Time Series Analysis for Causal Relationship among Top Cryptocurrencies

---

MTH517A: Time Series Analysis course project

Abir Naha(201257)

Arkonil Dhar(201279)

Koyel Pramanick(201333)

Suchismita Roy(201440)

August 2021 - November 2021

Acknowledgement

Our journey of accomplishing this project really involves many ones to whom we are highly obliged. We would like to express our deepest appreciation to all those who have provided us the possibility to complete this project. We give a special gratitude to our respected instructor Dr. Amit Mitra, Department of Mathematics and Statistics, IIT KANPUR, whose contribution in stimulating suggestion, valuable guidance, constructive criticism and encouragement helped us to coordinate our project.

We take the privilege to thank the authors and publishers of the various books we have consulted. Also thanks to the Wikipedia and various other free website from which we got help. At last we would like to thank our seniors and batch mates for their co-operation throughout the project. Without their guidance and supervision this project would not have been completed.

Abstract

When we have more than one correlated variables of interest, in many situations we are interested to check whether one variable is the cause of the other variable or not. Here We have a data on the daily prices of top Cryptocurrencies-Bitcoin, Ethereum and Binance coin. Our main aim is to check if the price of Bitcoin granger causes the price of Ethereum or Binance Coin. Firstly, we have done Augmented Dickey-Fuller Test for checking stationarity. Then we have used Engle-Granger Test and Johansen's Procedure to check whether the variables are cointegrated or not. Lastly, we have used Granger Causality Test to test whether the fluctuation of price of the Bitcoin is the cause of the fluctuation of price of other cryptocurrencies or not.

Contents

1 About Dataset **7**

2 Detection of Non-stationarity **7**

2.1 Augmented Dickey-Fuller Test 7

3 Detection of Cointegration **8**

3.1 Engle-Granger Method 8

3.1.1 Pre-test each variable 8

3.1.2 Limitations 9

3.2 Johansen's Procedure 9

3.2.1 The trace Test 10

3.2.2 The Maximum Eigenvalue Test 10

3.2.3 Limitations 11

4 Detection of Causal Relationship **11**

4.1 Granger Causality Test 12

5 Results and Interpretations **13**

5.1 Results from Stationarity Test 13

5.2 Results from Cointegration Test 15

5.2.1 Engle–Granger Test 15

5.2.2 Johansen's test: 15

5.3 Results from Granger Causality Test 15

Introduction

We are all familiar with the physical mode of currency which is the medium of paying price on exchange of some goods. But with the rapid Digitization over the past decade, online payment methods have became more convenient to use and parallelly decentralised currencies have also become popular which don't need a medium like banks to settle a transaction. So from where all this things began?

In the old barter system goods used to be exchanged against goods. But finding the exact match of goods was a problem. So to solve this problem, gold coins started coming into place where gold coins were exchanged for the goods that were being bought. But as the trade increased, carrying huge quantity of gold was a problem. Now, to solve this problem, governments agreed to issue currency notes whose value was based on gold. So instead of gold coins people could trade using paper.

But when the world war had started, the countries wanted full power to print as much money as they wanted to pay for their war costs. And that's when they decided to cut their ties to the gold standard and to solve this, currencies were pegged or attached to US dollars, and US dollar was in turn pecked to gold at a fixed rate.

Then came in the concept of Fiat money, a system which has been used globally with variable exchange rates between two currencies. Variable exchange rates basically denotes that the currency rates may change based on demand and supply. But the problem with Fiat money is that it gives central banks greater control over the economy because they can control how much money is being printed. Here is where the exact problem lies that if government prints too much of money, it will result in hyperinflation.

To solve this problem of unlimited supply of money came the concept of Bitcoin. Bitcoin was launched in 2009 wherein the supply of tokens would be capped would be limited at 21 million. But in future dollars, rupees and other currencies along with cryptocurrencies, will coexist rather than cryptocurrencies completely wiping off the existing non-cryptocurrencies.

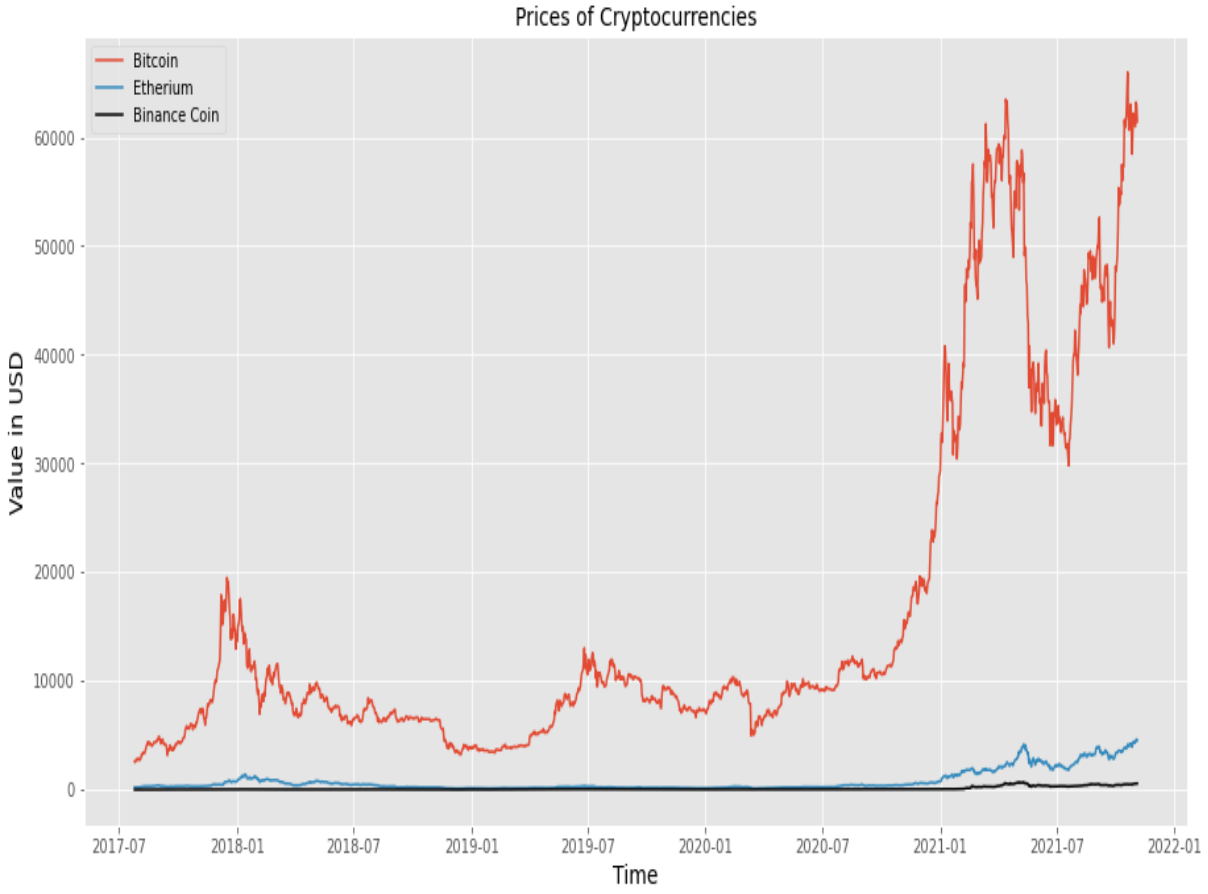
Bitcoin is the world's first cryptocurrency which was developed by a pseudonym Satoshi Nakamoto in 2009 and has been one of the most volatile asset in history. Suppose someone invested 100$ in around October 2010 when bitcoin was trading at 10 cents, the person would have bought 1000 bitcoin and the value of the asset as of November 5, 2021 would be 685 million USD! But along with it's attractive return it's pretty volatile and vulnerable to swing traders.

In July 2015 another cryptocurrency came live named etherium which was conceived in 2013 by programmer Vitalik Buterin. Etherium is now world's 2nd highest market capped crypto with a market capitalisation of nearly 512 Billion USD where Bitcoin stands at top with a market value of over 1 trillion USD.

Another cryptocurency name Binance Coin came into existence in July 2017 and already have reached a

market cap of 100 billion.

As a pioneer, Bitcoin have been most popular and most volatile of all the three highest marked valued cryptocurrency mentioned above. But a single bitcoin costs above 60 thousand USD or above 40 lakhs INR so it's hard for a common investor to buy one bitcoin. Where as Etherium and Binance Coin is currently trading for around 4000 and 600 USD respectively which is way less than the price of a bitcoin.



Objective

The main goal of our project is to search, establish and validate if there is a causal relationship among them and to see whether the price movement of the series in long run are similar or not.

We will first do a basic check of stationarity and then find if the time series are cointegrated among themselves or not and lastly we will check if the price of Bitcoin granger causes the price of Ethereum or Binance Coin.

1 About Dataset

1. The historic data of the cryptocurrencies have been collected from Yahoo finance website.

2. A historic data contains mainly 5 columns naming Open, High, Low, Close and Adjusted Close.

3. We only consider the close values and made a dataframe merging all the close values of 3 cryptocurrencies.

2 Detection of Non-stationarity

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and T be an index set. A real valued time series is a real valued function $X(t, \omega)$ defined on $T \times \Omega$ such that for a fixed t, $X(t, \omega)$ is a random variable defined on $(\Omega, \mathcal{F}, \mathcal{P})$. If X_t is a process whose statistical properties do not change over time, we call the process Stationary. Here by stationarity we mean weak stationary or covariance stationary. Here we assume that the first two moments and their joint moment do not change over time. Now to test stationarity, we will use Augmented Dickey-Fuller Test.

2.1 Augmented Dickey-Fuller Test

Augmented Dickey-Fuller test is widely used to detect non-stationary times series. It is an extention of Dickey-Fuller Test. Consider the model,

$$X_t = \phi X_{t-1} + \epsilon_t,$$

where, ϵ_t is a white noise process. The basic intuition behind the Dickey-Fuller Test for nonstationarity is to simply regress X_t on its (one period) lagged value X_{t-1} and find out if the estimated coefficient ϕ is statistically equal to 1 or not. However, the errors of the term in the DF test usually depicts serial correlation. To remove this drawback, we use Augmented Dickey-Fuller Test. In this project we will denote the first order difference using the sign 'Δ'. ADF test is applied to the model,

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \sum_{j=1}^{p} \delta_j \Delta X_{t-j} + \epsilon_t,$$

where α is intercept, β is the coefficient on a time trend series and p is the order of autoregressive process. ADF test can be done using three different models.

1. A pure random walk without drift, i.e., imposing the constraint $\alpha = 0$ and $\beta = 0$.

2. A random walk with a drift, i.e., imposing the constraint $\beta = 0$.

3. A model with deterministic trend with a drift i.e., $\alpha \neq 0$ and $\beta \neq 0$

The lag value p is to be determined using ACF lot or PACF plot. Alternately, we can use usual t-test or Akaike information criterion (AIC), Bayesian information criterion (BIC) information to estimate p too. The test procedure is described as follows.

Test Procedure :

Here we test the null hypothesis $H_0 : \gamma = 0$ against the alternative hypothesis $\gamma < 0$.

The **test statistic** is given by,

$$F_\gamma = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

where, $SE(\hat{\gamma})$ is the standard error of the $\hat{\gamma}$. As this is a left tail test,we will reject the null hypothesis against the alternative at 5% level of significance if the observed value of the test statistic is less than the critical values of Dickey Fuller t distribution and will conclude the series is stationary.

3 Detection of Cointegration

A time series process $X_t, t \in \mathbb{N}$ is said to be **integrated of order** d (symbolically, $X_t \sim I_d$) if d is the smallest integer $\ni \nabla^d X_t$ is a stationary process. Such a process is always non-stationary if $d > 0$. However, we can find a linear combinantion of two time series processes X_t and Y_t with same order of integration as a stationary process. In that case the processes X_t and Y_t are called co-integrated and are said to be connected through a model, known as error correction model (ECM). ECMs helps us to estimate the speed of a dependent variable at which it stabilizes to equilibrium when there is a change in other variables. Now we are discussing two tests used to detect co-integration.

3.1 Engle-Granger Method

Engle-Granger Test is a popular method to detect cointegration.[Ssekuma (2011)]

3.1.1 Pre-test each variable

Let $(X_t)_{t\in\mathbb{N}}$ and $(Y_t)_{t\in\mathbb{N}}$ be two nonstationary series of our interest. At first, we will fit a linear regression between $(X_t)_{t\in\mathbb{N}}$ and $(Y_t)_{t\in\mathbb{N}}$ using ordinary least square method. Here our model is given by,

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t,$$

where ϵ_t is the error term. The estimates of the parameter is given by, $\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y})}{\sum_{t=1}^{n}(x_t-\bar{x})^2}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. The estimated line is then given by,

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t.$$

If these variables are cointegrated then we will find strong linear relationship between the variables which will give the value of$\hat{\beta}$ falls between 0.5 to 1 or the plot of y against x will appear in an increasing or decreasing direction. Now the residuals can be estimated as,

$$\hat{\epsilon_t} = y_t - \hat{y}_t. \tag{1}$$

If the series of residual is stationary then the two series are cointegrated. To check stationarity we apply ADF test on the model,

$$\Delta\hat{\epsilon}_t = a_1\hat{\epsilon}_{t-1} + \epsilon_t,$$

The notation is as described above. We will proceed to test stationarity using ADF test by checking the significance of the estimates of a_1 using ADF test that is described in the previous section. The rejection of H_0 implies that residuals are stationary and this further implies that the variables under study is cointegrated.

3.1.2 Limitations

The main drawback of this test is using this method we can only test cointegration between two variables, i.e., only one cointegrating relationship can be verified. Now if we have a large number of variables under study, say r, then we need to perform the test ${}^rC_2$ times. Hence, we are proceeding for Johansen's Procedure to remove this drawback.

3.2 Johansen's Procedure

[Ssekuma (2011)] Johansen's Test which is named after Søren Johansen, is widely used for testing cointegration among a set of time series variables which are **integrated of order one i.e.,** I_1. Let's say, we have n time series of interest that are integrated of order one. Denote X_t as the $n \times 1$ vector whose components are the I_1 time series of our interest. At first we assume that it follows a Vector Autoregression (VAR) model of order p. The model is given by,

$$X_t = \Pi_1 X_{t-1} + \Pi_2 X_{t-2} + \cdots + \Pi_p X_{t-p} + u_t. \tag{2}$$

u_t is multivariate Gaussian noise term with null mean vector. This test can be used to test cointegration among two or more variables. Now we subtract X_{t-1} from both the sides of (2).

$$\begin{aligned} & X_t = \Pi_1 X_{t-1} + \Pi_2 X_{t-2} + \cdots + \Pi_p X_{t-p} + u_t \\ \implies & X_t - X_{t-1} = (\Pi_1 - I_n) X_{t-1} + \Pi_2 X_{t-2} + \cdots + \Pi_p X_{t-p} + u_t \\ \implies & \Delta X_t = \Gamma_1 \Delta X_{t-1} + \Gamma_2 \Delta X_{t-2} + \cdots + \Gamma_{p-1} \Delta X_{t-p+1} - \Pi X_{t-p} + u_t \end{aligned}$$

where, $\Gamma_1 = \Pi_1 - I_n$, $\Gamma_i = \Pi_i - \Gamma_{i-1} \quad \forall i$ and $\Pi = I_n - \Pi_1 - \Pi_2 - \cdots - \Pi_p$. The matrix Π is called the impact matrix as it determines the extent to which the system is cointegrated. For $i \in \{1, 2, \ldots, n\}$, i^{th} equation of the above system is

$$\Delta X_{it} = \gamma_{i1} \Delta X_{i(t-1)} + \gamma_{i2} \Delta X_{i(t-2)} + \cdots + \gamma_{1p-1} \Delta X_{i(t-p+1)} - \Pi' X_{i(t-p)} + u_{it}$$

Here γ_{ij} is i^{th} row of Γ_j. ΔX_{it}, i.e., the i^{th} component of X_t is stationary or I_0 as X_{it} is I_1. u_t is also I_0. So, $\Pi' X_{i(t-p)}$ is also stationary.

If none of the components of X_t are cointegrated, they must be zero. On the other hand, if they are cointegrated, all the rows of Π must be cointegrated but not necessarily distinct. This is because the number of distinct cointegrating vectors depends on the row rank of Π [Ssekuma (2011)]. Now there are two types of Johansen's procedure.

3.2.1 The trace Test

Here the null hypothesis is r variables are cointegrated, against alternative n variables are cointegrated.
Test Statistic: The test statistic is given by,

$$\mathcal{J}_{trace} = -T \sum_{r+1}^{n} ln(1 - \hat{\lambda_i}),$$

where T is the sample size and $\hat{\lambda}_i$ is the i^{th} largest canonical correlation. This test statistic does not follow any standard distribution. The critical values are calculated by the software, or are found out using tables. If the observed value of the test statistic is higher than the critical value, we will reject the null hypothesis.

3.2.2 The Maximum Eigenvalue Test

For this test we test the null hypothesis, the number of cointegrating vectors is r, against alternative hypothesis the number of cointegrating vectors is $(r + 1)$.

Test statistics: The test statistic is given by,

$$\mathcal{J}_{trace} = -T(1 - \hat{\lambda}_{r+1}).$$

Here as well, the test statistic does not follow any standard distribution. The critical values are calculated by the software, or are found out using tables. If the observed value of the test statistic is higher than the critical value, we will reject the null hypothesis.

For both the tests the test is performed for each $r = 0, 1, \ldots, (n-1)$. If the hypothesis $r = 0$ is accepted then there are no cointegrating equation and if rejected then there is at least one cointegrating equation. If then the test for $r = 1$ is accepted then we will conclude there is exactly one cointegrating equation; if rejected then at least two cointegrating equation are present. Likewise, if for some value of r^*, our test gets accepted, we will conclude that r^* is the number of cointegrating vectors. Which infers the series will become stationary after using a linear combination of r^* variables.

Further, we can deduce error correction model using this method. We can find a linear combination by using components of eigenvectors corresponding to the largest eigen value.

3.2.3 Limitations

There is an underlying assumption that the cointegrating vector remains constant during the period of study but it is not the true scenario. It may happen that in long run the scenario changes. The technical progress, economic crisis, changes in people behaviour and preferences etc may causes the change. If the sample period is long, this phenomenon occurs.

4 Detection of Causal Relationship

When we are interested in more than one variables, we are mostly interested in finding the answer, whether there is any relationship between these two series or not. If they are related then we further ask, is one series is the cause of the other, i.e., **is there any causal relationship between these series?** In this section, we will try to find the answer of this question. There are four kind of Causality.

1. unidirectional causality from X to Y,
2. unidirectional causality from Y to X,
3. if X causes Y and Y causes X (It is known as **feedback**),
4. no causal relationship between X and Y.

Further, if the present value of X effects the present value of Y, we call it X is Instantaneous Granger Causal to Y. We are not considering this case here.

To detect causality here we will use **Granger Causality Test**.

4.1 Granger Causality Test

By intuition, it is clear that, the present value of the variable is dependent on the previous values of the time series. Now, if time series X is the cause of the the series Y then the present value of Y must be effected by the previous values of X in addition to the past values of y. Hence, we can intuitively say, if the prediction of present value of Y using past values of both X and y is significantly different from the prediction of present value of Y using past values of Y only, then we should say that X is the cause of Y. Formally, there are two basic principles behind the test,

1. The cause of the process occurs before the effect.
2. The cause gives unique information about the future values of its effect.

Now we are formally describing the test. Li et al. (2017)

We are interested in the following two time series X and Y. Granger test is performed using these two regression model. Let L be our lag parameter.

The full model denoted by M_f uses the lagged values of both the time series X and Y which is given by,

$$Y_t = \sum_{l=1}^{L} \alpha_l Y_{t-l} + \sum_{l=1}^{L} \beta_l X_{t-l} \tag{3}$$

The reduced model denoted by M_r uses the lagged values the time series Y only, which is given by,

$$Y_t = \sum_{l=1}^{L} \alpha_l Y_{t-l} \tag{4}$$

We say X Granger-causes Y, if the full model performs "significantly better" than the reduced model. To establish this statistically, we use the following $F - test$, basically that is used to test the null hypothesis $\beta_l = 0 \quad \forall l$ against the alternative that the null is not true.

Test Statistic :

$$F = \frac{SSE_r - SSE_f}{SSE_f} \times \frac{n - d_f - 1}{d_f - d_r} \overset{H_0}{\sim} F_{(d_f - d_r),(n - d_r - 1)}. \tag{5}$$

Here SSE_f and SSE_r denotes the sum of square due to error of the full and the reduced model respectively.

If y_t denotes the observed value of Y at point t and $\hat{y}_t$ denotes the predicted value using the full model, then, $SSE_f = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2$. Similarly we can define SSE_r using the reduced model. d_f and d_r denote the number of independent variable in the full model and the reduced model. n is the total number of observations. Here $d_f = 2L$ and $d_r = L$.

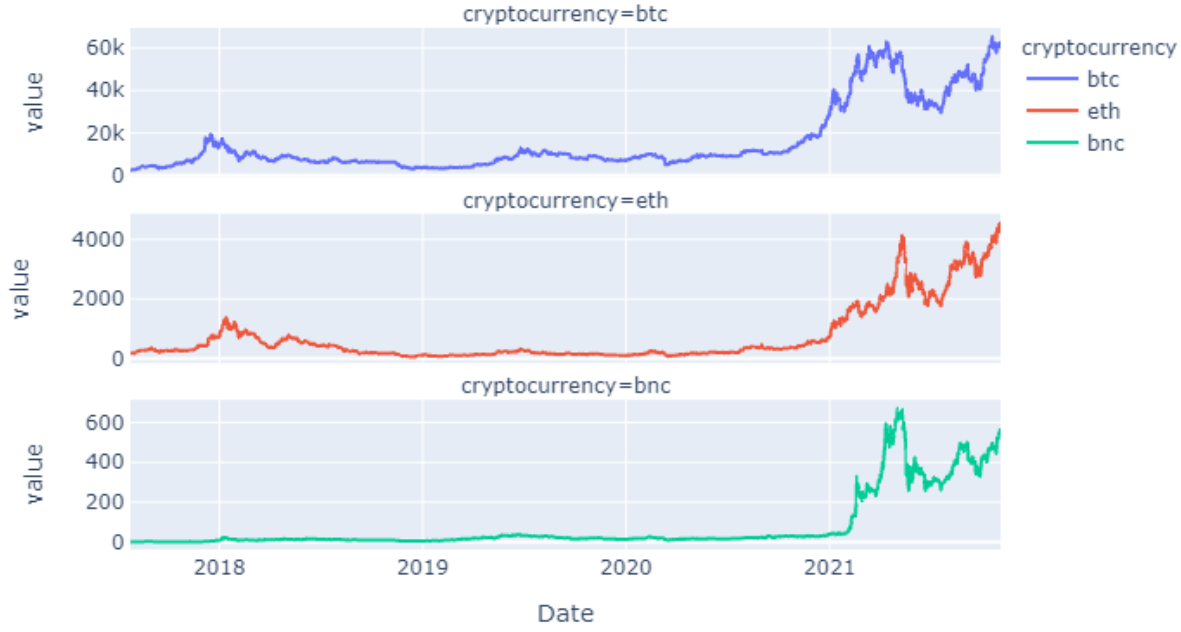
We will reject null hypothesis at $\tilde{\alpha}\%$ level if the observed value of $F > F_{\tilde{\alpha},(d_f-d_r),(n-d_r-1)}$, where $F_{\tilde{\alpha},(d_f-d_r),(n-d_r-1)}$ denotes the upper $\tilde{\alpha}\%$ point of $F_{(d_f-d_r),(n-d_r-1)}$ distribution.

5 Results and Interpretations

5.1 Results from Stationarity Test

The null hypothesis for Augmented Dickey Fuller test is H_0: Time series is non stationary vs H_1: Time series is not non-stationary.

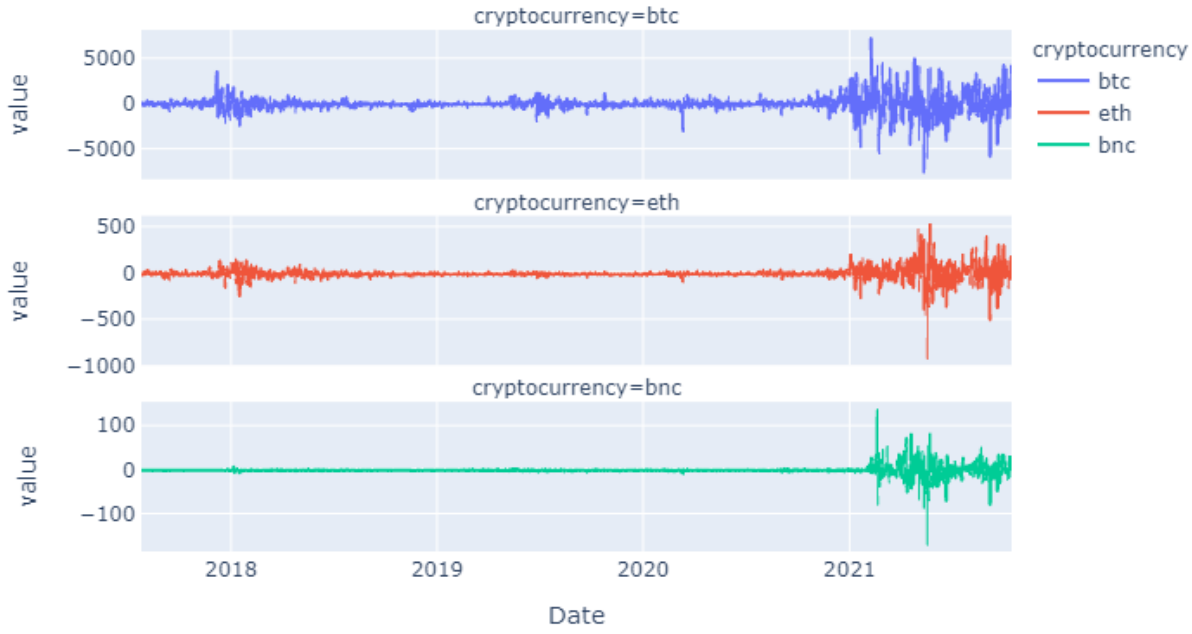
The result of the test on 3 time series are as follows:



Cryptocurrency	Test Statistic	p-value	Critical Values		
			1%	5%	10%
Bitcoin	0.538901	0.986009			
Ethereum	1.084446	0.995082	-3.435	-2.863	-2.568
Binance Coin	-0.345244	0.918851			

So after performing ADF test for stationarity we can conclude that the high p-value shows not enough evidence to reject the null hypothesis and we can hold our assumptions for the time series being non-stationary.

After that we will check if the first difference of our time series is stationary or not. The result of ADF test on the 3 series after first difference is as follows:



Cryptocurrency	Test Statistic	p-value	Critical Values		
			1%	5%	10%
Bitcoin	-7.639731	0.0000			
Ethereum	-11.222493	0.0000	-3.435	-2.863	-2.568
Binance Coin	-7.524771	0.0000			

So after performing ADF on the differenced series we can see very low value of the test statistics which supports the rejection of null hypotehesis and we can conclude that all the 3 time series has order of integration 1.

5.2 Results from Cointegration Test

5.2.1 Engle–Granger Test

- Bitcoin-Ethereum: p-value for the adf test = 0.02567
- Bitcoin-Binance: p-value for the adf test = 0.02988

In both the cases the null hypothesis is rejected, that is, $\{\epsilon_t\}$ is stationary. Hence both the pairs, Bitcoin-Ethereum and Bitcoin-Binance are co-integrated.

5.2.2 Johansen's test:

- Trace Test:

variable	Test Statistic	Critical Values		
		90%	95%	99%
r=0	79.9731	32.0645	35.0116	41.0856
r=1	36.2709	16.1616	18.3985	23.1485
r=2	0.2537	2.7055	3.8415	6.6349

- Maximum Eigen Value Test:

variable	Test Statistic	Critical Values		
		90%	95%	99%
r=0	43.7028	21.8731	24.2522	29.2631
r=1	36.0166	15.0006	17.1481	21.7465
r=2	0.2537	2.7055	3.8415	6.6349

We can see that in both the tests, the null hypothesis is rejected for $r = 0$ and $r = 1$ but accepted for $r = 2$. Hence it can be concluded that there are two cointegration equations possible.

5.3 Results from Granger Causality Test

The null hypothesis for the granger causality test is H_0: The time series x(t) does not granger cause y(t) i.e lagged values of x does not help predicting the variation in y. And the alternate hypothesis is that there is presence of granger causality.

In the below chart we have made a matrix which is filled with the p-value of the test where the p-value is for the testing if the column series granger causes the time series present in the row.

	Bitcoin	Etherium	Binance Coin
Bitcoin	1.0	0.0	0.0
Etherium	0.0	1.0	0.0
Binance Coin	0.0	0.0	1.0

A low p-value suggests rejection of the null hypothesis and we can conclude that all the time series granger causes each other. Since we are specifically interested if the Bitcoin series granger causes Etherium and Binance Coin or not so in that senario we can reject the null hypothesis and conclude that bicoin does granger cause Etherium and Binance Coin.

Conclusions

Our main objective was to check whether Bitcoin price affects other top crypotcurrencies or not and whether they exhibit similar stochastic trend in long run and we have established our assumption as true. Though each of our cointigration and causality test have their own drawbacks but in this scenario our test performed well and gave meaningful results.

Online References

1. Johansen test
2. Granger causality
3. The results and interpretaion of the tests is taken from the documentation of python statsmodels package.

References

[1] Li, Z., Zheng, G., Agarwal, A., Xue, L., and Lauvaux, T. (2017). *Discovery of Causal Time Intervals*, pages 804–812.

[2] Ssekuma, R. (2011). A study of cointegration models with applications.