THE FRAUD DETECTION OF NEW YORK PROPERTIES

Contents

1. Executive Summary

This report provides an analysis of the 2018 New York Property Valuation and Assessment Data, which contains 1,070,994 New York City property assessments. A total of 32 fields (columns) are provided by the dataset to describe each assessment further. The goal of the analysis is to identify the problematic property assessments that could possibly lead to a variety of property value frauds.

To more effectively gauge the fraud risk, a fraud score, which takes into account factors such as the monetary value and structure of a property, has been issued for each property assessment in the dataset. This report provides a detailed explanation of the five steps that generate the fraud score. The summary of the five steps are as follow:

(1) Data cleaning, which fills in the necessary missing fields with reasonable numbers
(2) Variable creation. 45 new variables are created for later use
(3) Z-scaling and Principal Component Analysis (PCA). Z-scaling is used to scale the 45 new variables. PCA is used for dimensionality reduction. Eight principal components are selected and Z-scaled again for model application.
(4) Model application. Two unsupervised models, Heuristic Algorithm and Autoencoder, are used to generate two scores (S1 and S2) for each property in the dataset.
(5) Quantile Binning, which combines S1 and S2 to produce the final fraud score for each property.

This report also investigates the potential misstatement of the assessments associated with the 10 highest-scoring properties. The investigation shows that the high fraud scores of the 10 property assessments are mainly due to some of the following reasons:

- FULLVAL, AVTOT, and AVLAND are unusually high
- Public property that is missing important information
- Property structure does not match property type on file

In addition to the top 10 highest-scoring property assessments, this report defines the top 1% records in the dataset that have the highest fraud score as potential fraudulent records. A breakdown of the top 1% records shows that:

- The majority of the top 1% records are located in Manhattan and Staten Island
- Most of the top 1% records are from Tax Class 2 and Tax Class 4

2. Description of Data

The dataset used for analysis is the 2018 New York Property Valuation and Assessment Data. It contains 1,070,994 New York City property assessments. A total of 32 fields (columns) are included in the dataset to provide further information and description of the property assessments.

The New York City Government first released the original property valuation data in 2011, for the purpose of calculating property tax and granting eligible properties exemptions or/and abatements. Since 2011, the dataset has been updated annually. It is collected, recorded, and managed by various New York City departments, including the Department of Building Reporting.

All 32 fields could be further divided into 14 numerical and 18 categorical fields. The two tables listed in this section will summarize the information provided through the 32 fields.

2.1 Numerical Field Table

	Field Name	Field Type	# of Records w/ Value	% Populated	# Unique Values	# Records w/ Zero	Mean	STD	Min	Max
1	LTFRONT	Numerical	1070994	100.00%	1297	169108	3.66E+01	7.40E+01	0.00	1.00E+04
2	LTDEPTH	Numerical	1070994	100.00%	1370	170128	8.89E+01	7.64E+01	0.00	1.00E+04
3	STORIES	Numerical	1014730	94.75%	112	0	5.01E+00	8.37E+00	1.00	1.19E+02
4	FULLVAL	Numerical	1070994	100.00%	109324	13007	8.74E+05	1.16E+07	0.00	6.15E+09
5	AVLAND	Numerical	1070994	100.00%	70921	13009	8.51E+04	4.06E+06	0.00	2.67E+09
6	AVTOT	Numerical	1070994	100.00%	112914	13007	2.27E+05	6.88E+06	0.00	4.67E+09
7	EXLAND	Numerical	1070994	100.00%	33419	491699	3.64E+04	3.98E+06	0.00	2.67E+09
8	EXTOT	Numerical	1070994	100.00%	64255	432572	9.12E+04	6.51E+06	0.00	4.67E+09
9	BLDFRONT	Numerical	1070994	100.00%	612	228815	2.30E+01	3.56E+01	0.00	7.58E+03
10	BLDDEPTH	Numerical	1070994	100.00%	621	228853	3.99E+01	4.27E+01	0.00	9.39E+03
11	AVLAND2	Numerical	282726	26.40%	58592	0	2.46E+05	6.18E+06	3.00	2.37E+09
12	AVTOT2	Numerical	282732	26.40%	111361	0	7.14E+05	1.17E+07	3.00	4.50E+09
13	EXLAND2	Numerical	87449	8.17%	22196	0	3.51E+05	1.08E+07	1.00	2.37E+09
14	EXTOT2	Numerical	130828	12.22%	48349	0	6.57E+05	1.61E+07	7.00	4.50E+09

Table 1.

2.2 Categorical Field Table

	Field Name	Field Type	# of Records w/ Value	% Populated	# Unique Values	# Records w/ Zero	Most Common Field Name
1	RECORD	Categorical	1070994	100.00%	1070994	0	All Different
2	BBLE	Categorical	1070994	100.00%	1070994	0	All Different
3	B	Categorical	1070994	100.00%	5	0	4
4	BLOCK	Categorical	1070994	100.00%	13984	0	3944
5	LOT	Categorical	1070994	100.00%	6366	0	1
6	EASEMENT	Categorical	4636	0.43%	13	0	E
7	BLDGCL	Categorical	1070994	100.00%	200	0	R4
8	TAXCLASS	Categorical	1070994	100.00%	11	0	1
9	EXT	Categorical	354305	33.08%	4	0	G
10	ZIP	Categorical	1041104	97.21%	197	0	10314
11	EXMPTCL	Categorical	15579	1.45%	15	0	X1
12	PERIOD	Categorical	1070994	100.00%	1	0	FINAL
13	VALTYPE	Categorical	1070994	100.00%	1	0	AC-TR
14	EXCD1	Categorical	638488	59.62%	130	0	1017
15	EXCD2	Categorical	92948	8.68%	61	0	1017

16	OWNER	Categorical	1039249	97.04%	863348	0	PARKCHESTER PRESERVAT
17	STADDR	Categorical	1070318	99.94%	839281	0	501 SURF AVENUE
18	YEAR	Date	1070994	100.00%	1	0	2010/11

Table 2.

Nine fields are selected to create the 45 variables: (1) FULLVAL, (2) AVLAND, (3) AVTOT, (4) LTFRONT, (5) LTDEPTH, (6) BLDFRONT, (7) BLDDEPTH, (8) STORIES, and (9) ZIP. A more detailed introduction of the nine fields, together with the rest 23 fields will be provided in the data quality report (DQR) attached in **Appendix 1**.

2.3 A Brief Introduction of the Nine Fields

Field 14

- Field Name: FULLVAL
- Field Type: Numerical, continuous
- Description: Total market value of property

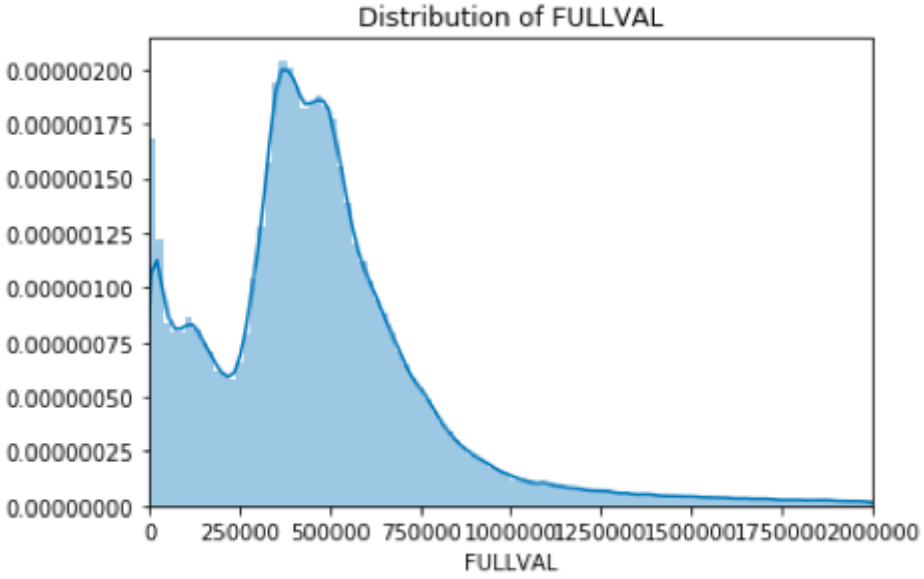


Figure 1.

Field 15

- Field Name: AVLAND
- Field Type: Numerical, continuous
- Description: Actual land value

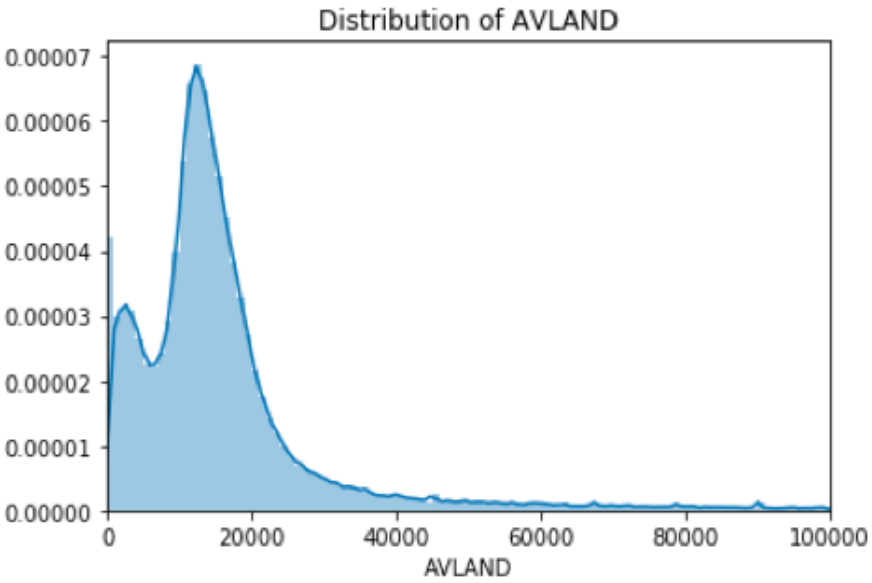


Figure 2.

Field 16

- Field Name: AVTOT
- Field Type: Numerical, continuous
- Description: Actual total value

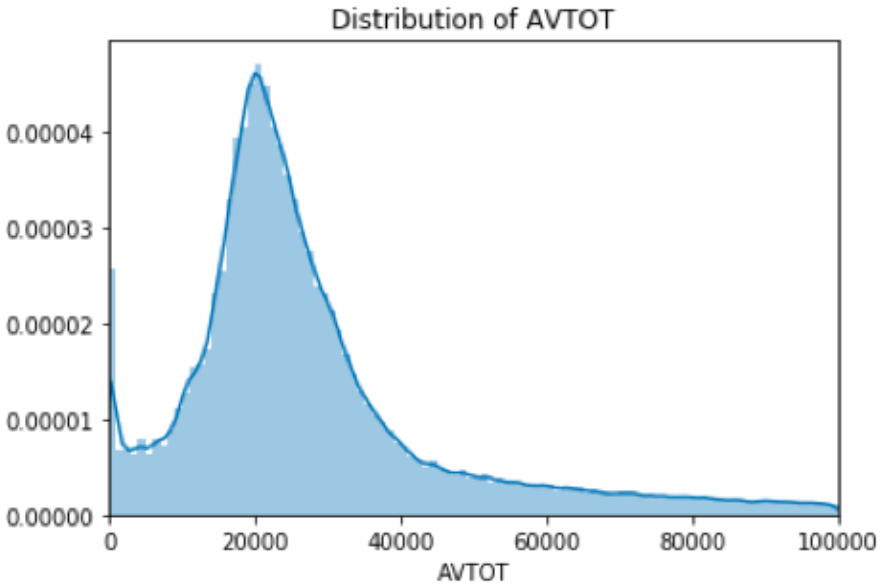


Figure 3.

Field 10

- Field Name: LTFRONT
- Field Type: Numerical, continuous
- Description: Lot frontage in feet

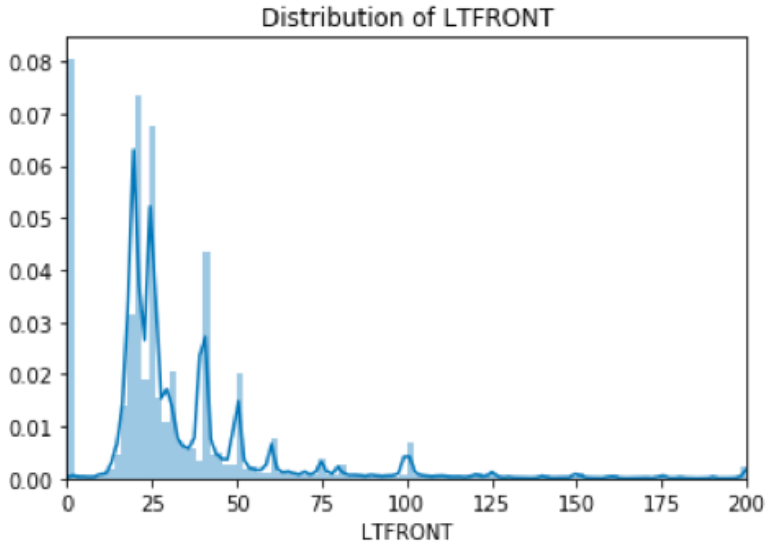


Figure 4.

Field 11

- Field Name: LTDEPTH
- Field Type: Numerical, continuous
- Description: Lot depth in feet

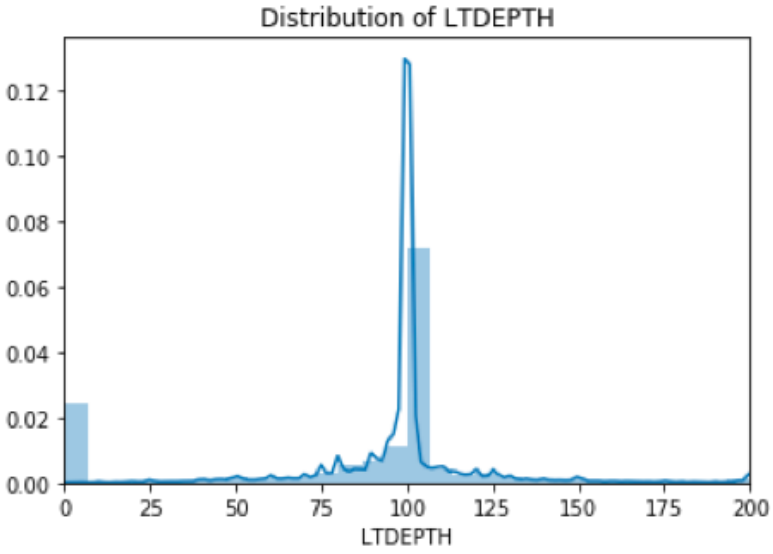


Figure 5.

Field 23

- Field Name: BLDFRONT
- Field Type: Numerical, continuous
- Description: Building frontage in feet

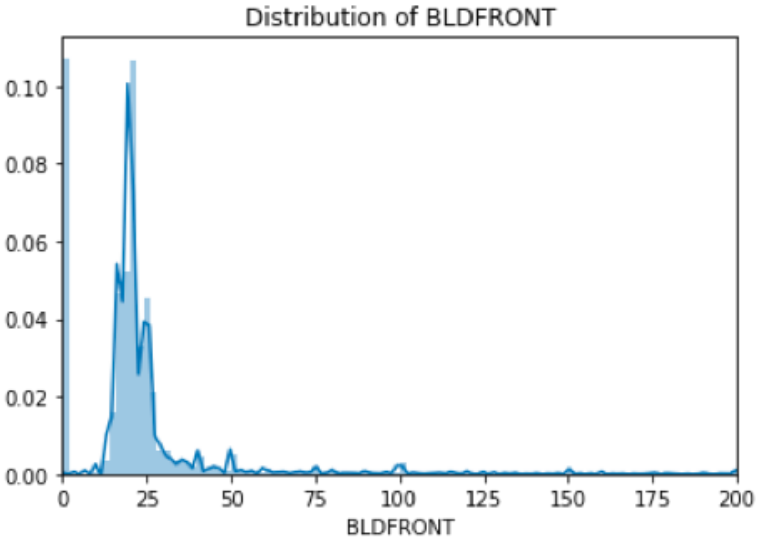


Figure 6.

Field 24

- Field Name: BLDDEPTH
- Field Type: Numerical, continuous
- Description: Building depth in feet

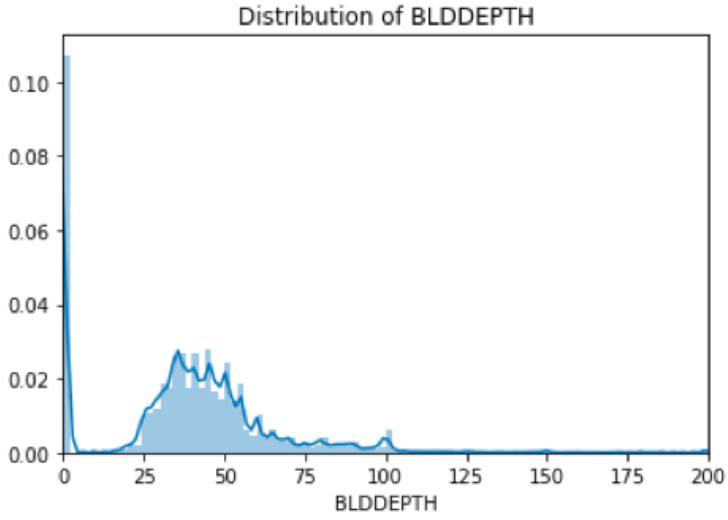


Figure 7.

Field 13

- Field Name: STORIES
- Field Type: Numerical, continuous
- Description: The number of stores in building
- Number of Unique Values: 111

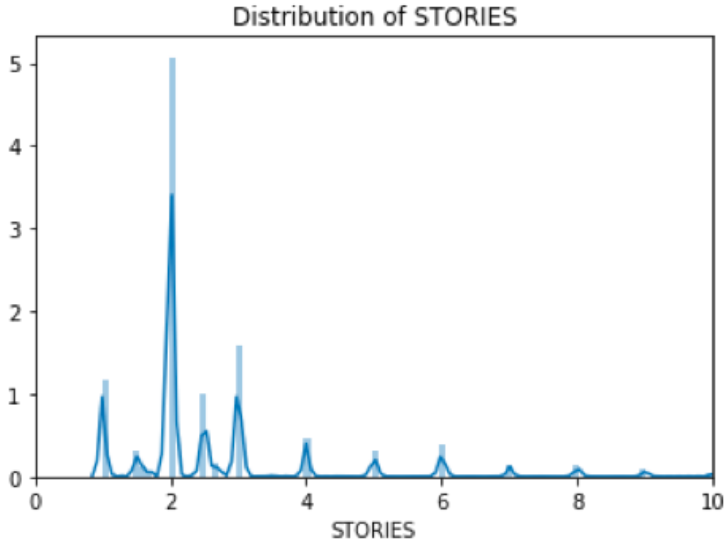


Figure 8.

Field 21

- Field Name: ZIP
- Field Type: Categorical
- Description: Postal zip code of property
- Number of Unique Values: 197

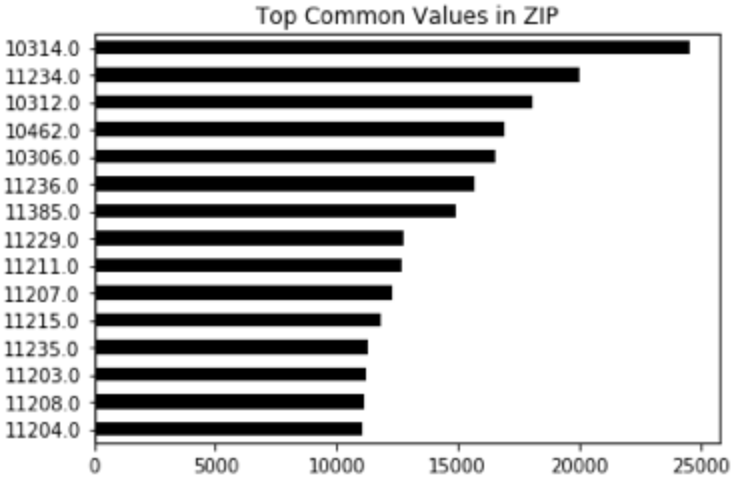


Figure 9.

3. Data Cleaning

In this section of the report, the methods used to fill in the missing values associated with the nine fields will be introduced.

3.1 BLDDPETH, BLDFRONT, LTDEPTH, and LTFRONT

Missing Data and % Populated

Both LTDEPTH and LTFRONT fields contain 15.8% zero value, with no NA value; BLDDPETH and BLDFRONT fields contain 21.4% zero value respectively, with no NA value.

Methodology

Aggregate by BLDGCL (represents for building class), and fill in zero value with median value in each class. Change records with missing value under classes with less than 10 records into NA. Aggregate by TAXCLASS (represents for property tax class), and fill in NA value with median value in each class.

3.2 ZIP

Missing Data and % Populated

ZIP field contains 2.8% of NA value, with no zero value.

Methodology

Sort dataset by BBLE (which is the concatenation of borough code, block code, Unique # within borough/block, easement), to ensure the proximity of missing zip codes to those nearby geographically. Fill NA values with previous BBLE's zip

3.3 STORIES

Missing Data and % Populated

STORIES field contains 5.3% of NA values, with no zero value.

Methodology

Aggregate by BLDGCL (represents for building class), and fill in NA value with median value in each class. Change records with missing value under classes with less than 10 records into NA. Aggregate by TAXCLASS (represents for property tax class), and fill in NA value with median value in each class.

3.4 FULLVAL, AVTOT, and AVLAND

Missing Data and % Populated

FULLVAL, AVTOT, and AVLAND fields contain 1.2% of zero values respectively, with no NA value.

Methodology

Aggregate by ZIP, TAXCLASS, and fill 0 by median value of each class. For the missing value under classes with less than 10 records, aggregate by B (represents for borough code), TAXCLASS, and fill 0 by median value of each class. Change records with missing value under classes with less than 10 records into NA.

After the previous steps, there is still over 600 records with NA values, these NA values come from two parts:

(1) records with zero value whose corresponding value has less than 10 records
(2) all data records for the corresponding B and TAXCLASS aggregation group are zero

Aggregate by B, ZIP, calculate LOT_AREA (which equals to LTFRONT * LTDEPTH) and a ratio (which equals to FULLVAL/ LOT_AREA), and fill NA values by multiplying LOT_AREA by ratio in corresponding group. Since for each TAXCLASS, LOT_AREA and field with missing value has high correlation (up to 99%), therefore, here LOT_AREA works as a proxy for TAXCLASS.

4. Variable Creation

4.1 The Logic of Variables Creation:

The basic logic lies in that we need to identify the anomalous property records, so that the best way to measure this abnormality is to detect the properties' 'unit price', which may like the prices of a unit property area. Based on the logic state above, we have the following steps to create variables:

Step 1:

Create variables that can measure the sizes of properties, there are totally 3 that we can get from existing data set:

- LOTAREA = LTFRONT * LTDEPTH
- BLDAREA = BLDFRONT * BLDDEPTH
- BLDVOL = BLDAREA * STORIES

These three variables measure the properties' lot area, building area and building volume.

Step 2:

Calculate 9 variables that can reflect 'unit price', which basically is the property value (including FULLVAL, AVLAND and AVTOT) divided by property area (volume):

- FULLVAL_LOTAREA: FULLVAL ÷ LOTAREA
- FULLVAL_ BLDAREA: FULLVAL ÷ BLDAREA
- FULLVAL_ BLDVOL: FULLVAL ÷ BLDVOL
- AVLAND_ LOTAREA: AVLAND ÷ LOTAREA
- AVLAND_ BLDAREA: AVLAND ÷ BLDAREA
- AVLAND_ BLDVOL: AVLAND ÷ BLDVOL
- AVTOT _ LOTAREA: AVTOT ÷ LOTAREA
- AVTOT _ AVLAND: AVTOT ÷ AVLAND
- AVTOT _ BLDVOL: AVTOT ÷ BLDVOL

Step 3:

For we need to consider the abnormality of each properties' values, and properties' values may differ because of the location and tax class. Therefore, here we need to rescale them according to the location, tax class and all (consider the average level of all the NYC properties). Here we choose the first 3 number of the zip code (zip3), the first 5 number of the zip code (zip5), borough code (B), tax class (TAXCLASS) and all records to group all the properties, and divide that 9 variables by the mean of corresponding groups (groups with same zip3, zip5, borough code or tax class).

The details are listed below.

4.2 The Detailed List of All Variables

Following is the list of all variables:

No.	Variables	Description (scale here also means divide)
1	FULLVAL_LOTAREA_ZIP3	FULLVAL per unit of lot area scaled by the mean of corresponding zip3 groups
2	FULLVAL_LOTAREA_ZIP5	FULLVAL per unit of lot area scaled by the mean of corresponding zip5 groups
3	FULLVAL_LOTAREA_B	FULLVAL per unit of lot area scaled by the mean of corresponding borough code groups
4	FULLVAL_LOTAREA_TAXCLASS	FULLVAL per unit of lot area scaled by the mean of corresponding tax class groups
5	FULLVAL_LOTAREA_ALL	FULLVAL per unit of lot area scaled by the mean of all records
6	FULLVAL_ BLDAREA_ZIP3	FULLVAL per unit of building area scaled by the mean of corresponding zip3 groups
7	FULLVAL_ BLDAREA_ZIP5	FULLVAL per unit of building area scaled by the mean of corresponding zip5 groups
8	FULLVAL_ BLDAREA_B	FULLVAL per unit of building area scaled by the mean of corresponding borough code groups
9	FULLVAL_ BLDAREA_TAXCLASS	FULLVAL per unit of building area scaled by the mean of corresponding tax class groups
10	FULLVAL_ BLDAREA_ALL	FULLVAL per unit of building area scaled by the mean of all records
11	FULLVAL_ BLDVOL_ZIP3	FULLVAL per unit of building volume scaled by the mean of corresponding zip3 groups
12	FULLVAL_ BLDVOL_ZIP5	FULLVAL per unit of building volume scaled by the mean of corresponding zip5 groups
13	FULLVAL_ BLDVOL_B	FULLVAL per unit of building volume scaled by the mean of corresponding borough code groups
14	FULLVAL_ BLDVOL_TAXCLASS	FULLVAL per unit of building volume scaled by the mean of corresponding tax class groups
15	FULLVAL_ BLDVOL_ALL	FULLVAL per unit of building volume scaled by the mean of all records

16	AVLAND_ LOTAREA_ZIP3	AVLAND per unit of lot area scaled by the mean of corresponding zip3 groups
17	AVLAND_ LOTAREA_ZIP5	AVLAND per unit of lot area scaled by the mean of corresponding zip5 groups
18	AVLAND_ LOTAREA_B	AVLAND per unit of lot area scaled by the mean of corresponding borough code groups
19	AVLAND_ LOTAREA_TAXCLASS	AVLAND per unit of lot area scaled by the mean of corresponding tax class groups
20	AVLAND_ LOTAREA_ALL	AVLAND per unit of lot area scaled by the mean of all records
21	AVLAND_ BLDAREA_ZIP3	AVLAND per unit of building area scaled by the mean of corresponding zip3 groups
22	AVLAND_ BLDAREA_ZIP5	AVLAND per unit of building area scaled by the mean of corresponding zip5 groups
23	AVLAND_ BLDAREA_B	AVLAND per unit of building area scaled by the mean of corresponding borough code groups
24	AVLAND_ BLDAREA_TAXCLASS	AVLAND per unit of building area scaled by the mean of corresponding tax class groups
25	AVLAND_ BLDAREA_ALL	AVLAND per unit of building area scaled by the mean of all records
26	AVLAND_ BLDVOL_ZIP3	AVLAND per unit of building volume scaled by the mean of corresponding zip3 groups
27	AVLAND_ BLDVOL_ZIP5	AVLAND per unit of building volume scaled by the mean of corresponding zip5 groups
28	AVLAND_ BLDVOL_B	AVLAND per unit of building volume scaled by the mean of corresponding borough code groups
29	AVLAND_ BLDVOL_TAXCLASS	AVLAND per unit of building volume scaled by the mean of corresponding tax class groups
30	AVLAND_ BLDVOL_ALL	AVLAND per unit of building volume scaled by the mean of all records
31	AVTOT _ LOTAREA_ZIP3	AVTOT per unit of lot area scaled by the mean of corresponding zip3 groups
32	AVTOT _ LOTAREA_ZIP5	AVTOT per unit of lot area scaled by the mean of corresponding zip5 groups

33	AVTOT _ LOTAREA_B	AVTOT per unit of lot area scaled by the mean of corresponding borough code groups
34	AVTOT _ LOTAREA_TAXCLASS	AVTOT per unit of lot area scaled by the mean of corresponding tax class groups
35	AVTOT _ LOTAREA_ALL	AVTOT per unit of lot area scaled by the mean of all records
36	AVTOT _ AVLAND_ZIP3	AVTOT per unit of building area scaled by the mean of corresponding zip3 groups
37	AVTOT _ AVLAND_ZIP5	AVTOT per unit of building area scaled by the mean of corresponding zip5 groups
38	AVTOT _ AVLAND_B	AVTOT per unit of building area scaled by the mean of corresponding borough code groups
39	AVTOT _ AVLAND_TAXCLASS	AVTOT per unit of building area scaled by the mean of corresponding tax class groups
40	AVTOT _ AVLAND_ALL	AVTOT per unit of building area scaled by the mean of all records
41	AVTOT _ BLDVOL_ZIP3	AVTOT per unit of building volume scaled by the mean of corresponding zip3 groups
42	AVTOT _ BLDVOL_ZIP5	AVTOT per unit of building volume scaled by the mean of corresponding zip5 groups
43	AVTOT _ BLDVOL_B	AVTOT per unit of building volume scaled by the mean of corresponding borough code groups
44	AVTOT _ BLDVOL_TAXCLASS	AVTOT per unit of building volume scaled by the mean of corresponding tax class groups
45	AVTOT _ BLDVOL_ALL	AVTOT per unit of building volume scaled by the mean of all records

Table 3.

5. Dimensionality Reduction

5.1 Z-Scale

Before doing Principal Component Analysis, we need to rescale all the variables because the PCA will set the original coordinate axis of the variable which has the largest variance. If we do not rescale all these variables, the initial data will annoy the PCA. Here we choose Z-Scale to rescale all the variables, and the z-score can be calculated as below:

$$z = \frac{x - \mu}{\sigma}$$

5.2 PCA (Principal Component Analysis)

The PCA (Principal Component Analysis) is a widely used dimensionality reducing method, whose basic rule is to find the variable with the largest variance each time and set that variable's direction as the coordinate system, then rotate the coordinate system to be orthogonal and find the variable with the second largest variance.

The picture below shows the process:

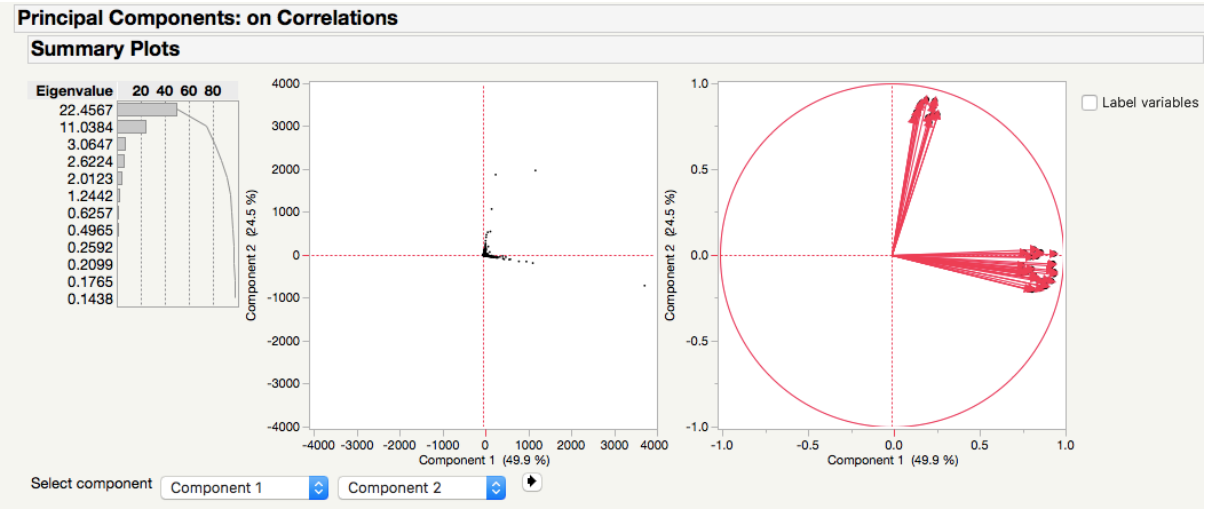


Figure 10.

In this way, we can find almost all the variables can be represented (expressed) by the first several PCs, which means after several rotations, the variance of remaining variables will be extremely small.

The chart below shows the total variance be explained (to what extent our variables can be explained):

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	22.457	49.904	49.904	22.457	49.904	49.904
2	11.038	24.53	74.434	11.038	24.53	74.434
3	3.065	6.81	81.244	3.065	6.81	81.244
4	2.622	5.828	87.071	2.622	5.828	87.071
5	2.012	4.472	91.543	2.012	4.472	91.543
6	1.244	2.765	94.308	1.244	2.765	94.308
7	0.626	1.391	95.699	0.626	1.391	95.699
8	0.497	1.103	96.802	0.497	1.103	96.802
9	0.259	0.576	97.378			
10	0.21	0.466	97.845			
11	0.176	0.392	98.237			
12	0.144	0.319	98.556			
13	0.129	0.288	98.844			
14	0.105	0.234	99.078			
15	0.084	0.187	99.265			
16	0.062	0.137	99.402			
17	0.046	0.103	99.505			
18	0.039	0.086	99.591			
19	0.035	0.078	99.669			
20	0.033	0.073	99.742			
21	0.021	0.046	99.788			
22	0.018	0.039	99.827			
23	0.014	0.03	99.858			
24	0.011	0.025	99.883			
25	0.009	0.019	99.902			
26	0.008	0.017	99.919			
27	0.007	0.016	99.935			
28	0.006	0.013	99.948			
29	0.005	0.01	99.958			
30	0.004	0.009	99.967			
31	0.003	0.007	99.974			
32	0.003	0.007	99.98			
33	0.003	0.006	99.986			
34	0.002	0.004	99.99			
35	0.001	0.003	99.993			
36	0.001	0.002	99.995			
37	0.001	0.001	99.997			
38	0.001	0.001	99.998			
39	0	0.001	99.999			
40	0	0	99.999			
41	0	0	99.999			
42	0	0	100			
43	0	0	100			
44	2.84E-05	6.31E-05	100			
45	2.17E-05	4.83E-05	100			
Extraction Method: Principal Component Analysis.						

Table 4.

We can easily find that the first eight PCs can explain 96.802% of all our variables, which is almost an ideal result and we need not to take more PCs because increasing number of PCs does not necessarily increase the percent of variance explained.

The scree plot below shows how the eigenvalue of our PCs, which means how important or to what extent they can represent our variables:

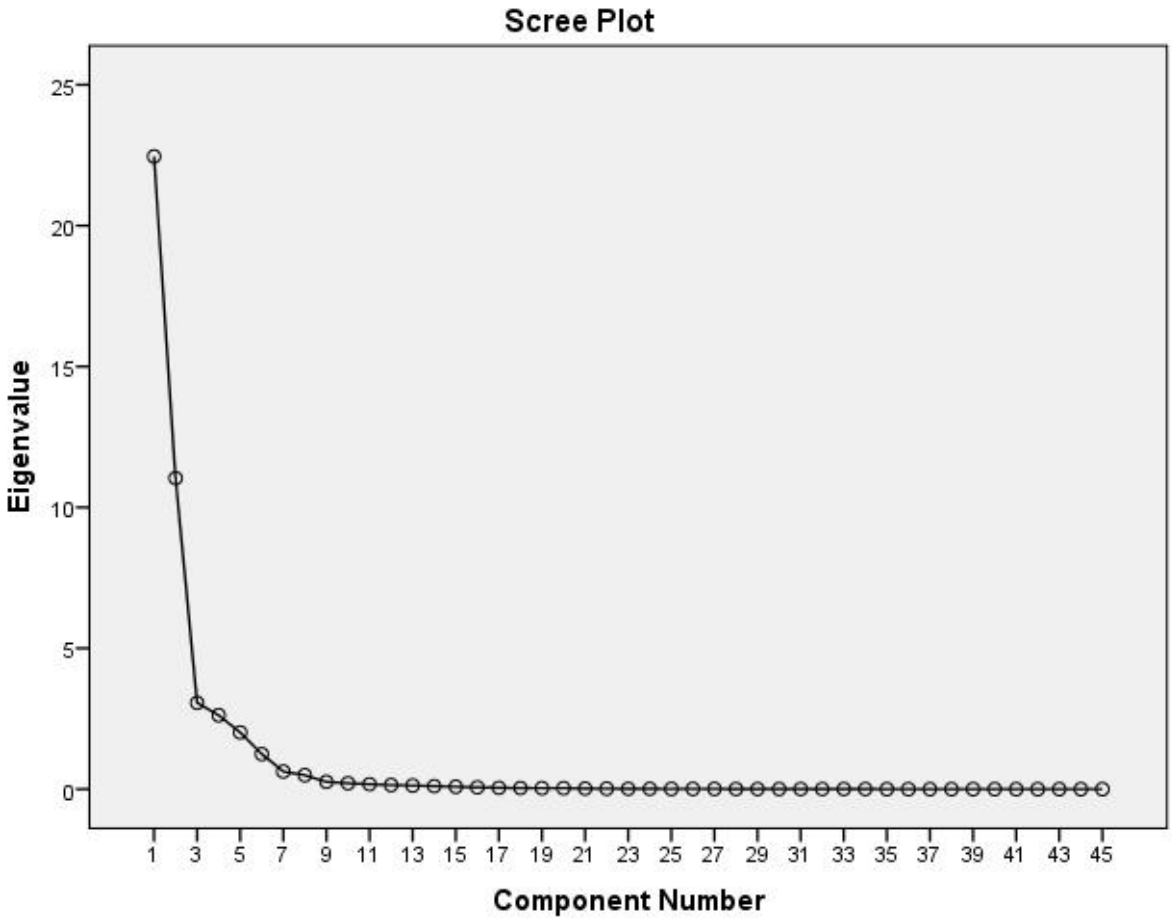


Figure 11.

5.3 Second Z-Scale

After PCA, we need to Z-Scale all the PCs again because the process of rotating coordinate system and making all the PCs orthogonal will set them different weight and here we want them to be equally important. Through the second Z-Scale, we now have all the PCs equally weighted and we can easily build the fraud score in the next section.

6. Algorithms:

To get our designated fraud score to identify fraud cases, we would like to integrate two fraud scores derived from two ways of modeling. Here are two methods applied to calculate fraud score:

(1) Heuristic algorithm (linear method)
(2) Autoencoder (non-linear method)

6.1 Method 1: Heuristic algorithm

To calculate the fraud score given this linear method, we applied the following formula:

$$H_i = \left(\sum_k \left| z_k^i \right|^n \right)^{\frac{1}{n}} \qquad where\ k = 1, 2, \dots, 8$$

since it calculates the distance each PC is away from the mean 0 so that we could detect anomalous records which would have higher H score.

Our team chose the n to be 2 and the formula would be updated as following:

$$H_i = \left(\sum_k \left| z_k^i \right|^2 \right)^{\frac{1}{2}} \qquad where\ k = 1, 2, \dots, 8$$

Using this formula, every record got its corresponding H which stands for its fraud score under the first method.

6.2 Method 2: Autoencoder

Besides employing linear algorithm to see which record might be fraud, we wanted to use a deep learning algorithm to detect the anomalous records. We applied autoencoder model. The reason we chose autoencoder is that this model would learn how records behave generally and once there is some record behaving differently, it might be a fraud case. We applied autoencoder to train a model with our z-scaled PCs created before as training dataset. The logic of autoencoder could be briefly summarized as following graph:

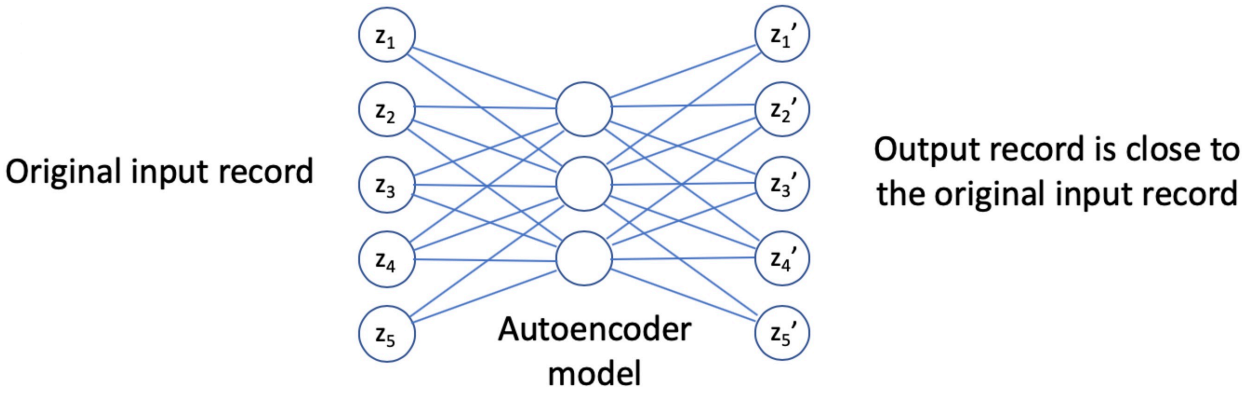


Figure 12.

Our inputted our z-scaled PCs went into Autoencoder to train a model. Then we applied this model into these z-scaled PCs to output certain representation values based on this model as it learns from the original data. Through comparing the original data and output data given the model we trained, we would detect anomalous records. Here we check the anomaly by checking the Reconstruction

Mean Square Error which would be our fraud score for this case: higher the corresponding Reconstruction MSE is, harder it is for the autoencoder to represent the output for our interested record, which indicates that this record might be a potential fraud case.

6.3 Fraud Score Integration

After computing two fraud scores based on two modeling methods, we would like to integrate to get a unique fraud score for each record through quantile binning. Through quantile binning for both score values would mitigate the importance or weight of each fraud score to affect the final fraud score. We utilized quantile binning with 1070994 bins where each record would be assigned into a bin. We first sorted our dataset based on our first fraud score derived from the Euclidean Distance in the ascending order and then assigned records into 1070994 bins given this order so that each record would have its own quantile bin noted as qs1. We applied the same process toward the dataset based on the Reconstruction MSE conducted from the autoencoder model and then each record would have its own quantile bin noted as qs2 again based on the second fraud score. After assigning quantile bins twice, we simply integrate these two values together to get our final fraud scores for all 1070994 records:

$$FS_i = qs1_i + qs2_i$$

7. Results

This section of the report discusses the result of the two unsupervised models that were used to generate S1 and S2, which were combined together through Quantile Binning to generate the final fraud score.

7.1 Overall Comparison

	Top 1% Records			**All Records**		
	Mean	Median	Std	Mean	Median	Std
LTFRONT	131	46	360	37	25	74
LTDEPTH	173	100	346	89	100	76
STORIES	10	5	10	5	2	8
FULLVAL	24617600	7000000	110487800	874265	447000	11582430
AVLAND	4240071	513000	40078720	85068	13678	4057260
AVTOT	10646170	1440000	67027270	227238	25340	6877529

Table 5.

We select top 1% records as our potential fraudulent records, which has 10710 records in total and compute mean, median and standard deviation of selected numerical fields. Comparing with the original records as shown in the table above, mean and median of potential fraudulent records

(10710 records) are much higher than those of fields in original records. Moreover, potential fraudulent records have a larger standard deviation. Specifically, for AVTOT, median from potential fraudulent records is about 57 times higher than that of original records.

The top owners of potential fraudulent records are mostly government and public authorities. Top 3 owners include Park and Recreation, City of New York and CNY/NYCTA. When only considering commercial corporations, top 3 owners are National Pass Rr Corp, Consolidated Edison C, and Penn Central Company.

The map below summarizes the count of potential fraudulent records in each zip code in New York City. Top 1% records are clustered in B1 (Manhattan) and B5 (Staten Island), which are Manhattan and Staten Island. Specifically, area of B5 and BLOCK 2620, area of B5 and BLOCK 2859 and area of B1 and BLOCK 1405 are top 3 areas with 67, 50, 48 counts respectively:

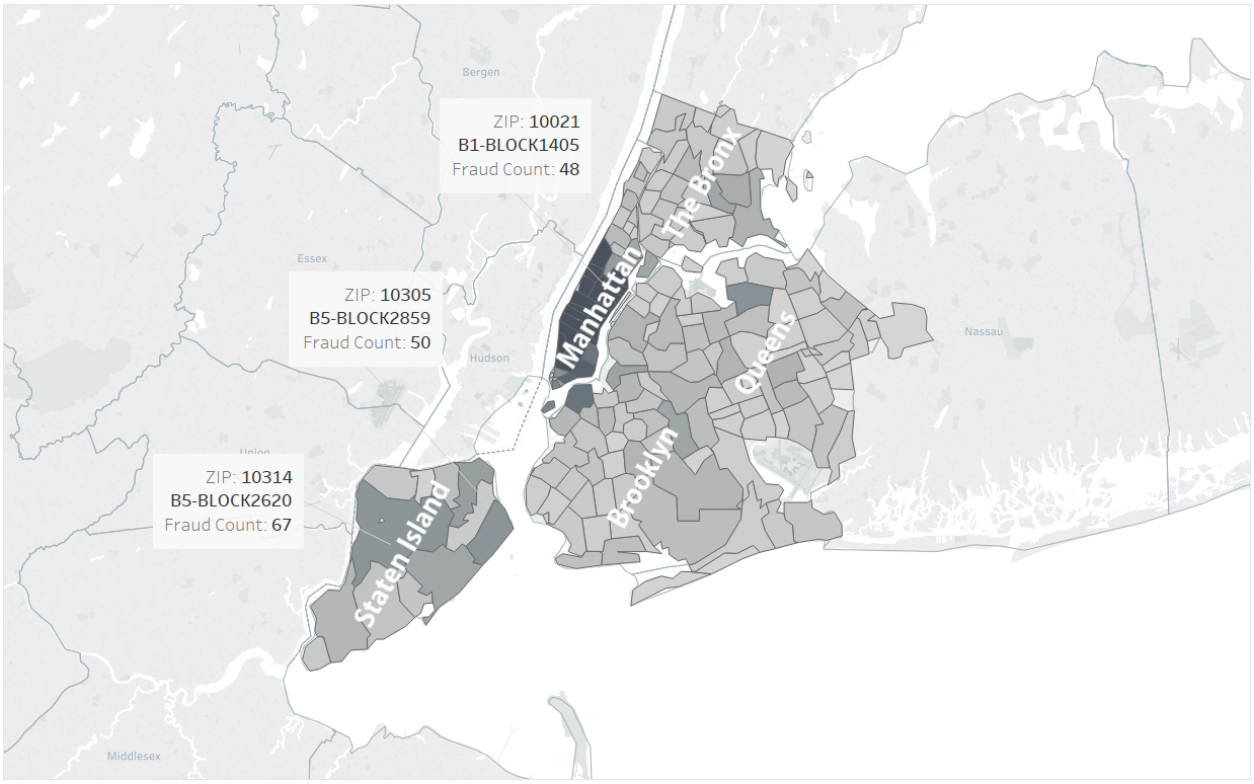


Figure 13.

7.2 Top 10 Records

Table 6 and *Table 7* provides an overview of the top 10 high-scoring records:

RECORD	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP	BLDFRONT	BLDDEPTH
632816	157	95	1	2.93E+06	1.32E+06	1.32E+06	11373	1	1
1067360	1	1	2	8.36E+05	2.88E+04	5.02E+04	10307	36	45
565392	117	108	NA	4.33E+09	1.95E+09	1.95E+09	NA	0	0
585118	298	402	20	3.44E+06	1.55E+06	1.55E+06	11101	1	1
132749	2	1	NA	0.00E+00	0.00E+00	0.00E+00	10025	0	0
85886	4000	150	1	7.02E+07	3.15E+07	3.16E+07	NA	8	8
585120	139	342	20	2.15E+06	9.68E+05	9.68E+05	NA	1	1
585439	94	165	10	3.71E+06	2.52E+05	1.67E+06	11101	1	1
690833	610	534	3	2.42E+08	1.04E+08	1.09E+08	11385	20	20
935158	136	132	8	1.04E+06	2.36E+05	4.68E+05	10301	1	1

Table 6.

RECORD	B	BLOCK	LOT	OWNER	TAXCLASS	BLDGCL
632816	4	1842	1	864163 REALTY, LLC	2	D9
1067360	5	7853	85		1	B2
565392	3	8590	700	U S GOVERNMENT OWNRD	4	V9
585118	4	420	1	NEW YORK CITY ECONOMI	4	O3
132749	1	1875	46	CNY/NYCTA	3	U7
85886	1	1254	10	PARKS AND RECREATION	4	Q1
585120	4	420	101		4	O3
585439	4	459	5	11-01 43RD AVENUE REA	4	H9
690833	4	3866	70	PARKS AND RECREATION	4	Q1
935158	5	13	60	RICH-NICH REALTY, LLC	2	D3

Table 7.

Figure 13 reveals that most of potential fraudulent records are in B1 and B5 (Manhattan and Staten Island). However, *Figure 14* below demonstrates that 50% of the 10 highest-scoring records are from B4 (Queens).

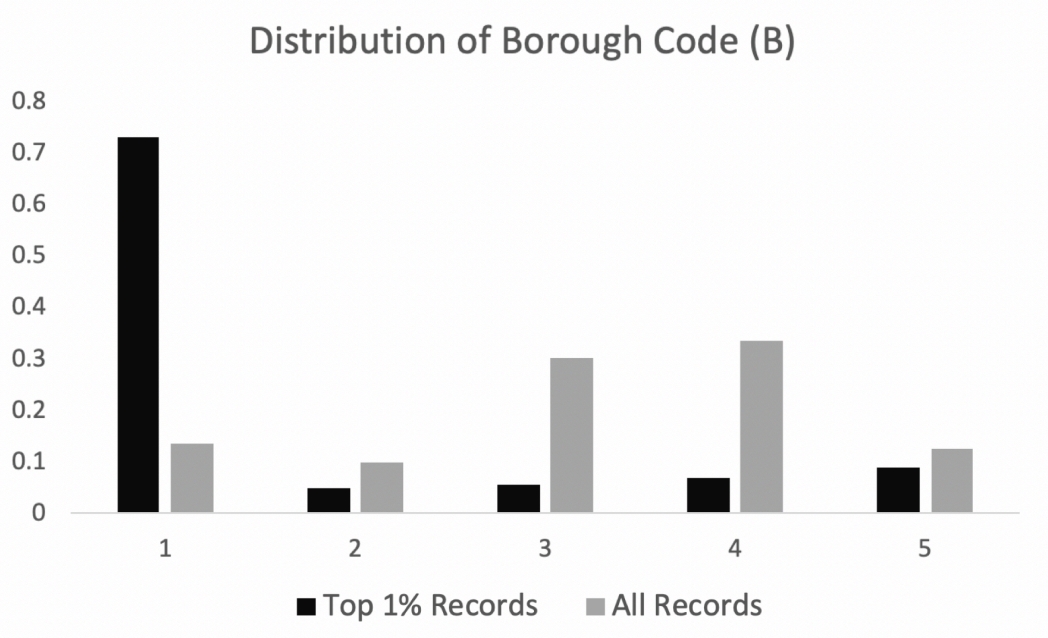


Figure 14.

Figure 15 below shows that top 1% records are clustered in TAXCLASS2 and TAXCLASS4, which are residential property with more than 3 units and all other real property including office buildings, factories, and etc. From our top 10 records, 60% are from TAXCLASS4, indicating fraudsters tend to classify property under this tax class.

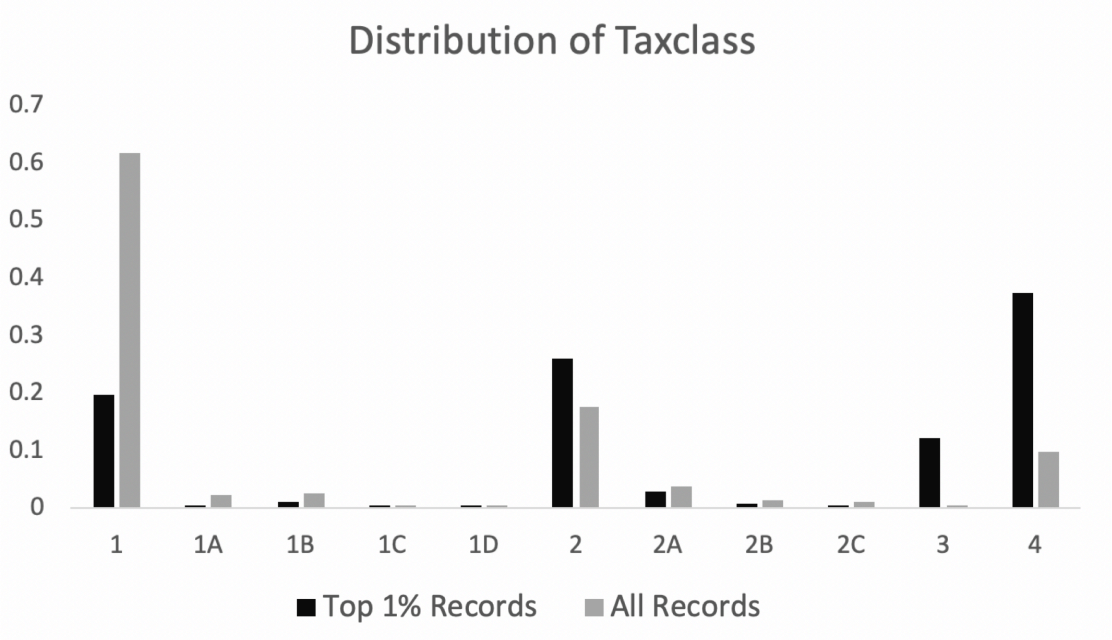


Figure 15.

7.3 Top 10 Records Breakdown

We used the following resources to locate each record and check its validity regarding property size and value:

- Zillow, an online real estate database company
- Apartable, a platform providing information on apartments
- StreetEast, a searchable online database of residential property listings and recorded sales in New York City

Record 632816

According to street address and zip code provided, we found that currently the property is a 7-story luxury apartment with 83 units built in 2015 and lot area matched with the one on record. Since we could not find any information before 2015 to verify the previous property, we looked into TAXCLASS, BLDFRONT, BLDDEPTH and STORIES. BLDFRONT and BLDDEPTH of 1 don't make much sense as TAXCLASS2 indicates it was an apartment. Moreover, STORIES of 1 was quite unusual for an apartment. Thus, this record was suspicious.

Record 1067360

No owner information available. With street address and zip code, we located the place to a residential area with houses of 1-3 stories, which matched TAXCLASS1's description. LOTDEPTH and LOTFRONT both matched the information from Zillow. However, BLDDEPTH and BLDFRONT are 1620 sq. ft compared to 2850 sq. ft from Zillow. In addition, home value at the same area had increased 53% from Jan 2011 to Jan 2019 on average. Using the same growth rate, the market value of this record should be roughly $646K, which was 200K lower than the market value on record. Thus, this record might need further investigation.

Record 565392

The lot area is estimated to be 13K sq. ft (0.29 acre). A general street address is provided without mentioning the specific building or house number. Zip code and stories number are not available. Moreover, this record has suspiciously high FULLVAL, AVLAND, and AVTOT. Since we have data of lot area, we researched the land price on Zillow and found lot/land of 0.27 acre for sale in Brooklyn Area is about $2M, far less than $2B on record for AVLAND. We also searched area around Flatbush Avenue and found a lot of 7K sq. ft is $1.4M, on average $200 per sq. ft. The record would have an estimated $2.6M land value, still far less than AVLAND on record. In addition, it was a US government owned property. Thus, all these factors make it look fraudulent.

Record 585118

We can locate this property with the owner, address, and zip code information provided in the dataset. However, BLDFRONT and BLDDEPTH of 1 don't make sense for a building that has 20 stories.

Record 132749

For this record, most of the numerical fields, except LTDEPTH and LTFRONT, are either missing or equal to 0. The address provided by the dataset is quite vague, but we do know that this property is owned by the New York City Transit Authority. It is suspicious that a property owned by a public authority is missing any one of FULLVAL, AVLAND, or AVTOT.

Record 85886

This property is located next to a park near the Joe Dimaggio Highway. It has unusually high FULLVAL, AVLAND and AVTOT, but relatively small BLDFRONT and BLDDEPTH.

Record 585120

Only street address is provided by the dataset for this record. There is no zip code and owner's information. The Building Class of this property indicates that it is a 20-story building. However, the dataset shows that BLDFRONT and BLDDEPTH of this property is only 1.

Record 585439

Using the street address provided in the dataset, 11-01 43 Avenue, and zip code 11101, we successfully located this property in Brooklyn, New York. It is in fact an eleven-story building, as opposed to a ten-story building that was described in the dataset. This property was renovated and became a hotel in 2011. It has a 5,000 sq. ft rooftop. So logically, it should have a relatively large BLDFRONT or/and BLDDEPTH. However, according to the dataset, this property's BLDFRONT and BLDDEPTH both are equal to 1.

Record 690833

This property is a three-story apartment that has lot area of 10,000,000 square feet and build area of 54,170 square feet. However, its LTFRONT (610ft), LTDEPTH (534ft), BLDFRONT (20ft), and BLDDEPTH (20ft) don't match its lot area and building area. Plus, it achieves abnormally high values in all value-related variables including FULLVAL, AVLAND, AVTOT, EXLAND and EXTOT.

Record 935158

This property was remodeled to an eleven-story apartment in 2012. It is suspicious that as an eight-story building, this property's BLDFRONT and BLDDEPTH are both equal to 1. Moreover, it has unusually high FULLVAL, AVLAND and AVTOT.

8. Conclusion

The 2018 New York Property Valuation and Assessment Data provides the information of 1,070,994 New York City property assessments. This report provides a thorough analysis of the 1,070,994 property assessments, with the purpose of identifying the suspicious assessments that might be involved with property value fraud.

Research reveals that fraudsters may manipulate the valuation of a property for financial gain. Hence, three variables that are associated with the monetary value of a property, together with five variables that reflect the structure of a property are combined to generate a comprehensive fraud score for each one of the 1,070,994 property assessments in the dataset. The three variables that are corresponding to the valuation of a property are: (1) FULLVAL – total market value of property, (2) AVTOT – actual total value, and (3) AVLAND – actual land value. The five structure-related variables are: (1) LTFRONT – lot frontage, (2) LTDEPTH – lot depth, (3) BLDFRONT – building frontage, (4) BLDDEPTH – building depth, and (5) STORIES – the number of stories.

45 expert variables were built by using different combinations of the seven variables. After Z-scaling and Principal Component Analysis (PCA), the 45 expert variables were further reduced to eight principal components (PC1, PC2, PC3, …, PC8). The eight principal components were Z-scaled again before plugging into two unsupervised models: (1) Heuristic Algorithm and (2) Autoencoder. The two models generated two scores (S1 and S2), which produced the final fraud score through Quantile Binning.

Based on the ranking of fraud score, 10 top-scoring property assessments were selected for further investigation. The investigation shows that the high fraud scores of the 10 property assessments are mainly due to some of the following reasons:

- FULLVAL, AVTOT, and AVLAND are unusually high, given the specific street address of the property, the property type, and the property structure.
- Public property that is missing important information, such as the specific address and valuation of the property.
- Property structure does not match property type. For example, an apartment building only has one story; a hotel building that has a sizeable rooftop only has a tiny BLDFRONT or/and BLDDEPTH.

Using Fraud score to detect potential property value fraud has been widely used in the financial industry. Fraud score, if well calculated, successfully translates abstract fraud risk into a comprehensive matric that can be well measured and benchmarked.

However, it must be remembered that the analysis introduced by this report only takes into consideration four variables in the dataset that are associated with the monetary value of the 1,070,994 properties. Although, five more variables that reflect the information about property type and structure are also added to the two supervising models, more variables could be factored in to generate a more dynamic and comprehensive fraud score.

It is also worth noting that the selection of the top 10 property assessments at the end of this report was mainly for demonstration purposes. In other words, the top 10 cutoff line was not derived from an academic research. In reality, the determination of cutoff line is a difficult task, which

requires a great deal of domain knowledge. This in fact mirrors a significant feature of most machine learning methods that subject-matter experts are still strongly needed for the final decision-making process.

9. Appendix (Data Quality Report)

DATASET DESCRIPTION

Name of Dataset: NY Property Valuation and Assessment Data
Description: This dataset gives information on New York property valuation and assessment for purpose to calculate property tax, grant eligible properties exemptions and/or abatements. It is provided by Department of Finance and owned by NYC OpenData. Data are updated annually and consist of 32 fields and 1070994 records.
Date created: September 2, 2011
Last updated: September 10, 2018
Link:https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8

SUMMARY OF ALL FIELDS

Numerical Fields (14 Fields)

	Field Name	Field Type	# of Records w/ Value	% Populated	# Unique Values	# Records w/ Zero	Mean	STD	Min	Max
1	LTFRONT	Numerical	1070994	100.00%	1297	169108	3.66E+01	7.40E+01	0.00	1.00E+04
2	LTDEPTH	Numerical	1070994	100.00%	1370	170128	8.89E+01	7.64E+01	0.00	1.00E+04
3	STORIES	Numerical	1014730	94.75%	112	0	5.01E+00	8.37E+00	1.00	1.19E+02
4	FULLVAL	Numerical	1070994	100.00%	109324	13007	8.74E+05	1.16E+07	0.00	6.15E+09
5	AVLAND	Numerical	1070994	100.00%	70921	13009	8.51E+04	4.06E+06	0.00	2.67E+09
6	AVTOT	Numerical	1070994	100.00%	112914	13007	2.27E+05	6.88E+06	0.00	4.67E+09
7	EXLAND	Numerical	1070994	100.00%	33419	491699	3.64E+04	3.98E+06	0.00	2.67E+09
8	EXTOT	Numerical	1070994	100.00%	64255	432572	9.12E+04	6.51E+06	0.00	4.67E+09
9	BLDFRONT	Numerical	1070994	100.00%	612	228815	2.30E+01	3.56E+01	0.00	7.58E+03
10	BLDDEPTH	Numerical	1070994	100.00%	621	228853	3.99E+01	4.27E+01	0.00	9.39E+03
11	AVLAND2	Numerical	282726	26.40%	58592	0	2.46E+05	6.18E+06	3.00	2.37E+09
12	AVTOT2	Numerical	282732	26.40%	111361	0	7.14E+05	1.17E+07	3.00	4.50E+09
13	EXLAND2	Numerical	87449	8.17%	22196	0	3.51E+05	1.08E+07	1.00	2.37E+09
14	EXTOT2	Numerical	130828	12.22%	48349	0	6.57E+05	1.61E+07	7.00	4.50E+09

Categorical Fields (15 Fields)

	Field Name	Field Type	# of Records w/ Value	% Populated	# Unique Values	# Records w/ Zero	Most Common Field Name
1	RECORD	Categorical	1070994	100.00%	1070994	0	All Different
2	BBLE	Categorical	1070994	100.00%	1070994	0	All Different
3	B	Categorical	1070994	100.00%	5	0	4
4	BLOCK	Categorical	1070994	100.00%	13984	0	3944
5	LOT	Categorical	1070994	100.00%	6366	0	1
6	EASEMENT	Categorical	4636	0.43%	13	0	E
7	BLDGCL	Categorical	1070994	100.00%	200	0	R4
8	TAXCLASS	Categorical	1070994	100.00%	11	0	1
9	EXT	Categorical	354305	33.08%	4	0	G
10	ZIP	Categorical	1041104	97.21%	197	0	10314
11	EXMPTCL	Categorical	15579	1.45%	15	0	X1
12	PERIOD	Categorical	1070994	100.00%	1	0	FINAL
13	VALTYPE	Categorical	1070994	100.00%	1	0	AC-TR
14	EXCD1	Categorical	638488	59.62%	130	0	1017
15	EXCD2	Categorical	92948	8.68%	61	0	1017

Other (3 Fields)

	Field Name	Field Type	# of Records w/ Value	% Populated	# Unique Values	# Records w/ Zero	Most Common Field Name
1	OWNER	Categorical	1039249	97.04%	863348	0	PARKCHESTER PRESERVAT
2	STADDR	Categorical	1070318	99.94%	839281	0	501 SURF AVENUE
3	YEAR	Date	1070994	100.00%	1	0	2010/11

FIELD DESCRIPTION

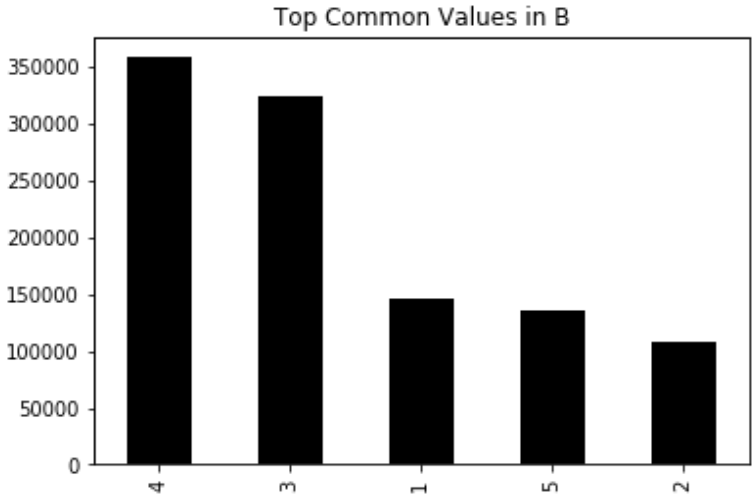
Field 1 RECORD
Field Name: RECORD
Field Type: Categorical
Description: Track data order
Most Common Values: No missing value and all values are different

Field 2 BBLE
Field Name: BBLE
Field Type: Categorical
Description: Concatenation of Borough, Block, Lot and Easement codes.
Most Common Values: No missing value and all values are different

Field 3 B
Field Name: B
Field Type: Categorical
Description: Borough codes

1	MANHATTAN
2	BRONX
3	BROOKLYN
4	QUEENS
5	STATEN ISLAND

Most Common Values:

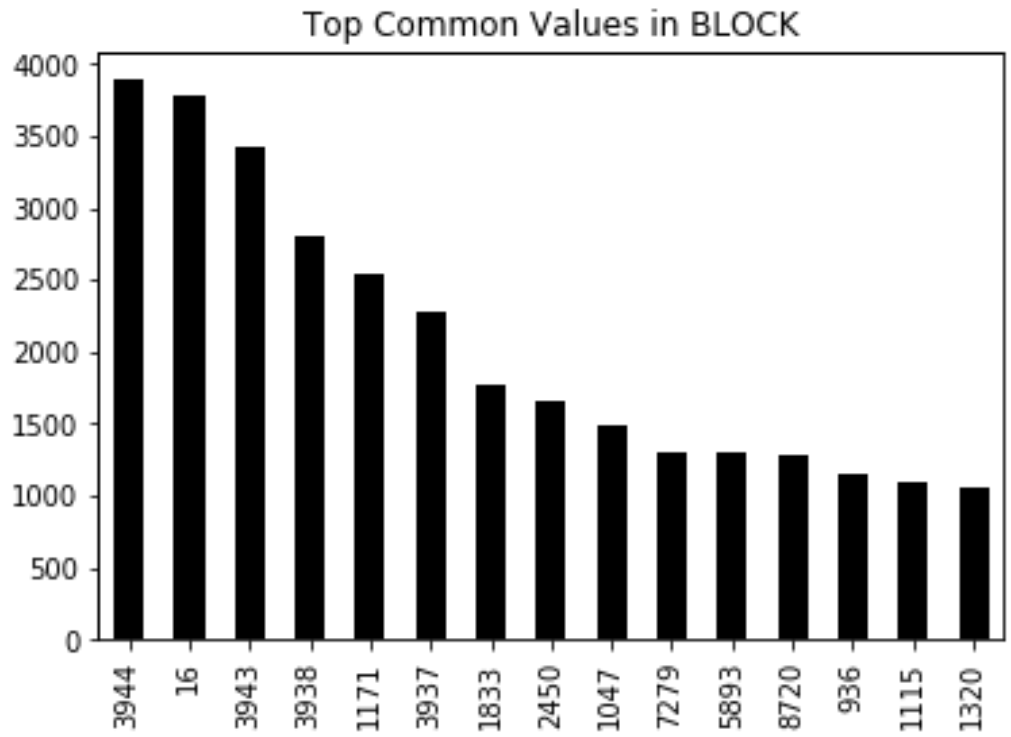


4 358046
3 323243
1 146220
5 136200
2 107285

Field 4 BLOCK
Field Name: BLOCK
Field Type: Categorical
Description: Valid block ranges by borough codes

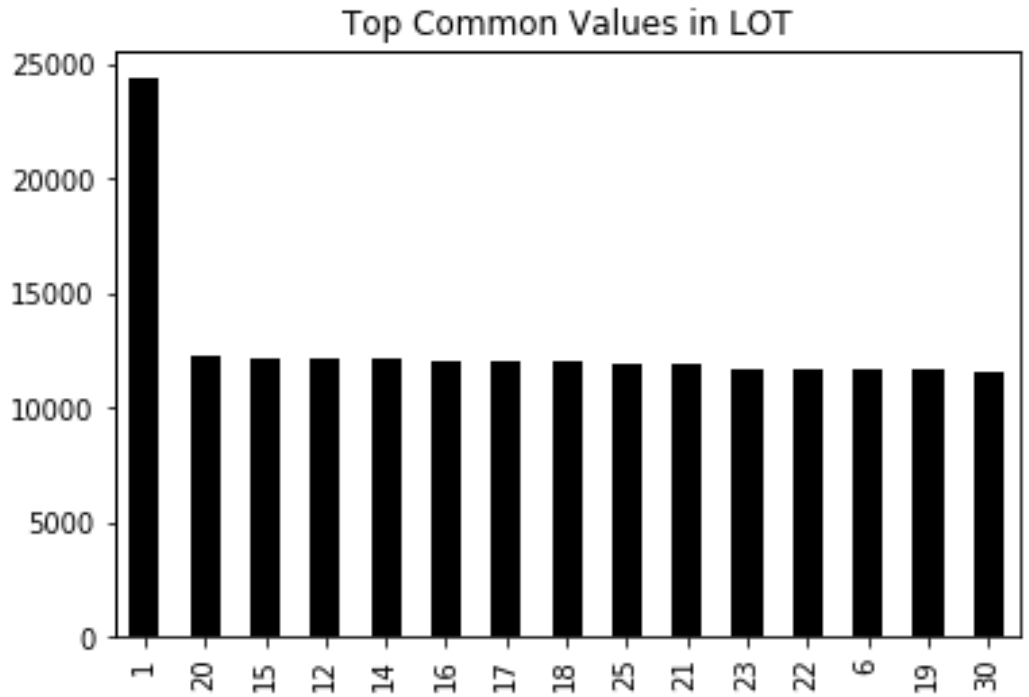
1-2,255	MANHATTAN
2,260-5,958	BRONX
1-8,955	BROOKLYN
1-16,350	QUEENS
1-8,050	STATEN ISLAND

Most Common Values:



3944	3888
16	3786
3943	3424
3938	2794
1171	2535
3937	2275
1833	1774
2450	1651
1047	1480
7279	1302
5893	1295
8720	1281
936	1151
1115	1090
1320	1049

Field 5 LOT
Field Name: LOT
Field Type: Categorical
Description: Unique numbers within borough and block codes
Most Common Values:

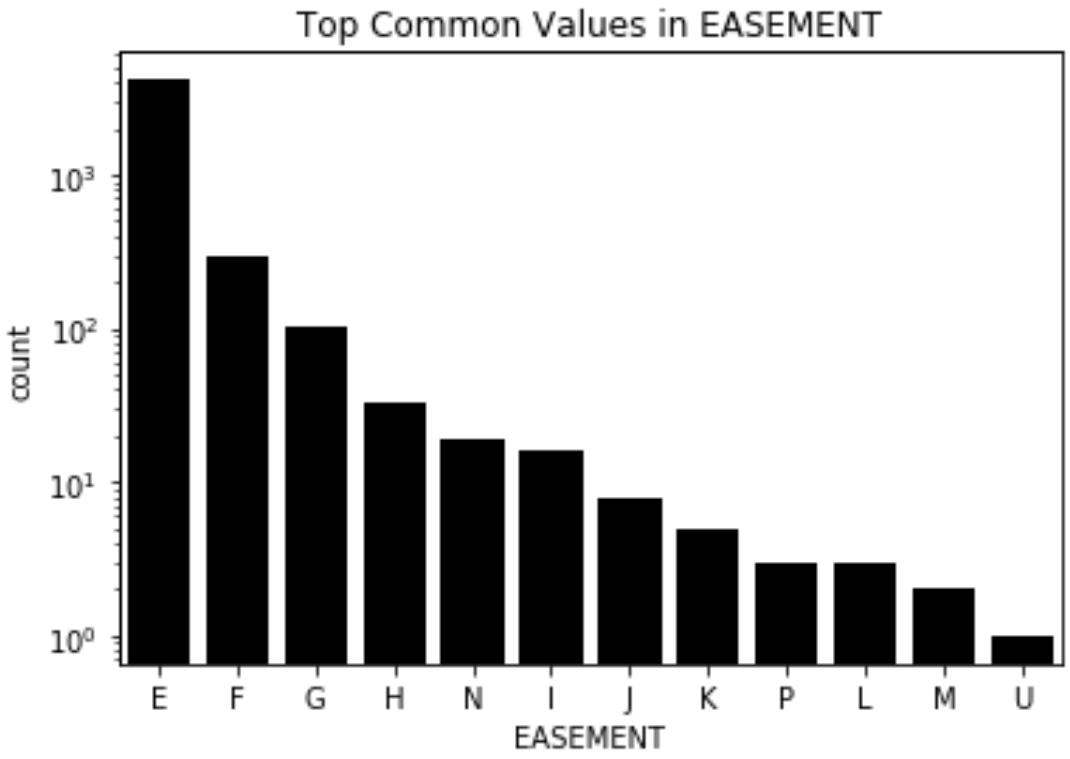


1	24367
20	12294
15	12171
12	12143
14	12074
16	12042
17	11982
18	11979
25	11949
21	11840
23	11705
22	11665
6	11646
19	11640
30	11596

Field 6 EASEMENT
Field Name: EASEMENT
Field Type: Categorical
Description:

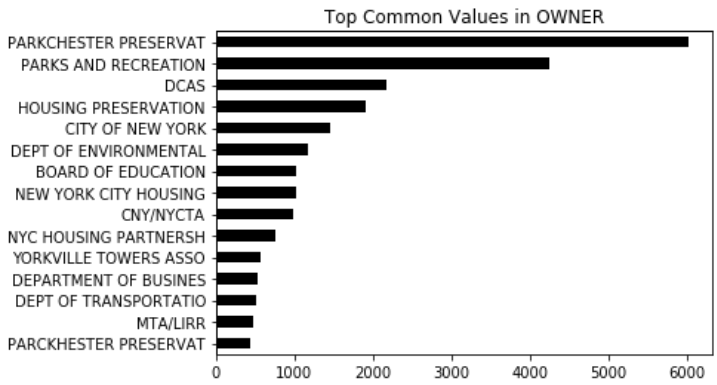
SPACE	Indicates the lot has no Easement.
'A'	Indicates the portion of the Lot that has an Air Easement
'B'	Indicates Non-Air Rights.
'E'	Indicates the portion of the lot that has a Land Easement
'F' THRU 'M'	Are duplicates of 'E'.
'N'	Indicates Non-Transit Easement
'P'	Indicates Piers.
'R'	Indicates Railroads.
'S'	Indicates Street
'U'	Indicates U.S. Government

Most Common Values:



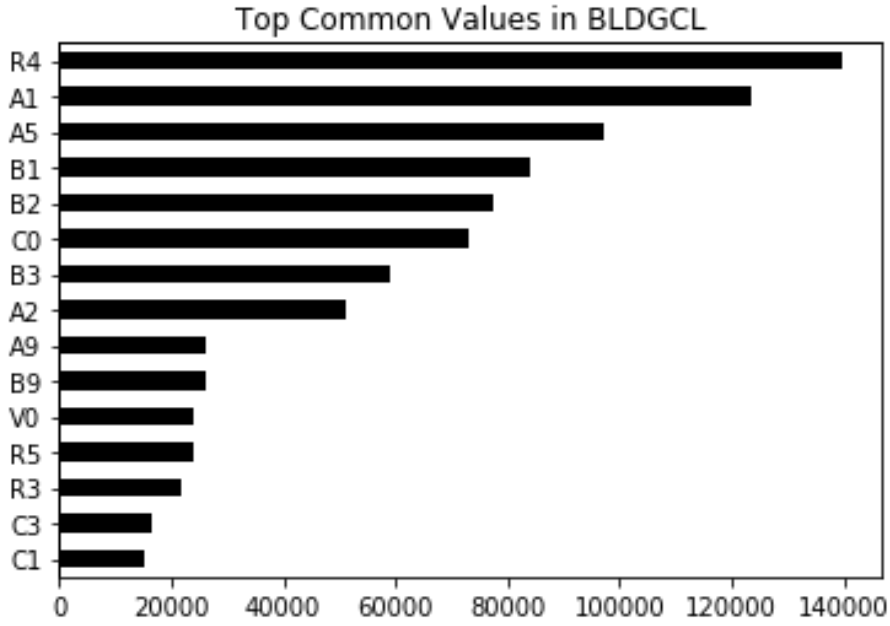
E	4148
F	296
G	102
H	33
N	19
I	16
J	8
K	5
P	3
L	3
M	2
U	1

Field 7 OWNER
Field Name: OWNER
Field Type: Text
Description: Owner's name of each unit of real estate
Most Common Values:



PARKCHESTER PRESERVAT	6021
PARKS AND RECREATION	4255
DCAS	2169
HOUSING PRESERVATION	1904
CITY OF NEW YORK	1450
DEPT OF ENVIRONMENTAL	1166
BOARD OF EDUCATION	1015
NEW YORK CITY HOUSING	1014
CNY/NYCTA	975
NYC HOUSING PARTNERSH	747
YORKVILLE TOWERS ASSO	558
DEPARTMENT OF BUSINES	527
DEPT OF TRANSPORTATIO	503
MTA/LIRR	467
PARCKHESTER PRESERVAT	439

Field 8 BLDGCL
Field Name: BLDGCL
Field Type: Categorical
Description: Building Class
Most Common Values:

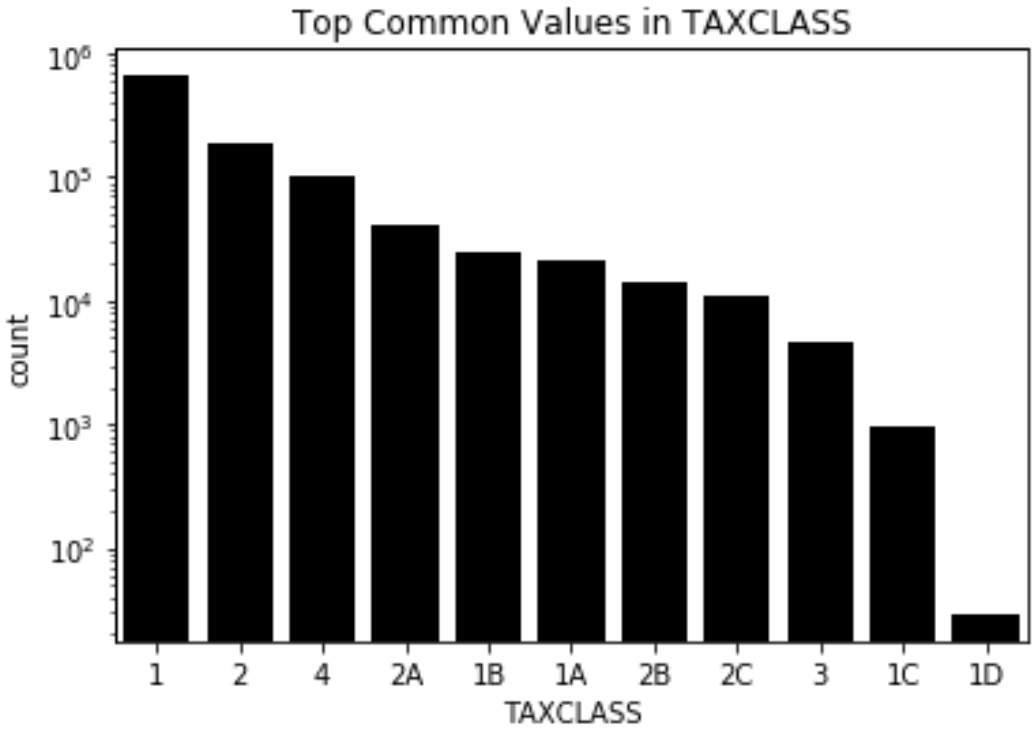


R4	139879
A1	123369
A5	96984
B1	84208
B2	77598
C0	73111
B3	59240
A2	51130
A9	26177
B9	26133
V0	24049
R5	23957
R3	21667
C3	16343
C1	15070

Field 9 TAXCLASS
Field Name: TAXCLASS
Field Type: Categorical
Description: Current property tax class code. There is a direct correlation between building class and tax class.

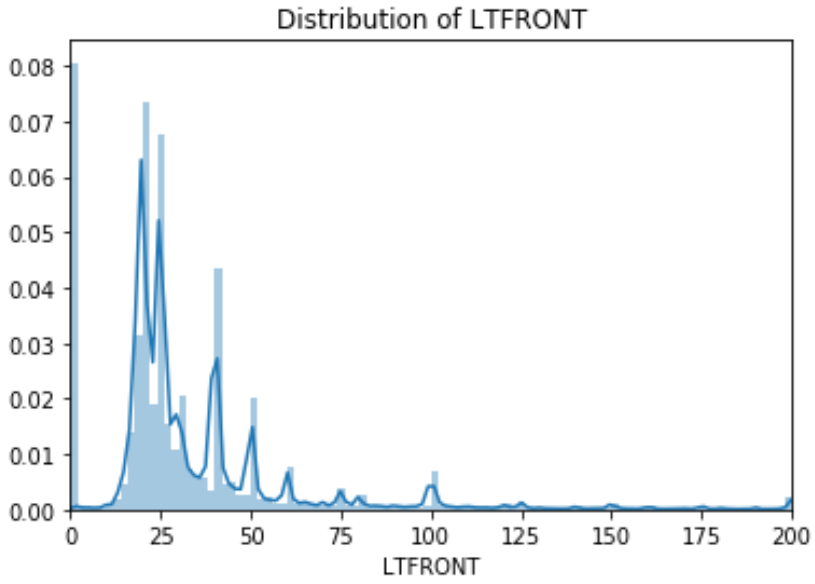
TAX CLASS 1 = 1-3 UNIT RESIDENCES
TAX CLASS 1A = 1-3 STORY CONDOMINIUMS
ORIGINALLY A CONDO
TAX CLASS 1B = RESIDENTIAL VACANT LAND
TAX CLASS 1C = 1-3 UNIT CONDOMINUMS
ORIGINALLY TAX CLASS 1
TAX CLASS 1D = SELECT BUNGALOW COLONIES
TAX CLASS 2 = APARTMENTS
TAX CLASS 2A = APARTMENTS WITH 4-6 UNITS
TAX CLASS 2B = APARTMENTS WITH 7-10 UNITS
TAX CLASS 2C = COOPS/CONDOS WITH 2-10 UNITS
TAX CLASS 3 = UTILITIES (EXCEPT CEILING RR)
TAX CLASS 4A = UTILITIES - CEILING RAILROADS
TAX CLASS 4 = ALL OTHERS

Most Common Values:

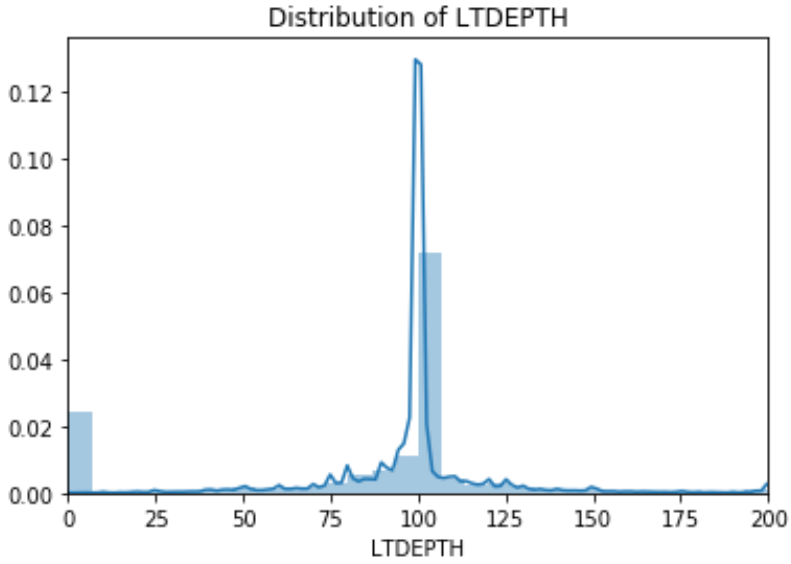


1	660721
2	188612
4	104310
2A	40574
1B	24738
1A	21667
2B	13964
2C	10795
3	4638
1C	946
1D	29

Field 10 LTFRONT
Field Name: LTFRONT
Field Type: Numerical, continuous
Description: Lot frontage (width) in feet
Distribution:



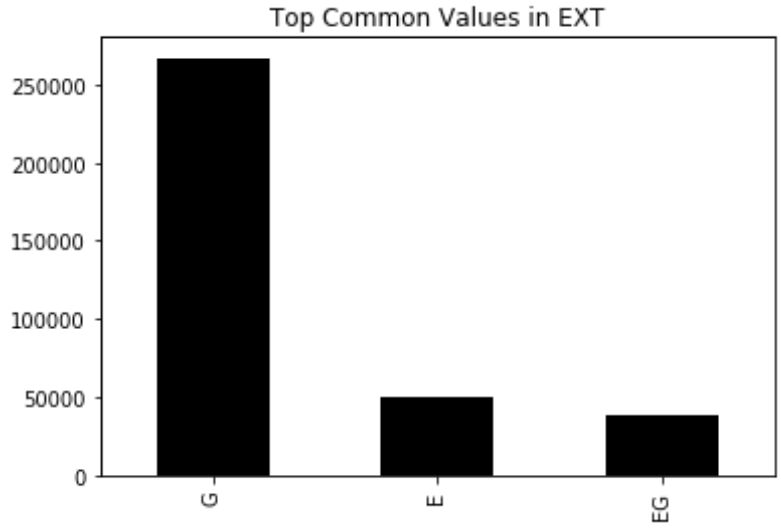
Field 11 LTDEPTH
Field Name: LTDEPTH
Field Type: Numerical, continuous
Description: Lot depth in feet
Distribution:



Field 12 EXT
Field Name: EXT
Field Type: Categorical
Description: Extension indicator

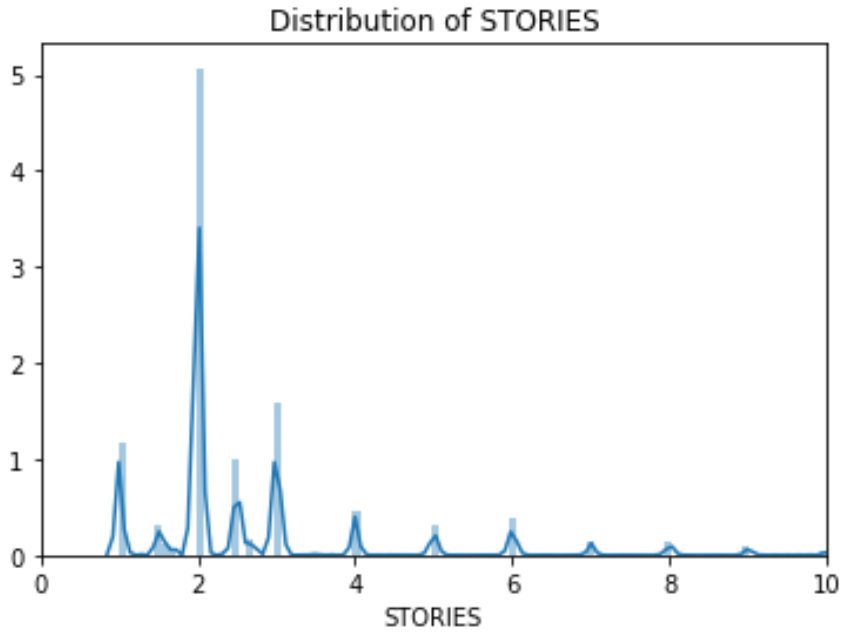
'E' = EXTENSION
'G' = GARAGE
'EG' = EXTENSION AND GARAGE

Most Common Values:

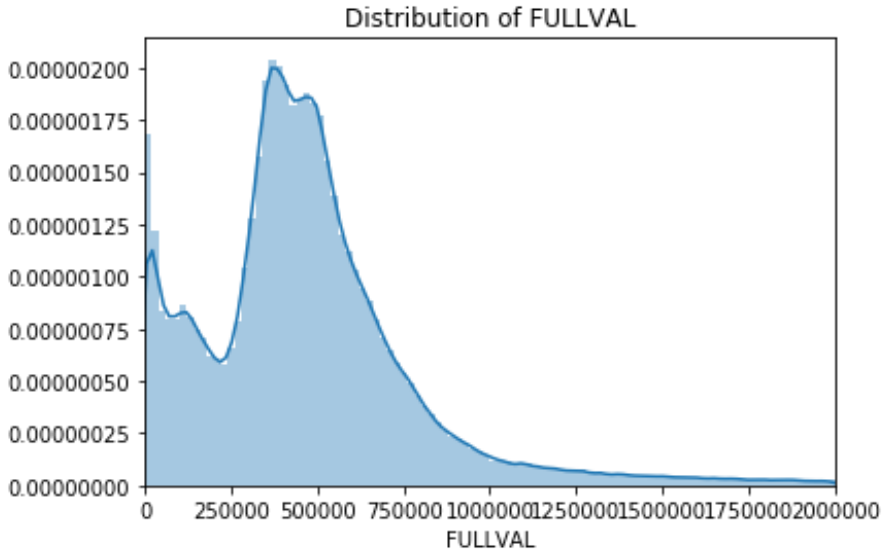


G 266970
E 49442
EG 37893

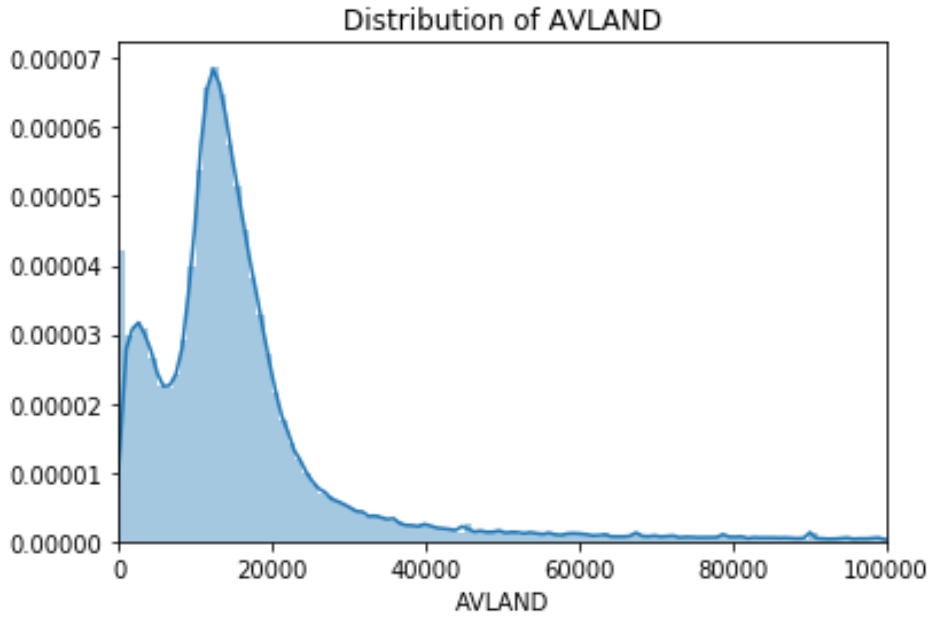
Field 13 STORIES
Field Name: STORIES
Field Type: Numerical, continuous
Description: The number of stories in the building
Distribution:



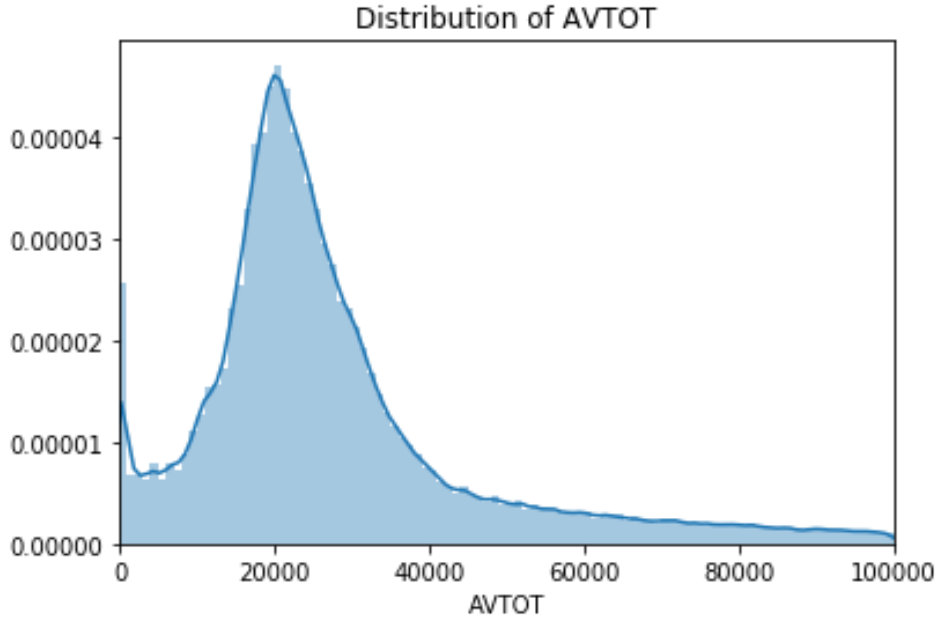
Field 14 FULLVAL
Field Name: FULLVAL
Field Type: Numerical, continuous
Description: Market value
Distribution:



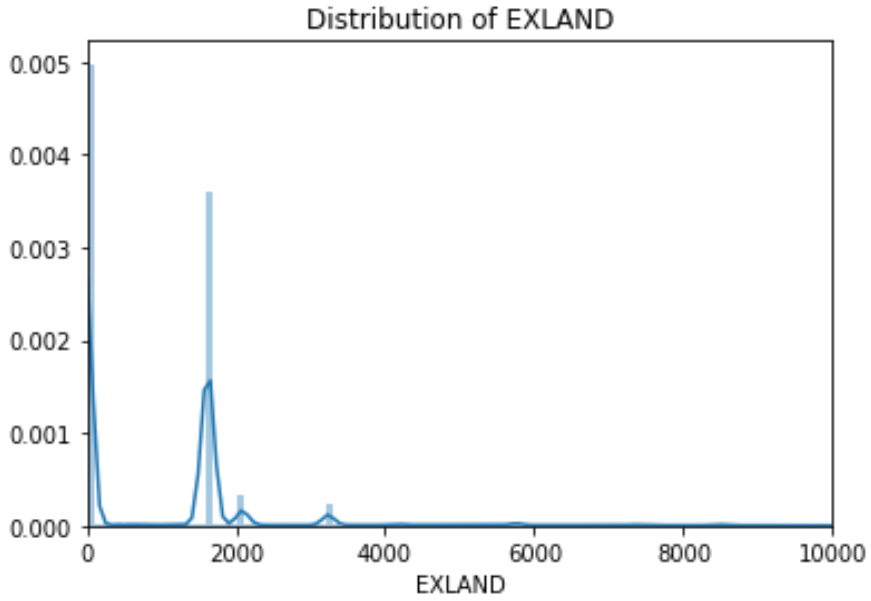
Field 15 AVLAND
Field Name: AVLAND
Field Type: Numerical, continuous
Description: Actual land value
Distribution:



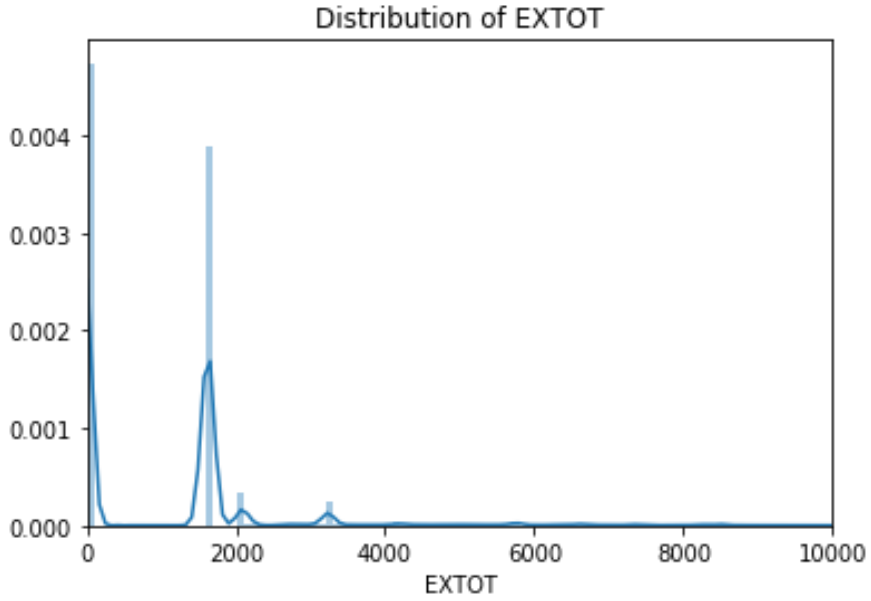
Field 16 AVTOT
Field Name: AVTOT
Field Type: Numerical, continuous
Description: actual total value
Distribution:



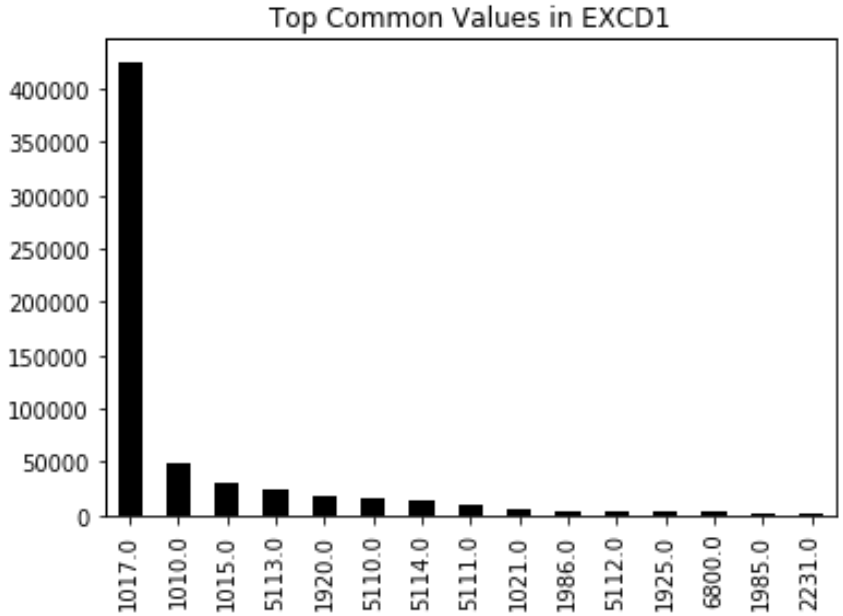
Field 17 EXLAND
Field Name: EXLAND
Field Type: Numerical, continuous
Description: actual exempt land value
Distribution:



Field 18 EXTOT
Field Name: EXTOT
Field Type: Numerical, continuous
Description: Actual exempt land total
Distribution:



Field 19 EXCD1
Field Name: EXCD1
Field Type: Categorical
Description: Exemption code 1
Most Common Values:

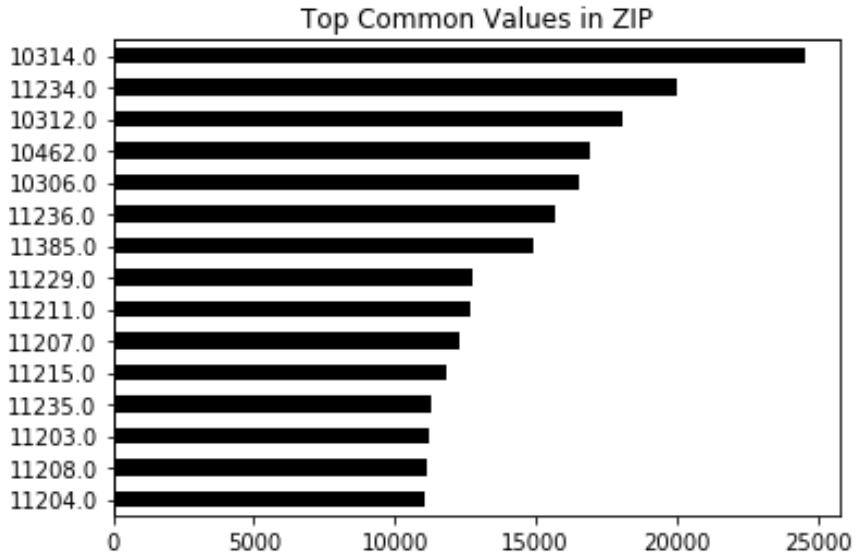


1017.0	425348
1010.0	49756
1015.0	31323
5113.0	23858
1920.0	17594
5110.0	16834
5114.0	14984
5111.0	10609
1021.0	6613
1986.0	4231
5112.0	4071
1925.0	3566
6800.0	3494
1985.0	2471
2231.0	1866

Field 20 STADDR
Field Name: STADDR
Field Type: Text
Description: Street address of the property
Most Common Values：

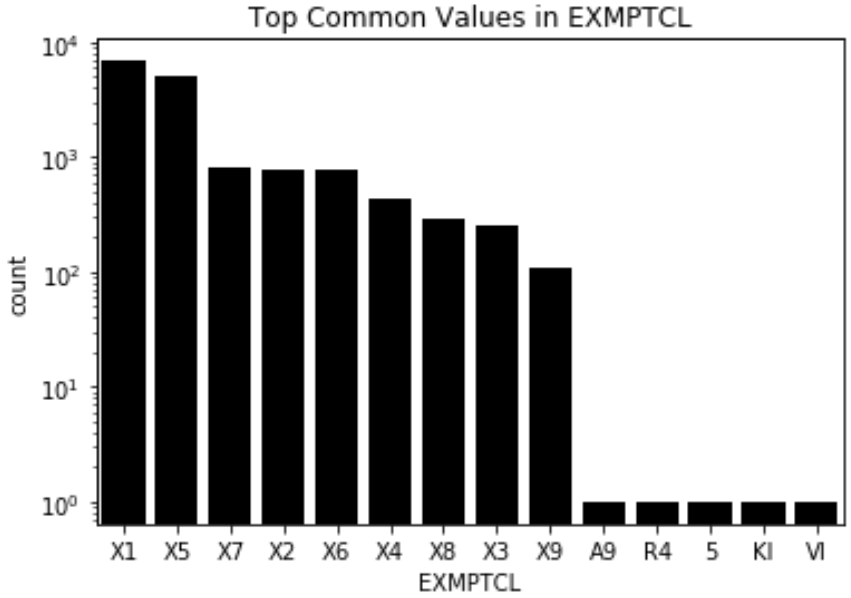
501 SURF AVENUE	902
330 EAST 38 STREET	817
322 WEST 57 STREET	720
155 WEST 68 STREET	671
20 WEST 64 STREET	657
1 IRVING PLACE	650
220 RIVERSIDE BOULEVARD	628
360 FURMAN STREET	599
200 EAST 66 STREET	585
30 WEST 63 STREET	562
2 BAY CLUB DRIVE	556
350 WEST 42 STREET	556
200 RECTOR PLACE	549
301 EAST 79 STREET	538
350 WEST 50 STREET	498

Field 21 ZIP
Field Name: ZIP
Field Type: Categorical
Description: Postal zip code of the property
Most Common Values:



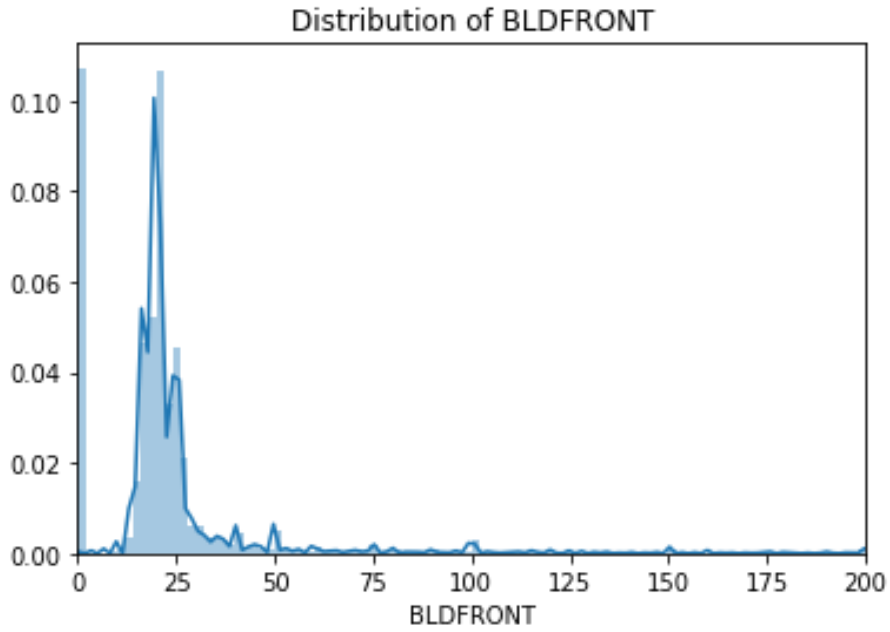
10314.0	24606
11234.0	20001
10312.0	18127
10462.0	16905
10306.0	16578
11236.0	15678
11385.0	14921
11229.0	12793
11211.0	12710
11207.0	12293
11215.0	11834
11235.0	11312
11203.0	11241
11208.0	11139
11204.0	11061

Field 22 EXMPTCL
Field Name: EXMPTCL
Field Type: Categorical
Description: Exempt class used for fully exempt properties only
Most Common Values:

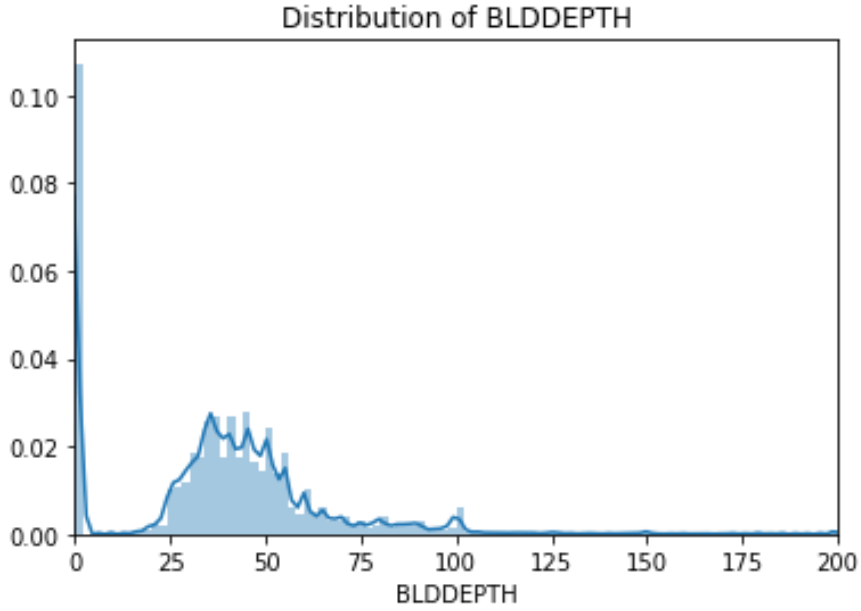


X1	6912
X5	5208
X7	820
X2	770
X6	764
X4	441
X8	292
X3	259
X9	108
A9	1
R4	1
5	1
KI	1
VI	1

Field 23 BLDFRONT
Field Name: BLDFRONT
Field Type: Numerical, continuous
Description: Building frontage (width) in feet
Distribution:

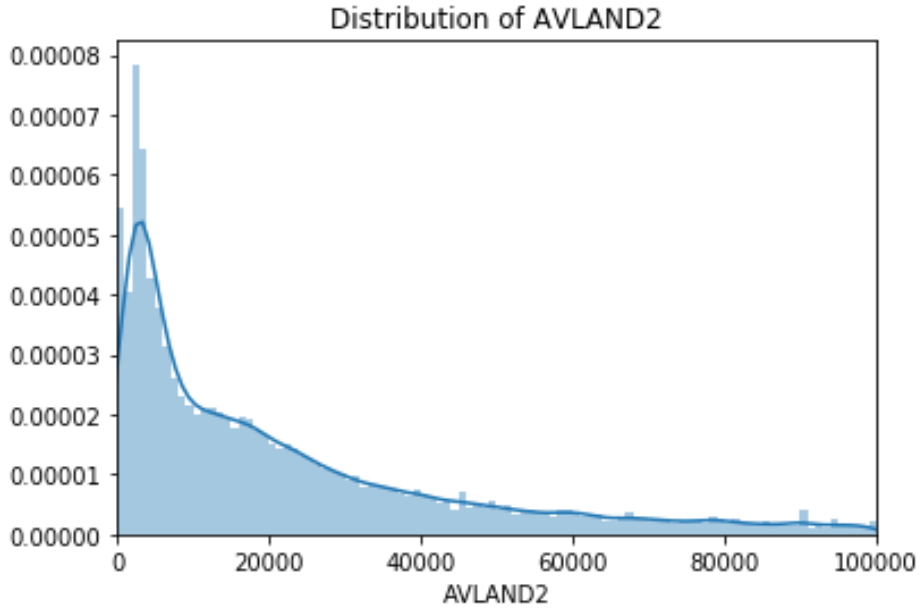


Field 24 BLDDEPTH
Field Name: BLDDEPTH
Field Type: Numerical, continuous
Description: Building depth in feet
Distribution:



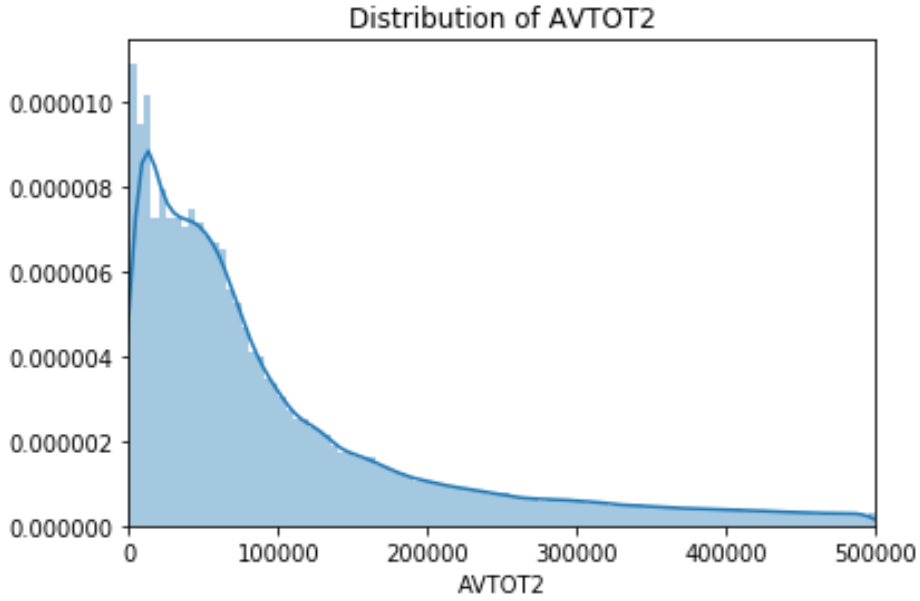
Field 25 AVLAND2

Field Name: AVLAND2
Field Type: Numerical, continuous
Description: Transitional land value
Distribution:

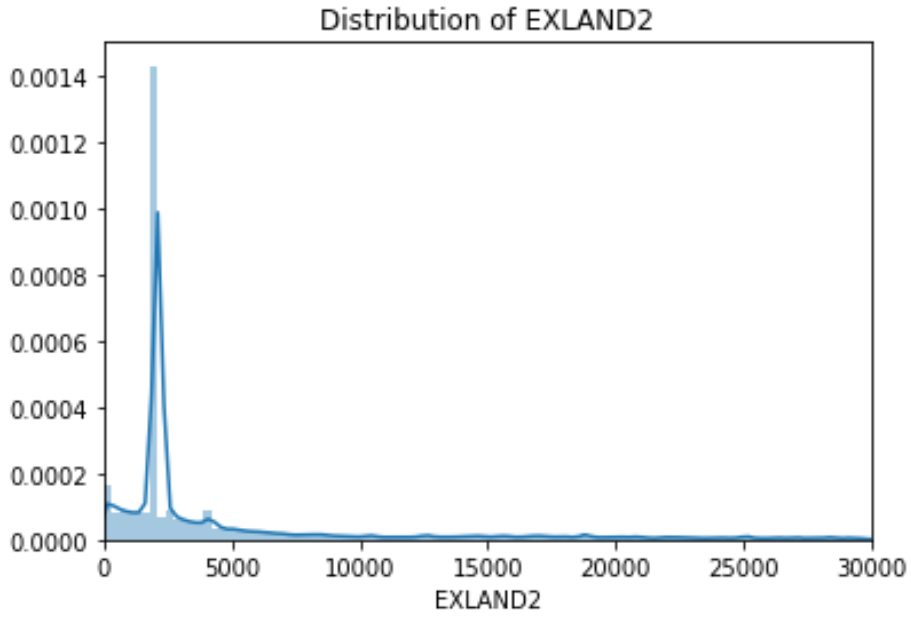


Field 26 AVTOT2

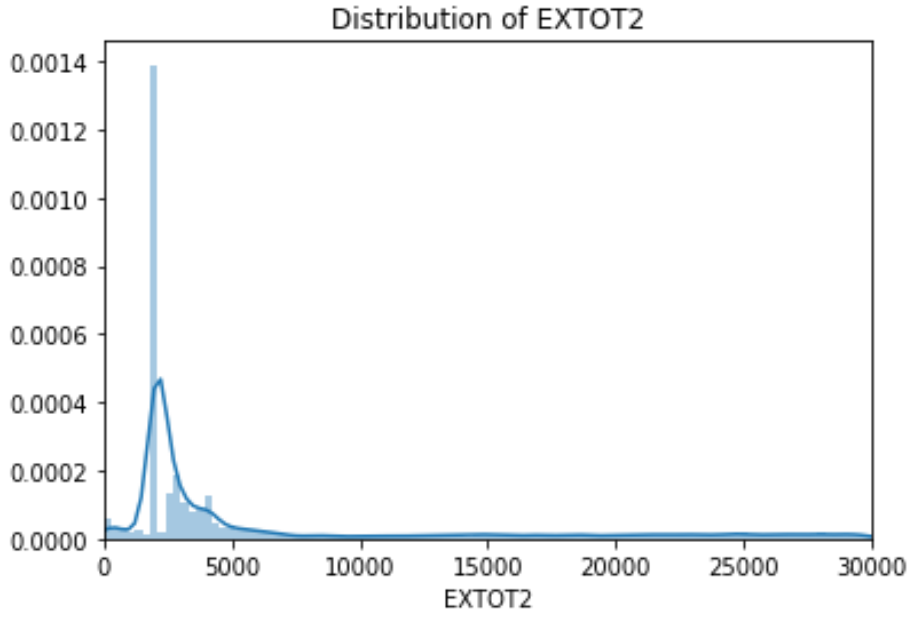
Field Name: AVTOT2
Field Type: Numerical, continuous
Description: Transitional total value
Distribution:



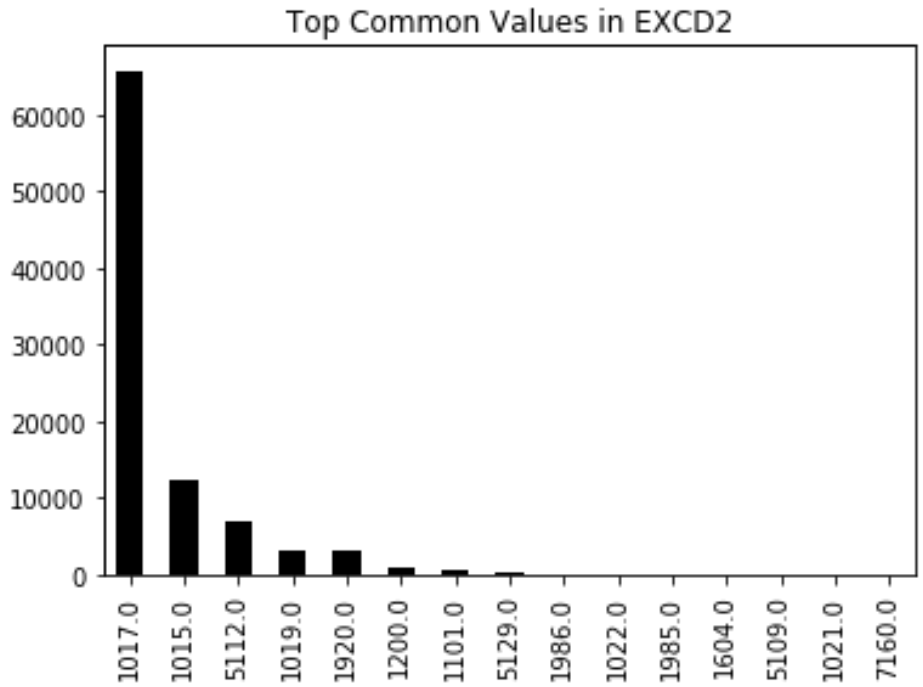
Field 27 EXLAND2
Field Name: EXLAND2
Field Type: Numerical, continuous
Description: Transitional exempt land value
Distribution:



Field 28 EXTOT2
Field Name: EXTOT2
Field Type: Numerical, continuous
Description: Transitional exempt land total
Distribution:



Field 29 EXCD2
Field Name: EXCD2
Field Type: Categorical
Description: Exemption code 2
Most Common Values:



1017.0	65777
1015.0	12337
5112.0	6867
1019.0	3178
1920.0	2961
1200.0	881
1101.0	494
5129.0	227
1986.0	35
1022.0	31
1985.0	21
1604.0	13
5109.0	11
1021.0	8
7160.0	7

Field 30 PERIOD
Field Name: PERIOD
Field Type: Categorical
Description: Indicator for the change period of the file
Most Common Values: No missing value and only one unique value – FINAL

Field 31 YEAR
Field Name: YEAR
Field Type: Date
Description: Year and month
Most Common Values: No missing value and only one unique value – 2010/11

Field 32 VALTYPE
Field Name: VALTYPE
Field Type: Categorical
Description: Value type
Most Common Values: No missing value and only one unique value – AC-TR