CRYPTEX Dataset: Fine-Grained Cryptocurrency Pricing Data from 2019 to 2024

Lucas Raicu[1], Stefan Donisa[2], Lucas Ciobanu[3], Lan Nguyen[4], and Ioan Raicu[5]

[1] Glenbrook South High School, Glenview, IL, USA 266520@glenbrook225.org
[2] John Hershey High School, Mount Prospect, IL, USA stefandonisa08@gmail.com
[3] Lisle High School, Lisle, IL, USA lucasciobanu@gmail.com
[4] Illinois Institute of Technology, Chicago, IL, USA lnguyen18@hawk.iit.edu
[5] Illinois Institute of Technology, Chicago, IL, USA iraicu@iit.edu

Abstract. When simulating financial models, fine-grained datasets significantly impact model effectiveness. Our research aimed to experiment with cryptocurrency time-series data, particularly candlestick data from exchanges like Binance. However, obtaining high-quality, one-second granularity data from exchange APIs and even online websites proved difficult, unintuitive, impossible, or expensive. This challenge inspired us to create an efficient Python-based framework that extracts the transaction history for 153 cryptocurrency trading pairs from Binance.us since September 2019. This data is then cleaned and summarized into a variety of sub-datasets ranging from yearly to one-second granularity candlesticks. To enable ease of data-sharing, we published a sample (4-year BTCUSDT trading pair data) dataset on Kaggle and the entirety of the datesets in CSV formatted files (261GB) on a publicly accessible web server that updates nightly.

Keywords: Datasets, Time-series, Candlestick, Binance, Cryptocurrencies, Bitcoin, Kaggle

1 Introduction

Bitcoin [2] is a decentralized digital currency and the first-ever cryptocurrency created in 2009 by an unknown person or group using the pseudonym Satoshi Nakamoto. It operates on a decentralized peer-to-peer network, known as the blockchain, without the need for a central authority or intermediary. Bitcoin represents a transformative innovation in the financial landscape, and its impact has extended beyond the world of finance, inspiring the development of thousands of other cryptocurrencies and driving discussions about the future of money and decentralized technologies.

Like Bitcoin, there are thousands of cryptocurrencies that have been developed with different approaches and different use-cases. Analyzing cryptocurrency pricing data is going to be of high interest to the Fintech community. Getting access to this price data is readily available through both graphs and numerical

datasets (e.g. CSV format). The sources that are available make it incredibly difficult to get OHLC (also known as candle stick data [1]) and volume data at fine-grained resolution below 1-min intervals.

2 Making a Case for Fine-Grained Data

Time-series data represent events or measurements, reflecting cryptocurrency prices over time. Candlestick plots visualize this data concisely (see figure 1), displaying essential values for a specific time period. Smaller candlestick data granularity means more opportunities to buy and sell, leading to greater potential profit.

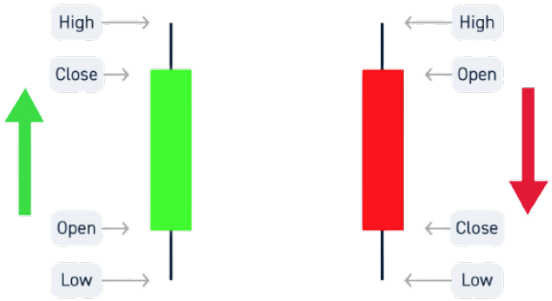

Fig. 1: Candlestick Basics

For example, in Table 1 we explored the BTCUSDT candlestick data from 2021; if one purchased one bitcoin on 1/1/21, the potential profit (assuming an all-knowing oracle that could identify the high and low of every candle stick) showed the highest profit with the 1-second granularity (nearly 400X higher than when using weekly candle stick data).

Table 1: Financial Performance for each Granularity

Period	**Potential Profit**	**Opportunities**
1-sec	$170M	28.9M
1-min	$37M	524K
1-hour	$5.4M	8.7K
1-day	$1.1M	365
1-week	$431K	52

Cryptocurrencies are decentralized digital assets, unlike centralized currencies like the U.S. dollar. Cryptocurrency exchanges, like Binance, facilitate buy-

ing, selling, and trading. Similar to traditional banks facilitating currency exchange, cryptocurrency exchanges act as clearinghouses, facilitating the buying, selling, and trading digital assets. One of these exchanges is Binance; but due to regulatory changes in 2019, only Binance.us is accessible to U.S. residents, and users from outside the U.S. must use Binance.com. To access the real-time cryptocurrency data from Binance, one must send HTTP requests coupled with their API key for authentication.

After a thorough review of the potential sources of data for 1-sec granularity candle sticks, we concluded that there were no sources that offered the 1-sec candle stick datasets for free. The only resource we found was from Binance.com through a web GUI interface. Downloading the BTCUSDT dataset took multiple tries, yielded 48 individual files (1-month period each), and did not separate the US-based transactions from the global transactions found at Binance.com.

3 CRYPTEX DATASET

This project aimed to download transactions (from September 2019 to May 2024) for all 153 trading pairs (see Figure 2) from Binance.us, and then generate candle stick data at various intervals such as 1-sec, 1-min, 1-hour, 1-day, 1-week, 1-month, and 1-year. All transactions and candle stick data is updated nightly for all trading pairs, and made available on a publicly available web server [6] hosted on the Mystic system[4]. We also published the Bitcoin dataset to Kaggle under "Fine-Grained Bitcoin Historical Dataset 2019-2023"[5].

1INCH, AAVE, ACH, ADA, ALGO, ALICE, ALPINE, ANKR, ANT, APE, API3, APT, ARB, ASTR, ATOM, AUDIO, AVAX, AXL, AXS, BAL, BAND, BAT, BCH, BICO, BLUR, BNB, BNT, BOND, BOSON, BTC, BTRST, CELO, CELR, CHZ, CLV, COMP, COTI, CRV, CTSI, DAI, DAR, DASH, DGB, DIA, DOGE, DOT, EGLD, ENJ, ENS, EOS, ETC, ETH FET, FIL, FLOKI, FLOW, FLUX, FORTH, FTM, GAL, GALA, GLM, GRT, GTC, HBAR, ICP, ICX, ILV, IMX, IOST, IOTA, JAM, KAVA, KDA, KNC, KSM, LAZI, LDO, LINK, LOKA, LOOM, LPT, LRC, LSK, LTC, LTO, MANA, MASK, MATIC, MKR, MXC, NEAR, NEO, NMR, OCEAN, OGN, OMG, ONE, ONT, OP, OXT, PAXG, POLYX, POND, PORTO, PROM, QNT, QTUM, RAD, RARE, REEF, REN, REQ, RLC, RNDR, ROSE, RVN, SAND, SANTOS, SHIB, SKL, SLP, SNX, SOL, STG, STORJ, SUSHI, SYS, T, THETA, TLM, TRAC, TUSD, UNI, USDC, VET, VITE, VOXEL, VTHO, WAVES, WAXP, XEC, XLM, XNO, XRP, XTZ, YFI, ZEC, ZEN, ZIL, ZRX

Fig. 2: 153 trading pairs available for download [6]

Our solution involves a 5-stage pipeline (see Figure 3) from historical transaction extraction, summarizing the data in candlestick comma-delimited text format, visualization of the data through interactive plots using plotly library, automation of the extraction process to handle failures, throttling due to lim-

itations of the Binance.us API, incremental daily updates of the datasets, and finally the hosting of the datasets on Kaggle and our own web server.

Fig. 3: Our Solution

Our approach compared to the manual process of downloading this data can be found in Figure 4).

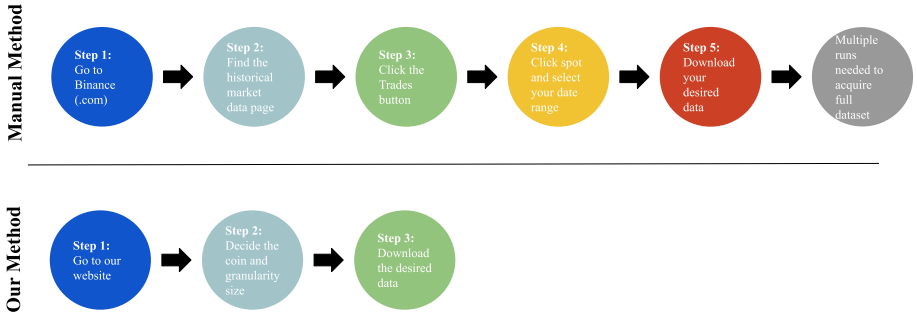

Fig. 4: Dataset download process: proposed solution vs. manually from Binance.com

Finer-grained data provides the necessary precision to build models for high-frequency trading. Automation for transaction history extraction this online data repository to automatically update across all 153 coins while making it easy for users of this dataset to use tools such as wget and curl to easily automate the download of this dataset in its entirety (currently 261 GB).

The raw data has 6 columns, starting with the timestamp (in Unix timestamp with seconds granularity), open, close, high, low, and volume. Each coin (e.g. BTCUSD pair) contains 7 files with different granularity of the candlestick data, from 1 second up to 1 year. Figure 5 shows the first samples in the weekly candlestick raw data.

The candlestick data is built from the transaction data that is shown in raw form in Figure 6.

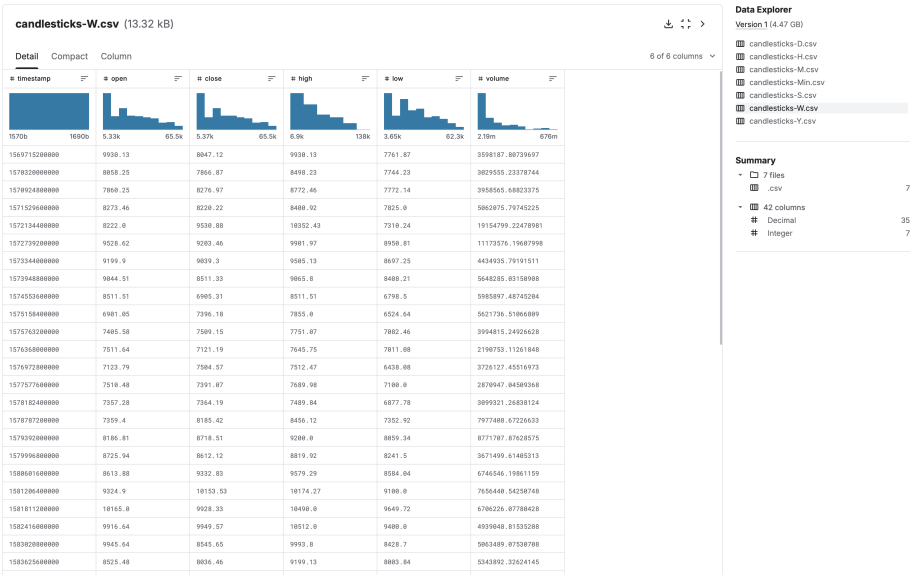

Fig. 5: BTCUSD weekly candlestick data on Kaggle

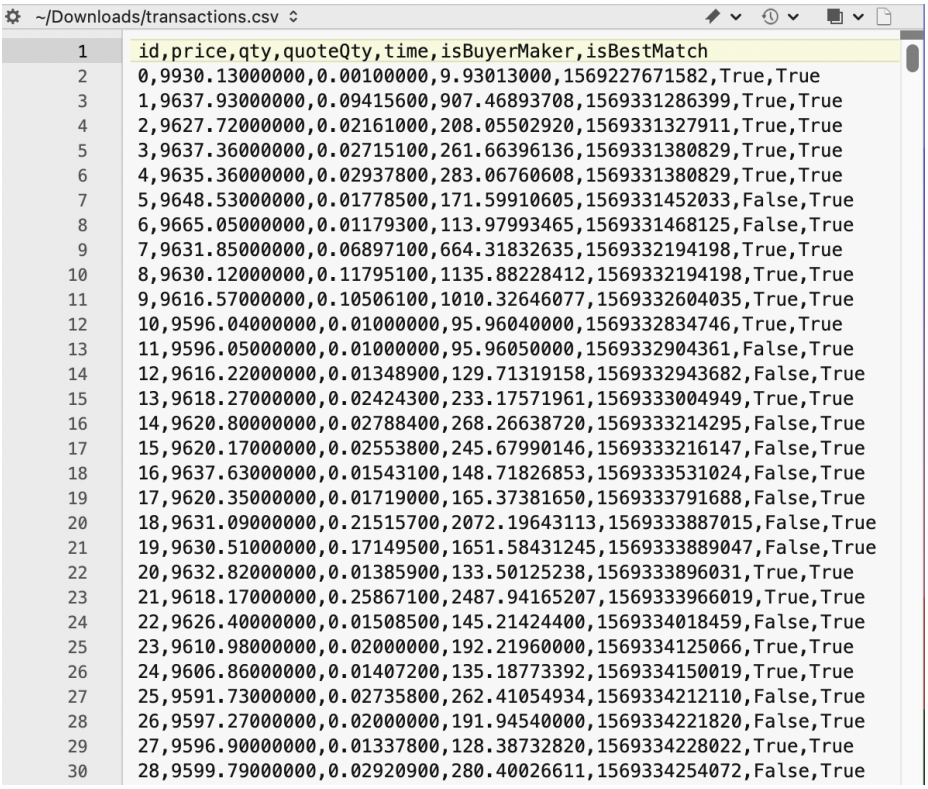

Fig. 6: BTCUSD transactions raw data

Here are some examples of popular crypto showing a summary view of the entire dataset from 2019 - 2024 (see figure 7).

(a) BTCUSD 2019-2024

(b) ETHUSD 2019-2024

(c) BNBUSD 2019-2024

(d) XRPUSD 2019-2024

(e) SOLUSD 2019-2024

(f) ADAUSD 2019-2024

(g) DOTUSD 2019-2024

(h) AVAXUSD 2019-2024

(i) DOGEUSD 2019-2024

Fig. 7: Monthly candlestick data along with volume for 9 cryptocurrencies (BTC, ETH, BNB, XRP, SOL, ADA, DOT, AVAX, DOGE) from 2019 to 2024

4 Evaluation

We initially considered GitHub to host our website and datasets; we found storage limitations (5 GB) insufficient for our data that currently is 261GB large. The framework we built runs continuously and gets daily updates to the datasets, including redrawing the plots. We run the website on the Mystic [4] Cloud at IIT that has 10GbE network connectivity and a 1.2TB RAID5 array spanning 6x SSD storage devices. To share our datasets, we made a website for all our data, utilizing HTML and CSS (see Figure 8).

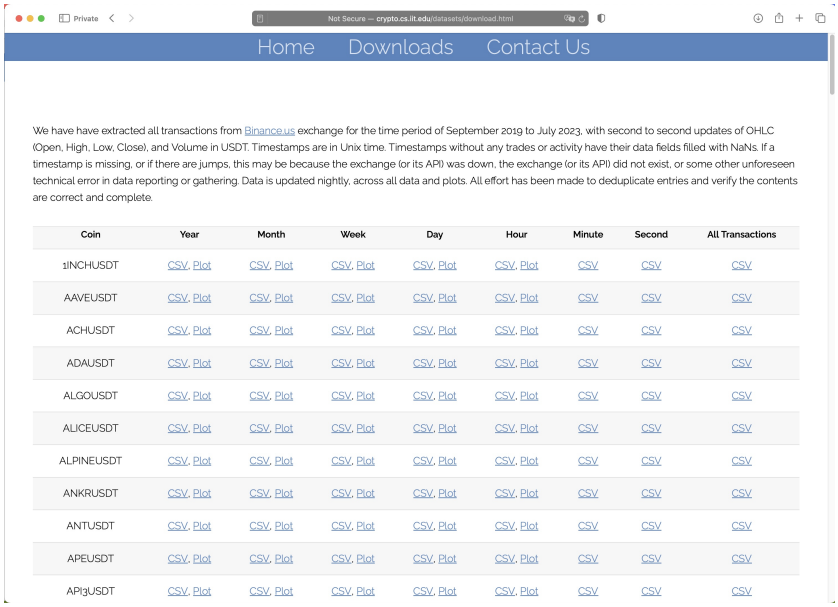

Fig. 8: The Website where all datasets are freely available and downloadable

We also established a page on Kaggle, where data scientists can access a data sample from Bitcoin since 09/2019 (see Figure 9).

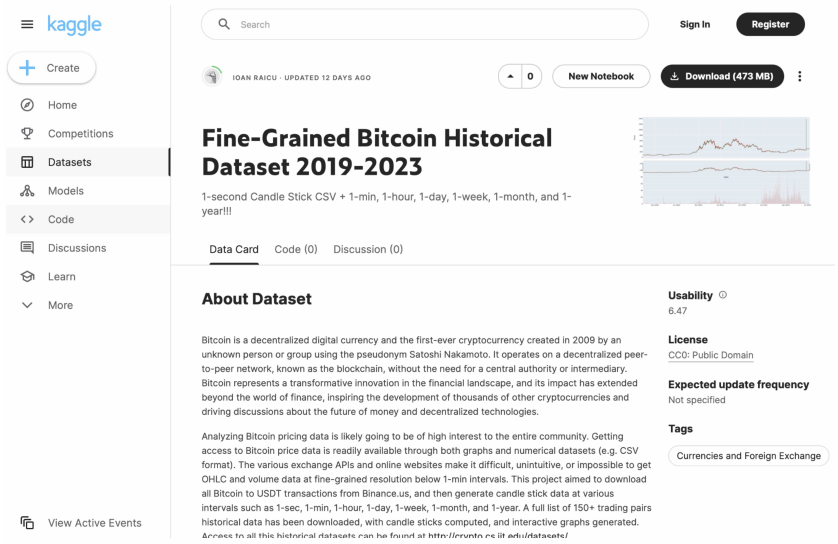

Fig. 9: The Kaggle project that allows a subset of data to be downloaded

We evaluated the total time needed to download 4-years of data on the BTCUSDT trading pair, using different approaches (see Table 2). The manual process from Binance.com took us about 1 hour and several attempts, including the manual merging of 48 files to assemble the final BTCUSDT candle stick data at different granularities. The Binance.us API is the approach we proposed, which involved the implementation of a Python-based framework that extracted all transactions, summarized the data with candle sticks, and generated interactive plots. Our approach took 660 seconds to download the entire 4-year dataset, including the summarizing of the data, and plot generation. Once we have the datasets published, others can get the same dataset using simple tools such as wget in a mere 51 seconds. We are experimenting with a time-series database XStore [3], that allows the efficient data extraction of subsets based on time index if the entire dataset is not required.

Table 2: Comparison of Data Collection Approaches

Source	**Approach**	**# of files**	**Time (s)**	**Info**
binance.com	Manual	48	3,600	User-friendly, long run time
binance.us	Manual	N/A	N/A	N/A
binance.com	API	N/A	N/A	Not possible without VPN
binance.us	API	1	660	Automated, requires programming
mystic.cs.iit.edu/datasets	XStore	1	70	Low latency, slower than wget
mystic.cs.iit.edu/datasets	wget	1	51	Fastest, no date range

5 Challenges, Conclusions, and Future Work

Throughout the project, we encountered multiple challenges. Jupyter Notebook had limitations with large files and rendering large and complex interactive plots, which prompted us to switch to a command-line interface. Additionally, we faced issues with Binance.com, due to access restrictions from a US-based IP address; therefore, we used a REST API to Binance.us and focused on US-based trading pair transactions. To address timezone issues, we used the Python's lambda library.

We have developed candlestick datasets at different granularity (1-sec, 1-min, 1-hour, 1-day, 1-week, 1-month, and 1-year) that are updated daily. These datasets encompass 153 trading pairs, spanning from September 2019 to the present day, and are now accessible to the public through our dedicated Kaggle page and website hosted on Mystic [4]. The entire dataset is accessible at http://mystic.cs.iit.edu/datasets.

In the future, we intend to collect global transaction data from Binance.com and utilize these datasets for backtesting trading strategies. Furthermore, we aim to explore data correlation between trading pairs to improve backtesting accuracy. Furthermore, we are considering the implementation of sentiment analysis to predict a coin's price in the future.

Bibliography

[1] Marshall, B.R., Young, M.R., Rose, L.C.: Candlestick technical trading strategies: Can they create value for investors? Journal of Banking & Finance **30**(8), 2303–2323 (2006)

[2] Nakamoto, S.: Bitcoin whitepaper. https://bitcoin.org/bitcoin.pdf (2008)

[3] Nguyen, L., Raicu, I.: Accelerating crud with chrono dilation for time-series storage systems (2023)

[4] Orhean, A., Ballmer, A., Koehring, T., et al.: Mystic: Programmable systems research testbed to explore a stack-wide adaptive system fabric. In: 8th Greater Chicago Area Systems Research Workshop (GCASR) (2019)

[5] Raicu, L., Donisa, S., Ciobanu, L., Nguyen, L., Raicu, I.: Fine-grained bitcoin historical dataset 2019-2023. https://www.kaggle.com/datasets/iraicu/fine-grained-bitcoin-historical-dataset-2019-2023 (2023), accessed: August 5, 2023

[6] Raicu, L., Donisa, S., Ciobanu, L., Nguyen, L., Raicu, I.: Cryptex dataset. http://mystic.cs.iit.edu/datasets/ (2025), accessed: January 12, 2025