Project 1

New York Property Data Fraud Report

DSO 562
Fraud Analytics

Group 6
Daniel Xu
Andrew Zhao
Kefan Luke Lu
Tao Xiong
Yuxing Scott Cai
Xinyi Alex Guo

February 12, 2020

Table of Contents

Executive Summary

This report evaluates The New York City Property Valuation and Assessment Data for the purpose of fraud detection utilizing unsupervised machine learning methods. The software used for analysis is Python, and the methodologies implemented include Principal Component Analysis (PCA) and Autoencoder.

The original data set comprises 1,070,994 records of properties across the city of New York with detailed information about their locations, estimated values, lot and building sizes, owners, number of stories, building and tax classes etc. The general analysis process includes preliminary data cleaning, creating expert variables, normalizing and dimensionality reduction, applying fraud algorithms, calculating and combining fraud scores, and identifying potential anomalies.

By implementing Heuristic Algorithm and Autoencoder, a combined fraud score is produced for each of the one million properties. Properties are rank-ordered by this score and those with high scores are deemed potentially anomalous. The report thus further investigates the top 10 records with the highest scores.

Thorough inspection on these anomalous properties shows that the most suspicious records tend to have conspicuously higher values in a number of fields with comparison to the majorities. In the meantime, our examination indicates that some of these anomalies could be accounted for by either historical or governmental, ‘bona-fide’ reasons. Whereas future investigations are suggested for identified properties purchased by individual buyers and those with unclear or missing information.

Data Overview

The City of New York Property Valuation and Assessment Dataset, publicly posted by the Department of Finance on the City of New York Open Data Website, represents property valuations and assessments conducted in the city as of November 2010 for the purpose of calculating property tax and granting eligible properties exemptions and/or abatements. The data was collected and recorded into the system primarily by various City employees such as property assessors, property exemption specialists, ACRIS reporting etc.

The original data set consists of 1,070,994 records of properties across the city of New York with 32 distinct fields containing detailed information about their locations, estimated values, lot and building sizes, owners, number of stories, building and tax classes etc. Among the 32 columns, there are 14 numerical fields, 17 categorical fields and 1 time-stamp field.

Below are summary tables and descriptions of the variables considered to be most pertinent to our analysis, excerpted from the data quality report which can be found in the Appendix.

Summary Statistics of Key Variables:

Field Name	Field Type	# Records w/ Value	% Populated	# Unique	# Zeros	Mean	St. Dev.	Min	Max
LTFRONT	Numeric	1070994	100.00%	1297	169108	36.6353	74.03283872	0	9999
LTDEPTH	Numeric	1070994	100.00%	1370	170128	88.86159	76.39628129	0	9999
STORIES	Numeric	1014730	94.75%	112	0	5.006918	8.365707394	1	119
FULLVAL	Numeric	1070994	100.00%	109324	13007	874264.5	11582430.99	0	6.15E+09
AVLAND	Numeric	1070994	100.00%	70921	13009	85067.92	4057260.056	0	2.67E+09
AVTOT	Numeric	1070994	100.00%	112914	13007	227238.2	6877529.306	0	4.67E+09
BLDFRONT	Numeric	1070994	100.00%	612	228815	23.04277	35.579696	0	7575
BLDDEPTH	Numeric	1070994	100.00%	621	228853	39.92284	42.70715468	0	9393

Summary Statistics of Key Numerical Fields

Field Name	Field Type	# Records w/ Value	% Populated	# Unique	Most Common
RECORD	Categorical	1070994	1	1070994	N/A
BBLE	Categorical	1070994	1	1070994	N/A
B	Categorical	1070994	1	5	4
BLOCK	Categorical	1070994	1	13984	3944
BLDGCL	Categorical	1070994	1	200	R4
TAXCLASS	Categorical	1070994	1	11	1
ZIP	Categorical	1041104	0.9721	197	10314

Summary Statistics of Key Categorical Fields

Field Name: **RECORD**

RECORD is an ordinal categorical variable for referencing a property in the data set. It has 1,070,994 unique values, ranging from 1 to 1,070,994. No duplicates nor missing values were found in the field.

Field Name: **BBLE**

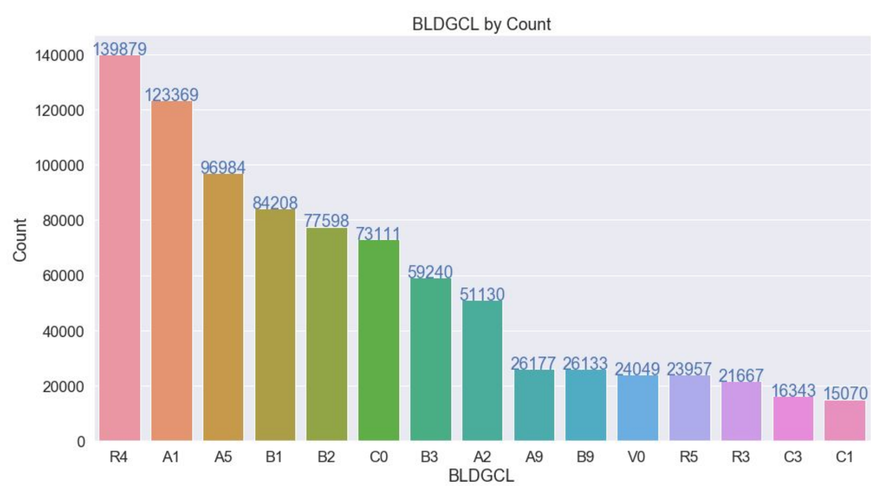
BBLE is a categorical variable representing the concatenation of Boro, Block, Lot, and Easement code. The Length is 10 or 11 alphanumeric. No duplicate nor missing combinations were found in the field.

Field Name: **B**

B is a categorical variable which takes 5 unique values, each standing for a borough to which a property belongs. Specifically, 1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island. No missing values were found in the field.

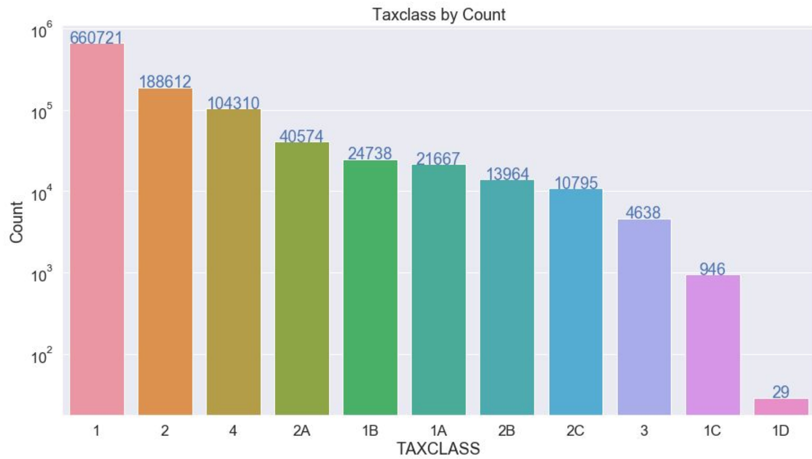
Field Name: **BLDGCL**

BLDGCL is an alphanumeric categorical variable with 200 unique levels indicating the building class of a property. Each level contains 2 digits – the first digit is a character from A to Z, the second digit is a number from 0 to 9. No missing values were found in the field. The top 15 BLDGCL is as follows:



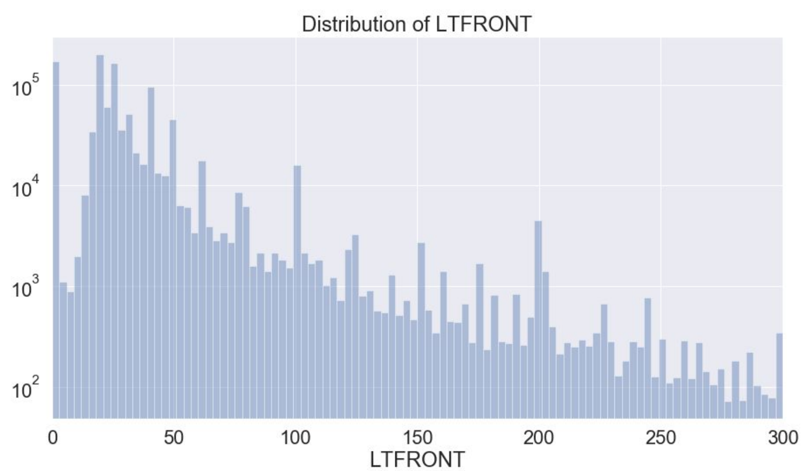
Field Name: **TAXCLASS**

TAXCLASS is an alphanumeric categorical variable with 11 unique levels indicating the tax class of a property - “1”, “1A”, “1B”, “1C”, “1D”, “2”, “2A”, “2B”, “2C”, “3”, and “4”. No missing values were found in the field. The rank ordered TAXCLASS is as follows:



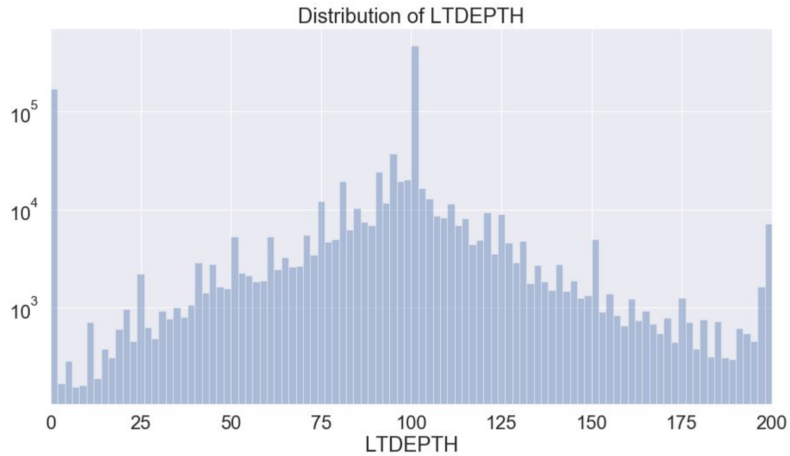
Field Name: **LTFRONT**

LTFRONT is a numeric variable measuring the lot frontage in feet, with 1,297 unique values ranging from 0 to 9999. There are 169,108 zero records, possibly indicating missing values. The LTFRONT distribution is as follows:



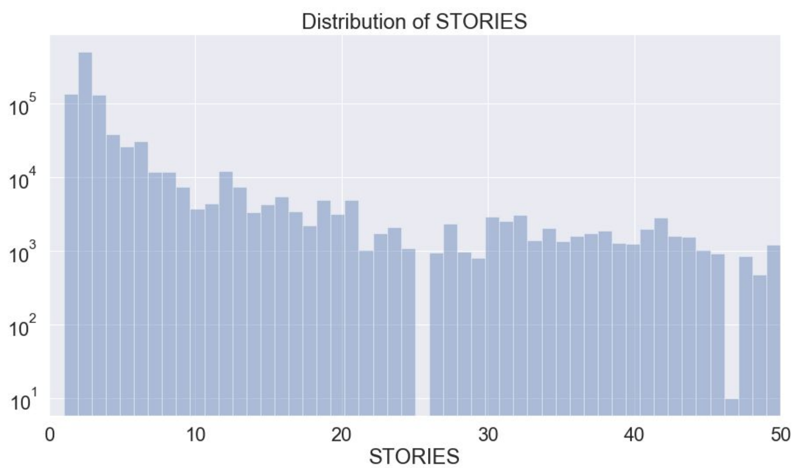
Field Name: **LTDEPTH**

LTDEPTH is a numeric variable measuring the lot depth in feet with 1,370 unique values ranging from 0 to 9999. There are 170,128 records of zero records, possibly indicating missing values. The LTDEPTH distribution is as follows:



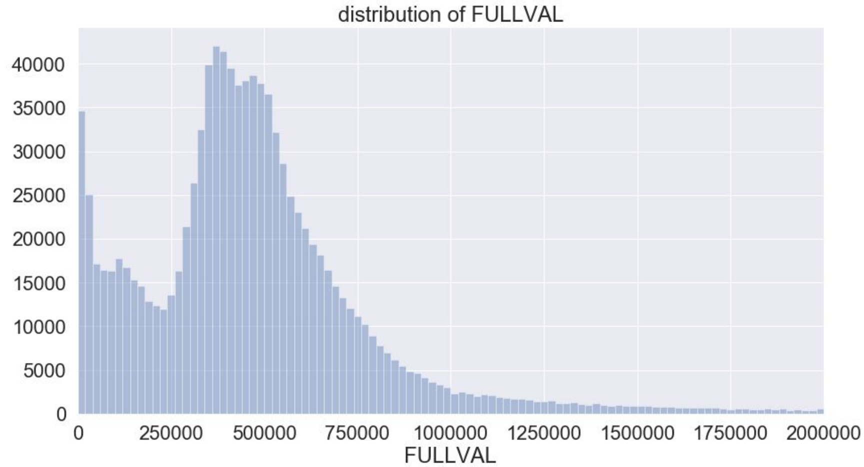
Field Name: **STORIES**

STORIES is a numeric variable measuring the number of stories of a property with 111 unique values ranging from 1 to 119. No missing values were found in the field. The STORIES distribution is as follows:



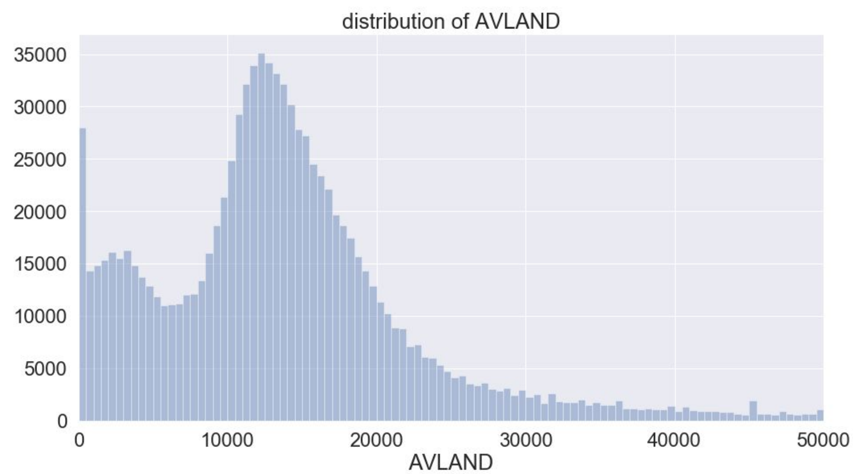
Field Name: **FULLVAL**

FULLVAL is a numeric variable representing the total market value of the property with 109,324 unique values ranging from 0 to 6,150,000,000. There are 13,007 records of zero records. The FULLVAL distribution is as follows:



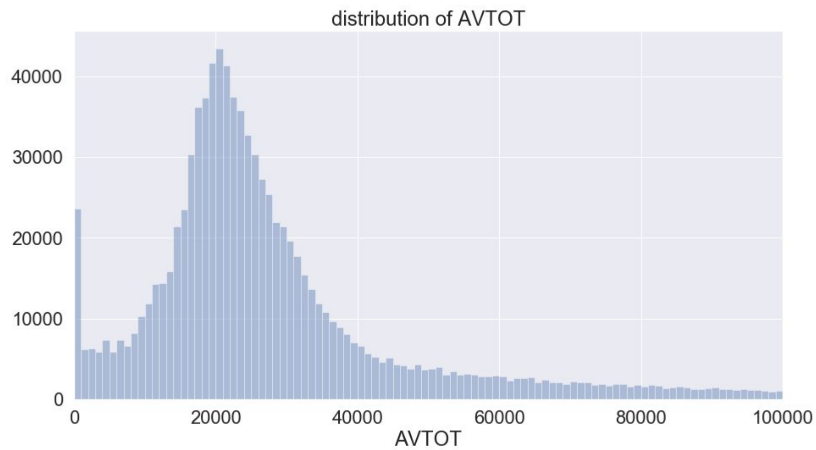
Field Name: **AVLAND**

AVLAND is a numeric variable representing the actual value of land with 70,921 unique values ranging from 0 to 2,668,500,000. There are 13,009 zero records found in the field. The AVLAND distribution is as follows:



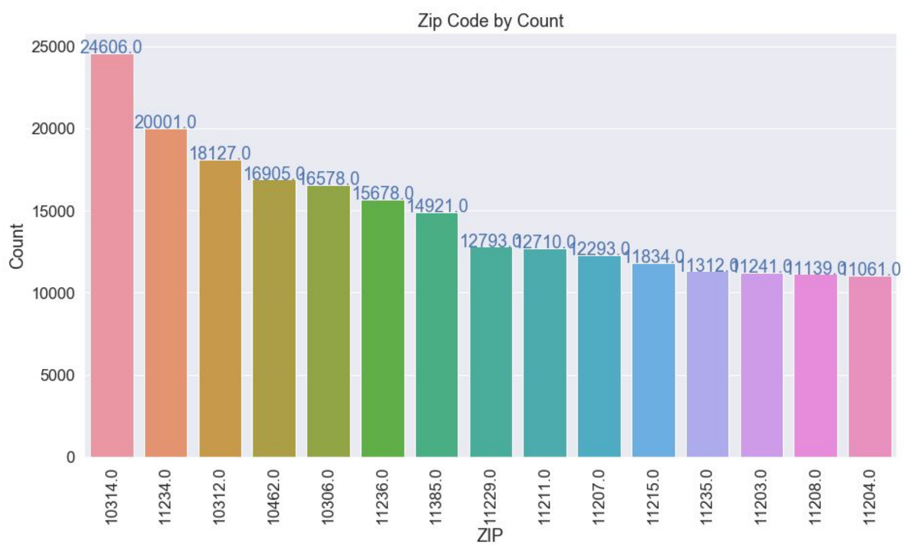
Field Name: **AVTOT**

AVTOT is a numeric variable representing the actual total value of a property with 112,914 unique values ranging from 0 to 4,668,308,947. There are 13007 zero records found in the field. The AVTOT distribution is as follows:



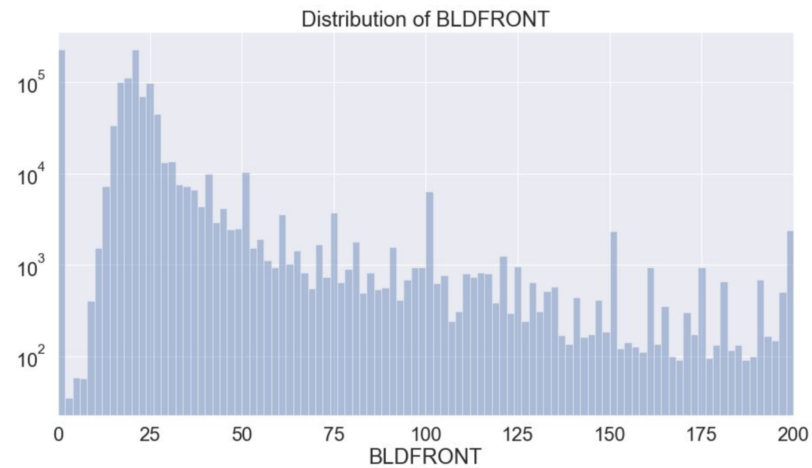
Field Name: **ZIP**

ZIP is a categorical variable representing the zip code in which a property is located with 196 unique values. There are 29,890 missing records found in the field. A LTFRONT of 0 may indicate missing value. The Top 15 Zipcodes by occurring frequency are as follows:



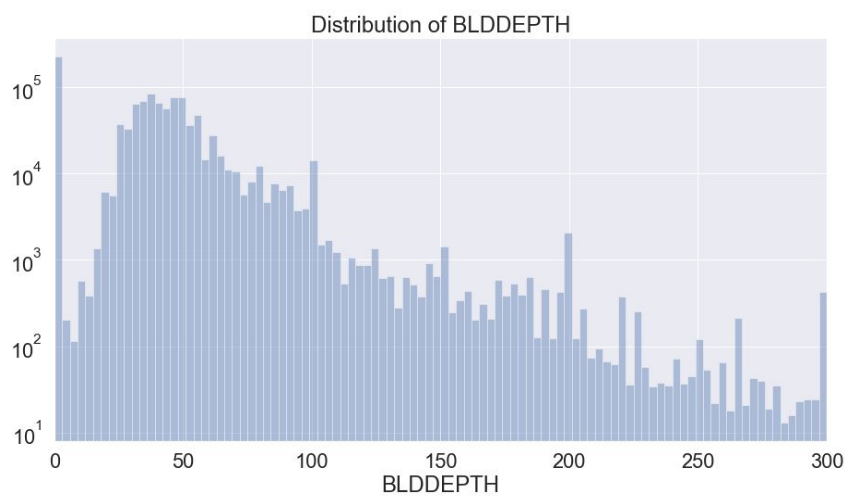
Field Name: **BLDFRONT**

BLDFRONT is a numeric variable measuring the building frontage in feet, with 612 unique values ranging from 0 to 7575. There are 228,815 zero records, possibly indicating missing values. The BLDFRONT distribution is as follows:



Field Name: **BLDDEPTH**

BLDDEPTH is a numeric variable measuring the building frontage in feet, with 621 unique values ranging from 0 to 9393. There are 228,853 zero records, possibly indicating missing values. The BLDDEPTH distribution is as follows:



Data Cleaning

Before we started to create our expert variables, preliminary data cleaning was conducted to prepare the dataset for analysis as shown below.

For the Building Class variable **BLDGCL**, since we considered the original 200 classification levels in this field to be too granular and some of them might contain only a few records, we only took the first character digit [A-Z] of the variable, narrowing it down to 26 building classes.

For the variable **ZIP**, there were originally 29,890 missing records, which we filled in by taking the modes of properties sharing the same **Boro** code and **Block** code. We believe this would serve as a relatively accurate estimate of where the property is located. For cases where fewer than five records remained after the grouping, we then took the mode within the same **Boro** code for approximation.

For the variable **STORIES**, 56,264 missing records were filled in by taking the arithmetic means aggregated by **ZIP** and **BLDGCL**.

For dimensional variables such as **LTFRONT**, **LTDEPTH**, **BLDFRONT**, **BLDDEPTH**, we believe there exists a certain ratio between the Lot Area (i.e, **LTFRONT * LTDEPTH**) and the Building Area (**BLDFRONT*BLDDEPTH**) for each Building Class:

$$BLD - LTRatio = \frac{BLDFRONT * BLDDEPTH}{LTFRONT * LTDEPTH}$$

So, for records with data on either LOT or BLD but missing data on the other, we calculated the average building-lot area ratio of its building class, and multiplied or divided it, depending on whether we had the LOT or BLD data. For instance, if we had a Class A family dwelling located in Queens with **BLDFRONT**=20, **BLDDEPTH**=30, and the average ratio for Class A = 0.8, the LOT area would be 20*30/0.8 = 750. In a similar fashion, we computed the average frontage-depth ratio of that building class, and along with the area data obtained above we were able to determine the respective values of the missing frontage and depth. Namely, if we had a front-depth ratio of 1, then with 85.71 LOT area, **LOTFRONT** = 9.26, **LOTDEPTH** = 9.26. For cases where none of these four fields were available, we used the arithmetic means aggregated by **ZIP** and **BLDGCL** for imputation. If there were fewer than 5 records in the group, we then took the mean aggregated by **BLDGCL** only.

For variables such as **FULLVAL**, **AVLAND**, **AVTOT**, since these three values would be our major target variables for testing anomalies, when replacing the empty and zero values we would want the filled-in values to be as innocuous as possible. So we only grouped them by **ZIP** and **BLDGCL** and took the arithmetic means, assuming in most cases property value does not vary much within the same class of properties (with similar attributes such as number of stories etc.) within the same zip code. Similarly, if there were fewer than 5 records in each group, we aggregated the fields by **BLDGCL** only.

Finally, we removed less informative and pure text fields for the sake of variable creation later on: **EASEMENT**, **STADDR**, **OWNER**, **LOT**, **PERIOD**, **YEAR**, and **VALTYPE**. For more precise predictions, we also removed less strong indicators that would not feed into our fraud detection model, such as **EXLAND**, **EXTOT**, **EXT**, **EXTOT2**, **EXLAND2**, **EXCD1**, **EXCD2**, **AVLAND2**, and **AVTOT2**, considering they might not serve as effectively as **FULLVAL**, **AVLAND** and **AVTOT**.

After the data cleaning process, we were left with the following 14 fields: **RECORD**, **BORO**, **BLOCK**, **BLDGCL**, **TAXCLASS**, **ZIP**, **LTFRONT**, **LTDEPTH**, **STORIES**, **BLDFRONT**, **BLDDEPTH**, **FULLVAL**, **AVTOT**, and **AVLAND**.

Variable Creation

Based on the imputed and adjusted variables mentioned above, we now proceed to creating 45 expert variables. These new variables would have very strong correlations among each other but we would then remove the correlations by dimensionality reduction techniques introduced in the next section.

The first three core variables that we created are:

S1 = LOTAREA = LTFRONT * LTDEPTH
S2 = BLDAREA = BLDFRONT * BLDDEPTH
S3 = BLDVOL = BLDAREA * STORIES

LOTAREA is the area of the lot for the property calculated from the lot frontage and lot depth, assuming that the lot of the property is a rectangle. Similarly, **BLDAREA** is the area of the property also assuming that the building is a rectangle with dimensions **BLDFRONT** and **BLDDEPTH**. Lastly, **BLDVOL** stands for the volume of the building, which is approximated by unit length of height per story multiplied by the number of stories multiplied by the building's area. We assumed that all properties have a similar height per story, therefore the unit length of height per story could be omitted from the formula.

Next, we created nine ratios where each of the three monetary variables outlined below were normalized by the above three calculated dimensional variables.

V1 = FULLVAL (Total market value of the property)
V2 = AVLAND (Actual value of land)
V3 = AVTOT (Actual total value)

The reason that we normalized them is because the interaction of these terms could outline some significance in our results. The calculations of the nine ratios are demonstrated below:

$$r_1 = \frac{V_1}{S_1} \qquad r_4 = \frac{V_2}{S_1} \qquad r_7 = \frac{V_3}{S_1}$$

$$r_2 = \frac{V_1}{S_2} \qquad r_5 = \frac{V_2}{S_2} \qquad r_8 = \frac{V_3}{S_2}$$

$$r_3 = \frac{V_1}{S_3} \qquad r_6 = \frac{V_2}{S_3} \qquad r_9 = \frac{V_3}{S_3}$$

After calculating the above ratios, we then calculated the aggregated averages by the following five groups:

ZIP5, ZIP3, TAXCLASS, B, ALL,
where **ZIP3** is simply the first three digits of **ZIP5**, and **ALL** is grouping by all records.

Having obtained the group averages, for each record, we then divided the ratio by their respective scale average factor. By doing this step, we have created all the 45 variables we need to proceed to the next step.

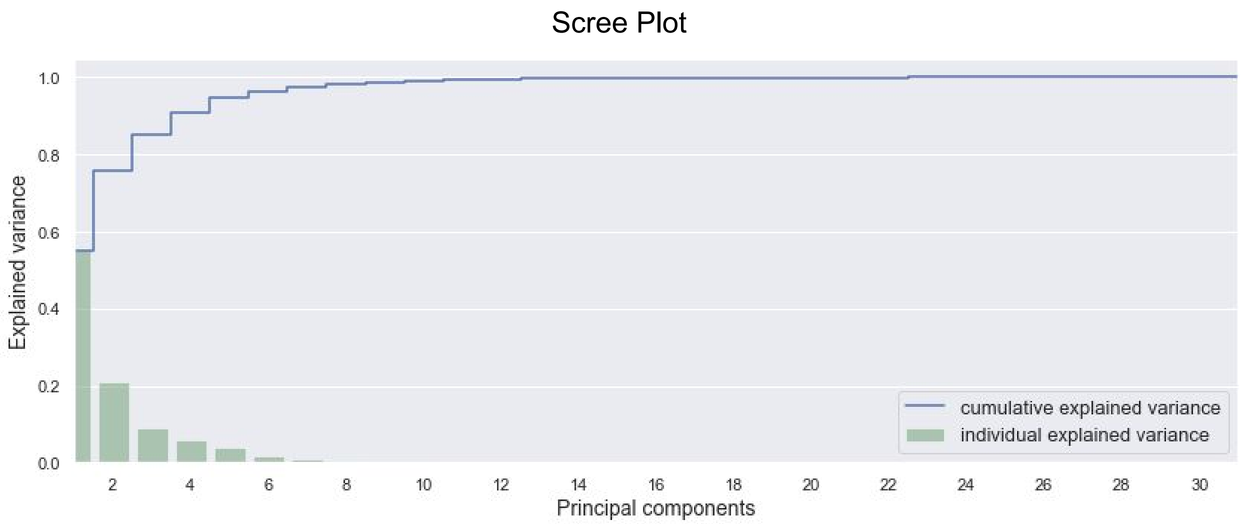
The following shows the calculation for each group <r i>g. The ratios for each record is divided by their respective average of each ratio i for each group g.

$$\frac{r_1}{< r_1 >_g}, \quad \frac{r_2}{< r_2 >_g}, \quad \frac{r_3}{< r_3 >_g}, \quad \ldots \quad \frac{r_9}{< r_9 >_g} \qquad g = 1, \ldots, 5$$

It is important to point out that due to the lack of expertise in the real estate domain, we assumed that other variables that were not used to calculate our 45 key variables do not contribute significantly when trying to identify anomalies. Therefore, we assumed that the variables we created using the size of the building and the market price of the property were enough to identify anomalous records.

Dimensionality Reduction

Before using Principal Component Analysis (PCA) to reduce dimensionality, StandardScaler from Scikit-Learn was used to Z scale the dataset to center the data and get the scales the same. After Z scaling, PCA was used to generate a scree plot and to visualize the result (see the chart below). Furthermore, the top 5 principal components (PC) were kept to account for around 90% cumulative explained variance. This step is for dimensionality reduction and removing correlations. Consequently, the retained PCs were Z scaled again to ensure all the PCs are equally important.



Algorithms

Score 1: Heuristic Function of the Z scores

After dimensionality reduction and scaling, the anomaly or fraud score was calculated by the distance from the origin with the function below. The value of n equals 2 was chosen for the Minkowski distance formula.

$$s_i = \left(\sum_k |z_k^i|^n \right)^{1/n}, \quad n \text{ anything}$$

The reason we chose n = 2 is because we assumed that every principal component has an equal weight on the overall distance. If we chose a bigger n, bigger zk's will outweigh the smaller zk's.

This method is a valid way of finding fraudulent records since fraudulent records tend to exist farthest from the origin of the dataset.

Score 2: Autoencoder

Besides the fraud score calculated from the heuristic function of the Z scores, we also used the autoencoder to find the outliers. An autoencoder is a special type of the neural network that attempts to copy the input values to the output values.

In this model, we used 4 layers in total, including 2 hidden layers. For each hidden layer, there are two neurons. The reason we chose 2 hidden layers and 2 neurons for each layer is that when the number of neurons in hidden layers is less than that of the input layer, the hidden layer could extract the essential pattern of the data and ignore noise. After applying the autoencoder to the data, we got a predicted value for each entry in the data frame. If a record is normal, the autoencoder can return values close to the input values. Otherwise, the predicted values would be far from the input values.

Finally, we calculated the difference between the original and predicted values (reproduction error) for each column, and calculated the Euclidean distance (of the reproduction errors) for each record and derived the fraud score. A high fraud score indicates that a record is an outlier (and likely to be fraudulent).

Combined Scores

After calculating the individual scores, we rank ordered each score from lowest (least likely outliers) to the highest (most likely outliers), assigning an index ranging from 0 to 1,070,993 for Score 1 and 2 of each record. Through this normalization process, both Score 1 and 2 became comparable (having equal weighting) and we simply combined them to get the final fraud score.

Score 1	
Raw	Rank-Order
0.025	1
0.343	2
0.462	3
1.5	4

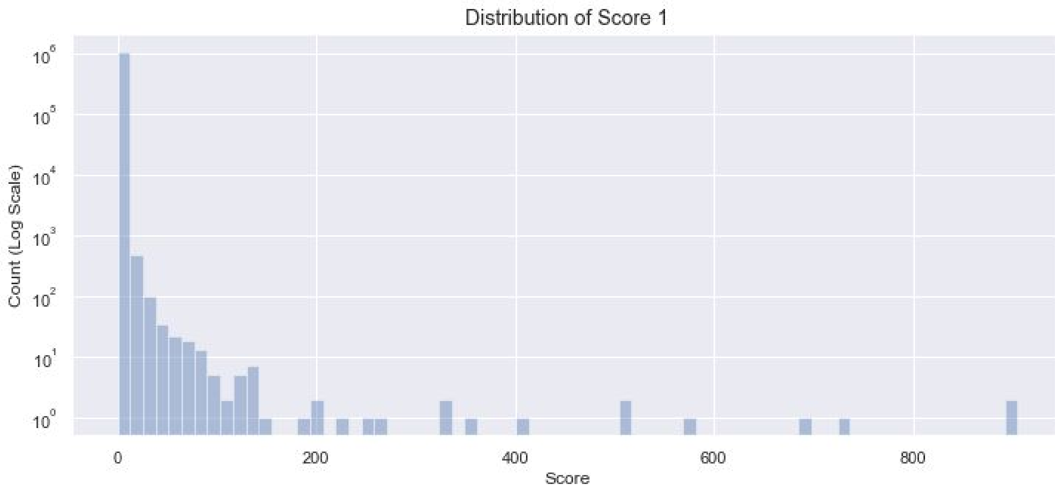
Score 2	
Raw	Rank-Order
5.25	1
6.88	2
9.23	3
9.24	4

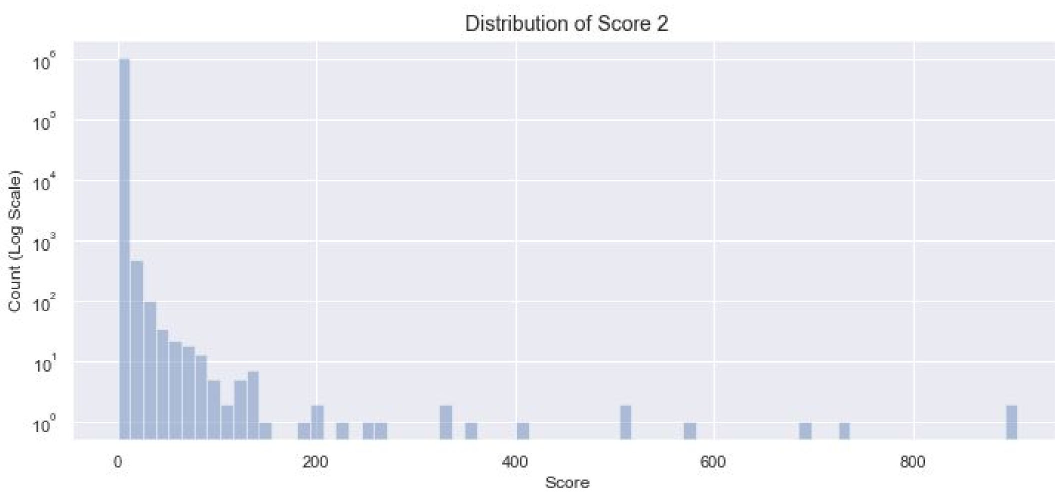
Combined Score
2
4
6
8

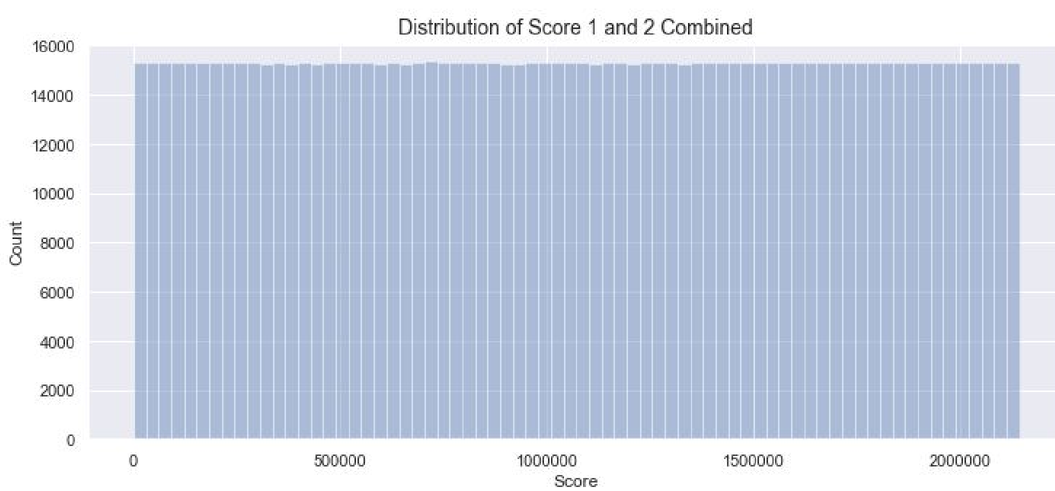
Finally, we sorted the dataset again by the combined score (highest to lowest), and analyzed the top 10 records below.

Results

Fraud Score Distributions







As expected, the distributions for both Score 1 and 2 are extremely right skewed and almost identical. As a result, the combined score distribution is essentially a straight line.

Top 10 Records

Below, we analyzed the top 10 records with the highest fraud score based on our algorithms.

RECORD	BBLE	B	BLOCK	OWNER	BLDGCL	TAXCLASS	EXMPTCL
7056	1000621001	1	62	BROOKFIELD PROPERTIES	R	4	-
565392	308590070o	3	8590	U S GOVERNMENT OWNRD	V	4	X1
776306	4080100001	4	8010	TONY CHEN	Q	4	-
337274	3021111001	3	2111	ONE HANSON PLACE COND	R	4	-
337275	3021111002	3	2111	HANSON PLACE PARTNERS	R	4	-
750816	4066610005E	4	6661	M FLAUM	V	1B	-
565398	3085910100	3	8591	DEPT OF GENERAL SERVI	V	4	X1
1053359	5063730001	5	6373	PARKS AND RECREATION	V	4	-
378985	3037711002	3	3771	N/A	R	2	-
6837	1000471001	1	47	120 BROADWAY HOLDINGS	R	4	-

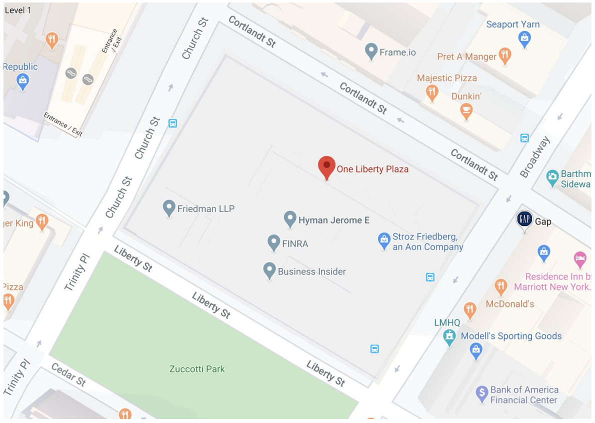
Below are detailed examinations of the 10 Records:

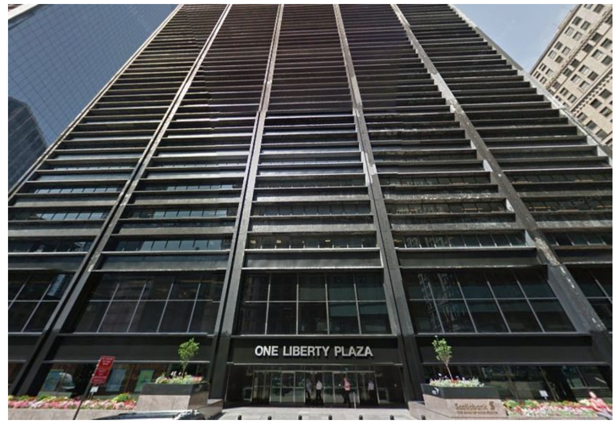
Record 7056
Owner: BROOKFIELD PROPERTIES
Address: 1 LIBERTY PLAZA

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R1	61	497	889	514	912
R4	31	509	815	504	884
R7	33	374	848	402	913

This property has an extremely high unit lot area market value, land value, and total value for its ZIP3, TAX and BORO class. However, upon further inspection, this property has a Lot Area of

1.36 sqft, and a building area of 16,437 sqft, which indicates that there may be some erroneous data for this record.





Record 565392
Owner: U S GOVERNMENT OWNRD
Address: FLATBUSH AVENUE

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R1	230	15	1	15	1
R2	395	838	877	843	909
R3	397	838	831	844	909
R4	606	90	3	84	3
R5	353	851	885	799	909
R6	355	851	853	799	909
R7	540	17	1	17	1
R8	356	851	894	799	909
R9	357	851	880	799	909

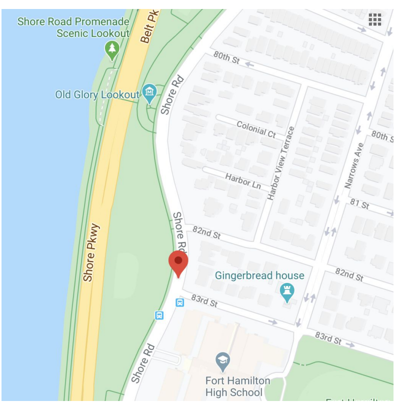
This property has extremely high values in a large number of our ratios. In particular ratios that pertain to the building area and volume are disproportionately high (R2, R3, R5, R6, R8, R9).

This record does not specify the exact location of the properties, so we cannot locate and investigate this property. However, based on the owner field, we know that this property is owned by the U.S. Government. In addition, this record belongs to exemption class X1, so we believe it may not be fraudulent.

Record 776306
Owner: TONY CHEN
Address: SHORE ROAD

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R1	609	40	4	55	5
R4	618	454	14	501	15
R7	656	135	4	131	5

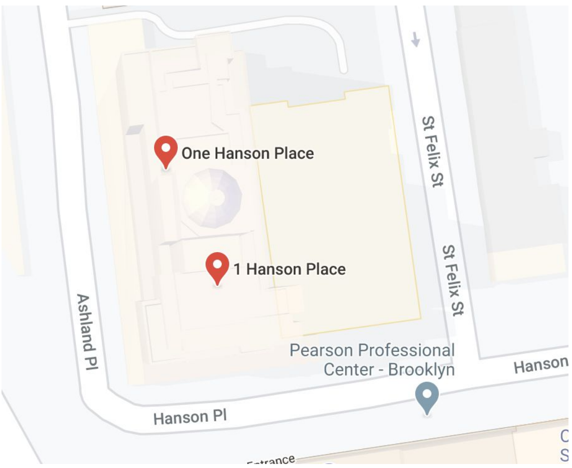
This properties' unit lot area market value, actual land value, and total value are higher than normal based on different classes (ZIP5, ZIP3 and BORO). Since this property seems to be owned by an individual, we believe this property may be fraudulent in nature (especially if we can confirm that it is residential).



Record 337274
Owner: ONE HANSON PLACE COND
Address: 1 HANSON PLACE

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R1	4	525	31	509	32
R4	2	382	13	355	14
R7	2	589	30	593	32

This property's unit lot area market value, unit lot area actual land value, and unit lot area actual value are deemed to be higher than average. However, upon further inspection, this is a historic building that has been retrofitted into a modern luxury apartment building. While the values are considered outliers for its ZIP3 class and BORO, they seem to be relatively normal when looking at the ZIP5 class, suggesting that it is located in a high valued area/neighborhood. Hence this property may not be a fraudulent case.



Record 337275
Owner: HANSON PLACE PARTNERS
Address: 1 HANSON PLACE

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R1	3	435	26	421	27
R4	2	316	11	293	12
R7	2	488	25	491	27

This property's unit lot area market value, unit lot area actual land value, and unit lot area actual value are deemed to be higher than average.

While the owners are different, this record is similar to the **Record 337274** in terms of the different valuations. This suggests that there may be a change in ownership for this building, hence this property may not be fraudulent.

Record 750816
Owner: M FLAUM
Address: VLEIGH PLACE

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R1	508	34	207	46	4
R4	363	267	355	294	9
R7	352	72	348	70	2

This property's unit lot area market value, unit lot area actual land value, and unit lot area actual value are higher than average. Again, there is no street number associated with this property, and the owner seems to be an individual, leading us to conclude that this property might be fraudulent.

Record 565398
Owner: DEPT OF GENERAL SERVI
Address: FLATBUSH AVENUE

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R2	211	448	468	451	485
R3	212	448	444	451	485
R4	9	1	0	1	0
R5	189	454	473	427	486
R6	190	454	456	427	486
R7	8	0	0	0	0
R8	190	455	478	427	486
R9	191	455	470	427	486

This properties' unit building area market value, unit building volume market value, unit lot area actual land value, unit building area actual land value,unit building volume actual land value, unit lot area actual value, unit building area actual value and unit building volume actual value higher than average. Since this is a government owned property, we deduce that it may not be fraudulent.

Record 1053359

Owner: PARKS AND RECREATION
Address: HOLDRIDGE AVENUE

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R2	239	6	0	6	0
R3	240	6	0	6	0
R4	15	5	0	5	0
R5	391	14	0	13	0
R6	393	14	0	13	0
R7	9	2	0	2	0
R8	394	14	0	13	0
R9	395	14	0	13	0

This property's unit building area market value, unit building volume market value, unit lot area actual land value, unit building area actual land value,unit building volume actual land value, unit lot area actual value, unit building area actual value and unit building volume actual value are higher than average.

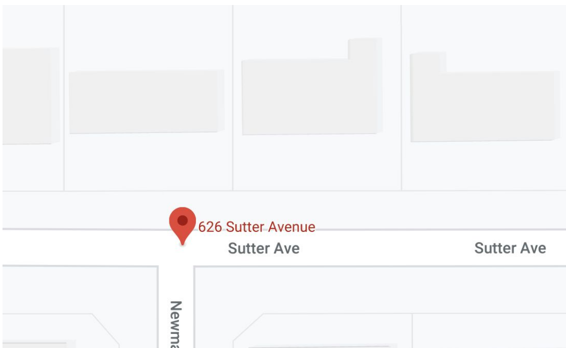
Upon further inspection, this parcel of land (along with many others in the area) are all owned by PARKS AND RECREATION, with valuations ranging from a few thousand dollars to 10s of millions of dollars. We are unsure of the validity of this owner, hence we believe it may be fraudulent.

Record 378985
Owner: No Record
Address: 626 SUTTER AVENUE

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R1	299	236	9	228	14
R4	127	279	8	259	10

R7	216	264	8	266	14

This property has a high unit lot area market value, land value, and total value for its ZIP5, ZIP3 and BORO class. However, upon further inspection, this property has a Lot Area of 1.48 sqft, and a building area of 0.37 sqft, which indicates that this property's information may be fraudulent. Online research also indicates this property is a low income apartment complex.



Record 6837
Owner 120 BROADWAY HOLDINGS
Address 120 BROADWAY

Ratio	ZIP5	ZIP3	TAXCLASS	BORO	ALL
R1	62	179	320	185	328
R4	28	202	324	200	351
R7	34	134	305	145	329

This property's unit lot area market value, unit lot area actual land value, and unit lot area actual value are higher than average. Online research indicates this is a historic building that has been retrofitted, with a large building area relative to its lot.

Conclusion

Summary

The goal of this project was to identify fraudulent records within the City of New York Property Valuation and Assessment Dataset. To do so, we first had to explore the data and understand the attributes as well as the quality of the data. At the conclusion of our data exploration process, we identified null, invalid and missing values, which we developed systematic methods to fill in. Once the dataset is complete, we created a series of 45 expert variables (dataset) which we believe will best help us identify fraudulent entries.

The next step in our fraud detection process was to normalize the dataset so that values among different columns are comparable. We did so by first normalizing (z-scale) the data so that all the columns are centered at 0. Next, we performed PCA on the data to remove the linear correlations and reduce the dimensionality. Finally used z-scaling to center the data again.

Afterwards, we combined the z-score of the 45 columns into Score 1 by calculating the Euclidean distance. We also ran the 45 columns through an Autoencoder and used the combined (among all the columns) reproduction error as Score 2. We rank-ordered both scores and combined them to arrive at our final fraud score.

Finally, we individually evaluated the top 10 records (ranked from highest to lowest fraud score) by hand and made a determination on whether the property record is truly fraudulent.

Future Work

If time allows in the future, we hope to continue expanding and improving our fraud detection model for this dataset. We believe there are three key areas we can improve on:

1. Methods to Fill Missing Values
    a. Identify better/more intelligent methods to fill missing values, whether it is through some sort of linear regression model or other more complex methods.
2. Scoring Mechanisms
    a. Explore other potential scoring/outlier detection mechanisms.
    b. Look to potentially assign different weightings to each score instead of weighing both scores equally when we combined them.
3. Expert Variables
    a. Consult with different domain experts to better understand the types of data we can use to identify fraudulent records.

Appendix

Data Quality Report: New York Property Data

Part I - Data Description

This dataset represents property valuations and assessments conducted in NYC as of November 2010 for the purpose of calculating property tax and granting eligible properties exemptions and/or abatements. The dataset consists of 1,070,994 records and 32 fields. The data was collected and recorded into the system primarily by various City employees such as property assessors, property exemption specialists, ACRIS reporting etc.

Part II - Summary Tables of Fields

Below are two summary tables listing all the fields for the property dataset. Specifically, Table 1 represents the numeric fields, while Table 2 represents the categorical and time/date fields.

Field name	Field type	# of non-NA	% populated	# unique values	# '0' records	Mean	Std	Min	Max
LTFRONT	numeric	1070994	100	1297	169108	36.64	74.03	0	9999
LTDEPTH	numeric	1070994	100	1370	170128	88.86	76.4	0	9999
STORIES	numeric	1014730	94.74656254	111	0	5.01	8.37	1	119
FULLVAL	numeric	1070994	100	109324	13007	874264.51	11582430.9	0	6150000000
AVLAND	numeric	1070994	100	70921	13009	85067.92	4057260.06	0	2668500000
AVTOT	numeric	1070994	100	112914	13007	227238.17	6877529.31	0	4668308947
EXLAND	numeric	1070994	100	33419	491699	36423.89	3981575.79	0	2668500000
EXTOT	numeric	1070994	100	64255	432572	91186.98	6508402.82	0	4668308947
BLDFRONT	numeric	1070994	100	612	228815	23.04	35.58	0	7575
BLDDEPTH	numeric	1070994	100	621	228853	39.92	42.71	0	9393
AVLAND2	numeric	282726	26.39846722	58591	0	246235.72	6178962.56	3	2371005000
AVTOT2	numeric	282732	26.39902745	111360	0	713911.44	11652528.9	3	4501180002
EXLAND2	numeric	87449	8.165218479	22195	0	351235.68	10802212.6	1	2371005000
EXTOT2	numeric	130828	12.21556797	48348	0	656768.28	16072510.1	7	4501180002

Table 1: Numeric Fields of NYC Property Data

Field name	Field type	# of non-NA	% populated	# unique values	Most Common Field
Record	categorical	1070994	100	1070994	NA
BBLE	categorical	1070994	100	1070994	NA
B	categorical	1070994	100	5	4
BLOCK	categorical	1070994	100	13984	3944
LOT	categorical	1070994	100	6366	1
EASEMENT	categorical	4636	0.432868905	12	E
OWNER	categorical	1039249	97.03593111	863346	PARKCHEST ER PRESERVAT
BLDGCL	categorical	1070994	100	200	R4
TAXCLASS	categorical	1070994	100	11	1
EXT	categorical	354305	33.08188468	3	G
EXCD1	categorical	638488	59.61639374	129	1017
STADDR	categorical	1070318	99.93688107	839280	501 SURF AVENUE
ZIP	categorical	1041104	97.20913469	196	10314
EXMPTCL	categorical	15579	1.454629998	14	X1
EXCD2	categorical	92948	8.678666734	60	1017
PERIOD	categorical	1070994	100	1	FINAL
YEAR	date/time	1070994	100	1	2010/11
VALTYPE	categorical	1070994	100	1	AC-TR

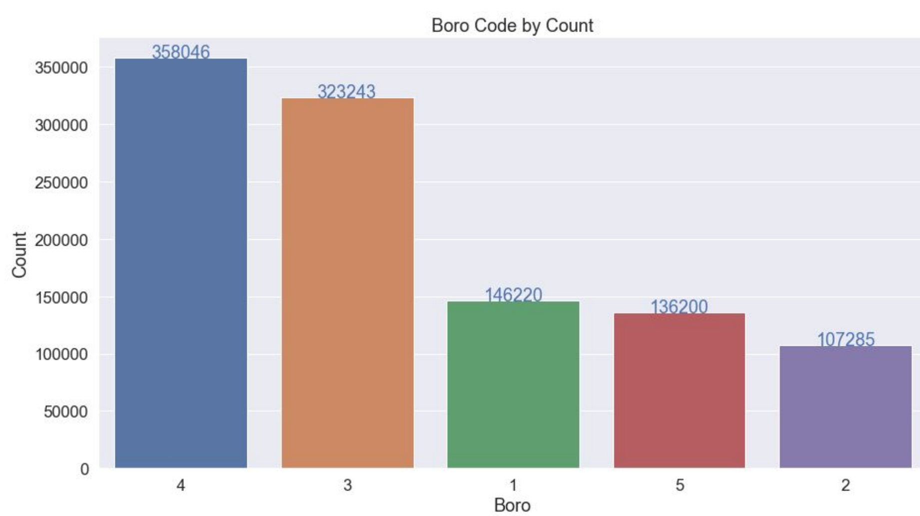
Table 2: Categorical Fields of NYC Property Data

Part III - Field Description

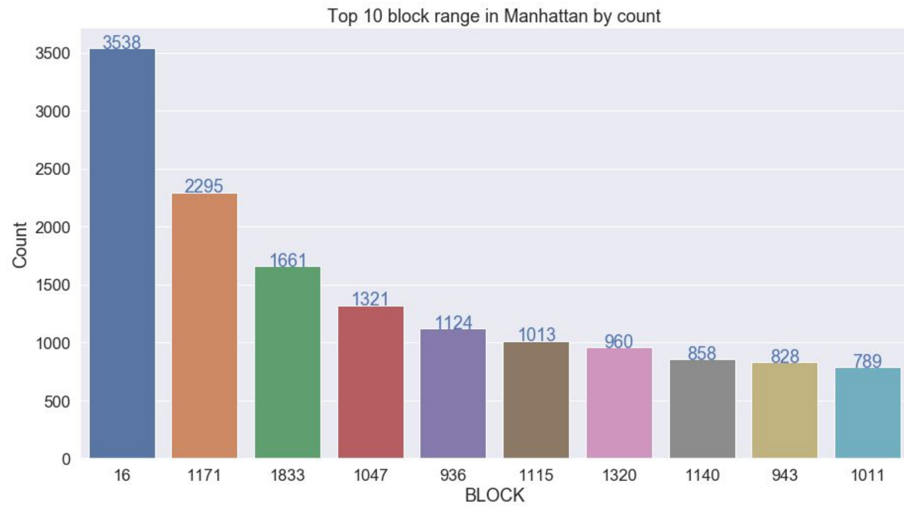
1. **Record**: Unique record numbers for properties in the data.

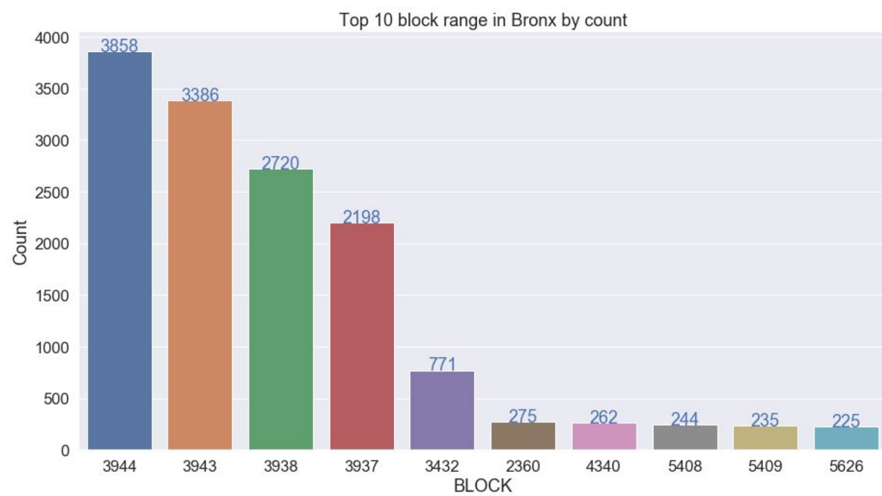
2. **BBLE**: Concatenation of Boro, Block, Lot, Easement code, Length is 11 alphanumeric. No duplicate combinations were found in the data.

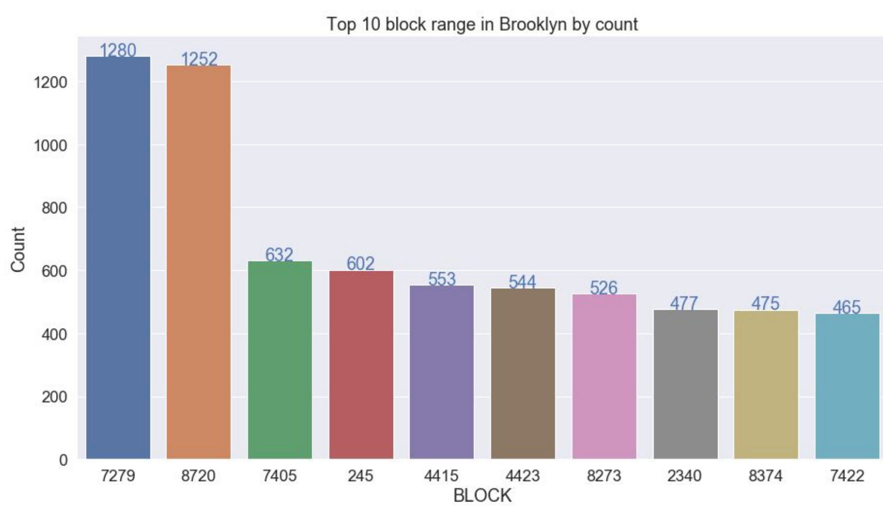
3. **B**: Stands for Boro codes, which are 1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island.

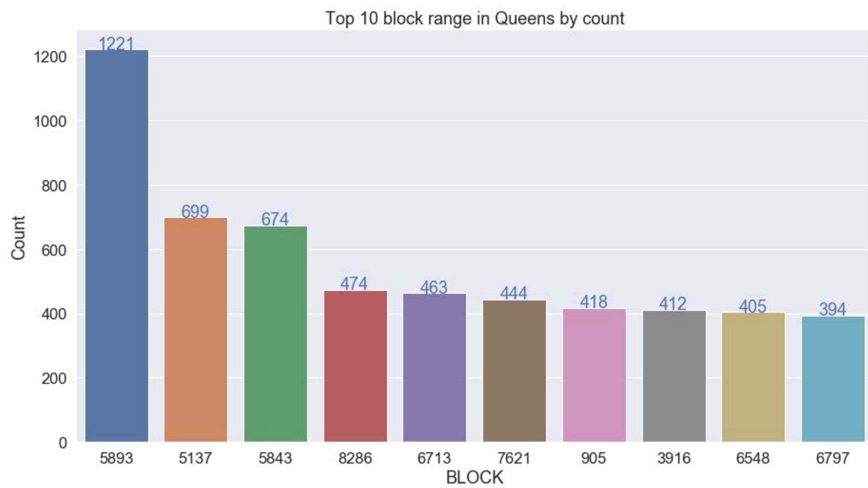


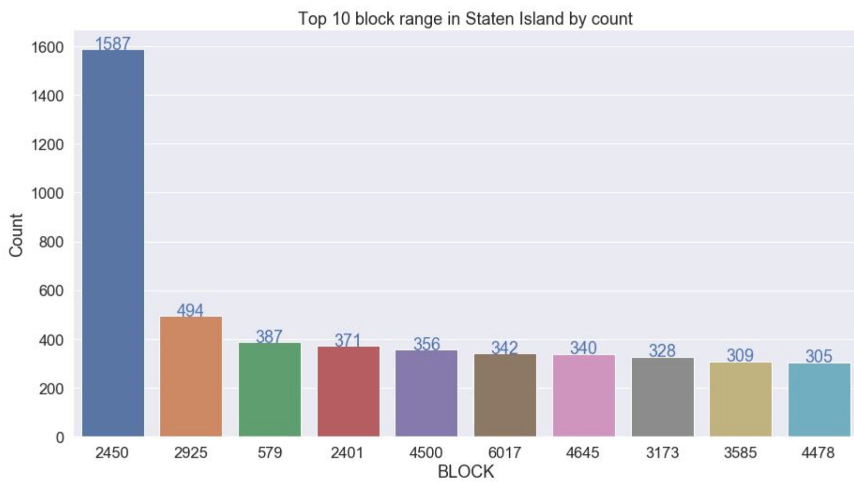
4. **BLOCK**: Valid block ranges by Boro, specifically: Manhattan 1 ~ 2,255, Bronx 2,260 ~ 5958, Brooklyn 1 ~ 8,955, Queens 1 ~ 16,350, Staten Island 1 ~ 8,050. Below are the Top blocks by Boro Code 1~5.



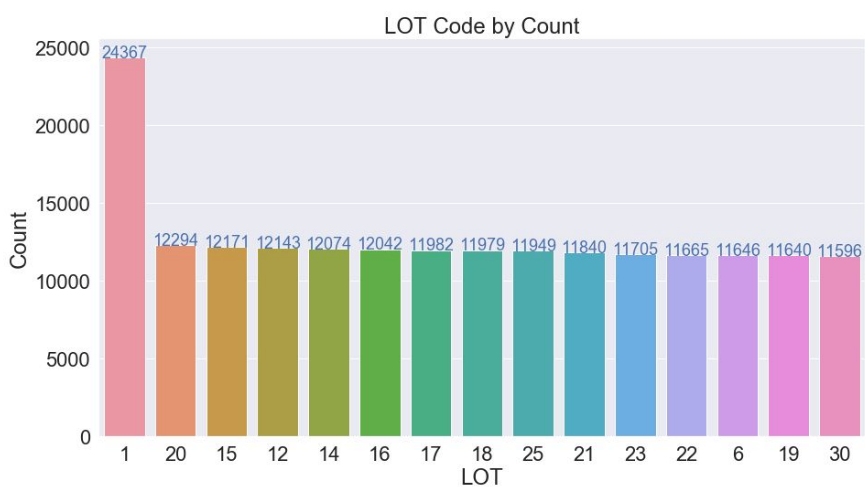






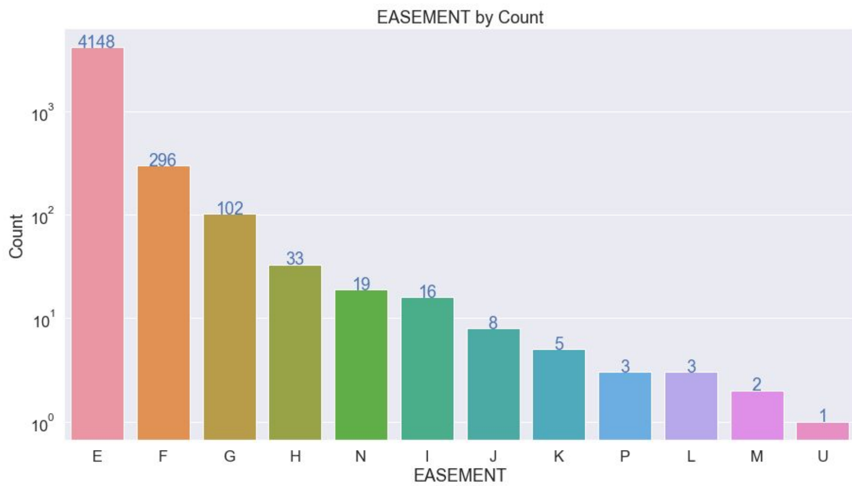


5. **LOT**: Unique numbers within Boro or Block.

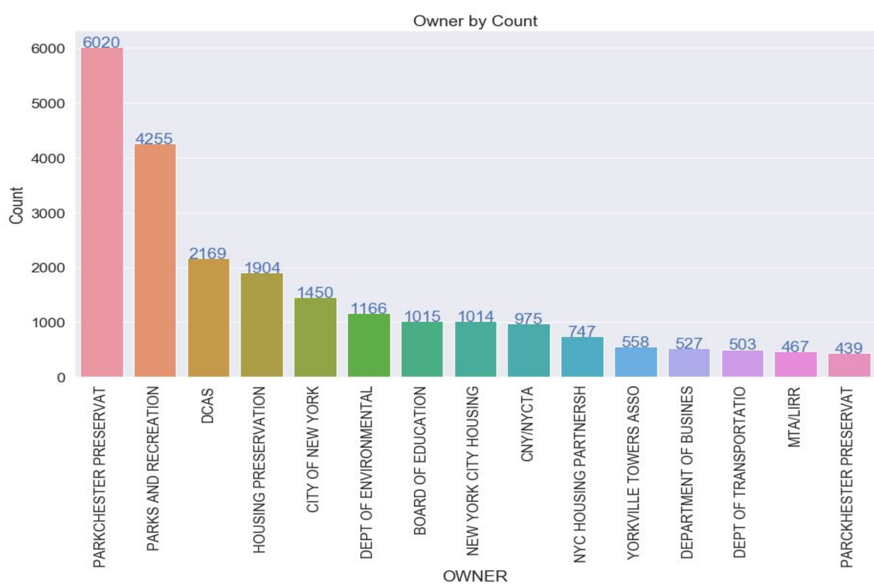


6. **EASEMENT**: SPACE: the lot has no Easement.
'A': the portion of the Lot that has an Air Easement.

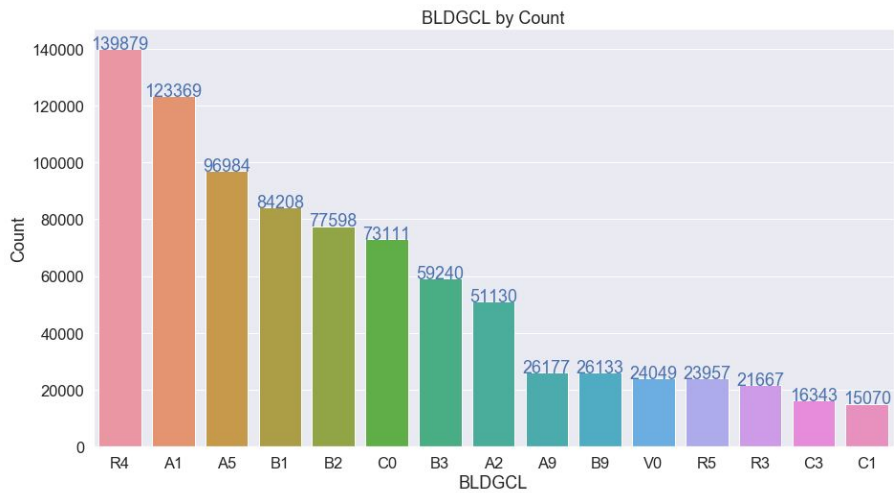
'B': Non-Air Rights.
'E': the portion of the lot that has a Land Easement.
'F' through 'M': duplicates of 'E'.
'N': Non-Transit Easement.
'P': Piers.
'R': Railroads
'S': Street.
'U': U.S. Government.



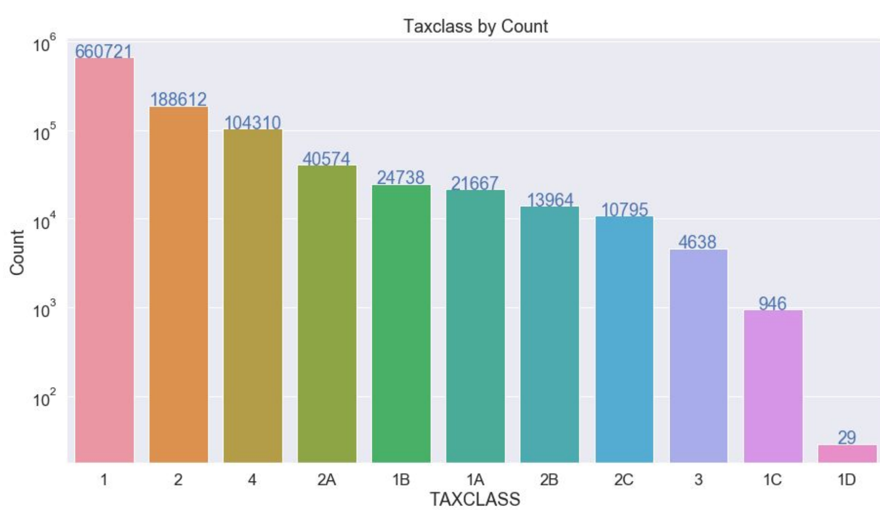
7. **OWNER**: Owners of properties in the data.



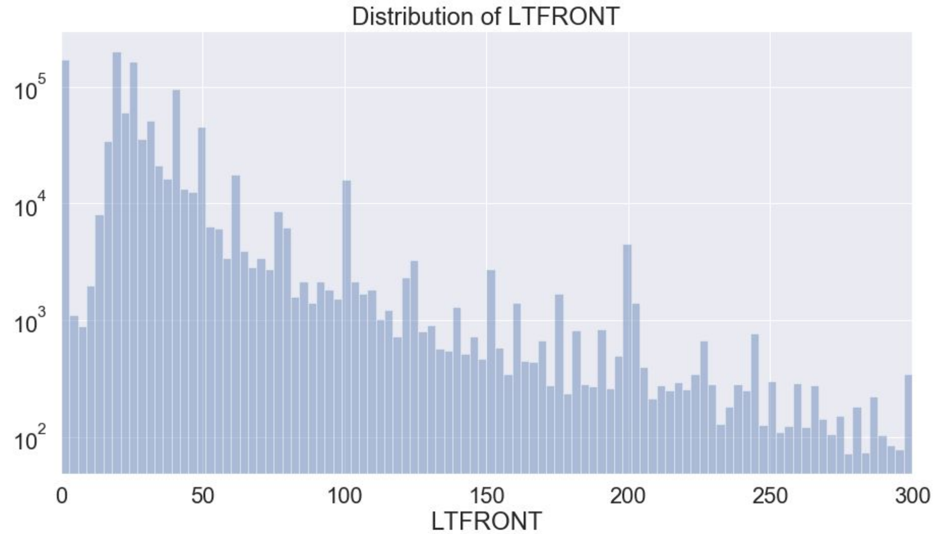
8. **BLDGCL**: Used to denote building class: Position 1 = ALPHA and Position 2 = NUMERIC.



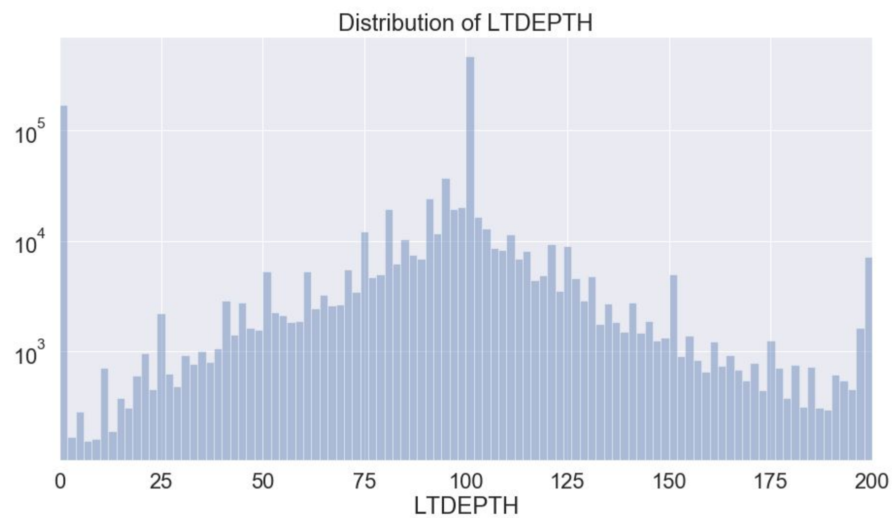
9. **TAXCLASS**: Current Property Tax Class Code by NYS Classification.
Tax Class 1 = 1-3 Unit Residences
Tax Class 1a = 1-3 Story Condominiums Originally A Condo
Tax Class 1b = Residential Vacant Land
Tax Class 1c = 1-3 Unit Condominiums Originally Tax Class 1
Tax Class 1d = Select Bungalow Colonies
Tax Class 2 = Apartments
Tax Class 2a = Apartments With 4-6 Units
Tax Class 2b = Apartments With 7-10 Units
Tax Class 2c = Coops/condos With 2-10 Units
Tax Class 3 = Utilities (Except Ceiling Rr)
Tax Class 4a = Utilities - Ceiling Railroads
Tax Class 4 = All Others



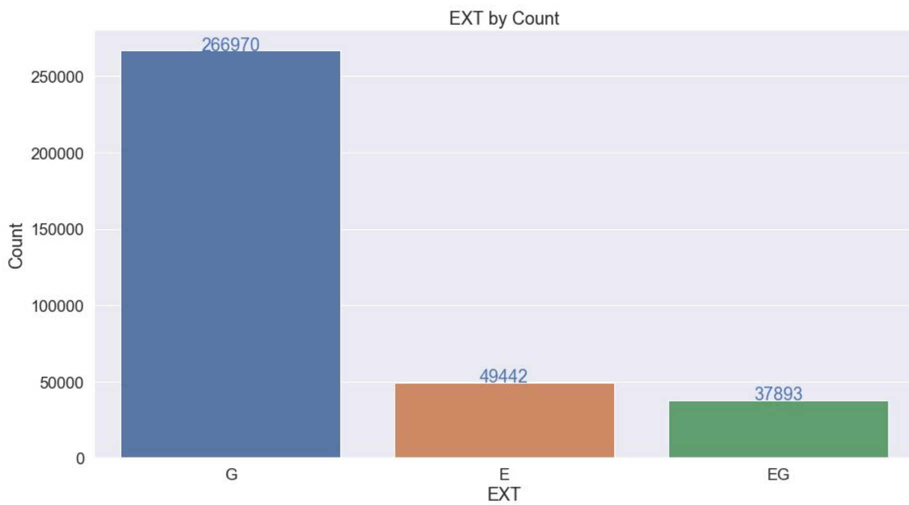
10. **LTFRONT**: Lot frontage in feet. The cut-off values for x to check for outliers were set to be between 0 and 300.



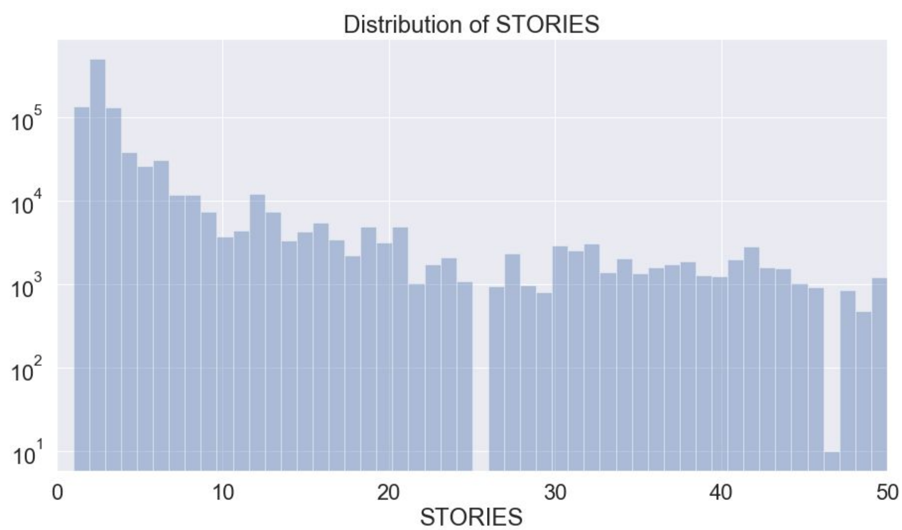
11. **LTDEPTH**: Lot depth in feet. The cut-off values for x to check for outliers were set to be between 0 and 200.



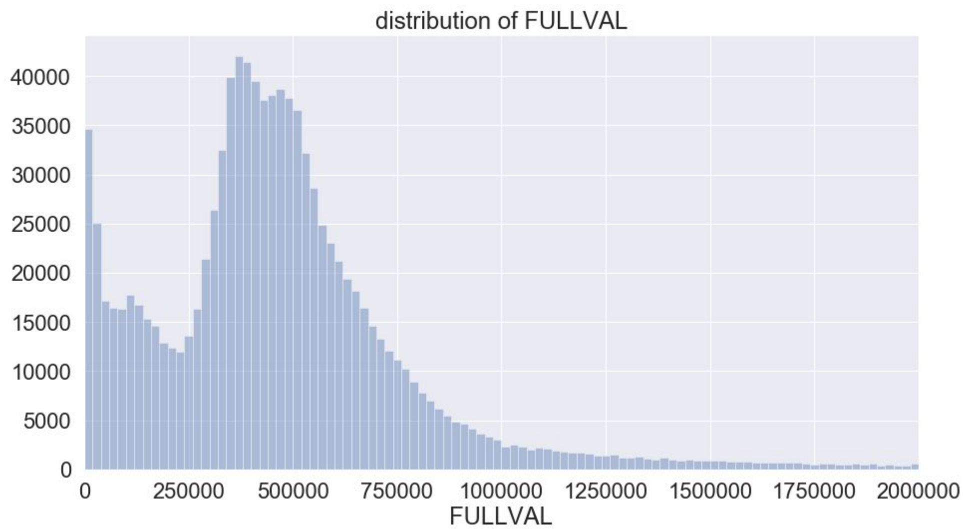
12. **EXT**: Extension, with ‘E’ = Extension, ‘G’ = Garage, ‘EG’ = Extension and Garage.



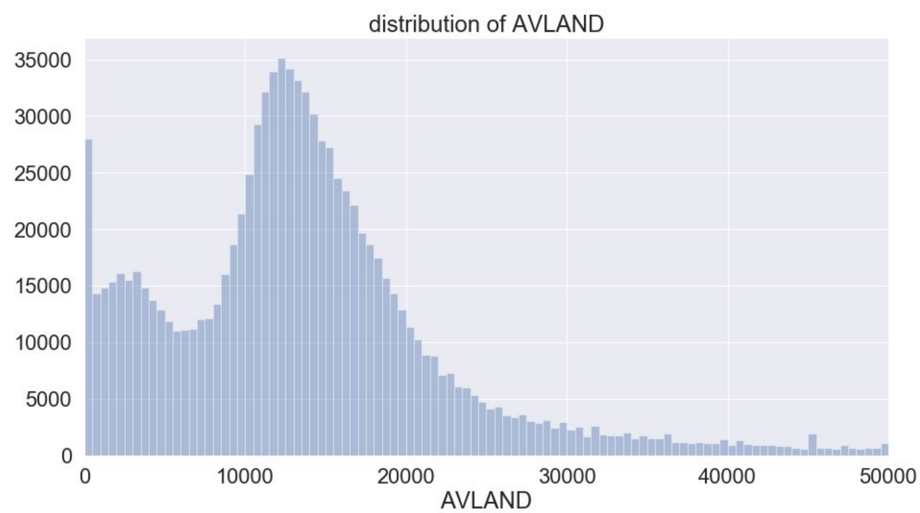
13. **STORIES**: Number of stories for the building. The cut-off values for x to check for outliers were set to be between 0 and 50.



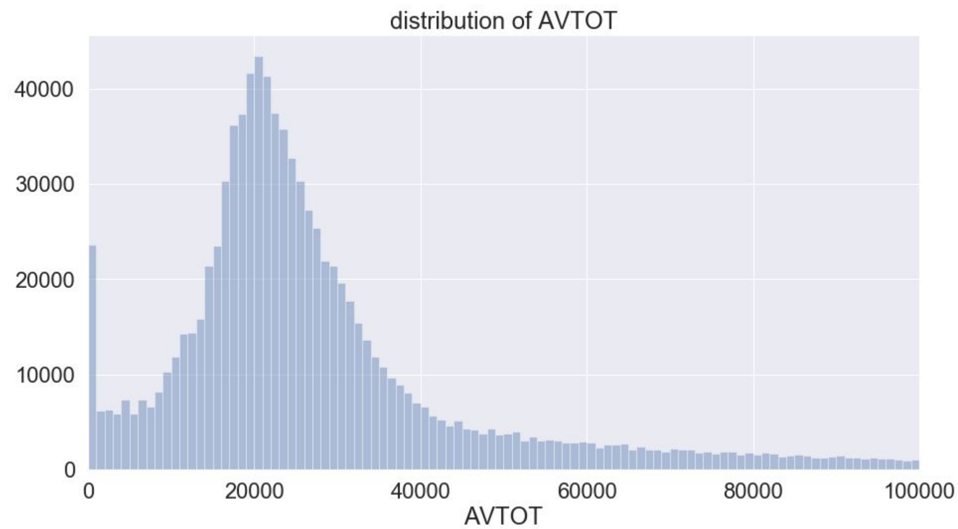
14. **FULLVAL**: Total market value of the property. The cut-off values for x to check for outliers were set to be between 0 and 2,000,000.



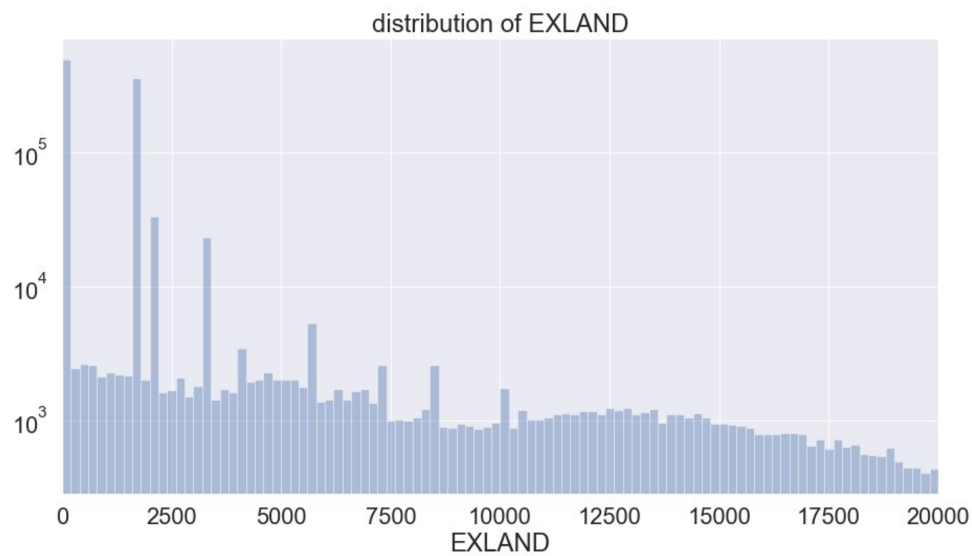
15. **AVLAND:** Actual land value. The cut-off values for x to check for outliers were set to be between 0 and 50,000.



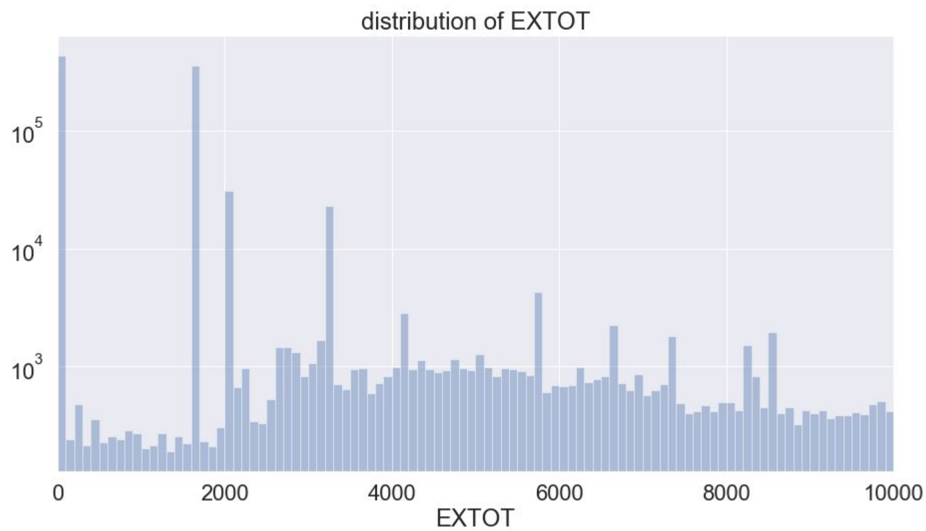
16. **AVTOT**: Actual total value. The cut-off values for x to check for outliers were set to be between 0 and 100,000.



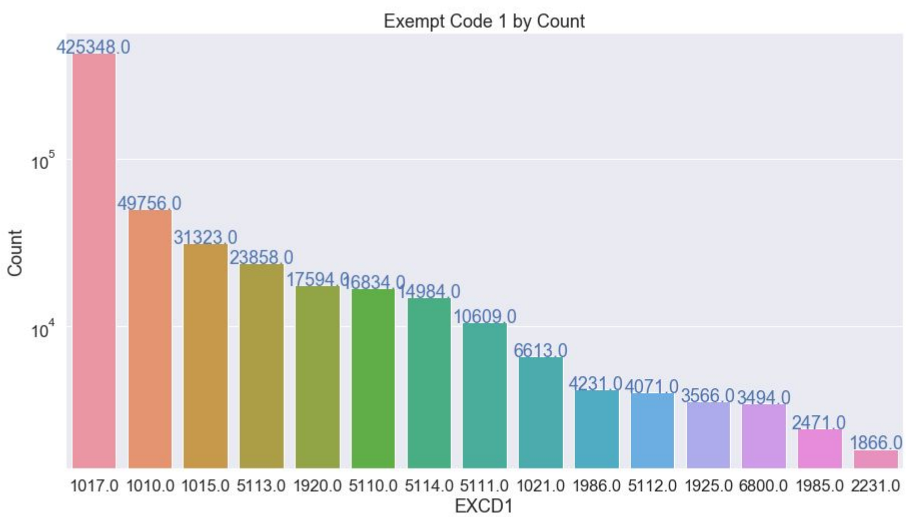
17. **EXLAND**: Actual exempt land value. The cut-off values for x to check for outliers were set to be between 0 and 20,000.



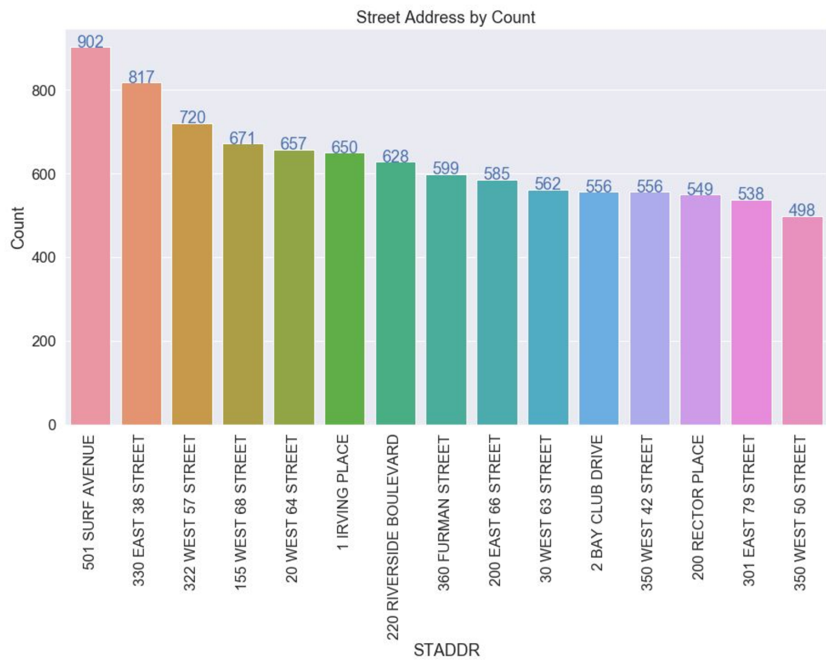
18. **EXTOT**: Actual exempt land total. The cut-off values for x to check for outliers were set to be between 0 and 10,000.



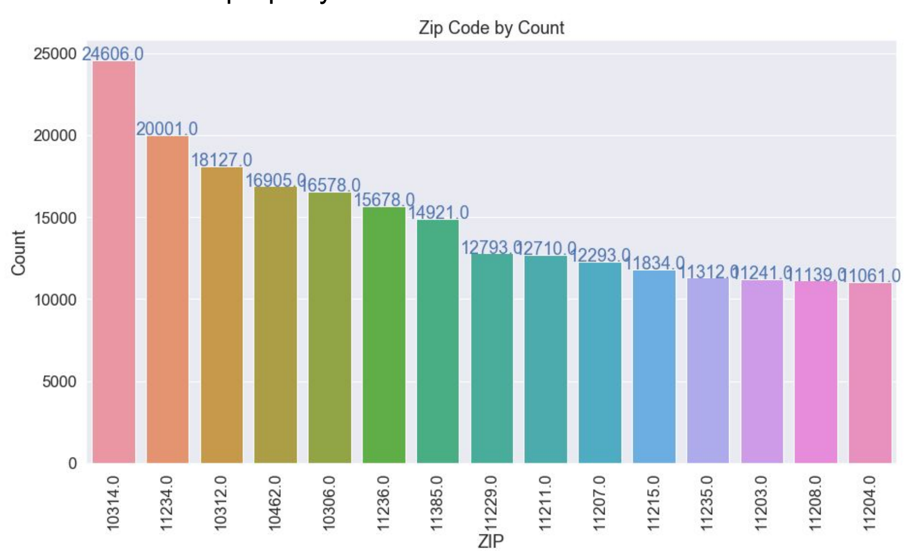
19. EXCD1: Exemption code 1.



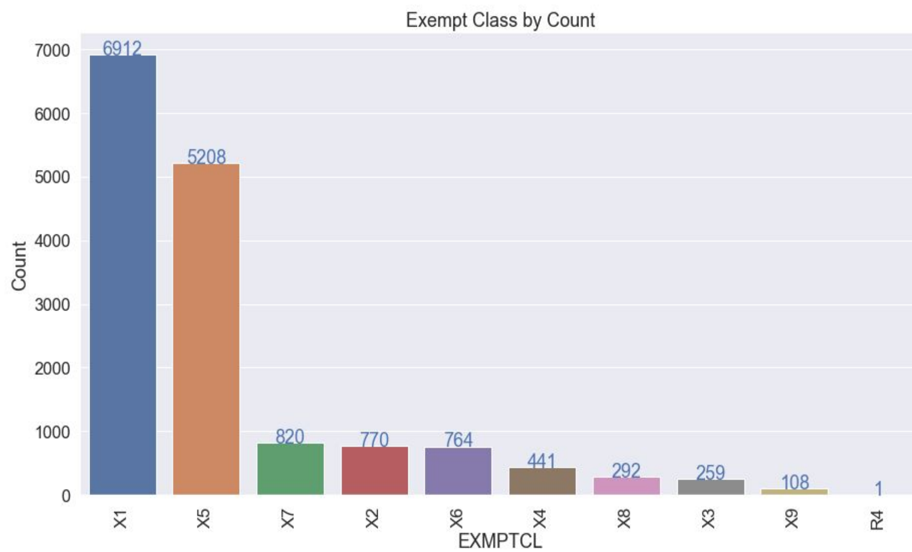
20. STADDR: Street Address for the property.



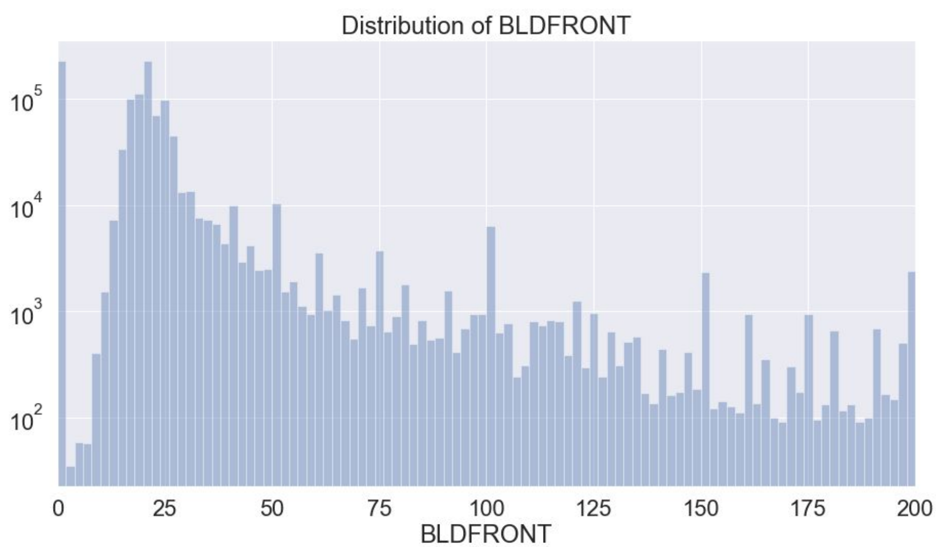
21. **ZIP**: Zip code in which the property is located.



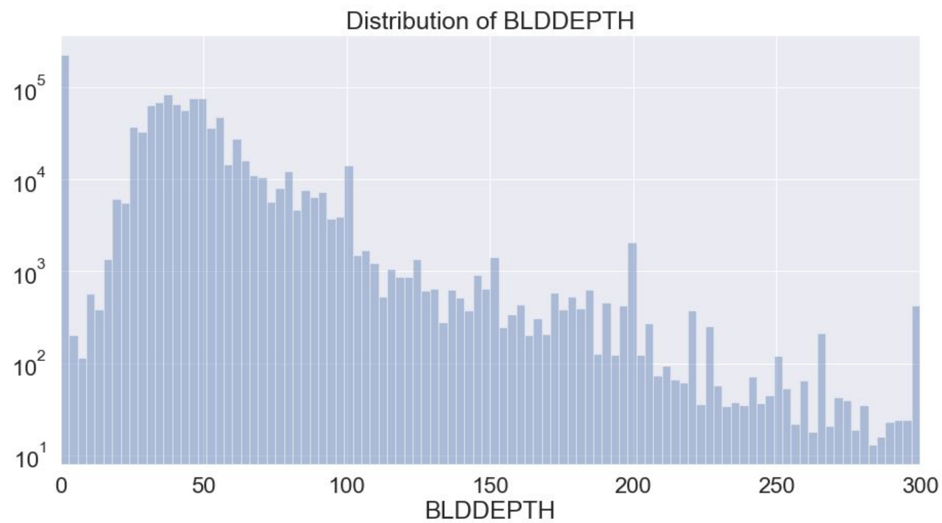
22. **EXMPTCL**: Exempt class used for fully exempt properties only.



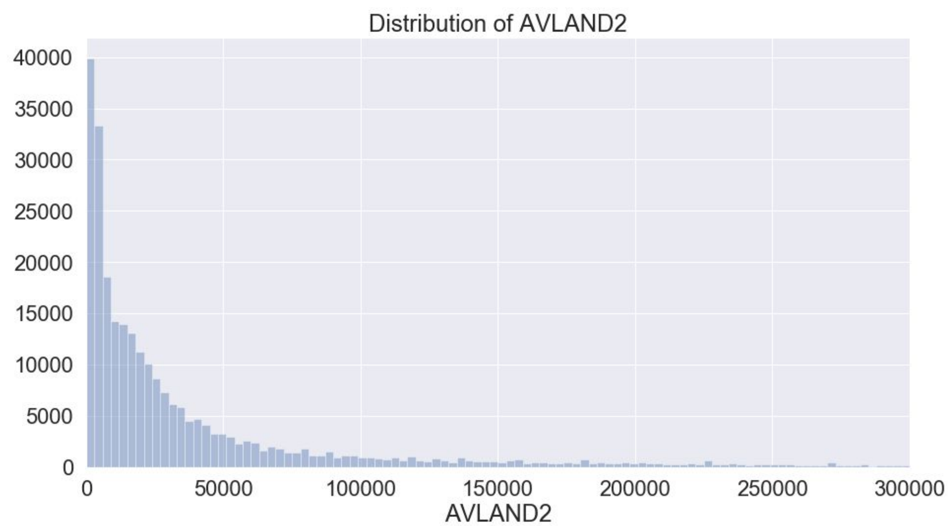
23. **BLDFRONT**: Building frontage in feet. The cut-off values for x to check for outliers were set to be between 0 and 200.



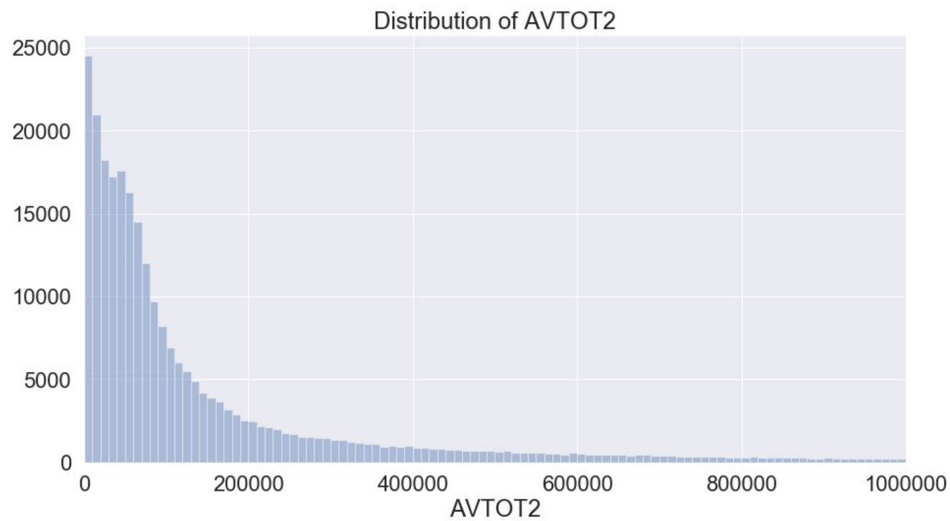
24. **BLDDEPTH**: Building depth in feet. The cut-off values for x to check for outliers were set to be between 0 and 300.



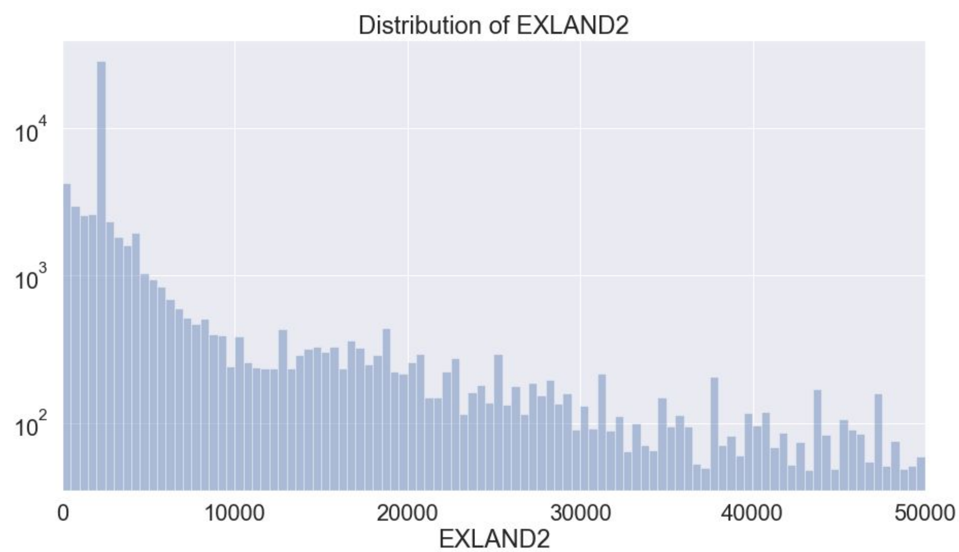
25. **AVLAND2**: Transitional land value. The cut-off values for x to check for outliers were set to be between 0 and 300,000.



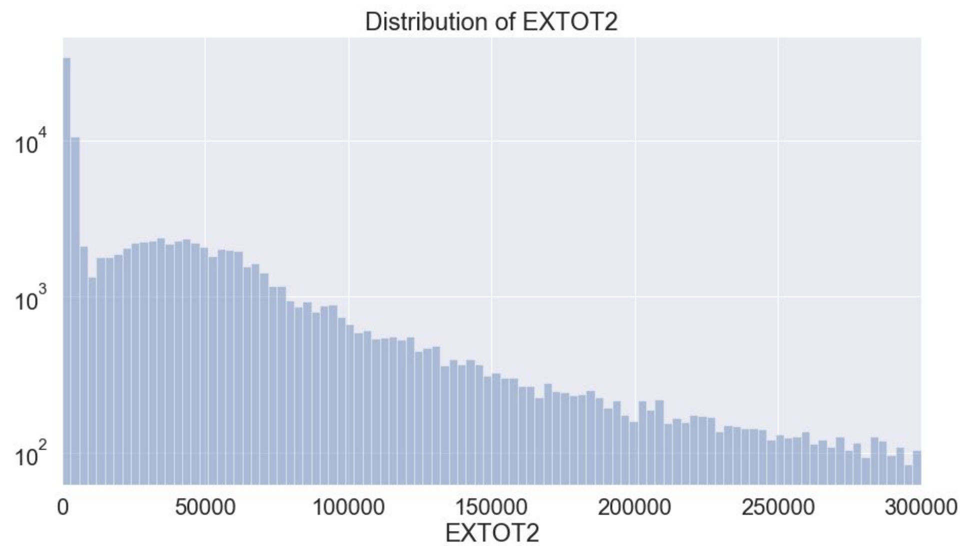
26. **AVTOT2**: Transitional total value. The cut-off values for x to check for outliers were set to be between 0 and 1,000,000.



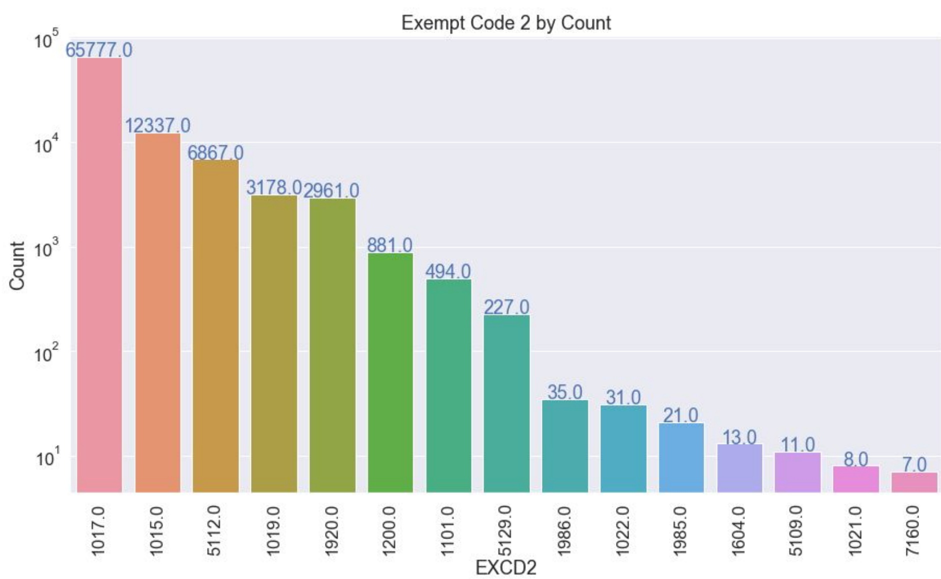
27. **EXLAND2**: Transitional exempt land value. The cut-off values for x to check for outliers were set to be between 0 and 50,000.



28. **EXTOT2**: Transitional exempt land total. The cut-off values for x to check for outliers were set to be between 0 and 300,000.



29. **EXCD2**: Used to denote Exemption Code 2.



30. **PERIOD**: Assessment period when file was created. It was found that all the data entries took the value of ‘Final’.

31. **YEAR**: Year when the assessment was conducted. All the data entries took the value of ‘2010/11’.

32. **VALTYPE**: Valuation type. All the data entries took the value of ‘AC-TR’.