Data Understanding for Synthetic Fraud Generation: A Diffusion Model Approach to Sparkov Fraud Dataset

Akash Murali, Anish Rao, Raghu Ram Sattanapalle

January 28, 2025

Source

The dataset used for this analysis is the *'Credit Card Transactions Fraud Detection Dataset'*, obtained from Kaggle.

- **Link:** `https://www.kaggle.com/datasets/kartik2112/fraud-detection/data`
- **Access:** Publicly available with no restrictions.

Structure and Metadata

The dataset has 1296675 rows and 23 columns Possible key features identified in the dataset include:

- **trans_date_trans_time:** Time and date of the transaction.
- **amt:** Transaction amount.
- **category:** Type of transaction.
- **job:** Job of the cardholder.
- **merch_lat** and **merch_long:** Coordinates of the transaction.

Other key features that correlate with fraudulent transactions may exist and will require further analysis.

Table 1: Statistical Summary of Selected Features in the Dataset

Feature	Count	Mean	Std	Min	25%	Max
amt	1,296,675	70.35	160.32	1.00	9.65	28,948.90
lat	1,296,675	38.54	5.08	20.03	34.62	66.69
long	1,296,675	-90.23	13.76	-165.67	-96.80	-67.95
city_pop	1,296,675	88,824.44	301,956.42	23.00	743.00	2,906,700.00
is_fraud	1,296,675	0.0058	0.0759	0.00	0.00	1.00
merch_lat	1,296,675	38.54	5.11	19.03	34.73	67.51
merch_long	1,296,675	-90.23	13.77	-166.67	-96.90	-66.95

Missing Data

- No missing values were found in this dataset.

Anomalies

- Each feature was plotted to check for anomalies.
- **Geographical data:** Consistent with no invalid coordinates. Hotspots of fraudulent activity were observed upon mapping the coordinates.

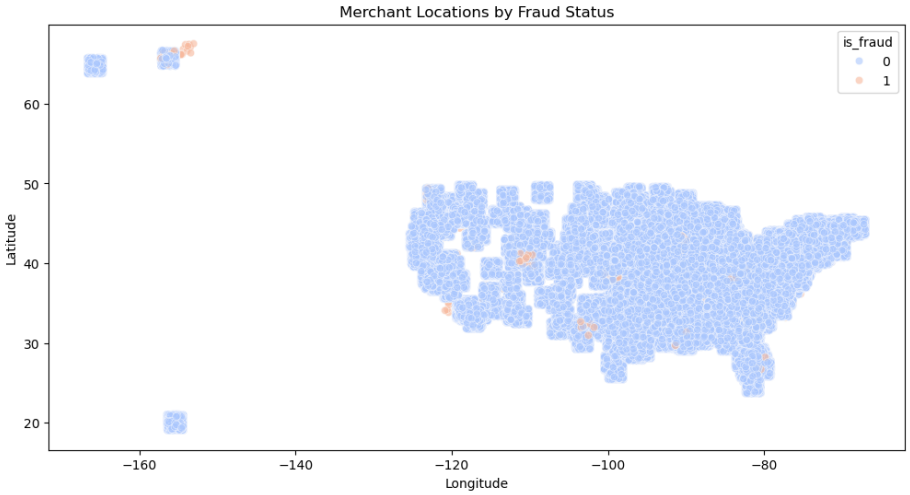

Figure 1: Merchant Locations by Fraud Status

- **Time data:** Transactions occur more frequently during the day, as expected.
- The dataset is logically consistent; no negative transactions or future dates were found.

Bias

- The dataset exhibits an extreme class imbalance, with only around 0.5% of transactions being fraudulent.

Figure 2: Class Imbalance

- Significant transaction volume discrepancies between states may introduce bias.

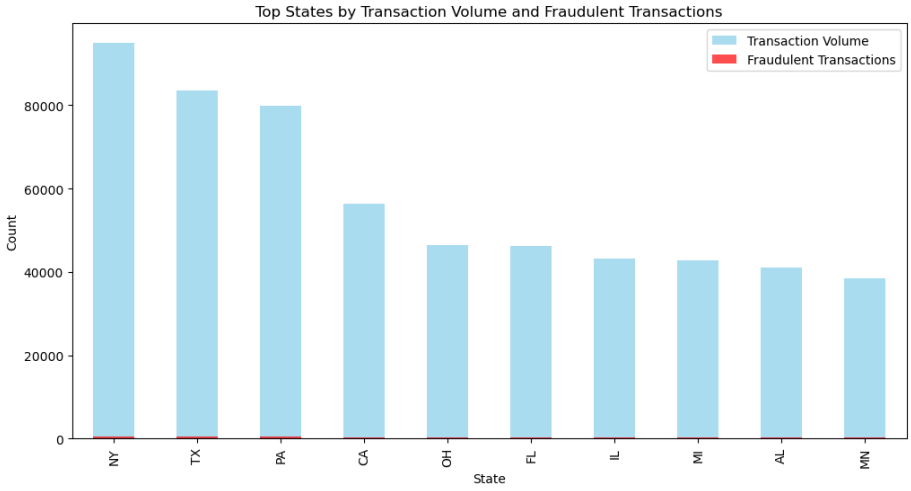

Figure 3: Transaction Volume by State

- During the first six months of the year, a higher prevalence of fraudulent transactions was observed, although the cause is unclear.

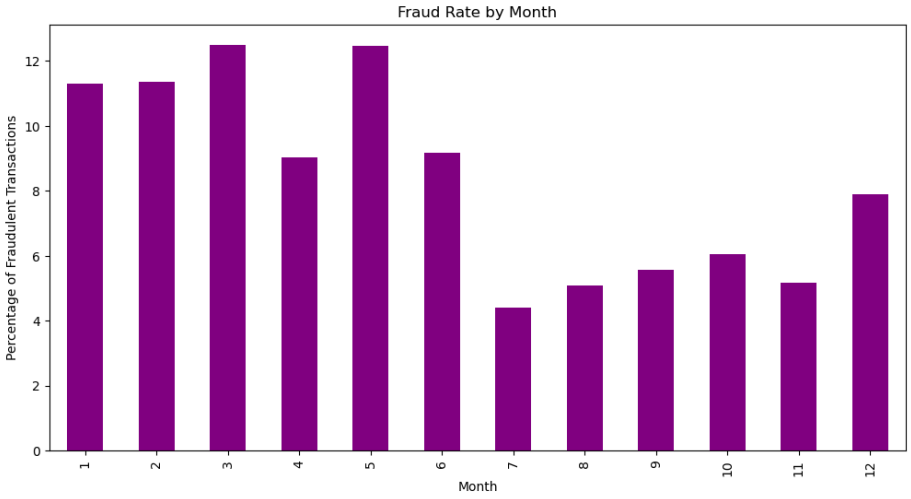

Figure 4: Fraud Rate by Month

- Furthermore, as can be seen in Figure 1, most fraudulent transactions are centered around certain pockets on the map. This could introduce some bias.

Distribution

- Most numerical features have skewed distributions, with a majority of values concentrated in specific ranges.
- Strong correlations were observed among geographical features such as *Lat*, *Long*, and *Zipcode*.
- There is a minor correlation between *amt* (transaction amount) and *is_fraud*.

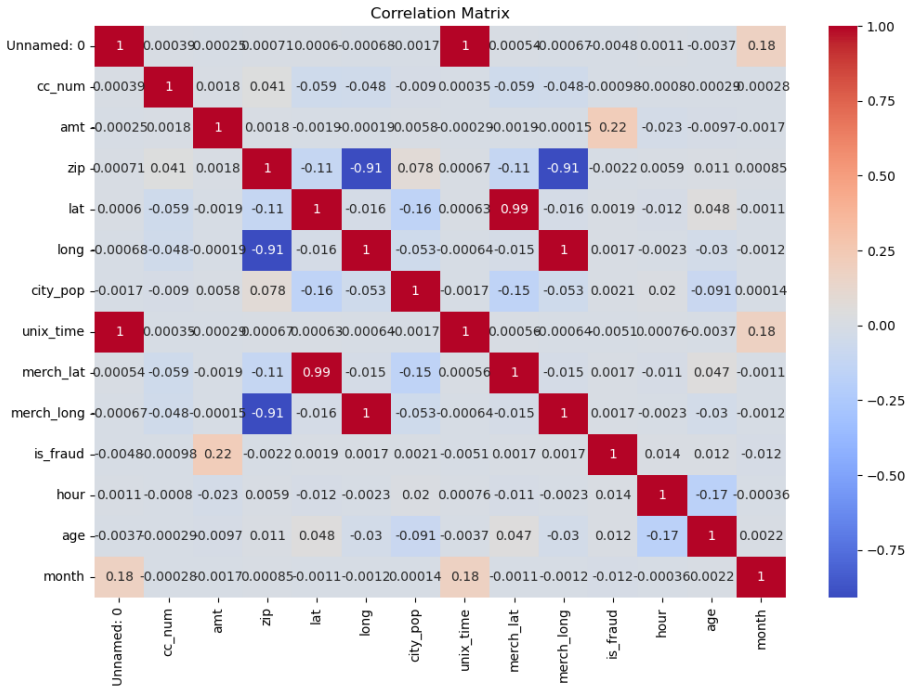

Figure 5: Correlation Matrix

Categorical Data

The dataset contains the following unique categories for each categorical variable:

Variable	Unique Categories
Category	14
Gender	2
State	51
Job	494

Table 2: Number of unique categories in each categorical variable.

Category Distribution for `category`

The distribution of the top categories in `category` is as follows:

Category	Percentage
Gas/Transport	10.15%
Grocery (POS)	9.54%
Home	9.49%
Shopping (POS)	9.00%
Kids/Pets	8.72%

Table 3: Top 5 categories in `category` distribution.

Category Distribution for `gender`

The `gender` feature is well-balanced, with the following distribution:

Gender	Percentage
Female (F)	54.74%
Male (M)	45.25%

Table 4: Distribution of `gender` feature.

Category Distribution for `state`

The `state` feature is highly imbalanced. The top states are:

State	Percentage
Texas (TX)	7.32%
New York (NY)	6.44%
Pennsylvania (PA)	6.16%
California (CA)	4.35%
Ohio (OH)	3.58%

Table 5: Top 5 states in `state` distribution.

Category Distribution for `job`

The `job` feature has 494 unique categories, with the top 5 being:

Job	**Percentage**
Film/Video Editor	0.75%
Exhibition Designer	0.71%
Naval Architect	0.67%
Surveyor (Land/Geomatics)	0.67%
Materials Engineer	0.64%

Table 6: Top 5 jobs in `job` distribution.

Fraud Transaction Percentage by Age Group

Fraud transaction percentages by age group are shown in the bar chart below:

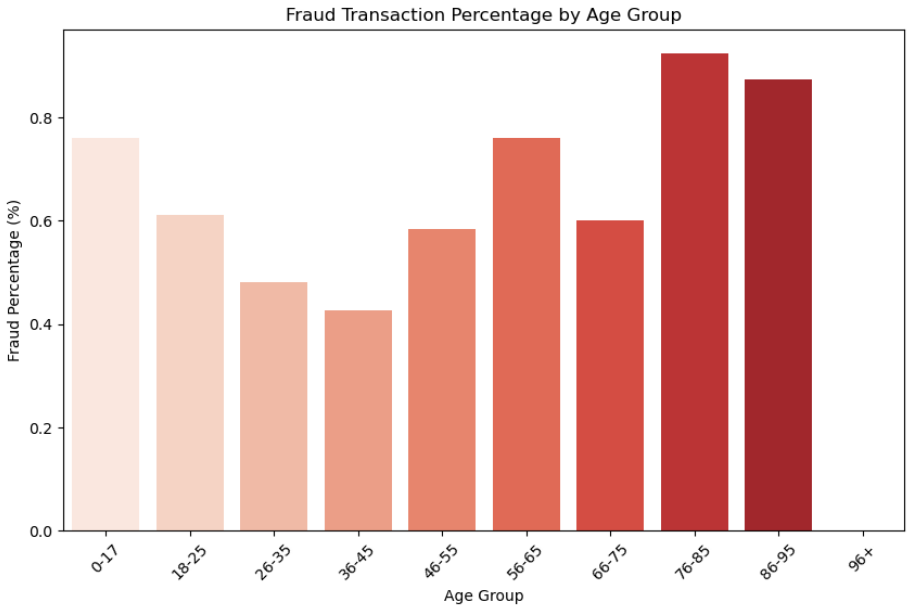

Figure 6: Fraud Transaction Percentage by Age Group.

Age Group	Fraud Percentage
0-17	0.7609
18-25	0.6103
26-35	0.4815
36-45	0.4258
46-55	0.5842
56-65	0.7612
66-75	0.6018
76-85	0.9234
86-95	0.8738
96+	0.0000

Table 7: Fraud Transaction Percentages by Age Group.

Category features analysis

- **Unique Categories:** The dataset contains a significant number of unique categories, particularly in the `job` feature (494 categories), indicating high diversity.
- **Balance:** While the `gender` feature is balanced, `state` and `job` features show significant imbalance, with a small number of categories dominating the distribution.
- **Fraud Analysis by Age Group:** Fraudulent transactions are highest among age groups 76-85 (0.92%) and 86-95 (0.87%), suggesting potential targeting or vulnerability in older age groups.

Ethical Considerations

The dataset contains multiple forms of Personally Identifiable Information (PII) and sensitive data that must be handled with care to comply with privacy regulations. The key sensitive attributes include:

- **cc_num (Credit card number):** Highly sensitive financial information.
- **first, last (Name fields):** Directly identifiable information.
- **dob (Date of birth):** Often used for identity verification.
- **street, city, state, zip (Address details):** Can reveal an individual's residence.
- **trans_num (Transaction ID):** May contain identifiers linking to personal financial records.

Alignment with Project Goals

The dataset aligns well with the project's objectives of fraud detection and class imbalance handling using diffusion models. Some of the important features for the project include:

- **Transaction Attributes:** ant, merchant, category, trans_date_trans_time, unix_time.
- **Personal and Geographic Features:** dob, gender, street, city, state, zip, lat, long.

The above features are crucial for data augmentation and provide sufficient information to train a robust classifier model.

Scalability

The dataset size and memory footprint can be managed on a standard modern computer (8GB RAM or more). Key considerations include:

- Reading the data into memory using tools like pandas should take a few seconds to minutes.
- The dataset does not require advanced techniques such as distributed processing to handle its size under typical scenarios.

These characteristics make the dataset manageable with the available resources, ensuring efficient processing without the need for specialized infrastructure.

Transformations

Fraud Transaction Analysis by Time

The following analyses provide insights into fraud and non-fraud transactions based on the hour of the day and the day of the week:

Fraud vs Non-Fraud Transactions by Hour of the Day

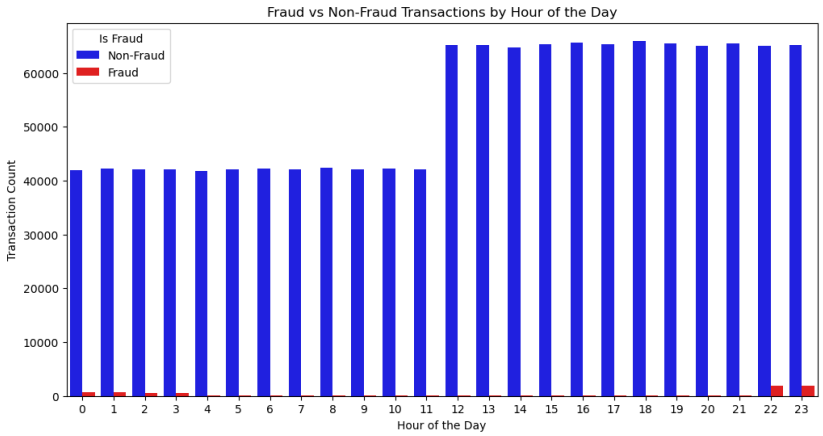

Figure 7: Fraud vs Non-Fraud Transactions by Hour of the Day.

Fraud vs Non-Fraud Transactions by Day of the Week

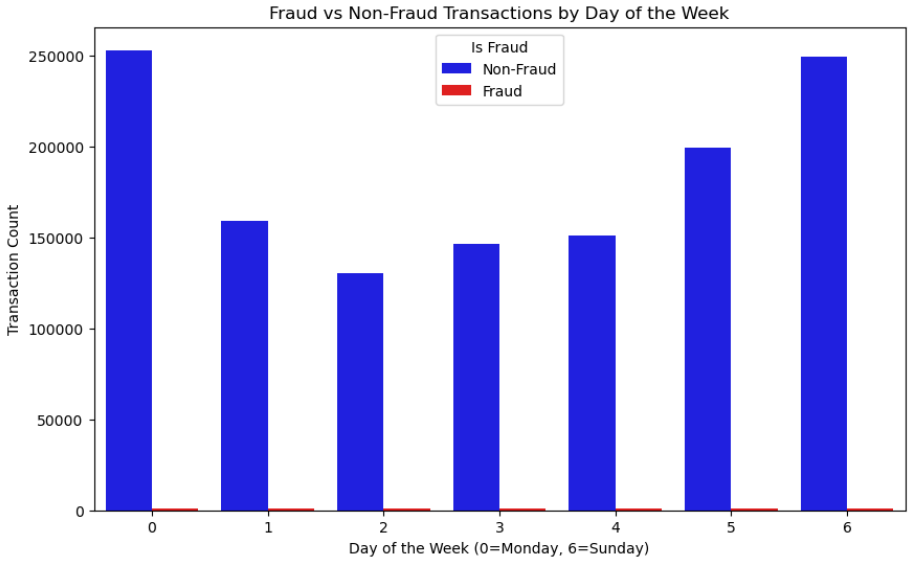

Figure 8: Fraud vs Non-Fraud Transactions by Day of the Week.

Preprocessing and Feature Engineering

- **Preprocessing:** Numerical features such as `amt` and `city_pop` have varying scales and units. Standardization or normalization is required to bring these features to the same scale.
- **Feature Engineering:** New features such as `transaction_hour` and `transaction_day` were created from `trans_date_trans_time` to enable time-based analysis of transactions.

These transformations enhance the dataset's utility for model training and provide additional insights for fraud detection.

Data Encoding

The following encoding techniques are recommended for categorical variables:

- **One-Hot Encoding:** Suitable for features with low cardinality, such as `category` (14 unique values) and `gender` (2 unique values).
- **Target Encoding:** Recommended for `state`, which has medium cardinality. This technique can capture fraud patterns effectively.
- **Frequency Encoding:** Useful for high cardinality features such as `job` and `merchant`, reducing dimensions while retaining information.
- **Label Encoding:** Applicable to `merchant` for tree-based models or `job` for scenarios where frequency encoding is not preferred.

Temporal Features

- Temporal features such as `trans_date_trans_time` have been transformed into `transaction_hour` and `transaction_day` for sequential analysis.

These encoding and transformation strategies ensure the categorical and temporal variables are efficiently utilized for model training and analysis

Predictive Power

Feature Predictive Analysis

- Numerical features show varying predictive strength:
  - Transaction amount `amt` demonstrates highest predictive power (MI: 0.0158) after standardization
  - Geographic features show moderate predictive value (lat: 0.00395, long: 0.00390)

  - City population provides limited signal (MI: 0.00302)
- Engineered temporal features enhance predictability:
  - Transaction hour and day derived from trans_date_trans_time
  - Enables capture of temporal fraud patterns

Feature Selection Strategy

- Dimensionality reduction necessary considering:
  - Varying predictive power across features
  - High-cardinality categorical variables requiring specific encoding strategies
  - Diffusion model efficiency requirements
- Recommended approach:
  - Retain standardized numerical features with strong MI scores
  - Incorporate encoded categorical features based on cardinality:
    * One-hot encoding for low-cardinality features (category, gender)
    * Target encoding for medium-cardinality features (state)
    * Frequency encoding for high-cardinality features (merchant, job)
  - Include engineered temporal features for sequence modeling

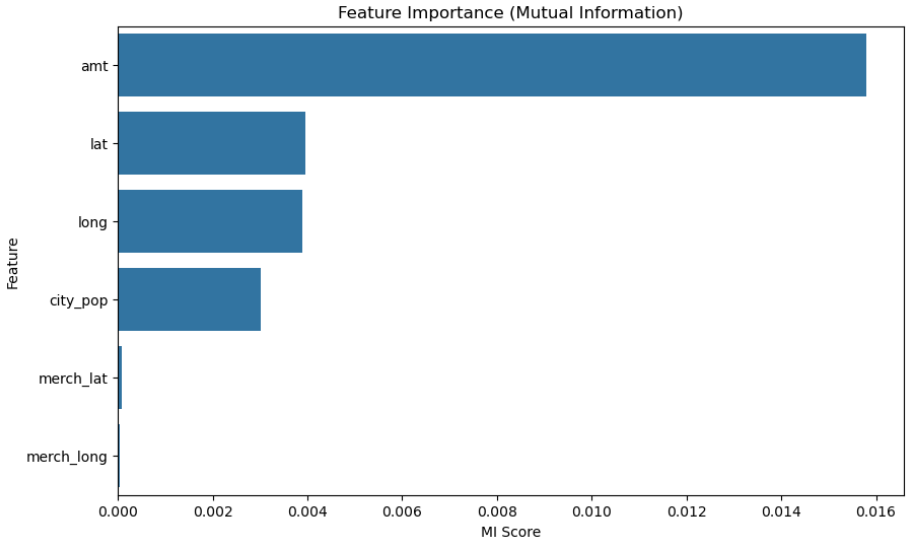

Figure 9: Feature Importance Analysis using Mutual Information Scores for Numerical Features

Target Variable

- **Target Definition**:
  - Binary classification problem with target variable `is_fraud`
    - 0: Legitimate transaction
    - 1: Fraudulent transaction
- **Class Distribution**:
  - Severe class imbalance:
    - Non-Fraud (0): 99.48% of transactions
    - Fraud (1): 0.52% of transactions
    - Imbalance ratio approximately 1:192
- **Special Handling Requirements**:
  - Class imbalance necessitates synthetic data generation
  - Diffusion model approach particularly suitable for:
    - Generating realistic minority class samples
    - Preserving feature relationships in synthetic data
    - Addressing class imbalance while maintaining data quality

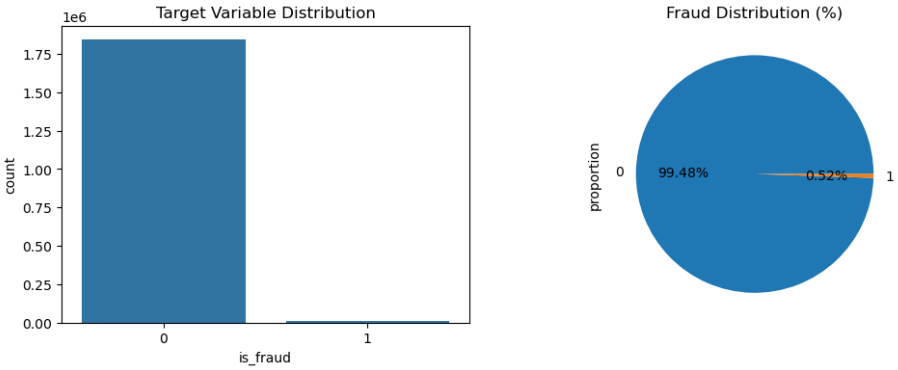

Figure 10: Distribution of Fraud vs Non-Fraud Cases

Validation Strategy

- **Dataset Usage for Diffusion Model**:
  - Full Dataset Utilization:

        * Use complete dataset to learn fraud patterns
        * Include both `fraudTrain.csv` and `fraudTest.csv`
        * Maximize pattern learning from all available fraud cases
    - Pattern Preservation Focus:
        * Temporal patterns (hour of day, monthly variations)
        * Spatial patterns (merchant and customer locations)
        * Transaction value distributions

- **Evaluation Strategy**:
    - Distribution Matching:
        * Compare generated samples vs real fraud distributions
        * Assess feature correlation preservation
        * Validate temporal and spatial pattern maintenance
    - Quality Metrics:
        * Statistical similarity measures
        * Feature relationship preservation
        * Geographic pattern consistency

- **Downstream Validation**:
    - Fraud Detection Performance:
        * Train models on real + synthetic data
        * Compare against real-data-only baselines
        * Evaluate detection improvement

Data Leakage

- **Generation-Specific Leakage Concerns**:
    - Identity Features:
        * `trans_num`: Should not be copied, generate new
        * `cc_num`: Need to generate novel numbers while preserving patterns
        * `merchant`: Maintain distribution without exact copying
    - Temporal Features:
        * `trans_date_trans_time`: Generate new timestamps
        * `unix_time`: Derive from generated timestamps
        * Preserve patterns without copying exact times

- **Pattern Preservation Strategy**:
    - Statistical Patterns:

        * Learn distribution of transaction amounts
        * Capture temporal patterns (time of day, week, month)
        * Maintain geographic relationships
    - Feature Relationships:
        * Preserve merchant-location correlations
        * Maintain realistic transaction timing patterns
        * Keep valid feature value combinations

- **Generation Guidelines**:
    - Identity Generation:
        * Create new, valid identifiers
        * Maintain realistic reuse patterns
        * Avoid exact copying of real values
    - Quality Control:
        * Verify generated samples are novel
        * Ensure pattern preservation
        * Validate feature relationship maintenance

Interpretability

- **Key Visualizable Patterns**:
    - Transaction Amount Patterns:
        * Clear visualization of amount distributions by fraud status
        * Box plots showing fraud transactions have different amount patterns
        * Outlier transactions visible for stakeholder review
    - Category-based Insights:
        * Top 5 categories most susceptible to fraud
        * `grocery_pos` and `shopping_net` showing highest fraud cases
        * Clear ranking for risk assessment
    - Temporal Risk Patterns:
        * Hourly fraud rate variations clearly visible
        * Significant spike in fraud rate during early morning hours
        * Lower risk periods during mid-day hours

- **Stakeholder Communication Value**:
    - Business Insights:
        * Identify high-risk transaction categories

    - Highlight dangerous time periods
    - Quantify amount-based risk patterns
  - Model Validation:
    - Compare synthetic data distributions with these baseline patterns
    - Verify preservation of key fraud characteristics
    - Assess quality of generated samples

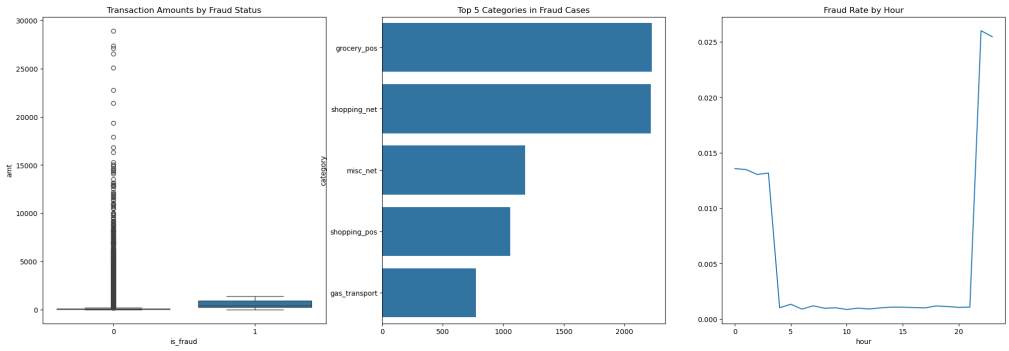

Figure 11: Key Interpretable Patterns: Transaction Amounts by Fraud Status (left), Top 5 Categories in Fraud Cases (center), and Fraud Rate by Hour (right)

Limitations

- **Dataset Constraints**:
  - Class Imbalance:
    - Severe imbalance (0.52% fraud cases)
    - Limited examples of fraudulent patterns
    - Challenges in learning diverse fraud behaviors
  - Temporal Coverage:
    - Limited to specific time period in 2020
    - May not capture seasonal fraud patterns
    - Potentially outdated fraud techniques
  - Feature Scope:
    - Missing modern fraud indicators:
      - Device fingerprinting
      - IP address information
      - User behavioral patterns

        * Limited merchant metadata
        * No transaction velocity features

- **Desired Additional Data**:
    - Technical Features:
        * Device identification data
        * Network/IP information
        * Browser fingerprinting
        * Authentication patterns
    - Behavioral Indicators:
        * Customer shopping patterns
        * Historical transaction velocities
        * Account age and history
        * Previous fraud attempts
    - Enhanced Metadata:
        * Merchant risk scores
        * Category-specific fraud rates
        * Geographic risk indicators