FRAUD DETECTION

PRESENTED BY:-

RAMA MISHRA

AGENDA

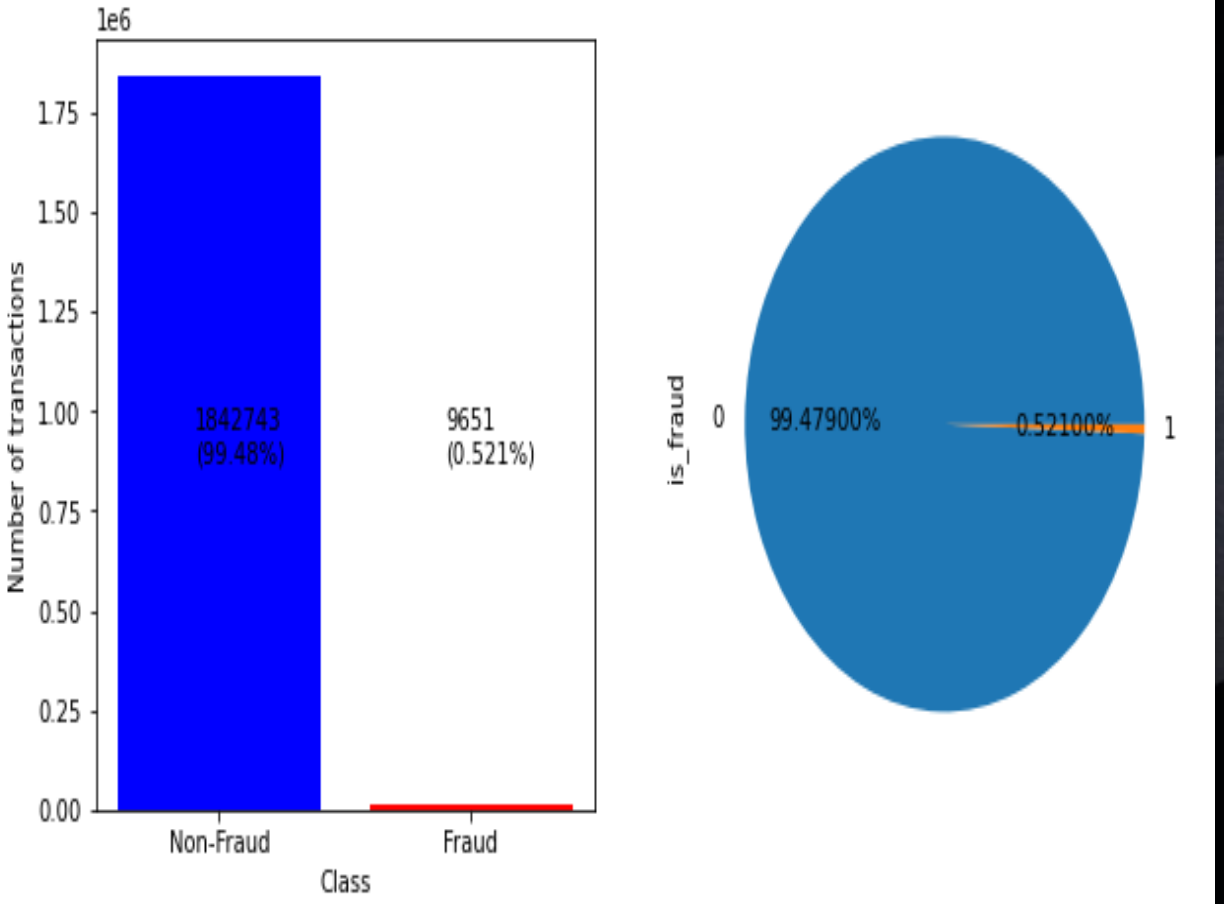
- OBJECTIVE
- OVERVIEW OF DATA
- INSIGHTS
- DATA MODELLING
- BUSINESS RECOMMENDATION
- COST BENEFIT ANALYSIS

Analyze the business impact of these fraudulent transactions and recommend the optimal ways that the bank can adopt to mitigate the fraud risks

Problem Statement

- Detecting credit card fraud using machine learning is a must in banking industry.
- They need to put proactive monitoring and fraud prevention mechanisms in place.
- Machine learning fraud detection algorithms are way more effective than humans.

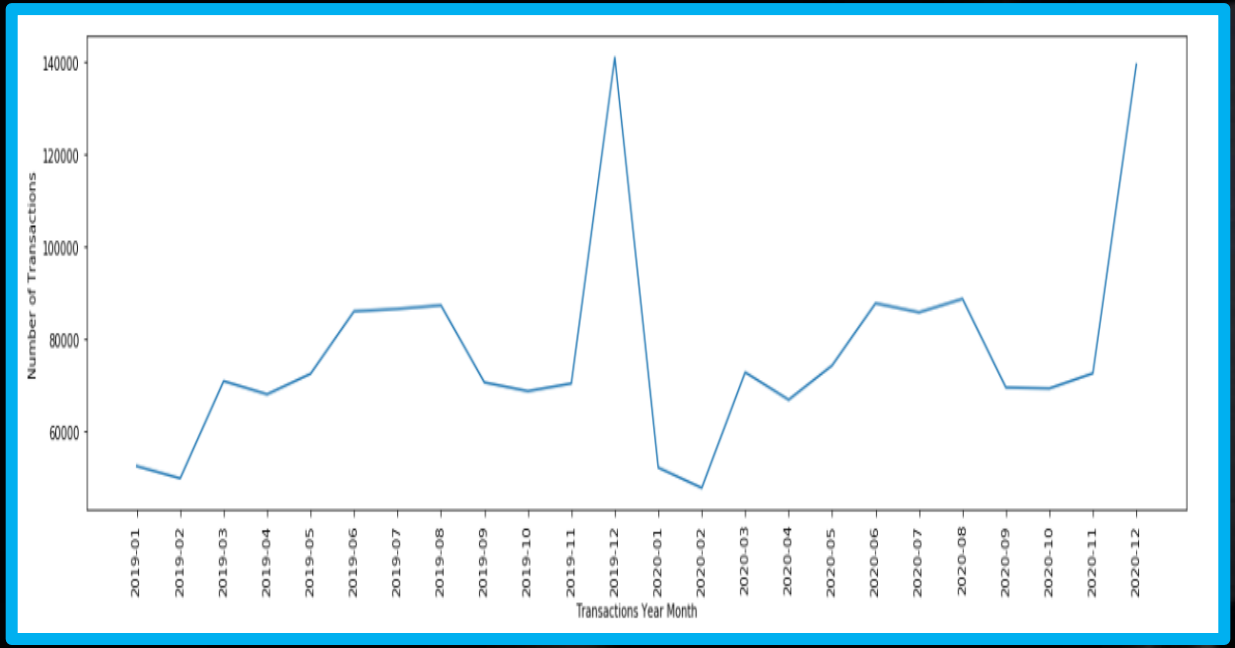
OVERVIEW OF DATA



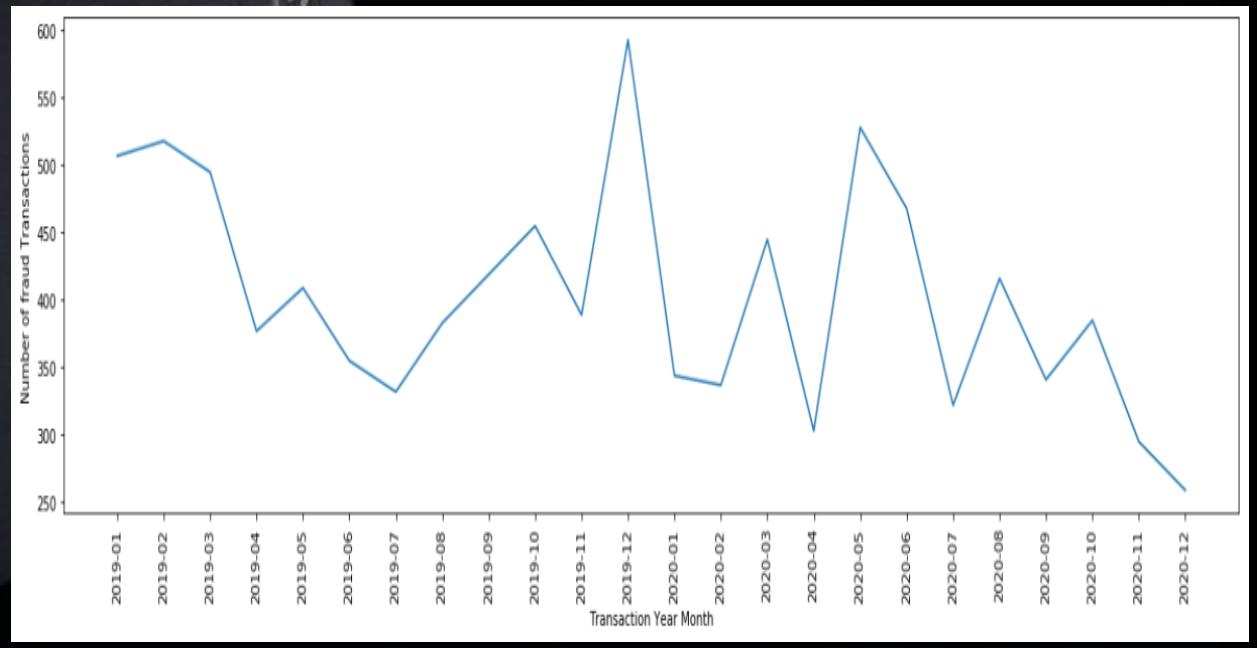
1. Fraud data = 9651 transaction (0.521%)
2. 2. Non-Fraud_data= 1842743 transactions(99.48%)

Visual representation of frequency of transactions and fraud_transactions month_year wise

No of transactions done month_year wise



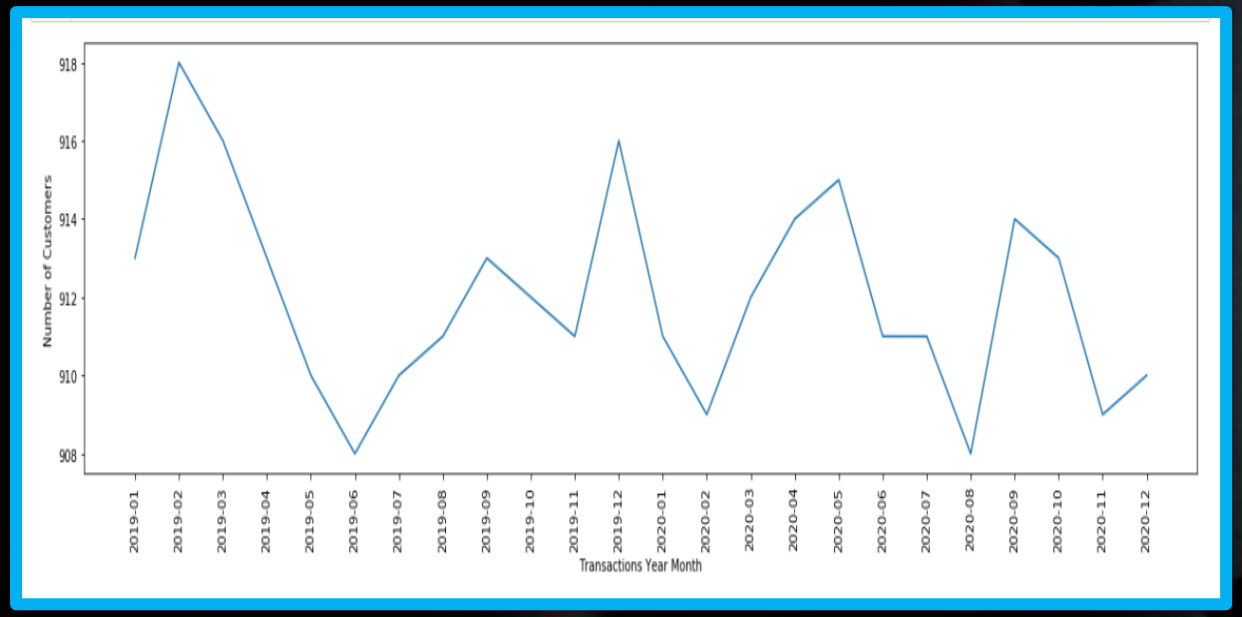
No of fraud transactions done month_year wise



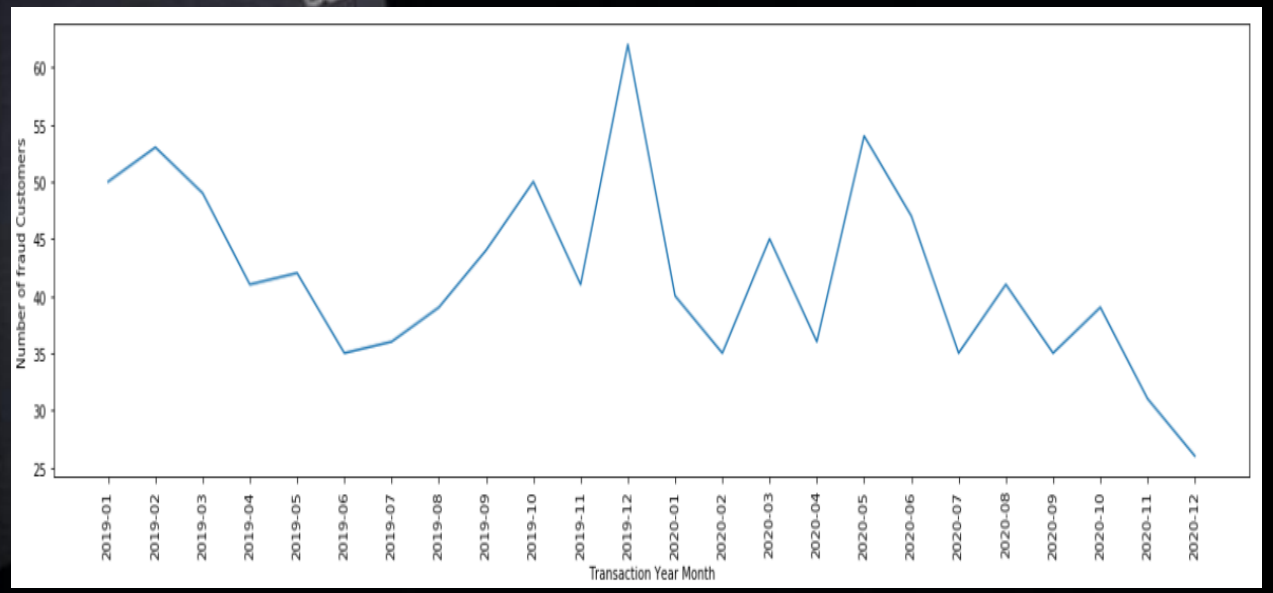
- No of transactions done month_year wise has increased at dec month of 2019 and 2020, in rest of the months there are normal transactions
- Its been noticed that fraud transactions were on higher side during new year eve. Specifically in Dec month . Later peaks were observed in march, may, august and October months of 2020.

Visual representation of frequency of customers and fraud customers month_year wise

No of customers did transactions month_year wise



No of customers did fraud transactions month_year wise



- February month has highest no of customers latterly decreased and then increased in September and December 2019.
- In 2020 may and september month has higher number of customers who did transactions.
- Transactions done by Fraud customers are majorly in Feb_19, Oct_19, Dec_19, March_20 and May_20.

- ✓ Fraud Transactions are done in both **gender male and female** card holders, whereas the count of fraud transaction are slightly on higher side regards to female card holders. But maximum amount spend for fraudulent transactions are more against Male Credit card holders.

- ✓ The maximum amount spend for fruad transactions belongs to credit card holders "0-20", "60-70" and "70-80" **age bin**. The increased count of fraud transactions are noticed in the age group of 50-60.

- ✓ The maximum amount spend for fraud transactions were **people with the job** of credit card holders of Researcher & Assistant.

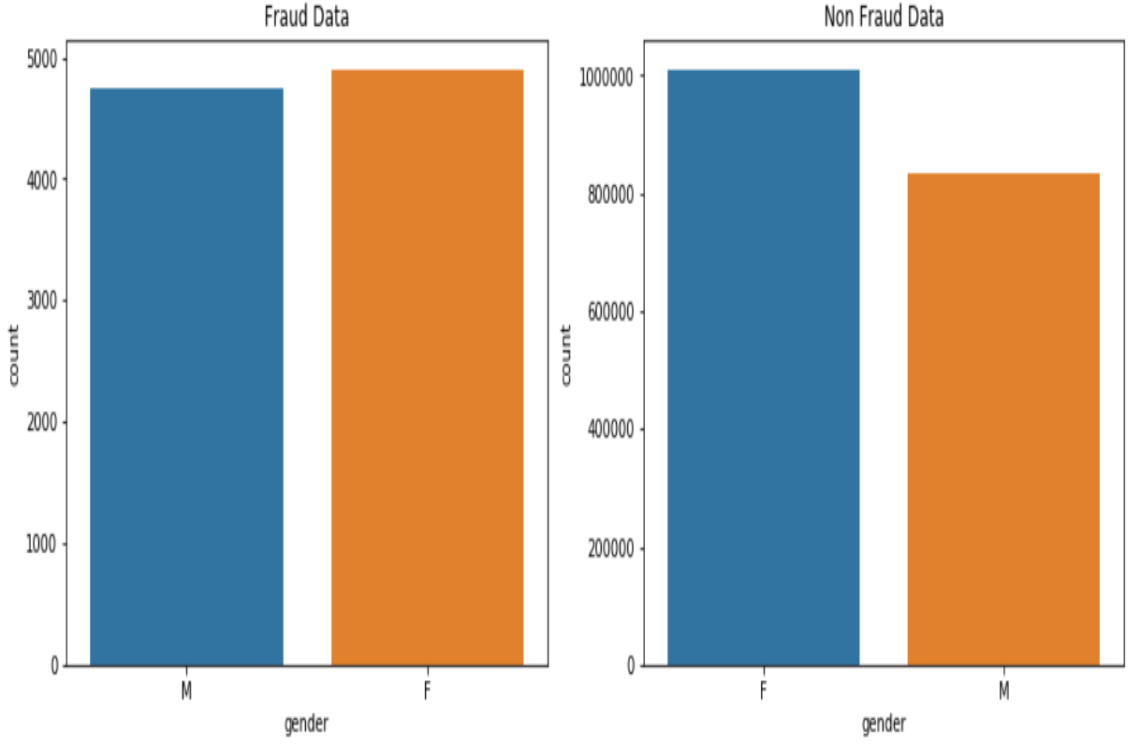
- ✓ Fraud Transactions are noticed more on **Saturday ,Sunday, Monday & Thursday** whereas the normal transaction observed are lesser on Saturday & Thursday and more on Sunday and Monday. Also as we see the spending pattern for fraud transactions are same throughout Week.

- ✓ Fraud transactions are majorly done during **odd hours** of the day i.e. between 22 - 3 Hr. But the maximum amount spend for doing fraudulent transactions are majorly between 12 to 23 Hr. Major fraudulent transactions happened at lower amount.

- ✓ Fraud transactions are done more at **grocery_pos, shopping_net, misc_net, shopping_pos, gas_transport Categories**.But the maximum fraud transaction amount spend were more on shopping_net, shopping_pos, misc_net category and entertainment.

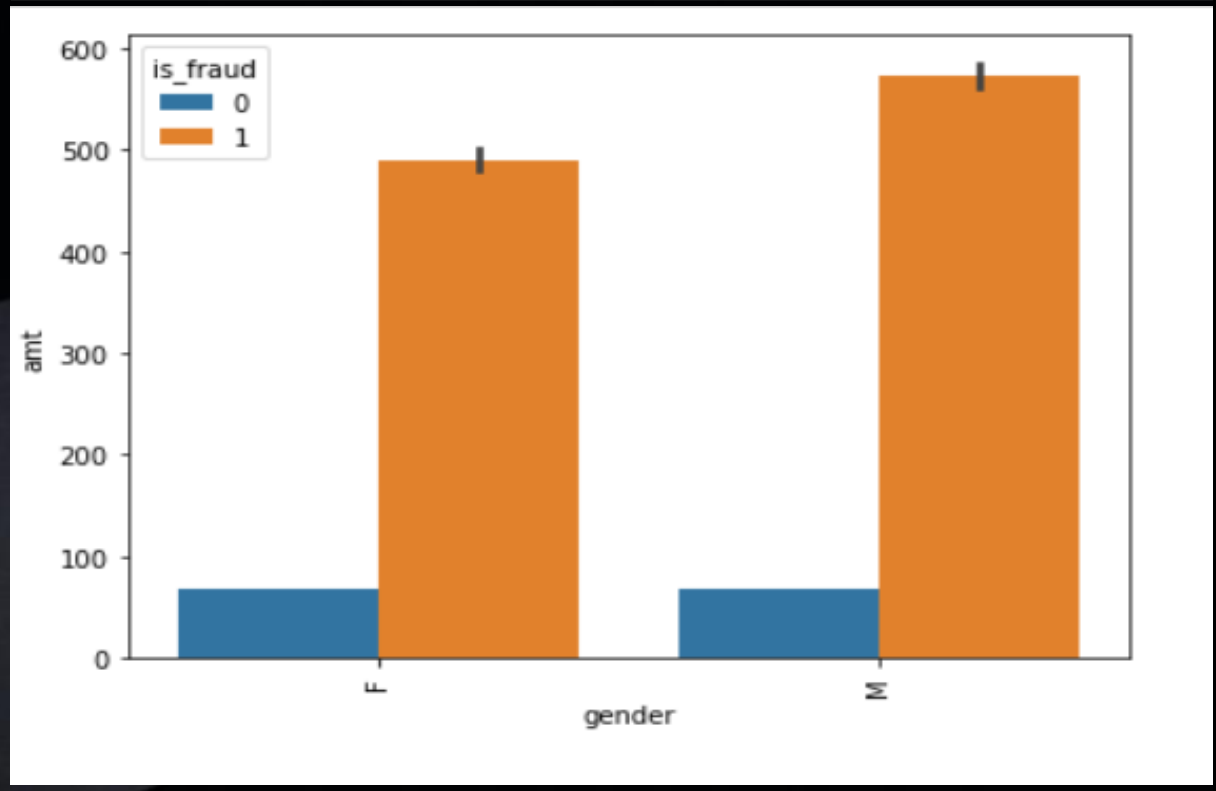
- ✓ Frauds transactions were observed more **in NY, TX and PA States**. The maximum amount spend for fraud transactions were at RI, HI, DE and VT State.

Fraud transaction based on gender



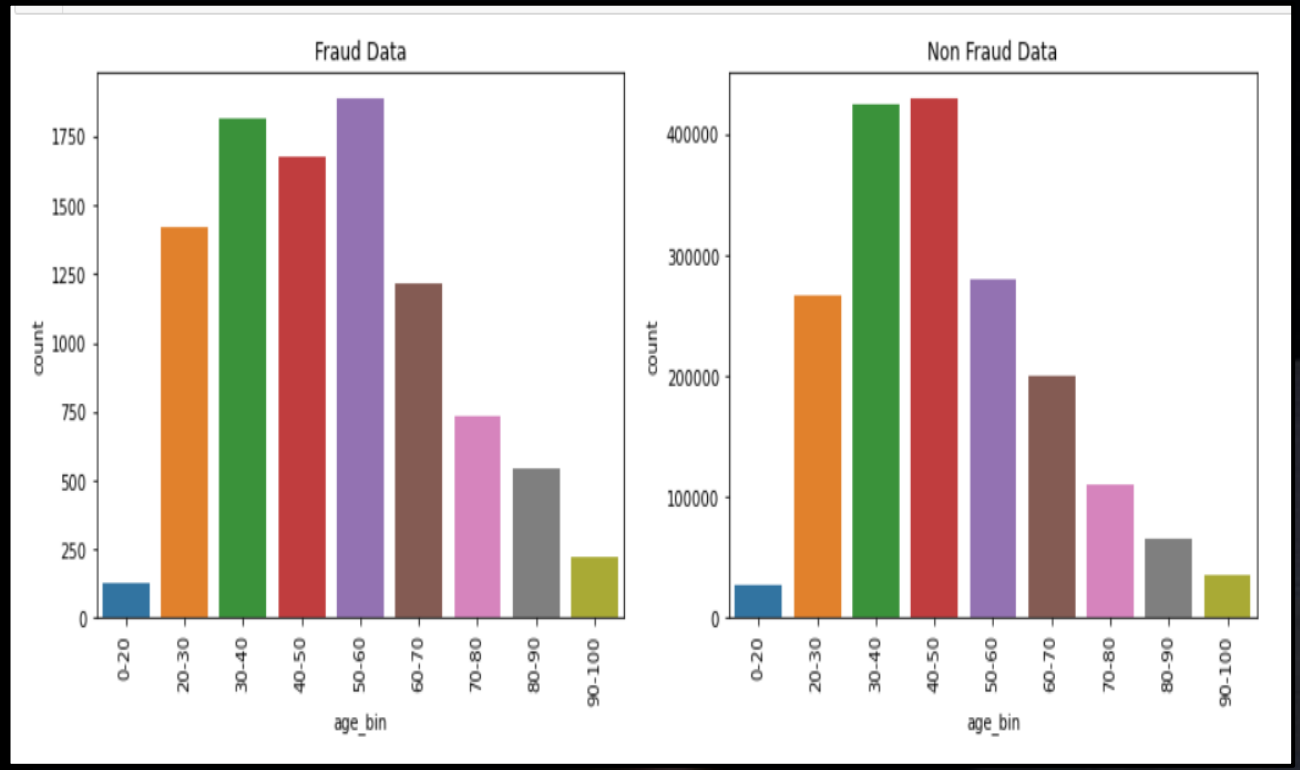
- Frauds Transcations are done on both male and female card holders
- But count of fraud transcations is slightly more on Female's credit card holders

Amount Spent in gender of credit card holders for Fraud Transactions data

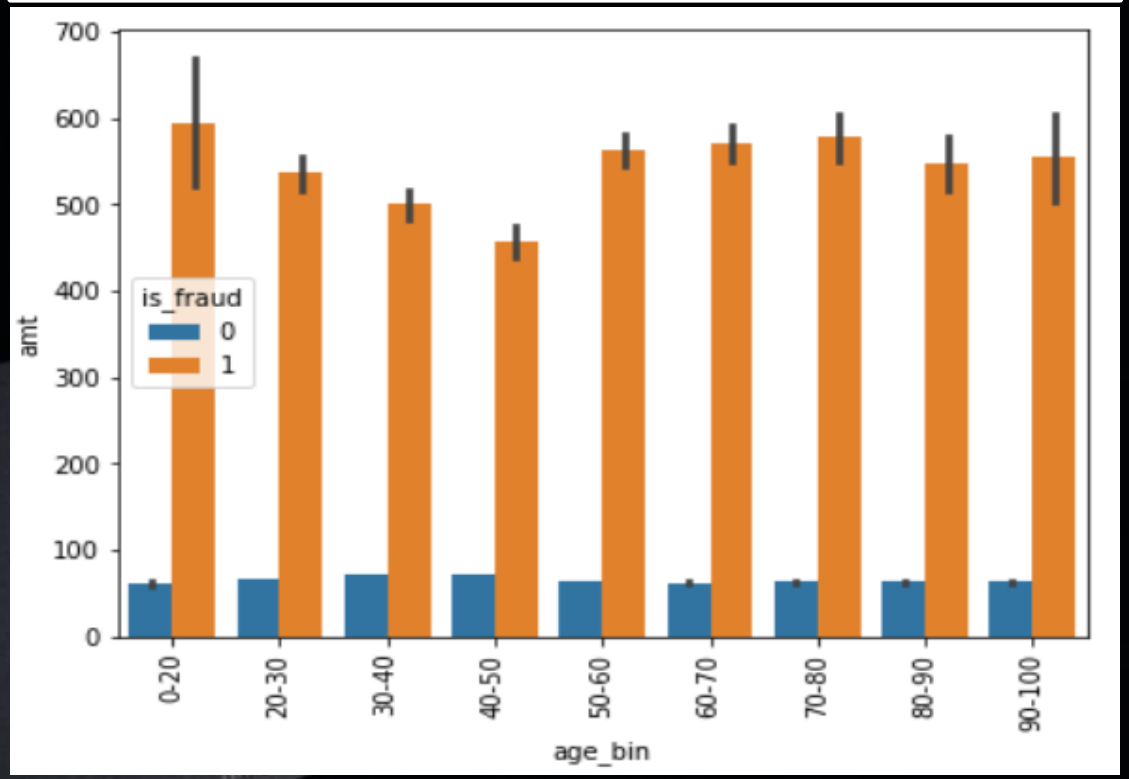


- The maximum amount spend for fruad transactions were done on Male's Credit card Holders.

Fraud Transaction in different Age_bins



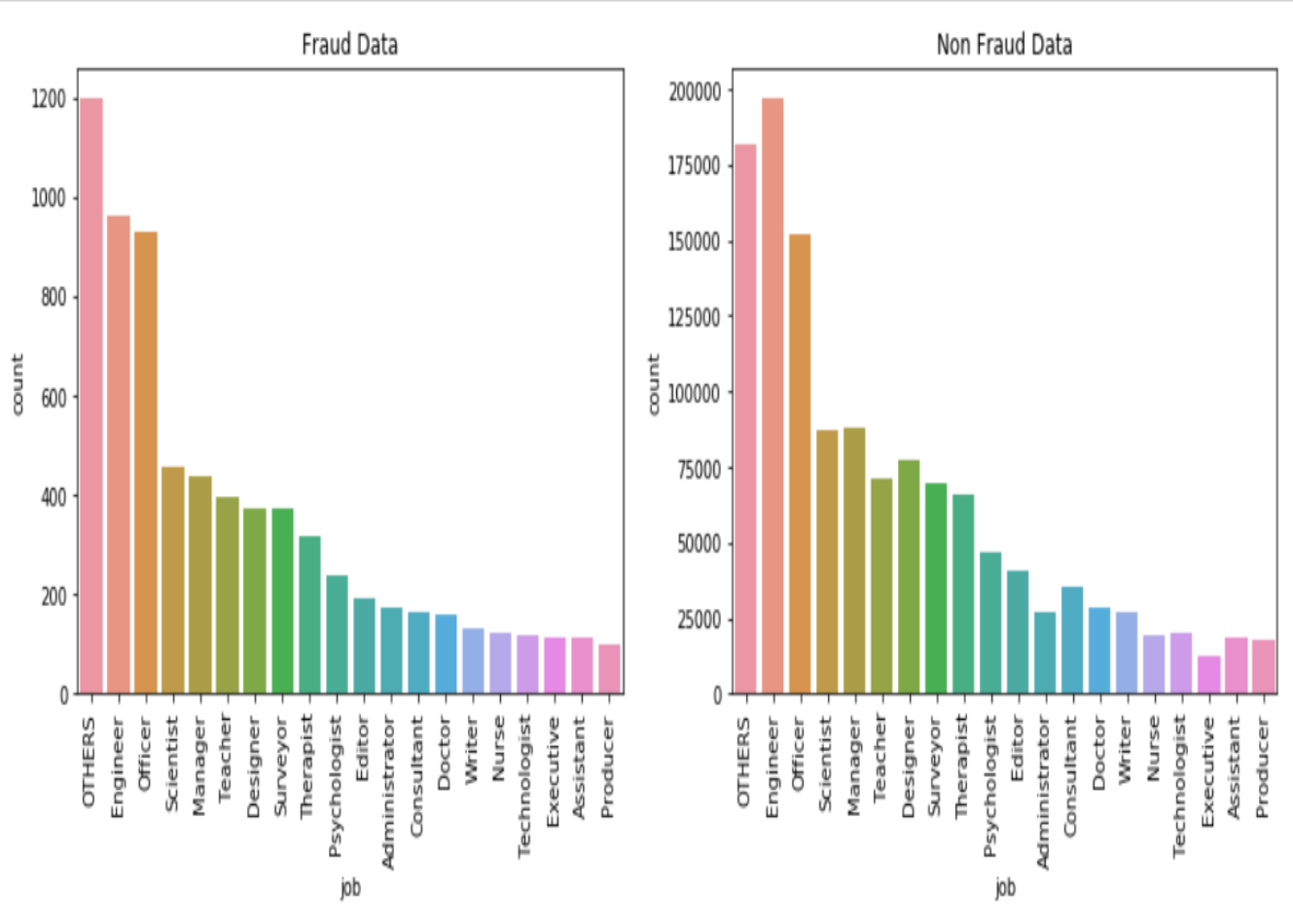
Amount Spent under age_bin for Fraud Transactions data



- The increased count of fraud transactions are noticed in the age group of 50-60 in comparison of non fraud transactional data by this group.
- The higher transactions and fraud rate are observed in the age group of 30-40
- Count of Frauds transactions are more on credit card holder's age group of 20 to 60

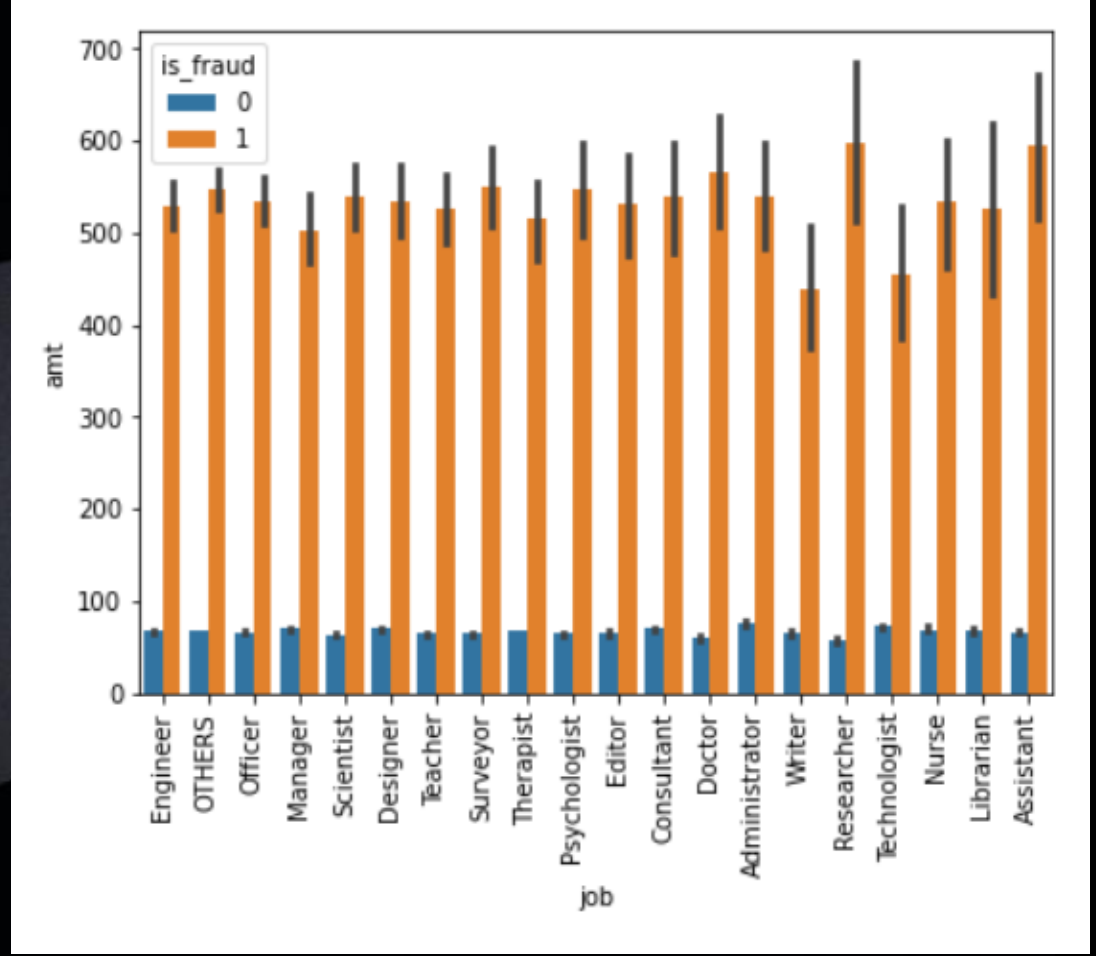
- The maximum amount spend for fruad transactions belongs to credit card holders "0-20", "60-70" and "70-80" age bin.

Fraud Transactions in different job



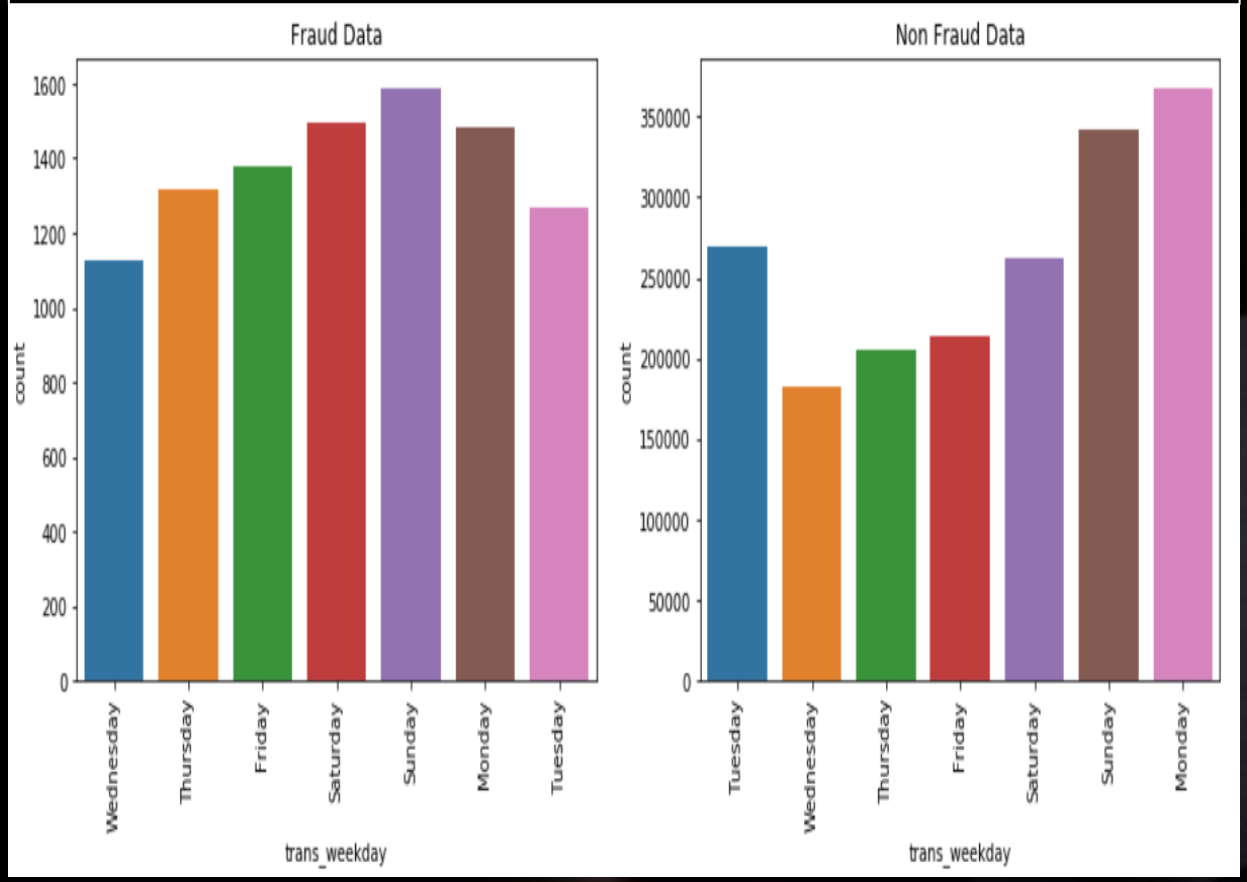
- Count of Frauds transactions are done more to people with occupations as Engineer, officer, others, scientist, Manager, Teacher.

Amount Spent of people with different job of credit card holders for Fraud Transactions data



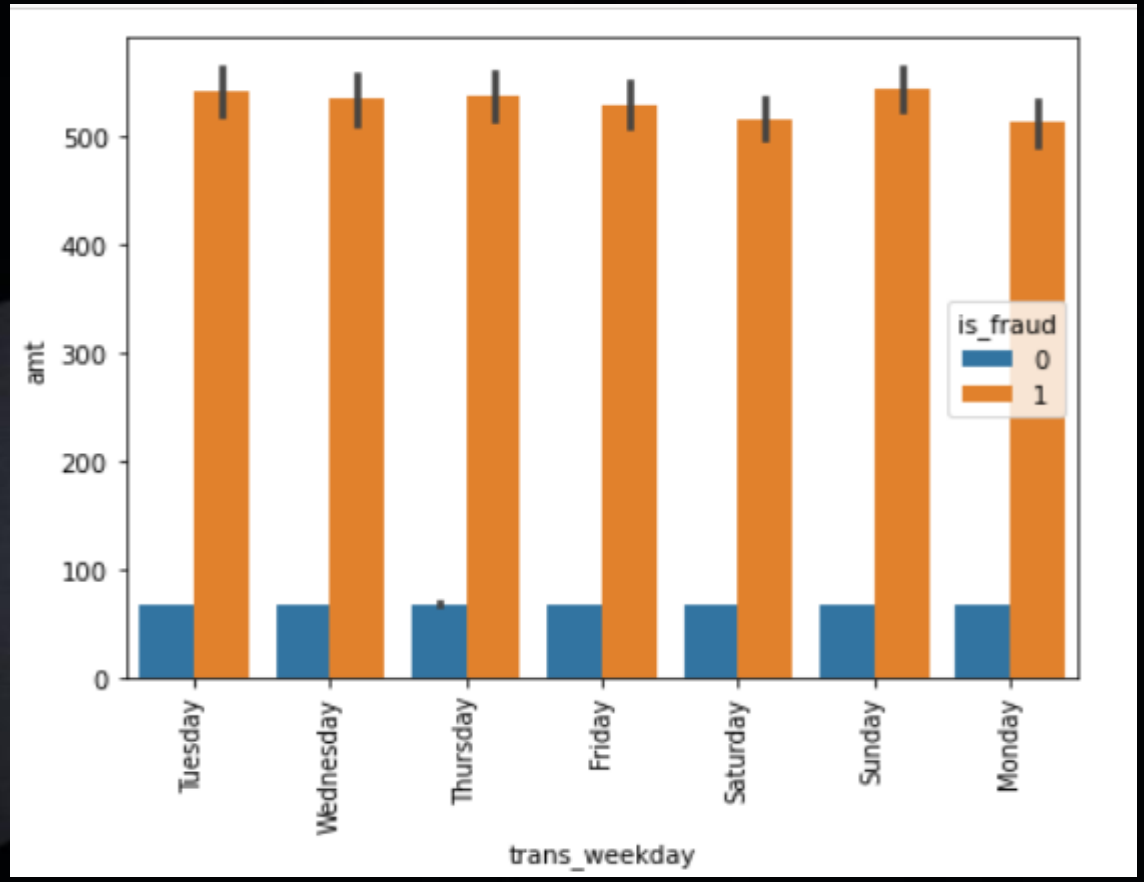
- The maximum amount spend for fraud transactions were of people with the job of credit card holders of Researcher, Assistant.

Fraud transaction in different weekdays



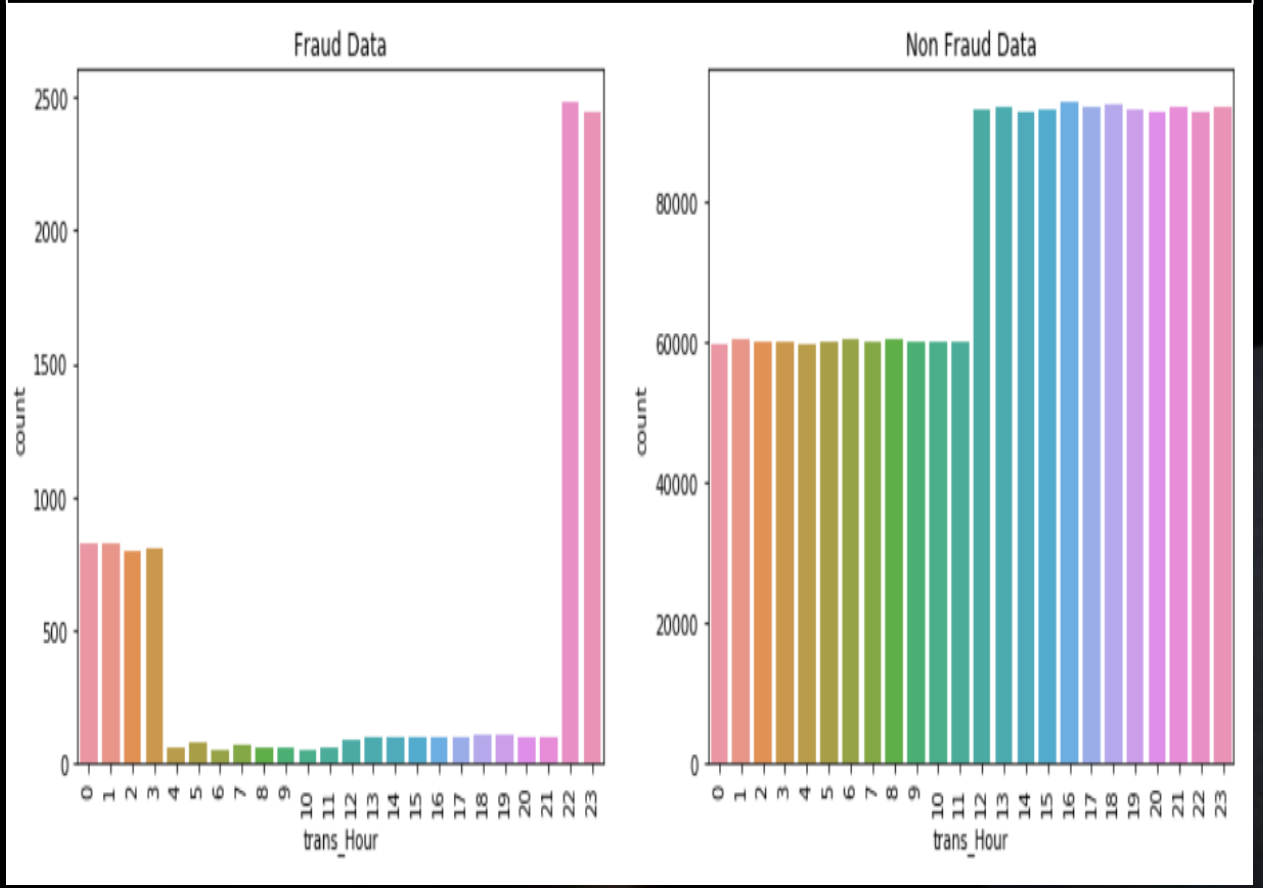
- Count of frauds transactions are more on Saturday, Sunday, Monday and Thursday whereas count of normal transaction are less on Thursday & Saturday and more on Sunday and Monday

Amount Spent at different trans_weekday for Fraud Transactions data



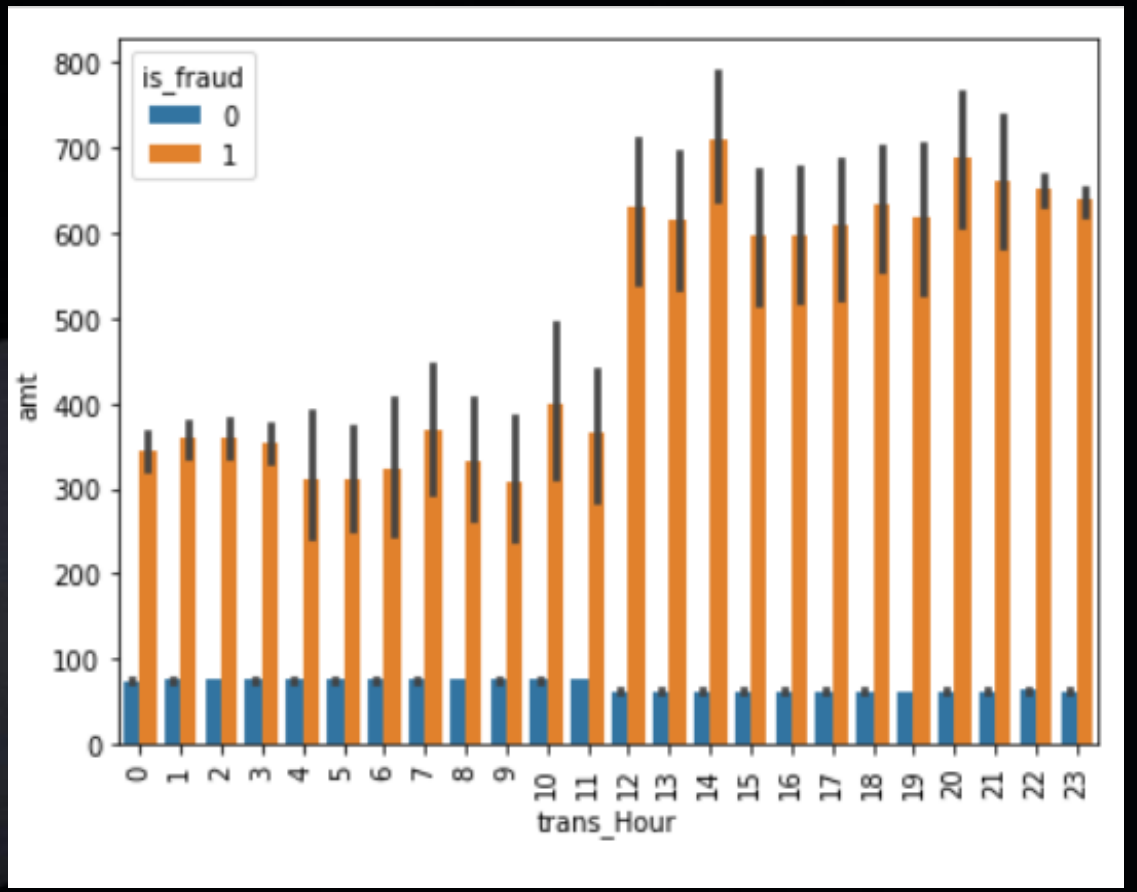
- Nearly same amount is being spend for fraud transactions throughout Week.

Hourly based Fraud Transactions



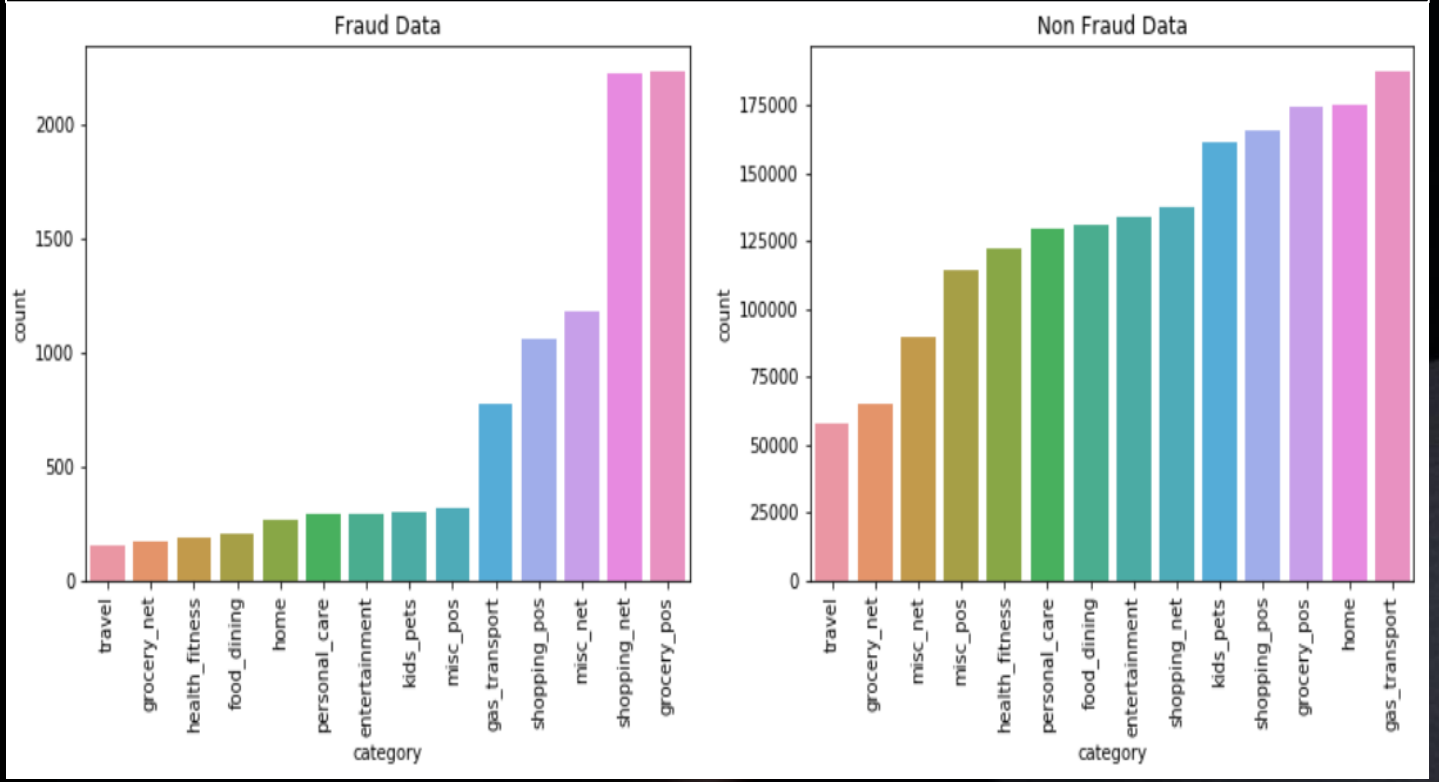
- Fraud transactions are done during odd hours of the day i.e. between 22 - 3 Hr

Amount Spent at different trans_Hour for Fraud Transactions data



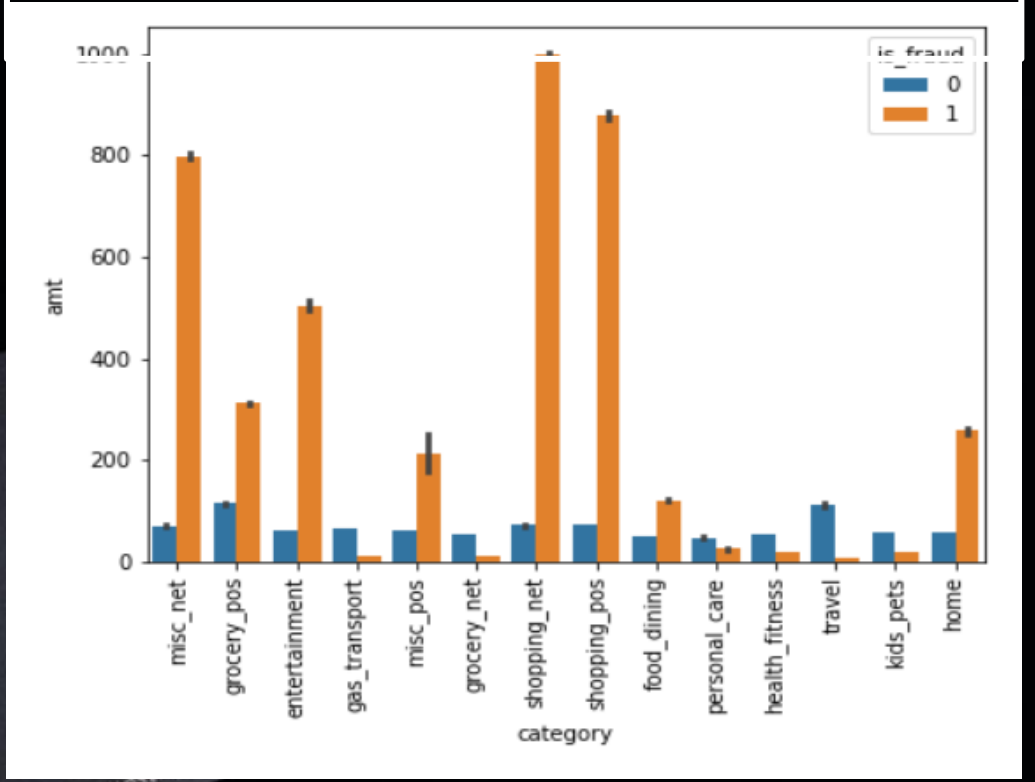
- The maximum amount spend for fraud transactions were done mostly between 12 to 23 Hr

Fraud Transactions in different categories



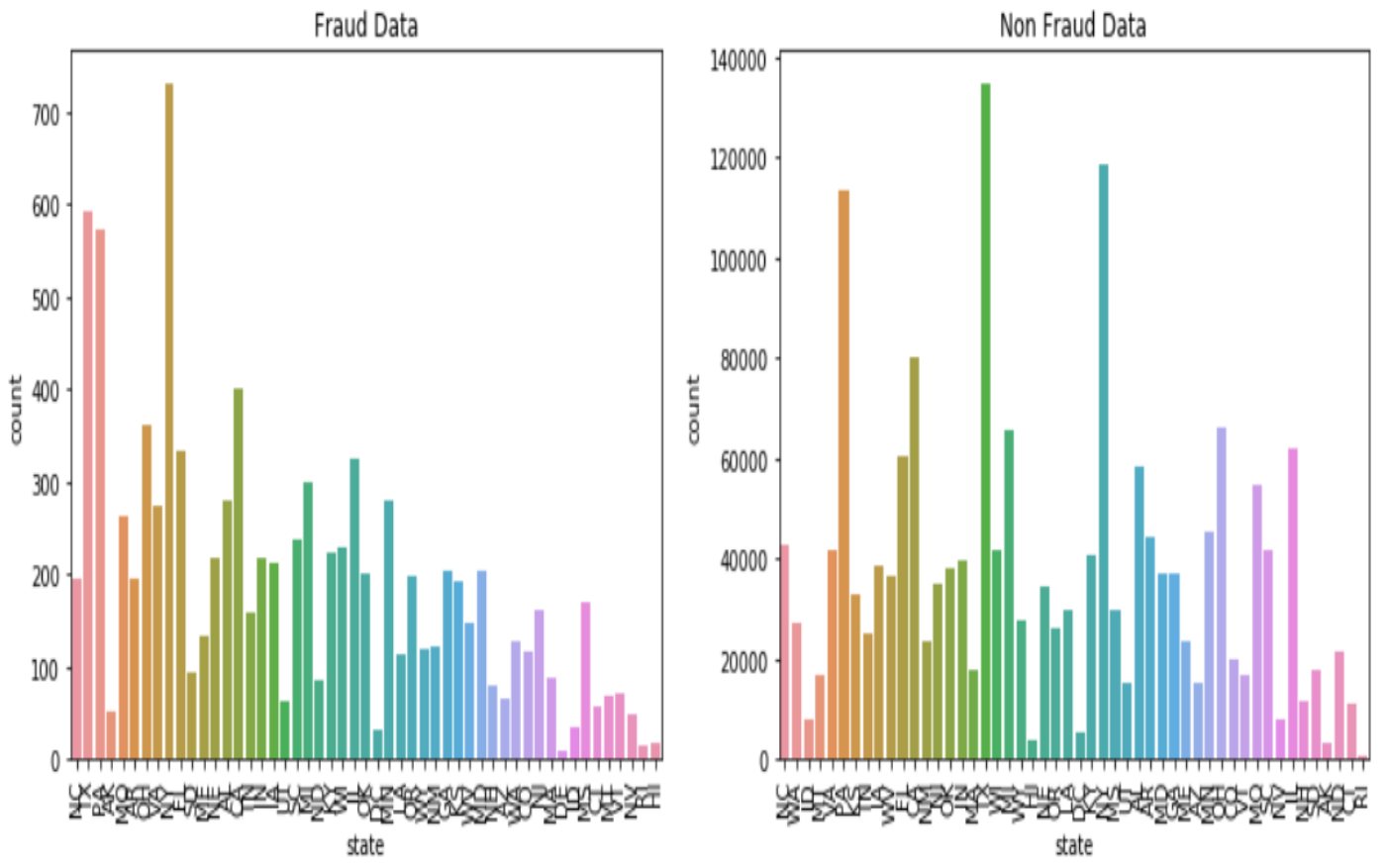
- Count of Fraud transactions are done more at grocery_pos, shopping_net, misc_net, shopping_pos, gas_transport Categories
- Count of Fraud transactions are more in shopping_net as compare to count of normal transaction are less.
- Count of Fraud transactions are slightly less in gas_transport, shopping_pos whereas count of normal transaction are more.

Amount Spent in different categories for Fraud Transactions data

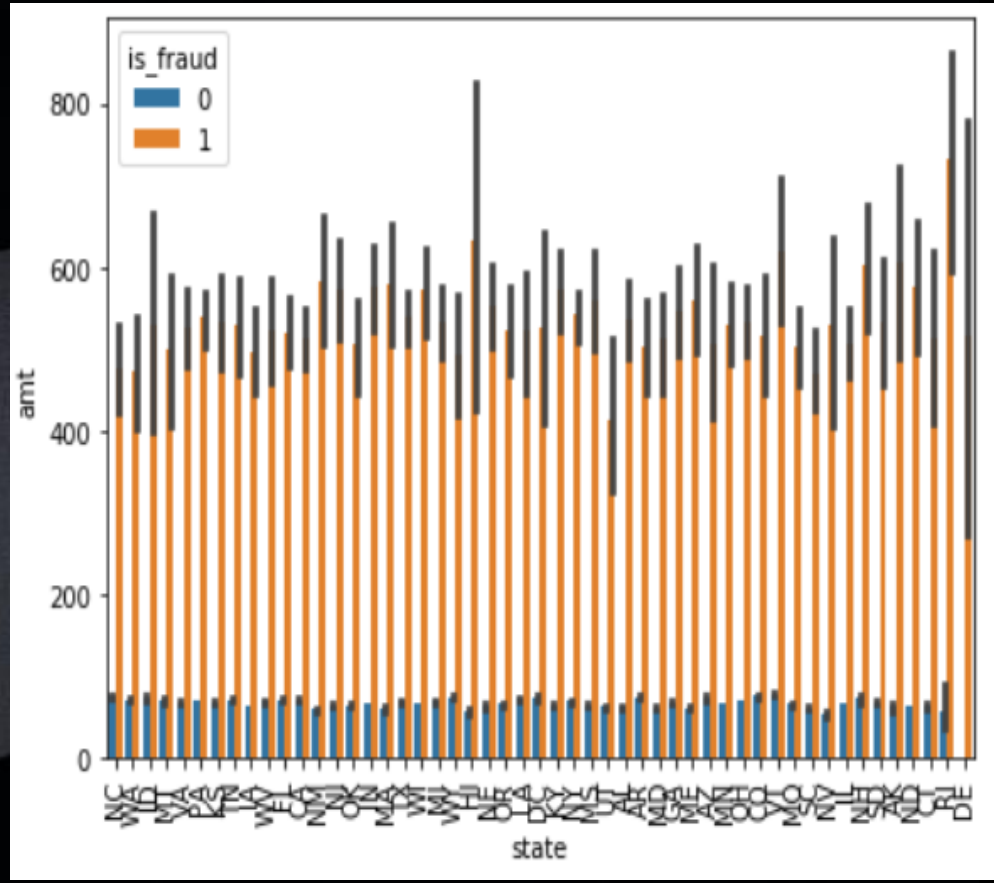


- The maximum amount spend were on shopping_net, shopping_pos, misc_net category and entertainment for fruad transactions.

Fraud Transactions in different State



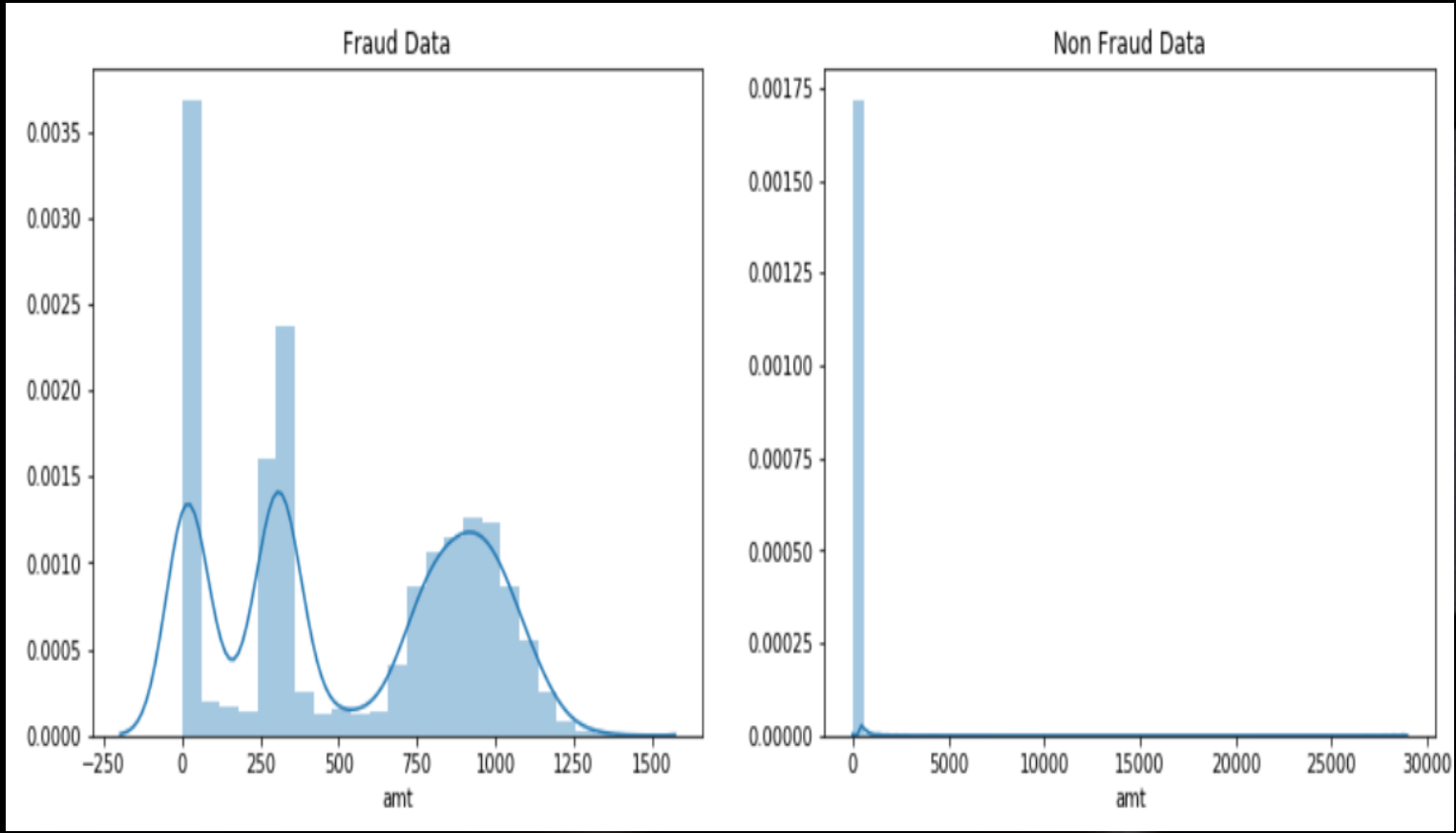
Amount Spent in different state for Fraud Transactions data



- Count of Frauds transactions are more in NY, TX and PA States.

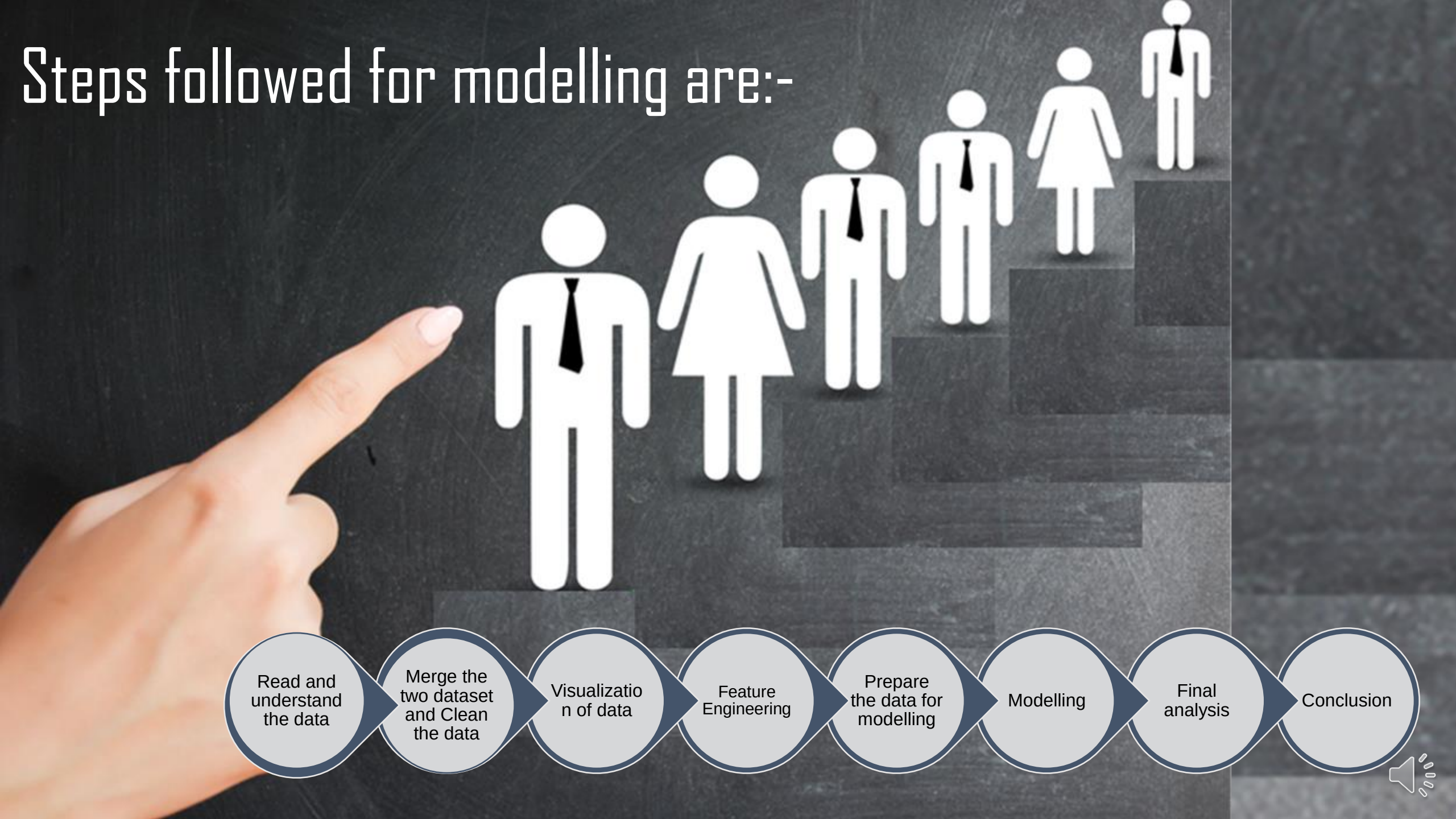
- The maximum amount spend for fraud transactions were at RI, HI, DE and VT State.

Distribution of Fraud amount and non Fraud amount.



Distribution of amount for fraud transaction is more widely spread as compare to non fraud transaction and major frauds happen at lower amounts.

Steps followed for modelling are:-



Model Consideration:

- Based on the accuracy, ROC, precision and recall of different models, we will consider XGBOOST (Hyperparameter Tuning) for SMOTE data as our final model.
- The test accuracy is 99.9%, recall is 89.6% and ROC is 99.8% .
- The recall for fraudulent transaction is 89.6%, which is highest among all other models. Since our buisness objective is more important to identify fraudulent transaction than the non-fraudulent transaction accurately. High recall means model will correctly identify almost all fraudulent transaction.
- Hence XGBOOST (Hyperparameter Tuning) model for SMOTE data is chosen based on its performance on Recall metric.

Compilation of models For Test data

1. Classification Report for Target (1) from Decision Tree (default Hyperparameter) on Test data

	precision(%) Positive predictive value(%)	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative predictive value(%)
SMOTE data =>	54.8	84.7	66.5	99.6	92.2	99.6	0.4	99.9
ADASYN data =>	56.3	84.1	67.4	99.6	91.9	99.7	0.3	99.9

Classification Report for Target (1) from Decision Tree (Hyperparameter Tunning) on SMOTE Test data

	precision(%) Positive predictive value(%)	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative predictive value(%)
SMOTE data	41.4	88.8	56.5	99.3	95.5	99.3	0.7	99.9

---

2. Classification Report for Target (1) from Random forest (Default Hyperparameters) on Test data

	precision(%) Positive predictive value(%)	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative predictive value(%)
SMOTE data	89.1	80.0	84.3	99.8	99.6	99.9	0.1	99.9
ADASYN data	90.6	77.6	83.6	99.8	99.5	100	0.0	99.9

Classification Report for Target (1) from Random Forest(Hyperparameter Tunning) on SMOTE Test data

	precision(%) Positive predictive value(%)	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative predictive value(%)
SMOTE data	65.0	85.9	74.0	99.7	99.5	99.8	0.2	99.9

---

3. Classification Report for Target (1) from XGBOOST (Default Hyperparameters) on Test data

	precision(%) Positive predictive value(%)	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative predictive value(%)
SMOTE data	85.2	89.6	87.4	99.9	99.8	99.8	0.1	99.9
ADASYN data	84.4	89.1	86.7	99.9	99.7	99.7	0.1	99.9

Classification Report for Target (1) from XGBOOST (Hyperparameter Tunning) on SMOTE Test data

	precision(%) Positive predictive value(%)	recall(%) (sensitivity)	f1-score(%)	Accuracy(%)	ROC(%)	Specificity(%)	False postive rate (%)	Negative predictive value(%)
SMOTE data	85.2	89.6	87.4	99.9	99.8	99.9	0.1	99.9

BUSINESS RECOMMENDATION

STRATEGIES RECOMMENDED TO BANK FOR ADOPTING OPTIMAL WAYS TO MITIGATE FRAUD RISKS BY DETECTING THE FRAUDULENT TRANSACTION BASED ON OUR MODEL OBSERVATIONS as follows:-

1. The fraudulent transaction probability of a transaction increases with increase in `hist_trans_avg_amt_24h` values. Based upon past spending pattern we have derived`hist_trans_avg_amt_24h` which is actually average amount spent through transactions in last 24 hours by the credit card holder's. So if comparable amount spent in last 24hrs v/s past spent data gets increased then its ideal for Bank to sent an SMS ALERT! to customer confirming about the transactions.

2. The fraudulent transaction probability of a transaction increases with increase in `weekday_Thursday`, `weekday_Saturday` and `weekday_Monday` values. As per the pattern model shows that major fraud transactions are noticed in `weekday_Thursday`, `weekday_Saturday` and `weekday_Monday`. So banks need to be extra cautious and high alert on this specific days to avoid fradulent transactions

3. The fraudulent transaction probability of a customer increases with increase in `amt` values. At any point in time if bank notices the nature of amount spent is higher then regular spending pattern in such cases bank should noticed the same at early stage by sending neccessary alerts to customers.

BUSINESS RECOMMENDATION

STRATEGIES RECOMMENDED TO BANK FOR ADOPTING OPTIMAL WAYS TO MITIGATE FRAUD RISKS BY DETECTING THE FRAUDULENT TRANSACTION BASED ON OUR MODEL OBSERVATIONS as follows:-

4. The fraudulent transaction probability of a customer increases with increase in `catg_home`, `catg_shopping_pos`, `catg_grocery_pos`, `catg_health_fitness`, `catg_gas_transport` values. Model predicted that major fraud transactions are occurred in the catg_home, catg_shopping_pos, catg_grocery_pos, catg_health_fitness`, `catg_gas_transports as these are the platform where any customer would spend large transactional amount so as fraudsters also follows the same trend. In such case its always recommended to bank to keep an eye on the track record of spend amount through FLASH SMS ALERT mentioning the detailed transaction history to respective credit card holders.

5. The fraud transactions are majorly done during odd hours of the day i.e. between 22 - 3 Hr so banks needs to ensure to send an SMS ALERT during such odd hours.

Cost Benefit Analysis

Part 1(For Whole data Before Modelling)

Cost Benefit Analysis

Questions	Answer
1. Average number of transactions per month	77183.0833333333
2. Average number of fraudulent transaction per month	402.125
3. Average amount per fraud transaction	530.6614122888819

Part 2 (For Whole data After Modelling)

Cost Benefit Analysis(Part 2)

Questions	Answer
1. Cost incurred per month before the model was deployed (b*c)	213392.2204
2. Average number of transactions per month detected as fraudulent by the model (TF)	405.125
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(TF*$1.5)	607.69
5. Average number of transactions per month that are fraudulent but not detected by the model (FN)	16.25
6. Cost incurred due to fraudulent transactions left undetected by the model (FN*c)	8623.24795
7. Cost incurred per month after the model is built and deployed (4+6)	9230.94
8. Final savings = Cost incurred before - Cost incurred after(1-7)	204161.2804

- The costs incurred before the model is deployed was $213392.
- So bank had to pay the entire transaction amount to the customer for every fraudulent transaction which accounted for a heavy loss to the bank.

- After the model is deployed **final savings were amounted to $ 204161**.

CONCLUSION

- The concept behind using machine learning for fraud detection is that fraudulent transactions have specific features that legitimate transactions.

- Machine learning helps these institutions-
  - ✓ To reduce time-consuming manual reviews.
  - ✓ Costly chargebacks and fees.
  - ✓ Denial of legitimate transactions.