# Data wrangling: Project Report

## The Relationship Between the Price of Bitcoin and the News on Financial Markets

Group 15

Bogdan-George Dragomir (bdr570), Jialong Mei (jmi620),
Rosalie Fischer (rfr540), & Emma Ståhlberg (esg780)

February 6, 2023

## 1 Research Question

Modern times introduce novel concepts and ideas that the general population finds intriguing and difficult to adjust to. Therefore, individuals seek reliable data from which solid conclusions can be drawn. One such topic explored in this paper is cryptocurrency, with Bitcoin serving as the earliest and most well-known example. Given that the entire cryptocurrency market is believed to be highly unpredictable, it may be useful to analyze current data in order to identify patterns in its erratic behaviour. Therefore, the present project, by applying existing procedures on different data sources, was designed to offer an answer to the following research question:

> *Is the price of Bitcoin influenced by news reports on the financial markets?*

## 2 Data Sources

To answer the aforementioned research question, we used multiple data to facilitate our journey. All the files mentioned can also be found in the folder of the project. The sources of our collected data are the following:

Yahoo Finance (Last accessed: 24/01/2023)[2]
Data collected: $BTC-USD-Prices.csv$, which covers multiple features relating to the price of one Bitcoin (such as highest and lowest price in a day etc) converted to US dollar. Using the filter options available on the website, we selected the maximum time possible covered in the data (September 17, 2014 - January 24, 2023). Moreover, each record in the data corresponds to one day.

Data World (additionally including CNBC) (last accessed: 24/01/2023)[1]
Data collected: *cnbc_news_dataset.csv*, which covers multiple features such as the urls of the financial news presented on the CNBC website, their titles, the authors of the articles, corresponding publication dates etc. All the urls presented in the dataset were written on the website called CNBC.

Our own dataset:
We created a pandas series which contains the sentiment of a news article (negative, neutral, positive). Each sentiment was drawn from each description of a news article, available in the *cnbc_news_dataset.csv*, after all the data wrangling methods necessary to transform our collected datasets were performed. More details are presented in section three.

## 3 Data Wrangling Methods

To successfully operate with our data, we used multiple data wrangling methods, ranging from simple acquisition of the date, cleaning the data and so on.

### 3.1 Data Acquisition

The first procedure that we did was to extract the information in a format that is useful for further processing. We used two available datasets in a CSV format (the data sets previously mentioned in section 2). The pandas method *read_csv* was used twice in this context.

### 3.2 Data Cleaning

Our datasets contain features that we considered not to be essential to reach a valid conclusion to the research question provided. Therefore, various cleaning actions were performed, so our datasets will contain only the elements that we believed are strictly necessary.

First of all, we will not use the columns *header_image*, *raw_description*, *scraped_at*, *keywords*, because those contain elements that were not needed. Thus, we used the function *drop* to modify the existing dataset. Also, since we already know that all the news were provided by CNBC, the column *publisher* is no longer a key element in our project. So, we deleted it as well.

Second of all, we thought that a well structured article on a website should have at least some details relating to the author of the article (such as his or her name, email address etc). At a first glance, we saw that some records do not have a known author. Considering that a specific news article which may produce some effects on the price of Bitcoin should have an author that can be researched to determine the knowledge the author has on the financial market, we decided to delete the records that do not have a current author written (those that have a *nan*). Thereby, we used the method *dropna*. At this point, we observed that the columns *short_description* and *description* were identical. Consequently, we chose to delete *short_description*.

Third of all, we observed that the news in the dataset *cnbc_news_dataset.csv* are not structured in a specific order (increasing or decreasing). Therefore, we decided to change the structure by organising the records in an increasing order, according to the date of publishing (given in the *published_at* feature). We did this using the function *sort_values*. Furthermore, after sorting the news by their dates of publishing, we see that the first article was written in 2006. Taking a close look at the $BTC - USD - Prices.csv$, the first price of Bitcoin in the data set is from September 17, 2014. Hence, the news that were published before this date are irrelevant, as we do not have available data to be analysed during those years. As a consequence, we decided to eliminate the records that contain news published before September 17, 2014, using *drop*. Also, we saw that the indexes are not placed in an organised order, so we used the method *reset_index*. Now the indexes are in increasing order.

At this point, where we mostly had structured our news dataset, we realised that our datasets with prices have some data that are beyond the maximum period that we have in the news dataset. Since we have news until 2021 and the prices until 2023, we deleted the records in the $BTC - USD - Prices.csv$ that had a date beyond the last date in *cnbc_news_dataset.csv* (again using the *drop* method). We also used the same method to drop the records of the prices that are before the first date in the news dataset.

### 3.3 Creation of Our Own Dataset & Appending it

We added a new column called *Sentiment* to the cnbc news dataset, that holds the sentiment that can be taken from each article presented in the dataset, after all the procedures presented above were completed. To facilitate our work and efficiency in terms of time management, we decided to use the description of each article to determine its sentiment. After a manual quick look through some articles, we decided that each description approximately matches the true sentiment of the article. Naturally, if we had used the entire text from an article, then the value obtained would have been more precise. Therefore, we adapted our reasoning based on the minimum and maximum values presented below.

We integrated a library called *TextBlob*, which can be used to determine the sentiment analysis of a text. Then, by iterating through each description in the *Description* column in the news dataset, we used the method *TextBlob* on each text. The output of the method is a value between 1 and -1, used to determine the sentiment analysis of the text (1 being most positive and -1 most negative). Nonetheless, after we ran our algorithm, we observed the minimum value computed was -0.14749206349206348 and the maximum value 0.33888888888888885.

So, we decided that a value between -0.05 and 0.05 would represent a neutral feeling, over 0.05 a positive feeling and under -0.05 a negative feeling. Finally, each adequate feeling was added to the series *sentiments_column*.

### 3.4 Data Merging

By using the command *pd.concat*, we successfully merged the dataframe *cnbc_news_dataset.csv* with the *sentiments_column*.

### 3.5 Data Processing

One idea that we thought is going to give us a specific conclusion revolves around checking the prices of Bitcoin for 15 days, near the date of each news in our news dataset (5 days before the date and 10 days after the date). We considered a range of 15 days to be able to analyze the prices, considering that the prices may take a few days to adjust, depending on the influence of each news (if any influence at all). After each reduced data frame of 15 records obtained for each news, using the principle mentioned above, we counted the number of prices that increased, and the number of prices that decreased. Then, by taking the index and the sentiment of each record, we compared the number of prices that increased or decreased. If the difference between the number of prices increased and the number of prices decreased was greater than 2, and the sentiment is "positive", then we considered that the news influenced the price (a positive sentiment made the prices increase for a few days). The same principle was used to determine if the news influenced the prices in the context of negative or neutral sentiment. Using a dictionary, we kept the id and sentiment of each news and the output of the prices (influenced, increased/decreased, or not influenced). The final step was using text processing (by means of regular expressions) to extract the sentiment and the outcome from the dictionary and compare them to see whether the news indeed influenced the prices. After this process, we counted the number of news that had an impact on the Bitcoin prices.

### 3.6 Text Processing: Regular Expressions

We used two patterns to extract the sentiment and the outcome for each key and value presented in the dictionary mentioned in the section "Data Processing". For the sentiment, we captured the text which contains the sentiment, and then we used another pattern to obtain the final output (increased, decreased, neutral). The used patterns can be reviewed in the ipynb notebook.

### 3.7 Data Visualizations

The plot below covers the habits of the Bitcoin prices over time (the first and last price included in the dataset, after cleaning the data). We can observe that the prices continuously increased with an accentuated performance registered from the 2020 onwards.


Bitcoin Prices Over Time

## 4 Conclusion

After all the data wrangling methods mentioned, we found that only 64 out of 233 news had a control over the prices of the Bitcoin, which represents only 27.46% of the total amount of news after cleaning our datasets. Therefore, this number suggests that the prices of Bitcoin are not influenced by the news on the financial markets. Nevertheless, this conclusion should only be considered as representative in the context offered by the data collected, together with the methods used to process it. Since we used only one source to gather information

about the financial news, there is no guarantee that our results essentially covers the true relationships between the news on the financial market and the prices of Bitcoin or any other cryptocurrency. If we had used more sources, different from CNBC, then the results probably would have been much more different (since there may be the presence of bias, that surely affects the results of any research). Furthermore, our text analysis used a simple library that we could not decide whether or not was extremely accurate with its results. If we had used the entire text from an article, or more advanced techniques or tools, then there is the probability that the results would have been more accurate. Moreover, there are multiple ideas that can be used to determine if there is a relationship between the prices and the news. If our solution provided the current results, maybe another one would provide a different number, which might invalidate the current answer to our research question.

# References


[1] CrawlFeeds. Cnbc news dataset - dataset by crawlfeeds, Oct 2021. URL `https://data.world/crawlfeeds/cnbc-news-dataset`.

[2] N.D. Bitcoin usd (btc-usd) price history amp; historical data, Jan 2023. URL `https://finance.yahoo.com/quote/BTC-USD/history?period1=1410912000&period2=1674518400&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true&guccounter=1`.