Does Crypto Pay Out ?

COMP 30780 Data Science in Practice

Cillian Dunne (18447676), Eoghan McGlinchey (18300093), Aakrit Shrestha (1748526)

School of Computer Science

University College Dublin

07/05/2021

An exploration of the cryptocurrency landscape, specifically looking to determine whether it is profitable enough to merit investment. The project looks to investigate: cryptocurrencies' relationships with each other, the common stigma that cryptos are too dangerous to invest in, Covid-19's effect on the market, and social media's effect on the market. The analysis takes place on twenty selected currencies. Results found include; a low level of interactivity between currencies, with notable exceptions. Stigmas holding true, cryptocurrencies are indeed wildly volatile, however, some patterns can be observed in certain tokens. Covid's effect on cryptocurrencies appears to be quite limited, largely not affecting the market's projected direction. Social media can be used to effectively identify trends using volume, with sentiment analysis providing a clearer picture.

Declaration. We Cillian Dunne (18447676), Eoghan McGlinchey (18300093), Aakrit Shrestha (1748526) declare that this assignment is our own work and that we have correctly acknowledged the work of others. This assignment is in accordance with University and School guidance[1] on good academic conduct in this regard.

[1] See https://www.cs.ucd.ie/sites/default/files/cs-plagiarism-policy_august2017.pdf

1.Introduction

The cryptocurrency market is expanding at an incredible rate, with records being set regularly. Bitcoin has landed at an astounding 1.1-trillion-dollars, with the total market cap of crypto hitting 2-trillion-dollars[2]. Due to this huge rise in crypto's value, its notoriety is at an all-time high – it seems as though crypto is on more minds than ever before. From those with a formal education in economics to individuals working in your local supermarket – everyone is trading these days. Never before has trading cryptocurrency been so accessible to the masses. Applications such as Coinbase and Binance have simplified trading to a simple click of a button. Revolut, a relatively new online banking option that has taken Ireland by storm (approximately a quarter of the country have a Revolut account[3]), has a range, albeit a limited selection, of cryptos that one can exchange Euros for extremely easily and portfolio can be very quickly be assembled with next to no effort on the part of the investor at all; the user can simply buy and sell and that's that. The fact that a huge amount of the population is stuck at home due to Covid-19 lockdown restrictions could also be a contributing factor to the ever-growing popularity of cryptocurrencies. The cryptocurrency market is expanding at an incredible rate, both in terms of value and fame. We intend to dive in to see if this boom is well-founded and to ask the question: Does Crypto Pay Out?

The key topics that will be explored are:

- Cryptocurrencies relationship with other cryptocurrencies.
- The statement, "Cryptos are too volatile to invest in".
- Covid-19's influence on cryptocurrency.
- Social media's influence on cryptocurrencies.

The sections that will be covered in this report are:

- Motivations & Objectives:
    - Detail previous work done in relation to our project.
    - Outline of research questions to be answered in later sections of the report.
- Data Wrangling
    - Explanation of data collection and data processing.
- Data Analysis & Results
    - A detailed look at the analysis and results for each research question.
- Discussion
    - Dialogue discussing project aspects such as ethics, reproducibility and potential limitations.
- Conclusion & Future Work
    - Reiteration of key findings and ideas for future work.
- Responsibilities
    - Summary of how the workload was split between the authors.
- Bibliography
- Appendix

[2]https://www.reuters.com/article/us-crypto-currency-marketcap-idUSKBN2BS1I7

[3]https://www.altfi.com/article/6571_1-in-5-irish-adults-have-a-revolut-card#:~:text=Digital%20banking%20service%20Revolut%20has,Irish%20adults%20are%20Revolut%20customers

2.Motivations & Objectives

Much talk has been around how bitcoin affects other crypto currencies[4] so one of the main goals of the paper is to show how different crypto currencies influence one another. Viewing different methods in answering that question with the use of bar plots, heatmaps and boxplots to give an overall answer. The other is it safer to invest long term or short-term answering it with the aid of seasonality trends giving a clear picture on how each currency acts. Also wanting to know how covid has affected the prices of the cryptocurrencies since covid has been defining part of everyone's lives. Could there be a drop in price or growth with all the areas covid has affected. With social media being a big part of everyone's lives could you use it to find trends in the crypto currencies.

2.1. Background & Motivations

Another group of researchers had done work on twitter sentiment analysis to predict the price of cryptocurrencies during the period of 2017 and early 2018. (Olivier et al., 2020) Their work was on nine largest currencies at the time being Bitcoin (BTC), Ethereum (ETH), XRP (XRP), Bitcoin Cash (BCH), EOS (EOS), Litecoin (LTC), Cardano (ADA), Stellar (XLM) and TRON (TRX). The method they used to collect the data was with a live stream crawler that continuously stored tweets in real-time. Totaling 22,912,039 tweets. For the financial information it was sourced from CoinMarketCap.

For the sentiment analysis use of the Valence Aware Dictionary and Sentiment Reasoner (VADER) algorithm. While taking a 12-step pre-processing plan to clean the text. Then applying the granger-causality test to find any trends on the set of tweets.

They found that for Bitcoin, Bitcoin-cash and Litecoin sentiment analysis is possible.Reason being why we picked this project was because social media is ever growing. Its influence has been far reaching. A simple google search of "social media's influence on" will give out multiple suggestions on varying topics. With a lot of media attention given to crypto currencies by popular media figures like Elon Musk (52.8m followers). Their influence cannot be ignored. With tweets seemingly causing a "crash" in price Bitcoin from uncredited (February 22, 2021)[5] Did a Tweet From Elon Musk Cause the Bitcoin Crash? This Is What We Know. Retrieved May 5th, 2021, from [2]. The topic being an interest for all of us I thought it would be fitting to study this.

2.2.Research Questions

2.2.1.RQ1: How much of an influence do the share prices of cryptocurrencies have on each other?

This research question investigated how the percentage daily high share price change of the top twenty cryptocurrencies by market cap on the 22nd of March 2021[6].

a) Do the share prices of cryptocurrencies have any influence on each other?
b) Do the share prices of cryptocurrencies have any influence on each other a number of days later?
c) Which cryptocurrencies are most influential?

[4] https://marketrealist.com/p/how-does-bitcoin-price-affect-other-cryptos/
[5] https://www.entrepreneur.com/article/365892
[6] https://coinmarketcap.com/

2.2.2.RQ2: Is it safer to invest in crypto short-term or long-term? (Cillian)

a) Can we find recurring trends?
b) Can we justify investing using either strategy?

2.2.3.RQ3: Has Covid-19 influenced cryptocurrencies?

a) What differences can be observed?
b) How have these cryptocurrencies been affected?

2.2.4.RQ4: Does social media influence cryptocurrency?

a) Do the volume of tweets correlate to the price ?
b) Can you spot trends with sentiment analysis ?

3.Data Wrangling

3.1.Data Acquisition

For research questions 1, 2, and 3, the historical cryptocurrency data we used came from an API key from cryptocompare.com. The key was free of charge and was easy to work with. The key could take the abbreviation of a cryptocurrency and the number of days for which you wished to collect historical data for. The historical data was collected by iterating over a list of cryptocurrencies and using the requests and json packages in python to get a response from the API key and create a JSON file with the historical data for each cryptocurrency. Within the JSON files, the objects used came from the "Data" array and the names used were "time", "high", "low", "open", and "close".

For research question 4 we used twint to scrape twitter avoiding the limitations of Twitter's API. A use of a keyword search was implemented which grabbed any tweets containing the word given. A limit of 50 likes was placed to help with the run times due to the high volume of tweets for Bitcoin. This was placed for all other currencies to keep the research consistent. All tweets were gathered during the period of 2019 to 2020. In total 155,345 tweets were gathered from 14 different currencies bitcoin (BTC), binancecoin (BNB), bitcoincash (BCH), cardano (ADA), dogecoin (DOGE), ethereum (ETH), Litecoin (LTC), miota (MIOTA), monero (XMR), ripple (XRP), stellar (XLM), tether (USDT), tezos (XTZ), vechain(VEN). The reason why these 14 were chosen was due to the 20 given from the API of cryptocompare.com they had enough of a following where the ISO code was not mistaken for another form of popular media on twitter.

3.2.Data Cleaning & Preparation

Cleaning and preparing the historical data for the cryptocurrencies was relatively easy. All that really had to be done was to convert the time from Unix Epoch time to UTC time. For RQ1, the important objects and names from each JSON file were used to create columns for a pandas data frame that contained the historical data for the JSON file.

```
df = pd.read_json(json)

        #create empty lists for each important name in the JSON file

        #populate each list iteratively

        for i in range(len(df)):

                ls1.append(df[i][‘important name1’])

                .

                .

                .

        #use populated lists to populate new data frame

        df = pd.DataFrame()

        df[‘col1 name’] = ls1

        .

        .

        .

        #convert time column from Epoch seconds to YYYY-MM-DD

        df[‘time_col’] = pd.to_datetime(df[‘time_col’], unit=’s’)

        #set time col to become the index

        df = df.set_index(‘time_col’)

        return df
```

Figure 1. Sample code to show how each JSON file was converted to a pandas data frame.

```
#function to convert json file to dataframe
def convert_to_df(json):
    #convert json file to dataframe object
    df = pd.read_json(json)

    #create empty lists to be populated iteratively
    utc_time = []
    high = []
    low = []
    opn = []
    close = []
    #populate lists iteratively
    for i in range(len(df['Data']['Data'])):
        utc_time.append(df['Data']['Data'][i]['time'])
        high.append(df['Data']['Data'][i]['high'])
        low.append(df['Data']['Data'][i]['low'])
        opn.append(df['Data']['Data'][i]['open'])
        close.append(df['Data']['Data'][i]['close'])

    #create refined dataframe from these lists
    df = pd.DataFrame()
    df['time'] = utc_time
    df['high'] = high
    df['low'] = low
    df['open'] = opn
    df['close'] = close

    #convert time column from Epoch to YYYY-MM-DD
    df['time'] = pd.to_datetime(df['time'], unit='s')

    #set the index to time
    df = df.set_index('time')

    #return the dataframe
    return df
```

Next, a helper function was made to create data frames that represented the daily share prices of all the cryptocurrencies for a given price type (high, low, open, close). This function used pandas.concat(join='inner') and it also used pandas.DataFrame.max().sort_values(ascending=False) to ensure the data frame created was sorted from left to right in descending order of cryptocurrencies peak share prices.

```
def create_overall(price_type, currencies):

    df = pd.concat(col, join='inner')

    #sort columns in descending order of highest value in each
    #column

    df = df.loc[:, df.max().sort_values(ascending=False).index]
```

```
        return df
```

Figure 2. Sample code to show how a data frame for a given share price type was created.

```
#function to create overall dataframes of all currencies for a given price type (high, low, open, close)
#and sort the columns in descending order of the highest value in each column
def create_overall(col, currencies):
    overall = pd.concat(col, axis = 1, join = 'inner')
    overall.columns = currencies

    #sort columns in dataframe in descending order of the highest value in each column
    overall = overall.loc[:, overall.max().sort_values(ascending=False).index]

    return overall
```

Finally, these functions were then used to create data frames representing the high, low, opening and closing daily share prices for all the cryptocurrencies. This was done by creating a data frame for each currency using the function in figure 1 and then populating lists for each share price type with the corresponding column in the dataframe. Then data frames for each share price type were created using the function in figure 2.

```
currencies_ls = [‘curr1’,…]

   #create lists to store given share price for all currencies

   price_type1 = []

   .

   .

   .

   #iteratively create dataframes for each cryptocurrency

   for currency in currencies_ls:

       #create name for dataframe of each currency

       df_name = currency+’df’

       #vars assigns the dataframe to the name of the dataframe

       vars()[df_name] = convert_to_df(currency+’.json’)

       #append each important column from the dataframe to the lists

       price_type1.append(vars()[df_name].price_type1)

       .

       .

       .

#create dataframes for the high, low, opening and closing #prices of all
cryptocurrencies

price_type1 = create_overall(high, currencies)
```

.

.

.

Figure 3. Sample code to show how data frames for each share price type were created.

```
#list of currencies
currencies = ['BTC','ETH','XRP','BCH','ADA','LTC','XEM','XLM','EOS','NEO','MIOTA','DASH',

#create lists to store data for all currencies
high_cols = []
low_cols = []
open_cols = []
close_cols = []

#iteratively create dataframes for each cryptocurrency
for currency in currencies:
    #assign the name of the df (e.g. BTC_df) as a string to the df_name variable
    df_name = currency+'_df'
    #concatenate '.json' to the end of the currency so the json file can be read
    currency+='.json'
    #use vars to assign the dataframe to the name of the dataframe
    vars()[df_name] = convert_to_df(currency)
    #append each column to overall lists
    high_cols.append(vars()[df_name].high)
    low_cols.append(vars()[df_name].low)
    open_cols.append(vars()[df_name].open)
    close_cols.append(vars()[df_name].close)

#create overall dataframes for the high, low, open and close prices of all currencies
#and sort the columns in descending order of the highest value in each column
overall_high = create_overall(high_cols, currencies)
overall_low = create_overall(low_cols, currencies)
overall_open = create_overall(open_cols, currencies)
overall_close = create_overall(close_cols, currencies)
```

Cleaning the twitter involved removing the columns which were not relevant. In total 37 columns are available for a single tweet. Reducing it down to 16. The removal of duplicates was performed by matching tweets by the date, time and id and filtering out non-English tweets. A new column was made which contained the currency name from where the tweet was originally pulled from. Then the hashtag column which contained the use of 301,369 hashtags, it contained duplicate tweets between different currencies. These duplicates were kept since one tweet could be directed at multiple currencies. Then to find the number of tweets for a selected currency. Search for a hashtag and group the selected ones together (see Table 1). Removing the duplicates and totalling up the amount in each day. Which was the basis for the answer of “do volume of tweets correlate to the price”

Table 1. Hashtags used in search of tweets for a certain currency

bitcoin	'btc','bitcoin'
binancecoin	'bnb','binancefutures','binance','binancetrading'
bitcoincash	'bch','bitcoincash'
cardano	'ada','cardano','cardanocommunity'
dogecoin	'doge','dogecoin','dogearmy'
ethereum	'eth','ethereum'
Litecoin	'lite','litecoin','ltc'
miota	'miota','iotatalks','iotacommunity'
monero	'monero','xmr'
ripple	'xrpcommunity','xrpthestandard','xrparmy','xrpcommmunity'
stellar	'xlm','stellar','stellarfamily','stellarlumens'
tether	'tether','usdt'
tezos	'tezos','xtz'
vechain	'vechain','vefam','vet','vechainthor'

For the sentiment analysis, we cleaned the text using the regular expression library. We used a 4-step action to remove the '@' mentions, then a removal of the '#' symbol and 'RT'. Finally, the hyperlinks. This was all performed on the hashtag data frame. Once the tweets have been cleaned, they would be added to a new column called 'tweet_clean'. On this resulting column with the use of 'text blob' producing three more columns each filled with a subjectivity, polarity and feeling score.

4.Data Analysis & Results

4.1.RQ1

4.1.1.Datasets

The data used to analyse this research question was taken from the JSON files created as described in section 3.1. These JSON files were then cleaned and prepared as outlined in section 3.2 to create data frames for each share price type over the past three years.

4.1.2.Approach

First, a function was created to plot a scatter plot with the correlation coefficient and r squared values as a legend to show the relationship between two cryptocurrencies. Then a function was created to plot small multiples of these scatter plots to show the relationship between one cryptocurrency and every other cryptocurrency. Upon analysing the correlation coefficients between the daily high share prices of Bitcoin and the rest of the cryptocurrencies, it was found that most of the correlation coefficients were suspiciously close to 1. Having presented this discovery in a progress update session, it was suggested that this was likely to have been caused by the fact that when correlating values in time series data, the trends are correlated instead of the values themselves. This problem was solved by taking each data frame that represented the various daily share price types for each cryptocurrency and converted the values to percentage changes using the .pct_change() function. In order for meaningful analysis to be carried out, the NaN values were dropped.

Next, a function was created to create a data frame that could take a cryptocurrency and populate its values as well as the values of all the cryptocurrencies, including itself a given number of days later. This was done for the purpose of creating visualisations that would depict relationships between changes in share prices of currencies with a shift of any amount of days. Interactive scatter plots were then created using the ipywidgets package for the purpose of making it quicker and easier to look at scatter plots between two cryptocurrencies.

The next functions created data frames, the values of which contained the correlation coefficients between currencies in the index and the currencies which were the column names. The same was done for currencies in the index and the currencies which were the column names a number of days later. These functions were created as helper functions to the functions that found the minimum, maximum, mean and median correlation coefficients for each cryptocurrency as well as the heatmap function which used the seaborn package to plot a heatmap to represent the strength of the correlations between each of the cryptocurrencies. To find the mean correlation coefficients, the correlation coefficients first had to be converted to z-scores and the mean z-scores were then converted to r-values. This is because “a correlation coefficient is a cosine, and cosines are not additive”[7]. An interactive bar chart and heatmap was then created so that the strengths of cryptocurrencies as well as the correlation coefficients between cryptocurrencies with various shifts in days could be observed at the click of a button.

The final function that was created was one that plotted a boxplot that visualised the pairwise correlation coefficients shifted by the number of days on the x-axis. This function used the function that created dataframes of the correlation coefficients for a given number of days later to populate a multidimensional list. This list then became the argument for the matplotlib.pyplot.boxplot() function.

[7] https://www.researchgate.net/post/average_of_Pearson_correlation_coefficient_values - Edel Garcia

4.1.3.Results

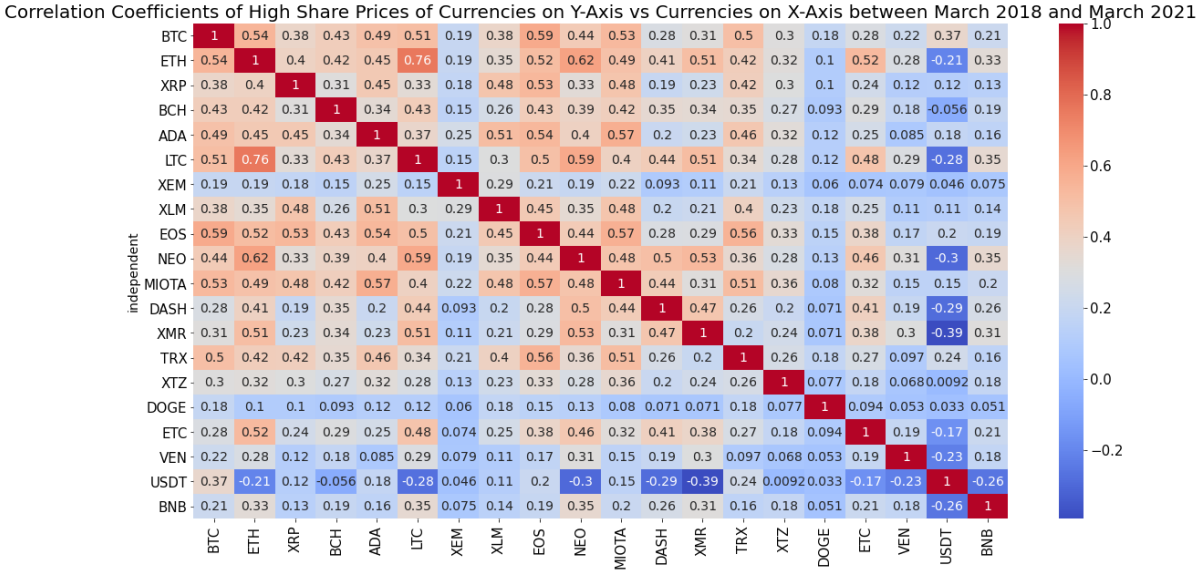

Figure 4. A heatmap showing the correlation coefficients between the percentage change in high share prices of cryptocurrencies on the y-axis versus the cryptocurrencies on the x-axis between March 2018 and March 2021.

The heatmap in figure 4 gives an overview of the influence each cryptocurrency has on each other. It is clear to see that overall, the correlation coefficients are not that strong, apart from the correlation coefficient between Ethereum and Litecoin which was 0.76. What we can see however is that the cryptocurrencies with the darkest squares (strongest correlation coefficients) are Bitcoin, Ethereum, Litecoin, EOS and MIOTA. Meanwhile, the lightest squares belong to XEM, Dogecoin and Tether.

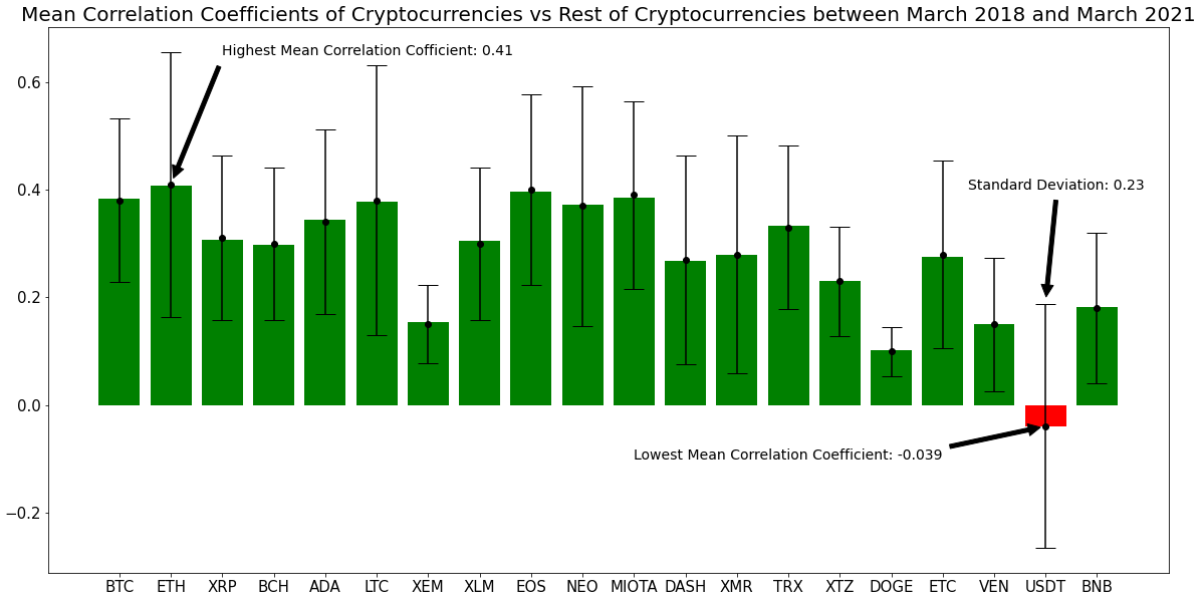

Figure 5. A bar chart showing the mean correlation coefficients of each cryptocurrency on the x-axis versus every cryptocurrency except for itself with error bars that represent the standard deviations between March 2018 and March 2021.

The bar chart in figure 5 shows that Ethereum was the most influential cryptocurrency on average with a mean correlation coefficient of 0.41. As well as that, since the standard deviation was so significant, it can be inferred that there were a number of correlation coefficients between Ethereum and other cryptocurrencies which were significantly higher or lower than 0.41. Meanwhile, Tether had the lowest mean correlation coefficient at 0.039. However, since the standard deviation was much larger than the mean itself at 0.23, it suggests that there were a few cases in which the correlation coefficient between Tether and other cryptocurrencies were significantly stronger than 0.039.

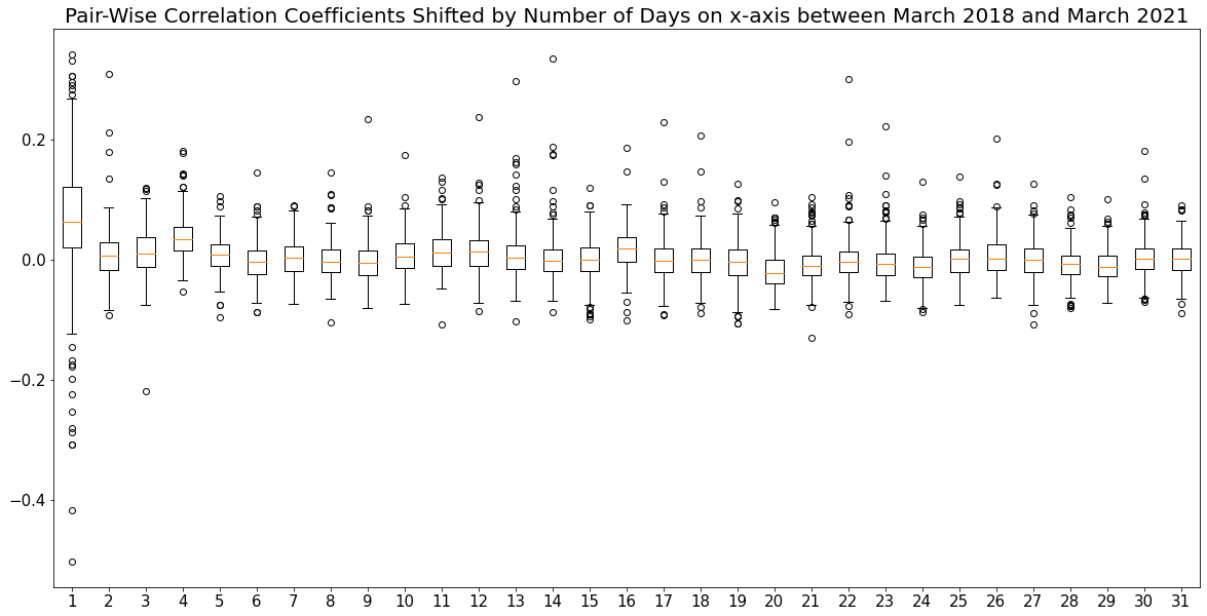

Figure 6. A box plot that shows the pairwise correlation coefficients between all the cryptocurrencies shifted by the number of days on the x-axis between March 2018 and March 2021.

The box plot in figure 6 gives an overview of how the lag in terms of number of days between cryptocurrencies affected the correlation coefficients between them. After a lag of two days and beyond, almost all the median correlation coefficients are extremely close to 0 with a few outliers that are not too significant. After shifts of one day and fourteen days, there are outliers that are more significant than shifts of other numbers of days, in particular after one day where there is an outlier of approximately -0.5. This correlation coefficient represented the relationship between Tether on one day and Tron a day later, however, nothing was found to suggest why this relationship was so strong in comparison to the rest of the relationships.

4.1.4.Discussion

These findings show that in general, the daily high share prices of cryptocurrencies had little effect on each other. This is illustrated by the overall lightness of colours of the squares in the heatmap in figure 4 and also by the bar chart in figure 5. After a lag of one day, the lag in number of days between cryptocurrencies seemed to have no effect on the correlation coefficients between them, as illustrated by the box plot in figure 6 in which the medians are all very close to 0 and the interquartile ranges and upper and lower quartiles all looked very similar.

4.2.RQ2

4.2.1.Datasets

See section 4.1.1.

4.2.2.Approach

The difficulty that arose when pondering solutions to this particular research question was how one could determine whether a cryptocurrency was best suited for long-term or short-term investment. Along with this issue, another question mark was how would one quantify this attribute? We began our analysis by looking at our time series data. Research, study, and of course trial-and-error were all done to help think of a method of evaluation for this research question and throughout this experimental period, time series decomposition was a concept that stuck. Time Series decomposition is a useful statistical operation that deconstructs a given time series into its core components, consisting of the following:

- Observed: The original time series data.
- Trend: The smoothed or rolling average of the time series data.
- Seasonal: The repeating or cyclic part of the time series data.
- Noise: The random variation found in the time series data.

Some attempt of manual decomposition is evident in our Jupyter notebooks; however, the final analysis utilizes the seasonal_decompose function found in the Python module called statsmodels.
Time Series decomposition requires a period to be specified. The selected period of analysis for this research question is one year. There are a few reasons for this, however, we must first explain how we are quantifying short-term and long-term trade effectiveness to properly justify our reasoning for selecting a yearly period for time series decomposition. This justification will be made at the end of this subsection.

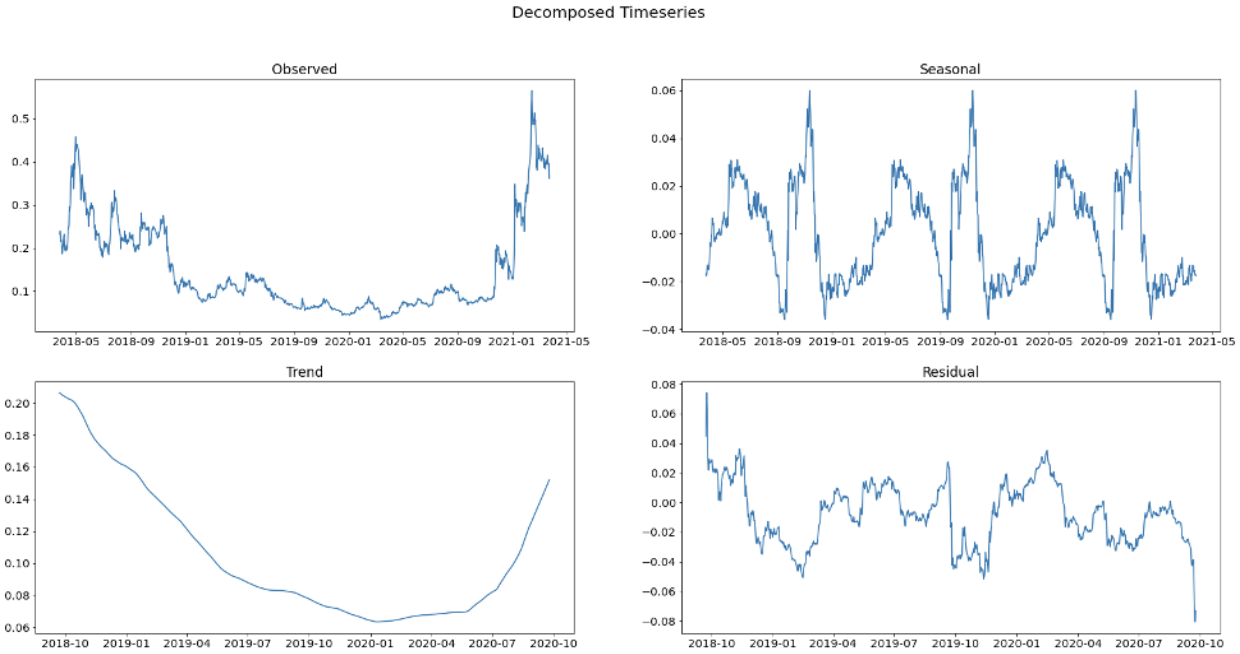

Figure 7. A time series decomposition of Stellar (XLM).

Now that we have access to the components discussed above, we can begin using them to attempt to answer our question at hand. The following formulae[4] will be used to quantify short-term and long-term trade effectiveness.

$$F_T = max(0, 1 - \frac{var(R_t)}{var(T_t + R_t)}) \qquad F_S = max(0, 1 - \frac{var(R_t)}{var(S_t + R_t)})$$

F_T : strength of the trend,
F_S : strength of the seasonality,
T_T : smoothed trend component,
S_T : seasonal component,
R_T : remainder component

The above formulas essentially compare the variance of the noise or remainder component with the variance of the trend and seasonal components. The two formulae compute strength values F_T, F_S for trend and seasonality respectively. These values are between 0 and 1 and form the basis for our results.

As mentioned previously, time series decomposition requires a specified period to break down the data concerning the given period. Experimentation was performed on different periods, the results of which can be observed below.

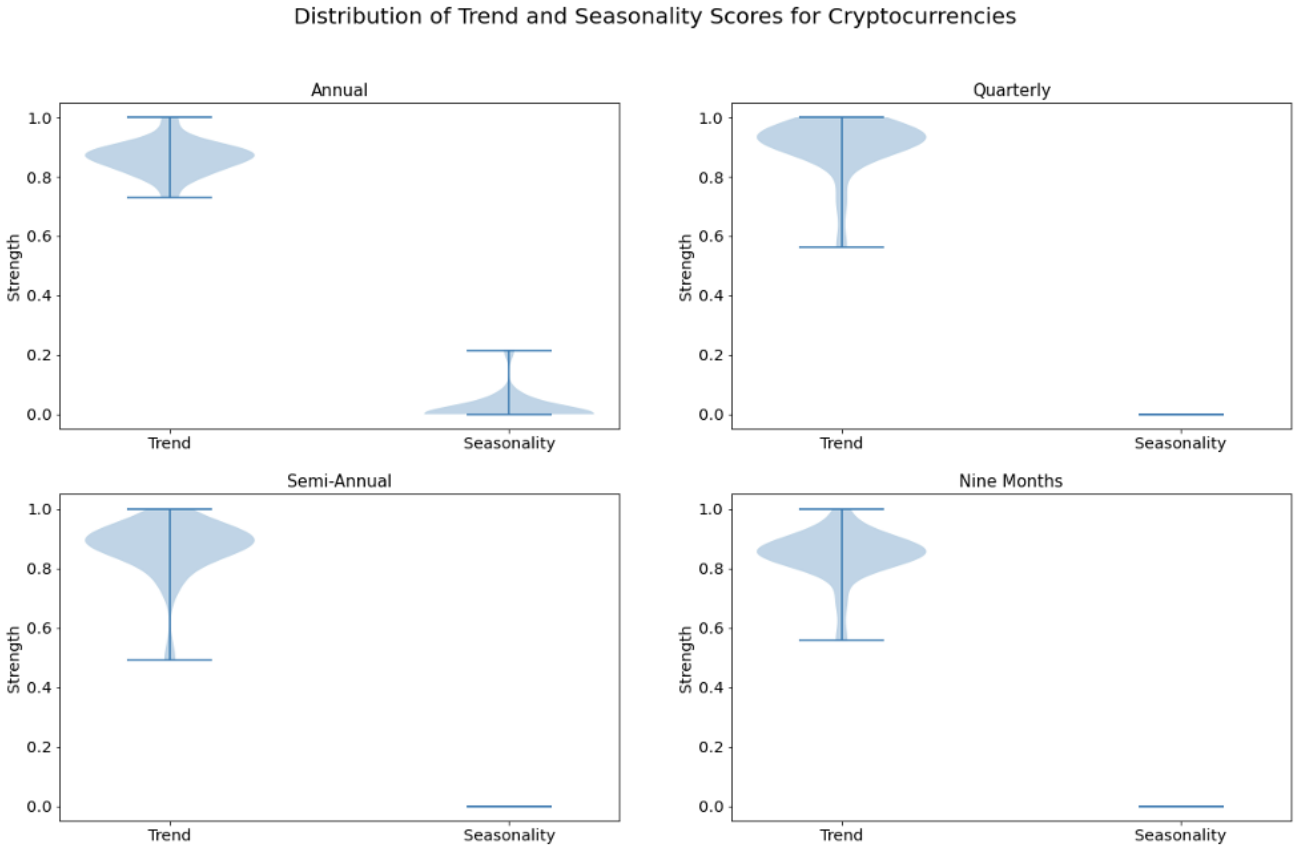

Figure 8. A matrix of violin plots showing the distribution of trend/seasonality strength values for different periods of tim[8]e series decomposition.

Trend strength varies interestingly across the different periods, however, seasonal strength is non-existent for quarterly, semi-annual, and nine-month periods. Seasonality can be observed for an annual period and so has been selected for our analysis.

[8] https://otexts.com/fpp2/seasonal-strength.html

4.2.3.Results

From Figure 8, we see that even in the annual violin plot that seasonality values for the cryptocurrencies being analyzed in this project are low, considerably lower than their trend scores.

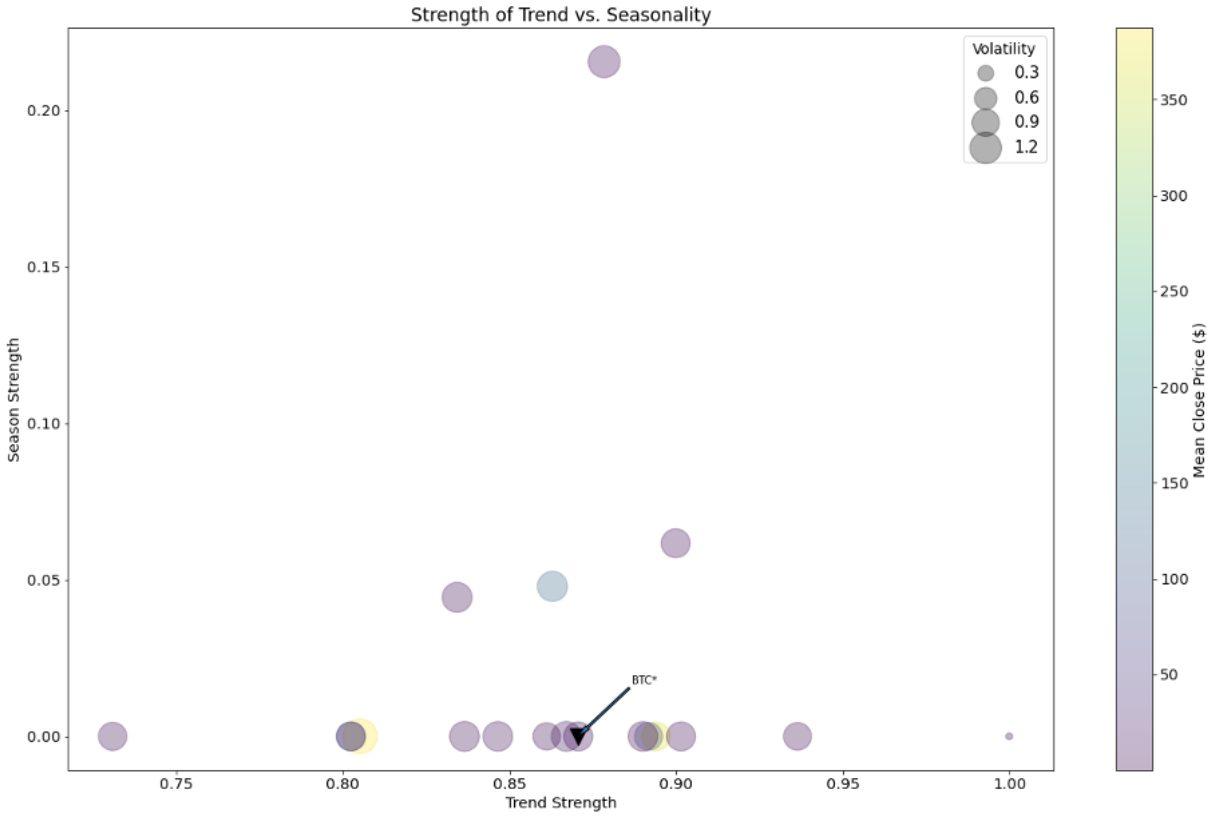

Figure 9. A scatter plot of cryptocurrencies that compares the strength scores for trend (x-axis) and seasonality (y-axis). Color and size are used to encode the mean close price and volatility of a crypto, respectively.

We observe in Figure 9 that most of the cryptocurrencies are strongly trended. However, when looking at seasonality strength, we note the opposite – the strength values are quite low, and that most of the cryptocurrencies are devoid of any seasonality at all.

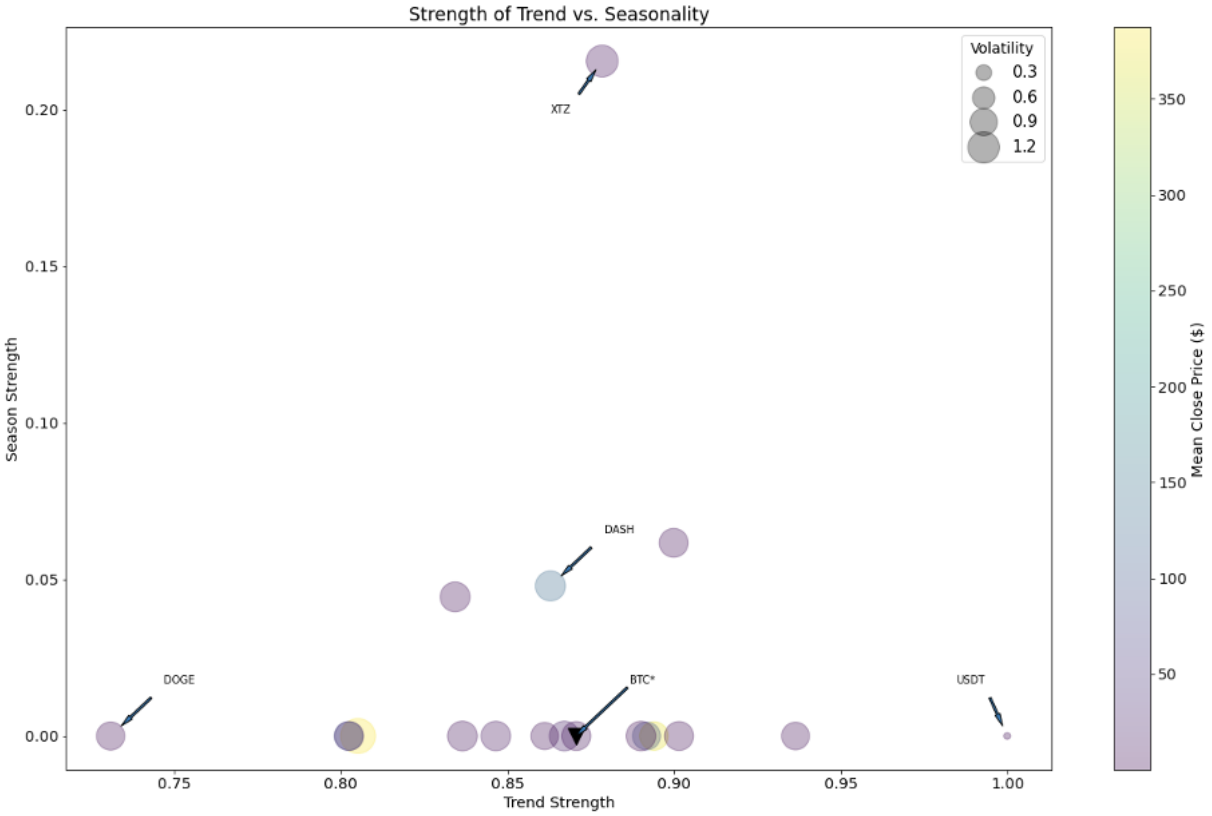

Figure 10. Annotated Figure 9 to highlight certain points of interest.

Taking a look at some of the outliers and other points of interest highlighted in Figure , we see that XTZ displays a significantly higher seasonality when compared to the rest of the dataset. However, an approximate 0.2 seasonality score still is not particularly dependable. Even with XTZ having the highest score in this aspect, the research here would conclude that short-term investment would not be a safe investment for cryptocurrencies, even XTZ.

Another point of intrigue is that of DASH. DASH is a relatively high-profile high-value cryptocurrency. It is interesting to see a high-value cryptocurrency display any seasonality and may point towards investors being savvy and buying/selling at similar times.

The penultimate point of interest is located at the leftmost extremity of the x-axis and is named DOGECoin. DOGE is known to be something of an anomaly in the crypto world, in that it was initially created as a joke, referencing the popular internet meme doge[9]. Despite its origin, DOGECoin has recently gained traction and is known to be volatile and unpredictable. For example, the most recent April Fool's Day in 2021 DOGECoin saw an increase of four times.

Finally, we observe USDT, a low volatility, value, and season strength cryptocurrency with the maximum trend strength value. USDT is an important example in the dataset as it demonstrates that a high trend value does not necessarily represent a positive or negative trend. USDT is a stable coin, meaning its value is pegged to another currency, in this case, USD AKA US Dollars.

[9] https://www.theverge.com/2013/12/31/5248762/doge-meme-rescue-dog-wow

4.2.4.Discussion

The results presented in the previous section have shown in essence that cryptocurrencies are indeed highly volatile and unpredictable. They display very little seasonality, though are, for the most part, highly trended. By the methodology outlined in the previous approach section, by having such low seasonal strength and such high trend values, cryptocurrencies, in general, are better suited and are safer when investing long-term.

Our findings align with our thinking ahead of analysis for this research question. It also aligns with the common phrase, “Cryptocurrencies are too dangerous to invest in”. Most of the cryptocurrencies display seasonality and as such coincide with this mindset.

The method of analysis that was outlined in the approach section of this report is an aspect that could be looked at regarding future work. Other methods of evaluation could be explored that may lead to different results.

4.3.RQ3

4.3.1.Datasets

For crypto data, see section 4.1.1.
Covid data gathered from CDC api. Used in order to analyse deaths over time.

4.3.2.Approach

Analysis first takes place by naively comparing the covid period against the period before. These periods are highlighted and then plotted individually. The distribution of volatilities are compared.

A function was created a dataframe, the values of which contained only data that was the national Covid deaths in the U.S.A. Two other functions were then created to create dataframes for a given cryptocurrency before and during Covid. Since the Covid deaths data in the U.S.A. was gathered on a weekly basis, in order to match the number of rows the Covid deaths data frame had, the share prices data in both of the other functions had to be taken every seventh day.

Next, a function was created to create a dataframe for a given share price type and cryptocurrency, which used the three functions above to create a dataframe. The columns of which were the share prices before and during covid, and in the case of the cryptocurrency being Bitcoin, a column for the number of Covid deaths was also created. The covid deaths column was not created in the instance of other cryptocurrencies, since the Covid deaths were so much larger than the share prices of any other cryptocurrency and as such, plotting these values on a time series would not be of any use from a visual perspective.

The final function that was created was an interactive time series, in which the share price type, the cryptocurrency and whether to use the percentage changes in share prices and Covid deaths were all options that could be chosen at the click of a button.

4.3.3.Results

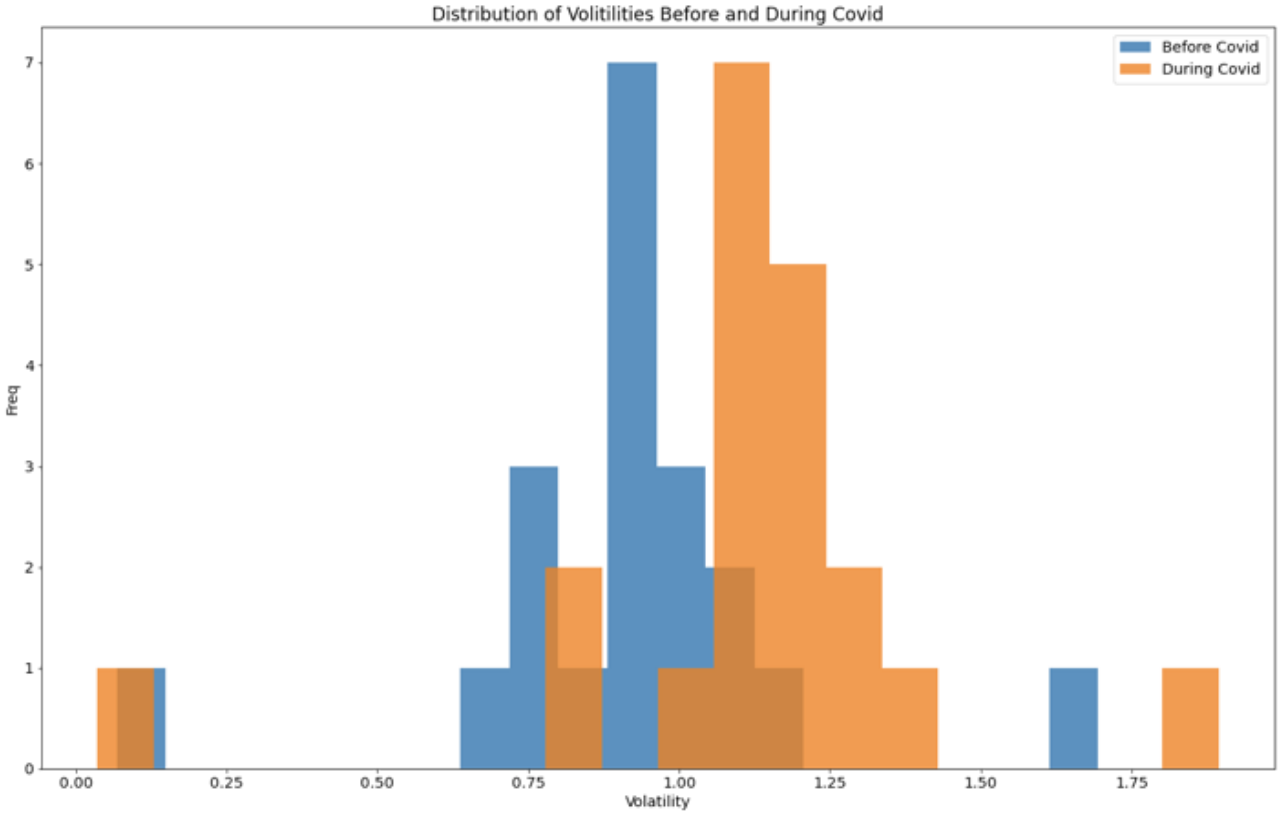

Figure 11. Histogram to show distribution of volatilities of currencies before and during the pandemic

From what can be seen from Figure 11, cryptocurrency volatility has increased quite significantly during the Covid-19 period. This increase can be observed nearly across all currencies. This could be due to the pandemic, however, the results here most certainly do not prove this, and could be an indication that the market, in general, has become more volatile.

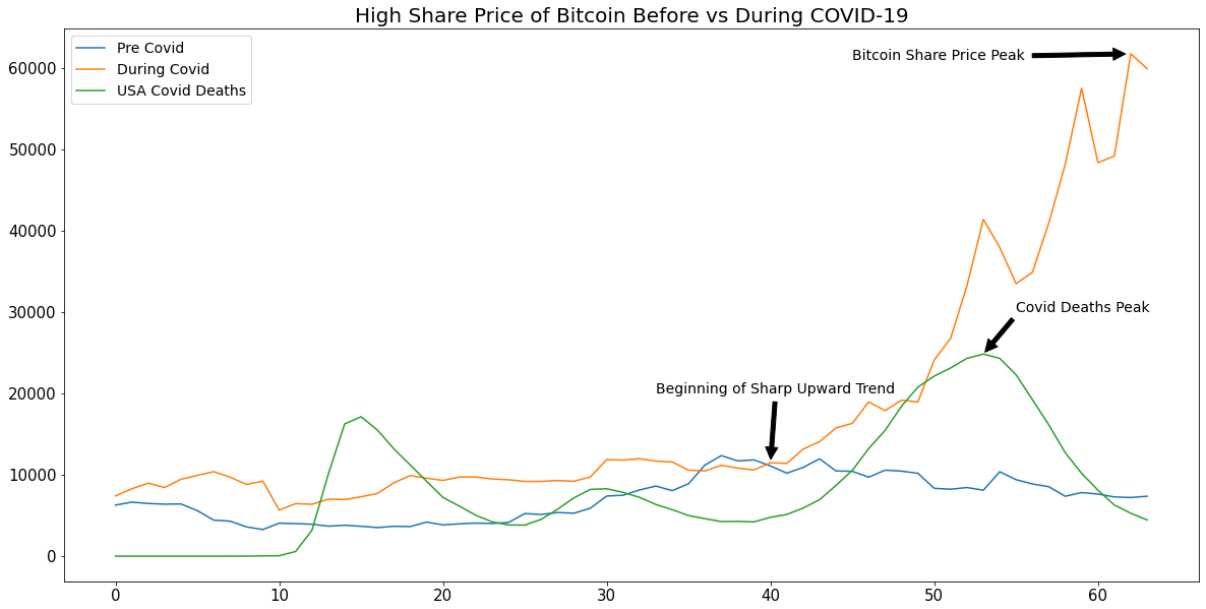

Figure 12. A time series showing how the share price of Bitcoin changed before and during Covid, as well as how the number of Covid deaths in the U.S.A. changed

The time series in figure 12 shows the share price of Bitcoin before and during Covid and the green line represents the number of Covid deaths in the U.S.A. The x-axis represents the number of weeks before and during Covid-19. The beginning of a sharp upward trend in the high share price of Bitcoin began at around week 40 of Covid in the U.S.A. Meanwhile, the Covid deaths underwent an upward trend at around the same time. However, the number of Covid deaths peaked at week 53, while the share price of Bitcoin continued to increase until around week 60 of Covid.

It was observed using the interactive time series function that many of the other cryptocurrencies were similar to the behaviour of Bitcoin before and during Covid-19, in that they began to increase in value during week 40 or 50 of Covid in the U.S.A.

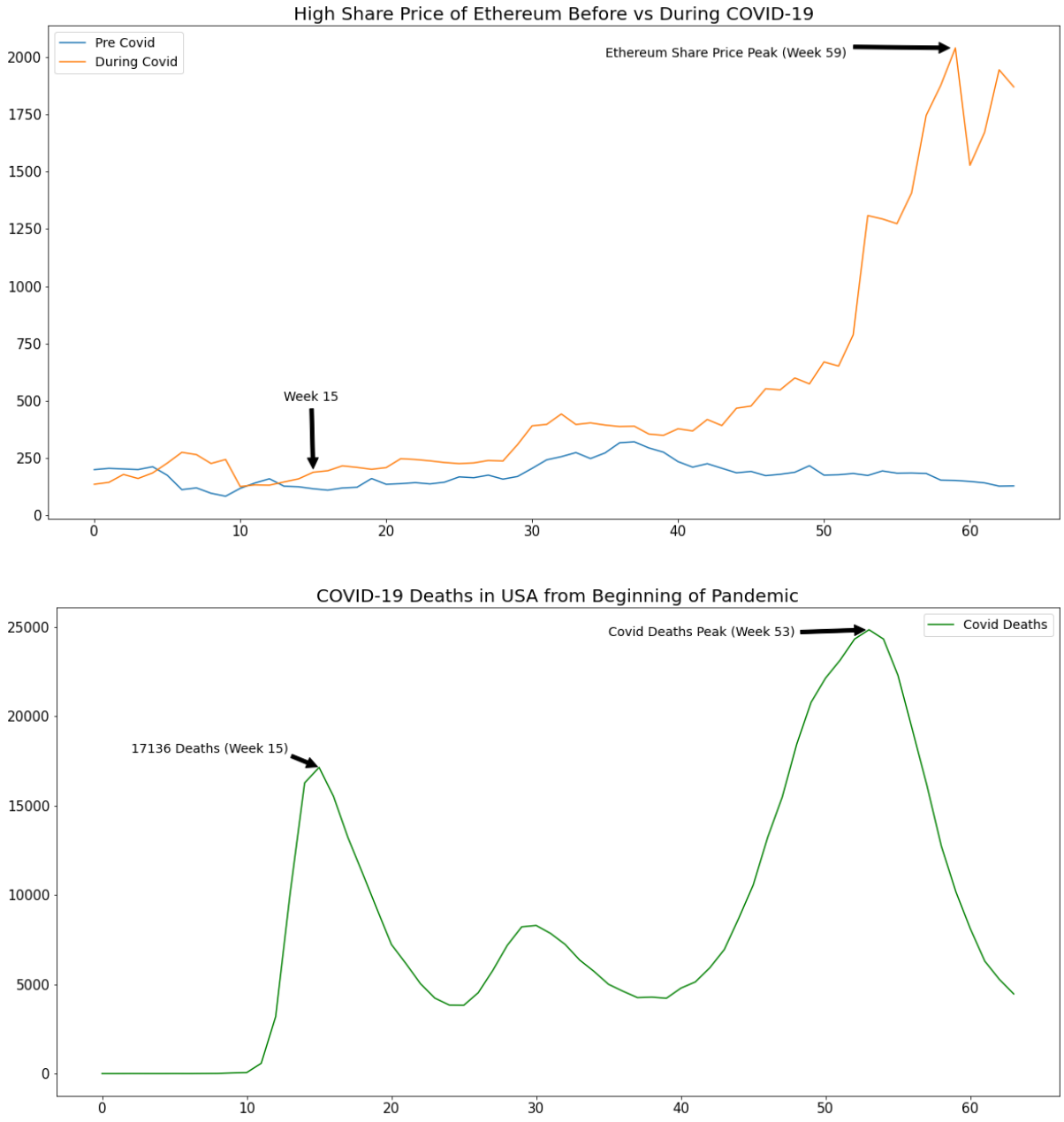

Figure 13. A time series showing how the share price of Ethereum changed before and during Covid, as well as how the number of Covid deaths in the U.S.A. changed

The time series in figure 13 shows the share price of Ethereum before and during Covid, as well as the number of Covid deaths in the U.S.A. below. While there was a peak in around week 15 in Covid

deaths, at the same time, the share price of Ethereum was undergoing an upward trend. As well as that, while the Covid deaths peaked in week 53, the share price of Ethereum continued to rise until week 59.

4.3.4.Discussion

Although a lot of the cryptocurrencies experienced a sharp increase in their share prices during Covid Since this happened between weeks 40 and 50 of Covid in the U.S.A. rather than from the beginning of Covid, it is safe to say that these increases in share prices were not due to Covid-19. This sentiment is shared when looking at the cryptocurrency volatility increase, in that it is very unlikely that this increase is tied to the Covid pandemic.

4.4.RQ4a

4.4.1.Datasets

4102_MASTER_FILE_CURRENCY.csv contains the closing price between 2019 and 2020 for the 14 cryptocurrencies

4105_hashtags_explode.csv contains the tweets in the same period with the hashtags column expanded upon

4.4.2.Approach

Started by altering the dataset containing the closing price by dropping the columns that were not necessary and setting the crypto column as a column header with the closing price as the values having the date as the index. For the tweet data a data manipulation was performed by only showing the tweets with the corresponding hashtag into temporary data frames. Then concatenating the temporary data frames together into one, dropping the duplicates while keeping the last tweet. Since it's only taking the sum and not the individual hashtag it will not matter which hashtag is left after dropping the tweet. Finally summing up the total amount of tweets for each day. As a x value it was assigned as the total sum of tweets for a hashtag group. For the y value the corresponding closing price on that given day. Using a linear regression function from scipy we found the regression line which gave the answer to the question.

4.4.3.Results

Currency	Regression line	Total Tweets	Highest Closing Price
Cardano*	0.75	3228	0.19060
Ethereum	0.65	6079	752.53000
Bitcoin	0.49	16892	28972.40000
Tezos	0.38	603	4.36400
Vechain	0.36	3503	11.91000
Litecoin	0.30	2817	141.74000
Stellar	0.30	619	0.20690
Binancecoin	0.29	2534	39.04000
Ripple	0.27	8954	0.69400
Dogecoin	0.20	435	0.00479
Bitcoincash	0.15	856	511.04000
Monero	0.14	232	167.85000
Miota	0.07	3283	0.51560
Tether	-0.08	1730	1.02800

Table 2. Regression line with total tweets and highest closing price

Table 2 shows the total amount of tweets during the 2019 and 2020 period and the growth of the tweets by the regression line. Note that while Cardano* is the highest it is with faulty data since only 2020 was pulled when searching exclusively for tweets. Resulting in less tweets in 2019 and gathering the cardano tweets from out other currencies during the 2019 period which included cardano hashtags giving it the appearance of a high growth. When filtering out 2019 cardano does show a high regression line. Ethereum had the most tweets as the price went up. Tether did not grow as the other currencies because of the nature of the coin. It being a stable coin the price does not deviate from 1 all too much. Meaning tweets for that coin wouldn’t necessarily grow due to the low fluctuation.

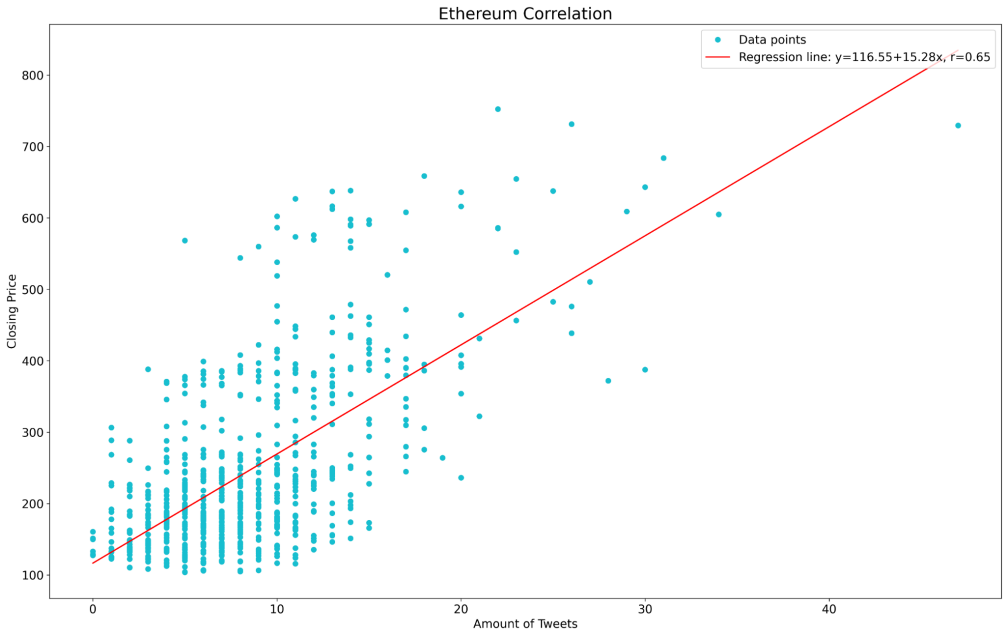

Figure 12. A linear regression plot of Ethereum

4.4.4.Discussion

The results have shown a weak positive correlation is present with all the crypto currencies except Ethereum, Cardano during the 2020 period and Tether where Ethereum and Cardano have a strong correlation. Proving as a crypto increases in its value the attention it receives from twitter increases alongside it.

The use of cleaning tweets by hashtags only may have skewed the results. If a word search of the tweet body were conducted possibly a greater volume of tweets could have been used to give a boarder picture. Due to the lack of tweets of Monero for example where with the results presented would imply it would have a greater number of tweets overall.

Reproduction of these results should fare well. Due to the knowledge of which hashtags were used. Possibly a limit of 50 likes removed would give a different view. People usually like a tweet if they enjoyed it. Assumptions can be made that a positive price increase would mean a higher volume of people tweeting on that day would receive more likes than of which of a price drop.

4.5.RQ4b

4.5.1.Datasets

4102_MASTER_FILE_CURRENCY.csv contains the closing price between 2019 and 2020 for the 14 cryptocurrencies

4108_SENTIMENT_explode.csv contains the tweets in the same period with the sentiment analysis applied

4.5.2.Approach

The tweets were separated once again by the hashtags. With the caveat of having the new sentiment columns added upon which contained the analysis column having values of positive, negative, or neutral. Each tweet is then separated by its value of the analysis column and then grouped by the date and summed up for that given date. The currency dataset is then joined with the altered tweets dataset creating 14 different datasets containing only the closing price of the relevant currency and tweets with the summed totals of each type of analysis tweet.

The use of the granger causality test was implemented by comparing the closing price as ‘x1’ to the total per day amount of an analysis value ‘x2’. Each test was done with a maximum of 100 lags due to memory issues when multiple granger tests were left in a notebook and no currency showed any increase of a P score going past 100. While the test is meant to show a relationship between two variables where one causes the other. This is not the case here. The intended use of the granger causality test in this research is to show the trend between the two variables. Once the test is completed the highest P score is documented and then applied to a graphing function where the tweets dataset is day shifted by the amount of lag shown to have the highest score.

4.5.3.Results

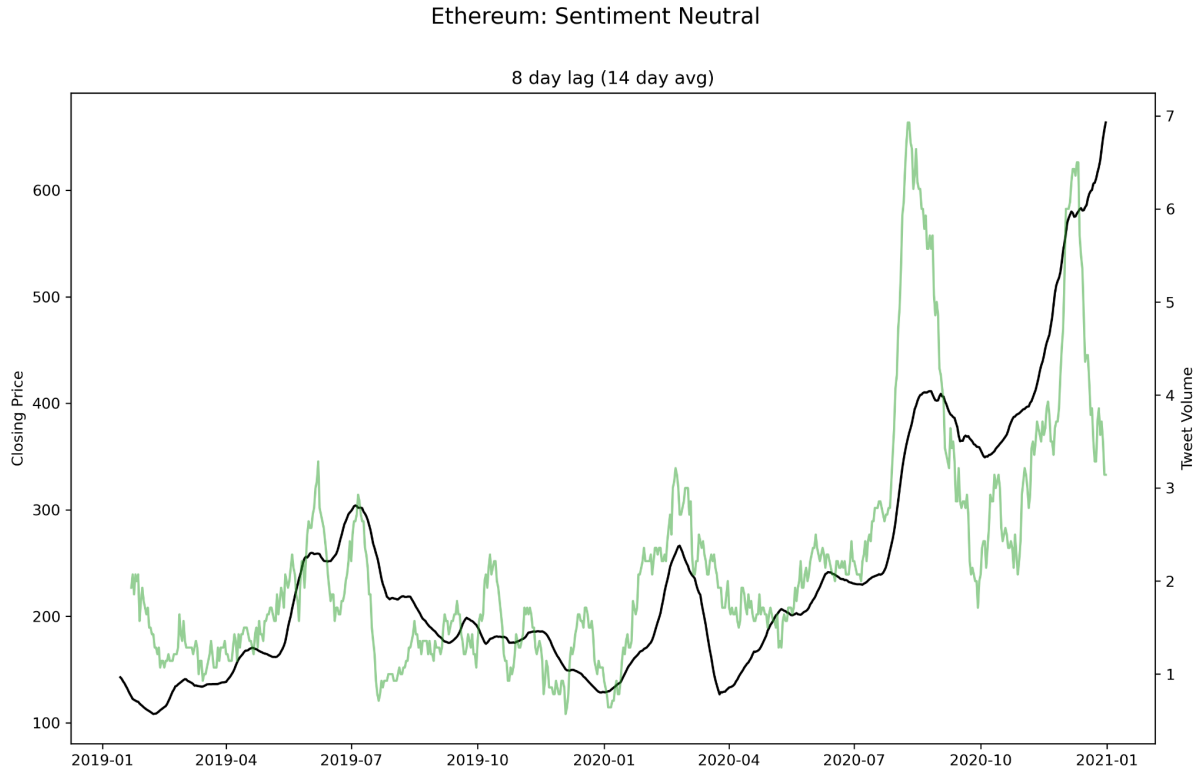

Figure 13. A time series showing the strong trends with the neutral sentiment of Ethereum

Ethereum, Bitcoin and Cardano are crypto currencies which you can spot trends with sentiment analysis. All three had strong trends with the neutral tweets. Ethereum and Bitcoin did not show any with the negative while Cardano did. For the positive Ethereum and Cardano showed strong trends while bitcoin did not. To note Cardano was only analysed during the 2020 period due to tweets only

being pulled from that year. While the rest of the currencies did not show any strong trends. None showed a strong negative trend excluding Cardano

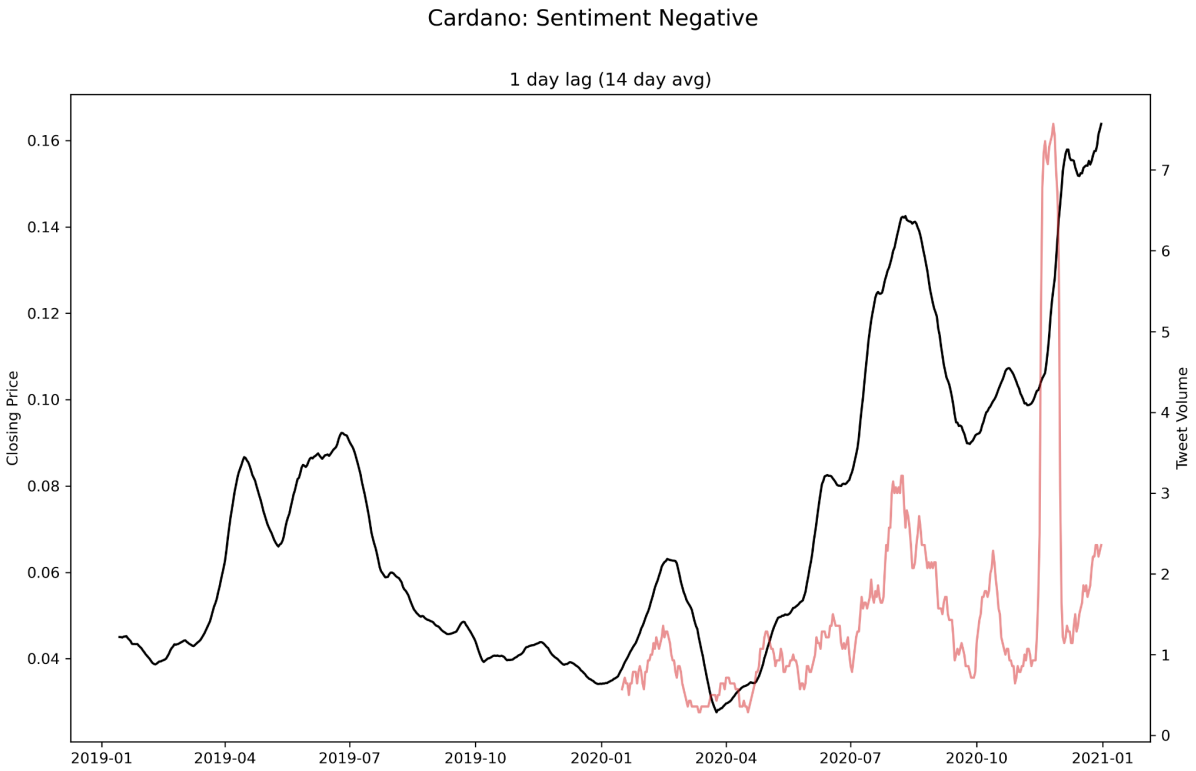

Figure 14. A time series showing the strong trend with the negative sentiment of Cardano with tweets only from 2020

4.5.4.Discussion

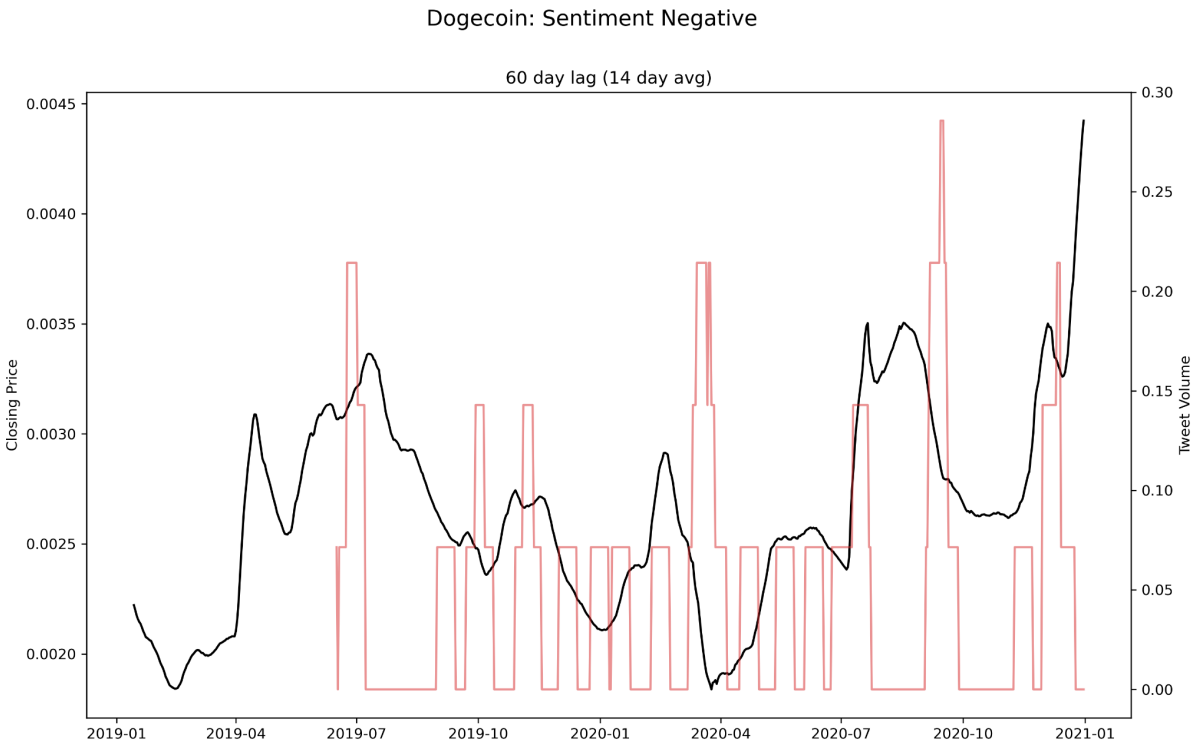

Figure 15. A time series showing affect a lack of tweets produces an inadequate graph

A few of the currencies did not have sufficient tweets for a proper analysis to be performed being tezos, stellar, monero and dogecoin. Many gaps had occurred where a jittery view was produced in the graph. As seen in figure 15.

	Positive	**Negative**	**Neutral**
Cardano	70.6	10.1	19.2
Miota	67.9	10.1	22.1
Stellar	67.2	10.5	22.3
Etheruem	61.5	11.5	27.1
Vechain	61.3	11.5	27.2
Litecoin	60.0	13.7	26.2
Bitcoincash	59.8	11.1	29.1
Dogecoin	58.9	9.2	32.0
Tezos	57.9	10.8	31.3
Bitcoin	57.8	16.2	26.0
Monero	57.8	12.1	30.2
Binancecoin	48.5	15.1	36.5
Ripple	47.5	11.5	41.0
Tether	32.8	13.5	53.7

Table 3. Percentage of sentiment values for each currency sorted by the positive sentiment

Table 3 shows that the currencies are not viewed in a negative light. Although a change in how the text were analysed may shift the neutral tweets into the other two types. Take note of tether and how it has the highest number of neutral tweets. It can be attributed to the nature of the stable coin.

The reasoning why the neutral tweets had the strongest trends and not the most percentage could be due to it not being tweets responding to the price itself but the interest in the currency. The positive may only rise when the closing price does and the negative being the opposite case. Further research could be done to view this

5.Discussion

5.1.Ethical Considerations

One ethical consideration in relation to this research is that research question three (4.3) was not investigated with the intention of profiting from a global pandemic. It is not something that we condone and the research question was only investigated for research and academic purposes. For research question four it was not our aim to gather the tweets in any ill mannered form where we could use any negative tweets against the user who produced them.

5.2.Reproducibility

The project environment can be reproduced as outlined in the myenv.yml file and the readme.txt file.

5.3.Limitations

For research questions one and three, only the daily high share prices of cryptocurrencies were really examined. With more time, daily opening, closing and low share prices would have gotten more attention. For research question three, the number of Covid deaths might have affected the share prices of cryptocurrencies a number of weeks later, with more time this could have been investigated.

As well as that, for research question four, the results would not be the same if the data was analysed for this year (2021). Had the steps to clean the data been known from the beginning, it would have been possible to get data relating to more cryptocurrencies for the purpose of linking tweets and hashtags to these cryptocurrencies and more time would have been spent analysing instead of cleaning the data.The threshold of only scraping tweets with a minimum of 50 likes may have hurt the coins in gathering the volume of tweets. Lesser-known coins might not be able to generate likes as bitcoin can but have a big enough following of people tweeting about it. In the end most of this research could be chalked up as which coin is most popular. Also, an unfair representation could have occurred. While giving a like in general is done when you enjoy a tweet. This could result in a skew in more positive tweets coming through for the sentiment analysis. If possible a new study could be conducted without the 50 likes limit. Having Cardano only have a viable 2020 period hurt the overall research since it can not be fairly assessed with the other currencies during the 2019 to 2020 period.

6.Conclusions & Future Work

In relation to research question one, we were a bit too ambitious in thinking we could find a significant link between share prices of various crypto currencies. Hence we were excited to find upon first examination that a lot of the correlation coefficients were close to 1. However, as outlined in section 4.1.2, this was a result of correlating the trends, rather than the values of the share prices themselves. Based on the results from section 4.1.3, we did not find any significant evidence to suggest that the share prices of cryptocurrencies have much of an influence on each other. Nor did we find any evidence to suggest that the share prices of crypto currencies on one day affect other crypto currencies a number of days later. With regards to future research, as mentioned in section 5.3, perhaps investigating daily opening, closing and low share prices may yield different conclusions. As well as that, perhaps the share prices of crypto currencies do influence each other but maybe a number of hours later, rather than a number of days later. So it may be worth investigating the hourly data of cryptocurrencies instead of just the daily data.

Regarding research question two, we initially aimed to investigate the common stigma that cryptocurrencies are too dangerous to invest in. To that end, we asked the question, is it safer to invest in crypto short-term or long-term? From the analysis performed, it was found that yes, cryptocurrencies are indeed extremely dangerous to invest in due to their extremely high volatility. For the most part, the currencies analyzed are highly trended, however, they displayed little to no seasonality. It was concluded that long-term buying is safer than short-term investment. The main suggestion for future work in this section is to do with the method of evaluation of a “safe” short/ long term trade. Variations of our work could uncover interesting results.

The research carried out for research question three revealed that Covid deaths in the U.S.A. did not affect the share prices of cryptocurrencies. However, as mentioned in section 5.3, the number of Covid deaths could have influenced the share prices of cryptocurrencies a number of weeks later. This is something that future researchers could examine.

For research question four it would be best to rename it as can trends be spotted using social media as the name would be more apt. The question of does social media influence cryptocurrency was not answered. The sub questions were answered though. Finding the volume of tweets do increase with the price is not something to be shocked by. It would make sense for more people to be interested in a currency which has a higher value so a correlation with the two was just proved. For the sentiment analysis to spot trends. Finding that it is possible but not for all currencies. Having Cardano have the strongest across the board was interesting but only for the 2020 period was a promising sight although if a proper data scrap was performed the results may have been different for an overall view. Having it performed with Ethereum it is possible to view a trend with the use of the granger causality test. The build upon this a limit of 50 likes to be removed I believe would be imperative to get a unbiased result reason being the same mentioned in the limitations

From the research performed, it has been shown that crypto does indeed pay out; however, there is a rather large caveat in that an individual requires a huge amount of luck to make it out of this dangerous market in one piece, so to speak. The cryptocurrency landscape has certainly earned its revered reputation and from this research, has even exceeded this notoriety.

7.Responsibilities

Cillian's primary research question was RQ2. He also worked with Eoghan to complete RQ3. Cillian wrote the following sections in the report: introduction, 4.2.RQ2, 4.3.RQ3 (with Eoghan), part of the conclusion.

Eoghan's primary research was conducted on RQ1 and worked with Cillian in investigating RQ3. Eoghan wrote the following sections in the report: 3, 4.1.RQ1, 4.3.RQ3 (with Cillian), 5 and part of the conclusion.

Aakrit primary research was conducted on RQ4. Aakrit wrote the following sections in the report: 2, 2.1, RQ4a, RQ4b and part of the conclusion.