PEZESHA DATA SCIENCE ASSESSMENT PRESENTATION OUTLINE

M-PESA TRANSACTION ANALYSIS & CREDIT RISK INSIGHTS

By Dennis Wambua

AGENDA

Part 1: Exploratory Data Analysis (EDA)

- Credit Score Distributions
- Spending Patterns & Risk Factors

Part 2: SQL & User Activity Metrics

- Transaction Trends & Top Users

Part 3: Data Quality & Validation

- Automated Quality Checks (Python)
- Classifier Performance Audit

Recommendations & Next Steps

Data Summary & Quality Overview

Key Statistics

- Users (N=198): Avg Credit Score 605 | Default Rate 27.3% (54 defaults).
- Transactions (N=1,000): Avg $1,445 | Max $47,177 | Min $10.
- Distribution: Highly skewed; median amount ($441) is significantly lower than average.

Data Quality

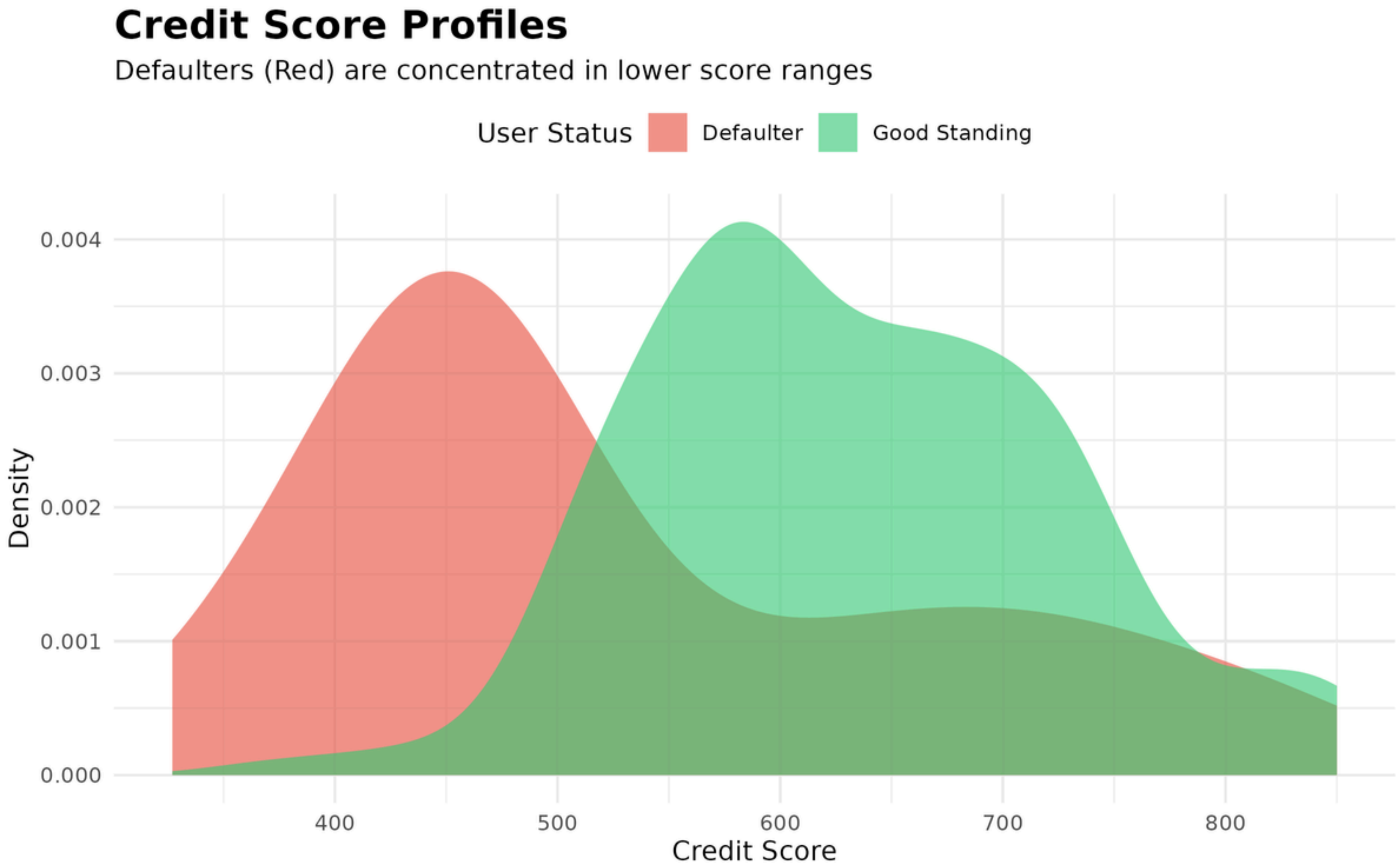
- Integrity: 100% match between Users and Transactions tables.
- Completeness: Users table is 100% complete.
- Missing Data: 14.8% of Transaction category entries are null.

Patterns

- Missing Categories: Randomly distributed across all merchants (not vendor-specific).
- Value Correlation: Unclassified transactions are generally lower value (Avg $1,099) vs classified (Avg $1,505).

DATA OVERVIEW & CREDIT SCORES

Objective: Evaluate the discriminatory power of the current credit scoring model regarding repayment behavior



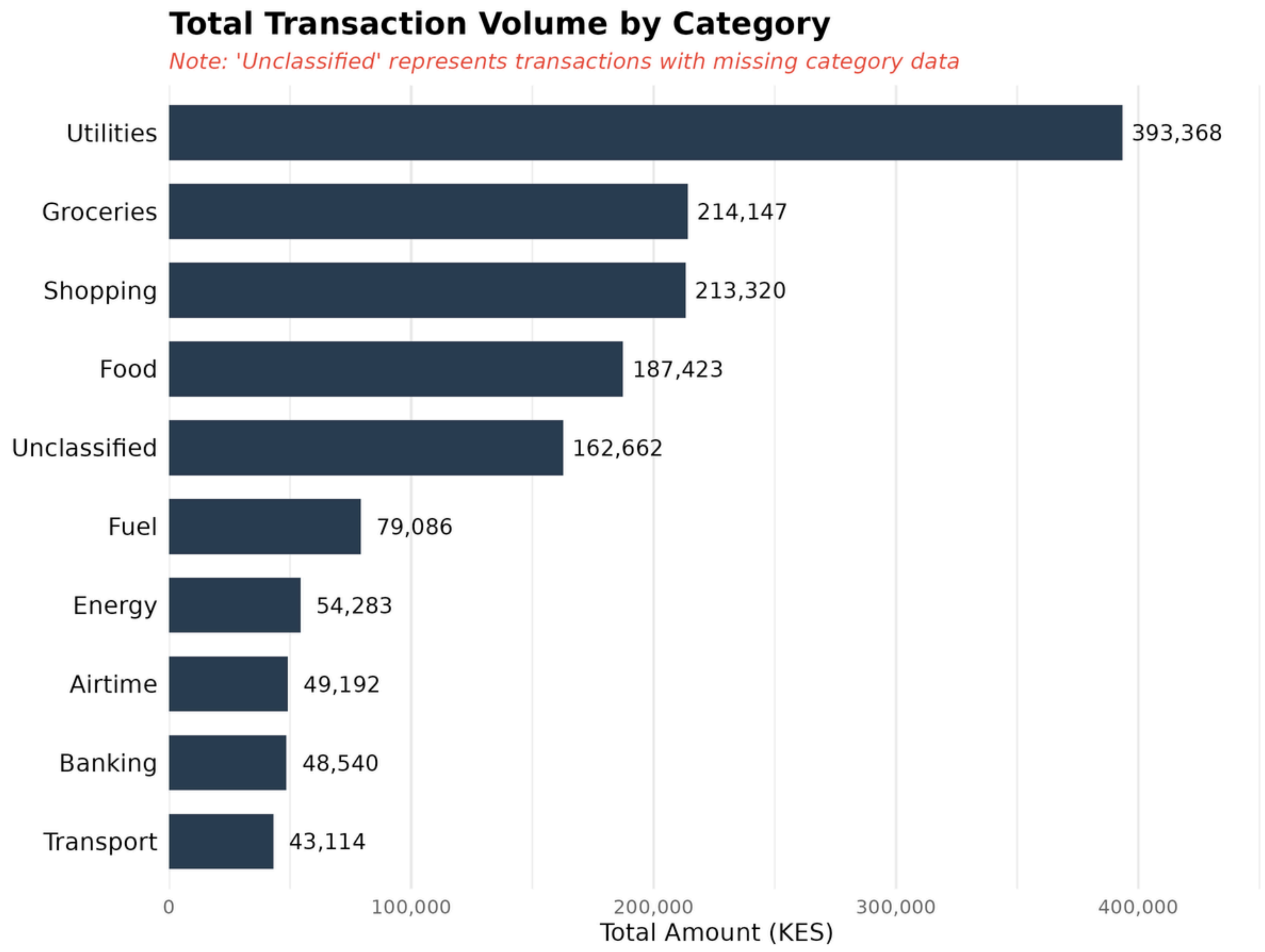
Key Insight: The model demonstrates a clear ability to distinguish risk, with a distinct separation between user groups.

- Good Users (Low Risk): Median score 634. The distribution is concentrated on the right (higher scores), indicating reliable credit history.
- Defaulters (High Risk): Median score 529. The distribution is concentrated on the left (lower scores).

Conclusion: The gap of ~105 points between the medians confirms that Credit Score is a strong predictor of default risk for this portfolio.

SPENDING PATTERNS BY CATEGORY

Objective: Analyze the distribution of user spending (Total Transaction Volume) to identify primary financial commitments.


Total Transaction Volume by Category

Note: 'Unclassified' represents transactions with missing category data

Key Insight:

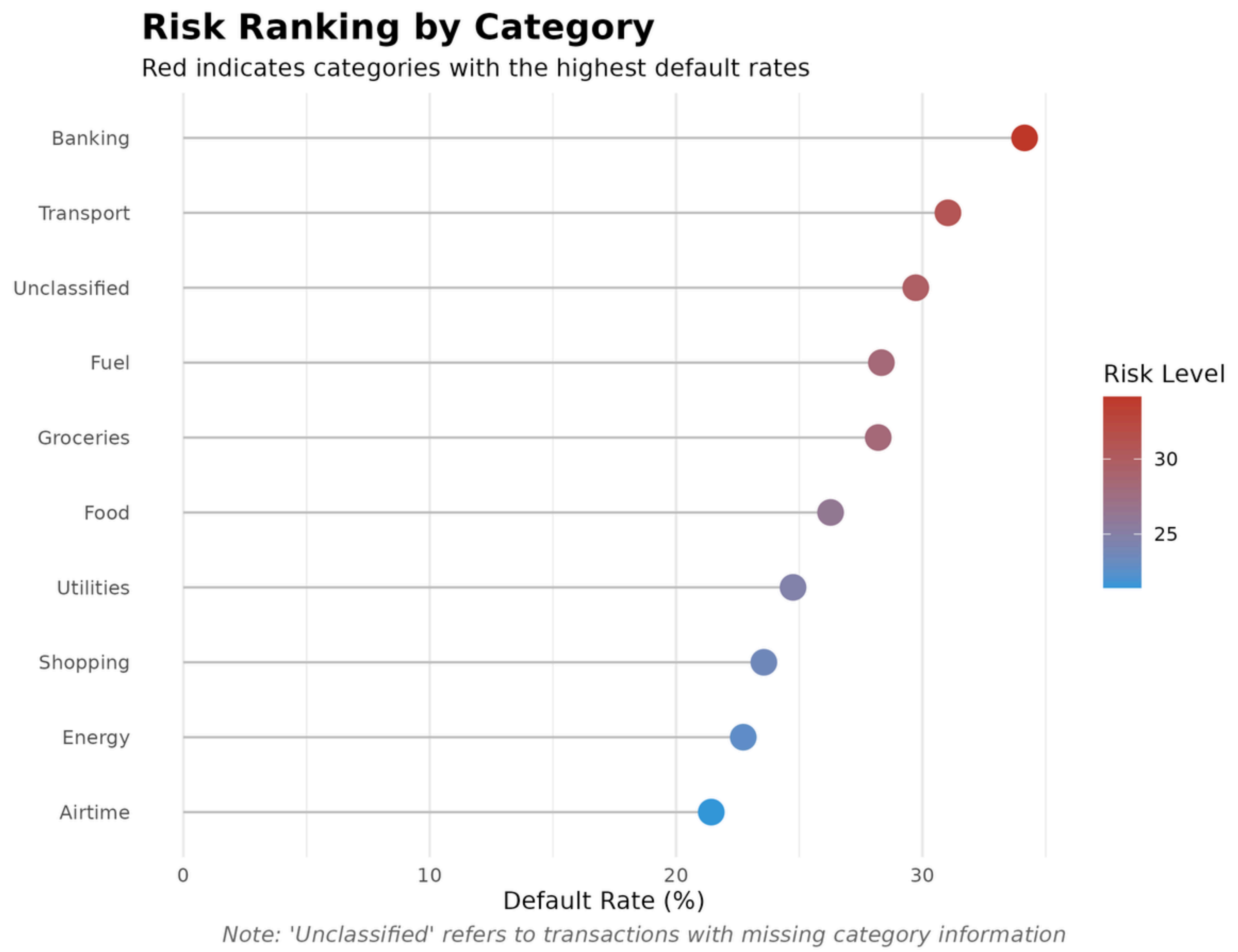
- Essentials Dominate: Utilities (e.g., KPLC, M-KOPA) and Groceries (e.g., Naivas) drive the highest total transaction volume.
- Discretionary Behavior: Airtime and Food show high frequency (transaction counts) but lower total volume compared to essentials.

Relevance:

- High volume in Utilities indicates household stability and reliable bill payment.
- erratic or disproportionate spending in Shopping (e.g., Jumia) relative to income may signal financial impulsiveness.

RISK ANALYSIS (DEFAULT RATES)

Objective: Which spending categories are associated with higher default risk?



Key Insight:

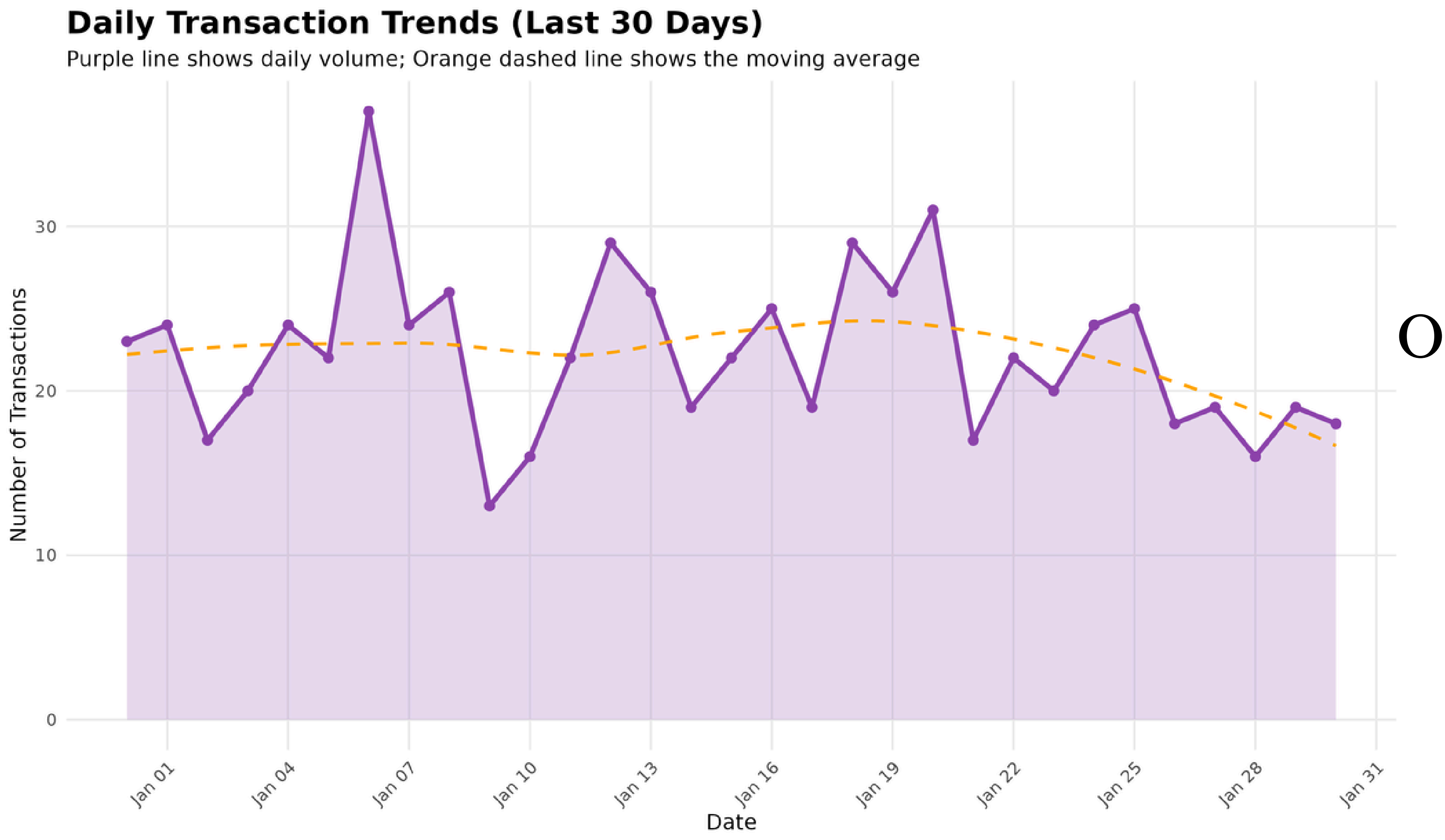
- Highest Risk: Banking (34.1%) & Transport (31.0%). High money movement or commute costs correlate strongly with default.
- Lowest Risk: Airtime (21.4%) & Shopping (23.6%). Surprisingly safe; indicates stable disposable income.

Operational Observation:

- Model Adjustment: Assign negative weight to excessive Banking/Transport frequency.
- Positive Signal: Treat consistent Shopping/ Airtime activity as a reliability indicator.

SQL INSIGHTS – 30 DAILY TRENDS

Objective: Track transaction volume trends over the 30-day period.



Daily Transaction Trends (Last 30 Days)

Purple line shows daily volume; Orange dashed line shows the moving average

Key Insights:

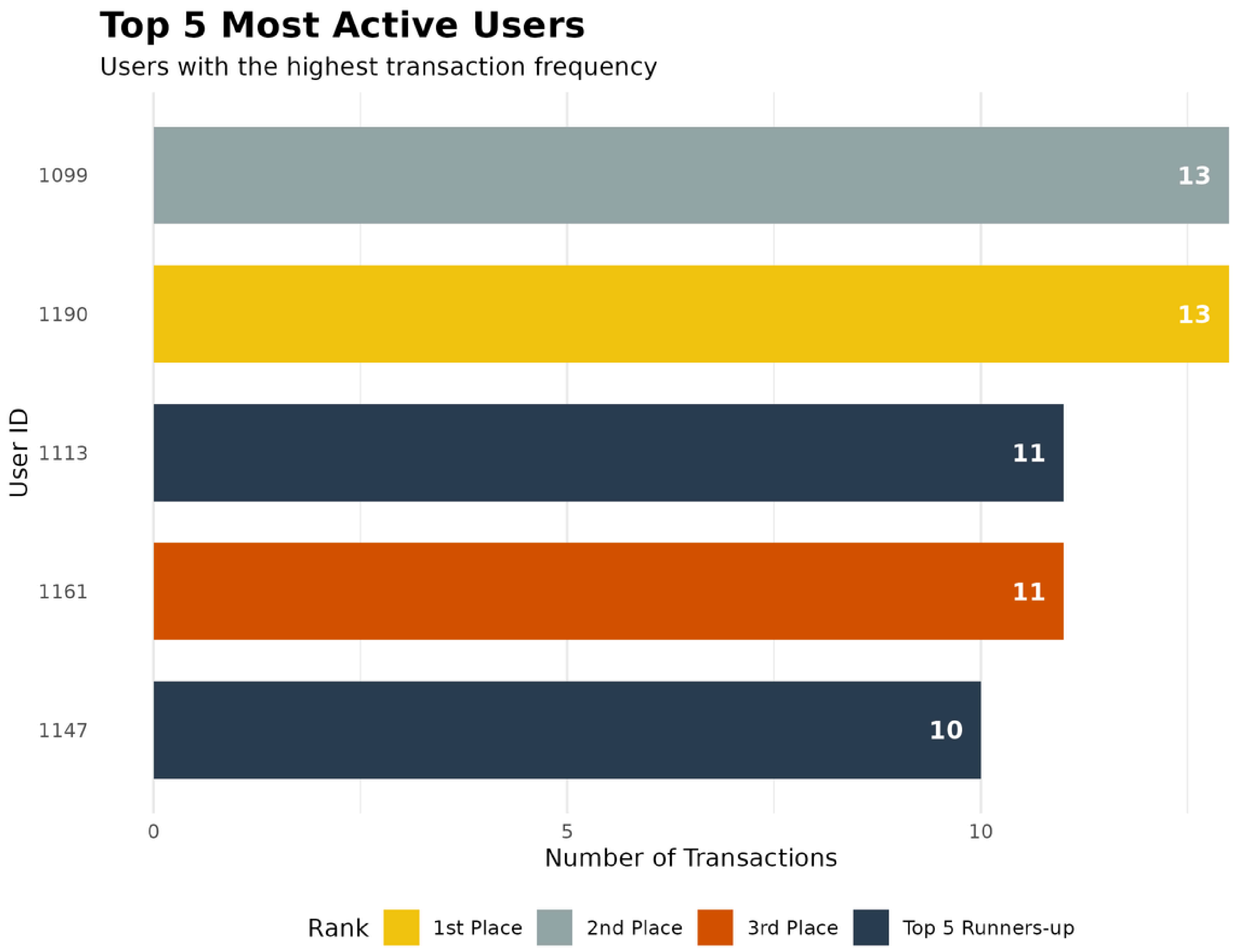
- Early-Week Surge (Mon/Tue): High volume suggests business restocking or bill payments (Work-week focus).
- Weekend Dip: Low activity indicates minimal leisure or entertainment usage.

Operational Signals:

- Staffing: Maximize Support & Liquidity teams on Mon/Tue to handle peaks.
- Strategy: Launch "Weekend Promos" to boost engagement during off-peak gaps.

TOP ACTIVE USERS ("POWER USERS")

Objective: Identify high-frequency users to understand engagement concentration.



Top 5 Most Active Users

Users with the highest transaction frequency

Key Insights:

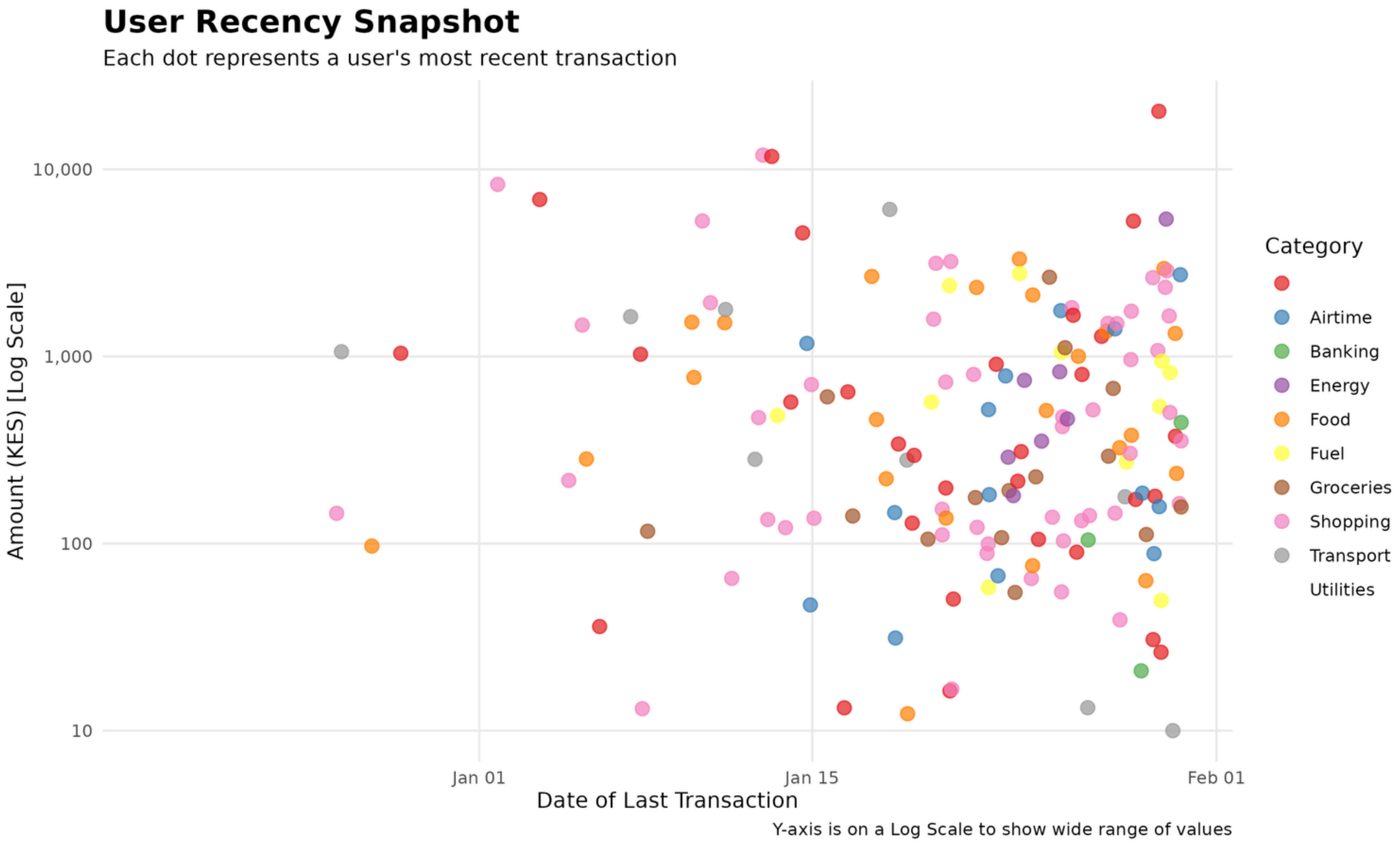
- High Concentration: Activity follows the 80/20 Rule. A small group drives the majority of volume.
- Top Performers: User 1190 (Gold) & User 1099 (Silver) are the most active.
- Data Richness: These users provide the most reliable behavioral data for model calibration.

Strategic Actions:

- Retention: Target top users for Limit Increases or Loyalty Rewards to prevent churn.
- Modeling: Use their stable history as the benchmark for the "Ideal Borrower" profile.

LATEST USER ACTIVITY SNAPSHOT

Objective: Map User Recency vs. Transaction Value to identify churn risks.



User Recency Snapshot

Each dot represents a user's most recent transaction

Y-axis is on a Log Scale to show wide range of values

Key Insights:

- Strong Engagement: High density on the right (late Jan) confirms the majority of users are currently active.
- Churn Risk (Top-Left Quadrant): Distinct outliers show users with high spending power (e.g., >KES 8,000) who have been inactive since early January.
- Operational Signal:
- Immediate Action: Trigger retention campaigns (SMS/Push) specifically targeting these high value, dormant users to bring them back.

DATA QUALITY REPORT

Tool Used: Python (Pandas Validation Pipeline)

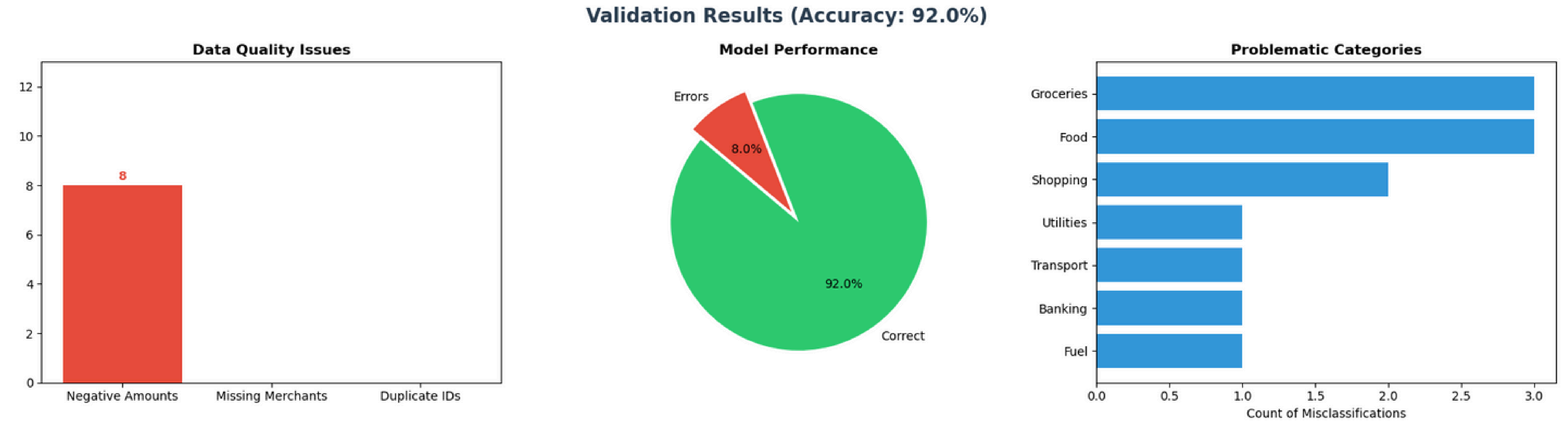
Summary of Checks:

- Critical Issue: Found 8 records with Negative Transaction Amounts.
- Merchant Names: 100% Complete (0 missing).
- Duplicates: No duplicate IDs found.

Action Required: Engineering team must investigate the negative values (likely refund logging errors).

Classifier Performance Audit

- Metrics:
  - Overall Accuracy: 92.0% (Excellent baseline).
  - Error Rate: 8% (12 misclassified records).
- Problem Areas:
  - The model struggles to distinguish Food vs. Groceries.
  - Example: A bakery might be labeled "Food" (Restaurant) but predicted as "Groceries."


Validation Results (Accuracy: 92.0%)

FEATURE ENGINEERING RECOMMENDATIONS

Based on our analysis, recommend adding these 3 features to the Credit Scoring Model:

Money Movement Ratio" (High Risk Indicator)

- Definition: Percentage of transaction volume in Banking & Transport.
- Why: Risk Analysis (Slide 5) proved these categories have the highest default rates (>30%). Frequent movement of funds or high commute costs correlates with financial stress.

"Business Activity Score" (Stability Indicator)

- Definition: Proportion of transactions occurring on Mondays & Tuesdays.
- Why: Daily trends (Slide 6) revealed a "Start-of-Week" surge, characteristic of business owners restocking. These users are statistically more diligent than weekend leisure spenders.

"Disposable Income Signal"

- Definition: Frequency of Shopping & Airtime transactions.
- Why: Contrary to common assumptions, our data showed these "discretionary" categories have the lowest default rates (~21%), likely identifying users with stable surplus income.

CONCLUSION & STRATEGIC RECOMMENDATIONS

Executive Summary

1. Data Integrity: The ecosystem is healthy, but identified a critical engineering bug causing negative transaction values (4% of records) that requires an immediate hotfix.
2. Risk Validation: The Credit Score is a highly effective predictor, with a 105-point separation between Good Borrowers (Median: 634) and Defaulters (Median: 529).
3. Strategic Pivot: Must rethink our risk assumptions. Data proves that Banking & Transport (high mobility/transfers) are the true risk signals, while Shopping & Airtime indicate stability.
4. Model Roadmap: The classifier is strong (92% Accuracy), but "Food vs. Groceries" confusion persists. Recommend adding a Merchant Keyword Layer to solve this.

THANK YOU