Understanding Market Dynamics in 2021:

General Sentiment and Outlier Behavior[1]

A report to Jefferies Group

Yinqiu (Linkin) Chen[2], Fei (Adam) Cheng[3], Kairan (Kevin) Gu[4],

Robert Lanni[5], Matthew Pulcini[6]

[1] Lehigh University, Bethlehem, PA 18015, USA. Course listing: "Fintech Capstone"; FIN-388.
Advisor: Burak Eskici. Last revised: May 9, 2022.
[2]Finance & Accounting, College of Business, Lehigh University. Email: yic321@lehigh.edu and chenyinqiuu@gmail.com
[3]Finance, College of Business, Lehigh University. Email: fec223@lehigh.edu and adamantiumx2099@gmail.com
[4]Computer Engineering & Finance, Lehigh University. Email: kag222@lehigh.edu and kairangu@gmail.com
[5]Finance, College of Business, Lehigh University. Email: rol222@lehigh.edu and rlannijr18@gmail.com
[6]Finance, College of Business, Lehigh University. Email: mcp224@lehigh.edu and matthewpulcini@gmail.com

Table of contents

Executive Summary

In this report, we are analyzing the market behavior in 2021, identifying outlier market behavior (in terms of price movements and volume), and testing if this correlates with aggregated sentiment measures by utilizing sentiment analysis and measuring event relevance in order to give an investment suggestion. The stock market in 2021 had high growth and volatility. There were major changes in the market dynamics, such as new fintech products for retail investors, cash influx to the markets, an increased number of retail investors, and behavior changes, such as gamification-like investment decisions. All these events make it increasingly more challenging to predict the stock price than ever. However, this also creates a perfect chance for us to test the correlation between a stock price change and news sentiment score change. We decided to focus on the Standard and Poor's 500, also known as S&P 500, companies as our sample. We used the (1.5 *IQR) method to determine the number of price outliers for each sector on a daily basis and compared it with the Federal Reserve Bank of San-Francisco Daily Sentiment Index(FRBS) to find the correlation between them. After thorough research, we do find some correlations between them. For some sectors, materials, for example, the correlation is stronger and for other sectors, utilities, for example, the correlation is weaker. But generally speaking, there is a positive correlation between FRBS daily sentiment index change and the S&P 500 companies' stock price change.

Introduction

When looking at the market, the individual investor is trying to find an edge, increase alpha, or see a trend that other investors may not have noticed yet. When looking at the equities market specifically in the past two years, it is hard to see consistent trends, as the market has been completely reshaped over the past two years due to the pandemic. In late 2020 and early 2021, prices of technology stocks soared, SPACs had a record year, and speculative equities flew high as normal value stocks rose slightly. It was a year where fundamental financial analysis became less useful, as companies were trading at extremely high valuations with financial multiples (EPS, P/E Ratio, P/B ratio, etc) that did not match the price. We found this to be very interesting as many investors believed these ratios as critical to determining the value of a company. We have seen large percentage gains/falls become more prevalent in 2020/2021 as well, allowing for more short-term profits.

Major Changes in 2021

New Players

Applications like Robinhood and Webull have gamification-like platforms that allow users to buy intricate financial instruments such as calls and puts. Many of the people buying the derivatives did not have a fundamental understanding of how they worked. These platforms had no way to determine the levels of knowledge and experience that the user possesses but still allowed them to be involved in complex trading strategies. These applications provided complicated methods of stock trading, allowed users to buy fractional shares, and also offered

commission-free trading. Allowing fractional shares makes it possible for retail traders to acquire equities that they could not previously get their hands on. For example, Tesla has a stock price of $1225 on Nov 5th, 2021 (Yahoo Finance). If people are not using the partial-purchasing method, many individual investors would not be able to purchase $TSLA since it is expensive and will dominate their portfolio allocation. Also, the commission-free policy encouraged more retail traders with small accounts. The traditional method requires a commission fee when people place an order and this fee is constant regardless of transaction size. Under this setting, institutional investors benefit more than individual investors because they typically place larger orders, therefore making the average commission fee for each order smaller than those of the retail investors. A third method these fintech companies used to attract new customers is that they do not require an open deposit account. The traditional brokerage platforms always require an open account deposit from hundreds of dollars to thousands. However, some fintech companies, Robinhood for example, don't require any open account deposits. This new policy enabled young adults and less financially stable individuals to get involved in the stock market. This helped enable the stock market to grow at rates that even analysts could not believe as there was more new cash entering the market than ever before.

These fintech companies gave many Americans an easy way to invest by "so-called" democratizing finance, something that was unprecedented. This increasing flexibility and ease of access to the financial markets catalyzed growth for investment platforms, introducing new users and a younger demographic to stock trading. This can be evidenced by using the most famous commission-free trading platform, Robinhood, as an example. The number of overall Robinhood users rose sharply as well during 2021, as they have had tremendous growth since 2018, almost doubling every year[1] (see graph below).

[1] https://www.statista.com/statistics/822176/number-of-users-robinhood/

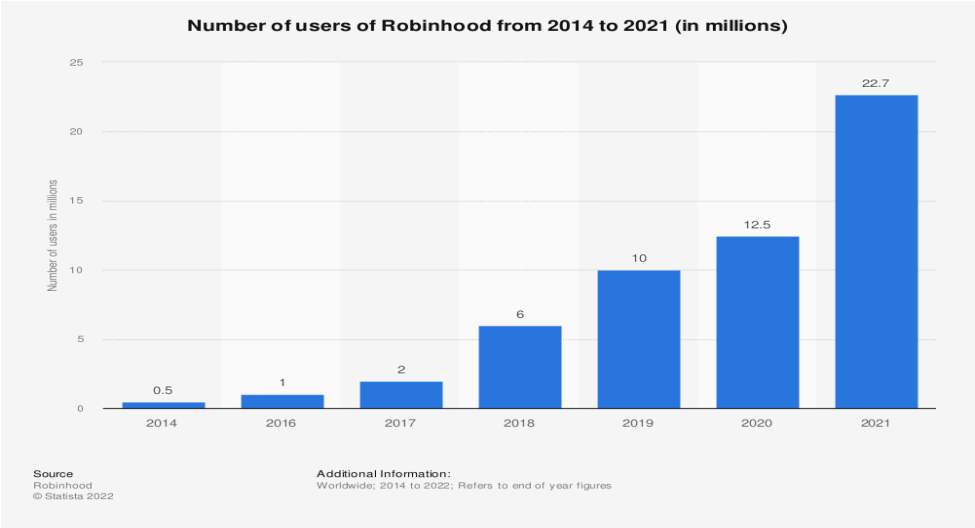


Figure 1 - number of Robinhood users from 2014 to 2021

In 2021, there were about 130M recorded households in the US, and Robinhood had 21M active accounts, equalling 1 in every 6 households. Robinhood was not the only company that experienced major growth, as many other fintech companies also experienced a huge user increase during 2020 and 2021. Most of these companies actually doubled their users during that time period. (see graph below)

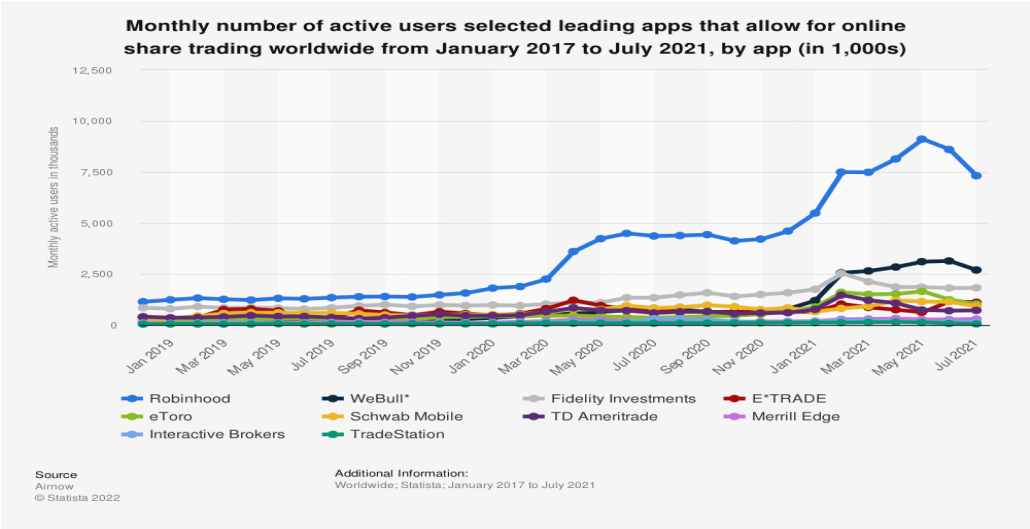


Figure 2 - Monthly user increase pattern for the top 10 fintech companies (Statista.com)

New Cash, More Cash

Lending rates and prime rates were at an all-time low in 2020 since 2015. In order to boost liquidity in the market and help get money flowing through the economy, these market conditions continued into 2021. The chart below shows the historical U.S. average lending rate. We can easily see that there is a huge drop from 2020 to the present.

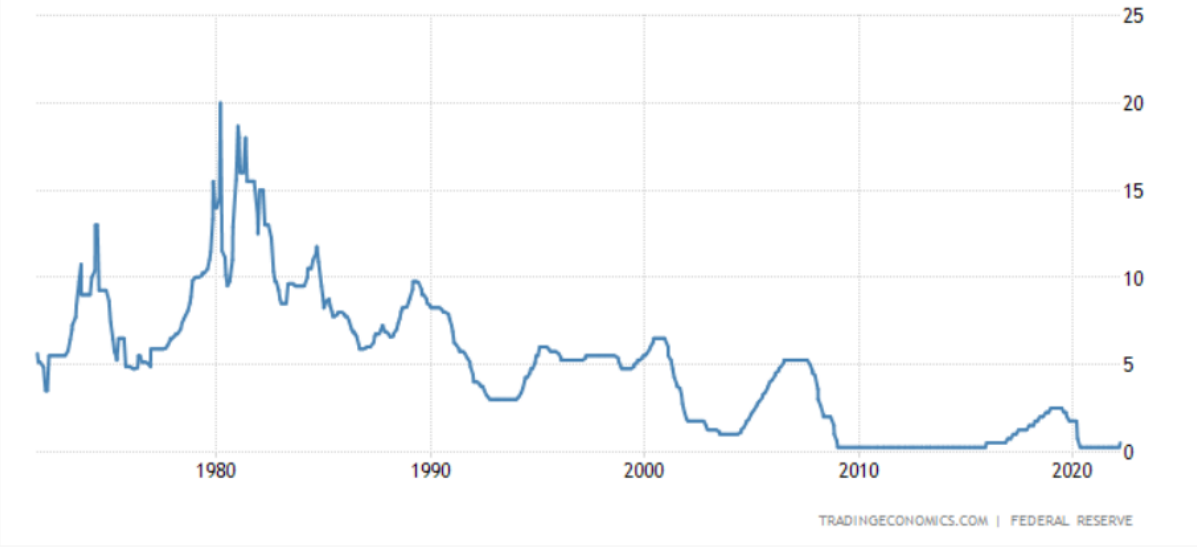


Figure 3 - The average lending rate in the US from 1970 to 2022, source: tradingeconomics.com

Coronavirus caused the federal government to pump an exuberant amount of money into unemployment funds, PPE payments, and the stimulus checks as this all contributes to a market climate that allows such a boom in late 2020 and 2021. In 2020 and 2021, the US government has provided over $850 billion direct payments to taxpayers. (Peter G Peterson Foundation, March 15, 2021). Based on a survey from Betterment, 46% of Americans said they invested part of their stimulus check into the stock market. Specifically, “Of those who invested their stimulus check, 70% invested at least half or less of their stimulus check.” (Zack Friedman, July 27, 2021). In short, regardless of how they invested their money (through institutional investors or retail investors), nearly $150 billion new cash went into the stock market.

Also, the flourishing of stock trading platforms allowed for the largest ever influx of money into the equities market in history. According to both MarketWatch and Bloomberg[2], the 2021 stock market took in more cash than the past 20 years combined[3].

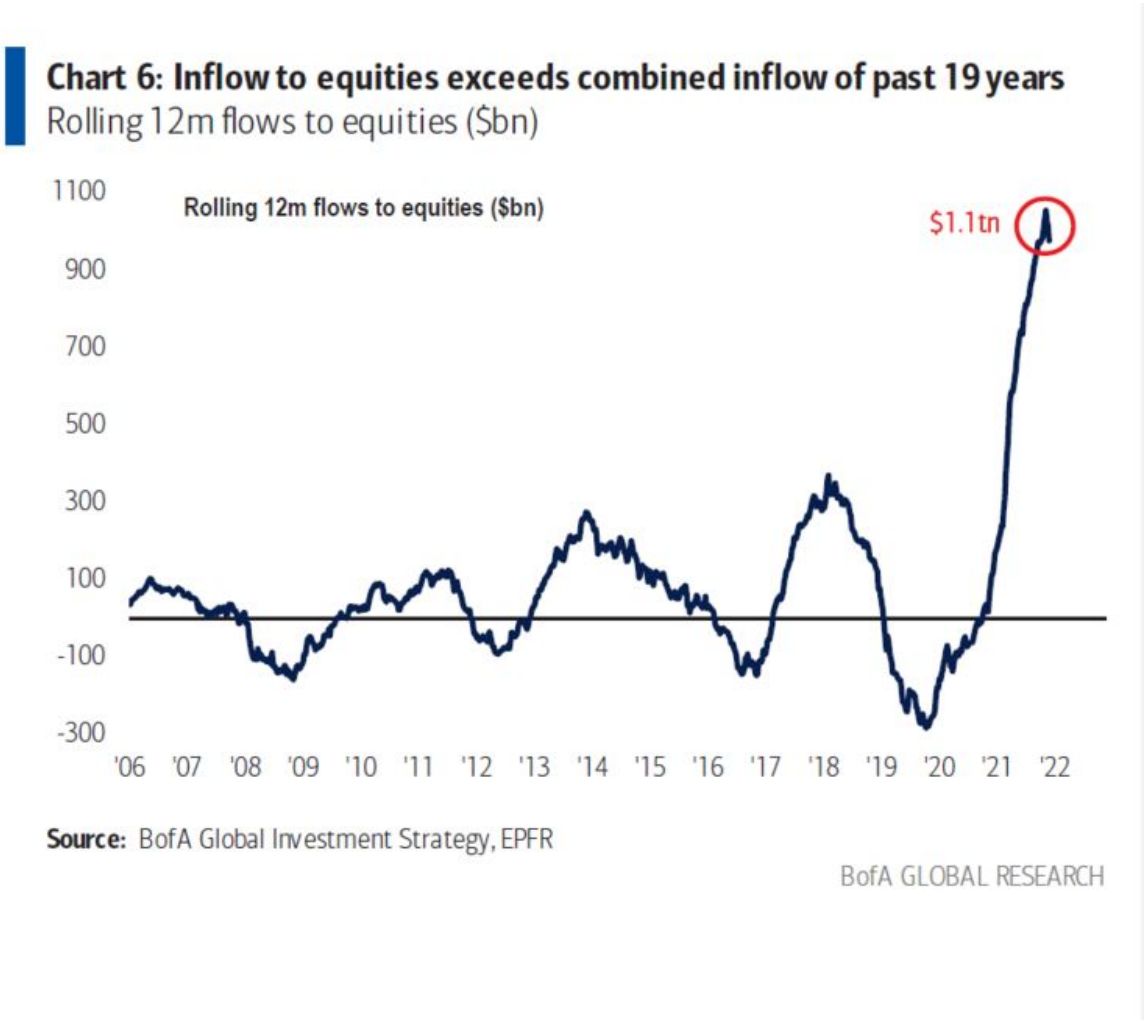


Figure 4 - cash inflows of stock market from 2006 to 2022

All these factors combined with each other contributed to the stock market booming in 2021.

[2] https://www.bloomberg.com/news/articles/2021-11-25/stock-funds-took-in-more-cash-in-2021-than-two-decades-combined

[3] https://www.marketwatch.com/story/the-1-trillion-that-has-flowed-to-global-stocks-in-2021-is-bigger-than-the-last-20-years-combined-11631273525

Big Events

The Stock Market in 2021 was extremely volatile, despite growth in many sectors. Structural changes and the new cash, as discussed above, resulted in unprecedented market behavior. For example, in January 2021, Gamestop soared as short sellers attacked the faltering company. This was the beginning of a crazy and lucrative time for investors involved with "meme stocks". Meme stocks became popular on reddit.com/wsb, a page where users would share their investment advice. A member of this community that goes by the name "DeepF*ckingValue" broke the norm for investors by refusing to sell his calls (no matter how deep ITM they were). This along with the vast short interest in a failing brick and mortar game sales company, allowed Gamestop to soar from around $9 a share all the way to $453.

Figure 5 - GameStop stock price for 2021 from Yahoo Finance

When Gamestop and AMC "broke" the stock market, it had more of a ripple effect than most people realize. Many stocks during this 4-5 month time frame broke normal valuation multiples like EBITDA, Price/Earnings ratio, and Earnings per Share, as speculative stocks based on almost none of them. This began a trend in the market where there was less investing

and more gamification if you will, where users are gambling more than investing on sound fundamentals.

Methods

Why S&P 500

In this report, we chose 2021 stock price data from companies in the S&P 500 as the main data for analysis. The S&P 500 is composed of the 500 largest equities by market capitalization. The key benefit of using companies in the S&P 500 is the wide market breadth that they span. These 500 companies cover all eleven Global Industry Classification Standards (GICS), with twenty-four industry groups. The S&P 500 is very diversified and in addition to its broad scope, the companies within the index are highly credible. In other words, these companies are all large, well-known, reputable, and audited, so they are unlikely to be posting unsound data. Unlike small-scale companies whose stock price and trading volume can be easily affected by fraud, insider trading, and other kinds of schemes. The market cap for S&P 500 companies is about $42.4 trillion, which means it is so hard for one person or one organization to affect the stock price under their own interest. Lastly, their performance will strongly influence the performance of the entire market. If the S&P 500 Index got a cough, the whole stock market would likely feel those same symptoms, except slightly more exacerbated than the S&P 500. With all of this in mind, any changes in the S&P 500 stock price and trading volume can be roughly considered as a reflection of the attitude change in the whole market.

The third advantage of using the S&P 500 companies is its accessibility. The S&P 500 Index is a well-known Index and it is also widely used worldwide. There are many well-known organizations analyzing S&P 500 companies. So, a lot of sentiment analysis about the S&P 500

is already posted online. In this situation, we will be able to observe previous models, especially those that fit our needs and are accurate.

Why 2021

In 2021, there were a multitude of events that contributed to the extremely volatile financial market. The continuation of COVID-19, the flourishing of meme stocks, and the boom of SPACs in 2021 made the stock market both extremely volatile and unpredictable. The strong volatility also led to record high profits across Investment Banks, like Goldman Sachs, J.P. Morgan, and Citi, who all posted record profits in 2021 according to The Financial Times[4]. Under this market condition, the traditional stock market model is becoming less useful as it becomes more unpredictable.

We searched to find another indicator that can explain why stock prices can fluctuate drastically and rapidly. We utilized this information to try and find a correlation between our market sentiment on specific sectors and why people decide to purchase equities at a specific time. Our hypothesis was that the stock price of one sector is related to the sentiment score of that sector, which in other words means the price change and the sentiment score change should be corresponding. In our opinion, since 2020, more and more individual investors are entering into the market than ever thought before. And based on a study conducted by Brunswick, “Moreover, most investors (88%) are making decisions based on information they have learned online ” (Maja Pawinska Sims, Jan 30, 2019). In 2021, there were many major news stories, which made it significantly easier to test if the market outlier behavior and the sentiment score changes correspond.

[4] https://www.ft.com/content/c3e4bcce-00fa-43fb-a427-625523cdebdb

Anecdotal Evidence

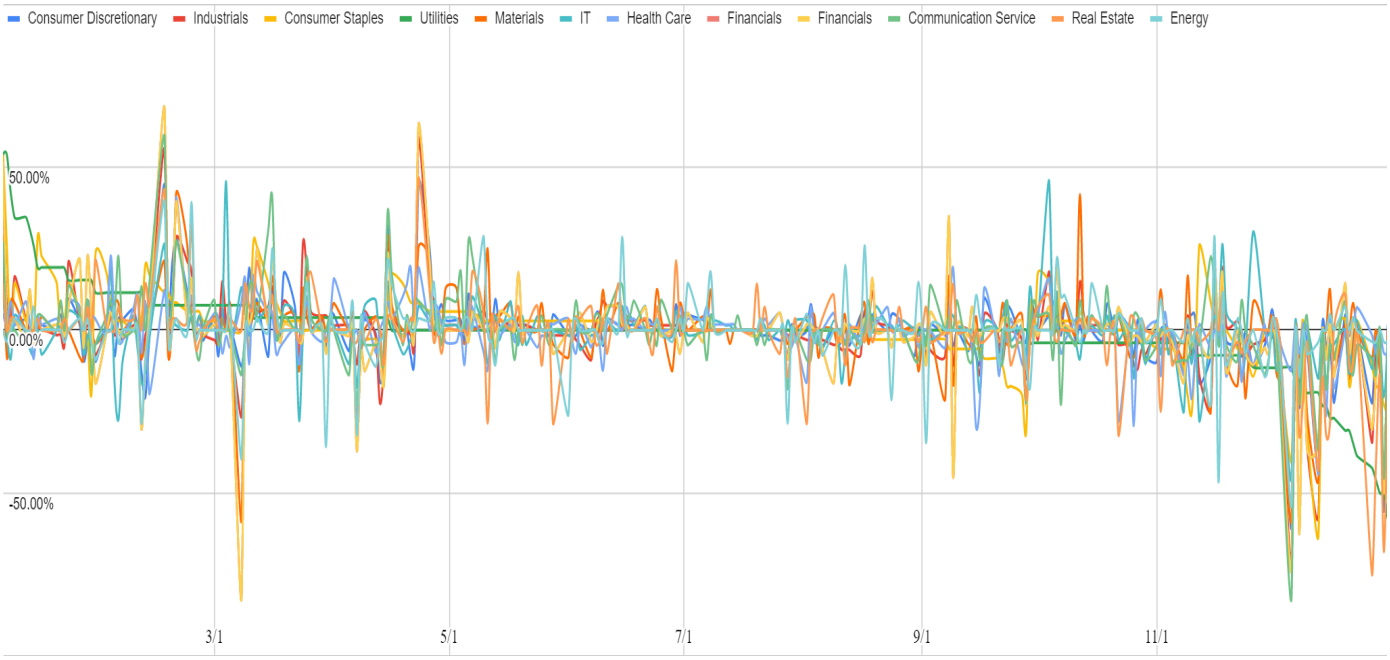


Figure 6 - comparison between 11 sectors % of outliers on one day in 2021

May 11th 2021

The S&P 500 lost 2.1% to 4,063 for its biggest drop since February, a weekly decline of more than 5%. One of the biggest concerns was Inflation, which accelerated at its fastest pace since 2008 last month along with the Consumer Price Index (CPI) spiking 4.2% from a year ago, compared to the Dow Jones estimate of 3.% increase in that category. The monthly gain was 0.8% , versus the expected 0.2%. Excluding volatile food and energy prices, the core CPI increased a whopping 3% from the same period in 2020 and 0.9% monthly basis. The respective estimates were 2.3% and 0.3% . Investors have been fearful that a pick up of inflation could squeeze margins in multiple sectors and lessen corporate profits. The lessening of large corporate profits normally is a sign of a recession as the trickle down economic effect is important to note. This fear is causing the stock market to slide downward in may as people believed the Federal

Reserve is behind in its fiscal policy decisions according to CNBC[5]. This fear and uncertainty in the market caused a sell off as we calculated 180 negative outliers on this day, with the market reaching its low point across all indexes just before market close. Tech Shares led the decline out of all sectors, with strength in the energy sector.

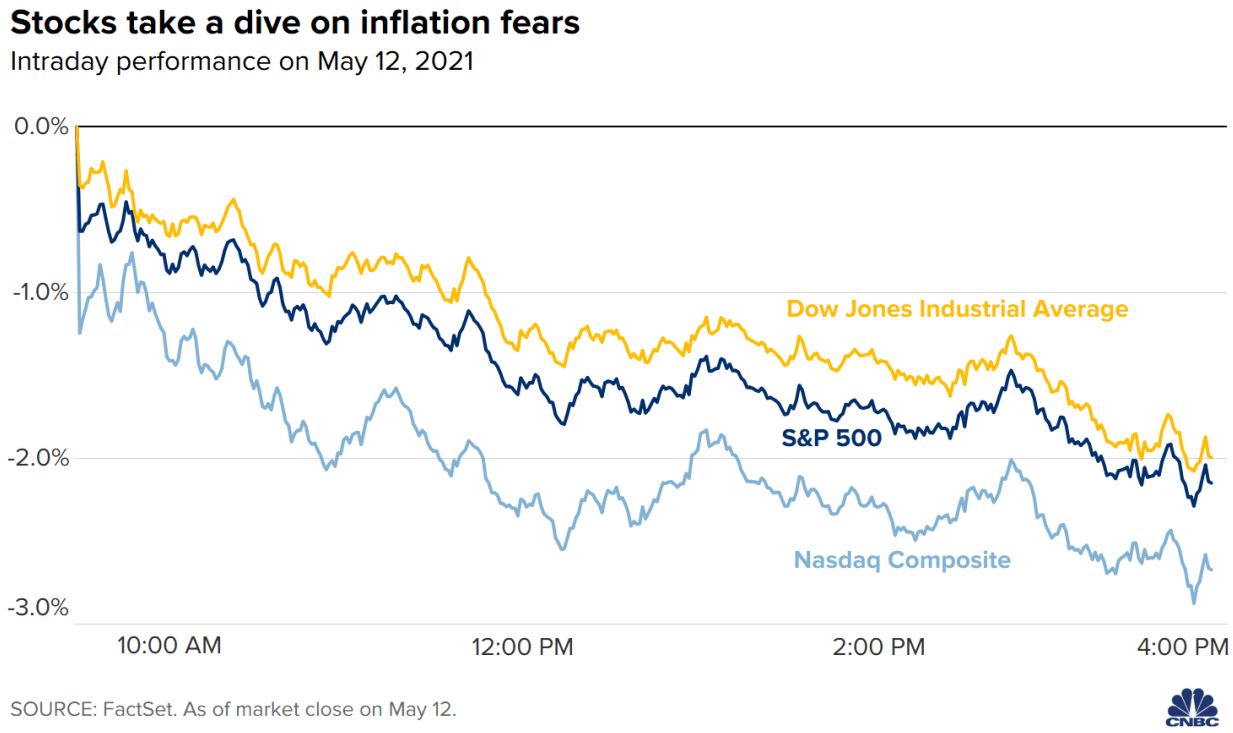


Figure 7 - comparison between DJI, S&P 500, and Nasdaq Composite on May 12, 2021

December 1st, 2021

In early December, there were two days where we found outlier correlations. On December 1st, 2021, we calculated 250 negative outliers. The fear-gauge CBOE Volatility Index (VIX) increased by 18.4%, to close at 27.19. This market selloff was most likely due to two factors; the news of a new coronavirus variant omicron and its potential risks along with Federal

[5] https://www.cnbc.com/2021/05/11/stock-market-futures-open-to-close-news.html

Reserve decisions. Fed Chairman Jerome Powell proposed that US lawmakers consider asset purchases quicker than planned.. Powell was worried about inflation however did not want to act afraid in front of the American people. The uncertainty seemed to be the greatest fear as both aspects of the day's news are based upon fear. The increasing volatility index is also an indicator of the correlation between fear and market volatility, specifically a negative move showing support for our thesis(Yahoo Finance, May 11th 2021) .

December 2nd, 2021

The S&P 500 declined 53.96 points, or 1.2%, to close at 4,513.04 on Wednesday, with only the utilities sector ending in the green, , while 10 of 11 major sectors closed in the red. The communication services sector led the decline with nearly 2% decline, followed by consumer discretionary that lost 1.9% in the session. The Nasdaq Composite Index closed at 15,254.05, after declining 283.64 points, or 1.8%, however, it did touch an intraday peak at 15,816.82. (Zacks Equity Research, December 2, 2021, Stock Market News for Dec 2, 2021)

On Wednesday, the fear-gauge CBOE Volatility Index (VIX) increased 14.5%, to close at 31.12. Declining issues outnumbered advancing ones on the NYSE by a 2.26-to-1 ratio, while a 2.96-to-1 ratio favored decliners on the Nasdaq. A total of 14.2 billion shares were traded yesterday, much higher than the last 20-session average of 11.3 billion. This is important to note as the rise in volume may show strength in the economy as “buying the dip” can be seen here. This reinforces our theory as once the federal reserve decided to step in and the market learned that the vaccines were effective against the Omicron Variant. This confidence in the global economy, along with vaccination numbers, and its ability to deter COVID-19 allowed the market

to rally. We found another example where back to back days have had tremendous volatility in our data analysis.

Data Generation

We generated our data from the python script attached below. To start with, we installed Yahoo Finance, Pandas, and Excel Writer Libraries. We then obtained a list of the S&P 500 stocks from Wikipedia. Additionally, we download the specific equities data, including low price, high price, and volume from January 1st of 2021 to January 1st of 2022 from Yahoo Finance. Finally, we created a Panda's data frame out of a Python dictionary and exported it into an Excel file.

```
# Import packages
import yfinance as yf
import pandas as pd
import xlsxwriter

# Read and print the stock tickers that make up S&P500
tickers = pd.read_html(
    'https://en.wikipedia.org/wiki/List_of_S%26P_500_companies')[0]
print(tickers.head())

# Get the data for this tickers from yahoo finance
data = yf.download(tickers.Symbol.to_list(),'2021-1-1','2022-1-1', auto_adjust=True)['Low']
# Create a Pandas DataFrame out of a Python dictionary
df = pd.DataFrame.from_dict(data)
# look at the Data
print(df.head())


df.to_excel("S&P500.xlsx")
```

Figure 8 - Detailed code shown above by reading the S&P 500 tickers from wikipedia

Descriptive Statistics

Market Outlier Behavior

The raw data consists of the security's open price, the close price, and the volume traded. There are 252 days where the stock market is open in 2021 and yields valid raw price and volume data represented by each column. There were initially 505 stocks from the raw data, yet a few are filtered out from further analysis due to lack of relevant data, mostly regarding either price change or sectorial information that Excel can provide.

By using the earlier method where the gap between opening and closing is the basis of outlier computation, it is important to be aware that the price percentage change may be influenced by after-market behaviors on the previous day and the premarket fluctuation on the current date. Essentially the price percentage change being referred to throughout this report is purely about the regular market hours price movement of a single date, typically from 9:30 AM eastern time to 4:00 PM on a trading day.

Volume percentage change of each stock is calculated by finding the difference between current day volume and previous day volume, divided by previous day volume. The next step is to identify the outliers among this price and volume change data. Our immediate consideration is to break down percentage change information into four quarters. Generating outliers for each stock for price percentage change and volume percentage change is the goal here, and we figure it should be more precise if we break down information into four quarters, each consisting of three months of data.

The function below is an example cell formula that computes the InterQuartile Range of either price percentage change or volume percentage change of a given stock in one quarter.

=(QUARTILE.INC('Q1'!B2:BJ2,3)-QUARTILE.INC('Q1'!B2:BJ2,1))

Our definition of an outlier is any data value that is beyond 1.5 times the IQR (interquartile range) from the mean, either positively or negatively, within a quarter. We compute the average price % change and volume % change for each stock in every quarter, used as the benchmark for outlier detection.

A coherent formula interpretation of our outlier is described below:

Positive outlier: Stock's quarter's mean plus [1.5 times (Q_3 - Q_1)]

Negative outlier: Stock's quarter's mean minus [1.5 times (Q_3 - Q_1)]

We identified both price outliers and volume outliers for each stock. In each sector, we added positive price outliers and negative price outliers to get the ratio of stocks that have an outlier in the sector on a specific and. And for the volume outliers, we are mainly focusing on the positive outlier. A low volume day for security is not particularly within our research interest.

Below is an example of how we evaluate whether a stock is an outlier on a certain date:

IFS('Q1'!B2>='1.5IQR$'!$R2,"1.5xIQR",'Q1'!B2<='1.5IQR$'!$N2,"-1.5xIQR")[6]

By using this formula, we are asking Excel to label "1.5xIQR" to positive outliers and to label "-1.5xIQR" to negative outliers, same for both price and volume data as long as their mean and 1.5 * IQR have been computed. Assigning "+1.5xIQR" instead of "1.5xIQR" does not seem flawed at first but when pasting cell texts to another sheet/file, Excel might interpret the "+" sign

[6] IFS statement allows us to input multiple criteria and only output value if the argument is deemed "true." Its syntax can be described below.

=IFS([Something is True1, Value if True1,Something is True2,Value if True2)

as an operator and therefore wipe out the text and display an error message. To visualize the outlier computation, we utilized conditional formatting that gives each positive outlier a green cell background and each negative outlier a red background with the white text color. The cell without an outlier were automatically denoted "#N/A."

We have run a COUNTIF function for each stock to track the total amount of outliers there are out of the 252 days. For price percentage change, on average there are 7.19 percent of dates or 18.1 days that exhibit outlier behaviors (either positive or negative), which is a reasonable amount that gives us sufficient resources to analyze further while keeping rigorous criteria to prevent regular market noises.

Correlation between Sectors

	Consumer Discretionary	Industrials	onsumer Staple	Utilities	Materials	IT	Health Care	Financials	munication Ser	Real Estate
Consumer Discretionary										
Industrials	82.53%									
Consumer Staples	42.08%	58.59%								
Utilities	31.30%	42.61%	55.57%							
Materials	67.01%	76.45%	46.41%	29.44%						
IT	54.61%	54.62%	29.18%	17.05%	34.76%					
Health Care	41.42%	46.33%	46.67%	29.28%	33.66%	37.63%				
Financials	74.71%	80.38%	46.60%	30.77%	72.79%	41.63%	32.22%			
Communication Service	68.87%	68.24%	56.33%	36.38%	59.08%	43.44%	38.36%	63.88%		
Real Estate	56.40%	62.70%	48.31%	46.71%	38.45%	39.97%	41.38%	54.34%	55.62%	
Energy	44.60%	42.71%	17.38%	6.20%	46.40%	27.64%	12.24%	55.54%	43.24%	25.06%

Figure 9 - Correlation between different sectors are shown above

Using the price outliers, both positive and negative ones, we have computed the percentage of stocks within each sector that exhibit outlier behaviors on each date. The method to derive sectoral information will be discussed later in this section. As seen at Figure *10* , we ran

correlations between sector's price outlier data and noticed that sectors behave quite differently in 2021 in terms of outlier percentage on each date. The average correlation coefficient of all relationships among 11 sectors is 46.72 percent (not weighted by the number of stocks within each sector), implying that it should be more informative if we break down our further study of S&P 500 stocks into sectors. It is also fascinating to note that within every sector, the amount of positive and negative outliers eventually cancel each other out, since the average percentages of the outliers in 2021 in every sector are all within 1 percent. This is unsimilar to the average price movement by sector in 2021, that stocks are expected to increase by at least five percent every year.

Microsoft Excel is equipped with a finance function to release sub-industry info of a given stock. After removing 6 stocks for which Excel could not provide sub-industry information, there are 47 sub-industries being labeled as 466 stocks from the Standard & Poor's 500 Index. We want to narrow these sub-industries down into sector using the GICS[7] methodology that classifies our industry labels into 11 sectors: Energy, Materials, Industrials, Utilities, Healthcare, Financials, Consumer Discretionary, and Consumer Staples, Information Technology, Communication Services, and Real Estate. In the January 2022 version of the methodology, the GICS structure is made of 11 sectors, 24 industry groups, 69 industries, and 158 sub-industries. Sectoral analysis may be meaningful if each sector demonstrates a different sensitivity toward sentiment data.

Before grouping stocks by industry, our outlier table is sorted alphabetically with stock tickers as the row headers and date as column headers. After assigning industry labels, we reorder the tickers and acquire a table stratified by sub-industry categories (sorted alphabetically

[7] *The Global Industry Classification Standard (GICS) was developed by MSCI in collaboration with S&P Dow Jones Indices to provide an efficient, detailed, and flexible tool for use in the investment process.*

within each sub-industry. This allows us to visualize the effect of sectoral analysis better, noticing that the same sector stocks tend to exhibit similar outlier tendencies on certain dates.

Sentiments

FRBS is the abbreviation for Federal Reserve Bank of San Francisco. It is a nonprofit public service institution that supports the economy through research, banking supervision, community development, and education. FRBS Daily Sentiment is one of the services they provide. The engine uses economics-related news articles from 24 major U.S. newspapers compiled by the news aggregator service Factiva to generate a sentiment score using lexical analysis. Lexical analysis is the process of converting a sequence of characters (such as in a computer program or web page) into a sequence of tokens (strings with an assigned and thus identified meaning). Individual article scores are then aggregated into a daily time-series measure of news sentiment. The index is constructed as a trailing weighted-average of time series, with weights that decline as the amount of time since article publication increases. The data is regularly updated at a weekly frequency.

In this report, we will compare the FRBS Daily Sentiment Index (represented as FRBS in the rest of the article) with % of stocks within each sector that had outlier behavior to see if there is a strong correlation between the news sentiments and stock price changes. The 2021 sentiment index is shown below.

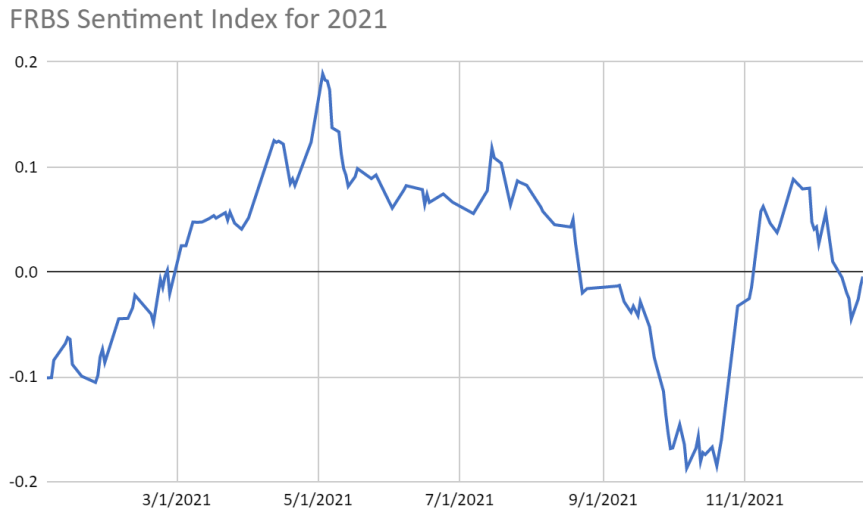


Figure 10 - Line chart of FRBS sentiment index for 2021

The negative sign means that on that day, the attitude toward S&P 500 companies shown by financial news is negative. And the positive sign means that the news about the S&P 500 companies on that day is positive. Our hypothesis here is that if there is a huge change in the sentiment index, there should also be a huge change in the stock prices, which can be represented by the high ratio of price outliers. Based on this hypothesis, we then calculated the % change of FRBS during 2021. The data graph is listed below.

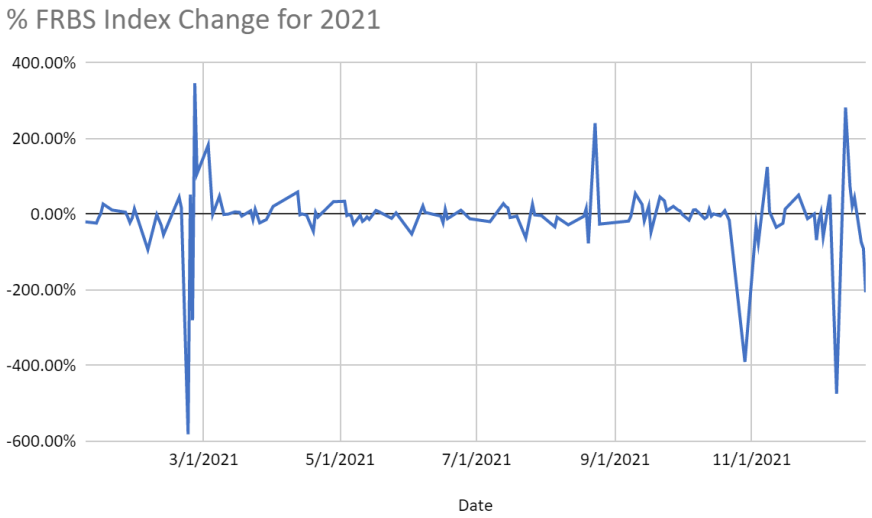


Figure 11 - % FRBS change for 2021

What we want to see in our model is around March 1, there should be many price outliers. And the ratio of outliers should be high for mid-August, late October, and late December because the percentage sentiment score change is very high during those days.

Analysis Results and Explanation

Sector (GICS)	CORREL w 0	CORREL w/o 0	FRBS difference w 0	FRBS difference w/o 0	diff 2 days Prior w/o 0
Communication Service	5.85%	7.68%	12.02%	13.19%	-5.69%
Consumer Discretionary	5.67%	2.22%	13.93%	15.30%	12.08%
Consumer Staples	8.74%	10.14%	6.62%	8.60%	-2.13%
Energy	-6.36%	-9.22%	2.40%	3.35%	-8.03%
Financials	0.46%	0.18%	8.05%	9.31%	-0.80%
Health Care	9.10%	10.45%	9.86%	11.36%	15.13%
Industrials	6.49%	6.49%	12.42%	9.61%	2.95%
Information Technology	5.92%	5.92%	4.24%	4.22%	5.39%
Materials	5.11%	6.58%	19.53%	27.99%	5.66%
Real Estate	7.56%	9.48%	12.61%	16.19%	7.75%
Utilities	0.73%	0.49%	0.87%	1.08%	2.56%

Figure 12 - Correlation table between stock price outliers for each sector and FRBS Daily Sentiment data.

Using the FRBS Daily Sentiment Data, we compiled a correlation table at the sector level that reflects how closely price outliers correlated with the sentiment scores. In the first two columns, we calculated the correlation between daily stock price outliers for a given sector and the FRBS Daily Sentiment Score. In the third and fourth columns, we check the correlation between daily stock price outliers and the day-to-day change in the FRBS Daily Sentiment Score (i.e., [FRBS(T) - FRBT(t-1)] / FRBS (t-1)) Day 2 minus Day 1). In the last column, we offset the price outlier data two days behind the change in FRBS sentiment data to account for a potential two-day lag between sentiment and price outlier results.

It can be evidenced that the correlation between the two measures is relatively low across the board, but there were some key standouts that indicate a possible relationship. For example, healthcare and consumer staples led the sectors with correlation scores of just over 10%. This

shows that out of the eleven sectors examined, healthcare and consumer staples are the likeliest to have stocks with price outliers resulting from dramatic news coverage. It is also important to note that the energy sector had a “negative correlation” with the FRBS data. One possible explanation could be because the energy sector experienced heavy losses in 2021, which resulted in many negative price outliers in our data set. Thus, as news coverage and sentiment increased, stock prices in the energy sector generally decreased. While that is a sufficient explanation for the negative direction of the correlation, it is also important to highlight the magnitude of the correlation, which was -9.22% in correlation without values of 0. This magnitude of the absolute value of 9.22% rivals the top movers of healthcare and consumer staples in terms of the percent change movement.

We also calculated the day-to-day change in FRBS sentiment data to compare with our price outliers, which yielded more positive correlation results. Shown by the columns with “FRBS Difference”, it can be seen that the day-to-day changes in daily sentiment aligned more with the price outliers than the actual daily sentiment. We believe there are a couple of reasons for this. . First, the news is something that is reacted to, which means that there is likely to be an element of time or lag involved. Only once investors have ample time to react to the news they can buy or sell a number of securities that can dramatically impact the price. Since the day-to-day change in sentiment shows how the tone of news changed, and therefore how people reacted, it drew a stronger correlation to price outliers. Second, the day-to-day change shows the direction the news is heading in, as opposed to the static measure of daily sentiment. In other words, daily sentiment shows how investors feel about the news, but those feelings that are being measured are likely already priced in. However, what is not priced in are the live changes in

sentiment, which are being measured by the day-to-day change in FRBS. Therefore, those increases and decreases in sentiment scores are a better reflection of how prices will move.

Lastly, we ran one final correlation analysis in an attempt to better capture the effect of lag in the market. The last column shows how we offset the price outlier data two days behind the day-to-day change in the FRBS sentiment data to see if accounting for a two-day lag yielded better correlation results. We ultimately saw varied results with no clear success or failure. For some sectors like healthcare and utilities, the correlation was highest in this measure, but lowest in other sectors such as consumer staples, financials, and industrials. However, these results may actually be quite informative , indicating that some sectors experience more news-related lag than others. Investors and trading algorithms could use this analysis to their advantage and attempt to implement better timing into their trades.

Using the sector level analysis for FRBS difference w/o 0, we can generate a number of sector-specific insights based on where correlation with sentiment is high and low. One sector that has high correlation with FBRS Daily Sentiment is Consumer Discretionary. Our analysis is that when consumers are feeling positive sentiment towards the market and economy, then they are more willing to spend their disposable income on discretionary goods. This is why discretionary goods are highly correlated with the FRBS sentiment data. A similar principle can be applied to the Materials sector. The materials sector primarily consists of items that are used in construction, building, and renovation, all of which tends to increase during times of high market sentiment. The sector with the lowest correlation to consumer sentiment is the Utilities sector. This makes a lot of sense, considering individuals need basic utilities such as electricity, gas, water, etc. for daily life and will pay those expenses regardless of their sentiment towards the market. Energy has the second lowest correlation to consumer sentiment, which follows the

same principle as utilities. Individuals will still need energy, power, and gas regardless of market performance and sentiment. Overall, our correlation data at the sector level is highly justifiable and intuitive.

Conclusion

In general, we found that there is a positive correlation between the FRBS sentiment index and the sectoral outlier behavior among S&P 500 stocks (except the energy sector). Different sectors respond to the sentiment score change differently. However, the general tendency is positive, which means when the sentiment score goes higher, the stock prices of the companies tend to go higher. The average correlation between the FRBS sentiment score and stock price change is about 4.58 percent, and the average correlation between the FRBS change to stock price change is about 10.93%. We prefer to use the percentage change correlation, which means 10.93% of the price change can be explained by the percentage change of sentiment score. This positive correlation supported our hypothesis that the sentiment scores and stock market price have a relationship. The stock market is volatile, and there are many factors that can cause a stock price change. Our research is not conclusive but it represents a possibility to explain the 2021 stock market behavior. We are looking forward to other research projects that explain the stock market price change and hope this report can be one that contributes to the study of how financial news and stock market behaviors are related.

Reference

Airnow. (August 25, 2021). Monthly number of active users selected leading apps that allow for online share trading worldwide from January 2017 to July 2021, by app (in 1,000s) [Graph]. In Statista. Retrieved April 26, 2022, from https://www.statista.com/statistics/1259822/global-etrading-app-monthly-active-users

Barbara Kollmeyer. (September 10, 2021). The $1 trillion that has flowed to global stocks in 2021 is bigger than the last 20 years combined. https://www.marketwatch.com/story/the-1-trillion-that-has-flowed-to-global-stocks-in-2021-is-bigger-than-the-last-20-years-combined-11631273525

United States Fed Fund Rate, Retrieved April 29, 2022, from https://tradingeconomics.com/united-states/interest-rate

Peter G. Peterson Foundation, March 15, 2021, What to know about all three rounds of coronavirus stimulus checks, https://www.pgpf.org/blog/2021/03/what-to-know-about-all-three-rounds-of-coronavirus-stimulus-checks#:~:text=The%20first%20round%20of%20stimulus%20payments%20were%20authorized%20under%20the,a%20total%20of%20%24292%20billion.

Zack Friedman, July 27, 2021, 46% Of Stimulus Checks Were Invested In The Stock Market?, https://www.forbes.com/sites/zackfriedman/2021/06/27/46-of-people-invested-their-stimulus-checks-in-the-stock-market/?sh=761fbd0e72f0

Maja Pawinska Sims, Jan 30, 2019, Study: Investors Increasingly Make Decisions Based On Digital & Social Media Sources, https://www.provokemedia.com/latest/article/study-investors-increasingly-make-decisions-based-on-digital-social-media-sources#:~:text=Moreover%2C%20most%20investors%20(88%25),investment%20decisions%20from%20digital%20sources.

Market Holidays & Trading Hours. NYSE. (n.d.). Retrieved May 2, 2022, from https://www.nyse.com/markets/hours-calendars

GICS - Global Industry Classification Standard. MSCI. (n.d.). Retrieved May 2, 2022, from https://www.msci.com/our-solutions/indexes/gics

Tesla stock price for 2021, Yahoofinance, Retried May 3, 2021, https://finance.yahoo.com/quote/TSLA/history?period1=1609459200&period2=1640908800&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

Shapiro, A. H., Sudhof, M., & Wilson, D. (2020, March 1). Measuring news sentiment. Federal Reserve Bank of San Francisco. Retrieved May 2, 2022, from https://www.frbsf.org/economic-research/publications/working-papers/2017/01/

Zacks Equity Research, December 2, 2021, Stock Market News for Dec 2, 2021, https://www.google.com/url?q=https://www.yahoo.com/video/stock-market-news-dec-2-142502118.html&sa=D&source=docs&ust=1652128610810854&usg=AOvVaw2m3qPi53u3AABOs47Eh0oE

Zacks Equity Research, May 12, 2021, Stock Market News for May 12, 2021, https://www.nasdaq.com/articles/stock-market-news-for-may-12-2021-2021-05-12

Appendix

Our Journey

Alternative Analysis

Initially, we sought to use the GDELT Global Knowledge Graph (GKG) Dataset since it tracks data points that have nearly perfect compatibility with the goal of our project. GDELT monitors the world's news to quantify and codify events, allowing the platform to generate a real-tune global news graph. Not only does the database offer exceptional data services and features, but it also has a clear purpose as well. GDELT seeks to provide global insight by using advancing technology to break down both language and access barriers. Additionally, GDELT offers a number of complementary data analysis capabilities that are either built-in or enabled through partnerships, such as our intended platform of Google BigQuery. Yet despite all of GDELT's qualities, their data files were far too large and complex for our machines to handle. Additionally, the compatibility between GDELT and Google BigQuery proved to be increasingly complex, and we had to alter our approach if we wanted to continue with GDELT.

We explored the GDELT by coding it in python. One of the challenges of dealing with GDELT is that its size makes implementation in a SQL database. It is easier to extract a subset of the data that we wish to work with, and do our actual data investigation afterward. We attempted to generate the data through the tutorial (https://nbviewer.org/github/JamesPHoughton/Published_Blog_Scripts/blob/master/GDELT%20Wrangler%20-%20Clean.ipynb) but it did not successfully produce the output.

1. We identified the files we need to download from the GDELT database. We extracted this list to collect and process the data.

```
import requests
import lxml.html as lh

gdelt_base_url = 'http://data.gdeltproject.org/events/'

# get the list of all the links on the gdelt file page
page = requests.get(gdelt_base_url+'index.html')
doc = lh.fromstring(page.content)
link_list = doc.xpath("//*/ul/li/a/@href")

# separate out those links that begin with four digits
file_list = [x for x in link_list if str.isdigit(x[0:4])]
```

2. We extracted relevant information (country, specific events, etc) from the downloaded file. Then, we created a set of files that mirrored GDELT in quantity and format.

```
import os.path
import urllib
import zipfile
import glob
import operator

local_path = '/Users/me/Desktop/GDELT_Data/'

fips_country_code = 'UK'

for compressed_file in file_list[infilecounter:]:
    print compressed_file,

    # if we dont have the compressed file stored locally, go get it. Keep trying if necessary.
    while not os.path.isfile(local_path+compressed_file):
        print 'downloading,',
        urllib.urlretrieve(url=gdelt_base_url+compressed_file,
                           filename=local_path+compressed_file)

    # extract the contents of the compressed file to a temporary directory
    print 'extracting,',
    z = zipfile.ZipFile(file=local_path+compressed_file, mode='r')
    z.extractall(path=local_path+'tmp/')

    # parse each of the csv files in the working directory,
    print 'parsing,',
    for infile_name in glob.glob(local_path+'tmp/*'):
        outfile_name = local_path+'country/'+fips_country_code+'%04i.tsv'%outfilecounter

        # open the infile and outfile
        with open(infile_name, mode='r') as infile, open(outfile_name, mode='w') as outfile:
            for line in infile:
                # extract lines with our interest country code
                if fips_country_code in operator.itemgetter(51, 37, 44)(line.split('\t')):
                    outfile.write(line)
            outfilecounter +=1

        # delete the temporary file
        os.remove(infile_name)
    infilecounter +=1
    print 'done'
```

3. Finally, we transferred the data we sampled into csv files. We used Pandas Dataframe to load the data, save them to a pickle and delete the temporary files. The Algorithms we used is simple: we build dataframe out of the temporary files and merge them into a big dataframe.

```
import glob
import pandas as pd

# Get the GDELT field names from a helper file
colnames = pd.read_excel('CSV.header.fieldids.xlsx', sheetname='Sheet
                         index_col='Column ID', parse_cols=1)['Field

# Build DataFrames from each of the intermediary files
files = glob.glob(local_path+'country/'+fips_country_code+'*')
DFlist = []
for active_file in files:
    print active_file
    DFlist.append(pd.read_csv(active_file, sep='\t', header=None, dty
                              names=colnames, index_col=['GLOBALEVENT

# Merge the file-based dataframes and save a pickle
DF = pd.concat(DFlist)
DF.to_pickle(local_path+'backup'+fips_country_code+'.pickle')

# once everythin is safely stored away, remove the temporary files
for active_file in files:
    os.remove(active_file)
```

Variable Definition

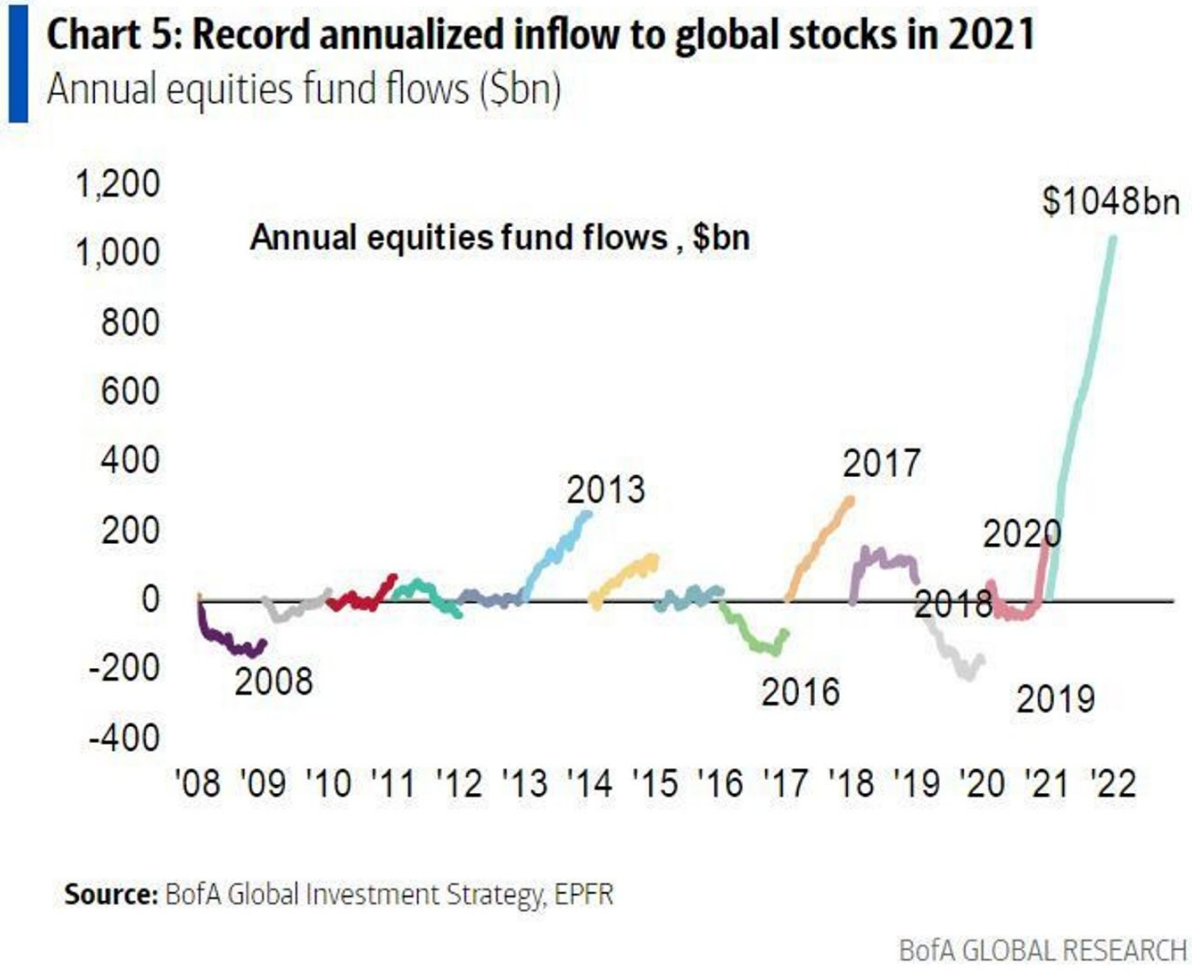
In this report, there are 12 variables used. Nine of them are about data properties. They are Ticker, Date, Sector, price_outlier, ratios_pri, ratio_vol, FRBS, %FRBS. These variables measure one characteristic of the data. For example, what date this data is on, and how many outliers are observed on that date. The remaining four variables measure correlation related data. They are Correlation w 0, Correlation w/o 0, %FRBS w o, % FRBS w/o 0. These four variables measure how strongly the outlier market behavior is associated with market sentiment change under different scenarios. Below is an explanation of these variables.

- Ticker: The symbols of the stock, which is the main identifier of a company in the stock market. It will only be used in the raw data, which means it will not be used in the final calculation and report.
- Sector: this title represents which sector a company belongs to. There are 11 sectors in total. They are Communication Services and Consumer Discretionary. Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate, and Utilities. The Criteria to identify the sector a company is in is GICS (Global Industry Classification Standard).
- Date: this variable is the key to linking sentiment scores and outliers' behaviors. It also enables us to create a time-series graph.
- Price outlier: this variable measures if the price change on a specific day for a specific stock is an outlier. We defined a price change of more than 1.5 IQR of the annual price change as an outlier.
- FRBS: FRBS represents the Federal Reserve Bank of San Francisco Sentiment Index. This variable measures the attitude of society toward news. The larger the number means

the more extreme the society's attitude is. A positive sign means that there is good news about the market on that day and a negative sign means there is bad news about the market.

- %FRBS: this variable measures the percentage change of the FRBS score from the previous. It will be used to calculate the %FRBS w o and % FRBS w/o 0.
- Ratio_pri: this represents what percentage of all companies in a sector has a price outlier on a specific day. It will be one of the two variables that are computed with FRBS to find the correlation between outliers and sentiment changes.
- Ratio_vol: this represents what percentage of all companies in a sector has a volume outlier on a specific day. It will be the other variable of the two variables that are computed with FRBS to find the correlation between outliers and sentiment changes.
- Correlation w 0: it measures the correlation between ratio_pri and FRBS. The way to calculate this variable is using the CORR function in Google Sheets. Under this variable, the days that ratio_pri equals 0 are not removed.
- Correlation w/o 0: it measures the correlation between ratio_pri and FRBS. The way to calculate this variable is using the CORR function in Google Sheets. Under this variable, the days that ratio_pri equals 0 are removed.
- %FRBS w 0: it measures the correlation between ratio_pri and %FRBS change from the previous day. The way to calculate this variable is using the CORR function in Google Sheets. Under this variable, the days that ratio_pri equals 0 are not removed.
- % FRBS w/o 0: it measures the correlation between ratio_pri and %FRBS change from the previous day. The way to calculate this variable is using the CORR function in Google Sheets. Under this variable, the days that ratio_pri equals 0 are removed.

Appendix Figure 1

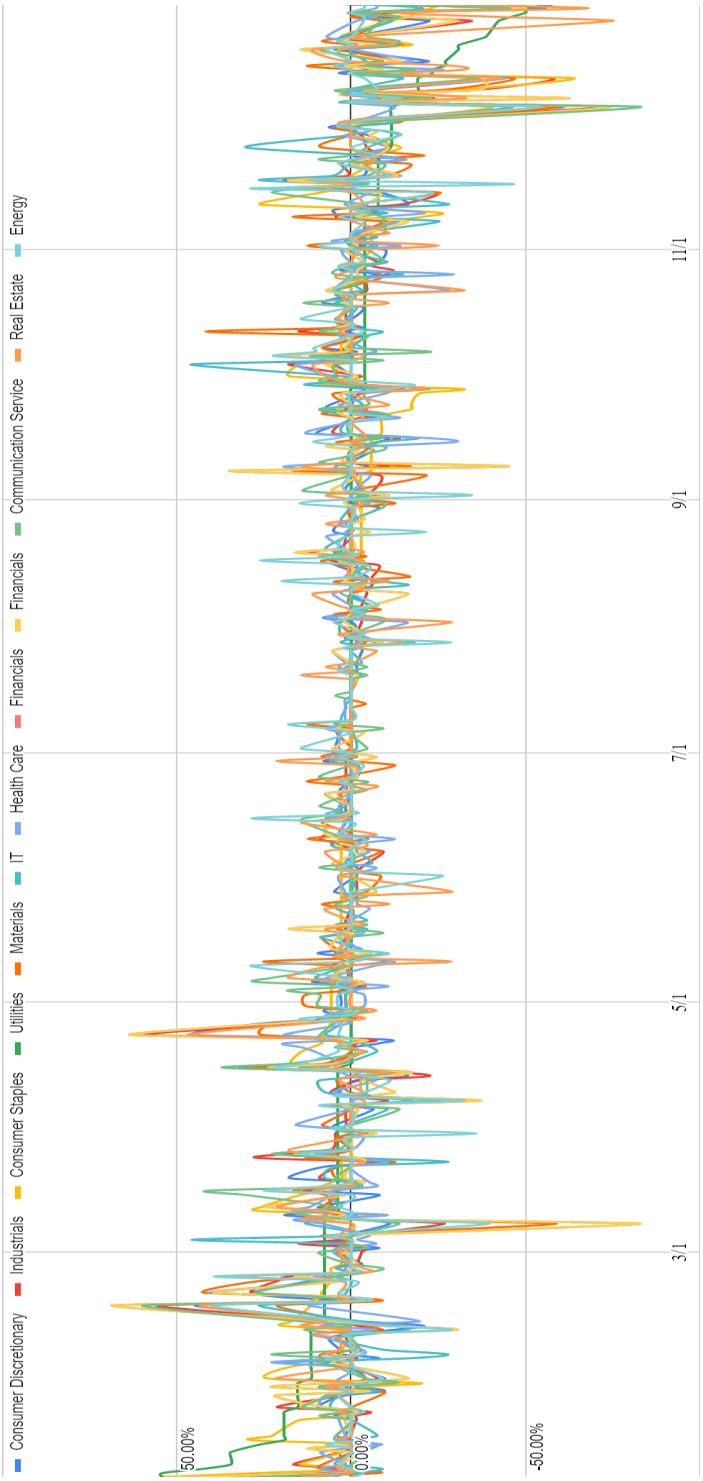


Appendix Figure 2

Sector (GICS)	*Sub-Industries*	SUM of Count of Sub
Information Technology	Software & IT Services	41
	Semiconductors & Semiconductor Equipment	16
	Computers, Phones & Household Electronics	6
	Electronic Equipment & Parts	5
Information Technology Total		68
Industrials	Machinery, Equipment & Components	26
	Professional & Commercial Services	14
	Aerospace & Defense	10
	Freight & Logistics Services	9
	Passenger Transportation Services	5
	Construction & Engineering	2
	Office Equipment	1
Industrials Total		67
Consumer Discretionary	Hotels & Entertainment Services	18
	Specialty Retailers	13
	Automobiles & Auto Parts	8
	Textiles & Apparel	6
	Homebuilding & Construction Supplies	6
	Diversified Retail	5
	Consumer Goods Conglomerates	4
	Leisure Products	3
Consumer Discretionary Total		63
Health Care	Healthcare Equipment & Supplies	29
	Healthcare Providers & Services	12
	Pharmaceuticals	10
	Biotechnology & Medical Research	6
Health Care Total		57
Financials	Banking Services	22
	Insurance	18
	Investment Banking & Investment Services	15
	Financial Technology (Fintech) & Infrastructure	1
Financials Total		56
Consumer Staples	Food & Tobacco	16
	Food & Drug Retailing	6
	Personal & Household Products & Services	5
	Beverages	5
	Household Goods	3
Consumer Staples Total		35
Real Estate	Residential & Commercial REIT	27
	Real Estate Operations	1
Real Estate Total		28
Utilities	Electrical Utilities & IPPs	17
	Multiline Utilities	6
	Renewable Energy	2
	Natural Gas Utilities	1
Utilities Total		26
Materials	Chemicals	13
	Containers & Packaging	6
	Metals & Mining	3
	Construction Materials	2
Materials Total		24
Communication Service	Media & Publishing	12
	Telecommunications Services	5
	Communications & Networking	5
Communication Service Total		22
Energy	Oil & Gas	14
	Oil & Gas Related Equipment and Services	6
Energy Total		20
Grand Total		**466**

Appendix Figure 3

Appendix Figure 4



We also explored Natural language processing for newspaper articles. The tutorial is from https://analyticsindiamag.com/how-to-scrape-summarize-convert-news-articles-into-text-files/. Basically, we achieved this by using web scraping and natural language processing (NLP).

1. We included the newspaper library and nltk library for NLP.

   from newspaper import Article

   import nltk

2. The *punkt* of nltk library is used to tokenize the sentences in order to be used for NLP. So we need to download *punkt* sentence tokenizer.

   nltk.download('punkt')

3. We include the URL of the newspaper sources
4. Set the language of the article which is to be scraped and summarized. Define an object for further use.

   article = Article(url, language="en") # en for English

5. Download, parse and perform NLP on the news article

   article.download()

   article.parse()

   article.nlp()

6. The article is now scraped and downloaded. We can print useful information on the console.

   print("Article Title:")

   print(article.title) #prints the title of the article

   print("\n")

   print("Article Text:")

```
print(article.text) #prints the entire text of the article
print("\n")
print("Article Summary:")
print(article.summary) #prints the summary of the article
print("\n")
print("Article Keywords:")
print(article.keywords) #prints the keywords of the article
```

7. The above result can be written in a text file. The following lines of codes are used to write tt into a text file

```
file1=open("NewsFile.txt", "w+")
file1.write("Title:\n")
file1.write(article.title)
file1.write("\n\nArticle Text:\n")
file1.write(article.text)
file1.write("\n\nArticle Summary:\n")
file1.write(article.summary)
file1.write("\n\n\nArticle Keywords:\n")
keywords='\n'.join(article.keywords)
file1.write(keywords)
file1.close()
```

```
#Resources from https://analyticsindiamag.com/how-to-scrape-summarize-convert-news-articles-into-text-files/
from newspaper import Article
import nltk
#install library for Nature Language processing
nltk.download('punkt')
#url link
url= 'https://economictimes.indiatimes.com/markets/stocks/news/jefferies-sees-87-bull-case-upside-in-shares-o
#Set the language of the article which is to be scraped and summarized. Define an object for further use.
article = Article(url, language="en") # en for English
article.download()
article.parse()
article.nlp()
#print information
print("Article Title:")
print(article.title) #prints the title of the article
print("\n")
print("Article Text:")
print(article.text) #prints the entire text of the article
print("\n")
print("Article Summary:")
print(article.summary) #prints the summary of the article
print("\n")
print("Article Keywords:")
print(article.keywords) #prints the keywords of the article
file1=open("NewsFile.txt", "w+")
file1.write("Title:\n")
file1.write(article.title)
file1.write("\n\nArticle Text:\n")
file1.write(article.text)
file1.write("\n\nArticle Summary:\n")
file1.write(article.summary)
file1.write("\n\n\nArticle Keywords:\n")
keywords='\n'.join(article.keywords)
file1.write(keywords)
file1.close()
```


```
#Resources from https://analyticsindiamag.com/how-to-scrape-summarize-convert-news-articles-into-text-files/
from newspaper import Article
import nltk
#install library for Nature Language processing
nltk.download('punkt')
#url link
url= 'https://economictimes.indiatimes.com/markets/stocks/news/jefferies-sees-87-bull-case-upside-in-shares-o
#Set the language of the article which is to be scraped and summarized. Define an object for further use.
article = Article(url, language="en") # en for English
article.download()
article.parse()
article.nlp()
#print information
print("Article Title:")
print(article.title) #prints the title of the article
print("\n")
print("Article Text:")
print(article.text) #prints the entire text of the article
print("\n")
print("Article Summary:")
print(article.summary) #prints the summary of the article
print("\n")
print("Article Keywords:")
print(article.keywords) #prints the keywords of the article
file1=open("NewsFile.txt", "w+")
file1.write("Title:\n")
file1.write(article.title)
file1.write("\n\nArticle Text:\n")
file1.write(article.text)
file1.write("\n\nArticle Summary:\n")
file1.write(article.summary)
file1.write("\n\n\nArticle Keywords:\n")
keywords='\n'.join(article.keywords)
file1.write(keywords)
file1.close()
```


8. Finally, we will get the following result with the URL used in this example saved into a text file.

```
Article Title:
Jefferies sees 87% bull case upside in shares of this electricals company


Article Text:
Higher spend on capex, housing and infra development in India

Faster ramp-up in FMEG, resulting in better margins

Subdued traction in capex, housing and infra development in India

Lower offtake in FMEG

Excess volatility in commodities impacting margins

Sharp volatility in commodities, RM procurement

Slowdown in capex/housing and

Rising competition impacting pricing
```

```
Article Summary:
It sees a 37 per cent upside potential in base case while 87 per cent in bull case.
In bear case, it sees a 38 per cent potential downside.“Our conviction is underpinned by robust prospects (revival in ca
pex, infra, housing), its market leadership (21 per cent organized share), a ramp-up in B2C FMEG (37 per cent sales CAGR
 in 5 years) and benefits accruing from strategic initiatives.
The company currently trades at a 25-40 per cent P/E discount to Havells and Whirlpool .”She has a base case target of R
s 3,300 on the counter.
Polycab's focus on in-house manufacturing and deeper penetration could drive +18 per cent sales CAGR over FY22-25.“Over
FY22-25e, we estimate Polycab's consolidated sales/PAT to post +14 per cent/30 per cent CAGR – despite factoring rising
input costs.
Copper volatility is typically a pass-through in the Cables and Wires industry,” she said.For the bull case scenario, Je
fferies says two things need to happen:Conversely, bear case scenario may kick in, if there isAmong key risks to the Jef
feries call are
```

```
Article Keywords:
['case', 'electricals', 'cagr', 'housing', 'sees', 'bull', 'rs', '87', 'capex', 'volatility', 'infra', 'upside', 'cent',
 'sales', 'shares', 'jefferies', 'company']
```