TABLE OF CONTENTS

EXECUTIVE SUMMARY

"Real Estate Fraud" is a broad term used to describe different types of frauds and deceits faced by real estate industry stakeholders. Incorrect report of land or property value is one of them, which could potentially connect with tax fraud problems.

This report commissions to examine the 2010/2011 New York property 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 cleaning and estimating missing variables
2. Creating new informative features
3. Dimensionality reduction through PCA process
4. Calculating fraud score using both heuristic algorithm and Autoencoder method

Each step will be explained in details through the report. Using both Autoencoder and Heuristic Algorithm, we selected total of 37 unique abnormal records with highest fraud scores, which could be classified as underlying real estate frauds.

Further investigation into those suspicious records shows that these abnormality can be summarized into several types, such as extremely high/low values, mismatching numbers, tax exemption cases, etc. Our assumption of these latent fraud records are mainly:

1. Falsely report land/property value to get high loans
2. Tax avoidance
3. Incorrect data input

The report, however, may include following limitations:

1. Limited understanding of New York property legislation
2. Lack of thorough understanding of certain variables
3. As an unsupervised learning model, someparameters of the models are not well-adjusted, and further optimization is needed

DATA DESCRIPTION

Overview from NYC Government

“The Offices of the City Register maintain the New York City public records for the Bronx, Brooklyn, Manhattan, and Queens. These records include Real and Personal Property transfers, interest, and ownership information. The Richmond County Clerk maintains all property records for Staten Island.

These records are open for inspection and must be recorded and corrected through the office of the Richmond Records dated after 1966 can be recorded and corrected through the Automated City Register Information online or at ACRIS terminals in City Register Offices. Records dated before 1966 can only be accessed City Register's Office in the borough where the property is located.”

File name: NY_property_data Source
Url: https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8
Data Provided by: Department of Finance (DOF)
Dataset Owner: NYC OpenData
Category: Housing & Development
Date Created: September 2, 2011 Metadata Last Updated: September 5, 2014
Data Volume: 1048575 Fields: 30 fields, 16 continuous, 12 categorical, 2 text
Details Variables: RECORD, BBLE, BLOCK, LOT, EASEMENT, OWNER, BLDGCL, TAXCLASS, LTFRONT, LTDEPTH, STORIES, FULLVAL, AVLAND, AVTOT, EXLAND, EXTOT, EXCD1, STADDR, ZIP, EXMPTCL, BLDFRONT, BLDDEPTH, AVLAND2, EXLAND2, EXLAND2, EXTOT2, EXCD2, PERIOD, YEAR, VALTYPE

Original Names with Description

Field Name	Type	Description
RECORD	categorical	Number of records
BBLE	categorical	Concatenation of BORO, BLOCK, LOT, EASEMENT
BLOCK	categorical	Valid block ranges by BORO Codes
LOT	categorical	Unique # of the property within BORO/BLOCK
EASEMENT	categorical	Used to describe easement
OWNER	text	Owner's name
BLDGCL	categorical	Building class
TAXCLASS	categorical	Tax class
LTFRONT	continuous	Lot Frontage in feet
LTDEPTH	continuous	Lot Depth in feet
STORIES	continuous	Number of stories for the building
FULLVAL	continuous	Total market value of the property
AVLAND	continuous	If not zero the Total Land Area
AVTOT	continuous	Assessed Value of the property
EXLAND	continuous	Part of Land Value that is tax exempted
EXTOT	continuous	Part of total assessed value that is tax exempted
EXCD1	continuous	No info
STADDR	text	Street Address
ZIP	categorical	Postal zip code of the property
EXMPTCL	categorical	Exempt Class used for fully exampt properties only
BLDFRONT	continuous	Building Frontage in feet
BLDDEPTH	continuous	Building Depth in feet
AVLAND2	continuous	No info
AVTOT2	continuous	No info
EXLAND2	continuous	No info
EXTOT2	continuous	No info
EXCD2	continuous	No info
PERIOD	categorical	No info
YEAR	date	Assessment year
VALTYPE	categorical	No info

DATA CLEANING

Rationale Behind

Due to the nature of this large, relatively messy dataset, data cleaning before proceeding to operation is essential to the success of this project. We decided to divide this dataset into two parts, based on the number of zeros and/or missing values of each column. Mainly we focus on seven key numeric variables which are AVLAND, AVTOT, FULLVAL, LTDEPTH, LTFRONT, STORIES, and ZIP. If a record has 3 or more missing values out of the 7 key variables, this record will be considered as a "inferior and highly suspicious fraud record". We decide to divide our dataset based on this threshold since filling the most of missing values of a record with the average of correspondent features may potentially transfer fraud record to normal record. Those records that have 3 or more missing values of the 7 key variables (total number of which is 16751) are separated from our dataset. In the end, we proceed with 1031834 records and implement machine learning methods to conduct fraud detections.

Steps

1. Count the number of zeros and/or NAs of the 7 key variables in each row. Record each entry
2. If count is less than 3, we store each record into a new dataset, which will be used for future analysis
3. For those count numbers that are greater than or equal to 3, we store them into another dataset, and set aside for later study
4. Exclude the following columns, since they are irrelevant to the analysis we are going to perform: EASEMENT, OWNER, STADDR, PERIOD, YEAR, VALTYPE - these are mostly textual or with only one category

We found 16,751 entries that have more than or equal to 3 zeros/NAs.

VARIABLE CREATION

Overview

In order to detect frauds, we need further information beyond the original dataset. With that in mind, variable creation is also an essential part of this project. Below is a detailed description of variables we've created in addition to the original ones.

Catalog of Expert Variables

Label	Mathematical Meaning
Var 01	Ratio of *FULVAL* and Average *FULVAL* of buildings grouped by *ZIP CODE*
Var 02	Ratio of *FULVAL* and Average *FULVAL* of buildings grouped by *TAX CLASS*
Var 03	Average *FULLVAL* per footprint (*LTFRONT * LTDEPTH*)
Var 04	Average *FULLVAL* per volume (*LTFRONT * LTDEPTH * STORIES*)
Var 13	Ratio of *AVLAND* and Average *AVLAND* of buildings grouped by *ZIP CODE*
Var 14	Ratio of *AVLAND* and Average *AVLAND* of buildings grouped by *TAX CLASS*
Var 15	Average *AVLAND* per footprint (*LTFRONT * LTDEPTH*)
Var 18	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by *ZIP CODE*
Var 19	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by *TAX CLASS*
Var 20	Average *AVTOT* per footprint (*LTFRONT * LTDEPTH*)
Var 23	Average *AVTOT* per volume (*LTFRONT * LTDEPTH * STORIES*)
Var 25	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by *TAX CLASS* and *STORIES*
Var 26	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by footprint and *STORIES*
Var 27	Ratio of *EXTOT / AVTOT* and Average *EXTOT/AVTOT* of the buildings grouped by *ZIP CODE*
Var 28	Ratio of *EXTOT / AVTOT* and Average *EXTOT/AVTOT* of the buildings grouped by *TAX CLASS*
Var 29	Ratio of *EXTOT/AVTOT* and Average *EXTOT/AVTOT* of the buildings with same volume (*LTFRONT*LTDEPTH*STORIES*)
Var 30	Ratio of *EXLAND / AVLAND* and Average *EXLAND / AVLAND* of the buildings grouped by *ZIP CODE*
Var 31	Ratio of *EXLAND / AVLAND* and Average *EXLAND / AVLAND* of the buildings grouped by *TAX CLASS*

Label	Mathematical Meaning
Var 32	Ratio of *EXLAND/AVLAND* and Average *EXLAND/AVLAND* of the buildings with same footprint (*LTFRONT*LTDEPTH)*
Var 33	Ratio of *FULVAL* and Average *FULVAL* of buildings grouped by *newBLOCK*
Var 34	Ratio of *AVLAND* and Average *AVLAND* of buildings grouped by *newBLOCK*
Var 35	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by *newBLOCK*
Var 36	Ratio of *EXTOT / AVTOT* and Average *EXTOT/AVTOT* of the buildings grouped by *newBLOCK*
Var 37	Ratio of *EXTOT / AVLAND* and Average *EXTOT/AVLAND* of the buildings grouped by *newBLOCK*
Var 38	Ratio of *FULVAL* and Average *FULVAL* of buildings grouped by *BLDGCL*
Var 39	Ratio of *AVLAND* and Average *AVLAND* of buildings grouped by *BLDGCL*
Var 40	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by *BLDGCL*
Var 41	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by *BLDGCL* and *STORIES*
Var 42	Ratio of *EXLAND/AVLAND* and Average *EXLAND/AVLAND* of the buildings grouped by *BLDGCL*
Var 43	Ratio of *FULVAL* and Average *FULVAL* of buildings grouped by *BFD (BLDFRONT*BLDDEPTH)*
Var 44	Average *FULVAL* per building footprint (*BLDFRONT * BLDDEPTH*)
Var 45	Average *FULVAL* per building volume (*BLDFRONT * BLDDEPTH * STORIES*)
Var 46	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by *BFD (BLDFRONT*BLDDEPTH)*
Var 47	Average *AVTOT* per building footprint (*BLDFRONT * BLDDEPTH*)
Var 48	Average *AVTOT* per building volume (*BLDFRONT * BLDDEPTH * STORIES*)
Var 49	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by *STORIES* and *BFD (BLDFRONT*BLDDEPTH)*
Var 50	Ratio of *AVTOT* and Average *AVTOT* of buildings grouped by *BFD (BLDFRONT*BLDDEPTH)* and *STORIES*
Var 51	Ratio of *EXTOT/AVTOT* and Average *EXTOT/AVTOT* of the buildings with same building volume (*BLDFRONT*BLDLTDEPTH*STORIES*)
Var 52	Ratio of *EXLAND/AVLAND* and Average *EXLAND/AVLAND* of the buildings with same building footprint (*BLDFRONT*BLDLTDEPTH*)

**The number encoded to each of the newly-created variable is NOT continuous since we deleted some of the variables which measure the data with identical meaning.*

** For variable 27 to 32, we are trying to measure the exemption ratio of each property and land. Since EXTOT stands for the exemption amount of the total property, which is AVTOT, we use EXTOT/AVTOT to measure the exemption ratio of that particular property. And EXLAND stands for the exemption amount of the total land value, which is AVLAND, we use EXLAND/AVLAND to measure the exemption ratio of that particular land.*

By comparing EXTOT/AVTOT or EXLAND/AVLAND of a particular property to the average EXTOT/AVTOT or EXLAND/AVLAND within the same Zip or Tax Class or same size, we are able to detect anomalies in the data.

After variable creation, we have 79 columns for each entry in our dataset, all of which except for RECORD (which will be later used to identify specific input after fraud detection), will be used towards our next step - principal component analysis.

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 (PC), 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 NYP 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" is 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 79 variables.
3. Plots and outputs:

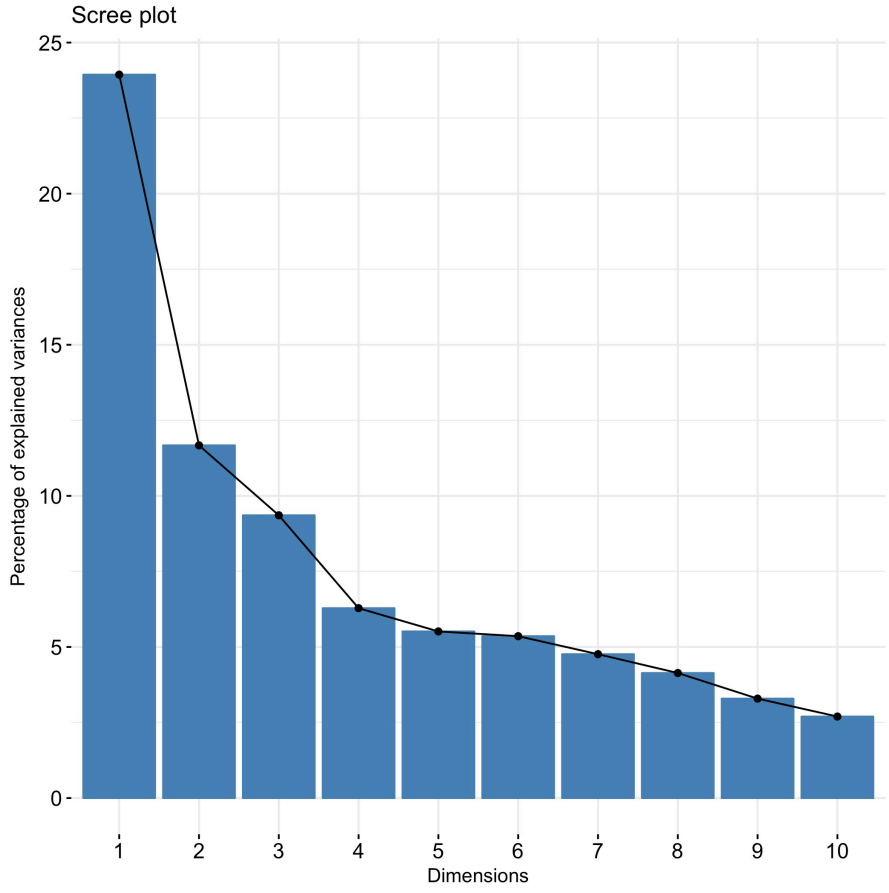


Figure 1. Scree Plot

The scree plot above can immediately show us how each PCs performs to explain the variation in the data. From PC 1 to PC 13 as a whole, the cumulative percentage of variation explained is 77.67%. PC 1 can explain up to 23.03% of the variation, and PC 2 has 10.49%.

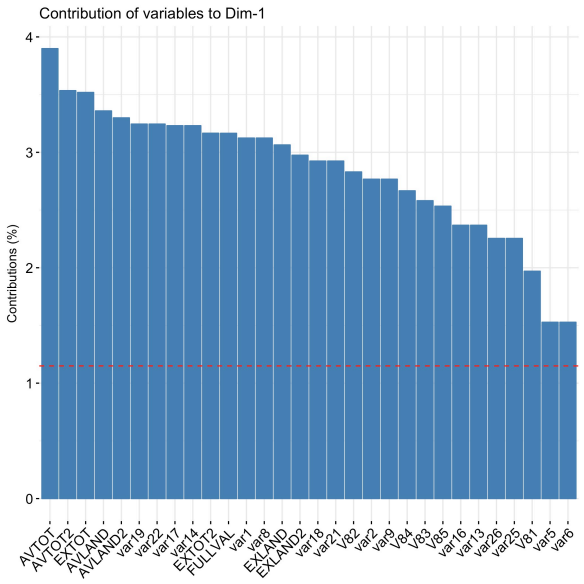


Figure 2. Contribution of Top 20 Variables to PC1

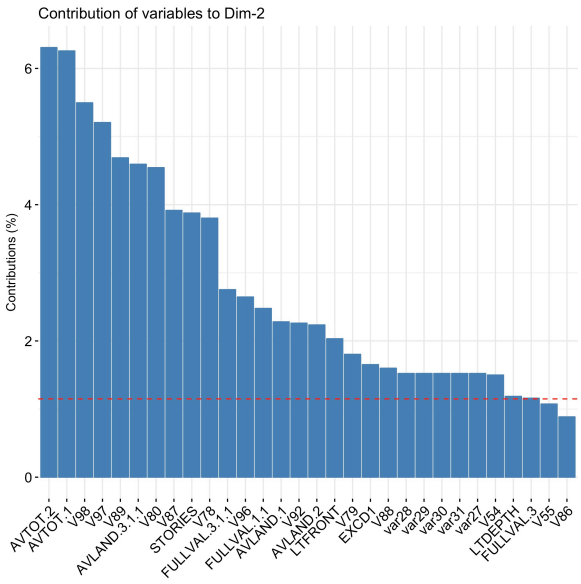


Figure 3. Contribution of Top 20 Variables to PC2

Figure 2 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 3 puts out the composition of PC2 by top-20-contributed variables. The red dashed line represents the average contribution of all variables. Top 10 variables accounts for more percentage of the composition.

CALCULATION OF FRAUD SCORE

Heuristic Algorithm

Since we have implemented z-scale before PCA, it saves us the effort to calculate Mahalanobis Distance (z-scale then euclidean). We will go for calculating the euclidean distance and constructing the algorithm directly.

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 13 PCs.

The distribution is shown as follow. We can observe that it generally right skewed and has a long tail.

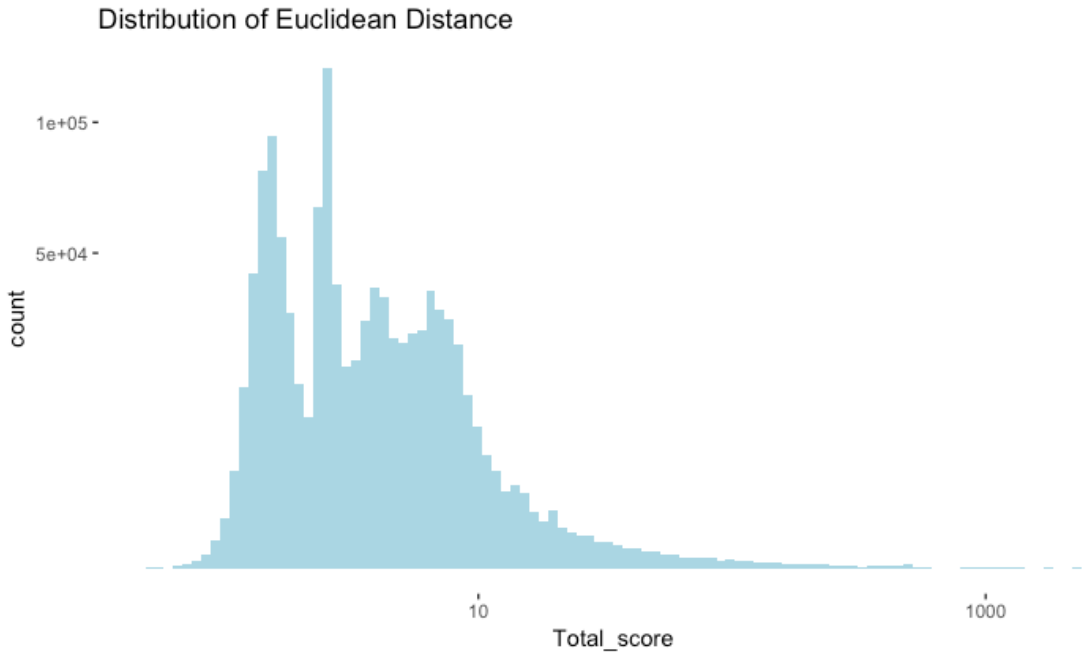


Figure 4. Distribution of Sum of Euclidean Distance

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

	dist	RECORD	dist.1	BBLE	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	EXLAND	EXTOT
1	1	376243	2376.479328	4142600001	14260	1	NA	LOGAN PROPERTY, INC.	T1	4	4910	0	3	374019883	1792808947	4668308947	1792808947	4668308947
2	2	294061	2312.930838	1011110001	1111	1	NA	CULTURAL AFFAIRS	Q1	4	840	0	NA	6150000000	2668500000	2767500000	2668500000	2767500000
3	3	866606	1801.46806	1011161001	1116	1001	NA	JPMORGAN CHASE BANK,	R5	4	0	0	7	63900000	20565000	28755000	0	0
4	4	497224	1319.794	1011320020	1132	20	NA	FORDHAM UNIVERSITY	W6	4	200	250	20	209000000	63000000	94050000	63000000	94050000
5	5	78804	1304.192403	3085900700	8590	700	NA	U S GOVERNMENT OWNRD	V9	4	117	108	NA	4326303700	1946836665	1946836665	1946836665	1946836665
6	6	315453	1131.238558	4009260001	926	1	NA	NEW YORK STATE DEPART	T1	4	3030	5948	1	5279000000	343800000	2375550000	343800000	2375550000
7	7	901790	1068.69595	4141400001	14140	1	NA	UNITED STATES OF AMER	V0	1B	999	999	NA	540143500	32408610	32408610	32408610	32408610
8	8	6949	1013.81142	1015101092	1510	1092	NA	BOXWOOD FLTD PARNTERS	R4	2	75	93	31	296508	22896	133429	0	0
9	9	612575	1009.118335	1015110001	1511	1	NA	969 PARK CORP	D4	2	175	193	12	20400000	4590000	9180000	228507	228507
10	10	888450	873.9649454	4151000700	15100	700	NA	U S GOVERNMENT OWNRD	V9	4	8000	2600	NA	1662400000	748080000	748080000	748080000	748080000

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.

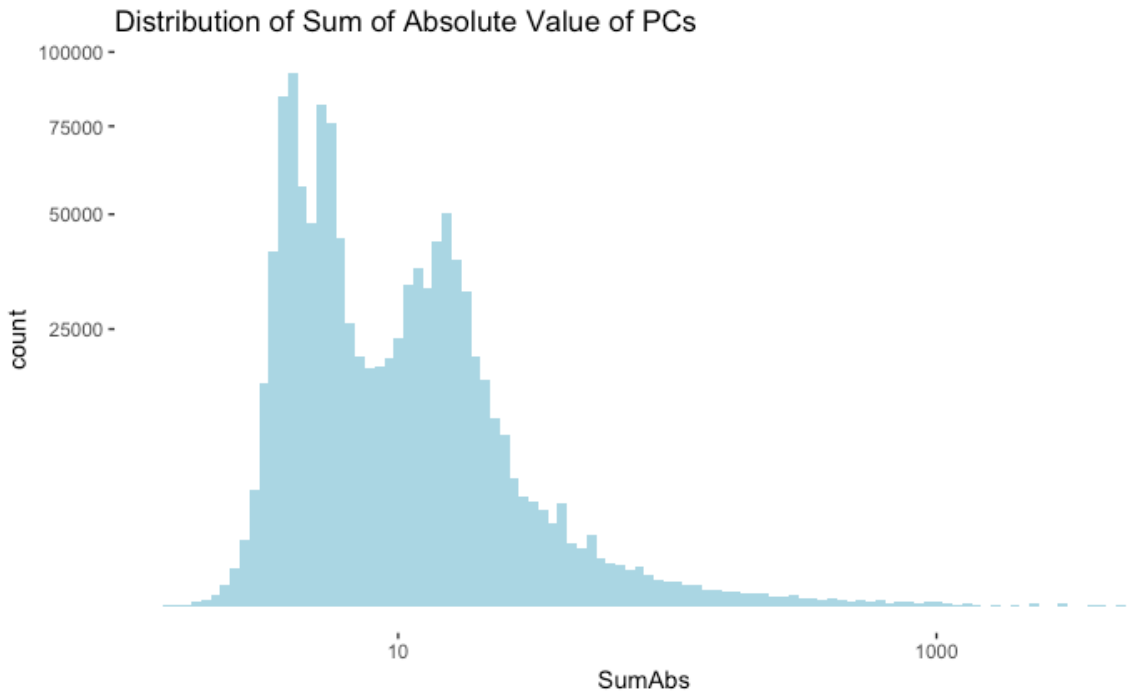


Figure 5. Distribution of Sum of Absolute Value of PCs

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

	sum	RECORD	sum.1	BBLE	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	EXLAND	EXTOT
1	1	376243	4965.306	4142600001	14260	1	NA	LOGAN PROPERTY, INC.	T1	4	4910	0	3	374019883	1792808947	4668308947	1792808947	4668308947
2	2	294061	4180.940	1011110001	1111	1	NA	CULTURAL AFFAIRS	Q1	4	840	0	NA	6150000000	2668500000	2767500000	2668500000	2767500000
3	3	866606	3914.031	1011161001	1116	1001	NA	JPMORGAN CHASE BANK,	R5	4	0	0	7	63900000	20565000	28755000	0	0
4	4	497224	2924.057	1011320020	1132	20	NA	FORDHAM UNIVERSITY	W6	4	200	250	20	209000000	63000000	94050000	63000000	94050000
5	5	901790	2850.838	4141400001	14140	1	NA	UNITED STATES OF AMER	V0	1B	999	999	NA	540143500	32408610	32408610	32408610	32408610
6	6	78804	2357.116	3085900700	8590	700	NA	U S GOVERNMENT OWNRD	V9	4	117	108	NA	4326303700	1946836665	1946836665	1946836665	1946836665
7	7	315453	2276.281	4009260001	926	1	NA	NEW YORK STATE DEPART	T1	4	3030	5948	1	5279000000	343800000	2375550000	343800000	2375550000
8	8	612575	2221.160	1015110001	1511	1	NA	969 PARK CORP	D4	2	175	193	12	20400000	4590000	9180000	228507	228507
9	9	6949	2213.881	1015101092	1510	1092	NA	BOXWOOD FLTD PARNTERS	R4	2	75	93	31	296508	22896	133429	0	0
10	10	888450	1892.939	4151000700	15100	700	NA	U S GOVERNMENT OWNRD	V9	4	8000	2600	NA	1662400000	748080000	748080000	748080000	748080000

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.

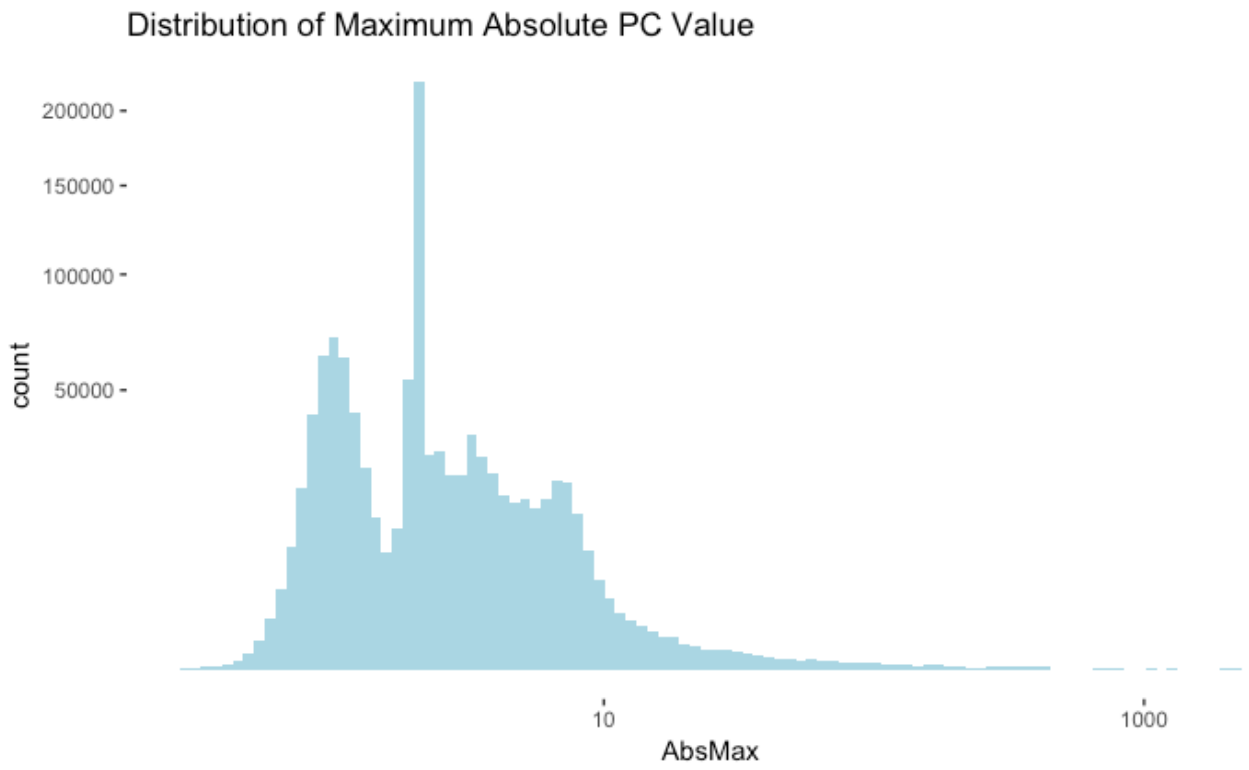


Figure 6. Distribution of Maximum Absolute PC Value

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

	max	RECORD	dist.1	BBLE	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	EXLAND	EXTOT
13	1	294061	2123.794589	1011110001	1111	1	NA	CULTURAL AFFAIRS	Q1	4	840	0	NA	6150000000	2668500000	2767500000	2668500000	2767500000
14	2	376243	2080.822974	4142600001	14260	1	NA	LOGAN PROPERTY, INC.	T1	4	4910	0	3	374019883	1792808947	4668308947	1792808947	4668308947
15	3	866606	1329.843752	1011161001	1116	1001	NA	JPMORGAN CHASE BANK,	R5	4	0	0	7	63900000	20565000	28755000	0	0
16	4	78804	1243.831826	3085900700	8590	700	NA	U S GOVERNMENT OWNRD	V9	4	117	108	NA	4326303700	1946836665	1946836665	1946836665	1946836665
17	5	315453	1043.568011	4009260001	926	1	NA	NEW YORK STATE DEPART	T1	4	3030	5948	1	5279000000	343800000	2375550000	343800000	2375550000
18	6	497224	1028.825239	1011320020	1132	20	NA	FORDHAM UNIVERSITY	W6	4	200	250	20	209000000	63000000	94050000	63000000	94050000
19	7	447396	823.3969779	3085910100	8591	100	NA	DEPT OF GENERAL SERVI	V9	4	466	1009	NA	2310884200	1039897890	1039897890	1039897890	1039897890
20	8	901790	782.2346166	4141400001	14140	1	NA	UNITED STATES OF AMER	V0	1B	999	999	NA	540143500	32408610	32408610	32408610	32408610
21	9	6949	761.2978325	1015101092	1510	1092	NA	BOXWOOD FLTD PARNTERS	R4	2	75	93	31	296508	22896	133429	0	0
22	10	612575	755.0853613	1015110001	1511	1	NA	969 PARK CORP	D4	2	175	193	12	20400000	4590000	9180000	228507	228507

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. Those records detected by the former two methods are in RECORD #: 6949, 78804, 294061, 315453, 376243, 497224, 612575, 866606, 888450, 901790. The only one different record popped out using the last method is 447396 rather than 888450. There may be one variable in 447396 that is extremely abnormal that needs further inspection.

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 13 principal directions and therefore compress features of each record to 13. 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 5. 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.

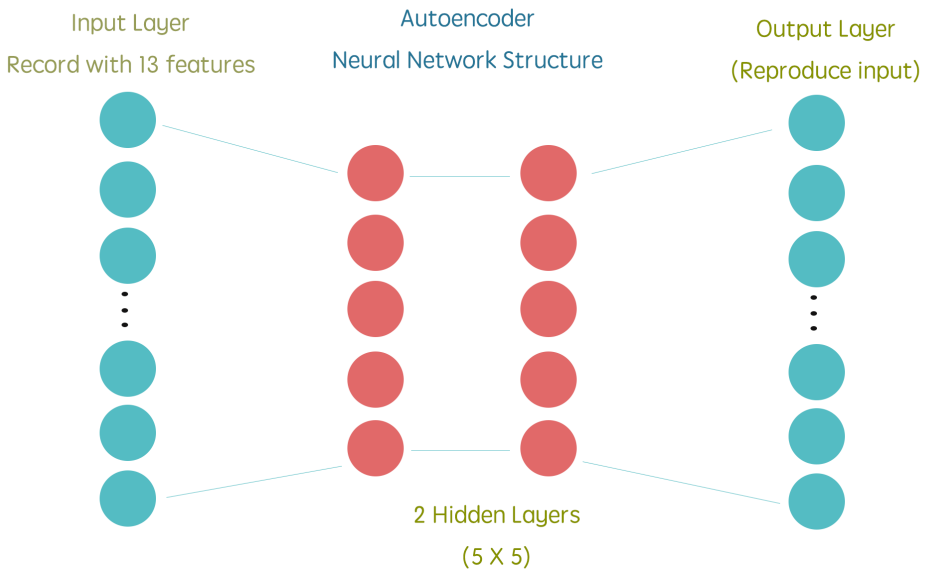


Figure 7. Autoencoder Neural Network Structure

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.

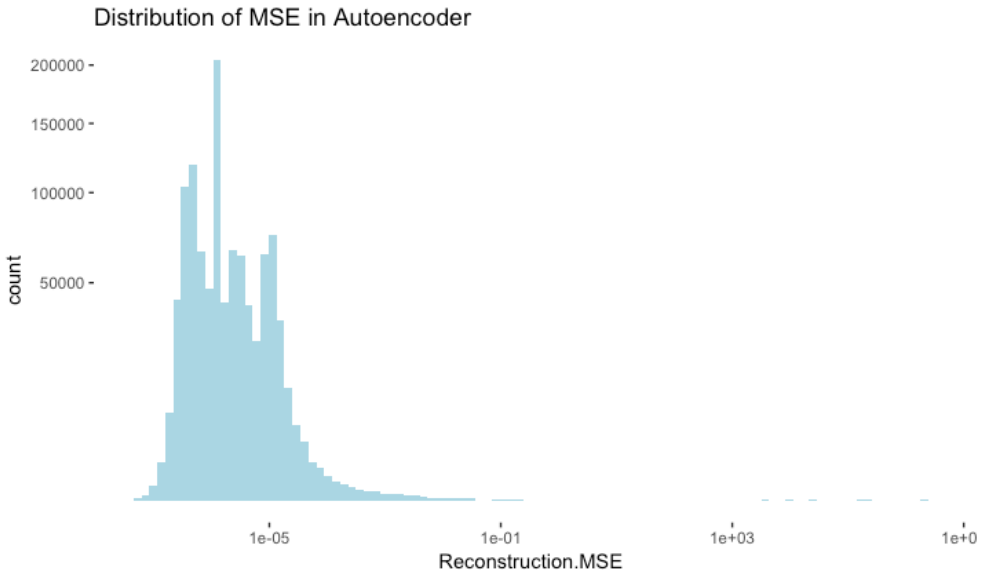


Figure 8: Distribution of MSE with Autoencoder

COMPARISON BETWEEN 2 CALCULATIONS

Summary

Compare the top 10,000 highest fraud score in both Euclidean Distance and Autoencoder and we find there is 76% overlapping between the two approaches.

In the Euclidean Distance approach, the top 10 fraud record is listed below,
Record: 376243, 294061, 866606, 497224, 78804, 315453, 901790, 6949, 612575, 888450

And in the Autoencoder approach, the top 10 fraud record is listed below,
Record: 294061, 78804, 315453, 497224, 447396, 511544, 888450, 376243, 866606, 6949

We have 80% overlapping in the first 10 record between the two approaches. This indicates two scoring approaches capture quite similar but not exactly identical patterns of the data.

INSIGHTS AND CONCLUSIONS

Most of anomalies in top 10 records contain plenty of missing records so that we extend records number to 30 for more insights.

1. Top 30 Anomalies of Euclidean Distance Approach

	RECORD	BBLE. (unique.identifier)	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT. (length)	LTDEPTH	STORIES	FULLVAL	AVLAND(assed.value.for.land.itself)	AVTOT(value.total)
1	6949	1015101092	1510	1092	NA	BOXWOOD FLTD PARNTERS	R4	2	75	93	31	296508	22896	133429
2	47631	3009631102	963	1102	NA	NA	R1	2C	0	0	4	3537033	9	13035
3	49808	3009631106	963	1106	NA	LAVAGNINO, DORIA	R1	2C	0	0	4	2456579	6	9053
4	78804	3085900700	8590	700	NA	U S GOVERNMENT OWNRD	V9	4	117	108	NA	4326303700	1946836665	1946836665
5	126050	4153760888	15376	888	NA	DCAS	V0	1B	2059	3189	NA	250277625	15016658	15016658
6	156402	2026050040	2605	40	NA	CITY OF NEW YORK	Y3	4	999	999	6	1021000000	83700000	459450000
7	194459	3009631104	963	1104	NA	NA	R1	2C	0	0	4	2456579	6	9053
8	294061	1011110001	1111	1	NA	CULTURAL AFFAIRS	Q1	4	840	0000	NA	6150000000	2668500000	2767500000
9	315453	4009260001	926	1	NA	NEW YORK STATE DEPART	T1	4	3030	5948	1	5279000000	343800000	2375550000
10	376243	4142600001	14260	1	NA	LOGAN PROPERTY, INC.	T1	4	4910	0000	3	374019883	1792808947	4668308947
11	403412	3009631108	963	1108	NA	ZYLLA, KARI A	R1	2C	0	0	4	3308103	9	12192
12	447396	3085910100	8591	100	NA	DEPT OF GENERAL SERVI	V9	4	466	1009	NA	2310884200	1039897890	1039897890
13	467075	4162400101	16240	101	NA	PARKS AND RECREATION	V0	1B	1450	1007	NA	212210578	12732635	12732635
14	476709	3009631103	963	1103	NA	NEAL, KELLY	R1	2C	0	0	4	2456579	6	9053
15	497224	1011320020	1132	20	NA	FORDHAM UNIVERSITY	W6	4	200	250	20	209000000	63000000	94050000
16	508448	3009631105	963	1105	NA	WANG, HONGYI	R1	2C	0	0	4	2456579	6	9053
17	511544	1014201439	1420	1439	NA	ROBERTS, HOLLY HARLAY	R4	2	0	0	21	186608	17072	83974
18	527478	3009631107	963	1107	NA	NA	R1	2C	0	0	4	3096943	8	11412
19	566900	3009631101	963	1101	NA	CROWLEY, IAN	R1	2C	0	0	4	3798604	10	13998
20	605798	1010400029	1040	29	NA	WWP OFFICE, LLC	O9	4	200	290	49	501000000	54900000	225450000
21	612575	1015110001	1511	1	NA	969 PARK CORP	D4	2	175	193	12	20400000	4590000	9180000
22	650467	5023590001	2359	1	NA	PARKS AND RECREATION	Q5	4	898	396	1	433200000	194850000	194940000
23	694877	4017870020	1787	20	NA	NA	Q6	4	1700	1635	6	1223500000	110025000	550575000
24	725233	3034750064	3475	64	NA	DCAS	V0	1B	186	2060	NA	22158000	97	97
25	866606	1011161001	1116	1001	NA	JPMORGAN CHASE BANK,	R5	4	0	0	7	63900000	20565000	28755000
26	888450	4151000700	15100	700	NA	U S GOVERNMENT OWNRD	V9	4	8000	2600	NA	1662400000	748080000	748080000
27	901790	4141400001	14140	1	NA	UNITED STATES OF AMER	V0	1B	999	999	NA	540143500	32408610	32408610
28	911958	4141400100	14140	100	NA	DEPT RE-CITY OF NY	V0	1B	999	999	NA	214450275	12867017	12867017
29	928226	2039371011	3937	1011	NA	PARKCHESTER PRESERVAT	R5	4	0	0	10	33300000	5346000	14985000
30	987389	2054670100	5467	100	NA	RUFFALO ENTERPRISES	V0	1B	503	999	NA	31236727	134	134

2. Top 30 Anomalies of Autoencoder Approach

	RECORD	BBLE. (unique.identifier)	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT. (length)	LTDEPTH	STORIES	FULLVAL	AVLAND(assed.value.for.land.itself)	AVTOT(value.total)
1	6949	1015101092	1510	1092	NA	BOXWOOD FLTD PARNTERS	R4	2	75	93	31	296508	22896	133429
2	47631	3009631102	963	1102	NA	NA	R1	2C	0	0	4	3537033	9	13035
3	49808	3009631106	963	1106	NA	LAVAGNINO, DORIA	R1	2C	0	0	4	2456579	6	9053
4	78804	3085900700	8590	700	NA	U S GOVERNMENT OWNRD	V9	4	117	108	NA	4326303700	1946836665	1946836665
5	126050	4153760888	15376	888	NA	DCAS	V0	1B	2059	3189	NA	250277625	15016658	15016658
6	156402	2026050040	2605	40	NA	CITY OF NEW YORK	Y3	4	999	999	6	1021000000	83700000	459450000
7	194459	3009631104	963	1104	NA	NA	R1	2C	0	0	4	2456579	6	9053
8	294061	1011110001	1111	1	NA	CULTURAL AFFAIRS	Q1	4	840	0000	NA	6150000000	2668500000	2767500000
9	315453	4009260001	926	1	NA	NEW YORK STATE DEPART	T1	4	3030	5948	1	5279000000	343800000	2375550000
10	376243	4142600001	14260	1	NA	LOGAN PROPERTY, INC.	T1	4	4910	0000	3	374019883	1792808947	4668308947
11	403412	3009631108	963	1108	NA	ZYLLA, KARI A	R1	2C	0	0	4	3308103	9	12192
12	447396	3085910100	8591	100	NA	DEPT OF GENERAL SERVI	V9	4	466	1009	NA	2310884200	1039897890	1039897890
13	467075	4162400101	16240	101	NA	PARKS AND RECREATION	V0	1B	1450	1007	NA	212210578	12732635	12732635
14	476709	3009631103	963	1103	NA	NEAL, KELLY	R1	2C	0	0	4	2456579	6	9053
15	497224	1011320020	1132	20	NA	FORDHAM UNIVERSITY	W6	4	200	250	20	209000000	63000000	94050000
16	508448	3009631105	963	1105	NA	WANG, HONGYI	R1	2C	0	0	4	2456579	6	9053
17	511544	1014201439	1420	1439	NA	ROBERTS, HOLLY HARLAY	R4	2	0	0	21	186608	17072	83974
18	527478	3009631107	963	1107	NA	NA	R1	2C	0	0	4	3096943	8	11412
19	566900	3009631101	963	1101	NA	CROWLEY, IAN	R1	2C	0	0	4	3798604	10	13998
20	605798	1010400029	1040	29	NA	WWP OFFICE, LLC	O9	4	200	290	49	501000000	54900000	225450000
21	612575	1015110001	1511	1	NA	969 PARK CORP	D4	2	175	193	12	20400000	4590000	9180000
22	650467	5023590001	2359	1	NA	PARKS AND RECREATION	Q5	4	898	396	1	433200000	194850000	194940000
23	694877	4017870020	1787	20	NA	NA	Q6	4	1700	1635	6	1223500000	110025000	550575000
24	725233	3034750064	3475	64	NA	DCAS	V0	1B	186	2060	NA	22158000	97	97
25	866606	1011161001	1116	1001	NA	JPMORGAN CHASE BANK,	R5	4	0	0	7	63900000	20565000	28755000
26	888450	4151000700	15100	700	NA	U S GOVERNMENT OWNRD	V9	4	8000	2600	NA	1662400000	748080000	748080000
27	901790	4141400001	14140	1	NA	UNITED STATES OF AMER	V0	1B	999	999	NA	540143500	32408610	32408610
28	911958	4141400100	14140	100	NA	DEPT RE-CITY OF NY	V0	1B	999	999	NA	214450275	12867017	12867017
29	928226	2039371011	3937	1011	NA	PARKCHESTER PRESERVAT	R5	4	0	0	10	33300000	5346000	14985000
30	987389	2054670100	5467	100	NA	RUFFALO ENTERPRISES	V0	1B	503	999	NA	31236727	134	134

We selected 30 records with highest fraud score using each method. There are total 37 unique records, of which 24 records are overlapping (64.86%). By looking into each observation, we detect several kinds of abnormality, and following conclusions:

Type of Abnormality	Records	Brief Description
Abnormally high value cases	694877	It's a huge but not tall property (1700*1635, 6 stories) with extremely high value(1.2B in FULLVAL, 0.11B in AVLAND and 0.5B in AVTOT). Besides, it's fully tax exempted and has no owner information, which makes it suspicious.
	294061	It has an abnormally high value(6.15B in FULLVAL, 2.67B in AVLAND and 2.77B in AVTOT) while the corresponding average values in its ZIP 10028 are 4151059 in FULLVAL, 1039836 in AVLAND and 1634642 in AVTOT
Abnormally low value cases	49808, 194459, 476709, 508448, 527478, 47631, 566900, 403412	These records have abnormally low AVLAND (less than 10) while their FULLVAL are over 2 million. This could be wrong input or data distortion. Also, these 8 properties share same address, they are from the same block (963) and are quite close to each other (lot numbers are 1101-1108)
Tax Exemption cases	78804, 294061, 447396, 888450, 901790, 464181, 694877, 650467, 156402, 725233, 987389, 911958, 126050, 467075	For all these records, their AVLAND equals EXLAND, and AVTOT equals EXTOT, which means that all the land & property value is tax exempted: 1. By looking at the owner, we notice that many of these property are owned by public sector, such as US Government, United States Of Amer, Parks&Recreation, City of NY, Dept Re-City of NY, etc., which is reasonable. (e.g. 78804, 888450, 901790, 464181,650467,156402, 467075) 2. However, those private properties that are still totally tax exempted may seem suspicious and require further investigations. (e.g. 725233, 987389)
Mismatching numbers cases	725233, 987389	There is huge difference between their assessed value AVTOT and market value FULLVAL, which is not reasonable.
	6949, 612572, 511544	Their BLDFRONT and BLDDEPTH are larger than its LOTFRONT and LOTDEPTH, which is not reasonable.

Type of Abnormality	Records	Brief Description
Abnormal TAXCLASS	814431, 89397, 531034, 464181, 47631, 196030, 195808, 78804, 294061, etc	Tax Class 4 (the definition of this tax class is others) does not contain the highest number of properties nor largest average property value but a high percentage of top fraud records. Plus, The property under this class is quite a mixture: there are both high business buildings(814431 89397 531034 464181 47631 196030 195808 found by Heuristic Methods) and public buildings (78804，294061 etc)

**Missing values: In fact, missing values in these anomaly records do not result in a high fraud score directly because when we build algorithm to calculate fraud score, we've already manipulated the dataset and changed missing fields to corresponding average values.*

With general business knowledge, we could make assumptions about possible causes of these abnormality:

1. Falsely report land/property value to get high loans. It's a common practice in real estate frauds that property owners report higher value than reality to be able to get high loans from finance institutions.
2. Tax avoidance. By underrating the property/land value, the owner will be able to avoid tax. Another tax avoidance method could be getting tax exemption illegally.
3. Incorrect data input. Based on our observations, some data abnormality could be caused by incorrect input or inaccurate evaluation.

Currently our conclusions are based on limited understanding of NY Property dataset as well as NY Real Estate Legislation. Further in-depth investigation could be conducted with more field knowledge and comprehensive information.

APPENDIX

Data Quality Report (DQR)

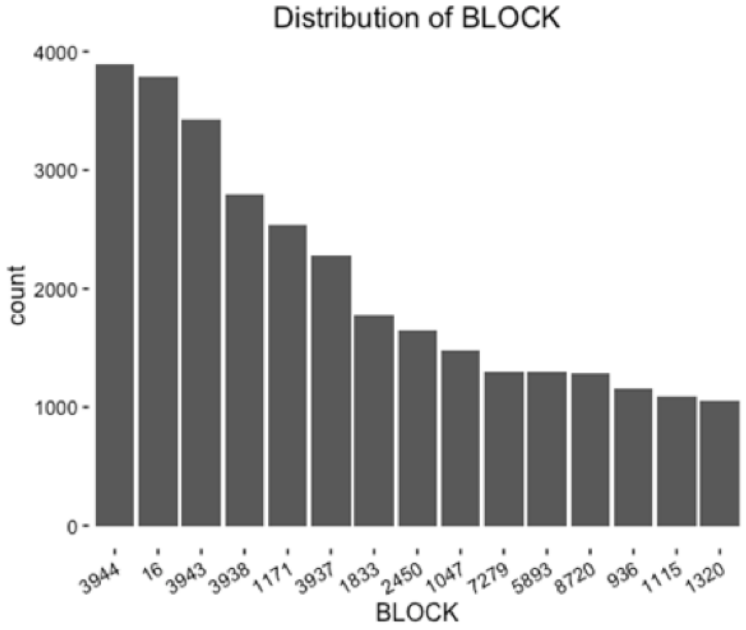
1. Summary Statistics for Numerical Variables:
   For 16 numerical variables in the data set, we provide a summary table that include mathematical statistics such as maximum and minimum value, first and third quartile, mean, median, number of NA, mode, standard deviation, number of unique value, and the percentage of populated records for that variable.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	mode	SD	Unique#	% populated
LTFRONT	1.00	20.00	25.00	43.12	40.00	9999.00	168867	20	78.62	1276	0.9264529
LTDEPTH	1.0	97.0	100.0	105.3	100.0	9999.0	169888	100	70.71	1335	0.9190918
STORIES	1.00	2.00	2.00	5.06	3.00	119.00	52142	2	8.43	111	0.999993
FULLVAL	4.00e+0	3.11e+05	4.50e+05	8.91e+05	6.23e+05	6.15e+09	12762	502000	11774390	1008276	0.9926174
AVLAND	1.000e+00	9.425e+03	1.375e+04	8.705e+04	1.983e+04	2.668e+09	12764	45000	4125933	70528	0.9900619
AVTOT	1.000e+00	1.866e+04	2.556e+04	2.336e+05	4.688e+04	4.668e+09	12762	16588	6993850	112293	0.9755419
EXLAND	1.000e+00	1.620e+03	1.620e+03	6.840e+04	3.240e+03	2.668e+09	484224	1620	5485336	33185	0.9702738
EXTOT	1.000e+00	1.620e+03	1.620e+03	1.559e+05	9.786e+03	4.668e+09	425999	1620	8536636	63804	0.9779192
EXCD1	1010	1017	1017	1605	1017	7170	425933	1017	1388.13	129	1
BLDFRONT	1.0	18.0	20.0	29.3	25.0	7575.0	224661	20	38.029	609	0.9993191
BLDDEPTH	1	35	44	51	55	9393	224699	40	42.424	619	0.9989731
AVLAND2	3.000e+00	5.705e+03	2.006e+04	2.464e+05	6.234e+04	2.371e+09	767609	2408	6199390	58169	0.9677328
AVTOT2	3.000e+00	3.401e+04	8.001e+04	7.161e+05	2.408e+05	4.501e+09	767603	750	11690165	110890	0.9894901
EXLAND2	1.000e+00	2.090e+03	3.053e+03	3.518e+05	3.142e+04	2.371e+09	961900	2090	10852484	21996	0.9587078
EXTOT2	7.000e+00	2.889e+03	3.712e+04	6.581e+05	1.066e+05	4.501e+09	918642	2090	16129808	48106	0.933104
EXCD2	1011	1017	1017	1372	1017	7160	957634	1017	1105.5	60	1

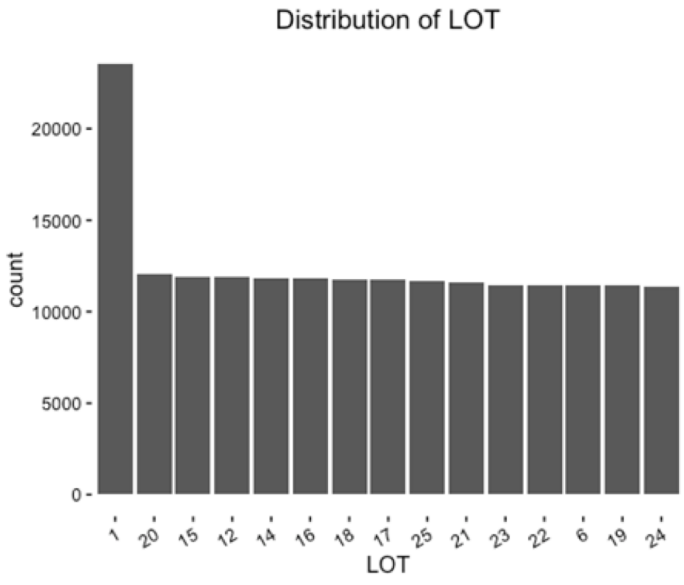
2. Detailed Information for Each Field:

Field Name	Description
RECORD	Categorical no metric. Tracking data order

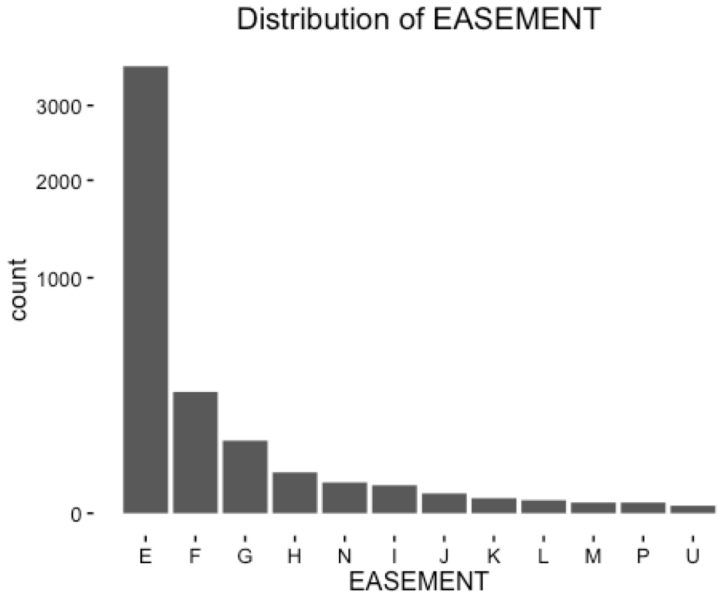
Field Name	Description
BLOCK	categorical no metric. Valid block ranges by boro



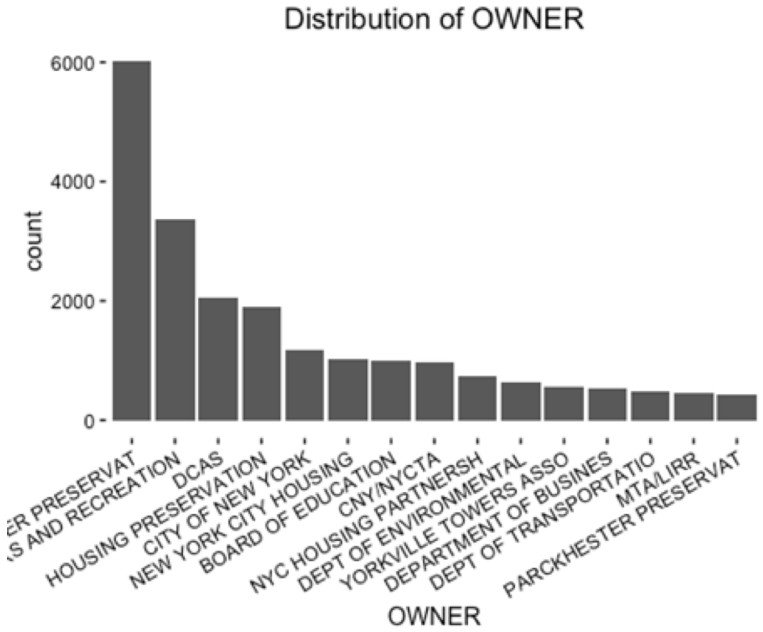
Field Name	Description
LOT	categorical no metric. Unique number within block or boro; Unique LOT number" 6366



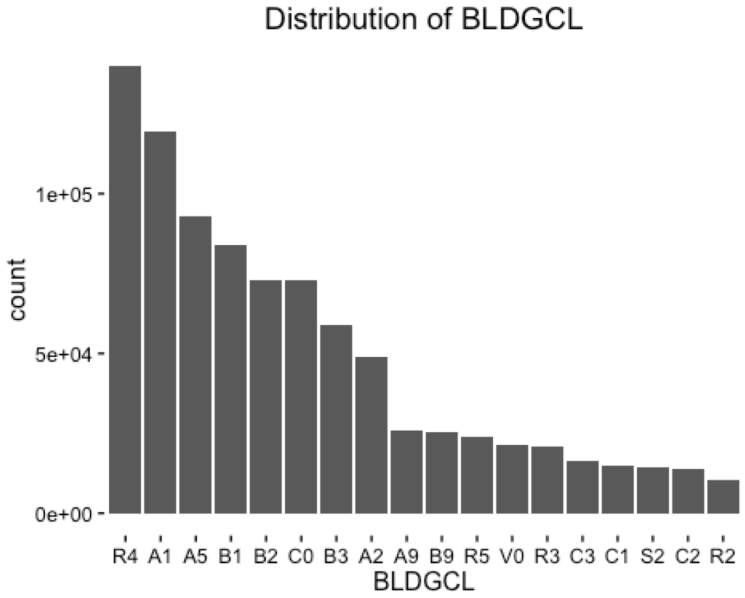
Field Name	Description
EASEMENT	categorical with metric, is a field that is used to describe easement



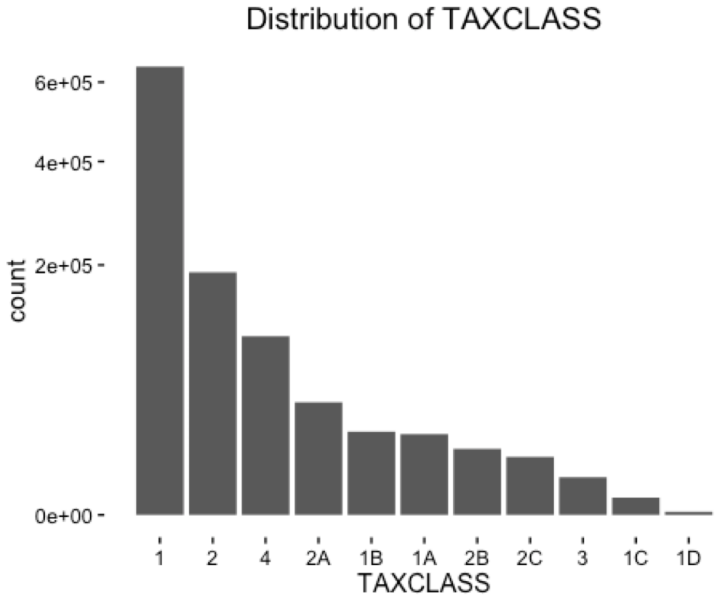
Field Name	Description
OWNER	character without metric. The Owner's Name



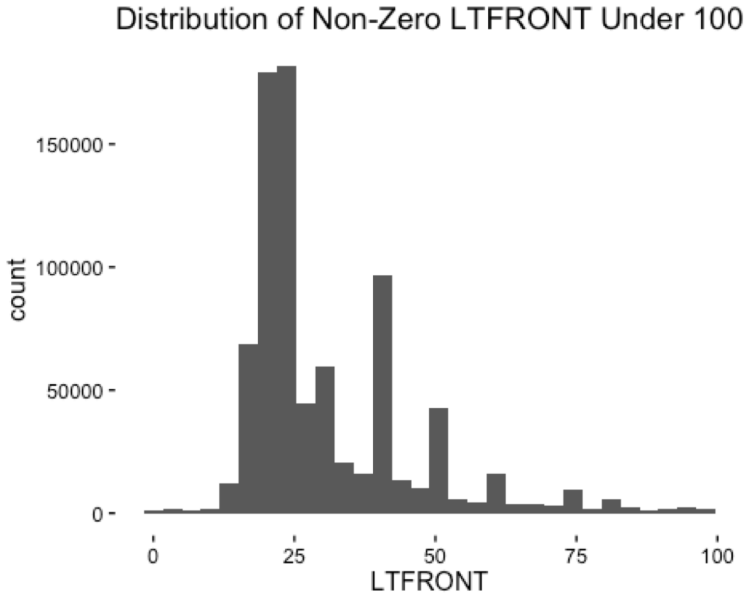
Field Name	Description
BLDGCL	categorical with metric. Building class



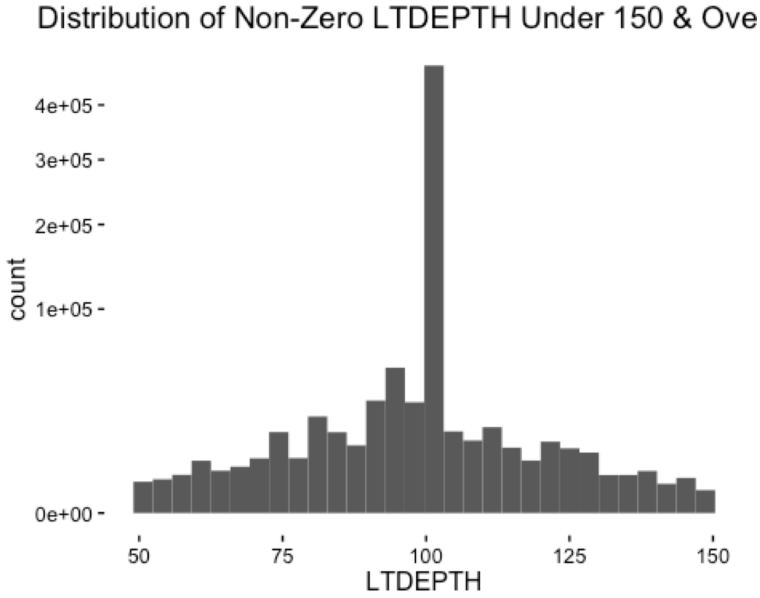
Field Name	Description
TAXCLASS	categorical with metric. Current Property Tax Class Code



Field Name	Description
LTFRONT	continuous with metric, Lot Frontage in feet

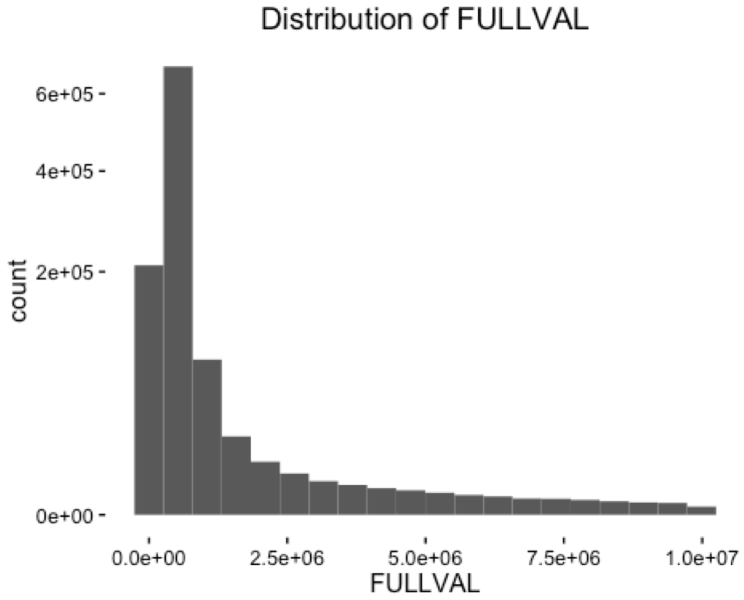


Field Name	Description
LTDEPTH	continuous with metric, Lot Depth in feet

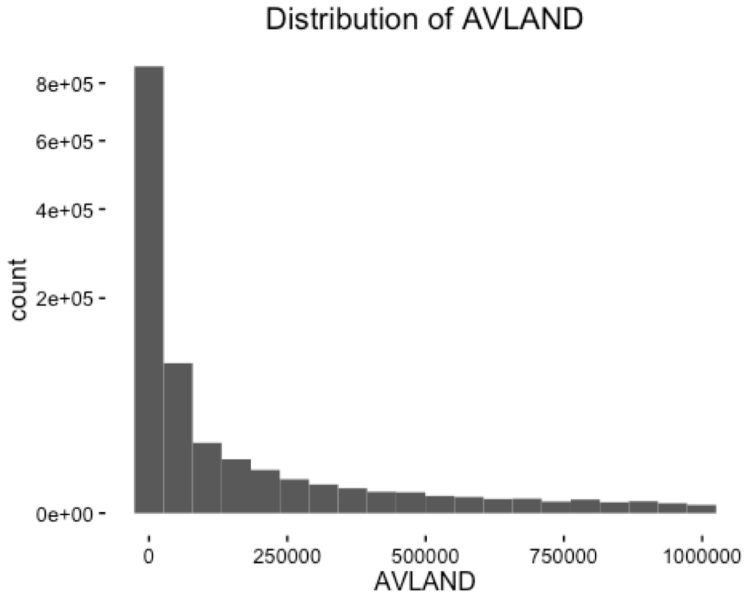


Field Name	Description
STORIES	continuous with metric. The number of stories for the building

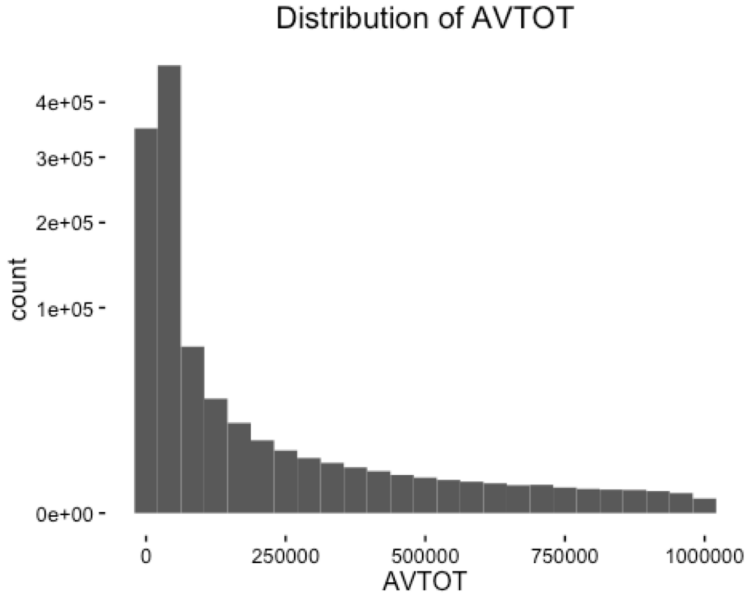
Field Name	Description
FULLVAL	continuous with metric, full value of the property



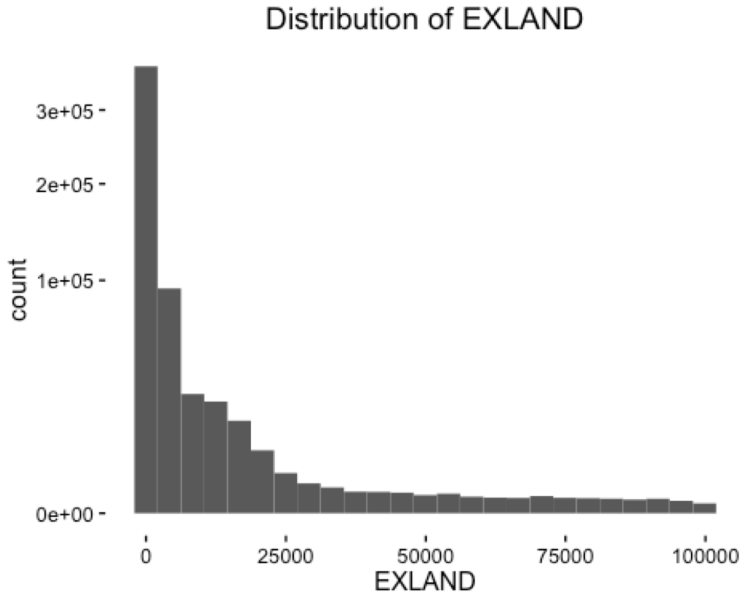
Field Name	Description
AVLAND	continuous with metric, Current year's total market value of the land



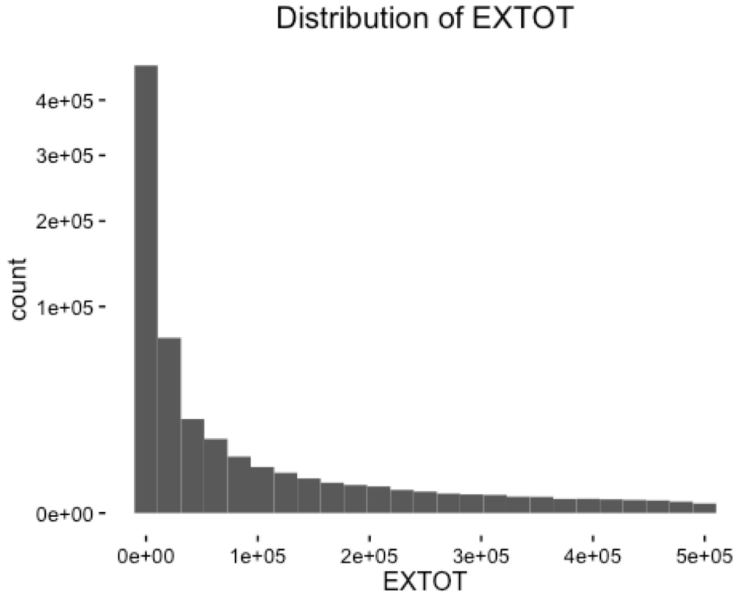
Field Name	Description
AVTOT	continuous with metric, Current year's total market value



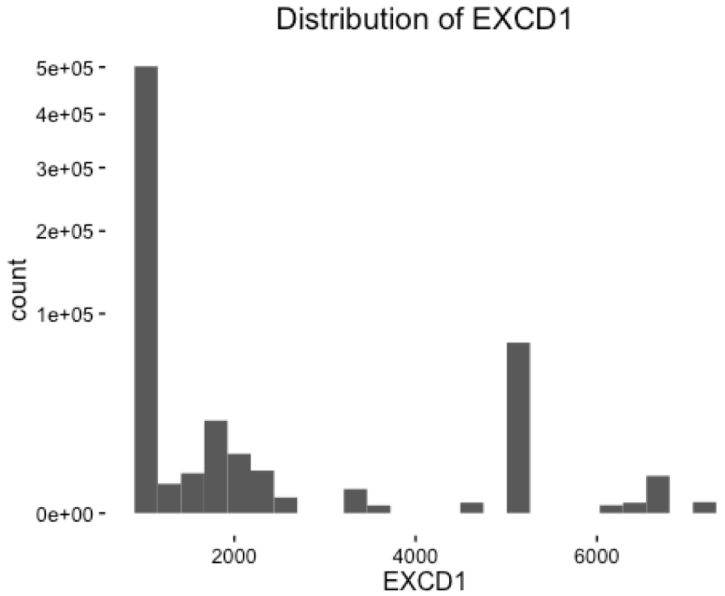
Field Name	Description
EXLAND	continuous with metric, Current Transitional Exempt Land Value



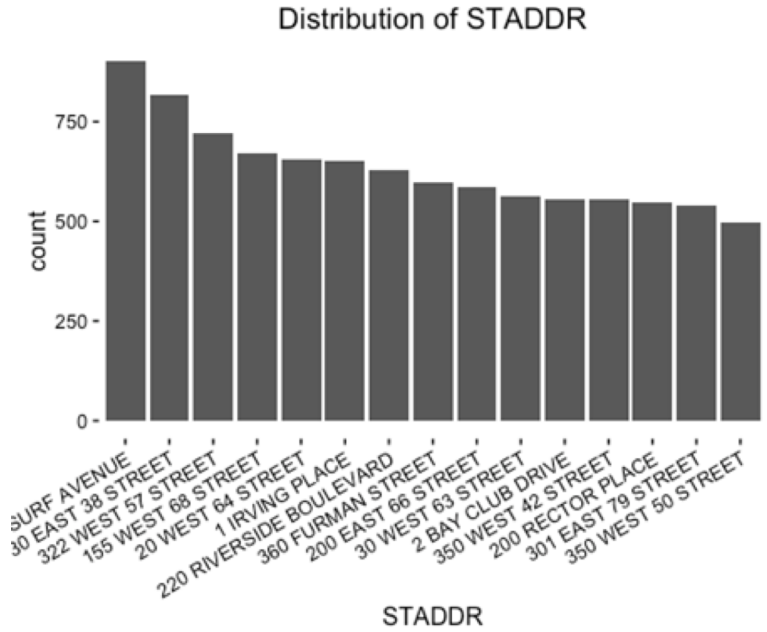
Field Name	Description
EXTOT	continuous with metric, Current Transitional Exempt Total Value



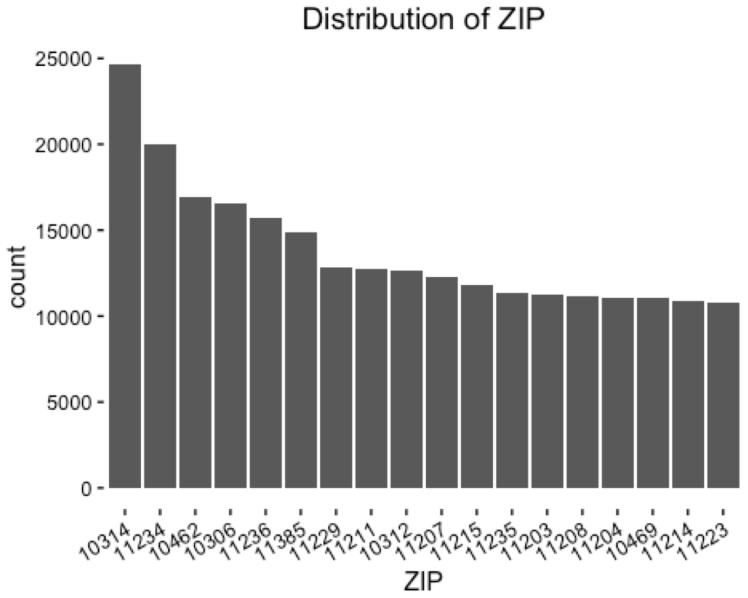
Field Name	Description
EXCD1	continuous with metric



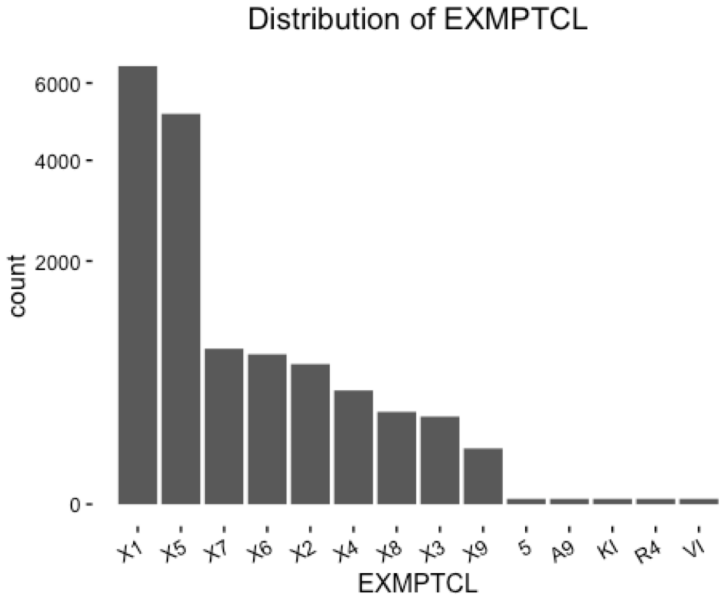
Field Name	Description
STADDR	character without metric, Street name for the property



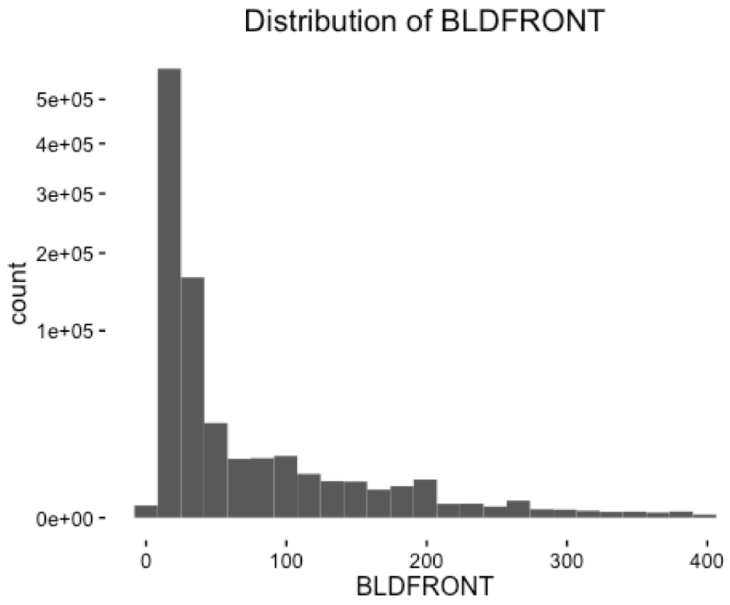
Field Name	Description
ZIP	categorical with metric, Postal Zip code of the property



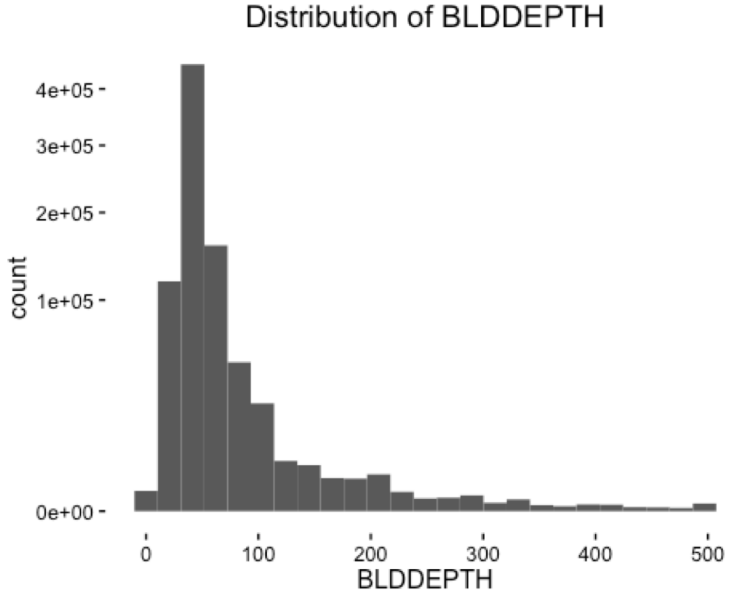
Field Name	Description
EXMPTCL	Categorical with metric, Exempt Class used for fully exempt properties only



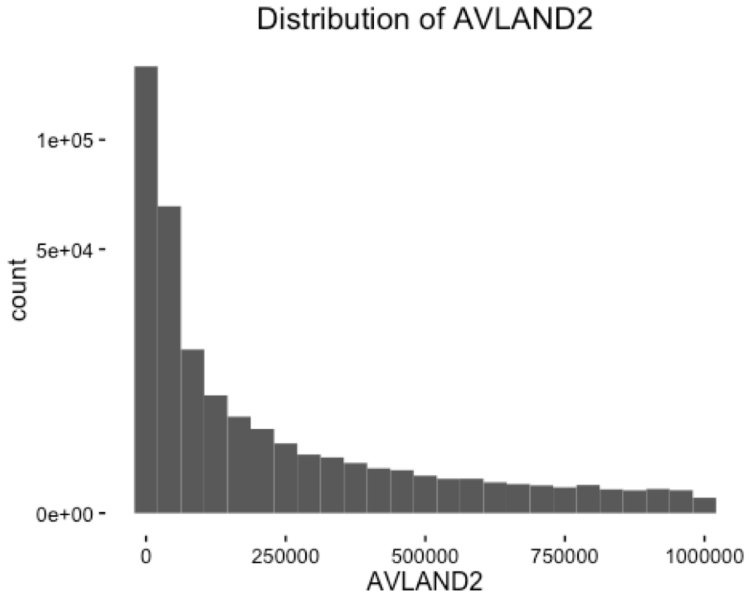
Field Name	Description
BLDFRONT	continuous with metric



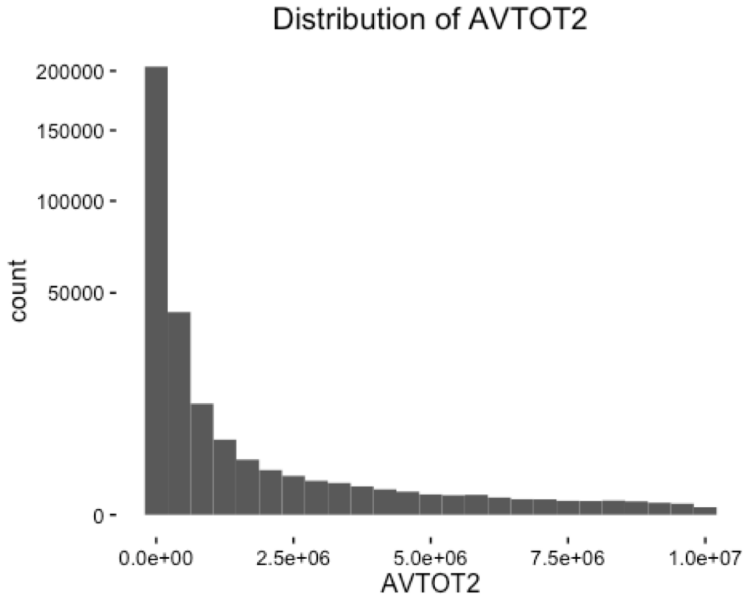
Field Name	Description
BLDDEPTH	continuous with metric



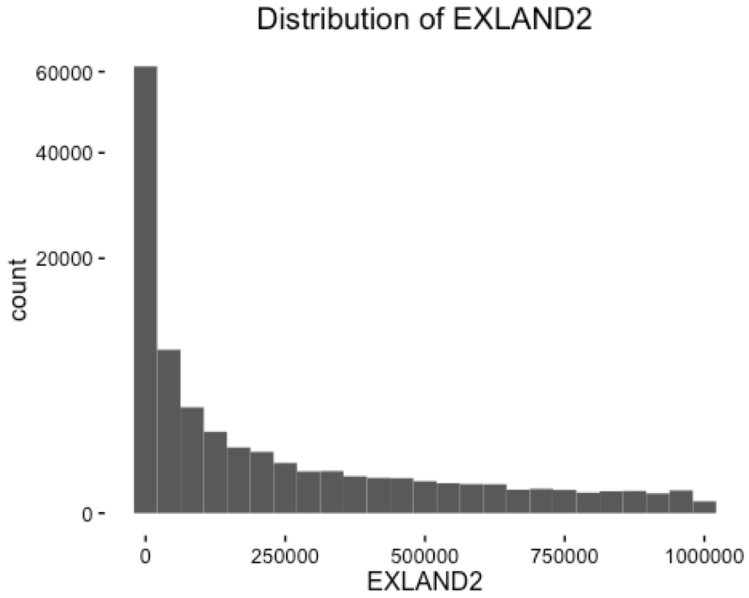
Field Name	Description
AVLAND2	continuous with metric



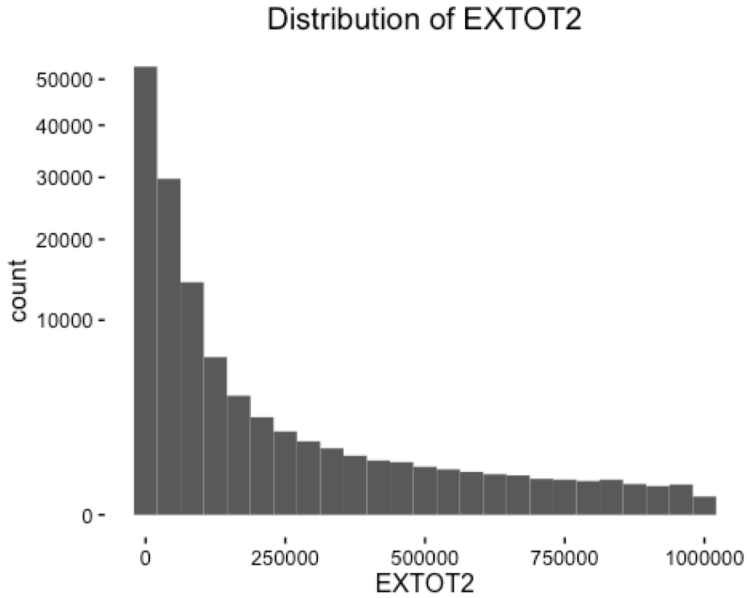
Field Name	Description
AVTOT2	continuous with metric



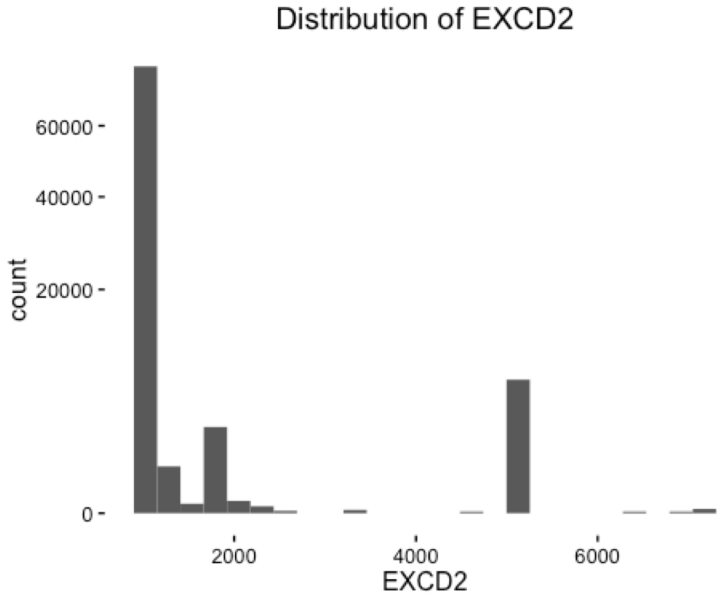
Field Name	Description
EXLAND2	continuous with metric



Field Name	Description
EXTOT2	continuous with metric



Field Name	Description
EXCD2	continuous with metric



Field Name	Description
PERIOD	Categorical no metric, Indicator for Change Period of the File

Field Name	Description
YEAR	categorical no metric, value = 2010/11, no missing data

Field Name	Description
VALTYPE	categorical no metric, Value = AC-TR, no missing data