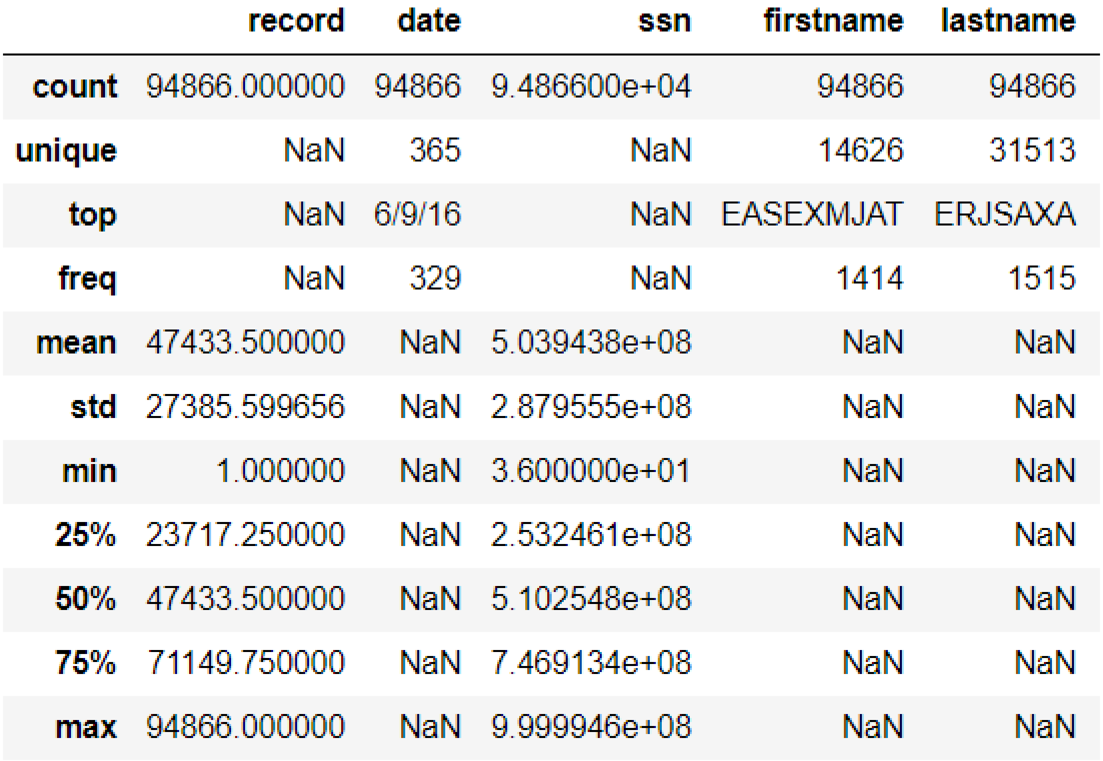
DATA QUALITY REPORT

File Description

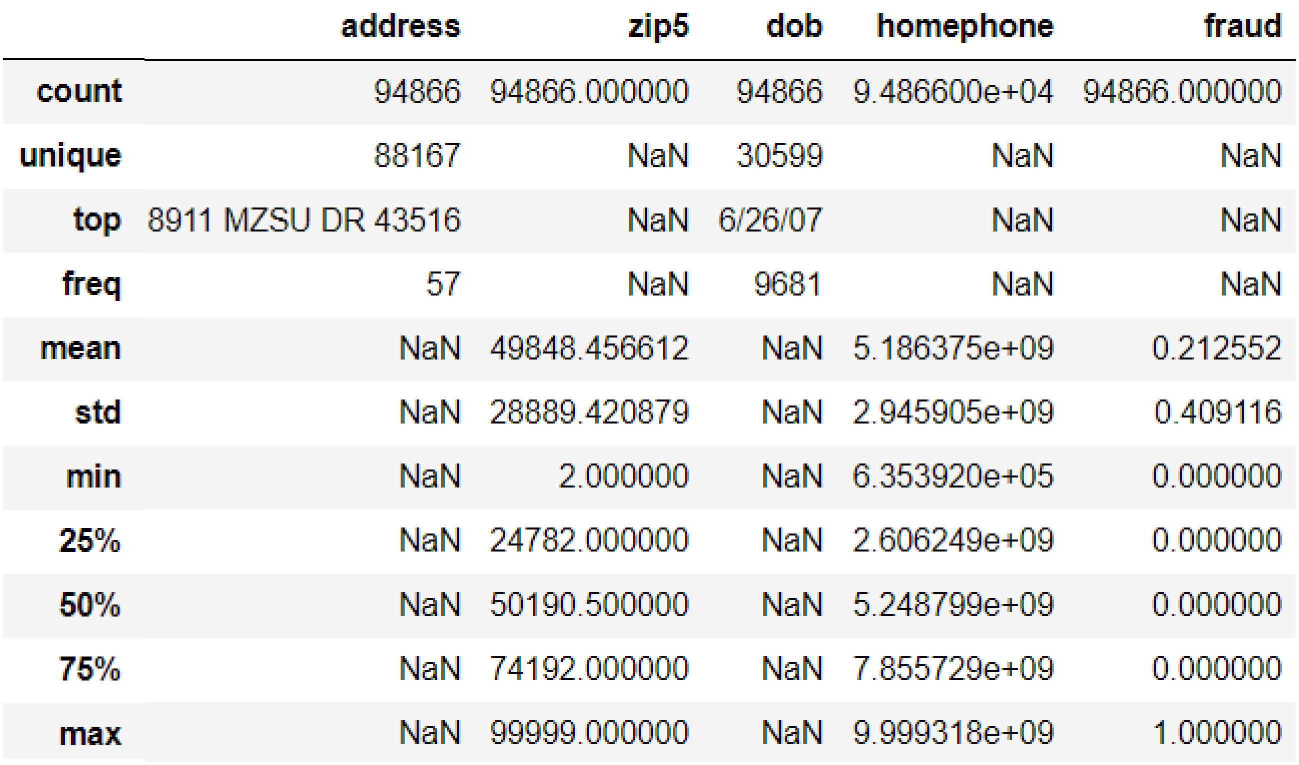
File Name: applications.csv
Description: File contains information people use while filling applications (for credit card, new phone etc.)
Number of Records: 94,866
Number of Features: 10 features

Data Description

Summary Statistics:



	record	date	ssn	firstname	lastname
count	94866.000000	94866	9.486600e+04	94866	94866
unique	NaN	365	NaN	14626	31513
top	NaN	6/9/16	NaN	EASEXMJAT	ERJSAXA
freq	NaN	329	NaN	1414	1515
mean	47433.500000	NaN	5.039438e+08	NaN	NaN
std	27385.599656	NaN	2.879555e+08	NaN	NaN
min	1.000000	NaN	3.600000e+01	NaN	NaN
25%	23717.250000	NaN	2.532461e+08	NaN	NaN
50%	47433.500000	NaN	5.102548e+08	NaN	NaN
75%	71149.750000	NaN	7.469134e+08	NaN	NaN
max	94866.000000	NaN	9.999946e+08	NaN	NaN



	address	zip5	dob	homephone	fraud
count	94866	94866.000000	94866	9.486600e+04	94866.000000
unique	88167	NaN	30599	NaN	NaN
top	8911 MZSU DR 43516	NaN	6/26/07	NaN	NaN
freq	57	NaN	9681	NaN	NaN
mean	NaN	49848.456612	NaN	5.186375e+09	0.212552
std	NaN	28889.420879	NaN	2.945905e+09	0.409116
min	NaN	2.000000	NaN	6.353920e+05	0.000000
25%	NaN	24782.000000	NaN	2.606249e+09	0.000000
50%	NaN	50190.500000	NaN	5.248799e+09	0.000000
75%	NaN	74192.000000	NaN	7.855729e+09	0.000000
max	NaN	99999.000000	NaN	9.999318e+09	1.000000

Feature Description:

1. Record

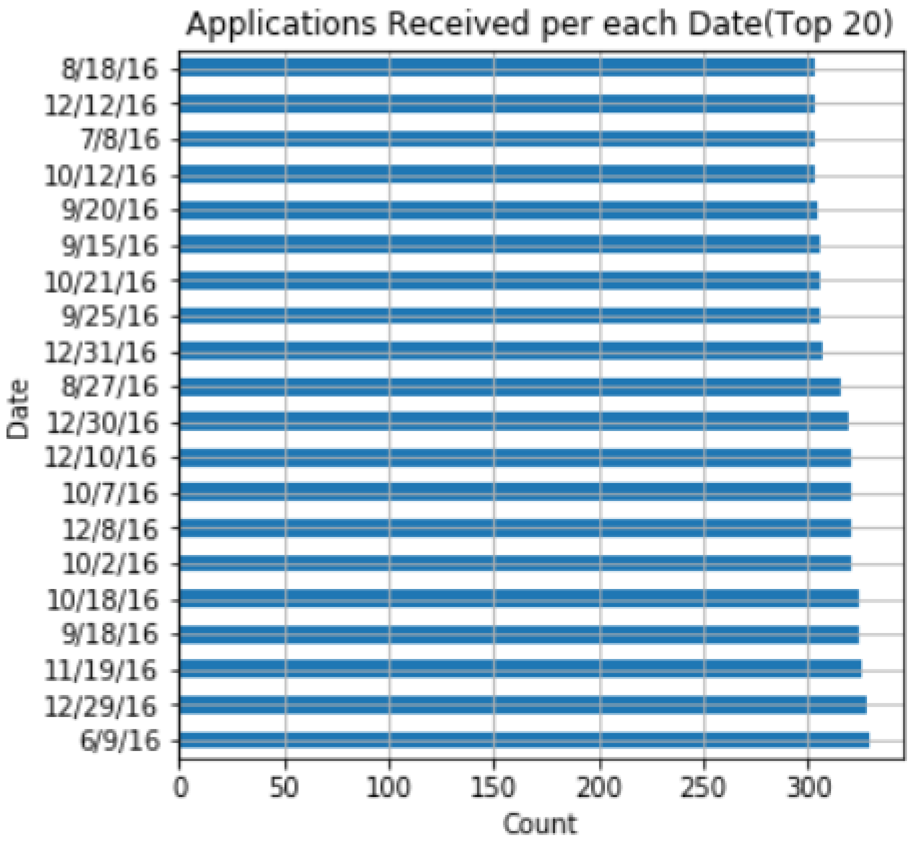
Description: Manually added field to uniquely identify each observation.
Percent Populated: 100%
Unique Values**:** 94,866

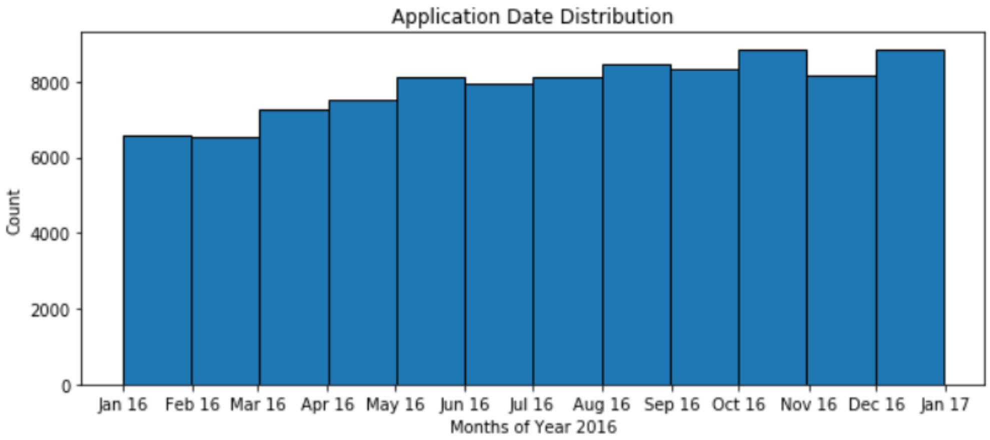
2. Date

Description: Date of applying.

Percent Populated: 100%

Unique Values: 365

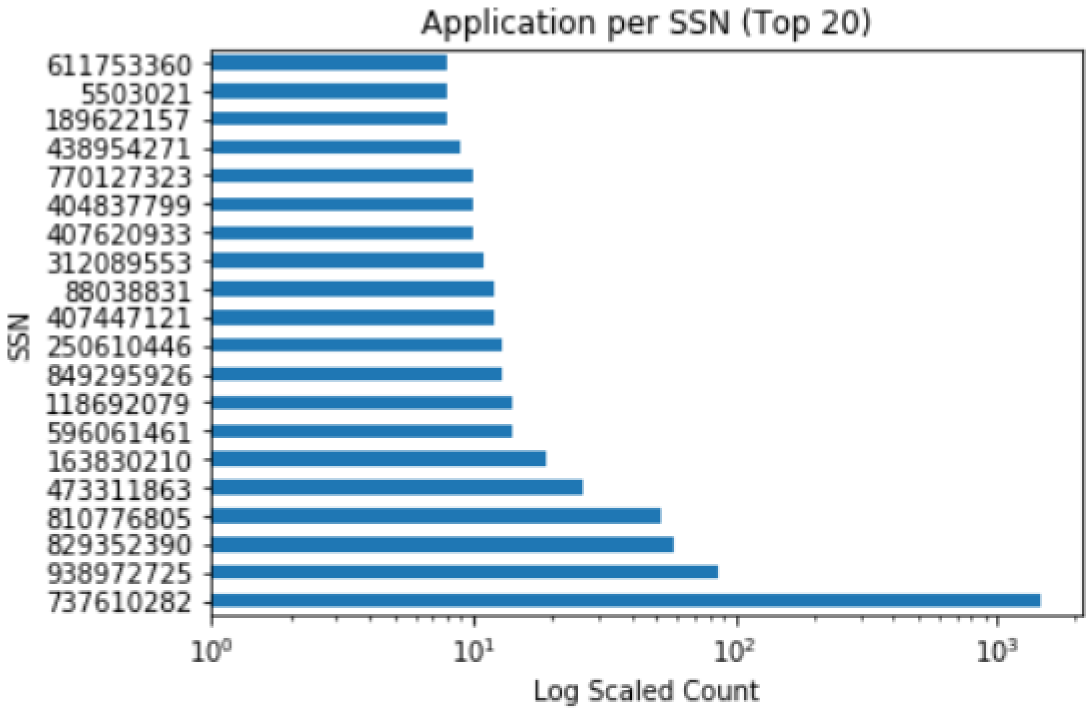




From the dataset, we can see that there is at least one entry for each day of the year.

3. **SSN**

Description: Hashed value of the SSN used by the person for application.
Percent Populated: 100%
Unique Values: 86771



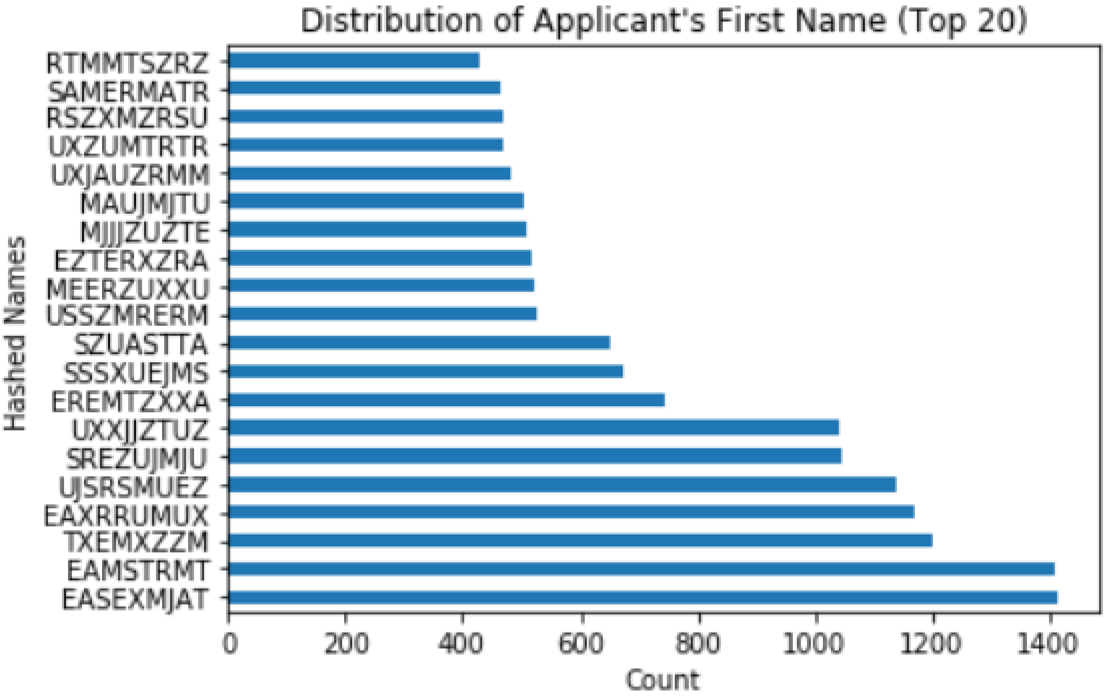
As we can see that one SSN “737610282” has abnormally large number of counts, we conclude that it is the frivolous entry people put in to avoid putting in there real SSN (like 000000000)

4. **First Name**

Description: Hashed value of the first name of the applicant.

Percent Populated: 100%

Unique Values: 14,626

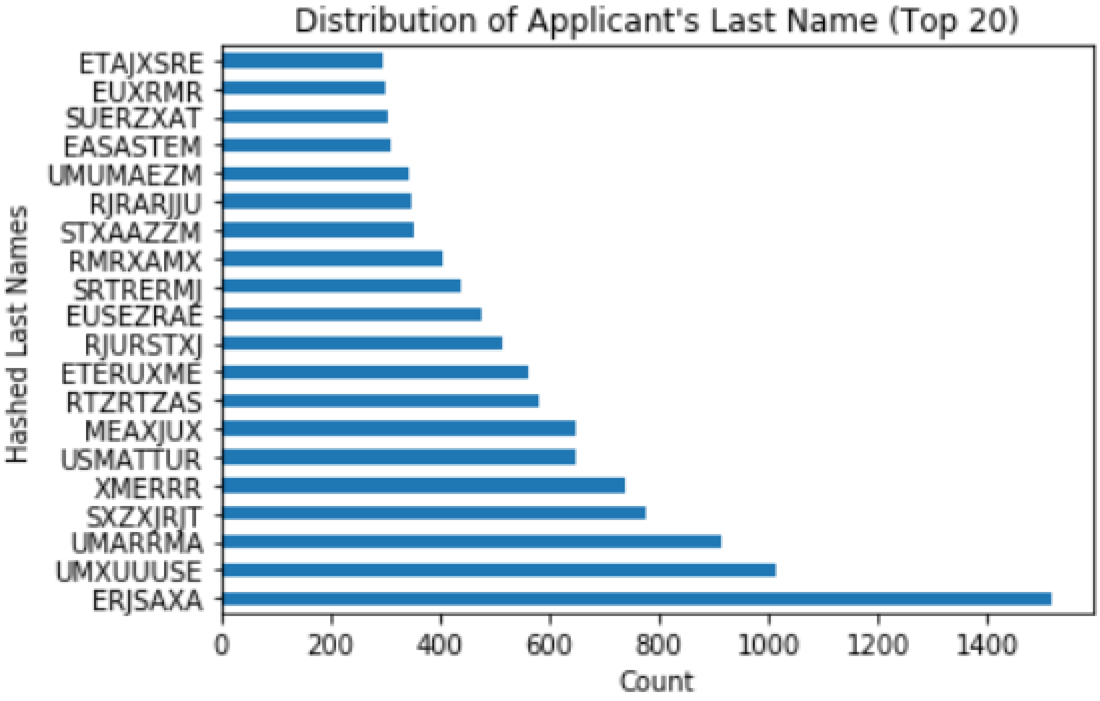


5. Last Name

Description: Hashed value of the last name of the applicant.

Percent Populated: 100%

Unique Values: 31,513

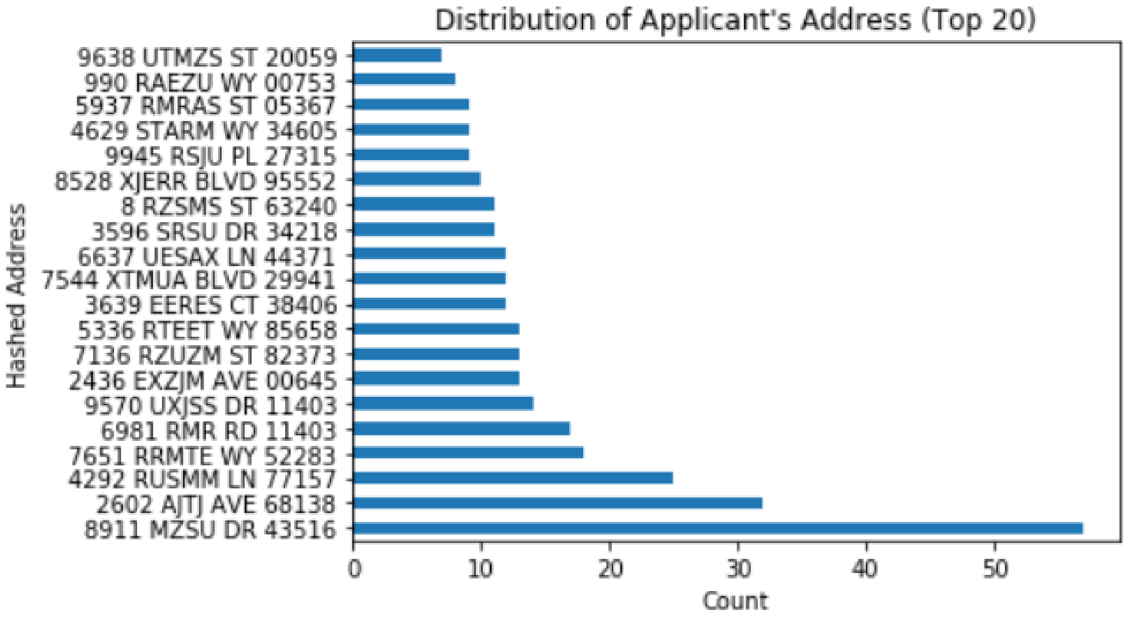


6. Address

Description: Hashed value of address filled by applicant.

Percent Populated: 100%

Unique Values: 88,167

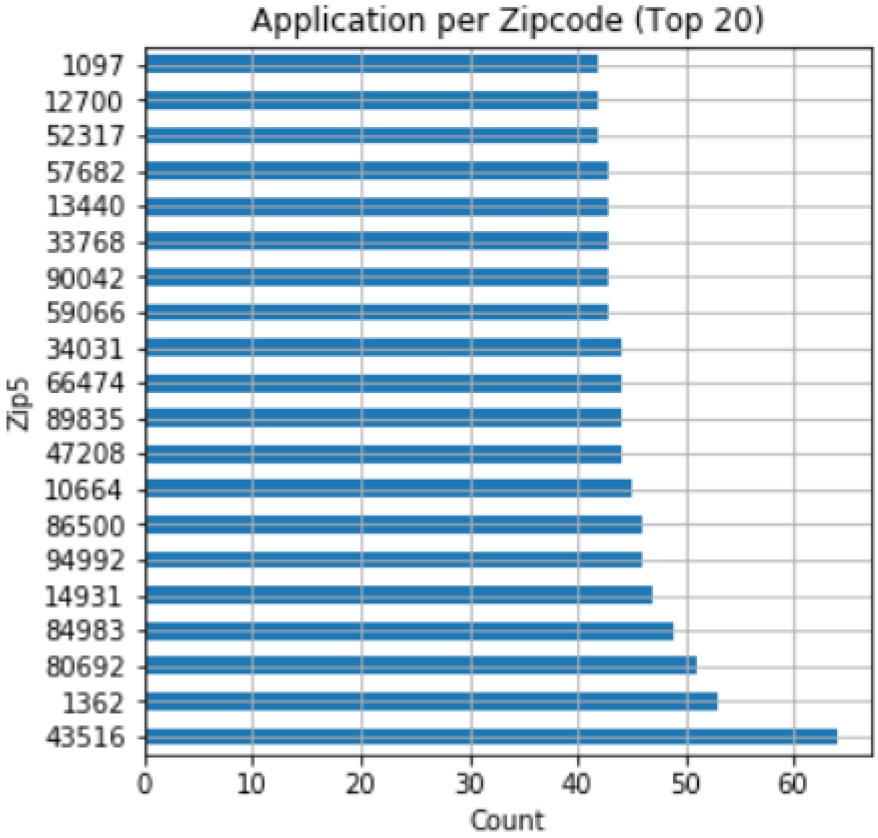


7. **Zip5**

Description: Zip of the applicant

Percent Populated: 100%

Unique Values: 15,855

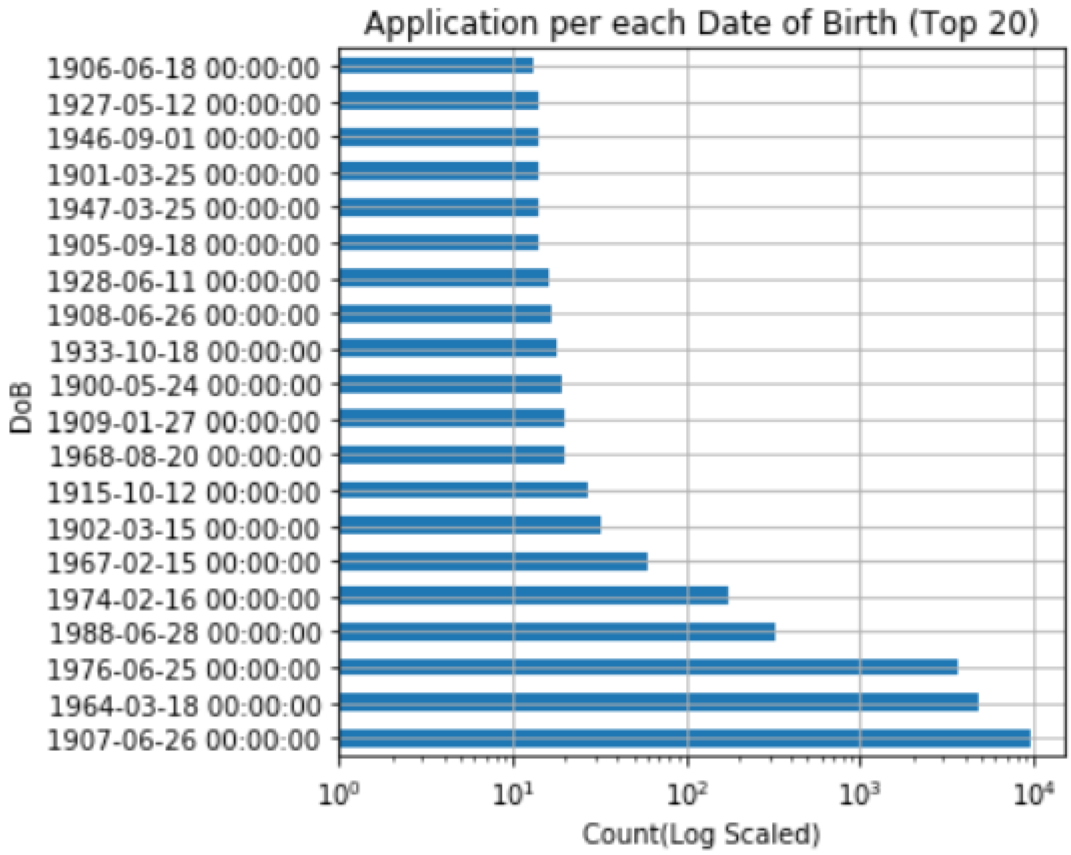


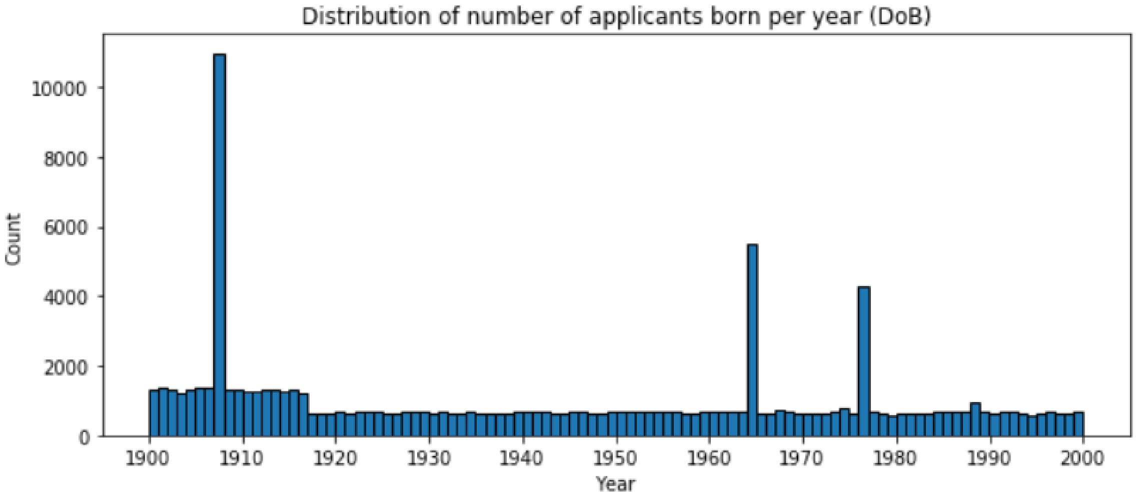
8. DOB

Description: Date of birth of applicant

Percent Populated: 100%

Unique Values: 30,599



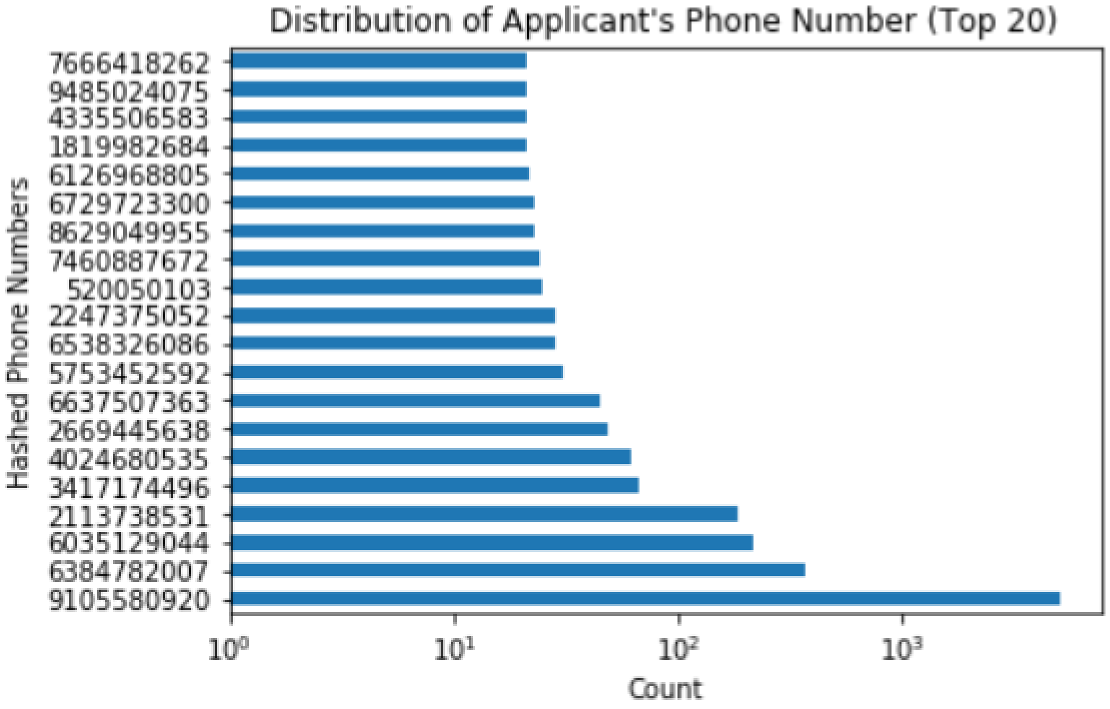


9. Home Phone

Description: Hashed value of phone number of applicant.

Percent Populated: 100%

Unique Values: 20,762



Some phone number may seem to have less digits as they have trailing zeroes. We also see that count of phone number 910-558-0920 is abnormally high. Hence, we conclude that it is a frivolous entry (like 999-999-9999) which people enter when unwilling to share real phone number.

10.Fraud

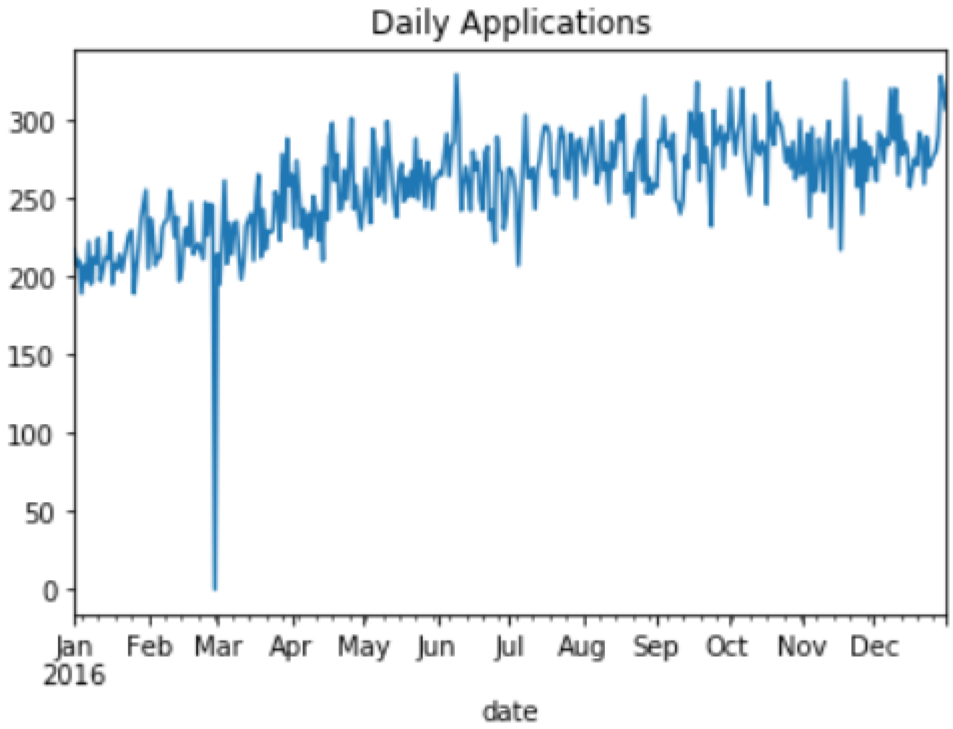
Description: Dependent variable (Telling whether the record is fraud or not)

Percent Populated: 100%

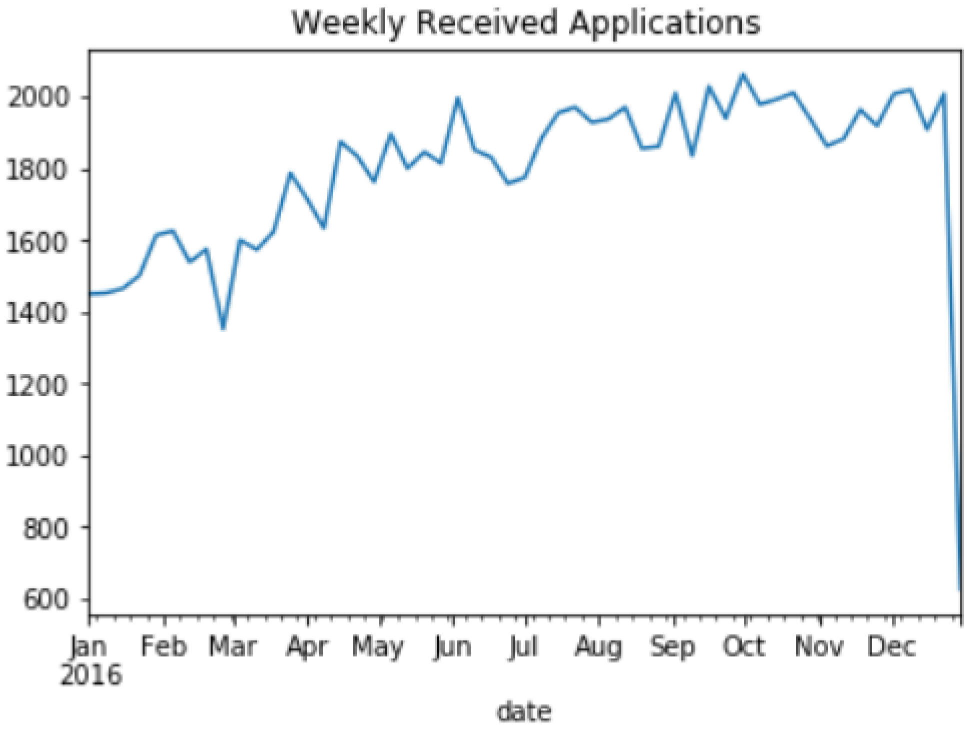
Unique Values: 2 (0 – No fraud, 1- Fraud)

Time Based Analysis:

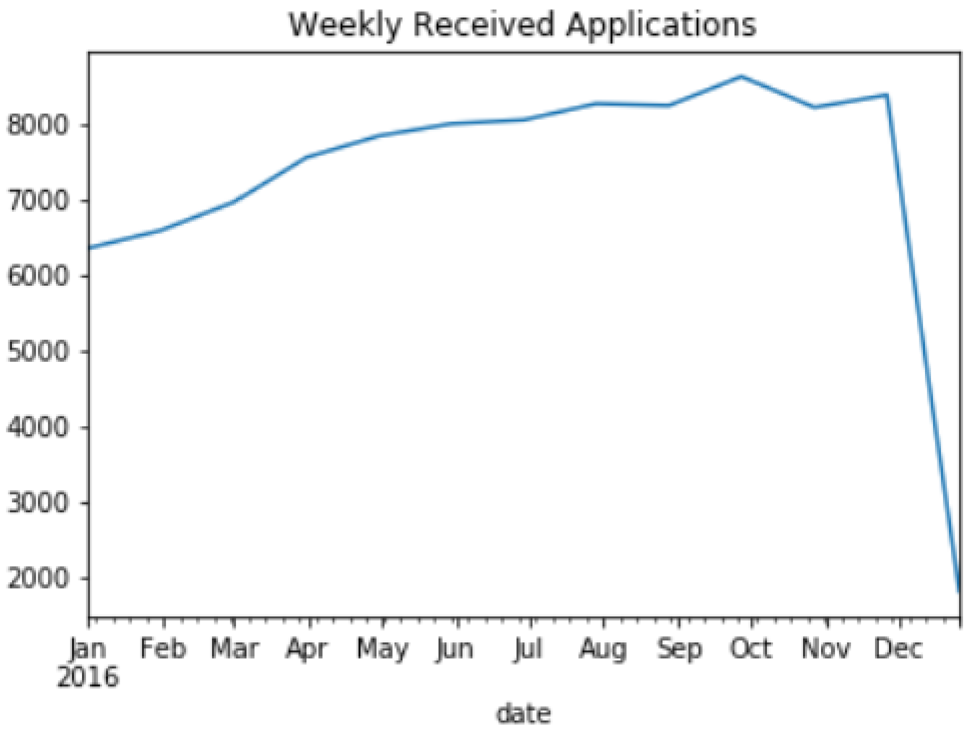
Number of applications received per day-



Number of applications received per week-



Number of applications received per month –


Weekly Received Applications
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
date