# Explanatory Data Analysis On Transaction Dataset

Bishnu Datt Badu

This project involves a comprehensive Exploratory Data Analysis (EDA) on a transactional dataset to uncover patterns in customer behavior and identify potential fraud indicators. The analysis aims to evaluate data quality, understand spending trends, and highlight anomalies across various features. Key focus areas include time-based and categorical insights. Visualizations and statistical summaries are used to support meaningful interpretations and drive further investigation.

## Handling Missing Values:

1. Since ”acqCountry” and ”merchantCountrycode” contains same values for each row but there are around 623 rows which have missing values for both columns. Hence these rows are removed.
2. Remaining value for ”merchantCountryCode” is imputed with the help of ”acq-Country” and vice-versa
3. Missing ”transactionType” values are imputed by most frequent type i.e ”PUR-CHASE”
4. Rows containing missing values for ”posEntryMode” and ”posConditionCode” are removed.

## EDA Questions

1. What are the most common mechants and merchantCategories by transaction count and amount?

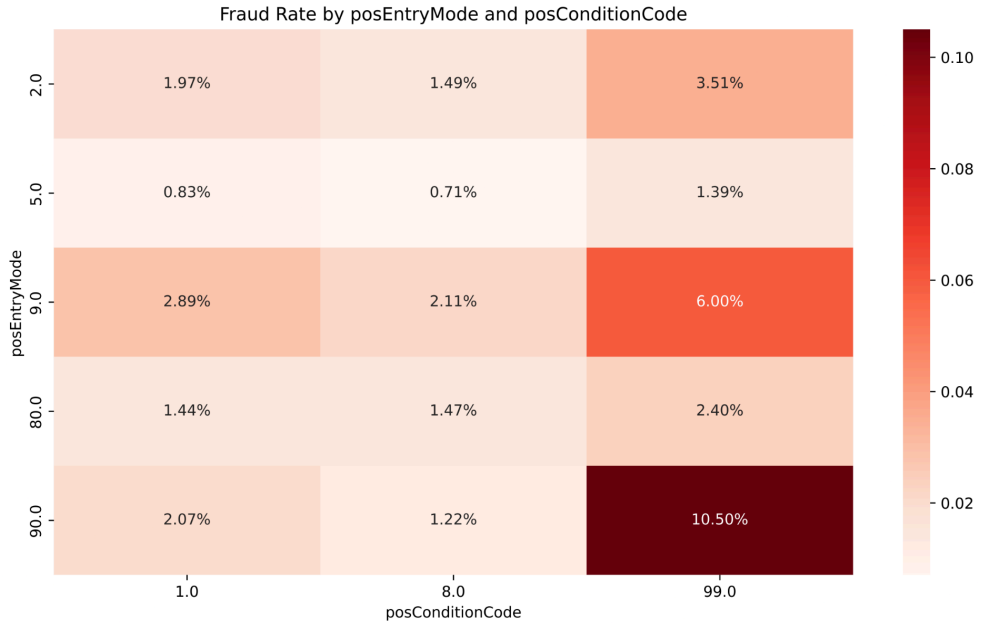
=> The top four merchant brands based on both transaction count and average transaction amount are **AMC**, **EZ Putt Putt**, **Lyft**, and **Uber**. The leading merchant categories by transaction count are **online_retail**, **fastfood**, **entertainment**, and **food**, whereas by average transaction amount, **entertainment** ranks above **fastfood**.

2. Is there a difference in spending behaviour between fraudulent and non-fraudulent transactions?

=> The average transaction amount for fraudulent transactions is quite higher (232.37) than that of non-fraudulent transactions (133.48). It means **fraud targets large transactions.**

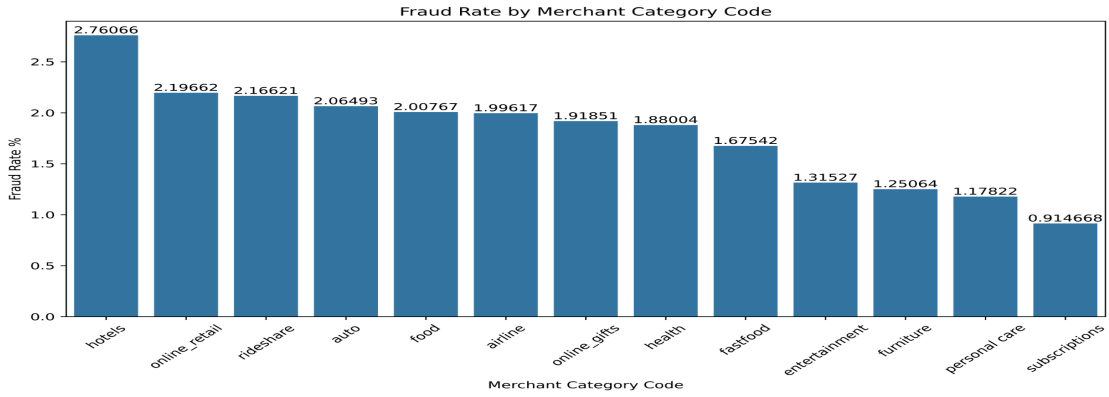
3. Which posEntryMode and posConditionCode combinations are more prone to fraud?

=> Combination of 90.0 posEntryMode and 99.0 posConditionCode is more prone to fraud(10.5% of fraud transactions among their total combined transactions).



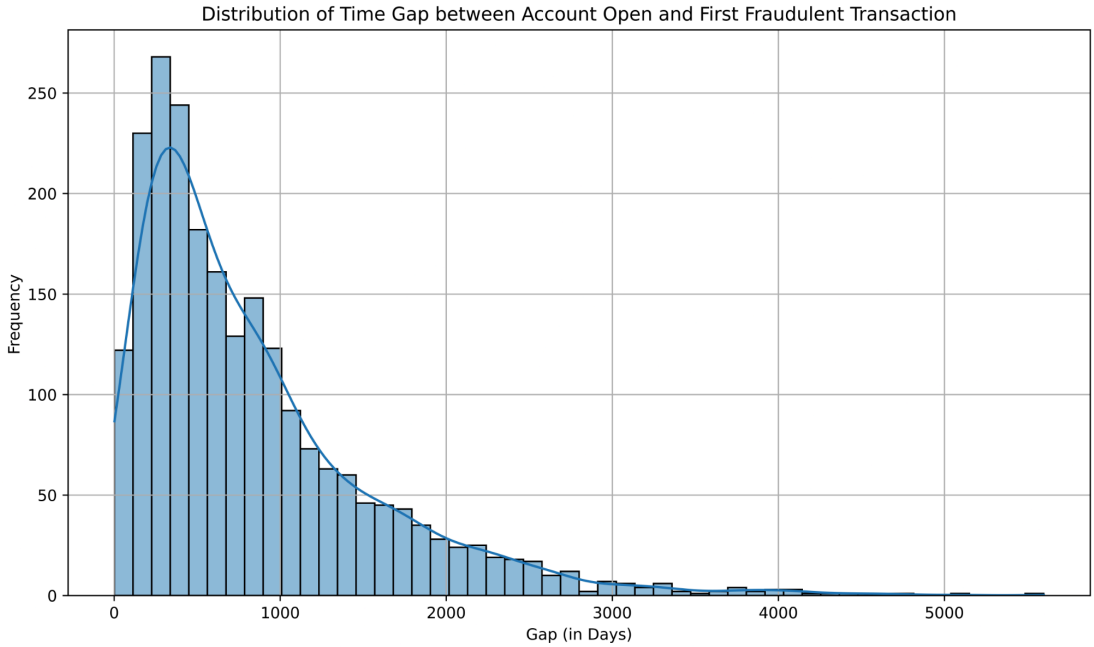

4. Are there specific merchant categories (merchantCategoryCode) that have higher fraud rate?

=> Although, there is no that much difference of fraud rate between merchant categories. Hotel is the common merchant category for fraud transactions followed by online_retail.




5. What is the average time gap between accountOpenDate and transactionDateTime for fraudulent transactions?

=> The average gap between account opening and first transaction is 840.89 days, but the median is 621 days, indicating a right-skewed distribution influenced by extreme outliers (up to 5599 days). After capping the top 25% of outliers, the adjusted average gap drops to 728.65 days, providing a more representative view of user behavior.




6. Do mismatches in cardCVV vs. enteredCVV and currentExpDate vs. expirationDateKeyInMatch correlate with fraud?

=> **CVV Mismatch**

Fraud rate when CVV matches: 1.70%

Fraud rate when CVV mismatches**:** 3.18%

There is a notable increase in fraud rate when CVV entered does not match the one on record. This suggests a **significant association between CVV mismatch and fraud,** implying that fraudsters are more likely to enter incorrect CVV codes.

(Result from Chi-square test)

**Expiration Date Mismatch**

Fraud rate when expiration date matches: 1.35%

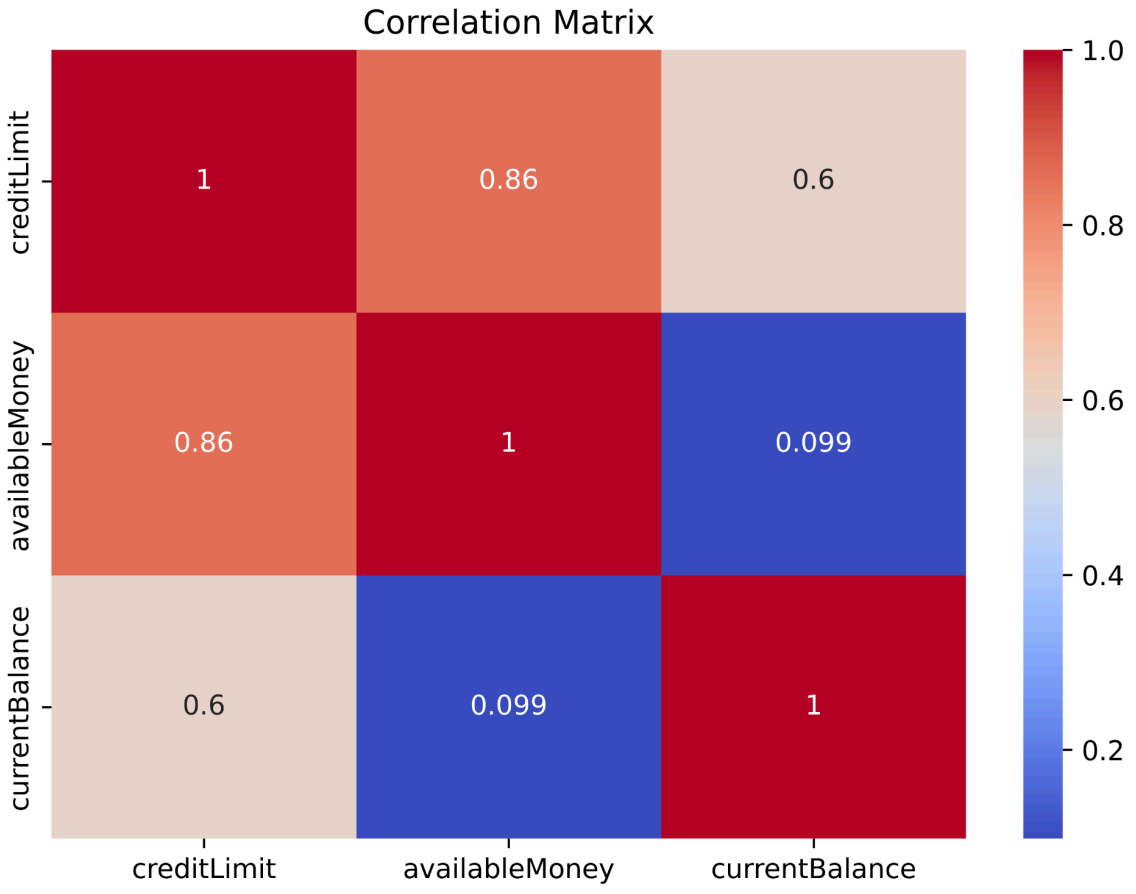
Fraud rate when expiration date mismatches: 1.72%

The difference in fraud rates due to expiration date mismatch is marginal. Therefore, **there appears to be no significant association between expiration date mismatch and fraud.**

7. Is there a correlation between creditLimit, availableMoney, and currentBalance?

=> There is positive correlation between creditLimit and availableMoney, indicating that accounts with higher credit limits tend to have more available funds. Similarly, a positive correlation is observed between creditLimit and currentBalance.

However, there is negligible correlation between availabeMoney and currentBalance. It means these balances may not move together consistently.



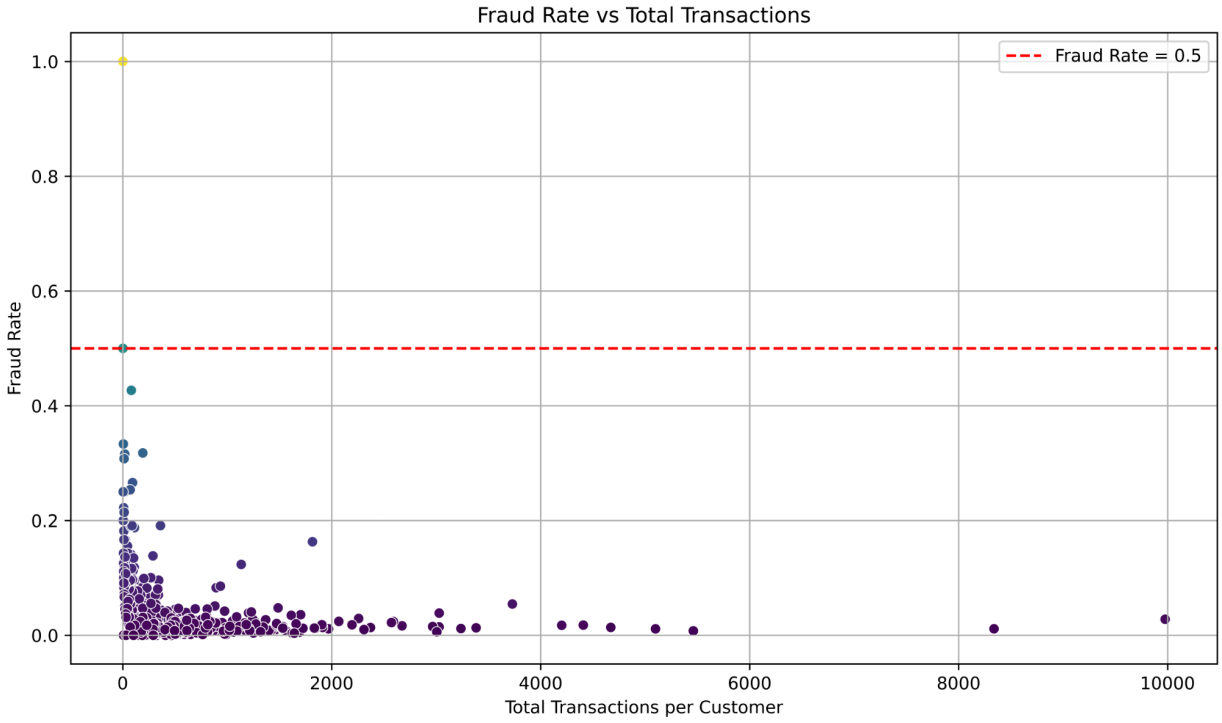

8. Do fraudulent transactions tend to have a certain transactionType or cardPresent flag?

=> There is no significant flag based on transaction type for fraud transactions, the fraud rate is : ADDRESS_VERIFICATION(0.531%) , PURCHASE (1.748%), and REVERSAL(1.79%).

Presence of card(1.51%) is less prone to fraud compared to no card(1.89%).

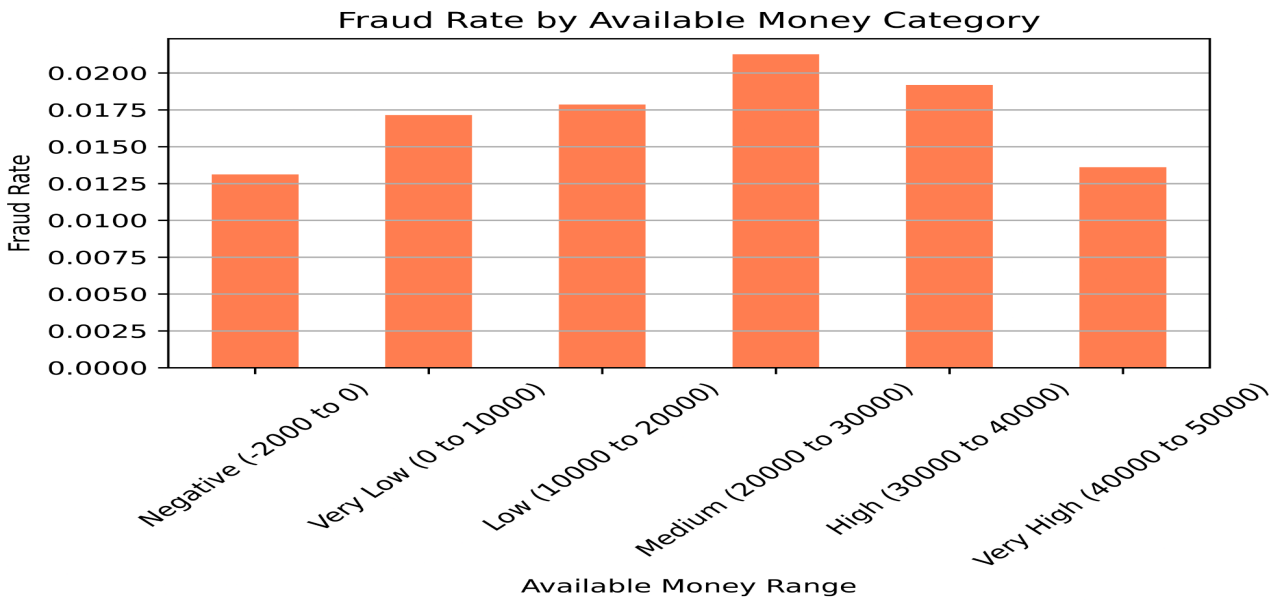
9. Are there repeated Frauds from the same customerID or accountNumber? What are their patterns?

=> Most of the fraud transactions are done by the accountNumber having less than overall 2000 transactions. The plot shows it all.




10. Is there a pattern between availableMoney and fraud?

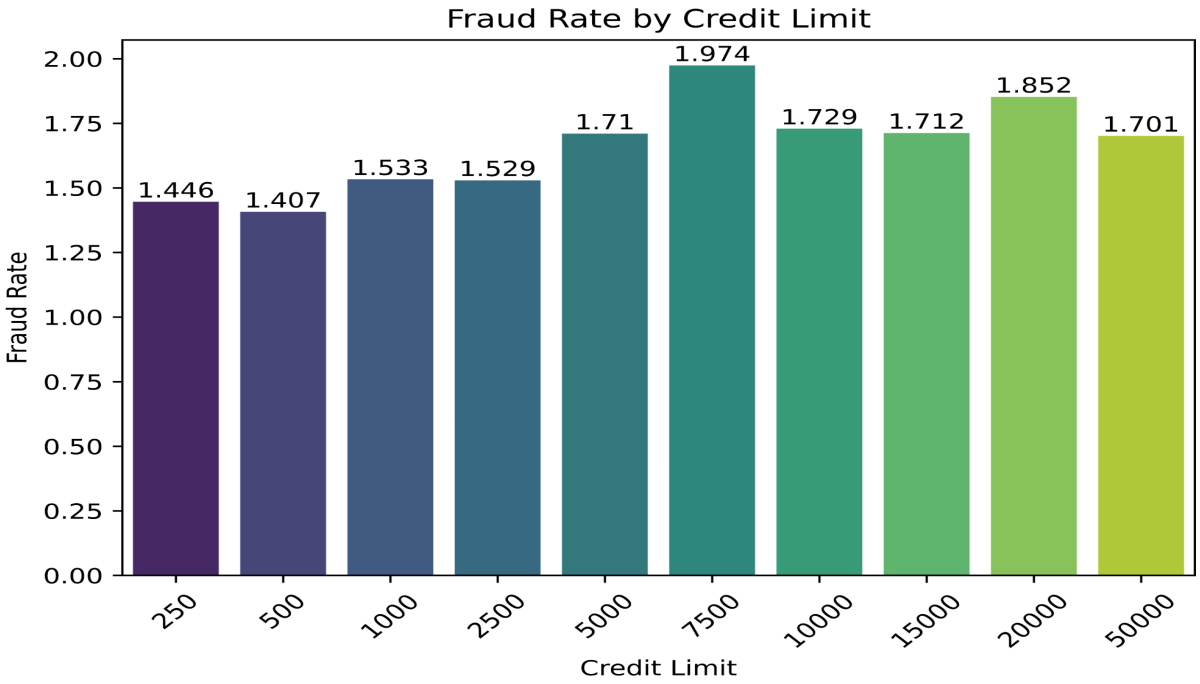
=> There are negative values in availableMoney, but there are only 34 fraud transactions associated with fraud transactions.
Medium amount (20000 to 30000) has higher fraud rate in comparison with other segments i.e 0.212%. There is no significant difference between fraud and non-fraud transactions based on availableMoney. (Mann-Whitney U Test)




11. Are customers with high credit limits more/less likely to face fraud?

=> Almost all creditLimit classes have close fraud-rate among which credit limit class 7500 has the highest fraud rate of 1.974%.

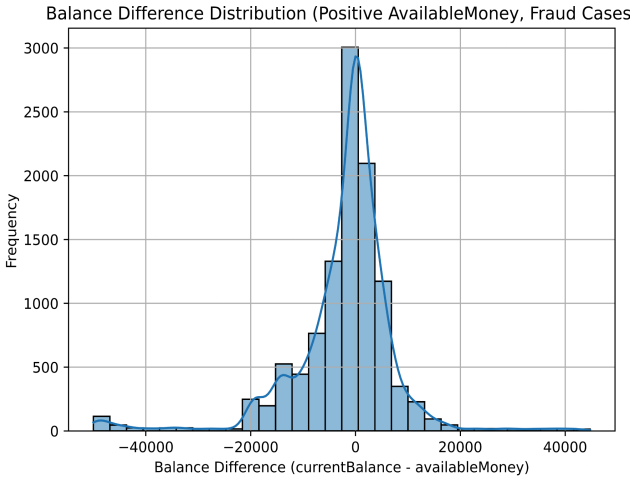
Chi-square test shows a significant association between creditLimit and fraud.

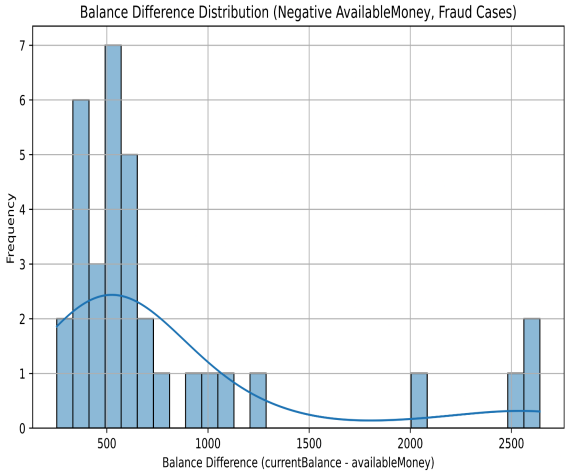


12. What is the average balance difference (current vs available) in fraudulent cases?

=> Fraud with negative available money often involves moderate balance differences, hinting at controlled fund misuse where as with positive available money, fraud patterns are more varied, often involving small manipulations.

A few extreme outliers may indicate larger or more deliberate fraud attempts.

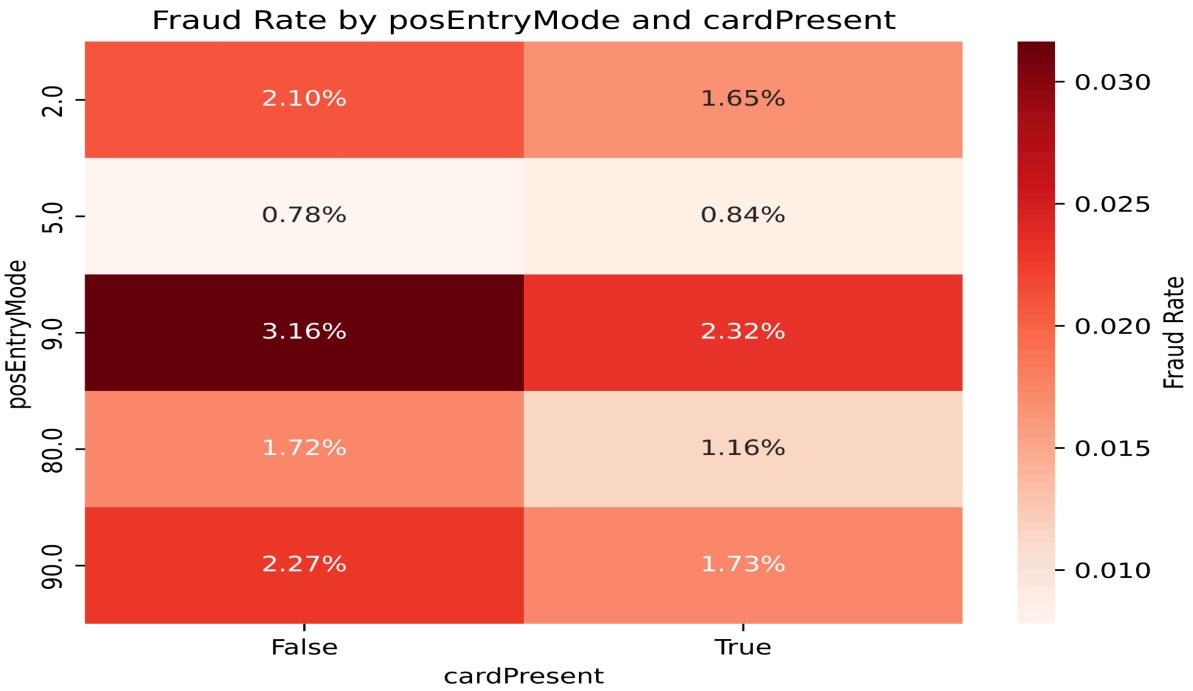



13. Does posEntryMode + cardPresent combo indicate more fraud?

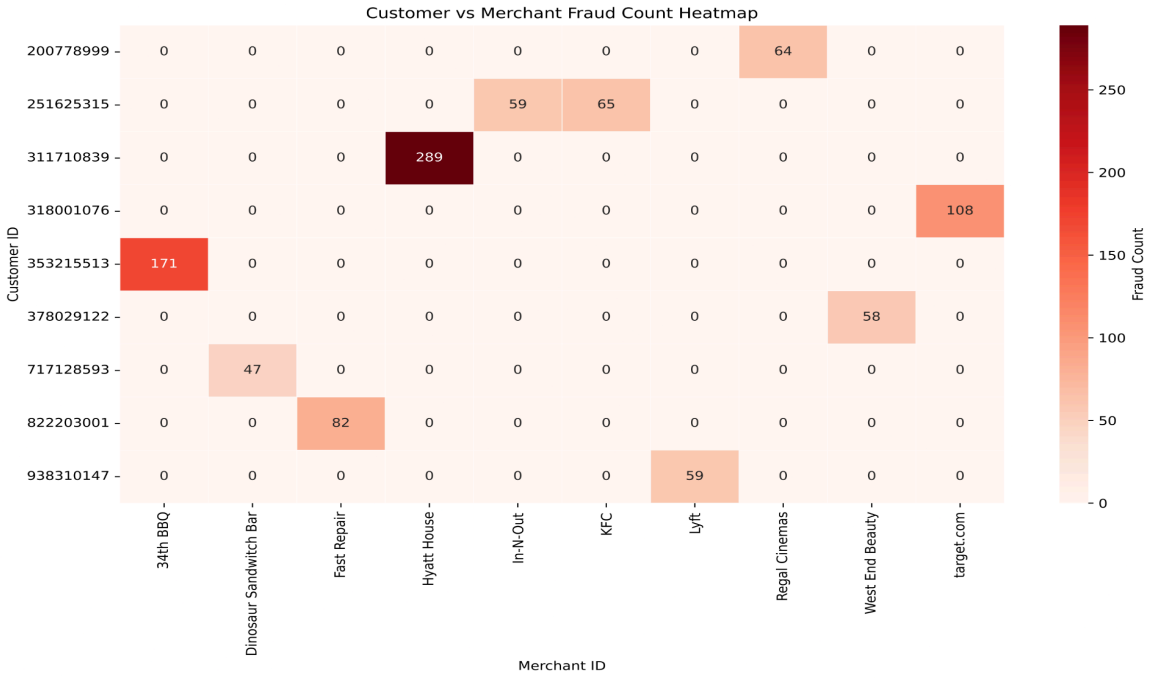
=> New feature is created by combining posEntryMode and cardPresent named as “pos_card_combo’ . The combination of 9.0 pos Entry mode with absence of card (3.16%) is more associated with fraud transactions followed by same pos entry mode but presence of card(2.32%). Overall We can declare posEntryMode 9.0 more active for fraud transactions.

Performing chi-square hypothesis test between ‘pos_card_combo’ and ‘isFraud’, the test shows a significant association between those column, denoting the feature good indicator for modeling.




14. Do certain customers get frauded more on specific merchants?

=> It is found that customerID (311710839) with Hyatt House merchant has almost all fraud transactions (289). SImilarly customerID(353215513) with BBQ also had all fraud transactions (171).

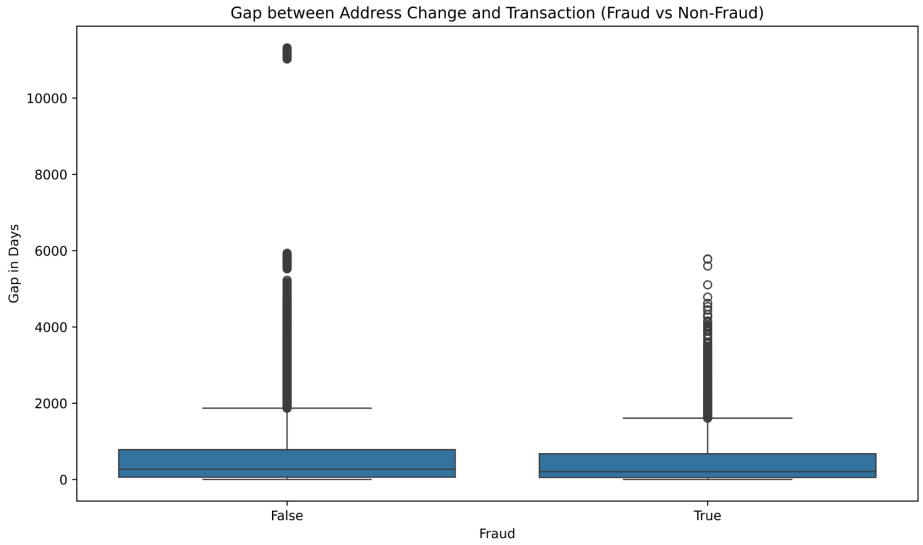


15. Is there a pattern between customer balance and transaction type in fraud cases?

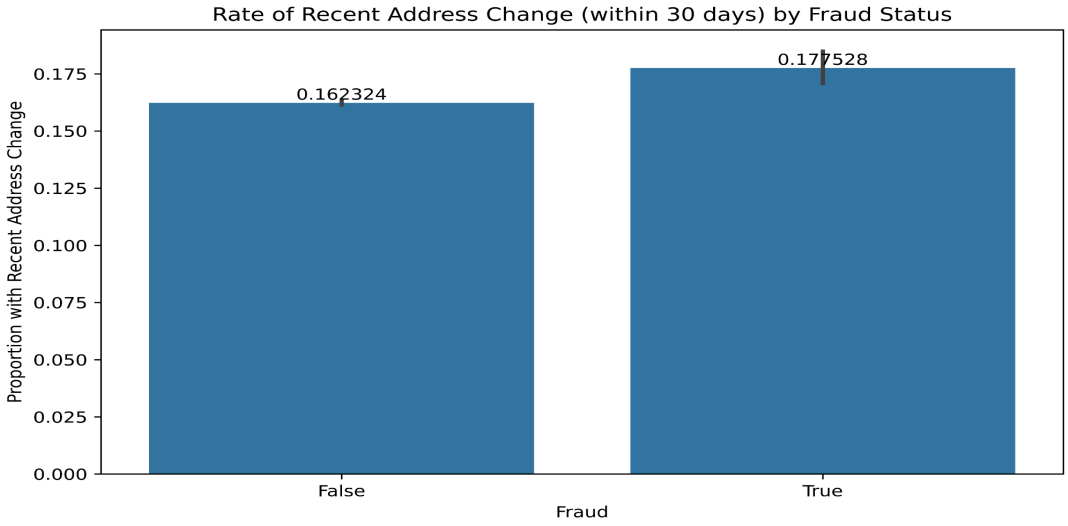
=> No Interpretation idea

16. How does the date of last address change associate with fraud transactions?

=> Non-fraud cases show many outliers with address changes delayed by decades (upto 30 years). Fraud cases tend to have fewer such extremes, indicating more recent address updates.




Fraudsters are slightly more likely to have changed their address within 30 days compared to non-fraudsters. Also the difference is visible, but not massive.

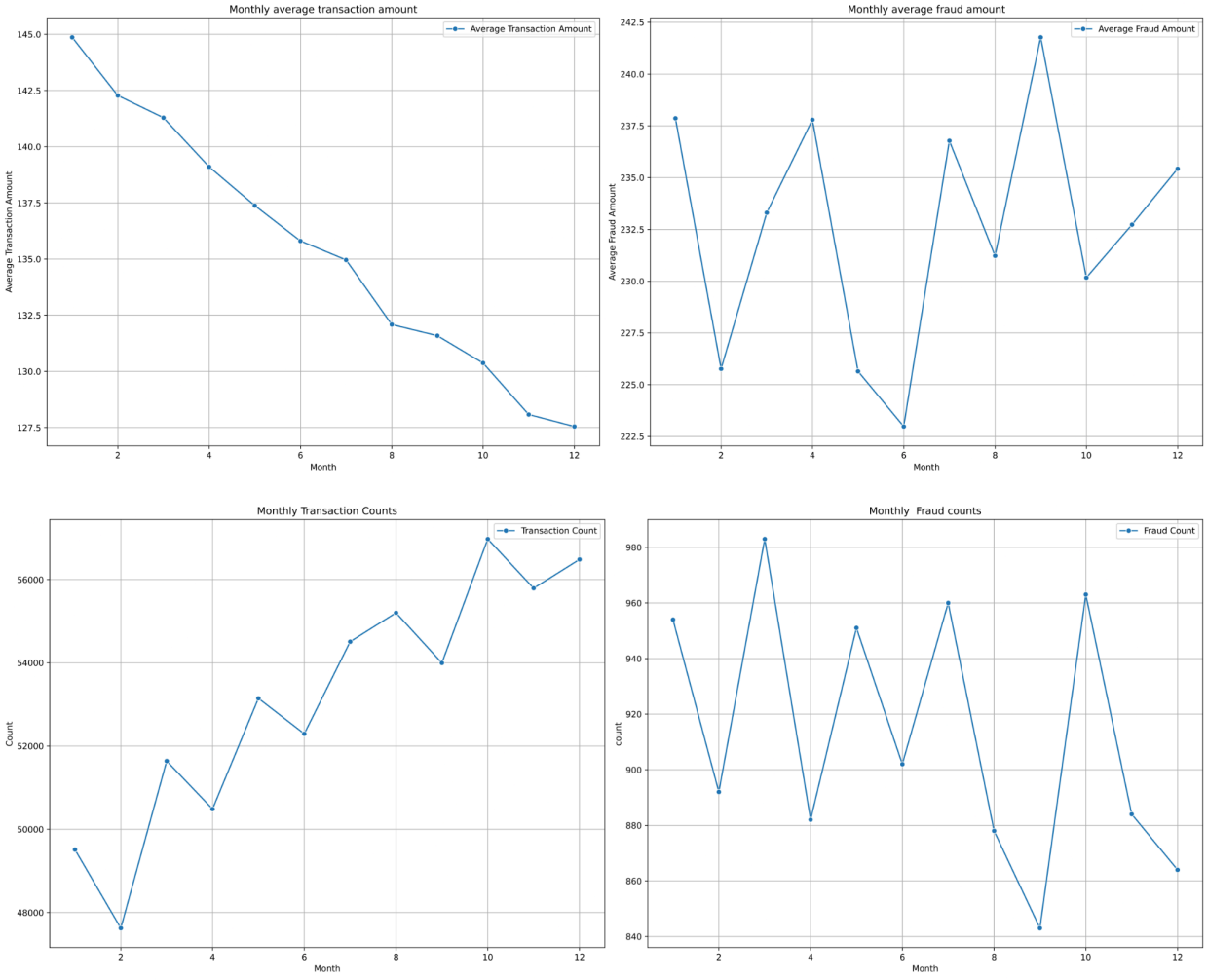


17. What time of day or day of week or which month sees more fraud based on average transaction amount and total transaction count?

=> Based on Month:

- October records the highest total transaction amount and count, while February has the lowest; however, average transaction amount steadily declines over the year.

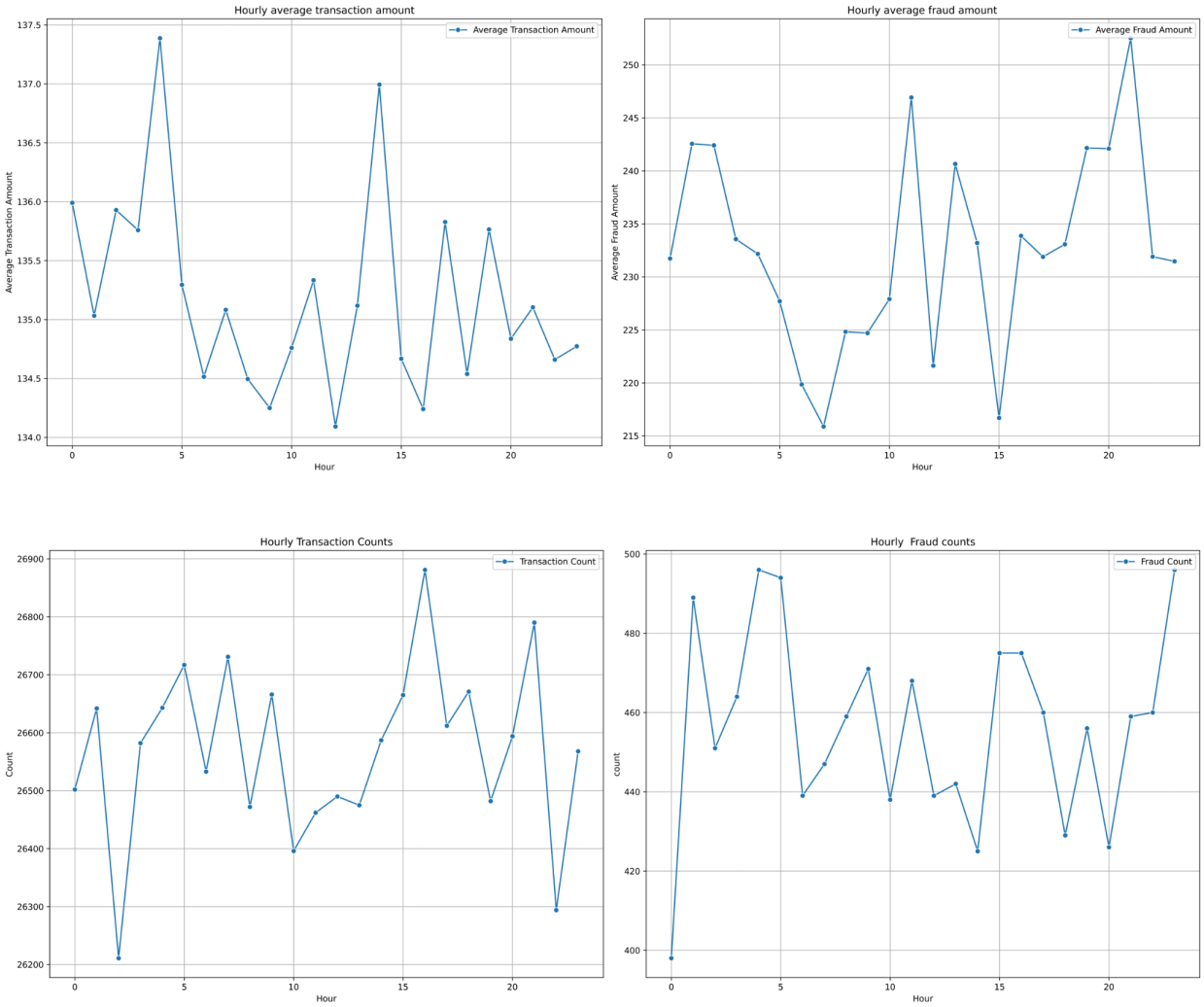
- Fraudulent transactions are evenly spread across months, with March having the most and September the fewest.




Based on Hour:

- Number of transaction is highest at 4 PM evening whereas lowest at 2 AM (morning)

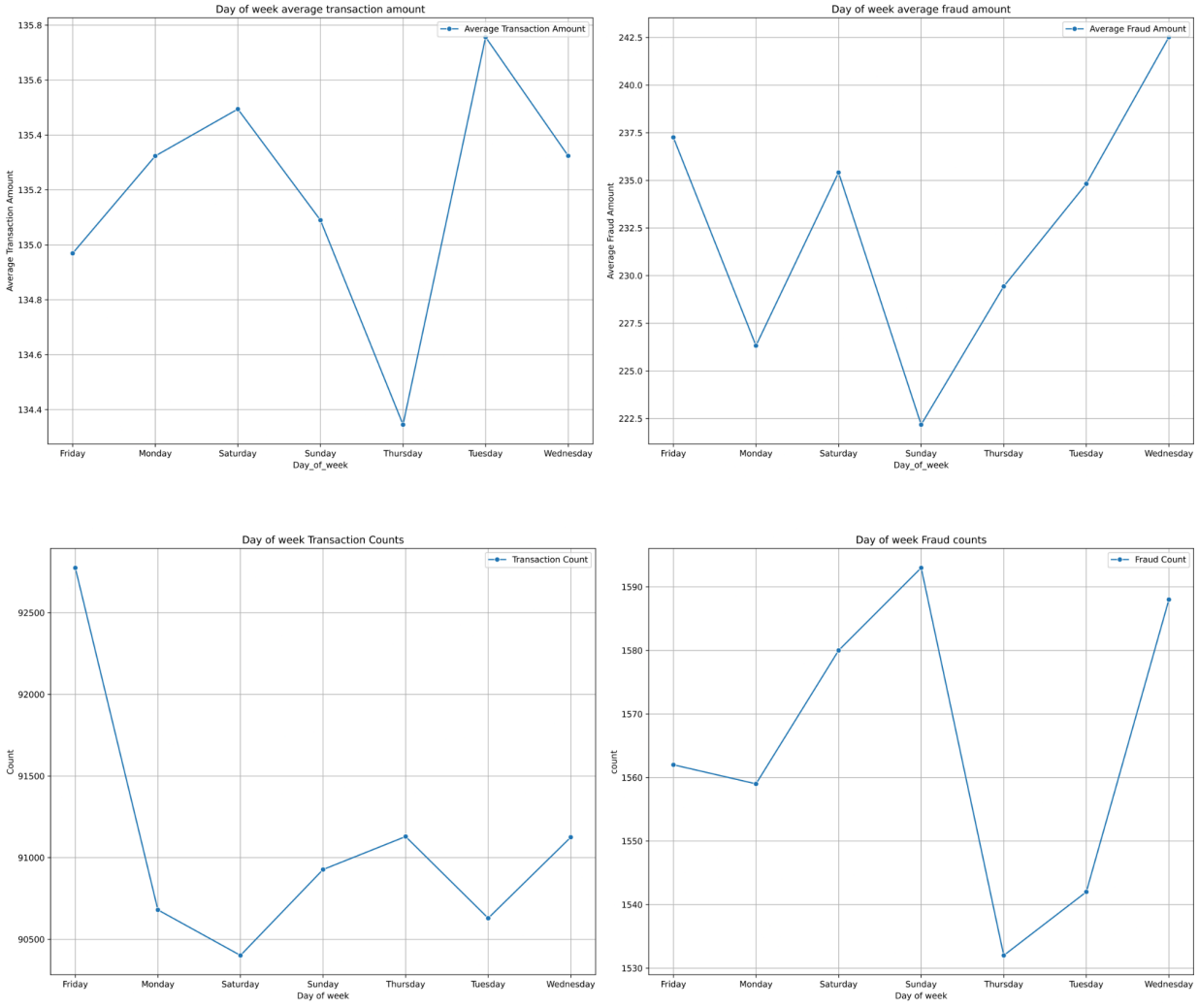
- Average amount of transaction is higher at 4 AM (morning) where as lowest at 12 PM (Noon)
- The count of fraud transactions is between 400 to 500 per hour.
- More fraud transactions take place around 11 pm, 4 am and 5 am whereas there are low fraud transactions around 12 Am.(Comparatively)




Based on day of week:

- The average transaction amount is quite equal for each day, but comparatively Tuesday has the highest value and THursday has the lowest.
- Transaction counts range from 90,000 to 93,000, peaking on Friday and dipping on Saturday.

- Fraud cases are relatively uniform across days, ranging between 1,530 and 1,590, with Sunday recording the highest number of fraud transactions and Thursday the lowest.



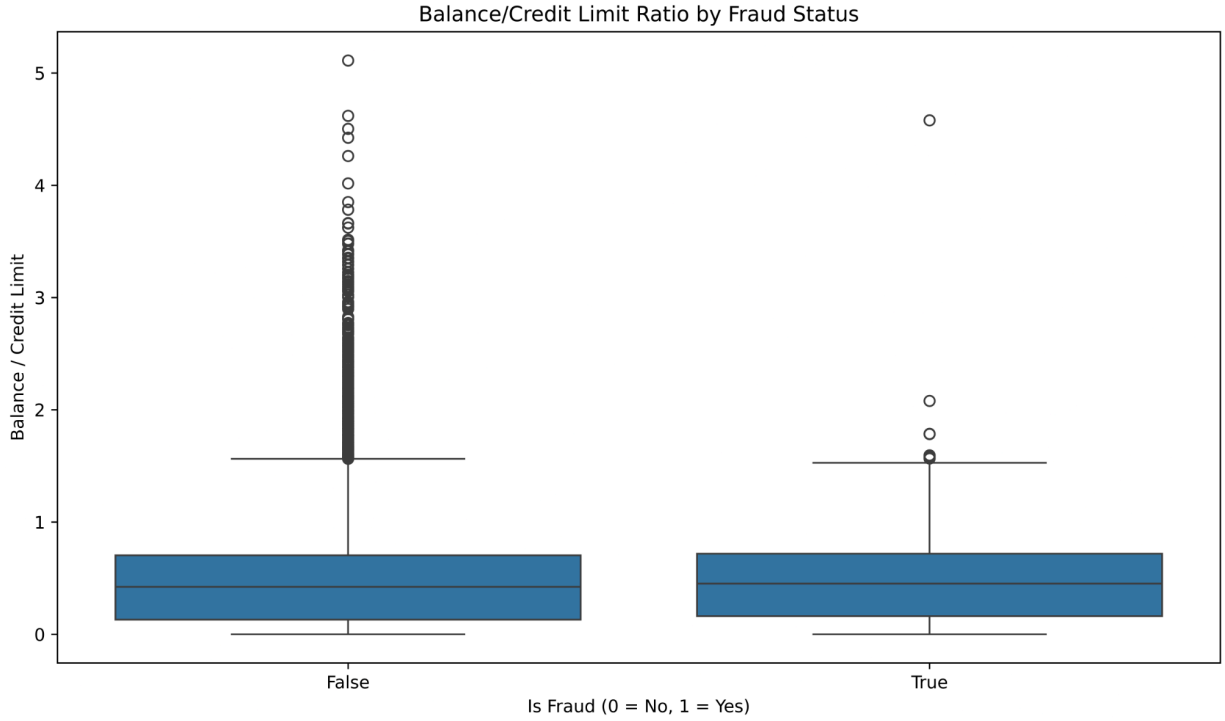

18. Does card last digit having length less than 4 are more prone to fraud?

=> There is no significant difference of fraud rate between credit card last digit having 4 digits(1.717) and not having 4 digits(1.73).

19. Does a high ratio of currentBalance/creditLimit correlate with fraud likelihood?

=> We can consider all transactions having a ratio greater than 2 as non-fraud transactions and it is found that only 2 transactions are fraud having ratio greater than 2.

Performing mannwhitneyu hypothesis testing suggests significant difference between fraud and non-fraud balance/creditLimit ratio distribution which makes it a good variable to modeling.




20. Are very small transactions (e.g., <$1) more likely to be fraudulent?

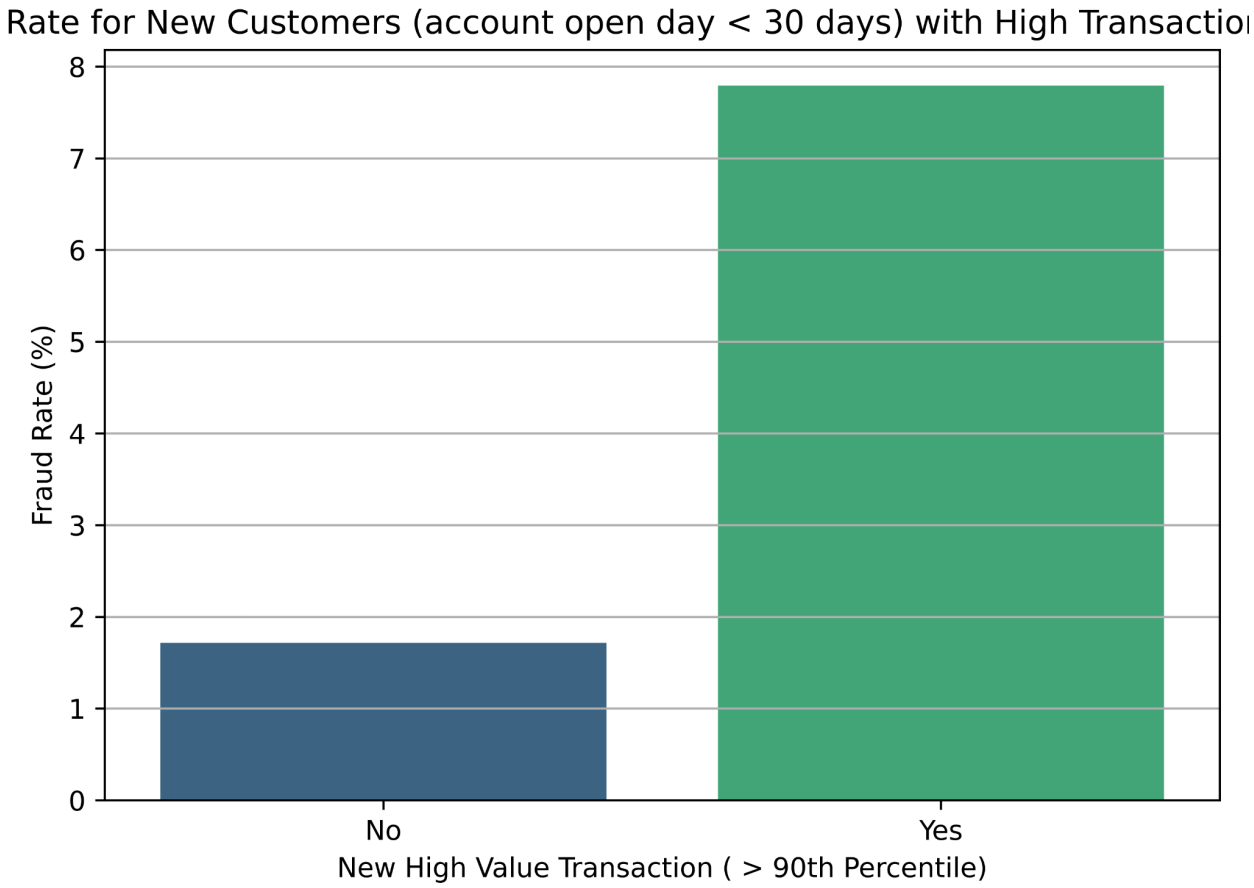
=> The fraud rate for small transactions i.e less than $1 is 0.49%, which means these small transactions aren't likely to be fraudulent.

21. Are new customers (accountOpenDate < 7 days) with high transactionAmount (> 90th percentile) disproportionately targeted?

=> Customers who are new (accounts opened within the last 30 days) and perform large transactions (above the 90th percentile) are 4.5x times more likely to commit fraud compared to others.

No fraud was observed in customers with accounts younger than 7 daus, hence the window was extended to 30 days for meaningful insight.

Although this segment is small in volume(only 154 transactions), the fraud concentration (~7.8%) is notably higher than the overall average.




22. Do customers with a lot of available money have different fraud rates when buying from popular merchants compared to less popular ones?

=> The dataset is filtered on the basis of availableMoney greater than third quartile and then the merchant brands are binned into three categories (“Low”, “Mid”, “high”). New column named ‘Merchant_volume’ is created which shows the category name for each merchant brand. Calculating fraud rates on each category, the fraud rate is found as (Low: 1.08%, Medium: 1.34%, High: 1.77%). There is no significance difference fraud rate per category.
But chi-square test indicates a statistically significant association due to the large sample size. Cramer’s V method is used to test the strength of chi square test and it shows a very weak association between merchant_volume and isFraud(0.0147).

23. Does combination of (acqCountry != merchantCountry) and cardCVV mismatches prone to fraud transactions?

=> A column “risky_combo” is created by combining cross border mismatch and cardCVV mismatch.
The fraud rate associated with risky_combo( True cross border mismatch and True cardCVV mismatch) is 3.20% and contrasting combo has 1.70%. It suggests the combination can be more prone to fraud transactions.

24. Which combination of merchantCategoryCode and merchant Brand is prone to fraud?

=> Since each merchant brand is associated with a single merchant category, we can't decide the propensity (incline to) toward fraud transactions based on the combination of these entities.