# Bitcoin price forecast

Gagik Khalafyan

May 2022

## Table of Contents

# Abstract

The cryptocurrency market is expanding with exponential rates, and bitcoin is one of the leading players there. Its price depends on many factors inside and outside the financial world, but also its values in previous periods can play an essential role.

The following study concentrates on analyzing the available data on bitcoin price and applying various time-series models to predict future values. Some models, like SARIMA and Holt-Winter's ETS don't provide significant results, but VAR models make a breakthrough and prove that they can be a convenient tool in the financial market.

## Introduction

Cryptocurrencies are, in fact, currencies that are designed with encrypted protocols and aim to reduce counterfeiting and prevent currency counterfeiting. The most important feature of digital currency is its decentralization. That is, no particular institution, organization, or government controls it.

After the stock market slump and the news of the strange price increase of some currency codes, many investors' attention was drawn to the digital currency market. Of all the cryptocurrencies, bitcoin is the most popular and has gained worldwide popularity. Ethereum is another digital currency that has the potential to grow, But many people still do not know it. Bitcoin is also a cryptocurrency based on the blockchain process whose identity is unknown. The most important feature of this currency code is its scarcity, which eliminates the possibility of inflation. Bitcoin is a decentralized currency accessible only to individuals and does not depend on institutions and banks.

Taking into account current market trends and knowing the future values of cryptocurrencies is essential that is why we tend to create a model that will be able to analyze the available data and supply accurate supply predictions for bitcoin price.

Knowing that information will be a powerful tool in the current cryptocurrency market and give individuals and companies a lot of competitive advantage.

The following two sections of the paper will represent the literature review and research methodology, after which we will move on to the core analysis, results and conclusions.

## Literature Review

Several reports were reviewed to correctly analyze the data and choose the appropriate approach for the analysis, where the following methods have been used:

- Modified model BART (Binary Auto-Regressive Tree), which was more accurate than the ARIMA for a short term period [2]
- Long Short Memory (LSTM) is highly efficient in predicting the volatile dynamics inherent in crypto-currency markets [3]
- Artificial Neural Network (ANN) is effective for longer-term history [4]

# Research methodology

The data for the research was taken from binance.com via Kaggle[1] in '.csv' format, and Python programming language was used to apply the required analysis. It initially contains observations for each minute, but it was aggregated to display daily data.

For the research, the following models have been applied:

- SARIMA(from the ACF and PACF plots, auto.arima() function result)
- Holt Winter's exponential smoothing
- VAR

# Analysis and results

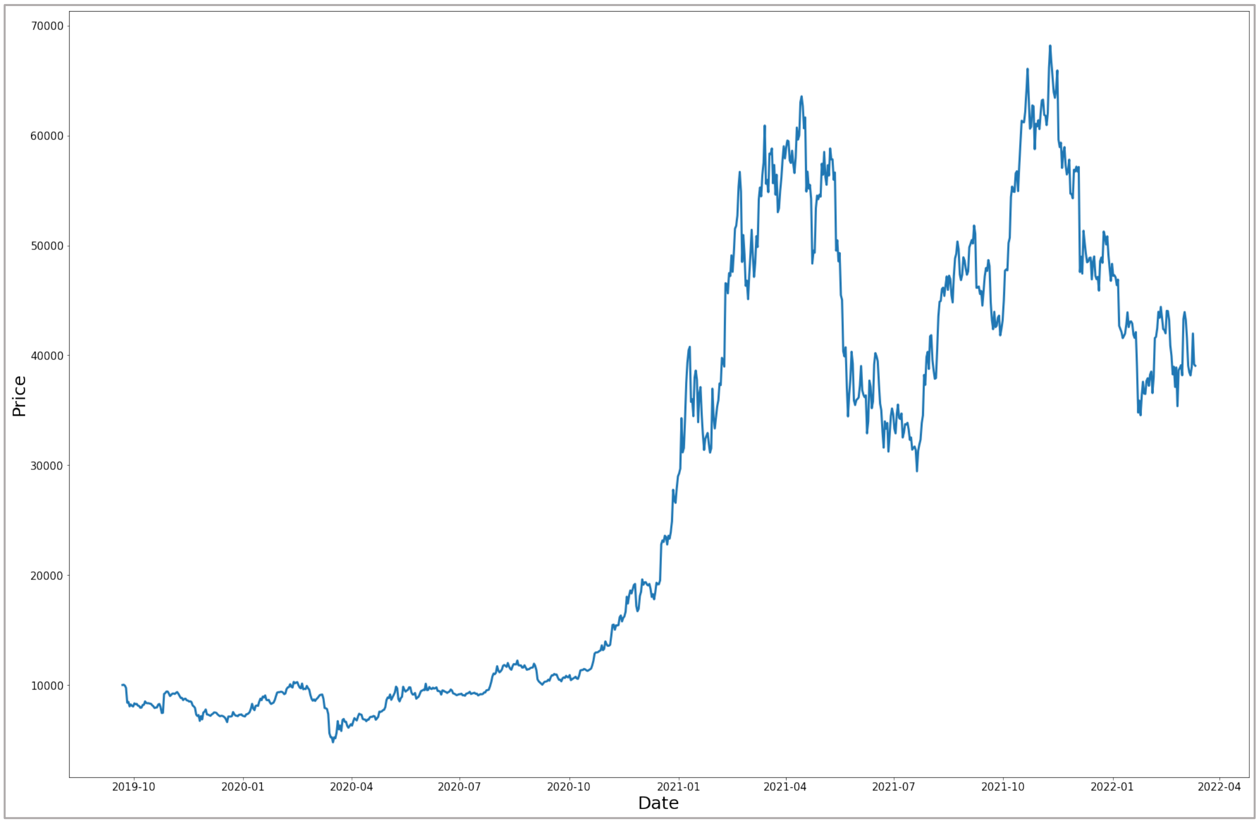


**Figure 1: Bitcoin price over time (BUSD)**

From the plot, it is clear that the series has a multiplicative trend. Still, there is no clear picture of seasonality, as only two regions seem to perform the same, From April 2021 to July 2021 and from July 2021 to January 2022, which is not enough to conclude that there is seasonality.

### - SARIMA

Before transforming the data, the p-value of the ADF test is 0.705, which is greater than 0.05, which means we need to accept the null hypothesis and conclude that our time series is not stationary.

For the KPSS test p-value is 0.01, which is less than 0.05, which means we need to reject the null hypothesis of the KPSS test and conclude that our time series is not trend stationery.

After applying first-order differencing to the data, the situation changes, and the series becomes stationary, meaning the SARIMA model can be created.

To identify dependence orders of the model, ACF and PACF of the transformed series will be created:

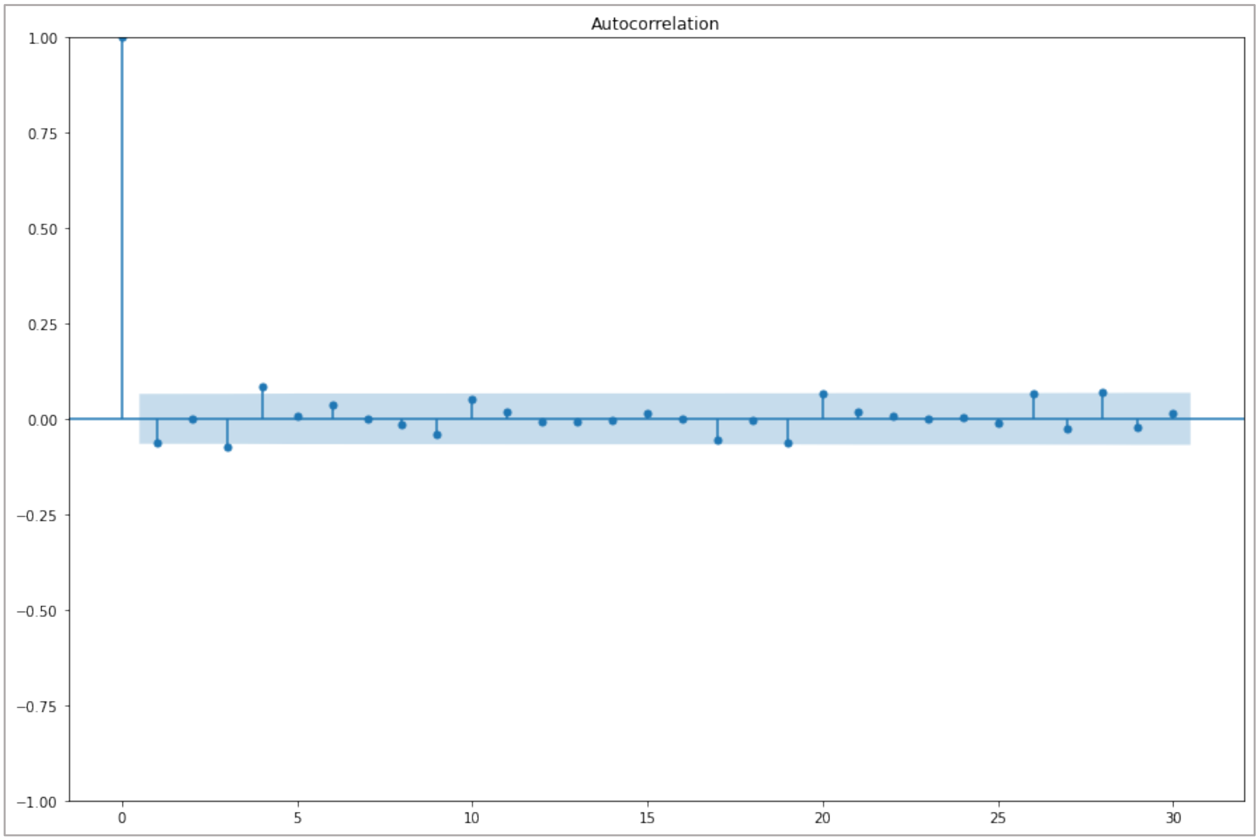


**Figure 2: ACF of the series**

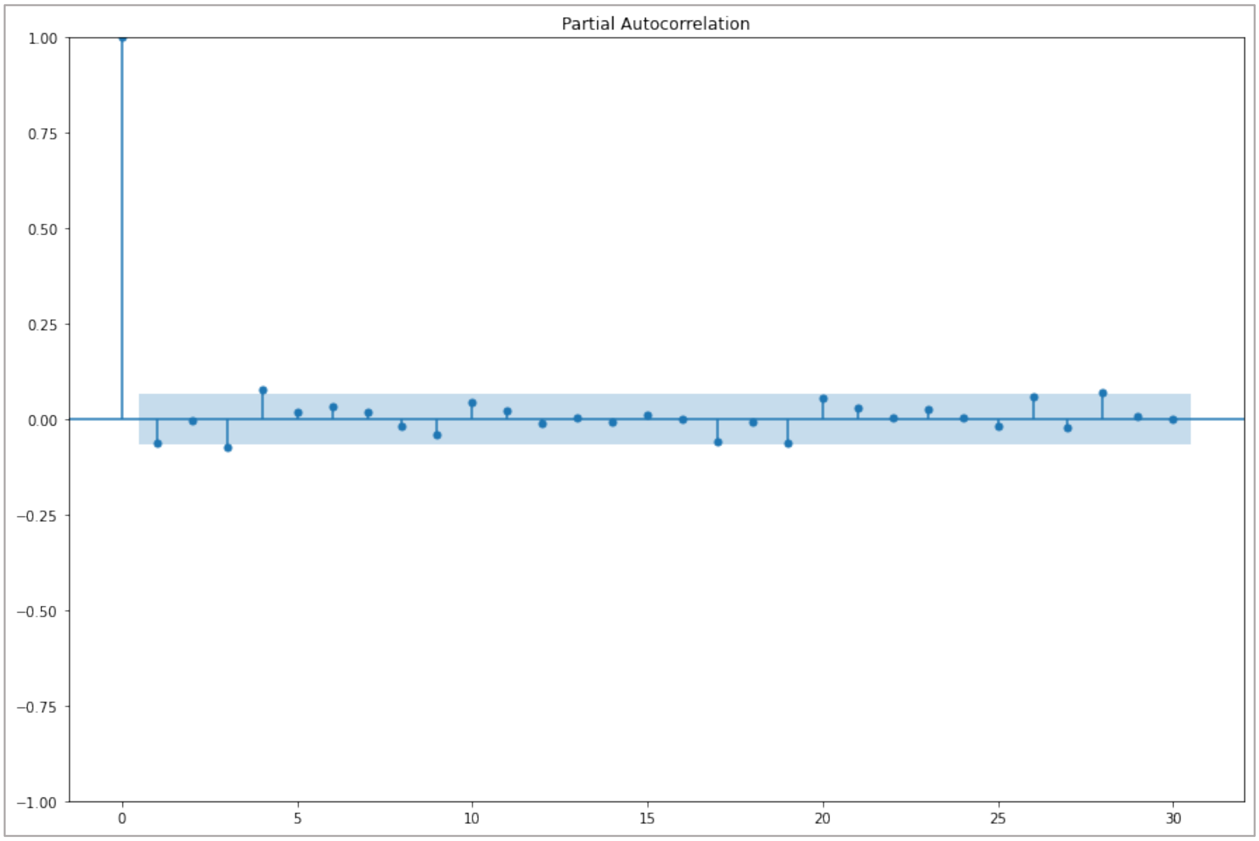
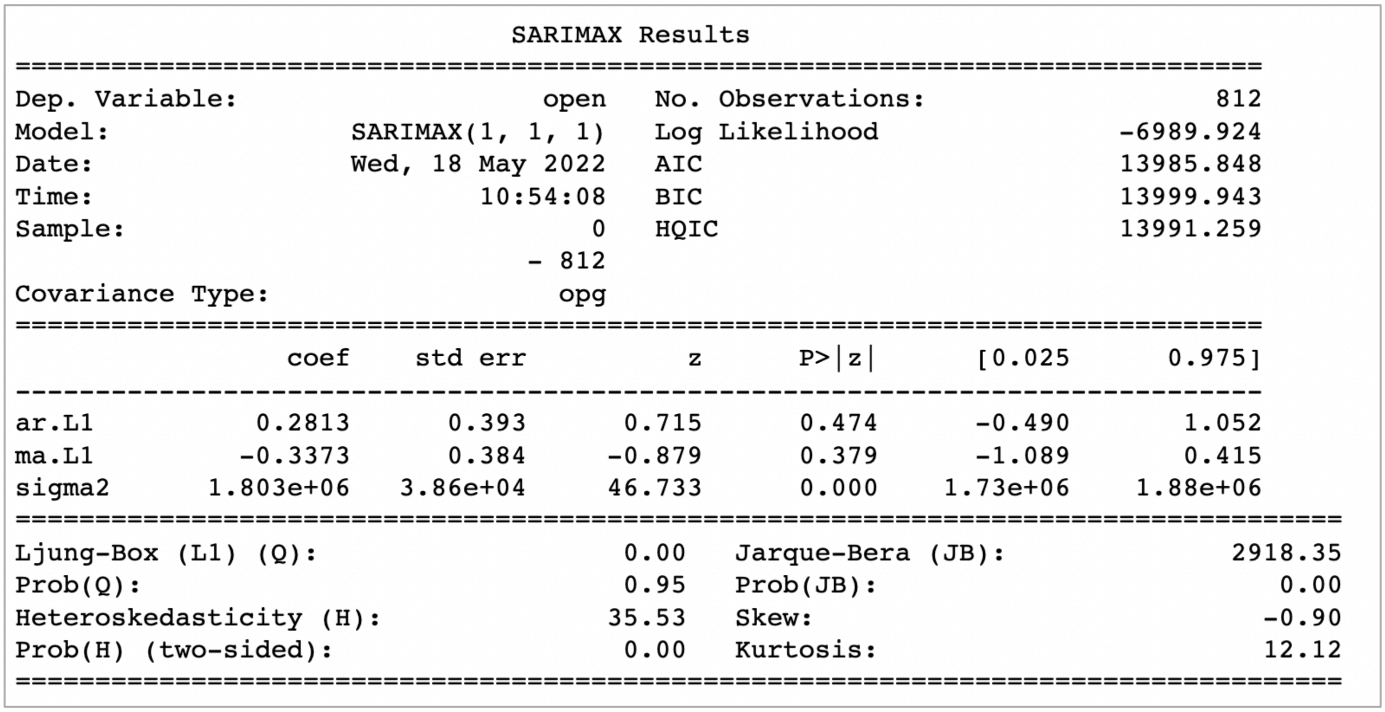


**Figure 3: PACF of the series**

We see that both ACF and PACF immediately cut off, having just a single lag on the edge of the threshold. Thus, we will use AR(1,1) model. Taking into account that we applied first-order differencing, the model will be ARIMA(1,1,1) also converting to SARIMA; we will have SARIMA(1,1,1)(1,0,1,30) as we don't have a seasonal component in the data.

Now we split the data into training and testing sets, dedicating 10% of it to the test set and initially fit a SARIMA model into it.

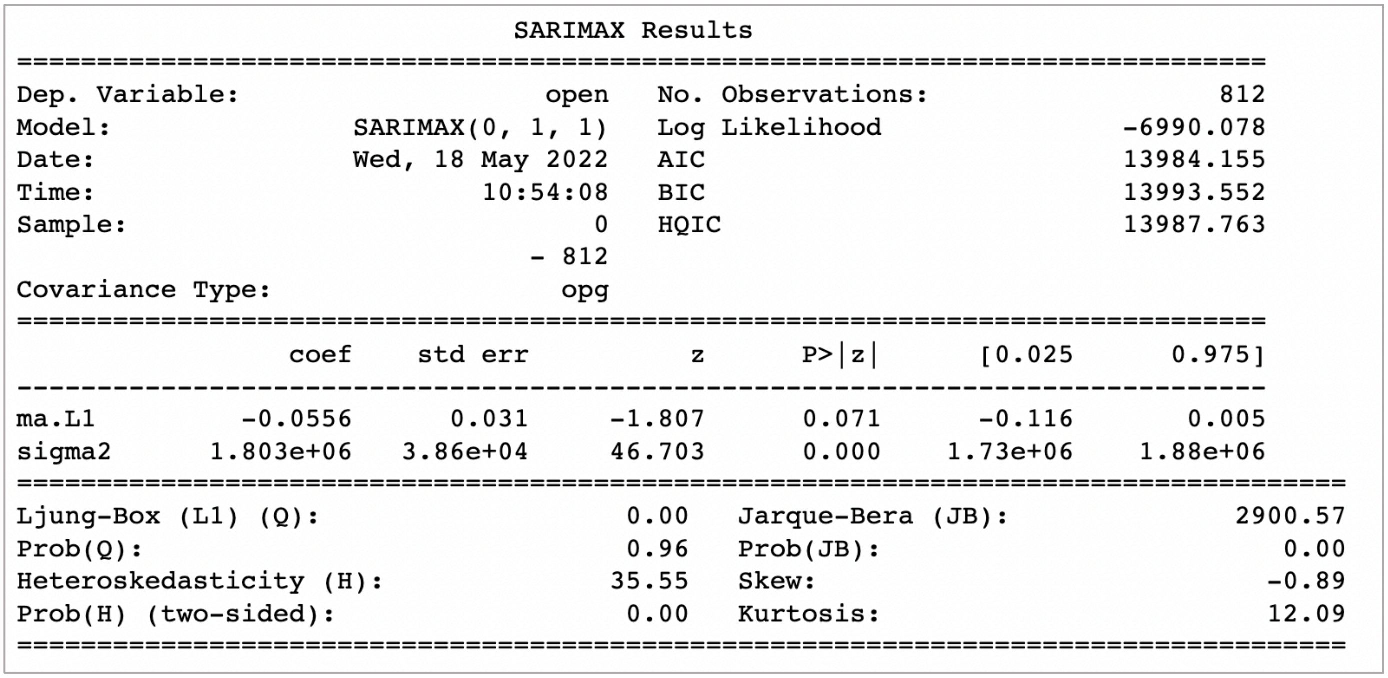
The SARIMA model provided the following results:



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   open   No. Observations:                  812
Model:               SARIMAX(1, 1, 1)   Log Likelihood               -6989.924
Date:                Wed, 18 May 2022   AIC                          13985.848
Time:                        10:54:08   BIC                          13999.943
Sample:                             0   HQIC                         13991.259
                                - 812
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2813      0.393      0.715      0.474      -0.490       1.052
ma.L1         -0.3373      0.384     -0.879      0.379      -1.089       0.415
sigma2      1.803e+06   3.86e+04     46.733      0.000    1.73e+06    1.88e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):              2918.35
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):              35.53   Skew:                            -0.90
Prob(H) (two-sided):                  0.00   Kurtosis:                        12.12
===================================================================================
```

**Table 1: SARIMAX results of the initially selected model**

We see that the coefficient for the AR term is not significant, so we can try switching to Sarima(0,1,1)(0,0,0,0) model.



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   open   No. Observations:                  812
Model:               SARIMAX(0, 1, 1)   Log Likelihood               -6990.078
Date:                Wed, 18 May 2022   AIC                          13984.155
Time:                        10:54:08   BIC                          13993.552
Sample:                             0   HQIC                         13987.763
                                - 812
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.0556      0.031     -1.807      0.071      -0.116       0.005
sigma2      1.803e+06   3.86e+04     46.703      0.000    1.73e+06    1.88e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):              2900.57
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):              35.55   Skew:                            -0.89
Prob(H) (two-sided):                  0.00   Kurtosis:                        12.09
===================================================================================
```

**Table 2: SARIMAX results of the adjusted model**

Now we have significant coefficients.

Residual diagnostics will also be applied to be fully confident in the selected model:

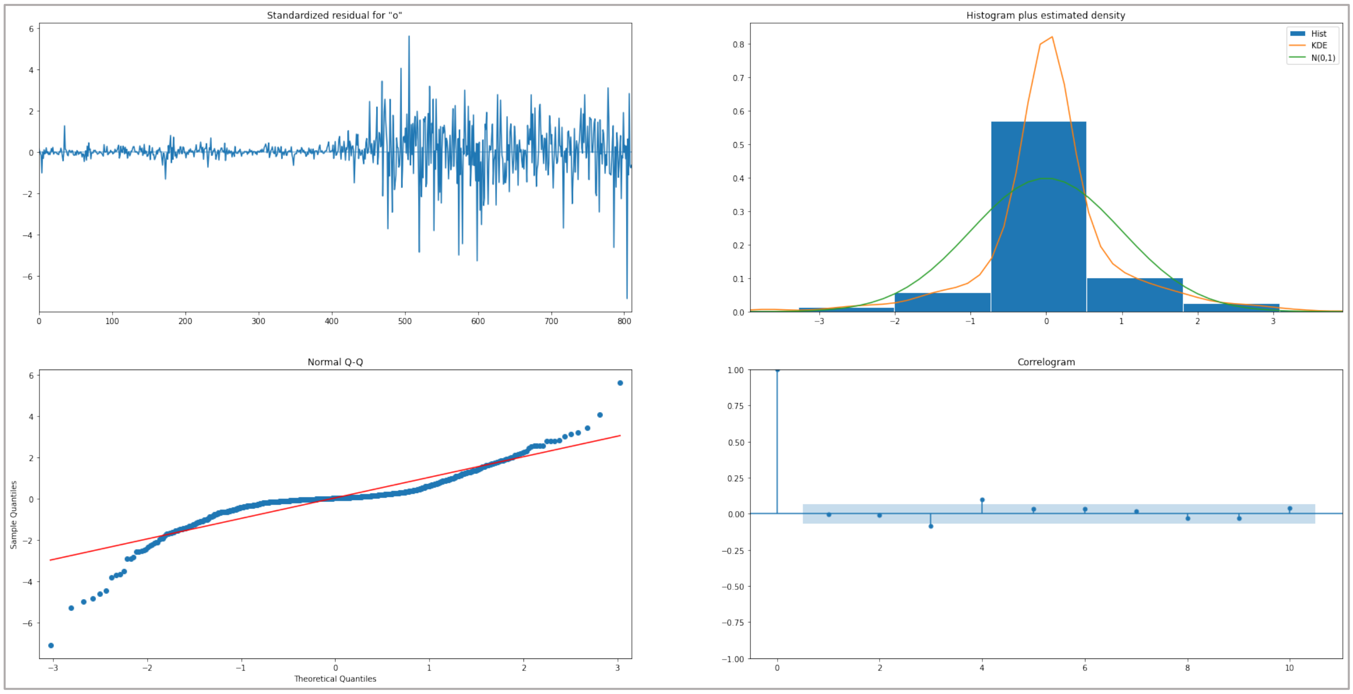


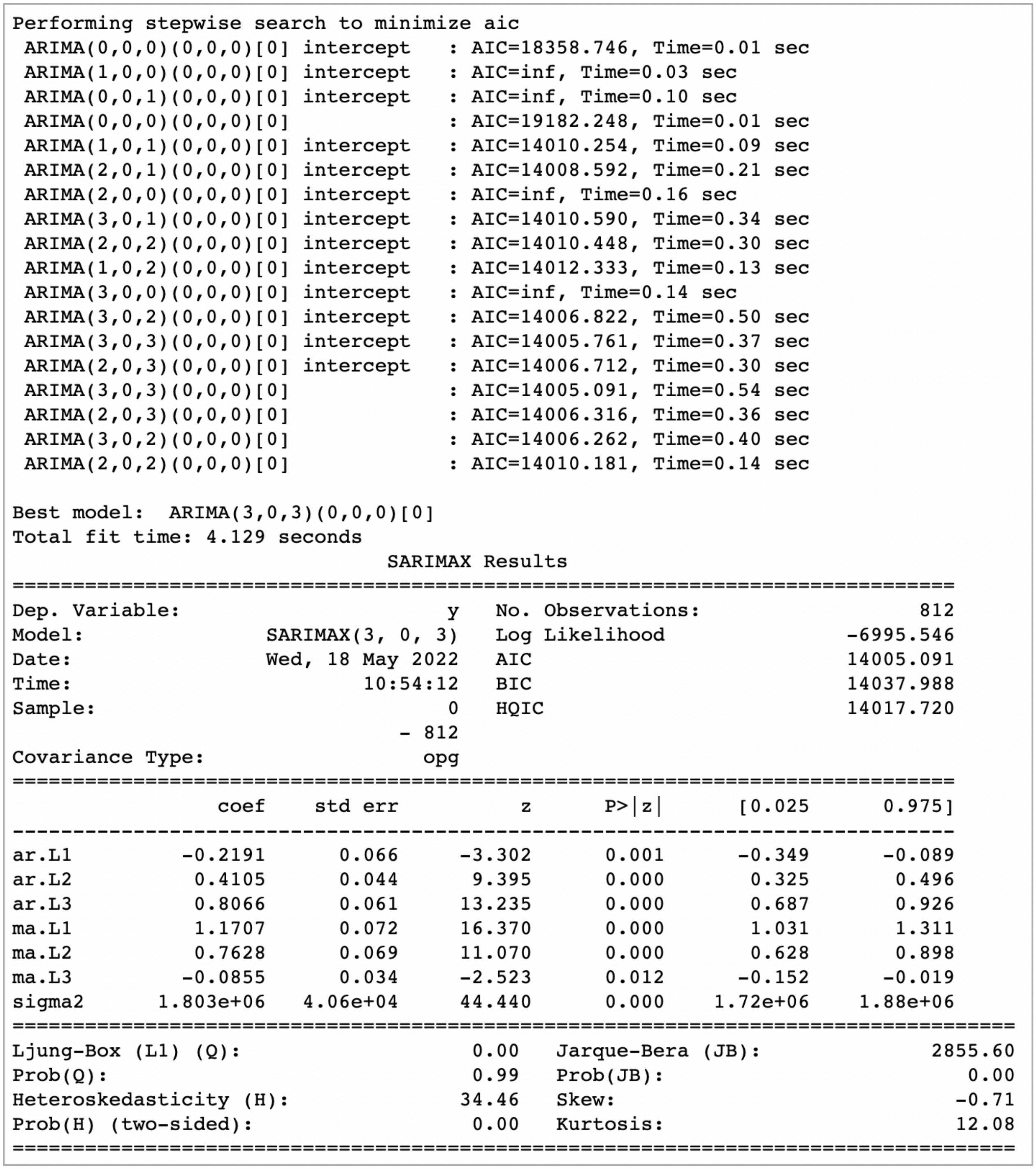
**Figure 4: Residual diagnostics**

From the first plot, we see that residuals have approximately 0 mean (though some refinements might be needed for the last part of them), then looking at the histogram, we can detect that they have an approximately normal distribution, looking at the QQ plot, we see that they are not very far from again indicating normal distribution, and from correlogram we see no significant lags (except the 4th one which is just a little high), which means the residuals are independent.

The p-value Ljung-Box test in the summary table is not greater than 0.05, which might be connected to non-seasonal data. The p-values of Heteroskedasticity and Jarque-Bera show the same thing.

To sum up the findings, we can say that the model's residuals are from white noise, meaning they are independently and identically distributed.

We also used auto. arima function to apply final validation of the selected model:



```
Performing stepwise search to minimize aic
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=18358.746, Time=0.01 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=inf, Time=0.03 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.10 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=19182.248, Time=0.01 sec
 ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=14010.254, Time=0.09 sec
 ARIMA(2,0,1)(0,0,0)[0] intercept   : AIC=14008.592, Time=0.21 sec
 ARIMA(2,0,0)(0,0,0)[0] intercept   : AIC=inf, Time=0.16 sec
 ARIMA(3,0,1)(0,0,0)[0] intercept   : AIC=14010.590, Time=0.34 sec
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=14010.448, Time=0.30 sec
 ARIMA(1,0,2)(0,0,0)[0] intercept   : AIC=14012.333, Time=0.13 sec
 ARIMA(3,0,0)(0,0,0)[0] intercept   : AIC=inf, Time=0.14 sec
 ARIMA(3,0,2)(0,0,0)[0] intercept   : AIC=14006.822, Time=0.50 sec
 ARIMA(3,0,3)(0,0,0)[0] intercept   : AIC=14005.761, Time=0.37 sec
 ARIMA(2,0,3)(0,0,0)[0] intercept   : AIC=14006.712, Time=0.30 sec
 ARIMA(3,0,3)(0,0,0)[0]             : AIC=14005.091, Time=0.54 sec
 ARIMA(2,0,3)(0,0,0)[0]             : AIC=14006.316, Time=0.36 sec
 ARIMA(3,0,2)(0,0,0)[0]             : AIC=14006.262, Time=0.40 sec
 ARIMA(2,0,2)(0,0,0)[0]             : AIC=14010.181, Time=0.14 sec

Best model:  ARIMA(3,0,3)(0,0,0)[0]
Total fit time: 4.129 seconds
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  812
Model:               SARIMAX(3, 0, 3)   Log Likelihood               -6995.546
Date:                Wed, 18 May 2022   AIC                          14005.091
Time:                        10:54:12   BIC                          14037.988
Sample:                             0   HQIC                         14017.720
                                - 812
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.2191      0.066     -3.302      0.001      -0.349      -0.089
ar.L2          0.4105      0.044      9.395      0.000       0.325       0.496
ar.L3          0.8066      0.061     13.235      0.000       0.687       0.926
ma.L1          1.1707      0.072     16.370      0.000       1.031       1.311
ma.L2          0.7628      0.069     11.070      0.000       0.628       0.898
ma.L3         -0.0855      0.034     -2.523      0.012      -0.152      -0.019
sigma2      1.803e+06   4.06e+04     44.440      0.000    1.72e+06    1.88e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):              2855.60
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):              34.46   Skew:                            -0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                        12.08
===================================================================================
```

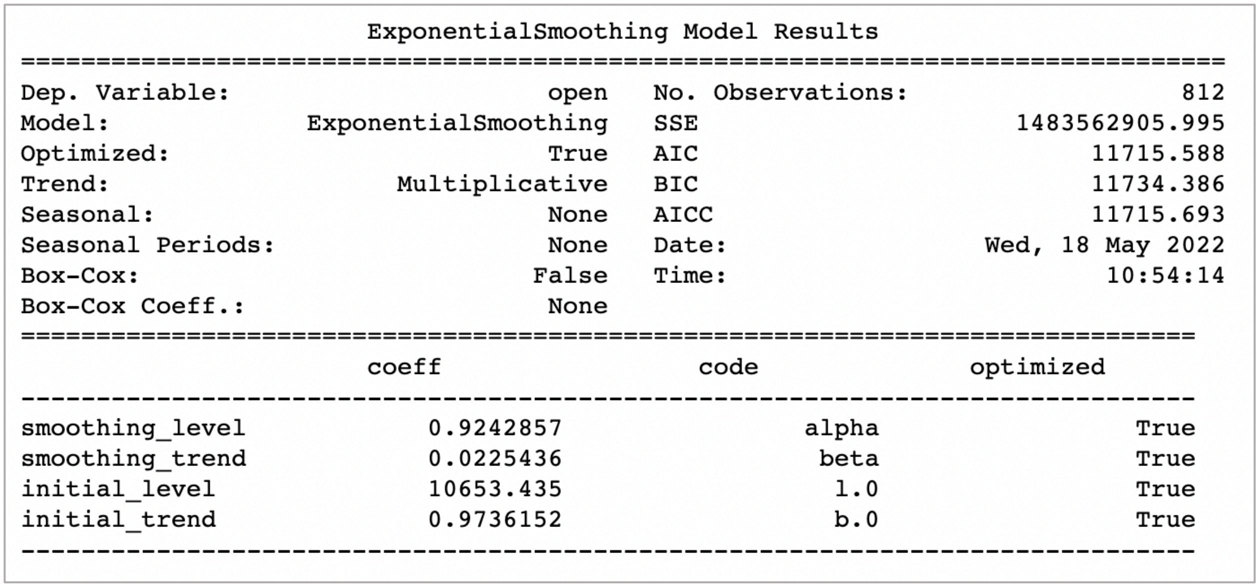
**Table 3: auto.arima results**

From the given results, we see that ARIMA(3,0,3)(0,0,0)[0]  is the best model based on AIC. The coefficients are also significant, and the results of residual diagnostic tests are ok (again, except Ljung-Box), so we can compare this model with our selection to determine the best one.

Now, as the final choice of the SARIMA model is made, the next step will be to apply Holt Winter's exponential smoothing and compare the forecasting results with those of the SARIMA.

## - Holt Winter

The first step will be the estimation of the train set:



```
                       ExponentialSmoothing Model Results
================================================================================
Dep. Variable:                     open   No. Observations:                  812
Model:             ExponentialSmoothing   SSE                     1483562905.995
Optimized:                         True   AIC                          11715.588
Trend:                   Multiplicative   BIC                          11734.386
Seasonal:                          None   AICC                         11715.693
Seasonal Periods:                  None   Date:                 Wed, 18 May 2022
Box-Cox:                          False   Time:                         10:54:14
Box-Cox Coeff.:                    None
==============================================================================
                       coeff                 code              optimized
------------------------------------------------------------------------------
smoothing_level            0.9242857                alpha                 True
smoothing_trend            0.0225436                 beta                 True
initial_level              10653.435                  l.0                 True
initial_trend              0.9736152                  b.0                 True
------------------------------------------------------------------------------
```

**Table 4: Holt-Winter's ETS results**

As the results are acceptable, forecasting will be applied to the two models:

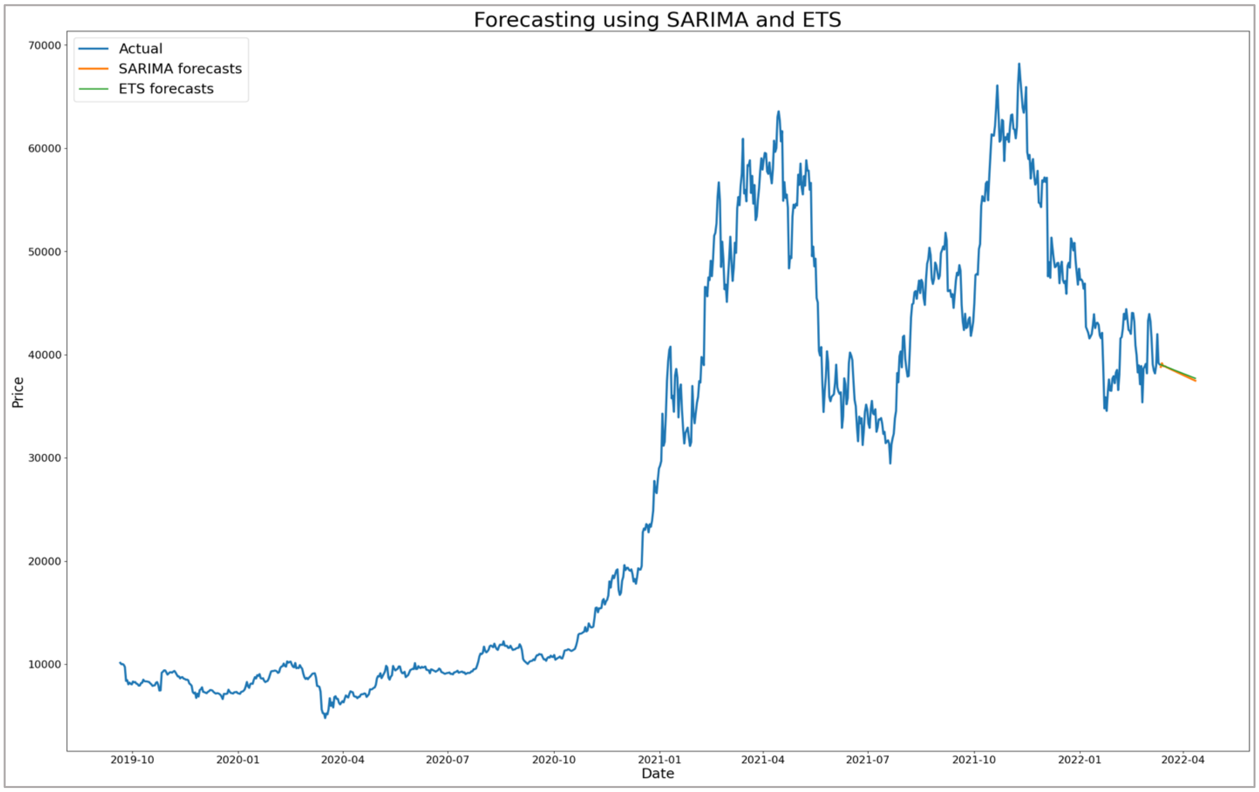


**Figure 5: Forecasting with SARIMA and Holt Winter's ETS**

Though ETS had much better MSE than ARIMA, their forecasts don't differ too much, and both are showing lines tending to go down without any volatility, which doesn't seem to be a correct forecast. That means that maybe the next step in the project will be breaking down the data into several parts and modeling them separately, as we can see that the values have other behavior for different periods. Thus, modeling them separately might get a better touch on seasonal components and have higher quality forecasts.

## - VAR

To create a VAR model, we will apply first-order differencing on the open value, volume, and the number of trades, after which we will check whether the resulting series is stationary or not:

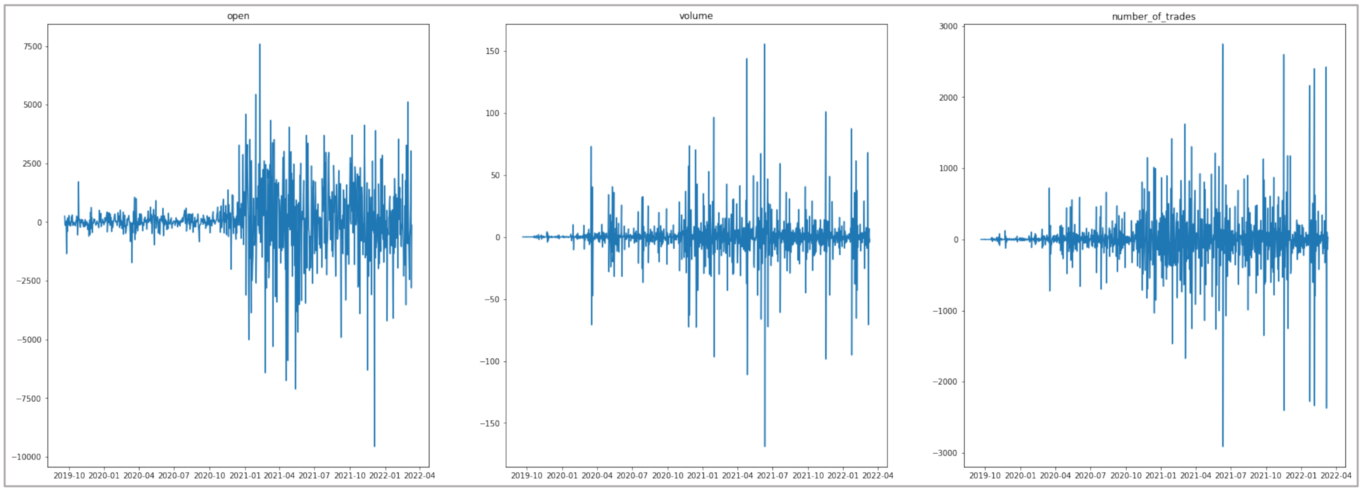
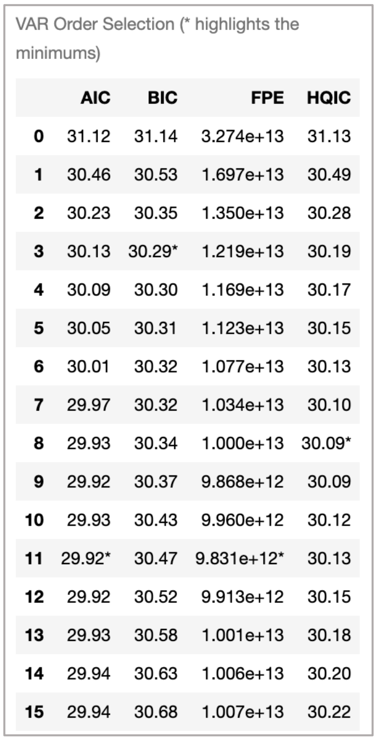


**Figure 6: Series of first-order differences of variables selected for VAR**

We can visually observe that all three series are stationary. Applying the ADF test on each, the results are again acceptable, as a p-value of each is less than 0.05.

The final step will be to decide the number of maximum lags for the VAR model:



VAR Order Selection (* highlights the minimums)

| | AIC | BIC | FPE | HQIC |
|---|---|---|---|---|
| 0 | 31.12 | 31.14 | 3.274e+13 | 31.13 |
| 1 | 30.46 | 30.53 | 1.697e+13 | 30.49 |
| 2 | 30.23 | 30.35 | 1.350e+13 | 30.28 |
| 3 | 30.13 | 30.29* | 1.219e+13 | 30.19 |
| 4 | 30.09 | 30.30 | 1.169e+13 | 30.17 |
| 5 | 30.05 | 30.31 | 1.123e+13 | 30.15 |
| 6 | 30.01 | 30.32 | 1.077e+13 | 30.13 |
| 7 | 29.97 | 30.32 | 1.034e+13 | 30.10 |
| 8 | 29.93 | 30.34 | 1.000e+13 | 30.09* |
| 9 | 29.92 | 30.37 | 9.868e+12 | 30.09 |
| 10 | 29.93 | 30.43 | 9.960e+12 | 30.12 |
| 11 | 29.92* | 30.47 | 9.831e+12* | 30.13 |
| 12 | 29.92 | 30.52 | 9.913e+12 | 30.15 |
| 13 | 29.93 | 30.58 | 1.001e+13 | 30.18 |
| 14 | 29.94 | 30.63 | 1.006e+13 | 30.20 |
| 15 | 29.94 | 30.68 | 1.007e+13 | 30.22 |

**Table 5: Max lag selection for the VAR model**

Gaining the required information, we get the following results after applying the VAR model:

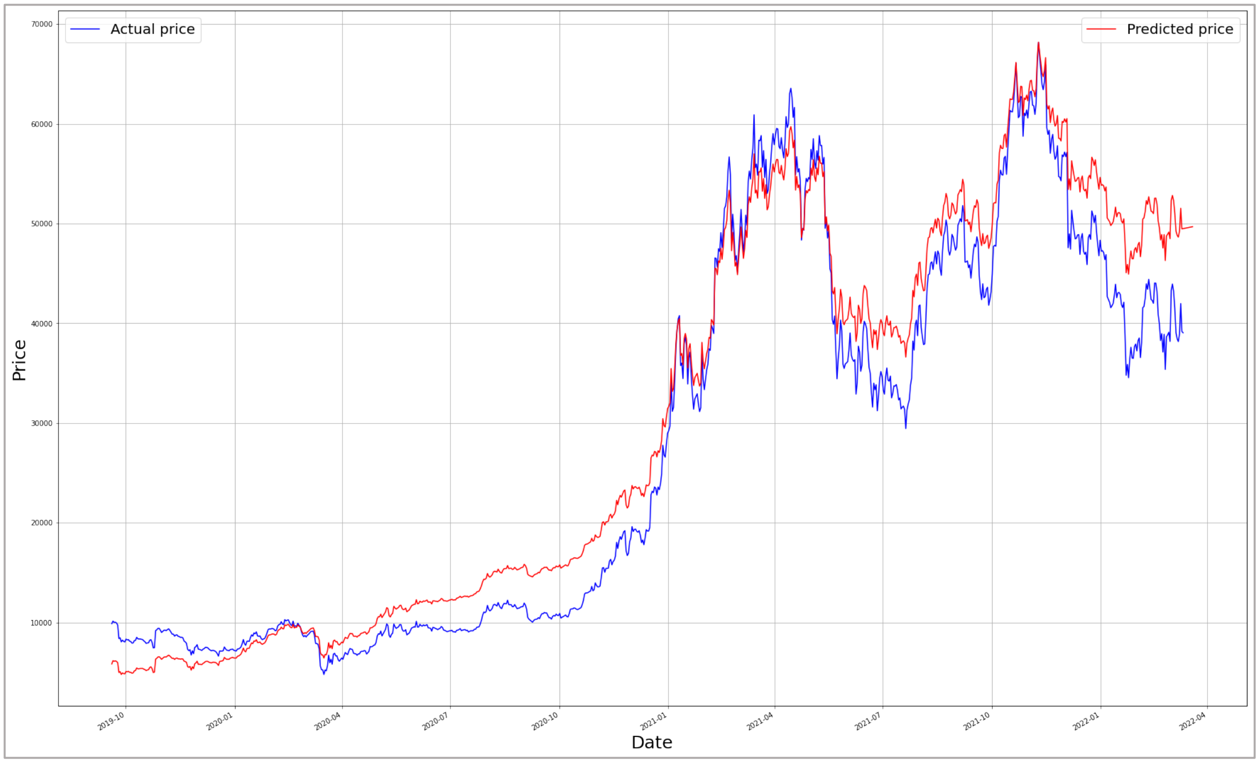


**Figure 7: Forecasting with VAR model**

We observe that the results are pretty good when using VAR. Thus, this model will be our final choice.

# Conclusion

An accurate forecast of any cryptocurrency is pretty tough as the market has only started gaining power a couple of years. There are no well-known tendencies in the market.

And that situation is confirmed when we try to forecast the future values using SARIMA or Holt-Winter's exponential smoothing. Those two methods don't provide any reliable results and don't have real value.

In contrast, when we apply VAR, the results are excellent, and even though we don't have an exact matching of the values, they are very close, and most importantly, the trend is correctly replicated; thus, it can be used for forecasting and getting the real deal from Bitcoin price changes.

# References

[1] Binance Full History
*https://www.kaggle.com/datasets/jorijnsmit/binance-full-history*

[2] Forecasting of Cryptocurrency Prices Using Machine Learning
*https://link.springer.com/chapter/10.1007/978-981-15-4498-9_12*

[3] A New Forecasting Framework for Bitcoin Price with LSTM
*https://ieeexplore.ieee.org/abstract/document/8637486*

[4] An Advanced CNN-LSTM Model for Cryptocurrency Forecasting
*https://www.mdpi.com/2079-9292/10/3/287*