

DATA CLEANING & PRE-PROCESSING

SUMMARY REPORT

CRYPTO CURRENCIES AROUND THE GLOBE



Submitted By	PGID
Akash Topdar	12120033
Lalit Kumar Sharma	12120091
Nikhileshwar AVS	12120088
Prajukta Pradhan	12120039
Ramteja Bandaru	12120048

Table of Contents

Executive Summary	3
The chosen domain and Seed sources.....	4
The structured and unstructured sources from open domain/ internal sources	5
Download/ crawl/ collect data from all the sources	6
Convert data from original sources (Webpages) to structured data fields.....	9
Data cleaning/pre-processing as needed	11
Observations/ Insights and Analysis on the data collected.....	14
Strategy to enhance the data with crowd sourcing methods	20
References and Sources used for this Assignment	23

Executive Summary

Problem Statement

To create a dataset using an end-to-end data collection and pre-processing pipeline for multiple crypto currencies that are widely traded across different platforms and use the dataset to derive meaningful insights

Proposed Solution

- Explored multiple websites (CoinGecko, Coincap, CoinMarketCap) to understand the available data that can be scrapped and can be used for the different use cases to run analytics
- Generally, the single seed data sources are readily available for Crypto data however all the sources are not mineable. So, it is necessary to explore each of these sources and understand which websites are mineable and which are not
- We majorly relied on the Web Crawling technique to get the desired minimal data and then cleaned it using different techniques so that it can be enhanced for further analysis
- With the help of the scrapped data, we added additional columns (7D_SMA, 20D_SMA, 50D_SMA, 100D_SMA , 100D_SMA, 24hr % Change, 7d % Change, 30d % Change, Close_INR, Close_EUR, Close_JPY, Close_GBP, Close_CAD) to get more clarity on how the data can be used to derive different insights and run predictive analysis

Challenges Faced/Brief Understanding of the Problem

- Majority of the crypto sites were non mineable
- All the key columns in the table were not minable, only few basic columns like Date, Open and Close data could be mined
- Only limited time (30 days) data was scrapable
- Had to calibrate the scrapper in a way such that it can surpass the data pulls for more than 30 days
- Some of the websites blocked our IP addresses
- Without APIs it was difficult to scrap the data for more than 1 coin for more than 30 days. If we had to mine for more coins and duration, we were prompted to purchase the APIs
- The real time conversion of USD value to other currencies was not possible because the data columns were not in a consumable format and had to be pre-processed for conversion

The chosen domain and Seed sources

DOMAIN OVERVIEW:

We as a group wanted to work on a domain where we can further implement the learnings of the AMPBA program. Equity market or crypto currency market is always a domain which requires a lot of data to make informed decisions. In the recent past, with an exponential growth in crypto market and launch of new crypto currency trading software in the market, the investors are always keen to get access to the historical data set which will enable them to enhance, diversify and grow their portfolio based on analysing different trends and patterns from the crypto data.

With the impact of Covid globally in 2020, the crypto market witnessed the growth of around 10% in 2020, year after year, compared to 2017-2019 growth. According to one of the reports published by Fortune Business Insights, the crypto market is expected to reach USD 1,902 Mn in 2028. Few developing countries even started acknowledging crypto currencies. Recently, the news of Russia deliberating on the use of crypto currencies in Internal Trade as payment method was also floating.

Popularity of crypto market started growing in 2020 and the year witnessed the increase in the new & credible players. Most of the new investors are young investors. From early 2020 to September 2021, saw almost 10-fold in increase in new investors. Based on the latest report published by crypto.com, we will witness a total of 1Bn crypto users by the end of December 2022.

After the rapid push for the crypto market in 2020, a lot of new crypto currencies also came in the market over the time and currently we have more than **18500** crypto currencies in existence. We have even witnessed a lot of new currencies coming into the market and go out of circulation as well.

Above mentioned were few of the key motivating factors that inspired us as a group to choose this domain(Crypto Currency market). We did a study on top 100 currencies now based on various legit sources & collected the data for last 6-7 years available. We wanted to come up with a solution where we can use different Data processing techniques to collect the data and make informed decision using the same.

SEED SOURCES:

For our chosen domain, we have many sources to fetch the data like **yahoo finance, Coingecko, CoinMarketCap, wazirx** and many others. After discussing with the active investors/traders in the crypto market, we went ahead with yahoo finance and Coingecko as our seed source.

Both yahoo finance and Coingecko had the same data for all the coins, hence we went ahead with data fetched from Coingecko as it is reliable source for crypto market compared to yahoo finance.

For our domain, data available was both fundamental and historical. So, we chosen the historical data for our analysis part.

Twitter is our unstructured data source from where we fetched the 50,000 tweets each for bitcoin and Ethereum.

When it comes to sentiment analysis then getting tweets from the twitter serves best. It comprises to a large user base as it is a part of crowdsourcing to make our dataset better. Analysing the tweets for sentiment analysis into 'Neutral', 'Positive' or 'Negative' is quite comfortable.

The structured and unstructured sources from open domain/ internal sources

The structured and unstructured sources from open domain/internal sources:

As we wanted to scrap the historical price data for the top 100 crypto currencies, almost all the sources in our domain had structured data.

Structured sources:

- Coingecko is our structured data source from where we scraped the historical data of 6 years.
- Reason behind choosing Coingecko is it being one of the reliable sources for crypto market.
- We were able to scrape the historical data from this platform unlike other sources where data was not minable for more than 30 days.
- We were able to scrap the data without any API from the platform.
- The platform had the crypto currencies price data from their inception.

Unstructured sources:

- We also fetched the tweet data from the twitter to do the sentiment analysis related to the top 2 crypto currencies
- Twitter is our unstructured data source from where we fetched the 50000 tweets each for bitcoin and Ethereum
- When it comes to sentiment analysis then getting tweets from the twitter serves best
- It comprises to large user base

Analyse the tweets for sentiment analysis into 'Neutral', 'Positive' or 'Negative' is quite comfortable.

Download/ crawl/ collect data from all the sources

STRUCTURED SOURCES:

Option 1 : Web Interface of Coin Market Cap which states that the data is not minable (highlighted in red)

We initiated the data collection process through <https://coinmarketcap.com/> which is one of the single sources of truth for the crypto data. However, while scraping the data, we hit the roadblock where we observed that the required data is not mineable and requires paid APIs to pull the data from backend html pages.

Screen Capture of Coin Market Cap : <https://coinmarketcap.com/historical/20220522/>

Historical Snapshot - 22 May 2022									
Market Cap:		Price:		Volume (24h):					
\$1 Billion+		\$100+		\$10 Million+					
USD						Previous Week View All			
Rank	Name	Symbol	Market Cap	Price	Circulating Supply	Volume (24h)	% 1h	% 24h	% 7d
1	Bitcoin	BTC	\$577,568,895,835.72	\$30,323.72	19,046,768 BTC	\$21,631,532,269.67	0.18%	3.03%	-3.13%
2	Ethereum	ETH	\$246,986,847,793.10	\$2,043.17	120,884,132 ETH	\$10,941,123,403.41	0.38%	3.48%	-4.78%
5	BNB	BNB	\$52,184,792,531.31	\$319.61	163,276,975 BNB *	\$1,604,297,405.63	0.34%	2.29%	2.44%
13	Wrapped Bitcoin	WBTC	\$8,398,827,737.60	\$30,364.20	276,603 WBTC *	\$164,247,022.27	0.20%	3.21%	-2.86%
24	Bitcoin Cash	BCH	\$3,793,751,185.91	\$198.93	19,070,475 BCH	\$2,711,327,935.56	-0.38%	3.47%	-7.27%
26	Monero	XMR	\$3,392,226,625.44	\$187.16	18,124,694 XMR	\$128,816,309.83	0.20%	6.35%	9.02%
42	Zcash	ZEC	\$1,547,109,678.34	\$107.13	14,441,281 ZEC	\$118,265,214.63	-0.15%	2.39%	-4.48%
44	Maker	MKR	\$1,403,744,453.34	\$1,435.86	977,631 MKR *	\$103,221,852.78	0.90%	2.65%	-8.39%
77	PAX Gold	PAXG	\$615,681,948.65	\$1,852.49	332,355 PAXG *	\$17,685,376.90	-0.03%	-0.07%	2.28%
110	yearn.finance	YFI	\$348,004,632.44	\$9,498.53	36,638 YFI *	\$54,775,674.53	0.43%	2.33%	-10.01%
146	Illuvium	ILV	\$196,804,038.00	\$302.37	650,861 ILV *	\$13,263,540.96	0.49%	3.29%	-6.53%
<div> <div>* Not Mineable</div> <div>Load More</div> <div> Previous Week View All </div> </div>									

Option 2 : Web Interface of Coin Base Cap which states that the platform will let the users to only Trade (highlighted in red)

We initiated the data collection process through <https://coinbase.com/> which is also one of the single sources of truth for the crypto data and widely used in the trading community. However, while scraping the data, we hit the roadblock where we observed that the platform is majorly designed for Trading and can only provide rolling 30 days of trading data for limited number of coins. The APIs did not let us pass the backend code to fetch the desired data set

Screen Capture of Coin Base : <https://coinbase.com/trade>

Name	Price	Change	Market cap	Watch
Bitcoin BTC	₹2,247,259.96	-0.23%	₹42.8T	Buy
Ethereum ETH	₹138,221.66	-1.21%	₹16.8T	Buy
Ethereum 2 ETH2	₹138,221.66	-1.59%	₹16.7T	
Tether USDT	₹77.48	+0.00%	₹5.7T	
USD Coin USDC	₹77.57		₹4.1T	Buy
BNB BNB	₹23,212.29	-2.37%	₹3.8T	
XRP XRP	₹30.29	+1.15%	₹1.5T	
Binance USD BUSD	₹77.65	-0.10%	₹1.4T	

Identity Verification Required
You're almost ready to trade. Please verify your personal information
[Verify your ID](#)

New on Coinbase
The Sandbox
Added 18 hours ago

Option 3 : Web Interface of Coin Gecko which was minable but had limited columns (highlighted in green)

After more research across different blogs and forums we came across Coin Gecko <https://coingecko.com/> which is widely used among the traders and is the single sources of truth for the crypto data. However, with limited number of mineable columns we had to work on enhancing this dataset for analytical purposes

Screen Capture of Coin Gecko : https://www.coingecko.com/en/coins/bitcoin/historical_data#panel

Overview Markets Converter **Historical Data** Bitcoin Halving Countdown

Bitcoin USD (Historical Data)

7d 14d 30d 60d 90d YTD 1y 2022-04-27 - 2022-05-27

Date	Market Cap	Volume	Open	Close
2022-05-27	\$559,066,076,344	\$31,972,044,692	\$29,347	N/A
2022-05-26	\$563,577,022,958	\$23,108,764,916	\$29,585	\$29,347
2022-05-25	\$564,885,248,022	\$23,486,436,837	\$29,655	\$29,585
2022-05-24	\$555,459,414,424	\$26,646,754,325	\$29,163	\$29,655
2022-05-23	\$578,090,387,085	\$17,199,930,155	\$30,351	\$29,163
2022-05-22	\$561,696,351,559	\$13,561,006,428	\$29,492	\$30,351
2022-05-21	\$557,201,657,281	\$25,980,877,920	\$29,257	\$29,492
2022-05-20	\$578,678,245,848	\$29,190,056,577	\$30,382	\$29,257
2022-05-19	\$547,927,169,401	\$26,219,431,165	\$28,772	\$30,382
2022-05-18	\$581,272,773,176	\$24,964,703,224	\$30,502	\$28,772
2022-05-17	\$569,801,943,005	\$28,418,179,221	\$29,924	\$30,502
2022-05-16	\$596,346,266,889	\$22,948,123,704	\$31,319	\$29,924
2022-05-15	\$573,777,699,744	\$25,759,546,764	\$30,189	\$31,319
2022-05-14	\$558,053,182,541	\$71,541,696,924	\$29,311	\$30,189
2022-05-13	\$555,266,177,877	\$61,338,080,284	\$29,126	\$29,311
2022-05-12	\$549,294,001,873	\$65,985,702,702	\$28,913	\$29,126

UNSTRUCTURED SOURCES:

To explore our social-media crowd sourcing strategy to make our dataset better we chose Twitter as one of our unstructured sources.

Packages used:

- 1- Snsrape.modules.twitter – snsraeper is a python library used for scrapping Social Networking Sites(SNS) . It is helpful to scrape things like user profiles , hashtags, searches and return relevant posts.
- 2- Got the tweets for keyword like 'bitcoin' from 21-03-2020 to 30-04-2022 and set tweet limit to be 50000.
- 3- Created a list of tweets and saved it into a list to be used in a Dataframe later.
- 4- Used TwitterSearchScrapecr().getititems() function to get the data available which are Date, tweet, tweet ID, username.

Challenges faced & how can we overcome it?

While trying to fetch the tweets between 21-03-2020 to 30-04-2022, Jupyter notebook kept running for hours. So, we decided to limit the tweets scraping to 50000 which can be considered as a good number in doing the sentiment analysis.

OR

We can write an optimised code to count the number of re-tweets to directly extract the maximum re-tweets on a particular date and then use appropriate method for sentiment analysis of traders around the world.

The end objective of the data collection exercise was to find a source where we can get access to the historical data for at least last 6 years. This is exactly the period when Crypto-currency trading became widely accepted and gained momentum across different investors and there was significant transaction volume and market-cap for the key currencies like Bitcoin, Ethereum, Tether, USD Coin, BNB etc.

- **BeautifulSoup** – It is a library used for scrapping data from the web where data is in either a HTML or XML format. It has HTML or XML parser for extracting data from a class-attribute hierarchy.
- **Requests** – A simple library to fetch data by sending a HTTP request to a URL, to fetch data.
- **JSON** – So that the final scraped data can be saved or viewed in JSON format.

After scrapping the data from Coin Gecko using the above-mentioned libraries, we save it into a JSON file.(data.json). This data set was further cleansed and enhanced which we will touch base upon in the next section

112835 rows x 6 columns

112835 rows x 6 columns

Challenges faced while collecting the data:

When we started collecting historic data from Coin Gecko, the major challenges we faced were:

Problem Statement	Proposed Solution
Inspect element on coingecko.com had 4 numerical columns on the page had same class attributes. While extracting only one column's value was getting converted	We took the help of geeksforgeeks.com to find the syntax for extracting columns with same class attribute name
Historical data was available for only current year and the existing code had to be recalibrated to pull the historical data	We recalibrated the Seed URL in way so that it can run dynamically across all 100 coins to extract the historical data. We created a list of URLs that can extract historic data for the 100 coins
Only 1 year block data could be extracted in each iteration for each coin	We used pagination to scrape data through each page using a 'for' loop. However, the run time was almost 3 mins through Jupyter notebook, we anticipate that any other powerful environment would fetch results faster. Fetching details of each coin takes about 10-15 secs, so for 100 coins it might take a few minutes to scrape the desired data.

Data cleaning/pre-processing as needed

Cleansing the dataset:

For cleansing of data, the below python libraries used were:

- **NumPy** – NumPy is fundamental package for doing mathematical and scientific operations on multidimensional arrays and Dataframe in our code.
- **Pandas** – This open -source python library was used to import and manage our large dataset. This library was useful for manipulation , processing and analysing the data.
- **Matplotlib** – It is a 2Dimensional plotting library in Python , which we used to our final analysis on our clean data.
- **Seaborn** – It is a part of Matplotlib library used for statistical graphs.

Using Pandas, we imported our raw data form JSON format into Dataframe format in Jupyter notebook.

Identifying and handling missing values:

We had 'N/A' as missing values in the data for the "latest" Close price data for every currency. The Close price for every currency on the most recent date was unavailable. We used "NumPy. Where" to find the no of missing values and then removed them from the dataset. Finally, these rows with Close prices as 'N/A' were removed as we had ample historic data to process and analyse.

Cleaning column data:

Columns – Market Cap , Open, Close, Volume had '\$' and ',' in the values. These special characters were removed using string operations in dataframe and convert from Object datatype to Float datatype.

After the above 2 data cleansing steps , our dataset looked like this:

Figure 2: Dataset after removing missing values and special characters

	Date	Currency	MarketCap	Volume	Open	Close
1	2022-05-16	bitcoin	5.963463e+11	2.294812e+10	31319.0	29924.0
2	2022-05-15	bitcoin	5.737777e+11	2.575955e+10	30189.0	31319.0
3	2022-05-14	bitcoin	5.580532e+11	7.154170e+10	29311.0	30189.0
4	2022-05-13	bitcoin	5.552662e+11	6.133808e+10	29126.0	29311.0
5	2022-05-12	bitcoin	5.492940e+11	6.598570e+10	28913.0	29126.0
6	2022-05-11	bitcoin	5.906444e+11	5.407136e+10	31027.0	28913.0
7	2022-05-10	bitcoin	5.761978e+11	5.836993e+10	30270.0	31027.0
8	2022-05-09	bitcoin	6.485143e+11	3.081058e+10	34070.0	30270.0
9	2022-05-08	bitcoin	6.787024e+11	1.913072e+10	35573.0	34070.0
10	2022-05-07	bitcoin	6.870934e+11	3.222329e+10	36116.0	35573.0

Adding relevant columns:

To make this data more useable we a few more columns so that traders can use the data to predict the market and take informed decisions accordingly for future buying or selling based on this data.

- **Simple Moving Average** : Since we scrapped 3 years of data, we added calculated columns for weekly, monthly, half year simple moving average to generate more meaningful trends. Using rolling mean function in pandas , the above columns were calculated

- **% Change Columns:** 3 Columns to assess the % change in 24h, 1 week and 1 month was also calculated for each crypto currency
- **Close Prices in different currency converts :** As the data scraped was in USD, 6 columns were added for different currencies. By converting US Dollar to Indian Rupee, Japanese yen , Euro, Pound, Canadian Dollars, Danish Krona, the dataset can be used by traders from different countries as well. As trading is done based on the Close price , so we converted the Close price in USD into above mentioned currencies, so that we will be able to analyse the behaviour of the Close price for a particular cryptocurrency in different countries

Below are the columns that we added in the data set for further analysis:

Category	Column Name	Description
Simple Moving Averages	7D_SMA	7-day Simple moving average of close prices
	20D_SMA	20-day Simple moving average of close prices
	50D_SMA	50-day Simple moving average of close prices
	100D_SMA	100-day Simple moving average of close prices
	200D_SMA	200-day Simple moving average of close prices
% Change	24h %change	% Change in closing price within 24 Hours window
	7d %change	% Change in closing price within 7 days window
	30d %change	% Change in closing price within 30 days window
Currency Conversions	Close_INR	Close Price Converted into INR
	Close_EUR	Close Price Converted into Euro
	Close_JPY	Close Price Converted into Yen
	Close_GBP	Close Price Converted into GBP
	Close_CAD	Close Price Converted into CAD

Figure 3: Adding relevant columns to the dataset

Close	7D_SMA	20D_SMA	50D_SMA	100D_SMA	200D_SMA	24h_%Change	7D_%Change	30D_%Change	Close_INR	Close_EUR	Close_JPY	Close_GBP
0.076758	0.077688	0.093758	0.102526	0.073594	0.077158	-0.066893	0.055702	-0.426023	5.9457	0.0718	9.7575	0.0612
0.077350	0.078234	0.092237	0.103192	0.073925	0.077074	0.007713	0.051952	-0.361693	5.9915	0.0724	9.8327	0.0616
0.064018	0.077357	0.089857	0.103618	0.074110	0.076927	-0.172359	-0.087478	-0.471122	4.9588	0.0599	8.1380	0.0510
0.069557	0.077561	0.087006	0.104144	0.074347	0.076814	0.086523	0.020885	-0.427420	5.3879	0.0651	8.8421	0.0554
0.053736	0.071852	0.083903	0.104316	0.074437	0.076602	-0.227454	-0.426480	-0.515879	4.1624	0.0503	6.8309	0.0428
0.045961	0.067089	0.080543	0.104254	0.074446	0.076358	-0.144681	-0.420440	-0.574742	3.5602	0.0430	5.8426	0.0366
0.050718	0.062585	0.077541	0.104282	0.074496	0.076101	0.103491	-0.383315	-0.549530	3.9286	0.0475	6.4473	0.0404
0.052738	0.059154	0.074839	0.104381	0.074580	0.075844	0.039828	-0.312932	-0.543469	4.0851	0.0494	6.7041	0.0420
0.055113	0.055977	0.072387	0.103680	0.074682	0.075604	0.045034	-0.287485	-0.506863	4.2691	0.0516	7.0060	0.0439
0.050270	0.054013	0.070262	0.102580	0.074682	0.075297	-0.087874	-0.214752	-0.547919	3.8939	0.0471	6.3903	0.0401

SAVING THE PRE-PROCESSED DATASET:

After removing the unnecessary and missing values and adding relevant columns that would consolidate our continuous date series, we saved it into a JSON file – (crypto.json)

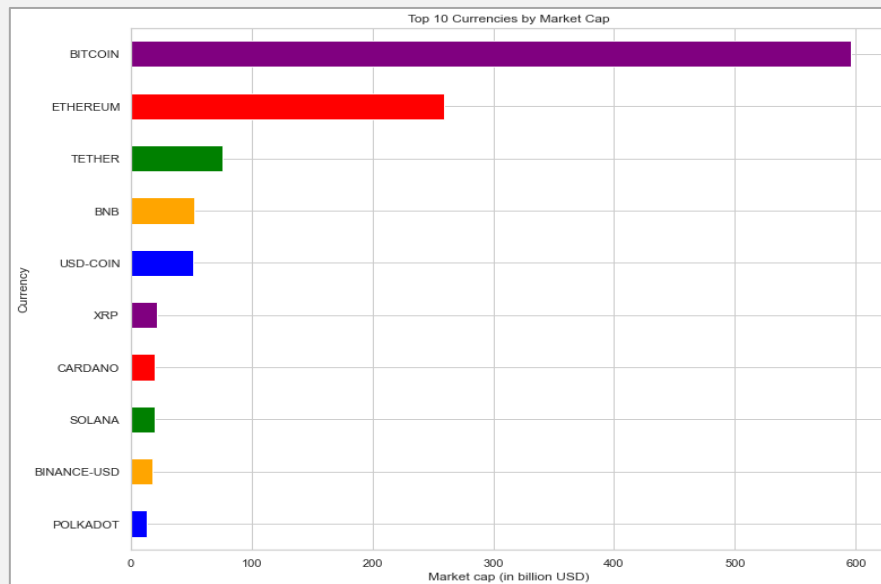
Final Output structure of our data.

Figure 4: Columns of the final structured dataset

```
Data columns (total 20 columns):
#      Column      Non-Null Count  Dtype
---  -
0     Date         110904 non-null   datetime64[ns]
1     Currency      110904 non-null   object
2     MarketCap     110904 non-null   int64
3     Volume        110904 non-null   float64
4     Open          110904 non-null   float64
5     Close         110904 non-null   float64
6     7D_SMA        110904 non-null   float64
7     20D_SMA       110904 non-null   float64
8     50D_SMA       110904 non-null   float64
9     100D_SMA      110904 non-null   float64
10    200D_SMA      110904 non-null   float64
11    24h_%Change  110904 non-null   float64
12    7D_%Change   110904 non-null   float64
13    30D_%Change  110904 non-null   float64
14    Close_INR    110904 non-null   float64
15    Close_EUR    110904 non-null   float64
16    Close_JPY    110904 non-null   float64
17    Close_GBP    110904 non-null   float64
18    Close_CAD    110904 non-null   float64
19    Close_DKK    110904 non-null   float64
dtypes: datetime64[ns](1), float64(17), int64(1), object(1)
```

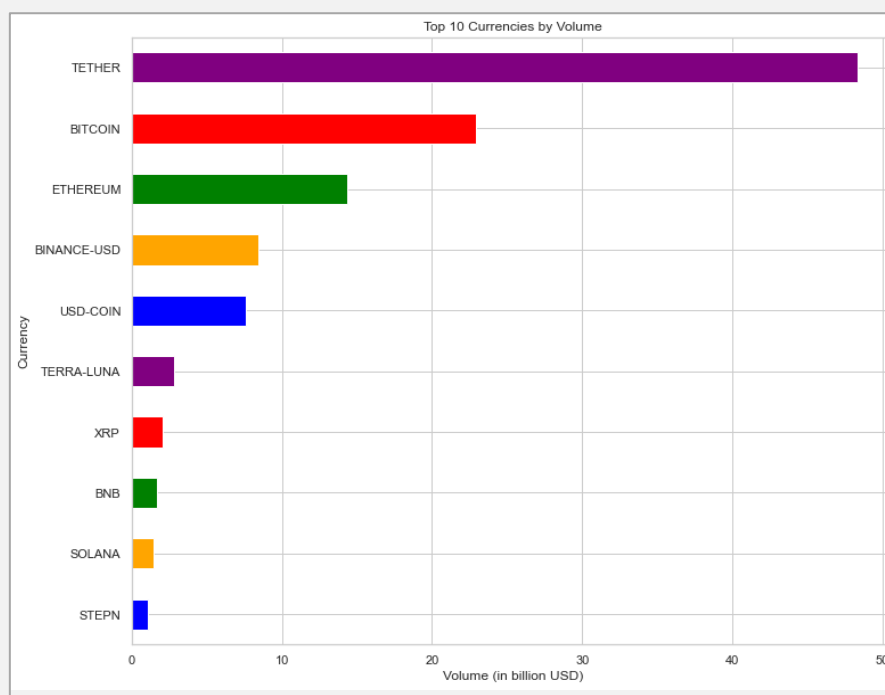
Observations/ Insights and Analysis on the data collected

1. Top 10 currencies by Market Cap (in Billion USD):



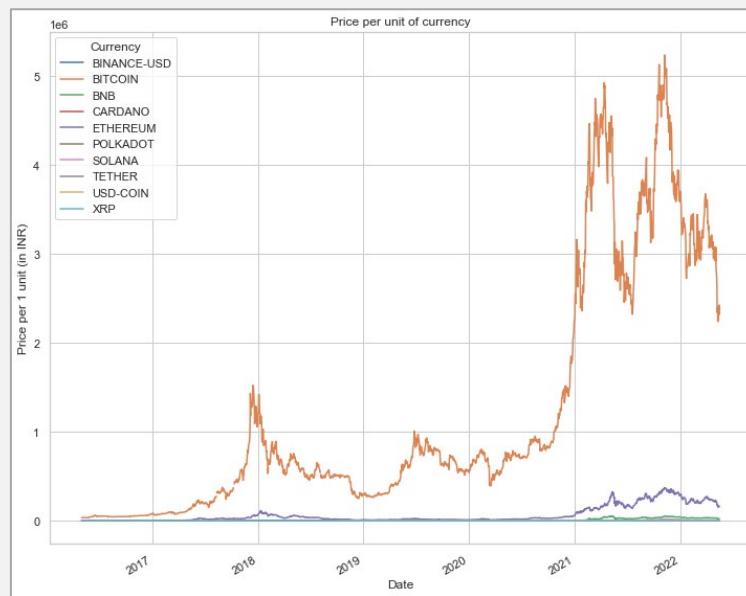
From the graph above , we see that from the data collected , the Highest Market Cap (in Billion USD) is of Bitcoin with more than 550 billion USD Market Cap and Polkadot is the crypto currency with the 10th Highest Market Cap of more than 10 billion USD till 2022.

2. Top 10 currencies by Transaction Volume:

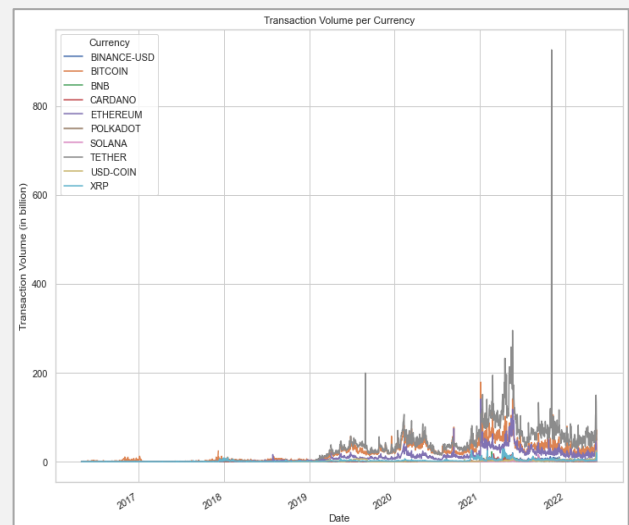
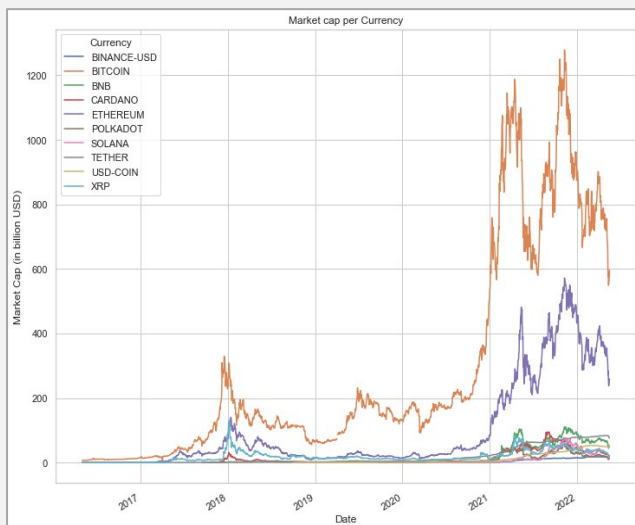


From the graph, Tether has the highest transactional Volume with more than 40 billion USD in total and STEPN has the 10th Highest Transactional Volume with more than 1 million USD in total.

3. Trends of Top 10 currencies (From 2017-2022):



The graph suggests that Bitcoin has had the highest Close Price since 2017 and XRP the 10th Highest. Bitcoin and Ethereum have had a significant rise in their Close Prices since 2021. But other crypto currencies have not seen a significant high in their Close Prices.



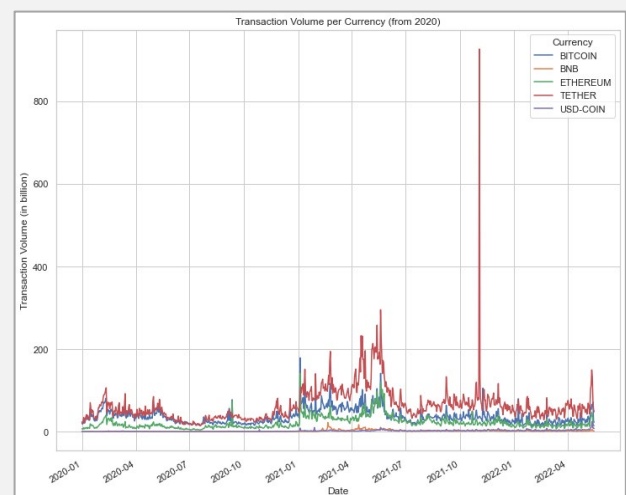
Where the Market Cap of Bitcoin is the highest, we see that Tether has the highest transaction Volume. This might be because the price of Bitcoin kept increasing significantly because of which traders must have bought them in less amounts. Whereas we see that Tether Prices has not increased significantly and it is dependent on Bitcoin Prices. Hence traders must have considered Tether as a multi-bagger option due to low price and high volume.

Where the Market Cap of Bitcoin is highest, the transaction Volume of Bitcoin is 2nd Highest. 2020 was a significant year in the world of crypto currency because of increase in demand. This provided a perfect environment for investing in crypto currency. Due to the global economic crisis, as the traditional currency and assets took a massive hit, investors turned to crypto-currency and hence the valuation of Bitcoin and Ethereum saw a meteoric rise as the big brand investors and companies stockpiled on digital currency.

4. Trends of Top 5 Currencies from 2020 onwards:



We see that since January 2021, there has been a significant rise in Bitcoin and Ethereum Prices, while BNB, Tether and USD-coin have been consistently stagnant till now. This might be due to the increasing popularity of Bitcoin and Ethereum, also consistent rise in Crypto-currency demand.



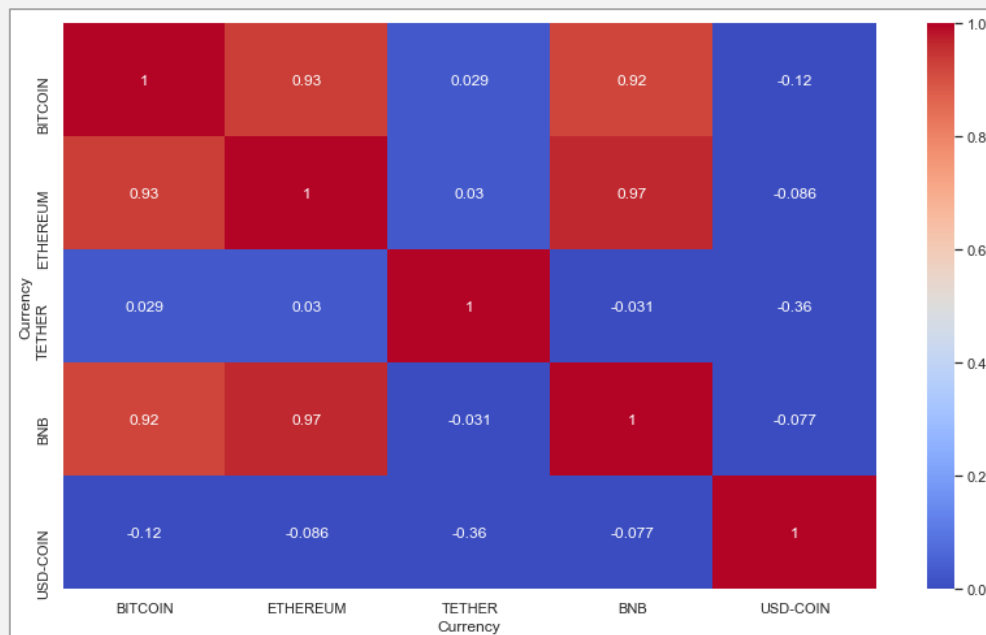
Since 2020, we see a huge increase in Bitcoin Market Cap as there was an increase in its Price as well. The 2 crypto currencies which saw a meteoric rise in their Market Cap were Ethereum and Bitcoin.

Also, the rate of increase of prices for both Bitcoin and Ethereum is significantly high. This must be due to high demand of these currencies.

In the Transaction Volume however we see that Tether had an increase in price post 2020 but not as significant as the Top 2 currencies, and hence investors have been buying this currency in large amounts.

We also see that the transaction volume of Bitcoin and Ethereum have been quite similar after 2021, due to the steep increase in their prices.

5. Correlation of Top 5 Currencies based on Bitcoin:



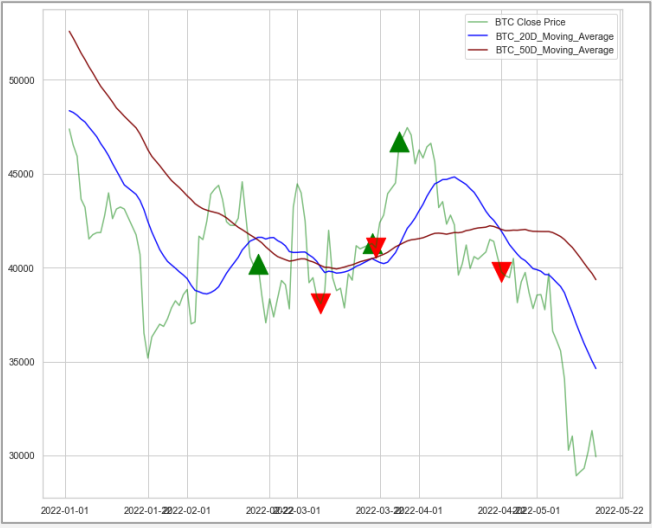
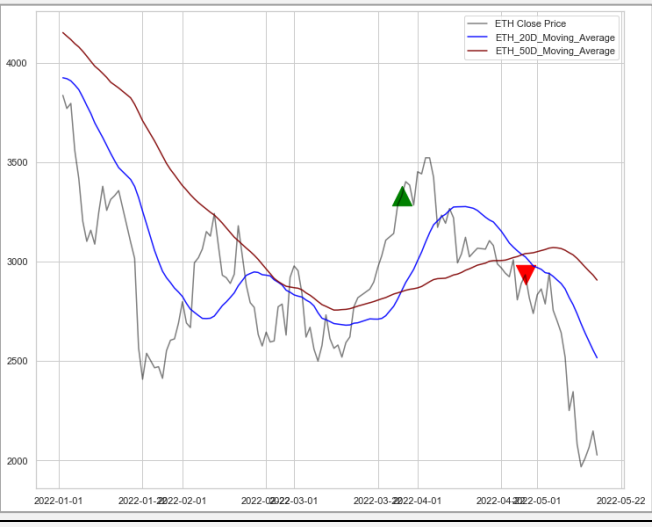
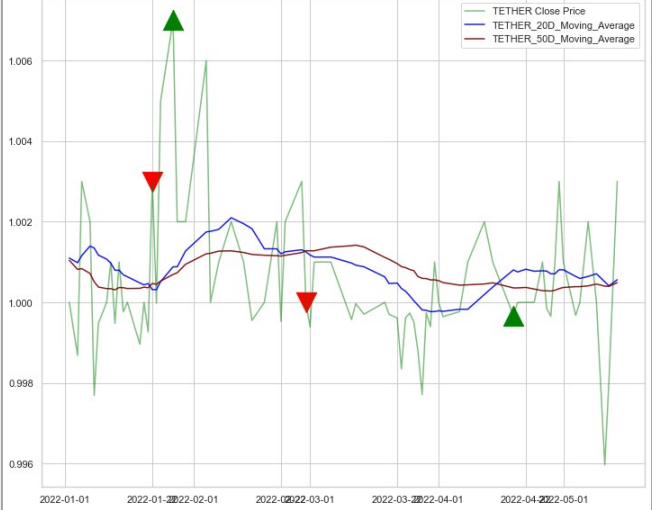
From the heat-map above , we see that BNB and ETHEREUM have more direct correlation with BITCOIN.

None of the other 4 crypto-currencies have a strong direct or indirect correlation with USD-COIN as the correlation is closer to 0 than 1 or -1.

Tether has an indirect correlation with Bitcoin. Hence the value of tether coin usually stays stable and hence is known as Stable coin.

BNB and Ethereum also have a direct correlation with each other.

6. Trading Analysis based on Moving Averages of Bitcoin, Ethereum and Tether:

Graph	Interpretation
	<p>The trading analysis chart is shown for BITCOIN, The buy and sell index have been determined based on the columns – 20D_SMA and 50D_SMA.</p> <p>When the 20-Day simple moving average > 50-day simple moving average on 1 day and 20D_SMA < 50D_SMA on previous day, an investor must 'Buy'. If the condition is completely opposite, then they must sell.</p> <p>Based on Moving average only, this chart is suggesting Selling as the 20D_SMA has decreased than the 50D_SMA.</p>
	<p>The trading analysis chart is shown for ETHEREUM, The buy and sell index have been determined based on the columns – 20D_SMA and 50D_SMA.</p> <p>Based on Moving average only, this chart is suggesting Selling as the 20D_SMA has decreased than the 50D_SMA.</p>
	<p>The trading analysis chart is shown for ETHEREUM, The buy and sell index have been determined based on the columns – 20D_SMA and 50D_SMA.</p> <p>Based on Moving average only, this chart is suggesting Buying as the 20D_SMA has increased than the 50D_SMA.</p>

In real world analysts not only, trade based on technical knowledge of Moving averages but also on trends, news , government policies etc. and investor sentiment analysis.

7. Prediction Analysis for 2028: Bitcoin and Ethereum:

Based on the Close prices we have from the last 6 years we can make a prediction of what the Close price might be in the next 6 years i.e., in 2028. So, we have considered the CAGR – Compound Annual Growth rate from 2020 to 2022.

2020 is the year when crypto-currency prices rose prominently than their significantly low prices till 2019. Hence predicting based on 2020 prices would give us more accurate values.

CAGR calculating function:

```
# Defining CAGR function which will help in calculating the CAGR
def CAGR(start, end, time):
    Growth_Rate = (end/start)**(1/time)-1
    return Growth_Rate

#Extracting Close price data for a particular date
start_btc = float(df_bitcoin.Close[df_bitcoin.Date == '2020-04-30'])
start_eth = float(df_ethereum.Close[df_ethereum.Date == '2020-04-30'])
print(start_eth)
print(start_btc)

205.56
8610.64

#Extracting Close price data for a particular date
end_btc = float(df_bitcoin.Close[df_bitcoin.Date == '2022-05-01'])
end_eth = float(df_ethereum.Close[df_ethereum.Date == '2022-05-01'])
print(end_eth)
print(end_btc)

2832.51
38538.0

#Calculating CAGR for ethereum in the last 2 years
eth_CAGR = CAGR(start_eth, end_eth, 2)
btc_CAGR = CAGR(start_btc, end_btc, 2)
print('Ethereum grows with a CAGR of {:.2%}'.format(eth_CAGR))
print('Bitcoin grows with a CAGR of {:.2%}'.format(btc_CAGR))

Ethereum grows with a CAGR of 271.21%
Bitcoin grows with a CAGR of 111.56%

# Defining forecast function
def forecast(end, CAGRr, years):
    forc = end + (1+CAGRr)**years
    return forc

# forecasting ethereum price in next 6 years based on CAGR calculated
years = 6
print('Forecast for ethereum in 6 years time', forecast(end_eth, eth_CAGR, years))
print('Forecast for bitcoin in 6 years time', forecast(end_btc, btc_CAGR, years))

Forecast for ethereum in 6 years time 5448.876189996257
Forecast for bitcoin in 6 years time 38627.65225552491
```

Hence, we can find that based on the CAGR the price around 1st May ,2028:

For Ethereum can go up to \$5,450

For Bitcoin can go up to \$38,627

Strategy to enhance the data with crowd sourcing methods

In recent years, crowdsourcing has emerged as a new area of research. Various users have found it to be a very intriguing and enjoyable feature to investigate due to its uniqueness. It is regarded as an innovative and dynamic web-enabled service platform that is suited to leveraging people's enormous potential via the internet. Crowdsourcing has been expedited and made easier thanks to technical advancements and extended collaboration tools such as platforms and social networks. Traders and Investors can provide critical and constructive feedback around the trading schemes through various online forums like Twitter, Facebook, RSS Feed, Blogs, News etc. This provides investor an opportunity to gain assessments on how to control risk while maximizing returns.

Through crowd sourcing the hobbyist & the professional traders can share their thoughts and strategic across different crypto currencies which can help the investor communities to decide & formulate their individual trading strategies. Using different ML techniques on the rich text data available on different social platforms the investors can understand the Sentiment of a currency and train a ML based trading model which continuously generates a LONG or SHORT signal and continuously learns over time and help the users to take informed decision.

Example of Platforms that uses Crowd Sourcing to enhance Crypto Trading

QuantConnect Community - <https://www.quantconnect.com/forum/discussions/1/interesting>

QuantConnect connects engineers all around the globe with market data and a cluster computer to test and design quantitative trading techniques in a variety of cryptocurrencies. The contributors develop and generate algorithm that help other traders to use the platform's knowledge infrastructure to launch their own hedge fund start-ups.

Twitter – <https://www.twitter.com>

Social Platforms like Twitter empower people to write context notes , facts and figures around the crypto currencies, these tweets add helpful and informative perspective for people from different points of view. Twitter has stressed the necessity of having contributors with a variety of perspectives, which are chosen not by demographics but by how they have rated previous related tweets. Any tweet that forms a strong point of view around any currency is generally Top rated/retweeted by the users. With the help of these tweets, we can enhance the Crypto Data by adding the flavour of Sentiment analysis that can be derived from these tweets and then run ML algorithms to make informed decisions around the Buy and Sell strategies in the crypto market.

Reddit - <https://www.reddit.com/r/CryptoCurrency/>

Reddit is another crypto-currency News blog which has about 4.9M members currently. It is a network of communities where people with common hobbies, interests and passions come together. It is a web content rating and discussion website. Hence , for millennial traders this website might give us insight into greed and fear index based on the crypto-currency discussion.

CoinTelegraph - <https://cointelegraph.com/>

CoinTelegraph is one of the oldest blogs founded in 2013 in this domain. This crypto-currency blog has covered numerous important events related to crypto world. Only limitation with this website is it readily has analysis done only on the Top 10 currencies in prices in real-time. Main topics covered in this are related to trading, mining, blockchain, news etc. This website might help us understand the popularity of the oldest coins in this domain.

Here is an **Example** how this dataset can be made more informative using crowd-sourcing websites:

Social media crowdsourcing is one of the best ways to find the traders' sentiments towards any given crypto currency.

Target: To enhance our dataset based on people's opinions and trends on social media.

Platform used: Twitter

New columns added: Using the Python library mentioned in Section-4, we found three more columns from the tweets extracted – Subjectivity, Polarity and Sentiments

Category	Column Name	Description
Sentiment Analysis	Subjectivity	Finding how influential someone's tweet can be, or how much a tweet is based on personal opinion than facts. Lies between [0,1]
	Polarity	How much opinion is contradictory or negative towards a certain topic. Lies Between [-1,1]
	Sentiments	Sentiment analysis – Positive , Neutral or Negative based on the above listed columns

Analysis on the new columns: For Bitcoin and Ethereum.

For instance, we took 2 of the most popular coins and did their sentiment analysis.

The data for **Bitcoin** looked like this :

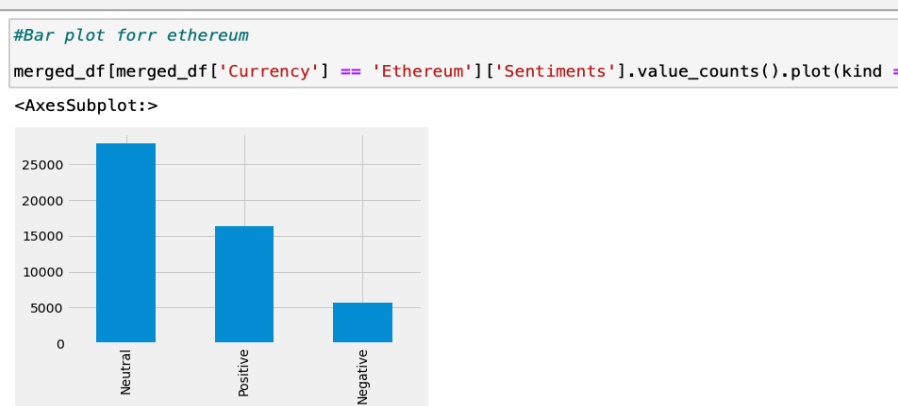
	Currency	Date	Clean_Tweets	Subjectivity	Polarity	Sentiments
0	Bitcoin	2022-04-29	Bitcoin (BTC) Price Update: Price: \$38,651 Last...	0.000000	0.000000	Neutral
1	Bitcoin	2022-04-29	Bitcoin price index USD EUR CNY GBP RUB	0.000000	0.000000	Neutral
2	Bitcoin	2022-04-29	silver ape alert emergency update. Jump in : L...	0.500000	0.136364	Positive
3	Bitcoin	2022-04-29	The Dilemma Faced by All Beginners in the Bitc...	0.535714	0.285714	Positive
4	Bitcoin	2022-04-29	Bitcoin Last Price \$38575 BTC 🚀 Daily Indicator...	0.066667	0.000000	Neutral

The data for **Ethereum** looked like this:

	Currency	Date	Clean_Tweets	Subjectivity	Polarity	Sentiments
0	Ethereum	2022-04-29	I know it may not feel like it but this is the...	0.294444	0.422222	Positive
1	Ethereum	2022-04-29	Bitcoin Last Price \$38575 BTC 🚀 Daily Indicator...	0.066667	0.000000	Neutral
2	Ethereum	2022-04-29	You know whose been swimming with their trunks...	0.600000	-0.100000	Negative
3	Ethereum	2022-04-29	How much Ethereum should one pleb own? 1 whole...	0.533333	0.333333	Positive
4	Ethereum	2022-04-29	Culpower found ethereum in a User vault at thl...	1.000000	1.000000	Positive

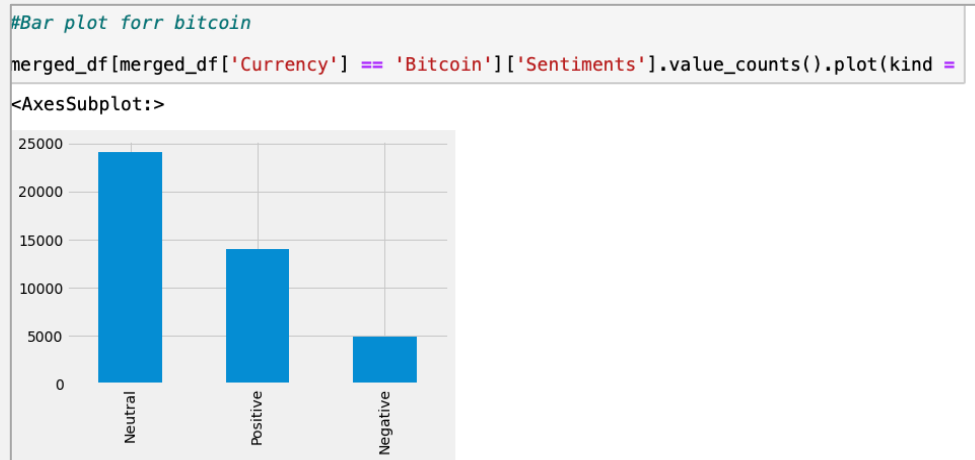
Analysis on Tweet sentiment for Ethereum:

As we can see from the below graph, a trader from this preliminary analysis would see that Ethereum has more Positive sentiments than Negative and as Neutral Sentiments are the highest might decide neither to sell nor to buy at this point based on the latest twitter trends. This is a sentimental analysis of the trends on #ethereum.



Similarly, Analysis on Tweet Sentiment for Bitcoin:

As we can see from the below graph, a trader from this preliminary analysis would see that Bitcoin has more Positive sentiments than Negative and as Neutral Sentiments are the highest might decide neither to sell nor to buy at this point based on the latest twitter trends. This is a sentimental analysis of the trends on #bitcoin.



Possible challenges ahead:

The possible challenges we might face going ahead is that we have historic price data for at most 6 years for the oldest crypto currencies. As their popularity will also be high the number of tweets is going to be in millions. If one coin has millions of tweets, then 100 coins will have billions of tweets. So, it might take a huge amount of time to scrape data from any crowd-sourcing website.

Possible solutions:

- Use retweet count as a part of our coding algorithm. Listing only top 5 most retweeted tweets per day and then do our sentiment analysis based on that
- Change our search hashtag which will help to zero-out sentiments and remove news items and scrape out only opinion tweets related to that hashtag
- Reduce our historic data timeline to be able to accommodate relevant tweets which show crowd opinion and sentiment for analysis
- We can do our sentiment analysis, add Bull(optimistic) & Bear(pessimistic) and Greed & Fear Index columns based on social media crowd-sourcing platforms and crypto-currency news blogs

References and Sources used for this Assignment

Sources

- <https://www.coingecko.com/>

GitHub Link – Final Submission for Group 3

- https://github.com/Nikhileshwar-AVS/DCPP_Group-3

References

- <https://www.towardsdatascience.com/cryptocurrency-market-analysis-with-web-scraping-61e4fd0b1c81>
- <https://towardsdatascience.com/algorithmic-trading-in-python-simple-moving-averages-7498245b10b>
- <https://www.socialmediatoday.com/content/how-use-twitter-crowdsourcing-and-simple-market-research>
- <https://www.geeksforgeeks.com>
- <https://www.quantconnect.com/forum/discussions/1/interesting>
- <https://www.youtube.com/watch?v=XOdrsdhWpKE>
- https://github.com/uforodavid/horrible_bosses