Credit Card Transaction Fraud

May 2018 - Project 3 Report - Supervised Fraud Algorithm

Executive Summary

Data Description

File Description:

The dataset is called ‘card transactions’ which contains 96,708 records and 10 variables. It includes information about the record number, card number, the date the card transactions made, merchant number, merchant description, merchant state, merchant zip code, transaction type, transaction amount and fraud condition.

File Name: card transactions.xlsx

Data Source: This dataset is the simulated data based on real card transactions records.

Number of Records**:** 96,708 records

Number of Fields: 10 variables in total:

	Recordnum	Cardnum	Date	Merchantnum	Merch Description	Merchant State	Merchant Zip	Transtype	Amount	Fraud
count	96708.000000	9.670800e+04	96708	93333	96708	95513	92052.000000	96708	96707.000000	96708.000000
unique	NaN	NaN	365	13090	13125	227	NaN	4	NaN	NaN
top	NaN	NaN	2010-02-28 00:00:00	930090121224	GSA-FSS-ADV	TN	NaN	P	NaN	NaN
freq	NaN	NaN	684	9310	1688	11990	NaN	96353	NaN	NaN
first	NaN	NaN	2010-01-01 00:00:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
last	NaN	NaN	2010-12-31 00:00:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	48354.500000	5.142201e+09	NaN	NaN	NaN	NaN	44709.817603	NaN	395.792713	0.010485
std	27917.339254	5.391327e+04	NaN	NaN	NaN	NaN	28376.097348	NaN	832.047787	0.101859
min	1.000000	5.142110e+09	NaN	NaN	NaN	NaN	1.000000	NaN	0.010000	0.000000
25%	24177.750000	5.142152e+09	NaN	NaN	NaN	NaN	20855.000000	NaN	33.450000	0.000000
50%	48354.500000	5.142196e+09	NaN	NaN	NaN	NaN	38118.000000	NaN	137.900000	0.000000
75%	72531.250000	5.142246e+09	NaN	NaN	NaN	NaN	63103.000000	NaN	427.665000	0.000000
max	96708.000000	5.142311e+09	NaN	NaN	NaN	NaN	99999.000000	NaN	47900.000000	1.000000

Field Name	Data Type
Recordnum	Numerical variable
Cardnum	Numerical variable
Date	Date variable
Merchantnum	Numerical variable
Merch Description	Text variable
Merchant State	Categorical variable
Merchant Zip	Numerical variable
Transtype	Categorical variable
Amount	Numerical variable
Fraud	Categorical variable

Fields Explanation:

Field 1- Field Name: Recordnum

Description: “record” is a numerical variable. It is the ordinal reference number for each record.

Missing Values: None. 100% populated

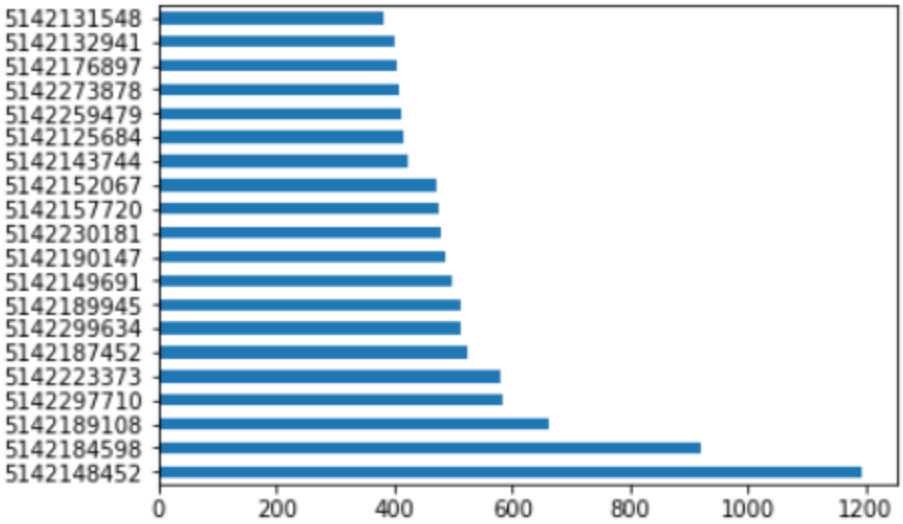
Unique Values: 96,708 unique values, ranging from 1 to 96,708. No repeated values or missing values exist.

Field 2- Field Name: Cardnumber

Description: “Cardnumber” is a numerical variable, indicating the card number that the transaction was made.

Missing Values: None. 100% populated.

Unique Values: 1644 unique values. Here is a representation of some of those unique values.

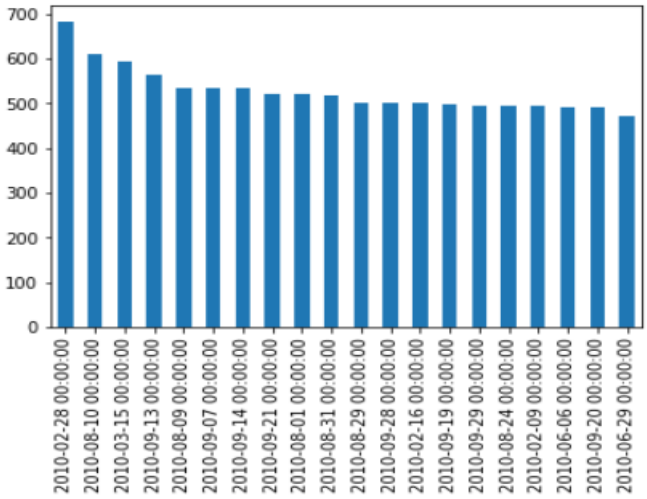


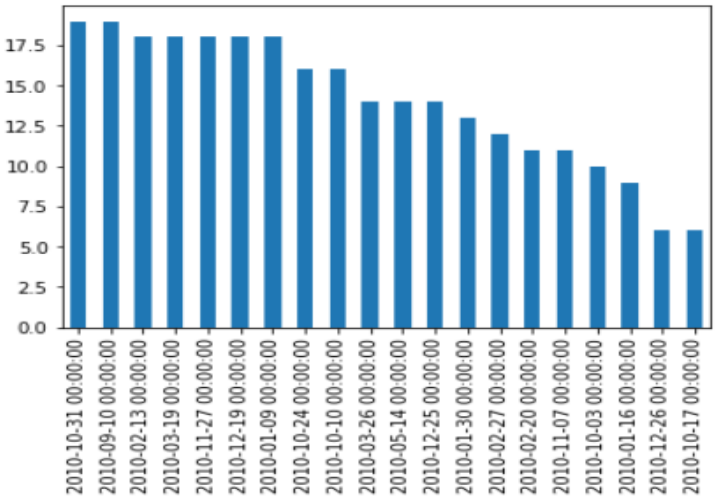
Field 3-Field Name: Date

Description: “Date” is a date variable, indicating the date of the transaction.

Missing Values: None. 100% populated.

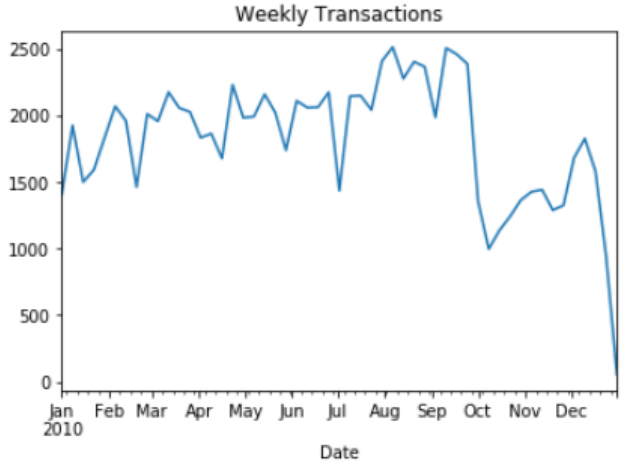
Unique Values: 365 unique values. The distribution bellow shows the top 20 frequent

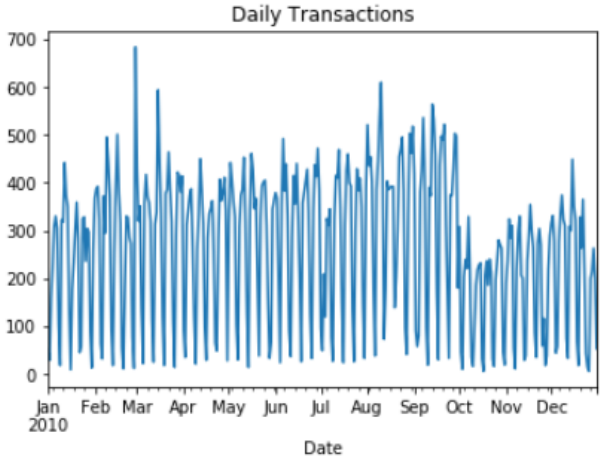


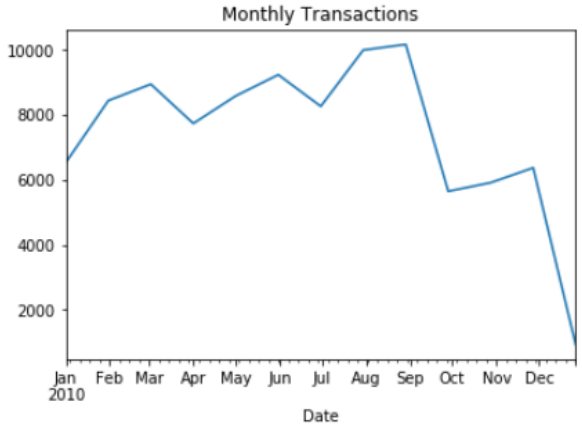


and Top 20 Dates bottom 20 which appeared in Bottom 20 Dates the data:

The following plots are daily, weekly and monthly transactions.





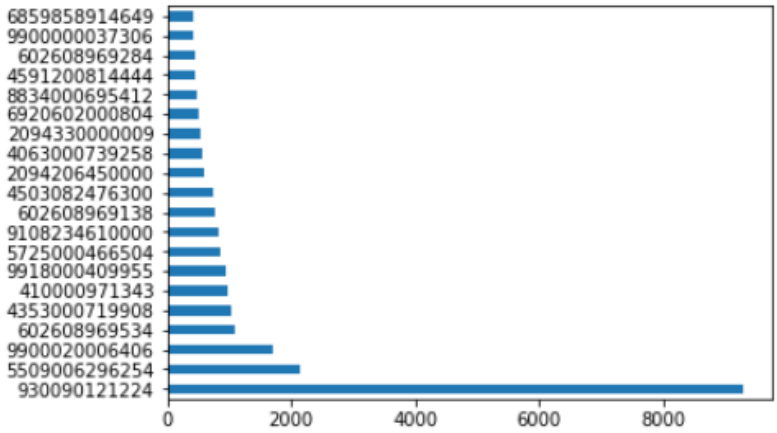


Field 4-Field Name: Merchantnum

Description: “Merchantnum” is a numerical variable, indicating the merchant number of the transactions.

Missing Values: Yes. This field has 3,375 missing values.

Unique Values: 96.5101% populated with 13,091 unique values. The top 20 frequent appeared merchant number records are shown below. Number 930090121224 in total appeared 9310 times, which is suspicious.

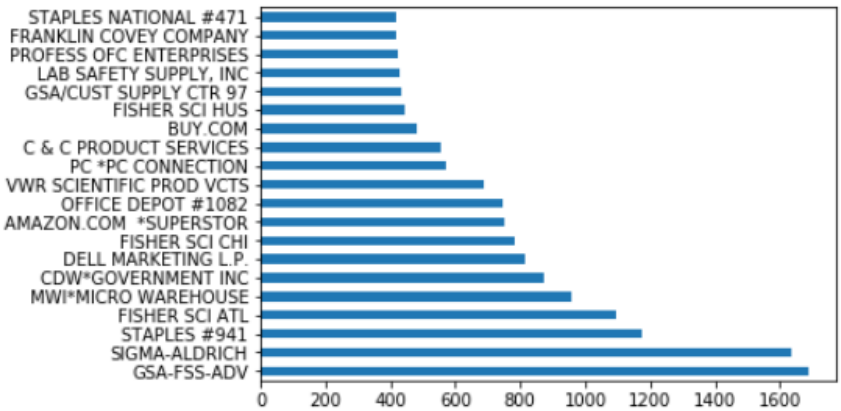


Field 5- Field Name: Merch Description

Description: “Merch Description” is a text variable, indicating the merchant information of transactions.

Missing Values: None. 100% populated.

Unique Values: 13,125 unique values. The top 20 frequent appeared merchant information are shown below.



* In our data exploration, we realized that some of the merchant description elements are numbers instead of text. Below we show a sample of these entries:

663	1
610	1
702	1
677	1
537	1
495	1
450	1
386	1
358	1
097	1
454	1
759	1

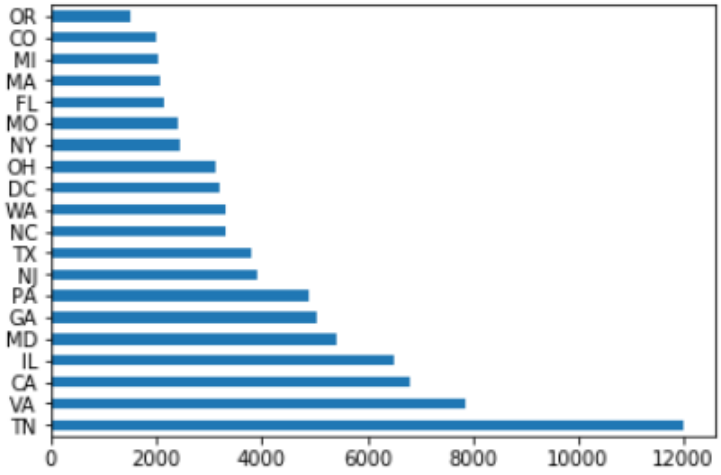
Field 6- Field Name: Merchant State

Description: “Merchant State” is a categorical variable that shows the state of transactions. This field has 1,195 missing values. What strange is that some records are with numbers, instead of state name.

Missing Values: Yes. 98.76% populated.

Unique Values: 228 unique values.

The top 20 most frequent records are distributed below:



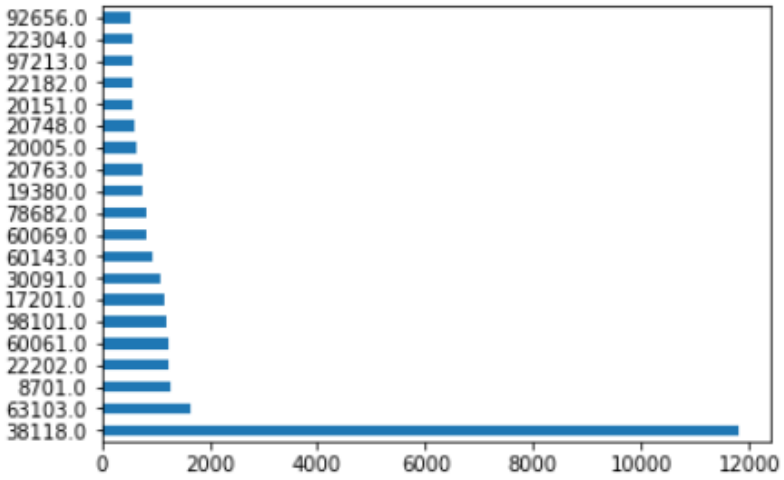
Field 7-Field Name: Merchant Zip

Description: “Merchant Zip” is a numerical variable that contains the zip code information of transactions. This field has 4,656 missing values.

Missing Values: Yes. 95.2% populated.

Unique Values: 4,568 unique records. The most frequent zip code “38118” appeared 11,823 times, which seems like a fraudulent zip code.

The distribution of the top 20 records:

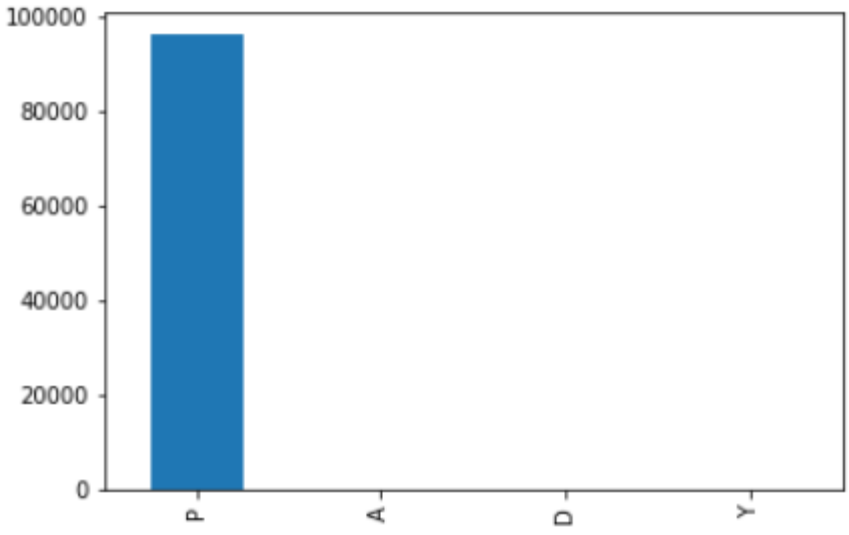


Field 8-Field Name: Transtype

Description: “Transtype” is a categorical variable that shows the transaction type.

Missing Values: No. 100% populated.

Unique Values: 4 types. No missing values. The type ‘P’ appeared 96,353 time, which is suspicious. The distribution of the top 20 records:

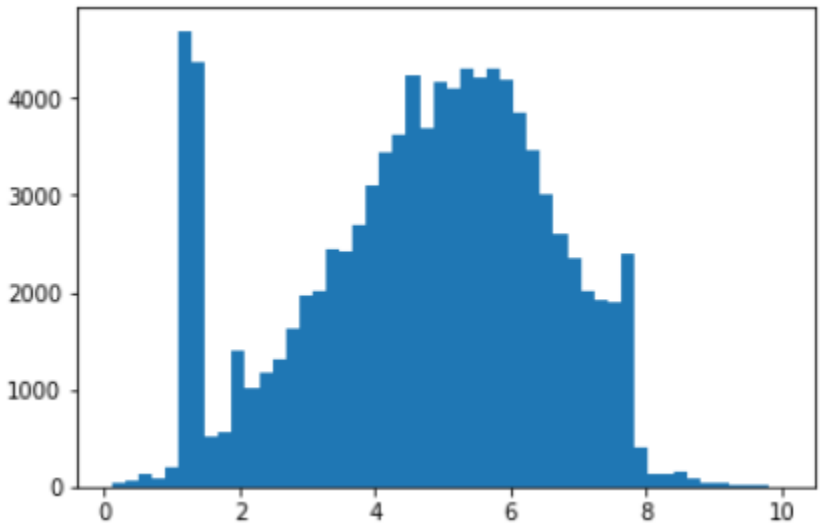


Field 9-Field Name: Amount

Description: “Amount” is a numerical variable that show the amount of different transactions.

Missing Values: 100% populated.

Unique Values: The distribution of amount is showed below:



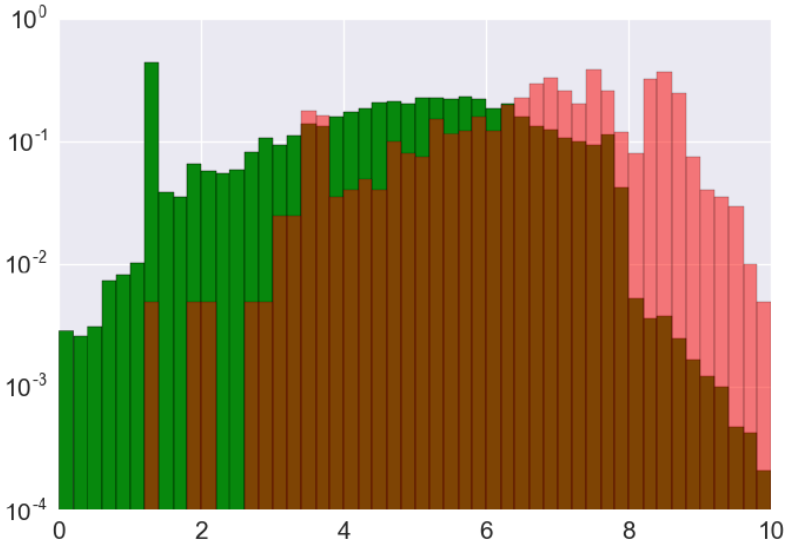
Field 10-Field Name: fraud

Description: “fraud” is a categorical variable that contains fraud condition of the recording.

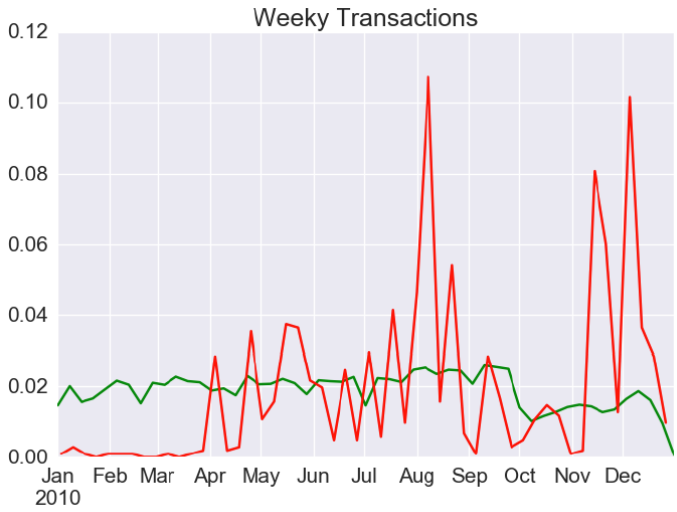
Missing Values: No. 100% populated.

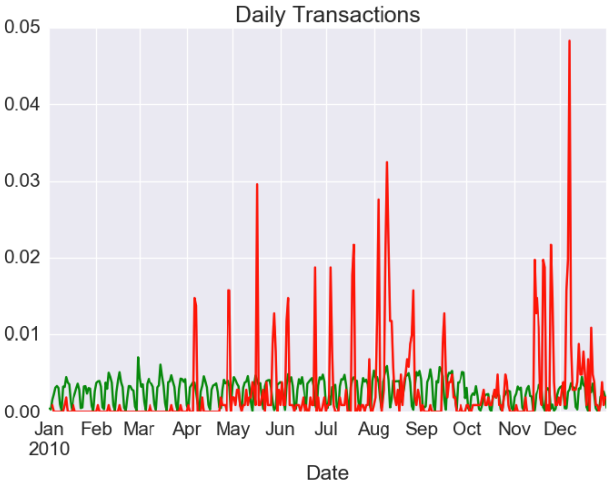
Unique Values: Two values in the data “0” and “1”. The distribution of the fraud condition:

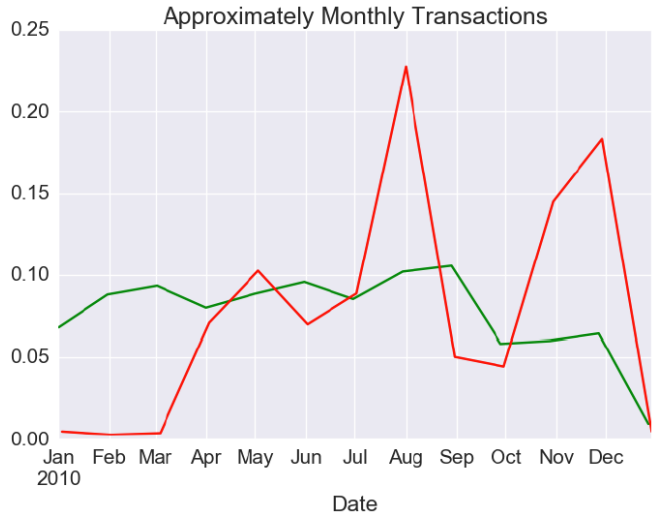
Red: Fraudulent Record
Green: Non-Fraud Record

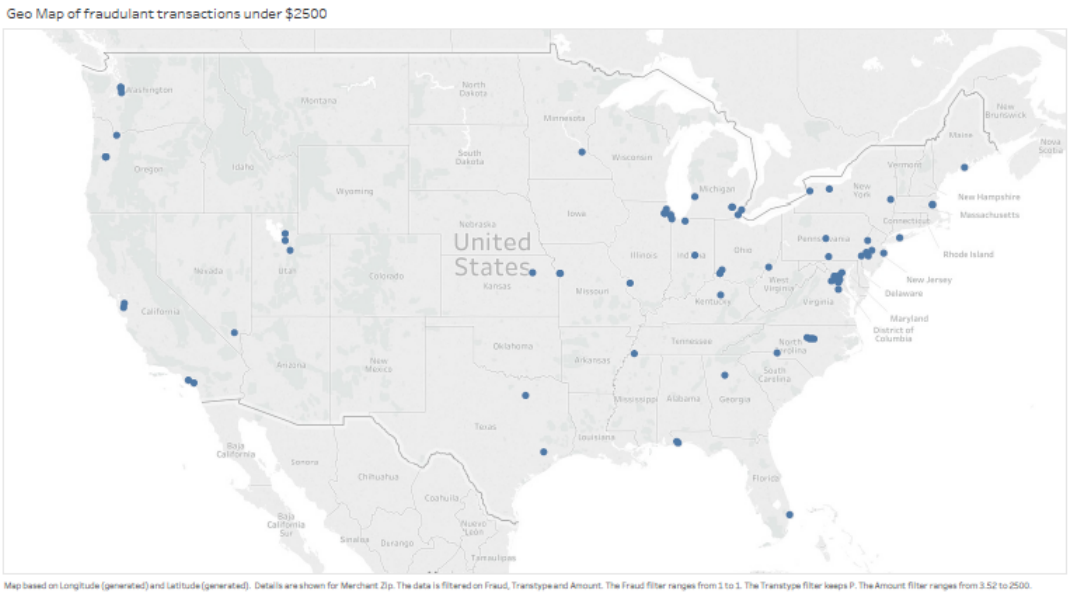


Here are more plots of fraudulent transaction in different time windows:









Here are further insights regarding dataset:

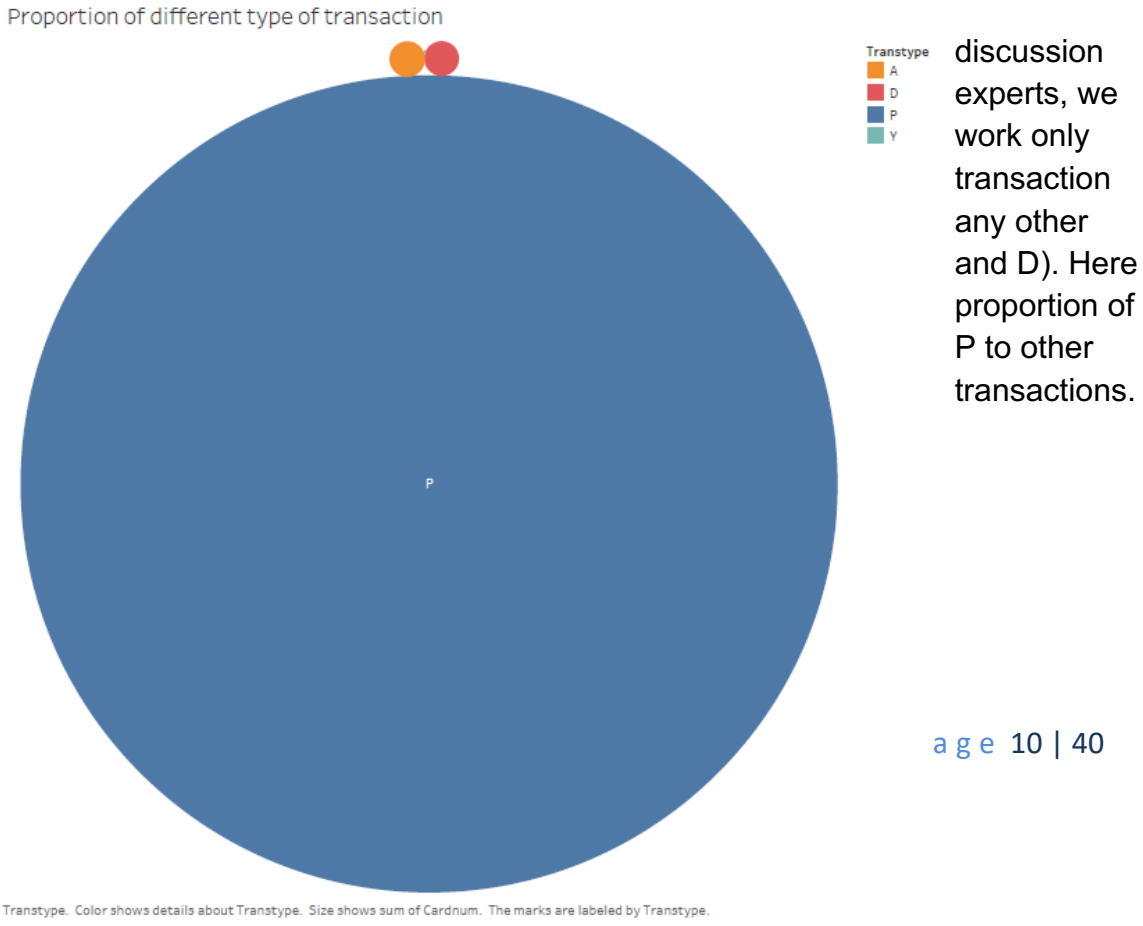
Data Cleaning

After exploring our dataset, we focused on addressing the missing values and out of character elements. There are three variables with missing values: Merchant Number, Merchant Zip Codes, and Merchant State. The below table summarizes our treatment of these missing values:

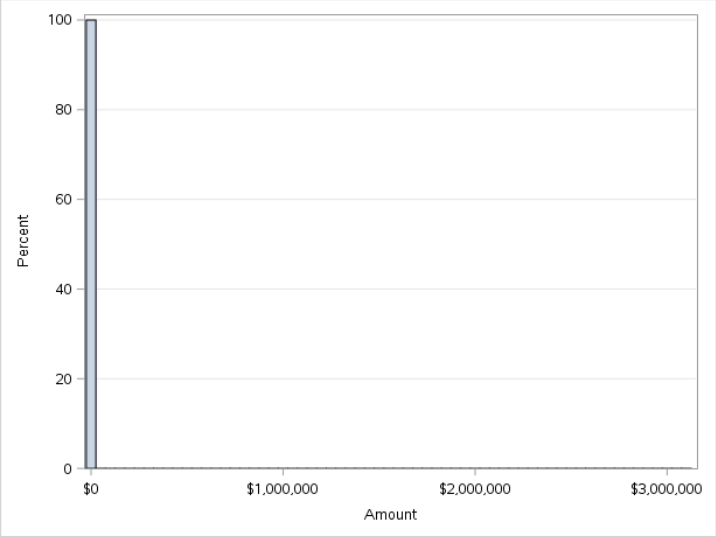
Name of the variable	Treatment of the missing variables
Merchant Number	In order to create unique number, we assigned number from 1 to 3375. Every identified missing merchantnum got a unique number in this range.
Merchant Zip	Replace with “Blank”
Merchant State	Replace with “Blank”

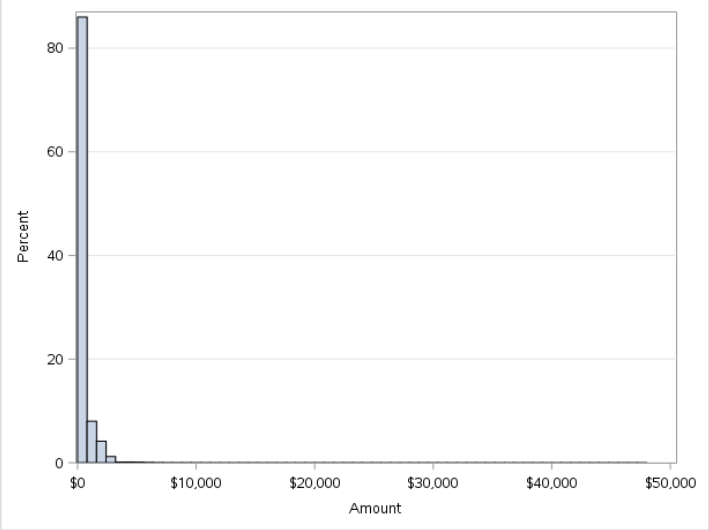
* It is important to note that we cannot replace the missing “Merchantnum” with “Blank” because we use this variable to link and group and it would create a problem going forward. Therefore, we need to make this a unique number.

- After discussion with industry experts, we decided to work only with P type transaction and remove any other type (A, Y, and D). Here is the proportion of transaction types of P to other transactions.



Furthermore, We noticed an out of character “Amount” done by “INTERMEXICO” Merchant. We removed the transaction which had really high transaction amount. There is a mistake in one of the entry’s in which amount is noted as “$3,102,045.53” This is a mistake in entering the dollar amount instead of foreign value. As seen here data is disproportionately distributed because of that value. After removing the outlier, the histogram changes to:





Expert variables

We built ---- variables using ----- given variables in the original data: Recordnum, Cardnum, Merchantnum, and fraud label from past data. Then, we focused on 4 core combination of core variables in the original data : “same_cardnum_fraud_xdays”, “same_merchant_fraud_xdays”, “same_cardnum_xdays”, and “same_merchant_xdays”. Then, we looked at how many times

the same core variables have shown up within past 1,3,7,14,30 days. Furthermore, we looked at ---------- which allowed us to create a fraud score for each “Recordnum”. All of our records were constructed to look only into the past data.

Feature Selection:

In order to increase the learning accuracy and reduce complexity, we proceed with feature selection process to select the most impactful variables and eliminate the rest. Most widely used and generally accepted method is the univariate K-S Test. The K-S test enables us to observe the distribution of each variable between goods and bads. K-S test allows us to test for differences in the shape of two sample distribution. The null hypothesis is that two samples drawn from populations should have same cumulative distribution functions. If null hypothesis is true, the defined difference should be zero and two sample follow the same distribution. The alternative hypothesis indicates that the data do not follow the same distribution. We used K-S test in R to realize this calculation and get the KS statistics for each

variables	description
same_cardnum_fraud_30	In the past 30 days, the count of fraud found for the same card number
same_cardnum_fraud_14	In the past 14 days, the count of fraud found for the same card number
same_cardnum_fraud_7	In the past 7 days, the count of fraud found for the same card number
same_cardnum_fraud_3	In the past 3 days, the count of fraud found for the same card number
same_merchantdesc_fraud_30	In the past 30 days, the count of fraud found for the same merchant description
same_merchantdesc_fraud_14	In the past 14 days, the count of fraud found for the same merchant description
same_merchantdesc_fraud_7	In the past 7 days, the count of fraud found for the same merchant description
same_merchantnum_fraud_3	In the past 3 days, the count of fraud found for the same merchant number
same_merchantdesc_fraud_3	In the past 3 days, the count of fraud found for the same merchant description
same_cardnum_fraud_1	In the past 1 day, the count of fraud found for the same card number
same_merchantdesc_fraud_1	In the past 1 day, the count of fraud found for the same merchant description
same_cardnum_7	In the past 7 day, the count of fraud found for the same card number
same_cardnum_3	In the past 3 day, the count of fraud found for the same card number
same_merchantdesc_3	In the past 3 days, the count of same merchant description

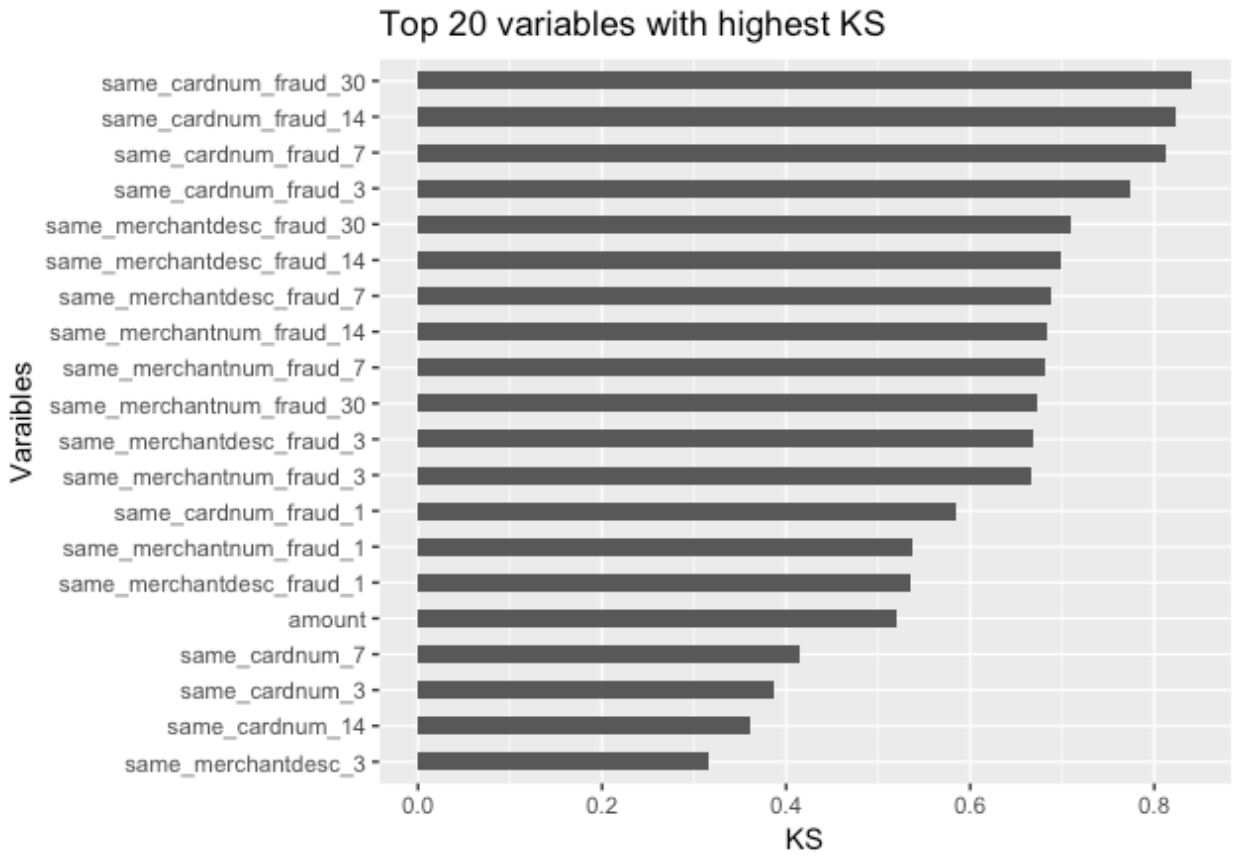
variable. We finally decided to select top 14 variables with highest KS statistics (shown in the graph below).

In regard to feature selection, we decided to choose 14 variables for two reasons:

1) The effectiveness of the variables after the 14^{Th} variables were much lower in comparison to the other variables

2) We decided to use less variables to avoid unnecessary complexity since we are dealing with small dataset.

Here is the effectiveness of the top 14 variables:

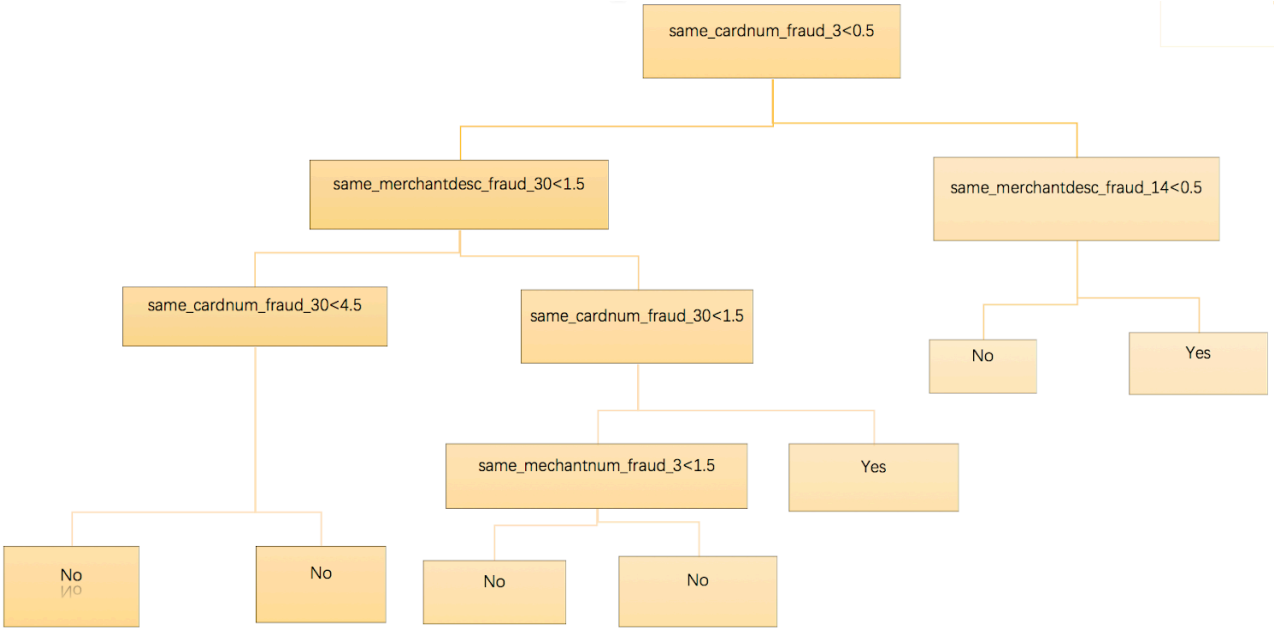


Models

In our analysis, we utilized 5 different models to be able to predict fraudulent transaction in the OOT segment of the data. These models are Decision Trees, Random Forest, Logistic Regression, Gradient Boosting Tree, and SVM. The first 10 months were used to train and test our models, and the last 2 months data was used to test the model in an environment which it has never seen before.

1) Decision Trees:

Decision trees divides the dataset into smaller and smaller data sets according to features introduced in the analysis until small enough set is achieved which describes only one feature. In this case, we used same_cardnum_fraud_3, same_merchantdesc_fraud_30, same_cardnum_fraud_30 same_merchantnum_fraud_3, same_merchantdesc_fraud_14 variables to construct the decision tree. We trained the decision trees on the training datasets. The trained tree has 7 terminal nodes. Below is an illustration of the decision tree:



FDR: Combined the predicted probability to be fraud with label of 'Yes' and reorder from high to low according label 'Yes':

Then calculated FDR of top 2% records on training data is 0.9314159; on test data is 0.9285714; and on OOT data is 0.6449704.

Confusion Matrix and Statistics:

Train	Predicted: NO	Predicted: YES
Actual: NO	58,166	169
Actual: YES	20	283

Test	Predicted: NO	Predicted: YES
Actual: NO	24,901	96
Actual: YES	6	128

OOT	Predicted: NO	Predicted: YES
Actual: NO	12,232	162
Actual: YES	15	176

2) Random Forest:

The decision trees suffer from high variance, so we try the other advanced tree model: random forest. We built the random forest tree in R on the training data and use the trained model to predict the label for each record in the testing data.

FDR: Combined the predicted probability to be fraud with label of 'Yes' and reorder from high to low according label 'Yes':

Then calculated FDR of top 2% records on training data is 0.9535398; on test data is 0.9642857; and on OOT data is 0.7455621.

Confusion Matrix and Statistics:

Train	Predicted: NO	Predicted: YES
Actual: NO	58,186	37
Actual: YES	0	415

Test	Predicted: NO	Predicted: YES
Actual: NO	24,895	41
Actual: YES	12	183

OOT	Predicted: NO	Predicted: YES
Actual: NO	12,227	49
Actual: YES	20	289

3) Logistic Regression:

In our search for best model, we applied the logistic regression model using our training dataset to predict the probability of fraud. We use the predict () function and set the type option be to "response" to tell R to output probabilities of the form P (Y = 1|X). Then we set the threshold is 0.5. It means that if the probability of one record to have the label of "download" is larger than 0.5, then the record would be predicted as "Fraud", otherwise "Not Fraud". The summary of the trained model is shown as below:

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -5.61864    0.11022 -50.978  < 2e-16
same_cardnum_fraud_30       0.16134    0.01504  10.724  < 2e-16
same_cardnum_fraud_14       0.01656    0.03035   0.546 0.585213
same_cardnum_fraud_7        0.06226    0.05283   1.179 0.238578
same_cardnum_fraud_3        0.91290    0.13176   6.928 4.25e-12
same_merchantdesc_fraud_30  0.05955    0.02796   2.129 0.033223
same_merchantdesc_fraud_14  0.02341    0.05214   0.449 0.653394
same_merchantdesc_fraud_7   0.05692    0.04527   1.257 0.208574
same_merchantnum_fraud_3    0.01825    0.07535   0.242 0.808638
same_merchantdesc_fraud_3   0.21205    0.08454   2.508 0.012128
same_cardnum_fraud_1        0.09049    0.12036   0.752 0.452160
same_merchantdesc_fraud_1   1.22207    0.13759   8.882  < 2e-16
same_cardnum_7             -0.01705    0.03086  -0.552 0.580655
same_cardnum_3             -0.28401    0.07847  -3.619 0.000295
same_merchantdesc_3        -0.12364    0.02881  -4.291 1.78e-05
```

In the table above, we can find that the significant variables are same_cardnum_fraud_30, same_cardnum_fraud_3, same_merchantdesc_fraud_30, same_merchantdesc_fraud_3, same_merchantdesc_fraud_1, same_cardnum_3, and same_merchantdesc_3.

FDR: Combined the predicted probability to be fraud with label of 'Yes' and reorder from high to low according label 'Yes':
Then calculated FDR of top 2% records on training data is 0.95089; on test data is 0.9446; and on OOT data is 0.6686391.

Confusion Matrix and Statistics:

Train	Predicted: NO	Predicted: YES
Actual: NO	58,166	169
Actual: YES	20	283

Test	Predicted: NO	Predicted: YES
Actual: NO	24,901	96
Actual: YES	6	128

OOT	Predicted: NO	Predicted: YES
Actual: NO	12,232	162
Actual: YES	15	176

4) Gradient Boosting Tree:

We further used boosting to make the predictors consecutively and not independently. In boosting the succeeding predictors learns from the mistakes of the prior predictors. The following steps were taken:

- We used the cross validation (a machine learning method) to choose most suitable parameters for model, the Tuning parameter 'n.trees' was held constant at a value of 0.001.

- Tuning parameter 'n.minobsinnode' was held constant at a value of 20.

- Accuracy was used to select the optimal model using the largest value.

- The final values used for the model were n.trees = 200, interaction.depth = 1, shrinkage = 0.001 and n.minobsinnode = 20.

- Using these parameters, we made a prediction model, and apply this model on train data and get the 'strange results'

Train	Predicted: NO	Predicted: YES
Actual: NO	58,186	452
Actual: YES	0	0

- In order to solve the strange thing, we used down sampeling to choose records again. After down sampeling, we apply the model again and get the final table (the training, the test, the oot), which are reasonable.

Confusion Matrix and Statistics:

Train	Predicted: NO	Predicted: YES
Actual: NO	367	22
Actual: YES	20	430

Test	Predicted: NO	Predicted: YES
Actual: NO	24,152	10
Actual: YES	755	214

OOT	Predicted: NO	Predicted: YES
Actual: NO	11,438	10
Actual: YES	809	328

we run the ‘pred_train_nroc = predict(train_nroc,train_apply,type="prob")’ and got the following table:

	No	Yes
1	0.9933646	0.006635381
2	0.9933646	0.006635381
3	0.9933646	0.006635381
4	0.9933646	0.006635381
5	0.9933646	0.006635381
6	0.9933646	0.006635381
7	0.9933646	0.006635381
8	0.9933646	0.006635381

Change the column ‘fraud’ in train_apply dataset into 0/1 variables (if the fraud is yes, then 1; if the fraud is no, then 0) and make it as.numerical:

Fraud
No
No
No
No
No
No
No
No
No
No

b
0
0
0
0
0
0
0
0
0
0

Combined column b with column ‘Yes’ and reorder from high to low according column ‘Yes’

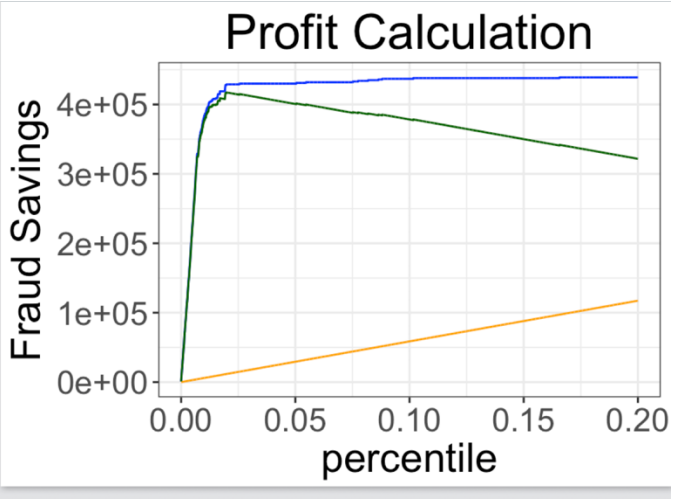
	b	a
1	1	0.1488621
2	1	0.1488621
3	1	0.1488621
4	1	0.1488621
5	1	0.1488621
6	1	0.1488621
7	1	0.1488621
8	1	0.1488621
9	1	0.1488621
10	1	0.1488621

FDR: Next, based on training data, we calculated the top 2% records and to see how many fraud are detected in top 2% recordings. We have calculated that the total fraud for all are 452. So the FDR for top 2% is 0.9380531.

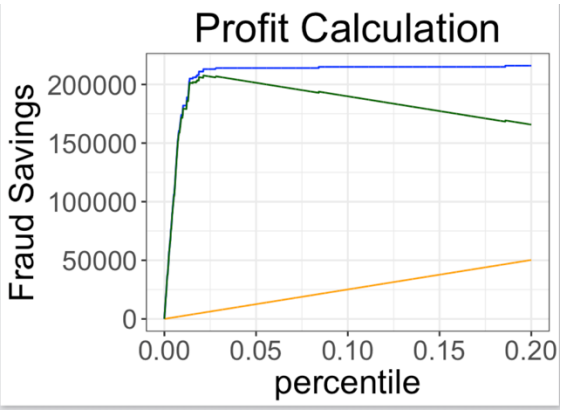
For test data and oot data, we repeat the above steps and calculated that in test data, the FDR for top 2% is 0.9464286. In oot data, the FDR for top 2% is 0.7218935.

Roi curve

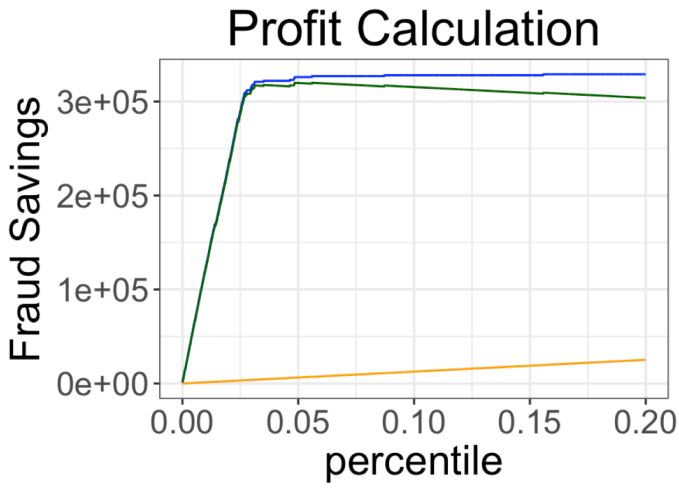
Train



Test



Oot



5) SVM:

We finally used the SVM model as follows:

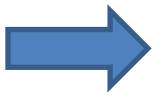
- We used SVM to make prediction model: Call: svm(formula = Fraud ~ ., data = train_apply, probability = TRUE)

- Parameters: SVM-Type: C-classification, SVM-Kernel: radial
- cost: 1 and gamma: 0.07142857
- Number of Support Vectors: 748 (353 395)

Next, calculated FDR

Change the column 'fraud' in train_apply dataset into 0/1 variables (if the fraud is yes, then 1; if the fraud is no, then 0) and make it as.numerical:

Fraud
No
No
No
No
No
No
No
No
No
No



b
0
0
0
0
0
0
0
0
0
0

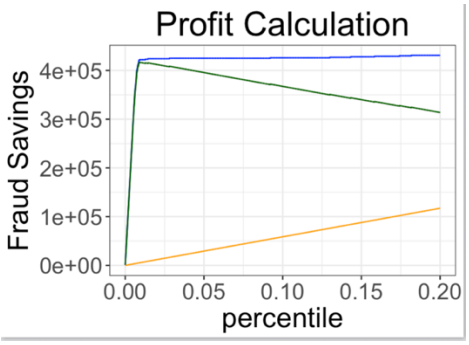
Combined column b with column 'Yes' and reorder from high to low according column 'Yes'

	b	a
1	1	1.0000000
2	1	1.0000000
3	1	1.0000000
4	1	1.0000000
5	1	1.0000000
6	1	0.9999999
7	1	0.9999998
8	1	0.9999995
9	1	0.9999995
10	1	0.9999995

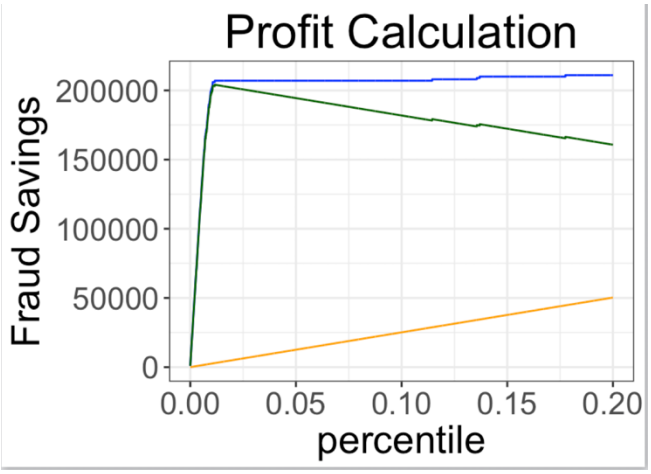
FDR: Then calculated FDR of top 2% records on training data is 0.9380531; on test data is 0.9241071; and on oot data is 0.7218935.

Roi curve

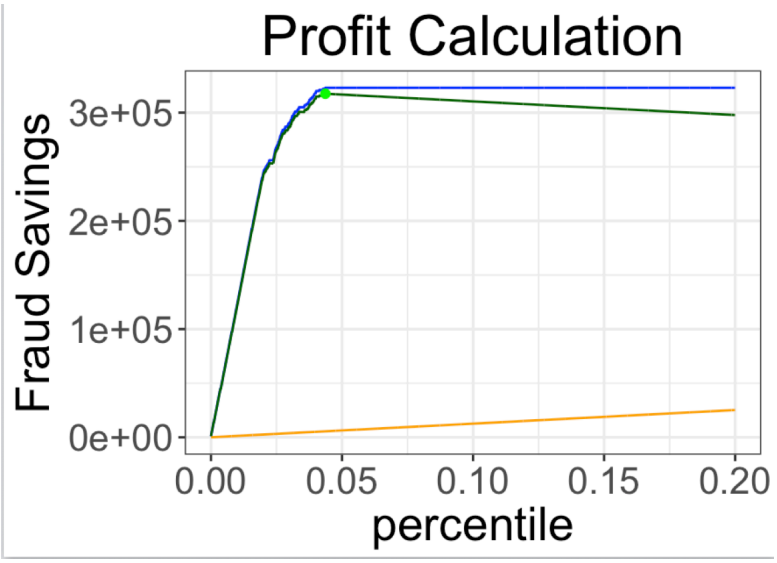
Train:



Test:



Oot



Results

The models' performance can be summarized in below table:

Comparison of FDR					
Dataset	**Gradient Boosting Tree**	**SVM**	**Decision Tree**	**Logistic Regression**	**Random Forest**
Train	93.81%	92.72%	93.14159%	94.46%	95.35398%
Test	94.64%	92.41%	92.85714%	95.089%	96.42857%
OOT	72.19%	69.18%	64.49704%	66.86391%	74.55621%

Based on our analysis, our best model performance belongs to Random Forest Model which was able to identify 74.56% of fraudulent transactions in Out of Time period.

Even though our model is not able to identify all of the fraudulent transaction, we are still able to offer saving of about $200K for top 5% of the transactions.

Appendix

Data Quality Report

Summary

File Description:

'card transactions' is a dataset containing the 96,708 records and 10 variables. It includes information about the record number, card number, the date the card transactions made, merchant number, merchant description, merchant state, merchant zip code, transaction type, transaction amount and fraud condition.

File Name:

card transactions.xlsx

Data Source:

This dataset is the simulated data based on real card transactions records, while some records are manipulated to be fraudulent for education purpose.

Number of Records:

96,708 records

Number of Fields:

10 variables in total:

Field Name	Data Type
Recordnum	Categorical variable
Cardnum	variable
Date	Date variable
Merchantnum	Categorical variable
Merch Description	Text variable
Merchant State	Categorical variable
Merchant Zip	Categorical variable
Transtype	Categorical variable
Amount	Numerical variable
Fraud	Categorical variable

Statistic Summary

	Recordnum	Cardnum	Date	Merchantnum	Merch Description	Merchant State	Merchant Zip	Transtype	Amount	Fraud
count	96708.000000	9.670800e+04	96708	93333	96708	95513	92052.000000	96708	96707.000000	96708.000000
unique	NaN	NaN	365	13090	13125	227	NaN	4	NaN	NaN
top	NaN	NaN	2010-02-28 00:00:00	930090121224	GSA-FSS-ADV	TN	NaN	P	NaN	NaN
freq	NaN	NaN	684	9310	1688	11990	NaN	96353	NaN	NaN
first	NaN	NaN	2010-01-01 00:00:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
last	NaN	NaN	2010-12-31 00:00:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	48354.500000	5.142201e+09	NaN	NaN	NaN	NaN	44709.817603	NaN	395.792713	0.010485
std	27917.339254	5.391327e+04	NaN	NaN	NaN	NaN	28376.097348	NaN	832.047787	0.101859
min	1.000000	5.142110e+09	NaN	NaN	NaN	NaN	1.000000	NaN	0.010000	0.000000
25%	24177.750000	5.142152e+09	NaN	NaN	NaN	NaN	20855.000000	NaN	33.450000	0.000000
50%	48354.500000	5.142196e+09	NaN	NaN	NaN	NaN	38118.000000	NaN	137.900000	0.000000
75%	72531.250000	5.142246e+09	NaN	NaN	NaN	NaN	63103.000000	NaN	427.665000	0.000000
max	96708.000000	5.142311e+09	NaN	NaN	NaN	NaN	99999.000000	NaN	47900.000000	1.000000

Note: Amount has the extreme value, which is 3,102,045.53, therefore considered it as outlier. Remove it before the statistic summary.

The population percentages for all variables are shown in the table below:

```
Recordnum            100.000000
Cardnum              100.000000
Date                 100.000000
Merchantnum           96.510113
Merch Description    100.000000
Merchant State        98.764321
Merchant Zip          95.185507
Transtype            100.000000
Amount               100.000000
Fraud                100.000000
dtype: float64
```

Fields Explanation

Field 1

Field Name: Recordnum

Description:

"Recordnum" is a categoriacal variable. It is the ordinal reference number for each record.

Unique Values:

100% populated with 96,708 unique values, ranging from 1 to 96,708. No repeated values or missing values exist.

Field 2

Field Name: Cardnumber

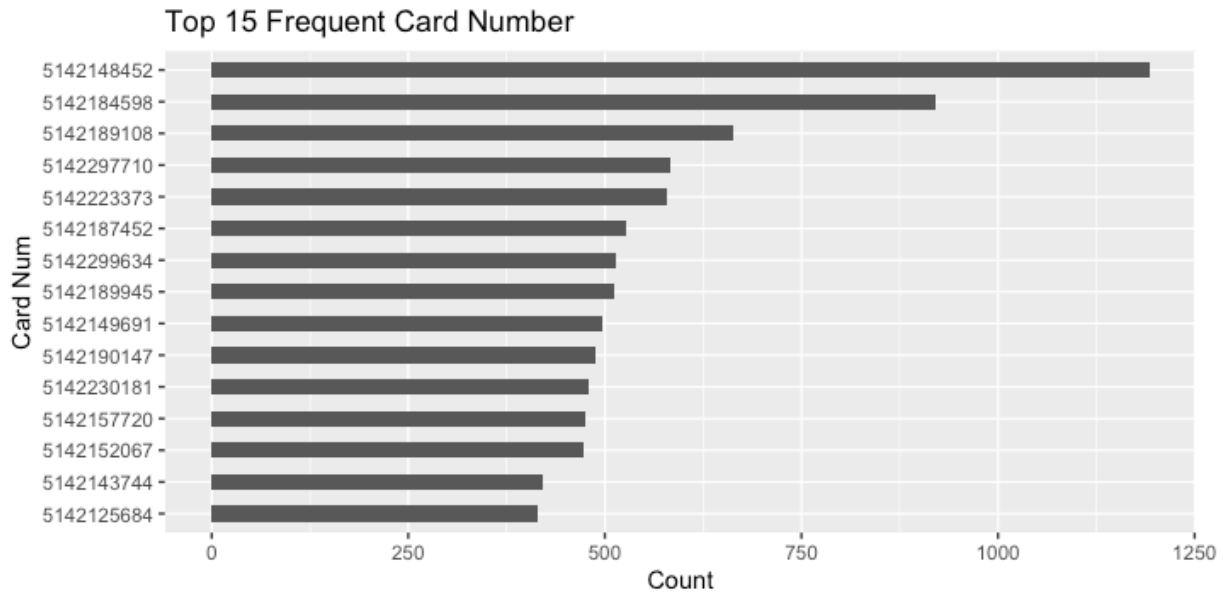
Description:

“Cardnumber” is a categorical variable, indicating the card number that the transaction was made.

Unique Values:

100% populated with 1644 unique values.

The following table shows the distribution of times of a card used, and the plot shows top 15 frequent card numbers.



Field 3

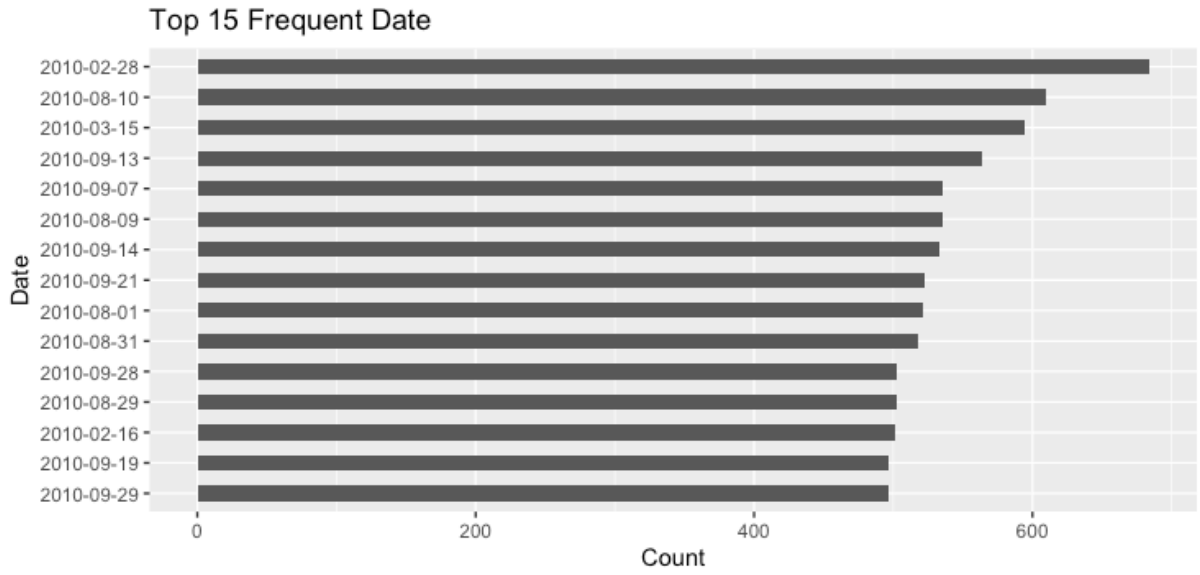
Field Name: Date

Description:

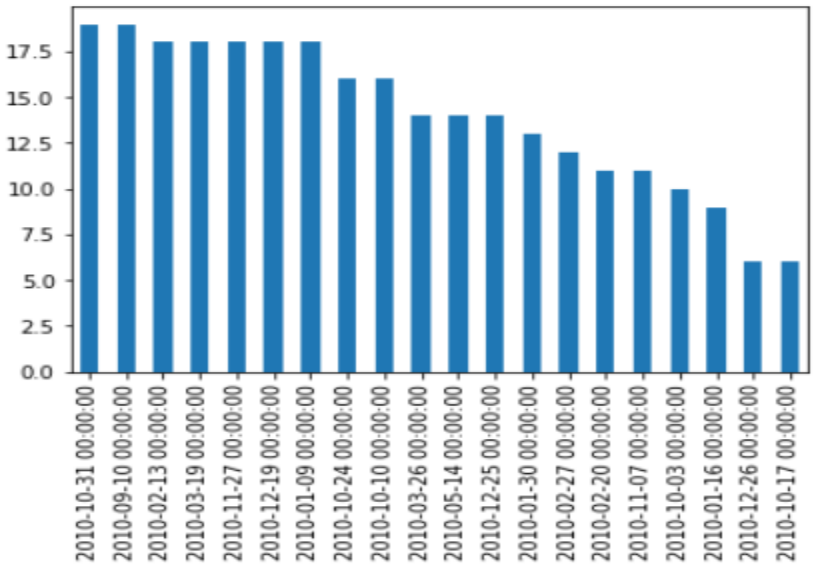
“Date” is a date variable, indicating the date of the transaction.

Unique Values:

100% populated with 365 unique values. The distribution of the top 15 frequent appeared dates are shown below.

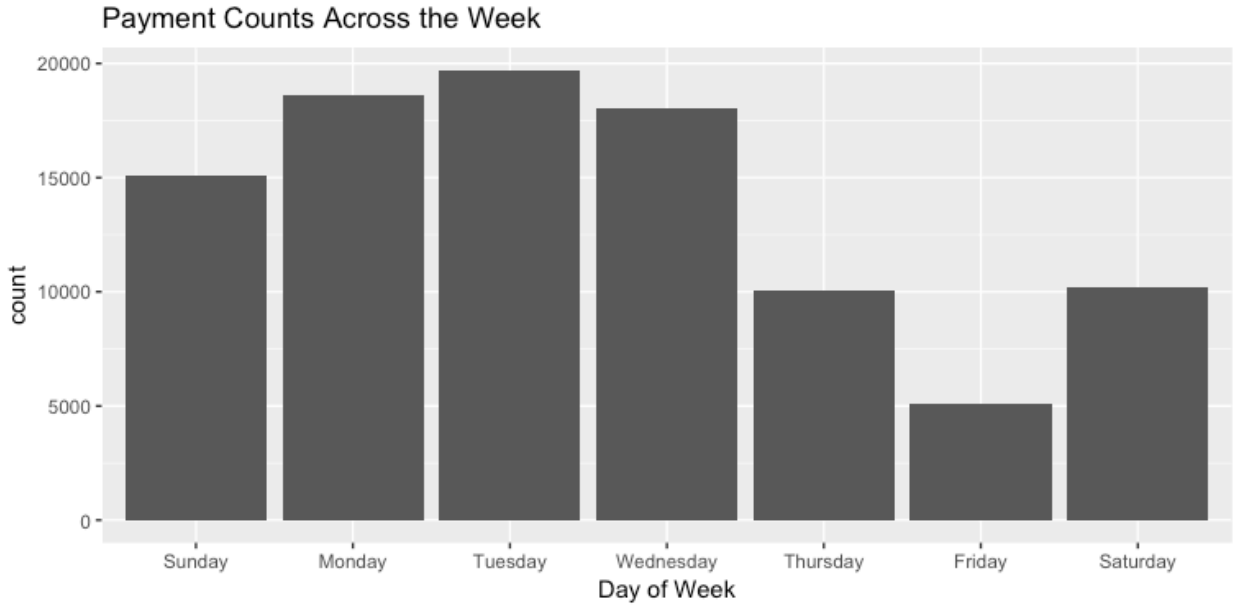


The distribution of the least 20 frequent appeared dates are shown below.

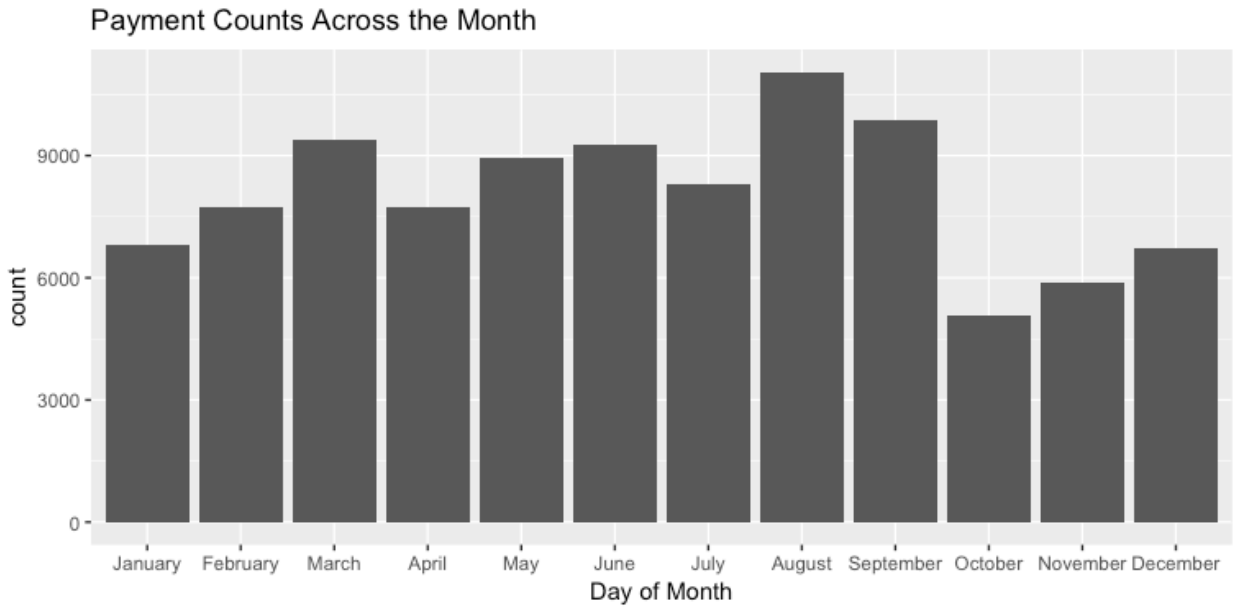


Plot the distribution of the months and day of week as below, we can see that the number of payments decreases in winter and the lowest number appears in October and Friday.

Day of week


Payment Counts Across the Week
20000
15000
10000
5000
0
count
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of Week

Day of month


Payment Counts Across the Month
9000
6000
3000
0
count
January
February
March
April
May
June
July
August
September
October
November
December
Day of Month

Field 4

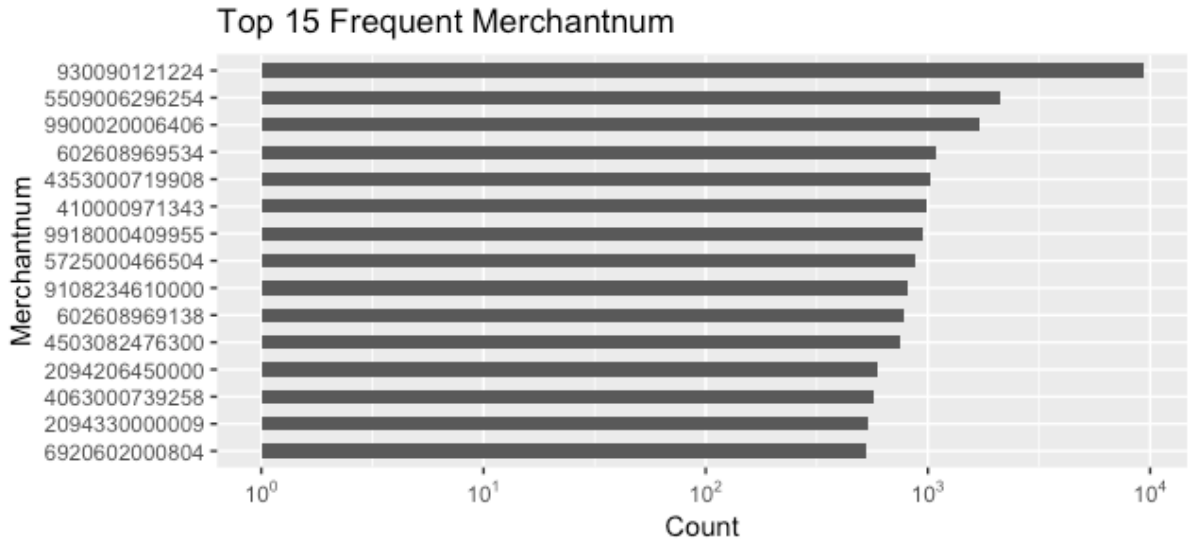
Field Name: Merchantnum

Description:

"Merchantnum" is a categorical variable, indicating the merchant number of the transactions. This field has 3,375 missing values.

Unique Values:

96.5101% populated with 13,091 unique values. The top 15 frequent appeared merchant number records are shown below. Number 930090121224 in total appeared 9310 times, which is suspicious.



Field 5

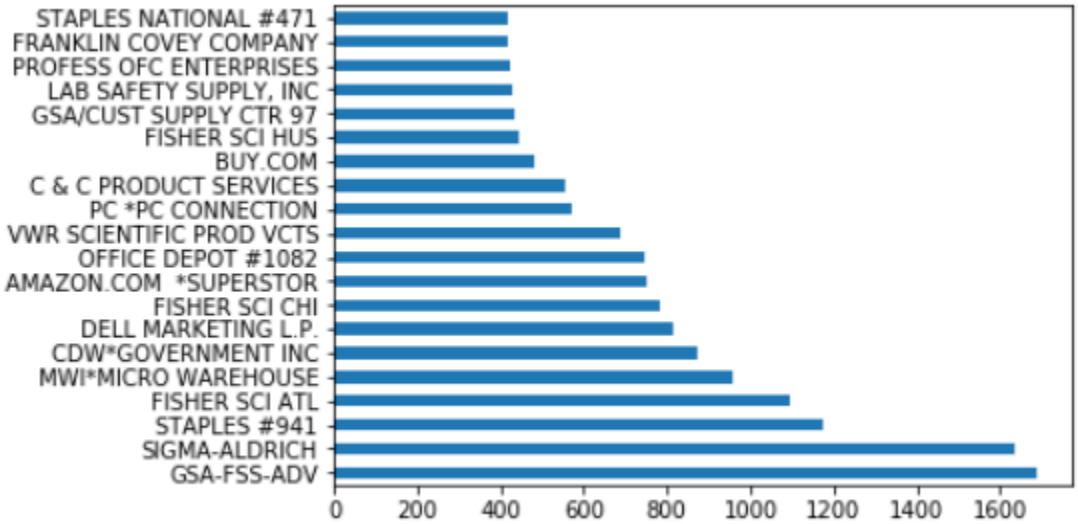
Field Name: Merch Description

Description:

"Merch Description" is a text variable, indicating the merchant information of transactions.

Unique Values:

100% populated with 13,125 unique values. The top 20 frequent appeared merchant information are shown below.



Field 6

Field Name: Merchant State

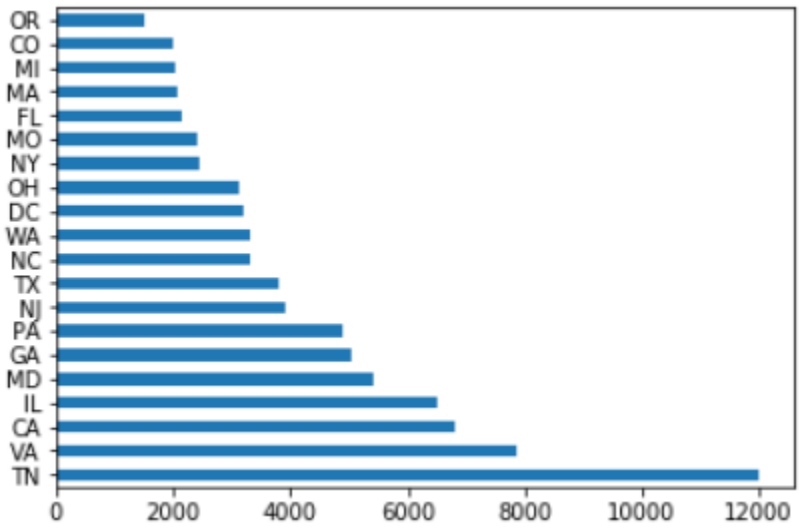
Description:

"Merchant State" is a categorical variable that shows the state of transactions. This field has 1,195 missing values.

Unique Values:

98.76% populated with 228 unique values.

The top 20 most frequent records are distributed below:



We noticed that some Merchant State is number instead of Character. It is unmoral.

```
TN     11990
VA      7872
CA      6817
IL      6508
MD      5398
GA      5025
PA      4899
NJ      3912
TX      3790
NC      3322
WA      3300
DC      3208
OH      3131
NY      2430
MO      2420
FL      2143
MA      2081
MI      2033
CO      1987
OR      1510
KS      1236
WI       953
CT       952
MN       939
UT       939
NH       908
NV       726
KY       520
RI       467
OK       411
       ...
952        1
870        1
874        1
876        1
879        1
357        1
354        1
544        1
297        1
546        1
```

Field 7

Field Name: Merchant Zip

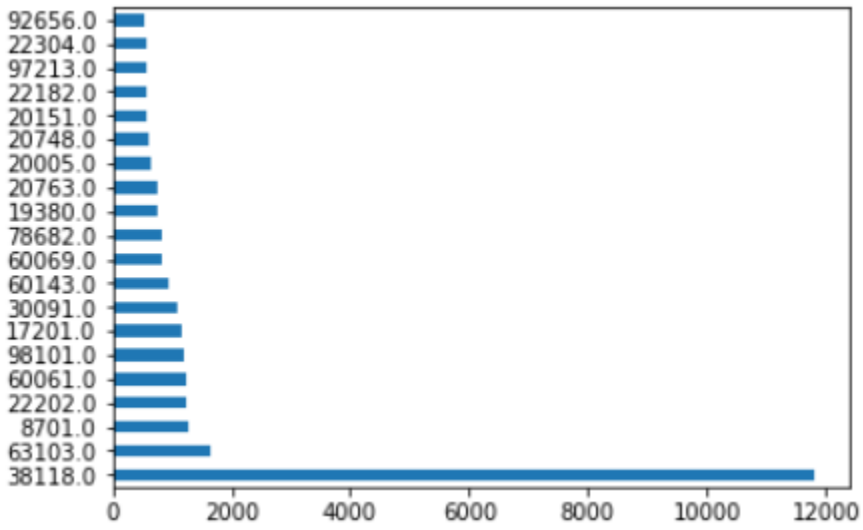
Description:

“Merchant Zip” is a categorical variable that contains the zip code information of transactions. This field has 4,656 missing values.

Unique Values:

95.2% populated with 4,568 unique records. The most frequent zip code “38118” appeared 11,823 times, which seems like a fraudulent zip code.

The distribution of the top 20 records:



Field 8

Field Name: Transtype

Description:

“Transtype” is a categorical variable that shows the transaction type.

Unique Values:

100% populated with 4 types. No missing values. The type ‘P’ appeared 96,353 time, which is suspicious.

Transtype

```
P    96353
A      181
D      173
Y        1
Name: Transtype, dtype: int64
```

Field 9

Field Name: Amount

Description:

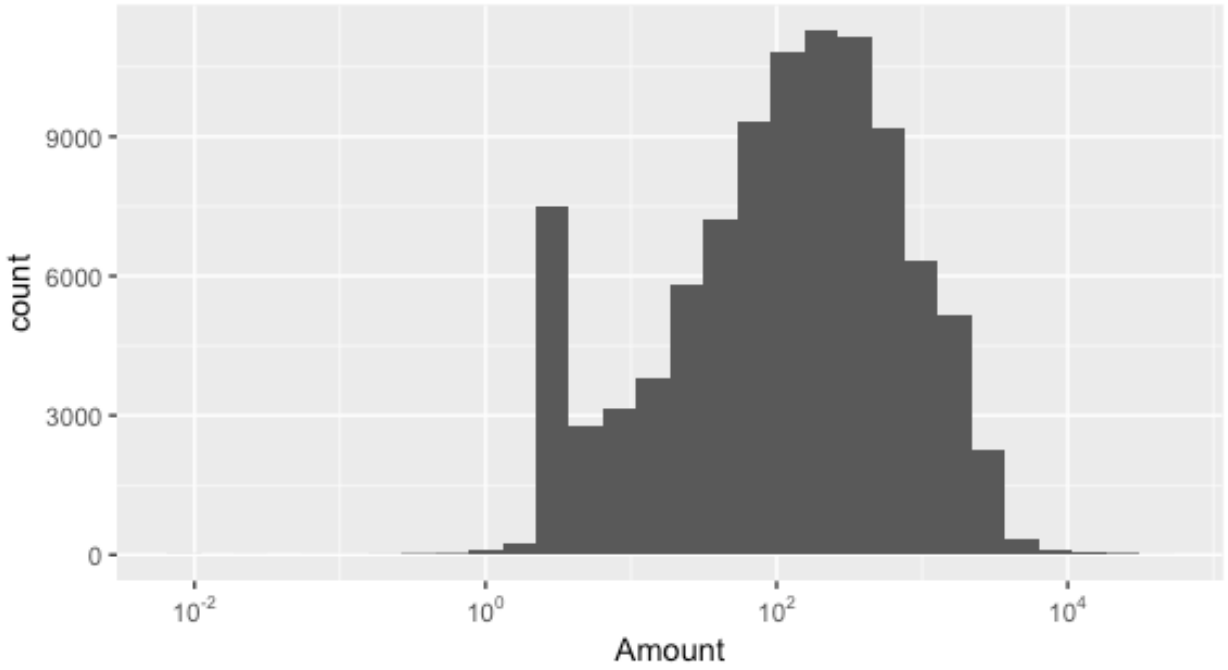
“Amount” is a numerical variable that show the amount of different transactions.

Unique Values:

The distribution of amount is showed below:

Amount has the extreme value, which is 3,102,045.53, therefore considered it as outlier. Remove it before the plotting the graph.

After eliminating the extreme large outlier, the graph:



Field 10

Field Name: fraud

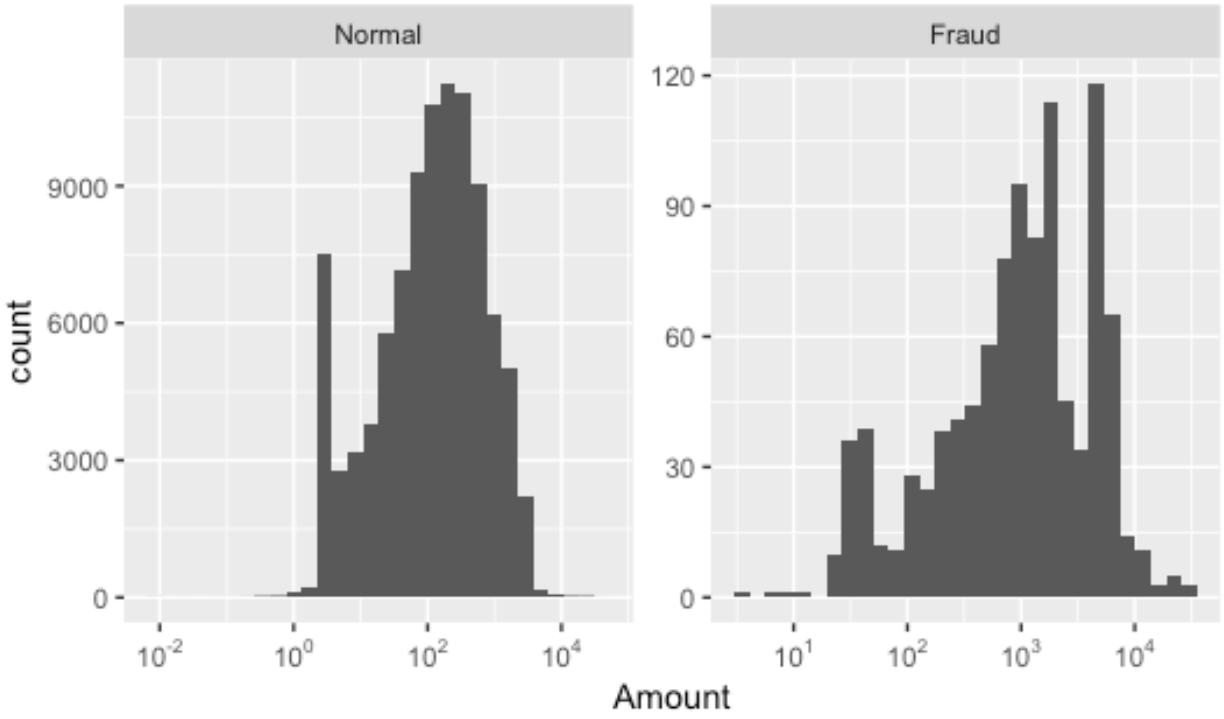
Description:

“fraud” is a categorical variable that contains fraud condition of the recording.

The distribution of the fraud condition:

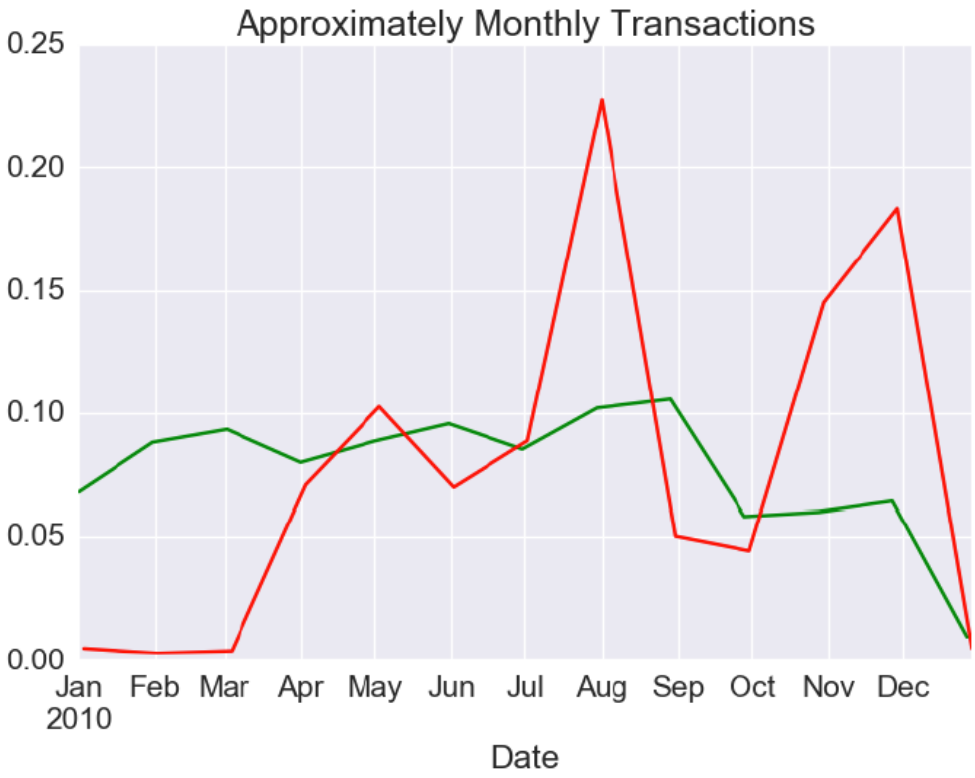
```
0     95694
1      1014
Name: Fraud, dtype: int64
```

the graph below shows the relation between Amount and Fraud. Left hand side shows the distribution of “Amount ”for “Normal” labeled records, right hand shows the distribution of “Amount ” for “Fraud” labeled records.



With fraud information, the following three plots show the monthly transactions.

The red line are fraudulent records with high possibility.


Approximately Monthly Transactions
0.25
0.20
0.15
0.10
0.05
0.00
Jan
2010
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Date