

# Data Quality Report

## 1. Data Description

The data is about **Application/ Identity Fraud**. The data is synthetic application data indicating PII of individual from US applications such as applications for credit card from January 1st, 2017, to December 31st, 2017. It has 1,000,000 rows and 10 fields.

## 2. Summary Table

### I. Numerical Table

Field Name	% Populated	Min	Max	Mean	Stdev	%Zero
Date	100.00	2017-01-01	2017-12-31	NA	NA	0.00
Dob	100.00	1900-01-01	2016-10-31	NA	NA	0.00

### II. Categorical Table

Field Name	% Populated	#Unique Value	Most Common Value
Record	100.00	1,000,000	Nan
SSN	100.00	835,819	999999999
First Name	100.00	78,136	EAMSTRMT
Last Name	100.00	177,001	ERJSAXA
Address	100.00	828,774	123 MAIN ST
Zip	100.00	26,370	68138
Home Phone	100.00	28,244	9999999999
Fraud Label	100.00	2	0

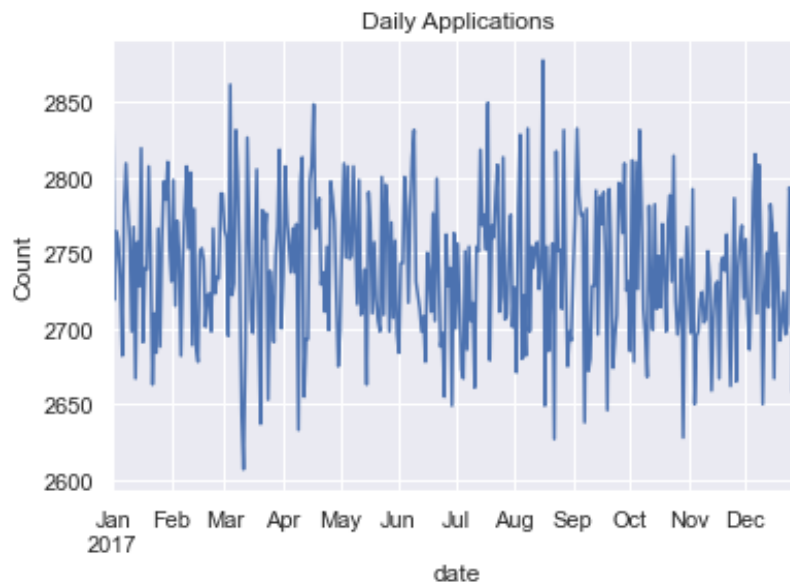
## 3. Visualization of Each Field:

### (1) Field Name: Record

Description: Record Number: Ordinal Unique positive integer for each record from 1 to 1,000,000.

## (2) Field Name: Date

Description: Date: The “date” field indicates the date on which the individual has raised a request for an application. The daily application distribution across time.



## (3) Field Name: Ssn

Description: SSN: SSN refers to the social security number of the applicant. The distribution indicates the top 15 most common SSNs used by applicants.

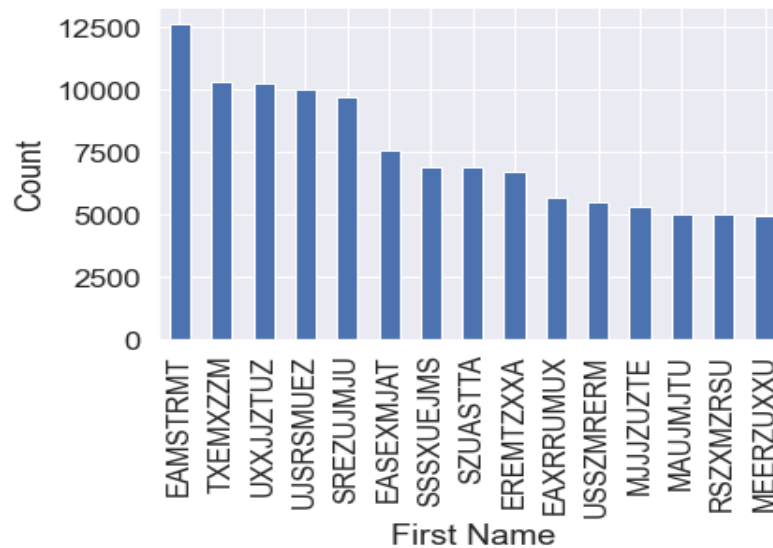
The most common value of the SSN is 999999999, the count is 16,935.



#### (4) Field Name: Firstname

Description: First Name: The field “firstname” refers to the First name of the applicant. The distribution indicates the top 15 most common First name used by applicants.

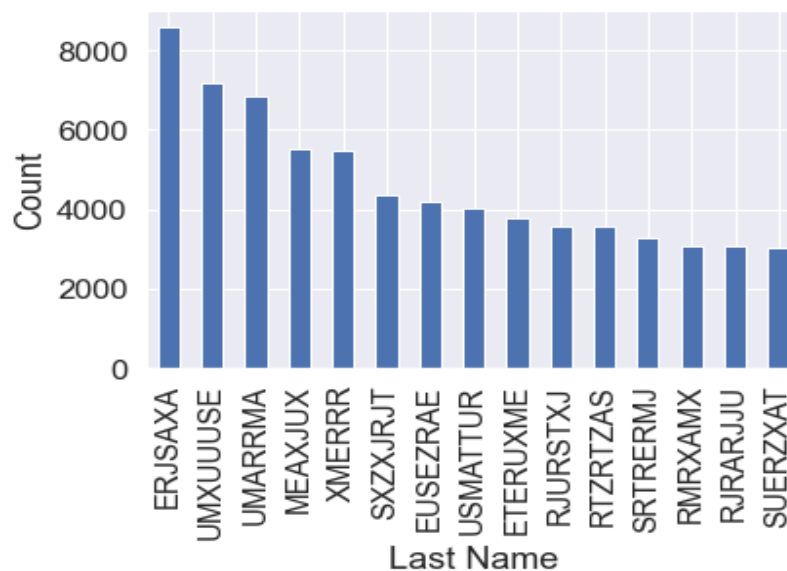
The most common value of the first name is EAMSTRMT, the count is 12,658.



#### (5) Field Name: Lastname

Description: Last Name: The field “lastname” refers to the Last name of the applicant. The distribution indicates the top 15 most common Last name used by applicants.

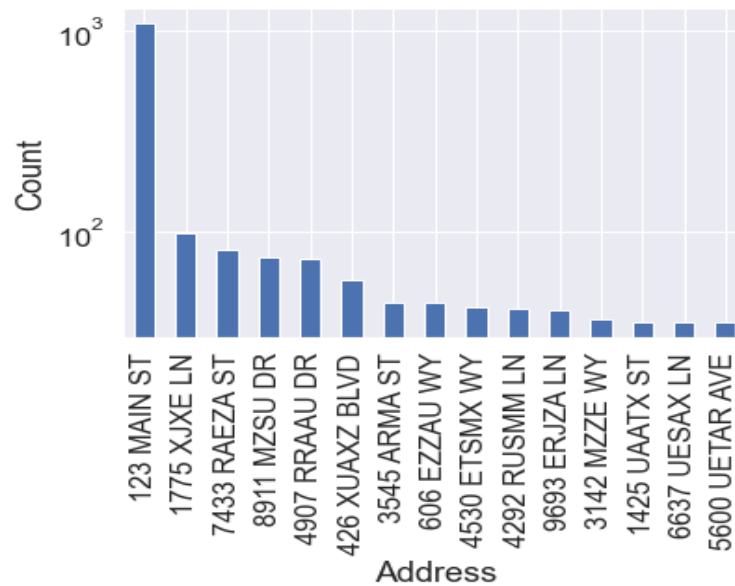
The most common value of the first name is ERJSAXA, the count is 8,580.



## (6) Field name: address

Description: Address: The field “address” refers to the Address of the applicant. The distribution indicates the top 15 most common Last name used by applicants.

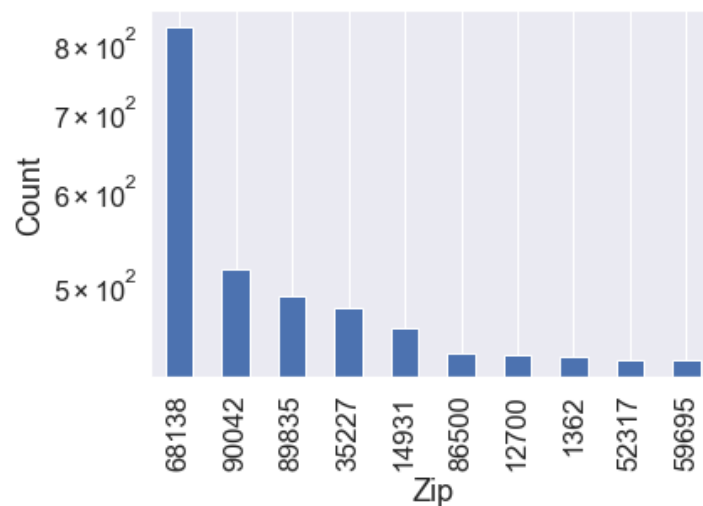
The most common value of the first name is 123 MAIN ST, the count is 1,079.



## (7) Field Name: Zip5

Description: Zip: The field “zip5” refers to the Zip code of the applicant. The distribution indicates the top 10 most common Zip code used by applicants.

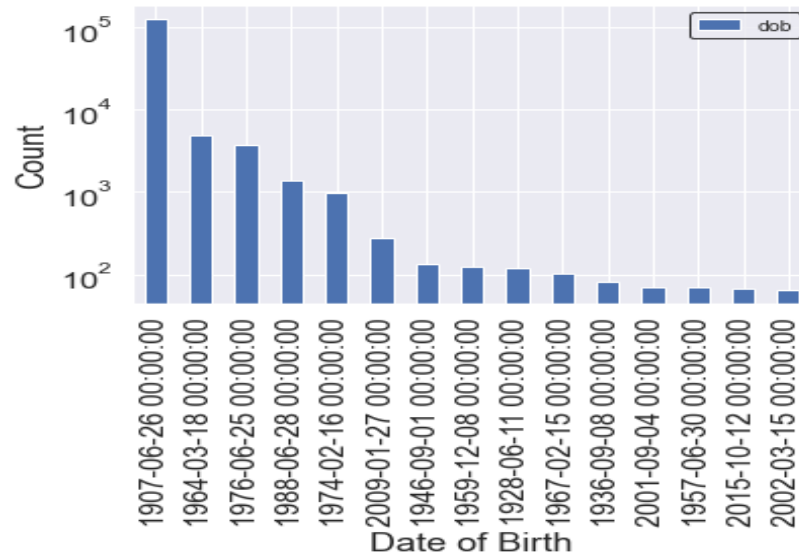
The most common value of the Zip code is 68138, the count is 823.



## (8) Field Name: Dob

Description: Date of Birth: The field “dob” refers to the Date of Birth of the applicant. The distribution indicates the top 15 most common Zip code used by applicants.

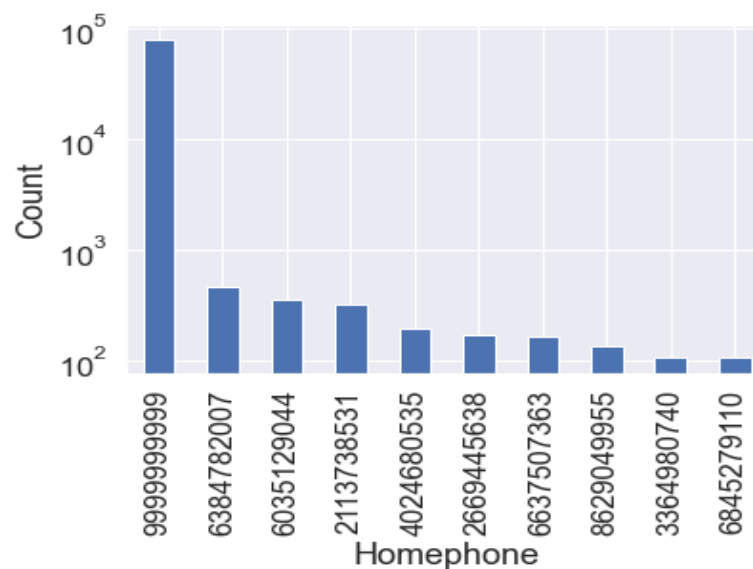
The most common value of the dob code is 1907-06-26, the count is 26,568.



## (9) Field Name: Homephone

Description: Home Phone: The field “homephone” refers to the phone number of the applicant. The distribution indicates the top 10 most common home phone used by applicants.

The most common value of the dob code is ‘999999999’, the count is 78,512.



(10) Field Name: Fraud label

Description: Fraud = 0 (no fraud label), Fraud=1 (fraud label)

The count of fraud\_label=0 (985,607) and fraud\_label =1 (14,393)

