Crypto Prophet: Investing Intelligently

Group members -

- Tejaswini Bosukonda
- Atharv Belsare
- Zachary Wallace

Introduction

For many years, people have been interested in stocks and their subsidiaries, including cryptocurrencies. This, of course, includes the desire to find trends in the values of said things. If one can predict these values over time, one can effectively make large sums of profit. Cryptocurrency is especially interesting in this regard because of its high-value fluctuation.

Cryptocurrency had a recent appearance only around a couple of decades ago, making it a novel topic. This means that while data research has already been done on the subject, there is a limited scope of said research. Most of the work done today is focused on predicting the prices of cryptocurrencies. Our focus, on the other hand, is identifying correlations, trends, and market volatility in helping make informed investment decisions.

Dataset

We mine the time-series data (data that changes over a period of time) over the prices and statistics of the various cryptocurrencies in order to determine the trends and correlations. The dataset named "Cryptocurrency Historical Prices" is taken from the Kaggle community. It consists of 23 .csv files each representing a Cryptocurrency. Few of them are Bitcoin, Dogecoin, Litecoin, Ethereum, Tether, etc.
Each of the dataset has the following features:

- Date - Date of observation
- Open - Opening price on the given day
- High - Highest price on the given day
- Low - Lowest price on the given day
- Close - Closing price on the given day
- Volume - Number of stocks sold on the given day
- Market Cap - Market Capitalization in USD

Correlation Analysis (Tejaswini)

Why do we need price trend analysis for cryptocurrencies?

Identifying the correlations among the various cryptocurrencies can be vital information in making an informed investment decision.

It can be helpful in risk management. Since cryptocurrencies are highly volatile, identifying the correlations among the various cryptocurrencies can help the investors diversify their portfolio by investing in multiple cryptocurrencies which are not strongly correlated, thereby preventing the loss of money due to sharp changes in price of one cryptocurrency.

Also, it can be helpful in identifying trading strategies. We can study the behavior of cryptocurrencies and come up with strategies. For example, if two cryptocurrencies are strongly positively correlated, and if one of their prices is on an increasing trend, it makes sense to invest in the other also. On the

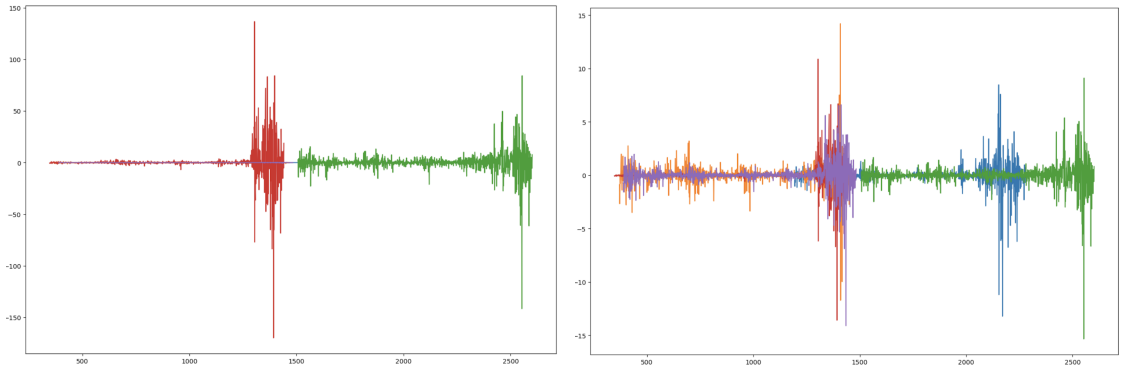
other hand, if they are strongly negatively correlated, investors may want to short one when the other is performing poorly.

Lastly, it can be helpful for arbitrage opportunities. The correlation information can provide opportunities for arbitrage traders to profit by exploiting price discrepancies between different cryptocurrencies. If two cryptocurrencies have a strong positive correlation, for example, and one is temporarily undervalued relative to the other, an arbitrage trader could buy the undervalued cryptocurrency and simultaneously sell the overvalued one to profit from the price difference.

Preprocessing and Analysis

Before we perform any techniques, we need to preprocess the data. We start by losing all the cryptocurrencies which have less than 3.5yrs of data. This is done because in order to identify a trend or establish a correlation, we would want significant history for better accuracy. We end up with a total of 15 cryptocurrency datasets after dropping the irrelevant datasents. We then pick the last 3.5 years of data from all the datasets so that it is uniform.

After this, we compute the price change between all the cryptocurrencies over the entire period. The price change is given as "Closing Price - Opening price". We base all of our further computations off of this price change because it provides a simple and intuitive measure of the degree to which the price of a particular cryptocurrency has changed over a given period of time. On visualizing this (Figure 1(a)) value (over five datasets), we see that the data has different scales.



(a) *(b)*

Figure 1: (a) Various scales of price change values (b) Normalized price change values

Therefore, we go ahead and normalize all the datasets using StandardScaler() method. Figure 1(b) shows the normalized datasets.

To identify the correlations among all the datasets, we create a dataframe that consists of the price change values of all the cryptocurrencies (as columns) over the entire 3.5 year period (as rows).

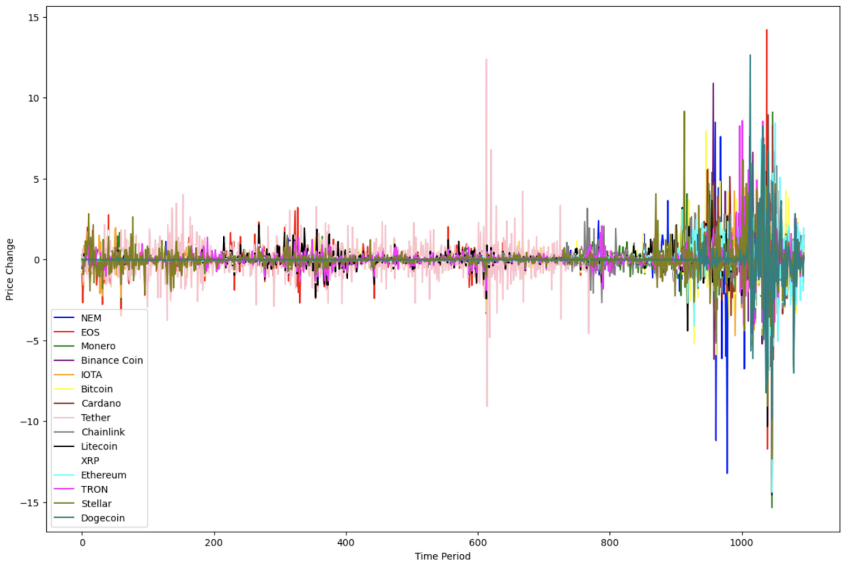
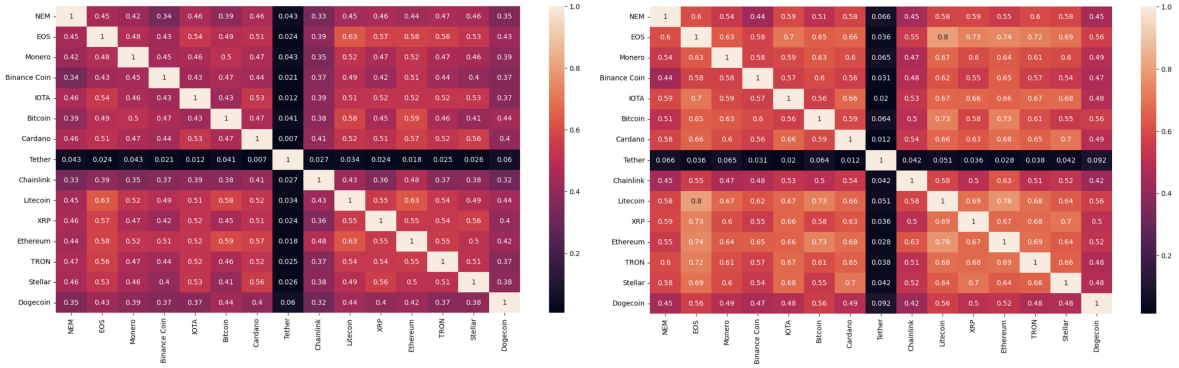


Figure 2: Price changes of all 15 cryptocurrencies

We then apply two correlation models namely “Kendall” and “Spearman” to compute the correlations and plot a heatmap to visualize the results. Both the correlations give similar results.



(a) *(b)*

Figure 3: (a) Results of Kendall correlation (b) Results of Spearman correlation

Clustering

Why do we need clustering for cryptocurrencies?

Clustering is performed to group cryptocurrencies that move in a similar pattern, allowing investors to diversify their portfolio by investing in multiple cryptocurrencies that are not highly correlated with each other. By identifying the clusters of cryptocurrencies, investors can reduce the risk of losses and improve their chances of making profits.

Methodology

We initially started out by performing clustering based on the recent market cap trend of the various cryptocurrencies. We decided to drop this idea because the market cap value is a generalized value that represents the complete value of the company’s shares of stock. It only depends on the stock price and the company’s shares which doesn’t give vital information like price variation as explained above

in the preprocessing section. Therefore, we use the price change values computed as part of the correlation section to base the formation of clusters.

We then use the concept of Principal Component Analysis (PCA) to reduce the dimension of the data and identify the most important features that contribute to the price change of the cryptocurrencies. We set the number of dimensions as 2 and use the method on the price change dataframe we created earlier to obtain the reduced data.

After this, we use the K-Means clustering algorithm on the reduced data. To decide on the number of clusters, we use the elbow technique over cluster counts ranging between 1 and 15. On observing the results, we set the number of clusters to 4.

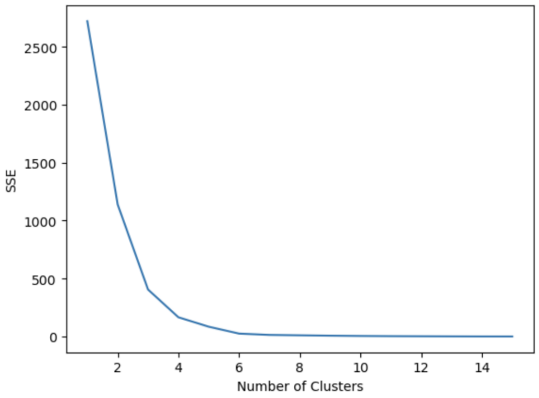
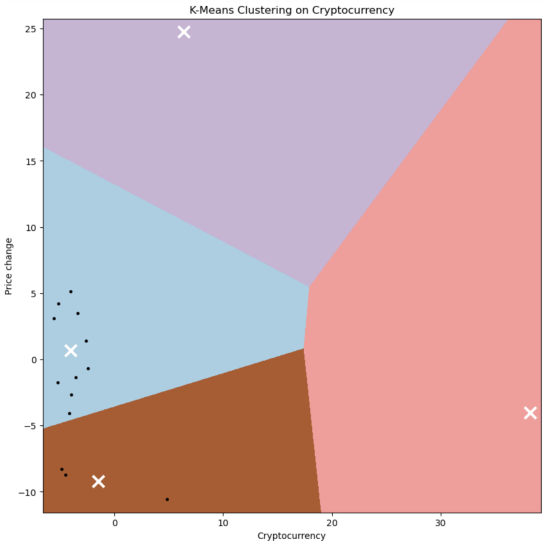


Figure 4: Results of Elbow technique for clustering

On implementing the K-Means algorithm with 4 clusters, we get the following results.



(a)

labels	Cryptocurrencies
0	EOS
0	Monero
0	Binance Coin
0	IOTA
0	Bitcoin
0	Cardano
0	Chainlink
0	Litecoin
0	Ethereum
0	TRON
1	Tether
2	Dogecoin
3	NEM
3	XRP
3	Stellar

(b)

Figure 5: Results of K-Means clustering. In (a) the blue cluster corresponds to label 0, pink cluster to label 1, purple cluster to label 2, and brown cluster to label 3.

Price Trend Analysis (Atharv)

Why do we need price trend analysis for cryptocurrencies?

The cryptocurrency market is highly volatile and can be challenging to predict, with prices fluctuating rapidly and often without apparent cause. As a result, investors in the cryptocurrency market require tools that can help them understand the behavior of the market and make more informed decisions about buying and selling cryptocurrencies. One such tool is price trend analysis, which involves analyzing historical price data to identify patterns and trends that may indicate future price movements.

Investors typically use technical analysis tools such as candlestick charts, moving averages, and support and resistance levels to conduct price trend analysis. These tools can help investors to identify key price levels and trends in the market, as well as potential opportunities for profit.

For example, candlestick charts can be used to identify patterns such as bullish and bearish engulfing patterns, dojis, and hammers. These patterns can indicate market sentiment changes and may predict future price movements. Moving averages can also be used to identify trends in the market, with longer-term moving averages typically used to identify the overall trend and shorter-term moving averages used to identify potential profit opportunities.

In short, price trend analysis is an essential tool for investors in the cryptocurrency market, as it can provide valuable insights into market behavior and help investors make more informed decisions about buying and selling cryptocurrencies. By analyzing historical price data and using technical analysis tools, investors can identify patterns and trends in the market and potential profit opportunities. While price trend analysis is not foolproof and cannot predict the future with certainty, it can help investors to mitigate risk and make more informed decisions in the highly volatile cryptocurrency market.

How Predicting Market Trends Can Help Investors?

In investing, predicting market trends is essential for maximizing profits and minimizing risks. Predicting market trends involves analyzing economic and financial data to identify patterns and relationships that can help investors make informed investment decisions. This report explores how predicting market trends can help investors, including timing investments, maximizing returns, minimizing risk, and developing a long-term investment strategy.

Timing Investments

Timing investments is a crucial aspect of investing. Predicting market trends can help investors identify the right time to buy or sell investments.

Maximizing Returns

Predicting market trends can also help investors maximize their returns. By analyzing economic and financial data, investors can identify investments likely to perform well in the current market environment.

Minimizing Risk

Predicting market trends can also help investors minimize their risks. By identifying potential risks in the market, investors can develop strategies to mitigate them.

How Can Regression Be Useful in Predicting Market Trends?

Regression analysis is a statistical technique used to examine the relationship between two or more variables. In predicting market trends, regression analysis can be a valuable tool for identifying patterns, quantifying relationships, forecasting future trends, and mitigating risks. This report explores why regression can help predict market trends and how it can benefit investors.

Using regression analysis combined with other analytical tools and market knowledge, investors can make more informed decisions about their investments and achieve their financial goals. By understanding how regression analysis can help predict market trends, investors can enhance their investment strategies and increase their chances of success.

Methodology

Simple Linear Regression

We used the data on the opening and closing prices for multiple cryptocurrencies, including Bitcoin, Ethereum, and Litecoin, retrieved from the Kaggle dataset. We then used linear regression analysis to predict closing prices using the opening prices as the independent variable. This technique involves fitting a straight line to the data, representing the best fit between the two variables.

We combined all the cryptocurrency files into one extensive dataset and then split the data set into a training and testing set. The training dataset comprised 75% of the data, while the testing dataset comprised 25% of the data. This division allowed us to validate the accuracy of the model's predictions.

As seen in Figure 6, predicted prices generated by the linear regression model and comparing them with the actual prices, we discovered that while the model produced estimates that were near the actual values, the accuracy of the predictions was suboptimal.

	Date	Name	Actual	Predicted
10/5/20	23:59	AAVE	53.219243	53.741457
10/6/20	23:59	AAVE	42.401599	44.156844
10/7/20	23:59	AAVE	40.083976	39.339247
10/8/20	23:59	AAVE	43.764463	43.050632
10/9/20	23:59	AAVE	46.817744	46.805604
...	...	...	...	...
7/2/21	23:59	XRP	0.656763	0.647744
7/3/21	23:59	XRP	0.672888	0.668513
7/4/21	23:59	XRP	0.694945	0.692861
7/5/21	23:59	XRP	0.654300	0.662884
7/6/21	23:59	XRP	0.665402	0.668544

Figure 6: Linear Regression Output

Multiple Linear Regression

To improve the model's accuracy of predictions, we shifted to multiple linear regression. Multiple linear regression is a technique used to analyze the relationship between multiple predictor variables and a target variable, allowing for more complex relationships to be captured than what is possible with simple linear regression. Another advantage of Multiple Linear Regression is that it can help reduce the influence of outliers by considering the impact of multiple predictors simultaneously. This

is because outliers may be because of a single predictor, but when multiple predictors are considered together, the influence of any single predictor is diluted.

Similar to the Linear Regression model, the training dataset comprised 75% of the data, while the testing dataset comprised 25%. For this model, the predictor variables are Open, High, Low, Volume, and Market cap, and the target variable is Close.

Figure 7 shows the output of actual Close prices and predicted Close prices.

	Date	Name	Actual	Predicted
10/5/20	23:59	AAVE	53.219243	53.741457
10/6/20	23:59	AAVE	42.401599	44.156844
10/7/20	23:59	AAVE	40.083976	39.339247
10/8/20	23:59	AAVE	43.764463	43.050632
10/9/20	23:59	AAVE	46.817744	46.805604
...	...	...	...	...
7/2/21	23:59	XRP	0.656763	0.647744
7/3/21	23:59	XRP	0.672888	0.668513
7/4/21	23:59	XRP	0.694945	0.692861
7/5/21	23:59	XRP	0.654300	0.662884
7/6/21	23:59	XRP	0.665402	0.668544

Figure 7: Multiple Linear Regression Output

As we can see in Figure 7. The multiple Linear regression model performs much better than the Simple Linear Regression model. Hence, using Multiple Linear regression improved the prediction accuracy and gave results close to the actual prices.

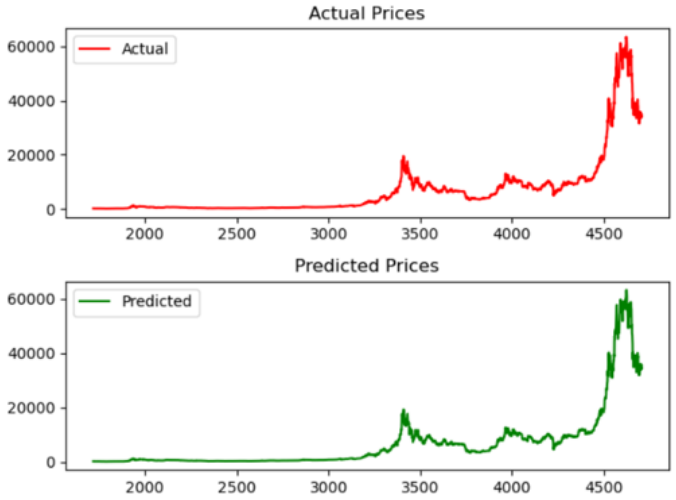


Figure 8: Actual vs Predicted prices for Multiple Linear Regression

Time-Series Regression Analysis

Time series analysis is a powerful tool for predicting cryptocurrency prices because cryptocurrencies exhibit the unique characteristic of being highly volatile and prone to sudden changes in price. Time series analysis is particularly well-suited to analyze and predict the behavior of such dynamic and constantly evolving markets. Various factors often influence the prices of cryptocurrencies, which can cause prices to fluctuate rapidly and unpredictably, making it difficult to predict future prices using traditional statistical methods. Time series analysis can help to capture these dynamic relationships by analyzing patterns and trends in past prices and using them to forecast future prices.

The main intention behind using time series regression analysis is to predict future data. In the above regression models, we split the data into train and test data, trained the model, and tested it on the data

we already have. But in reality, we would need a forecast of future data, and a time series forecast gives us predictions of what the data might look like.

The Time-Series model used ARIMA (Autoregressive integrated moving average) works by modeling the autocorrelation in a time series, using the combination of autoregression (AR) and moving average (MA) components, and incorporating differencing to make the series stationary.

The dataset we currently have has data only up till 7/6/21. We used the Time Series Regression on Bitcoin (BTC) data and predicted Low, Close, and High prices for the next seven days (7/7/21 – 7/13/21).

	High		Close		Low
2021-07-07	35213.038859	2021-07-07	34310.743971	2021-07-07	33314.731373
2021-07-08	35321.564439	2021-07-08	34243.570656	2021-07-08	33293.164191
2021-07-09	35213.778601	2021-07-09	34108.743472	2021-07-09	33295.556450
2021-07-10	35322.424129	2021-07-10	34090.247316	2021-07-10	33310.413480
2021-07-11	35320.688868	2021-07-11	34033.035518	2021-07-11	33325.277604
2021-07-12	35321.177554	2021-07-12	33922.877999	2021-07-12	33340.148822
2021-07-13	35215.746443	2021-07-13	33939.302059	2021-07-13	33355.027134

Figure 9: Forecasted High, Close, and Low prices for BTC

The forecasting model gives accurate forecasts, confirmed by comparing them with the actual values available on the internet archive. The forecast gives us a reasonable estimate of how prices will be over the period and can help us identify when would be a good time to sell or buy.

Thus, regression analysis is a valuable tool for predicting cryptocurrency prices because it enables investors to analyze the complex relationships between different variables that can affect the price of cryptocurrencies. By analyzing historical data and identifying patterns and trends in cryptocurrency prices, regression models can be used to make informed predictions about future prices. Moreover, regression analysis can also help investors to identify which variables have the most significant impact on cryptocurrency prices, allowing them to make more informed investment decisions. With the help of regression analysis, investors can better understand the market dynamics of cryptocurrencies and develop effective trading strategies to capitalize on market opportunities.

Overall, regression analysis is an essential tool for predicting cryptocurrency prices. Still, it should be used with other analytical and investment strategies to develop a comprehensive understanding of the market and make informed investment decisions.

Volatility (Zachary Wallace)

The volatility of cryptocurrencies is a well-known facet that makes investments very risky in most cases. Part of building an efficient portfolio for investment is analysis of the volatility of our options. A problem appears though: how can we quantitatively measure the volatility rankings of each cryptocurrency?

Pre-Analysis

We need a measure so a client's portfolio can include safety rankings of investments. To start our analysis, we can analyze some basic representations of our data. That is, we can simply look at the change in market capital over time. This will display the value of the coin over time since the market capital represents the overall value of the cryptocurrency.

The following are some of the cryptocurrencies with their change in market capital over time. We will look at the subset {Aave, EOS, Bitcoin, Etherium} of all the cryptocurrencies to reduce clutter:

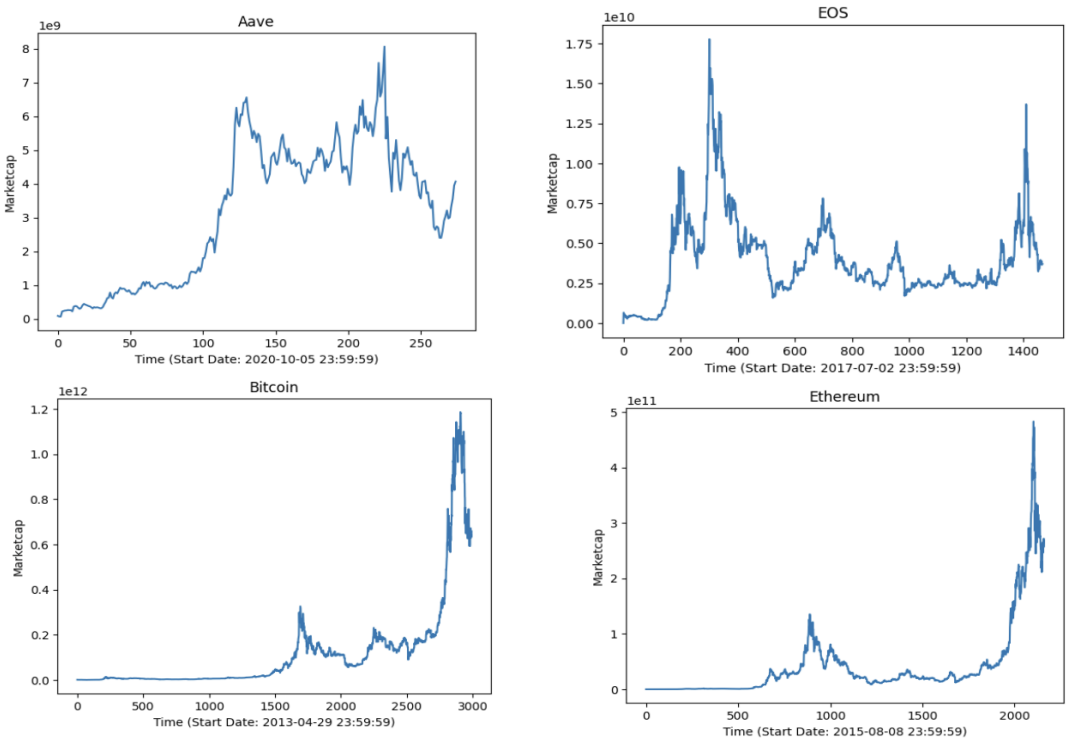


Figure 10: Visualization of change in market capital

Notes on the visuals:

- The "Time" on the x-axis refers to an index of days since the start date
- Refer to the multiple of e at the top left for significant digits of market capital

Volatility with Clustering

With this information, we can now apply data mining techniques to visualize the volatility of the cryptocurrencies. For the first set of trials, clustering was the first technique attempted. The motivation was to cluster the data into similar points of data such that outliers would be more visible.

In order to graph, we used the aforementioned market capital along with a calculation of the absolute difference in market opening and closing values. The reason behind this was to provide a market capital value in relation to the difference in market values per day. Along with each, we created a distribution of this absolute difference to model the spread of the data. The results of that experiment gave graphs for most cryptocurrencies that appeared similar to the graph here (showing Monero as an example):

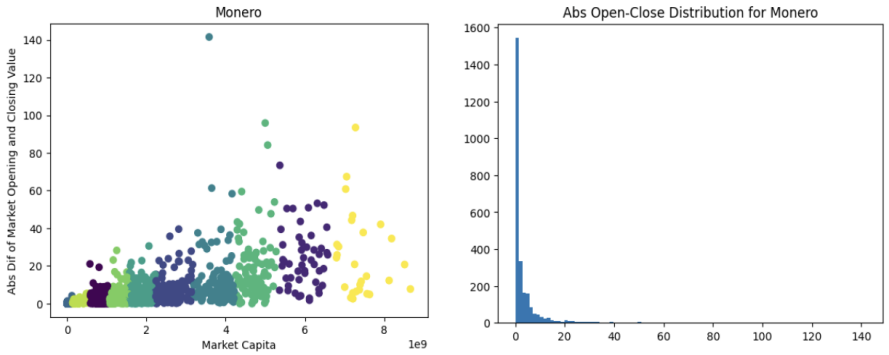


Figure 11: Clustering of Monero

- The type of clustering used wa K-Means++, with cluster count of 10
- The distribution is made with a simple histogram to show density

Performing clustering on this resulting data certainly gave some interesting results. From observation of this and other more apparent volatile cryptocurrencies like Bitcoin, for example (appendix for reference), it seems like there is a correlation in the cluster's sparsity and volatility. That is, the closer together points are in clusters as we move to the right, the less volatile a cryptocurrency is.

Consequently though, this posed a huge problem that could not help with our question: how can this quantitatively show someone that this would be a safer choice? There is no definite ranking here that easily compares its volatility to other cryptocurrencies'. Therefore, we need to move on to the next experiment.

Volatility with Linear Regression

In this second experiment, we took a more simple approach to finding a ranking of the volatility. To do this, we started by revisiting the original plots of market capital over time. These original plots, where we can visually observe volatile periods in the currencies, would be the answer to finding our results.

The next data mining technique applied here was simple linear regression. Due to the values of time being a 1-Dimensional listing of values, a non-simple variant would not be usable. Our implementation of simple linear regression would not be to predict upcoming values though – since that would be extremely ineffective. Rather, our implementation was designed to create a linear prediction line to represent the "ideal" progression of market capital over time. This "ideal" change over time would clearly be unobtainable by any stock or cryptocurrency. That is why we can measure a cryptocurrency's volatility by its ability to *approach* this line of ideality. Therefore, we can have our implementation return a cryptocurrency's Sum of Squared Errors; a complete measure of each point's error margin from the line of ideality. Moreover, this means the larger the SSE for a cryptocurrency, the further from an ideal growth it is – that is the more volatile it is. The results of this experiment appeared as such:

Volatility Ranking (Least volatile first)
Aave: 4.927684116452311e+20
Crypto.comCoin: 6.579150343708819e+20
Cosmos: 7.402420592533324e+20
Solana: 2.5581808938639483e+21
WrappedBitcoin: 2.7977898113083596e+21
Monero: 3.698677827000237e+21
Uniswap: 4.6217970355708866e+21

Tron: 4.86894382854183e+21
NEM: 6.815825388652295e+21
Iota: 6.839162260122751e+21
EOS: 9.752834242038242e+21
ChainLink: 1.079080995246876e+22
Stellar: 1.4607611795527912e+22
USDCoin: 1.6288969090900261e+22
Polkadot: 2.276102375036336e+22
Litecoin: 2.595721590866329e+22
Dogecoin: 1.6124637551556124e+23
Cardano: 1.8402754587748878e+23
Tether: 2.135532905010077e+23
BinanceCoin: 3.2131908124477094e+23
XRP: 3.6717479913621346e+23
Ethereum: 6.976073019517479e+24
Bitcoin: 7.426109654380517e+25

Thus, we can quantitatively conclude the rankings of volatility for each of the cryptocurrencies. It was personally surprising to learn some of these rankings. Bitcoin being the most risky to invest in is not too surprising though, given its popularity.

How do these analyses work together?

Let's take a small example to understand how the work combines -

- To begin with, visualizations under the volatility section are used to gain an understanding of the historical behavior of cryptocurrencies. This is followed by data preprocessing and correlation analysis to refine the data and identify patterns among various cryptocurrencies.

- To obtain a diversified portfolio we can use the clustering algorithm and choose which cryptocurrencies we want to invest in based on the clustering results. A cluster is formed by cryptocurrencies that follow similar trend, thus based on market conditions we can have a cluster that is highly volatile and the investor might not want to invest in currencies belonging to that one cluster, since all the cryptocurrencies in that cluster are likely to follow the same market trend. This will help in diversifying portfolios and reducing risk of losing money.

- Based on the chosen cryptocurrencies, regression analysis is used to forecast future prices. Investors can use this information to make informed decisions about when to invest. For example, buying when the forecasted price is low and selling when the forecasted price is high can help maximize profits.

- In conclusion, by utilizing data mining techniques such as clustering and regression analysis, investors can make informed decisions about cryptocurrency investments and create a diversified portfolio that minimizes risk and maximizes returns.

Conclusion

Predicting the behavior of cryptocurrencies is a complex task that requires careful consideration of a number of different factors. One of the main challenges in cryptocurrency prediction is the volatility of data, which can be influenced by a wide range of external factors and can change rapidly over time. To address this challenge, it is important to use sophisticated algorithms and techniques that can analyze large amounts of data and identify meaningful patterns and trends.

Another important consideration in cryptocurrency prediction is the use of clustering and correlation analysis, which can help to identify relationships between different cryptocurrencies and the factors that influence their behavior. By clustering cryptocurrencies based on similar patterns of price fluctuations, it is possible to gain insights into the underlying factors that drive market movements and make more accurate predictions about future trends.

Trend analysis is another important tool in cryptocurrency prediction, as it allows us to identify patterns and trends in market behavior over time. By analyzing historical data and identifying key trends and patterns, we can gain insights into the underlying factors that drive market movements and make more informed predictions about future price movements.