Identity Fraud Identification

Data Quality Report
For: Dr. Stephen Coggeshall

By:
Shreerang Javadekar | Vandik Zaveri | James Lee
Liana Liang | Yadi Gong | Wendy Mu

Index:

1. Preface

1.1 Introduction

Following is a data quality report for Dr. Stephen Coggeshall, Professor at the University of Southern California, and owner of the dataset of "Applications Dataa" for Fraud Identification.

The intent of this report is to align with the owner of the dataset the primary exploratory analysis done by Mr. Shreerang Javadekar to proceed further with building a supervised learning model to help identify fraudulent transactions.

1.2 Brief on the Dataset

The dataset on hand consists of 1,000,000 applications made using 835,819 unique Social Security Numbers. The uniqueness of these 835,819 Social Security Numbers, however, is questionable.

We have a total of 10 unique fields in the dataset (including record number) that describe personal identification details of every single application – details including name, Social Security Number, Address, Zipcode and Phone Number. Details on the date when the application was made are also available. A label is also available that indicates whether the application is fraudulent or not.

Every field is 100% filled. There are no blank entries in the dataset. Further details have been mentioned in the Summary Table.

2. Summary Tables

Below is a summary table for all the fields in the dataset. All the fields have been treated as categorical.

Total number of records: 1,000,000
Total number of fields: 10.
Unique Identifiers for Every Recods: Record Number (Range: 1-1,000,000).

Field/Metric	# Records with a Value	% Populated	# Unique Categories	Most Common Value
Record	1,000,000	100%	1,000,000	Every value appears just once.
Date (date)	1,000,000	100%	365	“2016-08-16”
Social Security Number (ssn)	1,000,000	100%	835,819	“999-99-9999”
First Name (firstname)	1,000,000	100%	78,136	“EAMSTRMT”
Last Name (lastname)	1,000,000	100%	177,001	“ERJSAXA”
Address (address)	1,000,000	100%	828,774	“123 MAIN ST”
Zipcode (zip5)	1,000,000	100%	26,370	“68138”
Date of Birth (dob)	1,000,000	100%	42,673	“1907-06-26”
Phone Number (homephone)	1,000,000	100%	28,244	“999-999-9999”
Fraud Label (fraud_label)	1,000,000	100%	2	“0”

3. Exploratory Analysis

Below is an exploratory analysis for every single field in the dataset. Adequate visualizations have been added to supplement the description.

3.1 Record (record)

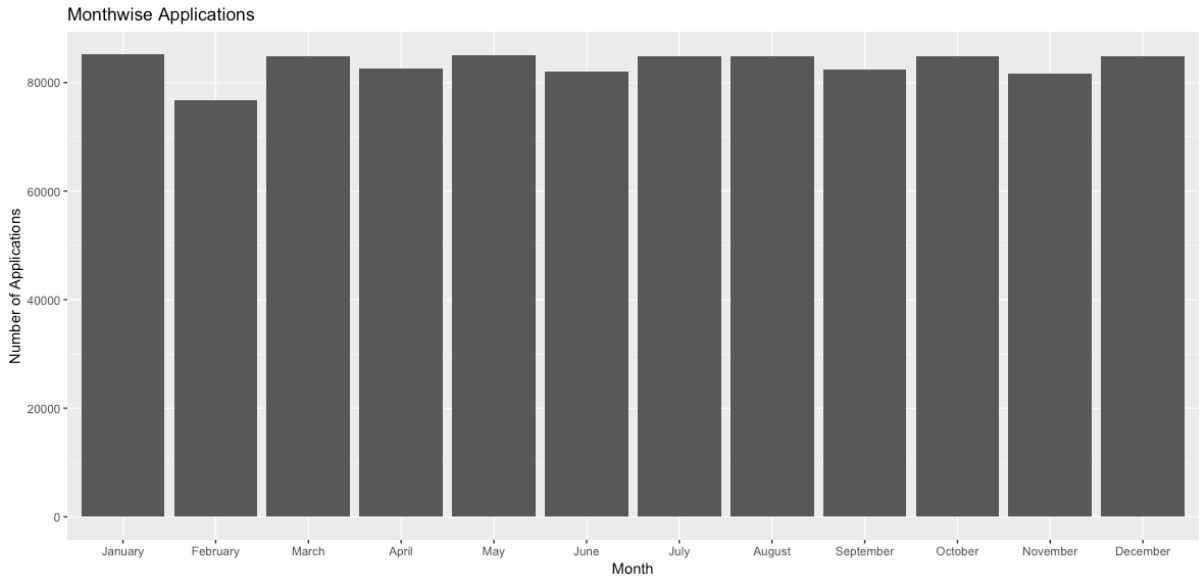
“record” is a unique identifier for every entry. The record indexing is numerical. It starts at 1 and ends at 1000000 with step increases of 1.

3.2 Date (date)

"date" is an indicator for the date when the application was filed.

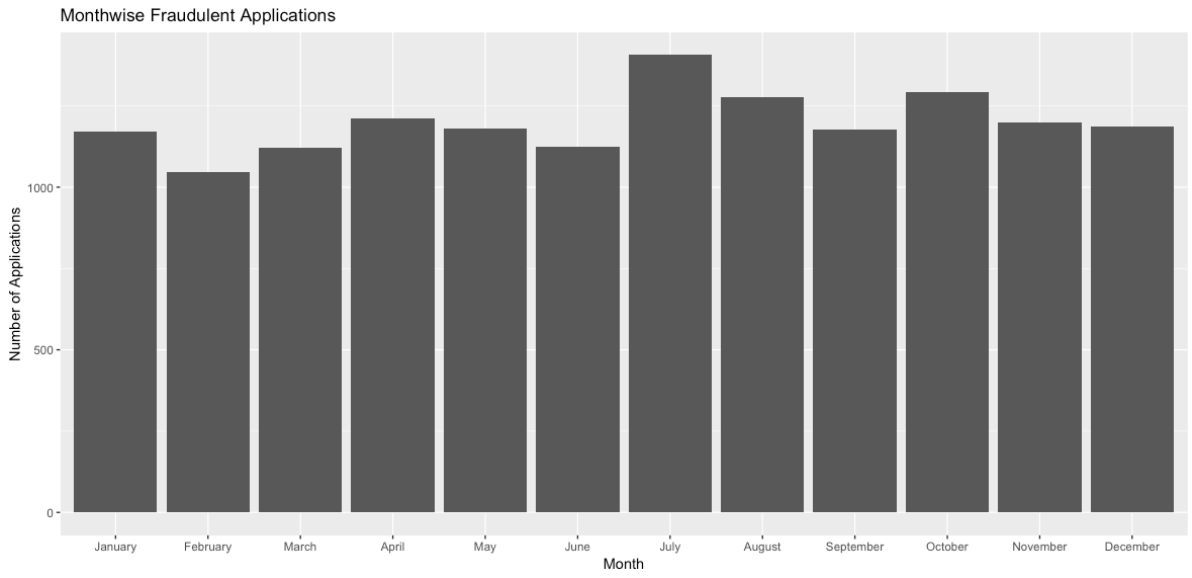
100% of this field is filled. The dates in this filed are the 366 days of the leap year 2016. However, there are no applications for the date of 02/29/2016.

A monthwise distribution of the number of applications (both fraudulent and non-fraudulent) is as below:

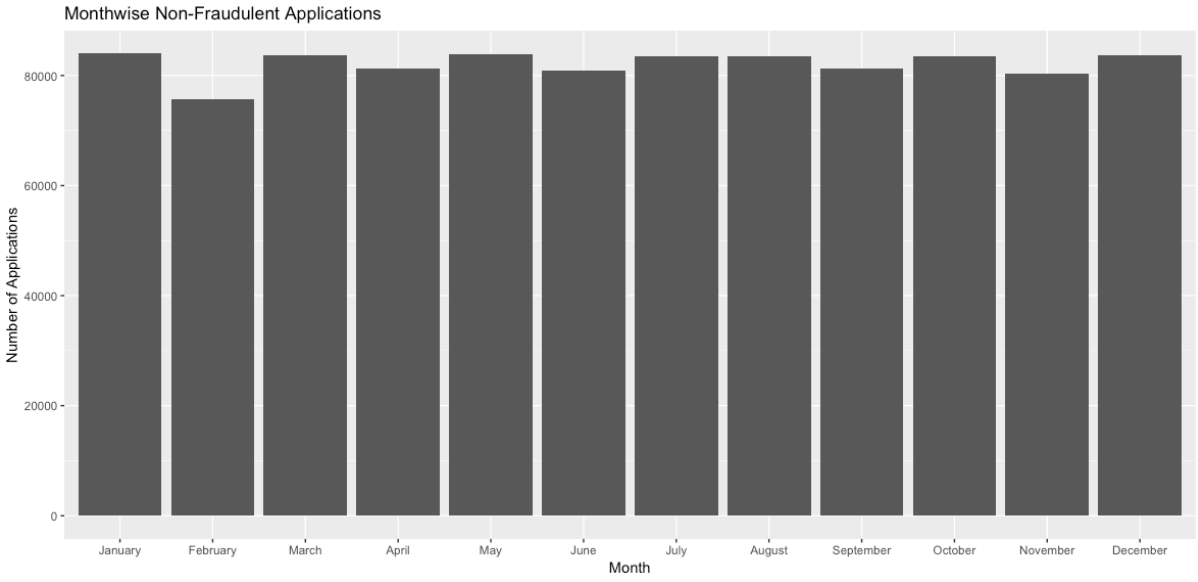


The distribution for number of applications seems to be uniform across all months with just a small dip in the month of February.

For fraudulent applications the distribution is as below:

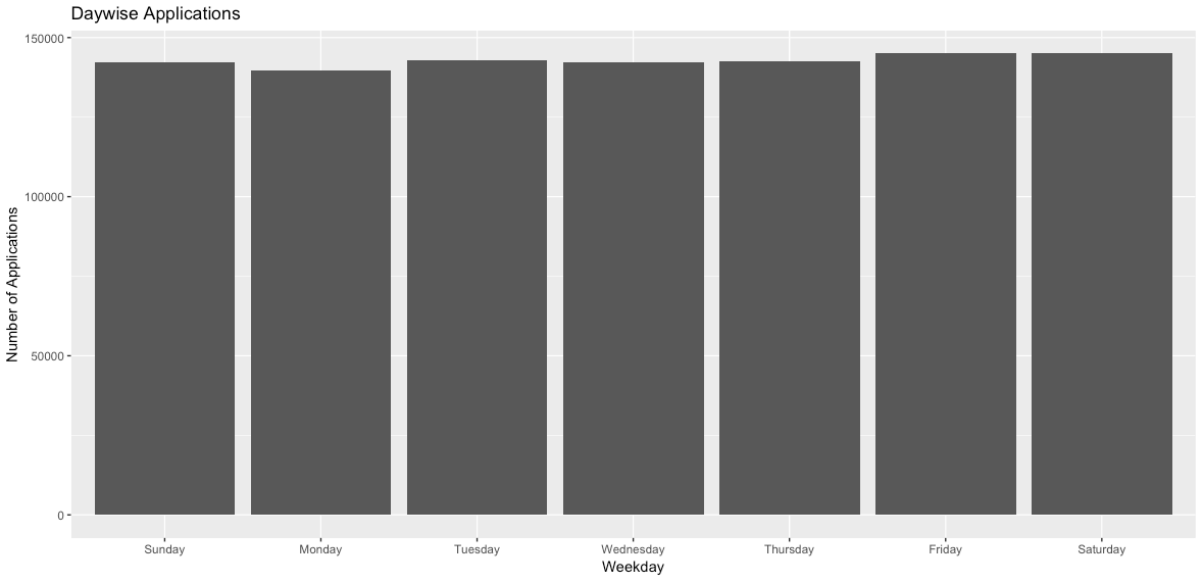


The month of July seems to have witnessed the highest number of fraudulent applications.

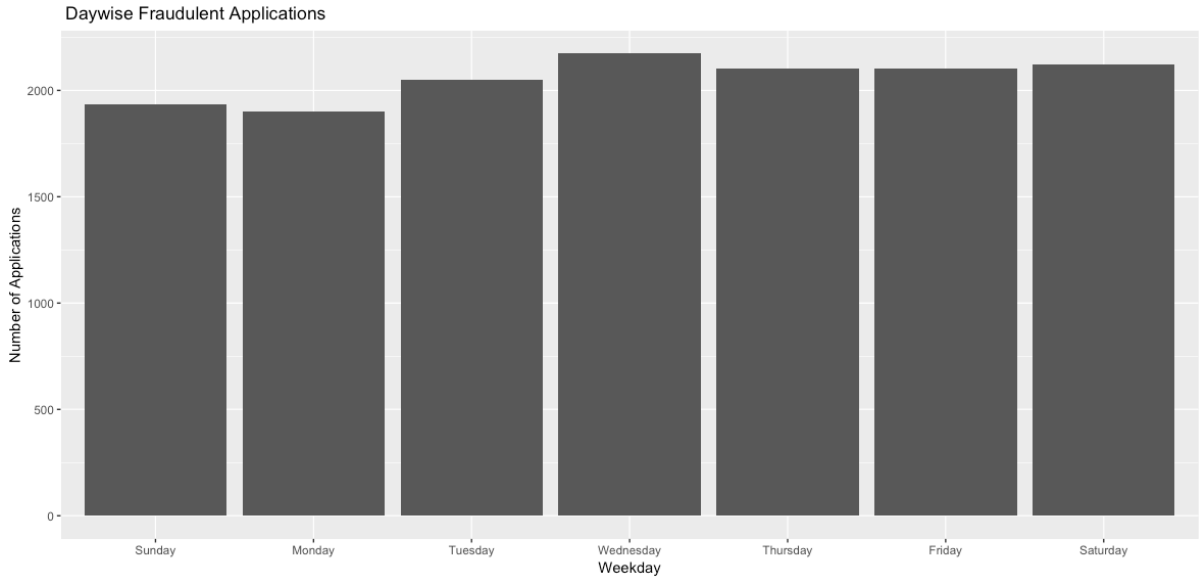


There is no significant trend in the non-fraudulent applications across months.

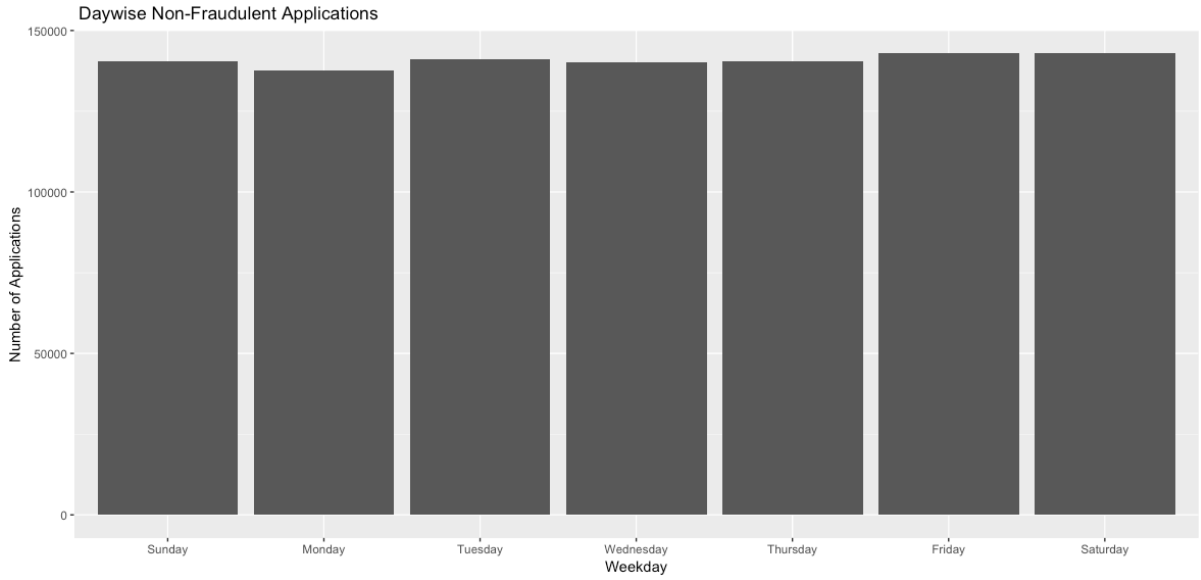
A daywise distribution of all the transactions is as follows:



For fraudulent applications, the daywise distribution is as below:

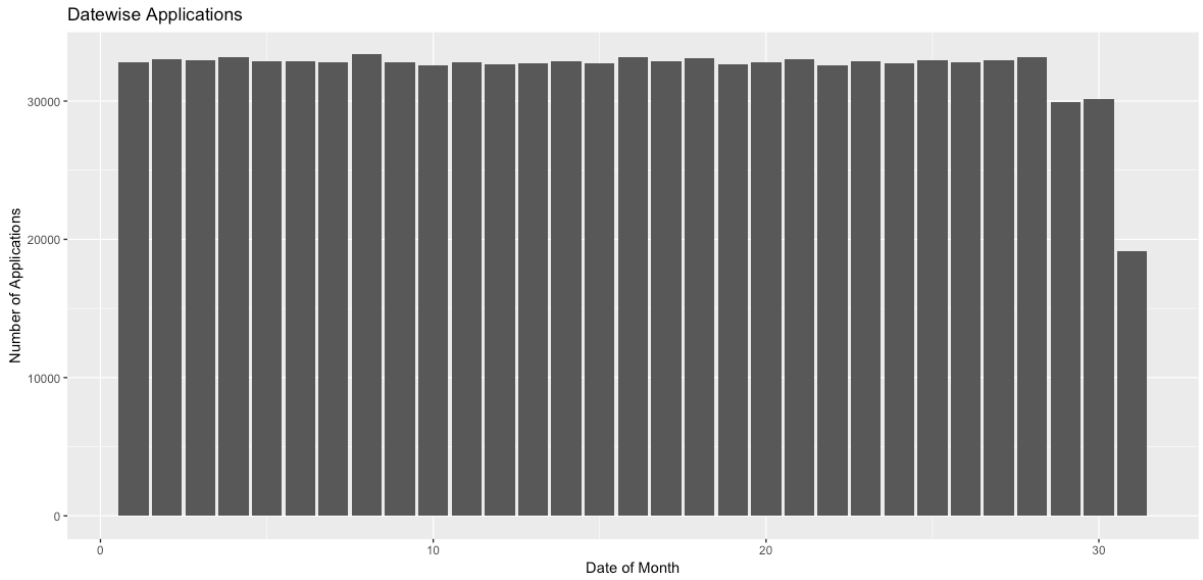


For non-fraudulent applications, the daywise distribution is as below:



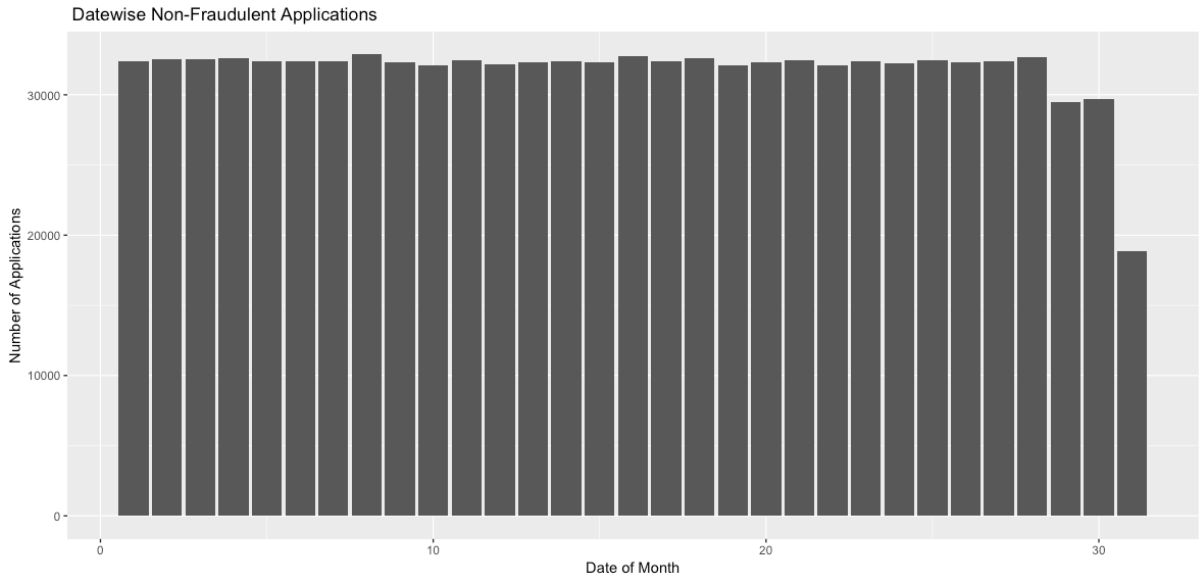
There is no significant trend in the daywise distribution across both fraudulent and non-fraudulent applications.

A datewise distribution of the number of applications (both fraudulent and non-fraudulent) is as below:

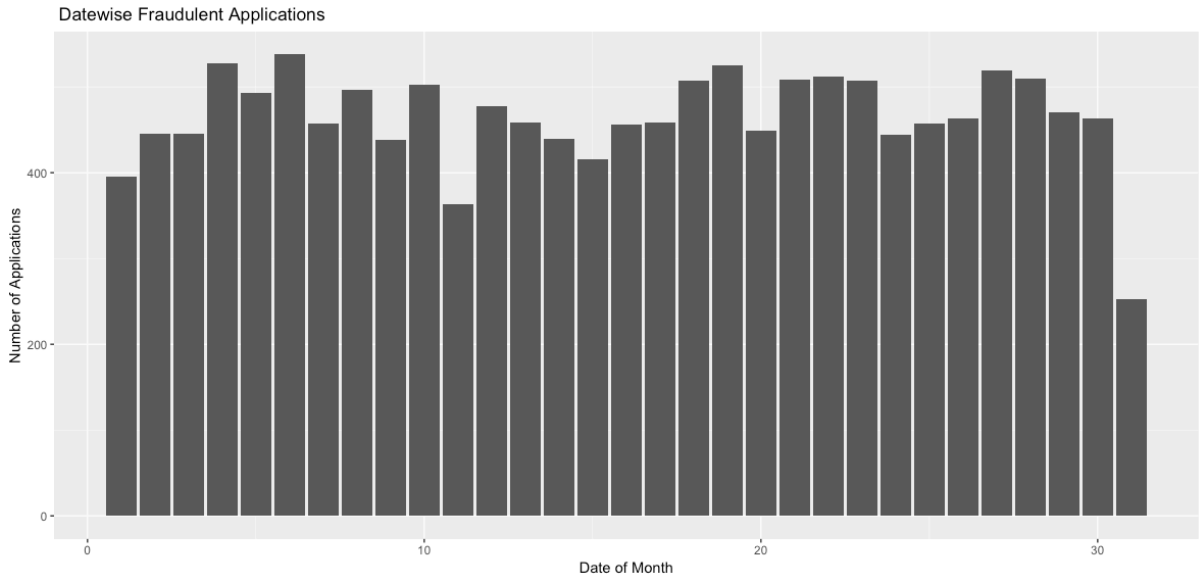


There seems to be a dip in the number of transactions nearing the month. For 31st, the number is very low because of the few months that have the 31st date.

For non-fraudulent applications, the distribution is as below:



For fraudulent applications, the distribution is as below:

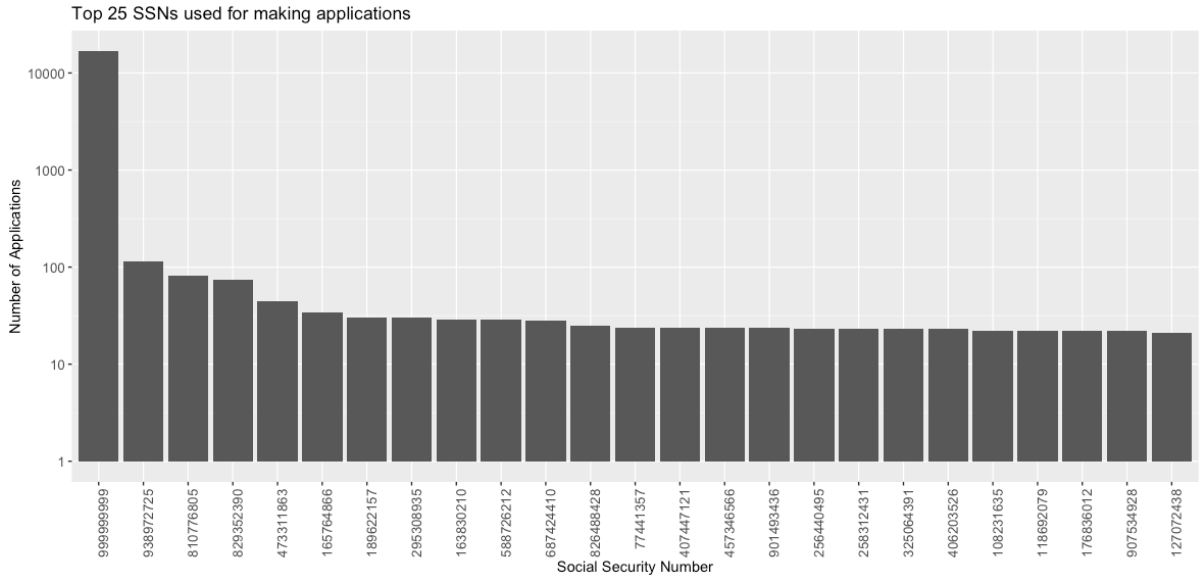


3.3 Social Security Number (ssn)

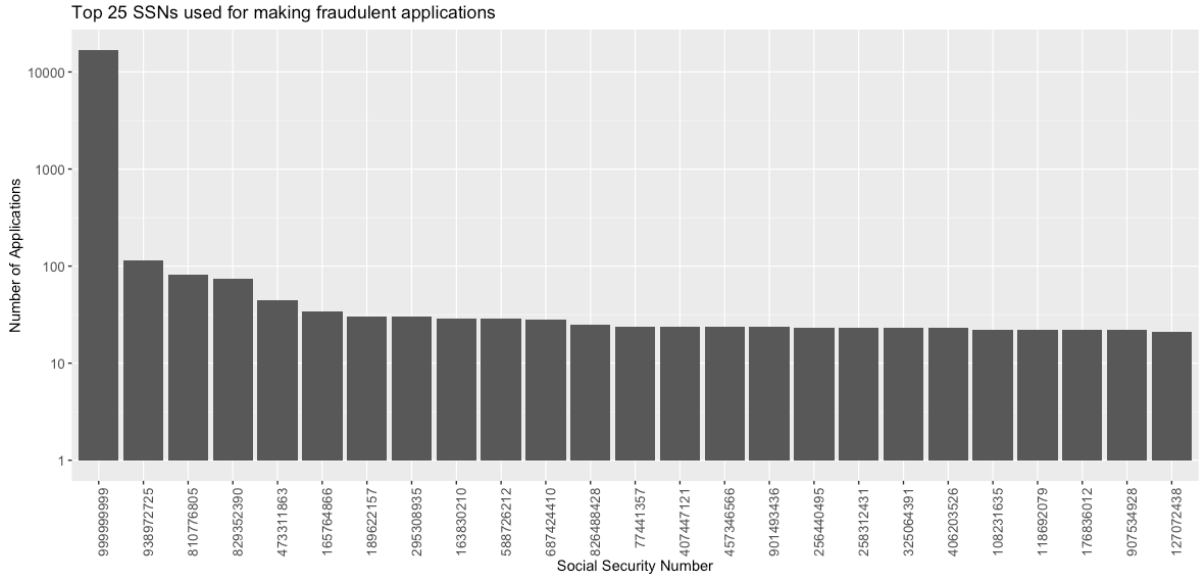
"ssn" is an indicator of the Social Security Number used for making the application.

There are 835,819 unique Social Security numbers in the dataset. Their uniqueness and authenticity, however, is questionable. 100% of the dataset is filled.

A distribution of the top 25 Social Security numbers used for making the applications is as follows. The Y-axis is in log scale.



The top 25 Social Security Numbers used for making fraudulent applications is as follows:



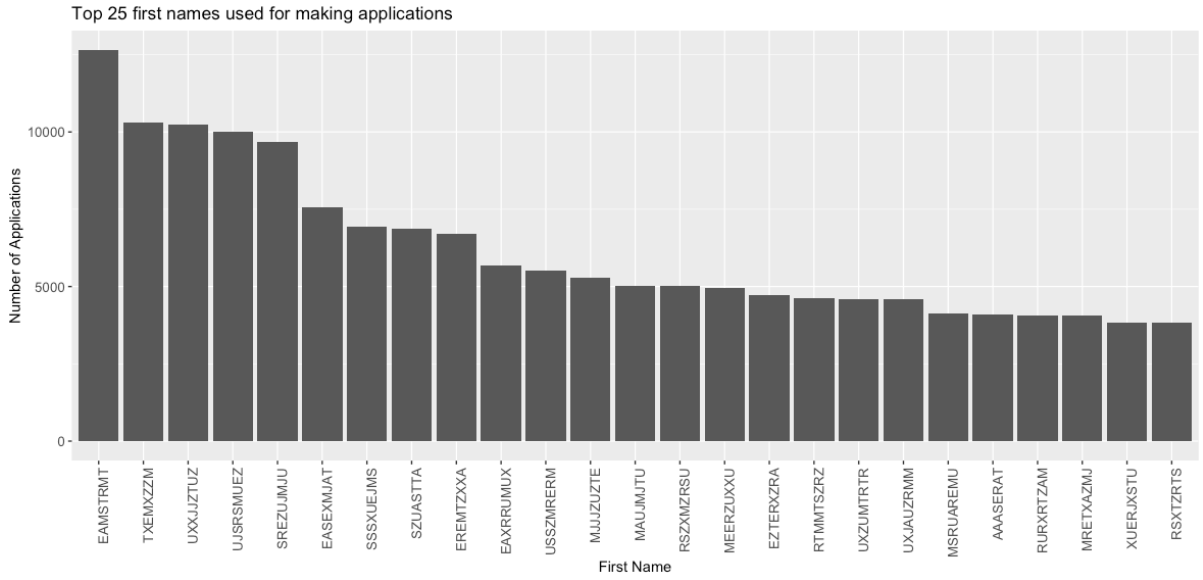
It is clear that the "999-99-9999" is a proxy number being used for making applications to mask one's identity.

3.4 First Name (firstname)

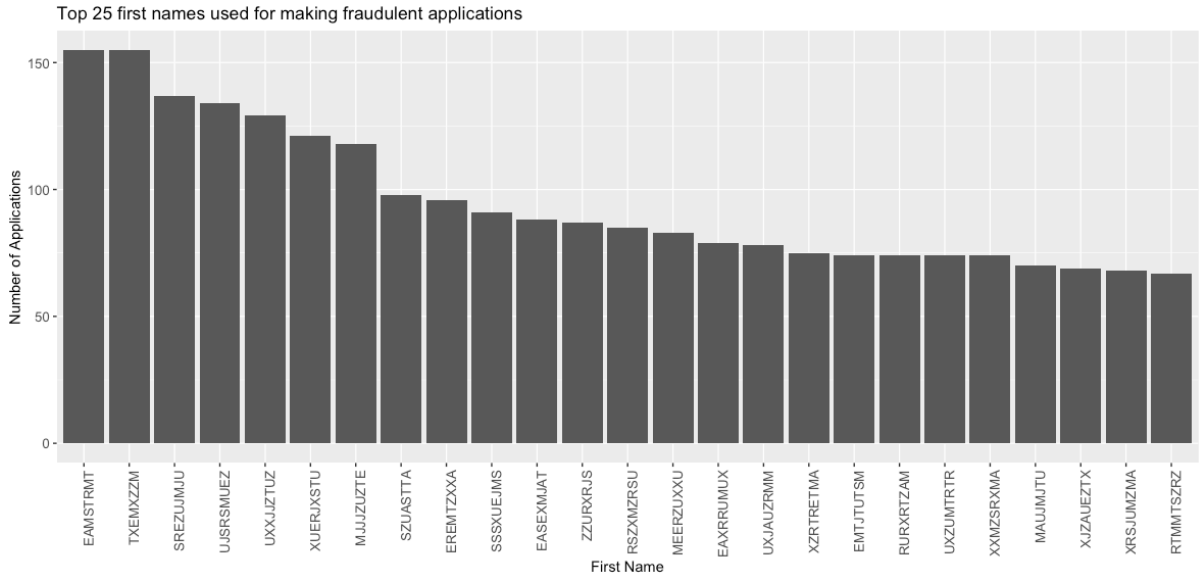
"firstname" is an indicator for the first name of the applicant.

There are 78,136 unique first names in the dataset. 100% of the fields are filled in this column.

A distribution of the top 25 first names with the highest number of applications is as follows:



A distribution of the top 25 first names with the highest number of fraudulent applications is as follows:

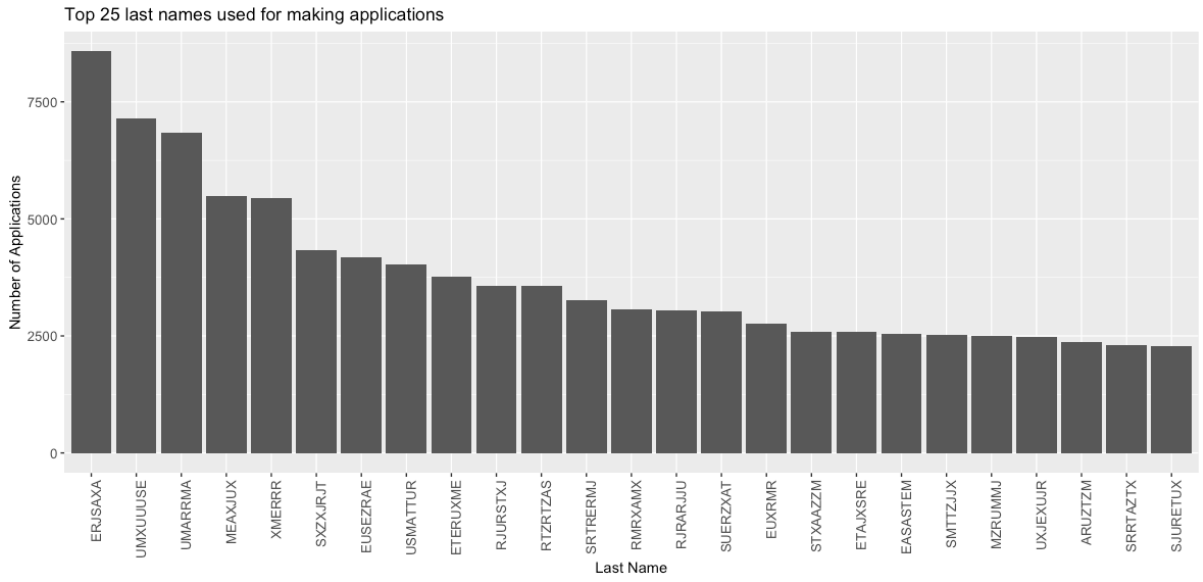


3.5 Last Name (lastname)

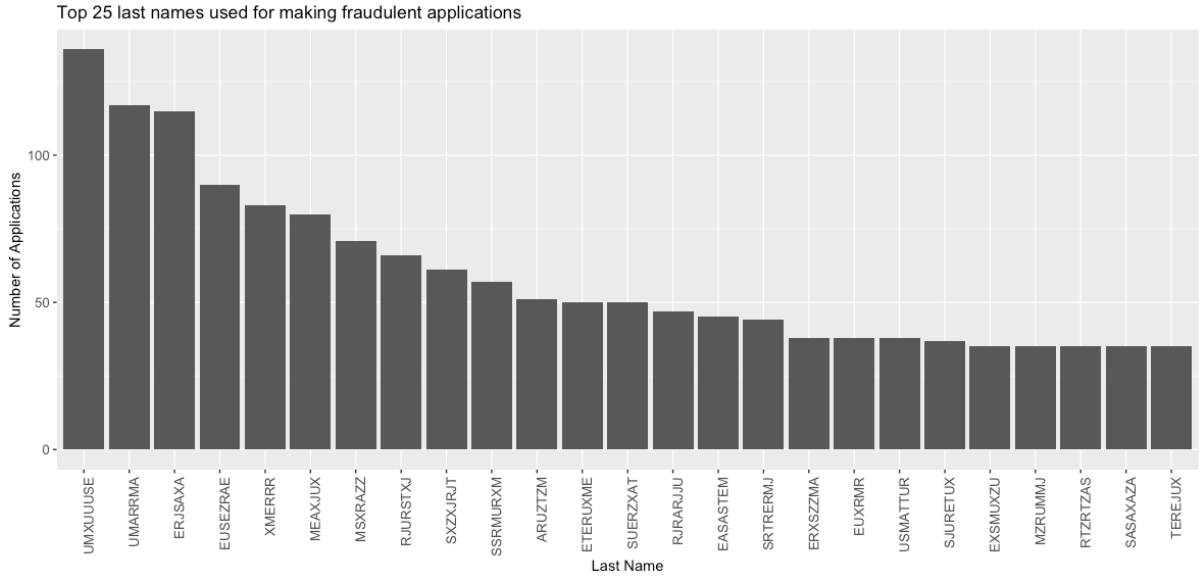
"lastname" is an indicator of the last name used while making the application.

There are 177,001 unique last names in the dataset. 100% of the dataset is filled

A distribution of the top 25 last names used for making the applications is as follows:



For fraudulent applications, the distribution is as below:

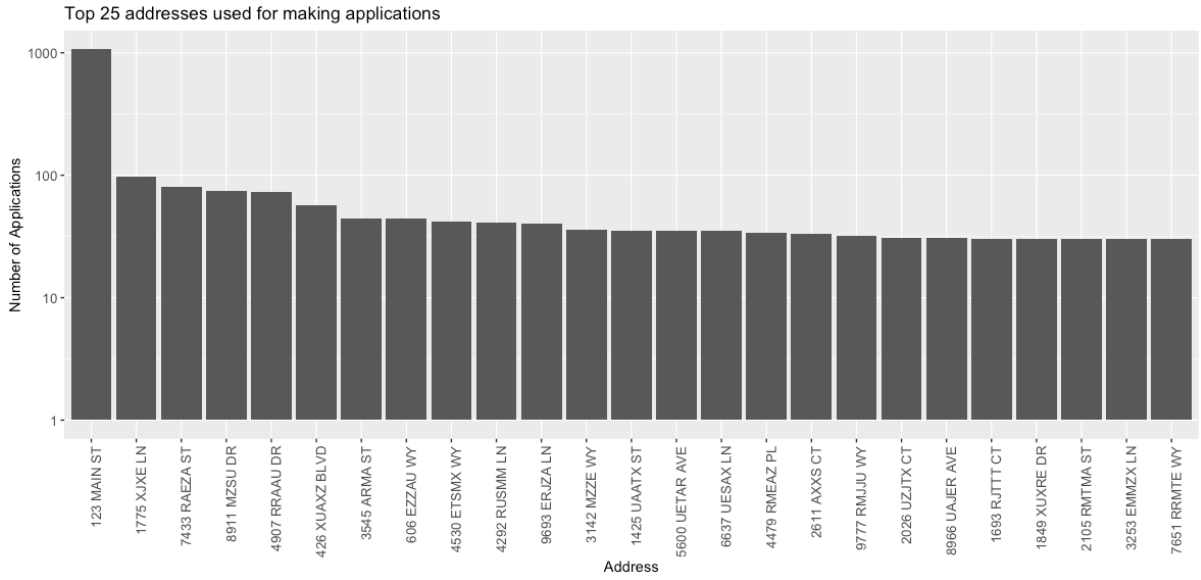


3.6 Address (address)

"address" is an indicator of the Address of the applicant filing the application.

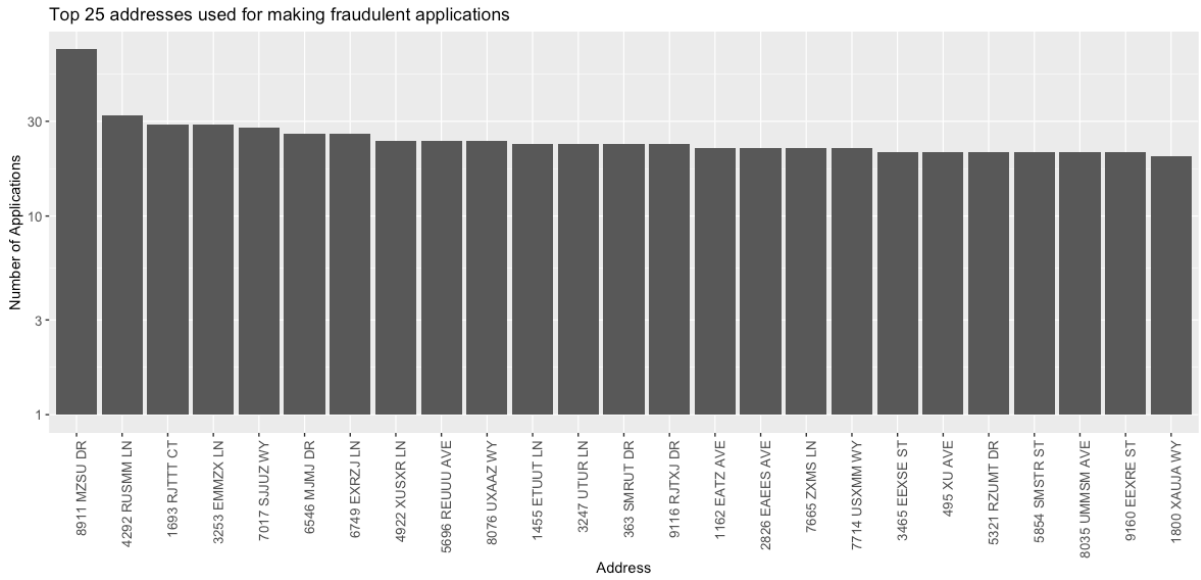
There are 828,774 unique addresses in the dataset. 100% of the fields are filled in this column.

A distribution of the top 25 addresses with the highest number of applications is as follows:



"123 MAIN ST" seems to be a proxy address being used while filing applications.

A distribution of the top 25 addresses with the highest number of fraudulent applications is as follows:



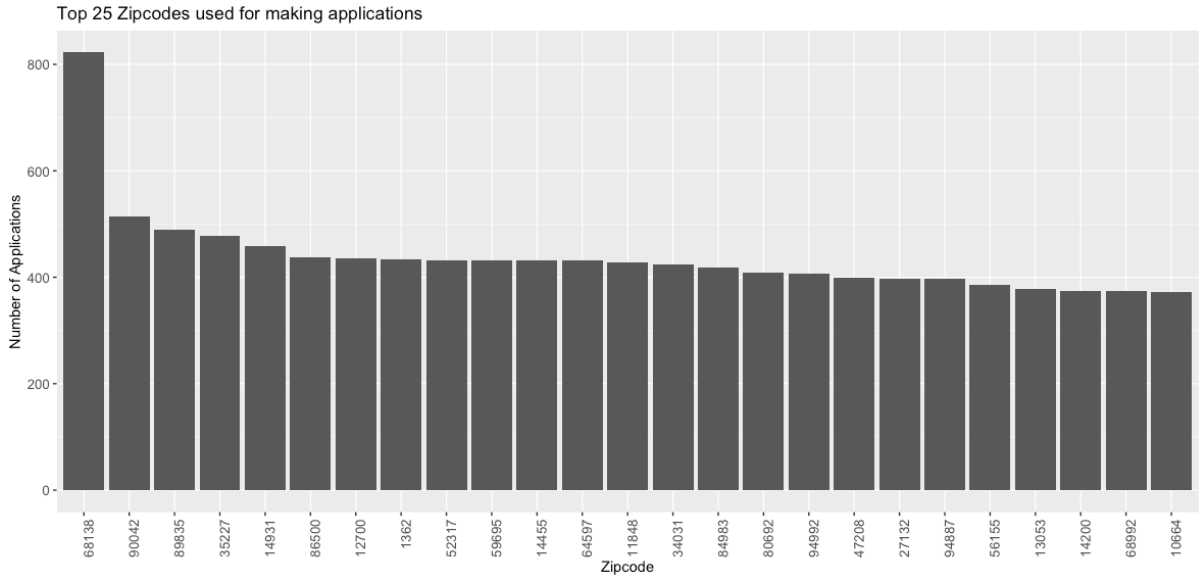
The proxy address surprisingly doesn't appear in the fraudulent applications.

3.7 Zipcode (zip5):

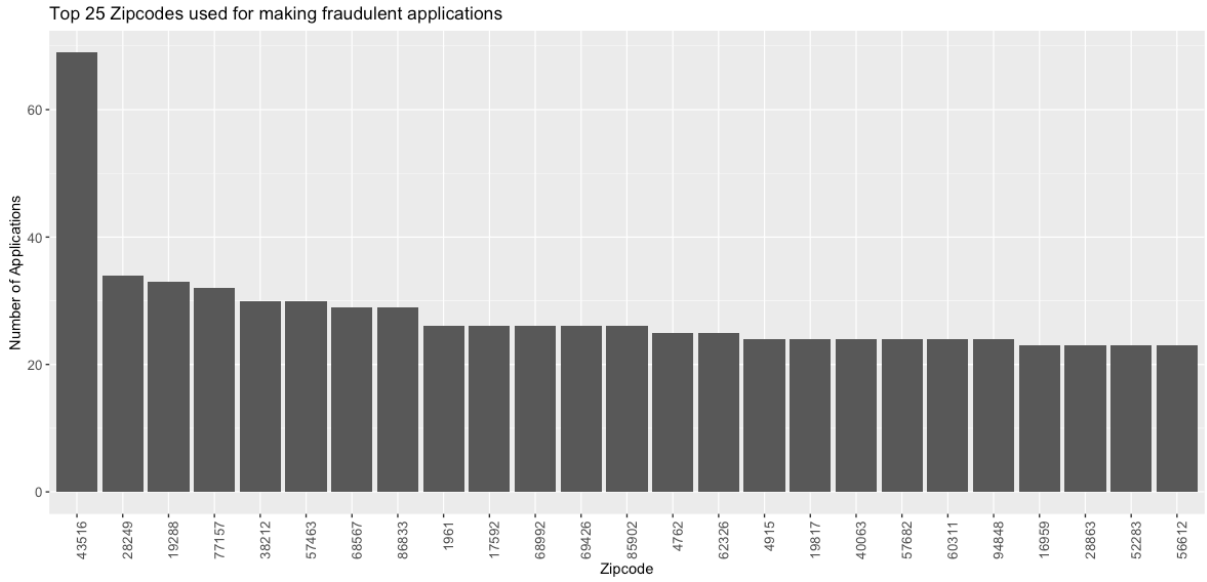
“zip5” is an indicator of the Zipcode of the applicant filing the application.

There are 26,370 unique zipcodes in the dataset. 100% of the fields are filled in this column.

A distribution of the top 25 zipcodes with the highest number of applications is as follows:



A distribution of the top 25 zipcodes with the highest number of fraudulent applications is as follows:

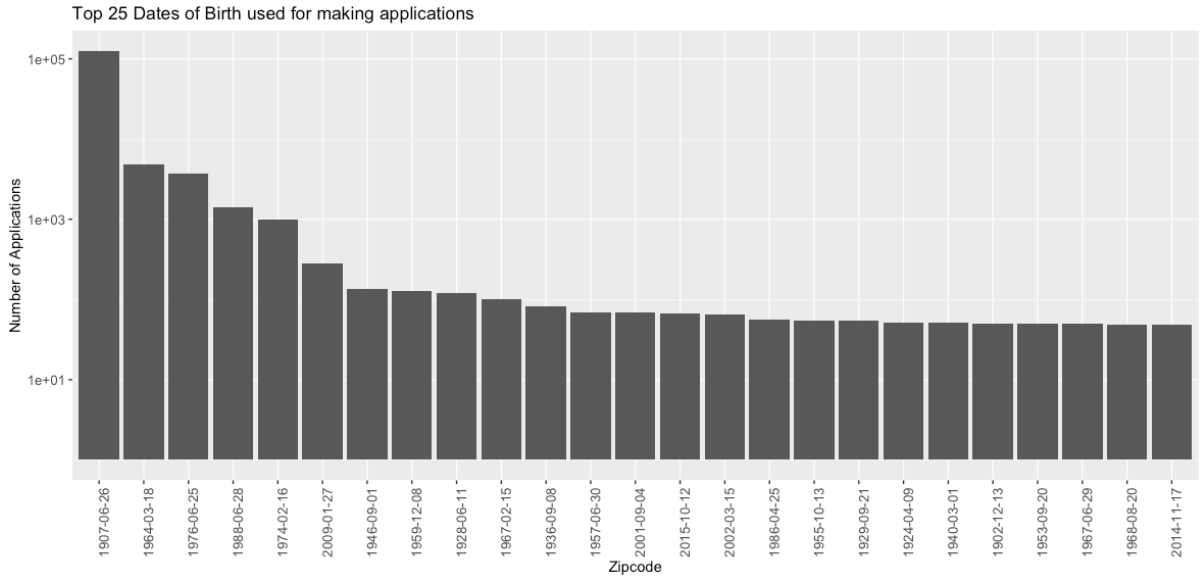


3.8 Date of Birth (dob):

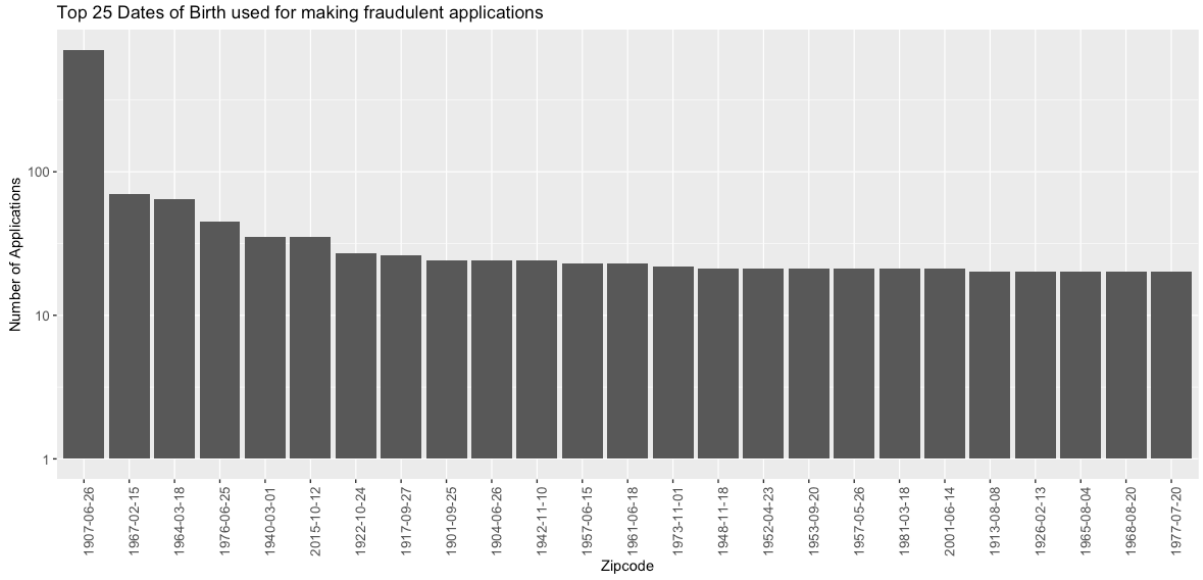
"dob" is an indicator of the Date of Birth of the applicant filing the application.

There are 42,673 unique dates of birth in the dataset. 100% of the dataset is filled.

A distribution of the top 25 birthdates used for filing applications is as follows:



A distribution of the top 25 birthdates used for filing fraudulent applications is as follows:



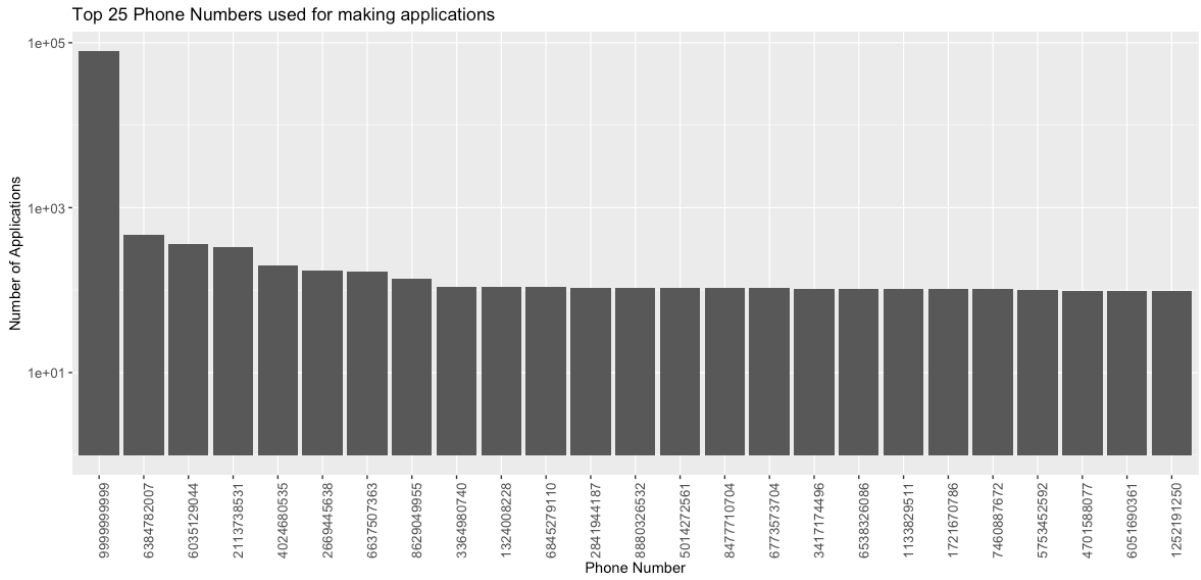
"1907-06-26" seems to be a proxy date used for birthdate. It is unlikely for people with a greater than 100 years of age to exist to make any applications.

3.9 Phone Number (homephone):

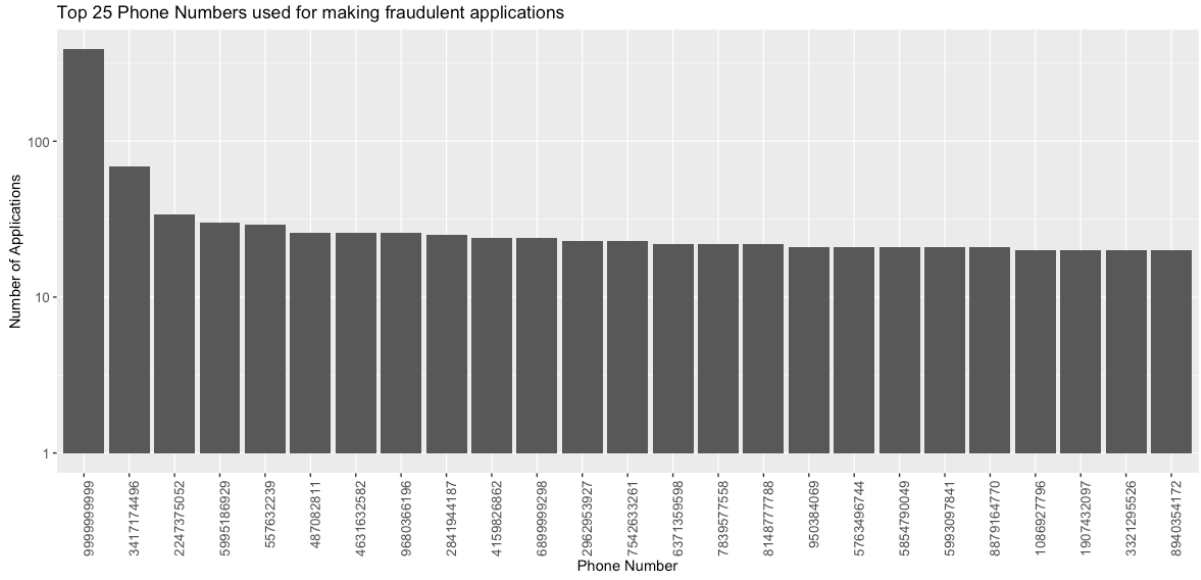
“homephone” is an indicator of the Phone Number of the applicant filing the application.

There are 28,244 unique phone numbers in the dataset. 100% of the dataset is filled.

A distribution of the top 25 phone numbers used for filing applications is as follows:



A distribution of the top 25 phone numbers used for filing fraudulent applications is as follows:



“999-999-9999” is clearly a proxy number being used for phone number while filing applications.

3.10 Fraud Indicator (fraud_label):

“fraud_label” is an indicator to indicate whether the application is fraudulent or not. It is a binary classifier with two categories.

100% of this column is filled.

A summary of this field is as below:

Fraud Category	Category Description	Count
0	Non-fraudulent application	985,607
1	Fraudulent application	14,393

4. Appendix

The above data quality report has been created with the intent of aligning all the relevant parties on the distribution and key statistical summaries of all the fields in the dataset.

The entire analysis is based on the dataset and dictionary shared by Dr. Stephen Coggeshall on 03/28/2019.

The approval of Dr. Stephen Coggeshall on the report implies Mr. Shreerang Javadekar can proceed with building a supervised model to identify trends in fraudulent transactions.

Signature: _____
Date: _____