An Interactive S&P 500 Dashboard!

Alyssa Benjamin, Katelyn Donn, Monisha Kapadia, Avril Mauro, Xining Xu

Northeastern University, Boston, MA, USA

Abstract

The S&P 500 is a collection of 500 large-cap company stocks and its price data has historically been used to assess the performance and health of the U.S. market. Given the strong relevance of the S&P 500 but also its complexity and opportunities of nuance in its interpretation, our group aimed to create an interactive dashboard that displays various visualizations of stock data over time to help beginners understand movement and correlations in the market. Our goal is to create a convenient and user-friendly dashboard that gives the user freedom to explore stock performance through different time frames and industries of their choice. In order to implement these features, we used libraries such as Plotly, Dash, and Dash Bootstrap Components. We predicted industries such as Healthcare and Technology to flourish over time, while others like Real Estate may fluctuate with market recessions. Using our dashboard, we were able to draw a multitude of industry-specific conclusions as well as see the impact of global events like COVID-19 on individual stocks over time.

Introduction & Hypothesis

Finance and Data Science go hand in hand, and as a group full of business and data science majors, we were naturally drawn to the intersection of the two and were led to the stock market. Smart investors leverage data to maximize the return on their investment. For those just starting out, however, the stock market can be overwhelming and confusing, making it difficult to know where to even start. This often prevents many people, especially students, from beginning their investing journey. It is crucial, however, to start investing at a young age in order to grow one's personal wealth. We are motivated by the goal of bridging this knowledge gap and making it easy and convenient to learn more about market trends for those with little to no knowledge of the stock market.

There is great practicality to dashboards for the use of finance. Dashboards are commonly used to communicate information in an easily-digestible manner, and the best part is that they are interactive. Thus, we decided to leverage this data science tool to glean insights on financial data for the purpose of this project. Our dashboard allows users to specify specific time frames and industries to examine closing price data for. For context, the S&P 500 is divided among 11 industry sectors according to the Global Industry Classification Standard. These sectors are: Energy, Materials, Industrials, Utilities, Healthcare, Financials, Consumer Discretionary, Consumer Staples, Information Technology, Communication Services, and Real Estate.

The results we hypothesized are related to performance in the above sectors. For example, we predict that the Healthcare and Information Technology industries will grow the most over time. The Healthcare industry has seen an influx of investment [3], especially from the public sector over the last few decades because of an increase in medical developments as well as the rise in popularity of medical care programs. The Information Technology

industry, similarly, has grown in recent years because of the vast amount of jobs created within the industry and the development of new tech giants like Tiktok and BeReal [4]. Additionally, we predict that the Real Estate industry will be hit hard throughout the years because of the recessions from the Housing Crisis in the United States, which could potentially impact how the Real Estate industry performs in latter years on our dashboard [2]. For individual stocks, we're not entirely sure how individual companies will perform against one another, but judging off of the current market in 2023, we can assume that companies that have flourished in the last 20 years like Apple and Uber most likely have seen strong growing trends within the S&P 500.

Our dashboard is effective because it reduces data overload. While financial websites like Yahoo Finance provide thorough data on specific stocks, there is often an excess of that data (i.e. beta, volume, p/e ratio etc.) that can be overwhelming to a beginner who is unfamiliar with a stock quote. That is why we decided to focus on one specific measure (adjusted closing price) to build a foundation for beginners who use our dashboard. Another weakness of existing financial websites like Yahoo Finance is the lack of creative visualization between stocks and industries. Price over time for a single stock is often the only visual available, but our dashboard allows users to see bubble charts, candlestick charts, bullet charts, tables, and also read about each industry as they explore.

Methods

All of the data used within the dashboard were extracted as CSV files from Kaggle. The S&P 500 Historical Data has a CSV file containing the daily open, close, high, and low prices on an ETF representative of the S&P 500 index from 1927 to 2020 [**2**]. The S&P 500 Stock Price with Financial Statement dataset has a CSVcontaining the daily closing prices of each stock, denoted by its ticker label, from 2010 to 2022 [**6**]. We will utilize this data to conduct a time series analysis of stock performance over a maximum 10 year window. The Top Tech Companies Stock Price Dataset has a .csv file containing the company names and 11 industry sectors (GICS) for each stock according to their ticker label in the S&P 500 [**5**]. We read these files into our Python environment using the pandas library. Our main data frames were as follows:

- **index_df** → daily open, close, high, and low prices for an ETF which is representative of the entire S&P index from 2010 to 2020 [**2**]
- **price_df** → daily price data for each stock, denoted by ticker label, from 2010 to 2022 [**6**]
- **industry_df** → ticker label, company name, and industry sector for each stock in the S&P 500 [**5**]

We used the **pandas merge** method on data frames **price_df** and **industry_df** to create **close_df** which shows the daily adjusted closing price, ticker label, and industry sector for each stock from 2010 to 2022. This merge allowed us to filter price data by industry sectors. Using this dataframe, we used a function called **num_filter()**, imported from **utils.py**, written by group member, Avril, to filter our data on a specific time frame between two user-inputted years. In order to calculate our aggregated average closing price for each sector, we used the **pandas groupby** and mean methods to group our data on sector and calculate average mean closing price. Although our raw data does not include null values, the user may observe null values in the Table (Figure 1) on the dashboard. This is because the pandas groupby and mean methods output null values for stocks which did not have an IPO (Initial Public Offering) in the first user-inputted year during filtering. These null values decrease as the time frame becomes more recent because all stocks will have eventually launched an IPO. For the purposes of standardizing our analysis, we only focused on adjusted closing prices for all stocks and filtered the dataframes to

only include price data from that column. All our visualizations take this into account except for the Candlestick Graph (Figure 3) which utilizes open, close, high, and low prices altogether.

Analysis & Visualizations

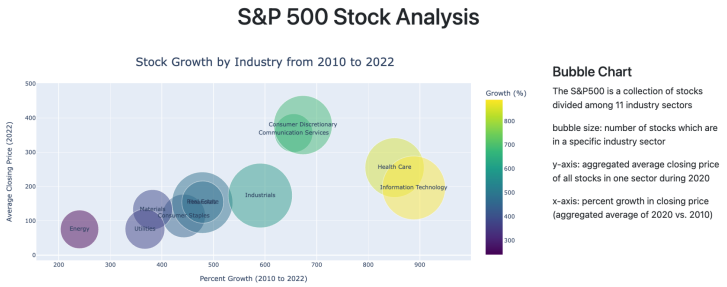

Figure 1 *Bubble chart showing the growth of each industry in the S&P 500 over the span of 10 years (2010 to 2020).*

When a user launches our dashboard, the first visualization they will see is a bubble chart where each bubble represents one of the 11 official S&P 500 industry sectors. The x-axis represents the percent growth from 2010 to 2020 in the aggregated average adjusted closing price of all stocks in one industry. The y-axis represents the aggregated average adjusted closing price of all the stocks in each sector during 2020. The bubble size correlates to the number of stocks within the industry, so a larger bubble represents a greater amount of stocks in that industry in the entire S&P 500. For example, the Information Technology Sector is represented by a relatively large bubble whereas Energy is small, meaning there are more Information Technology stocks than Energy stocks in the S&P 500. The color bar on the right represents helps us differentiate industries by their growth, mirroring the x-axis measurements. Darker colored bubbles on the left side of the graph are lower growth, lighter colors on the right side are higher growth. Bubbles more towards the top of the graph, such as Consumer Discretionary, have higher average closing prices; this makes sense for this sector since luxury goods often fall in this category.

Figure 2 *Bullet Chart and Table showing the tickers, company names, and average adjusted closing price for stocks within a user-specified timeframe and industry.*

The next part of the dashboard is user-interactive. The Table and Bullet Chart (Figure 2) show the change in closing prices between two years. The boxes above the table are controls that allow the user to define the timeframe for this analysis; the user can pick any year between 2010 and 2022. To the right of those year inputs is a drop-down menu that allows the user to define the industry sector to pull stocks from. The next part of the dashboard is user-interactive. The Table and Bullet Chart (Figure 2) show the change in closing prices between two years. The boxes above the table are controls that allow the user to define the timeframe for this analysis; the user can pick any year between 2010 and 2022. To the right of those year inputs is a drop-down menu that allows the user to define the industry sector to pull stocks from. The Table displays Sector, Ticker, Company Name, Average Adjusted Close for Year 0, Average Adjusted Close for Year 1, and the Change between those two years (red being negative, green being positive). For the Table, selecting “All” will show all stocks in the S&P 500 while selecting a specific sector will only show stocks in that sector. The Bullet Chart also corresponds to these controls and shows the change in aggregate closing price of all stocks in that sector between the two user-inputted years. If

"All" is selected, the change will display for the entire S&P 500 Index. The dollar amount is represented by the large number ($75.5), and the smaller number next to the green triangle represents the overall increase (red triangle for decrease) of the average closing price over the time period. Finally, the bottom right chart contains various information about individual stocks.

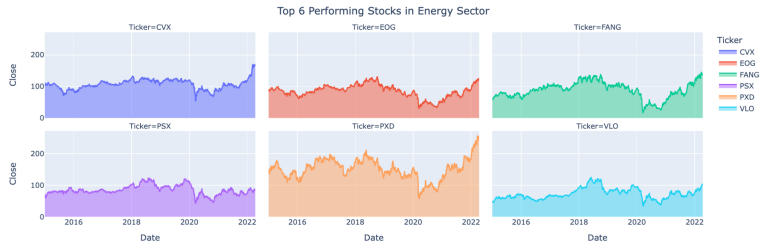

Figure 3 *The area subplot graph depicts the closing price over time of the top 6 performing stocks calculated within each sector controlled by the user-inputted industry/ time range.*

For the next set of visualizations, top 6 stocks were calculated based on the input from industry dropdown. If "All" was selected, the top 6 stocks were calculated from the overall S&P 500, but otherwise top 6 stocks from an individual industry were calculated. These calculations were done by first filtering by industry if necessary and taking the mean adjusted closing price to sort the stocks from highest to lowest and the top 6 were sliced. The individual graphs consist of the closing price of the stock plotted over time. The legend corresponds to the individual color of each graph that represents a separate company or ticker. The users can reference the table from Figure 2 to determine more about the company the ticker abbreviation represents in this graph. The purpose of the visualization is to help users determine which stocks are propelling the growth of a certain sector and how their performance could have fluctuated over time. The user can also define a specific time range to recalculate the top 6 stocks to better explore the historical specific data and drivers of industry. The graphs will also re-plot based on new time ranges to see the exact fluctuations in closing price that determined the stock's status as one of the highest of the industry.

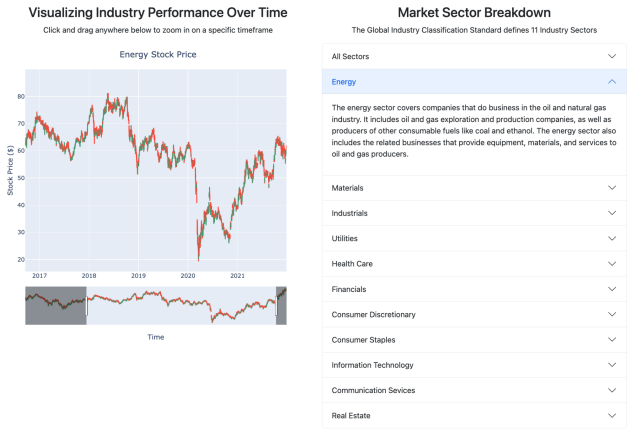

Figure 4 *This figure displays both the candlestick chart based on the entire S&P 500 or a specific sector along with the user defined time range and an accordion component that gives users an explanation of each sector*

The candlestick chart on the left is a common financial visualization that allows the open, close, high, and low prices to be plotted over time. The red lines dictate a decrease in price while green depicts an uptick. The visualization comes in with a built-in range slider underneath the plot for users to be able to explore specific and smaller time frames easily. A chart like this is much more informative as opposed to a typical line graph of a stock's price at only close or open and users can get a better understanding of stock fluctuations even throughout a singular day. The chart takes in user input for industry and will either plot the total price of the S&P 500 or a specific industry by calculations of the mean open, low, close, and high prices of all the companies in the S&P 500 that belong to the selected industry. The chart also takes user input for a specific time range of years between 2010 and 2020 to once again reinforce the user's understanding of historical or current stock fluctuations. On the right there is an accordion component that lists all of the sectors we implemented as filters on the dashboard which we followed the GIC standards for. For users without a strong foundation in stock sectors and the standards followed that made these sector classifications, the accordion item provides a description for each industry based on whatever sector is selected.

Conclusions

The main limitation of our dashboard is its lack of access to real-time data or data outside of our current years constraints of 2010 to 2022 (only 2020 for the ETF representing the S&P Index). Though there are a good 10 or 15 years of overlap within the datasets, once you surpass those years, the charts and graphs start to populate with null values, which can be frustrating. Having access to real time data (like on Yahoo Finance) would be incredibly beneficial to our product and our end user as it would be helpful to have a dashboard that updates as stock prices update. Then, the user would be able to stay up-to-date with current market trends rather than simply analyzing past trends to predict future markets (which can be very unreliable due to ever-shifting market prices).

The conclusions from our project drew from all industries and time periods. For example, for each industry, there's a different breakdown in terms of trends throughout the years. Information Technology clearly has the highest growth out of all the industries, where a lot of the focus of the growth is in recent years, as seen by the individual stocks. This makes sense as the Information Technology industry consists of companies like Apple, Adobe, Mastercard, and Oracle, which are all companies with high brand recognition who have products with steady day to day usage. The Energy sector has low growth and low average price over time, but it's the most volatile industry over all, which is incredibly interesting to see. This is mainly because of how sensitive the energy sector is to changes in resources such as price of oil, demand and supply of energy, and government regulations, just to name a few. Additionally, because so many consumers are reliant on fuel in their day to day lives, it's a difficult commodity to substitute out if the prices of gas or oil fluctuates. That limitation on consumers allows the market to behave more recklessly because of their constant, dependable consumer base. The Consumer Discretionary industry has the highest price overall, which makes sense considering it consists of a lot of luxury brands that cater to high-paying individuals. The Industrial industry is the most steady throughout as it perfectly falls in the middle of the growth and price chart. It's interesting to note that the top 4 out of 6 stocks in the overall S&P 500 are from the Consumer Discretionary industry, with the last 2 falling into the Communication Services industry. As this is measured by the prices with the highest stocks, we're assuming in our conclusions that higher stock price equates to better company performance. During COVID-19, it's interesting to see how certain industries went up (notably, Consumer Staples, Information Technology, and Health Care), some industries stayed relatively stable (Communication Services), and the remaining industries took massive hits (i.e. Industrials, Real Estate, and Consumer Discretionary).

Overall, it's incredibly interesting seeing how our original hypothesis compares to our conclusions and how the nuances of time played into the conclusions we drew.

Author Contributions

Alyssa Benjamin …… Report, Graph

Katelyn Donn ……. Report, Github, Graph

Monisha Kapadia ……... Graphs, Report

Avril Mauro …… Graphs, Dashboard Layout, Report

Xining Xu …… Report, Graph

References

1. *GICS - Global Industry Classification Standard*. MSCI. (n.d.). Retrieved April 19, 2023, from https://www.msci.com/our-solutions/indexes/gics

2. Han, Henry. *S&P 500 Historical Data.* Kaggle, 5 Nov. 2020, https://www.kaggle.com/datasets/henryhan117/sp-500-historical-data.

3. Huang, MeiChi. *A Nationwide or Localized Housing Crisis? Evidence from Structural Instability in US Housing Price and Volume Cycles.* Springer. *SpringerLink*, 31 May 2018, https://link.springer.com/content/pdf/10.1007/s10614-018-9822-9.pdf.

4. Levit, Katharine, et al. *Trends In U.S. Health Care Spending, 2001.* HealthAffairs, Jan. 2003, https://www.healthaffairs.org/doi/epdf/10.1377/hlthaff.22.1.154.

5. Mantero, Tomas. *Top Tech Companies Stock Price.* Kaggle, 24 Nov. 2020, https://www.kaggle.com/datasets/tomasmantero/top-tech-companies-stock-price?select=List%2Bof%2BSP%2B500%2Bcompanies.csv.

6. Park, Hanseo. *S&P 500 Stocks Price with Financial Statement.* Kaggle, 18 Apr. 2022, https://www.kaggle.com/datasets/hanseopark/sp-500-stocks-value-with-financial-statement.

7. The S&P Sectors. Corporate Finance Institute, 14 Mar. 2023, https://corporatefinanceinstitute.com/resources/valuation/the-sp-sectors/.

8. Wolf, Michael, and Dalton Terrell. *The High-Tech Industry, What Is It and Why It Matters to Our Economic Future.* Beyond the Numbers, Cornell University, 1 May 2016, https://ecommons.cornell.edu/bitstream/handle/1813/78715/BLS_BTN_The_high_tech_industry.pdf?sequence=1.

9. *GICS - Global Industry Classification Standard*. MSCI. (n.d.). Retrieved April 19, 2023, from https://www.msci.com/our-solutions/indexes/gics