Assignment 1: Project Data Mosaic

Group Number: 34

Student IDs:

Ashar Nasir	24280027	Worked on the analytical aspect of the data, and report creation.
Altaf Hussain	24280081	Wrote the code for the whole data pipeline, including data extraction, collection, transformation and load.

Github Link: https://github.com/ashar0800/assignment1

Overview of Your Topic: Why did you choose it? What data do you expect to see?

We chose Cryptocurrency and Blockchain because of our interest in Quantitative Finance and the Cryptocurrency Market.

We expect to observe insights from the extracted data that will correlate with the state of the cryptocurrency market cycle. During a bull run, we expect certain metrics to inflate and similarly during a bear run, some of the other metrics would inflate. It is up to us how we define those metrics based on the domain knowledge.

Data Collection Process: Summarize the steps you took for each source and any challenges (API rate limits, incomplete data, TOS constraints).

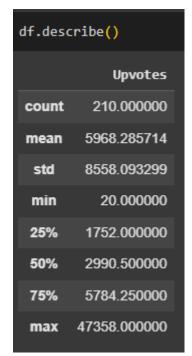
-> We fetched data from Reddit, Kaggle Datasets, Yahoo Finance and Google Trends.

The data collected from these sources had these issues:

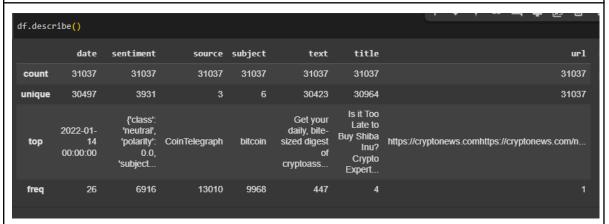
- API rate limit and restriction on how much data can be fetched concurrently.
- Deleted and removed posts
- Bot generated data can lead to noise
- Without Authentication, we cannot access all kinds of data.
- Compliance and Privacy laws pose ethical and legal challenges.
- User consent is not incorporated in the operation.

Initial Observations: Generate a summary of the datasets using pandas. Add the screenshot of the console output of pandas DataFrame in the document.

The data collected from Reddit, Kaggle, Yahoo and Google Trends was analysed using Pandas describe() statement. The resulting screenshots are pasted below:



Reddit Data



Kaggle Dataset

historical_data.describe()									
	0pen	High	Low	Close	Volume	Dividends	Stock Splits		
count	185.000000	185.000000	185.000000	185.000000	1.850000e+02	185.0	185.0		
mean	81489.408193	83142.968454	80007.674937	81698.831651	4.760960e+10	0.0	0.0		
std	17714.963518	18091.900923	17223.483955	17661.542469	2.724488e+10	0.0	0.0		
min	53949.085938	54838.144531	52598.699219	53948.753906	1.240347e+10	0.0	0.0		
25%	63184.339844	64443.707031	62442.152344	63193.023438	2.935094e+10	0.0	0.0		
50%	90536.812500	91868.742188	88741.664062	90558.476562	4.064664e+10	0.0	0.0		
75%	97508.382812	99014.679688	95747.226562	97508.968750	5.780592e+10	0.0	0.0		
max	106147.296875	109114.882812	105291.734375	106146.265625	1.492189e+11	0.0	0.0		

Yahoo Finance Data df2.describe() Bitcoin Blockchain Cryptocurrency 54.000000 54.000000 count 54.000000 45.537037 2.944444 mean 1.629630 17.509796 0.301985 0.708341 std 27.000000 2.000000 1.000000 min 31.250000 25% 3.000000 1.000000 50% 37.000000 3.000000 1.500000 75% 54.000000 3.000000 2.000000 100.000000 4.000000 3.000000 max

What AI product will you make using this data?

-> We can create a real time investment recommendation system based on either long term data for long term investors or short term data for short term investors.

Google Trends

How does collecting from multiple sources help or hinder data quality? What conflicts or discrepancies might you face?

- -> Collecting data from multiple sources can greatly improve the quality and coverage of data and the analysis offered by it. More data leads to better predictive analytics..
- -> We do face some conflicts and discrepancies, as detailed below:
 - For data aggregation, we do not have common ground established to view the whole data in unison.
 - Different data sources might come with data in different formats, which would require extensive preprocessing.
 - The aggregation may lead to inconsistent data
 - The aggregation may lead to duplicates.

Can you think of ways to store and combine all of this data?

We can store and combine this kind of data in different ways. Some of them are:

- 1. Storing data using pandas framework
- 2. Store and combine data into a database (MySQL etc)
- 3. Storing data in a file (CSV etc)

