UNIVERSITY OF PISA

DEPARTMENT OF COMPUTER SCIENCE

MASTER DEGREE IN COMPUTER SCIENCE

Mapping the peaks and depths of crypto exchanges with Machine Learning

Supervisors:

Prof. Laura Emilia Ricci

Dr. Damiano Di Francesco Maesa

Presented by:

Luca Santarella

Academic Year 2022/2023

Abstract

The thesis aims to conduct a data analysis regarding cryptocurrency exchanges using statistical tests often used in financial data such as Benford's Law, trade size clustering on round numbers and analysis of the correlation relationship between trading volume and transaction data. We also propose a machine learning approach to detect anomalies in trading volumes with convolutional autoencoders using unsupervised learning, as well as an attempt to normalize the trading volumes reported using web traffic data of the exchanges' websites. The techniques employed can be used as good indicators of potentially fraudulent activities such as wash trading and volume inflation which benefit dishonest cryptocurrency exchanges at the expense of regulated cryptocurrency exchanges which follow mandatory regulatory compliance and by creating a false perception of the cryptocurrency market for customers.

Acknowledgements

I would like to express my sincere gratitude to my advisors Laura Ricci and Damiano Di Francesco Maesa for their invaluable guidance and support throughout my master's program. Their expertise and encouragement helped me to complete this research and write this thesis.

I am grateful to my loving girlfriend Ginevra who supported me and encouraged me during difficult times.

I would also like to thank my friends and family for their love and support during this process, without them, this journey would not have been possible.

Contents

Chapter 1

Introduction

Interest in blockchain technology and cryptocurrencies has grown over the years since the 2008 paper “A peer-to-peer electronic cash system” [1] from the mysterious author Satoshi Nakamoto that marked the beginning of the new blockchain era. The diffusion of cryptocurrencies has reached also ordinary people, who do not possess particular technology expertise, typically through mass media such as TV news, articles, and blogs. An NBC News poll in 2022 [2] reveals that one in five United States inhabitants has invested in, traded, or used a cryptocurrency, meaning 20% of the USA population has indeed come in contact with the crypto world, this data shows how pervasive cryptocurrencies have become over time in our life and that there is an actual heed in the public that needs to be preserved. The public perception of cryptocurrencies can be somehow distorted by the frequent news articles that report crypto scams such as Ponzi schemes, crypto loggers, and fake crypto services to name a few [3], as well as illicit activities that use cryptocurrencies such as money laundering [4], fundraising of terrorist organizations [5], and dark web marketplaces such as the infamous Silk Road which use digital currencies as a payment method [6]. Cryptocurrencies carry a misconception that they are mostly associated with crime, the reality is that a Chainalysis[1] report shows that only a very tiny percentage (0.24%) of all cryptocurrency activity is related to illegal activities [7]. A 2023 poll made by Pew Research Center also shows that the majority of USA citizens do not

[1] https://www.chainalysis.com/

trust that the current ways to trade, invest and use cryptocurrencies are safe, in fact for the 88% of the interviewed who did hear at least a little about cryptocurrencies the 75% of them are not confident or not very confident about the reliability and safety of the ways available to interact with cryptocurrencies.

Cryptocurrency exchanges are platforms that allow their users to buy and sell cryptocurrencies for another digital currency or fiat currency (e.g. Euro, USD, etc.), they act as intermediaries between buyers and sellers, and their role is vital to the crypto ecosystem as they are often the entry point for those who want to trade digital currencies. Numerous news of bankruptcies of crypto exchanges such as the failure of Mt. Gox in 2014 [8], which at the time was the biggest Bitcoin exchange, and more recently in 2022 the bankruptcy of FTX [9] have diminished the credibility of crypto exchanges in the eyes of the public. These bankruptcies are not limited to crypto exchanges but also to crypto service providers such as the crypto lender Celsius in 2022 [10] and the crypto hedge fund Three Arrow Capital that was forced in liquidation in 2022 [11]. Thus, despite being one of the most essential "players" in the crypto industry with their primary function of facilitating the crypto market, they still remain untrusted third parties, that operate in a centralized way which brings potential risks related to the delegation of funds and possible market manipulations.

For this reason, it is vital to be able to distinguish honest crypto exchanges that operate by the book and follow laws and regulations from dishonest crypto exchanges that manipulate the market and try to boost their profits causing unfair competition. Dishonest crypto exchanges perform illegal forms of market manipulation such as wash trading, that is the inflation or fabrication of fake trading volumes by creating fictitious buy and sell orders for a certain coin [12]. Amiram et al. [13] state that the way that an exchange performs wash trading is by creating and processing immediately fake orders created by its own wallets and therefore not placed by real users. Another possible way is to create incentives for users to trade among themselves by eliminating trading fees for "top-tier" trades. The effectiveness of wash trading is examined by Cong et al. that affirm that "liquidity begets liquidity", i.e. crypto exchanges have a strong economic incentive in faking their trading volumes. This is due to the fact that the trading volume of an exchange is one of the primary parameters taken into consideration when choosing an exchange. Therefore, a higher trading volume increases the popularity of the exchange that ranks better on crypto data aggregators such as CoinMarketCap[2] and CoinGecko[3] which, as an indirect result, brings an increase in profits from the transaction fees due to the attraction of more customers.

In this thesis we present our study of this phenomenon. The data collection process was performed with a Python script which uninterruptedly collected transaction data for the major cryptocurrencies pair (BTC/USD, ETH/USD, ADA/USD, XRP/USD and ETH/BTC) in real-time using the WebSocket API of the exchanges Kraken, Gemini, and Binance, representing trusted exchanges, and for the exchanges LBank, Hotcoin, Bitmex, BTSE, and Changelly. Moreover, trading volumes of the top 200 exchanges have been collected hourly from the CoinGecko API. This phase lasted over a year, with trading volumes which ranged from January 2022 to June 2023 and transaction data collected for three months in the time frame April 2023 - June 2023.

Our goal is to use statistical tests and a modern Machine Learning (ML) approach, which is novel to the literature in fake trading data analysis to the best of our knowledge, to understand possible malicious activities, anomalies, and out-of-the-ordinary scenarios in exchanges by examining transaction data and trading volumes, two of the most characterizing features of an exchange. In our experiments, we use trusted exchanges (on the criteria of U.S. regulation) as a reference group for the values of the tests which are then compared to the rest of the exchanges. We analyze the volume partition of the pairs examined in the crypto exchanges to understand if there is a correspondence with the cryptocurrency dominance of the crypto market. For most of the exchanges, the most traded coin is Bitcoin followed by Ethereum, the partition of traded volume of these two coins reflects the average BTC and ETH dominance of the market, except for the exchange BTSE which has a bigger market for ETH with respect to BTC and the exchange LBank which has a comparable amount of trades among the pairs BTC-USD, ETH-USD, ADA-USD, and ETH-BTC.

Benford's Law describes the distribution of the first significant digit in naturally generated numerical data sets, it states that the probability of a number starting with a certain digit

[2] https://coinmarketcap.com

[3] https://coingecko.com

decreases logarithmically as the digit increases. This law can be used as a statistical test to understand if the trades reported are human-generated or not, the result is that exchanges such as LBank, BTSE and Hotbit present a significative deviation from Benford's Law, signaling possible anomalies.

Next, we analyze the Pearson correlation between trading volumes and the number of trades for each one of the exchanges, the idea is that these two attributes usually have a positive correlation in a regular market. We find out that the exchanges LBank, HitBTC, and Changelly have a moderate trading volume-transactions correlation, whereas Hotcoin has a low correlation, the remaining exchanges show a high or very high correlation instead, as anticipated.

Real transaction data made by actual users should present accumulations of trades on round amounts, this is due to the human need for cognitive reference points, in fact, a person is more inclined to trade a quantity with a round number, e.g. a quantity which has 0 as the ending digit. In order to quantify the difference between the number of trades with round numbers with respect to unrounded numbers we used the Student's t-test. The results show that the exchanges Bitmex, Hotcoin, HitBTC, and Changelly surprisingly do not show significant differences between the number of trades with round and unrounded numbers.

Finally, we try to estimate real trading volume by using web traffic data about the website visits of the exchanges as an indicator of the popularity of the exchange. A reference group of trusted exchanges is used to determine a baseline of volume per visit, then this value is multiplied by the number of visits of the top 50 exchanges by trading volume, the result obtained is an adjusted volume that relies on web traffic data, then the ratio adjusted volume to reported volume is calculated for every exchange. We discover that 32% of the top 50 exchanges examined showed a strong indicator of possible dishonest practices.

We create a convolutional autoencoder model which is used to perform anomaly detection on the trading volumes. The model is trained on a dataset that does not contain anomalies and then it is applied to new data instances to reconstruct the trading volume, if the reconstruction error is greater than a threshold then the data point is considered an outlier. Mapping anomalies of the trading volumes could be crucial to identify possible unexpected scenarios or to gain insight into the particular exchange. A baseline trend made by trusted exchanges is computed and used to detect "legitimate" anomalies that were mapped with price movements of Bitcoin, these data points were discarded as they are not considered significant anomalies. The result of the detection anomaly shows that Coinstore and WhiteBIT are among the exchanges with the highest number of anomalies. Web traffic data is once again used to determine if the sudden increases in trading volume are justified, in the case of Coinstore we find that the increase in website visits, which happened six days after, could have been caused by the increase in trading volume reported, meaning that the higher reported trading volume attracted more users due to a better ranking on websites like CoinMarketCap and CoinGecko and an increase in popularity. On the other hand, WhiteBIT shows a remarkable increase in trading volume which was backed by an increase in website visits, furthermore, we are able to also identify the cause of the following uptrend which was mostly caused by a real-life marketing campaign.

Although the techniques that have been employed do not give direct evidence of market manipulations they can be used as good indicators for possible situations of market manipulation conducted from the crypto exchanges. The failure of various tests could signal a potentially suspicious activity from an exchange, which could be addressed with further detailed analysis and more close inspections regarding the exchange activities. In our analysis, our findings identified the exchanges LBank and HitBTC as strong suspects of wash trading due to failing three out of five tests presenting numerous anomalies identified in the transaction data as well as in the trading volumes.

The rest of this thesis is structured as follows: Chapter 2 gives a background on blockchain technology, Chapter 3 defines crypto exchanges and explains how they work, Chapter 4 gives an overview of the most prominent work on data analysis in crypto exchanges for dishonest practices, Chapter 5 details the data collection phase, Chapter 6 explains the techniques adopted and their results, and finally Chapter 7 draws the final conclusions regarding the results obtained.

Chapter 2

Blockchain Technology

A blockchain is a decentralized and distributed ledger made by an ever-growing list of records (transactions) that is immutable, append-only and secured by cryptographic proof. The central concept of a blockchain is that the consensus reached among the participants of the distributed network of peers is trustless. This means that the transactions recorded on the ledger are agreed upon by everyone without relying on a central authority or a third party.

2.1 Bitcoin

Blockchain technology is mainly known for its most famous use case of cryptocurrencies, the most notorious one being the cryptocurrency Bitcoin, which was proposed by a mysterious author or group of authors under the pseudonym of Satoshi Nakamoto in a whitepaper in 2008 [1]. Satoshi proposed in their paper a payment protocol system that implements a public, permanent and decentralized ledger, the idea was to create a peer-to-peer version of electronic cash that would allow payment transactions between two parties without the need for a central authority or a financial institute like a bank. The term Bitcoin can be abbreviated as BTC or with the sign ₿ which is usually used to refer to the monetary unit of the cryptocurrency.

The main aspects of the protocol are the address used in the protocol, which is an identifier of the user in the distributed network (note that a person can own multiple addresses), the wallet, which is the set of credentials (private and public keys) used to manage and access the funds of the user, it is often a software application or a hardware device but it can be also a paper wallet (simply a sheet of paper where are printed the keys). The wallet should always be kept secure because it provides ownership and control of the funds, meaning that the loss of the credentials will result in a loss of the bitcoins associated with the wallet. Transactions are a fundamental part of the Bitcoin protocol, they are defined as the exchange of funds between two users and they are represented as digital signatures that prove the ownership of bitcoins and specify the recipient's public address (see Figure 2.1). These transactions are verified in a process called mining, which is the set of operations for the maintenance of the blockchain, the process consists in verifying the transactions and then adding the block to the blockchain by demonstrating a Proof-of-Work (PoW), which involves searching for a value that when hashed, with the hashing algorithm SHA-256, the hash begins with a certain number of zero bits. The miners, that are participants of the Bitcoin network that use powerful computers to solve the Proof-of-Work problem are then rewarded for adding a new block with an amount of bitcoin which is halved every four years (the reward is currently 6.25 bitcoin and there have been three halving events on November 28 2012, July 9 2016, and May 11 2020).

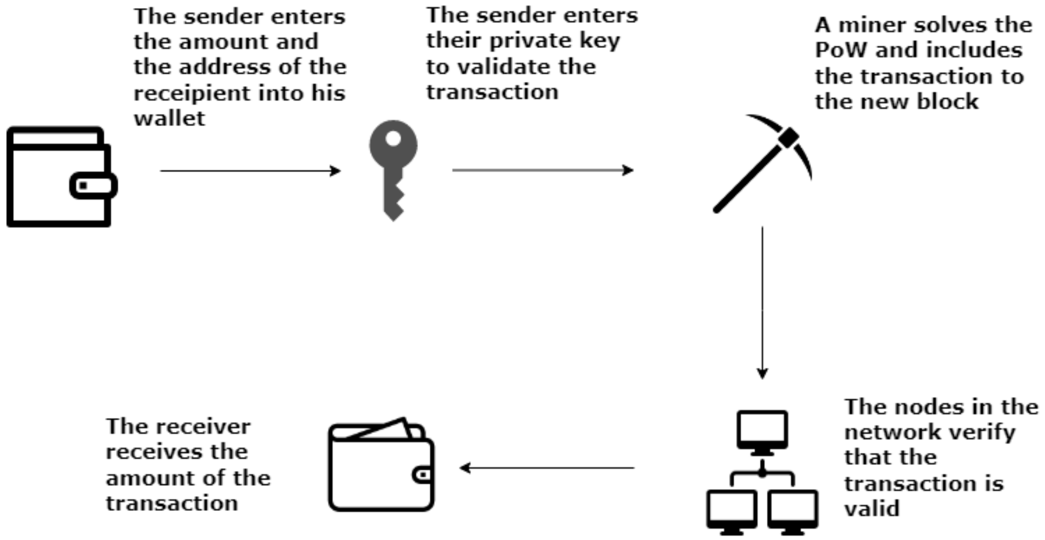

The sender enters the amount and the address of the receipient into his wallet
The sender enters their private key to validate the transaction
A miner solves the PoW and includes the transaction to the new block
The nodes in the network verify that the transaction is valid
The receiver receives the amount of the transaction

Figure 2.1: A scheme of a Bitcoin transaction.

Nakamoto proposed that the network should run as follows:

1. New transactions are sent to every node in the network in broadcast.
2. Each node collects new transactions into a block.
3. Each node works on finding a proof-of-work for its block, which is deliberately difficult.
4. When a node solves the proof-of-work, it broadcasts the block to all nodes, and the node receives a reward for finding the proof-of-work.
5. Nodes accept the block only if all transactions in it are valid and not already spent.
6. Nodes express their acceptance of the block by working on creating the next block in the chain, using the hash of the accepted block as the previous hash.

Bitcoin has a fixed limited supply, this is an essential feature that distinguishes it from traditional fiat currencies (such as Euro, U.S. dollars, etc.), in fact, the total number of bitcoins that can ever exist is capped at twenty-one million. The limited supply of bitcoins creates a sense of scarcity as it happens for other goods such as gold (Bitcoin is often called the "digital gold"), this scarcity along with an increasing demand can lead to a potential increase in the value of Bitcoin over time. The Bitcoin protocol offers a higher level of privacy w.r.t the usual financial systems, although payments are publicly available to anyone in the network, the identities of the participants are not explicitly exposed, instead, the addresses are used to represent users' identities. It is important to note that all transactions associated with an address could technically be linked to a specific individual using techniques such as blockchain analysis and data correlation to analyze transaction patterns and reveal the link between an address and the user identity, as it happened with the data analysis company Chainalysis[1], that used its tools to identify, together with the Israeli National Bureau for Terrorist Financing (NBCTF), 40 addresses involved in financing terrorist-type activities [14].

[1] https://www.chainalysis.com/

2.2 Ethereum

The second most famous digital currency is Ethereum which expanded the concept of blockchain beyond the simple payment protocol and store of value as Bitcoin and it introduced the concept of smart contracts which are used to build decentralized applications (DApps). Ethereum was proposed in 2014 by Vitalik Buterin in his whitepaper "A next-generation smart contract and decentralized application platform" [15] and later created by a team of programmers in 2015.

The main difference between Ethereum and Bitcoin resides in the fact that Ethereum is programmable, so it is possible to build and deploy decentralized applications on its network [16]. A smart contract is a self-executing computer program that lives on the Ethereum blockchain, that automatically executes predefined actions or agreements when certain conditions are met, i.e. they are triggered by a transaction from a user (or another contract). A huge feature of the Ethereum blockchain is the Turing-Completeness, which means it can solve any computational problem given enough resources and time. This property allows programmers to develop solutions on the Ethereum blockchain leveraging the Ethereum Virtual Machine (EVM), which is a runtime environment that serves as the execution engine for smart contracts, the EVM is designed to act as one single entity maintained by thousand of connected computers running an Ethereum client, this allows to execute smart contracts across all nodes in a consistent way. The consensus mechanism adopted in the Ethereum blockchain used to be the Proof-of-Work system which worked similarly to the one present in the Bitcoin protocol but since 15 September 2022 with the so-called "The Merge", Ethereum migrated to a Proof-of-Stake (PoS) system [17]. PoS relies on validators who hold and "stake" their Ether (monetary unit in Ethereum) to secure the network and validate transactions. The transition to PoS has been done for efficiency and energy-saving reasons, as the Ethereum Foundation stated, the transition reduced energy consumption by 99.95% [17].

Ethereum offers a wide community of developers and decentralized applications which offer a vast range of services and functionalities, such as:

1. **DeFi**

   Decentralized Finance (DeFi) is one of the most promising kinds of DApp, platforms such as Compound[2], Aave[3] and MakerDAO[4] provide decentralized lending and borrowing services and financial services such as speculation on price movements using derivatives, trading of crypto in a decentralized way, insurance against risks and earning interest in a savings account. More generally speaking, DeFi offers financial instruments (monetary contracts between two parties) without relying on traditional intermediaries such as banks, brokerages, conventional exchanges and financial institutions. The decentralized nature of this kind of services allows for an open and permissionless environment without the classical eligibility criteria needed from banks (such as a good credit score). Another considerable benefit lies in the interoperability among these services, since they are standardized in a smart contract it is possible to have composability of these services which can be exploited to create new and innovative financial applications.

2. **Fungible Token**

   Smart contracts allow the creation of ERC-20 (Ethereum Request for Comments 20) tokens, which is a token standard that implements an API (Application Programming Interface) for tokens within smart contracts. An ERC-20 token is a type of digital asset that is fungible, meaning that they have a property that makes each token exactly the same (in type and value) as another token, they can represent a wide range of digital assets, e.g. utility tokens, security tokens, stablecoins (coins that are pegged 1:1 to another asset such as the U.S. dollar, USDT is the most famous stablecoin).

3. **NFTs**

   An NFT (Non-Fungible Token) is a type of digital asset utilizing standard ERC-721 and ERC-1155. As the name suggests, this kind of token is non-fungible meaning that these types of tokens are unique and identifiable, i.e. they are not equally exchangeable as for the traditional ERC-20 tokens. NFTs are used to represent ownership and proof of

[2] https://compound.finance

[3] https://aave.com

[4] https://makerdao.com/it/

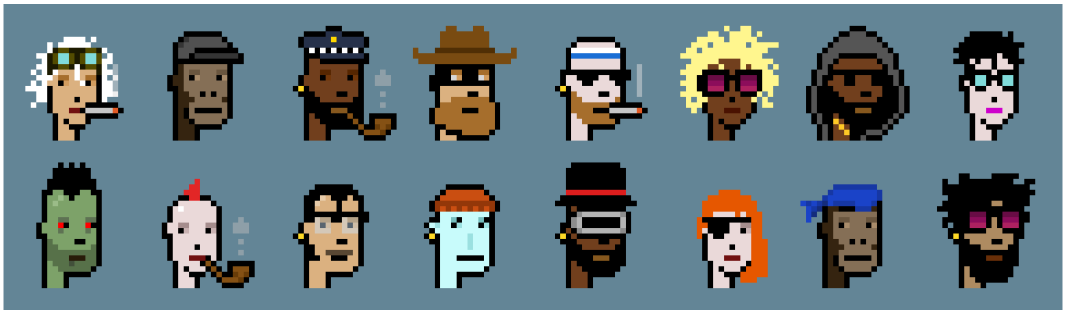


Figure 2.2: The Cryptopunks NFT collection.

authenticity of a unique item, the ownership is recorded on the blockchain and can be transferred by the owner, in this way NFTs can be traded or sold between two parties. One of the most common use case of NFTs is digital art, a major example is the collection of CryptoPunks (see Figure 2.2) released in 2017 by Larva Labs[5] which reached a floor price of around 60 ETH in November 2022 [18].

4. **Governance Dapps**

   Governance DApps are applications that enable users to vote on proposals and it allows decentralized decision-making giving access to community discussion boards where new proposals can be submitted and voted by the community of a certain DApp. The governance is token-based, meaning that individuals hold tokens that grant them voting and decision power over a proposal, usually the voting system is proportional to the number of tokens held by users. The decision system is implemented into a smart contract facilitating voting, proposal submission and making it more reliable and transparent the outcome of the voting process. One example of a governance-based system is a DAO (Decentralized Autonomous Organization) which is an organization that operates without a hierarchical structure as it happens in most traditional organizations but instead leverages the decentralized nature of the governance process.

5. **Gaming**

   The gaming world is another interesting field that has been affected by the diffusion of smart contracts. In-game items now can be represented on a blockchain with the

[5] https://www.larvalabs.com/cryptopunks

possibility of introducing important features such as transparency, real ownership and programmability into game mechanics. The ownership of a certain item can be now represented as NTFs and in-game currency can be represented as an ERC-20 token, giving an incentive to play for gamers. Play-to-earn games leverage this concept by creating a tokenized economy within the game, players are rewarded with tokens (which have real-life value) when reaching certain achievements. Furthermore, the usage of smart contracts can ensure fairness within games giving proof of real randomness in random number generation which allows players to verify that certain mechanics inside the game are actually fair.

2.3 Blockchain Applications

The adoption of the blockchain concept has spread in many other areas [19], other than finance, such as:

1. **Supply Chain**
   In this context the supply chain management system, which logs the movements of a certain good or product, can be implemented as a digital ledger database that has the properties of being transparent, reliable and can be agreed on without a third party involved, all of these properties make it suitable for auditing purposes in an open distributed system. One of the various sectors in the supply chain area which can benefit from this kind of technology is the food processing supply chain for the traceability of food with many systems that already have deployed a solution on the market, e.g. the supermarket chain Carrefour pioneered this field in 2018 with a food-tracking system based on blockchain technology [20].

2. **Internet of Things**
   The Internet of Things (IoT) field can also benefit in terms of data security because a blockchain can provide a distributed, incorruptible and tamper-resistant ledger that can be used to manage transactions or logs in a distributed and trustworthy manner, with a

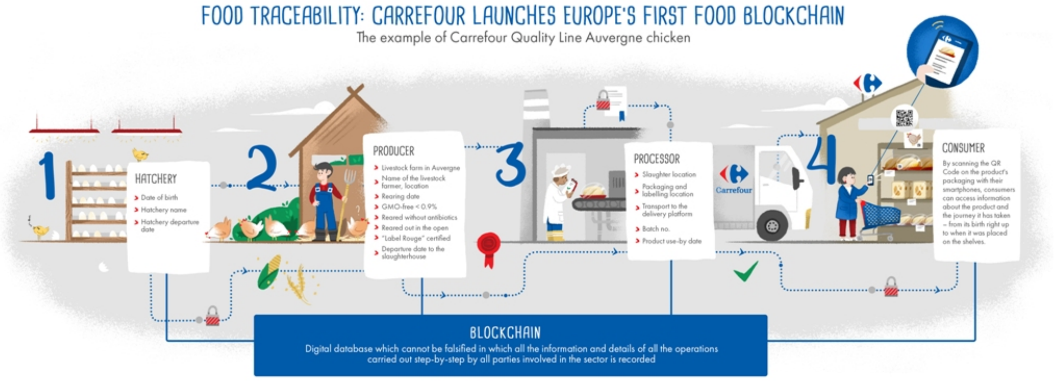

FOOD TRACEABILITY: CARREFOUR LAUNCHES EUROPE'S FIRST FOOD BLOCKCHAIN
The example of Carrefour Quality Line Auvergne chicken
1
HATCHERY
Date of birth
Hatchery name
Hatchery departure date
2
PRODUCER
Livestock farm in Auvergne
Name of the livestock farmer, location
Rearing date
GMO-free < 0.9%
Reared without antibiotics
Reared out in the open
"Label Rouge" certified
Departure date to the slaughterhouse
3
PROCESSOR
Slaughter location
Packaging and labelling location
Transport to the delivery platform
Batch no.
Product use-by date
Carrefour
4
CONSUMER
By scanning the QR Code on the product's packaging with their smartphones, consumers can access information about the product and the journey it has taken – from its birth right up to when it was placed on the shelves.
BLOCKCHAIN
Digital database which cannot be falsified in which all the information and details of all the operations carried out step-by-step by all parties involved in the sector is recorded

Figure 2.3: Food Traceability (image by Carrefour Group).

focus on data integrity and reliability. In fact, data sensors can easily be integrated with the blockchain with a system that allows sensors to directly write on the blockchain [21].

3. **Healthcare System**

   In the medical field, there are devices and technologies which can be used to track and monitor the health condition of patients by reading the human body attributes. This data can be used in a blockchain to preserve the privacy of patients and store it in a ledger format, the idea is that a smart contract can be written in such a way that some terms and condition is applied when certain data is available, in this way the smart contract can trigger some events based on the terms. More generally, blockchain technology can improve interoperability, through the data exchange between healthcare providers and the security of healthcare data by securely storing patient records.

4. **Identity Management**

   Identity management solutions are systems that are designed to manage and control digital identities in an organization or a general system. Also in this case, the use of a blockchain can facilitate the management of digital identities, users can have control over their personal data (self-sovereign identity) [22] and decide to whom to share it, reducing the risk of identity theft and fraud. This system can simplify and reduce the time spent in KYC (Know Your Customer) procedures for user verification.

Chapter 3

Cryptocurrency Exchanges: An Overview

3.1 Cryptocurrency Exchanges

Cryptocurrency exchanges are online platforms that allow users to buy, sell and trade cryptocurrencies. They are a fundamental part of the cryptocurrency ecosystem because they provide a way for people and companies to exchange their digital assets for other cryptocurrencies or fiat currencies (traditional money such as euros or U.S. dollars).

A crypto exchange works in two possible ways: as a matching platform, which acts as an intermediary between buyers and sellers and in return it profits from a transaction fee from both parties or as a market maker which is defined as "Any intermediary who creates a market for a financial obligation" [23], that financially benefits from the bid-ask spread, which is defined as the difference between the bid (buy orders) price and the ask (sell orders) price, [24]. In the case of a matching platform, the exchange works through an order book, which is an electronic list of buy and sell records from users, that is used to keep track of the bid price, which is the maximum price that a buyer is willing to pay for a certain asset and the ask price, which is the minimum price that a seller is willing to take for a certain asset [25].

The trade is executed when a buyer is willing to buy a determined asset by paying the best ask price or a seller is willing to sell at the best bid price, in order-book based exchanges there are two main types of orders: market orders, which are orders that are processed immediately by buying or selling at the market price of the asset, whereas, for the limit orders the trade is executed when a certain price is reached for the asset [26]. Exchanges that act as market makers profit from the bid-ask spread by playing the role of intermediary, e.g. they quote a bid price of 100 units which is the price that they are willing to pay to buy the asset and an ask price of 110 units which is the price that they are willing to accept for the sale of an asset. Thus, the bid-ask spread is computed as the difference between ask price (110 units) and bid price (100 units), which represents the margin profit of the exchange.

Using a cryptocurrency exchange is similar to using a traditional stock exchange, users must first create an account, which typically involves providing personal information and undergoing a KYC (Know Your Customer) process which is a legally required verification process of the users and the origin of the funds used in the platform. Once the account is set up, users can deposit funds using a variety of methods, such as bank transfers, credit card payments or crypto transfers. These funds can then be used to buy cryptocurrencies at the current price of the market or they can be used to place an order limit which is an order that goes through when certain conditions are met (e.g. the price of the coin reach a certain price or variation).

3.2 Cryptocurrency Exchange Features

There are many different cryptocurrency exchanges that have their own unique features and services provided, some exchanges only focus on crypto-to-crypto trades and provide advanced features for trading while others are more open to new and/or occasional users. The main characteristics to consider in a crypto exchange are the following:

1. **Type of Exchange**

   The nature of the exchange which can be a traditional CEX (Centralized EXchange) where there is a single entity, such as a company, behind the exchange. These are the most common type of exchanges, they don't require any technical background to use. CEXs usually do not charge anything for the registration but they take up a fee for every transaction, whether it is a buy, a sell or a withdrawal from the account. The other type of exchange is DEX (Decentralized EXchange), where the exchange is programmed as a smart contract on a blockchain such as Ethereum, in this case, users can trade with each other without intermediaries and in complete transparency.

2. **Trading Volume**

   One of the most important aspects of an exchange is the trading volume, it expresses the total amount of assets traded in the last 24 hours (usually expressed in USD) for the exchange. This value can be used as a metric to quantify the health and popularity of the exchange.

3. **Fees**

   One of the factors to consider is the fee charged by the exchange for their services, the fee is usually a percentage of the transaction value, a flat fee or a combination of both. Fees are usually structured based on the maker/taker system, meaning that the exchange sets a different fee based on who creates a new order (maker) and who fills up the order (taker). In many cases, the fee for the transaction is around 0.1 - 0.2 % but many exchanges offer a fee tier system based on the volume traded by the user, e.g. if the user achieves a certain amount of volume traded the exchange will lower the fees for the next trades.

4. **Number of assets**

   The number of assets is another important metric to evaluate the exchange, users will prefer exchanges with a high amount of markets, i.e. pair of assets that the exchange will let you trade.

5. **Reputation**

   The reputation of an exchange is a somewhat more subjective metric, websites such as CoinMarketCap[1], CoinGecko[2] provide a score based on web traffic, average liquidity,

[1] https://coinmarketcap.com

[2] https://coingecko.com

confidence that the volume reported by an exchange is legitimate. Other factors that affect the reputation of an exchange are the security measures taken, the compliance with the authorities regarding the regulations and licenses, and the reviews of the users posted on forums and social media can also give an overall picture of the exchange.

6. **Security**

   Security is another important aspect to keep in consideration when choosing a cryptocurrency exchange. Since cryptocurrencies are often the target of hacks and scams, it is crucial to select an exchange that has strong security measures in place, this may include things like two-factor authentication, cold storage of assets, and frequent audits. State-of-the-art security measures in handling the funds of users are crucial for a reputable exchange. Typically a system of cold wallets, which are a type of cryptocurrency wallet that securely stores the private keys offline, are used to store the majority of funds. The exchange needs to have security measures also to prevent theft of funds from the users, such as the usage of two-factor authentication or measures to inform users about possible phishing attacks.

7. **Proof of Reserve**

   Many exchanges try to be more transparent about their funds by providing a Proof of Reserve which is an independent audit conducted by a third party that makes sure that the exchange holds the assets it claims to on behalf of its clients. The reserve data provided can be an audit made by a third party that verifies that the reserves of the exchange match the funds of the users or by providing the on-chain addresses containing the financial reserves for the exchange.

3.3 Decentralized Exchanges

The introduction of smart contracts and programmability into the Ethereum blockchain has given the possibility to build decentralized exchanges, which are a type of cryptocurrency exchange platform that allows peer-to-peer transactions without a central authority as it happens for centralized exchanges. A major benefit of using a DEX lies in the fact that the user has full control over their funds, this is due to the nature of the smart contract which holds and controls the assets being traded and automatically executes trades when certain conditions are met. In this way there is no need to transfer the funds to the exchange before the trade takes place, thus, reducing the risk associated with using CEXs such as loss or thefts of funds [27], cyber attacks [28], scams [3] or a possible bankruptcy of the crypto exchange as it happened in November 2022 with the crypto exchange FTX [9]. Another critical aspect in favor of decentralized exchanges is that the pseudo-anonymity of the blockchain is still preserved when directly trading on-chain, as opposed to the use of centralized cryptocurrency exchanges which need to gather your personal information and implement KYC procedures by law. Overall, DEXes provide important benefits such as a lower counterparty risk (the user has still control over their funds without relying on a central authority), potentially lower fees and a more vast array of trading pairs whose liquidity would not be sufficient to be listed on centralized crypto exchange [29]. A DEX (we use Uniswap[3] as an example) operates as an automated liquidity protocol powered by a constant product formula $x * y = k$ which is kept balanced through the trades. In this formula, k is the product, that must remain invariant, of the pairs' reserve balances of the tokens x and y. For every pair of tokens in the DEX there is a liquidity pool made up of the reserves of the two tokens of the pair (e.g. UNI-ETH will have a liquidity pool made up of ETH tokens and UNI tokens and it will mint UNI-ETH). Every user of the DEX can deposit tokens into the liquidity pool (LP) and become a liquidity provider, by doing so the depositor will receive a certain amount of pool tokens, which are tokens that track the shares of LP and that can be redeemed at any time. The smart contract that regulates the pair of tokens will act as an automated market-maker that will preserve the constant formula. The fee applied by the Uniswap protocol is 0.3% which is added to reserves, this affects the formula because the invariant k will increase for every trade. The fee generated for every transaction is divided among the liquidity providers which can use their liquidity tokens (that were generated when they deposited liquidity into the LP) to get their shares by burning them, since the liquidity tokens are effectively ERC-20 tokens, they can also be traded or sold normally.

[3] https://uniswap.org

Chapter 4

Wash Trading

The great importance of the role of crypto exchanges in the market lead to a raising interest in the academic world regarding these platforms and how they affect the crypto market. The topics that are more frequently discussed in the literature are the following: scams and risks associated with crypto exchanges, forecasting of the cryptocurrency exchange rates and analysis of the competition of crypto exchanges with a focus on fake trading practices. The following are the most prominent papers concerning the analysis of crypto exchanges in the context of fake trading.

1. **Crypto Wash Trading** [30]

   In this paper, Cong et al. work on the phenomenon of wash trading in the context of crypto trading. This deceiving technique is mainly employed by unregulated crypto exchanges which try to boost the real amount of their volume traded or by directly fabricating totally false volumes. The economic benefits of wash trading for exchanges are multiple, as the authors say, "liquidity begets liquidity", for this reason, crypto exchanges inflate trading volumes to have a better ranking on aggregator websites such as CoinMarketCap, CoinGecko, Nomics, and also to create a better image of the company on various forums such as Bitcointalk or Reddit. As a consequence, the more popular and famous the exchange becomes the more users are attracted to it, the main profits of the company operating the exchange are through the transaction fees, for this reason having more users generates more volume and consequently more profits. The authors use structured statistical tests which are performed on data about the trades made on the platforms. The analysis aims to compare the trading activity done in the exchanges examined with statistical and behavioral patterns which are well-known in trading. The total amount of exchanges taken in exam is 29, these are divided into regulated and unregulated and also in two tiers based on the web traffic of the website which somehow represents the popularity of the exchange.

One of the techniques used is Benford's Law, which is a well-known statistical benchmark that is used to detect fraud in macroeconomic, accounting and engineering fields[31]. Benford's law describes the distribution of the first digit in naturally generated datasets as more likely to be a small number. In general, the law states that the probability of the first digit of a number being a particular digit d is given by: P(N is the first significant digit) = $log_{10}(1+N^{-1})$, for example, the probability of 1 being the first significant digit is 30.10%. This analysis of the distribution of the first significant digit was performed on the transactions of each exchange, the result is that all the regulated exchanges re-

$digit\, d$	$P(d)$
1	30.10%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

Table 4.1: The probability of the digit decreases as the digit increases.

spect Benford's Law whereas nine of the unregulated exchanges present some form of discrepancy in the distribution which indicates a form of wash trading.

The second analysis technique performed by the authors is the trade size clustering test, which consists in finding out if the transactions on the exchanges tend to cluster on round prices, e.g. in the paper, the base unit (smallest unit) for BTC transaction is set to 10^{-4} BTC (which at the time corresponded to approximately one dollar), clusters of transactions should appear on multiples of the base unit. Authentic trades made on the exchange will have these clusters on round prices since it is a behavior generally observed by traders in financial markets. The results show that the regulated exchanges once again pass the test showing clusters on the multiples of the base unit, some of the unregulated ones show abnormal patterns.

The third test examines the tail of the trade size distribution of the transactions of the exchanges taken in exam, the goal is to check whether they fit the power law which is usually observed in financial data. Also in this case regulated exchanges behave as the power law predicts whereas some of the unregulated exchanges fail to follow this distribution.

Finally, the authors try to quantify the volume of fake trading in the cryptocurrency market, the approach is to use rounded trades which tend to be authentic human trades as indicators of real volume and unrounded trades which typically represent programmed trades made by bots to indicate wash trading volume (the authors take into account also legitimate algorithmic trading using a benchmark ratio). Since the ratio of rounded trades/unrounded trades should estimate the ratio of authentic trades/fake trades, this value can be used to quantify the wash trading. In this methodology, the authors also quantify the level of roundness of the trades, which is a qualitative parameter that takes into account the trade size, e.g. 1.01 BTC will have a higher level of roundness than 2.123 BTC. In this way they are able to compare the level of roundness on regulated and unregulated exchanges, thus identifying authentic unrounded trades which are discarded in the estimation of wash trading volume. The final estimation is that more than 70% of the traded volume reported by unregulated exchanges is made of wash trades, the authors conclude that more popular and well-known exchanges tend to have low or no wash trading activities at all, whereas less prominent exchanges have a more financial incentive to engage in these wash trading activities.

2. **Direct Evidence of Bitcoin Wash Trading** [32]

   Aloosh et al. provide direct evidence of "fake volumes" in crypto exchanges in contrast to the most popular techniques in detecting wash trading which rely solely on indirect estimation. The detection and characterization of wash trading is made by using leaked internal data which is used as "ground truth" of the now closed bitcoin exchange Mt. Gox. This data is made of 16 million records of buy and sell transactions that took place when the exchange was in activity, which represents 8 million trades with a unique trade ID, the authors found out that 115 thousand of these have the same trader ID on both sides (buy and sell). The authors discuss how important it is to verify whether the accusations of fake volumes are indeed true. This would allow to create a discussion on regulations that are applied to crypto exchanges and it would make it possible for the consumer to have a fairer comparison among centralized crypto exchanges. The researchers evaluated various indirect estimation techniques using this internal data. They discovered that the techniques which use Benford's Law, trade size clustering, deviations from the log-normal distribution and finally the E-Divisive with Medians (EDM) [33] which is a technique to reveal structural breaks, i.e. mean shifts or trend shifts in a time series, are all good indicators of wash trading activity. They also demonstrated how indirect estimation techniques based on power law tail distributions are not reliable estimation techniques. The authors conclude by analyzing the impact of wash trading on the crypto market, finding out that wash trading has a small impact on Bitcoin returns and market liquidity but a significant brief increase in fee revenue for the exchange.

3. **Competition and Product Quality: Fake Trading on Crypto Exchanges** [13]

   The authors investigate how the manipulation of information, such as the trading volumes reported, affects the competition among crypto exchanges and what are the short and long-term effects of such practice. In order to do this, statistical tests and machine learning techniques are used to attempt to uncover possible manipulation and (fake) trading practices from crypto exchanges. Moreover, the paper discusses how the decision to mislead customers about product quality involves a trade-off between short-term gains and long-term reputation damage.

One of the types of analysis used is the EDM algorithm, which can identify structural breaks in trading volume and transaction data. This algorithm, which is the base for the BreakoutDetection algorithm from Twitter [34], is able to detect unusual patterns in trading and most importantly is robust to the presence of short-lived abnormalities in the series, which makes it suitable for cryptocurrency data which commonly tends to have frequent peaks despite being a naturally generated dataset. Another measure used to detect wash trading is the deviation of the frequencies of the first significant digit from Benford's Law. A deviation from Benford's Law, which models the distribution of the digits of a naturally generated dataset, may indicate abnormalities in the data series. The authors use the MAD (Mean Absolute Deviation) measure between the theoretical Benford's Law and the distribution that was observed in the data retrieved from the crypto exchange, in this way, it is possible to quantify objectively the deviation observed. The final technique adopted is the distance from log-normal distribution, in this analysis the authors state that without any kind of manipulation from the crypto exchange, the trading volume should follow a log-normal distribution, so any notable deviation from such distribution could indicate a possible manipulation in the data. In this case, to measure the distance the Kolmogorov-Smirnov (KS) statistic is adopted, this measure can quantify the distance between a sample of cumulative distribution function (c.d.f.) and the c.d.f. of a reference distribution (the log-normal distribution in our case). Finally, the three measures are aggregated into a unified measure of fake trading by extracting the principal components of the measures described. In conclusion, the paper examines the short-run and longer-run effects of wash trading, the evidence is that, in the short run, inflating trading volume gives a great benefit to the crypto exchange because it increases real trading volume and investors initially cannot distinguish between real and fake trading volume. However, in the longer run, fake trading negatively affects an exchange's web popularity and the revenue from the trading fees.

4. **Do Cryptocurrency Exchanges Fake Trading Volumes? An Empirical Analysis of Wash Trading Based on Data Mining** [35]

   The paper proposes a data mining approach to analyze both off-chain data (transactions on the crypto exchange) and on-chain data (transactions on the blockchain), as well as the web visits of the website of the exchanges to detect those exchanges which engage in fake trading activities. The main goal is to establish whether crypto exchanges fake their trading volumes, how they do it and if it can be detected. The authors developed several metrics based on the on-chain and off-chain transactions combined with the page-view, which is the index that shows the number of times the website is visited and it is used as an indicator for the popularity of the crypto exchange. The metrics, which are divided into off-chain and on-chain types, are compared to analyze if the exchange examined is faking trades because the off-chain data can be easily falsified but this is not true for the on-chain data which is a reflection of the true state of the crypto exchange.

   The analysis reveals that some exchanges, such as Huobi, have a bigger suspicion of inflating the trading volume whereas others, like Binance, are more honest and they have a lower probability of falsifying the trading volumes. Regarding the way in which the exchanges fake their trading volumes, the results indicate that different exchanges use different automatic programs to inflate trading volumes, furthermore, the paper identifies Gate.io, Kucoin and MXC as exchanges that concealed the wash trading more effectively than Huobi which may use a more easily recognizable faking strategy. In conclusion, the authors state that it is important to identify and let it be known about these misleading conducts from the crypto exchange, which can deceive the customers and manipulate the market to a certain degree.

5. **WTEYE: On-chain wash trade detection and quantification for ERC20 cryptocurrencies** [36]

   The wash trading phenomena occurs also on Decentralized EXchanges (DEX) which operate on blockchains like Ethereum, Binance Smart Chain (BSC) [1] and Polygon [2]. For

[1] https://www.bnbchain.org/en/smartChain

[2] https://polygon.technology

this reason, Cui et al. developed two algorithms that aim to identify and quantify the wash trade activities based on the ERC20 token on-chain data. The focus is on DeFi (Decentralized Finance) projects such as Uniswap[3], PancakeSwap[4] and dYdX [5] which can have an economic incentive to engage in wash trading because it can help in rebuilding network structure and it has a deep impact on the fluctuation of token price. The algorithm proposed is based on transaction graph network modeling, with the transactions modeled as directed flows in the graph. Furthermore, the authors give a rigorous mathematical definition of the wash trades which is integrated into the graph model.

The first step of the algorithm is to build the graph G model using the addresses and amounts of the trade. The graph is defined as G = (V, E), where V is the set of traders' addresses and E is the set of trades, the weights of each edge are equal to the amount of the trade. Given the graph in input, the algorithm checks the presence of circles using the DFS (Depth First Search) traversal algorithm since wash trade is always given by collusive cliques. If the transactions created from circle nodes satisfy the condition of the definition of wash trade then the trades are labeled as suspicious.

In the second algorithm, the authors make the assumption that a node has a higher probability to make wash trades with its neighbors, this is done to take the key nodes as input and to detect wash trades in the neighbor list iteratively. As in the other algorithm, transactions that satisfy the mathematical criteria imposed are labeled as wash trades.

The experiments made revealed that for the majority of ERC-20 tokens, the amount of wash trade is over 15% of the volume, moreover, the experiments made on the UNI token suggest that over 30% of transactions involving UNI could be wash trades. In conclusion, the paper discusses how the high cost of the transaction fees on the blockchain prevents significant inflation in the trading volume of ETH and BTC. When a new token is launched on the crypto market the creators of such coin can give an initial boost of the trading volume to increase future earnings and the number of trades of the token, this practice clearly deceives new potential investors who are fooled by the real value

[3] https://app.uniswap.org

[4] https://pancakeswap.finance

[5] https://trade.dydx.exchange/

of the token. Furthermore, the experiments suggest that the wash trades are mainly employed by whale addresses, this insight can be crucial to detect wash trade in a more efficient manner.

Chapter 5

Data Collection and Preprocessing

The data collected during the analysis involves features of the crypto exchanges which can help identify possible inconsistencies within their behavior. The most insightful information comes from the transactions and trading volumes reported by the exchanges, as well as the web traffic for the exchanges which can give an overview of their popularity.

Trading Volumes

Concerning the trading volumes, which are defined as the total quantity of cryptocurrencies traded on the exchange (usually expressed in USD), the data collection phase started in January 2022 and has been carried out until December 2022, collecting over nine thousand hourly data points for every exchange for the whole year of 2022. The collection has been performed on a Raspberry Pi Zero W[1] which automatically requested the trading data every hour using *cron*, which is a job scheduler that can be used to schedule commands on a determined time period. This data has been collected from the CoinGecko API, which is a data aggregator that collects and ranks cryptocurrencies, exchanges and NFTs. In order to continuously collect the temporal series a Python script that acts as a data collector has been set to automatically make API requests every hour for the trading volumes of the top crypto exchanges and then store

[1] https://www.raspberrypi.com/products/raspberry-pi-zero-w/

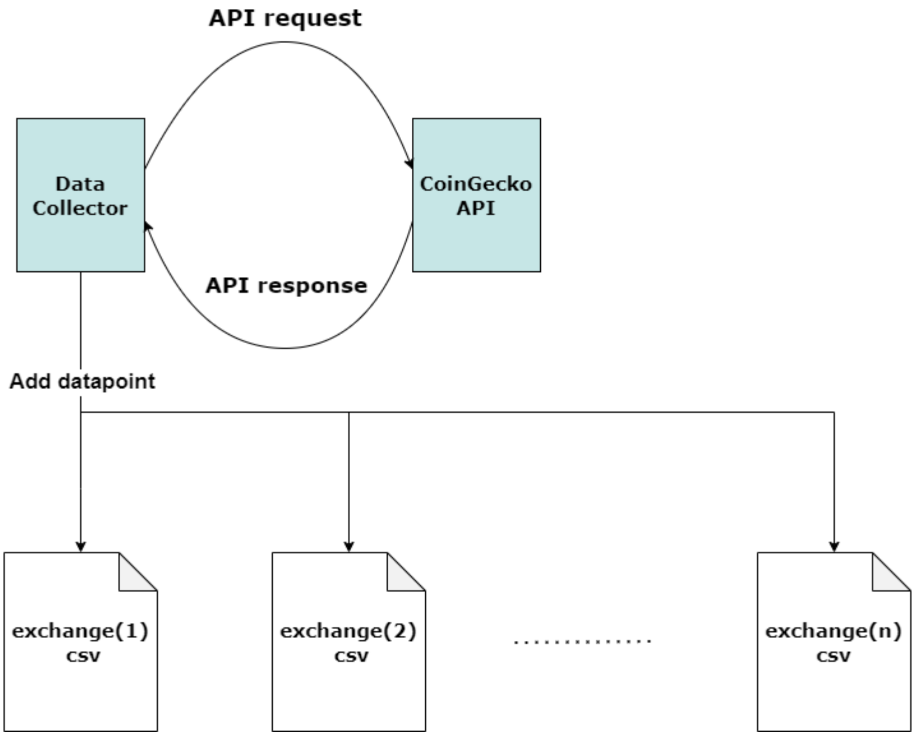


Figure 5.1: Data collector for trading volumes.

it in a CSV (Comma Separated Value) file (see figure 5.1), the total dimension of all CSV files is roughly 75 MB. The trading volume dataset contains a temporal series for every exchange analyzed where every data point contains the trading volume in USD for every hour of the day and its timestamp. The REST API of CoinGecko offers an endpoint to retrieve a list of the top crypto exchanges supported by the aggregator and additional information for each one of them. The script makes an API request every hour, it gets the trading volume for the last 24 hours which is expressed in BTC and then multiplies it for the current value of BTC, then it stores it in a new row of the file associated with the determined exchange. Hourly data of the trading volumes allow a more fine granularity analysis where potential sudden increases or decreases in a short interval (typical of the cryptocurrency market) can be better captured and studied.

Transaction Pairs

Given the continuous flow of data needed to collect the real-time transactions of the exchanges the public WebSocket API provided by the exchanges has been used to continuously retrieve the transactions. The transactions involved in the analysis are the ones associated with the pairs: **BTC-USD**, where we include every transaction of Bitcoin against the U.S. dollar or the stablecoins USDT, USDC or BUSD which are pegged to the U.S. dollar, **ETH-USD** (Ethereum against USD), **ADA-USD** (Cardano against USD), **XRP-USD** (Ripple against USD) and **ETH-BTC** (Ethereum against Bitcoin). These four cryptocurrencies are the coins with the highest market capitalization[2] according to CoinMarketCap and they are also the most traded crypto, this allows us to have a significant subset of transactions on which to perform the analysis.

A WebSocket is a communication protocol that provides full-duplex communication channels over a single, long-lived connection between a client and a server. This kind of communication is more suitable for real-time data collection because it provides a stream of data of the new instant updates about the transactions, it has a smaller overhead since when the initial connection is established the data exchange becomes more efficient and, furthermore, it guarantees a persistent communication unlike traditional HTTP connections, which are stateless and short-lived. The transaction data has been collected in real-time for the whole time frame ranging from April 2023 to June 2023, the data collection has been performed on a desktop computer with an i7-11700K @ 3.6 GHz processor and 32 GB of RAM and the total amount of data collected corresponds to roughly 13 GB. Both scripts used for gathering trading volumes and transaction data were written in Python 3.11.

The attributes of the transaction dataset (see table 5.1) are the following: *timestamp* of the transaction (expressed in Unix time in milliseconds), *price* of the cryptocurrency (expressed in the monetary unit of the quote currency, e.g. for the pair BTC-USD the price will be expressed in USD), *amount* of the transaction (expressed in the monetary unit of the base currency in the pair, e.g. for the pair BTC-USD the amount is expressed in BTC), *type* of transaction, which

[2] The market capitalization is an indicator of the value of crypto, it is calculated by multiplying the total number of coins that have been mined by the price of a single coin at any given time.

TIMESTAMP	PRICE	AMOUNT	TYPE	TID
1681336317189	0.00076132	29965.97	sell	170967411110
1681336322855	0.00089553	29967.97	buy	170967414591
1681336402716	0.0109	29965.97	sell	170967458808
1681336410973	0.00491	29965.97	sell	170967462710

Table 5.1: Example of transaction dataset from Gemini

can be "buy" or "sell" according to whom initially placed the order and finally *tid* which is the transaction id.

Examined Exchanges

The crypto exchanges which were considered for the transaction data are divided into two categories. The first is the trustworthy group, which is the group of exchanges that are regulated in the U.S. (one of the most strict regulations in the crypto market) and possess a good reputation built on years and years of service. These exchanges are heavily regulated and need to be transparent about their operations since they can receive frequent audits and must adhere to regulatory compliance. The trustworthy group is composed of the exchanges Kraken [37], Gemini [38], and Binance[3] [39]. The other group of exchanges considered was chosen based on low scores of the spot exchange score of CoinMarketCap (CMC), which corresponds to a ranking system, where attributes such as web traffic, liquidity, trading volume, and confidence in the legitimacy of trading volumes are considered along with more qualitative factors such as longevity, reputation, and user feedback to classify the exchange and to give a more transparent and reliable ranking system. The low scores given to the exchanges could indicate a possible bad reputation for the exchange. The group comprises exchanges that do not have a license in the U.S. to operate as crypto exchanges, thus, being less monitored, are LBank,

[3] Binance technically does not possess a license for the U.S. market but it opened a second crypto exchange named "Binance US" which is licensed and can operate in the U.S.

Hotcoin, Bitmex, BTSE, HitBTC, Hotbit[4], and Changelly.

Table 5.2 describes the features of the crypto exchanges examined. The volume reported is the total amount for the month of May 2023 whereas the web traffic refers to the number of website visits that the exchange received in May 2023 and the spot exchange score is the score that CMC gave to the exchanges. The web traffic data has been collected from the website SimilarWeb, which is a service that ranks and analyzes the web traffic of popular websites[5]. The scores of the exchanges go from 0 to 10, where a spot exchange score above 6.0 is defined as "Good", from 6.0 to 4.0 as "Average" and below 4.0 as "Poor" [40].

		WEB	EXCHANGE	FOUNDING		
NAME	VOLUME	TRAFFIC	SCORE	YEAR	OFFICE	TRUST
Binance	$215.9B	61.5M	9.9	2017	Cayman Islands	✓
Kraken	$12B	4.8M	8.2	2011	U.S.A.	✓
Gemini	$674M	1.5M	6.8	2014	U.S.A.	✓
LBank	$22.9B	11.3M	6.7	2016	China	✗
Hotcoin	$86.8B	7.3M	5.9	2017	Australia	✗
Bitmex	$11.9M	892K	5.0	2014	Seychelles	✗
BTSE	$15.8B	6.7M	4.8	2018	Virgin Islands	✗
HitBTC	$14.9B	304.4K	4.2	2013	Virgin Islands	✗
Hotbit	$2.3B	1M	3.3	2018	Estonia	✗
Changelly	$11.6B	1.5M	3.2	2020	Seychelles	✗

Table 5.2: Overview of the crypto exchanges considered.

The data that is usually provided for each crypto exchange are well described in their API documentation, we will use the crypto exchange Kraken as reference[6]. There is a distinct characterization of private data, which is data concerning the operations on the exchange

[4] Note that Hotbit ceased its operations on 22 May 2023, meaning that some data is missing

[5] https://www.similarweb.com

[6] https://docs.kraken.com/rest/

made by the user, some examples of this kind of data are the balance of the user, open and closed orders, trade history, and trade volume. Besides private data, crypto exchanges often offer also reporting services about the trades which gives an overview of the operations made with various coins. This can facilitate the analysis of their internal trades on the crypto market. The API provides also calls to create or cancel new trades on the market, these can be used to automate operations when certain conditions of the market are met and more generally it allows the creation of trading bots that can operate on the market by buying or selling certain assets in an independent way.

The public market data involves data about the crypto exchange which is accessible by everyone, the main measures are:

- **Tickers**

  A symbol ticker is an acronym used to identify a pair of assets (e.g. BTC-USD is the ticker for the pair Bitcoin - USD), associated with the symbol ticker are data regarding the pair such as traded volume in the last 24 hours, asks, bids, last trade closed, number of trades in the last 24 hours, and opening price of the day.

- **OHLC**

  Open High Low Close (OHLC) data represents the fluctuation over time of the price of a certain asset. This kind of data is used to create candlestick charts (an example is shown in figure 5.2).

  The x-axis represents the time, every candle represents a time span of 4 hours (the most common time spans used are 1m, 15m, 1h, 4h, 1d) and the y-axis represents the price of the coin. In order to create a single candle (see figure 5.3) the following price points are needed: *Open* (1) - the first price point for the time span selected, *High* (2) - the highest price point recorded during the whole time span, *Low* (3) - the lowest price point recorded throughout the time span and *Close* (4) - the last price point recorded for the pair.

  The relationship among the open, high, low, and close price points determines the appearance of the candle. The distance between the open and the close is called the body

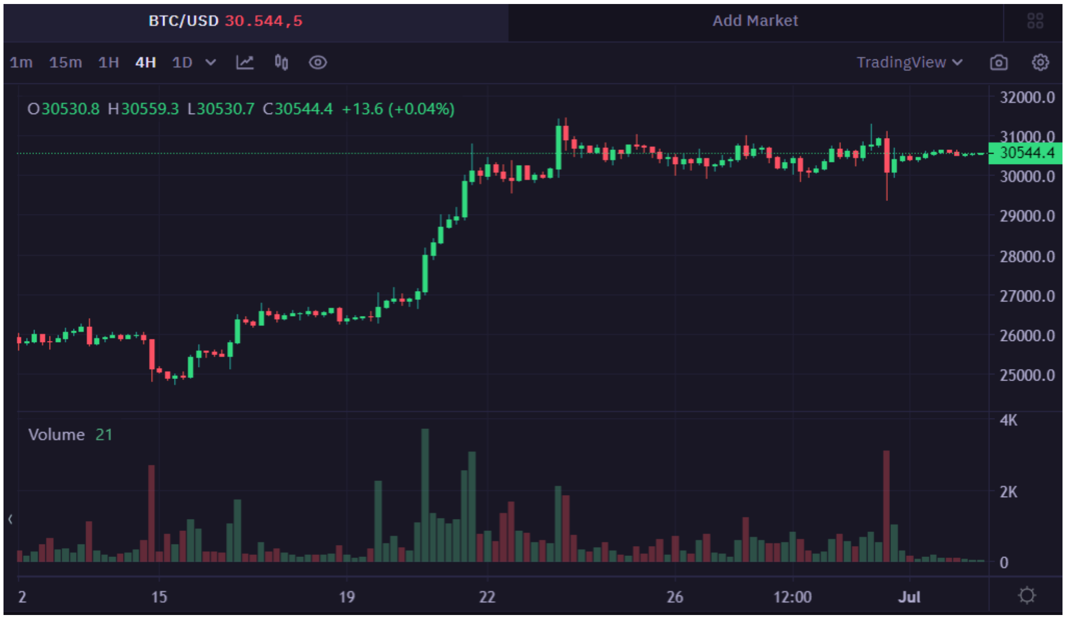


Figure 5.2: A candlestick chart for the pair BTC-USD (Kraken).

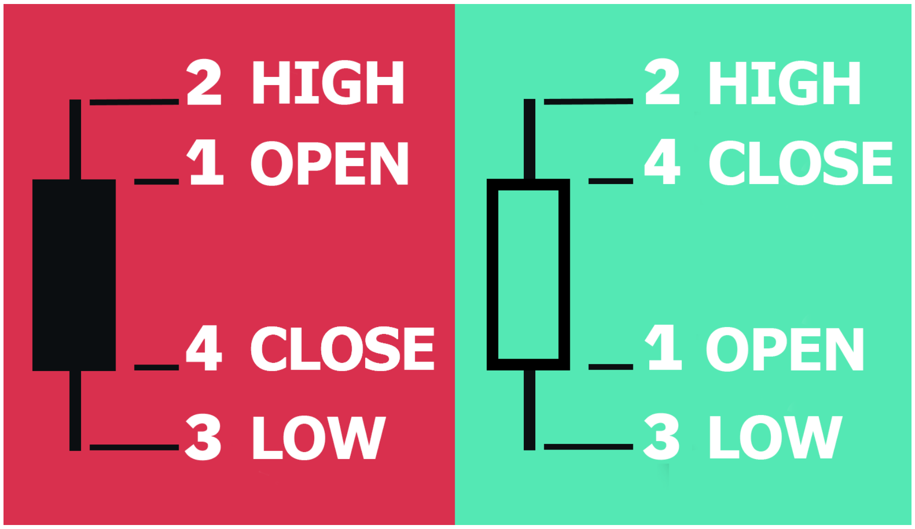
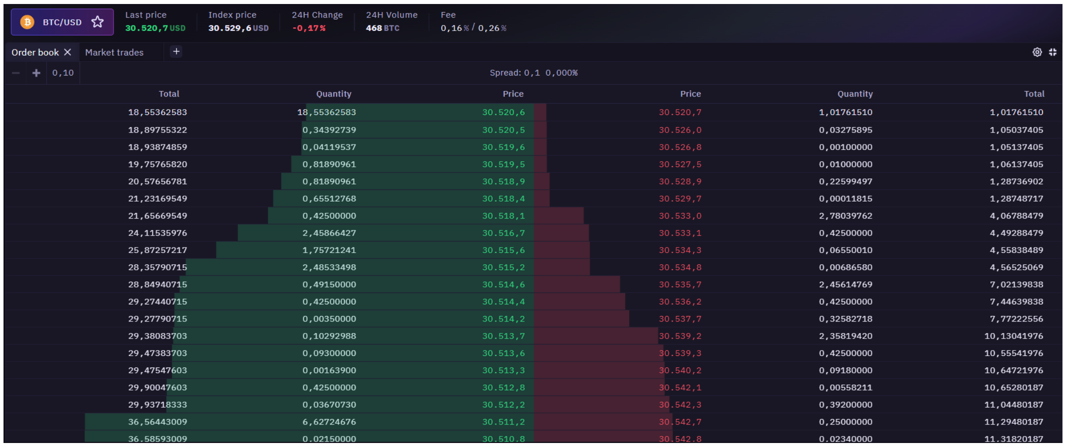


Figure 5.3: A candlestick depicted with its price points.

of the candle, whereas the distance between the body and the high/low is called the wick or shadow and finally the distance between the high and the low of the candle is called the range of the candlestick. If the body of the candle is green it means that the asset closed at a price that is higher than the open, instead, if it is colored in red it means that the close price point was lower than the open.

- **Order book**



BTC/USD | Last price 30.520,7 USD | Index price 30.529,6 USD | 24H Change -0,17% | 24H Volume 468 BTC | Fee 0,16% / 0,26%

Order book × | Market trades | +

− + 0,10 | Spread: 0,1 0,000%

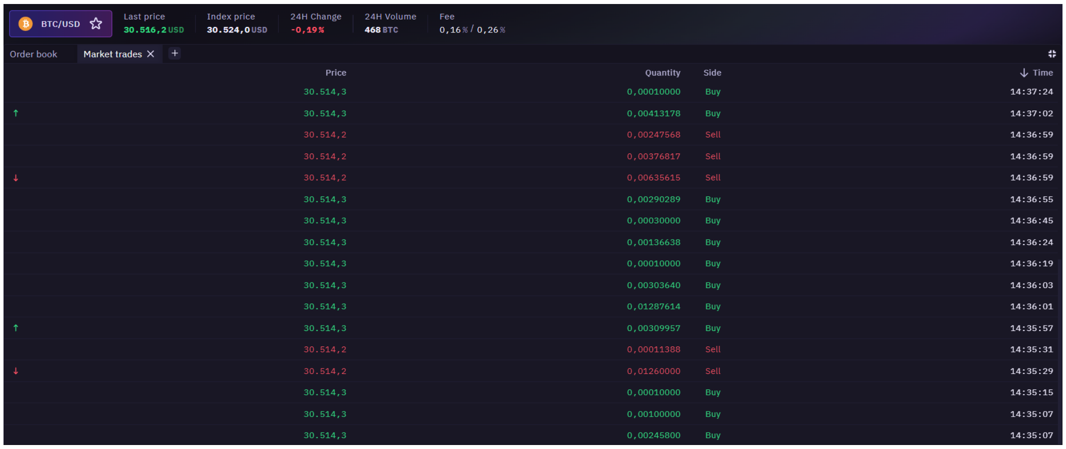
Total	Quantity	Price	Price	Quantity	Total
18,55362583	18,55362583	30.520,6	30.520,7	1,01761510	1,01761510
18,89755322	0,34392739	30.520,5	30.526,0	0,03275895	1,05037405
18,93874859	0,04119537	30.519,6	30.526,8	0,00100000	1,05137405
19,75765820	0,81890961	30.519,5	30.527,5	0,01000000	1,06137405
20,57656781	0,81890961	30.518,9	30.528,9	0,22599497	1,28736902
21,23169549	0,65512768	30.518,4	30.529,7	0,00011815	1,28748717
21,65669549	0,42500000	30.518,1	30.533,0	2,78039762	4,06788479
24,11535976	2,45866427	30.516,7	30.533,1	0,42500000	4,49288479
25,87257217	1,75721241	30.515,6	30.534,3	0,06550010	4,55838489
28,35790715	2,48533498	30.515,2	30.534,8	0,00686580	4,56525069
28,84940715	0,49150000	30.514,6	30.535,7	2,45614769	7,02139838
29,27440715	0,42500000	30.514,4	30.536,2	0,42500000	7,44639838
29,27790715	0,00350000	30.514,2	30.537,7	0,32582718	7,77222556
29,38083703	0,10292988	30.513,7	30.539,2	2,35819420	10,13041976
29,47383703	0,09300000	30.513,6	30.539,3	0,42500000	10,55541976
29,47547603	0,00163900	30.513,3	30.540,2	0,09180000	10,64721976
29,90047603	0,42500000	30.512,8	30.542,1	0,00558211	10,65280187
29,93718333	0,03670730	30.512,2	30.542,3	0,39200000	11,04480187
36,56443009	6,62724676	30.511,2	30.542,7	0,25000000	11,29480187
36,58593009	0,02150000	30.510,8	30.542,8	0,02340000	11,31820187

Figure 5.4: The order book containing the asks and bids for the pair BTC-USD (Kraken).

Every ask and bid offer is listed in the order book (see figure 5.4) which keeps track throughout the time of the buy and sell orders present on the platform for a certain pair. The order book is divided into two parts on the left side in green the bid offers (buy transactions) and on the right side the ask offers (sell transactions), each entry is made up by the price and the amount.

- **Trades**

  The recent trades for a given pair can be retrieved on demand (see figure 5.5). In each transaction made on the exchange there are the timestamp, price, amount and type of transaction (buy or sell).



BTC/USD | Last price 30.516,2 USD | Index price 30.524,0 USD | 24H Change -0,19% | 24H Volume 468 BTC | Fee 0,16% / 0,26%

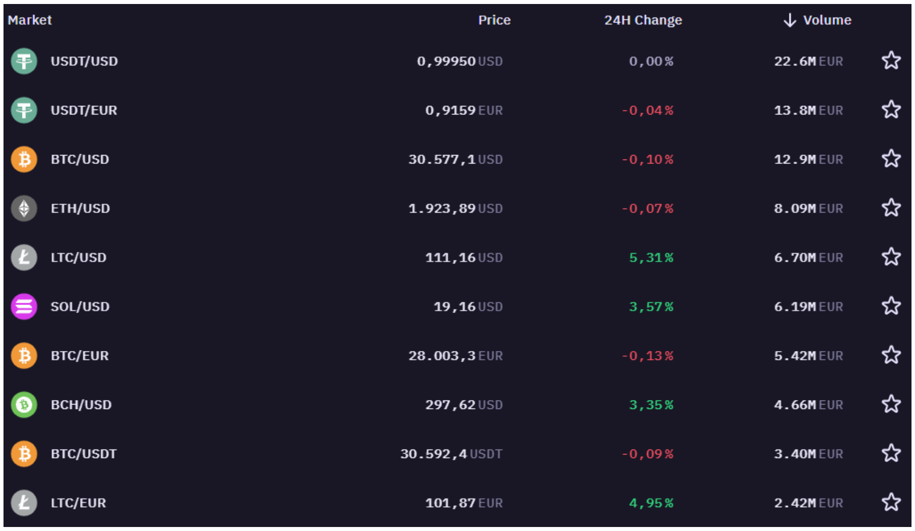
Order book | Market trades × | +

	Price	Quantity	Side	↓ Time
	30.514,3	0,00010000	Buy	14:37:24
↑	30.514,3	0,00413178	Buy	14:37:02
	30.514,2	0,00247568	Sell	14:36:59
	30.514,2	0,00376817	Sell	14:36:59
↓	30.514,2	0,00635615	Sell	14:36:59
	30.514,3	0,00290289	Buy	14:36:55
	30.514,3	0,00030000	Buy	14:36:45
	30.514,3	0,00136638	Buy	14:36:24
	30.514,3	0,00010000	Buy	14:36:19
	30.514,3	0,00303640	Buy	14:36:03
	30.514,3	0,01287614	Buy	14:36:01
↑	30.514,3	0,00309957	Buy	14:35:57
	30.514,2	0,00011388	Sell	14:35:31
↓	30.514,2	0,01260000	Sell	14:35:29
	30.514,3	0,00010000	Buy	14:35:15
	30.514,3	0,00100000	Buy	14:35:07
	30.514,3	0,00245800	Buy	14:35:07

Figure 5.5: The most recent trades for the pair BTC-USD (Kraken).

- **Tradable pairs**

  It is possible to request the list of tradable pairs supported by the exchange (see figure 5.6) along with additional information about each pair.



Market	Price	24H Change	↓ Volume	
USDT/USD	0,99950 USD	0,00%	22.6M EUR	☆
USDT/EUR	0,9159 EUR	-0,04%	13.8M EUR	☆
BTC/USD	30.577,1 USD	-0,10%	12.9M EUR	☆
ETH/USD	1.923,89 USD	-0,07%	8.09M EUR	☆
LTC/USD	111,16 USD	5,31%	6.70M EUR	☆
SOL/USD	19,16 USD	3,57%	6.19M EUR	☆
BTC/EUR	28.003,3 EUR	-0,13%	5.42M EUR	☆
BCH/USD	297,62 USD	3,35%	4.66M EUR	☆
BTC/USDT	30.592,4 USDT	-0,09%	3.40M EUR	☆
LTC/EUR	101,87 EUR	4,95%	2.42M EUR	☆

Figure 5.6: Some of the cryptocurrency tradable pairs supported by Kraken.

The data collected has been analyzed to check for consistency and integrity. For the trading volume data which ranged in a time frame of a complete year, there were some missing data points due to network or power outages during the collection phase. This was resolved through a data integration process where the missing data for the exchanges were retrieved by another endpoint of the CoinGecko API which provides daily trading volumes for a determined exchange. In this way, it was possible to partially recover the missing data, and a linear interpolation of the missing data points was performed to complete the temporal series.

Chapter 6

Crypto Exchanges Data Analysis

After the preprocessing phase, we proceeded to analyze the data collected in order to discover possible hidden patterns of dishonest behavior and reveal significant insights into the state of the exchange market. The analysis performed used the data collected on transactions and trading volumes since they are possible indicators of suspicious activities such as wash trading.

6.1 Statistical Analysis

6.1.1 Cryptocurrency Dominance

The techniques used rely on both statistical and machine learning approaches, with the goal of exploiting both views to have a more comprehensive understanding of the data in question. Table 6.1 gives an overview of the total number of transactions of the pairs BTC-USD, ETH-USD, ADA-USD, XRP-USD and ETH-BTC for each one of the ten crypto exchanges analyzed.

For most of the exchanges, the BTC-USD pair is the most traded one, followed by the pair ETH-USD as second and the others with a smaller relative percentage. These percentages are measured considering only these four cryptocurrencies but, nevertheless, they tend to respect

	TRADES	BTC-USD	ETH-USD	ADA-USD	XRP-USD	ETH-BTC
Binance	114M	51.1%	30.6%	8.29%	5.34%	4.7%
Kraken	2.9M	63.1%	22.3%	7.16%	4.86%	2.56%
Gemini	704K	59.5%	37.2%	-	-	3.29%
LBank	9.6M	26.6%	23.6%	20.1%	8.84%	20.9%
Hotcoin	12M	42.8%	36.9%	8.17%	7.88%	4.29%
Bitmex	6.5M	60.6%	30.7%	3.2%	5.47%	-
BTSE	9.9M	27%	39.9%	15.9%	14%	3.25%
HitBTC	3M	30.1%	30%	10.9%	15.1%	13.8%
Hotbit	639K	41.8%	16.4%	13.9%	14.3%	13.6%
Changelly	2.6M	37.3%	21.5%	11.7%	15%	14.6%

Table 6.1: Traded volume and the partition of trades among the pairs.

the BTC dominance of the whole market. The crypto market coin dominance is a ratio of that currency market capitalization to the cumulative market capitalization of all cryptocurrencies (or rather the top 125 coins). It is often used to have an idea of how big a coin is in terms of market capitalization and adoption. In figure 6.1 the dominance chart for the months of April, May and June 2023 is shown, the average BTC dominance is close to 47% and the average ETH dominance is around 19%.

In our data, the only exception that does not follow this trend is the BTSE exchange which has a bigger market for Ethereum with respect to Bitcoin and LBank which has a comparable share of markets among the pairs BTC-USD, ETH-USD and ETH-BTC. As expected, Binance which has the highest trading volume in the market is also the one with the highest number of trades (114 million trades) and Gemini, which is the smallest exchange in our dataset in terms of traded volume, is also the one with fewer trades (704 thousand trades). The remaining exchanges have a comparable number of trades which range in the order of millions of trades.

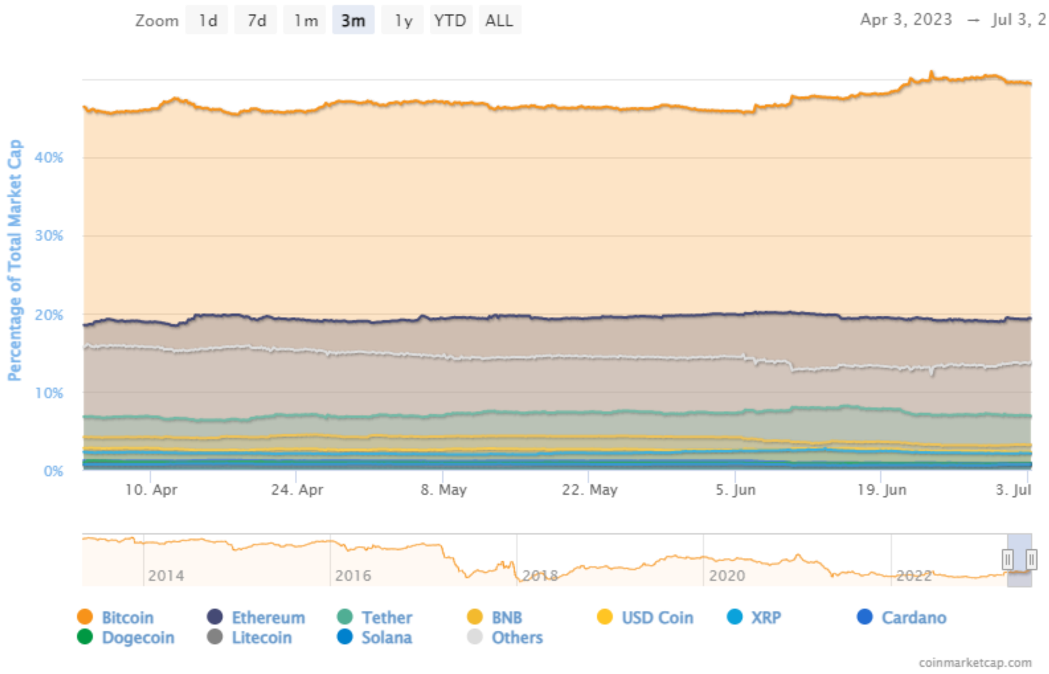


Figure 6.1: BTC dominance has been around 47% for April-May-June 2023.

6.1.2 Benford's Law Test

Next, we performed the Benford's Law [41] test on the transaction data. The goal of this statistical test is to discover if the trades of the dataset adhere to Benford's Law (also known as "First Digit Law" or "Law of Anomalous Numbers") which describes the distribution of the first significant digit in naturally generated numerical data sets. The law states that the probability of a number starting with a certain digit decreases logarithmically as the digit increases (see table 4.1).

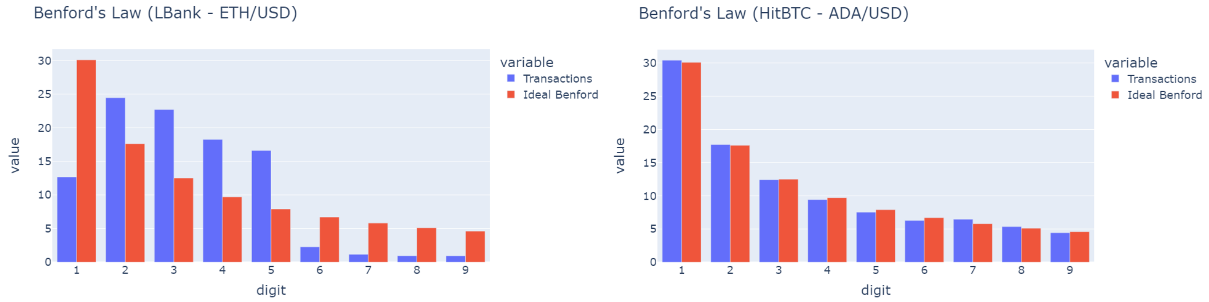
This statistical test is often used to detect anomalies or potential fraud in numerical data because deviations from the distribution described by Benford's Law could indicate manipulated data or irregularities in the dataset. In our experiment the attribute used to perform the first digit test is the trade size of the transactions which is computed by multiplying the price of the coin by the amount traded, this gives us the value of the transaction. In order to have a quantitative result that could be used to compare the exchanges, we computed the

	BTC-USD	ETH-USD	ADA-USD	XRP-USD	AVERAGE
Binance	2.65%	1.84%	0.86%	1.08%	1.6%
Kraken	1.72%	1.47%	1.8%	0.89%	1.47%
Gemini	3.18%	3.43%	-	-	3.3%
LBank	6.75%	7.62%	5.18%	2.39%	5.48%
Hotcoin	0.94%	0.53%	0.76%	0.42%	0.66%
Bitmex	2.85%	2.52%	1.07%	1.52%	1.99%
BTSE	6.06%	5.46%	3.53%	4.09%	4.78%
HitBTC	2.22%	4.05%	0.3%	1.53%	2.03%
Hotbit	4.99%	4.52%	3.25%	3.52%	4.07%
Changelly	2.66%	5.9%	0.25%	1.51%	2.58%

Table 6.2: Deviations from Benford's Law for every pair of the exchanges.

average deviation for all digits of a single pair and the average of all the pairs for a single exchange (shown in table 6.2). More formally the deviation for a single pair is computed as $dev = \frac{1}{9}\left(\sum_{i=1}^{9}|freq_i - ideal_i|\right)$ where i is the digit, $freq_i$ is the actual frequency of the digit i and $ideal_i$ is the ideal frequency of digit i according to the Benford's Law. The total average deviation is simply the average of all the deviations of the pairs.

The results of the test indicate that LBank is the exchange with the highest deviation from Benford's Law, the average deviation in percentage for the ETH-USD trades is a 7.62% and the average deviation for the BTC-USD is 6.75% and the total average deviation is 5.48%. On the other side, the exchange with the lowest deviation is Hotcoin which has an average deviation of 0.66%. Overall, the most "suspicious" exchanges are LBank, BTSE and Hotbit (which are highlighted in table 6.2 which have a significantly higher deviation to the Benford's Law w.r.t the other exchanges. This means that there could be some anomalies inside the transaction data of these exchanges. Figure 6.2 shows on the left part a bar plot that depicts the actual distribution of the first significant digit (in blue) of the transaction data for the pair ETH-USD for the exchange LBank and in red the ideal distribution dictated by Benford's Law and the right part of the figure shows the bar plot for HitBTC for the pair ADA-USD which shows an almost perfect overlap of the two distributions.



(a) Distribution of first digit (LBank - ETH/USD). (b) Distribution of first digit (HitBTC - ADA/USD).

Figure 6.2: Comparison of the distribution of the first significant digit (blue) with Benford's Law (red).

It is easily noticeable how the bars have different heights, in fact, the frequency of the digit 1 as the first significant digit is less than 15% whereas according to Benford's Law should have been around 30%. On the right side, the figure shows the same data for the exchange HitBTC, in this case, it is easy to notice how the distribution follows the First Digit Law.

6.1.3 Transactions-Volume Correlation

The next experiment consists in checking the correlation between the number of transactions and trading volume reported by the exchange. To do this we use the Pearson Correlation Coefficient (PCC) which is a statistical measure that quantifies the linear relationship between two variables, it is usually denoted with the symbol ρ (or r in some cases) and it could range from -1 to +1. The measure is expressed as the ratio between the covariance, which is the joint variability of two variables and the product of their standard deviations, which is a measure of the amount of dispersion of a set of values, $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$. The value of the PCC reaches the value of 1 when we have a perfect positive correlation, i.e. when one variable increases the other increases in the same way, whereas when the value of the PCC is -1 it indicates that we have a perfect negative correlation, finally a value of 0 indicates no linear relationship whatsoever. In our test the two variables in question are the daily frequency of the transactions and the daily trading volume, the goal is to understand if there is a positive correlation between the two as is expected due to the nature of traded volume. In fact, as the number of transactions increases, we expect also an increase in the amount of traded volume which in our test would result in a value of ρ close to 1. Note that this is not always the case, the linear relationship holds true generally for high-liquid markets, such as major stock exchanges or crypto markets in our case, in contrast in less liquid markets or for certain complex financial instruments the relationship could not be linear. Even in high-liquid markets, there could be instances of a non-linearity relationship between the frequency of trades and the traded volume. For example, just a few large trades could notably increase the traded volume without a corresponding increase in the number of trades. Similarly, a large number of smaller transactions could increase the number of transactions without affecting the traded volume [42]. In our experiment we grouped and calculated the frequencies of the transactions for each day of the time span, after getting the dataset of the frequency of transactions per day we computed the Pearson correlation coefficient compared with the traded volume. This process has been repeated for all the pairs of the exchanges in exam. Table 6.3 shows the results obtained from the analysis, the group of trustworthy exchanges which are: Binance, Kraken and Gemini have respectively an average coefficient of 0.97, 0.92 and 0.79.

Although the interpretation of the value of ρ is somewhat dependant on the nature of data and on the purpose of the test, a good rule of thumb scale is the one reported in table 6.4 which sets the value 0.7 or higher the threshold for a high correlation [43].

According to this scale, the values produced from the trustworthy group indicate a high correlation between the daily number of trades and the daily trading volume, the other exchanges which falls in this category are BTSE, which has a strong correlation with an average value of $\rho = 0.94$ and Hotbit with an average value of $\rho = 0.76$. The exchanges with the lowest values of correlation are Hotcoin with $\rho = 0.39$, which according to table 6.4 indicates a low correlation between the two variables and Changelly, LBank and HitBTC with respectively 0.53, 0.57 and 0.57 which falls on the low end of the moderate correlation tier. Finally, the remaining exchanges have values that fall on the high end of the moderate correlation tier.

	BTC-USD	ETH-USD	ADA-USD	XRP-USD	AVERAGE
Binance	0.96	0.98	0.99	0.97	0.97
Kraken	0.92	0.95	0.91	0.91	0.92
Gemini	0.81	0.77	-	-	0.79
LBank	0.77	0.57	0.47	0.5	0.57
Hotcoin	0.44	0.45	0.33	0.35	0.39
Bitmex	0.82	0.84	0.36	0.73	0.68
BTSE	0.96	0.93	0.95	0.93	0.94
HitBTC	0.64	0.52	0.46	0.67	0.57
Hotbit	0.75	0.88	0.71	0.7	0.76
Changelly	0.53	0.69	0.29	0.62	0.53

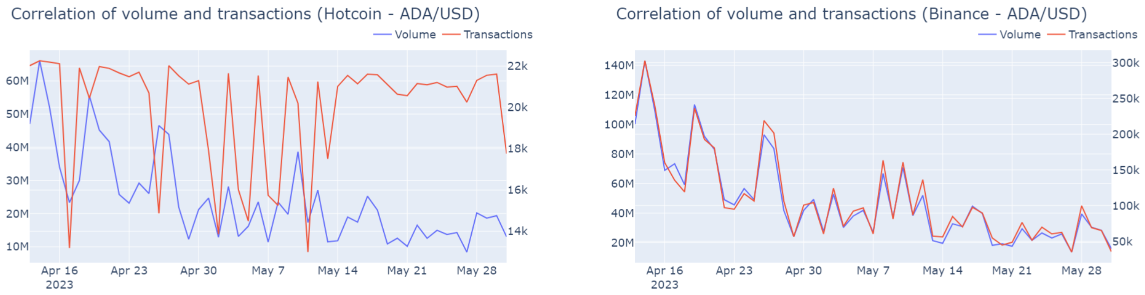
Table 6.3: Pearson correlation coefficient of number of transactions and traded volume.

SIZE OF ρ	**INTERPRETATION**
0.9 to 1	Very high correlation
0.7 to 0.89	High correlation
0.5 to 0.69	Moderate correlation
0.3 to 0.49	Low correlation
0 to 0.29	Little if any correlation

Table 6.4: Strength of correlation.

Figure 6.3 represents the comparison between the highest correlation value (0.99) for the pair ADA/USD for Binance (on the right side) and the lowest correlation value (0.33) for the pair ADA/USD for Hotcoin (on the left side).

Binance has a trading volume and a number of transactions series which almost overlap, the linear relationship is evident, whereas the exchange Hotcoin presents two lines that do not correlate at all. This could be an indicator of some kind of anomaly since the volume series



(a) Correlation of $\rho = 0.33$ (Hotcoin).

(b) Correlation of $\rho = 0.99$ (Binance).

Figure 6.3: Comparison of the correlation of traded volume (in blue) and transactions (in red) for Binance and Hotcoin.

of both exchanges tend to have a similar shape with Hotcoin having roughly half of the volume (peak at $140M for 14/04 for Binance corresponds to the peak of $66M for the market ADA/USD) but the transactions amount of are a magnitude of order smaller than those of Binance with a completely different shape (besides the last part of the plot which more or less tends to follow the trend).

6.1.4 Traded Amounts Distribution

Next, we analyzed the trade size of the transactions which is the attribute *amount* present in the data set which identifies the quantity of coins exchanged. Real transactions made by real people tend to have round numbers for the number of coins exchanged in a transaction, this behavior is due to the human need for cognitive reference points (such as multiples of 10) [44], for example, a person would be more incline to buy or sell a quantity with a round number ending with the digit 0 instead of an unrounded number. Round numbers act as a cognitive shortcut for convenience and efficiency, another name for this phenomenon is the so-called round-number bias. This behavioral pattern could be exploited to discover if the transactions of the exchange are generated by humans and represent real trades or if they are automatically generated by some algorithm to boost the volumes through wash trading or just by fabricating transactions. The goal of the test is to check if there are accumulations of trades sizes on quantities which are round numbers, a honest exchange should manifest these bumps in transaction frequencies on multiples of 10.

For our test we used different base units for each pair, e.g. for Bitcoin we grouped the transactions in bins that are multiples of 0.001 BTC (which at the time corresponded roughly to $30) up until 0.1 BTC, in this way we could focus on fine-granularity which is the majority of trades (for Kraken the 80% of BTC-USD trades have a trade size less than 0.1 BTC). The base unit for Ethereum is 0.01 ETH (which at the time corresponded roughly to $20) with bins going up to 10 Ether, for Cardano and Ripple the base unit used is 10 ADA and XRP (which at the time corresponded roughly to $40) going up to 1000.

To better quantify this "bumpiness" property (the presence of clusters of transactions on round numbers) of the exchange we used the Student's t-test for comparing the trade size at round numbers against the nearby unrounded numbers. The Student's t-test is a statistical test that is used to compare the means of two groups of data to check if there is a significant difference between them. This test is named after William Sealy Gosset who first determined it in 1908 and published it under the pseudonym "Student" [45]. There are two types of t-tests: independent t-test, where we compare the means of two independent groups to determine if there is a significant difference between them, this is the case for our experiment, the other type is the paired samples t-test where we compare the means of two related samples, it is usually used when the same subject is measured twice under different conditions. The t-test quantifies the difference between the arithmetic means of the two samples expressed as the t-value, which is a ratio of the difference between the sample means to the variation within the samples, the formula for independent two samples t-test with the same variance is $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}}$ where n is the sample size and s_p is the pool standard deviation (which is used for estimating the variance) [46]. A higher t-value indicates that the difference of the means of the group in the nominator is greater than the pooled standard error in the denominator, this implies that we have a more significant difference between the two groups. The degrees of freedom, which indicates the maximum number of independent values that can vary, for this kind of testing is $df = 2n - 2$. In order to perform the Student's t-test first we compute the t-value, then we choose a level of significance (often denoted as α), in our case we will choose the standard $\alpha = 0.05$, this level of significance determines if the observed difference is statistically significant, in other terms α is the probability to reject the null hypothesis despite having the null hypothesis being actually true [47]. The null hypothesis for our test is that the difference between the frequencies of trades at round numbers and nearby unrounded numbers is zero. Once the t-value, level of significance and degrees of freedom have been set, it is possible to determine if the null hypothesis is rejected or accepted by comparing the t-value against the according value of the Student's t table which lists the critical value of t. Therefore, if our computed t is greater than the critical value of t then we can reject the null hypothesis and we can conclude with a 95% of confidence that the two groups do have a significant difference between them, instead if the t-value computed is less than the critical t value then we accept the null hypothesis meaning that we do not have enough evidence to conclude that there is a significant difference.

Our expectation for the experiment is that honest exchanges containing real transactions will reject the null hypothesis, this is due to the accumulations of trades on round numbers which would cause a significant difference with the nearby unrounded trades taken in exam. According to the Student's t-table, the critical t-value for a confidence level of 95% with degrees of freedom equal to 16[1] is 2.12, this value will be used as a threshold to evaluate if reject or not the null hypothesis. Table 6.5 shows the t-values computed for the transactions of every pair of the exchanges, the highlighted exchanges are the ones which have a t-value smaller than 2.12 for every pair.

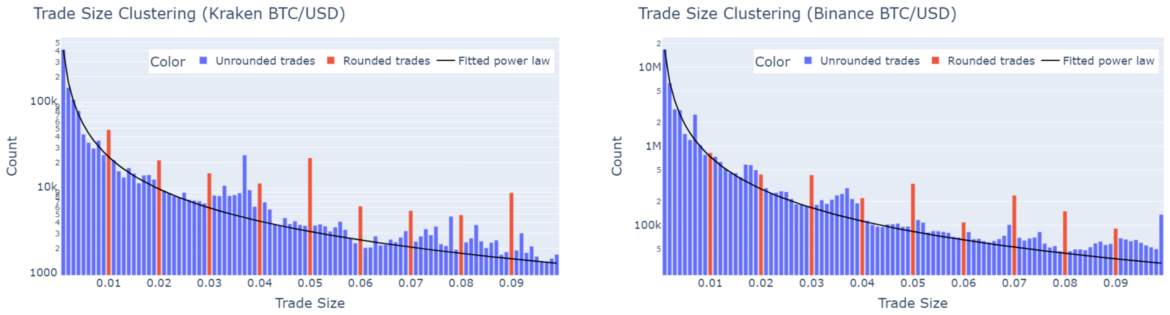
The trusted exchanges are once again in line with expectations regarding the transaction data. As a matter of fact, the t-values are greater than the critical t-value, therefore proving the significative difference between rounded trades against unrounded trades. Figure 6.4 depicts the plots for the trade size clustering for Kraken and Binance for the BTC/USD, a qualitative assessment of the plots can reveal the "bumpiness" property regarding the rounded trades as well as a trend of the amount of transactions which fits a power law.

The power law distribution (also known as Pareto distribution) is a probability distribution

[1] $n = 9$ in our data set, therefore $df = 9 * 2 - 2 = 16$

	BTC-USD	ETH-USD	ADA-USD	XRP-USD
Binance	3.6	4.74	4.4	3.6
Kraken	4.47	7.1	5.024	4.13
Gemini	3.12	3.82	-	-
LBank	2.15	0.8	1.53	0.11
Hotcoin	1.5	1.5	0.92	1.67
Bitmex	0.52	2.07	0.41	0.48
BTSE	29.3	1.24	2.5	1.23
HitBTC	1.75	1.29	0.69	1.42
Hotbit	1.05	2.33	1.23	0.31
Changelly	1.58	1.78	0.39	1.51

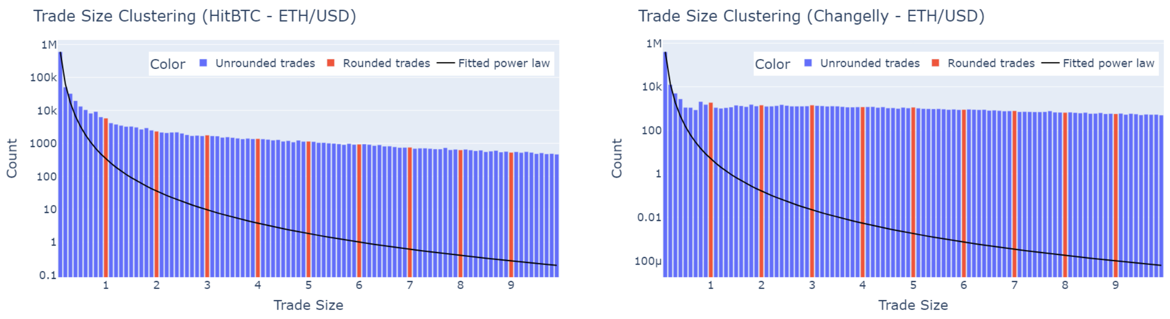
Table 6.5: t-values for the pairs of each exchange.



(a) Trade size clustering for BTC/USD (Kraken).

(b) Trade size clustering for BTC/USD (Binance).

Figure 6.4: Kraken and Binance show accumulations on rounded trades and follow a power law trend (vertical axis is log-scaled).



(a) Trade size clustering for ETH/USD (HitBTC). (b) Trade size clustering for ETH/USD (Changelly).

Figure 6.5: HitBTC and Changelly do not show accumulations on rounded trades and present a uniform distribution (vertical axis is log-scaled).

where the frequency of an event is inversely proportional to the power of its magnitude. The distribution is expressed in the form of $P(x) = Cx^{-\alpha}$ where $P(x)$ is the probability of the event with frequency x, C is a constant and α is the exponent that determines the shape of the distribution. Power law distributions are often present in economics and finance in data involving market trades, stock market returns, income and wealth, etc. [48]. It is reasonable to expect the same behavior in our transaction data set where the Pareto principle should in theory apply, the idea is that the majority of transactions should be more concentrated in the area of the plot with smaller trade size transactions. The power law distribution and the trade size clustering are visibly respected for the trustworthy exchanges, the same cannot be said about the remaining exchanges because as table 6.5 shows the majority fails in at least one of the pairs. Furthermore, the distribution of the trade sizes depicted in 6.5 is apparently a uniform distribution that does not reflect the Pareto principle, in fact, a small high peak is shown only on the first part of the plot followed by a more “flat” behavior. Note that this cannot be proven as direct evidence of manipulation of data, but certainly raises doubts about the legitimacy of the data examined, it cannot be excluded that there are possible irregularities with the data which could be intentional or not.

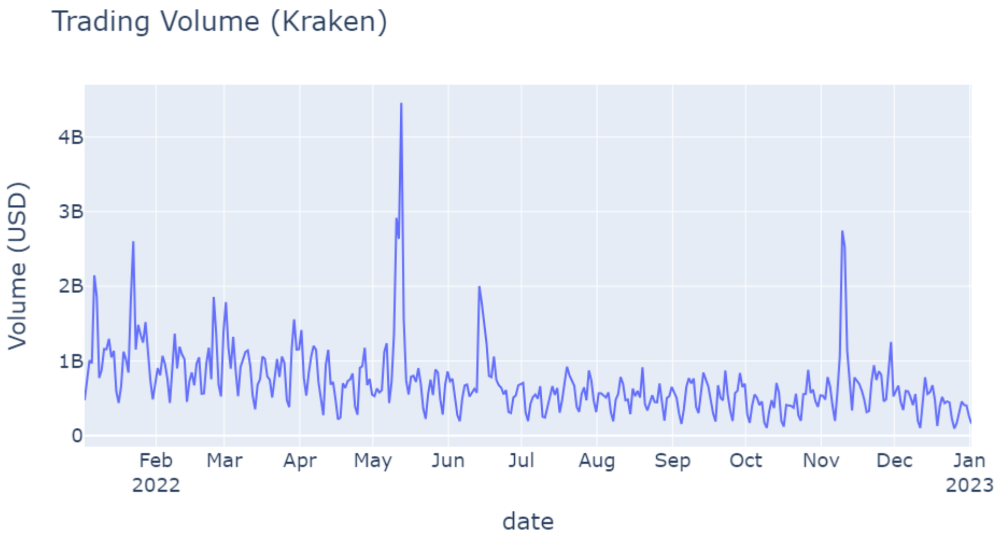


Figure 6.6: Trading Volume (Kraken).

6.2 Anomaly Detection with ML

6.2.1 Trading volume - Bitcoin Price Association

For our ML analysis, we used trading volume data instead of transactions data, this dataset is composed of 47 exchanges whose hourly trading volumes (expressed in USD) have been collected for the whole year of 2022. The fine-granularity of our dataset (hourly volumes instead of the more standard daily volumes) allows us to better capture sudden rises in volume which tend to happen when a sudden price movement is recorded.

A common pattern in trading volume data is the presence of sudden peaks which are often associated with sudden price movements in the market, e.g. a real-life event could have an impact on the value of Bitcoin, thus, the crypto market "moves" towards an uptrend which motivates traders to speculate by creating more buy orders that implies an increase in the trading volume. The same principle applies to real-life events which result in a decrease of value for a coin, meaning that there will be more sell orders and an increase of the trading volume.

Figure 6.6 shows the trading volume for the exchange Kraken throughout 2022, an example of a sudden peak is noticeable in the month of May where the volume reached a peak of $4.4 billion USD in contrast to an average trading volume of $716 million USD. The reason behind this burst is the news of the de-pegging of the stablecoin UST from the US dollar [49], an event that shocked the crypto market with BTC suffering a 20.08% drop in value during May 2022 [50].

6.2.2 Trusted Exchanges Correlation

The next analysis aims to compare the trading volumes of trusted exchanges which are combined to define a baseline against the trading volumes of other exchanges, in this way, it is possible to understand if there are notable differences and how the comparison holds out. Firstly, we define once again the group of trustworthy exchanges by expanding the group list, this is due to the fact that having a broader group would allow a more comprehensive analysis. The trustworthy group is composed of the exchanges Kraken, Coinbase and Bitstamp which are regulated in the U.S. and Bitfinex, Poloniex and OKX which despite not possessing a U.S. license are well-known exchanges that have operated for many years and built a good reputation, as a further indicator of trust we have based the decision on the exchange score provided by CoinMarketCap which is a score used to evaluate the legitimacy of an exchange's trading volume. Binance and Gemini have not been included in the trustworthy group because they significantly differ in size with magnitudes of order of difference w.r.t. the other exchanges. Binance and Gemini have an average volume of respectively $15 billion USD and $102 million USD against the average volume of the trusted group of $875 million USD. Figure 6.7 shows a heat map[2] of the Pearson Correlation Coefficient (PCC) values between the trading volumes of trusted exchanges and in figure 6.8 the heat map of the correlation between the trading volume of the baseline against the other exchanges.

Volumes of the exchanges are all influenced by the same factors related to the crypto market,

[2] A heat map is a kind of graphical representation of the magnitude of a certain value using colors, usually a darker color signifies a higher value

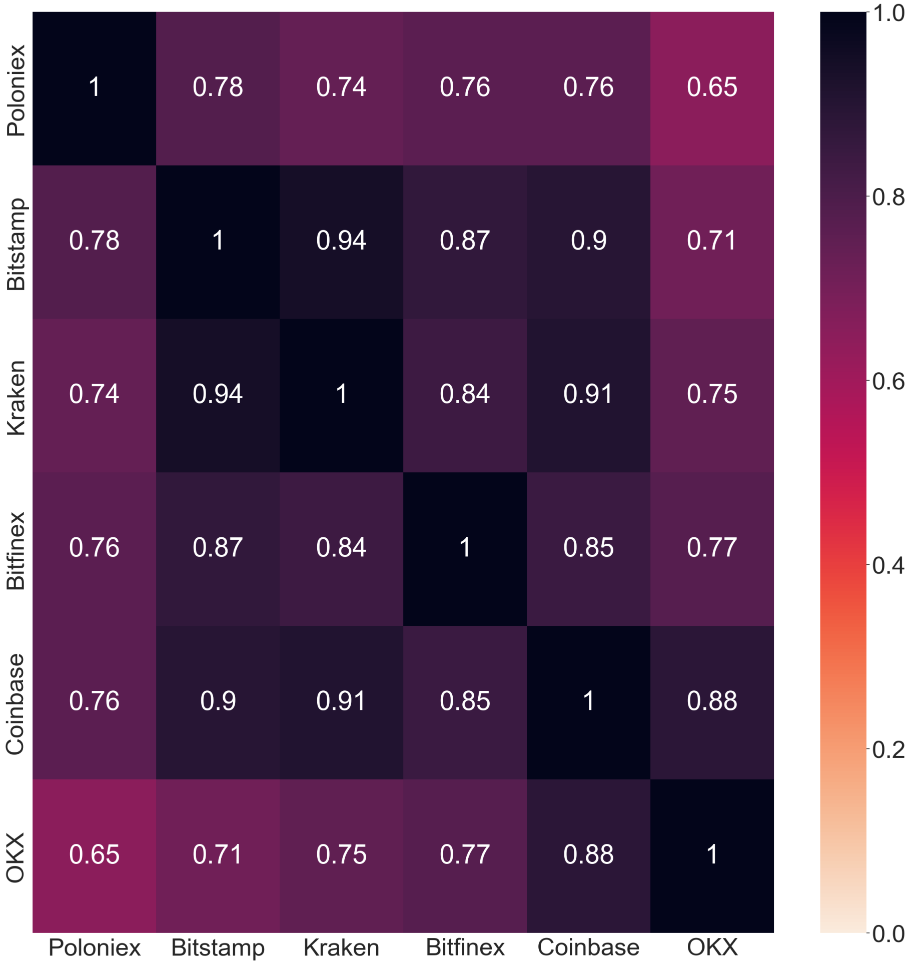


Figure 6.7: Correlation of the trading volume among the trusted exchanges.

consequently, exchanges should in theory follow the same trend but with different magnitudes based on the size and market share of the exchange. Note that it is unreasonable to expect that the trading volume between two exchanges possess a behavior perfectly identical but it is expected to have at least a moderate correlation between them since they are both under the same market conditions.

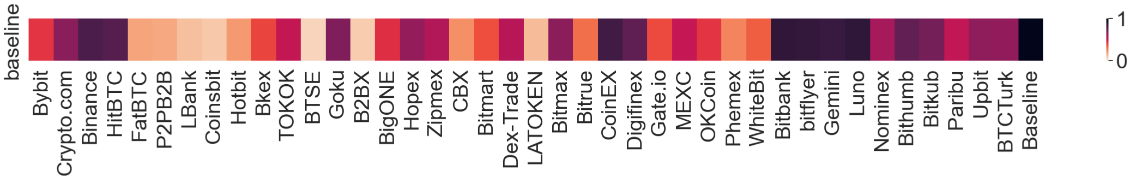


Figure 6.8: Correlation of the trading volume of the baseline against other exchanges.

6.2.3 Convolutional Autoencoder Model

Next, we built a reconstruction convolutional autoencoder model to perform anomaly detection on the temporal series characterized by trading volumes. An autoencoder is a type of neural network that learns the encoding function which is responsible to encode the input data into a lower-dimensional representation (known also as latent/hidden space) and the decoding function which tries to reconstruct the original input data from the lower-dimensional representation [51]. The model is structured into an encoder and decoder (figure 6.9 represents a graphical schema of the model), the two components of the model allow the learning of an efficient representation.

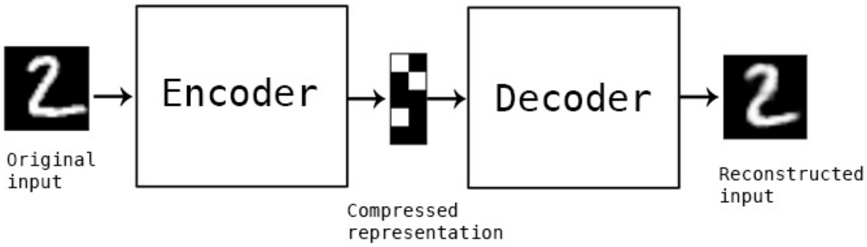


Figure 6.9: Graphical representation of an autoencoder model (image from The Keras Blog).

An autoencoder performs unsupervised learning where the dataset is unlabelled, our temporal series for the trading volumes is such an example. This kind of model is mainly used for tasks such as dimensionality reduction - where the model tries to learn a mapping between the high-dimensional input data and the low-dimensional representation, so the encoder can be used to reduce the dimensionality in tasks like data visualization, e.g. PCA (Principal Component Analysis) is a technique which can be used to visualize multidimensional data by exploiting dimensionality reduction [52], data denoising - autoencoders learn how to remove noise from the original data which can be helpful in case of noisy or incomplete datasets [53] and feature extraction - which is used to capture the most relevant information from some data in order to reduce the dimensionality [54]. Autoencoders can also be used for anomaly detection, which is a process typical of the data mining field where we try to determine anomalies in the dataset. The model is trained on normal datasets which do not present anomalies and then it is used to reconstruct new data instances, if the reconstruction error is greater than a certain threshold then the data point is considered an outlier or anomaly in the dataset [55].

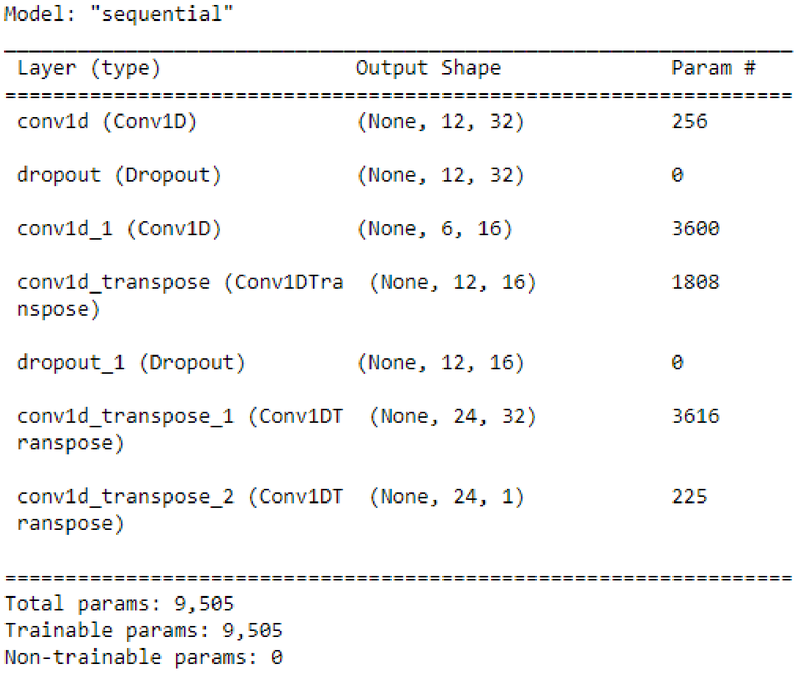
Mapping anomalies of the trading volumes could be crucial to identify possible unexpected scenarios or to gain insight into the particular exchange. Market manipulation could be a kind of anomaly, unusual peaks and depths in the trading volume could be caused by illicit activities such as: pump-and-dump schemes, which is a form of market manipulation where a group of people artificially inflate the price of a certain asset by inviting other people in buying the asset at the same time to sell it at a higher price [56] [57], insider trading, where for example an actor inside a company trades some assets based on non-public information or material [58]. Besides possible market manipulation, another kind of anomaly is sudden price movements which result in sudden increases in trading volume. These kinds of price movements are usually motivated by the beginning of a new trend for a certain coin price, a change in the sentiment of the market or real-life events such as new regulations regarding the crypto market.

The framework used for our neural network is Keras[3], which is a deep learning framework written in Python that runs on top of the machine learning platform TensorFlow[4]. The model is structured as shown in figure 6.10 as a standard convolutional autoencoder with two layers of 1-d convolution for the encoder, which is commonly used for temporal data [59], and a dropout layer in between and accordingly the decoder has two layers of 1-d deconvolution for the reconstruction of the input data.

The activation function used is the ReLU (Rectified Linear Unit) and the optimization algorithm used is Adam with a learning rate of 0.005 and MSE (Mean Squared Error) as the loss function. The model is trained for 100 epochs with a batch size of 128, the learning curve of the training for the model used the exchange Binance can be seen in 6.11a as well as the reconstruction error for a single datapoint in 6.11b.

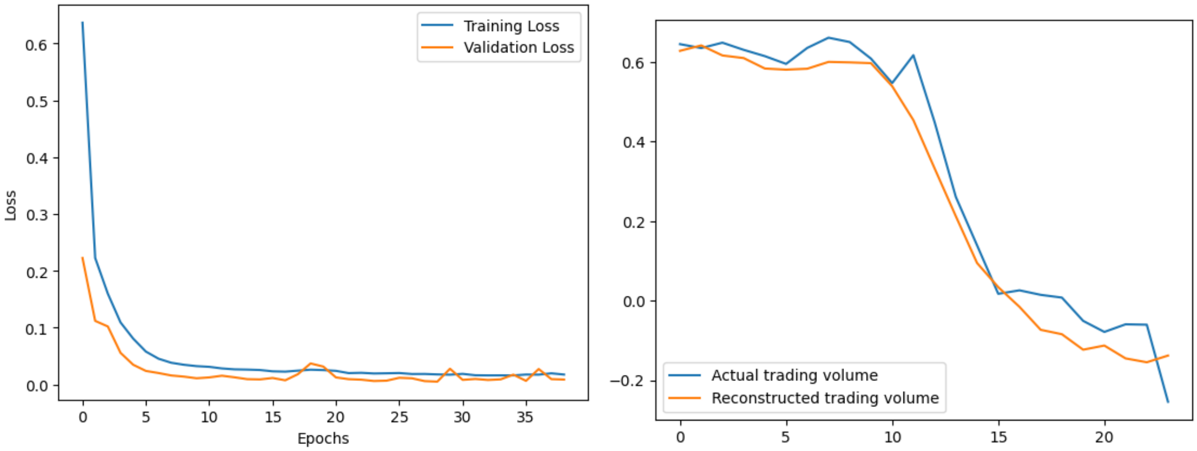
[3] https://keras.io/about/

[4] www.tensorflow.org/



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 conv1d (Conv1D)             (None, 12, 32)            256       
                                                                 
 dropout (Dropout)           (None, 12, 32)            0         
                                                                 
 conv1d_1 (Conv1D)           (None, 6, 16)             3600      
                                                                 
 conv1d_transpose (Conv1DTra  (None, 12, 16)           1808      
 nspose)                                                         
                                                                 
 dropout_1 (Dropout)         (None, 12, 16)            0         
                                                                 
 conv1d_transpose_1 (Conv1DT  (None, 24, 32)           3616      
 ranspose)                                                       
                                                                 
 conv1d_transpose_2 (Conv1DT  (None, 24, 1)            225       
 ranspose)                                                       
                                                                 
=================================================================
Total params: 9,505
Trainable params: 9,505
Non-trainable params: 0
```

Figure 6.10: Summary of the convolutional autoencoder model in Keras.



(a) Learning curve for the training on Binance. (b) Reconstruction error for a single data point.

6.2.4 Anomaly Detection Analysis

In order to perform the anomaly detection we first normalized the values of the dataset and partitioned the dataset into training data, which is made by a temporal series of 10 days with no anomalies and test data which consists of the whole temporal series, then after creating and training the model with Keras, the autoencoder was able to reconstruct the original input.

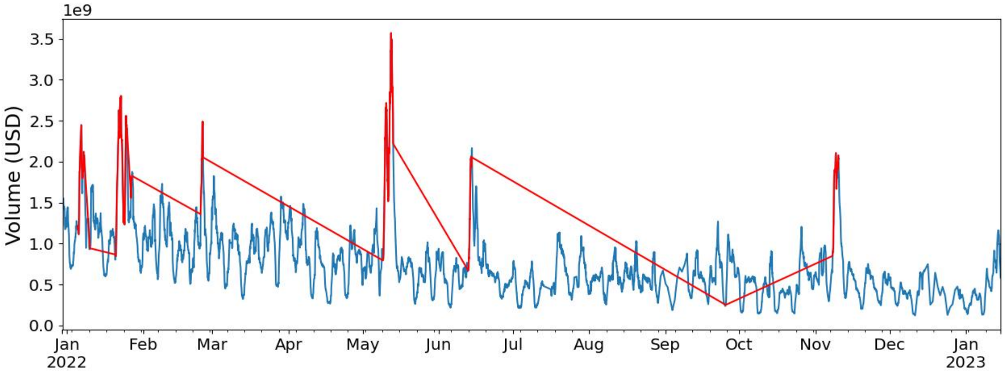


Figure 6.12: Legitimate outliers (in red) corresponding to price movements.

Thus, the detection of the anomaly is based on the reconstruction of the data, first, we find the maximum value of the MAE (Mean Absolute Error) loss on training data, this value represents the worst possible error that the model can make in reconstructing the input, so we can use it as a threshold to detect anomalies. In fact, if the model encounters data points that have a reconstruction loss greater than the MAE, it means that the data point does not belong to the trend. Finally, the model is applied to the test data, in which the trading volume of the exchange and the anomalies which have been found are plotted. The data time frame covered is the whole year of 2022 and the first six months of 2023.

The baseline trend made by the trusted exchanges has been used to detect anomalies that were motivated by price movements, so being legitimate outliers and not anomalies they have been discarded from the process of anomaly detection of other exchanges. The data points with the associated price movement expressed as the price change from the highest high price point to the lowest low or vice versa for a decrease[5]) which were discarded are (shown in figure 6.12) and listed in table 6.6.

Some of the exchanges with the highest percentage of anomalies are Coinstore with 2.53% and Whitebit with 2.57%. The results of the anomaly detection process are shown in figure 6.13.

For the exchange Coinstore the anomalies in trading volume span from 11/04/23 to 15/04/22,

[5] BTC price movements are chosen as being the most dominant coin on the crypto market, therefore, having a greater magnitude in the market

DATE	PRICE CHANGE
05/01/22 - 07/01/22	-12.68%
21/01/22 - 27/01/22	-11.97%
24/02/22 - 25/02/22	-11.7%
09/05/22 - 12/05/22	-22.54%
13/06/22 - 14/06/22	-21.14%
07/11/22 - 09/11/22	-26.1%
14/03/23 - 15/03/23	+20.3%

Table 6.6: Discarded data points with the dates and the price movement associated.

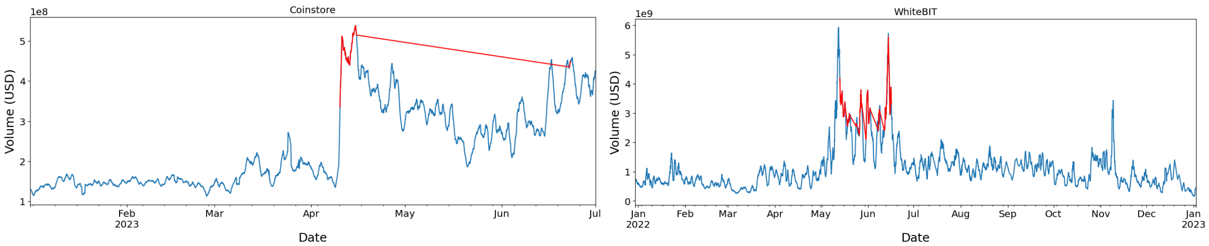


Figure 6.13: Anomalies detected (in red) for Coinstore and Whitebit.

with a sudden peak on the first day and an overall increase in traded volume for the following weeks, the same scenario happens for the traded volume for WhiteBIT where the time frame ranges from 14/05/22 to 14/06/22, also in this case we have a sudden peak followed by a structural increase in trading volume.

In order to assess the “legitimacy” of these anomalies we use traffic data regarding the website daily visits[6] of the exchanges where available (data ranging from April 2023 to June 2023) as a benchmark to test whether the significant increase in trading volume is supported by a respective increase in website accesses. Google Trends data[7] tracks the interest over time for a certain topic, thus, it can be used to track the interest for exchanges and to check whether the increase in trading volume is matched with an increase in Google searches, we use this

[6] Web traffic is provided by SimilarWeb (similarweb.com/it/)

[7] https://trends.google.it/trends/

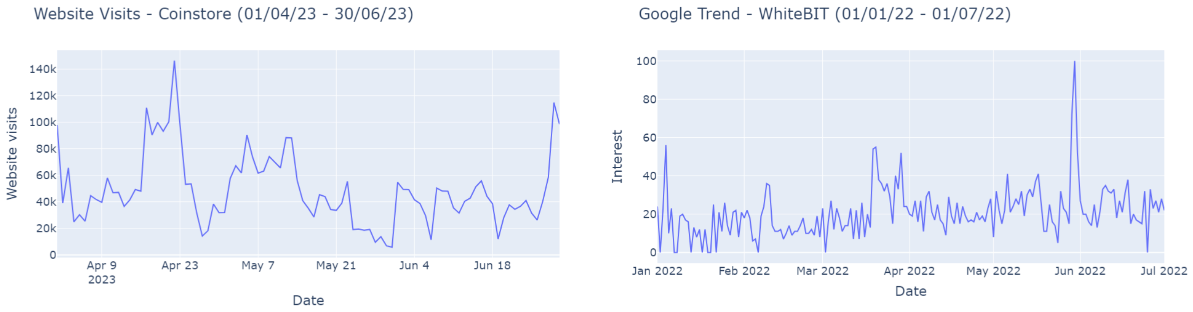


Figure 6.14: Website visits for Coinstore and Google Trend for WhiteBIT

data when more precise web traffic data is not available. Figure 6.14 shows on the left side the trend for the web traffic of Coinstore (data ranges from April 2023 to June 2023) and on the right side the interest over time for the exchange WhiteBIT (data ranges from January to July 2022).

Web traffic data regarding Coinstore have somewhat correspondence with the peak in trading volume reported, although with a slight delay, in fact, the peak in trading volume is registered on 11/04/23 whereas the increase in visits is on 17/04/23, suggesting that the higher number of website visits could be the cause of the increase in volume (meaning that possibly a higher reported volume attracted more users due to better ranking on websites like CoinMarketCap and Coingecko) and not a phenomenon which happened concurrently. Furthermore, the increase in trading volume is remarkably high compared to the increase in web traffic, volume started from an average of $135M USD and reached a peak $500M USD totaling a 5x increase whereas the number of website visits increase only of 2x (going from 47K daily visits to 110K) for a shorter period of time. On the other hand, WhiteBIT showed a considerable increase in trading volume on 09/05/22 which is motivated by a -22.54% change in price for Bitcoin (see 6.6), which is followed by an uptrend for the following month. This prolonged increase is likely due to the publicity of the news that WhiteBIT donated almost $1M USD to Ukraine by winning an online auction for the Eurovision's 2022 trophy that was won by the Ukrainian artists [60]. This is corroborated also by the Google Trends data that shows a great peak in the interest of Whitebit on 30/05/22 which coincides with the day of the diffusion of the news, note that a greater interest on WhiteBIT does not mean necessarily higher traffic for the website but it can partly justify the higher trading volume. These two examples show how anomaly detection for trading volumes can be used to filter possible anomalous scenarios that can be investigated in detail with external data.

6.3 Estimate of Normalized Volumes

For the last analysis we use web traffic data[8]about the monthly number of website visits of the exchanges to normalize the volumes reported. The trading volumes are in the time frame of the month of May 2023, the dataset contained a total of 340 exchanges which have been filtered to the top 50 by reported traded volume. The goal is to try to estimate the real volumes by using the web traffic as a proxy for the actual popularity and market share of the exchange, web traffic should reveal if the trading volume is proportional to the amount of users of the exchange. This method excludes users who access the services of the exchange through different means such as: the mobile application or trades made using the exchange API, we are aware that this constitutes a simplification but the aim of the goal of this technique is to give a rough estimate of the normalized volume. The normalized volumes are computed by firstly gathering the daily trading volumes and web traffic, then for each one of the exchanges we compute the volume per visit which is then compared with the one computed from the baseline trend. In order to compute the volume per visit we compute the value $vol_{vis} = \frac{vol_{rep}}{\#visits}$, where vol_{vis} is the volume per visit, vol_{rep} the volume reported and $\#visits$ the number of visits of the website. The volume per visit of the baseline trend made by trusted exchanges[9] is computed as $\mu = \frac{1}{n}\sum_{i=1}^{n} vol_{vis_i}$, the value computed is $\mu = \$2670$ USD. This value is used to give an estimate of the real normalized volume by computing $vol_{norm} = \mu * \#visits$, where vol_{norm} is the normalized volume. Finally, we compute the ratio of the normalized volume over the reported volume by computing $\alpha = \frac{vol_{norm}}{vol_{rep}}$, this coefficient is used to determine in what quantity the reported volume is actually genuine. In figure 6.15 are depicted the exchanges

[8] Web traffic is provided by SimilarWeb (similarweb.com/it/)

[9] The trustworthy group of exchange is once again composed by Kraken, Coinbase, Bitfinex, Bitstamp, Poloniex and OKX.

in exam with their reported volume (in blue) and the normalized volume (in red). The chart shows that the normalized volume of the majority of the exchanges equals or even surpasses the reported volume, whereas a non-inconsiderable minority shows a very low reported volume compared to the normalized one. Furthermore, figure 6.16 shows the ratio α for the top 50 exchanges, since it is reasonable to expect them to have a normalized volume that is at least half of the reported volume, we set arbitrarily the threshold of "suspicious" volume to a ratio which is less than 0.6. The exchanges in red in figure 6.16 are those that exchanges which have a ratio less than 0.6 and that indicate a possible situation of volume inflation. The result of the analysis is that 9 exchanges out of the top 50 exchanges have a ratio less than 0.1, showing a strong indicator of possible dishonest practices whereas 7 exchanges have a ratio between 0.1 and 0.6, the remaining 33 exchanges show a volume which is in line with the estimate. The exchanges which do not have a reported volume that fits the normalized volume for this test are reported in table 6.7 together with the coefficient α.

	α
Nami	0.0024
Bitubu	0.009
Bullish.com	0.01
CoinDCX	0.017
BTC-Alpha	0.058
Bibox	0.065
Bit.com	0.073
HitBTC	0.083
Bitrue	0.092
Coinstore	0.28
Dcoin	0.34
Bitforex	0.42
AZBit	0.47
Pionex	0.48
Changelly	0.52
Yobit	0.59

Table 6.7: Exchanges which have a coefficient less than 0.6.

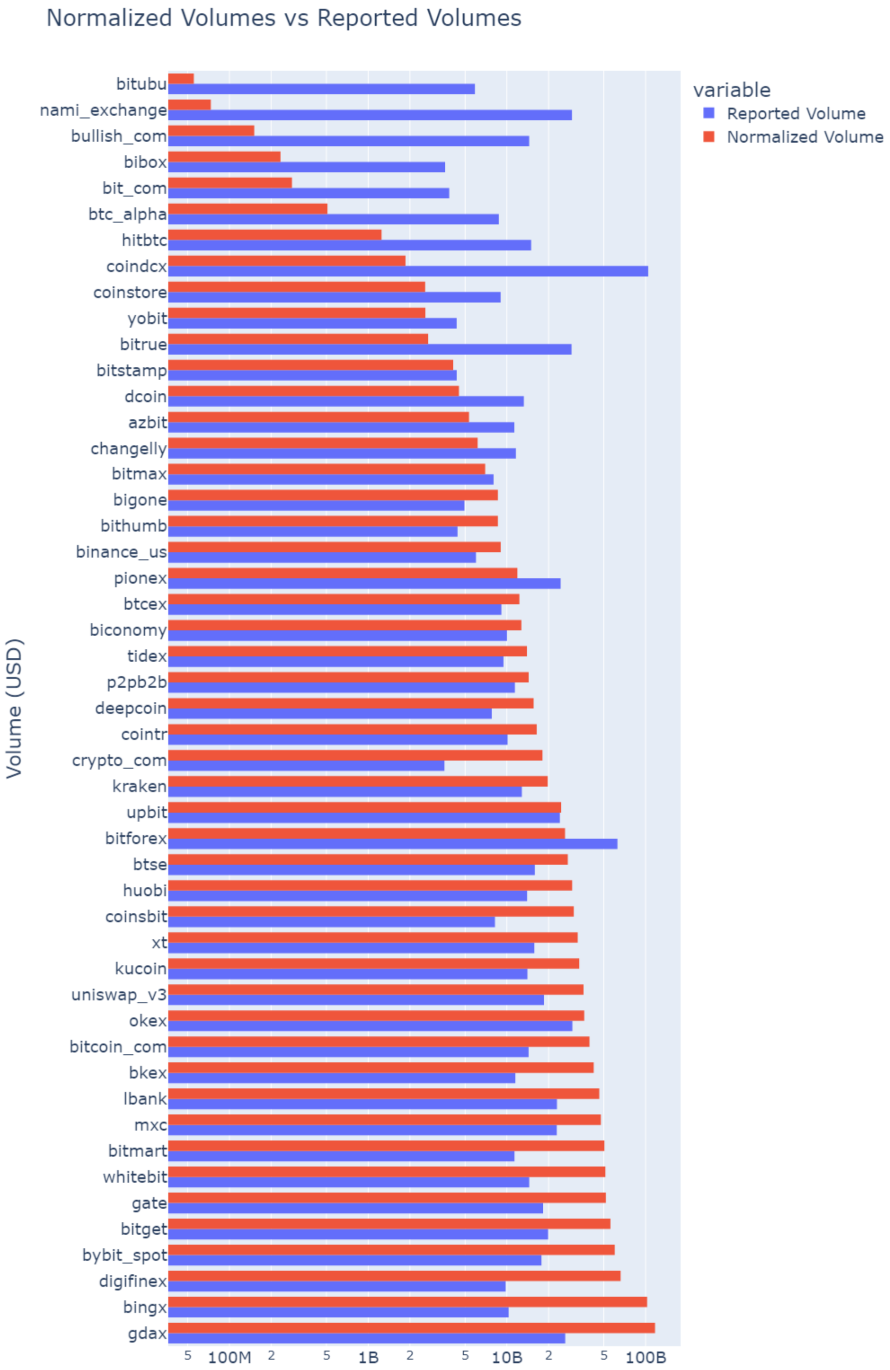


Figure 6.15: The top 50 exchanges with reported volume (in blue) and normalized volume (in red) (horizontal axis is log-scaled).

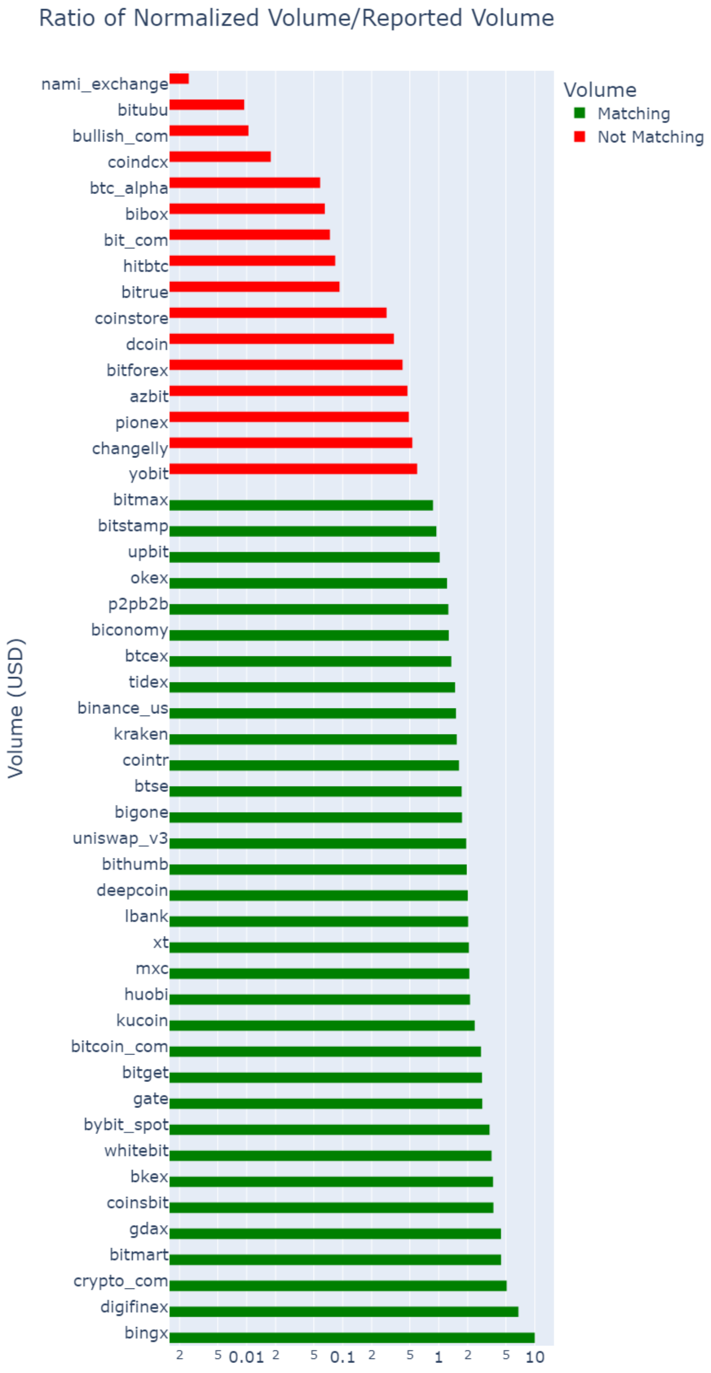


Figure 6.16: Ratio of normalized volume over reported volume (horizontal axis is log-scaled).

Chapter 7

Conclusion

Our work focuses on discovering dishonest practices such as wash trading and fake volumes employed by crypto exchanges by using public market data like trading volumes, transaction data, and web traffic data. The analysis conducted shows discrepancies between the values of the trusted exchanges group that was used as a reference for real, transparent data and the group of exchanges suspected of potential forms of market manipulations. Several statistical tests are performed on the data collected for the overall period between January 2022 and June 2023, the results of these tests represent indicators of possible anomalies present in the data.

The first test is the association between BTC dominance, which was close to 47%, and the partition of volumes in an exchange. The results indicate that BTSE does not follow BTC dominance, the BTC/USD market (27%) is smaller than the market for ETH/USD (39.9%), furthermore, LBank has a suspiciously comparable amount of market share among BTC/USD (26.6%), ETH/USD (23.6%), ADA/USD (20.1%) and ETH/BTC (20.9%).

Benford's Law test is then applied to determine if the transactions were naturally generated or suspected automated trades. The law defines the distribution of the first significant digit of the value of the trades as more likely to be a smaller value. The test shows that the exchanges with the highest deviation from Benford's Law (i.e. the less likely human-generated data series) are LBank with an average of 5.48%, BTSE (4.78%), and Hotbit (4.07%).

Next, we examine the Pearson correlation (r) between the number of transactions and the trading volume of every exchange, the expected result of this test is to find a positive correlation between the two measures. We find that the exchange Hotcoin has the lowest correlation coefficient (r=0.39) showing low correlation, whereas Changelly (r=0.53), LBank (r=0.57) and HitBTC (r=0.57) show moderate correlation, instead, the remaining exchanges all show high or very high correlation (r > 0.68). The low correlation between the two features could indicate a doubtful situation where either transaction data or trading volume present some kind of irregularity.

The distribution of traded amounts is examined to establish if transaction data could have been manipulated by fabricating non-existent trades that are not associated with users. A genuine transaction dataset should have an accumulation of transactions on amounts with round numbers, in fact, it is human nature to use cognitive reference points for the amount traded. This is why we employ the Student's t-test to find out if there is a significant difference between the group of transactions with round numbers and unrounded numbers. In the case of honest exchanges we expect to have these "bumps" on round numbers, thus, having t-values, which indicate the strength of difference, greater than the t-critical value (in our test it is 2.12). The results demonstrate that in the exchanges HitBTC, Hotcoin, Bitmex, and Changelly the accumulations of trades on round numbers do not occur in any of the market pairs analyzed, showing that these exchanges could potentially have faked their transactions data.

Next, we propose a ML approach by using a convolutional autoencoder to perform anomaly detection on trading volumes where the detection of the anomaly is based on the reconstruction error of the data. In the case of a data point that presents a reconstruction error higher than a threshold, the data point is labeled as an anomaly. This process is used first on a baseline formed by the trading volumes of the trusted exchanges, this allows us to map the peaks in trading volume with the price movement of Bitcoin that cause them. These anomalies in trading volume that were motivated by external factors are discarded in the anomaly detection process of the exchanges. Furthermore, we integrate web traffic and Google Trends data to discover the cause of the anomalies which were present in Coinstore and WhiteBIT (two of the exchanges with the highest percentages of anomalies). For the exchange Coinstore there is a correspondence during a peak in trading volume, which was registered six days before a peak in the number of website visits. This fact could signify that the peak in reported volume by Coinstore is the source of the peak in the number of website visits, meaning that the higher reported trading volume potentially attracted more customers and thus more profits. On the other hand, for WhiteBIT the considerable increase in trading volume is backed by web traffic data. Finally, we try to estimate adjusted volume by computing a baseline created by averaging the trading volume of trusted exchanges and by computing the legitimate expected volume per visit and then multiplying this value by the website visits of the exchanges. A ratio between adjusted volume and reported volume is then used to determine the exchanges that could have faked their volumes. The result is that nine of the crypto exchanges in the top 50 rankings have an estimated volume that is less than 10% of the reported trading volume.

In general, our findings are that exchanges such as LBank and HitBTC are the ones with the highest probability of faking their volumes. They appear to do so by creating fake orders, for example automatically created by bots as LBank could have done since it failed many of the statistical tests on transaction data. They also appear to be fabricating the reported volume so that it is not comparable with the number of website visits. This could have happened with HitBTC which has an unrealistic number of website visits (304K for May 2023) for an exchange of the magnitude of billions of dollars of trading (reported volume for May 2023 is $14.9B). These scenarios call for a more attentive and detailed regulation of crypto exchanges as well as a need to inform customers about exchanges that perform these illegal activities and these forms of market manipulations that can alter the perception of the crypto market.

Bibliography

[1] Satoshi Nakamoto and A Bitcoin. A peer-to-peer electronic cash system. *Bitcoin. "https://bitcoin.org/bitcoin.pdf"*, 2008.

[2] CNBC. *One in five adults has invested in, traded or used cryptocurrency, NBC News poll shows.* https://www.cnbc.com/2022/03/31/cryptocurrency-news-21percent-of-adults-have-traded-or-used-crypto-nbc-poll-shows.html, 2022. Accessed: 2023-07-11.

[3] Massimo Bartoletti, Stefano Lande, Andrea Loddo, Livio Pompianu, and Sergio Serusi. Cryptocurrency scams: analysis and perspectives. *Ieee Access*, 9:148353–148373, 2021.

[4] Valeriia Dyntu and Oleh Dykyi. Cryptocurrency in the system of money laundering. *Baltic Journal of Economic Studies*, 4(5):75–81, 2018.

[5] Amit Majumder, Megnath Routh, and Dipayan Singha. A conceptual study on the emergence of cryptocurrency economy and its nexus with terrorism financing. In *The Impact of Global Terrorism on Economic and Political Development: Afro-Asian Perspectives*, pages 125–138. Emerald Publishing Limited, 2019.

[6] Nicolas Christin. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd international conference on World Wide Web*, pages 213–224, 2013.

[7] Chainalysis. *2023 Crypto Crime Trends: Illicit Cryptocurrency Volumes Reach All-Time Highs Amid Surge in Sanctions Designations and Hacking.*

https://blog.chainalysis.com/reports/2023-crypto-crime-report-introduction/, 2023. Accessed: 2023-07-11.

[8] Wired. *The Inside Story of Mt. Gox, Bitcoin's $460 Million Disaster.* https://www.wired.com/2014/03/bitcoin-exchange/, 2014. Accessed: 2023-07-11.

[9] BBC. *Crypto giant FTX collapses into bankruptcy.* https://www.bbc.com/news/business-63601213. Accessed: 2023-07-11.

[10] Reuters. *Major crypto lender Celsius files for bankruptcy.* https://www.reuters.com/technology/crypto-lender-celsius-files-bankruptcy-2022-07-14/, 2022. Accessed: 2023-07-11.

[11] Sky News. *Crypto hedge fund Three Arrows Capital plunges into liquidation.* https://news.sky.com/story/crypto-hedge-fund-three-arrows-capital-plunges-into-liquidation-12642402, 2022. Accessed: 2023-07-11.

[12] Tom CW Lin. The new market manipulation. *Emory LJ*, 66:1253, 2016.

[13] Dan Amiram, Evgeny Lyandres, and Daniel Rabetti. Competition and product quality: fake trading on crypto exchanges. *Available at SSRN 3745617*, 2020.

[14] The Cryptonomist. *Chainalysis helps Israeli authorities block terrorist activity and seize $1.7 million in crypto.* https://en.cryptonomist.ch/2023/06/29/chainalysis-against-terrorist-activities/, 2023. Accessed: 2023-07-11.

[15] Vitalik Buterin et al. A next-generation smart contract and decentralized application platform. 2014. *URL: https://github. com/ethereum/wiki/wiki/White-Paper*, 2019.

[16] Ethereum Foundation. *What is Ethereum.* https://ethereum.org/en/what-is-ethereum/. Accessed: 2023-07-11.

[17] Ethereum Foundation. *The Merge.* https://ethereum.org/en/roadmap/merge/. Accessed: 2023-07-11.

[18] The Cryptonomist. *NFT CryptoPunks outperform BAYC in value.* https://en.cryptonomist.ch/2022/11/16/nft-cryptopunks-outperform-bayc-in-value/. Accessed: 2023-07-11.

[19] Bhabendu Kumar Mohanta, Soumyashree S Panda, and Debasish Jena. An overview of smart contract and use cases in blockchain technology. In *2018 9th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–4. IEEE, 2018.

[20] Carrefour. *The Food Blockchain.* https://www.carrefour.com/en/group/food-transition/food-blockchain. Accessed: 2023-07-11.

[21] Xu Wang, Xuan Zha, Wei Ni, Ren Ping Liu, Y Jay Guo, Xinxin Niu, and Kangfeng Zheng. Survey on blockchain for internet of things. *Computer Communications*, 136:10–29, 2019.

[22] Yang Liu, Debiao He, Mohammad S Obaidat, Neeraj Kumar, Muhammad Khurram Khan, and Kim-Kwang Raymond Choo. Blockchain-based identity management systems: A review. *Journal of network and computer applications*, 166:102731, 2020.

[23] B. Brindley, J. Law, and J. Smullen. *A Dictionary of Finance and Banking.* Oxford Paperback Reference. OUP Oxford, 2008.

[24] Borsa Italiana. *GLOSSARIO FINANZIARIO - BID-ASK SPREAD.* https://www.borsaitaliana.it/borsa/glossario/bid-ask-spread.html, 2023. Accessed: 2023-07-11.

[25] U.S. SECURITIES and EXCHANGE COMMISSION. *Bid Price/Ask Price.* https://www.investor.gov/introduction-investing/investing-basics/glossary/ask-price, 2023. Accessed: 2023-07-11.

[26] Bitstamp USA Inc. *How does a crypto exchange work?* https://www.bitstamp.net/learn/crypto-trading/how-does-a-crypto-exchange-work/, 2023. Accessed: 2023-07-11.

[27] Danny Nelson John Biggs. *Upbit Is the Seventh Major Crypto Exchange Hack of 2019*. https://www.coindesk.com/upbit-is-the-sixth-major-crypto-exchange-hack-of-2019, 2019. Accessed: 2023-07-11.

[28] Abhishta Abhishta, Reinoud Joosten, Sergey Dragomiretskiy, and Lambert JM Nieuwenhuis. Impact of successful ddos attacks on a major crypto-currency exchange. In *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, pages 379–384. IEEE, 2019.

[29] Lindsay X Lin, Eric Budish, Lin William Cong, Zhiguo He, Jonatan H Bergquist, Mohit Singh Panesir, Jack Kelly, Michelle Lauer, Ryan Prinster, Stephenie Zhang, et al. Deconstructing decentralized exchanges. *Stanford Journal of Blockchain Law & Policy*, 2(1):58–77, 2019.

[30] Lin William Cong, Xi Li, Ke Tang, and Yang Yang. Crypto wash trading. Technical report, National Bureau of Economic Research, 2022.

[31] Cindy Durtschi, William Hillison, Carl Pacini, et al. The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting*, 5(1):17–34, 2004.

[32] Arash Aloosh and Jiasun Li. Direct evidence of bitcoin wash trading. *Available at SSRN 3362153*, 2019.

[33] Nicholas A James, Arun Kejariwal, and David S Matteson. Leveraging cloud data to mitigate user experience from 'breaking bad'. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3499–3508. IEEE, 2016.

[34] Twitter Inc. *Breakout detection in the wild*. https://blog.twitter.com/engineering/en_us/a/2014/breakout-detection-in-the-wild. Accessed: 2023-07-11.

[35] Jialan Chen, Dan Lin, and Jiajing Wu. Do cryptocurrency exchanges fake trading volumes? an empirical analysis of wash trading based on data mining. *Physica A: Statistical Mechanics and its Applications*, 586:126405, 2022.

[36] Wei Cui and Cunnian Gao. Wteye: On-chain wash trade detection and quantification for erc20 cryptocurrencies. *Blockchain: Research and Applications*, 4(1):100108, 2023.

[37] Kraken. *Is Kraken licensed or regulated?* https://support.kraken.com/hc/en-us/articles/360031282351-Is-Kraken-licensed-or-regulated-, 2023. Accessed: 2023-07-11.

[38] Gemini. *Is Gemini licensed and/or regulated?* https://support.gemini.com/hc/en-us/articles/204734485-Is-Gemini-licensed-and-or-regulated-, 2023. Accessed: 2023-07-11.

[39] Binance. *Licenses, Registrations and Other Legal Matters*. https://www.binance.com/en/legal/licenses, 2023. Accessed: 2023-07-11.

[40] CoinMarketCap. *Exchange Ranking*. https://support.coinmarketcap.com/hc/en-us/articles/360052030111-Exchange-Ranking, 2023. Accessed: 2023-07-11.

[41] Rachel M Fewster. A simple explanation of benford's law. *The American Statistician*, 63(1):26–32, 2009.

[42] Xiaoqing Eleanor Xu and Chunchi Wu. The intraday relation between return volatility, transactions, and volume☒. *International Review of Economics & Finance*, 8(4):375–397, 1999.

[43] Agustin Garcia Asuero, Ana Sayago, and AG González. The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1):41–59, 2006.

[44] Tao Chen. Round-number biases and informed trading in global markets. *Journal of Business Research*, 92:105–117, 2018.

[45] Damir Kalpić, Nikica Hlupić, and Miodrag Lovrić. *Student's t-Tests*, pages 1559–1563. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[46] Tae Kyun Kim. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540–546, 2015.

[47] Angus Brown. The strange origins of the student's t-test. *Physiology News Magazine*, 16:19, 2008.

[48] Xavier Gabaix. Power laws in economics and finance. *Annu. Rev. Econ.*, 1(1):255–294, 2009.

[49] The Cryptonomist. *What really happened to Terra LUNA and UST?* https://en.cryptonomist.ch/2022/05/12/what-happened-terra-luna-ust/, 2022. Accessed: 2023-07-11.

[50] Cointelegraph. *Bitcoin falls below $27KtoDecember2020lowsasTether’spegslipsunder$0.99.* https://cointelegraph.com/news/bitcoin-falls-below-27k-to-december-2020-lows-as-tether-stablecoin-peg-slips-under-99-cents, 2022. Accessed: 2023-07-11.

[51] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

[52] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[53] Omar M Saad and Yangkang Chen. Deep denoising autoencoder for seismic random noise attenuation. *Geophysics*, 85(4):V367–V376, 2020.

[54] Yesi Novaria Kunang, Siti Nurmaini, Deris Stiawan, Ahmad Zarkasi, et al. Automatic features extraction using autoencoder in intrusion detection system. In *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 219–224. IEEE, 2018.

[55] Oleksandr I Provotar, Yaroslav M Linder, and Maksym M Veres. Unsupervised anomaly detection in time series using lstm-based autoencoders. In *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, pages 513–517. IEEE, 2019.

[56] Tao Li, Donghwa Shin, and Baolian Wang. Cryptocurrency pump-and-dump schemes. *Available at SSRN 3267041*, 2021.

[57] Jiahua Xu and Benjamin Livshits. The anatomy of a cryptocurrency {Pump-and-Dump} scheme. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1609–1625, 2019.

[58] Mervyn King and Ailsa Roell. Insider trading. *Economic policy*, 3(6):163–193, 1988.

[59] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398, 2021.

[60] Bloomberg. *Eurovision Winner Sells Trophy for $900,000 to Buy Drones for Ukraine.* https://www.bloomberg.com/news/articles/2022-05-30/crypto-exchange-drops-900-000-on-eurovision-mic-to-buy-drones, 2022. Accessed: 2023-07-11.