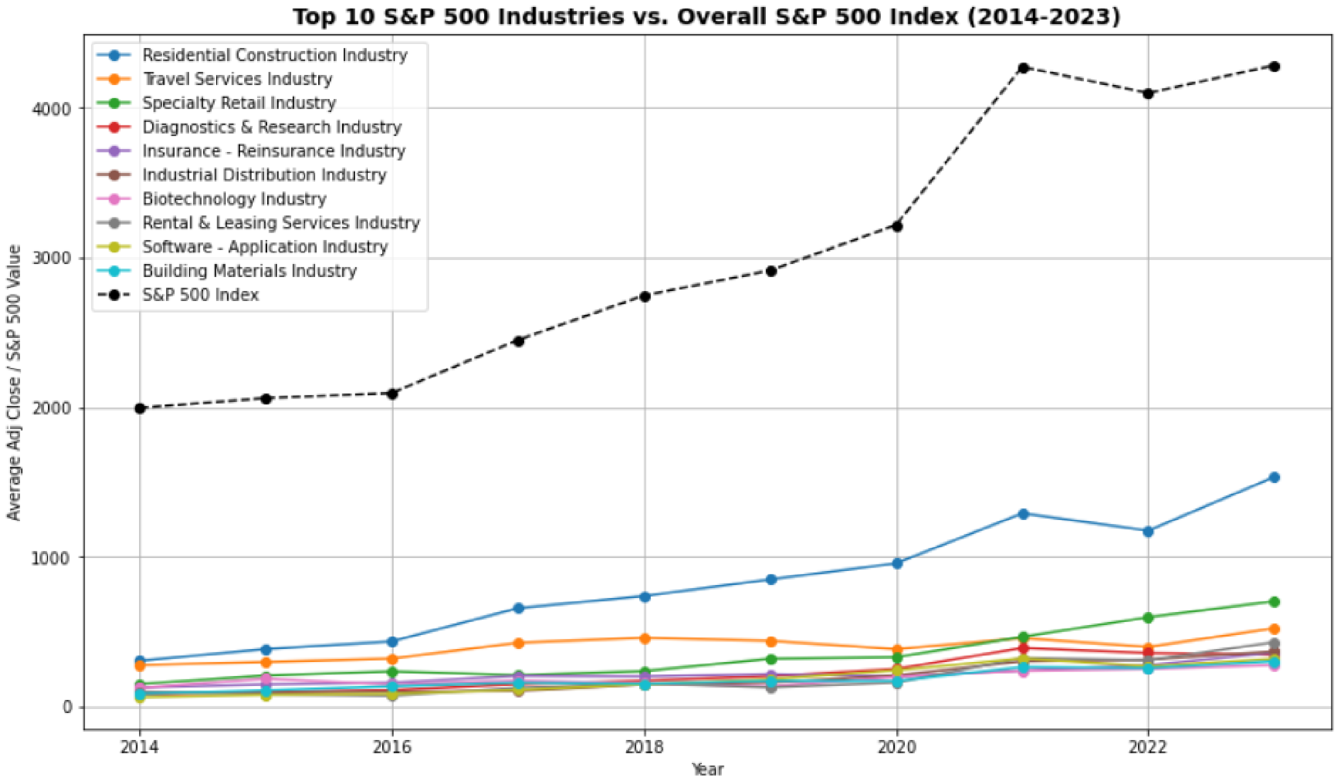


EDA - 4. Industries Performance vs S&P 500 Index

Q4. **How do specific industries within the S&P 500 perform compared to the overall index?**

Insights

- Most industries and the overall S&P 500 index show an upward trajectory, indicating growth over the period.
- While all industries exhibit growth, the rates differ significantly. Some industries, like Travel Services, Residential Construction, and Specialty Retail, experienced rapid growth, others, such as Software Application and Insurance-Reinsurance, grew at a slower pace.
- Residential Construction saw the most substantial growth, with the values increasing significantly compared to the S&P 500 index.
- Several industries, particularly Residential Construction and Travel Services, consistently outperformed the S&P 500 index, suggesting they were less affected by broader market fluctuations.

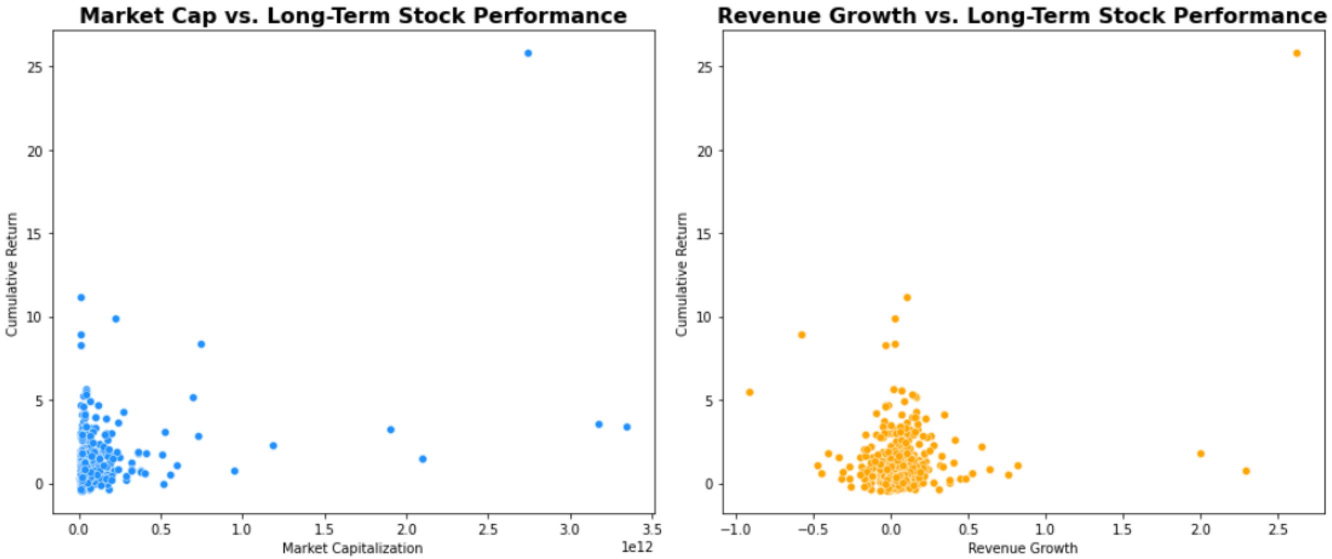


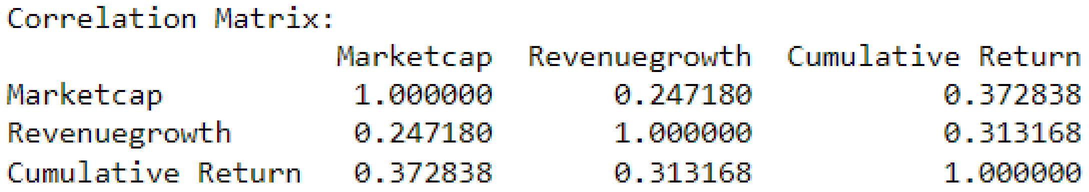
EDA - 5. Market Cap and revenue growth vs long-term stock performance

Q5. **How does the market capitalization and revenue growth of S&P 500 companies correlate with long-term stock performance?**

Insights

- Moderate positive correlations in market capitalization (0.37) and revenue growth (0.31) with long-term cumulative returns. This implies that companies with larger market caps and higher revenue growth tend to have better long-term performance, but these factors are not strongly predictive on their own.
- The scatter plot for market cap shows some outliers with very high cumulative returns despite having smaller market caps. This indicates that smaller companies can still achieve significant growth, making them potential high-risk, high-reward opportunities.





```
Correlation Matrix:
                   Marketcap  Revenuegrowth  Cumulative Return
Marketcap           1.000000       0.247180           0.372838
Revenuegrowth       0.247180       1.000000           0.313168
Cumulative Return   0.372838       0.313168           1.000000
```

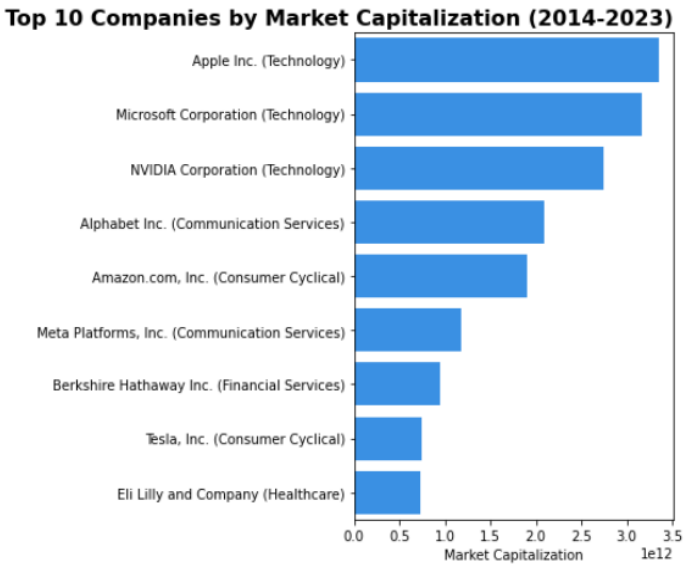
* Calculating the cumulative return for each company over the period (from the first available date to the last) using the adjusted close price.

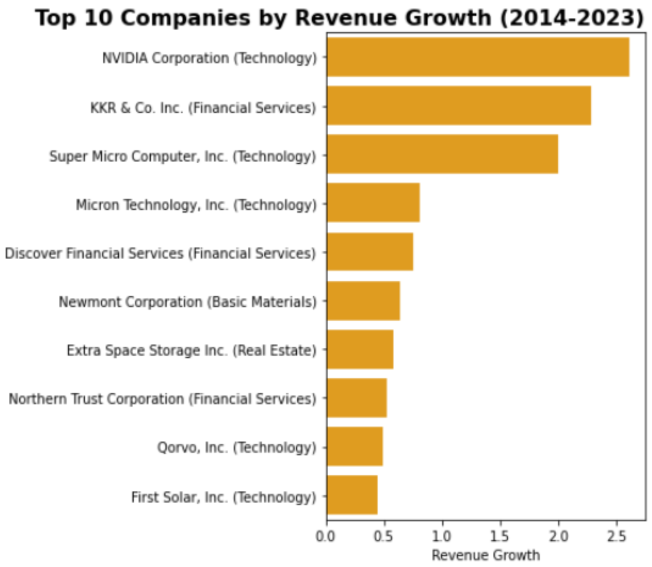
EDA - 6. Top-performing companies

Q6. **What are the top-performing companies in terms of revenue growth and market capitalization for the last 10 years**

Insights

- Companies like Apple, Microsoft, and NVIDIA lead the list, indicating that technology companies have been the most valuable in terms of market cap over the past decade.
- Companies like NVIDIA, Super Micro Computer, Micron Technology showing significant growth that suggests that technology companies not only have large market caps but are also rapidly expanding their revenues.
- Companies like Alphabet (Google) and Meta Platforms (Facebook) rank high in market capitalization, showing the significant role of digital advertising and social media companies.
- Berkshire Hathaway represents the Financial Services sector in the top 10 by market capitalization and indicates the strength and stability of large financial firms in the market.



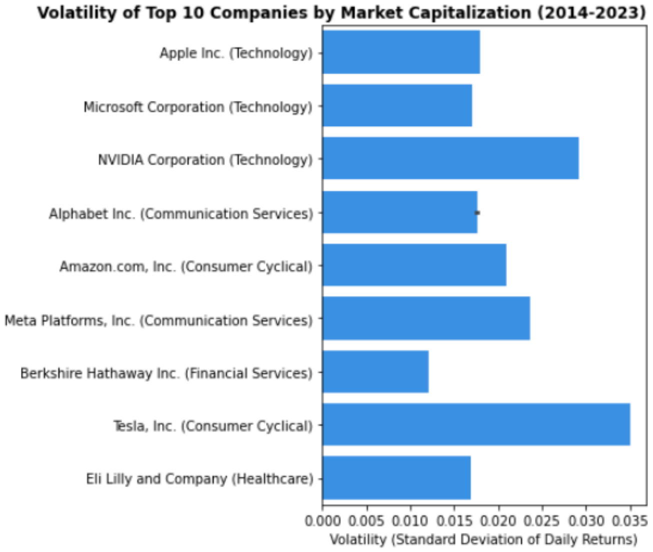


EDA - 7. Risk profiles of top-performing companies

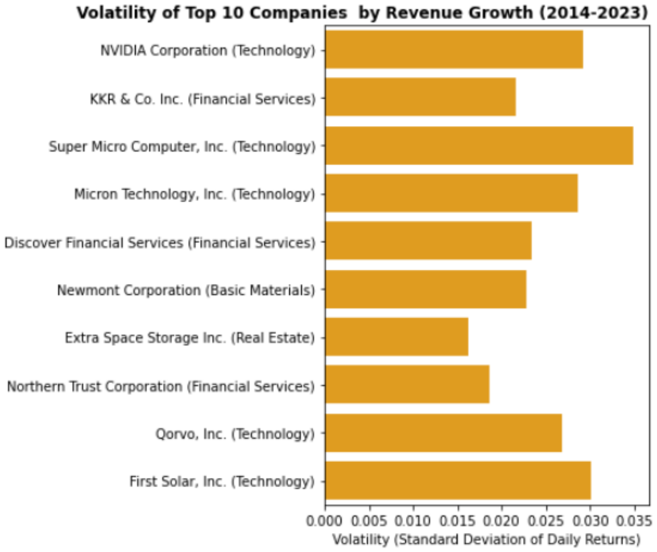
Q7. **What are the historical volatility and risk profiles of top-performing companies?**

Insights

- Volatility in Top Companies by Market Capitalization
- Companies Apple and Microsoft show relatively low volatility, making them potentially lower-risk investments within the technology sector.
- NVIDIA, while also a tech giant, shows higher volatility compared to Apple and Microsoft, which tend to have more market fluctuations.
- Tesla, a leader in a industry of electric vehicles, is more volatile than the other top market cap companies indicating uncertainty or high expectations around innovation and growth.
- Berkshire Hathaway (Financial Services) show lower volatility, indicating more stable stock prices.



Volatility of Top 10 Companies by Market Capitalization (2014-2023)



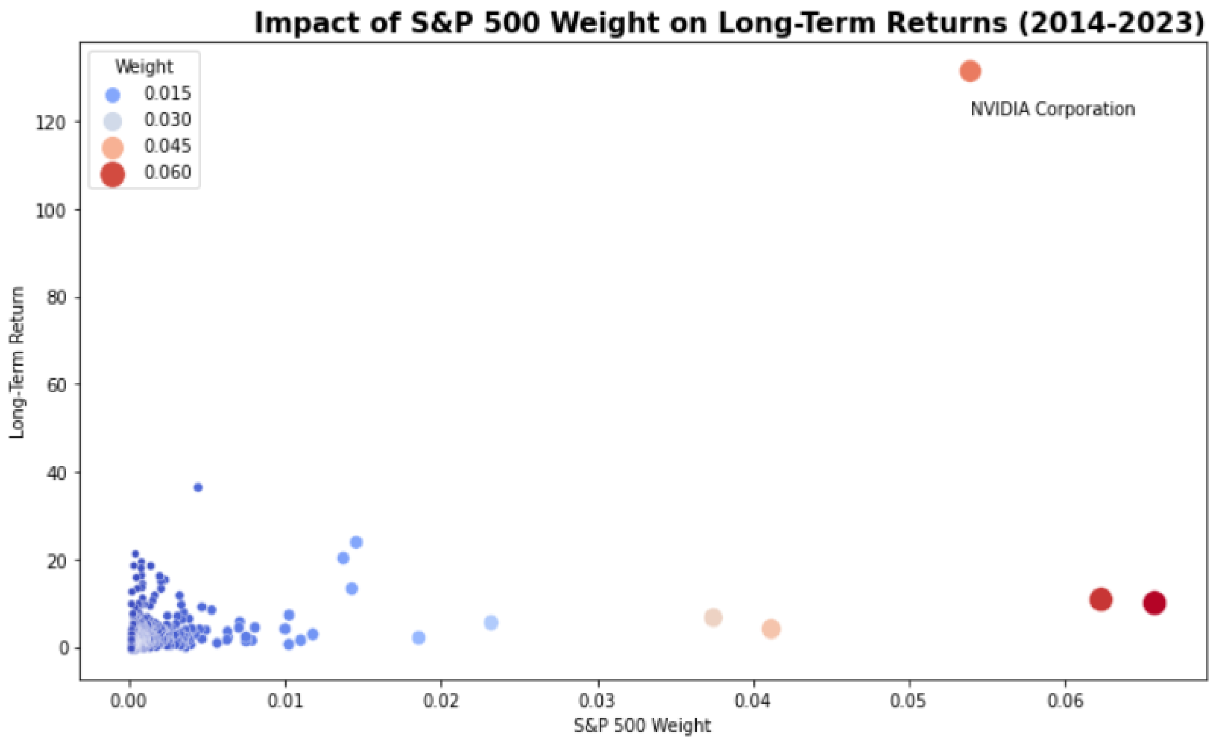
Volatility of Top 10 Companies by Revenue Growth (2014-2023)

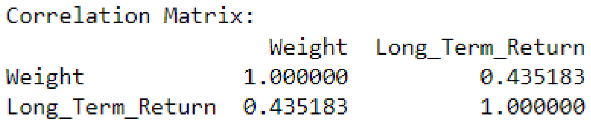
EDA - 8. Impact of Market Weight on Returns

Q8. **How does a company’s weight in the S&P 500 influence its long-term returns? Do higher-weight companies tend to perform better?**

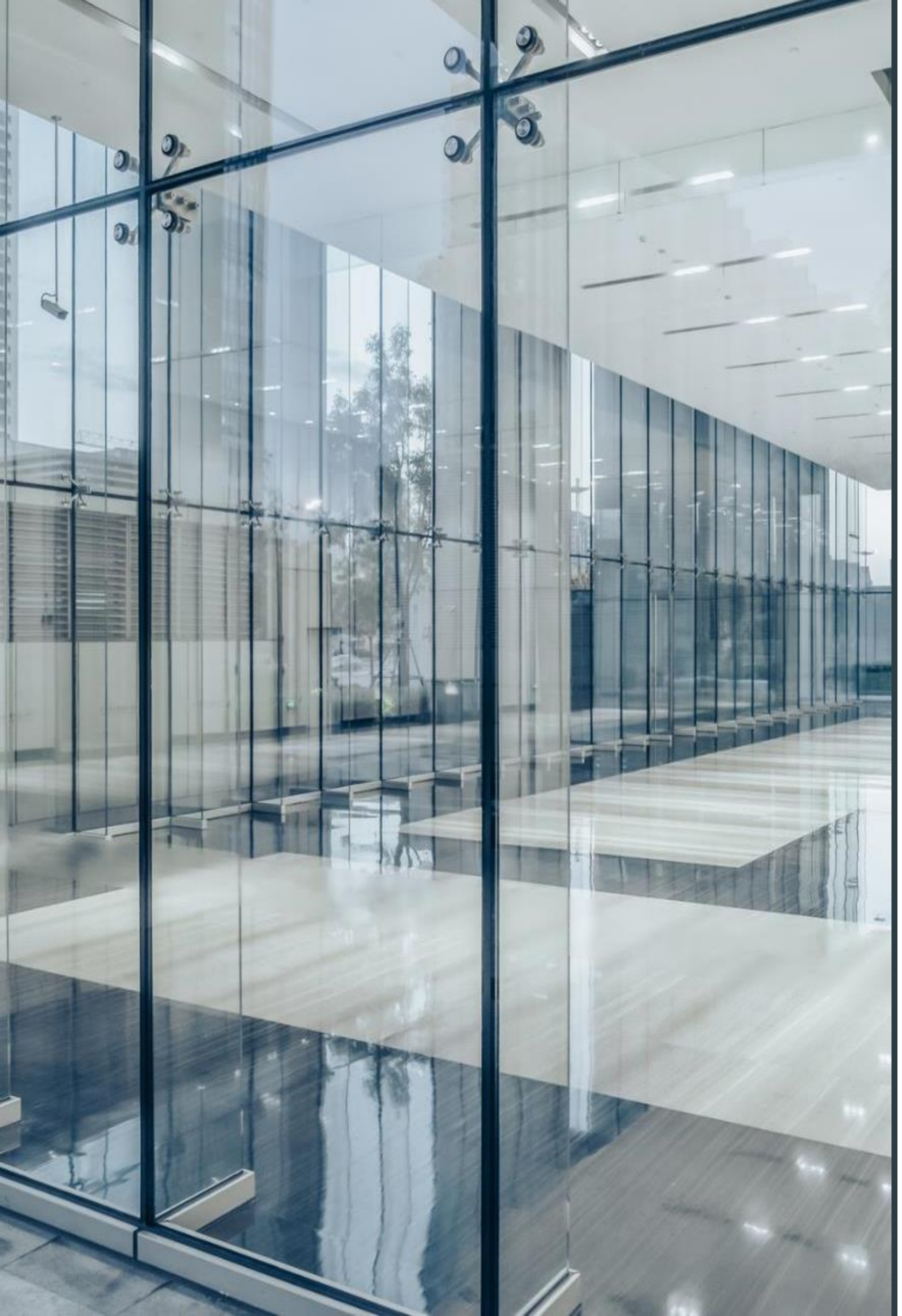
Insights

- The moderate positive correlation (0.4) between S&P 500 weight and long-term returns suggests that as a company's weight in the S&P 500 increases, its long-term returns tend to increase as well.
- The majority of stocks are clustered at the lower left of the plot, indicating that most companies in the S&P 500 have relatively low weights (below 0.02) and lower long-term returns (below 40%).
- Outlier performance: There's a notable outlier, NVIDIA Corporation, with a weight around 0.05 and an exceptionally high long-term return of about 130% significantly outperformed others over the period.
- The relationship between weight and returns doesn't appear to be strictly linear, suggesting other factors likely influence long-term performance.





```
Correlation Matrix:
                    Weight  Long_Term_Return
Weight            1.000000          0.435183
Long_Term_Return  0.435183          1.000000
```



Summary

- The S&P 500 index and its stocks have shown a consistent upward trend in adjusted close prices, reflecting robust growth and recovery, particularly in the aftermath of the COVID-19 pandemic.
- The Technology sector has emerged as the dominant performer, both in terms of market capitalization and revenue growth, highlighting its pivotal role in driving overall market value.
- While most industries within the S&P 500 have shown growth, there are notable differences in growth rates. Industries like Residential Construction and Travel Services have outperformed the broader market, suggesting potential areas for targeted investment.
- The analysis of risk profiles indicates that companies with high market caps, such as Apple and Microsoft, generally exhibit lower volatility, making them potentially lower-risk investments. In contrast, companies like Tesla and NVIDIA show higher volatility, reflecting market uncertainties and high growth expectations.
- Overall, these insights provide valuable guidance for long-term investors looking to make informed decisions based on historical trends, sectoral performance, and risk profiles within the S&P 500.

THANK YOU

Python / EDA Final Project : S&P 500 Index Analysis

CONTACT

in Rina Irene Rafalski
@ rinaraf@gmail.com