NY Property Tax Fraud Identification

Image via iStock

Fangyu Lo, MSBA

Qing Gong, MSBA

Biyun Jing, MSBA

Chenyue Wang, MSBA

Ryan DiBenedetto, MSBA

April 27th, 2021

Project Advisor:
Professor Stephen Coggeshall

Table of Contents

Executive Summary

In recent years, the booming real estate market in New York City has the city concerned about a possible increase in tax fraud attempts. An increase could cost government agencies millions or even billions of dollars in lost revenues. The city was highly concerned about property owners misrepresenting property attributes to reduce the amount of taxes owed. As a result, a fraud analysis project was initiated to investigate properties with possible fraudulent attributes and to produce recommendations regarding these records.

This project will be focusing on investigating and identifying the anomalies and fraudulent properties in about a million NY property records. The goal of this project is to apply a robust fraud analytics methodology to identify fraudulent property values. The New York Property Data was cleaned by identifying and removing irrelevant records (exclusions) and filling in missing values (imputation). Moreover, 45 new variables were created by representative measures to detect unusual values. Next, we carried out a dimensionality reduction to convert the created variables to the Z scale and using PCA to remove correlations and reduce dimensions, scaling again after PCA. Later, we created fraud model algorithms to apply two outlier detection methods and get two fraud scores. We then combined the two scaled fraud scores as the final fraud score. Lastly, we analyzed a sample of the final fraud scores generated from the previous step and investigated potential fraudulent properties.

The properties we analyzed based on fraud scores and modeling results are listed below. We recommend further investigation if the property is identified as a highly fraudulent one according to our methodology.

Table 0. property result analysis

Item	Record	Recommendation
1	917942	Investigate
2	684704	Investigate
3	1065870	Make correction
4	1059883	Investigate
5	151044	Not investigate
6	252834	Investigate
7	771150	Investigate
8	330292	Investigate
9	104411	Investigate
10	109067	Not investigate

1. Data Description

1.1 Dataset Description

The "NY property data.csv" is a dataset representing NYC properties assessments for the purpose of calculating property taxes, granting eligible properties exemptions, and/or abatements. Data is collected and entered into the system by various city employees. The dataset is provided by the NYC Department of Finance and some of its details are presented below. For more information on the data, please see Appendix I for reference.

- Name: Property Valuation and Assessment Data
- Purpose: NYC properties assessments to calculate property tax, grant eligible properties, exemptions, and/or abatements
- Source: NYC Open Data
- Time: Sep-2018
- Number of Fields: 32
- Number of Records: 1,070,994

1.2 Data Summary

The variables in the dataset contain both categorical variables and numeric variables. We summarized the categorical data in Table 1.2.1 and the numeric variables in Table 1.2.2. For more information on the data summary, please see Appendix I for reference.

Table 1.2.1 Categorical Variable

Item	Column name	# of records	% populated	Unique values	Most common field value
1	RECORD	1,070,994	100.00%	1,070,994	N/A
2	BBLE	1,070,994	100.00%	1,070,994	N/A
3	B	1,070,994	100.00%	5	4
4	BLOCK	1,070,994	100.00%	13,984	3944
5	LOT	1,070,994	100.00%	6.366	1
6	EASEMENT	4,636	0.43%	12	E
7	OWNER	1,039,249	97.04%	863,347	PARKCHESTER PRESERVAT
8	BLDGCL	1,070,994	100.00%	200	R4
9	TAXCLASS	1,070,994	100.00%	11	1
10	EXT	354,305	33.08%	3	G
11	EXCD1	638,488	59.62%	129	1017
11	STADDR	1,070,318	99.94%	839,280	501 SURF AVENUE
13	ZIP	1,041,104	97.21%	196	10314
14	EXMPTCL	15,579	1.45%	14	X1
15	EXCD2	92,948	8.68%	60	1017
16	PERIOD	1,070,994	100.00%	1	FINAL
17	YEAR	1,070,994	100.00%	1	2010/11
18	VALTYPE	1,070,994	100.00%	1	AC-TR

Table 1.2.2 Numeric Variable

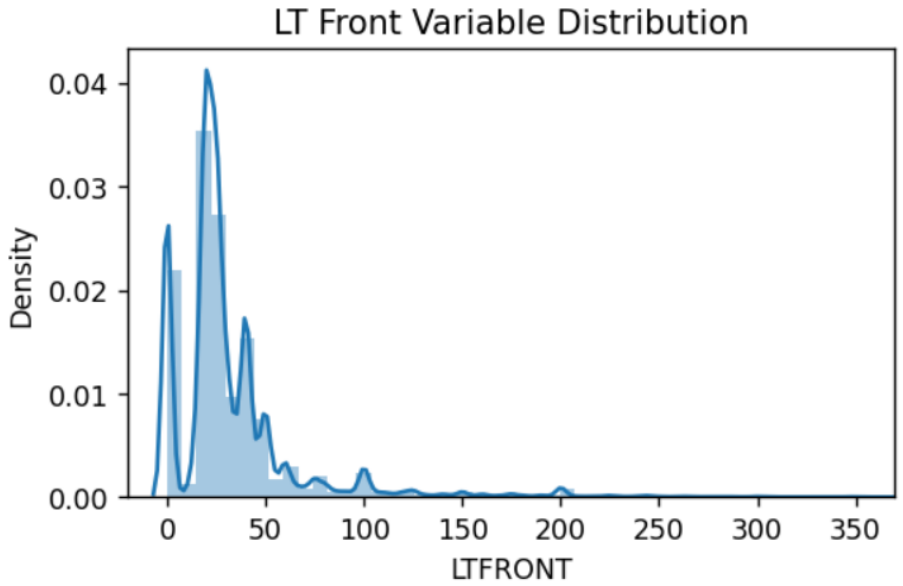
Item	Column name	# of records	% populated	Unique values	Mean	Standard deviation	Min value	Max value	# of zeros
1	LTFRONT	1,070,994	100.00%	1,297	36.64	74.03	0	9,999	169,108
2	LTDEPTH	1,070,994	100.00%	1,370	88.86	76.40	0	9,999	170,128
3	STORIES	1,014,730	94.75%	111	5.01	8.37	1	119	0
4	FULLVAL	1,070,994	100.00%	109,324	874,264.51	11,582,430.99	0	6,150,000,000	13,007
5	AVLAND	1,070,994	100.00%	70,921	85,067.92	4,057,260.06	0	2,668,500,000	13,009
6	AVTOT	1,070,994	100.00%	112,914	227,238.17	6,877,529.31	0	4,668,308,947	13,007
7	EXLAND	1,070,994	100.00%	33,419	36,423.89	3,981,575.79	0	2,668,500,000	491,699
8	EXTOT	1,070,994	100.00%	64,255	91,186.98	6,508,402.82	0	4,668,308,947	432,572
9	BLDFRONT	1,070,994	100.00%	612	23.04	35.58	0	7,575	228,815
10	BLDDEPTH	1,070,994	100.00%	621	39.92	42.71	0	9,393	228,853
11	AVLAND2	282,726	26.40%	58,591	246,235.72	6,178,962.56	3	2,371,005,000	0
12	AVTOT2	282,732	26.40%	111,360	713,911.44	11,652,528.95	3	4,501,180,002	0
13	EXLAND2	87,449	8.17%	22,195	351,235.68	10,802,212.67	1	2,371,005,000	0
14	EXTOT2	130,828	12.22%	48,348	656,768.28	16,072,510.17	7	4,501,180,002	0

1.3 Histograms of Important Fields

In this project, we are mainly interested in analyzing fraud in specific value variables. Thus, before introducing our methodology and analysis, we first presented the histogram of the value fields for examining their distribution.

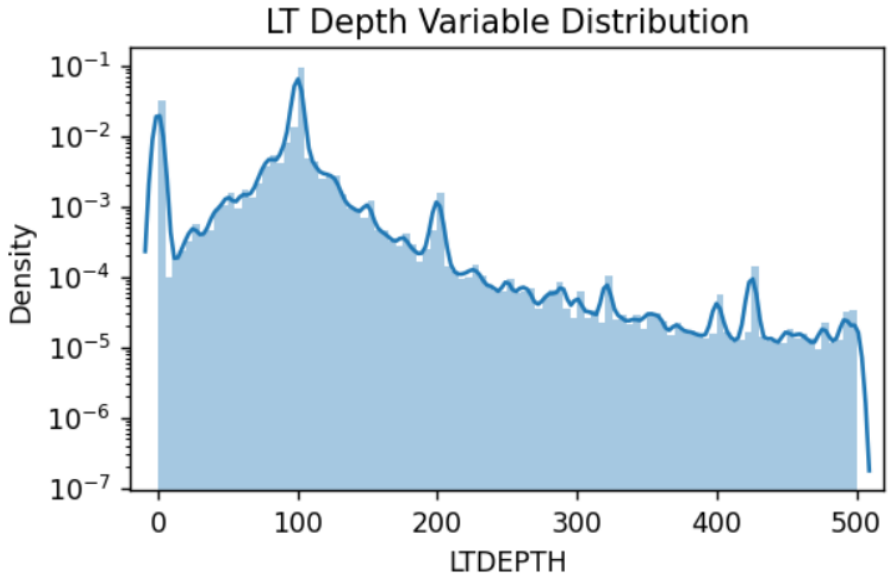
Field 1.

- Name: LTFRONT
- Description: lot frontage in feet
- Histogram: (data in the histogram excludes outliers > 370 and is 99.45% populated)



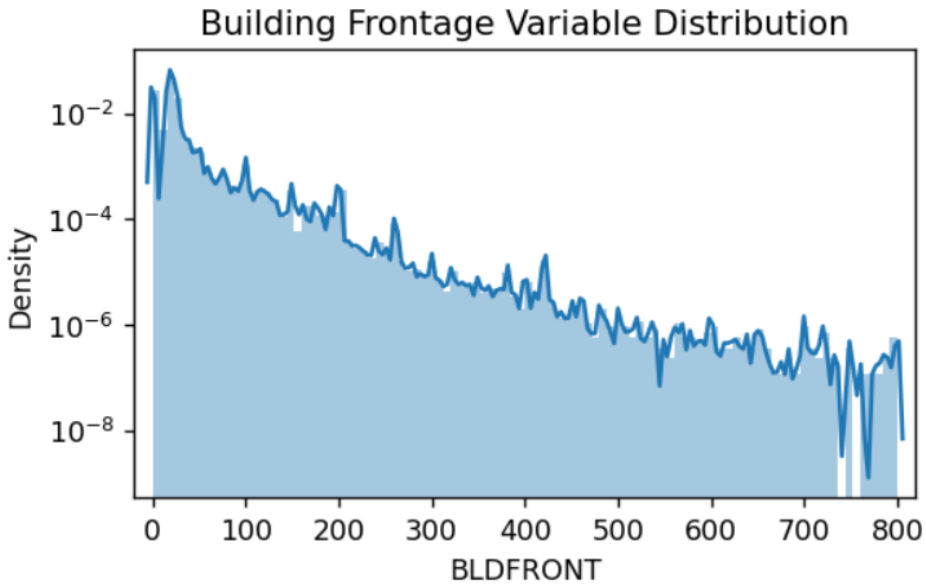
Field 2.

- Name: LTDEPTH
- Description: lot depth in feet
- Histogram: (data in the histogram excludes outliers > 500 and is 99.67% populated)



Field 3.

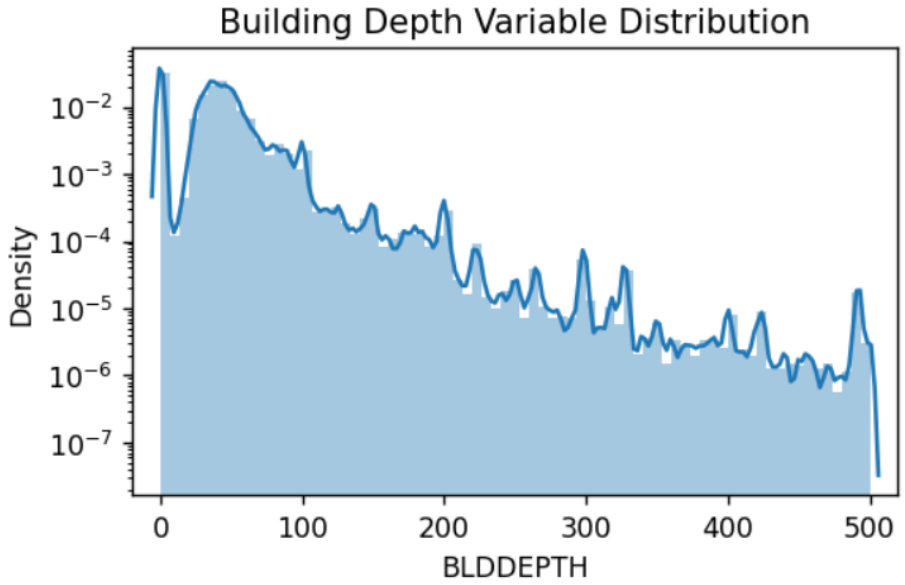
- Name: BLDFRONT
- Description: building frontage in feet
- Histogram: (data in the histogram excludes outliers > 800 and is 99.996% populated)



Field 4.

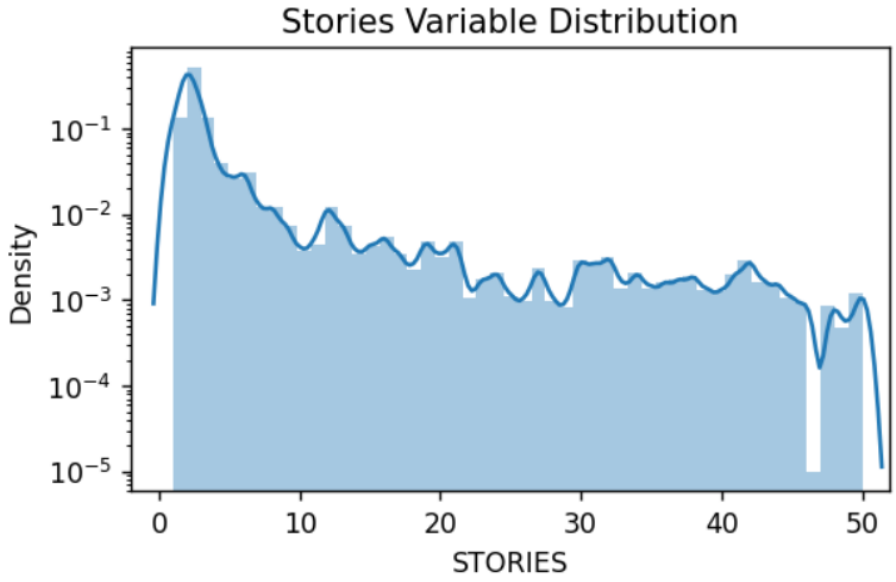
- Name: BLDDEPTH

- Description: building depth in feet
- Histogram: (data in the histogram excludes outliers > 500 and is 99.92% populated)



Field 5.

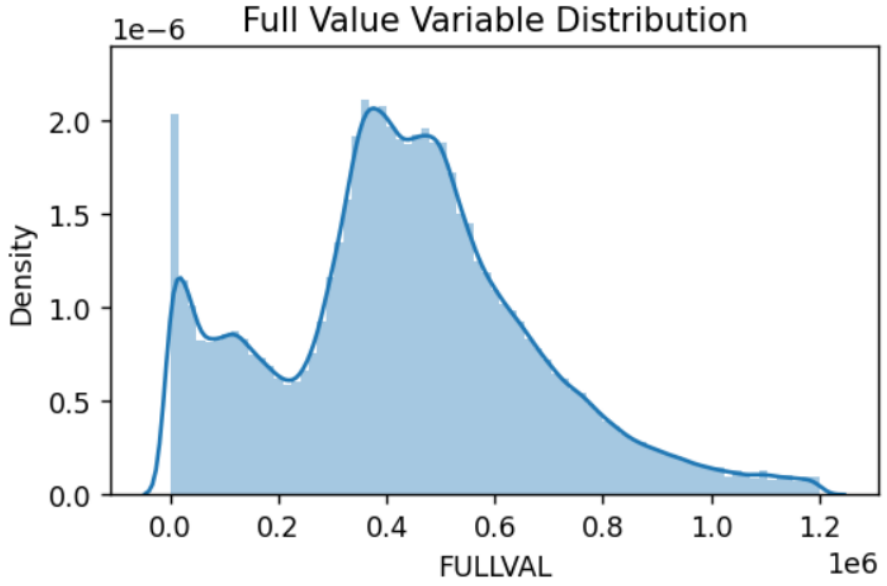
- Name: STORIES
- Description: the number of stories for the building (# of floors)
- Stories Histogram: (data in the histogram excludes outliers > 50 and is 99.5% populated)



Field 6.

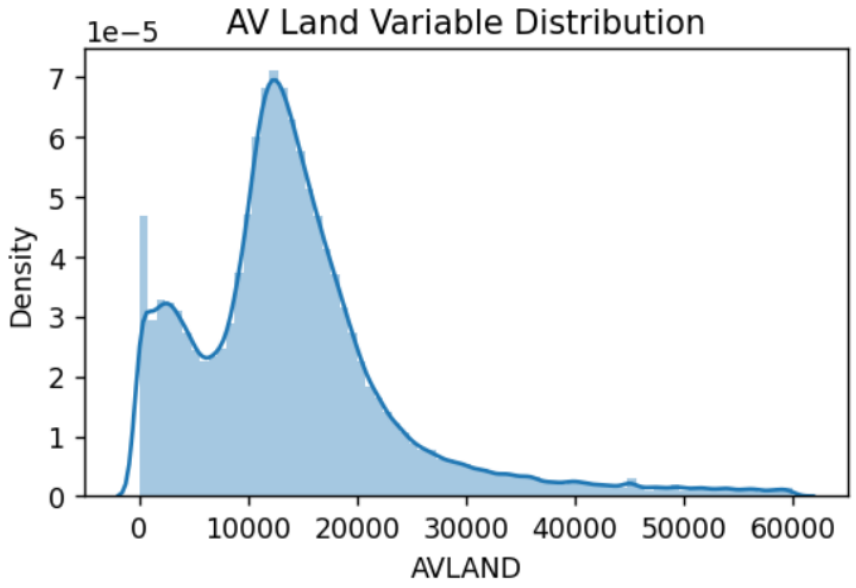
- Name: FULLVAL
- Description: total market value of the property

- Histogram: (data in the histogram excludes outliers > 1,200,000 and is 93.28% populated)



Field 7.

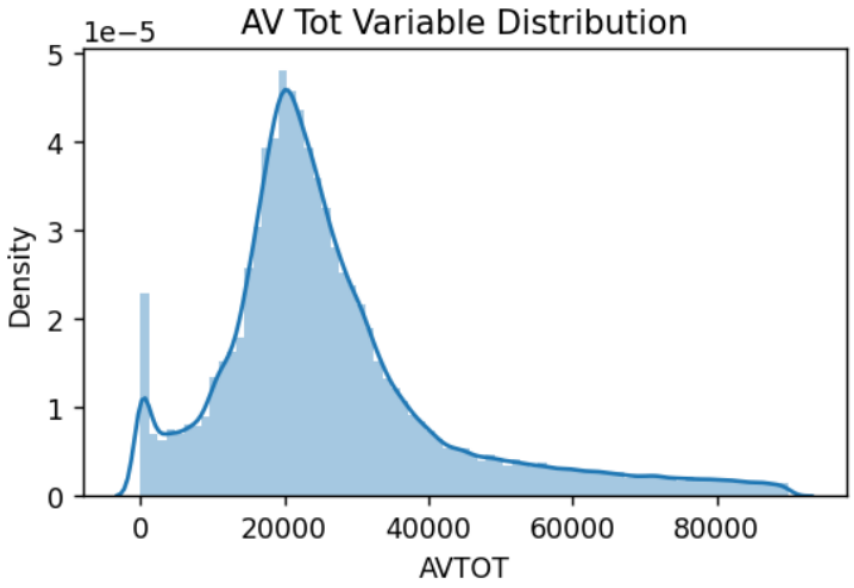
- Name: AVLAND
- Description: assessed land value
- Histogram: (data in the histogram excludes outliers > 60,000 and is 91.62% populated)



Field 8.

- Name: AVTOT
- Description: assessed total value

- Histogram: (data in the histogram excludes outliers > 90,000 and is 84.99% populated)



2. Data Cleaning

2.1 Data Exclusions

This data set includes some records that we are not interested in as part of this property fraud analysis. 24,168 records that are shown owned by the city, state, or federal government have been removed and excluded from this project.

2.2 Data Imputations

In this data set, many fields are containing missing values. We chose to fill 9 fields that are important to build algorithms.

The first is the zip code (ZIP). Since the record in the original dataset has already been sorted by zip code, we were able to fill in 11,372 records with the new zip using both the record before and the record after. For the rest, we replaced missing zip codes with the previous record's value that is available.

Secondly, we imputed the three value variables (FULLVAL, AVLAND, AVTOT). For these 3 fields, we grouped each value by tax class (TAXCLASS) and filled in a missing value with the mean of that group.

Then, we dealt with the missing value of the stories of the building (STORIES). For field STORIES, we also grouped values by TAXCLASS and filled them in with each mean.

Last but not least, we imputed the depth and frontage of the lots and buildings (LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH). Before filling in missing values, we started by replacing all 0s and 1s with NAs. Then, we calculated the mean value for each tax class and filled it in with the mean of that group.

3. Variable Creation

The representative variables allow us to build fraud models without noise and to help us analyze data in a more meaningful and interpretable way. Based on expert knowledge, we created 45 variables using the property size and value fields. We will explain more of the methodology in this section, and we will move the records only with these created variables to the next stage of the fraudulent property analysis.

We first created 3 size value fields, lot area, building area, and building volume. The reason we built size values is that usually, the bigger the size of the property, the more valuable the property should be. We multiplied the lot frontage (LTFRONT) by lot depth (LTDEPTH) to get the lot area and multiplied the building frontage (BLDFRONT) by building depth (BLDDEPTH) to get the building area. For the building volume, we multiplied the newly created building area value field by stories (STORIES).

- Lot area (S_1) = LTFRONT * LTDEPTH
- Building area (S_2) = BLDFRONT * BLDDEPTH
- Building volume (S_3) = S_2 * STORIES

Secondly, to get the standardized metric of property value per size unit, we created 9 ratio value variables using assessed and total land value as well as the size value fields. To be more specific, for each of the 3 property values (FULLVAL, AVLAND, AVTOT), we divided it by each of the 3 size values created earlier. Thus, we got 9 ratio variables (Formula 3.1) and were named r1 to r9.

- V_1 = FULLVAL
- V_2 = AVLAND
- V_3 = AVTOT

Formula 3.1 ratio variables

$$r_1 = \frac{V_1}{S_1} \quad r_4 = \frac{V_2}{S_1} \quad r_7 = \frac{V_3}{S_1}$$
$$r_2 = \frac{V_1}{S_2} \quad r_5 = \frac{V_2}{S_2} \quad r_8 = \frac{V_3}{S_2}$$
$$r_3 = \frac{V_1}{S_3} \quad r_6 = \frac{V_2}{S_3} \quad r_9 = \frac{V_3}{S_3}$$

(extracted from the MGTA 463 Fraud Analytics course taught by Professor Stephen Coggeshall)

Finally, since the location and property type are vital factors for property value estimation, we grouped the 9 ratios by 5-digit ZIP code, 3-digit ZIP code, tax class, and borough code to get the geography and tax class characteristics. We calculated the average of each ratio for each group, and we divided each of the 9 ratio variables by them to get 36 (9*4) grouped average variables. These variables will help us understand the value difference between areas and tax classes, and we could compare them with individual property values.

Figure 3.1 grouped average variables

Ratio var.
r1
r2
r3
r4
r5
r6
r7
r8
r9

+

Scale factors	Description
zip5	5-digit zip code
zip3	3-digit zip code
taxclass	Tax class
boro	Borough code

At the end of the variable creation, we combined the 9 ratio variables and the 36 grouped average variables as the 45 variables used in the further fraudulent property analysis. For more information and the statistics of the 45 variables, please see Appendix II.

4. Dimensionality Reduction

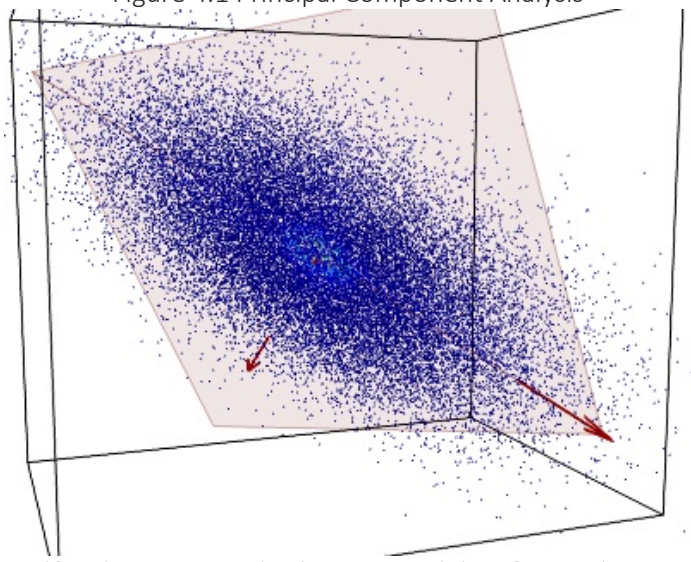
Dimensionality reduction is used to reduce the number of data dimensions and to remove strong correlations. After building the 45 expert variables, we applied dimensionality reduction to bring the number of columns down and convert the sphere to a circle (or ellipse) in the two-dimensional space. We used Principal Component Analysis (PCA), linear dimensionality reduction, to transform the data.

Before using PCA, we need to scale the data to get all the dimensions scaled similarly. We applied Z-scale (the formula below) in this step since this scaling method allows us to center the data and make all dimensions on the same footing. In other words, Z-scaled dimensions measure how many standard deviations away from the population mean, and each of their values is an individual measure of unusualness. We called the new Z-scaled variables "Z scores." They are ready to be moved on to the Principal Component Analysis.

$$z_i = \frac{x_i - \mu_i}{\sigma_i}$$

We then applied PCA on the scaling data to remove the correlated dimensions. PCA works by finding the dominant directions in the data and rotating the coordinate system along these directions through an angle θ (Formula 4.1 and Figure 4.2). If two variables x1 an x2 are highly correlated, they could be replaced by a single Principal Component (PC). We looked at the eigenvalues and decided to keep the top 6 Principal Components (PCs). We rewrote the original data in terms of this new rotated coordinate system. These PCs would then represent each record in the new reduced space. We threw away the higher PCs to finish reducing the dimensions of the data matrix.

Figure 4.1 Principal Component Analysis

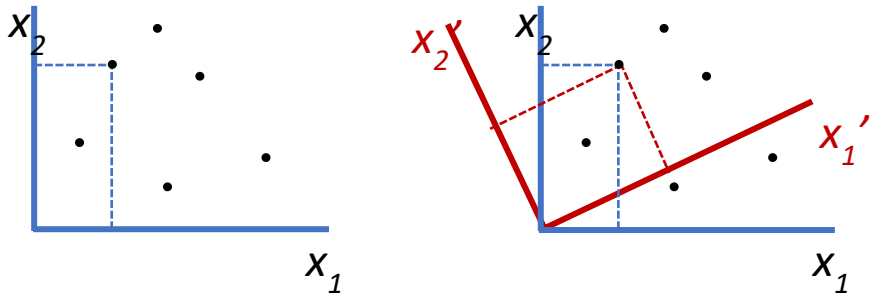


(extracted from the MGTA 463 Fraud Analytics course taught by Professor Stephen Coggeshall)

Formula 4.1 PCA new coordinate system

$$x_1' = x_1 \cos\theta + x_2 \sin\theta$$
$$x_2' = -x_1 \sin\theta + x_2 \cos\theta$$

Figure 4.2 PCA rotation and linear transformation of coordinates



(extracted from the MGTA 463 Fraud Analytics course taught by Professor Stephen Coggeshall)

After the dimensionality reduction, we Z scaled the data again to make all the remaining Principal Components equally important. These Z scores of the data fields can tell the tendency of an outlier of each variable. We will input them into our fraud model algorithms.

5. Fraud Model Algorithms

After the dimensionality reduction, we applied two different algorithms to find abnormal records. We then combined two scores from the methods to calculate a final anomaly score.

The first score is calculated based on a heuristic algorithm, where we take the Z scores we calculated in the previous part and combine them using a sum of absolute values squared and taking square roots. We used the following formula:

$$s_{1i} = \left(\sum_k |PCz_{ik}|^p \right)^{1/p}$$

In this formula, i is for each record, and k is for each of the Z- scaled Principal Components. We chose the p in the distance formula to be 2, which is the Euclidean distance.

For the second score, we calculated it using Auto-encoder on the Z-scaled Principal Components. Score 2 is the difference between the input and the output vectors of the autoencoder, also known as the reconstruction error. We first compressed the data records from 6 to 3 dimensions and expanded them back to 6 dimensions. The model would generate a significantly larger error for the abnormal records than for the regular records. Similar to the calculation of score 1, we also chose the Euclidean distance, where p equals 2. The formula is listed below:

$$s_{1i} = \left(\sum_k |PCz'_{ik} - PCz_{ik}|^p \right)^{1/p}$$

After calculating both scores, we took the average of two scores to get the final score. Finally, we sorted and ranked the records using the final score in descending order.

6. Results

By the two fraud model algorithms and the fraud scores, we got a ranking list of unusual records. We will present and examine the top 5 unusual records generated by the models. Also, we chose another 5 records that we believe are worth presenting and discussing.

Record #1: 917942
The property, reported at 154-68 BROOKVILLE BOULEVARD with 3 stories, is actually a 7-story building in Queens, New York. The building is owned by LOGAN PROPERTY INC with 62 feet frontage and 86 feet depth, and its assessed land value and total value are $1.8 billion and $4.7 billion. These numbers are far higher than the average valuations of similar properties with the same tax class, borough code, and building dimensions. This probably due to the misrepresentation in building size. Given these massive errors, this property needs further investigation.

Record #2: 684704
The property, owned by W RUFERT, is a 4-story structure in Queens, New York with an incomplete street name (69 STREET). The property value is $374k and the value of lot front and lot depth are 2 and 2 feet respectively, which doesn't make sense for a property of that valuation. The smaller lot front and lot depth value and unusually higher property valuation are abnormal compared to the 3 properties in the same zip area with the same tax class and BORO code. Due to this, this property record requires further investigation.

Record #3: 1065870
The street name for this property is Hylan Boulevard and located in zip area 10309. The total market value, assessed land value, and assessed total value of this property are $290 million, $17 million, and $17 million respectively. These values are not reasonable for a property with a building frontage of 39.5 feet and a building depth of 82.67 feet. Compared with properties with similar characteristics in this neighborhood, this property has significantly higher valuations. While the property may have unusual values, it appears to be public property and may not warrant investigation. Correction of these values may be necessary.

Record #4: 1059883
The property was reported as a 5-story structure on Sagona Ct with an unknown owner on Staten Island in New York. However, there is no 5-story structure on Sagona Ct, which consists of only 2 story structures. Also, the lot frontage and lot depth are both values of 5 feet, which is extremely unrealistic given the sizes of all properties on the street. Moreover, this property has extremely high values compared with 2 similar properties in the neighborhood. Given a large number of errors, this record needs to be investigated.

Record #5: 151044
The valuation of this property is $1.6 billion. However, the value of building front and building depth, 62 and 86 feet respectively, seems suspiciously small for a property with such a high value. Comparing 3 properties with similar building sizes in the neighborhood, the property's full value,

assessed land value, and assessed total value are much higher. Nevertheless, the property denotes Yankee Stadium, which would explain the structure's unusual valuation given its unusual use, and likely does not require further investigation.

Record #6: 252834
This property is located at 4510E DOUGLAS AVENUE, a 4-story structure with an unknown owner in the Bronx, New York. It was ranked eleventh by the fraud algorithm and has unusual values for lot frontage and depth when compared to the entire dataset, both zip code groups, its tax class, and its borough. Also, the lot size is quite large and nowhere near the 6 feet by 6 feet indicated in the records. This error requires further investigation.

Record #7: 771150
This property has unusually low building frontage and depth variables with regards to its full value, assessed value, and assessed total value. Upon further investigation, it appears that the values for building depth and building frontage may accurately represent the structure, which appears to be an approximately 1800 sqft home and does not appear to be worth $29,100,000. However, this structure is a part of a conjoined housing community with a few other similar structures in the immediate surrounding area. The monetary values may represent the entire community's estimated value and the size values represent the actual size of this single unit alone. This requires further investigation.

Record #8: 330292
This is a 1 story structure in Brooklyn, New York, owned by the PRATT INSTITUTE. The values of 3 and 5 respectively for the building's frontage and depth are highly atypical for a building with this level of full value, assessed value, and assessed total value. Further investigation shows that this structure is a learning institution with multiple buildings that are clearly larger than the indicated 3 feet by 5 feet, with all visible buildings having more than 1 story. Given that this appears to be a private institution that is not public property, this record requires further investigation.

Record #9: 104411
This property represents a 34-story structure in New York City, New York, owned by AMTO REALTY, INC. While the building may have slightly low values for lot size, lot depth, and building volume, the building frontage and depth variables of 18 feet and 52 feet respectively are unusually low and inaccurate. The building appears to be closer to 60 feet by 50 feet, which is much larger than the current reported area. Given this large discrepancy, this record needs investigation.

Record #10: 109067
This structure has odd values for its building depth and building frontage variables given its full value, assessed total value, and assessed value. However, upon further investigation, this building's dimensions appear to be approximately accurate. The structure's size of 18 feet by 52 feet appears to represent the actual size well, with a small amount of unused land behind the structure. The structure is also located along the riverbed, which likely makes the property value higher than average. An investigation could be conducted into this record but appears unnecessary and is not recommended.

7. Summary and Conclusions

If New York taxes only consider property values alone, most of the records in this study are probably acceptable when the residents pay adequate or more tax amounts. Given that most of the sample records seem to have reasonable valuations for property values, the impact on taxes is likely minimal. New York may even be benefitting from residents paying too much in property taxes. However, if it is not the case that New York taxes only on property valuations alone, more actions could be taken to reduce tax fraud.

Some records might have data errors: a multi-story building located in 5 feet by 5 feet lot area, or 8 feet by 8 feet building located in a large lot over one acre. Gathering additional information from other sources could be done to examine if these unusual records are reasonable or acceptable. For example, a crosslink could be established between property sales records and the property database to ensure that property values are up to date. Through census surveys, the New York government could ask questions and assess property information to bring corrected information. Moreover, New York could send out surveys to property owners to collect data directly and compare it to existing records. If a discrepancy is found, property surveyors could be dispatched to investigate the property from public areas to ensure that the information reported is correct.

Next, different variables can be created using some of the other variables in the dataset. Variables such as exemptions counts, exemption classes, and building classes were unused. Variables grouped by these values could yield additional information for unusual records. Valuations based on city block and boro could also be helpful, especially with area and stories. A 1 story building right next to a 100 should raise questions.

Altogether, this investigation has yielded promising results for enabling quick investigation of these records, further improvements could be made to fine-tune results and to assess other vital elements of these records. Regarding the analyzed records, no firm accusations were made regarding potential fraud. However, there are several oddities regarding building and lot sizes and building valuations. Analyzed records numbers #1, #2, #4, #6, #7, #8, and #9 require further investigation. Record #3 requires corrections. Records #5 and #10 do not require further investigation.

Appendix I. Data Quality Report

Appendix I-1. Data Set Description

- Name: Property Valuation and Assessment Data
- Purpose: NYC properties assessments to calculate property tax, grant eligible properties, exemptions, and/ or abatements
- Source: NYC Open Data
- Time period: November 17th, 2010
- Number of fields: 32
- Number of records: 1,070,994

Appendix I-2. Data Set Summary

- Numeric Fields

Item	Column name	# of records	% populated	Unique values	Mean	Standard deviation	Minimum value	Maximum value	# of zeros
1	LTFRONT	1,070,994	100.00%	1,297	36.64	74.03	0	9,999	169,108
2	LTDEPTH	1,070,994	100.00%	1,370	88.86	76.40	0	9,999	170,128
3	STORIES	1,014,730	94.75%	111	5.01	8.37	1	119	0
4	FULLVAL	1,070,994	100.00%	109,324	874,264.51	11,582,430.99	0	6,150,000,000	13,007
5	AVLAND	1,070,994	100.00%	70,921	85,067.92	4,057,260.06	0	2,668,500,000	13,009
6	AVTOT	1,070,994	100.00%	112,914	227,238.17	6,877,529.31	0	4,668,308,947	13,007
7	EXLAND	1,070,994	100.00%	33,419	36,423.89	3,981,575.79	0	2,668,500,000	491,699
8	EXTOT	1,070,994	100.00%	64,255	91,186.98	6,508,402.82	0	4,668,308,947	432,572
9	BLDFRONT	1,070,994	100.00%	612	23.04	35.58	0	7,575	228,815
10	BLDDEPTH	1,070,994	100.00%	621	39.92	42.71	0	9,393	228,853
11	AVLAND2	282,726	26.40%	58,591	246,235.72	6,178,962.56	3	2,371,005,000	0
12	AVTOT2	282,732	26.40%	111,360	713,911.44	11,652,528.95	3	4,501,180,002	0
13	EXLAND2	87,449	8.17%	22,195	351,235.68	10,802,212.67	1	2,371,005,000	0
14	EXTOT2	130,828	12.22%	48,348	656,768.28	16,072,510.17	7	4,501,180,002	0

- Categorical Fields

Item	Column name	# of records	% populated	Unique values	Most common field value
1	RECORD	1,070,994	100.00%	1,070,994	N/A
2	BBLE	1,070,994	100.00%	1,070,994	N/A
3	B	1,070,994	100.00%	5	4
4	BLOCK	1,070,994	100.00%	13,984	3944
5	LOT	1,070,994	100.00%	6.366	1
6	EASEMENT	4,636	0.43%	12	E
7	OWNER	1,039,249	97.04%	863,347	PARKCHESTER PRESERVAT
8	BLDGCL	1,070,994	100.00%	200	R4
9	TAXCLASS	1,070,994	100.00%	11	1
10	EXT	354,305	33.08%	3	G

Item	Column name	# of records	% populated	Unique values	Most common field value
11	EXCD1	638,488	59.62%	129	1017
11	STADDR	1,070,318	99.94%	839,280	501 SURF AVENUE
13	ZIP	1,041,104	97.21%	196	10314
14	EXMPTCL	15,579	1.45%	14	X1
15	EXCD2	92,948	8.68%	60	1017
16	PERIOD	1,070,994	100.00%	1	FINAL
17	YEAR	1,070,994	100.00%	1	2010/11
18	VALTYPE	1,070,994	100.00%	1	AC-TR

Appendix I-3. Data Field Exploration

Field 1.

- Name: RECORD
- Description: unique identifier of each entry in the data

Field 2.

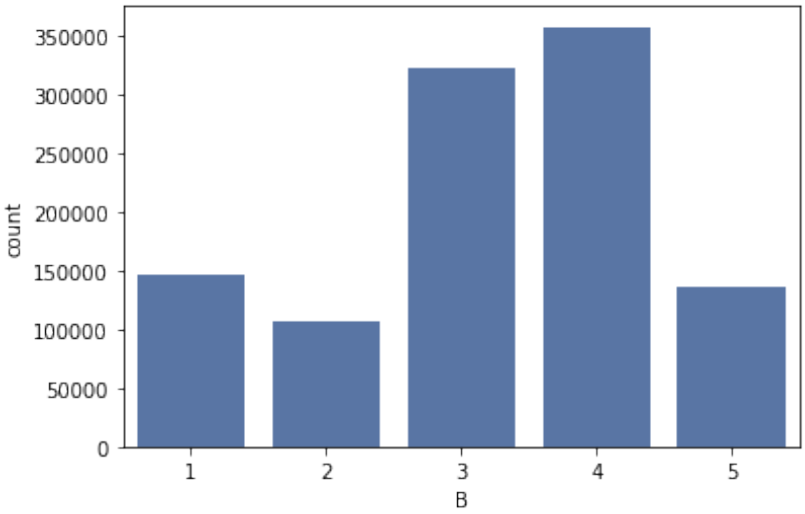
- Name: BBLE
- Description: concatenation of borough code (B), block code (BLOCK), lot (LOT), and easement (EASEMENT)

Field 3.

- Name: B
- Description: Borough Code

Borough Code	Representation
1	Manhattan
2	Bronx
3	Brooklyn
4	Queens
5	Staten Island

- Count plot:

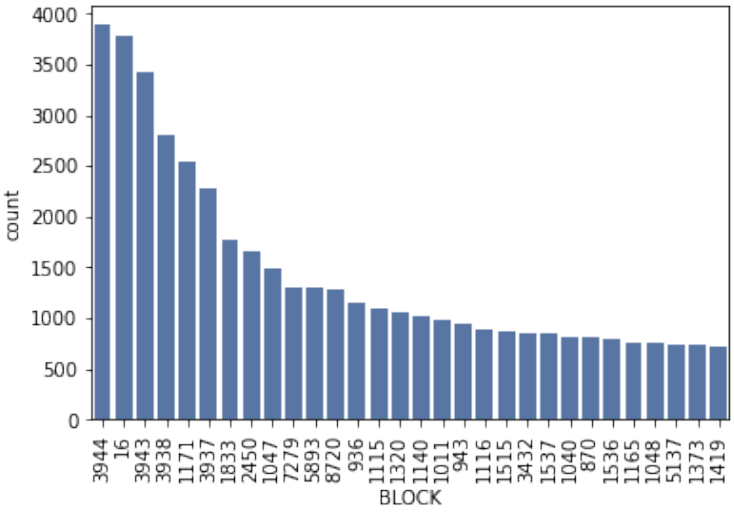


Field 4.

- Name: BLOCK
- Description: valid block ranges by borough

Borough Code	Representation	Block range
1	Manhattan	1 to 2,255
2	Bronx	2,260 to 5,958
3	Brooklyn	1 to 8,955
4	Queens	1 to 16,350
5	Staten Island	1 to 8,050

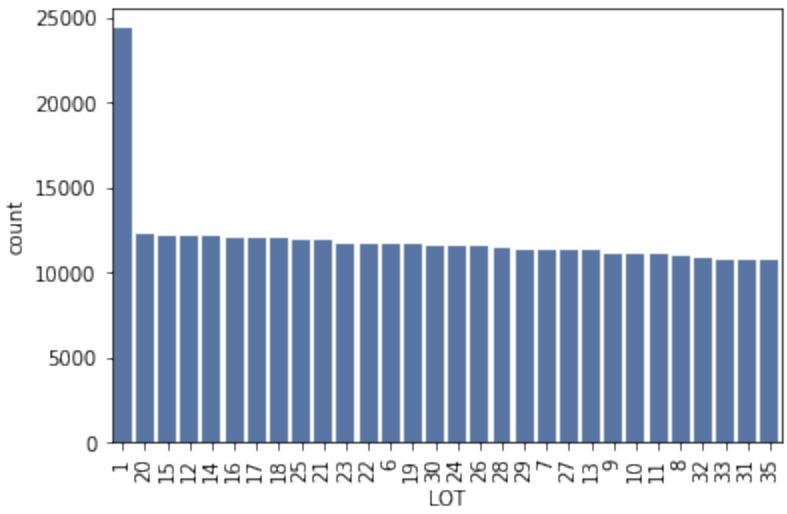
- Count plot: (top 30 categories)



Field 5.

- Name: LOT

- Description: unique number within a borough/ block
- Count plot: (top 30 categories)

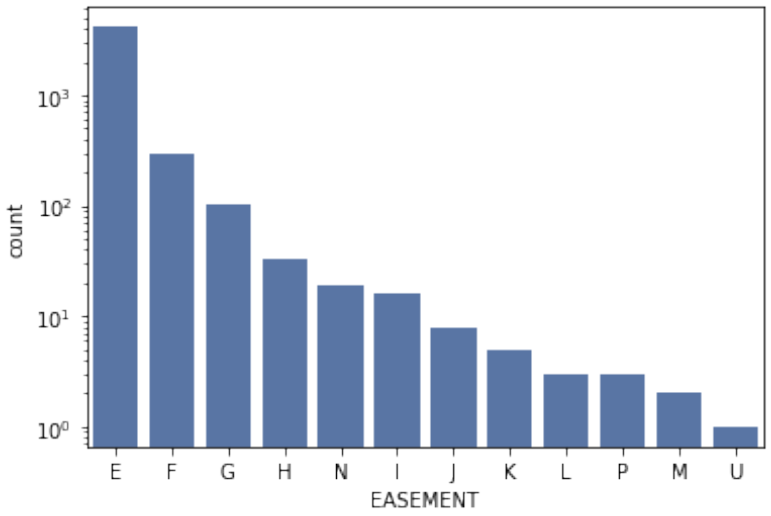


Field 6.

- Name: EASEMENT
- Description: a field used to describe easement

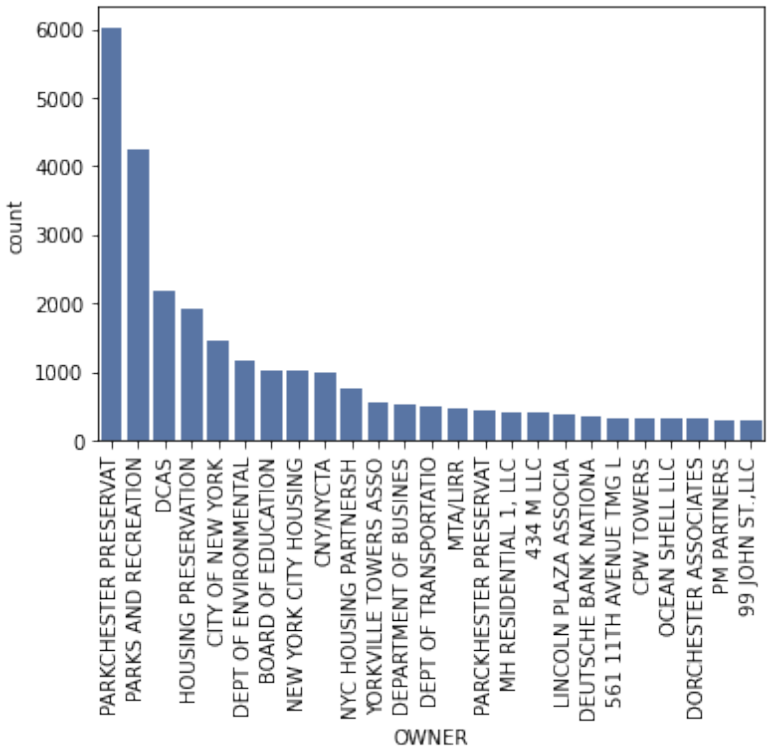
EASEMENT value	Indication
Space	Indicates the lot has no Easement
A	Indicates the portion of the Lot that has an Air Easement
B	Indicates Non-Air Rights.
E	Indicates the portion of the lot that has a Land Easement
F thru M	Are duplicates of 'E'
N	Indicates Non-Transit Easement
P	Indicates Piers
R	Indicates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

- Count plot:



Field 7.

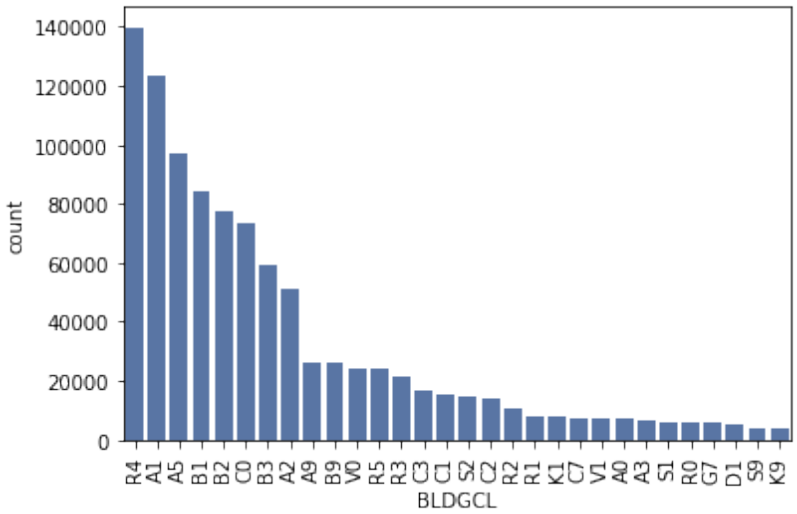
- Name: OWNER
- Description: owner's name
- Count plot: (top 20 categories)



Field 8.

- Name: BLDGCL

- Description: building class
- Count plot: (top 30 categories)

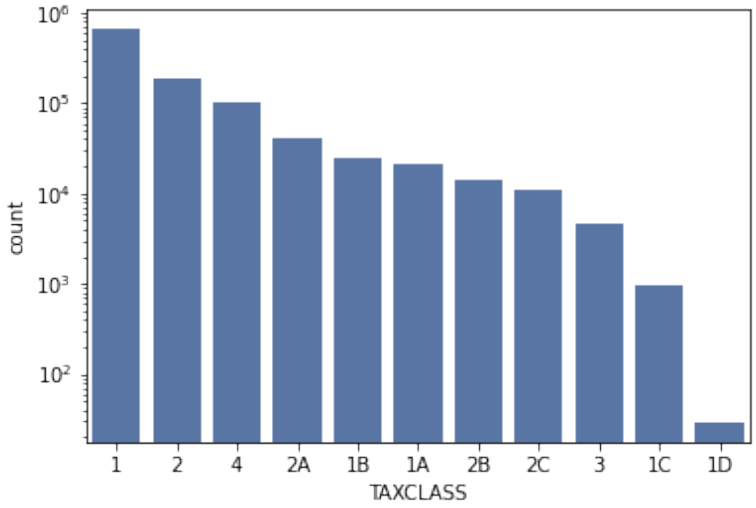


Field 9.

- Name: TAXCLASS
- Description: current property tax class code (NYS Classification)

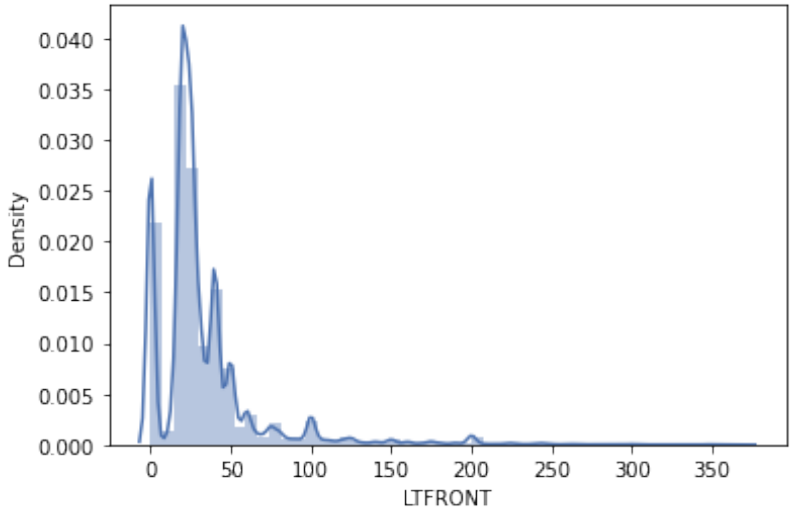
TAXCLASS value	Representation
1	1-3 unit residences
1A	1-3 story condominiums
1B	Residential vacant land
1C	1-3 unit condominiums
1D	Select bungalow colonies
2	Apartments
2A	Apartments with 4-6 units
2B	Apartments with 7-10 units
2C	Coops/condos with 2-10 units
3	Utilities (except ceiling railroads)
4A	Utilities - ceiling railroads
4	All others

- Count plot:



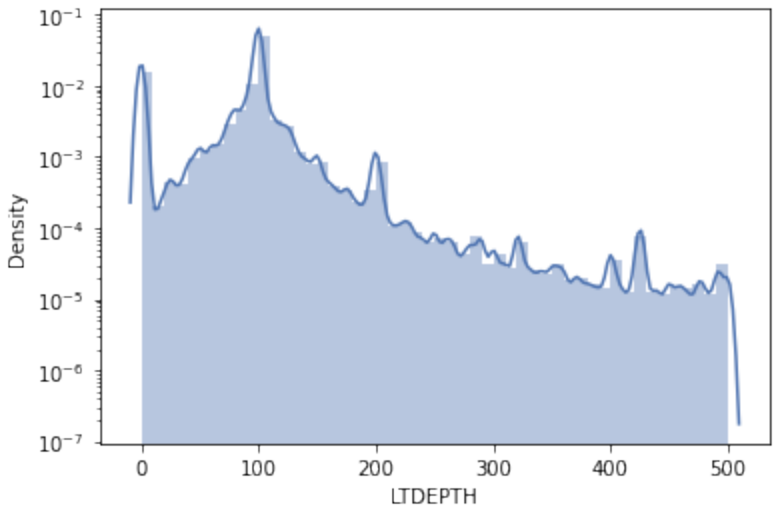
Field 10.

- Name: LTFRONT
- Description: lot frontage in feet
- Histogram: (data in the histogram excludes outliers > 370 and is 99.45% populated)



Field 11.

- Name: LTDEPTH
- Description: lot depth in feet
- Histogram: (data in the histogram excludes outliers > 500 and is 99.67% populated)

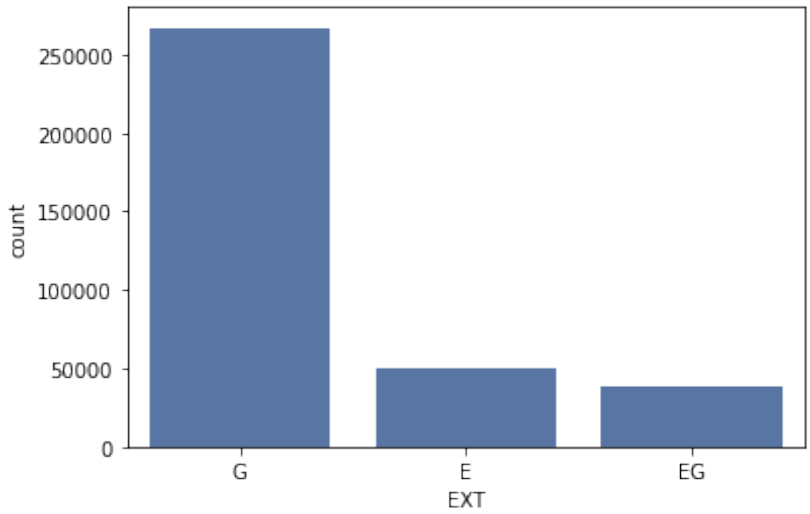


Field 12.

- Name: EXT
- Description: extension

EXT value	Representation
E	Extension
G	Garage
EG	Extension and Garage

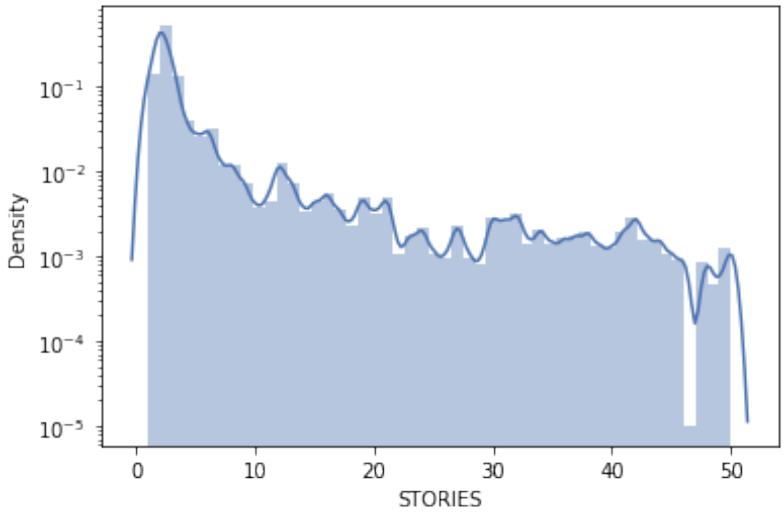
- Count plot:



Field 13.

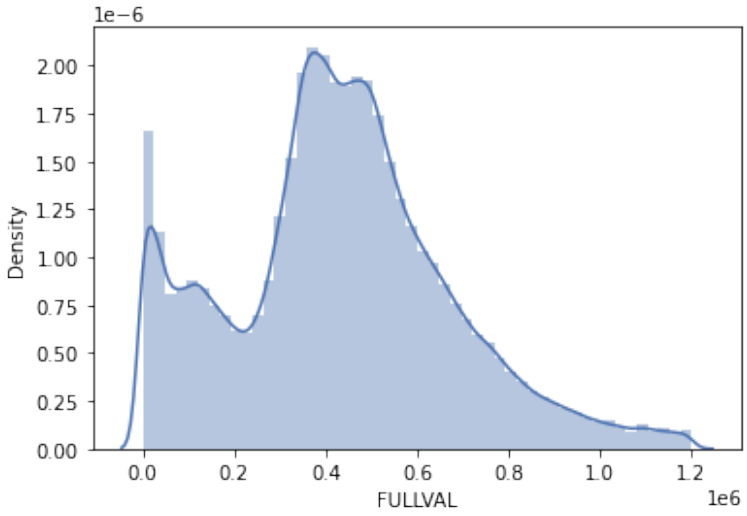
- Name: STORIES
- Description: the number of stories for the building (# of floors)

- Histogram: (data in the histogram excludes outliers > 50 and is 99.5% populated)



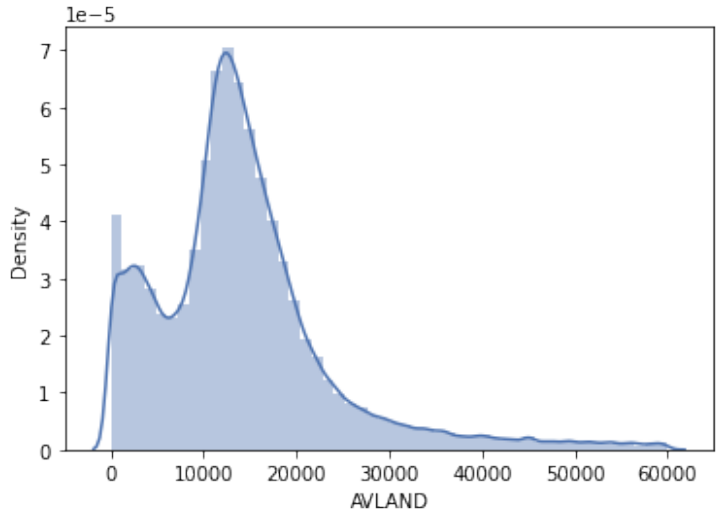
Field 14.

- Name: FULLVAL
- Description: total market value of the property
- Histogram: (data in the histogram excludes outliers > 1,200,000 and is 93.28% populated)



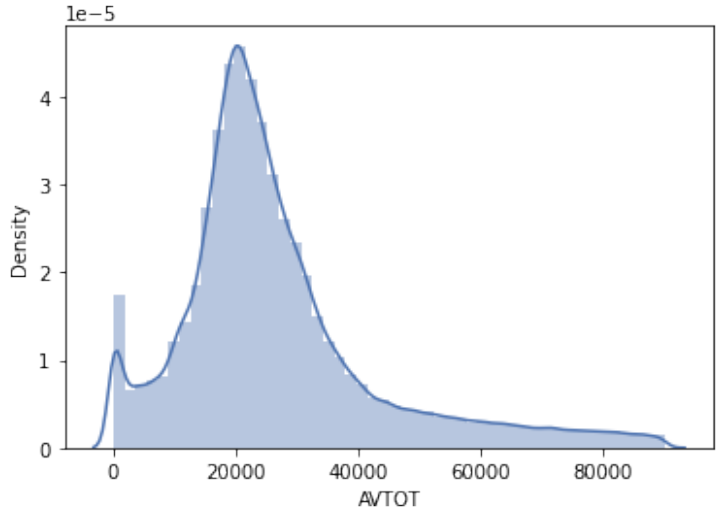
Field 15.

- Name: AVLAND
- Description: assessed land value
- Histogram: (data in the histogram excludes outliers > 60,000 and is 91.62% populated)



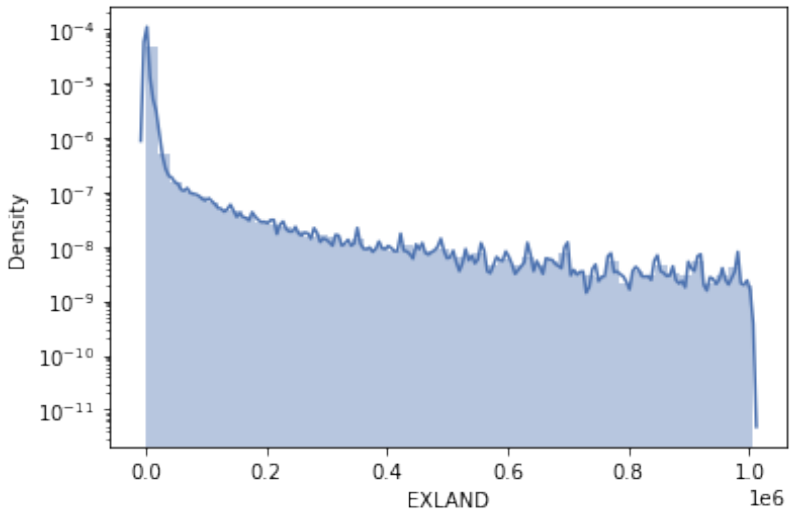
Field 16.

- Name: AVTOT
- Description: assessed total value
- Histogram: (data in the histogram excludes outliers > 90,000 and is 84.99% populated)



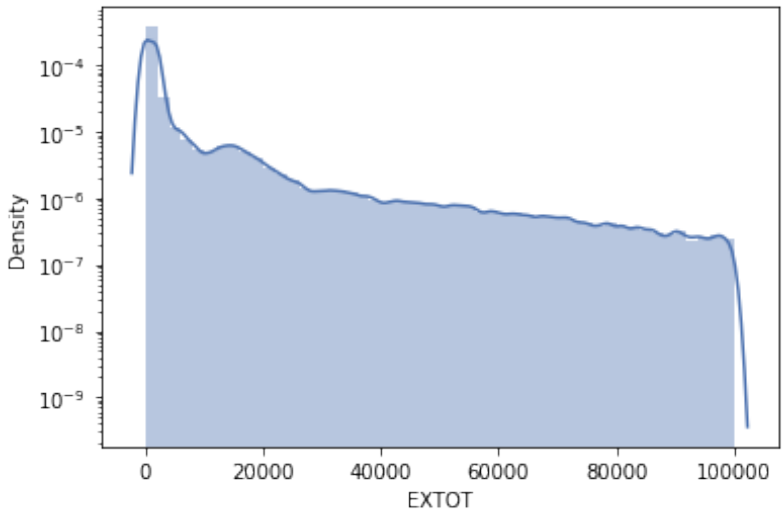
Field 17.

- Name: EXLAND
- Description: exempt land value
- Histogram: (data in the histogram excludes outliers > 1,005,000 and is 99.63% populated)



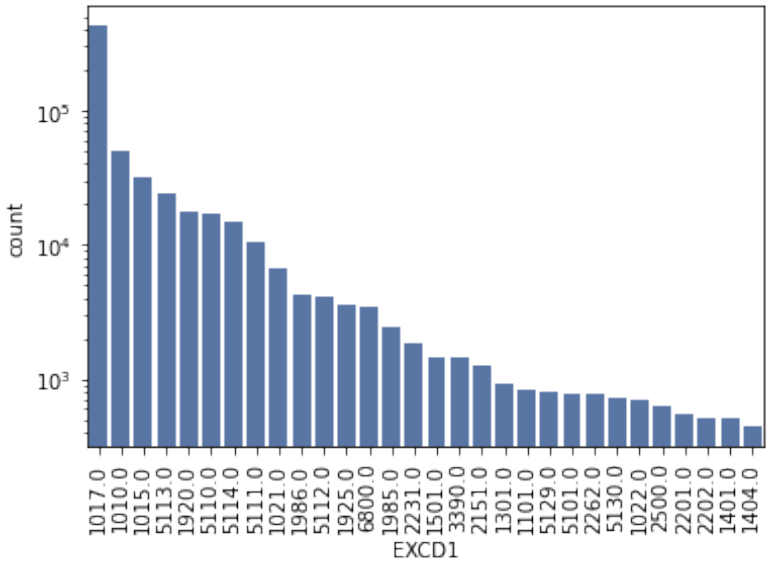
Field 18.

- Name: EXTOT
- Description: exempt total value
- Histogram: (data in the histogram excludes outliers > 100,000 and is 96.45% populated)



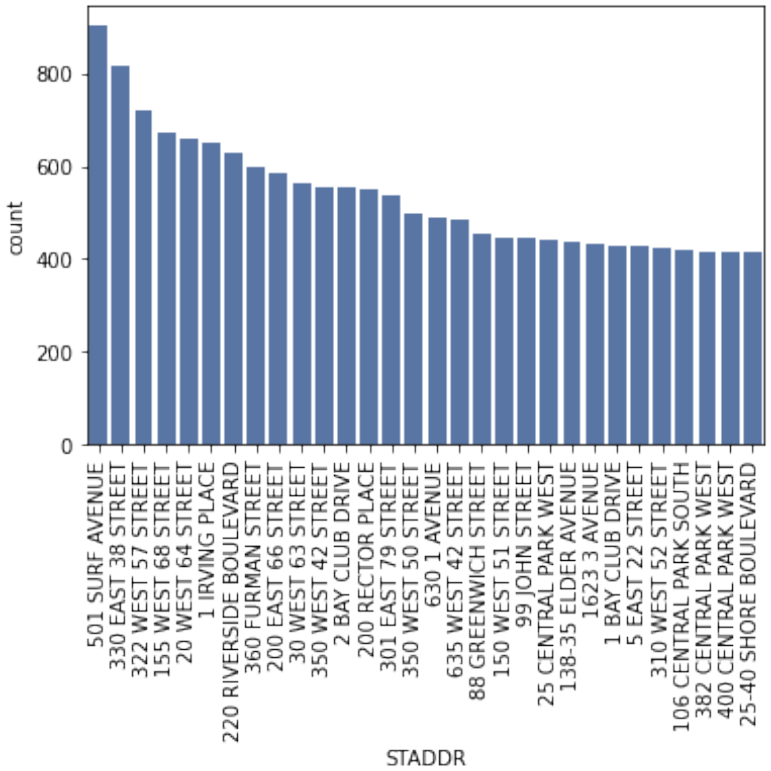
Field 19.

- Name: EXCD1
- Description: N/A
- Count plot: (top 30 categories)



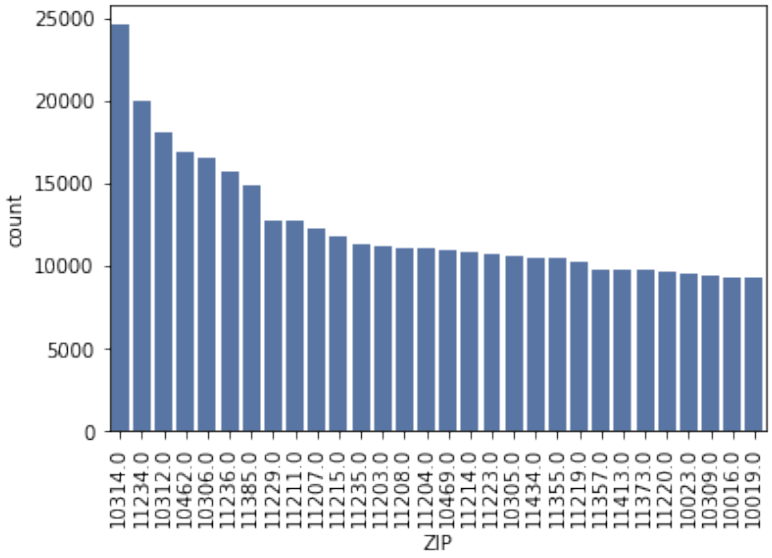
Field 20.

- Name: STADDR
- Description: street name for the property
- Count plot: (top 30 categories)



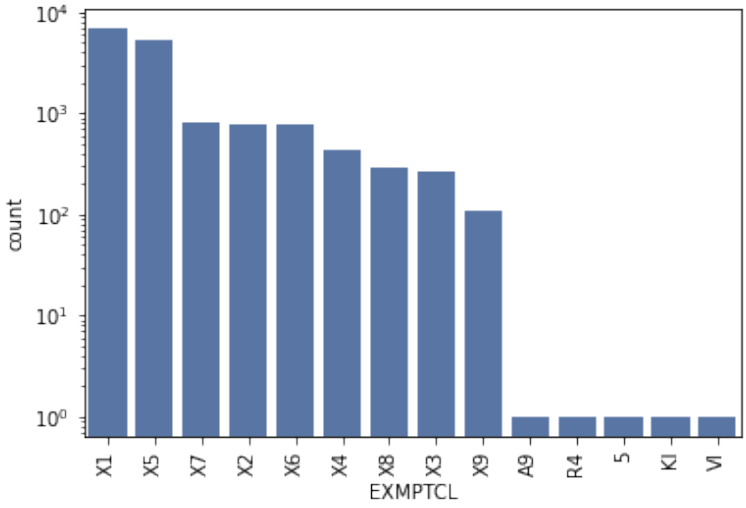
Field 21.

- Name: ZIP
- Description: postal zip code of the property
- Count plot: (top 30 categories)



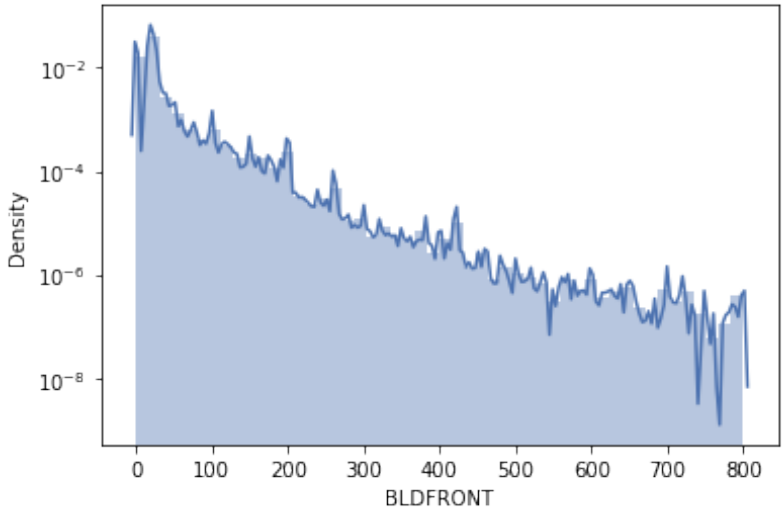
Field 22.

- Name: EXMPTCL
- Description: exempt class used for fully exempt properties only
- Count plot:



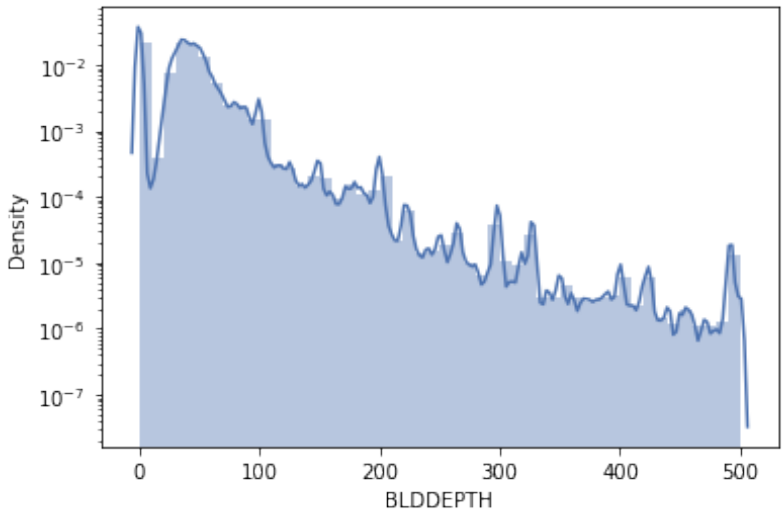
Field 23.

- Name: BLDFRONT
- Description: building frontage in feet
- Histogram: (data in the histogram excludes outliers > 800 and is 99.996% populated)



Field 24.

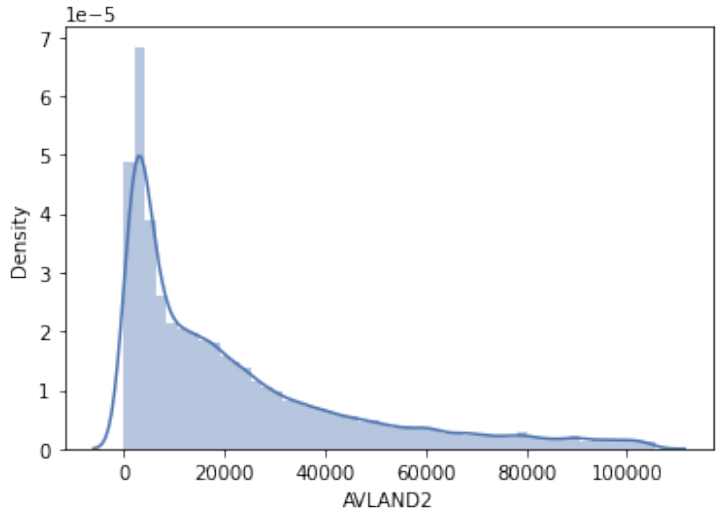
- Name: BLDDEPTH
- Description: lot depth in feet
- Histogram: (data in the histogram excludes outliers > 500 and is 99.92% populated)



Field 25.

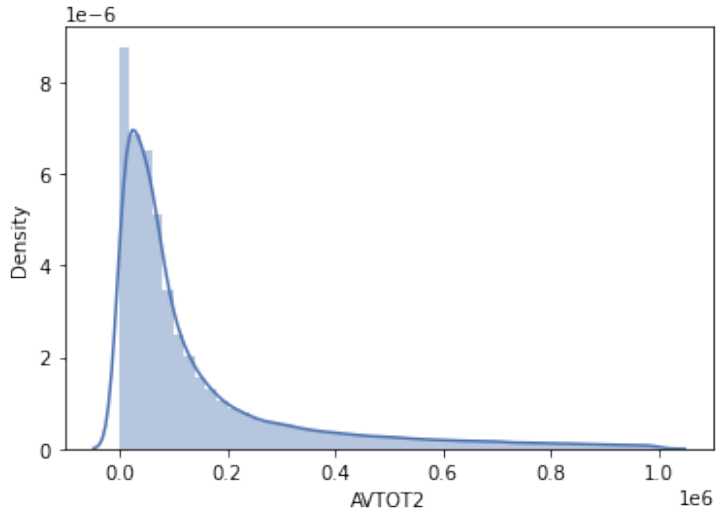
- Name: AVLAND2
- Description: new market value of the land

- Histogram: (data in the histogram excludes outliers > 105,500 and is 81.93% populated)



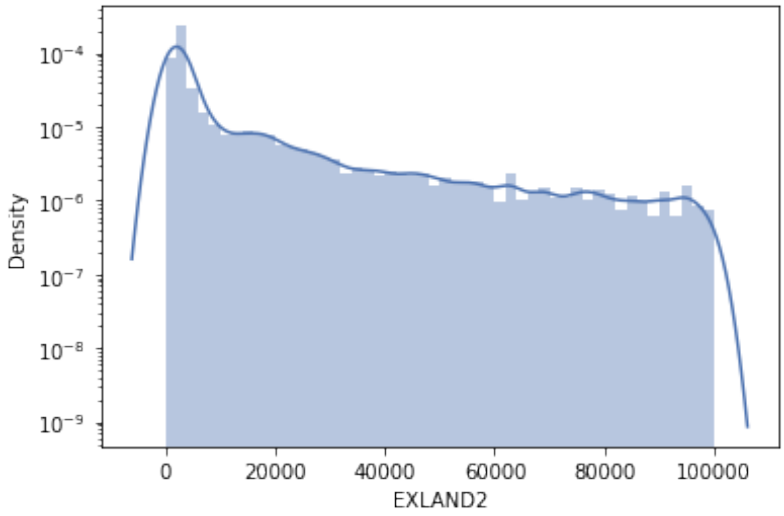
Field 26.

- Name: AVTOT2
- Description: new total market value
- Histogram: (data in the histogram excludes outliers > 1,000,000 and is 91.62% populated)



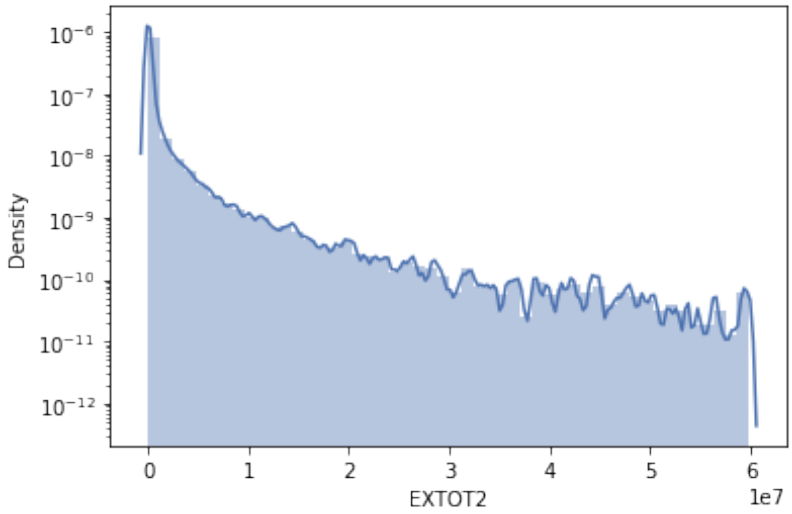
Field 27.

- Name: EXLAND2
- Description: new exempt land value
- Histogram: (data in the histogram excludes outliers > 100,000 and is 84.01% populated)



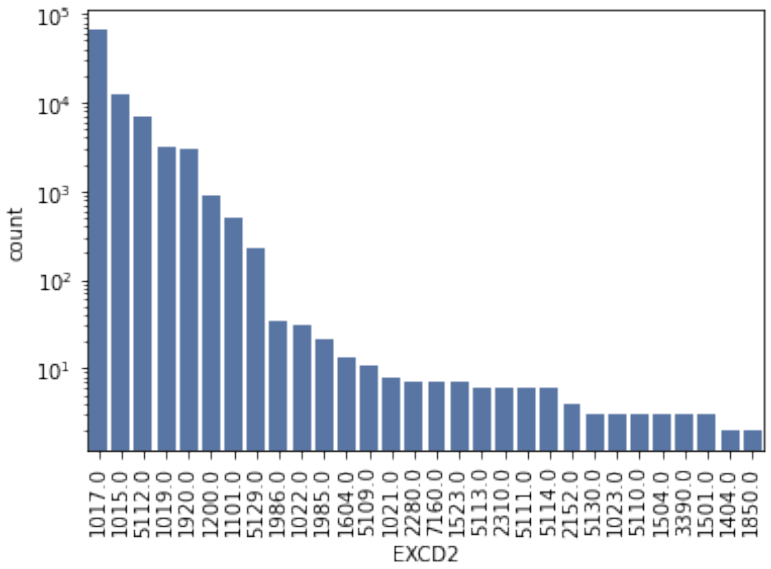
Field 28.

- Name: EXTOT2
- Description: new exempt total value
- Histogram: (data in the histogram excludes outliers > 60,000,000 and is 99.90% populated)



Field 29.

- Name: EXCD2
- Description: N/A
- Count plot: (top 30 categories)



Field 30.

- Name: PERIOD
- Description: indicator for the change period of the file (all records have the same value)

Field 31.

- Name: YEAR
- Description: year and month when the data was most recently updated (all records have the same value)

Field 32.

- Name: VALTYPE
- Description: the parcel's values are reflected in another lot (all records have the same value)

Appendix II. Variable Statistics

Item	Variable	Count	Mean	Std.	Min	25%	50%	75%	Max
1	r1	1046826	213.57	532.56	0.00039058	76.26	152.86	243.18	138637.34
2	r2	1046826	536.42	1027.44	0.00074662	211.82	500	683.82	310551.57
3	r3	1046826	244.08	498.82	8.98E-05	82.5	239.06	336.23	227500
4	r4	1046826	10.46	69.98	8.33E-06	2.35	4.6	7.23	22249.8
5	r5	1046826	22.82	378.51	0.00020961	6.2	14.98	20.29	334636.37
6	r6	1046826	10.04	178.33	1.04E-05	2.02	7.15	10.04	111545.46
7	r7	1046826	26.83	203.87	8.33E-06	5.28	8.58	13.53	64711.85
8	r8	1046826	50.46	932.2	0.00030625	17.47	26.98	36	871362.22
9	r9	1046826	19.07	327.02	6.05E-05	7.09	12.68	17.58	290454.07
10	r1_zip5	1046826	1	2.16	2.23E-06	0.46	0.92	1.23	747.76
11	r2_zip5	1046826	1	2.13	1.31E-06	0.44	0.96	1.26	658.74
12	r3_zip5	1046826	1	2.73	3.94E-07	0.39	0.93	1.24	956.75
13	r4_zip5	1046826	1	5.13	1.66E-06	0.28	0.75	1.1	2456.16
14	r5_zip5	1046826	1	8.17	4.11E-06	0.34	0.8	1.08	5004.51
15	r6_zip5	1046826	1	10.01	9.21E-07	0.25	0.73	1.02	4200.7
16	r7_zip5	1046826	1	5.81	8.23E-07	0.29	0.63	0.95	3237.54
17	r8_zip5	1046826	1	7.37	7.10E-06	0.44	0.74	1	5527.75
18	r9_zip5	1046826	1	7.91	2.23E-06	0.4	0.71	0.96	4864.74
19	r1_zip3	1046826	1	2.4	2.05E-06	0.43	0.86	1.23	727.14
20	r2_zip3	1046826	1	2.12	1.25E-06	0.4	0.95	1.27	756.25
21	r3_zip3	1046826	1	2.65	2.94E-07	0.35	0.93	1.27	912.3
22	r4_zip3	1046826	1	6.39	9.34E-07	0.28	0.69	1.06	2051.04
23	r5_zip3	1046826	1	17.12	4.26E-06	0.3	0.77	1.05	15414.36
24	r6_zip3	1046826	1	16.14	9.57E-07	0.22	0.7	1	9790.65
25	r7_zip3	1046826	1	7.31	3.88E-07	0.26	0.54	0.83	2407.6
26	r8_zip3	1046826	1	22.53	7.93E-06	0.42	0.7	0.93	21881.14
27	r9_zip3	1046826	1	16.31	2.68E-06	0.39	0.68	0.93	14530.83
28	r1_taxclass	1046826	1	2.93	1.49E-06	0.37	0.68	1.06	1119.59
29	r2_taxclass	1046826	1	3.74	1.89E-06	0.49	0.78	1.06	1525.68
30	r3_taxclass	1046826	1	3.48	9.20E-07	0.49	0.81	1.11	923.91
31	r4_taxclass	1046826	1	3.32	4.64E-06	0.32	0.71	1.1	1438.08
32	r5_taxclass	1046826	1	6.44	9.50E-06	0.33	0.82	1.12	4260.03
33	r6_taxclass	1046826	1	6.31	7.96E-07	0.3	0.82	1.14	3169.93
34	r7_taxclass	1046826	1	2.95	1.71E-06	0.38	0.73	1.12	1286.4
35	r8_taxclass	1046826	1	6.13	1.00E-05	0.45	0.84	1.12	4680.65
36	r9_taxclass	1046826	1	5.63	4.90E-06	0.45	0.84	1.15	4204.08

Item	Variable	Count	Mean	Std.	Min	25%	50%	75%	Max
37	r1_boro	1046826	1	2.37	2.32E-06	0.42	0.84	1.24	727.14
38	r2_boro	1046826	1	2	1.32E-06	0.4	0.95	1.27	756.58
39	r3_boro	1046826	1	2.54	3.13E-07	0.35	0.92	1.28	932.87
40	r4_boro	1046826	1	6.31	9.34E-07	0.27	0.67	1.06	2054.62
41	r5_boro	1046826	1	17.15	4.41E-06	0.3	0.77	1.05	15466.54
42	r6_boro	1046826	1	16.04	9.94E-07	0.22	0.7	1	9483.22
43	r7_boro	1046826	1	7.28	3.88E-07	0.26	0.51	0.81	2407.6
44	r8_boro	1046826	1	21.93	8.19E-06	0.41	0.7	0.93	21241.27
45	r9_boro	1046826	1	15.73	2.78E-06	0.39	0.67	0.94	13819.33