Crypto trends:
Google Trends Analysis
& Predictive Modeling

Written by Aleksandr Lazutin

lazzzzutin@gmail.com

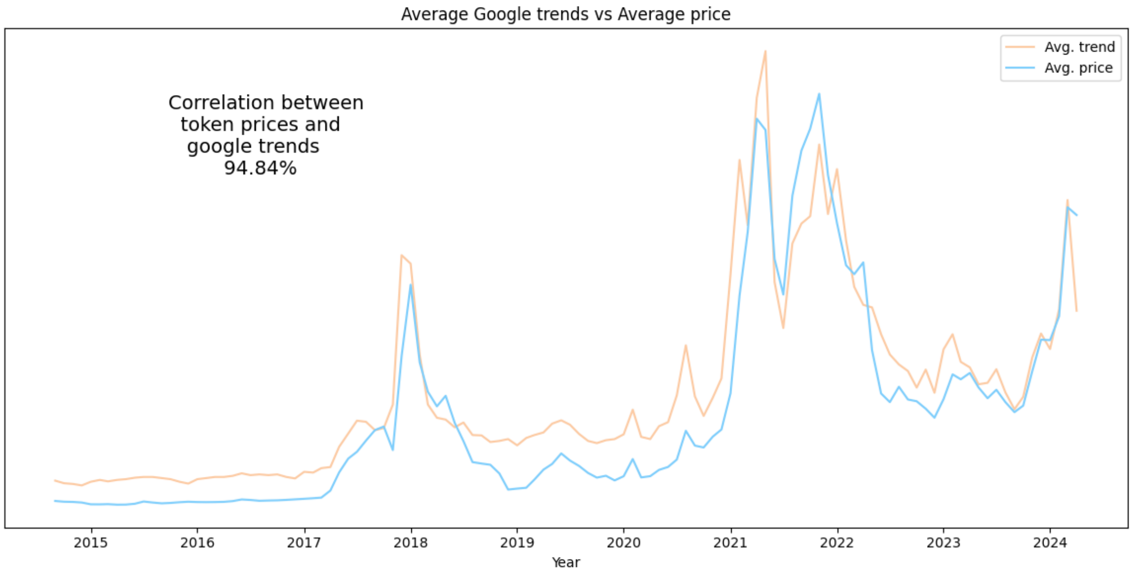
Introduction

World of cryptocurrency filled with a huge amount of data. Some of this data originates from exchanges, where crypto is trading (price, volume, open interest), some emerges from the blockchain (transactions/second, transaction fees, blockchain difficulty) and some of it comes from people, showing interest in particular currencies. (Google trends, social networks).

In this report I will dive deeper into the last category, often referred to as sentiment indicators. Sentiment indicators are easy way to understand how a vast majority of people are feeling right now. Arguably, the most popular is Fear and Greed indicator.

It shows the prevailing emotion on the market at the moment. It does that by analysing different factors: Volatility, Market Momentum/Volume, Social Media, Dominance, Google Trends.

Google Trends is incredibly useful for determining interest in cryptocurrency. It appears that search popularity is highly correlated with price action.

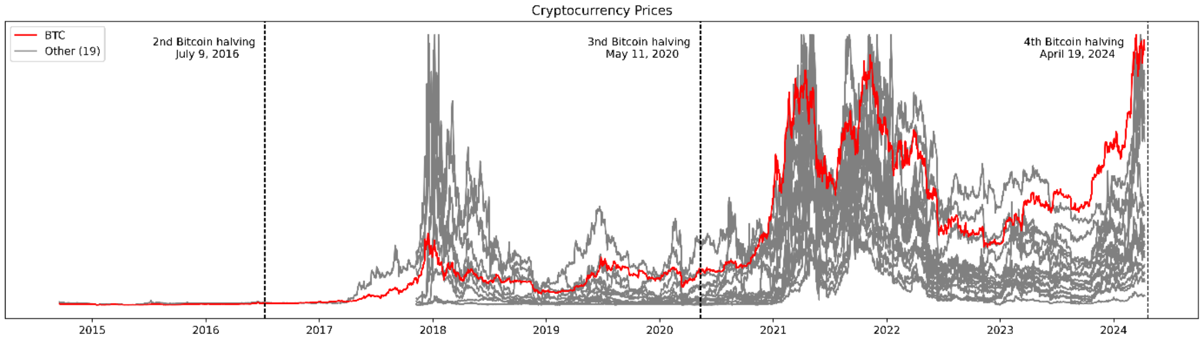


On this picture I've visualized the average price alongside Google Trends statistics for all cryptocurrencies, provided in the dataset (I extended the original trends dataset, grabbing more data from Google Trends). Correlation of 94.84% means that changes in price are closely mirrored by corresponding shifts in search popularity, suggesting a strong link between market sentiment and cryptocurrency price dynamics.

Data Analysis

Exploratory data analysis

Provided dataset contains data about price action, volume and search interest for 20 cryptocurrencies.

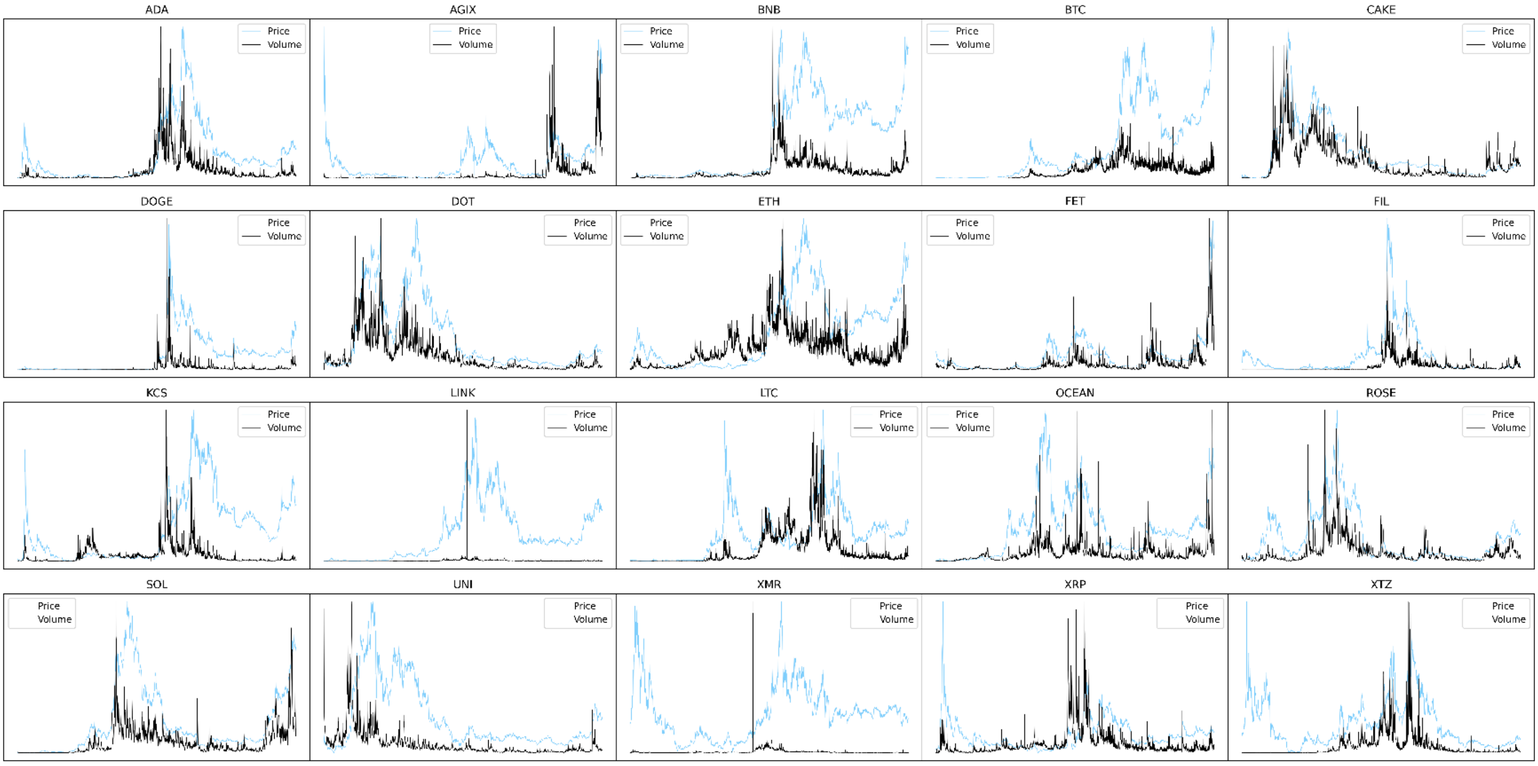


It's possible to see on this picture correlation between Bitcoin price and prices of other cryptocurrencies.

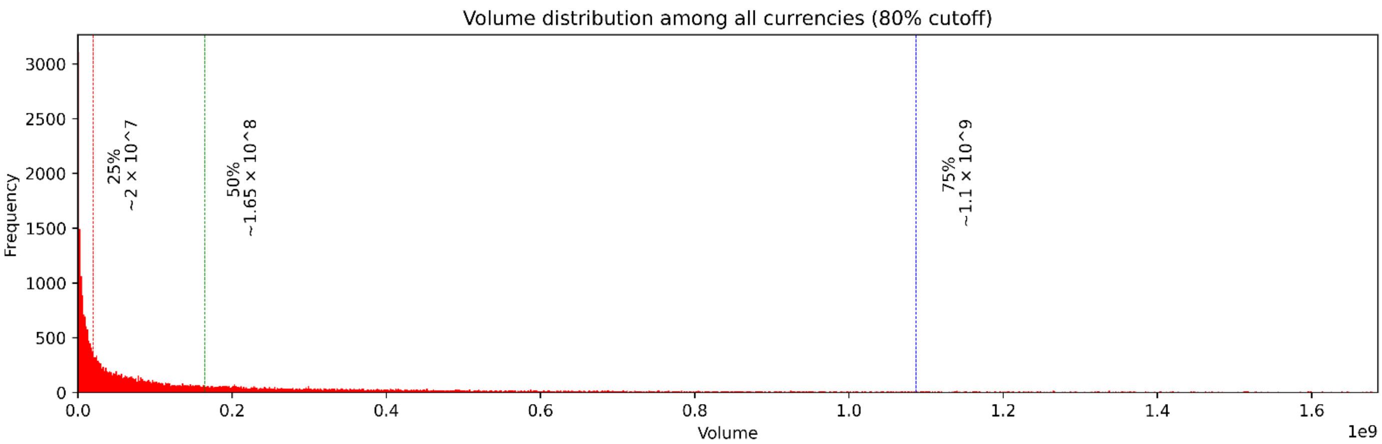
Here's picture with [correlation matrix] for some of them:

	BTC	ETH	BNB	ADA	XRP	LTC	XMR	DOGE	SOL	DOT	AVAX	GOLD	SP500
BTC	1												
ETH	0.89	1											
BNB	0.79	0.81	1										
ADA	0.77	0.81	0.78	1									
XRP	0.59	0.63	0.59	0.71	1								
LTC	0.75	0.78	0.75	0.75	0.58	1							
XMR	0.7	0.68	0.64	0.61	0.46	0.6	1						
DOGE	0.65	0.69	0.68	0.65	0.52	0.64	0.53	1					
SOL	0.76	0.79	0.75	0.79	0.63	0.7	0.58	0.63	1				
DOT	0.78	0.81	0.8	0.85	0.64	0.77	0.63	0.66	0.77	1			
AVAX	0.79	0.84	0.82	0.81	0.64	0.74	0.63	0.68	0.82	0.84	1		
GOLD	0.21	0.2	0.21	0.18	0.09	0.19	0.19	0.11	0.16	0.15	0.17	1	
SP500	**0.53**	0.53	0.47	0.51	0.38	0.46	0.41	0.38	0.42	0.51	0.48	0.19	1

Very often price change is coupled with volume increase due to increased interest and demand.



The picture below illustrates the distribution of daily trading volumes for all currencies, displaying 80% of the data. Cut-off is made to highlight the exponential pattern characterizing this distribution.



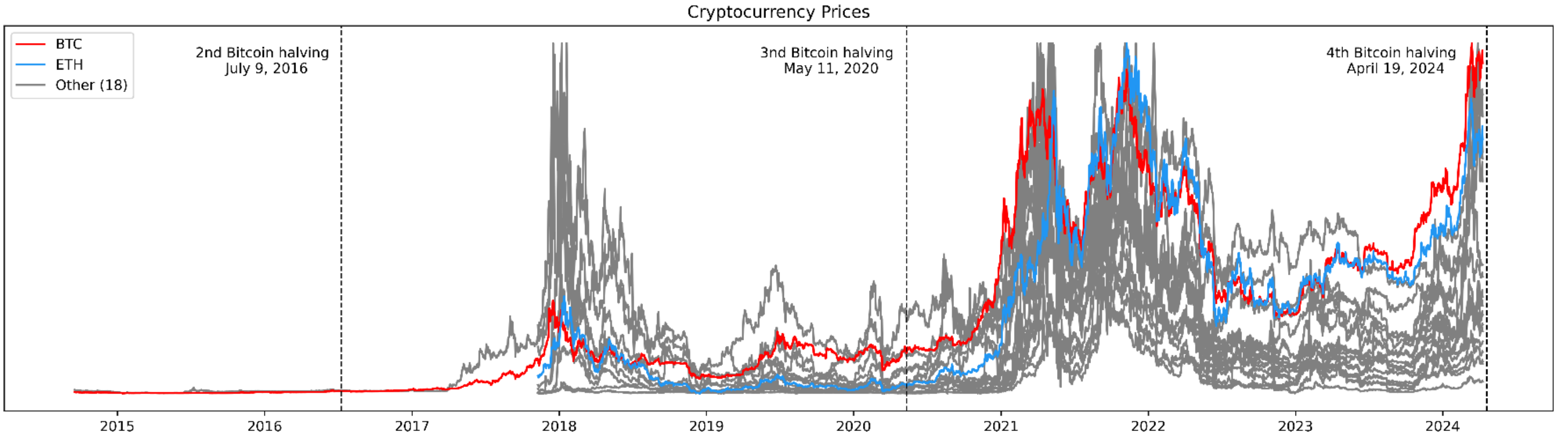
Half of all daily trading volumes is less than 165 million dollars, while the other half is widely spread with the maximum value of 350 billion dollars, recorded on February 26, 2021, for BTC.

Past vs Current Trends Cycles

Historically, all cryptocurrency cycles have been closely tied to Bitcoin halvings. Bitcoin halving is an event that occurs approximately every 4 years where the reward for mining new bitcoins is cut in half, reducing the rate at which new bitcoins enter circulation and potentially impacting the price. As of now, there have been four halving events:

1. November 28, 2012 — Reward down: 50 BTC to 25 BTC
2. July 9, 2016 — Reward down: 25 BTC to 12.5 BTC
3. May 11, 2020 — Reward down: 12.5 BTC to 6.25 BTC
4. April 19, 2024 — Reward down: 6.25 BTC to 3.125 BTC

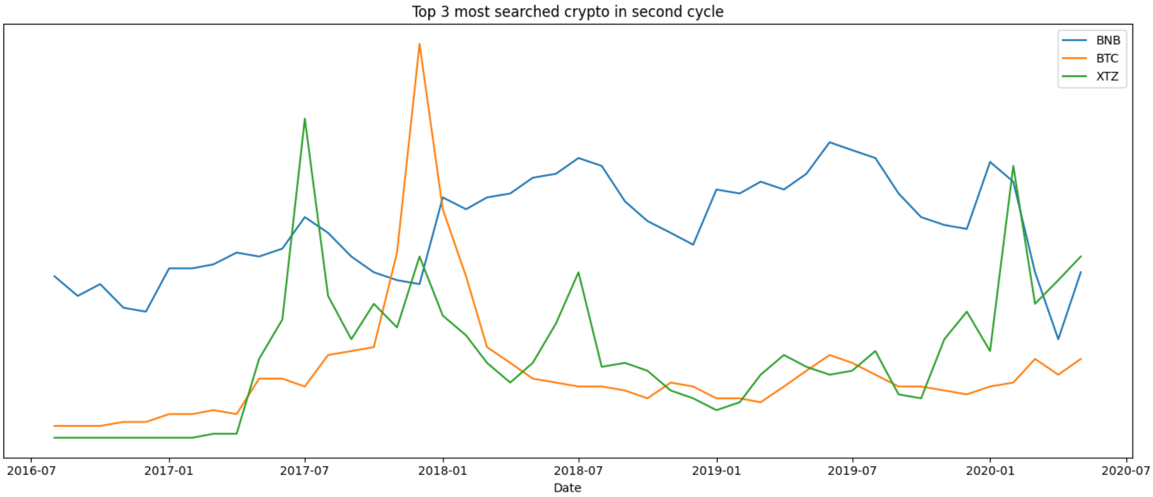
For the purpose of analysis, I have focused on the 2nd and 3rd halving cycles, as they provide more data, and most cryptocurrencies in the dataset were launched during 2nd cycle.



Here are the mean Google Trends values for the top 10 most googled cryptocurrencies during the second cycle:

BNB	XTZ	BTC	XMR	AGIX	XRP	OCEAN	SOL	LINK	LTC
53.61	21.15	17.17	14.96	12.17	10.76	10.43	9.5	9.43	8.43

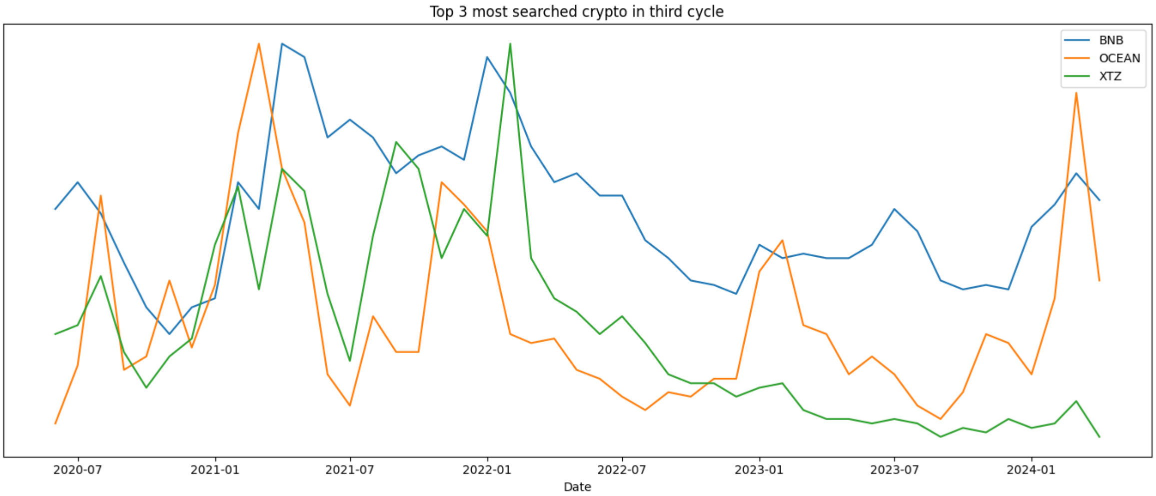
Top 3 most searched crypto in second cycle



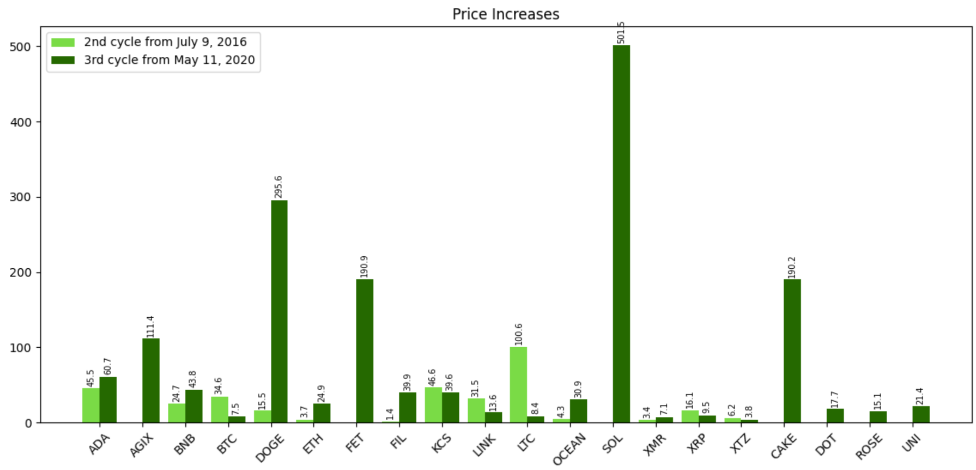
Top 10 most googled cryptocurrencies during the third cycle:

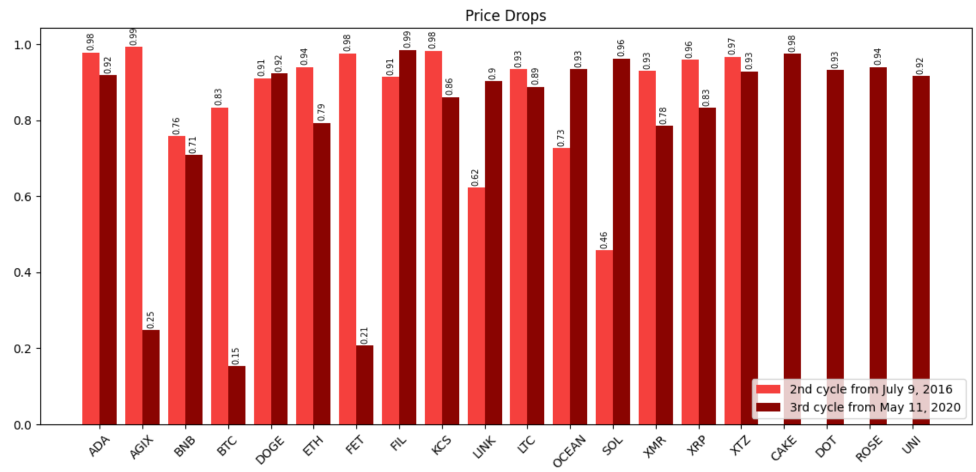
BNB	OCEAN	XTZ	LINK	SOL	DOT	BTC	XRP	UNI	ADA
62.43	38.62	36.02	35.77	33.85	31.4	29.53	28.11	26.77	26.11

Top 3 most searched crypto in third cycle



On the next pictures I will present the fluctuations in cryptocurrency prices during the second and third cycles using the following format: 7.41 indicates that the price increased by a factor of 7.41 from the lowest point, while 0.78 indicates that the cryptocurrency lost 0.78 times its value from the highest point.



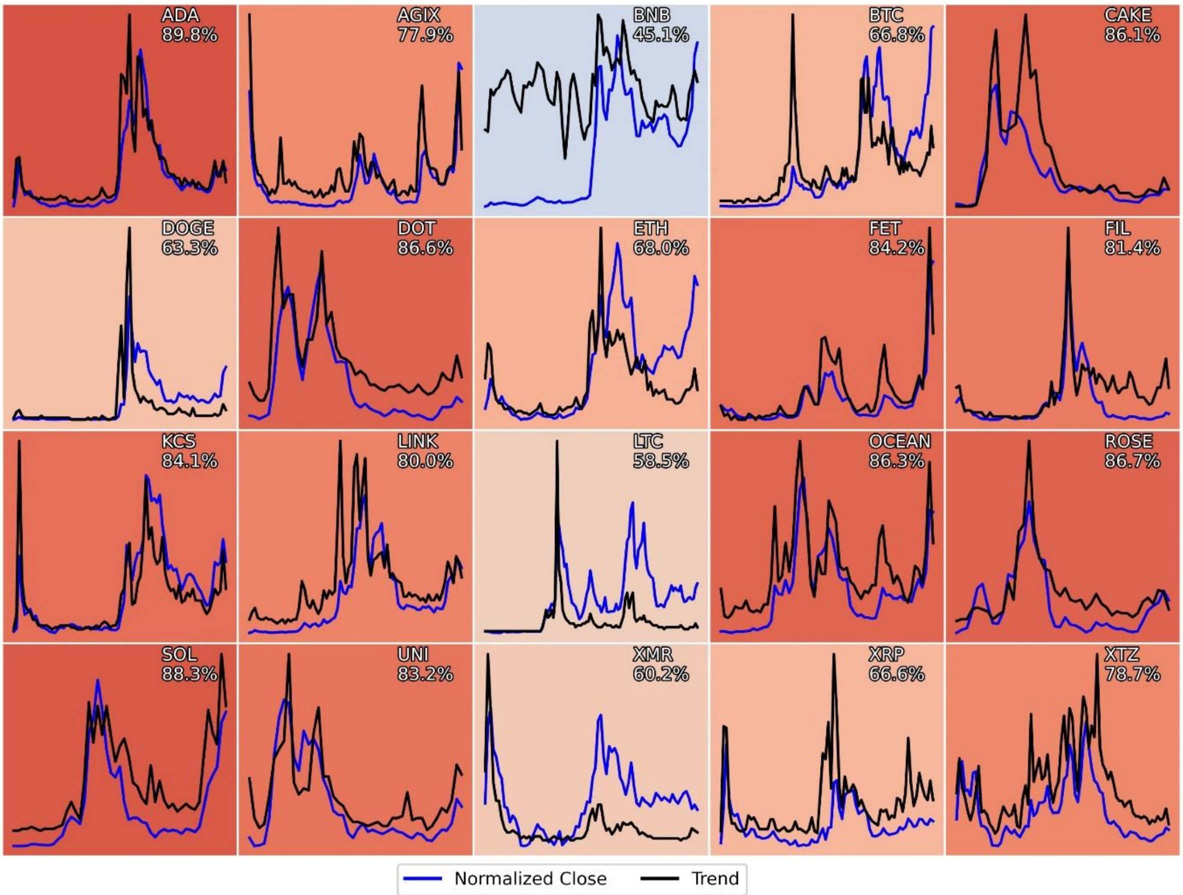


In the second cycle among 20 cryptocurrencies, Litecoin emerges as the growth leader, while Solana takes the lead in the third cycle. The majority of cryptocurrencies experienced a decline of over 90% from their peak values during both cycles.

Correlations

Token price vs Trends

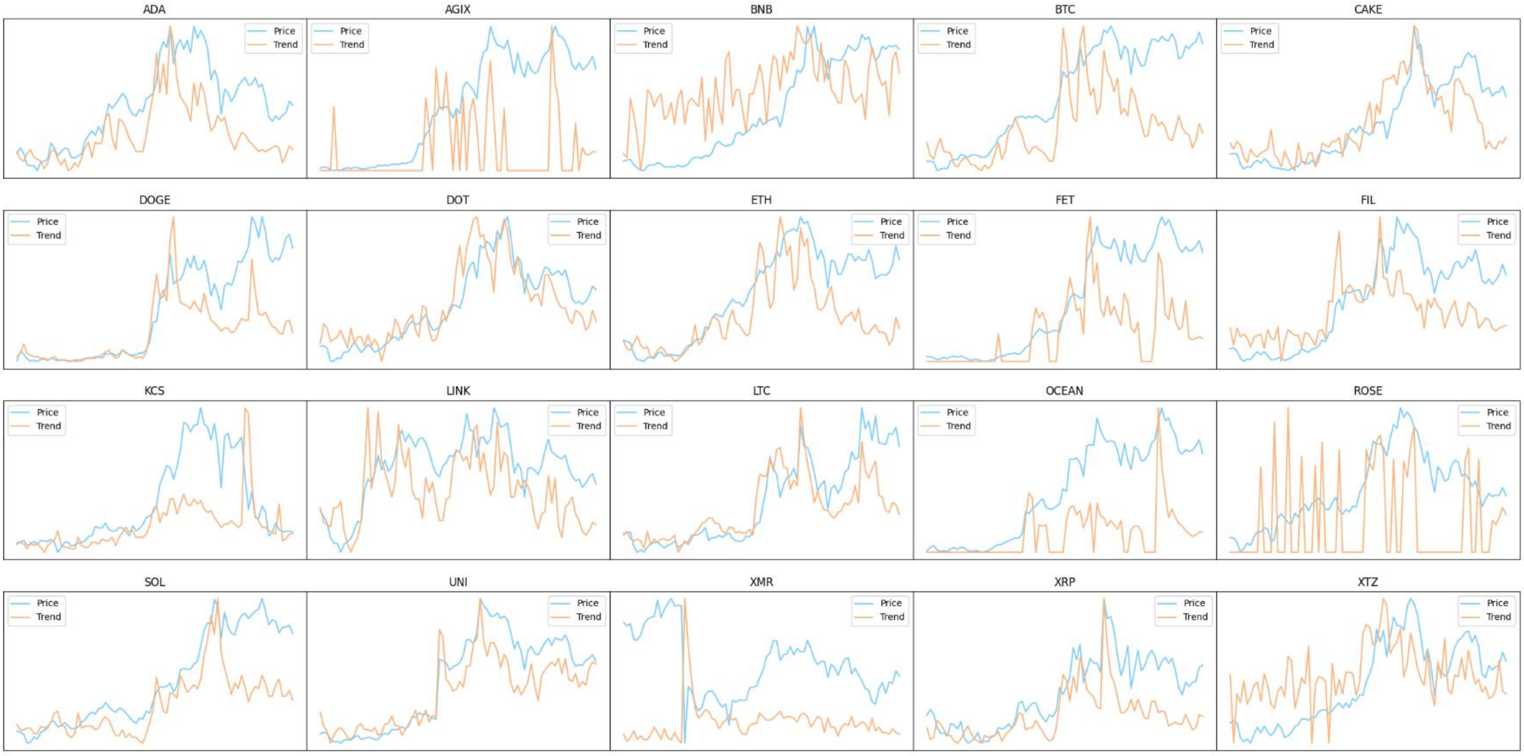
As demonstrated in the introduction section, there exists a strong correlation between token prices and web search popularity. On average, for all cryptocurrencies, this correlation stands at an impressive 94%, indicating a remarkably strong relationship between these two factors.



The heatmap above displays the correlation values for every cryptocurrency in the dataset. The prices have been normalized to ensure that both trend values and prices are on the same scale.

It's ambiguous whether price fluctuations drive trend alterations or if trends influence price movements. In the next section, we'll explore this relationship to uncover the leading factor.

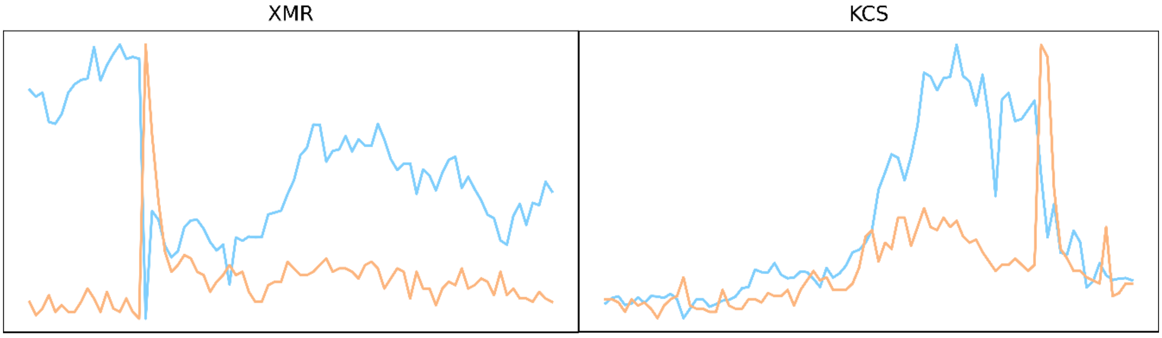
Time Lag



The picture above shows the prices and trends for each day starting from January 19, 2024. Sometimes, the price changes first, and other times, the trend changes earlier. Occasionally, both change on the same day. I think there might be one thing causing both the price and trend changes. My guess is that it's the amount of attention a currency gets on social media and in the news.

Additionally, there's another intriguing observation regarding XMR and KCS.

Token Patterns



On February 6th, Monero's price fell by about 40%, but at the same time, online searches for it shot up from 18 to 100. This happened because Binance removed Monero from exchange.

On March 26th, KCS, dropped by 15% because KuCoin was accused of breaking money-transferring rules by U.S. prosecutors. Meanwhile, online searches for KCS jumped from 30 to 100.

These incidents highlight that there isn't always a linear dependency between online searches and the price. While online search trends can provide valuable insights into market sentiment, they do not always directly translate into price movements. External events, such as regulatory actions or exchange delistings, can exert significant downward pressure on prices despite increased public interest.

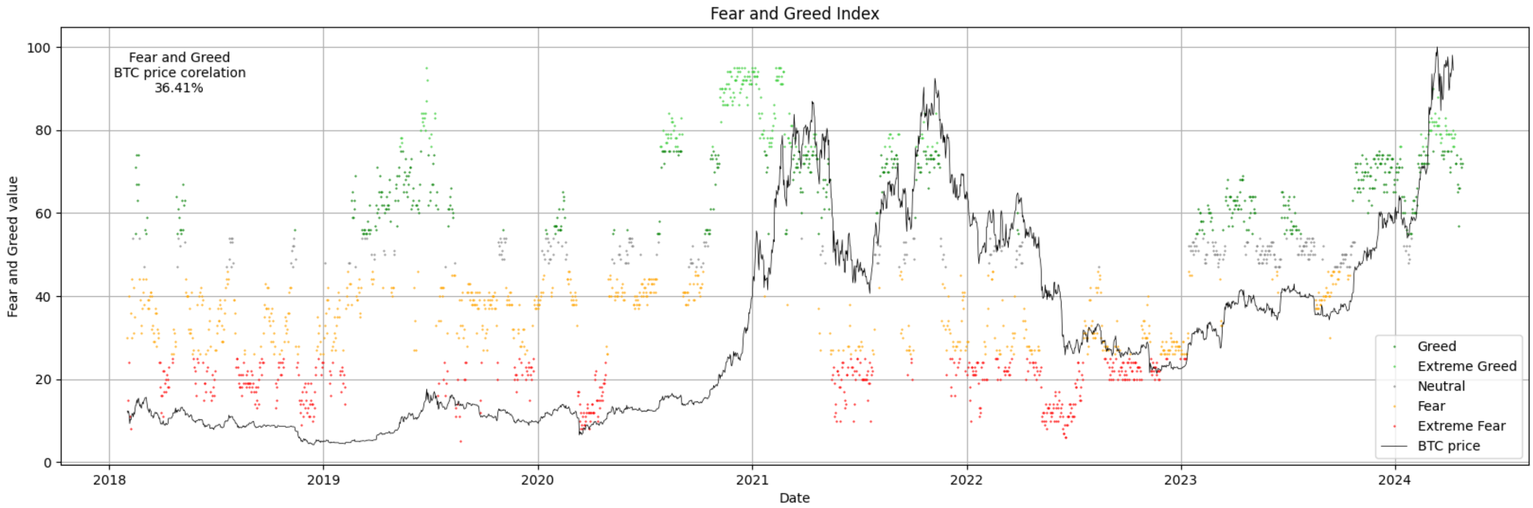
However, these occurrences are relatively uncommon, and their influence on the cryptocurrency market is usually limited—unless such an event involves Bitcoin. This is because most cryptocurrencies are strongly correlated with Bitcoin's price movements.

Additional Data Sources

I used three data sources to enrich my analysis:

- Fear and Greed index - [link]
- Tweets per day - [link]
- News sentiment analysis from my previous project - [link]

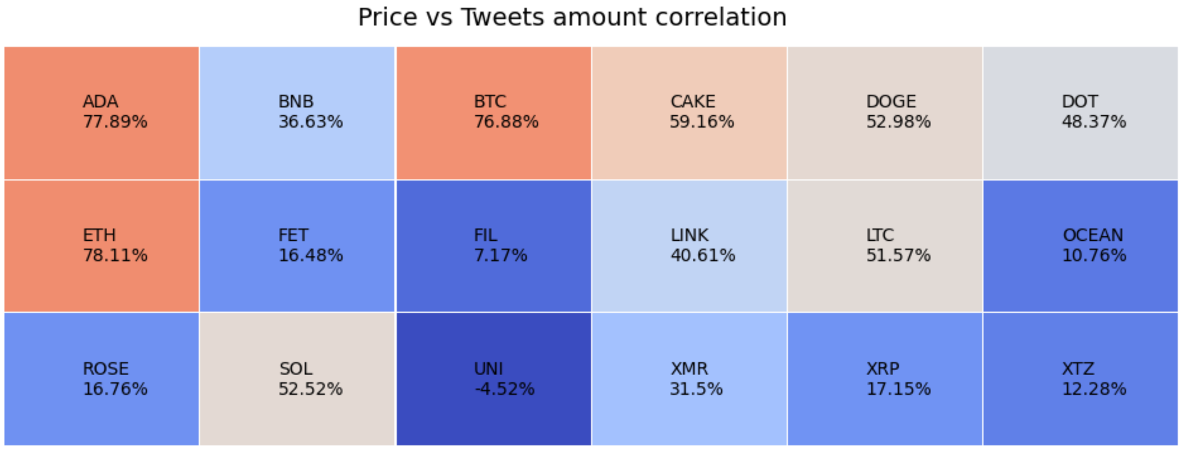
Fear and Greed index



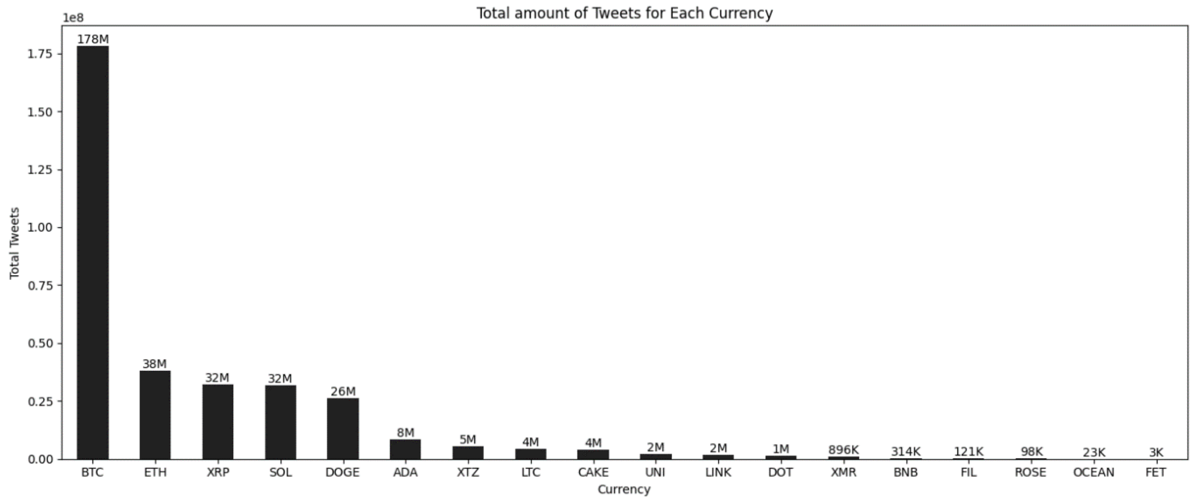
The Fear and Greed index, while not closely tied to Bitcoin's price, is still a very handy tool. It helps to find the lowest and highest points in the market and tells us whether people are feeling optimistic (bullish) or pessimistic (bearish) about it.

Tweets per day

Twitter updated its API costs, making it more challenging to collect data. As a result, the latest available date on the source site is March 14, 2023. Therefore, some of the conclusions drawn from the analysis may be outdated and incorrect at the present time.



The number of tweets per day doesn't show a strong correlation with the price, except for BTC, ADA, or ETH. This suggests that for these specific currencies, social media hype tends to influence price movements.



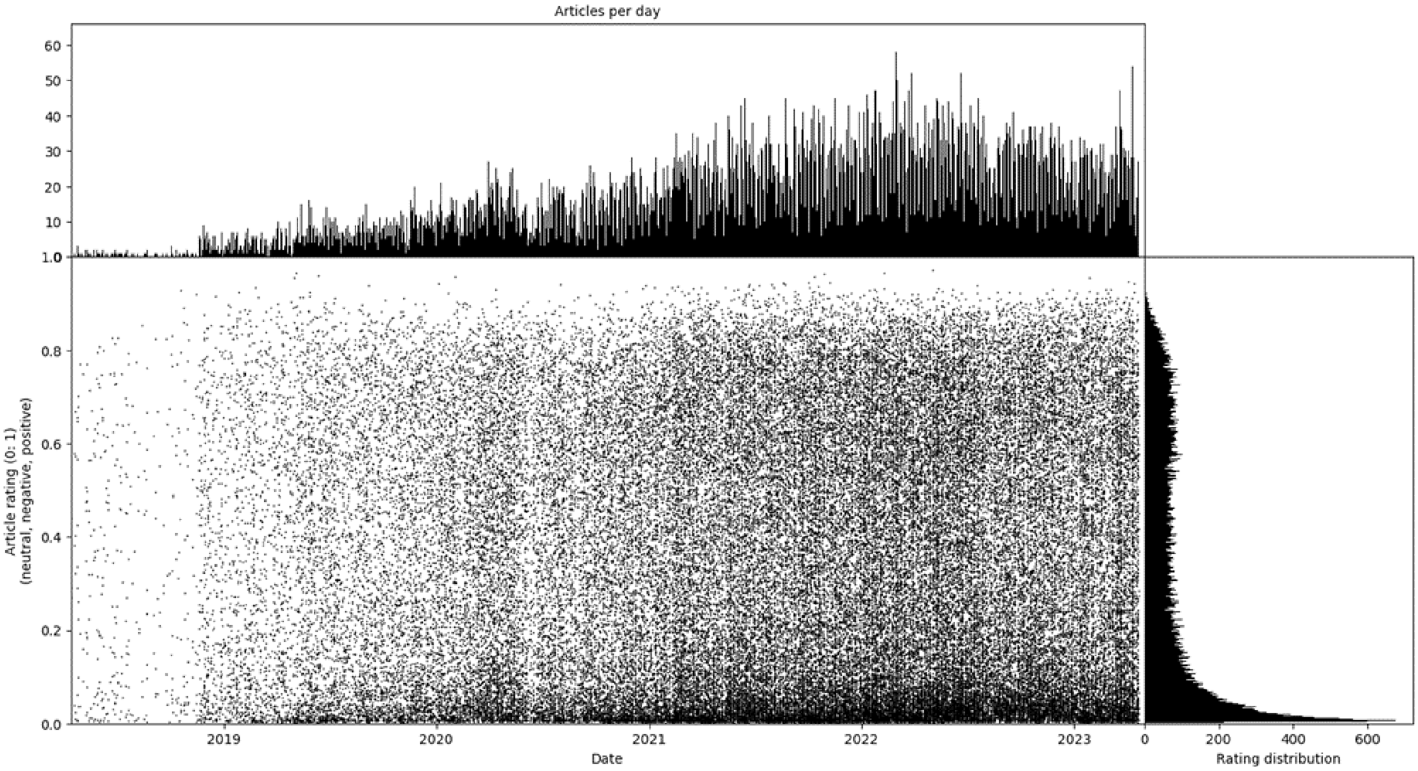
Even though BTC and ETH are the most talked-about cryptocurrencies on Twitter, ADA only ranks sixth with a total of 8 million tweets. This indicates that the quantity of tweets alone is not the sole determinant of influencing price movements.

News sentiment analysis

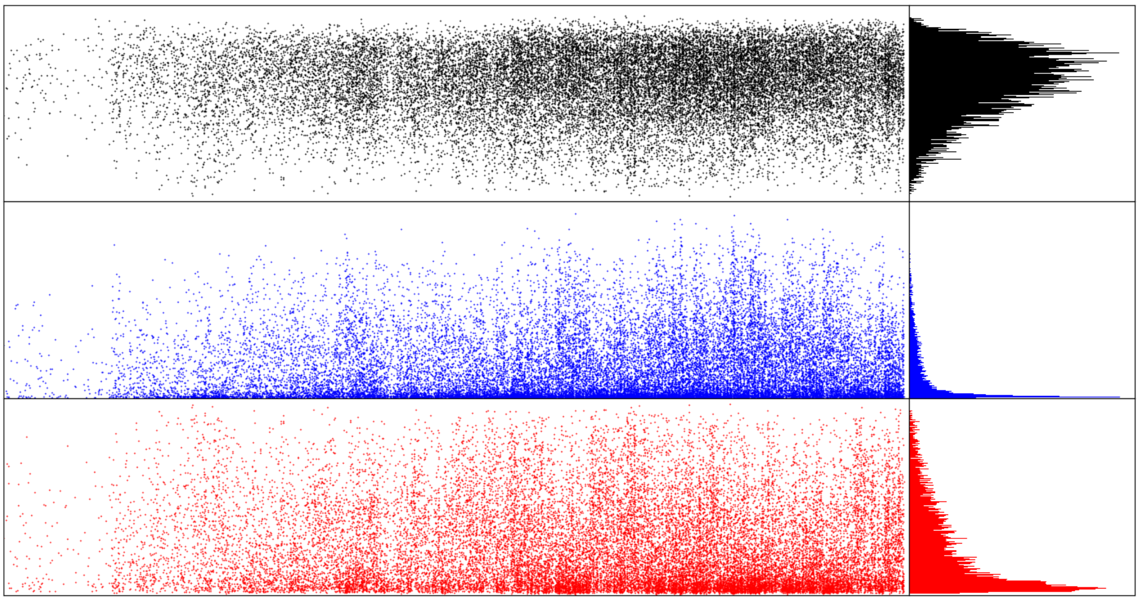
This project was made 8 months ago, so it's possible that some of the conclusions may be outdated and incorrect at the present time.

Here's a brief overview of the project creation process:

1. Gathered over 28,000 news articles from a crypto-news aggregator.
2. Preprocessed the data and filtered out any invalid entries.
3. Used two neural networks to analyze the sentiment of the articles.
4. Conducted a comparative analysis between the two neural networks.
5. Wrote a report summarizing the findings, along with relevant graphs and visuals.



On the picture above there is a rating distribution for BTC related news.



Black – neutral rating distribution, blue – negative rating distribution, red – positive rating distribution.

Neutral rating is assigned more frequently compared to positive and negative ratings, and positive rating is more frequent than negative rating.

It seems that there are more positively rated news articles when the price of BTC is rising, and more negatively rated news when the price is falling. This observation could serve as another sentiment indicator, supporting the theory that price movements are influenced by popular news media.

Machine Learning Model

I built three different ML models to evaluate their performance and determine which one performs better.:

1. Linear regression
2. Random forest
3. Feedforward neural network

For training purposes I've picked ADA because it has highest Trend/Volume and Trend/Price correlation. The main challenge for training is the small dataset, which has length of only 262 samples.

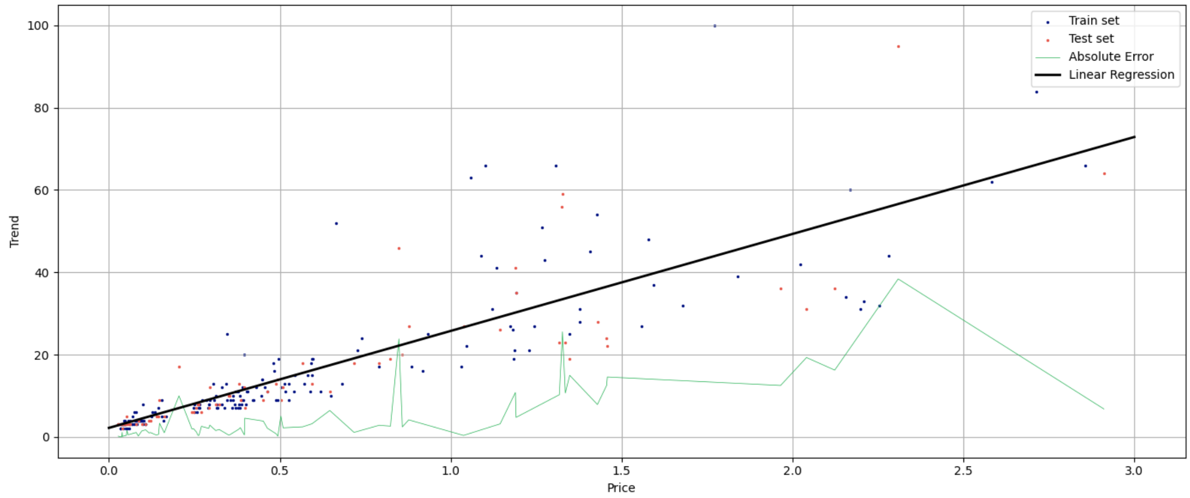
The data will be divided into two categories: training data (70%) and test data (30%).

Initially, I will attempt to predict values without feature engineering, using the original dataset. Then, I will create three features: Price Change, Volume Change, and Trend Change, and compare the performance to determine the most effective approach.

Without feature engineering

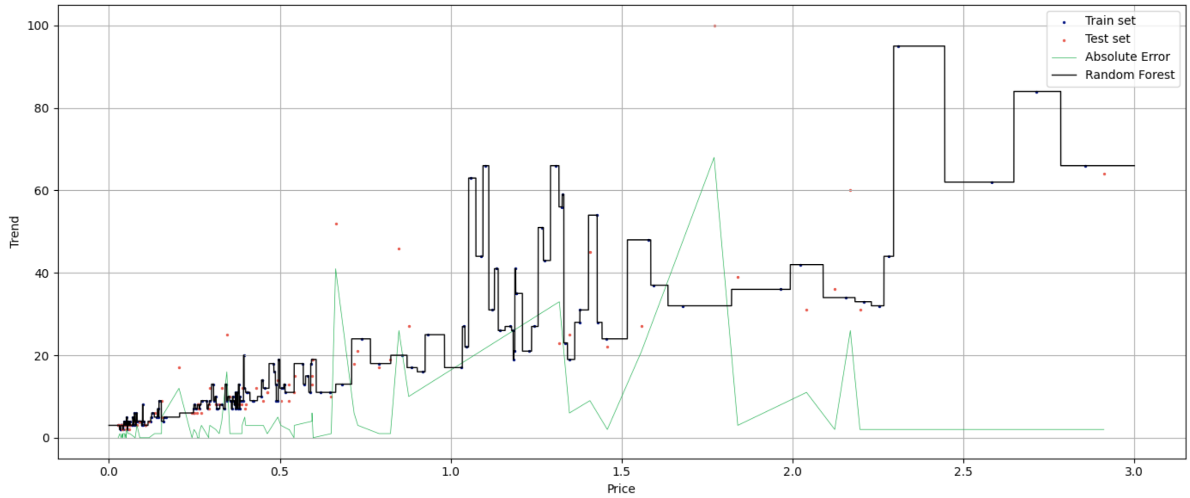
	Date	Trend	Currency	Open	High	Low	Close	Volume
0	2019-04-07	4	ADA	0.090161	0.091286	0.088536	0.090484	100804121
1	2019-04-14	4	ADA	0.083731	0.085479	0.081372	0.084643	82514758
2	2019-04-21	3	ADA	0.076556	0.077978	0.071989	0.074106	89301952
3	2019-04-28	3	ADA	0.070189	0.070646	0.068020	0.068591	42979831
4	2019-05-05	3	ADA	0.066907	0.067966	0.065709	0.066146	49098473

Linear Regression



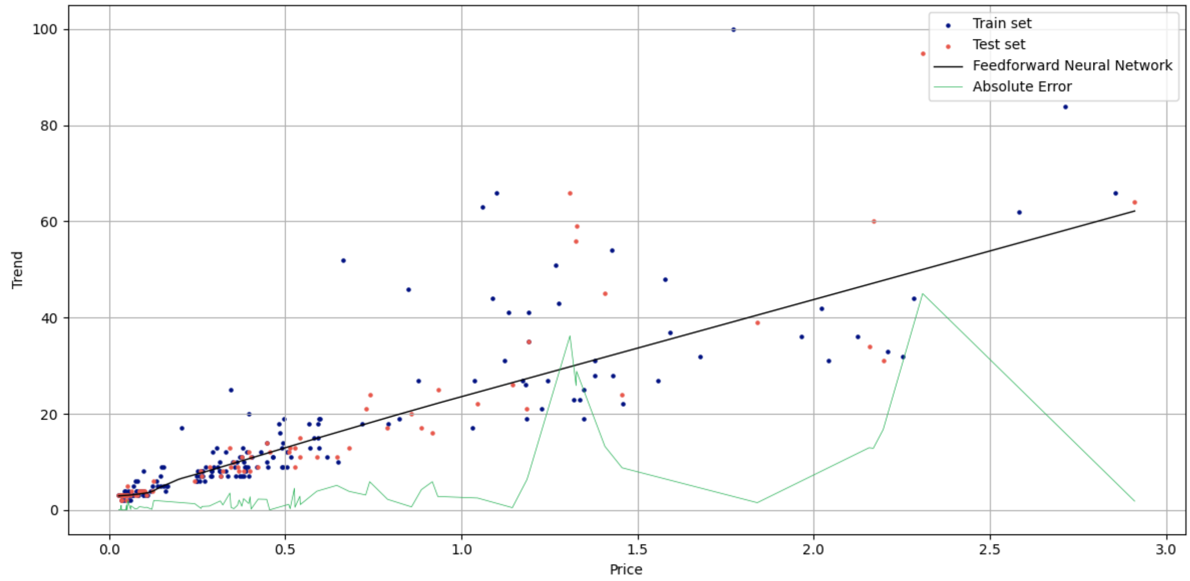
In the image above, you will see test and train sets distribution. Linear regression is denoted by a wide, black line. The linear regression is a function that attempts to estimate a result by selecting a line that minimizes the error on the training set. The narrow green line represents the absolute error, indicating how far every test point is from the forecast line.

Random Forest



Random Forest is a method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees. The narrow green line represents the absolute error, indicating the deviation of each test point from the predicted values by the Random Forest model.

Feedforward neural network



Feedforward Neural Network consists of multiple layers of nodes, where information moves in only one direction—from the input nodes, through the hidden nodes, to the output nodes. The narrow green line represents the absolute error, showing the deviation of each test point from the predicted values by the Feedforward Neural Network model.

The model consists of three layers, each with 500 neurons, utilizing the ReLU activation function. Mean squared error is employed as the loss function, with the Adam optimizer used for training. On average, the model is trained for 500 epochs.

Results

	Linear Regression		Random Forest		Feedforward NN	
	No shuffle	Shuffle	No shuffle	Shuffle	No shuffle	Shuffle
Price	1.72	3.85	14.68	11.44	1.59	3.66
Volume	1.62	6.21	17.68	16.44	1.47	5.04
Price +Volume	1.56	3.19	12.36	8.6	1.52	2.6

Mean absolute error (Goggle Trends)

Here's a comparison of the regression algorithms measured in mean absolute error (MAE), where MAE = 2 indicates that predictions on average are ±2 points off from the correct answer. "No shuffle" indicates that training and test data were located in chronological order, while "shuffle" indicates that the training set and test set were selected randomly.

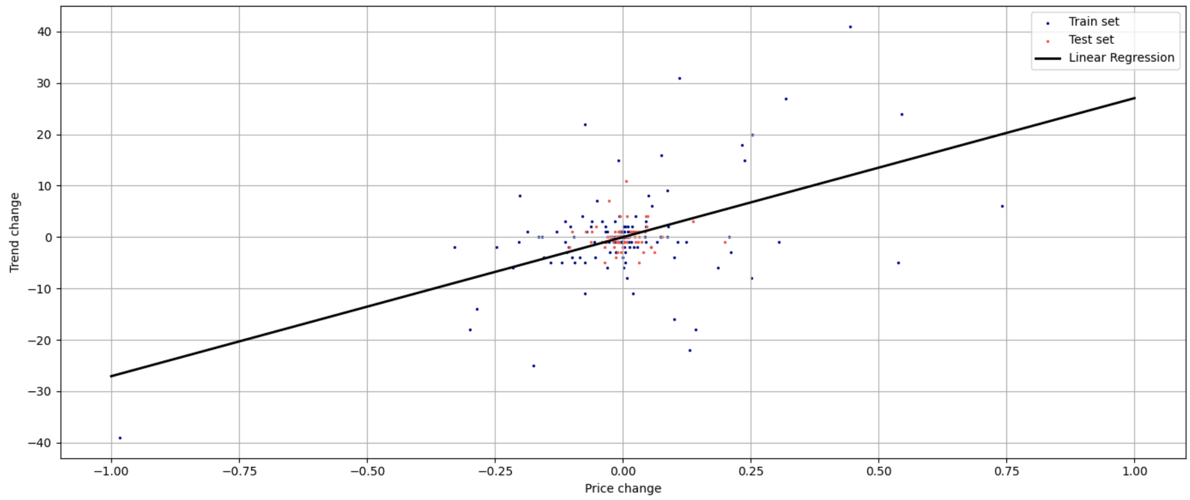
The table suggests that, on average, the feedforward neural network performed slightly better than linear regression. However, this improvement is quite small and mainly stems from the feedforward neural network's distinction in a specific area (Price < 0.25), where it achieved slightly better results.

With feature engineering

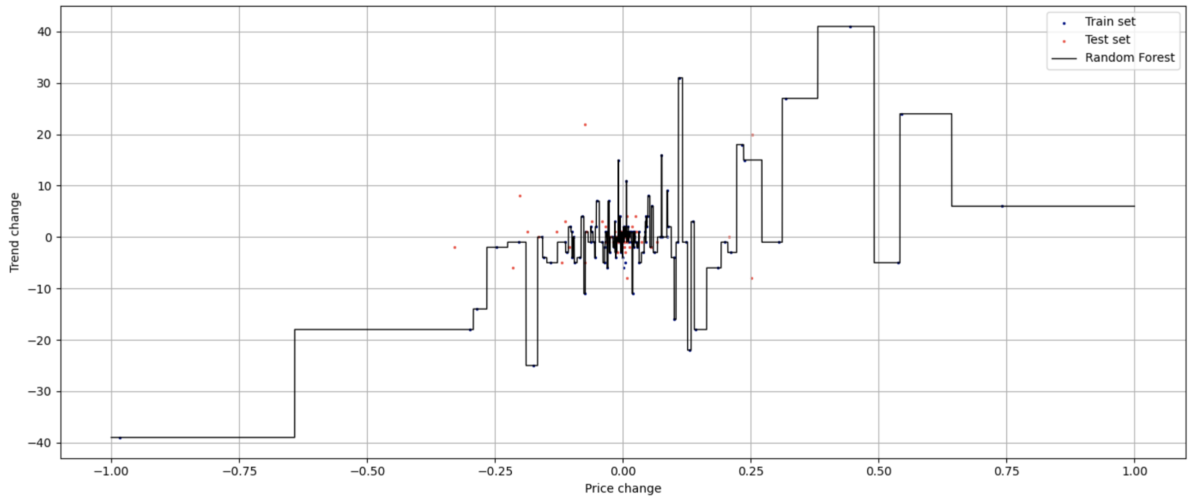
	Date	Price_Change	Volume_Change	Trend_Change
1	2019-04-14	-0.005841	-18289363.0	0.0
2	2019-04-21	-0.010537	6787194.0	-1.0
3	2019-04-28	-0.005515	-46322121.0	0.0
4	2019-05-05	-0.002445	6118642.0	0.0
5	2019-05-12	0.003940	91925316.0	2.0

Predicting trend changes based on price and volume changes might indeed be simpler.

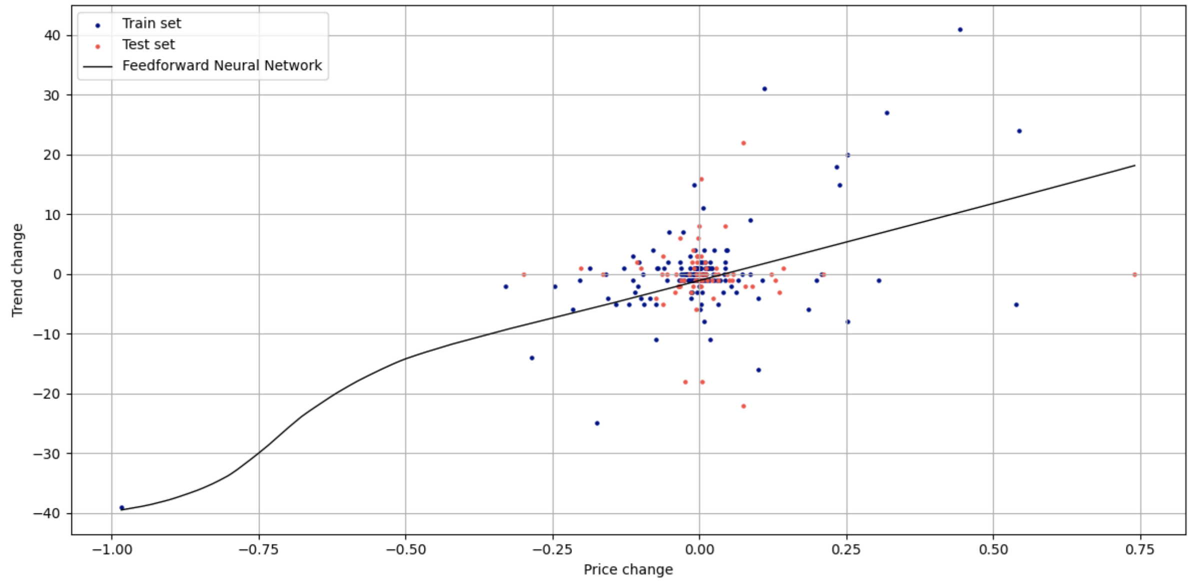
Linear Regression



Random Forest



Feedforward neural network

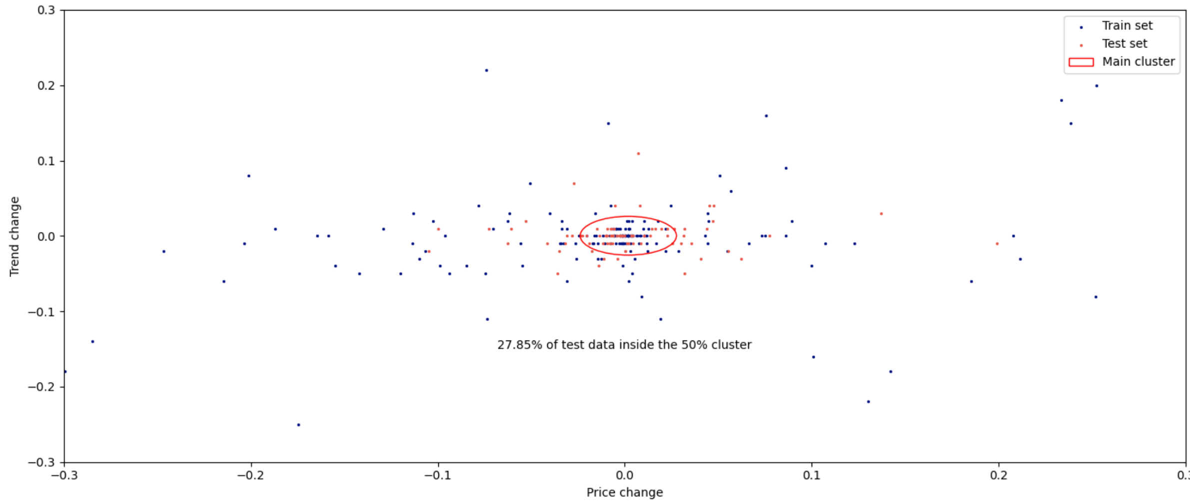


Results

	Linear Regression		Random Forest		Feedforward NN	
	No shuffle	Shuffle	No shuffle	Shuffle	No shuffle	Shuffle
Price change	1.58	3.07	2.9	4.89	1.5	2.83
Volume change	1.39	2.91	2.91	4.32	1.37	2.87
Price + Volume	1.5	2.91	2.15	3.84	1.41	2.54

Mean absolute error (Goggle Trends)

The Feedforward Neural Network comes out on top again, but only slightly ahead of Linear Regression. Random Forest shows a notable improvement, largely due to higher data density. The lower mean absolute error (MAE) without shuffling can be attributed to the fact that 30% of the latest data is located close to the main cluster.



The proximity of the test data to the mean value naturally results in a lower mean absolute error

Predicting trend changes based on price and volume changes proves to be a more accurate approach compared to predicting trends based on absolute values. However, the difference between the two methods is not as significant as anticipated.

Even without prior Google Trends data, it's possible to predict the trend value for the next week with a precision of ±3 %, which I find satisfactory.

To make predictions more accurate, I suggest gathering more data and using a type of neural network called a Recurrent Neural Network (RNN). RNNs are good at understanding patterns in data that comes in a sequence. They can remember past information to make better predictions about what might happen next.

Conclusion

In conclusion, here are the key findings:

- Google Trends ratings show a strong correlation with cryptocurrency prices.
- Cryptocurrency prices are closely tied to the price of Bitcoin (BTC).
- Cryptocurrency cycles often see significant drops, with many cryptos losing more than 80% from their cycle highs.
- There is no significant time lag between price movements and changes in Google Trends ratings.
- Google Trends ratings do not always follow price movements; in some cases, it's the opposite, especially in rare negative situations like delisting from an exchange.
- The number of tweets about a cryptocurrency does not consistently correlate with price movements and may not accurately reflect its popularity.
- The most precise models for predicting Google Trends ratings are Linear Regression and Feedforward Neural Networks, which yield comparable results.
- Feature engineering can help improve the precision of predictions.

Data sources:

- Google trends - https://trends.google.com/trends/
- Fear and Greed index - https://alternative.me/crypto/fear-and-greed-index/
- Correlation matrix - https://www.blockchaincenter.net/en/crypto-correlation-tool/
- Tweets per day - https://bitinfocharts.com/comparison/tweets-btc.html#alltime
- News sentiment analysis - https://github.com/doxlix/BTC_news_santiment