DSO 562 FRAUD ANALYTICS
Project 2

Professor Stephen Coggeshall

# Supervised Fraud Model
## *On Transaction Data*

*Team 1*

*Kalvin Tran | Aviroop Ghosal | Mohamad Ganji*
*Zhaojun (Pauline) Liu | Mark Orgel | Vu Duong*
*Jiaqi Zhu | Saif Rehman | Zhihan (Han) Li*

Mar/27/2019

# Table of Contents

# Executive Summary

In this paper, we outline the process used to develop an optimum model to identify fraudulent credit card transactions ("Transaction Data") made by employees from a Tennessee government agency in the calendar year 2010. Before cleaning and analyzing the data, an initial Data Quality Report ("DQR") was created where we reviewed the distribution and frequency of the fields (Please see Exhibit 1 for the DQR). We then proceeded to clean the data by filling in empty cells (Please see the Data Cleaning Section for more information regarding the cleaning methodology) and created 267 candidate variables (Please see the Variable Creation Section for more information regarding the created variables).

To reduce the number of variables considered for modeling, the data set was broken into three (3) parts: training, testing, and out of time (OOT) sets. The Kolmogorov-Smirnov method (KS) and Fraud Detection Rate (FDR) were then utilized to identify useful variables for predicting fraud. However, only the KS Score was utilized to filter fields as the FDR returned similar results. Of the 267 candidate variables, 26 variables were selected to model fraud.

Next, 7 different statistical models were used to predict fraud - logistic regression, K-nearest neighbor, decision tree, random forest, gradient boosting, AdaBoost, and neural network. Of the 7 models, the random forest model was considered the best model due to (i) its high FDR for both training and testing data sets and (ii) its low difference between the FDRs of both the training and testing data sets. Although the decision tree model had a higher FDR on the training data set than the random forest model, its difference with the testing data set was far greater (Approximately 20.65% for the decision tree model and approximately 10.53% for the random forest model). It was assumed that the random forest model was less likely overfitting the training data set and thus, much more generalizable for out of time fraud detections.

Predicting frauds for the OOT data set resulted in an FDR of approximately 51.4% for 3% of the population. Based upon a saving of $2,000 for every fraud caught and a loss of $50 for every false positive, the decision tree model resulted in a total savings of $170,000 at 3%. As shown in the Model Algorithms section, the client should set a score cutoff at approximately 4%, which would save approximately $178,150.

# Description of Data

Our analysis was performed using the Excel file called "card transactions.xlsx". This data was made available on Blackboard by Professor Stephen Coggeshall for education purpose.

The data reviewed in this report originally includes 10 different fields for every transaction. Of the 10 fields, 1 field is numeric and 9 are categorical. The dataset identified 96,753 transactions. Transactions marked as Non-fraud are collected in real life while the rest 1,059 fraud transactions are created by the professor for education purpose based on the characters of real fraud.

An aggregate summary of the fields' summary is presented in **Figure 1** below.

***FIGURE 1: Field Summary***

| Field Name | Field Type | Field Description | Term, Acronym, or Code Definitions | Number of valid records | Percentage of valid record | # records with value zero |
|---|---|---|---|---|---|---|
| Recnum | Categorical | File key | | 96,753 | 100.0% | 0 |
| Cardnum | Categorical | Card Number | 10-digit number | 96,753 | 100.0% | 0 |
| Date | Categorical | Transaction Date | Format: YYYY-MM-DD | 96,753 | 100.0% | 0 |
| Merchnum | Categorical | Merchant Number | | 93,378 | 96.51% | 0 |
| Merch description | Categorical | Merchant's Description | | 96,753 | 100.0% | 0 |
| Merch state | Categorical | State where Merchant is located | | 95,558 | 98.76% | 0 |
| Merch zip | Categorical | Merchant's Zip | | 92,097 | 95.19% | 0 |
| Transtype | Categorical | Transaction Type | | 96,753 | 100.0% | 0 |
| Amount | Numeric | Purchase Amount | | 96,753 | 100.0% | 0 |
| Fraud | Categorical | Mark of Fraud | 1 = Fraud<br>0 = Not Fraud | 96,753 | 100.0% | 95,694 |

Summary statistics for the numeric variables and categorical variables are presented below in **Figure 2** and **Figure 3**.

***FIGURE 2: Numeric Variables Statistics Summary***

| Field | Records | Unique Values | Mean | Median | Mode | Min | Max | Std. Dev. |
|---|---|---|---|---|---|---|---|---|
| Amount | 96,753 | 34,909 | 427.88 | 137.98 | 3.62 | 0.01 | 3,102,045.53 | 10,006.14 |

***FIGURE 3: Categorical Variables Statistics Summary***

| Field Name | Number of valid records | Most Common Field Value | Percentage | Number of Unique Field Values |
|---|---|---|---|---|
| Recnum | 96,753 | N/A | N/A | 96,753 |
| Cardnum | 96,753 | 5142148452 | 1.23% | 1,645 |
| Date | 96,753 | 2010-02-28 | 0.7% | 365 |
| Merchnum | 93,378 | 930090121224 | 9.97% | 13,092 |
| Merch description | 96,753 | GSA-FSS-ADV | 1.74% | 13,126 |
| Merch state | 95,558 | TN | 12.59% | 228 |
| Merch zip | 92,097 | 38118 | 12.89% | 4,568 |
| Transtype | 96,753 | P | 99.63% | 4 |
| Fraud | 96,753 | 0 | 98.91% | 2 |

# Summary Distributions of Most Important Variables

Below, we have identified several fields that were critical for our analysis: Amount, Merch Zip, Merch State, and Date. These are the primary variables that we used to create new variables and predict fraud. Details about these values and their distributions within the dataset are as follows:

## Amount

Amount is a numeric field that indicates the credit card transaction amount in dollars. **Figure 4** below shows a histogram of Amount under $25 bins in the range of $0 to $5,000. The bin and range combination were chosen to provide an appropriate view of the distribution. Furthermore, **Figure 5** is a box-and-whisker plot of Amount.

***FIGURE 4: Histogram of Amount for Credit Card Transactions in 2010***

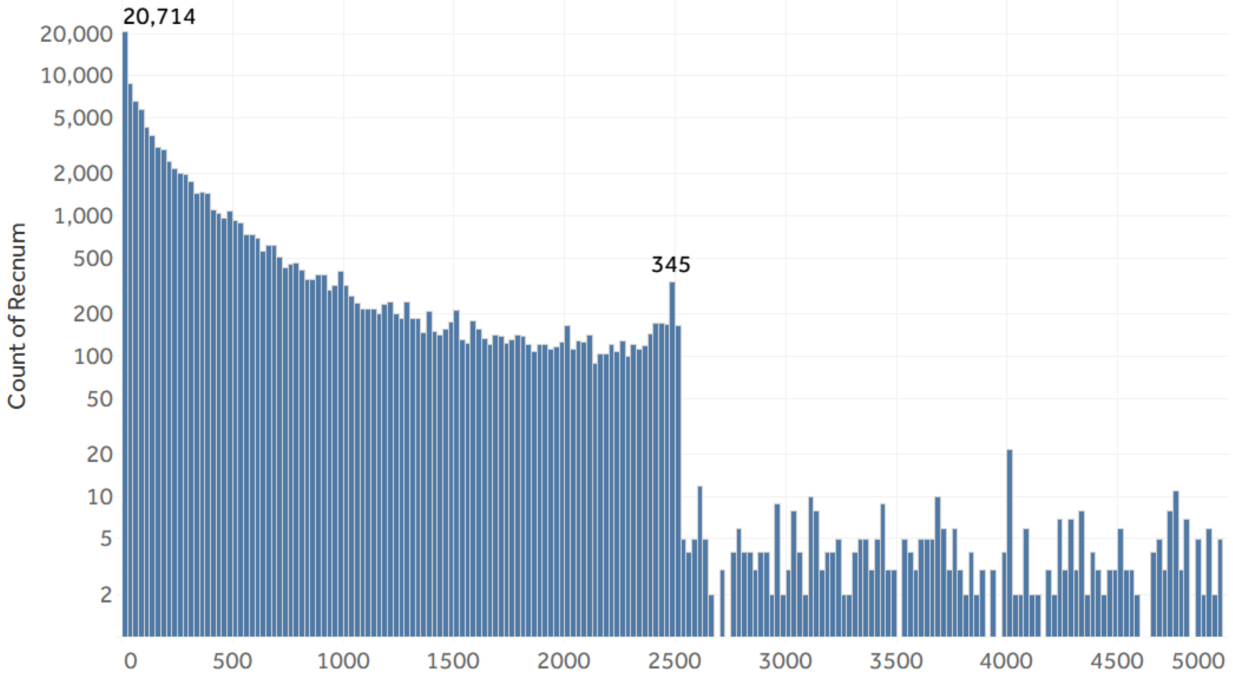
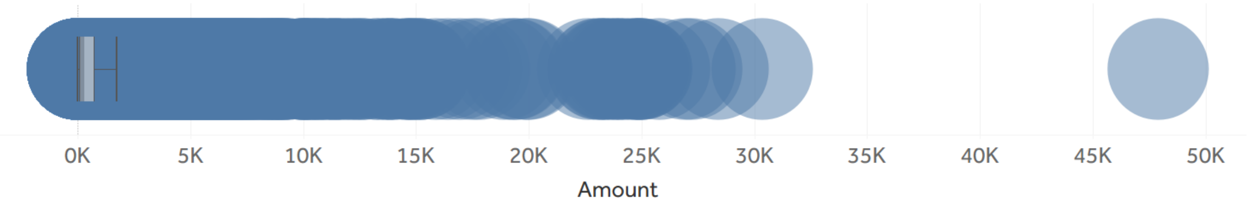


***FIGURE 5: Box-and-Whisker Plot of Amount for Credit Card Transactions in 2010[1]***

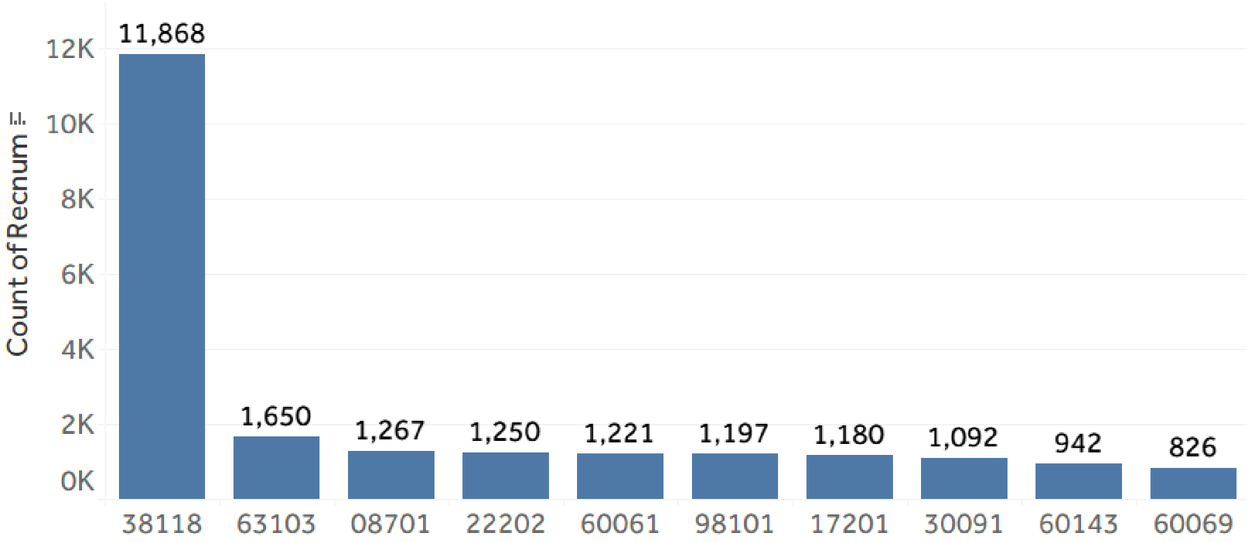
[1] Graph 12 excludes an outlier with the value of $3,102,046



## Merch Zip

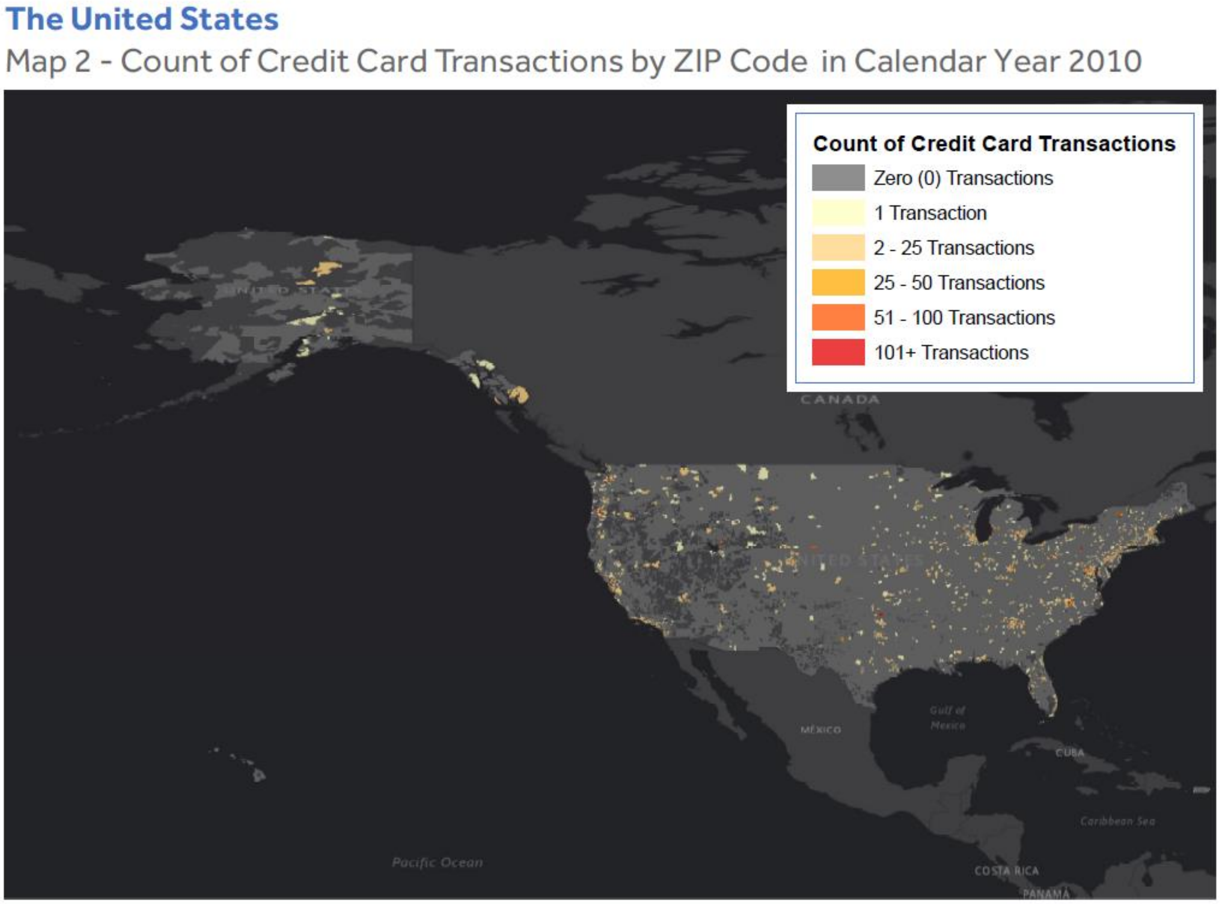
Merch Zip is a categorical field that identifies the ZIP code for the merchant providing the good or service in a transaction. As shown in the summary, there are 92,097 records with ZIP codes, and among them, there are 4,567 unique ZIP codes. **Figure 6** below shows the top ten (10) zips with the most credit card transactions in the calendar year 2010.

***FIGURE 6: Top 10 Merch Zips for Credit Card Transactions in 2010***



Like Merch State, Merch Zip was mapped to identify where the highest number of credit card transactions were by ZIP code. Please see **Figure 7** for the map of the concentration of transactions by ZIP code. The map shows that there are more ZIP codes with credit card transactions along the coast. Again, this may just be due to coastal regions being more likely to have a larger population for credit card transactions.

*FIGURE 7: Count of Credit Card Transactions by Zip*



Not all ZIP codes were able to be mapped due to limitations in available ZIP codes in shapefiles and invalid ZIP codes in the database. As a result, 11,951 transactions were left unmatched. Please see Exhibit 4 in the DQR for more information regarding unmatched ZIP codes.

## Merch State

Merch State is a categorical field that indicates the home state of the merchant providing the good or service for the transaction. As shown in the summary, there are 95,558 records for Merch State with 227 unique records. **Figure 8** shows the concentration of credit card transactions in the calendar year 2010 by state. Only 91,978 transactions were mapped (Approximately 95%). The remaining 4,775 transactions were not mapped due to limitations based on the data. Please see Exhibit 2 in DQR for more information. Below is **Figure 9** which shows the top ten (10) states with the most credit card transactions in the calendar year 2010.

***FIGURE 8: Count of Credit Card Transactions by State***

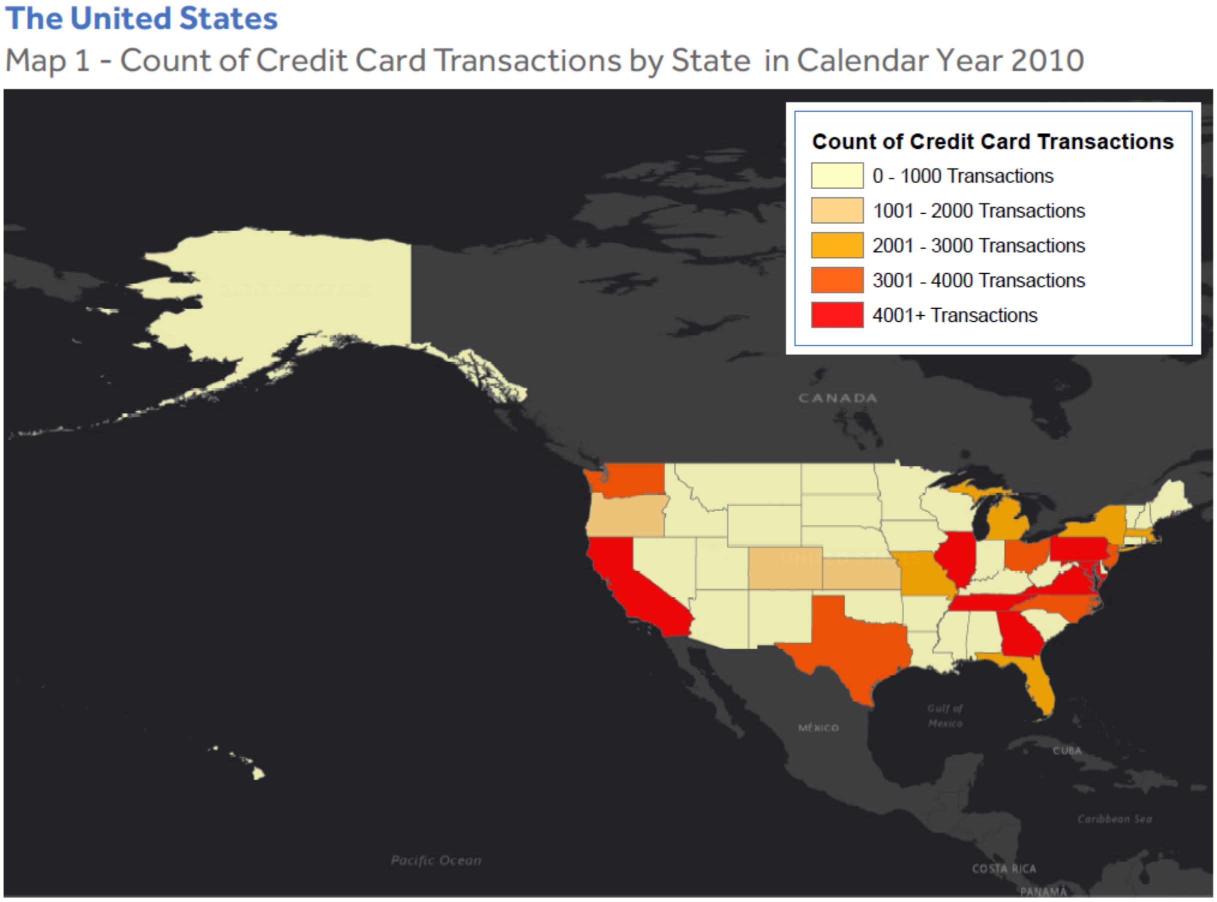
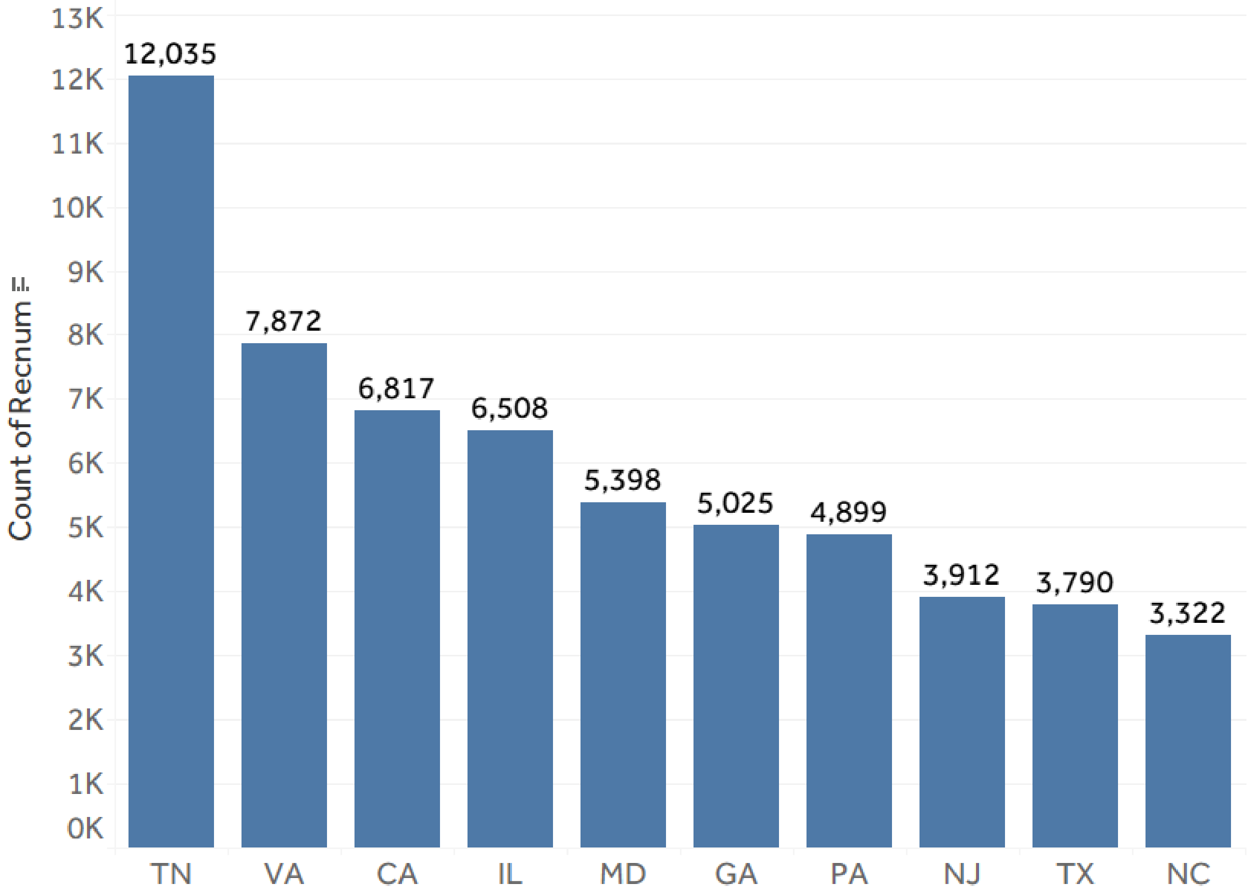


***FIGURE 9: Top 10 Merch States for Credit Card Transactions in 2010***



## Date

Date is a categorical field used to identify the month, day, and year a transaction took place in the format of mm/dd/yy. As shown in the summary, there are 96,753 records across 365 days within the calendar year of 2010. **Figure 10 through 13** provide more information on transaction trends in 2010.

***Figure 10: Total Transaction Count by Day in 2010***

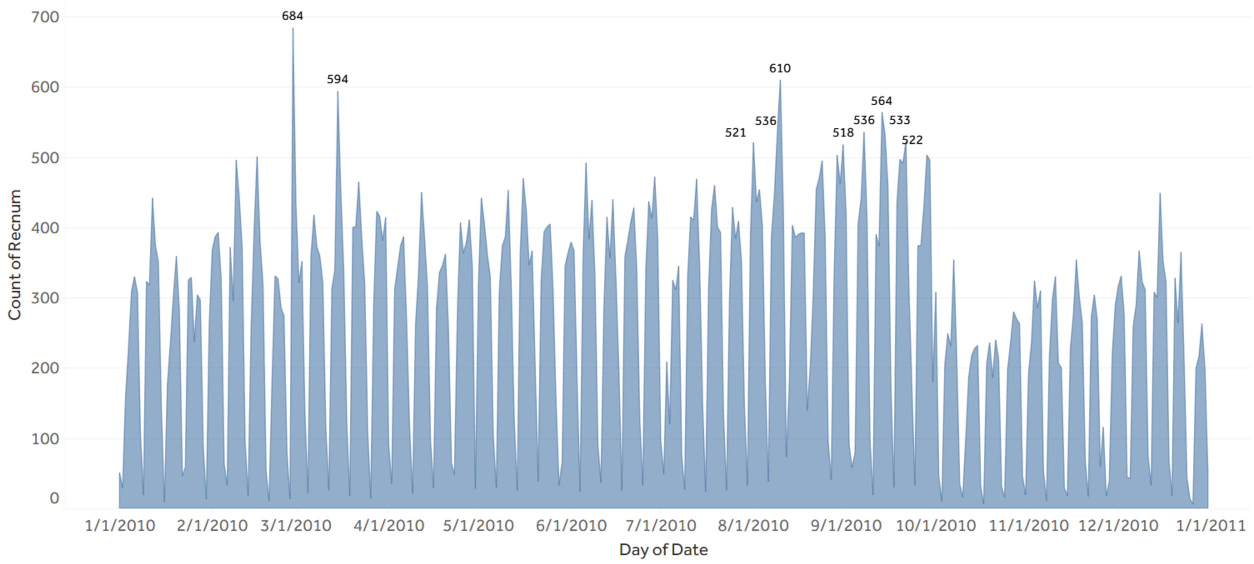


***Figure 11: Total Transaction Count by Week in 2010***

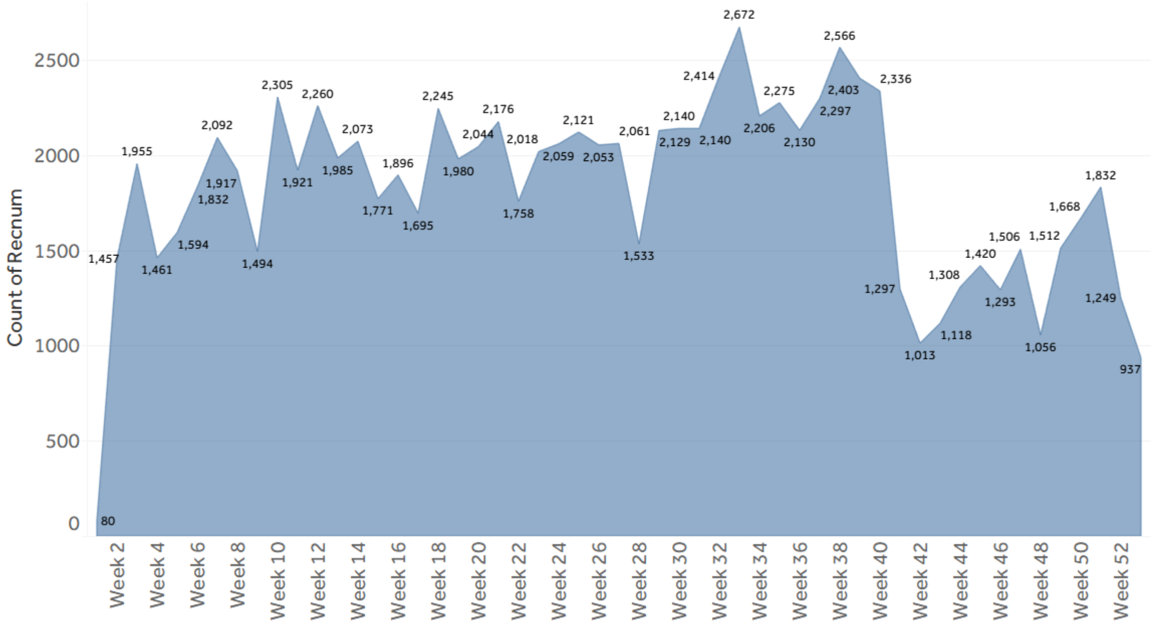


***Figure 12: Total Transaction Count by Month in 2010***

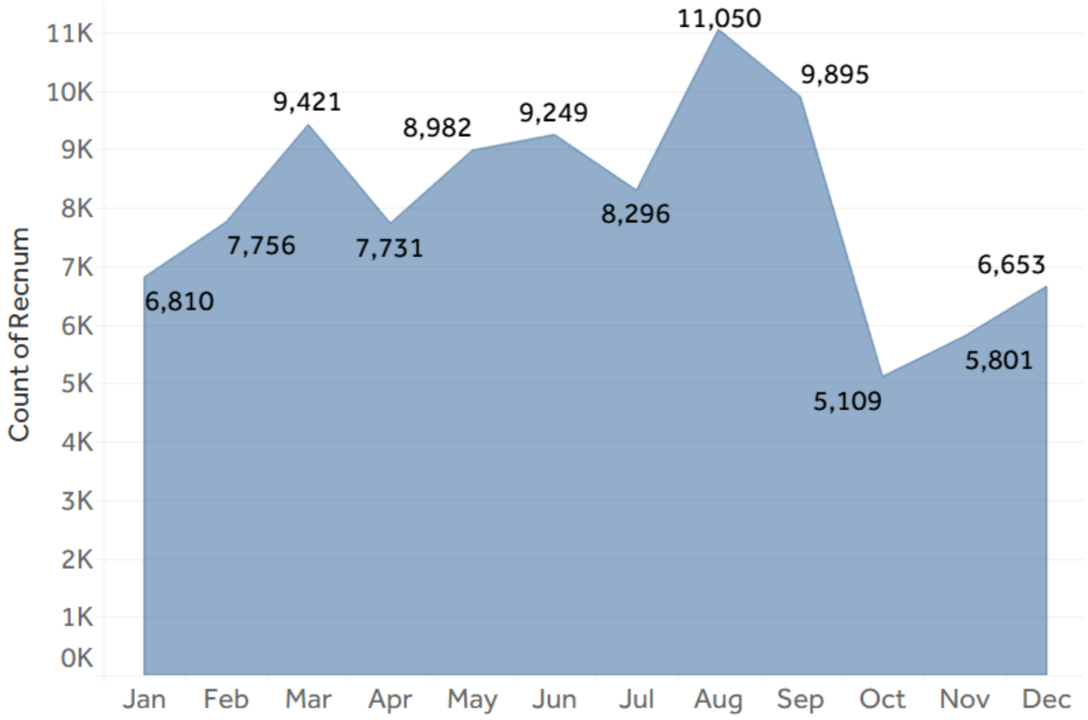
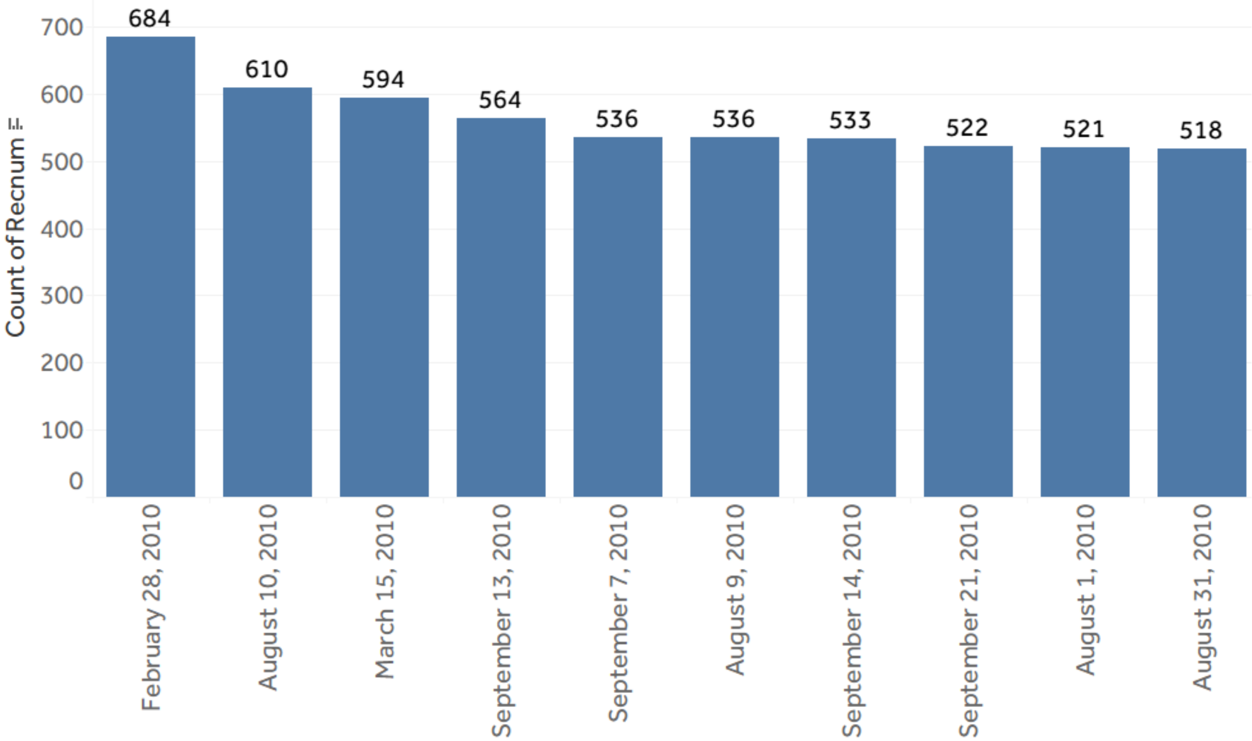


***Figure 13: Top 10 Days for Credit Card Transactions in 2010***



Based on **Figure 10**, transaction trends appear to follow a cyclical trend where there is a sharp decline in the number of transactions at the end of the week. One potential reason for this may be that individuals may tend to make more credit card transactions during the week to take care of their chores (which includes grocery shopping) and during the weekend they may have an increased propensity to stay home. However, explaining this phenomenon would require further research.

Based on **Figure 11** and **12**, the last three months of 2010 experienced significantly lower quantities of credit card transactions. During the winter season, individuals may go out less and be less likely to go shopping. Furthermore, many of the credit transactions taking place on the Eastern portion of the United States. This may have a further impact on the number of credit card transactions as the weather during the winter season may reduce the propensity for individuals to go out. However, explaining this phenomenon would require further research. Alternatively, another reason that the last three months experienced significantly lower quantities of credit card transactions is due to the nature of governmental budgets and allocations. Governmental agencies and employees tend to be more conservative with their budget allocations during the earlier portion of the fiscal year.

Lastly, based on **Figure 10** and **13**, the top ten credit card transactions are clustered around the months of August and September. As mentioned above in the previous paragraph, this may be due to increased shopping in preparation for the winter period and/or holidays. Again, further research would be required to explain this. Aside from the main cluster occurring in the months of August and September, February 28, 2010 was the day with the highest number of transactions.

# Data Cleaning

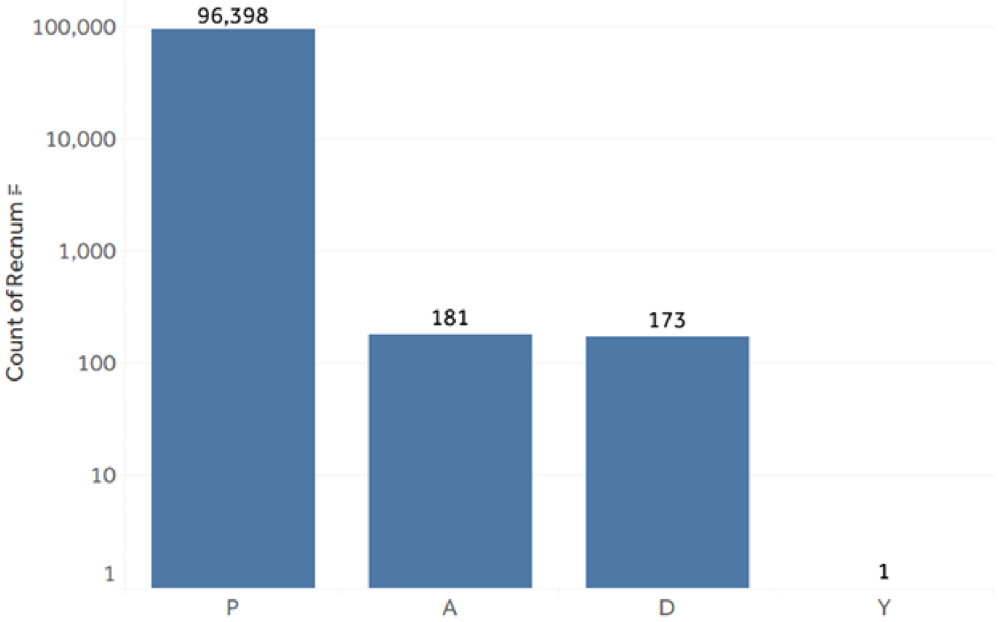
The original Credit Transaction data included missing values for three (3) fields: Merchnum, Merch State and Merch Zip. Extensive effort was put forth into determining reasonable estimates for missing values. **Figure 14** shows the count of missing data for each of these fields.

***FIGURE 14: Missing Data***

| Field | Count for Missing Data |
|---|---|
| Merchnum | 3,375 |
| Merch state | 1,195 |
| Merch zip | 4,656 |

First, we kept only Type P transactions. **Figure 15** shows the distribution of transactions under different types. Only 96,398 transactions were kept for the next step.

***Figure 15: Total Credit Card Transactions by Transtypes in 2010***



After removing one outlier whose transaction amount is greater than $ 3,000,000, we utilized Python in order to create new variables and fill in missing values. Please note that within the three fields processed in this section, there are no records that have value zero (0). Missing fields were filled with the modes by the aggregates and if there were still missing fields remaining, the aggregate step was escalated. For example, missing fields for Merch State was filled in with the mode aggregated by ZIP code and if there were still missing fields, it was escalated to the mode of all transactions. This process was repeated for ZIP and Merchnum. Please see the following for the escalation process:

Merch State:

1. Aggregated by Zip
2. Aggregated by all

Zip:

1. Aggregated by Cardnum and Merch State
2. Aggregated by Merch State
3. Aggregated by all

Merchnum:

1. Aggregated by Cardnum and Merch State
2. Aggregated by Merch State
3. Aggregated by all

After applying the process to fill in missing values, the data set no longer has any missing data.

# Candidate Variables

To fully explore the hidden information and relationship between fraud and fields, we extensively created:

1. One Hundred (100) new variables to study transactions under the same card or with the same merchant (Mean, Median, Max, Sum, and Count of transactions over the past 1, 3, 7, 14 and 30 days);
2. One hundred and Fifty (150) new variables to study the transactions with the same Zip code or in the same State (Mean, Median, Max, Sum, and Count of transactions over the past 1, 3, 7, 14 and 30 days);
3. Five (5) new variables to study days since variables (Current date minus date of most recent transaction with the same card, merchant, card at this merchant, card in this Zip and card in this state);
4. Six (6) new variables to study the velocity change in the same card (Average and actual velocity change over the past 7, 14 and 30 days);
5. Six (6) new variables to study the velocity change in the same merchant (Average and actual velocity change over the past 7, 14 and 30 days);

In the end, we created two hundred and sixty-seven (267) new variables.

***FIGURE 16: Variable Creation***

| *Variable Group* | *Variable Names* |
|---|---|
| Average transaction amount by card in 1, 3, 7, 14, and 30 days | mean_Cardnum_1d<br>mean_Cardnum_3d<br>mean_Cardnum_7d<br>mean_Cardnum_14d<br>mean_Cardnum_30d |
| Actual transaction amount/average transaction amount by card in 1, 3, 7, 14, and 30 days | Actual/mean_Cardnum_1d<br>Actual/mean_Cardnum_3d<br>Actual/mean_Cardnum_7d<br>Actual/mean_Cardnum_14d<br>Actual/mean_Cardnum_30d |
| Maximum transaction amount by card in 1, 3, 7, 14, and 30 days | max_Cardnum_1d<br>max_Cardnum_3d<br>max_Cardnum_7d<br>max_Cardnum_14d<br>max_Cardnum_30d |
| Actual transaction amount/maximum transaction amount by card in 1, 3, 7, 14, and 30 days | Actual/max_Cardnum_1d<br>Actual/max_Cardnum_3d<br>Actual/max_Cardnum_7d<br>Actual/max_Cardnum_14d<br>Actual/max_Cardnum_30d |

| *Variable Group* | *Variable Names* |
|---|---|
| Median transaction amount by card in 1, 3, 7, 14, and 30 days | median_Cardnum_1d<br>median_Cardnum_3d<br>median_Cardnum_7d<br>median_Cardnum_14d<br>median_Cardnum_30d |
| Actual transaction amount/Median transaction amount by card in 1, 3, 7, 14, and 30 days | Actual/median_Cardnum_1d<br>Actual/median_Cardnum_3d<br>Actual/median_Cardnum_7d<br>Actual/median_Cardnum_14d<br>Actual/median_Cardnum_30d |
| Total transaction amount by card in 1, 3, 7, 14, and 30 days | sum_Cardnum_1d<br>sum_Cardnum_3d<br>sum_Cardnum_7d<br>sum_Cardnum_14d<br>sum_Cardnum_30d |
| Actual transaction amount/Total transaction amount by card in 1, 3, 7, 14, and 30 days | Actual/sum_Cardnum_1d<br>Actual/sum_Cardnum_3d<br>Actual/sum_Cardnum_7d<br>Actual/sum_Cardnum_14d<br>Actual/sum_Cardnum_30d |
| Count of transaction by card in 1, 3, 7, 14, and 30 days | count_Cardnum_1d<br>count_Cardnum_3d<br>count_Cardnum_7d<br>count_Cardnum_14d<br>count_Cardnum_30d |
| Actual transaction amount/Count of transaction by card in 1, 3, 7, 14, and 30 days | Actual/count_Cardnum_1d<br>Actual/count_Cardnum_3d<br>Actual/count_Cardnum_7d<br>Actual/count_Cardnum_14d<br>Actual/count_Cardnum_30d |
| Average transaction amount at merchant in 1, 3, 7, 14, and 30 days | mean_Merchnum_1d<br>mean_Merchnum_3d<br>mean_Merchnum_7d<br>mean_Merchnum_14d<br>mean_Merchnum_30d |
| Actual transaction amount/average transaction amount at merchant in 1, 3, 7, 14, and 30 days | Actual/mean_Merchnum_1d<br>Actual/mean_Merchnum_3d<br>Actual/mean_Merchnum_7d<br>Actual/mean_Merchnum_14d<br>Actual/mean_Merchnum_30d |

| ***Variable Group*** | ***Variable Names*** |
|---|---|
| Maximum transaction amount at merchant in 1, 3, 7, 14, and 30 days | max_Merchnum_1d<br>max_Merchnum_3d<br>max_Merchnum_7d<br>max_Merchnum_14d<br>max_Merchnum_30d |
| Actual transaction amount/maximum transaction amount at merchant in 1, 3, 7, 14, and 30 days | Actual/max_Merchnum_1d<br>Actual/max_Merchnum_3d<br>Actual/max_Merchnum_7d<br>Actual/max_Merchnum_14d<br>Actual/max_Merchnum_30d |
| Median transaction amount at merchant in 1, 3, 7, 14, and 30 days | median_Merchnum_1d<br>median_Merchnum_3d<br>median_Merchnum_7d<br>median_Merchnum_14d<br>median_Merchnum_30d |
| Actual transaction amount/Median transaction amount at merchant in 1, 3, 7, 14, and 30 days | Actual/median_Merchnum_1d<br>Actual/median_Merchnum_3d<br>Actual/median_Merchnum_7d<br>Actual/median_Merchnum_14d<br>Actual/median_Merchnum_30d |
| Total transaction amount at merchant in 1, 3, 7, 14, and 30 days | sum_Merchnum_1d<br>sum_Merchnum_3d<br>sum_Merchnum_7d<br>sum_Merchnum_14d<br>sum_Merchnum_30d |
| Actual transaction amount/Total transaction amount at merchant in 1, 3, 7, 14, and 30 days | Actual/sum_Merchnum_1d<br>Actual/sum_Merchnum_3d<br>Actual/sum_Merchnum_7d<br>Actual/sum_Merchnum_14d<br>Actual/sum_Merchnum_30d |
| Count of transaction at merchant in 1, 3, 7, 14, and 30 days | count_Merchnum_1d<br>count_Merchnum_3d<br>count_Merchnum_7d<br>count_Merchnum_14d<br>count_Merchnum_30d |
| Actual transaction amount/Count of transaction at merchant in 1, 3, 7, 14, and 30 days | Actual/count_Merchnum_1d<br>Actual/count_Merchnum_3d<br>Actual/count_Merchnum_7d<br>Actual/count_Merchnum_14d<br>Actual/count_Merchnum_30d |

| *Variable Group* | *Variable Names* |
|---|---|
| Average transaction amount by card at merchant in 1, 3, 7, 14, and 30 days | mean_Cardnum_Merchnum_1d<br>mean_Cardnum_Merchnum_3d<br>mean_Cardnum_Merchnum_7d<br>mean_Cardnum_Merchnum_14d<br>mean_Cardnum_Merchnum_30d |
| Actual transaction amount/average transaction amount by card at merchant in 1, 3, 7, 14, and 30 days | Actual/mean_Cardnum_Merchnum_1d<br>Actual/mean_Cardnum_Merchnum_3d<br>Actual/mean_Cardnum_Merchnum_7d<br>Actual/mean_Cardnum_Merchnum_14d<br>Actual/mean_Cardnum_Merchnum_30d |
| Maximum transaction amount by card at merchant in 1, 3, 7, 14, and 30 days | max_Cardnum_Merchnum_1d<br>max_Cardnum_Merchnum_3d<br>max_Cardnum_Merchnum_7d<br>max_Cardnum_Merchnum_14d<br>max_Cardnum_Merchnum_30d |
| Actual transaction amount/maximum transaction amount by card at merchant in 1, 3, 7, 14, and 30 days | Actual/max_Cardnum_Merchnum_1d<br>Actual/max_Cardnum_Merchnum_3d<br>Actual/max_Cardnum_Merchnum_7d<br>Actual/max_Cardnum_Merchnum_14d<br>Actual/max_Cardnum_Merchnum_30d |
| Median transaction amount by card at merchant in 1, 3, 7, 14, and 30 days | median_Cardnum_Merchnum_1d<br>median_Cardnum_Merchnum_3d<br>median_Cardnum_Merchnum_7d<br>median_Cardnum_Merchnum_14d<br>median_Cardnum_Merchnum_30d |
| Actual transaction amount/Median transaction amount by card at merchant in 1, 3, 7, 14, and 30 days | Actual/median_Cardnum_Merchnum_1d<br>Actual/median_Cardnum_Merchnum_3d<br>Actual/median_Cardnum_Merchnum_7d<br>Actual/median_Cardnum_Merchnum_14d<br>Actual/median_Cardnum_Merchnum_30d |
| Total transaction amount of certain card by card at merchant in 1, 3, 7, 14, and 30 days | sum_Cardnum_Merchnum_1d<br>sum_Cardnum_Merchnum_3d<br>sum_Cardnum_Merchnum_7d<br>sum_Cardnum_Merchnum_14d<br>sum_Cardnum_Merchnum_30d |
| Actual transaction amount/Total transaction amount by card at merchant in 1, 3, 7, 14, and 30 days | Actual/sum_Cardnum_Merchnum_1d<br>Actual/sum_Cardnum_Merchnum_3d<br>Actual/sum_Cardnum_Merchnum_7d<br>Actual/sum_Cardnum_Merchnum_14d<br>Actual/sum_Cardnum_Merchnum_30d |

| Variable Group | Variable Names |
|---|---|
| Count of transaction by certain card at merchant in 1, 3, 7, 14, and 30 days | count_Cardnum_Merchnum_1d<br>count_Cardnum_Merchnum_3d<br>count_Cardnum_Merchnum_7d<br>count_Cardnum_Merchnum_14d<br>count_Cardnum_Merchnum_30d |
| Actual transaction amount/Count of transaction by card at merchant in 1, 3, 7, 14, and 30 days | Actual/count_Cardnum_Merchnum_1d<br>Actual/count_Cardnum_Merchnum_3d<br>Actual/count_Cardnum_Merchnum_7d<br>Actual/count_Cardnum_Merchnum_14d<br>Actual/count_Cardnum_Merchnum_30d |
| Average transaction amount by card in merchant zip code in 1, 3, 7, 14, and 30 days | mean_Cardnum_Merch state_1d<br>mean_Cardnum_Merch state_3d<br>mean_Cardnum_Merch state_7d<br>mean_Cardnum_Merch state_14d<br>mean_Cardnum_Merch state_30d |
| Actual transaction amount/average transaction amount by card in merchant zip code in 1, 3, 7, 14, and 30 days | Actual/mean_Cardnum_Merch state_1d<br>Actual/mean_Cardnum_Merch state_3d<br>Actual/mean_Cardnum_Merch state_7d<br>Actual/mean_Cardnum_Merch state_14d<br>Actual/mean_Cardnum_Merch state_30d |
| Maximum transaction amount by card in merchant zip code in 1, 3, 7, 14, and 30 days | max_Cardnum_Merch state_1d<br>max_Cardnum_Merch state_3d<br>max_Cardnum_Merch state_7d<br>max_Cardnum_Merch state_14d<br>max_Cardnum_Merch state_30d |
| Actual transaction amount/maximum transaction amount by card in merchant zip code in 1, 3, 7, 14, and 30 days | Actual/max_Cardnum_Merch state_1d<br>Actual/max_Cardnum_Merch state_3d<br>Actual/max_Cardnum_Merch state_7d<br>Actual/max_Cardnum_Merch state_14d<br>Actual/max_Cardnum_Merch state_30d |
| Median transaction amount by card in merchant zip code in 1, 3, 7, 14, and 30 days | median_Cardnum_Merch state_1d<br>median_Cardnum_Merch state_3d<br>median_Cardnum_Merch state_7d<br>median_Cardnum_Merch state_14d<br>median_Cardnum_Merch state_30d |
| Actual transaction amount/Median transaction amount by card in merchant zip code in 1, 3, 7, 14, and 30 days | Actual/median_Cardnum_Merch state_1d<br>Actual/median_Cardnum_Merch state_3d<br>Actual/median_Cardnum_Merch state_7d<br>Actual/median_Cardnum_Merch state_14d<br>Actual/median_Cardnum_Merch state_30d |

| *Variable Group* | *Variable Names* |
|---|---|
| Total transaction amount of certain card by card in merchant zip code in 1, 3, 7, 14, and 30 days | sum_Cardnum_Merch state_1d<br>sum_Cardnum_Merch state_3d<br>sum_Cardnum_Merch state_7d<br>sum_Cardnum_Merch state_14d<br>sum_Cardnum_Merch state_30d |
| Actual transaction amount/Total transaction amount by card in merchant zip code in 1, 3, 7, 14, and 30 days | Actual/sum_Cardnum_Merch state_1d<br>Actual/sum_Cardnum_Merch state_3d<br>Actual/sum_Cardnum_Merch state_7d<br>Actual/sum_Cardnum_Merch state_14d<br>Actual/sum_Cardnum_Merch state_30d |
| Count of transaction by card in merchant zip code in 1, 3, 7, 14, and 30 days | count_Cardnum_Merch state_1d<br>count_Cardnum_Merch state_3d<br>count_Cardnum_Merch state_7d<br>count_Cardnum_Merch state_14d<br>count_Cardnum_Merch state_30d |
| Actual transaction amount/Count of transaction by card in merchant zip code in 1, 3, 7, 14, and 30 days | Actual/count_Cardnum_Merch state_1d<br>Actual/count_Cardnum_Merch state_3d<br>Actual/count_Cardnum_Merch state_7d<br>Actual/count_Cardnum_Merch state_14d<br>Actual/count_Cardnum_Merch state_30d |
| Average transaction amount by card in merchant state in 1, 3, 7, 14, and 30 days | mean_Cardnum_Merch state_1d<br>mean_Cardnum_Merch state_3d<br>mean_Cardnum_Merch state_7d<br>mean_Cardnum_Merch state_14d<br>mean_Cardnum_Merch state_30d |
| Actual transaction amount/average transaction amount by card in merchant state in 1, 3, 7, 14, and 30 days | Actual/mean_Cardnum_Merch state_1d<br>Actual/mean_Cardnum_Merch state_3d<br>Actual/mean_Cardnum_Merch state_7d<br>Actual/mean_Cardnum_Merch state_14d<br>Actual/mean_Cardnum_Merch state_30d |
| Maximum transaction amount by card in merchant state in 1, 3, 7, 14, and 30 days | max_Cardnum_Merch state_1d<br>max_Cardnum_Merch state_3d<br>max_Cardnum_Merch state_7d<br>max_Cardnum_Merch state_14d<br>max_Cardnum_Merch state_30d |
| Actual transaction amount/maximum transaction amount by card in merchant state in 1, 3, 7, 14, and 30 days | Actual/max_Cardnum_Merch state_1d<br>Actual/max_Cardnum_Merch state_3d<br>Actual/max_Cardnum_Merch state_7d<br>Actual/max_Cardnum_Merch state_14d<br>Actual/max_Cardnum_Merch state_30d |

| *Variable Group* | *Variable Names* |
|---|---|
| Median transaction amount by card in merchant state in 1, 3, 7, 14, and 30 days | median_Cardnum_Merch state_1d<br>median_Cardnum_Merch state_3d<br>median_Cardnum_Merch state_7d<br>median_Cardnum_Merch state_14d<br>median_Cardnum_Merch state_30d |
| Actual transaction amount/Median transaction amount by card in merchant state in 1, 3, 7, 14, and 30 days | Actual/median_Cardnum_Merch state_1d<br>Actual/median_Cardnum_Merch state_3d<br>Actual/median_Cardnum_Merch state_7d<br>Actual/median_Cardnum_Merch state_14d<br>Actual/median_Cardnum_Merch state_30d |
| Total transaction amount of certain card by card in merchant state in 1, 3, 7, 14, and 30 days | sum_Cardnum_Merch state_1d<br>sum_Cardnum_Merch state_3d<br>sum_Cardnum_Merch state_7d<br>sum_Cardnum_Merch state_14d<br>sum_Cardnum_Merch state_30d |
| Actual transaction amount/Total transaction amount by card in merchant state in 1, 3, 7, 14, and 30 days | Actual/sum_Cardnum_Merch state_1d<br>Actual/sum_Cardnum_Merch state_3d<br>Actual/sum_Cardnum_Merch state_7d<br>Actual/sum_Cardnum_Merch state_14d<br>Actual/sum_Cardnum_Merch state_30d |
| Count of transaction by card in merchant state in 1, 3, 7, 14, and 30 days | count_Cardnum_Merch state_1d<br>count_Cardnum_Merch state_3d<br>count_Cardnum_Merch state_7d<br>count_Cardnum_Merch state_14d<br>count_Cardnum_Merch state_30d |
| Actual transaction amount/Count of transaction by card in merchant state in 1, 3, 7, 14, and 30 days | Actual/count_Cardnum_Merch state_1d<br>Actual/count_Cardnum_Merch state_3d<br>Actual/count_Cardnum_Merch state_7d<br>Actual/count_Cardnum_Merch state_14d<br>Actual/count_Cardnum_Merch state_30d |
| Current date minus date of most recent transaction with same card/merchant/card at this merchant/card in this zip code/ card in this state | Days_since_per_Cardnum<br>Days_since_per_Merchnum<br>Days_since_per_Cardnum_Merchnum<br>Days_since_per_Cardnum_Merch zip<br>Days_since_per_Cardnum_Merch state |
| Average daily count of transactions with the same card over the past 7, 14, 30 days | Avg_daily_count_Cardnum_7d<br>Avg_daily_count_Cardnum_14d<br>Avg_daily_count_Cardnum_30d |

| *Variable Group* | *Variable Names* |
|---|---|
| Count of transactions with the same card over the past 1 day/average daily count of transactions with the same card over the past 7, 14, 30 days | Actual/Avg_daily_count_Cardnum_7d<br>Actual/Avg_daily_count_Cardnum_14d<br>Actual/Avg_daily_count_Cardnum_30d |
| Average daily count of transactions with the same merchant over the past 7, 14, 30 days | Avg_daily_count_Merchnum_7d<br>Avg_daily_count_Merchnum_14d<br>Avg_daily_count_Merchnum_30d |
| Count of transactions with the same merchant over the past 1 day/average daily count of transactions with the same merchant over the past 7, 14, 30 days | Actual/Avg_daily_count_Merchnum_7d<br>Actual/Avg_daily_count_Merchnum_14d<br>Actual/Avg_daily_count_Merchnum_30d |

# Feature Selection Process

Data before October 1, 2010 was randomly divided into training and testing sets. Data of the first 30 days in 2010 was removed because of the lack of prior data to calculate the candidate variables.

## Filter Stage

To reduce the dimensionality, we performed the feature selection process including methods of filter and wrapper. The main tools used in Filter stage are the Kolmogorov-Smirnov method (KS) and Fraud Detection Rate (FDR). KS is a robust measure of how well two distributions are separated (in this case, Fraud vs Non-fraud). When plotting the cumulative distribution of goods and bads sorted by the field we want to test, KS is the maximum of the difference of cumulative and it doesn't matter which sides of the plot we start.

***Figure 17: KS Algorithm***

$$KS = max \int_{x_{min}}^{x} [P_{good} - P_{bad}]dx$$

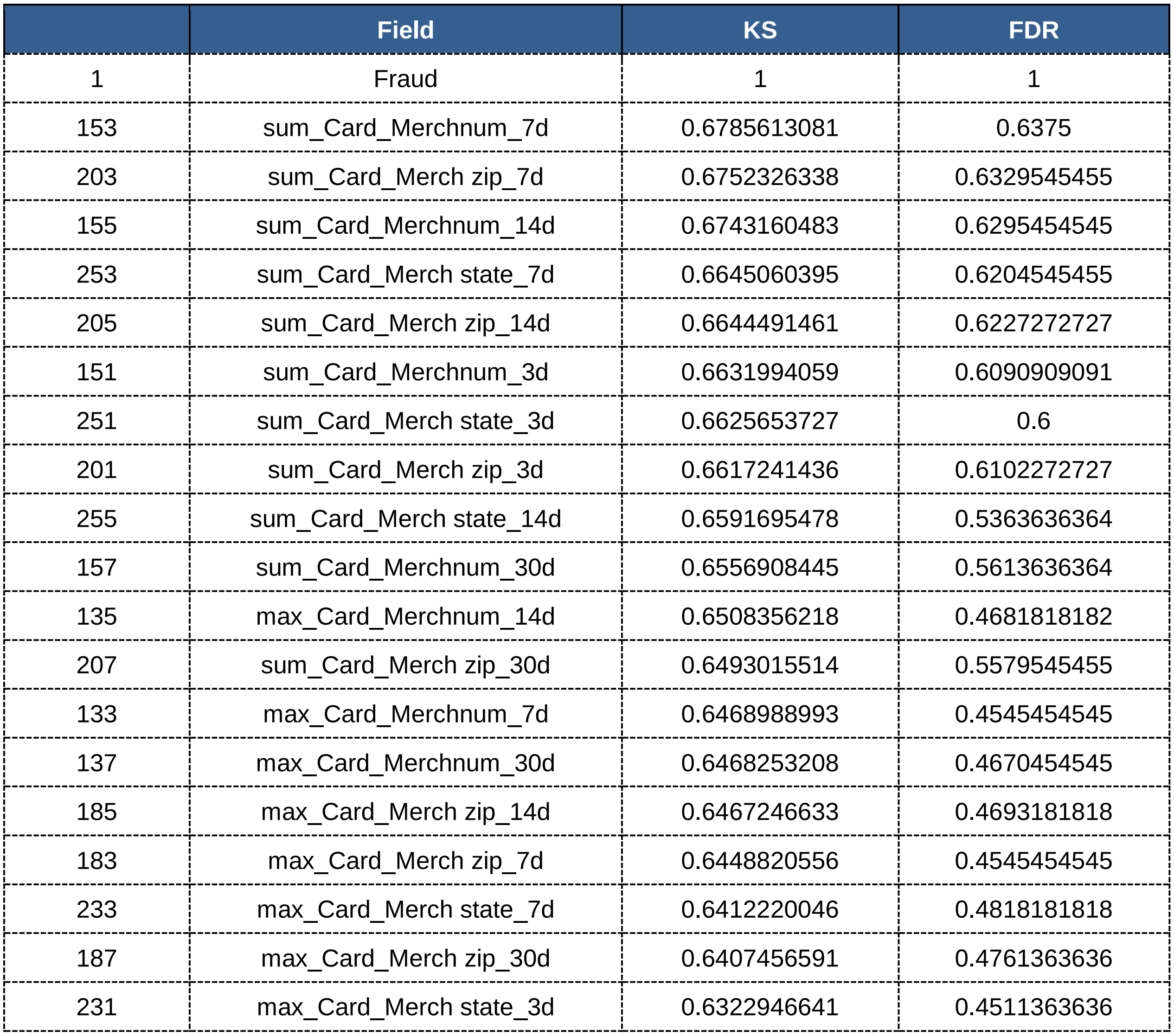
$$KS = max \sum_{x_{min}}^{x} [P_{good} - P_{bad}]$$

FDR is the other powerful tool that measures what percentage of all the frauds are caught at a particular examination cutoff location. In our case, we set the population rate at 3%. For example, FDR 75% means that the model catches 75% of overall fraud by only checking the top 3% most probable (based on probability score/propensity) fraudulent transactions of the total population.

Two fields were added into this stage as the comparison: "Fraud" from the original data and "Random" which contained only random numbers. "Fraud" should have perfect KS and FDR while "Random" is supposed to perform poorly in the test. In our case, we used only KS to filter fields while FDR was used as the reference since FDR returned a similar ranking.

136 variables were left for wrapper stage based on their KS score. The top 20 of them are listed in **figure 18**.

***FIGURE 18: Top 20 KS And FDR***



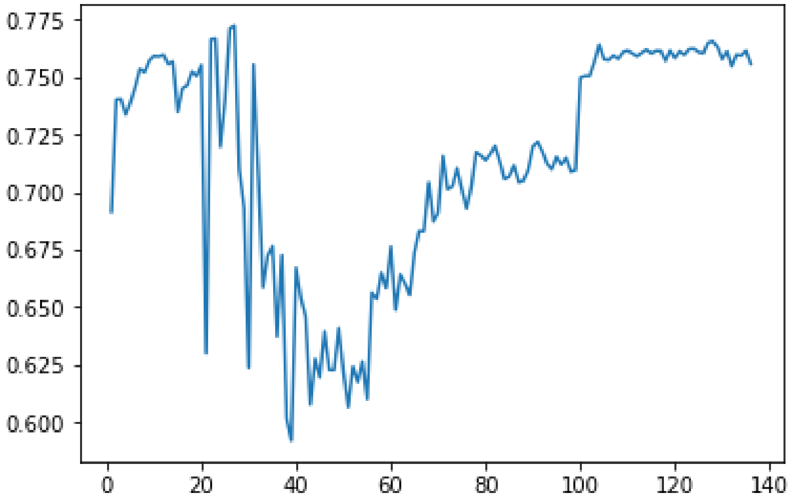
| | Field | KS | FDR |
|---|---|---|---|
| 1 | Fraud | 1 | 1 |
| 153 | sum_Card_Merchnum_7d | 0.6785613081 | 0.6375 |
| 203 | sum_Card_Merch zip_7d | 0.6752326338 | 0.6329545455 |
| 155 | sum_Card_Merchnum_14d | 0.6743160483 | 0.6295454545 |
| 253 | sum_Card_Merch state_7d | 0.6645060395 | 0.6204545455 |
| 205 | sum_Card_Merch zip_14d | 0.6644491461 | 0.6227272727 |
| 151 | sum_Card_Merchnum_3d | 0.6631994059 | 0.6090909091 |
| 251 | sum_Card_Merch state_3d | 0.6625653727 | 0.6 |
| 201 | sum_Card_Merch zip_3d | 0.6617241436 | 0.6102272727 |
| 255 | sum_Card_Merch state_14d | 0.6591695478 | 0.5363636364 |
| 157 | sum_Card_Merchnum_30d | 0.6556908445 | 0.5613636364 |
| 135 | max_Card_Merchnum_14d | 0.6508356218 | 0.4681818182 |
| 207 | sum_Card_Merch zip_30d | 0.6493015514 | 0.5579545455 |
| 133 | max_Card_Merchnum_7d | 0.6468988993 | 0.4545454545 |
| 137 | max_Card_Merchnum_30d | 0.6468253208 | 0.4670454545 |
| 185 | max_Card_Merch zip_14d | 0.6467246633 | 0.4693181818 |
| 183 | max_Card_Merch zip_7d | 0.6448820556 | 0.4545454545 |
| 233 | max_Card_Merch state_7d | 0.6412220046 | 0.4818181818 |
| 187 | max_Card_Merch zip_30d | 0.6407456591 | 0.4761363636 |
| 231 | max_Card_Merch state_3d | 0.6322946641 | 0.4511363636 |

## Wrapper Stage - Stepwise Selection Methods

A wrapper method has a model "wrapped" around the process. Our chosen model is Logistic Regression. At each step in the model selection process, we applied cross-validation 3 fold and Area Under the ROC Curve (AUC) to evaluate model performance. After running it twice and comparing AUC, we decided to go with the 26 best variables out of 136 candidate variables selected from the filter method.

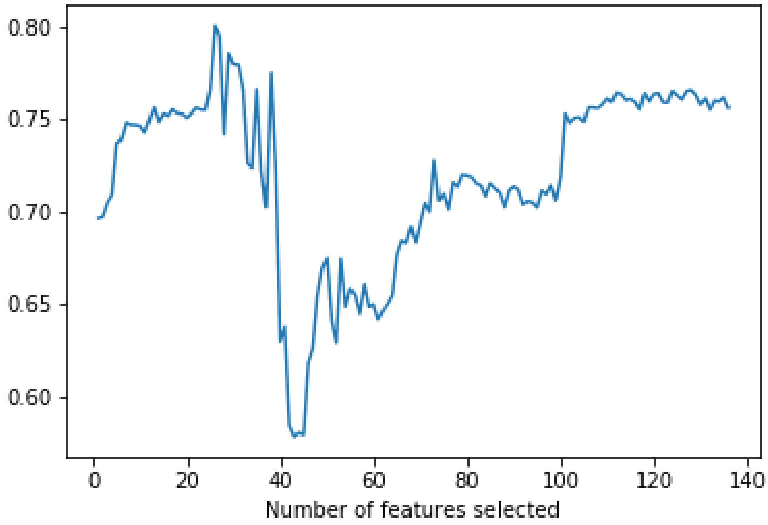
First try: AUC was maximized at about 77.5% with 27-feature model

***FIGURE 19: AUC Under the First Logistic Model***



Second try (chosen): AUC was maximized at about 80% with 26-feature model

***FIGURE 20: AUC Under the Second Logistic Model***



## Final Selected Variables

***FIGURE 21: 26 Selected Variables***

| No. | Variable Name |
|---|---|
| 1 | Actual/max_Cardnum_Merch state_3d |
| 2 | Actual/max_Cardnum_Merch zip_7d |
| 3 | Actual/mean_Merchnum_14d |
| 4 | Actual/mean_Merchnum_30d |
| 5 | Actual/mean_Merchnum_7d |
| 6 | Actual/median_Merchnum_30d |
| 7 | Actual/sum_Cardnum_1d |
| 8 | Actual/sum_Cardnum_3d |
| 9 | Actual/sum_Cardnum_Merch state_3d |
| 10 | Actual/sum_Cardnum_Merch state_7d |
| 11 | Actual/sum_Cardnum_Merch zip_3d |
| 12 | Actual/sum_Cardnum_Merch zip_7d |
| 13 | Actual/sum_Cardnum_Merchnum_7d |
| 14 | Avg_daily_count_Cardnum_14d |
| 15 | Days_since_per_Cardnum_Merch state |
| 16 | Days_since_per_Cardnum_Merch zip |
| 17 | Days_since_per_Cardnum_Merchnum |
| 18 | count_Cardnum_1d |
| 19 | count_Cardnum_Merch zip_14d |
| 20 | count_Cardnum_Merchnum_3d |
| 21 | max_Cardnum_Merch zip_7d |
| 22 | max_Cardnum_Merchnum_30d |
| 23 | max_Merchnum_30d |
| 24 | max_Merchnum_7d |
| 25 | mean_Merchnum_14d |
| 26 | median_Merchnum_7d |

# Model Algorithms

We tried multiple algorithms for classifying fraud and non-fraud: Logistic Regression, KNN, Decision Tree, Random Forest, Gradient Boosting, AdaBoost and Neural Network. The outputs of these algorithms were then combined using a table to compare their performance. We used the dataset after October 31, 2010 for the out-of-time validations, and randomly split the rest of the dataset into training (70%) and testing set (30%).

## Model 1: Logistic Regression

Logistic regression is used to describe properties of data and to explain the relationship between a dependent binary variable and one or more independent variables, used commonly in classification problems. It fits an S-curve (sigmoid function) that can map any real value input to a value between 0 and 1. We used grid search to find the best penalty method and the inverse of regularization strength, based on "accuracy" scoring method. We set the max iteration count to 200. Our training and testing FDRs were just under 60% for this model, and our out of time FDR was 27.37%, making this one of the weakest models we built.

## Model 2: KNN

The K-nearest Neighbors algorithm can be used on both classification and regression problems. In the k-NN classification model, the class of observation is determined by a plurality vote of its k neighbors. Every observation will be assigned to the class most common among k nearest neighbors. Also, the weight of all the points in each neighborhood is uniform. The parameter in KNN is usually the number of neighbors. 5 nearest neighbors were used. The FDR for training dataset was approximately 100%, but for testing, FDR dropped to only 77%. It indicated the model had an overfitting issue, possibly due to the low count of neighbors used.

## Model 3: Decision Tree

Decision Tree algorithm divides the data into rectangular spaces/ boxes and assigns a score to each box. The algorithm finds optimal cutoff points for each dimension that reduces impurities on either end of the split point. It is a rule-based algorithm that follows a graphical tree structure. The optimal candidate cutoff points are measured by impurities in resulting 2 boxes, that are added together to weighted by number in each box, and thus the cutoff point resulting in lowest impurity is selected. For continuous inputs, variance is used. This process of selecting the optimal cutoff point is repeated across all dimensions, and the location at a particular dimension at which the lowest impurity is obtained results in the optimal cutoff point, and then the same process is repeated in each of the resulting boxes. This continues until a stopping criterion of total impurity/minimum points/leaf is reached. The Decision Tree model parameters used were Gini impurity for split criteria and the minimum number samples in each box after one split was set to 15, the maximum depth was set to 14. Our Decision Tree model resulted in an FDR of 94.7% on training, 74.05% on testing, and 44.19% on out of time data. Although it resulted in a high FDR on the out of time data, due to its unstable and prone to overfitting nature, we decided to not select this as our best model.

## Model 4: Neural Network

This model mimics the way in which human brain processes information. Neural network is a deep learning algorithm that learns data representation very well and maps the input data to output data fairly accurately. Neural network is composed of input layer, hidden layer and output layer. Each layer consists of nodes that are interconnected with nodes in other layers. This algorithm has been inspired by biological neurons of the brain. The input data is fed into the input layers as input vector (1xn), this input vector is multiplied by weight matrix (nxm), where n is the number of features/number of nodes in the input layer and m is the number of nodes in the hidden layer, this is further added to bias in each of these nodes, hence obtaining a vector that is passed through an activation function (sigmoid/linear/tanh..), thus resulting in a vector that is passed to the output layer. The output layer nodes have the output data, and this output of the transfer function is mapped to it. The loss function is calculated (the actual labeled data values - the function output), these errors are back propagated to each node, the derivative of these errors with respect to the weights are calculated, thus weights are adjusted for each record. This is how the neural network algorithm reads through the entire data incrementally and then through multiple epochs, it achieves a fairly accurate classification or regression model. The important parameters in neural networks are the number of hidden layers, number of nodes, activation function, epochs, learning rate, and number of iterations. The number of neurons in the hidden layer of our Neural Network model is 100. L2 penalty parameter is set at 0.001, maximum 200 iterations, constant learning rate of 0.001, Relu as our activation function and solver as adam optimizer. Since we used a simple neural network architecture, we didn’t attain better FDRs.

## Model 5: Random Forest

Random Forest is an ensemble decision tree algorithm. It builds multiple trees by using random-chosen subsets of variables. The decision tree models are fit into these samples, based on higher probability score of the data point being classified in a class, each decision tree model votes on the majority class, thus the class with the highest mode for that record thus is selected as the class for that record in the random forest model. This algorithm is more robust than a decision tree model since even a slight modification in the training data would change the decision tree cutoff point and thus would provide a different result. The important parameters of this model are the number of trees, the maximum depth of each tree, and the minimum number of samples of each split, and we adjusted these parameters to improve the performance of the model. The number of trees used is 16 and we used gini impurity to measure the quality of the split. Minimum 20 samples were required to split at each node and the maximum depth of the trees was set to 8 nodes. Due to stable predictions, higher performance, reduced variance and reduced overfitting nature, this turned out to be our best model, as it resulted in an FDR of 51.4% on the out of time data and performed well on training (87.7%) and testing (79.8%) without overfitting.

## Models 6 & 7: Boosted Trees

Boosting is another ensemble technology. It can also improve the performance of the basic decision tree model by building many weak prediction models and eventually produce a stronger one. The parameters in boosted trees are loss function, the number of trees and maximum depth. In this project, we build adaptive boosting and gradient boosting classification models. Both these two boosting models build multiple trees by firstly randomly pick up a subset of observations with equal weights. But they perform differently when they produce the next model:

For adaptive boosting, it will use this first model to fit the whole dataset and identify the misclassified points. When creating the next model, higher weights will be given to the misclassified points so that they will be more likely to be picked up for to next model. The algorithm will repeat this process until there is no change on the loss function. Number of trees was set to 50. As shown in figure 20, the FDRs for training, testing and out of time were 76.87%, 66.54% and 36.87%.

Contrary to adaptive boosting, which increases the weights of misclassified points at every interaction, gradient boosting tries to fit the new model to the residual errors made by the previous model. The objective is to find the bests split to minimize the error. This process will be repeated until the loss function does not change. We ran model with 100 trees and 2 for the number of minimum samples in each box after every split. As shown in figure 20, the FDRs for training, testing and out of time were 83.19%, 73.15% and 31.28%.

Thus, the boosted trees performed fairly well, but the random forest model outperformed them.

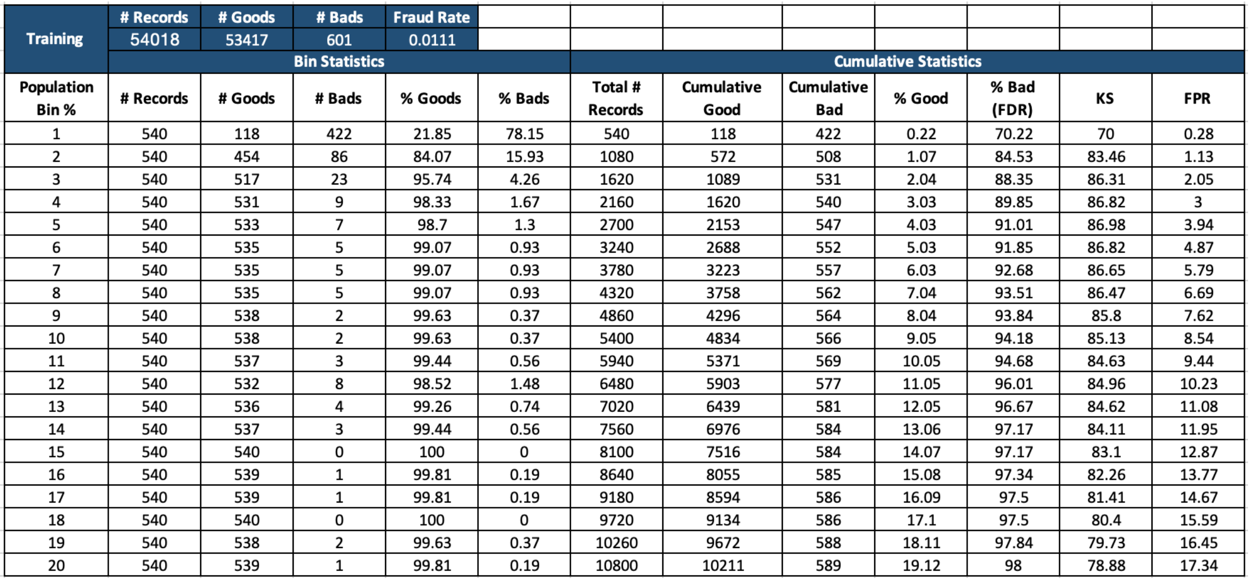
## Our Best Model

For each algorithm, we adjusted the parameters and a random-selected subset of variables to fit the model. The highest accuracy was attained by the Random Forest model. We set the number of trees as 16, the maximum depth at 8 and the minimum number of samples remain after each split was 20. We used the grid search method to find the best parameters for our model. The FDRs of our best model were 87.69%, 79.77% and 51.4% for training, testing and out of time data. The difference between training and testing was 7.92%. It indicated that the model was relatively robust for further predictions. Our baseline model was the logistic regression model and the FDRs were 58.74%, 56.03% and 27.37% for training, testing and out of time data respectively.

***FIGURE 22: FDR at 3% of Each Model***

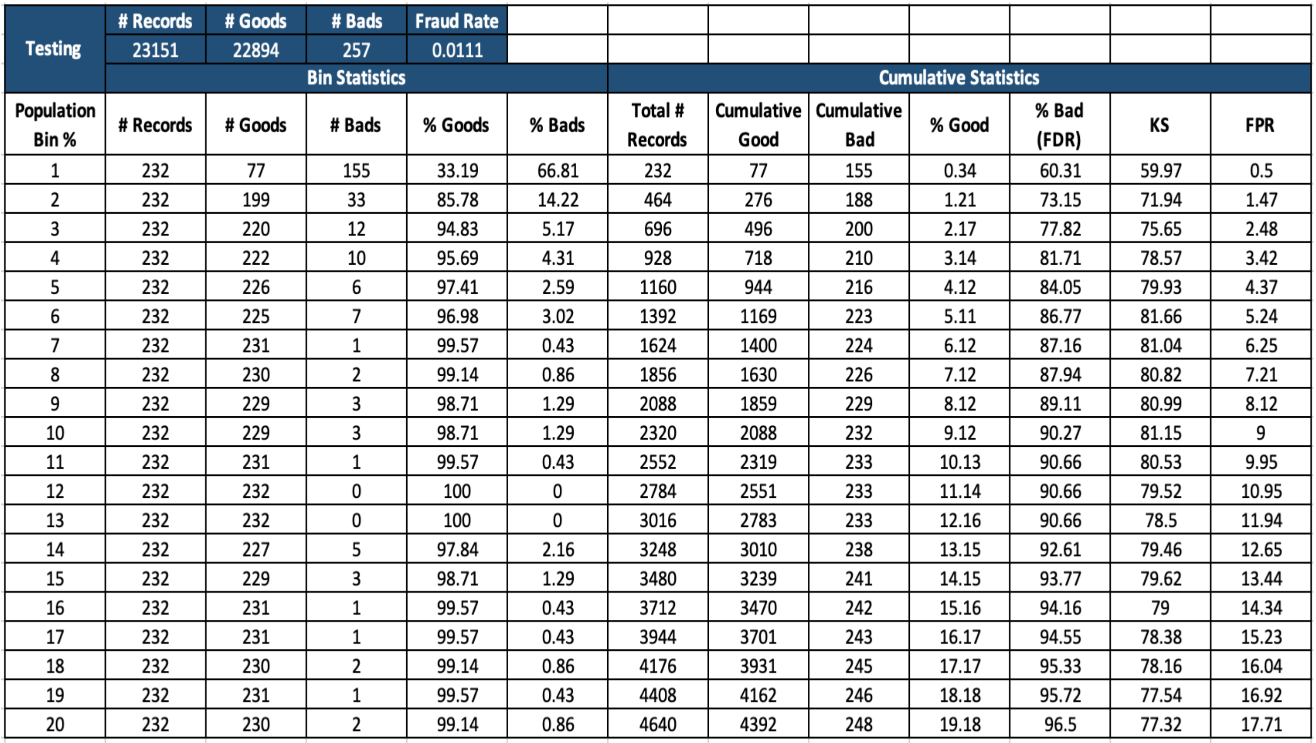
| Model Name | Training | Testing | Out of Time |
|---|---|---|---|
| Logistic Regression | 58.74 | 56.03 | 27.37 |
| KNN | 100.00 | 77.82 | 22.35 |
| Decision Tree | 94.70 | 74.05 | 44.19 |
| Random Forest | 87.69 | 79.77 | 51.40 |
| Gradient Boosting | 83.19 | 73.15 | 31.28 |
| AdaBoost | 76.87 | 66.54 | 36.87 |
| Neural Network | 60.23 | 50.98 | 26.82 |

***FIGURE 23: The Performance of The Best Model at Training***



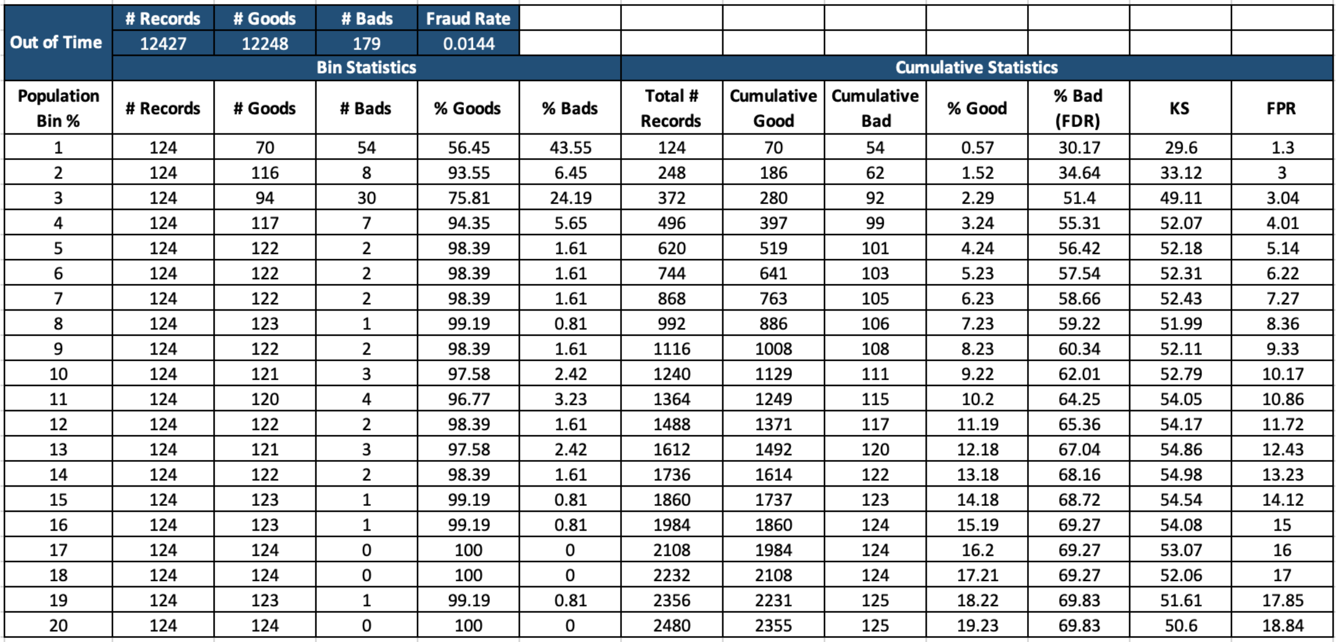
| Training | # Records | # Goods | # Bads | Fraud Rate | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 54018 | 53417 | 601 | 0.0111 | | | | | | | | |
| | Bin Statistics | | | | | Cumulative Statistics | | | | | | |
| Population Bin % | # Records | # Goods | # Bads | % Goods | % Bads | Total # Records | Cumulative Good | Cumulative Bad | % Good | % Bad (FDR) | KS | FPR |
| 1 | 540 | 118 | 422 | 21.85 | 78.15 | 540 | 118 | 422 | 0.22 | 70.22 | 70 | 0.28 |
| 2 | 540 | 454 | 86 | 84.07 | 15.93 | 1080 | 572 | 508 | 1.07 | 84.53 | 83.46 | 1.13 |
| 3 | 540 | 517 | 23 | 95.74 | 4.26 | 1620 | 1089 | 531 | 2.04 | 88.35 | 86.31 | 2.05 |
| 4 | 540 | 531 | 9 | 98.33 | 1.67 | 2160 | 1620 | 540 | 3.03 | 89.85 | 86.82 | 3 |
| 5 | 540 | 533 | 7 | 98.7 | 1.3 | 2700 | 2153 | 547 | 4.03 | 91.01 | 86.98 | 3.94 |
| 6 | 540 | 535 | 5 | 99.07 | 0.93 | 3240 | 2688 | 552 | 5.03 | 91.85 | 86.82 | 4.87 |
| 7 | 540 | 535 | 5 | 99.07 | 0.93 | 3780 | 3223 | 557 | 6.03 | 92.68 | 86.65 | 5.79 |
| 8 | 540 | 535 | 5 | 99.07 | 0.93 | 4320 | 3758 | 562 | 7.04 | 93.51 | 86.47 | 6.69 |
| 9 | 540 | 538 | 2 | 99.63 | 0.37 | 4860 | 4296 | 564 | 8.04 | 93.84 | 85.8 | 7.62 |
| 10 | 540 | 538 | 2 | 99.63 | 0.37 | 5400 | 4834 | 566 | 9.05 | 94.18 | 85.13 | 8.54 |
| 11 | 540 | 537 | 3 | 99.44 | 0.56 | 5940 | 5371 | 569 | 10.05 | 94.68 | 84.63 | 9.44 |
| 12 | 540 | 532 | 8 | 98.52 | 1.48 | 6480 | 5903 | 577 | 11.05 | 96.01 | 84.96 | 10.23 |
| 13 | 540 | 536 | 4 | 99.26 | 0.74 | 7020 | 6439 | 581 | 12.05 | 96.67 | 84.62 | 11.08 |
| 14 | 540 | 537 | 3 | 99.44 | 0.56 | 7560 | 6976 | 584 | 13.06 | 97.17 | 84.11 | 11.95 |
| 15 | 540 | 540 | 0 | 100 | 0 | 8100 | 7516 | 584 | 14.07 | 97.17 | 83.1 | 12.87 |
| 16 | 540 | 539 | 1 | 99.81 | 0.19 | 8640 | 8055 | 585 | 15.08 | 97.34 | 82.26 | 13.77 |
| 17 | 540 | 539 | 1 | 99.81 | 0.19 | 9180 | 8594 | 586 | 16.09 | 97.5 | 81.41 | 14.67 |
| 18 | 540 | 540 | 0 | 100 | 0 | 9720 | 9134 | 586 | 17.1 | 97.5 | 80.4 | 15.59 |
| 19 | 540 | 538 | 2 | 99.63 | 0.37 | 10260 | 9672 | 588 | 18.11 | 97.84 | 79.73 | 16.45 |
| 20 | 540 | 539 | 1 | 99.81 | 0.19 | 10800 | 10211 | 589 | 19.12 | 98 | 78.88 | 17.34 |

### *FIGURE 24: The Performance of The Best Model at Testing*



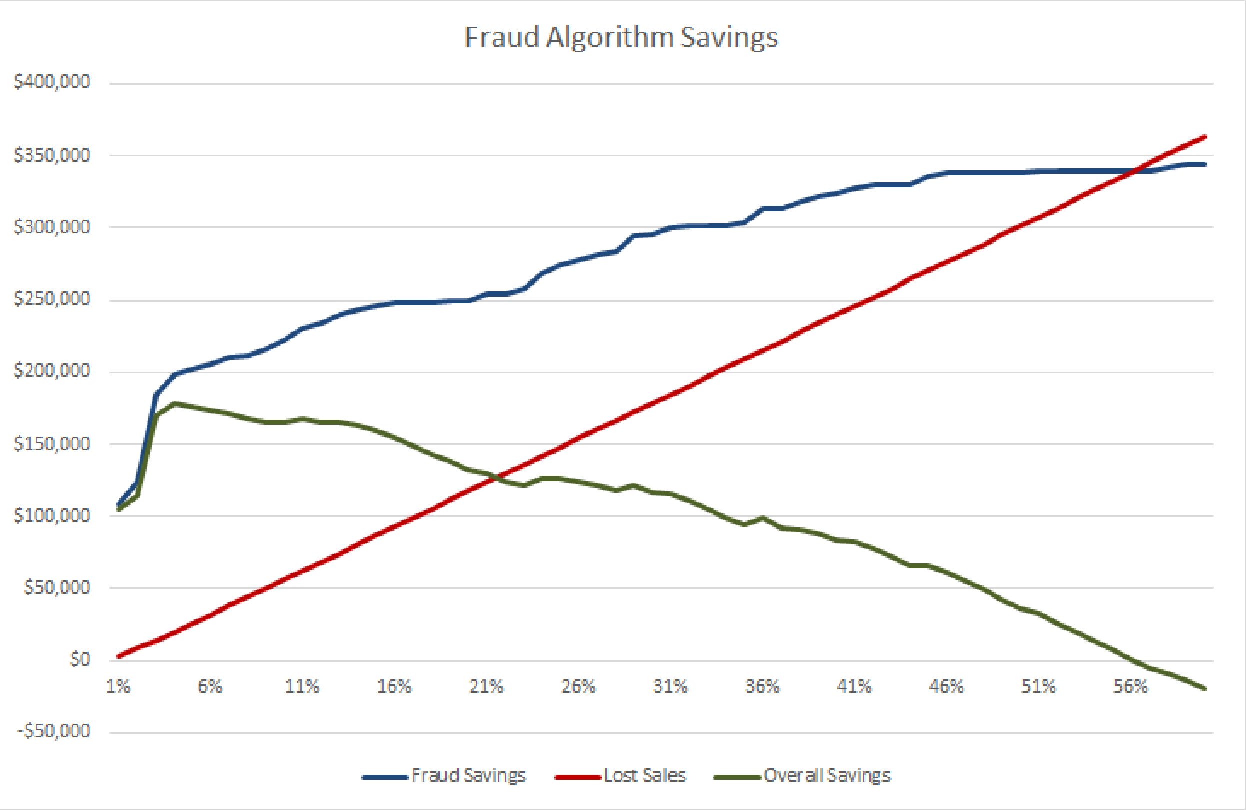
| Testing | # Records | # Goods | # Bads | Fraud Rate | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 23151 | 22894 | 257 | 0.0111 | | | | | | | | |
| | Bin Statistics | | | | | Cumulative Statistics | | | | | | |
| Population Bin % | # Records | # Goods | # Bads | % Goods | % Bads | Total # Records | Cumulative Good | Cumulative Bad | % Good | % Bad (FDR) | KS | FPR |
| 1 | 232 | 77 | 155 | 33.19 | 66.81 | 232 | 77 | 155 | 0.34 | 60.31 | 59.97 | 0.5 |
| 2 | 232 | 199 | 33 | 85.78 | 14.22 | 464 | 276 | 188 | 1.21 | 73.15 | 71.94 | 1.47 |
| 3 | 232 | 220 | 12 | 94.83 | 5.17 | 696 | 496 | 200 | 2.17 | 77.82 | 75.65 | 2.48 |
| 4 | 232 | 222 | 10 | 95.69 | 4.31 | 928 | 718 | 210 | 3.14 | 81.71 | 78.57 | 3.42 |
| 5 | 232 | 226 | 6 | 97.41 | 2.59 | 1160 | 944 | 216 | 4.12 | 84.05 | 79.93 | 4.37 |
| 6 | 232 | 225 | 7 | 96.98 | 3.02 | 1392 | 1169 | 223 | 5.11 | 86.77 | 81.66 | 5.24 |
| 7 | 232 | 231 | 1 | 99.57 | 0.43 | 1624 | 1400 | 224 | 6.12 | 87.16 | 81.04 | 6.25 |
| 8 | 232 | 230 | 2 | 99.14 | 0.86 | 1856 | 1630 | 226 | 7.12 | 87.94 | 80.82 | 7.21 |
| 9 | 232 | 229 | 3 | 98.71 | 1.29 | 2088 | 1859 | 229 | 8.12 | 89.11 | 80.99 | 8.12 |
| 10 | 232 | 229 | 3 | 98.71 | 1.29 | 2320 | 2088 | 232 | 9.12 | 90.27 | 81.15 | 9 |
| 11 | 232 | 231 | 1 | 99.57 | 0.43 | 2552 | 2319 | 233 | 10.13 | 90.66 | 80.53 | 9.95 |
| 12 | 232 | 232 | 0 | 100 | 0 | 2784 | 2551 | 233 | 11.14 | 90.66 | 79.52 | 10.95 |
| 13 | 232 | 232 | 0 | 100 | 0 | 3016 | 2783 | 233 | 12.16 | 90.66 | 78.5 | 11.94 |
| 14 | 232 | 227 | 5 | 97.84 | 2.16 | 3248 | 3010 | 238 | 13.15 | 92.61 | 79.46 | 12.65 |
| 15 | 232 | 229 | 3 | 98.71 | 1.29 | 3480 | 3239 | 241 | 14.15 | 93.77 | 79.62 | 13.44 |
| 16 | 232 | 231 | 1 | 99.57 | 0.43 | 3712 | 3470 | 242 | 15.16 | 94.16 | 79 | 14.34 |
| 17 | 232 | 231 | 1 | 99.57 | 0.43 | 3944 | 3701 | 243 | 16.17 | 94.55 | 78.38 | 15.23 |
| 18 | 232 | 230 | 2 | 99.14 | 0.86 | 4176 | 3931 | 245 | 17.17 | 95.33 | 78.16 | 16.04 |
| 19 | 232 | 231 | 1 | 99.57 | 0.43 | 4408 | 4162 | 246 | 18.18 | 95.72 | 77.54 | 16.92 |
| 20 | 232 | 230 | 2 | 99.14 | 0.86 | 4640 | 4392 | 248 | 19.18 | 96.5 | 77.32 | 17.71 |

### *FIGURE 25: The Performance of The Best Model at Out of Time*



| Out of Time | # Records | # Goods | # Bads | Fraud Rate | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 12427 | 12248 | 179 | 0.0144 | | | | | | | | |
| | Bin Statistics | | | | | Cumulative Statistics | | | | | | |
| Population Bin % | # Records | # Goods | # Bads | % Goods | % Bads | Total # Records | Cumulative Good | Cumulative Bad | % Good | % Bad (FDR) | KS | FPR |
| 1 | 124 | 70 | 54 | 56.45 | 43.55 | 124 | 70 | 54 | 0.57 | 30.17 | 29.6 | 1.3 |
| 2 | 124 | 116 | 8 | 93.55 | 6.45 | 248 | 186 | 62 | 1.52 | 34.64 | 33.12 | 3 |
| 3 | 124 | 94 | 30 | 75.81 | 24.19 | 372 | 280 | 92 | 2.29 | 51.4 | 49.11 | 3.04 |
| 4 | 124 | 117 | 7 | 94.35 | 5.65 | 496 | 397 | 99 | 3.24 | 55.31 | 52.07 | 4.01 |
| 5 | 124 | 122 | 2 | 98.39 | 1.61 | 620 | 519 | 101 | 4.24 | 56.42 | 52.18 | 5.14 |
| 6 | 124 | 122 | 2 | 98.39 | 1.61 | 744 | 641 | 103 | 5.23 | 57.54 | 52.31 | 6.22 |
| 7 | 124 | 122 | 2 | 98.39 | 1.61 | 868 | 763 | 105 | 6.23 | 58.66 | 52.43 | 7.27 |
| 8 | 124 | 123 | 1 | 99.19 | 0.81 | 992 | 886 | 106 | 7.23 | 59.22 | 51.99 | 8.36 |
| 9 | 124 | 122 | 2 | 98.39 | 1.61 | 1116 | 1008 | 108 | 8.23 | 60.34 | 52.11 | 9.33 |
| 10 | 124 | 121 | 3 | 97.58 | 2.42 | 1240 | 1129 | 111 | 9.22 | 62.01 | 52.79 | 10.17 |
| 11 | 124 | 120 | 4 | 96.77 | 3.23 | 1364 | 1249 | 115 | 10.2 | 64.25 | 54.05 | 10.86 |
| 12 | 124 | 122 | 2 | 98.39 | 1.61 | 1488 | 1371 | 117 | 11.19 | 65.36 | 54.17 | 11.72 |
| 13 | 124 | 121 | 3 | 97.58 | 2.42 | 1612 | 1492 | 120 | 12.18 | 67.04 | 54.86 | 12.43 |
| 14 | 124 | 122 | 2 | 98.39 | 1.61 | 1736 | 1614 | 122 | 13.18 | 68.16 | 54.98 | 13.23 |
| 15 | 124 | 123 | 1 | 99.19 | 0.81 | 1860 | 1737 | 123 | 14.18 | 68.72 | 54.54 | 14.12 |
| 16 | 124 | 123 | 1 | 99.19 | 0.81 | 1984 | 1860 | 124 | 15.19 | 69.27 | 54.08 | 15 |
| 17 | 124 | 124 | 0 | 100 | 0 | 2108 | 1984 | 124 | 16.2 | 69.27 | 53.07 | 16 |
| 18 | 124 | 124 | 0 | 100 | 0 | 2232 | 2108 | 124 | 17.21 | 69.27 | 52.06 | 17 |
| 19 | 124 | 123 | 1 | 99.19 | 0.81 | 2356 | 2231 | 125 | 18.22 | 69.83 | 51.61 | 17.85 |
| 20 | 124 | 124 | 0 | 100 | 0 | 2480 | 2355 | 125 | 19.23 | 69.83 | 50.6 | 18.84 |

***FIGURE 26: Fraud Saving Plot***



Assuming a $2,000 gain for every fraud that's caught and a $50 loss for every false positive, the overall savings becomes negative when the cutoff point goes higher than 56%. Considering all business goals and risk taste, the client should choose a cutoff point below or equal 56%. We recommend a cutoff point at 4% where the overall savings is maximized at approximately $178K.

# Conclusions

In this project, we investigated nearly 100,000 credit card transactions made by employees of a Tennessee government agency throughout 2010, each with a label indicating whether or not the transaction was fraudulent. Using variables generated from this data and selected for predictive significance, we developed a model that can accept future transaction data and determine whether or not the transaction is likely fraudulent, with a degree of accuracy high enough to both ensure a large percentage of total fraud will be caught, while also not declining an overly large number of legitimate transactions.

Our final model used a random forest supervised machine learning method to predict likely fraudulent transactions. Using this model, we were able to predict 51.4% of the total out of time frauds at 3% of the total transactions. Although our second strongest model built using decision trees had a stronger training prediction rate (94.7% vs. 87.7%), there are signs of overfitting when comparing its FDR rates at 3% of the total transactions between the training and testing data sets, and the out of time prediction rate was lower. For these reasons, and we feel that the random forest model is likely to have the strongest predictive accuracy when used to detect future frauds. Based on our OOT prediction, we recommend a cutoff of 4% of the total credit card transactions which was estimated to save approximately $178K over the period of the OOT data.

In order to further improve our model, there are several adjustments we could make as part of the future research initiative. First, separating the transactions into clusters and then building independent models for each cluster should result in higher overall FDR since the characteristics of fraud by segment are likely to be quite different. Second, expanding the timeframe of data for all sets, training, testing, and particularly out of time would allow us to better identify models that will be strong predictors of future fraud. This would also give us a more definitive answer on whether the random forest driven approach did indeed produce the strongest model.

Finally, working with the project stakeholders to develop an in-depth understanding of the implications of declaring a transaction fraudulent in this context could help us build a model that better achieves those goals. For example, because fraud on these accounts involves many government entities, it is possible that detecting yet not declining potentially fraudulent transactions in order to prosecute fraudulent actors is more desirable than merely preventing the fraud. In that case, the risk of false positives would be low and our team could take a higher acceptable false positive rate into consideration when building the strongest model.

## Appendix

# Data Quality Report

# Table of Contents

# 1. Introduction

This Data Quality Report ("DQR" or "Report") is a review of credit card transaction data ("Transaction Data") provided by Professor Stephen Coggeshall ("Professor") for the University of Southern California course, DSO 562: Fraud Analytics. Although the origin and time of the data is kept unknown for anonymity, the data represents credit card transactions within calendar year 2010 (January 1, 2010 to December 31, 2010).

### A. Purpose of Collecting Credit Card Transaction Data

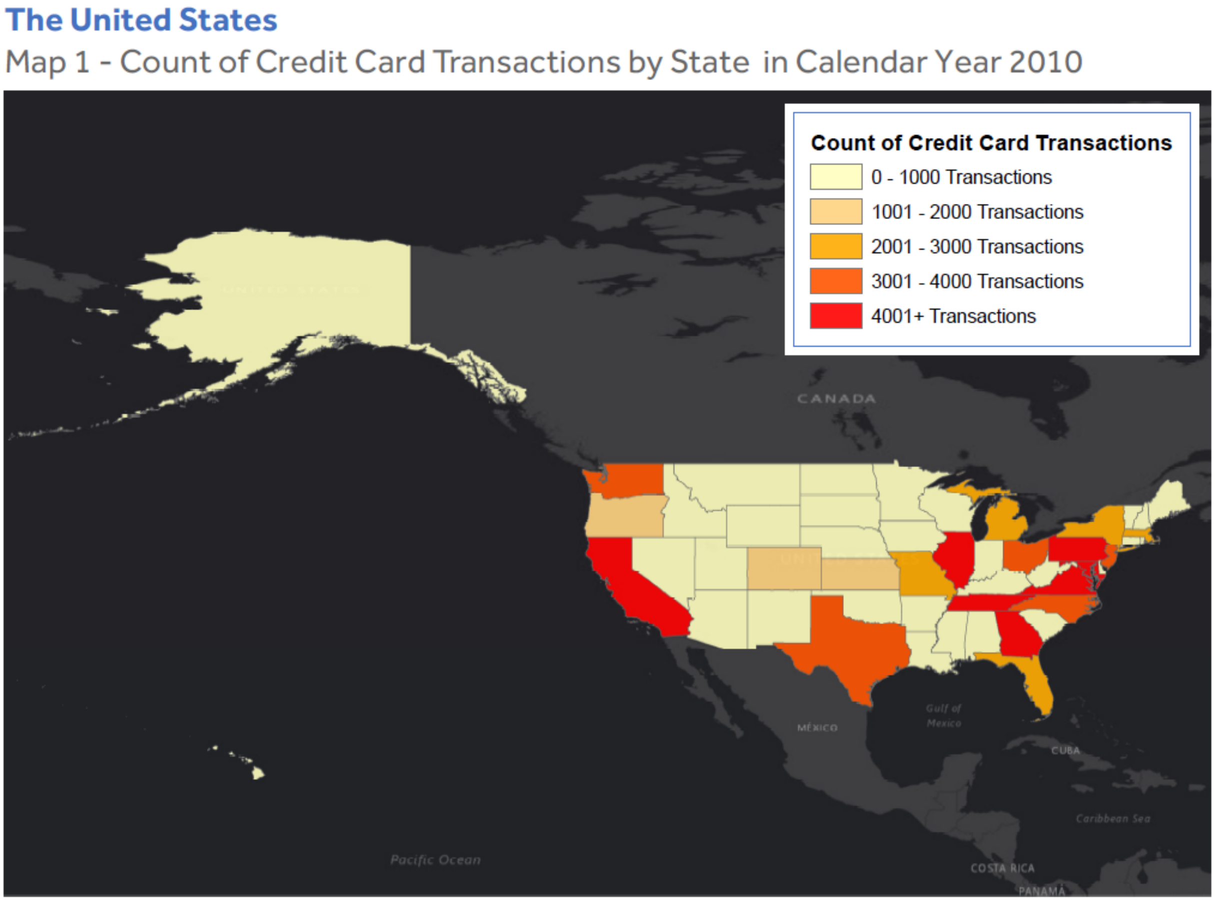
Transaction Data is financial data collected by credit card companies and acts as a record of the transfer of funds for the purchase of goods or services. By recording actual purchases, Transaction Data is useful for analyzing and predicting consumer behavior. This can lead to myriad of uses such as analyzing an individual's financial risk or improve an advertising/marketing campaign.

However, since the Transaction Data relies on credit card purchases, it is subject to inaccuracies due to instances of fraud. Credit card companies can use its Transaction Data to identify potential cases of fraud and reduce these instances fraudulent transactions.

### B. Data Summary

The Transaction Data is a record of credit card transactions within the United States and was provided by the Professor for the purpose of identifying fraud using supervised learning. It is a record 96,753 transactions within calendar year 2010 and contains ten (10) fields. Of which, only one (1) is numeric. The remaining nine (9) fields are categorical. Please see **Map 1** on the following page for the geographical merchant concentration of the data and **Exhibit 1** for the definitions of each field. Due to limits based on the understanding of the fields and the quality of the entries, **Map 1** has been drawn based on the existing data as is prior to cleaning. For more information regarding the missing data, please see **Exhibit 2** more information regarding missing and other merchant state information.

***Map 1: Count of Credit Card Transactions by State***



# 2. Field Summary

As mentioned in Section 1, there are ten (10) fields within the Transaction Data.

### A. Aggregate Summary

Below is **Table 1**, showing a summary of each field's field type, number of records, percent populated, count of unique values, count of records with the value zero (0), and the percentage with the value zero (0).

***Table 1: Aggregate Summary***

| Field | Field Type | Records | % Populated | Unique Values | Records with value zero | % with value zero |
|---|---|---|---|---|---|---|
| Recnum | Categorical | 96,753 | 100.00% | 96,753 | 0 | 0.00% |
| Cardnum | Categorical | 96,753 | 100.00% | 1,645 | 0 | 0.00% |
| Date | Categorical | 96,753 | 100.00% | 365 | 0 | 0.00% |
| Merchnum | Categorical | 93,378 | 96.51% | 13,091[1] | 231 | 0.24% |
| Merch Description | Categorical | 96,753 | 100.00% | 13,126 | 3[2] | 0.00% |
| Merch State | Categorical | 95,558 | 98.76% | 227 | 0 | 0.00% |
| Merch Zip | Categorical | 92,097 | 95.19% | 4,567 | 0 | 0.00% |
| Transtype | Categorical | 96,753 | 100.00% | 4 | 0 | 0.00% |
| Amount | Numeric | 96,753 | 100.00% | 34,909 | 0 | 0.00% |
| Fraud | Categorical | 96,753 | 100.00% | 2 | 95,694[3] | 98.91% |

[1] It is assumed that entries with no characters other than spaces (' ') are considered blank

[2] It is assumed that a Merch Description of '0000000000000000000000' is equivalent to zero (0)

[3] Although 95,694 tuples contain zero (0) in the Fraud field, a zero (0) indicates that the transaction was not an instance of fraud

### B. Numeric Field Summary

As stated in Section 1 and shown in **Table 1**, there is only one (1) numeric field in the Transaction Data, the Amount. Below is **Table 2**, showing a summary of the Amount's mean, median, mode, min, max, and standard deviation.

***Table 2: Numeric Summary***

| Field | Records | Mean | Median | Mode | Min | Max | Standard Deviation |
|---|---|---|---|---|---|---|---|
| Amount | 96,753 | $427.89 | $137.98 | $3.62 | $0.01 | $3,102,045.53 | $10,006.14 |

### C. Categorical Field Summary

As stated in Section 1 and shown in **Table 1**, there are nine (9) categorical fields in the Transaction Data. Below is **Table 3**, showing a summary of the categorical fields' most common entry its percentage of the total records.

***Table 3: Categorical Summary***

| Field | Records | Unique Values | Most Common Entry | Count of Entry | Percentage of Total |
|---|---|---|---|---|---|
| Recnum | 96,753 | 96,753 | N/A[1] | N/A[1] | N/A[1] |
| Cardnum | 96,753 | 1,645 | 5142148452 | 1,192 | 1.23% |
| Date | 96,753 | 365 | 2/28/2010 | 684 | 0.71% |
| Merchnum | 93,378 | 13,091 | 930090121224 | 9,310 | 9.62% |
| Merch Description | 96,753 | 13,126 | GSA-FSS-ADV | 1,688 | 1.74% |
| Merch State | 95,558 | 227 | TN | 12,035 | 12.44% |
| Merch Zip | 92,097 | 4,567 | 38118 | 11,868 | 12.27% |
| Transtype | 96,753 | 4 | P | 96,398 | 99.63% |
| Fraud | 96,753 | 2 | 0 | 95,694 | 98.91% |

[1] Since every entry is a unique entry, common entry and its associated fields are not applicable.

# 3. Field Details

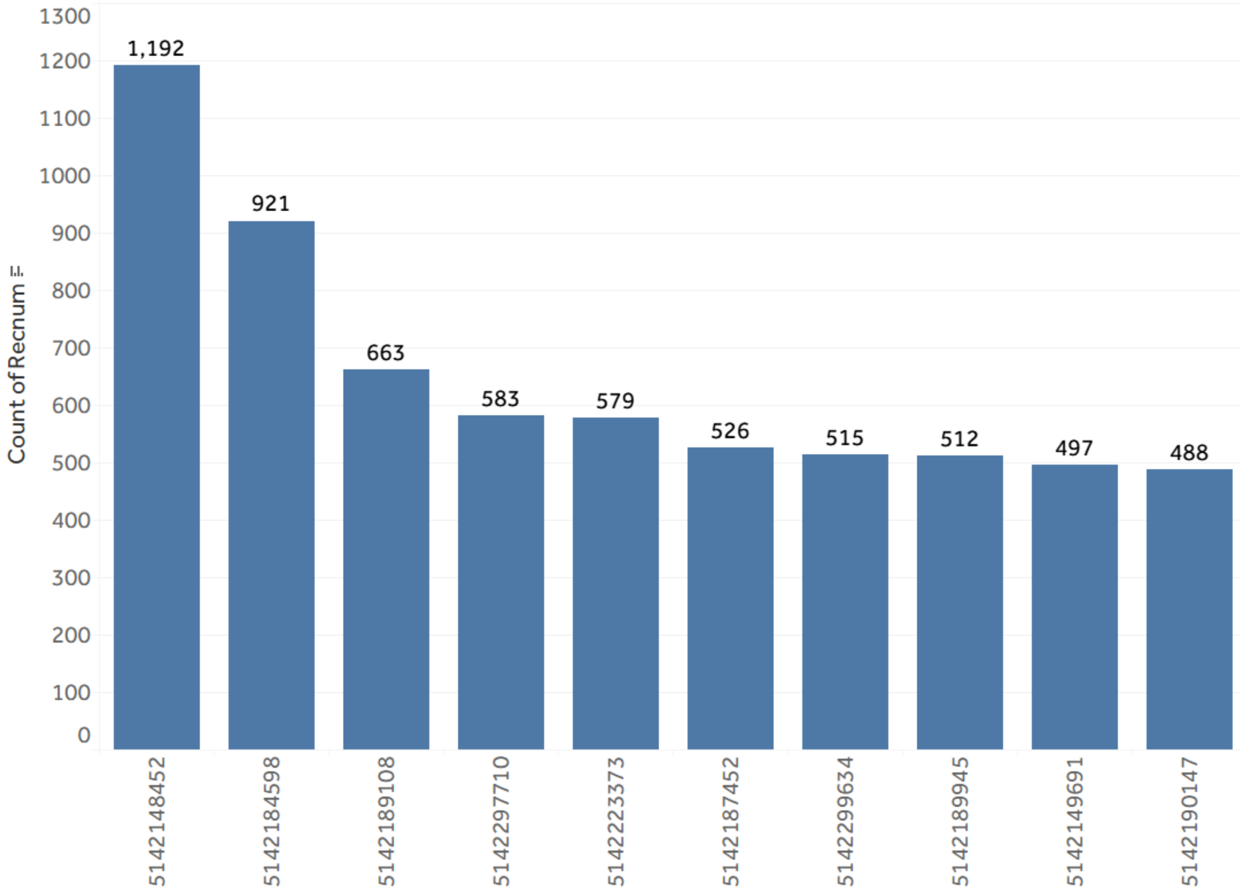
The Transaction Data has ten (10) different fields. Below is a short description of every field including additional information such as graphs, tables, and/or maps to provide greater insights into the data.

## 3.1. Recnum

Recnum is a nominal categorical field that provides a unique ordinal identifier for each tuple, which translates to 96,753 unique records. Since each tuple is assigned a unique identifier, the field follows a uniform distribution. A graph or table would not provide any greater insight into the data and thus, was not included as part of the report.

## 3.2. Cardnum

Cardnum is a categorical field used to indicate the credit card number for each transaction. As shown on **Table 3**, there are 96,753 records and among those records, there are 1,192 unique credit card numbers. Interestingly, all values for Cardnum have 10 digits, which does not follow standard number of digits used by the four (4) major credit card networks (Visa, MasterCard, American Express, and Discover). Thus, it may be likely that the original credit card numbers were modified to create anonymity while still maintaining its use as an identifier for a specific credit card user. For more information on the structure of credit card numbers, please see **Exhibit 3**. **Graph 1** below shows the top ten (10) credit card numbers with the most transactions.

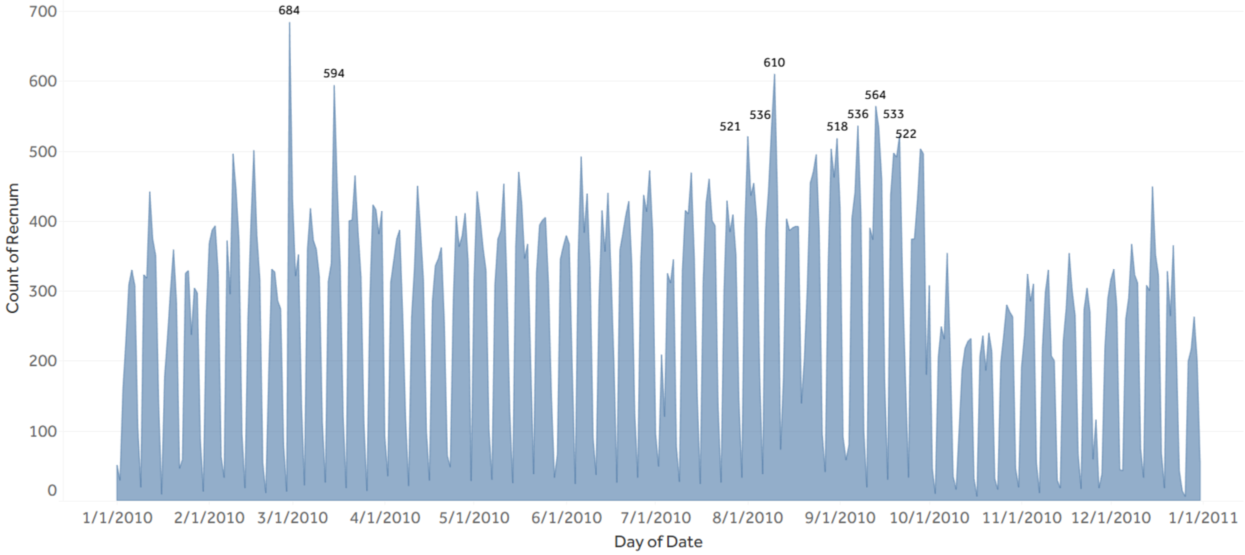


***Graph 1: Top 10 Cardnum with the Most Transactions***

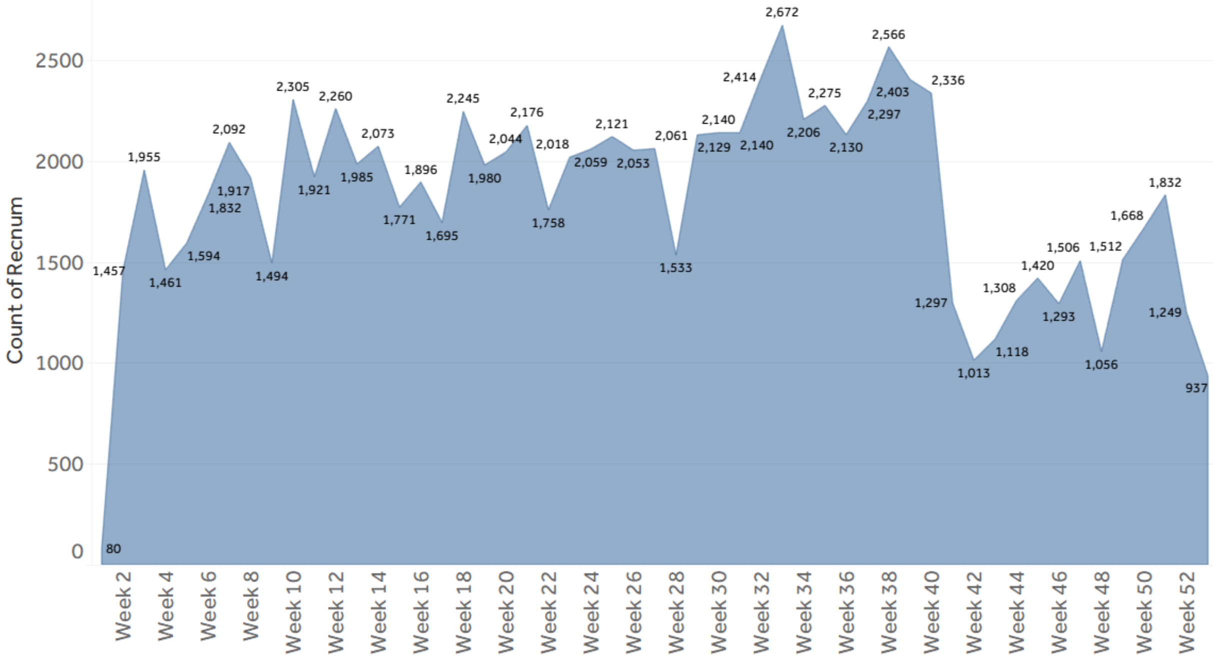
Based on **Graph 1**, Cardnum 5142148452 has the highest number of transactions at 1,192.

## 3.3. Date

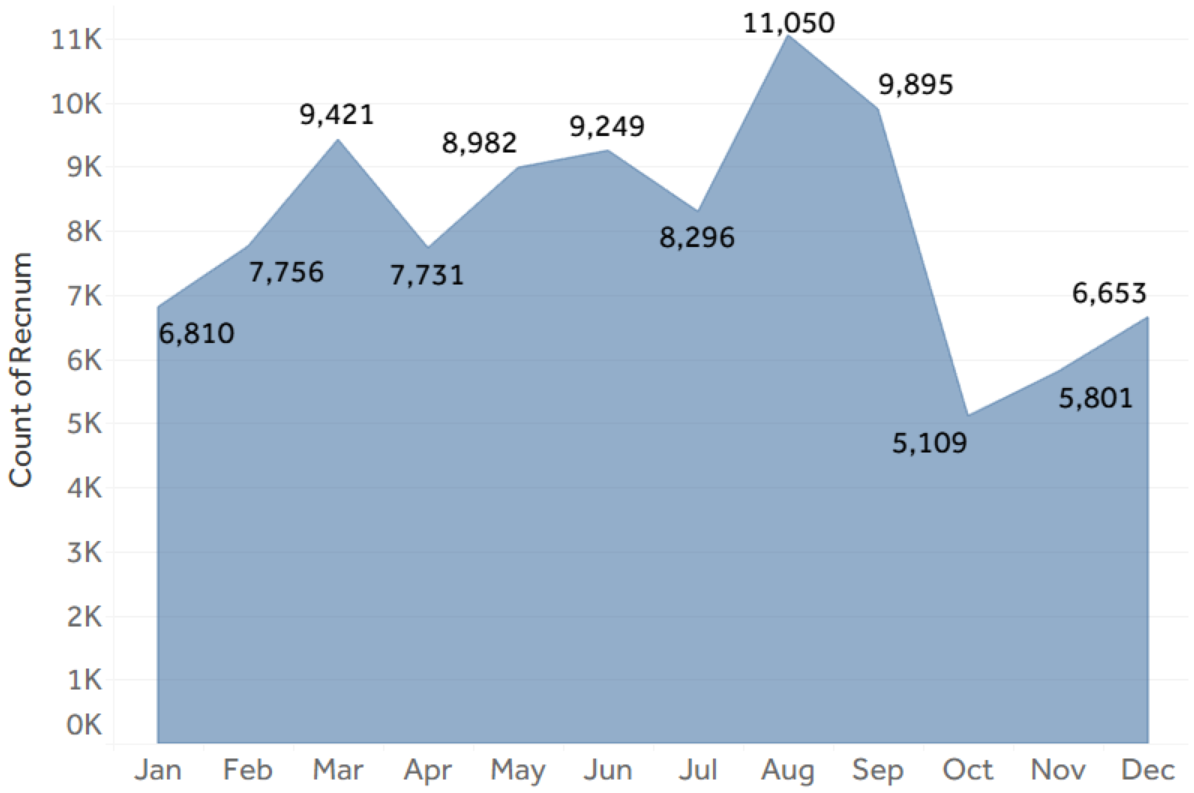
Date is a categorical field used to identify the month, day, and year a transaction took place in the format of mm/dd/yyyy. As shown on **Table 1**, there are 96,753 records across 365 days within the calendar year of 2010. **Graphs 2 through 5** provide more information on transaction trends in 2010.



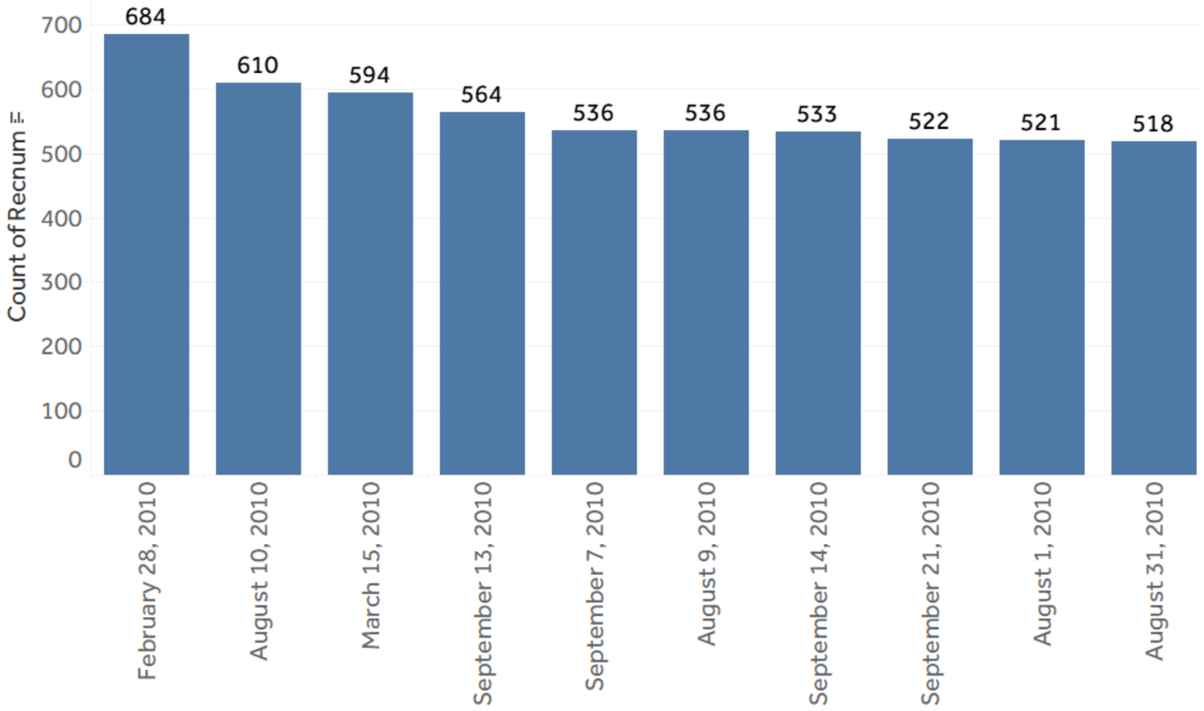
***Graph 2: Total Transaction Count by Day in 2010***



***Graph 3: Total Transaction Count by Week in 2010***



***Graph 4: Total Transaction Count by Month in 2010***



***Graph 5: Top 10 Days for Credit Card Transactions in 2010***

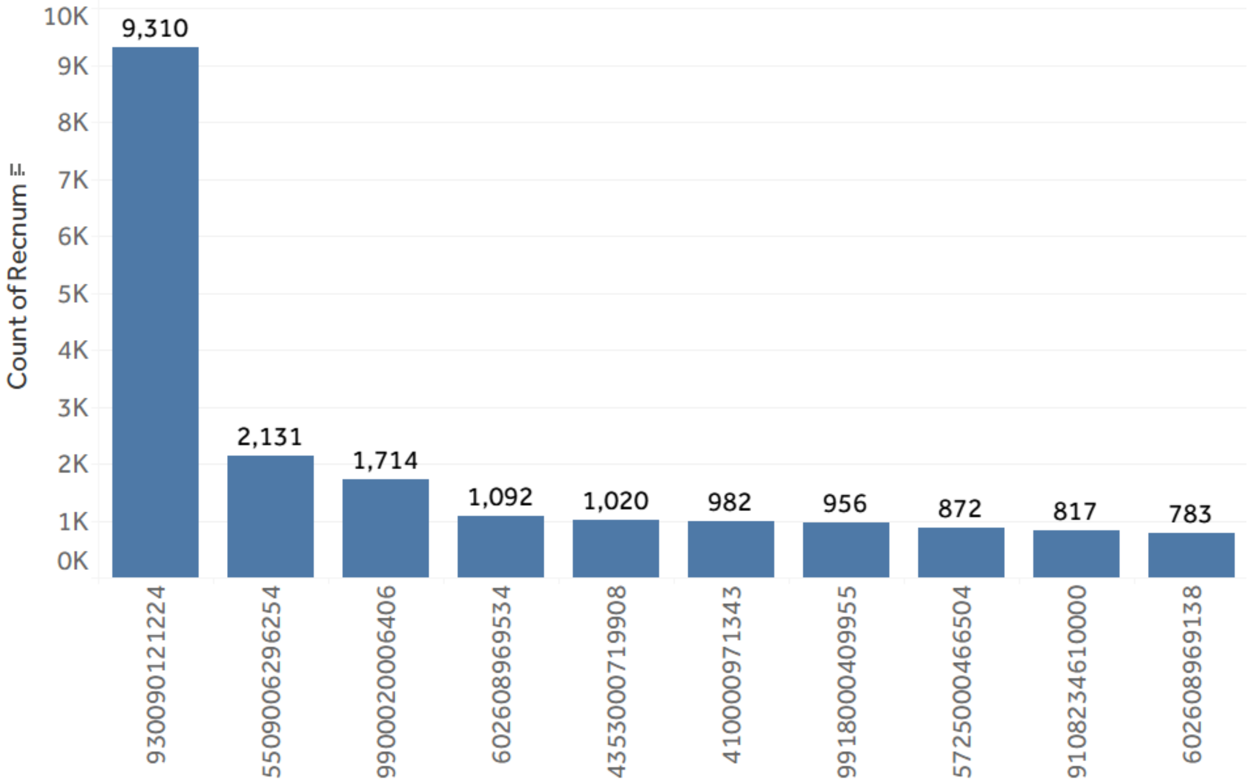
Based on **Graph 2**, transaction trends appear to follow a cyclical trend where there is a sharp decline in the number of transactions at the end of the week. One potential reason for this may be that individuals may tend to make more credit card transactions during the week to take care of their chores (which includes grocery shopping) and during the weekend they may have an increased propensity to stay home. However, this phenomenon would require further research to validify.

Based on **Graphs 3** and **4**, the last three months of 2010 experienced significantly lower quantities of credit card transactions. During the winter season, individuals may go out less and be less likely to go shopping. Furthermore, many of the credit transactions taking place on the Eastern portion of the United States (as shown in **Map 1** and discussed in **Section 3.F**). This may have a further impact on the number of credit card transactions as weather during the winter season may reduce the propensity for individuals to go out. However, this phenomenon would require further research to validify.

Lastly, based on **Graphs 2** and **5**, the top ten credit card transactions are clustered around the months of August and September. As mentioned above in the previous paragraph, this may be due to increased shopping in preparation for the winter period and/or holidays. Again, further research would be required to further validify this. Aside from the main cluster occurring in the months of August and September, February 28, 2010 was the day with the highest number of transactions.

## 3.4. Merchnum

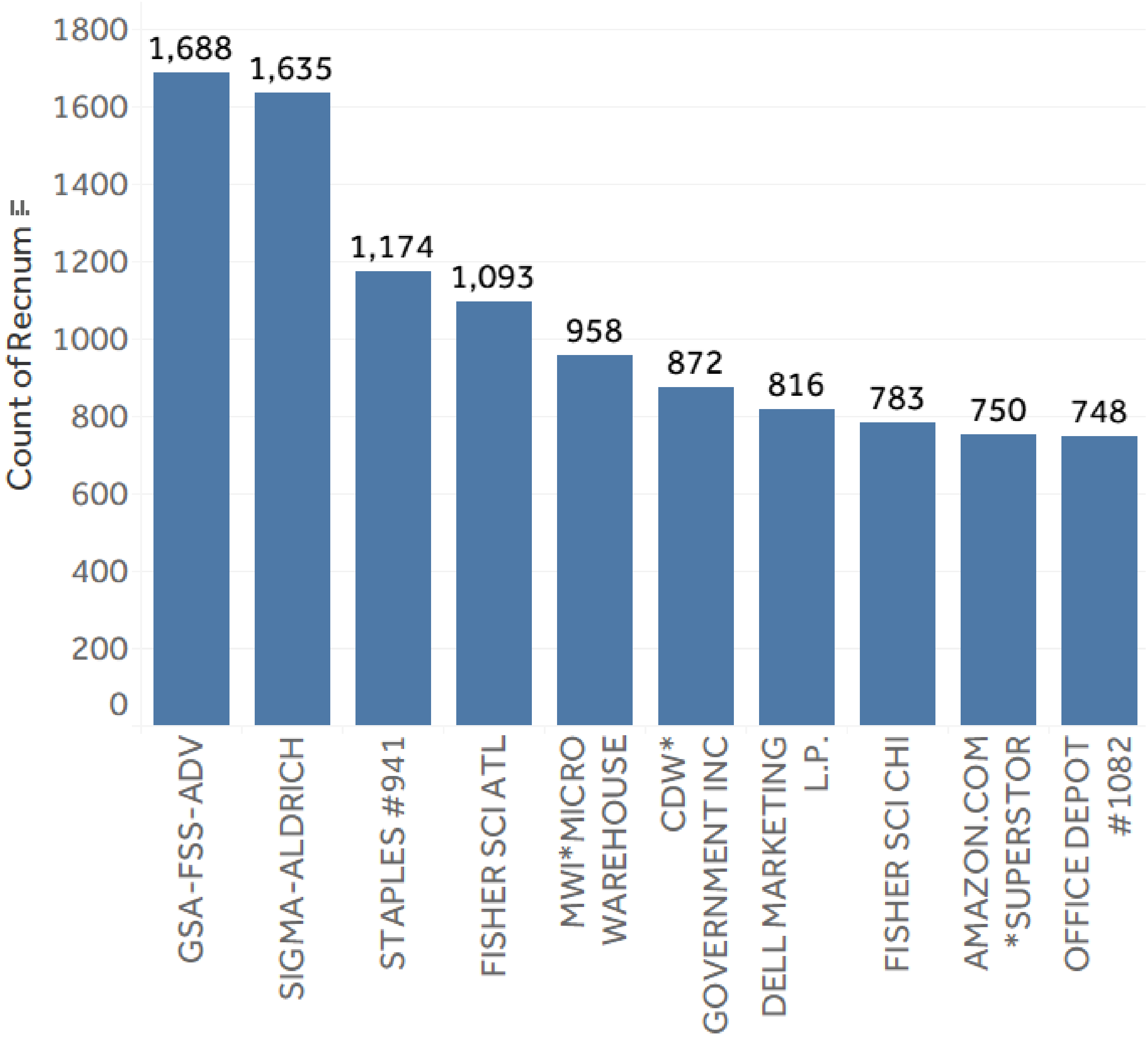
Merchnum is a categorical field that identifies the merchant that provided the good or service for the transaction. As shown on **Table 3**, there are 93,378 transactions that identified a merchant and across those 93,378 transactions, 13,092 unique merchants were identified. **Graph 3** below shows that the merchant with the Merchnum 930090121224 is the most common merchant with 9,310 transactions within the calendar year 2010.



*Graph 6: Top 10 Merchnums for Credit Card Transactions in 2010*

## 3.5 Merch Description

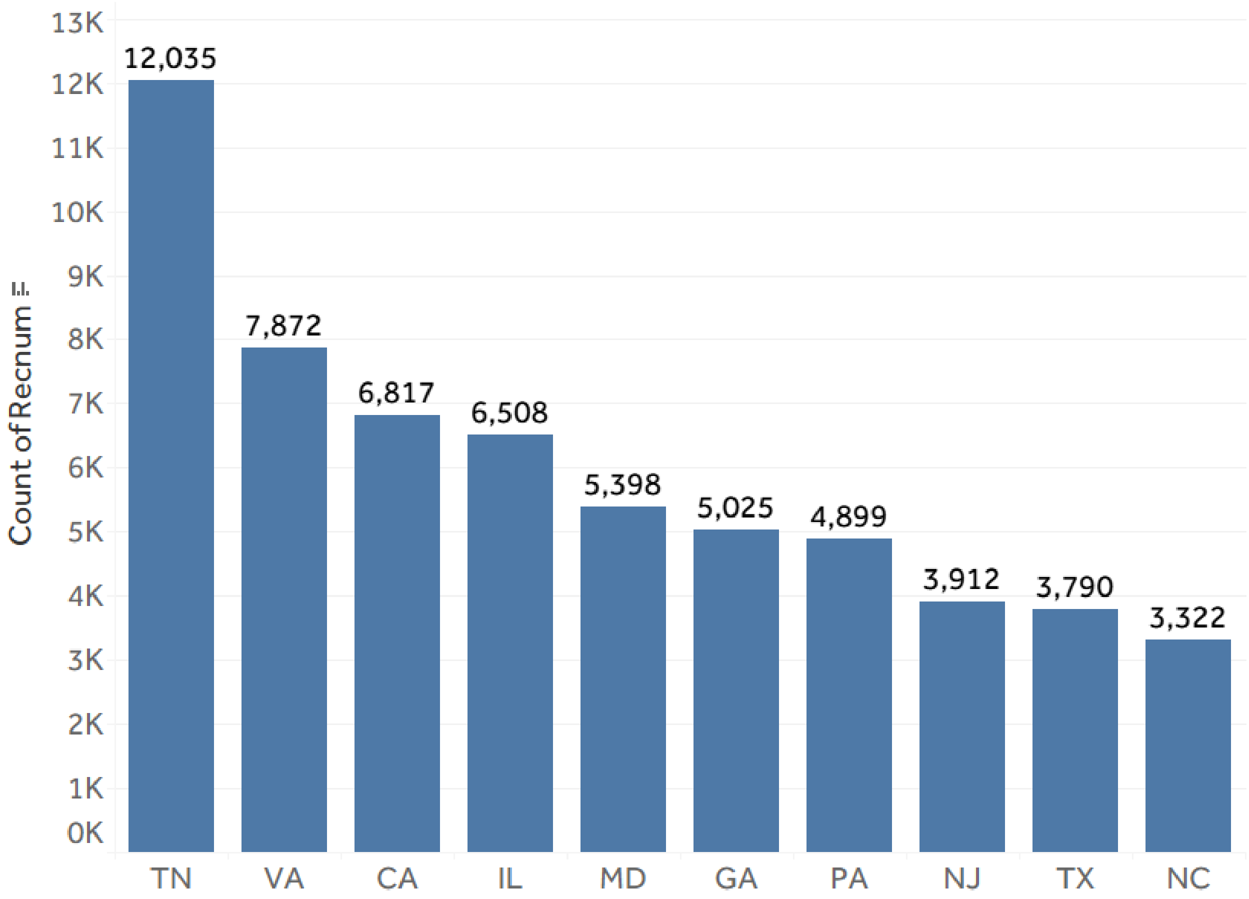
Merch Description is a categorical field that provides extra detail regarding the merchant and the transaction. As shown in **Table 1**, Transaction Data has 96,753 records for the Merch Description field with 13,126 unique records. Of the 13,126 unique records, GSA-FSS-ADV is the most common record. Below is **Graph 7** which shows the top ten (10) Merch Descriptions for credit card transactions in calendar year 2010.



***Graph 7: Top 10 Merch Descriptions for Credit Card Transactions in 2010***

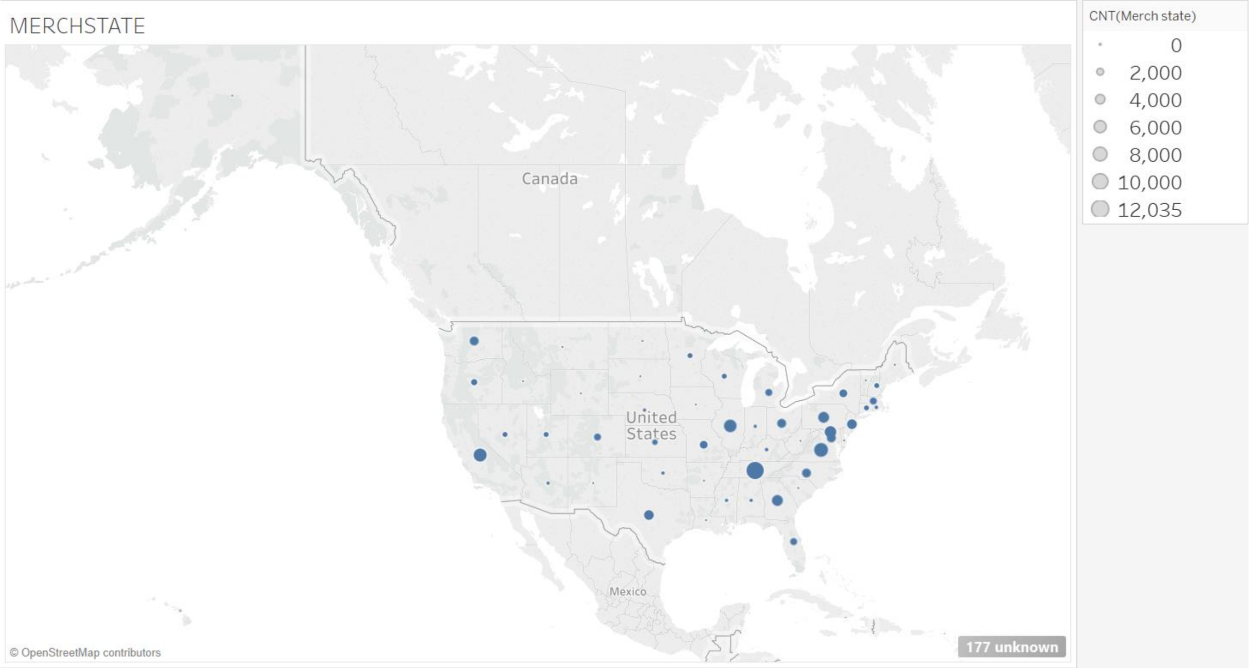
## 3.6 Merch State

Merch State is a categorical field that indicates the home state of the merchant providing the good or service for the transaction. As shown in **Table 1**, there are 95,558 records for Merch State with 227 unique records. **Map 1** on page 2 shows the concentration of credit card transactions in calendar year 2010 by state. Only 91,978 transactions were mapped (Approximately 95%). The remaining 4,775 transactions were not mapped due to limitations based on the data. Please see **Exhibit 2** for more information. Below is a **Graph 8** which shows the top ten (10) states with the most credit card transactions in calendar year 2010.



***Graph 8: Top 10 Merch States for Credit Card Transactions in 2010***

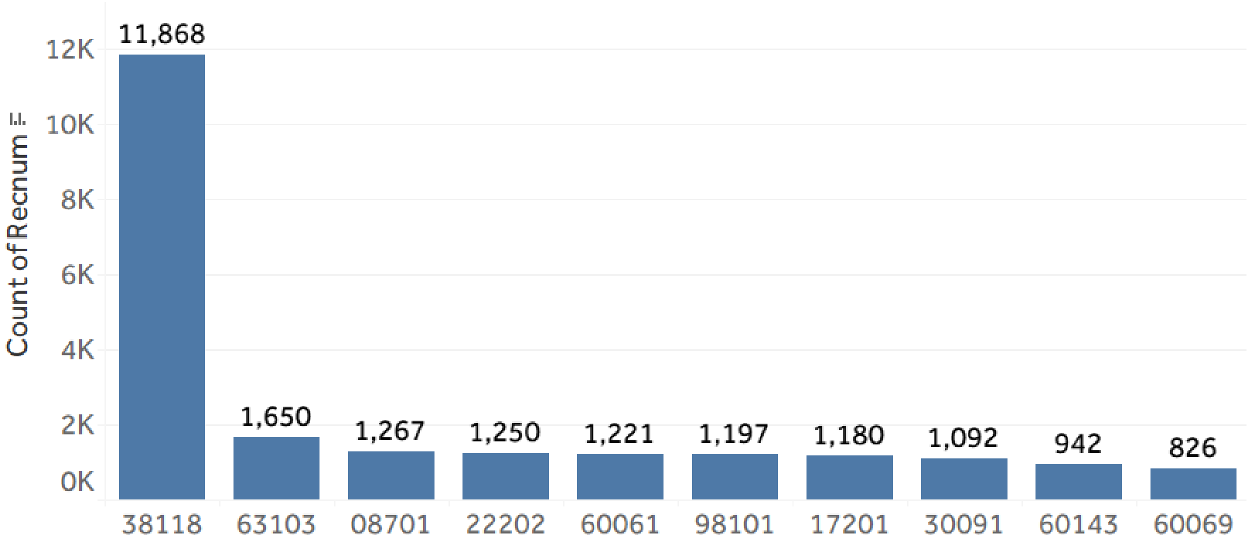
Based on **Map 1** it appears that there are a higher number of credit card transactions for coastal U.S. states. This might be likely because the coastal states may have a higher population able to have credit card transactions.



***Graph 9: Merchants Distribution***

## 3.7 Merch Zip

Merch Zip is a categorical field that identifies the ZIP code for the merchant providing the good or service in a transaction. As shown on **Table 1**, there are 92,097 records with ZIP codes and among them, there are 4,567 unique ZIP codes. **Graph 10** below shows the top ten (10) zips with the most credit card transactions in calendar year 2010.

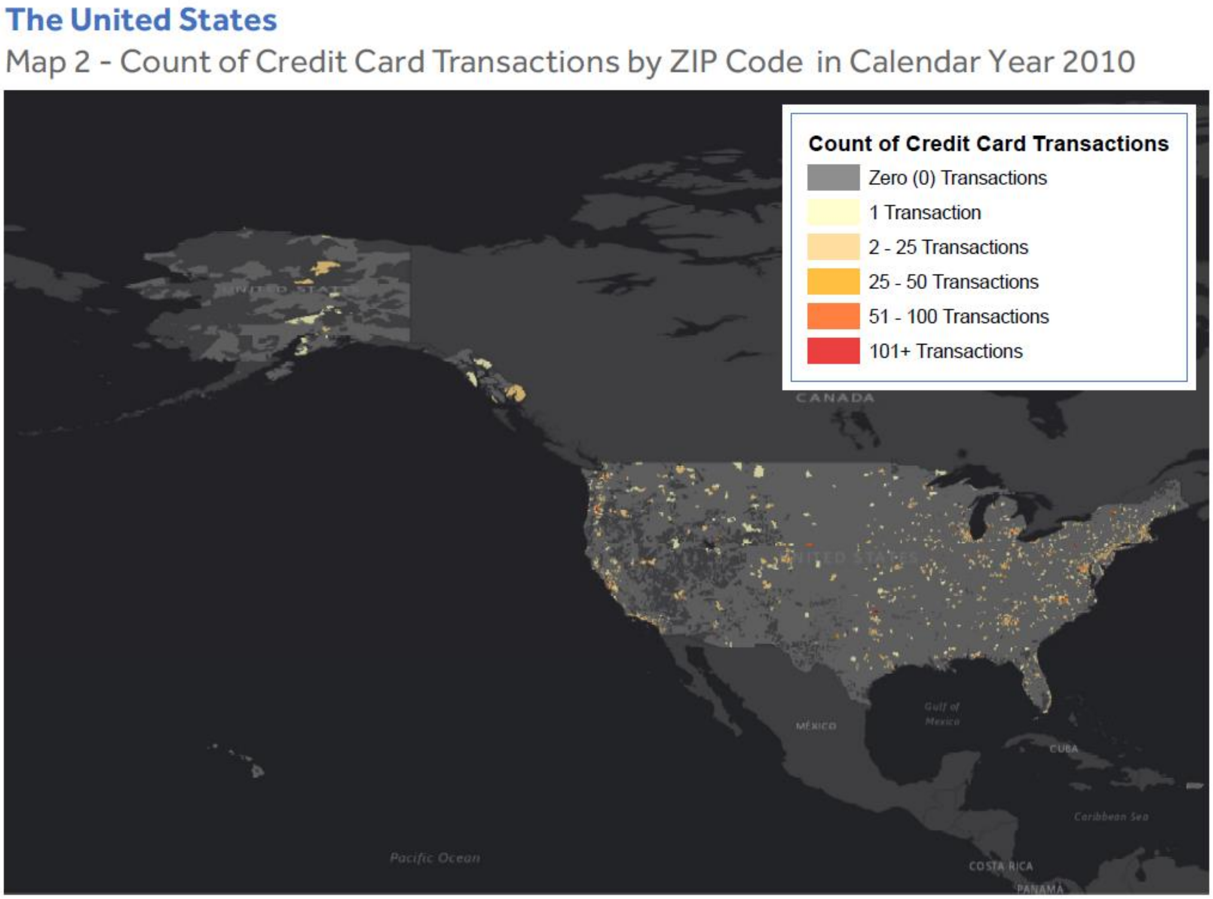


***Graph 10: Top 10 Merch Zips for Credit Card Transactions in 2010***

Like Merch State, Merch Zip was mapped to identify where the highest number of credit card transactions were by ZIP code. Please see **Map 2** on the following page for the map of the concentration of transactions by ZIP code. Similar to **Map 1**, **Map 2** shows that there are more ZIP codes with credit card transactions along the coast. Again, this may just be due to coastal regions being more likely to have a larger population for credit card transactions.

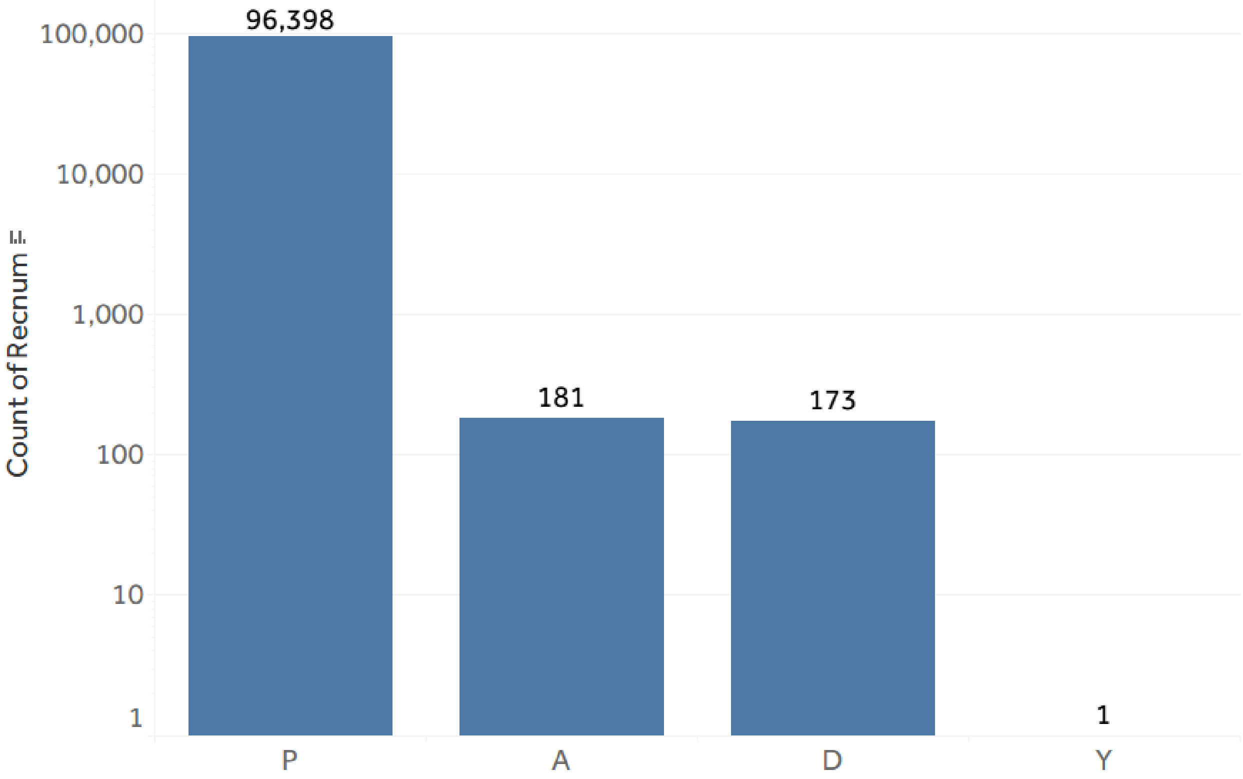
Not all ZIP codes were able to be mapped due to mapping limitations. 11,951 transactions were left unmatched. Please see **Exhibit 4** for more information regarding unmatched ZIP codes.

***Map 2: Count of Credit Card Transactions by Zip Code***



## 3.8 Transtype

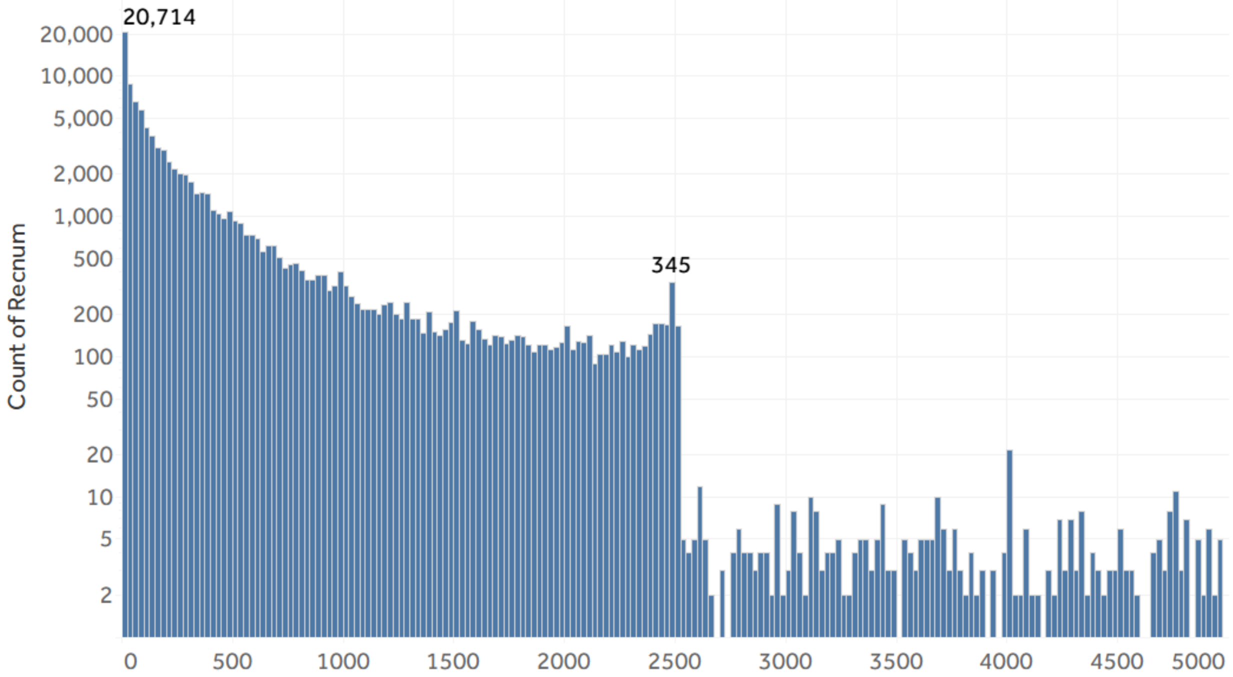
Transtype is a categorical field that identifies the type of credit card transaction. There are four types of transactions (A, D, P, and Y), but the actually meaning behind these labels are unknown. However, As identified in **Table 3**, P is the most common type of credit card transaction with 96,398 records out of 96,753 records being classified as P. **Graph 11**, below shows the total number of credit card transactions by Transtypes in calendar year 2010.



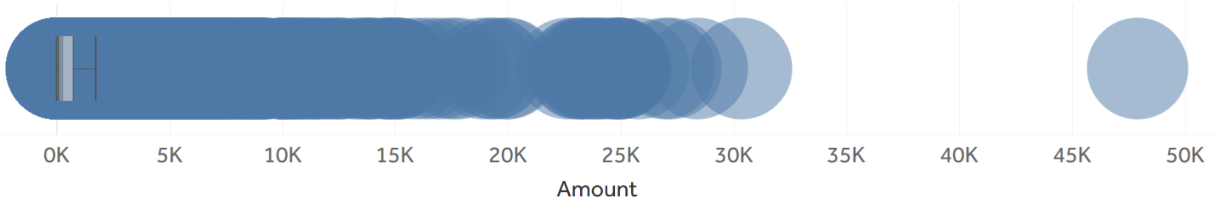
***Graph 11: Total Credit Card Transactions by Transtypes in 2010***

## 3.9 Amount

Amount is a numeric field that indicates the credit card transaction amount in dollars. **Graph 12** below shows a histogram of Amount under $25 bins in the range of $0 to $5,000. The bin and range combination were chosen to provide an appropriate view of the distribution. Furthermore, **Graph 13** is a box-and-whisker plot of Amount.



***Graph 12: Histogram of Amount for Credit Card Transactions in 2010***



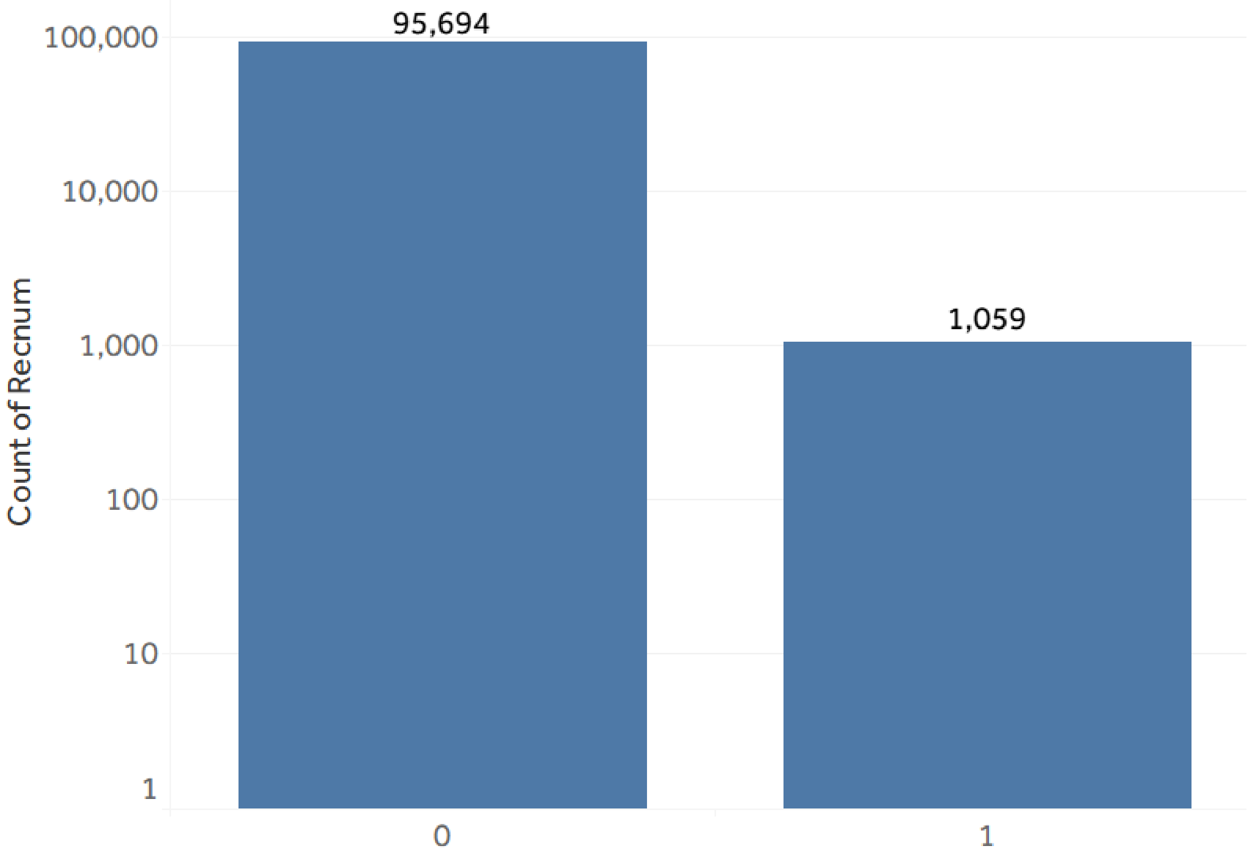
***Graph 13: Box-and-Whisker Plot of Amount for Credit Card Transactions in 2010[1]***

[1] Graph 12 excludes an outlier with the value of $3,102,046

Based on **Graphs 12** and **13**, Amount appears to be skewed to the right and the box-and-whisker plot shows that there are many outliers. With an upper whisker at $1,729, there are 5,060 outliers.

## 3.10 Fraud

Fraud is a categorical field that indicates whether a credit card transaction was fraudulent (value = 1). **Graph 14** below shows the count of fraudulent and non-fraudulent credit card transactions.



***Graph 14: Count of Fraud for Credit Card Transactions in 2010***

Based on **Graph 14**, approximately 1.09% of the credit card transactions in calendar year 2010 were fraudulent.

# 4. Exhibits

## Exhibit 1: Field Definitions

**Recnum:** Record number is an identification number unique to every transaction

**Cardnum:** Card number is the credit card number used for the transaction

**Date:** Date happens to be the day of the transaction. It is in the format of mm/dd/yyyy

**Merchnum:** Merchant number is the identification number that the entity providing the good or service to the credit card holder. It is a unique business identification number.

**Merch Description:** Merchant description provides more information regarding the name of the entity providing the service to the credit card holder.

**Merch State:** Merchant state is the state where the business providing the good or service resides. Until more information is provided, it is unclear whether Merchant state is helpful in identifying where the credit card transaction took place.

**Merch Zip:** Merchant ZIP code is the ZIP code where the business providing the good or service resides. Until more information is provided, it is unclear whether Merchant ZIP code is helpful in identifying where the credit card transaction took place.

**Transtype:** There are four types of transactions (A, D, P, and Y), but the actually meaning behind these labels are unknown.

**Amount:** Dollar amount paid through the transaction

**Fraud:** This binary field indicates whether the transaction was actually a case of fraud.

## Exhibit 2: Uncertain Merchant State Information

There were a variety of reasons that made it difficult to map some of the transactions. Some of the reasons include:

- Numerical IDs with unknown meaning;
- Letter combinations that do not necessarily match to US states; and
- Blank entries

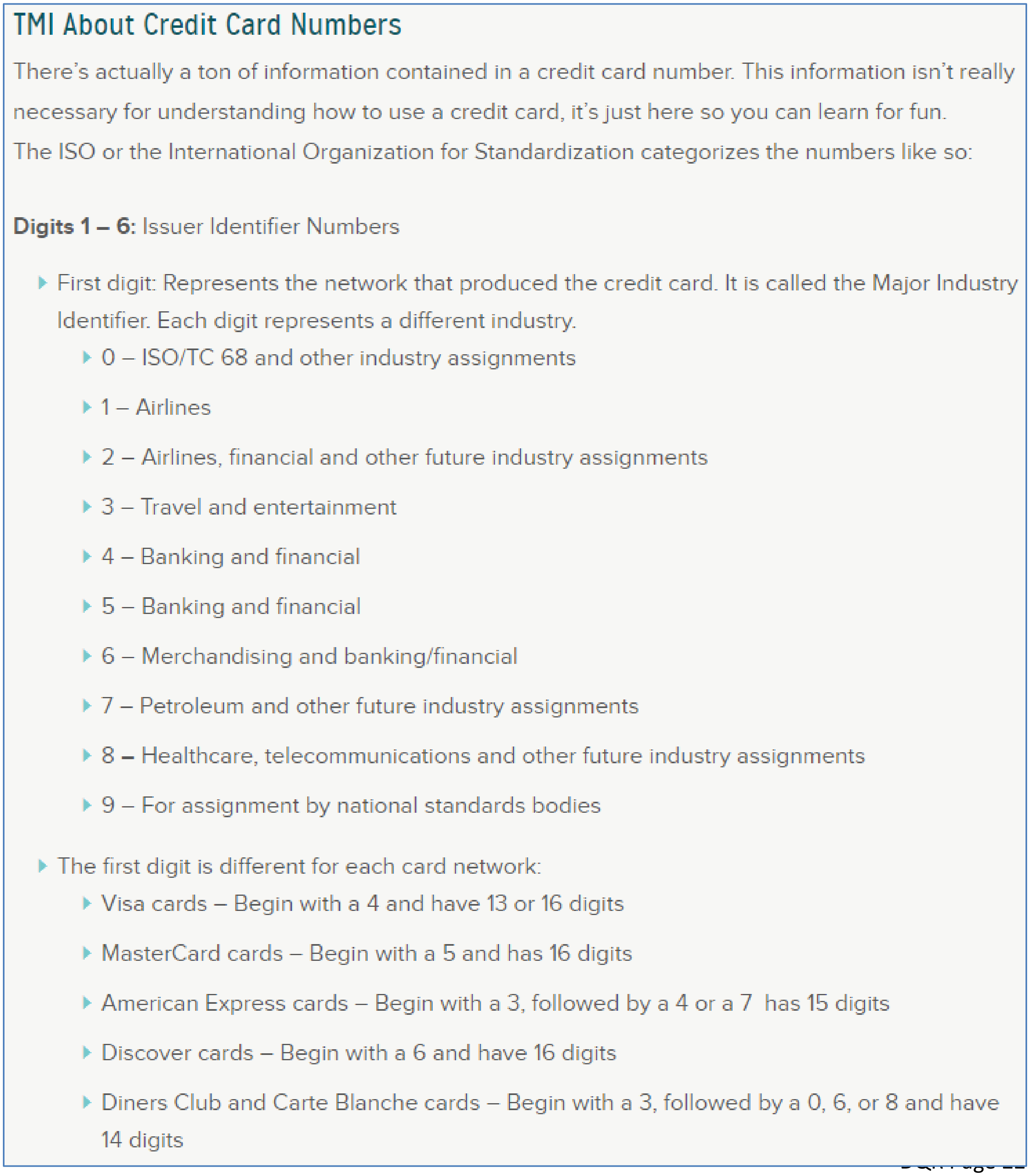
The last two categories make up the bulk of the reasons for not match transactions to a state.

| State | Count |
|---|---|
| Numerical ID | 180 |
| AB | 5 |
| BC | 23 |
| DC | 3,208 |
| MB | 3 |
| NS | 5 |
| ON | 137 |
| PQ | 14 |
| QC | 4 |
| US | 1 |
| (blank) | 1,195 |
| **Total** | **4,775** |

# Exhibit 3: Credit Card Structure

**Source:** Harkness, B. (2018, November 14). Anatomy of a Credit Card. Retrieved February 26, 2019, from https://www.creditcardinsider.com/learn/anatomy-of-a-credit-card/

The article used standards from the American National Standards Institute and the ISO or the International Organization for Standardization

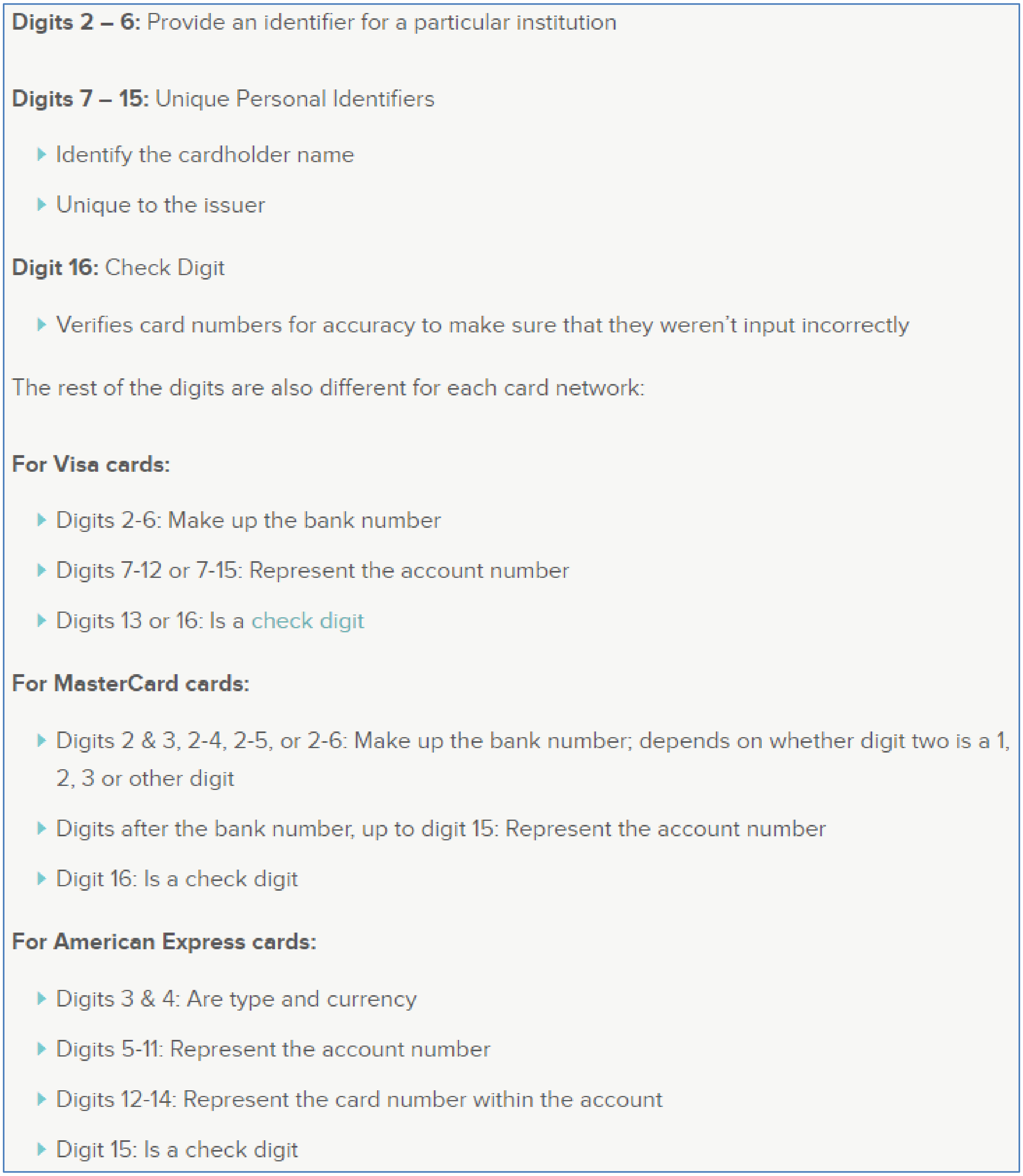


## TMI About Credit Card Numbers

There's actually a ton of information contained in a credit card number. This information isn't really necessary for understanding how to use a credit card, it's just here so you can learn for fun.
The ISO or the International Organization for Standardization categorizes the numbers like so:

**Digits 1 – 6:** Issuer Identifier Numbers

- First digit: Represents the network that produced the credit card. It is called the Major Industry Identifier. Each digit represents a different industry.
    - 0 – ISO/TC 68 and other industry assignments
    - 1 – Airlines
    - 2 – Airlines, financial and other future industry assignments
    - 3 – Travel and entertainment
    - 4 – Banking and financial
    - 5 – Banking and financial
    - 6 – Merchandising and banking/financial
    - 7 – Petroleum and other future industry assignments
    - 8 – Healthcare, telecommunications and other future industry assignments
    - 9 – For assignment by national standards bodies
- The first digit is different for each card network:
    - Visa cards – Begin with a 4 and have 13 or 16 digits
    - MasterCard cards – Begin with a 5 and has 16 digits
    - American Express cards – Begin with a 3, followed by a 4 or a 7  has 15 digits
    - Discover cards – Begin with a 6 and have 16 digits
    - Diners Club and Carte Blanche cards – Begin with a 3, followed by a 0, 6, or 8 and have 14 digits



**Digits 2 – 6:** Provide an identifier for a particular institution

**Digits 7 – 15:** Unique Personal Identifiers

- Identify the cardholder name
- Unique to the issuer

**Digit 16:** Check Digit

- Verifies card numbers for accuracy to make sure that they weren't input incorrectly

The rest of the digits are also different for each card network:

**For Visa cards:**

- Digits 2-6: Make up the bank number
- Digits 7-12 or 7-15: Represent the account number
- Digits 13 or 16: Is a check digit

**For MasterCard cards:**

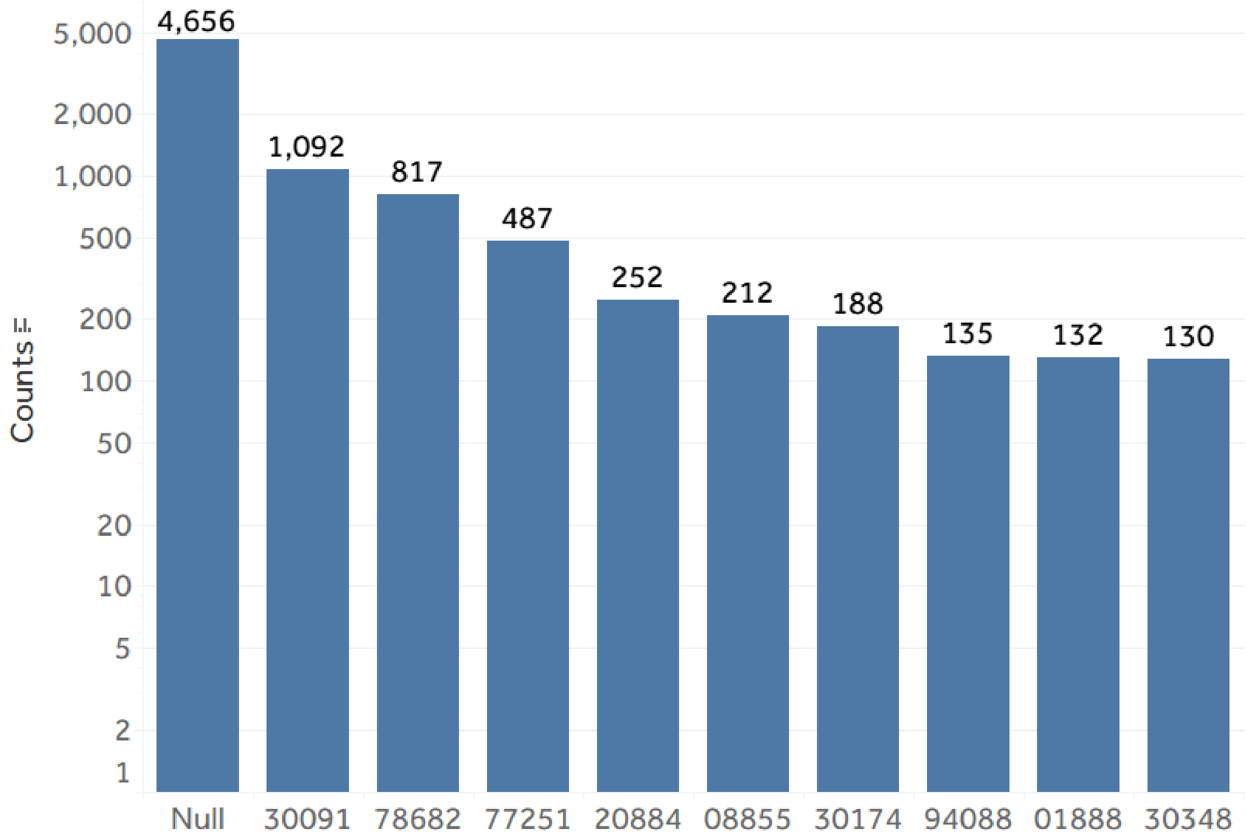
- Digits 2 & 3, 2-4, 2-5, or 2-6: Make up the bank number; depends on whether digit two is a 1, 2, 3 or other digit
- Digits after the bank number, up to digit 15: Represent the account number
- Digit 16: Is a check digit

**For American Express cards:**

- Digits 3 & 4: Are type and currency
- Digits 5-11: Represent the account number
- Digits 12-14: Represent the card number within the account
- Digit 15: Is a check digit

## Exhibit 4: Uncertain Merchant ZIP Code Information

11,951 transactions were not matched to a location on a map. On of the largest reasons for this is due to 4,656 of those transactions not having an entry. Even with entries, not all fields were matched to the ZIP code information provided by the US Census Bureau. More research is required to identify whether or not certain ZIP codes are missing from the provided census information or if the ZIP code itself is invalid.



***Graph 15: Top 10 Unmatched Fields for Merch Zip***