EXPLORING ARBITRAGE OPPORTUNITIES
IN CRYPTO

by

Dmytro Golovenko

A thesis submitted in partial fulfillment of the requirements for the degree of

MA in Business and Financial Economics

Kyiv School of Economics

2025

Thesis Supervisor: Dr. Matkovskyy Roman ____________________

Approved by

Head of the KSE Defense Committee, Professor Besedina Elena

Date ______________________________

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CEX Centralized exchanges

DEX Decentralized exchanges

USDT Tether coin. A stable coin for United States dollar.

BNB Binance Coin. A coinf of the Binance crypto exchange.

Market Cap Market capitalization. (measured in usd$)

Chapter 1. Introduction

The rapid rise of cryptocurrencies, in both capitalization and adoption attempts many researches to write articles about how traditional finance findings translate or fail to translate in crypto.

Cryptocurrency market is characterized by 24/7 trading(in contrast to stocks market), low barriers to entry, and global accessibility, making it attractive for individuals, invest funds and researches.

Figure 1: Market Capitalization

One of the questions in financial economics is the persistence of arbitrage opportunities. In other words: Price discrepancies across markets that can be exploited for profit without risk.

If such opportunities do exist, it supports the argument that markets are not always efficient at determining prices. This statement may be evaluated as such that contradicts the efficient markets hypothesis: [market efficiency: CFA institute]

While arbitrage has been studied extensively in the context of equity and foreign exchange markets [Mitchell, M., Pulvino, T., & Stafford, E. (2002). Limited arbitrage in equity markets. Journal of Finance, 57(2), 551–584.

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. Journal of Finance, 52(1), 35–55.
Lewellen, J. (2015). "The Cross-section of Expected Stock Returns". Critical Finance Review. 4(1): 1–44.] cryptocurrency markets remain relatively underexplored, despite their structural features (fragmented exchanges, varying liquidity, differing fee structures) that may create or eliminate arbitrage possibilities.

This paper asks: Do arbitrage opportunities exist in centralized cryptocurrency exchanges (CEXs)? Addressing this question is relevant for several groups:

Individual traders, who are interested in researching or trying to exploit arbitrage opportunities. They could benefit from the provided framework for opportunities identifiction and the provided code for it.

Academics, who study how identification of arbitrage opportunities and the opportunity size.

This paper makes three primary contributions to the growing literature on cryptocurrency market efficiency and arbitrage:
First, it constructs a dataset of triangular arbitrage cycles using 1-second Binance Spot data. Binance is the largest and most liquid spot market globally as of may 2025. The paper incorporates taker-side execution prices and per-second aggregated prices and

statistics. The analysis covers Binance trading pairs available during the selected period (may 2025)

Second, it introduces a methodological framework that combines triangular arbitrage detection with a predictive classification model. Unlike prior studies that focus solely on identifying instantaneous arbitrage or comparing theoretical returns of proposed strategies, this paper explicitly examines whether such opportunities persist into the next second after they were identified. That provides a new angle to look on the persistence of this opportunities, and slippage.

Third, it provides empirical evidence on the frequency, profitability, and persistence of triangular arbitrage in centralized cryptocurrency exchange(Binance). Arbitrage profitability is evaluated using taker-side high price, ensuring a conservative estimate of the possible occurred profit.

When reviewing the older studies and models, it was easy to see that that papers had to rely on less granular data due to lower data availability and lower compute resources at that time. Given the context it limited the researchers in the available methods and numbers of questions that can be answered using that data.

The analysis in this paper is intentionally restricted to Binance Spot exchange data. It does not examine cross-exchange arbitrage, derivative-spot price deviations, or arbitrage between exchanges. Execution is modeled exclusively through taker-side prices, and the study does not attempt to model market-making strategies. These boundaries ensure that the results reflect arbitrage opportunities that are realistically observable and measurable using publicly available 1-second historical data, without speculating about other advanced strategies outside the mentioned scope.

In this paper chapter 2: reviews the academic literature on arbitrage. Chapter 3 outlines the applied methodology, and data used in chapter 4. Chapter 4 presents descriptive statistics for both the arbitrage dataset and the engineered model features. Chapter 5 focuses on the results of the methodology: model estimates, coefficients. And in Chapter 6 one can find the summary of the paper and recommendations.

Chapter 2. Industry Overview and Related Studies

On **22 June 2025** the combined value of all listed crypto-assets was **≈ $3.14 trillion** - more than double the level at the start of 2024. The biggest contributor to the market volume is bitcoin(BTC). By 22.06.2025 data it accounted for $2.13 trillion (**65 % of market cap**). **[coinmarketcap crypto index]**

The two next biggest contributor currencies are Ethereum and Tether: 9% for Ethereum(ETH) and around 6% for tether(USDT).

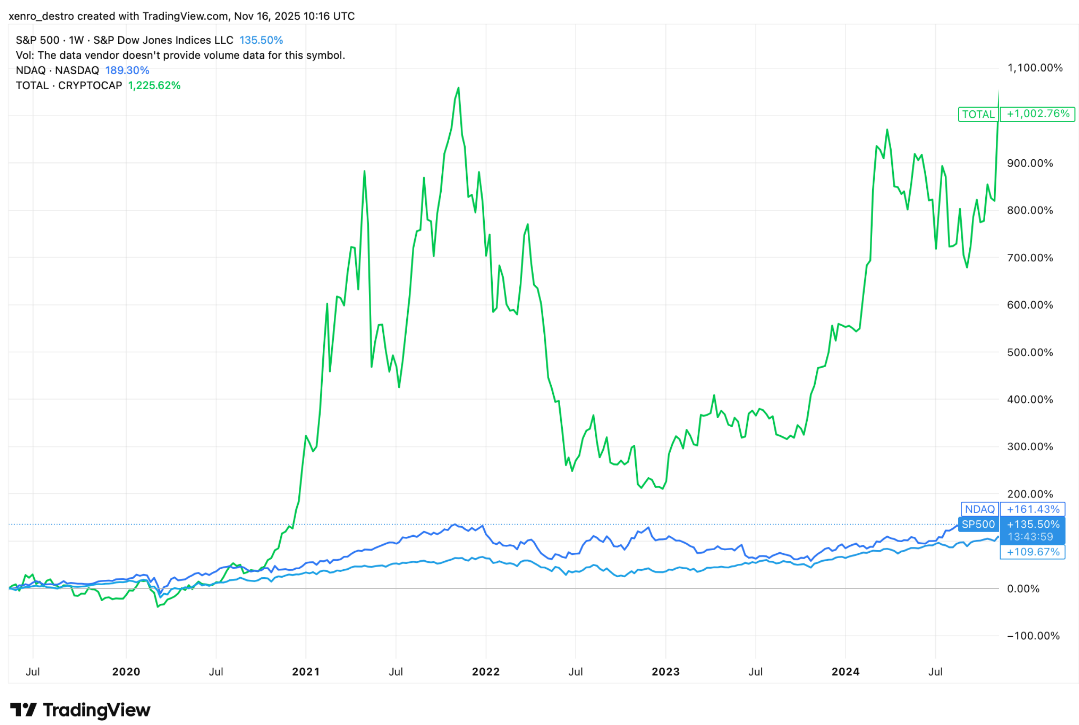
Table 1: Top 10 currencies by capitalization [coinmarketcap , 22.06.2025]:

Rank	Name	Symbol	Market Cap
1	Bitcoin	BTC	$2.13T
2	Ethereum	ETH	$291.15B
3	Tether	USDT	$156.76B
4	XRP	XRP	$128.59B
5	BNB	BNB	$90.74B
6	Solana	SOL	$76.24B
7	USD Coin	USDC	$61.81B
8	TRON	TRX	$25.77B
9	Dogecoin	DOGE	$24.74B

10	Cardano	ADA	$20.08B

The crypto market emerged relatively new when compared to the traditional finance, banking or commodities exchanges. However it demonstrates impressive volumes (according to tradingview chart below). Another feature that is unique to crypto market when compared to the stock market is ease of trade. (due to lower fees, less regulation, 24/7 trade and a large range of assets)

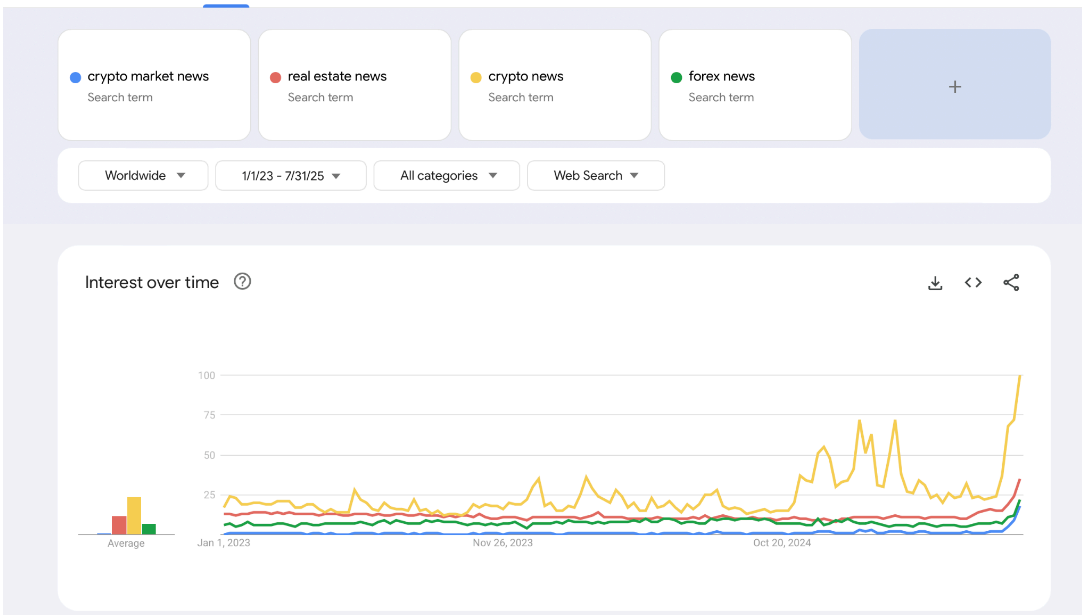
Figure 2: Market volumes



Source: tradingview.com, chart created by author.

Tickets used: SP500, NDAQ, TOTAL: CRYPTOCAP

Figure 3: Google trends



Unlike the stock markets - markets are open 24/7. And individuals face low barriers to entry. For example, an individual can open an account on CEX for free. And to conduct a few trades one would need just a few dollars.

DeFi vs CeFI:

DeFI = decentralized finance. CeFi = centralized finance. CEX = centralized exchange (e.g. Binance, MEX, Bybit)

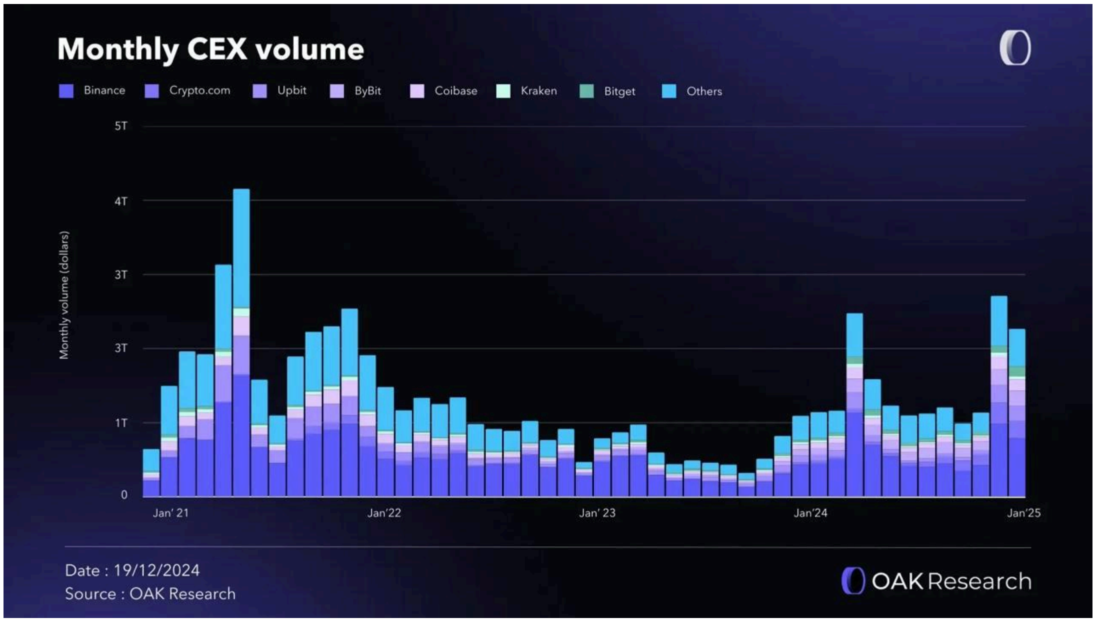


Figure 4, Monthly CEX volume (Source: OAK research)

Centralised exchanges (CeFi) remain the biggest group of players in terms of SPOT trade. (source = OAK research)

Centralised exchanges (CEXs) still have the larger share then DEXs. both CoinGecko's Q1 2025 industry report and OAK research provide estimates on the volume of transaction for ten largest CEX. According to OAK research, top 10 CEX by size finished the q4 2024 year with 5.4 trl usd . Binance remanded the largest centralized exchange by trade volume. (from +-35% to 40% across the year)

CEXs continue to attract traders because they 1) offer deep order books, 2) high liquidity due a higher share of trades 3) less prone to geographic restrictions and regulatory changes.. These structural advantages mean that, even when daily average fees are factored in, triangular or latency-based arbitrage inside a single CEX can still be profitable. This question is examined in the study.

Chart 4, Oak research, CEX vs DEX: Report from 2024 and forecasts for 2025, full version

At the same time DEX is growing. For example here, OAK research shows a chart of share of trades. And we can clearly see how DEX is gaining market share in terms of Futures trades. A broader survey confirms the trend: Binance's spot share has fallen from 52 % in late 2023 to around 40 % today. cointelegraph.com[1]

However CEX still remains the dominant player for Spot exchange. Dex share for spot exchange still remains at around 10% share in 2025. (10% in terms of volume of trade converted to USDT)

Source: coingecko

arbitrage opportunities

There are several factors why in crypto arbitrage opportunities can arise:

1. Noise trading: Prices may temporarily move away from fundamental values because of uninformed or sentiment-driven trades.[Shleifer & Vishny, 1997]
2. Due to Information asymmetry. Some traders may receive information earlier or have better data than others, which creates short-lived price differences before the market adjusts. [Grossman & Stiglitz (1980)]
3. Due to differences in expectations for traders. Even if all information was public, traders may interpret it differently. Or act on it differently. Each trader may have a different risk preferences, financial constraints, etc.

These frictions echo the broader "limits to arbitrage" framework proposed by Shleifer & Vishny (1997)[2] and suggest that measurable, if fleeting, mis-pricings should exist even in large venues like Binance—precisely the phenomenon investigated in this paper.

Additionally to that factors, there are a lot of the assets and cryptocurrencies traded, among the many exchanges. And test a hypothesis whether this opportunities exists, In this paper I decided to focus on the the CEX aspect of the arbitrage, on the biggest Crypto Exchange - Binance.

The main reason for focusing on binance are:

1) data: Binance platform provides the data with up to 1 second intervals

2) I've already seen the papers on DEXes and this side seems more researched to me.

Related studies:

The topic of arbitrage is a long-established one in conventional financial economics, and a considerable body of literature exists addressing its presence, constraints, and feasibility of exploitation. Despite the differences between conventional and cryptocurrency markets, the same basic questions of consideration remain. This section reviews the most relevant studies and situates the present work within this literature.

Arbitrage in Traditional Financial Markets

Foundational studies by *Shleifer and Vishny (1997)* introduced the concept of **limits to arbitrage**, pointing to the persistence of market mispricing's because of possible constraints on arbitrageurs. Later, Mitchell, Pulvino, & Stafford (2002) offered concrete empirical evidence about the possible presence of arbitrage opportunities even within the equity market. And constraints of capital, liquidity, and risk. These works form the basis for explaining why, if triangular arbitrage exists, in crypto, necessarily, its trading could remain systematically untapped.

Also related to this topic is a study by Lewellen (2015), where the author analyzed the role of return and mispricing in equity markets through cross-sectional techniques. Although this study is unrelated to the topic of crypto-currency arbitrage, this example shows that inefficiencies can be statistically analyzed for larger data sets. This Paper takes a relevant tack by focusing on empirical identification

Arbitrage in Cryptocurrency Markets

Although there is limited academic literature studying cryptocurrency arbitrage, a small set of papers studies price anomalies across decentralized platforms. *Vadym Pakholchuk's* work (2024) analyses arbitrage opportunities within decentralized exchanges through the lens of machine learning algorithms applied to one-minute data, pinpointing inefficiencies created by the rules of price formation of automated market makers. The strategy of this research relies on a cycle-building method, similarly to how it is done in this article. Nevertheless, there are several key differences between the current study and the studies noted previously:

1. it focuses on centralized exchange (CEX) data rather than AMMs,
2. it uses 1-second data, offering significantly higher temporal resolution, and
3. it extends the problem by evaluating the probability persistence, rather than only detecting opportunities.

This distinction is crucial because persistence, rather than one-time detection, determines practical exploitability.

Another relevant line of research involves mathematical frameworks for detecting and testing arbitrage, such as the work of *Stelios Arvanitis*, who models stochastic arbitrage using set-estimation and advanced statistical testing. While mathematically intensive, his methods provide a theoretical foundation for potential future extensions of the present work, particularly with respect to improving detection algorithms or evaluating conditions under which arbitrage is statistically significant.

Industry Dashboards

One of the tools to get data on crypto is dune.com. The platform functions as a website that is connected to Dune database that is created from parsing data from blockchain. Then individuals may access that data, create and share dashboards from it. One of such dashboards on the of arbitrage is "CEX - DEX Arbitrage" dashboard.

This dashboard is different from our paper and papers mentioned earlier in the sense that it uses a bottom up approach and finds the users who already profited from activities that were identified as "arbitrage" by the authors of the dashboard. It is not a perfect representation of arbitrage traders, but it is a good dashboard for presenting evidence that there are actually traders who profited from arbitrage. In my opinion this can be used as evidence supporting the theory that arbitrage opportunities in crypto exist, or as counterexample to theory saying the opposite.

Dashboard: https://dune.com/rig_ef/cex-dex-dash

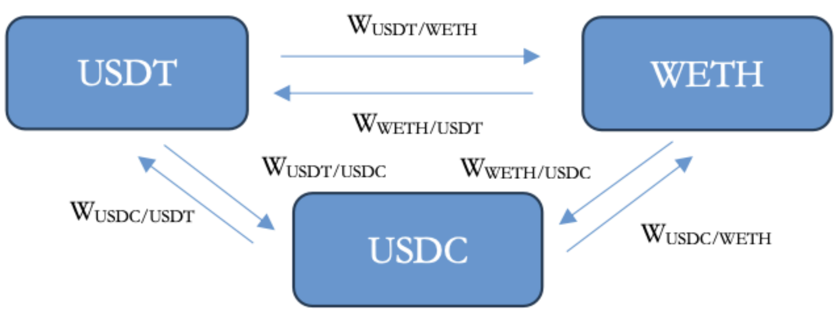
Chapter 3. Methodology

The methodology consists of two essential parts:

Model/algorithm 1:

The goal of the model, algorithm is to find all opportunities for each time period. Then for each token keep up to n best trades.

- at first we consider only 3 trades setup, and look at if they can lead to a profit. This phenomena is already know by the formal name "Bellman- Ford cycle"



The empirical analysis uses Binance Spot market data sampled at a 1-second frequency. The dataset covers the period from 01 may 2025 to 31 may 2025, and includes all trading pairs active on Binance during that interval. All timestamps are recorded in microseconds, consistent with Binance historical data formats.

It is important to mention that in my methodology I use the taker price, so there is no speculation on if a particular new trade would happen under an order that could've been placed according to the strategy. Every trade is in the dataset is real, and was executed in a given timeframe.

Construction of One-Second Price Snapshots

The underlying price and trade data were not artificially interpolated or smoothed;

instead, they rely on the existing 1-second rows available in the dataset. For each symbol I construct directed currency edges by joining the 1-second candles with the symbol table (“available_symbols”). The following view(table) definition generates two directed edges per trading pair for each 1-second interval:

(1) a “base asset => quote asset” edge using the direct quote (based on the high price of the second),

(2) a reverse “quote => base asset” edge using the inverse of the high price.

For each trading pair and each 1-second interval, the datasource provides an option to choose from: "open", "close", high, low. My algorithm uses the “high price” of the candle for buy orders for sell orders as the execution rate for constructing the directed edge. The use of the high and low prices serves as an approximation of worst case price scenario. A taker order is filled immediately against resting liquidity in the order book, and therefore transacts at the least favorable price available at that moment.

Using the high price ensures that the profitability of any triangular cycle is not overstated. If a cycle remains profitable even when evaluated at the worst execution price within the second, then it is more likely to represent an actionable opportunity. In contrast, using last-traded prices or midpoint quotes could introduce optimistic bias by implicitly assuming optimistic execution. The high-price approach therefore prioritizes robustness and reduces the likelihood of identifying false positives in the arbitrage set.

While this approach provides some robustness, it is still not a direct representation of the available arbitrage opportunities. The fact that 3 pairs of tokens where priced in a certain way, and a certain number of transactions happened does not definitely mean that it was possible to place an order that would be executed with a 100% probability.
This uncertainty in the ability to execute an order led me to try a new approach. As a proxy measure to figure out if the opportunity would persist I constructed additional column, that shows if the opportunity persist in the next (1 second) time interval. It is

later discussed in Model 2 and “Chapter 5: Results” sections of the research paper.

Figure 6: triangular arbitrage example

source: APPLYING MACHINE LEARNING MODELS FOR ARBITRAGE OPPORTUNITIES ON DEX, Vadym Pakholchuk, 2024.

To be more specific, my algorithm includes

1) the “fee” parameter as well. The “fee” parameter is included by adding a condition to the algorithm, to filter out the pairs that produce gross profit <= 0.3% after 3 trades. The algorithm does not assume any particular fee, but rather filters out opportunities that result in a low gross profit trades (<=0.3%). The resulting dataset can be later aggregated or filtered even more in the analysis part.

2) “filtering conditions”:

2.1 at least 1 trade must occur in each particular second for each particular 3 currency pair to be included in the “pairs” dataset.

2.2 the starting currency for the 1st buy/sell order is in one of five most popular coins: 'USDT','BTC','BUSD','BNB','TRY'.

The whole file for the algorithm is available in github:

https://github.com/gybzd/binance_arbitrage_research/blob/main/main_folder/generate_triangle_arbitrage_parquet.py

The resulting dataset is 3790 rows long (N=3790), and has a following structure:

Open time	a	b	c	r_ab	r_bc	r_ca	cycle_rate	pct_pnl	path_symbols	P1 trade count	P2 trade count	P3 trade count
17460576 0500	USDT	TRY	CATI	38.55	0.23	0.11	1.00	0.34	{USDTTRY,CATITRY,CATIUS	12	8	45

0000									DT}			
1746 0576 0500 0000	TRY	CATI	US DT	0.23	0.11	38.55	1.00	0.34	{CATITR Y,CATIUS DT,USDT TRY}	8	45	12
1746 0576 0700 0000	BTC	USDC	HB AR	94,18 8.62	5.49	0.00	1.00	0.39	{BTCUSD C,HBARU SDC,HBA RBTC}	34	1	1
1746 0576 0700 0000	BTC	BNB	HB AR	157.0 1	3,294 .78	0.00	1.00	0.36	{BNBBTC ,HBARBN B,HBARB TC}	1	1	1
1746 0576 0700 0000	BNB	HBAR	BT C	3,294 .78	0.00	157.0 1	1.00	0.36	{HBARBN B,HBARB TC,BNBB TC}	1	1	1

Table: pairs structure

-Model 2:

Classification model:

Explained variable: fact of profitable trades occurring at >= in next time interval(1 second). In other words: if the opportunity is present in one particular second for one particular set of currencies and trades - how sure can one be that it would remain in the next second.

Feature/variables engineering:

Predictors are constructed using lagged market activity from previous seconds.

For each triangular cycle, the following features are created:

Lags for each pair for 1 second: p1_trade_count_lag1, p2_trade_count_lag1, p3_trade_count_lag1.

Lags for each pair for the 2 and 3 seconds:

P1_trade_count_lag2, p2_trade_count_lag2, p3_trade_count_lag2, p1_trade_count_lag3, p2_trade_count_lag3, p3_trade_count_lag3.

Additionally, boolean variables indicate whether the same cycle was profitable in the immediate past:

is_profitable_prev1_sec

These features represent **short-term liquidity and activity** of the arbitrage cycle. Higher trade counts imply deeper liquidity and a greater chance that similar rate configurations persist across seconds. This assumption is further tested in chapter 5: results part. Additionally, feature values are standardized using z-score normalization before fitting the model.

Model variables and coefficients are present in the "Descriptive Analysis" part.
This approach allows us to move from the regression problem that is already discussed in "Lewellen (2015)" to a Classification model.
Classification models answer a different question when compared to the regression models.
Logistic regression is generally considered as a standard approach to hypothesis testing in statistics within classification model problem space. And Logistic regressions are still often used as a base model.

Classification models can be evaluated by a set of metrics defining their efficiency:

Chapter 4. Data

4.1 Data sources:

To be able to conduct the analysis of arbitrage opportunities on 1 second intervals the 2 things are needed:

1) transactions conducted
2) transactions recorded (data)

As mentioned earlier in the industry review chapter: binance is the largest CEX in by may 2025, and it was chosen for the analysis. Binance provides an interface to download the data in 1second intervals.

Data operations by steps:

The process of working with data can be categorized into such steps:

1)Obtaining the data from API
2) Cleaning and Transforming the data
3) Feature engineering and analysis

4.2 Data Collection Process:

All the next steps in data where obtained from this essential tables from binance:
https://data.binance.vision/index.html?prefix=data/spot/daily/klines/
An example of the data saved is:


```
[
  {
    "open_time": 1717201392000,
    "open": 0.00789700,
    "high": 0.00789700,
    "low": 0.00789700,
```

"close": 0.00789700,
"volume": 7.50000000,
"close_time": 1717201392999,
"quote_volume": 0.05922750,
"trade_count": 1,
"taker_base_volume": 0.00000000,
"taker_quote_volume": 0.00000000,
"symbol_enum": "ETHFIBNB"
}
]

Each kline includes:

- open – first trade price of the second
- high – maximum executed trade price
- low – minimum executed trade price
- close – last trade price
- volume – total base asset volume
- quote_volume – total quote asset volume
- trade_count – number of executed trades
- taker_base_volume / taker_quote_volume – volume of taker-initiated trades

Ultimately, there is 330 gb file (41.3 bln rows). And a smaller version, 33gb file with 2 bln rows ror arbitrage research purposes. The second file covers a period of 01.05.2025 - 31.05.2025 (31 days), (n=12 993 446) while the first one is designed for 1 year.

Data was obtained from through the binance data API using a custom python script that is available in attached github repository in the main.py file. The script uses a klines endpoint of "https://data-api.binance.vision/api/v3/klines" and has 2 key steps:

1) connecting to a local database on duckdb to allow data transfer

2)obtaining all available symbols ("get_all_available_symbols" method)

3)cycle that iterated through each symbols.

-this part of the code ensures that data for each available day is parsed into a json response and then stored in a separate file in a specified folder a "pkl" format. Storing in a different files increases the number of files needed for analysis but reduces the average size of each file, which allows the file to fit into RAM memory of computer/server used for analysis.

After the data is loaded on the a computer by python, it is then becomes available in a database for analysis and further transformations such as "feature engineering" from the previous chapter or even analysis from "Chapter 5: Results". Another enabler for that is the fact that files are initially stored in multiple files. This separation of files again boost performance in a database stage.

After the files are loaded in a database, they are once again.

Cleaning and transforming the data:

After obtaining the data it is processed and transformed inside DuckDB.

4.3 Cleaning and Preprocessing

Base vs quote confusion:

Binance encodes pairs as <BASE><QUOTE> (e.g., ETHFI**BNB**), in the algorithm part the base and quote asset were "flipped" to obtain price not just for selling but for buying the asset as well.

Constructing Directed Exchange Edges

src => dst (base => quote) uses the high price (worst taker-side buy).

dst => src (quote => base) uses the inverted value(1/ "rate") of the high price.

Example of flippage:

open_time	src	dst	rate	side	symbol	trade_date	trade_count
1746202000000000	ALT	TRY	1.21	ask	ALTTRY	2025-05-02	0
1746202001000000	ALT	TRY	1.21	ask	ALTTRY	2025-05-02	0
1746202000000000	TRY	ALT	0.826446281	bid	ALTTRY	2025-05-02	0
1746202001000000	TRY	ALT	0.826446281	bid	ALTTRY	2025-05-02	0

Table 2, Raw data example

To arrive at a table demonstrated earlier in the "model" 2 section, the data needs to be transformed. Transformations, and the logic of model/algorithm 1 are executed inside DuckDB and by utilizing DuckDB SQL syntax. All transformations are available in duckdb_code.sql file, and enum_conversion_code.sql file responsible for converting text names to enum* variables.

*enum stands for enumerated type, a datatype that allows to store the text more efficiently.

Feature engineering and analysis:

Most features are created inside DuckDB as well. As mentioned in the methodology, models use a transformed dataset. Also they use lag of variables and an engineered variable "is_profitable_prev1_sec". Such variables are calculated inside DuckDB by utilizing window functions capabilities.

After all transformations in DuckDB, the resulting dataset is exported into Python for: statistical analysis, model training, cross-validation, and other analysis.

Chapter 5. Results

5.1. Descriptive Analysis

While doing the literature review, I found out that the DeFi data was more researched by the similar method to what I was interested in. To be more specific, Vadym Pacholchuk, already applied a "Bellman- Ford cycle" to DeFI data. However one of the problems with DeFi research was a fact that there were no 1 second intervals available, only 1 minute intervals.

My initial hypothesis stated that there would be no daily arbitrage opportunities. Indeed, the output of the algorithm showed the same.

As for the 1 second intervals - the situation is very different. When we apply the algorithm to the data, we get completely different results:

Table 3: Sample arbitrage paths

open_time	path_symbols	rate_ab	rate_bc	rate_ca	rates multiplied	P&L
2025-05-01T 07:01:25Z	{USDTTRY => CETUSUSDT => CETUSTRY}	0.026	5.043	8.300	1.09	8.74%
2025-05-01T 07:01:25Z	{CETUSUSDT=> CETUSTRY=> USDTTRY}	5.043	8.300	0.026	1.09	8.74%
2025-05-01T 07:01:25Z	{USDCTRY=> CETUSUSDC=> CETUSTRY}	0.026	5.030	8.300	1.08	8.47%
2025-05-01T	{MOVEUSDT=>	4.840	0.226	1.000	1.09	9.35%

16:56:04Z	MOVEUSDC=> USDCUSDT}					
2025-05-01T 16:56:42Z	{MOVEUSDT=> MOVEUSDC=> USDCUSDT}	4.817	0.227	1.000	1.09	9.31%
2025-05-02T 06:15:55Z	{ONDOUSDT=> ONDOTRY=> USDTTRY}	1.103	38.900	0.026	1.11	11.38%

The table above demonstrates the trades that my algorithm was able to find on the 1s intervals. We are not yet talking about how easy it is to execute on this trades, but we already see that it is technically possible to find this pairs, which was not possible on daily intervals.

Now, when we know that it's technically possible to find such pairs, we introduce a few more variables to reduce the "noise" opportunities that can't really be executed or will not lead us to a profit.

There are 2 main problems to consider that may lead to an opportunity to become not valid:

1)**Slippage** happens when the trade goes through at a different price than ones expected. This often occurs on decentralized exchanges, where all trades are market orders that can change the price. Slippage can reduce or even erase profits, especially for arbitrage traders.

Source: https://www.investopedia.com/terms/s/slippage.asp

2)**High fees**(including fees for low-volume on account level)

This part doesn't need a long explanation. If 3 trades generated a profit of 0.2%, while we payed a 1% of fees, then we lost 0.8% of our starting amount.

To account for this problems I introduced a few elements in my paper.

To ensure that the opportunities survive trading fees, I looked at the binance fees:

Table 4: Binance spot fee tiers

Level	30-Day Trade Volume (USD*)	and/or	BNB Balance	Maker / Taker
Regular User	< 1,000,000 USD	or	≥ 0 BNB	0.1000% / 0.1000%
VIP 1	≥ 1,000,000 USD	and	≥ 25 BNB	0.0900% / 0.1000%
VIP 2	≥ 5,000,000 USD	and	≥ 100 BNB	0.0800% / 0.1000%
VIP 3	≥ 20,000,000 USD	and	≥ 250 BNB	0.0400% / 0.0600%
VIP 4	≥ 75,000,000 USD	and	≥ 500 BNB	0.0400% / 0.0520%
VIP 5	≥ 150,000,000 USD	and	≥ 1,000 BNB	0.0250% / 0.0310%
VIP 6	≥ 400,000,000 USD	and	≥ 1,750 BNB	0.0200% / 0.0290%
VIP 7	≥ 800,000,000 USD	and	≥ 3,000 BNB	0.0190% / 0.0280%
VIP 8	≥ 2,000,000,000 USD	and	≥ 4,500 BNB	0.0160% / 0.0250%
VIP 9	≥ 4,000,000,000 USD	and	≥ 5,500 BNB	0.0110% / 0.0230%

The table indicates that trading fees start from 0.1% and can be reduced to so far as to 0.01% fee / trade.

In my algorithm I made a separate parameter for that, and used 0.3% as a base value. On top of the fees, lets not forget that we need to make profit, so we would look for trades with higher levels than base.

Table 5: Pairs by P&L bucket

pairs_cnt	pnl_group
177981	0-3%
580	3-5%
143	6-9%
35	9-12%
27	15%+
8	12-15%

*Table assumes "taker-taker-taker" leg type, and that all trades would execute immediately, under prices for the particular second.

When we group all possible opportunity pairs by 3% buckets (upper bound is >=, includes the 3%) then we can see that the bigger the profit, the less pairs there is. And after that filter for at least 0.3% of profit, our avg pnl = 0.45%. Median: 0.37%. So once again - the bigger the gain, less opportunities we have.

But more importantly - we need to be sure that we would make a profit. (Slippage problem).

To avoid the slippage problem I used the following tools:

1) trade has to happen in the same moment, same second.

-If we have to wait for the specific pair to cash out - we can't be sure it would happen.

2) trade counts by pair.

- to be sure that the trade can be executed, I filtered the data so that every pair in the algorithm had atleast 1 trade for the each currency used.

3) a proxy metric: "likelihood of the successful execution".

- having the data for the next trades we can successfully check whether the same opportunity for arbitrage remained possible in the next second or seconds.

Table 6: Profitability-persistence example (t+1 / t+2)

open_time	**path_symbols**	**rates multipled**	**P&L**	**is profitable next second**	**is profitable in next2 seconds**
2025-05-01T07:01:25Z	{USDTTRY,CETUSUSDT,CETUSTRY}	1.09	8.74%	FALSE	TRUE
2025-05-01T07:01:25Z	{CETUSUSDT,CETUSTRY,USDTTRY}	1.09	8.74%	FALSE	TRUE
2025-05-01T07:01:25Z	{USDCTRY,CETUSUSDC,CETUSTRY}	1.08	8.47%	FALSE	TRUE
2025-05-01T16:56:04Z	{MOVEUSDT,MOVEUSDC,USDCUSDT}	1.09	9.35%	FALSE	TRUE
2025-05-01T16:56:42Z	{MOVEUSDT,MOVEUSDC,USDCUSDT}	1.09	9.31%	**TRUE**	FALSE
2025-05-02T06:15:55Z	{ONDOUSDT,ONDOTRY,USDTTRY}	1.11	11.38%	TRUE	TRUE
2025-05-02T06:15:55Z	{USDTTRY,ONDOUSDT,ONDOTRY}	1.11	11.38%	TRUE	TRUE

2025-05-02T06:15:55Z	{USDCTRY,ONDOUSDC,ONDOTRY}	1.11	11.02%	TRUE	FALSE
2025-05-02T06:16:11Z	{ONDOUSDT,ONDOTRY,USDTTRY}	1.05	5.05%	FALSE	FALSE
2025-05-02T06:16:11Z	{USDTTRY,ONDOUSDT,ONDOTRY}	1.05	5.05%	FALSE	FALSE

From here we can see that not every pair could be utilized after we spotted it.

While one may argue that it's not a showing a complete picture, it is still a great proxy to us. Once we agreed to say that opportunity is possible to execute if it remained for the next second, then we can model the factors that influence it.

To reiterate: if we have a binary classifier "is profitable next second", then we can model the expected probability % of the outcome we want. (the likelihood of opportunity to work) .

opportunities, count	**opportunities, share%**	**is profitable next second**
169,885	95.03%	FALSE
8,889	4.97%	TRUE

Table 7: profitability summary table 1

This table gives us a much different picture from the tables with opportunities. It shows that only 5% of opportunities can be executed next second. So once we identified something, it may be gone very soon. And if we submitted an order, a lot can change fast,. So there is no guarantee that in that split seconds the opportunity may be gone too.

The 2 main reasons for the opportunity to be gone are:

1) the prices changed

2) not enough liquidity to guarantee a trade.

While we can't do anything about the price change, the liquidity issue can be solved by arbitrage performers in a few ways. For example:

1)connecting binance trading account in a system of other(DeFi +CeFI) exchanges

2)replacing a 1 pair in the cycle for combinations of a 2-3 pairs.

So the question of volatility problem vs price change problem remains open.

5.2. Predicting opportunities that would "remain":

Given that we have a new proxy, it is finally possible to try to predict if the opportunity would remain:

Table 7: Base Model output

```
Part 1: Simple Model
==================================================
Optimization terminated successfully.
         Current function value: 0.191743
         Iterations 7

 Statsmodels Summary:
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:               178774
Model:                          Logit   Df Residuals:                   178767
Method:                           MLE   Df Model:                            6
Date:                Wed, 03 Dec 2025   Pseudo R-squ.:                 0.03011
Time:                        11:36:32   Log-Likelihood:                -34279.
converged:                       True   LL-Null:                       -35343.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept              -3.0353      0.011   -267.539      0.000      -3.058      -3.013
p1_trade_count_lag1     0.0011      0.000      2.758      0.006       0.000       0.002
p2_trade_count_lag1     0.0031      0.000     14.041      0.000       0.003       0.004
p3_trade_count_lag1     0.0018      0.000     13.246      0.000       0.002       0.002
p1_trade_count_lag2     0.0044      0.000     10.608      0.000       0.004       0.005
p2_trade_count_lag2     0.0031      0.000     12.933      0.000       0.003       0.004
p3_trade_count_lag2     0.0024      0.000     15.518      0.000       0.002       0.003
=======================================================================================
ROC AUC: 0.6556
```

While the model has some significant variables and is better than random(AUC>0.5), it is still not good at predicting the fact whether the opportunity mat stays.

Table 8: profitability summary table 2

	Predicted Profitable	**Predicted Not Profitable**
Actual Profitable	158	8731
Actual Not Profitable	195	169690

While I identified that lags of number can be helpful in terms of predicting whether opportunity would be available to exploit in the next second, this area of the research requires more effort and hypothesis validation.

In summary: the model provided better then random performance (AUC>0.5), but the signal are weak. This interpretation is consistent with other research [Lewellen, Jonathan W]

Chapter 6. Conclusions and Recommendations

6.1 Restatement of Research Objective

This research's purpose is to determine if arbitrage opportunities exist with a large centralized exchange of cryptocurrencies (Binance). The research also evaluated if arbitrage opportunities exist long enough to take advantage of through taker-side execution; considering short-term persistence, researchers aim to use market-microstructure variables (i.e., recent trade activity and recent prices) to predict the short-term persistence of arbitrage. The study is intended to evaluate both the current state of triangular arbitrages and their persistence through a combination of triangular arbitrage detection and persistence classification models.

6.2 Summary of Key Findings

The empirical findings prove that opportunities for triangular arbitrage actually occur on Binance Spot, sometimes yielding gross percentage profits above a multiple of percent. Moreover, these opportunities tend to be transient, as about 95% of detected cycles become unprofitable within the next second. It can be hypothesized that, although inefficiencies emerge momentarily, they get corrected almost instantly, making them less amenable to exploitation. These findings, thus, tend to provide evidence of a highly unstable price structure, possibly due to natural volatility, or made up for by corrective forces.

The classification model formalizes this variability: its predictive ability is poor, even for recall, and its AUC is merely moderate, suggesting that persistence is mostly a random event from second-level aggregated data.

Additionally, liquidity constraints make up a significant aspect. Even if a profit-making cycle is available, a lack of trading volume can make it difficult to execute to its full capacity. The classification model agrees that this is unstable, since its predictive accuracy is poor, according to low recall and moderate figures for accuracy of the ROC curve, suggesting that this phenomenon is difficult to predict using market data with up to second lag. If arbitrage opportunities can be identified, they tend to be temporary.

6.3 Implications for Market Efficiency

The findings have several implications for discussions of market efficiency, but they do not allow for a definitive conclusion in favor of or against the Efficient Markets Hypothesis (EMH). On one hand, the presence of observable triangular arbitrage opportunities suggests that Binance Spot does not maintain perfect price alignment across all trading pairs at all times. This alone indicates deviations from the strongest form of market efficiency.

However, the extremely short lifespan of these opportunities: most disappear within one second suggests rapid adjustment dynamics. This behavior is consistent with, but not uniquely explained by the limited-arbitrage framework of Shleifer & Vishny (1997), in which mispricings may arise but are difficult or costly to exploit. The data do not completely reveal why the opportunities vanish. The short-lived nature of the cycles could reflect:

1) natural price volatility at the one-second granularity,
2) insufficient liquidity to execute trades at published prices,
3) active arbitrageurs eliminating the imbalance
4) order-book microstructure noise
5) latency and slippage effects not observable in the dataset.

Therefore, while the results show that arbitrage is detectable yet difficult to exploit, this cannot be taken as strong evidence of EMH.

6.4 Practical Implications for Traders and Market Participants

The results of this study indicate that triangular arbitrage opportunities on Binance Spot are highly sensitive to short-term price movements. Since most opportunities disappear within one second, traders attempting to act on them must operate with fast data access and the ability to submit trades quickly. While this does not require extreme infrastructure, it does imply that delays of several seconds would make most opportunities untradable.

Liquidity considerations also play a central role. Even when a triangular cycle appears profitable based on the taker-side high price, actual execution may differ because the available traded volume at that moment may be insufficient. When there is a lack of liquidity, the actual price traded may differ from the quoted price, thus diminishing or even forfeiting the apparent profit. Essentially, a lack of sufficient volume for a side of a cycle can cause a failure to execute a trade.

Arbitrage opportunity disappearance can be explained through two basic processes:

1. Price movements within the second, which can quickly eliminate the price imbalance.
2. Changes in available traded volume, which may reduce the feasibility of executing the full cycle.

Although the data sets analyzed within this study lack depth of book information or sub-second trade data, information on a 1-second interval is still relevant. Much of the existing academic work relies on data of this nature, and from this investigation, patterns relevant to the rate of change of arbitrage opportunities can be observed. However, to pinpoint the difference between volatility and liquidity constraints, precision of a different nature is required.

There are possible methods to reduce these shortcomings. For example, traders can increase their capacity to finish a cycle of transactions by aggregating the available liquidity from a number of platforms, such as a centralized or a decentralized exchange. In certain instances, trading directly between two assets can be replaced by trading through a chain of assets to tap a larger pool of available liquidity.

Overall, these findings indicate that, although triangular arbitrage exists on CEX markets, to take advantage of such opportunities, one must be responsive as well as aware of the liquidity dynamics. The transient nature of these opportunities is a good indication of the significance of quality of execution and that a slippage of even a short amount of time or lack of enough volume can make a difference.

Recommendations for further research:

1) Including multiple CEXs such as Coinbase, Kraken, and OKX can be beneficial for analyzing cross-market arbitrage opportunities, synchronized price processes, and fragmented liquidity.

It would be possible to determine if trading profits move from one platform to another, if common vs. unique inefficiencies of CEXs can be identified, or if dynamics of triangular arbitrage depend on the fact that transactions cross trading platforms.

2) Integration of CEX and DEX Markets

A decentralized exchange (DEX) is driven by different microstructural processes, uses automated market makers (AMMs), and demonstrates different slippage and fee patterns. It is possible to conduct a unified study of arbitrage opportunities for different market structures by combining CEX and DEX data.

Moreover, such integration can also provide information about the interaction of liquidity between the centralized and decentralized systems, and if arbitrageurs help to stabilize these two.

3) Survival Analysis of Arbitrage Opportunities

In this study, the persistence is approximated by binary 'next-second' indicators. However, future researchers can utilize survival techniques to model persistence through the duration of 'arbitrage' opportunities.

6.5 Limitations

Although the findings of this research provide a wealth of knowledge regarding the presence and short-term characteristics of triangular arb opportunities found on the Binance institutional trading platform, there are many restrictions involved with making sense of these outcomes.

Firstly, the model is based on aggregated trade data at a 1-second interval, as opposed to a complete order book. Although this is a higher level of detail than that of most previous models, it is nevertheless a partial, aggregated description of the truth. It is not a precise description of the order book at the time of a trading opportunity detection. Moreover, this being the case, the assessment is not capable of determining the depth of available trades at each price, nor if the whole cycle of the upcoming opportunity can be performed at such a rate.

Secondly, the high price on the taker side is a conservative benchmark for execution. Although this makes the likelihood of exaggerating the profitability lower, this metric fails to incorporate the actual second-by-second price movement and the sequence of trade execution. It is possible that the actual execution price differs according to the timing of the trade.

Thirdly, the data set fails to provide cross exchange information. All the above findings relate only to the opportunity for a single exchange, Binance. Any opportunity that depends on multiple exchanges or is a result of synchronized pricing across multiple exchanges is not relevant to this research.

Fourth, the classification model relies on second-level features that depend on recent trade numbers and price movements. Without the ability to process sub-second data or order book level statistics, neither can the model include additional possible relevant predictors such as instantaneous bid-ask spreads, passive liquidity, market impact assessments, or positioning. It is therefore difficult to characterize the microstructure drivers of persistent arbitrage opportunities.

Finally, this analysis is not meant to replicate other real-world concerns, such as handling partial fills, delays from a network, or API errors. These concerns can be relevant to actual trading, but they are not possible to quantify from this data.

Overall, these shortcomings do not negate the significance of empirical results but rather emphasize the scope within which the empirical conclusions have to be understood. These shortcomings also provide scope for future research, possibly through the consideration of additional sources of information, multiple exchange datasets, or enhanced models of executions.

REFERENCES

Cointelegraph Article, Binance loses ground to rivals, DEX's on the rise: Report https://cointelegraph.com/news/binance-crypto-com-lose-ground-dex-rise

Shleifer and Vishny, 1997: The Limits of Arbitrage. The Journal of Finance Vol. 52, No. 1 (Mar., 1997) https://www.jstor.org/stable/2329555

Lewellen, Jonathan W., The Cross Section of Expected Stock Returns (August 22, 2014). Forthcoming in Critical Finance Review, Tuck School of Business Working Paper No. 2511246, Available at SSRN https://ssrn.com/abstract=2511246 or http://dx.doi.org/10.2139/ssrn.2511246

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. American Economic Review, 70(3), 393–408. Available at SSRN: https://ssrn.com/abstract=228054

Vadym Pakholchuk (2024), Arbitrage opportunities on decentralized exchanges: application of graph neural networks, master thesis paper at Kyiv School of Economics, https://kse.ua/wp-content/uploads/2025/09/Vadym_Pakholchuk_ARBITRAGE-OPPORTUNITIES.pdf

Oak research, CEX vs DEX: Report from 2024 and forecasts for 2025, report available at https://www.oakinvest.fr/oakresearch-en

Coingecko, Q1 2025 industry report available at: https://assets.coingecko.com/reports/2025/CoinGecko-2025-Q1-Crypto-Industry-Report.pdf

Appendix

Appendix A: Charts

Figure 1: Market Capitalization

https://www.tradingview.com/symbols/TOTAL/?timeframe=12M, screenshot 03.09.25

Figure 2: Market volumes

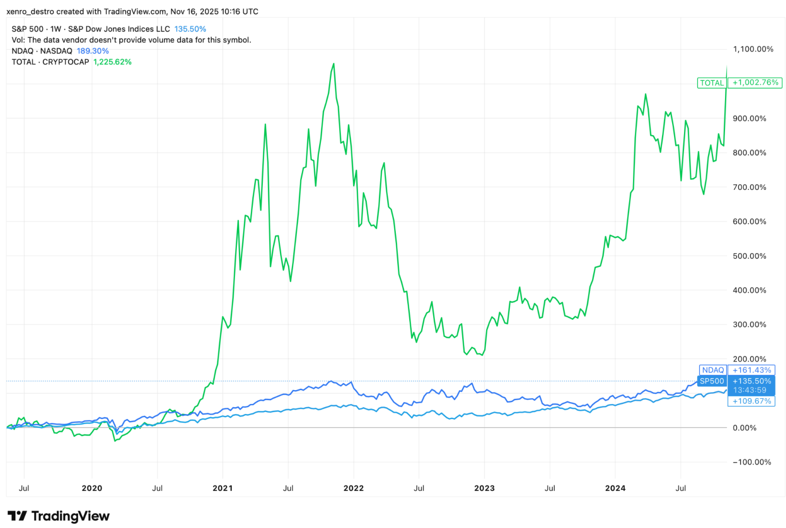
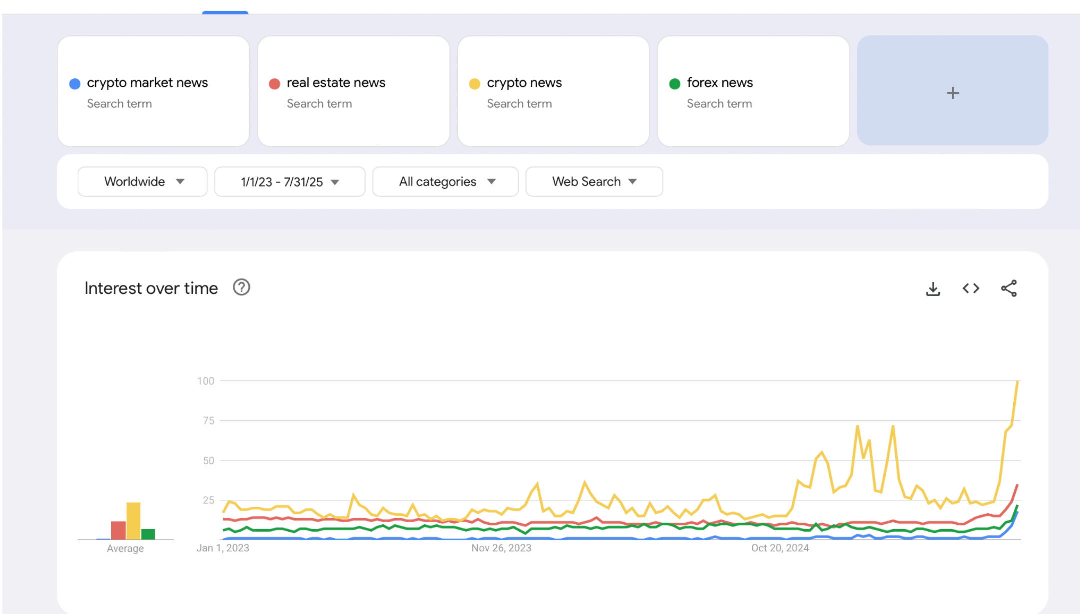


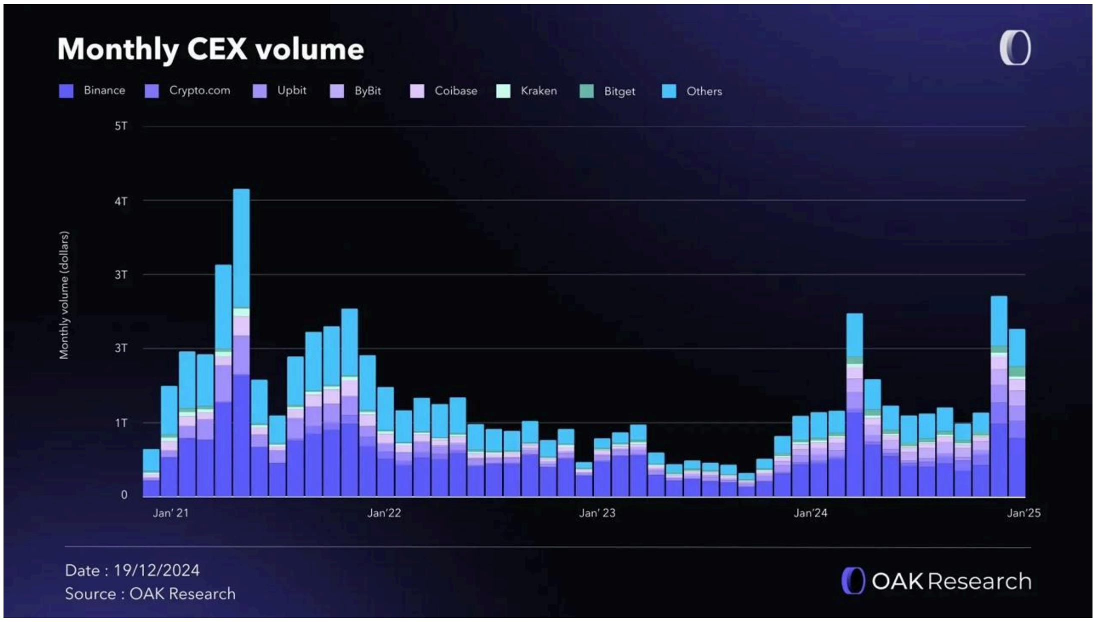
Figure 3, google trends



Url:

https://trends.google.com/trends/explore?date=2023-01-01%202025-07-31&q=crypto%20market%20news,real%20estate%20news,crypto%20news,forex%20news

Figure 4: Monthly CEX volume



Source, Source: Oak research, CEX vs DEX: Report from 2024 and forecasts for 2025, full version

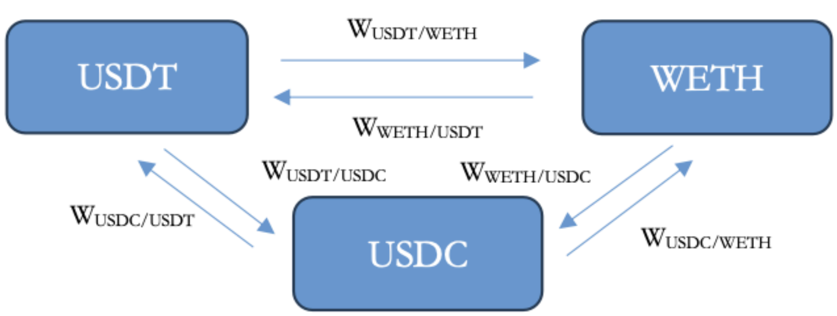
Url: https://www.oakinvest.fr/oakresearch-en

Figure 5: DEX to CEX Futures trade volume

Source: Oak research, CEX vs DEX: [Report](#) from 2024 and forecasts for 2025, [full version](#)

Url: https://www.oakinvest.fr/oakresearch-en

Figure 6: triangular arbitrage example



Appendix B: Tables

Table 1,

Data: https://coinmarketcap.com/ , 25.06.25

Rank	Name	Symbol	Market Cap
1	Bitcoin	BTC	$2.13T
2	Ethereum	ETH	$291.15B
3	Tether	USDT	$156.76B
4	XRP	XRP	$128.59B
5	BNB	BNB	$90.74B
6	Solana	SOL	$76.24B
7	USD Coin	USDC	$61.81B
8	TRON	TRX	$25.77B
9	Dogecoin	DOGE	$24.74B
10	Cardano	ADA	$20.08B

Table 2: Raw data example

open_time	src	dst	rate	side	symbol	trade_date	trade_count
1746202000000000	ALT	TRY	1.21	ask	ALTTRY	2025-05-02	0

1746202001000000	ALT	TRY	1.21	ask	ALTTRY	2025-05-02	0
1746202000000000	TRY	ALT	0.826446281	bid	ALTTRY	2025-05-02	0

Source: example from custom dataset based on data from binance.

Table 3: Sample arbitrage paths

open_time	path_symbols	rate_ab	rate_bc	rate_ca	rates multiplied	P&L
2025-05-01T07:01:25Z	{USDTTRY => CETUSUSDT => CETUSTRY}	0.026	5.043	8.300	1.09	8.74%
2025-05-01T07:01:25Z	{CETUSUSDT=> CETUSTRY=> USDTTRY}	5.043	8.300	0.026	1.09	8.74%
2025-05-01T07:01:25Z	{USDCTRY=> CETUSUSDC=> CETUSTRY}	0.026	5.030	8.300	1.08	8.47%
2025-05-01T16:56:04Z	{MOVEUSDT=> MOVEUSDC=> USDCUSDT}	4.840	0.226	1.000	1.09	9.35%
2025-05-01T16:56:42Z	{MOVEUSDT=> MOVEUSDC=> USDCUSDT}	4.817	0.227	1.000	1.09	9.31%

2025-05-02T 06:15:55Z	{ONDOUSDT=> ONDOTRY=> USDTTRY}	1.103	38.900	0.026	1.11	11.38%

Table 4: Binance spot fee tiers

Level	30-Day Trade Volume (USD*)	and/or	BNB Balance	Maker / Taker
Regular User	< 1,000,000 USD	or	≥ 0 BNB	0.1000% / 0.1000%
VIP 1	≥ 1,000,000 USD	and	≥ 25 BNB	0.0900% / 0.1000%
VIP 2	≥ 5,000,000 USD	and	≥ 100 BNB	0.0800% / 0.1000%
VIP 3	≥ 20,000,000 USD	and	≥ 250 BNB	0.0400% / 0.0600%
VIP 4	≥ 75,000,000 USD	and	≥ 500 BNB	0.0400% / 0.0520%
VIP 5	≥ 150,000,000 USD	and	≥ 1,000 BNB	0.0250% / 0.0310%
VIP 6	≥ 400,000,000 USD	and	≥ 1,750 BNB	0.0200% / 0.0290%

Table 5: Pairs by P&L bucket

pairs_cnt	**pnl_group**
177981	0-3%
580	3-5%

143	6-9%
35	9-12%
27	15%+
8	12-15%

Table 6: Profitability-persistence example (t+1 / t+2)

open_time	path_symbols	rates multipled	P&L	is profitable next second	is profitable in next2 seconds
2025-05-01T07:01:25Z	{USDTTRY,CETUSUSDT,CETUSTRY}	1.09	8.74%	FALSE	TRUE
2025-05-01T07:01:25Z	{CETUSUSDT,CETUSTRY,USDTTRY}	1.09	8.74%	FALSE	TRUE
2025-05-01T07:01:25Z	{USDCTRY,CETUSUSDC,CETUSTRY}	1.08	8.47%	FALSE	TRUE
2025-05-01T16:56:04Z	{MOVEUSDT,MOVEUSDC,USDCUSDT}	1.09	9.35%	FALSE	TRUE
2025-05-01T16:56:42Z	{MOVEUSDT,MOVEUSDC,USDCUSDT}	1.09	9.31%	**TRUE**	FALSE
2025-05-02T06:15:55Z	{ONDOUSDT,ONDOTRY,USDTTRY}	1.11	11.38%	TRUE	TRUE
2025-05-02T06:15:55Z	{USDTTRY,ONDOUSDT,ONDOTRY}	1.11	11.38%	TRUE	TRUE
2025-05-02T06:15:55Z	{USDCTRY,ONDOUSDC,ONDOTRY}	1.11	11.02%	TRUE	FALSE
2025-05-02T0	{ONDOUSDT,ONDOTRY,U	1.05	5.05%	FALSE	FALSE

6:16:11Z	SDTTRY}				
2025-05-02T06:16:11Z	{USDTTRY,ONDOUSDT,ONDOTRY}	1.05	5.05%	FALSE	FALSE

Table 7: Base Model output:

```
...  Part 1: Simple Model
     ==================================================
     Optimization terminated successfully.
              Current function value: 0.191743
              Iterations 7

      Statsmodels Summary:
                                Logit Regression Results
     ==============================================================================
     Dep. Variable:                 target   No. Observations:               178774
     Model:                          Logit   Df Residuals:                   178767
     Method:                           MLE   Df Model:                            6
     Date:                Wed, 03 Dec 2025   Pseudo R-squ.:                 0.03011
     Time:                        11:36:32   Log-Likelihood:                -34279.
     converged:                       True   LL-Null:                       -35343.
     Covariance Type:            nonrobust   LLR p-value:                     0.000
     =======================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
     ---------------------------------------------------------------------------------------
     Intercept              -3.0353      0.011   -267.539      0.000      -3.058      -3.013
     p1_trade_count_lag1     0.0011      0.000      2.758      0.006       0.000       0.002
     p2_trade_count_lag1     0.0031      0.000     14.041      0.000       0.003       0.004
     p3_trade_count_lag1     0.0018      0.000     13.246      0.000       0.002       0.002
     p1_trade_count_lag2     0.0044      0.000     10.608      0.000       0.004       0.005
     p2_trade_count_lag2     0.0031      0.000     12.933      0.000       0.003       0.004
     p3_trade_count_lag2     0.0024      0.000     15.518      0.000       0.002       0.003
     =======================================================================================
     ROC AUC: 0.6556
```

Table 8: profitability summary table 2

	Predicted Profitable	**Predicted Not Profitable**
Actual Profitable	158	8731
Actual Not	195	169690

Profitable		

Appendix C: Code

Github page with code for the project:

https://github.com/gybzd/binance_arbitrage_research/tree/main/main_folder