Data Quality Report, Project 2

Chengjun Liu, 7116099552

# Data Description

Dataset Name: Product Application Identity Data

Dataset Source: Simulation based on real-life Application Identity Data

Release Date: N/A

Number of Columns: 10

Number of Records: 1000000

Summary

Overview:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Data Type | Non-null Records | Description |
| record | Integer | 1000000 | Number of Records |
| date | Integer(Datetime) | 1000000 | Application date |
| ssn | Integer | 1000000 | Social security number |
| firstname | Text string | 1000000 | First name of applicant |
| lastname | Text string | 1000000 | Last name of applicant |
| address | Text string | 1000000 | Applicant address |
| zip5 | Integer | 1000000 | Zip code in 5 digits |
| dob | Integer(Datetime) | 1000000 | Date of Birth |
| homephone | Integer | 1000000 | Home Phone Number |
| fraud\_label | Integer(Binary) | 1000000 | Whether application is fraud(1) or not(0) |

Because all variables in integers are regarded and analyzed as categorical variables, we include integer variables in the categorical variable table:

Summary of Integer Variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Non-null Record | Non-null Percentage | Min | Max |
| record | 1000000 | 100% | 1 | 1000000 |
| date | 1000000 | 100% | 20160101 | 20161231 |
| ssn | 1000000 | 100% | 36 | 999999999 |
| zip5 | 1000000 | 100% | 2 | 99999 |
| dob | 1000000 | 100% | 19000101 | 20161031 |
| homephone | 1000000 | 100% | 593799 | 9999999999 |
| fraud\_label | 1000000 | 100% | 0 | 1 |

Summary of Categorical Variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Non-null Record | Non-null Percentage | Number of Unique Values | Most frequent category |
| record | 1000000 | 100% | 1000000 | None |
| date | 1000000 | 100% | 365 | 20160816 |
| ssn | 1000000 | 100% | 835819 | 999999999 |
| firstname | 1000000 | 100% | 78136 | EAMSTRMT |
| lastname | 1000000 | 100% | 177001 | ERJSAXA |
| address | 1000000 | 100% | 828774 | 123 MAIN ST |
| zip5 | 1000000 | 100% | 26730 | 68138 |
| dob | 1000000 | 100% | 42673 | 19070626 |
| homephone | 1000000 | 100% | 28244 | 9999999999 |
| fraud\_label | 1000000 | 100% | 2 | 985607 |

Data Field Exploration

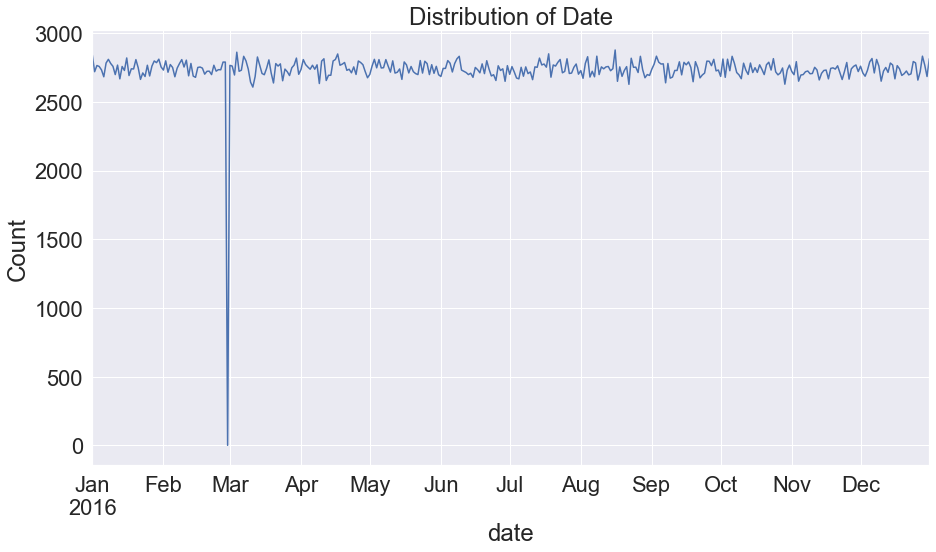
Variable 1

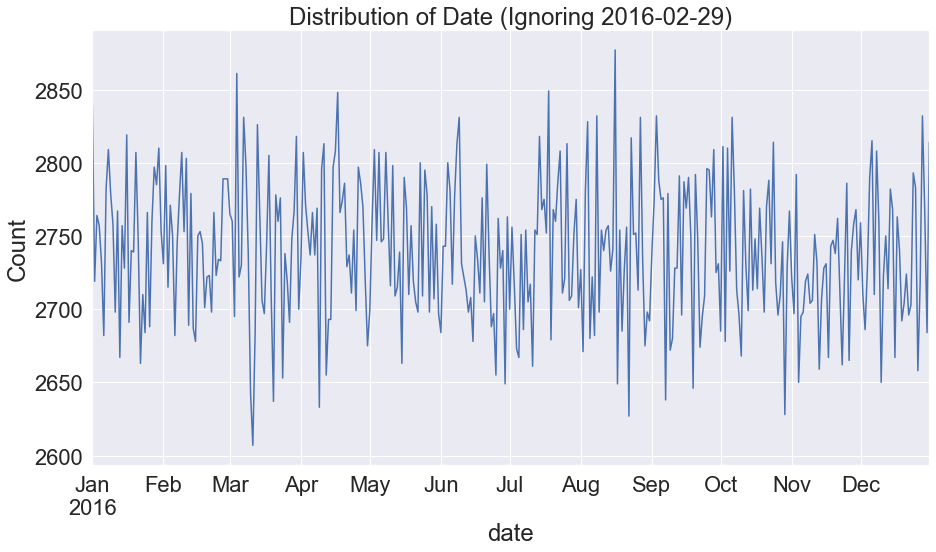
Name: date

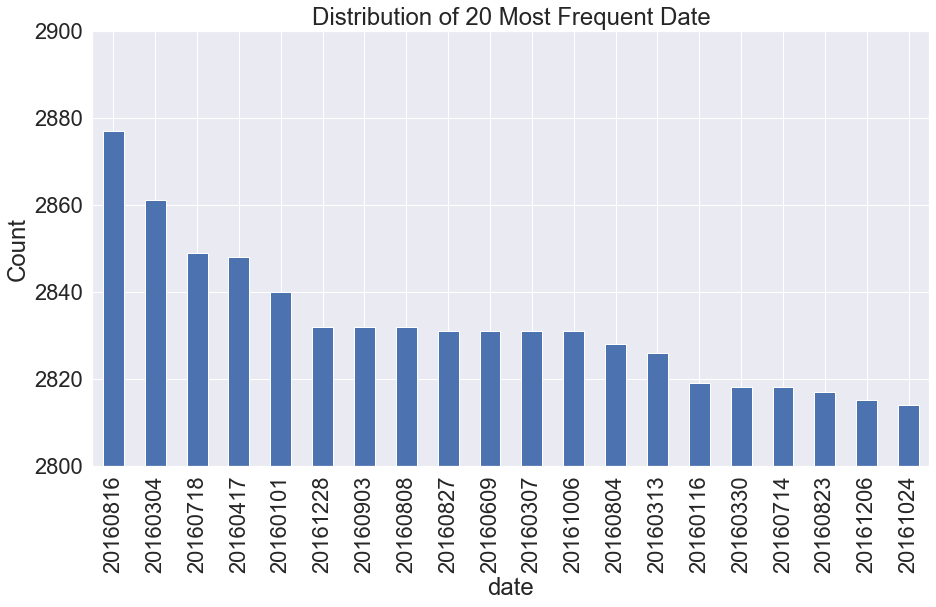
Description: The date that application is submitted.

Because there are no applications on Feb 29, 2016, we ignore Feb 29, 2016 in the second plot to have a better view of the distribution.

To show the distribution of the whole time and the most frequent dates, corresponding plots are here.



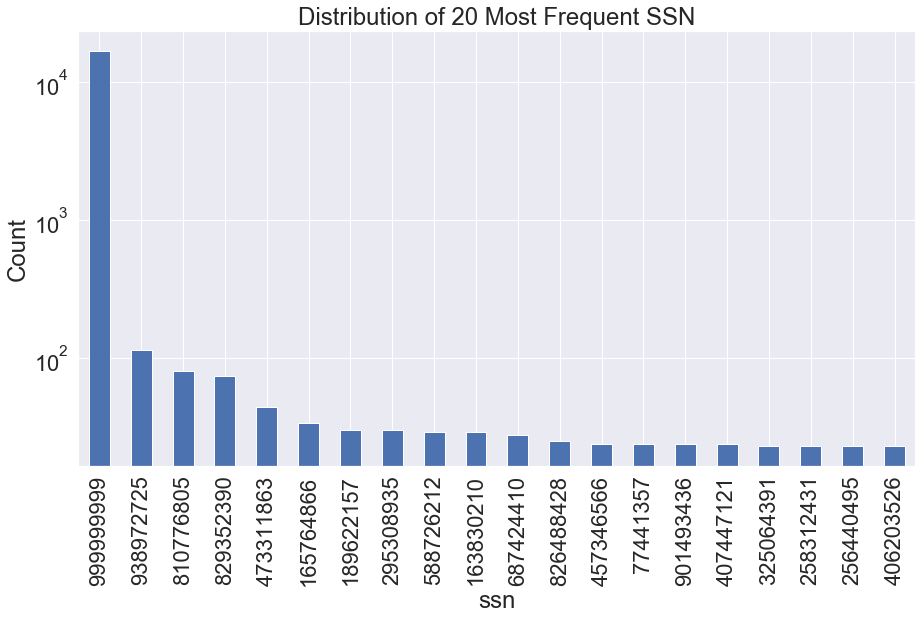




Variable 2

Name: ssn

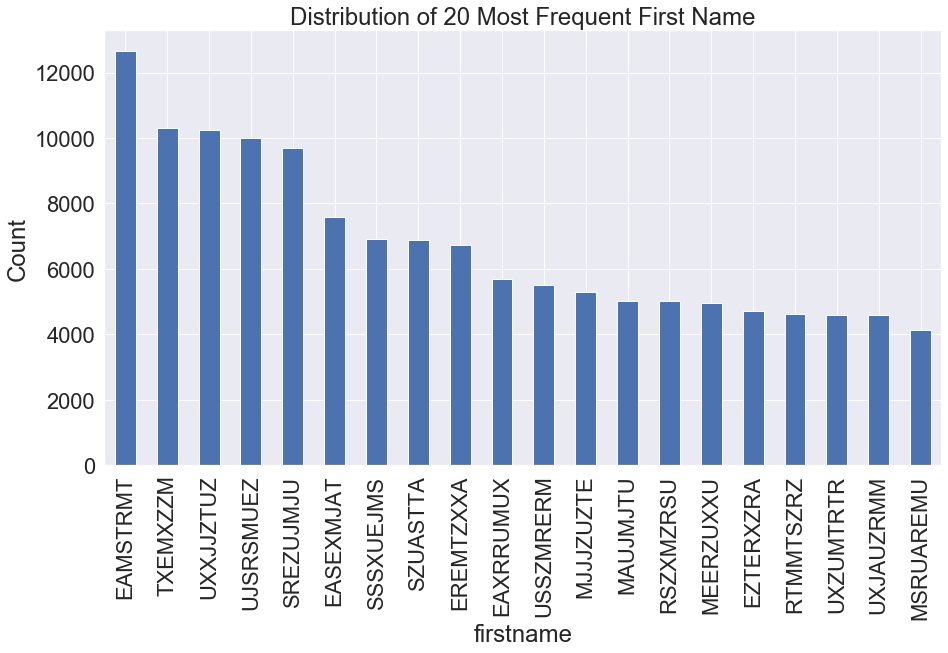
Description: Social Security Number of the applicant.



Variable 3:

Name: firstname

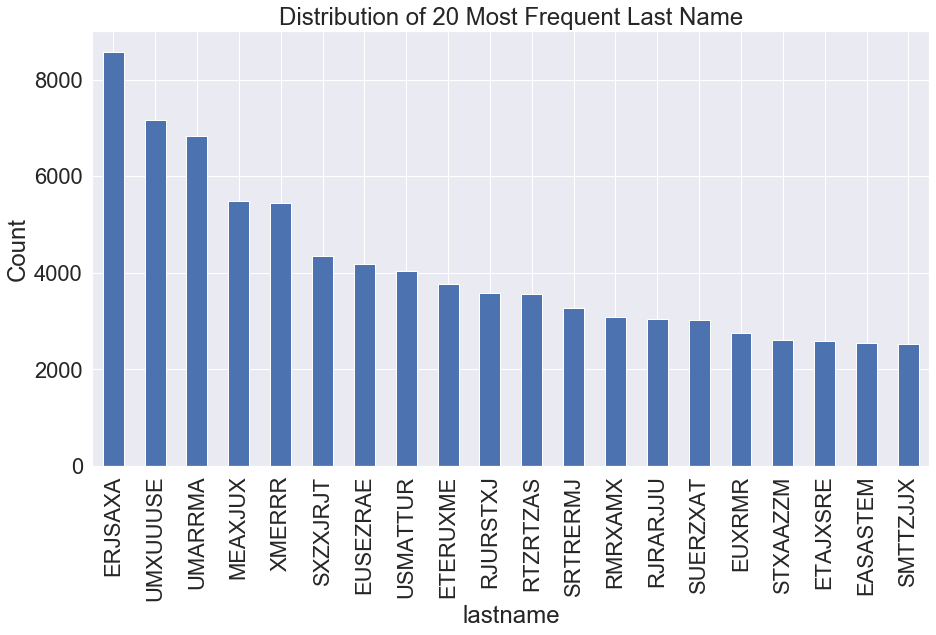
Description: First name of the applicant.



Variable 4

Name: lastname

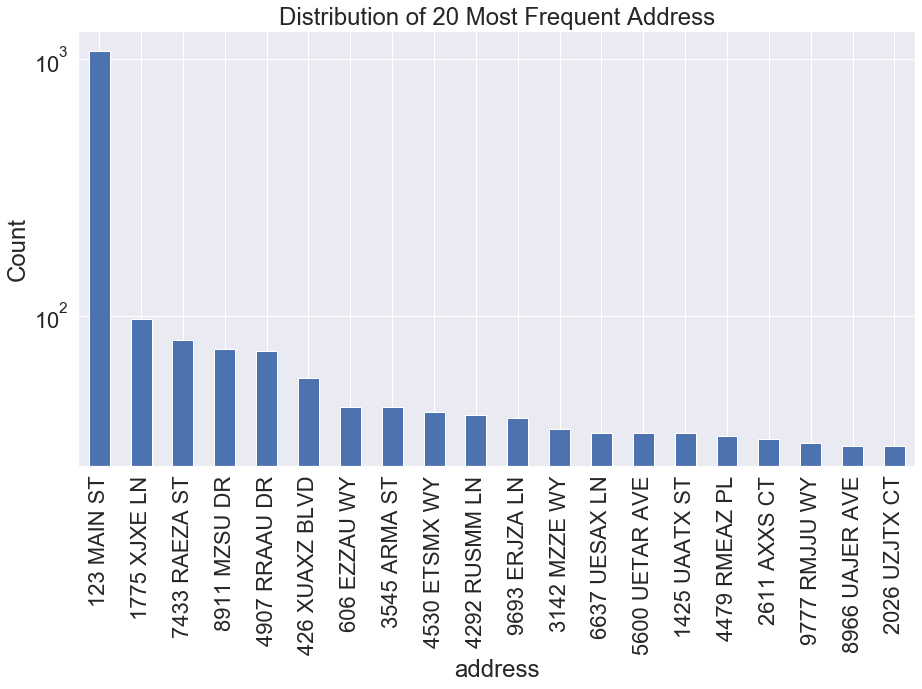
Description: Last name of the applicant.



Variable 5

Name: address

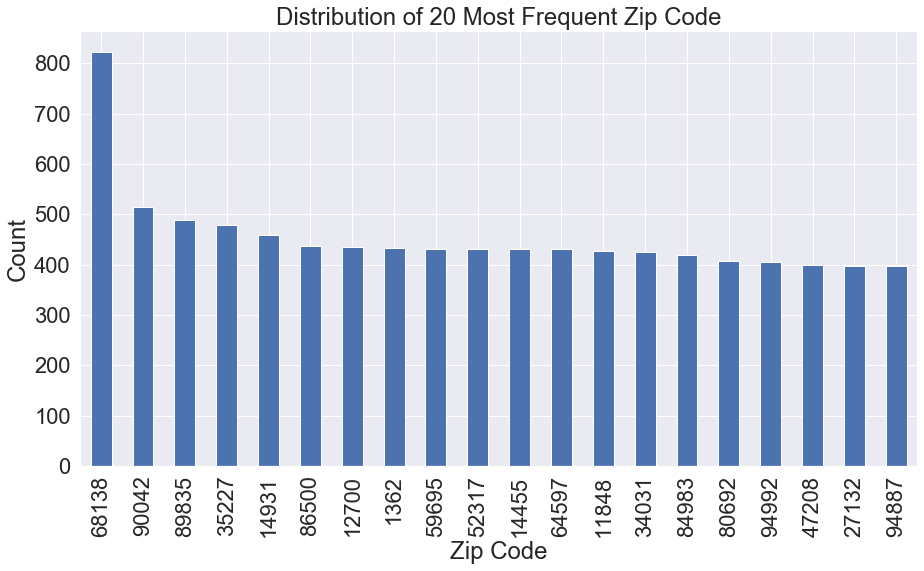
Description: Home address of the applicant.



Variable 6

Name: zip5

Description: Zip code of the home address of the applicant, in the format of 5 digits.

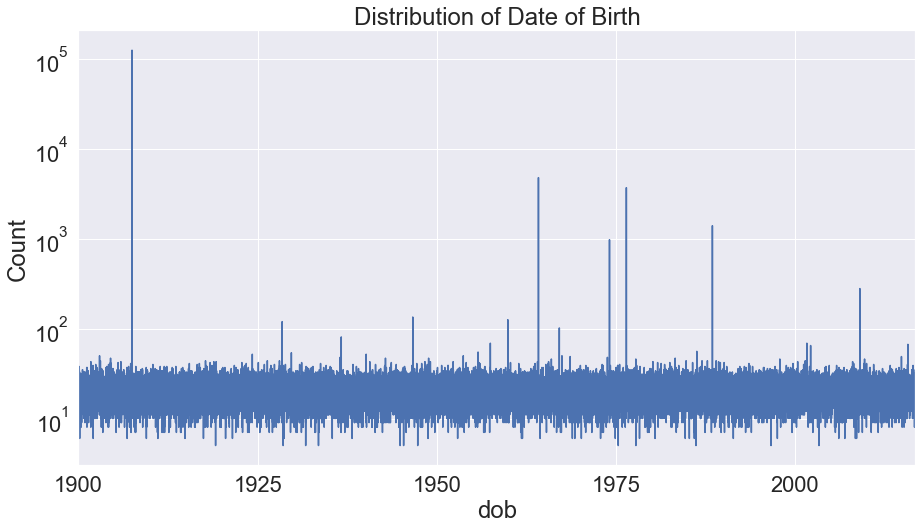


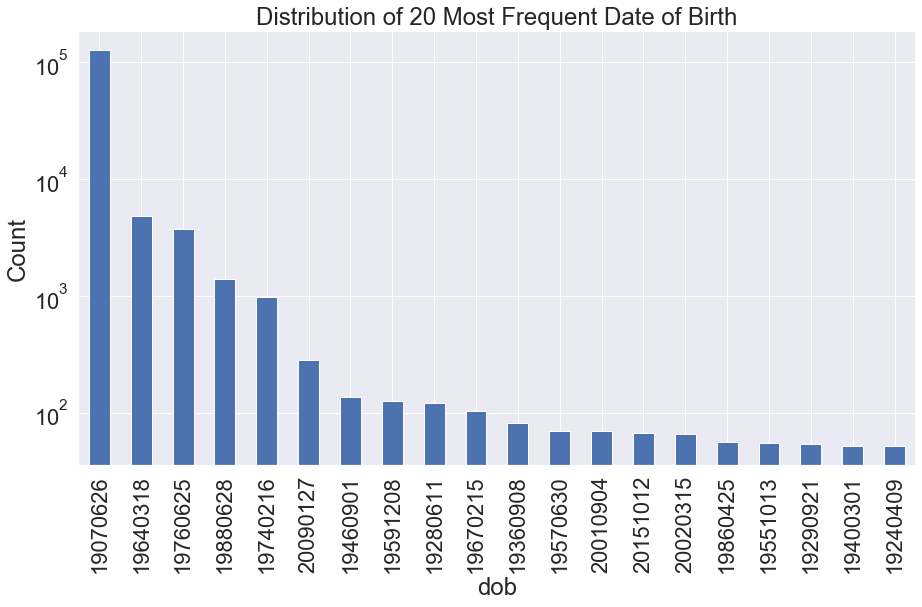
Variable 7

Name: dob

Description: Date of Birth of the applicant.

To show the distribution of the whole time and the most frequent dates, corresponding plots are here.

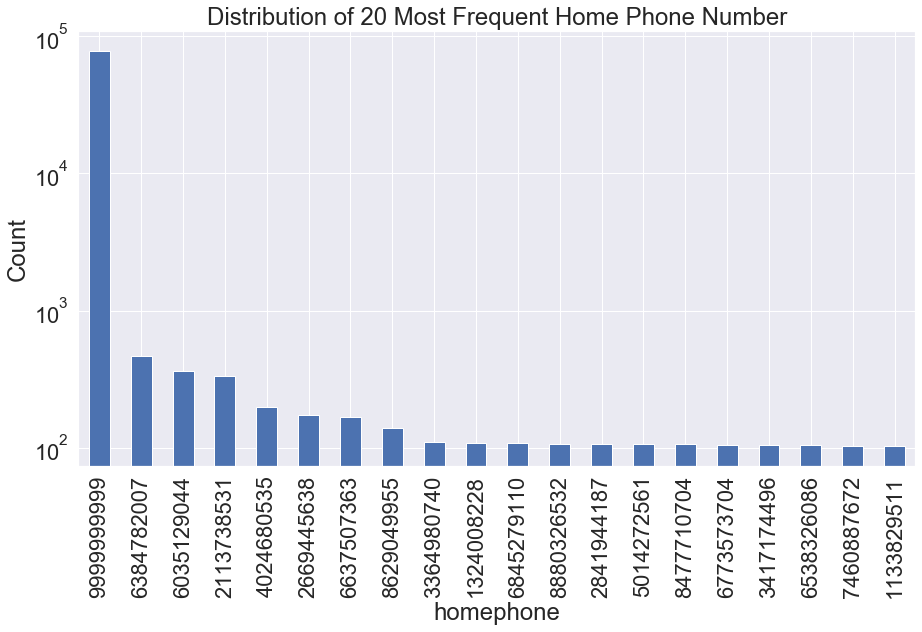




Variable 8

Name: homephone

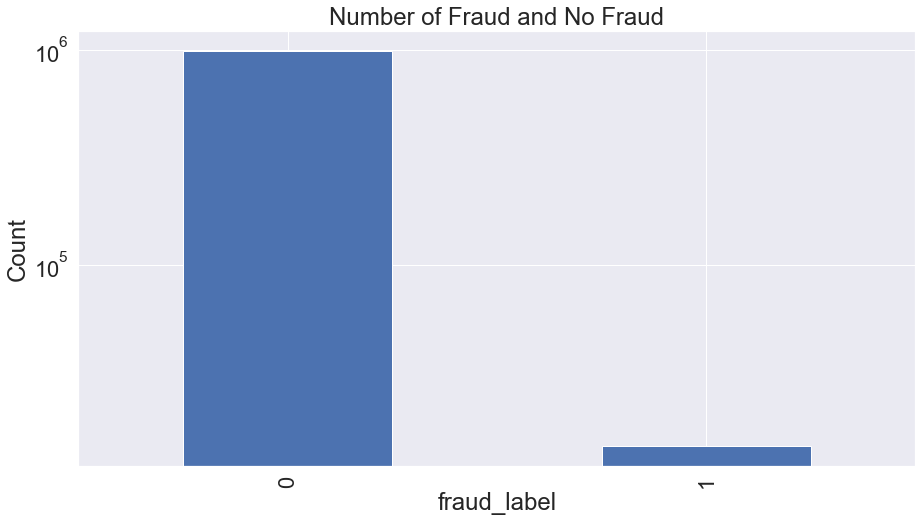
Description: Home phone number of the applicant.



Variable 9

Name: fraud\_label

Description: Whether the application is a fraud (0 means no, 1 means yes).



In addition, to explore relationship of fraud and date, we plot the distributions of fraud and no fraud and highlight dates with most fraud and no fraud.

