USC Marshall
School of Business

Data Quality Report

Product Application Data

DSO 562 Fraud Analytics
Chutong Yan
February 15, 2020

Table of Contents

1. High-level File Description

The *application data* file stores personal information submitted by product applicants during year 2016 (2016-01-01 to 2016-12-31). The dataset contains 1,000,000 records and 10 fields (including field *record*, a unique key to identify each record). *Application data.csv* is not a real dataset. Rather, it is manually generated by one of Professor Coggeshall's ex-colleagues.

2. Summary Statistics Table

The dataset contains 9 categorial fields (excluding *record*). The summary statistics of these fields are provided in Table 1.

Table 1 Summary Table of Categorical Fields

Name	Type	# Not null	% Populated	# Unique	# zero	Most Common Field Value
date	Date/time	1,000,000	100.00%	365	0	2016-08-16
ssn	Categorical	1,000,000	100.00%	835,819	0	999999999
firstname	Categorical	1,000,000	100.00%	78,136	0	EAMSTRMT
lastname	Categorical	1,000,000	100.00%	177,001	0	ERJSAXA
address	Categorical	1,000,000	100.00%	828,774	0	123 MAIN ST
zip5	Categorical	1,000,000	100.00%	26,370	0	68138
dob	Categorical	1,000,000	100.00%	42,673	0	19070626
homephone	Categorical	1,000,000	100.00%	28,244	0	9999999999
fraud_label	Categorical	1,000,000	100.00%	2	985,607	0

* Field *record* is not specified in the table above because summary statistics of unique record key is not meaningful.

3. Explorations of Fields

3.1 record

record field is the unique integer label for each record, ranging from 1 to 1,000,000.

3.2 date

date field represents the dates of product applications. The table below shows the frequency of the top 15 values of *date* field.

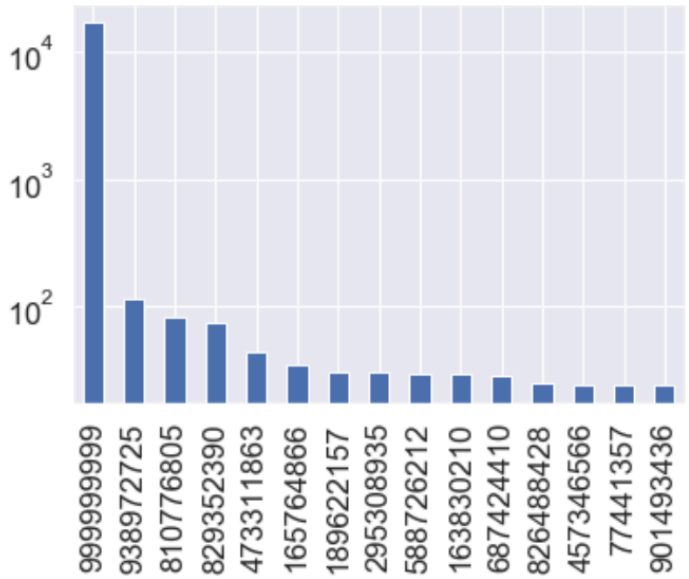
Table 2 Frequency of Top 15 values of *date* field

date	Frequency
2016-08-16	2,877
2016-03-04	2,861
2016-07-18	2,849
2016-04-17	2,848
2016-01-01	2,840
2016-12-28	2,832
2016-09-03	2,832
2016-08-08	2,832
2016-08-27	2,831
2016-06-09	2,831
2016-03-07	2,831
2016-10-06	2,831
2016-08-04	2,828
2016-03-13	2,826
2016-01-16	2,819

3.3 ssn

ssn field represents the social security numbers submitted by product applicants. The bar plot below shows the frequency of the top 15 values of *ssn* field.

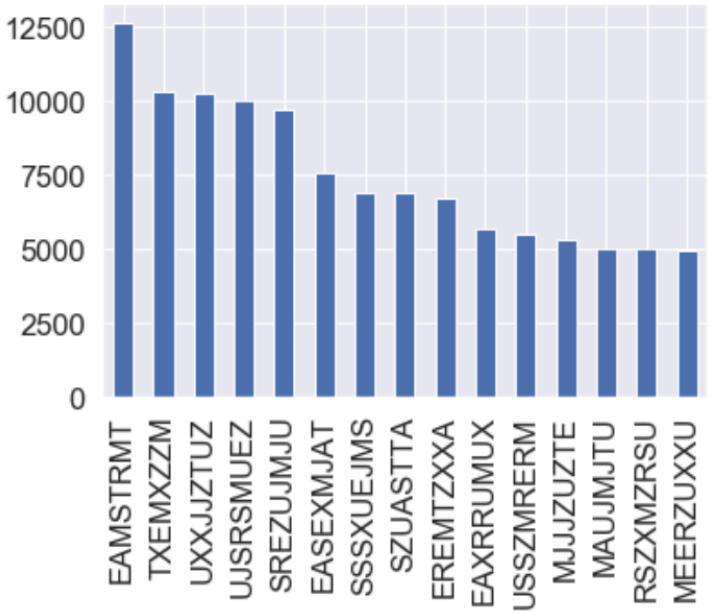
Figure 1 Frequency of Top 15 values of field *ssn*



3.4 firstname

firstname field represents the first names submitted by the applicants. The bar plot below shows the frequency of the top 15 values of *firstname* field.

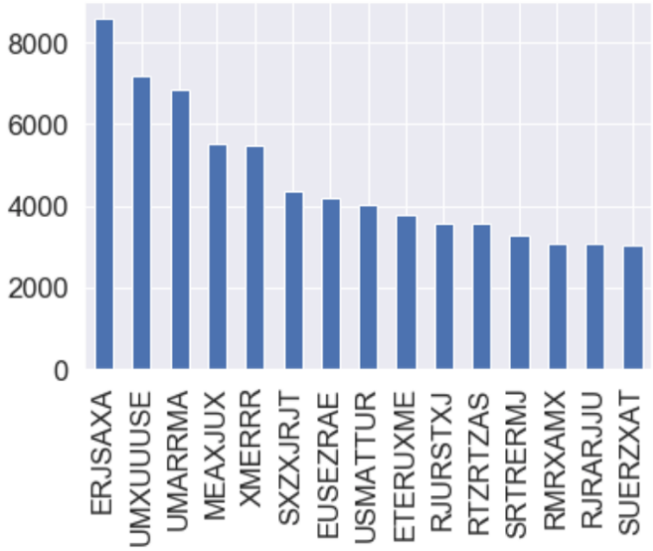
Figure 2 Frequency of Top 15 values of field *firstname*



3.5 lastname

lastname field represents the last names submitted by the applicants. The bar plot below shows the frequency of the top 15 values of *lastname* field.

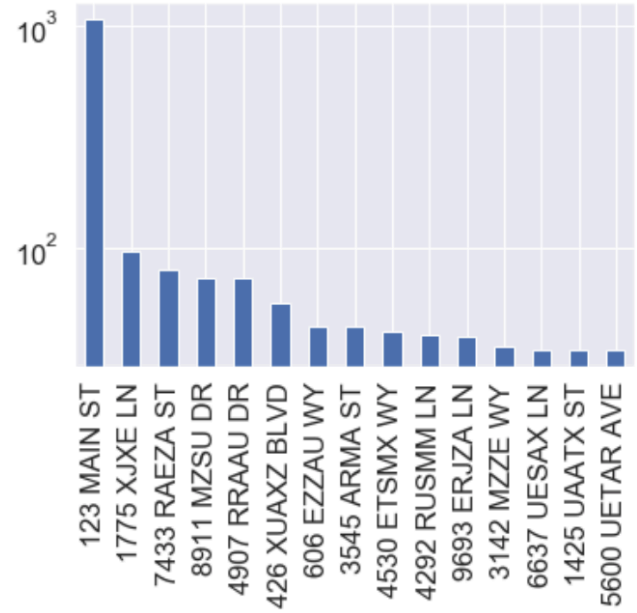
Figure 3 Frequency of Top 15 values of field *firstname*



3.6 address

address field represents the addresses submitted by the applicants. The bar plot below shows the frequency of the top 15 values of *address* field.

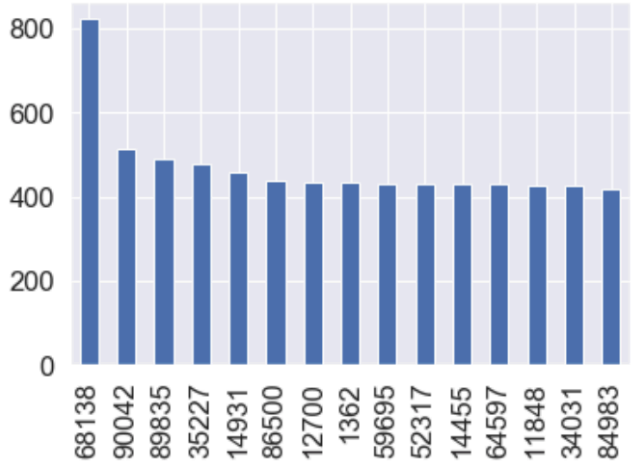
Figure 4 Frequency of Top 15 values of field *address*



3.7 zip5

zip5 field represents the zip codes submitted by the applicants. The bar plot below shows the frequency of the top 15 values of *zip5* field.

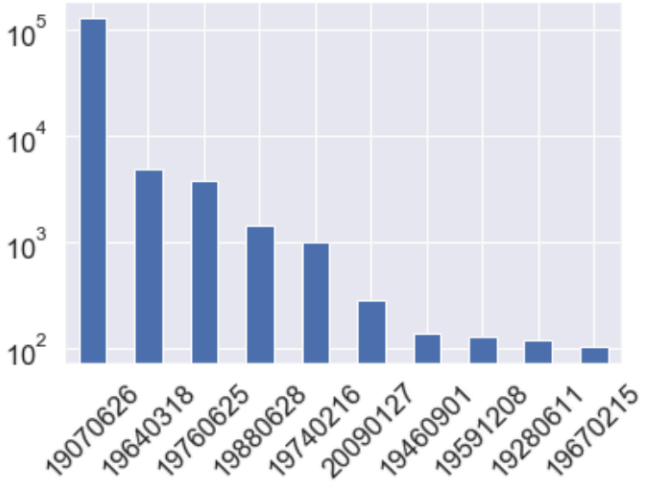
Figure 5 Frequency of Top 15 values of field *zip5*



3.8 dob

dob field represents the dates of birth submitted by the applicants. The bar plot below shows the frequency of the top 10 values of *dob* field.

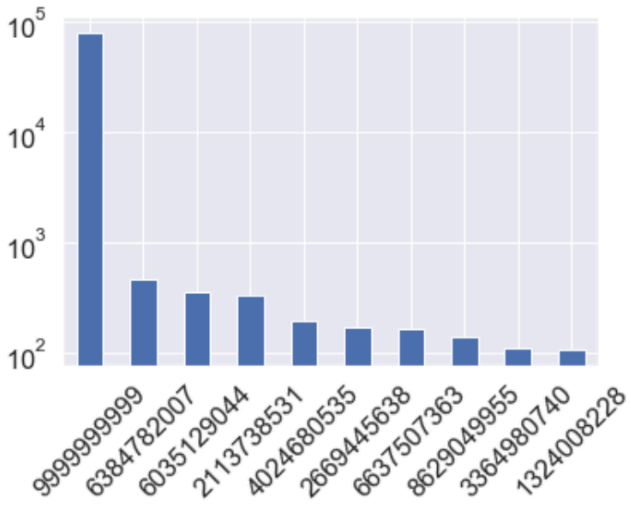
Figure 6 Frequency of Top 10 values of field *dob*



3.9 homephone

homephone field represents the home phone number submitted by the applicants. The bar plot below shows the frequency of the top 10 values of *homephone* field.

Figure 7 Frequency of Top 10 values of field *homephone*



3.10 fraud_label

fraud_label field represents whether or not an application is identified as fraud, where 1 = fraud and 0 = not fraud. The table below shows the count of values of *fraud_label* field.

Table 3 Count of values of field *fraud_label*

fraud_label	Count
0	985,607
1	14,393