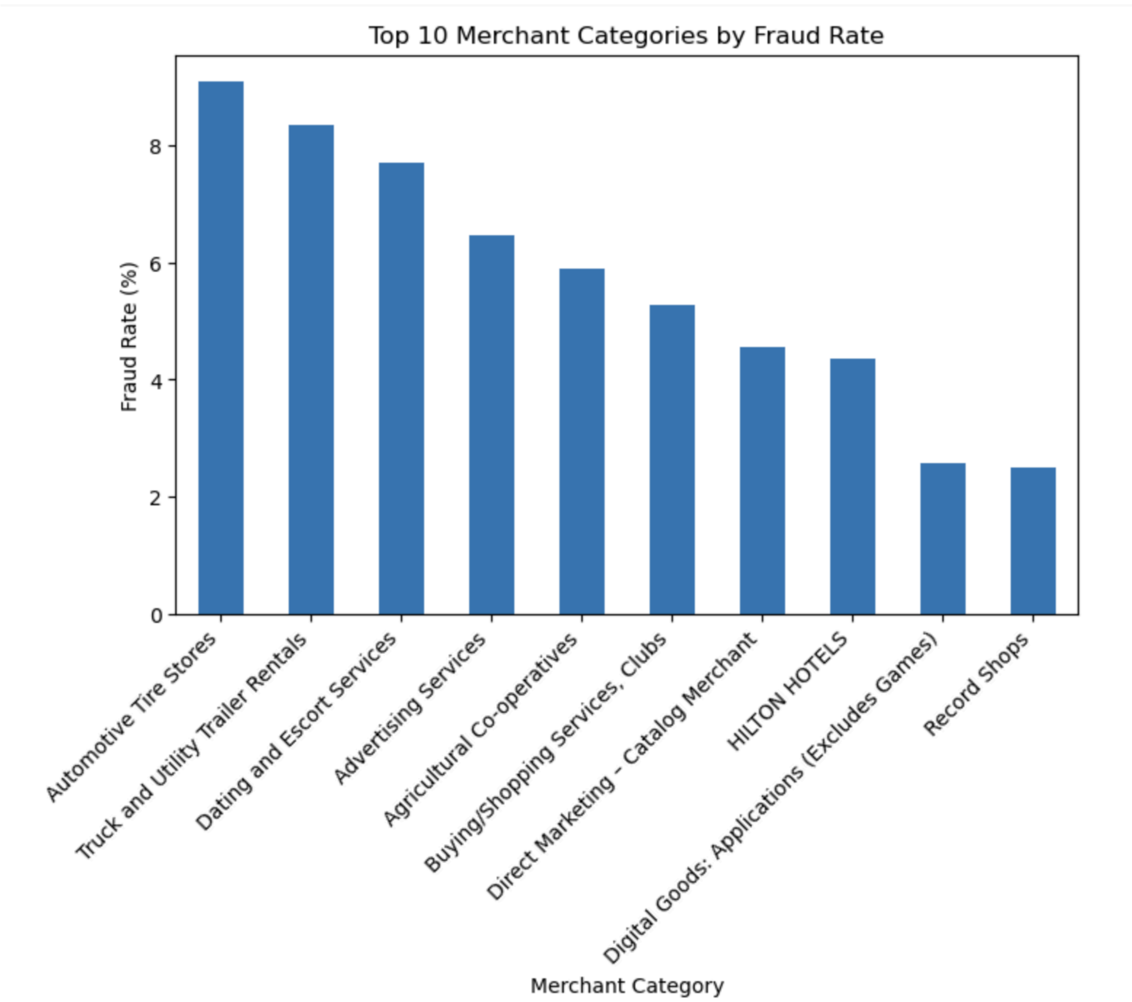
Bivariate Analysis

How does each feature correlate to the fraud indicator?

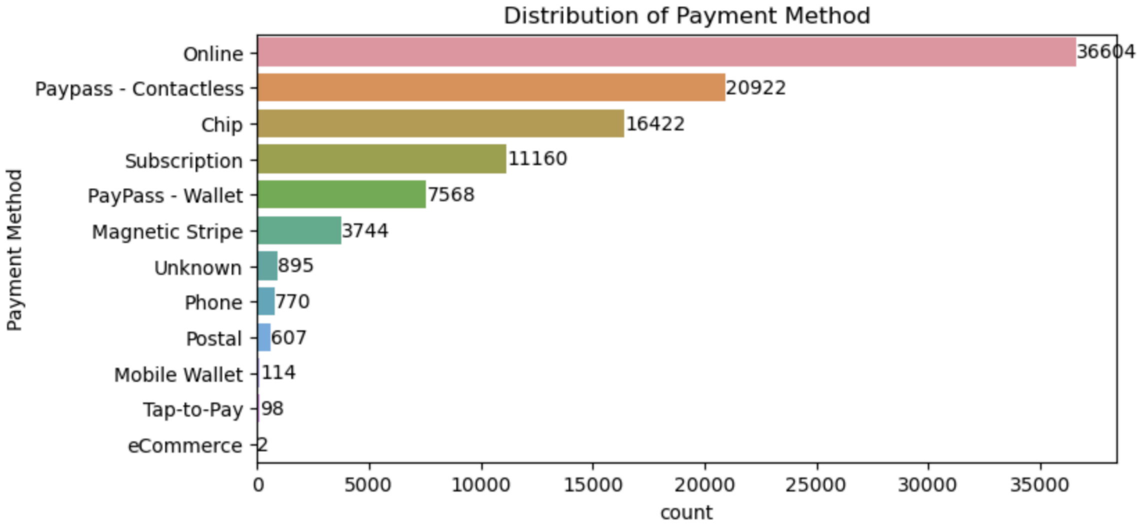
Merchant Category

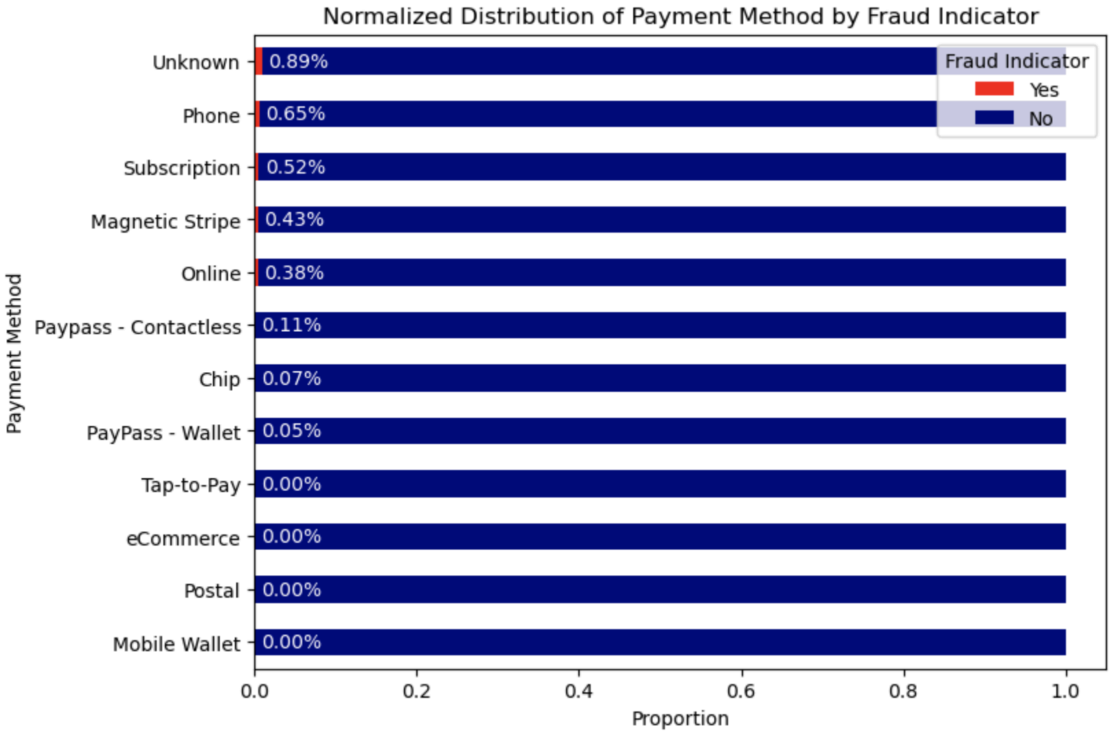
```
Merchant Category
HO JO INN, HOWARD JOHNSON                                                  100.000000
Roofing – Contractors, Sheet Metal Work – Contractors, Siding – Contractors  50.000000
COMFORT INNS                                                                50.000000
LOEWS HOTELS                                                                33.333333
FARIFIELD INN                                                               28.571429
HOMEWOOD SUITES                                                             25.000000
Automotive Tire Stores                                                       9.090909
Truck and Utility Trailer Rentals                                            8.333333
Dating and Escort Services                                                   7.692308
Advertising Services                                                         6.462585
dtype: float64
```

Susceptible to fraud: Inns, hotels, and services.



Payment Method





Merchant Location (May be insignificant)

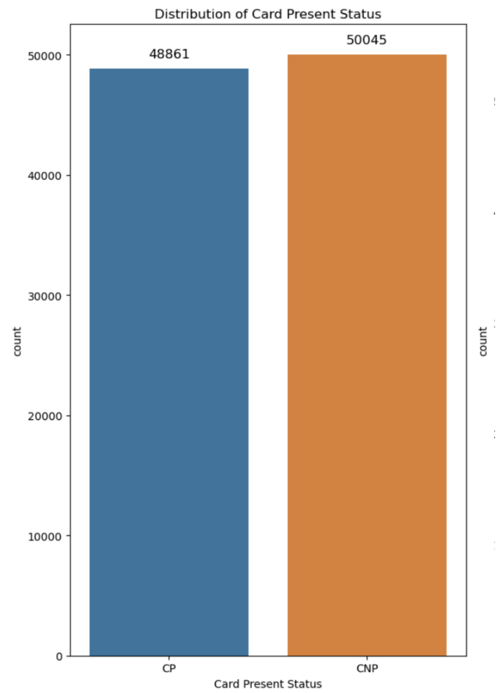
```
Merchant Location
LBN    38.888889
SGP     7.389163
IND     6.896552
ECU     5.000000
PER     4.761905
ISL     3.125000
DOM     3.030303
DEU     2.439024
BRA     2.208202
ITA     1.016260
dtype: float64
```

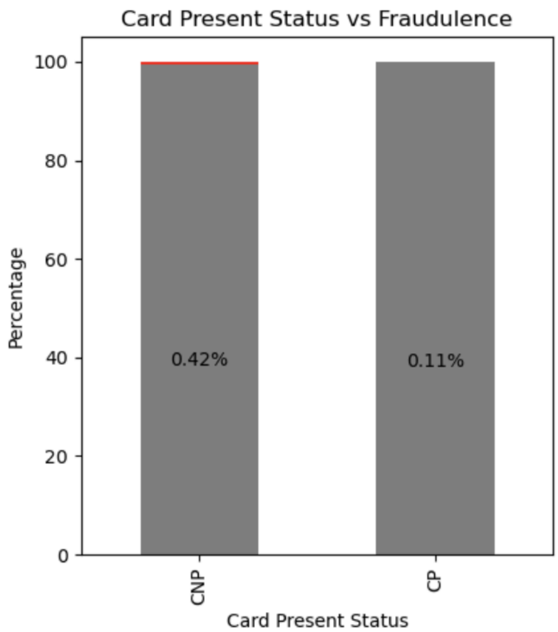
From previous visualizations, we know that 0.267931% of transactions are flagged for fraud. Similarly, we calculate the average per merchant location to compare with these high fraudulence locations.

```
# Average pct of fraudulence per merchant location
avg_fraud_rate_all = fraud_rate.mean()
print(f"Average % of fraudulence per merchant location: {avg_fraud_rate_all:.6f}%")
```

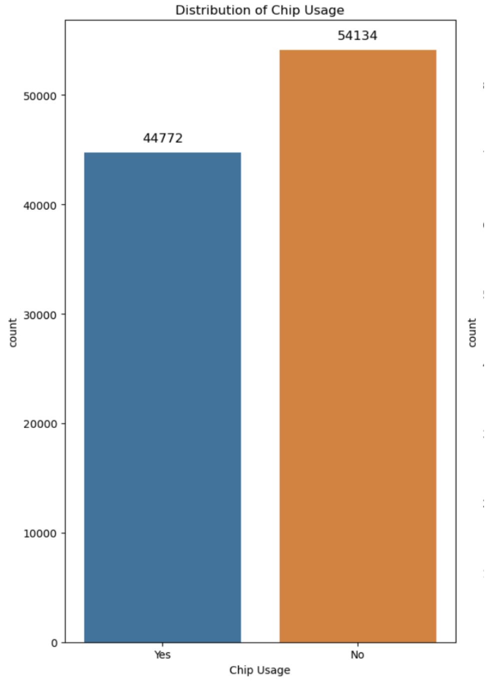
```
Average % of fraudulence per merchant location: 1.182484%
```

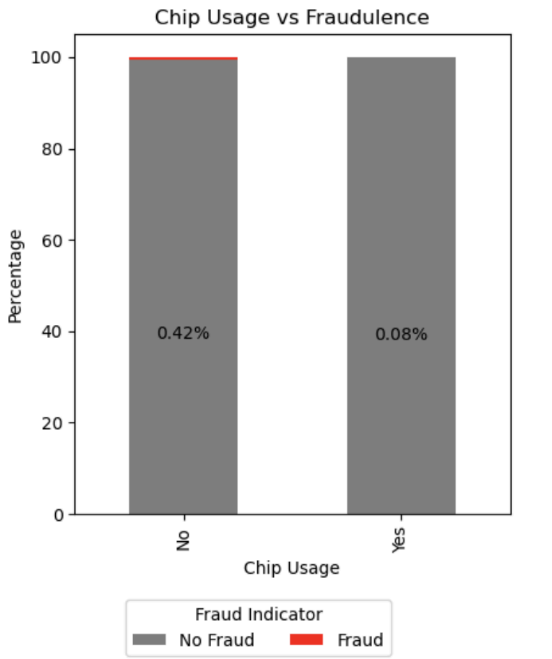
Card Present Status (Yes/No)



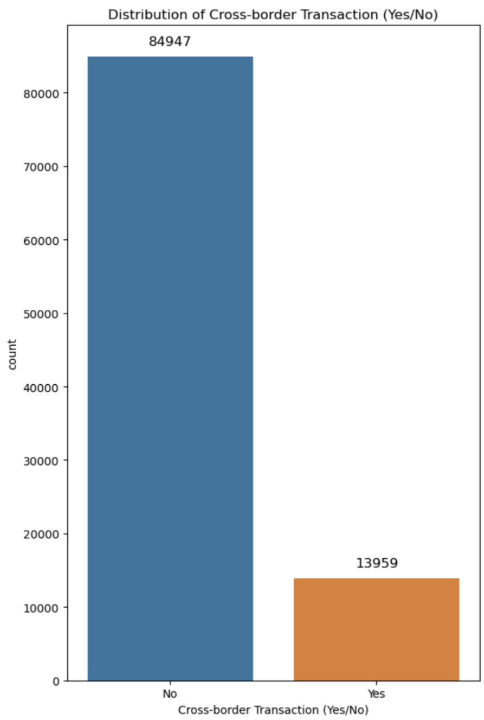


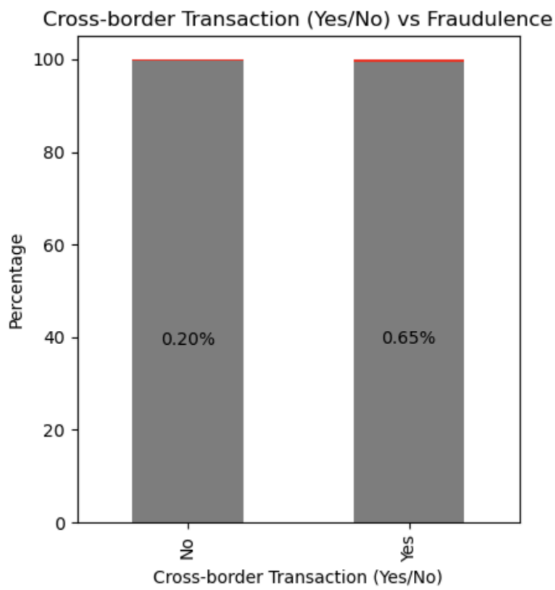
Chip Usage (Yes/No)



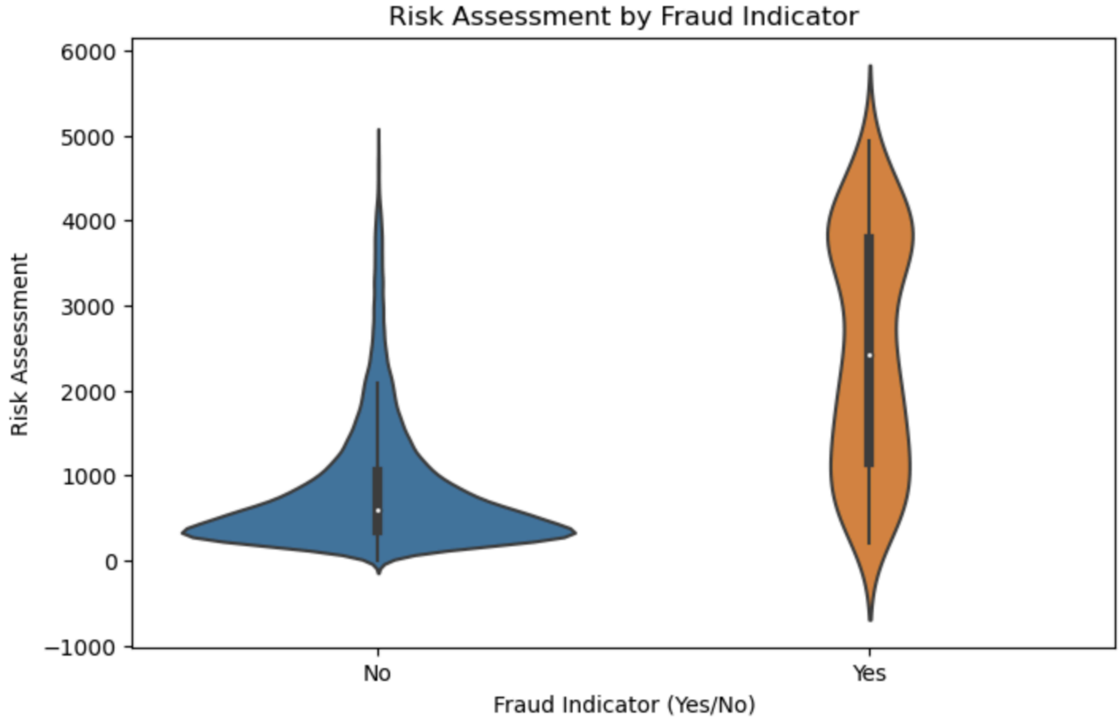


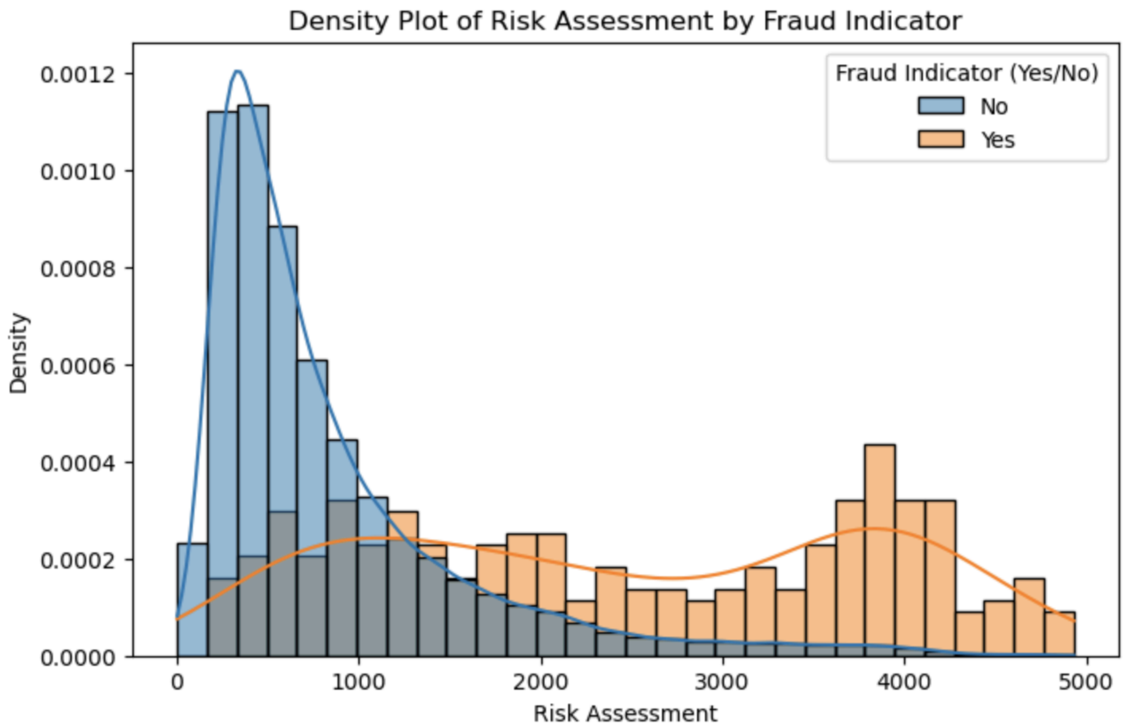
Cross Border Transaction (Yes/No)

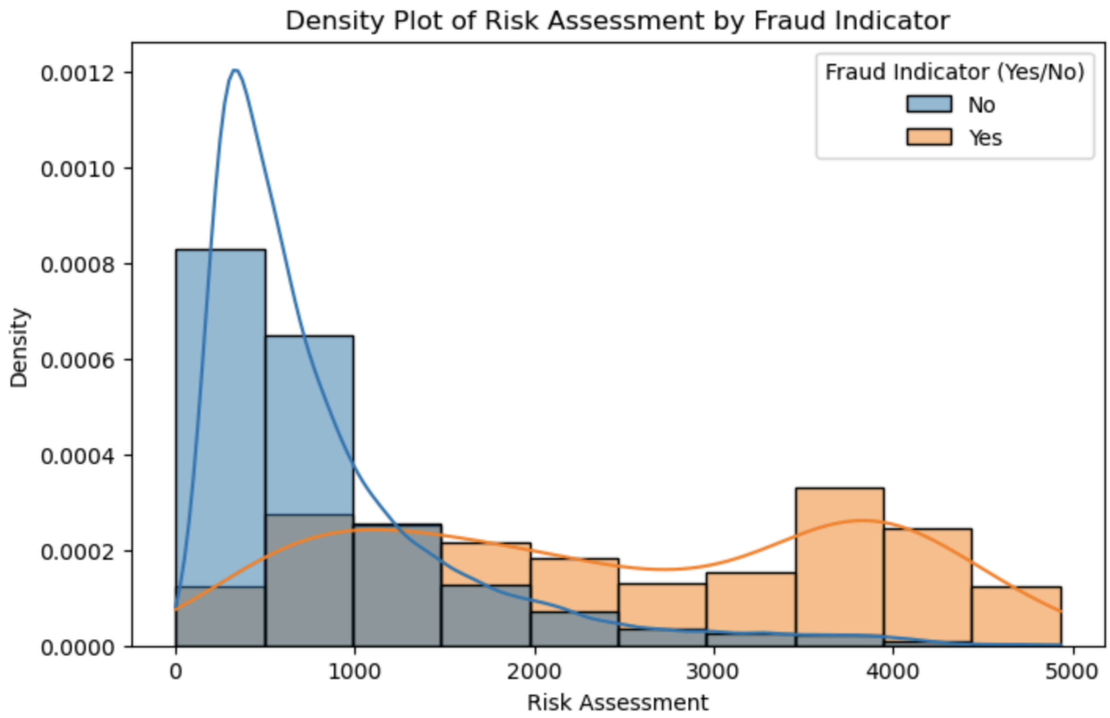


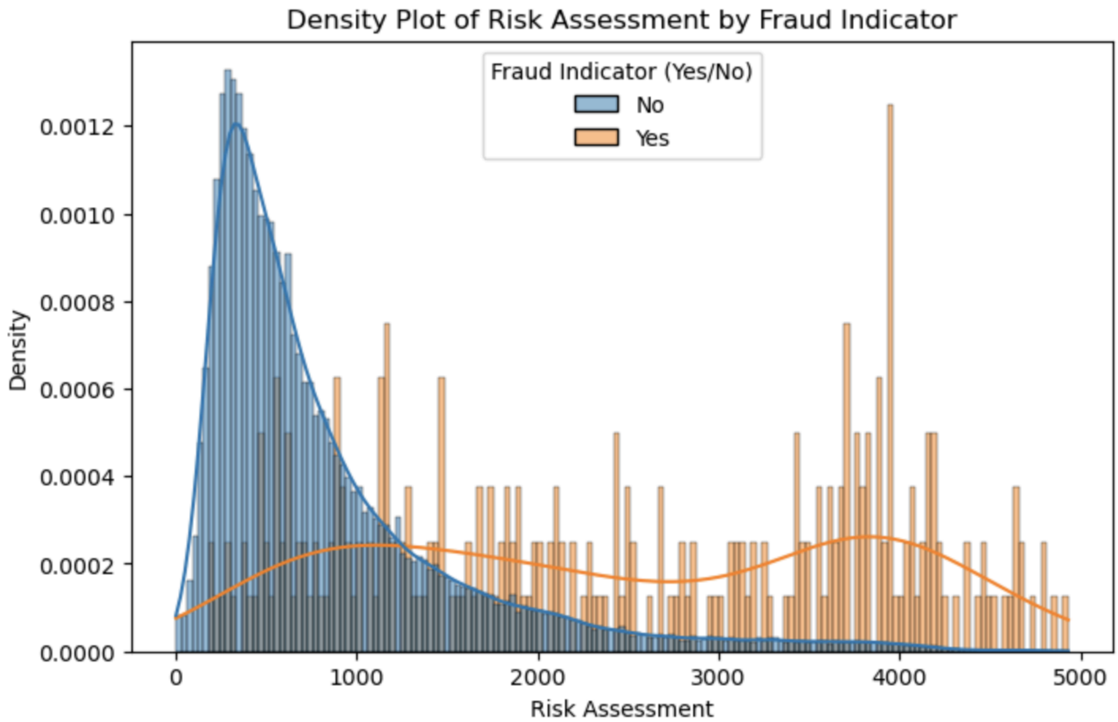


Risk Assessment









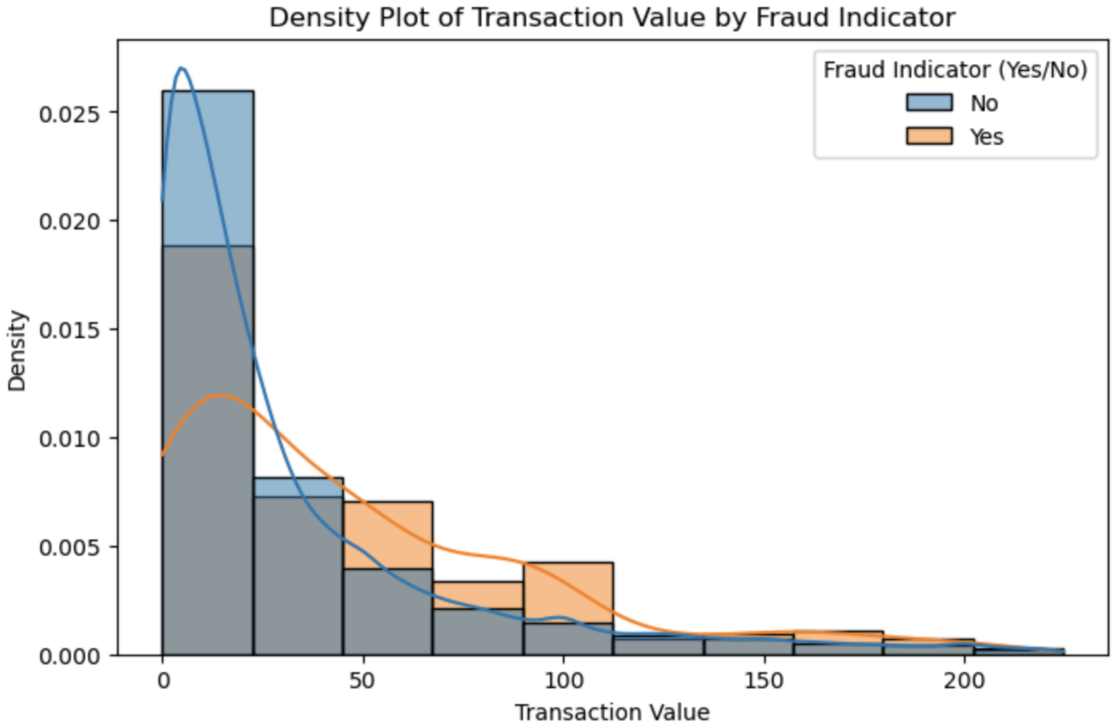
RISK ASSESSMENT

- **Distribution Shape:** The 'Risk Assessment' for non-fraudulent transactions (No) is tightly concentrated around lower values with a narrow distribution, indicating that most non-fraudulent transactions have a low-risk assessment score. The distribution for fraudulent transactions (Yes) is wider, suggesting a greater variability in the risk assessment scores for fraudulent transactions.

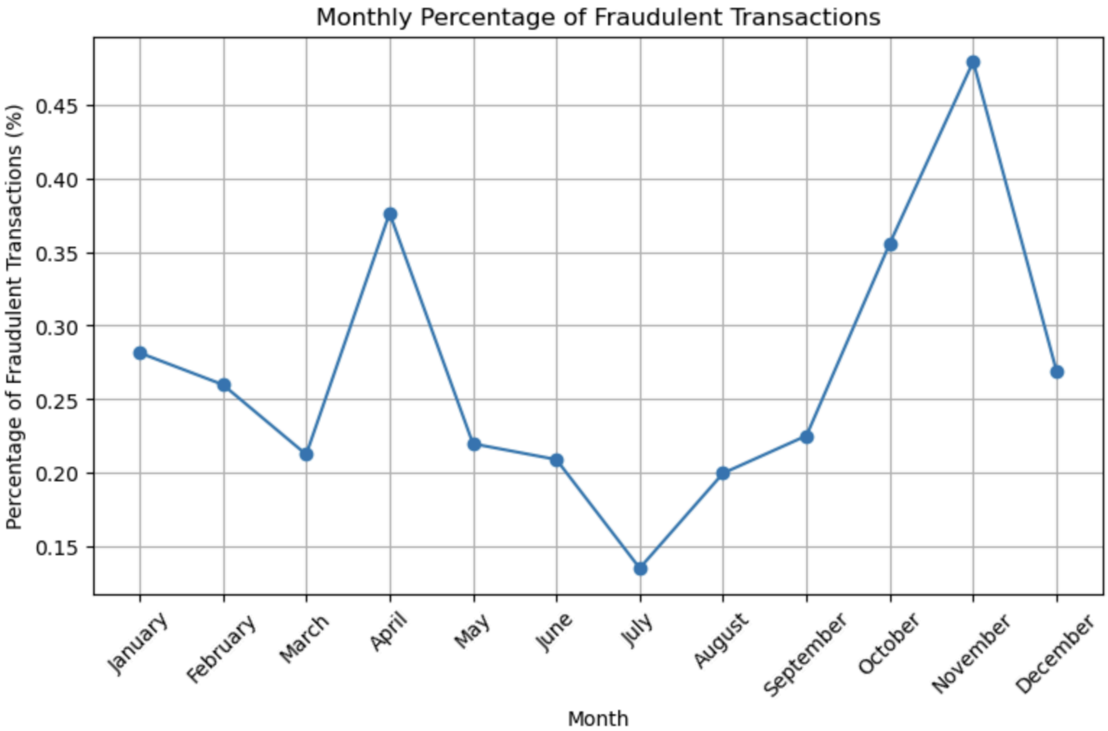
- **Range:** The range of 'Risk Assessment' scores for fraudulent transactions is broader than for non-fraudulent ones, indicating that fraudulent transactions can have a wide range of risk scores, but tend to have higher scores on average.
- **Median and Quartiles:** The median (indicated by the white dot) for fraudulent transactions is higher than for non-fraudulent ones, which aligns with the expectation that transactions deemed more risky are more likely to be fraudulent.
- **Outliers:** The plot for non-fraudulent transactions shows fewer outliers compared to the plot for fraudulent transactions, indicating that most non-fraudulent transactions adhere to a lower risk profile.

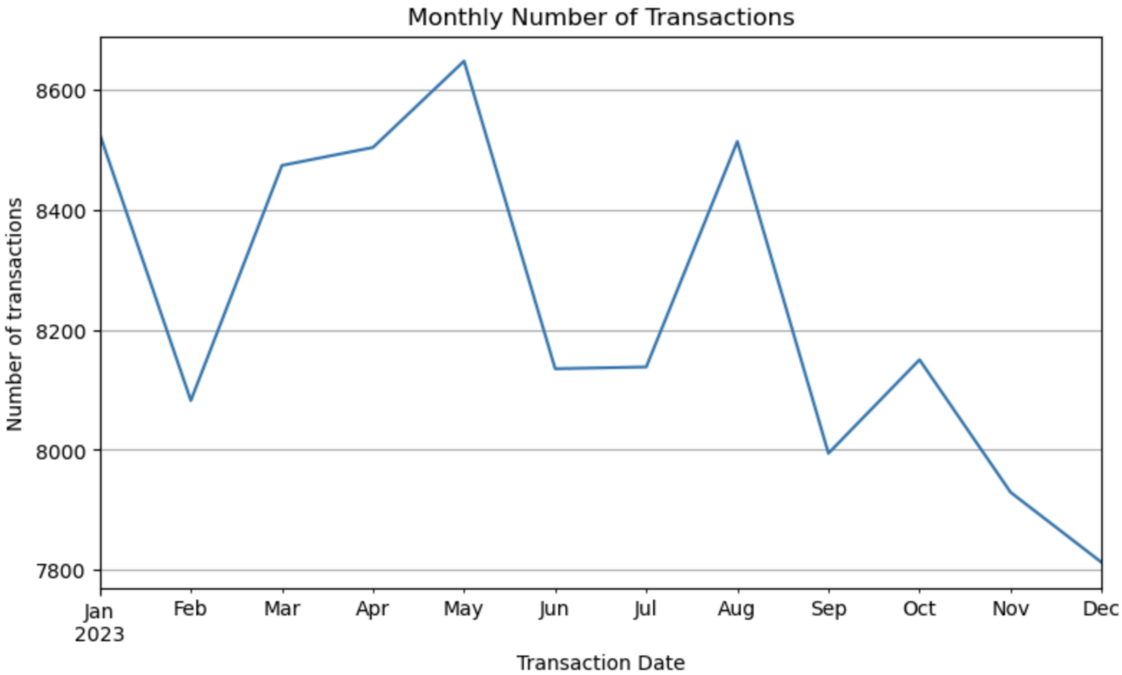
Transaction Values

Transaction values seemingly does not reflect upon the fraudulence of a particular transaction in a linear relationship.



Temporal Analysis





Cramer’s V-Test

Is the occurrence of fraud independent of a particular feature, and if so, by what degree?

Chi-Square Test for Independence

```
Card Present Status - p-value: 5.383848817728829e-21
Chip Usage - p-value: 2.1958704289262313e-24
Cross-border Transaction (Yes/No) - p-value: 6.5051876113018414e-21
Risk Assessment - p-value: 0.0
Payment Method - p-value: 2.0386422879988157e-21
Transaction Value - p-value: 0.0
Merchant Location - p-value: 0.0
```

So what do these p-values tell us?

1. **'Card Present Status', 'Chip Usage', 'Cross-border Transaction', 'Payment Method'** display **very small p-values**. These extremely small numbers suggest that the likelihood of observing the data if the null hypothesis were true (no association) is extremely low. In practical terms, these results indicate that there is a very strong statistical significance and a likely association between each of these categories and the occurence of fraud. The distribution of fraud indicators varies significantly across the different levels of these categorical variables.
2. **'Risk Assessment', 'Transaction Value', 'Merchant Location'** display a **p-value of 0** (likely due to rounding). This suggests a certain statistical association between these variables and the 'Fraud Indicator (Yes/No)'.

This is where the V Test comes in–we estimate the strength of the correlations.

Cramer’s V Test

```
Card Present Status: 0.029724385577123456
Chip Usage: 0.032245345011337606
Cross-border Transaction (Yes/No): 0.02966066849583715
Risk Assessment: 0.3032596563816111
Payment Method: 0.03390723967705102
Transaction Value: 0.298090525518079
Merchant Location: 0.14172863438503566
Merchant Category Code (MCC): 0.13796010489253901
```

How do we interpret these values?

1. ***WEAK ASSOCIATIONS:*** **Card Present Status (0.0297), Chip Usage (0.0322)**, **Cross-border Transaction (0.0298)**, and **Payment Method (0.0339)** have very low Cramér's V values, suggesting these variables have very weak associations with the

variable they were compared against. These factors might not be strong predictors on their own for the variable of interest in your analysis.

2. ***MODERATELY WEAK ASSOCIATIONS:*** **Merchant Location (0.1417)** and **Merchant Category Code (MCC) (0.1379)** have slightly higher but still relatively low Cramér's V values. There is a weak association with the variable they were compared against, indicating they have a bit more influence than the previously mentioned variables but still a limited predictive power.
3. ***MODERATELY STRONG ASSOCIATIONS:*** **Risk Assessment (0.3033)** and **Transaction Value (0.2983)** stand out with the highest Cramér's V values among those listed, indicating a moderate association with the variable they were compared against. This suggests that these variables have a more substantial relationship and could be more significant predictors in your analysis.

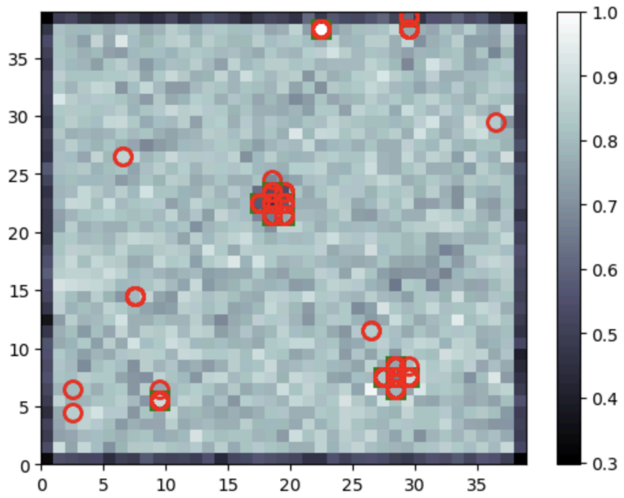
Self-Organized Map (SOM) Analysis

The Self-Organized Map is a **clustering** algorithm that leverages neural networks to iteratively group similar data points together. It has been proven to produce the highest accuracies between clustering algorithms when predicting credit fraud (G et al., 2018).

References:

G, A., K, M., Reddy, B. K. K., Iyengar, N. Ch. S. N., & Caytiles, R. D. (2018). Analyzing the performance of Various Fraud Detection Techniques. *International Journal of Security and Its Applications*, *10*(5), 21–36. https://doi.org/10.14257/ijsia.2018.12.5.03

SOM U-Matrix



Map of all the neurons (clusters) where each cluster is a square; clusters can have very little transactions or very many according to how they arrange themselves. Similar transactions will naturally group together.

How to interpret this:

- Darker shades: boundaries between different clusters.
- Lighter shades: clusters of similar data points.
- **Areas with dense clusters of red circles**: common patterns or features among fraudulent transactions. These neurons have learned to represent the "profile" of fraudulent activity within your dataset.
- **Isolated red circles:** indicate outliers or less common types of fraudulent transactions.
- **Neurons with mixed red and green shapes**: represent ambiguous areas or transitional zones between fraudulent and non-fraudulent transactions.

Threshold = 0.1

Cluster	Features
(17, 22)	Fraud Rate: 0.113234253361641 Feature Analysis: Payment Method mode: 2.9999999999999996 Merchant Location mode: 127.0 Card Present Status mode: 0.0

	Chip Usage mode: 0.0 Cross-border Transaction (Yes/No) mode: 0.0 Merchant Category mode: 134.0 Transaction Value mean: 70.5966053314461 Transaction Value median: 22.52 Risk Assessment mean: 1282.073130455296 Risk Assessment median: 877.0000000000002
(22, 37)	Fraud Rate: 0.10582010582010581 Feature Analysis: Payment Method mode: 2.9999999999999996 Merchant Location mode: 46.0 Card Present Status mode: 0.0 Chip Usage mode: 0.0 Cross-border Transaction (Yes/No) mode: 1.0 Merchant Category mode: 331.0 Transaction Value mean: 75.0776402116402 Transaction Value median: 15.889999999999999 Risk Assessment mean: 1901.989417989418 Risk Assessment median: 1717.0
(9, 5)	Fraud Rate: 0.17276014463640015 Feature Analysis: Payment Method mode: 2.9999999999999996 Merchant Location mode: 46.0 Card Present Status mode: 0.0 Chip Usage mode: 0.0 Cross-border Transaction (Yes/No) mode: 1.0 Merchant Category mode: 331.0 Transaction Value mean: 103.15316191241463 Transaction Value median: 17.0 Risk Assessment mean: 1866.0498192044997 Risk Assessment median: 1647.0

Summative Analysis

- Predominant payment method (mode approximately 3) → Online
- Card not present (0.0), and no chip usage (0.0), suggesting a pattern of remote transactions without physical card verification.
- Risk assessments notably high across all clusters, with means over 1000 and medians also high, reflecting the transactions' perceived riskiness.
- Mode for "Cross-border Transaction (Yes/No)" alternates between 0.0 and 1.0 across clusters; both domestic and cross-border transactions prone to fraud.
- Transaction values vary, with means ranging from approximately 70 to 103, and medians significantly lower.

Training using Random Forest & Gradient Boosting

Due to the fact that we have many categorical variables and it is not computationally viable to OneHotEncode some of these variables (such as Merchant Location), we choose to select between the **Random Forest Classification** or **Gradient Boosting** models, as it handles non-linear relationships and interactions between categorical features well.

Random Forest Classification

```
              precision    recall  f1-score   support

           0       0.91      0.91      0.91        53
           1       0.91      0.91      0.91        53

    accuracy                           0.91       106
   macro avg       0.91      0.91      0.91       106
weighted avg       0.91      0.91      0.91       106
```

Accuracy: 0.9056603773584906

Feature Importances:
Risk Assessment: 0.3948
Transaction Value: 0.2168
Merchant Category Code (MCC): 0.1224
Merchant Category: 0.1167 -> Highly correlated, may have overfit
Payment Method: 0.0491
Cross-border Transaction (Yes/No): 0.0428
Chip Usage: 0.0352
Card Present Status: 0.0223

Gradient Boosting (CatBoost)

```
CatBoost Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.80      0.82        56
           1       0.79      0.82      0.80        50

    accuracy                           0.81       106
   macro avg       0.81      0.81      0.81       106
weighted avg       0.81      0.81      0.81       106
```

Accuracy: 0.8113207547169812

Random Forest has a higher accuracy of 90.5%, compared to the 81.1% accuracy of CatBoost. The Random Forest also leads in precision, recall, and F1-scores for both classes. This suggests that Random Forest is suited for the task at hand, as it is able of not only identifying correct instances but also reducing the chances of false alarms and misses.