IST 707 Final Project

Credit card fraud detection

Meichan Huang

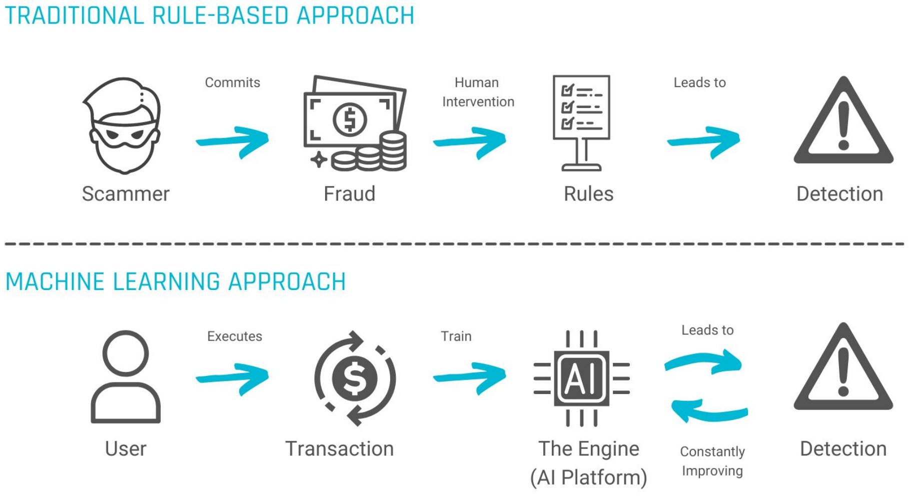
Credit Card Frauds

Credit card fraud is a significant issue that affects consumers, businesses, and financial institutions globally.

It involves unauthorized and illegal use of a credit card to make purchases or withdraw funds.

The implications of credit card fraud are far-reaching and can have various impacts:

- Financial loss
- Identity theft
- Impact on credit scores
- Business damage



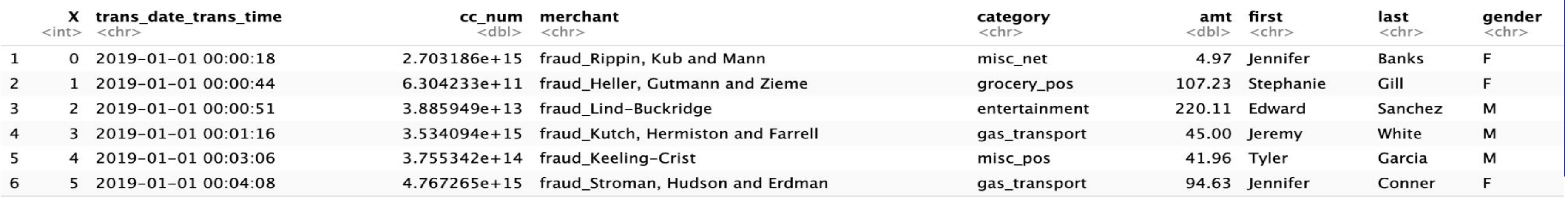
The dataset

This dataset comprises **simulated credit card transactions legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants.**

Data source: https://www.kaggle.com/datasets/kartik2112/fraud-detection

The primary objective of this dataset is to facilitate the development of fraud detection algorithms and models to identify potentially fraudulent transactions.

The dataset contains 555,719 rows of observations, which has 23 columns of variables, 12 of which are qualitative data:



	X	trans_date_trans_time	cc_num	merchant	category	amt	first	last	gender
	<int>	<chr>	<dbl>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>
1	0	2019-01-01 00:00:18	2.703186e+15	fraud_Rippin, Kub and Mann	misc_net	4.97	Jennifer	Banks	F
2	1	2019-01-01 00:00:44	6.304233e+11	fraud_Heller, Gutmann and Zieme	grocery_pos	107.23	Stephanie	Gill	F
3	2	2019-01-01 00:00:51	3.885949e+13	fraud_Lind-Buckridge	entertainment	220.11	Edward	Sanchez	M
4	3	2019-01-01 00:01:16	3.534094e+15	fraud_Kutch, Hermiston and Farrell	gas_transport	45.00	Jeremy	White	M
5	4	2019-01-01 00:03:06	3.755342e+14	fraud_Keeling-Crist	misc_pos	41.96	Tyler	Garcia	M
6	5	2019-01-01 00:04:08	4.767265e+15	fraud_Stroman, Hudson and Erdman	gas_transport	94.63	Jennifer	Conner	F

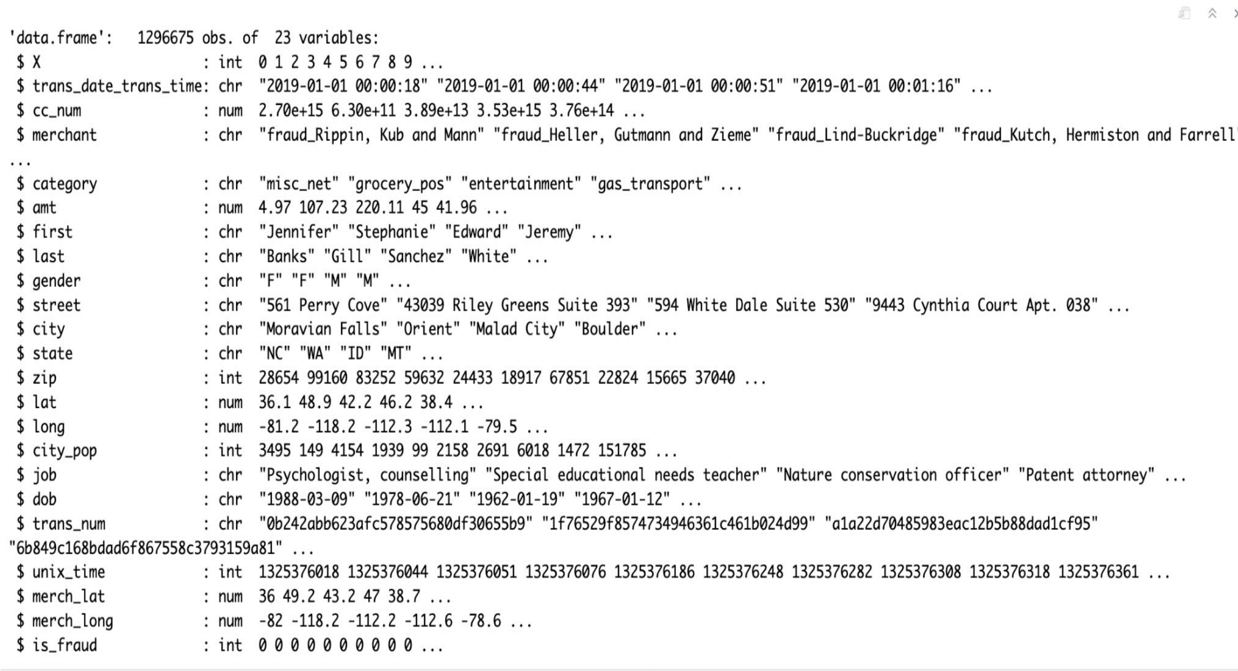
Business questions:

1. Which factors are most strongly associated with fraudulent transactions?
   a. In which transaction categories does fraud occur most frequently?
   b. Are certain customer demographics (like age or gender) more susceptible to fraud?
   c. Which geographic locations have higher instances of fraud?
   d. During what times (months, days, and specific hours) do fraudulent transactions occur most often?
2. Which machine learning model is the most effective in identifying and preventing credit card fraud through the analysis of transaction features? Is it Logistic Regression, Random Forest, or Support Vector Machine (SVM)?

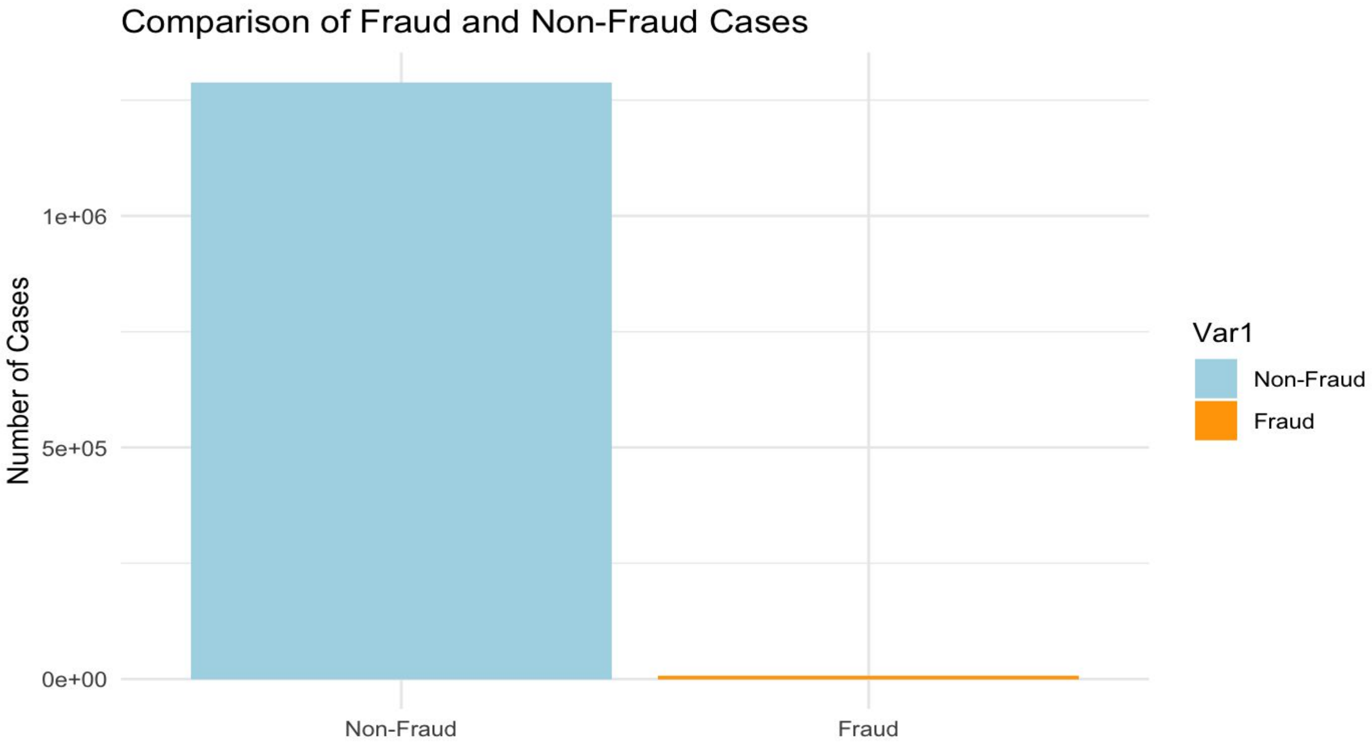
EDA analysis - structure of the train data

The structure showed that:

-
- Data transformation needed:
    - Transform categorical variables to numeric
- Remove redundant columns



```
'data.frame':	1296675 obs. of  23 variables:
 $ X                    : int  0 1 2 3 4 5 6 7 8 9 ...
 $ trans_date_trans_time: chr  "2019-01-01 00:00:18" "2019-01-01 00:00:44" "2019-01-01 00:00:51" "2019-01-01 00:01:16" ...
 $ cc_num               : num  2.70e+15 6.30e+11 3.89e+13 3.53e+15 3.76e+14 ...
 $ merchant             : chr  "fraud_Rippin, Kub and Mann" "fraud_Heller, Gutmann and Zieme" "fraud_Lind-Buckridge" "fraud_Kutch, Hermiston and Farrell
...
 $ category             : chr  "misc_net" "grocery_pos" "entertainment" "gas_transport" ...
 $ amt                  : num  4.97 107.23 220.11 45 41.96 ...
 $ first                : chr  "Jennifer" "Stephanie" "Edward" "Jeremy" ...
 $ last                 : chr  "Banks" "Gill" "Sanchez" "White" ...
 $ gender               : chr  "F" "F" "M" "M" ...
 $ street               : chr  "561 Perry Cove" "43039 Riley Greens Suite 393" "594 White Dale Suite 530" "9443 Cynthia Court Apt. 038" ...
 $ city                 : chr  "Moravian Falls" "Orient" "Malad City" "Boulder" ...
 $ state                : chr  "NC" "WA" "ID" "MT" ...
 $ zip                  : int  28654 99160 83252 59632 24433 18917 67851 22824 15665 37040 ...
 $ lat                  : num  36.1 48.9 42.2 46.2 38.4 ...
 $ long                 : num  -81.2 -118.2 -112.3 -112.1 -79.5 ...
 $ city_pop             : int  3495 149 4154 1939 99 2158 2691 6018 1472 151785 ...
 $ job                  : chr  "Psychologist, counselling" "Special educational needs teacher" "Nature conservation officer" "Patent attorney" ...
 $ dob                  : chr  "1988-03-09" "1978-06-21" "1962-01-19" "1967-01-12" ...
 $ trans_num            : chr  "0b242abb623afc578575680df30655b9" "1f76529f8574734946361c461b024d99" "a1a22d70485983eac12b5b88dad1cf95"
"6b849c168bdad6f867558c3793159a81" ...
 $ unix_time            : int  1325376018 1325376044 1325376051 1325376076 1325376186 1325376248 1325376282 1325376308 1325376318 1325376361 ...
 $ merch_lat            : num  36 49.2 43.2 47 38.7 ...
 $ merch_long           : num  -82 -118.2 -112.2 -112.6 -78.6 ...
 $ is_fraud             : int  0 0 0 0 0 0 0 0 0 0 ...
```

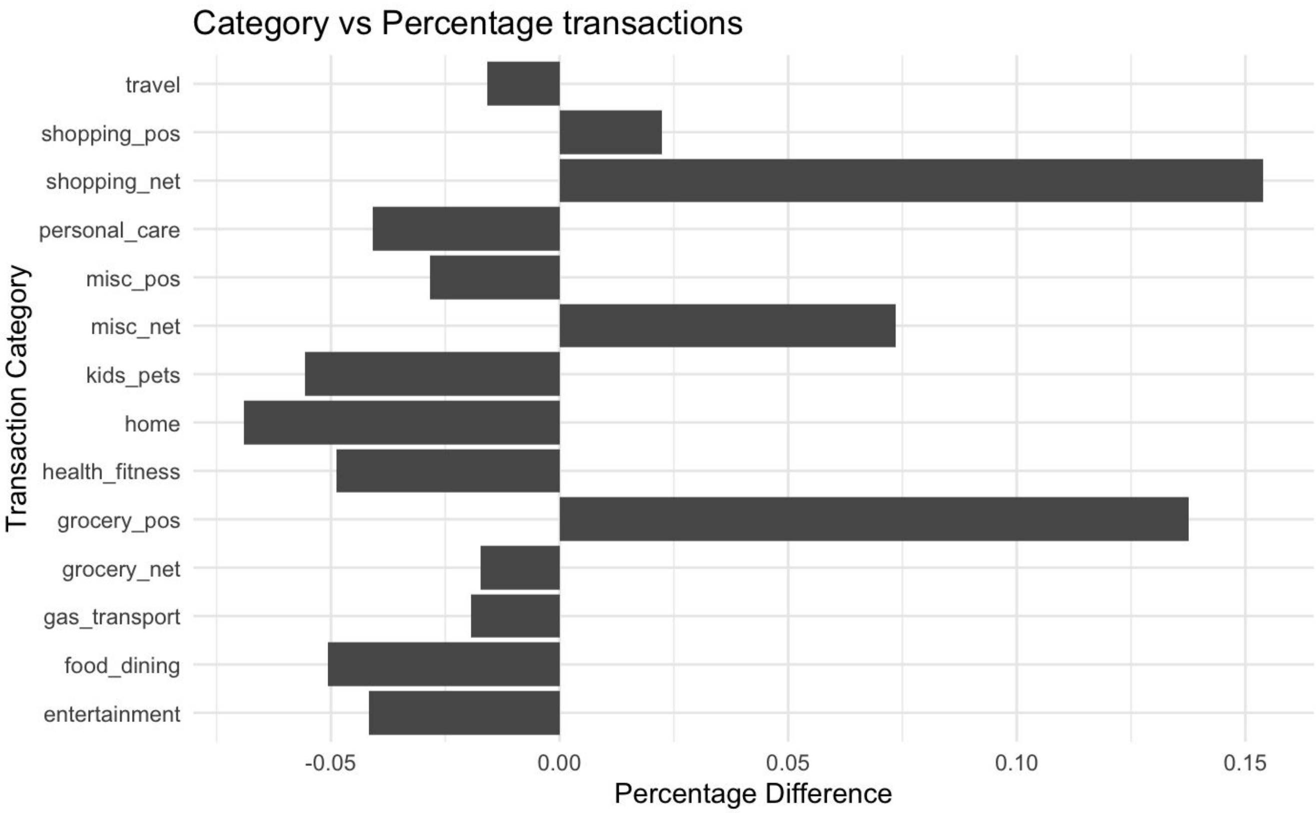


Imbalance dataset:
How to process?

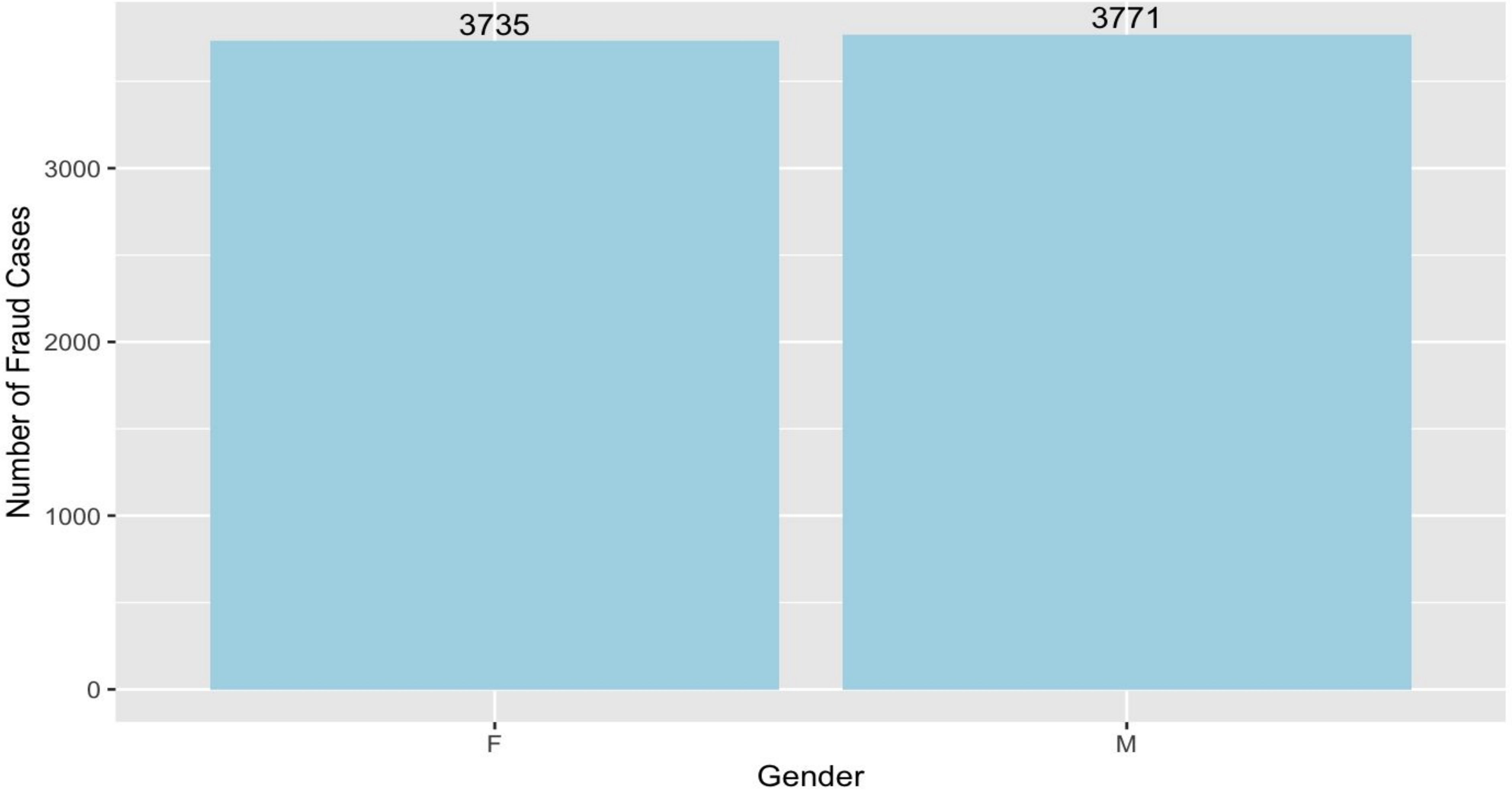
Transaction by category

The results showed, four categories of transactions are fraud-prone:

- shopping_pos,
- shopping_net,
- misc_net, and
- grocery_pos

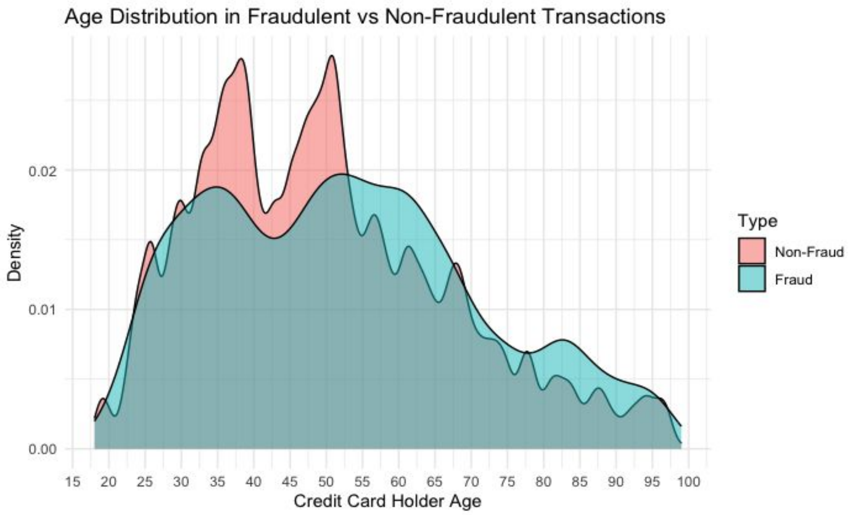


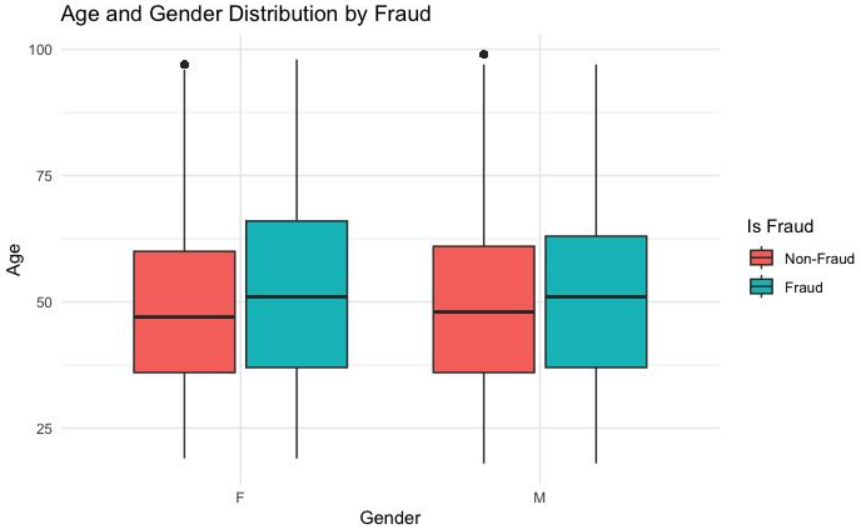
Fraud Cases by Gender

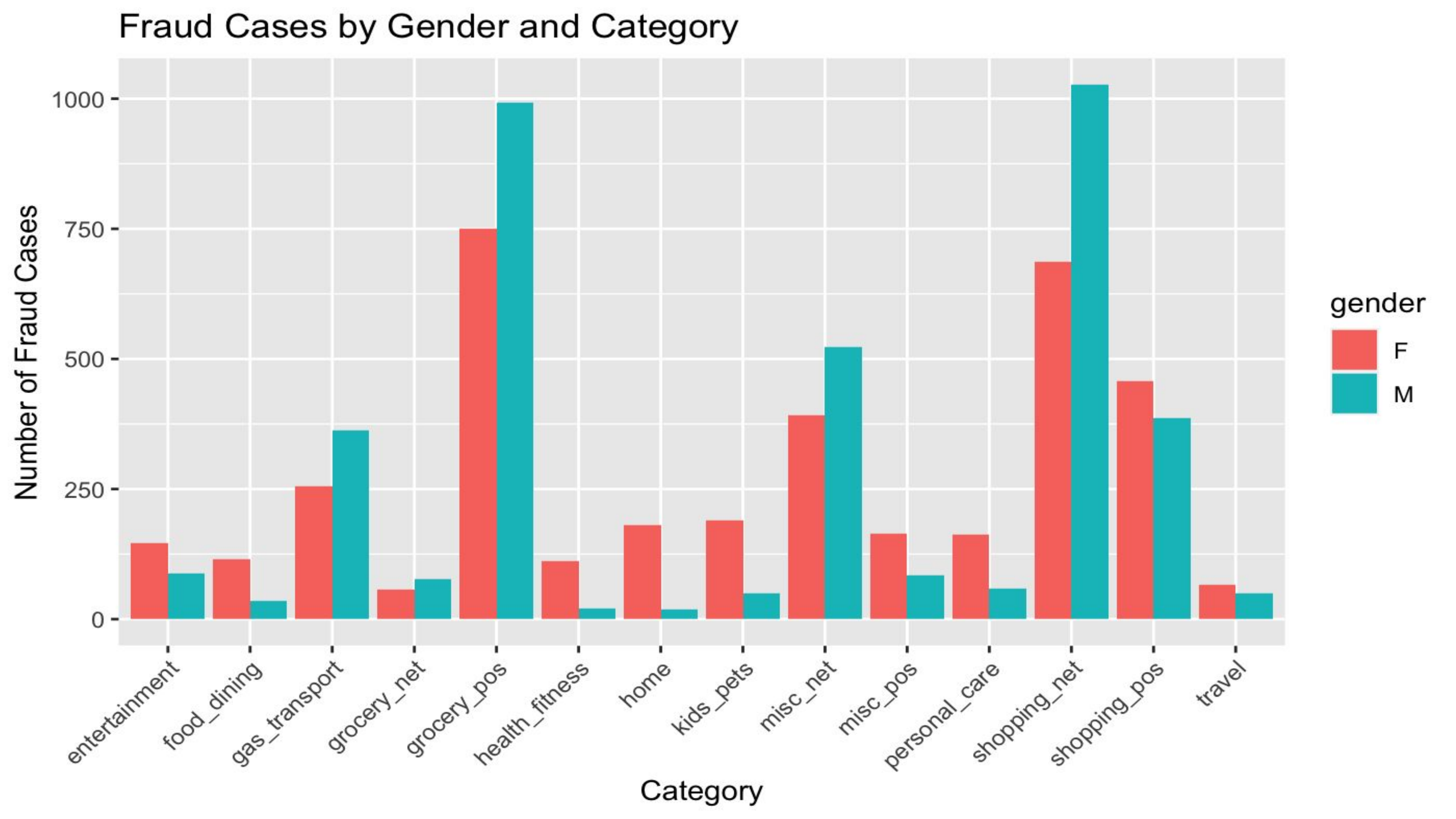


Age and Fraud

- In transactions that are not fraudulent, there are notable peaks at ages 35 and 40, followed by another peak at around 50 years.
- For fraudulent transactions, the age distribution is more even, with noticeable peaks at age 35 and a range from 50 to 55 years.
- Gender differences don't significantly impact the occurrence of fraud. However, in fraudulent cases, females show a broader range of ages compared to males.



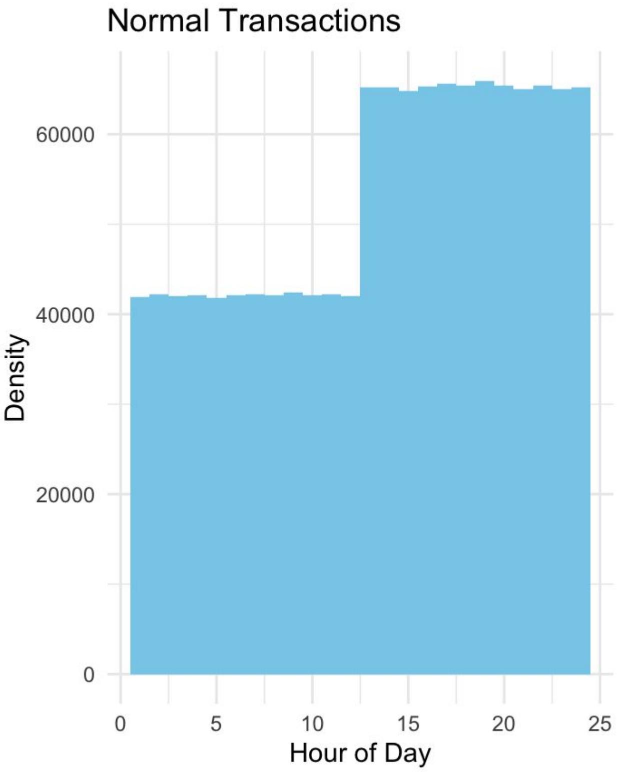



Fraud Cases by Gender and Category
Number of Fraud Cases
1000
750
500
250
0
entertainment
food_dining
gas_transport
grocery_net
grocery_pos
health_fitness
home
kids_pets
misc_net
misc_pos
personal_care
shopping_net
shopping_pos
travel
Category
gender
F
M

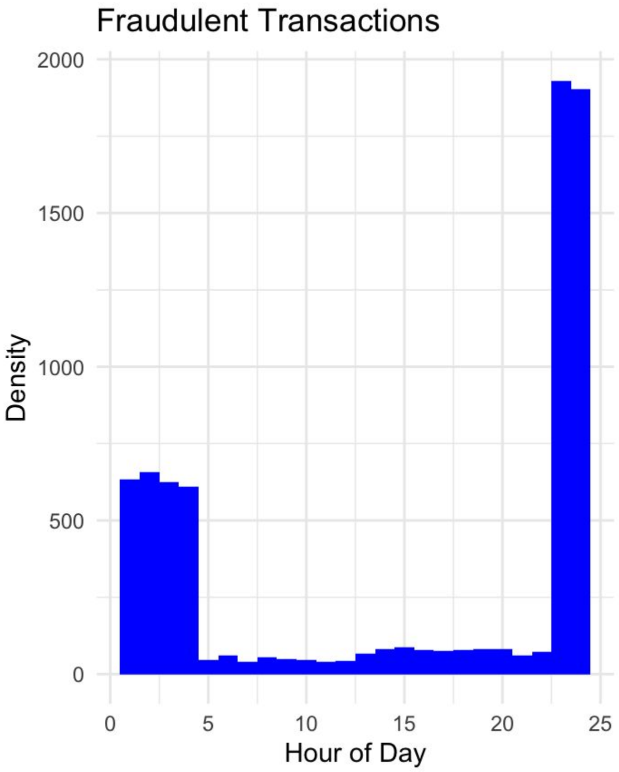
Transaction time and fraud

The hours of the day that are likely to have fraud are:

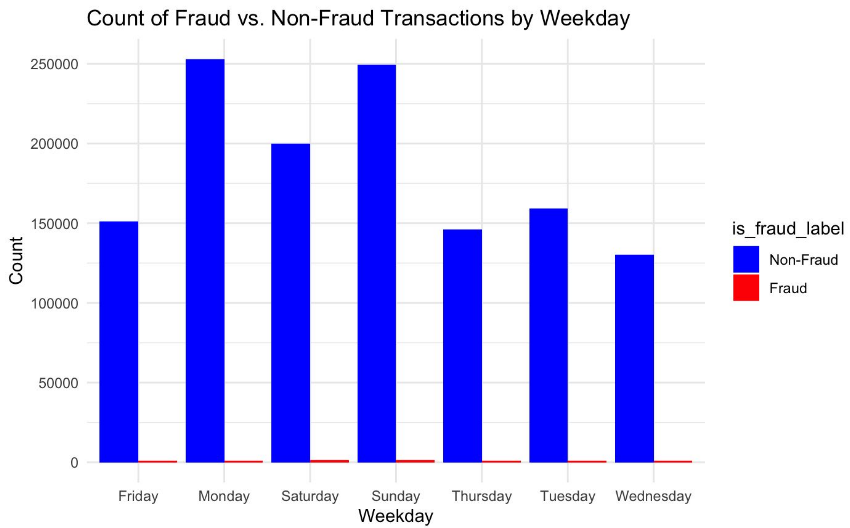
(1) Midnight: 0-5 a.m.
(2) Late night 22-24 p.m.

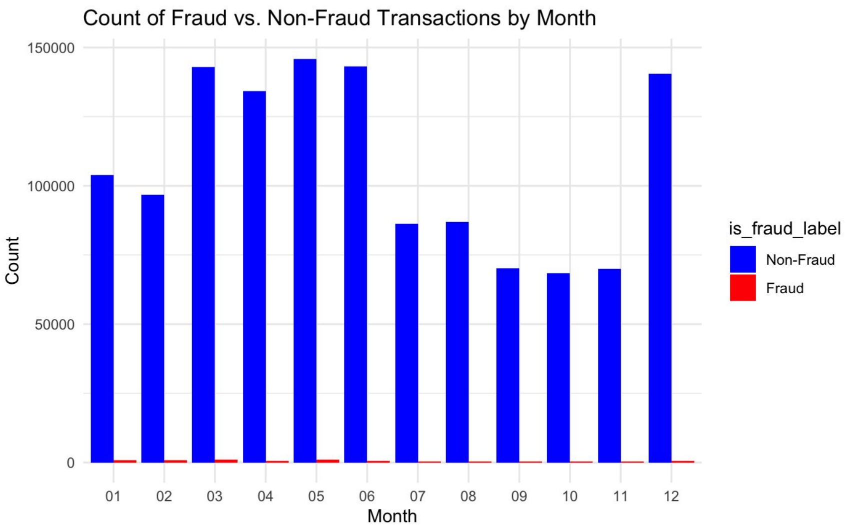


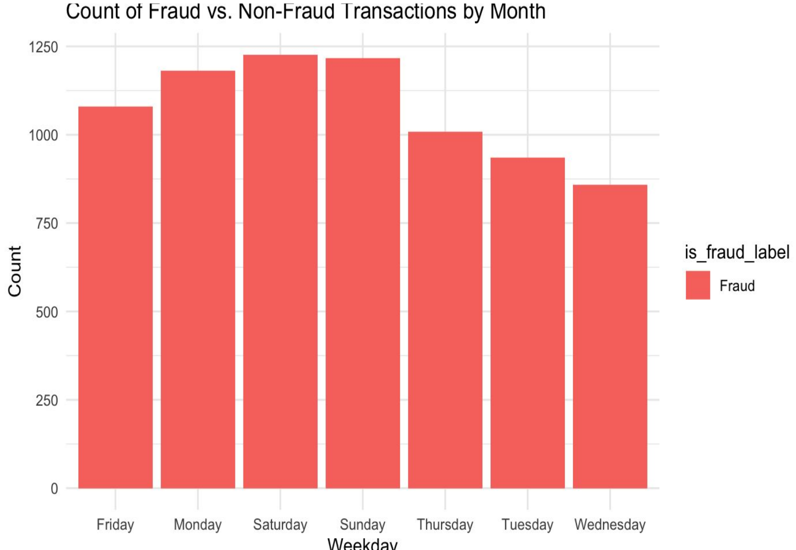
Normal Transactions

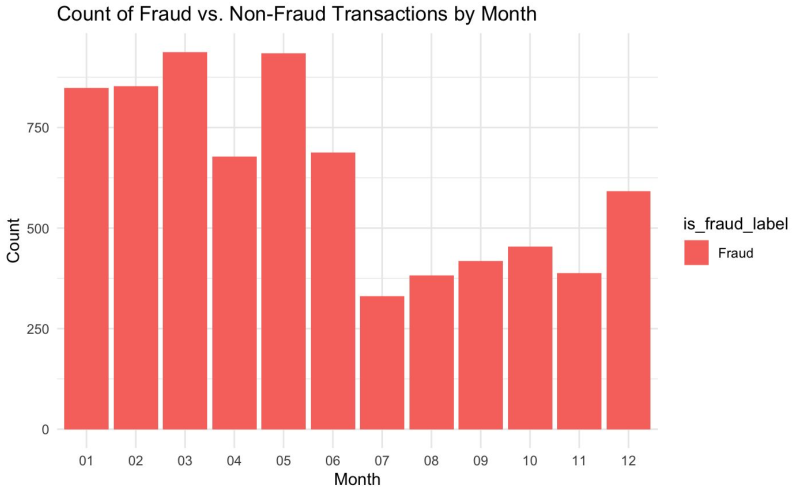


Fraudulent Transactions

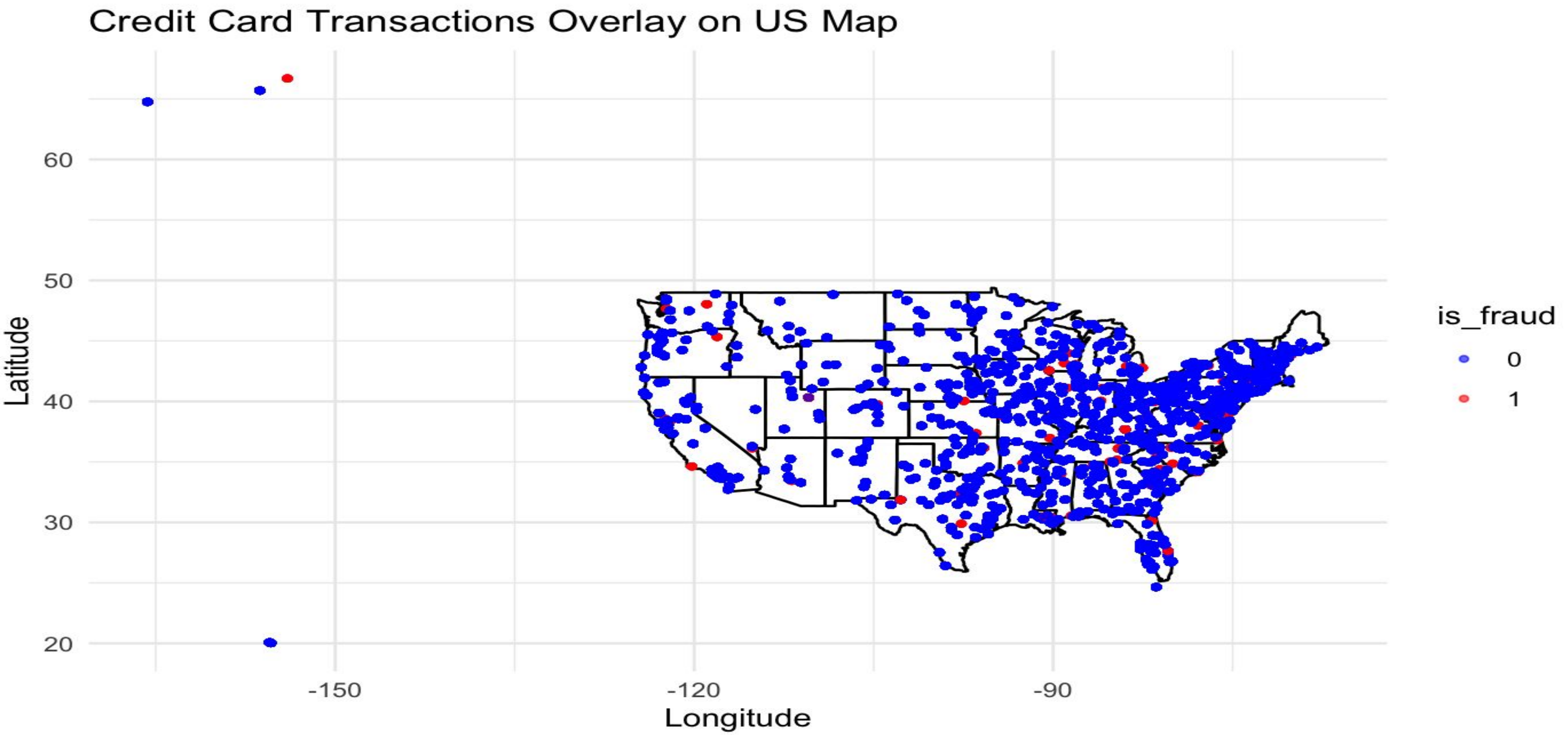

Count of Fraud vs. Non-Fraud Transactions by Weekday


Count of Fraud vs. Non-Fraud Transactions by Month

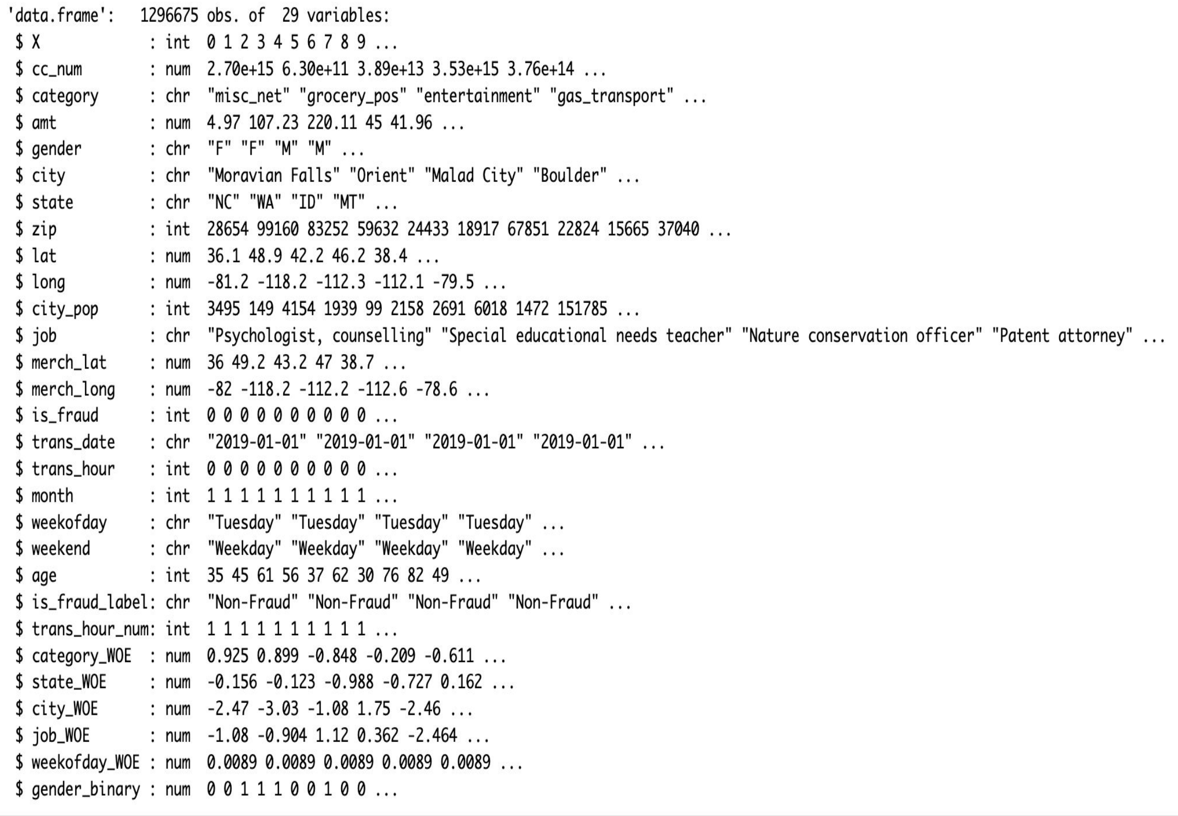

Count of Fraud vs. Non-Fraud Transactions by Month


Count of Fraud vs. Non-Fraud Transactions by Month

Fraud vs. Non-fraud transactions in US



Data Transformation: categorical variables



```
'data.frame':	1296675 obs. of  29 variables:
 $ X             : int  0 1 2 3 4 5 6 7 8 9 ...
 $ cc_num        : num  2.70e+15 6.30e+11 3.89e+13 3.53e+15 3.76e+14 ...
 $ category      : chr  "misc_net" "grocery_pos" "entertainment" "gas_transport" ...
 $ amt           : num  4.97 107.23 220.11 45 41.96 ...
 $ gender        : chr  "F" "F" "M" "M" ...
 $ city          : chr  "Moravian Falls" "Orient" "Malad City" "Boulder" ...
 $ state         : chr  "NC" "WA" "ID" "MT" ...
 $ zip           : int  28654 99160 83252 59632 24433 18917 67851 22824 15665 37040 ...
 $ lat           : num  36.1 48.9 42.2 46.2 38.4 ...
 $ long          : num  -81.2 -118.2 -112.3 -112.1 -79.5 ...
 $ city_pop      : int  3495 149 4154 1939 99 2158 2691 6018 1472 151785 ...
 $ job           : chr  "Psychologist, counselling" "Special educational needs teacher" "Nature conservation officer" "Patent attorney" ...
 $ merch_lat     : num  36 49.2 43.2 47 38.7 ...
 $ merch_long    : num  -82 -118.2 -112.2 -112.6 -78.6 ...
 $ is_fraud      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ trans_date    : chr  "2019-01-01" "2019-01-01" "2019-01-01" "2019-01-01" ...
 $ trans_hour    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ month         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ weekofday     : chr  "Tuesday" "Tuesday" "Tuesday" "Tuesday" ...
 $ weekend       : chr  "Weekday" "Weekday" "Weekday" "Weekday" ...
 $ age           : int  35 45 61 56 37 62 30 76 82 49 ...
 $ is_fraud_label: chr  "Non-Fraud" "Non-Fraud" "Non-Fraud" "Non-Fraud" ...
 $ trans_hour_num: int  1 1 1 1 1 1 1 1 1 1 ...
 $ category_WOE  : num  0.925 0.899 -0.848 -0.209 -0.611 ...
 $ state_WOE     : num  -0.156 -0.123 -0.988 -0.727 0.162 ...
 $ city_WOE      : num  -2.47 -3.03 -1.08 1.75 -2.46 ...
 $ job_WOE       : num  -1.08 -0.904 1.12 0.362 -2.464 ...
 $ weekofday_WOE : num  0.0089 0.0089 0.0089 0.0089 0.0089 ...
 $ gender_binary : num  0 0 1 1 1 0 0 1 0 0 ...
```

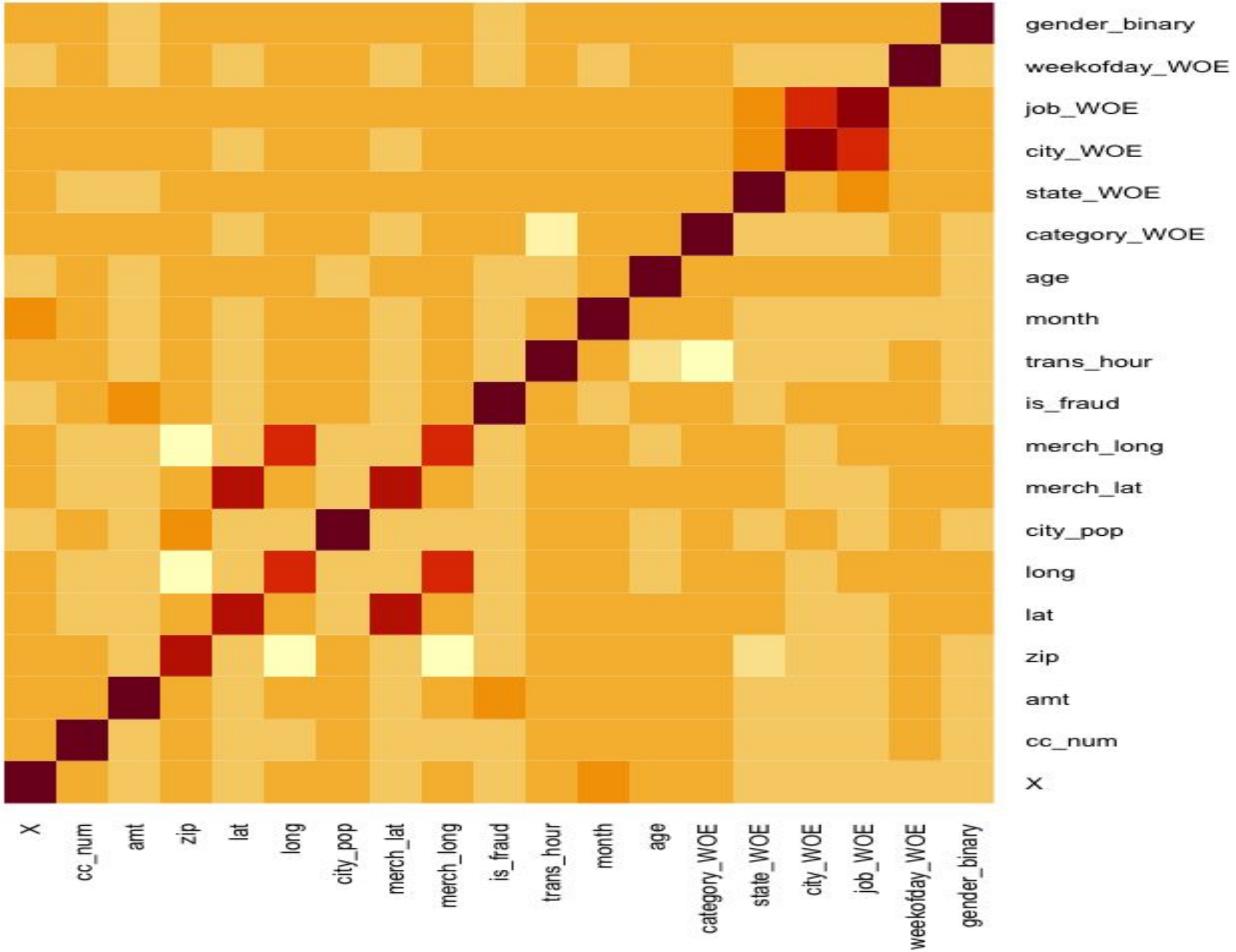
Two types of categorical data transformation are used here:

(1) Dummy coding for binary variable: gender

(2) WoE(Weight of Evidence) for highly cardinal variables such as city, state, month, weekday, job

$$WoE = \left[\ln \left(\frac{Distr\ Goods}{Distr\ Bads} \right) \right] * 100$$

Correlation Matrix Heatmap

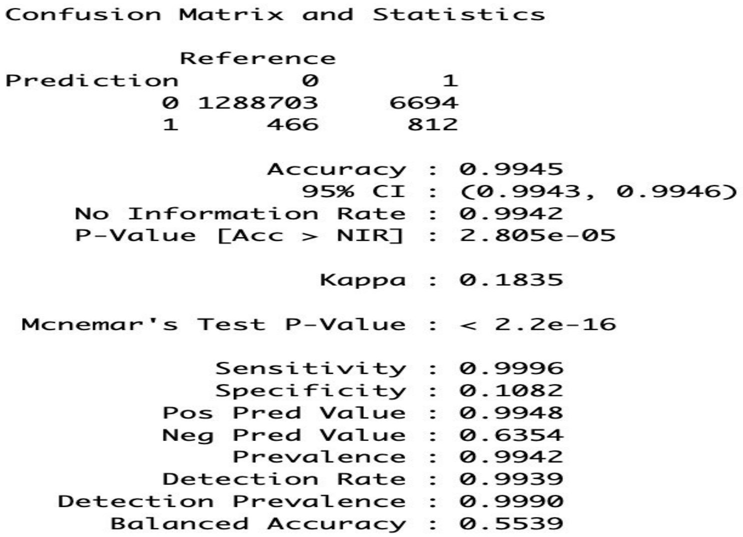


Model 1: Logistic Regression

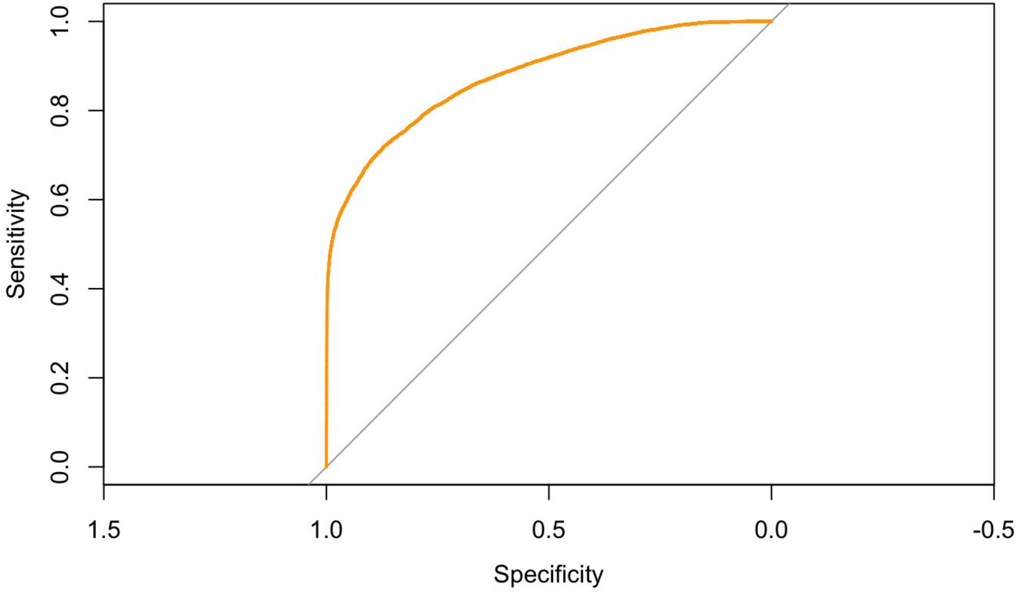
model <- glm(is_fraud ~ ., data = train1, family = binomial())

Performance matrix:

AUC: 0.8778



```
Confusion Matrix and Statistics

          Reference
Prediction       0       1
         0 1288703    6694
         1     466     812

               Accuracy : 0.9945
                 95% CI : (0.9943, 0.9946)
    No Information Rate : 0.9942
    P-Value [Acc > NIR] : 2.805e-05

                  Kappa : 0.1835

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9996
            Specificity : 0.1082
         Pos Pred Value : 0.9948
         Neg Pred Value : 0.6354
             Prevalence : 0.9942
         Detection Rate : 0.9939
   Detection Prevalence : 0.9990
      Balanced Accuracy : 0.5539
```



```
Area under the curve: 0.8778
```

An AUC of 0.8778 suggests that the model has a high ability to differentiate between the two classes (e.g., fraud and non-fraud). It means that there's an 87.78% chance that the model will correctly distinguish between a positive and a negative instance when randomly picking one of each.

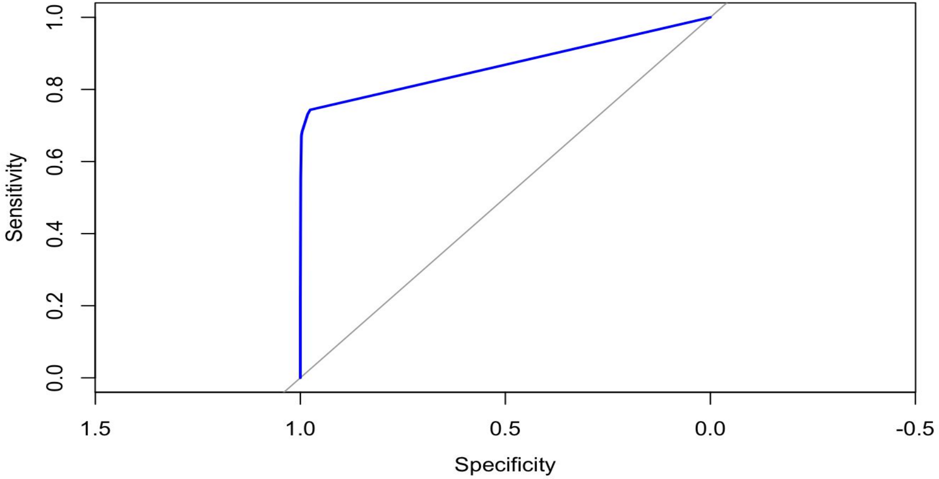
Model 2: Decision Tree

Model: Decision Tree with 3 CV folds

model2 <- train(is_fraud ~ amt + category_WOE + age, data = train1, method = "rpart", trControl = control, metric = "ROC")

Performance matrix: AUC: 0.867

The results were similar to that of the logistic regression, which still leaves room for improvement.



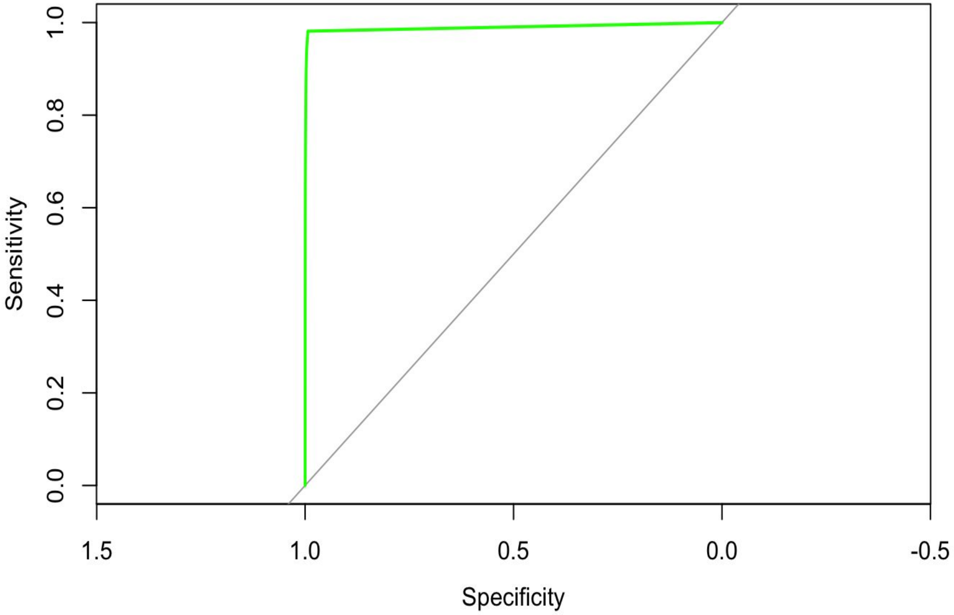
```
Area under the curve: 0.8674
```

Model 3: Random Fores

model_rf = randomForest(as.factor(is_fraud) ~ amt + category_WOE + age, data = train1, ntree=50)

Performance matrix: AUC: 0.98

Results: The best model so far in identifying fraud, with AUC of 0.98. However, it still needs work as I was not able to examine the precision and accuracy of the model yet…



```
Area under the curve: 0.9898
```

Conclusion and limitations

Conclusions:

(1) Which factors are most strongly associated with fraudulent transactions?

(a) - shopping_pos, shopping_net, misc_net, and grocery_pos are more prone to fraud.

(b) Older customers are more susceptible to fraud.

(c) States on the east, midwest, and south are more susceptible to frauds.

(d) Transactions happened at Midnights, on Mondays, Saturdays, and Sundays, and January, February, March, and May are more likely to be fraudulent transactions.

(2) With the current modeling, random forest seems to perform the best, reaching AUC = 0.98; Logistic Regression follows, with AUC = 0.88; Decision Tree performed the worst, with AUC = 0.87

Limitations:

(1) Data might need to be more balanced, because the AUC are still not ideal, which could indicate lower performance and false classification of fraud as non-fraud;

(2) Tuning of the model might be needed