Cryptocurrencies: A Time Series Analysis Through Logistic Regression, Prophet, Cointegration, Clustering, And Geometric Brownian Motion

Santino Luppino, Patrick Besser, Zheng Li, Luke Ditton, Yitao Huang
Advisor: Khaldoun Khashanah

05/09/2022

Key Words: Cryptocurrencies, Time Series, Technical Analysis, Fundamental Analysis, Assets

Classification: Time Series Analysis: Analysis of specific data points over a period of time. Fundamental Analysis: A method of analyzing an asset's intrinsic value by analyzing related financial and economic factors.

0. Abstract

The cryptocurrency market has experienced a high level of returns and a high level of volatility through the past few years. There is still a sense of mystery behind the cryptocurrency markets and what actually makes the prices move. Data will be collected for the different time series of Bitcoin (BTC), Ethereum (ETH), Binance Coin (BNB-USD), XRP USD (XRP-USD), Litecoin USD (LTC-USD), and DogeCoin (DOGE-USD) through a 3 year period. This study will involve the use of Logistic Regression, Prophet, Cointegration, Clustering, and Geometric Brownian Motion to analyze if the select group of cryptocurrencies exhibit times series characteristics similar to that of the equity markets. The Logistic Regression model created from these 5 cryptocurrency models proved to be efficient enough in making predictions regarding the movement of the returns either up or down; BNB stood out from the study with the highest accuracy of all the cryptocurrencies under this same model. The Prophet models created from the five cryptocurrencies all show very small mean squared error. Among the given Prophet models, the one with Bitcoin turns out to have the highest accuracy and the model with Ripple has the least accuracy. Overall, Prophet seems to be a great tool for fitting and predicting cryptocurrencies prices.K-means clustering was used to find similarities within our data. We found that BNB and XRP are often grouped separately, and BTC, ETH, and LTC have similar returns and were grouped together. We used Cointegration testing to determine whether the 5 cryptocurrencies exhibit cointegration relationships. We found that all 5 cryptocurrencies are cointegrated and that every possible pair of cryptocurrencies is cointegrated. Lastly, Geometric Brownian Motion did not model the cryptocurrency price paths accurately since the logarithmic returns are not from a normal distribution. Overall, our study demonstrated that the likelihood of extreme returns affected the models ability to make accurate predictions and it was discovered that UUP, SPY, USO, TNX, VIX are the most influential side factors on the price movement in the sample of cryptocurrencies. The Prophet model demonstrated its accuracy in making predictions relative to Geometric Brownian Motion prediction models.

1. Introduction

Cryptocurrencies have evolved as one of the world’s most intriguing, yet perplexing global assets, which present both an investment opportunity and an alternative to the traditional monetary system. The intriguing allure of this market is embedded within the traditionally higher returns compared to the equity market. For example, Bitcoin’s market capital grew from 1 billion in 2013 to 1200 billion in 2021. Moreover, the enticing liquidity of the cryptocurrency market opens opportunities for the average joe to join in on this cryptocurrency frenzy. Mobile applications such as Robinhood and Coinbase added to the popularity of the cryptocurrency market by introducing low barriers of entry to the cryptocurrency market for the average investor. The perplexing component of the cryptocurrency market involves the comprehension of how these markets actually function. The mystery behind these markets involve the comprehension of the meaning of both fundamental and technical analysis that pertain specifically to these markets.

Understanding this huge currency market is necessary but complicated because it has a much smaller time period of existence for us to study compared to the equities market. Since the "players", who affect the ups and downs of the price, of the cryptocurrency market are the same as the "players" of the equities market, it is reasonable to apply a widely accepted method in the equities market - Time Series Analysis - on the cryptocurrency market. This study will use Logistic Regression, Prophet, Cointegration, and Clustering as a means of time series analysis for the different cryptocurrencies. Both Logistic Regression and Prophet will use a variety of different features as inputs in order to make predictions on the price movements of the chosen cryptocurrencies. Cointegration analysis seeks to determine which cryptocurrencies influence another cryptocurrency, as well as how much influence each has. Lasty, clustering will be used in conjunction with outside factors such as the VIX, US Dollar Index, and interest rates to determine which cryptocurrencies are closely related as a means of sector analysis.

This perplexing problem caused by the lack of fundamental analysis and the volatile behavior of the cryptocurrency market leads to the question: Is it possible to accurately predict the price movements of cryptocurrencies? The goal of this research is to identify relationships between cryptocurrencies and the side factors typically used to evaluate the equities market through different time series methods. Knowing more about the predictability of cryptocurrencies could help investors make more informed decisions and boost investor confidence in the market.

2. Literature Review

Do-Hyung et al. (2019) suggests that the cryptocurrency market has been shown to have time series characteristics that are also demonstrated in the equity market through a recent study from 2019. This study applied the long short-term memory (LSTM) model to classify the time series for cryptocurrency and involved collecting price data for BTC (Bitcoin), ETH (Ethereum), XRP (Ripple), BCH (Bitcoin Cash), LTC (Litecoin), DASH (Dash), and ETC (Ethereum Classic) and the KRW (Korean Won). The LSTM model used the price tensors of close price, high price, low price, and transaction volume. The model is used to determine whether the weight value is maintained by adding cell states in an LSTM cell. The test shows that XRP and DASH were predictable, while BCH was hard to predict. Moreover, this study discovered that window size and unit of time were not significant to prediction performance, therefore this suggests that the LSTM might have not been the best methodology for this study. This points out the need for a different form such as linear regression, regression trees, or other types of neural networks that could be a better predictor of cryptocurrencies.

Qureshi et al. (2020) addresses the question of whether the cryptocurrencies have an index or benchmark that they follow or whether the cryptocurrencies price movements are independent from each other. This research study investigated the multiscale interdependencies among Bitcoin, Ethereum, Ripple, Litecoin, and Bitcoin Cash) by using continuous wavelet transformation (CWT) analysis. This kind of analysis involves frequency analysis over the time series. The study showed that Ripple and Ethereum were the benchmarks that were moving the other cryptocurrencies; however, coherence is relatively unstable at higher frequencies, while the lower frequencies were significantly stable. Similar analysis can be performed on the cryptocurrency market by utilizing Ripple and Ethereum as the benchmark. The nature of the cryptocurrencies relative to these two benchmarks can be discovered through regression models while using the benchmarks as the dependent variable.

Liu et al.(2020) show that the cross-section of cryptocurrencies can be meaningfully analyzed using standard asset pricing tools. The data collected the information from over 200 major exchanges and provides daily data on opening, closing, high, low prices, volume and market capitalization (in dollars) for most of the cryptocurrencies. The study applied the Fama and French (1996) three-factor model. The paper shows that a three-factor model with the cryptocurrency market factor, a cryptocurrency size factor, and a cryptocurrency momentum factor explains the excess returns of the market.

Zhu (2021) shows that investor attention is indeed the granger cause of Bitcoin market both in return and realized volatility, and the shock from investor attention can last for several weeks in the Bitcoin market. The data used is the Bitcoin prices from July 1, 2013 to May 31, 2020 in daily frequency. Then use the Standard Granger to compare the time-series data with the Google search intensity of Bitcoin. The study proved that investor attention indeed affects the Bitcoin market both in return and realized volatility, respectively

While cryptocurrencies are interconnected, they are independent from other equities. Liu and Tsyvinski (2018) showed that the risk-return tradeoff of cryptocurrencies (in particular Bitcoin, Ripple, and Ethereum) is distinct from those of stocks, currencies, and precious metals. Cryptocurrencies have no exposure to the returns of other equities, but they can be predicted by factors specific to cryptocurrency markets. By constructing proxies for investor attention, the researchers proved that both positive and negative investor attention predict correlating returns for Bitcoin, Ripple, and Ethereum. This all goes to show that crypto is not affected by traditional factors, but is an entirely new and independent market.

Cryptocurrency and the blockchain technology itself are great technological advancements that have a lot of promise, but as they currently exist they are an environmental disaster. Cryptocurrencies are mined using either a proof-of-work (most popular) or proof-of-stake system. Proof-of-work is used for currencies like Bitcoin and Ethereum, and it has a much higher energy cost. Goodkind et al. (2018) estimated the cost of health and climate damages of energy uses for four cryptocurrencies. They found that in 2018, $1 of Bitcoin value created was responsible for $0.49 in damages in the US, and $0.37 in China. Since the value of these coins has increased since then, and proof-of-work scales energy costs to the price of the currency, we can assume that the human health and pollution damages have increased.

Göttfert investigates whether the six largest cryptocurrencies by market capitalization exhibit cointegration (2019). Cointegration is a technique used to find a possible correlation between time series processes in the long term (Corporate Finance Institute, 2020). If cointegration pairs are found investors will have more insight on how to interpret and act on the cryptocurrency market. It will also grant a better understanding of the price formations of cryptocurrencies. The Johansen cointegration test and the Engle-Granger two step analysis for cointegration are used to evaluate possible cointegration pairs between Bitcoin and altcoins, Ethereum, Ripple (XRP), Bitcoin Cash, Litecoin and EOS. Estimates of potential long-run relationships are made using a Vector Error Correction Model (VECM). The results of this study show that Bitcoin is cointegrated with Bitcoin Cash, Ethereum, Litecoin, and Ripple, all of the altcoins examined excluding EOS. The VECM results imply that the price of Bitcoin has a long-run effect on the prices of Bitcoin Cash, Ethereum, Litecoin, and Ripple that is statistically significant. This study utilized daily closing price data, but additional period lengths could provide more insight into the relationships between the cryptocurrencies.

Malladi and Dheeriya test the hypothesis of whether cryptocurrency prices are partly determined by the global stock indices, gold prices, and fear gauges such as the VIX and the US Economic Policy Uncertainty Index by conducting a time series analysis of returns and volatilities of Bitcoin and of Ripple (2021). To determine the relationships between returns and volatility of Bitcoin and Ripple, they used the autoregressive-moving-average model with exogenous inputs model (ARMAX), Generalized Autoregressive Conditionally Heteroscedastic (GARCH) model, Vector Autoregression (VAR) model, and Granger causality tests. Their findings suggest that the Bitcoin crash of 2018 could have been modeled and explained using their methods and that forecasts of the direction and magnitude of Bitcoin volatility are more precise than those of forecasted Bitcoin returns. They also found that the returns of gold and global stock markets do not have a causal relationship with Bitcoin but do influence smaller cryptocurrencies like Ripple. These smaller cryptocurrencies can have a causal effect on Bitcoin prices. The last finding is that Bitcoin prices are primarily influenced by measures of fear like the GVOL and VIX and other cryptocurrency returns like Ripple. This research is significant because an investor in any asset needs to be aware of the relationship and influences between traditional assets and cryptocurrencies.

Like mentioned above, neural networks are an effective approach to predict the fluctuations of cryptocurrency's effectiveness. The paper "Volatility Analysis of Bitcoin Price Time Series" Pichl, Kaizoji (2017) focuses on the price of Bitcoin in terms of different countries' currencies and their volatility over five years. Pichl and Kajzoji showed that the Heterogeneous Autoregressive model for realized volatility is working quite well on the Bitcoin dataset. More importantly, they constructed a feed-forward neural network with two hidden layers using ten days moving window sampling daily return as predictors. Then they used this neural network to estimate the next-day logarithmic return. The model demonstrates predictions that are roughly catching rare fluctuations. This paper then recommends more sophisticated machine learning models for higher accuracy. This paper helps us to decide avoiding complicated machine learning methods since the prediction results are not significantly higher but the running time is.

Yenidogan et al. (2018) used Prophet and Arima methods on R to forecast Bitcoin prices. The Prophet is a large scale time series forecast model strongly accommodated with missing data and is good at capturing the shifts in the trend within the data. They compared the Prophet model with a non-seasonal ARIMA model. Due to the characteristics of time series data, the paper divides their data into three sets in order: train, validation, and test. They also constructed a correlation matrix to find the most three correlated features. The paper's result shows that in both validation set and testing set, Prophet tends to outperform the ARIMA model. This paper helps us to decide using the Prophet model as one of the methods to analyze the cryptocurrency data.

One other inevitable behavior of cryptocurrencies is that financial price bubbles are commonly seen in cryptocurrency prices. And these bubbles have been linked with the epidemic-like spread of an investment idea. The paper "Predicting Cryptocurrency Price Bubbles Using Social Media Data and Epidemic Modelling" Phillips, Gorse (2017) aims to use the behavior of online social media indicators to predict such financial price bubbles using an established hidden Markov model that was previously used for detecting epidemic outbreaks. Phillips and Gorse used the model to build a trading strategy which turned out to outperform a buy and hold strategy based on the historical data. The importance of this paper's finding is that it opens our eyes and provides us a different perspective on the problem. It also teaches us social media data mining skills and proves the effectiveness of cross-curricular approach.

3.Methodology

3.1 Logistic Regression

Logistic Regression is a statistical regression model that uses a logistic function to model a binary variable[1]. Logistic Regression is used when the dependent variable is categorical. The dependent variable will be binary for 1 (if the logarithmic return of the cryptocurrency is greater than 0 and 0 otherwise, this will be set as our decision boundary for the logistic regression. The inputs of this Logistic Regression will be the lagged returns for the chosen cryptocurrency. The amount of lagged returns will be set at 5 days for the cryptocurrencies. Regarding the inputs, the

[1] This is an article that gives a detailed overview of what a linear regression is. See this link for more details: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

returns will be lagged by 5 days for each respective cryptocurrency. The logistic regression formula used in this study is as follows:

$$log(\frac{p_{1,t}}{1-p_{1,t}}) = \beta_0 + \beta_{1,1}r_{1,t-1} + \beta_{1,2}r_{1,t-2} + \beta_{1,3}r_{1,t-3} + \beta_{1,4}r_{1,t-4} + \beta_{1,5}r_{1,t-5} \quad (1)$$

This formula that has been created demonstrates the 5 lags that will be used from our data for each of the cryptocurrencies. An Autoregressive model will not be used because of the possibility of autocorrelation problems and the fact that the logistic regression model does not take models as inputs. Once the returns are lagged, they will be placed in a data frame along with the direction for each cryptocurrency. This will allow for the analysis if the cryptocurrency is up or down due to the lagged returns of what happened the 5 previous days. After our model is created, the accuracy of this model and the confusion matrix will be computed to test whether this type of time series analysis is viable for cryptocurrencies.

3.2 Prophet

Prophet is a tool developed by Benjamin Letham and Sean J. Taylor at Facebook. It is usually used for large scale time series forecasting. Prophet accounts for seasonality, trends, and holidays[2]. This is especially a good model for Bitcoin since Bitcoin has some seasonalities: a few months tend to be more profitable than other months, Saturdays have had the highest upside, and Thursdays have had the largest drops (Interdax 2020). Prophet in theory is similar to a generalized additive model. It fits trends, seasonality and holidays. Prophet's core is the sum of four functions: growth(g), seasonality(s), holiday(h), and error(e). For Prophet's growth function, it provides three options: [2]linear growth, logistic growth, and flat. Due to the characteristic of cryptocurrencies' trend, we would choose the linear growth option for the growth function. The seasonality function is just a Fourier Series, and it takes in consideration of time. For the holiday function, Prophet has a built-in list of dates of US holidays which the model will check for each holiday's effect on the input and modify the final forecast. The error function aims to catch the randomness that is not accommodated by the model. The model's is expressed as a function below:

$$y(t) = g(t) + s(t) + h(t) + e(t) \quad (2)$$

For the inputs, we will pick a long time period of Bitcoin and other crypto currencies prices and then split them into training and testing sets. The training data will be the input for Prophet. Then we will run Prophet and check its mean squared error and compare it to its mean square error with side factors since we will also be using other side information as predictors to get a more accurate Prophet model. This step may take the most of times, since we need to make sure that these side information have prediction power and they also have to be as independent to each other as possible. Examples are VIX index, interest rates, and etc. These side predictors will be

[2] Tune, Paul. "Time Series Modeling of Bitcoin Prices." Medium. Towards Data Science, February 20, 2021. https://towardsdatascience.com/time-series-modeling-of-bitcoin-prices-5133edfec30b.

determined by the clustering matrix. We will only choose the factors that have high prediction power for our crypto currencies time series data. The output for the model will be a forecast of future cryptocurrencies' price fluctuations and it will be compared with the logistic regression from above.

3.3 Cointegration

(add cointegration quantitative description)

Cointegration is a test used to establish if there is a correlation between several time series in the long term (Göttfert 2019)[3]. The method was developed in 1987 by Robert Engle and Clive Granger. If two or more non-stationary time series are integrated together in a way that they cannot deviate from equilibrium in the long term, then they are cointegrated. The tests are also used to identify degrees of sensitivity of two variables to the same average price over a time period. The three main methods of cointegration are the Engle-Granger Two-Step Method, Johansen Trace test, and Johansen Maximum Eigenvalue test. The Engle-Granger Two-Step method begins by generating residuals based on the static regression and then testing the residuals for the presence of unit roots. It then uses the Augmented Dickey-Fuller Test (ADF) or other tests to test for stationarity units in time series. The Engle-Granger method will show the stationarity of the residuals if the time series is cointegrated. This calculation can be expressed as:

$$A(L)\,\Delta y_t = \gamma + B(L)\,\Delta x_t + \alpha(y_{t-1} - \beta_0 - \beta_1 x_{t-1}) + \nu_t .$$

The Johansen test is used to test cointegrating relationships between multiple non-stationary time series data with large sample sizes. Compared to the Engle-Granger test, the Johansen test allows for multiple cointegrating relationships. The Johansen test will be utilized to determine if there are cointegration relationships present in the system of the five cryptocurrencies being studied. To perform a cointegration test, we will acquire long run non-stationary time series data for the securities we are assessing. Next we load the dataset into R studio and declare the time series objects. Next we bind the variables into a system and select the lag criteria. We then run the Johansen Trace test on our system and interpret the output of statistical critical values. By interpreting the output we can determine how many cointegration relationships are present in the system.

3.4 Correlation Clustering

The correlations discovered in the cointegration method will be used to partition data points into groups based on their similarities. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. The distance between nodes will be computed and stored into a matrix. The calculation can be express as the:

[3] http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-161079

$$\sum_{i \in C_k} \sum_{j=1}^{P} (x_{ij} - \underline{x}_{kj})^2 \quad (3)$$

Where C_k is the each cluster of K, and $\underline{x}_{kj}$ is the mean feature j in cluster C_k. $x_{ij} - \underline{x}_{kj}$ is Euclidean distance. For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster. Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance). Furthermore, this process only calculates the local optimum instead of the global optimum. Therefore, we must run it over and over again until the result does not change.

The correlations from the assets will be mapped using correlations then outputted into a network. This will be utilized using the clustering packages made available in R. Once our clustering model is created, indices such as the S&P500 and the US Dollar Index will be used to test whether they can be viewed as benchmarks to the cryptocurrency market as a whole. We are going to apply K-Mean Clustering in this case, since we only care about the relationship between factors we chose and the cryptocurrency. By seeing which groups the cryptocurrencies and other factors fall into, we can gain insight into the latent similarities between different kinds of assets.

3.5 Geometric Brownian Motion

Geometric Brownian Motion (GBM) is a continuous stochastic process in which the logarithm of the model follows Brownian Motion with drift. This follows the following equation:

$$S_t = S_0 * e^{(\mu_i - \frac{\sigma^2}{2}) * t + \sigma * W(t)} \quad (4)$$

, where Wt is Brownian Motion. The logarithmic returns and adjusted close prices for the cryptocurrencies will be used for this study. The training set for the data will be price data from 01/01/2018 to 06/20/2021 and the testing set will be data from 07/01/2021 to 12/31/2021. The calibration technique will be used on both the return (μ) and volatility (σ) of our data in order to minimize the discrepancy between theoretical value and the actual value in practice. The mean and standard deviation will be calibrated in the following way:

$$\sigma = \sqrt{365} * s \text{ and } \mu = 365 * m + \frac{\sigma^2}{2} \quad (5)$$

, with s being the standard deviation from the time series and m being the mean from the time series. 365 days will be used for the calibration since cryptocurrencies are traded everyday. This GBM simulation method will be used on the 6 individual cryptocurrencies used in this study to create price paths over the training period. Each cryptocurrency will be calibrated based on the training sample of the data. 1,000 simulations will be created for each of the 6 selected cryptocurrencies. Each simulation will run from the last price in the training set. The mean of each day in the simulation will be computed and used as our “best fit” simulation for the purposes of this study. The “best fit” will be a single time series composed of the average of each simulated day. This will potentially deliver a “theoretical average” of the simulations. The mean square error between the simulated returns and actual returns will be used as a metric to assess model accuracy. Lastly, Jarque-Bera tests for normality will be used to evaluate whether the logarithm of the returns follows a normal distribution. This will validate or invalidate the choice of using Geometric Brownian Motion as a means of predicting stock prices.

4. Data

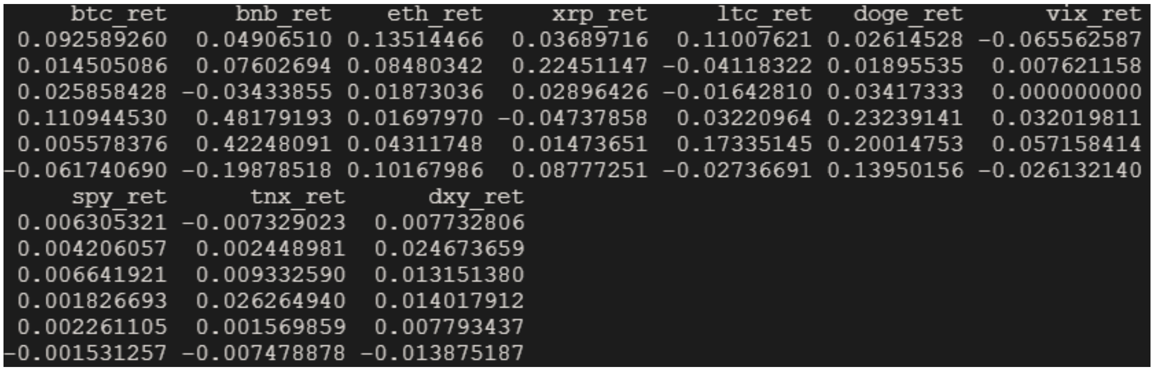
The study will use the adjusted close prices for 5 different cryptocurrencies, the VIX index, interest rates, and the dollar index. The cryptocurrencies used in this study are Bitcoin (BTC), Ethereum (ETH), Binance Coin (BNB-USD), XRP USD (XRP-USD), Litecoin USD (LTC-USD), and Dogecoin (DOGE-USD). The data will be taken over the period from 2018-01-01 to 2021-12-31 due to the availability of the data. Moreover, the frequency of the data will be daily for the adjusted close prices. The data will be collected by using the quantmod package in the R programming language library. The data will be downloaded from Yahoo Finance by using the getSymbols function in R. The data is free to obtain, so there will be no data associated with data collection in our study. Once the data is collected, the logarithmic returns will be computed for each of the cryptocurrencies.

The first part of our study involves performing a logistical regression analysis in order to determine if the lagged log returns from previous days have any predictive insight on the future returns. The data will be split into a 50% training and 50% testing split. The order will be the amount of days that will be lagged back will be set to 5 day for the purpose of this study. The lagged returns for each of the cryptocurrencies along with the direction of the log returns will be stored in a data frame and used to create a logistic regression. Once the model is created, the testing data will be compared to the predicted returns from our model. The accuracy and confusion matrix will be used for our interpretation.

As a contrast to the logistical regression model, we will perform a prophet model analysis. The goal is to see how well the prophet fits on the cryptocurrency data with some side factors and compare the results with the logistical regression model. In total, five different prophet models will be constructed, each with a different cryptocurrency dataset. The inputs for each will be the log return of each cryptocurrency. In addition, the choice of side factors will be determined based on the analysis of the cointegration model. The data will be split into 50% training and 50% testing. For each model, the daily seasonality will be set to false since our data is in days. The seasonality mode will be set to multiplicative because the seasonality volatility is increasing each year. Once the model is constructed, the mse will be evaluated and compared with the logistical regression model.

We will also gather additional data to be used as factors in our analysis. In the same time range as the cryptocurrencies, we will download daily data for the VIX volatility index, the federal funds rate, and the USDX dollar index from Yahoo finance and FRED. For the data downloaded from Yahoo finance, we will use the getSymbols function in the quantmod package in R to call the data. This data will be used directly as factors in Prophet to help forecast data and also in Cointegration.

We will conduct a cointegration test to find cointegration relationships between Bitcoin, Ethereum, Binance Coin, XRP, Litecoin, and the aforementioned additional variables. The daily adjusted close prices of the cryptocurrencies and daily variable data will be the inputs for the model. The Johansen trace test is our cointegration method as it tests cointegration relationships between multiple non-stationary time series data with large sample sizes. Once we have generated the correlations from the cointegration method, they will be clustered together into a network. The distance between nodes will be computed and stored into a matrix. The correlations from the assets will be mapped using correlations then outputted into a network. Once the clusters are created, the S&P500 and US Dollar Index will be used in these clusters to determine whether these can be used as benchmarks or not. A snapshot of the data header can be found bellow:

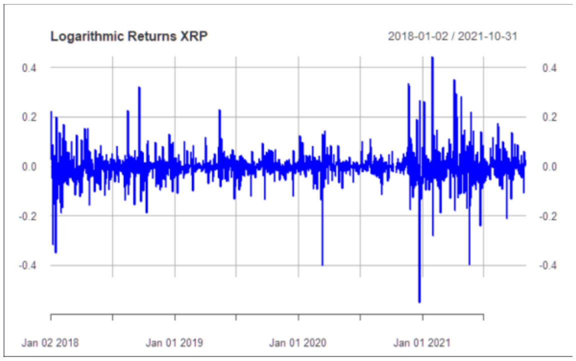


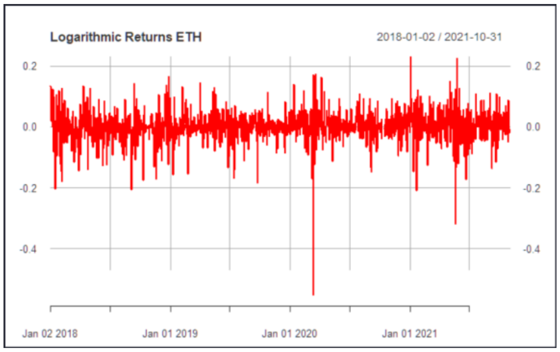
```
    btc_ret     bnb_ret    eth_ret     xrp_ret     ltc_ret   doge_ret      vix_ret
 0.092589260  0.04906510 0.13514466  0.03689716  0.11007621 0.02614528 -0.065562587
 0.014505086  0.07602694 0.08480342  0.22451147 -0.04118322 0.01895535  0.007621158
 0.025858428 -0.03433855 0.01873036  0.02896426 -0.01642810 0.03417333  0.000000000
 0.110944530  0.48179193 0.01697970 -0.04737858  0.03220964 0.23239141  0.032019811
 0.005578376  0.42248091 0.04311748  0.01473651  0.17335145 0.20014753  0.057158414
-0.061740690 -0.19878518 0.10167986  0.08777251 -0.02736691 0.13950156 -0.026132140
     spy_ret      tnx_ret      dxy_ret
 0.006305321 -0.007329023  0.007732806
 0.004206057  0.002448981  0.024673659
 0.006641921  0.009332590  0.013151380
 0.001826693  0.026264940  0.014017912
 0.002261105  0.001569859  0.007793437
-0.001531257 -0.007478878 -0.013875187
```

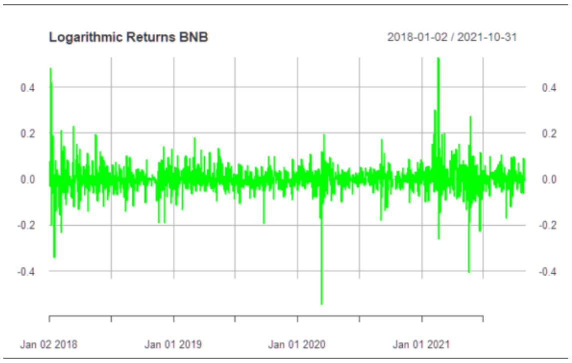
5. Results

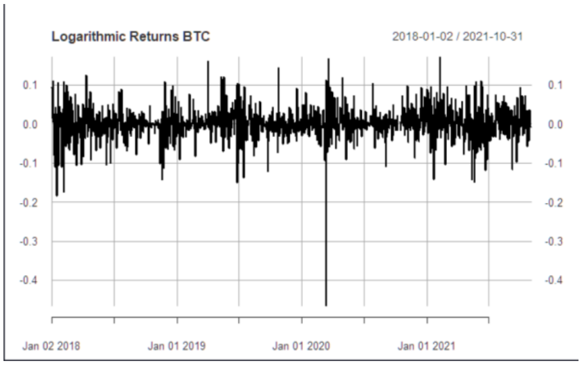
5.0 Logarithmic Returns

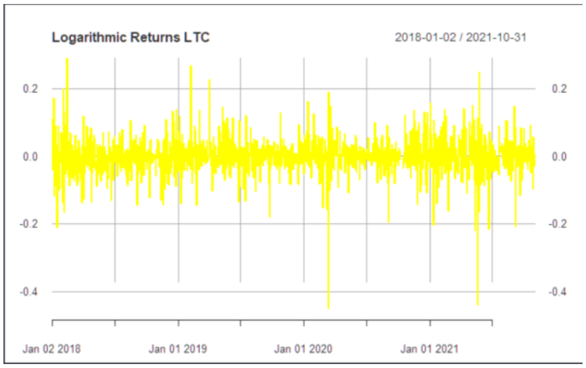
The logarithmic returns will be computed for the returns of Bitcoin (BTC), Ethereum (ETH), Binance Coin (BNB-USD), XRP USD (XRP-USD), Litecoin USD (LTC-USD), and DogeCoin (DOGE-USD). The log returns will be computed using R and the graphs for each of the logarithmic returns will be computed and depicted below:

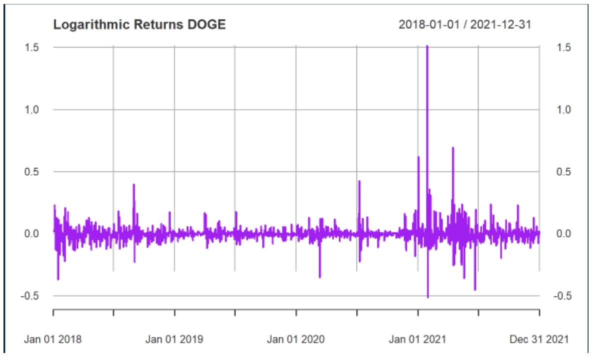



Logarithmic Returns ETH
2018-01-02 / 2021-10-31
0.2
0.0
-0.2
-0.4
Jan 02 2018
Jan 01 2019
Jan 01 2020
Jan 01 2021


Logarithmic Returns BNB
2018-01-02 / 2021-10-31
0.4
0.2
0.0
-0.2
-0.4
Jan 02 2018
Jan 01 2019
Jan 01 2020
Jan 01 2021


Logarithmic Returns BTC
2018-01-02 / 2021-10-31
0.1
0.0
-0.1
-0.2
-0.3
-0.4
Jan 02 2018
Jan 01 2019
Jan 01 2020
Jan 01 2021


Logarithmic Returns LTC
2018-01-02 / 2021-10-31
0.2
0.0
-0.2
-0.4
Jan 02 2018
Jan 01 2019
Jan 01 2020
Jan 01 2021



5.1 Logistic Regression

The Logistic Regression will be performed on the direction of the log returns for each of the logarithmic returns and the lagged returns for each respective cryptocurrency. The direction of the log returns will be determined by the following criteria: if the return is greater than 0, then a 1 is assigned to the return and a 0 otherwise. Since our previous attempt at Logistic Regression results in only the Down predictor being useful for interpretation, we will lower the lower the amount of categories in order to narrow down our results. The lagged returns will be set to 5 days for each of the cryptocurrencies, which will be the input to our model. This model was tested on a variety of different lags ranging from 0 to 10. The calibrated mu and sigma will be used for this model. From our experiments, the accuracies and confusion matrices were roughly similar in composition to each other. The accuracies from the model for BTC, BNB, ETH, XRP, LTC, and Doge are 50.69%, 73.08%, 48.63%, 53.43%, 50.82%, and 52.47%. The confusion matrices for each of the logistic regression models for each of the cryptocurrencies can be shown below:

BTC	**0**	**1**	**ETH**	**0**	**1**
0	**54**	**54**	**0**	**128**	**149**
1	**305**	**315**	**1**	**225**	**226**
BNB	**0**	**1**	**DOGE**	**0**	**1**
0	**251**	**89**	**0**	**248**	**206**
1	**107**	**281**	**1**	**140**	**134**

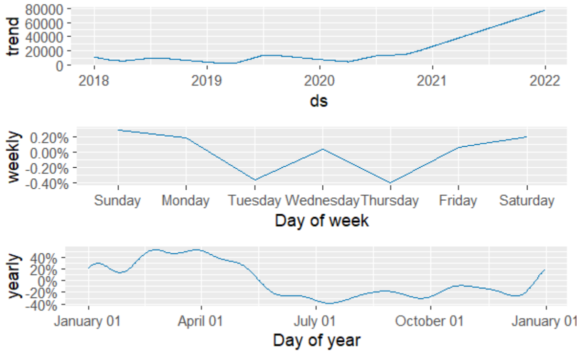
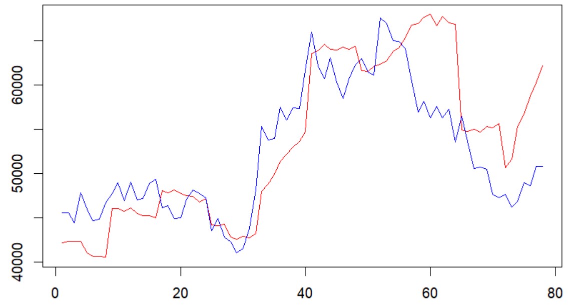
XRP	**0**	**1**	**LTC**	**0**	**1**
0	**245**	**201**	**0**	**236**	**240**
1	**138**	**144**	**1**	**118**	**134**

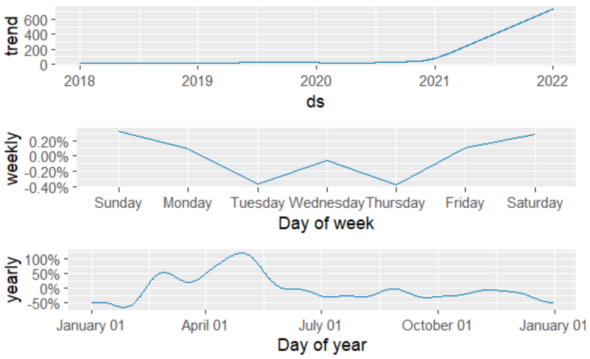
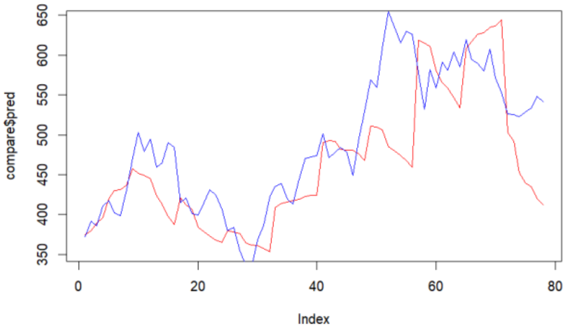
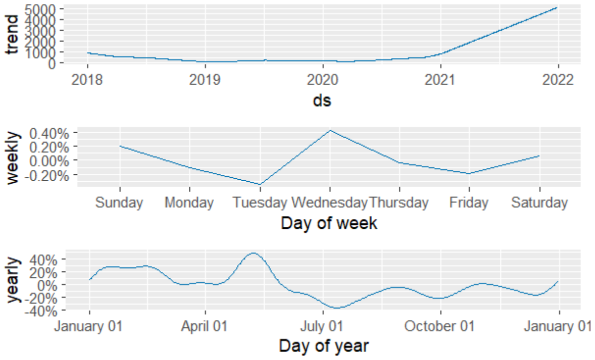
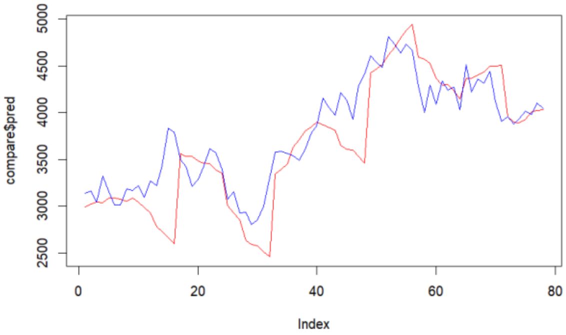
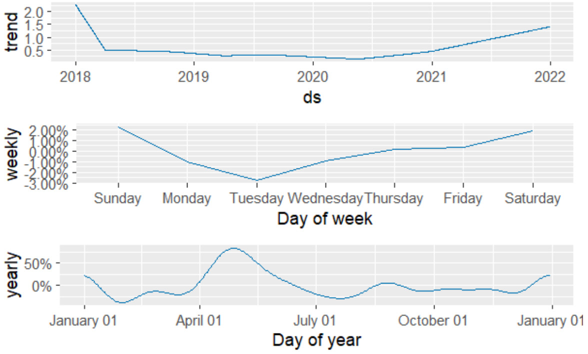
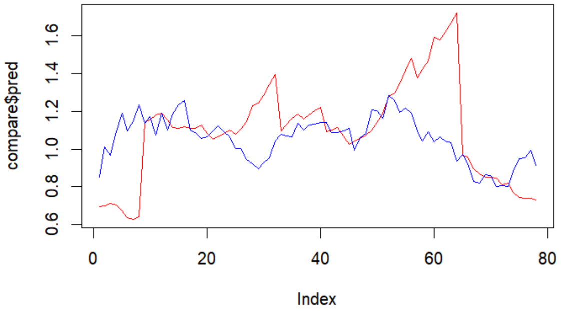
It is important to recognize that BNB stood out as the only crypto that had an accuracy that was better than that of a coin flip. The standard deviations of BTC, BNB, ETH, XRP, LTC, and Doge are 3.89%, 5.52%, 5.03%, 5.82%, 5.42%, and 6.02%. In terms of standard deviation it is important to recognize that BNB does not stand out among the other cryptocurrencies, which is not telling of the true nature behind its high accuracy. Moreover the Kurtosis computed for BTC, BNB, ETH, XRP, LTC, and Doge is 2.74, 14.91, 1.95, 11.32, 8.19, 9.10. Kurtosis does offer a contradicting explanation for its accuracy in the logistic regression model considering BNB has the highest kurtosis, yet the highest accuracy in the model.

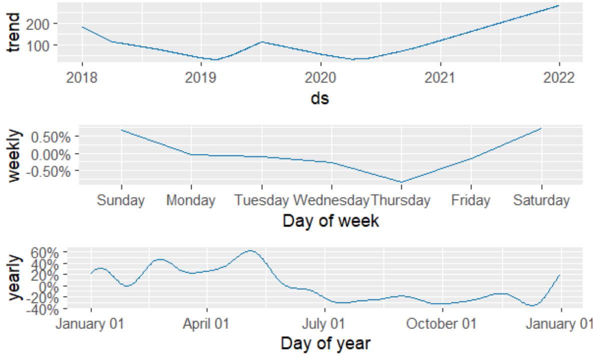
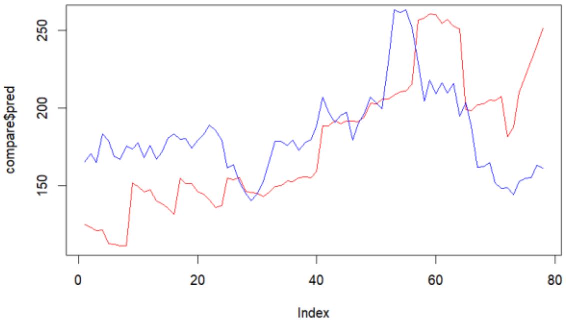
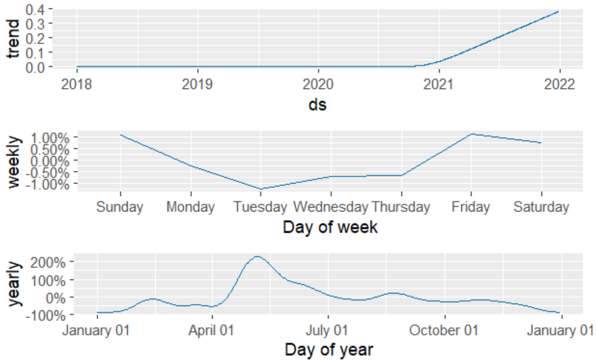
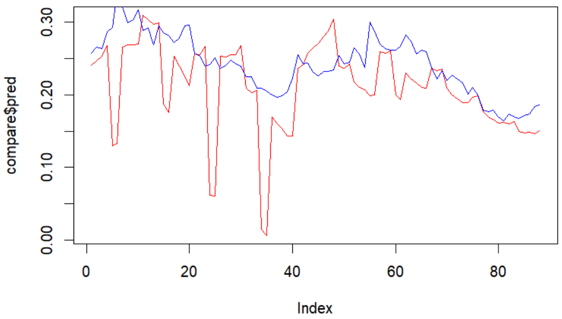
5.2 Prophet

Six different Prophet models each with a different cryptocurrency data were runed. The prices were the inputs and the forecasting period is 80 days. In addition to these, the models also takes three side factors: UUP, SPY, USO, which were given from the results of cointegration method. In order to compare the mean squared error, I had to normalize the input which was to take the log returns of the cryptocurrencies prices and the predicted prices. For each cryptocurrencies, a graph with trend analysis will be shown, a mean squared error of Prophet's predictions with only historical data as input, a mean squared error of Prophet's predictions with side factors as input as well, and a graph with the model's predicted price with side factors are shown below:

Prophet: Trend Observation and Price Predictions

Bitcoin: 	**MSE: 0.00228 MSE(With Side): 0.00146** 

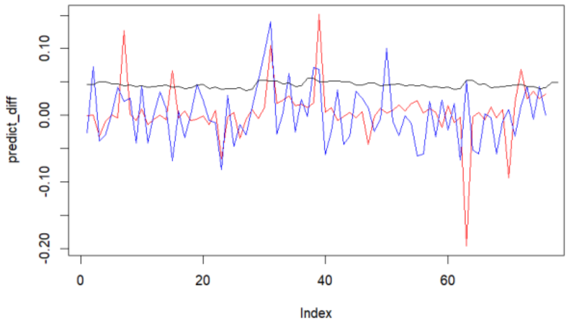
BNB:  	**MSE: 0.00605 MSE(With Side): 0.00218** 
ETH:  	**MSE: 0.00311 MSE(With Side): 0.00228** 
XRP:  	**MSE: 0.00380 MSE(With Side): 0.00514** 
LTC:	**MSE: 0.00355 MSE(With Side): 0.00239**

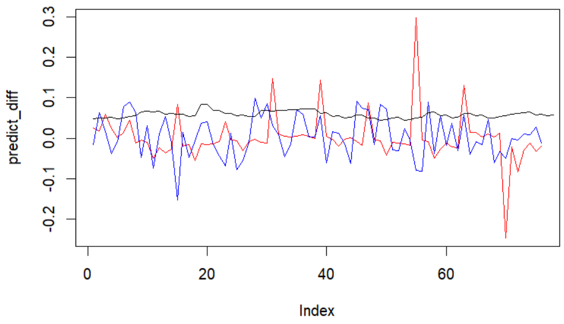
	
DOGE: 	**MSE: 0.00237 MSE(With Side): 0.00349** 

The graphs above show that all cryptos have had an upward trend in the most recent 2 years. They all seem to have a lower return on Tuesdays and Thursdays of each week. They also seem to do well in the second quarter of each year. For the graphs on the right, the red lines show the Prophet's predictions and the blue lines show the real price movements. The Prophet's predictions seem to be well aligned with the real crypto price movements. Adding the side factors, most of the models seem to have a lower mean squared error except for XRP and DOGE. One possible explanation for this is that these two cryptos' prices are both really small - less than 1 dollar. However, their market capitalization is not far from the other cryptocurrencies, which means that the number of coins for these two cryptos are way more than the other cryptos. Due to the small price and large number of coins, the effect of the side factors may be reduced to minimum. The graphs below show how Garch model's prediction tell us about Prophet's predictions:

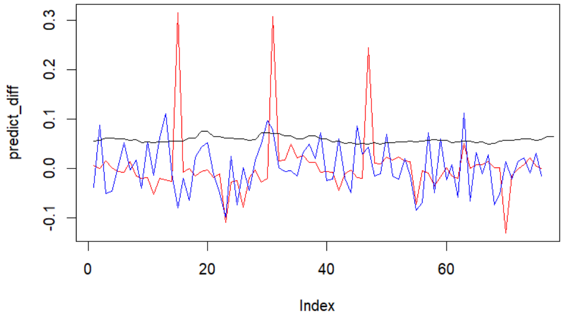
Garch Log Return Volatility Predictions with Prophet

BTC: 0.043	BNB: 0.056

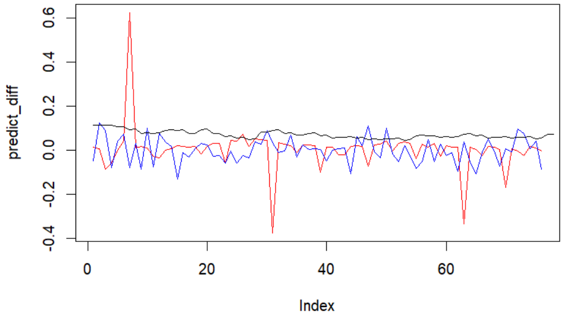




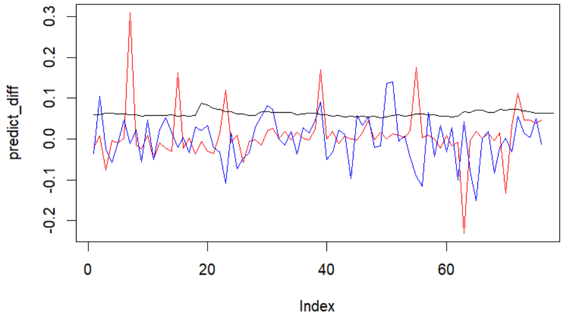
ETH: 0.058



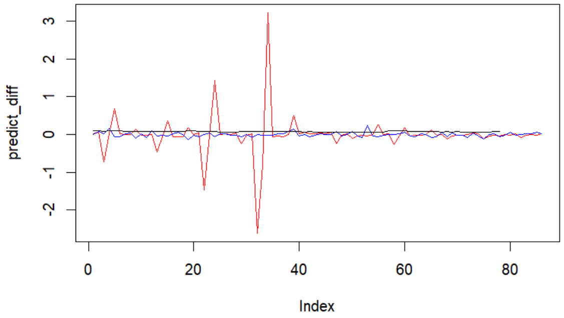
XRP: 0.069



LTC: 0.061



DOGE: 0.074



Garch model is a classic statistical model that describe time series data's volatility of the error terms. In the graphs, the black lines are the Garch prediction of future log return volatility. From the graphs, the Garch lines seem to well cover most of the fluctuations in both real price and Prophet's predicted price movements. There seems to be more spikes in Prophet's predictions. This may be due to the fact that the Prophet models were conducted with a rolling window. Whenever Prophet's prediction deviates from the real price movements, the next rolling window will correct the predictions which leads to some large changes in price movements. Again, XRP and DOGE's predictions were a bit more off compared to the others. This may be due to the same reason mentioned above - small price level and large amount of coins in the market.

In general, Prophet seems to be a very solid model for forcasting cryptocurrencies' price fluctuations. Prophet's log return predictions have an average mean squared error of 0.0028 with side factors. On average, the three side factors, UUP, SPY, USO, improve the model's mean squared errors by 0.0006.

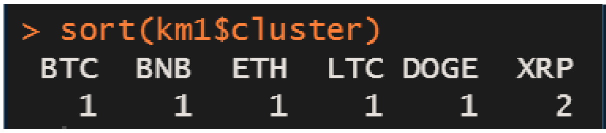
5.3 Clustering

We performed k-means clustering with three groups on the lagged adjusted returns of 5 cryptocurrencies. This kind of clustering places data into groups based on their means. We also included the lagged returns of the US dollar index, the US Treasury yield spread, the volatility index (VIX), and the S&P 500 (SPY) as additional data.

Without the additional factors, we clustered with 2 groups and 3 groups. In both cases, Ripple (XRP) is in its own group. In the three group model, XRP is also in its own group, and so is Binance Coin (BNB). This could allow us to view our other results from the context of those three currencies being somewhat similar.

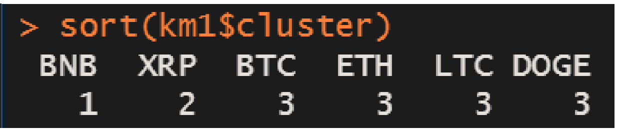
Clustering without additional factors:

2 groups:



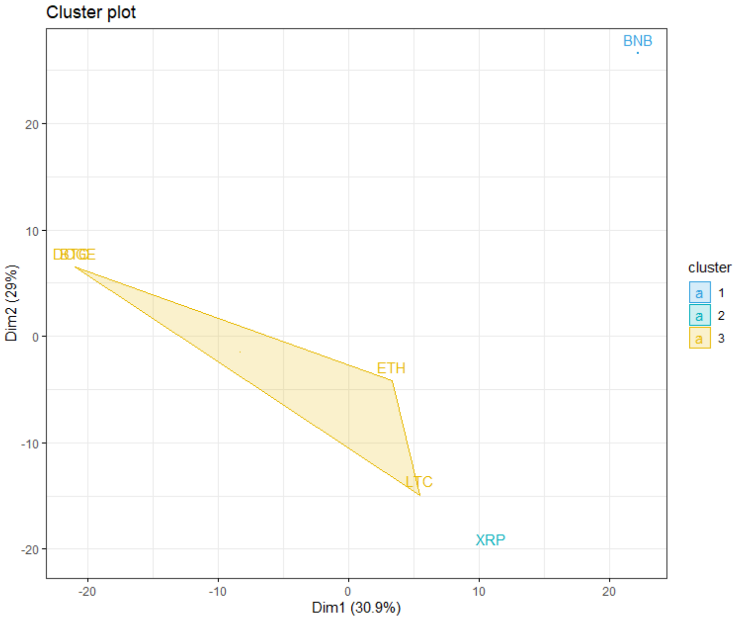
```
> sort(km1$cluster)
 BTC  BNB  ETH  LTC DOGE  XRP
   1    1    1    1    1    2
```

3 groups:



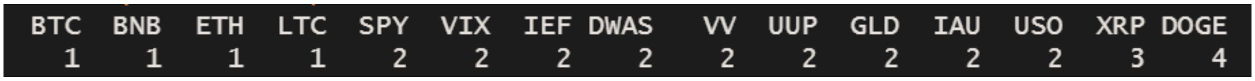
```
> sort(km1$cluster)
 BNB  XRP  BTC  ETH  LTC DOGE
   1    2    3    3    3    3
```

We also performed a visualization of the crypto cluster:



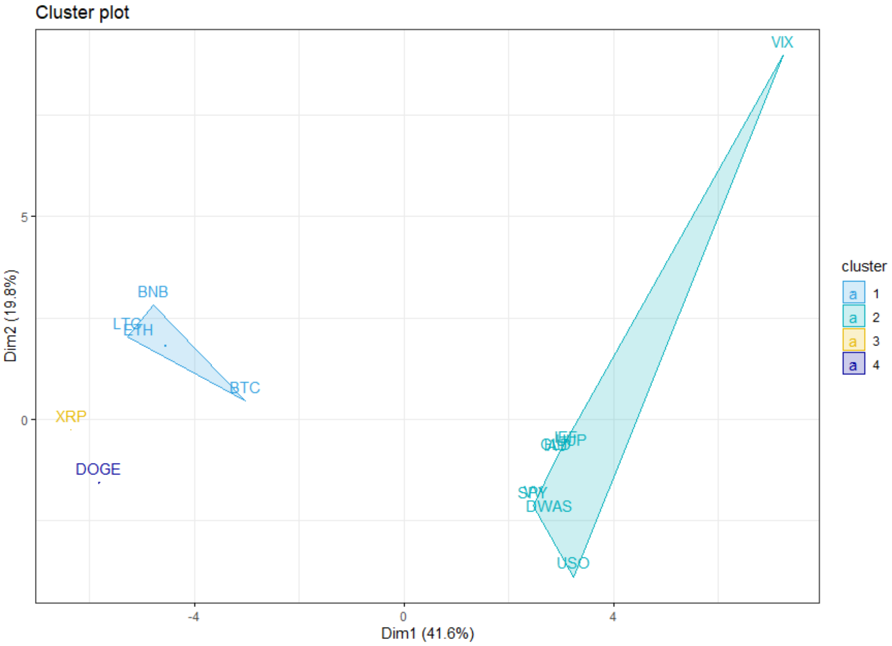
With the additional factors, we clustered with 4 groups. None of the side factors were grouped in with the cryptocurrencies, and Ripple and Dogecoin are in their own separate groups.

Clustering with additional factors:



```
 BTC  BNB  ETH  LTC  SPY  VIX  IEF DWAS   VV  UUP  GLD  IAU  USO  XRP DOGE
   1    1    1    1    2    2    2    2    2    2    2    2    2    3    4
```

Visualization with additional factors:



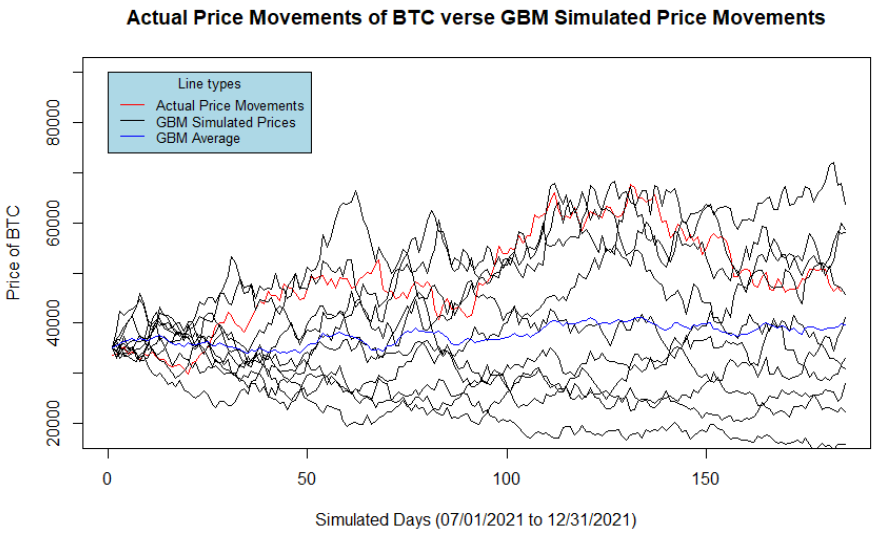
5.4 Cointegration

We performed a cointegration test to determine if there are multiple, if any, cointegration relationships in a system of 6 cryptocurrencies. We ran the Johansen trace test and gathered the results below. To interpret the results we examine the section of the console output labelled “Values of critical test statistic and critical values of test”. “r = 0” is the null hypothesis stating that there are no cointegration relationships present. To reject the null we must prove that the test statistic is larger than the critical values of the 1, 5, and 10 percent confidence level. Since this is true for “r = 0” we can reject the null. The same holds true for the null hypothesis that there are not 5 or more cointegration relationships. We also ran the same test on multiple systems of pairs where every possible combination of cryptocurrency pairs has been accounted for. All of these tests suggested that there are cointegration relationships between each pair. We also incorporated our side factors into the cointegration model to determine which factors are most influential in the price movements of cryptocurrencies. We found that UUP, SPY, and USO are the most influential.

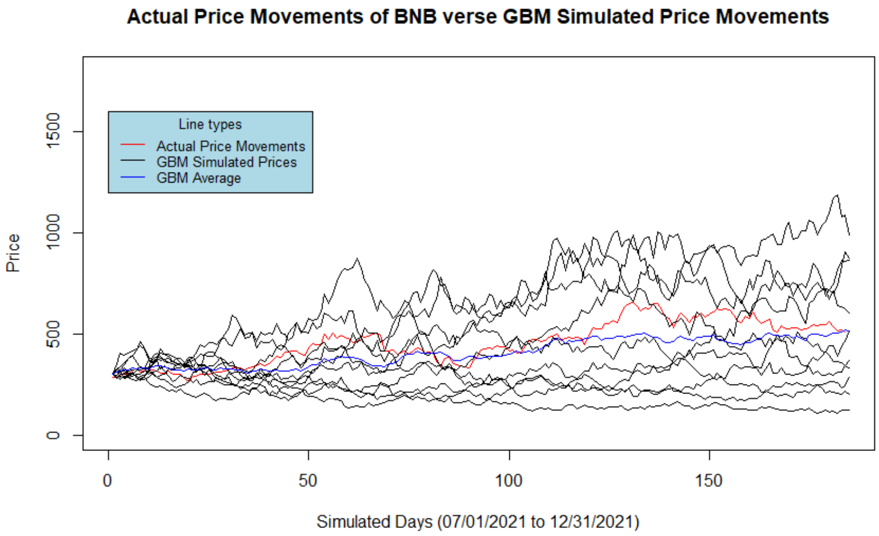
5.5 Geometric Brownian Motion

Geometric Brownian Motion was performed on the selected group of cryptocurrencies. Each graph displays the actual price movements of the cryptocurrency over the time period, the first 10 simulations from the model, and the theoretical average of 1,000 simulations.

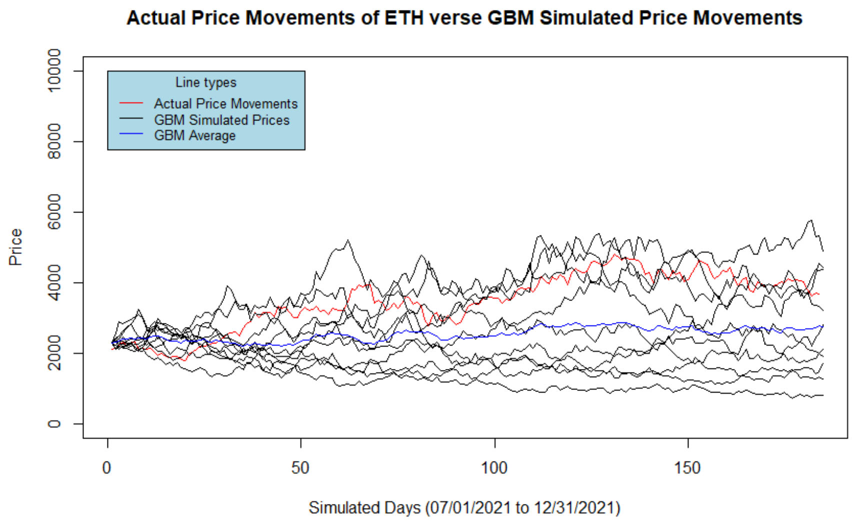
1. **BTC Simulation**



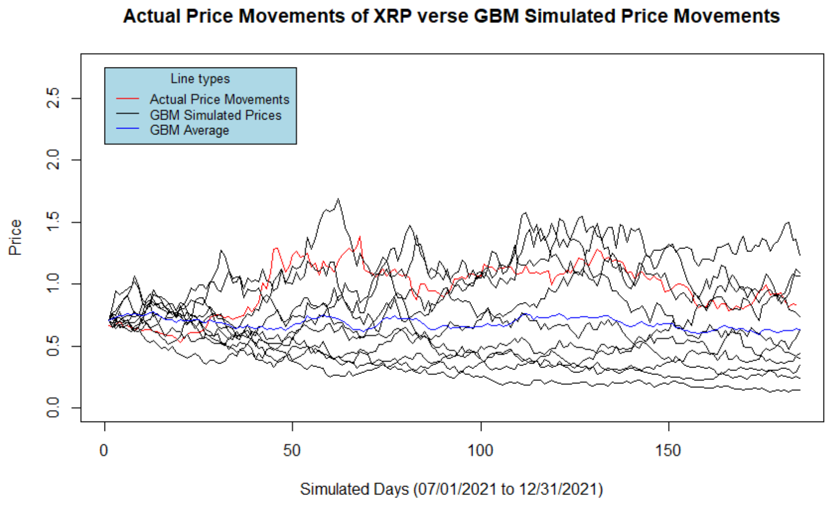
2. **BNB Simulation**



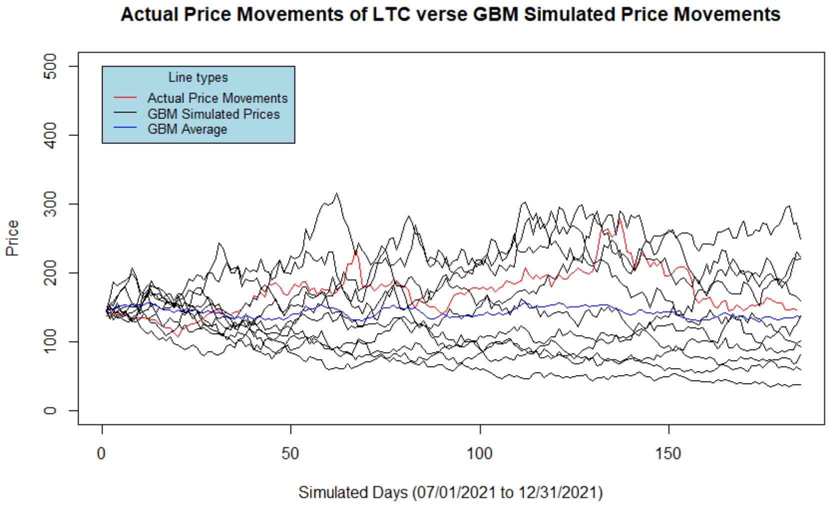
3. ETH Simulation



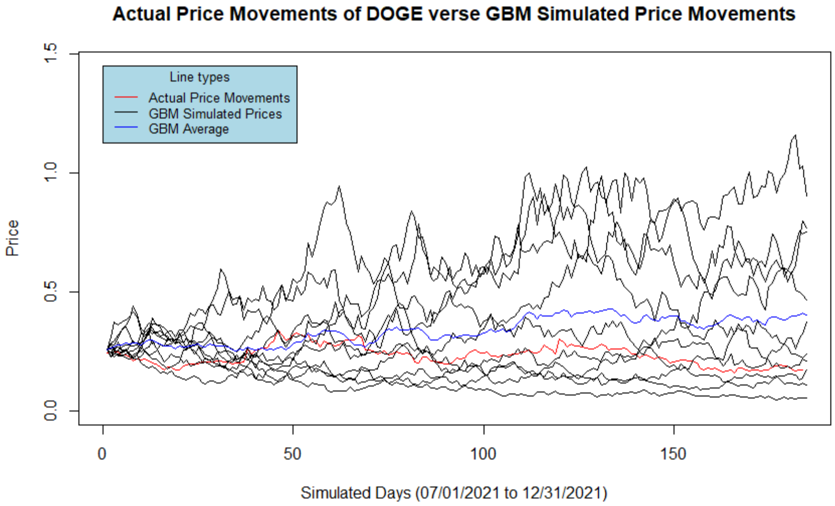
4. XRP Simulation



5. LTC Simulation

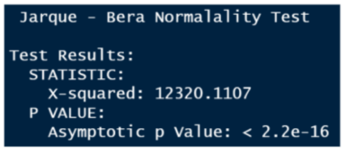


6. DOGE Simulation



The simulations show the inaccuracies of the Geometric Brownian Motion on the entire group of cryptocurrencies. The mean square error for BTC, BNB, ETH, XRP, LTC, & DOGE is 0.001297681, 0.002241178, 0.002164144, 0.003012867, 0.002740599, & 0.003690543 respectively. Each time series of the logarithmic returns were tested for normality using the Jarque-Bera test for normality. The test has the null hypothesis that the log of the returns follows a normal distribution and the alternative hypothesis is that the log returns do not follow a normal distribution. The results of the normality tests can be shown below:

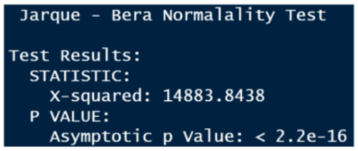
BTC



```
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 12320.1107
  P VALUE:
    Asymptotic p Value: < 2.2e-16
```

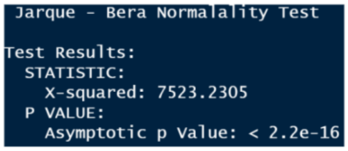
BNB



```
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 14883.8438
  P VALUE:
    Asymptotic p Value: < 2.2e-16
```

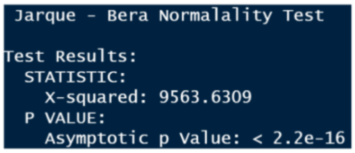
ETH



```
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 7523.2305
  P VALUE:
    Asymptotic p Value: < 2.2e-16
```

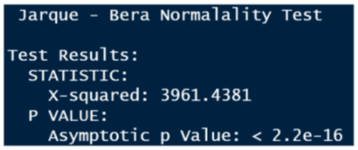
XRP



```
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 9563.6309
  P VALUE:
    Asymptotic p Value: < 2.2e-16
```

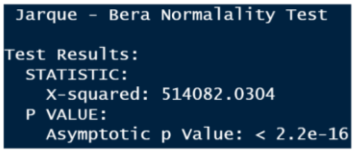
LTC



```
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 3961.4381
  P VALUE:
    Asymptotic p Value: < 2.2e-16
```

DOGE



```
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 514082.0304
  P VALUE:
    Asymptotic p Value: < 2.2e-16
```

The results of the test for normality show that all of the p-values were statistically significant. Therefore, the assumption that the logarithm of the returns follow a normal distribution can be rejected.

6. Analysis

The logistic regression for the cryptocurrencies highlights a trend in the data. This trend involves 5 of our cryptocurrencies sitting around a 50% level of accuracy for their respective models, which is essentially just a coin flip. Unlike the previous study that explored trends in the cryptocurrency market, this methodology proves that the returns of these cryptocurrencies are just a guess of random chance. The lone cryptocurrency that exceeded the 50% level of accuracy was BNB, which had an accuracy of 73.08%. Moreover, the rest of the cryptocurrencies were fairly spread out in terms of the down and up days that they predicted whether the model prediction was correct or not. As seen in the confusion matrix, the model was correct and incorrect in its predictions for up days and down days roughly the same, hence why the accuracy is roughly 50%. The only one of our logistic regressions that give insightful information is BNB. With an impressive 74% accuracy the model did have its faults in getting the up prediction wrong 97 times and the down prediction wrong 83 times, but the model was correct in its 1 and 0 predictions 271 times and 243 times respectively.

It appears that out of all of our models the logistic regression model for BNB was the most accurate and presents an intriguing money-making opportunity. This could be explained by the fact that BNB had the highest Kurtosis compared to the other cryptos in this study. A higher kurtosis suggests that an investor might expect to experience a higher frequency of higher returns (positive or negative). Since BNB has the highest Kurtosis it suggests that its distribution does have more significant return swings as the other cryptocurrencies. This reason suggests that further research of the cryptos that have similar values of kurtosis could potentially lead to similar logistic regression models in terms of accuracy as that of BNB. Although the other cryptocurrencies had above 50% accuracy, these models still perform up to par with our intuition that cryptocurrency returns follow a pattern of simple random chance. This model demonstrates that cryptocurrencies embody a similar functionality and interoperability of time series analysis and classification as the equity market.

The Prophet model shows how well the model Prophet fits these six cryptocurrencies. After comparing the mean squared error for each model's predictions, the one with Bitcoin turns out to have the highest accuracy with mean squared error of 0.0028. Whereas the model with Ripple has the least accuracy with mean squared error of 0.0051. The mean squared errors, the graphs, and the Garch models all show that the Prophet model fits the cryptocurrencies time series data pretty well, and they are able to create solid price predictions. In addition to these results, the seasonality & trend graphs show that all six cryptocurrencies have a general upward trend in the recent two years, which shows a general trend of interest in the market. What's more interesting is that all six cryptocurrencies tend to have a higher return in the second quarter of each year. This result may affect investors' decision-making when considering investing in cryptocurrencies.

K-means clustering gives us insight into similarities within the data. Without the additional factors, we clustered with 2 groups and 3 groups. In both cases, Ripple (XRP) is in its own group. This coin has seen a bit lower return compared to other coins such as bitcoin, so this makes sense. In the three-group model, BNB is also in its own group, and BTC, ETH, LTC, and DOGE are grouped together. Clustering in this way can help to show us which securities move together. BTC and ETH being grouped together makes sense since they are the two most popular cryptocurrencies, but LTC also being included with those two shows that it may have a similar profile.When side factors were added, the cryptocurrencies were entirely separate from the other securities. We used a variety of security types and sectors, but nothing else seems to have a similar return profile to the cryptocurrencies. This further cements how unique they are, and emphasizes why this research is necessary. Overall, clustering allows us to view our other results with the added context of similarity within our data.

Cointegration tests grant an understanding of the relationship between cryptocurrencies. If the test statistic is greater than the critical values, we can reject the null hypothesis that there are no cointegration relationships within the system. Since this is true for "r = 0" we can reject the null. The same holds true for the null hypothesis that there are not 4 or more cointegration relationships. The results of the cointegration test on the system of the five cryptocurrencies we are studying suggests that all of the currencies are cointegrated with one another. The same test was run on independent systems of possible cryptocurrency pairs. All of these tests suggested that there are cointegration relationships in each pair. The most influential relationships are all currencies excluding BTC are heavily influenced by BTC and that BNB is influenced by both ETH and LTC. We also incorporated our side factors into the cointegration model to determine which factors are most influential in the price movements of cryptocurrencies. We found that UUP, SPY, and USO, are the most influential. These side factors should be considered by investors when making investment decisions regarding cryptocurrencies.

The traditional pricing model that is typically used for pricing financial instruments based on the Black and Scholes model, which is fundamentally constructed off Geometric Brownian Motion proved to be a poor model of prediction. It is evident that ⅚ of the theoretical averages from the predictions finished below the actual price movements. The outlier of Dogecoin had its theoretical average finish above the actual price movement. All of these models were hurt by the high level of volatility in the data and the level of kurtosis in the log returns. The mean square errors that were computed from Geometric Brownian Motion, on average, were 5 times larger than that of the mean square error for Prophet. This demonstrates the Prophet model’s predictive power that it has over the traditional pricing model that is used to price securities in the equity market.

7. Conclusion

Throughout the research study, the fact that the select group of cryptocurrencies had extreme levels of returns influenced the results in the study. This idea was shown through the high levels of kurtosis and the high levels in volatility over the testing and training period. The kurtosis in the distributions led to the skewness of the returns. BNB had the highest kurtosis among the group of cryptocurrencies. BNB proved to have some of the most predictive power in the Logistic Regression and Prophet model. Additionally, through K-means clustering it was determined that cryptos and other types of securities would not be grouped together, which is a result of the highly volatile returns of cryptocurrency. In GBM, which requires a log-normal distribution, BNB performed the worst out of the group of cryptocurrencies. GBM as a whole displayed that each of the select groups of cryptocurrency does not follow a normal distribution. Compared to the mean squared errors of GBM, Prophet models’ mean squared errors on average are 5 times smaller. In general, Prophet seems to be a very solid tool to forecast future crypto prices and analyze crypto trends and seasonalities. From the results of GBM and Prophet, cryptos seem to be highly volatile, although long-term trends can be captured, day-to-day price fluctuations are still highly unpredictable. Due to the high levels of volatility, risk-averse investors should not invest in the cryptocurrency market.

In the future, many things can be done to extend the research horizon. Different cryptocurrencies can be added to our model as there are many different types of cryptocurrencies with different utilities like stable coins, currencies to facilitate smart contracts, and decentralized finance, and thus, exhibit different pricing behaviors. The introduction of Dogecoin showed that some of our models can be heavily skewed when currencies outside of the typical store of value utility are introduced. Introducing these different types of cryptocurrencies will allow us to better predict any type of cryptocurrency in the future. Different side factors like the data from different sectors of stock can be analyzed and utilized for the Prophet model. This may improve the predictability of Prophet and help understand more about the driving factors of cryptocurrency. Future studies can also explore how well Prophet is by comparing it with other predictive models. GBM is classic but not well known for predicting future fluctuations. Although Prophet beats GBM, it is not necessarily the best model for crypto predictions. Lastly, the time period that the cryptocurrencies in the study could be expanded. Due to the limitations in the availability of the data, the time series models could only use 3 years of data for the purposes of training.

References

Fabozzi, Frank J. 2014. The Basics of Financial Econometrics : Tools, Concepts, and Asset Management Applications. The Frank J. Fabozzi Series. Hoboken, New Jersey: John Wiley & Sons.

Yenidogan, I., Cayir, A., Kozan, O., Dag, T., & Arslan, C. (2018). Bitcoin forecasting using Arima and prophet. *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. https://doi.org/10.1109/ubmk.2018.8566476

Phillips, Ross C., and Denise Gorse. “Predicting Cryptocurrency Price Bubbles Using Social Media Data and Epidemic Modelling.” *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017. https://doi.org/10.1109/ssci.2017.8280809.

Pichl, Lukáš, and Taisei Kaizoji. “Volatility Analysis of Bitcoin Price Time Series.” *Quantitative Finance and Economics* 1, no. 4 (2017): 474–85. https://doi.org/10.3934/qfe.2017.4.474.

Kwon, Do-Hyung, Ju-Bong Kim, Ju-Sung Heo, Chan-Myung Kim, and Youn-Hee Han. “Time Series Classification of Cryptocurrency Price Trend Based on a Recurrent LSTM Neural Network.” Journal of Information Processing Systems. Korea Information Processing Society, June 1, 2019. http://jips-k.org/q.jips?cp=pp&pn=680.

Qureshi, Saba, Muhammad Aftab, Elie Bouri, and Tareq Saeed. “Dynamic Interdependence of Cryptocurrency Markets: An Analysis across Time and Frequency.” Physica A: Statistical Mechanics and its Applications. Elsevier B.V., August 19, 2020. https://www.sciencedirect.com/science/article/abs/pii/S0378437120305641?via%3Dihub.

Liu, Weiyi, et al. “Common Risk Factors in the Returns on Cryptocurrencies.” *Economic Modelling*, vol. 86, 2020, pp. 299–305., doi:10.1016/j.econmod.2019.09.035.

Zhu, Panpan, et al. “Investor Attention and Cryptocurrency: Evidence from the Bitcoin Market.” *PLOS ONE*, vol. 16, no. 2, 2021, doi:10.1371/journal.pone.0246331.

Yukun Liu & Aleh Tsyvinski, 2018. "Risks and Returns of Cryptocurrency," NBER Working Papers 24877, National Bureau of Economic Research, Inc.

Goodkind, A. L., Jones, B. A., & Berrens, R. P. (2020). Cryptodamages: Monetary value estimates of the air pollution and human health impacts of cryptocurrency mining. *Energy Research & Social Science*, *59*, 101281. https://doi.org/10.1016/j.erss.2019.10128

“Cointegration.” Corporate Finance Institute, 1 May 2020, https://corporatefinanceinstitute.com/resources/knowledge/other/cointegration/#:~:text=Cointegration%20is%20a%20technique%20used,and%20the%20Phillips%2DOuliaris%20test.

Malladi, R.K., Dheeriya, P.L. Time series analysis of Cryptocurrency returns and volatilities. J Econ Finan 45, 75–94 (2021). https://doi.org/10.1007/s12197-020-09526-4

Fama, Eugene F., and Kenneth R. French. “Common Risk Factors in the Returns on Stocks and Bonds.” *Journal of Financial Economics* 33, no. 1 (1993): 3–56. https://doi.org/10.1016/0304-405x(93)90023-5.

Ďurka, P., & Pastoreková, S. (2012). ARIMA vs. ARIMAX – which approach is better to analyze and forecast macroeconomic time series? . *Proceedings of 30th International Conference Mathematical Methods in Economics*.

Yenidogan, Isil, Aykut Cayir, Ozan Kozan, Tugce Dag, and Cigdem Arslan. “Bitcoin Forecasting Using Arima and Prophet.” *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 2018. https://doi.org/10.1109/ubmk.2018.8566476.

Vosoughi, AliReza. *ON THE RELATIONSHIP BETWEEN STOCK RETURNS AND EXCHANGE RATES: TESTS OF GRANGER CAUSALITY*, https://www.academia.edu/4696166/ON_THE_RELATIONSHIP_BETWEEN_STOCK_RETURNS_AND_EXCHANGE_RATES_TESTS_OF_GRANGER_CAUSALITY?auto=citations&from=cover_page.

Göttfert, J. (2019). Cointegration among cryptocurrencies : A cointegration analysis of Bitcoin, Bitcoin Cash, EOS, Ethereum, Litecoin and Ripple (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-161079

Interdax. (2020, July 24). *Seasonality in bitcoin: Examining almost a decade of Price Data*. Medium. Retrieved October 30, 2021, from https://medium.com/interdax/seasonality-in-bitcoin-examining-almost-a-decade-of-price-data-abf47b1421cb.