**DATA QUALITY REPORT FOR APPLICATION DATA**

**Team members**

Yimeng Huang [yimenghu@usc.edu](mailto:yimenghu@usc.edu)

Lingxiao Lyu [llyu@usc.edu](mailto:llyu@usc.edu)

Haotian Wu 1613217870 [hwu41198@usc.edu](mailto:hwu41198@usc.edu)

Jinze Xin [jinzexin@usc.edu](mailto:jinzexin@usc.edu)

Yijia Xu [yijiaxu@usc.edu](mailto:yijiaxu@usc.edu)

Ruoyan Zhou [ruoyanzh@usc.edu](mailto:ruoyanzh@usc.edu)

**Description**

**Dataset Name**: Applications Data

**Dataset Purpose**: assessments on product application data. It’s a synthetic data set built from studying the statistical properties of more than a billion applications. It is built so as to reproduce the important univariate and multivariate field distributions of real data.

**Data Source**: N/A

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

**Number of Fields**: #10

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

**Summary Table**

**Categorical Fields:**

**Table

Description automatically generated**

**Field Description & Visualization**

Field 1:

Name: record

Description: Unique identifier of each entry in the data.

Field 2:

Name: date

Description: Date of application.

First 30 days’ count of applications:

Background pattern

Description automatically generated

Count application by month:

Chart

Description automatically generated

Count application by weekday:

A picture containing icon

Description automatically generated

Count application by day:

Background pattern

Description automatically generated

Field 3:

Name: ssn

Description: Social Security Number of applicants. 999999999 is counted as frivolous value

Top 20 counts of SSN in log scale:

Histogram

Description automatically generated

Field 4:

Name: firstname

Description: First name of applicants.

Top 20 counts of first name in log scale:Chart, histogram

Description automatically generated

Field 5:

Name: lastname

Description: Last name of applicants.

Top 20 counts of last name in log scale:Chart, histogram

Description automatically generated

Field 6:

Name: address

Description: Addresses of applicants.

Top 20 counts of address in log scale: **Histogram

Description automatically generated**

Field 7:

Name: zip5

Description: 5-digit zip code of applicants.

Top 20 counts of zip5 in log scale:Histogram

Description automatically generated with medium confidence

Field 8:

Name: dob

Description: Date of Birth of applicants.

Applicants’ Date of birth in chronological order, and in log scale:

**Chart

Description automatically generated**

Field 9:

Name: homephone

Description: Home phone number of applicants. 999999999 is counted as frivolous value

Top 20 counts of Homephone in log scale:

**Shape, rectangle

Description automatically generated**

Field 10:

Name: fraud label

Description: Whether the application was fraud or not.

‘1’ Indicates the application was fraudulent.

‘0’ indicates the application was non-fraudulent

Count of fraud label in log scale:

Shape

Description automatically generated with low confidence