**DATA QUALITY REPORT**

**PROJECT 2:**

**APPLICATION FRAUD ANALYTICS**

**Prepared for:**

Professor Coggeshall

USC Marshall DSO 562 Fraud Analytics

Homework 4

**Prepared by:**

Nanchun Shi

February 17, 2020

**TABLE OF CONTENTS**

[**PART I - INTRODUCTION**](#_jlgkpo7oov1a) **------------------------------------------------------------------------------ 2**

[**PART II - SUMMARY OF DATASET**](#_f5z4j27jbf5k) **------------------------------------------------------------------ 2**

[**PART III - FIELD DESCRIPTION**](#_8lvwnqf88br7) **--------------------------------------------------------------------- 3**

# 

# PART I - INTRODUCTION

The dataset of interest “applications data” contains information on 1,000,000 product application records received from January 1st, 2016 to December 31th, 2016. The dataset contains 10 fields including record number, application date, applicant information, and fraud label, which indicates if the application is considered to be an identity fraud.

## **PART II - SUMMARY OF DATASET**

**Below are the summary table for the 10 fields in the dataset:**

**Categorical fields:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Field Name** | **# of Records** | **% Populated** | **# Unique Values** | **Most Common Value** |
| **record** | 1000000 | 100% | 1000000 | - |
| **date** | 1000000 | 100% | 365 | 20160816 |
| **ssn** | 1000000 | 100% | 835819 | 999999999 |
| **firstname** | 1000000 | 100% | 78136 | EAMSTRMT |
| **lastname** | 1000000 | 100% | 177001 | ERJSAXA |
| **address** | 1000000 | 100% | 828774 | 123 MAIN ST |
| **zip5** | 1000000 | 100% | 26370 | 68138 |
| **dob** | 1000000 | 100% | 42673 | 19070626 |
| **homephone** | 1000000 | 100% | 28244 | 9999999999 |
| **fraud\_label** | 1000000 | 100% | 2 | 0 |

## 

## 

## 

## **PART III - FIELD DESCRIPTION**

**1.**

**Field name:** record

**Field type:** int64

**Description:** unique index number for each record, ascending from 1 to 1000000.

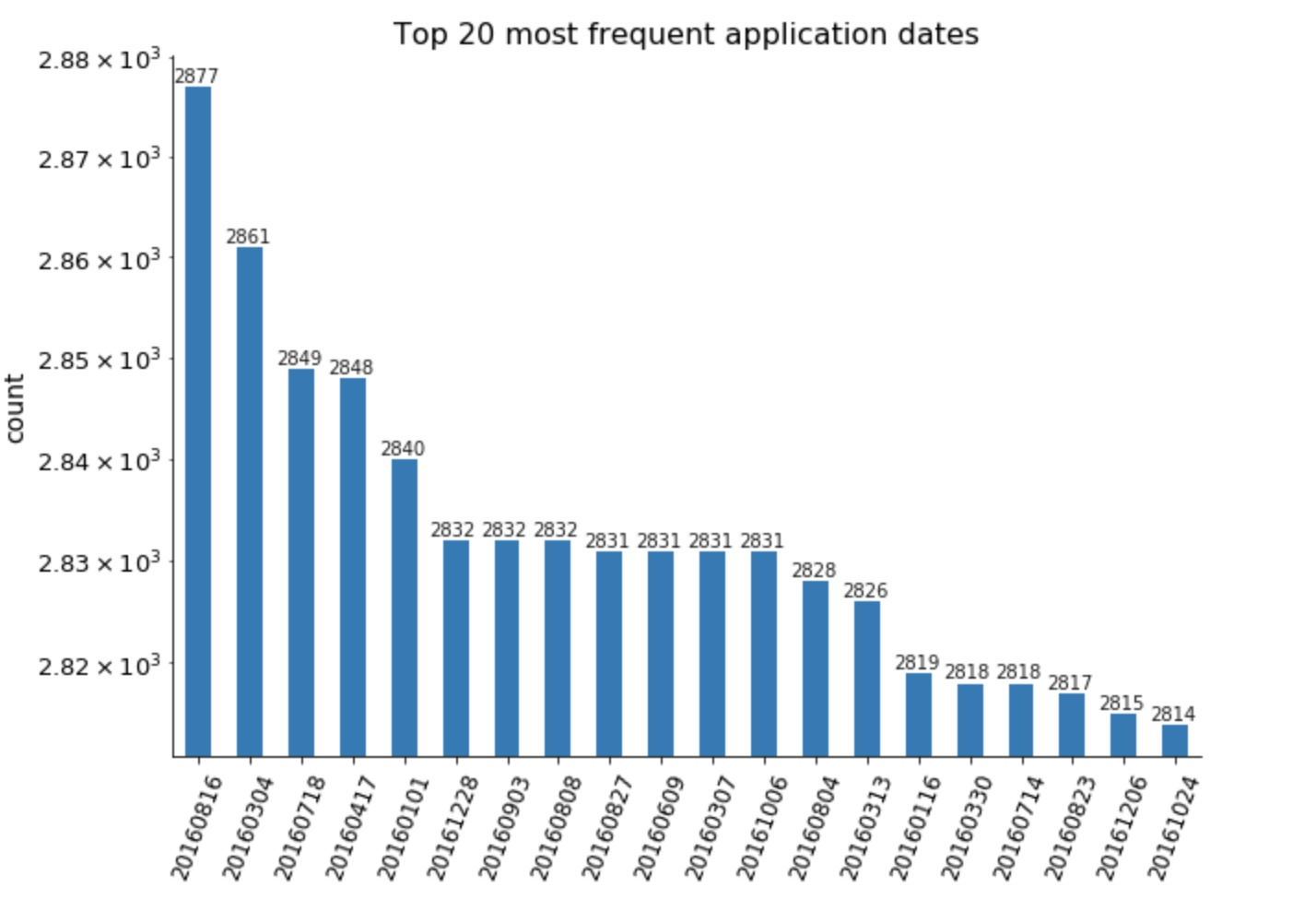
**2.**

**Field name:** date

**Field type:** int64

**Description:** application date of each record.

Top 20 most frequent dates:



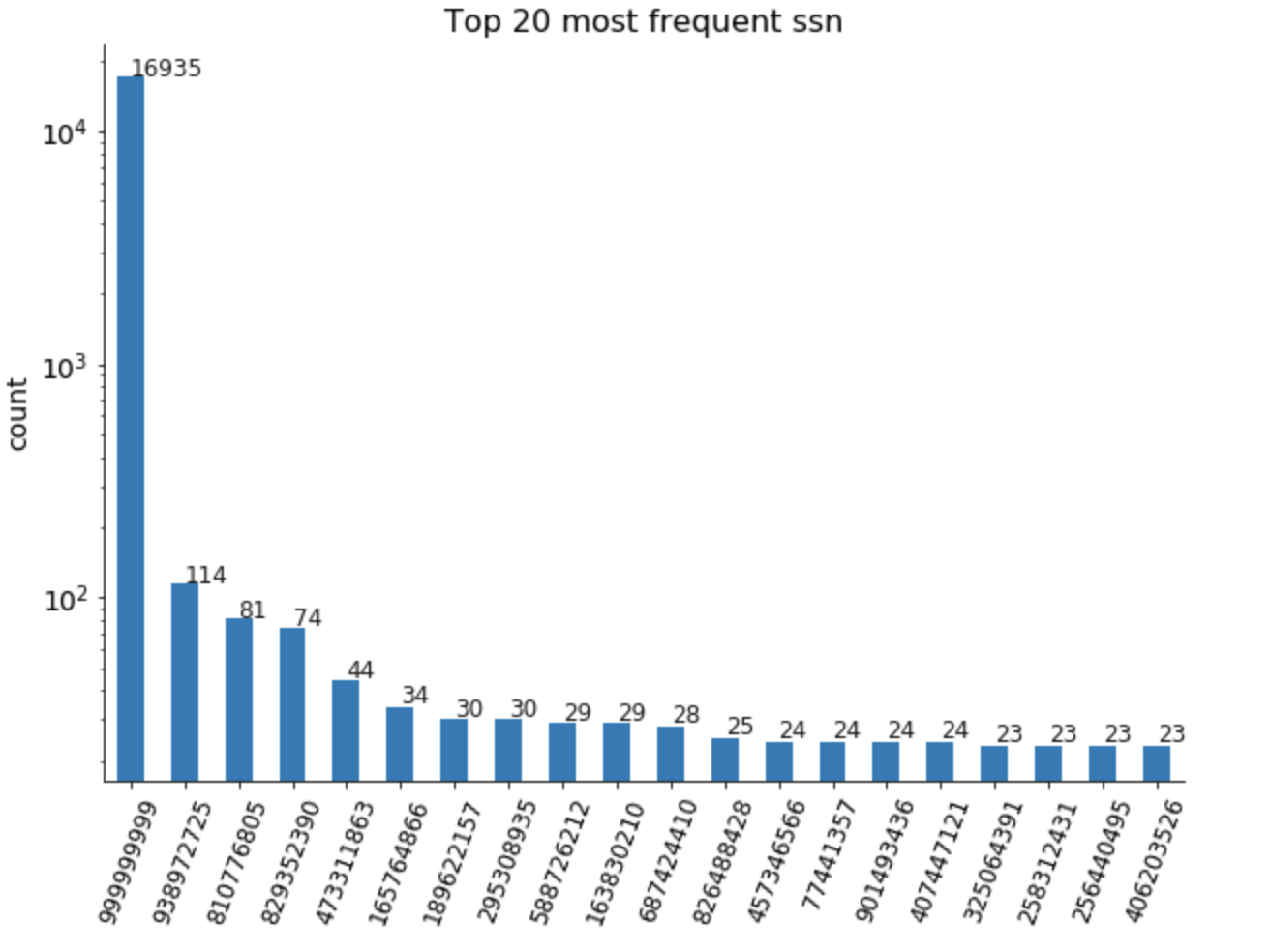
**3.**

**Field name:** ssn

**Field type:** int64

**Description:** social security number of the applicant.

Top 20 most frequent ssn:



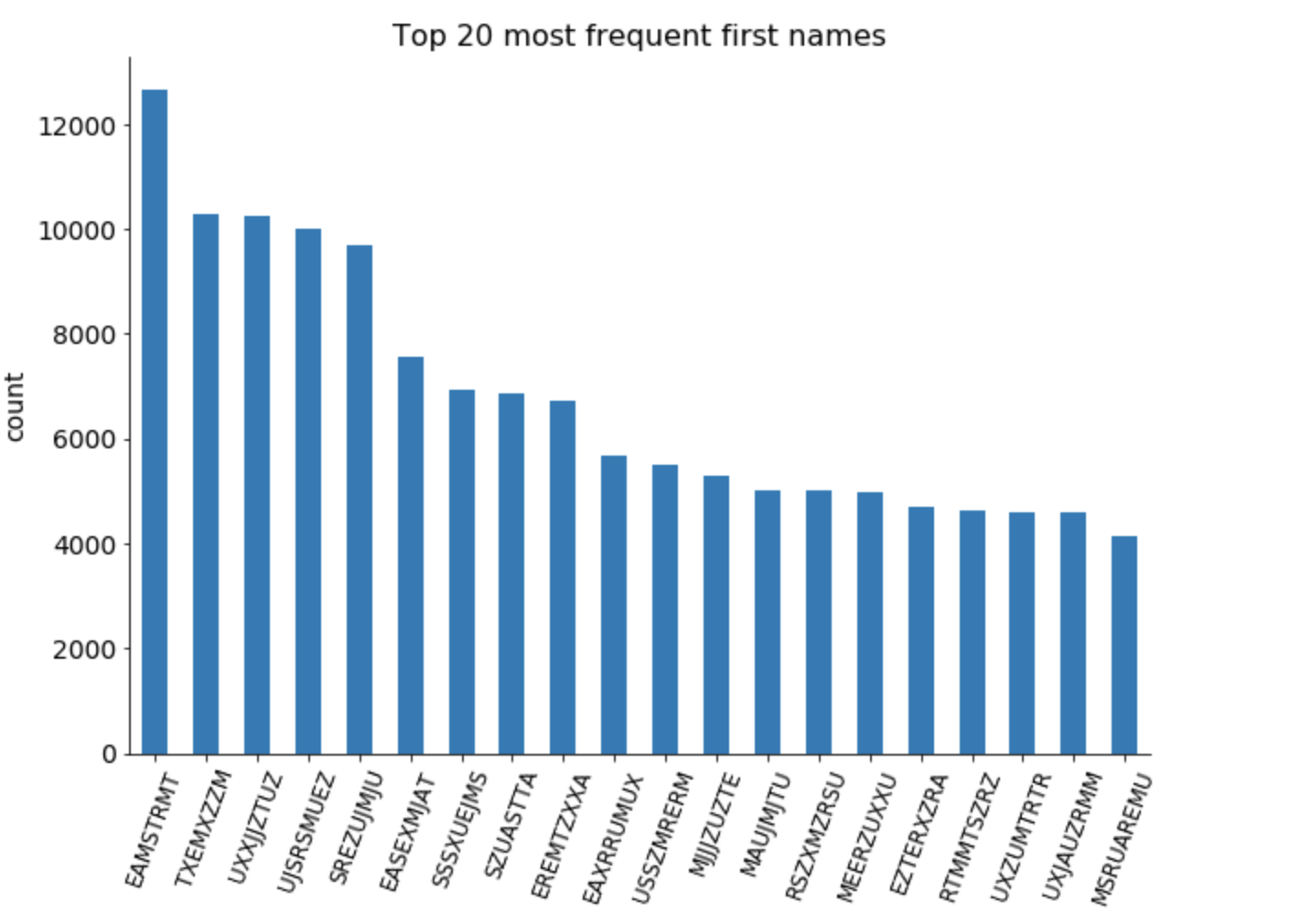
**4.**

**Field name:** firstname

**Field type:** object

**Description:** first name of the applicant.

Top 20 most frequent first names:

****

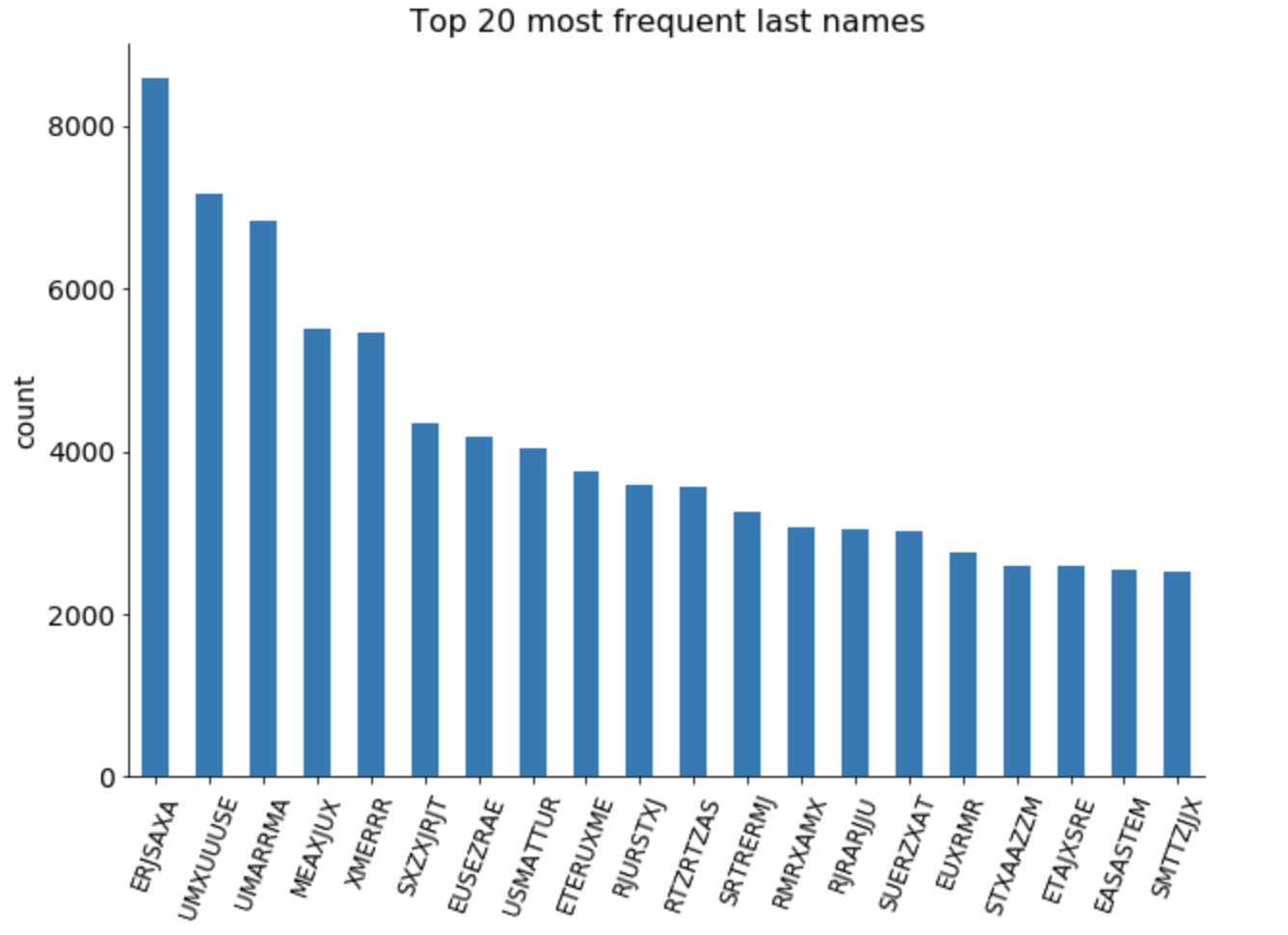
**5.**

**Field name:** lastname

**Field type:** object

**Description:** last name of the applicant.

Top 20 most frequent last names:



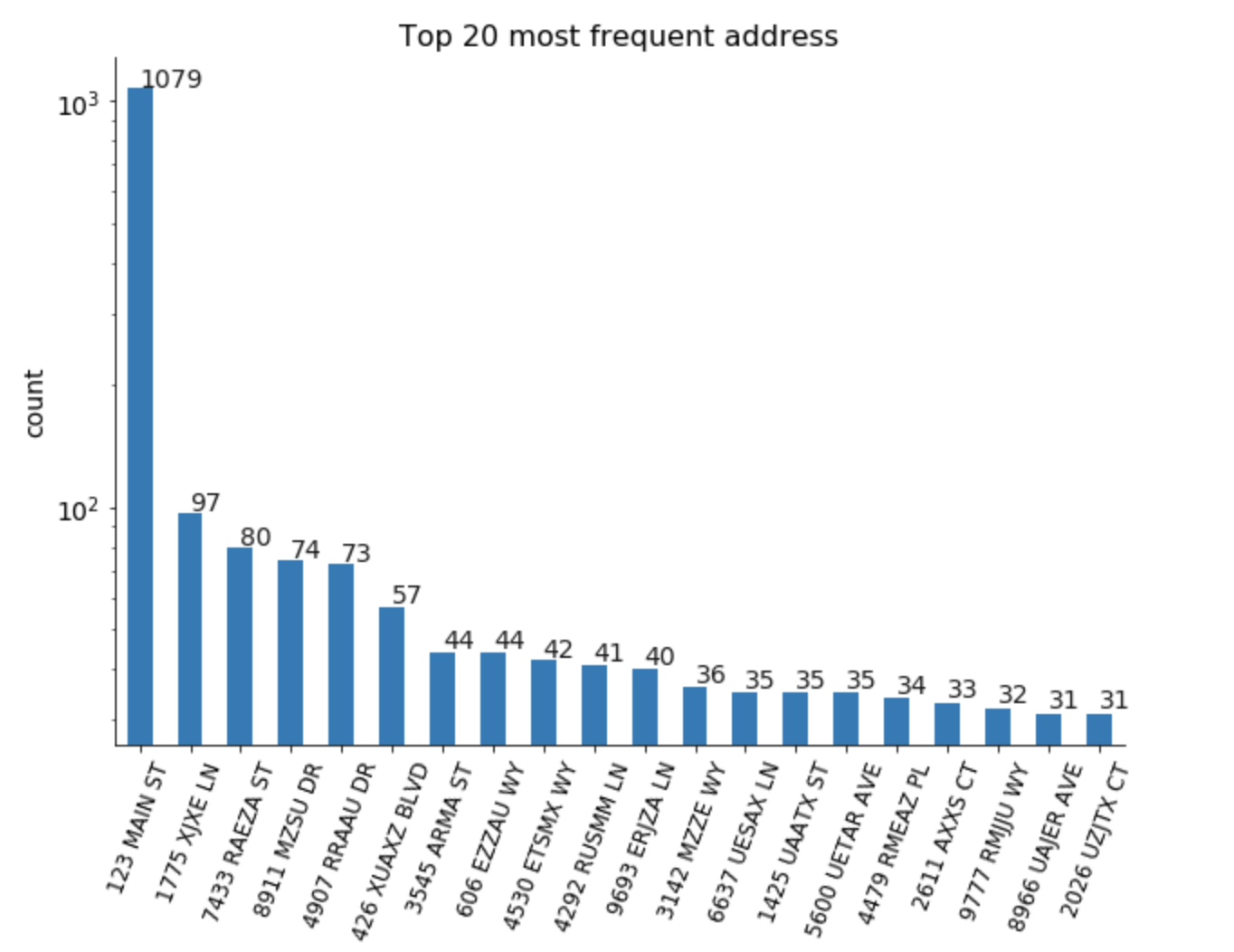
**6.**

**Field name:** address

**Field type:** object

**Description:** address of the applicant.

Top 20 most frequent address:



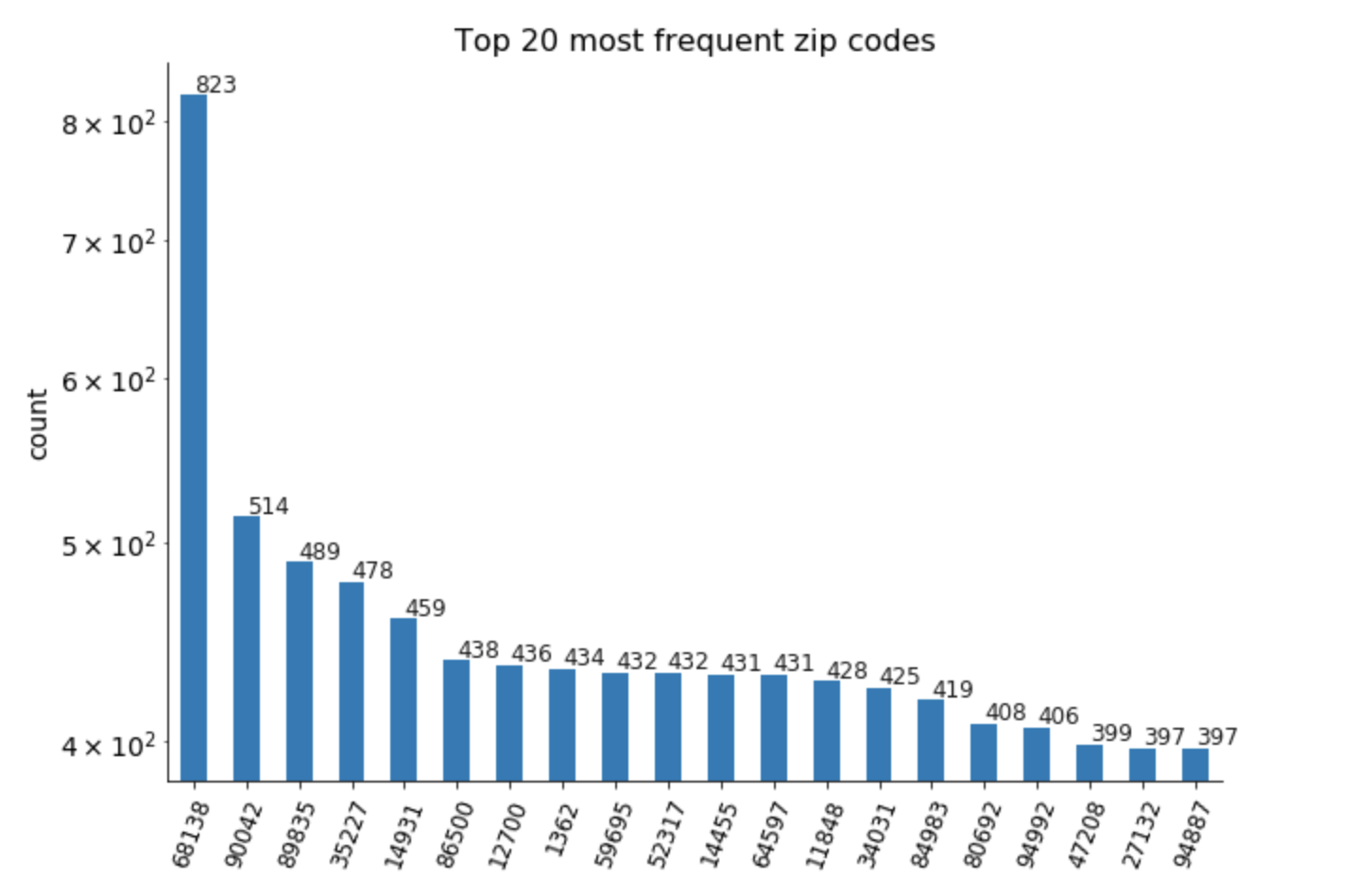
**7.**

**Field name:** zip5

**Field type:** int64

**Description:** 5-digit zip code of the applicant.

Top 20 most frequent zip codes:

****

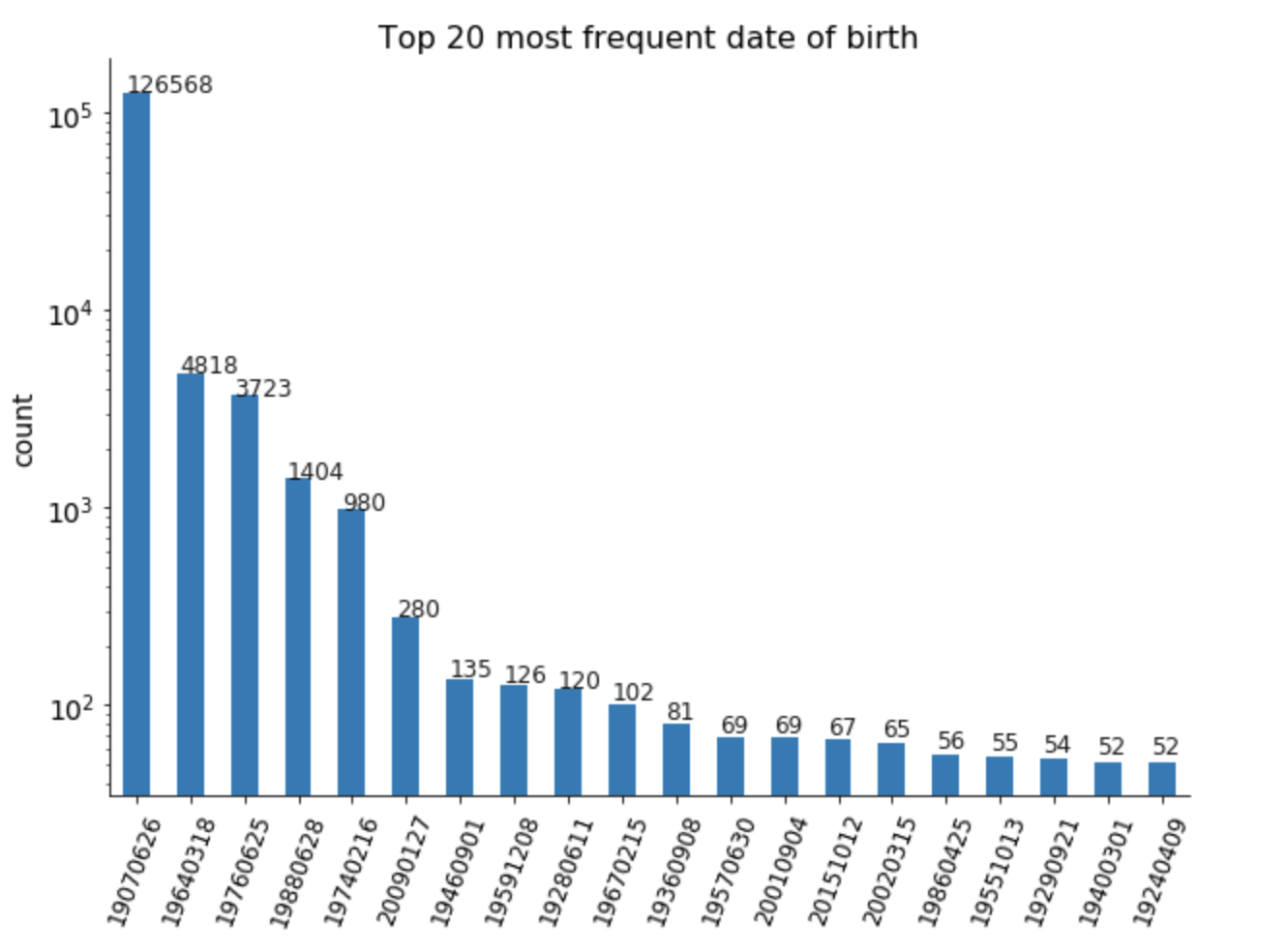
**8.**

**Field name:** dob

**Field type:** int64

**Description:** date of birth of the applicant.

Top 20 most frequent date of birth:



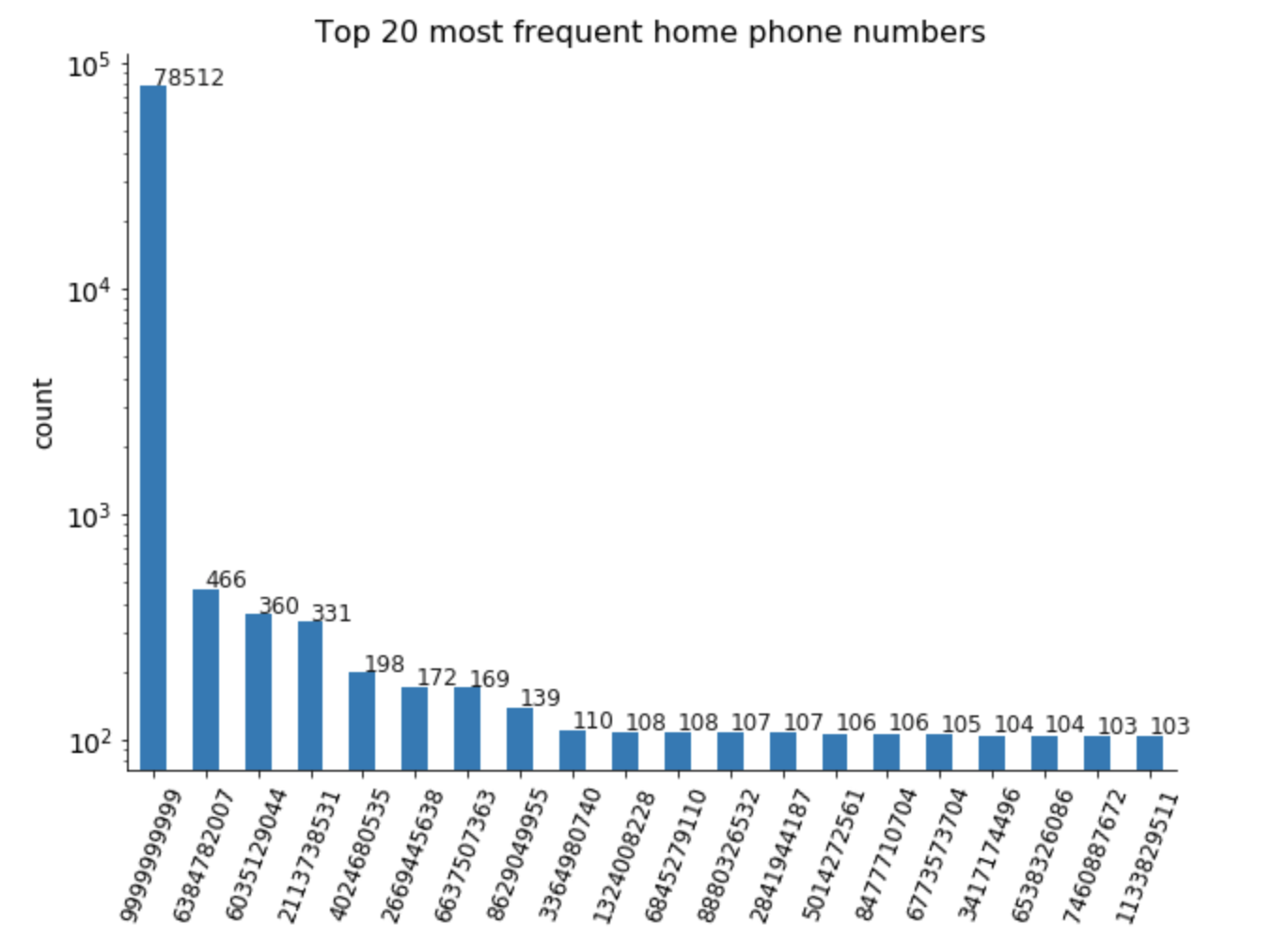
**9.**

**Field name:** homephone

**Field type:** int64

**Description:** home phone number of the applicant.

Top 20 most frequent home phone number:

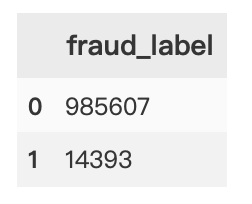


**10.**

**Field name:** fraud\_label

**Field type:** int64

**Description:** binary indicator; 1 means the record is considered as a fraud, 0 otherwise.

Counts for two labels: