Data Quality Report

Section 1 – Data Overview

This dataset contains credit card transaction records, covering the period from **January 1, 2010** to **December 31, 2010** (12 months). The data consists of **98,393 transaction records** across 10 fields, including transaction details (including credit card number, datetime information and transaction type), merchant information and fraud indicators. The dataset appears to track purchases and potentially other transaction types, with approximately **2.53%** of transactions flagged as **fraudulent**.

Section 2 – Field Summary Tables

Categorical Fields:

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
Date	98,393	100.00%	0	365	2/28/10
Merchnum	94,970	96.52%	0	13091	930090121224
Merch description	98,393	100.00%	0	13126	GSA-FSS-ADV
Merch state	97,181	98.77%	0	227	TN
Transtype	98,393	100.00%	0	4	P
Recnum	98,393	100.00%	0	98393	1
Fraud	98,393	100.00%	95,901	2	0 (Not Fraud)
Merch zip	93,664	95.19%	0	4567	38118
Cardnum	98,393	100.00%	0	1645	5142148452

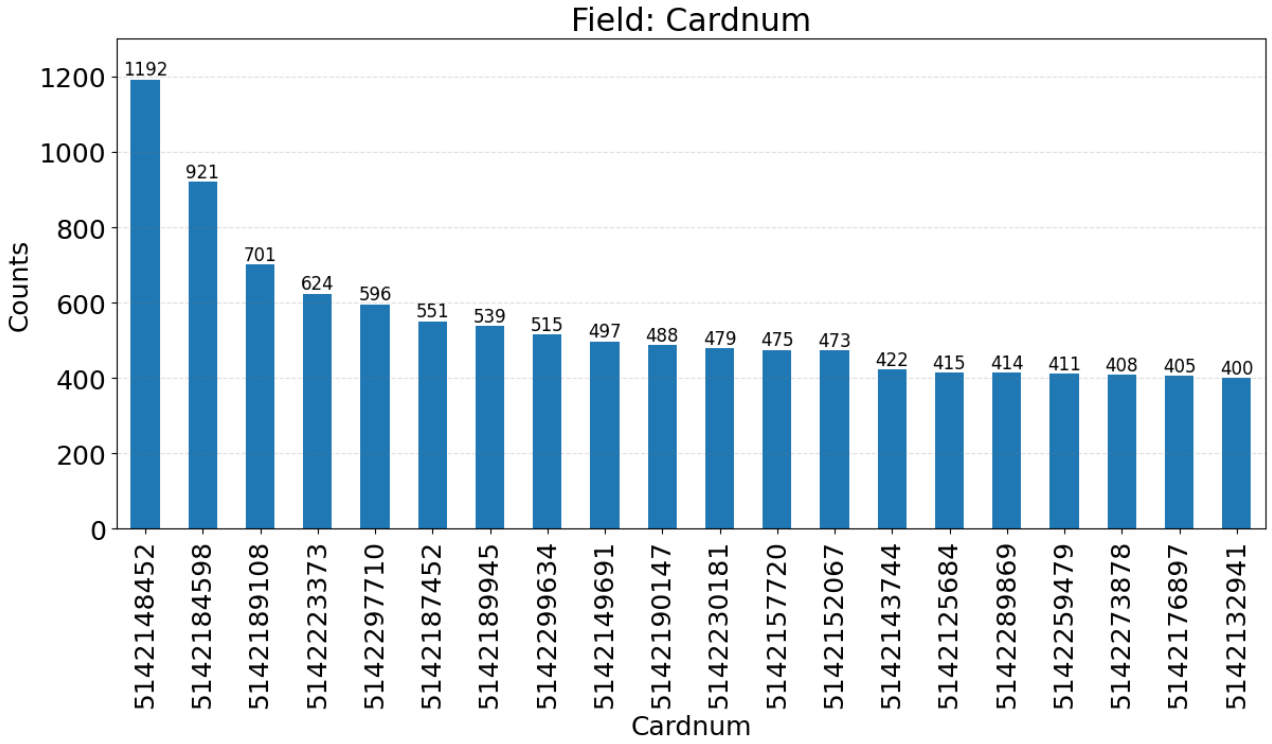
Numeric Fields:

Field Name	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Std. Dev.	Most Common
Amount	98,393	100.00%	0	0.01	3,102,045.53	424.290926	9,922.44	3.62

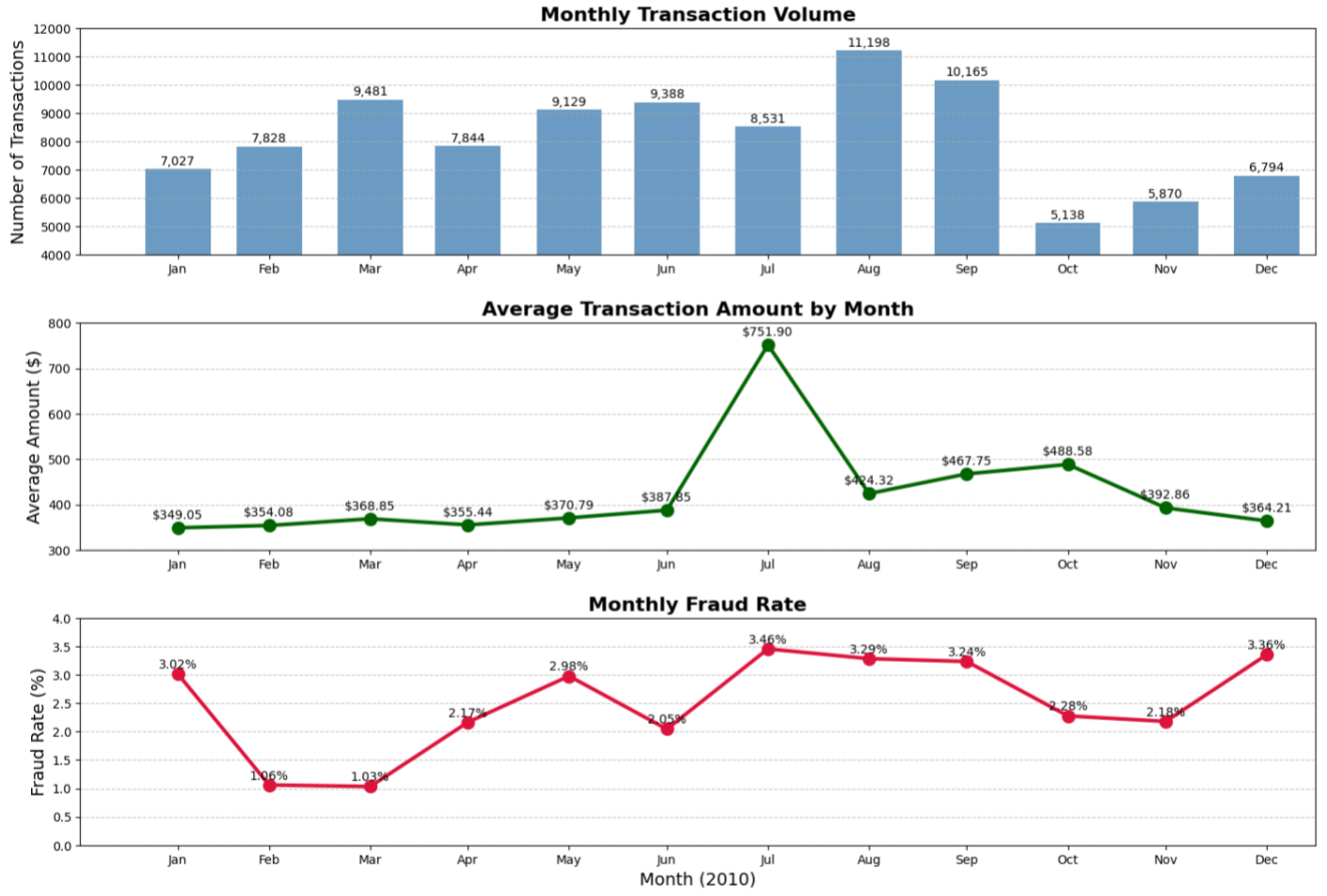
Section 3 – Field Analysis

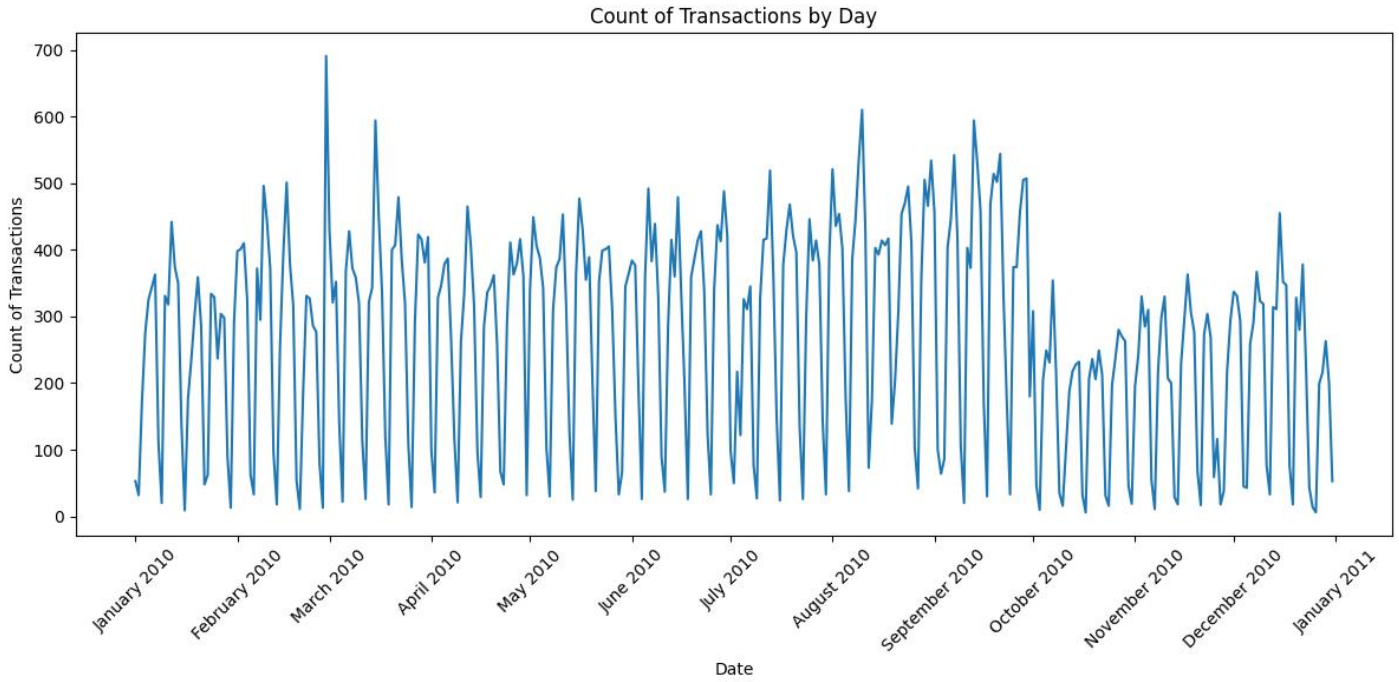
1. **Recnum**: The Recnum field serves as a sequential primary key identifier, ranging from 1 to 98,393, with each record having a unique value. This confirms that each transaction in the dataset is uniquely identifiable with no duplicates. While technically a numeric value, it functions as an index rather than a measurable quantity, making it more appropriate to treat as a categorical field for analytical purposes.

2. **Cardnum:** Credit card numbers are unique identifiers. Although they consist only of digits, you wouldn't add or average them. They should be represented as categorical (or as a string) rather than numeric. The most frequent card (5142148452) appears in multiple transactions, indicating repeated use by the same cardholder. With 1,645 unique card numbers across 98,393 transactions, the average cardholder has approximately 60 transactions in the dataset. This field is crucial for identifying patterns of card usage and potential clustering of fraudulent activities across specific cards.

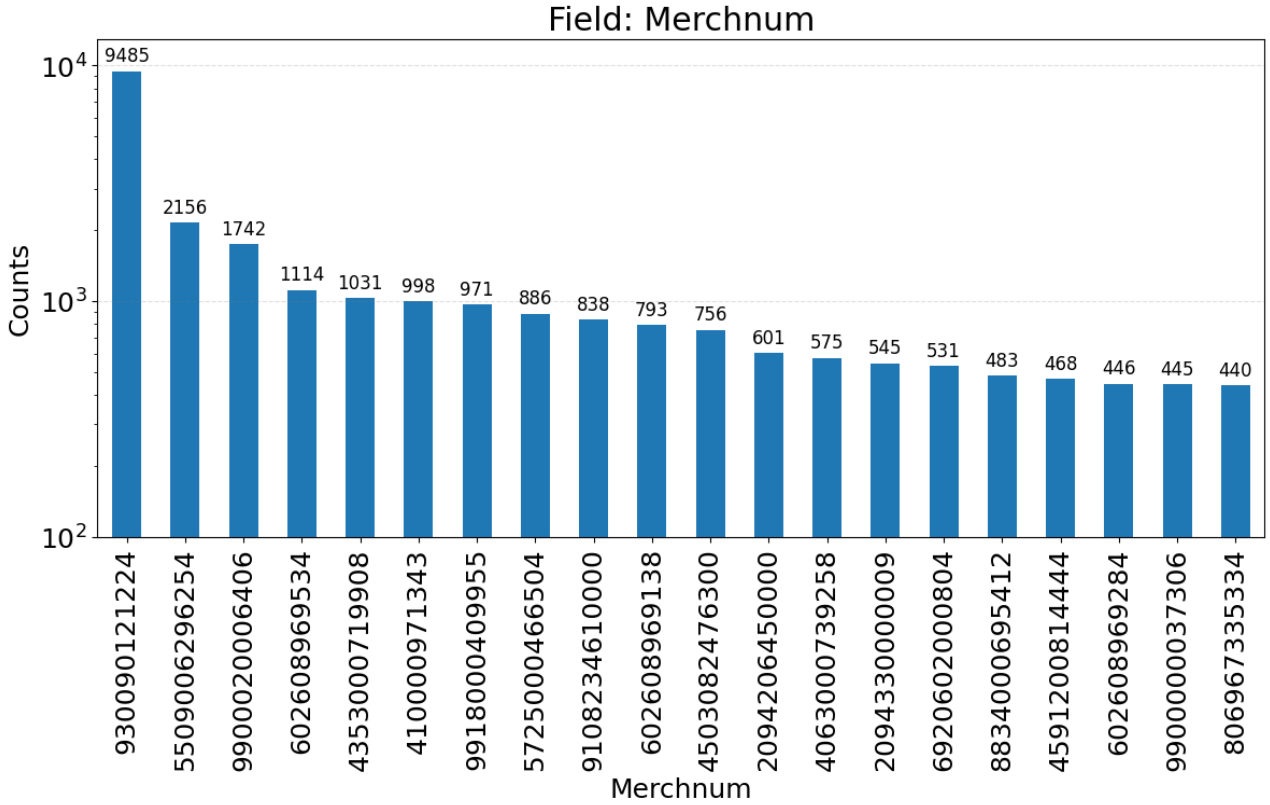


3. **Date:** Dates have an inherent order and can be converted to a numerical timestamp if the analysis requires time series computations. However, dates usually are handled with date-time tools or by extracting components (month, day, year). They are not “numbers” in the arithmetic sense, so we may treat the raw date field as categorical (or, more precisely, as a date variable). The Date field captures transaction dates in MM/DD/YY format, spanning from January 1, 2010, to December 31, 2010. The time-series visualizations reveal distinct seasonal patterns, with August and September showing the highest transaction volumes (11,198 and 10,165 transactions respectively). There's a notable drop in transaction volume during October, which could indicate a seasonal business cycle or data collection anomaly. The average transaction amounts show a dramatic spike in July ($751.90), nearly twice the monthly average, suggesting large seasonal purchases or potentially suspicious activity during this period. The monthly fraud rate fluctuates between 1.03% and 3.89%, with peaks in January, May, July, and December, potentially indicating seasonal fraud patterns around holidays.

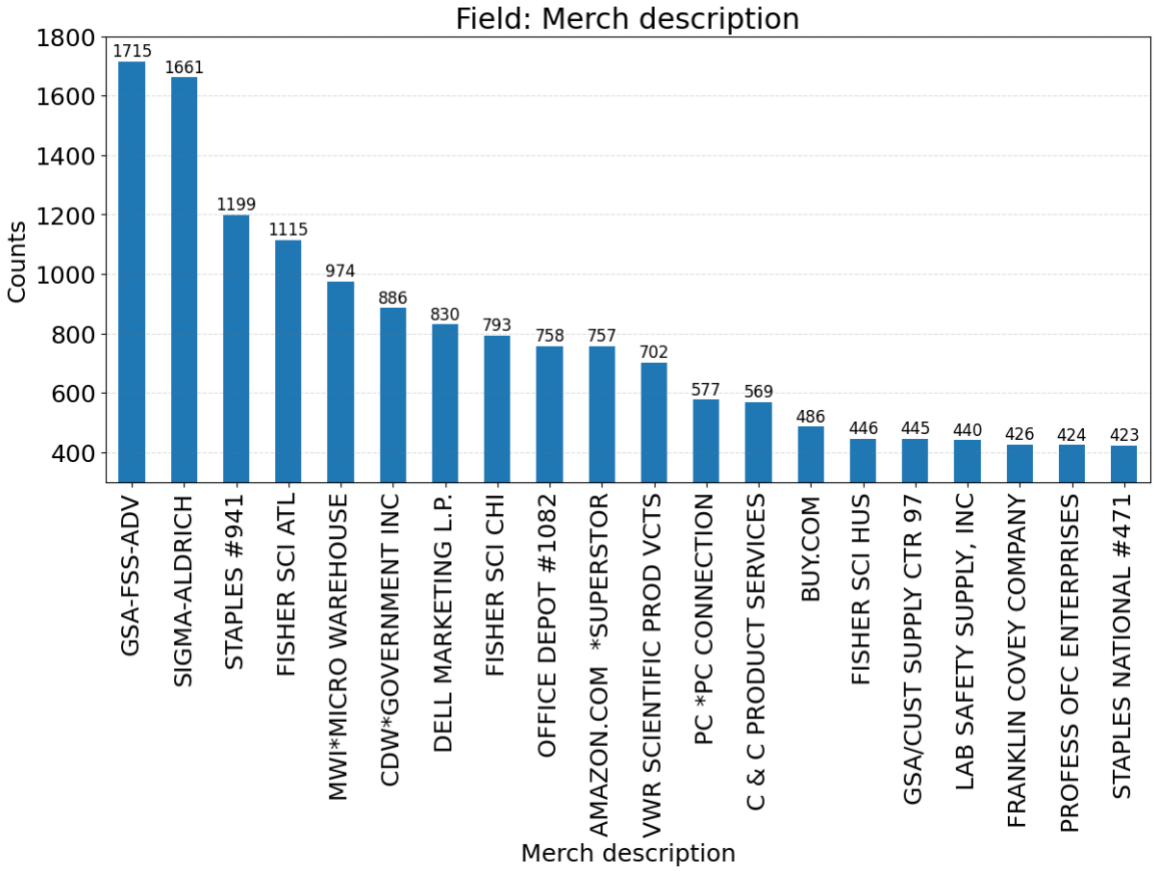




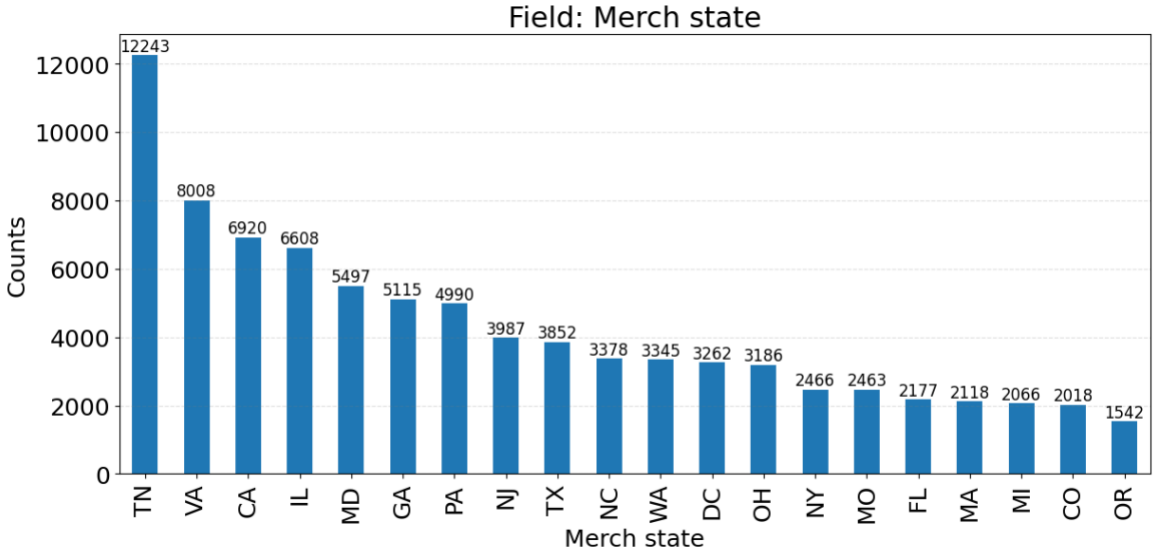
4. **Merchnum:** This field, which likely represents a unique merchant number, is an identifier. Even though it might be stored numerically, the number is a code, not something you would add or average. Thus, it should be treated as categorical. The distribution is highly skewed, with the top merchant (930090121224) accounting for 9,485 transactions (approximately 10% of all transactions). This concentration suggests a dominant relationship with specific merchants. The logarithmic scale in the visualization highlights the long-tail distribution characteristic of merchant activity.



5. **Merch description:** This is a text field that describes the merchant. It is inherently categorical. The Merch description field provides text descriptions of merchants, with 13,126 unique descriptions across all 98,393 records. The most common merchants are "GSA-FSS-ADV" (1,715 transactions) and "SIGMA-ALDRICH" (1,661 transactions). There is a diversity of merchants which provides rich contextual information for understanding transaction patterns.

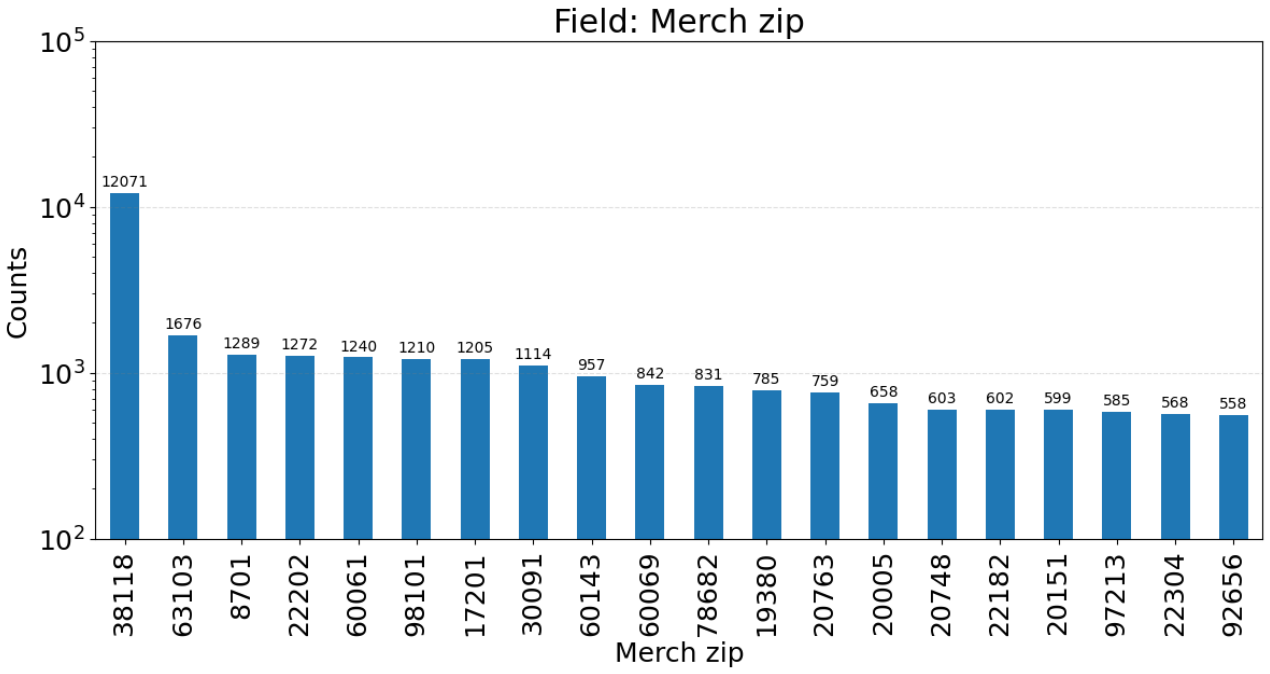


6. **Merch state:** Merch state indicates the U.S. state where merchants are located, with 227 unique values across 97,181 records (98.77% of transactions). Tennessee (TN) dominates the distribution

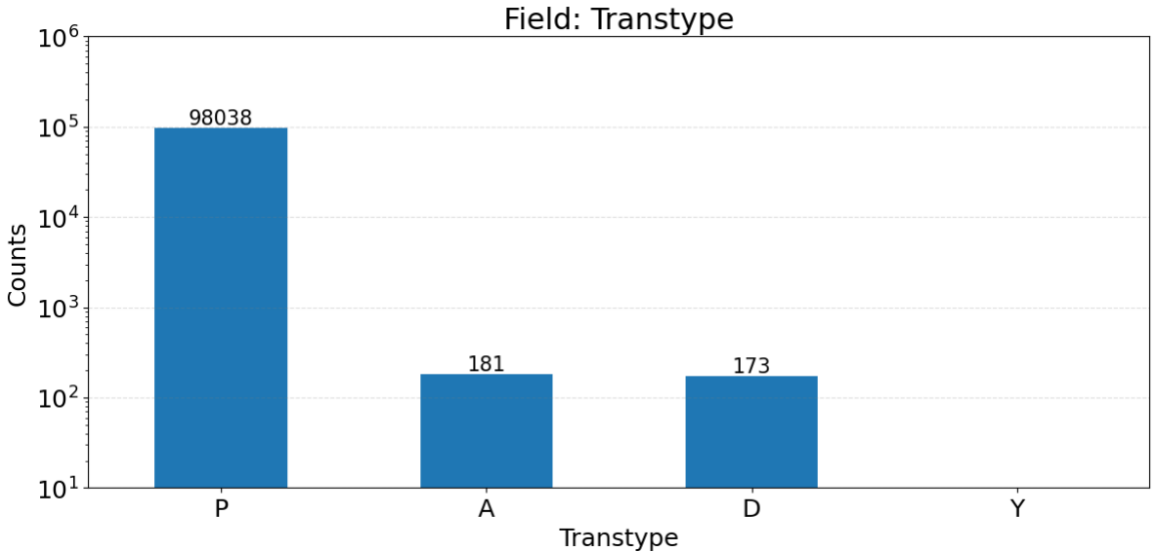


with 12,243 transactions, followed by Virginia (VA) with 8,008 and California (CA) with 6,920. This strong geographic concentration suggests the cardholder organization likely has significant operations or relationships in these states. The presence of 227 unique values (exceeding the 50 U.S. states plus territories) indicates potential data quality issues, such as non-standard state codes or international locations being captured in this field. The geographic distribution could help identify regional fraud patterns or unusual transaction locations that might indicate compromised cards.

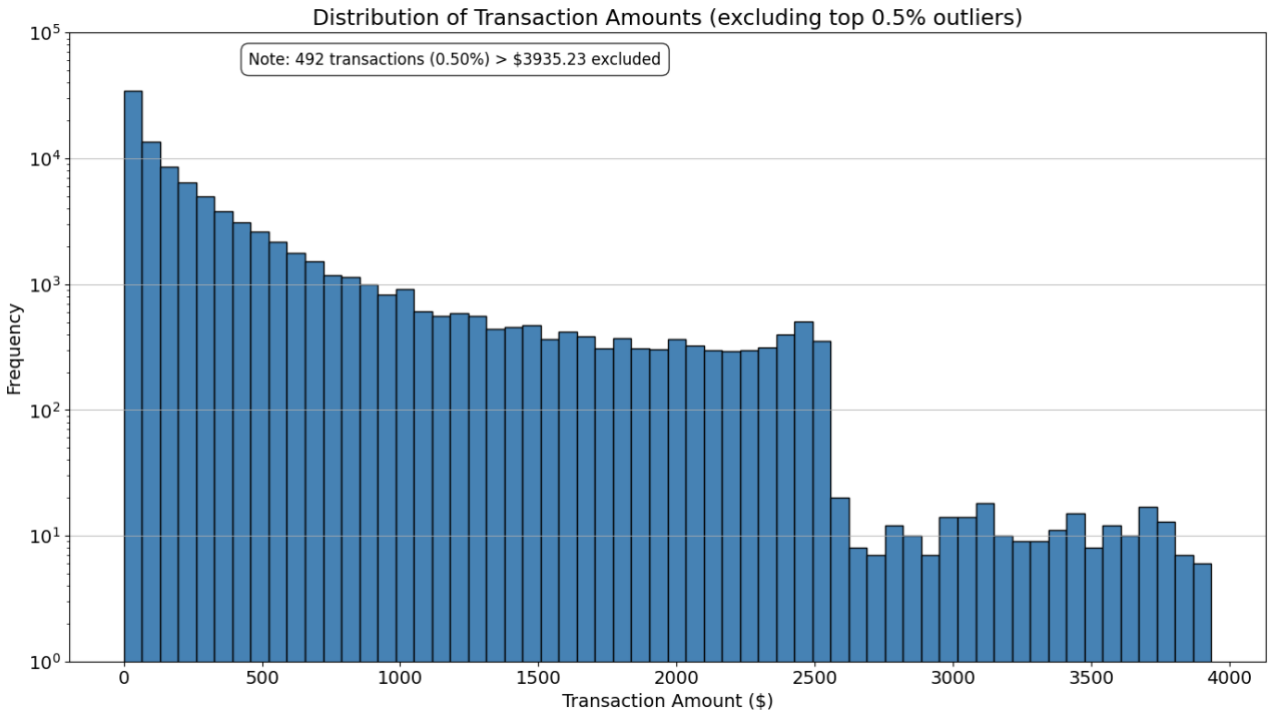
7. **Merch zip:** Merch zip contains ZIP codes for merchant locations, populated in 93,664 records (95.19% of transactions) with 4,567 unique values. The distribution shows extremely high concentration in certain ZIP codes, with 38118 (Memphis, TN) accounting for 12,071 transactions (nearly 13% of all transactions with ZIP codes). This corresponds with the high frequency of Tennessee in the Merch state field. The log-scale visualization reveals clusters of activity in specific regions, providing granular geographic information beyond state-level analysis. The 4.81% of missing ZIP codes, combined with the 1.23% of missing state information, suggests some geographic data quality issues that might impact location-based fraud analysis.



8. **Transtype:** Transtype categorizes the type of transaction with four unique values across all 98,393 records. It is a label and should be treated as categorical. The overwhelming majority (98,038 or 99.6%) are coded as "P" (likely indicating Purchases), with minimal occurrences of "A" (181 transactions) and "D" (173 transactions), and a single "Y" transaction. The extreme imbalance in this distribution indicates the dataset predominantly captures purchase transactions rather than other activities like refunds, cash advances, or balance transfers. Understanding the meaning of these codes is essential for transaction analysis, as different transaction types carry different risk profiles and behavioral patterns. The rare transaction types ("A", "D", and "Y") might require special attention in fraud analysis as unusual patterns.



9. **Amount:** Amount represents the monetary value of each transaction, ranging from $0.01 to $3,102,045.53. It is a continuous value on which calculations (sums, averages, etc.) are meaningful; therefore, it should be numeric. The distribution is heavily right-skewed, with 99.5% of transactions below $3,935.23 (the remaining 0.5% being extreme outliers). The log-scale visualization shows distinct transaction behavior patterns, with a sharp decline after $2,500, potentially reflecting authorization limit policies or natural spending thresholds. The most common transaction amount is $3.62, occurring in numerous transactions and likely representing a standard fee or small purchase. Large transactions significantly skew the average amount upward. The presence of multi-million dollar transactions requires some investigation.



10. **Fraud:** Although Fraud is recorded as 0 or 1, it is a binary indicator (whether the transaction is fraudulent or not). This field is used as the target variable in classification. It can be stored as numeric (binary) for some algorithms, but conceptually it is categorical because the "0" and "1" represent classes rather than measurable quantities. Only 2,492 transactions (2.53%) are flagged as fraudulent, creating a significant class imbalance typical in fraud detection problems. The time-series visualization of fraud counts by day reveals temporal patterns, with several significant fraud spikes in January, July, and August 2010. These spikes could indicate coordinated fraud attacks or periods of enhanced fraud detection activity. Daily fraud counts range from 0 to over 50 transactions, suggesting considerable volatility in fraud occurrence.