DATA QUALITY REPORT

DQR Prepared By	Chinmayi Bengaluru Prakash
Initials	CBP
Date Created	February 16, 2022
Version No.	1.0

1. File Description

This section provides a high-level description of the dataset file and the dataset.

The table below provides general information on the dataset file:

Data File Name	"applications data.csv"
Data File Description	The dataset represents applications for credit cards and cell phones. The dataset contains date of application received, personal identity information (First Name, Last Name, SSN, Date of Birth, Address, Phone Number), zip code and Fraud Label. Fraud Label categorizes a record as good (not fraud) and bad (fraud).
Granularity	A record for each application.

The table below provides general information on the dataset:

Time Period of Data	January 01, 2016 - December 31, 2016
Number of Fields (Columns)	10
Number of Records (Rows)	1000000
Number of Numeric Fields	2
Number of Categorical Fields	8

2. Fields Summary Tables

This section provides information on two field summary tables listing all fields with summary information.

2.1 Numeric Fields Table

The following table lists all the numeric fields along with their summary statistics:

Field Name	% Populated	Min	Max	Mean	Standard Deviation	% Zero
date	100.00%	2016-01-01	2016-12-31	NA	NA	0.00%
dob	100.00%	1900-01-01	2016-10-31	NA	NA	0.00%

2.2 Categorical Fields Table

The following table lists all the categorical fields along with their summary statistics:

Field Name	% Populated	# Unique Values	Most Common Value
record	100.00%	1,000,000	-
ssn	100.00%	835,819	999999999
firstname	100.00%	78,136	EAMSTRMT
lastname	100.00%	177,001	ERJSAXA
address	100.00%	828,774	123 MAIN ST
zip5	100.00%	26,370	68138
homephone	100.00%	28,244	9999999999
fraud_label	100.00%	2	0

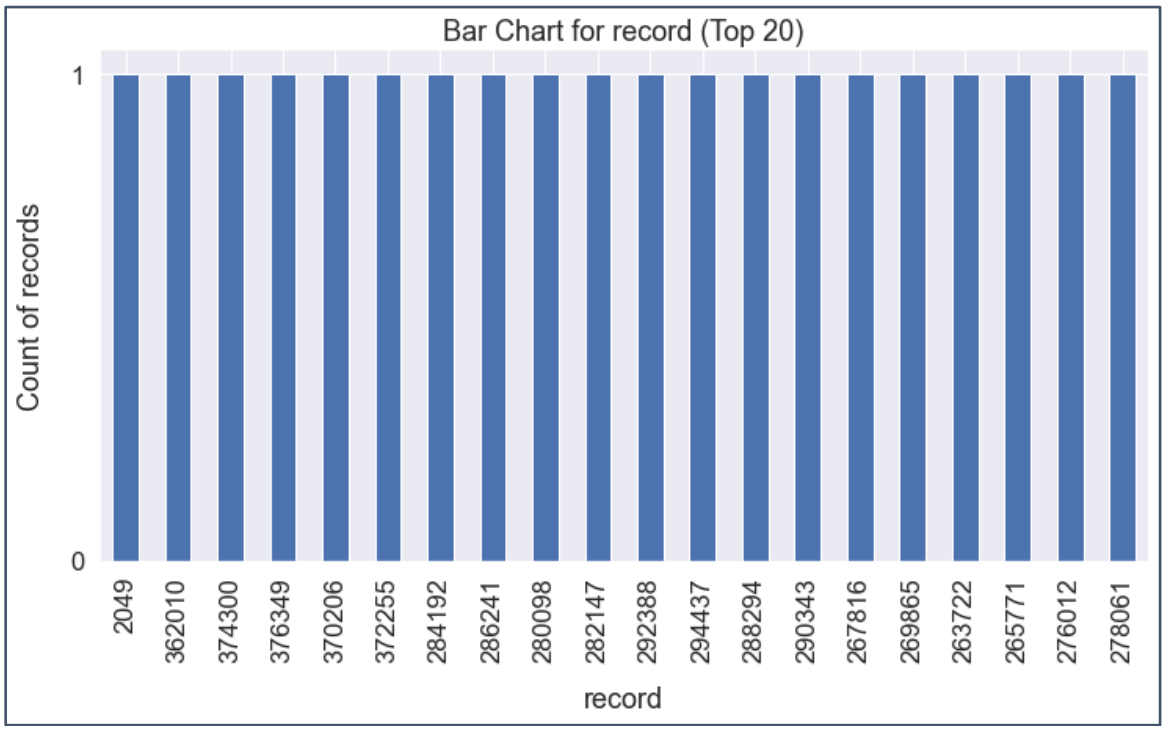
3. Description of the fields

This section provides a brief description for each of the 10 fields:

3.1 record

Field Name	record
Field Description	**Categorical field** with no missing values.
	This field denotes a unique number for each application.
	This field has a unique value for each row.

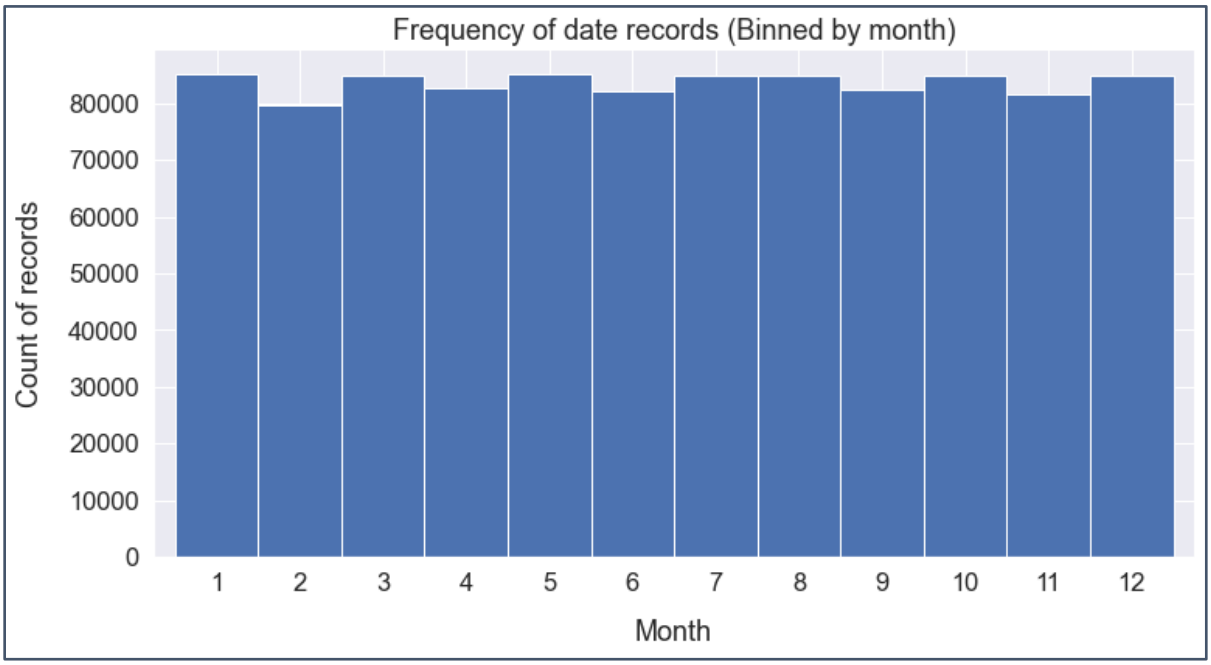
The bar chart below displays top 20 values of "record" field:



3.2 date

Field Name	date
Field Description	**Numerical field** with no missing values.
	This field denotes date of application received in YYYY-MM-DD format.
	The data ranges from 2016-01-01 to 2016-12-31.

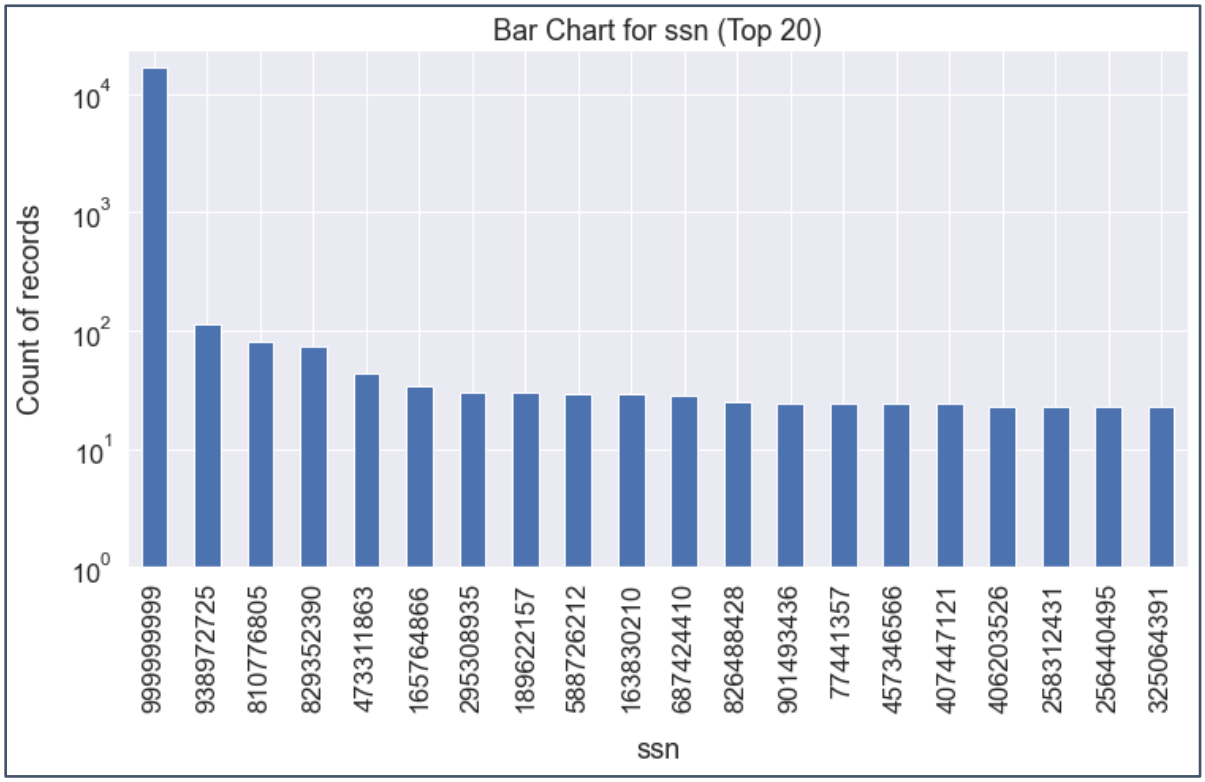
The following is the distribution of values of "date" field binned by 12 months of 2016:



3.3 ssn

Field Name	ssn
Field Description	**Categorical field** with no missing values.
	This field denotes the 9-digit SSN (Social Security Number) of the applicant.
	This field has 835,819 unique values.
	The most common value is 999999999.

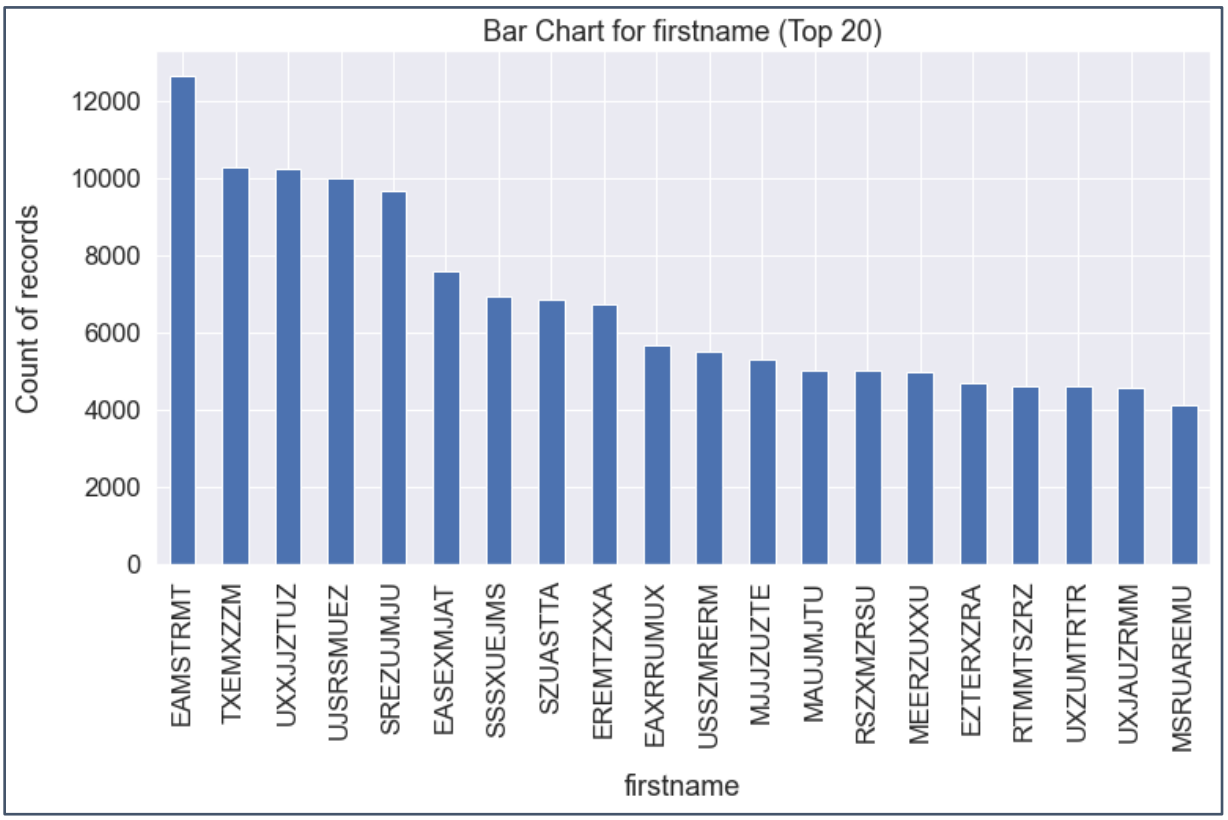
The bar chart below displays top 20 values of "ssn" field:



3.4 firstname

Field Name	firstname
Field Description	**Categorical field** with no missing values.
	This field denotes the first name of the applicant.
	This field has 78,136 unique values.
	The most common value is EAMSTRMT.

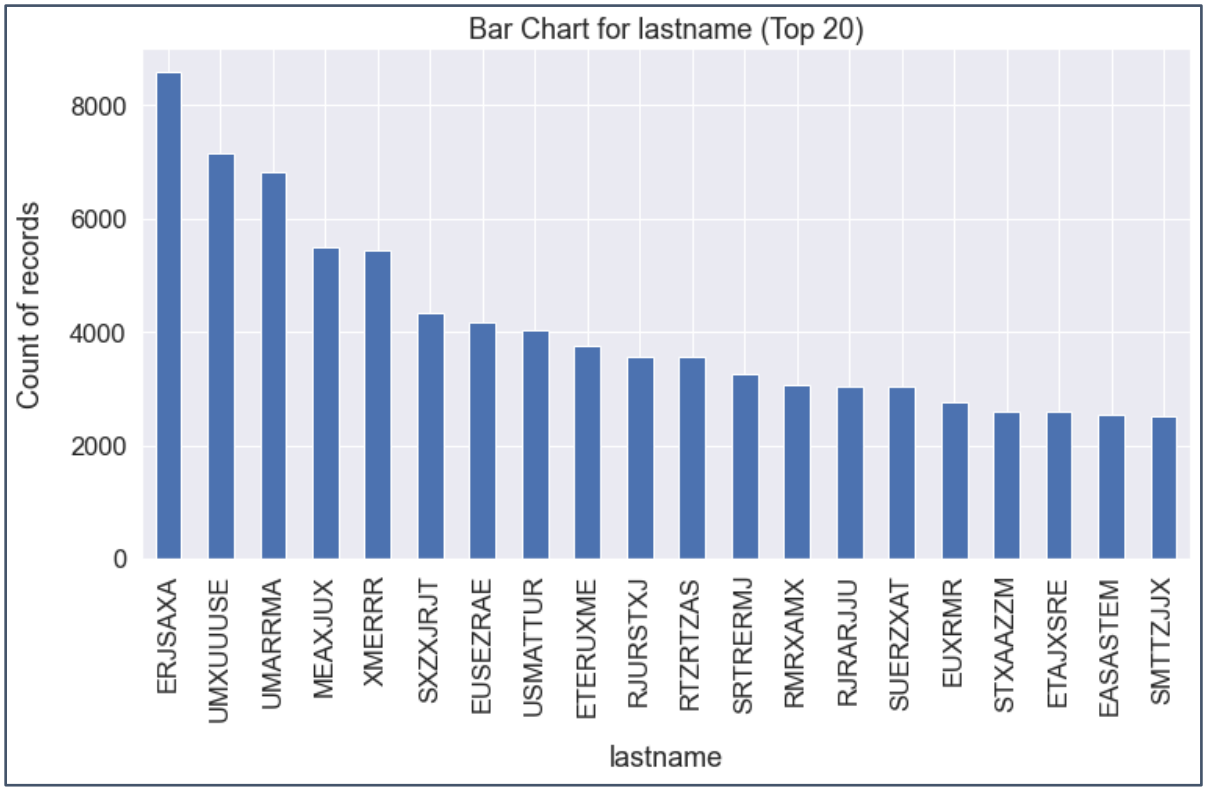
The bar chart below displays the top 20 unique values of "firstname" field:



3.5 lastname

Field Name	lastname
Field Description	**Categorical field** with no missing values.
	This field denotes the last name of the applicant.
	This field has 177,001 unique values.
	The most common value is ERJSAXA.

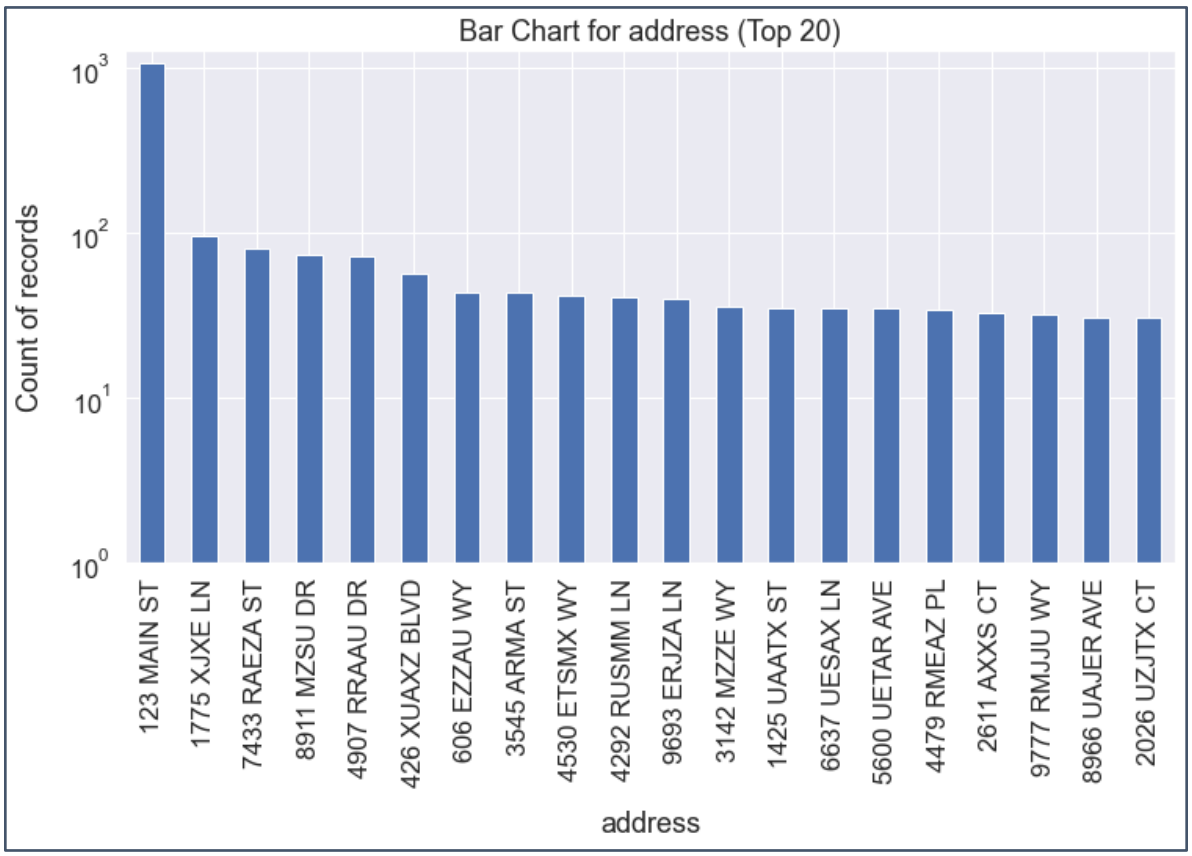
The bar chart below displays the top 20 unique values of "lastname" field:



3.6 address

Field Name	address
Field Description	**Categorical field** with no missing values.
	This field denotes the address of the applicant.
	This field has 828,774 unique values.
	The most common value is 123 MAIN ST.

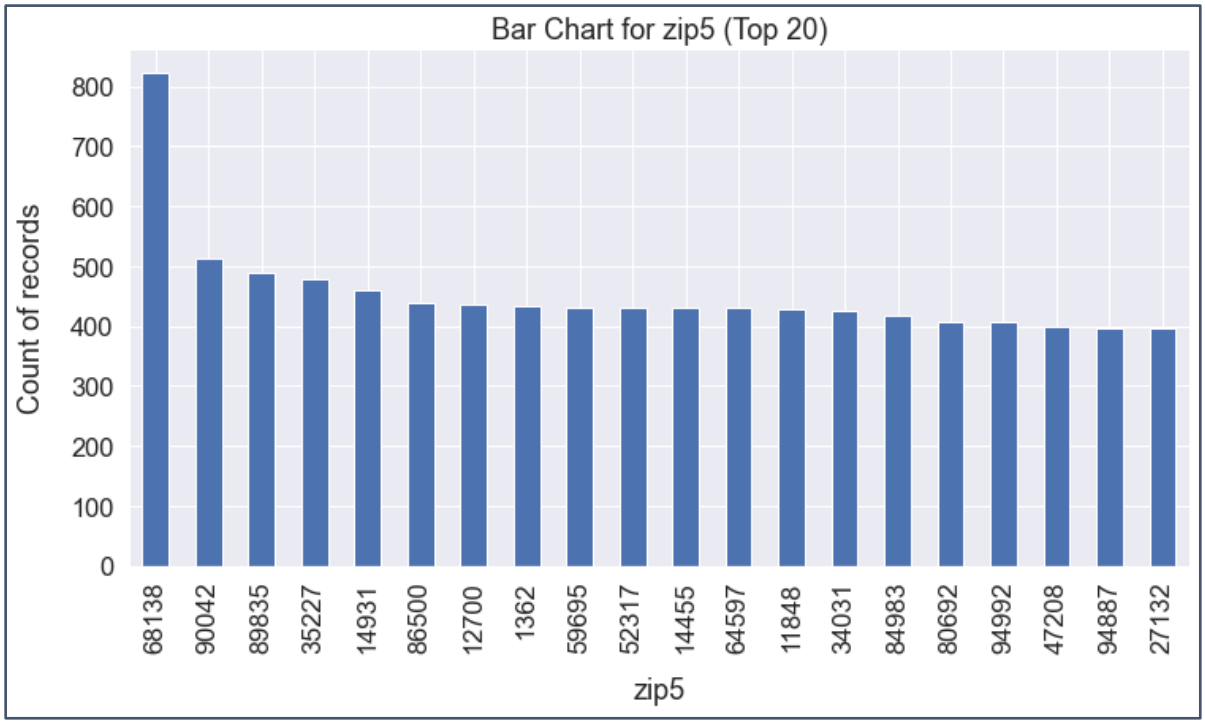
The bar chart below displays the top 20 unique values of "address" field:



3.7 zip5

Field Name	zip5
Field Description	**Categorical field** with no missing values.
	This field denotes the 5-digit zip code where the applicant lives (address).
	This field has 26,370 unique values.
	The most common value is 68138.

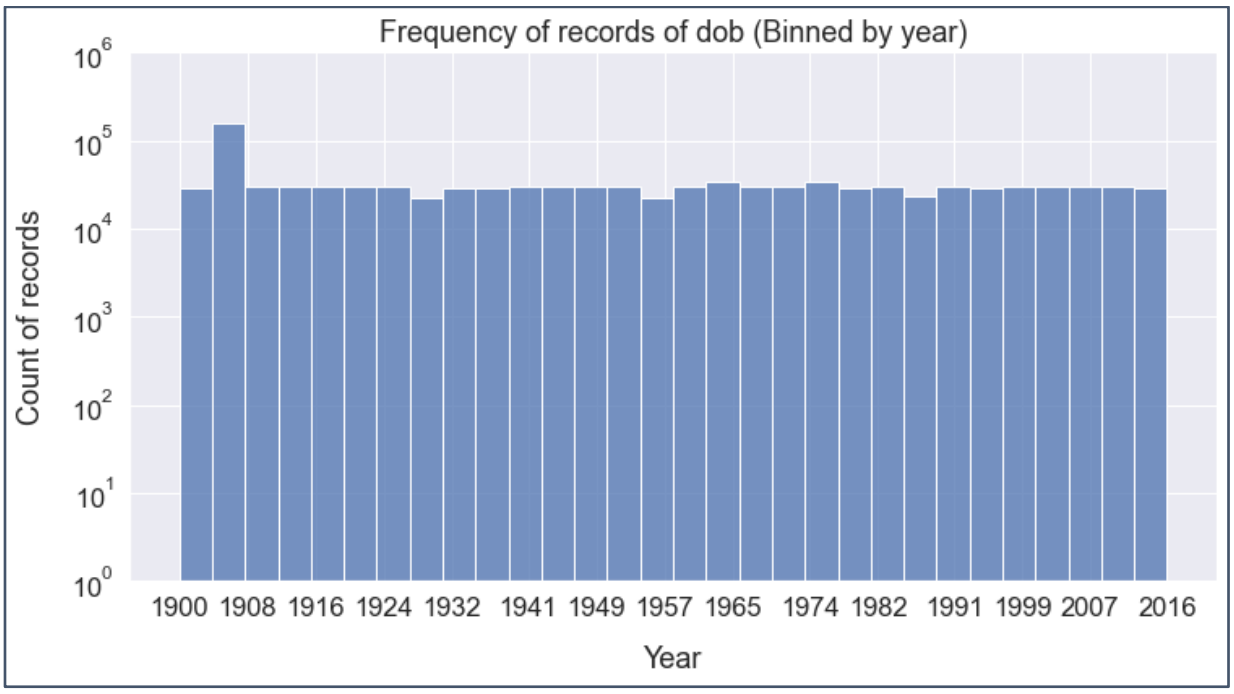
The bar chart below displays the unique values of "zip5" field:



3.8 dob

Field Name	dob
Field Description	**Numerical field** with no missing values.
	This field denotes the date of birth of the applicant in YYYY-MM-DD format.
	The data ranges from 1900-01-01 (oldest applicant) to 2016-10-31 (youngest applicant).

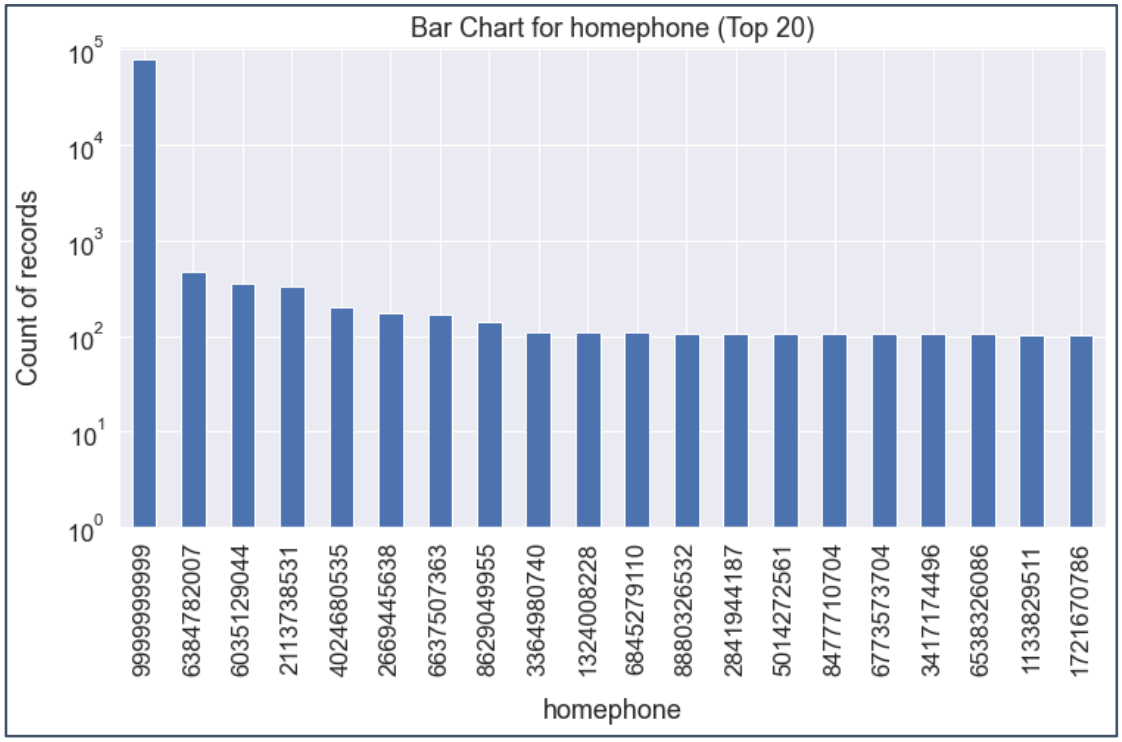
The following is the distribution of values of "dob" field binned by year of birth from 1900 to 2016:



3.9 homephone

Field Name	homephone
Field Description	**Categorical field** with no missing values.
	This field denotes the 10-digit phone number of applicant.
	This field has 28,244 unique values.
	The most common value is 9999999999.

The bar chart below displays the top 20 unique values of "homephone" field:



3.10 fraud_label

Field Name	fraud_label
Field Description	**Categorical field** with no missing values.
	This field denotes whether a record is fraudulent or not. 0 - Not fraud 1 - Fraud
	This field has 2 unique values.
	The most common value is 0.

The bar chart below displays the unique values of "fraud_label" field: