Analysis & Dashboard of PaySim Financial Transaction Data

Scott Goley

Goals

- A write-up of your methodology

- Key insights or trends from the data

Topics to cover

- Explain what steps you took to validate and cleanse the data.

- What did you do to explore trends or patterns in the data?

- How did you decide which insights to highlight from the data?

- What additional information or context might have been useful in forming your conclusions?

Project Materials

- Github Repo

- Streamlit Cloud Dashboard

Validation & Cleansing

- I made a transaction surrogate key to handle for duplication
  - `hash( datetime || Type || nameOrig || nameDest || amount )`
- I confirmed that dataset does not contain obvious duplicates
- I confirmed the time range of dataset (approximately covering a 31 day period truncated to hours)
- I made an adjustment from `step` to a datetime data type for easier graphing
- I made an adjustment on tx amount based on apparent direction (Transfer = Credit from Origin, Debit to Destination)
- I made new balances calculations on both the Origin & Destination accounts.
- I made boolean field to determine if the given vs the calculated Origin & Destination balances were the same.

Explore

- The dataset contains approximately 6.36M total transactions

- It contains about 6.35M sending accounts and 2.72M receiving accounts

- 8.21K were flagged via the `isFraud` boolean variable

- With $12B* in flagged transactions, this appoximates to 1% of the total 1.14T* in notional value and 0.1% of transaction count.
  - *absolute notional values*

Dashboard

Dashboard

Dashboard
Tx Profiling
Payment Network
Balances
Waterfall

Table Preview

Top 100

High Level Metrics

# Trx	$ Trx	# Orig	# Dest	# Flagged Trx	$ Flagged Trx
6.36M	1.14T	6.35M	2.72M	8.21K	$12.06B

Transaction Counts Over Time

Fraud Status
■ Normal
■ Fraud

Count of Tx

Jan 5 2025 · Jan 12 · Jan 19 · Jan 26

Transaction Date/Time
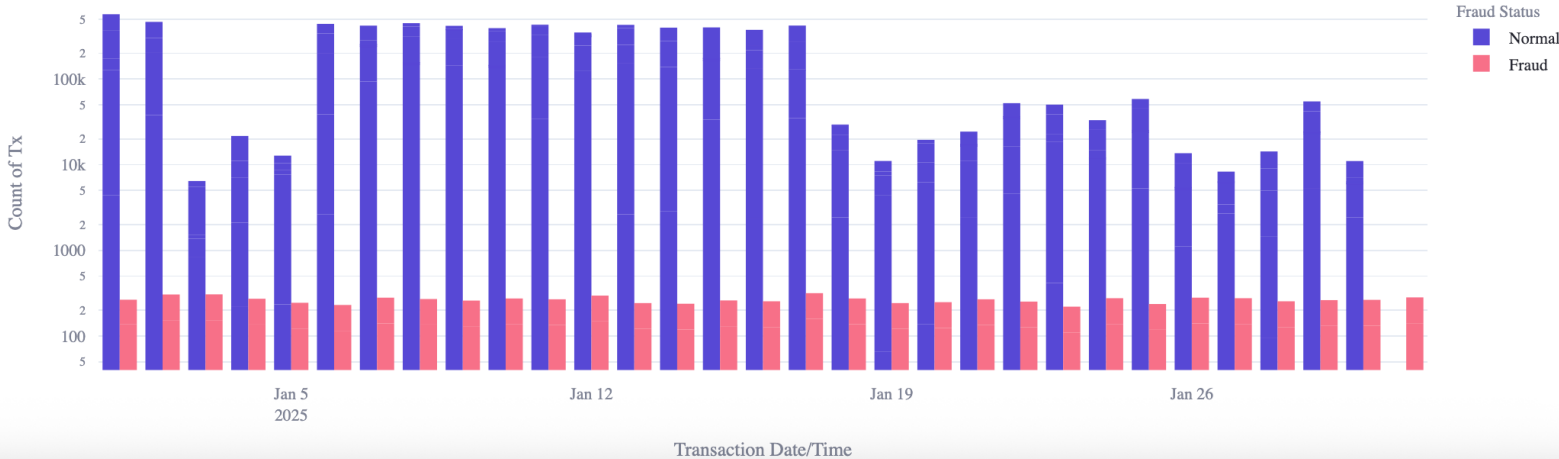
Explore - Dashboard

- The dashboard itself features most of these metrics as a headline for frequent viewing
- The primary featured chart is targeted to viewing proportion of fraudulent transactions over time vs non fraud.

Deploy ⋮

Dashboard

🔍 Tx Profiling

📶 Payment Network

⚖ Balances

🌊 Waterfall

Table Preview

20 Each Type ⌄
Top 100
Latest 100
20 Each Type
100 Fraud

Transaction Preview

	datetime	tx_sk	step	type	abs_amount	amount	amount_orig	amount_dest	nameOrig	old_bal
15	2025-01-06 14:00:00	9475374537790904600	134	CASH_IN	81851.26	-81851.26	81851.26	81851.26	C1580308479	
16	2025-01-13 18:00:00	13072596225520433450	306	CASH_IN	295749.8	-295749.8	295749.8	295749.8	C750549587	
17	2025-01-07 16:00:00	2325340325350179525	160	CASH_IN	122601.32	-122601.32	122601.32	122601.32	C1954687067	
18	2025-01-14 17:00:00	14683110611851008299	329	CASH_IN	596808.18	-596808.18	596808.18	596808.18	C861727853	
19	2025-01-08 09:00:00	11653449511108615657	177	CASH_IN	98720.87	-98720.87	98720.87	98720.87	C33822345	40
20	2025-01-01 18:00:00	7099673570642276674	18	CASH_OUT	156744.17	156744.17	156744.17	156744.17	C351656668	
21	2025-01-08 13:00:00	14081744815798476362	181	CASH_OUT	78950.86	78950.86	78950.86	78950.86	C686572989	
22	2025-01-09 12:00:00	5491555955472968267	204	CASH_OUT	82996.26	82996.26	82996.26	82996.26	C1966436401	
23	2025-01-17 13:00:00	16691914000372288433	397	CASH_OUT	487100.39	487100.39	487100.39	487100.39	C573888147	
24	2025-01-07 12:00:00	14491794174156991324	156	CASH_OUT	149359.09	149359.09	149359.09	149359.09	C414152429	
25	2025-01-29 11:00:00	8894064117424315380	683	CASH_OUT	392951.51	392951.51	392951.51	392951.51	C1050147742	

Explore - Preview

- The table version of the dataset is configured to make frequently accessed filters available for review & download:
    - i. Top 100 Value Tx
    - ii. Latest 100 Tx
    - iii. 20 Random Each Tx Type
    - iv. 100 Fraud
- Or potentially many other variations on request.
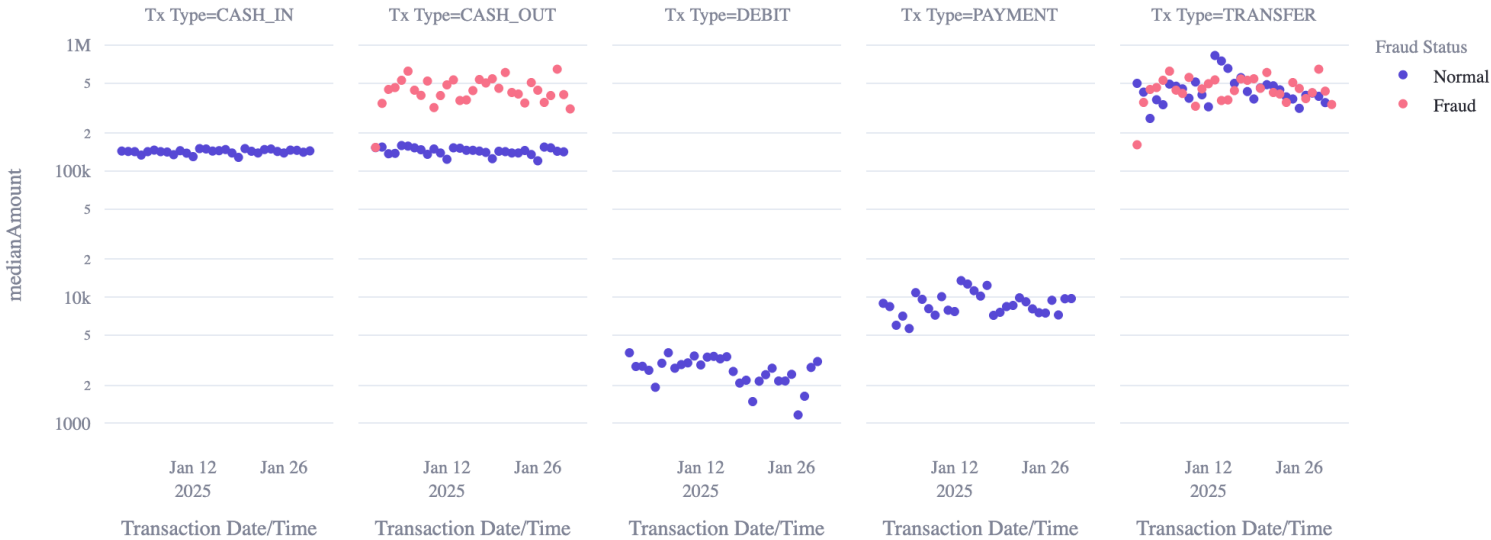
📊 Dashboard

🔍 **Tx Profiling**

📶 Payment Network

⚖️ Balances

🌊 Waterfall

🔍 Tx Profiling

Descriptive Analysis

Transaction Counts Over Time (median)

Explore - Profiling

- For high level profiling of the transactions and a few characteristics of the population vs fraudulent transactions:
    - i. All occur as part of the `CASH_OUT` or `TRANSFER` tx types.
    - ii. Of the `CASH_OUT` transactions, they tend to feature a significantly higher median and higher average notional value than non-fraudulent transactions.
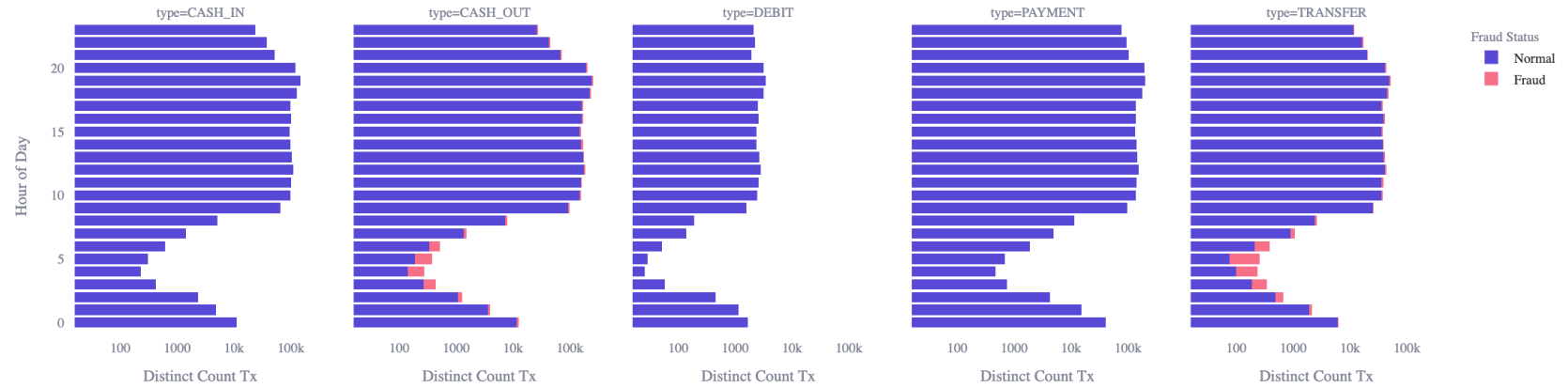
Dashboard
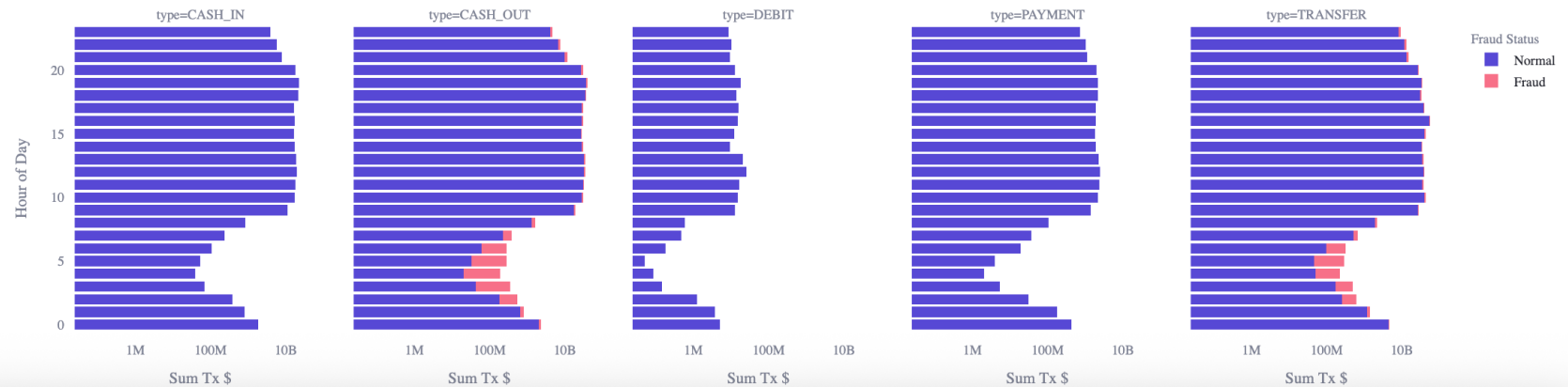
🔍 **Tx Profiling**

📶 Payment Network

⚖️ Balances

🌊 Waterfall

Transactions Counts by Hour of Day

type=CASH_IN	type=CASH_OUT	type=DEBIT	type=PAYMENT	type=TRANSFER

Fraud Status

■ Normal
■ Fraud

Hour of Day

Distinct Count Tx Distinct Count Tx Distinct Count Tx Distinct Count Tx Distinct Count Tx

Transactions $ Amounts by Hour of Day

type=CASH_IN	type=CASH_OUT	type=DEBIT	type=PAYMENT	type=TRANSFER

Fraud Status

■ Normal
■ Fraud

Hour of Day

Sum Tx $ Sum Tx $ Sum Tx $ Sum Tx $ Sum Tx $

Explore - Hours

- An additional feature of flagged transactions are:
    - i. A significant proportion occur at a low point in transaction volume.
        - a. Due to my datetime adjustment, this appears to be between the 3am – 7am hours of the day but requires further context to confirm.
    - ii. This is true of both `CASH_OUT` and `TRANSFER` types of flagged transactions.

Transaction Network Graph

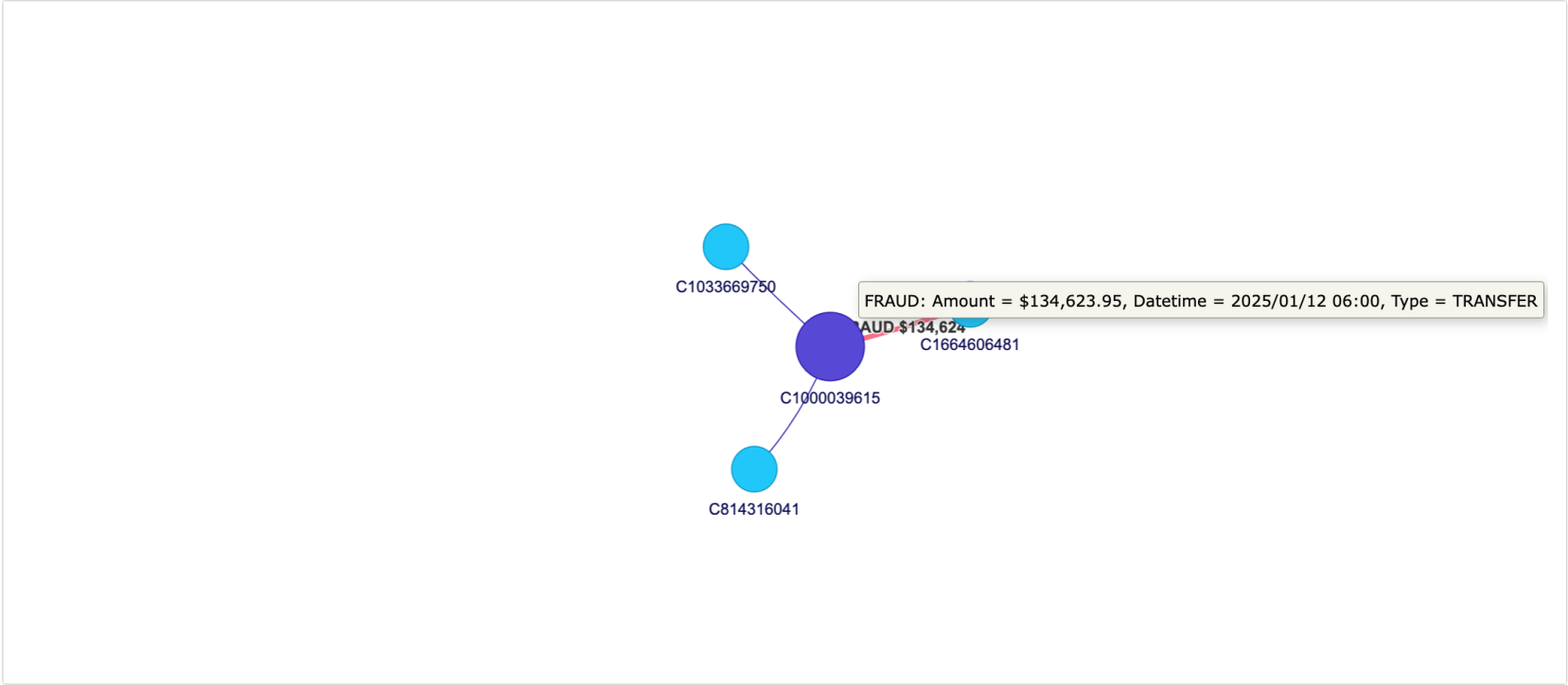FRAUD: Amount = $134,623.95, Datetime = 2025/01/12 06:00, Type = TRANSFER

Sidebar navigation:
- Dashboard
- Tx Profiling
- **Payment Network**
- Balances
- Waterfall

Fraud Receivers
C1000039615

Graph labels:
- C1033669750
- C1664606481
- C1000039615
- C814316041
- AUD $134,624

Transaction Details

	datetime	tx_sk	step	type	abs_amount	amount	amount_orig	amount_dest	nameOrig	old_balance_orig	new_balance_orig	post_balance_orig	trust_orig_balance
0	2025-01-12 06:00:00	6311493832270830241	270	TRANSFER	134623.95	134623.95	134623.95	134623.95	C1664606481	134623.95	0	269247.9	☐
1	2025-01-14 11:00:00	3104809758410863632	323	CASH_OUT	135185.79	135185.79	135185.79	135185.79	C814316041	10287	0	145472.79	☐
2	2025-01-17 10:00:00	9718973262544582494	394	CASH_IN	110991.07	-110991.07	110991.07	110991.07	C1033669750	11359	122350.07	122350.07	☑

Explore - Graph

- Finally – I produced a graph exploration view so that single actors which were either the sender or receiver of a flagged `isFraud` can be deeply inspected in the context of their other transactions & counterparties.

- One of the most important characteristics observed here is that nearly every transaction originates from a single location though some share a common destination.
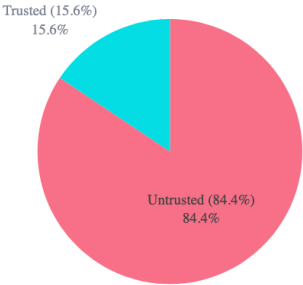
Dashboard

Tx Profiling

Payment Network

Balances

Waterfall
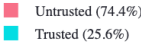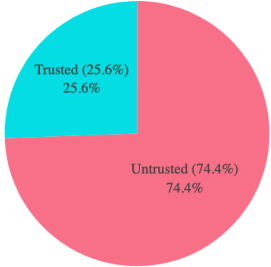
⚖️ **Balances**

Balance Trust Analysis

Origin Balance Trust

Origin Balance Trust Distribution

Trusted (15.6%)
15.6%

Untrusted (84.4%)
84.4%

■ Untrusted (84.4%)
■ Trusted (15.6%)

Destination Balance Trust

Destination Balance Trust Distribution

Trusted (25.6%)
25.6%

Untrusted (74.4%)
74.4%

■ Untrusted (74.4%)
■ Trusted (25.6%)
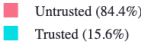
Fraudulent Transaction Details

	datetime	tx_sk	type	amount	abs_amount	nameOrig	post_balance_orig	amount_orig	nameDest	amount_dest	post_balance_dest
0	2025-01-01 04:00:00	17149662971944459024	TRANSFER	10000000	10000000	C7162498	22930418.44	10000000	C945327594	10000000	10000000
1	2025-01-01 04:00:00	16153842662465233597	CASH_OUT	10000000	10000000	C351297720	20000000	10000000	C766681183	10000000	10000000
2	2025-01-01 19:00:00	18117189727304724915	TRANSFER	10000000	10000000	C416779475	21861008.32	10000000	C380259496	10000000	10000000

Explore - Balances

- This perspective shows that very few of the balance calculations on accounts match between the original and calculated balances and could possible be an indicator the `isFraud` flag.

- If we assume that fraudulent transaction in the same exact amount are linked, the movement becomes obvious:
    i. Initiate a `TRANSFER` to an account of bad actor control
    ii. Initiate a `CASH_OUT` transaction for the full quantity captured.

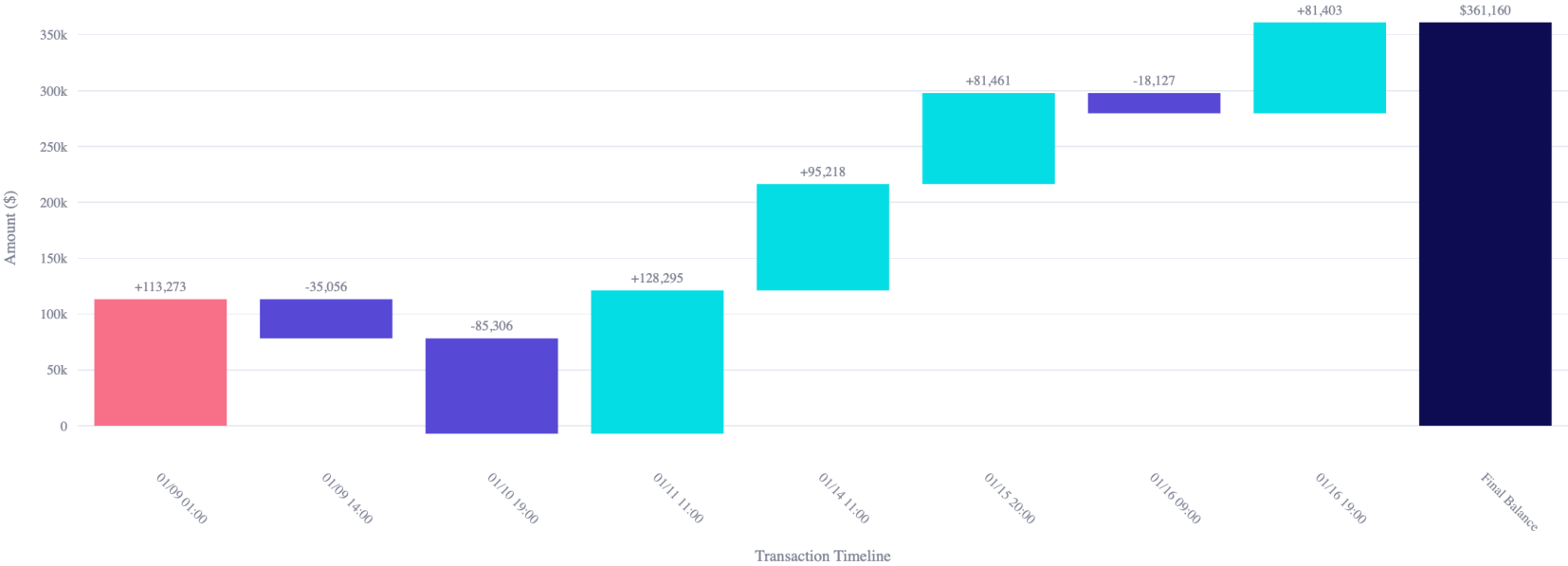## 💧 **Transaction Waterfall Chart**

Analysis for: C1005714733

Account Balance Waterfall Chart

Account Balance Changes for C1005714733

C1005714733 ▾

Total Inflow	Total Outflow	Net Change	Fraud Transactions
$499,649.78	$138,489.53	$361,160.25	1

Transaction Details

Explore - Waterfall

- This perspective shows on the receiving / destination account side how various transactions are received by what could possibly be a merchant or aggregator account.

- The appearance is:
  - All fraudulent transactions initiate as the only recent transaction for a given individual account
  - A `TRANSFER` operation takes place to a destination account
  - One or more `CASH_OUT` transactions take place from the destination. (Hallmark layering)

Additional Context

- Would be helpful to get account classification between personal individual vs business, operational omnibus etc.

- Would be helpful to compare vs additional time periods.

- Would be helpful to get more datetime granularity or even a transaction index to ensure exact order rather than inferring order since lots of the `TRANSFER` & `CASH_OUT` pairs happen within the same time step.