| FULL LEGAL NAME | LOCATION (COUNTRY) | EMAIL ADDRESS | MARK X FOR ANY NON-CONTRIBUTING MEMBER |
|---|---|---|---|
| Jefferson Bien-Aimé | USA | jefferson.bienaime@gmail.com | |
| Marin Stoyanov | Bulgaria | azonealerts@gmx.com | |
| | | | |


| **Statement of integrity:** By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above). | |
|---|---|
| **Team member 1** | Jefferson Bien-Aimé |
| **Team member 2** | Marin Stoyanov |
| **Team member 3** | |

| Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.<br>**Note:** You may be required to provide proof of your outreach to non-contributing members upon request. |
|---|
| |

## Definition

In the realm of statistics and data science, a time series is considered non-stationary if its statistical properties change over time. It means that the values of the time series do not fluctuate around a constant mean or with a constant variance (Hyndman & Athanasopoulos, 2018). For non-stationarity time-series, the mean, the variance, or the covariance across different time intervals varies. We can formalized it as follow :

Let Y be a given time series, and $Y_t$ denote the value of the series at time t. The series is non-stationary if there exists any time point t, such that $E(Y_t) \neq \mu$ or $Var(Y_t) \neq \sigma^2$, where $E(Y_t)$ denotes the expected value (mean) of $Y_t$, $Var(Y_t)$ is the variance of $Y_t$, μ represents the series' mean, and $\sigma^2$ is the series' variance (Hyndman & Athanasopoulos, 2018).

David Hitchcock from the university of South Carolina put together this equation

$$Y_t = \mu_t + X_t$$

where :

- $Y_t$ is the non-stationary time series at time t.
- $\mu_t$ is the mean of the time series at time t, which is a function of time. Sometimes, it may represent some deterministic trend.
- $X_t$ is a stationary time series with mean zero. (Hitchock, 2018)

One common type of non-stationarity is a unit root non-stationarity, which can be expressed as:

$$X_t = X_{t-1} + e_t$$

where:

- $X_t$ is the time series at time t,
- $X_{t-1}$ is the value of the series at the previous time step, and
- $e_t$ is a random error term.

## Description

In essence, non-stationarity describes a scenario where key statistical properties of a dataset—such as the mean, variance, or covariance—change over time. Some examples in the real-world financial data, includes stock, interest rates, treasury bills, foreign exchange, GDP. They are influenced by factors like economic shifts, investor sentiment, and corporate events, government intervention.

There are several forms of non-stationarity time series. The most common are the following :

1. **Random Walk** : It is given by the following equation ($Y_t = Y_{t-1} + e_t$). It predicts that the value at time "t" will be equal to the last period value plus a stochastic (non-systematic) component that is a white noise, which means εt is independent and identically distributed with mean "0" and variance "$\sigma^2$."
2. **Random Walk with Drift** : It is given by the following equation ($Y_t = \alpha + Y_{t-1} + e_t$). It will be the same equation as the random walk but plus a constant, or drift (α).
3. **Deterministic Trend** : It is given by the following equation ($Y_t = \alpha + \beta_t + e_t$). The value at time "t" is regressed on a time trend (βt) with a drift component (α) and a white noise, which means εt is independent and identically distributed with mean "0" and variance "$\sigma^2$."
4. **Random Walk with Drift and Trend**: It is given by the following equation ($Y_t = \alpha + Y_{t-1} + \beta_t + e_t$). It is the combination of a random walk with a drift component (α) and a deterministic trend ($\beta_t$). It specifies the value at time "t" by the last period's value, a drift, a trend, and a stochastic component. (Iordanova, 2022)

There are several others non-stationarity process for example :

- Seasonal Non-stationary Process where the series has predictable and repeated patterns over certain time periods. For example, stocks for utility companies, sales for companies during holiday.
- Structural Break where the series has experienced a one-time change in the parameters of the process. For example, we can talk about the financial crisis of 2008 for stocks, COVID-19 for interests rates.

## Demonstration

For our analysis, we selected the stocks data for (SNEX) stock from January 1, 2017, to June 30, 2023. We will also use the bitcoin BITCOIN/USD and Nasdaq100 with the same time interval.

We prepared the data and performed the Augmented Dickey-Fuller (ADF) test—a standard statistical test for non-stationarity—using the Python programming language, alongside pandas, numpy, and statsmodels libraries (McKinney, 2010; Seabold & Perktold, 2010). We will be also using some other test such as The Phillips-Perron and the KPSS — Kwiatkowski–Phillips–Schmidt–Shin which is also considered to be resilient to autocorrelation and heteroskedasticity.

We will consider the following :
$H_0$ : The time series is non-stationary
$H_1$ : The time series is stationary

The full demonstration is in the following link :

https://colab.research.google.com/drive/10s5ttWaMIsng7-ZNisVPcwvCzsn07ToB

The output from the ADF test, the KPSS and Phillip-Perron(PP) tests for SNEX Stocks are as followed :

| Results of ADF Test: | | Results of KPSS Test: | |
|---|---|---|---|
| Test Statistic | -1.050024 | Test Statistic | -1.050024 |
| P-value | 0.734585 | P-value | 0.734585 |
| #Lags Used | 10.000000 | #Lags Used | 10.000000 |
| Number of Observations Used | 1622.000000 | Number of Observations Used | 1622.000000 |
| Critical Value (1%) | -3.434388 | Critical Value (1%) | -3.434388 |
| Critical Value (5%) | -2.863324 | Critical Value (5%) | -2.863324 |
| Critical Value (10%) | -2.567720 | Critical Value (10%) | -2.567720 |

Table 1 : ADF and KPSS Test

```
Results of Phillips-Perron Test:
    Phillips-Perron Test (Z-tau)
=====================================
Test Statistic              -1.035
P-value                      0.740
Lags                         25
-------------------------------------

Trend: Constant
Critical Values: -3.43 (1%), -2.86 (5%), -2.57 (10%)
```

Table 2 : PP Test

Table 1 and 2 returned the statistics of –1.05, -1,05 and -1,035 and a critical value at 5% of -2.86, -2.86 and -2.86 respectively for ADF, KPSS and PP test. All test Statistics are negative and the critical value at 5%, being below the standard 0.05 significance level, allows us to accept the null hypothesis of the series being non-stationary. Furthermore, the series also contains a unit root test.

## Diagram

Visual exploration is a crucial step in preliminary data analysis. We used matplotlib to graph the original time series data and its rolling mean and standard deviation for a visual check on non-stationarity (Hunter, 2007). We have also plotted not only the rolling mean for 20 days.

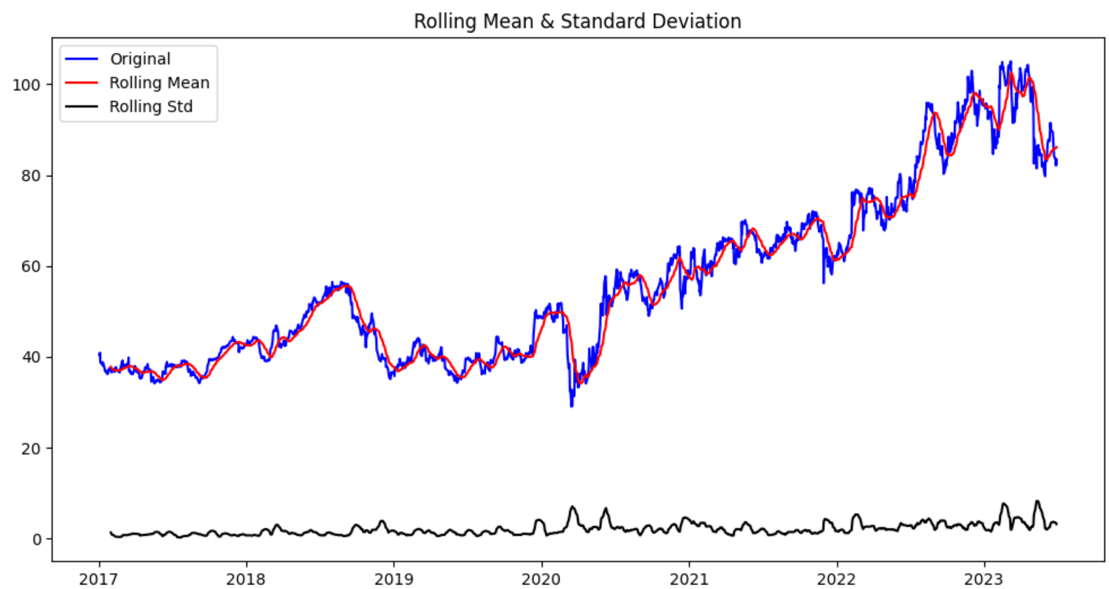



Figure 1: Rolling mean and standard deviation 20 days

The plots revealed that both the rolling mean and standard deviation varied over time, supporting the non-stationarity hypothesis.

## Diagnosis

As a further diagnostic tool, we generated the following Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

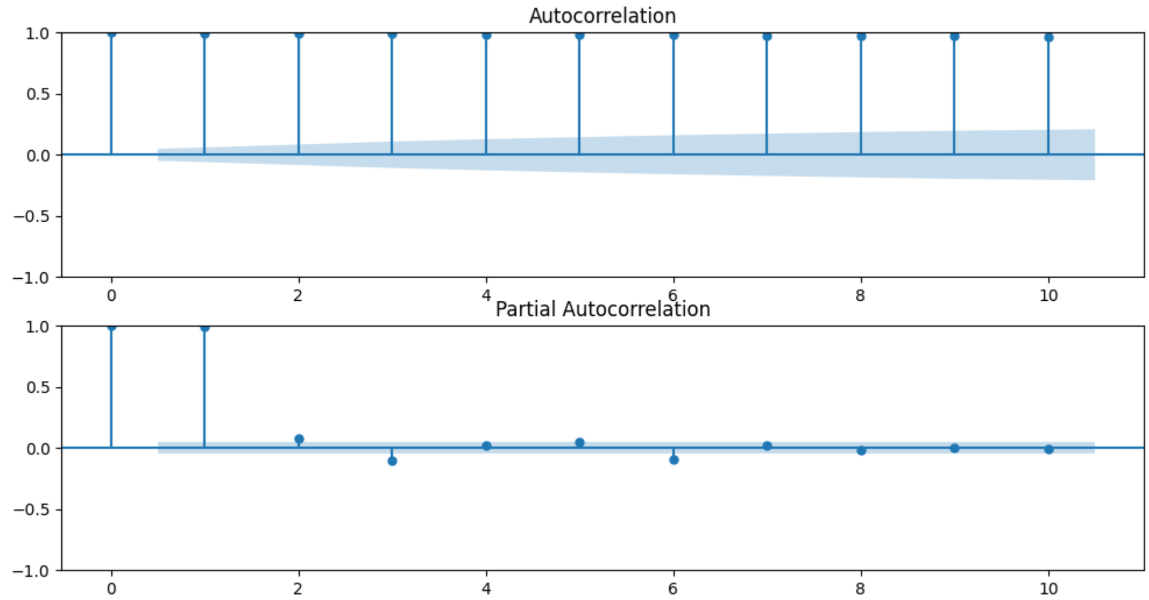



Figure 2 : ACF and PACF

The ACF plot showed significant autocorrelation at several lags, further reinforcing the assumption of non-stationarity.

## Damage

Non-stationarity presents various problems when it comes to modeling time series data. Many traditional forecasting methods—like Autoregressive (AR), Moving Average (MA), and ARMA models—assume stationarity. When this assumption is violated, these models can generate misleading or inaccurate predictions (Hyndman & Athanasopoulos, 2018).

The high variability exhibited in the rolling mean and standard deviation indicates the presence of high volatility in the data.

## Directions

Possible approaches to address non-stationarity include differencing the series or applying transformations like logarithmic or square root transformations. These strategies aim to stabilize the series' mean (Box, Jenkins, Reinsel, & Ljung, 2015).

Alternatively, we could consider using models designed to handle non-stationarity, such as Autoregressive Integrated Moving Average (ARIMA) or Seasonal ARIMA (SARIMA) models.

## Deployment

Upon addressing the non-stationarity or implementing a model capable of handling such data, we can utilize the model for various purposes, such as predicting future stock returns or bitcoin assets, assessing risk, or optimizing investment portfolios.

For example, an investor or fund manager could employ the model to project future returns, basing their portfolio adjustments on the expected future performance. The model would serve as a key tool for decision-making in buying, selling, or holding assets.

Finally, it's important to note that addressing non-stationarity is only one aspect of the modeling process. Real-world factors such as market volatility, sudden economic changes, public sentiments, or unexpected events could impact actual assets returns and must be considered alongside model forecasts in any investment strategy.

**BONUS**

The first part with the unit roots is analogous so we will skip it and then we will concentrate over the new things that we have learned in this week's lectures.

**Definition**

- **Dickey Fuller Test -** this is the base for the Augmented Dickey Fuller Test and it can be described like this:

$$x_t = \boldsymbol{\alpha} x_{t-1} + \upsilon_t \qquad \text{where } \upsilon_t \text{ is white noise}$$

And after some mathematical manipulations it can be expressed like this:

$$\triangledown x_t = \gamma x_t + \upsilon_t \qquad \text{where } \gamma = (\boldsymbol{\alpha} - 1)$$

Dickey Fuller test can be set up like this:

$H_0 : \gamma = 0$ there is a unit root

$H_1 : \gamma \neq 0$ the time series is stationary

These are the 3 main versions of Dickey Fuller tests:

1. $\triangledown x_t = \gamma x_t + \upsilon_t$ This test can be written like this:

$H_0 : \gamma = 0$ there is a unit root

$H_1 : \gamma \neq 0$ the time series is stationary

**2.** $\nabla x_t = \gamma x_{t-1} + \alpha_0 + \upsilon_t$

a) $(\phi_1)$ : $H_0 : \gamma = 0,\ \alpha_0 = 0$

$H_1 : H_0\ is\ not\ true$

b) $(\tau_2)$ : $H_0 : \gamma = 0$

$H_1 : H_0\ is\ not\ true$

**3.** $\nabla x_t = \gamma x_{t-1} + \alpha_0 + \alpha_2 t + \upsilon_t$

a) $(\phi_2)$ : $H_0 : \gamma = 0,\ \alpha_0 = 0,\ \alpha_2 = 0$

$H_1 : H_0\ is\ not\ true$

b) $(\phi_3)$ : $H_0 : \gamma = 0,\ \alpha_0 = 0$

$H_1 : H_0\ is\ not\ true$

c) $(\tau_3)$ : $H_0 : \gamma = 0$

$H_1 : H_0\ is\ not\ true$

- Augmented Dickey Fuller test

1. $\nabla x_t = \gamma x_{t-1} + \sum_{t=2}^{p} \beta_i \nabla x_{t-i+1} + \upsilon_t$

$(\tau_1)$ : $H_0 : \gamma = 0$

**2.** $\nabla x_t = \gamma x_{t-1} + \alpha_0 + \sum_{t=2}^{p} \beta_i \nabla x_{t-i+1} + \upsilon_t$

$(\phi_1)$ : $H_0 : \gamma = 0,\ \alpha_0 = 0$

$(\tau_2)$ : $H_0 : \gamma = 0$

**3.** $\nabla x_t = \gamma x_{t-1} + \alpha_0 + \alpha_2 t + \sum_{t=2}^{p} \beta_i \nabla x_{t-i+1} + \upsilon_t$

$(\phi_2)$ : $H_0 : \gamma = 0, \alpha_0 = 0, \alpha_2 = 0$

$(\phi_3)$ : $H_0 : \gamma = 0, \alpha_0 = 0$

$(\tau_3)$ : $H_0 : \gamma = 0$

- Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

For short this test is called KPSS and is used as a complement to unit root tests. This test shows is if there is a time trend.

$H_0$ : *the time series is trend stationary*

$H_1$ : *the time series is not trend stationary*

Usually both ADF and KPSS tests are used together to better understand the time series behavior.

a) If the ADF test cannot reject $H_0$ and the KPSS test rejects $H_0$ , then the time series is non-stationary
b) If the ADF test rejects $H_0$ and the KPSS test cannot reject $H_0$ , then the time series is trend stationary

- **Vector Autoregressive Model (VAR)**

This is a time series model that is used to forecast two or more time series. The formula for a 2 time series with a lag 2 VAR model looks like this:

$$x_t = \alpha_0 + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \beta_1 y_{t-1} + \beta_2 y_{t-2} + e_t$$

$$y_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t$$

Where $e_t$ ~ white noise$(0, \sigma_e^2)$ and $\varepsilon_t$ ~ white noise$(0, \sigma_\varepsilon^2)$ and they are not autocorrelated

The VAR model has the following key features:

a) It is not a simultaneous model – this means that these equations can be solved separately and not at the same time.
b) $e_t$ and $\varepsilon_t$ are contemporaneously correlated – this means that they can be correlated and can be affected by exogenous factors
c) $x_t$ and $y_t$ are stationary and ergodic
d) Each equation can be estimated by OLS estimation

- 

**Description**

We also chose Bitcoin and Nasdaq because there is a correlation between them and this can be used for further forecasting.

**Demonstration**

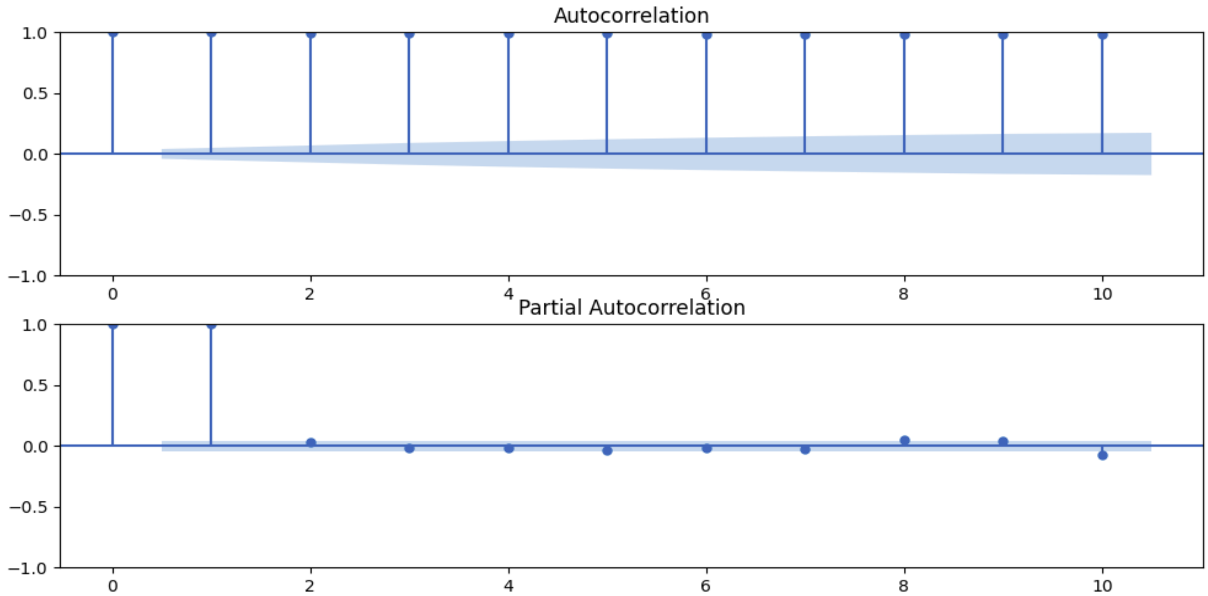
The part with the unit root test is analogous so we will skip it.After implementing the unit root tests we can see that the data should be differenced with 1 lag and then we can implement the VAR model. Also because the prices are very far away from each other, so additional normalization was done by applying the logarithm function. Also there was something else that is important for proper modeling. Bitcoin daily prices are collected every day while Nasdaq daily prices are collected only during the weekdays so a proper filtering was done and the prices of Bitcoin during the weekends were dropped, so that the 2 datasets have the same length.

A cointegration test was run before and after differencing and logging the data and the p-value before lagging the data was 0.11 so there is no cointegration, but after logging the data the p-value dropped below 0.05 so after these data manipulations there is cointegration after all.

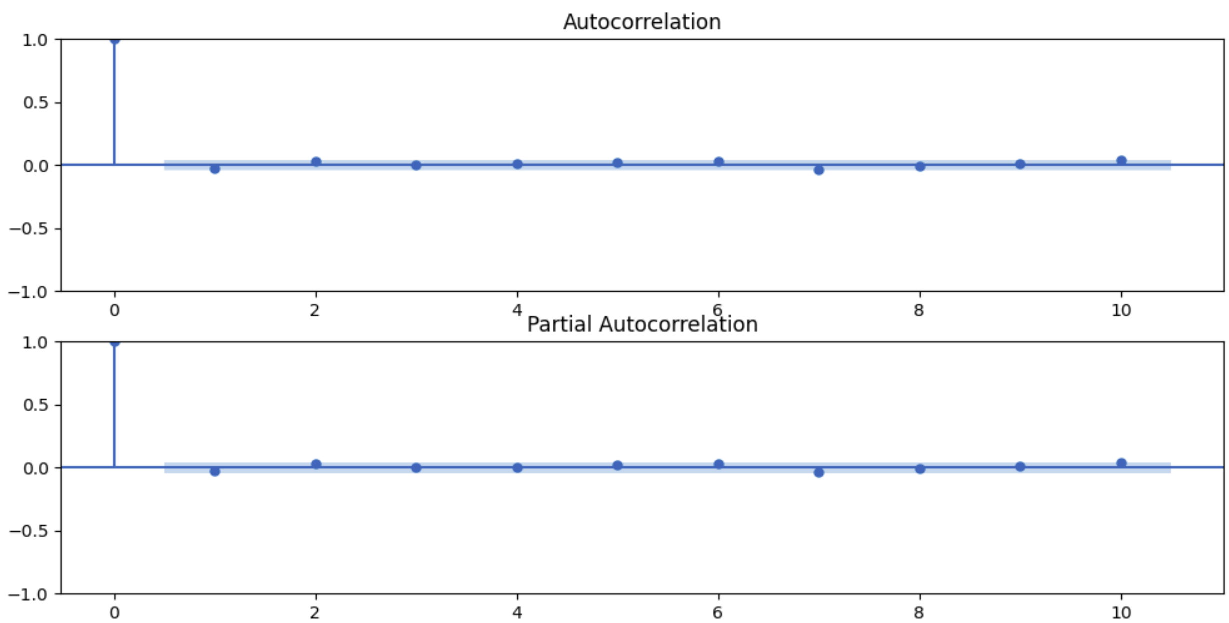
After all these data preparations a VAR(1) model was run on both the pure closing price data and the transformed data and then according forecast was made.

**Diagram and Diagnossis**

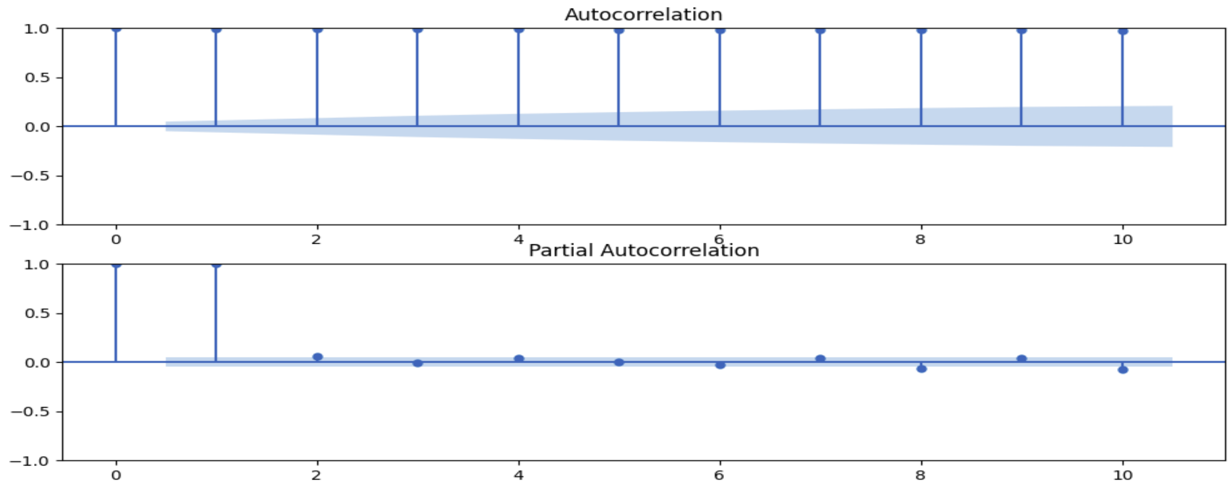
ACF and PACF plot of Bitcoin closing price with no differencing:



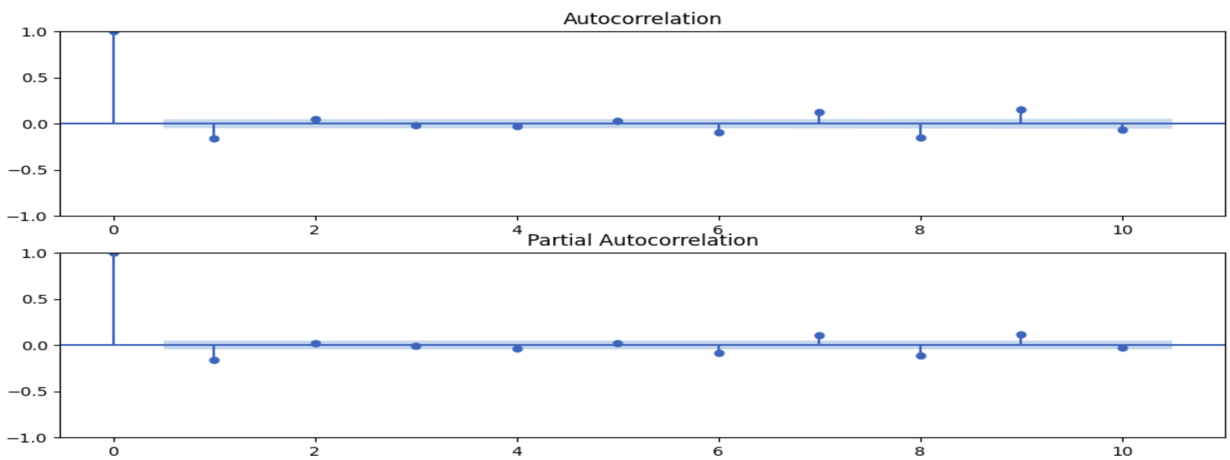

And after differencing:

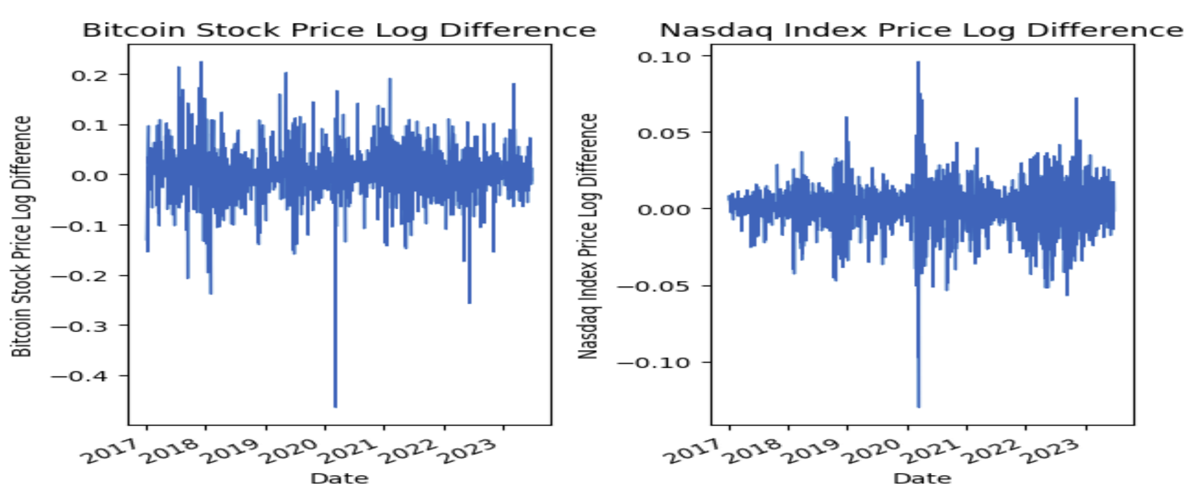



ACF and PACF plot of Nasdaq closing price with no differencing:




And after differencing:

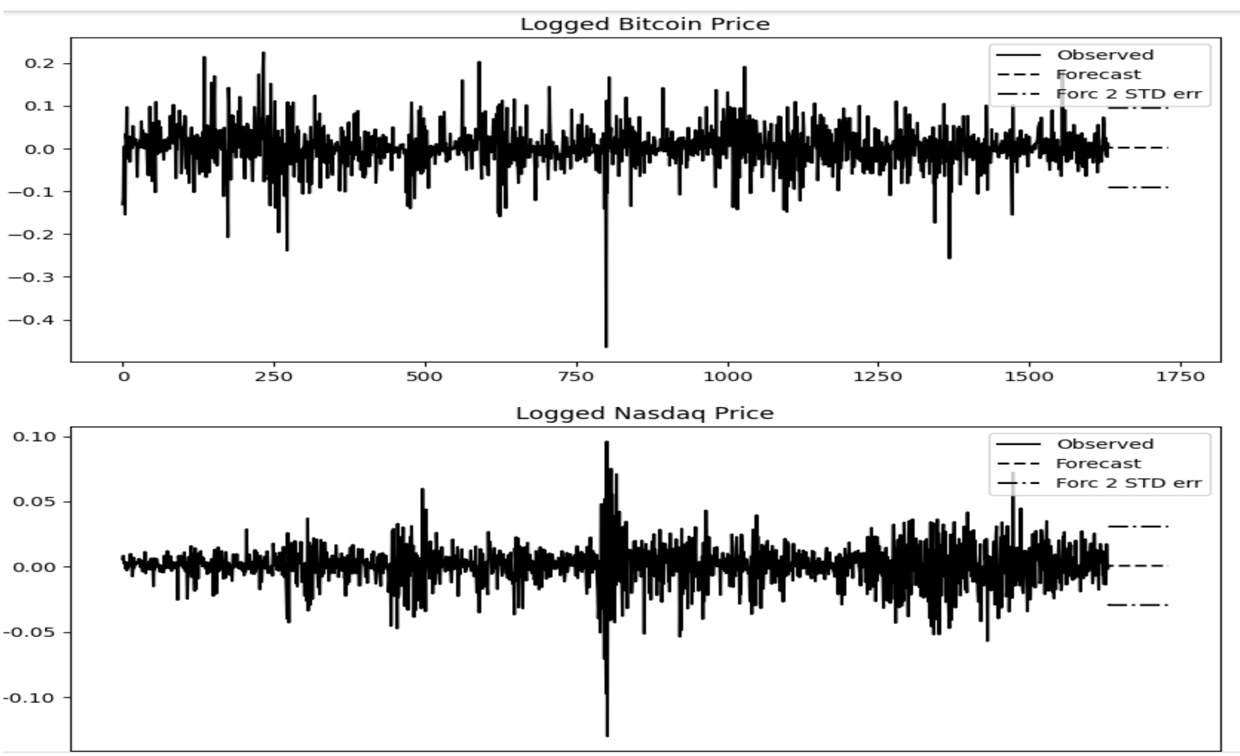


```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Wed, 05, Jul, 2023
Time:                     14:41:21
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                   -14.5261
Nobs:                     1630.00    HQIC:                  -14.5386
Log likelihood:           7235.21    FPE:                4.81695e-07
AIC:                     -14.5460    Det(Omega_mle):     4.79927e-07
--------------------------------------------------------------------
Results for equation Logged Bitcoin Price
=========================================================================================
                             coefficient       std. error           t-stat            prob
-----------------------------------------------------------------------------------------
const                           0.002148         0.001176            1.827           0.068
L1.Logged Bitcoin Price        -0.009371         0.025621           -0.366           0.715
L1.Logged Nasdaq Price         -0.062026         0.079370           -0.781           0.435
=========================================================================================

Results for equation Logged Nasdaq Price
=========================================================================================
                             coefficient       std. error           t-stat            prob
-----------------------------------------------------------------------------------------
const                           0.000789         0.000376            2.101           0.036
L1.Logged Bitcoin Price        -0.002848         0.008184           -0.348           0.728
L1.Logged Nasdaq Price         -0.158323         0.025354           -6.244           0.000
=========================================================================================

Correlation matrix of residuals
                        Logged Bitcoin Price  Logged Nasdaq Price
Logged Bitcoin Price                1.000000             0.263011
Logged Nasdaq Price                 0.263011             1.000000
```

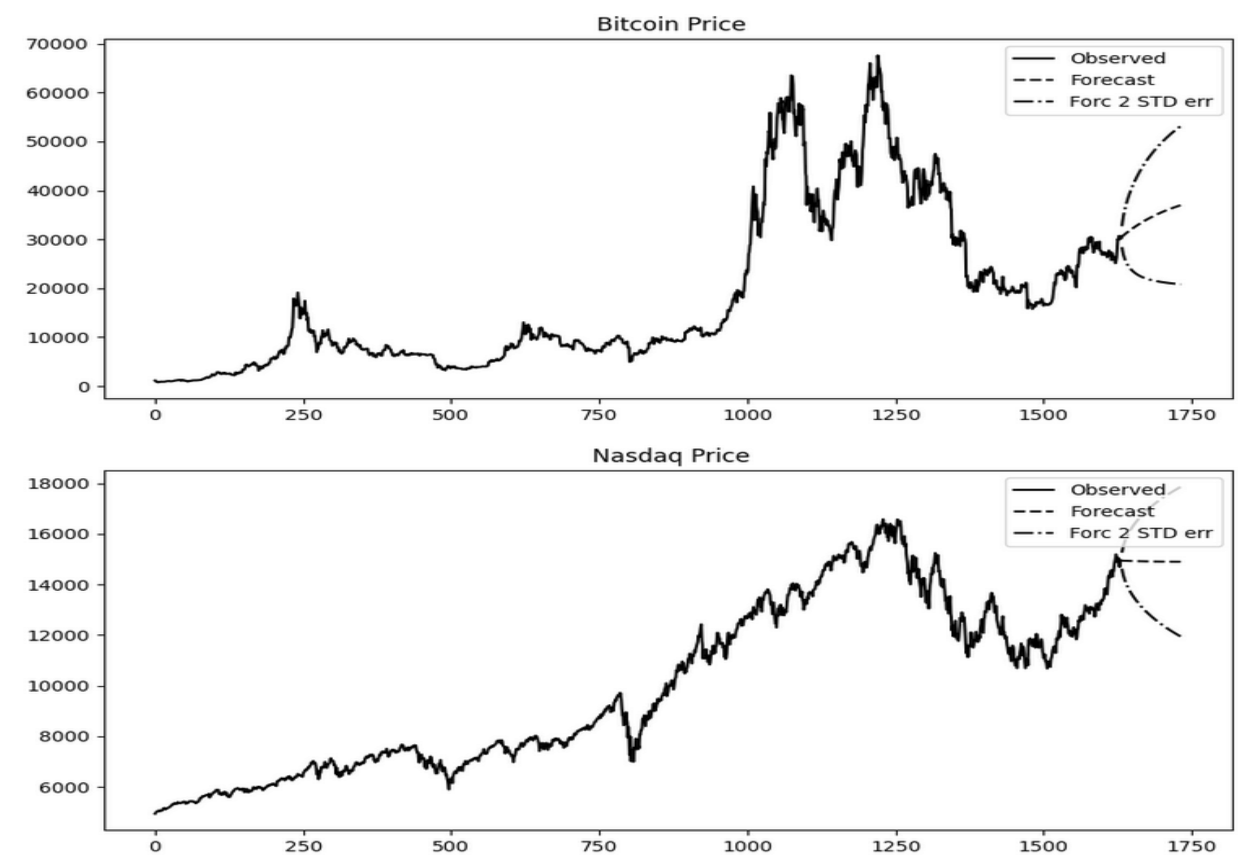


```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Wed, 05, Jul, 2023
Time:                     14:41:22
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                    24.0714
Nobs:                     1631.00    HQIC:                   24.0589
Log likelihood:          -24236.6    FPE:                2.78910e+10
AIC:                      24.0516    Det(Omega_mle):     2.77887e+10
--------------------------------------------------------------------
Results for equation  Bitcoin Price
======================================================================================
                      coefficient       std. error           t-stat              prob
--------------------------------------------------------------------------------------
const                 -202.293961       121.823723           -1.661             0.097
L1. Bitcoin Price        0.990447         0.003528          280.764             0.000
L1. Nasdaq Price         0.039843         0.017353            2.296             0.022
======================================================================================

Results for equation  Nasdaq Price
======================================================================================
                      coefficient       std. error           t-stat              prob
--------------------------------------------------------------------------------------
const                   20.759401        18.073111            1.149             0.251
L1. Bitcoin Price        0.000126         0.000523            0.240             0.810
L1. Nasdaq Price         0.998291         0.002574          387.781             0.000
======================================================================================

Correlation matrix of residuals
                   Bitcoin Price   Nasdaq Price
 Bitcoin Price          1.000000       0.299015
 Nasdaq Price           0.299015       1.000000
```

A VAR(1) model for the NOT transformed closing prices of Bitcoin and Nasdaq



Forecast of the real (NOT transformed) closing prices of Bitcoin and Nasdaq

**Damage**

A VAR model can be used to analyze the dynamic relationships among multiple time series variables. While VAR models can be useful for understanding the relationships among variables, there are several challenges and problems that can arise when using them.

- the potential for omitted variable bias. This problem can arise because in order to have the same length of the 2 datasets we dropped some of the bitcoin values and in case they were important this will bias the estimated relationships among the included values.
- Non-stationarity. This can lead to spurious relationships among variables and then misleading results and conclusions
- The residuals should be normally distributed, otherwise it may indicate that the model is misspecified or that there are issues with the data. Non-normality of the residuals can also affect the validity of hypothesis tests and confidence intervals based on the model.
- Nonlinearities in the residuals of a VAR model can indicate that the model is not correctly set or that there are nonlinear relationships among the variables that are not being captured. This can lead to biased and inconsistent parameter estimates, and can lead to wrong confidence intervals and conclusions.

**Directions**

Data preparation and exploratory data analysis are always recommended. We can cite an old but very important saying: “Garbage in, garbage out”, which means that without proper data preparation, manipulations and cleaning we cannot rely on the model’s outputs. About removing outliers, it is difficult to say because they may be important and when removed they can totally distort the model and it results. So before removing them we should make additional measures and run the model with and without them so that we would see the difference and then take a decision whether to remove them or not.

**Deployment**

We would use the model as a guidance by looking at the forecast that it brings and this is how we would know where approximately the price should fluctuate (+/- a standard deviation or 2). This is just an approximation and nothing is for 100% sure. Thanks to these predictions and forecasts we can create an estimation of potential entry points as well as stop loss and take profit levels.

## References

1. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
2. Iordanova, T. (2022, January). An Introduction to Non-Stationary Processes. Retrieved from Investopedia: https://www.investopedia.com/articles/trading/07/stationary.asp
3. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.
4. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
5. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.
6. Seabold, S., & Perktold, J. (2010, June). Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference (Vol. 57, pp. 61-67).