



# EDA Report - Fraud Transactions

---

## Dataset overview

- Rows: 10,000
- Columns: transaction\_id, amount, transaction\_hour, merchant\_category, foreign\_transaction, location\_mismatch, device\_trust\_score, velocity\_last\_24h, cardholder\_age, is\_fraud
- Missing values: none detected
- Fraud rate: 151 / 10,000 (1.51%)

## Key findings

- Amount distribution is heavily right-skewed with a long tail of high-value transactions.
- Most transactions are in categories: Grocery, Electronics, Travel, Clothing, Food.
- Fraud is rare (~1.5%), creating a significant class imbalance.
- Higher velocity (transactions in last 24h) and foreign transactions/location mismatches show higher fraud proportions in exploratory checks.
- Device trust score tends to be lower for fraud cases.

## Recommended analyses and actions

- Feature engineering: rolling-velocity (7/30 day), merchant risk score, geospatial distance between billing and transaction location, time-since-last-transaction, cardholder spending patterns.
- Modeling: use class-weighted algorithms or resampling (SMOTE) plus appropriate cross-validation (stratified). Optimize for business metric (cost-based thresholding).
- Monitoring: implement data-drift and performance monitoring (e.g., Evidently, scheduled checks), set alerts for sudden changes in fraud rate or feature distributions.

## Business impact

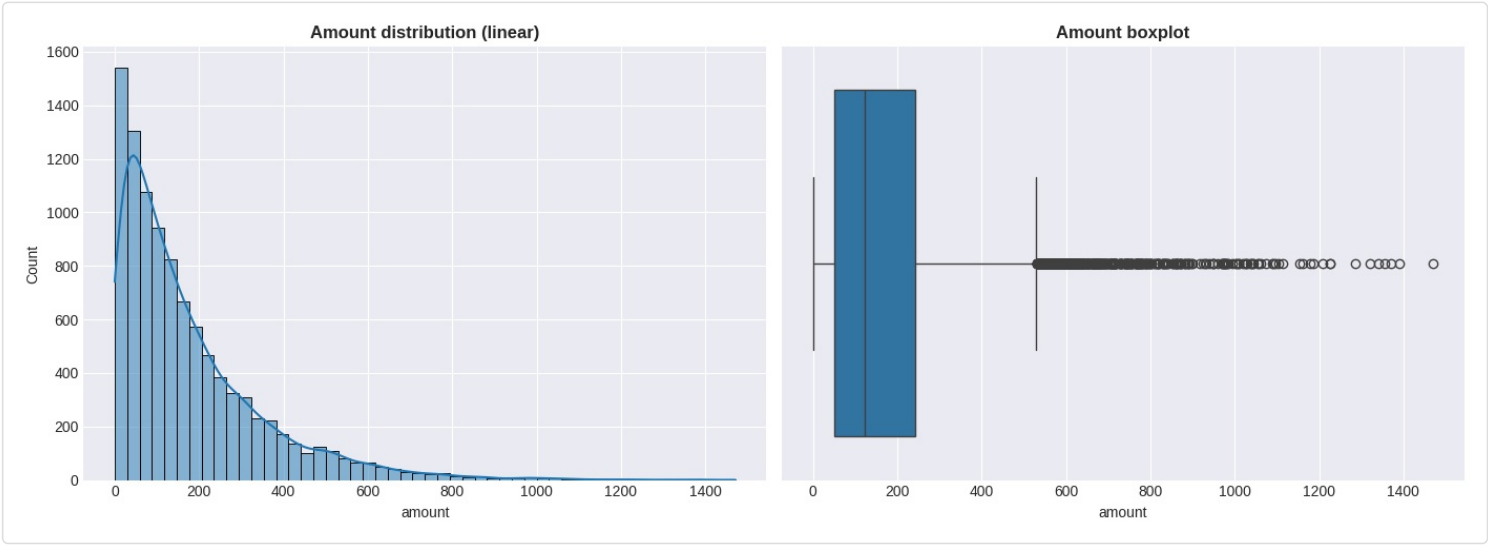
- Improved detection reduces chargeback costs; however high false positives impact customer experience.
- Use cost-sensitive thresholds: weigh the cost of missed fraud vs. customer friction to set operational alerting thresholds.
- Prioritize features that improve precision (device trust, velocity, foreign/location mismatch) in real-time scoring to limit disruption.

## Next steps

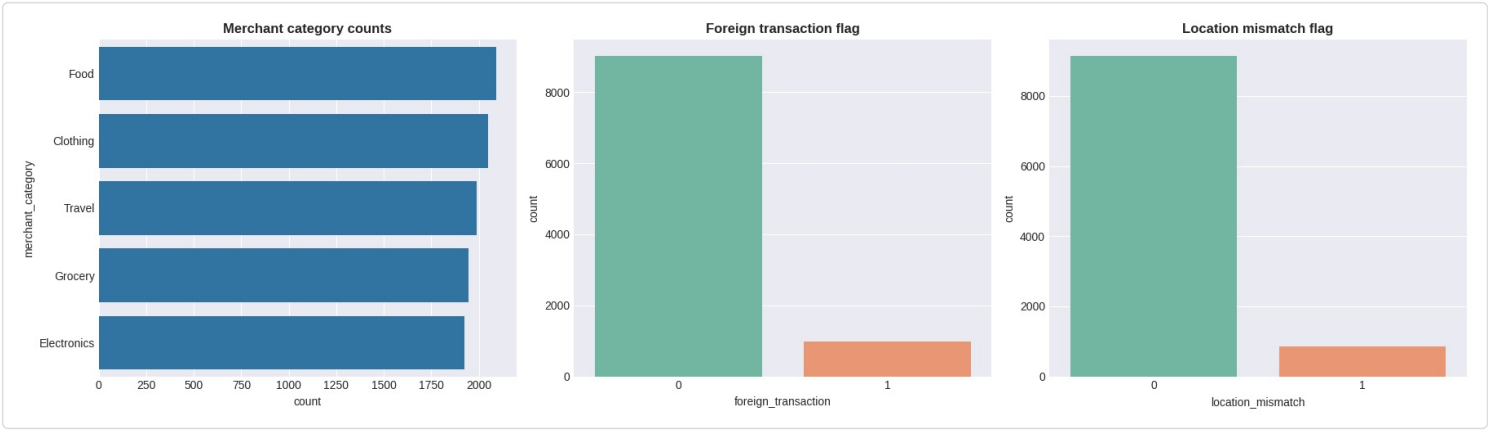
1. Run deeper bivariate analysis and statistical tests for significance.
2. Build a baseline model and compute cost-based metrics.
3. Generate feature importances and SHAP explanations for high-impact features.
4. Implement a monitoring pipeline and automate weekly EDA reports.

## Visualizations

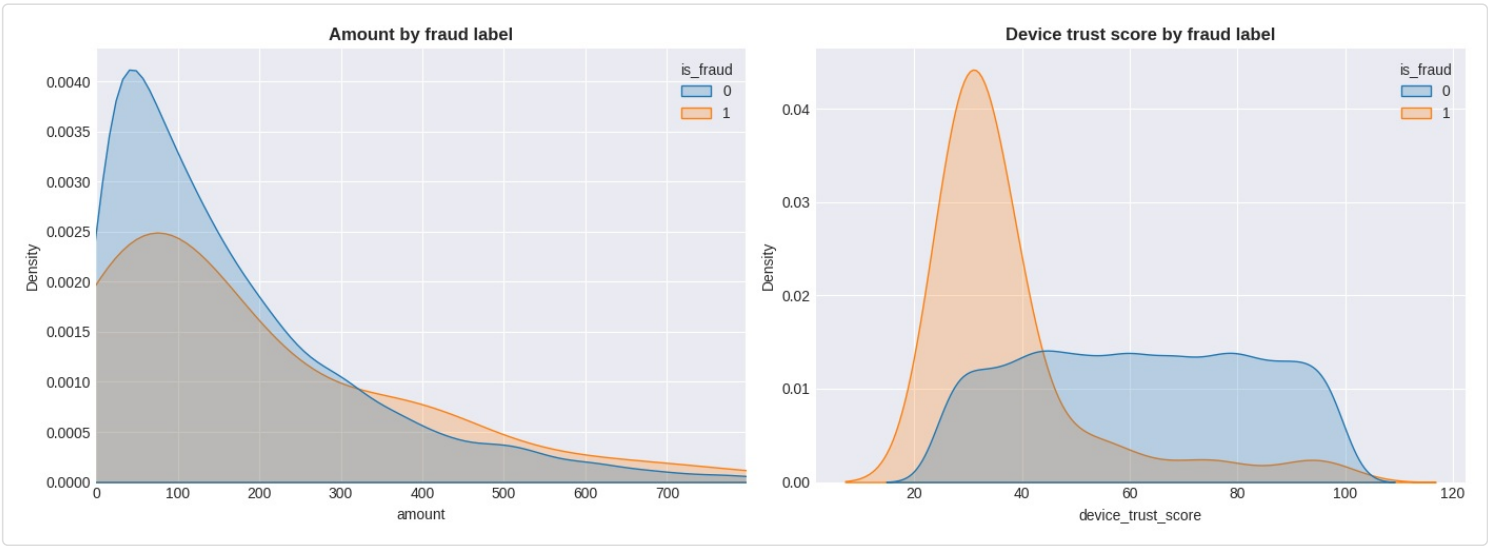
### 01 Amount Dist



02 Categorical Dist

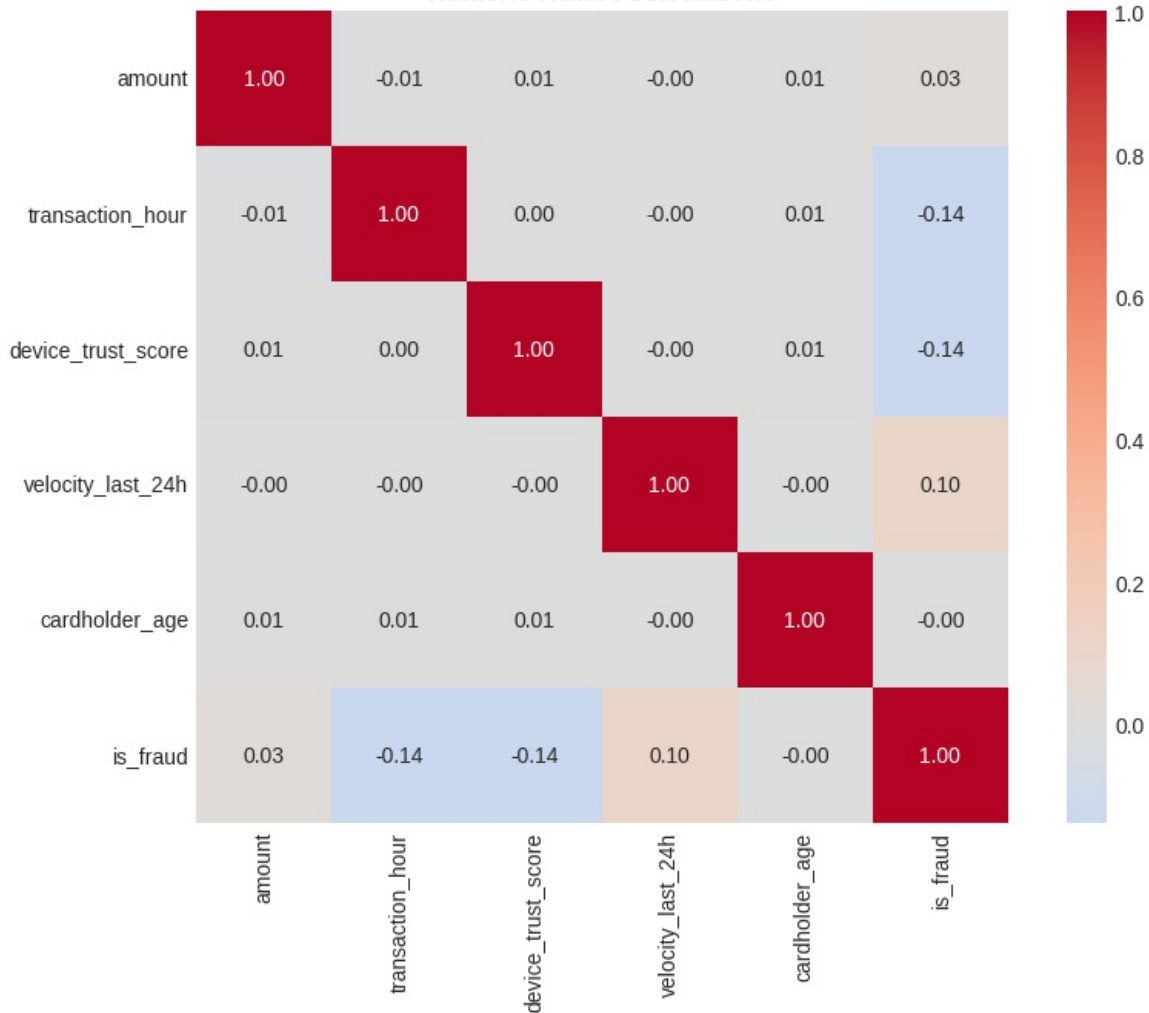


03 Fraud Features

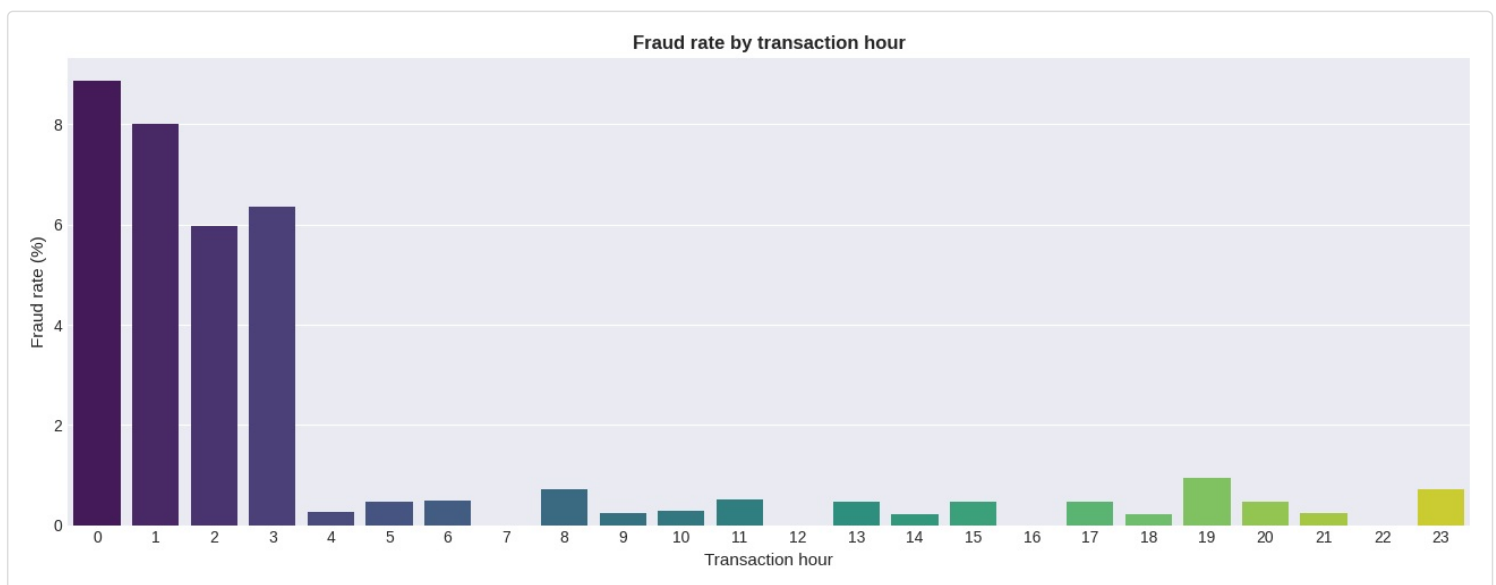


04 Correlation Heatmap

**Numeric feature correlations**

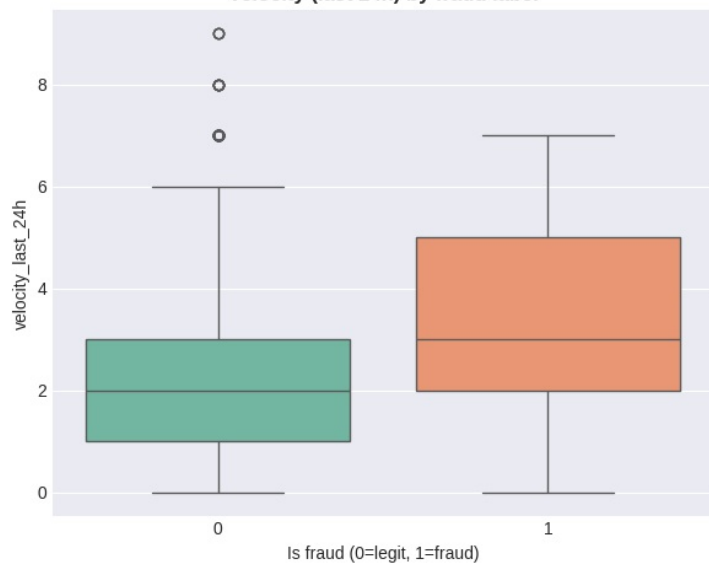


## 05 Hourly Fraud Rate



## 06 Velocity Device Analysis

Velocity (last 24h) by fraud label



Device trust score by fraud label

