Data Quality Report (DQR)

Application Data

The Application Data is a dataset of product application with PII (Personal Identifying Information – the fields that identify the person). It covers the whole year of 2016 (01/01/2016 – 12/31/2016) and consists of 10 fields with 1,000,000 records.

SUMMARY TABLES

TABLE FOR NUMERIC FIELDS

Field Name	% Populated	Min	Max	Mean	Stdev	% Zero
date	100	2016-01-01	2016-12-31	N/A	N/A	0
dob	100	1900-01-01	2016-10-31	N/A	N/A	0

TABLE FOR CATEGORICAL FIELDS

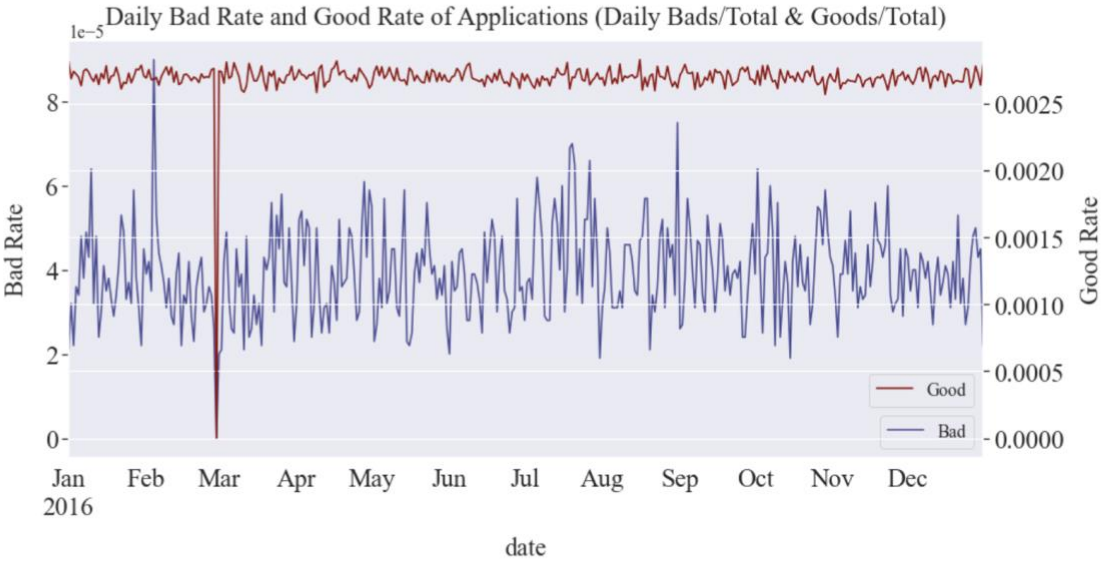
Field Name	% Populated	# Unique Values	Most Common Value
record	100	1,000,000	N/A
ssn	100	835,819	999999999
firstname	100	78,136	EAMSTRMT
lastname	100	177,001	ERJSAXA
address	100	828,774	123 MAIN ST
zip5	100	26,370	68138
homephone	100	28,244	9999999999
fraud_label	100	2	0

DESCRIPTION OF EACH FIELD

record (Categorical) -- Each record has a unique number.

date (Numeric - Date)

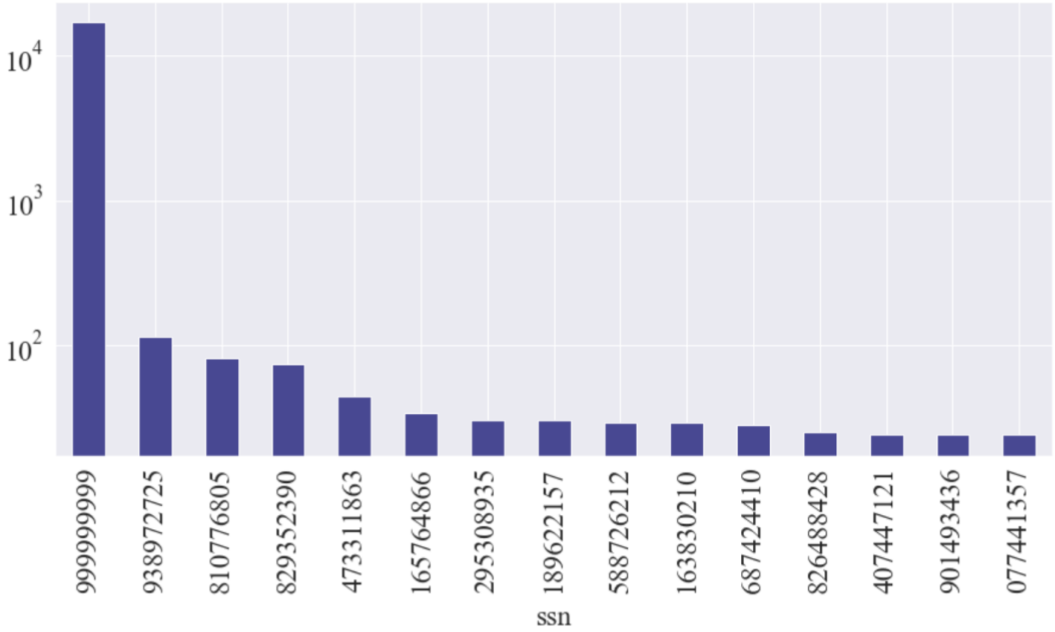
Date. The minimum date is 2016-01-01 and the maximum date is 2016-12-31.



ssn (Categorical)

SSN (Social Security number). There are 835,819 unique SSNs and the most common value is 999999999.

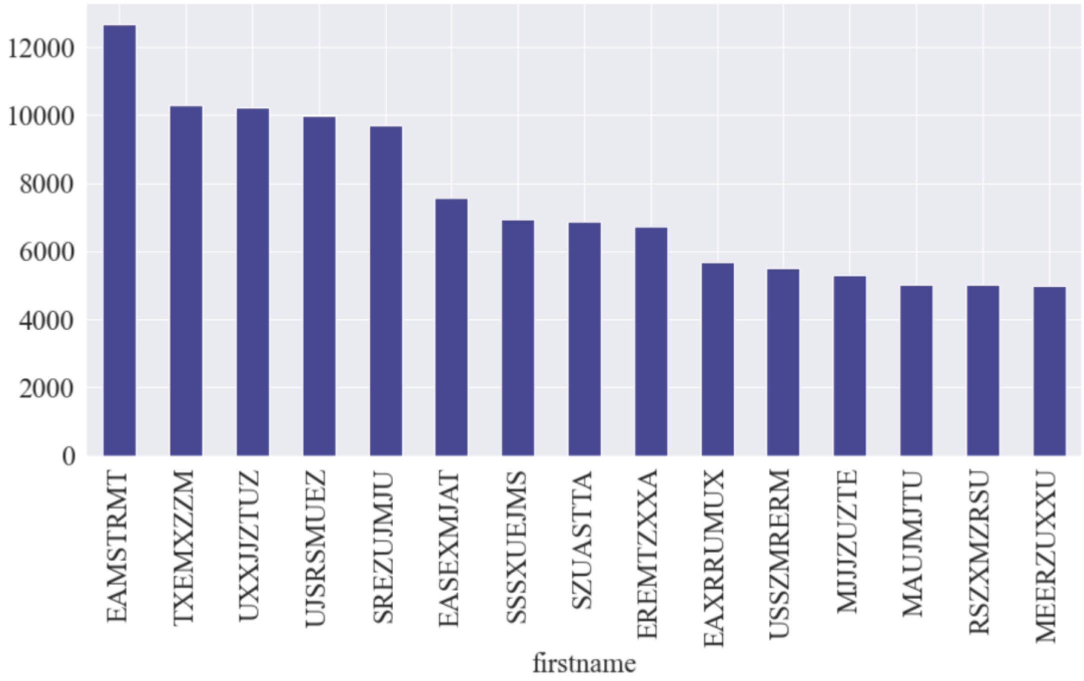
Bar Plot of ssn (Top 15)



firstname (Categorical)

First Name. There are 78,136 unique first names and the most common value is EAMSTRMT.

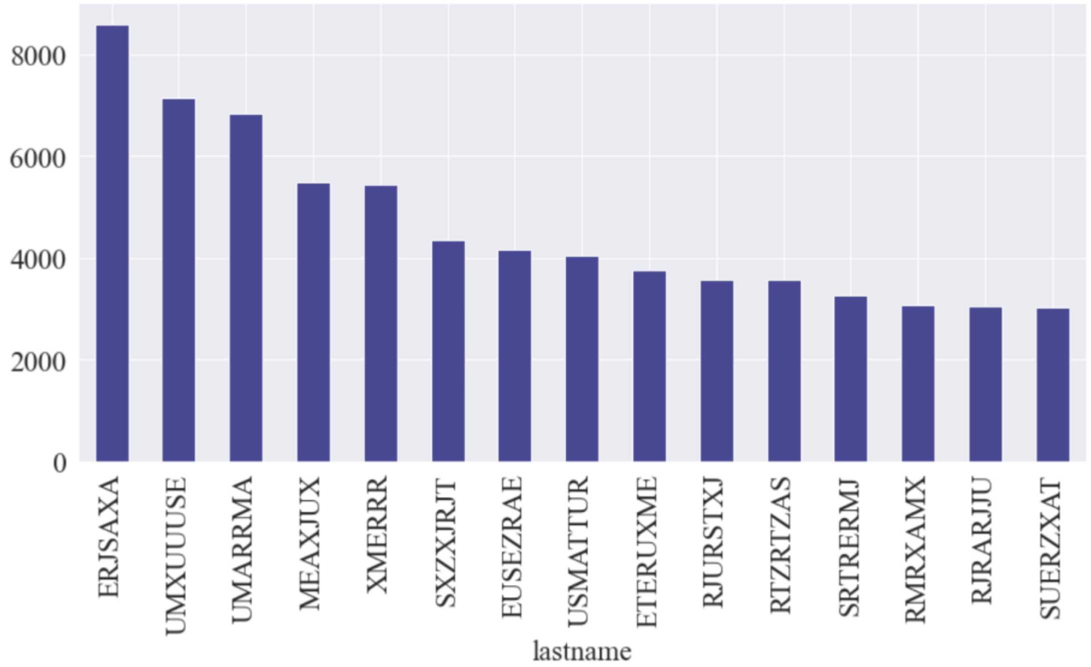
Bar Plot of firstname (Top 15)



lastname (Categorical)

Last Name. There are 177,001 last names and the most common value is ERJSAXA.

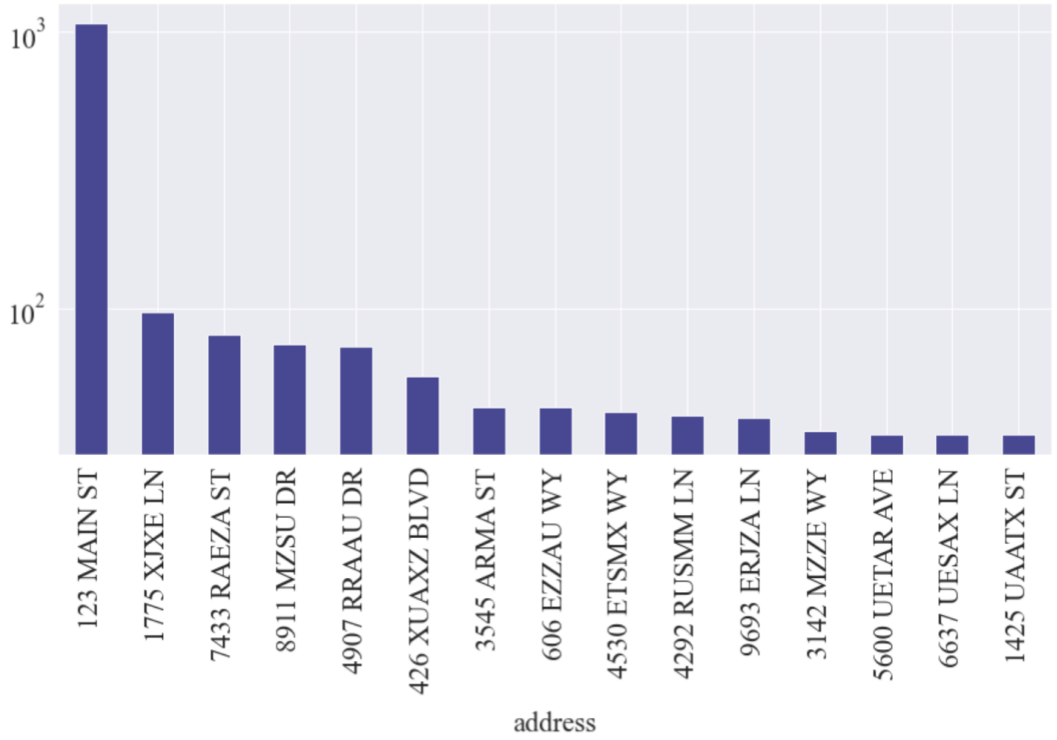
Bar Plot of lastname (Top 15)



address (Categorical)

Address. There are 828,774 unique addresses and the most common value is 123 MAIN ST.

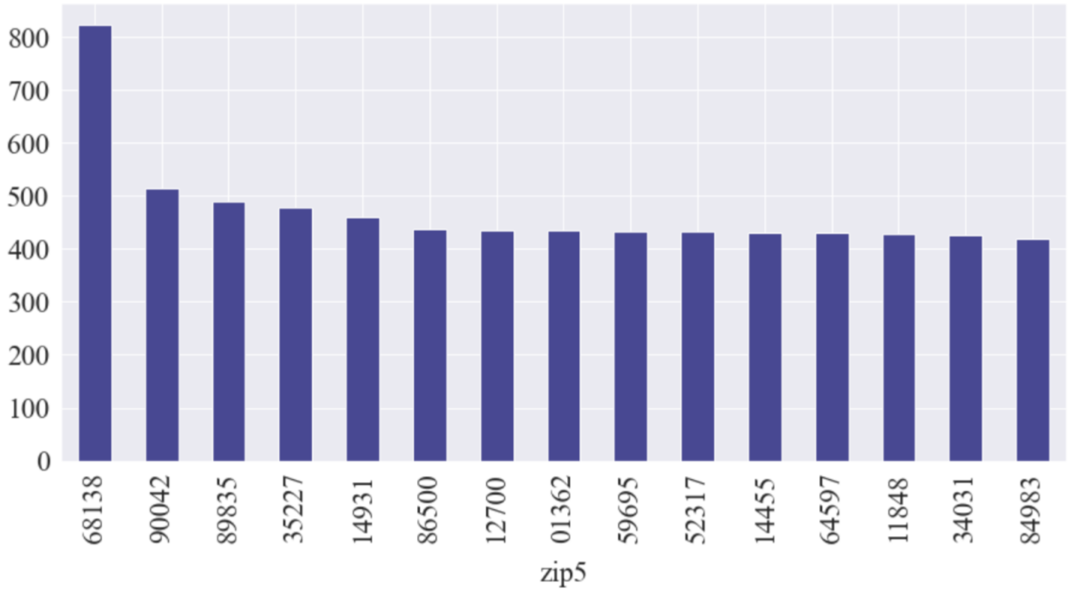
Bar Plot of address (Top 15)



zip5 (Categorical)

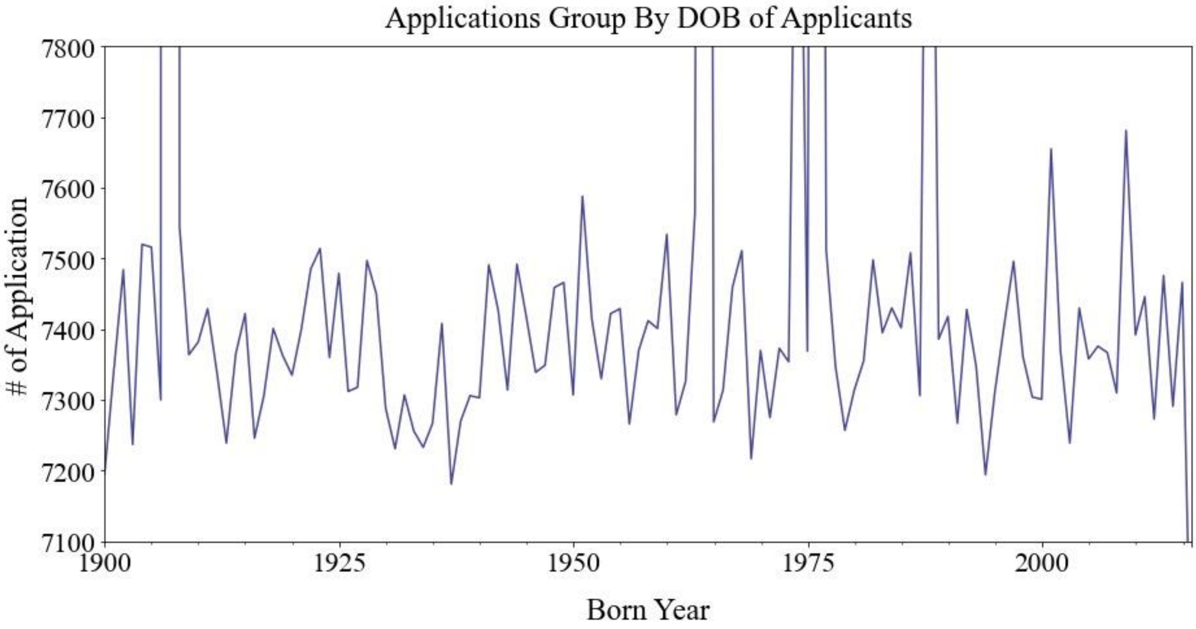
Zip Code. There are 26,370 zip codes and the most common value is 68138.

Bar Plot of zip5 (Top 15)



dob (Numeric - Date)

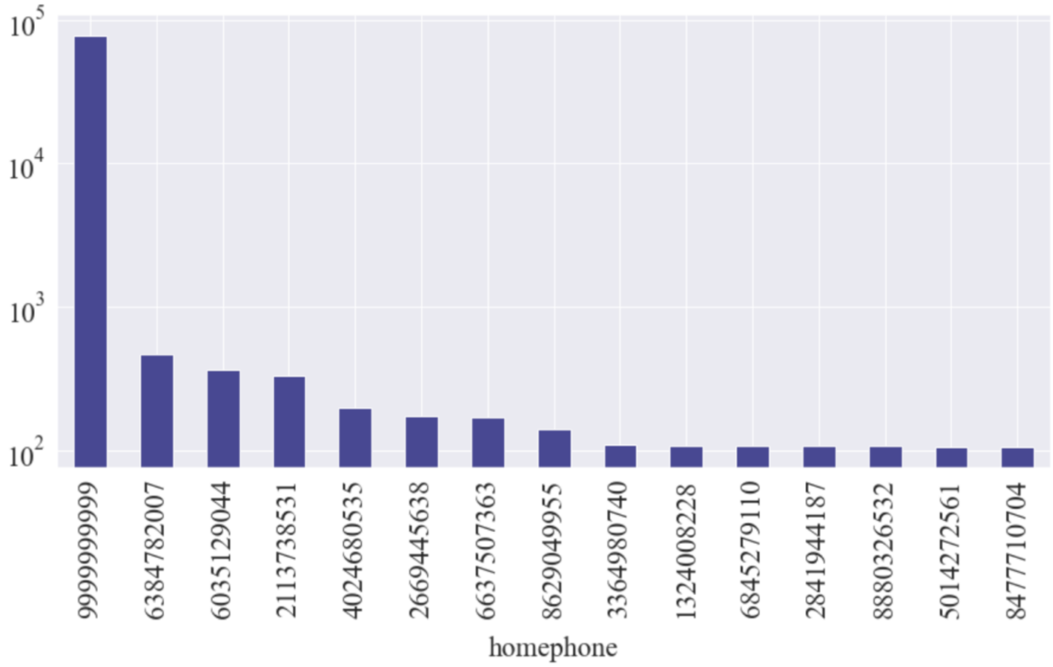
Date of Birth. The minimum date of birth is 1900-01-01 and the maximum date of birth is 2016-10-31.



homephone (Categorical)

Home Phone Number. There are 28,244 unique values and the most common value is 9999999999.

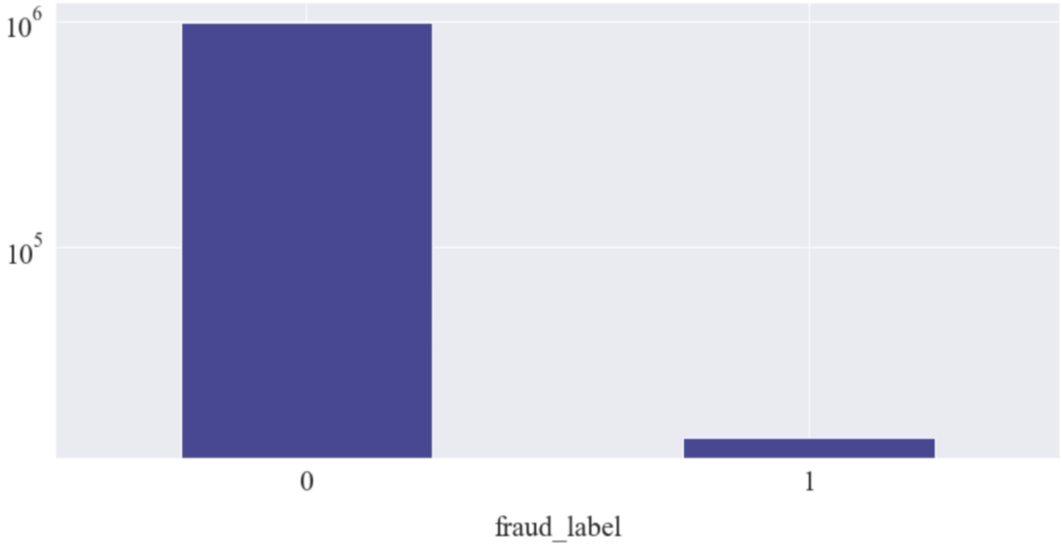
Bar Plot of homephone (Top 15)



fraud_label (Categorical)

Label of Fraud. There are 2 unique values where 1 means fraud and 0 means no fraud. The most common value is 0.

Bar Plot of fraud_label


10^6
10^5
fraud_label