Data Quality Report

1. **Data description**

The data includes the Personal Identifying Information (PII) from the applications. There are in total 10 PII fields. The records are collected from January 1st 2017 to December 31st 2017, with the total of 1,000,000 rows. The objective is to look for identity fraud and identify fraud algorithms from the PII fields.

2. **Summary Tables**

a. Numerical Table

Field	% Populated	Min	Max	Mean	Standard Deviation	%Zero
Date	100%	2017-01-01	2017-12-31	n/a	n/a	0%
Dob	100%	1900-01-01	2016-10-31	n/a	n/a	0%

b. Categorical table

Field	% Populated	# Unique Values	Most Common Value
Record	100%	1,000,000	/
SSN	100%	835,819	999999999
Firstname	100%	78,136	EAMSTRMT
Lastname	100%	177,001	ERJSAXA
Address	100%	828,774	123 MAIN ST
Zip5	100%	26,370	68138
Homephone	100%	28,244	9999999999
Fraud_Label	100%	2	0

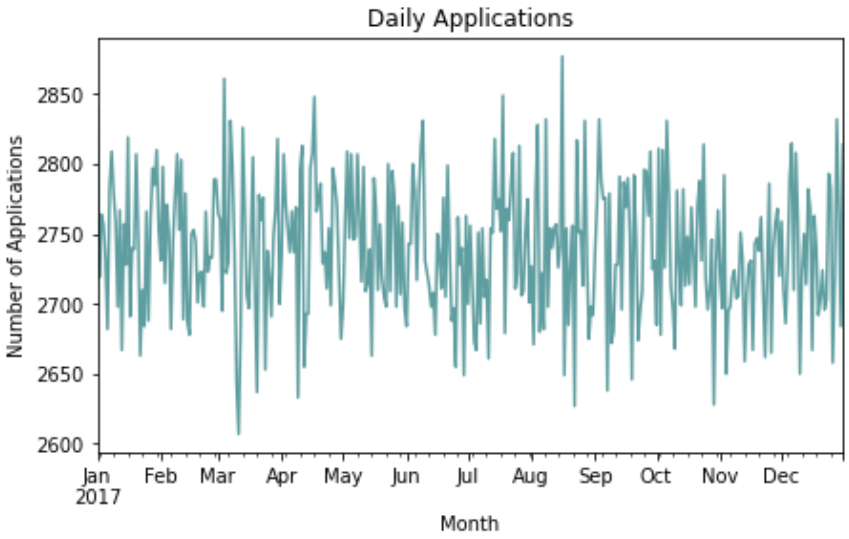
3. **Data Visualization**

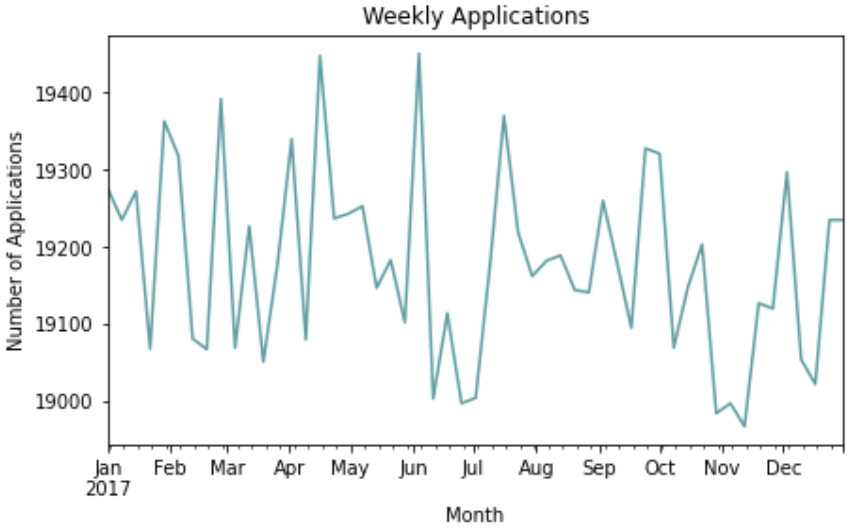
a. **Record**

Record numbers are ordinal numbers from 1 to 1,000,000. Each row is one unique positive number.

b. **Date**

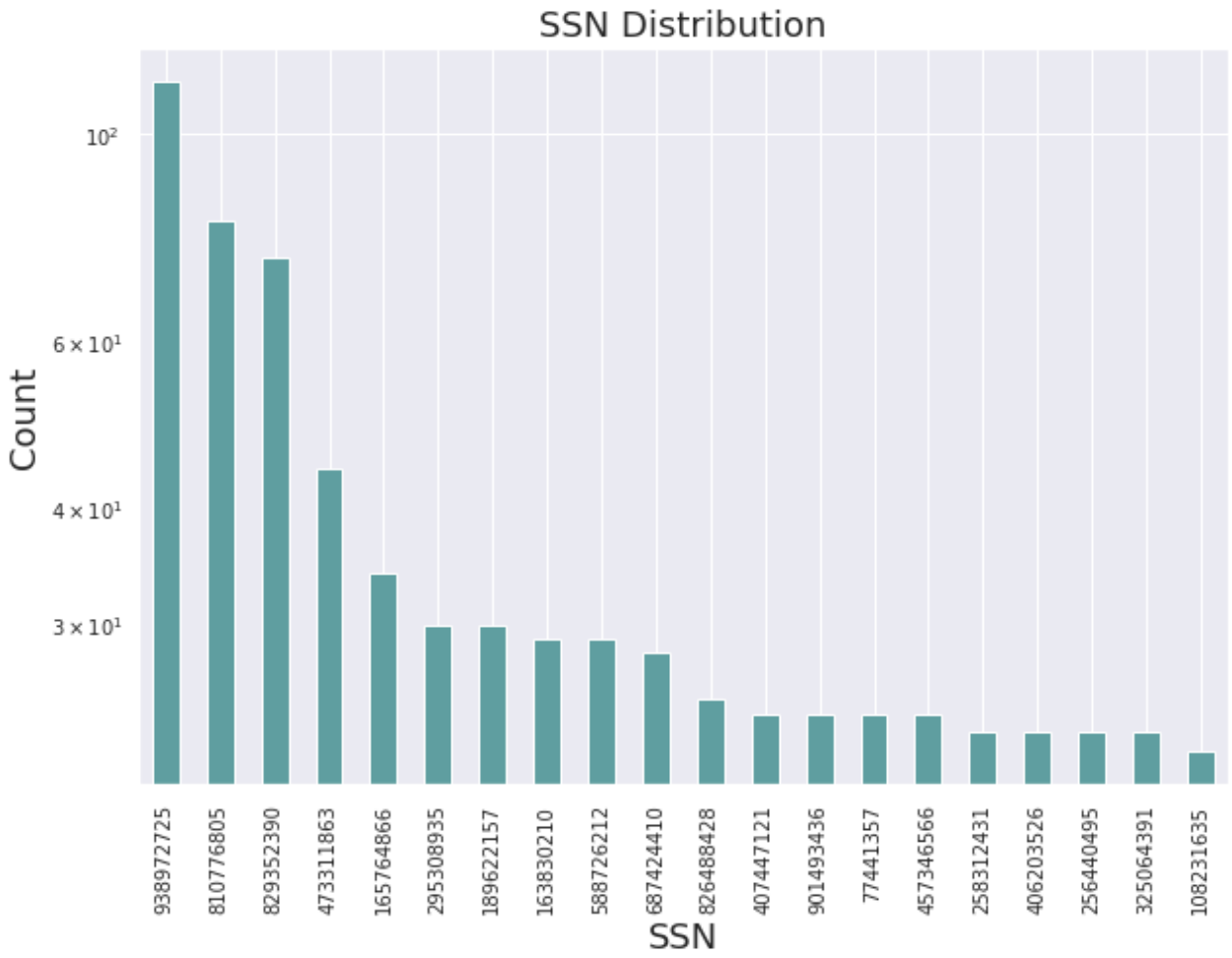
The date of the application is recorded for each application. From this, we can observe the frequency of the applications coming in by months.





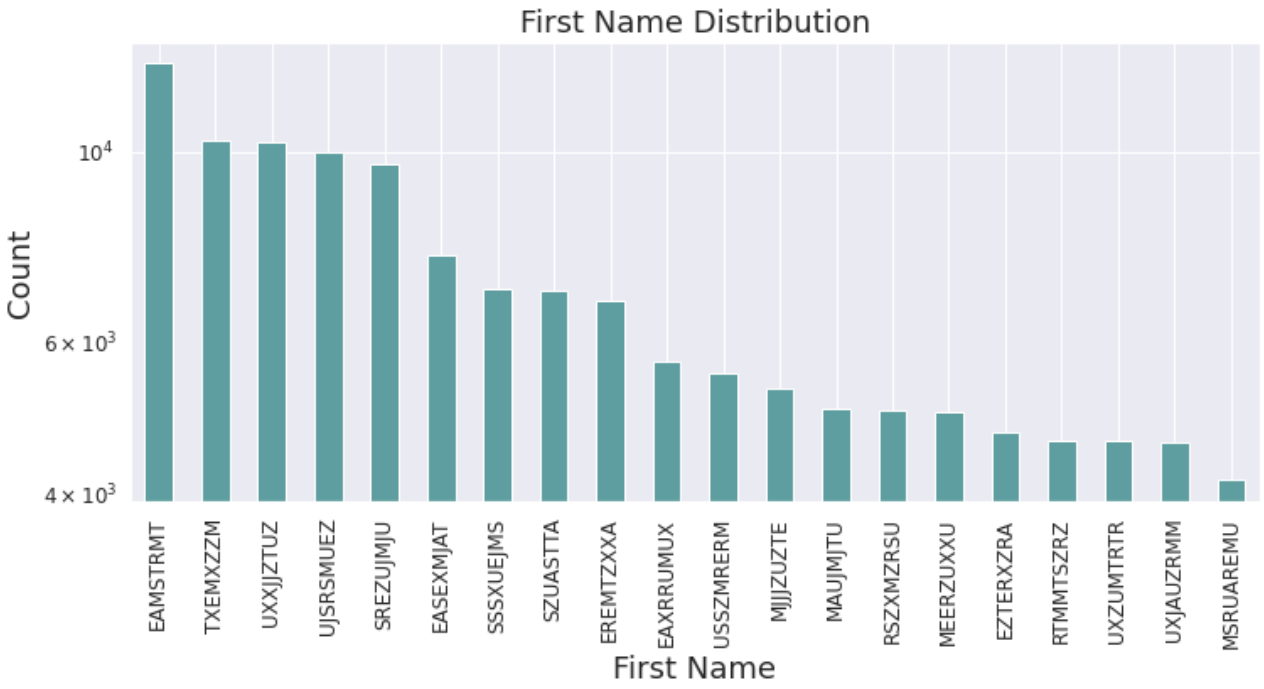
c. **SSN**

SSN records each applicant's social security number. The most common value is 999999999 with 16,935 records, which is possibly a default number when a customer chooses not to fill in their SSN. There are certainly other SSN numbers that are used to apply for multiple applications as shown in the chart below. The chart shows the 20 most common values of SSN, among a total of 835,819 unique SSN.



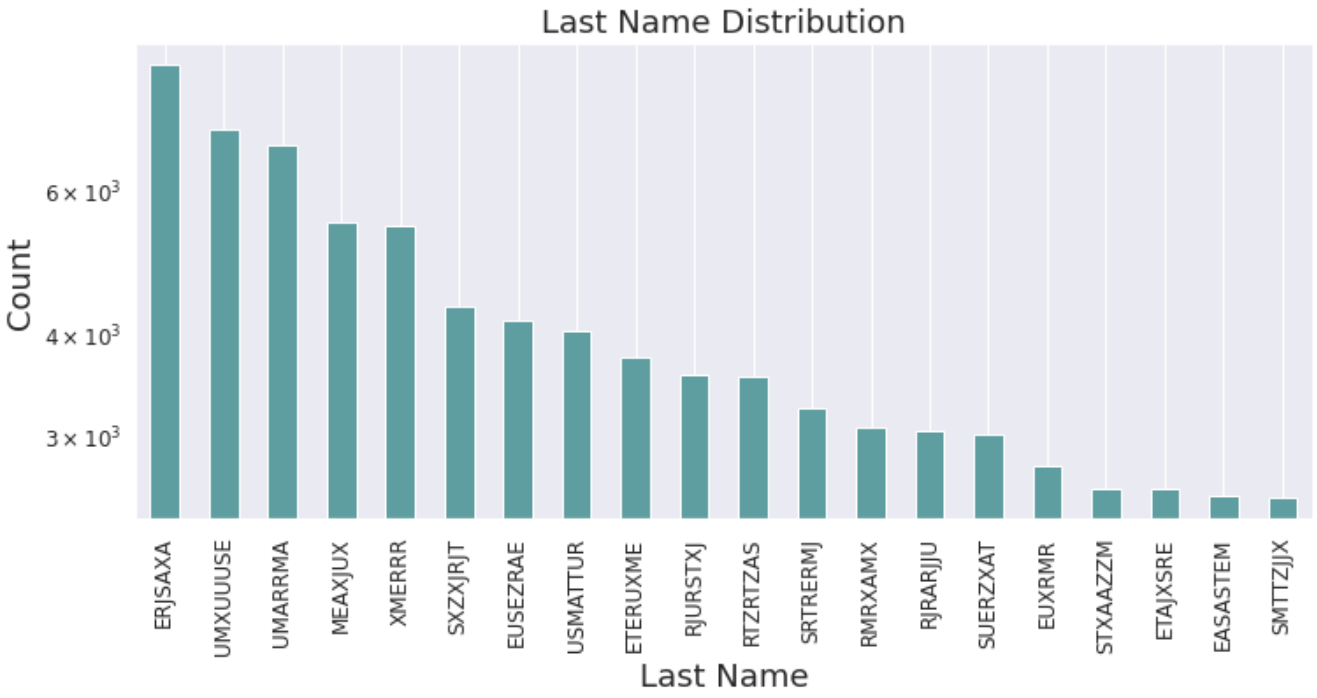
d. **Firstname**

The most common first name is EAMSTRMT with 12,658 records. It could be a good indicator of fraud identity when combining with other information. Yet, having the same first name is not uncommon. This chart shows the 20 first names with the highest count, among a total of 78,136 unique first names.



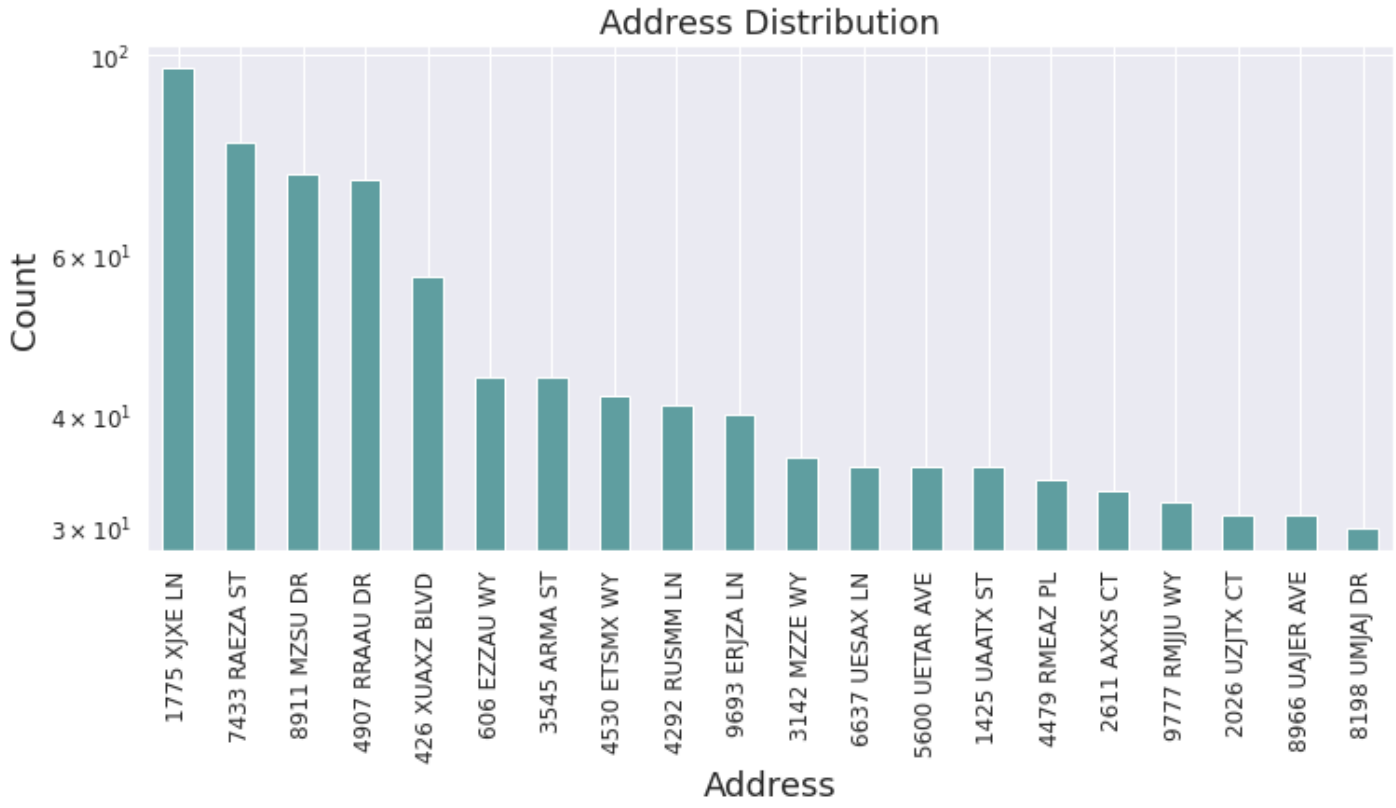
e. **Lastname**

The most common last name is ERJSAXA with 8,580 records. This chart shows the 20 last names with the highest count, among a total of 177,001 unique last names. Last name itself is not an accurate factor to evaluate fraud. However, the combination of last name, first name and other fields could potentially do it.



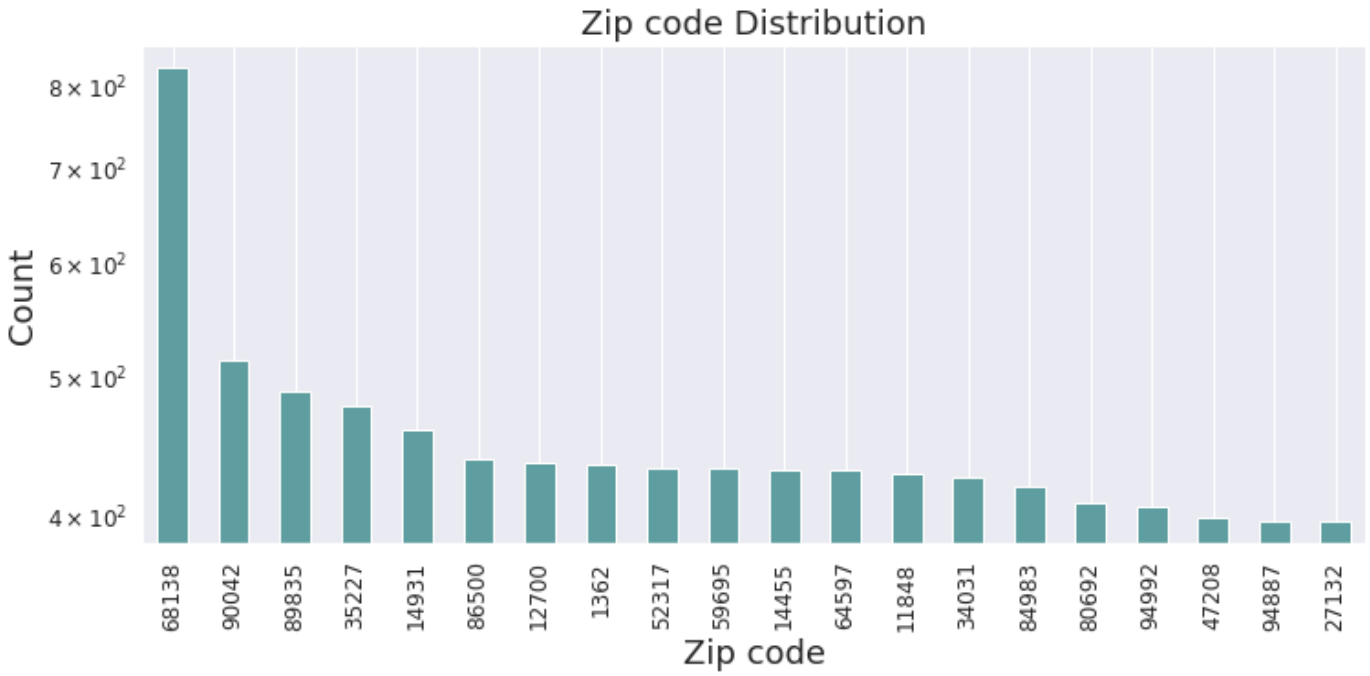
f. **Address**

123 Main St is the most common value with 1,079 records. This again could be a default value when a customer chooses not to fill in their address. This chart shows the 20 addresses with the highest count, among a total of 828,774 unique addresses.



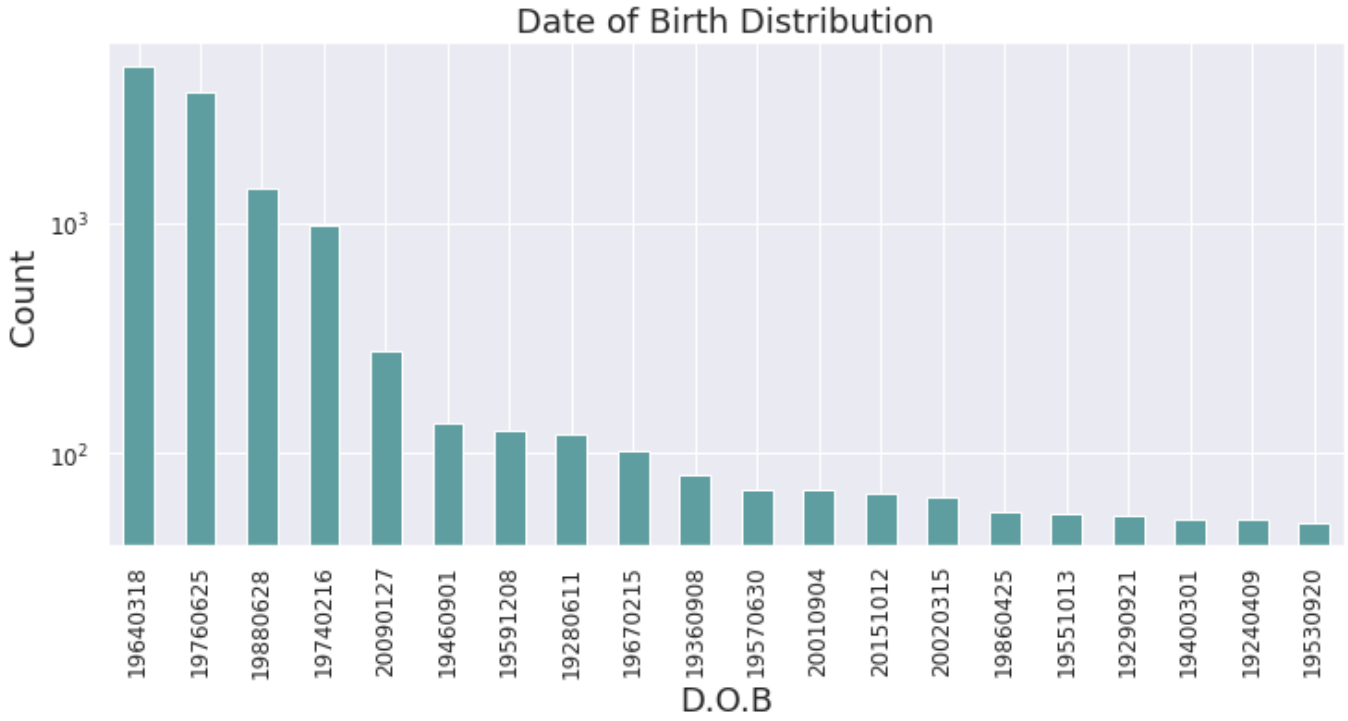
g. **Zip5**

Zip code ‘68138’ is the most common value with a total of 823 records. This chart shows the 20 zip codes with the highest count, among a total of 26,370 unique zip codes.



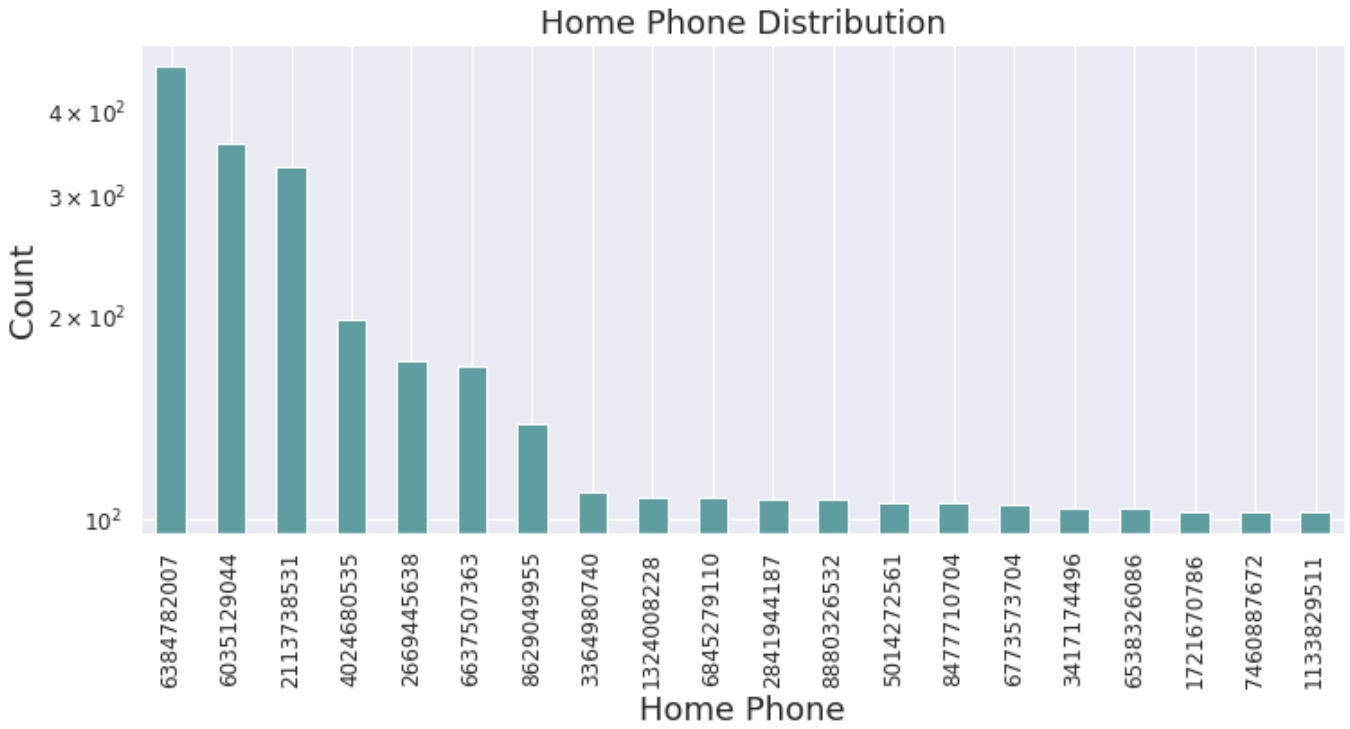
h. **DOB**

‘19070626’ is the most common value with 126,568 records. This again could be a default value when a customer chooses not to fill in their date of birth. This chart shows the 20 addresses with the highest count, among a total of 42,673 unique addresses.



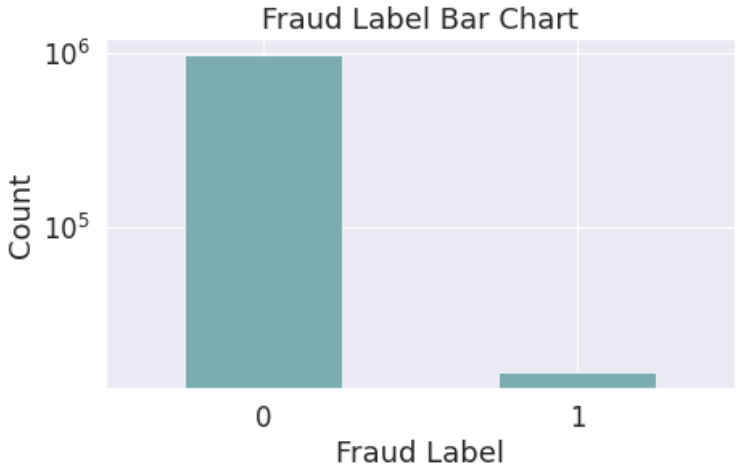
i. **Homephone**

The most common value of home phone number is 999999999 with 78,512 records, which is possibly a default number when a customer chooses not to fill in their home phone number. There are certainly other home phone numbers that are used to apply for multiple applications. This chart shows the 20 home phone numbers with the highest count, among a total of 28,244 unique addresses.



j. **Fraud_label**

Fraud value is a Boolean field with a value of 0 or 1, where 1 represents a potential fraud and 0 means it is not a potential fraud. The percentage of the applications identified as a potential fraud is around 1.4%.



Below charts demonstrates monthly frequency of ‘good’ and ‘bad’ applications; bad indicates a potential fraud and good indicates otherwise. In the chart, red line represents the frequency of a ‘bad’ application and green line represents the frequency of a ‘good’ application. ‘Good’ applications appear to have a relatively steady frequency while ‘bad’ applications tend to fluctuate more.