Unsupervised Model on NYC Property Data

Minglong Cen
Rui Cao
Rui Liu
07/17/2017

Content

Part I Executive Summary.....3

Part II Data Description 4

Part III Data preparation.....13

Part IV Variable Creation.....14

Part V Algorithm.....15

Part VI Results.....19

Appendix.....25

Part I: Executive Summary

Property fraud is the fastest growing white-collar crime according to the FBI. Everyone that owns property should be concerned.

Possible modes of property fraud: Assessed value too low, Person claiming an exemption for more than one property, claiming exemption on rental property, Other unusual things to look for, Mistakes (values too high or too low), Names or addresses incorrect (maybe too many entries) and so on.

This report aims to detect potential New York Property fraud based on the public provided NYC Property Valuation and Assessment Data file. For fraud detection purpose, we conducted unsupervised machine learning methods, including Principal Component Analysis and Autoencoder Algorithm in R.

The original NYC Property Valuation and Assessment data includes 1,048,575 number of distinct records with information of property features, ownership, etc. We conducted following steps in R: data exploration, data cleansing, expert variable creation, data scaling and model building. We had two models: Principal Component Analysis and Autoencoder. For Principal Component Analysis, we chose top 6 principal components, which explained more than 90% of the overall variance. Those 6 PCs were selected for the second z-scaling and heuristic method to obtain the final fraud score of each record (Euclidean distance). We used Autoencoder to reproduce the z-scaled PC scores with 2 hidden layers. Each hidden layer has 10 nodes. The fraud scores were measured as the difference between the original input record and the autoencoder output.

We explored the highest 0.1% fraud scores (1040 number of records) and identified the overlapped records in two distinct methods. After that we did statistical analysis to gain some business insights.

We got top 10 overlapped abnormal records and we think they could be our best candidates for potential property fraud.

Our report may include the following limitations:

1.We do not have comprehensive understanding of NY property legislations and variables;

2.We only have 15 expert variables which only include 10 fields information of the dataset;

3.When we handle the missing value, we use mode and median to replace, this may cause some bias.

Part II: Data Description

The City of New York Property Valuation and Assessment Data file derives from NYC open data source created by the Department of Finance on the City of New York, which collects public records for Bronx, Brooklyn, Manhattan, and Queens. The datasets contain 1,048,575 number of distinct records with information on property features, ownership info, etc. There are 30 variables of this dataset including 16 numeric variables, 12 categorical variables and 2 characters. All the records are taken from November 2010.

Following is description of important fields and the full data quality report will be shown in appendix.

Description of Important Fields:

(1) TAXCLASS

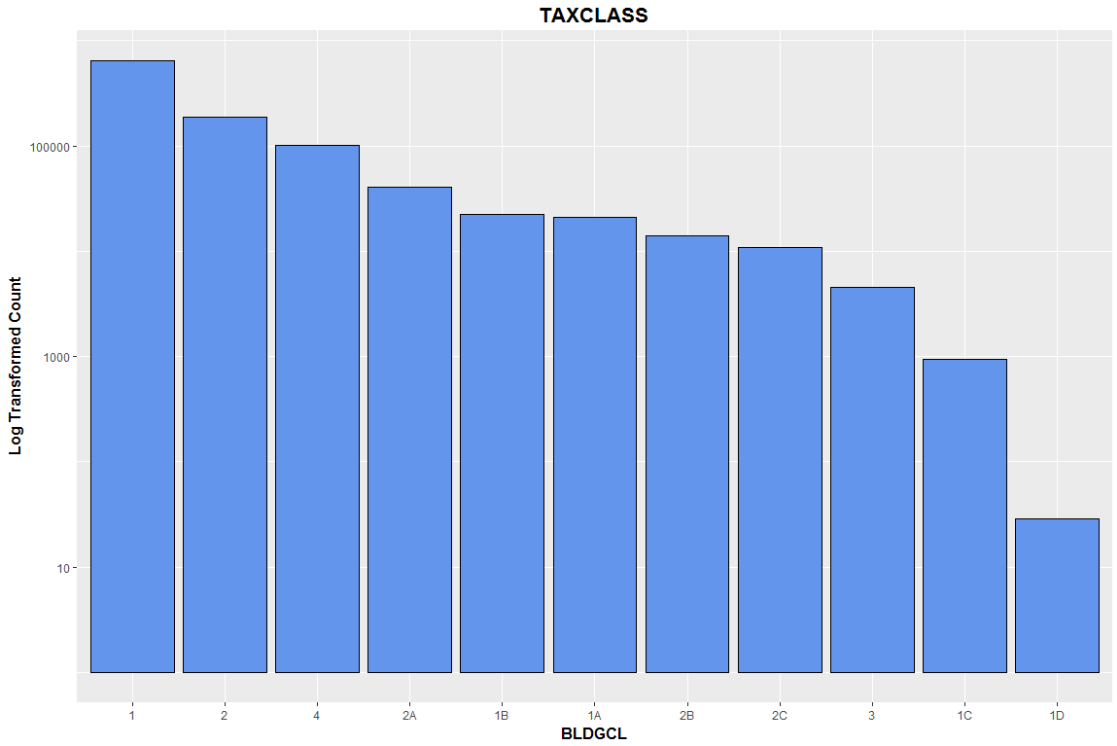
Description: categorical with metric, indicates current Property Tax Class

Number of Missing Values: 0

Number of Unique Values: has 11 unique levels: "1", "1A", "1B", "1C", "1D", "2", "2A", "2B", "2C", "3", "4"

%populated: 100%

Histogram:



(2) LTFRONT

Description: continuous variable, indicates the length of lot frontage in feet

Number of Missing Values: 0

Number of Unique Values: 1277, ranging from 0 to 9999, including 168867 records of 0 LTFRONT

%populated: 100%

mean: 36.17

min: 0

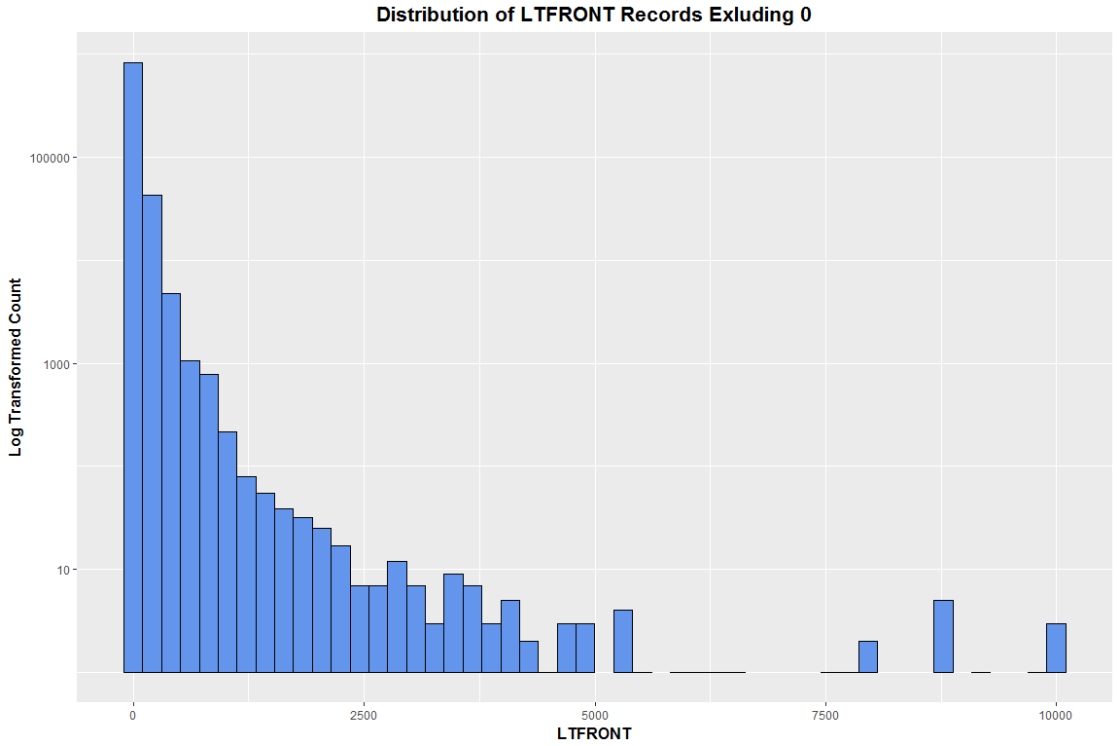
max: 9999

median: 25

mode: 20

sd: 73.73

Histogram:



(3) LTDEPTH

Description: continuous variable, indicates the length of lot depth in feet

Number of Missing Values: 0

Number of Unique Values: 1336, ranging from 0 to 9999, including 169888 records of 0 LTDEPTH

%populated: 100%

mean: 88.28

min: 0

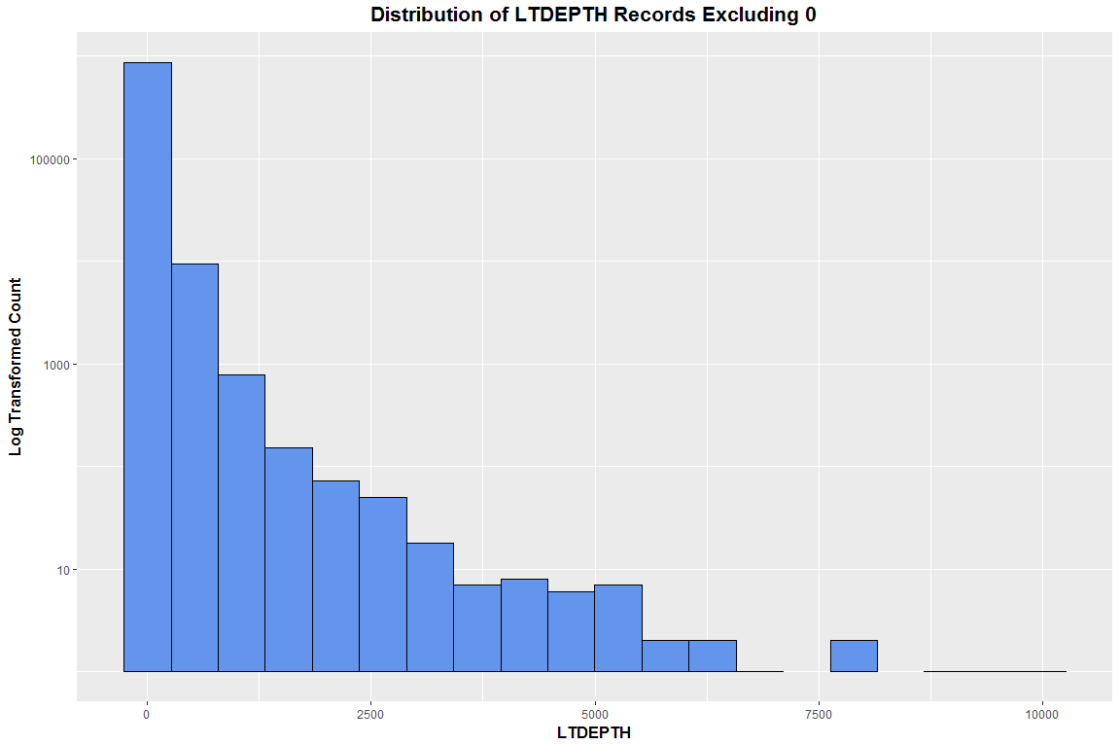
max: 9999

median: 100

mode: 100

sd: 75.48

Histogram:



(4) STORIES

Description: continuous variable, indicates the number of stories of the property

Number of Missing Values: 52142

Number of Unique Values: 112, ranging from 1 to 119

%populated: 95.03%

mean: 5.06

min: 1

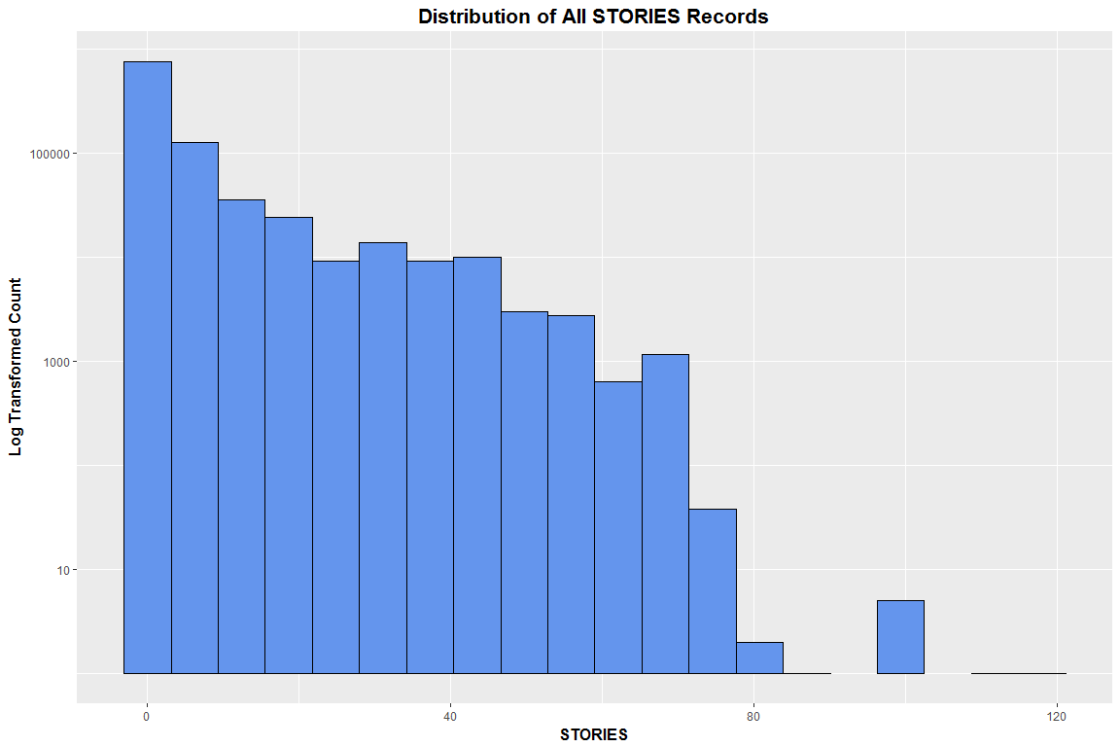
max: 119

median: 2

mode: 2

sd: 8.43

Histogram:



(5) FULLVAL

Description: continuous variable, represents the full market value of the property

Number of Missing Values: 4

Number of Unique Values: 108274, ranging from 0 to 1663775000, including 12762 records of 0 FULLVAL

%populated: 99.99%

mean: 863261.68

min: 0

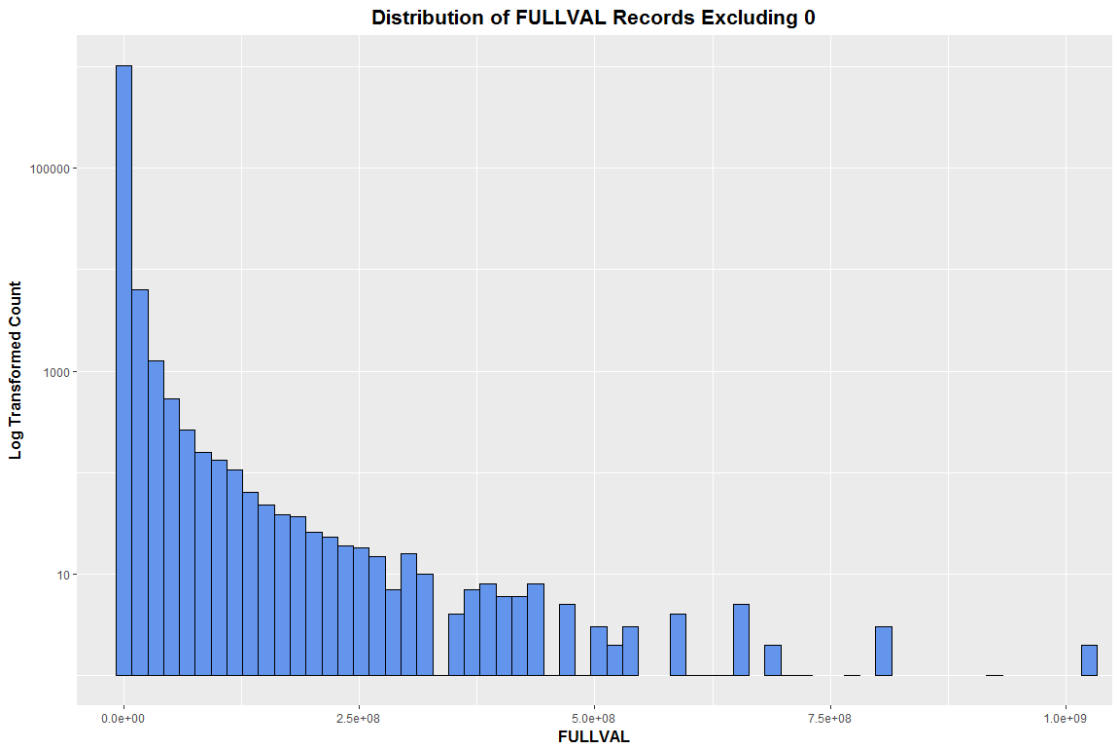
max: 1663775000

median: 446000

mode: 502000

sd: 7169292.45

Histogram:



(6) AVLAND

Description: continuous variable, represents the assessed value of the land

Number of Missing Values: 1

Number of Unique Values: 70529, ranging from 0 to 1946836665, including 12764 records of 0 AVLAND

%populated: 100%

mean: 83450.22

min: 0

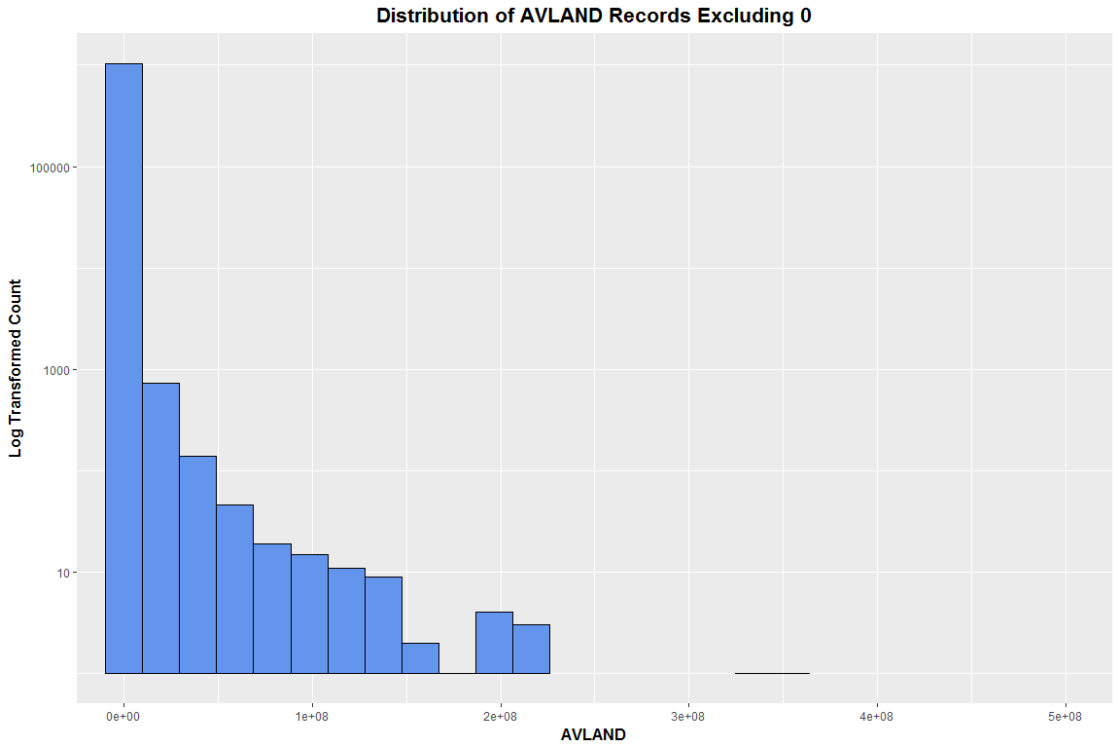
max: 1946836665

median: 13646

mode: 45000

sd: 3166324.32

Histogram:



(7) AVTOT

Description: continuous variable, represents current year's actual total market value of the property

Number of Missing Values: 3

Number of Unique Values: 112292, ranging from 0 to 1946836665, including 12762 records of 0 AVTOT

%populated: 99.99%

mean: 221401.97

min: 0

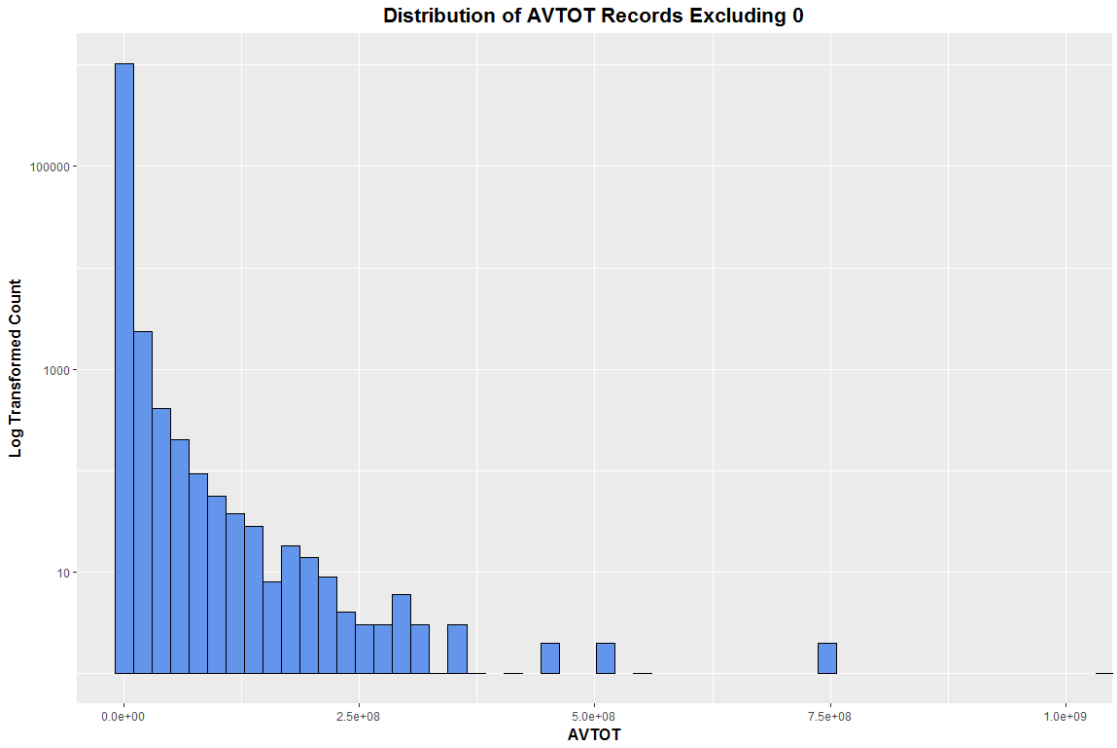
max: 1946836665

median: 25339

mode: 16588

sd: 3854074.47

Histogram:



(8) ZIP

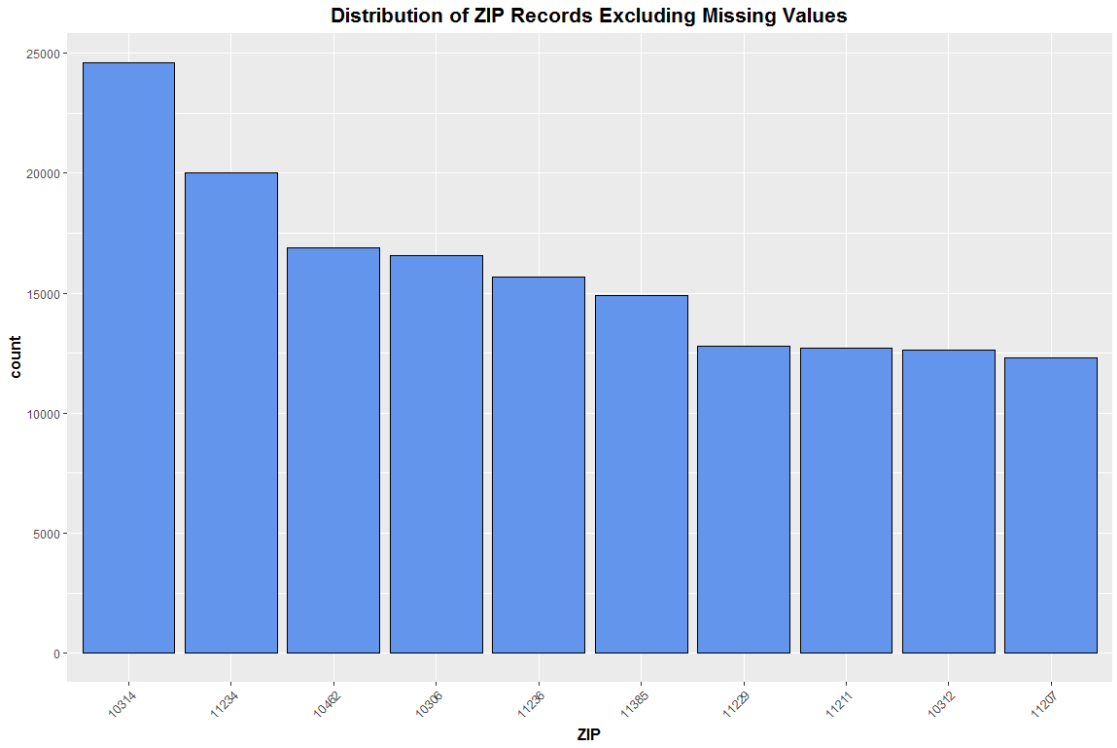
Description: categorical with metric variable, represents the zip code of the property

Number of Missing Values: 26356

Number of Unique Values: 197

%populated: 97.49%

Histogram:



(9) BLDFRONT

Description: continuous variable, represents the length of building frontage of in feet

Number of Missing Values: 0

Number of Unique Values: 610, ranging from 0 and 7575, including 224661 records of 0 BLDFRONT

%populated: 100%

mean: 23.02

min: 0

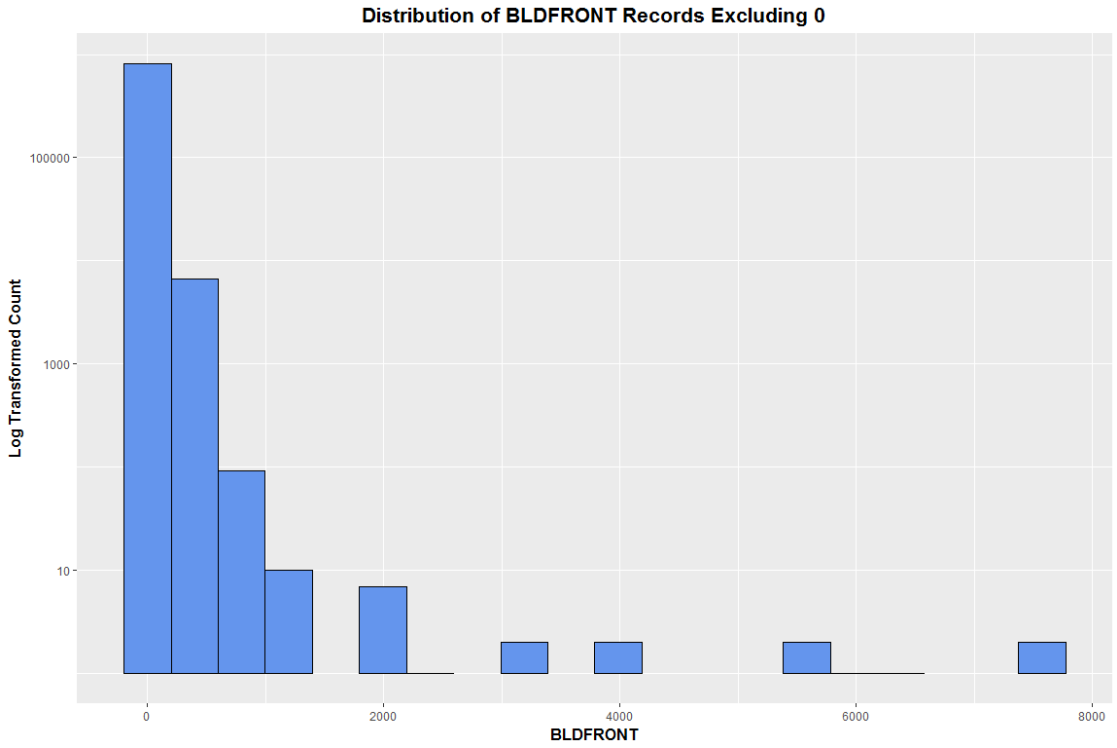
max: 7575

median: 1017

mode: 20

sd: 1388.13

Histogram:



(10) BLDDEPTH

Description: continuous variable, indicates the length of building depth in feet

Number of Missing Values: 0

Number of Unique Values: 620, ranging from 0 to 9393, including 224699 records of 0 BLDDEPTH

%populated: 100%

mean: 40.07

min: 0

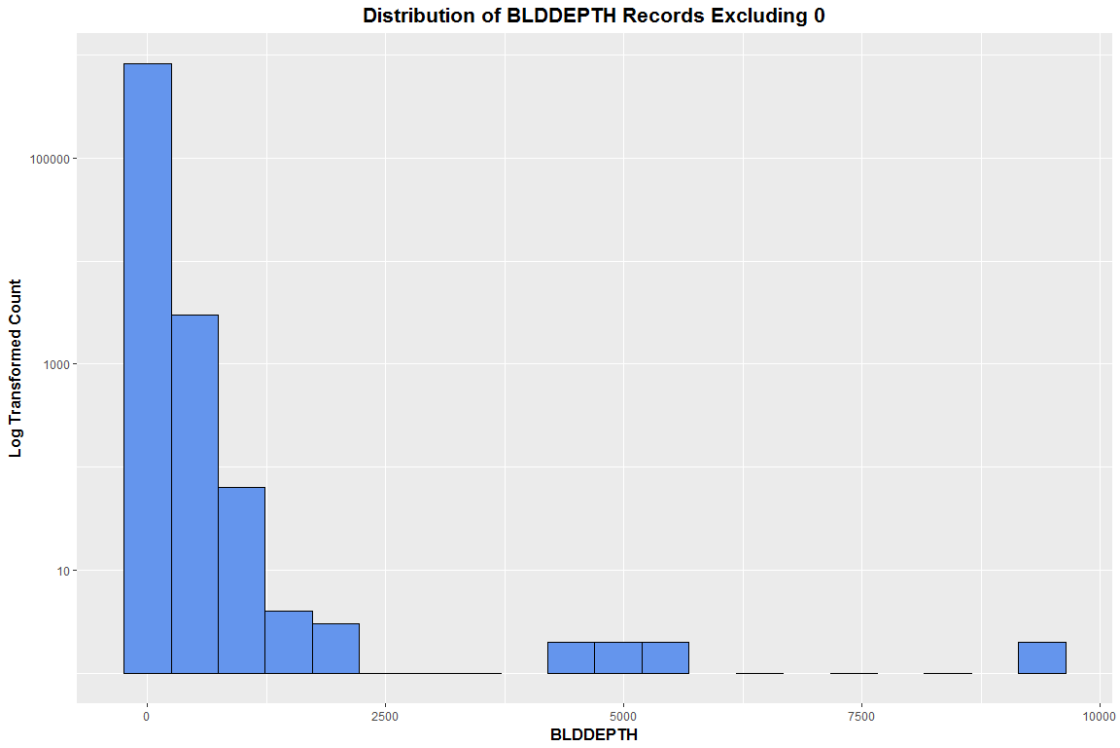
max: 9393

median: 39

mode: 40

sd: 43.04

Histogram:



Part III: Data Preparation

1. Handling Missing Values:

We used the median value for numeric variables and the most frequent value (the mode) for nominal variables. More specifically, if the continuous variables are normally distributed, the replacement of NA would be the mean of the variable values; if the continuous variables are skewed distributed, the replacement would be the median.

2. Combining Existing Variables:

We multiplied existing variables to define 3 new variables: LTAREA, FOOTPRINT, BLDVOLUME.

(1) LTAREA = LTFRONT * LTDEPTH

Describe the area of the land.

(2) FOOTPRINT = BLDFRONT * BLDDEPTH

Describe the area of the building.

(3) BLDVOLUME = LTFRONT * LTDEPTH * STORIES

Describe the volume of the building.

These three new numerical variables will be used in next part to calculate the ratios which will help us compare the FULLVAL, AVLAND and AVTOT.

Part IV: Variable Creation

The original dataset has 30 variables. To detect fraud efficiently and gain useful information, we build 15 expert variables. Following is the description of the formulas and logic for creating the model variables.

Var01 = FULLVAL / LTAREA

Average Full value of building per area.

Var02 = AVLAND / LTAREA

Average assessed value of land per area.

Var03 = AVTOT / LTAREA

Average assessed value of property per area.

Var04 = FULLVAL / BLDVOLUME

Average Full value of building per volume.

Var05 = AVLAND / BLDVOLUME

Average assessed value of land per volume.

Var06 = AVTOT / BLDVOLUME

Average assessed value of property per volume.

Var07 = FULLVAL / FOOTPRINT

Average Full value of building per footprint.

Var08 = AVLAND / FOOTPRINT

Average assessed value of land per footprint.

Var09 = AVTOT / FOOTPRINT

Average assessed value of property per footprint.

Var10-12: < VAL > i / < VAL > j, i belongs to j tax class

Ratio of FULLVAL and average FULLVAL grouped by tax classes.

Ratio of AVLAND and average AVLAND grouped by tax classes.

Ratio of AVTOT and average AVTOT grouped by tax classes.

Var13-15: < VAL > i / < VAL > j, i belongs to j zip bin

Ratio of FULLVAL and average FULLVAL grouped by zip bins.

Ratio of AVLAND and average AVLAND grouped by zip bins.

Ratio of AVTOT and average AVTOT grouped by zip bins.

Part V: Algorithm

Following the creation of expert variables that quantify signals for various fraud modes, we used several steps to get the fraud score.

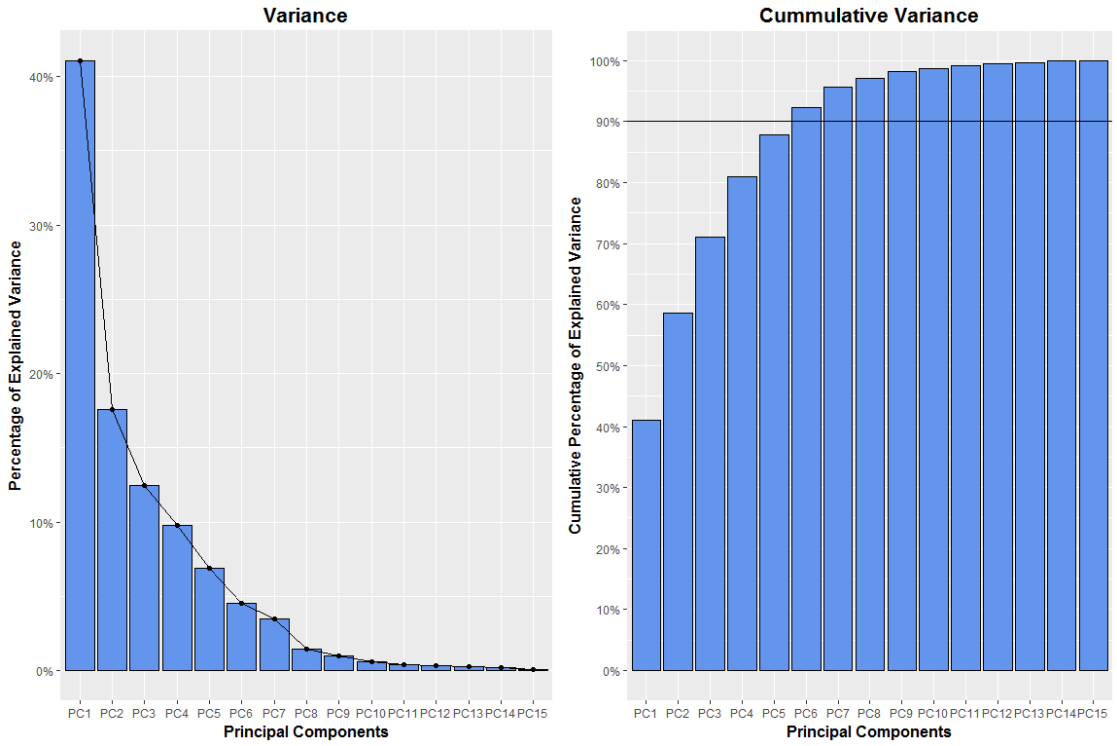
Step1: Z-scale to prepare for feature selection/dimensionality reduction

We scaled the 15 expert variables to make sure different scale do not affect their contribution to the PCA (by using the argument "scale=TRUE" in prcomp() function).

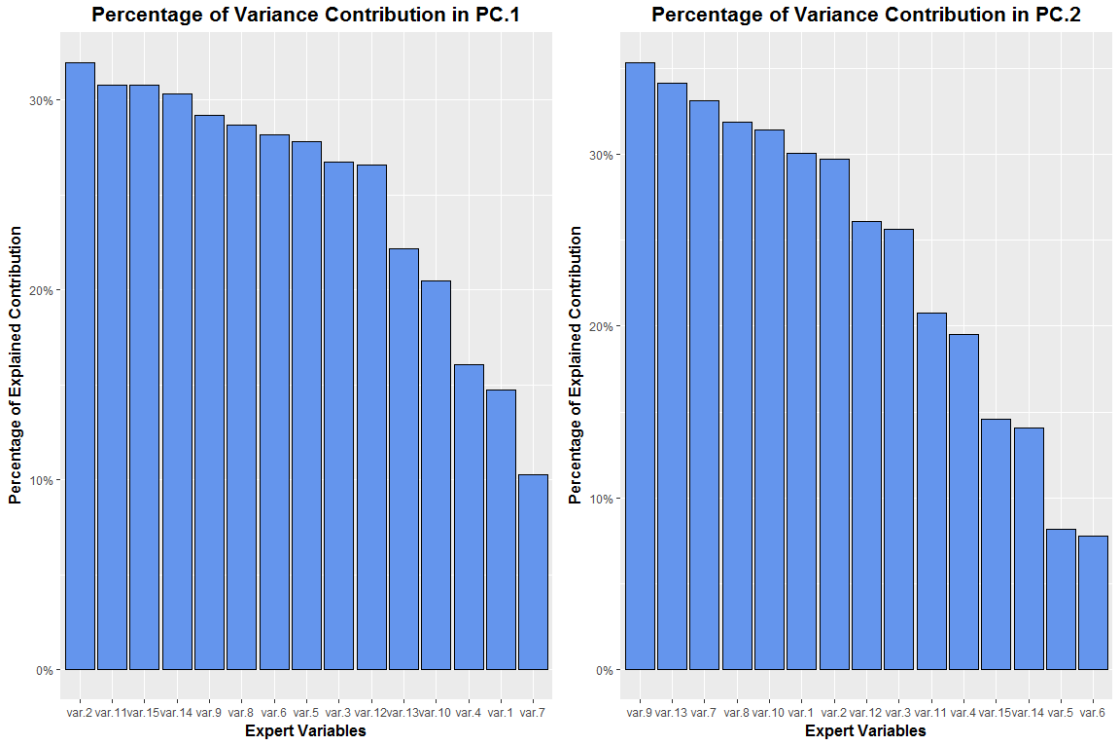
Step2: Reduce dimensions via PCA

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation). Commonly, PCA is used to explain the variance-covariance structure of a set of variables through linear combinations. It is often used as a dimensionality-reduction technique.

We performed PCA with prcomp() function in R to get the rotation matrix between 15 principal components and the z-scaled expert variables, as well as a matrix indicative of the linear relationships of each record and 15 principal components. Then we keep the first 6 PCs which account for more than 90% variance of the whole dataset to represent each record in this new reduced space of the PCs.



The left plot above shows us how much percentage of variance in the data is explained by each PC. Each principal component explains diminishing percentage of the variance. The right plot shows PC1 to PC6, as a whole, explains more than 90% of variance within the new reduced space. So, we choose those 6 PCs to perform our fraud score calculation.



The left graph indicates the contribution of 15 expert variables to PC1. The right graph indicates the contribution of 15 expert variables to PC2.

Step3: Z-scale the data fields again

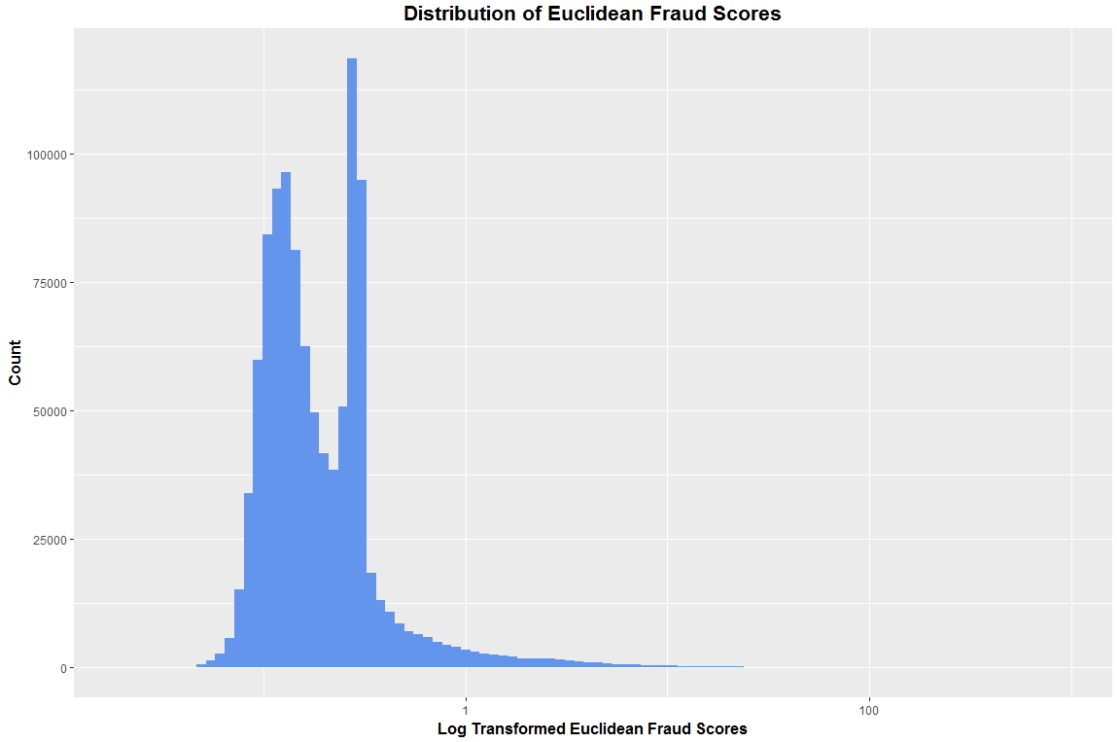
We used scale (..., center=TRUE, scale=TRUE) function to z-scale the PCs we generated.

Step4: Calculate Fraud Scores

Then we decided to use two methods to calculate fraud scores. The first one is combining the z-scaled PCs with a heuristic algorithm. The second one is reconstructing the z scaled PCs through an autoencoder.

1. Heuristic Algorithm

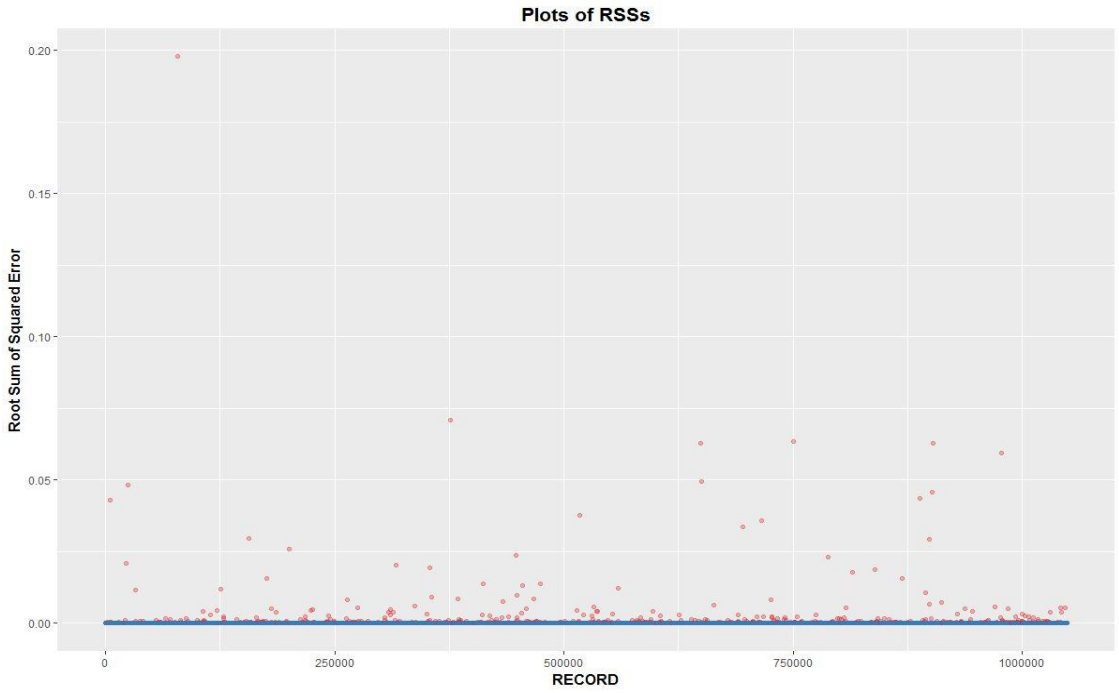
Euclidean distance is the distance between two points in Euclidean space. For each record, we squared the values for each of the principal components, and summed them up. Next, we rooted the sum of the 6 squared PC values for each row to get the fraud score, which is the Euclidean distance. The distribution of fraud scores is shown below:

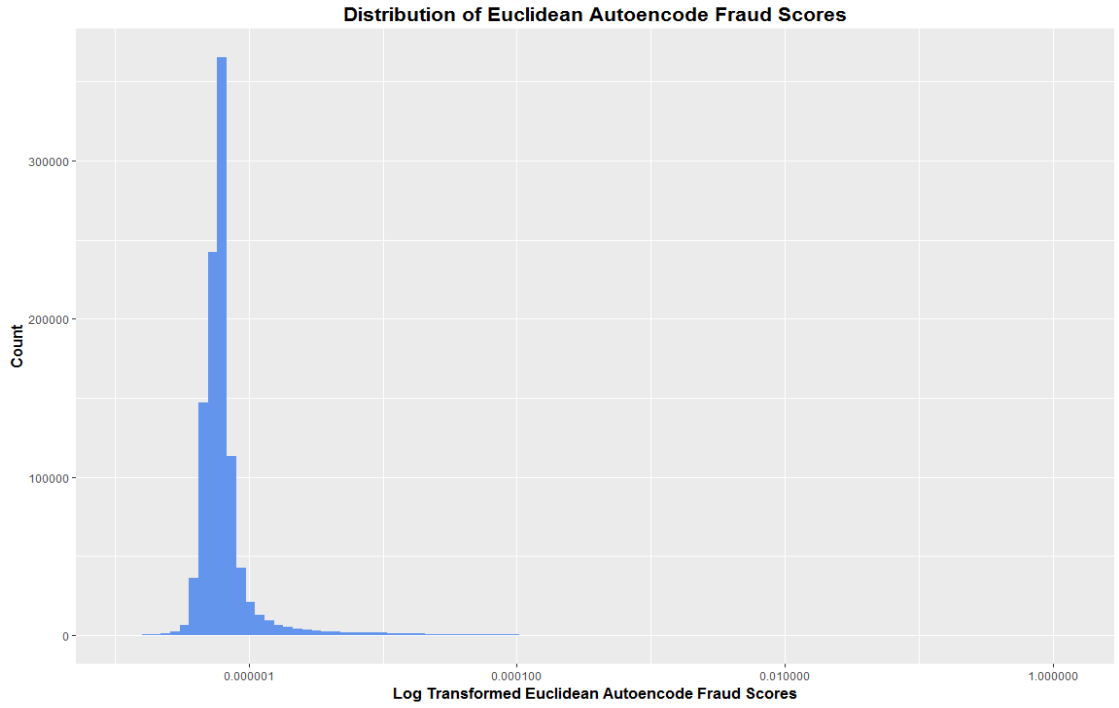


In the graph above, we can see the distribution of Euclidean distance is a long-tailed graph which depicts how the number of log-transformed fraud scores distributed across the value range. although with two peaks, more than 90% of the records have low fraud scores (log fraud scores < 1). As a result, we might have to suspect the observations with high fraud scores.

2. Autoencoder

We trained autoencoder on a 6 * 1,048,575 dataset to reproduce the z scaled PC records. We used 2 hidden layers (10 nodes for each layer); each hidden layer has a length of 10. We used package "h2o" in R to implement this method. We define the fraud score as the distance between the original input record and the autoencoder output record with the form of Euclidean distance. The distribution of the fraud scores is shown below:





Part VI: Results

Overview:

Comparing the top 0.1% (1048 records) highest fraud scores in PCA Euclidean Distance and Autoencoder, we found there is approximately 92.94% (974/1048) overlapping between these two methods. So, we selected these overlapped records to do the statistical analysis.

The following table shows the statistical information for top 0.1% overlapped fraud scores and full dataset:

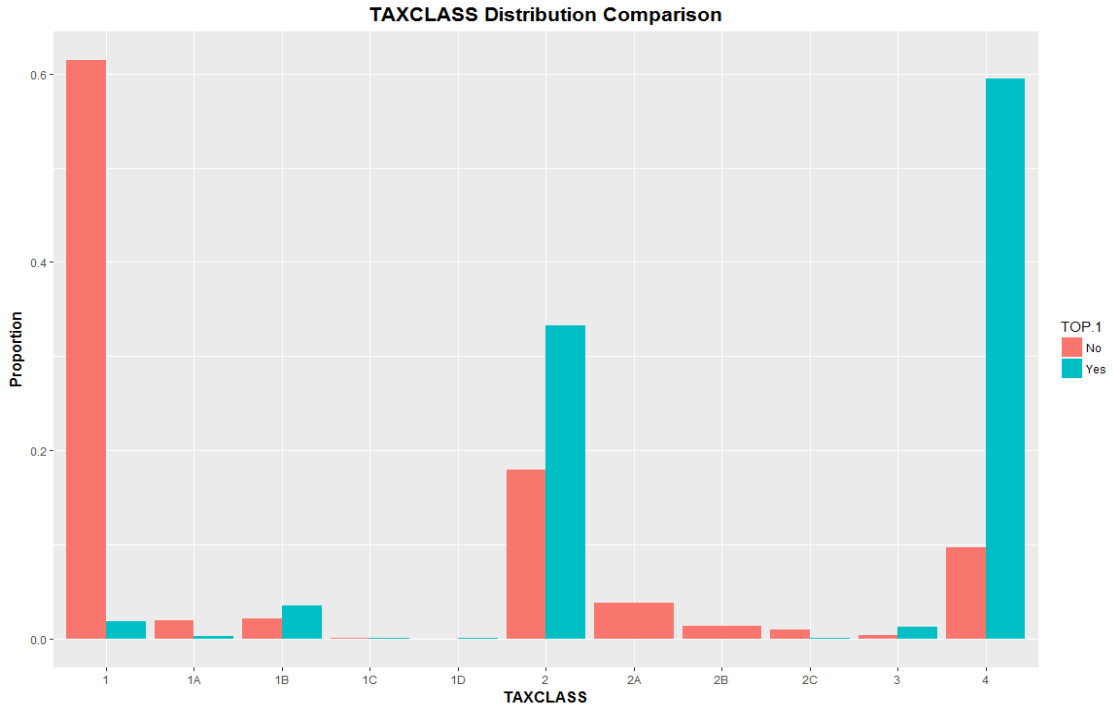
	range	sum	median	mean	std.dev
LTFRONT	9998	4.8e+05	200.0	499.16116	952.79954
LTDEPTH	9998	4.7e+05	204.0	484.71591	818.51157
STORIES	84	1.9e+04	17.0	19.34401	17.12555
FULLVAL	1663740000	1.3e+11	72236005.0	133722899.59814	168244050.55155
AVLAND	1946836505	2.6e+10	9000000.0	26696187.45764	99250663.64258
AVTOT	1946836505	6.0e+10	31570650.0	61649156.74587	101925566.79957
EXLAND	1946835446	1.7e+10	581365.5	17631840.73244	98785392.32288
EXTOT	1946835045	3.2e+10	9267658.0	33036812.23037	94408812.78742
LTAREA	99980000	7.6e+08	44880.0	786778.40496	4722160.50709
FOOTPRINT	2755325	3.4e+07	12712.5	35287.38843	107004.61936
BLDVOLUME	124061759	2.6e+09	767312.5	2702324.44835	8421482.99227

	range	sum	median	mean	std.dev
LTFRONT	9998	4.2e+07	25	40.200355	72.31734
LTDEPTH	9998	1.1e+08	100	104.478227	64.76314
STORIES	118	5.1e+06	2	4.911033	8.24600
FULLVAL	1663774996	9.1e+11	450000	868736.960091	7168789.12693
AVLAND	1946836664	8.8e+10	13751	83617.544800	3166318.75645
AVTOT	1946836664	2.3e+11	25560	221712.492073	3854052.11891
EXLAND	1946836664	3.7e+10	1620	35015.011627	3066648.71598
EXTOT	1946836664	8.8e+10	1620	83845.118101	3131401.34142
LTAREA	99980000	6.6e+09	2500	6313.654055	154716.49656
FOOTPRINT	71151974	2.1e+09	880	2036.254858	99621.22153
BLDVOLUME	124061759	4.0e+10	6300	38133.638591	378403.25268

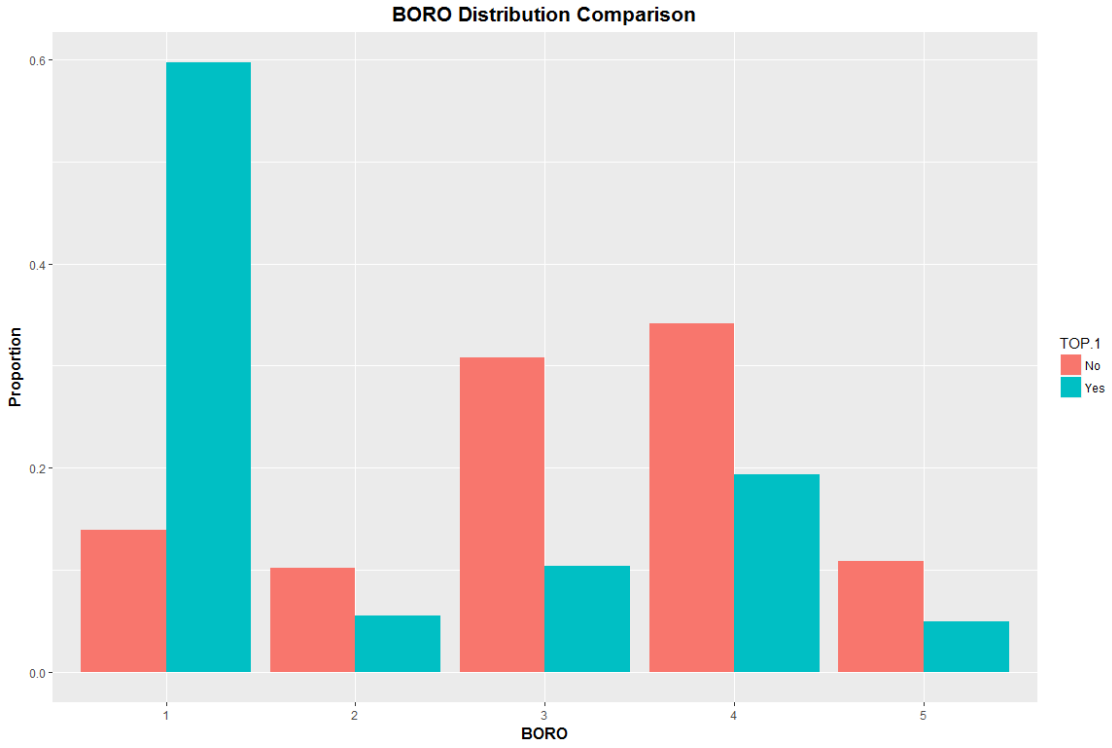
Insights:

(1) We find that the top 0.1% has higher mean, median, std value of FULLVAL, LTAREA, BLDVOLUME and FOOTPRINT compared to full dataset, which means the potential fraud properties are usually the big buildings and higher value buildings of the cities.

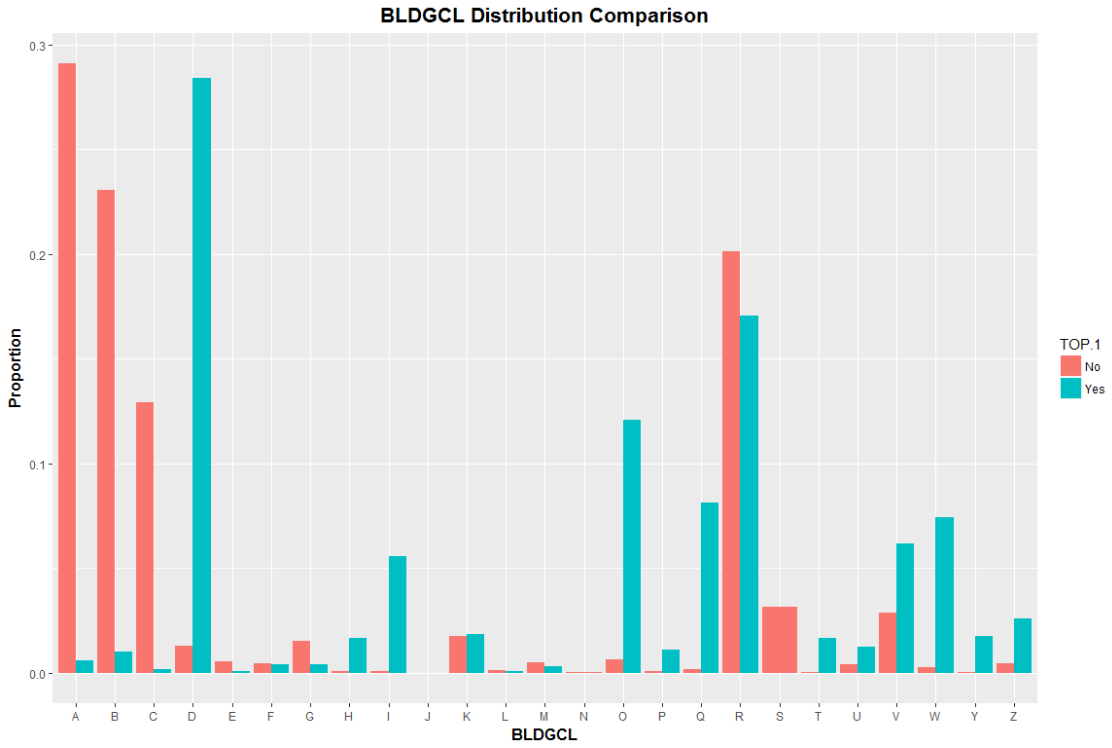
(2) As for the tax class distribution, top 0.1% has 64% tax class 4 properties, while in full dataset, only 10% properties in tax class 4. This may be caused by the fact that tax class 4 has the lowest rates. Tax class 1 rate is 19.991%, Tax class 2 rate is 12.892% and Tax class 3 rate is 10.934%. The following graph is TAXCLASS distribution comparison:



(3) The following graph is the distribution of BORO for the top 0.1% and full dataset. For top 0.1%, over 60% of the properties are in BORO 1, which is Manhattan, but in full dataset, only 14% properties in BORO 1. Manhattan has higher house and land price which may be the reason for this distribution. Also, BORO 2 is Bronx, BORO3 is Brooklyn, BORO4 is Queens and BORO5 is Staten island.



(4) The following graph is BLDGCL distribution. The top 0.1% has higher proportion of BLDGCL D. This may be caused by the fact that BLDGCL D are elevator apartments which have higher values.



Further Exploration:

In the PCA Euclidean Distance method, the top 10 fraud records are 78804, 648675, 902256, 977471, 750447, 376243, 24586, 5393, 901790, 888450.

In the Autoencoder Euclidean Distance method, the top 10 fraud scores are 78804, 376243, 750447, 977471, 648675, 902256, 650467, 24586, 901790, 888450.

We had 90% overlapping in the top 10 fraud record. So, we selected top 10 matching records as best candidates for potential fraud.

Top 10 records:

RECORD	FULLVAL	AVLAND	AVTOT
78804	450000	1946836665	1946836665
648675	450000	13751	25560
902256	450000	13751	25560
977471	3443400	1549530	1549530
750447	251989	1001	8934
376243	374019883	1792808947	25560
24586	3712000	252000	1670400
5393	2930000	1318500	1318500
901790	540143500	32408610	32408610
888450	1662400000	748080000	748080000

(1) Record: 78804

Its AVLAND equals EXLAND, and AVTOT equals EXTOT, which means it is tax exempted. Owners can avoid tax by getting tax exemption illegal.

(2) Record: 648675

It has an abnormally low value 25560 in AVTOT while the corresponding mean value of the same zip area is 248437.Also, the AVLAND and AVTOT is abnormally low, as the FULLVAL is 45000. This may be caused by wrong input.

(3) Record: 902256

It has an abnormally low value 25560 in AVTOT while the corresponding mean value of the same zip area is 67713.8. Also, the AVLAND and AVTOT is abnormally low, as the FULLVAL is 45000. This may be caused by wrong input.

(4) Record: 977471

It has an abnormally high value 3443400 in FULLVAL,1549530 in AVLAND and 1549530 in AVTOT, while the corresponding mean values in zip 11101 area are 935539 in

FULLVAL,101610 in AVLAND and 363798 in AVTOT. Owners may report higher value to get higher loan from the financial institutions.

(5) Record: 750447

It has an abnormally low value in FULLVAL, AVLAND and AVTOT while the corresponding mean value of FULLVAL, AVLAND and AVTOT in the same zip area is 711749,52231 and 87013. Also, the AVLAND and AVTOT is abnormally low, as the FULLVAL is 251989. This may be caused by wrong input.

(6) Record: 376243

It has an abnormally high value 374019883 in FULLVAL,1792808947 in AVLAND and 25560 in AVTOT, while the corresponding mean values in zip 11101 area are 496650 in FULLVAL,289090 in AVLAND and 41662 in AVTOT. Owners may report higher value to get higher loan from the financial institutions.

(7) Record: 24586

It has an abnormally high value 3712000 in FULLVAL,252000 in AVLAND and 1670400 in AVTOT, while the corresponding mean values in zip 11101 areas are 935539 in FULLVAL,101610 in AVLAND and 363798 in AVTOT. Owners may report higher value to get higher loan from the financial institutions.

(8) Record: 5393

It has an abnormally high value 2930000 in FULLVAL,1318500 in AVLAND and 1318500 in AVTOT, while the corresponding mean values in zip 11101 areas are 705678 in FULLVAL,33563 in AVLAND and 141866 in AVTOT. Owners may report higher value to get higher loan from the financial institutions.

(9) Record: 901790

Its AVLAND equals EXLAND, and AVTOT equals EXTOT, which means it is tax exempted. Owners can avoid tax by getting tax exemption illegal.

(10) Record: 888450

Its AVLAND equals EXLAND, and AVTOT equals EXTOT, which means it is tax exempted. Owners can avoid tax by getting tax exemption illegal.

These 10 fraud records can be classified into 3 types:

1.Falsely report land property value to get higher loan;

2.Tax avoidance. Owners can avoid tax by getting tax exemption illegal;

3.Incorrect data input.

Our analysis has some limitations. Later if we can have more comprehensive understanding of New York property legislations, maybe we can detect more useful and accurate fraud records.

Appendix: DQR

(1) RECORD

Description: categorical no metric variable for unique identification of records

Number of Missing Values: 0

Number of Unique Value: 1048575, ranging from 1 to 1048575

%populated: 100%

(2) BBLE

Description: categorical no metric variable for unique identification of properties, concatenation of BORO code, BLOCK code, EASEMENT code and LOT code

Number of Missing Values: 0

Number of Unique Values: 1048575, ranging from 1 to 1048575

%populated: 100%

(3) BLOCK

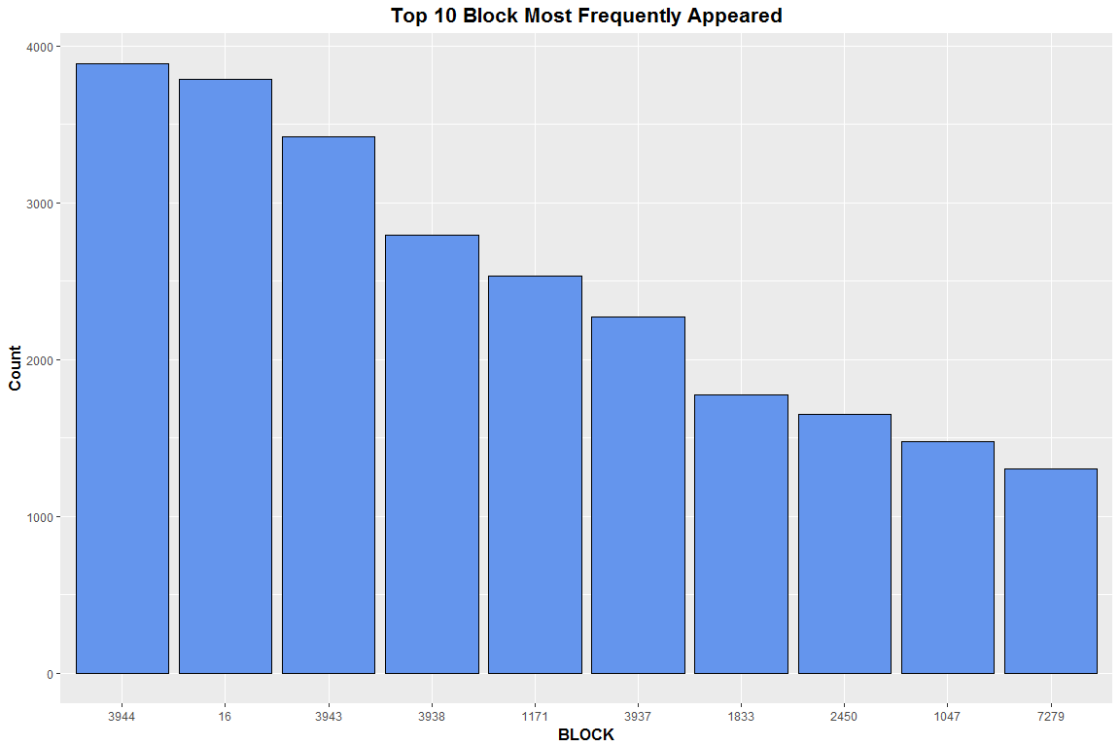
Description: categorical with metric variable, valid block ranges by BORO

Number of Missing Values: 0

Number of Unique Values: 13949, ranging from 1 to 16350

%populated: 100%

Histogram:



(4) LOT

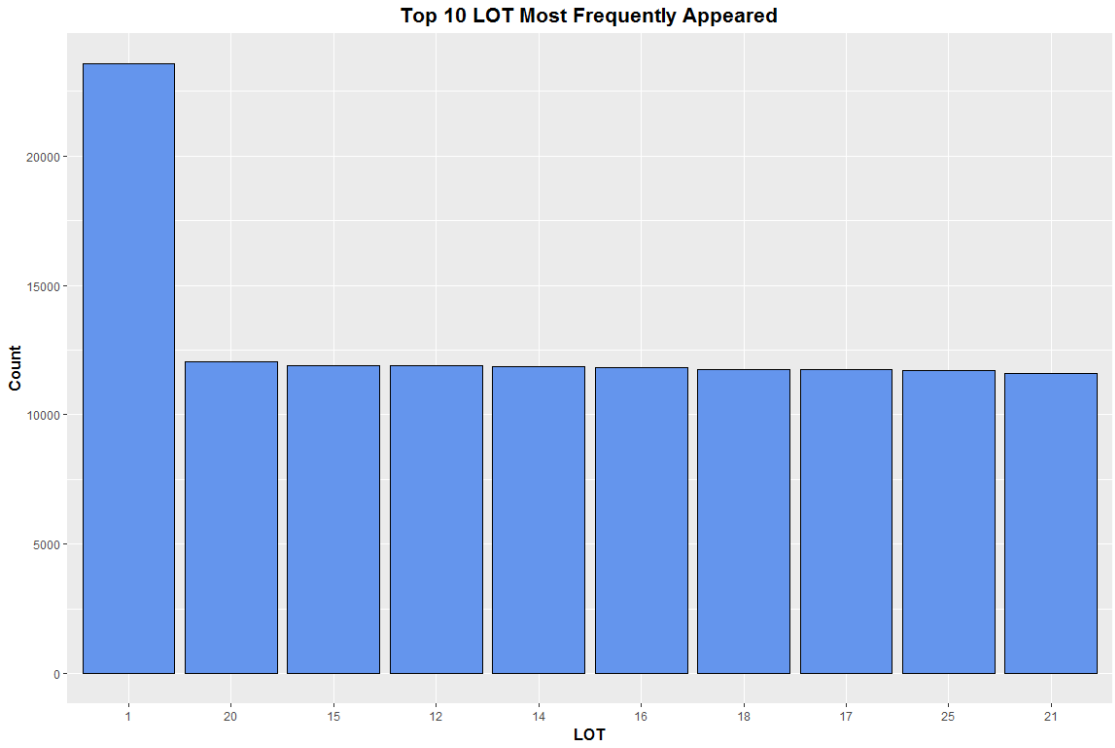
Description: categorical no metric variable, Unique number within BLOCK or BORO

Number of Missing Values: 0

Number of Unique Values: 6366, ranging from 1 to 9978

%populated: 100%

Histogram:



(5) EASEMENT

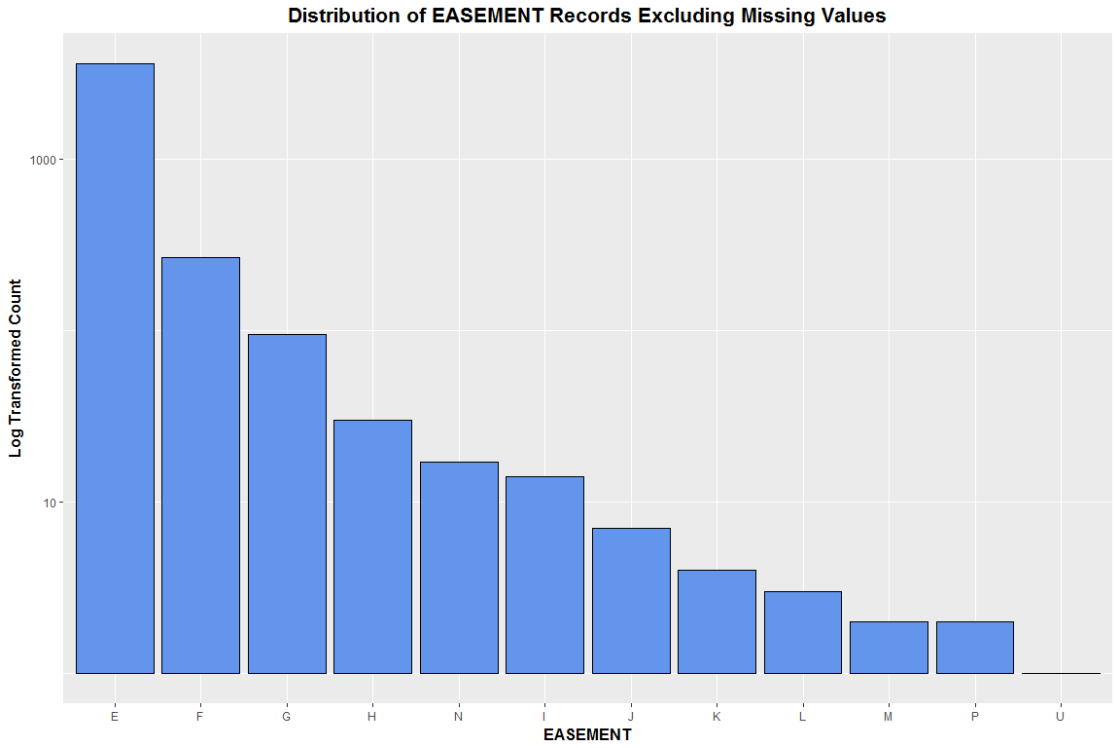
Description: categorical no metric, is a field that is used to describe easement

Number of Missing Values: 1044532

Number of Unique Values: EASEMENT has 13 levels:"","E","F","G","H","I","J","K","L","M","N","P","U"

%populated: 38.56%

Histogram:



(6) OWNER

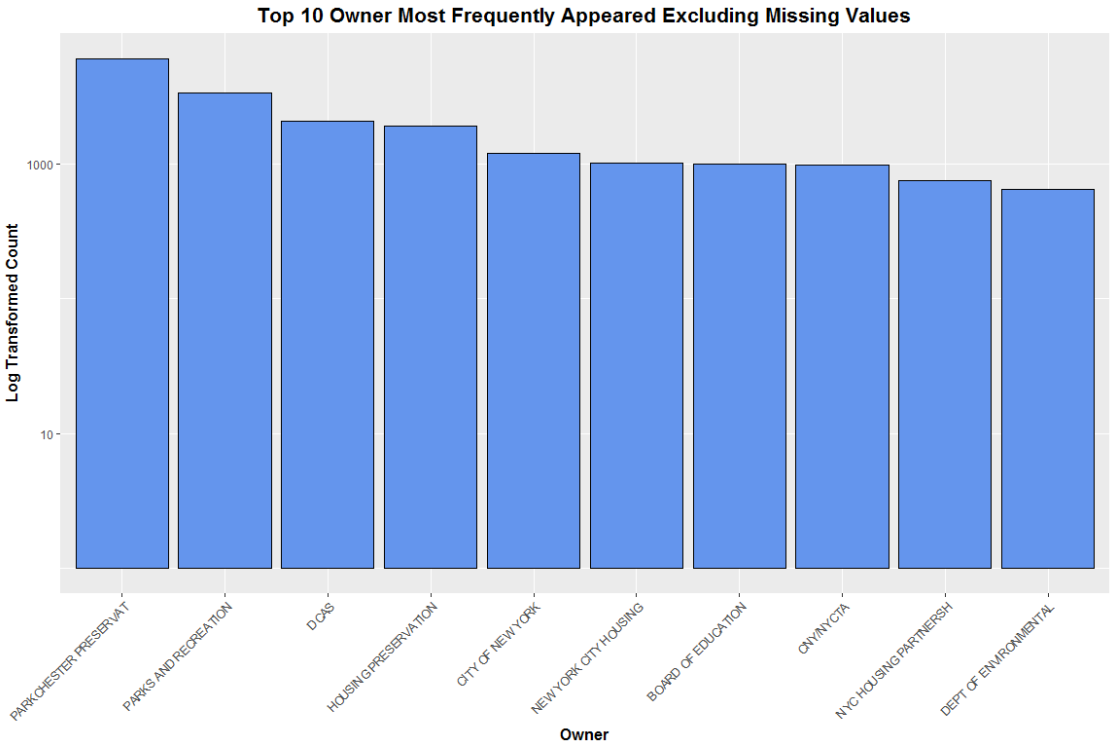
Description: categorical no metric, used to indicate the Owner's Name

Number of Missing Values: 31081

Number of Unique Values: 847055

%populated: 97.04%

The top 10 most frequently occurred OWNER names are:



(7) BLDGCL

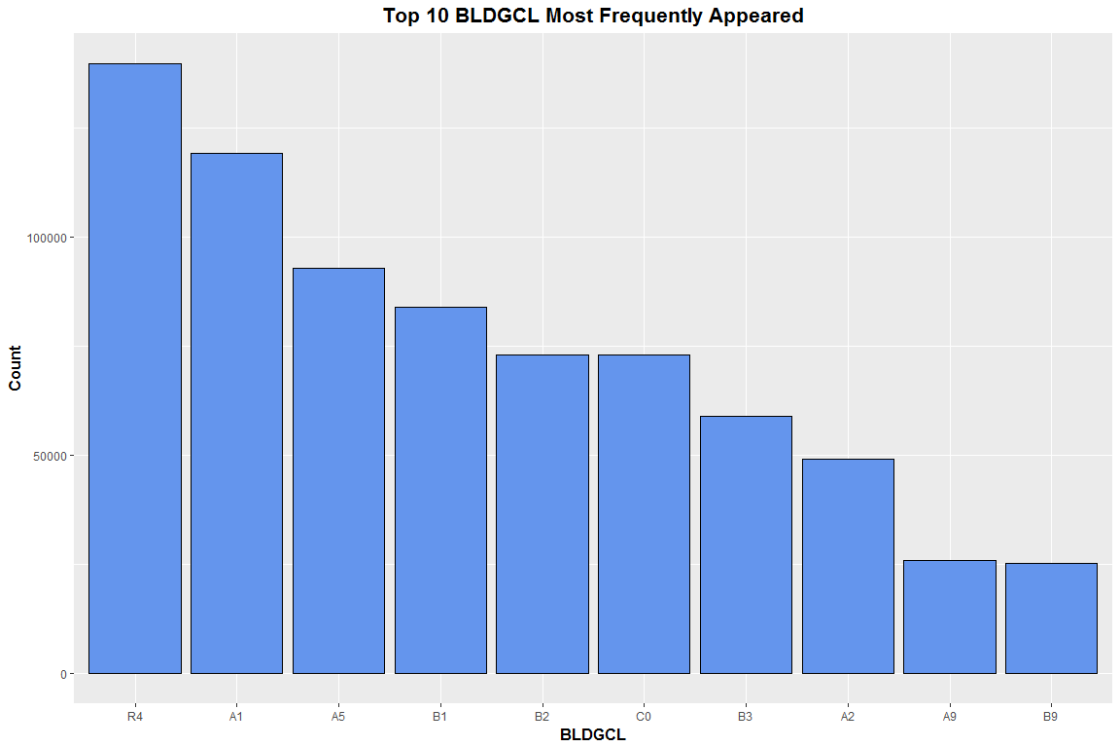
Description: categorical with metric, indicates building class

Number of Missing Values: 0

Number of Unique Values: 200

%populated: 100%

Histogram:



(8) TAXCLASS

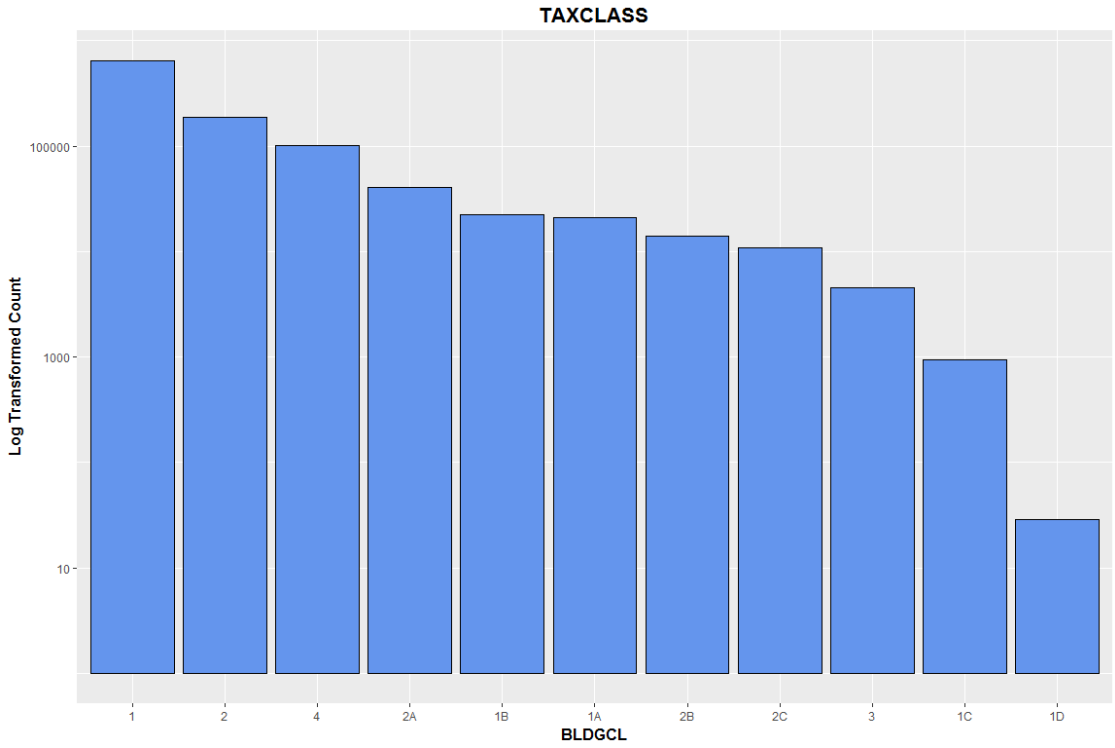
Description: categorical with metric, indicates current Property Tax Class

Number of Missing Values: 0

Number of Unique Values: has 11 unique levels:
"1","1A","1B","1C","1D","2","2A","2B","2C","3","4"

%populated: 100%

Histogram:



(9) LTFRONT

Description: continuous variable, indicates the length of lot frontage in feet

Number of Missing Values: 0

Number of Unique Values: 1277, ranging from 0 to 9999, including 168867 records of 0 LTFRONT

%populated: 100%

mean: 36.17

min: 0

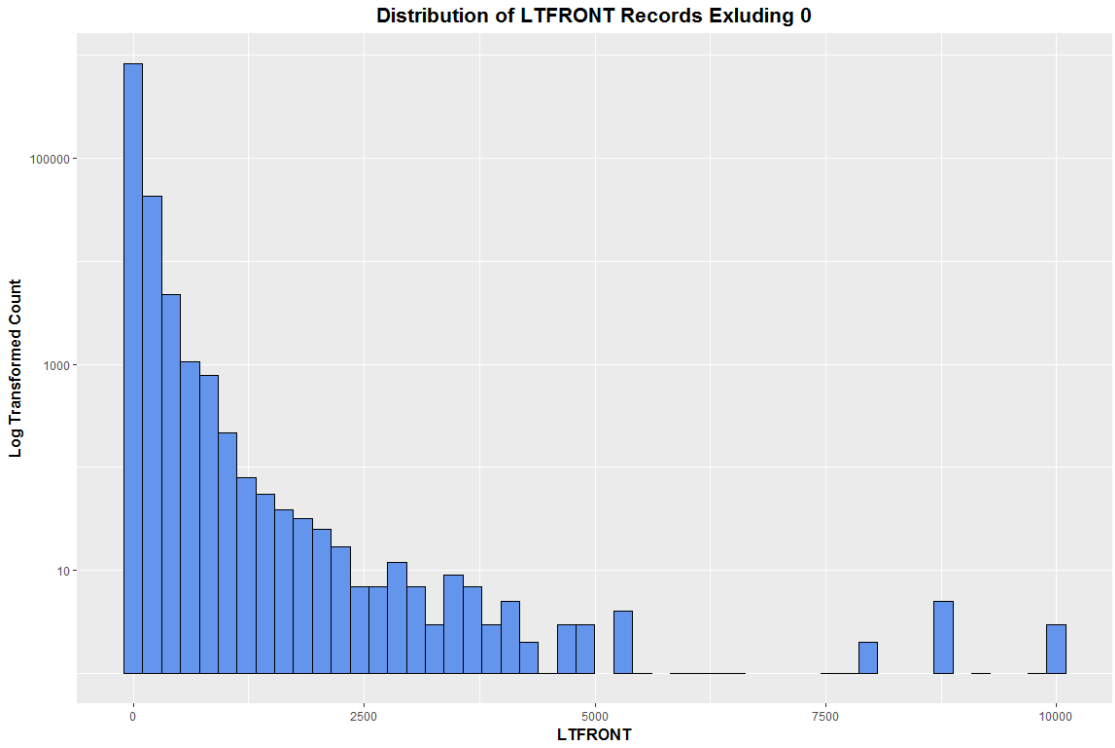
max: 9999

median: 25

mode: 20

sd: 73.73

Histogram:



(10) LTDEPTH

Description: continuous variable, indicates the length of lot depth in feet

Number of Missing Values: 0

Number of Unique Values: 1336, ranging from 0 to 9999, including 169888 records of 0 LTDEPTH

%populated: 100%

mean: 88.28

min: 0

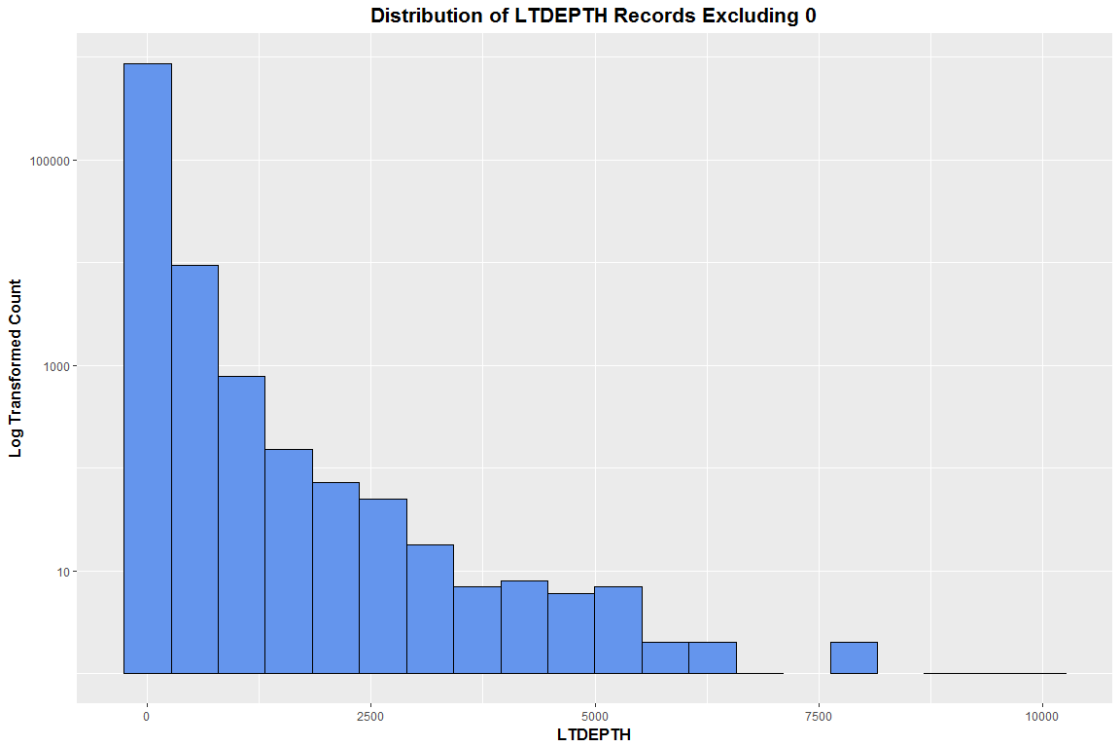
max: 9999

median: 100

mode: 100

sd: 75.48

Histogram:



(11) STORIES

Description: continuous variable, indicates the number of stories of the property

Number of Missing Values: 52142

Number of Unique Values: 112, ranging from 1 to 119

%populated: 95.03%

mean: 5.06

min: 1

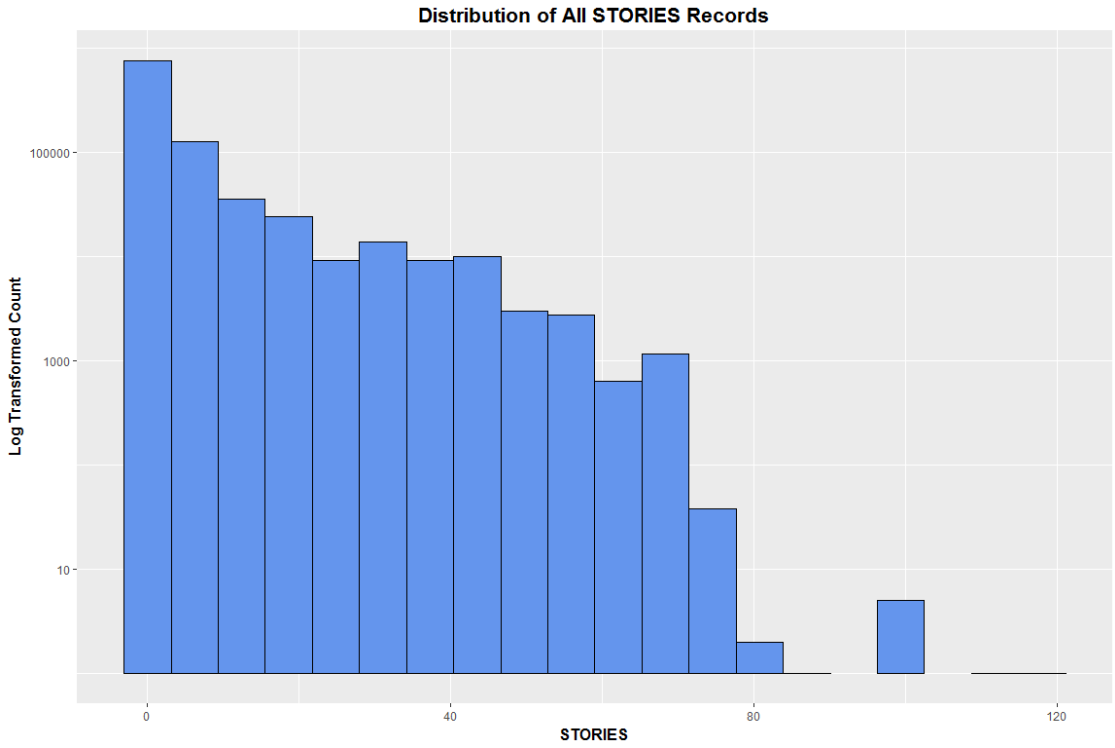
max: 119

median: 2

mode: 2

sd: 8.43

Histogram:



(12) FULLVAL

Description: continuous variable, represents the full market value of the property

Number of Missing Values: 4

Number of Unique Values: 108274, ranging from 0 to 1663775000, including 12762 records of 0 FULLVAL

%populated: 99.99%

mean: 863261.68

min: 0

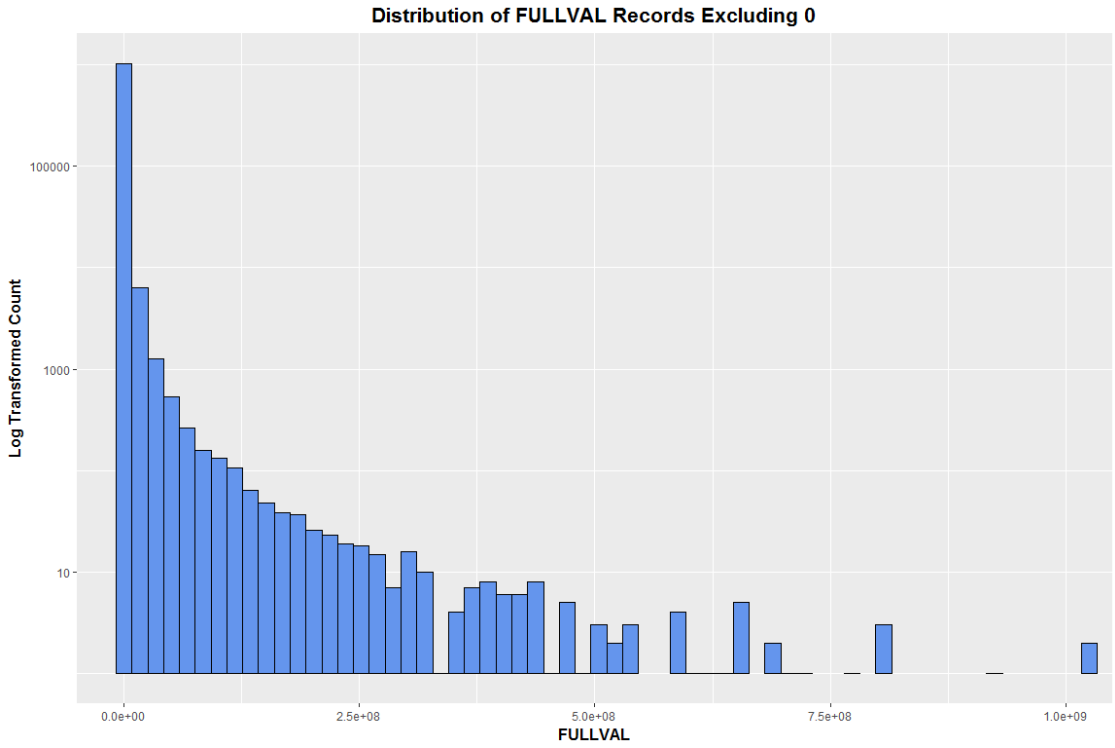
max: 1663775000

median: 446000

mode: 502000

sd: 7169292.45

Histogram:



(13) AVLAND

Description: continuous variable, represents the assessed value of the land

Number of Missing Values: 1

Number of Unique Values: 70529, ranging from 0 to 1946836665, including 12764 records of 0 AVLAND

%populated: 100%

mean: 83450.22

min: 0

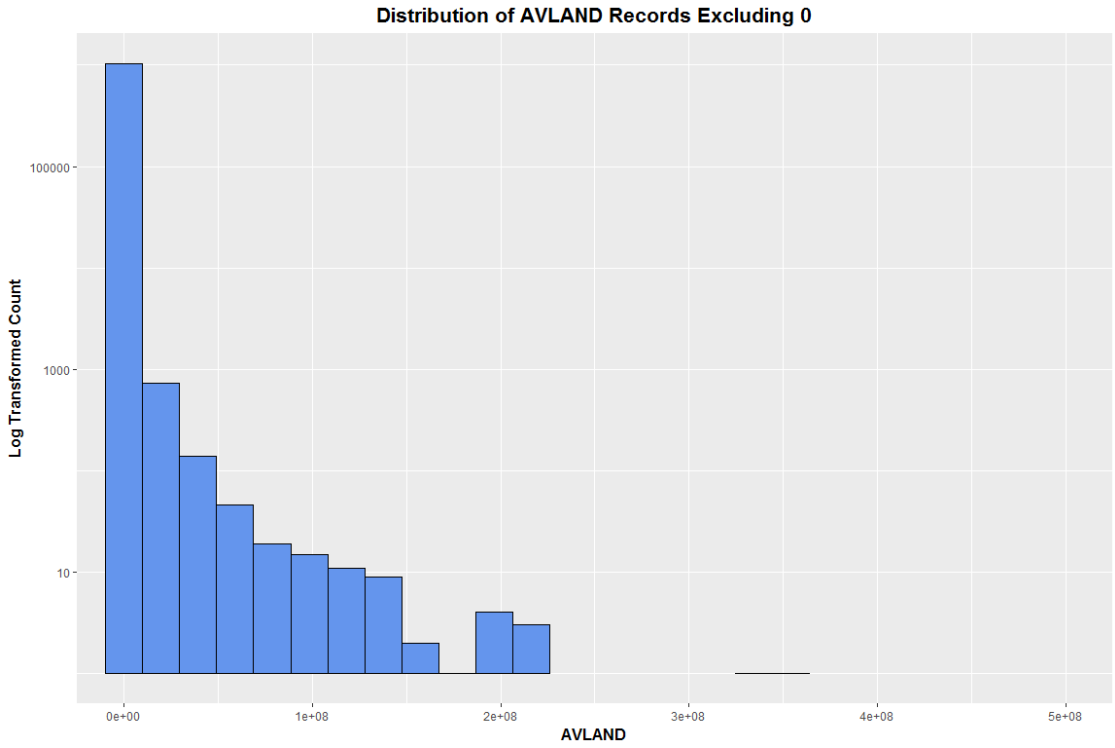
max: 1946836665

median: 13646

mode: 45000

sd: 3166324.32

Histogram:



(14) AVTOT

Description: continuous variable, represents current year's actual total market value of the property

Number of Missing Values: 3

Number of Unique Values: 112292, ranging from 0 to 1946836665,including 12762 records of 0 AVTOT

%populated: 99.99%

mean: 221401.97

min: 0

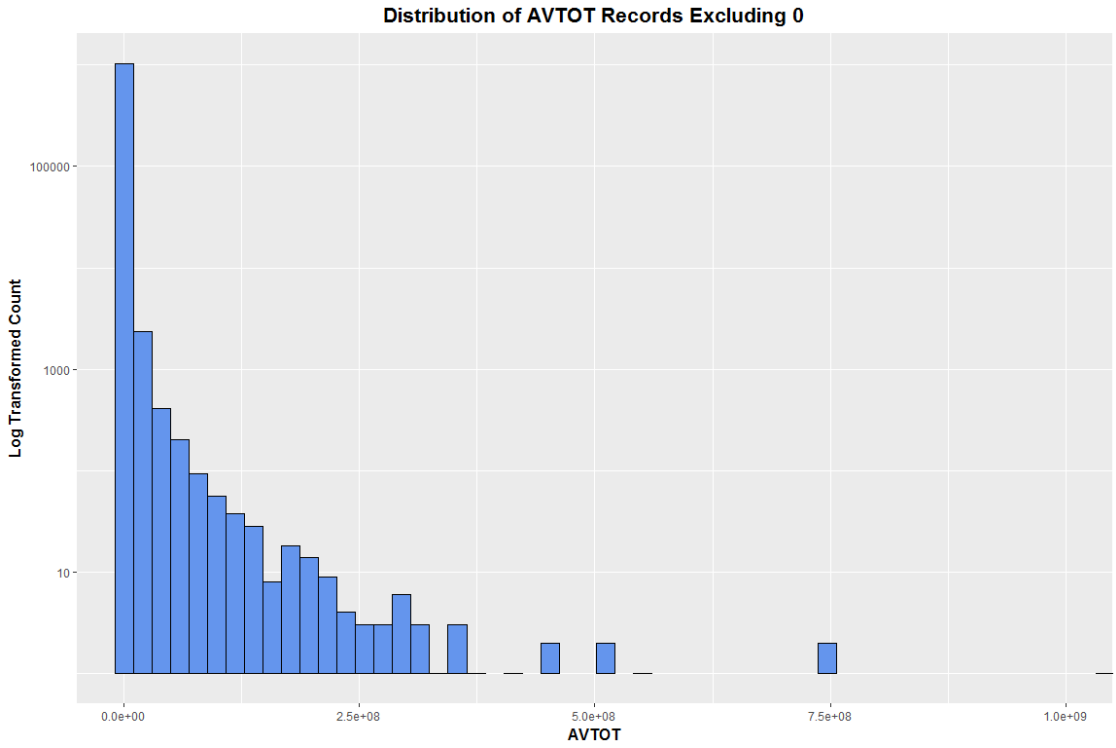
max: 1946836665

median: 25339

mode: 16588

sd: 3854074.47

Histogram:



(15) EXLAND

Description: continuous variable, represents the actual value of the exempt land

Number of Missing Values: 1

Number of Unique Values: 33186, ranging from 0 to 1946836665,including 484224 records of 0 EXLAND

%populated: 100%

mean: 34266.94

min: 0

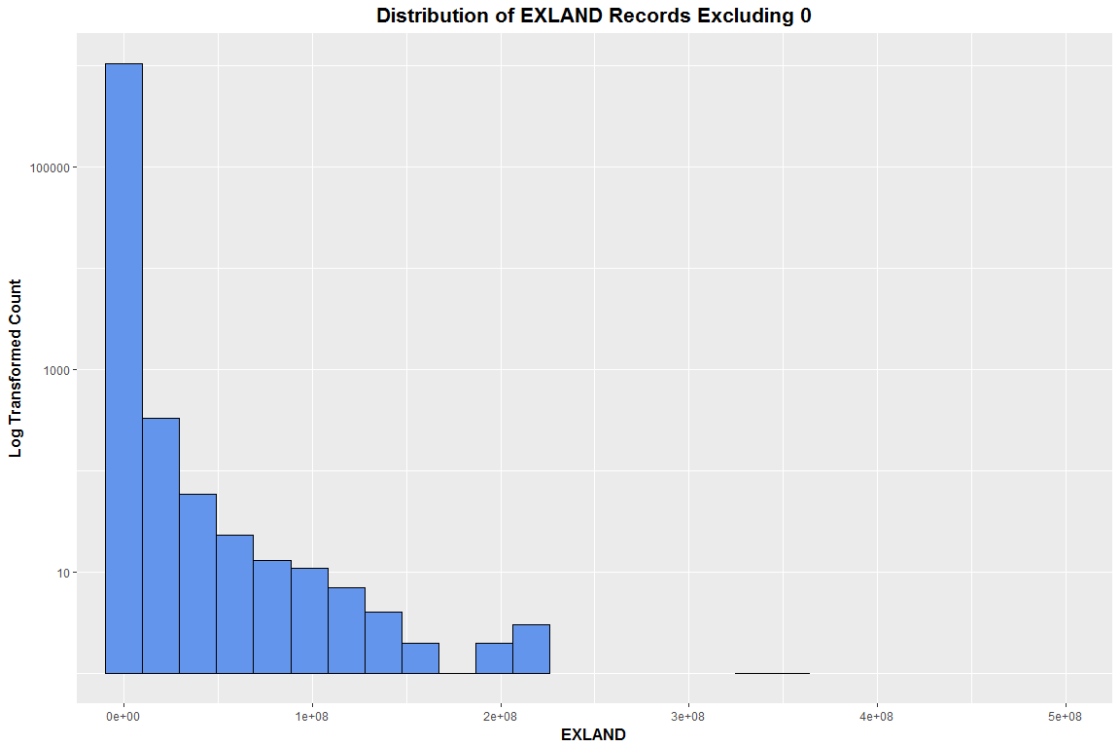
max: 1946836665

median: 1620

mode: 1620

sd: 3066658.43

Histogram:



(16) EXTOT

Description: continuous variable, represents the total value of the exempt property

Number of Missing Values: 3

Number of Unique Values: 63803, ranging from 0 to 1946836665,including 425999 records of 0 EXTOT

%populated: 99.99%

mean: 83187.20

min: 0

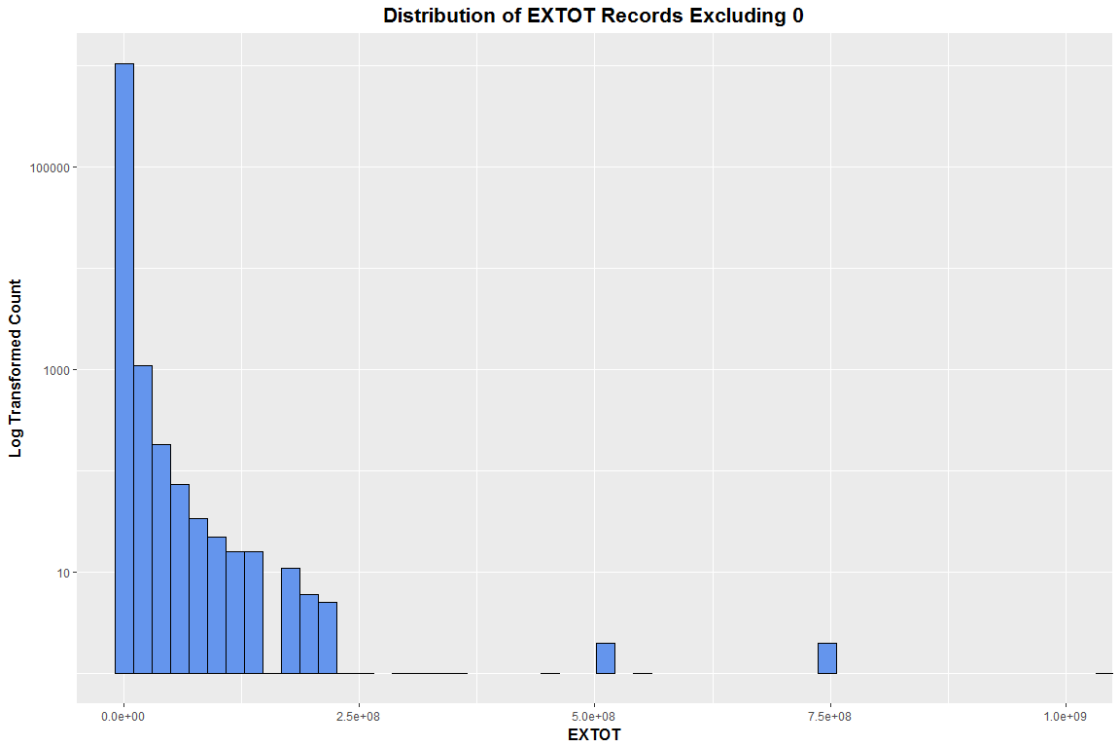
max: 1946836665

median: 1620

mode: 1620

sd: 3131423.2

Histogram:



(17) EXCD1

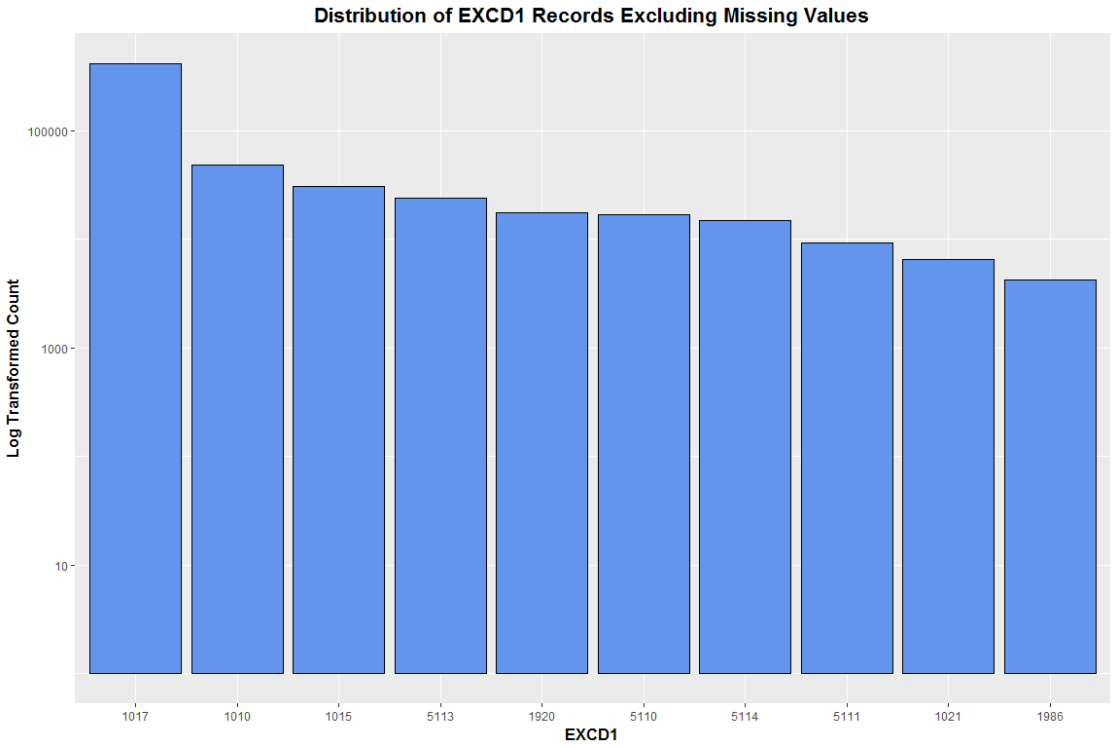
Description: categorical with metric variable, represents the code for the exempt reasons

Number of Missing Values: 425933

Number of Unique Values: 130, ranging from 1010 to 7170

%populated: 59.38%

Histogram:



(18) STADDR

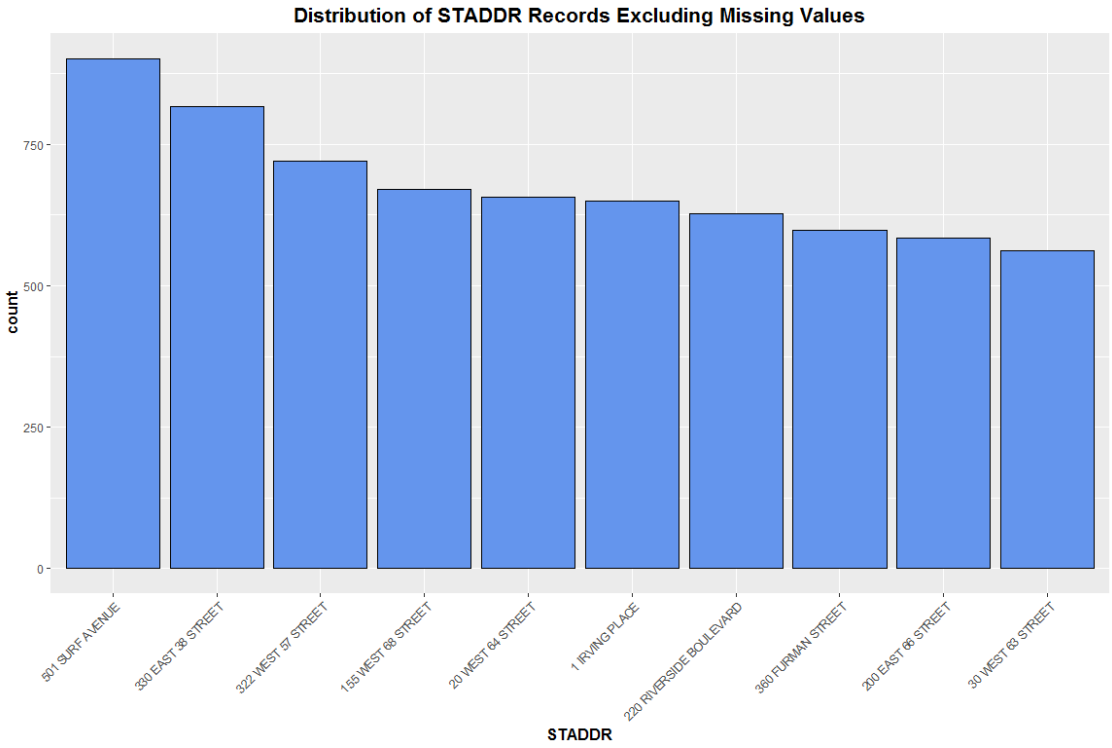
Description: categorical no mertric variable, represents the street address of the property

Number of Missing Values: 641

Number of Unique Values: 820638

%populated: 99.94%

The top 10 most frequently occurred STADDR values are:



(19) ZIP

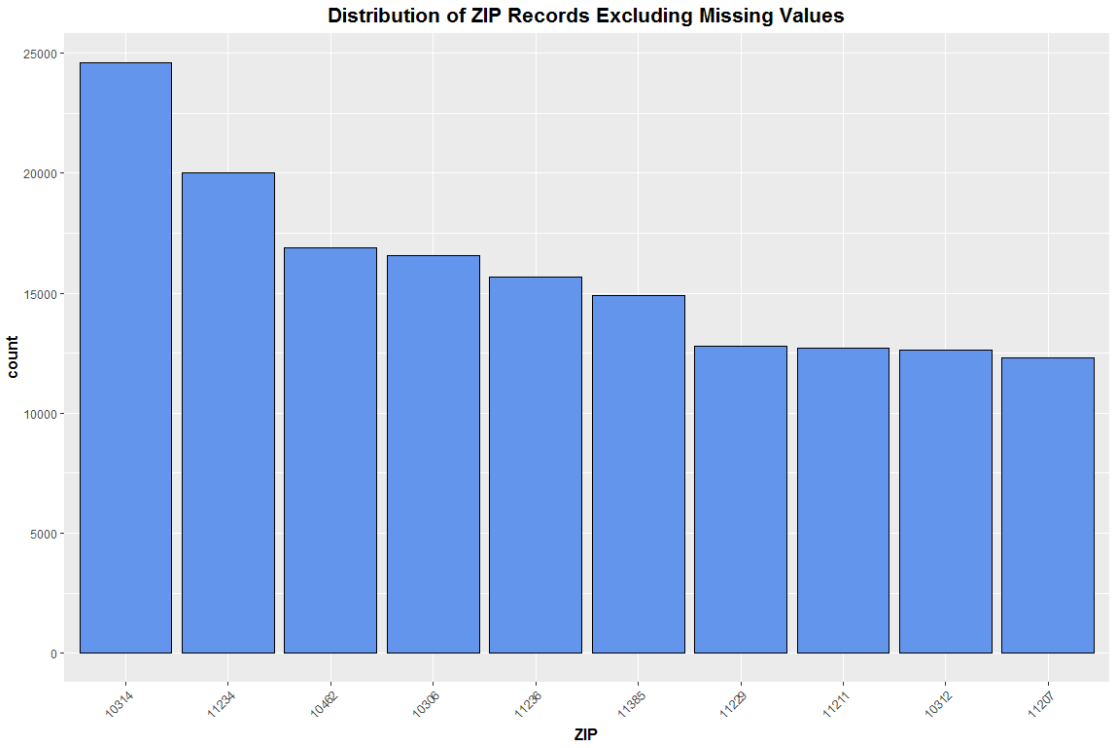
Description: categorical with metric variable, represents the zip code of the property

Number of Missing Values: 26356

Number of Unique Values: 197

%populated: 97.49%

Histogram:



(20) EXMPTCL

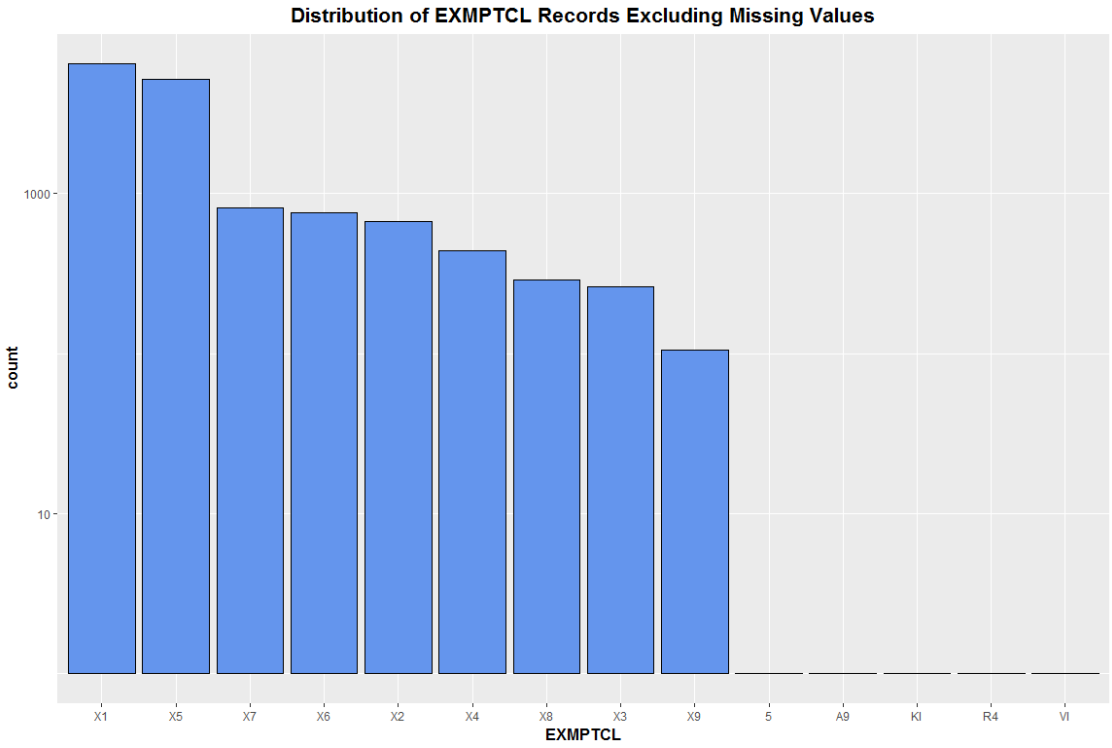
Description: categorical with metric variable, represents the exempt class

Number of Missing Values: 1033583

Number of Unique Values: has 15 levels:"","5","A9","KI","R4","VI","X1","X2","X3","X4","X5","X6","X7","X8",and "X9"

%populated: 1.43%

Histogram:



(21) BLDFRONT

Description: continuous variable, represents the length of building frontage of in feet

Number of Missing Values: 0

Number of Unique Values: 610, ranging from 0 and 7575, including 224661 records of 0 BLDFRONT

%populated: 100%

mean: 23.02

min: 0

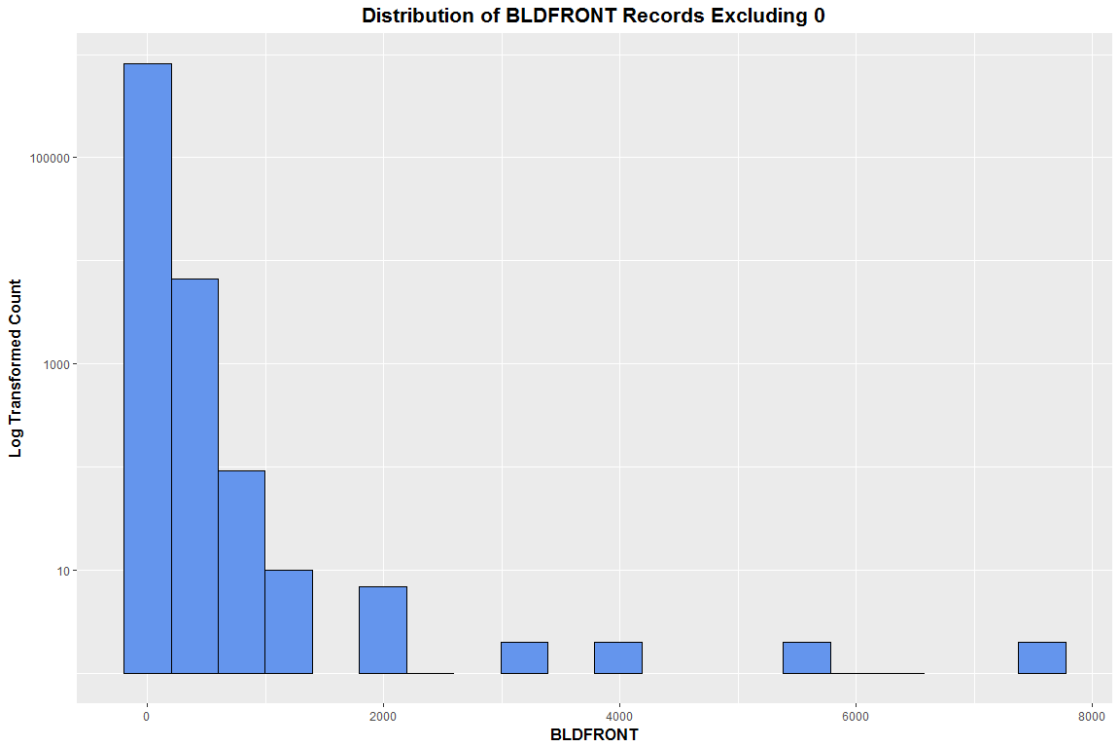
max: 7575

median: 1017

mode: 20

sd: 1388.13

Histogram:



(22) BLDDEPTH

Description: continuous variable, indicates the length of building depth in feet

Number of Missing Values: 0

Number of Unique Values: 620, ranging from 0 to 9393,including 224699 records of 0 BLDDEPTH

%populated: 100%

mean: 40.07

min: 0

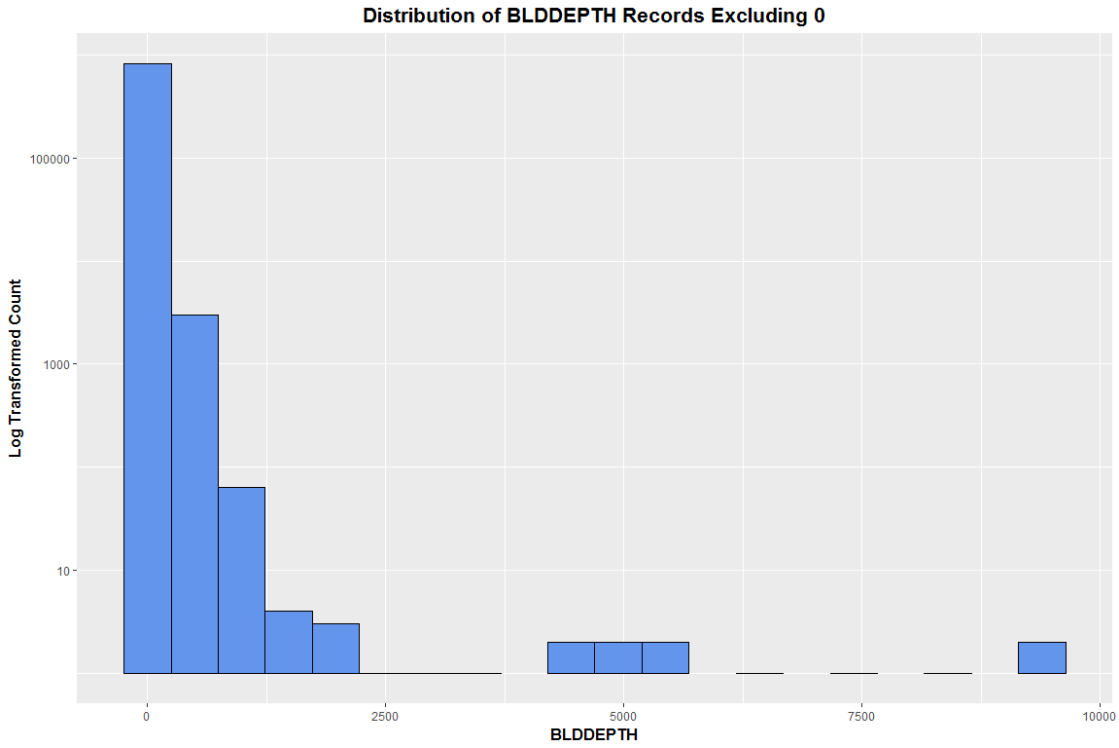
max: 9393

median: 39

mode: 40

sd: 43.04

Histogram:



(23) AVLAND2

Description: continuous variable, indicates the transitional value of the land

Number of Missing Values: 767610

Number of Unique Values: 58169, ranging from 3 to 1644454002

%populated: 26.79%

mean: 237927.57

min: 3

max: 1644454002

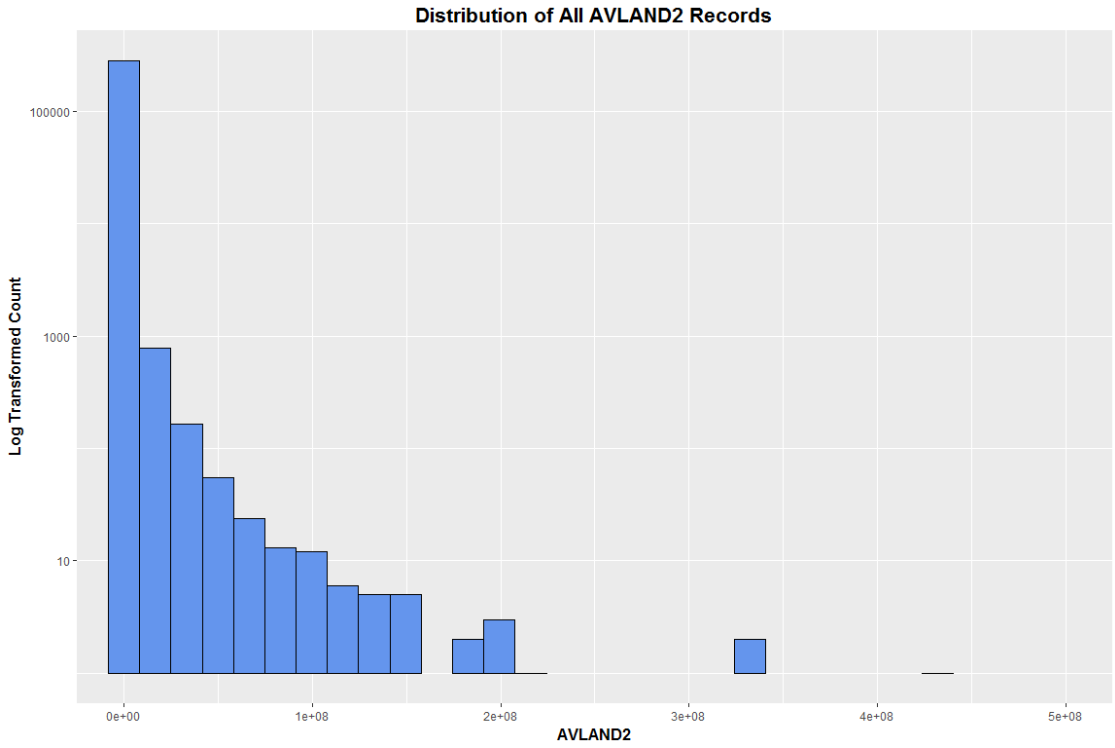
median: 20059

mode: 2408

sd: 4292804.11

Histogram:

```
[1] 16694
```



(24) AVTOT2

Description: continuous variable, indicates the transitional total value of the property

Number of Missing Values: 767605

Number of Unique Values: 110889, ranging from 3 to 1629810000

%populated: 26.80%

mean: 691290.29

min: 3

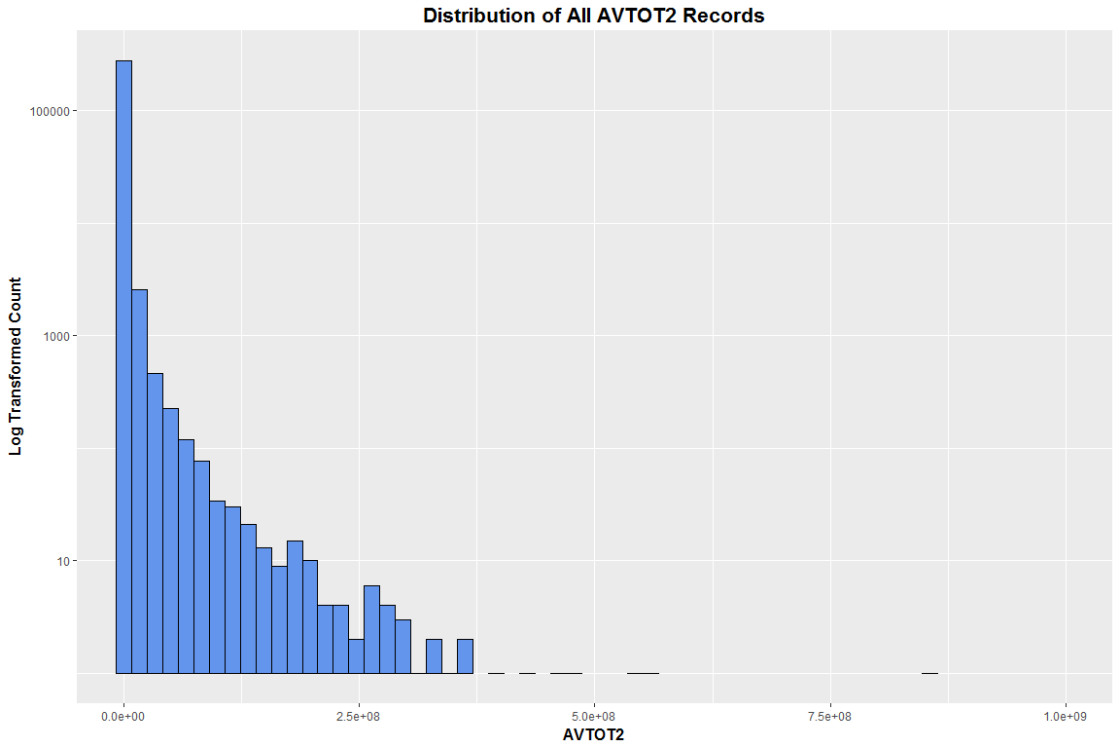
max: 1629810000

median: 80010

mode: 750

sd: 6554318.73

Histogram:



(25) EXLAND2

Description: continuous variable, indicates the transitional value of the exempt land

Number of Missing Values: 961901

Number of Unique Values: 21996, ranging from 1 to 1644454002

%populated: 8.27%

mean: 324450.83

min: 1

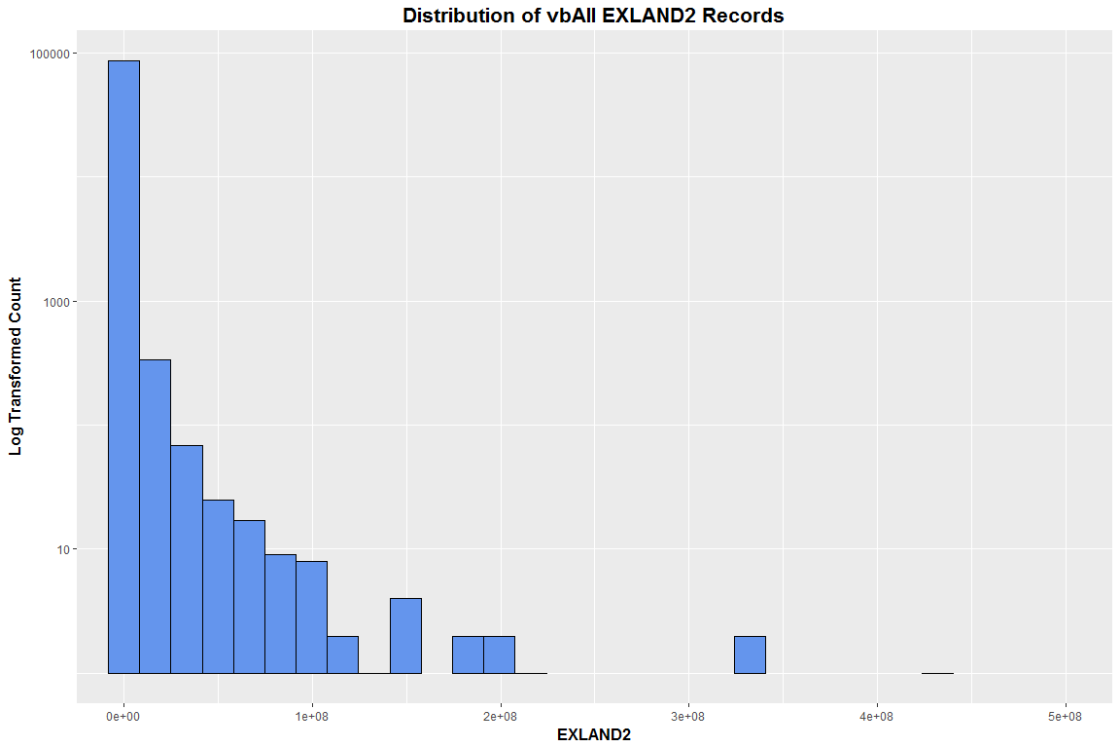
max: 1644454002

median: 3053

mode: 2090

sd: 7275688.6

Histogram:



(26) EXTOT2

Description: continuous variable, indicates the transitional total value of the exempt property

Number of Missing Values: 918644

Number of Unique Values: 48105, ranging from 7 to 1629810000

%populated: 12.39%

mean: 604510.03

min: 7

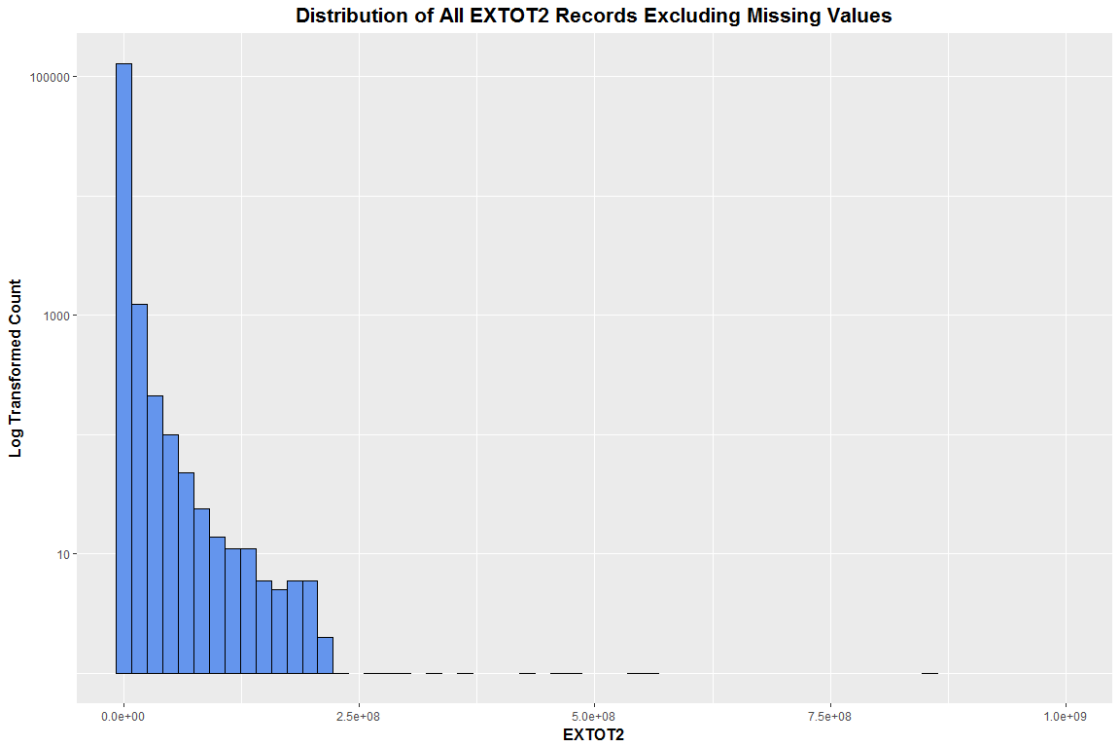
max: 1629810000

median: 37114

mode: 2090

sd: 7585515.34

Histogram:



(27) EXCD2

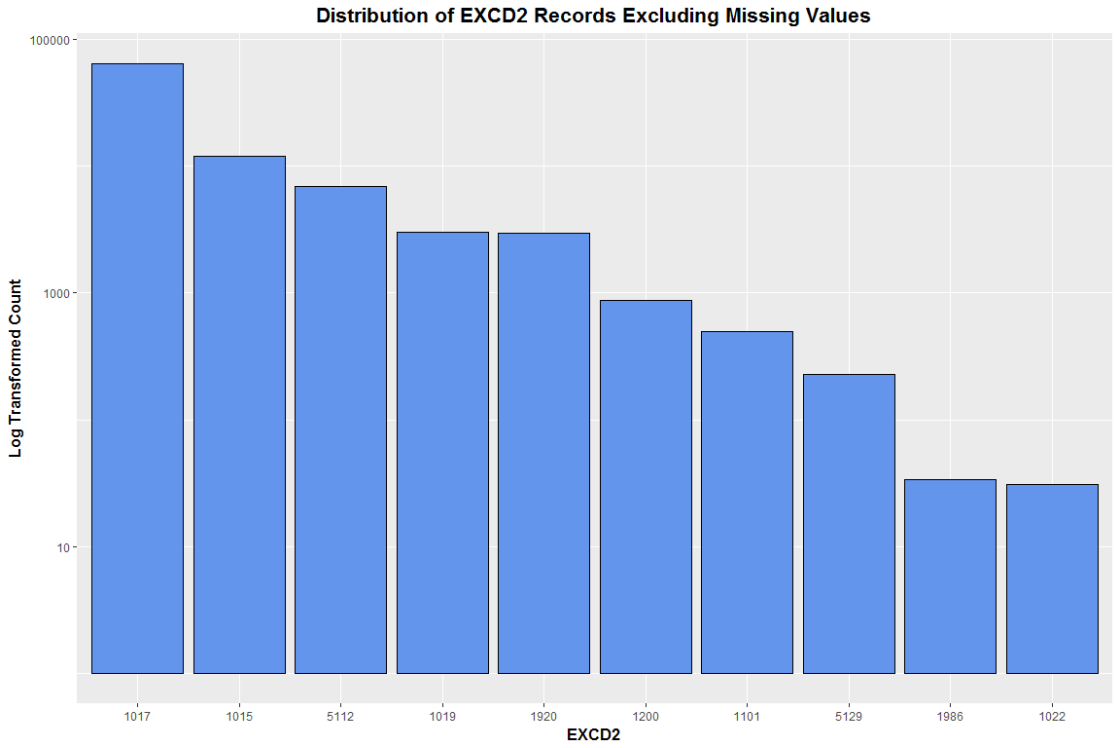
Description: categorical with metric variable, indicates the code for the exempt reasons for EXLAND2 and EXTOT2 records

Number of Missing Values: 957634

Number of Unique Values: 61, from 1011 to 7160

%populated: 8.67%

Histogram:



(28) PERIOD

Description: categorical variable, indicates the assessment period when the file was created

Number of Missing Values: 0

Number of Unique Values: 1, all have "FINAL" in the PERIOD field

%populated: 100%

(29) YEAR

Description: categorical variable, indicates the time that the record is made

Number of Missing Values: 0

Number of Unique Values: 1, all have "2010/11"in the YEAR field

%populated: 100%

(30) VALTYPE

Description: categorical variable, indicates the valid type of the record

Number of Missing Values: 0

Number of Unique Values: 1, all have "AC-TR" in the VALTYPE field

%populated: 100%