****

**University of Southern California**

Marshall School of Business

**Data Quality Report**

*Homework 5*

**Student: Wei Tang**

**Course:** DSO562 Fraud Analytics

**Instructor:** Dr. Stephen Coggeshall

**Date:** 03/02/2018

**BASIC INFORMATION**

|  |  |
| --- | --- |
| **Dataset** | Personally Identifiable Information (PII) |
| **Records** | 94,866 |
| **Columns** | 10 categorical variables |
| **Time period** | 01/01/2016 – 12/31/2016 |
| **Resource** | Simulated by Professor Stephen Coggeshall |

**SUMMARY TABLE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **Variables** | **# of Unique values** | **Count** | **Percentage populated** |
| Categorical | record | 94,899 | 94,866 | 100% |
| date | 365 | 94,866 | 100% |
| ssn | 86,771 | 94,866 | 100% |
| firstname | 14,626 | 94,866 | 100% |
| lastname | 31,513 | 94,866 | 100% |
| address | 88,167 | 94,866 | 100% |
| zip5 | 15,855 | 94,866 | 100% |
| dob | 30,599 | 94,866 | 100% |
| homephone | 20,762 | 94,866 | 100% |
| fraud | 2 | 94,866 | 100% |

**DATA ANALYSIS**

**record** – order number of each record

# of values: 94,866

# of unique values: 94,899

Distribution: Start from 1, increasing by 1 each time, with no repetition number; Increase to 94,866 in the end

**date** – date of generating records

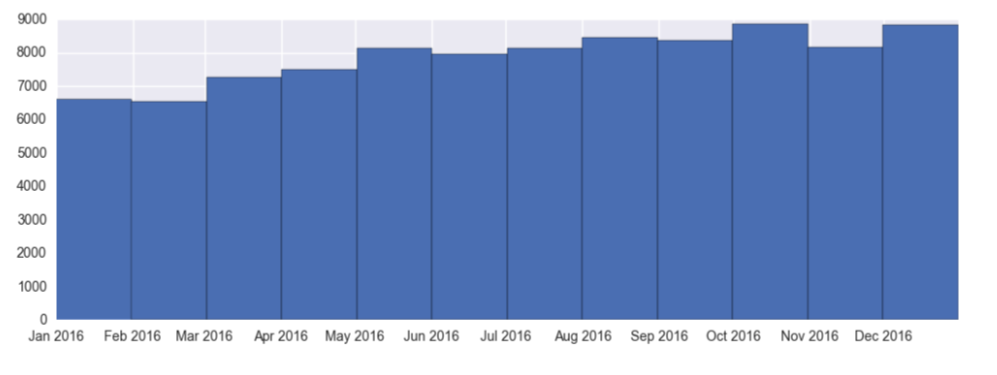
# of values: 94,866

# of unique values: 356

Top15 values:

|  |  |  |
| --- | --- | --- |
| **rank** | **date** | **counts** |
| 1 | 06/09/16 | 329 |
| 2 | 12/29/16 | 328 |
| 3 | 11/19/16 | 325 |
| 4 | 09/18/16 | 324 |
| 5 | 10/18/16 | 324 |
| 6 | 10/02/16 | 320 |
| 7 | 12/10/16 | 320 |
| 8 | 12/08/16 | 320 |
| 9 | 10/07/16 | 320 |
| 10 | 12/30/16 | 319 |
| 11 | 08/27/16 | 315 |
| 12 | 12/31/16 | 307 |
| 13 | 09/25/16 | 306 |
| 14 | 10/21/16 | 305 |
| 15 | 09/15/16 | 305 |

Monthly distribution:



**ssn** – social security number of each record

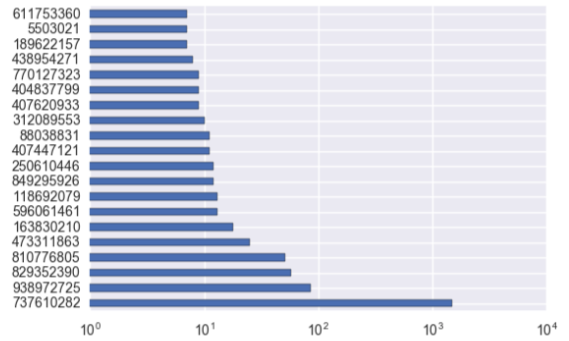
# of values: 94,866

# of unique values: 86,771

Top15 values:

|  |  |  |
| --- | --- | --- |
| **rank** | **ssn** | **counts** |
| 1 | 737610282 | 1478 |
| 2 | 938972725 | 85 |
| 3 | 829352390 | 57 |
| 4 | 810776805 | 51 |
| 5 | 473311863 | 25 |
| 6 | 163830210 | 18 |
| 7 | 596061461 | 13 |
| 8 | 118692079 | 13 |
| 9 | 849295926 | 12 |
| 10 | 250610446 | 12 |
| 11 | 407447121 | 11 |
| 12 | 88038831 | 11 |
| 13 | 312089553 | 10 |
| 14 | 407620933 | 9 |
| 15 | 404837799 | 9 |

Log scale histogram of TOP 15 values:



**firstname** – first name of each record

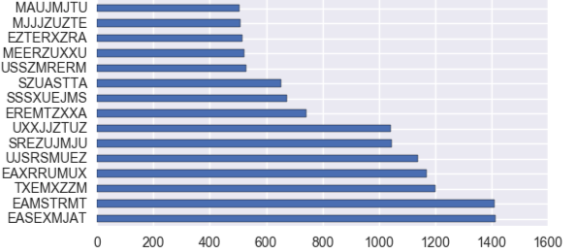
# of values: 94,866

# of unique values: 14,626

Top15 values:

|  |  |  |
| --- | --- | --- |
| **rank** | **firstname** | **counts** |
| 1 | EASEXMJAT | 1414 |
| 2 | EAMSTRMT | 1411 |
| 3 | TXEMXZZM | 1200 |
| 4 | EAXRRUMUX | 1170 |
| 5 | UJSRSMUEZ | 1138 |
| 6 | SREZUJMJU | 1044 |
| 7 | UXXJJZTUZ | 1042 |
| 8 | EREMTZXXA | 742 |
| 9 | SSSXUEJMS | 675 |
| 10 | SZUASTTA | 653 |
| 11 | USSZMRERM | 529 |
| 12 | MEERZUXXU | 523 |
| 13 | EZTERXZRA | 516 |
| 14 | MJJJZUZTE | 510 |
| 15 | MAUJMJTU | 504 |

Histogram of TOP 15 values:



**lastname** - last name of each record

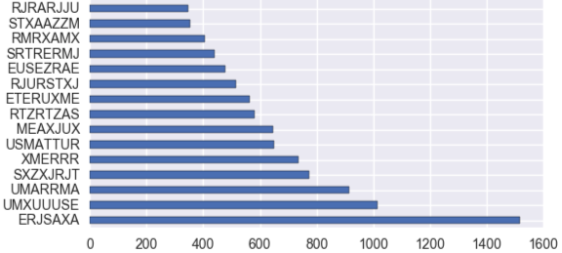
# of values: 94,866

# of unique values: 31,513

Top15 values:

|  |  |  |
| --- | --- | --- |
| **rank** | **lastname** | **counts** |
| 1 | ERJSAXA | 1515 |
| 2 | UMXUUUSE | 1013 |
| 3 | UMARRMA | 913 |
| 4 | SXZXJRJT | 775 |
| 5 | XMERRR | 737 |
| 6 | USMATTUR | 649 |
| 7 | MEAXJUX | 645 |
| 8 | RTZRTZAS | 582 |
| 9 | ETERUXME | 562 |
| 10 | RJURSTXJ | 515 |
| 11 | EUSEZRAE | 476 |
| 12 | SRTRERMJ | 438 |
| 13 | RMRXAMX | 405 |
| 14 | STXAAZZM | 352 |
| 15 | RJRARJJU | 348 |

Histogram of TOP 15 values:



**address** - address of each record

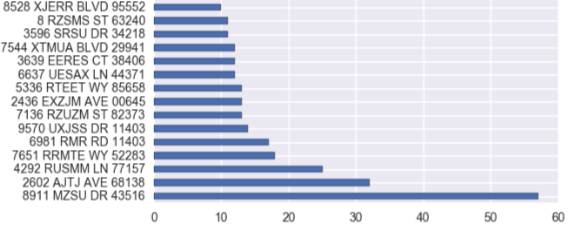
# of values: 94,866

# of unique values: 88,167

Top15 values:

|  |  |  |
| --- | --- | --- |
| **rank** | **address** | **counts** |
| 1 | 8911 MZSU DR 43516 | 57 |
| 2 | 2602 AJTJ AVE 68138 | 32 |
| 3 | 4292 RUSMM LN 77157 | 25 |
| 4 | 7651 RRMTE WY 52283 | 18 |
| 5 | 6981 RMR RD 11403 | 17 |
| 6 | 9570 UXJSS DR 11403 | 14 |
| 7 | 7136 RZUZM ST 82373 | 13 |
| 8 | 5336 RTEET WY 85658 | 13 |
| 9 | 2436 EXZJM AVE 00645 | 13 |
| 10 | 7544 XTMUA BLVD 29941 | 12 |
| 11 | 6637 UESAX LN 44371 | 12 |
| 12 | 3639 EERES CT 38406 | 12 |
| 13 | 3596 SRSU DR 34218 | 11 |
| 14 | 8 RZSMS ST 63240 | 11 |
| 15 | 8528 XJERR BLVD 95552 | 10 |

Histogram of TOP 15 values:



**zip5** – zip code of each record

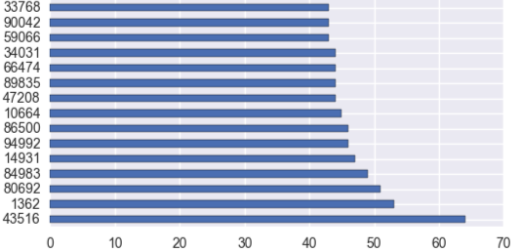
# of values: 94,866

# of unique values: 15,855

Top15 values:

|  |  |  |
| --- | --- | --- |
| **rank** | **zip5** | **counts** |
| 1 | 43516 | 64 |
| 2 | 1362 | 53 |
| 3 | 80692 | 51 |
| 4 | 84983 | 49 |
| 5 | 14931 | 47 |
| 6 | 94992 | 46 |
| 7 | 86500 | 46 |
| 8 | 10664 | 45 |
| 9 | 47208 | 44 |
| 10 | 89835 | 44 |
| 11 | 66474 | 44 |
| 12 | 34031 | 44 |
| 13 | 59066 | 43 |
| 14 | 90042 | 43 |
| 15 | 33768 | 43 |

Histogram of TOP 15 values:



**dob** – date of birth of each record

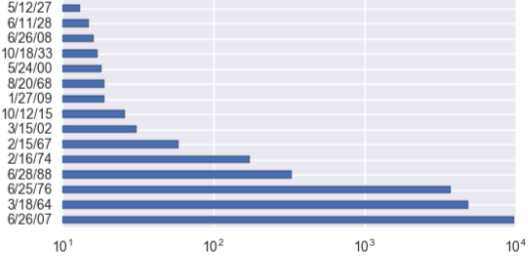
# of values: 94,866

# of unique values: 30,599

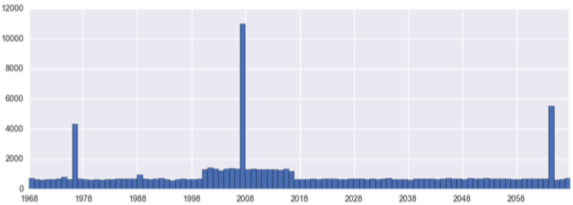
Top15 values:

|  |  |  |
| --- | --- | --- |
| **rank** | **dob** | **counts** |
| 1 | 6/26/07 | 9681 |
| 2 | 3/18/64 | 4808 |
| 3 | 6/25/76 | 3698 |
| 4 | 6/28/88 | 330 |
| 5 | 2/16/74 | 173 |
| 6 | 2/15/67 | 59 |
| 7 | 3/15/02 | 31 |
| 8 | 10/12/15 | 26 |
| 9 | 8/20/68 | 19 |
| 10 | 1/27/09 | 19 |
| 11 | 5/24/00 | 18 |
| 12 | 10/18/33 | 17 |
| 13 | 6/26/08 | 16 |
| 14 | 6/11/28 | 15 |
| 15 | 5/12/27 | 13 |

Histogram of TOP 15 values:



Yearly distribution:



**homephone** – home phone number of each record

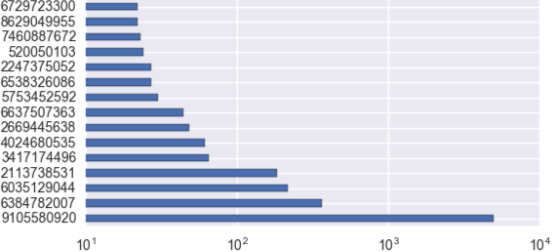
# of values: 94,866

# of unique values: 20,762

Top15 values:

|  |  |  |
| --- | --- | --- |
| **rank** | **homephone** | **counts** |
| 1 | 9105580920 | 4974 |
| 2 | 6384782007 | 364 |
| 3 | 6035129044 | 215 |
| 4 | 2113738531 | 184 |
| 5 | 3417174496 | 65 |
| 6 | 4024680535 | 61 |
| 7 | 2669445638 | 48 |
| 8 | 6637507363 | 44 |
| 9 | 5753452592 | 30 |
| 10 | 6538326086 | 27 |
| 11 | 2247375052 | 27 |
| 12 | 520050103 | 24 |
| 13 | 7460887672 | 23 |
| 14 | 8629049955 | 22 |
| 15 | 6729723300 | 22 |

Histogram of TOP 15 values:



**fraud** – judgement of each record, whether it is fraud or not

# of values: 94,866

# of unique values: 2

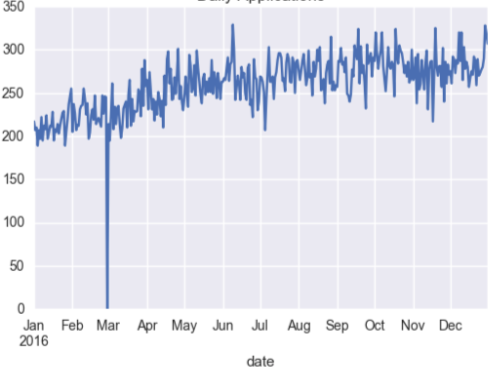
Values distribution:

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **fraud** | **counts** | **percentage** |
| 1 | 0 | 70702 | 78.7% |
| 2 | 1 | 20164 | 21.3% |

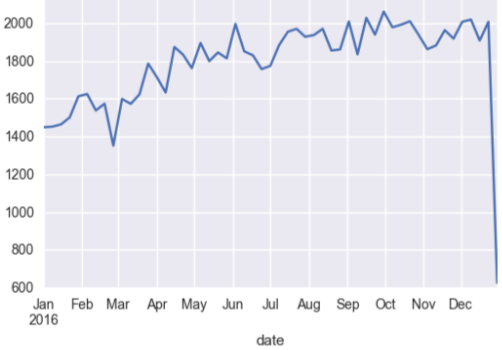
1 means fraud, so 21.3% of records are fraud.

**FURTHER ANALYSIS**

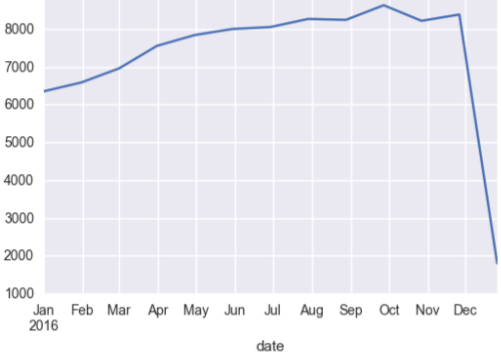
**# of transactions per day**



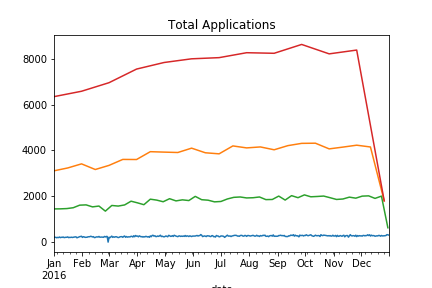
**# of transactions per week**



**# of transactions per month**



**# of transactions monthly versus semimonthly versus weekly versus daily**



• monthly

• semimonthly

• weekly

• daily