The Relationship of Percentage Change of between Altcoins and Bitcoin:

"Can movements of the top Altcoins predict Bitcoin's percentage change?"

Lamae A. Maharaj

Honors Seminar Thesis

in

Economics & Crypto Assets

For:

City University of New York: Queens College

Department of Economics

2022

Acknowledgements:

I want to extend a sincere thanks to my advisor, Professor Jennifer Roff of the CUNY Queens College Department of Economics. She has advised me on this thesis and helped me tremendously. Dr. Roff has taught me how hard research can get. The lesson that she has given me in these couple of months is that sometimes we can find a diamond by looking for a ruby. Without her coaching, I want to say that this would not have been successful. I want to say many thanks to Dr. Adam D. Kapelner of the Department of Mathematics for giving me the inspiration and knowledge on machine learning and data science. I would also like to thank Dr. Suleyman Taspinar for his Python exercises and my first exposure to econometrics. Many thanks to the Department of Economics at CUNY Queens College. I want to dedicate this research paper and my efforts to both my parents, my little sister and all my loved ones. Without their love and support, none of this would have been possible.

The Relationship of Percentage Change of between Altcoins and Bitcoin

"Can movements of the top Altcoins predict Bitcoin's percentage change?"

By: Lamae Maharaj

May 2022

Abstract

This paper explored the relationship between the price percentage movement of the altcoins (Ethereum, Solana, Ripple, Cardano, and Binance Coin) and its effect and its use in prediction Bitcoin's percentage price movement. Recent literature suggests that Bitcoin's performance can manipulate market movement in the investment world of crypto assets. This paper reverse engineers the idea of smaller assets manipulating the price of Bitcoin through complex algorithms and traditional statistical modeling techniques. The goal is to find a relationship between Bitcoin and altcoins to increase prediction power of the models discussed.

1. Introduction

This paper examines the percentage change of the top 5 altcoins and the correlation of these changes with Bitcoin. Altcoins are cyrptocurrencies that are not Bitcoin, including Ethereum, Solana, Ripple, Binance Coin, and Cardano. I am interested in exploring the relationship between movements in price changes among cryptocurrencies. The percentage change in price will be the main variable within this experiment. This feature does not exist within the dataset as a given variable.

This variable will be engineered from within the features of the original data set.

Utilizing Yahoo Finance API via Python packages, I first explore data patterns in pricing between Bitcoin and the top give altcoins. I subsequently employ models with increasing predictive power via machine learning algorithms such as the Linear Regression, Random Forest, and Polynomial Regression. Exploring decision trees within the Random Forest Regression algorithm can create robust predictive models used in the financial industry and other advanced algorithms within the spectrum of machine learning.

This paper proceeds as follows. Section 2 will discuss previous experiments like the one discussed in this paper. Similar methodologies are explored together with theoretical ideologies that influenced this experiment. Section 3 will discuss the data and the data wrangling techniques. Section 4 will show the mathematical methodologies used in the experiment. Section 4 will analyze machine learning models that can produce predictions using percentage change of the crypto assets. Section 5 will be the results of the experiment together with selecting the best model for predictive analytics. The model comparison based on Root Mean Squared Error and R-Squared values will determine how well the model can be used to predict this phenomenon. Section. 6 will be the conclusion and in Section 7 will hold all visuals and tables that hold results for regressions and other models.

2. Literature Review

Ciain (2018) experiments on trading volumes within the selected coins and currency compensation to create a table for descriptive statistics. The table later categorizes which coin has a correlation and which does not. However, the intensity of the relationship is not a factor on the table. Discussions about coin mining play a role within the experiment but minimally. Coins such as Dogecoin and PeerCoin stay in the analysis, but these coins were popularly known to be meme tokens. Meme tokens are foundational only due to their social media popularity.

Yang (2020) experiments on modeling and quantifying the underlying characteristics of cryptocurrencies. The experiment consisted of Noise-Assisted Multivariate Empirical Mode Decomposition (NA-MEMD) to interpret currency formations and behaviors within respected groups. The experiment gave birth to a view of investment strategies that grouped coins to invest in based on their behaviors. Important variables include investor sentiment, micro-structure, effects of significant events, and fundamental values in trading.

Aysan (2020) analyses long-term relations between currencies. The experimental team observed closing price data to build investment strategies. The experiment becomes critical to risk analysis, and investments by grouping assets come into play. The versatile capabilities of investment firms are enhanced since the grouped assets are all labeled by high and low volatility. Such ideas are deemed successful as they are useful in algorithmic trading.

3. Data

I collected data from Yahoo Finance. It was extracted from the Yahoo Finance API and accessible through the Python package yfinance. The data is in a time series format and consists of prices between the dates of January 1st, 2018, up until May 1st, 2022. This format consists of Bitcoin, Ethereum, Solana, Ripple, Binance Coin, and Cardano prices. These are the top five most popular altcoins within the market.

The data was orderly since it is pulled from the Yahoo Finance application programming interface. Some columns were dropped, such as Index Values, opening prices labeled as Open, daily highs and lows, dividends, stock splits, and an extra column of dates. The remaining columns include the index labeled dates and Close consisting of the daily close of the coin's price. The percentage change from the function within the Pandas library helps find the metric needed for further calculation. This function is the pandas.pct_change(), which takes the previous day's closing, and its percentage change is now assigned to a new column. Some rows were dropped due to NA values due to errors. These NA values occurred in mainly in the prices of Solana since Solana was only released in April of 2019.

The data lived in a relational database built on the popular SQL (Structured Query Language). The database was hosted and successfully built on the PostgreSQL servers to store data orderly. The uncleaned data sums up to be around 300 megabytes. Nevertheless, keeping the cleaned data safe and away from computer crashes and other permanently damaging situations. The pandas.info() function shows the data types that the values were to proceed with the analysis. The prices labeled "Closed" and percentage change labeled "CoinNamePCH" were float values. Floats are numbers than allow to carry decimal values instead of integer data type, which allows only integers. After most data wrangling, the respective percent changes join a new data frame where only each coin's dates and percentage change are ready for analysis.

A statistical table is shown in Section 7A. The count of each variable shows a total of 751 observations. The crypto currency with the highest percentage change is Binance Coin (BIN) which is around 69%. The minimum price change across the time frame is Bitcoin which is only at an 18% price percentage change.

In Section 7A, the coin with the highest standard deviation is Solana (SOL). This coin came out much later than the others. The crypto coin that had the lowest price drop is Ripple (XRP) which was -42.33%. There are many outliers within this data set. However, if the models that are being made were to be in use for future predictions, there would be lesser room for error in extrapolation. Extrapolation is where a model is attempting to predict for values that it has never been trained to deal with. Most models are interpolated models and predict based on data it was built on.

4. Mathematical & Modeling Methodology

The analysis begins by examining the relationship of the percentage change of all coins. The regression tool from the statsmodels API calculates regressions. The equation begins with altcoin change in percentage multiplied by some x which equals Bitcoin’s change in percent as follows:

*Bitcoin*D% = *Ethereum*D%*x

*Bitcoin*D% = *Solana*D%*x

*Bitcoin*D% = *BinanceCoin*D%*x

*Bitcoin*D% = *Cardano*D%*x

*Bitcoin*D% = *Ripple*D%*x

The formula below shows the regression model that is used on the table in Section 8B:

BitcoinD% = EthereumD%*x + SolanaD%*x + BinanceCoinD%*x + CardanoD%*x + RippleD%*x

This model is entered in the Ordinary Least Squares function statsmodels.OLS.fit() to fit the needed values and produce a summary. Other methodologies are used for prediction purposes. The next algorithm shown in Section 7B1 is known as the Random Forest Regression. This algorithm is a supervised learning method that makes use of the trees to calculate. The number of trees produced within this model to decide of based on the best prediction model. These trees are constructed in random way as given the name Random Forest. Each tree is made from sampled rows. At each node a different sample of features are selected for splitting the data for training purposes. This algorithm uses a technique called bagging where an ensemble algorithm fits multiple models and selected the best fit.

The next algorithm that is tested within this data set is the polynomial regression. This algorithm factors in the relationships between the independent variable and the dependent variable and models the data to the nth degree polynomial in respect to x. The polynomial usually finds a pattern very quickly as the candidate set of functions are expanded from the function set of linear models. These error types also make the algorithms above vulnerable. Typical polynomial models are depicted as such in Section 7B2.

These models will be trained with 80% of the collected data and be tested with 20% of the collected data. To test model accuracy and allow the model to learn from the data, sklearn's library called model_selection will be in use. The function called train_test_split will split the data frame in two pieces of data. One being the training set and one being the testing set. The training set will consist of two parts. The X_train will be the x values consisting of price percentage change of the top five altcoins and y_train will be y values of the price percentage change of Bitcoin. The model will be fit to these two pieces of the split. Then the X_test will be predicted by entering it back into the trained model to produce the y_predict. The newly y_predict variable will be used in the

RMSE calculation to test for prediction accuracy for each of the models. A picture of the how the data can be split for testing is shown in Section 7B3.

All models will be compared to each other by using two important metrics. The first being R-Squared. The R-Squared metrics shows the level of variance for the dependent variables in a regression model. Below the SSE (Sum of Squared Regression Error) is the difference between the observed values and the predicted values. The Sum Squared Total Error is the squared difference between the observed variable and its mean.

The second metric will be the RMSE (Root Mean Squared Error). The RMSE metric shows the standard deviation of prediction errors. The RMSE is the square root of the sum of differences between the squared value of the predicted and the actual values divided by the number of observations. The formula for the RMSE is shown in Section 7B5. This metric used for predictive models throughout machine learning and statistical modeling.

5. Results

The visualizations below show Bitcoin's percentage change together with the price percentage change of the comparing altcoins overt the same time frame in the data above. Bitcoin is showed in orange while Ethereum is showed in the blue. The first visual in Section 7C shows Ethereum's movement, then Solana, Ripple, Binance Coin and finally Cardano. The visuals show the price percentage change frequencies of Bitcoin and all other altcoins. The visuals show the change in percentage of price as the y axis over the time labeled in the x axis.

The results show that the linear model that the R-Squared shows a 62.9% which is the percentage of how well observed outcomes are replicated with the created model. This metric shows that the percentage change based on price can be a useful mode. The coefficients are shown below together with a section of the regression table. ETHD% = 0.4470, which is the largest coefficient amongst all the coefficients in the model. The linear model is shown below as:

BTCD% = 0.447(ETHD%) + -0.0280(SOLD%) + 0.0516(ADA%D) + 0.1044(BIND%) + 0.081(XRPD%).

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0150	0.082	-0.182	0.856	-0.177	0.147
ETH	0.4470	0.027	16.781	0.000	0.395	0.499
SOL	-0.0280	0.012	-2.396	0.017	-0.051	-0.005
ADA	0.0516	0.019	2.758	0.006	0.015	0.088
BIN	0.1044	0.018	5.750	0.000	0.069	0.140
XRP	0.0481	0.015	3.268	0.001	0.019	0.077

This newly found model can show relationships to the BTCD%. As the percentage of change of the independent crypto currencies move up, the larger the impact the movements make on the percentage change of Bitcoin's price. The table below shows the summary of the linear regression from the OLS model. The Root Mean Squared Error is the standard deviation of the prediction errors. The RMSE returned to be 2.25 for this linear model. The statsmodels API python package was used to create the OLS linear model and its summary. The summary can be found in Section 7D1.

The next model that will be the polynomial regression. The results table is shown in Section 7D2 with the created model. This model is less useful than the linear model. The R-Squared of the polynomial regression model is at 0.9% which shows that expanding the candidate set of functions to fit the data is not needed. Therefore, the linear regression model seems to be the best fit model between the two. The data is not well represented by this model. It is the case that when the degrees of polynomials increase the model becomes less useful. The RMSE for this model returns to be 3.6887 as opposed to the much better 2.25 produced by the linear model above. The statsmodels API python package was used to create the OLS polynomial model and its summary.

The Random Forest Regression Algorithm produces coefficients that are different from the linear model above. This algorithm was taken out of the sklearn library and imported as the Random Forest Regressor. The coefficients for ETHD% is 0.67794654, ADAD% is 0.10599741, XRPD% is 0.06338565, SOLD% is 0.05599518, and finally BIND% is 0.11447866. The R-Squared for this model is 87.86%. This means that through the Random Forest Algorithm, a better fit model was found to predict the price percentage change of Bitcoin. The Root Mean Squared Error is the standard deviation of the prediction errors. The RMSE for this model returns to be 2.173839. This shows that the Random Forest Algorithm outperforms the linear model and the polynomial regression model. The model produced by the Random Forest is shown below. The table for this model is found in Section 7E3.

***BTC*D% = 0.67794654(*ETHD%*) +0.10599741(*ADAD%*) + 0.06338565(*XRPD%*) + 0.05599518(*SOLD%*) + 0.11447866(*BIND%*)**

Based on the analysis, it is the case that price percentage change of the top five altcoins do influence Bitcoin’s price percentage change. Based on performance, the best prediction model for this dataset is the Random Forest. Although there is room for increasing performance level, the Random Forest Regression is the highest performing predicting model. The second-place prediction model is the Linear Regression model and lastly would be the polynomial regression model. What can be considered as a misspecification error in the polynomial regression is underfitting. The model may or may have not fit the data correctly as opposed to other models like the linear regression.

6. Conclusion

Based on the findings through regression models, Bitcoin's percentage change in price can be predicted by the price change in percentage of the top five popular altcoins. I have shown regression models that detect correlations with the movements of the altcoins. The machine learning models used, and the training split data technique code can be found on my GitHub page that is provided underneath the references. Predictive models for price percentage change can better be predicted with more technical models like the Long-Short Term Memory models. Due to the small-time frame and lack of resources, better models can be created for future purposes.

The interesting part of this research shows that the oldest crypto currencies are tied to Bitcoin's movement while the younger, Solana, is not too correlated like the others. This may not be the best family of coins to build a portfolio around, but for prediction purposes they can manipulate Bitcoin. I will continue to modify these models because it is my area of interest. However, I will proceed to use more advanced artificially intelligent machine learning algorithms to increase prediction accuracy. I may even retrain the models using as much data as possible. My findings currently suggest that the price percentage change of the top five altcoins can influence Bitcoin's price percentage change and even be useful enough to predict it.

7. Tables & Visualizations

7A: Statistics table over the entire data set.

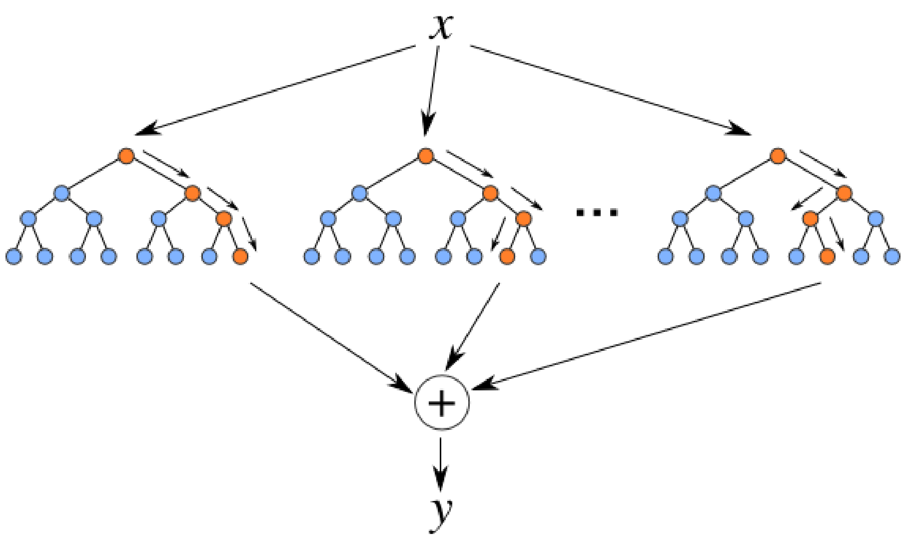
Date	BTC	ETH	SOL	ADA	BIN	XRP
2018-01-01	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-02	9.701106	14.470236	NaN	7.401285	5.028879	3.758630
2018-01-03	1.461080	8.850306	NaN	37.960389	7.899160	25.171107
2018-01-04	2.619566	1.890688	NaN	3.191741	-3.375565	2.938779
2018-01-05	11.733293	1.712468	NaN	-10.282646	61.897286	-4.627370
...	...	...	...	...	...	...
2022-04-27	2.947888	2.871182	2.711067	1.564681	1.546830	1.654823
2022-04-28	1.357522	1.661904	-0.196321	0.399118	3.901528	-1.219499
2022-04-29	-2.926557	-4.131479	-4.251519	-4.610106	-3.357616	-5.114509

7B: Regression Tables for Altcoins & Bitcoins

Regressions For Altcoins & BTC

	Dependent variable:				
	BTC				
	(1)	(2)	(3)	(4)	(5)
Ethereum	0.585*** (0.018)				
Solana		0.167*** (0.015)			
Cardano			0.353*** (0.018)		
Binance Coin				0.363*** (0.018)	
Ripple					0.273*** (0.017)
Constant	0.003 (0.086)	0.139 (0.124)	0.084 (0.109)	0.072 (0.107)	0.191* (0.115)
Observations	751	751	751	751	751
R2	0.593	0.144	0.341	0.359	0.267
Adjusted R2	0.593	0.142	0.340	0.358	0.266
Residual Std. Error (df = 749)	2.335	3.388	2.972	2.931	3.134
F Statistic (df = 1; 749)	1,092.648***	125.543***	387.566***	419.761***	272.711***
Significant Levels			*p<0.1; **p<0.05; ***p<0.01		

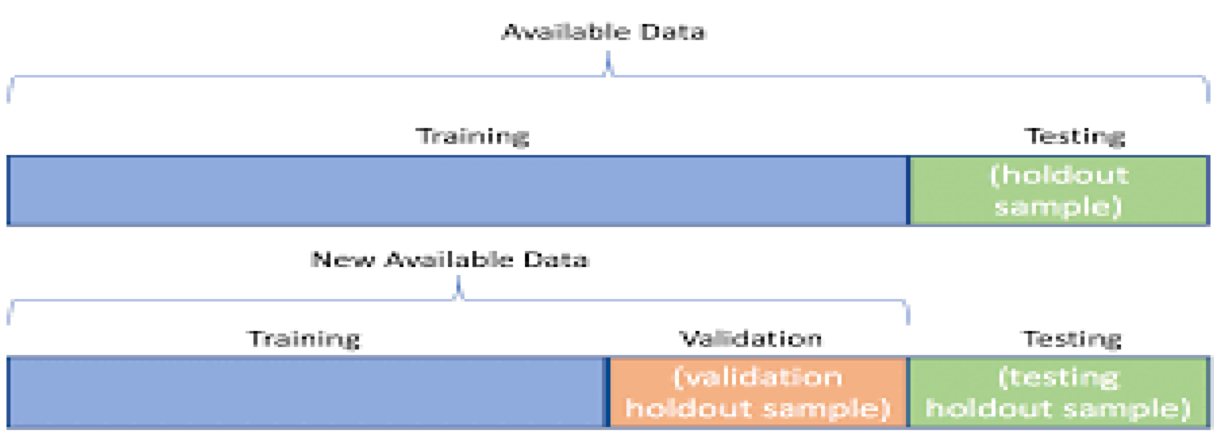
7B1: The Random Forest Regression Tree Model



7B2: Polynomial Regression Model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

7B3: Data Splitting Visual:



7B4: R-Squared:

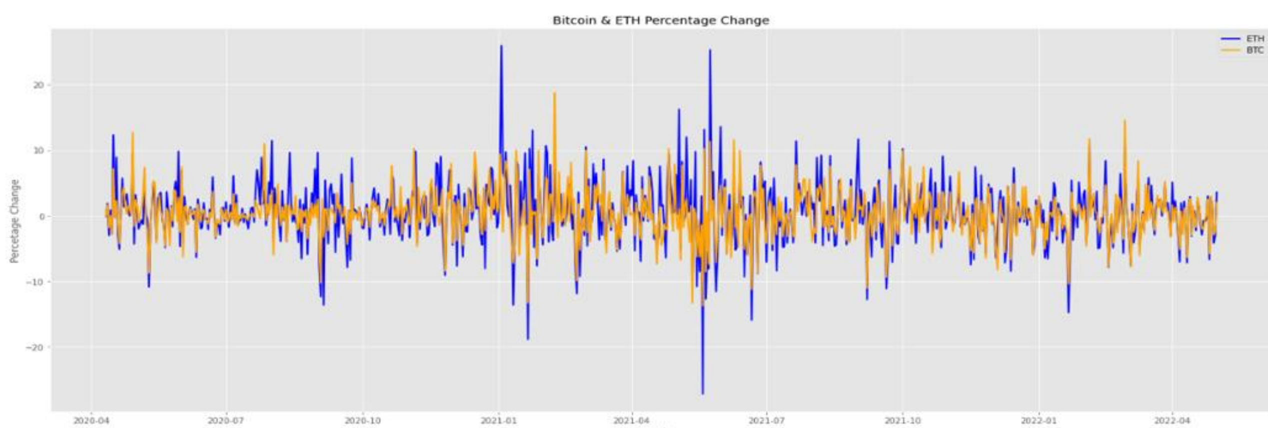
$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

7B5: Root Mean Squared Error:

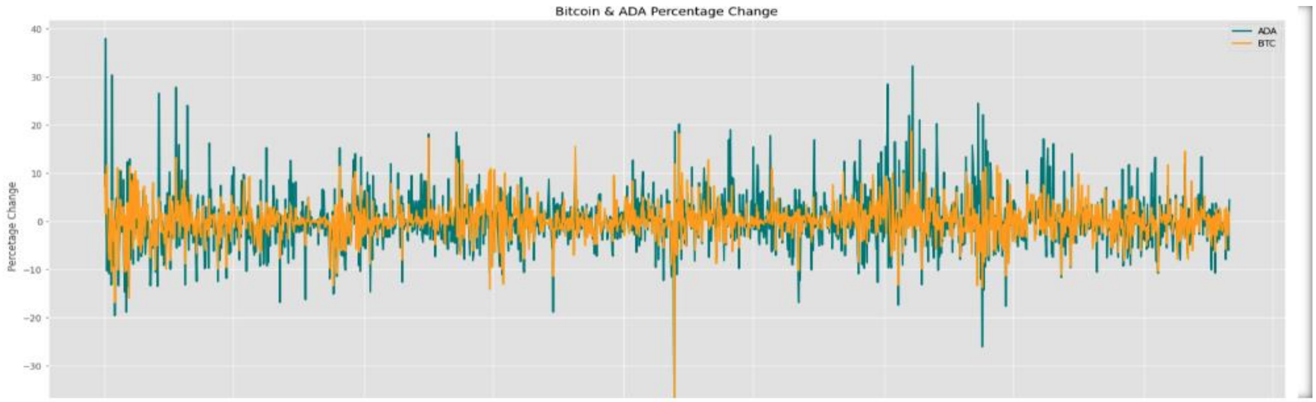
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

7C: Data Visualizations:

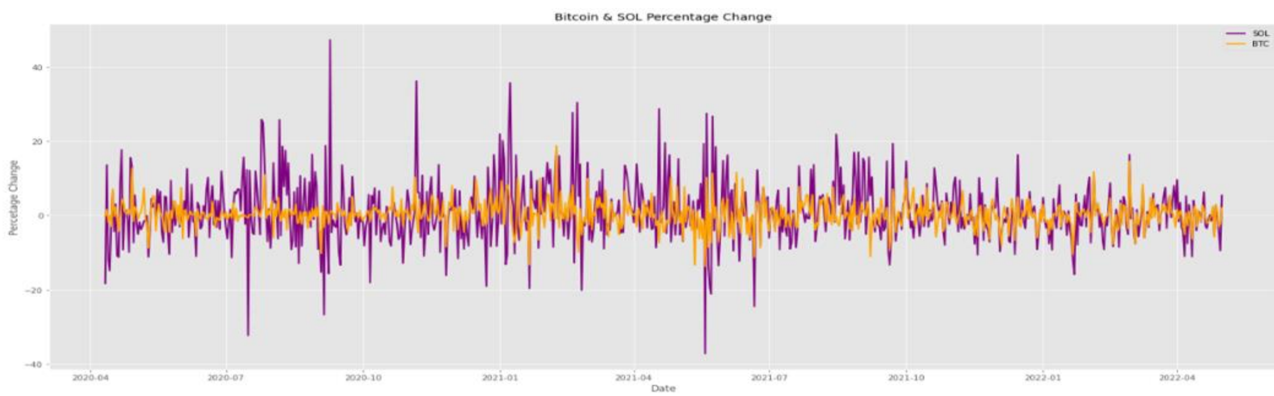
Bitcoin = Orange, Ethereum = Blue



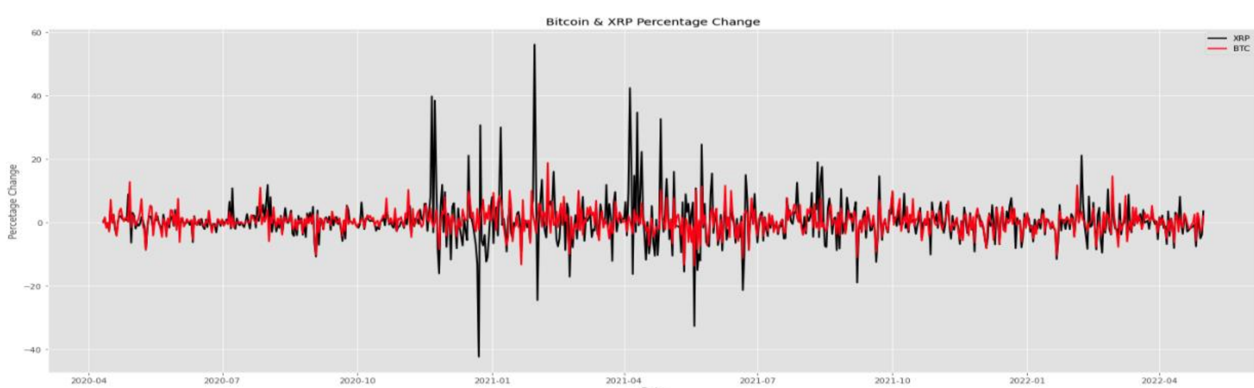
Bitcoin = Orange, Cardano = Teal



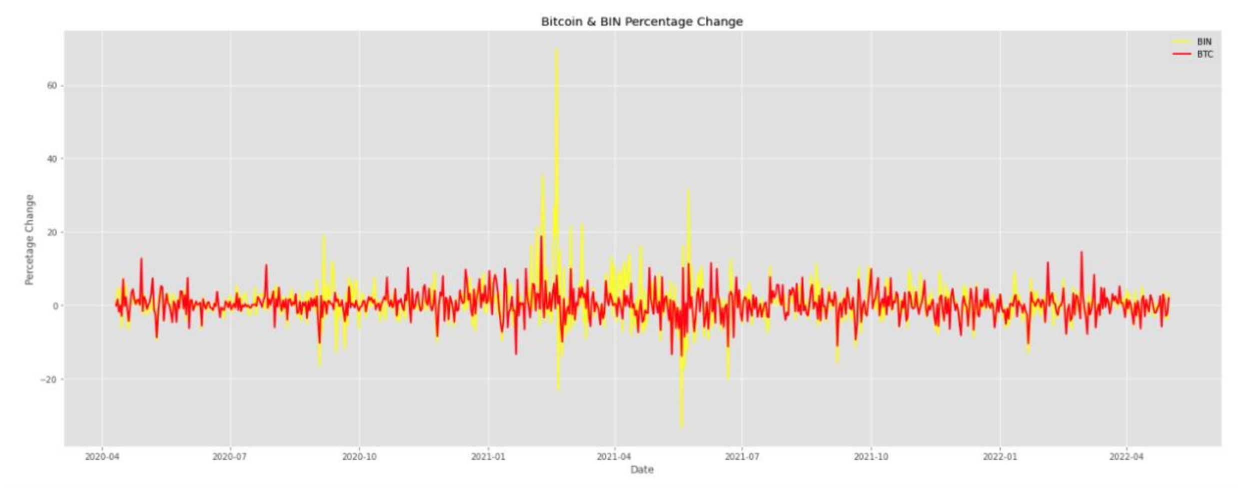
Bitcoin = Orange, Solana = Purple



Bitcoin = Red, XRP = Black



Bitcoin = Red, Binance Coin = Yellow



7D1: BTC Linear Regression Table with all Altcoins:

Regressions For Altcoins & BTC

	Dependent variable:
	BTC
Ethereum	0.447***
	(0.027)
Solana	-0.028**
	(0.012)
Cardano	0.052***
	(0.019)
Binance Coin	0.104***
	(0.018)
Ripple	0.048***
	(0.015)
Constant	-0.015
	(0.082)
Observations	751
R2	0.629
Adjusted R2	0.626
Residual Std. Error	2.236 (df = 745)
F Statistic	252.523*** (df = 5; 745)
Significant Levels	*p<0.1; **p<0.05; ***p<0.01

7D2: Polynomial Regression Table with BTC & Altcoins

OLS Regression Results

==

Dep. Variable:	**BTC**	**R-squared:**	**0.009**
Model:	**OLS**	**Adj. R-squared:**	**0.003**
Method:	**Least Squares**	**F-statistic:**	**1.387**
Date:	**Wed, 11 May 2022**	**Prob (F-statistic):**	**0.227**
Time:	**11:05:47**	**Log-Likelihood:**	**-2035.7**
No. Observations:	**751**	**AIC:**	**4083.**
Df Residuals:	**745**	**BIC:**	**4111.**
Df Model:	**5**		
Covariance Type:	**nonrobust**		

==

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	**0.4030**	**0.154**	**2.612**	**0.009**	**0.100**	**0.706**
I(np.power(ETH, 2))	**-0.0042**	**0.003**	**-1.483**	**0.139**	**-0.010**	**0.001**
I(np.power(SOL, 2))	**-0.0007**	**0.001**	**-0.791**	**0.429**	**-0.003**	**0.001**
I(np.power(ADA, 2))	**-0.0003**	**0.002**	**-0.177**	**0.860**	**-0.004**	**0.003**
I(np.power(BIN, 2))	**0.0014**	**0.001**	**1.985**	**0.047**	**1.57e-05**	**0.003**
I(np.power(XRP, 2))	**4.491e-05**	**0.001**	**0.063**	**0.950**	**-0.001**	**0.001**

7D3: Random Forest Regression Table with BTC & Altcoins:

Model:	Random Forest Regression
Coefficients:	Values:
ETH:	0.67794654
ADA:	0.10599741
XRP:	0.06338565
SOL:	0.05599518
BIN:	0.11447866
RSME:	2.173839
R-Squared:	0.8786

References:

Ciaian, P., Rajcaniova, M., & Kancs. (2018). Virtual relationships: Short- and long-run evidence from BitCoin and altcoin markets. *Journal of International Financial Markets, Institutions and Money*, *52*, 173–195. https://doi.org/10.1016/j.intfin.2017.11.001

Aysan, A. F., Khan, A. U. I., & Topuz, H. (2021). Bitcoin and Altcoins Price Dependency: Resilience and Portfolio Allocation in COVID-19 Outbreak. *Risks*, *9*(4), 74. https://doi.org/10.3390/risks9040074

Yang, B., Sun, Y., & Wang, S. (2020). A novel two-stage approach for cryptocurrency analysis. *International Review of Financial Analysis*, *72*, 101567. https://doi.org/10.1016/j.irfa.2020.101567

Demir, E., Simonyan, S., García-Gómez, C.-D., & Lau, C. K. M. (2020). The asymmetric effect of bitcoin on altcoins: evidence from the nonlinear autoregressive distributed lag (NARDL) model. *Finance Research Letters*, 101754. https://doi.org/10.1016/j.frl.2020.101754

Link to code:

https://github.com/lamaemaharaj/EconomicsResearchHonors