# Data Quality Report

#### **High-Level Data Description**

This dataset contains credit card transaction records from a U.S. government organization, likely based in Tennessee. It consists of **98,393 records** and **10 fields**. Each row represents a single transaction, including information such as card number, merchant details, transaction type, date, amount, and a fraud flag. The dataset spans from **January 1, 2010 to December 31, 2010**, covering a full calendar year of transaction activity. The “Fraud” column is a binary indicator, where 1 denotes fraudulent transactions, and 0 denotes legitimate ones.

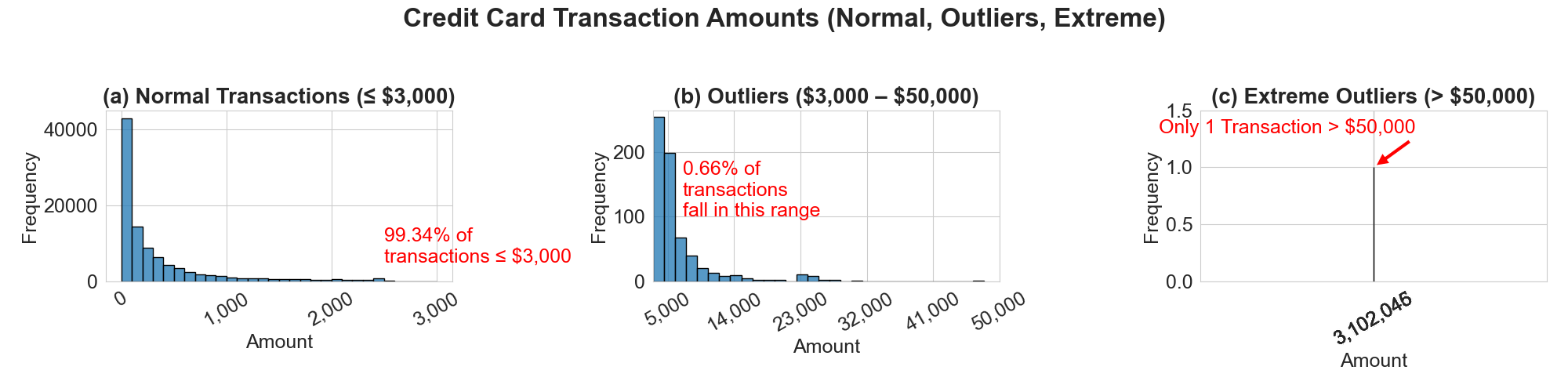
**Numeric Fields Summary**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | Field Type | # Records with Values | % Populated | # Zeros | Min | Max | Standard Deviation | Mean | Most Common |
| Amount | Numeric | 98,393 | 100.00 | 0 | 0.01 | 3,102,046 | 9,922.44 | 424.29 | 3.62 |

**Categorical Fields Summary**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | Field Type | # Records with Values | % Populated | # Zeros | # Unique Values | Most Common |
| Merch description | Categorical | 98,393 | 100.00 | 0 | 13,126 | GSA-FSS-ADV |
| Recnum | Categorical | 98,393 | 100.00 | 0 | 98,393 | 1 |
| Cardnum | Categorical | 98,393 | 100.00 | 0 | 1,645 | 5142148452 |
| Merchnum | Categorical | 94,970 | 96.52 | 0 | 13,091 | 930090121224 |
| Merch state | Categorical | 97,181 | 98.77 | 0 | 227 | TN |
| Date | Categorical | 98,393 | 100.00 | 0 | 365 | 2/28/2010 12:00:00 AM |
| Transtype | Categorical | 98,393 | 100.00 | 0 | 4 | P |
| Fraud | Categorical | 98,393 | 100.00 | 95,901 | 2 | 0 |
| Merch zip | Categorical | 93,664 | 95.19 | 0 | 4,567 | 38118 |

**Distribution Plots**



### **Figure 1. Field: Amount**

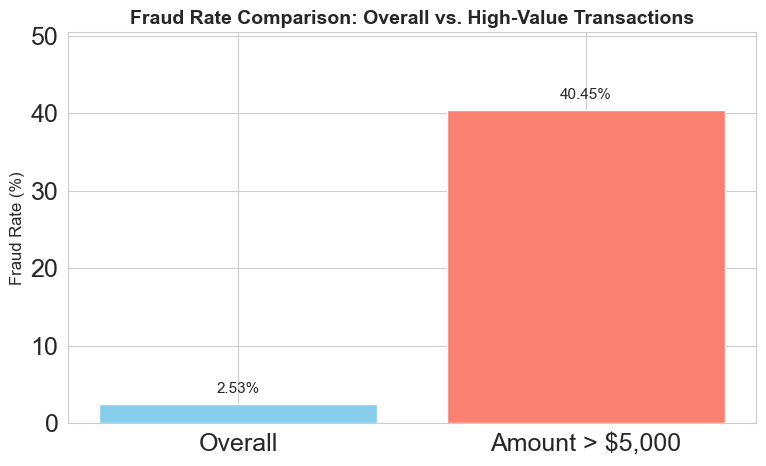
### Distribution of transaction amounts, segmented into normal (≤ $3,000), outlier ($3,000–$50,000), and extreme outlier (> $50,000) transactions.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Recnum | Date | Merch description | Merchnum | Merch state | Merch zip | Transtype | Amount | Fraud |
| 53492 | 7/13/10 | INTERMEXICO |  |  |  | P | $3,102,045.53 | 0 |

**Table 2.** Summary of the transaction with the maximum recorded amount ($3,102,045.53). This transaction occurred in July 2010, was not labeled as fraud, and is missing key merchant fields.

**Outlier Transaction – Field: Amount**  
The transaction with the highest recorded amount ($3,102,046.53) occurred on July 13, 2010 and was associated with the merchant "INTERMEXICO". This transaction was not flagged as fraudulent, but notably, it is missing several key merchant fields (Merchnum, Merch state, and Merch zip). The date falls in mid-summer, consistent with the overall seasonal spike in transaction volume. The incompleteness of this record, despite its unusually high value, may warrant further review.

**Fraud vs. Transaction Amount**  
The overall fraud rate in the dataset is 2.53%. However, among transactions greater than $5,000, the fraud rate jumps to 40.45% — more than 16 times higher. This highlights Amount as a highly predictive feature and suggests that high-value transactions may be subject to tighter scrutiny or automated risk scoring.

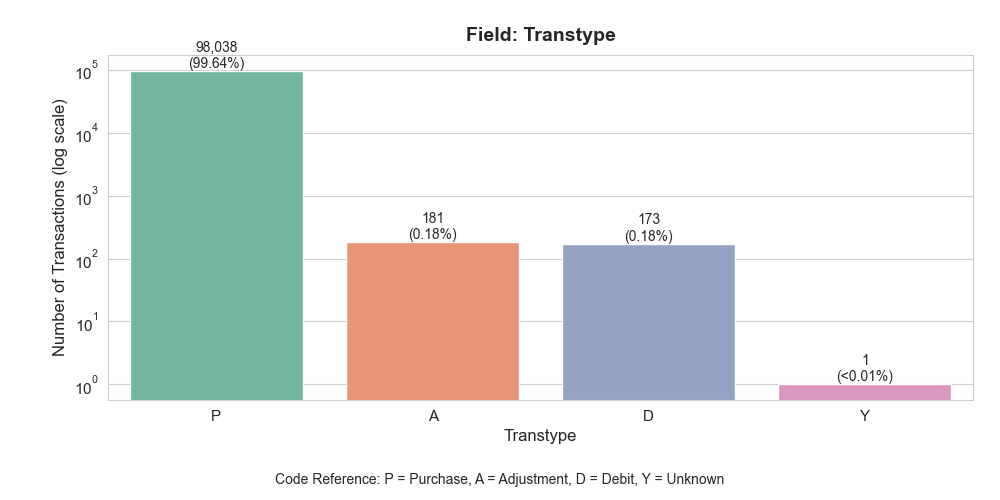


**Figure 2. Fraud Rate Comparison: Overall vs. High-Value Transactions**  
The fraud rate increases sharply for transactions over $5,000, jumping from 2.53% overall to 40.45%. This highlights the Amount field as a highly predictive feature and reinforces the importance of value-based thresholds in fraud modeling.

****

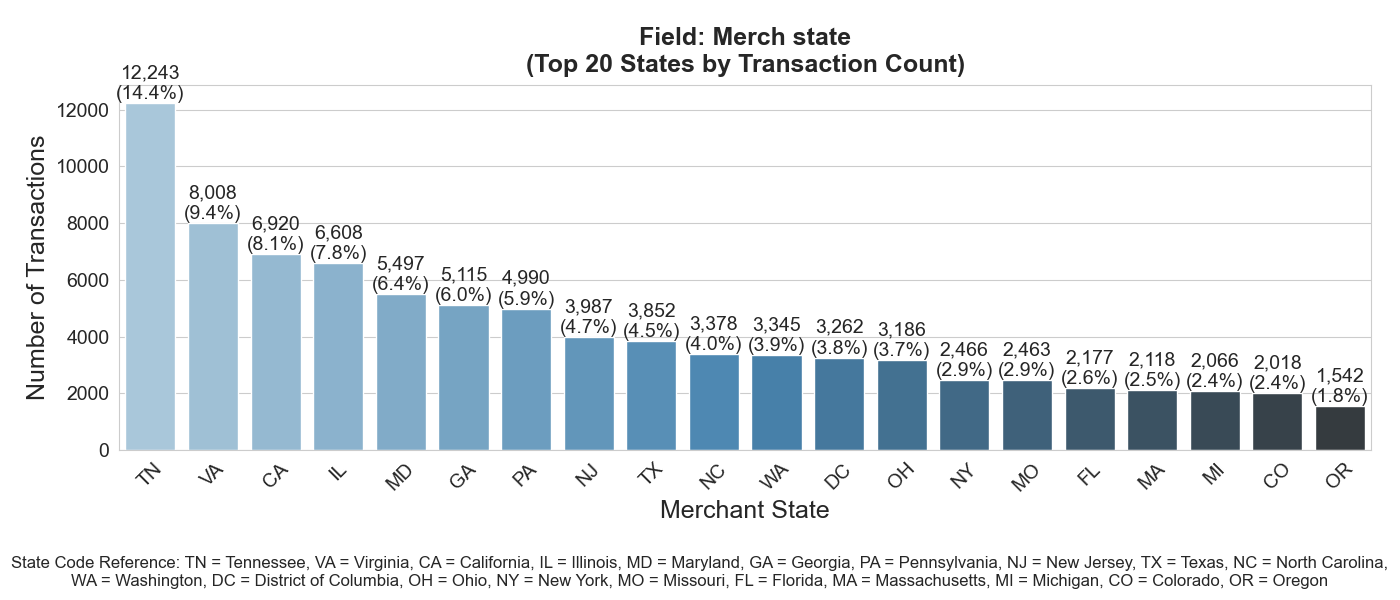
### **Figure 3. Field: Fraud**

### Distribution of the Fraud field, showing heavy class imbalance with very few fraudulent transactions (label = 1).



### **Figure 4. Field: Transtype**

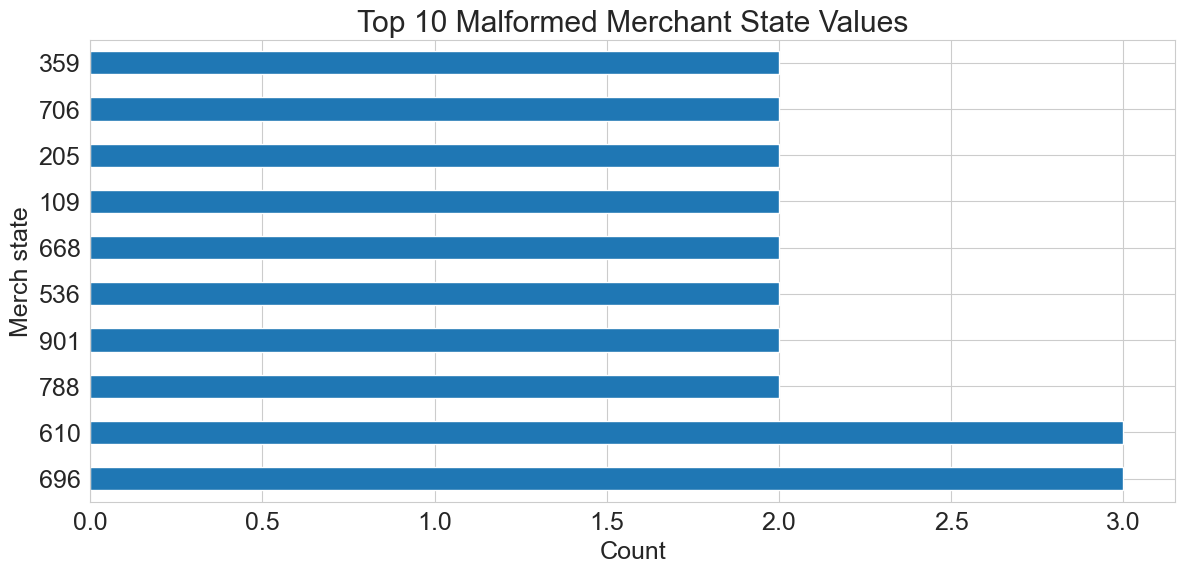
### Distribution of transaction types. Most transactions are type 'P', with very few of other types.

****

### **Figure 5. Field: Merch state**

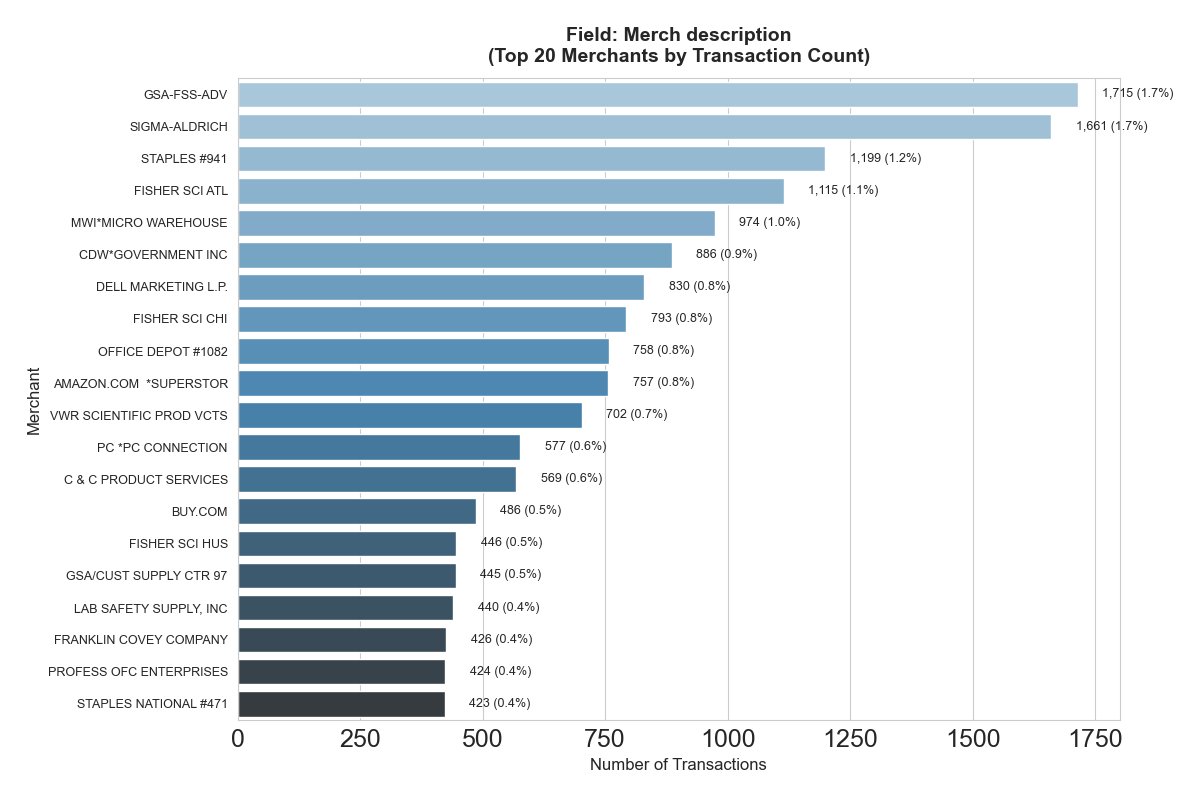
### Top 20 most frequent merchant states, showing concentration in a few regions like TN, VA, and CA.

**Data Quality Note – Field: Merch state**  
The Merch state field contains significant data quality issues. Of the entries identified as malformed (i.e., not valid 2-letter U.S. state codes), **87% are missing**, and the remaining **168 distinct values** include numeric strings such as 696, 610, and 022. These values are not valid state abbreviations and may represent misclassified zip codes or vendor codes. This subset should be cleaned or excluded prior to modeling to ensure the geographic information used is meaningful and reliable.



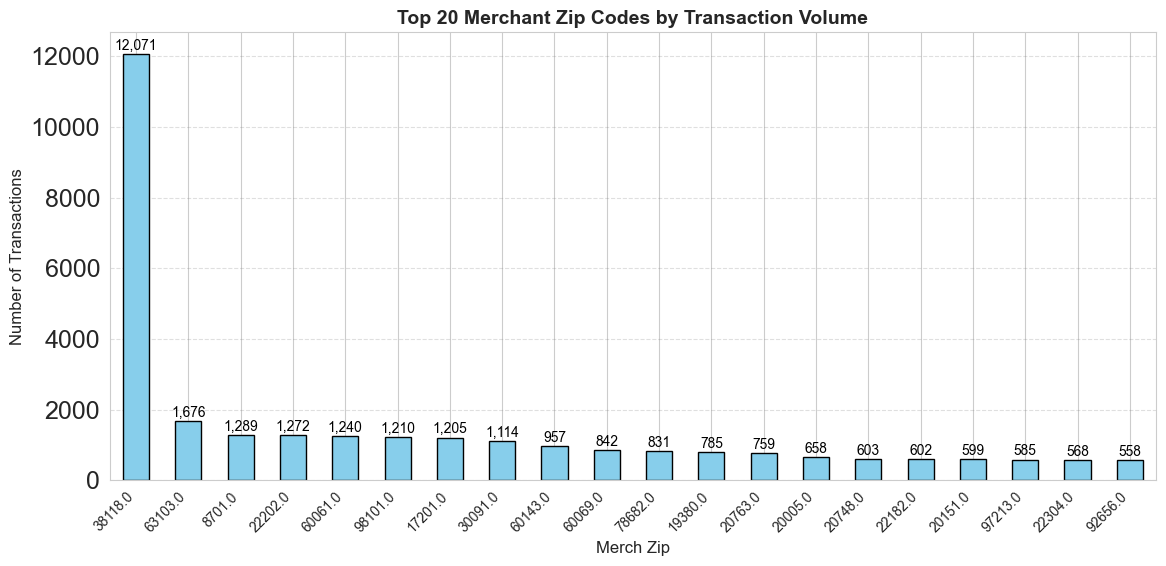
### **Figure 6. Field: Merch state (Malformed Values)**

### Top 10 malformed entries in the Merch state field. These values are not valid U.S. state codes and appear to be numeric placeholders or misclassified fields (e.g., vendor codes or truncated zip codes). This subset contains 168 unique values, 87% of which are missing. Cleaning or exclusion is recommended.

****

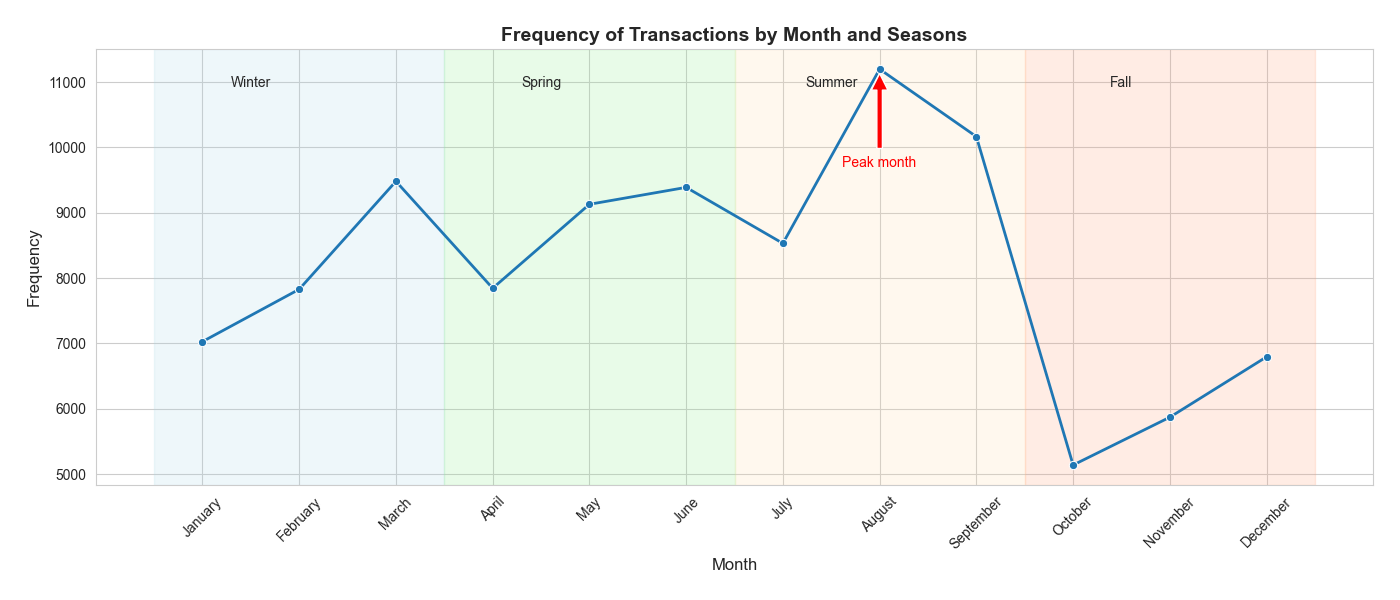
### **Figure 7. Field: Merch description**

### Top 20 merchant descriptions by transaction count. Most transactions are associated with a small number of high-volume vendors.



**Figure 8: Top 20 merchant zip codes by transaction volume.**

**Field: Merch zip**  
The 38118 zip code dominates the transaction volume, with 12,071 transactions, accounting for **12.9%** of the dataset. The remaining zip codes show much smaller transaction volumes, with the next most frequent being 63103 (1.8%). This is likely due to **concentration of vendor activity** or a **centralized processing location** tied to a specific region (e.g., Memphis, TN). This heavy skew should be considered when performing any geographic or transaction pattern analysis, as it may not represent the broader national or regional distribution.

****

### **Figure 9. Field: Date**

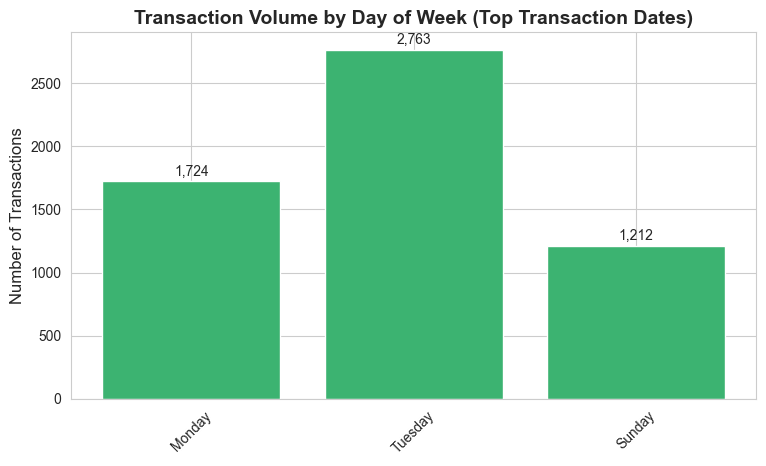
### Monthly frequency of transactions with seasonal grouping. Transaction volume peaks in August, consistent with end-of-fiscal-year procurement cycles.

The dataset shows **continuous daily transaction activity** with **no gaps**. This suggests that the system records at least one transaction per day, potentially due to automated charges, system pings, or a very active account base. Transactions are recorded on **all seven days of the week**, indicating continuous system operation, including weekends. This suggests a mix of automated and business-as-usual processes, with no inactive days across the dataset.

**Time Component – Field: Date**  
All transactions are timestamped at 00:00:00, indicating that time-of-day data is unavailable. This suggests that either the time was not recorded or was stripped/reset during data processing. As a result, temporal analysis in this dataset is limited to the date level only.

**Unusual Concentration – Field: Date**  
The most common transaction date is **February 28, 2010**, with 691 transactions — approximately **0.70% of all records**. While not far above other high-volume days (e.g., August 10 with 610), the consistency of vendors on this date (e.g., FedEx, Sigma-Aldrich) and the low fraud rate (1%) suggest a potential **batch posting** or **end-of-month reconciliation**. This date may act as a placeholder or represent grouped transactions processed on the same day.

**Day-of-Week Pattern – Field: Date**  
Among the top transaction days, a clear pattern emerges: **Tuesdays** are the most frequent transaction day (2,763 records), followed by **Mondays** and **Sundays**.



**Figure 10. Transaction Volume by Day of Week (Top 10 Transaction Dates)**  
Among the top transaction days, most fall on Mondays and Tuesdays. This suggests that transaction activity is concentrated early in the workweek, likely due to batch processing or scheduled uploads following weekends.

**Excluded Fields**

The following fields were excluded from distribution plots due to their nature:

* Recnum, Cardnum, and Merchnum: Identifier fields with high cardinality or all-unique values.