CREDIT CARD FRAUD DETECTION

By Arnab Saha

AGENDA

- Problem Background
- Problem Understanding & Impact
- Insights
- Model Evaluation
- Cost Benefit Analysis
- Conclusion
- Business Recommendation

PROBLEM BACKGROUND

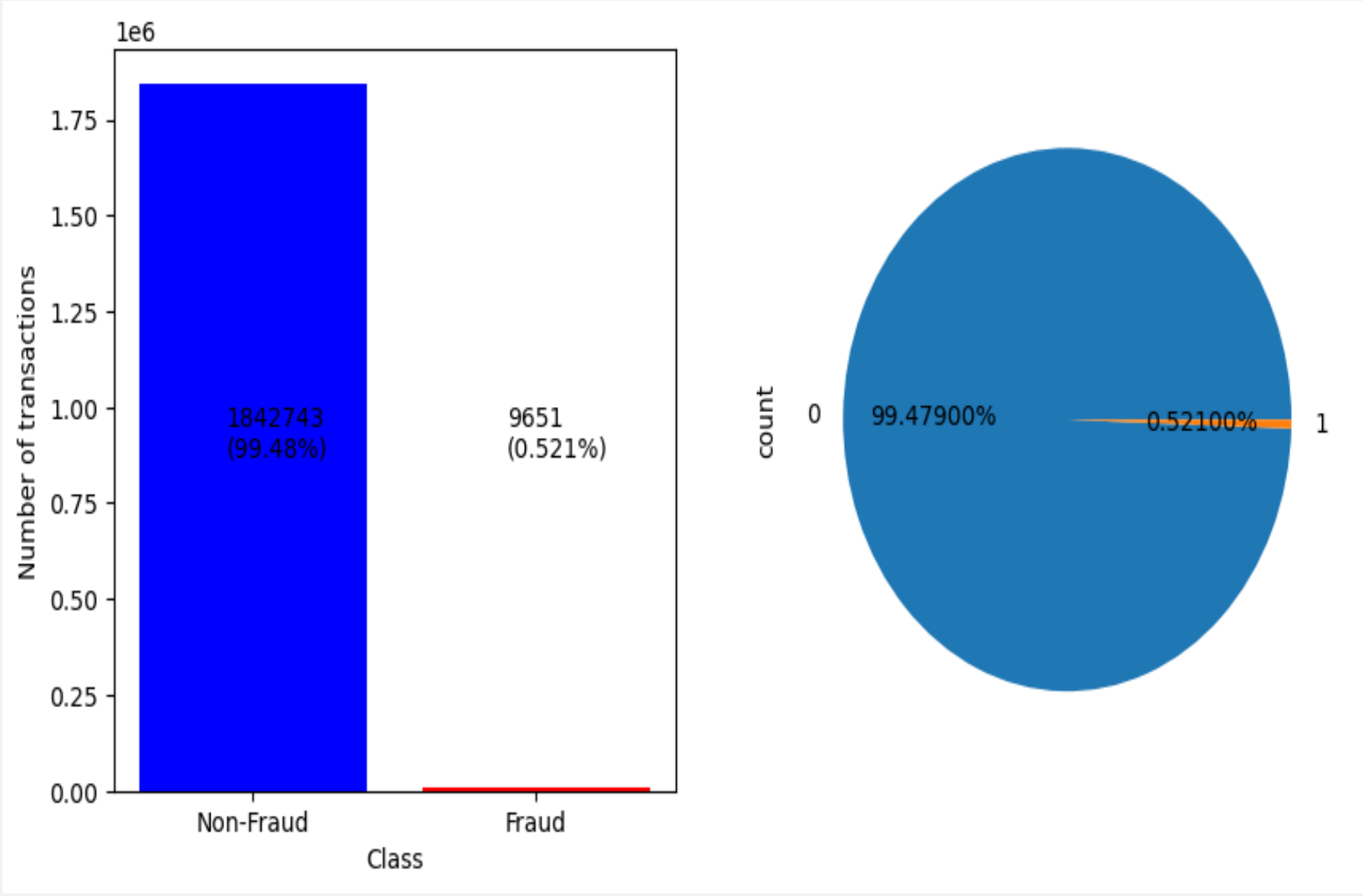
- Large number of fraudulent transactions is causing huge revenue loss and profitability crisis.
- Fraudsters has been using stolen/lost cards to do fraudulent activities.
- Fraudsters are doing ATM skimming at various POS terminals which don’t have proper security system.
- The goal is to track data breaches on time to prevent further losses and retaining high profitable customers.

PROBLEM UNDERSTANDING & IMPACT

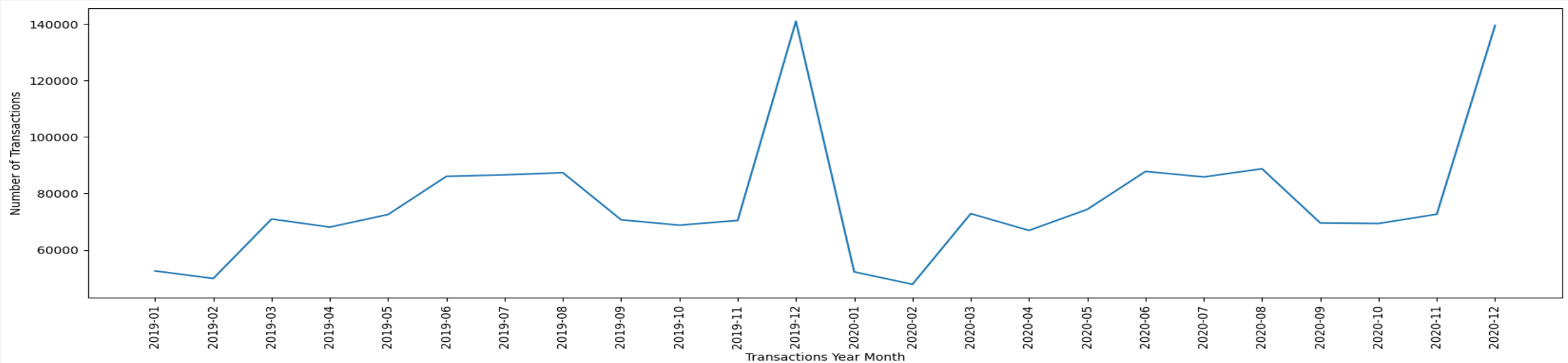
- Fraudsters, skimmers are getting unauthorized access to a customer’s credit card info and committing fraudulent transactions.
- Fraudulent cases are happening on online and sometimes offline.
- Online fraud are hard to detect due to resources and evidence constraints.
- Most of fraudulent activities are happening in the non-peak and odd hours of the day.
- Because of the fraudulent activities banks is suffering from financial losses, losing trust and unable to retain customers.

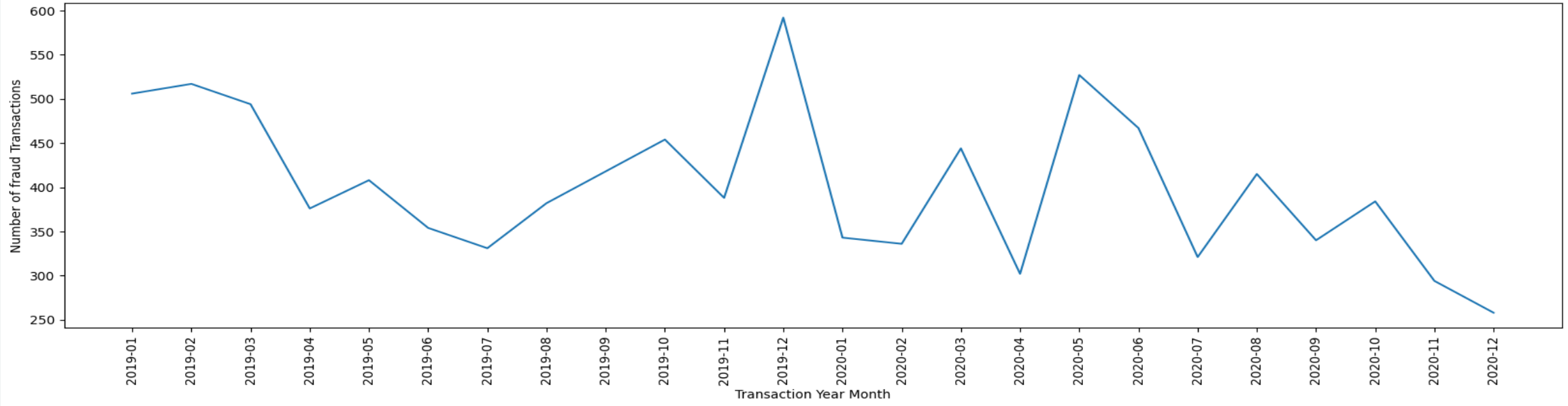
OVERVIEW OF DATA

- Fraud Data = 9651 transaction(0.521 %)
- Non-Fraud Data = 1842743 transaction(99.48 %)
- The data seems to be imbalance.



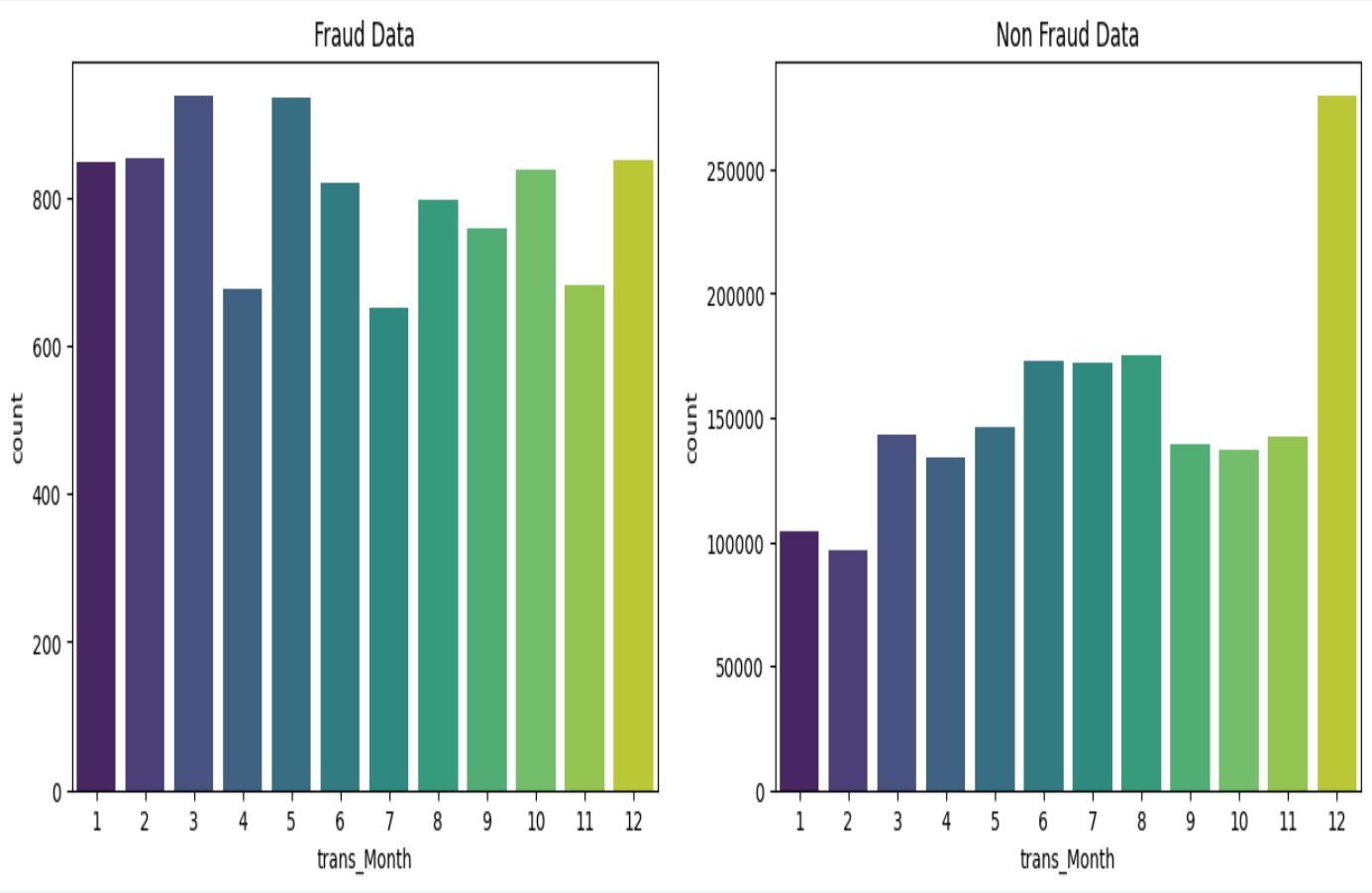
FREQUENCY OF TRANSCTION MONTHLY WISE

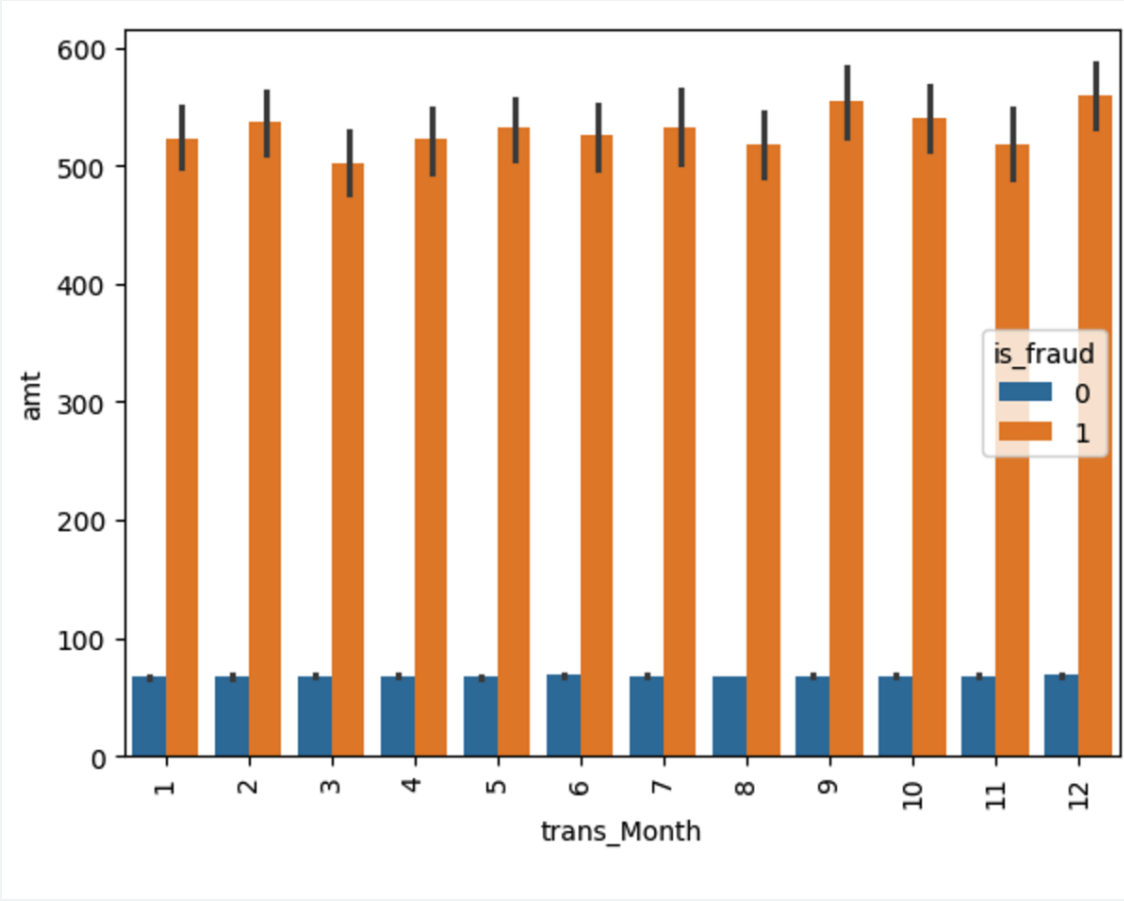




Number of Fraud transactions are more on new year eve and specifically in Dec month, followed by May, August.

ANALYSIS BASED ON TRANSACTIONS MONTH

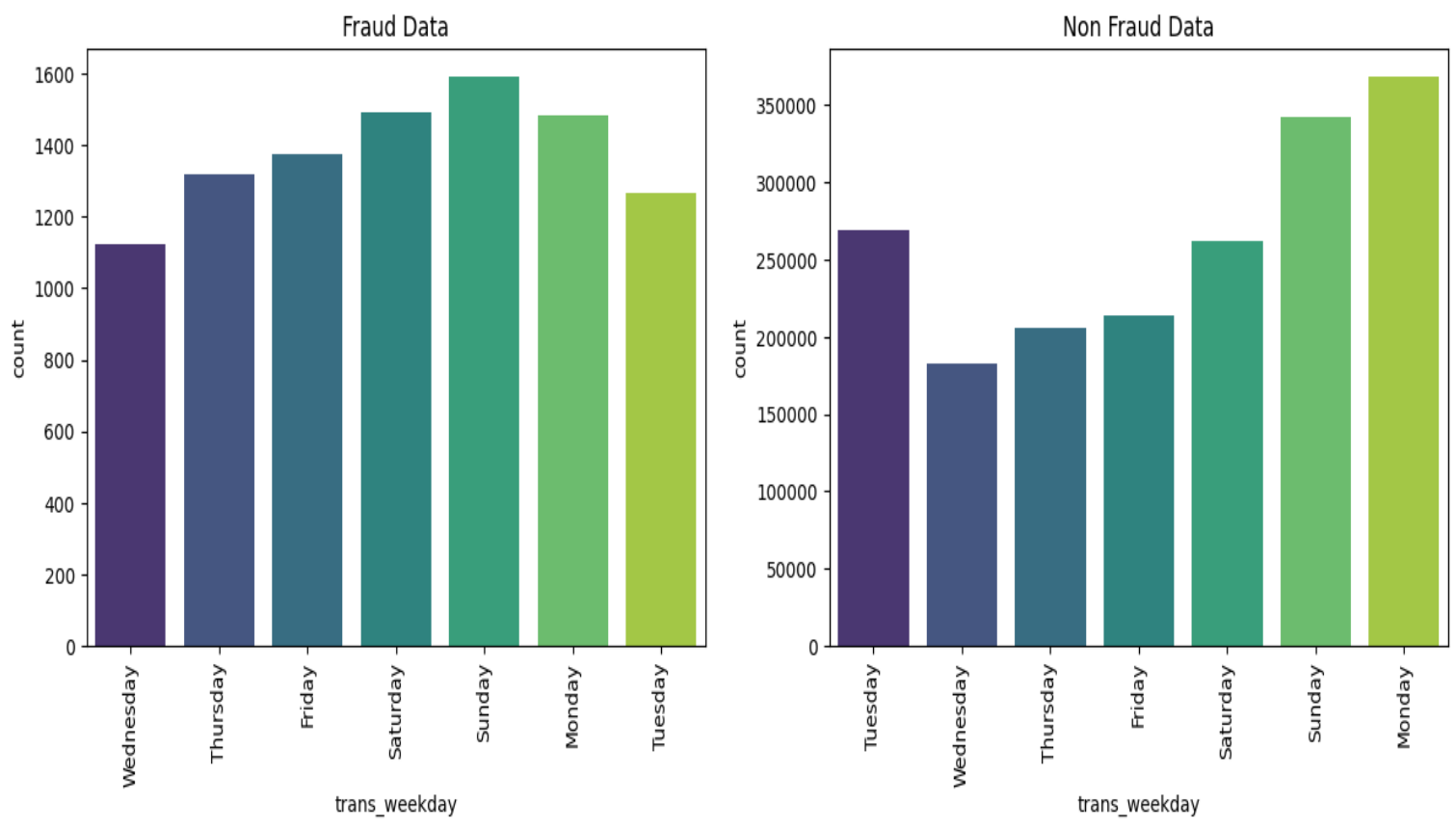


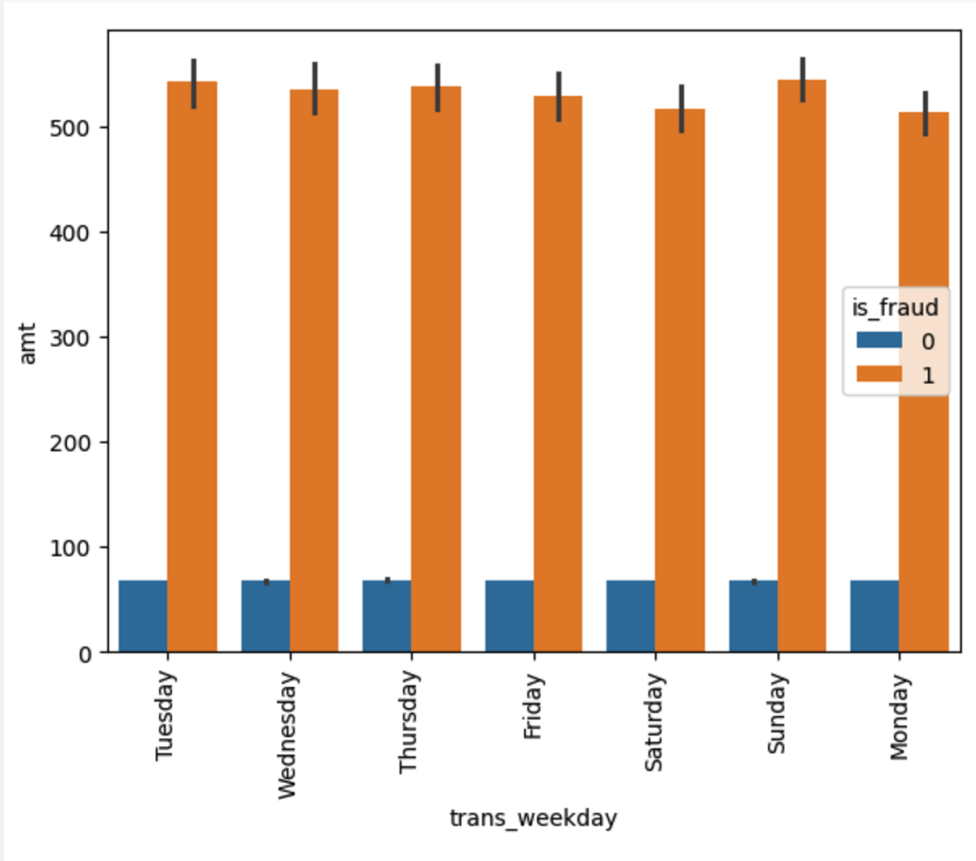


Fraudulent transactions are more in 3rd and 5th month where the normal transaction is less in respectively month.

The amount spent on fraudulent transactions are similar through out the month .

ANALYSIS BASED ON TRANSACTIONS WEEK

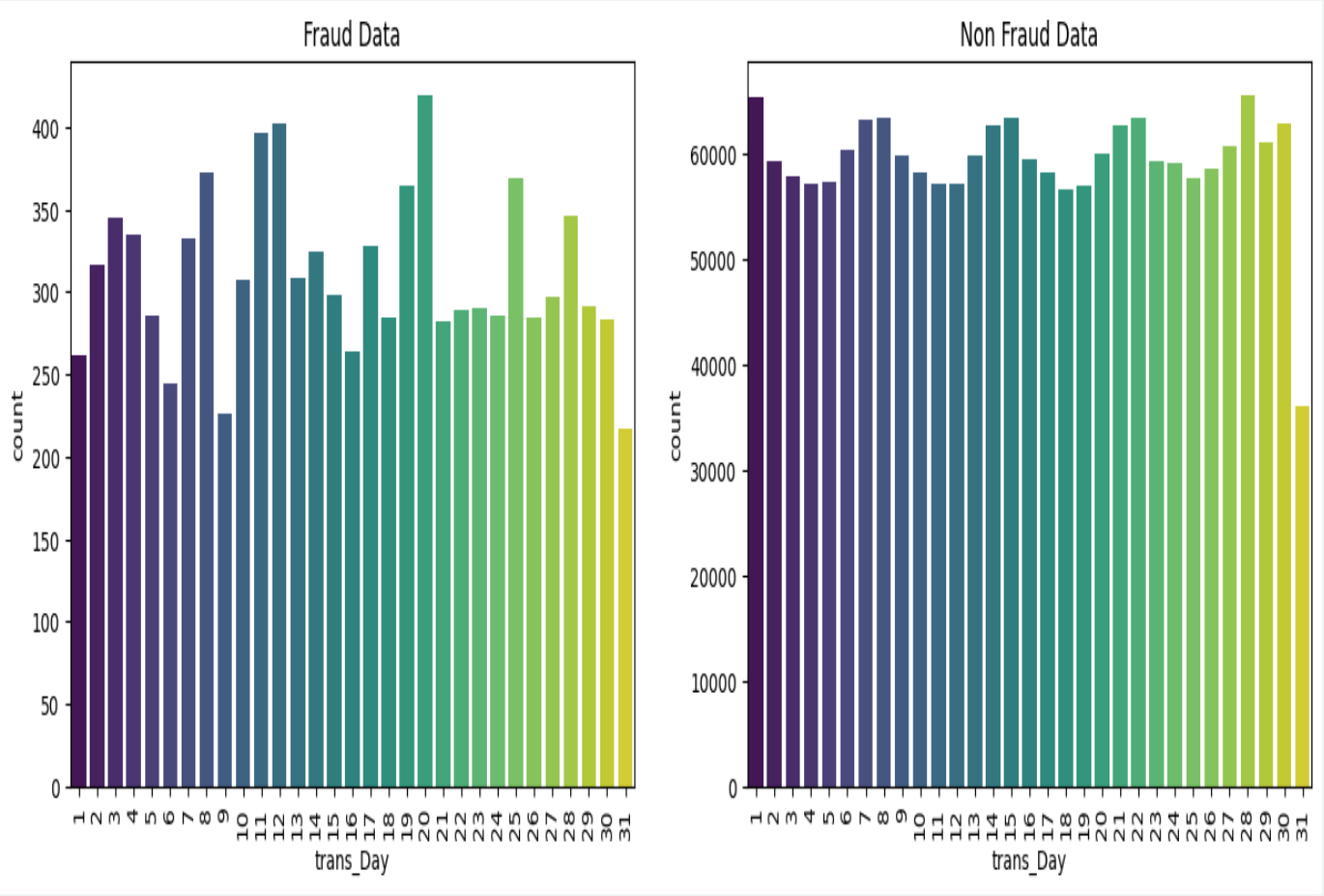




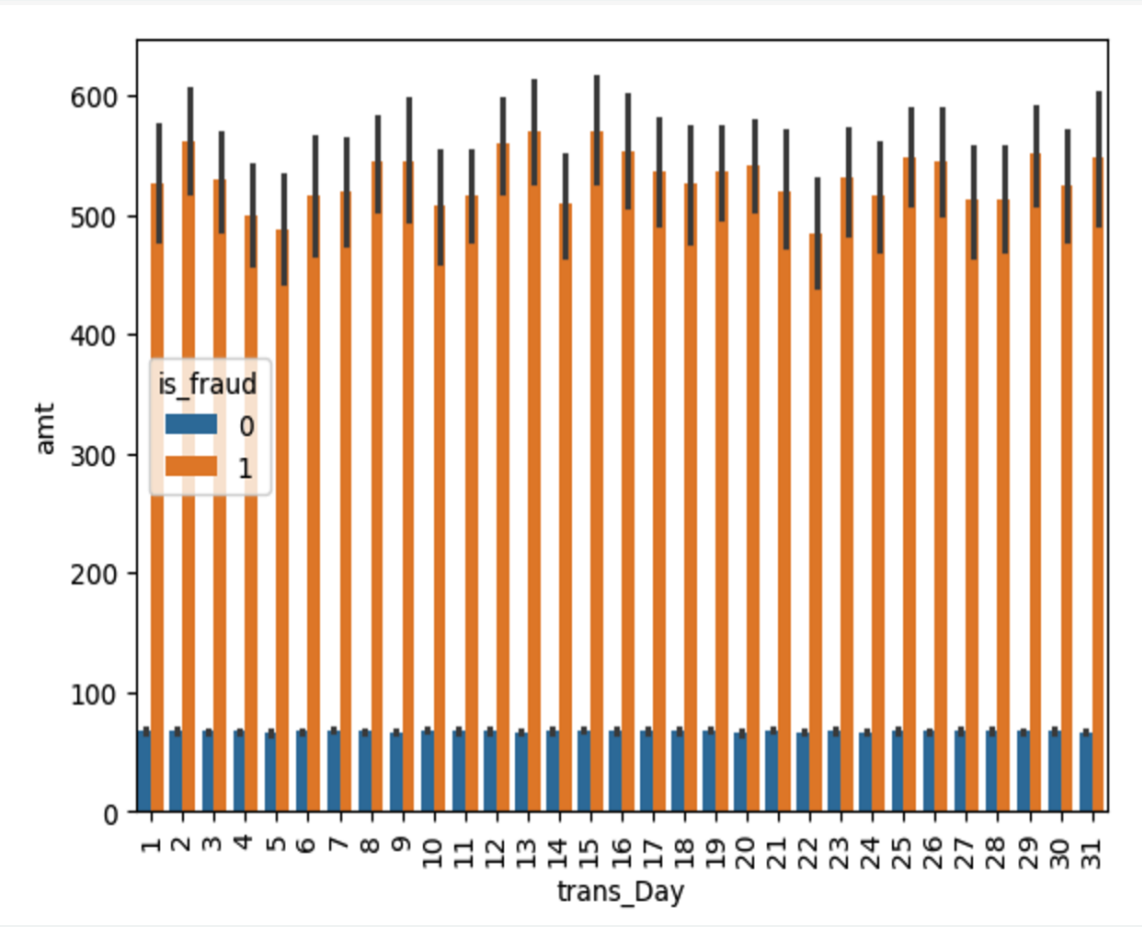
Fraudulent transactions are more on Saturday, Sunday and Monday compared to normal transaction.

The amount spent on fraudulent transactions are similar through out the week .

ANALYSIS BASED ON TRANSACTIONS DAYS

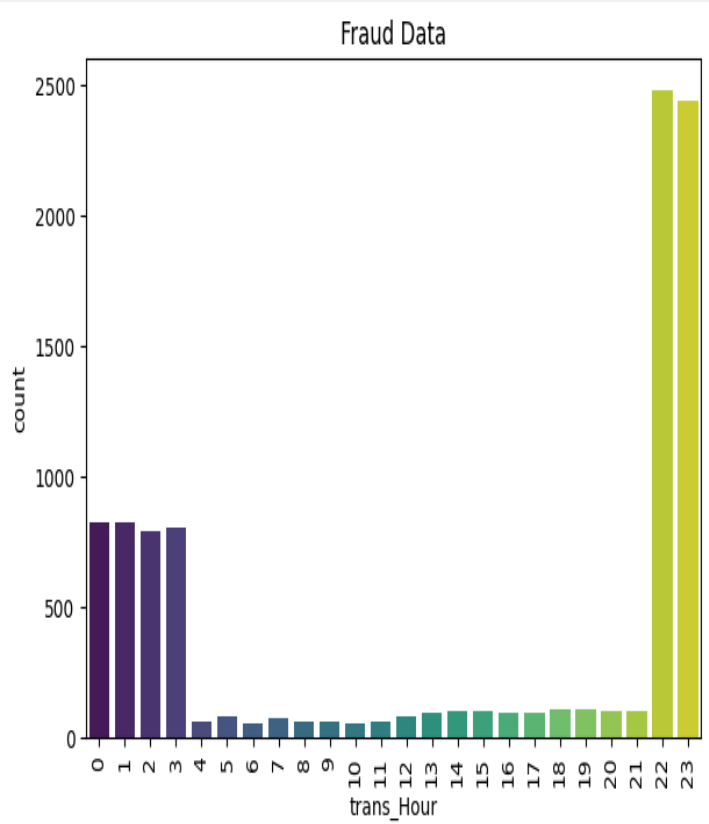


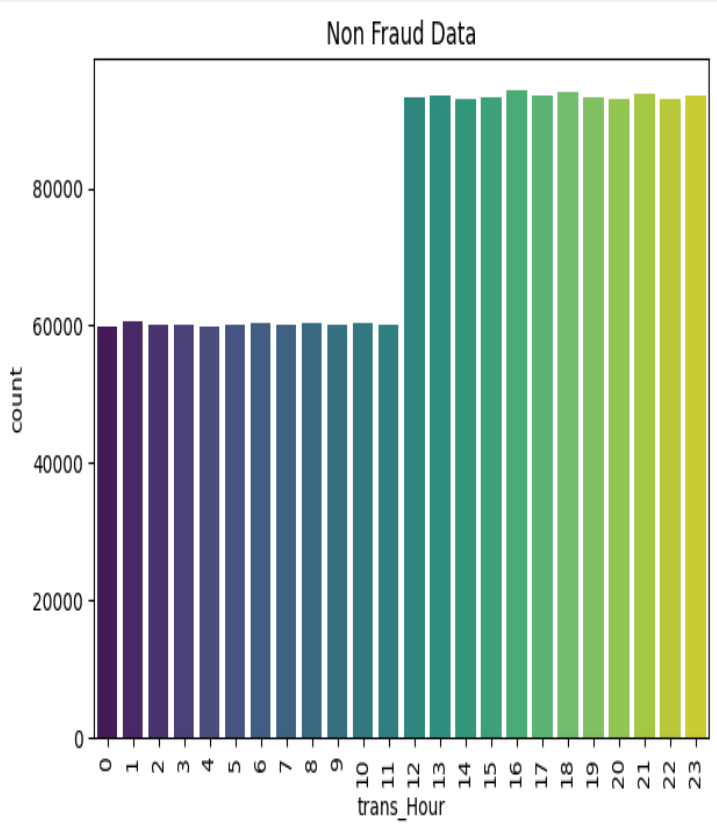
Fraudulent transactions are more in number on 11th, 12th and 20th days of the month.

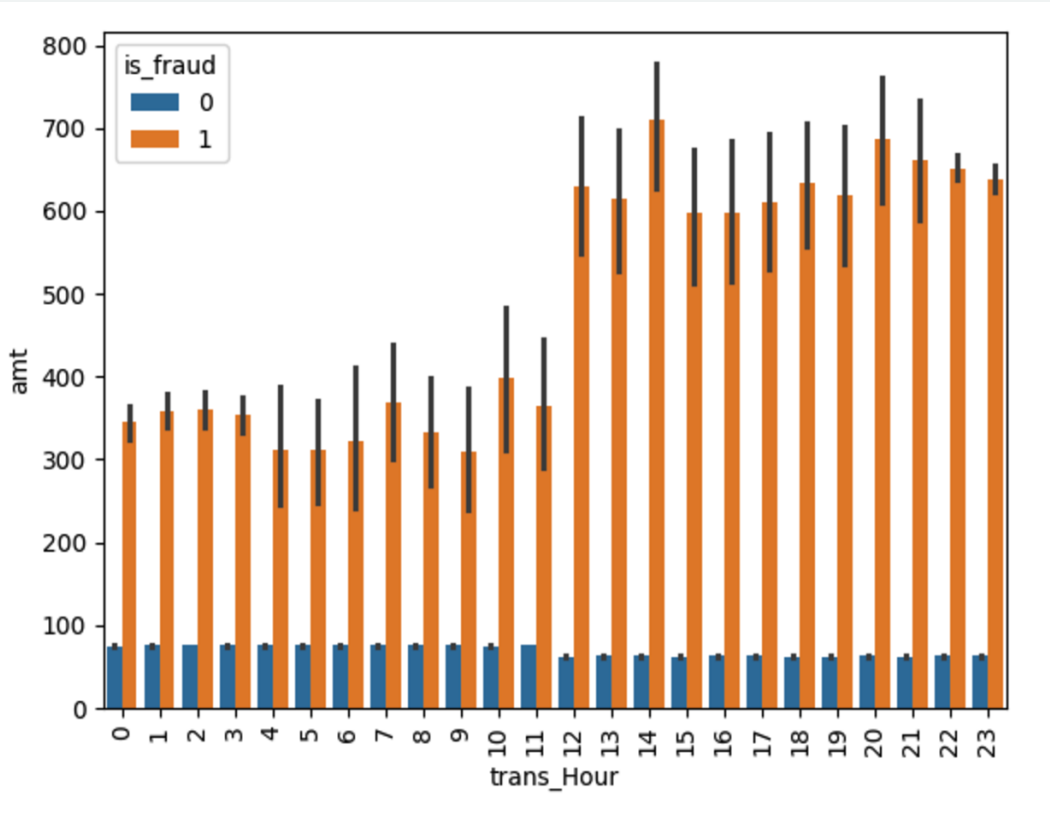


The amount spent on fraudulent transactions are similar through out the days of the month.

ANALYSIS BASED ON TRANSACTIONS HOURS



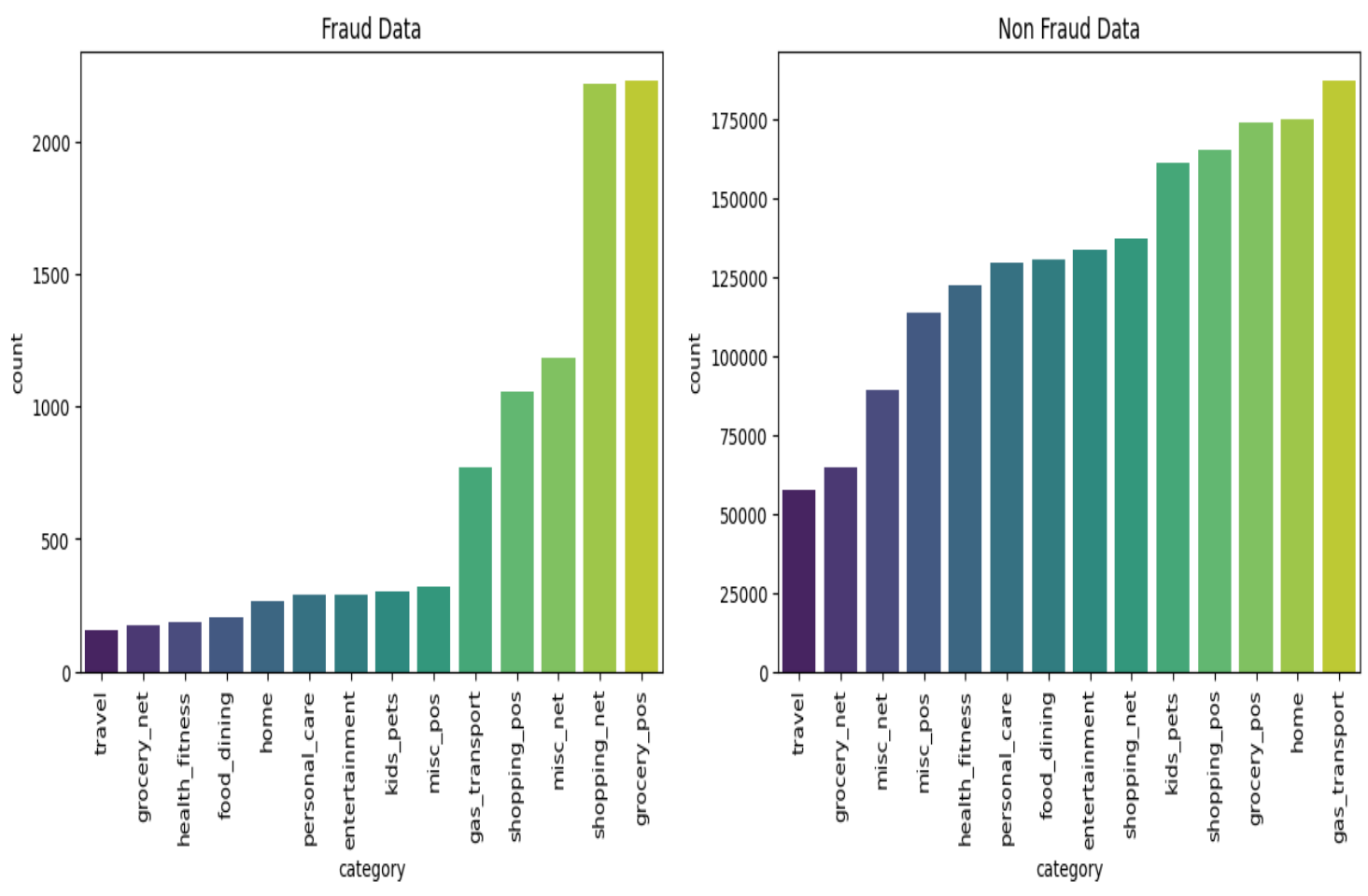


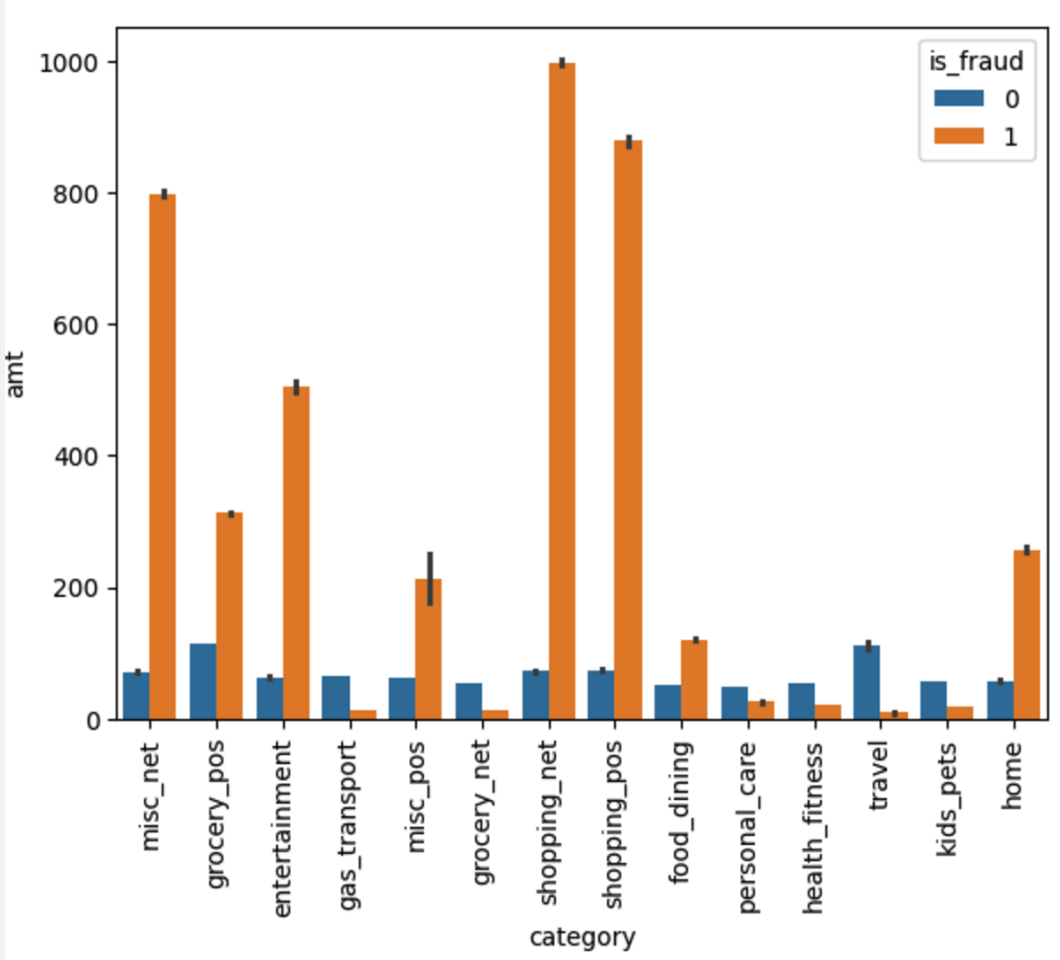


Fraudulent transactions are more in number at odd hours of the day(between 22:00-03:00)

The amount spent on fraudulent transactions is more between 12:00 to 23:00.

ANALYSIS BASED ON DIFFERENT CATEGORIES

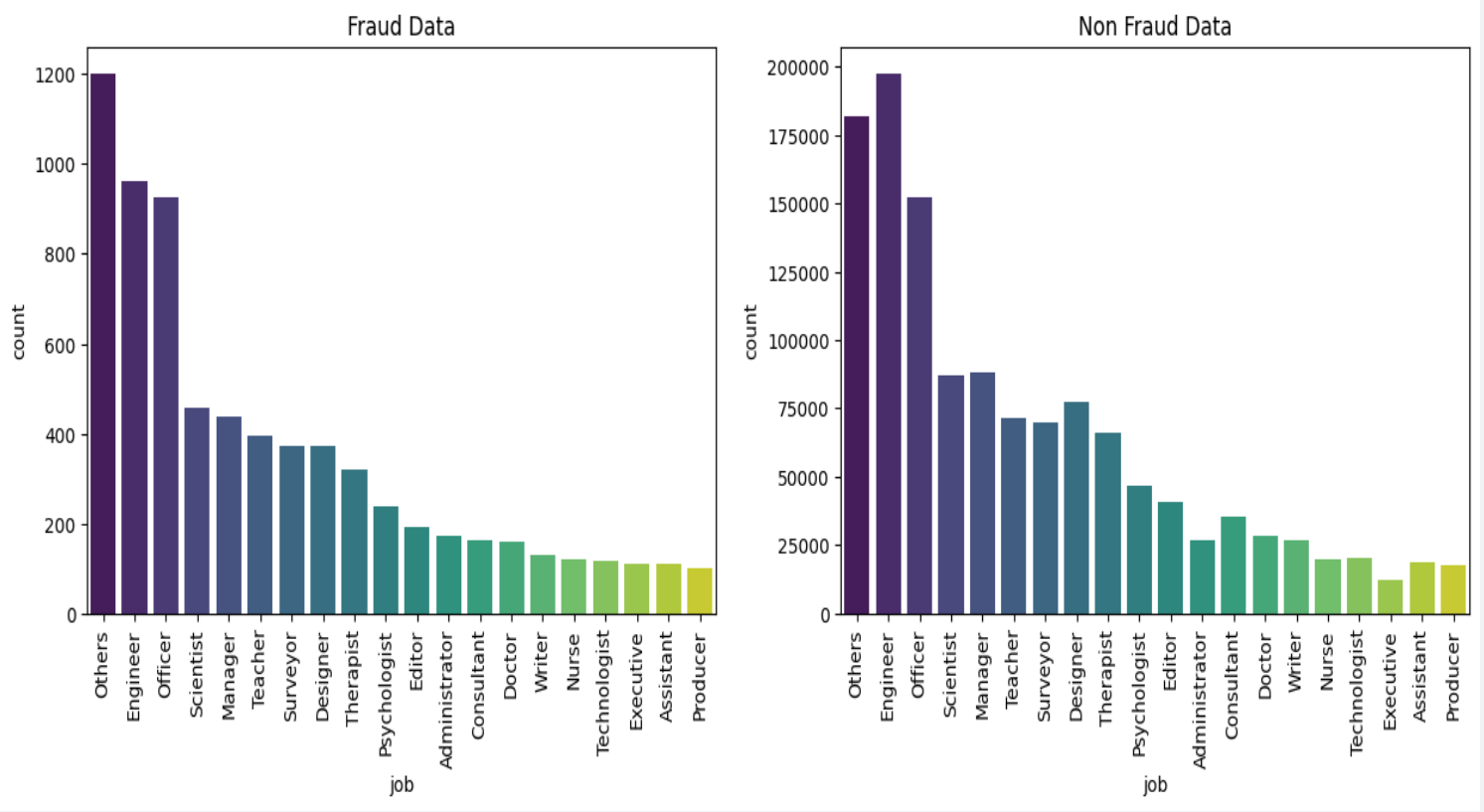


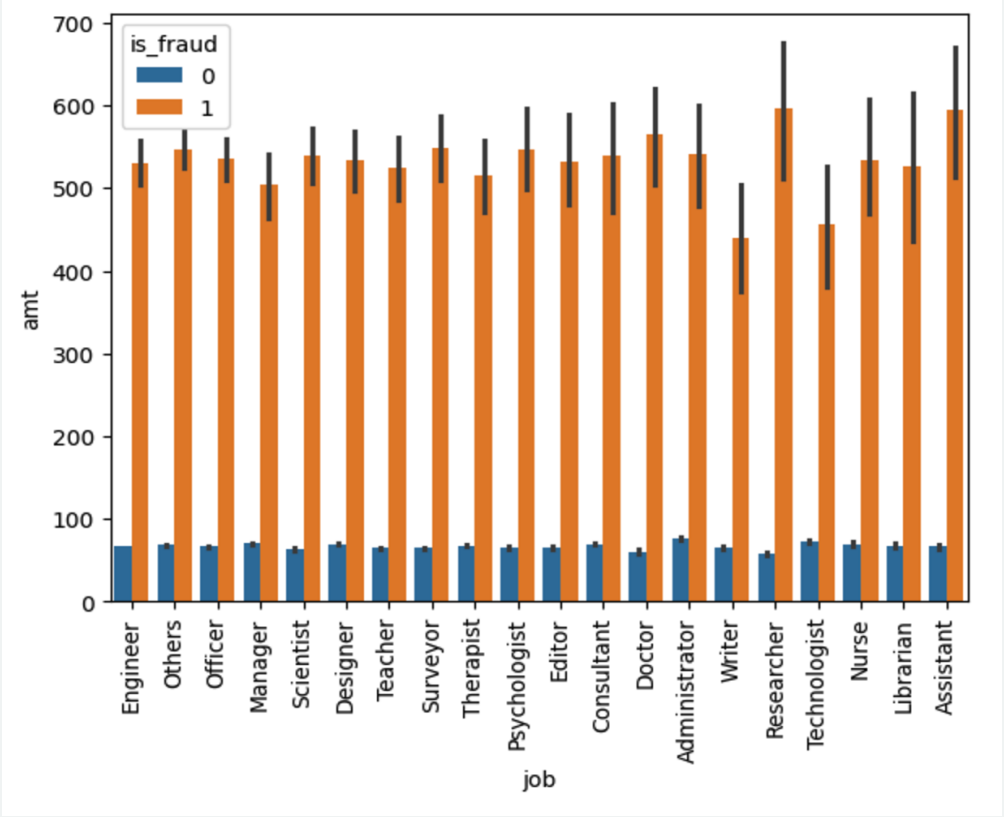


Fraudulent transactions are more frequent in the grocery_pos, shopping_net, and misc_net categories.

The highest fraudulent amount is spent on shopping_net, followed by shopping_pos and misc_net.

ANALYSIS BASED ON DIFFERENT JOB CATEGORIES

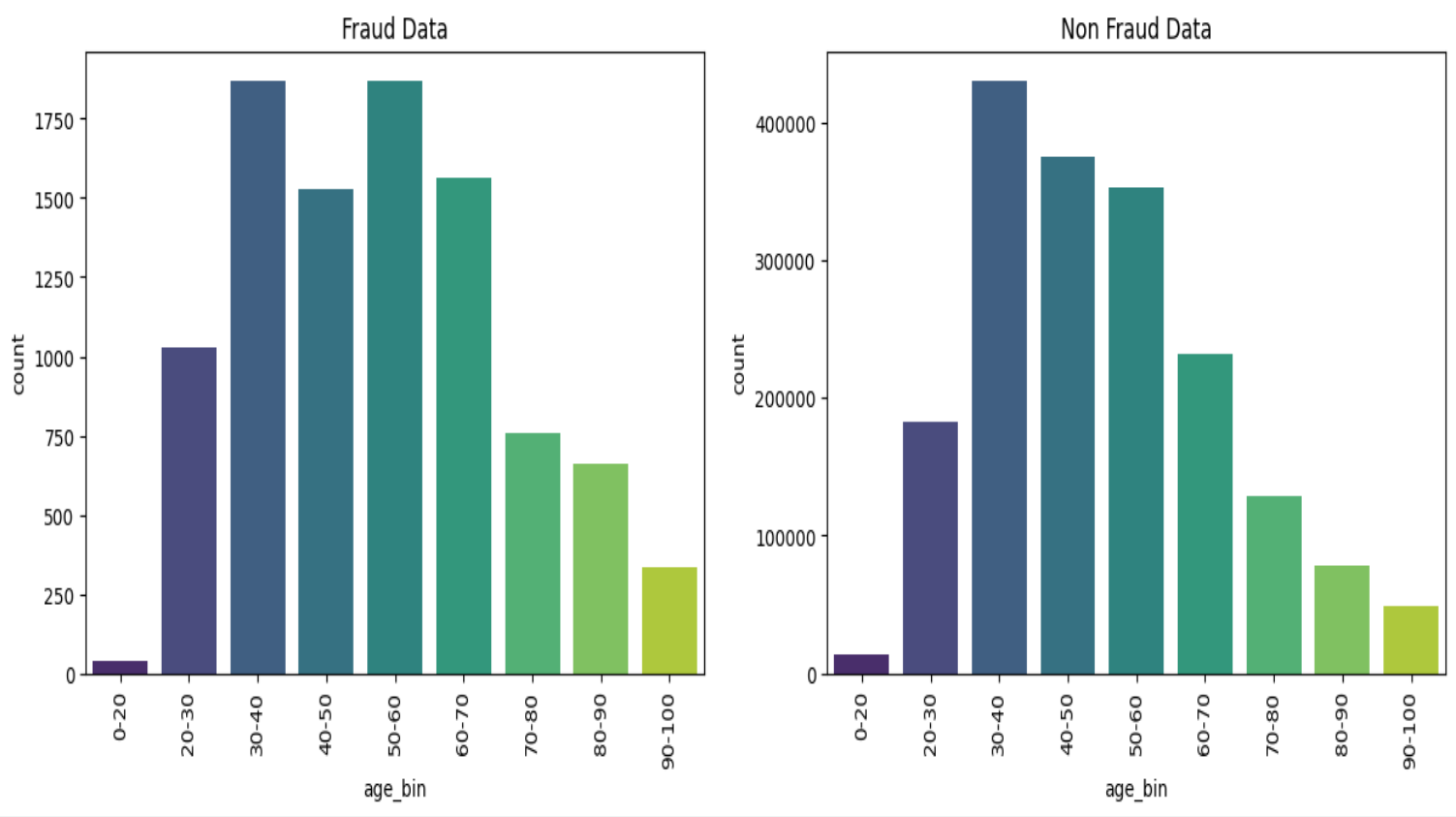


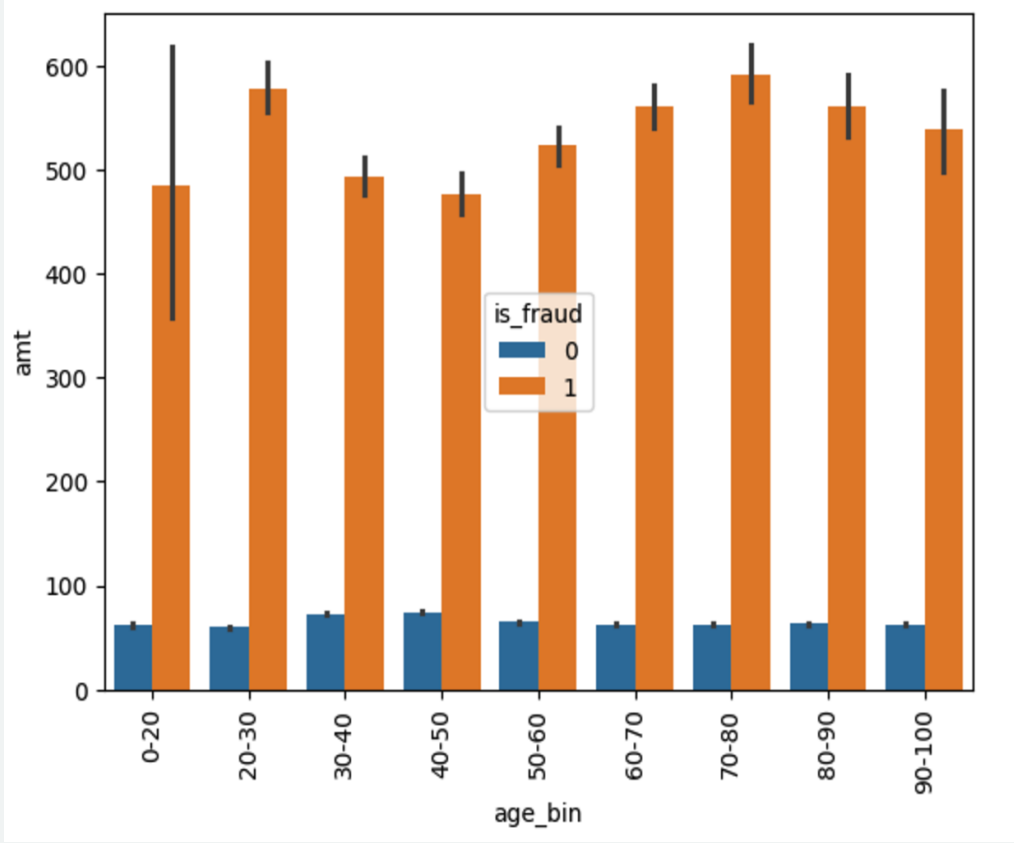


Fraudulent transactions is highest in Others job categories followed by Engineer, Officer.

The amount spent on fraudulent transactions is highest on credit card holder under the Researcher, Assistant job categories.

ANALYSIS BASED ON AGE GROUP

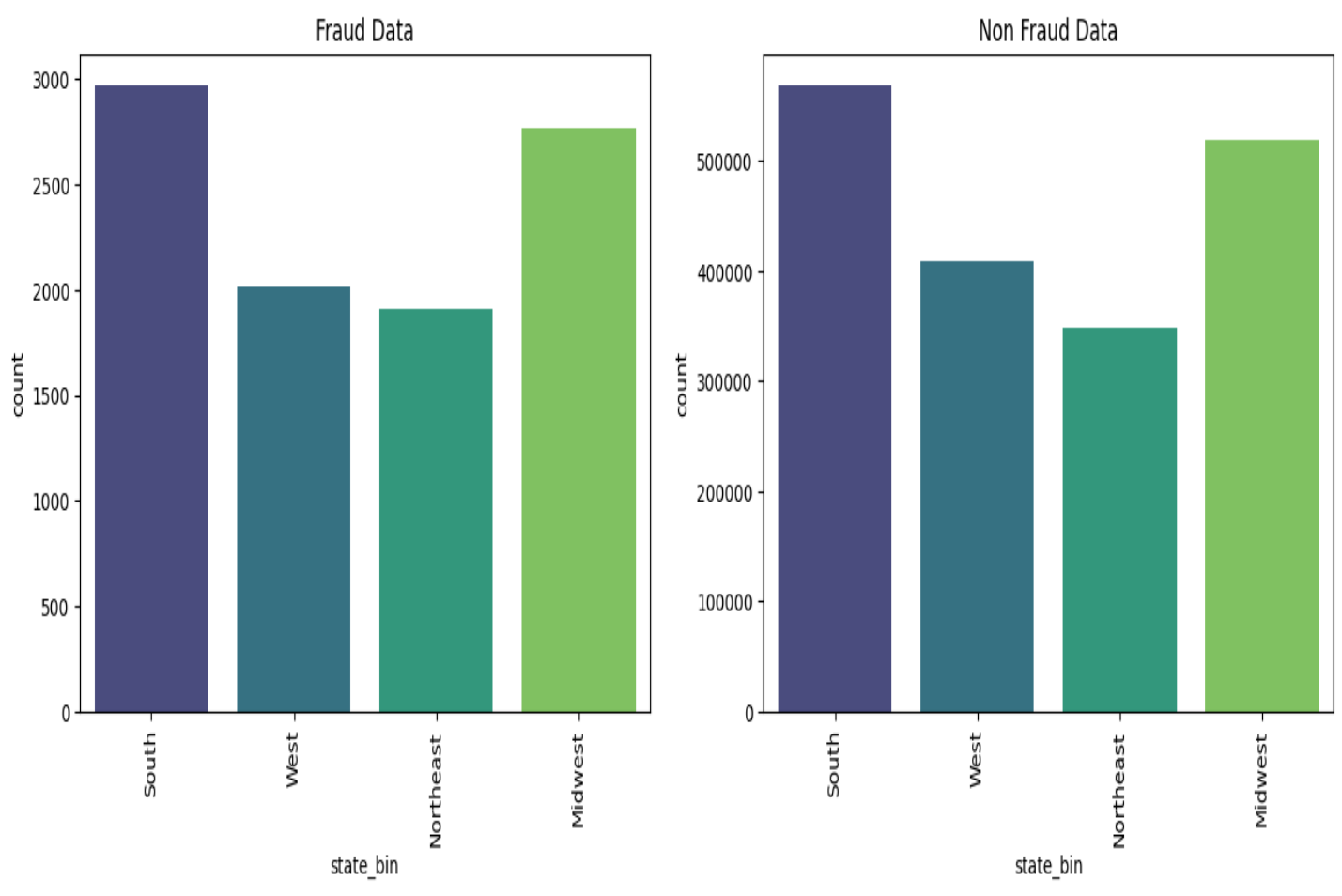




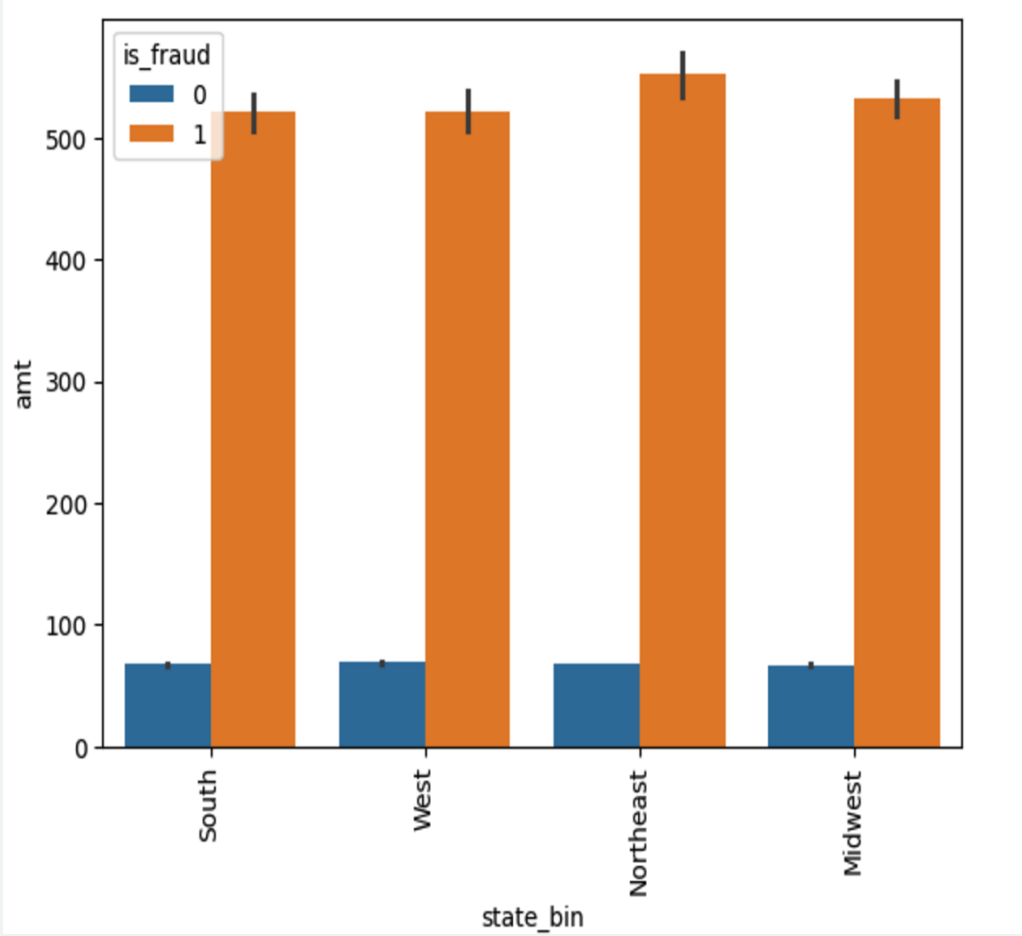
Fraudulent transactions are more in the age group 50-60, age group 30-40 and age group 20-60

The amount spent on fraudulent transactions is highest on credit card holders “70-80” followed by “20-30”.

ANALYSIS BASED ON DIFFERENT REGIONS

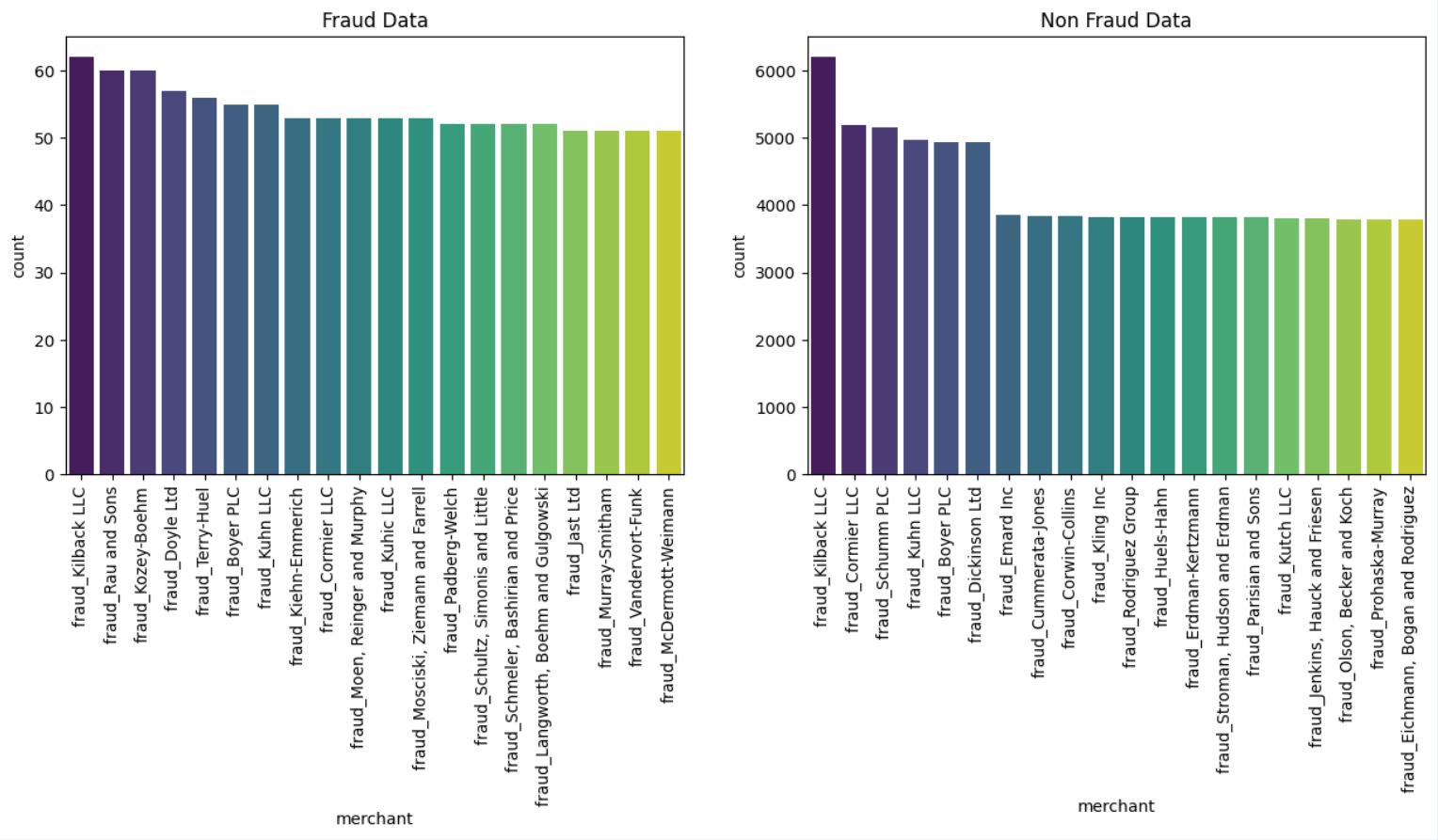


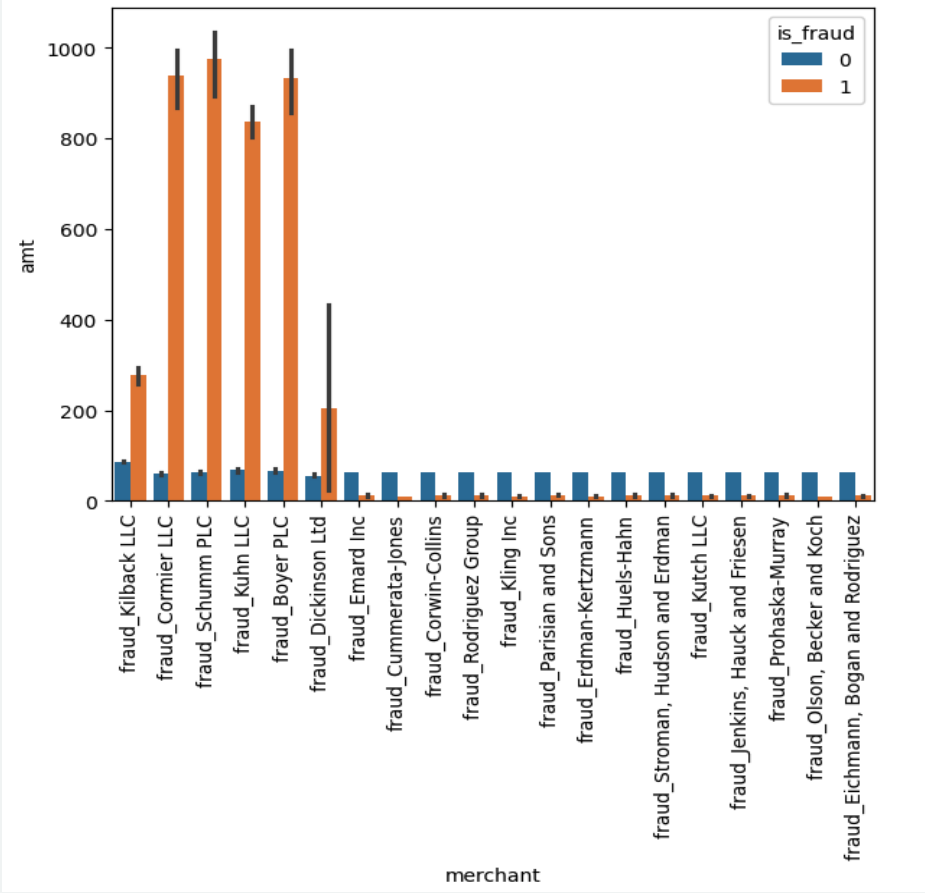
The number of fraudulent transactions is higher in the South and Midwest regions.



The amount spent on fraudulent transactions is similar across all regions, except for the Northeast region, which shows a slightly higher expenditure.

ANALYSIS BASED ON DIFFERENT MERCHANT



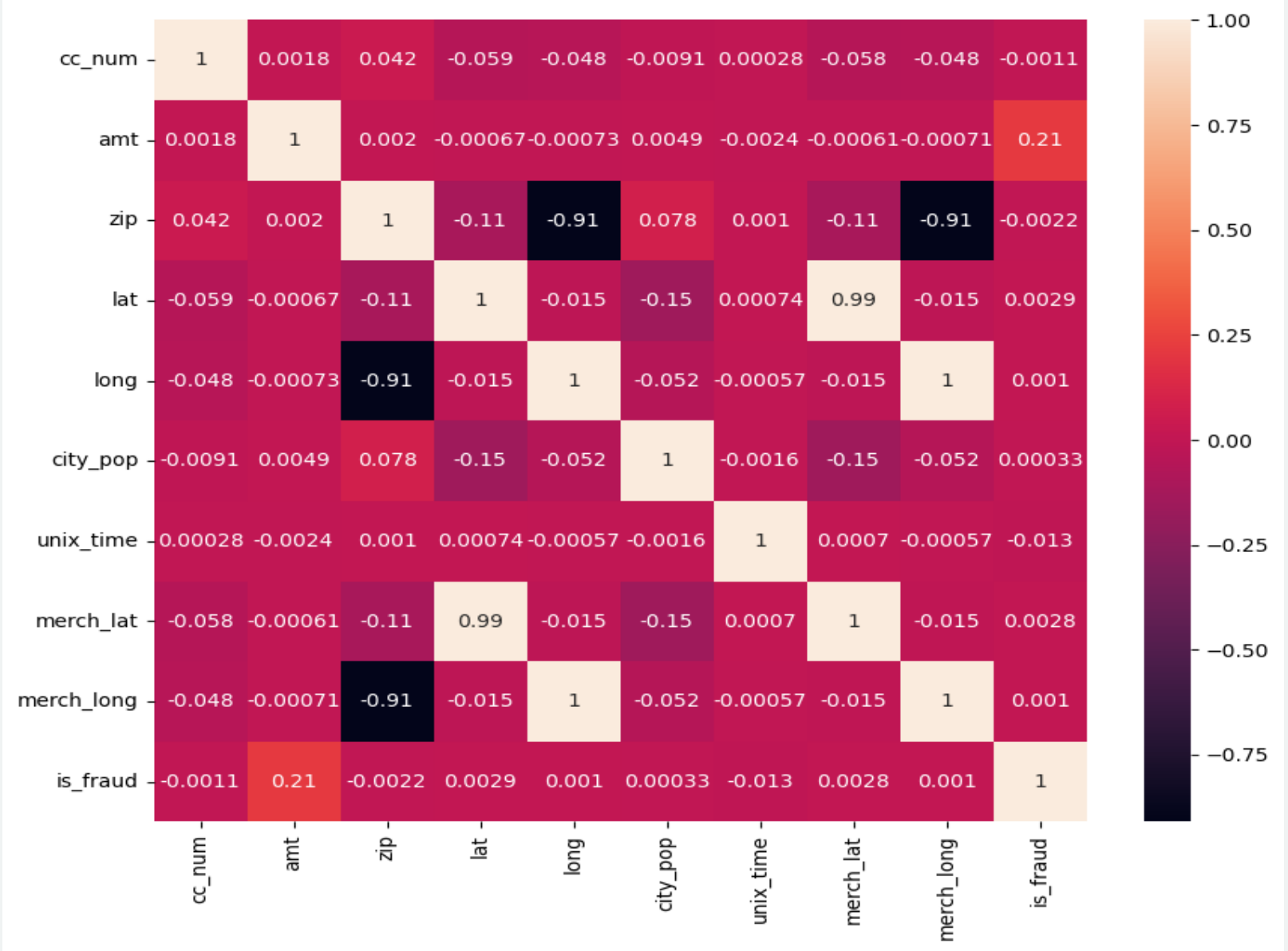


The fraud transactions is more in number at 'fraud_Kilback LLC', followed 'fraud_Rau and Sons', 'fraud_Kozey-Boehm'.

The amount spend for fraud transactions is highest at 'fraud_Schumm PLC', followed by 'fraud_Kilback LLC', 'fraud_Cormier LLC' Merchant.

MULTI VARIATE ANALYSIS

- The amt is positively correlated with is_fraud.
- The lat is positively correlated with merch_lat.
- The zip is negatively correlated with the long.



MODELLING STATS

Alogrithm	Accuracy	Recall	F1-Score	Precision	Specificity
Decision Tree with default SMOTE	99.6	84.7	66.4	54.6	99.6
Decision Tree with default ADASYN	99.6	83.6	67.0	55.9	99.7
Decision Tree with Tuned SMOTE	99.1	87.4	51.2	36.3	99.2
Random Forest with default SMOTE	99.8	79.7	84.2	89.3	99.9
Random Forest with default ADASYN	99.8	78.3	83.9	90.5	100
Random Forest with Tuned SMOTE	99.6	84.7	68.2	57.1	99.6
XGBoost with default SMOTE	**99.8**	**90.7**	**86.3**	**82.3**	**99.9**
XGBoost with default ADASYN	99.8	89.9	84.6	79.8	99.9
XGBoost with Tuned SMOTE	**99.8**	**90.7**	**86.3**	**82.3**	**99.9**

Both the **XGBoost model with default SMOTE** and the **XGBoost model with tuned SMOTE** have the same results. Since our goal is to detect fraudulent transactions, we will choose the model with the higher recall metric.

MODEL EVALUATION

- Based on the accuracy, ROC, precision, and recall of different models, we have selected **XGBoost (with Hyperparameter Tuning) for SMOTE data** as our final model.

- The model achieves a test accuracy of **99.8%**, a **recall of 90.7%**, and an **ROC of 99.8%**.

- Given that our business objective prioritizes the identification of fraudulent transactions over non-fraudulent ones, a high recall is critical. With a recall of 90.7%, the XGBoost model accurately identifies nearly all fraudulent transactions, making it the best choice for our needs.

- Therefore, **XGBoost (Hyperparameter Tuning) on SMOTE data** is chosen based on its superior performance in detecting fraud, as reflected by its high recall metric.

Clasification Report for Decision Tree on Test data on default Hyperparameter

	precision(%) Positive predictive value	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative (%) predictive value
SMOTE data	54.6	84.7	66.4	99.6	92.2	99.6	0.4	99.9
ADASYN data	55.9	83.6	67.0	99.6	91.6	99.7	0.3	99.9

Clasification Report for Decision Tree on Test data on SMOTE Hyperparameter Tunning

	precision(%) Positive predictive value	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative (%) predictive value
SMOTE data	36.3	87.4	51.2	99.1	94.7	99.2	0.8	99.9

2. Clasification Report for Random forest on Test data on default Hyperparameter

	precision(%) (Positive predictive value)	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative (%) predictive value
SMOTE data	89.3	79.7	84.2	99.8	99.5	99.9	0.1	99.9
ADASYN data	90.5	78.3	83.9	99.8	99.6	100	0.0	99.9

Clasification Report for Random Forest on Test data on SMOTE Hyperparameter Tunning

	precision(%) Positive predictive value	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative (%) predictive value
SMOTE data	57.1	84.7	68.2	99.6	99.0	99.7	0.3	99.9

3. Clasification Report for XGBoost on Test data on default Hyperparameter

	precision(%) (pp value)	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative (%) predictive V
SMOTE data	82.3	90.7	86.3	99.8	99.8	99.9	0.1	100.0
ADASYN data	79.8	89.9	84.6	99.8	99.7	99.9	0.1	99.9

Clasification Report for XGBoost on Test data on SMOTE Hyperparameter Tunning

	precision(%) +ve predictive value	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	-ve (%) predictive
SMOTE data	82.3	90.7	86.3	99.8	99.8	99.9	0.1	100.0

COST BENEFIT ANALYSIS

- Cost Analysis On Whole Data

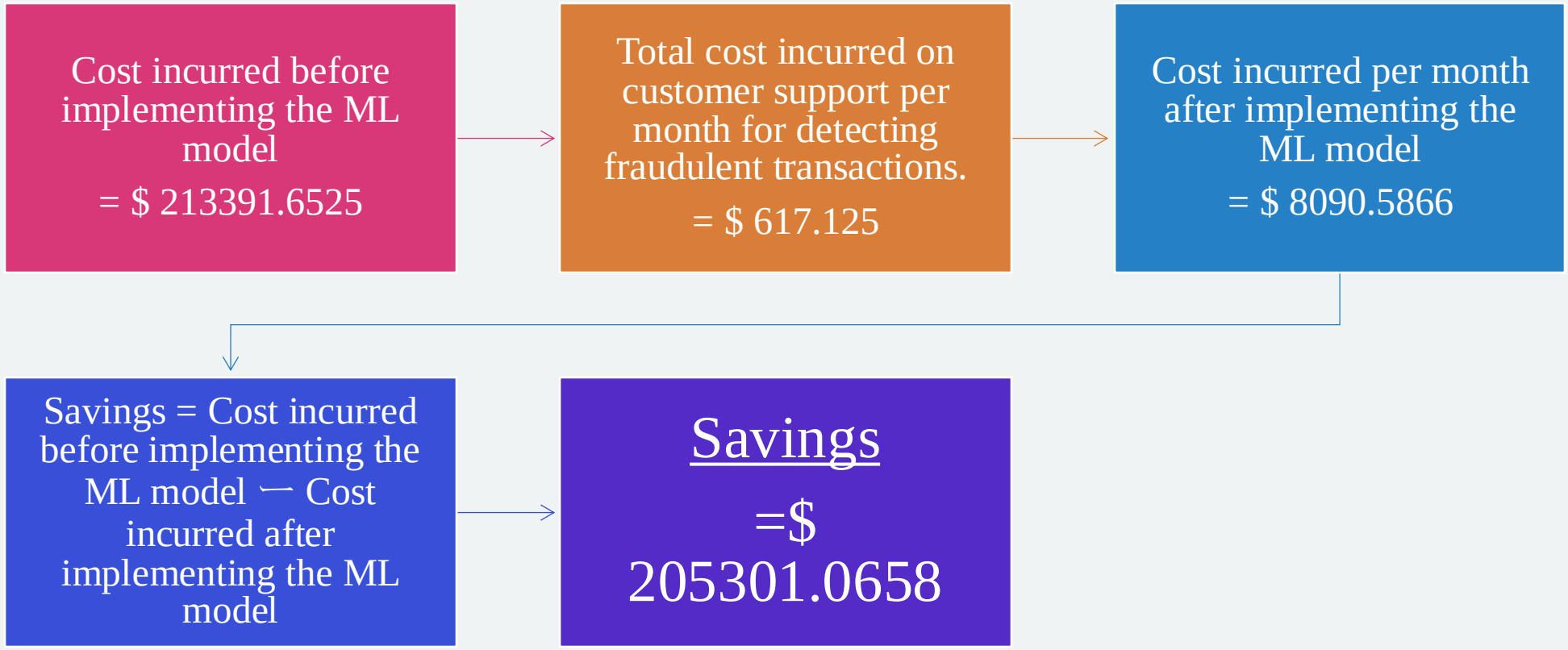
Sr. No.		
1)	Avearge number of transactions per month	77183.0
2)	Avearge number of fraudulent transactions per month	402.0
3)	Average amount per fraud transaction	$530.66

- Cost Analysis After Modelling

Sr. No.		
1)	Cost incurred per month before the model was deployed	$213391.6525
2)	Average number of transactions per month detected as fraudulent by the model	411.0
3)	Cost of providing customer executive support per fraudulent transaction detected by the model	$1.5
4)	Total cost of providing customer support per month for fraudulent transactions detected by the model	$617.125
5)	Average number of transactions per month that are fraudulent but not detected by the model	14.0
6)	Cost incurred due to fraudulent transactions left undetected by the model	$8623.4616
7)	Cost incurred per month after the model is built and deployed	$8090.5866
8)	Final savings = Cost incurred before - Cost incurred after	$205301.0658

CONCLUSION

- Conclusion ⇨ Profit predicted after implementing the ML model.



- There has been a 95.67% reduction in the amount paid by the bank to customers for losses due to fraudulent transactions, thanks to the implementation of the ML model.

BUSINESS RECOMMENDATION

1) The probability of fraudulent transactions increases with a higher hist_trans_avg_amt_24h value, which represents the average amount spent by the credit card holder in the last 24 hours. If the amount spent in the last 24 hours significantly exceeds the historical spending pattern, it is crucial for the bank to send an SMS alert to the customer, confirming the transaction activity and ensuring its legitimacy.

2) The probability of fraudulent transactions increases on Thursday, Saturday, and Monday, as these days show a higher incidence of fraud based on pattern analysis. Therefore, banks should exercise heightened vigilance and take extra precautions on these specific days to reduce the risk of fraudulent transactions.

3) The probability of fraudulent transactions increases with higher transaction amounts. If a bank detects that a customer's spending exceeds their typical spending pattern, it should promptly identify this anomaly and send the necessary alerts to the customer at an early stage.

4) The probability of fraudulent transactions increases in categories such as catg_home, catg_shopping_pos, catg_grocery_pos, catg_health_fitness, and catg_gas_transport. The model indicates that these categories are common targets for both legitimate high-value transactions and fraudulent activities, as fraudsters often follow similar spending patterns. Therefore, it is recommended that banks closely monitor spending in these categories and proactively send FLASH SMS ALERTS to customers, detailing the transaction history and ensuring the legitimacy of the charges.

5) Fraudulent transactions typically occur during off-peak hours, specifically between 22:00 and 03:00. Therefore, banks should prioritize sending SMS alerts for any transactions during these timeframes to enhance security.