Data Analyzing Project on Stock Trading Strategies for Individual Investors based on the studies of USDX and ETFs

Qing Zhu and Bite Xiong

Abstract

Stock investment is an efficient approach for individual investors to fight against inflation and protect their assets. However, even low-risk stocks experience substantial price drops when the stock market crashes, exposing individual investors to high risks of losing their assets. In this project, by analyzing the stock market during the covid-19 pandemic, we try to seek stock investment strategies that can help individual investors reduce their stock investment risks during this period. Specifically, we investigate the correlation between US Dollar Index and three ETFs (QQQ, XLV and XLE), during the starting and recovery stage of the covid-19 pandemic. We implement Risk Analysis and Prophet to forecast the price of USDX. We also use Time Series Analysis and ARIMA model to forecast QQQ, XLV and XLE by analyze their own previous price.

Introduction

Personal assets can lose their value quickly with time due to inflation. To fight against inflation, individual investors can try different investment approaches. Among all these options, investing stocks is ideal for individual investors with limited savings given its higher potential return rate than bank saving accounts or investment on bounds, and zero service fee brought by the stock dealer such as Robinhood.

However, the higher potential return rate from stock markets does not come for free: individual investors can lose their whole assets if they focus on high-risk stocks. Besides, even low-risk stocks still suffer dramatic price fluctuations when the overall US stock market experiences a downtrend. For example, during the starting stage of the covid-19 pandemic, almost all stocks

(excluding options) started to encounter sharp price decreases, with individual stock investors experiencing dramatic investment loss.

Since such a situation could occur again in the future, as individual investors, we want to explore a stock investment strategy to reduce the risks when the overall US stock market is moving downward. Specifically, we want to explore:

1. During the downtrend of the US stock market (e.g., the starting stage of the covid-19 pandemic), what sectors (e.g., finance, industrial and airlines ETFs) should individual investors invest in to protect their stock accounts from the sharp drop as the overall US stock market?
2. During the uptrend of the US stock market (e.g., the recovering stage of the covid-19 pandemic), what sectors should individual investors invest in so that their accounts can bounce back as fast as (or even quicker than) the overall US stock market?

We hope that our investigation for these two questions can provide some helpful guidance to individual investors to reduce their risks in stock investments if situations similar to covid-19 occur again in the future.

To make our problem workable, we make the following simplifications:

1. We will focus on the stock performance during the covid-19 pandemic for two reasons. First, unlike

   other stock market crashes in stock history, the stock crash since 2020 is not triggered by a specific sector, which provides an ideal case to observe how different stock sectors respond to the overall market signals. Second, the US stock market is bouncing back from the covid-19 pandemic. This period provides us an opportunity to explore investment strategies during the recovery period.

2. We need to seek proxy data that can represent the overall performance of the US stock market and different sectors. For this purpose, we use the US Dollar Index (USDX) as a proxy to quantify the overall US stock market. USDX is a measurement of the price of

the U.S. dollar relative to the price of a basket of six currencies of the US's most significant trading partners. Experienced investors recognized its correlative relationship with the US stock market. We also select the three ETFs (QQQ, XLV, XLE) to represent the performance of different stock sectors.

Methodology

Firstly, we download the data from finance.yahoo.com and investing.com. Then we use python to do data Validation & Cleansing. After that we use pandas to do the Data Analysis. The Data Visualization is done by seaborn. Finally, we forecast the data and make our conclusion.

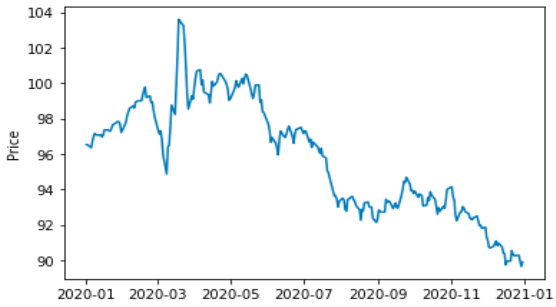
Analysis

1. Exploratory Data Analysis

1-1. Data Filtering

To Begin with, we download the USDX on investing.com and ETF: QQQ, XLE, XLV on finance.yahoo.com from 01/01/2020 to 01/01/2021. The line chart of the USDX is in Figure1.

Figure 1: USDX Price



According to our proposal, we want to explore the period of data influenced by the covid-19. Figure1 present a severe fluctuation between 2020-02-01 to 2020-05-01, which is the same time as the covid-19 spread in the US, so we choose this interval to do the exploratory data analytics.

1-2. Return Analysis

The first thing we want to know is that what is the performance of the index we chose because the return is the most important thing for the investor.

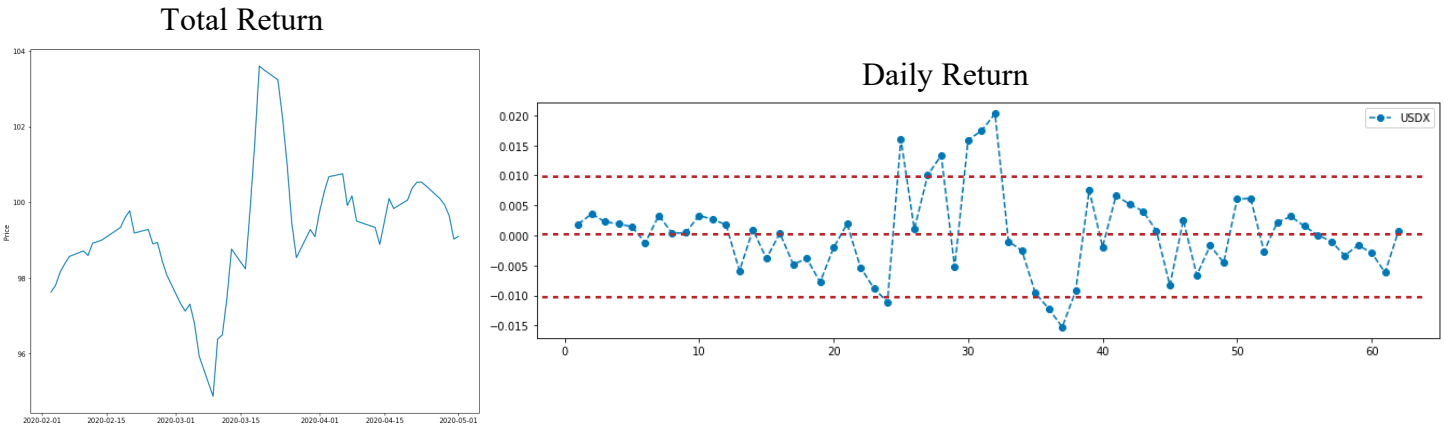


Figure 2: USDX Return Analysis

According to Figure2, we find that during the chosen period, USDX have a Total Return of 2%. This is the only index that have positive return. From the Daily Return, we can see that most of the point is located between (-0.01,0.01). This is much milder compared to the chosen ETF index.

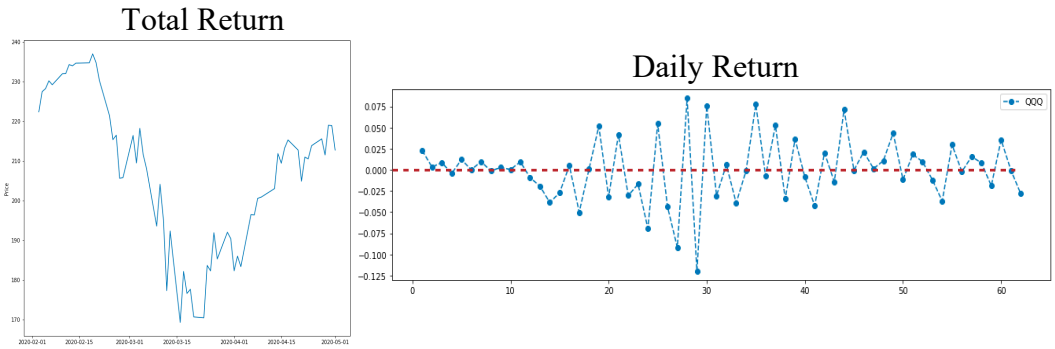
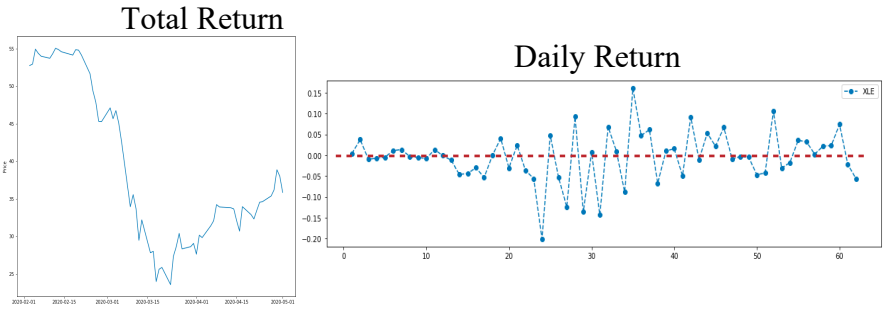


Figure 3: QQQ Return Analysis

According to Figure3, we find that during the chosen period, QQQ has a Total Return of -4%. What's more, QQQ's largest plunge and maximum rebound is 29% and 29%. From the Daily

Return, we can see that the range of the point is bigger than ETF index. There is one day of nearly -12.5% decrease in QQQ.

Figure 4: XLE Return Analysis



According to Figure4, we find that during the chosen period, XLE have a Total Return of -32%. What's more, XLE's maximum plunge and maximum rebound is 57% and 65%. From the Daily Return, we can see that there are more points lies under the 0-scale line. The range of the point is similar to QQQ.

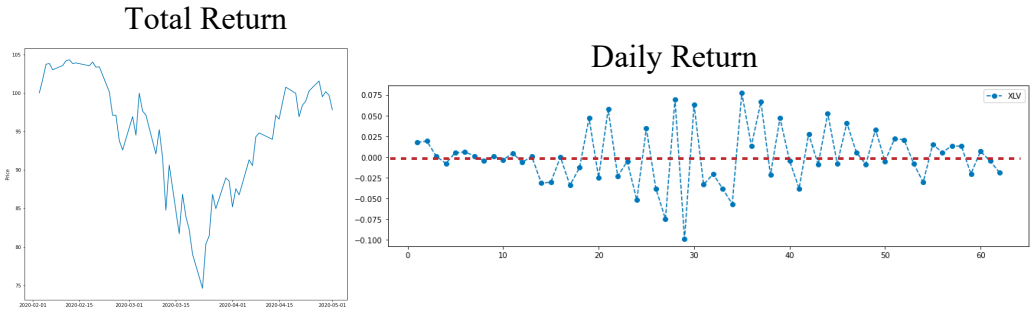
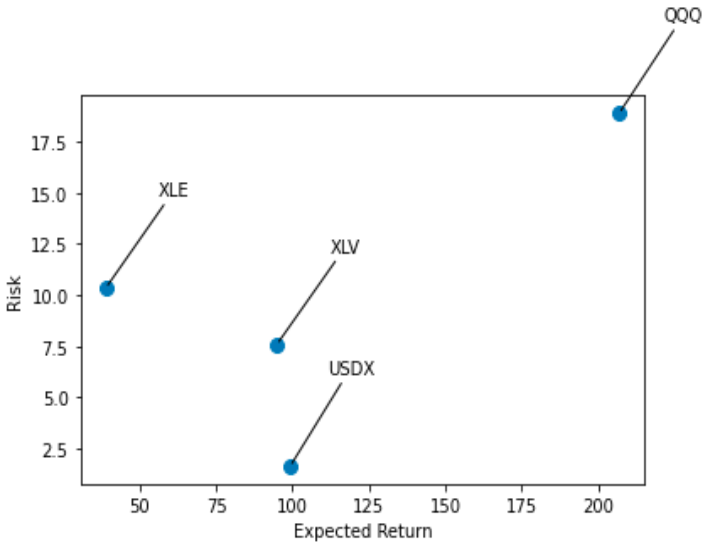


Figure 5: XLV Return Analysis

According to Figure5, we find that during the chosen period, XLV have a Total Return of -2%. What's more, XLE maximum plunge and maximum rebound is 28% and 36%. XLV Total Return is a 'V' Shape. From the Daily Return, we can see that there are more points lies close to the 0-scale line.

1-3. Risk Analysis

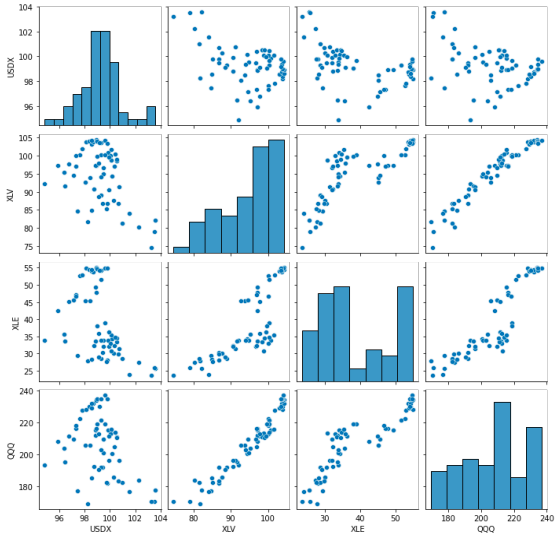
Figure 6: Risk Analysis of USDX, XLE, XLV, QQQ



In Figure6, we find that in the chosen period, QQQ has the highest Expected Return, XLE has the lowest Expected Return. What's more, QQQ has the highest Risk, USDX has the lowest Risk. In the chosen period, QQQ is an ETF that has both high Risk and high Expected Return. USDX has a good Expected Return and extremely low Risk.

1-4. Correlation Analysis

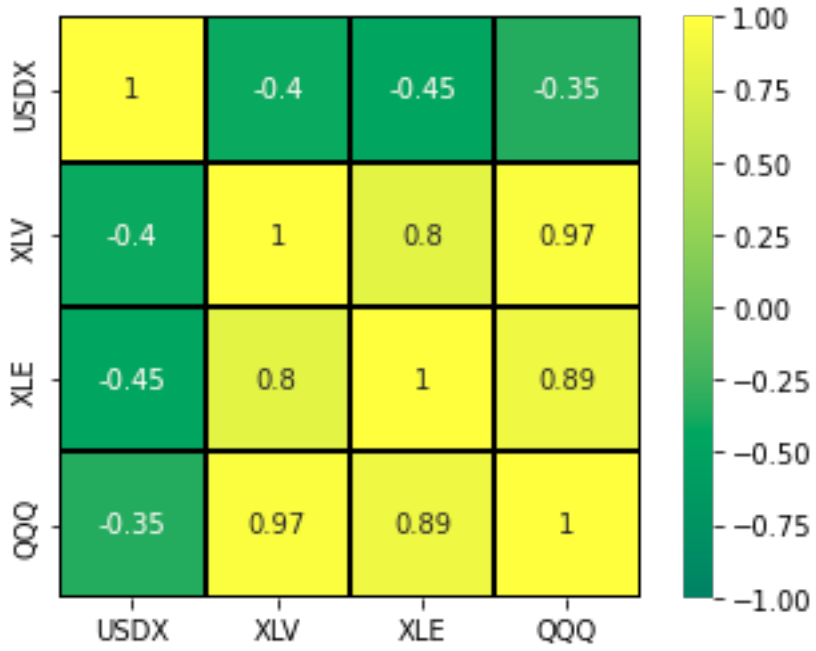
Figure 7: Pair Plot of USDX, XLE, XLV, QQQ



In Figure7, the histogram is the frequency of four indexes price. The histogram of USDX is bell-shaped. The histogram of XLV and QQQ is skewed towards the right side. The histogram of XLE is skewed towards the left side.

The scatter plot indicates the correlation between each index. For example, the graphs among QQQ, XLV, XLE are similar. The points are distributed from the bottom left corner to the upper right corner, which indicates positive correlation. However, graphs of three ETFs with USDX show that the points are distributed from upper left corner to the bottom right corner, which indicates negative correlation.

Figure 8: Heat Map of USDX, XLE, XLV, QQQ



In Figure8, we can see the concrete correlation value between each index. QQQ and XLV have the highest positive value of correlation equal to 0.97. XLE and USDX have the highest negative value of correlation equal to -0.45.

2. Two different Methods of Analysis

2-1. Forecasting USDX price by Prophet Algorithm

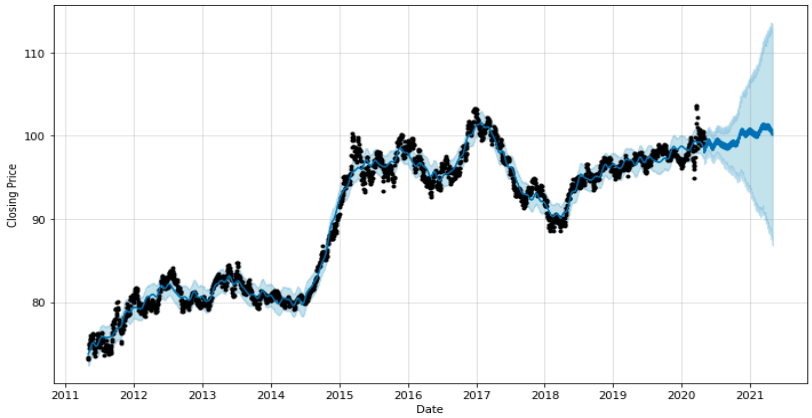


Figure 8: Forecasting Price of USDX

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well, thus is a good choice to forecast the trend of USDX. The most commonly used model for forecasting predictions is the auto-regressive model, which is the same as we used for the forecasting. In order to suite the Prophet Algorithm, we select the period of time from 05/01/2011 to 05/01/2020. Then we create the Prophet Model and plot the USDX Forecast for Period of One year. According to Figure8, the price of USDX on 04/30/2021 would be close to 100. Comparing to the actual value of USDX, which is 91.27, the error of the forecasting is about 9.57%.

2-2. Time Series Analysis – ARIMA model

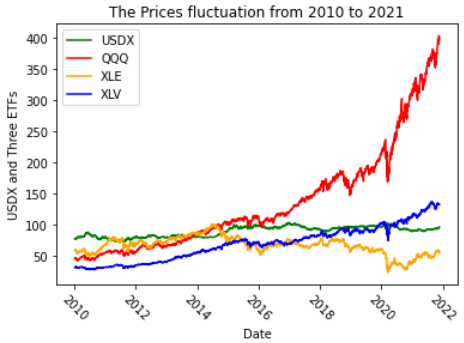


Figure 9: 10-year prices of USDX, QQQ, XLE and XLV

After implementing regression model and Prophet, we investigated the potential to predict ETF asset price with the asset's own previous price data. To investigate this capability, we need to understand the theories of Time Series and the ARIMA model.

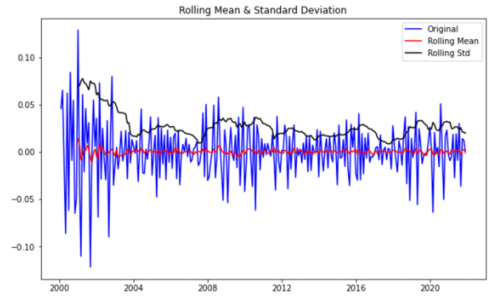
In Time Series, data has 2 variables: time and the variable we want to predict. We observe data in equal intervals, like days, weeks, or months. Our ETFs satisfy these requirements, and they also have trends – upward and/or downward movement of data. The trend of USDX is not obvious, but we still tried. They could be Time Series for our prediction analysis. However, Time Series has to be stationary before being applied in models.

From Figure 9, we can hardly see the correlation between the four assets visually. This is another reason that we want to explore them separately.

After several tests and trials, we chose to utilize the following data to conduct our analysis and develop our ARIMA model:
USDX daily data from 2010-01-01 to 2021-11-23 (250 days as lags and rolling window),
QQQ 21-year monthly data from 2000-01-01 to 2021-12-03 (12 months as lags and rolling window),
XLE 21-year monthly data from 2000-01-01 to 2021-12-03 (12 months as lags and rolling window),
XLV 21-year monthly data from 2000-01-01 to 2021-12-03 (12 weeks as lags and rolling window).

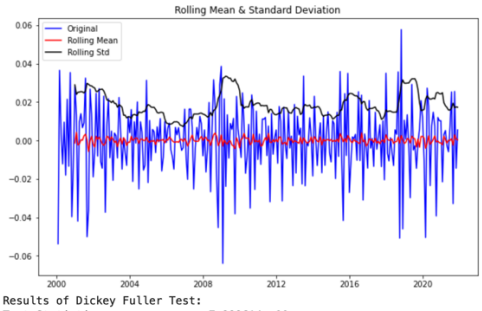
2-2-1. Remove Trend and Make Time Series Stationary

We use Log Scale Tranformation, Exponential Decay Transformation and Time Shift Transformation to detrend our Time Series. Then we use ADCF Test (Augmented Dicky-Fuller Test) to test for the stationaries. The null hypothesis is "Time Series is not stationary." We tested for all 4 assets. QQQ and XLV are stationary, but XLE and USDX are not stationary.



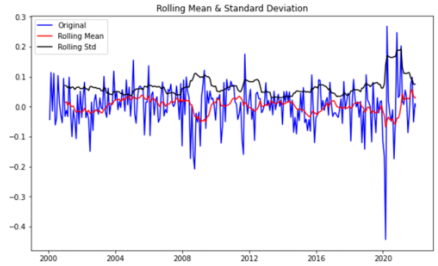

```
Results of Dickey Fuller Test:
Test Statistic                -7.323997e+00
p-value                        1.173998e-10
#Lags Used                     1.400000e+01
Number of Observations Used    2.480000e+02
Critical Value (1%)           -3.456996e+00
Critical Value (5%)           -2.873266e+00
Critical Value (10%)          -2.573019e+00
dtype: float64
```

Figure 10: ADCF for QQQ.



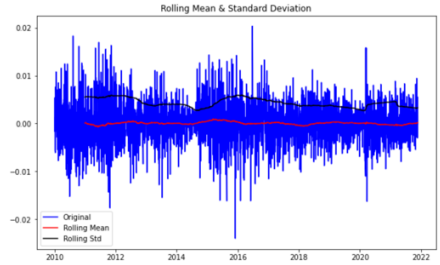

```
Results of Dickey Fuller Test:
Test Statistic                -7.999614e+00
p-value                        2.353154e-12
#Lags Used                     1.400000e+01
Number of Observations Used    2.480000e+02
Critical Value (1%)           -3.456996e+00
Critical Value (5%)           -2.873266e+00
Critical Value (10%)          -2.573019e+00
dtype: float64
```

Figure 11: ADCF for XLV.




```
Results of Dickey Fuller Test:
Test Statistic                -1.668329e+01
p-value                        1.509967e-29
#Lags Used                     0.000000e+00
Number of Observations Used    2.630000e+02
Critical Value (1%)           -3.455461e+00
Critical Value (5%)           -2.872593e+00
Critical Value (10%)          -2.572660e+00
dtype: float64
```

Figure 12: ADCF for XLE.




```
Results of Dickey Fuller Test:
Test Statistic                -1.625276e+01
p-value                        3.631480e-29
#Lags Used                     1.000000e+01
Number of Observations Used    2.961000e+03
Critical Value (1%)           -3.432560e+00
Critical Value (5%)           -2.862517e+00
Critical Value (10%)          -2.567290e+00
dtype: float64
```

Figure 13: ADCF for USDX.

From Figure 10, we can see the Test Statistic is less than the Critical Value, hence, we reject null hypothesis. P-value is so small, we can also reject null hypothesis, which means QQQ has got stationary, and we can utilize ARIMA model now. According to the bad test result of XLE and USDX, we will apply ARIMA model to QQQ and XLV.

2-2-2. Find order (p, d, q) for our ARIMA model

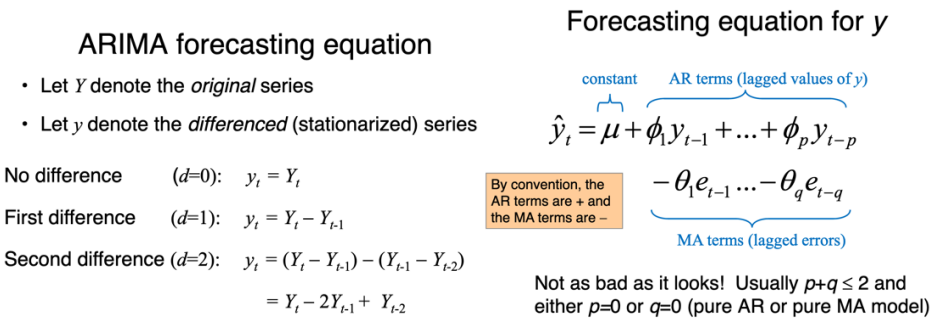


Figure 14 and Figure 15: ARIMA forecasting equation. (Nau, 2014)

ARIMA means Auto-Regressive Integrated Moving Average. P is AR terms lagged values of y. Q is MA terms lagged errors. D is integrated difference.

We displayed plots of the autocorrelation function (ACF) and partial autocorrelation function to estimate p and q based on the visualization's content and compared the Residual Sum of Squares to compare different model versions.

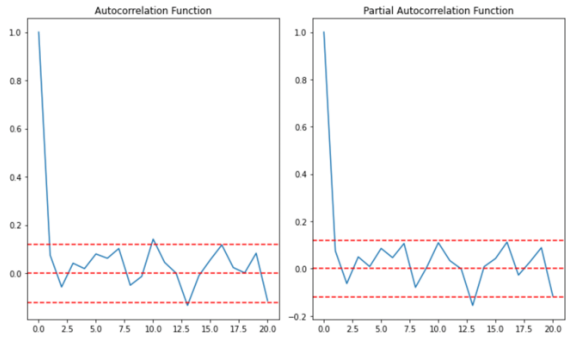


Figure 16: ACF and PACF for QQQ

We see that the curve touches y=0 and it crosses the first line at x=2 or 1. Thus, from theory, Q = 2 or 1 From the PACF graph, we see that the curve touches the y=0 line and crosses the first line at x=2 or 1. Thus, from theory, P = 2 or 1. We use the same theory to determine the Q and P for XLV (Figures 17).

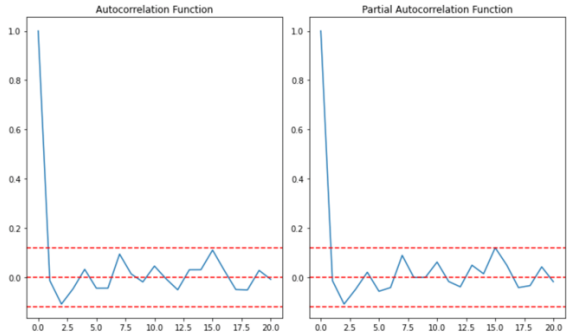


Figure 17: ACF for q and PACF for p of XLV

2-2-3. ARIMA Model Fit

We take order (2, 1, 2) as parameters to utilize ARIMA model for QQQ, and order (2, 1, 1) as parameters to as parameters to utilize ARIMA model for XLV. Both display smaller RSS (RSS-QQQ = 1.2114. RSS-XLV = 0.4298) than their AR and MA model individually. Hence, our ARIMA model performs better.

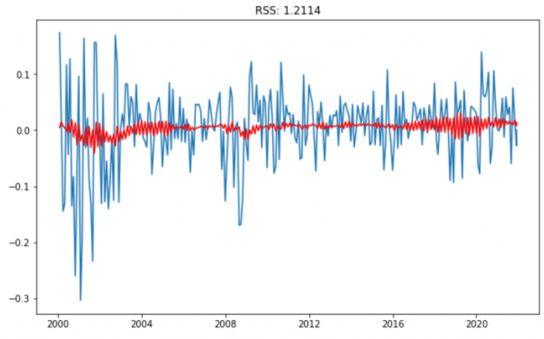


Figure 18: ARIMA fit for QQQ.

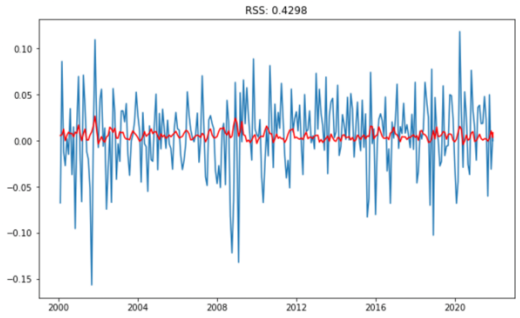
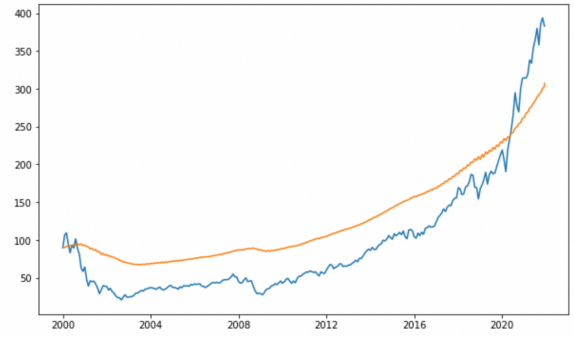


Figure 19: ARIMA fit for XLV.

2-2-4. ARIMA Model Prediction



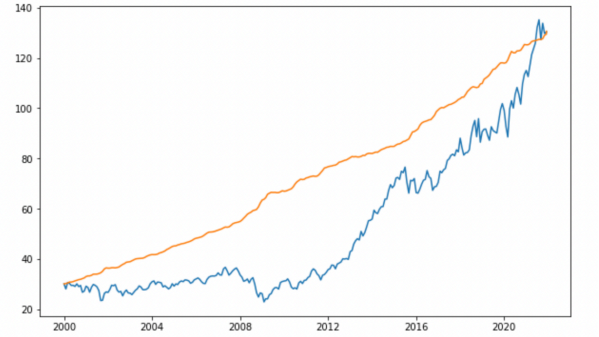


Figure 20: ARIMA prediction for QQQ. Figure 21: ARIMA prediction for XLV.

From Figure 20 and 21, we can see our model catch all the upward trend of each ETF, but it is not accurate for narrow down time intervals.

2-2-5. ARIMA Model Forecasting

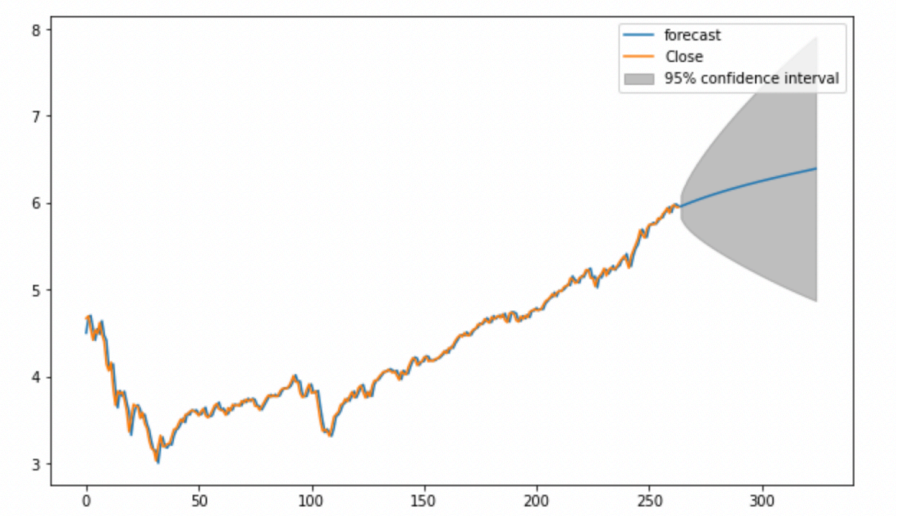


Figure 22: ARIMA forecasting for QQQ.

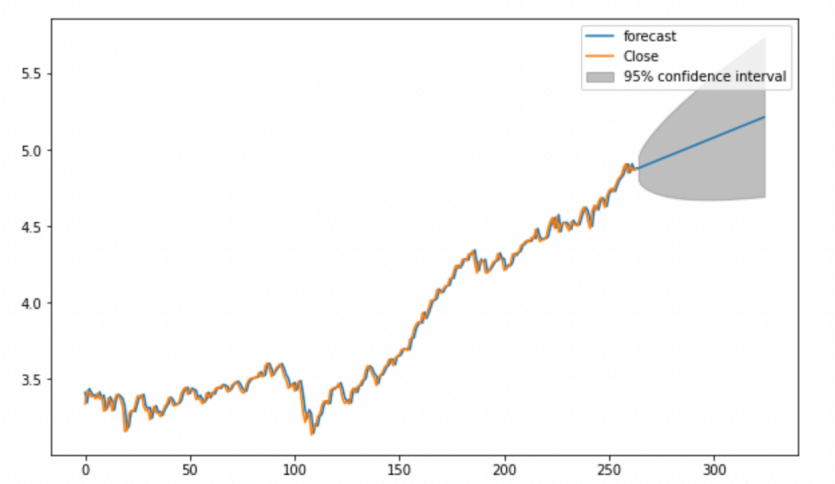


Figure 23: ARIMA forecasting for XLV.

We can see from the forecasting plots that QQQ and XLV will follow their upward trend in the future but still have 95% chance to drop in the confidence interval which may incur some price drop.

3. Conclusions

In **Return Analysis**, we find that USDX is the only index that have positive Total Return during the chosen period of time. QQQ is an ETF which tracks the Nasdaq-100 Index, ranks in the top 1% of large-cap growth-funds, so its daily return fluctuate severely. What's more, QQQ holds the stocks of companies such as Apple, Amazon, and Facebook which were not influenced by the covid-19 seriously, so its price was not plunge in the period of time. The XLV holds the stocks of companies of Health Care, so its price was also not plunging a lot. The XLE holds the company of Energy, and it had the largest negative value in Total Return.

In **Risk Analysis**, we find that in the chosen period, QQQ is an ETF that has both high Risk and high Expected Return. USDX has a good Expected Return and extremely low Risk. QQQ is more suitable for the investors who want to gain high return and can afford high risk at the same time. USDX is a more stable investment tool for the risk-averse investors.

In **Correlation Analysis**, we find that USDX has negative correlation with QQQ, XLE and XLV. We would like to suggest the investors use USDX as a tool to hedge the risk of QQQ, XLE and XLV in their portfolio.

In **Forecasting USDX Price**, we find that the predict price is higher than the actual value, which means that we underestimate the influence of the covid-19. In future research, we suggest that more variables should be counted in the analysis of USDX.

In **ARIMA** Model, we can see QQQ and XLV fit ARIMA very well. However, USDX cannot be Time Series and cannot fit ARIMA model, probably because its price has almost no trend. XLE cannot fit Time Series and ARIMA too, probably because its price follow an upward trend and a downward trend.

References

1. Binu. 'Time Series Data Analysis Stock Price Code_12'

https://www.kaggle.com/biphili/time-series-data-analysis-stock-price-code-12#3.Risk-Analysis

2. Rohan Kumar. Predicting Stock Prices Using Facebook's Prophet Medelhttps://medium.com/analytics-vidhya/predicting-stock-prices-using-facebooks-prophet-model-b1716c733ea6

3. Nau R. "Lecture notes on forecasting" https://people.duke.edu/~rnau/Slides_on_ARIMA_models--_Robert_Nau.pdf. 2014.

4. Chatterjee A."Time Series For beginners with ARIMA" https://www.kaggle.com/freespirit08/time-series-for-beginners-with-arima. July 2018.