Find Anomalies in NY Property Data

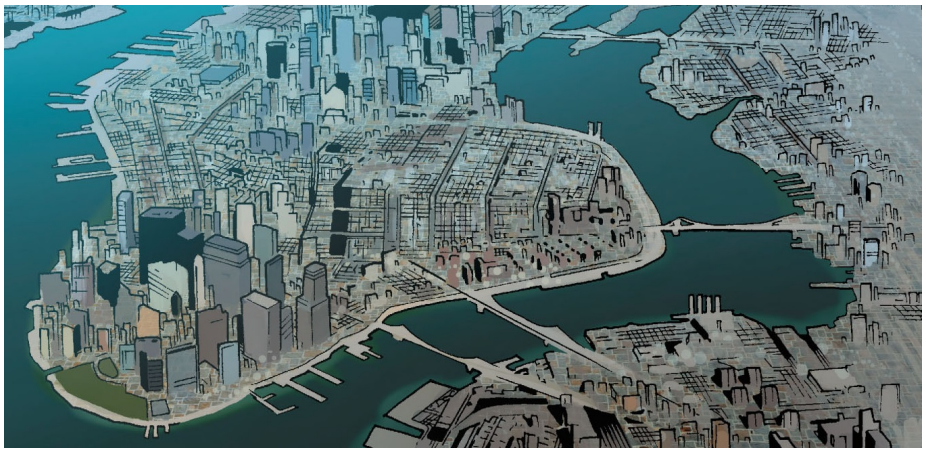
Heng Chi

April 2021

Contents

1 Executive Summary

Our goal of this project is to analyze the real estate data of New York City (NYC) to identify property tax fraud in the city. The data set is provided by the NYC Open Data. The time period of our data set is November 17, 2010, which contains 32 fields and 1,070,994 records.



We first walked through the data set for a data quality report, including data cleaning, visualization, and description. The missing values were filled also. We calculated the following variables for each of the 32 fields *mean, standard deviation, min, max, number of populated values, percent populated, number of unique values, and number of records equal to zero.*

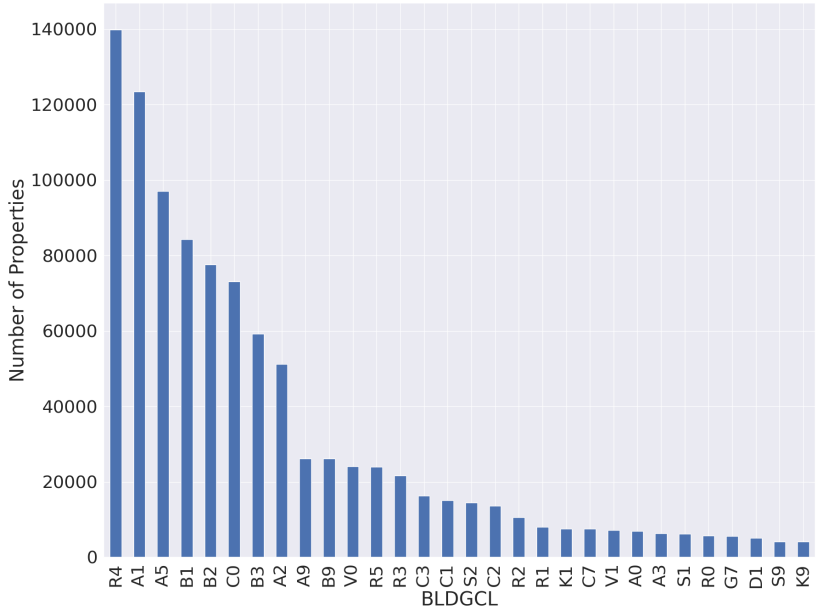
Then, we created 45 new variables with histogram to help us build up algorithm. They were created based on the average of three sizes normalized lot area, building area, and building volume in five groups (zip5, zip3, tax class, borough, and no group).

After that, we applied Dimensionality Reduction to improve the 45 variables data set. We created and modified two algorithms. For each records, we executed our algorithms and produced two scores that can evaluate the likelihood of the record being a fraud. One of the model used Principal Component Analysis with orthogonal transformation. The other model was a neural network autoencoder. Finally, We combined the two scores into one final score and ranked the new scores in descending order to form the result. The final conclusion is conducted based on the top ten final score.

2 Description of Data

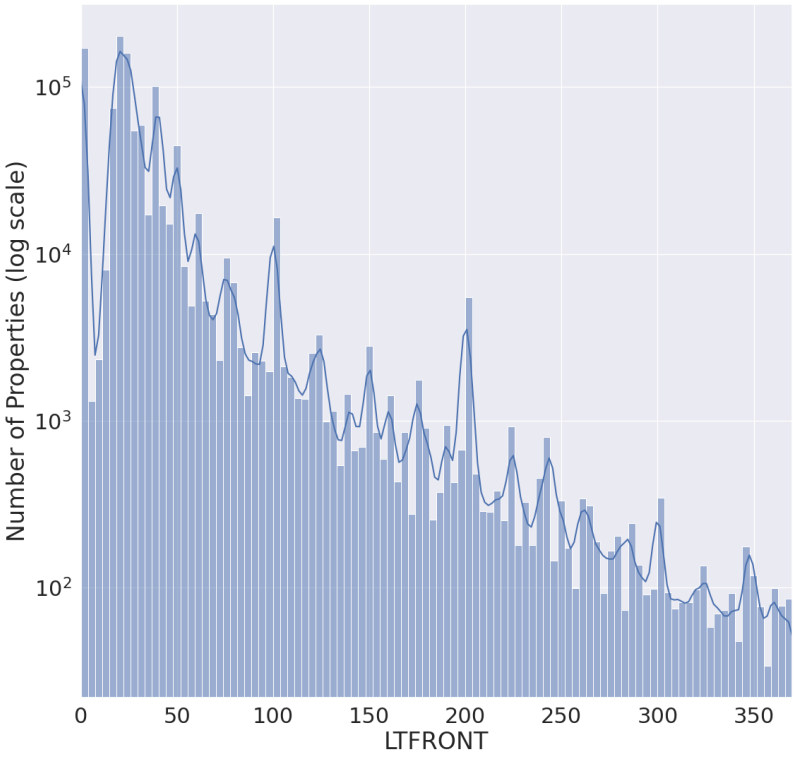
BLDGCL

Two-Character length data filed including the NYC building classification code. The first character in the building classification code is a letter and the second character is a number. All records in the dataset contain a building classification code. The bar chart below shows the top 30 building classification with the highest number of properties in the dataset. The result indicates that R4 is the most common code among the 200 unique values in BLDGCL field.



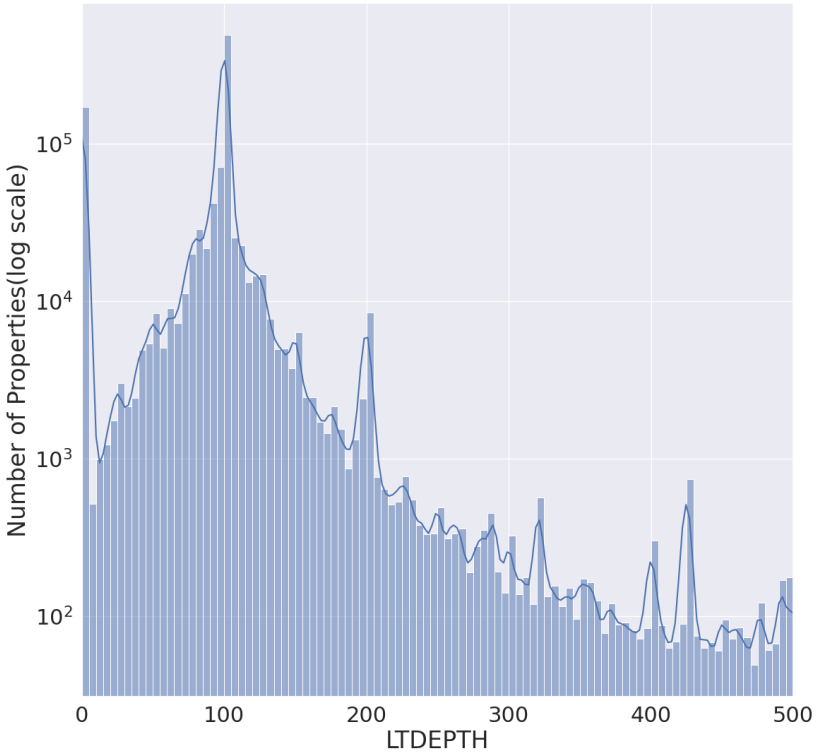
LTFRONT

Integer data field including the measurement of the property lot's front width in feet. All records in the dataset contain a front width lot measurement. Excluding those property records with a measurement bigger than 370, the distribution below shows the front width lot measurement for the property records in the dataset. The most common measurement of lot's front width is zero feet with 169108 property records having this erroneous value. The second common measurement of lot's front width is 20 feet with 135178 property records.



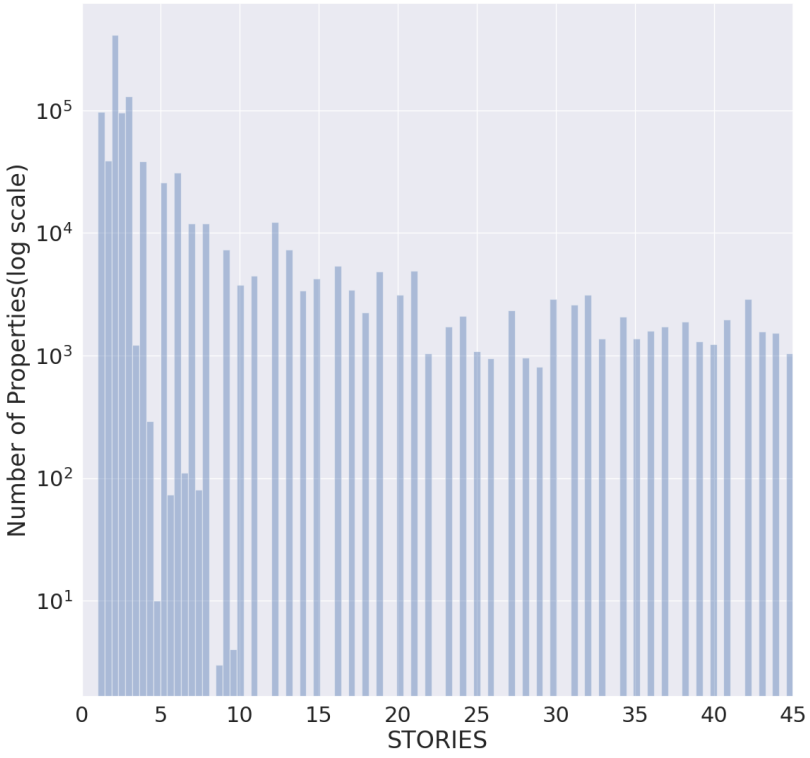
LTDEPTH

Integer data field including the depth of the lot measured in feet. All records in that dataset contain a depth of the lot measurement. Excluding those property records with a measurement bigger than 500, the distribution below shows the lot depth measurements for the property records in the dataset. The most common lot depth for the property is 100 feet with 464541 property records. The second most common lot depth is zero feet with 170128 property records.



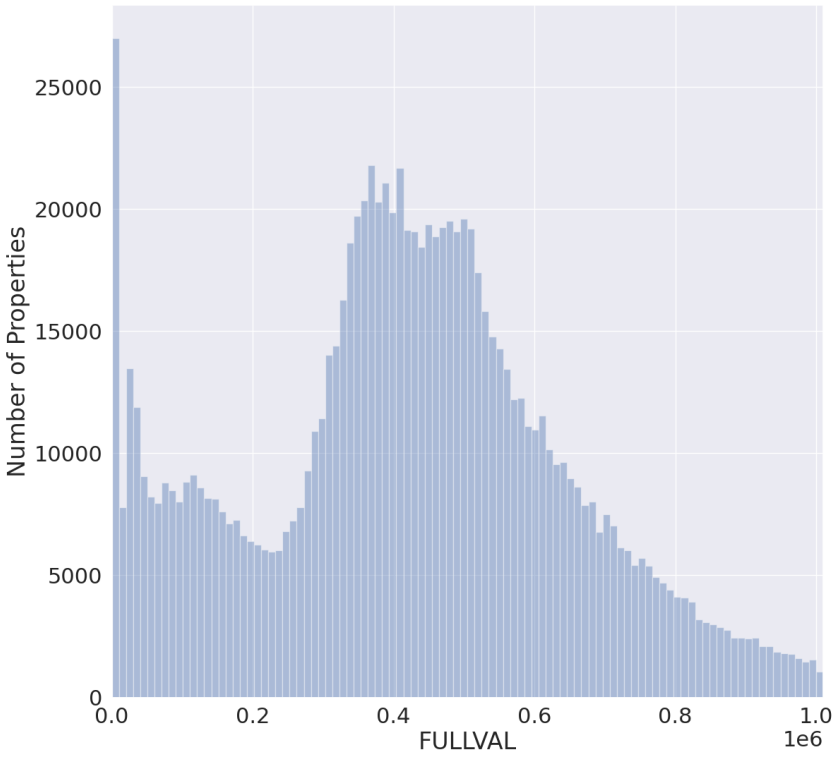
STORIES

Numerical data field including the number of stories for the building. This data field has 56264 property records with no value for the number of stories in the dataset. Excluding those property records with a measurement bigger than 45, the distribution below shows the number of stories data field for the property records. The most common number of stories is 2 with 415092 property records. The highest number of stories for a building in the dataset is 119.



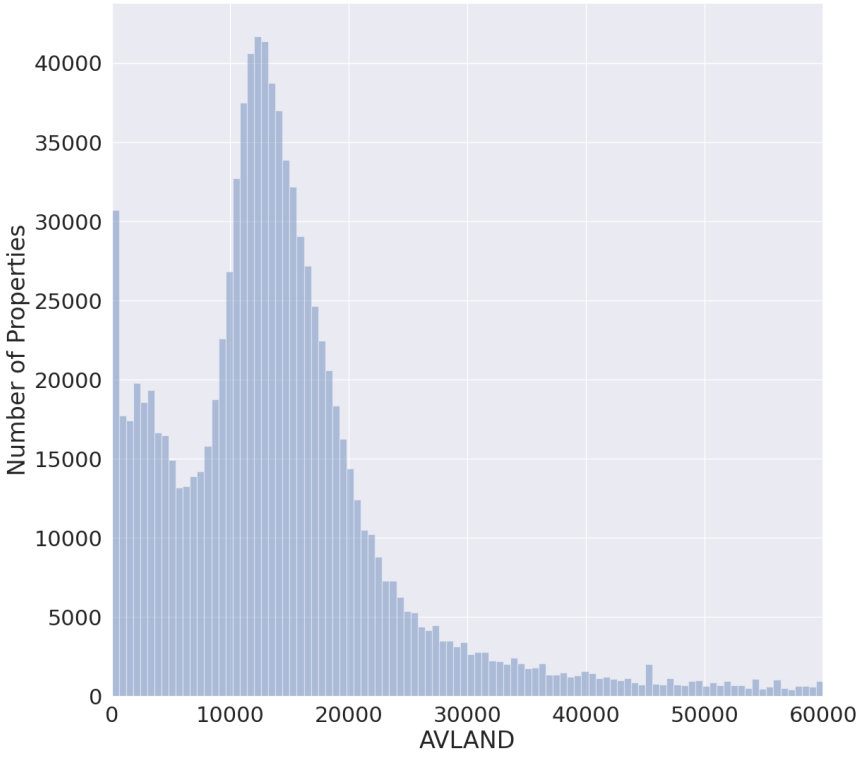
FULLVAL

Numerical data field including the full market value of the property. All records in the dataset contains a full market value of the property and the range is from 0 to 6.15 billion. Excluding those property records with a measurement bigger than 1009999, the distribution plot below shows the histogram of the full market value of the property data field for the property records in the dataset.



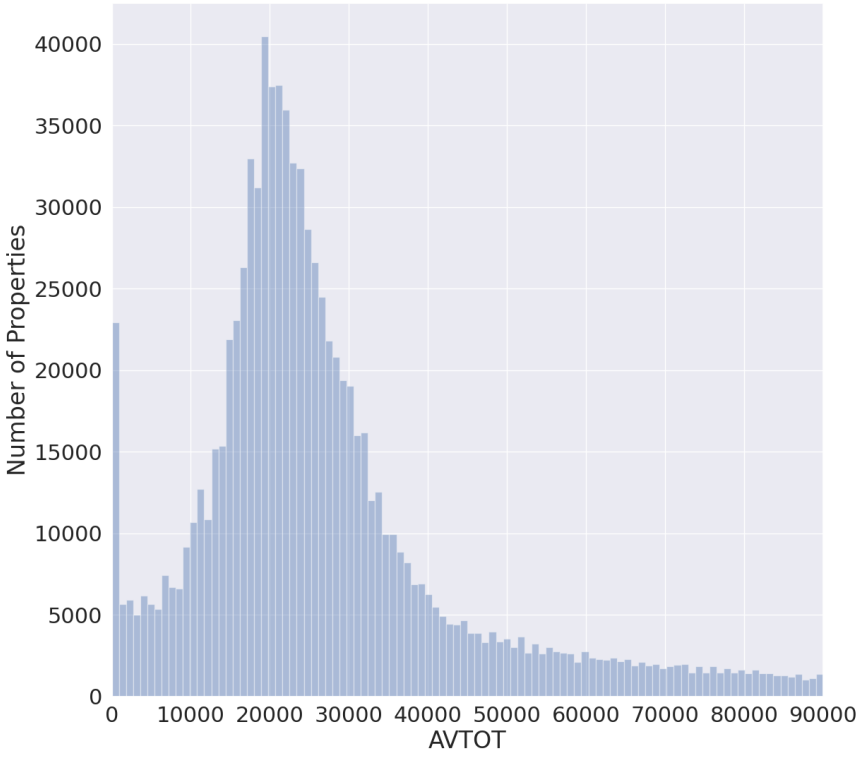
AVLAND

Numerical data field including the actual value of the land on the property. All records in the dataset contain the value for this data field and the range is from 0 to 2.6685 billion. Excluding those property records with a measurement bigger than 60000, the distribution plot below shows the histogram of the actual value of the land data field for the property records in the dataset.



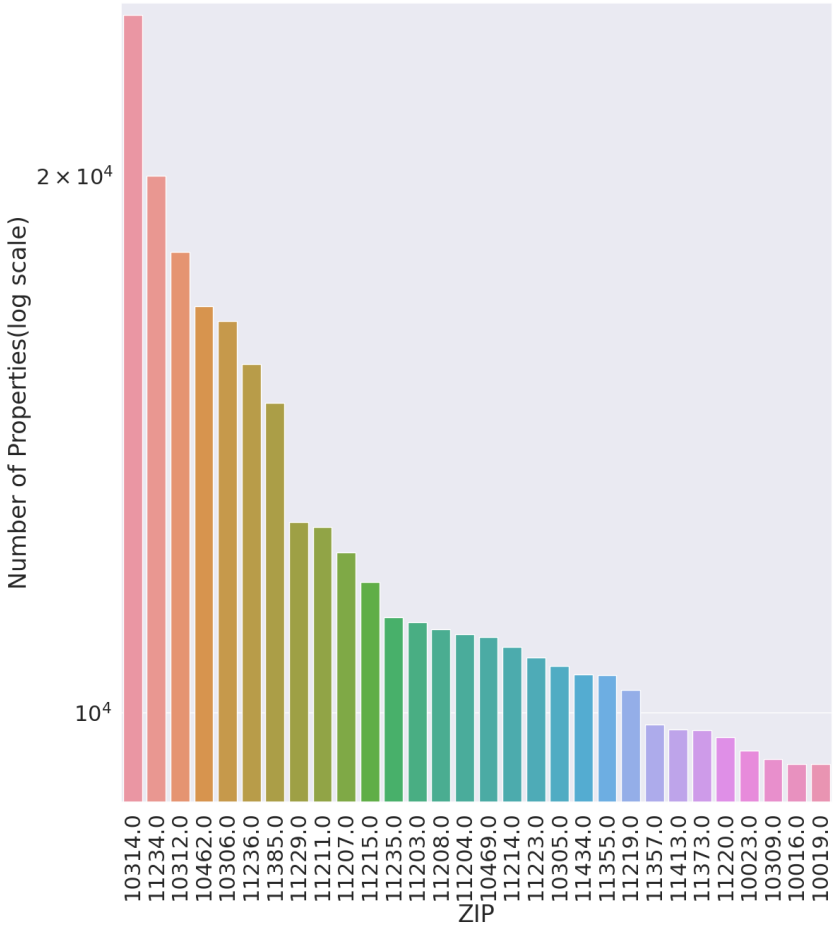
AVTOT

Numerical data field containing the actual total value of the property. All records in the dataset contain the value for this data field and the range is from 0 to 4.6683 billion. Excluding those property records with a measurement bigger than 90000, the distribution plot below shows the histogram of the actual total value data field for the property records in the dataset.



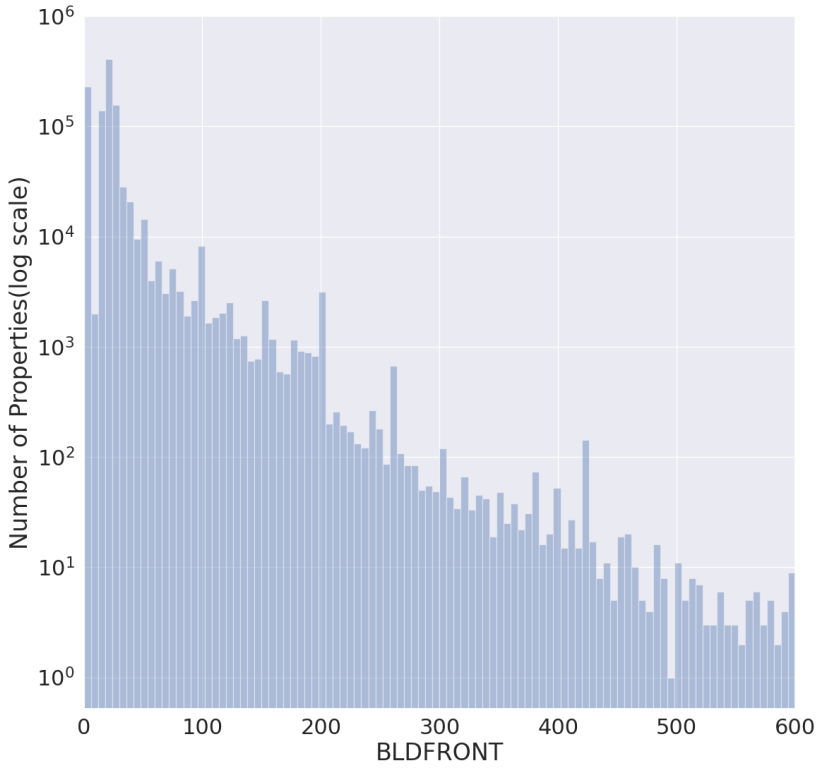
ZIP

Categorical data field including the zip code for the property record. The bar chart below shows the top 30 zip code with the most property records in the dataset. There are 196 different zip codes in total.



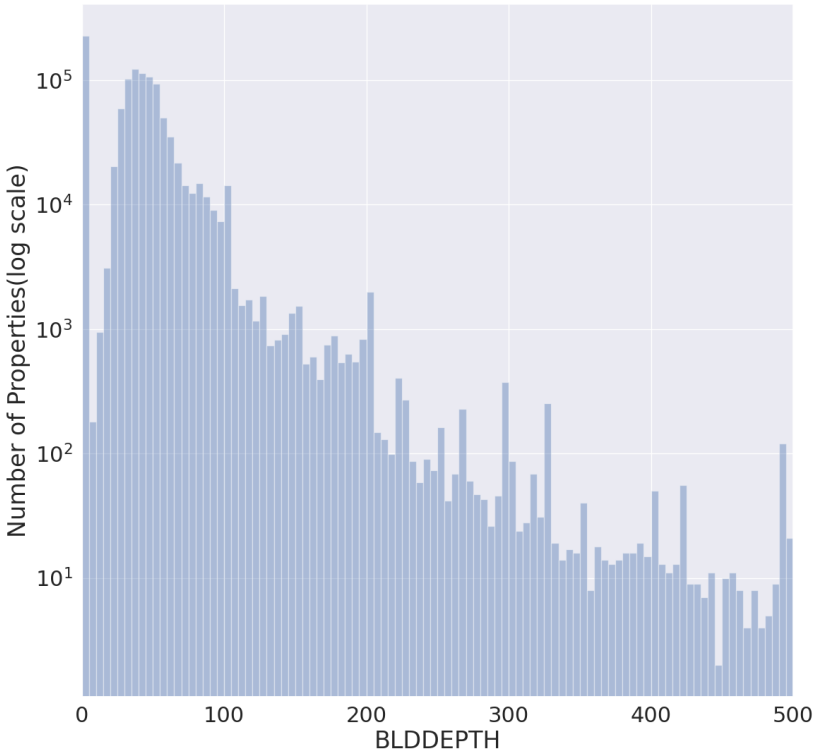
BLDFRONT

Integer data field including the front width of the building measured in feet. All records in the dataset contain a front width of the measured building. Excluding those property records with a measurement bigger than 600 feet, the distribution plot below shows the building front width data field for the property records in the dataset. The most common measurement is zero feet with 228815 property records having this erroneous value. The second common measurement is 20 feet with 195101 property records. A majority of the buildings have the front width of 26 feet or less.



BLDDEPTH

Integer data field including the depth of the building measured in feet. All records in the dataset contain the depth of measured building. The longest measurement for the front width of the building is XX .Excluding those property records with a measurement bigger than 500 feet, the distribution plot below shows the building depth data field for the property records in the dataset. The most common measurement is zero feet with 228853 property records. The second common measurement is 40 feet with 48775 property records. A majority of the buildings have a depth of 60 feet or less.



3 Data Cleaning

We cleaned the data before we made the variables. First we identified and removed certain records that are bad or useless, then we filled in missing values based on our understanding of the data set.

3.1 Exclusions

Our goal is to identify potential property tax fraud, the properties owned by city, state or federal governments are not in our consideration. We excluded these properties before we do any variables calculations, since they would skew the statistics and variable values.

What we did was searching the field "OWNERS" and removed the records that contains some particular descriptions. For example: "CITY OF NEW YORK" or "BOARD OF EDUCATION". After this phase, 24,168 records were removed.

3.2 imputation

When going through the data set, we found that a lot of the records had at least one field either contents "missing" or showed zero value. Although "missing" might be an important piece of information, we decided not to use that information. Instead, we replaced the missing field value with something reasonable. The fields that we applied the imputation are: ZIP, FULLVAL, AVLAND, AVTOT, STORIES, LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH.

ZIP

After the exclusion, there are 21,772 records missing the zip code. However, we noticed that the records are already mostly sorted by zip code. We believe this pattern is an idiosyncrasy of how the data is collected and stored.

FULLVAL AVLAND AVTOT

There were about 13,000 records missing these three property value amounts. That was about less than 1%. For these three, missing could be NAs or the "0" value. Hence we need to fill in both the NAs and the "0"s.

To fill these missing values, we first group the data by TAXCLASS, since that was a good description of the nature of the buildings. After that, we calculated the mean value of each of the properties for each group and used the value to replace the NAs and the "0"s.

STORIES

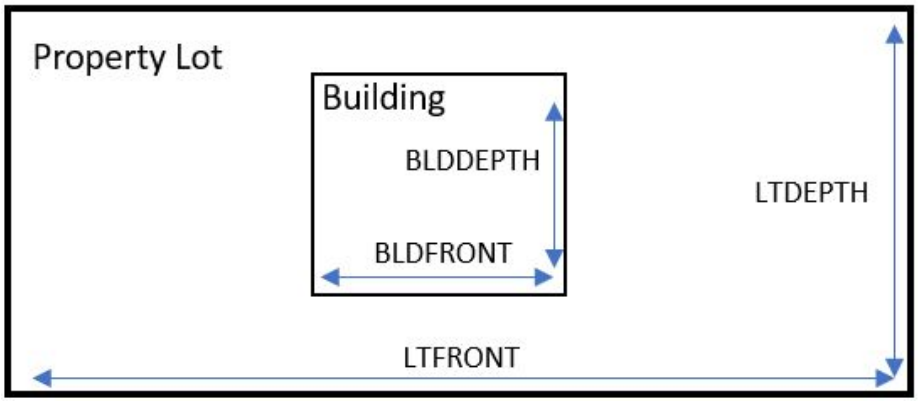
There were 43,968 records missing STORIES (the number of stories in a building). That was about 5% of the records. No records had the value zero for this field. One thing noticeable was the field values were not always integers.

A building could be 1.5 stories tall, meaning part of the building was one story and the other part was two stories.

To fill these missing values, again, we first group the data by TAXCLASS. After that, we calculated the mean value of stories for each group and used the value to replace the NAs.

LTFRONT LTDEPTH BLDFRONT BLDDEPTH

These four variables are the lot and building sizes, below is a picture to remind us what these fields mean.



From the picture above, we learned that "0" are equivalent as missing, since a property will not exist with any of these four variable being "0". Additionally, we also treated the sizes of "1" as missing for the similar reason. Under such definition, we found that about 200,000 records were missing some or all of the lot and building sizes. Which was about 20% of the total.

Field	Number of "0" or "1" value
LTFRONT	169108+839=169947
LTDEPTH	170128+127=170255
BLDFRONT	228815+77=228892
BLDDEPTH	228853+59=228912

To fill these missing values, we first replace all the 0's and 1's to NAs. Then, we applied the mean function (which ignores NAs) to calculate the group-wise average of the non-missing records. Lastly, we replaced the NAs by the results we calculated and finished our imputation.

4 Variable Creation

This dataset provided some variables such as different property dollar values and measurement variables. And to determine if the valuation was fraudulent or not, we focused on these variables and checked if the assessed value was an outlier or not. Therefore, we decided to further create some other variables based on our dataset, so that the model we created would be able to learn more patterns and predict the fraud from our data. And in this part, we created 45 variables, and we will introduce them in the following.

Size Related Variables Creation

To assess property value, we created three variables in the beginning, which are related to size. We not only consider the size of the building, but also the size of the lot. Therefore these three variables are Area_{LOT}, Area_{BLD}, and Vol_{BLD}, which refer to the area of the lot, the area of the building, and the volume of the building. Because these three variable are expected to be related to the property values. To calculate the area of the lot (Area_{LOT}), we multiplied the lot frontage size (LTFRONT) with the lot depth size (LTDEPTH).

$$S_1 = Area_{LOT} = LTFRONT \cdot LTDEPTH$$

To calculate the area of the building (Area_{BLD}), we multiplied the building frontage size (BLDFRONT) with the building depth size (BLDDEPTH).

$$S_2 = Area_{BLD} = BLDFRONT \cdot BLDDEPTH$$

To calculate the volume of the building (Vol_{BLD}), we multiplied the building area (Area_{BLD}) with the number of stories of the building (STORIES)

$$S_3 = Vol_{BLD} = S_2 \cdot STORIES$$

After we created these three variables related to size, we further calculated the normalized values for the assessment values based on these values.

Variables Normalization

In the normalization part, three assessment values were normalized by each of the three variables we just created. The three assessment values were FULLVAL, AVLAND, and AVTOT. FULLVAL is the market value of the property, AVLAND is the assessed value of the property, and AVOTOT is the assessed total value of the property. In the part of normalization, we created nine variables. Three assessment values:

$$V_1 = FULLAL$$

$$V_2 = AVLAND$$

$$V_2 = AVTOT$$

Normalized them by the size variables:

$$r_1 = \frac{V_1}{S_1} \qquad r_2 = \frac{V_1}{S_2} \qquad r_3 = \frac{V_1}{S_3} \tag{1}$$

$$r_4 = \frac{V_2}{S_1} \qquad r_5 = \frac{V_2}{S_2} \qquad r_6 = \frac{V_2}{S_3} \tag{2}$$

$$r_7 = \frac{V_3}{S_1} \qquad r_8 = \frac{V_3}{S_2} \qquad r_9 = \frac{V_3}{S_3} \tag{3}$$

Group and Average

After the normalization, we grouped the data by ZIP5, ZIP3, TAXCLASS, B (borough), and ALL (no group), and calculated the averages of these nine variables after grouping. By grouping, we were able to divide the dataset by location and type of the building. And it helps us to better assess if a property assessment value is considered an outlier or not. ZIP3 is a shorten version of ZIP5, so that we can get more general locations other than locations that were too specific. We labelled each group g_n, where the range of n was 1 to 5. And for each group, we calculated the average for each r_i, which got us 45 variables.

$$\frac{r_1}{< r_1 > g_1}, \frac{r_2}{< r_2 > g_1}, \frac{r_3}{< r_3 > g_1}, ... \frac{r_7}{< r_7 > g_5}, \frac{r_8}{< r_8 > g_5}, \frac{r_9}{< r_9 > g_5} \tag{4}$$

After we created 45 variables, we went ahead to the part of the dimensionalty reduction.

5 Dimensionality Reduction

Dimensionality Reduction with PCA

Principal Component Analysis(PCA) is a dimensionality reduction method that we applied to the NY property data to compute the most influential features. By doing so, we reduced the amount of variables with potentially trading off some accuracy.

In our dimensionality reduction process, we have the following four steps: standardization with z score, PCA with sklearn package, feature selection, and standardization with z score again.

Step 1: Z-score standardization

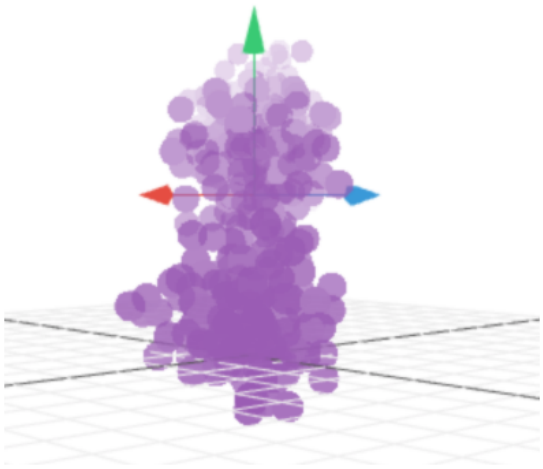
Machine learning algorithms are sensitive to the scaling difference. We used Z score standardization to prepare data for PCA and the following machine learning study. We normalized the data based on mean and standard deviation. The Z-score measures how many standard deviations a given data is away from the mean, while the overall average of normalized data is 0. If the normalized data is positive, that given data is greater than the mean. If the normalized data is negative, the given data is smaller than the mean.

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

z : standard score
x: observed value
$\mu : mean of the sample$
$\sigma : standard deviation of the sample$

Step 2: PCA with sklearn package

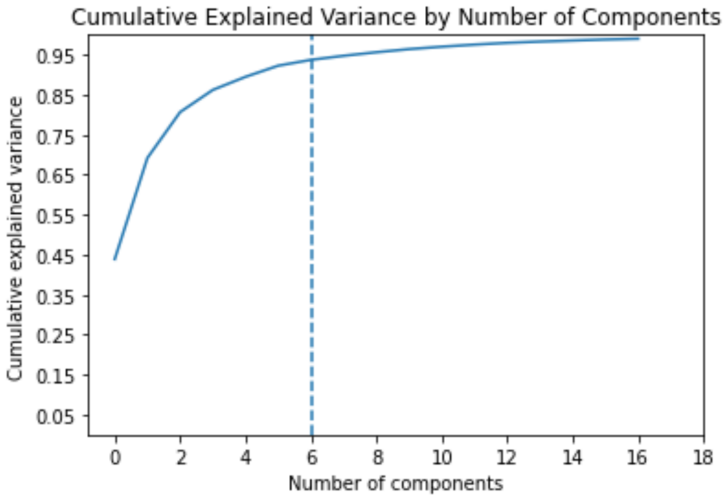
The sklearn's PCA function performs linear dimensionality reduction by projecting the Singular Value Decomposition (SVD) of the original data to a lower dimensional space. In the picture below, the purple points represent the original data and the arrow represents the extracted features. Mathematically, the arrows represent the eigenvectors from the data's covariance matrix. The specific values along the eigenvectors are called eigenvalues, which is the output values for each PC dimension.



(Picture reference: https://setosa.io/ev/principal-component-analysis/)

Step 3: Feature selection

The figure below presents the cumulative explained variance for different numbers of chosen components. As the number of PCs increases, the variance that can be explained increases. Since with the top 6 components more than 90% variance is covered, we consider the top 6 components as a good trade off of explained variance and complexity. We chose the top 6 PCs to move forward.



Step 4: Z-score standardization again

Since our generated new features might contain scaling difference, we used Z-score to standardize our data again for the final machine learning algorithm.

Data dimension change

The input data before dimension reduction has 45 fields with dimension specific meanings such as zip and TAXCLASS. The data after PCA contains 6 fields (the top 6 dimensions). They do not represent any specific meanings anymore, instead, the new dimensions represent linear combinations of old dimensions.

6 Fraud Model Algorithms

Score 1: Mahalanobis-like distance

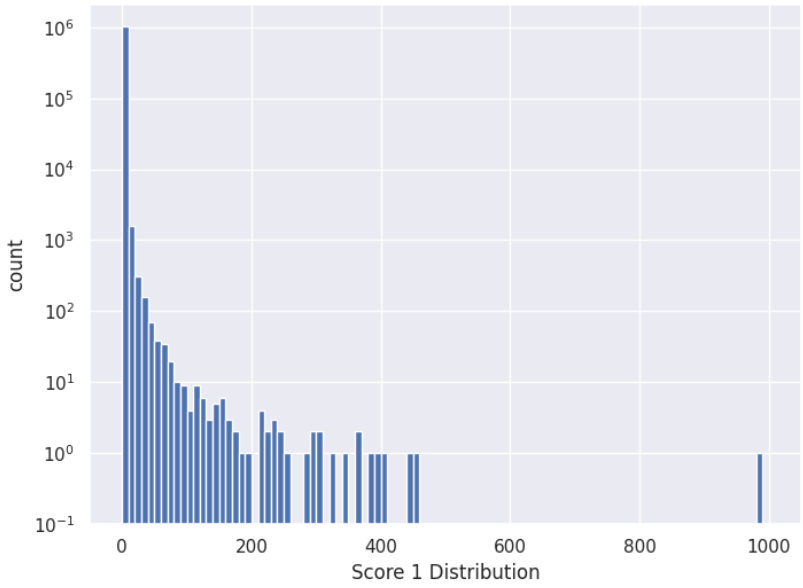
The Mahalanobis distance takes into account the different scales and correlations. It draws equal contours by scaling based on the correlations and different standard deviations. It then measures how many standard deviation contour lines it crosses between points.

After Dimensionality Reduction, our data reduced to a matrix of a million records by six columns. All the records are z-scaled, which are called z-scores. It saved us the effort to calculate the Mahalanobis, which require z scale, PCA, and z scale again.

Score 1 used Minkowsky distance to the origin. Minkowski distance is a metric in Normed vector space. We choose power(p) in the distance formula to be 2, which is the Euclidean distance. All the records are centered, which means that the average is zero, so the measure of outlier is from looking at how far a record is from the origin. We can proceed with the heuristic algorithm to calculate the fraud score by detecting outlier using z-scores. This will update the distance(Fraud Score) formula as below :

$$S_1 i = (\sum_k |PC_{Zik}|^p)^{1/p}, p = 2$$

When there is an anomaly(unusual record), it would have an unusual z-score that lead $S_1 i$ to be large

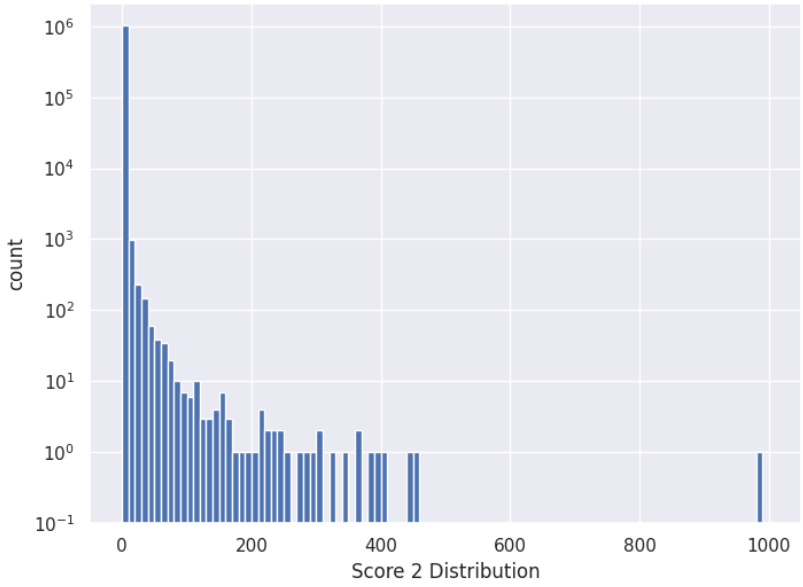


Score 2: Autoencoder Reproduction Error

Score 2 used Minkowsky distance on the difference between the original input record vector and the autoencoder output vector of the autoencoder(autoencoder error). Here we also choose the Euclidean distance(power 2). To compute the output vector, we did an additional step at first, where we trained an autoencoder on the dataset. An autoencoder is a neural network that is trained to reconstruct it's own input, in our case, to reproduce the z scaled PC records. While training the autoencoder, we compressed the data records from 6 down to 3 dimensions, and expanded it back to 6.

$$S_1 i = (\sum_k |PC_{Z'_{ik}} - PC_{Zik}|^p)^{1/p}, p = 2$$

When there is an anomaly(unusual record), the difference between the input and output vector would be large, which would generate a large $S_1 i$

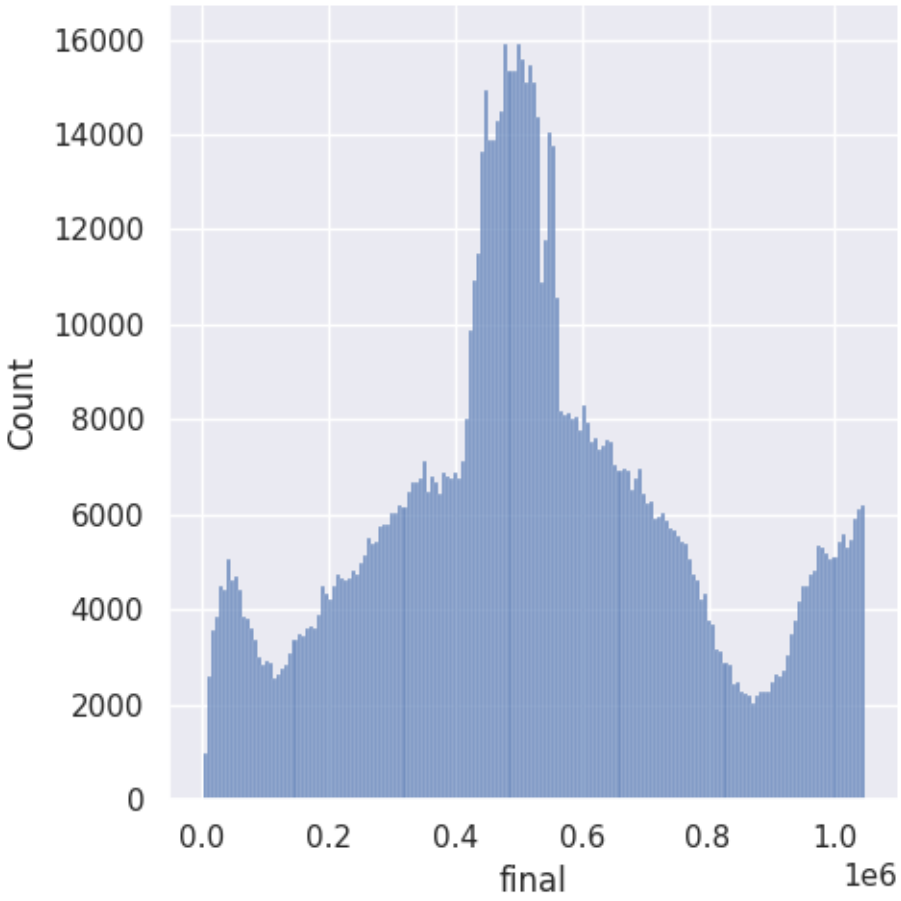


Final Score: Weighted Score

Final Score is the combination of score 1 and score 2 by the average rank order. Generally, when we combine scores, we don't assume the distributions

are the same, so we used quantile binning/ranking ordering to scale score 1 and score 2 to put them on the same footing.

$$FinalScore = (Score_1 * 0.5 + Score_2 * 0.5)$$



7 Results

	RECORD	BBLE	B	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	EXT	STORIES	FULLVAL	AVLAND
898846	917942	4142600001	4	14260	1	NaN	LOGAN PROPERTY, INC.	T1	4	4910.0	124.443774	NaN	3.000000	3.740199e+08	1.792809e+09
667828	684704	4036590105	4	3659	105	NaN	W RUFERT	V0	1B	2.0	2.000000	NaN	4.000000	3.738399e+05	9.185395e+03
1042136	1065870	5076440001	5	7644	1	NaN	PEOPLE OF THE ST OF N	V0	1B	2891.0	1488.000000	NaN	4.000000	2.901746e+08	1.741048e+07
1036415	1059883	5069770012E	5	6977	12	E	NaN	Z7	4	5.0	5.000000	NaN	5.517257	2.772747e+06	4.449960e+05
148125	151044	2024930001	2	2493	1	NaN	NaN	Q6	4	798.0	611.000000	NaN	6.000000	1.663775e+09	7.875000e+07
39267	39770	1007980066	1	798	66	NaN	GREENHORN DEVELOPMENT	D9	2	75.0	98.000000	NaN	13.000000	1.020000e+07	1.741500e+06
115122	116647	1015410021	1	1541	21	NaN	MF ASSOCIATES OF NEW	D6	2	25.0	75.000000	NaN	35.000000	1.610000e+08	1.921500e+07
11988	12076	1001790001	1	179	1	NaN	15 WORTH STREET PROPE	G6	4	74.0	150.000000	NaN	1.000000	2.610000e+06	1.170000e+06
33294	33751	1006940042	1	694	42	NaN	GUIDARA, FRANK	D4	2	122.0	98.000000	NaN	15.000000	1.440000e+07	5.400000e+05
648382	665158	4030720001	4	3072	1	NaN	ST JOHNS CEMETERY	Z8	4	1412.0	2532.000000	NaN	1.000000	2.935500e+07	1.314000e+07

RECORD 917942

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	109.239271	138.324292	261.519058	168.541827
full_bld	494.747786	544.002382	394.377553	563.936096
full_vol	271.630345	273.052250	147.095784	287.130052
assess_lot	5.783570	4.934896	43.735991	6.158232
assess_bld	66.867008	21.709386	78.552603	21.636150
assess_vol	26.554009	11.393055	35.188668	11.762399
tot_lot	9.152644	9.441580	118.589818	13.203304
tot_bld	157.634241	39.822520	186.162557	41.022134
tot_vol	59.705963	19.988820	69.088592	21.017958

Map Track:

From the documents on NYC financial and Google map, this is an inn hotel near Kennedy International Airport.

Unusual Part:

Wrongly choose T1, airport, as tax class. Even though the land's tenant is an airport group, this land still needs to be characterized as a hotel according to law.

This is an 8-story building but it was wrongly marked as 3-story. [1]

Qualification:

The reason this property has a higher unit acre value than its geographical neighbors may because it is a protrusion of its borough and zip11422. In another word, zip11422 is not an expensive place but property 917942 is expensive since 917942 is close to airport. A closer land near airport can cause the premium land price. However, it cannot explain why it also has a higher value among other hotels. After all, inn are not a fancy hotel and this land is not located at some expensive districts, like Manhattan.

RECORD 684704

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	215.047198	190.716414	83.476772	168.541827
full_bld	484.353056	595.300976	114.498483	563.936096
full_vol	234.044601	305.424763	28.624893	287.130052
assess_lot	6.863968	6.824555	1.784743	6.158232
assess_bld	19.005676	21.600438	2.813071	21.636150
assess_vol	10.087295	11.923529	0.703278	11.762399
tot_lot	15.379640	14.936828	1.785817	13.203304
tot_bld	36.389169	41.158874	2.814200	41.022134
tot_vol	18.243144	21.368615	0.703560	21.017958

Map Track:

Cannot locate by vague address, BBLE or NYC public documents.

Unusual Part:

Excessive and abnormal high unit acre price compared with groups mean unit price.

[1] For all abnormal data of STORIES, DEPTH or FRONT in this report, it may be caused because of a wrong input, missing value or deliberate fraud. Further detection is needed.

RECORD 1065870

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	148.323667	152.528063	83.476772	152.528063
full_bld	503.737080	467.705704	114.498483	467.705704
full_vol	263.333805	255.387365	28.624893	255.387365
assess_lot	7.247025	8.924127	1.784743	8.924127
assess_bld	22.479704	18.476264	2.813071	18.476264
assess_vol	11.368824	10.096300	0.703278	10.096300
tot_lot	15.990394	21.502513	1.785817	21.502513
tot_bld	36.500527	31.731270	2.814200	31.731270
tot_vol	19.057706	17.453270	0.703560	17.453270

Map Track:

Cannot locate by vague address, BBLE or NYC public documents.

Unusual Part:

Abnormal high unit price compared with its logical neighbors' price.

The assessed value is not expensive compared with its tax class peers, but the market value is far more expensive. This is unusual.

RECORD 1059883

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	148.323667	152.528063	261.519058	152.528063
full_bld	503.737080	467.705704	394.377553	467.705704
full_vol	263.333805	255.387365	147.095784	255.387365
assess_lot	7.247025	8.924127	43.735991	8.924127
assess_bld	22.479704	18.476264	78.552603	18.476264
assess_vol	11.368824	10.096300	35.188668	10.096300
tot_lot	15.990394	21.502513	118.589818	21.502513
tot_bld	36.500527	31.731270	186.162557	31.731270
tot_vol	19.057706	17.453270	69.088592	17.453270

Map Track:

Cannot locate by vague address, BBLE or NYC public documents.

Unusual Part:

Abnormal high unit price compared with its logical neighbors.
This property has a mere tax class, Easement.

RECORD 151044

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	153.578725	143.799774	261.519058	143.617634
full_bld	489.109780	410.646119	394.377553	410.467056
full_vol	152.702001	202.628054	147.095784	202.649005
assess_lot	12.237007	6.026712	43.735991	6.016192
assess_bld	34.227917	17.701464	78.552603	17.698662
assess_vol	16.617155	10.120273	35.188668	10.120771
tot_lot	47.015153	16.378639	118.589818	16.350790
tot_bld	167.227367	39.715413	186.162557	39.693424
tot_vol	49.334731	19.465649	69.088592	19.457085

Map Track:

Yankee stadium

Unusual Part:

Abnormal high unit price compared with its logical neighbors.
The size of the building is obviously wrong. This may be caused by missing value or fraud, so further investigation is necessary.

Qualification:

Considering this is a landmark of NYC, a dozens time higher value may sounds reasonable. However, the full value per acre is excessively high.

RECORD 39770

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	436.752055	368.767090	84.618372	359.235203
full_bld	735.991011	530.576697	117.980182	517.942490
full_vol	191.425588	114.436158	18.050409	111.912748
assess_lot	27.607065	34.732913	9.387500	33.469382
assess_bld	53.374867	49.223571	13.312831	47.494149
assess_vol	18.331070	10.851404	2.071942	10.445557
tot_lot	83.572458	95.865416	38.078667	92.859664
tot_bld	148.347090	132.236298	53.091500	128.007681
tot_vol	33.453062	22.577242	8.122780	21.781626

Map Track:

An elevator apartment.

Unusual Part:

Abnormal high unit price compared with its logical neighbors.
Building size is not 8 $\times$ 8 feet, but a 25 $\times$ 35 feet apartment. Further investigation is needed to know whether it is a deliberate wrong data input or not.

RECORD 116647

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	356.233271	356.430647	84.618372	359.235203
full_bld	514.640942	514.812772	117.980182	517.942490
full_vol	119.884639	119.873028	18.050409	111.912748
assess_lot	25.777156	25.815896	9.387500	33.469382
assess_bld	35.097501	35.134738	13.312831	47.494149
assess_vol	7.853854	7.853887	2.071942	10.445557
tot_lot	76.576727	76.679049	38.078667	92.859664
tot_bld	96.093731	96.192905	53.091500	128.007681
tot_vol	16.913765	16.914512	8.122780	21.781626

Map Track:

An elevator apartment.

Unusual Part:

Abnormal high z-score of unit price compared with its logical neighbors.

RECORD 12076

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	314.134646	368.767090	261.519058	359.235203
full_bld	409.120611	530.576697	394.377553	517.942490
full_vol	109.119143	114.436158	147.095784	111.912748
assess_lot	25.391483	34.732913	43.735991	33.469382
assess_bld	42.494369	49.223571	78.552603	47.494149
assess_vol	19.525460	10.851404	35.188668	10.445557
tot_lot	80.041333	95.865416	118.589818	92.859664
tot_bld	109.683252	132.236298	186.162557	128.007681
tot_vol	31.658282	22.577242	69.088592	21.781626

Map Track:

A parking lot at Manhattan

Unusual Part:

Abnormal high z-score of unit price compared with its logical neighbors. Wrong size. This is a 50×100 feet park, not 5×5 feet as marked.

Qualification:

Location at core area of Manhattan may explain why it is so expensive compared with other parks. Besides, wrong size may explain why it is so expensive compared with other properties at Manhattan. To know whether it is a deliberate wrong data input or not needs further investigation.

RECORD 33751

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	436.752055	368.767090	84.618372	359.235203
full_bld	735.991011	530.576697	117.980182	517.942490
full_vol	191.425588	114.436158	18.050409	111.912748
assess_lot	27.607065	34.732913	9.387500	33.469382
assess_bld	53.374867	49.223571	13.312831	47.494149
assess_vol	18.331070	10.851404	2.071942	10.445557
tot_lot	83.572458	95.865416	38.078667	92.859664
tot_bld	148.347090	132.236298	53.091500	128.007681
tot_vol	33.453062	22.577242	8.122780	21.781626

Map Track:

Fremin Gallery

Unusual Part:

Abnormal high z-score of unit price compared with its logical neighbors.

Wrongly documented as 8 feet $\times$ 10 feet building, but it is actually a 30 feet $\times$ 60 feet building.

Qualification:

Part of this building belongs to a highway park according to the public documents on NYC government website. This may explains why the owner declared that building size is 8 feet $\times$ 10 feet.

RECORD 665158

The ratio of unit value of this property to group mean unit value. The bigger the number is, more expensive the property is than similar buildings:

	zip5	zip3	taxclass	borough
full_lot	223.579889	190.716414	261.519058	168.541827
full_bld	689.508480	595.300976	394.377553	563.936096
full_vol	368.785230	305.424763	147.095784	287.130052
assess_lot	6.770377	6.824555	43.735991	6.158232
assess_bld	32.236330	21.600438	78.552603	21.636150
assess_vol	22.870134	11.923529	35.188668	11.762399
tot_lot	12.337739	14.936828	118.589818	13.203304
tot_bld	49.215813	41.158874	186.162557	41.022134
tot_vol	32.536085	21.368615	69.088592	21.017958

Map Track:

St. John Cemetery.

Unusual Part:

Abnormal high z-score of unit price compared with its logical neighbors.

There is another cemetery just miles away with a normal price. It makes the data of St. John Cemetery even more suspicious. No. 665158 property has no reason to have a much more expensive value than other cemetery. It is also not reasonable that the unit price of this cemetery is higher than that of nearby residential buildings.

8 Summary and Conclusions

8.1 Summary

A comprehensive analysis of NYC property assessment values was performed to calculate fraud scores to determine fraud. Principal Component Analysis and other variable engineering methods were used to make the model have an outstanding performance on fraud detection. Then, two fraud scores were calculated and the two scores were combined into one final fraud score. The top ten records were further analyzed using z scaled values and map track to determine which variables looked unusual. Following-up on these results with professional tax inspection can help determine whether these records are actually fraudulent. This model also uses imputation methods to fill-in some automatic value, causing some abnormal records. Using a better method other than imputation is what we can further do to distinguish between fraudulent data and missing data.

8.2 Unusual buildings have abnormal high unit price or abnormal low area

	zip5	zip3	taxclass	borough
full_lot	314.134646	368.767090	261.519058	359.235203
full_bld	409.120611	530.576697	394.377553	517.942490
full_vol	109.119143	114.436158	147.095784	111.912748
assess_lot	25.391483	34.732913	43.735991	33.469382
assess_bld	42.494369	49.223571	78.552603	47.494149
assess_vol	19.525460	10.851404	35.188668	10.445557
tot_lot	80.041333	95.865416	118.589818	92.859664
tot_bld	109.683252	132.236298	186.162557	128.007681
tot_vol	31.658282	22.577242	69.088592	21.781626

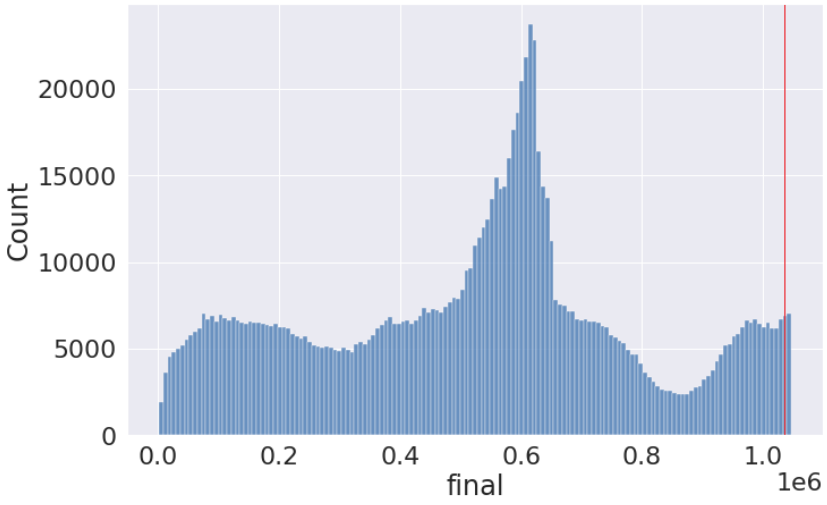
All properties with high final fraud scores have a trend that they all have much higher unit value than the unit value ('r1' to 'r9') of their geographical neighbors or that of their tax class neighbors. This may be caused by the abnormal small depth and front, or by excessive high price of land. Especially for the top 10 suspicious properties, the unit price of these properties is hundreds of times the unit price of the similar properties. This is why they are suspicious. It is unlikely that they will be much more expensive or have much smaller area than similar buildings.

8.3 Suspicious properties have boosting full value

	zip5	zip3	taxclass	borough
full_lot	314.134646	368.767090	261.519058	359.235203
full_bld	409.120611	530.576697	394.377553	517.942490
full_vol	109.119143	114.436158	147.095784	111.912748
assess_lot	25.391483	34.732913	43.735991	33.469382
assess_bld	42.494369	49.223571	78.552603	47.494149
assess_vol	19.525460	10.851404	35.188668	10.445557
tot_lot	80.041333	95.865416	118.589818	92.859664
tot_bld	109.683252	132.236298	186.162557	128.007681
tot_vol	31.658282	22.577242	69.088592	21.781626

Another trend of top suspicious properties is that their full market value is usually hundred of times the unit price of their neighbors, while their assessed value are usually dozens of times the unit price of their neighbors. In another word, assessed price is expensive but market value are even more expensive. In this way, false reporting the market value may be another way to detect fraud.

8.4 Top100 abnormal buildings are far less adequate to detect all fraudulent properties



From the histogram of the last fraud score of the model, simply focusing on the top 100 properties by fraud scores(right part of red line) may be not enough. There are still many buildings or lands with high fraud scores. Fraud detection needs to be extended to all buildings with a fraud score bigger than 800,000.

9 Appendix: Data Quality Report(DQR

9.1 Data Overiew

Dataset Name: Property Valuation and Assessment Data Dataset Purpose: NYC properties assessments to calculate property tax, grant eligible properties, exemptions and/or abatements Data Source: NYC Open Data Time Period: Nov 17, 2010 Number of Fields: 32 Number of Records: 1070994

9.2 Summary Tables

Numeric Field

	# of Records	# Populated	Unique Values	Mean	Standard Devision	Max Value	Min Value	# Zeros
LTFRONT	1070994.0	100.00	1297	36.635301	7.403284e+01	9.999000e+03	0.0	169108
LTDEPTH	1070994.0	100.00	1370	88.861594	7.639628e+01	9.999000e+03	0.0	170128
STORIES	1014730.0	94.75	112	5.006918	8.365707e+00	1.190000e+02	1.0	0
FULLVAL	1070994.0	100.00	109324	874264.505434	1.158243e+07	6.150000e+09	0.0	13007
AVLAND	1070994.0	100.00	70921	85067.918672	4.057260e+06	2.668500e+09	0.0	13009
AVTOT	1070994.0	100.00	112914	227238.168711	6.877529e+06	4.668309e+09	0.0	13007
EXLAND	1070994.0	100.00	33419	36423.890692	3.981576e+06	2.668500e+09	0.0	491699
EXTOT	1070994.0	100.00	64255	91186.981682	6.508403e+06	4.668309e+09	0.0	432572
EXCD1	638488.0	59.62	130	1602.014232	1.384227e+03	7.170000e+03	1010.0	0
BLDFRONT	1070994.0	100.00	612	23.042770	3.557970e+01	7.575000e+03	0.0	228815
BLDDEPTH	1070994.0	100.00	621	39.922836	4.270715e+01	9.393000e+03	0.0	228853
AVLAND2	282726.0	26.40	58592	246235.719265	6.178963e+06	2.371005e+09	3.0	0
AVTOT2	282732.0	26.40	111361	713911.436173	1.165253e+07	4.501180e+09	3.0	0
EXLAND2	87449.0	8.17	22196	351235.684273	1.080221e+07	2.371005e+09	1.0	0
EXTOT2	130828.0	12.22	48349	656768.281904	1.607251e+07	4.501180e+09	7.0	0

Categorical Field

	# of Records	% Populated	Most Common Field Values	Unique Values
RECORD	1070994.0	100.00	N/A	1070994
BBLE	1070994	100.00	N/A	1070994
B	1070994.0	100.00	4	5
BLOCK	1070994.0	100.00	3944	13984
LOT	1070994.0	100.00	1	6366
EASEMENT	4636	0.43	E	13
OWNER	1039249	97.04	PARKCHESTER PRESERVAT	863348
BLDGCL	1070994	100.00	R4	200
TAXCLASS	1070994	100.00	1	11
EXT	354305	33.08	G	4
ZIP	1041104.0	97.21	10314.0	197
EXCD1	638488.0	59.62	1017.0	130
STADDR	1070318	99.94	501 SURF AVENUE	839281
EXMPTCL	15579	1.45	X1	15
EXCD2	92948.0	8.68	1017.0	61
YEAR	1070994	100.00	2010/11	1
VALTYPE	1070994	100.00	AC-TR	1

9.3 Data Field Exploration

Field 1: RECORD

Description: Categorical data field including an integer representing the unique record number identifier from 1 to 1070994.

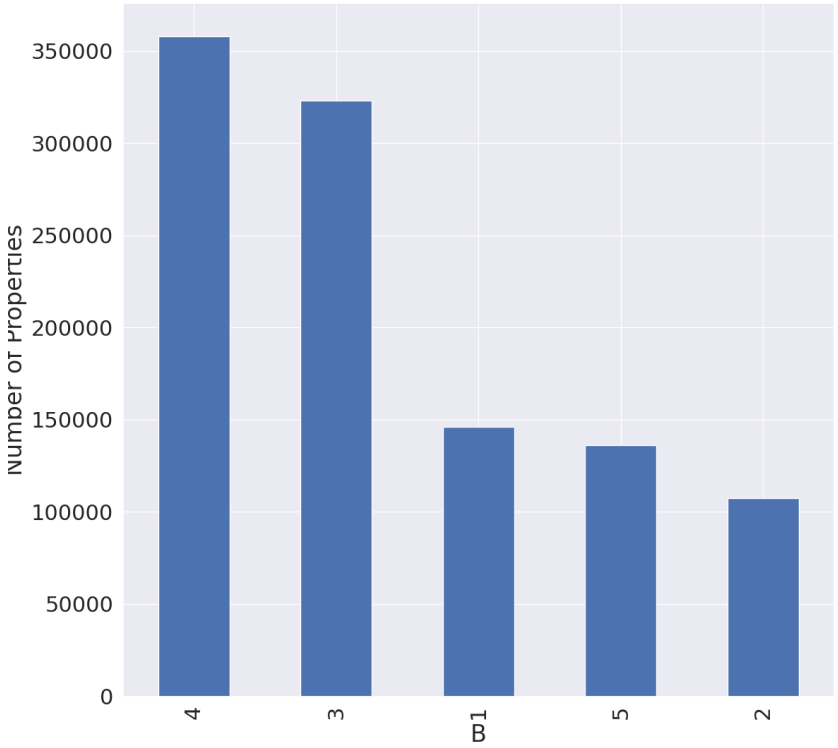
Field 2: BBLE

Description: Categorical data field including the concatenation of borough code, block code and lot.

Field 3: B

Description: Categorical data field including NYC's borough code.The bar chart below shows the total number of property records per borough in the dataset. Queens has the greatest number of property records, while Bronx has the least number of property records.

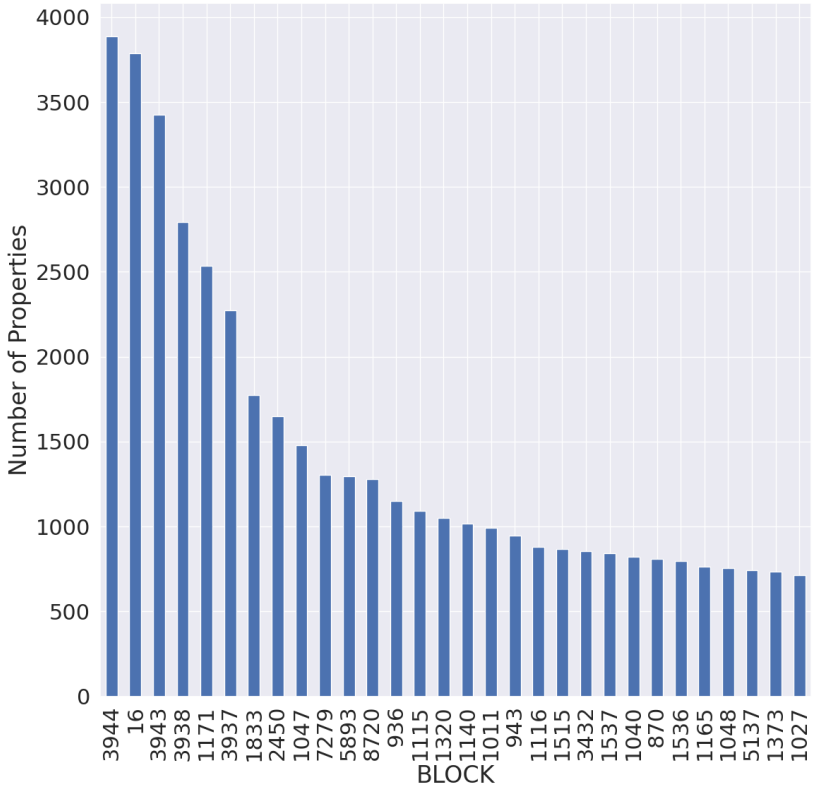
Borough Code	Borough Name
1	Manhatan
2	Bronx
3	Brooklyn
4	Queens
5	Staten Island



Field 4: BLOCK

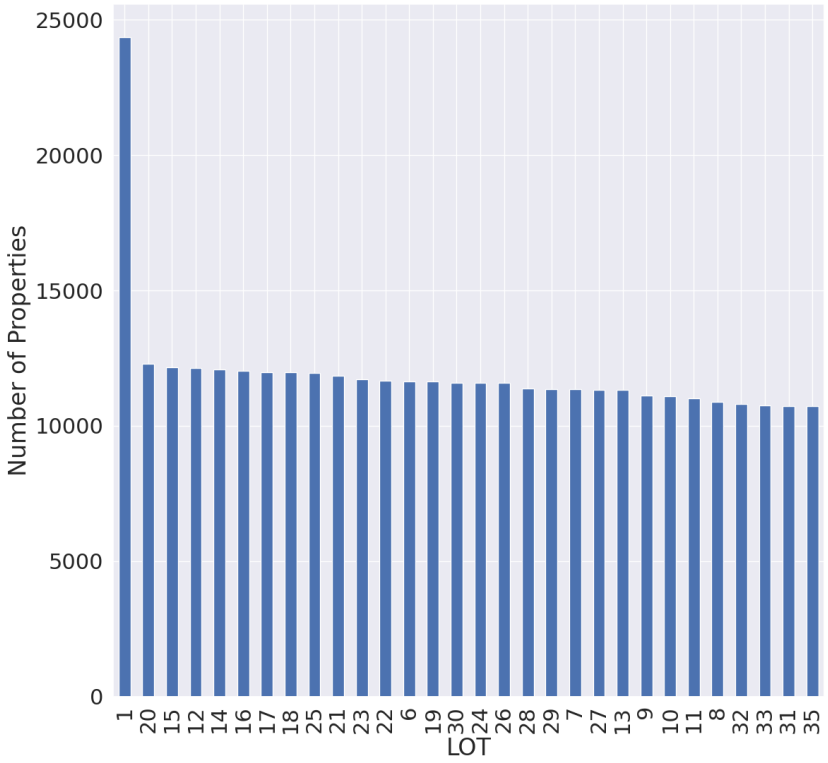
Description: Categorical data filed including valid block ranges by borough in the dataset. All records contain a block number and the boroughs have the following valid block ranges.

Field	Number of "0" or "1" value
Borough Name	Valid Block Range
Manhattan	1 to 2255
Bronx	2260 to 5958
Brooklyn	1 to 8955
Queens	1 to 16350
Staten Island	1 to 8050



Field 5: LOT

Description: Categorical data field including integer values for the unique lot number within the block of the borough. All records in the dataset contain a lot number. The bar chart below shows the top 30 lot numbers with the most property records in the datset.

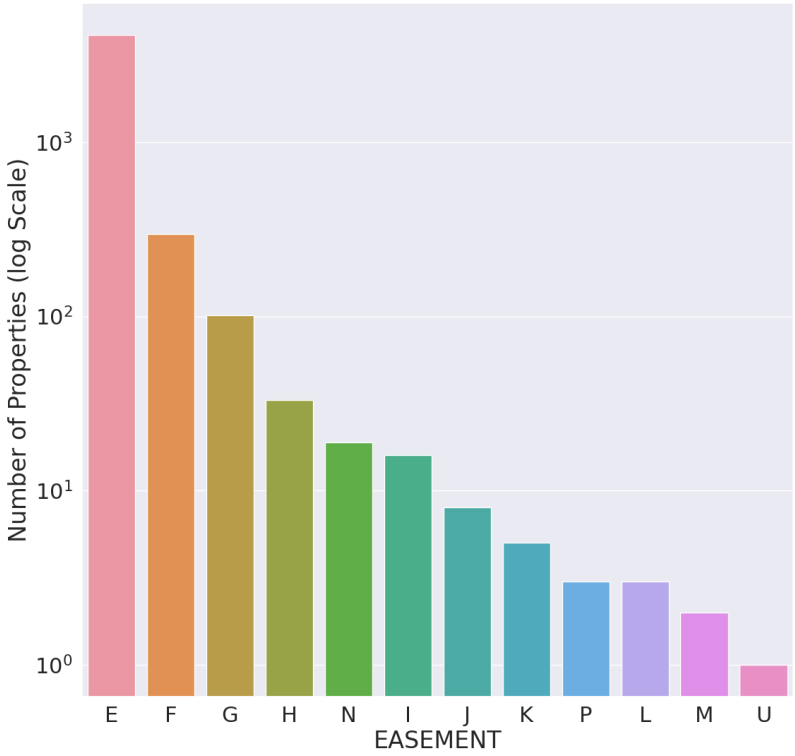


Field 6: EASEMENT

Description: Categorical data field including a letter code for any real estate easements authorized. Only 4936 records in the dataset contain an easement code. The easement letter codes are described in table.

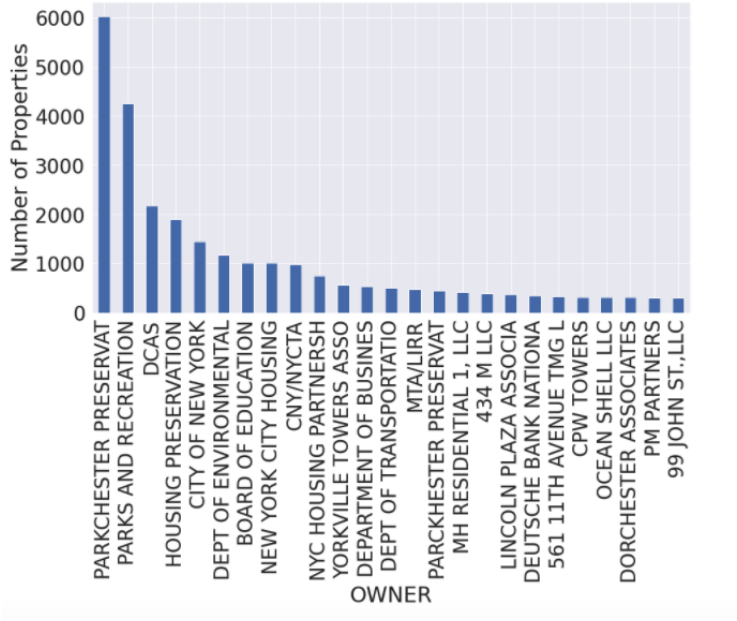
Easement Code	Description
SPACE	Indicates the lot has no Easement
'A'	Indicates the portion of the Lot that has an Air Easement
'B'	Indicates Non-Air Rights
'E'	Indicates the portion of the lot that has a Land Easement
'F' THRU 'M'	Duplicates of 'E'
'N'	Indicates Non-Transit Easement
'P'	Indicates Piers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

The bar char below shows the number of property records of each easement code in the datset. 'E' is the most common values and indicates that the property record has a Land Easement.


Number of Properties (log Scale)
10^3
10^2
10^1
10^0
E
F
G
H
N
I
J
K
P
L
M
U
EASEMENT

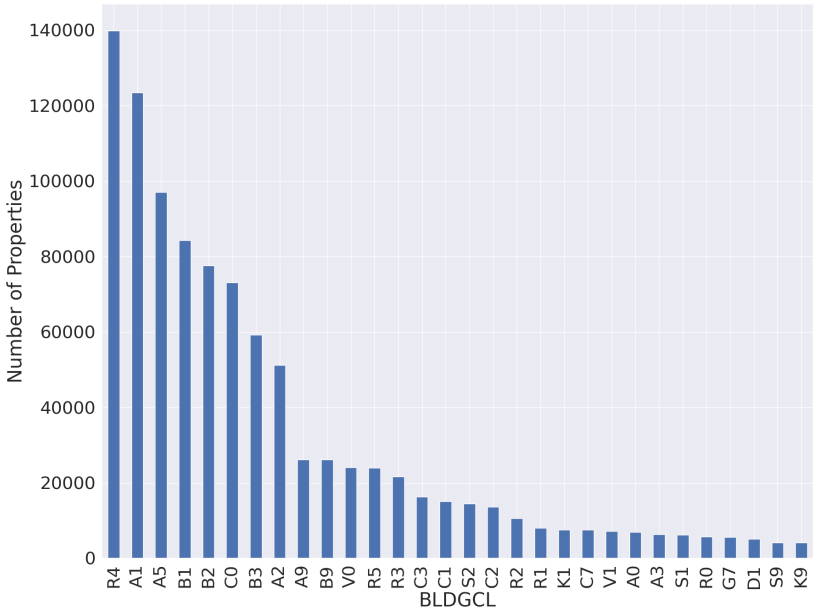
Field 7: OWNER

Description: Categorical data field including the name of the owner. There are 31745 property records without an owner listed. The bar chart below shows the top 25 owners with the most property records in the dataset. Parkchester Preservation Management has the most property records among the 863346 property owners in the data set.



Field 8: BLDGCL

Description: Two-Character length data filed including the NYC building classification code. The first character in the building classification code is a letter and the second character is a number. All records in the dataset contain a building classification code. The bar chart below shows the top 15 building classification with the highest number of properties in the dataset. The result indicates that R4 is them most common code among the 200 unique values in BLDGCL field.



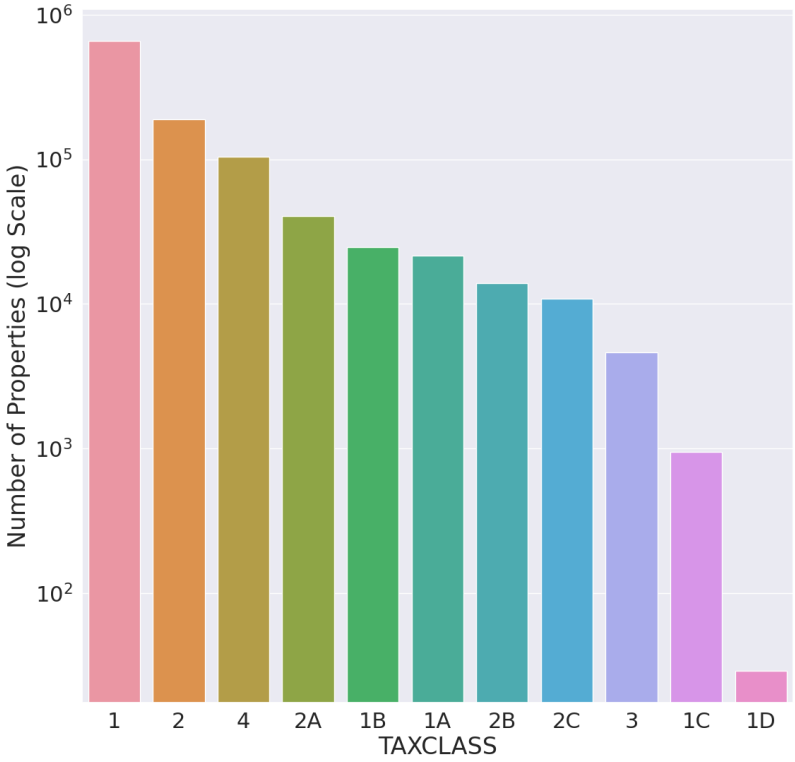
Field 9: TAXCLASS

Description: Two-Character length alphanumeric data field including the tax classification code. All records in the dataset contain a tax classification code and the valid classification codes are as follows.

Tax Classification Code	Building Type
1A	1-3 UNIT RESIDENCES
1B	1-3 STORY CONDOMINUMS
1C	UNIT CONDOMINUMS
1D	SELECT BUNGALOW COLONIES
2	APARTMENTS
2A	APARTMENTS WITH 4-6 UNITS
2B	APARTMENTS WITH 7-10 UNITS
2C	COOPS/CONDOS WITH 2-10 UNITS
3	UTILITIES-CEILING RAILROADS
4A	UTILITIES-CEILING RAILROADS
4	ALL OTHERS

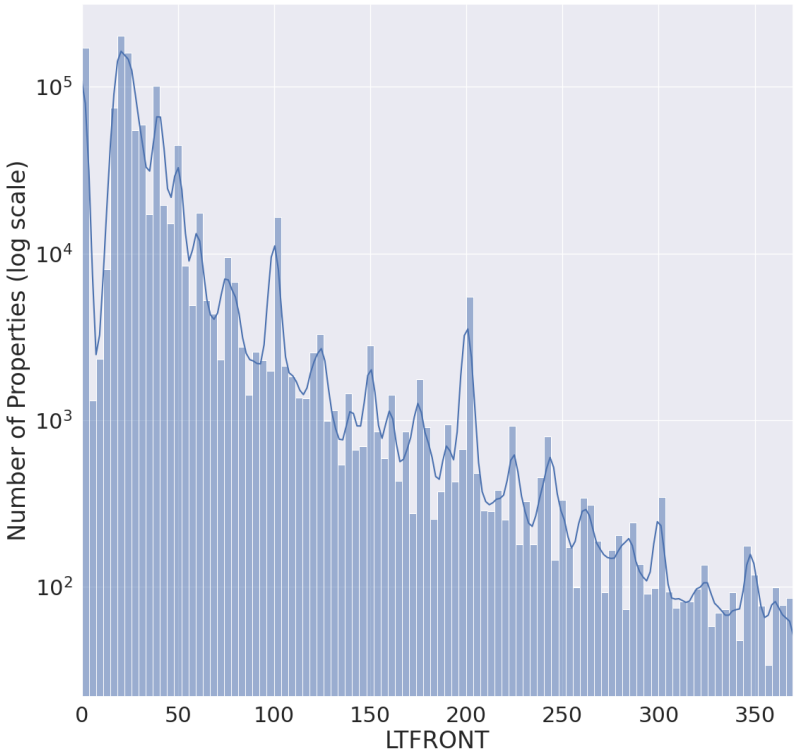
In addition, the first character of the tax classification code is based on the following classification code.

The bar chart below shows all tax classification codes in the dataset and tax classification code 1 is the most common value for this data field.


10^6
10^5
10^4
10^3
10^2
Number of Properties (log Scale)
1
2
4
2A
1B
1A
2B
2C
3
1C
1D
TAXCLASS

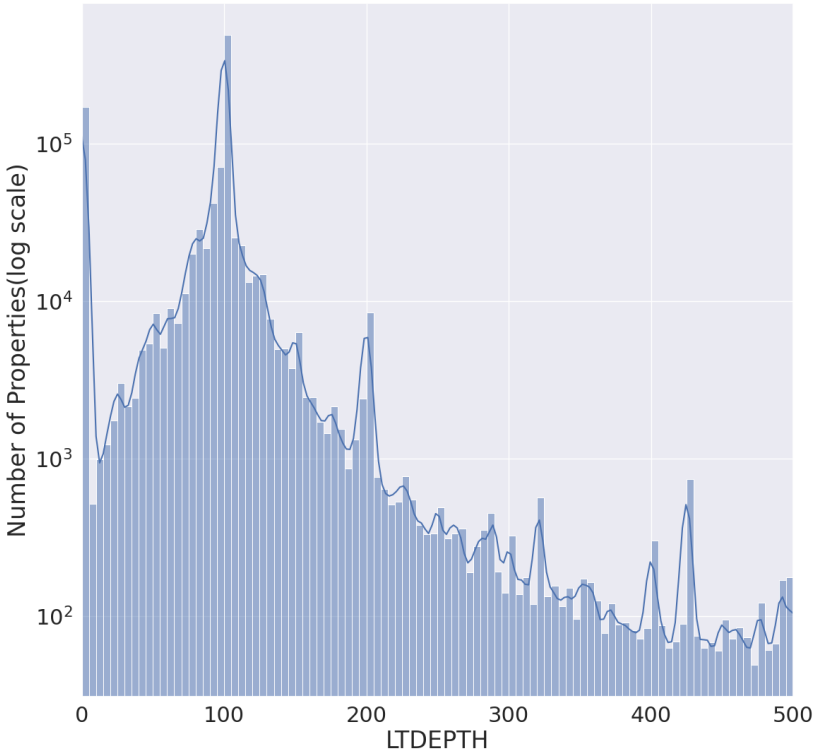
Field 10: LTFRONT

Description: Integer data field including the measurement of the property lot's front width in feet. All records in the dataset contain a front width lot measurement. Excluding those property records with a measurement bigger than 370, the distribution below shows the front width lot measurement for the property records in the dataset. The most common measurement of lot's front width is zero feet with 169108 property records having this erroneous value. The second common measurement of lot's front width is 20 feet with 135178 property records.



Field 11: LTDEPTH

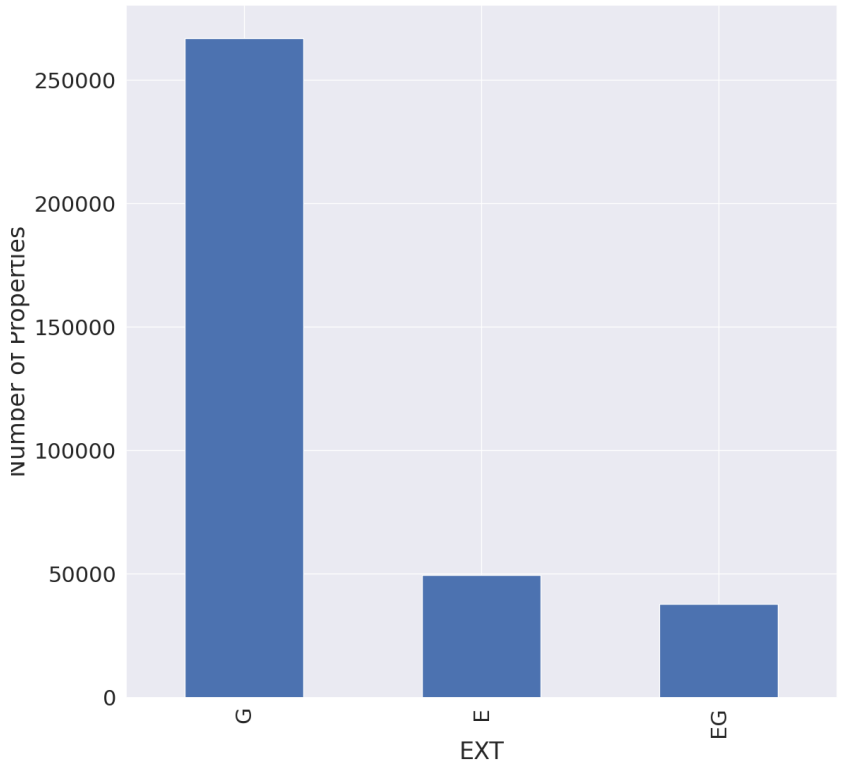
Description: Integer data field including the depth of the lot measured in feet. All records in that dataset contain a depth of the lot measurement. Excluding those property records with a measurement bigger than 500, the distribution below shows the lot depth measurements for the property records in the dataset. The most common lot depth for the property is 100 feet with 464541 property records. The second most common lot depth is zero feet with 170128 property records.



Field 12: EXT

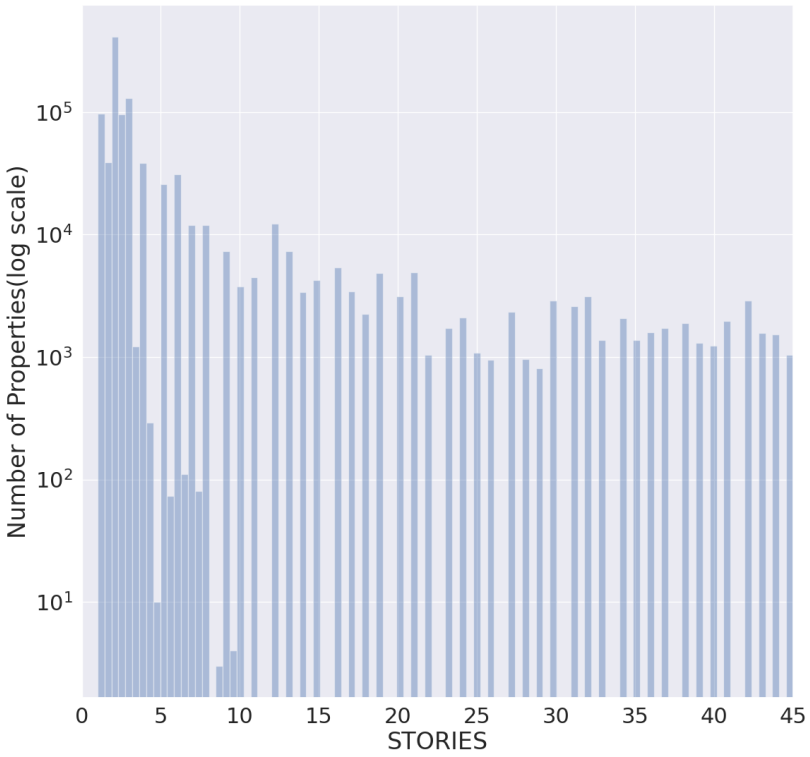
Description: Categorical data field including a letter code for the type of extension. Only 354305 property records contain an extension code. The letter codes are representative as follows:

Extension Code	Description
'E'	EXTENSION
'G'	GARAGE
'EG'	EXTENSION AND GARAGE



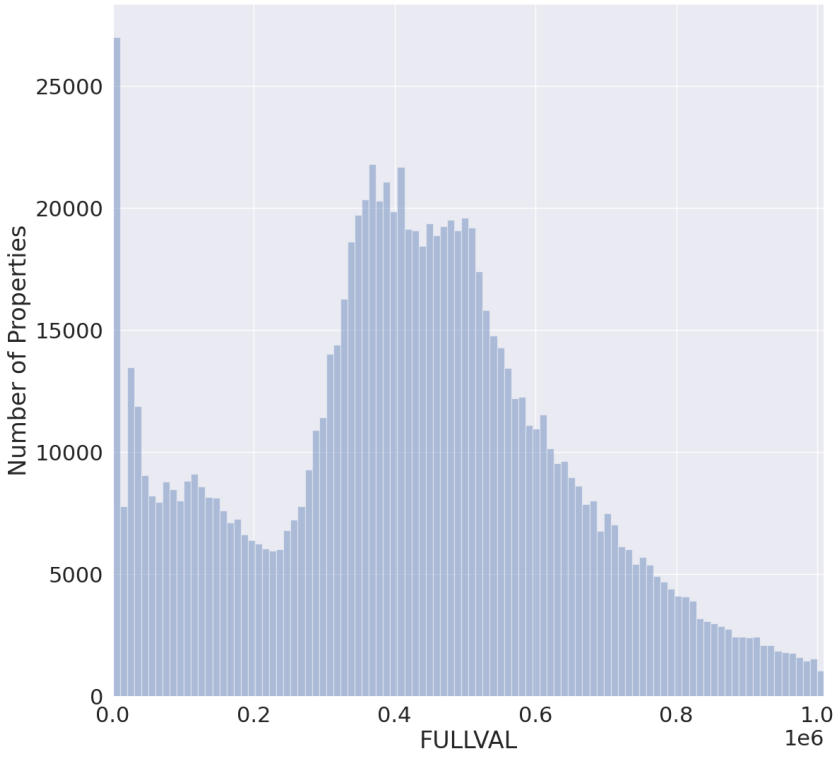
Field 13: STORIES

Description: Numerical data field including the number of stories for the building. This data field has 56264 property records with no value for the number of stories in the dataset. Excluding those property records with a measurement bigger than 45, the distribution below shows the number of stories data field for the property records. The most common number of stories is 2 with 415092 property records. The highest number of stories for a building in the dataset is 119.



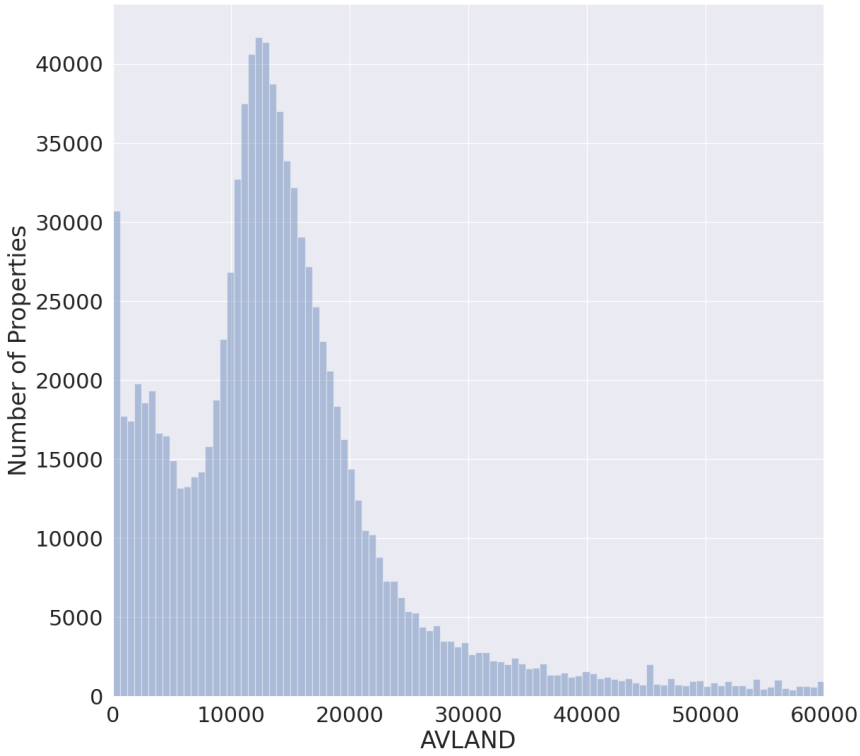
Field 14: FULLVAL

Description: Numerical data field including the full market value of the property. All records in the dataset contains a full market value of the property and the range is from 0 to 6.15 billion. Excluding those property records with a measurement bigger than 1009999, the distribution plot below shows the histogram of the full market value of the property data field for the property records in the dataset.



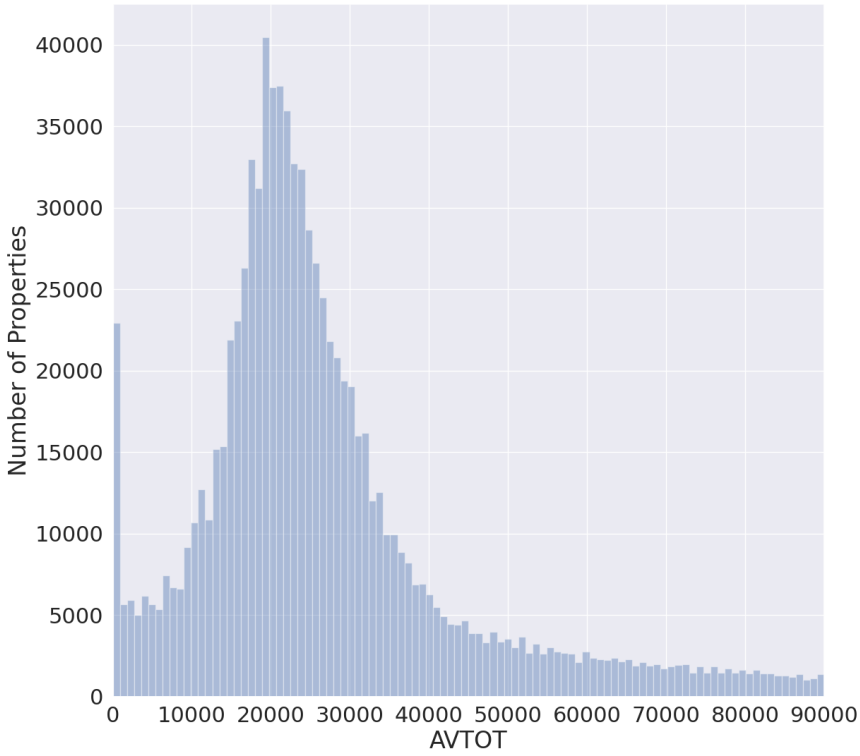
Field 15: AVLAND

Description: Numerical data field including the actual value of the land on the property. All records in the dataset contain the value for this data field and the range is from 0 to 2.6685 billion. Excluding those property records with a measurement bigger than 60000, the distribution plot below shows the histogram of the actual value of the land data field for the property records in the dataset.



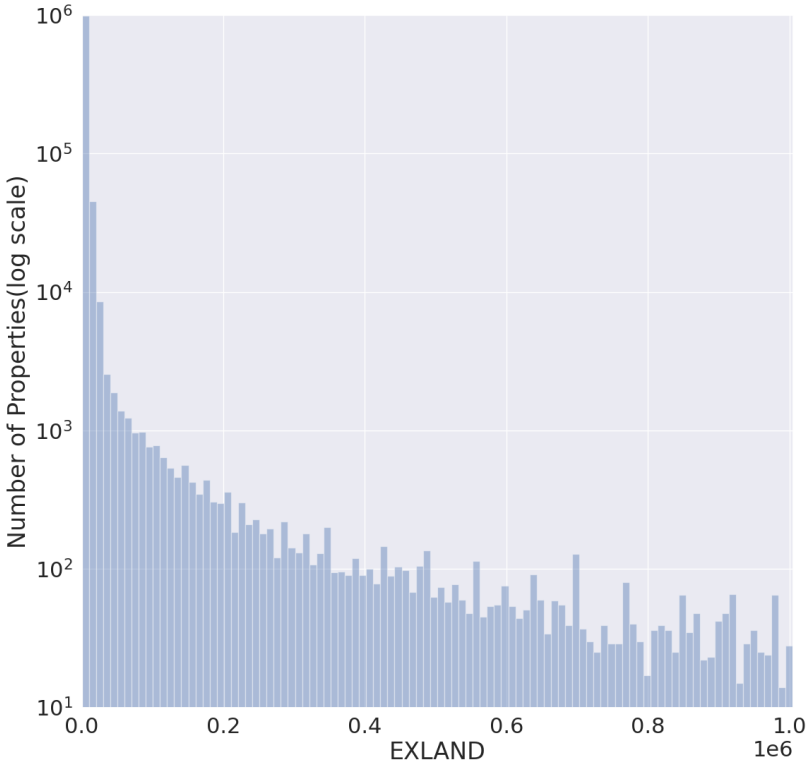
Field 16: AVTOT

Description: Numerical data field containing the actual total value of the property. All records in the dataset contain the value for this data field and the range is from 0 to 4.6683 billion. Excluding those property records with a measurement bigger than 90000, the distribution plot below shows the histogram of the actual total value data field for the property records in the dataset.



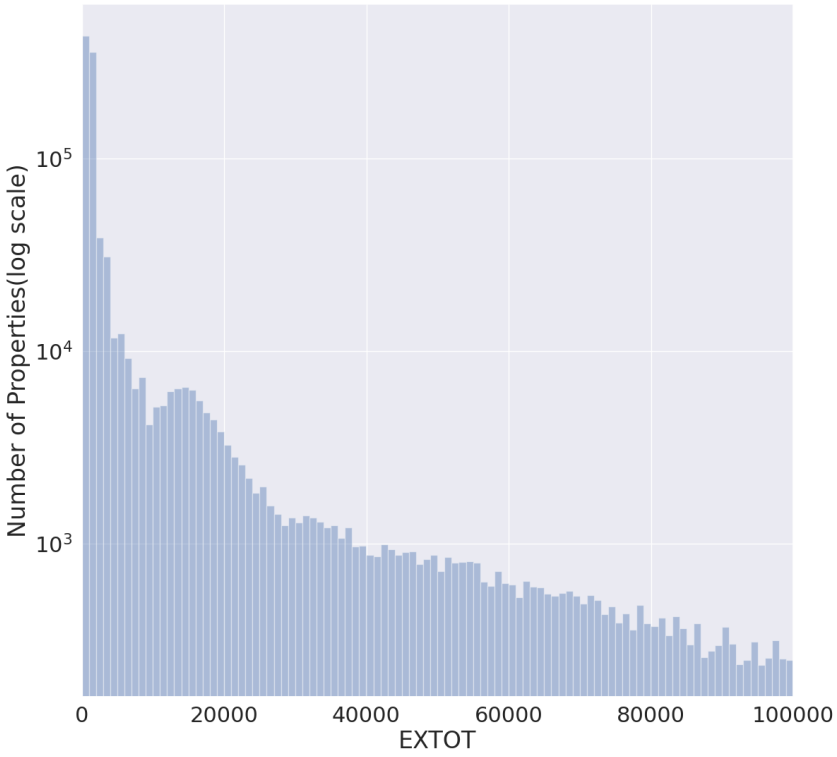
Field 17: EXLAND

Description: Numerical data field including the actual value of the exempt land. All records in the dataset contain the actual value and the range is from 0 to 2.6685 billion. Excluding those property records with a measurement bigger than 1005000, the graph below shows the histogram of the actual value of exempt land data field in the dataset.



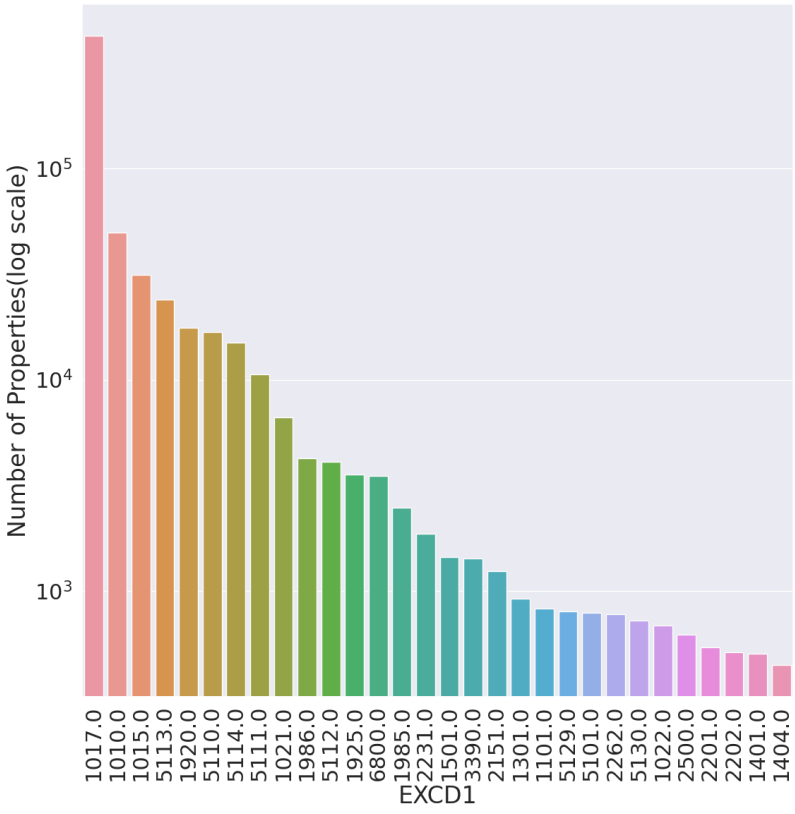
Field 18: EXTOT

Description: Numerical data field including the actual total value of the exempt land. All records in the dataset contain the actual total value and the range is from 0 to 4.6683 billion. Excluding those property records with a measurement bigger than 100000, the graph below shows the histogram of the actual total value of the exempt land data field in the dataset.



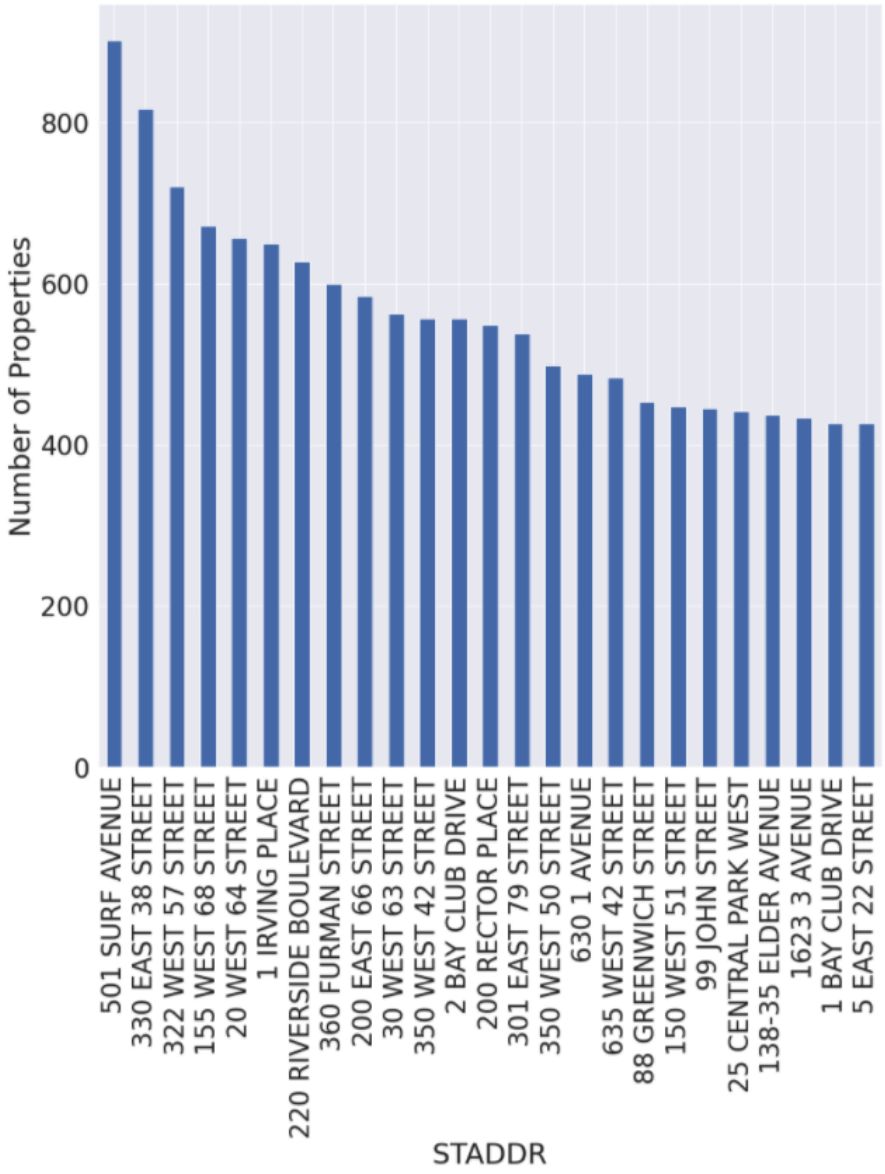
Field 19: EXCD1

Description: Categorical data field including exemption code 1 for the property. Only 638488 property records contain the exemption code 1 for the property in the dataset. The bar chart below shows the top 30 exemption codes with most property records. There is 129 unique values for these codes and 1017 is the most common codes.



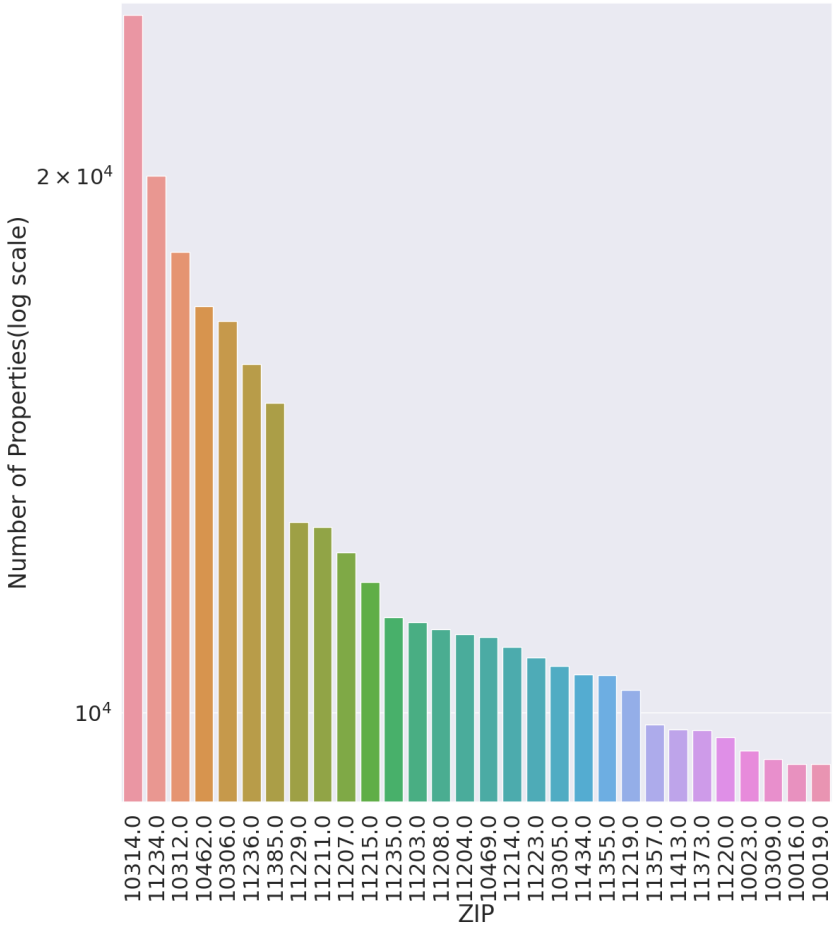
Field 20: STADDR

Description: Categorical data field including the street address of the property. Only 1070318 property records contain the street address. The bar chart below shows the top 25 street addresses with the most property records in the dataset.



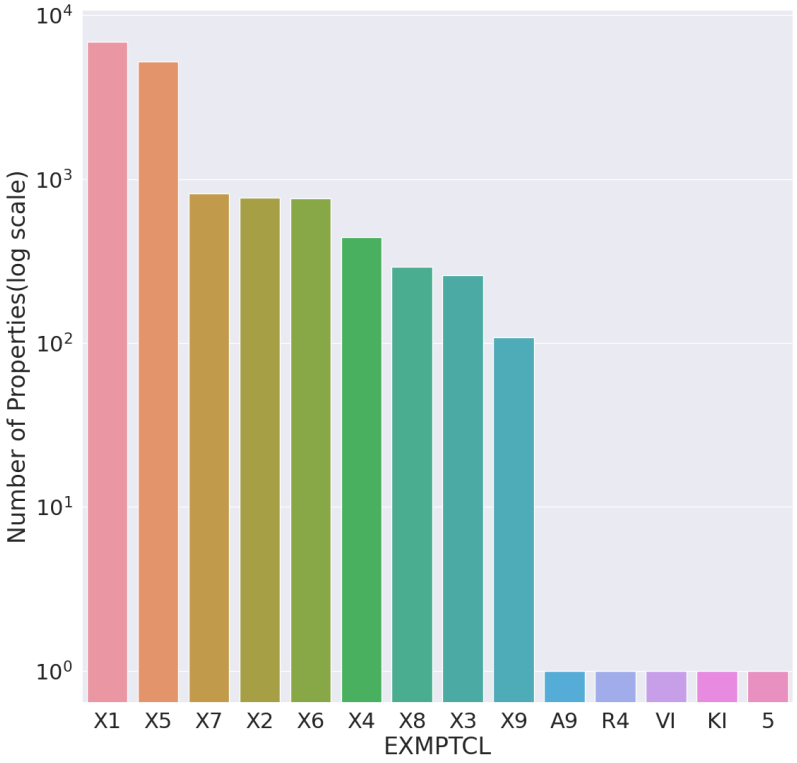
Field 21: ZIP

Description: Categorical data field including the zip code for the property record. The bar chart below shows the top 15 zip code with the most property records in the dataset. There are 196 different zip codes in total.



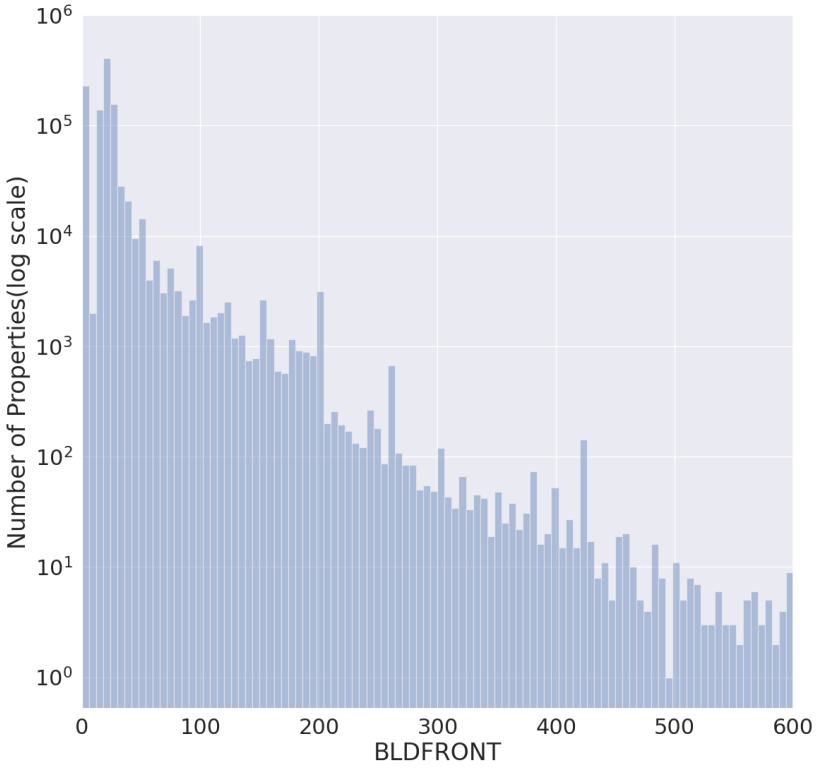
Field 22: EXMPTCL

Description: Two-Character alphanumeric data field including the exemption classification code used for fully exempt properties. Only 15579 property records contains the exemption classification code in the dataset. The bar chart below shows all exemption classification codes.



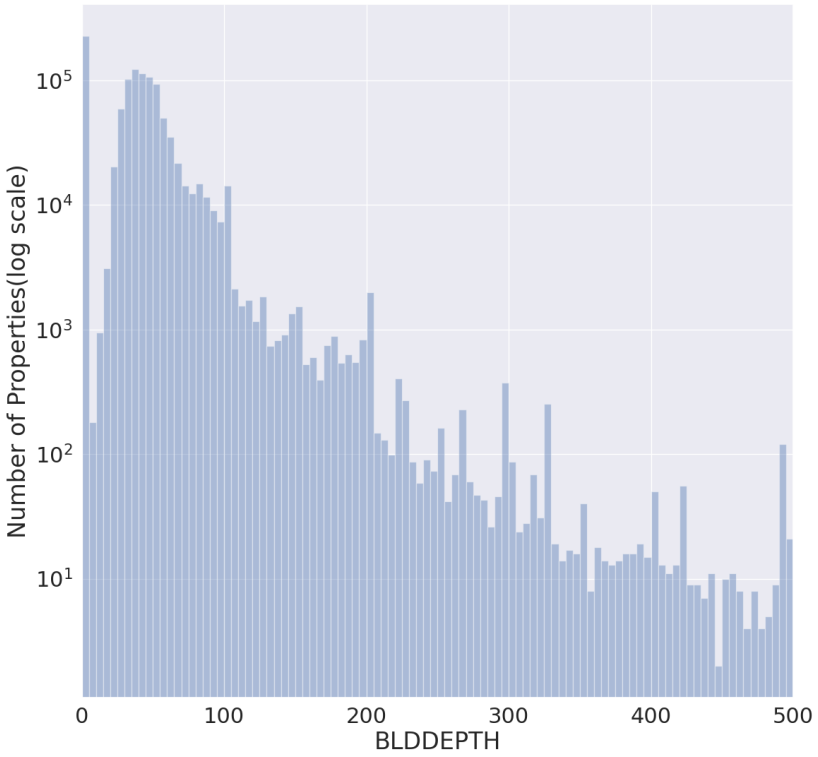
Field 23: BLDFRONT

Description: Integer data field including the front width of the building measured in feet. All records in the dataset contain a front width of the measured building. Excluding those property records with a measurement bigger than 600 feet, the distribution plot below shows the building front width data field for the property records in the dataset. The most common measurement is zero feet with 228815 property records having this erroneous value. The second common measurement is 20 feet with 195101 property records. A majority of the buildings have the front width of 26 feet or less.



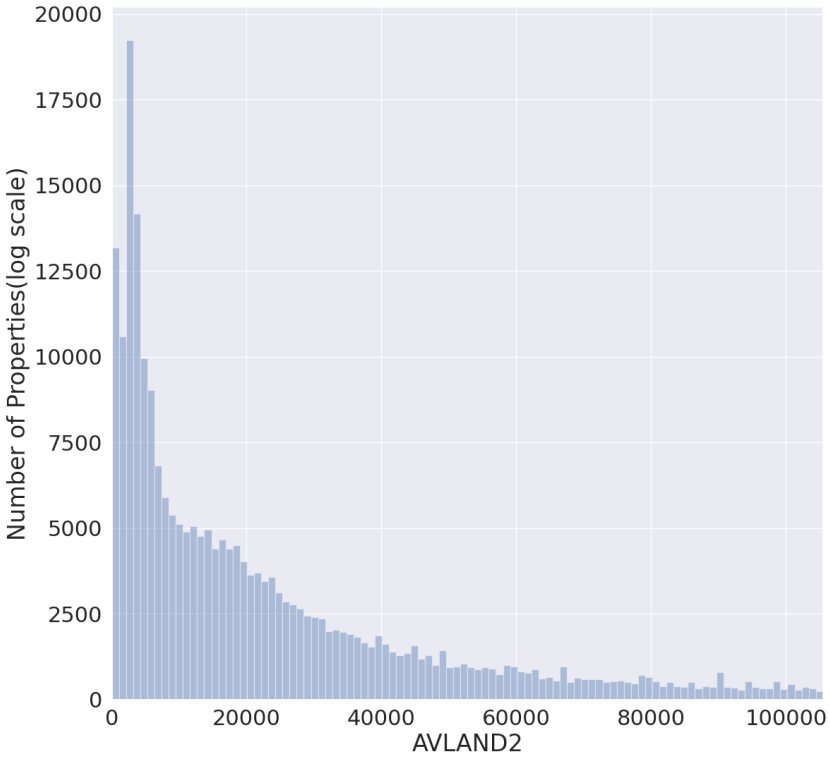
Field 24: BLDDEPTH

Description: Integer data field including the depth of the building measured in feet. All records in the dataset contain the depth of measured building. The longest measurement for the front width of the building is XX .Excluding those property records with a measurement bigger than 500 feet, the distribution plot below shows the building depth data field for the property records in the dataset. The most common measurement is zero feet with 228853 property records. The second common measurement is 40 feet with 48775 property records. A majority of the buildings have a depth of 60 feet or less.



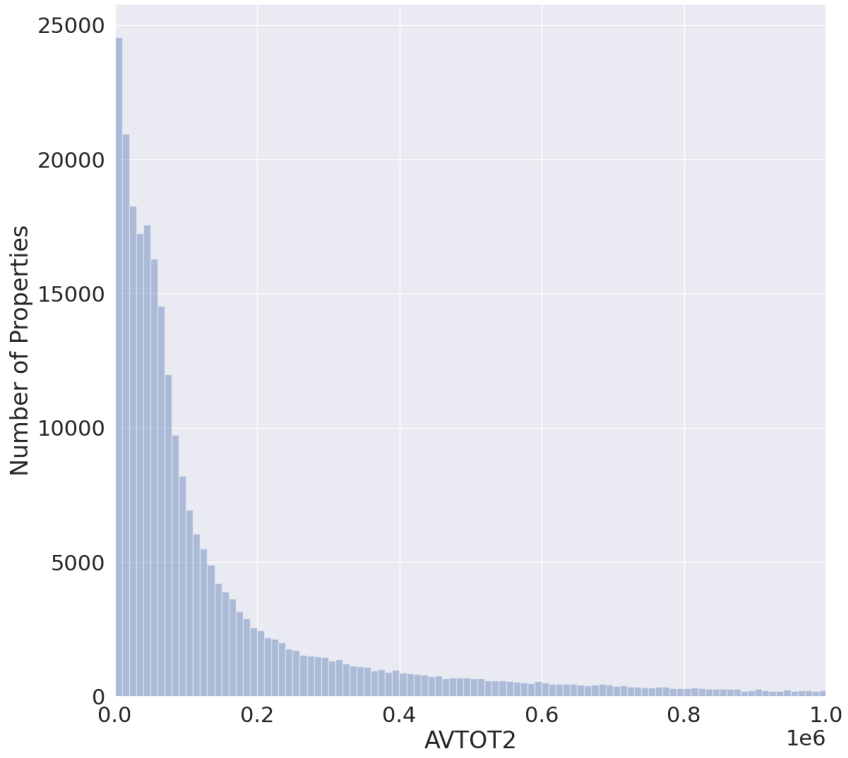
Field 25: AVLAND2

Description: Numerical data field including the transitional assessed land value. Excluding those property records with a measurement bigger than 105500, the graph below shows the histogram of the transitional assessed land value in the dataset.



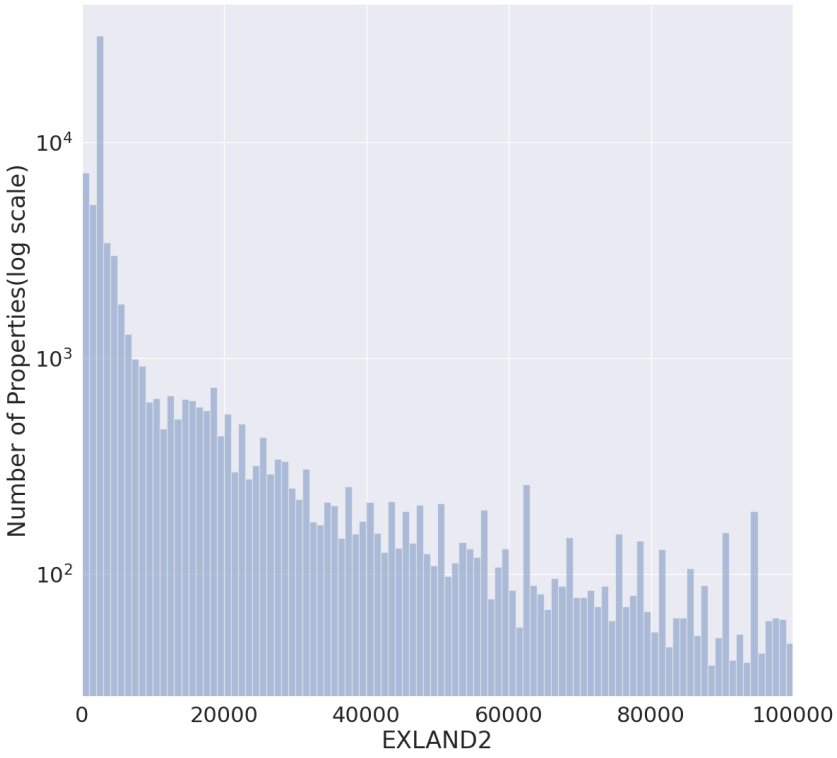
Field 26: AVTOT2

Description: Numerical data field including the transitional assessed value for exempt land. Only 87449 property records contain this data field and the range of values is from 1 to 2.371 billion. Excluding those property records with a measurement bigger than 1000000, the graph below shows the histogram of the transitional assessed value for exempt land value in the dataset.



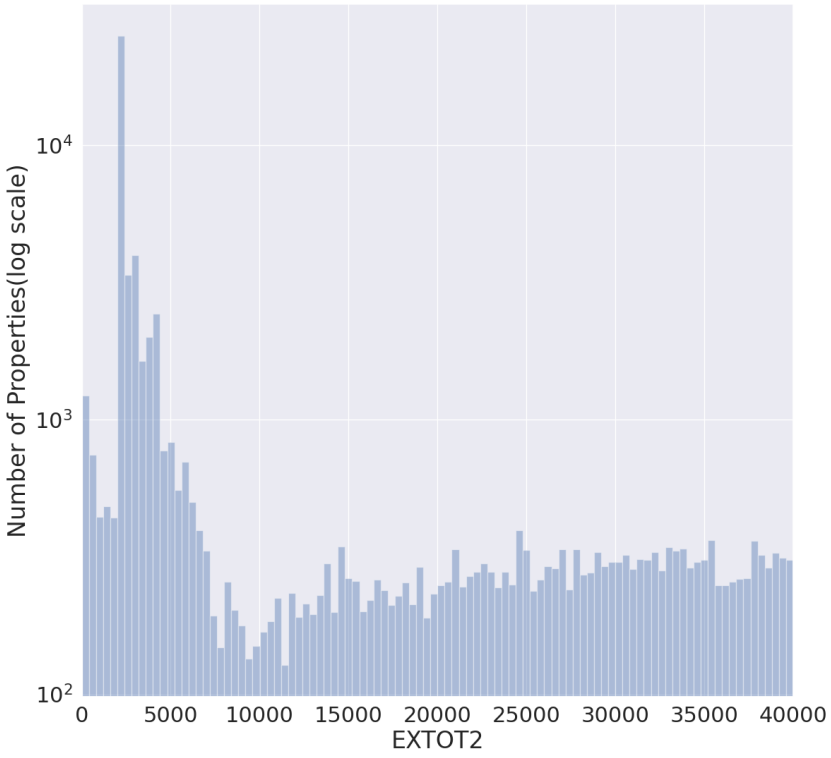
Field 27: EXLAND2

Description: Numerical data field including the transitional assessed value for exempt land. Only 87449 property records contains this value and the range of values is from 1 to 2.371 billion. Excluding those property records with a measurement bigger than 100000,the graph below shows the histogram of the transitional assessed value for exempt land data field in the dataset.



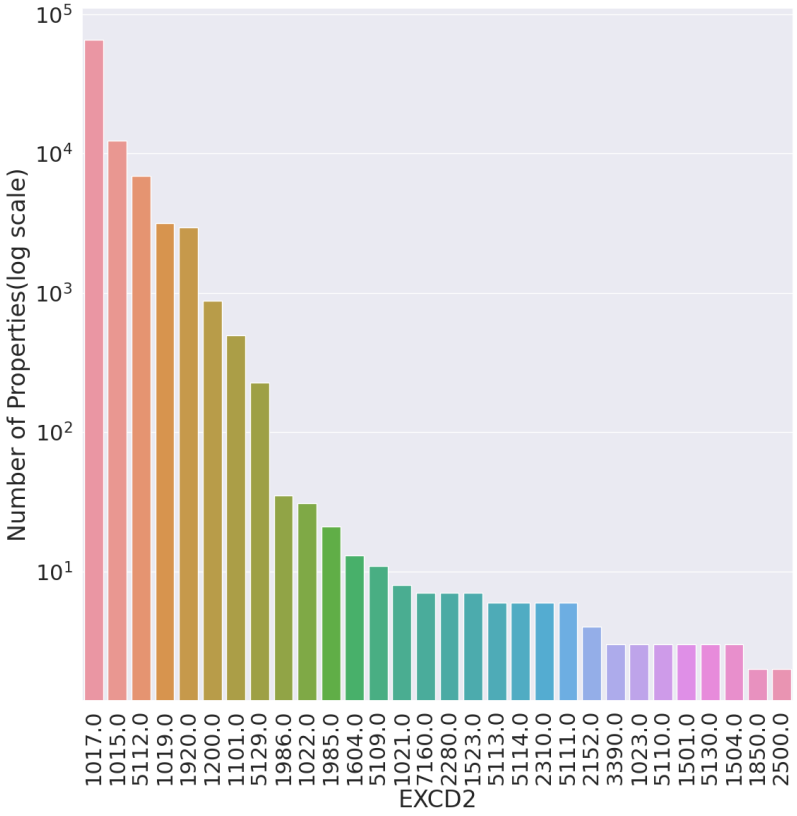
Field 28: EXTOT2

Description: Numerical data field including the transitional assessed total value for exempt land. Only 130828 property records contain this data field and the range of values is from 7 to 4.50118 billion. Excluding those property records with a measurement bigger than 40000, the graph below shows the transitional assessed total value for exempt land value in the dataset.



Field 29: EXCD2

Description: Categorical data field including exemption code 2 for the property. Only 92948 property records contain this data field. The bar char below shows the top 30 exemption codes with the most property records in the dataset.



Field 30: PERIOD

Description: A data field containing the assessment period when the file was created as described in the Access Database provided by NYC Open Data; however, all records in the data set imply contain the same value of "final" in the data field.

Field 31: YEAR

Description: A data field containing the assessment year for the property based on the NYC's data (beginning from July 1st of the calendar year and ending on the June 30th of the following calendar year). All records in the data set contain the same value of "2010/11" in the data field.

Field 32: VALTYPE

Description: A categorical data field with no description provided by the data owners (NYC Open Data) or the NYC Department of Finance; however, "VALTYPE" is often associated with the "value type" of a particular data field. This data field appears to have little value to the data set at the time, because all rows in the data set share the same value of "AC-TR" in this field.