

*Erasmus School of Economics*  
*University of Rotterdam*

---

---

*The impact of Twitter sentiment on bitcoin financials*

---

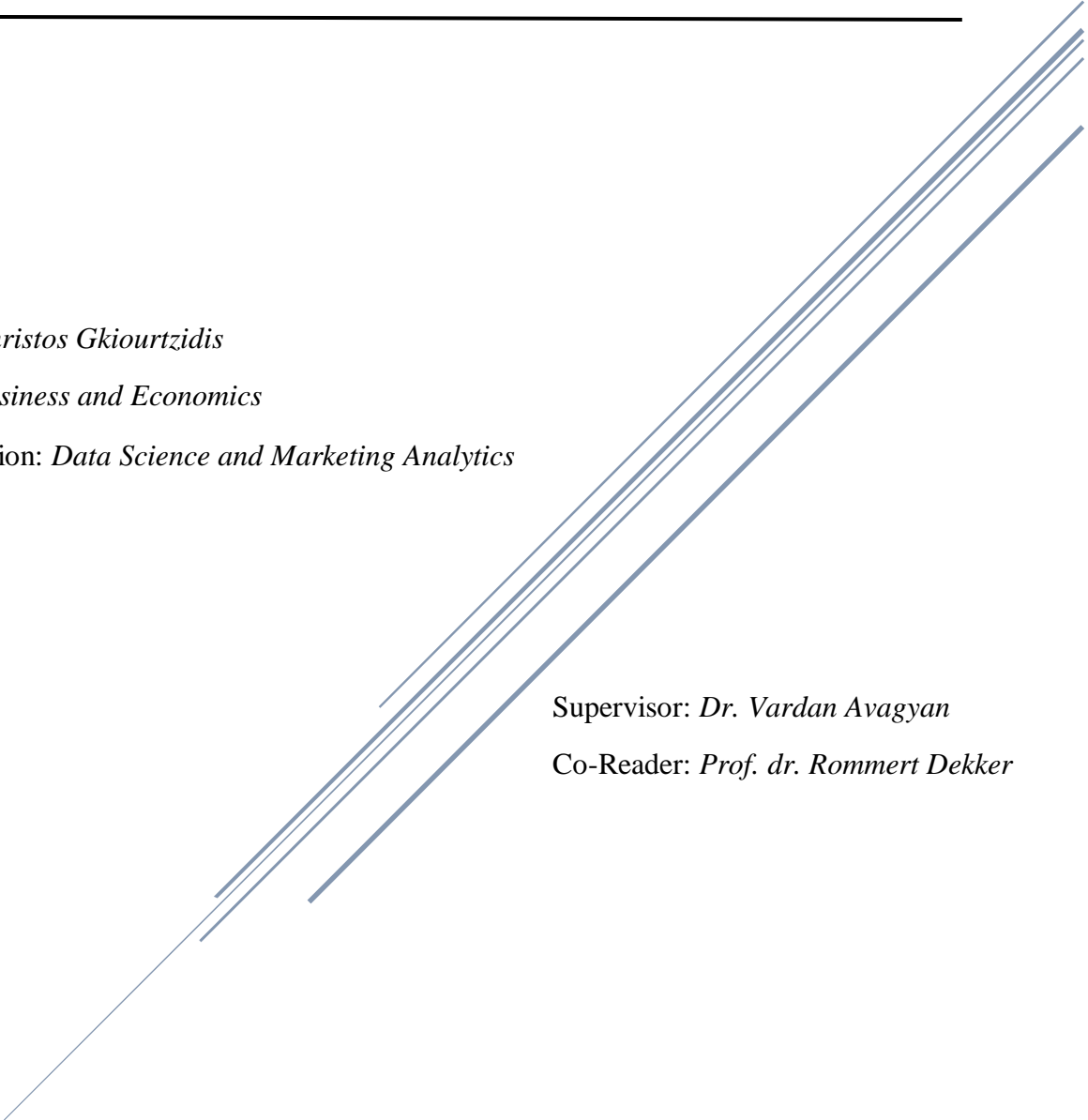
Author: *Christos Gkiourtzidis*

Master: *Business and Economics*

Specialization: *Data Science and Marketing Analytics*

Supervisor: *Dr. Vardan Avagyan*

Co-Reader: *Prof. dr. Rommert Dekker*



# Abstract

The present study investigates the relation between Twitter sentiment and bitcoin financials. After successfully cleaning and detecting bots, the posts are examined in order to extract the overall sentiment of each post. We focus on the period between 28/02/2021 and 23/06/2021. During that period, the price of bitcoin had many ups and downs, thus we want to know whether one of the reasons for this movement was Twitter. After extracting the overall sentiment of the period, we conduct a time-series analysis in order to test some hypotheses. After using VAR models, the results show that sentiment does not affect the price of bitcoin during that period, but affects the bitcoin returns with a lag of 2 past days. Moreover, evidence show that the tweet volume affects trading volume. After conducting the granger causality test, we can infer that the relationships are one way which means that a variable X granger causes another variable Y but not the other way-around. Finally, we use predictive models to forecast the future returns. By doing that we try to examine which model has better performance when it comes to predictions. The results show that ARIMA has the best predictive performance among VAR and SVM for this particular analysis. Based on the results, we provide relevant recommendations for investment companies and individual investors and traders.

*Keywords: Sentiment, Vader, VAR, Bitcoin, Twitter, time series.*

# Table of Contents

<b>1</b>	<b>Introduction</b> .....	<b>4</b>
1.1	Problem Statement and Research Questions. ....	4
1.2	Academic Relevance .....	5
1.3	Managerial Relevance .....	7
<b>2</b>	<b>Literature Review</b> .....	<b>8</b>
2.1	User-Generated-Content (UGC) .....	8
2.2	Social media .....	9
2.2.1	Social media: <i>Twitter</i> . ....	10
2.3	Cryptocurrencies. ....	11
2.4	Sentiment Analysis. ....	12
<b>3</b>	<b>Conceptual model</b> .....	<b>14</b>
3.1	Hypothesis rationale. ....	14
<b>4</b>	<b>Data and Methods</b> .....	<b>16</b>
4.1	Data collection. ....	16
4.1.1	Twitter data. ....	16
4.1.2	Cryptocurrency financial data. ....	17
4.2	Data Processing Approach. ....	18
4.2.1	Text processing .....	18
4.2.2	Sentiment Analysis (VADER) .....	19
4.3	Econometric Methods. ....	19
4.3.1	Akaike information Criteria (AIC) .....	19
4.3.2	Augmented Dickey-Fuller test (ADF) .....	20
4.3.3	Vector Autoregression (VAR) .....	20
4.3.4	Granger Causality test. ....	21
4.3.5	Impulse Response test. ....	21
<b>5</b>	<b>Results</b> .....	<b>22</b>
5.1	Simple Analysis. ....	23
5.2	Descriptive statistics. ....	24

5.2.1	Tweet volume. . . . .	24
5.2.2	Trading volume. . . . .	25
5.3	Pairwise correlations. . . . .	27
5.4	Lag selection. . . . .	28
5.5	Augmented Dickey-Fuller test. . . . .	29
5.6	Vector Autoregression model( VAR) . . . . .	30
5.7	Granger causality test. . . . .	32
5.8	Impulse Response Function. . . . .	34
5.9	Summary of main findings. . . . .	35
<b>6</b>	<b>Further Analysis: <i>Predictions</i>. . . . .</b>	<b>37</b>
6.1	ARIMA. . . . .	37
6.1.1	ARIMA results. . . . .	38
6.1.2	VAR results. . . . .	38
6.2	Support Vector Machines (SVM) . . . . .	39
6.2.1	Support Vector Machines results. . . . .	40
6.2.2	Future returns. . . . .	41
<b>7</b>	<b>Conclusion and Discussion. . . . .</b>	<b>42</b>
7.1	Main findings and discussion. . . . .	42
7.2	Managerial implications. . . . .	43
7.3	Limitations and future research. . . . .	44
	<b>References . . . . .</b>	<b>45</b>
	<b>Appendix. . . . .</b>	<b>48</b>

# 1 Introduction

## 1.1 Problem Statement and Research Questions

Technology has evolved during the past years, and it has become much easier for people around the world to generate and share content across social media platforms. While social media platforms have changed during this period, the number of users has skyrocketed. The introduction of web 2.0 technologies has developed a new era in social life, making the use of the internet and social media activities easier for everyone. That means that people are not just receiving content, but they also create and share content among the other users.

That revolution has changed many aspects of everyday life. From a financial view of point, people -ordinary users and not just investors- are sharing posts and opinions on social media platforms related to stocks, trades, and investments. So, there is much information laying in the social networks ready to be analyzed not only by businesses but also by individuals and stakeholders. The most known social media platform which is used for this kind of research is *Twitter*.

Twitter and other social media networks are great sources of information that researchers and institutions use to collect and analyze public opinions (Kouloumpis et.al, 2011). That could happen with many methods but the most known one is sentiment analysis which is part of data mining. With the data mining techniques, researchers gain information from the text and bring meaningful insights. The need for social information gain derived mostly because people, nowadays, use the internet almost every day (Pang & Lee, 2008). Online reviews, recommendations, ratings, and opinions about services have become an important asset for companies that try to enhance their already existing products, try to market new products, and try to place their products better than the competitors. But companies are not the only stakeholders that care about this information. Individuals, traders, and investors are also interested in information for financial purposes. People prefer the sentiment analysis methods because is faster and easier to extract financial information and main opinions through social media networks.

For this study, we examine cryptocurrencies and specifically, bitcoin. Cryptocurrencies were created in 2009 and are digital assets that are not issued or supervised by governments or institutions. This particular market is interesting and hides many opportunities for future

development. Cryptocurrencies are operated in a unique system called blockchain which makes the transactions secure and fast. This market became famous in the last two decades so not much is known about this trend (LIU & WU,2022). One way to extract information is the sentiment analysis on social platforms which may help people learn more about the crypto market. The market is driven by unknown factors and that can be seen from the extreme market volatility (White,2015).

People believe that internet and its features, such as social media, can influence other people. They also think that information from social media can be used to predict future returns. This study tries to infer the magnitude of the relationship between social media and bitcoin. The main question is whether Twitter information drives the price of bitcoin or whether the price of bitcoin drives the sentiment of people. For this reason, the sentiment and the volume of twitter are examined as Twitter information along with the bitcoin financials.

Finally, this study tries to find the best way to predict future returns by conducting different predictive models. By doing that, we want to examine which model has better performance when it comes to predicting the future returns of bitcoin. To do that, methods such as ARIMA, VAR, and SVM will be used in order to find the one with the best predictive power. For that reason, the next questions will be answered:

- (1) *What is the relationship between bitcoin tweets and bitcoin price?*
- (2) *What is the magnitude of the potential relationship between Twitter and bitcoin?*
- (3) *Do positive sentiment increase the price or the returns of bitcoin?*
- (4) *Which method predicts the returns of bitcoin with higher accuracy?*

The next section provides an extensive review of the literature surrounding our topic of interest, about the bitcoin, as a cryptocurrency and as a financial market, the user-generated-content, and the methods of gathering social sentiment based on social media platforms, while also examining the academic and managerial relevance of the study.

## 1.2 Academic Relevance

Nowadays, extracting data from social media networks has become significant to make meaningful insights for businesses and individuals. The process can be done with sentiment analysis methods from users of the net, and their online content. That content could be posts, messages, and opinions. There are many social platforms to which that sentiment could be applied but the most famous one is Twitter.

This thesis contributes to the already existing literature in many ways. First, prior literature showed that twitter information can be used to create time-series data by extracting meaningful variables such as sentiment and volume to look for relations to other subjects, which has been successful for elections (Tumasjan et al., 2010) and stock market returns (Li et al., 2018; Bollen et al., 2011). In this study, we try to find whether social media has informative value in determining market movements. This is because we focus on a new trending market in which we don't know whether social media estimators can be considered as related to the prices of cryptocurrencies.

Second, this analysis uses a combination of different methods to develop a representative dataset. That's because sometimes tweets are identified as bots, which must be detected and removed to have a clean dataset to analyze. It is believed that more than 25% of the posts are identified as bots which most probable are generated for scam, commercial and phishing reasons. For this study, almost 250.000 posts were detected as bots and removed.

Third, this thesis also contributes to the knowledge of the interaction between the cryptocurrency market and social measures. For this study, we make use of time series data, and thus it's important to apply a method of analysis that is a good fit for this data. That's why for this study we use a vector autoregression (VAR) model to predict variables on lagged values of multiple variables (Jarocinski and Mackowiak, 2017).

Cryptocurrencies have become known in the past two decades and not much work has been done to understand the market. That's why sentiment analysis is believed to have a large impact on crypto prices. This study attempts to find the relationship between sentiment from Twitter and bitcoin price if there is any.

Finally, this study contributes to the already existing work of people that want to predict the future bitcoin financials based on sentiment analysis.

### **1.3 Managerial Relevance**

As mentioned before, the cryptocurrency market is a new trend that became famous during the last two decades. That means that little research has been done about the market and specifically what drives the price of cryptocurrencies, the volume, and the sentiment of people. So, individuals and companies are not provided with the amount of information needed for investment or other purposes.

This thesis could enhance the knowledge of companies and investors by examining to what extent social media content can be used for investment purposes. This thesis is relevant for individuals because more and more people are using social media networks to search for information, relevant to cryptocurrency trades and transactions. That means that if there is any kind of relation between Twitter and bitcoin price, then people would know to look more for news and information on Twitter channels, pages, and posts. Thus, if Twitter is positive about bitcoin in a certain period, that would probably mean that this period is good for investing in bitcoin.

On the other hand, if we prove that sentiment is relevant and important for the cryptocurrency market that would give investors a dynamic way to plan their investments. This thesis also contributes to managers in the following way. First, they could use sentiment analysis techniques to measure people's feelings for different products and services. By doing that, they would know what people think and feel about their products and/or services. Thus, they could act by developing the product or making changes to be approachable to the customers.

Moreover, this study could also help researchers to build models to collect information automatically and then predict future prices of cryptocurrencies, and other products and services, too. By constructing AI systems to predict future prices, people in companies, save time and money because of automation. But the most important thing they gain is information. That information can be sold to stakeholders and can be used for investment purposes.

There are many ways that this analysis can be used, and many people can be benefited from their purposes. Individuals can use this information not only for investment reasons but also as a source of knowledge. On the other hand, managers can use this information to know what people think and feel about their products/services and, of course, for investment purposes.



## **2 Literature review**

This section provides a theoretical framework of several topics of relevant research for this study. An analysis of already existing literature is conducted on the following topics: social media and its influence, user-generated content, cryptocurrencies with a focus on Bitcoin, and finally sentiment analysis techniques. First, the literature about user-generated content is analyzed, followed by the social media influence. Then, cryptocurrencies are analyzed before we review the sentiment analysis techniques.

### **2.1 User-Generated-Content (UGC)**

By the end of the 20<sup>th</sup> century, new technologies have improved the already existing web services with new, more intelligent services and have enabled users to contribute to social media and interact with other people across the media world (Naab & Sehl,2017). Nowadays, users produce content and could customize, share, and develop it among the other users (Naab & Sehl 2017; Daugherty et. al 2008). UGC is mainly characterized by a feeling of personal contribution by simply creating, receiving, and sending content. For example, that kind of contribution could be comments on online articles or products, reviews on blogs, and many more (Naab & Sehl 2017; Krumm et.al 2008).

Daugherty et.al (2008) have advised that user-generated content is very useful and influential, especially in the final stage of the customer's journey, where the customer is one step before making a purchase or taking a decision upon something. Truson et.al (2010) suggested that the social influence of a person occurs when their behavior, characteristics, and beliefs are adapted by another person with the same attitudes. The influence could be caused because of emotional content, informational content, or other reasons. Meire et.al (2019) found that while emotional content has a positive influence on digital engagement, informational content has a larger positive influence on the outcome. Their study showed that social media can be used as an effective way of marketing tool. Moreover, the higher the number of activities and connections a user has, the higher the influence (Trusov et.al 2010).

Finally, Banerjee et.al (2021) studied a trend that occurs on social media platforms called question and answer(Q&A). That form is a common form of UGC that allows users to ask

questions about products/services and receive responses either from the platform itself or the other users. The whole concept of UGC can be beneficial not only for the creator of the content but also for the society in general such as other users, customers, and companies. The former enjoys the recognition that receives from the others, while the latter gains information/opinions and turn them into meaningful insights (Krumm et.al 2008).

## **2.2 Social media**

Social media is a form of communication based on the internet. Many social media platforms allow users to interact with each other by sharing, creating, and sending/receiving content. This kind of content could be blogs, micro-blogs, photos, messages, and more. Billions of people around the world are using social media platforms, especially nowadays when access to the internet is easier. But social media platforms are also a way of information gaining. Many people share their opinions about products or services and other people have the first sense of a particular thing they want to try/buy. Many people claim that social media platforms have a strong influence on people's lives and choices (Peng et.al 2018).

Peng et.al (2018) studied that social media platforms keep the potential to reshape the consumer's way of generating, spreading, and sharing content because of their strong ability to connect users. Moreover, they suggested that improving the content that is shared on platforms could lead to more efficient marketing campaigns by the firms. That comes along with the study of Schweidel & Moe (2014) which showed that the outcomes the marketing researchers obtain from the social media platforms may lead to a better understanding of customer habits and also, influence them in positive or negative ways.

Moreover, in these studies, other researchers studied the effect of emotion on social media and its influence (Lee 2021; Berger et.al 2012). Lee (2021) studied emotionality which is the language conveying that can be expressed by photos, punctuation, and strong words in social media posts. Lee (2021) tried to understand social media marketing by exploring the emotion contained in the post of the users and whether this emotion can affect the brand status. He found that reduced emotionality can increase adjustment to high-status communication norms. On the other hand, Berger et.al (2012) expressed the necessity of emotion in content to become viral.

Their study showed that positive content and positive emotion are more viral than negative ones, but the relationship between the emotion and the social transmission is more complex than valence alone. People share information with their feelings inside because they think that this information is useful for others (Berger et.al 2012; Peng et.al 2018; Shweidel & Moe 2014).

There is no doubt that internet-based platforms can have a huge amount of data and that information can be used to gain information through peoples' opinions. People are sharing their opinions on social media platforms about many topics, such as economic, social, and political topics. Some researchers studied the emotion of people when talking about the stock market and economic issues.

Antweiler & Frank (2004) studied whether posts and social media content posted on stock boards, affect the stock market itself. Their study indicated a strong positive correlation between posts and stock volume. They also found that message posting helps to predict volatility. Finally, they discovered that those posts have a positive correlation on the price but, on the following day. Moreover, another researcher Tetlock (2007) found that the content of the stock market news could be linked to investors' psychology. The study showed that high media pessimism could lead to low market returns, whereas high media pessimism predicts downward pressure on market price.

All these studies show that there is much information on social media platforms and people can use it to gain information not only about products and/or services but also for economic purposes, such as investments. Nowadays, there are many social media platforms, but most of the time (depending on the case) people choose to make their analysis or gain information through *Twitter*, which is discussed in the next section.

### **2.2.1 Social media: *Twitter***

Many people believe that Twitter is the best social media platform among the others to make a social sentiment analysis. Twitter ([www.twitter.com](http://www.twitter.com)) is a social networking and online news platform in which people can post "tweets" which are microblogging messages. They can post on their profile or other users' profiles, or even on company pages and groups. People can follow other users, can share content with them, can interact on posts and photographs. But what differentiates microblogs from normal blogging is that microblogging fulfills the need of the user to share information or opinions at an even faster rate. Twitter has become one of the leading platforms for

individuals and businesses to communicate and share information on a variety of topics. Nowadays, Twitter is also used as a mean for people to express their opinion not only for social content but also for economic. Many people believe that Twitter posts can contain meaningless information when it comes to economic decisions because people are driven by their own thoughts.

Kouloumpis et.al (2011) found that social platforms such as Twitter have been benefited by the growth of the internet and companies/media organizations are seeking ways to mine information about the opinions and feelings of the people. They also found that not all parts of the "speech" in posts are meaningful for sentiment analysis. That research comes along with the findings of Antweiler & Frank (2004) who found that there is financially relevant information in posts.

Other studies showed that researchers used Twitter data to predict different outcomes. Huberman (2010) used Twitter posts to predict future box-office revenues. The study shows that social media can be used for many real-world issues and future predictions. Moreover, Bollen et.al (2011) have built a model to predict the future movement of the DJ index based on Twitter messages with an accuracy of 87.6%.

These studies examined the prediction power of Twitter and found many different and informative results. Making use of Twitter data resulted in relatively accurate predictions and many outcomes for future analysis. It is undyeable that Twitter has a strong power over sentiment because of the large number of people that are sharing their thoughts on the platform. That is the reason researchers and not only, try to mine information from Twitter for future development purposes.

## **2.3 Cryptocurrencies**

Less than two decades ago, a new form of digital asset named cryptocurrency has been built. People claim that this digital asset can provide value to the holder in combination with interesting insights into their behavior (White,2015). Cryptocurrencies are a P2P digital asset that is encrypted. The advantage of this currency over the traditional form is that cryptocurrencies are not controlled by governments or companies, while they cannot be exchanged for any government fiat money (White,2015).

The very first cryptocurrency, called *Bitcoin*, was introduced back in 2008 by an anonymous person called "Satoshi Nakamoto" and came into existence in 2009. Since then, there was a dramatic rise in the cryptocurrency market. Nowadays, there are more than 50 million active investors that trade cryptocurrencies with more than 70 exchanges around the world. Through these years, there were significant ups and downs in both the price and volume of the currencies (Makarov & Schoar,2020).

Bitcoin is the first and currently biggest cryptocurrency with a market capitalization of \$912bn as of March 2022. Nowadays, the estimations of different cryptocurrencies are roughly 2.000, named Altcoins (alternative coins to Bitcoin). Bitcoin accounts for roughly 40% of the total market valuation of all cryptocurrencies combined. (Coinmarketcap, March 2022).

The most impressive and unique feature of cryptocurrency is the technology that supports it. The unique technological feature of Bitcoin is the public distributed ledger, called *blockchain*, containing all Bitcoin transactions. It's more like a financial book that holds all the information that has been done from the begging of the coin. That means that it is not necessary for central authorities to handle the process, creating a new way of decentralized finance. The blockchain needs to be maintained and that happens by a huge network of nodes running the Bitcoin protocol (LIU,2022).

Cryptocurrencies are one of the most innovative assets that human beings ever created. Cryptos have changed the lives of many people around the world and in combination with the fast-growing technology, they are developing even more. We do not know which are the limits of this new market and if that market will be useful for not only the economic development but also the social development. That's why people need to understand more about cryptocurrencies and make meaningful insights for a better estimation of this unique market. One of the ways to do that is by sentiment analysis on social media platforms in order to extract the opinions and feelings of the people.

## **2.4 Sentiment Analysis**

Sentiment analysis is the interpretation and classification of emotions within a text, to identify the attitude towards a particular subject as positive, negative, or neutral. Sentiment analysis is a powerful tool that is used to a wide range of problems that are of interest to human beings, such

as social, economic, political, psychological, and many more. Sentiment analysis is an active study in the field of natural language processing (NLP) that analyzes opinions, feeling, attitudes, and emotions (Pang & Lee,2008). Natural language processing is an area of computer science, AI, and linguistics based on computers that explains how computers understand and analyze natural texts provided by a human. Cambria & White supported that NPL is a theory-motivated range of techniques done by computers for the analysis and interpretation of human language. They also proposed that in a web where users share and generate content, the need for sensible computation and opinion mining is increasing dramatically

Many sentiment techniques can be implemented. The most common techniques are machining learning, lexicon, and knowledge-based methods. The lexicon method assumes that the sum of each word, which is based on a predefined list of words and measured on polarity or strength, gives us the sentiment (Liu,2010). On the other hand, the most common machine learning techniques for sentiment analysis are Naïve Bayes, Support Vector Machines (SVM), and Entropy, which are methods that require a trained dataset and then, the evaluation of specific features to be classified. Finally, in the knowledge-based method, the text is classified in a particular sentiment class based on words such as happy, sad, and afraid which we find in the text. This technique makes use of a lexicon that allocates a score to a huge number of words. Then, by checking the presence of these words in a piece of text word-by-word, a sentiment score is given (Cambria et al., 2017).

A very useful and important rule-based lexicon named “Valence Aware Dictionary and Sentiment Reasoner” has introduced by Hutto and Gilbert (2014). This lexicon uses both qualitative and quantitative methods, while a list of lexical features is constructed, primarily for sentiment analysis in the microblogging context. It has been shown that VADER performs better than eleven other sentiment analysis tools. VADER can detect the polarity and sentiment intensity in texts (Hutto & Gilbert,2014). Moreover, VADER was developed by Hutto & Gilbert as a solution to slang (“*Gotta*”, “*ASAP*”), special symbols, emojis, and writing style which is often used in a social media context.

Sentiment analysis is widely used in opinion mining from social media platforms and not only. This kind of analysis is very important because it requires computers to analyze and understand patterns and insights through texts. In a fast-developing world where the opinion and

feelings of the people matter for many purposes and topics such as economic, social, and political, individuals and institutions need to gain information to improve their already existing products and services. Thus, sentiment is a valuable tool that anyone can use for many real-world related problems.

## **3 Conceptual model**

### **3.1 Hypothesis rationale**

The first hypothesis is about tweet sentiment and the returns of bitcoin. Bollen et.al (2011) have shown from their studies that there is a positive relationship between investor sentiment and the stock market

Moreover, they found that sentiment analysis on Twitter data can have high predictive power on the returns. They also collected tweets on public mood states, which later used to predict the movement of the Dow Jones Industrial Average (DJIA) with an accuracy of 86,7%.

Sprenger et al. (2014) examined the relationship between social content and stock market returns. The researchers found that positive sentiment is related to increasing stock prices. Moreover, Sprenger et al. (2014) found that if the market is dominated by bulls (bullishness), then stock prices are increased. When the researchers performed lagged regressions, however, they found that positive sentiment could not be used to predict stock prices.

Even though research indicates a relation between investor sentiment and the returns, the magnitude of this relationship is not yet determined. We can infer that the relationship between sentiment and returns are tighter in markets where the volatility is high, thus it seems logic to assume that returns and sentiment are related in the cryptocurrency market. Thus, the first hypothesis will be:

*H1: Tweet sentiment influences positively the bitcoin's return*

The second hypothesis is about the tweet sentiment and the price of bitcoin. Therefore, the second hypothesis investigates whether the price of bitcoin is influenced by the overall sentiment around bitcoin captured by the social sentiment analysis done on Twitter. Previous research has shown

that bitcoin-related tweets can predict the movement of bitcoin's price with an accuracy of 62,48% (Sattarov et. al,2020).

Even though Sprenger et. al (2014) found that positive sentiment could not be used to predict stock prices, the cryptocurrency market is relatively new and needs more investigation and research to have more robust results.

This hypothesis tries to find whether there is any relationship between investors' social sentiment measured through Twitter and bitcoin price. Moreover, this hypothesis tests if tweet sentiment can be a predictor of future prices. After that a possible magnitude of the relationship could be inferred .Thus, the second hypothesis will be:

*H2: Tweet sentiment influences positively the bitcoin's price*

The third hypothesis is about tweet volume and bitcoin volume. Bollen et al. (2011) found that tweet volume can be used as a predictor of future election results and stock market financials. Sometimes, an increase in online message volume shows that some new information is discussed. In the financial world, that could possibly mean a signal to start trading.

Antweiler & Frank (2004) showed that the volume of internet economic-related messages can predict trading volume. To do that, they examined the information they found on Yahoo! Finance and Raging Bull boards for 45 companies of the Dow Jones Industrial Average. The results demonstrate that trading volume was successfully predicted by message volume. Moreover, the magnitude of the relationship was positive and statistically significant.

Sprenger et al. (2014) showed that the trading volume of stocks on the following day is influenced by the volume of posts. They came to that result after analyzing more than 250.000 posts every day based on the relationship between posts and financial related information.

Many researchers also found that message volume is correlated with the trading volume and when that happened people started trading. That happened because people thought that the increase of the message volume means a signal of trading. Thus, the third hypothesis will be:

*H3: Tweet volume influences positively the bitcoin's trading volume*



## 4 Data and Methods

This section provides information about the different data collection and the methodology that we in the study. The first part is about the data that we use for the analysis and explains how and where they were obtained from. Then, the process part of the data is explained before discussing in the final section, the methodology that we use for this study.

### 4.1 Data collection

#### 4.1.1 Twitter data

For the Twitter data, the main dataset that we use is obtained through Kaggle ([www.kaggle.com](http://www.kaggle.com)), contains approximately 3 million tweets from 06/02/2021 till today, and it's updated on weekly basis. Another dataset that we use is from Kaggle ([www.kaggle.com](http://www.kaggle.com)) too because there is some missing information on the first dataset. Our period of interest is from **28/02/2021 to 23/06/2021**. In that period there were many fluctuations, and the main question is whether these occurred because of the Twitter posts and general social media platforms.

The dataset contains information about the name of the user, the number of followers, the data of register, the origin of the person, the time of the post, whether the user is verified or not, and whether the post is a retweet or original one. The data is collected based on the hashtag (#*"#BTC"*, *"#Bitcoin"*). Duplicates and bot tweets are removed.

Another variable that we use is the volume of tweets containing bitcoin information. We obtain this variable from *Bitinfocharts* ([www.bitinfocharts.com](http://www.bitinfocharts.com)) which provides information about many interesting things for cryptocurrencies. For example, we can find how many tweets were posted on a particular day containing the word "bitcoin" or any other cryptocurrency we want to examine.

As mentioned before, because of some missing days another dataset from Kaggle is used to find the sentiment of these days. Also, in line with this dataset, a very useful website is considered for those missing sentiment days (<https://app.intotheblock.com>). This page contains much information about the cryptocurrencies such as social media sentiment, financial information, mining information, and network information.

Table 1: Descriptive summary statistics of the whole period (06/02/2021 till 01/05/2022)

Cryptocurrency	Total tweet volume	Total cleaned tweet volume	Average daily tweets (whole period)	Average Sentiment
<b>Bitcoin (BTC)</b>	3.199.794	1.989.592	27.585	0.1375

Table 1 demonstrates the descriptive statistics of the whole period for the Twitter data. As we see, the total number of tweets for that period (06/02/2021 till 01/05/2022) was 3.199.794.

Of these approximately 3 million tweets, 1.989.592 tweets were cleaned. That happened because there are out of the period of interest that we want to examine. The average number of daily tweets is 27.585, whereas the sentiment of the period is 0.1375.

Table 2: Descriptive summary statistics of the period of interest

Cryptocurrency	Total tweet volume	Total cleaned tweet volume	Average daily tweets	Average Sentiment	Min-Max sentiment
<b>Bitcoin (BTC)</b>	210.202	89.152	1812	0.1893	0.0410 - 0.3780

On the other hand, table 2 shows the same descriptive statistics but for the period of interest that we want to examine. For the period between 28/02/2021 till 23/06/2021 (116 days), the total number of tweets are 210.202 out of 89.152 were cleaned because of NA values in some variables such the location, hashtag, etc., bots, and empty cells after cleaning the dataset. The average sentiment for the period is 0.1893, whereas the range of the sentiment is from 0.0410 to 0.3780.

#### 4.1.2 Cryptocurrency financial data

The cryptocurrency data is obtained from Coinmarketcap ([www.coinmarket.com](http://www.coinmarket.com)). Coinmarketcap is one of the most popular cryptocurrency websites containing information on all the most known cryptocurrencies. Coinmarketcap contains information about more than 1.800 different cryptocurrencies. The website provides CSV files as well as an application programming interface (API) that can be used to obtain financial information about bitcoin. The website has information about the price, the market capitalization, 24-hour volume, circulation, and the 24-hour change.

For our analysis, we obtained the daily price of bitcoin. The bitcoin volume is also obtained from Coinmarketcap. For this analysis, we use a CSV file that contains the dates (28/02/2021 to 23/06/2021), the "Open" price which indicates the price that opens in the market, and the "High" and "Low" which indicate the highest and the lowest prices of bitcoin in 1 day, the "Price" which is the close price of the day, the volume of trading in USD and the change of the price in percent

(%). Table 3 summarizes all this information and provides some descriptive statistics of the dataset.

Table 3: Descriptive summary statistics of bitcoin data

Cryptocurrency	Mean price (USD)	Range of Price (USD)	Mean Volume of Bitcoin (USD in thousands)	Mean Change in price (%)	Range of change in bitcoin price (%)
Bitcoin (BTC)	49.428	31.692 – 63.541	126.88k	-0.1641	-(14.4000)-11.8300

## 4.2 Data Processing Approach

This section discusses the methods that are used to transform the variables from the raw data to meaningful data in order to be used for the analysis. The first subsection addresses how the initial dataset of tweets is reduced after cleaning bot activity as well as tweets irrelevant to the intended period of interest. The second subsection addresses the methods used to clean the text for the analysis and that happens because there are many symbols, words, numbers, etc. that are meaningless for the analysis. The third and final subsection describes how sentiment is used with VADER to extract the sentiment of the posts.

### 4.2.1 Text processing

Twitter data requires special treatment to become a cleaned and ready to use dataset. That happens because it contains many emojis, special symbols, links, URLs, etc. The first and the most difficult part is to exclude the bots. Bots and spam messages most of the time contain noise that does not help at all to make the analysis, that's why they are removed. If someone retrieves the data through API, then bot detection is easy. For this study, most of the bots are removed manually. That happened because many cells were left empty after an extensive cleaning process. Then, most of the remaining posts are checked manually both on the content and the username.

Then, the next steps for the data cleaning are: (1) messages are cleaned from stopwords, and (2) emojis are removed, even though VADER recognizes them and has sentiment points for them, too. Then, (3) special symbols such @#\$\$%^&\*() are removed, (4) mentions and URLs are removed, (5) numbers are removed since they are not meaningful, and (6) whitespaces, both

leading and trailing are removed. Then, the (7) empty cells are replaced by NA and removed. Finally, some bot names such as “Brett Murphy” are removed.

The only thing that we did not remove is the punctuation because it’s important for sentiment analysis. Moreover, VADER works well with punctuation and performs well in sentences, too (Hutto and Gilbert, 2014).

## **4.2.2 Sentiment analysis (VADER)**

Social media content is very different from the typical content that's why it's very challenging for the traditional sentiment analysis techniques. This content is rich in "slang" and abbreviated language which a typical method cannot deal with. Moreover, the meaning of a word in a sentence sometimes is different from the original meaning of the word. VADER deals well on these types of occasions. That's why VADER is widely chosen for social sentiment analysis (Hutto and Gilbert, 2014). Right after the cleaning of the dataset, the VADER algorithm is used to extract the sentiment. This algorithm computes a normalized weighted compound score between -1 and 1 for every cell of the dataset.

## **4.3 Econometric methods**

### **4.3.1 Akaike Information Criteria (AIC)**

A very important econometric element in a time series analysis is the estimation of the lag length of the autoregressive process. That’s important because we want to test whether information obtained from tweeter affects financial data. For this analysis, we use Akaike Information Criterion (AIC) for determining the lag length (Akaike, H. 1973). We use the Akaike Information Criterion to test the quality of the statistical models. Specifically, this method is mainly used to calculate the predictive power of a model.

When it comes to the optimal number of lags, we chose the one with the lowest AIC score. Akaike is also used to measure the performance of a model when information is taken out of the model. The estimator indicates an approximation of the information that is lost when a model is used, which afterward is used to choose the optimal lag. The result of the AIC is then used in the next step of the time series process which includes the stationarity of the variables, the autoregressive model, and the causality test.

### 4.3.2 Augmented Dickey-Fuller test (ADF)

Dickey and Fuller (1979) developed a unit-root test to determine whether the endogenous variables are stable, which means if they fluctuate around a fixed mean, or if they are moving around without a fixed mean and can deviate permanently from previous levels. To consider a time series stationary, statistical metrics such as variance, and mean should remain stable over time. That's a common thing to test because many units in a time series do not remain constant. For that reason, we use an Augmented Dickey-Fuller test (ADF) to check all the variables, because non-stationary variables can cause issues to our model and analysis (Pauwels et. al,2002). Also, stationarity is a basic requirement for the next steps of the analysis which include the autoregressive model and the causality test.

Many times, not all the variables are stationary. In that case, we can transform the time series into a stationary time series by differencing the time series variable

$$X(t) = X(t) - X(t - 1) \quad (1)$$

then, we can continue our analysis with the transformed stationary variables. The null hypothesis for this test supports non-stationarity on a 5% level.

### 4.3.3 Vector Autoregression (VAR)

Sims (1980) developed the vector autoregression model for three main purposes: (1) forecasting economic time series; (2) designing and evaluating economic models; (3) evaluating the consequences of alternative policy actions. This method tests the relationship between the variables in the vector. Models such as VAR are different from the simple univariate autoregressive model (AR) since it allows for the inclusion of multiple explanatory variables in the model. The base of the theory behind VAR is that each variable is a function of past lags of itself and other variables (Dekimpe and Hanssens,1999). An issue that arises with the VAR model with different variables used in the analysis is whether VAR can be used in the case of variables being cointegrated. It is believed that in short term, VAR models perform better than a vector error correction model (VECM) or a cointegrated VAR model. Moreover, it is proved by Naka and Tufte (1997) that the results of impulse response functions for VAR models did better in the short term compared to VECM models. Thus, for short-term effects on cryptocurrency financials such

as returns, price, and trading volume. Thus, the VAR model is preferred. A VAR model is typically given by the equation:

$$y_t = y_{t-1} + \dots + y_{t-n} + \text{error}_t \quad (2)$$

where  $y_t$  is a vector of variables,  $\text{error}_t$  assigns an error term to the forecast, and  $n$  stands for the number of lags in the model.

As discussed before, a basic requirement is to first make all the variables stationary to proceed with the VAR. The problem with non-stationary data in a VAR model is that it results in false test statistics, which is misleading when estimating a model. For this thesis, we want to measure the effect on a dependent variable following a change in the independent variable that's why the effect of the Twitter sentiment is going to be studied on the return, trading volume, price, and tweet volume.

#### 4.3.4 Granger Causality test

Another step of the analysis is to test the Granger causality test. The granger causality test is implemented to assess whether the independent variables provide a lagged influence on the dependent variable (Granger, 1969). Generally, this method is used for time series can be used to forecast another time series. The main assumption of granger causality is that if variable X causes Y, then changes in X will happen before the changes in Y (Granger and Newbold, 1986). If variable X is useful for predicting Y, then we say that variable X granger causes variable Y. Moreover, the output of the vector autoregression model is the input of the Granger causality test. The equation of the granger causality test is given by:

$$y_t = a_0 + \varphi y_{t-1} + \varphi_n y_{t-2} + \dots + a_m y_{t-m} + b_1 x_{t-1} + \dots + b_q x_{t-q} + \text{error}_t \quad (3)$$

This model implies that the last period's value of  $x$  has explanatory power for the current value of  $y$ . The coefficient  $b_1$  is a measure of the influence of  $x_{t-1}$  on  $y_t$ . If  $b_1 = 0$ , then past values of  $x$  do not affect  $y$ .

#### 4.3.5 Impulse response test

The impulse response function is mainly used on the results of the estimated VAR models to interpret the coefficients. That happens because variables in a VAR models depend on each other

and it's difficult to examine the individual influence of a single variable. The result of the impulse response test can determine the effect of an increase in one variable on the target variable over a certain period. The test also examines the bootstrap confidence bands for a 95% confidence interval so that when the upper and lower band carry the same sign, the response is interpreted as statistically significant at the 95% confidence level.

Sims (1980) introduced a method that can determine, the orthogonalized impulse response coefficients. That method is very helpful because it allows finding contemporaneous relations in the target and independent variables. The result of taking these contemporaneous relations into account is visualized in the IRF plot, where a plot that deviates from the starting point of zero shows a contemporaneous relation to the target variable. The formula of the impulse response function is given by:

$$y_t = Ay_{t-1} + e_t \quad (4)$$

That's the formula for one lag. To find, assume, the effect of the  $n$ -th element of the vector of shocks upon the  $n$ -th element of the state vector 2 periods later, which is a particular impulse response, write down the above equation of evolution one period lagged:

$$y_{t-1} = Ay_{t-2} + e_{t-1}. \quad (5)$$

Using this in the original equation of evolution, we have:

$$y_t = A^2y_{t-2} + Ae_{t-1} + e_t \quad (6)$$

then using the twice lagged equation to obtain:

$$y_t = A^3y_{t-3} + A^2e_{t-2} + Ae_{t-1} + e_t \quad (7)$$

From this, the effect of the  $n$ -th component of  $e_{t-2}$  upon the  $n$ -th component of  $y_t$  is the  $i$ ,  $n$  element of the matrix  $A^2$ .

## 5 Results

This chapter describes the results obtained during the analysis. First, a simple analysis is made along with some descriptive statistics of the variables. Then, pairwise correlations are conducted

which describe the associations between the variables. Afterward, lag selection, Vector Autoregression (VAR) model, a Granger causality test are used to model the VAR results and assess which variables provide a lagged influence on other variables.

## 5.1 Simple analysis

For the simple analysis, we use linear regression to take a look into the relationship between some variables. Linear regression is a supervised learning technique that is used to create models between a response variable and one or several predictor variables. Linear regression is a simple yet very informative and preferred method for analysis. However, in this analysis VAR models are more suitable due to the fact they can capture the dynamics of time series data and they also they allow feedback to occur between the variables in the model.

First, linear regression with "sentiment" as a dependent variable has been run. Then, another linear regression but this time with "price" as a dependent variable has been run, while a third model with tweet volume has been also run. Table 4 summarizes all the information from the regressions.

Table 4: Linear Regressions of the models.

	Sentiment	Price	Tweet Volume
<b>(Intercept)</b>	-23.04 ***	13.33 ***	13.54 ***
	(1.68)	(0.49)	(1.62)
<b>Price</b>	1.35 ***		-0.49 ***
	(0.14)		(0.14)
<b>Vol.</b>	-0.00	-0.09 ***	0.03
	(0.06)	(0.03)	(0.04)
<b>tweet volume</b>	0.90 ***	-0.21 ***	
	(0.09)	(0.06)	
<b>returns</b>	0.00	0.02	-0.01
	(0.06)	(0.03)	(0.04)
<b>sentiment</b>		0.33 ***	0.50 ***
		(0.03)	(0.05)
<b>R<sup>2</sup></b>	0.70	0.58	0.48
<b>Adj. R<sup>2</sup></b>	0.69	0.57	0.46

Signif. Codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 '' 1

As we see from table 4, the first model where the sentiment is the dependent variable has  $R^2$  and  $Adj. R^2$  is equal to 0.69. That means that 0.69 of the total variance is explained through this model. Moreover, variables such as "open", and "low" were kept out of the model. Then, we can infer



that “*price*” and “*tweet volume*” are positively correlated and very significant on a 5% level ( $p=9.85e-16$  and  $p=3.91e-16$ , respectively).

On the other hand, model 2 shows the regression model when “Price” is the dependent variable. For this model, the  $R^2$  and  $Adj. R^2$  is equal to 0.57. In this model, sentiment is positively correlated and highly significant on a 5% level ( $p=9.85e-16$ ), whereas bitcoin volume and tweet volume are both negatively correlated with the price and significant on a 5% level ( $p=0.000950$  and  $p=0.000511$ , respectively).

Finally, the third model has the tweet volume as the dependent variable. As we can see from table 4, the price has a negative and significant effect on tweet volume at a 5% level ( $p=0.000511$ ), whereas sentiment has a positive effect on tweet volume at a 5% level ( $p=3.91e-16$ ).

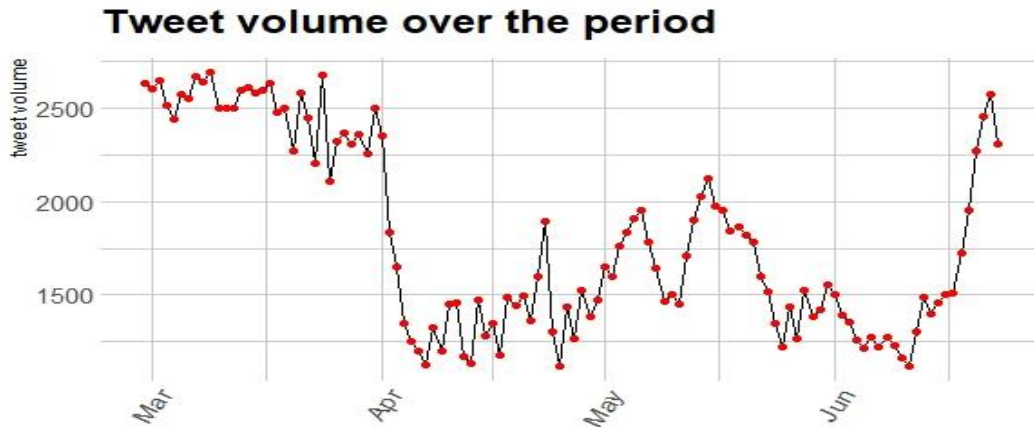
All in all, linear regression provides informative results for the relationships between the variables, but at the same time there are many problems when it comes to be used for a times-series analysis. One of them is the autocorrelation, which violates independence assumptions. Thus, for this analysis, linear regression is not a suitable method, instead, VAR models are preferred. VAR models can capture the relationship between the variables as they change over time, something that is not possible to examine with a simple linear regression.

## **5.2 Descriptive statistics**

### **5.2.1 Tweet volume**

Tweets are collected between the period of the 28th of February 2021 and the 23rd of June 2021, a total of 116 days. Figure 1 shows the daily tweet volume of bitcoin over the entire period. Just before April, we can see that the number of tweets fell dramatically, that happens because the returns of the next day’s decrease too. After that, the number of tweets are fluctuated until the end of June which are increased again to the levels of March. Figure 8 shows the average number of tweets per day, whereas figure 9 demonstrates the sentiment per day.

Figure 1: Tweet volume



### 5.2.2 Trading volume

The trading volume of the coin is measured in thousands of dollars. Trading volume is the number of transactions made in a day in USD. Figure 2 shows the volume of bitcoin transactions over the period, whereas figure 3 shows a time series plot with bitcoin trading volume and tweet volume during the period of interest. From figure 2, it is worth mentioning the decrease in trading volume in the first days of April which happens due to the decrease in tweet volume in the same period. Moreover, the peak of the trading volume happens sometime in the middle of May because that was the day with the lowest returns for bitcoin (-14%).

Figure 2: Trading volume of bitcoin

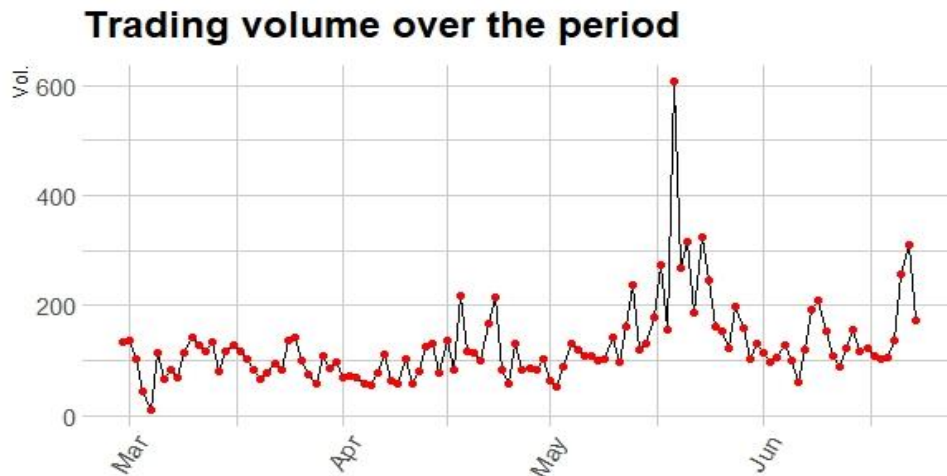


Figure 3: Time series plot of tweet and trading volume

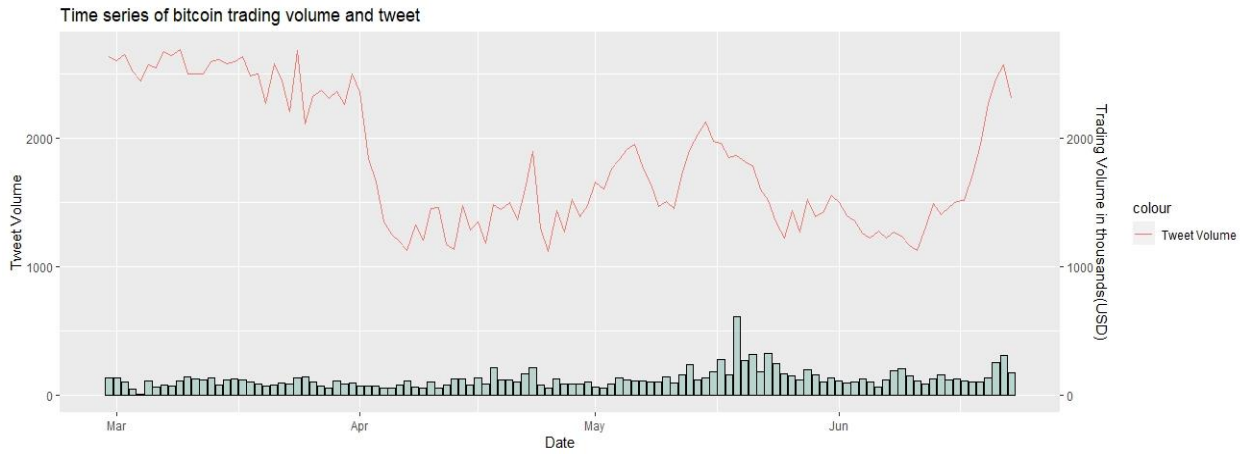
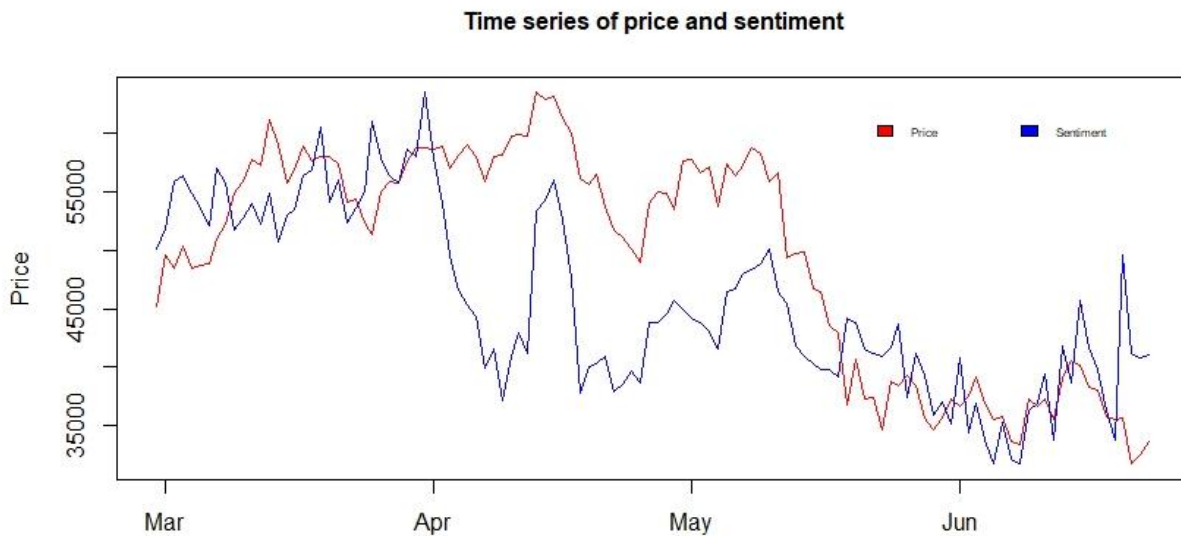


Figure 3 demonstrates the time series of tweet volume along with the trading volume. As we can see from the figure, the pattern of trading volume quite follows the one of tweet volume. For instance, beginning of April, as the tweet volume decreases, the same happens with the trading volume. The same happens with the pattern at the end of June where the trading volume increases as the number of tweets increases.

Figure 4: Time series plot of price and sentiment



All the tweets are collected and given a sentiment score between -1 and 1. Figure 4 demonstrates a time-series figure between the price of bitcoin and the sentiment of the period. At a first look, it seems that the price of bitcoin follows the sentiment, for example, in middle of the April an increase in price drives the sentiment up, too. The same pattern can be seen at the beginning of May. But, in the next sections, we will discuss which variable is more important in predicting the other one, if applicable.

After a first look at the figures, it seems that some variables follow the same pattern as some other variables. For instance, trading volume follows the same pattern as the tweet volume, whereas the price of bitcoin follows the same pattern as the pattern of sentiment. What is left now, is to test which variable follows the pattern of the other one and whether this relationship is one way or also the other way around.

What follows is a first look at the variable relationships with a pairwise correlation which indicates not only the magnitude of the relationship but also how strong it is.

### **5.3 Pairwise correlations**

Pairwise correlations provide a first look into the relationship between the variables. Table 5 demonstrates the pairwise correlations between the variables of bitcoin. The correlations provide a first insight into the strength and direction of the variable relationships.

Table 5 shows that price is significantly correlated with all the price-related variables because of collinearity. But the price is also positively correlated with tweeter sentiment which also is significant at a 5% level ( $r=0.63$  and  $p<0.001$ ). Moreover, tweet sentiment is not correlated with bitcoin's returns. Their relationship is quite neutral ( $r=0.16$ ) and non-significant ( $0.579$ ). Finally, tweet volume is highly correlated to trading volume. We can see that because the relationship is positive ( $r=0.69$ ) and very significant on a 5% level ( $p<0.001$ ).

After a first look at the pairwise correlations table, we can observe some interesting relationships between the variables such as the pairs of price and sentiment or the pair of tweet volume and trading volume. In the next sections, these relationships will be discussed deeper to understand better the meaning of their relationships.

Table 5: Pairwise correlation between the variables. (Could be also in appendix)

Parameter1	Parameter2	r	CI	CI_low	CI_high	t	df_error	p
Price	Open	0,974215	0,95	0,962929	0,9820959	46,10249	114	<b>3,4E-74***</b>
Price	High	0,988214	0,95	0,983003	0,9918341	68,92703	114	<b>2,36E-93***</b>
Price	Low	0,989225	0,95	0,984458	0,9925357	72,14382	114	<b>1,52E-95***</b>
Price	Vol.	-0,45866	0,95	-0,59151	-0,301555	-5,51094	114	<b>3,59E-06***</b>
Price	Change %	0,161779	0,95	-0,02116	0,3342372	1,750383	114	0,579198
Price	sentiment	0,632738	0,95	0,509157	0,7307559	8,724298	114	<b>5,34E-13***</b>
Price	tweet volume	0,202922	0,95	0,021397	0,3714944	2,212644	114	0,289148
Open	High	0,991093	0,95	0,987147	0,9938316	79,46166	114	<b>3,2E-100***</b>
Open	Low	0,98187	0,95	0,973891	0,9874265	55,30625	114	<b>8,36E-83***</b>
Open	Vol.	-0,40725	0,95	-0,54882	-0,242975	-4,76094	114	<b>8,56E-05***</b>
Open	Change %	-0,06056	0,95	-0,24023	0,1231131	-0,64782	114	1
Open	sentiment	0,60016	0,95	0,469181	0,705303	8,011146	114	<b>1,94E-11***</b>
Open	tweet volume	0,191308	0,95	0,009317	0,3610326	2,081051	114	0,35702
High	Low	0,983506	0,95	0,976238	0,9885635	58,05573	114	<b>4,16E-85***</b>
High	Vol.	-0,39739	0,95	-0,54054	-0,23187	-4,62369	114	0,00014
High	Change %	0,041935	0,95	-0,14146	0,2225505	0,448142	114	1
High	sentiment	0,621466	0,95	0,495265	0,7219799	8,469615	114	<b>1,95E-12***</b>
High	tweet volume	0,207946	0,95	0,02664	0,3760076	2,269877	114	0,276043
Low	Vol.	-0,5253	0,95	-0,64577	-0,379309	-6,5913	114	<b>2,4E-08***</b>
Low	Change %	0,082068	0,95	-0,10177	0,2604871	0,879212	114	1
Low	sentiment	0,60947	0,95	0,480551	0,7126045	8,207978	114	<b>7,34E-12***</b>
Low	tweet volume	0,1791	0,95	-0,00333	0,3499882	1,943689	114	0,435181
Vol.	Change %	-0,24725	0,95	-0,41104	-0,067995	-2,72446	114	0,08947
Vol.	sentiment	-0,26077	0,95	-0,42297	-0,082366	-2,88401	114	0,061042
Vol.	tweet volume	-0,00453	0,95	-0,18669	0,1779335	-0,04837	114	1
Change %	sentiment	0,160366	0,95	-0,02261	0,3329481	1,734691	114	0,579198
Change %	tweet volume	0,044528	0,95	-0,13892	0,2250179	0,475901	114	1
sentiment	tweet volume	0,685847	0,95	0,57549	0,7716723	10,06235	114	<b>4,37E-16***</b>

Signif. Codes: 0 '\*\*\*\*' 0.001 '\*\*\*' 0.01 '\*\*' 0.05 '\*' 0.1 '\*'

## 5.4 Lag selection

In the previous sections, linear regression and pairwise correlation were discussed. The former shows some kind of linear relationship between some variables, whereas the latter only shows the strength of the relationship between paired variables. To have a better understanding of the model, some other time-series properties are needed to understand better the variables in the model.

The first step is to determine the lags in the model. Lag length selection refers to the number of previous observations in a time series that will be used as predictive variables in the model. One of the most basic and informative criteria to use for comparing the quality of the model is the Akaike Information Criterion (AIC). The model estimates the lag with four different measures, Akaike's information (AIC) criterion, the Hannan-Quinn (HQ) criterion, the Schwarz (SC) criterion, and the Final Prediction Error. The optimal lag length was chosen from the lowest AIC score because AIC is better when handling small datasets.

Table 7 demonstrates the optimal lag length selection per criterion, whereas table 11 shows all the scores of the metrics. The optimal lag length is 2 (max lag was set to 10) and will be used for the next steps of the analysis which are the Augmented Dickey Fuller test, Granger Causality test, and Vector Autoregression model.

Table 7: Lag length selection

Akaike Information Criterion (AIC)	2
Hannan-Quinn (HQ)	1
Schwarz (SC)	1
Final Prediction Error (FPE)	2

## 5.5 Augmented Dickey-Fuller test

After the lag selection, a very important step is to test whether the variables are stationary. That happens because non-stationary variables can cause problems statistical problems in the time series analysis. A time series is stationary when metrics such as mean, variance, and autocorrelation remain constant over time. For this analysis, the variables price, tweet volume, trading volume, and sentiment are log-transformed to control for scaling and allow to compute elasticities.

To begin with, this test uses 2 lags as selected as optimal from AIC. The null hypothesis for this test is that the variable is non-stationary. That occurs when the p-value is greater than 5%. If a variable has a p-value greater than 5% needs to be transformed to proceed to the next steps of the analysis. As mentioned in the previous section, the transformation happens by differencing the

time series variable. Then, if the differencing variable proves to be stationary, it can be used for further analysis.

As we can infer from table 8, three variables (price, sentiment, and tweet volume) need to be transformed into their first difference because their p-value is greater than 5%. On the other hand, variables such as a change in price and trading volume prove to be stationary over time. Table 8 shows the results for the ADF test on a 5% level.

Table 8: ADF test

<b>Time series variable</b>	<b>P-value</b>
<b>Price</b>	0.4374
<b>Trading Volume</b>	<b>&lt;0.01***</b>
<b>Change of price in % (Returns)</b>	<b>&lt;0.01***</b>
<b>Sentiment</b>	0.0749
<b>Tweet Volume</b>	0.7879

The three non-stationary variables are transformed to their difference and a second ADF test is conducted for which all the p-values are smaller than 5%. Thus, the variables can be used for further analysis.

## **5.6 Vector Autoregression model (VAR)**

The time series regression model for this analysis is the vector autoregression model (VAR). After, selecting the optimal number of lags and transforming the non-stationary variables to stationary, the requirements are fulfilled to use VAR. This model is used to examine the relationships between the time series variables. For a better understanding of the results, VAR is modeled using the Granger causality test and the impulse response function. Every variable is displayed in a formula based on its lagged values and the lagged values of the other variables. Also, VAR is an excellent tool to use when dealing with time series data (Dekimpe and Hanssens,1999).

Table 9: Vector Autoregression model (VAR)

	Price	Sentiment	Tweet Volume	Change (Returns %)	Trading Volume
Price.l1	0.111 (0.112)	<b>1.219 *</b> (0.636)	0.184 (0.261)	<b>99.147 ***</b> (0.394)	<b>-2.657***</b> (0.897)
Sentiment.l1	0.019 (0.019)	<b>0.518***</b> (0.106)	0.017 (0.043)	<b>0.196 ***</b> (0.066)	<b>0.275 *</b> (0.149)
Tweet. Volume. l1	0.032 (0.042)	0.215 (0.241)	<b>-0.194*</b> (0.099)	<b>0.295 *</b> (0.150)	<b>1.059 ***</b> (0.341)
Change.l1	0.002 (0.028)	0.014 (0.159)	0.054 (0.065)	0.056 (0.099)	0.288 (0.225)
Trading. Volume. l1	0.010 (0.013)	0.009 (0.074)	0.017 (0.031)	0.022 (0.046)	<b>0.383 ***</b> (0.105)
Price.l2	0.077 (2.790)	1.315 (15.898)	5.464 (6.525)	4.332 (9.847)	31.910 (22.437)
Sentiment.l2	0.004 (0.019)	0.063 (0.106)	0.021 (0.044)	<b>0.210 ***</b> (0.066)	<b>0.255 *</b> (0.150)
Tweet. Volume. l2	0.019 (0.045)	0.195 (0.258)	0.004 (0.106)	0.180 (0.160)	0.266 (0.364)
Change.l2	0.001 (0.001)	0.005 (0.006)	0.0002 (0.002)	<b>-0.009**</b> (0.004)	0.012 (0.008)
Trading. Volume. l2	0.004 (0.012)	0.007 (0.066)	0.015 (0.027)	0.060 (0.041)	0.121 (0.094)
const	0.063 (0.055)	0.068 (0.314)	0.143 (0.129)	0.281 (0.195)	<b>2.295 ***</b> (0.443)
R <sup>2</sup>	0.077	0.219	0.056	0.999	0.470
Adjusted R <sup>2</sup>	0.012	0.144	0.036	0.999	0.419
Residual Std. Error (df = 103)	0.047	0.269	0.110	0.166	0.379
F Statistic (df = 10; 103)	0.863	<b>2.895 ***</b>	0.612	<b>8,535.628 ***</b>	<b>9.149 ***</b>
Note:	*p<0.1; **p<0.05; ***p<0.01				

Table 9 shows the vector autoregression model for bitcoin financials. The model demonstrates the effect of tweet sentiment, tweet volume, price, returns, and trading volume on their own lagged variables. The models that we focus on are the ones with price, returns, and trading volume as



dependent variables. We focus on these models because they have the answers to the hypothesis that we test.

As we can see from table 9, there is no significant effect on the lagged variables and the price which means no variables affects today's price. On the other hand, we can infer from table 9 that the sentiment of the previous day (lagged variable of sentiment) has a significant and positive effect on today's returns ( $r=0.196296$  and  $p=0.00343$ ). Moreover, not only does one day's sentiment has a positive effect on today's returns but also two days' sentiment has also a positive and significant effect on returns ( $r=0.210105$  and  $p=0.00185$ ). In addition, table 9 indicates that the returns of the previous two days have a negative and significant effect on today's returns ( $r=-0.009180$  and  $p=0.01327$ ).

Trading volume seems to be positively affected by the previous day's tweet volume ( $r=1.05913$  and  $p=0.002427$ ). Moreover, it seems that trading volume is also affected by its own first lag ( $r=0.38319$  and  $p=0.000413$ ). From these results, we can conclude that hypothesis 1 is supported because there is a strong and positive relationship between sentiment and bitcoin returns, whereas for hypothesis 2 there is not enough evidence to support it. Finally, hypothesis 3, is also supported because there is a positive and significant effect between tweet volume and trading volume.

In addition to these results, we also find that trading volume is negatively affected by the price of the previous day which is also significant at a 5% level ( $r=-2.65684$ ,  $p=0.003811$ ). Moreover, this can be supported by the Granger causality test which states that price granger causes trading volume at a 5% level ( $p=0.003066$ ) but not the other way around. In addition, IRF figure 8 shows that the first day of the response on a price impulse is positive but insignificant, whereas the second day is still positive but significant. In the period between days 3 and 5 the response is negative and significant. Furthermore, trading volume seems to be affected by its own lagged variable in a positive and significant way ( $r= 0.38319$ ,  $p= 0.000413$ ).

## **5.7 Granger causality test**

This section explores the Granger causality tests between the time series variables. By conducting these tests, we examine whether a time series variable has a lagged effect on another time series variable. The hypothesis behind Granger causality tests is that if a time series A affects time series

B, then changes in A will happen before changes in B(Granger and Paul. 1986). The optimal lag length is already chosen by Akaike Information Criteria (n=2). The null hypothesis of the tests is that the time series A does not cause time series B to granger cause itself. Knowing the value of a time series A at a given lag is valuable for forecasting the value of a time series B at a later period is referred to as “*Granger causes.*”

Table 10 demonstrates the relationships between the time series variables. First of all, we can infer that tweet volume granger causes trading volume but not the other way around on a 5% level of significance ( $p=0.002588$  and  $p=0.6708$ , respectively). That means that we can reject the null hypothesis of the test and infer that knowing the tweet volume is valuable for predicting the trading volume.

On the other hand, table 10 indicates that there is no granger cause effect either way for sentiment and price. Neither time series variable is useful to predict the future values of the other.

Finally, we can assume that the sentiment granger causes the price change (returns) and reject the null hypothesis because the p-value of the test is smaller than 5% ( $p=0.0003029$ ). On the hand, the other way around is not supported because the p-value is greater than 5% ( $p=0.791$ ). All in all, we can say that knowing the sentiment of the people we can predict the future values of the bitcoin’s returns.

Table 10: Granger causality tests

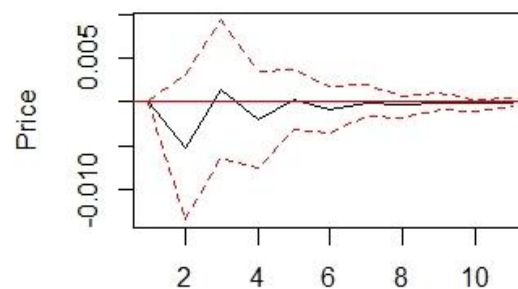
Dependent Variable	Independent Variable	F	Prob (>F)
Trading Volume	Tweet Volume	6.2944	<b>0.002588***</b>
Tweet Volume	Trading Volume	0.4008	0.6708
Price	Sentiment	0.792	0.4555
Sentiment	Price	1.8367	0.1642
Change	Sentiment	8.7352	<b>0.0003029 ***</b>
Sentiment	Change	0.235	0.791

## 5.8 Impulse Response Function

The impulse response function (IRF), of a dynamic system, is its response when tested with a brief signal, called impulse. That means that an impulse response is the reaction of a system to some external change. In our case, we use the impulse response function to examine what is the response of the variables when there is a change in an exogenous variable. The next plots show the reactions of variables such as the price in an external change of sentiment (figure 5), the response of returns in a sentiment shock (figure 6) and the response of trading volume to a shock from tweet volume (figure 7).

Figure 5: Sentiment impulse on Price

Orthogonal Impulse Response from Sentiment

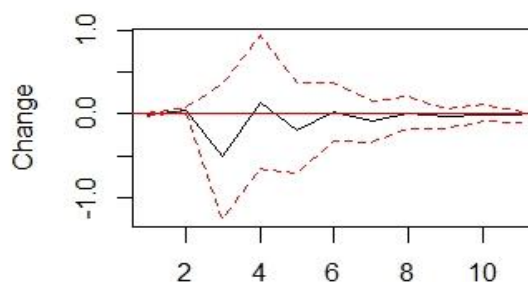


95 % Bootstrap CI, 500 runs

Figure 5 shows the response of price to a sentiment shock. The response is negative and significant the next day, whereas on the third day the response is positive and significant on a 5% level. After that, the response of prices to a sentiment shock tapers off in the next few days.

Figure 6: Sentiment impulse on returns

Orthogonal Impulse Response from Sentiment



95 % Bootstrap CI, 500 runs

Figure 6 demonstrates the responses of returns in a possible sentiment shock. The response is quite positive on the first and second day, but only on the first day, the response is significant at a 5% level. Again, the response of returns to a sentiment shock tapers off the upcoming days and it seems to be insignificant until day eleven.

Figure 7: Tweet volume impulse on trading volume

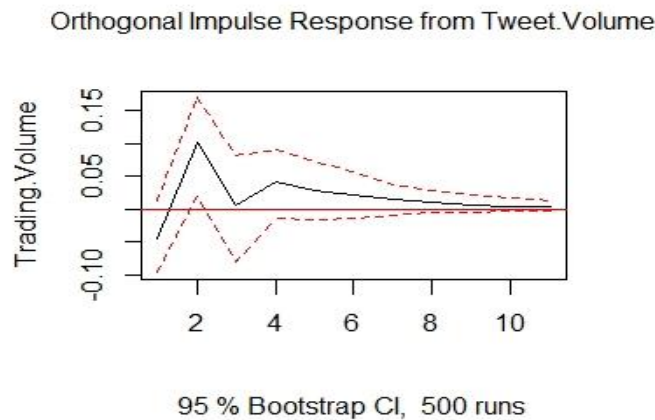


Figure 7 shows the response of trading volume on a tweet volume impulse. The shock of trading volume is positive and significant the next day. In the upcoming days, the shock remains positive but insignificant a 5%. However, after the seventh day, the shock of trading volume from tweet volume remains positive but it becomes significant.

## 5.9 Summary of main findings

This section provides the results of the main findings in combination with the evidence supporting or rejecting each hypothesis. The results have shown that tweeter information is an important predictor of cryptocurrencies financials. The results provide enough information to accept both hypotheses 1 and 3, whereas they also indicate that hypothesis 2 is not supported. The results and the discussion of each hypothesis are provided below.

First, evidence for a relation between tweet sentiment and bitcoin returns has been found. The VAR models indicate a strong association between sentiment and bitcoin returns which is also

positive. Also, this relation is positive and significant for a duration of 2 lagged days. Moreover, the Granger causality test indicates that sentiment is important for predicting future returns but not the other way around. Finally, the impulse response function indicates a positive effect of sentiment on the returns in the first two days. Thus, hypothesis 1 which states that tweet sentiment is positively related to bitcoin's returns can be therefore supported.

Second, this study did not find enough evidence to support hypothesis 2 which states that tweet sentiment is positively related to bitcoin's price. The VAR models did not indicate any relationship between the tweet sentiment and the price of bitcoin. In addition, that was also supported by the Granger causality test which indicates no effect of sentiment on price and the other way around. On the other hand, IRF shows a negative and significant effect the next day but it's not enough to support the second hypothesis. Thus, for hypothesis 2 there is not enough evidence.

Third, this analysis has found evidence that the tweet volume is positively associated with the trading volume of bitcoin. As obtained from VAR models, the effect of tweet volume on trading volume is positive and significant with a lag of 1 day. Also, the Granger causality test indicates that tweet volume granger caused trading volume but not the other way around. Finally, the IRF demonstrates that the effect of tweet volume is positive and significant the next day, whereas in the upcoming days, it remains positive but insignificant. However, after some days the effect becomes positive and significant again. Thus, we can infer that hypothesis 3 which states that tweet volume is positively related to trading volume can be supported.

## **6 Further Analysis: Predictions**

This section provides an analysis that aims to test the prediction and performance power for the bitcoin's returns with two different methods: time-series and machine learning. For this analysis, we use ARIMA for the time series prediction, whereas SVM is preferred for the machine learning prediction. One of the most important analysis that someone can make are the predictions of the future values. Then, the prediction of the future values of the variables can be used for various matters. For example, investors could invest with more confidence if they know that models have a great performance power with high accuracy. Also, for academic purposes, knowing the

predictive power of certain models could enhance the already existing knowledge of the cryptocurrency market and help in order to improve it even further.

## 6.1 ARIMA

Autoregressive (AR) integrated(I) moving average (MA) is a powerful statistical times series analysis model which is mainly used to either understand better the time-series data or predict future trends. That model is autoregressive because it tries to predict future trends based on the past values of the variables. Furthermore, an autoregressive integrated moving average model is a form of regression analysis that uses the power of one dependent variable relative to some other variables that change during the time (Box et.al, 2015).

ARIMA models can be used as another approach to time series forecasting. Exponential smoothing and ARIMA models are the two most widely used approaches to time series forecasting. The exponential smoothing models are describing the trend and seasonality in the data, and ARIMA models try to describe the autocorrelations in the data.

ARIMA is preferred for time-series predictions because it identifies the appropriate number of lags and the times that the variables have been differenced, while the forecasting is done by plugging in time series data for the variable of interest. The equation for ARIMA in our model is given by:

$$y_t = \mu + \phi_1 y_{t-1} + error_t \quad (8)$$

where  $Y$  regressed on itself lagged by one period,  $Y_{t-1}$  is the lagged value of itself and  $\mu$  is a constant. The model that is used is ARIMA (1,0,0), so AR (1), where 1 is lag order, 0 is the degree of differencing and 0 is the order of the moving average, which means that we have a first-order autoregressive model. That model was selected after examining the best performance models between AIC, BIC and AICc criteria (see table 12). Akaike's Information Criterion (AIC), which was useful in selecting predictors for regression, is also useful for determining the order of an ARIMA model.

### 6.1.1 ARIMA results

After conducting the ARIMA test we can infer from table 12 that based on the negative and significant at a 5% level coefficient, we will have a mean-reverting behavior in the future such as an alternation of the sign, and also that  $Y$  will be below the mean in the next period.

Table 12: ARIMA model results

Estimate	Std. Error	z value	Pr(> z )
-0.196660	0.090934	-2.1627	<b>0.03057*</b>

Figure 9 demonstrates the predicted trend of bitcoin's returns in the future. As predicted before by the ARIMA model, the mean of the future value is negative (-7.119097).

The most important thing when doing a prediction test is to examine the performance of the model. The metrics that are used for this analysis are RMSE which measures the average error performed by the model in predicting the outcome of an observation. and MAE which measures like RMSE the error but it's less sensitive to outliers compared to RMSE. The RMSE and MAE can be used together to measure the variation in the error in a set of forecasts. The RMSE for this model is 4.52698, whereas the MAE is 3.38375 as can be seen in table 15.

Table 13: Performance metrics of ARIMA forecast

RMSE	4.52698
MAE	3.38376

### 6.1.2 VAR results

VAR models are also cable of predicting future values. After conducting the VAR prediction model, we can infer that based on the results, which can see at table 14, VAR models have the worst predictive performance.

Table 14: Performance metrics of VAR forecast

RMSE	8.91
MAE	7.40

From these results, we can conclude that VAR's predictive power is not enough to predict with confidence the future values of the bitcoin's returns.

### 6.2 Support Vector Machine (SVM)

Support Vector Machine or SVM could be a supervised machine learning method that's used for both classification and regression problems. The goal of the tactic is to search out a hyperplane in an N-dimensional space that classifies the information points. If the amount of input features is 2, then the hyperplane is simply a line. If the quantity of input features is three, then the hyperplane becomes a 2-D plane. SVM is preferred instead of other machine learning models because is more productive in high dimensional spaces, SVM uses kernel tricks to unravel non-linear problems and SVM handles outliers better.

The SVM map the data of input space into a high-dimensional feature space, to solve many difficult problems that cannot be solved by the linear method in the original sample space. Compared with other machine learning methods, SVM has a simpler training process and better generalization capability. SVM also can be used in field of time series prediction and its performance is quite good. The formula for SVM is given by:

$$f(\mathbf{x}) = (\omega \cdot \Phi(\mathbf{x})) + \underline{b} \text{ with } \Phi: \mathbb{R}^k \rightarrow \mathcal{F}, \omega \in \mathcal{F}_1 \quad (9)$$

In SVR the basic idea is to map the data  $x$  into a high dimensional feature space  $F$  via a nonlinear mapping  $\Phi$  and do linear regression in this space, where  $b$  is a threshold. But  $\omega \cdot \Phi(\mathbf{x})$  would have to be computed in this high dimensional space, which is usually intractable, we have to use a trick. Thus, formula 9 is then written:



$$f(\mathbf{x}) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})) + b = \sum_{i=1}^l (\alpha_i - \alpha_i^*) k(\mathbf{x}_i, \mathbf{x}) + b \quad (10)$$

In which we introduced a kernel function  $\sum_{i=1}^l (\alpha_i - \alpha_i^*) (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}))$ . For example, a common kernel could be RBF:  $k(x, y) = \exp(-\| \mathbf{x} - \mathbf{y} \|^2 / (2\sigma^2))$ .

Kim (2003) used SVM to forecast financial statistics and claimed that this method is promising for statistical analysis forecasting because they use a risk function consisting of the empirical error and a regularized term which comes from the structural risk minimization principle. This thesis applies SVM to measure the performance power of the model when it comes to predicting the future values of bitcoin's returns.

### 6.2.1 Support Vector Machine results

After successfully splitting the data into training and testing sets, the test is ready to be conducted. In addition, repeated cross-validation is used to estimate the performance of a machine learning algorithm or configuration on a dataset. Repeated k-fold cross-validation provides a way to improve the estimated performance of a machine learning model. This involves simply repeating the cross-validation procedure multiple times and reporting the mean result across all folds from all runs.

The results of the SVM metrics for predicting the future values of bitcoin's returns after tuning the hyperparameters **c** and **sigma** of the model can be seen in table 15. As we can compare it to the relative numbers of ARIMA's performance, we can infer that both models do a very good job when it comes to predicting future values. On the one hand, ARIMA has a better MAE, whereas for SVM, the RMSE is better than ARIMA's.

Table 15: Performance metrics of SVM forecast

RMSE	4.633785
MAE	3.173609

All in all, when it comes to comparing the two models, there are no major differences between them. When the performance metrics are concerned both models are doing well, but as it seems from the results, we can say that ARIMA has slightly better performance than SVM and

VAR. In addition, the interpretation of ARIMA is more informative because of the statistical information that the results display while at the same time it provides nice visualizations of the prediction. Both models have advantages and disadvantages, the selection of the model should be based on the needs of the analysis.

### 6.2.2 Future returns

After successfully running different predictive models for the returns of bitcoin, we find that the ARIMA model performs slightly better than the others. That would be an informative and useful insight for investors, managers and all the stakeholders. Knowing that a time series model performs great when it comes to predicting future values, all these people could use it in order to make decisions for future investments, short or long-term depending on the needs of the people, or any other personal purposes.

For this analysis, we use the results from ARIMA to make a table with the future returns of bitcoin, for the period of the next 10 days. That could help the stakeholders to know the predictive results of the returns and use them in any way they think is useful. Table 16 shows the results of the returns.

Table 16: Bitcoin returns

<b>Day</b>	<b>Returns (%)</b>
<b>1</b>	-7.119
<b>2</b>	1.400
<b>3</b>	-2.753
<b>4</b>	5.414
<b>5</b>	-1.064
<b>6</b>	2.094
<b>7</b>	-4.118
<b>8</b>	<b>8.099</b>
<b>9</b>	<b>1.592</b>
<b>10</b>	<b>3.132</b>

As we can infer from table 16 the next day comes with a negative return of -7.2%, whereas the day after, the returns become positive 1,4%. For the upcoming days, the returns fluctuate until day 8 when the highest possible return occurs. From day 8 until day 10 the returns are positive, so these days seem good in order to take some possible investment actions.

## **7 Conclusion and Discussion**

This section concludes the results of the present study and discusses the findings. The first subsection discusses the main findings of the analysis. Then, the implications of the results relevant to management are discussed, while the final section presents several limitations, as well as suggestions for future research.

### **7.1 Main findings and discussion**

The present study aims to answer the research questions by testing some hypotheses for whether twitter information is important for cryptocurrency financials and if applicable to predict future returns of bitcoin. After conducting the research, the results show that tweet sentiment and tweet volume have a significant effect on cryptocurrency financials and especially on bitcoin returns and trading volume. The results provide evidence to support both hypotheses 1 and 3, whereas hypothesis 2 is not supported. In addition, the fact that the positive effect of tweet sentiment on returns and tweet volume on bitcoin's trading volume is robust after conducting all the possible tests.

After testing all the hypotheses and conducting all the possible tests we can answer the research questions stated in the begging. For the first question stated, "*What is the relationship between bitcoin tweets and bitcoin price?*", the VAR and granger causality tests do find a negative relationship between these variables, but it is insignificant at a 5% level. For the second question "*What is the magnitude of the potential relationship between Twitter and bitcoin*", the present study does find a positive and significant relationship between tweet sentiment and tweet volume with the returns of bitcoin and the trading volume. That relationship is statistically significant at a 5% level, and it is one way, meaning that Twitter makes the cryptocurrency financials move and not the other way around. Then, for the third question "*Do positive sentiment increase the price or the returns of bitcoin?*", this study finds that, indeed, positive sentiment can potentially increase

the returns of bitcoin. This study also finds that the lagged value of sentiment (1 and 2 days before) can be useful to predict the values of the current returns. That means that if we know yesterday's sentiment, we can predict today's returns as well as the upcoming day. Finally, this study also examines different methods in order to find the best model to predict the future returns of bitcoin. After conducting three different methods and models, this study finds that ARIMA does perform better when it comes to predicting the future returns of bitcoin.

To conclude, the research goal of this study is to measure the extent to which Twitter information can be used to explain cryptocurrency financials and predict future values of certain variables. We can say that microblogs seem to have predictive power on bitcoin financials. From this study, we can infer that tweet sentiment and tweet volume are important measures to explain some features of bitcoin and predict future values. This study also shows that ARIMA is powerful when it comes to predicting the future returns of bitcoin. Moreover, tweet volume can be used to predict the trading volume of bitcoin.

## **7.2 Managerial implications**

The results of this study are mainly relevant for investment companies and individuals who would like to invest in or trade cryptocurrencies and specifically, bitcoin. The study shows that tweet sentiment and tweet volume can be used to generate returns in the future while at the same time to predict trading volume of cryptocurrencies.

The results suggest that tweet sentiment can provide significant information to predict future returns, whereas tweet volume can be used to predict the trading volume of bitcoin. The results allow financial professionals to use bitcoin-related tweets as useful indicators of bitcoin future returns. Moreover, stakeholders can use the predictive models in order to find future possibly values. Thus, tweet sentiment and tweet volume can be used to develop trading strategies for the bitcoin market.

Moreover, since the returns of bitcoin and trading volume are influenced by interactions on Twitter, companies could follow online trends more and take actions based on these trends. For instance, investment companies can conduct sentiment analysis not only on Twitter but also on many social media platforms in order to have more robust results.

Furthermore, this action could be also implemented by companies in order to improve their existing services. Knowing the sentiment of people about a brand, could help managers to change or adjust the service on peoples' needs.

All in all, Twitter does provide information about cryptocurrency financials. However, since this market is relatively new and not much research has been previously done, investors and all stakeholders must use these findings with caution.

### **7.3 Limitations and future research**

The present study comes with several limitations that create interesting possibilities for future research. First, this study focuses and analyzes only the impact of Twitter information on bitcoin financials while many other platforms contain relevant information about sentiment such as Reddit, Facebook, or Instagram. Thus, future research could use these platforms to combine or compare the results and find interesting new findings.

Second, a limitation that comes along with the first one is that only English tweets were considered for this study. Due to the lack of knowledge and the computational time of translating other languages, this study analyzed only English posts. A possible future research could consider other languages such as Chinese and generally Asian countries because they are a substantial part of the cryptocurrency market.

Third, this study did not hold power for exogenous news as a variable. The news could play an important role in the price direction and thus should somehow be controlled to have a better understanding of the cryptocurrency market. For this analysis, collecting all the relevant news for a period of 5 months could be very time-intensive and complicated. Thus, future research could use relevant news in order to control for any price changes.

Fourth, for this study, the seasonality trend was not examined. That happened since the period was just 5 months and was unclear for having any pattern in that period. But Cyrus Ip (2019), examined the seasonality of bitcoin for 10 years. He found out that bitcoin has the best average monthly performance in *April, May, October, and November*, whereas markets were relatively quiet in *December, Q1*, and especially during *summertime*. Thus, future research may also consider these periods of the season when conducting bitcoin financial analysis.

All in all, despite these limitations, this study provides important information about the relationship between Twitter and Bitcoin financials which can be used in various fields for either academic or managerial purpose.

## References

- Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, 1(3), 1.
- Akaike, H. (1973) "Information theory and an extension of the maximum likelihood principle" in *2nd International Symposium on Information Theory* by B. N. Petrov and F. Csaki, eds., *Akademiai Kiado*: Budapest.
- Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59 (3), 1259–1294.
- Banerjee S, Dellarocas C, Zervas G(2021). Interacting User-Generated Content Technologies: How Questions and Answers Affect Consumer Reviews. *Journal of Marketing Research*.58(4):742-761.
- Bollen, J., Huina M., and Xiaojun Z. (2011). Twitter Mood Predicts the Stock Market. *Journal of Computational Science* 2(1): 1–8.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2), 48-57.
- Daugherty, T., Eastin, M. S., & Bright, L. (2008). Exploring consumer motivations for creating user-generated content. *Journal of interactive advertising*, 8 (2), 16–25.
- Davidov, D., Tsur, O., and Rappoport, A. (2010). Enhanced sentiment learning using Twitter hashtags and smileys. pages 241–249. *23rd International Conference on Computational Linguistics (COLING 2010)*.
- Dekimpe Marnik G., and Hanssens Dominique M., (1999), "Sustained Spending and Persistent Response: A New Look at Long-Term Marketing Profitability," *Journal of Marketing Research*, 36(November), 397–412.
- Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74 (366a), 427–431.
- Fuller, W. A. (2009). Introduction to statistical time series, Vol. 428, *John Wiley & Sons*.
- Granger Clive W., and Newbold Paul (1986), *Forecasting Economic Time Series*, 2d ed. *New York: Harcourt Brace Jovanovich*.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.
- Hasbrouck, J. (2003). Intraday Price Formation in U.S. Equity Index Markets. *The Journal of Finance*, 58(6), 2375–2399.
- Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4), 18-28.

- Huberman, B., Asur, S. (2010). Predicting the Future with social media. *2010 international conference on web intelligence and intelligent agent technology*.
- Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *In Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59-68.
- Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! *In Proceedings of the international AAAI conference on web and social media* (Vol. 5, No. 1, pp. 538-541).
- Krumm, J., Davies, N., & Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, 7 (4), 10–11.
- Kwartler, T. (2017). Text mining in practice with R. *John Wiley & Sons*.
- Lee, J. K. (2021). Emotional Expressions and Brand Status. *Journal of Marketing Research*, 58(6), 1178–1196.
- LIU, Y., TSYVINSKI, A. and WU, X., 2022. Common Risk Factors in Cryptocurrency. *The Journal of Finance*.
- Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.
- Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297-303.
- Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 135(2), 293-319.
- Makarov, Igor, and Antoinette Scholar. 2019. "Price Discovery in Cryptocurrency Markets." *AEA Papers and Proceedings*, 109: 97-99.
- Manchanda, P., Packard, G., & Pattabhiramaiah, A. (2015). Social dollars: The economic impact of customer participation in a firm-sponsored online customer community. *Marketing Science*, 34(3), 367-387.
- Marek Jarociński, Bartosz Maćkowiak; Granger Causal Priority and Choice of Variables in Vector Autoregressions. *The Review of Economics and Statistics* 2017; 99 (2): 319–329.
- Meire, M., Hewett, K., Ballings, M., Kumar, V., & Van den Poel, D. (2019). The Role of Marketer-Generated Content in Customer Engagement Marketing. *Journal of Marketing*, 83(6), 21–42.
- N, Mishra, and C. K. Jha. (2012) "Classification of Opinion Mining Techniques." *International Journal of Computer Applications* 56 (13).



- Naab, T. K., & Sehl, A. (2017). Studies of user-generated content: A systematic review. *Journalism*, 18(10), 1256–1273.
- Naka, A. and Tufte, D. (1997). Examining impulse response functions in cointegrated systems, *Applied economics* 29(12): 1593–1603.
- Narayanan, A., Bonneau, J., Felten, E., Miller, A., & Goldfeder, S. (2016). Bitcoin and cryptocurrency technologies: a comprehensive introduction. *Princeton University Press*.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2), 1-135.
- Pauwels, K., & Weiss, A. (2008). Moving from Free to Fee: How Online Firms Market to Change Their Business Model Successfully. *Journal of Marketing*, 72(3), 14–31.
- Pauwels, Koen, Dominique Hanssens, and S. Siddarth (2002), “The Long-term Effects of Price Promotions on Category Incidence, Brand Choice and Purchase Quantity,” *Journal of Marketing Research*, 39 (November), 421-439.
- Peng, J., Agarwal, A., Hosanagar, K., & Iyengar, R. (2018). Network Overlap and Content Sharing on Social Media Platforms. *Journal of Marketing Research*, 55(4), 571–585.
- Sattarov, O., Jeon, H.S., Oh, R. and Lee, J.D. (2020), “Forecasting bitcoin price fluctuation by Twitter sentiment analysis”, In 2020 *International Conference on Information Science and Communications Technologies (ICISCT)*, IEEE, pp. 1-4.
- Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A Joint Model of Sentiment and Venue Format Choice. *Journal of Marketing Research*, 51(4), 387–402.
- Sims, C. A. (1980). Macroeconomics and reality, *Econometrica: Journal of the Econometric Society* pp. 1–48.
- Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20 (5), 926–957.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168.
- Trusov, M., Bodapati, A. V., & Bucklin, R. E. (2010). Determining Influential Users in Internet Social Networks. *Journal of Marketing Research*, 47(4),
- Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment, *Fourth international AAAI conference on weblogs and social media*.
- White, L. H. (2015). The market for cryptocurrencies. *Cato J.*, 35, 383.

# Appendix

Figure 8: Sentiment per day

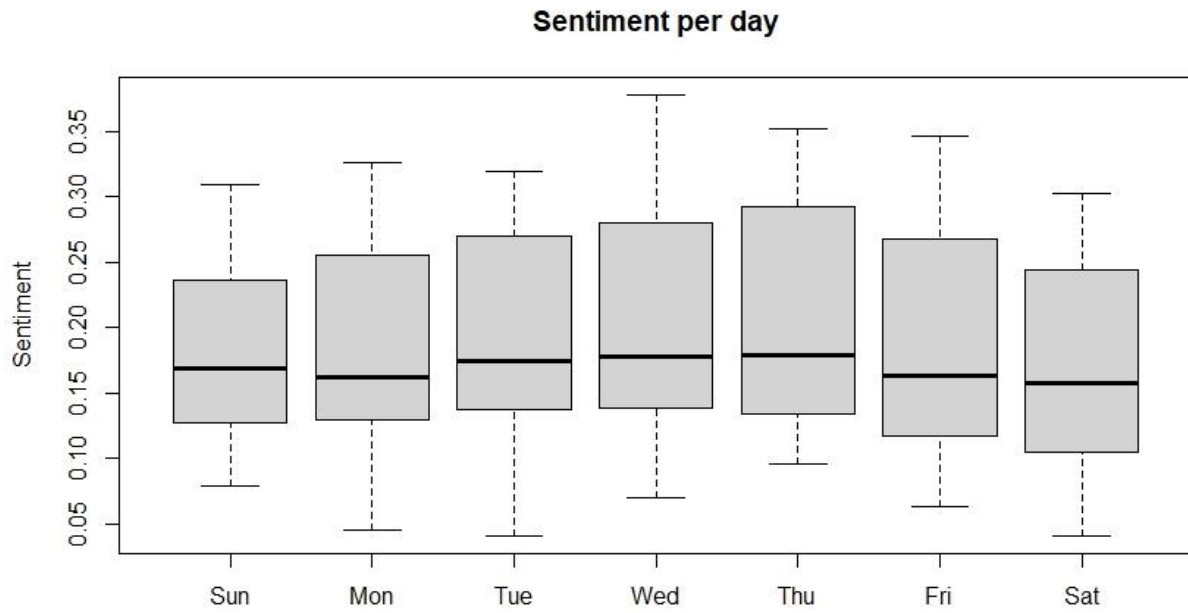


Figure 9: Tweet volume per day

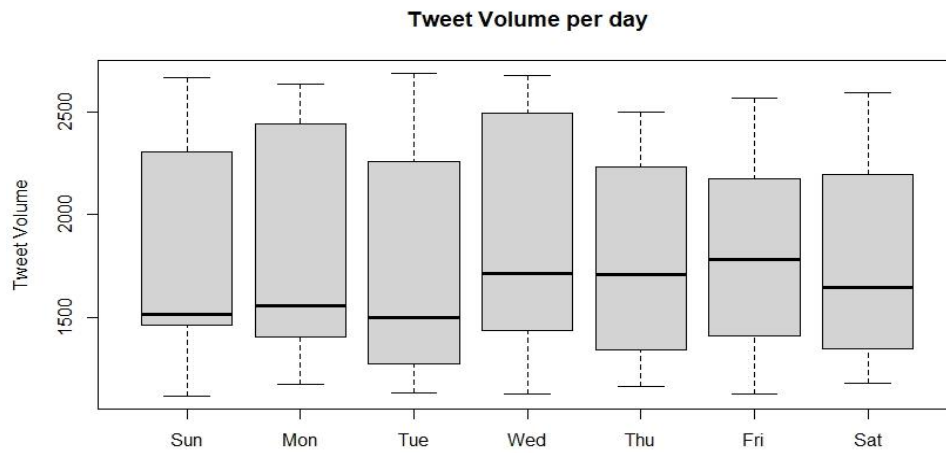


Table 6: Variables and data sources

Variables	Definition	Data Source
<i>Control Variables</i>		
Date/Day/Week	The days/weeks of the period of interest	Kaggle.com
<i>Independent Variables</i>		
Tweet sentiment	The average sentiment per day from tweeter posts	Kaggle.com/app.intotheblock.com
Tweet Volume	The average tweets per day	Kaggle.com/ bitinfocharts.com
<i>Dependent Variables</i>		
Price	The closing price of bitcoin	Coinmarketcap.com
Change	Change/returns of bitcoin price in %	Coinmarketcap.com
Trading volume	The trading volume of bitcoin in thousands (USD)	Coinmarketcap.com

Table 11: Lag selection scores

Lags	1	2	3	4	5	6	7	8	9	10
<b>AIC</b>	2.7113	<b>2.7103</b>	2.7365	2.7620	2.7903	2.7980	2.8164	2.8157	2.8270	2.8425
<b>HQ</b>	2.7419	2.7663	2.8180	2.8689	2.9227	2.9558	2.9997	3.0245	3.0612	3.1022
<b>SC</b>	2.7867	2.8485	2.9375	3.0258	3.1170	3.1874	3.2686	3.3308	3.4049	3.4833
<b>FPE</b>	5.9646	5.9227	7.7571	1.0160	1.3823	1.5482	1.9611	2.0942	2.5795	3.4149

Figure 8: Impulse from price

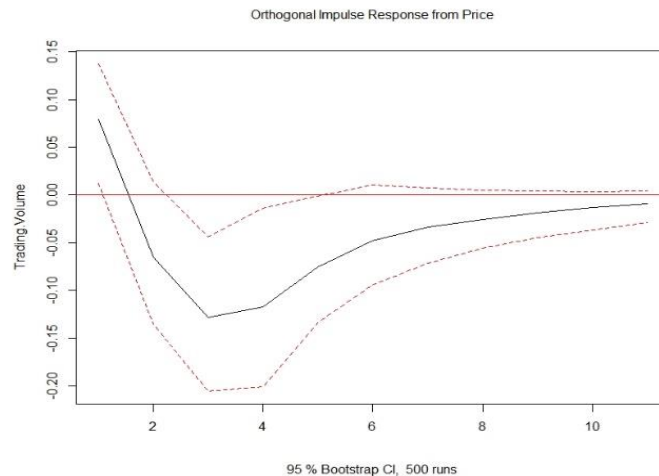


Table 12: Arima model selection

ARIMA MODEL	Performance (AIC)
ARIMA (2,0,2)	Inf
ARIMA (0,0,0)	688.0026
ARIMA (1,0,0)	685.3407
ARIMA (0,0,1)	686.1176
ARIMA (0,0,0)	686.1492
ARIMA (2,0,0)	686.4944
ARIMA (1,0,1)	685.5226
ARIMA (2,0,1)	687.4858
ARIMA (1,0,0)	683.566
ARIMA (2,0,0)	684.6614
ARIMA (1,0,1)	683.6891
ARIMA (0,0,1)	684.3481
ARIMA (2,0,1)	685.6611

Figure 9: Bitcoin's returns (ARIMA)

