SARS-CoV-2 impact analysis on the US stock market

Young Hoon Kim, Tyler Chi, Raghav Kaushik

July 2021

Abstract

Since the first reported case of SARS-CoV-2 (COVID-19) in December 31, 2019 in Wuhan, China, COVID-19 has affected our lives significantly. In this study, we are aiming to understand whether the impact of COVID-19 to the US stock market has been different among stocks of major industries, further analyzing trend similarities among different industries in the stock trading realm. In addition, we will provide an interdependent forecasting model in an effort to provide insights of industry stock trends into the future during the COVID period. Along with a series of exploratory data analyses to understand the trend similarities, we will also employ Granger causality test to check whether any industry "Granger-causes" other industry's stock movements, and construct a forecasting model with the Vectorized Autoregressive model.

1. Introduction

United States stock market has often been used as an indicator to understand both social and economic movement in the world. By no means does the stock market dictate the economic well-being of a country, but it does shine light on how consumers and individuals react to current affairs. On January 21 2020, the Centers for Disease Control and Prevention (CDC) confirmed the first US coronavirus case, and by March 13th President Trump declared COVID-19 a national emergency. The US stock market began to transition from being bullish to "a bear market hitting its trough on March 23, 2020, when the S&P 500 index fell to its low of 2,237.40—a drop of 34%" [1]. As unemployment claims rose to over 3.2 million, public sentiment and outlook towards life began to diminish. On March 27th President Trump signed the CARES Act, a $2 trillion relief package which included direct payments to over 60 million Americans. With stimulus checks gradually rolled out coupled with the news of vaccines being developed, investor sentiment began looking optimistic. From March 2020 to March 2021, "the Dow Jones, S&P 500, and Nasdaq have soared 76, 76, and 95 percent respectively, making...one of the best 365-day stretches since World War II" [2]. Evidently, COVID-19 has affected the US stock market and this has led us to analyze the stock behavior in different industries based on the progression of COVID-19. In this paper, we will examine whether or not COVID-19 has had an equal impact across all stocks of all sectors.

Going into this project, we knew that there was a lot of stock market volatility during the coronavirus pandemic. Also, it definitely seemed like some industries suffered more than others, and some industries actually thrived during Covid. For the purpose of this research paper, we split the pandemic into 2 phases, because we were interested in looking at how industries behaved when Covid first came, how industries behaved after people got used to Covid, and when vaccines started to get distributed.

The coronavirus response director Deborah Birx, MD, declared in early August that we were entering a new phase of the pandemic. This decision came as the United States started to get a hold of the pandemic, and vaccines were starting to enter their Phase 3 trials [3]. We decided to use this as the dividing point in the pandemic, to define "Phase 1" and "Phase 2" for the purpose of our research. The idea was that industry behavior might differ between these two phases. For this research paper, we will refer to the beginning of Covid (early January) to this dividing point as "Phase 1", and the period after this dividing point to present day will be referred to as "Phase 2".

2. Preliminary Knowledge

For the past century, people have been putting great effort and time to understand time series in an effort to foresee the future. Since the future is always shrouded with uncertainty, people have been devising methods to understand the past, present, future. There are 2 high-level ways of approaching time series, which are parametric and non-parametric ways. Parametric approach attempts to create a model that mimics the movement of time series on a basis of a parametric function, whereas non-parametric approach employs more of a computational approach to construct a model that resembles the time series with hopes of it forecasting the future correctly. For the parametric approach in

time series forecasting, it is important that the time series is stationary.

2.1. Stationarity

Stationary process is a process whose unconditional joint probability distribution does not change when shifted in time. In other words, for stationary process, even as time progresses its mean stays the same. If a process is strictly stationary and has finite variance, then the covariance function must depend only on the time lag. [4] Weaker form of stationarity can be achieved if

- The mean function is constant over time
- $\gamma_{t,t-k} = \gamma_{0,k}$ for all time t and lag k

In this paper, we are going to assume the weaker form of stationarity when we check for stationarity of time series in COVID-19, stock, and Google Trends data.

2.2. Autoregressive (AR) Process

Autoregressive processes are essentially regressions on themselves. $AR(p)$ process is a linear combination of the p most recent past values of itself plus e_t that explains everything new in the series at time t that is not explained by the past values. [5] pth-order of autoregressive process Y_t can be expressed as the following: $Y_t = AR(p)$ where

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + e_t \quad (1)$$

2.3. Vector Autoregression (VAR)

VAR model requires stationarity of the series as prerequisites. It describes the evolution of a set of k variables over time. VAR models are characterized by their order, which is the number of time lags in the past that the model uses. The first order of VAR, VAR(1), can be expressed as the following:

$$\begin{aligned} y_{1,t} &= c_1 + \Phi_{11,1} y_{1,t-1} + \Phi_{12,1} y_{2,t-1} + \ldots + \Phi_{1k,1} y_{k,t-1} + \varepsilon_{1,t} \\ y_{2,t} &= c_2 + \Phi_{21,1} y_{1,t-1} + \Phi_{22,1} y_{2,t-1} + \ldots + \Phi_{2k,1} y_{k,t-1} + \varepsilon_{2,t} \\ &\vdots \\ y_{k,t} &= c_k + \Phi_{k1,1} y_{1,t-1} + \Phi_{k2,1} y_{2,t-1} + \ldots + \Phi_{kk,1} y_{k,t-1} + \varepsilon_{k,t} \end{aligned} \quad (2)$$

Above equation is a expanded view of VAR(1) that shows each series being modeled by it's own lag and other series' lag. This set of equations are of order 1 because they only contain up to one lag of each of the predictors. If you are to increase the number of lags to p for the VAR model, which will be expressed as VAR(p), the pth order VAR model equation is:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + e_t \quad (3)$$

where y_{t-p} indicates the variable's value at pth lag of y_t, c indicates the constant intercept, A_p indicates a time-invariant matrix, and e_t indicates error terms. A VAR model is particularly useful when you need to understand the influence of multiple time series to one another; furthermore, it can forecast into the future based on the model constructed using those multiple time series.

2.4. Granger causality

The basis behind VAR is that each time a series in the system can predict its time series with past values of itself with other relevant time series. Granger Causality test provides the existence of this relationship among time series data at hand. When a series x can be used to predict a series y, x can be said to "Granger-cause" y. Key thing to notice is that even though x "Granger-cause" y, it does not prove that there is an actual causal relationship between x and y. In other words, x can be used as one of the predictors to predict y because x "Granger-cause" y, it does not mean x causes y to happen.

The null hypothesis of the Granger Causality Test is the following:

$$H_0 : \mathbb{P}[Y(t+1) \in A \mid \mathcal{I}(t)] \neq \mathbb{P}[Y(t+1) \in A \mid \mathcal{I}_{-X}(t)] \quad (4)$$

where $\mathbb{P}$ refers to probability, A is an arbitrary non-empty set, and $\mathcal{I}(t)$ and $\mathcal{I}_{-X}(t)$ respectively denote the information available as of time t in the entire universe, and that in the modified universe in which X is excluded. If the above hypothesis is accepted, we say that X Granger-causes Y. [6] In other words, Granger's causality tests the null hypothesis that the coefficients of past values in the regression equation is 0. Therefore, if p-value obtained from the test is less than the significance level of 0.05, then, you can safely reject the null hypothesis.

3. Data

For this project, we were interested in the following industries: Tech, Finance (fin), Travel, Vaccine Companies (vac), Retail, and Energy Companies (eng). We picked a couple stocks for each industry:

- Tech: Apple, Amazon, Microsoft, Google, and Salesforce
- Finance: Bank of America, J.P. Morgan, Goldman Sachs, Wells Fargo, and Citigroup
- Travel: United Airlines, American Airlines, Delta Airlines, Expedia Group, and Uber
- Vaccine: Johnson n Johnson, Moderna, Pfizer, and AstraZeneca
- Retail: Macy's, Walmart, Gap, Nordstrom, and Target

- Energy: Chevron, Exxon Mobil, First Solar, General Electric

3.1. Datasets

This study pulled data from several sources: The Yahoo-Finance API was used to pull stock information on the stocks that we were interested in. This API provides daily historical data of the stock, such as the opening price, closing price, and dividend information.

We used data from the Centers for Disease Control and Prevention (CDC) to get the Covid-19 Death Data. This data was very specific, as it included which state, submission date, the total number of cases, the number of confirmed cases, as well as the new number of cases for that day.

We used Google Trends to look into the global search interests of ”Covid” and ”Vaccine”. Google Trends provided this data as a CSV, with the daily global search interests of the two terms.

3.2. Data preprocessing

In general, we were more interested in longer term trends (weekly) rather than day to day trends. This is because there can sometimes be a lag time in how people respond to the news. Another issue is that Covid data and Google search Data comes daily, whereas stock data comes only on trading days (Monday through Friday, if there are no holidays that week.) Because of this, if would be very difficult to analyze daily trends, as a spike in Covid deaths over the weekend might not reflect in the stock market until Monday, when the stock market opens (assuming there is no holiday.) Because of this, we grouped all the data into weeks, starting on Monday, because Monday is the first day of the trading week.

The CDC Covid Data was originally separated by state. However, the rest of our data is not state-specific, so we just aggregated it to represent the United States as a whole. We then grouped the rows by week, and summed up the values (specifically the new case count, which is what we were most interested in). By doing this, we were able to get the number of new cases per week for the entire United States (weeks starting on Monday.)

We were able to get the Google Trends data in a weekly format, however the date format was different from the date formats in the Covid and Stock data, so the format had to be converted.

4. Analysis

In this section, we will share the results of various analyses done to answer poised questions earlier.

4.1. Exploratory Data Analysis

4.1.1 COVID-19 data

In order to first understand the relationship between COVID-19 data and stocks data we first summarized the grouped CDC data. In figure 2 you can see the statistics of the grouped CDC data that is represented of all the states in US. We can see that at the 50% percentile, the median of new cases and new deaths are 310,882 and 5,614 respectively.

	new_case	new_death
count	7.800000e+01	78.000000
mean	4.334906e+05	7697.512821
std	4.324213e+05	6444.007214
min	4.300000e+01	0.000000
25%	1.570230e+05	3844.750000
50%	3.108820e+05	5614.000000
75%	4.587092e+05	10510.500000
max	1.758429e+06	24942.000000

Figure 1: Statistics summary of grouped CDC data

In figure 3 when we graph new COVID cases over time (weekly), we can see the graph peak slightly after January 2021 with the rise of the slope starting around October 2020. Based on the graph we see 3 notable spikes up until February 2021 and these are likely to be attributed to multiple factors such as poor political leadership, holidays, and summer traveling (public getting impatient and fed up staying indoors).

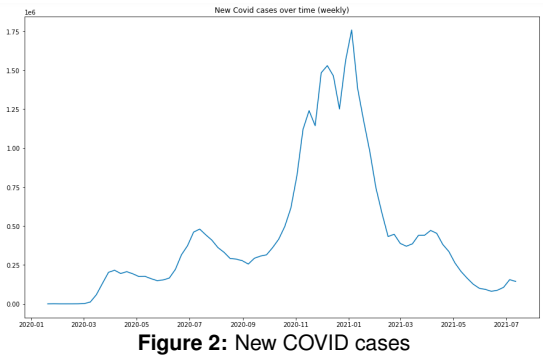

Figure 2: New COVID cases

In the Figure 4 COVID new deaths (weekly) graph, it illustrates the similar patterns and behaviors on how the public became infected with new cases. We clearly observe that the number of deaths was much steeper in the first peak around April 2020 (compared to the peak of new cases in

the same time frame) until it reaches the max peak around February 2021, which again is attributed to the similar reasons mentioned above for new cases.

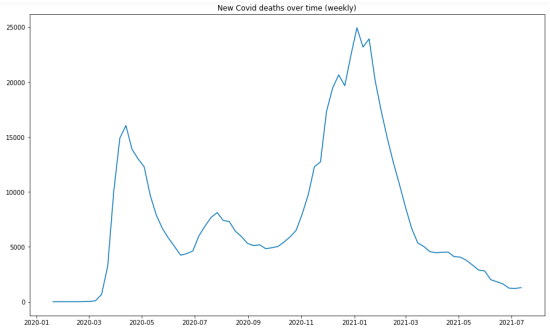

Figure 3: New COVID deaths

When we look at a simple time series of stock data based on % price changes from end of day price and opening day price (beginning of week and of week) in figure 5, we see some interesting trends. First, most of the industries have rapid spikes starting on March 2020, which is in line with the timeline of President Trump declaring a national emergency and COVID cases rising. We also observe that stocks had sharp negative % changes towards the summer of 2020, which exemplifies investors realizing that summer travel/holiday plans would naturally increase COVID cases and deaths, thus shorting the market. Adding on, at first glance it seems most of the industries slightly taper off with volatility after the summer of 2020, but the time series doesn't allow us to actually compare the true magnitude of dips and rises based on % changes which is why we go deeper into industry comparisons in the next sections.

4.1.2 Trend analysis between industries

We were interested in seeing if there was a significant difference in industry stock behavior between 2 specific phases: Phase 1 was the start of COVID-19 to late July, when we started getting vaccines approved by the FDA. The second phase was from early August to present day. We were specifically interested in seeing how stocks correlated with each other. For example, two industries equally impacted by COVID-19 might behave similarly - they would probably just move with the overall market. So if the overall market was going up, then similarly behaved industries might also both move up with the market. However, if one industry started significantly deviating from another industry, we could infer that they were behaving differently.

- **Tech vs Travel**

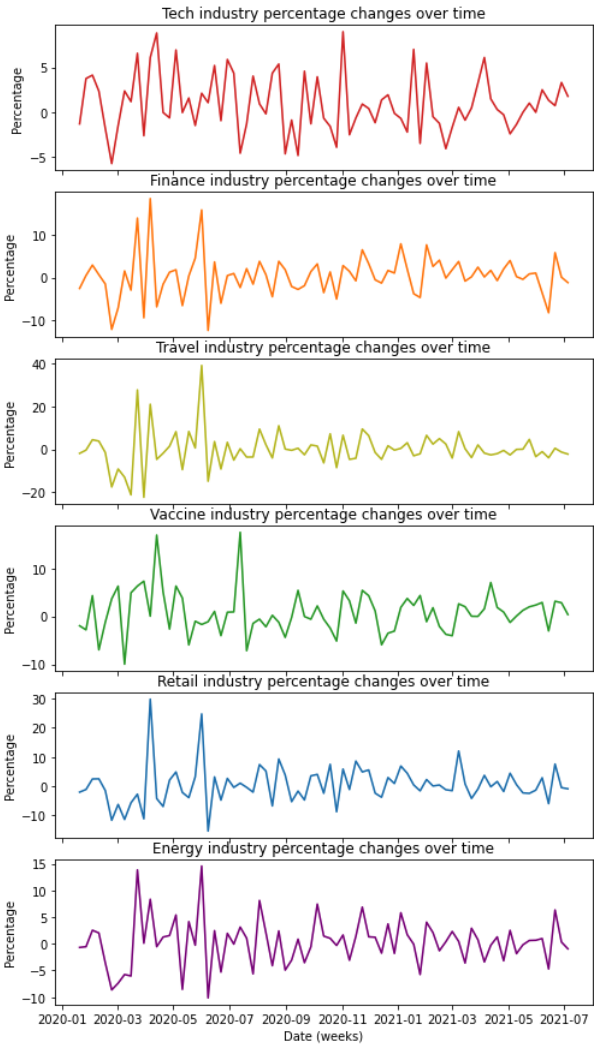

Figure 4: Stock movement by industries (weeks)

Travel was probably one of the hardest hit industries, especially early on in the pandemic. Especially because people were not traveling much, the travel industry suffered.

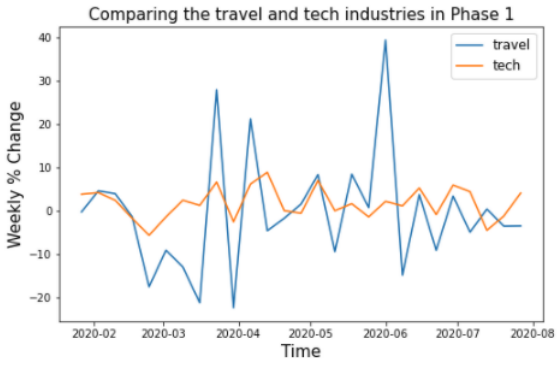

Figure 5: Comparing the tech vs travel industries in Covid phase 1

From Figure 5, it is clear that in the early phase of Covid, the travel industry stocks were much more volatile than tech stocks. However, Phase

2 tells a much different story.

Figure 6: Comparing the tech vs travel industries in Covid phase 2

It seems like in the Phase 2 of Covid, these two stocks were equally volatile, not necessarily correlating strongly with each other though. It is possible that in Phase 2, people started traveling more, and investors became more confident in investing in travel stocks.

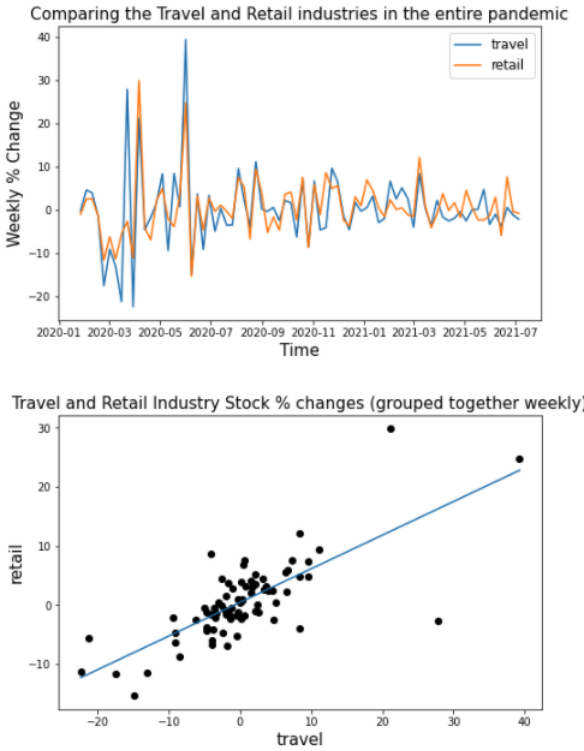

Figure 7: Comparing the retail and travel industries in the entirety of Covid

While travel and tech seemed to differ greatly, especially in the early phase of Covid, it seems like travel and retail seemed to match each other pretty closely over the entirety of Covid. This does make logical sense that the two industries would match pretty closely. Especially when Covid was extremely rampant, people were unwilling to go out to travel, and for similar reasons they were unwilling to go out shopping as much as they used to.

4.1.3 Taking into account Google Trends

We were interested in seeing how stocks from the various industries correlated with two specific google search terms: ”Covid” and ”Vaccine”. We thought that possibly when people were very worried about Covid, its search frequency would increase greatly, and at the same time there might be more fear in the marketplace. We also surmised that as Vaccines were beginning to become FDA approved, people might become more confident in the stock market. These reasons are why we chose the terms ”Covid” and ”Vaccine.” Similar to the rest of the data, we grouped the google trend data weekly, and expressed each week as a percentage change from the last week.

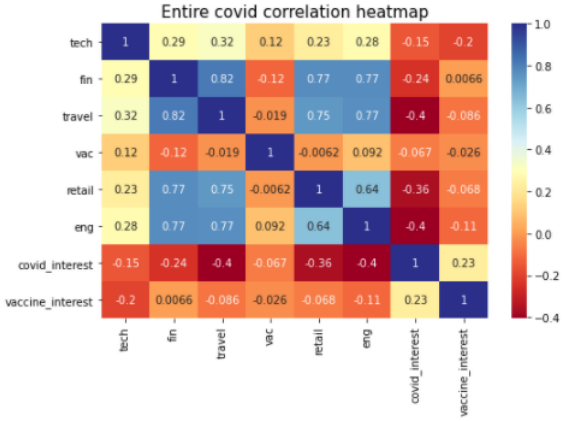

Figure 8: Correlation values of industry performance alongside how frequently ”Covid” and ”Vaccine” were searched on Google

Looking at Figure 8, we can see the correlation values between the various industry performances, as well as search terms. It seems like the term ”vaccine” did not correlate very strongly with any of the industry performances.

The search term ”Covid” did have stronger correlations than ”Vaccine” though. It had strong negative correlations with the retail and energy industries. It makes sense that it would have a strong negative correlation with the retail industry. As people were become more worried about Covid, and searching it on Google more, they would also be more worried about going out to shop, thus hurting retail companies.

Looking at Figure 9, the strongest correlation is between the new death delta and the Google Trend Covid interest delta. It makes sense that as weekly deaths increased, the search interest of ”Covid” on Google would also increase.

Figure 10 shows a correlation heatmap of all of the weekly delta variables that we studied: industry stock movement, CDC data (case and death count), and Google Trend data around the search terms ”Covid” and ”Vaccine”. Between the Google Trend data and the actual case/death count, the

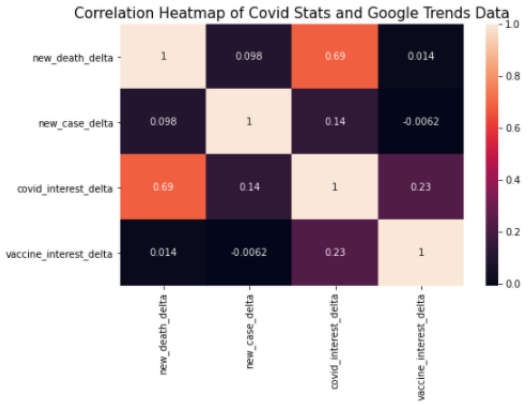

Figure 9: Correlation values of covid statistics alongside how frequently "Covid" and "Vaccine" were searched on Google

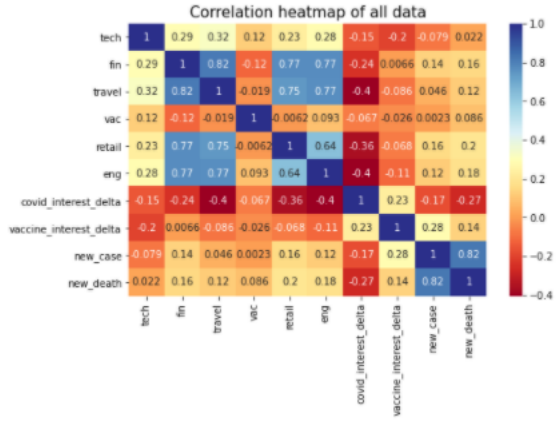

Figure 10: Correlation values of all the research data

Google Trend data was more strongly correlated with the stock market.

4.1.4 Stationarity Check

Before moving on to the Granger causality test, we needed to determine if the set of time series among different industries that we have for analyses are stationary. For the first effort of it, we plotted each industry's lag plot to see if we can capture any non-stationary characteristics from the plot.

Even after observing Figure 11, it was still uncertain whether any particular industry (or industries) was non-stationary or not. Tech and energy industries' lag plots seemed to be very sparsely populated, finance industry lag plot seems to be aggregated towards (1.5, 0) in $(y(t), y(t+1))$ field, and the rest seemed to be behaving very similarly towards (0,0). Since we were not able to clearly determine stationarity of each time series, we decided to approach with a different method.

In order to statistically determine stationarity of each industry's difference in movement, we used the Augmented Dickey Fuller Test (ADF test). ADF test's hypothesis are the following:

$$H_0 : \phi_1 = 1$$

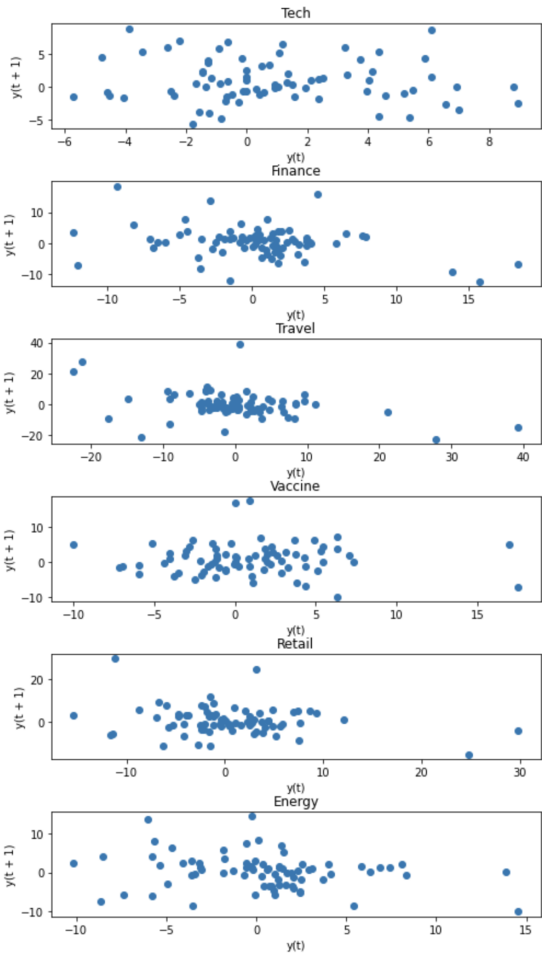

Figure 11: Lag plot of industries before differencing

$$H_A : \phi_1 < 1$$

where

$$Y_t = \mu + \Sigma_{i=1}^{P} \phi_i Y_{t-i} + \varepsilon_t$$

and

$$\Delta Y_t = \mu + \delta Y_{t-1} + \Sigma_{i=1}^{P} \beta_i \Delta Y_{t-i} + \varepsilon_t$$

Here, the null hypothesis is signifying that if ϕ_1 equals to 1 (in other words, if ϕ_1 is a unit root), then the process is non-stationary. Whereas, the alternative hypothesis says that if ϕ_1 is less than 1 (in other words, if ϕ_1 is not a unit root), then we can reject the null hypothesis and conclude that the process is stationary.

In order to test the statistical significance of the stationarity, it measures the t-statistic of the $\hat{\delta}$, which is

$$t_{\hat{\delta}} = \frac{\hat{\delta}}{se(\hat{\delta})}$$

and compare the above t-statistic with the Dickey Fuller distribution to test the statistical significance.

- $t_{\hat{\delta}} < DF_{critical}$: Reject H_0
- $t_{\hat{\delta}} > DF_{critical}$: Fail to reject H_0

The following figures of tables help us understand the result of the ADF test. From Figure 12, we can observe that all industries rejected the null hypothesis except for the finance industry. Since tech, travel, vaccine, retail, and energy rejected the null hypothesis, we could conclude that they all had stationary processes. Finance industry, however, failed to reject the null hypothesis with a p-value of 0.141897, so in order to provide an environment for us to conduct the Granger causality test, we needed to transform the finance time series into a stationary process.

	tech	fin	travel	vac	retail	eng
ADF Statistics	-9.616886	-2.399431	-12.938402	-8.801258	-5.602897	-10.707106
p-value	0.000000	0.141897	0.000000	0.000000	0.000001	0.000000

Figure 12: Augmented Dickey Fuller Test before differencing

Transforming a non-stationary process into a stationary process can be done in multiple different ways, but we decided to start with the simplest approach, simple differencing. Fortunately, our basis data was already differenced when we aggregated the stock movement data. Since it would not have provided comparable data if the stock prices were simply summed up (due to different pricing for each stock in the market), we aggregated the percentage difference of each stock within the industry for the week. This provided, in a way, first difference effect on the series, which we assume was the reason for other industries' stationarity. Inspired from this initial data preprocessing effort, we applied the same way of differencing technique to the finance series by simply taking the first order difference to the finance time series.

	tech	fin	travel	vac	retail	eng
ADF Statistics	-9.500838	-6.281635	-12.852146	-8.819093	-5.554990	-10.641742
p-value	0.000000	0.000000	0.000000	0.000000	0.000002	0.000000

Figure 13: Augmented Dickey Fuller Test after differencing

From Figure 13, we can observe that the differencing effort to transform the finance time series from non-stationary process to stationary process had worked perfectly. Figure 13 shows that for all industries, we were able to reject the null hypothesis and conclude that our data was now stationary.

Lag plots in Figure 14 shows the effect of differencing to the finance time series process. As we mentioned earlier, finance industry tended to have a slight inclination of its mean around 1.5 before differencing; however, after the differencing, we can observe that its data points are now more centered around 0 with smaller variance except for a few outliers.

4.2. Individual time series

Since we were planning on building a model that is capable of forecasting the future with the given

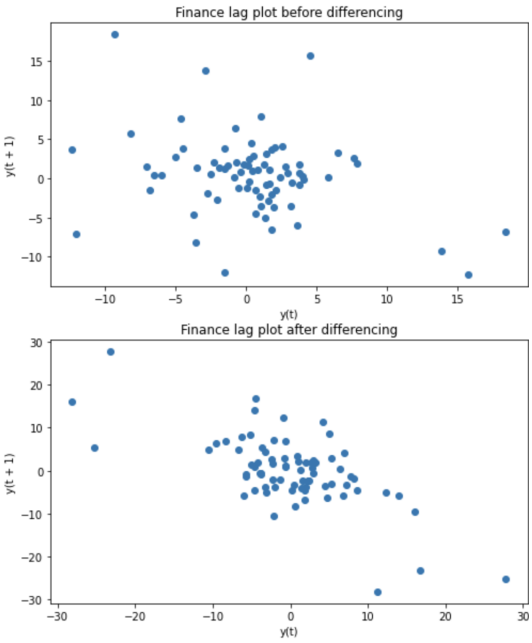

Figure 14: Finance lag plot before and after differencing

multiple time series, we decided to check whether each time series can be expressed with individual AR processes. Since the vectorized autoregressive model bases itself on building a multivariate AR model with the given set of time series, it was natural to check whether each time series were AR processes. Initially, we checked the autocorrelation function plot (ACF plots).

From Figure 15, we cannot gain much information since most of the time series seem to fall under the 95% confidence interval instantly after Lag 0, except for finance and travel data. However, since it demonstrates a pattern of decay as lag increases, we could assume that we were dealing with AR processes. To be more confident about the possibility of individual AR processes and determine the correct lag for each industry's time series, we used Pearson Correlation Coefficient tests for each industry.

Hypothesis test for Pearson Correlation Coefficients of the samples have the following hypotheses:

- H_0 :True Pearson Correlation Coefficient $(\rho = 0)$ based on the value of the sample coefficient r_{xy}
- H_A :True Pearson Correlation Coefficient $(\rho \neq 0)$ based on the value of the sample coefficient r_{xy}

where Pearson Correlation Coefficient of the samples are calculated with the following equation:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (5)$$

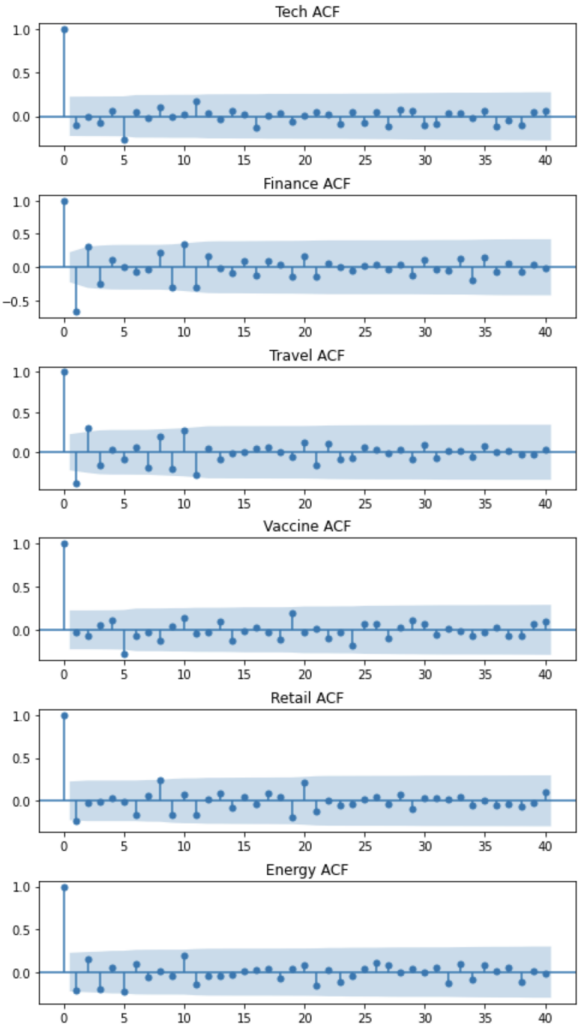

Figure 15: ACF plots for all industries

Figure 16 shows the test results for each industry. From the test result, we could observe that possible AR processes for each industry include:

- Tech: AR(5)
- Finance: AR(1), AR(2), AR(3), AR(9), AR(10), AR(11)
- Travel: AR(1), AR(2), AR(9), AR(10), AR(11)
- Vaccine: AR(5)
- Retail: AR(1), AR(8)
- Energy: AR(5), AR(10)

with p-values smaller than 0.05.

4.3. Granger Causality Test

Following the individual time series analysis, we checked interdependent predictability with the Granger causality test. Even though this test does not directly indicate the causal relationship between each time series, we assumed that verifying Granger causality among industries would help

	Industry	Lag	corr coef	p-value
0	tech	5	-0.275705	0.019955
1	fin	1	-0.669624	0.000000
2	fin	2	0.307969	0.007600
3	fin	3	-0.245140	0.036588
4	fin	9	-0.333964	0.005746
5	fin	10	0.409394	0.000642
6	fin	11	-0.397483	0.001043
7	travel	1	-0.387286	0.000598
8	travel	2	0.299681	0.009487
9	travel	9	-0.246632	0.044223
10	travel	10	0.347605	0.004240
11	travel	11	-0.388915	0.001366
12	vac	5	-0.293995	0.012827
13	retail	1	-0.227543	0.049609
14	retail	8	0.263451	0.029954
15	eng	5	-0.243664	0.040591
16	eng	10	0.243925	0.048414

Figure 16: Pearson Correlation Coefficients

show a certain level of similarity between two industries for all industries. Furthermore, "Granger-causal" relationship among industries would set grounds for having a robust VAR model.

	tech_x	fin_x	travel_x	vac_x	retail_x	eng_x
tech_y	1.000000	0.000600	0.066500	0.039800	0.000000	0.009500
fin_y	0.000300	1.000000	0.000000	0.006100	0.000000	0.000100
travel_y	0.000100	0.000000	1.000000	0.000000	0.000000	0.000100
vac_y	0.000000	0.001200	0.000000	1.000000	0.000000	0.000000
retail_y	0.000800	0.001800	0.000000	0.000200	1.000000	0.000000
eng_y	0.000200	0.008500	0.000400	0.000000	0.014000	1.000000

Figure 17: Granger Causality Test p-value matrix

From Figure 17, it clearly shows that most industries have "Granger-causal" relationship among one another except for $tech_y$ and $travel_x$. For pairs that show p-value smaller than 0.05 in the p-value matrix in Figure 17, we can safely reject the null hypothesis of the Granger causality test that the coefficients of past values of other time series in the regression equation is 0, which shows "Granger-causal" relationships among one another. For $tech_y$ and $travel_x$ case, we can interpret that travel stock movement does not "Granger-cause" tech stock movements. In other words, travel stock movement did not have enough predictive power over predicting tech stock movements.

5. Interdependent forecast

For interdependent forecasting, we employed the VAR model. In order to be robust for the model design, we conducted the Cointegration test first.

5.1. Cointegration Test

Cointegration test helps to establish the presence of a statistically significant connection between 2 or more time series. When there are 2 or more time series, where a linear combination of them has an order of integration (d) less than that of the individual series, the collection of series is said to be cointegrated. [7]

	tech	fin	travel	vac	retail	eng
Test Stat	142.09	91.81	55.97	36.13	19.25	6.85
Confidence Interval (95%)	83.9383	60.0627	40.1749	24.2761	12.3212	4.1296
Significance (Test stat > CI)	True	True	True	True	True	True

Figure 18: Cointegration test result

Cointegration test results for all industries in Figure 18 shows that all industries are cointegrated. This sets a robust basis for the VAR model that we constructed.

5.2. Vectorized Autoregression model

Before diving into building the VAR model, it was necessary to find the correct lag value. Since VAR model's credibility and accuracy depend heavily on the right lag value, it was crucial for us to find the correct lag for the model.

5.2.1 Lag for VAR

	AIC	BIC	FPE	HQIC
Lag_1	17.209216	18.507009	29,873,591.098409	17.727411
Lag_2	17.307447	19.736056	33,583,452.286344	18.276249
Lag_3	17.360579	20.937461	37,285,645.530761	18.786029
Lag_4	17.048821	21.791875	30,387,278.864772	18.937045
Lag_5	16.802900	22.730484	28,919,404.858705	19.160112
Lag_6	16.975749	24.106691	48,197,595.484511	19.808246
Lag_7	16.258555	24.612171	41,562,488.248065	19.572714
Lag_8	14.762742	24.358849	24,694,555.886465	18.565014
Lag_9	12.203914	23.062847	11,615,532.332519	16.500821
Lag_10	-33.838557	-21.695926	0.000000	-29.040426
Lag_11	-362.026466	-348.578717	0.000000	-356.720467
Lag_12	-369.179990	-354.405134	0.000000	-363.359428
Lag_13	-370.311430	-354.186892	0.000000	-363.969571
Lag_14	-370.119672	-352.622276	0.000000	-363.249753
Lag_15	-369.147022	-350.252971	0.000000	-361.742266

Figure 19: AIC, BIC, FPE and HQID statistics

Figure 19 shows the table of AIC, BIC, FPE and HQIC metric for the VAR model constructed with different lags. Minimum AIC and BIC values were most importantly considered in selecting the correct lag. From the table, we could find that both AIC and BIC values dip at lag 13 and start to climb up. Therefore, we decided on building a VAR model with lag 13, VAR(13).

5.2.2 Durbin-Watson Test

After constructing and fitting the VAR(13) model, we conducted a final test: Durbin-Watson test to find out if there was any correlation among the time series in the residuals from the VAR model. If the model was fitted correctly, then the test results from the Durbin-Watson test should not display any significant correlation of those industries in the residuals. We could observe no significant

	Durbin-Watson statistic
tech	2.36
finance	2.57
travel	2.24
vaccine	1.91
retail	2.05
energy	2.35

Table 1: Caption: Durbin-Watson test result

Durbin-Watson statistic that was out of the ordinary from Table 1 (0: negative correlation, 2: no correlation, and 4: positive correlation).

5.3. Predictions

Figure 20 shows the predicted vs actual values of the model. From the figure, we could observe that although the model forecast is sometimes off or late compared to the actual movement in stocks, it shows predictions that are close to its actual values. VAR prediction result for vaccine industry best mimicked the actual movement overall.

6. Further Studies

Especially now as the Delta Covid Variant has been in the news, it would be interesting to track stock performance against the Google Trends Data for "Delta Variant". In the context of our current research, this could be considered as "Phase 3". We might see that the two industries impacted most in Phase 1 (Retail and Travel) might not be as impacted in this Phase 3, as people might have gotten used to Covid being around, and won't take the Delta Variant as seriously.

In addition, further studies on various approaches for forecasting model can be done. Ranging from employing traditional machine learning model by columnizing the time series to advanced neural network combination approach would be interesting to study for. Including extra environmental variable and additional feature engineering effort can be studied to further improve the forecasting model.

Figure 20: Prediction vs actual

7. Conclusion

With the exception of Tech, all of the industries we studied seemed to have substantially more volatility in the first phase of Covid than the second phase of Covid. This is especially evident between tech and travel. In the first phase of Covid, travel was much more volatile than tech. However, they appeared to be similarly volatile in the second phase. Although their volatility seemed to be very loosely correlated, having a correlation coefficient of only 0.21.

On the other hand, retail and travel seemed to correlate strongly with each other over the course of the pandemic, having a correlation coefficient of 0.74. This does make sense, as the travel industry suffered for the same reasons as the retail industry – where people in lockdown weren't going out.

One of the most interesting findings from this project was that over the course of the pandemic, the strongest correlation was found between finance and travel. Initially we surmised that the strongest correlation would be between travel and retail, given how similar those industries are.

We found that Google Trends data had a stronger correlation with the industry performance than the case/death count data from the CDC. For example, the Google Trend data was a much stronger indicator of the retail industry than the CDC death count data. This stands to reason, as people are more influenced by what they see online, than by actual numbers that they don't see.

Although the main purpose of this project was to look at how the Google Trends and CDC data correlated with the stock market, there were interesting findings when juxtaposing the Google Trends and CDC data. The death counts were a stronger predictor of people searching for "Covid" than the case counts. This does make sense, as death counts in the news are a lot more sensational than case counts, leading to more people looking up "Covid" on Google.

Granger-causality test showed each industry's stock movement's interdependency. With the findings from the Granger causality test and the results from the cointegration test, we were able to build a VAR model to mimic the actual stock movements of various industries. Even though the result was not 100% accurate in prediction, we believe that with further studies in feature engineering and selecting the right model for prediction, prediction accuracy can be improved in the future.

References

[1] VRC. Covid-19: Event timeline.

[2] Felix Richter. Stocks emerge from covid crash with historic 12-month run.

[3] The American Journal of Medical Care. A timeline of covid-19 developments in 2020.

[4] Jonathan D. Cryer and Kung-sik Chan. *Time series analysis: with applications in R*, pages 16–17. Springer, 2011.

[5] Jonathan D. Cryer and Kung-sik Chan. *Time series analysis: with applications in R*, page 66. Springer, 2011.

[6] C.W.J. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[7] Søren Johansen. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59:1551–1580, 1991.