Data Quality Report

High-level description:

This is a data quality report on a dataset provided by the stakeholder for identity fraud analysis. The dataset contains identity data used by individuals to make a purchase or other transactions from January 1st, 2016, to December 31st, 2016. It has 1,000,000 rows and 10 fields.

Summary statistics table:

Field Name	% Populated	# Unique Values	Most Common Value
record	100	1,000,000	NA
date	100	365	8/16/2016
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
dob	100	42,673	6/26/1907
homephone	100	28,244	9999999999
fraud_label	100	2	0

Table 1. Statistical summary of the dataset.

Individual field summary:

1. Date:

This field indicates the date on which the individual has raised a request for a transaction.

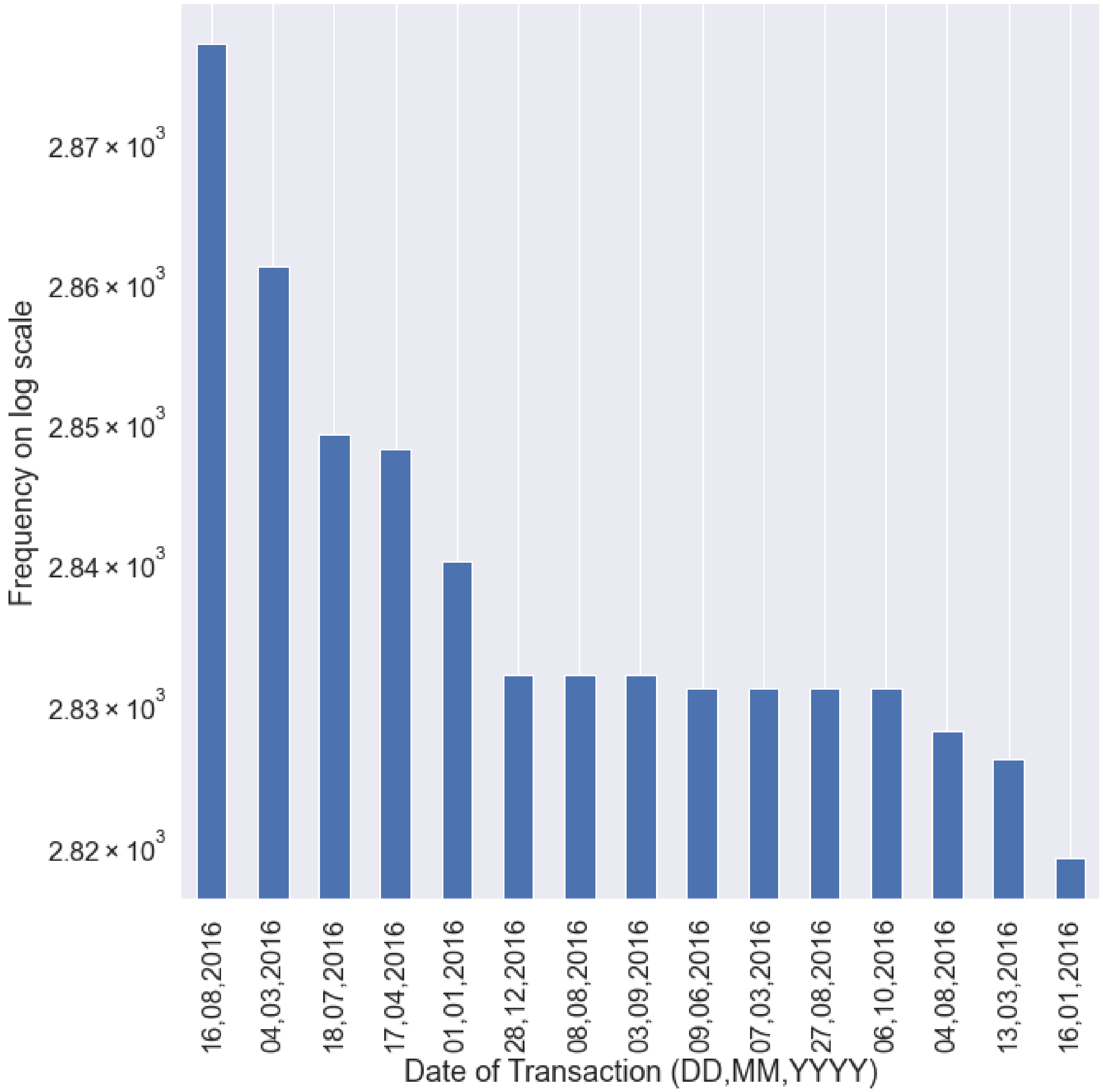

Fig. 1. Represents the first 15 dates with the highest entries on a logarithmic scale.

2. SSN:

This field indicates the SSN number of the individual who has raised a request for a transaction.

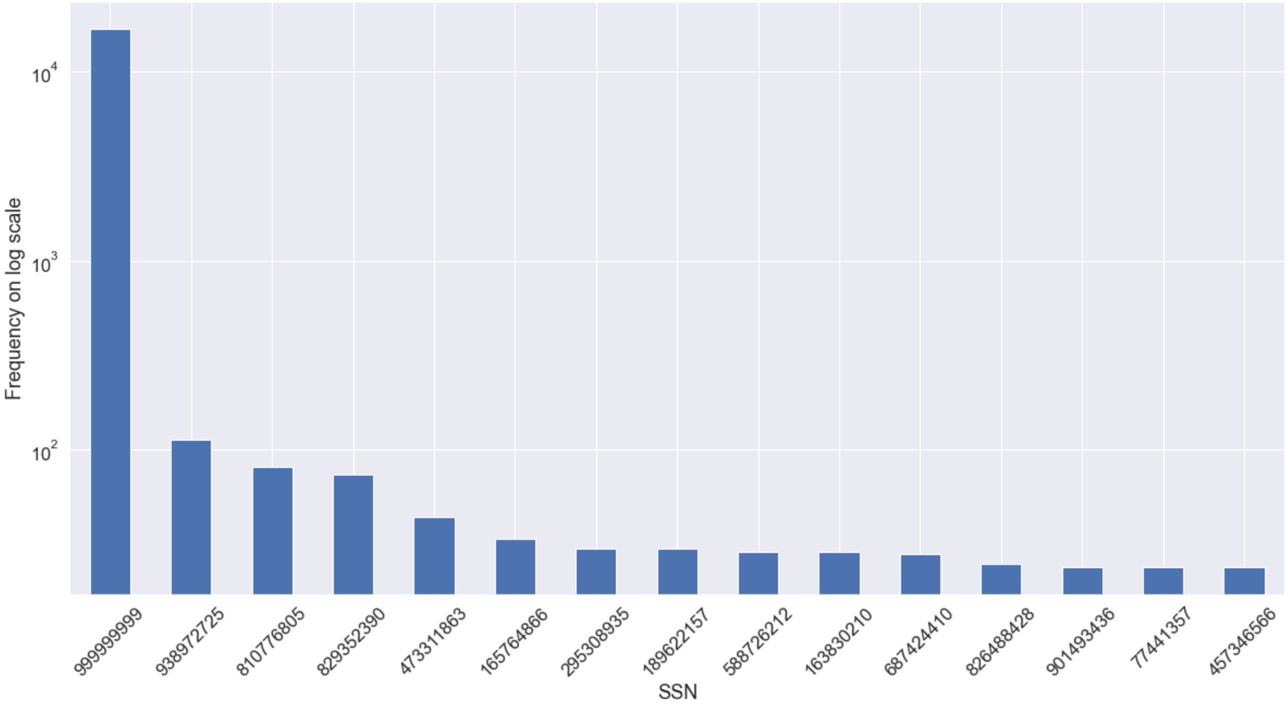

Fig. 2. Represents the first 15 SSNs with the highest entries on a logarithmic scale.

3. First name:

This field indicates the first name of the individual who has raised a request for a transaction.

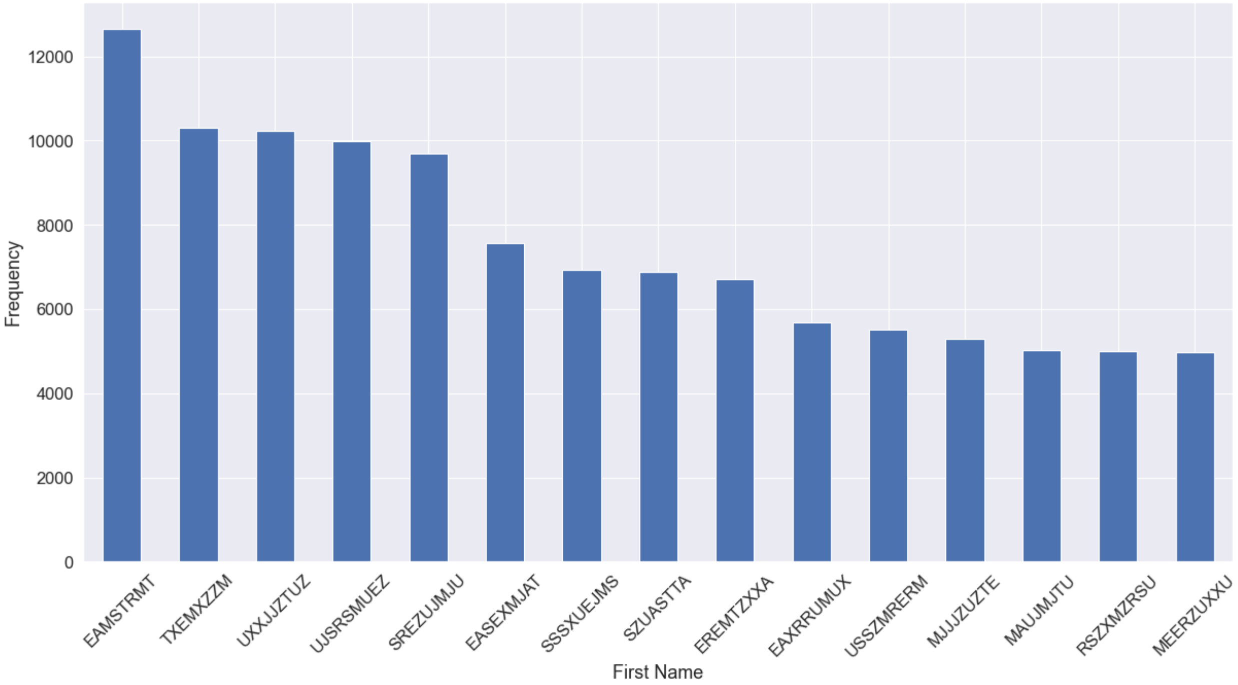

Fig. 3. Represents the first 15 first names with the highest entries in the dataset.

4. Last name:

This field indicates the last name of the individual who has raised a request for a transaction.

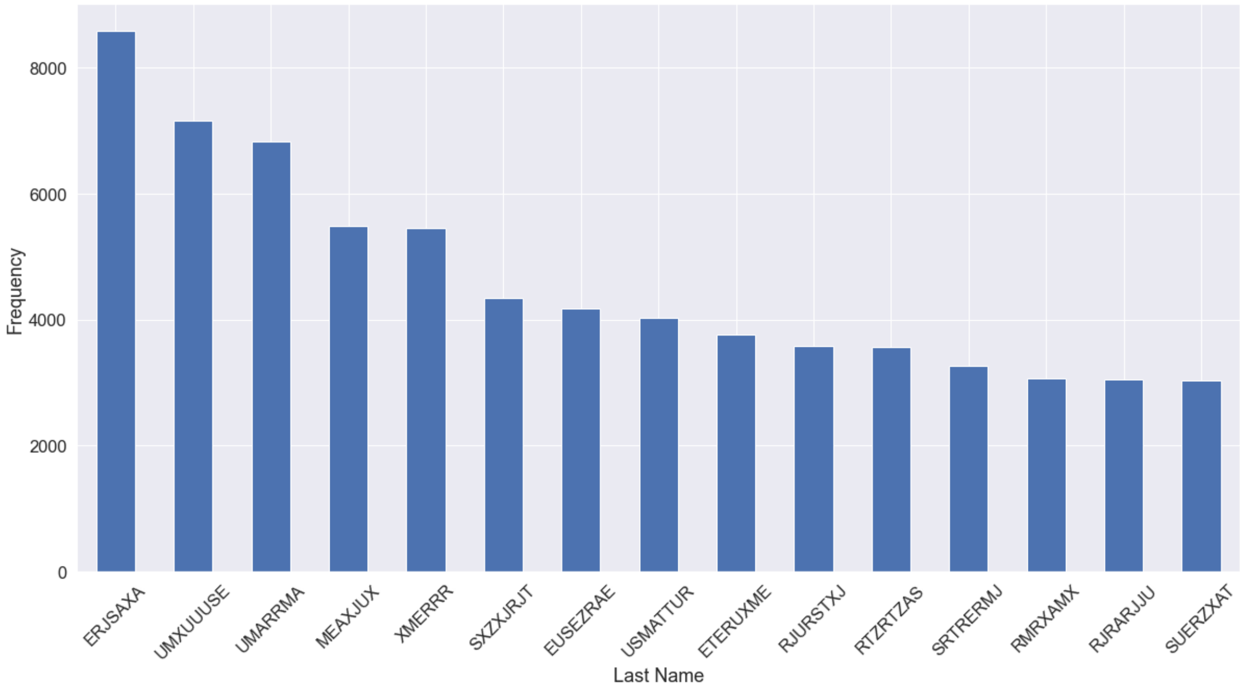

Fig. 4. Represents the first 15 last names with the highest entries in the dataset.

5. Address

This field indicates the address of the individual who has raised a request for a transaction.

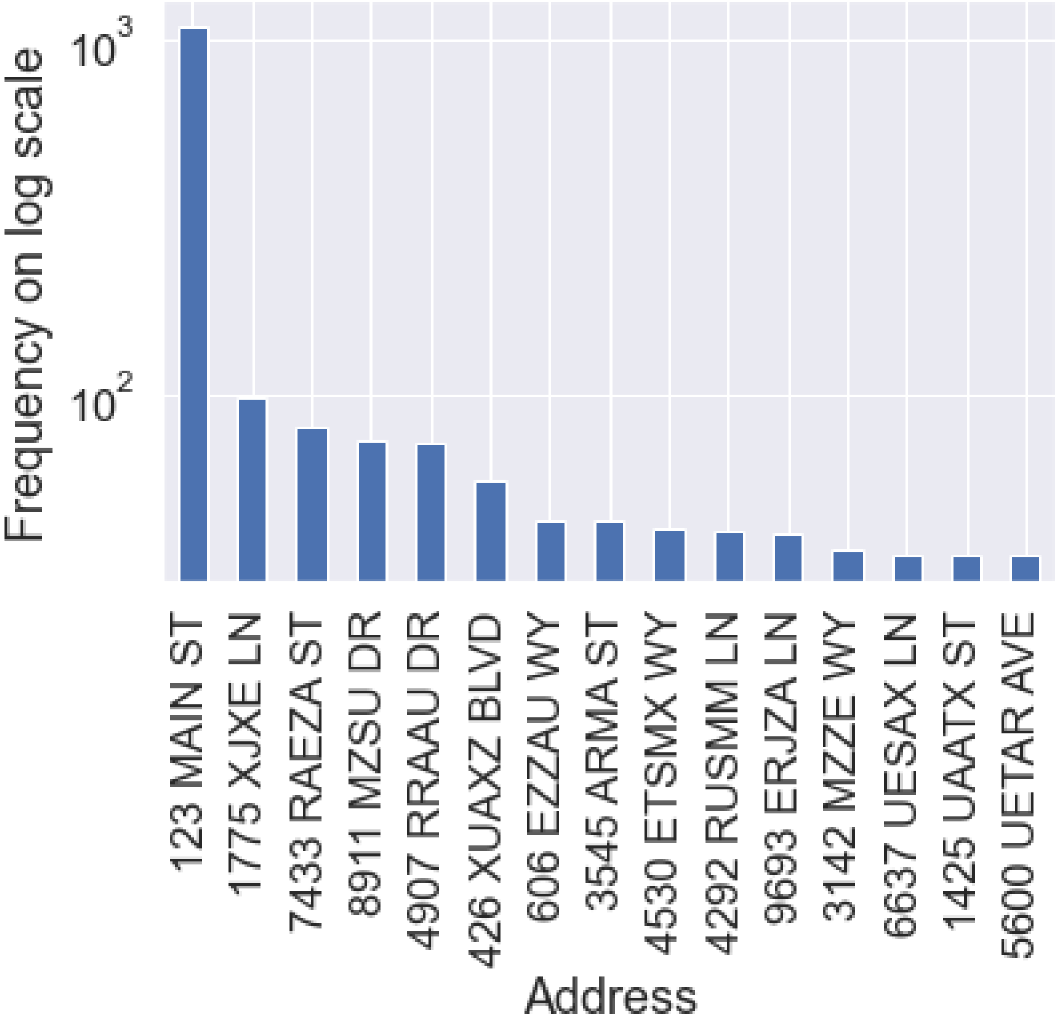

Fig. 5. Represents the top 15 most used addresses on a logarithmic scale.

6. Zip5

This field indicates the five-digit zip code of the individual making the transaction.

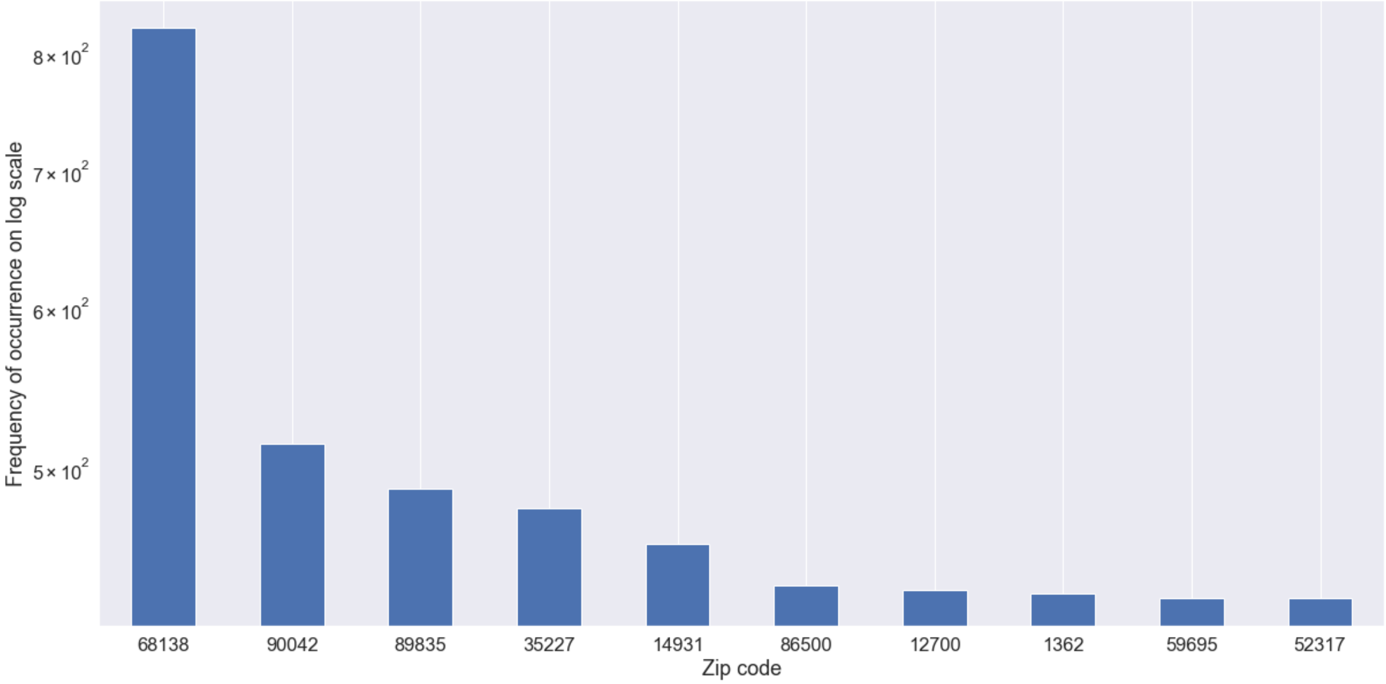

Fig. 6. Represents the top 10 most used zip codes on a logarithmic scale.

7. DoB:

This field indicates the date of birth of the individual making the transaction.

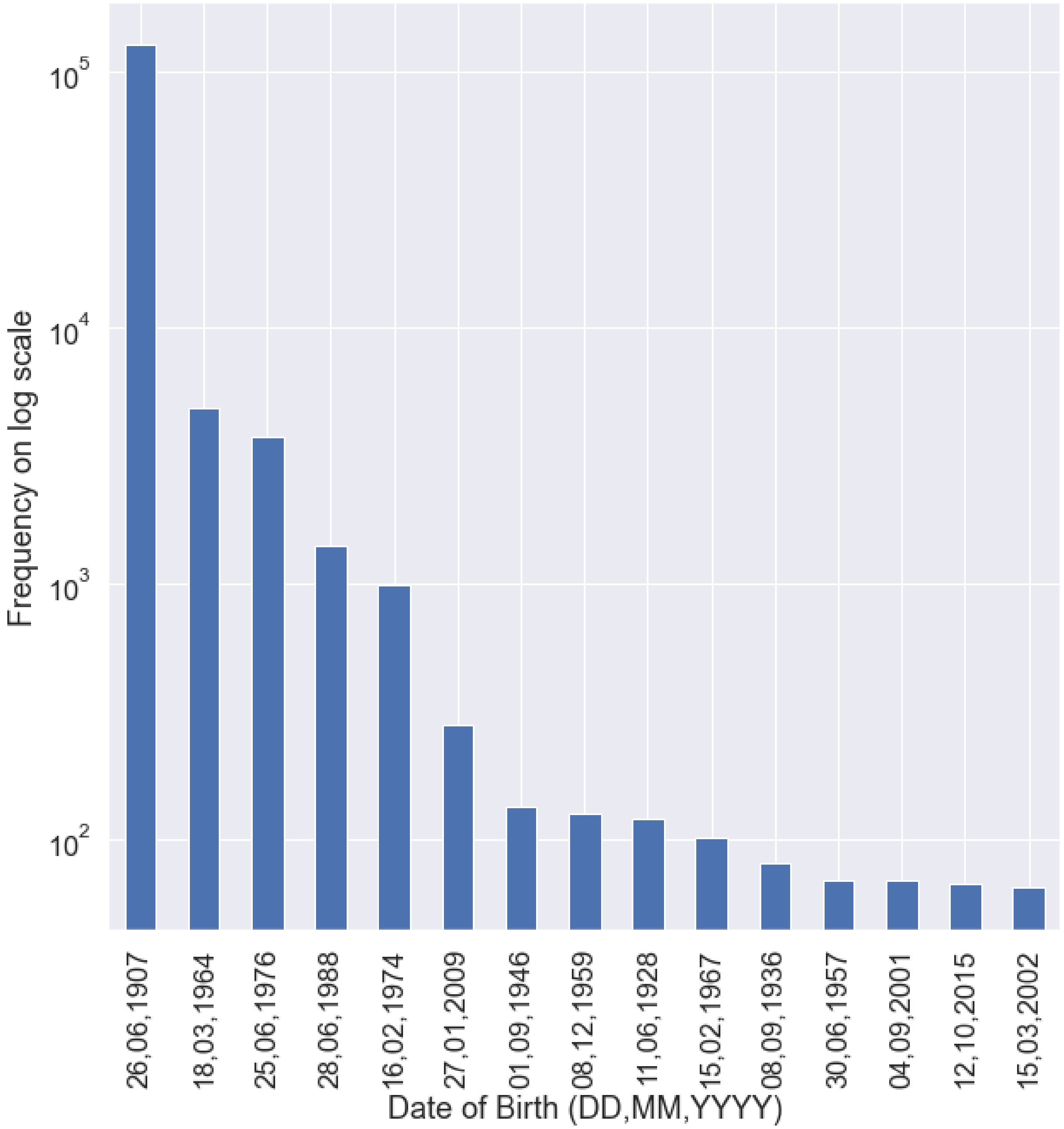

Fig. 7. Represents the top 15 most frequently occurring date of births on a logarithmic scale

8. Homephone:

This field indicates the home phone number of the individual making the transaction.

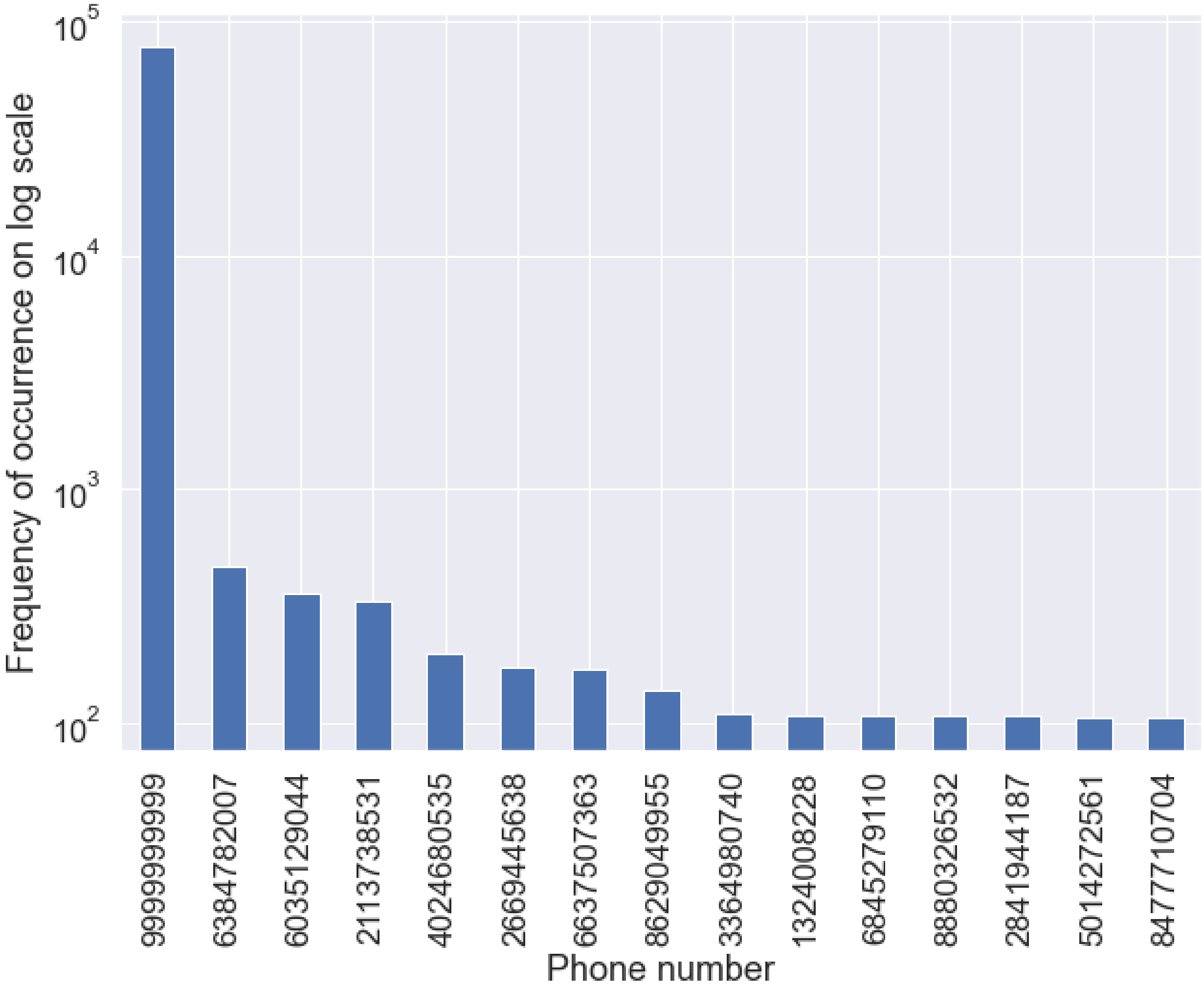

Fig. 8. Represents the top 15 most frequently home phone numbers on a logarithmic scale

9. Fraud Label,

This field indicates whether the entity is a fraudulent activity or not. 1 means that the entity is a fraud activity and 0 represents that the entity is not a fraud activity.

Value	Number of Occurrence	Percentage of total population
1	985607	98.56%
0	14393	1.44%

Table 2. Represent the frequency of unique fraud_label values.

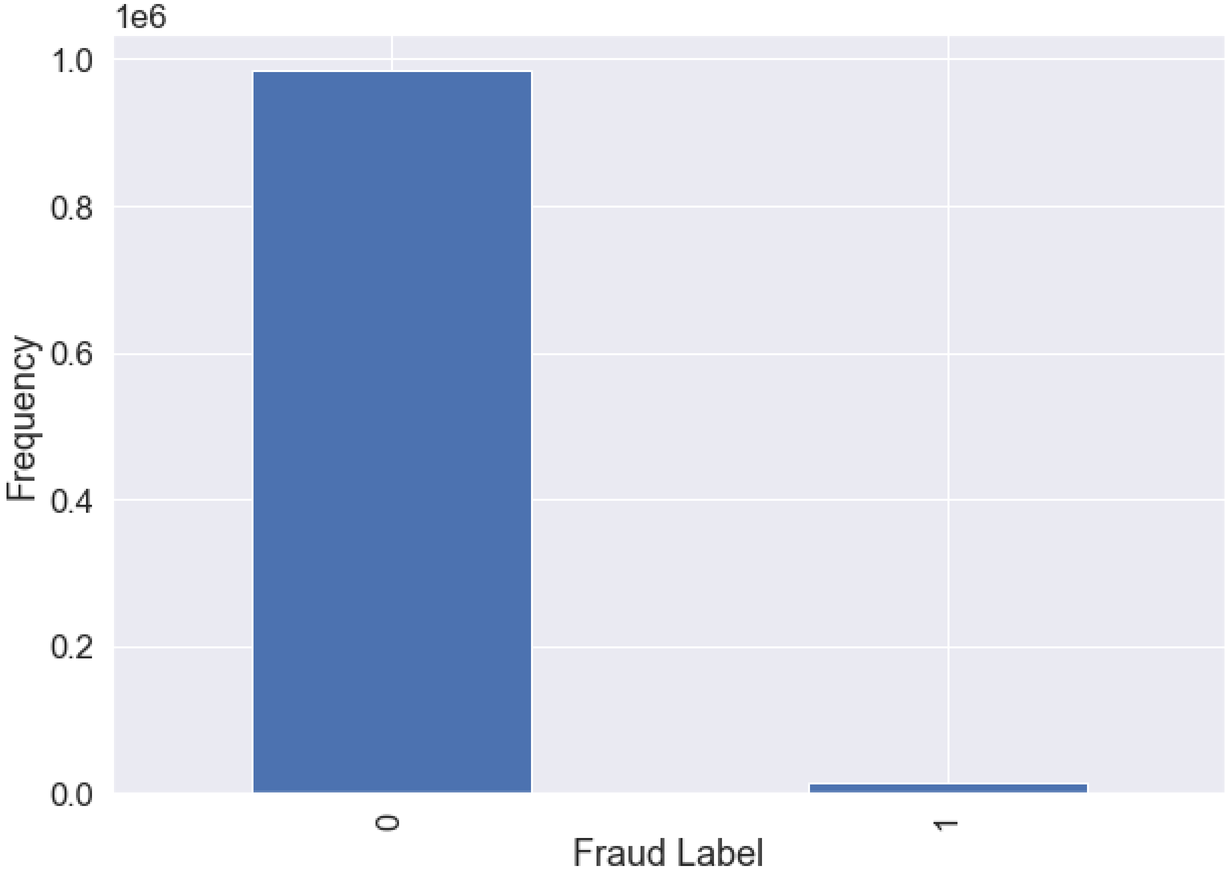

Fig. 9. Represents the frequency of unique fraud_label values.