Introduction

Financial markets are extremely difficult to predict, and various sectors of the financial markets are affected in different ways and by different things. The best way to look at the performance of the various market sectors is to look at sector Exchange Traded Funds (ETF) to determine how the sectors of the market move. In order to see how a market sector moves, we can look at various outside factors and economic indicators. These factors influence market sectors as well as help us predict how a market sector will perform in the future. A good economic environment is important. When the economic environment is good, so is company profitability. This is important to many companies whose shares trade on the stock market. Additionally, when the economy is expanding and doing well, this usually means that more people are investing and buying goods and services. When the economy is not doing well, this usually means that people are not investing or buying goods and services. This can cause a market sector to decrease in value, but, as we have seen in certain black swan events like the current pandemic, it may drive some market sectors through the roof. In this paper, we will look at five factors (Variable X_1: Interest Rates, Variable X_2: Inflation, Variable X_3: Gross Domestic Product (GDP), Variable X_4: National Unemployment Rates, and Variable X_5: Total Daily Number US COVID cases) to determine which factors influence the ETF value the most and which are the best predictor variables that can be used by financial institutions to predict ETF value.

Research

Our research has very much pointed to the assumption that COVID-19 has had a negative impact on the global economy and financial markets. This is due to significant reductions in income, a rise in unemployment, and disruptions in the transportation, service, and manufacturing industries. This is a result of the disease mitigation measures that have been implemented around the world. Proactive international actions are required to save lives and protect economic prosperity.

In our analysis, we will take a look at the data of major markets and try see if COVID-19 affected the global economy and financial markets as stated in our research.

Research Questions

1. Which variables are significant factors in determining stock market value?
2. Has the total daily number of US COVID cases affected the stock market value since March 2020 until March 2021?

Initial Model Variables:

1. Variable Y: Stock Market Value = Dependent Variable
2. Variable X_1: Interest Rates = Independent Variable
3. Variable X_2: Inflation= Independent Variable
4. Variable X_3: Gross Domestic Product (GDP) = Independent Variable
5. Variable X_4: National Unemployment Rates = Independent Variable
6. Variable X_5: Total Daily Number US COVID Cases = Independent Variable

Hypothesis

H_o: COVID-19 did affect the global markets.
H_1: COVID-19 did not affect the global markets.

Exploratory Data Analysis

When we conduct our initial exploratory data analysis on the data and we take a look at a linear model where the NASDAQ closing = X1 (Interest Rates) + X2 (Inflation Rates) + X3 (GDP Rate) + X4 (Unemployment Rates) + X5 (Number of New COVID Cases) in the Data Summary 1: NASDAQ Closing Model. What we see is that the p-value for all of the variables in the equation are all less than 0.05. This means that they are statistically significant in this relationship. However, when we look at the R2 value, it is 0.7765 which means that the model is not as strong as we would like it to be. This is a clue that we will need to conduct further tests on this data set to fully determine how to make the model stronger.

Data Summary 1: NASDAQ Closing Model

```
summary(marketLM)

##
## Call:
## lm(formula = nasdaq$Close ~ interestRates$Rate + inflationRate$Rate +
##     GDP$`GDP Rate` + unemploymentRate$Rate + COVID$new_cases)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3147.16  -447.47   -64.91   598.31  2178.98
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.589e+04  3.960e+02  40.137  < 2e-16 ***
## interestRates$Rate    -3.234e+03  3.548e+02  -9.117  < 2e-16 ***
## inflationRate$Rate    -8.145e+02  2.201e+02  -3.701 0.000255 ***
## GDP$`GDP Rate`        -6.731e+02  6.444e+01 -10.446  < 2e-16 ***
## unemploymentRate$Rate -6.928e+02  4.201e+01 -16.492  < 2e-16 ***
## COVID$new_cases        5.327e-03  1.055e-03   5.051  7.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 867 on 305 degrees of freedom
## Multiple R-squared:  0.7765, Adjusted R-squared:  0.7728
## F-statistic: 211.9 on 5 and 305 DF,  p-value: < 2.2e-16
```

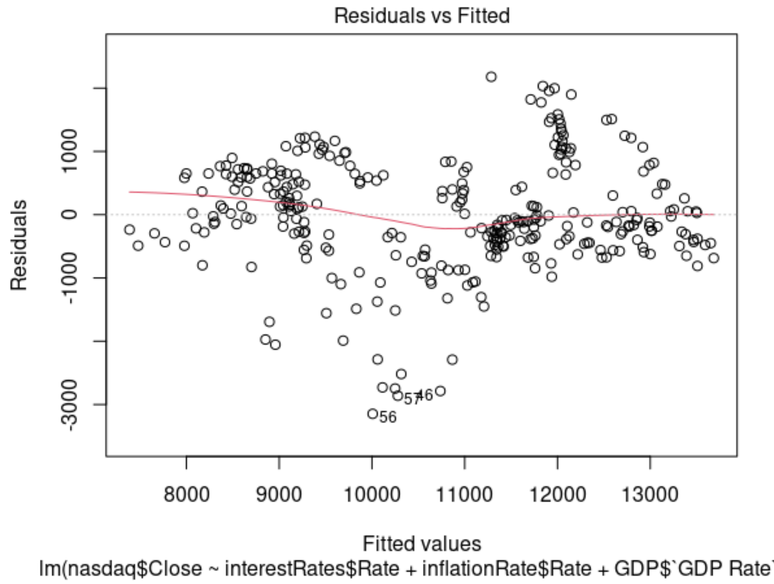

Figure 1: NASDAQ Residuals vs. Fitted

Here in Figure 1: NASDAQ Residuals vs. Fitted Values we see that linearity seems to hold reasonably well, as the red line is close to the dashed line. the fitted vs residuals plot, which allows us to detect several types of violations in the linear regression assumptions. We can also note the heteroskedasticity: as we move to the right on the x-axis, the spread of the residuals seems to be increasing. Finally, points 46, 56, and 57 may be outliers, with large residual values.

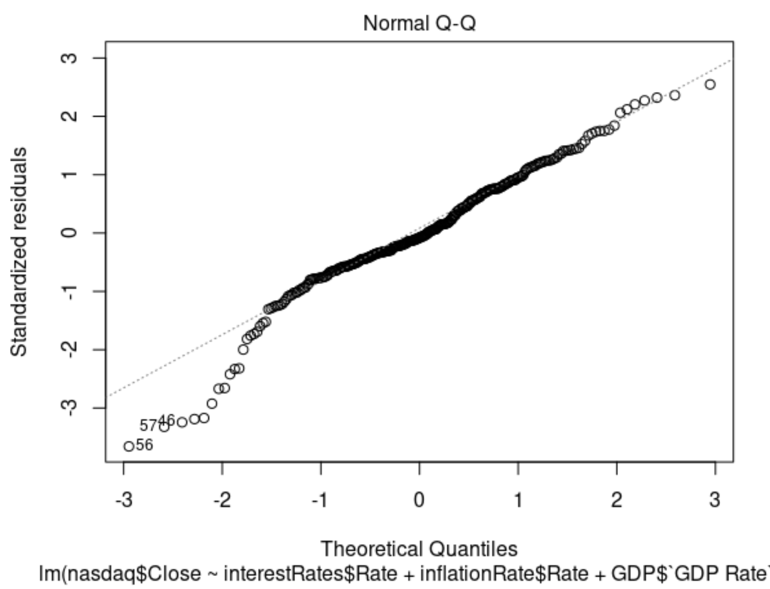

Figure 2: NASDAQ Normal Q-Q Plot

The Normal Q-Q plot shows if residuals are normally distributed. It's good if residuals are lined well on the straight dashed line. The Normal Q-Q Plot for Figure 2: NASDAQ Normal Q-Q shows many observations deviating from the dashed line. Additionally, the observations numbered as 46, 56, and 57 look a little off. These observations may be a potential problem. Looking at the QQ Plot and Standardized Residual Plot, we can identify some elements of weakness such as outliers such as 46, 56, and 57. Points are labeled as outliers in small to moderate size data sets if the standardized residual for the point falls outside of the interval of -2 to 2. Identification and examination of any outliers is a key part of regression analysis. This is important because an outlier can also be classified as a bad leverage point. A bad leverage point is a leverage point where standardized residuals falls outside the interval from -2 to 2. More from the standardized residual plot. There is a clear random pattern evident in this plot. Therefore, model1 may not be the best model and I will need to find other models to compare it to.

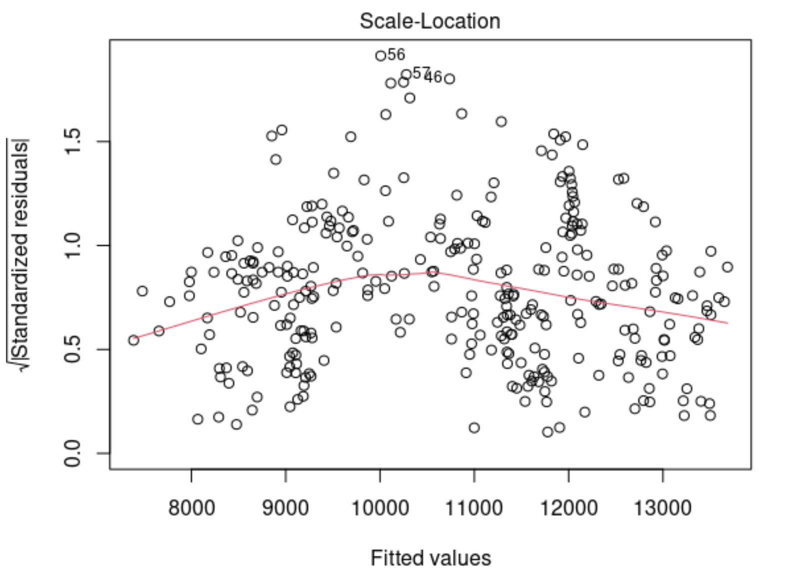

Figure 3: NASDAQ Scale- Location Plot

Scale-Location plot for Figure 3: NASDAQ Scale-Location Plot shows whether residuals are spread equally along the ranges of input variables (predictor). The assumption of equal variance (homoscedasticity) could also be checked with this plot. If we see a horizontal line with randomly spread points, it means that the model is good. However, in the Scale-Location Plot for Figure 3: NASDAQ Scale-Location Plot, there seems to be a pattern that is not random. Hence, the variance is not equal. This means that we need to conduct a transformation process to fix the issue with Figure 3: NASDAQ Scale-Location Plot.

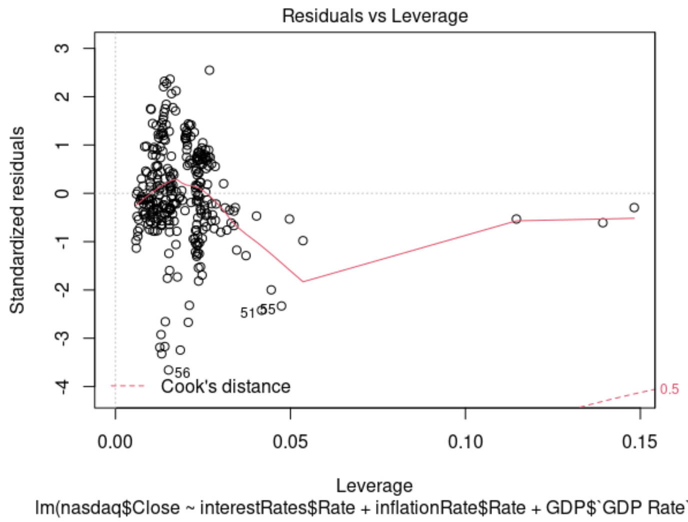

Figure 4: NASDAQ Residuals vs. Leverage

In the Figure 4: NASDAQ Residual vs. Leverage Plot for the NASDAQ Model we are looking at how the spread of standardized residuals changes as the leverage, or sensitivity of the fitted y-hat to a change in y, increases. Firstly, this can also be used to detect heteroskedasticity and non-linearity. The spread of standardized residuals shouldn't change as a function of leverage. Here the spread appears to decrease, indicating heteroskedasticity. Second, points with high leverage may be influential. Deleting them would change the model a lot. For this we can look at Cook's distance, which measures the effect of deleting a point on the combined parameter vector. Cook's distance is the dotted red line here, and points outside the dotted line have high influence. In this case there are no points outside the dotted line. Therefore, we do not see any points of high influence. However, we also see points 51, 55, and 56 that are identified as possible outliers.

When the assumption of data normally distributed is violated or the relationship between the dependent and independent variables in case of linear model are not linear, such as the NASDAQ Model, in such situations some transformations methods that may help the data set follow a normal distribution. Box Cox is one such transformation method. In the next part of this research and analysis, we will show two different models in order to best compare which model is best of the three models.

Data Analysis Plan

1. Gather data from relevant ETFs, the NASDAQ, interest rates, inflation rates, GDP, unemployment rate, and COVID.
2. Clean the data in excel to make it useful for our analysis
3. Put the data into R with read excel package to prepare for analysis
4. Filter all data to include only relevant dates where COVID was a present threat
5. Develop preliminary linear models
6. Use ANOVA to check for significant variables in each model
7. Update the models to include on the significant variables

8. Check the summaries of the models to ensure they are all accurate models (high adjusted R-squared)
9. Make relevant transformations to increase model validity
10. Plot the models to check for leverage points and outliers and determine what to do with these points
11. Added variable plots which (avp) which will show us which variables are significant if lines are non-flat.
12. Check for multicollinearity using vif unction.

Models We Plan to Use

Yes, we have used a transformed variable using BoxCox transformation in which if lambda is zero then we take the log of the variable and if lambda is not equal to zero then we transform using formula lambda 1v / lambda. We can also transform variables using the power transform function.

Model Development: Exploratory Analysis and Model Comparison

We have used the standard diagnostic model using plot function which produces 4 graphs. We also plan to include added variable plots (avp) which show us which variables are significant. We also want to check for multicollinearity using vif function.

Model 1:

NASDAQ Closing Value= X_1 (Interest Rates) + X_2 (Inflation Rates) + X_3 (GDP Rate) + X_4 (Unemployment Rates) + X_5 (Number of New COVID Cases)

Data Summary 1: NASDAQ Closing Model

```
summary(marketLM)

##
## Call:
## lm(formula = nasdaq$Close ~ interestRates$Rate + inflationRate$Rate +
##     GDP$`GDP Rate` + unemploymentRate$Rate + COVID$new_cases)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3147.16  -447.47   -64.91   598.31  2178.98
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.589e+04  3.960e+02  40.137  < 2e-16 ***
## interestRates$Rate    -3.234e+03  3.548e+02  -9.117  < 2e-16 ***
## inflationRate$Rate    -8.145e+02  2.201e+02  -3.701 0.000255 ***
## GDP$`GDP Rate`        -6.731e+02  6.444e+01 -10.446  < 2e-16 ***
## unemploymentRate$Rate -6.928e+02  4.201e+01 -16.492  < 2e-16 ***
## COVID$new_cases        5.327e-03  1.055e-03   5.051  7.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 867 on 305 degrees of freedom
## Multiple R-squared:  0.7765, Adjusted R-squared:  0.7728
## F-statistic: 211.9 on 5 and 305 DF,  p-value: < 2.2e-16
```

Figure 5: Model 1 Summary Output

NASDAQ Closing Model. What we see is that the p-value for all of the variables in the equation are all less than 0.05. This means that they are statistically significant in this relationship. However, when we look at the R2 value, it is 0.7765 which means that the model is not as strong as we would like it to be. This is a clue that we will need to conduct further tests on this data set to fully determine how to make the model stronger.

Becaue the variable output from the information above is showing that all variables are significantly significant, we need to take a closer look at the variables. To check for multicollinearity we will use vif function.

```
vif(marketLM1)
 interestRates$Rate   inflationRate$Rate      GDP$`GDP Rate`
          3.890649             5.059490             7.554023
unemploymentRate$Rate      COVID$new_cases
          7.543277             1.973101
```

As a rule of thump cut-off for removing variable is 5 hence drop GDP Rate because it is the largest variable value of all the variables. From this information we created a second Model (Model 2) and see if taking GDP out of the model makes the model better.

Model 2: NASDAQ Closing Value= X_1 (Interest Rates) + X_2 (Inflation Rates) + X_4 (Unemployment Rates) + X_5 (Number of New COVID Cases)

However, we are still not sure if Model 1 is better than Model 2. So we compare the two models using ANOVA. What this output tells us is Model 1 is better than Model 2.

```
Analysis of Variance Table

Model 1: nasdaq$Close ~ interestRates$Rate + inflationRate$Rate + unemploymentRate$Rate +
    COVID$new_cases
Model 2: nasdaq$Close ~ interestRates$Rate + inflationRate$Rate + GDP$`GDP Rate` +
    unemploymentRate$Rate + COVID$new_cases
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1    306 311279797
2    305 229254625  1  82025172 109.13 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #Plotting models
```

Figure 6; ANOVA Comparison of Model 1 and Model 2

However, we still have an issue with X_3 (GDP Rate). Taking X_3 (GDP Rate) made the model worse not better. This tells us that we still need to modify X_3 (GDP Rate). So, then we then take the log of X_3 (GDP Rate) and create Model 3, below, from this information.

Model 3:

NASDAQ Closing Value= X_1 (Interest Rates) + X_2 (Inflation Rates) + LOG(X_3 (GDP Rate)) + X_4 (Unemployment Rates) + X_5 (Number of New COVID Cases)

After developing Model 3, the next step is to compare Models 1, 2, and 3 with one another. To do this we will use linear regression output, linear regression plots, and mmp plots.

Model 1

```
Call:
lm(formula = nasdaq$Close ~ interestRates$Rate + inflationRate$Rate +
    GDP$`GDP Rate` + unemploymentRate$Rate + COVID$new_cases)

Residuals:
     Min       1Q   Median       3Q      Max
-3147.16  -447.47   -64.91   598.31  2178.98

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.589e+04  3.960e+02  40.137  < 2e-16 ***
interestRates$Rate    -3.234e+03  3.548e+02  -9.117  < 2e-16 ***
inflationRate$Rate    -8.145e+02  2.201e+02  -3.701 0.000255 ***
GDP$`GDP Rate`        -6.731e+02  6.444e+01 -10.446  < 2e-16 ***
unemploymentRate$Rate -6.928e+02  4.201e+01 -16.492  < 2e-16 ***
COVID$new_cases        5.327e-03  1.055e-03   5.051  7.6e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 867 on 305 degrees of freedom
Multiple R-squared:  0.7765,    Adjusted R-squared:  0.7728
F-statistic: 211.9 on 5 and 305 DF,  p-value: < 2.2e-16
```

Model 2

```
Call:
lm(formula = nasdaq$Close ~ interestRates$Rate + inflationRate$Rate +
    unemploymentRate$Rate + COVID$new_cases)

Residuals:
    Min      1Q  Median      3Q     Max
-3665.5  -411.2   -65.3   518.4  2381.0

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.297e+04  3.256e+02  39.827  < 2e-16 ***
interestRates$Rate    -1.971e+03  3.880e+02  -5.080 6.59e-07 ***
inflationRate$Rate    -1.544e+03  2.428e+02  -6.356 7.49e-10 ***
unemploymentRate$Rate -3.236e+02  2.641e+01 -12.251  < 2e-16 ***
COVID$new_cases        7.292e-03  1.207e-03   6.040 4.46e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1009 on 306 degrees of freedom
Multiple R-squared:  0.6965,    Adjusted R-squared:  0.6925
F-statistic: 175.5 on 4 and 306 DF,  p-value: < 2.2e-16
```

Figure 7: Linear Regression Output for Model 1 &2

Model 3

```
Call:
lm(formula = nasdaq$Close ~ interestRates$Rate + inflationRate$Rate +
    log(GDP$`GDP Rate`) + unemploymentRate$Rate + COVID$new_cases)

Residuals:
    Min      1Q  Median      3Q     Max
-1954.5  -321.5   117.2   442.4  1144.6

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.334e+04  2.901e+02  45.987  < 2e-16 ***
interestRates$Rate    -1.658e+03  2.886e+02  -5.746 2.73e-08 ***
inflationRate$Rate     5.310e+02  1.744e+02   3.044  0.00259 **
log(GDP$`GDP Rate`)   -2.562e+03  1.071e+02 -23.909  < 2e-16 ***
unemploymentRate$Rate -4.087e+02  3.785e+01 -10.796  < 2e-16 ***
COVID$new_cases        6.453e-03  7.637e-04   8.449 2.75e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 609.4 on 242 degrees of freedom
  (63 observations deleted due to missingness)
Multiple R-squared:  0.8826,    Adjusted R-squared:  0.8801
F-statistic: 363.7 on 5 and 242 DF,  p-value: < 2.2e-16
```

Figure 8: Linear Regression Output for Model 3

From the Linear Regression output in Figure 7 and Figure 8, we can see that Model 1 has an R-Sqared value of 0.7765 in comparison with Model 2 R-Sqared value of 0.6965. Taking a variable away in Model 2 shows that taking away variables does not make the model better. This lead us to then take the LOG of GDP. This was Model 3 which ended up being a very good model with an R-Squared Value of 0.8826.

Model 1

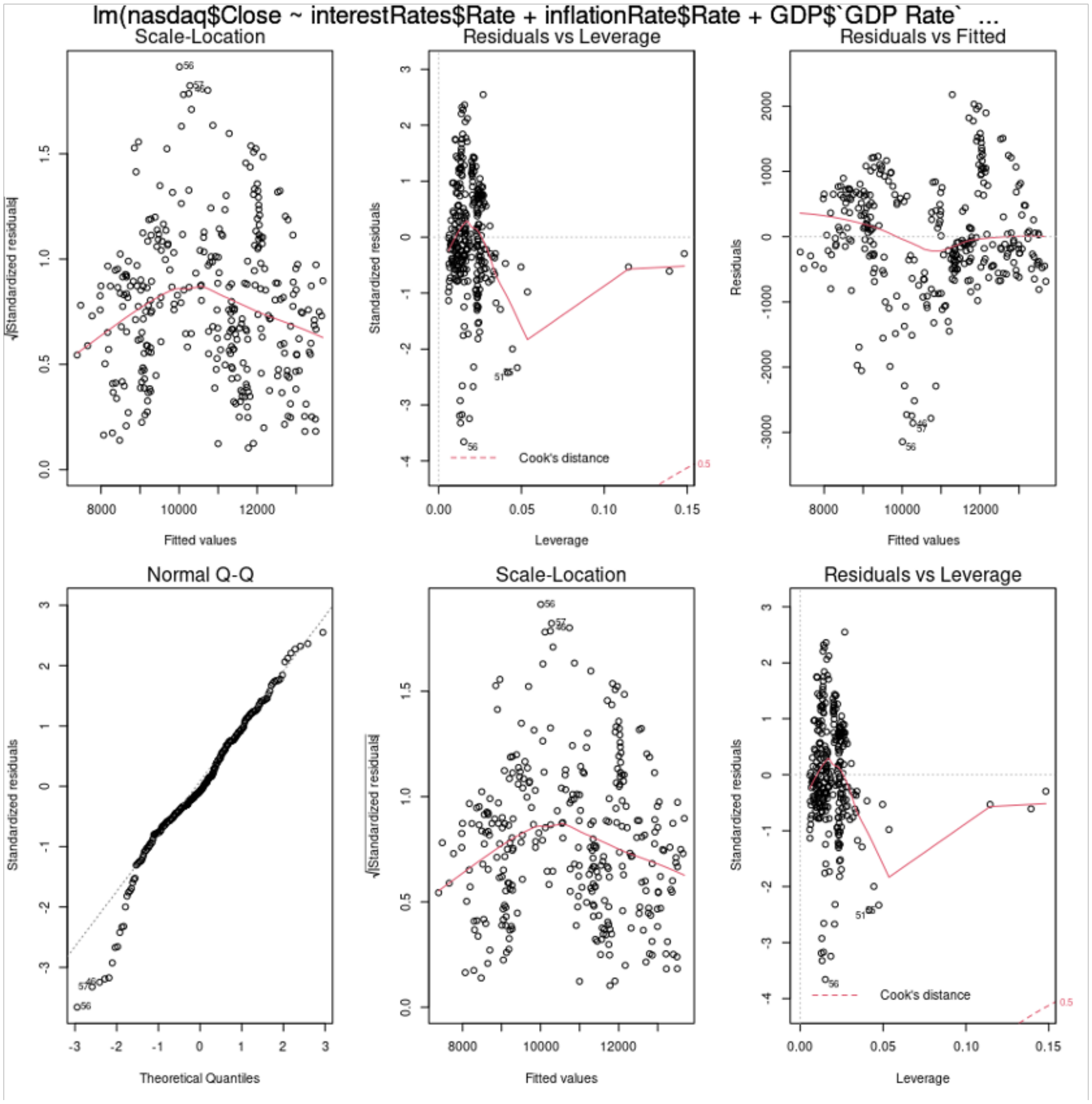

Figure 9: Linear Regression Plots for Model 1

Model 2

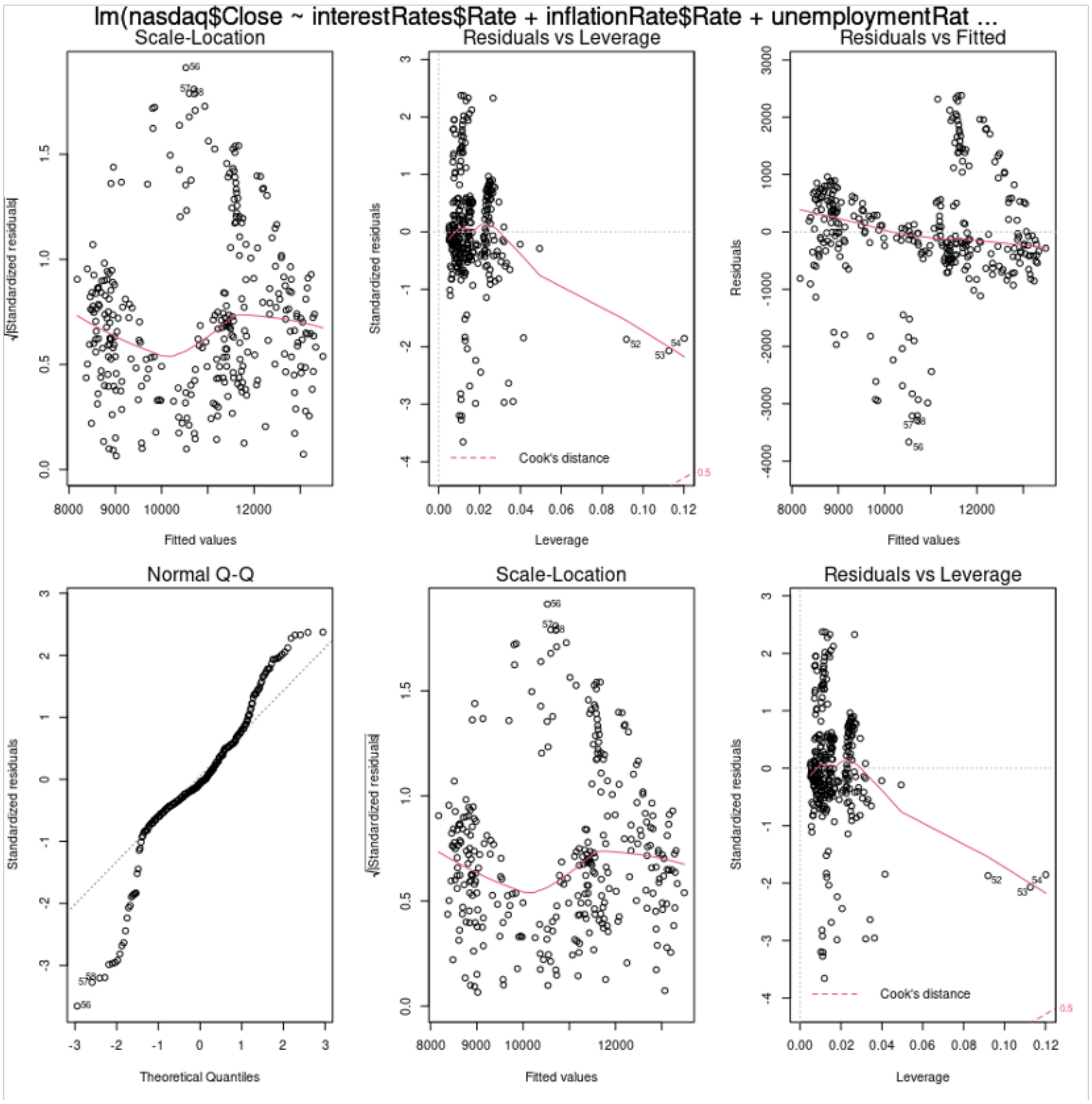

Figure 10: Linear Regression Plots for Model 2

Model 3

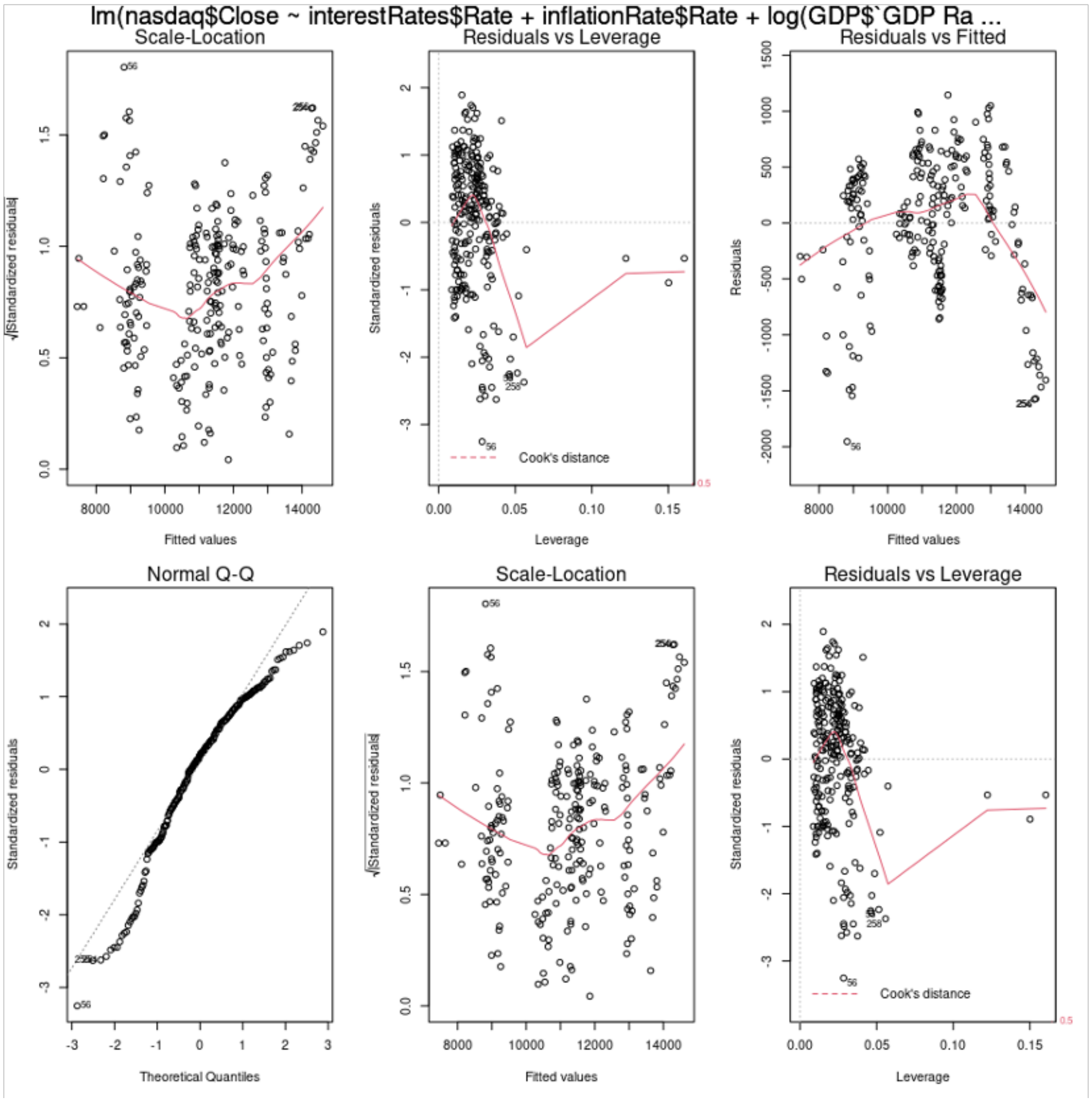

Figure 11: Linear Regression Plots for Model 3

Model 1

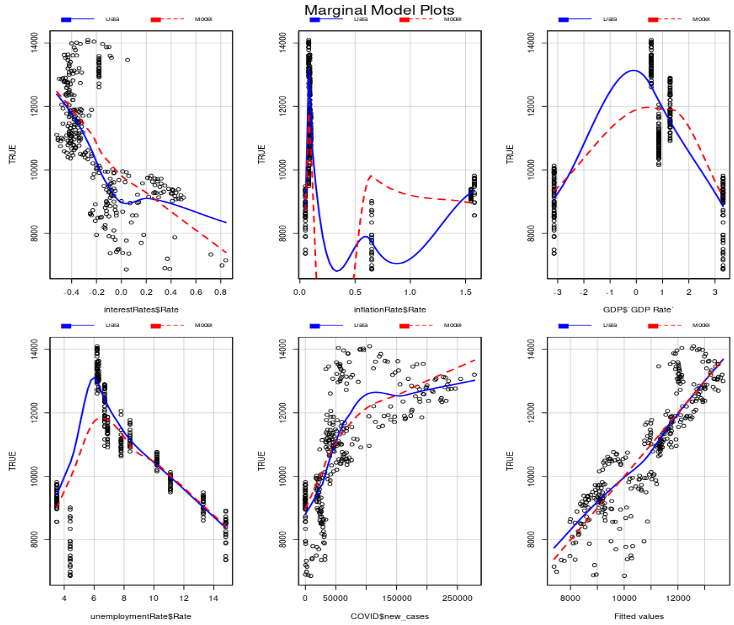

Figure 12: Marginal Model Plot (mmp) for Model 1

Model 2

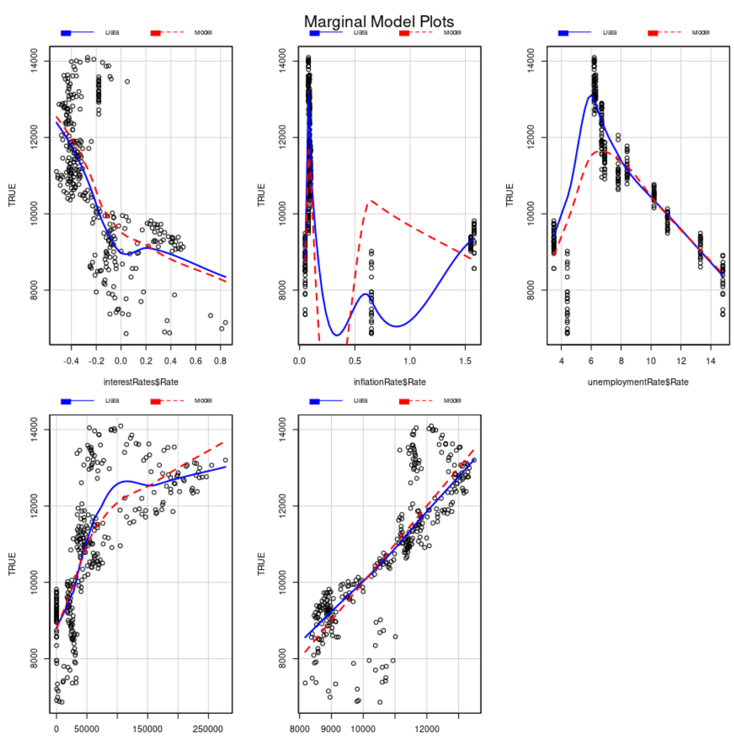

Figure 13: Marginal Model Plot (mmp) for Model 2

Model 3

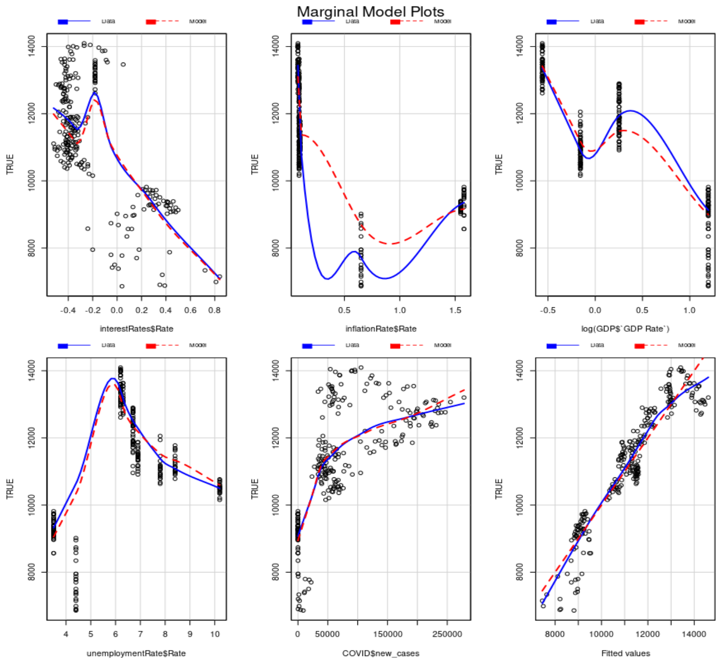

Figure 14: Marginal Model Plot (mmp) for Model 3

To check for model validity for each model, we used Marginal Model Plot (mmp). When assessing each model using mmp, they are all very similar, where all models are approaching the data in the direction of the data which means that the data is a good fit. However, we can see that Model 3 is a better fit than both Model 1 or Model 2.

Conclusion

We can conclude that we can reject null hypothesis and say that COVID-19 and other external factor did have an effect on the financial market.

To check for model validity for each model, we used Marginal Model Plot (mmp). When assessing each model using mmp, they are all very similar, where all models are approaching the data in the direction of the data which means that the data is a good fit. However, we can see that Model 3 is a better fit.

Model 3 is the best model of the three models with an R-Squared Value of 0.8826. Model 1 has an R-Squared value of 0.7765 in comparison with Model 2 R-Squared value of 0.6965. Taking a variable away in Model 2 shows that taking away variables does not make the model better. This lead us to then take the LOG of GDP.

Discussion

Our Analysis shows that COVID-19 did have an effect on the financial Market.

More research is required to see if other markets were also affected.

Limitations

We only looked at the NASDAQ. We need to look at other sectors of other industries such as the technology, health care, industrial, energy, communications, materials, and real estate for further analysis. This will give us a better picture on the how and what markets COIVD-19 has affected these market spaces.

We looked at some major market factors, but we did not look at all factors in the market. If we look at more factors, this may help strengthen the accuracy of our models.

Future Work

We need to look at other sectors of other industries such as the technology, health care, industrial, energy, communications, materials, and real estate for further analysis. This will give us a better picture on the how and what markets COIVD-19 has affected these market spaces.

Look at additional factors or variables that may affect the market to see if this strengthens the models.

List of Data Resources

Data	Variable	Date Accessed	Site Accessed
NASDAQ Composite(^IXIC)	Y	3/29/21	https://finance.yahoo.com/quote/%5EIXIC/history?period1=1522281600&period2=1616976000&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true
Inflation Rates	X1	3/29/21	https://fred.stlouisfed.org/graph/?g=cN69
Interest Rate Data	X2	3/29/21	https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=reallongtermrateAll
GDP	X3	3/29/21	https://fred.stlouisfed.org/graph/?g=cN69
Unemployment Data	X4	3/29/21	https://fred.stlouisfed.org/series/UNRATE
COVID Case Data	X5	3/29/21	https://ourworldindata.org/coronavirus/country/united-states

List of Research Resources

L. Bauer, K. Broady, W. Edelberg, J. O'Donnel, "Ten Facts About COVID-19 and the US Economy, accessed on 14 April 2021 (https://www.brookings.edu/research/ten-facts-about-covid-19-and-the-u-s-economy/).

A Pak, O. Adegboye, A. Adekunle, K. Rahnman, E. McBryde, and D. Eisen, Economic Consequences of the COVID-19 Outbreak: the Need for Epidemic Preparedness, accessed on 14 April 2021 (https://www.frontiersin.org/articles/10.3389/fpubh.2020.00241/full).

Shared Folder

FE 541 Final Project