Data Quality Report

1. Description

The Credit Card Transaction Data contains information about instances of credit card transactions and if that transaction was flagged as fraudulent or not. It contains about 96,753 records and 10 fields for each record. These fields include fields like Recnum, Cardnum, Date, Merchnum, Merch description, Merch state, Merch zip, Transtype, Amount and a fraud label with 1 if the record is fraudulent or 0 otherwise. The dataset has 1,059 records with fraudulent transactions and 95,694 records classified as not fraudulent.

2. Summary Tables

The following table summarizes the fields in the data and describes their distribution. As all the field seem to be categorical in nature, a single table has been constructed.

2.1 Categorical Fields

Field Name	Records with Values	Percentage Populated	Unique Values	Records with Zero	Most Common Value
Cardnum	96753	100	1645	0	5142148452
Date	96753	100	365	0	2010-02-28
Merchnum	93378	96.51	13092	0	930090121224
Merch description	96753	100	13126	0	GSA-FSS-ADV
Merch state	95558	98.76	228	0	TN
Merch zip	92097	95.18	4568	0	38118
Transtype	96753	100	4	0	P
Fraud	96753	100	2	95694	0

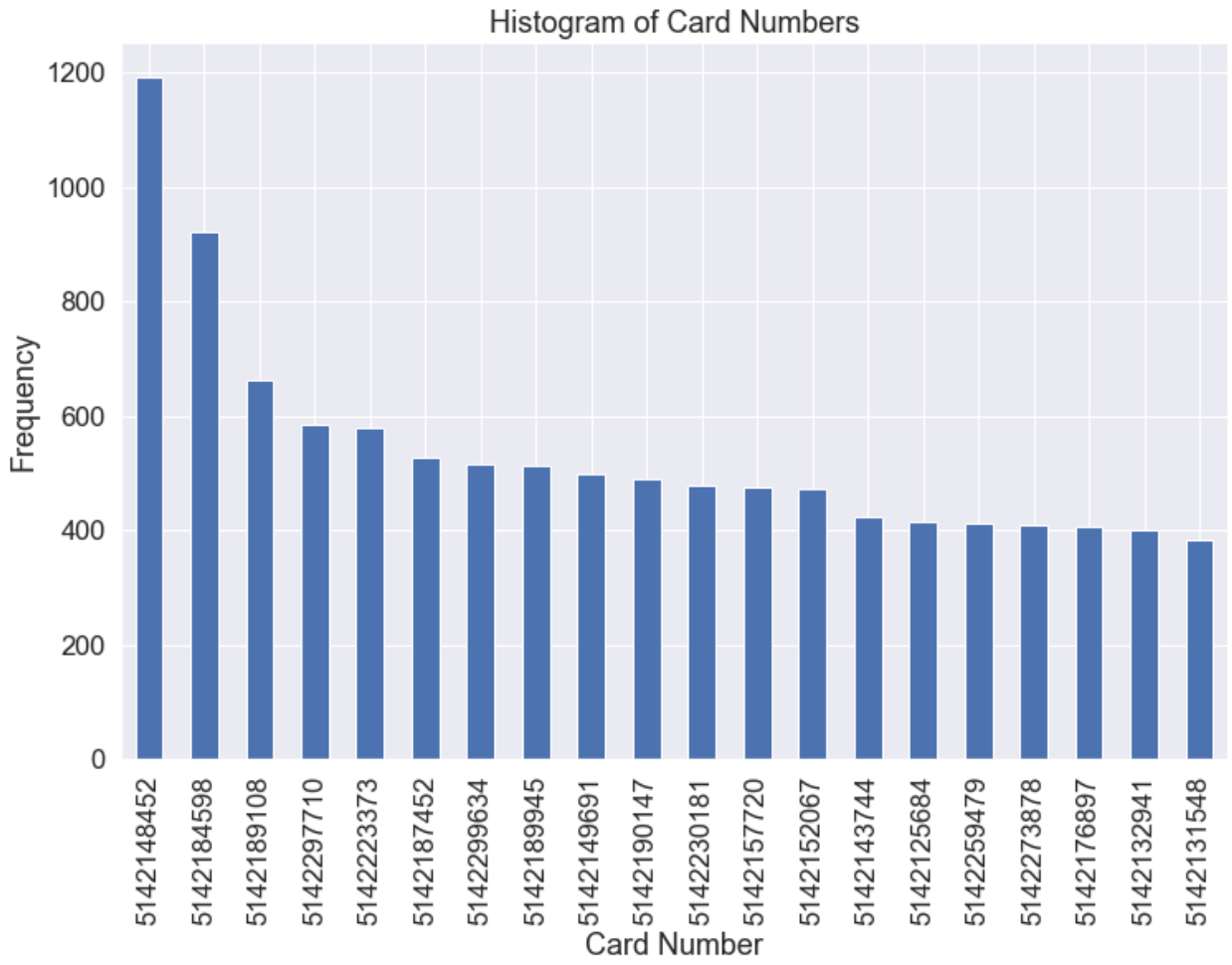
2.2 Continuous Fields

Field Name	Records with Values	Percentage Populated	Records with Zero	Mean	Standard Deviation	Min	Max
Amount	96753	100	0	427.88	10006.14	0.01	3102045.53

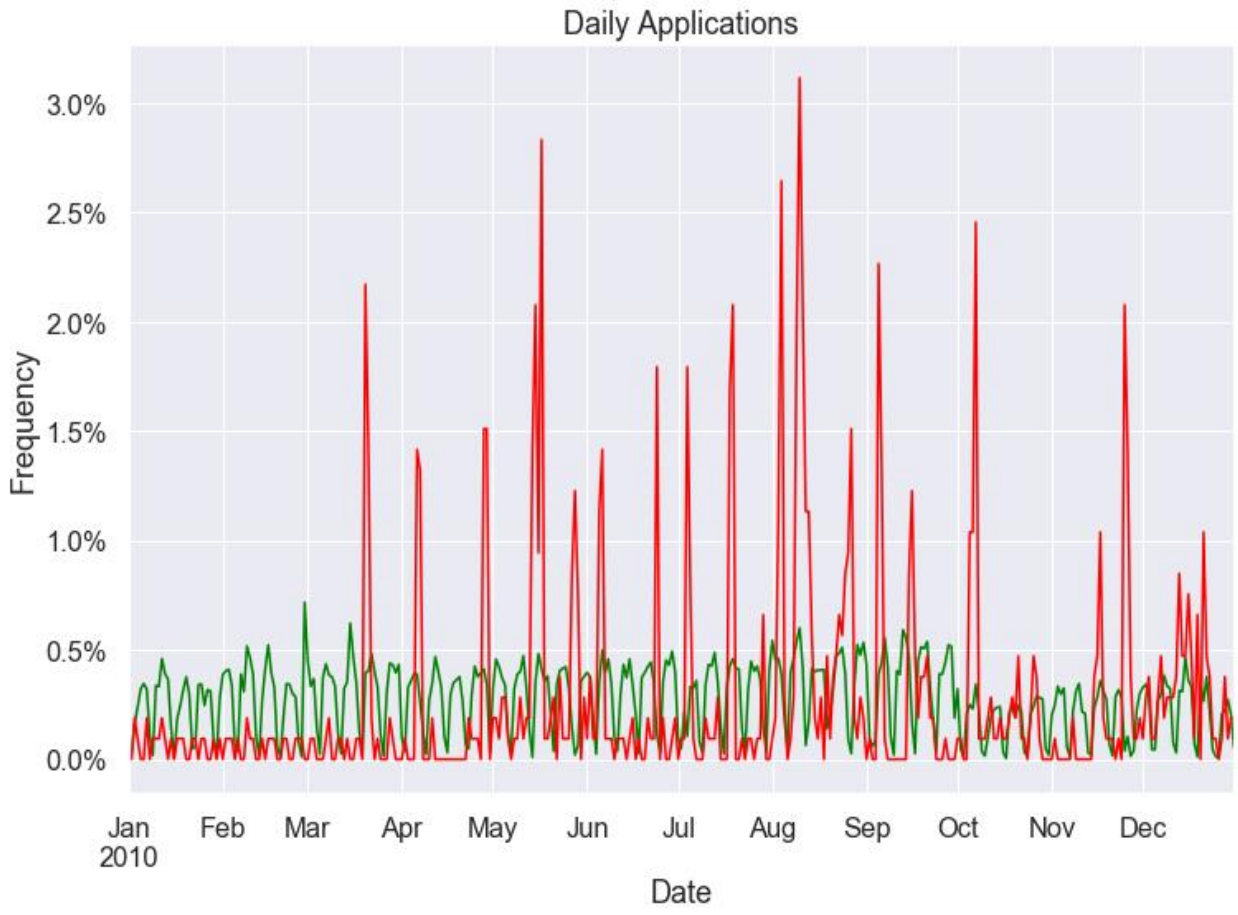
3. Fields

3.1 Recnum: It is an integer field that uniquely identifies each record in the data. It ranges from 1 to 96,753.

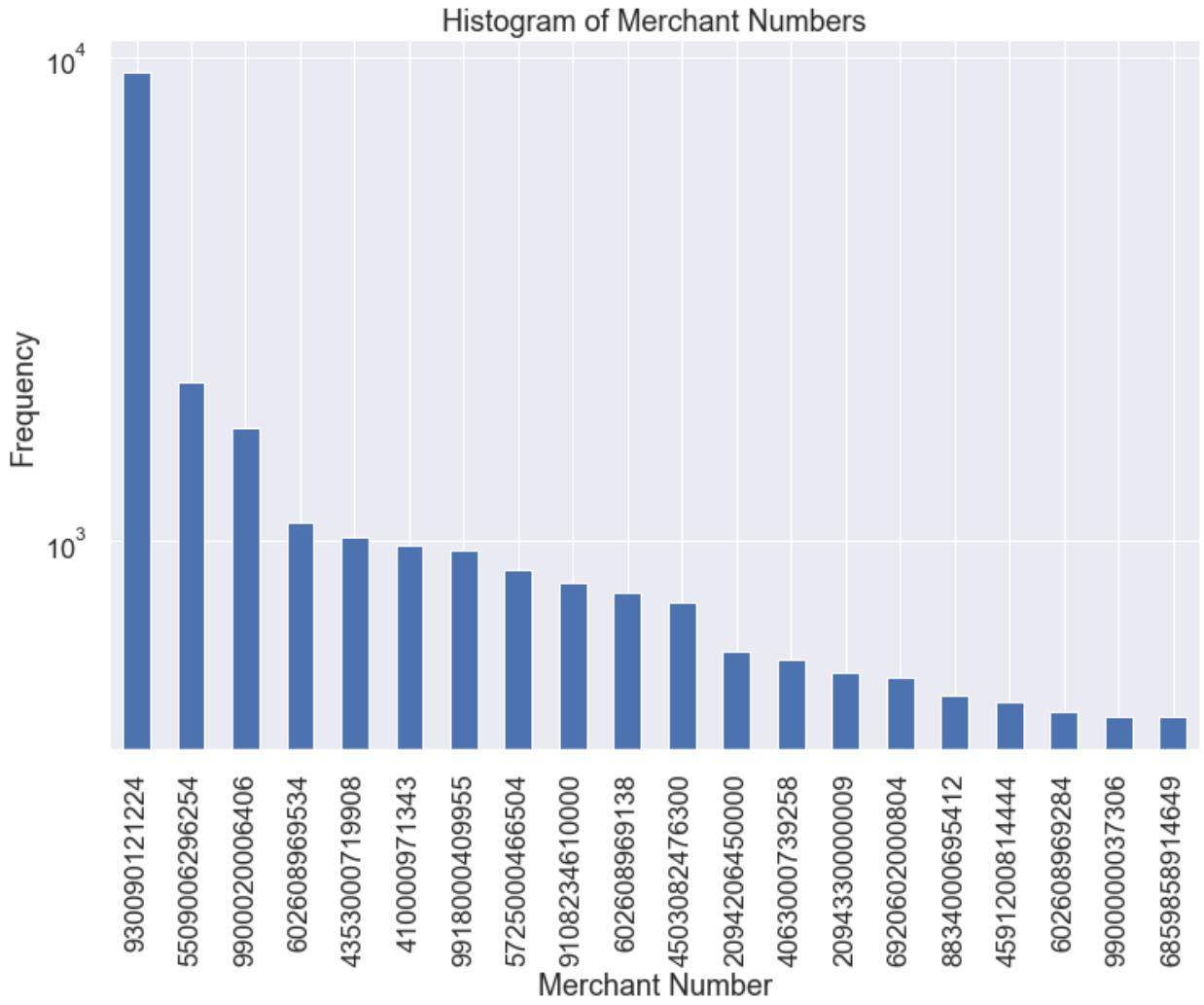
3.2 Cardnum: This field contains the card number of the card used in the transaction. A single card number can be seen for multiple transactions. The following histogram shows the distribution for the top 20 most common card numbers in the data. It can be observed that the transactions for the cards with numbers “5142148452” and “5142184598” are much greater than the rest which is odd and needs to be investigated.



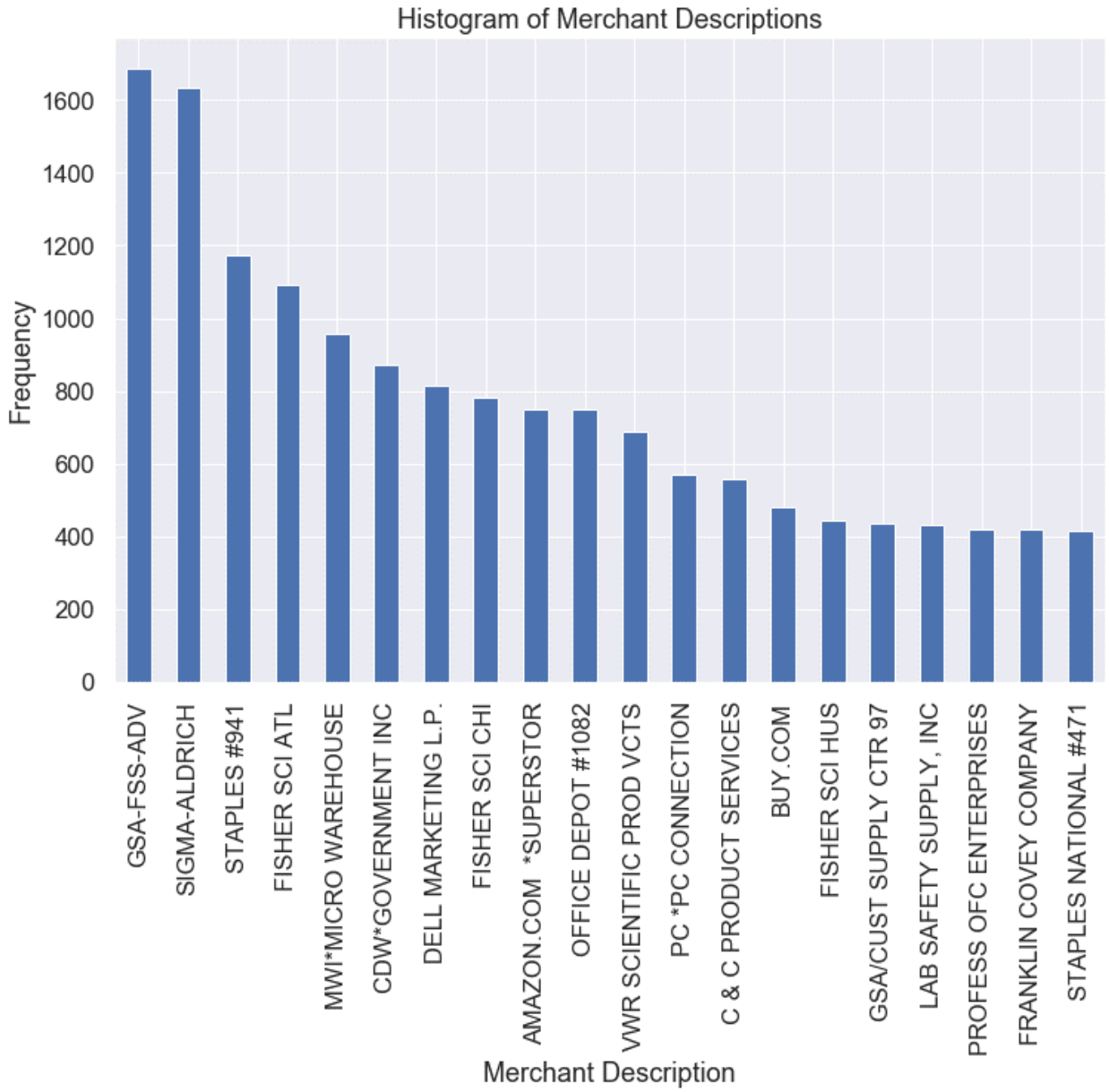
3.3 Date: This field contains the date of the transaction. It is stored as a datetime object. The following histogram shows the daily distribution of transactions over the given time period. The red line indicates fraudulent applications and the red line indicated the nonfraudulent applications.



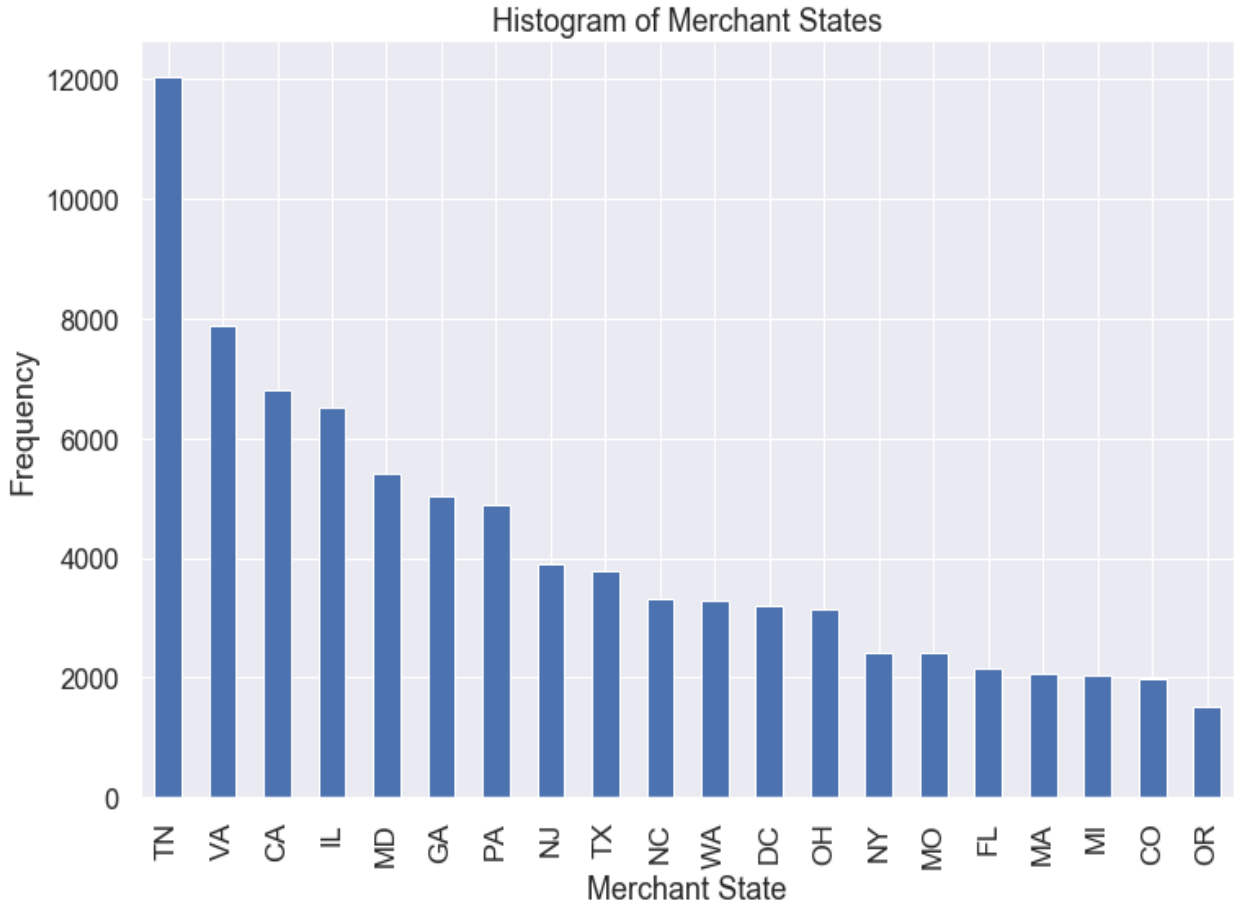
3.4 Merchnum: This field contains number that uniquely identify merchants involved in the transactions. The following histogram shows the frequency distribution of the top 20 most common merchant numbers in the data. It can be observed that the number of transactions for the merchant number "930090121224" is much greater than that for the rest of the merchant numbers which is odd and needs to be investigated.



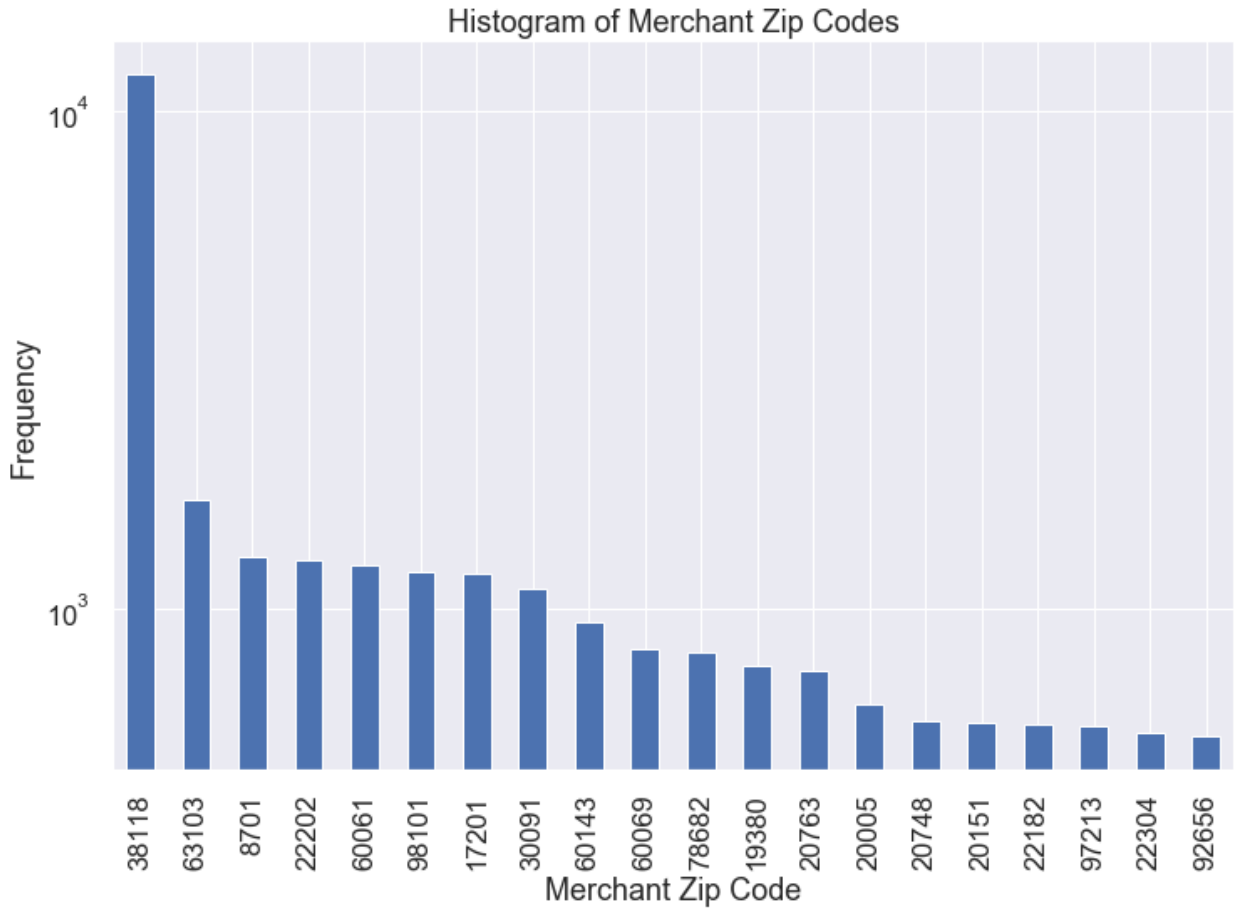
3.5 Merch description: This field gives a description of the merchant where the transaction has occurred. The following histogram shows the frequency distribution of the top 20 most common merchant descriptions in the data.



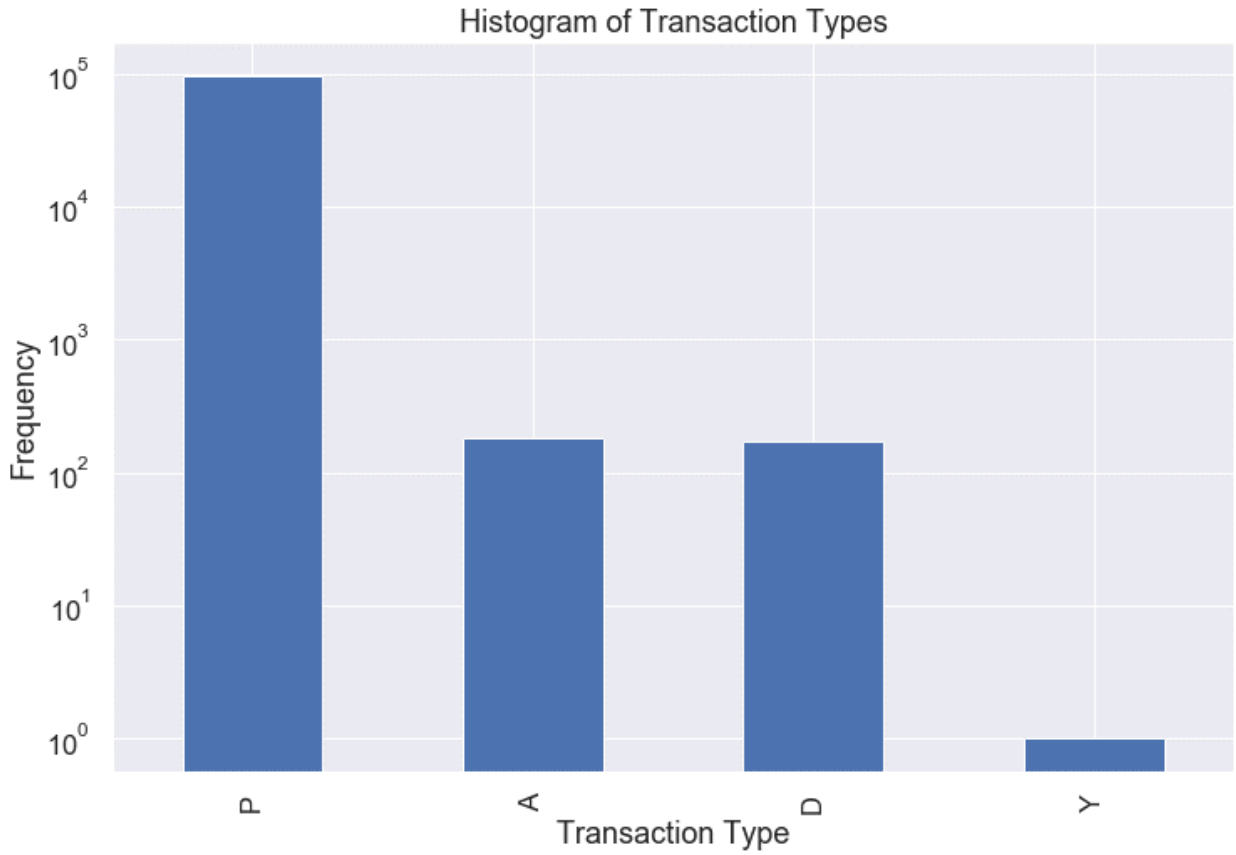
3.6 Merch state: This field indicates the state in which the merchant involved in the transaction is located. The following histogram shows the frequency distribution of the top 20 most common merchant states in the data.



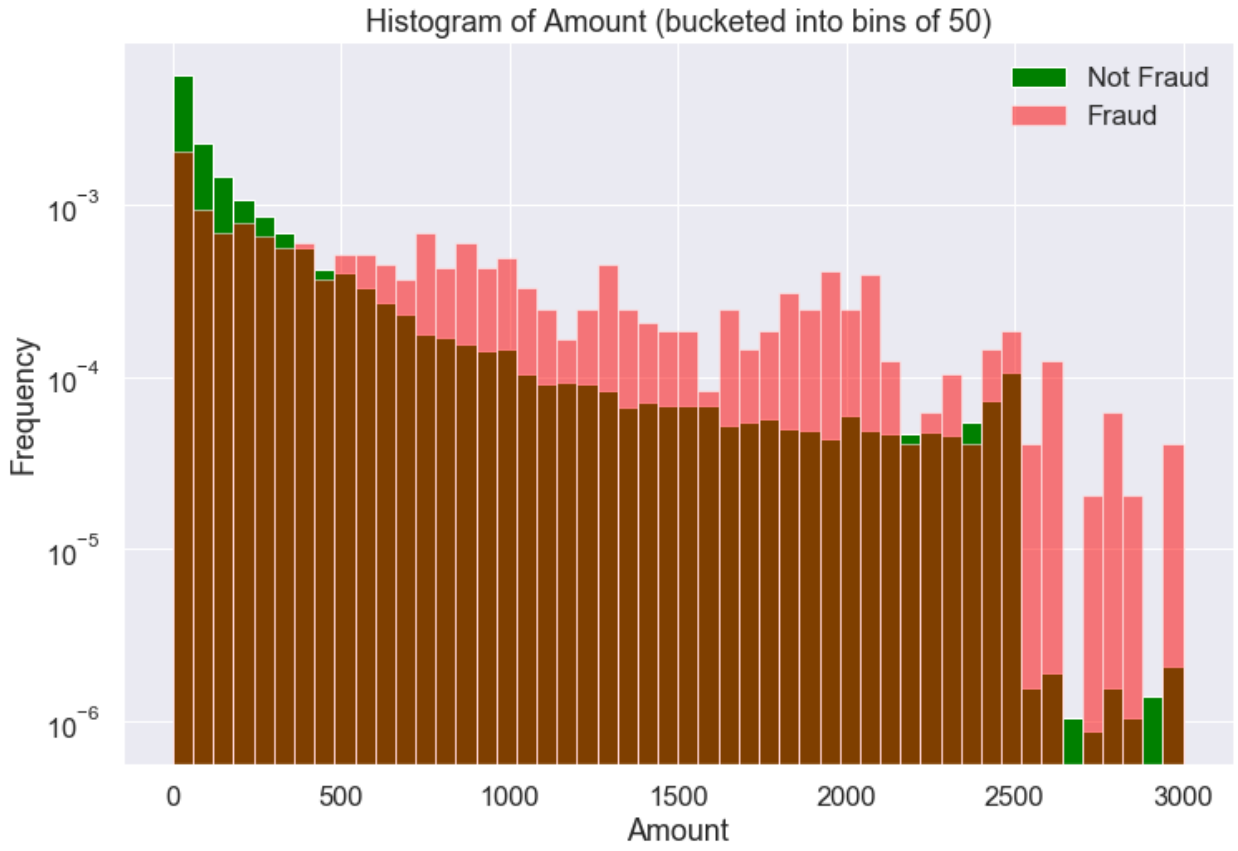
3.7 Merch zip: This field contains the zip code of the merchant involved in the transaction. The following histogram shows the frequency distribution of the top 20 most common merchant states in the data. It can be observed that the number of transactions for the merchant zip code “38118” is much greater than that for the rest of the merchant zip codes which is odd and needs to be investigated.



3.8 Transtype: This field refers to the transaction type of the record. The codes in this field could mean the following: P for Prenote, A for Authorization, D for Delayed Capture and there is not enough information figure out Y. The following histogram shows the frequency distribution of the Transaction Types in the data.



3.9 Amount: This field contains the amount of money involved in the transaction in the data. It can be assumed the currency involved is dollars (USD). The following histogram shows the frequency distribution of the Amount bucketed into bins of size 50. The non-fraudulent distribution is green in color and the fraudulent distribution in shown in red.



3.10 Fraud: This field has a label that classifies if a record is fraudulent or not. A value of 1 indicates that the record is fraudulent and a value of 0 indicates that the record is not fraudulent. The data has 1,059 fraudulent records and 95,694 non-fraudulent records.