Analysis of Factors Affecting S&P500 and Forecasting a Company's Valuation in the Index

A Regression and Time Series Analysis Project

Amaan Vora (RUID – 216005070)
Jahnavi Shah (RUID – 218009800)
Noopur Singh (RUID – 220008221)
Ganesh Raj K (RUID – 219001760)

INDEX

INTRODUCTION

Stock market, involves a large number of investors, buyers and sellers. If the stock market rises, then a country's financial development would be high and vice-versa. The United States of America has one such financial index called the S&P500. While it is not an index which can be used to determine stock exchanges, it does track the stock market activity of the top 500 companies around the world. It is one of the most commonly followed equity index and is considered by many experts as best representation of US stock market.

In recent years, we can use huge amount of data and analyze those due to the development of computer technology. In finance, forecasting stock market is considered to be one of most difficult tasks to do till now because of the stochastic behaviors and complex dependencies of stock market. While we do have multiple means of analyzing and forecasting the various trends of the stock market to better understand its position and our investment in it, for decades we were at a loss of a conclusive methodology to perform the aforementioned task.

Stock prediction has been a phenomenon since machine learning was introduced. But very few techniques became useful for forecasting the stock market as it changes with the passage of time. Supervised algorithms can only do so much because they rely on the completeness and the variation in data to produce a more accurate result. Lest we have a proper dataset, the Regression model would fail. In recent times, there has been a new wave of forecasting that relies upon time being a comparative factor within the algorithm's prediction process.

For our project, we plan on analyzing the S&P500 dataset from 2019 to 2022 and get an understanding of the world events that shaped the current form of the dataset. We also dig deep into the nitty-gritties of the dataset within individual sectors to better understand the impact of individual sectors and companies within the index. In the end, we perform a time-series forecast on a particular company's stock to understand the future of that stock within the index.

ABSTRACT

Our primary goal with the analysis and prediction of the S&P500 Financial Index is to get a deeper understanding of the world factors that affect the Index in its current form and how these factors would affect the stocks of any company moving forward. We intend to perform a deep-dive analysis into how this index functions and then proceed to comprehend the various cogs that propel the market forward. For this, it would be difficult for us to analyze a cumulative dataset and make our inferences on the same as it would not represent the whole picture in a decisive manner. Therefore, we plan to formulate individual datasets for each of the 11 sectors in the S&P500 Index – Communication

Services, Consumer Discretionary, Consumer Staples, Energy, Financials, Healthcare, Industrials, Information Technology, Materials, Real Estate and Utilities. Post this, we conclude through our analysis of the individual sectors how there are certain companies that affect the market more, and are affected more by world events. We also take one particular sector and perform a separate analysis on the companies within that sector to get an even further understanding of the individual company's role in shaping up the Index. Finally, we conclude our project by drawing inferences from these analyses and performing a Time Series Forecasting on a company's data to understand how it will be affected in the future.

PROBLEM FORMULATION

Our problem statement stems fundamentally from the current nature of the S&P500 Index. Given multiple world events that shook the fabrics of financial economics, it stands to reason that many experts are at odds given not just the present conditions of the Index, but also its projection in the near future. This is an Index that decides not just the fate of companies, but also the US Dollar, which is listed on the Index. Therefore, it is vital for us to understand and analyze the events that shaped up the Index this way. We plan on using our understanding and apply it to a forecasting procedure via Time Series, where we would predict the future of Pfizer's portfolio within the index. Pfizer is a multinational conglomerate pharmaceutical manufacturer that supplies its pharmaceuticals across the globe, and so its position in the market is essential to determine the public perception and the trend it will follow in the future.

DATASET SOURCE

The amount of data on the internet regarding the S&P500 Index is limitless. It would be futile of us to take in the entirety of the historical data from the inception of the Index. Moreover, through an introductory understanding of the entirety of the index, we realize that there is a certain shelf life in terms of years of how much data actually affects the current state of the index. Therefore, we decide to only take the most recent data into account. Post 2019, there have been significant events that have made the top financial experts wary of the current state of financial economics. Some events are listed as follows

- The COVID Pandemic
- The United States Election
- Russia – Ukraine War
- The Crisis in Yemen
- Elon Musk's purchase of Twitter

These events have significantly impacted companies and their perception amongst eager investors around the globe. It has made the prediction of the index extremely difficult through traditional means, as it is near impossible to predict human sentiment towards a company or its portfolio. Therefore, we start by understanding the dataset.

For our dataset, we turned to Kaggle, a Data Science repository containing the most popular datasets used by Data Scientists around the world. We were able to extract a vast dataset containing the following attributes –

- **Date** - in the following format: Y-M-D
- **Open** - the price of a stock when the market opens (in USD)
- **High** - The highest price reached that day
- **Close** - The day's lowest price
- **Volume** - It denotes the number of shares traded

Date	Low	Open	Volume	High	Close	Adjusted Close
1/2/2019	40.45540619	40.91081619	26430315	41.27134705	41.0341568	35.73406219
1/3/2019	39.80075836	41.02466965	28503533	41.11954498	39.88614655	34.73432922
1/4/2019	40.09487534	40.26565552	27145348	41.12903214	40.79696274	35.52750778
1/7/2019	40.66413879	40.86337662	20995469	41.3852005	41.01517868	35.71754074
1/8/2019	40.93927765	41.32827377	19677231	41.46110153	41.20493317	35.88277817
1/9/2019	40.98671722	41.23339844	20107580	41.43263626	41.11954498	35.8084259
1/10/2019	39.87665939	41.00569153	39731162	41.01517868	40.14231491	34.95741272
1/11/2019	39.99051285	40.16128922	21064506	40.6831131	40.6831131	35.4283638
1/14/2019	40.00948715	40.37001801	17605173	40.44591904	40.19924164	35.00698471
1/15/2019	40.25616837	40.41745758	25865371	41.02466965	40.54079819	35.3044281
1/16/2019	39.87665939	40.40797043	30169380	40.5977211	39.95256042	34.79217529
1/17/2019	39.68690872	39.80075836	24177284	40.47438431	40.29411697	35.08961105
1/18/2019	40.09487534	40.66413879	38618455	40.6831131	40.3510437	35.13919449

Data Representation of the first 20 days of Pfizer Stocks

It would consume a humongous amount of time and resources to analyze each and every company's data to understand the entirety of the S&P500 Index. We therefore shift our focus to the cumulative S&P500 Index and generate a dataset for the same –

Date	Open	High	Low	Close	Adjusted Close	Volume
11/29/2022	3964.19	3976.77	3937.65	3957.63	3957.63	3546040000
11/28/2022	4005.36	4012.27	3955.77	3963.94	3963.94	3615430000
11/25/2022	4023.34	4034.02	4020.76	4026.12	4026.12	1706460000
11/23/2022	4000.3	4033.78	3998.66	4027.26	4027.26	3279720000
11/22/2022	3965.51	4005.88	3956.88	4003.58	4003.58	3887990000
11/21/2022	3956.23	3962	3933.34	3949.94	3949.94	3850690000
11/18/2022	3966.39	3979.89	3935.98	3965.34	3965.34	4037360000
11/17/2022	3919.26	3954.33	3906.54	3946.56	3946.56	4051780000
11/16/2022	3976.82	3983.09	3954.34	3958.79	3958.79	4165320000
11/15/2022	4006.41	4028.84	3953.17	3991.73	3991.73	5015310000
11/14/2022	3977.97	4008.97	3956.4	3957.25	3957.25	4561930000
11/11/2022	3963.72	4001.48	3944.82	3992.93	3992.93	5593310000
11/10/2022	3859.89	3958.33	3859.89	3956.37	3956.37	5781260000

Data Representation of the last 20 days of the S&P500 Index

However, we do understand that within this dataset, it would be difficult for us to make conclusive inferences. While it would undoubtedly hasten the analysis, it would only provide a modest, superficial view of the data. We therefore need a sector wise division of the data at hand to gain a deeper understanding of the factors involved and how they would ultimately affect each and every player within the Index.

For this, we consider 11 broad sectors that are often used by financial experts around the globe to gain a deeper understanding of the Index –
Communication Services, Consumer Discretionary, Consumer Staples, Energy, Financials, Healthcare, Industrials, Information Technology, Materials, Real Estate and Utilities.
We sectionalize our dataset and consolidate a sample of 5 companies for each sector, which would give us a sampled understanding of said sector. This division helps give a more fruitful representation of each stock and its performance on the world stage. Conversely, it also lets us know how perception and events around the globe individually affect them.

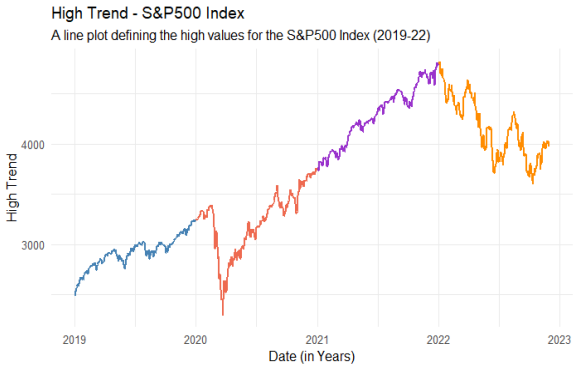
In order to consolidate the parameters in our datasets of 5 companies, we take a normalized statistic in lieu of weighted mean. Normalization is a scaling technique that helps the consolidation process by scaling all values within our purview into a preset scale. For the purposes of our dataset, we use a routine normalization technique and set our scale limits from 0 to 10. This helps us generate a dataset for each of the 11 sectors, consisting of 5 companies that act as a sample for the entire sector as a whole.

Date	mean_open	mean_low	mean_high	mean_close	mean_volume	mean_adjusted_close
1/2/2019	2.875410152	2.734594594	3.034031447	2.923372746	2.361104078	1.488088942
1/3/2019	2.673687712	2.274686522	2.747343049	2.348681078	3.329164485	0.993343503
1/4/2019	2.502328144	2.476538029	2.821450768	2.715547681	2.848023888	1.323371641
1/7/2019	2.711058339	2.600606811	2.892912667	2.73245413	2.526102834	1.337791928
1/8/2019	2.83465695	2.666498293	2.94386348	2.854239057	2.376244607	1.448805043
1/9/2019	2.861486968	2.710017076	2.92865257	2.763541993	2.172202666	1.368634105
1/10/2019	2.724609294	2.503460519	2.858238857	2.726822185	2.355540434	1.33866457
1/11/2019	2.694444737	2.603798741	2.789410753	2.769367541	1.879621187	1.374342845
1/14/2019	2.692235932	2.535413059	2.728245281	2.598427966	2.239708775	1.22069539
1/15/2019	2.615498215	2.56885045	2.902497941	2.845881464	2.264216358	1.444621204
1/16/2019	2.845939304	2.732696197	2.972002757	2.772850008	2.368371033	1.380784595
1/17/2019	2.73318534	2.68713811	2.957657592	2.907350794	2.310503329	1.501300572
1/18/2019	3.002974116	2.872637649	3.104816134	3.031266474	2.88755037	1.614816278

Data Representation of the first 20 days of the Healthcare Sector

METHODOLOGY

We begin our analysis by plotting the values of the trends of the S&P500. Plotting these trends gives us an understanding of the flow of stock prices. But it is not a good indicator of the actual behavior of stock on the index. Therefore, we also plot a candle-stick chart which gives us a better idea of high, low, open and close figures.

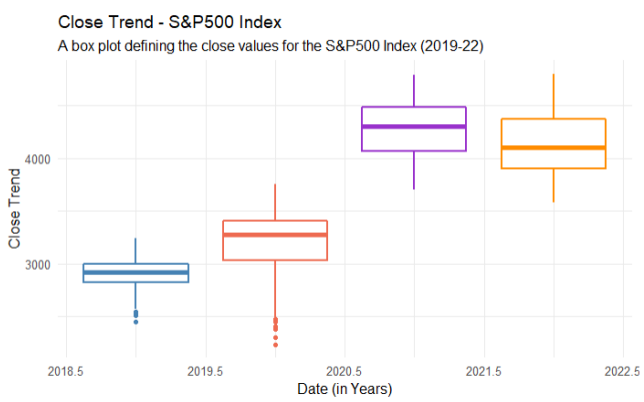


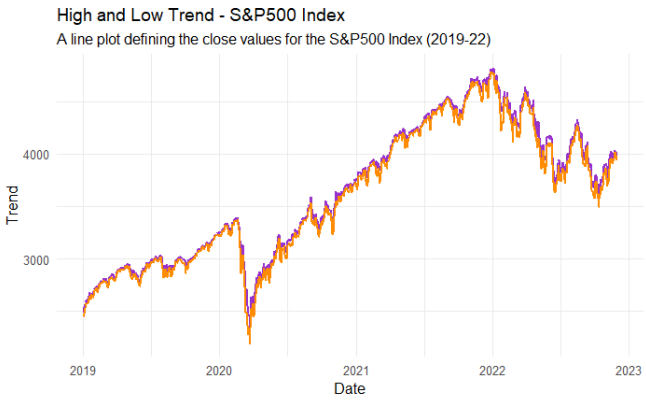
High Trend of S&P500 Index

The candlestick chart is a graphing technique That allows for the representation of all 4 Indicators on the same graph. From our graphs, we understand that there are certain points across the x-axis where there are definite dips in value. COVID is a fine example of one such dip in March 2020. November 2020 is another dip of a similar kind which was right around the time the US declared its election results. While 2021 had a relatively steady increase in the Index, 2022 saw an overall dip throughout the year due to the Russia-Ukraine war. But this does not give us an accurate representation of the individual companies or sectors that were affected by these events.

Candle Stick Chart for the 4 years of S&P500 Index

We also map out other trends in the dataset. In the Close trend boxplot, there is little variance among the majority of values. Additionally, the High and Low trend graph, while it gives us some understanding of the distribution of shares on any given day, it does not give us a good inference. Therefore, we move to the sector-wise analysis of the Index.

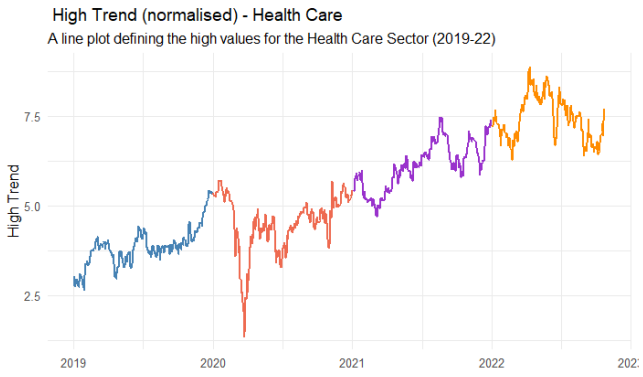


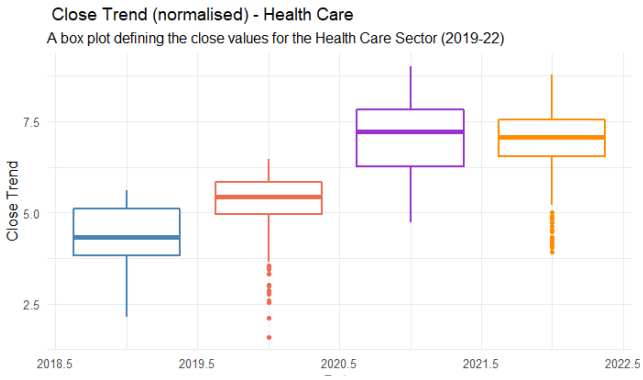


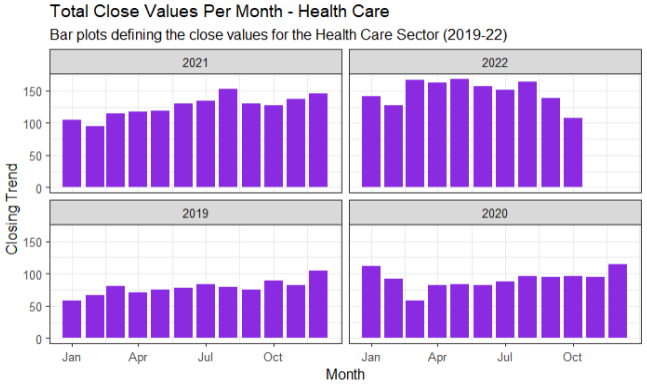
Data Representation of the first 20 days of Pfizer Stocks

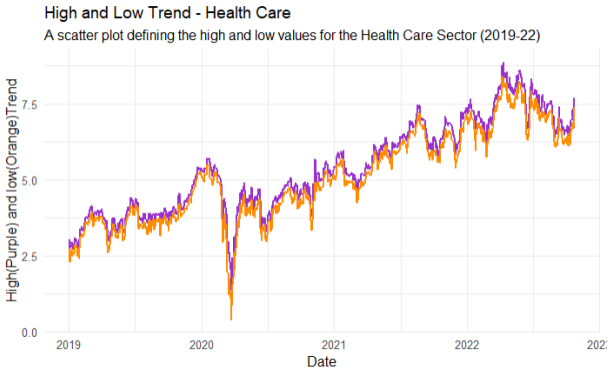
Within the Sector-wise Analysis, we develop the following graphs –

- A line graph depicting the High trend in the Index for given years
- A box plot depicting the Close trend in the Index for given years
- A sectional bar plot depicting a sum of Close trends for a given month in a year
- A line graph depicting the high and low trend for given years
- A line graph depicting the variance in trends for given years









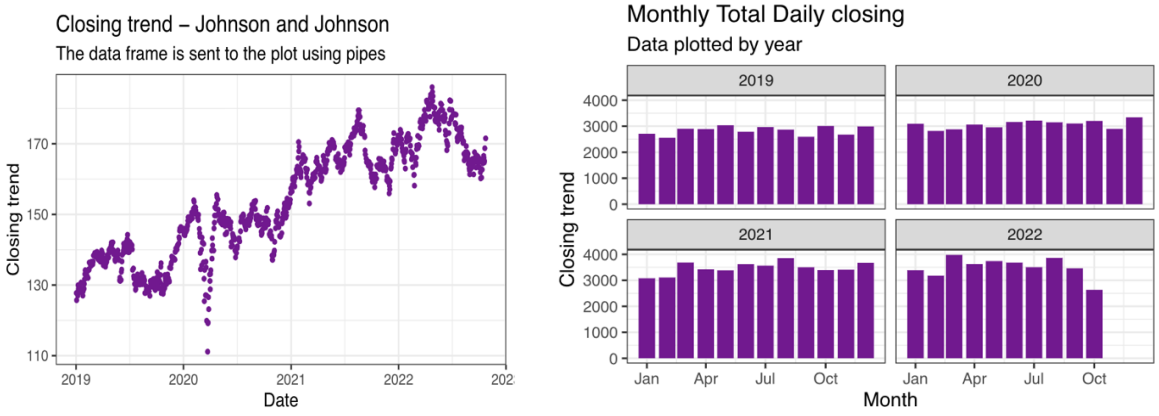
Representation of various trends of individual sectors in S&P500 Index

From our graphs, we understand that there is a huge disparity among different sectors for the same time period. While there are several sectors that were affected by COVID or the Russia Ukraine war, there were quite a few sectors that benefitted from the same. Healthcare is one such sector that benefitted in the aftermath of the Russia-Ukraine war. Given the huge supply of medicines that resurged within normal trading routes post COVID, this industry saw multiple fold increments in share price per company. Energy was a sector that was severely hampered by the war, owing to Russia closing several pipelines.

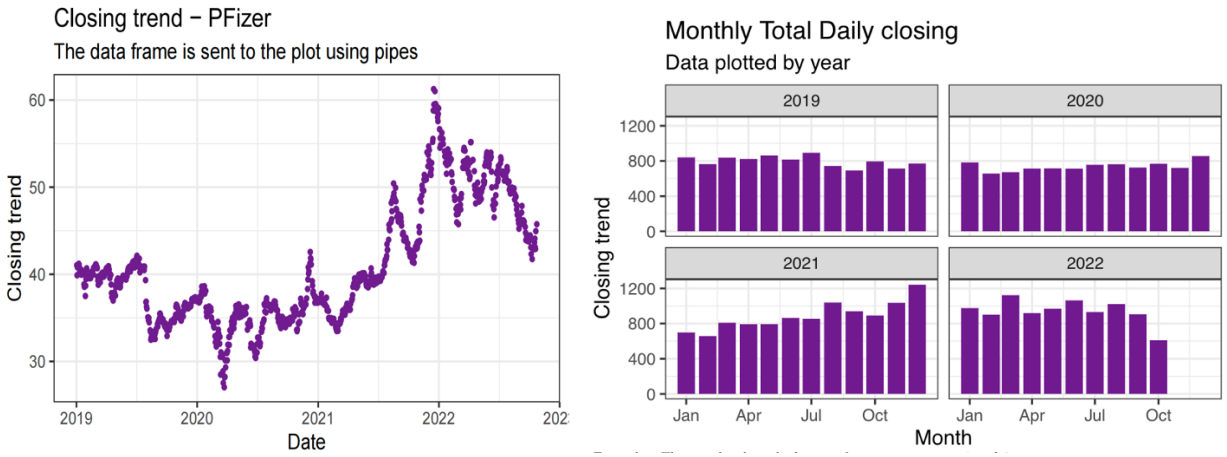
We were also able to notice significant changes in the closing trends of the dataset. Closing prices determine the price of stock of any company at the end of the market day. This in turn helps us understand the effect any particular event had on the totality of the day in which these stocks were traded.

We also get certain inferences from the comparison of high and low data in our graph. A comparison between these two trends gives us a rough estimate of the volume of shares that were traded on any particular day. There are two main lull periods that we can estimate from our graphs – a huge disparity between high and low prices at the start of COVID, and another disparity on February 7 2022, the day when Russia invaded Ukraine with over 200,000 soldiers. These days indicate a disparity because there were huge dumps of shares in the market which were never bought.

Since we see how a sector is getting affected yearly, we can carry out comparative analysis on the companies within the sector to understand which companies are mostly responsible for upward/downward trend. We have carried out the comparative analysis within HealthCare Sector on Companies Johnson & Johnson, Pfizer Inc., Merck & Co. Inc. Bristol-Myers Squibb Company, Stryker Corporation.



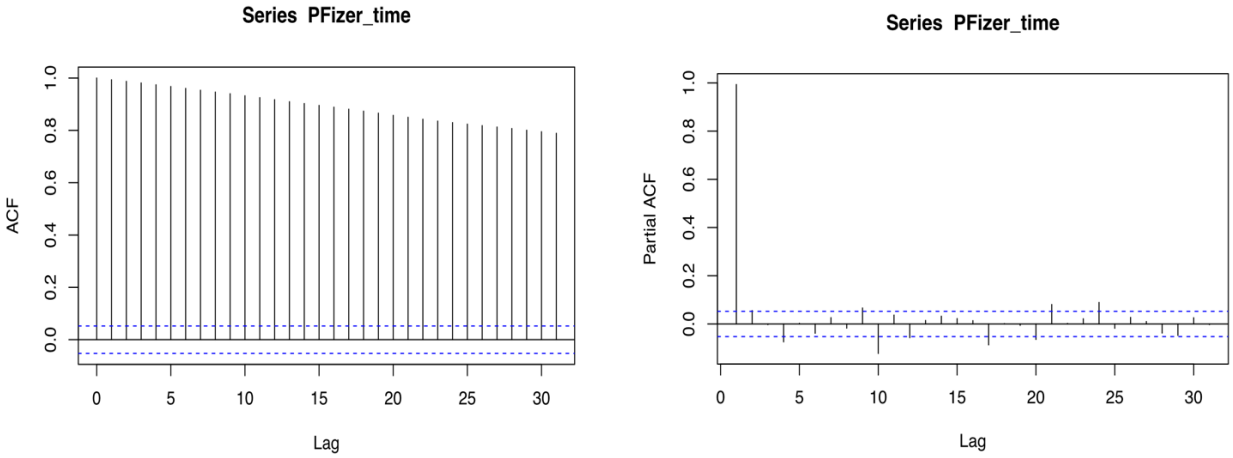
Representation of Trends of Johnson & Johnson Stock



Graphical Representation of Pfizer Stocks

Since we have made many of our analysis from the data, we will now be forecasting the data for the coming year. We have selected Pfizer data to closely observe and forecast the trend for coming year(s).

We should check first if the dataset is time series dataset and if not then we will have to convert it to TS. We have checked the stationarity of the data to see if it can be used for time series forecasting. To do so, we have checked auto correlation function i.e., ACF, partial auto correlation function i.e., PACF and ASD for which the response is follows



Graphical Representation of ACF and Partial ACF of Pfizer Stock

```
adf.test(PFizer_time)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  PFizer_time
## Dickey-Fuller = -2.295, Lag order = 11, p-value = 0.4534
## alternative hypothesis: stationary
```

From above, we see that for the ACF, the values are shooting higher than the blue dotted horizontal lines which says that they are not stationary. For partial auto correlation function, we see that it doesn't show much out of range values. But again, with Augmented Dickey-Fuller Test, we see that the p-value is not less than or equal to 0.05 which says that the data is definitely not stationary.

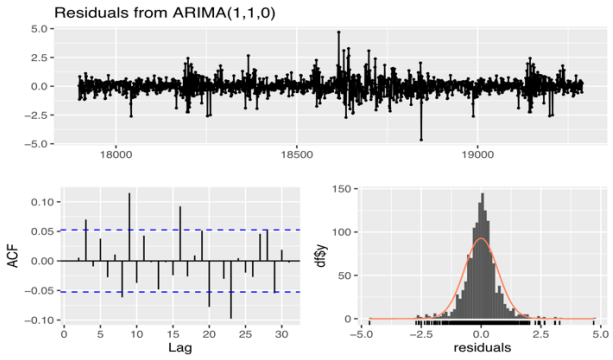
To fit a good and best ARIMA model, we have used auto ARIMA function to run all possible sets and choose the best fitting model

```
#Fit an ARIMA model

fit_arima<-auto.arima(PFizer_time,ic="aic",trace = TRUE)

##
##  Fitting models using approximations to speed things up...
##
##  ARIMA(2,1,2) with drift         : 2963.241
##  ARIMA(0,1,0) with drift         : 2967.46
##  ARIMA(1,1,0) with drift         : 2962.824
##  ARIMA(0,1,1) with drift         : 2964.56
##  ARIMA(0,1,0)                    : 2965.52
##  ARIMA(2,1,0) with drift         : 2964.356
##  ARIMA(1,1,1) with drift         : 2964.615
##  ARIMA(2,1,1) with drift         : 2964.81
##  ARIMA(1,1,0)                    : 2960.872
##  ARIMA(2,1,0)                    : 2962.419
##  ARIMA(1,1,1)                    : 2962.663
##  ARIMA(0,1,1)                    : 2962.628
##  ARIMA(2,1,1)                    : 2962.881
##
##  Now re-fitting the best model(s) without approximations...
##
##  ARIMA(1,1,0)                    : 2961.844
##
##  Best model: ARIMA(1,1,0)
```

Code Representation of ARIMA Model

After running the trace, we see that ARIMA (1,1,0) is the best fitting model. We checked the residuals and graph for ARIMA (1,1,0) as follows

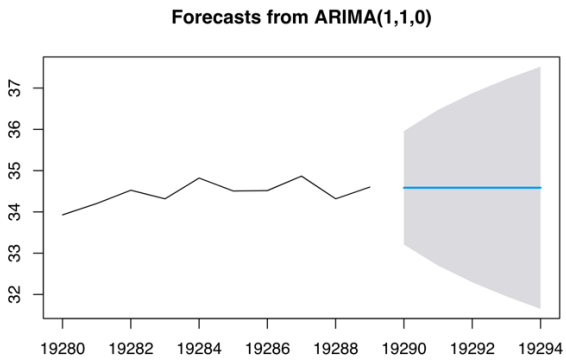
```
Series: PFizer_time
ARIMA(1,1,0)

Coefficients:
          ar1
      -0.0597
s.e.   0.0268

sigma^2 = 0.4913:  log likelihood = -1478.92
AIC=2961.84   AICc=2961.85   BIC=2972.32

Training set error measures:
                     ME      RMSE       MAE         MPE     MAPE      MASE
Training set -0.00487852 0.7004096 0.4828574 -0.02973778 1.229508 0.9988933
                    ACF1
Training set 2.756663e-05
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,0)
## Q* = 35.931, df = 9, p-value = 4.079e-05
##
## Model df: 1.   Total lags used: 10
```

Graphical Representation of ARIMA Model

Finally, we have used our ARIMA model to forecast for next five years with confidence interval as 95%. And to test if our model predicts correctly, we have used Box test of type Ljung-Box.




```
Box.test(fcst$resid,lag=5,type="Ljung-Box")

##
##  Box-Ljung test
##
## data:  fcst$resid
## X-squared = 8.9857, df = 5, p-value = 0.1096
```

Graphical Representation of ARIMA Forecast

We see here that the p-value is not below 0.05 and hence it says that our model doesn't have much correlation and is a good forecast.

CONCLUSION

The S&P500 Index is an Index used by experts around the world to gain an understanding of the current stock market and how the top companies affect the same. Through our work, we understand that this particular concept works both ways – companies are affected by world events and world events affect companies. Perception held by people for or against a company is also a huge factor in the prosperity or the downfall of a company and its reputation. The past 4 years have been a cause for concern for many economists. But our initial understanding of the Index was proven wrong on many occasions. A deeper dive into each sector showed that certain companies were affected more than others for any given event. And Pfizer's forecasting gave us a conclusive answer for its future. We therefore conclude a successful Time Series Analysis and Prediction of the S&P500 Financial Index for the years 2019 to 2022.