Executive Summary

Exploratory Data Analysis: Bank Transaction Dataset

Anthony Roca | aroca@charlotte.edu | January 21, 2026

1. Dataset Overview and Data Quality

This analysis examines a synthetic bank transaction dataset containing 2,512 transaction records across 16 features, sourced from Kaggle for fraud detection pattern exploration. The dataset demonstrates exceptional data quality with zero missing values, complete type consistency, and validated integrity across all fields. All 2,512 TransactionIDs are unique, no negative transaction amounts exist, and categorical variables maintain expected value sets across three channels (Online, ATM, Branch) and two transaction types (Debit, Credit).

Metric	Value	Insight
Total Records	2,512	Modest dataset size
Unique Accounts	495	~5 transactions per account
Missing Values	0	100% data completeness
Avg Transaction	$297.59	High variance ($291.95 SD)

2. Critical Findings: Synthetic Data Characteristics

Multiple analytical dimensions reveal patterns inconsistent with real-world banking data, confirming the synthetic nature of this dataset:

Geographic Distribution Anomalies

- New York City ranks 28th out of 43 cities with only 58 transactions, despite being the largest US city and financial capital
- Fort Worth leads with 70 transactions, followed by Charlotte (66) and Omaha (66), creating an inverse relationship between city size and transaction volume
- Transaction distribution shows remarkable uniformity across all 43 cities, inconsistent with real-world metropolitan concentration patterns

Temporal Pattern Constraints

- 100% of transactions occurred on weekdays (Monday-Friday) with zero weekend activity, highly atypical for modern banking where ATM and online channels operate continuously

- Monday represents 42.6% of all transactions (1,070 records), creating extreme concentration on a single weekday
- Transaction times concentrated in narrow 3-hour afternoon window (4pm-6pm) with zero activity during overnight hours (10pm-6am)
- PreviousTransactionDate column contains uniform November 2024 timestamps while TransactionDate spans 2023, indicating reference timestamp rather than actual prior transaction dates

3. Business Intelligence Insights

Despite synthetic constraints, the dataset enables meaningful exploration of fraud detection methodologies and pattern analysis techniques:

Account Activity Patterns

- Transactions per account show narrow distribution (1-12 range) with mean of 5.07 and median of 5.0, indicating balanced activity across the customer base rather than the power-law distribution typical of real banking
- Slight positive skewness (0.380) suggests some accounts with above-average activity, but lacks the extreme tail behavior seen in real-world datasets

Multi-Channel Engagement

- Near-equal channel split: Branch (34.6%), ATM (33.2%), Online (32.3%) indicates customers utilize multiple banking channels without strong preference
- Transaction amounts remain consistent across channels, suggesting transaction size does not correlate with channel selection
- Median transaction durations and login attempts show minimal variation across channels, contrary to expectations where online transactions typically exhibit different behavioral patterns

Device and IP Address Sharing

- 681 unique devices serve 495 accounts (1.38 accounts per device average), with some devices associated with up to 9 different accounts
- 592 unique IP addresses serve 495 accounts (1.21 accounts per IP average), indicating moderate cross-account sharing
- Cross-account patterns could represent legitimate scenarios (family sharing, business terminals) or fraudulent activity signals for detection model training

Merchant Ecosystem

- 100 unique merchants across 2,512 transactions demonstrate distributed transaction landscape without extreme concentration

- Account-level merchant diversity indicates customers engage with multiple merchants, reflecting normal consumer spending patterns

4. Recommended Analytical Follow-Ups

Fraud Detection Model Development

- Engineer features combining device/IP sharing patterns with transaction velocity metrics to identify potential account takeover scenarios
- Develop anomaly detection models focusing on deviations from established temporal patterns, despite synthetic constraints on timing distribution
- Create merchant-account affinity networks to identify unusual transaction routing or merchant concentration anomalies

Feature Engineering Opportunities

- Construct rolling window statistics (7-day, 30-day) for transaction frequency, amount volatility, and channel switching behavior
- Calculate balance-to-transaction ratios and balance velocity metrics to identify accounts with unusual financial dynamics
- Generate cross-dimensional interaction features combining temporal patterns, geographic attributes, and device metadata

Model Architecture Considerations

- Test ensemble methods (Random Forest, Gradient Boosting) against neural network architectures for handling mixed categorical and numerical features
- Implement class imbalance handling techniques given typical fraud detection scenarios with rare positive cases
- Evaluate graph-based approaches leveraging device-account-merchant network structures for fraud propagation detection

Validation Strategy

- Design time-based validation splits respecting temporal ordering to prevent data leakage and ensure model generalization
- Establish performance metrics beyond accuracy including precision, recall, F1-score, and ROC-AUC given fraud detection priorities
- Conduct sensitivity analysis on synthetic data artifacts to understand model robustness and transferability to real-world scenarios