Lab 1 - Credit Card Fraud Wrangling and EDA

Name: Xinhui Qian

Date: 2025/04/16

1 Introduction

This dataset captures detailed information about credit card transactions, providing a comprehensive view of individual financial interactions. In the project we aim to gain insights from various predictors of fraudulent credit card transactions, which is crucial in the financial industry.

2 Data

2.1 Basic Information

- **Number of rows**: 786,363
- **Number of columns**: 30
- **Duplicate columns**: None found

2.2 Data Types

The variables below encompass crucial details ranging from transaction specifics and card information to account characteristics and potential fraud indicators.

Variables	Data Type	Description
`accountNumber`	int64	A unique identifier for the customer account associated with the transaction
`creditLimit`	float64	The maximum amount of credit available to the customer on their account

Variables	Data Type	Description
`availableMoney`	float64	The amount of credit available to the customer at the time of the transaction
`transactionDateTime`	object	The date and time of the transaction
`transactionAmount`	float64	The amount of the transaction
`merchantName`	object	The name of the merchant where the transaction took place
`acqCountry`	object	The country where the acquiring bank is located
`merchantCountryCode`	object	The country where the merchant is located
`posEntryMode`	float64	The method used by the customer to enter their payment card information during the transaction
`posConditionCode`	float64	The condition of the point-of-sale terminal at the time of the transaction
`merchantCategoryCode`	object	The category of the merchant where the transaction took place
`currentExpDate`	object	The expiration date of the customer's payment card
`accountOpenDate`	object	The date the customer's account was opened
`dateOfLastAddressChange`	object	The date the customer's address was last updated
`cardCVV`	int64	The three-digit CVV code on the back of the customer's payment card
`enteredCVV`	int64	The CVV code entered by the customer during the transaction
`cardLast4Digits`	int64	The last four digits of the customer's payment card
`transactionType`	object	The type of transaction
`currentBalance`	float64	The current balance on the customer's account
`cardPresent`	bool	Whether or not the customer's payment card was present at the time of the transaction

Variables	Data Type	Description
`expirationDateKeyInMatch`	bool	Whether or not the expiration date of the payment card was entered correctly during the transaction
`isFraud`	bool	Whether or not the transaction was fraudulent

3 Data Preprocessing

3.1 Columns with Partial Missing Data

The table below shows the percentage of missing values in columns that contain incomplete data.

Column Name	Missing Values	Percentage
acqCountry	4,562	0.58%
merchantCountryCode	724	0.09%
posEntryMode	4,054	0.52%
posConditionCode	409	0.05%
transactionType	698	0.09%

I analyzed the transaction data before and after removing records with missing values. The original dataset had 786,363 records with 12,417 fraud cases (1.5790% fraud rate). After cleaning by removing entries with missing values, I kept 776,668 records with 11,966 fraud cases (1.5407% fraud rate). This process removed 9,695 records (1.23% of the original dataset). I tested whether the 2.43% decrease in fraud rate was statistically significant and got a p-value of 1.0000, indicating that the difference is not significant at the standard α=0.05 level. This means removing records with missing values didn't meaningfully change the overall fraud pattern in the dataset.

After deleting the missing values:

Metric	Original Data	After Cleaning	Change
Total Records	786,363	776,668	-9,695 (-1.23%)
Fraud Cases	12,417	11,966	-451 (-3.63%)

Metric	Original Data	After Cleaning	Change
Fraud Rate	1.5790%	1.5407%	-0.0383% (-2.43%)

I conducted statistical analysis to determine if the observed change in fraud rates could be attributed to random variation or if it represents a meaningful shift in the data distribution.

Test Parameter	Value	Interpretation
P-value	1.0000	Exceeds standard threshold (0.05)
Statistical Significance	No	The difference is not statistically significant
Confidence Level	95%	Standard level for statistical testing

The removal of records with missing values had minimal impact on the overall fraud rate. The 2.43% decrease wasn't statistically significant, suggesting my data cleaning approach didn't introduce bias into the fraud detection process. This confirms that using the cleaned dataset for modeling is appropriate and shouldn't negatively affect fraud detection algorithm performance.

3.2 Columns with Complete Missing Data

I deleted columns that were 100% missing.
The following columns had no data at all:

- echoBuffer
- merchantCity
- merchantState
- merchantZip
- posOnPremises
- recurringAuthInd

3.3 Outliers Detections

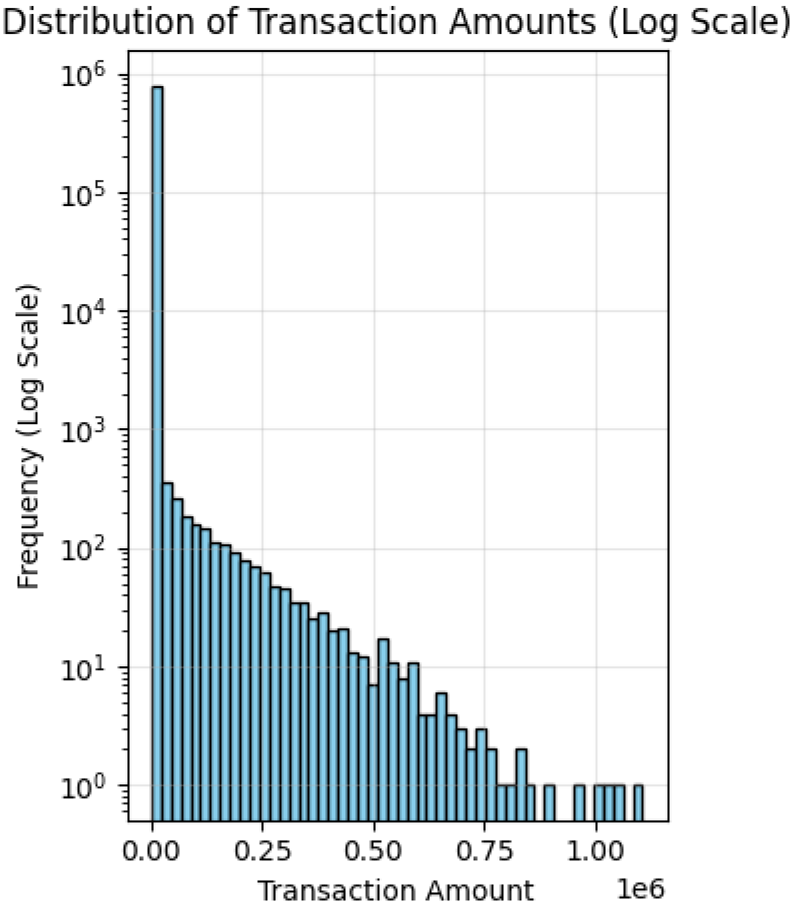
I used the IQR method to calculate bounds and identify outliers. I found that `creditLimit`, `transactionAmount`, `availableMoney`, `posEntryMode`, `posConditionCode`, and `currentBalance` contain outliers.
After investigating the plots and understanding the real meaning of the data, I decided these outliers should not be removed as they represent legitimate values in the context of credit card transactions.

3.4 Datetime Data

For date format data, I transformed them into `Datetime64` format so I could calculate time differences and uncover relationships between dates and fraud patterns.

3.5 Other Data

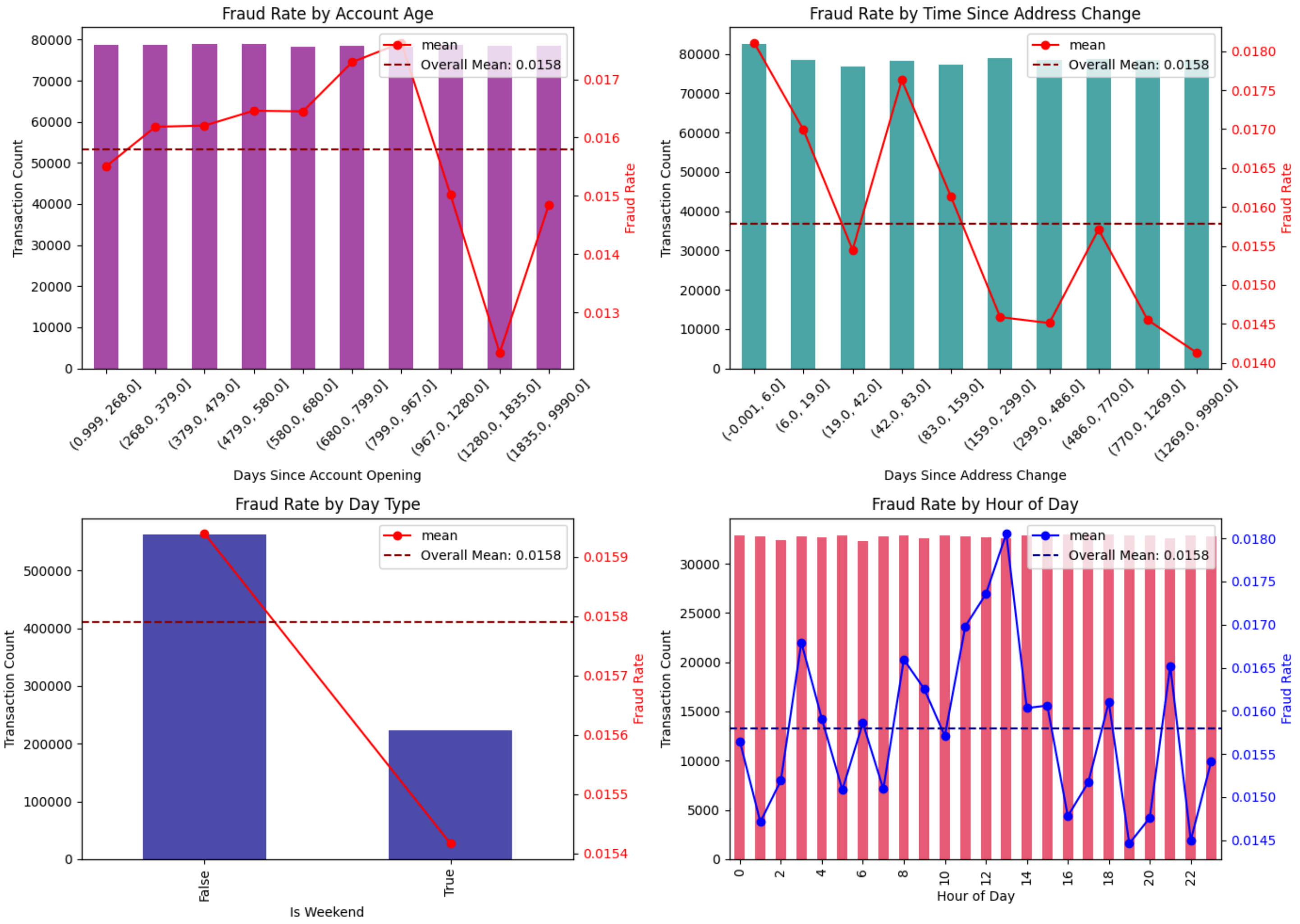
For the `cardLast4Digits`, I fill the fourth digit with `0` and transform data related to CVV into `string` type data.
For the `transactionAmount` variable, I applied a logarithmic transformation to normalize the distribution and enhance the visibility of patterns in the visualizations, as transaction amounts typically follow a right-skewed distribution. From the plot, we can see large deviation between transaction amount=, and the number decrease with the amount increase.

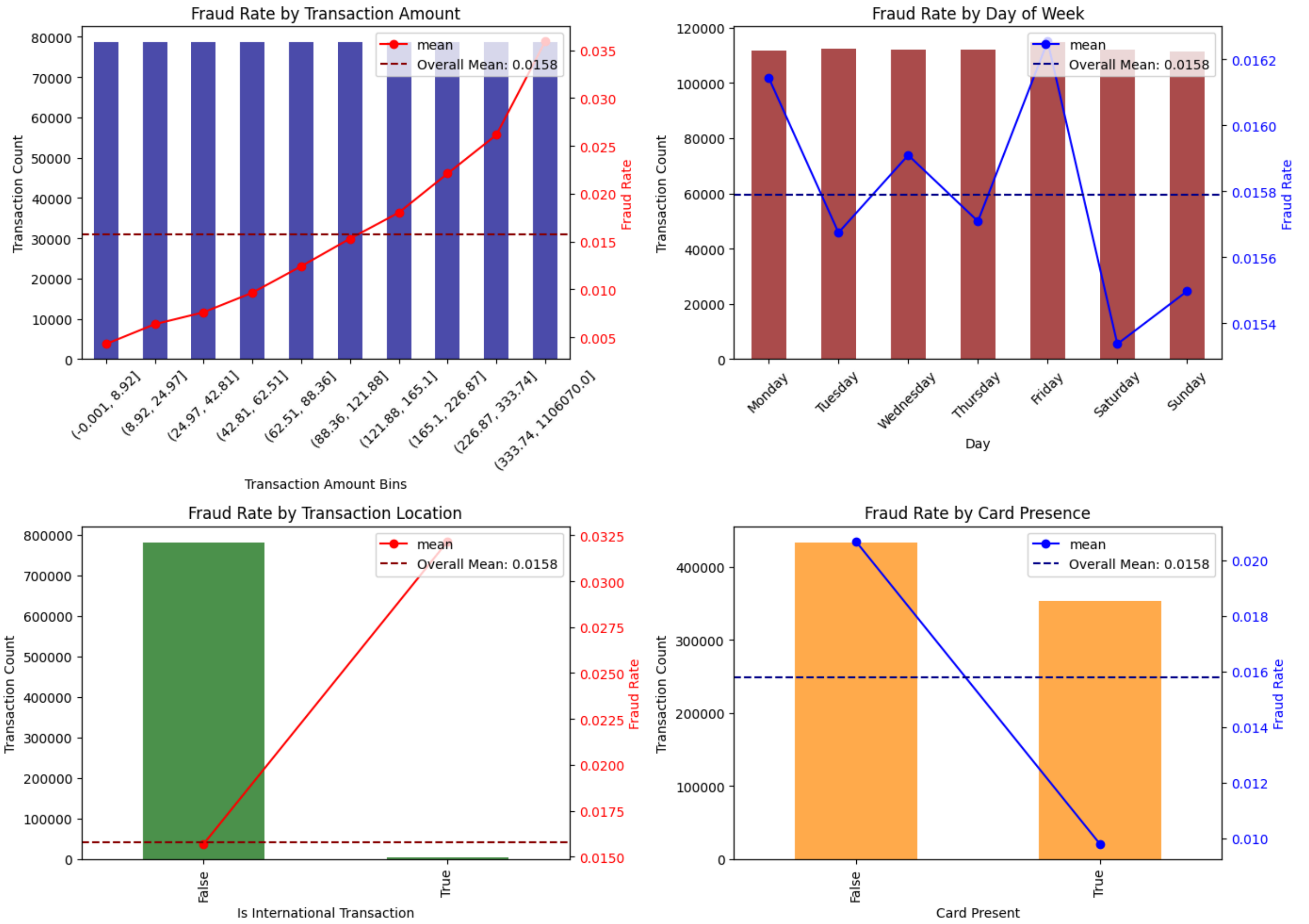


4 EDA Results

4.1 Time-Based and basic Patterns

My temporal analysis of transaction data revealed several significant fraud patterns:


Fraud Rate by Account Age
Transaction Count
0
10000
20000
30000
40000
50000
60000
70000
80000
Fraud Rate
0.013
0.014
0.015
0.016
0.017
mean
Overall Mean: 0.0158
(0.999, 268.0]
(268.0, 379.0]
(379.0, 479.0]
(479.0, 580.0]
(580.0, 680.0]
(680.0, 799.0]
(799.0, 967.0]
(967.0, 1280.0]
(1280.0, 1835.0]
(1835.0, 9990.0]
Days Since Account Opening
Fraud Rate by Time Since Address Change
Transaction Count
0
10000
20000
30000
40000
50000
60000
70000
80000
Fraud Rate
0.0140
0.0145
0.0150
0.0155
0.0160
0.0165
0.0170
0.0175
0.0180
mean
Overall Mean: 0.0158
(-0.001, 6.0]
(6.0, 19.0]
(19.0, 42.0]
(42.0, 83.0]
(83.0, 159.0]
(159.0, 299.0]
(299.0, 486.0]
(486.0, 770.0]
(770.0, 1269.0]
(1269.0, 9990.0]
Days Since Address Change
Fraud Rate by Day Type
Transaction Count
0
100000
200000
300000
400000
500000
Fraud Rate
0.0154
0.0155
0.0156
0.0157
0.0158
0.0159
mean
Overall Mean: 0.0158
False
True
Is Weekend
Fraud Rate by Hour of Day
Transaction Count
0
5000
10000
15000
20000
25000
30000
Fraud Rate
0.0145
0.0150
0.0155
0.0160
0.0165
0.0170
0.0175
0.0180
mean
Overall Mean: 0.0158
0
2
4
6
8
10
12
14
16
18
20
22
Hour of Day



- **Account Age**: Newer accounts (open within 3 years) show higher fraud rates compared to established accounts (>3 years).
- **Address Changes**: Fraud rates spike immediately after address changes and gradually decrease over time, with the highest fraud rates occurring within 30 days after a change.
- **Day of Week**: I found no clear pattern indicating whether weekends or weekdays have higher fraud rates, as the transaction volume isn't balanced between these categories and the difference is not significant (0.0005).
- **Hour of Day**: Fraud rates peak during 11am-1pm and reach their lowest point during evenings (7-8pm).
- **Fraud Rate by Transaction Amount**: The top-left graph shows fraud rates across different transaction amount bins. There's a clear upward trend as transaction amounts increase, with the highest fraud rates in the $1000-$2000 range. This suggests fraudsters target higher-value transactions to maximize gains before detection.
- **International Transaction Indicator**: The middle-left graph shows international transactions have a substantially higher fraud rate than domestic ones. This aligns with industry knowledge that cross-border transactions present higher fraud risks due to jurisdictional complexities and verification challenges.

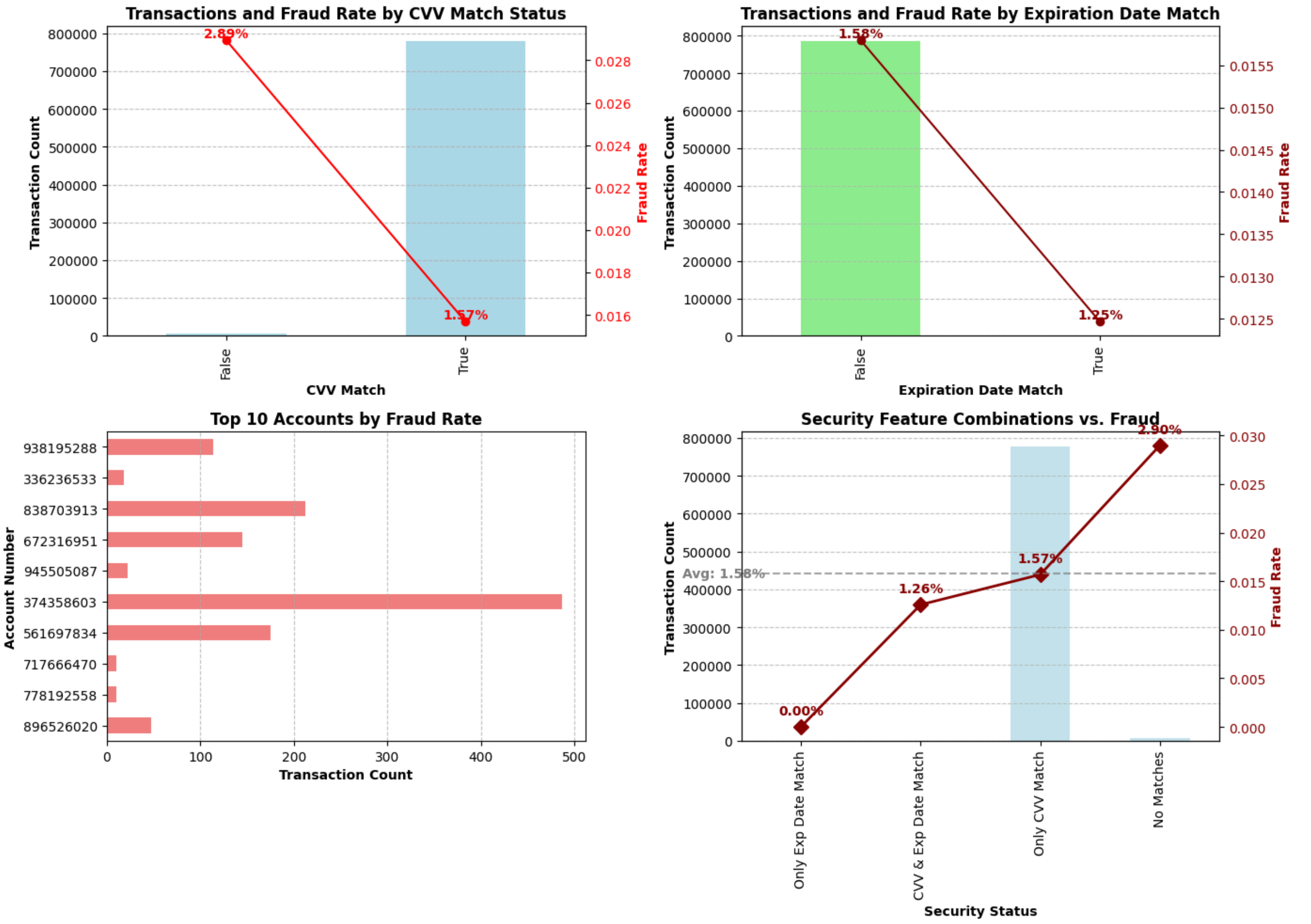
- **Card Present vs. Card Not Present**: The middle-right graph illustrates that card-not-present transactions (like online purchases) have a significantly higher fraud rate than card-present transactions. This reflects the additional security challenges in verifying remote transactions where physical card verification isn't possible.
- **Fraud Rate by Day of Week**: The bottom-right graph shows fraud rates across different days of the week. Weekend days (particularly Sunday) show elevated fraud rates compared to weekdays, with Wednesday having the lowest fraud rate. This weekly pattern may reflect differences in monitoring capabilities or consumer behavior patterns that fraudsters exploit.

Summary

This reveals several clear risk patterns in our dataset. Newer accounts (less than 3 years old) and accounts with recent address changes show significantly higher fraud rates. The data indicates fraud is concentrated in high-value transactions between $1000-$2000, international purchases, and card-not-present scenarios like online shopping. Time-based patterns suggest fraud peaks during midday hours (11am-1pm) and on weekends (especially Sunday), while reaching lowest levels during evening hours (7-8pm) and midweek (particularly Wednesday). These patterns suggest fraudsters strategically target situations with higher potential payoffs and weaker verification protocols, potentially exploiting predictable variations in bank monitoring throughout the week.

4.2 Security Features and Fraud Patterns

My analysis of security features revealed critical insights into fraud patterns:



- **CVV Match Status**: The top-left graph shows a dramatic difference in fraud rates between transactions with mismatched CVV (2.84%) versus matched CVV (1.57%). This nearly 2x difference highlights how important CVV verification is for fraud prevention. While there are fewer transactions with mismatched CVV, their fraud rate is significantly higher.
- **Expiration Date Match**: The top-right graph shows that expiration date mismatches also correlate with higher fraud rates (1.58%) compared to matched expiration dates (1.25%). While the difference is less dramatic than with CVV, it's still a meaningful fraud indicator.
- **Top 10 Accounts by Fraud Rate**: The bottom-left graph identifies specific high-risk accounts with elevated fraud activity. Account #374358603 shows the highest transaction count among high-fraud accounts, suggesting a potentially compromised account with sustained fraudulent usage. The varying patterns across these accounts indicate that fraud manifests differently across users.
- **Security Feature Combinations**: The bottom-right graph provides a comprehensive view of how different security feature combinations affect fraud rates:
    - When no security features match (neither CVV nor expiration date), the fraud rate spikes dramatically to 2.09% - nearly double the dataset average.
    - When only CVV matches but expiration date doesn't, the fraud rate is 1.26%.

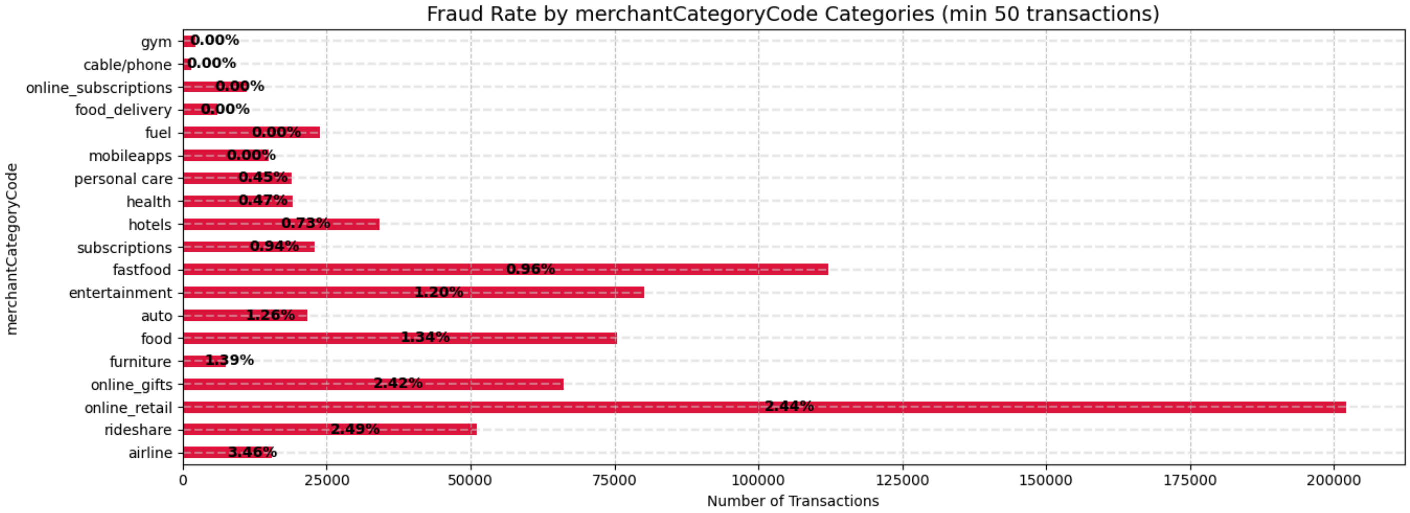
- When only expiration date matches but CVV doesn't, the fraud rate is 0%, but since the transaction count is 0 here, this scenario doesn't provide meaningful insights.
- The average fraud rate across the dataset is 1.58% (shown by the horizontal dashed line).

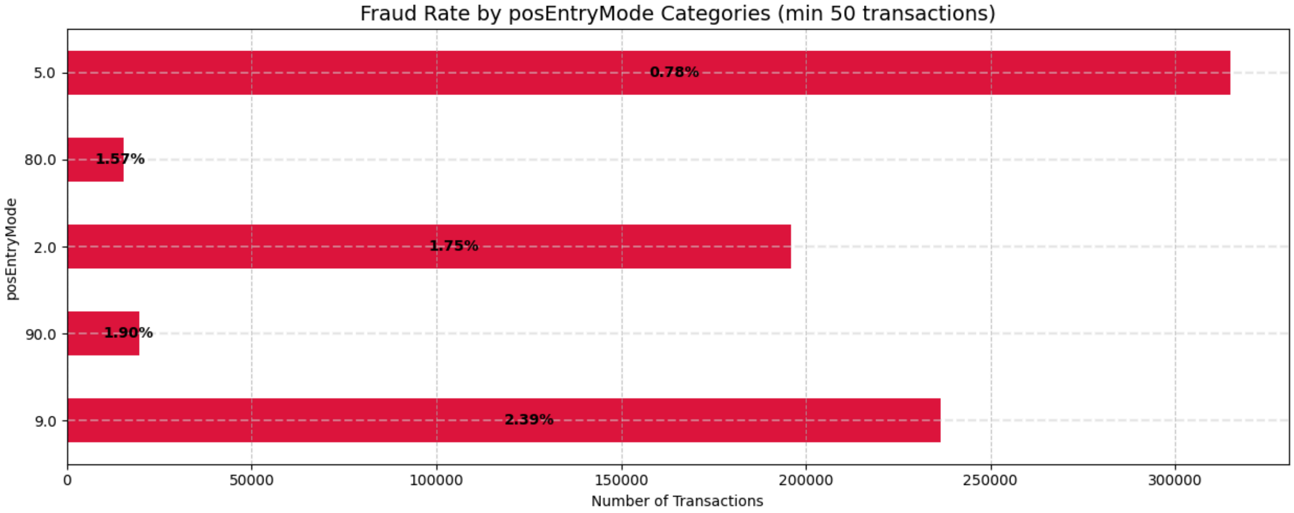
Summary

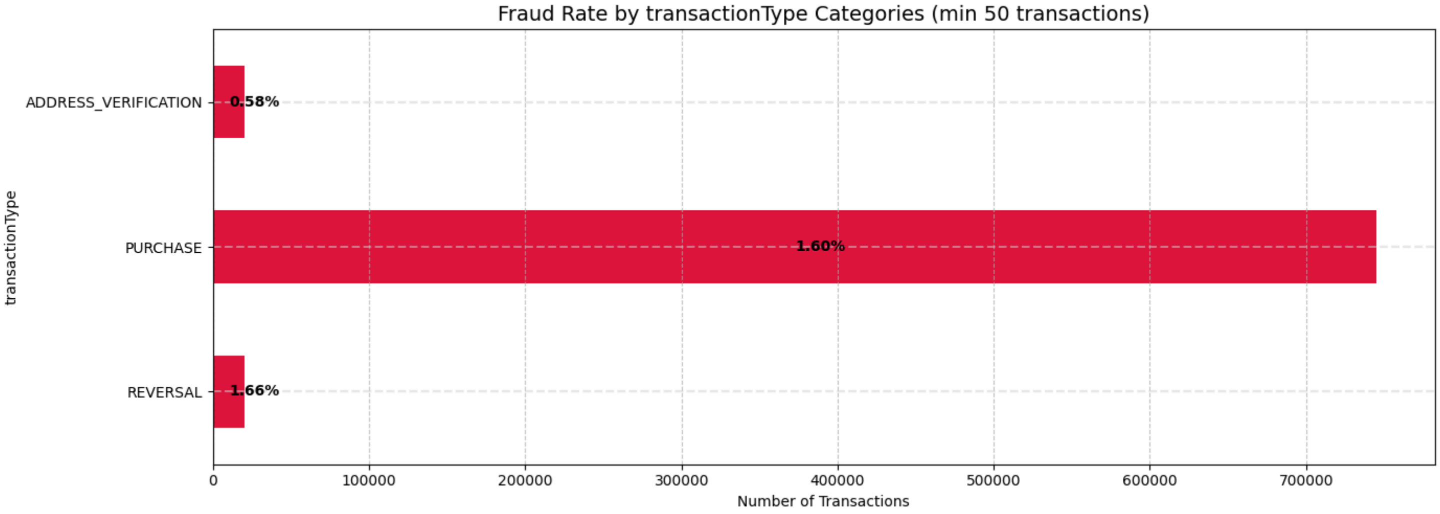
My findings strongly suggest that CVV verification is a more powerful fraud indicator than expiration date matching, and that combining both security features provides the strongest fraud prevention. This highlights the importance of implementing multiple security checks during transaction processing.

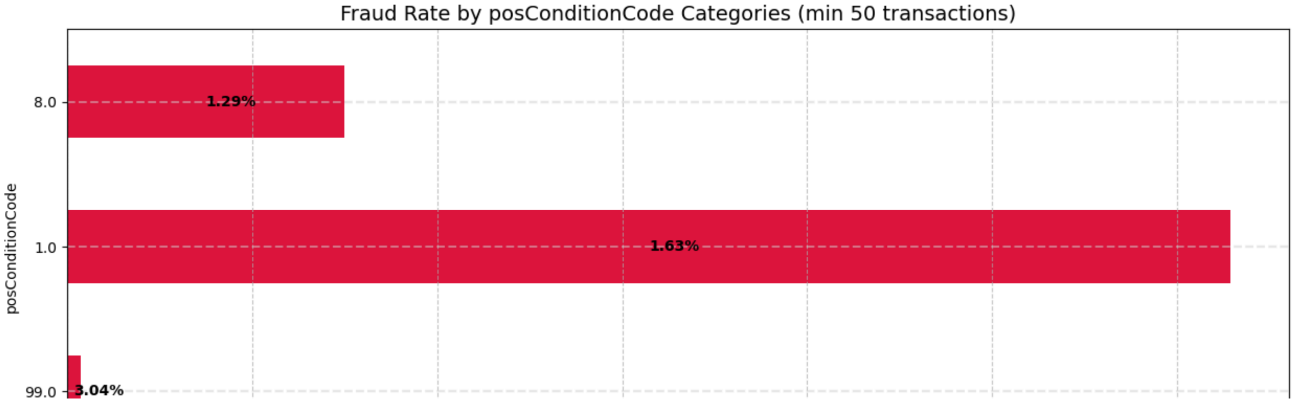
4.3 Categorical Analysis

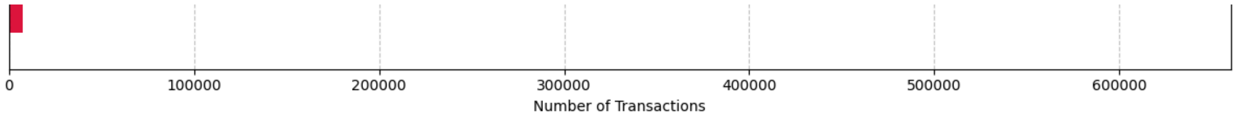
For the categorical variables, I only select the categories which have over 50 transactions in total for the statistical significance purpose and this help to better illustrate the main trend.

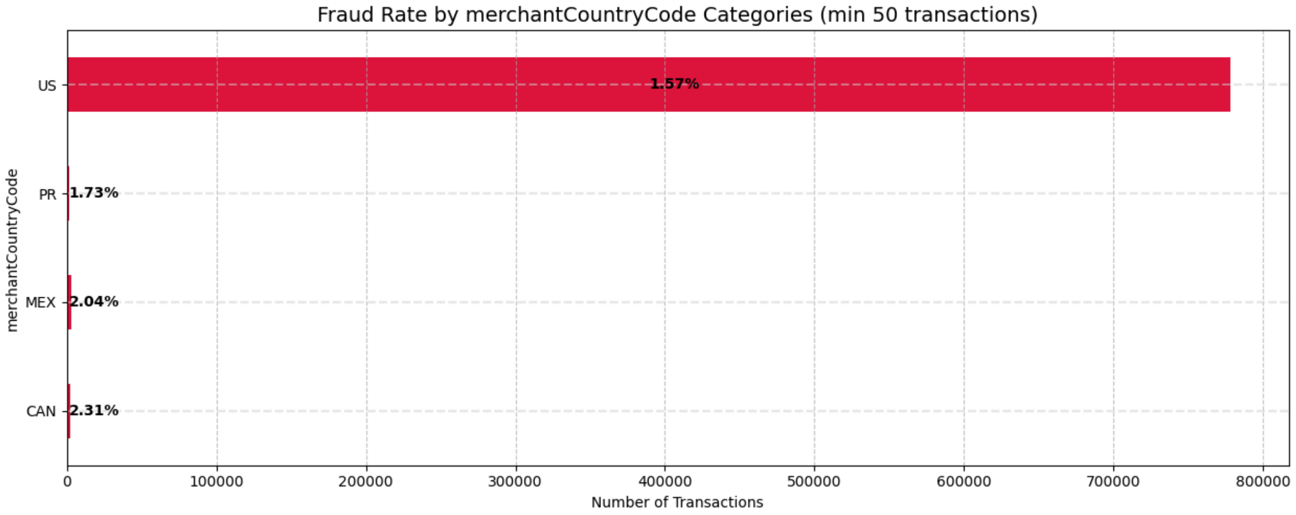

Fraud Rate by merchantCategoryCode Categories (min 50 transactions)


Fraud Rate by posEntryMode Categories (min 50 transactions)


Fraud Rate by transactionType Categories (min 50 transactions)


Fraud Rate by posConditionCode Categories (min 50 transactions)





The analysis of merchant categories and transaction characteristics revealed critical patterns in fraud distribution:

Merchant Category

- **High-Risk Categories**: Airlines (3.46%), online retail (2.44%), rideshare (2.49%), and online gifts (2.42%) show the highest fraud rates, all exceeding 2.4%.
- **Medium-Risk Categories**: Food (1.34%), entertainment (1.20%), auto (1.26%), and furniture (1.39%) display moderate fraud rates between 1-2%.
- **Low-Risk Categories**: Several categories show minimal fraud activity (≤0.5%), including gyms, cable/phone, online subscriptions, food delivery, fuel, and mobile apps.
- **Transaction Volume Impact**: While online retail has the highest transaction volume (~200,000 transactions), it maintains a high fraud rate (2.44%), suggesting persistent vulnerability despite scale.

Point of Entry

- As is shown in the picture, Mode 9 and 2 has both high volume and high fraud rate and Mode 90 has high volume of transaction and high fraud rate.
- By searching for the meaning, this shows that E-commerce transaction and Basic Magnetic stripe read are two ways most frequently used and have high fraud rate compared with other methods.

Transaction Type

- **Type Variations**: REVERSAL transactions show a slightly higher fraud rate (1.66%) than PURCHASE transactions (1.60%), while ADDRESS_VERIFICATION transactions have a substantially lower fraud rate (0.58%).

- **Volume Considerations**: PURCHASE transactions dominate the dataset with over 700,000 transactions, making their 1.60% fraud rate particularly impactful in absolute terms.

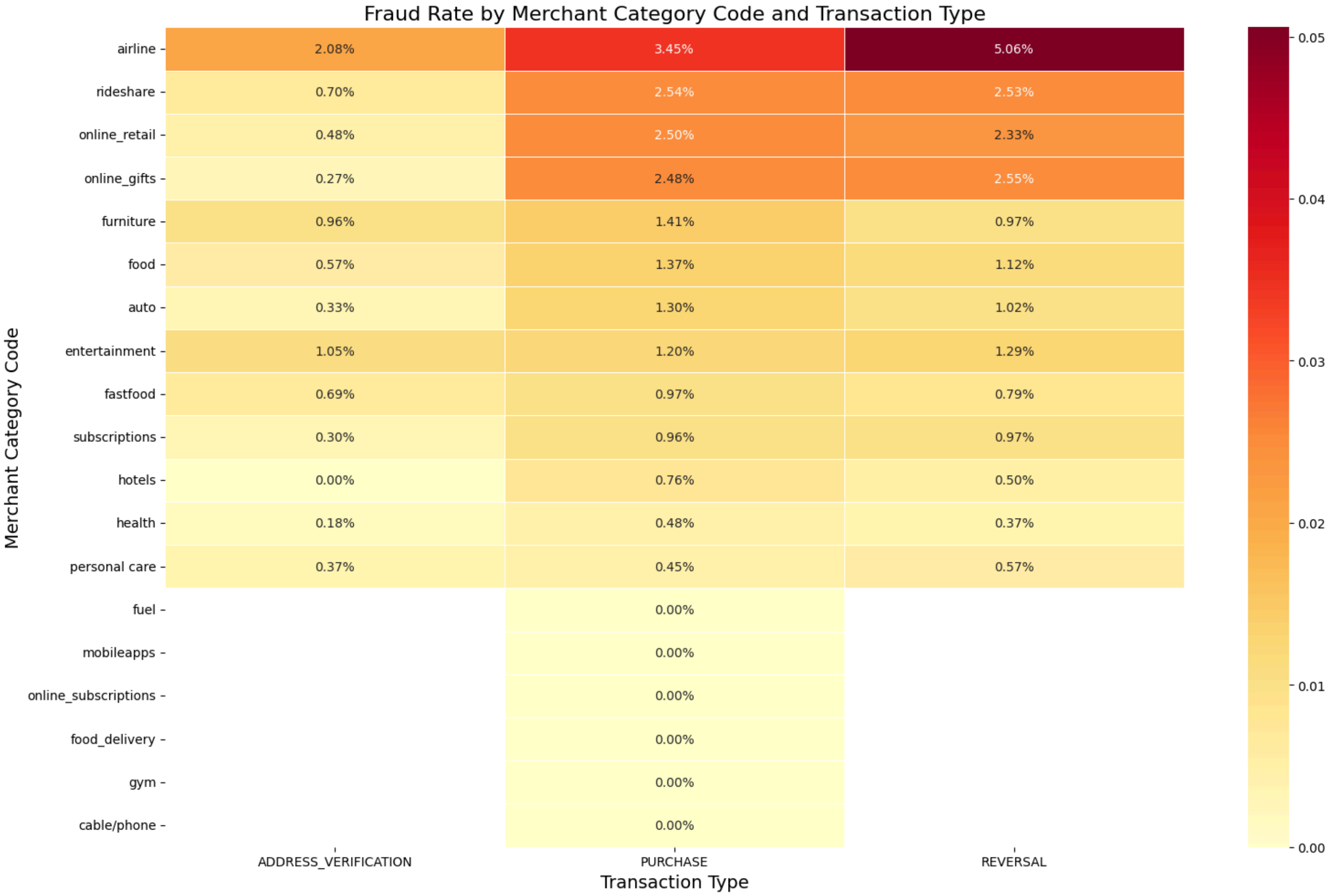
posCondition

- **Condition Code Impact**: POS condition code 1(Cardholder not present) shows the highest transaction volume and a fraud rate of 1.63%.
- **Risk Variations**: Condition code 8(Mail/telephone order (includes Visa phone and recurring transactions)) has a lower fraud rate (1.29%) despite significant transaction volume, while code 99 shows the lowest fraud rate (0.04%) with minimal transactions.

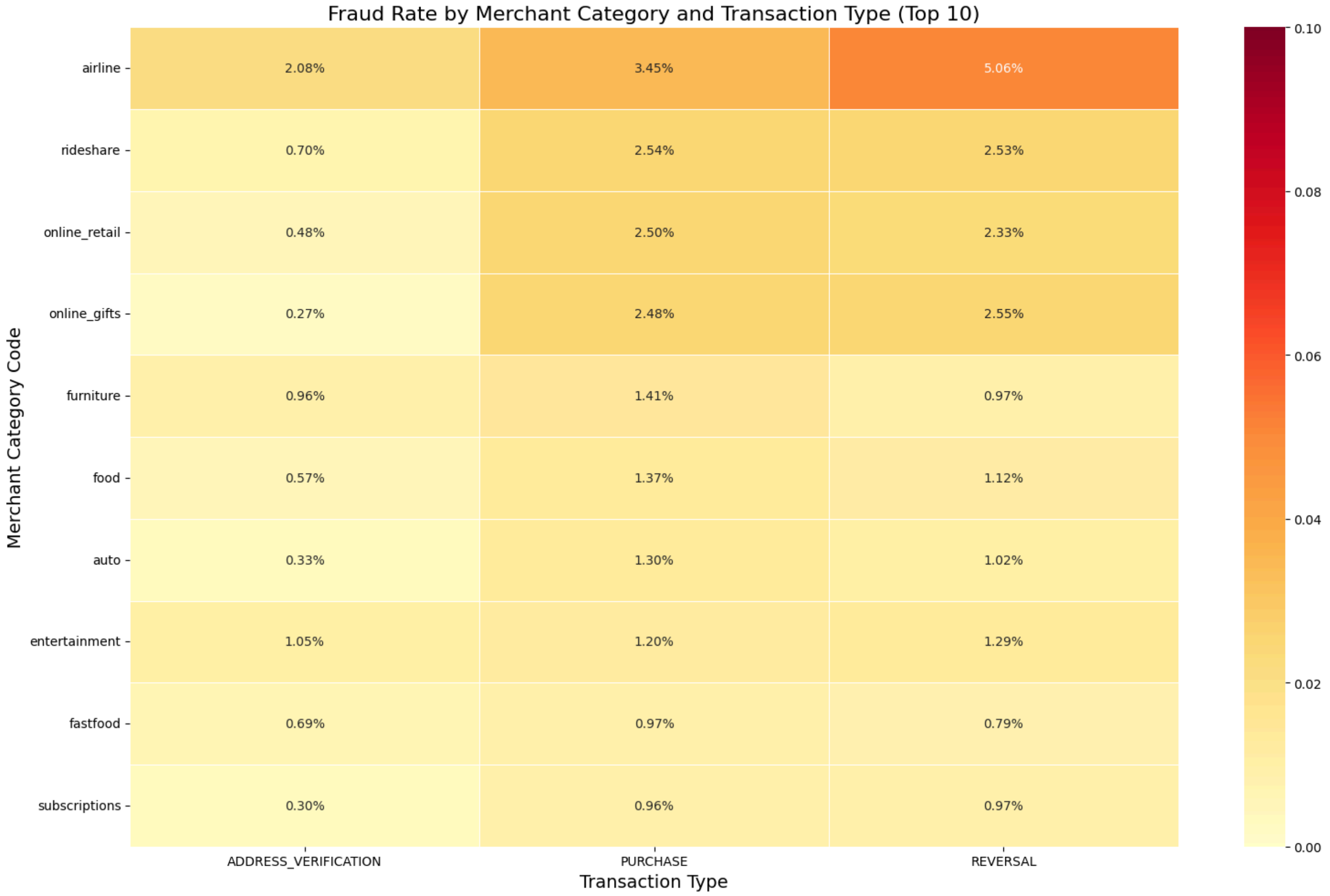
Merchant Country

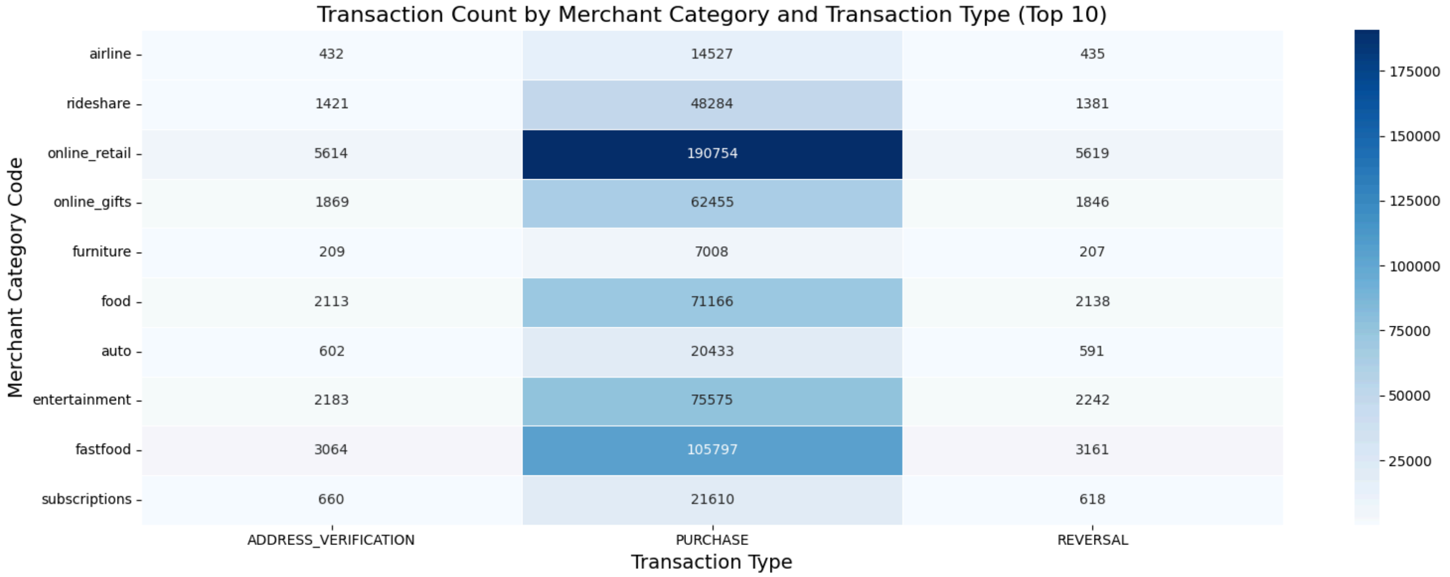
- **Geographic Risk Patterns**: Transactions with Canadian merchants show the highest fraud rate (2.31%), followed by Mexican merchants (2.04%), Puerto Rican merchants (1.73%), and US merchants (1.57%).
- **Volume Considerations**: While US merchants account for the vast majority of transactions, they show the lowest fraud rate among countries analyzed, suggesting better fraud controls or detection systems in the US market.
- **Cross-Border Transactions**: The significantly higher fraud rates in international transactions (CAN, MEX, PR) compared to domestic (US) transactions highlight the elevated risk in cross-border commerce.

Fraud rate by merchant category and transaction type



By plotting the heatmap of fraud rate by merchant category and transaction type, I uncovered further relationships. Then I zoomed in to focus on where the fraud rate is high:





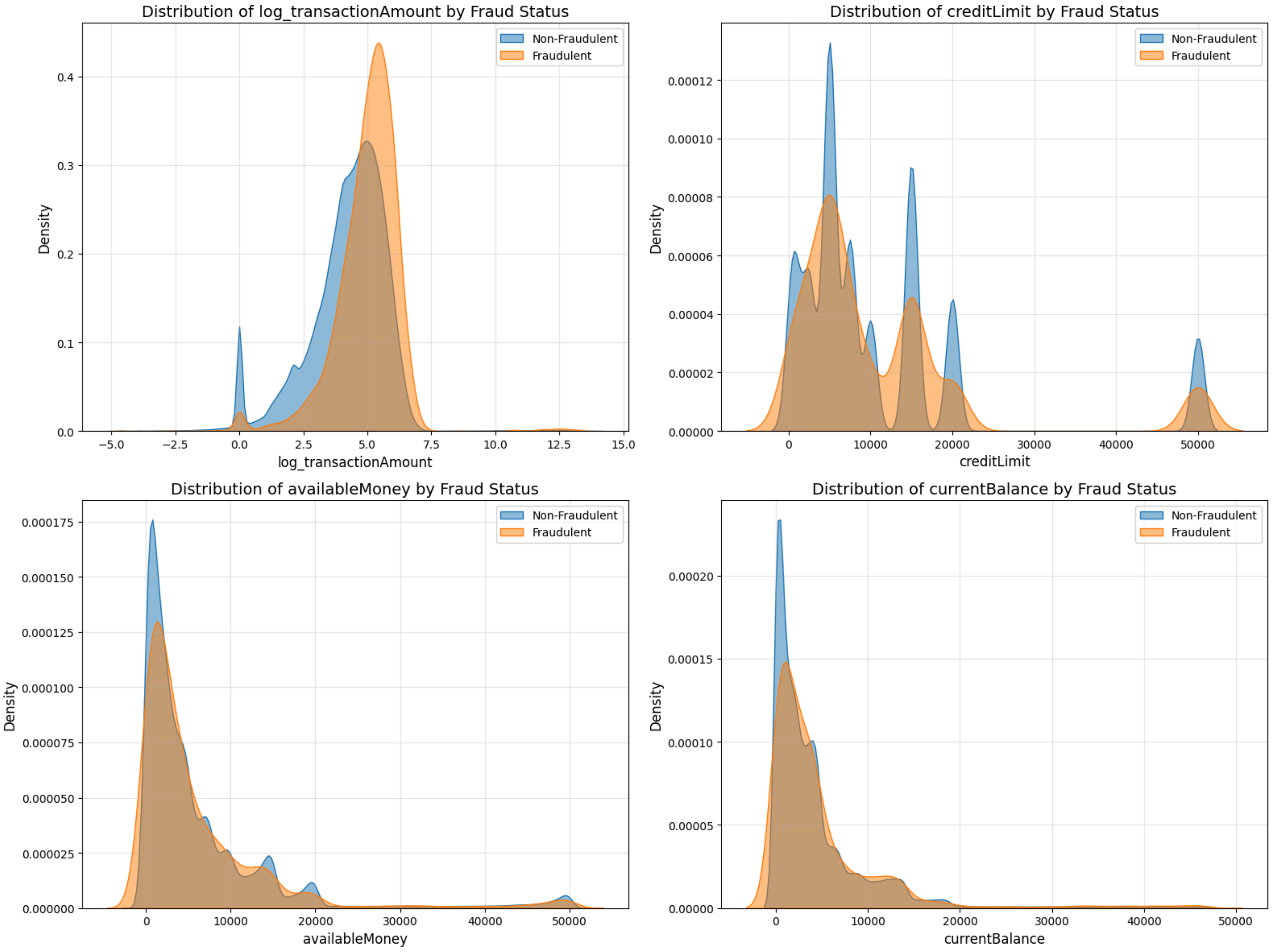
The heatmap analysis reveals that airline transactions have the highest fraud rates (2.08%-5.06%), with REVERSAL transactions being most vulnerable. Online categories consistently show elevated fraud (2.33%-2.55%), while PURCHASE transactions dominate volume (particularly online_retail at 190,754 transactions), making their fraud patterns most impactful despite sometimes lower percentage rates than REVERSALS. This analysis identifies two critical security focus areas: airline REVERSAL transactions and high-volume online retail PURCHASES.

Summary

Categorical analysis reveals distinct fraud patterns across different transaction attributes: high-risk merchant categories include airlines, online retail, rideshare, and online gifts (all exceeding 2.4% fraud rates); E-commerce transactions and basic magnetic stripe reads show both high volume and high fraud rates; REVERSAL transactions have slightly higher fraud rates than PURCHASE transactions; transactions where cardholders are not present show elevated risk; and international transactions (particularly with Canadian merchants at 2.31%) demonstrate significantly higher fraud rates than domestic US transactions (1.57%). These findings highlight specific vulnerabilities in the payment ecosystem that could be targeted for enhanced security measures and monitoring.

4.4 Conditional density plots for numerical variables

The Kernel Density Estimation (KDE) plots for key numerical variables revealed important patterns that differentiate fraudulent from non-fraudulent transactions:


Distribution of log_transactionAmount by Fraud Status

Distribution of creditLimit by Fraud Status

Distribution of availableMoney by Fraud Status

Distribution of currentBalance by Fraud Status

Summary

Specifically for log_transactionAmount, I noticed the density plot for fraud transactions notably shifts right compared to non-fraud transactions, indicating larger amounts are more prone to fraud. This confirms my earlier findings about transaction amount being a significant fraud indicator.

4.5 Multi-Swipe Transaction Analysis

A multi-swipe transaction analysis was conducted to identify potentially suspicious repeated transactions. Transactions were considered multi-swipes if they occurred within 5 minutes, had the same amount, and were from the same account.

```
Overall multi-swipe transaction report:
Multi-Swipe Transaction Analysis:
Conditions: Same account, same amount, within 5 minutes
Total transactions: 786363
Multi-swipe transactions: 13297
Percentage of multi-swipe transactions: 1.69%
Percentage of total dollar amount in multi-swipe transactions: 1.80%
Fraud Analysis in Multi-Swipe Transactions:
Fraudulent multi-swipe transactions: 230
Percentage of multi-swipe transactions that are fraudulent: 1.73%
Overall fraud rate in all transactions: 1.58%
Multi-swipe transactions are 1.10x more/less likely to be fraudulent
```

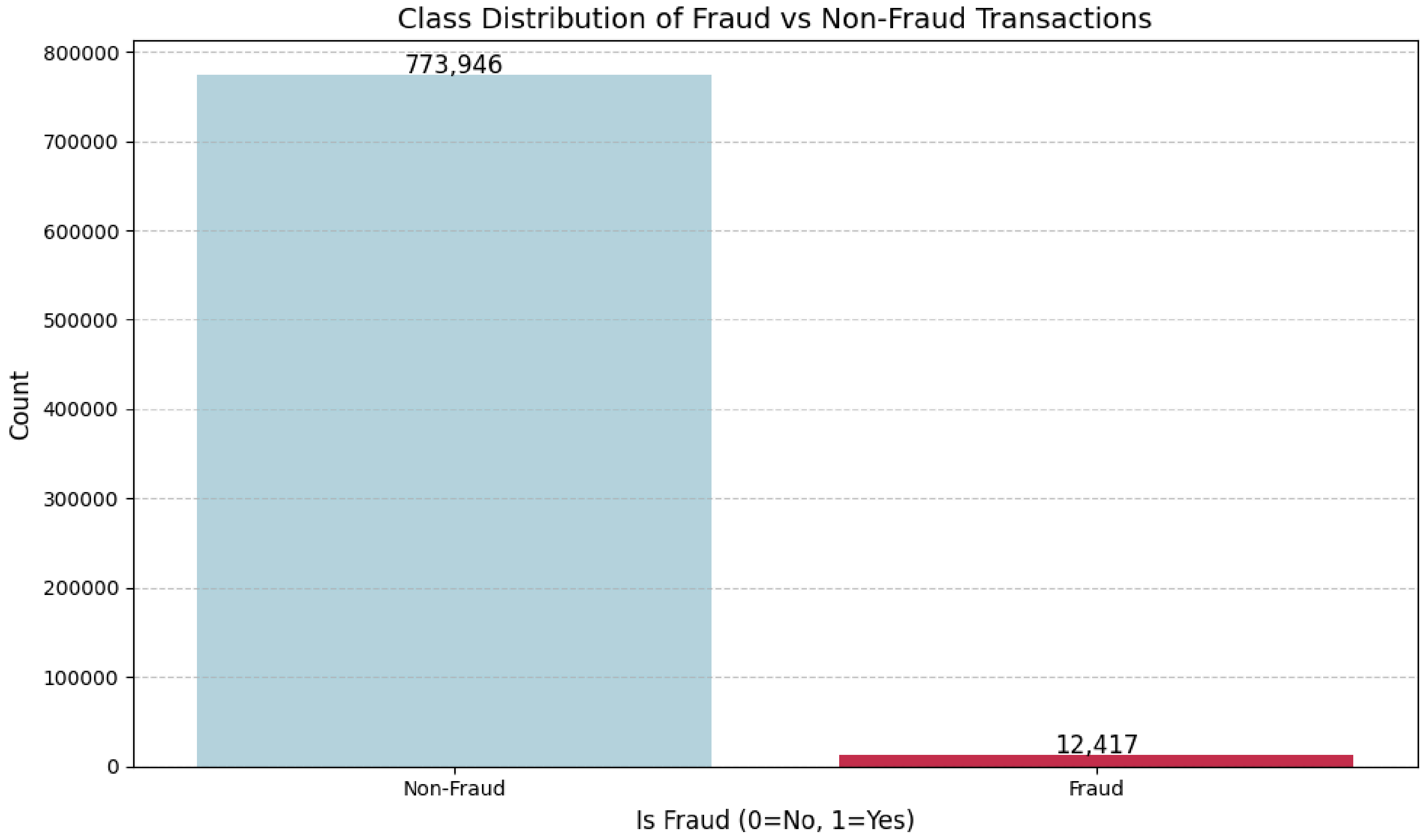
The multi-swipe transaction analysis reveals that out of 786,363 total transactions, 13,297 (1.69%) were identified as multi-swipes—defined as transactions from the same account with identical amounts occurring within a 5-minute window.

Summary

These multi-swipe transactions account for 1.80% of the total dollar volume. Notably, the fraud rate within multi-swipe transactions is 1.73% (230 fraudulent transactions), which is 1.10 times higher than the overall fraud rate of 1.58% in the dataset. This modest elevation in fraud risk suggests that while multi-swipe patterns aren't strongly indicative of fraud, they do represent a subtle risk factor that could signal behaviors such as card testing, transaction splitting to avoid detection thresholds, or legitimate repeated purchases that inadvertently share characteristics with fraudulent patterns.

4.6 Class Imbalance Analysis

The dataset exhibits significant class imbalance, which is common in fraud detection scenarios:



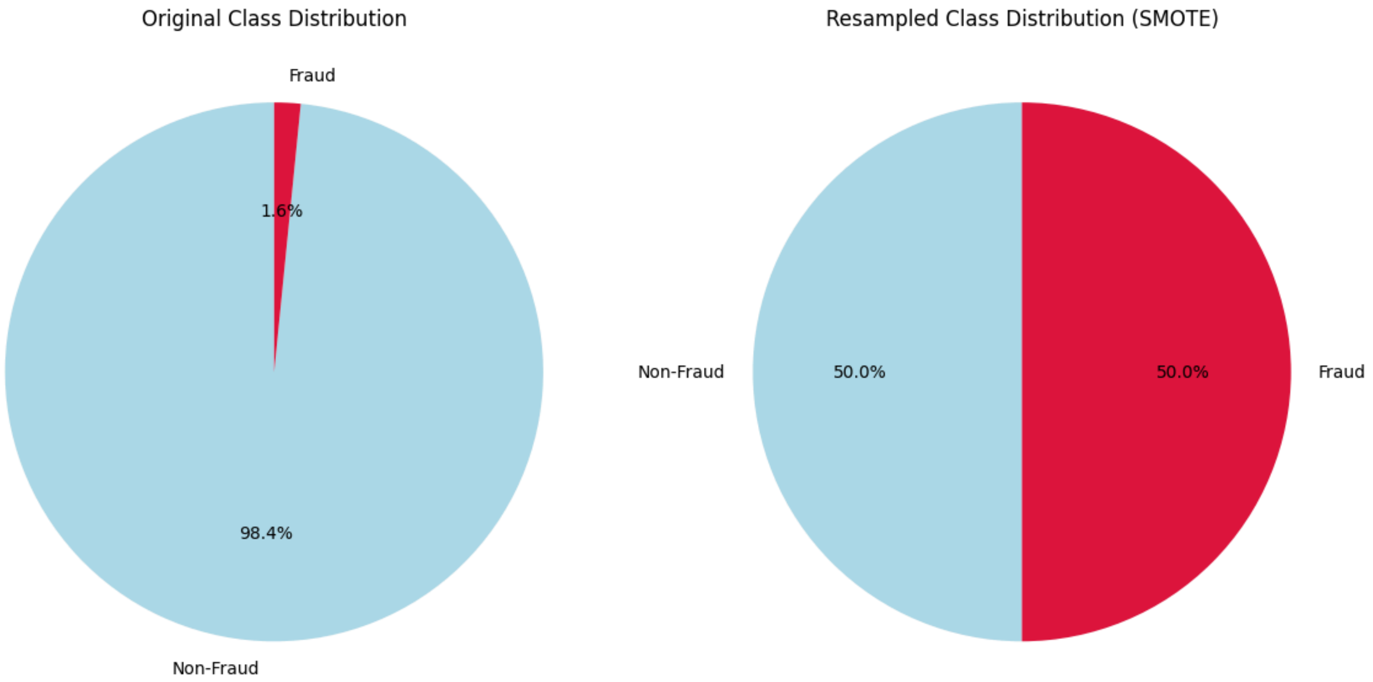
```
- Total transactions: 776,668
- Fraudulent transactions: 11,966
- Fraud percentage: 1.54%
- Imbalance ratio: 1:65
```

Summary

The bar chart clearly illustrates the severe imbalance between the classes, with the fraudulent transactions barely visible when plotted on the same scale as non-fraudulent ones. This severe imbalance presents challenges for modeling, as most standard algorithms tend to favor the majority class (non-fraudulent transactions) at the expense of correctly identifying the minority class (fraudulent transactions).

Class Imbalance Mitigation using SMOTE method

To address the class imbalance issue, Synthetic Minority Over-sampling Technique (SMOTE) is an approach that is easy to implement.



- Original class distribution: 98.46% non-fraud, 1.54% fraud
- Resampled class distribution: 50% non-fraud, 50% fraud

Summary

The pie charts illustrate the effect of applying SMOTE, transforming the tiny slice representing fraudulent transactions into an equal half of the dataset. SMOTE effectively balanced the class distribution by generating synthetic examples of the minority class (fraudulent transactions), which can help improve model performance, particularly in terms of recall for the fraud class.

5 Summary

I analyzed credit card fraud patterns in a dataset of 786,363 transactions.

After preprocessing (removing missing values and transforming date fields), several key fraud indicators is revealed: newer accounts (under 3 years) and recently changed addresses (within 30 days) have higher fraud rates; fraud peaks during midday hours (11am-1pm); higher-value transactions ($1000-$2000 range) show increased fraud likelihood; international and card-not-present transactions demonstrate substantially higher risk; and security feature mismatches (especially CVV) strongly correlate with fraud. The analysis also identifies high-risk merchant categories (airlines, online retail, rideshare) and transaction types (REVERSAL transactions).

With only 1.54% of transactions being fraudulent, the dataset shows significant class imbalance, which could be addressed using SMOTE to create balanced training data for fraud detection models.

6 Future Work

For the fraud detection problem, SMOTE may create synthetic samples in feature space that may not represent realistic fraud patterns. Also it may not work well with highly dimensional data typical in fraud detection. So maybe we should try ensembling method which can deal with this and use metrics like AUC, Recall rate in the future.