DATA QUALITY REPORT

Submitted By – Siddharth Jain

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

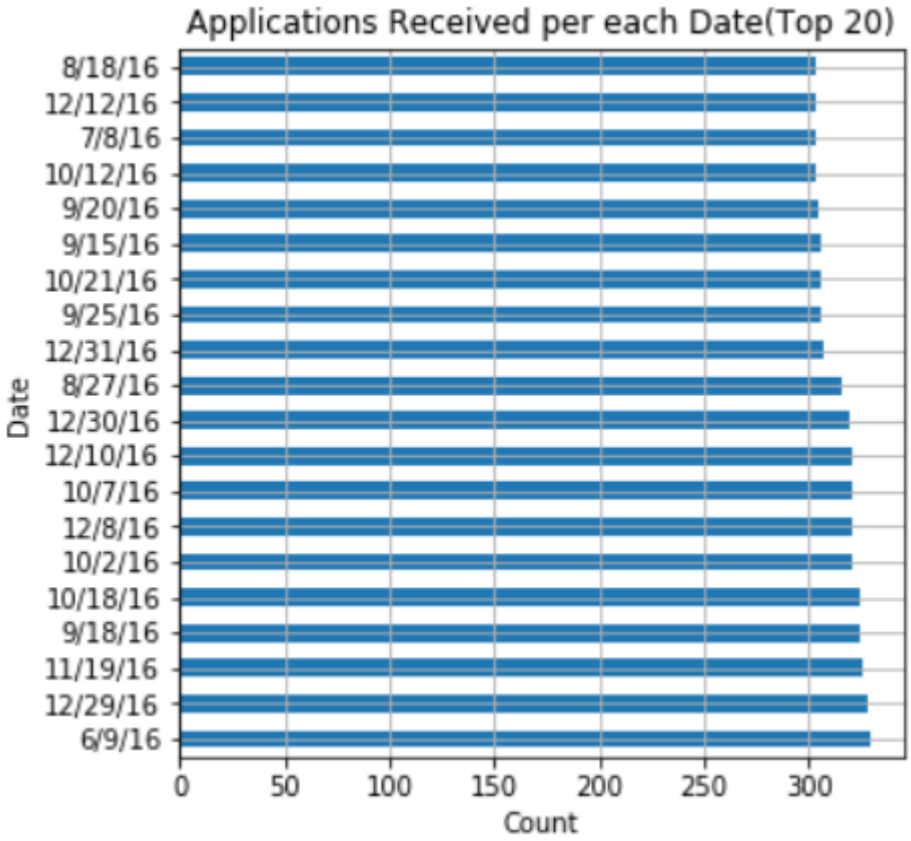
Data Description

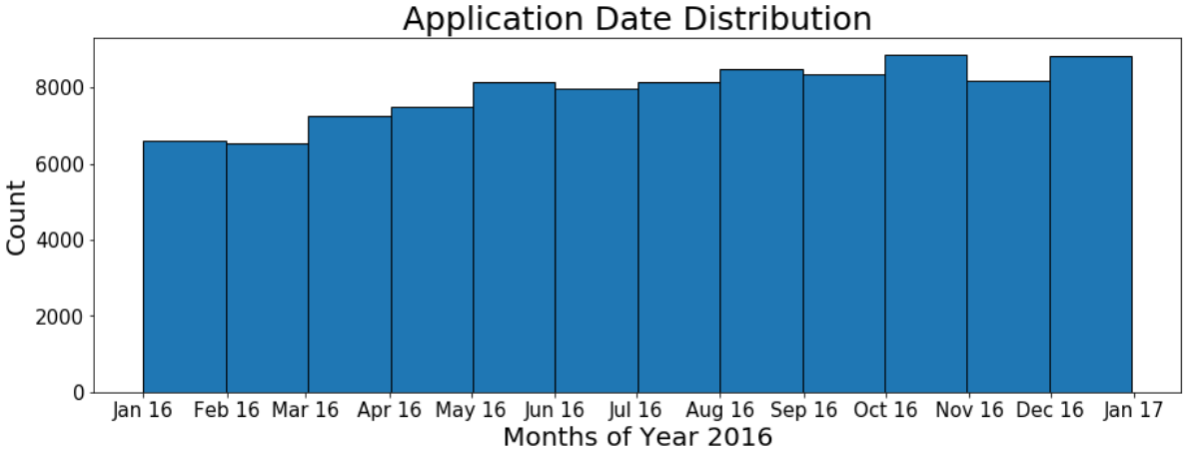
VARIABLE NAME	VARIABLE TYPE
Record	Numeric
Date	Date
SSN	Categorical/String
First Name	Categorical/String
Last Name	Categorical/String
Address	Categorical/String
Zip5	Categorical/String
Home Phone	String (Phone Number)
Date of Birth	Date
Fraud	Categorical

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

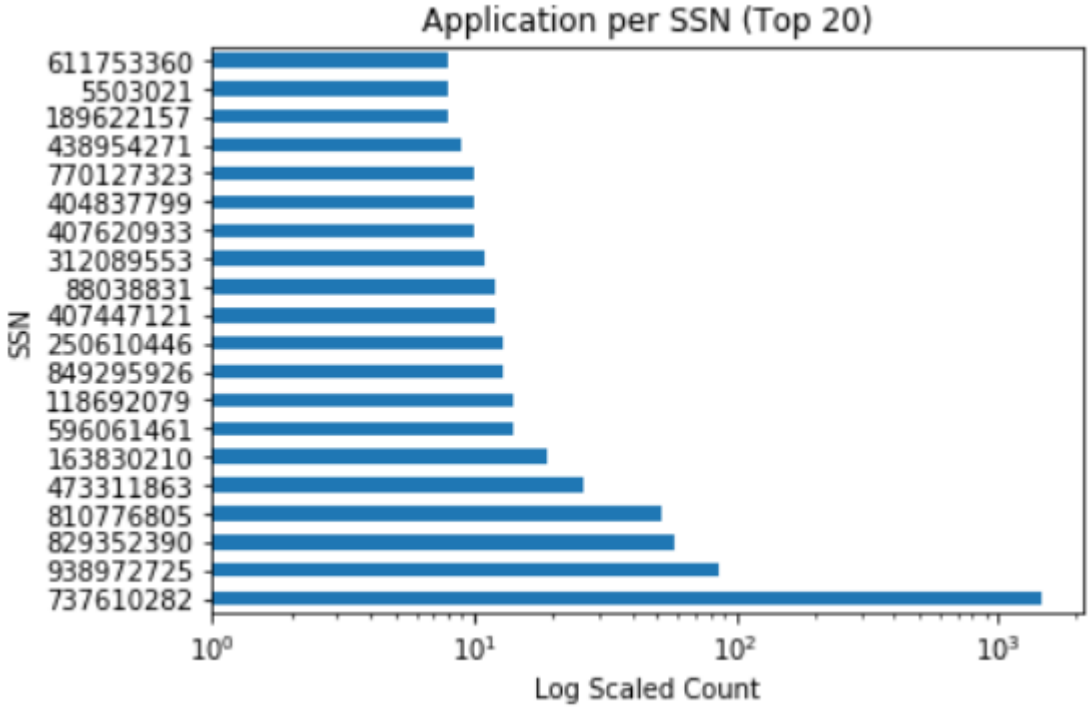




From the dataset, we can see that there is at least one entry for each day of the year.

3. SSN

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



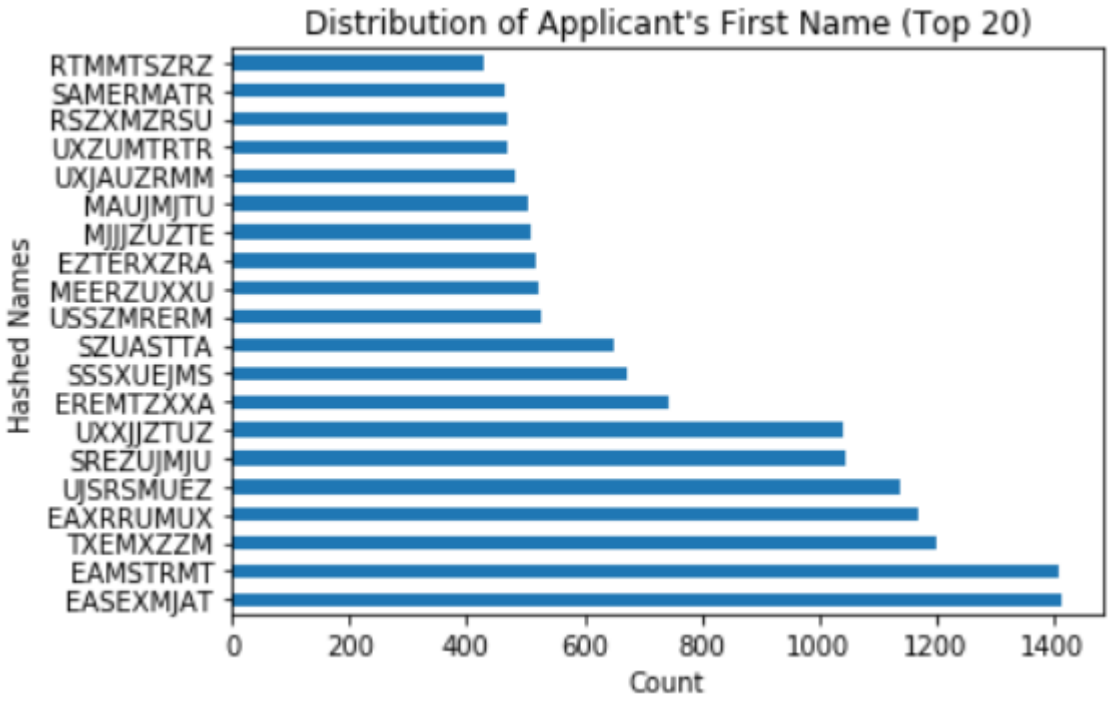
As we can see that one SSN "737610282" has abnormally large number of counts, we conclude that it is the frivolous entry people put in to avoid putting in there real SSN (like 000000000)

4. **First Name**

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

   Percent Populated: 100%

   Unique Values: 14,626

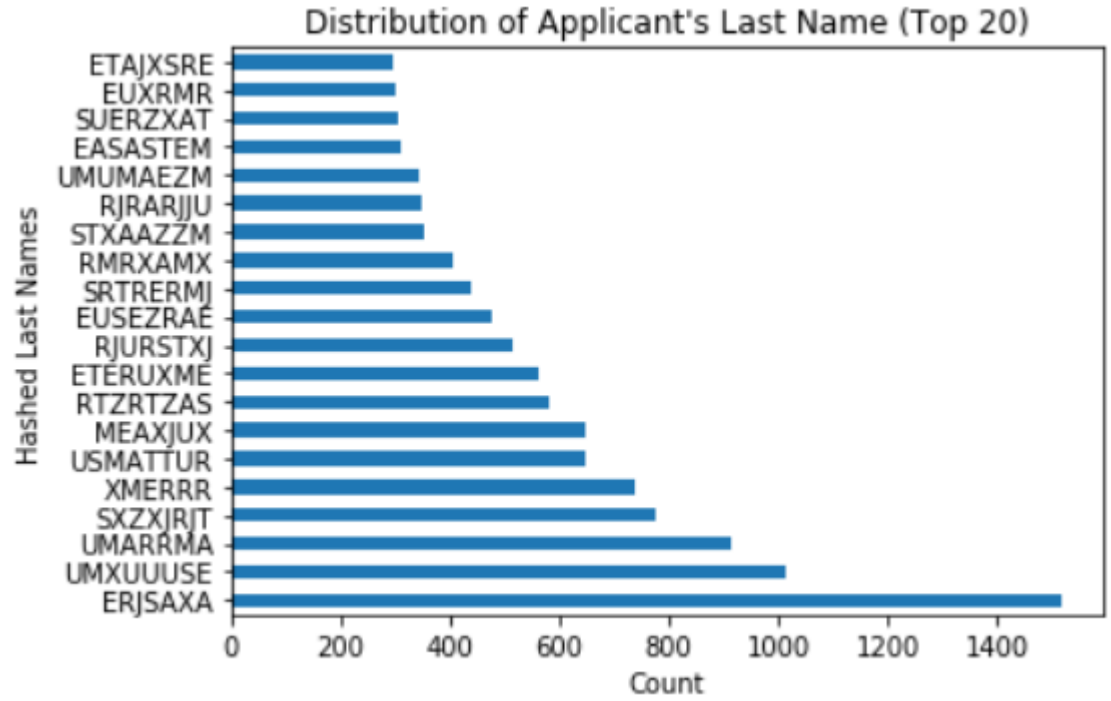


5. Last Name

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

Percent Populated: 100%

Unique Values: 31,513

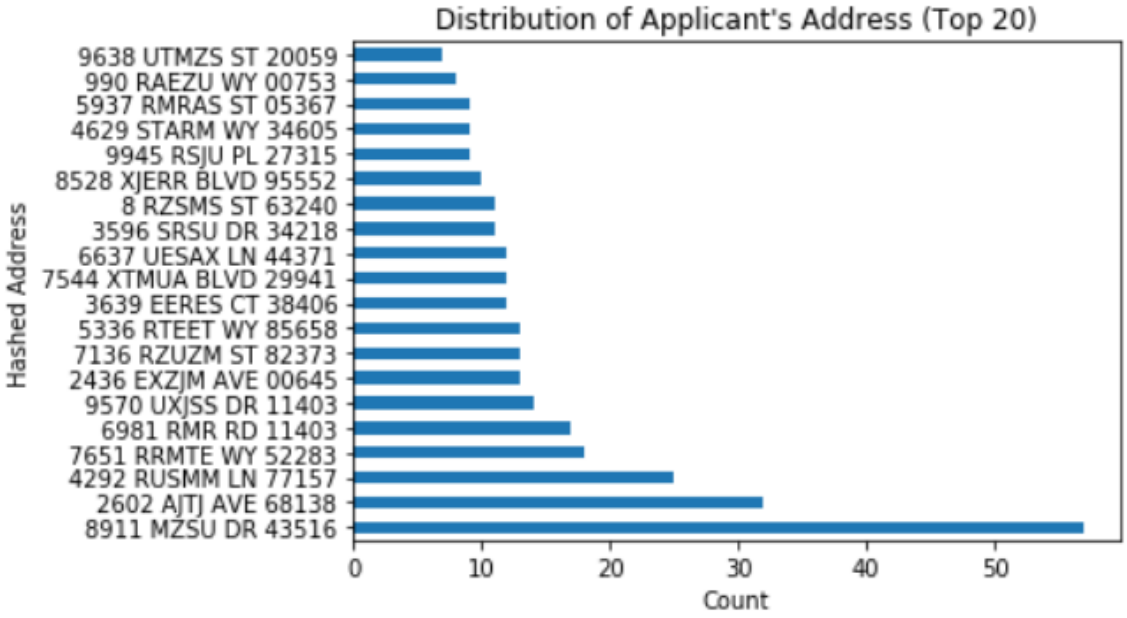


6. **Address**

Description: Hashed value of address filled by applicant.

Percent Populated: 100%

Unique Values: 88,167

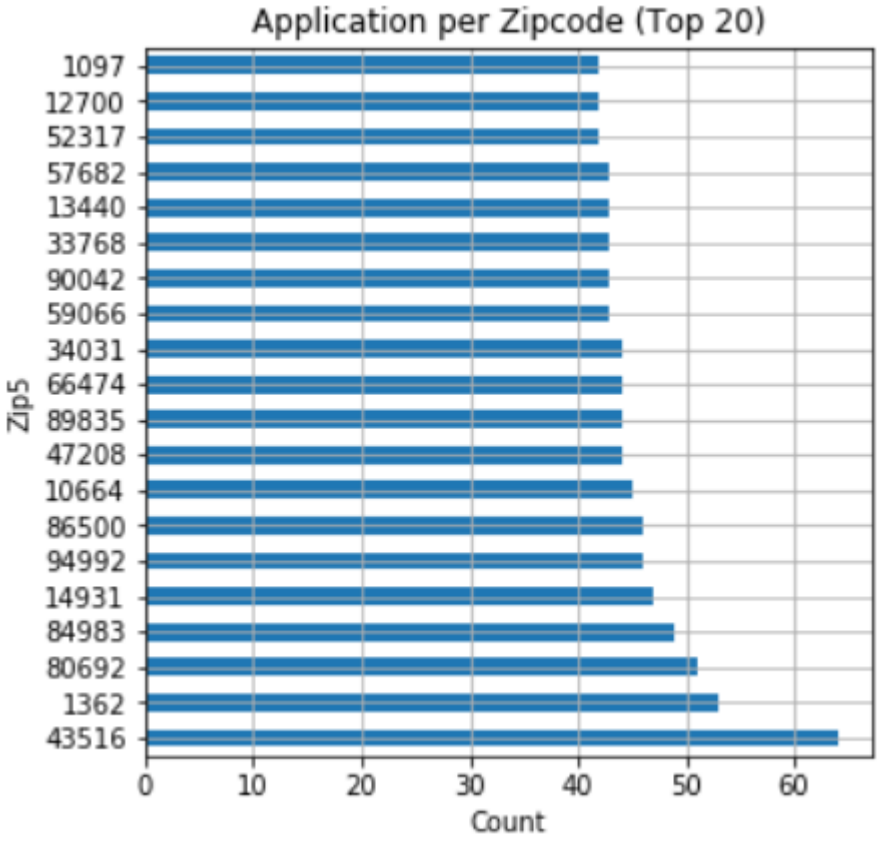


7. **Zip5**

Description: Zip of the applicant

Percent Populated: 100%

Unique Values: 15,855

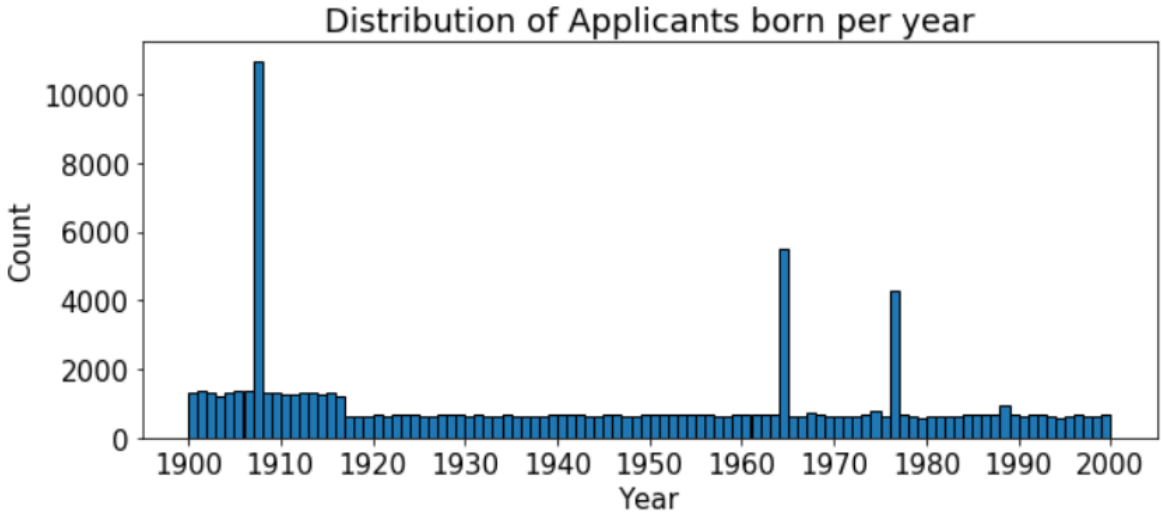


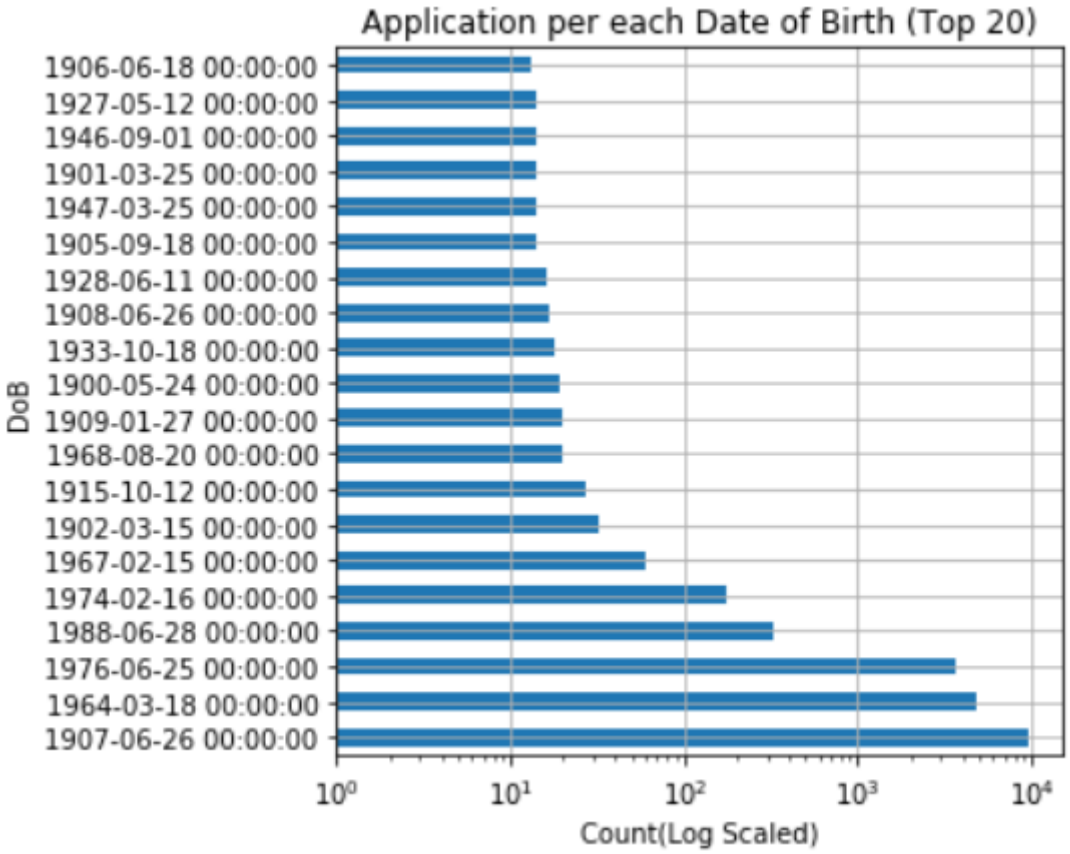
8. DOB

Description: Date of birth of applicant

Percent Populated: 100%

Unique Values: 30,599



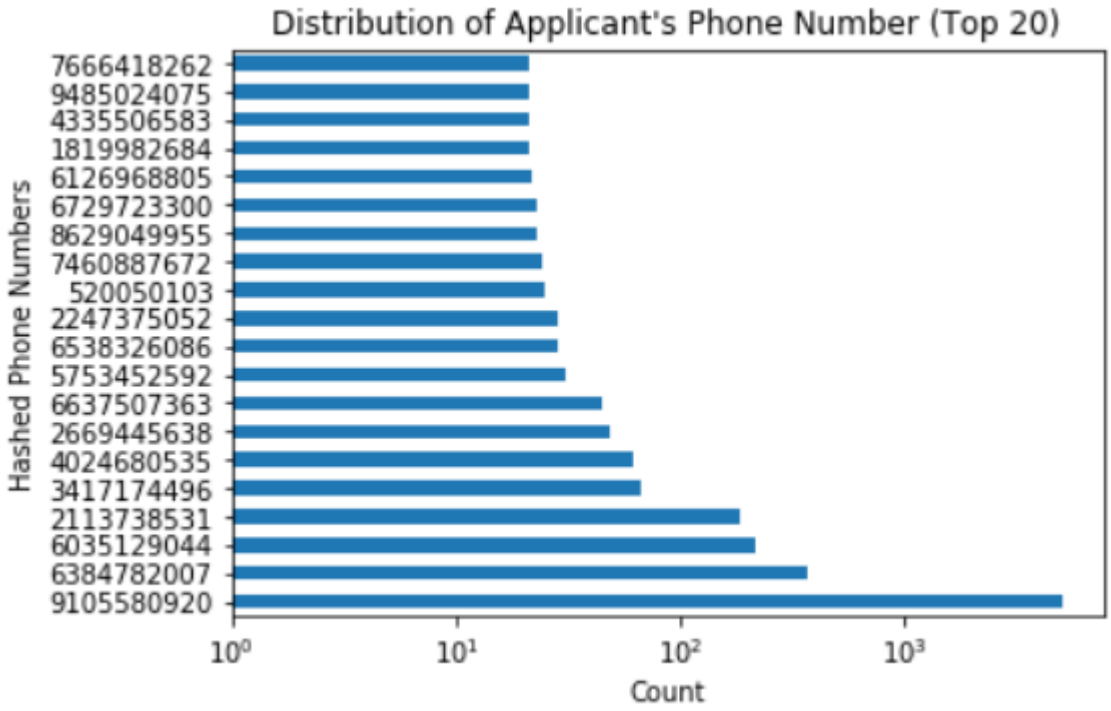


9. Home Phone

Description: Hashed value of phone number of applicant.

Percent Populated: 100%

Unique Values: 20,762



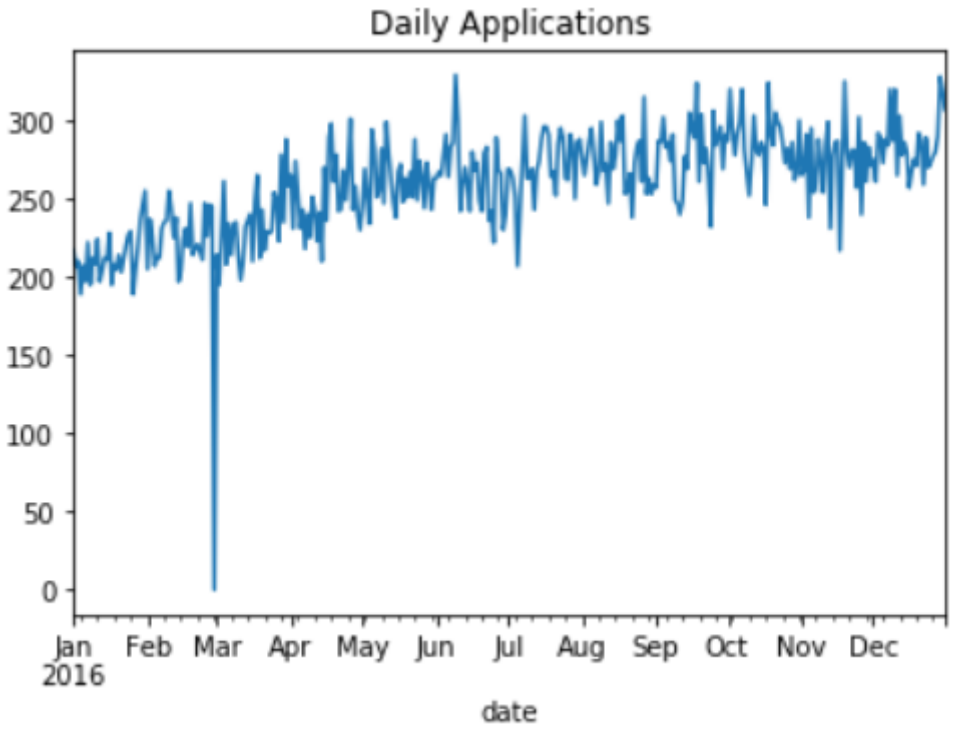
Some phone number may seem to have less digits as they have trailing zeroes. We also see that count of phone number 910-558-0920 is abnormally high. Hence, we conclude that it is a frivolous entry (like 999-999-9999) which people enter when unwilling to share real phone number.

10.Fraud

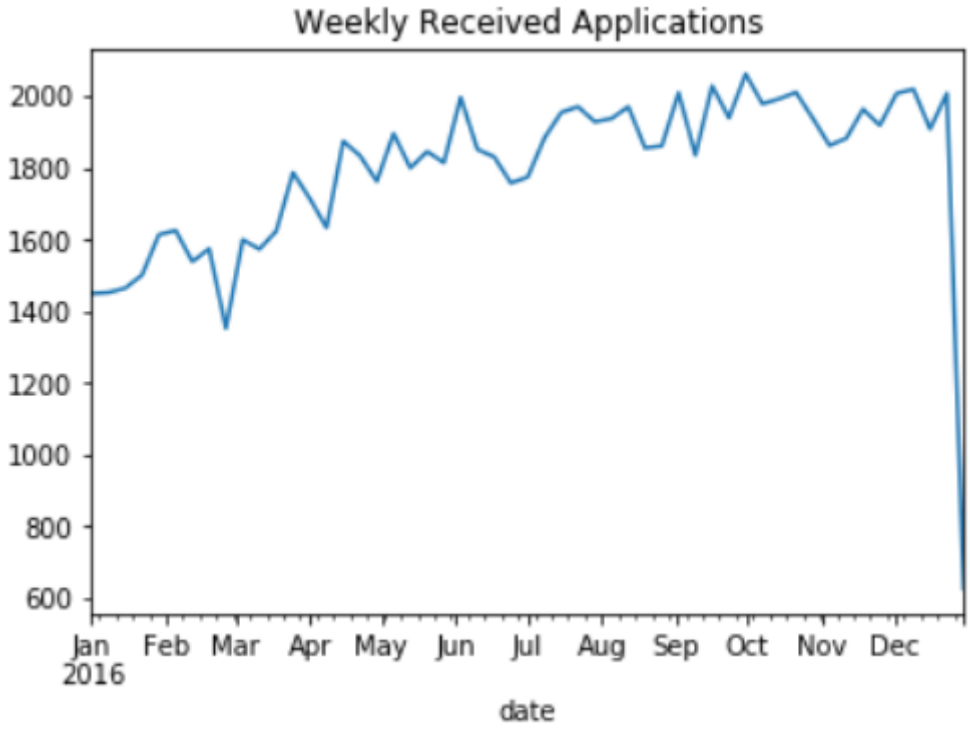
Description: Dependent variable (Telling whether the record is fraud or not)
Percent Populated: 100%
Unique Values: 2 (0 – No fraud, 1- Fraud)

Time Based Analysis:

Number of applications received per day-



Number of applications received per week-



Number of applications received per month –