

Preliminary PaySim Data Analysis

*Overview of Preliminary Key Findings and Trends from
the PaySim dataset*



Outline

I. Objective

II. Dataset Description

III. Key Findings

- I. Transaction Types**
- II. Value of Fraud Cases**
- III. Timing**
- IV. Who was involved**
- V. Flagged Cases**

IV. Conclusion

Background – Objective

Objective:

To analyze the PaySim dataset on mobile money transactions and to provide an overview of key findings, interesting trends, and additional data desired



Dataset

01

PaySim Dataset – Description of variables and assumptions

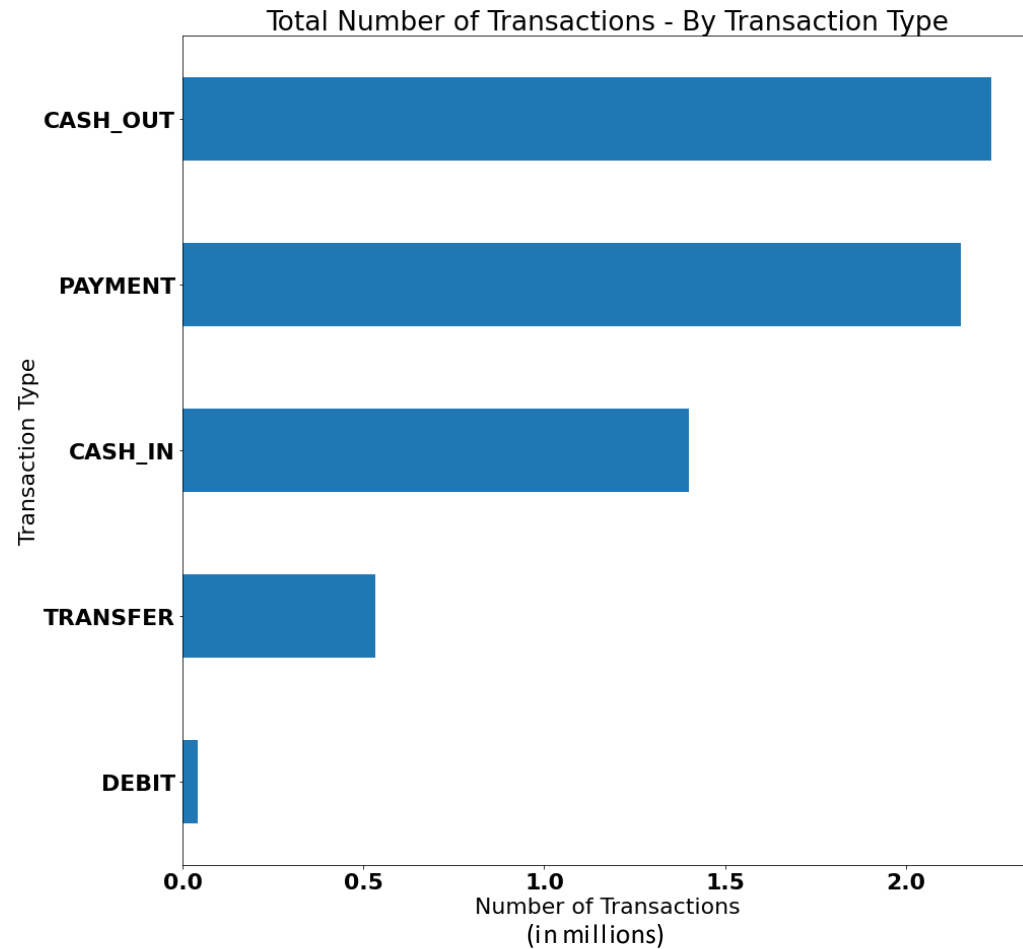
Column Name	Description
Step	Unit of time. 1 step is 1 hour of time
Type	Type of transaction
CASH_IN	Increase in balance of account by paying in cash to a merchant
CASH_OUT	Withdrawal of cash from a merchant; decrease in balance of account
DEBIT	Sending of money from mobile money service to a bank account
PAYMENT	Process of paying for good/service to merchants
TRANSFER	Process of sending money to another user of the service
Amount	Amount of the transaction in local currency
nameOrig	Customer who started the transaction
oldbalanceOrig	Initial balance before the transaction
newbalanceOrig	New balance after the transaction
nameDest	Customer who is the recipient of the transaction
oldbalanceDest	Initial balance before the transaction. This information is not available for customers that start with M (Merchants)
newbalanceDest	New balance after the transaction. This information is not available for customers that start with M (Merchants)
isFraud	Fraudulent transactions.
isFlaggedFraud	Transaction flagged as fraudulent
Variable(s) Created	
Date	Date of transaction (Assumption: January 2021 used)
Dayofweek	Day of week of transaction
Hourofday	Hour of day of transaction

- Dataset of mobile money transactions in an African country
- One month of transactions
- 6,362,620 total transactions

Assumptions and Variables Created

- Assumed transaction month – Jan. 2021
- Variables: Date, Dayofweek, and Hourofday created using “Step” variable

PaySim Dataset – Majority of transactions were either ‘cash out’ or ‘payment’

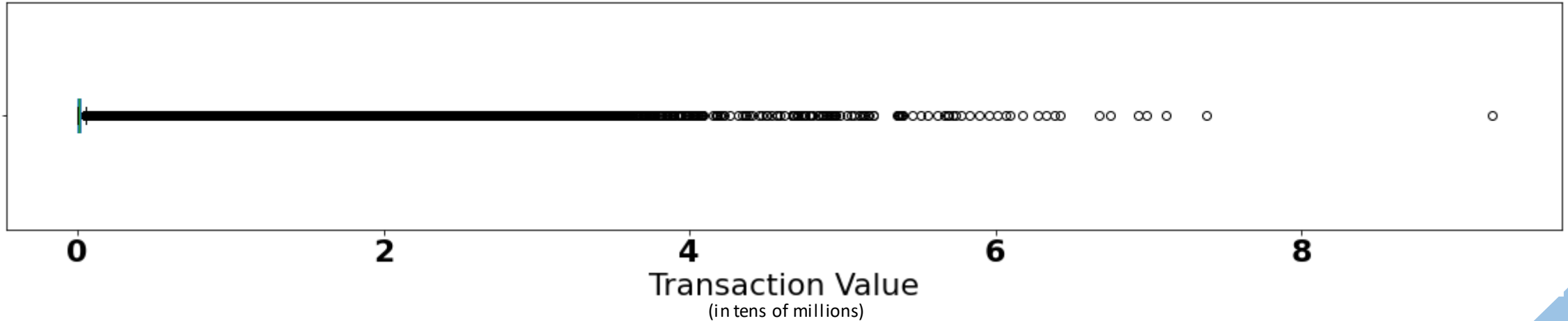


Proportion of total transactions

- **Cash out:** 35.2%
- **Payment:** 33.8%
- **Cash in:** 22.0%
- **Transfer:** 8.4%
- **Debit:** 0.7%

PaySim Dataset – Wide range of transaction amounts

Box Plot of Transaction Amounts



- **Wide range of transaction amounts** — min: 0; max: 92,445,520
- **Median:** 74,871
- **75th percentile:** 208,721

Dataset Discrepancies

1. Several transactions over 200,000 are not flagged as fraud.

Dataset description states that transactions greater than 200,000 should be flagged

2. Several cases where transaction amounts do not add up properly:

- Several “transfer” transactions where new balance destination account decreases *after* transaction occurs
- In some cases of transaction type “debit”, destination account decreases when it should increase

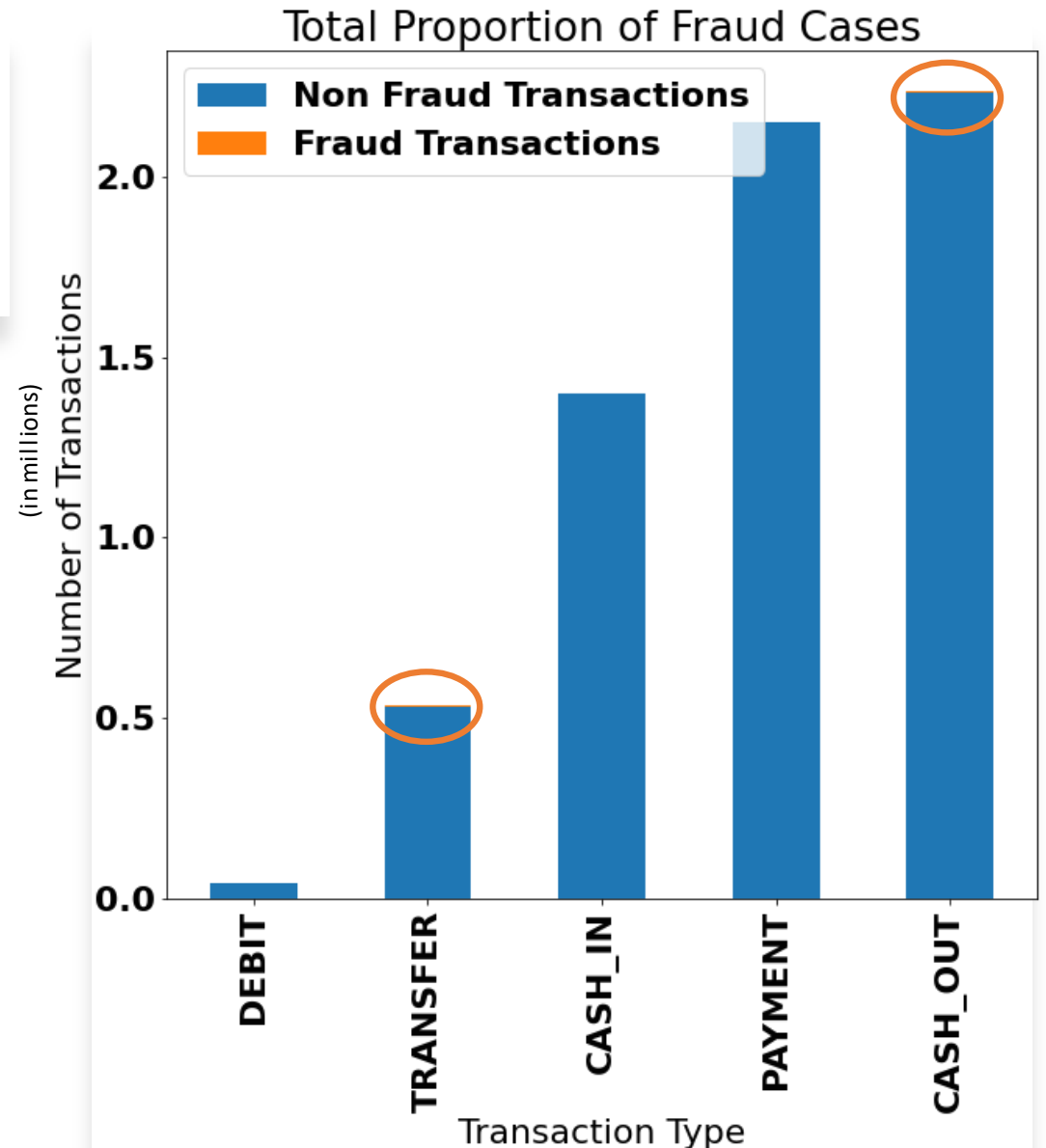
3. Sixteen cases where amount of transaction was 0; all were ‘cash out’ transactions.

**Due to dataset discrepancies, findings in this presentation should be considered preliminary*

Key Findings – Transaction Types 02

Fraudulent cases only occurred among two types of transactions – Transfer and Cashout

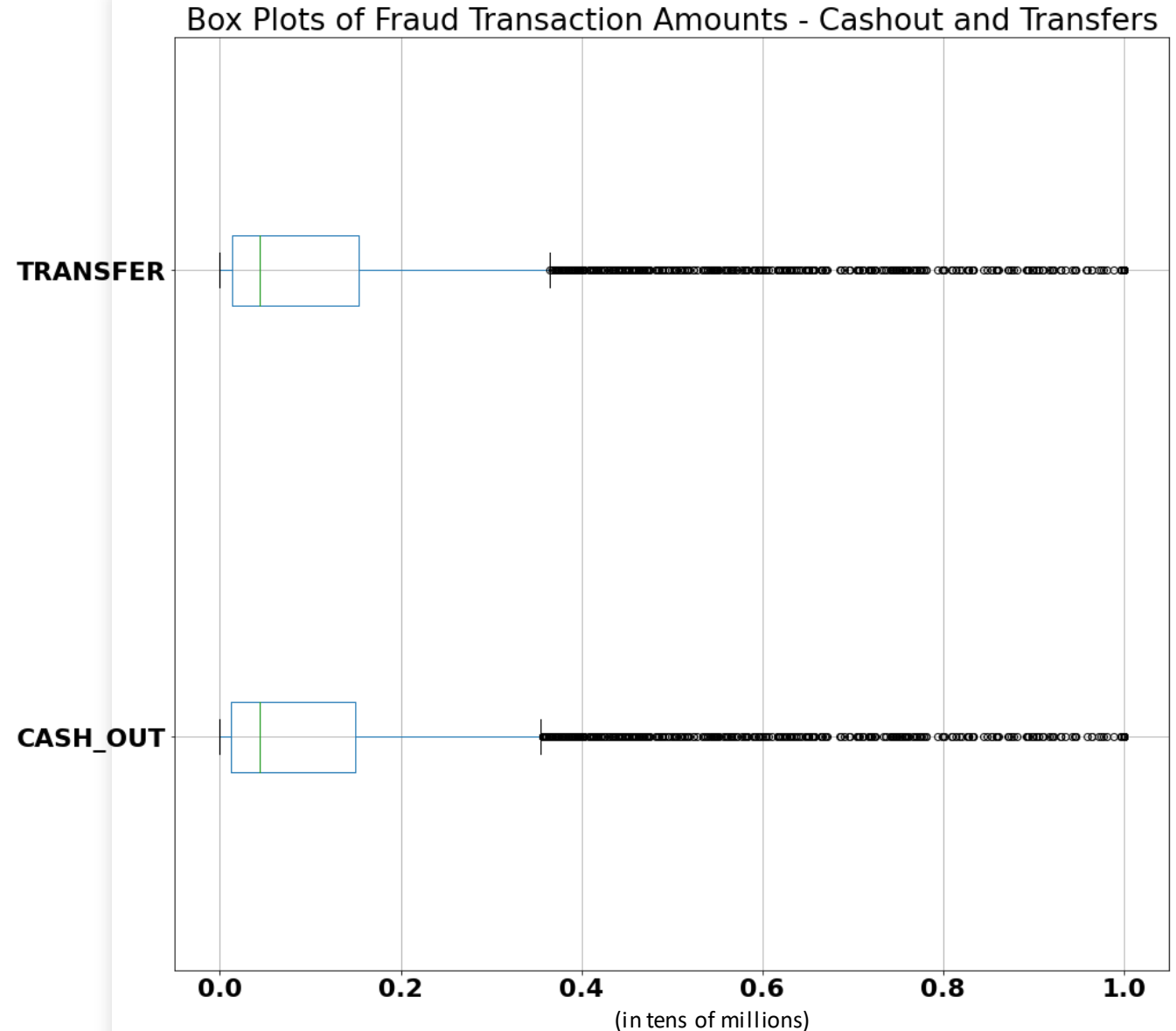
- **0.13 percent** of all transactions were fraudulent
- Out of 2,237,500 cashouts, **4,116 were fraudulent** (.77 percent)
- Out of 532,090 transfers, **4,097 were fraudulent** (.18 percent)



Key Findings – Value of Fraud Cases 03

Overall large ranges in the amount of fraud transactions

- **Total fraud amounts ranged anywhere from 0 to 10,000,000** (in 16 fraud cashout cases, amount of transaction was 0)
- **Average fraud amount:**
 - Transfer: 1,480,892
 - Cashout: 1,455,103
- **Median fraud amount:**
 - Transfer: 445,705
 - Cashout: 435,516





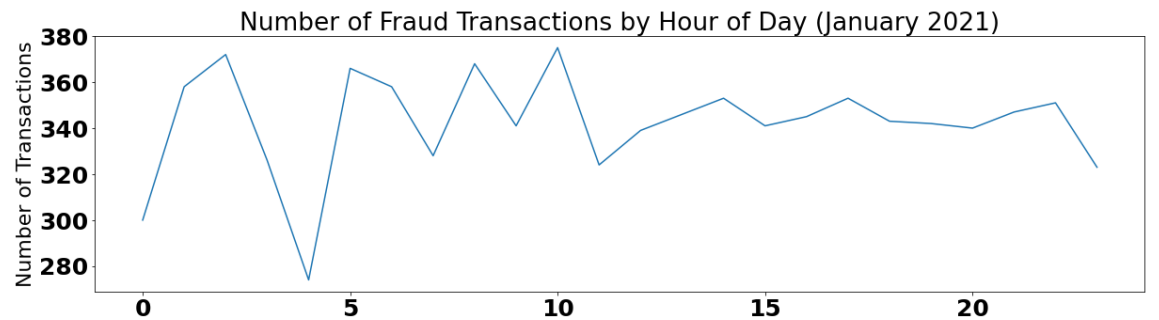
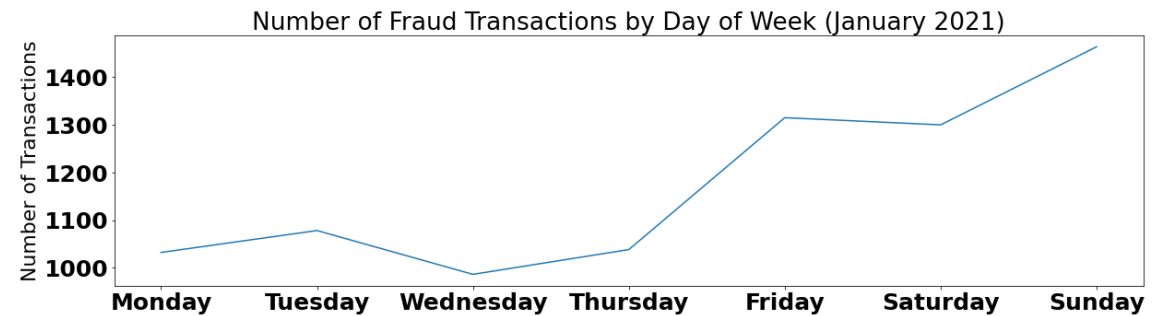
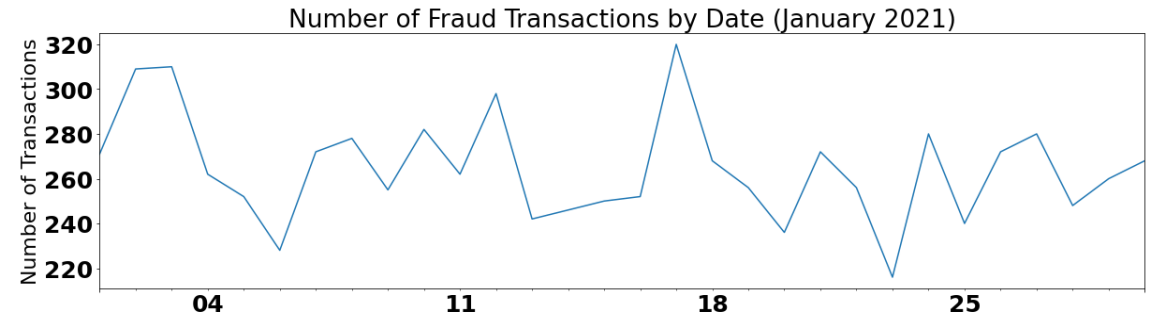
Key Findings – Timing

04

Number of fraud transactions seemed to peak during certain time periods

Patterns by time period:

- ***Date:*** Fraud transactions peaked in the beginning and the middle of the month
- ***Day of week:*** Fraud transactions tended to peak between Friday-Sunday
- ***Hour of the day:*** No dominant time of day when fraud transactions were more prevalent

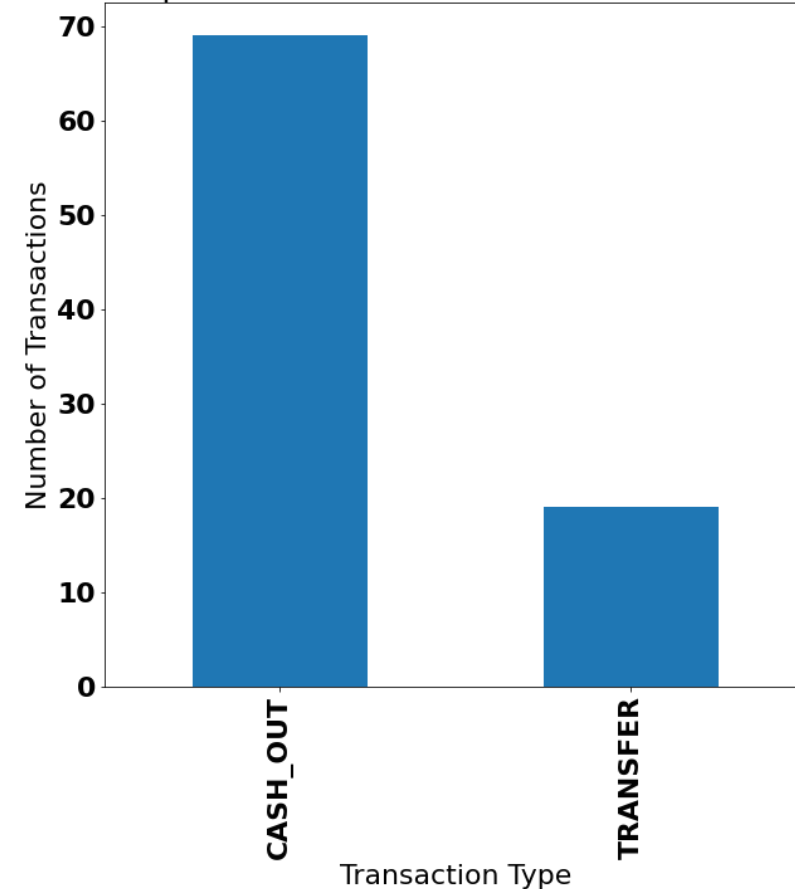



Key Findings – Who was involved? 05

Only customers “C” were involved in fraudulent transactions

- **Merchants “M” were not involved** in fraudulent transactions
- Out of 8,213 fraud cases, there were **88 cases** where destination account (nameDest) was involved twice (fraud committed twice by single user)

Distribution of Multiple Fraud Transactions with Same Destination Account





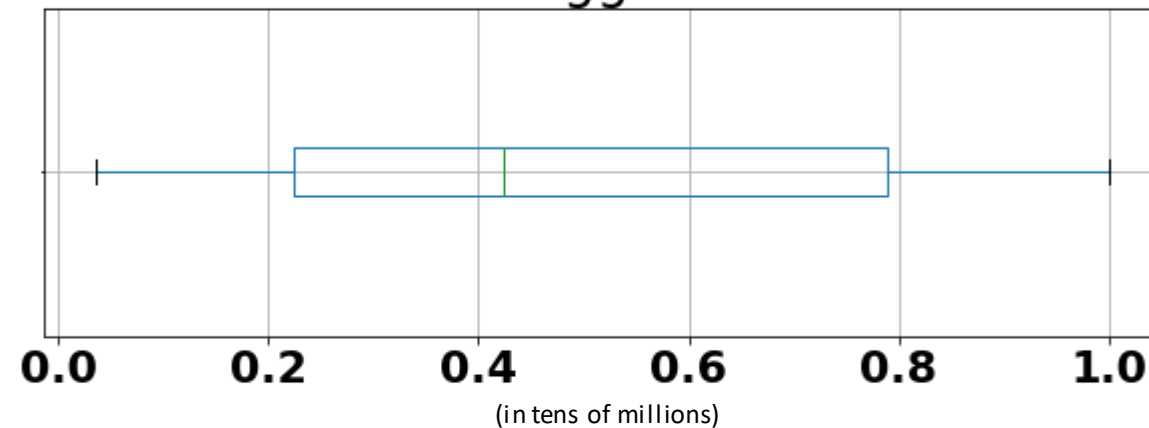
Key Findings – Flagged Cases

06

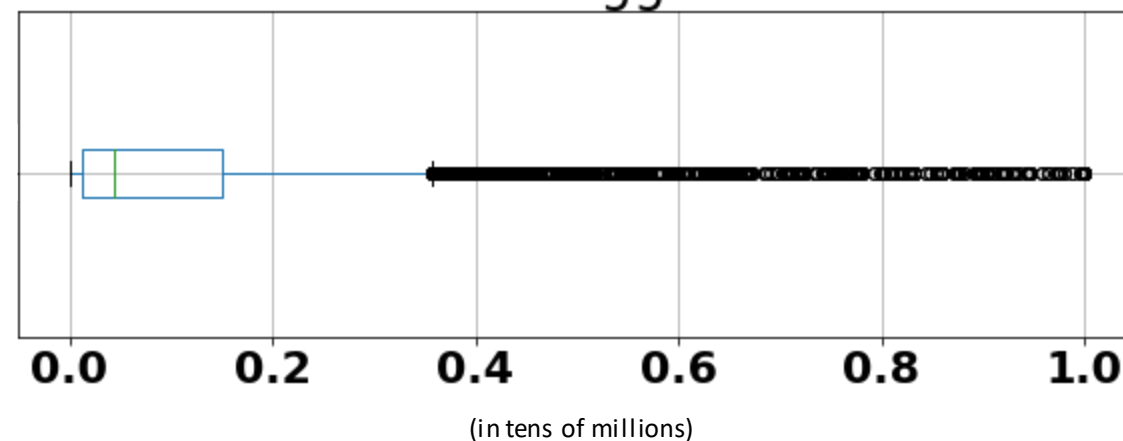
Out of 8,213 fraud cases, only 16 were flagged; mean and median amount of flagged transactions were higher than non-flagged

- **All 16 flagged cases** were “Transfer” transactions
- **Basic Stats of Flagged Cases - Amount:**
 - Range: 353,874 to 10,000,000
 - Average: 4,861,598
 - Median amount: 4,234,245
- **Basic Stats of Non Flagged Cases - Amount:**
 - Range: 0 to 10,000,000
 - Average: 1,461,343
 - Median amount: 438,983
- **No false positives** – all flagged cases were indeed fraud

Distribution of Flagged Fraud Amounts



Distribution of Non Flagged Fraud Amounts



Conclusion

07

Summary of Preliminary Key Findings

1. Out of over 6 million transactions, 0.13 percent of transactions were fraudulent; fraud found only among 2 types of transactions: cashout and transfer
2. Total amount of fraud transactions ranged between 0 and 10,000,000
3. Fraud transactions peaked in the beginning and middle of the month; most fraud transactions also took place between Friday-Sunday
4. Only customers “C” were involved in fraud transactions; 88 fraud cases involved repeat offenders
5. Out of the 8,213 fraud cases, only 16 were flagged as fraud; however, mean and median amount of flagged transactions were higher than non flagged ones

Going forward, additional information/data could prove helpful to validating and identifying more trends

Additional information/data can include...

1. **Larger dataset in terms of time and overall size.** Current data set is scaled down to 1/4th of the original dataset and limited to 1 month (744 steps)
2. **Location data.** (IP address and address of merchants/kiosks used)
3. **Additional customer and merchant data.** (address, phone number, email, duplicate accounts)