Correlation Between Bitcoin Pricing Movements And Sentiment of Public Discourse - Progress Report

Chris Messer, Erin Abbott, Devyn Byrd

# 1 Introduction

Bitcoin is a digital currency that was created in 2009. It is a decentralized form of currency, meaning it is not regulated by any government or central bank. Bitcoin transactions are recorded on a public ledger, called the blockchain, and they are secured using cryptography. Bitcoin can be used to purchase goods and services, or it can be held as an investment.

Like all currencies, the value of a Bitcoin is only what everyone agrees it is. Because there is no inherent worth to a Bitcoin, it is not backed by any physical commodity or government entity, the value in relation to the U.S. dollar can fluctuate greatly. Measuring from the first time a transaction was recorded to exchange physical goods for an amount of Bitcoin (now known as pizza day) to the peak price in 2021, the USD price equivalent increased from $.003 to $61,000 per Bitcoin. The price of bitcoin is highly volatile and can change drastically over short periods of time. This is due to a number of factors, including speculation, market forces, news about the currency, and the availability of exchanges. Bitcoin's price is also influenced by supply and demand. When demand for bitcoin increases, the price typically rises, and when demand decreases, the price usually drops.

Communities have formed across the internet both in-favor of a cryptocurrency revolution and against adoption of this new age tech. For example, communities on the online forum reddit.com, or general discourse on Twitter.com. This sentiment of this discussion is often a reflection of the price of bitcoin. This is intuitive, as people are more likely to speak negatively of an asset when it is underperforming, and positively when it is over-performing. Using machine learning and data processing tools available today, we can easily measure the sentiment of this discussion in a quantifiable way.

The use of sentiment analysis to extract valuable insights from customer feedback has been present since the 1950s, and since then it has continually evolved. Social media such as Twitter and Reddit provide diverse exposure to businesses, allowing them to connect to customers, receive feedback, and use sentiment analysis to improve or evolve their products and services. As many people are deeply invested in the cryptocurrency markets and regularly post technical analyses and thoughts, these posts can have an effect on the market.

In this paper, we explore the notion that the relationship between Bitcoin pricing movements and sentiment of public discourse around bitcoin is one of causation rather than correlation. E.g. the sentiment around Bitcoin actually drives the price of Bitcoin. We hypothesize that by looking backwards a period of time *n* from time *t,* we can predict the direction the price moved from *t-n* to time *t.* This paper is meant to serve as a progress update on our efforts in this research question, the challenges we have faced, and the next steps for our analysis.

# 2 Methodology

This study has been divided into four parts. First, we discuss the extraction of the data we will analyze using sentiment analysis, and how we approached “tagging” the qualitative discussion around Bitcoin with a quantitative measure of sentiment. Second, we perform some exploratory analysis on the now prepared datasets to explore the shape of the data, and determine whether there is any correlation between our two data populations, from Reddit and from Twitter. Third, we will perform a series of regression analyses to attempt to build a model that can appropriately predict pricing movements of Bitcoin, and analyze each model’s appropriateness. Finally, we will perform a retrospective analysis and determine “how well” our models have performed.

# 3 Data Preparation

Our analysis will focus on three sets of data:

* Tweets about Bitcoin
* Bitcoin Pricing information
* Reddit comments that contain the word Bitcoin

## 3.1 Twitter Data

Our twitter data contains two sets of data we will use for the analysis. It contains the sentiment tags for tweets about Bitcoin, and also the Bitcoin pricing information.

On first inspection, it appears the data has some missing data. We will need to be sure to handle this before running our analysis.

First, we separated the twitter data columns from the Bitcoin pricing data. Then, since we want to analyze the data at different time intervals, we aggregated the data at the day level rather than just the hour interval and stored it as a new dataframe.

## 3.2 Bitcoin Pricing information

As mentioned above, the Twitter data also contained information on Bitcoin Open/High/Low/Close (OHLC) at each interval. However, as noted, we are missing rows of data in that data set, which includes the Bitcoin data at those hours. As such, we determined it appropriate to seek a secondary source for Bitcoin pricing data.

One challenge we faced was locating an OHLC Bitcoin dataset that was at the hour level, and most datasets were aggregated at the day level. For the hour interval we determined it appropriate to exclude the missing hours from the analysis entirely. Looking at the data summary above, only 578 records out of 12k were missing. Further, even if we brought in the Bitcoin pricing data for the missing hours, we would still be missing the independent variables for those hours.

As the twitter data was aggregated at the daily level, we determined it was appropriate to pull in Bitcoin prices at the daily interval as well. For this, we used the R package “crypto2”. This will allow us to align the aggregated daily twitter data with daily Bitcoin OHLC data.

## 3.3 Bitcoin Data

The dataset from reddit did not come as clean as the Twitter dataset. As such, several steps were performed to extract, transform, and load the dataset into model-ready format. This portion of the analysis was performed in Python to take advantage of the natural language processing packages.

The data was sourced from Kaggle, and contained comments on reddit.com from 2012-2019, with each row representing the comments. To prepare the reddit data from the kaggle dataset,there were a few transformations that needed to be done. Firstly, we aligned the reddit comments time frame with the kaggle dataset for the tweets about Bitcoin. The twitter dataset contains only tweets from August 2017 to January 2019, and the reddit comments reached all the way back to 2012. Therefore, we trimmed the reddit comment dataset to match the twitter dataset.

After filtering on date, it was noted that 4% of the records had blank values. These records did not appear to be corrupted records on import, but instead were comments attached to other comments. Because these comments had null values for the datetime they were posted, we would be unable to include them in our analyses.

Next, we inspected the data and determined there were other formatting concerns. Specifically, we should be sure to handle:

* urls
* special characters
* new lines
* foreign languages
* numbers (typically do not add context to the sentiment)

Urls, special characters, new lines, and numbers were removed from the data. Next, we used a machine learning package to do the language detection. Because the sentiment model we selected only works on english language, we removed the non-english comments (~2% of total comments.)

#### 3.3.1 Sentiment Tagging

There are many different approaches to sentiment analysis, from simple rule based approaches to deep learning and recurrent neural networks. To decide on a model, we considered two factors, accuracy of model and consistency with other data sets.

Regarding accuracy of model, we consulted a whitepaper, Social media sentiment analysis for cryptocurrency market that covers accuracy of different sentiment analysis models when trained on crypto currency content across twitter and reddit. The paper found that of 21 models tested, VaderSentement was one of the strong performing models. Because this is the same model used in the twitter data kaggle set, we opted to use this one for the reddit sentiment analysis as well. Using this package in Python, we are able to input a sentence, and the sentiment score is output indicating whether the phrase is positive/negative/neutral.

Finally, we aggregated the data. The original data was at a per comment basis, and the twitter data was aggregated at hourly intervals. So we summed up the total positive, negative, and neutral comments at each hour, in aggregate, and across different subreddits.

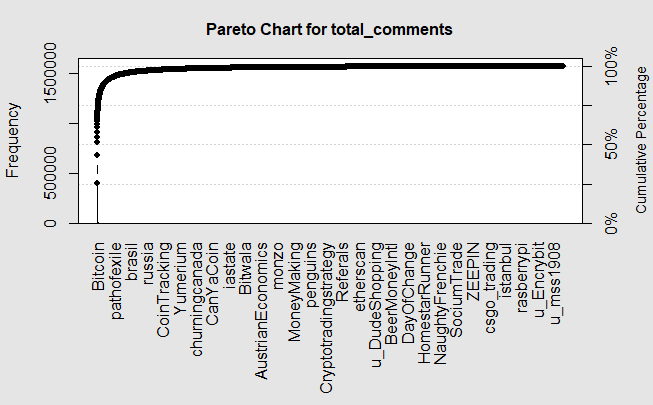
#### 3.3.2 Further Transformations in R

After preprocessing and tagging of the data was complete, further transformations in R occurred.

First, we aggregated the hourly data at a daily interval, to match the sets we created for the Bitcoin and twitter datasets. Next, we determined we wanted to use the data from reddit both in the aggregate (all comments across all communities) and also at the disaggregated level (treating comments in different communities as different independent variables.)

To accomplish this, we would need to create a new column for each different community, and for each community, three columns would be needed (positive, negative, and neutral comment count). A quick look at the summary of this data revealed there were too many communities to accomplish this. There are 11k communities in the dataset, and the majority of communities had a very low number of comments. If we were to create three columns for each community, we would end up with 33K columns, which would likely lead to severe overfitting of our model. As such, we determined it appropriate to examine which grouping of communities would lead to the most number of comments.

To accomplish this, we created a pareto chart, with the communities enumerated on the x-axis, and the cumulative percentage of comments they contribute to the total on the y-axis. This quickly shows us a very small percentage of the communities contribute to a majority of the total comments.



After exploring the underlying data, we noted that 10 communities contributed 65% percent of the total comments. As such, we elected to only use these communities for further investigation. These communities were then cast as columns onto the hourly and daily interval datasets.

The result of all of these operations is six data frames: an hourly and daily interval aggregation file for the twitter data, the reddit data, and the Bitcoin OHCL data.

These six data frames were then joined on day (for the daily interval set) and day and hour (for the hourly interval set) to make two dataframes, one for hourly interval and one for daily interval. These two final data frames were saved into csv files for reference, and will be then reloaded into a new workbook for further analysis.

#### 3.3.3 Final Transformations

Lastly, we will need to add a few columns to be our predictor variables. The goal of this analysis is not to “predict” the price of Bitcoin at a given interval given the sentiment of the hour. Being able to predict the price of Bitcoin at the top of the hour wouldn’t be useful if we have to wait until the top of the hour to know the sentiment of all tweets and reddit comments during that period.

Further, we also don’t want to just shift the close price down *n* number of rows, because we are also not attempting to predict the price of Bitcoin *at all*. The scope of this analysis is to predict the *correlation* between *price movements* and public sentiment. Thus, we are much more interested here in the sentiment of the comments/tweets during the hour and how they impact the direction the price moves in the next period.

To accomplish this, we add three columns:

* lagged\_price - the close price *n* periods ahead
* price\_dif - the difference between the close price of the observations and the lagged price
* price\_dir - the *direction* of the price movement. 0 for negative moves and 1 for positive moves

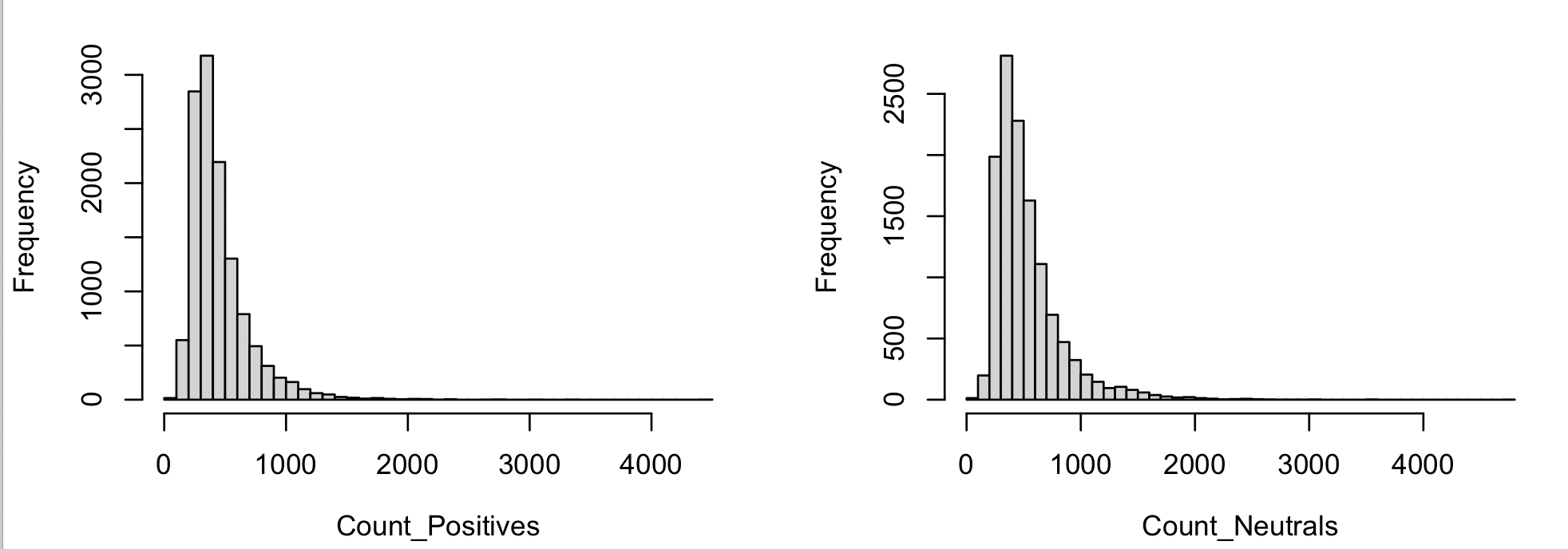
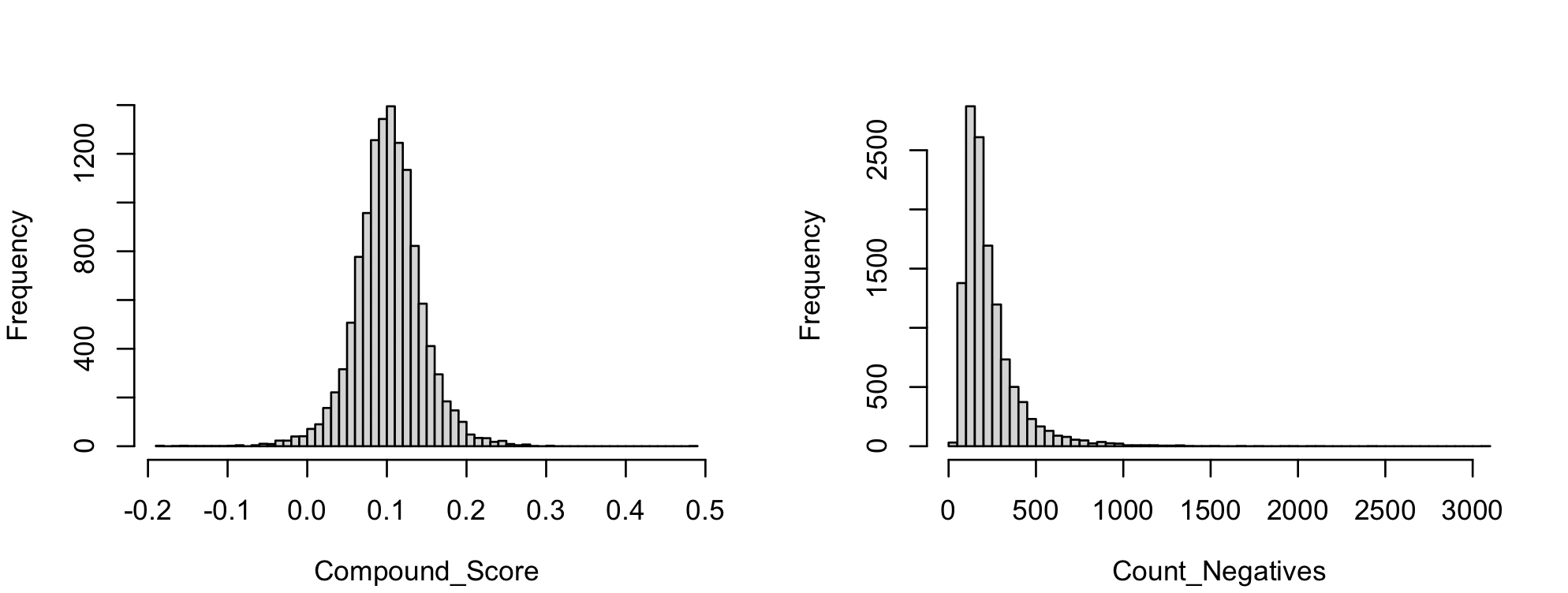
We wrapped these three transformations into a function, so that we can pass either a daily or hourly dataframe and a value of *n*  to make transformations on the fly. We will use this later on to try many different values of *n* to see which produces the best model.

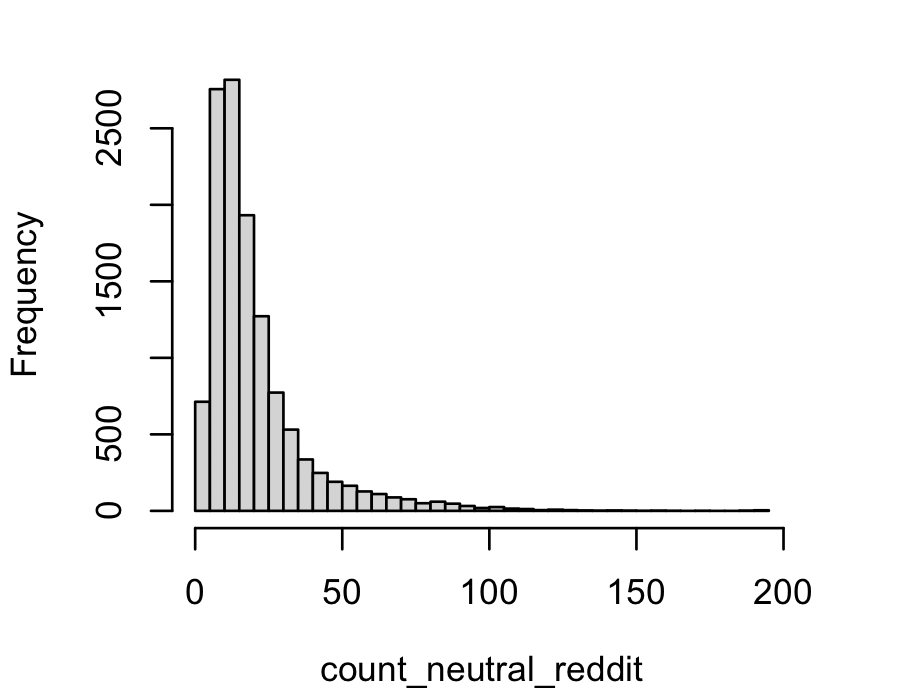
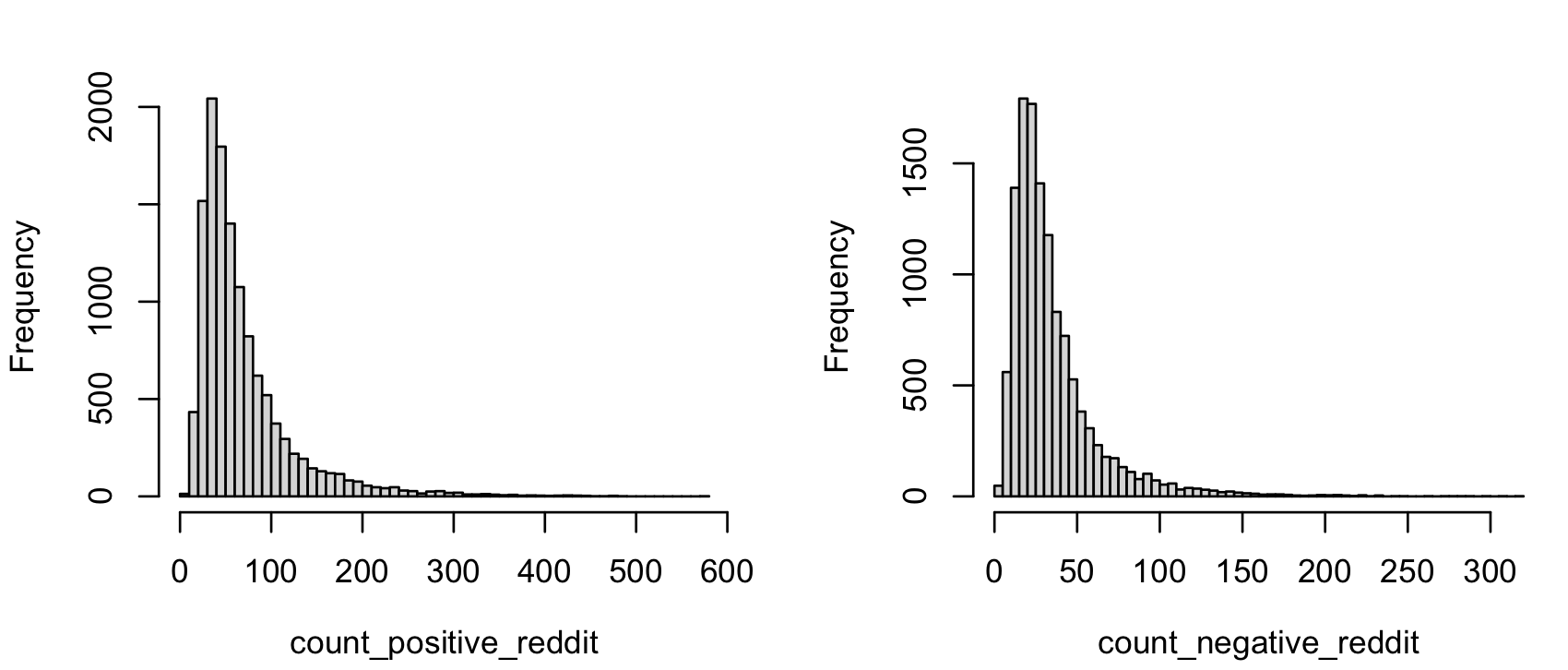
# 4 Data Exploration

All data exploration up to this point has only been done on the hourly datasets, so additional exploration of the transformed data as described above will be done in coming weeks. The main goal of the data exploration was to visualize the variables we have, see if any initial outliers stand out, and look at some initial correlations to verify if any variables seem to have the relationship we are expecting.

4.1 Variable Visualization

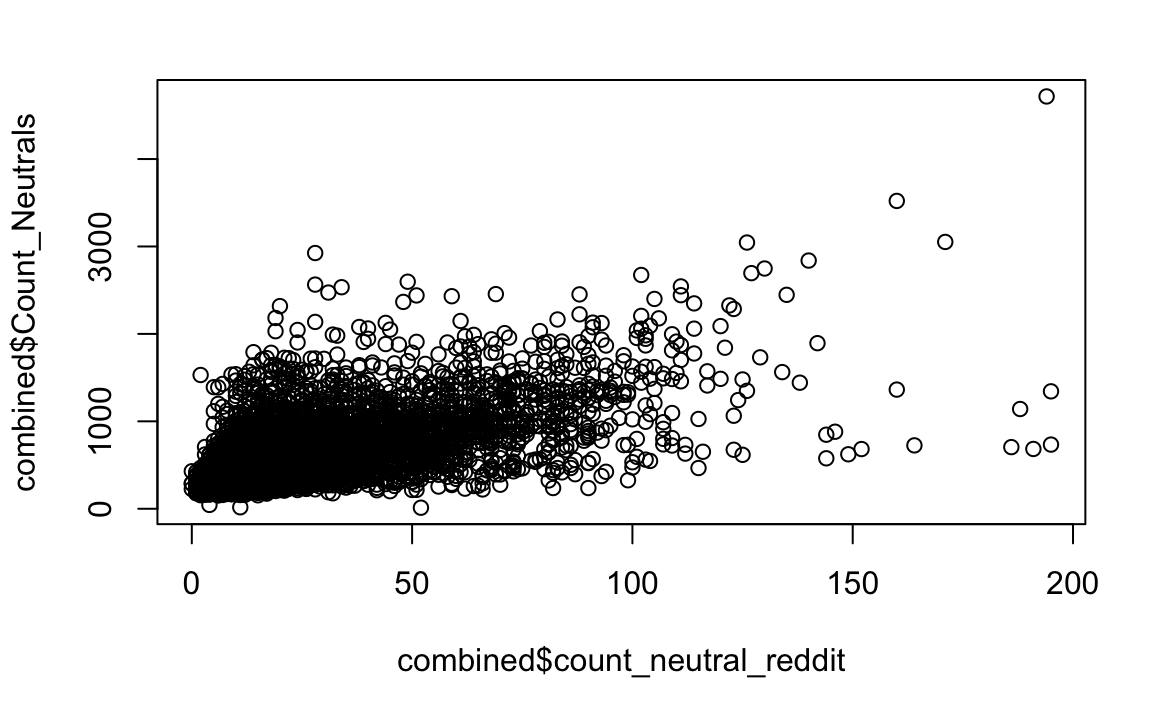
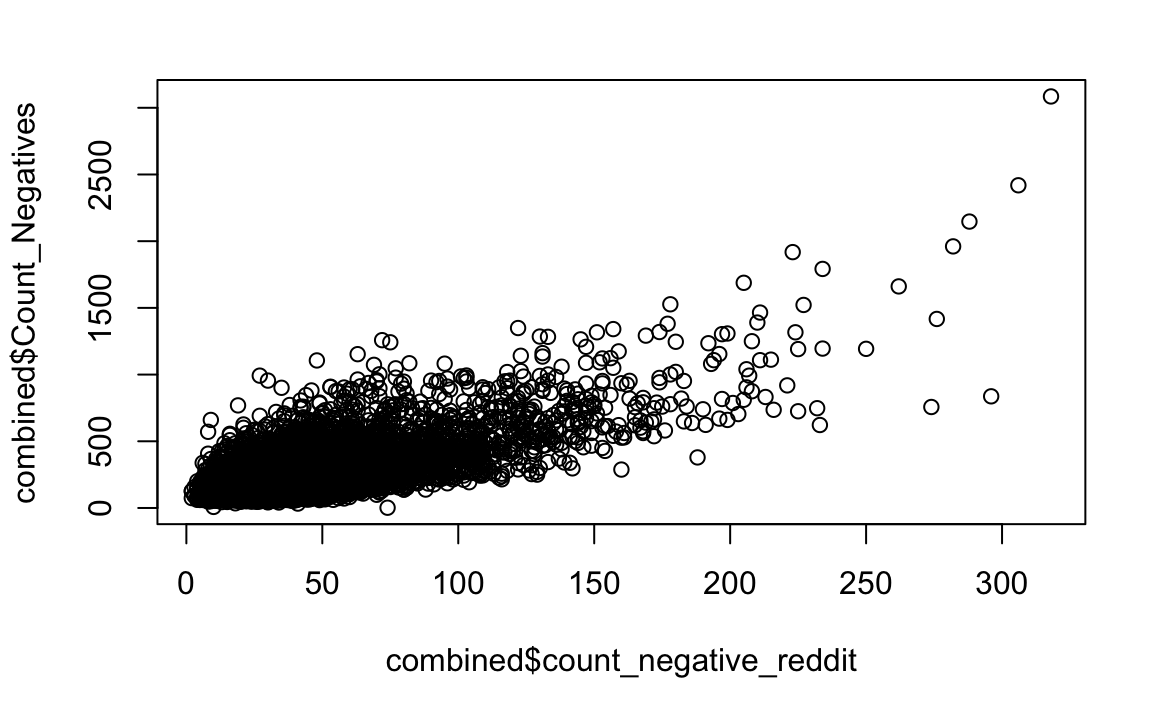
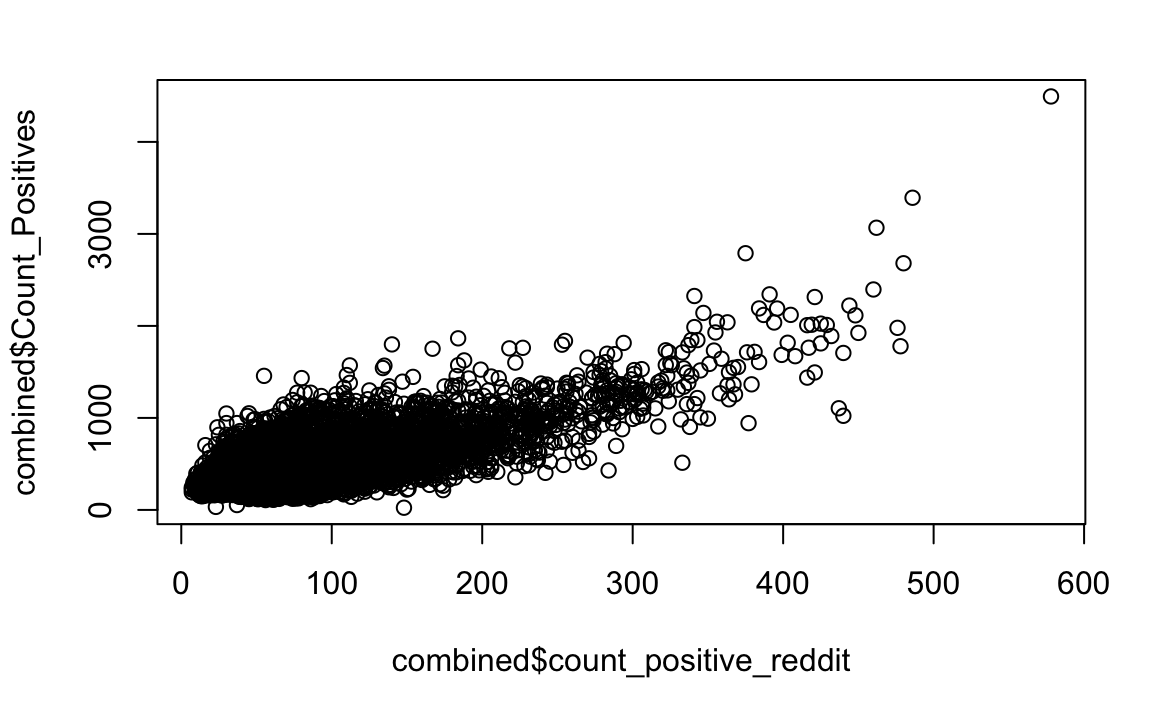
The first thing we did was visualize the distribution of our variables from the Twitter data using histograms to see if any appeared to have outliers that may need to be addressed. We can see that while the overall sentiment score (Compound\_Score) appears to be normally distributed, all of our count variables are skewed which indicates there are some high outlier values that may need to be imputed or removed prior to using the data in our models.



We repeated that exercise for the count variables from the Reddit data where we saw the same trend.

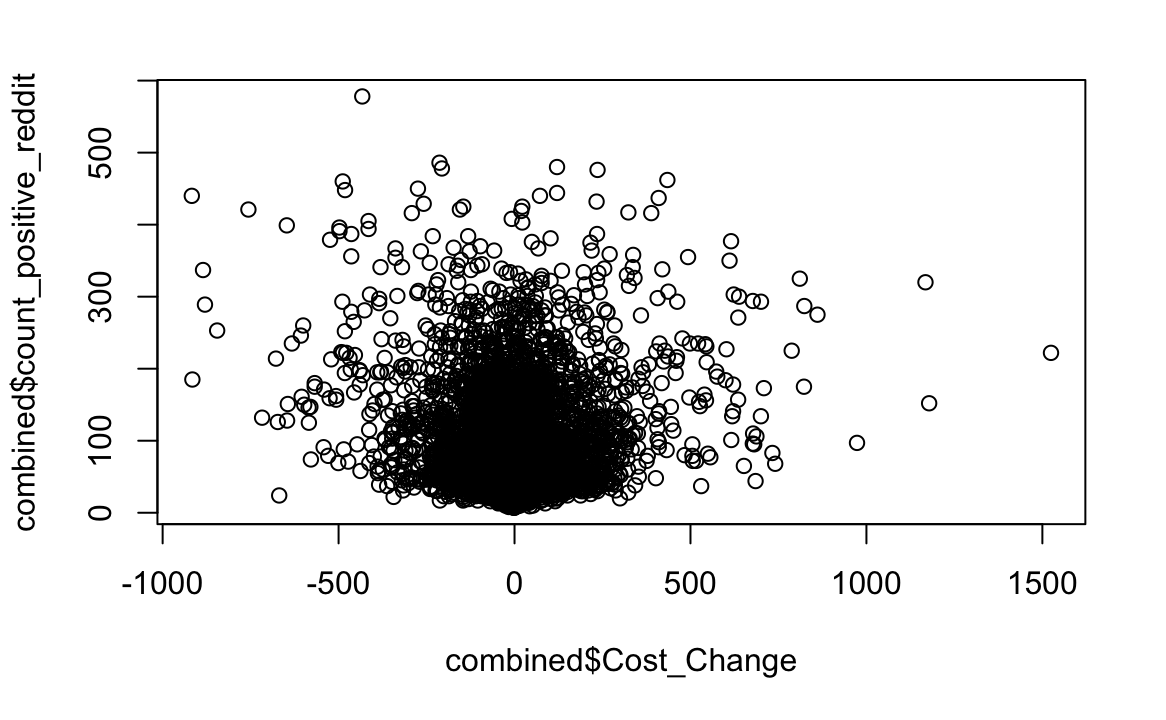
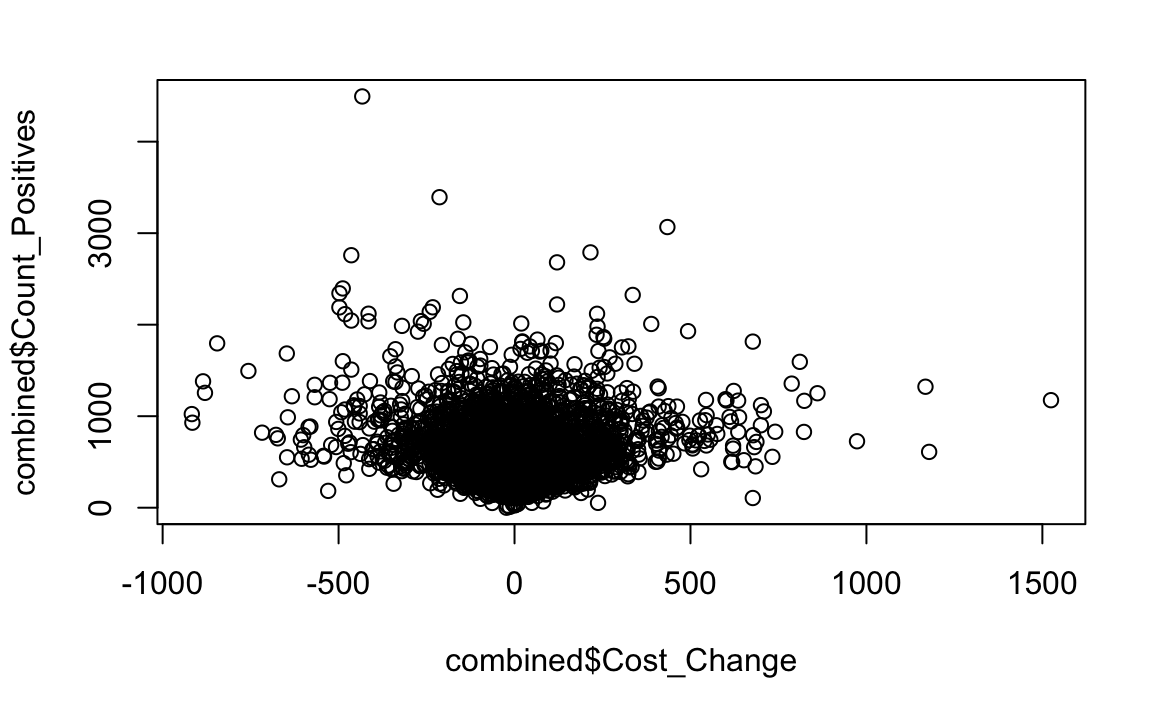
4.2 Variable Correlation

We first merged the hourly Twitter dataset with the hourly Reddit dataset. We then various variable combinations in order to verify if they have the relationship we are anticipating. We first compared the positive, negative, and neutral counts from each dataset to see if they are positively correlated. Our hypothesis was that if there are a higher number of positive tweets, for example, there would also be a higher number of positive reddit comments.



In general, the positive and negative comments do seem to have a positive correlation, while the neutral appears to have a weaker correlation. For this reason, multicollinearity could be present in models that use combinations of these variables, so this is something we will have to keep in mind.

The last thing we did was plot a few of these variables against the change in closing price of Bitcoin from one hour to the next (lag n-1) in order to see if any had a clear linear relationship. None of the combinations we tried had an obvious linear relationship. Below we see these plots for the count of positive tweets as well as the count of positive reddit comments.



However, this does not mean that they will not prove to be good predictors. They may need to be used in conjunction with other variables or we could need to use a different lag in order to calculate the change in Bitcoin price. We may also need to perform non-linear transformations on some of the variables. All of these options will be explored in upcoming stages of the project.

# 5 Data Analysis

## 5.1 Linear Regression

We will attempt to predict how far up or down the price of Bitcoin will move based on the sentiment.

A couple of challenges we anticipate to face:

* As the price of Bitcoin increases over the observation period, swings in the price will likely grow too, even if the same amount of people are tweeting/commenting about Bitcoin. We may need to introduce an interaction term to account for this.
* Regression is not meant to predict the price of the stock. Rather, we are attempting to measure the variance in the price of Bitcoin that is accounted for by public sentiment. This is a difficult concept to grasp and could cloud our analysis if not accounted for correctly.

## 5.2 Logistic Regression

Here, we will use the price\_dir variable, and try to predict the odds that the price moves up or down based on the independent variables.

# 

## Works Cited

Social media sentiment analysis for cryptocurrency market ... - arxiv. (n.d.). Retrieved March 27, 2023, from https://arxiv.org/pdf/2204.10185

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). Chapter 1: Affective Computing and Sentiment Analysis. In A Practical Guide to Sentiment Analysis (pp. 1–2). essay, Springer.