Final Project: Data Preparation

Kesav Adithya Venkidusamy

Bellevue university - Master of Science in Data Science

Course Name: DSC540-T301 Data Preparation (2223-1)

Assignment: Final Project Milestone 1

Instructor: Dr Catherine Williams

Due Date: 01/09/2022

Topic: Cryptocurrency Analysis

A cryptocurrency is a tradable digital asset or digital form of money, built on blockchain technology that only exists online. Cryptocurrencies use encryption to authenticate and protect transactions, hence their name. There are currently over a thousand different cryptocurrencies in the world, and many see them as the key to a fairer future economy. Out of numerous currencies available out there, I have considered 9 famous currencies based on market capitalization for my analysis. This analysis is to find out the feasibility of investment in cryptocurrency.

Datasets

As part of first milestone of this term project, I have considered three difference data sources that have different file types of information as mentioned below. I also established the relationship between them.

1. CSV file data source

First data source is csv file containing historical data for 9 cryptocurrencies which I considered for my analysis. I extracted the data from Yahoo finance for each cryptocurrency and consolidated into one. A yahoo finance link to extract historical data for Bitcoin (BTC) is given below.

https://finance.yahoo.com/quote/BTC-USD/history?p=BTC-USD

Crypto_USD.csv

2. API

Second data source is API as mentioned below which returns the current/latest price for all the cryptocurrencies present out there. In addition, I have provided another API link which returns the latest for the symbol passed as parameter.

API Link 1: https://pro-api.coinmarketcap.com/v1/cryptocurrency/listings/latest

API Link 2: https://pro-api.coinmarketcap.com/v1/cryptocurrency/quotes/latest

3. Website data

Third data source is website data. The link mentioned below has the metadata information like founder, symbol, hash algorithm and program language of implementation about the major cryptocurrencies in the tabular format.

Website Link: https://en.wikipedia.org/wiki/List_of_cryptocurrencies

Metadata and relationship for datasets

1. CSV Dataset

Metadata

CSV data source has three years of history data for below cryptocurrencies.

- Bitcoin – BTC
- Bitcoin Cash – BCH
- Ethereum – ETH
- Dogecoin – Doge
- Ethereum Classic – ETC
- Cardano – ADA
- Litecoin – LTC
- Neo – Neo
- Tether – USDT

The dataset has following fields

- Date - Date
- Open - Open price for the day for the cryptocurrency
- High - High price for the day for the cryptocurrency
- Low - Low price for the day for the cryptocurrency
- Close - Close price for the day for the cryptocurrency
- Adj Close - Adjusted close price for the day for the cryptocurrency
- Volume - Transaction Volume for the day for the cryptocurrency
- Symbol - Cryptocurrency symbol

Relationship

"**Symbol**" field which represents the symbol for cryptocurrency is used as a common key to make a join with other two datasets. The symbol field contains "-USD" word which needs to be removed to make a join with other datasets.

2. API Dataset

Metadata

The below API returns the array/list of cryptocurrency objects mentioned in the table for the parameters passed.

https://pro-api.coinmarketcap.com/v1/cryptocurrency/listings/latest

id	integer	The unique CoinMarketCap ID for this cryptocurrency.
name	string	The name of this cryptocurrency.
symbol	string	The ticker symbol for this cryptocurrency.
slug	string	The web URL friendly shorthand version of this cryptocurrency name.

cmc_rank	integer	The cryptocurrency's CoinMarketCap rank by market cap.
num_market_pairs	integer	The number of active trading pairs available for this cryptocurrency across supported exchanges.
circulating_supply	number	The approximate number of coins circulating for this cryptocurrency.
total_supply	number	The approximate total amount of coins in existence right now (minus any coins that have been verifiably burned).
market_cap_by_total_supply	number	The market cap by total supply. *This field is only returned if requested through the aux request parameter.*
max_supply	number	The expected maximum limit of coins ever to be available for this cryptocurrency.
last_updated	string <date>	Timestamp (ISO 8601) of the last time this cryptocurrency's market data was updated.
date_added	string <date>	Timestamp (ISO 8601) of when this cryptocurrency was added to CoinMarketCap.
tags	String	Array of tags associated with this cryptocurrency. Currently only a mineable tag will be returned if the cryptocurrency is mineable. Additional tags will be returned in the future.
Platform	Map	Metadata about the parent cryptocurrency platform this cryptocurrency belongs to if it is a token, otherwise null.
Quote	Map	A map of market quotes in different currency conversions. The default map included is USD.

The details for the below API is mentioned below.

https://pro-api.coinmarketcap.com/v1/cryptocurrency/quotes/latest

id	integer	The unique CoinMarketCap ID for this cryptocurrency.
name	string	The name of this cryptocurrency.
symbol	string	The ticker symbol for this cryptocurrency.
slug	string	The web URL friendly shorthand version of this cryptocurrency name.
is_active	integer [0 .. 1]	1 if this cryptocurrency has at least 1 active market currently being tracked by the platform, otherwise 0. A value of 1 is analogous with listing_status=active.
is_fiat	integer [0 .. 1]	1 if this is a fiat
cmc_rank	integer	The cryptocurrency's CoinMarketCap rank by market cap.
num_market_pairs	integer	The number of active trading pairs available for this cryptocurrency across supported exchanges.
circulating_supply	number	The approximate number of coins circulating for this cryptocurrency.
total_supply	number	The approximate total amount of coins in existence right now (minus any coins that have been verifiably burned).
market_cap_by_total_supply	number	The market cap by total supply. *This field is only returned if requested through the aux request parameter.*
max_supply	number	The expected maximum limit of coins ever to be available for this cryptocurrency.
date_added	string <date>	Timestamp (ISO 8601) of when this cryptocurrency was added to CoinMarketCap.
tags	Array	Array of tags associated with this cryptocurrency. Currently only a mineable tag will be returned if the cryptocurrency is mineable. Additional tags will be returned in the future.
Quote	Map	A map of market quotes in different currency conversions. The default map included is USD.

Relationship

"**Symbol**" field present in both the APIs which represents the symbol for cryptocurrency is used as a common key to make a join with other two datasets.

3. Website Dataset

Metadata

The metadata for the website dataset has been mentioned below. This contains information about major cryptocurrencies.

Release	Integer	Release year
Currency	String	Cryptocurrency name
Symbol	String	Cryptocurrency Symbol
Founder	String	Founder of cryptocurrency
Hash Algorithm	String	Hashing algorithm
Programming language of implementation	String	Programming language of implementation of cryptocurrency
Consensus mechanism	String	Consensus mechanism
Notes	String	Notes of cryptocurrency

The screenshot from website for few cryptocurrencies from the website is mentioned below.

Before 2013

Release	Currency	Symbol	Founder(s)	Hash algorithm	Programming language of implementation	Consensus mechanism	Notes
2009	Bitcoin	BTC,[2] XBT, ₿	Satoshi Nakamoto[nt 1]	SHA-256d[3][4]	C++[5]	PoW[4][6]	The first and most widely used decentralized ledger currency,[7] with the highest market capitalization.[8]
2011	Litecoin	LTC, Ł	Charlie Lee	Scrypt	C++[9]	PoW	One of the first cryptocurrencies to use scrypt as a hashing algorithm.
	Namecoin	NMC	Vincent Durham[10][11]	SHA-256d	C++[12]	PoW	Also acts as an alternative, decentralized DNS.
2012	Peercoin	PPC	Sunny King (pseudonym) [citation needed]	SHA-256d[citation needed]	C++[13]	PoW & PoS	The first cryptocurrency to use both PoW and PoS functions.

Relationship

"**Symbol**" field present in both the APIs which represents the symbol for cryptocurrency is used as a common key to make a join with other two datasets.

Challenges

Below is the list of challenges I faced while creating the datasets.

1. For CSV file, I am not able to find one single source of truth of file having historical data for the cryptocurrencies. So, I have to extract the data for each cryptocurrency one by one and consolidate all the dataset into one file.
2. In addition, after consolidation, I am not able to find any field to relate with other datasets. So, I had to add a variable called "Symbol" which denotes the symbol of cryptocurrency to establish the relationship
3. Finding a suitable website and getting the HTML object was easy. But scanning through the HTML content and identifying the right tag to start scraping was initially a challenge
4. Finding a suitable website and getting the HTML object was easy. But scanning through the HTML content and identifying the right tag to start scraping was initially a challenge
5. Getting API dataset was bit challenging as there are only few APIs available for free to get the latest price for cryptocurrencies.

Learning

Thorough inspection of the code behind a web page (HTML) and API is important to understand the structure and key tags in the web page. Without understanding of this data structure, coding a program is very difficult.