Data Quality Report

1. Data Description

This is the synthetic applications data showing identity information for each applicant during the Year 2017. It has 10 fields and 1000000 records.

2. Summary Tables

(1) Numerical Table

Field Name	% Populated	Min	Max	Mean	Stdev	% Zero
date	100.00	2017-01-01	2017-12-31	N/A	N/A	0.00
dob	100.00	1900-01-01	2016-10-31	N/A	N/A	0.00

(2) Categorical Table

Field Name	% Populated	# Unique Values	Most Common Value
record	100.00	1000000	N/A
ssn	100.00	835819	999999999
firstname	100.00	78136	EAMSTRMT
lastname	100.00	177001	ERJSAXA
address	100.00	828774	123 MAIN ST
zip5	100.00	26370	68138
homephone	100.00	28244	9999999999
fraud_label	100.00	2	0

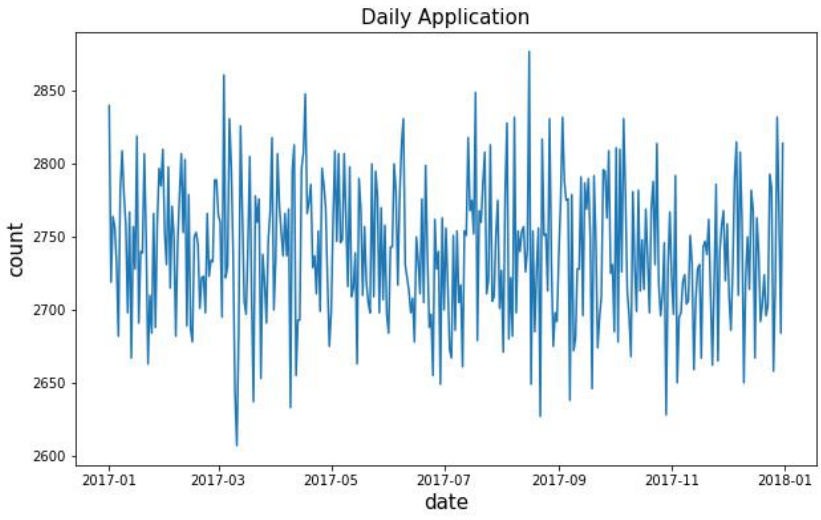
3. Visualization of Each Field

(1) Field Name: record

Description: Ordinal unique positive integer for each record ranging from 1 to 1000000.

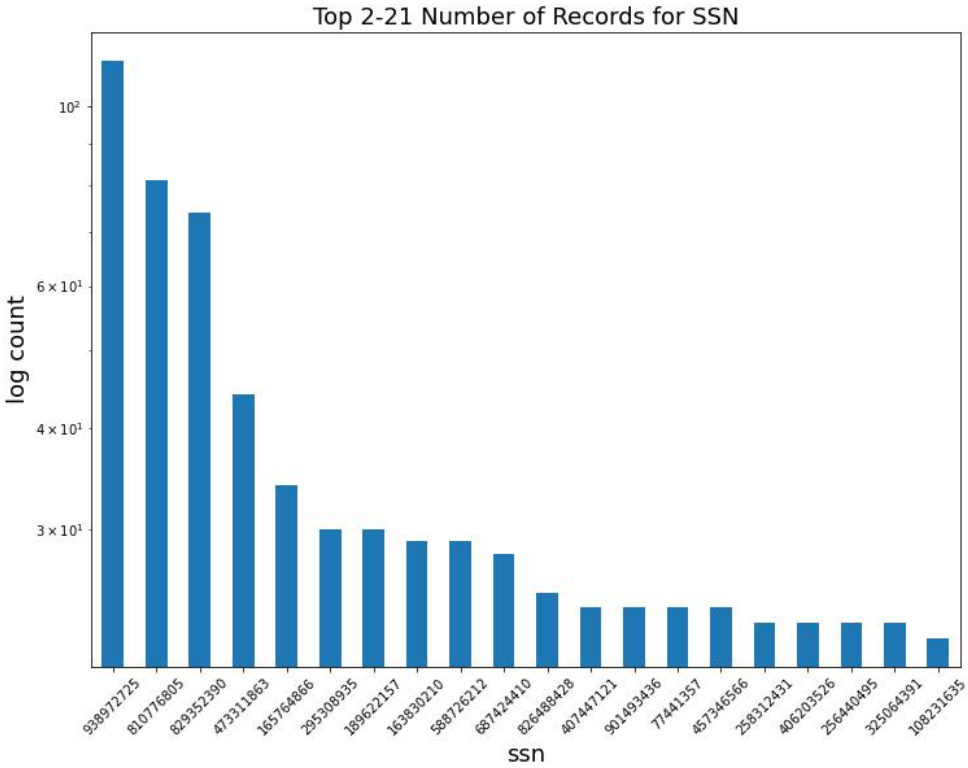
(2) Field Name: date

Description: The date for each application during Year 2017 with distribution shown below.



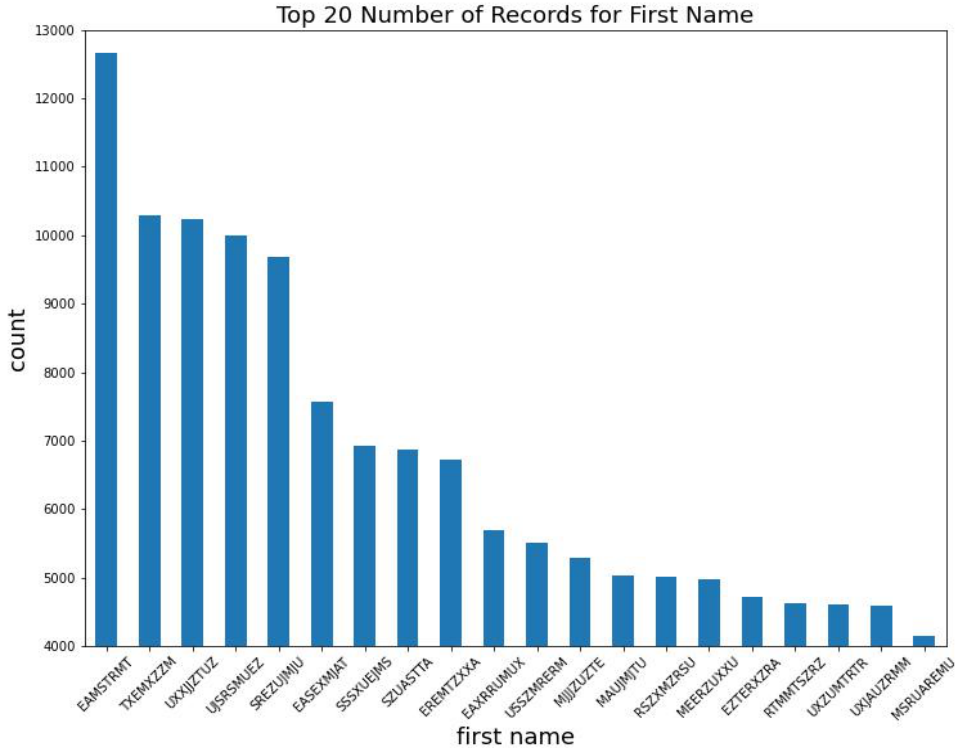
(3) Field Name: ssn

Description: SSN for each application. The most common value is 999999999 with 16935 counts, extremely higher than others. The subsequent top 20 frequent SSNs are presented below.



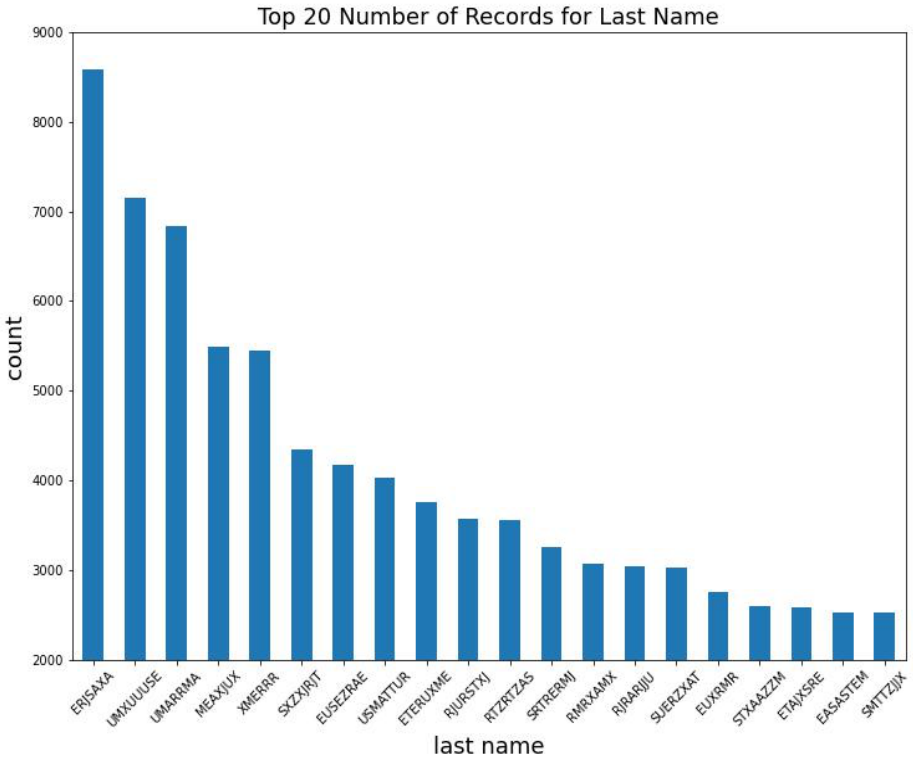
(4) Field Name: firstname

Description: The first name for each application. The top 20 are listed below.



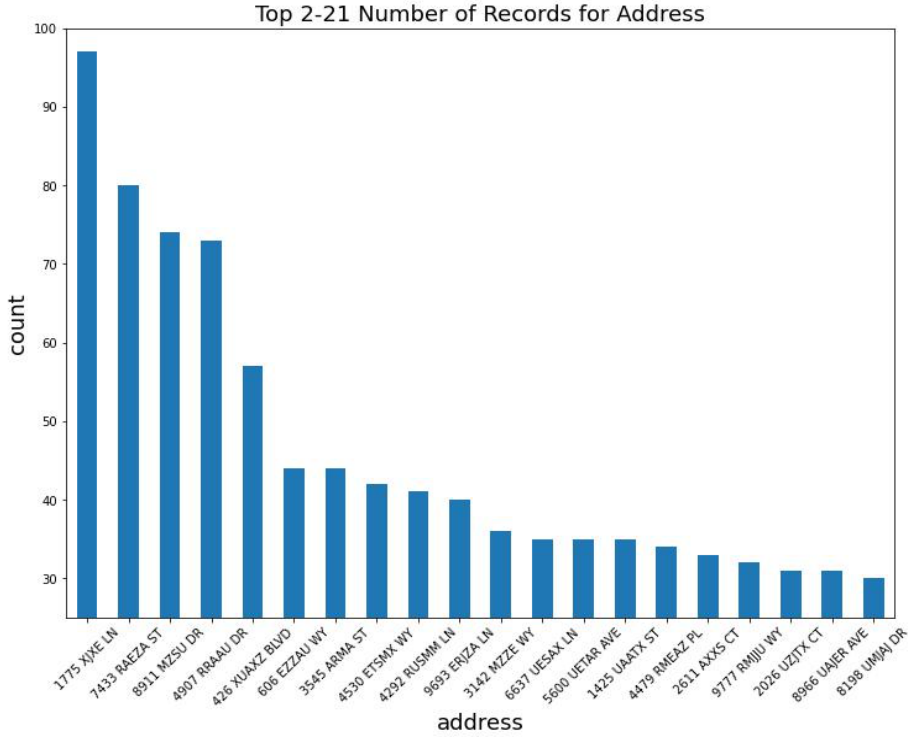
(5) Field Name: lastname

Description: The last name for each application with top 20 presented below.



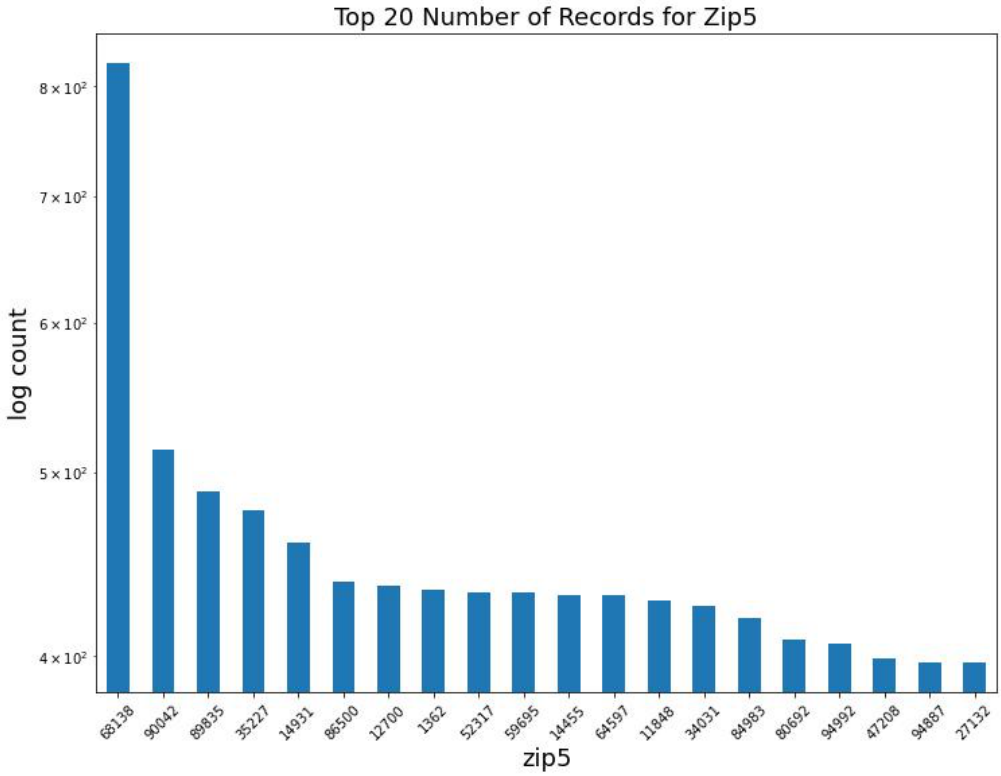
(6) Field Name: address

Description: Street address. The most common value is 123 MAIN ST with 1079 counts, significantly higher than others. The next 20 most common values are listed below.



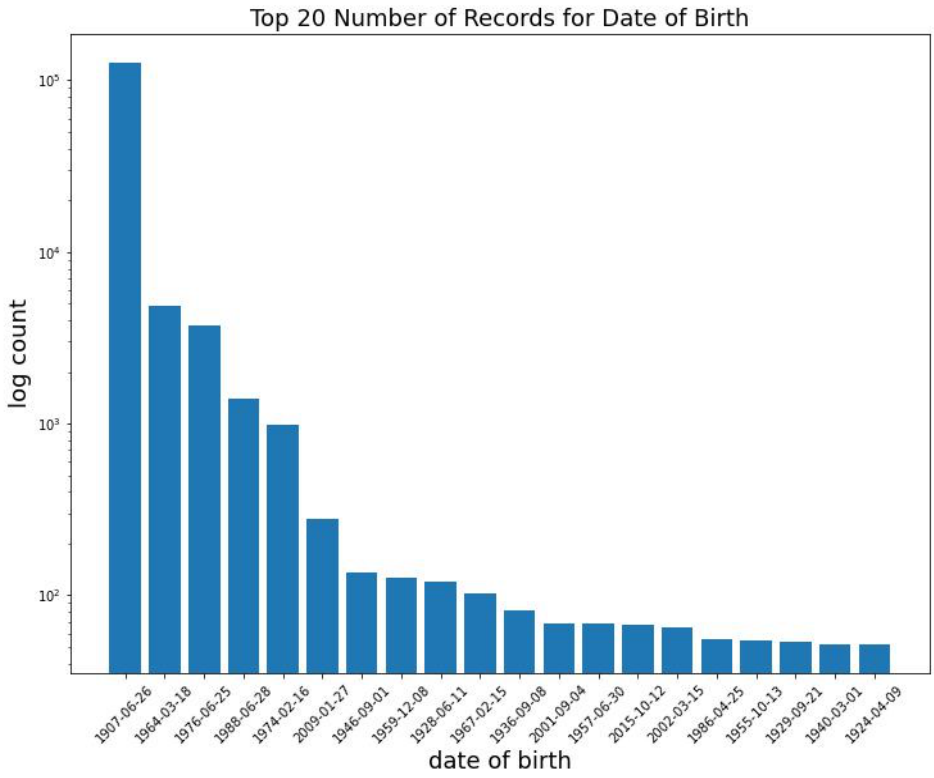
(7) Field Name: zip5

Description: 5-digit zip code. The most common value is 68138 with 823 counts. The top 20 frequent zip codes are visualized below.



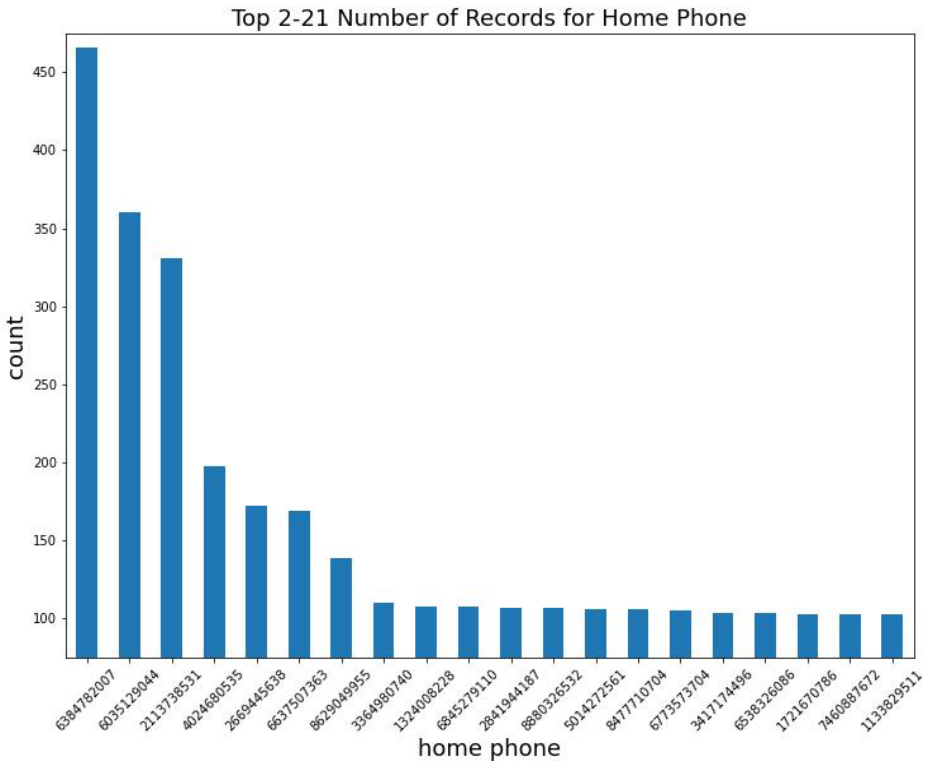
(8) Field Name: dob

Description: The field data of birth covers from 1900-01-01 to 2016-10-31. The most common value is 1907-06-26 with 126568 records. Top 20 log counts for dob are plotted.



(9) Field Name: homephone

Description: Home phone number. The most common value is 9999999999 with 78512 records, much higher than others. The top 2-21 frequent phone numbers are plotted.



(10) Field Name: fraud_label

Description: The field is binary with 1 indicating fraud, 0 otherwise. It has 985607 records of 0 and 14393 records of 1. The log distribution is shown in the figure below.