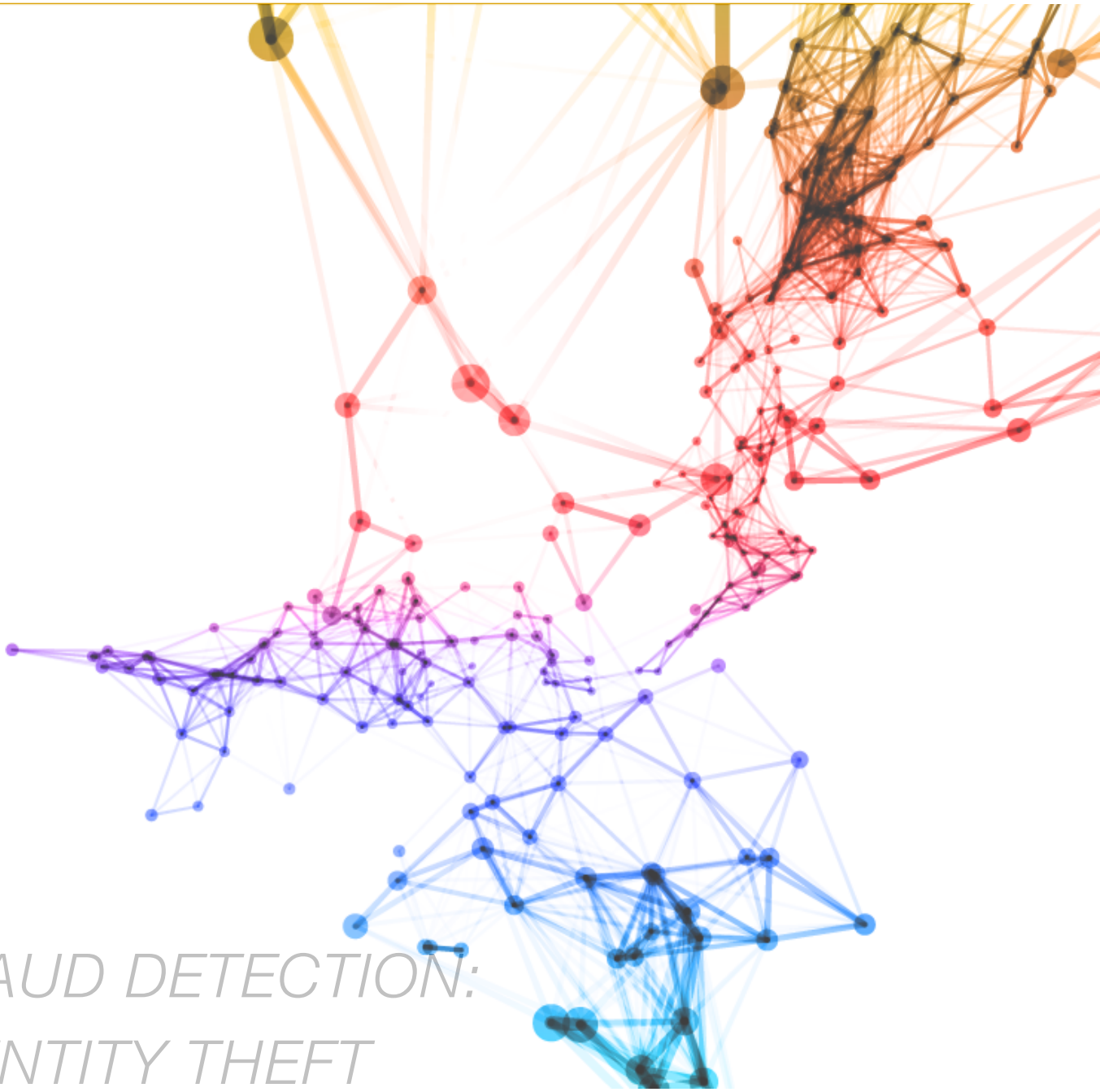


FRAUD DETECTION: IDENTITY THEFT

Presented by: Daniel Cheng, Eva Zhong, Hermosa Liu, Jimmy Chen, Judy Yan, Zack Chou
Directed by: Prof. Stephen Coggeshall
March 19, 2017

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Nowadays, Synthetic Identity Theft has been emerging as a major fraud activity over the past decade. Usually, fraudsters will use this fictitious identity to apply for credit card, open deposit accounts or obtain other important identity documents.The size of the synthetic identity theft business is estimated to be in the billions per year across North America.

Synthetic identity theft is fraud involving the use of a fictitious identity. Identity thieves create new identities using a combination of real and fabricated information, or sometimes entirely fictitious information. Typically, fraudsters will use a real Social Security Number (SSN) and pair it with a name not associated with that number, but there are also other ways. In some cases, an identity fraudster may create a completely fake identity with a phony SSN, name and address.

This report commissions to examine the 100,000 credit card application data, detect abnormality and potential fraud in the dataset. All data manipulation and analysis are conducted in R. Featured analysis methods include Principal Component Analysis (PCA), Heuristic Algorithm and Autoencoder. Major steps of analysis include:

1. Data manipulation and Creating new informative features
2. Dimensionality reduction through PCA process
3. Calculating fraud score using both heuristic algorithm and Autoencoder

Using both Autoencoder and Heuristic Algorithm, we selected top 10 abnormal records by each method with highest fraud scores, which could be classified as underlying synthetic identity frauds.

Further investigation into those suspicious records indicates that the abnormality is mainly due to frequent and repeated occurrence in the past. Our assumption of these latent fraud records are mainly:

1. Fraudsters tend to submit more applications than real applicants. It might because that they want to have a higher chance of being approved.
2. Although fraudsters may forge information such as name and SSN and change them time to time, but they also tend to keep inputting same real information for contact purpose, such as phone number and address that they have access to.

The report, however, may include following limitations:
1. Not using the full dataset since we can only calculate fraud scores from Jan 21st going forward
2. The overlapping rate of top records between 2 methods is not high enough

DATA DESCRIPTION

Overview of Credit Card Application

File name: applications 100k.csv
Data Provided by: Professor Stephen Coggeshall
Data Volume: 100,000 records
Fields: 9 fields (3 text; 2 dates; 4 categorical)
Time Frame: 01/01/2015 - 12/31/2015

Original Names with Description

Field Name	Type	Description
record #	categorical	Index of records
date	date	date of the application activity
ssn	categorical	social security number of the applicant
firstname	text	firstname of the applicant
lastname	text	lastname of the applicant
address	text	address of the applicant
zip5	categorical	zip code of the address of the applicant
dob	date	date of birth of the applicant
homephone	categorical	home phone number of the applicant

Application Date Distribution

From the Figure 1 below, we can see that credit card applications distribute quite evenly throughout year 2015.

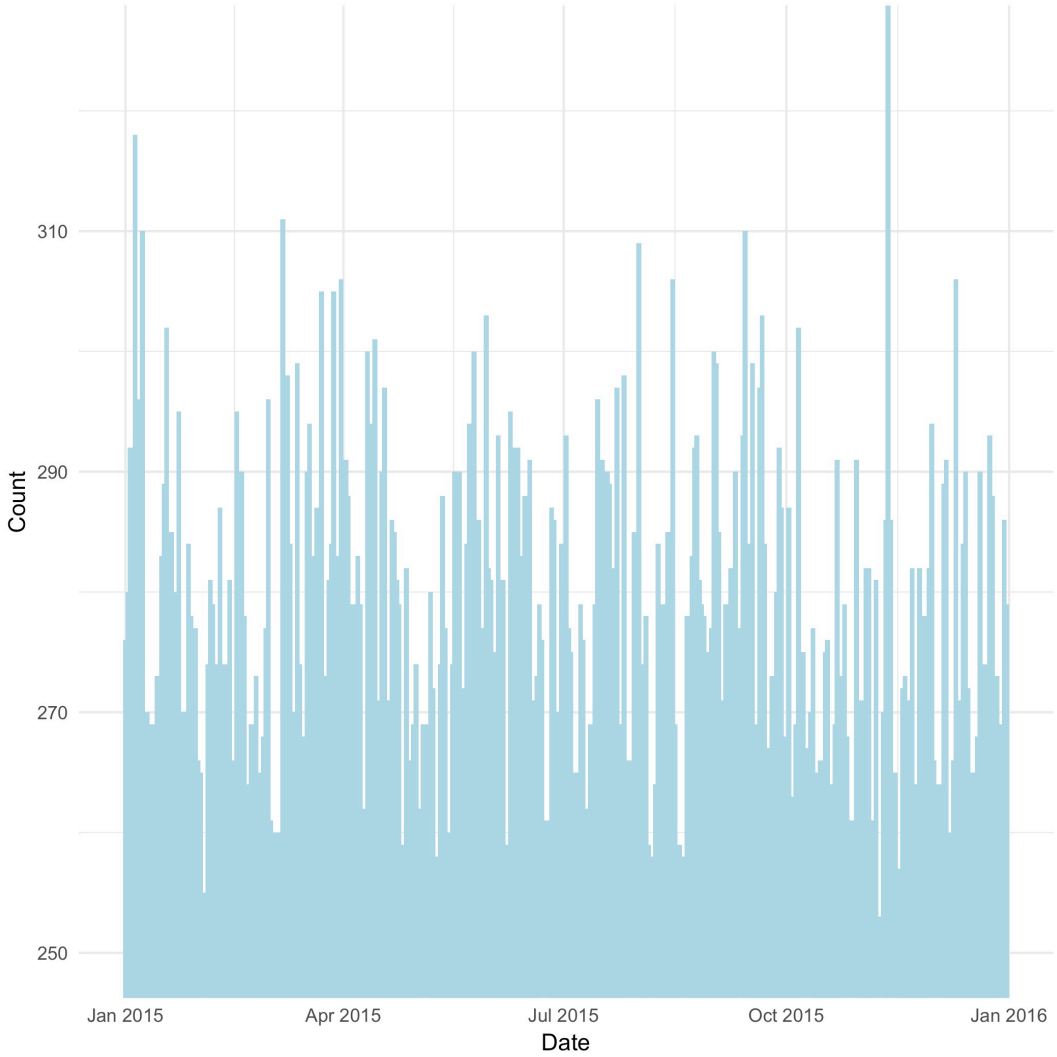


Figure 1. Application Date Distribution

VARIABLE CREATION

Overview

In order to detect frauds, we need further numeric information beyond the original dataset. With that in mind, variable creation is an imperative part of this project. The following describes in detail how we have created variables in addition to the original ones.

We created 73 new variables in total, and these newly-built variables can be divided to two types. On one hand, we want to know how often an specific credit card application sent by the same applicant within a certain period of time, for example, 3 days, 7 days, 14 days or even 21 days. The more frequent one specific information by the same applicant is captured, the more likely this application is fraudulent behavior. In order to detect this kind of deeds, we use both match of single field to create variable 13 to 28 and combination of two fields to derive variable 29 to 68. On the other hand, we would like to know how much time since the last time we record specific application information. The shorter the discretion of time between two application attempts is, the more likely that the applicant manages to do some tricks. Likewise, single field is used to build variable 69 to 72 and two associated fields are utilized for variable 73 to 82.

**Note that for the second type of variable, which is the number of days we observe applications sent by the same applicant, if only one application was observed from an applicant, then we assign value 999 as a holder to that field. The rationale behind this is that the larger the number, the less likely that record represents a fraud act.*

Catalog of New Variables

No. of Variable	Variable Name	Mathematical Meaning
10	fullname	full name of the applicant (firstname and lastname)
11	person	fullname and date of birth of the applicant
12	yearday	day of year of the application date (e.g. 01/01/2015 is 1 and 12/31/2015 is 365)
13	3_ssn	how many times a specific ssn appears within 3 days before the record was created
14	3_address	how many times a specific address appears within 3 days before the record was created
15	3_homephone	how many times a specific homephone appears within 3 days before the record was created

No. of Variable	Variable Name	Mathematical Meaning
16	3_person	how many times a specific fullname+dob appears within 3 days before the record was created
17	7_ssn	how many times a specific ssn appears within 7 days before the record was created
18	7_address	how many times a specific address appears within 7 days before the record was created
19	7_homephone	how many times a specific homephone appears within 7 days before the record was created
20	7_person	how many times a specific fullname+dob appears within 7 days before the record was created
21	14_ssn	how many times a specific ssn appears within 14 days before the record was created
22	14_address	how many times a specific address appears within 14 days before the record was created
23	14_homephone	how many times a specific homephone appears within 14 days before the record was created
24	14_person	how many times a specific fullname+dob appears within 14 days before the record was created
25	21_ssn	how many times a specific ssn appears within 21 days before the record was created
26	21_address	how many times a specific address appears within 21 days before the record was created
27	21_homephone	how many times a specific homephone appears within 21 days before the record was created
28	21_person	how many times a specific fullname+dob appears within 21 days before the record was created
29	3_ssn_address	how many times a specific ssn+adress appears within 3 days before the record was created
30	3_ssn_zip	how many times a specific ssn+zip appears within 3 days before the record was created
31	3_ssn_homephone	how many times a specific ssn+homephone appears within 3 days before the record was created
32	3_ssn_person	how many times a specific ssn+fullname+dob appears within 3 days before the record was created
33	3_address_zip	how many times a specific address+zip appears within 3 days before the record was created

No. of Variable	Variable Name	Mathematical Meaning
34	3_address_homephone	how many times a specific address+homephone appears within 3 days before the record was created
35	3_address_person	how many times a specific address+fullname+dob appears within 3 days before the record was created
36	3_zip_homephone	how many times a specific zip+homephone appears within 3 days before the record was created
37	3_zip_person	how many times a specific zip+fullname+dob appears within 3 days before the record was created
38	3_homephone_person	how many times a specific homephone+fullname+dob appears within 3 days before the record was created
39	7_ssn_address	how many times a specific ssn+adress appears within 7 days before the record was created
40	7_ssn_zip	how many times a specific ssn+zip appears within 7 days before the record was created
41	7_ssn_homephone	how many times a specific ssn+homephone appears within 7 days before the record was created
42	7_ssn_person	how many times a specific ssn+fullname+dob appears within 7 days before the record was created
43	7_address_zip	how many times a specific address+zip appears within 7 days before the record was created
44	7_address_homephone	how many times a specific address+homephone appears within 7 days before the record was created
45	7_address_person	how many times a specific address+fullname+dob appears within 7 days before the record was created
46	7_zip_homephone	how many times a specific zip+homephone appears within 7 days before the record was created
47	7_zip_person	how many times a specific zip+fullname+dob appears within 7 days before the record was created
48	7_homephone_person	how many times a specific homephone+fullname+dob appears within 7 days before the record was created
49	14_ssn_address	how many times a specific ssn+adress appears within 14 days before the record was created
50	14_ssn_zip	how many times a specific ssn+zip appears within 14 days before the record was created
51	14_ssn_homephone	how many times a specific ssn+homephone appears within 14 days before the record was created

No. of Variable	Variable Name	Mathematical Meaning
52	14_ssn_person	how many times a specific ssn+fullname+dob appears within 14 days before the record was created
53	14_address_zip	how many times a specific address+zip appears within 14 days before the record was created
54	14_address_homephone	how many times a specific address+homephone appears within 14 days before the record was created
55	14_address_person	how many times a specific address+fullname+dob appears within 14 days before the record was created
56	14_zip_homephone	how many times a specific zip+homephone appears within 14 days before the record was created
57	14_zip_person	how many times a specific zip+fullname+dob appears within 14 days before the record was created
58	14_homephone_person	how many times a specific homephone+fullname+dob appears within 14 days before the record was created
59	21_ssn_address	how many times a specific ssn+adress appears within 21 days before the record was created
60	21_ssn_zip	how many times a specific ssn+zip appears within 21 days before the record was created
61	21_ssn_homephone	how many times a specific ssn+homephone appears within 21 days before the record was created
62	21_ssn_person	how many times a specific ssn+fullname+dob appears within 21 days before the record was created
63	21_address_zip	how many times a specific address+zip appears within 21 days before the record was created
64	21_address_homephone	how many times a specific address+homephone appears within 21 days before the record was created
65	21_address_person	how many times a specific address+fullname+dob appears within 21 days before the record was created
66	21_zip_homephone	how many times a specific zip+homephone appears within 21 days before the record was created
67	21_zip_person	how many times a specific zip+fullname+dob appears within 21 days before the record was created
68	21_homephone_person	how many times a specific homephone+fullname+dob appears within 21 days before the record was created
69	last_ssn	number of days since the last same ssn was observed

No. of Variable	Variable Name	Mathematical Meaning
70	last_address	number of days since the last same address was observed
71	last_homephone	number of days since the last same homephone was observed
72	last_person	number of days since the last same fullname+dob was observed
73	last_ssn_address	number of days since the last same ssn+address was observed
74	last_ssn_zip	number of days since the last same ssn+zip was observed
75	last_ssn_homephone	number of days since the last same ssn+homephone was observed
76	last_ssn_person	number of days since the last same ssn+fullname+dob was observed
77	last_address_zip	number of days since the last same address+zip was observed
78	last_address_homephone	number of days since the last same address+homephone was observed
79	last_address_person	number of days since the last same address+fullname+dob was observed
80	last_zip_homephone	number of days since the last same zip+homephone was observed
81	last_zip_person	number of days since the last same zip+fullname+dob was observed
82	last_homephone_person	number of days since the last same homephone+fullname+dob was observed

DATA MANIPULATION

Frivolous Values Manipulation

Frivolous values occur when someone fills out the format usually with simple yet meaningless value to complete a form. In our data, 4 frivolous values are given as below:

Field Name	Frivolous Value
dob	19070626
address + zip	2602 AJTJ AVE 68138
ssn	737610282
homephone	9105580920

Frivolous values can lead to serious distortion of the interpretation from insights. To deal with this problem, we manage to replace the values of our newly-created variables that have frivolous value involved, and we have the following procedures:

1. Exclude records with frivolous values and calculate the means for each artificially created variable
2. Replace newly created fields of records containing frivolous values with pre-calculated mean values from procedure 1

Correlation Test

A correlation matrix is necessary before we input our data into PCA for there might be some highly correlated columns which will affect the result and computation time of our algorithms, so we manually excluded variables which have correlations of 1 with each other. As we examine the correlation matrix, numbers of those variables we chose to discard are 28 (21_person), 50 (14_ssn_zip), 59 (21_ssn_address), 62 (21_ssn_person), 65 (21_address_person), 67 (21_zip_person), 68 (21_homephone_person), 79 (last_address_person). Our data ends up with 62 variables.

PRINCIPAL COMPONENT ANALYSIS (PCA)

Introduction

Principal component analysis (PCA) is an unsupervised dimension-reduction technique used to transform high-dimensional data, prior to fit in a machine learning algorithm, into a smaller dimensional subspace which still contains most of the information. By lowering dimensions and projecting data onto a new orthogonal-rotated coordinate system, PCA can help us to easily summarize the variations (informations) in a dataset with the first 2 principal components (PC) and gives us insights which is hard to capture in its original status.
The output dataset of PCA that we are going to use is a square symmetric square matrix composed of principal components, which corresponds to a linear combination of the original variables, on its column and the original variables on its row. The first PC accounts for as much of the variation in the dataset with highest Eigenvalue, and the succeeding PCs explains the remaining variation in a decreasing order with decreasing Eigenvalues.

Purposes

PCA is necessary for our analytics is as following:

1. Dimensionality Reduction:
   Plenty of expert variables were created but many of which could measure related or identical properties and are thus redundant. After PCA, we are able to summarize the Applications data with fewer and representative characteristics.
2. Variables Retaining:
   PCA captures the variables which explain the most variation of the data to form a new set of variables (PCs) without discarding the original fields in the dataset (minimum loss of information).
3. Better Understanding of Data Variation:
   Before PCA, we conducted mean normalization of all the input records for the sake of easiness of understanding the result. Consequently, the new coordinate system that PCA generates has the origin at the center of the data so that we can take further advantage of this to visualize and detect potential frauds by calculating the distance between certain records and the origin.
4. Preparation of Adopting Following Algorithms:
   The transformed data that we obtain from PCA is the crucial element of our fraud algorithms, which are respectively Euclidean Distance and Autoencoder.

Procedures & Visualizations

1. Statistical Tools Used:
   In this project, we use R as the major statistical modeling tool to perform PCA. In terms of packages in R, "factoextra" and "FactoMineR" are chosen to perform the task.
2. Programming Procedures:
   Step1: Convert NAs into 0. Since the PCA function will automatically deleted observations with NA
   Step 2: Specify the arguments "scale.unit = TRUE" so that R can automatically standardize all variables
   Step 3: Execute the PCA function with 34 variables
3. Plots and outputs:

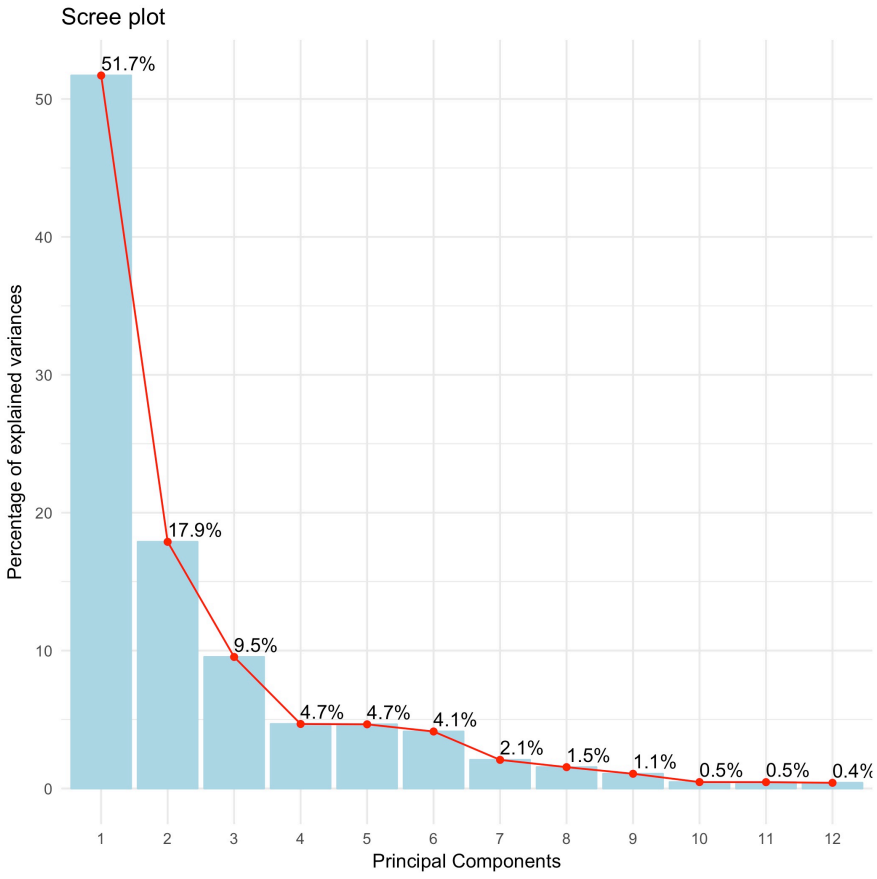


Figure 2. Scree Plot of PCA

The scree plot on the left can immediately show us how each PCs performs to explain the variation in the data. From PC 1 to PC 12 as a whole, the cumulative percentage of variation explained is 99.60%. PC 1 can explain up to 51.70% of the variation, and PC 2 has 17.90%. Finally, we decided to cut our data to PC 10.

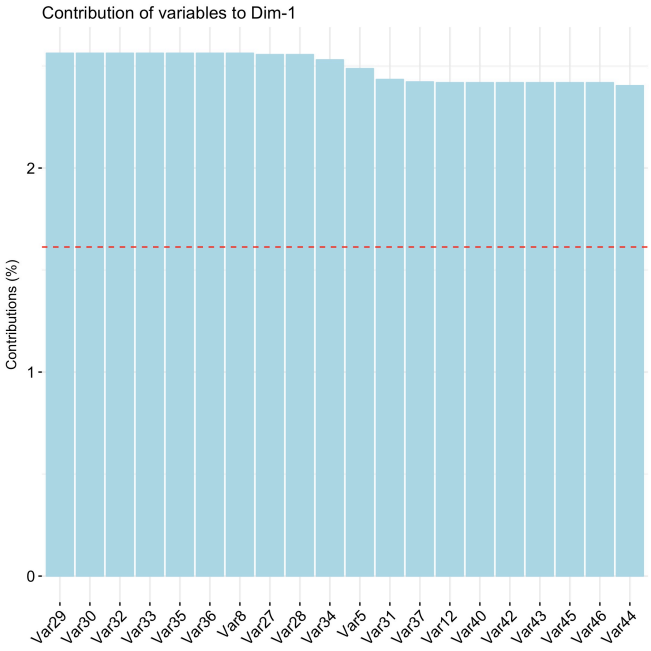


Figure 3. Contribution of Top 20 Variables to PC1

Figure 3 indicates the composition of PC1 by top-20-contributed variables. The red dashed line represents the average contribution of all variables. PC1 has relatively flat contribution distribution for all variables.

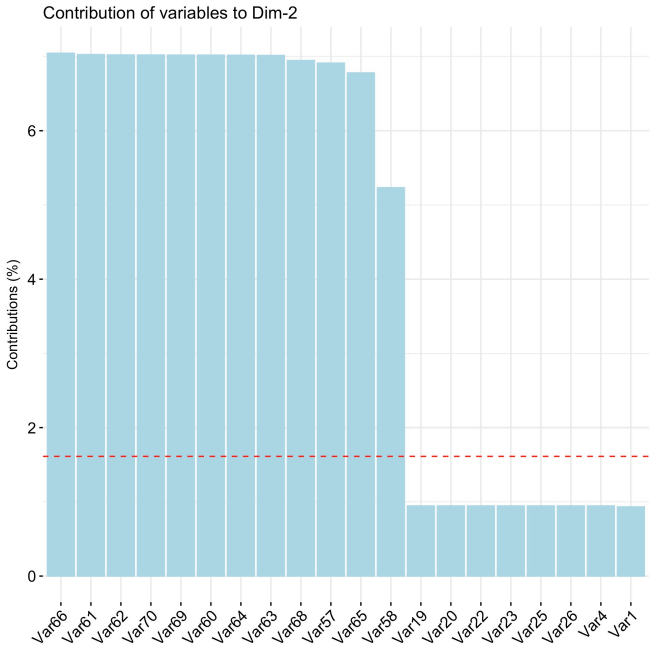


Figure 4. Contribution of Top 20 Variables to PC2

Figure 4 puts out the composition of PC2 by top-20-contributed variables. The red dashed line represents the average contribution of all variables. Top 12 variables contribute more to PC2.

CALCULATION OF FRAUD SCORE

Heuristic Algorithm

We have implemented Mahalanobis Distance (z-scale then euclidean) for this part of algorithm construction.

1. Euclidean Distance
   The Euclidean distance or Euclidean metric is the "ordinary" (i.e. straight-line) distance between two points in Euclidean space. The Euclidean norm, or Euclidean length, or magnitude of a vector measures the length of the vector. And it takes the following form:

$$\|\mathbf{P}\| = \sqrt{p_1^2 + p_2^2 + .. + p_n^2} = \sqrt{P \times P}$$

   Thus, in our case, the distance would be: $\sqrt{PC1^2 + PC2^2 + ... + PC13^2}$ as we choose 10 PCs.
   The distribution is shown as follow. We can observe that it generally right skewed and has a long tail.

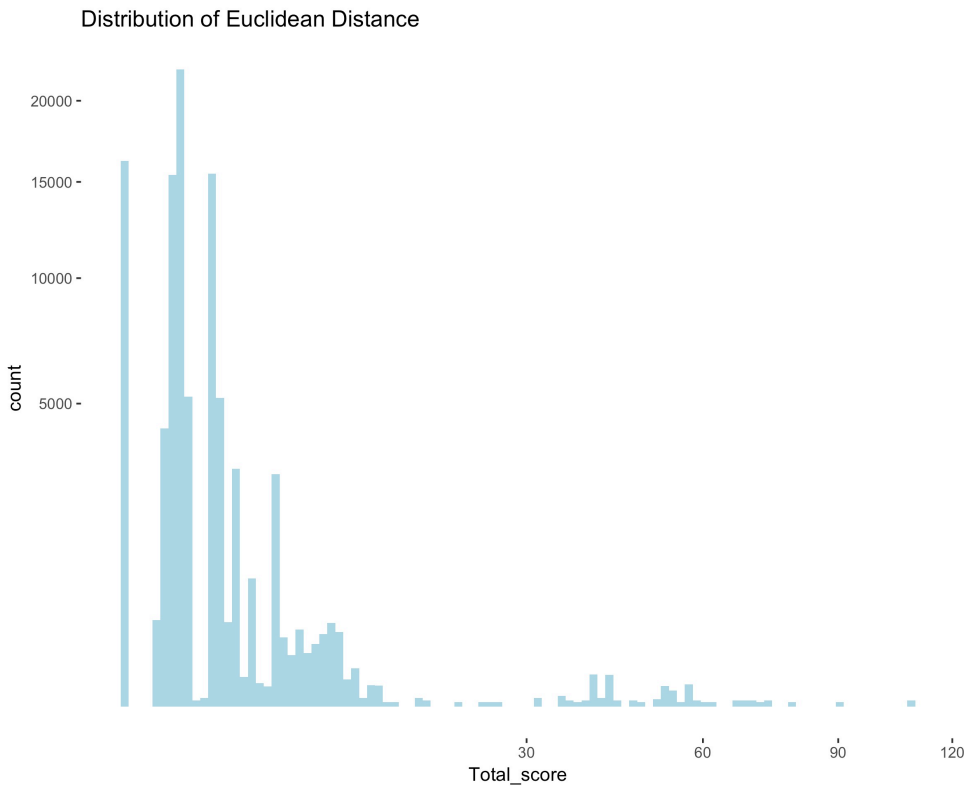


Figure 5. Distribution of Sum of Euclidean Distance

The main information of top 10 records with highest fraud score are shown as follow.

record	date	ssn	firstname	lastname	address	zip5	dob	homephone
92975	20151206	466088456	MRXZETSSM	UXEAEEMM	2602 AJTJ AVE	20045	20011102	6291914317
81534	20151024	236033166	SAEMRSZTT	RRRAZMEM	2602 AJTJ AVE	73044	19790210	9002498071
45115	20150613	646783682	XAMJEEMEJ	RESRJZUT	520 RXJZZ BLVD	68165	19621214	8789038124
82477	20151028	393296221	SMESZZZXM	REUSAUZA	2602 AJTJ AVE	30916	19030523	3769502961
68942	20150908	468575344	XUJXXZSUE	SAJJJAXA	2602 AJTJ AVE	98397	19530207	1136388978
71178	20150916	873258499	XMMZMAETM	EERMERJ	2602 AJTJ AVE	44145	20080512	2019407715
10072	20150206	641813750	SMJSMAUJJ	EASRUUTT	2602 AJTJ AVE	33722	19180519	7143573880
8542	20150131	2608661	RATAURZRT	ETAEXMJ	2602 AJTJ AVE	30224	19740324	5568704443
51260	20150706	677890626	SAMXMEZMS	EXUSUEZR	2602 AJTJ AVE	18836	19570930	1712559411
45667	20150615	790302381	UJSRSMUEZ	SXTMSRJS	7097 STTZE ST	73289	19180913	616958150

2. Sum of Absolute Value of PCs

   In the second algorithm, we choose to add up the absolute value of each PC to measure fraud score, and the distribution is shown as follow. In this graph we can observe that the distribution is still right skewed while with two peak values, just like bimodal distribution.

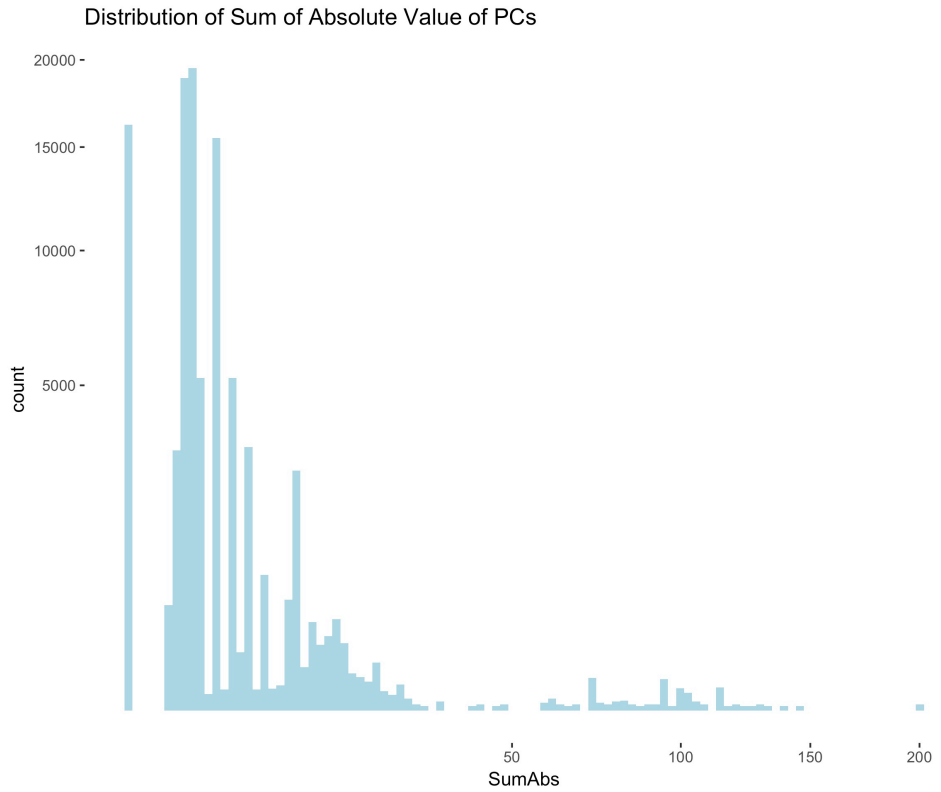


Figure 6. Distribution of Sum of Absolute Value of PCs

The main information of top 10 records with highest fraud score are shown as follow.

record..	date	ssn	firstname	lastname	address	zip5	dob	homephone
81534	20151024	236033166	SAEMRSZTT	RRRAZMEM	2602 AJTJ AVE	73044	19790210	9002498071
92975	20151206	466088456	MRXZETSSM	UXEAEEMM	2602 AJTJ AVE	20045	20011102	6291914317
45115	20150613	646783682	XAMJEEMEJ	RESRJZUT	520 RXJZZ BLVD	68165	19621214	8789038124
82477	20151028	393296221	SMESZZZXM	REUSAUZA	2602 AJTJ AVE	30916	19030523	3769502961
8542	20150131	2608661	RATAURZRT	ETAEXMJ	2602 AJTJ AVE	30224	19740324	5568704443
10072	20150206	641813750	SMJSMAUJJ	EASRUUTT	2602 AJTJ AVE	33722	19180519	7143573880
68942	20150908	468575344	XUJXXZSUE	SAJJJAXA	2602 AJTJ AVE	98397	19530207	1136388978
71178	20150916	873258499	XMMZMAETM	EERMERJ	2602 AJTJ AVE	44145	20080512	2019407715
97482	20151222	785465235	RURUTMTJU	RXRERTUT	2602 AJTJ AVE	49688	19180428	4100050507
51260	20150706	677890626	SAMXMEZMS	EXUSUEZR	2602 AJTJ AVE	18836	19570930	1712559411

3. Maximum Absolute PC Value

   As some records with only one or few abnormal value may be left over if we look at the sum of 13 PCs at the same time, we this time measure the fraud score by take the absolute value to each PC and choose the maximum one. From the graph below, we can see that the distribution is rather uneven with several ups and downs.

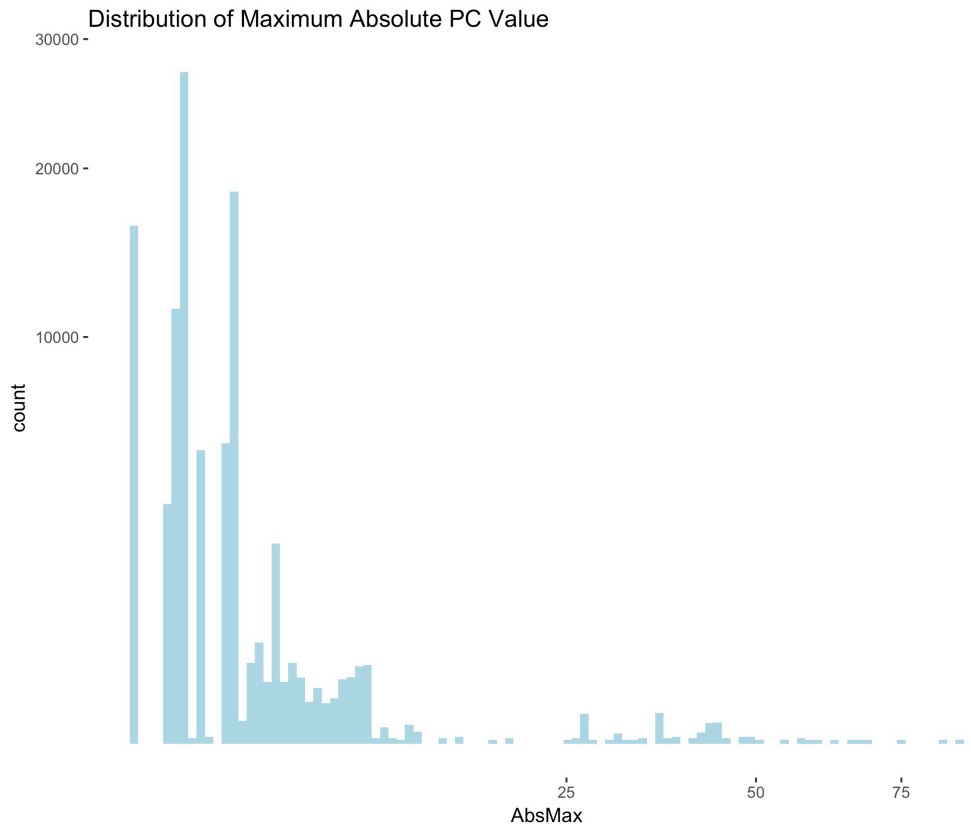


Figure 7. Distribution of Maximum Absolute PC Value

The main information of top 10 records with highest fraud score are shown as follow.

record..	date	ssn	firstname	lastname	address	zip5	dob	homephone
45115	20150613	646783682	XAMJEEMEJ	RESRJZUT	520 RXJZZ BLVD	68165	19621214	8789038124
81534	20151024	236033166	SAEMRSZTT	RRRAZMEM	2602 AJTJ AVE	73044	19790210	9002498071
92975	20151206	466088456	MRXZETSSM	UXEAEEMM	2602 AJTJ AVE	20045	20011102	6291914317
82477	20151028	393296221	SMESZZZXM	REUSAUZA	2602 AJTJ AVE	30916	19030523	3769502961
45667	20150615	790302381	UJSRSMUEZ	SXTMSRJS	7097 STTZE ST	73289	19180913	616958150
54012	20150716	764983006	XSUTSXRXT	SASAXAZA	7019 SSEA PL	44145	19270116	1737899096
68942	20150908	468575344	XUJXXZSUE	SAJJJAXA	2602 AJTJ AVE	98397	19530207	1136388978
71178	20150916	873258499	XMMZMAETM	EERMERJ	2602 AJTJ AVE	44145	20080512	2019407715
67718	20150904	301715768	STRUTAET	ERJSAXA	2602 AJTJ AVE	8705	19820414	2591792474
51260	20150706	677890626	SAMXMEZMS	EXUSUEZR	2602 AJTJ AVE	18836	19570930	1712559411

4. Summary
   Although using euclidean distance, sum of absolute value of PCs and maximum absolute PC value give different fraud score distribution, the records they detect have high overlap ratio. There is only one different record in the former methods. Those overlapped records detected by the former two methods are in RECORD #: 92975, 81534, 45115, 82477, 68942, 71178, 10072, 8542, 51260

Autoencoder

An Autoencoder neural network is an unsupervised learning approach that applied back propagation. It has very interesting features. It sets the target values that equals to the inputs and reproduce the input data by doing this. The hidden layer in the neural network enables non-linear transformations that different from PCA. With those non-linear features, it enables Autoencoder to possess a more powerful performance than linear approaches.

We implement Autoencoder to score the fraud of our record. We project each record on the top 10 principal directions and therefore compress features of each record to 10. This new dataset with feature number of 13 serves as the input dataset of Autoencoder. We use two hidden layers in the neural network and each hidden layer has a length of 4. We use a package call “h2o” in R to implement this approach and the output is also a dataset which has the same dimension with the input dataset.

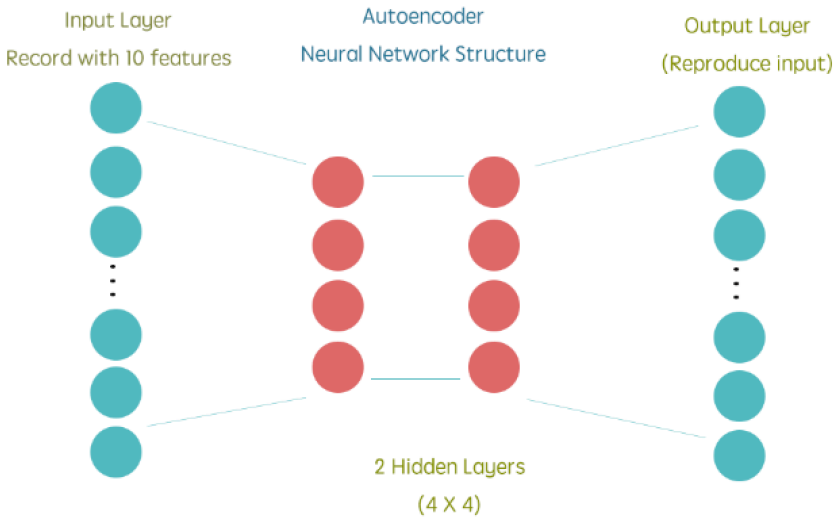


Figure 8. Autoencoder Neural Network Structure

The Autoencoder trains data in its neural network and discovered a pattern within the data. Output explains how well that each record corresponds with the pattern. We define the fraud score as the mean square error (MSE). It measures the distance between the input dataset and the output dataset of each record. And figure below shows the distribution of MSE.

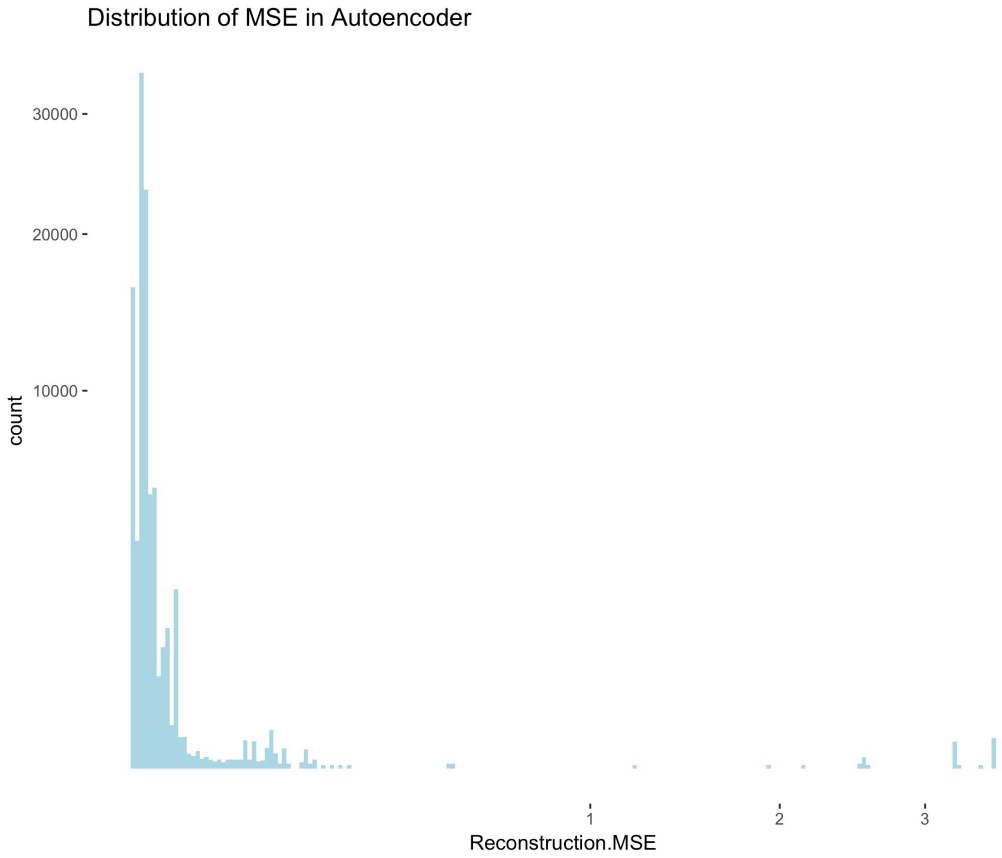


Figure 9. Distribution of MSE with Autoencoder

The main information of top 10 records with highest fraud score are shown as follow.

record	date	ssn	firstname	lastname	address	zip5	dob	homephone
89852	20151124	597458192	UZEEJUUAE	UXJUSTSR	9832 RMJME ST	66273	19161013	804556630
10982	20150209	3313643	XAEEAUJUS	UEMMSJMS	721 SAMZA DR	8106	19630429	2113738531
91874	20151202	705091123	EZXUMUUMR	ERRASRTX	6888 XJUXR LN	55194	19931208	6384782007
6046	20150122	665908664	SSUTSUUMX	SZMSRZMS	2778 EMUTZ AVE	5952	19530530	6410726346
6143	20150123	383335253	MSRUAREMU	RJURMAZS	7440 ZTRX BLVD	82122	19910821	2341561415
6594	20150124	927297824	SXXSZJMM	ESSXURAT	4692 RTJZ AVE	53690	19540512	5583968023
6782	20150125	57081832	RTMMTSZRZ	RSSZRUTE	5356 UAZZU AVE	62263	19410331	1099821057
7001	20150126	98332843	UUMEERXXZ	RSSATUMA	2882 EZXJR WY	39664	20040729	7689360769
7234	20150127	750055478	XETRJXESR	EARTMSMR	1898 UUEAE RD	4085	19700701	8570233435
7308	20150127	613960795	STRXXSXMR	EMXXSSUA	9320 ETJZJ ST	98295	20110321	5179499608

COMPARISON BETWEEN 2 CALCULATIONS

Summary

Compare the top 1,000 highest fraud scores in both Euclidean Distance and Autoencoder and we find there is 34.80% overlapping between the two approaches.

In the Euclidean Distance approach, records with top 10 fraud record are listed below,
Record: 92975, 81534, 45115, 82477, 68942, 71178, 10072, 8542, 51260, 45667

And in the Autoencoder approach, records with top 10 fraud record are listed below,
Record: 89852, 10982, 91874, 6046, 6143, 6594, 6782, 7001, 7234 7308

INSIGHTS AND CONCLUSIONS

1. Top 10 Anomalies of Euclidean Distance Approach

record	date	ssn	firstname	lastname	address	zip5	dob	homephone
92975	20151206	466088456	MRXZETSSM	UXEAEEMM	2602 AJTJ AVE	20045	20011102	6291914317
81534	20151024	236033166	SAEMRSZTT	RRRAZMEM	2602 AJTJ AVE	73044	19790210	9002498071
45115	20150613	646783682	XAMJEEMEJ	RESRJZUT	520 RXJZZ BLVD	68165	19621214	8789038124
82477	20151028	393296221	SMESZZZXM	REUSAUZA	2602 AJTJ AVE	30916	19030523	3769502961
68942	20150908	468575344	XUJXXZSUE	SAJJJAXA	2602 AJTJ AVE	98397	19530207	1136388978
71178	20150916	873258499	XMMZMAETM	EERMERJ	2602 AJTJ AVE	44145	20080512	2019407715
10072	20150206	641813750	SMJSMAUJJ	EASRUUTT	2602 AJTJ AVE	33722	19180519	7143573880
8542	20150131	2608661	RATAURZRT	ETAEXMJ	2602 AJTJ AVE	30224	19740324	5568704443
51260	20150706	677890626	SAMXMEZMS	EXUSUEZR	2602 AJTJ AVE	18836	19570930	1712559411
45667	20150615	790302381	UJSRSMUEZ	SXTMSRJS	7097 STTZE ST	73289	19180913	616958150

2. Top 10 Anomalies of Autoencoder Approach

record	date	ssn	firstname	lastname	address	zip5	dob	homephone
89852	20151124	597458192	UZEEJUUAE	UXJUSTSR	9832 RMJME ST	66273	19161013	804556630
10982	20150209	3313643	XAEEAUJUS	UEMMSJMS	721 SAMZA DR	8106	19630429	2113738531
91874	20151202	705091123	EZXUMUUMR	ERRASRTX	6888 XJUXR LN	55194	19931208	6384782007
6046	20150122	665908664	SSUTSUUMX	SZMSRZMS	2778 EMUTZ AVE	5952	19530530	6410726346
6143	20150123	383335253	MSRUAREMU	RJURMAZS	7440 ZTRX BLVD	82122	19910821	2341561415
6594	20150124	927297824	SXXSZJMM	ESSXURAT	4692 RTJZ AVE	53690	19540512	5583968023
6782	20150125	57081832	RTMMTSZRZ	RSSZRUTE	5356 UAZZU AVE	62263	19410331	1099821057
7001	20150126	98332843	UUMEERXXZ	RSSATUMA	2882 EZXJR WY	39664	20040729	7689360769
7234	20150127	750055478	XETRJXESR	EARTMSMR	1898 UUEAE RD	4085	19700701	8570233435
7308	20150127	613960795	STRXXSXMR	EMXXSSUA	9320 ETJZJ ST	98295	20110321	5179499608

**In our data cleaning process, we replace frivolous value with neutral value and thus our final top 10 anomalies in both methods do not contain frivolous value.*

In the data cleaning process, we replaced frivolous values with neutral values, thus our final top 10 anomalies in both methods do not contain frivolous values.

To further understand why these records are considered "fraud" by the algorithm, we look into which variables in such records are most abnormal, as listed below. While on average there should be no repetitive occurrence of certain variable value in past 3/7/14/21 days, we noticed that for the selected records, they contain variables that appear multiple times in past several days. For example, record 82477 has its address appeared 12 times in the past 21 days. Also in some records, variables such as Address even appear more than once in the same day (minimum number of days since last saw is 0).

FRAUD DETECTION — IDENTITY THEFT

Top 10 records by Heuristic Method

Record Number	Which variables in this record appeared most times in past 3 days		Which variables in this record appeared most times in past 7 days		Which variables in this record appeared most times in past 14 days		Which variables in this record appeared most times in past 21 days		Min. days since we last saw certain variables in this record	
	Max times	Variables or combinations	Max times	Variables or combinations	Max times	Variables or combinations	Max times	Variables or combinations	Min days	Variables or combinations
92975	3	Address	5	Address	6	Address	10	Address	0	Address
81534	3	Address	5	Address	8	Address	11	Address	0	Address
45115	1	Address/Homephone/SSN+Address/SSN+Zip/Address+Zip	1	SSN/Address/Homephone/SSN+Address/SSN+Zip/Address+Zip	2	Homephone	2	Homephone	1	SSN/Address/SSN+Address/SSN+Zip/Address+Zip
82477	1	Address	5	Address	7	Address	12	Address	4	Address
68942	1	Address	4	Address	8	Address	10	Address	0	Address
71178	0		1	Address	6	Address	10	Address	3	Address
10072	2	Address	3	Address	6	Address	8	Address	0	Address
8542	2	Address	3	Address	5	Address	6	Address	1	Address
51260	2	Address	2	Address	2	Address	4	Address	1	Address
45667	1	Address/Address+Zip	1	Address/Address+Zip	1	Address/Address+Zip	1	Address/Address+Zip	2	Address/Address+Zip
Compare with average value	0.0033		0.0075		0.0159		0.0375		925.76(*we use 999 to indicate that the variables never appear in the past)	

Top 10 records by Autoencoder

Record Number	Which variables in this record appeared most times in past 3 days		Which variables in this record appeared most times in past 7 days		Which variables in this record appeared most times in past 14 days		Which variables in this record appeared most times in past 21 days		Min. days since we last saw certain variables in this record	
	Max times	Variables or combinations	Max times	Variables or combinations	Max times	Variables or combinations	Max times	Variables or combinations	Min days	Variables or combinations
89852	2	homephone	3	homephone	3	homephone	3	homephone	2	homephone
10982	2	homephone	3	homephone	3	homephone	3	homephone	1	homephone
91874	2	homephone	3	homephone	3	homephone	4	homephone	1	homephone
6046	1	homephone	1	homephone	1	homephone	1	homephone	Hasn't appeared in the past	
6143	1	homephone	1	homephone	1	homephone	1	homephone	Hasn't appeared in the past	
6594	1	homephone	1	homephone	1	homephone	1	homephone	Hasn't appeared in the past	
6782	1	homephone	1	homephone	1	homephone	1	homephone	Hasn't appeared in the past	
7001	1	homephone	1	homephone	1	homephone	1	homephone	Hasn't appeared in the past	
7234	1	homephone	1	homephone	1	homephone	1	homephone	Hasn't appeared in the past	
7308	1	homephone	1	homephone	1	homephone	1	homephone	Hasn't appeared in the past	
Compare with average value	0.0033		0.0075		0.0159		0.0375		925.76(*we use 999 to indicate that the variables never appear in the past)	

Assumption and Further Discussion:

1.Normally, people who apply for credit card won't submit multiple applications within a month. Frequent application is an important symbol of “fraud”, and can be caught by our algorithm.

2. Records with high fraud score captured by the heuristic approach tend to repeat more times compared to the Autoencoder approach.

3. By observing the major variables that occur repeatedly, we conclude that Address and Homephone are two pieces of information that people subconsciously don’t change. Possible reason could be: even though they conduct frauds, they need a stable address where they can have access to relevant documents, as well as a stable number that can keep them updated.

APPENDIX

Data Quality Report (DQR)

I. FILE DESCRIPTION

File Name:

Applications Data

Number of Records:

100,000 Records

Fields:

Total of 9 Variables (9 Categorical Variables)

Time Frame:

From "2015-01-01" to "2015-12-31"

II. INFORMATION FOR EACH FIELD

Record

1. Description:
   The variable serves as primary to identify each record.

Characteristics	Value
Minimum	1
1st Quantile	25001
Median	50000
3rd Quantile	75000
Maximum	100000

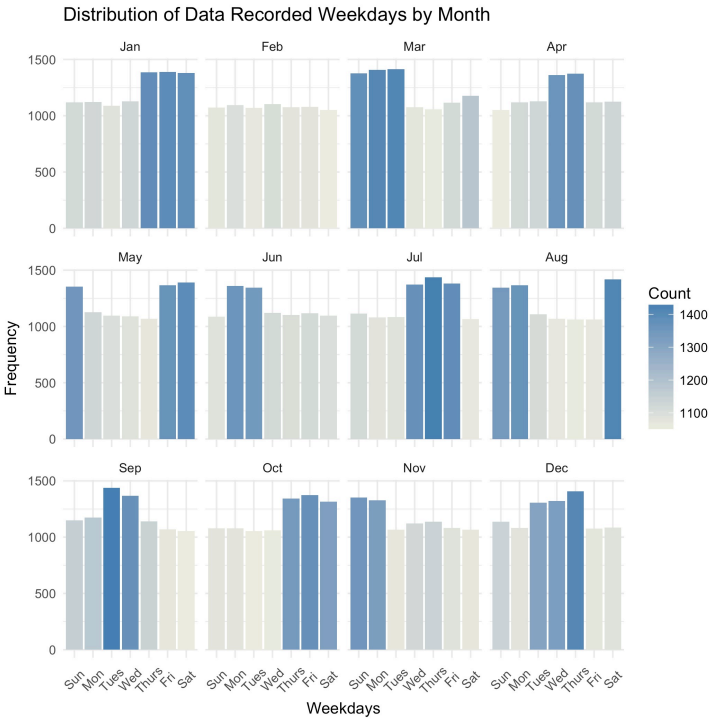
date

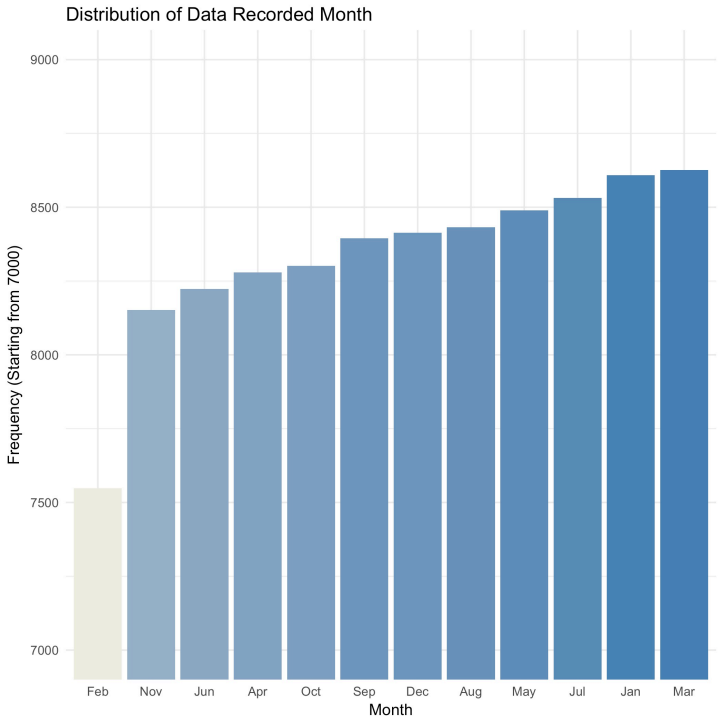
1. Description:
   Date range of the data recorded

2. Summary Statistics:

Minimum	1st Quantile	Median	3rd Quantile	Maximum
2015-01-01	2015-04-01	2015-07-01	2015-09-30	2015-12-31

3. Visualization:





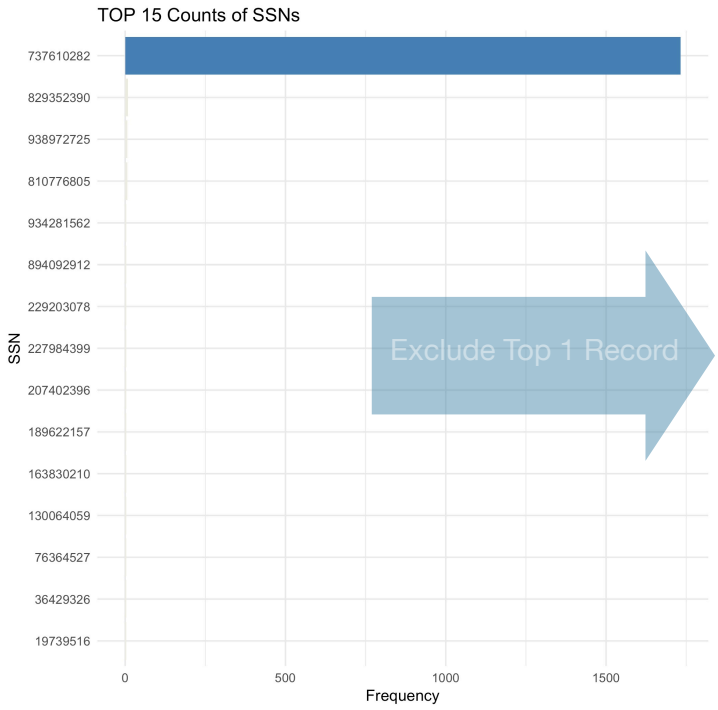
ssn

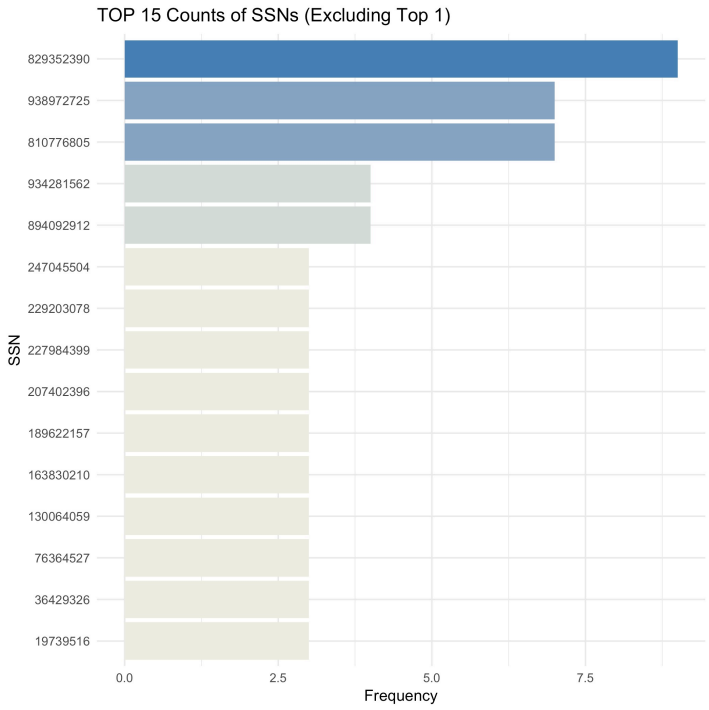
1. Description:
   Social security number of each observation

2. Summary Statistics:

Number of Unique Value
96535

3. Visualization:





SSN of 737610282 is a frivolous value with unparalleled number of counts, so we removed this record to create the plot on the right, but I did not delete the original plot to show its anomaly.

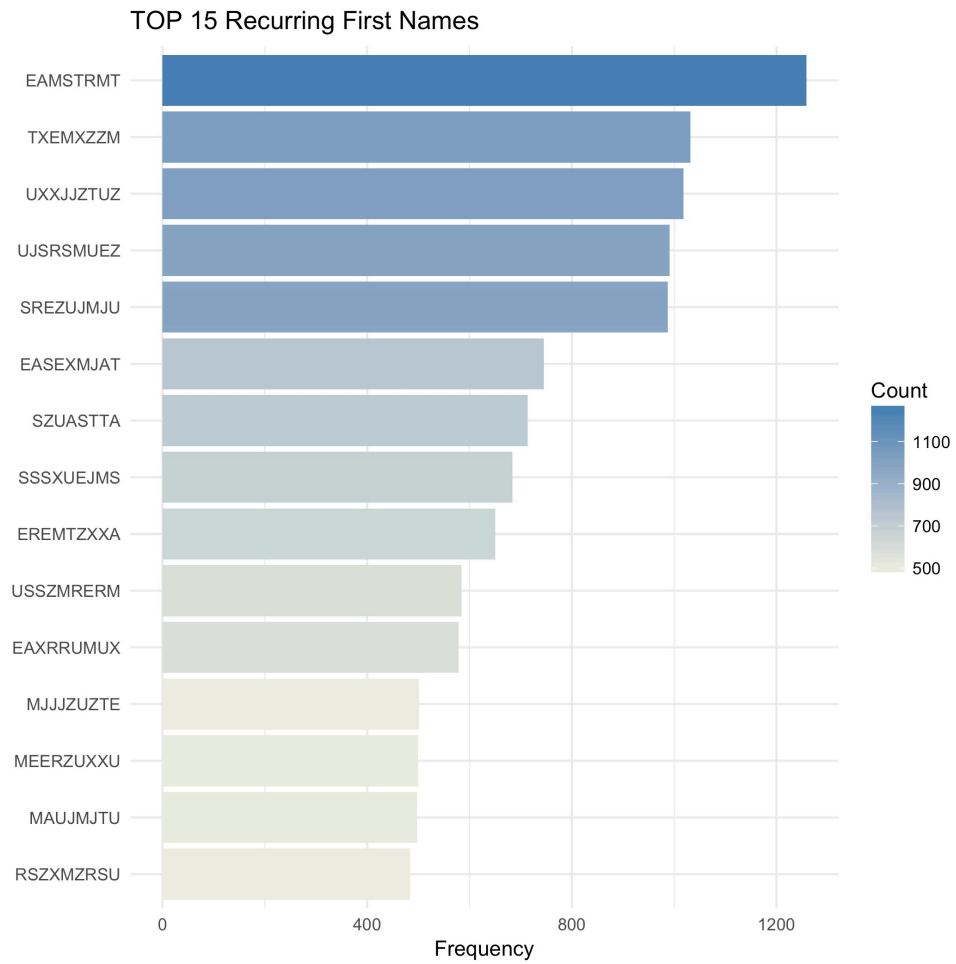
firstname

1. Description:
   The first names of each observation

2. Summary Statistics:

Number of Unique Value
16576

3. Visualization:



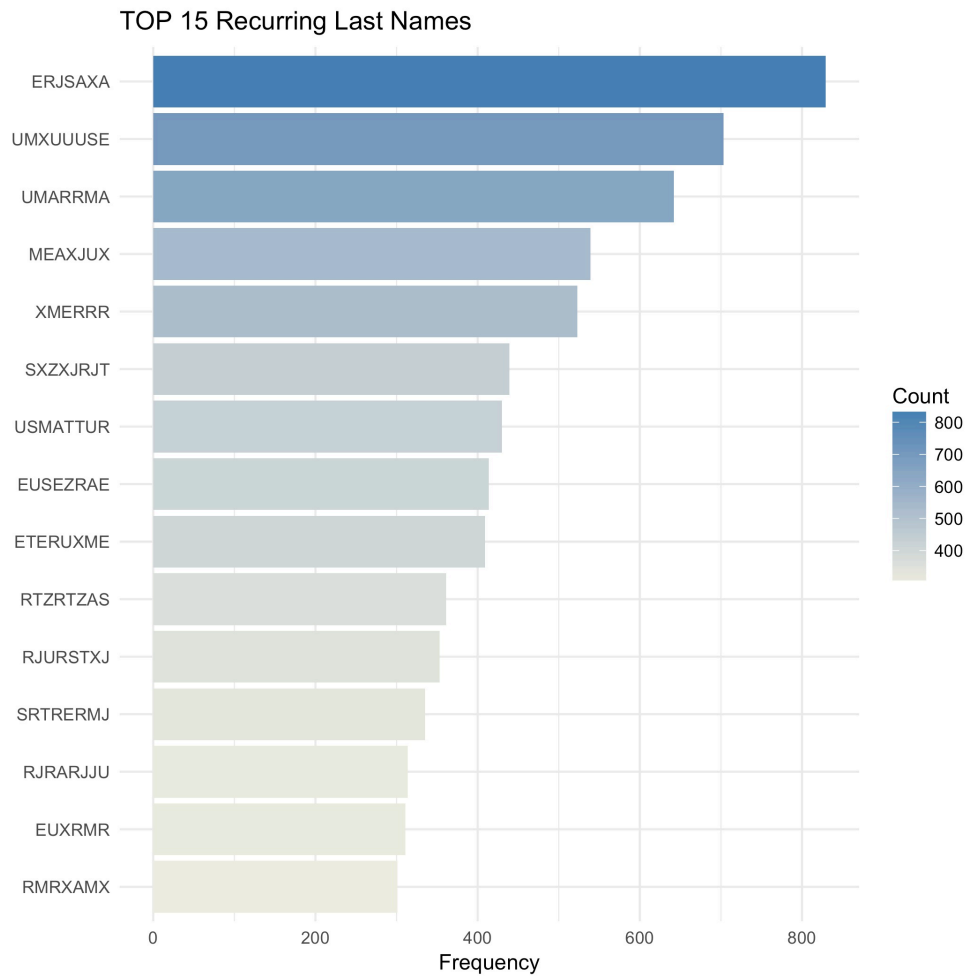
lastname

1. Description:
   The last names of each observation

2. Summary Statistics:

Number of Unique Value
36312

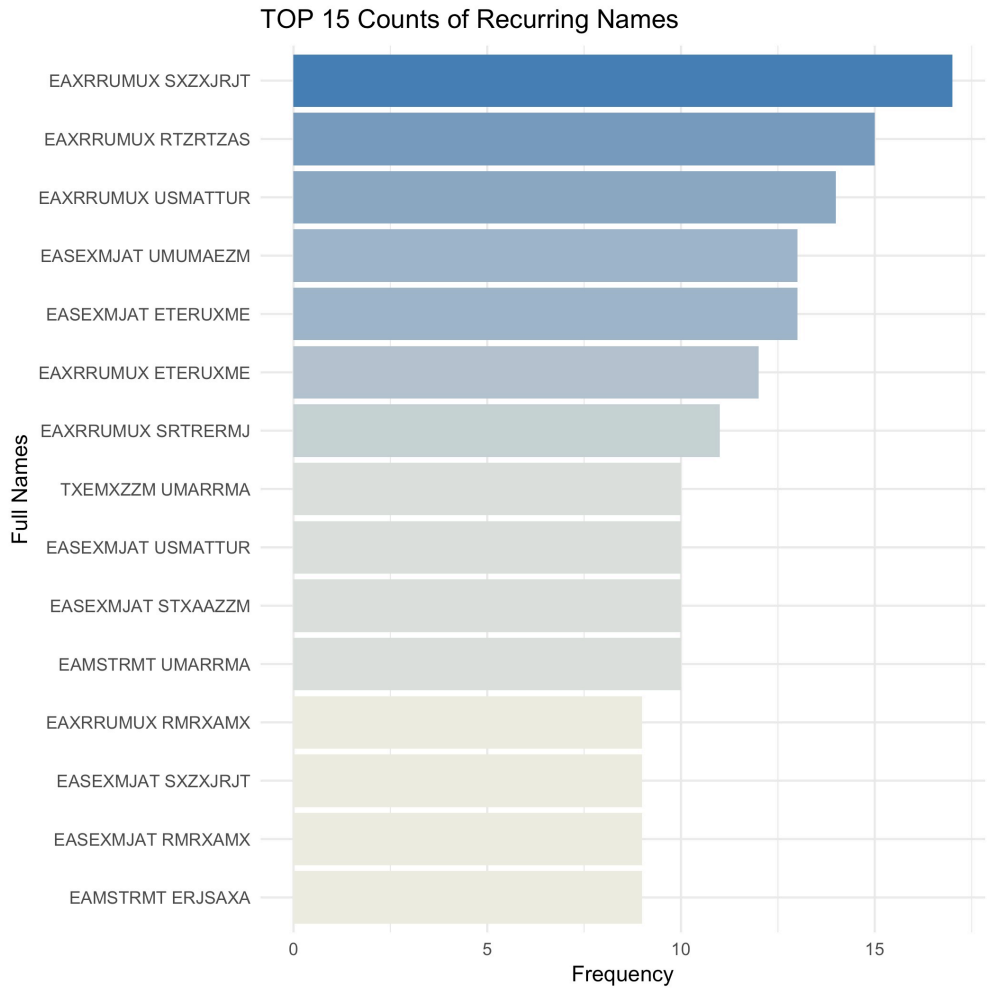
3. Visualization:



firstname + lastname

1. Description:
   Combined Variable from firstname and lastname

2. Visualization:



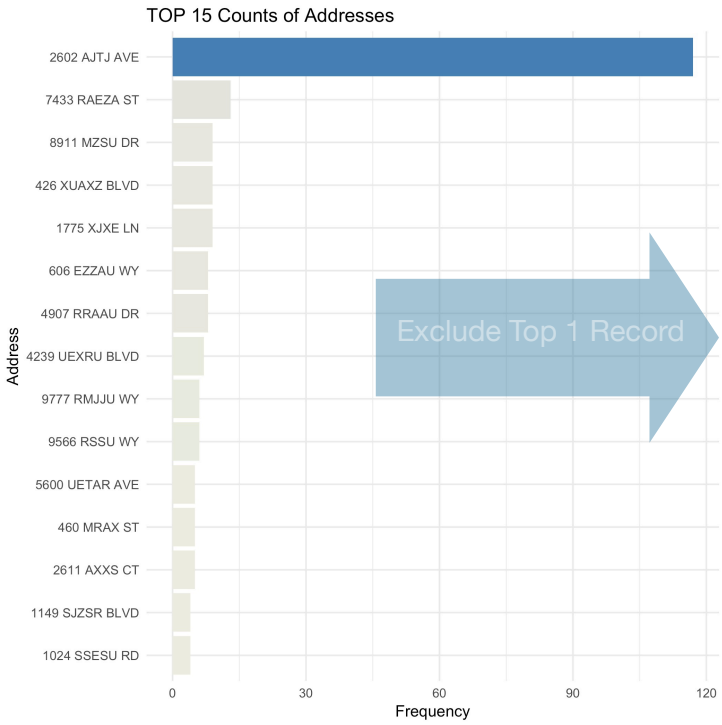
address

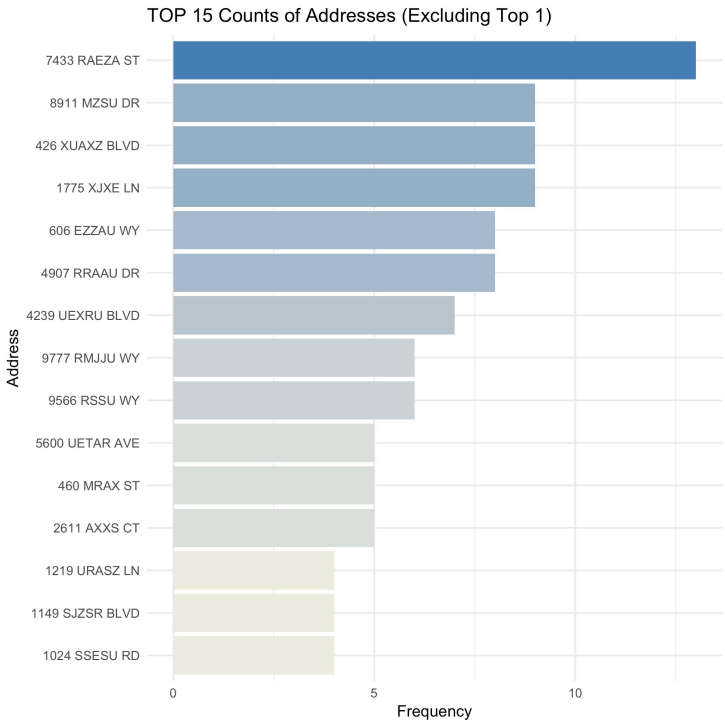
1. Description:
   The addresses of each observation

2. Summary Statistics:

Number of Unique Value
97563

3. Visualization:





Address of 2602 AJTJ AVE is a frivolous value with unparalleled number of counts, so we removed this record to create the plot on the right, but I did not delete the original plot to show its anomaly.

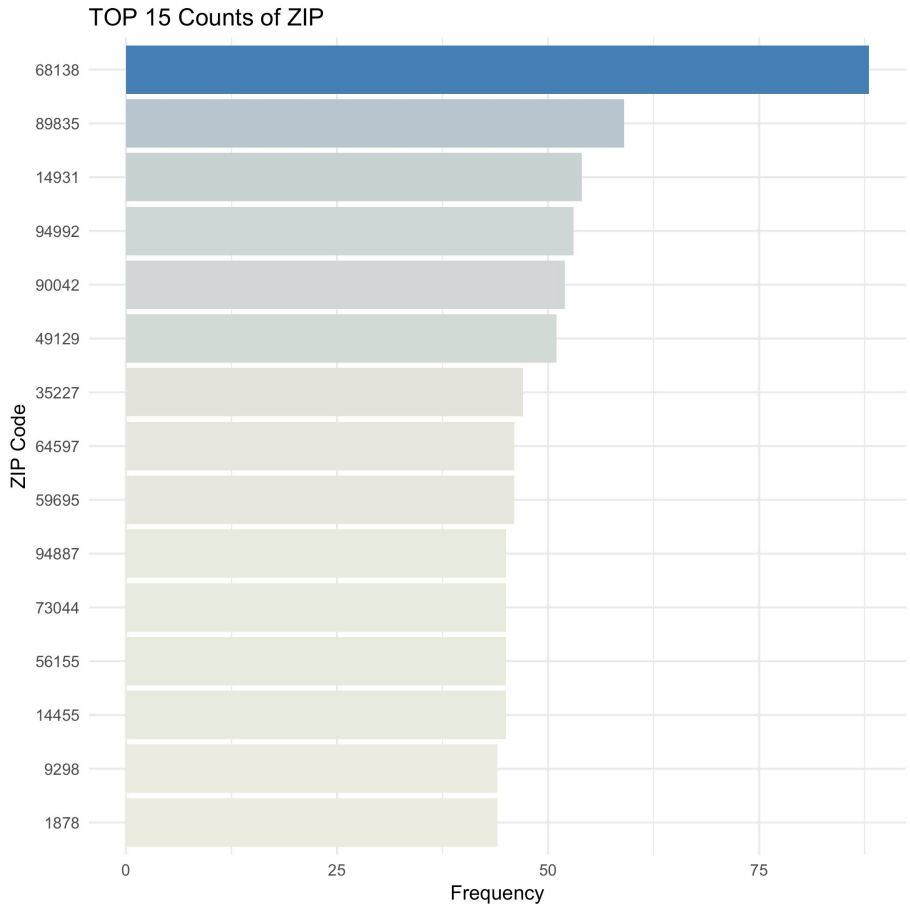
zip5

1. Description:
   The zip codes of each observation's address

2. Summary Statistics:

Number of Unique Value
16547

3. Visualization:



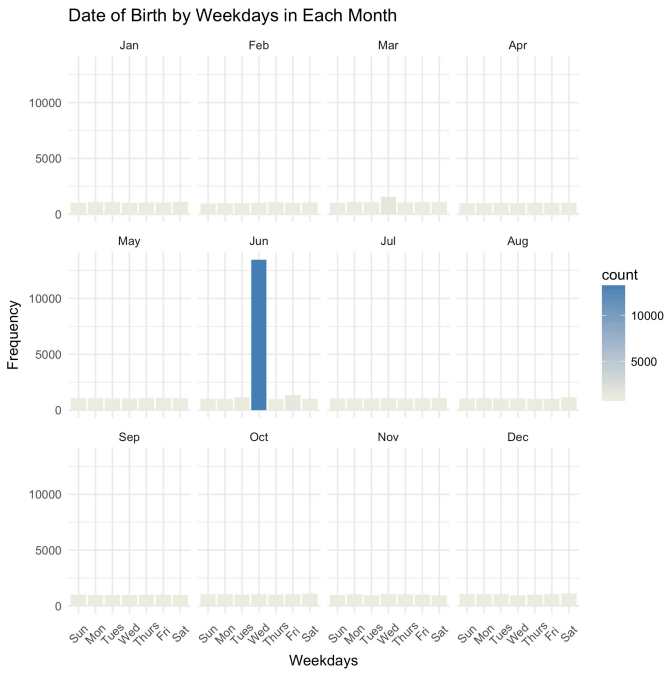
dob

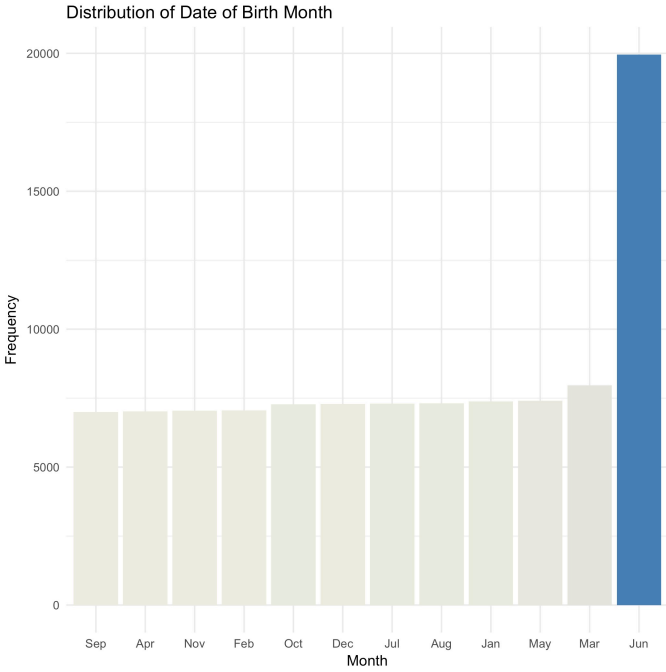
1. Description:
   The date of birth of each observation

2. Summary Statistics:

Number of Unique Value
36816

3. Visualization:





Date of birth on Wednesday in June (06/26/1907) is a frivolous value with unparalleled number of counts comparing to others.

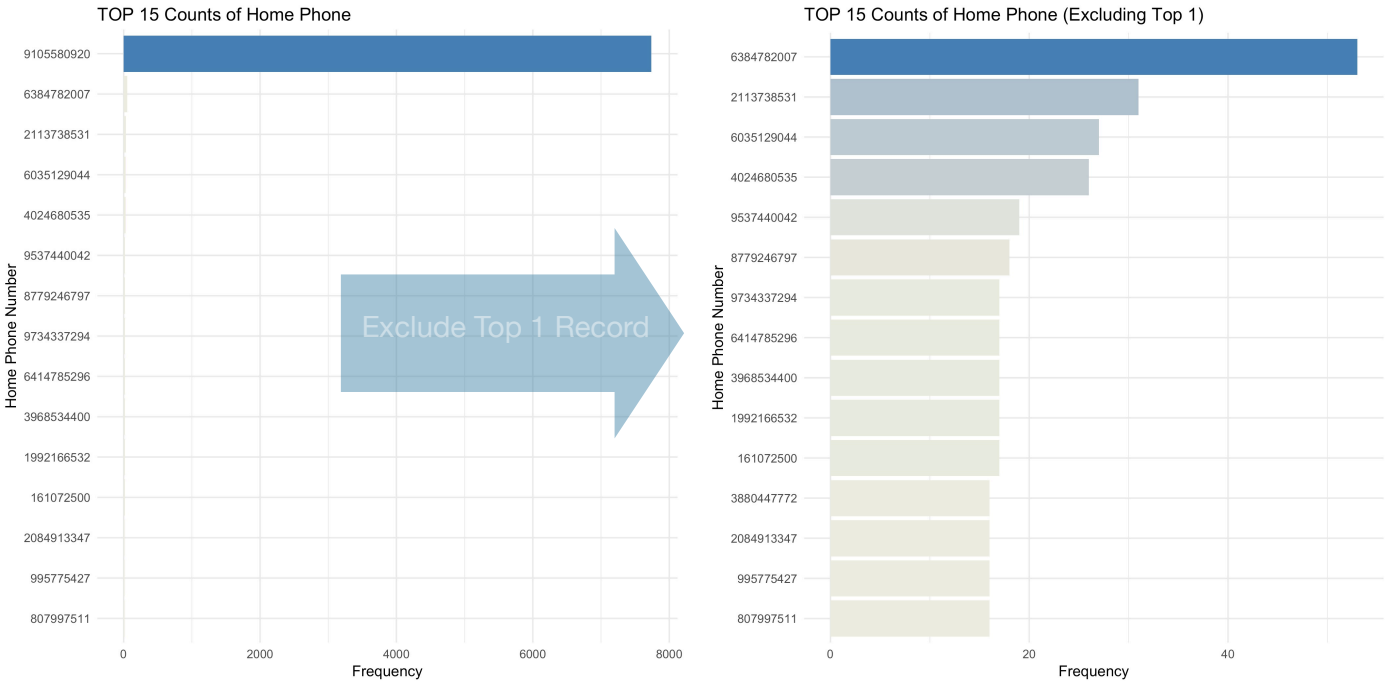
homephone

1. Description:
   The phone numbers of each observation

2. Summary Statistics:

Number of Unique Value
22181

3. Visualization:



Phone number of 9105580920 is a frivolous value with unparalleled number of counts, so I removed this record to create the plot on the right, but I did not delete the original plot to show its anomaly.