THE CORRELATION OF ONLINE SENTIMENT AND THE TOP 100 CRYPTOCOINS BASED ON MARKETCAP

David Sciola[†], Nic Alarcon-Belanger[‡] and Yunas Magsi[§]
[†] 101082459, davidsciola@cmail.carleton.ca
[‡] 101066600, nicalarconbelanger@cmail.carleton.ca
[§] 101115159, yunasmagsi@cmail.carleton.ca

Cryptocurrency · Sentiment · Social Media

Abstract—**The objective of this report is to build upon the data collected in Phase II of the data science project, and provides detailed insight on the analysis conducted to answer the proposed research question of "Is there a correlation between online sentiment and the top 100 cryptocoins based on marketcap?" This report highlights the methodologies used to collect sentiment and crypto coin price data from Reddit and CoinMarketCap for the duration of one year of historical data. The collected data was used to perform several analysis types including sentiment analysis, correlation analysis, and predict future price action for each of the 100 analyzed crypto coins. The sentiment analysis resulted in several plots for each crypto coin of growing and decreasing sentiment over time. The correlation between sentiment online and crypto performance proved to be only slightly correlated. Finally, future price prediction models are not yet validated as they are made using data up to the current day as of publishing this report. The entire project is replicable with the compiled code available at the Github Repository linked here: https://github.com/AedynLadd/sysc4906-termProject along with the raw and cleaned data here: https://drive.google.com/drive/folders/1Tok4DHnzozXk81gdE1iu65B9ulrKpmJ3**

I. Introduction

This report has been created to provide details on Phase III of the CryptOutliers data science project. This report builds upon the two previous reports created for Phase I and Phase II of the project. Following the introduction, a background is provided, outlining the work completed in Phase I and Phase II, including the motivation for the project. Next, the problem definition is provided for the analysis question: "Is there a correlation between online sentiment and the top 100 cryptocoins based on marketcap?" Design and methodology are then considered, detailing the approaches taken to solving answering the question, including data collected, steps taken to clean and analyze the data collected from Phase II, analysis techniques and risks/ethical concerns. Following the design and methodology section, the results of the analysis are provided in conjunction with a discussion summarizing the findings. Finally, the report is concluded with recommendations to further guide the project in the future.

II. Background

With the advancement of social media in the past few decades, it has provided global connectivity and is fundamental to our society. In January 2021, individuals associated with the reddit community r/WallStreetBets, collectively bought and held onto equity shares of Gamestop ($GME), causing the price to exponentially grow which resulted in large hedge funds losing billions of dollars[1].

Due to the widespread use of social media, it has accelerated the rise of cryptocurrencies, with the first cryptocoin known as Bitcoin, being invented in 2009 [2]. Since 2009, Bitcoin has become the most popular cryptocurrency available to the public, with the highest cryptocoin market capitalization of $880 Billion USD as of April 2nd, 2022.

The use of sentiment analysis [3] to predict future fluctuations in the economy is not a new thing. Twitter has been used in the past to provide insight into markets behavior [4] and upon seeing how communities in reddit have been a driving force for market performance in the case of Gamestop, this report explores various subreddits to determine the overall sentiment of topics as they relate to the cryptocurrencies market performance.

The planned work differs from other previously done studies as the intent of this project is to focus on how sentiment analysis relates to and predict cryptocoin performance, as opposed previous studies being focused on the equity markets.

Phase I was focused on planning the high-level direction of the project, mapping the phases of problem statement, data collection, data cleaning, data analysis, and visualization/reporting. The project overview from Phase I has since been refined and is shown in Figure 1: Refined project overview.

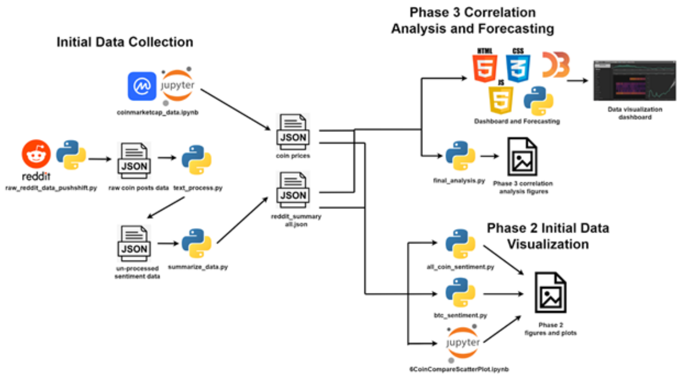


Fig. 1: Refined project overview

To compare social media's impact on cryptocurrency performance, two data sets were collected during Phase II of the project: Historical cryptocoin prices and social media sentiment. Historical cyrptocoin data was scraped from coinmarketcap.com using Jupyter Notebook script and online API. The script produced separate JSON files detailing the past 365 days performance of the top 100 cryptocoins by market capitalization size. Social media sentiment data was collected on reddit.com on the largest crypto-related subreddits by user-size. Data from each subreddit was first collected into individual JSON files with information containing posts titles, main body of text, upvotes, downvotes, and overall ranking. The data was then cleaned with multiple scripts such that sentiment scores on the basis of positivity, negativity, and neutrality were created for each of the 100 crypto coins across each subreddit investigated on a daily frequency.

III. Problem Definition

Inspired by r/Wallstreetbets historical influence, the purpose of this project is to attempt to draw a relation between online sentiment of users on Reddit and their influence on cryptocurrencies market performance. Specifically, the correlation of online sentiment and the top 100 cryptocoins based on marketcap is being investigated. Along with this analysis, a model will be created to forecast what future values of these coins will be based off the previous sentiment values. Market performance is defined using daily metrics including opening price, closing price, high of the day price, low of the day price, and volume. Volume is defined by the total dollar-value of transactions that occurred for the given time period. Market capitalization is defined by the a coins current price multiplied by the number of coins in circulation.

To further investigate into experiment. The following data will be used: the top 100 trending coins from coinmarketcap (ex. Bitcoin, Ethereum, XRP, Dogecoin, and 96 more). Reddit contains well over 2 million subreddits. To be able to parse a reasonable amount of data for the experiment, the sentiment analysis will be scraped from the following subreddits: CryptoCurrency, bitcoin, btc, CryptoMarkets, CryptoCurrencyTrading, Crypto General, blockchain, ico, icocrypto, and WallStreetBets.

IV. Design and Methodology

To solve the problem we have defined in the previous section, there are two major goals in mind. Firstly, tools are needed to be created capable of scouring reddit and coinmarketcap for data. And secondly, we needed to identify what information was relevant to us, and determine, how using this collected data, we can predict future expected prices.

Our success criteria is defined by the level of correlation our model achieves in its validation stage. Since the price of coins can fluctuate between changes in cents to thousands of dollars per day, we look to determine our models success through metrics such as its forecast bias, approximate residual forecast error, and its approximate mean absolute error. The successful model will be that of which has the lowest approximate mean absolute error across all cryptocurrencies measured.

A. Data Collection and Cleaning Process

To begin investigating the problem, we started by gathering social media data from Reddit embedded with the sentiment of users, as well as data related to the historical performance of various cryptocurrencies through CoinMarketCap.

1) CoinMarketCap: To gather data related to both the current and historical performances of a wide set of cryptocurrencies we used data gathered from CoinMarketCap.

The collection of data from CoinMarketCap begins by first gathering data on the most popular coins at the time (ranks 1-100) by using an api call at the address:

https://api.coinmarketcap.com/data-api/v3/map/all

Following this data collection is run for each of the coins individually to gather the necessary information on their historical performance. This is done using the unique identifiers for each coin as described by data collected in the previous step as parameters to a request at this address.

https://api.coinmarketcap.com/data-api/v3/cryptocurrency/historical

After making requests to these API endpoints, the data is aggregated and formatted to be used later on for future analysis.

2) Reddit Data: To gather sentiment data we turned to Reddit to get a wide range of possible data pertaining to the general public's thoughts and feelings on the performance of different cryptocurrencies. Some assumptions are made about the data collected, these assumptions are that scores indicate a polar level of agreement with the the post, and that all entities interacting with the site are unique.

While collecting data from Reddit, we chose to identify a cross-section of different subreddits we thought would most impact the performance of trade. These subreddits have a high amount of users, as well as main topics revolving around the trade of these coins. The data is first collected from each subreddit individually and stored in separate files - these files contain information including a posts title, main body of text, score, and overall ranking. By saving all this data individually the result is many smaller files that are quicker to parse through as interests change over time.

Separate scripts concatenates all these files and performs some analysis on the corpus. First, the title and main body of the each record is saved to a single entity as some posts have more descriptive titles than the data contained within them. Along with this we also save the scores of each post, this is taken as an indicator of how many unique individuals are in agreement with the sentiment we gather about said post.

One unique hurdle encountered is limits imposed by the reddit API. The native reddit api limits users to requesting the last 1000 items stored only. This worked fine for some subreddits such as r/icocrypto which have lower traffic, however subreddits such as r/CryptoCurrency have roughly 500 posts per day and 2 days worth of data is not sufficent for our use. To remedy this we turned to a database called pushshift. The pushshift API allows users to gather reddit data from the endpoint.

https://api.pushshift.io

Using this new API we were able to run exactly the same reddit searching script by just altering the endpoint and were able to collect up to 365 days worth of sentiment data for each subreddit we were interested in.

3) Data Cleaning: The data collected from CoinMarketCap was perfect as is for our purposes, no cleaning was needed other than some minor reformatting to make future analysis easier.

The data collected from reddit however was unorganized, specifically when it comes to the main text body that will be used in future sentiment analysis processes that will be discussed later on.

To clean reddit data we started by stripping bodies of text of any hyperlinks, and non-ascii characters. Following this we removed punctuation from the text and the text body is made to lowercase, then searches are made on the main body to identify what keywords are present. Keywords being noted here are the symbol (eg. BTC), the name (eg. Bitcoin), and the slug (eg. bitcoin). Identified keywords in the textual body are grouped by the symbol associated with that specific keyword (eg. Bitcoin/bitcoin map to the symbol BTC).

B. Analysis

1) Sentiment Analysis: To obtain sentiment values of a post we use the Valance Aware Dictionary for Sentiment Reasoning - VADER - a machine learning model used for text sentiment analysis which is both sensitive to the polarity and intensity of emotions found within a text post. The VADER model returns four fields: positivity, negativity, neutrality, and a compound score of these values.

2) Correlation analysis: The primary objective of the analysis was to determine if there exists any correlation or relationship between the coin prices obtained from CoinMarketCap and the sentiment values of those coins obtained from Reddit.

Before any analysis was to be done, the data from CoinMarketCap and Reddit needed to be converted to non-stationary data. This is due to the fact that both the time series data of the coin prices and sentiment values can potentially be stationary as is/if left unchanged. Stationary data simply means that the mean and amount of variance is constant over time, allowing for linear regression and correlation. Alternatively, the mean and amount of variance can change over time and this effect can render correlation conclusions like the linear regression and correlation coefficient calculations invalid since these calculations rely on the fact that the data has a constant mean and variance. Stationarity plays a key role when it comes to forecasting and predicting the future coin prices[5][6] .

To convert the non-stationary time series data to stationary data, the difference between the values of coin prices and sentiment values was used instead. This differencing removes the effect of trends or seasonality and ensure that the data has consistent mean and variance values. Note that taking the difference of a single time period does not always provide enough detail to fully ensure that the data is stationary. One possible method to test if data is stationary is the Augmented Dickey-Fuller test[9]. This test was used to determine if the coin price and sentiment data for each of the 100 coins was stationary to begin with. If the data was non-stationary then the difference would be continually taken until the data eventually became stationary[7][8].

Once the data is fully transformed to stationary data, a Durbin-Watson Test[9] was then performed on the data to ensure that there is no autocorrelation present within the data since the assumption of linear regression is that there is no autocorrelation. Any autocorrelation within the data means that there is a correlation between the data and the same data shifted by an amount which in turn invalidates the results of linear regressions performed with that data[10].

In the actual data analysis, the first correlation metric applied is Pearson's correlation coefficient. While Pearson's cannot be directly applied to time series data, if the difference of the data is taken, Pearson's can be utilized to quantify correlation. With that, the co-variance is also calculated to further quantify the amount of correlation[11].

Another method used to quantify the amount of correlation is cross-correlation. Cross-correlation is similar to normal correlation coefficient calculations except the difference is that the amount correlation is quantified for a range of offset or lag periods. This approach not only produces a cross-correlation coefficient, it also determines the amount of lag that produces that maximum amount of correlation[12].

Yet another method used to quantify the amount of correlation between coin prices and sentiment values is linear regression. The linear regression was performed on the stationary data for the coin prices and sentiment values because applying the linear regression directly to the non-stationary time series data is an invalid technique. The slopes of the lines of best fit are then be observed and any positive slopes indicate positive correlations between the coin prices and sentiment values.[12]

Finally, the last method used to quantify the amount of correlation between coin prices and sentiment values is cointegration. In particular the augmented Engle-Granger two-step cointegration test[9] is used to test whether or not cointegration is present between each of the 100 coins and their respective sentiment values. Any cointegration present is a good indicator of correlation between the two time series.

3) Vector Autoregression: In order to analyze the two time series we have, we made use of a statistical model called Vector autoregression (VAR) to analyze our multivariate time series. The goal is that sentiment and daily opening coin price influence each other through time. The order which gives our model the lowest Akaike information criteria (AIC) is then used to select the order of our model.

For the purpose of this report, the order selected is based off results found for only bitcoin. Future analysis would determine lowest AIC for each coin individually, however in hopes of creating a generalize model for predicting the future price based on sentiment alone using just the order from bitcoin seemed appropriate. The order of our VAR model based on lowest AIC among 10 iterations is a value of 3.

Then sentiment and open cost data is de-trended using its second difference, we are then able to create our validation model by predicting the ”N minus 7” days worth of future data. To measure our models accuracy, a variety of metrics are used as previously outlined. We define success for our model by having a value of less than 0.2 for our approximate mean average error across all coins evaluated.

C. Risks and Ethical Concerns

The Primary source of data used in the experiment is data extracted from Reddit. Reddit offers a free API with the condition that its usage is conducted with an open-source code and for a non-commercial product. The sentiment analysis of Reddit does not store any personal Reddit usernames to keep the users anonymous. With that being said, the CryptOutliers have fully intended to publish the project to be open source during and after the SYSC4906 course concludes.

The supplementary data extracted from CoinMarketCap offers a free API when registered with their developer account. The CryptOutliers currently have no intent for using it on a commercial intent, therefore are within the bounds of the CoinMarketCap terms and conditions.

To minimize risk and liability for both the users accessing the open-source project and the developers, a disclaimer is placed on the dashboard. This is to ensure that the users are informed before entering the website that the data is not intended to be used as financial advice at any point and to conclude their own research before conducting any operations.

V. Results and Discussion

This section aims to discuss the results of our tests related to the covariance of the variables being analyzed, their stationarity, and the results of the time series forecasting through the use of a Vector Autoregression model.

A. Correlation Output

From the correlation analysis, below will show the results that were generated.

The model below in Figure 2 shows a simple scatter plot which contains the coin price in CAD for 365 days vs the respective reddit sentiment score values for those 365 days.

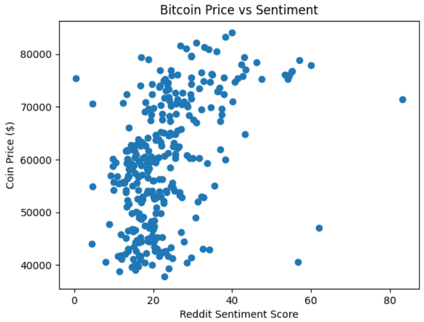


Fig. 2: Sentiment Score With Price

Below in Figure 3 is another scatter plot, except this scatter plot contains the difference in coin prices vs the differences in sentiment values. As mentioned previously, in order to perform correlation analysis on the two time series, the data must be stationary and this was accomplished by continually taking the differences of the data until the data passed the Augmented Dickey Fuller test.

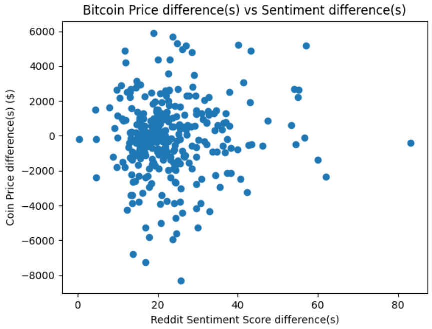


Fig. 3: Sentiment Score With Price Difference

Finally, in Figure 4 is the linear regression line of best fit applied to the previous scatter plot figure. The idea behind this figure is to be able to visualize the positive slope of the line of best fit where a positive slope would indicate a positive correlation. However, while the slope for this particular Bitcoin linear regression is positive, the actual slope value is relatively low which indicates a weak positive correlation. This trend of weak positive slopes was also common for all the other coins.

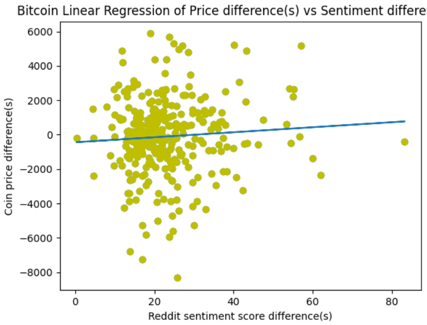
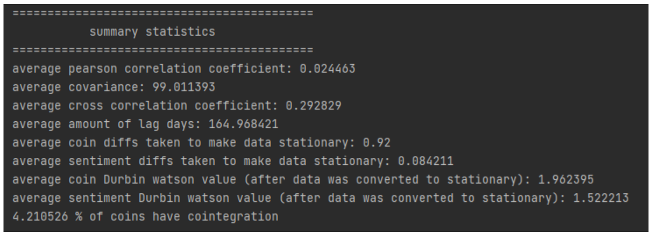


Fig. 4: Line of Best Fit Between Sentiment and Price

The overall generated value from the summary of the statistics of the list of 100 coins are as shown below. With that, the values are the averages across all 100 coins that produce an overall representation of the correlation between coin prices and sentiment values as a whole.



```
===========================================
          summary statistics
===========================================
average pearson correlation coefficient: 0.024463
average covariance: 99.011393
average cross correlation coefficient: 0.292829
average amount of lag days: 164.968421
average coin diffs taken to make data stationary: 0.92
average sentiment diffs taken to make data stationary: 0.084211
average coin Durbin watson value (after data was converted to stationary): 1.962395
average sentiment Durbin watson value (after data was converted to stationary): 1.522213
4.210526 % of coins have cointegration
```

Fig. 5: Summary of 100 Coins

From the results, the average Pearson correlation coefficient and average covariance were 0.024463 and 99.011393. While the average covariance value seems plausible, the average Pearson's correlation coefficient seems unexpectedly low considering the fact that Pearson's value ranges between -1 and 1. This unexpectedly low value however, goes conjointly with positive but relatively low observed slope values for the linear regressions.

The cross correlation the coefficient was 0.29 which indicates a slightly positive correlation between coin prices and sentiment values. While this coefficient value of 0.29 was expected, the average amount of lag days of 164.96 days was unexpected. This value of 164.96 days indicates that on average, an offset of 164.96 days produced the highest cross correlation coefficient across all the coins. While an amount of lag days is to be expected, 164.96 days is an unexpectedly high value. Somewhere within the range of 0-10 lag days seems more plausible.

Consequently on average the coin price sentiment data needed to be differentiated 0.92 times in order for it to become stationary data. This value of 0.92 indicates that overall the coin prices were non-stationary to begin with and this phenomenon is plausible in time series data like coin prices which can vary greatly over time.

Meanwhile the sentiment data only needed to be differentiated an average of 0.084211 times before it became stationary data. This simply indicates that the sentiment data was overall already stationary to begin with which is perfectly plausible.

In terms of the average Durbin Watson test values, the averages were 1.962395 for coin prices and 1.522213 for sentiment values. Since these values are relatively close to 2, this indicates that overall there is no autocorrelation present within the data.

Lastly the cointegration tests, 4.210526 % of the coins passed the cointegration test which means cointegration is present in 4.210526 % of the comparisons of the coins with their respective sentiment values. This value is relatively low, however, there is still some cointegration present which with the fact that overall correlation was detected between the coin prices and sentiment values but at low rates for most of the other methods used to quantify correlation.

B. Time series forecasting through Vector Autoregression

Upon running analysis using our vector autoregressive model, below are the results we've from our model specific to Bitcoin and Ethereum.

As seen in Figure 6 and 7 below, the model we created does a decent job in of predicting future values based off sentiment for bitcoin, we do however have a large forecasting error that needs to be taken into account. Other models such as that for ethereum depicted below are less optimized. The figures below depict the predicted values in dark blue, compared against the actual data in light blue. The forecast is given with an error window with upper and lower bounds defined by the lines flanking our forecast in red.

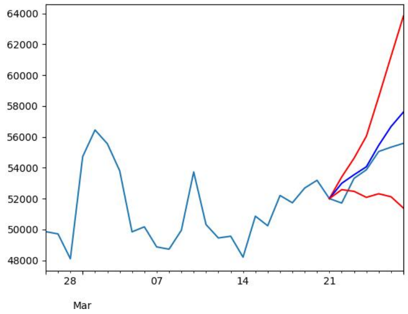


Fig. 6: VAR forecasting for Bitcoin

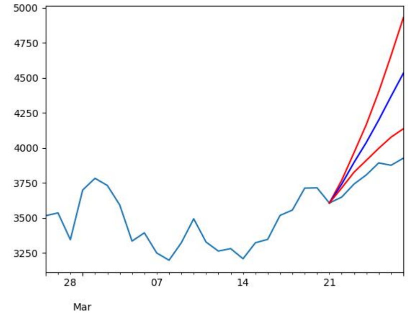


Fig. 7: VAR forecasting for Ethereum

The model analyzes data from a window of 135 days, and attempts to predict 7 days worth of data. Across all coins the average approximate MAE is 0.12 % and our forecast bias is -0.014 %. We use percentage as the value for measuring success of our model because cryptocurrencies range widely in their price, and it was the simplest way to compare the models accuracy.

A future goal would be to create an ensemble model to predict future prices. By using a combination of different statistical models, and machine learning models such as a RNN - more specifically an Long-Short Term Memory (LSTM) - we could take the accumulated result of all the different models and come to a more concise conclusion about what the forecasted values would be.

VI. Conclusion

Positive but relatively low correlation values were observed across the different methods used to quantify the relationship between coin prices and sentiment values. While slightly lower than expected, these tests indicate that sentiment values from Reddit and cryptocurrency prices are correlated. The correlation indicates that a causation relationship between sentiment values and coin prices may exist; however, future research is needed to confirm this hypothesis.

Being a time-limited project, only the daily opening prices were explored in terms of coin performance. Planned future work should include investigating the correlation between sentiment and the remaining price metrics of closing price, high of the day price, low of the day price, and volume. Additionally, further research can be conducted into other methods of correlating time-series data, while considering that online sentiment will either lead or lag price action.

REFERENCES

[1] J. Chung, ”Hedge Fund Melvin Lost $6.8 Billion in a Month. Winning It Back Is Taking a Lot Longer.”, WSJ, 2022. [Online]. Available: https://www.wsj.com/articles/melvin-plotkin-gamestop-losses-memestock-11643381321.

[2] Time.com, 2022. [Online]. Available: https://time.com/nextadvisor/investing/cryptocurrency/what-is-bitcoin/: :text=Bitcoin%20was

[3] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis, 2011. https://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf.

[4] Tim Smith and Charlene Rhinehart. Market sentiment, 2021. https://www.investopedia.com/terms/m/marketsentiment.asp.

[5] S. Chaudhari, “Stationarity in time series analysis explained using python,” Quantitative Finance amp; Algo Trading Blog by QuantInsti, 17-Feb-2021. [Online]. Available: https://blog.quantinsti.com/stationarity/. [Accessed: 01-Apr-2022].

[6] “Spurious regression ,” Youtube. [Online]. Available: https://www.youtube.com/watch?v=nt1h8iR0Sac. [Accessed: 04-Apr-2022].

[7] S. Wu, “Stationarity assumption in time series data,” Medium, 04-Jul-2021. [Online]. Available: https://towardsdatascience.com/stationarity-assumption-in-time-series-data-67ec93d0f2f. [Accessed: 03-Apr-2022].

[8] “Python tutorial: Making time series stationary,” Youtube, 03-Mar-2020. [Online]. Available: https://www.youtube.com/watch?v=bP1fbXd$_XSk.[Accessed: 01-Apr-2022]. Zach,\backslash Augmenteddickey-fullertestinPython(withexample),"Statology,25-May-2021.[Online].Available: https://www.statology.org/dickey-fuller-test-python/.[Accessed: 01-Apr-2022].$

[9] “Autocorrelation,” Corporate Finance Institute, 12-Apr-2021. [Online]. Available: https://corporatefinanceinstitute.com/resources/knowledge/other/autocorrelation/. [Accessed: 01-Apr-2022].

[10] Zach, “How to perform a Durbin-Watson Test in python,” Statology, 21-Jan-2021. [Online]. Available: https://www.statology.org/durbin-watson-test-python/. [Accessed: 29-Mar-2022].

[11] Glen$_bGlen_b\backslash HowtousePearsoncorrelationcorrectlywithTimeSeries,"StackExchange,01-Nov-1962.[Online].Available: https://stats.stackexchange.com/questions/133155/how-to-use-pearson-correlation-correctly-with-time-series.[Accessed: 01-Apr-2022]. A.Hayes,\backslash Whatiscross-correlation?,"Investopedia,30-May-2021.[Online].Available: https://www.investopedia.com/terms/c/crosscorrelation.asp.[Accessed: 02-Apr-2022].$

[12] “What is the difference between correlation and linear regression?,” GraphPad, 03-Oct-2019. [Online]. Available: https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/. [Accessed: 02-Apr-2022].