Data Quality Report

Dataset: Card Transactions

Description: This data is a simulated representation of 96708 card transaction requests during 2010. A typical record contains information about the card and merchant details, along with geographical location of the merchant, the dollar amount being requested for transaction and whether it is fraud or not.

Number of Records: 96708

Number of Variables: 10

Summary Statistics

Variable	Description	Percent Populated	Number of Unique Entries	Mean	Standard Deviation	Min	25%	50%	75%	Max	Mode	Frequency of Modal Entry
Recordnum	Record number	100%	96708	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cardnum	Card number	100%	1644	NA	NA	NA	NA	NA	NA	NA	5142148452	1192
Date	Date of transaction	100%	365	NaN	NaN	NaN	NaN	NaN	NaN	NaN	28-02-10	684
Merchantnum	Merchant number	96.51%	13091	NA	NA	NA	NA	NA	NA	NA	930090121224	9310
Merch Description	Merchant description	100%	13125	NaN	NaN	NaN	NaN	NaN	NaN	NaN	GSA-FSS-ADV	1688
Merchant State	Merchant state	98.76%	228	NaN	NaN	NaN	NaN	NaN	NaN	NaN	TN	11990
Merchant Zip	Merchant's ZIP	95.19%	4568	44709.82	28376.1	1	NaN	NaN	NaN	99999	38118	11823
Transtype	Type of transaction	100%	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	P	96353
Amount	Transaction amount	100%	34876	427.865	10008.47	0.01	33.5	137.9	427.715	3102046	3.62	4283
Fraud	Fraud or not (binary)	100%	2	0.010485	0.101859	0	0	0	0	1	0	95694

Description and Visualization

1. ***Recordnum*** is a categorical variable. It works as the ordinal reference number for each property record. There are 96708 records overall. Each row is a unique number (i.e., an identifier) and hence, visualization is not required.

2. ***Cardnum*** refers to the card number used at the point of sale.

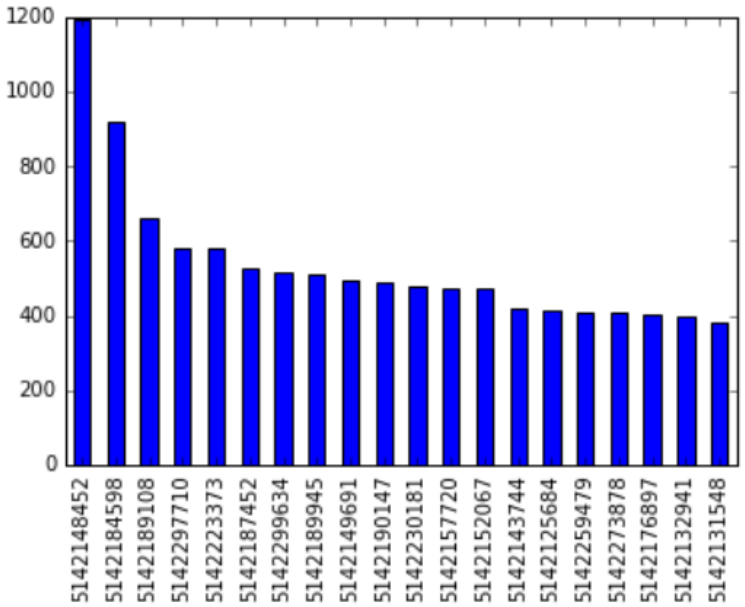

Figure 1: 20 most frequently occurring card numbers

3. ***Date*** refers to the date on which the transaction was requested.

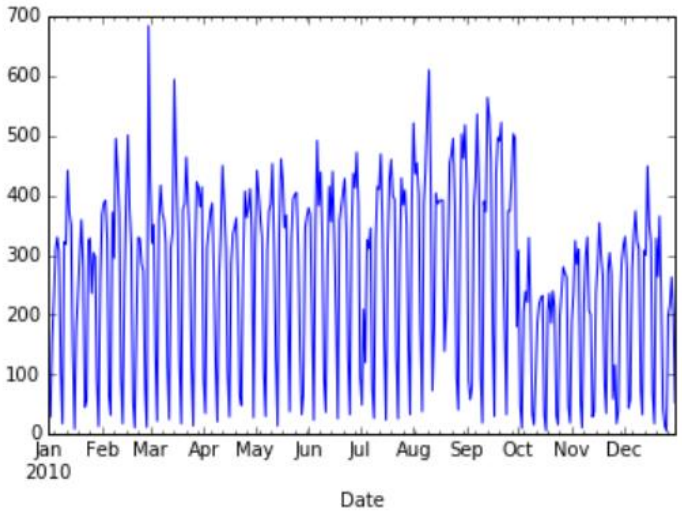

Figure 2: Daily transaction requests (2010)

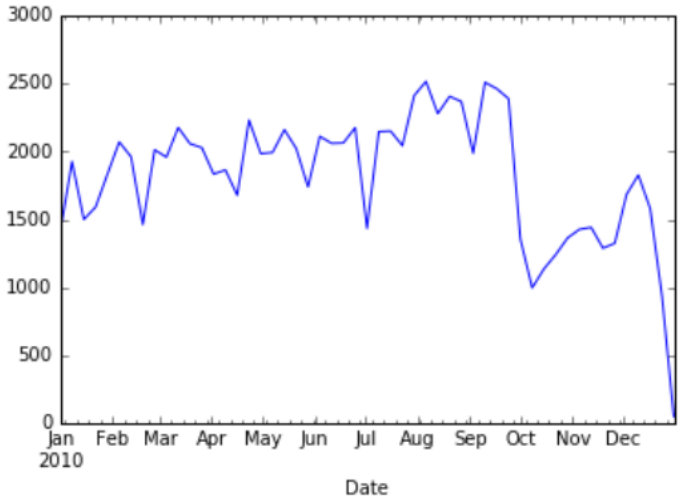

Figure 3: Weekly transaction requests (2010)

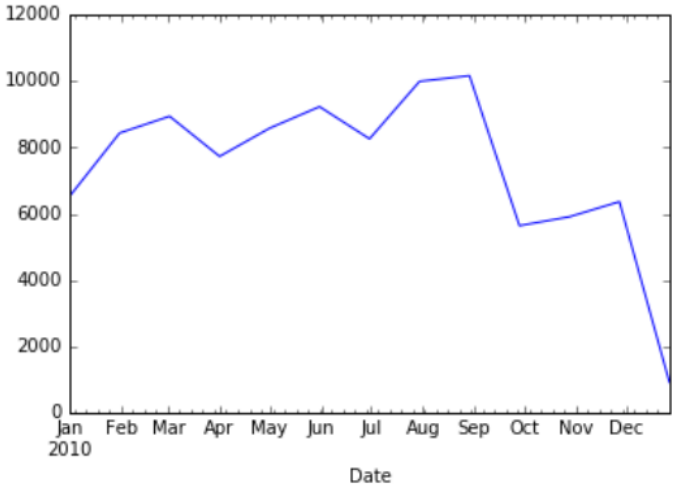

Figure 4: Monthly transaction requests (2010)

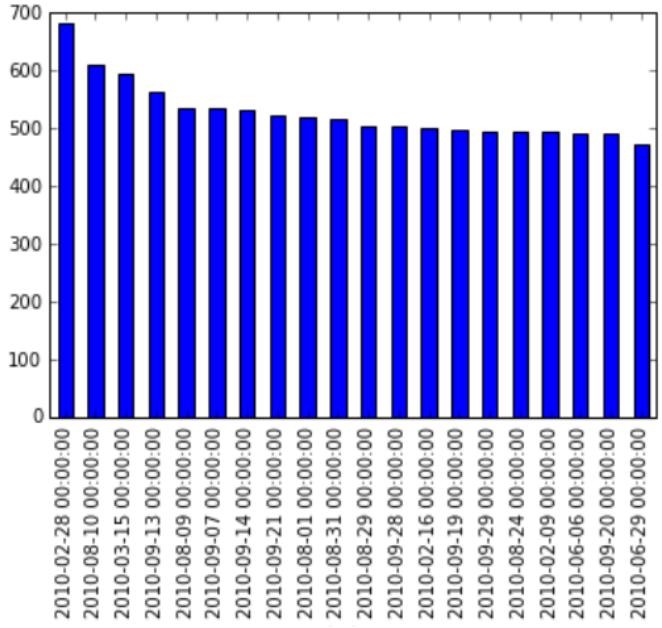

Figure 5: Top 20 days in terms of volume of transactions

4. ***Merchantnum*** refers to unique merchant IDs associated with each transaction.

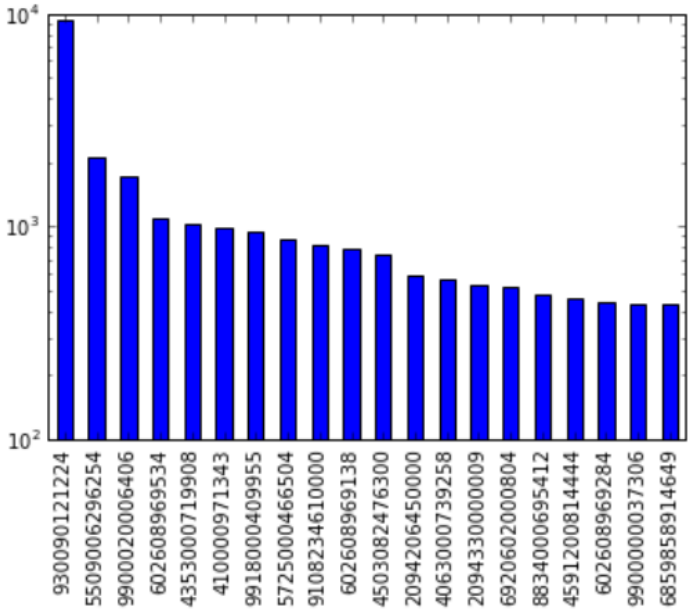

Figure 6: 20 most frequently occurring merchant IDs (log-scaled)

5. ***Merch Description*** refers to the name of the merchant which carried out the transaction. 20 most frequently occurring merchants have been shown below.

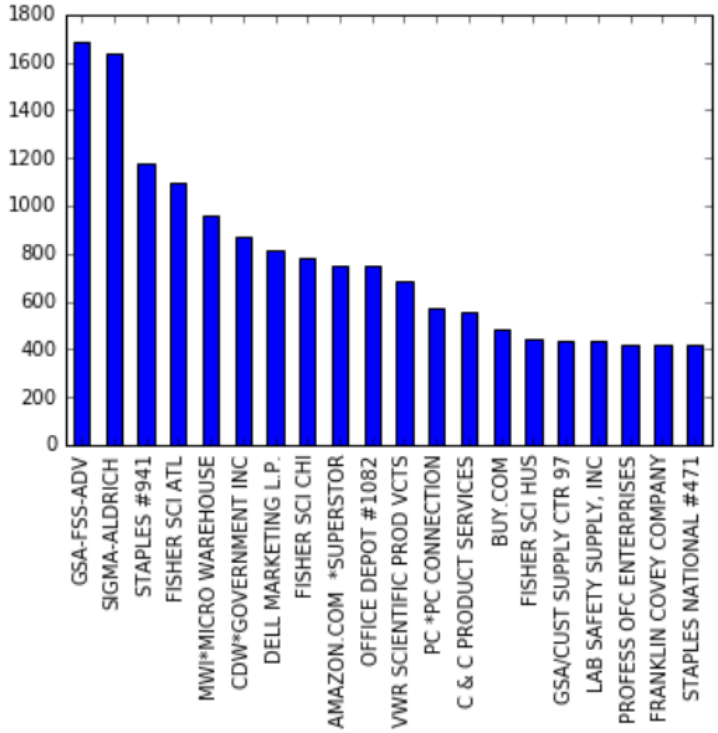

Figure 7: 20 most frequently occurring merchants

6. ***Merchant State*** refers to the state/territory in which the merchant is located. States are abbreviated using 2-letter codes; there are other territories are represented using 3-digit numeric codes.

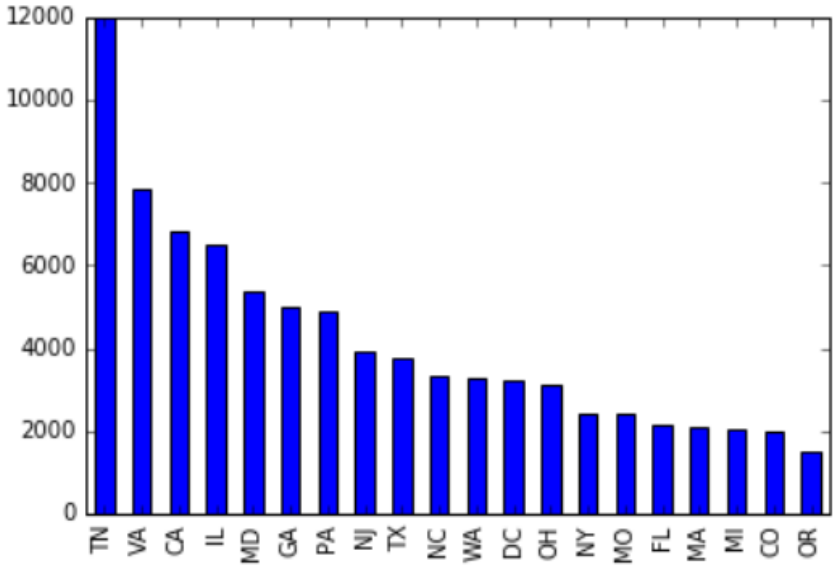

Figure 8: 20 states from where most transactions were requested

7. ***Merchant Zip*** contains information about each merchant's ZIP code. Below is a graph of the top 20 most frequently occurring ZIP codes.

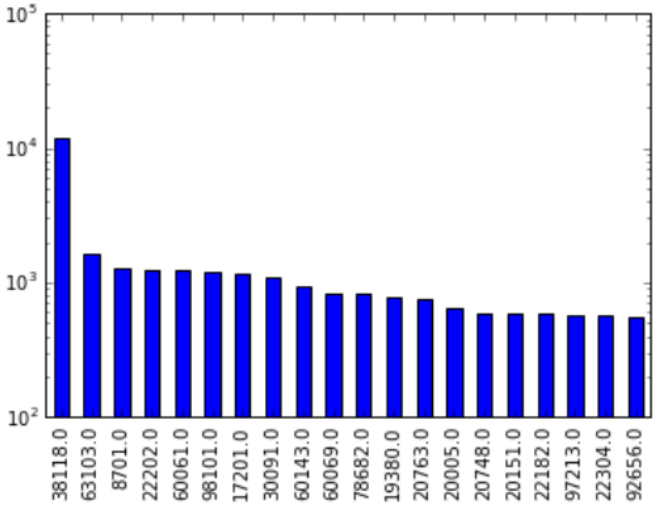

Figure 9: 20 most frequently occurring ZIP codes (log-scaled)

8. ***Transtype*** is a categorical variable referring to the type of transaction. There are four different types of transactions in the dataset.

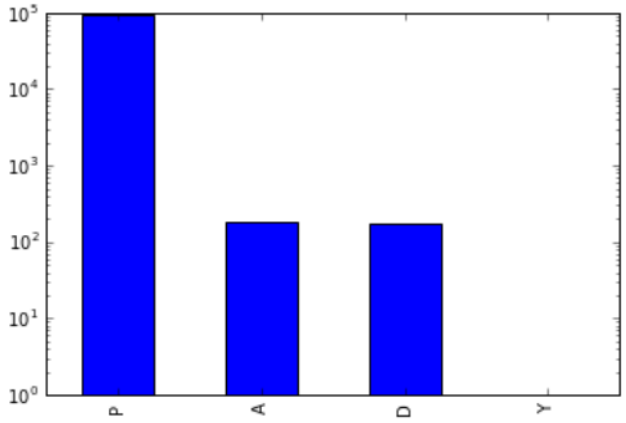

Figure 10: Log-scaled frequency distribution of transaction types

9. ***Amount*** contains information about the dollar amount corresponding to each transaction. Below is a graph of the top 20 most frequently occurring transaction amounts.

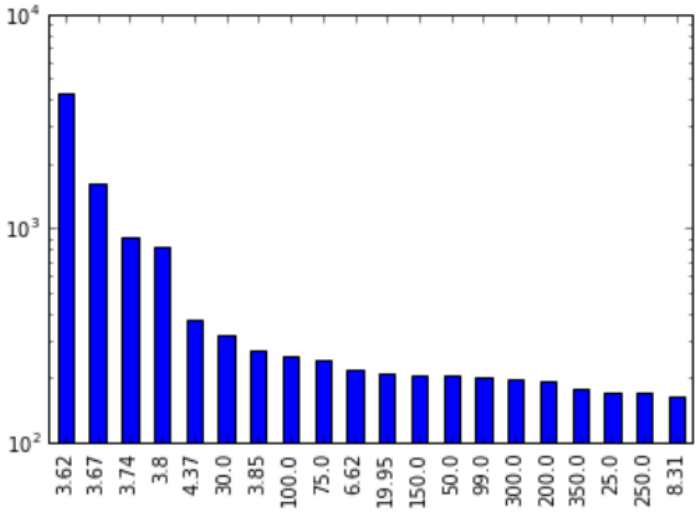

Figure 11: 20 most frequently occurring amounts (log-scaled)

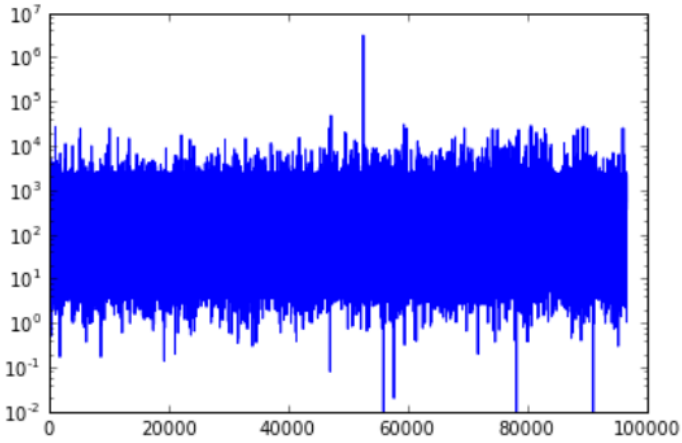

Figure 12: Distribution of transaction amounts (log-scaled)

10. ***Fraud*** contains information whether the transaction was determined to be fraud or not. Of the total 96708 observations, we observed that 95694 observations are recorded as legitimate, whereas 1014 transactions were marked as fraudulent.

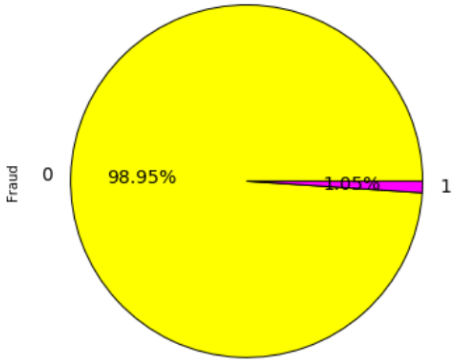

Figure 13: Pie chart showing percentage of fraudulent transaction requests

Conclusions

This dataset is to be used to build a fraud detection model for a future project. Having tabulated and visualized all the variables, it is evident that there's a small fraction of anomalies in the data. Also, some data cleaning will likely be required before we can apply mathematical models. Going forward, our aim is to build a supervised learning model to detect fraud with reasonable accuracy.