201200044 MANAGING BIG DATA PROJECT

ANALYSIS OF FINANCIAL MARKETS

Looking into stock markets as an investment option

AUTHORS

S3294870 EMMA CAÑAVATE QUERO

S2112132 ERIC WAN

S2261294 MIGUEL DE LA CRUZ CABELLO

S3053385 SIMONE GUALANDI

DATE

31ST JANUARY 2024

UNIVERSITY OF TWENTE.

1. INTRODUCTION

In today's financial world, the stock market holds a central position, playing a crucial role in the global economy. The stock market serves as a dynamic platform where investors, traders and entities participate in buying and selling securities. This interaction creates the paths of businesses and has an important impact on economic trends.

The development of stock markets was fueled by the rise of joint-stock companies [1]. These entities, which allowed investors to share ownership through the purchase of shares, became a popular mechanism for pooling capital to fund large-scale ventures, such as exploration and trade expeditions. Over time, the idea of buying and selling shares evolved, leading to the establishment of more organized and regulated stock exchanges. As of September 2023, the top five exchanges by market capitalization are the NYSE, NASDAQ, the Shanghai Stock Exchange in China, Euronext, and the Japan Exchange Group.

The understanding and analysis of stock markets requires a comprehensive approach. Investors and analysts employ a wide variety of tools and indicators to make informed decisions. Monitoring stock markets involves different examinations through key indicators such as stock prices, trading volumes and market indices. Price-earnings ratios, earnings per share, and dividend yields are some of the fundamental metrics that contribute to understanding the financial evolution of listed companies. Moreover, technical analysis, which involves studying historical price patterns and market trends, is another prevalent method for market overview. In addition, macroeconomic indicators, including interest rates, inflation rates, and geopolitical events, are vital components in the economic context that influences stock market movements.

Dealing with limited historical data is a big deal when it comes to analyzing the stock market. However, understanding how markets behaved in the past is crucial, but not all exchanges have equally good records. Major exchanges like the New York Stock Exchange (NYSE) or NASDAQ usually have a lot of historical data, making it easier to study past market trends. On the other hand, smaller or newer exchanges might not have as much data to work with, what makes them harder to analyze, comprehend and predict. This limitation underscores the importance of focusing analytical efforts on major exchanges, where a wealth of historical data allows for a better and reliable insight. For this reason, this report has focused on five markets: NYSE, NASDAQ, SP500, Forbes2000 and the Indian stock market.

In the next two sections, we delve into some background information and related work, offering insights into existing research and methodologies within the domain. Following this, the methodology section outlines the approach taken for data preparation, emphasizing the steps involved in collecting, processing, and organizing the vast datasets essential for stock market analysis. Subsequently, we present the results of our analyses, unveiling the patterns, trends, and insights derived from the application of big data techniques to stock market data.

2. BACKGROUND INFORMATION

2.1 ADJUSTED PRICE

The adjusted closing price is a calculation adjustment made to a stock's closing price [2]. The original closing price is the final price at which a stock is traded at on a specific day. However, the original closing price does not represent the most accurate valuation of the stock or security since it will not account for any actions that could have caused the price to shift. Therefore, an adjusted closing price will include any adjustments that need to be made to the price.

2.2 VOLATILITY

Volatility is commonly measured using the standard deviation, and historical volatility is often calculated based on the returns of a financial instrument such as a stock over a specific period. The general formula for calculating historical volatility using standard deviation is as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2}{n-1}}$$

Where Xi is each individual return, $\overline{X}$ is the average daily return, and n is the total number of daily returns. When calculating the volatility, a high or a low value can be obtained. High volatility is often associated with increased market risk and uncertainty. Prices can suffer quick and unpredictable movements, making it challenging for investors to accurately predict market trends. In contrast, low volatility indicates a rather stable market, i.e., fluctuations in stock prices are relatively consistent.

2.3 VOLUME-PRICE CORRELATION

One way we can find constructive price patterns is by identifying those in which volume rises on rallies, falls or declines [3]. The correlation between price and volume development is considered an important indicator. It is often used alongside trend analysis and other indicators based on the stock price.[4]

The theory is that when price and volume move together, the stock is acting correctly. Therefore, the higher the correlation, the better acting the stock is. If the correlation was strongly negative, we may see a constructive short forming [4]. A negative correlation, where prices move in the opposite direction of trading volume, can be a cautionary signal. It may indicate that significant price movements are not supported by widespread market participation. Moreover, for example, if prices are rising, but volume is decreasing, it could suggest weakening market conviction and the potential for a reversal.

Positive correlation between price development and volume development is a sign of strength. A positive correlation between price and volume suggests that significant price movements are accompanied by higher trading volumes. This often serves to confirm the strength and sustainability of a prevailing trend. In technical analysis, a surge in both price and volume can be indicative of a potential breakout. This occurs when a stock's price moves decisively beyond a key level, and the accompanying increase in volume suggests strong market conviction. Conversely, a reversal in the prevailing trend may be confirmed by a decrease in price accompanied by a decline in trading volume. This can signal a loss of market interest and a potential shift in sentiment.

There are different factors that can influence the correlation between price and volume. The first cause can be the different news and events. These can affect both price and volume. Unexpected developments may lead to a surge in volume and price volatility. Another cause can be earnings reports, since they often coincide with increased trading activity. Positive or negative surprises in earnings can result in significant price movements accompanied by elevated volumes. Additionally, investor “sentiment” plays a crucial role. High volume during price increases may indicate bullish sentiment, while high volume during price declines may suggest bearish sentiment.

2.4 ANNUAL RETURN

The annual return expresses a stock’s increase in value over a designated period. Information regarding the current price of the stock and the price at which it was purchased is required to calculate it [5]. The purchase price must be adjusted accordingly if any splits have occurred.

The simple return percentage is calculated first when the prices are determined, with that figure ultimately being annualized. The simple return is the current price minus the purchase price, divided by the purchase price.

3. RELATED WORK

This chapter provides an overview of some existing knowledge in managing big data for stock analysis and other information about stock market prediction and analysis. This section delves into relevant studies, scholarly articles, and established practices to establish a contextual framework for this research. By examining the literature, the aim is to identify key themes, methodologies and findings that have shaped the understanding of big data applications in stock market analysis.

The paper from Peng [6] focuses on understanding how investor sentiment impacts stock market volatility, utilizing the volatility decomposition theory proposed by Pollet and Wilson. Through a big data strategy, it conducts a comparative analysis involving web news emotion index, web search volume, social network emotion index, and social network heat index. After correlation and Granger causality tests, significant indicators correlated with the financial market are identified for forecasting analysis. The constructed market volatility index, based on investor sentiment, exhibits a strong predictive ability for stock market volatility turning points, particularly anticipating declines one to two days in advance. The study introduces practical insights for stock market volatility prediction and aids in financial market risk aversion.

The research from Wu [7] focuses on the microstructure of the financial market, specifically dealing with the substantial amount of data related to individual trades and multiple levels of quotes. Analyzing such data is computationally intensive, often posing challenges for financial academics and regulators. However, leveraging tools from data-intensive scientific research, the study demonstrates the effectiveness of various techniques for market data analysis.

3.1 RESEARCH QUESTIONS

RQ 1: What is the correlation between trading volume and stock price fluctuations over time in major stock exchanges?

Our primary research question is about the correlation between trading volume and stock price fluctuations across major stock exchanges. The essence of our investigation lies in understanding how significant changes in trading volume in each period, including both buying and selling activities, correspond to fluctuations in stock prices.

RQ 2: In which historical time periods have financial markets experienced the highest levels of stock price volatility and to which events can they be linked?

The second research question explores the historical time periods when financial markets experienced the highest and lowest levels of stock price volatility. This question naturally emerges from our exploration of trading volume dynamics, as we aim to pinpoint distinctive historical epochs marked by significant fluctuations in stock prices.

RQ 3: In what extent volume-price correlation and annual return ratio can lead to investment decisions?

This research question into the relationships between volume-price correlation and annual return ratio, and their collective impact on investment decisions. By analyzing the volume-price correlation table that highlights the most consistent performers, we seek to identify patterns that signal stability and potential growth. Simultaneously, this research considers the insights revealed by the annual return ratio, aiming to establish a link between these metrics and overall investment viability.

The objective is to provide a nuanced understanding of how these indicators interact, potentially attracting investors toward more astute and data-driven strategies in the financial market.

4. METHODOLOGY

4.1 DATASETS

The search for a suitable dataset involved looking at stock market datasets from Kaggle. Among all of them, the chosen dataset is the most comprehensive and reliable and included key U.S. stock exchanges and indexes such as the Forbes 2000, NYSE, NASDAQ, and S&P 500 [8]. While other Kaggle datasets were considered, the chosen dataset stood out because of its high level of maintenance and low error frequency, deduced from the discussion section.

To enable a more global analysis of stock market trends, a dataset from another country's stock market was chosen. Despite its incompleteness, a dataset of the Indian Stock Market, containing data for only a few years, was selected [9].

4.2 DATA UNDERSTANDING

Both datasets are in .csv format, offering a structured and easily accessible form of data. Upon closer inspection, it becomes apparent that while the names of the headers in each dataset differ, they essentially denote the same categories of information.

A notable difference can be seen when examining the precision of the value entries. The dataset associated to the U.S. market displays a higher level of detail, with very specific decimal values. In contrast, the Indian market dataset has a lower level of detail, limiting its entries to just two decimal points. Though this difference in data granularity is apparent, it will not have a significant impact on the processing and analysis of the data.

Date	Low	Open	Volume	High	Close	Adjusted Close
15-12-1980	0.121652	0.122210003	175884800	0.122210003	0.121652	0.094663337
16-12-1980	0.112723	0.113280997	105728000	0.113280997	0.112723	0.087715253
17-12-1980	0.115512997	0.115512997	86441600	0.116071001	0.115512997	0.089886256

Table 1: CSV-Snippet from the Apple stock

Furthermore, there is a difference in the frequency of data updates. The real-time ticker for the Indian stock market operates at a much higher frequency, with data entries being updated at one-minute intervals.

timestamp	open	high	low	close	volume
2017-01-02 09:15:00+05:30	340	340	340	340	11
2017-01-02 09:16:00+05:30	340	340	340	340	0
2017-01-02 09:17:00+05:30	340	340	340	340	0

Table 2: CSV-Snippet from the Aarti Industries Ltd stock

Due to the similar and clean structure of the datasets, extensive data preparation steps were skipped.

4.3 DATA PROCESSING

Our main approach is to process and transform the data before constructing our analyses was the following:

- For each file path, we read the CSV into a DataFrame and select the relevant columns needed for each analysis.
- Perform transformations to the data to extract the year from the date format and derive the company name from the file path and add them to new columns.

- Create various windows for aggregating data by company and year. This way we perform the calculations for each company separately.

Volatility

The volatility represents the fluctuations in the price of a stock over a specified time. To calculate this, all available data of each stock exchange/index was utilized. In the following, the rough procedure is described.

- The daily return for each stock was determined by dividing the difference between today's closing price and yesterday's closing with yesterday's closing price.
- The average of the daily returns for each company and for each year were calculated.
- The squared difference between the daily return and the average daily return were determined.
- Lastly, the result of the square root of the squared difference were grouped by year and rounded to two decimal places.

Volume-Price Fluctuation Correlation

Correlation can be understood as a coefficient measuring the linear relationship between two variables, ranging between -1 and 1. In this context, the correlation between volume and price fluctuations can provide insights about how volume changes are related to movements in the price of a stock. Our approach was the following:

- For each file path we read the CSV into a DataFrame and select the "Date", "Open", "Close", and "Volume" columns.
- For each row we create a column "Price Fluctuation" and subtract the opening price from the closing price.
- Then we group the data by company and year and perform a correlation using the pyspark corr function.

$$Correlation = corr("Volume", "Price\ Fluctuation")$$

- Next, we rank the companies based on their correlation for each year, and the top 50 companies are selected to keep track.
- We identify the 10 companies that appear the most frequently in the top 50 across years.
- And lastly, we convert the "top companies" DataFrame into an RDD containing the collection of those top 10 companies and download CSV files with the yearly correlation of each of those 10 companies.

Annual Return

As explained above, annual return can be understood as the percentage change in a stock price over a one-year period. Our approach to calculate the annual return was the following:

- For each file path, read the CSV into a DataFrame and select the "Date" and "AdjustedClose" columns.
- To begin with, we identify the first and last adjusted close prices of a stock for each company for each year. Here we use the adjusted close price as it accounts for factors like dividends or stock splits.
- We then create a new column for the annual return. It is calculated by finding the percentage change between these two prices.

$$Annual\ Return = \left(\frac{col(LastAdjustedClose) - col(FirstAdjustedClose)}{col(FirstAdjustedClose)}\right) * 100$$

- Once we calculate the annual returns, we rank the companies based on their annual returns for each year, and the top 50 companies are selected are selected to keep track.
- We identify 10 companies that appear the most frequently in the top 50 across years.
- Lastly, we convert the "top companies" DataFrame into an RDD containing the collection of those top 10 companies and download CSV files with the yearly annual returns of each of those 10 companies.

5. RESULTS

5.1 VOLATILITY

Figure 1 illustrates the volatility of different stock market exchanges and indexes in a yearly fashion. Next to the name of the stock market/index, which can be found on the left side of the graph, is a numerical scale ranging from 0 to a maximum of 10%. Please note that the scale is adjusted for each stock separately such that the trends can be more easily detected.

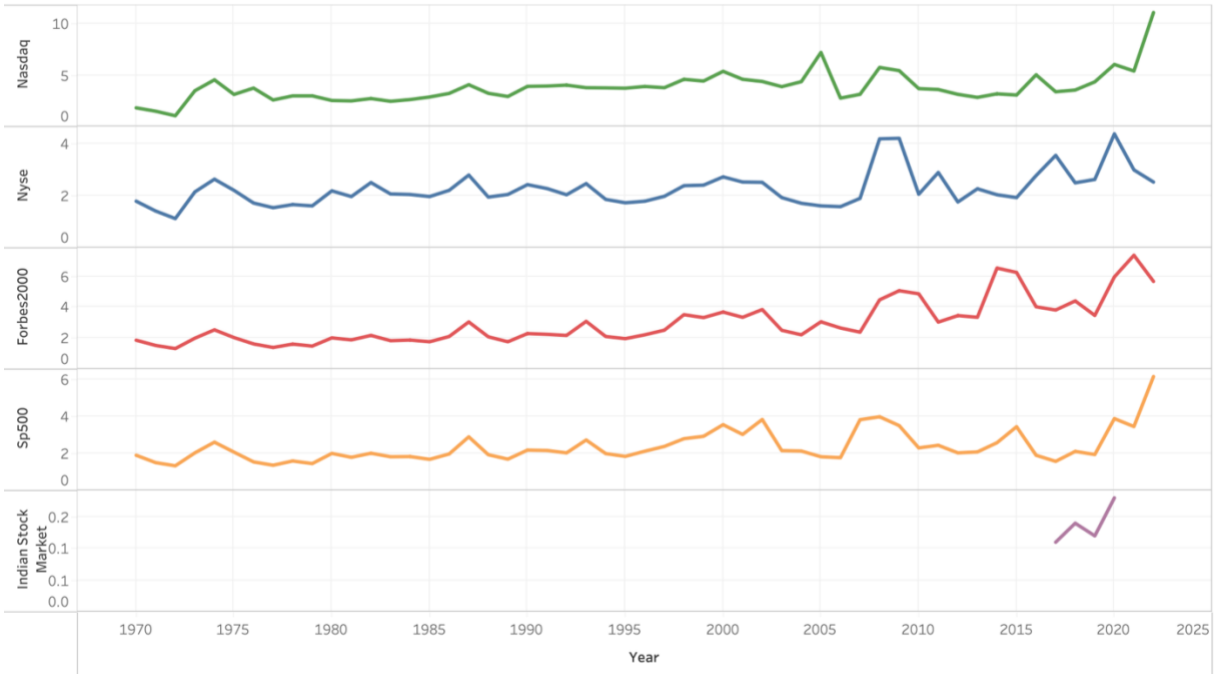


Figure 1: Volatility of Indian & US Stock Markets/Indexes

5.2 COMPANIES' CORRELATIONS AND RETURN ANALYSIS APPEARANCES AND RANKS

Figures 2, 3, 4, and 5 illustrate the most resilient companies throughout the years, as they appeared on the top 50 companies with highest volume-price fluctuation correlations and return analysis more times than any other companies.

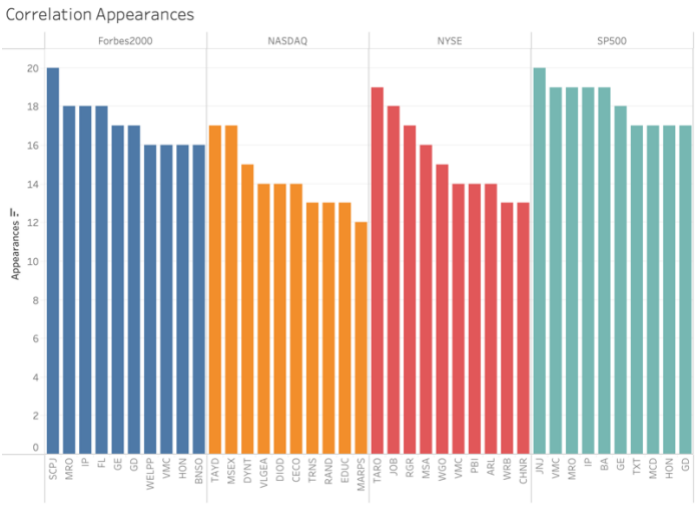


Figure 2: Correlation appearances for top 10 companies in US Stock Markets/Indexes

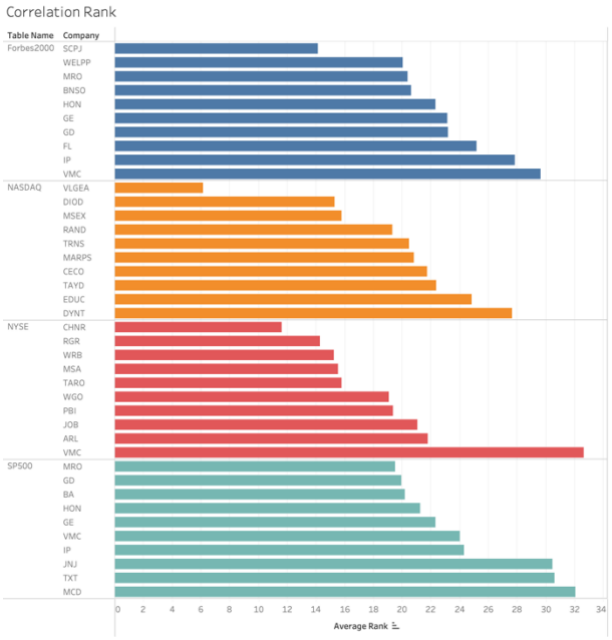


Figure 3: Correlation average ranks for top 10 companies in US Stock Markets/Indexes

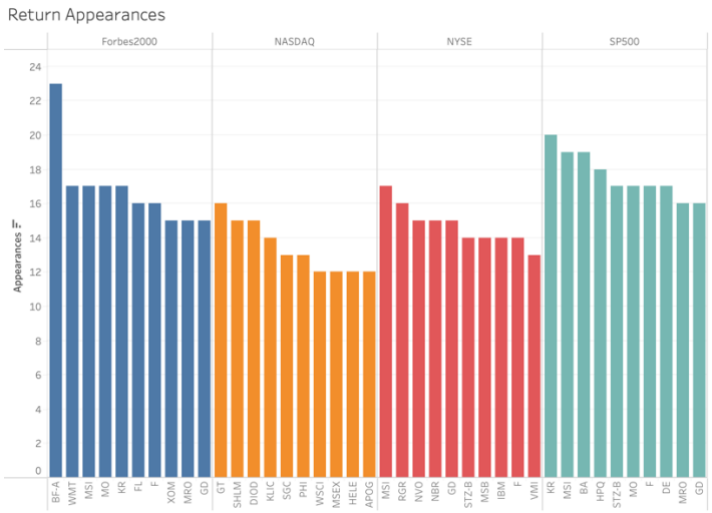


Figure 4: Return appearances for top 10 companies in US Stock Markets/Indexes

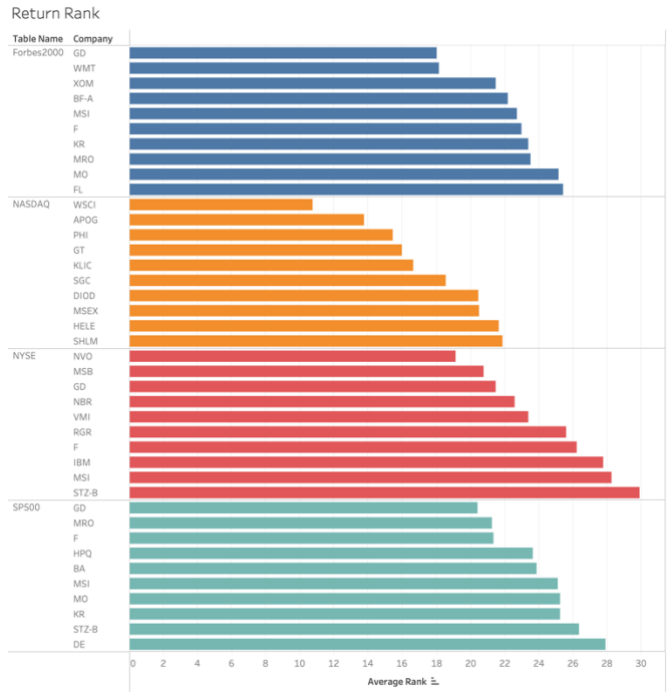


Figure 5: Return average ranks for top 10 companies in US Stock Markets/Indexes

6. CONCLUSION

6.1 RQ 1: VOLUME-PRICE CORRELATION

After conducting an in-depth analysis of the correlation between stock prices and related trading volumes, we have pinpointed a select group of companies that emerge as particularly reliable and resilient over time. This group of companies not only consistently exhibits a strong correlation between price fluctuations and volume traded but also shows a robust ability to withstand market volatility and maintain a stable trajectory. These firms, previously highlighted in the histograms section, provide insight into the traits that mark lasting success in the financial markets.

6.2 RQ 2: HISTORICAL VOLATILITY ANALYSIS

Throughout history there have been extremely meaningful volatility spikes across major financial markets. To mention a few, three different peaks in volatility and their corresponding events will be elaborated in the following.

BLACK MONDAY (1987)

The 1987 stock market crash in the United States was in large part blamed on 'program trading', the first technology/financial engineering-driven crash of its kind. However, massive speculative excesses were built up prior to the crash. This played a significant role in the decline of stock prices and the massive spike in volatility.[10]

Even though the day when the big decline in major averages happened came as a surprise, as has been the case with most other major volatility cycles, it didn't exactly come out of nowhere.

Once again, before the peak, the volatility was growing in an uptrend. In no time did the volatility spiked in the wake of the Monday collapse in stock prices before easing off and eventually dropping off in the following months. Finally, the market stabilized again.

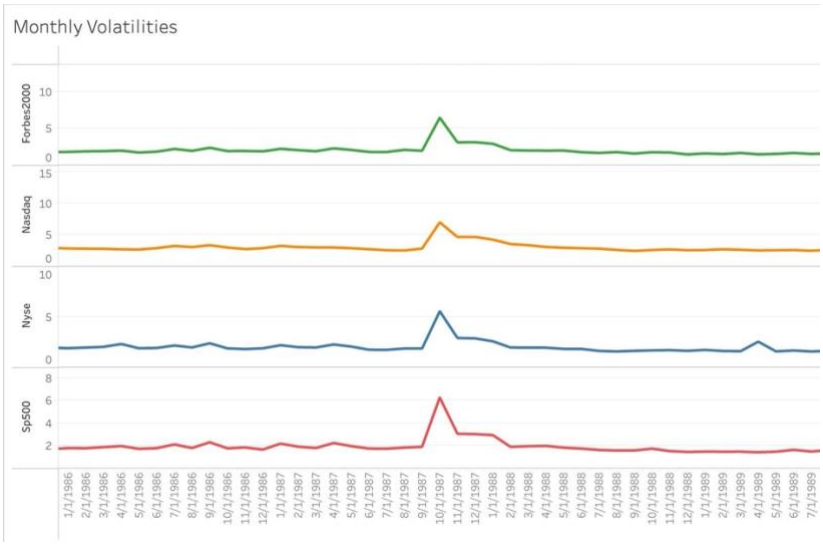


Figure 6: Volatility of US stock exchanges/indexes during 1986 - 1999

GREAT FINANCIAL CRISIS (GFC, 2008)

The Great Financial Crisis was driven by irresponsible banking practices on Wall St. The decline from 2007 to 2009 was the largest plunge in both stocks and the economy since the Great Depression, but it did not come without some type of warning that a major blow-up in volatility could be in the works [10].

Just before the big pic that can be seen in the figure below during the fall of 2008, two-week volatility was already at 41%. From there, the S&P 500 fell another 27% in about five weeks. In the year following, volatility normalized with two weeks. By analyzing the year 2007, the volatility was generally higher before the big spike in 1008. Once the panic hit and the market confidence came back, volatility died down.

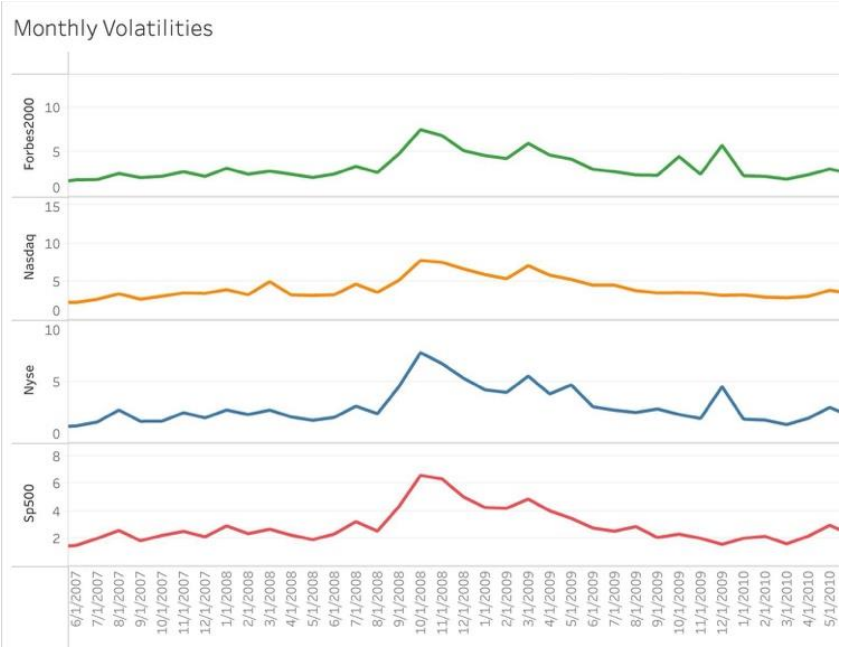


Figure 7: Volatility of US stock exchanges/indexes during 2007 - 2010

COVID-19 (2020-2022)

Another historical event that due to the pandemic's impact, investors and markets face a high degree of uncertainty regarding both the physical and financial. The latest COVID-19 outbreak results in extraordinary volatility around all the leading financial markets. [11] Starting from February 2020, several shock waves were noticed in stock markets due to the current outbreak. This can be seen in the figure below, where different peaks during the period that the pandemic lasted. Different build ups and deflations are built during the time. This could be because of the rapid and unexpected spread of the virus and the uncertainty among investors. Moreover, during the virus's effect, many countries implemented strict lockdowns and restrictions, which results in an increment in business closures and low economy activity. These measures had important consequences for corporate earnings and economic indicators, contributing to increased market volatility.

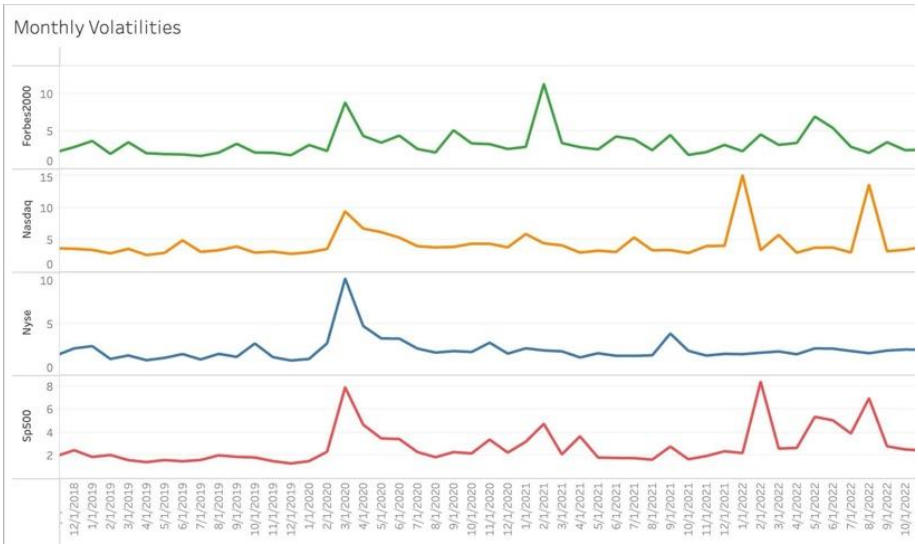


Figure 8: Volatility of US stock exchanges/indexes during 2018 - 2022

6.3 RQ 3: ANNUAL RETURN RATIO & INVESTMENT DECISIONS

The findings from the third question are notably intriguing. Indeed, five companies (BA, MSEX, FL, GD and RGR) featured in the volume-price correlation table also appear among the top performers in the annual return table. This overlap may imply a significant relationship: companies exhibiting robust volume-price relationships also tend to yield substantial annual returns, making them an attractive and potential investment for the future based on these past data results.

7. FUTURE WORK

Some suggestions that could be made as future work could be to focus on advancing the forecasting capabilities of key metrics, including volatility, annual return, and volume-price correlation. An expansion of the project could also involve the inclusion of datasets from diverse international markets, specifically Germany, India, and China. This could help to discover other market dynamics, enabling a more comprehensive understanding of global financial trends.

Additionally, the project could extend its analytical reach to understand different financial markets, particularly delving into cryptocurrencies. By doing this suggestion and combining various metrics and exploring diverse markets, predictive models could improve and make it easier to adapt to the movement of global finance.

8. REFERENCES

[1] Beattie, A. The Birth of Stock Exchanges. Investopedia. Retrieved January 13, 2024, from https://www.investopedia.com/articles/07/stock-exchange-history.asp

[2] CFI Team. Adjusted Closing Price. Corporate Finance Institute. Retrieved January 13, 2024, from https://corporatefinanceinstitute.com/resources/equities/adjusted-closing-price/

[3] Positive correlation between price and volume development - India - Investtech. (n.d.). Retrieved January 18, 2024, from https://www.investtech.com/main/market.php?CountryID=91&p=staticPage&fn=helpItem&tbReport=h_PVCPos#:~:text=Correlation%20between%20price%20and%20volume%20development%20is%20considered%20an%20important

[4] Price and Volume Correlation (n.d.). Retrieved January 18, 2024, from

https://seekingalpha.com/article/244906-price-and-volume-correlation

[5] Chen, J. (n.d.). Annual Return. Investopedia. Retrieved January 18, 2024, from

https://www.investopedia.com/terms/a/annual-return.asp

[6] D. Peng, "Analysis of Investor Sentiment and Stock Market Volatility Trend Based on Big Data Strategy," *2019 International Conference on Robots & Intelligent System (ICRIS)*, Haikou, China, 2019, pp. 269-272, doi: 10.1109/ICRIS.2019.00077.

[7] Wu, Kesheng et al. 'A Big Data Approach to Analyzing Market Volatility'. 1 Jan. 2013: 241 – 267.

[8] Stock Market Data (NASDAQ, NYSE, S&P500). (n.d.). Retrieved December 20, 2023, from https://www.kaggle.com/datasets/paultimothymooney/stock-market-data

[9] Stock Market India. (n.d.). Retrieved January 9, 2024, from

https://www.kaggle.com/datasets/hk7797/stock-market-india

[10] Robinson, P. (n.d.). Historical Volatility: A Timeline of the Biggest Volatility Cycles. DailyFX. Retrieved January 25, 2024, from

https://www.dailyfx.com/education/volatility/historical-volatility.html

[11] Zhang, N., Wang, A., Haq, N.-U. -, & Nosheen, S. (2021). The impact of COVID-19 shocks on the volatility of stock markets in technologically advanced countries. Economic Research-Ekonomska Istraživanja, 1–26.

https://doi.org/10.1080/1331677x.2021.1936112

9. APPENDIX

9.1 RUNNING TIME ON CLUSTER

Volatility Analysis

```
real    4m6.556s
user    2m50.434s
sys     0m16.425s
```

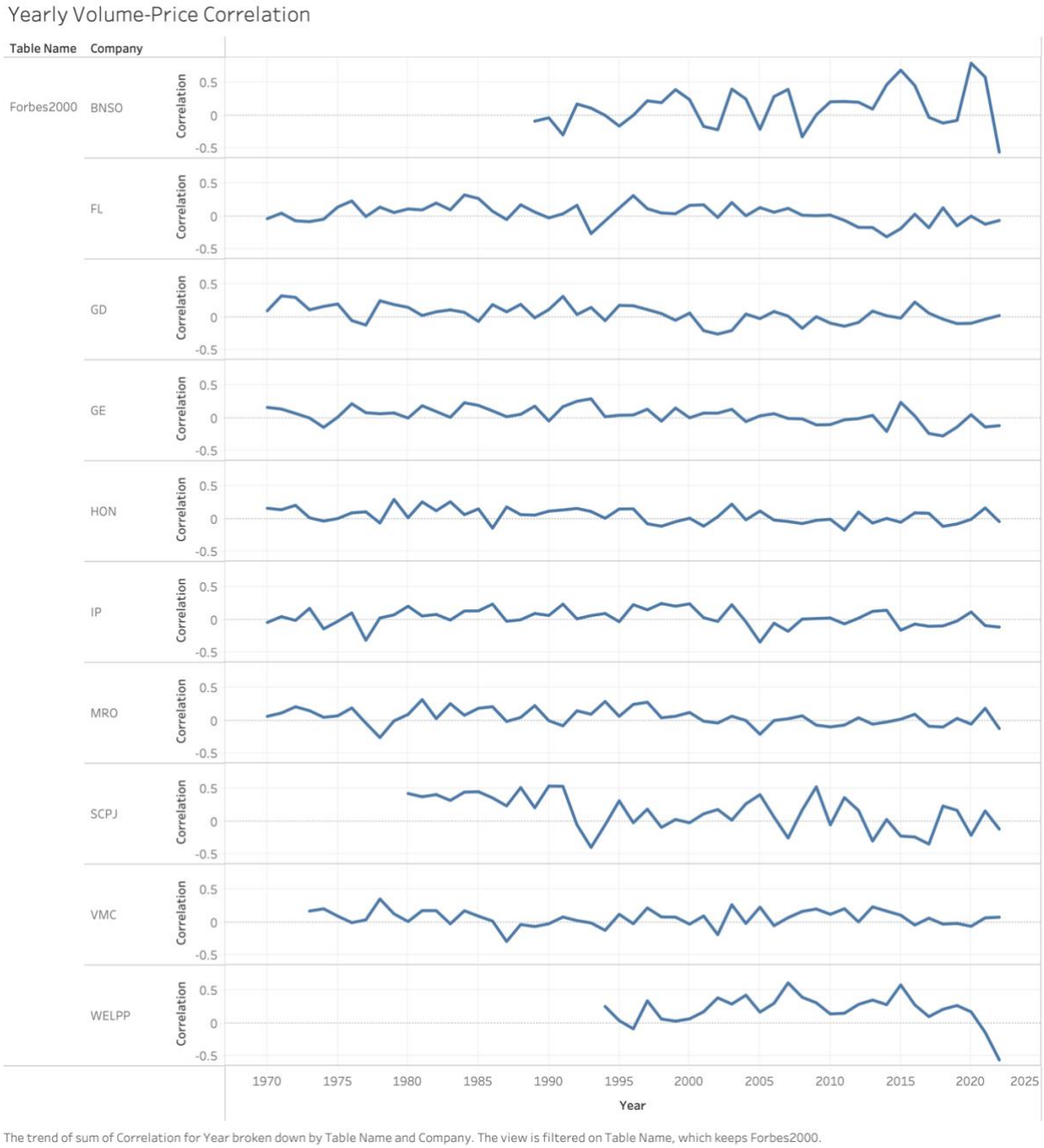
Volume-Price Correlation

```
real    2m40.173s
user    2m18.613s
sys     0m13.745s
```

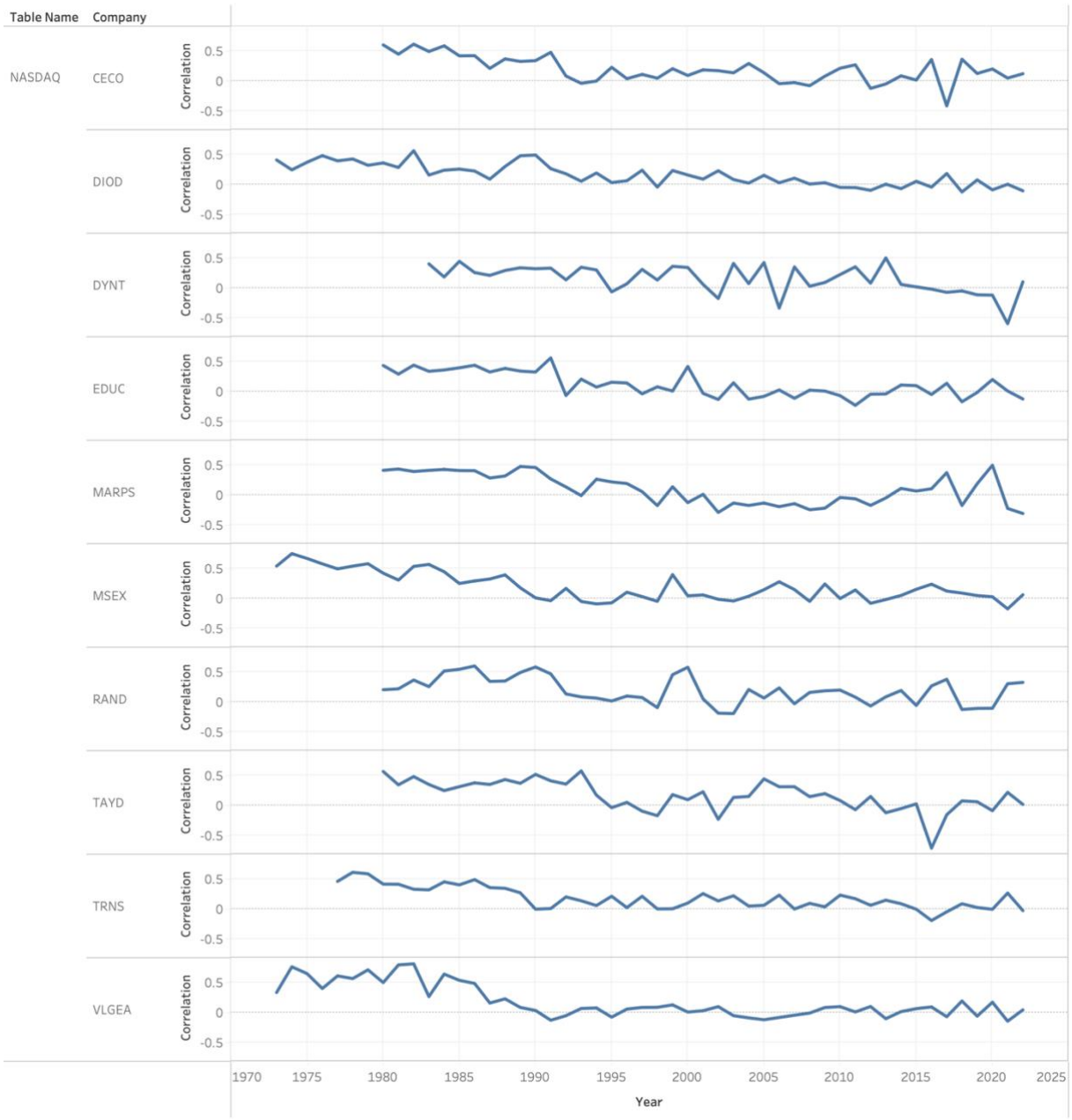
Annual Return Analysis

```
real    5m15.043s
user    2m22.214s
sys     0m13.391s
```

9.2 YEARLY VOLUME PRICE CORRELATIONS FOR ALL US MARKETS

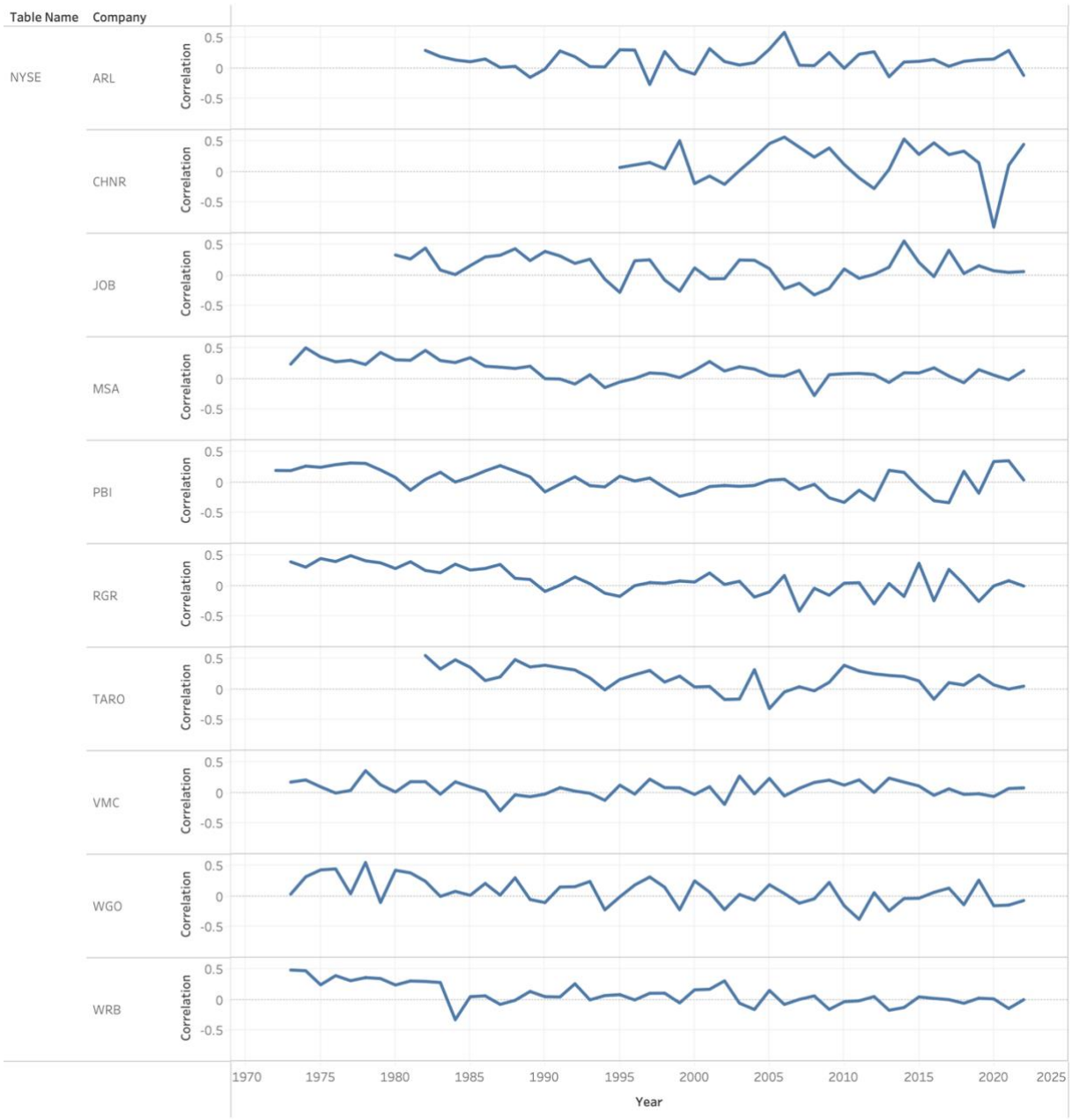


Yearly Volume-Price Correlation



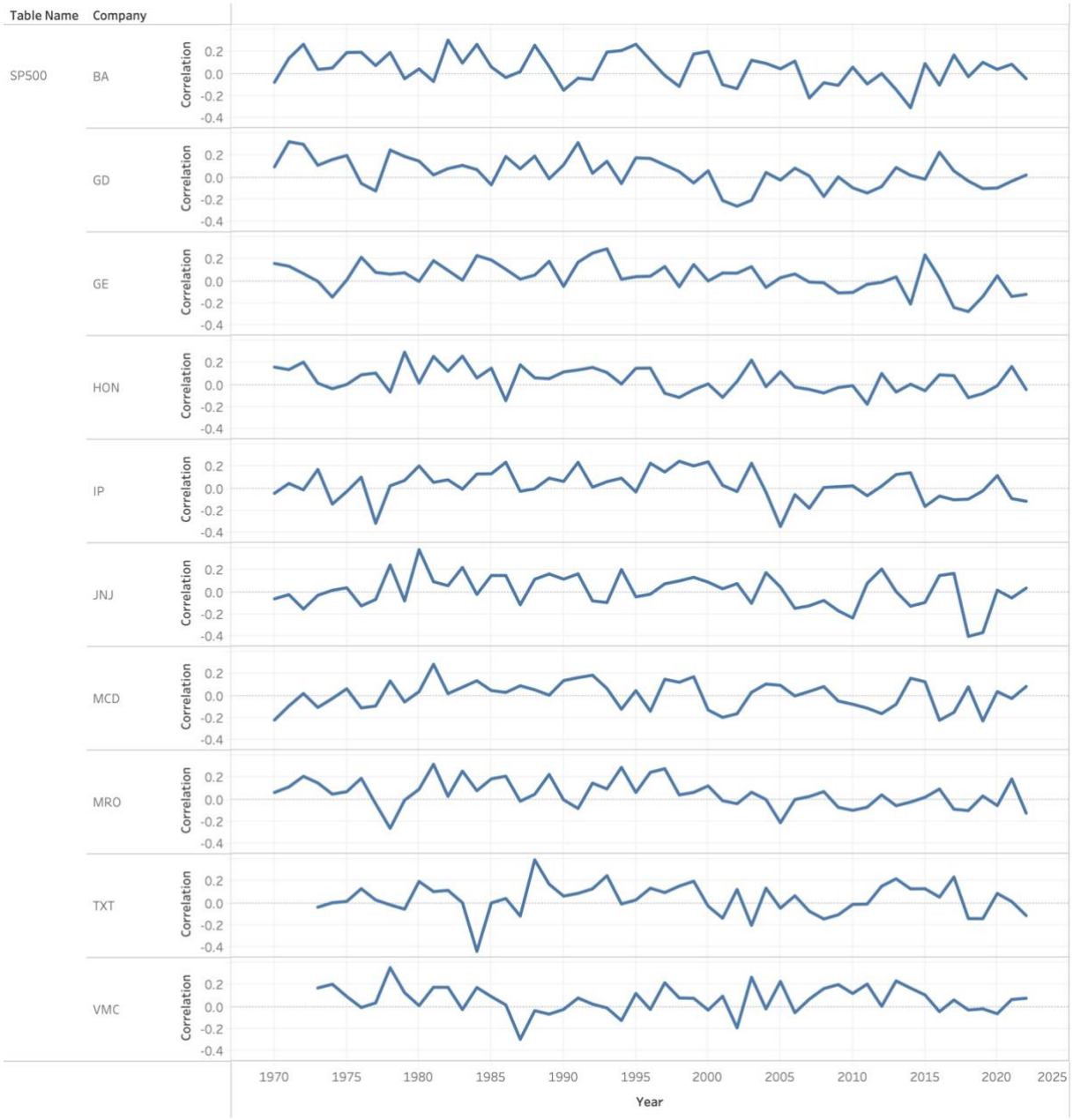
The trend of sum of Correlation for Year broken down by Table Name and Company. The view is filtered on Table Name, which keeps NASDAQ.

Yearly Volume-Price Correlation



The trend of sum of Correlation for Year broken down by Table Name and Company. The view is filtered on Table Name, which keeps NYSE.

Yearly Volume-Price Correlation



The trend of sum of Correlation for Year broken down by Table Name and Company. The view is filtered on Table Name, which keeps SP500.