Bitcoin-USD Prices and Prediction via SARIMA

Group member: Yijun Shen (1/3), Zehua Cheng (1/3), Yiwei Wang (1/3)

Date: 12/11/2022

Fall 2022

Introduction

The time series analysis is used to understand the tendency and pattern of a given data over time. In this project, we are trying to focus on analyzing and visualizing the future trends of the capital market using the ARIMA model.

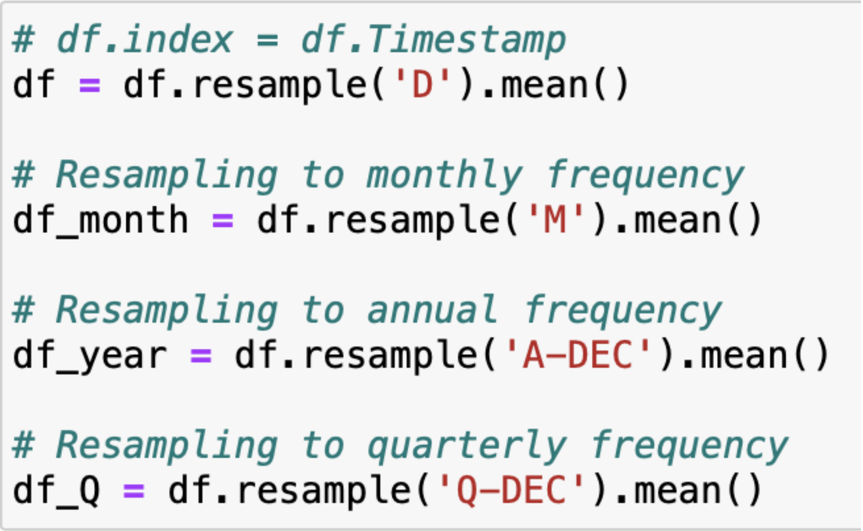
Capital market is one of the largest industries on earth, including currencies, equities, and derivatives. The percentage return of the capital market is normally assumed as Gaussian distribution. However, due to the continuous frequent impacts caused by various other industries, the capital market cannot be predicted directly through Gaussian distribution. Therefore, in order to conduct the prediction of the return rate of the capital market, ARIMA model is applied to Bitcoin-USD data to forecast the future tendency of the cryptocurrency.

Data selection

For the project, in order to reflect the macro perspective of the whole capital market through a single asset, we decided to focus on the price prediction of Bitcoin-USD, since cryptocurrencies are not only one of the most heated topic in recent years, but they are also known for their high volatilities and correlations to the economic factors such as risk-free interest rate. Bitcoin, among all other cryptocurrencies, has relatively high liquidity, therefore it becomes the perfect choice for our study. We obtained our dataset by employing the yahoo finance data package in Python to import BTC-USD's daily adjusted price from 2017-01-01 to 2022-07-31.

Data processing

For the data processing and data cleaning part of the project, we initially resample our existing daily Bitcoin price dataset to monthly, quarterly, and annually data to check the smoothness of each data. Figure 1 represents the code we used to modify the existing dataset.



```
# df.index = df.Timestamp
df = df.resample('D').mean()

# Resampling to monthly frequency
df_month = df.resample('M').mean()

# Resampling to annual frequency
df_year = df.resample('A-DEC').mean()

# Resampling to quarterly frequency
df_Q = df.resample('Q-DEC').mean()
```

Figure 1

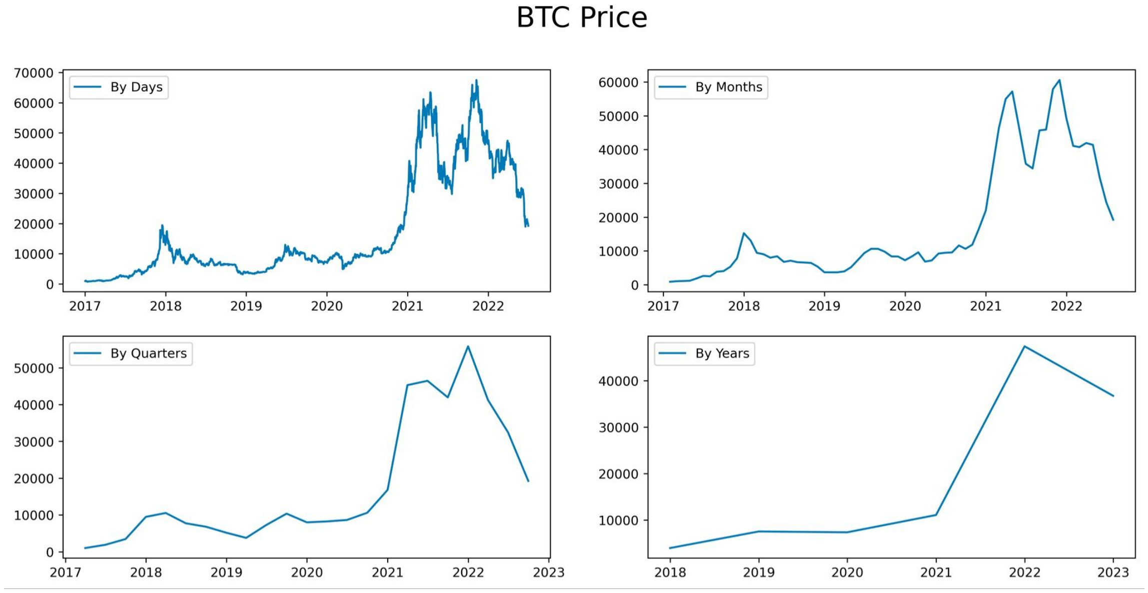


Figure 2

From Figure 2, according to the 4 plot, it is clear that monthly data is the smoothest data, which should be used for further modeling.

After data modification, the next step of the project is to check for stationarity of the series. We employed the statsmodels.api package in Python to plot the decomposition of the series and conduct the Dickey-Fuller test to check if our dataset is stationary.

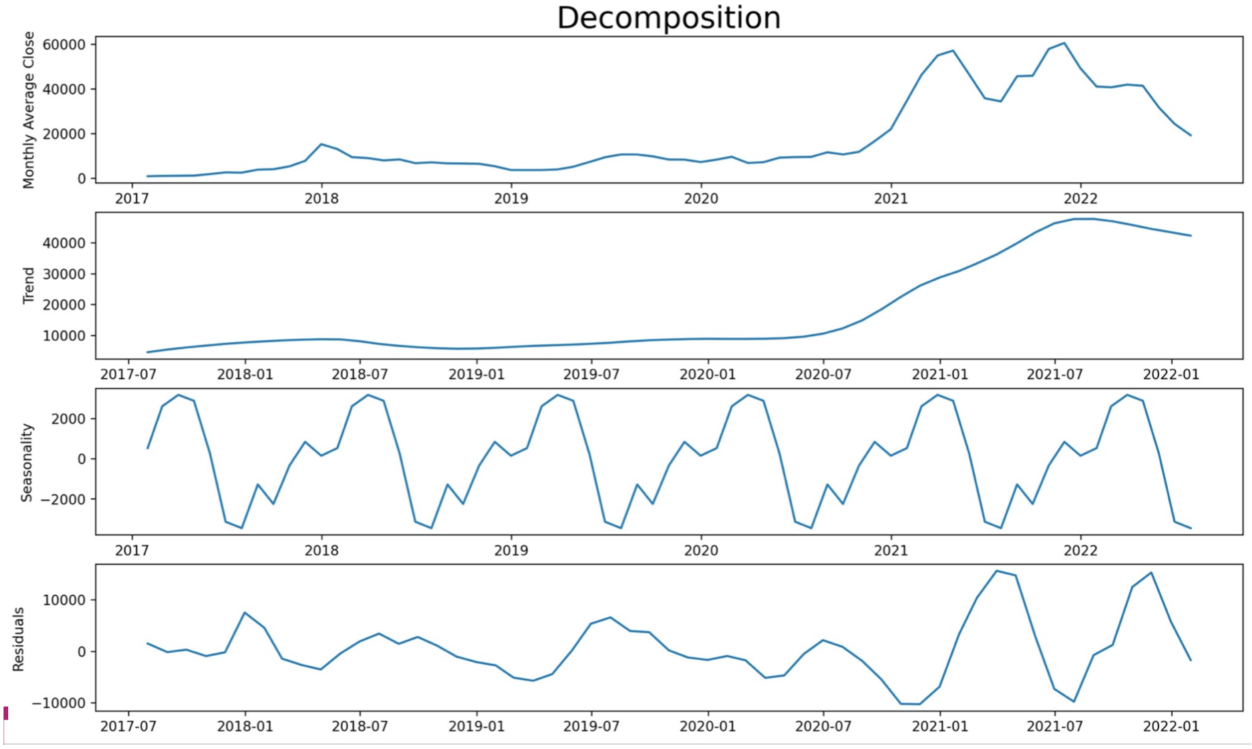
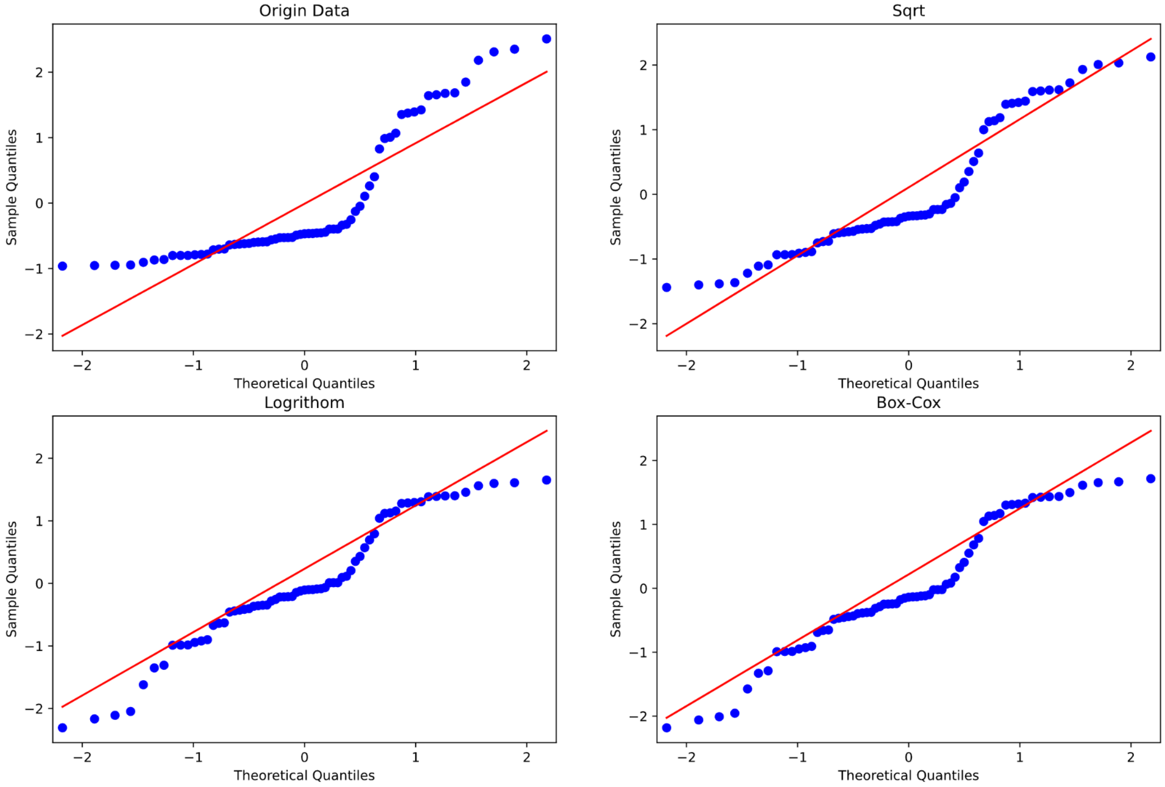


Figure 3

From figure 3, we found that its trend component is not a constant term, so is the seasonal component. For the Dickey-Fuller test, the p-value resulted in 0.48, suggesting that we cannot reject the null hypothesis that assumes the series is not stationary.

Furthermore, we are going to conduct MLE to estimate the coefficients of our model. However, a prerequisite of MLE is that the dataset has to follow a normal distribution. In order to ensure the qualification of our model, we need to check if our dataset is normally distributed. We conduct four normalization tests and qq-plots for the original monthly dataset, the square root of the dataset, the logarithm of the dataset, and the Box-Cox transformation of the dataset.


Origin Data

Sqrt

Logrithom

Box-Cox

Figure 4

Figure 4 contains four different qq-plots of the dataset. According to the qq-plot, the original data points are not on the diagonal, which means monthly data is not normally distributed. Furthermore, the P-value of normalization test of our original monthly dataset is smaller than 0.01. In other words, we can reject the normal distributed hypothesis. Compared with the other three P-values of normalization test, the logarithm transformation of the dataset has the largest P-value of 0.7, which means the data has the best normality. Thus, we finalized our dataset as the logarithm transformed data. For the Box-Cox transformed data, we optimize our data with $\lambda = 0.06$ which the below function.

$$x_{\text{transform}}(\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln x, & \lambda = 0 \end{cases}$$

Model Fitting

After the transformation, the P-value of the ADF test for transformed data is 0.1818, which means that the series are still not stationary as suggested by the Fuller test result. Since the series also exhibits seasonality, we also conducted seasonal difference to further remove seasonality of the series. Since the series also exhibits trends, we wanted to conduct regular difference to remove the trend in the series. After this step, the P-value of the ADF test for seasonal and regular differenced transformed dataset is 0.000576, which is perfect for further analysis. The decomposition graph ensures that we have successfully removed trend and seasonality of our dataset.

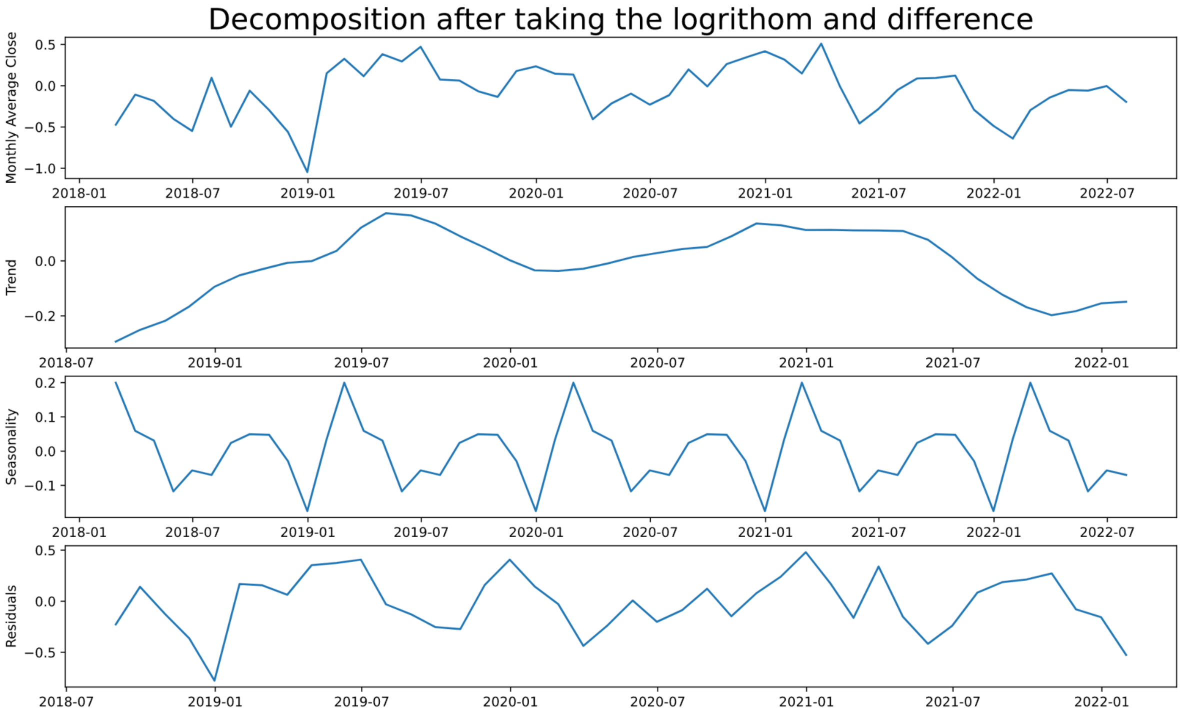


Figure 5

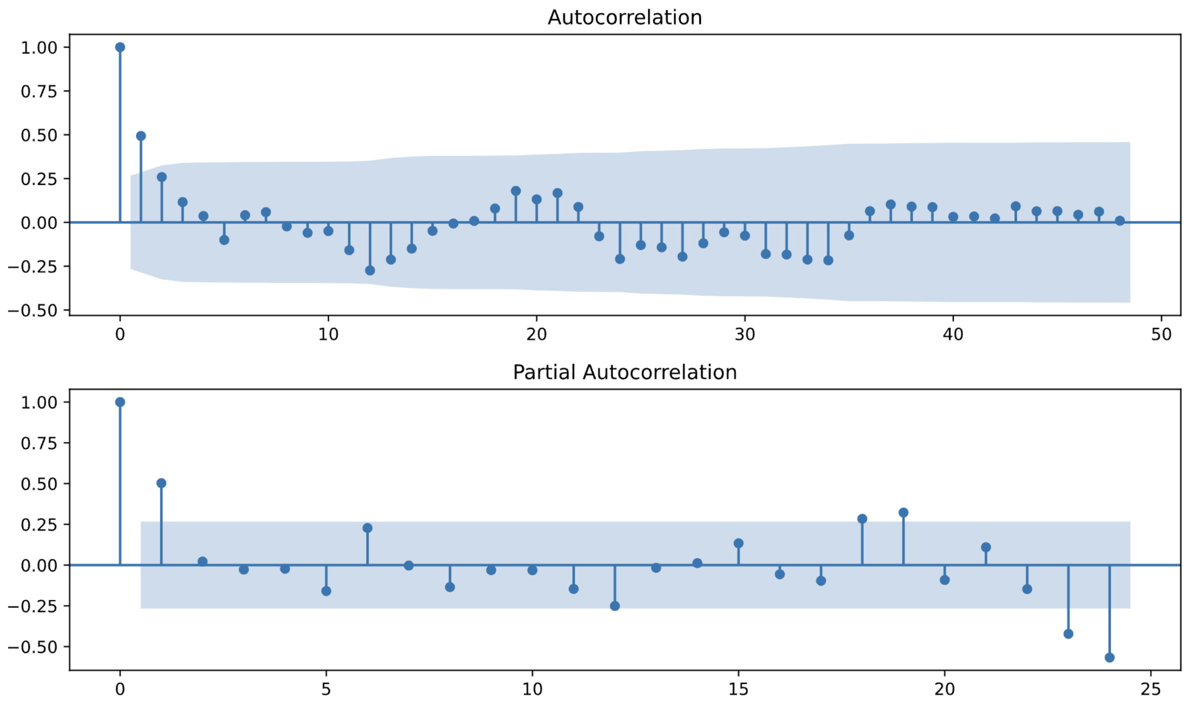
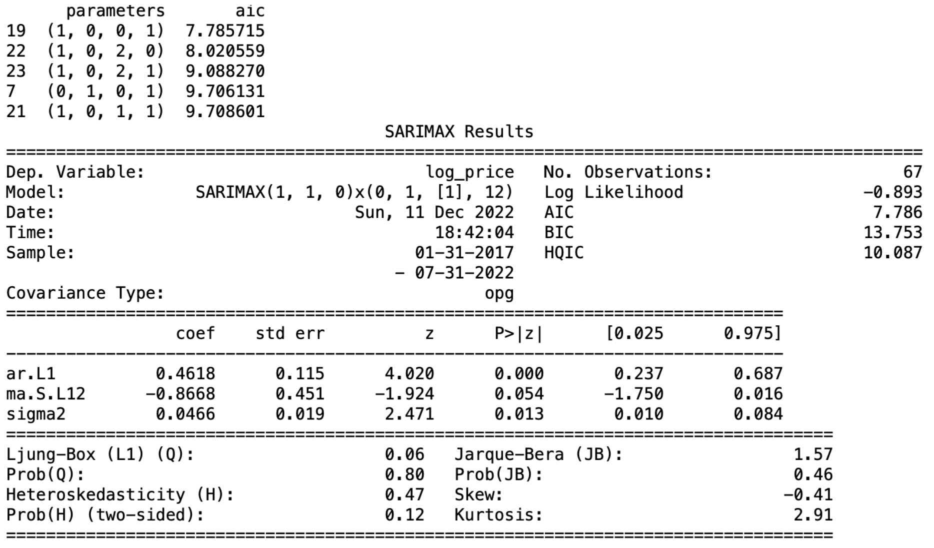


Figure 6

Figure 6 represents the ACF and PACF of the transformed data. We are trying to initial approximation of parameters using Autocorrelation and Partial Autocorrelation Plots. From the ACF graph, it was cut off after lag 1, so we may be able to conclude that p equals to 1. The PACF graph shows a tendency of exponential decay, which means that there is seasonality in the model. We need to further confirm our model by calculating AIC.



```
      parameters       aic
19  (1, 0, 0, 1)  7.785715
22  (1, 0, 2, 0)  8.020559
23  (1, 0, 2, 1)  9.088270
7   (0, 1, 0, 1)  9.706131
21  (1, 0, 1, 1)  9.708601
                                      SARIMAX Results                                       
============================================================================================
Dep. Variable:                            log_price   No. Observations:                   67
Model:             SARIMAX(1, 1, 0)x(0, 1, [1], 12)   Log Likelihood                  -0.893
Date:                              Sun, 11 Dec 2022   AIC                              7.786
Time:                                      18:42:04   BIC                             13.753
Sample:                                  01-31-2017   HQIC                            10.087
                                       - 07-31-2022                                         
Covariance Type:                                opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4618      0.115      4.020      0.000       0.237       0.687
ma.S.L12      -0.8668      0.451     -1.924      0.054      -1.750       0.016
sigma2         0.0466      0.019      2.471      0.013       0.010       0.084
===================================================================================
Ljung-Box (L1) (Q):                   0.06   Jarque-Bera (JB):                 1.57
Prob(Q):                              0.80   Prob(JB):                         0.46
Heteroskedasticity (H):               0.47   Skew:                            -0.41
Prob(H) (two-sided):                  0.12   Kurtosis:                         2.91
===================================================================================
```

Figure 7

According to AIC, the best model is Seasonal $ARIMA(1,1,0)\times(0,1,1)_{12}$, specifically in the form as $\Phi(B12)\phi(B)\nabla12\nabla x_t = \Theta(B12)\theta(B)w_t$. The backshift operator of our model is $(1 - 0.4618B)\nabla12\nabla x_t = (1 + 0.8668B12)w_t$ with white noise's variance $\sigma_w^2 = 0.04660$.

In order to test the goodness of our model, we are constructing the Box-Ljung test. We obtain the P-value of the Box-Ljung test with lag equals to 6 as 0.998 and when lag equals to 12 as 0.153, which prove that the residuals of our model are white noise. Thus, the model is a good fit to our dataset.

Prediction

In the last step, we wanted to predict the near future performance of Bitcoin. We decided to predict the next 6 month performance using the model described above. The following figure 8 is our prediction of the future price of Bitcoin and fitness of previous dataset.

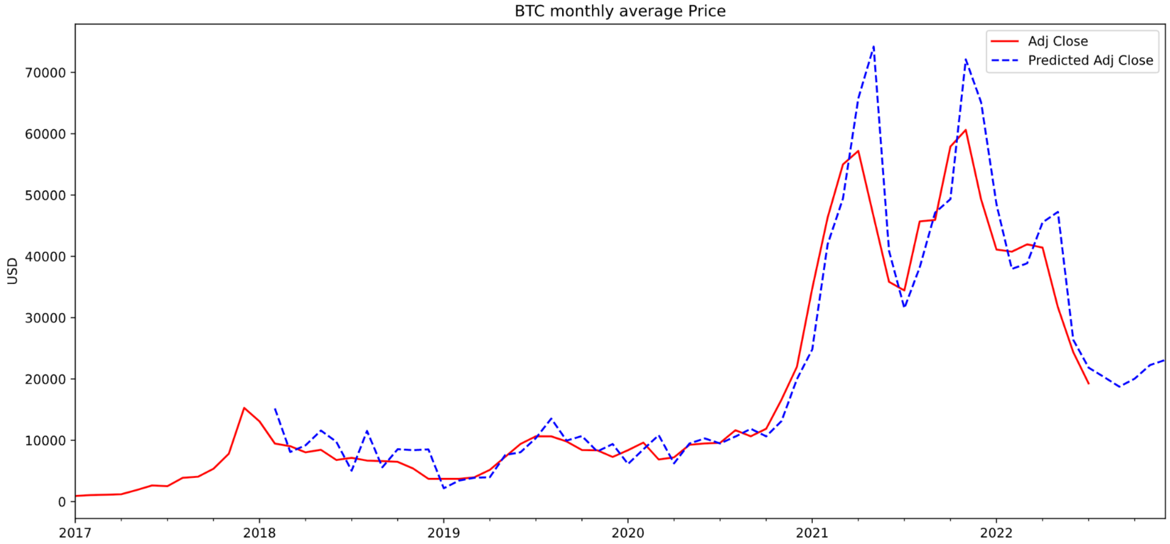


Figure 8

Conclusion & Potential Improvements

Bitcoin is a highly volatile investment asset. It is surprising for us to find out that it does exhibit trends and seasonality, providing us another potential aspect to understand and to predict the price/value of it. Moreover, the usage of Python built-in packages and data libraries make the analysis and prediction processes possible. Meanwhile, we also realize that there are a lot of other factors besides the historical price of Bitcoin that can be imperative variables of the future price tendency of Bitcoin, but because of time limitation and main focus of time series analysis, we are unable to dig into the details of each independent variable and conduct further analysis. For further improvements, we could employ other machine learning models such as Neural Network, to decompose Bitcoin prices from another angle.