DSO 562: Fraud Analytics

Data Quality Report on Applications Data

Yurong Jiang

6879765856

High-Level Description of Data

The dataset contains application information of the New York City. The data is collected to evaluate the possibility of identity fraud. The data has 1000000 rows and 10 columns. All fields are categorical.

Summary Table of All fields

Field	Type	Non-null	Unique Values	Most Common Field Value	Occurrence
record	Categorical	1,000,000	1000000	N/A	NA
date	Categorical	1,000,000	365	20160816	2,877
ssn	Categorical	1,000,000	835819	999999999	16,935
firstname	Categorical	1,000,000	78136	EAMSTRMT	12,658
lastname	Categorical	1,000,000	177001	ERJSAXA	8,580
address	Categorical	1,000,000	828774	123 MAIN ST	1,079
zip5	Categorical	1,000,000	26370	68138	823
dob	Categorical	1,000,000	42673	19070626	126,568
homephone	Categorical	1,000,000	28244	9999999999	78,512
fraud_label	Categorical	1,000,000	2	0	985,607

Short Description and Picture of Each Field

Field 1

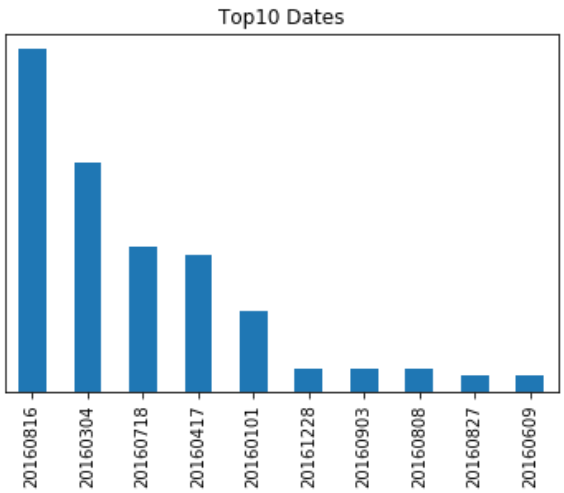
Field Name: record (int64)
Description: record is a categorical variable. It is a unique value to identify each record. Each value in record is unique.

Field 2

Field Name: date
Description:
date is a categorical variable representing the date of each transaction.

Unique Value:
date has 365 unique values. No missing value exists. The distribution is shown below, top 10 categories are listed below:

Category	Count
20160816	2877
20160304	2861
20160718	2849
20160417	2848
20160101	2840
20161228	2832
20160903	2832
20160808	2832
20160827	2831
20160609	2831



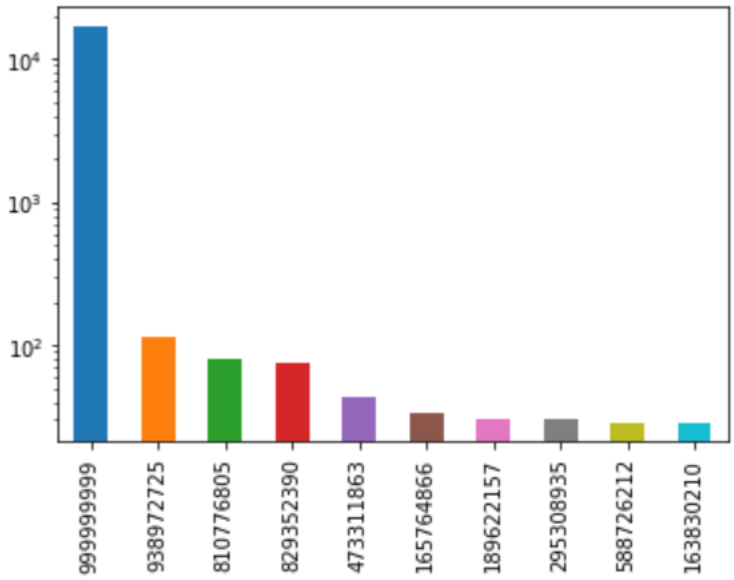
Field 3

Field Name: ssn (categorical, dtype: int64)

Description:

ssn is a categorical variable representing the social security number of the applicant.
ssn has 835819 unique values. No missing value exists. The distribution is shown below, top 10 categories are listed below. The following bar chart shows the log count and top 10 ssn.

Category	Count
999999999	16935
938972725	114
810776805	81
829352390	74
473311863	44
165764866	34
189622157	30
295308935	30
588726212	29
163830210	29



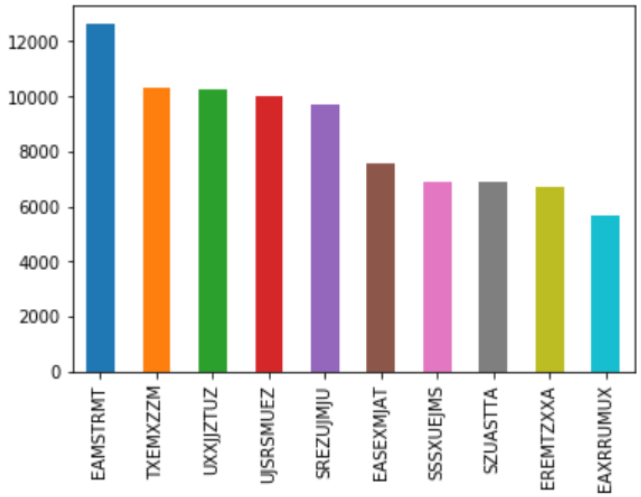
Field 4

Field Name: firstname (categorical, dtype: object)

Description:

firstname is a categorical variable representing the first name of the applicants. firstname has 78136 unique values. No missing value exists. The distribution is shown below, top 10 categories are listed below. The following bar chart shows the count and top 10 firstname.

Category	Count
EAMSTRMT	12658
TXEMXZZM	10297
UXXJJZTUZ	10235
UJSRSMUEZ	9994
SREZUJMJU	9688
EASEXMJAT	7576
SSSXUEJMS	6923
SZUASTTA	6878
EREMTZXXA	6717
EAXRRUMUX	5686

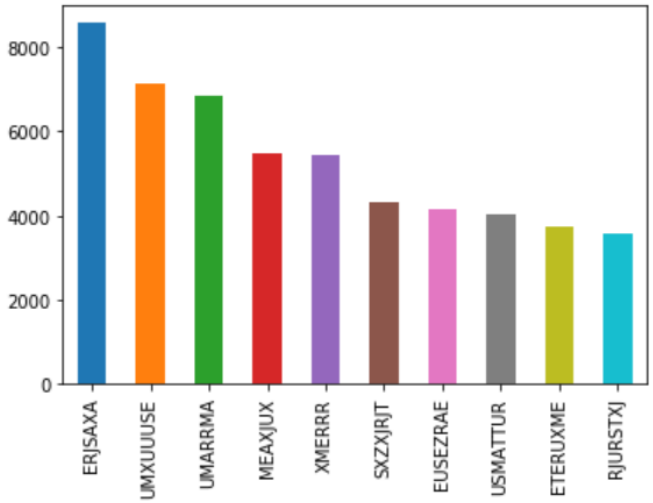


Field 5

Field Name: lastname (categorical, dtype: object)

Description:
lastname is a categorical variable describing the last name of the applicants. lastname has 177001 unique values. No missing value exists. The distribution is shown below, top 10 categories are listed below. The following bar chart shows the count and top 10 lastname.

Category	Count
ERJSAXA	8580
UMXUUUSE	7156
UMARRMA	6832
MEAXJUX	5492
XMERRR	5451
SXZXJRJT	4340
EUSEZRAE	4173
USMATTUR	4036
ETERUXME	3762
RJURSTXJ	3575

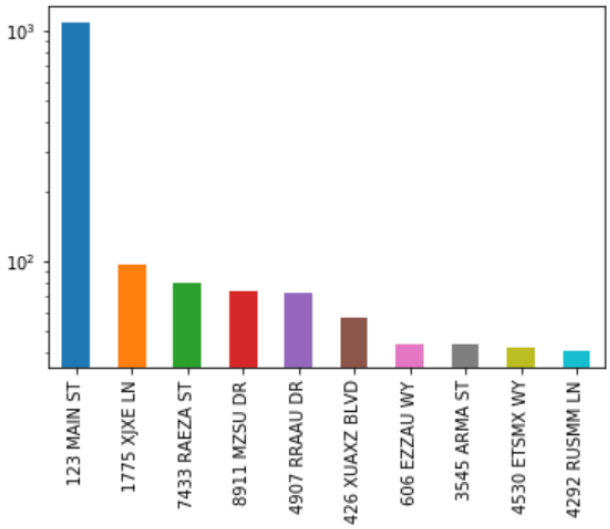


Field 6

Field Name: address (categorical, dtype: object)

Description:
address is a categorical variable describing the address of the applicant. address has 828774 unique values. There are no missing values. Below are top 10 categories in descending order. The following bar chart shows the log count and top 10 addresses.

Category	Count
123 MAIN ST	1079
1775 XJXE LN	97
7433 RAEZA ST	80
8911 MZSU DR	74
4907 RRAAU DR	73
426 XUAXZ BLVD	57
606 EZZAU WY	44
3545 ARMA ST	44
4530 ETSMX WY	42
4292 RUSMM LN	41

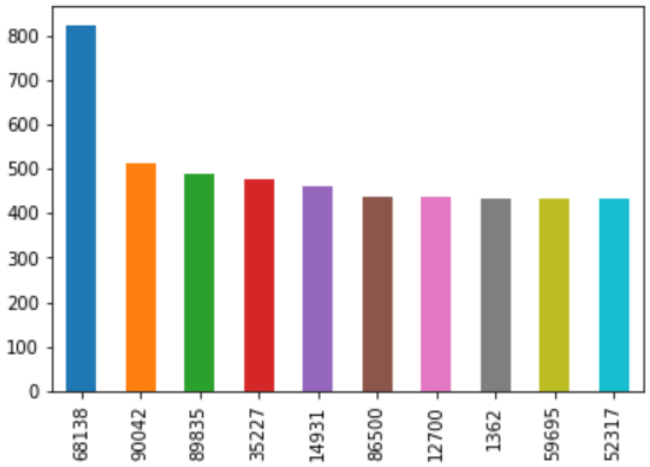


Field 7

Field Name: zip5 (categorical, dtype: int64)

Description:
zip5 is a categorical variable describing the 5-digit zip code of the applicants. zip5 has 26370 unique values. There are no missing values. Below are top 10 categories. The following bar chart shows the count of top 10 zip codes.

Category	Count
68138	823
90042	514
89835	489
35227	478
14931	459
86500	438
12700	436
1362	434
59695	432
52317	432



Field 8

Field Name: dob (categorical, dtype: int64)
Description:
dob is a categorical variable describing the date of birth of the applicants. dob has 42673 unique values. There are 0 missing values. Below are top 10 categories in descending order.

Category	Count
19070626	126568
19640318	4818
19760625	3723
19880628	1404
19740216	980
20090127	280
19460901	135
19591208	126
19280611	120
19670215	102

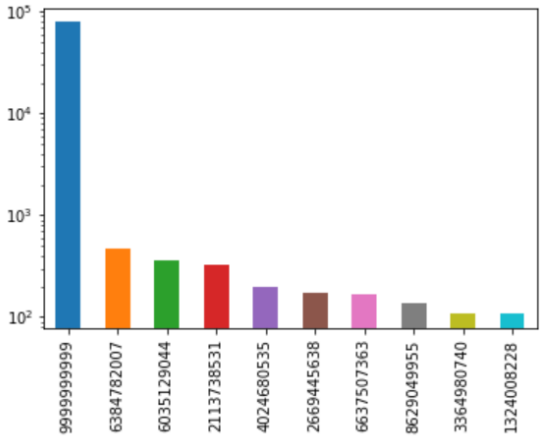
Field 9

Field Name: homephone

Description:

homephone is a categorical variable describing the home phone number of the applicants. homephone has 28244 unique values. Below are top 10 categories.

Category	Count
9999999999	78512
6384782007	466
6035129044	360
2113738531	331
4024680535	198
2669445638	172
6637507363	169
8629049955	139
3364980740	110
1324008228	108



Field 10

Field Name: fraud_label

Description:

fraud_label is a categorical variable denoting whether an application is fraud.

It has 2 unique values, 0 and 1. There are 985607 zeros and 14393 ones.

Category	Count
0	985607
1	14393