**DATA QUALITY REPORT FOR APPLICATION DATA**

**Description**

**Dataset Name**: Product Applications Data

**Dataset Purpose**: Build a supervised fraud detection model using records of identity information from applications for credit cards and cell phones

**Data Source**: From an identity fraud prevention company. Synthetic data made from analysis on over a billion real U.S. applications over about 10 years to reproduce the important univariate and multivariate field distributions of real data.

**Time Period**: 01-01-2016 to 12-31-2016

**Number of Fields**: #10

**Number of Records**: 1,000,000

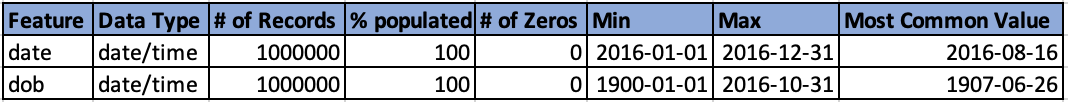
**Summary Table**

**Categorical Field:**

Table

Description automatically generated

**Date/time Field:**

****

**Data Field Exploration:**

Field 1: record

Description: Unique identifier of each application in the dataset.

Field 2: date

Description: Date of applications submitted. The plot shows that there is not a significant pattern that people prefer any specific dates to submit the applications.

Chart, histogram

Description automatically generated

Field 3: ssn

Description: The SSN number that applicant used to apply the credit cards or cell phones. In general, SSN should not change for a specific person. The plot shows the top 15 frequent SSN are used for applications. There are several SSN have been used too many times, which is strange.

Chart, histogram

Description automatically generated

Field 4: firstname

Description: First name the applicants used for applications. The plot shows the top 15 frequent first name.

Chart, bar chart

Description automatically generated

Field 5: lastname

Description: Last name the applicants used for applications. The plot shows the top 15 frequent last name.

Chart, bar chart

Description automatically generated

Field 6: address

Description: Physical address the applicants used for applications. The plot shows the top 15 frequent addresses. Several addresses appear too many times which may need further investigation.

Chart, histogram

Description automatically generated

Field 7: zip5

Description: Zip code of the address that applicants used for applications. The plot shows the top 15 frequent zip code.

Chart, histogram

Description automatically generated

Field 8: dob

Description: Date of birth that applicants used for applications. The histogram plot shows the distribution of birthday of applicants (bins = 29, every 4 years).

Chart, histogram

Description automatically generated

Field 9: homephone

Description: Home phone that applicants used for applications. The plot shows the top 15 frequent home phone number.

Chart, histogram, waterfall chart

Description automatically generated

Field 10: fraud\_label

Description: The fraud label of each application. If it is a fraud, the label will be 1. Otherwise, it will be 0. There are total 985607 applications labeled as 0, and 14393 applications labeled as 1.

Chart, bar chart

Description automatically generated