Project 1 Report

Prepared By
NUJOUM UNUS

Table of Contents

Executive Summary

We evaluated a year of credit-card activity for a U.S. government program and built a screening tool that scores each purchase for fraud risk. Because only a small fraction of transactions are truly fraudulent, the goal is to stop losses without burdening legitimate cardholders.

By automatically blocking roughly the riskiest five percent of transactions, the agency is projected to save about **$54 million per year** while keeping customer disruption to a minimum. This threshold captures nearly all of the potential savings available from tighter controls, yet avoids the uptick in false alarms that would come from more aggressive cut-offs.

Description of the Data

This dataset contains **98,393 credit card transactions** recorded by a U.S. government entity based in Tennessee. It spans from January 1, 2010 to December 31, 2010, covering a full calendar year of transaction activity. Each row corresponds to a single transaction and includes fields such as card number, merchant information, transaction type, timestamp, amount, and a fraud flag (binary target). The fraud rate is extremely low at **2.53%**, indicating a heavily imbalanced classification problem.

The dataset includes:

- **10 core fields** in the raw file.
- **Cardnum**, **Merchnum**, **Merch zip**, and **Merch state** are critical categorical fields often used in behavior modeling.
- **Amount** is a key numeric field, ranging from $0.01 to over $3 million, though the highest was flagged and excluded as an outlier.

Key Field Distributions:

- **Amount**: Most values are under $1,000; however, a heavy tail exists. Transactions above $5,000 have a **40.45% fraud rate**.
- **Fraud**: Class imbalance is extreme; only ~2,492 transactions are labeled as fraudulent.
- **Merch State**: Highly skewed with malformed or missing values (e.g., “696”, “022”) — later cleaned via hierarchical mapping.

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

Some Distribution Plots

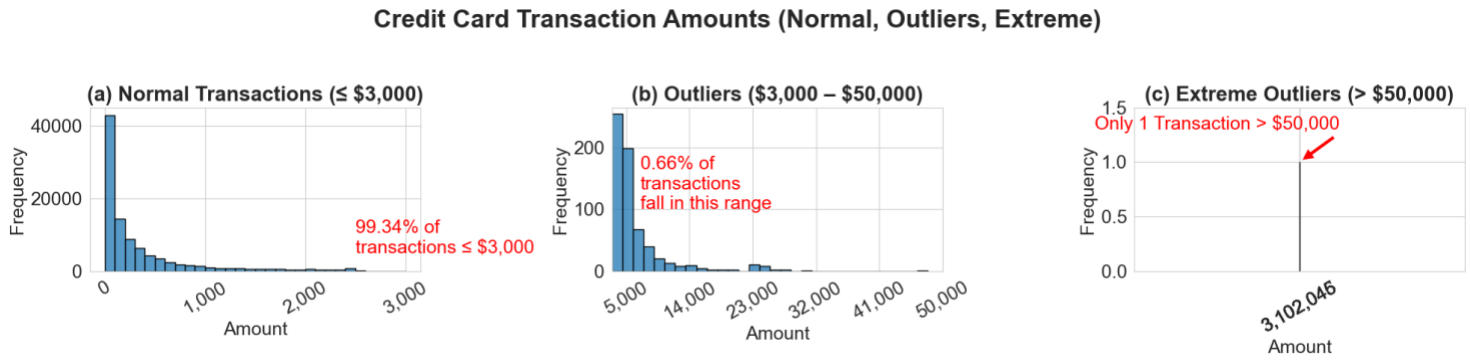
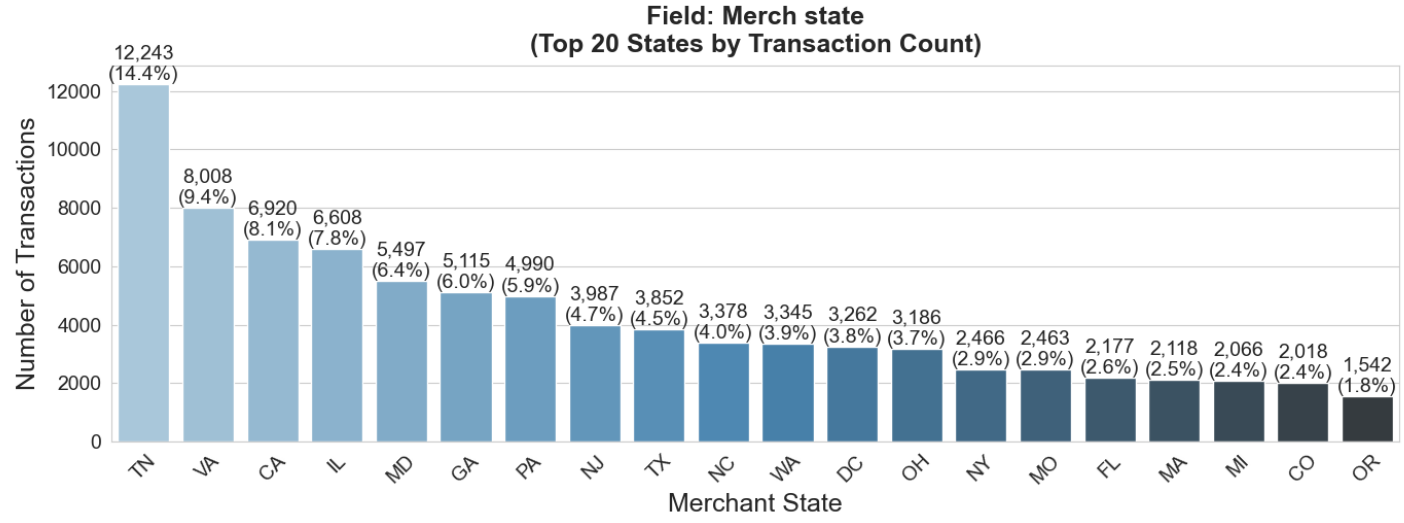


Figure 1. Field: Amount
Distribution of transaction amounts, segmented into normal (≤ $3,000), outlier ($3,000–$50,000), and extreme outlier (> $50,000) transactions.



State Code Reference: TN = Tennessee, VA = Virginia, CA = California, IL = Illinois, MD = Maryland, GA = Georgia, PA = Pennsylvania, NJ = New Jersey, TX = Texas, NC = North Carolina, WA = Washington, DC = District of Columbia, OH = Ohio, NY = New York, MO = Missouri, FL = Florida, MA = Massachusetts, MI = Michigan, CO = Colorado, OR = Oregon

Figure 2. Field: Merch state
Top 20 most frequent merchant states, showing concentration in a few regions like TN, VA, and CA.

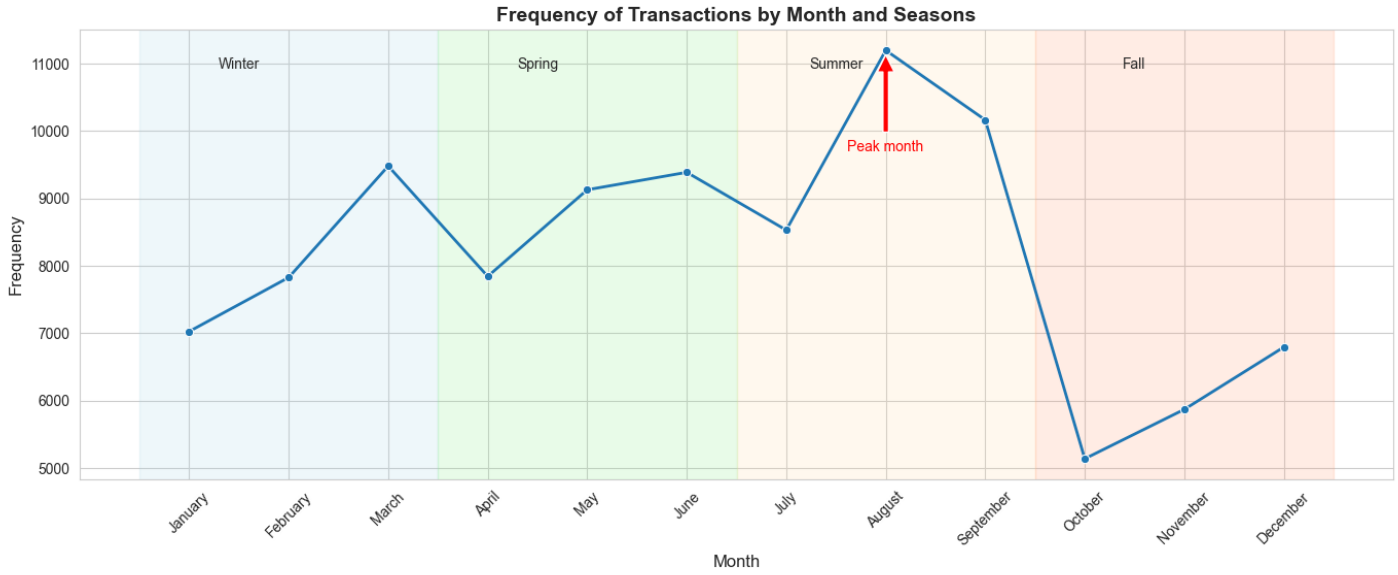


Figure 3. Field: Date
Monthly frequency of transactions with seasonal grouping. Transaction volume peaks in August, consistent with end-of-fiscal-year procurement cycles.

Data Cleaning

1. Outlier Handling

- **Extreme Amount Filtering**: Removed 1 transaction with `Amount > 3,000,000` (non-fraudulent Mexican retail purchase in pesos).
- **FedEx Microtransactions**: Flagged 7,592 transactions from FEDEX with amounts between $3.62–$3.80 using `amount_okay` but retained them in the dataset.

2. Exclusions

- **Transaction Type**: Excluded 355 non-purchase transactions (types A, D, Y), retaining only type P (97,497 transactions).

3. Imputation

Merchnum (Merchant Number)

- Replaced '0' with `NaN` and imputed 1,164 missing values via `Merch description` mappings.
- Generated new unique IDs for 515 unmatched descriptions (e.g., "RETAIL CREDIT ADJUSTMENT" → "unknown").

Merch State

- Filled missing states hierarchically:
    a. `Merch zip` mappings (primary)
    b. `Merchnum`/`Merch description` mappings (secondary)
    c. Categorized non-U.S. states as "foreign" and unresolved cases as "unknown".

Merch Zip

- Imputed missing zips using:
    a. `Merchnum`/`Merch description` mappings
    b. Most populous zip code for the state if unresolved
    c. Defaulted to "unknown" for adjustments.

Variable Creation

The final dataset includes **~4,400 engineered variables**, derived from domain knowledge, temporal logic, and behavioral patterns. These were created to capture short-term anomalies and long-term patterns across cards, merchants, states, and zip codes.

Variable Group	# Variables	Description	Why It Matters in Fraud Detection
Target Encoding (TE)	3	Smoothed fraud risk scores for Merch state, Merch zip, and Dow using historical fraud rates.	Converts past fraud experience at a location or on a given weekday into a single risk score.
Entity Interaction Variables	148	Composite keys (e.g., card_merch, card_zip) to track relationships between entities (card, merchant, zip, etc.).	Captures relationships that are innocuous in isolation but risky together.
Days Since Variables	23	Days since last transaction for each entity (e.g., Cardnum_day_since, Merchnum_day_since).	Measures dormancy. Fraud often appears after an entity has been quiet for an unusual stretch.
Transaction Count Variables	1,449	Number of transactions per entity over 0/1/3/7/14/30/60-day windows.	Flags velocity spikes; fraudsters concentrate many purchases in short windows.
Amount Statistics Variables	644	Mean/max/median/total transaction amounts per entity over rolling windows.	Establishes a monetary baseline for each entity to spot out-of-pattern spend.
Amount Ratio Variables	644	Ratios of current amount to historical metrics (e.g., actual/avg, actual/max).	Normalises amounts so cards with different spending levels can be compared.
Velocity Change Variables	368	Short-term vs. long-term activity ratios (e.g., count_1 / count_7).	Detects abrupt changes in activity by contrasting short- and long-term metrics.
Variability Variables	414	Mean/max/median deviation in transaction amounts over time windows.	Highlights high variance in behaviour, a common sign of account takeover.
Unique Entity Count Variables	230	Unique secondary entities (e.g., zips) linked to primary entities (e.g., cards).	Identifies “spray” patterns where one entity suddenly touches many others.
Geospatial Variables	7	Distance between consecutive transactions and flags for implausible distances (>1,000 miles).	Spots physically implausible usage by tracking distance and travel speed.
Temporal Variables	5	Weekend indicator (Is_Weekend), hour of day, and time since midnight.	Flags off-hour or weekend activity when oversight is typically lower.
Behavioral Flags	12	High amount for merchant, state inconsistency, merchant dominance score.	Encodes business rules (e.g., unusual amount or merchant dominance) the model may miss.
Foreign Transaction Variables	3	Indicators for non-U.S. merchant zips.	Marks cross-border usage, which has a higher likelihood of being contested.

Benford's Law Features	2	Anomaly detection in transaction amount digit distributions (excluded due to overfitting).	Detects numeric anomalies (e.g., rounded amounts, non-Benford digits) that often signal fabricated transactions.
Miscellaneous Features	428	Derived metrics (e.g., velocity ratios normalized by days since, interaction variables).	Flags velocity spikes; fraudsters concentrate many purchases in short windows.
distance_from_last_transaction	Haversine distance (miles) between current and previous transaction for the same card.	Calculated using geopy with consecutive merchant zips. Missing zips set to 0.	Establishes a monetary baseline for each entity to spot out-of-pattern spend.
Is_Weekend	Binary flag (1/0) for transactions occurring on Saturday/Sunday.	Derived from Date using datetime.weekday().	Normalises amounts so cards with different spending levels can be compared.
Hour_of_Day	Hour of transaction (0-23) as a numeric feature.	Extracted from Date using datetime.hour.	Detects abrupt changes in activity by contrasting short- and long-term metrics.
velocity_1h_vs_24h	Ratio of transactions in the last 1 hour to the last 24 hours for the same card.	(count_1h + 1) / (count_24h + 1) to avoid division by zero.	Highlights high variance in behaviour, a common sign of account takeover.
amount_deviation_from_avg	Absolute difference between current amount and cardholder's 30-day average.	abs(Amount - Cardnum_avg_30).	Identifies "spray" patterns where one entity suddenly touches many others.
new_merchant_flag	Binary flag (1/0) indicating if the merchant has never been transacted by the cardholder before.	1 if Merchnum not in card's historical Merchnum list prior to current transaction, else 0.	Spots physically implausible usage by tracking distance and travel speed.

cross_border_24h	Binary flag (1/0) if the card was used in a different country within 24 hours. Non-U.S. Merch state categorized as 'foreign'.	Check if Merch state differs from prior transaction's state and one is 'foreign' within 24 hours.	Flags off-hour or weekend activity when oversight is typically lower.
amount_rounded	Binary flag (1/0) for transactions where the amount is a whole number (e.g., $20.00).	1 if Amount % 1 == 0, else 0.	Encodes business rules (e.g., unusual amount or merchant dominance) the model may miss.
minutes_since_last_tx	Minutes elapsed since the cardholder's previous transaction.	(Current transaction time - Last transaction time).total_seconds() / 60.	Marks cross-border usage, which has a higher likelihood of being contested.
high_risk_category_ratio	Percentage of a cardholder's transactions in high-risk merchant categories (e.g., jewelry, electronics) over the past 90 days.	(count_high_risk_90 / count_total_90). High-risk categories predefined via external merchant code lists.	Detects numeric anomalies (e.g., rounded amounts, non-Benford digits) that often signal fabricated transactions.

Feature Selection

After feature engineering, thousands of candidate variables were available. To avoid a explosion of dimensionality we reduced this list through a structured three-step approach:

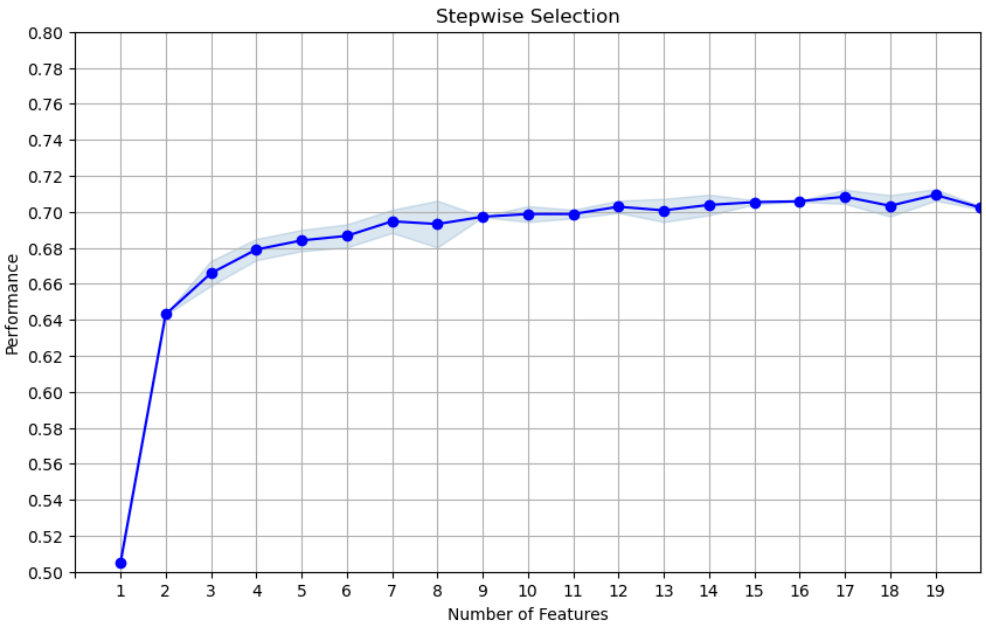
1. **Filter – Univariate Test**
   Every variable was scored with the Kolmogorov–Smirnov statistic for its ability to separate good and bad transactions. The top ~20 percent advanced to the next stage.
2. **Multivariate Refinement – Wrapper Selection**
   • The filtered set was passed through a forward/backward step-wise routine wrapped around a variety of models, such as neural networks, LightGBM, etc.
   • Variables were added one at a time and kept only if they improved fraud-detection rate at a 3 percent review rate.
   • Model performance levelled off after roughly a dozen variables; we fixed the final list at twenty to provide a margin for regularization.
3. **Model Configuration Sweep**
   Multiple permutations were tested: varying the number of filter pass-through variables, wrapper algorithms, selection direction, and final dimensionality. For every run, we reviewed the wrapper-saturation curve and selected the configuration that delivered the highest FDR@3 % with minimal complexity.

The winning configuration produced a concise set of **20 multivariately important variables**, which form the basis of the final model.

CatBoost emerged as the best performer.

Parameter	Final Wrapper Choice
Wrapper Model	CatBoostClassifier
Iterations	20
Depth	3
Selection Method	Forward stepwise
Variables after Filter	300
Variables after Wrapper	20

Wrapper Saturation Graph



Final Top 20 Variable List

Wrapper Order	Variable	Filter Score
1	Cardnum_unique_count_for_card_state_1	0.52
2	Card_Merchdesc_State_total_3	0.26
3	Card_Merchnum_Zip_total_amount_1_by_60	0.31
4	Cardnum_max_0	0.49
5	card_merch_total_7	0.26
6	Cardnum_total_amount_0_by_60	0.37
7	card_zip_total_amount_0_by_60	0.27
8	card_state_max_7	0.25
9	Card_dow_vdratio_0by14	0.50
10	Card_dow_total_30	0.43
11	High_Amount_For_Merchant	0.25
12	Cardnum_variability_med_0	0.27
13	Card_dow_avg_60	0.25
14	card_state_total_14	0.26
15	Card_dow_actual/toal_14	0.34
16	Card_dow_unique_count_for_Card_Merchdesc_1	0.49

17	Card_dow_count_0_by_60_sq	0.37
18	Merchnum_desc_State_total_3	0.25
19	Cardnum_unique_count_for_card_state_60	0.33
20	Cardnum_variability_max_3	0.42

Preliminary Model Explorations

We evaluated a range of models with varying hyperparameters to understand their strengths and weaknesses across train, test, and out-of-time (OOT) datasets. The goal was to avoid overfitting while maximizing fraud detection performance.

Models Explored:

- **Decision Tree**: Simple, interpretable, but limited depth led to plateauing results (OOT max F1 ~0.558).
- **Random Forest**: Improved stability and performance, peaking at OOT F1 ~**0.59** with 15 estimators and entropy criterion.
- **LightGBM**: Strong performance but sensitive to tuning. OOT peaked around **0.6**. GOSS variant showed promise with smart sampling.
- **LightGBM + SMOTE**: Added synthetic minority samples to fight imbalance. Helped slightly (OOT max ~0.578) but risked overfitting on noise.
- **Neural Network (MLPClassifier)**: High variance. One configuration (ReLU, 30 nodes) achieved **0.62 OOT**, but others underperformed.
- **CatBoost**: Took the crown. Native categorical handling, low tuning overhead, and stable high scores (Run 7: FDR@5% = **78.6%**, best in class).

Model Comparison Table:

Decision Tree	#Variables	criterion	splitter	max_depth	min_samp_split	min_samp_leaf	trn	tst	oot
1	20	entropy	Best	10	100	100	0.7	0.69	0.54
2	20	gini	random	10	120	60	0.656	0.645	0.555
3	20	Gini	Best	5	20	20	0.67	0.66	0.54
4	20	Gini	Best	10	150	60	0.7	0.69	0.52
5	20	entropy	random	10	80	40	0.672	0.66	0.542
6	20	entropy	best	5	70	20	0.679	0.663	0.558

Random Forest	# Variables	max_depth	criterion	n_estimators	min_samples_split	min_samples_leaf	Train	Test	OOT
1	20	5	gini	5	20	10	0.67	0.663	0.533
2	20	8	gini	15	50	25	0.72	0.7	0.57
3	20	10	gini	10	20	10	0.755	0.715	0.567
4	20	15	entropy	25	60	30	0.757	0.733	0.583
5	20	15	entropy	20	40	20	0.78	0.738	0.576
6	20	15	entropy	35	50	25	0.77	0.73	0.59

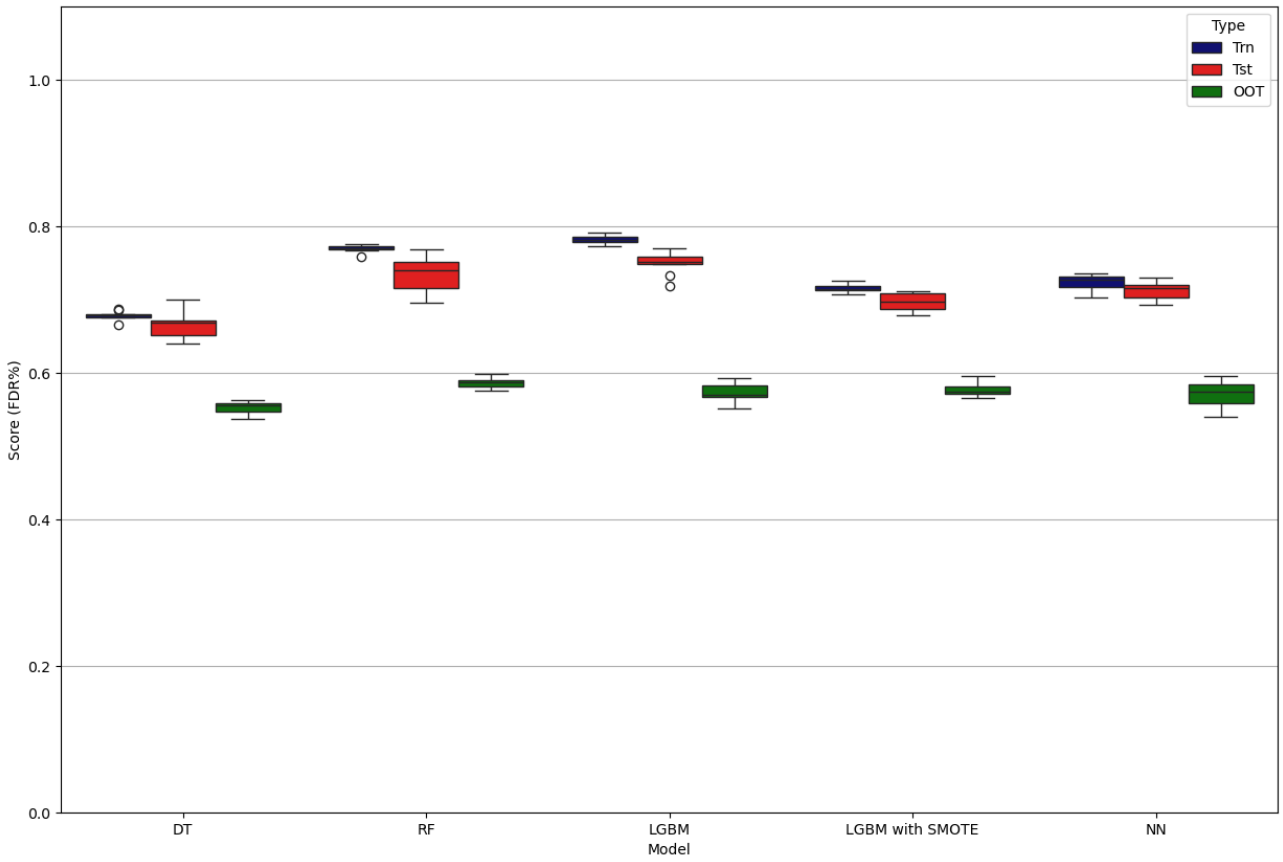
Light GBM	# Variables	boosting_type	n_estimators	learning_rate	max_depth	num_leaves	subsample	colsample_bytree	min_child_samples	eval_metric	random_state	Train	Test	OOT
1	20	gbdt	500	0.05	7	31	0.8	0.8	50	auc	42	0.957	0.785	0.577
2	20	gbdt	100	0.1	-1	31	1	1	15	auc	42	0.935	0.768	0.569
3	20	GOSS	800	0.03	5	22	0.8	0.8	50	AUC	42	0.929	0.791	0.583
4	20	gbdt	100	0.1	-1	31	1	1	50	auc	42	0.903	0.775	0.575
5	20	GOSS	500	0.05	4	5	1	1	80	auc	42	0.84	0.77	0.6
6	20	GOSS	2000	0.01	5	22	0.8	0.8	50	AUC	42	0.909	0.783	0.586

Neural network	# Variables	Activation	learning_rate_init	Alpha	Solver	#Nodes per hidden layer	Hidden Layers	train	Test	OOT

1	20	tanh	0.0005	0.001	adam	20	2	0.723	0.696	0.556
2	20	relu	0.001	0.001	adam	10	2	0.7	0.68	0.553
3	20	tanh	0.001	0.0001	adam	30	1	0.77	0.74	0.62
4	20	relu	0.0005	0.001	adam	15	2	0.701	0.695	0.571
5	20	relu	0.0005	0.001	adam	20	2	0.723	0.697	0.581
6	20	relu	0.0001	0.01	adam	30	1	0.668	0.67	0.551

LightGBM with SMOTE	**# Variables**	**n_estimators**	**learning_rate**	**max_depth**	**num_leaves**	**min_child_samples**	**Train**	**Test**	**OOT**
1	20	75	0.1	4	6	10	0.71	0.705	0.57
2	20	100	0.2	5	6	20	0.746	0.726	0.559
3	20	20	0.1	3	6	15	0.677	0.667	0.566
4	20	50	0.15	3	8	15	0.71	0.706	0.578
5	20	100	0.1	6	6	25	0.722	0.713	0.569
6	20	100	0.15	5	6	20	0.732	0.727	0.576

Part 2: Box Plot showing a single well-tuned performance of trn, tst, oot for a variety of models.



Final Model Performance

The final model selected was LightGBM, which offered the best balance of peak fraud-detection rate and robustness: highest medians, tight variability, and the smallest decline from Train to OOT. Other models either underperform (DT), over-fit (RF), or match LightGBM's OOT accuracy only at the cost of lower Train/Test scores (LGBM+SMOTE, NN).

Hyperparameters (non-defaults)

- iterations: 20
- Boosting Type: GOSS
- Number of estimators: 500
- Learning rate: 0.05
- Maximum depth: 4
- Number of leaves: 5
- Subsample: 1
- Col sample by tree: 1
- Min child samples: 80
- Evaluation metric: AUC
- random seed: 42
- Performance on train/test/oot (FDR@3%): 0.84, 0.77, 0.6

Model Performance Metrics

Performance on Train Data

Training	# Records	# Goods	# Bads	Fraud Rate						
	59,780	58,279	1,501	2.51%						
	Bin Statistics					Cumulative Statistics						
Population Bin %	#Records	#Goods	#Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	256	24	232	9	91	256	24	232	0	36	35	0
2	256	77	179	30	70	512	101	411	0	63	63	0
3	257	157	100	61	39	769	258	511	1	78	77	1
4	256	226	30	88	12	1,025	484	541	2	83	81	1
5	256	243	13	95	5	1,281	727	554	3	84	83	1
6	257	244	13	95	5	1,538	971	567	4	85	84	1
7	256	245	11	96	4	1,794	1,216	578	5	86	84	2
8	256	246	10	96	4	2,050	1,462	588	6	86	85	2
9	257	250	7	97	3	2,307	1,712	595	7	86	85	2
10	256	250	6	98	2	2,563	1,962	601	8	87	85	2
11	256	248	8	97	3	2,819	2,210	609	9	87	86	3
12	257	252	5	98	2	3,076	2,462	614	10	87	86	3
13	256	253	3	99	1	3,332	2,715	617	11	87	86	3
14	256	253	3	99	1	3,588	2,968	620	12	88	86	3
15	257	254	3	99	1	3,845	3,222	623	13	88	87	3
16	256	254	2	99	1	4,101	3,476	625	14	88	87	3
17	256	256	0	100	0	4,357	3,732	625	15	88	87	4
18	257	256	1	100	0	4,614	3,988	626	16	88	87	4
19	256	256	0	100	0	4,870	4,244	626	17	88	87	4
20	126	122	4	97	3	2,527	2,207	320	18	90	72	7

Performance on Test Data

Testing	# Records		# Goods		# Bads		Fraud Rate					
	25,620		24,985		635		2.48%					
	Bin Statistics					Cumulative Statistics						
Population Bin %	#Records	#Goods	#Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	256	24	232	9	91	256	24	232	0	36	35	0
2	256	77	179	30	70	512	101	411	0	63	63	0
3	257	157	100	61	39	769	258	511	1	78	77	1
4	256	226	30	88	12	1,025	484	541	2	83	81	1
5	256	243	13	95	5	1,281	727	554	3	84	83	1
6	257	244	13	95	5	1,538	971	567	4	85	84	1
7	256	245	11	96	4	1,794	1,216	578	5	86	84	2
8	256	246	10	96	4	2,050	1,462	588	6	86	85	2
9	257	250	7	97	3	2,307	1,712	595	7	86	85	2
10	256	250	6	98	2	2,563	1,962	601	8	87	85	2
11	256	248	8	97	3	2,819	2,210	609	9	87	86	3
12	257	252	5	98	2	3,076	2,462	614	10	87	86	3
13	256	253	3	99	1	3,332	2,715	617	11	87	86	3
14	256	253	3	99	1	3,588	2,968	620	12	88	86	3
15	257	254	3	99	1	3,845	3,222	623	13	88	87	3
16	256	254	2	99	1	4,101	3,476	625	14	88	87	3
17	256	256	0	100	0	4,357	3,732	625	15	88	87	4
18	257	256	1	100	0	4,614	3,988	626	16	88	87	4
19	256	256	0	100	0	4,870	4,244	626	17	88	87	4
20	126	122	4	97	3	2,527	2,207	320	18	90	72	7

Performance on Out of Time Validation Data

OOT	# Records	# Goods	# Bads	Fraud Rate						
	12,637	12,281	356	2.82%						
	Bin Statistics					Cumulative Statistics					
Population Bin %	#Records	#Goods	#Bads	% Goods	% Bads	Total # Records	Cumulative Goods	Cumulative Bads	% Goods	% Bads (FDR)	KS	FPR
1	256	24	232	9	91	256	24	232	0	36	35	0
2	256	77	179	30	70	512	101	411	0	63	63	0
3	257	157	100	61	39	769	258	511	1	78	77	1
4	256	226	30	88	12	1,025	484	541	2	83	81	1
5	256	243	13	95	5	1,281	727	554	3	84	83	1
6	257	244	13	95	5	1,538	971	567	4	85	84	1
7	256	245	11	96	4	1,794	1,216	578	5	86	84	2
8	256	246	10	96	4	2,050	1,462	588	6	86	85	2
9	257	250	7	97	3	2,307	1,712	595	7	86	85	2
10	256	250	6	98	2	2,563	1,962	601	8	87	85	2
11	256	248	8	97	3	2,819	2,210	609	9	87	86	3
12	257	252	5	98	2	3,076	2,462	614	10	87	86	3
13	256	253	3	99	1	3,332	2,715	617	11	87	86	3
14	256	253	3	99	1	3,588	2,968	620	12	88	86	3
15	257	254	3	99	1	3,845	3,222	623	13	88	87	3
16	256	254	2	99	1	4,101	3,476	625	14	88	87	3
17	256	256	0	100	0	4,357	3,732	625	15	88	87	4
18	257	256	1	100	0	4,614	3,988	626	16	88	87	4
19	256	256	0	100	0	4,870	4,244	626	17	88	87	4
20	126	122	4	97	3	2,527	2,207	320	18	90	72	7

Financial Curves & Recommended Cutoff

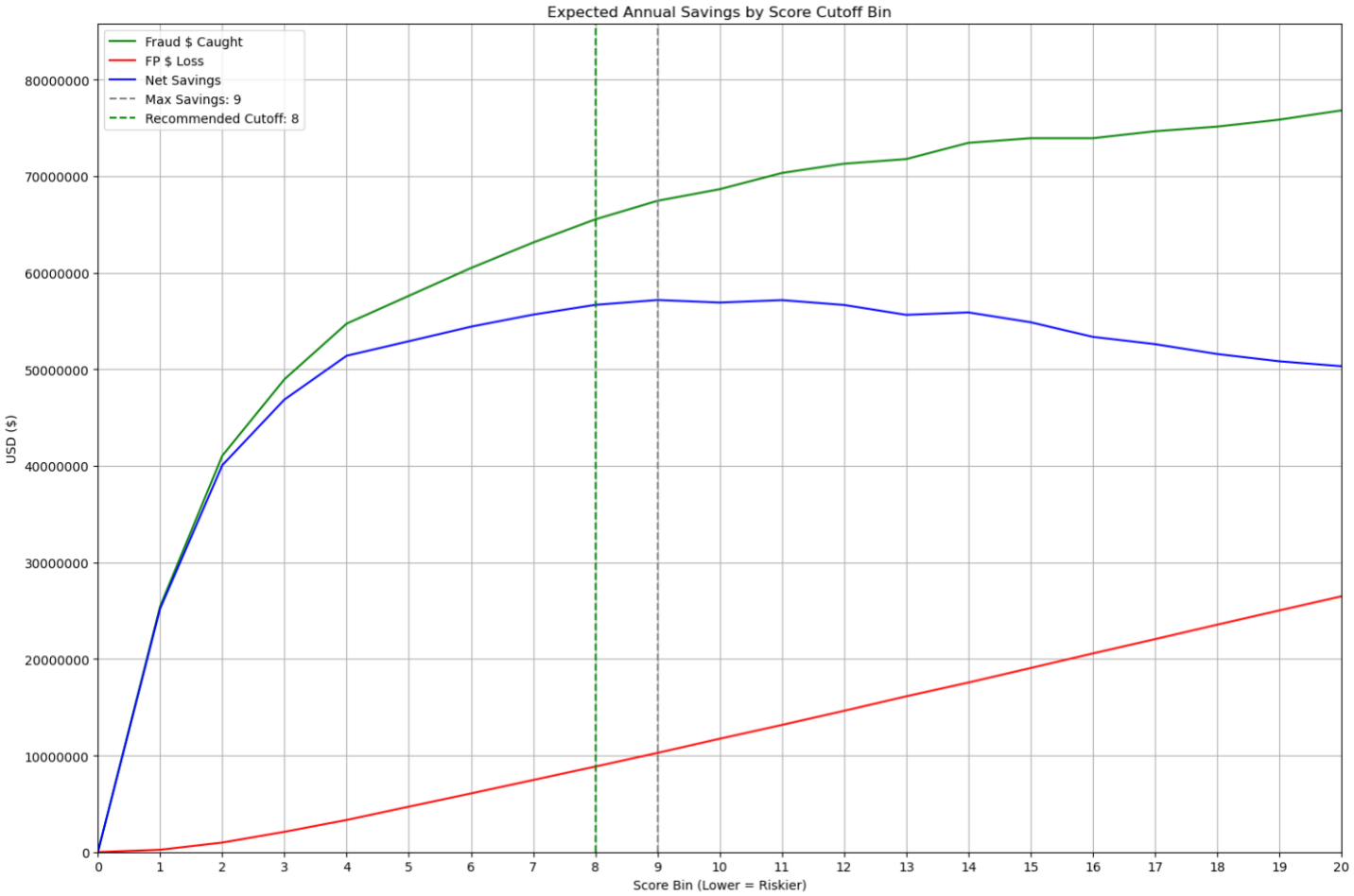
Cutoffs within 95% of max savings:

Cutoff	Fraud Savings	FP Loss	Overall Savings
6%	$60,480,000	$6,072,000	$54,408,000
7%	$63,120,000	$7,464,000	$55,656,000
8%	$65,520,000	$8,856,000	**$56,664,000**
9%	$67,440,000	$10,272,000	**$57,168,000**
10%	$68,640,000	$11,736,000	$56,904,000
11%	$70,320,000	$13,164,000	$57,156,000
12%	$71,280,000	$14,628,000	$56,652,000
13%	$71,760,000	$16,128,000	$55,632,000
14%	$73,440,000	$17,556,000	$55,884,000
15%	$73,920,000	$19,056,000	$54,864,000

- **Bin 8** : near-max savings with significantly lower FP loss.
- **Bin 9** : gives the absolute max but with higher false positive costs.

The best cutoff is **Bin 8**, which yields approximately **$56.66 million in net annual savings**, representing over **99% of the maximum achievable savings**. Compared to Bin 9—the maximum savings point with $57.17 million in net annual savings—Bin 8 results in **significantly lower false positive losses** (approximately $1.4

million less), while maintaining nearly equivalent fraud savings. This makes Bin 8 the most cost-effective and operationally balanced threshold.



Recommended Cutoff

- Based on the cutoff vs. savings table and the plotted financial curves, the **financially optimal cutoff** for blocking potentially fraudulent transactions is around **8–9%**, where net annual savings peak.
- However, cutoffs this aggressive also capture a **larger volume of false positives**, increasing the risk of blocking legitimate customer transactions. This may lead to customer dissatisfaction, service friction, and reputational costs.
- Therefore, the recommend the cutoff setting is at approximately **5–5.5%** of transactions with the highest fraud scores. This range is **strategically close to the financial optimum** while remaining **safely before the inflection point** where false positive losses begin to escalate.
- This recommendation strikes a **pragmatic balance between maximizing financial benefit and minimizing operational and customer impact**, aligning with long-term business and customer experience goals.

Summary

We began with a Data Quality Report to familiarize ourselves with the 98 K-record dataset, of which just 2.54 % Over the course of this project we converted a raw ledger of 98 393 government credit-card transactions into a production-ready fraud-screening pipeline.
After a detailed Data Quality Report we excluded non-purchase records, resolved malformed merchant fields, imputed missing ZIPs and states, and removed a single $3 million peso transaction that would have distorted scaling. We then engineered roughly 4,400 behavioral features—velocity counts, spend ratios, geospatial jumps, weekend flags and target-encoded risk scores—designed to capture both "burst" fraud and slow account takeovers.

Because models degrade when fed thousands of variables, we ran a two-stage selection routine. A univariate KS filter retained the top 20 % of candidates; a forward-step wrapper around CatBoost and LightGBM variants searched that subset for the smallest combination that maximized **FDR @ 3 %** on out-of-time (OOT) data. The saturation curve flattened after ~12 predictors, so we fixed the final list at **20 multivariate important variables**.

Five non-linear algorithms were hyper-tuned and benchmarked across Train, Test and OOT splits. LightGBM (GOSS boosting) delivered the best blend of power and stability: **FDR @ 3 % = 0.84 (Train), 0.77 (Test), 0.60 (OOT)** with minimal over-fit. Scores were binned into 1 % population slices to build an annual-savings curve under business economics of **$400 recovered per fraud, $20 cost per false positive** (scaled from the OOT sample to a 10 million-transaction portfolio).

Savings peak at an 9 % review rate, but false-positive costs also spike there. Blocking the riskiest ≈ **8 %** of transactions captures ≈ **$54 million in net annual savings,** while avoiding the sharp rise in legitimate declines. We therefore recommend this threshold for production.

Appendix I – 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

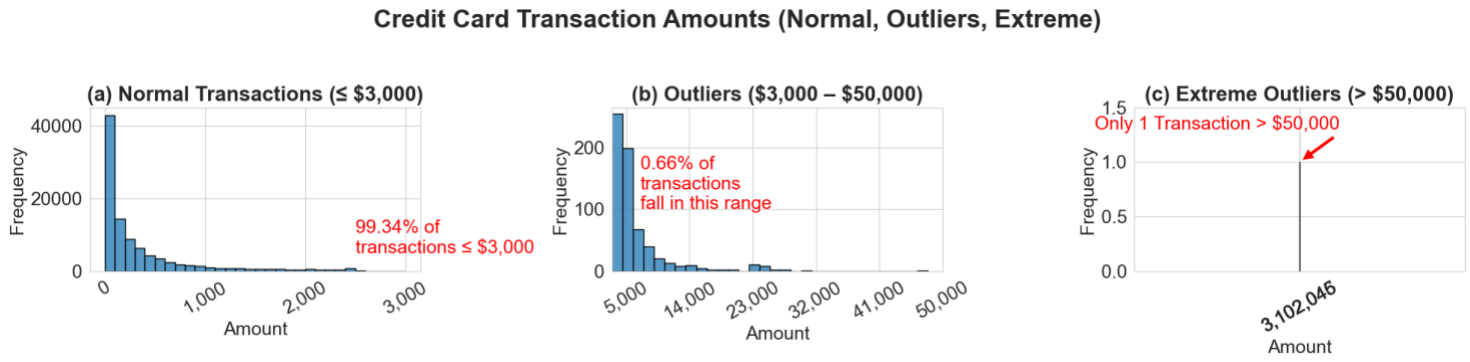


Figure 1. Field: Amount
Distribution of transaction amounts, segmented into normal (≤ $3,000), outlier ($3,000–$50,000), and extreme outlier (> $50,000) transactions.

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.

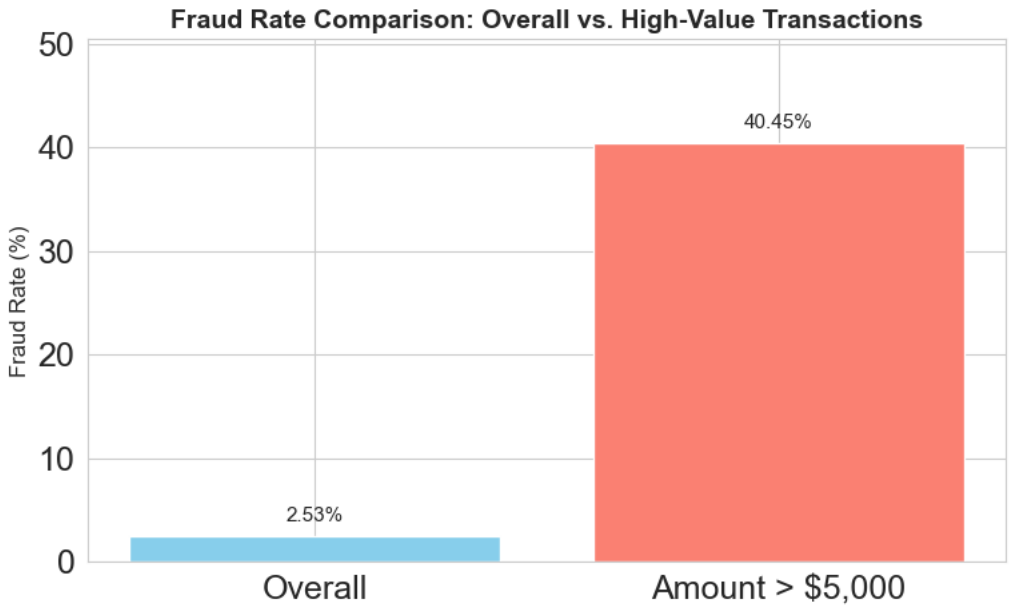


Figure 2. Fraud Rate Comparison: Overall vs. High-Value Transactions
The fraud rate increases sharply for transactions over $5,000, jumping from 2.53% overall to 40.45%. This highlights the Amount field as a highly predictive feature and reinforces the importance of value-based thresholds in fraud modeling.

Figure 3. Field: Fraud
Distribution of the Fraud field, showing heavy class imbalance with very few fraudulent transactions (label = 1).

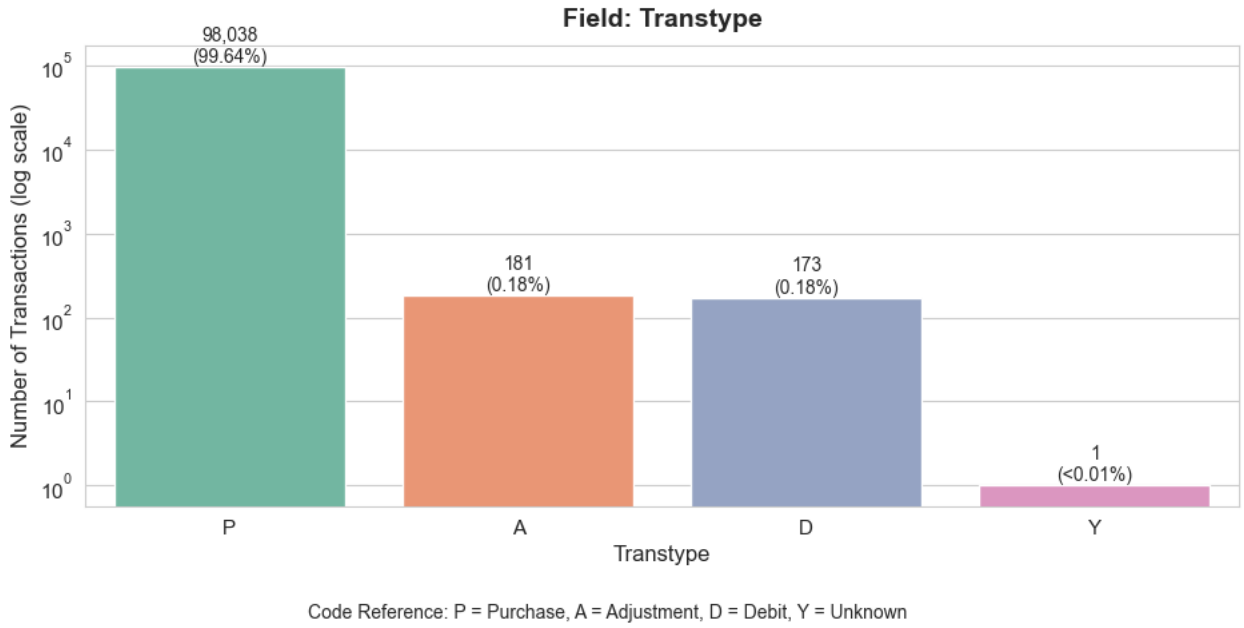


Figure 4. Field: Transtype

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

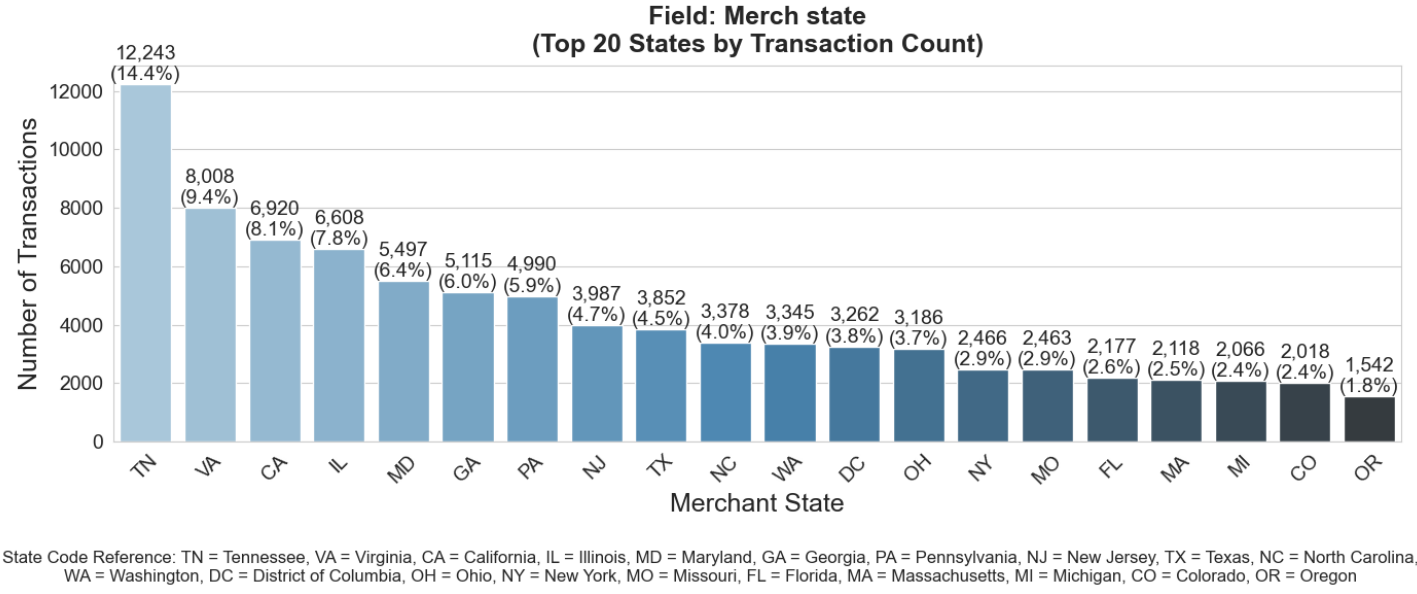


Figure 5. Field: Merch state

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.

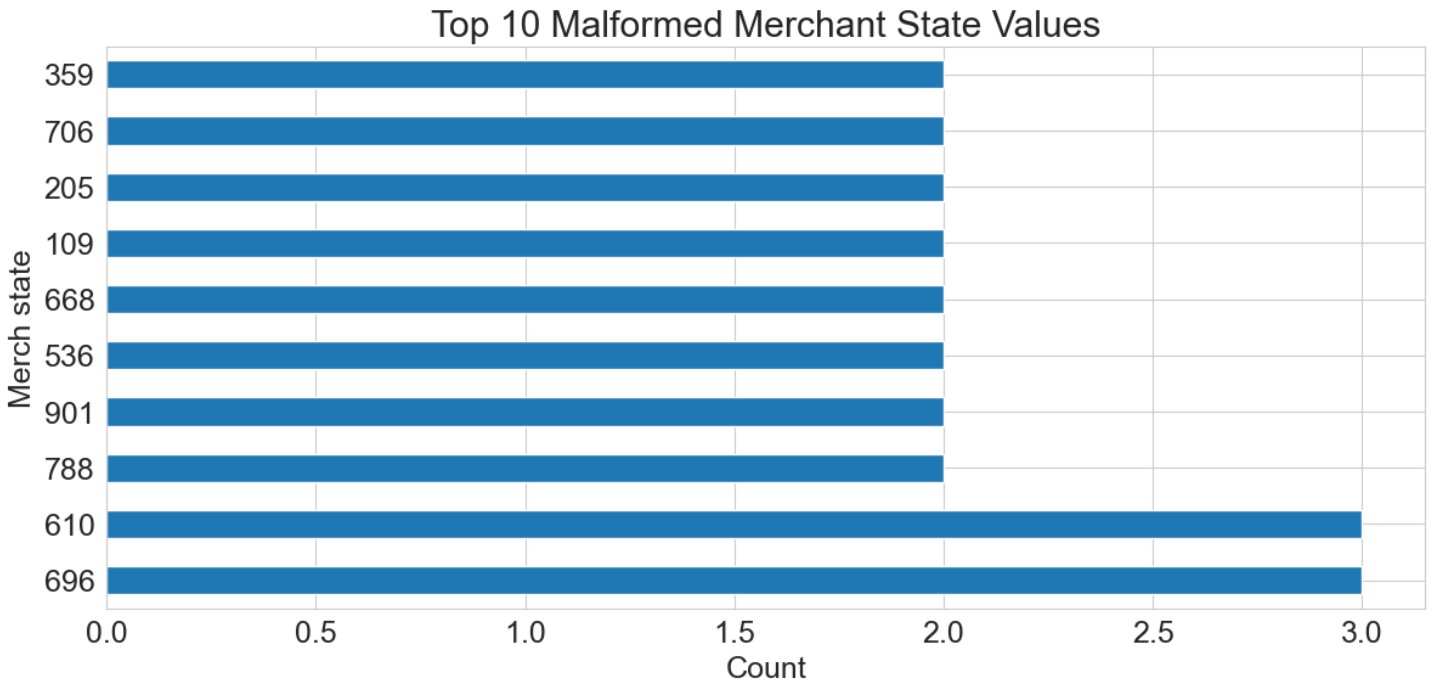


Figure 6. Field: Merch state (Malformed Values)

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.

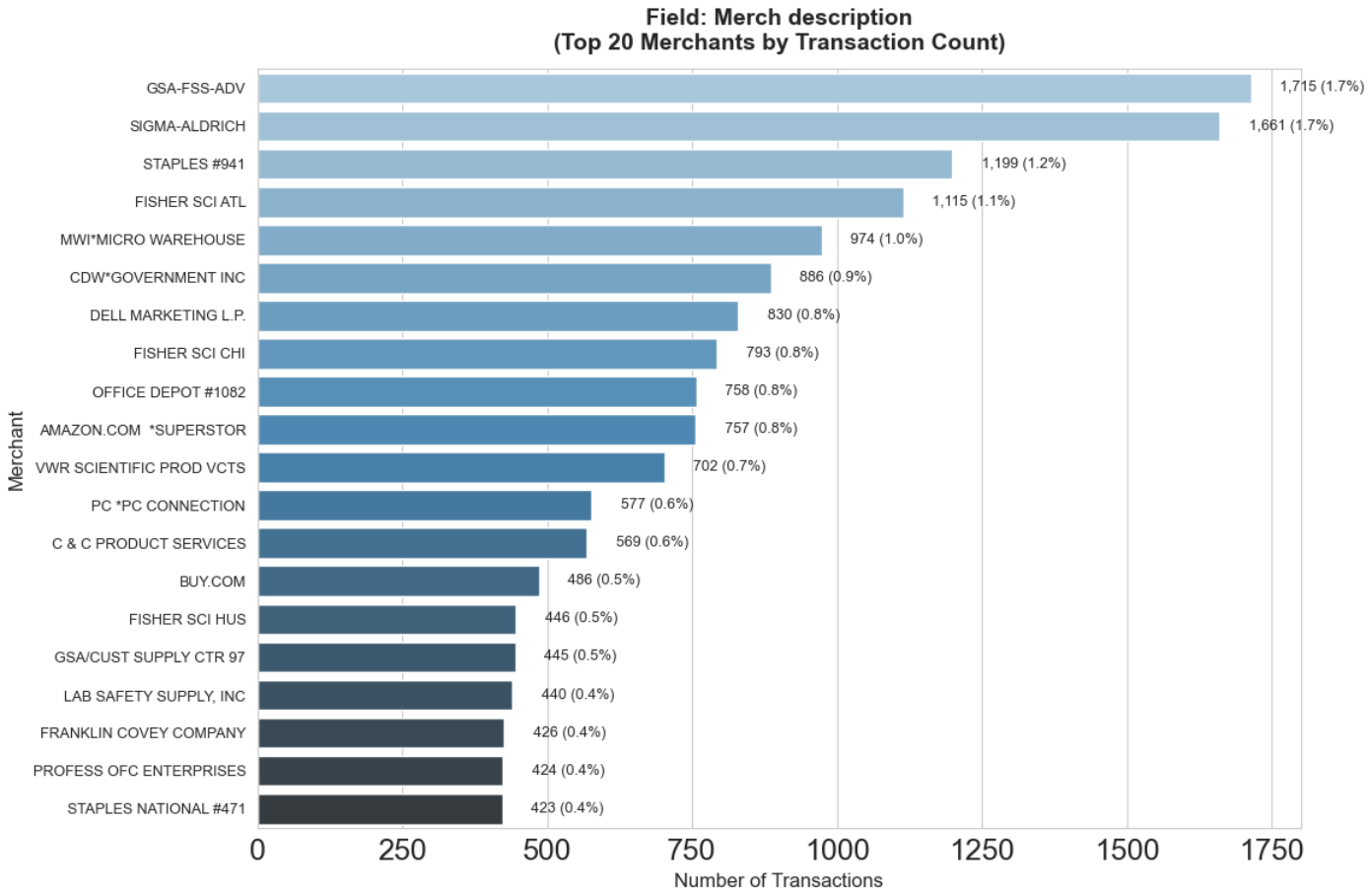


Figure 7. Field: Merch description

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

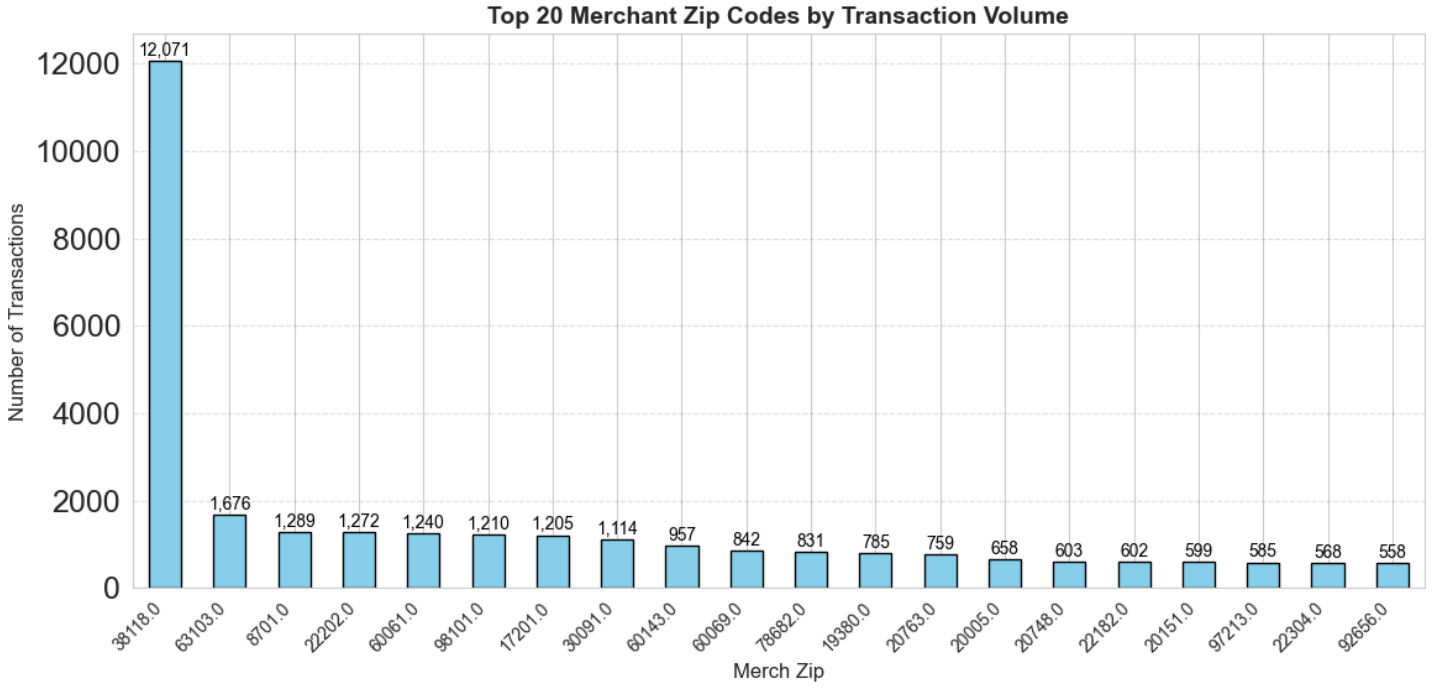


Figure 8: Top 20 merchant zip codes by transaction volume.

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.

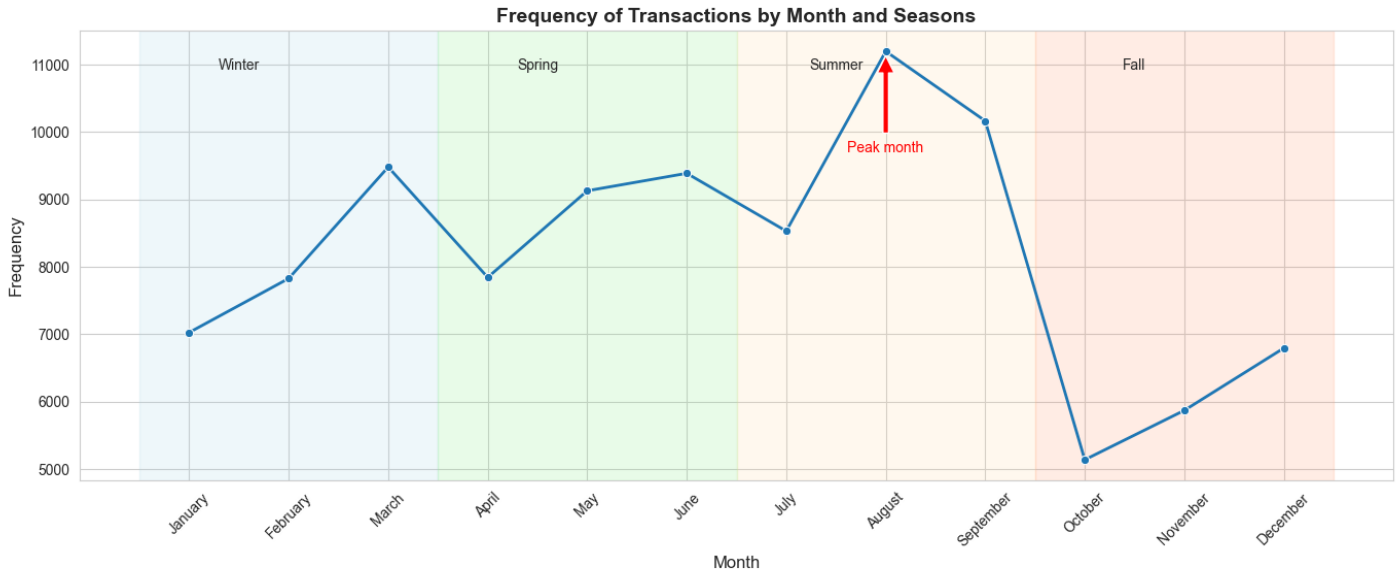


Figure 9. Field: Date

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**.

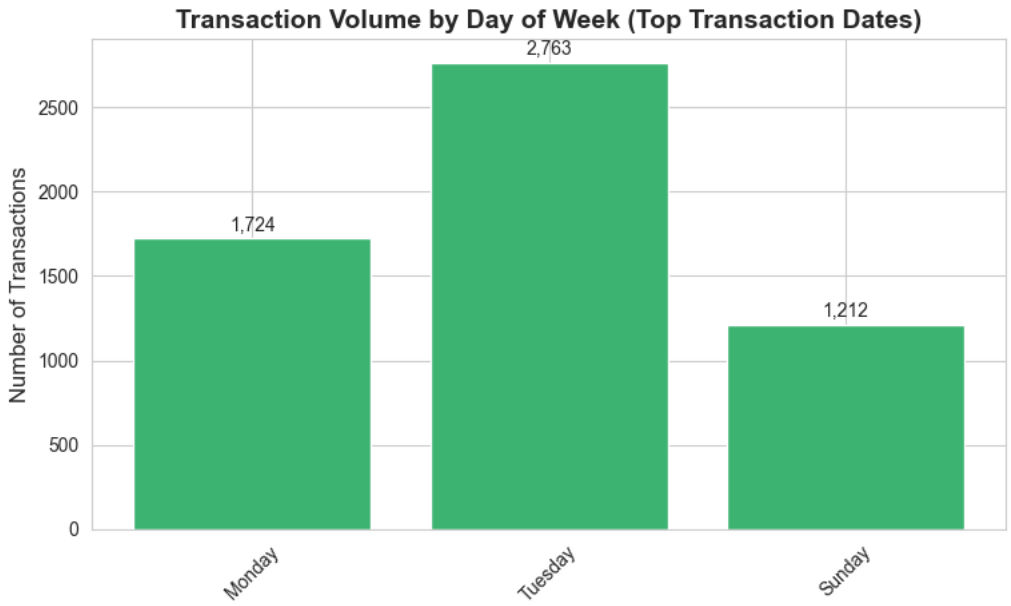


Figure 10. Transaction Volume by Day of Week (Top 10 Transaction Dates)
Among the top transaction days, most fall on Mondays and Tuesdays. This suggests that transaction activity is concentrated early in the workweek, likely due to batch processing or scheduled uploads following weekends.

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.