**Financial Transaction Fraud Detection Analysis**

**Jahnavi Chintakindi**

**Instructor: Dr. Gahangir Hossain**

**Data Visualization INFO 5709**

**Data Visualization – Term Project**

**12/11/2024**

**Introduction**

In the digital age, financial fraud has become increasingly sophisticated, with criminals exploiting various transaction types and timing patterns. This project analyzes a synthetic dataset of financial transactions to identify patterns and characteristics of fraudulent activities. The goal is to understand how factors such as transaction amounts, timing, and account balances correlate with fraudulent behavior.

Through the visualizations, we examine transaction amounts, timing patterns, and account behaviors to identify potential fraud indicators. This understanding can help develop more targeted approaches to fraud detection in financial systems.

The project combines Python analysis with Tableau visualizations to get a thorough view of how fraudulent transactions differ from legitimate ones. We look at both broad patterns and specific transaction characteristics to build a complete picture of fraudulent behavior in financial systems.

**Dataset**:

The dataset is collected from Kaggle's PaySim synthetic financial dataset. The data simulates mobile money transactions based on real-world patterns, providing a month of financial logs generated using the PaySim mobile money simulator. This dataset contains transaction records with various attributes that help understand patterns of both legitimate and fraudulent activities.

Dataset source: <https://www.kaggle.com/datasets/ealaxi/paysim1>

**Attributes**:

Total of 11 attributes are used to describe the transactions:

* step: Maps to hour of transaction (1-744 representing 30 days)
* type: Transaction type (TRANSFER, CASH\_OUT, DEBIT, PAYMENT, CASH\_IN)
* amount: Transaction amount
* nameOrig: Customer initiating transaction
* oldbalanceOrg: Origin account balance before transaction
* newbalanceOrig: Origin account balance after transaction
* nameDest: Recipient of transaction
* oldbalanceDest: Destination account balance before transaction
* newbalanceDest: Destination account balance after transaction
* isFraud: Binary fraud indicator (1 = fraud, 0 = legitimate)
* isFlaggedFraud: System flag for suspicious transaction

**Tools**:

* Python
  + pandas for data manipulation
  + matplotlib and seaborn for visualization
  + numpy for numerical operations
* Tableau for interactive visualizations

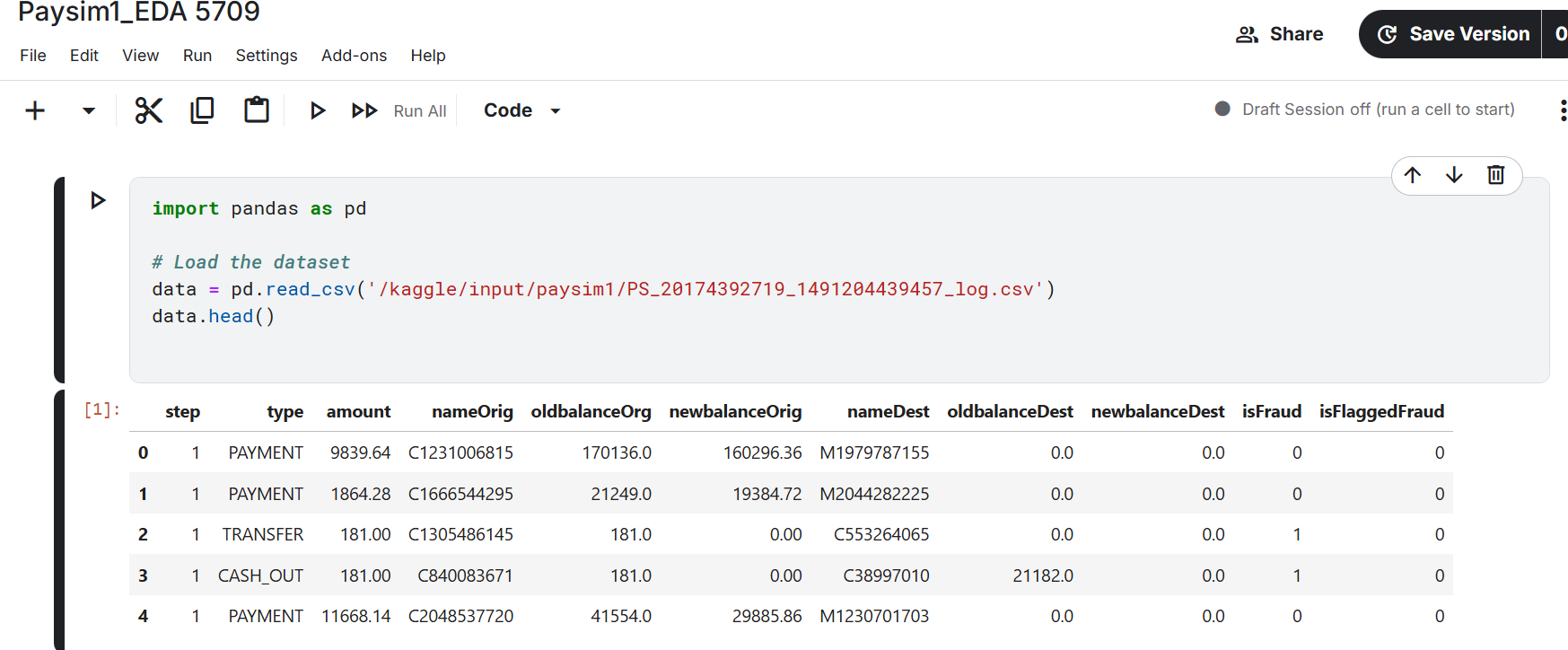
**Data Cleaning:**

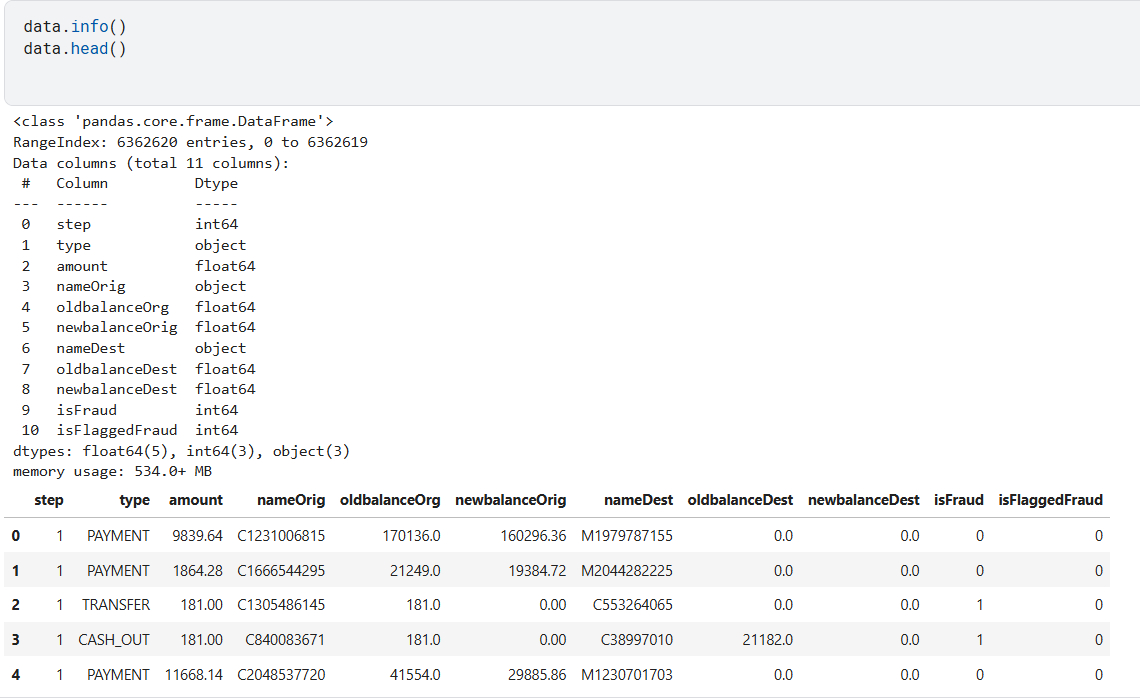
* Missing latitude and longitude attributes were replaced with the city lats and longs
* Created derived features:
  + hour: Extracted from 'step' attribute for temporal analysis
  + balance\_diff: Calculated balance changes
* Standardized transaction types
* Verified data consistency and removed duplicates

**Exploratory Data Analysis:**

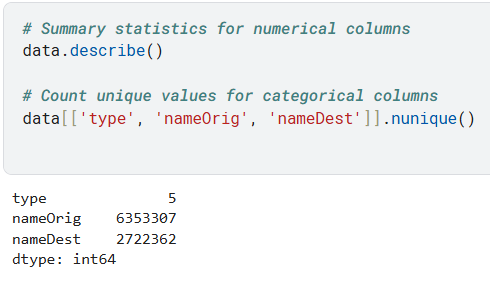
**Initial data exploration using Python:**

**Step1**: Loading the PaySim dataset using pandas and showing the first 5 rows of transaction data including payment types, amounts, and balances.

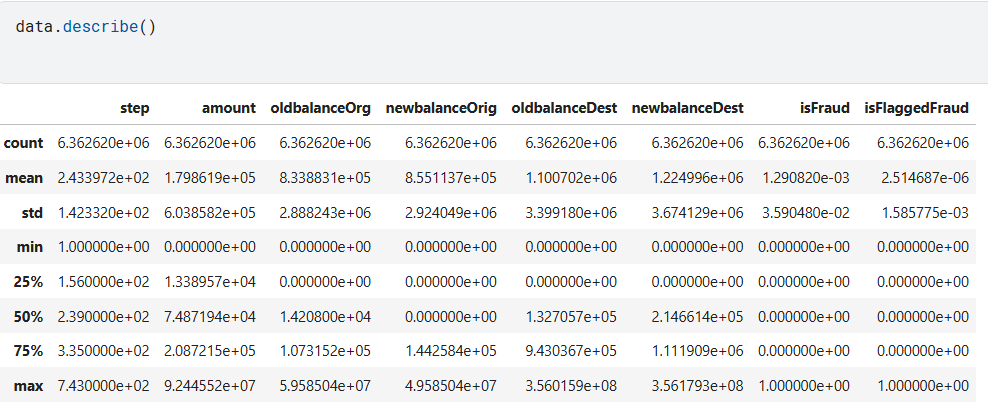


**Step2**: Overview of dataset structure showing 6.3M transactions with 11 columns including their data types and memory usage.

**Step 3:** Quick analysis showing 5 transaction types, 6.3M unique origin accounts, and 2.7M unique destination accounts.



**Step 4:** Descriptive statistics showing transaction amounts range from 0 to 92M with a mean of 179K and a fraud rate of 0.129%.



**Step 5**: Histogram showing right-skewed distribution of transaction amounts on a logarithmic scale with most transactions concentrated in lower amounts.

A screenshot of a computer

Description automatically generated

**Based on initial exploration, three hypotheses were developed:**

1. Fraudulent transactions involve higher amounts than legitimate ones (H1)
2. Fraud is concentrated in specific transaction types - TRANSFER and CASH\_OUT (H2)
3. Fraudulent transactions are more likely during early morning hours (H3)

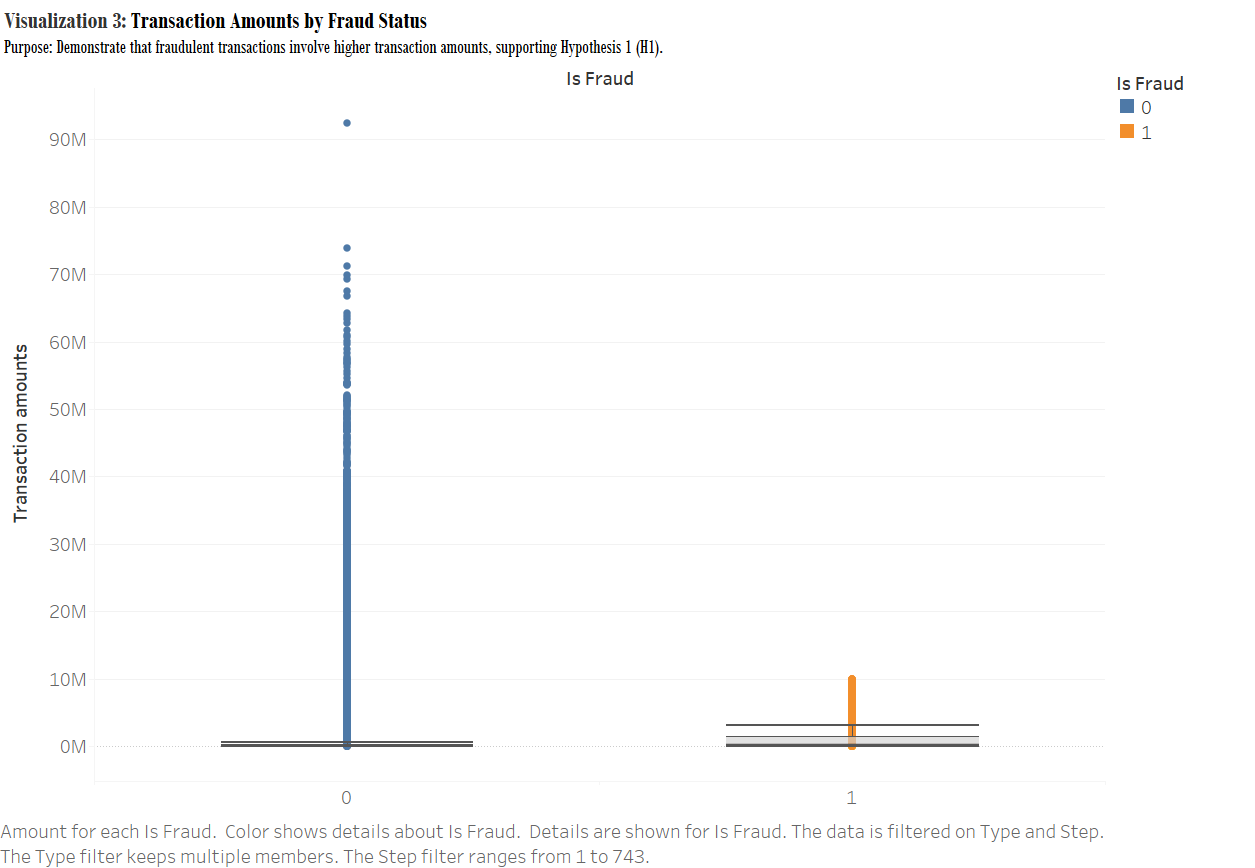
## **Graph-breakthrough-Result**

### **1. Transaction Amounts by Fraud Status (H1)**

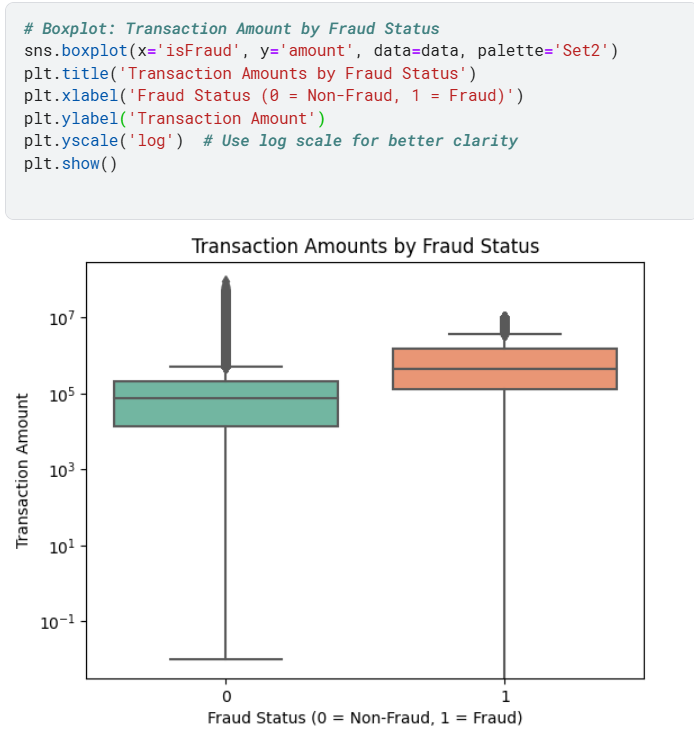
**Tableau Visualization**

**Visualization 3**: Transaction Amounts by Fraud Status

**Purpose**: Demonstrate that fraudulent transactions involve higher transaction amounts.



**Python Analysis:**



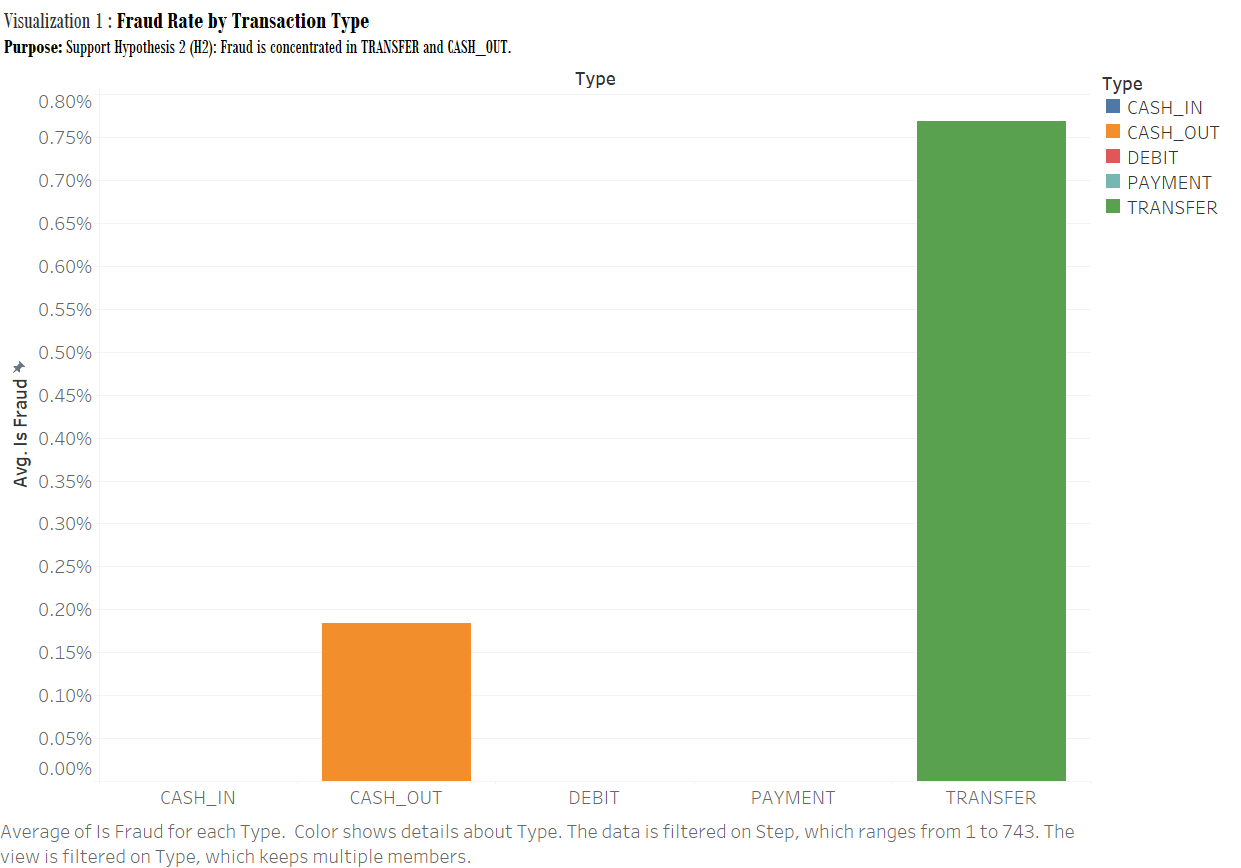
**Result**: The visualization confirms that fraudulent transactions typically involve higher amounts. The box plots show a clear distinction in amount distributions between legitimate and fraudulent transactions.

### **2. Fraud Rate by Transaction Type (H2)**

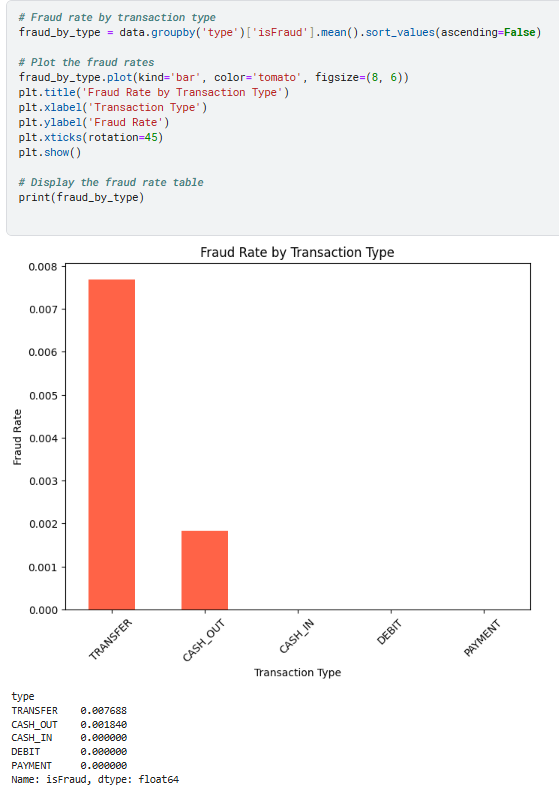
**Tableau Visualization**

**Visualization 1**: Fraud Rate by Transaction Type

**Purpose**: Support hypothesis that fraud is concentrated in TRANSFER and CASH\_OUT transactions.



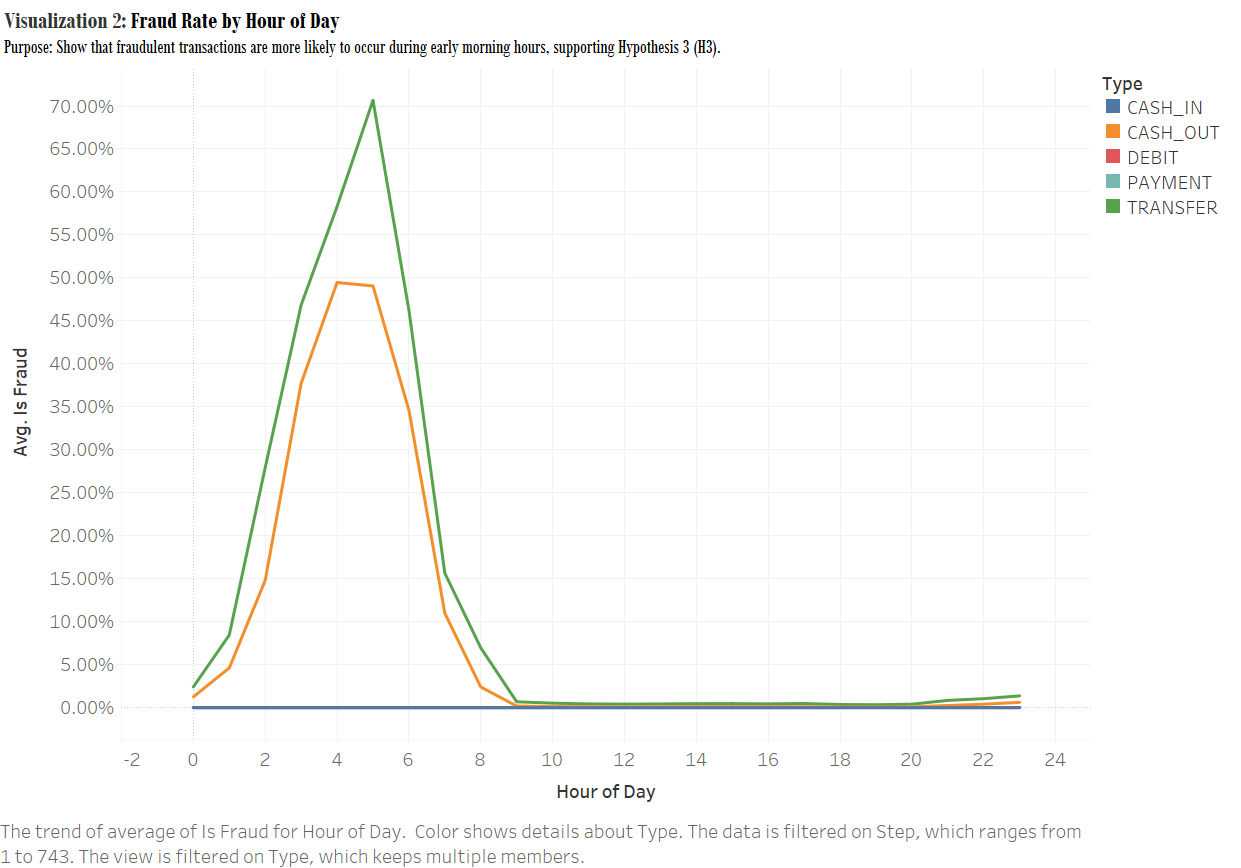
**Python Analysis**



**Result**: The analysis shows TRANSFER transactions have the highest fraud rate (0.8%), followed by CASH\_OUT (0.2%). Other transaction types show minimal fraud risk.

### **3. Temporal Fraud Patterns (H3)**

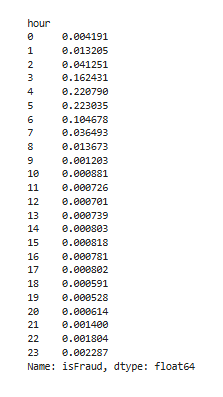
**Tableau Visualization** [Visualization 2: Fraud Rate by Hour of Day] Purpose: Show that fraudulent transactions are more common during early morning hours.



**Python Analysis:**

A graph with a line and dots

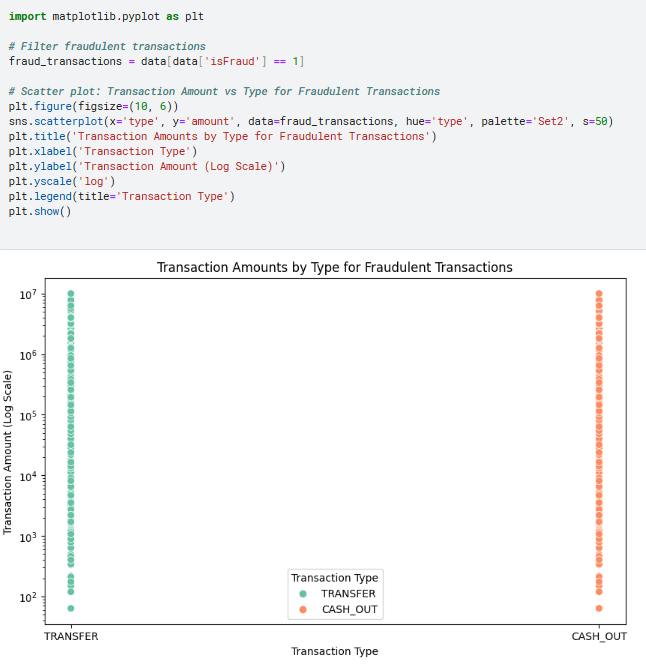
Description automatically generated with medium confidence

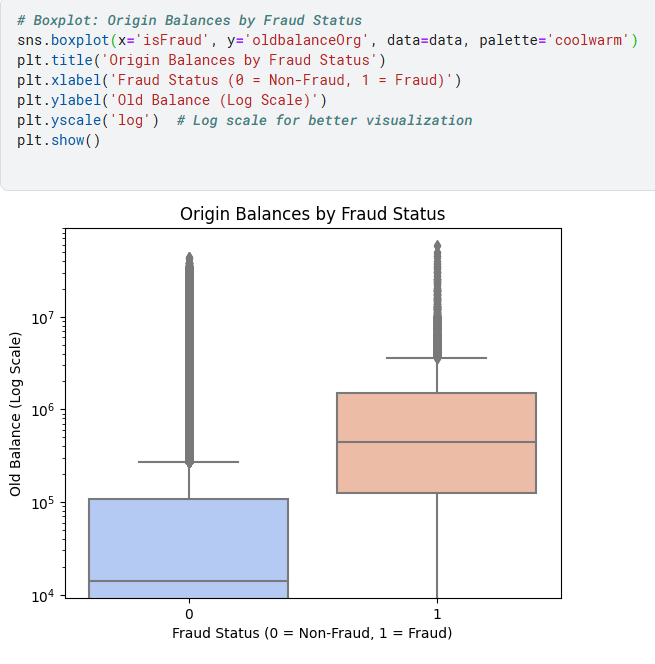


**Result**: Clear peak in fraudulent activity between 2-5 AM, with maximum fraud rate around 3-4 AM (approximately 22%).

### **Additional Insights: Account Balance Analysis**

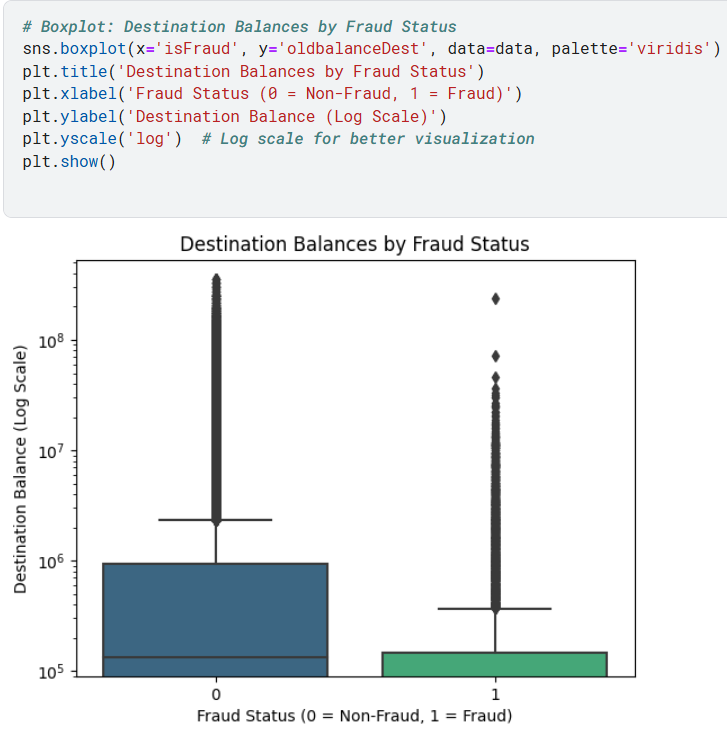
**Origin Balance Analysis**





This box plot visualization depicts the origin balances by fraud status. The y-axis shows the old balance on a logarithmic scale, while the x-axis distinguishes between non-fraudulent (0) and fraudulent (1) transactions. The plot provides a clear comparison of the origin balances between these two fraud statuses.

**Destination Balance Analysis**

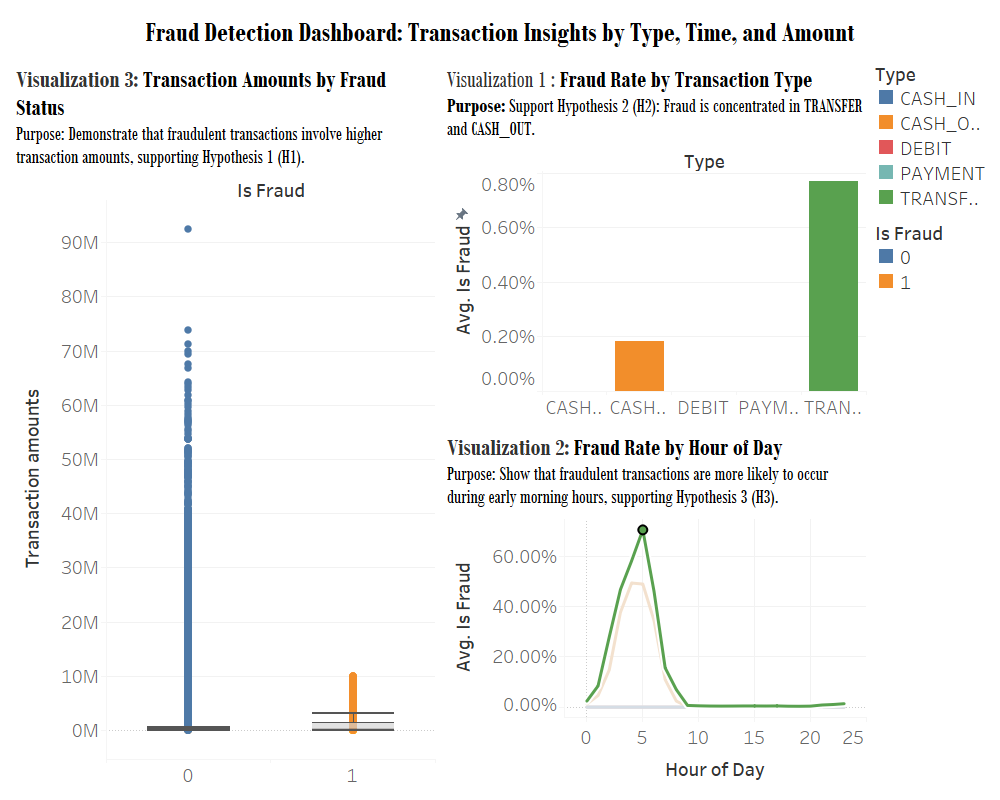


**Result**: Fraudulent transactions show distinct patterns in account balances:

* Higher origin account balances in fraudulent transactions
* Lower destination account balances in fraudulent cases
* Wider range of balance variations in fraudulent transactions

## **Key Findings**

**Tableau Dashboard:**



1. **Transaction Amount Patterns:** 
   1. Fraudulent transactions consistently show higher amounts
   2. Clear separation in amount distributions between fraud and non-fraud
   3. Higher variance in fraudulent transaction amounts
2. **Transaction Type Analysis:** 
   1. TRANSFER and CASH\_OUT are primary vehicles for fraud
   2. Other transaction types show minimal fraud risk
   3. TRANSFER transactions show highest fraud concentration
3. **Temporal Patterns:** 
   1. Early morning hours (2-5 AM) show highest fraud rates
   2. Peak fraud activity around 3-4 AM
   3. Business hours show minimal fraud activity
4. **Balance Patterns:** 
   1. Fraudulent transactions typically originate from high-balance accounts
   2. Destination accounts in fraud cases show lower balances
   3. Balance patterns provide strong fraud indicators

## **Discussions**

* The findings support all three initial hypotheses
* Clear patterns emerge in transaction timing, amounts, and types
* Account balance patterns provide additional fraud indicators
* Results suggest multiple approaches for fraud detection:
  + Enhanced monitoring during high-risk hours
  + Transaction amount thresholds
  + Account balance pattern monitoring
  + Transaction type risk scoring

## **Conclusion**

This project successfully identified key patterns in fraudulent financial transactions:

* Temporal patterns showing increased risk during early morning hours
* Transaction type vulnerabilities in TRANSFER and CASH\_OUT
* Clear amount and balance patterns distinguishing fraudulent transactions
* Multiple indicators available for fraud detection systems

## **References**

1. Lopez-Rojas, E., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. In The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus.
2. PaySim. (2017). Synthetic Financial Datasets For Fraud Detection [Dataset]. Kaggle. <https://www.kaggle.com/datasets/ealaxi/paysim1>