DATA QUALITY REPORT

By:

Sudhanshu Rai

Section 1 High Level Description

- The dataset contains information about various customers who either committed fraud or didn't do it.
- There are 1 million rows and 10 fields, containing information such as date of transaction, SSN, first name, last name, provided address, date of birth, etc.
- It also contains y label as fraud label containing zero and ones.
- Data is highly imbalanced, with only 1.42% fraud labels as one while others are zero.
- There are no missing values in the dataset.

Section 2 Field. Summary Tables

The following section contains two tables, based on categorical and numerical values, providing some basic insights about different fields.

2.1 Numerical Fields

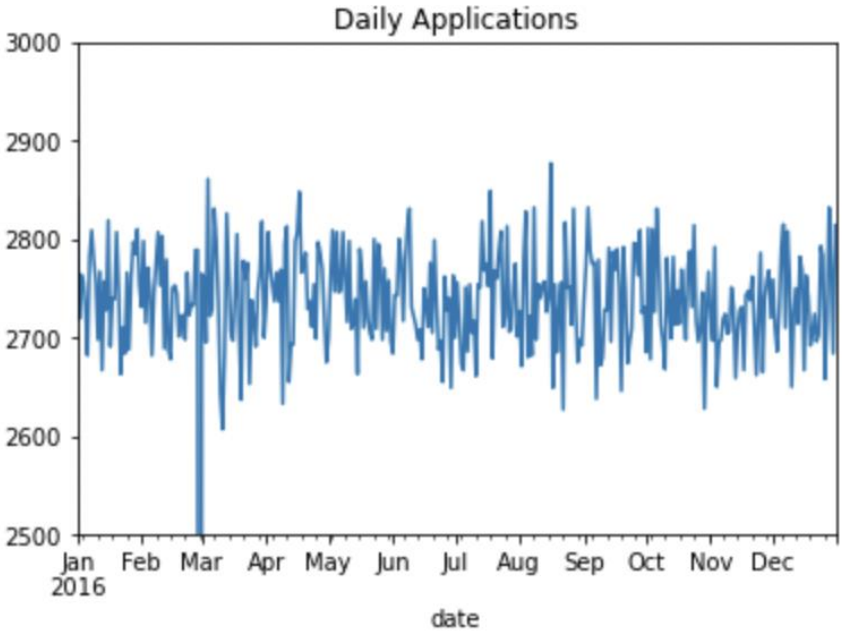
Field Name	% Populated	Min	Max	Mean	Standard deviation	% Zero
DATE	100%	2016-01-01	2016-12-31	2016-07-01 16:01:18	105 days 14:25:43	0%
DOB	100%	1900-01-01	2016-10-31	1952-02-26 04:27:50	13035 days 15:46:28	0%

2.2 Categorical Fields

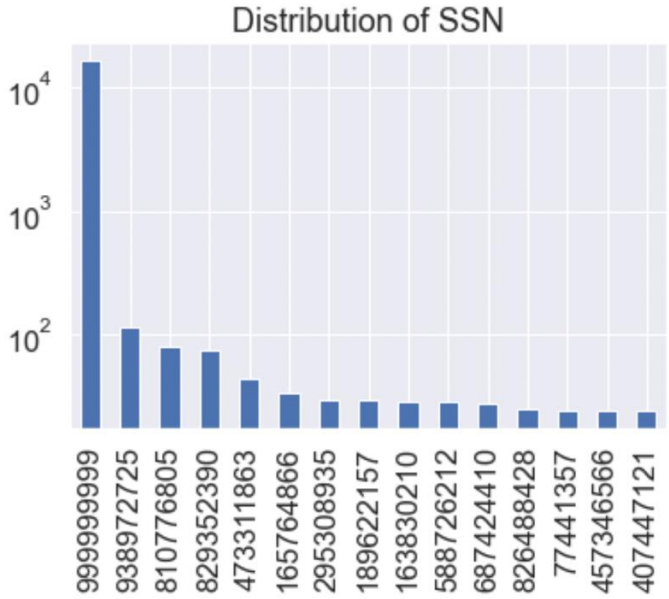
Field Name	% Populated	% Unique Values	Most Common Values
RECORD	100%	100%	All with frequency 1
SSN	100%	83.58%	999999999
FIRSTNAME	100%	7.81%	EAMSTRMT
LASTNAME	100%	17.70%	ERJSAXA
ADDRESS	100%	82.87%	123 MAIN ST
ZIP5	100%	2.64%	68138
DOB	100%	4.26%	19070626
HOMEPHONE	100%	2.82%	9999999999
FRAUD_LABEL	100%	0.002%	0

Section 3 Graphs and Charts

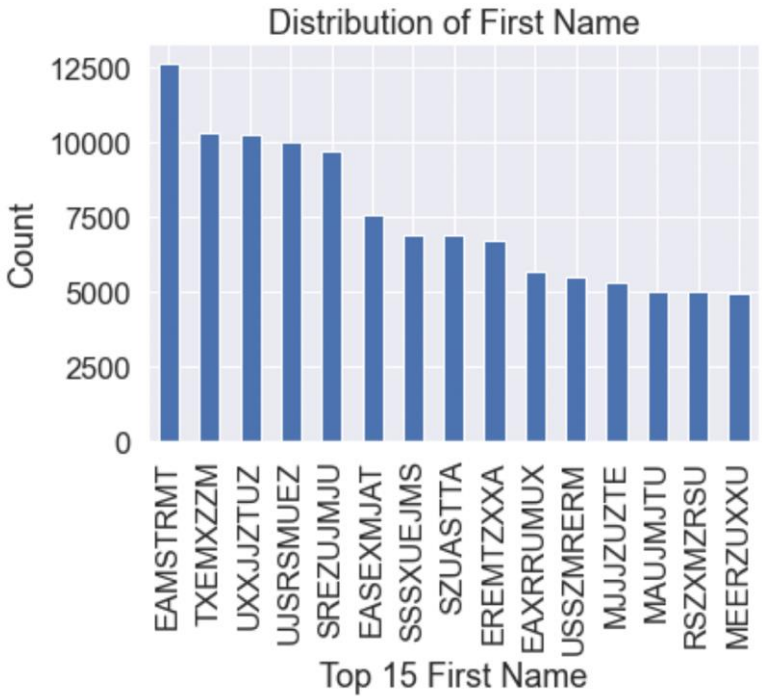
RECORD: This is index value of the table and contains incremental values.
DATE: This is numerical value, containing the dates of transaction. All the transactions are in year 2016.



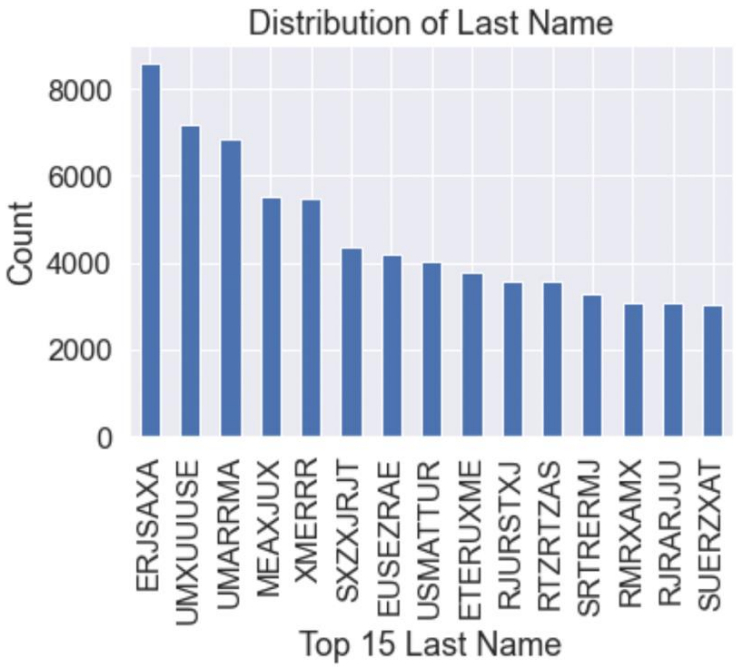
SSN: Social security number, containing 83% unique values. Also containing around 17,000 values as 999999999. Which seems like a fraud or dummy value.



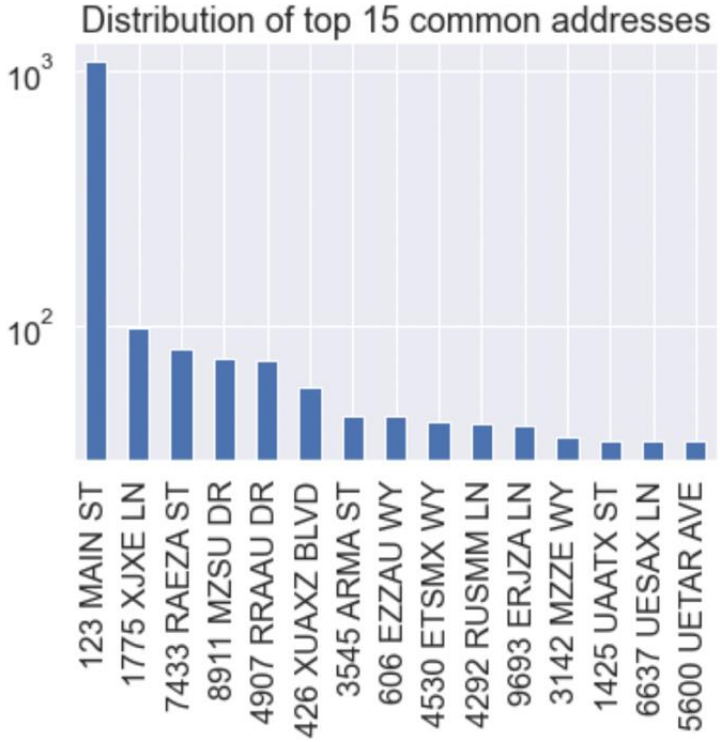
FIRSTNAME: Only around 8% of unique first names, requires some searching and attention. The field is 100% populated but low unique values is a flag. Also top 15 values seem like random keyboard types and makes no sense as the first name.



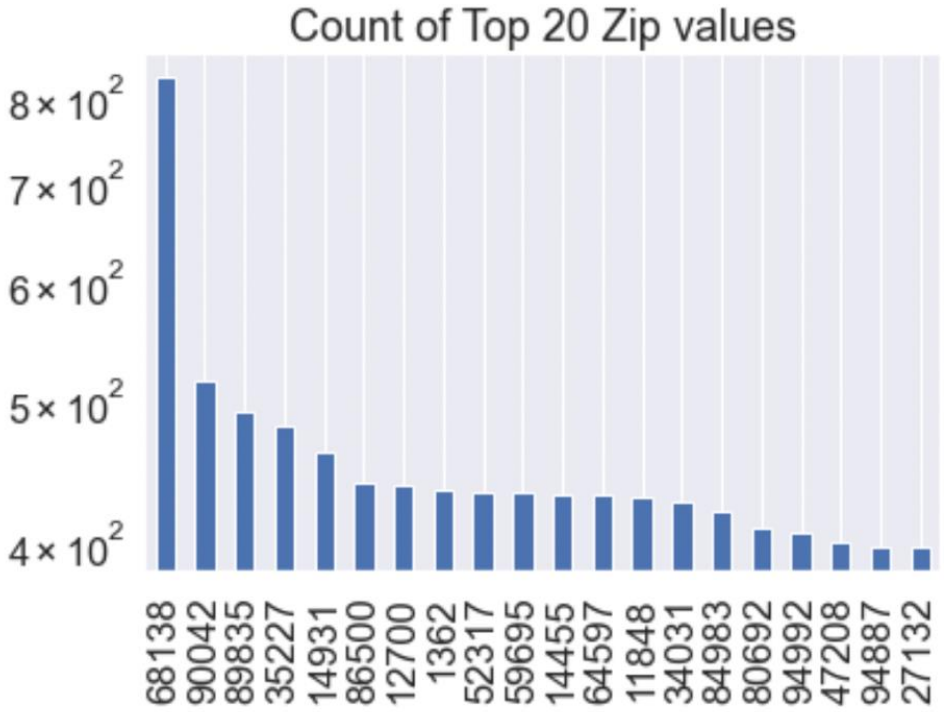
LASTNAME: Just like first name, last name also has low unique values percentage of 17%. Also, the top 15 last names look like randomly typed texts.



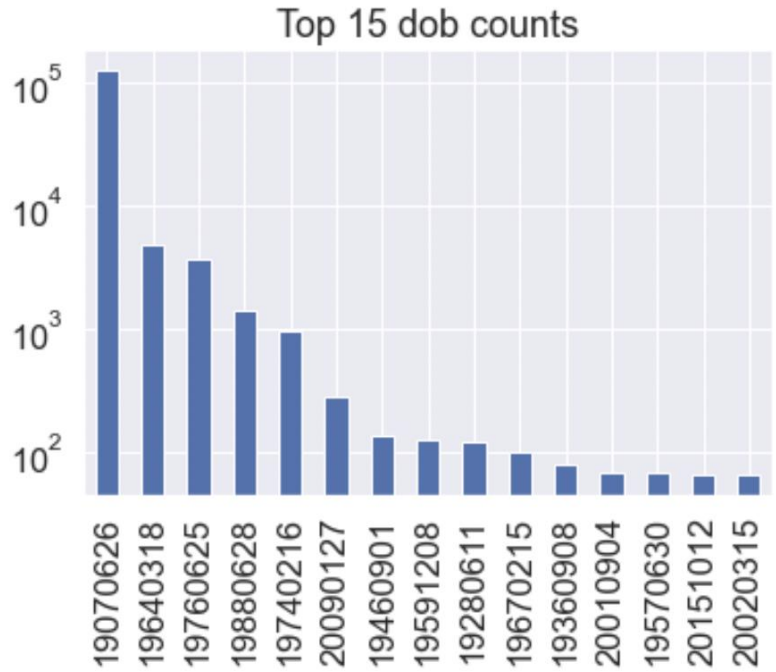
ADDRESS: This data overall looks normal, with around 82% of unique values. Only the largest or the most common value looks suspicious, which is '123 Main Street'.



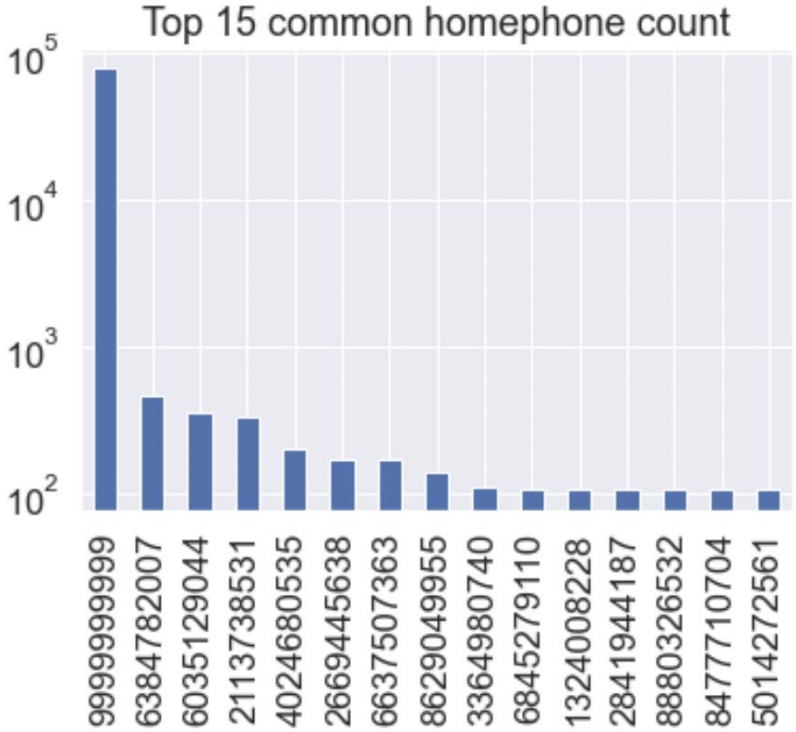
ZIP5: Another categorical variable, containing surprisingly only 2.5% values as unique. The most common value is 68138, which is **Omaha, Nebraska,** with an estimate population of only around 12 thousand people, has the highest number of records of 823.



DOB: Date of birth field containing 4.2% unique values. Also containing around 80,000 records, where the applicants' age is less than 10 years of age which is suspicious. Also, whopping 126,568 applicants born in 1907, which is around age of 109 years.



HOMEPHONE: Categorical field containing only 2.8% unique values. With over 78,000 values as 9999999999 which is very suspicious.



FRAUD_LABEL: This is the y variable or the dependent variable. This is highly imbalanced field with most records as non-fraud.