UNSUPERVISED FRAUD DETECTION MODEL REPORT

NY Property Data

February 13, 2020

Prepared by Team 7:
Ji Jie
Joycy Liu
Angela Ma
Emma Yang
Jenny Zhang
Wanqiu Zheng
Xindi Zhu

Content

1. Executive Summary

The purpose of this project is to look for anomalies among property records in the "Property Valuation and Assessment Data" file using unsupervised models and find out potential fraudulent properties through investigating further on the top ten most possible properties to have committed property tax frauds and researching on reasons that these properties turn out to be anomalous and therefore detected in the models.

The "Property Valuation and Assessment Data" file is a real estate assessment property dataset that consists of properties located in New York City. The document is mainly used to calculate property tax and grant eligibility to the properties that qualify for exemptions and/or abatements. The file is updated annually and collected and entered by the City employees such as property assessors and property tax specialists. The properties in the file were assessed in November 2010 and were from 5 main areas including Manhattan, Bronx, Brooklyn, Queens, and Staten Island.

The report begins with a detailed description of the dataset. The full data quality assessment is included as an appendix of this report. As for the size of the file, the dataset has 32 fields, including both numerical and categorical fields, and 1,070,994 records. Among the 32 fields, 14 of them are numerical fields while 18 are categorical.

Followed by the description of data and distributions, the report explains how each numerical variable has been cleaned and filled in for missing values and zero values. We also included an additional 45 variables we calculated and utilized in the model. All data used are then converted into z scores in order to be compared on the same scale. Principal component analysis (PCA) is performed after the conversion to remove correlations, and dimensions are set to be reduced to six, by our choice. Then the six dimensions are again z-scaled, ready to be used for the models.

As for the detection process, we used two methods to select anomalous properties. The first method, heuristic algorithm, calculates the distance of z-scale from the origin. The second method is the autoencoder. Both methods give a score for each record where a higher score indicates a more abnormal record. We decided to combine the two methods by giving their ranks equal weight, 50%, to generate a combined score with quantile binning.

Based on the new combined score, we selected the top ten properties with the highest anomaly score to further investigate. Through further research, some properties have reasonable explanations such as government properties for high anomaly scores. From a conservative point of view to prevent frauds, there are still six properties that seem suspicious after research and therefore worth more professional investigation for potential property tax frauds.

2. Description of Data

2.1 Description

The dataset represents the property valuation and assessment data of New York City. The data is mainly used for property assessments in NYC, calculating property tax and granting eligibility to the properties that qualify for exemptions and/or abatements. It is provided by the Department of Finance on NYC Open Data, and the properties in the file were assessed in November 2010. There are 32 columns and 1,070,994 records in this dataset.

2.2 Summary Statistics

Numerical variables:

Field Name	Field Type	# unique values	# of records with value	% populated	# of records with value zero	Mean	Standard Deviation	Min	Max
LTFRONT	numerical	1297	1070994	100	169108	36.64	74.03	0	9999
LTDEPTH	numerical	1370	1070994	100	170128	88.86	76.40	0	9999
FULLVAL	numerical	109324	1070994	100	13007	874264.51	11582430.99	0	6150000000
AVLAND	numerical	70921	1070994	100	13009	85067.92	4057260.06	0	2668500000
AVTOT	numerical	112914	1070994	100	13007	227238.17	6877529.31	0	4668308947
EXLAND	numerical	33419	1070994	100	491699	36423.89	3981575.79	0	2668500000
EXTOT	numerical	64255	1070994	100	432572	91186.98	6508402.82	0	4668308947
BLDFRONT	numerical	612	1070994	100	228815	23.04	35.58	0	7575
BLDDEPTH	numerical	621	1070994	100	228853	39.92	42.71	0	9393
STORIES	numerical	112	1014730	94.75	0	5.01	8.37	1	119
AVLAND2	numerical	58592	282726	26.40	0	246235.72	6178962.56	3	2371005000
AVTOT2	numerical	111361	282732	26.40	0	713911.44	11652528.95	3	4501180002
EXLAND2	numerical	22196	87449	8.17	0	351235.68	10802212.67	1	2371005000
EXTOT2	numerical	48349	130828	12.22	0	656768.28	16072510.17	7	4501180002

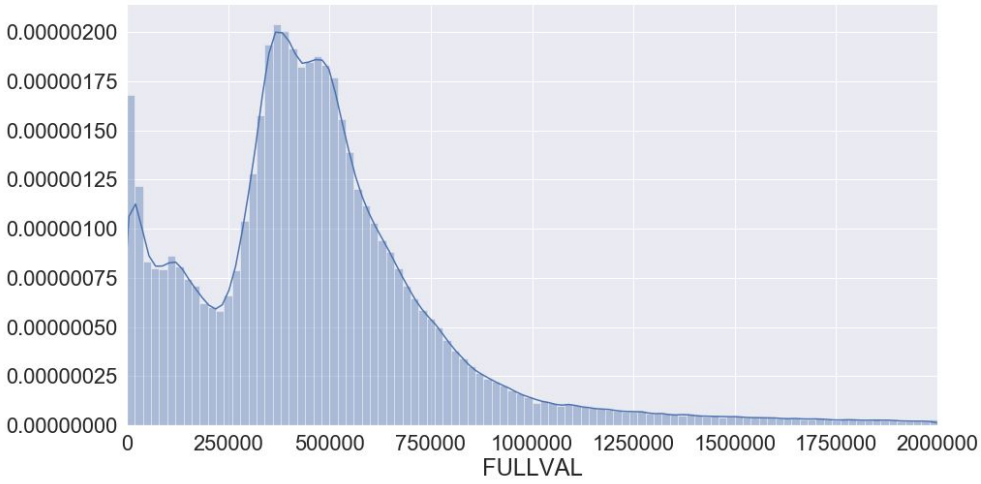
Categorical Variables:

Field Name	Field Type	# unique values	# of records with value	% populated	most common field value
RECORD	categorical	1070994	1070994	100	n/a
BBLE	categorical	1070994	1070994	100	n/a
B	categorical	5	1070994	100	4
BLOCK	categorical	13984	1070994	100	3944
LOT	categorical	6366	1070994	100	1
EASEMENT	categorical	13	4636	0.43	E
OWNER	categorical	863347	1039249	97.04	PARKCHESTER PRESERVAT
BLDGCL	categorical	200	1070994	100	R4
TAXCLASS	categorical	11	1070994	100	1
EXT	categorical	4	354305	33.08	G
EXCD1	categorical	130	638488	59.62	1017
STADDR	categorical	839281	1070318	99.94	501 SURF AVENUE
ZIP	categorical	197	1041104	97.21	10314
EXMPTCL	categorical	15	15579	1.45	X1
EXCD2	categorical	61	92948	8.68	1017
PERIOD	categorical	1	1070994	100	FINAL
YEAR	date/time	1	1070994	100	2010/11
VALTYPE	categorical	1	1070994	100	AC-TR

2.3 Variable Distributions

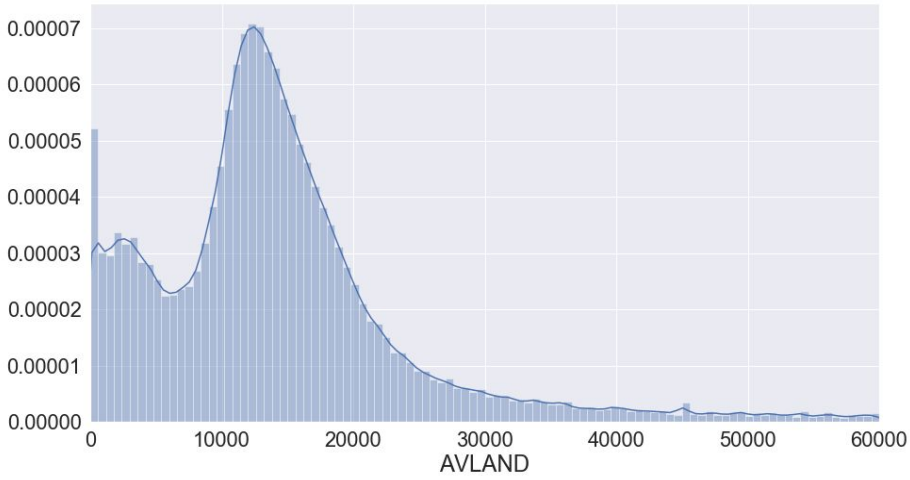
There are a total of 32 variables given in the dataset although only three of the distributions, FULLVAL, AVLAND, and AVTOT, are shown below. For each of the detailed descriptions and distributions, please refer to the data quality report in the appendix at the end.

Field Name: FULLVAL
Description: FULLVAL represents the market value of the property. There are 109,324 unique values and no missing values. There are 13,007 zero values that will be filled in with reasonable values. The mean is 874,264.5 and the standard deviation is 11,582,431. For better demonstration, we kept the data that have FULLVAL greater than 2,000,000, and the data in the histogram is 96.29% populated among the records with values.



Field Name: AVLAND

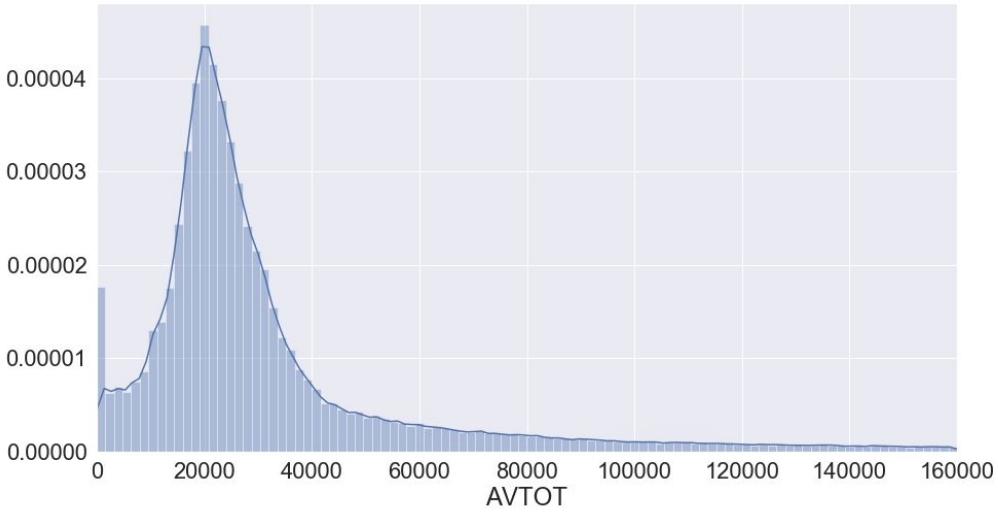
Description: AVLAND represents the actual land value of the property. There are 70,921 unique values and no missing values. There are 13,009 zero values that will be filled in with reasonable values. The mean is 85,067.9 and the standard deviation is 4,057,260. For better demonstration, we kept the data that have AVLAND greater than 60,000, and the data in the histogram is 91.61% populated among the records with values.



Field Name: AVTOT

Description: AVTOT represents the actual total value of the property. There are 112,914 unique values and no missing values. There are 13,007 zero values that will be filled in with reasonable values. The mean is 227,238.2 and the standard deviation is 6,877,529. For better demonstration,

we kept the data that have AVTOT greater than 160,000, and the data the histogram is 90.05% populated among the records with values.



3. Data Cleaning

There are nine variables we used directly from the dataset to get started on the fraud detection process: ZIP, STORIES, FULLVAL, AVLAND, AVTOT, LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH. As for data cleaning, we tried to fill in missing values and zero values in these nine variables using logical thinking and reasonable arrangements.

Field	Percentage of missing value	Percentage of 0
ZIP	2.79%	0.00%
STORIES	5.26%	0.00%
FULLVAL	0.00%	1.21%
AVLAND	0.00%	1.21%
AVTOT	0.00%	1.21%
LTFRONT	0.00%	15.79%
LTDEPTH	0.00%	15.89%
BLDFRONT	0.00%	21.36%
BLDDEPTH	0.00%	21.37%

ZIP

There are 20,890 missing values, which is around 2.80% of the total records, for this ZIP field. For the missing ZIPs in the dataset, we applied the street address field (STADDR), which contains 676 missing values, around 0.06% of total records, in Google Maps API to find the corresponding zip code. For ZIPs missing both ZIP and STADDR values or those Google Maps API could not return the zip code due to ambiguous street address, we aggregated ZIP by BLOCK and B, and used the mode value of each group to fill in the missing zip-codes.

STORIES

For the missing values in STORIES, we filtered out all the records with valid STORIES value and grouped them by TAXCLASS. The reason for this choice is due to the definition and classification of TAXCLASS. Property in NYC is divided into 4 classes:

- Class 1: Most residential property of up to three units and most condominiums that are not more than three stories.
- Class 2: All other property that is not in Class 1 and is primarily residential (rentals, cooperatives and condominiums).
    - Sub-Class 2a (4 - 6 unit rental building)
    - Sub-Class 2b (7 - 10 unit rental building)
    - Sub-Class 2c (2 - 10 unit cooperative or condominium)
    - Class 2 (11 units or more)
- Class 3: Most utility property.
- Class 4: All commercial and industrial properties and all other properties not included in tax classes 1, 2 or 3.

Therefore, we filled in the missing values with the mode, which is the most common value, in each group.

FULLVAL

FULLVAL only has zero values but no null values. However, zero values do not make sense in FULLVAL and the mean will be distorted by the zero values in FULLVAL, so all zero values have to be replaced with null values and later filled with positive values. The first step is to differentiate groups with more than 5 records from the rest after grouped by TAXCLASS and B and assign a label for those groups. For groups that have more than five records, we used the median aggregated by TAXCLASS and B to replace the zero values. For groups that have fewer than 5 records, we used the median aggregated by only TAXCLASS to replace the zero values.

AVLAND

Similar to FULLVAL, AVLAND only has zero values but no null values. However, zero values do not make sense in AVLAND, so all zero values have to be replaced and filled with positive values. We first grouped by TAXCLASS and B to see how many records are within each group.

For groups that have more than 5 records, we used the median aggregated by TAXCLASS and B to replace the zero values. For groups that have fewer than 5 records, we used the median aggregated by only TAXCLASS to replace the zero values.

AVTOT

Similar to FULLVAL, AVTOT only has zero values but no null values. However, zero values do not make sense in AVTOT, so all zero values have to be replaced and filled with positive values. We first grouped by TAXCLASS and B to see how many records are within each group. For groups that have more than 5 records, we used the median aggregated by TAXCLASS and B to replace the zero values. For groups that have fewer than 5 records, we used the median aggregated by only TAXCLASS to replace the zero values.

LTFRONT

We first separated the original data set into two groups: LTFRONT= 0 and LTFRONT ≠ 0. There are about 15.79 % of the records that have zero values, which were considered as missing values that needed to be filled. We used the median of "LTFRONT ≠ 0" group aggregated by ZIP and TAXCLASS that has records more than 5. For the group that has fewer than 5 records, we then used the median of aggregated by B & TAXCLASS that has more than 5 records. For the group that still has fewer than 5 records, we then used the median of aggregated by only TAXCLASS.

LTDEPTH

We used similar steps as LTFRONT. We first separated the original dataset into two groups: LTDEPTH=0 and LTDEPTH ≠ 0. There are about 15.89 % of records that have value 0, which we consider as missing values. We used the median of "LTDEPTH ≠ 0" group aggregated by ZIP & TAXCLASS that has records more than 5. For the group that has fewer than 5 records, we then used the median of aggregated by B and TAXCLASS that has more than 5 records. For the group that still has fewer than 5 records, we used the median aggregated by TAXCLASS.

BLDFRONT

We first separated the original dataset into two groups: BLDFRONT=0 and BLDFRONT ≠ 0. To fill in the BLDFRONT that has zero values (missing values), we used the median of "BLDRONT ≠ 0" group aggregated by ZIP and TAXCLASS that has records more than 5; if the group has fewer than 5 records, we then used the median of aggregated by B and TAXCLASS that has records more than 5; if the group has fewer than 5 records, we then used the median of aggregated by TAXCLASS.

BLDDEPTH

We then separated the original dataset into two groups: BLDDEPTH=0 and BLDDEPTH ≠ 0. To fill in the BLDDEPTH that has zero values (missing values), we used the median of "BLDDEPTH ≠ 0" group aggregated by ZIP and TAXCLASS that has records more than 5; if the group has fewer than 5 records, we then used the non-zero median of aggregated by B and TAXCLASS that has records more than 5; if the group has fewer than 5 records, we then used the median of aggregated by TAXCLASS.

4. Variable Creation

For the model building process to detect potential property frauds in the dataset, we created 45 new variables with the 11 variables in the original dataset, including LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, STORIES, ZIP, TAXCLASS, B, FULLVAL, AVLAND, AVTOT.

The variables that we created were all different transformations of three important variables as the following:

V_1 = FULLVAL – the full value of the building
V_2 = AVLAND – the assessed value of land
V_3 = AVTOT – the assessed value of property

The steps for creating these 45 variables are as follows:

1. Create three sizes for each property
   - V_4 = LOTAREA = LTFRONT * LTDEPTH
   - V_5 = BLDAREA = BLDFRONT * BLDDEPTH
   - V_6 = BLDVOL = BLDAREA* STORIES

2. Create 9 variables by normalizing FULLVALUE, AVTOT, AVLAND, respectively, by the LOTAREA, BLDAREA and BLDVOL for each record

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

3. Group above nine ratios by 5 criteria (ZIP5, ZIP3, TAXCLASS, B, ALL), in order to compare the ratios with properties under similar condition. The meanings of the 5 criteria are as follows:
    a. ZIP5 – zip code
    b. ZIP3 – 3-digit zip code
    c. TAXCLASS – tax class of the property
    d. B – borough
    e. ALL – grouped by ZIP, TAXCLASS, and B.

4. Obtain the averages of the nine ratios for each group in step 4;
5. Divide each of the nine ratios by the mean ratio value in that group based on ZIP5, ZIP3, TAXCLASS, B and ALL.

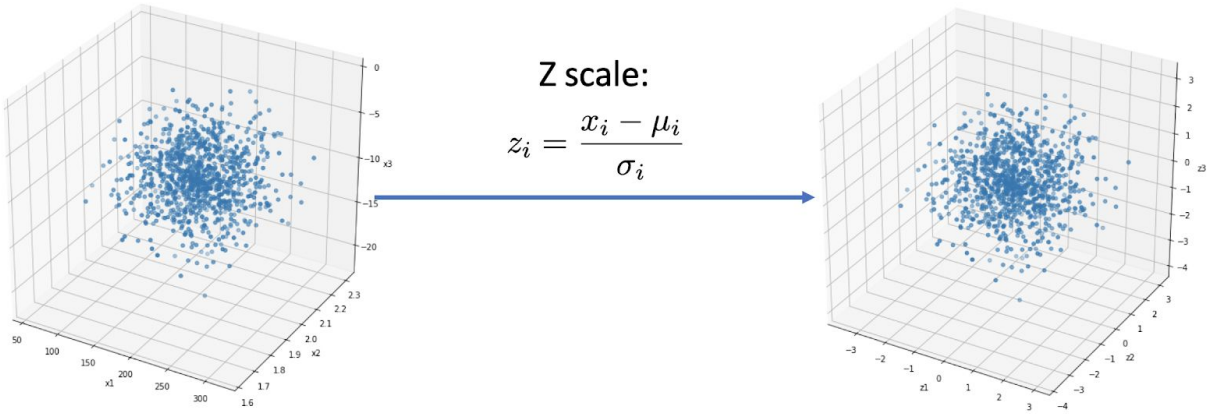
For each property record, we ended up with 45 variables which would be used in calculating the fraud score for each property later in the process. Since we expected similar ratios of the property value to property area for properties in a given location and building type, we performed a Principal Components Analysis (PCA) for these 45 variables to reduce dimensionality as the following step.

5. **Dimensionality Reduction**

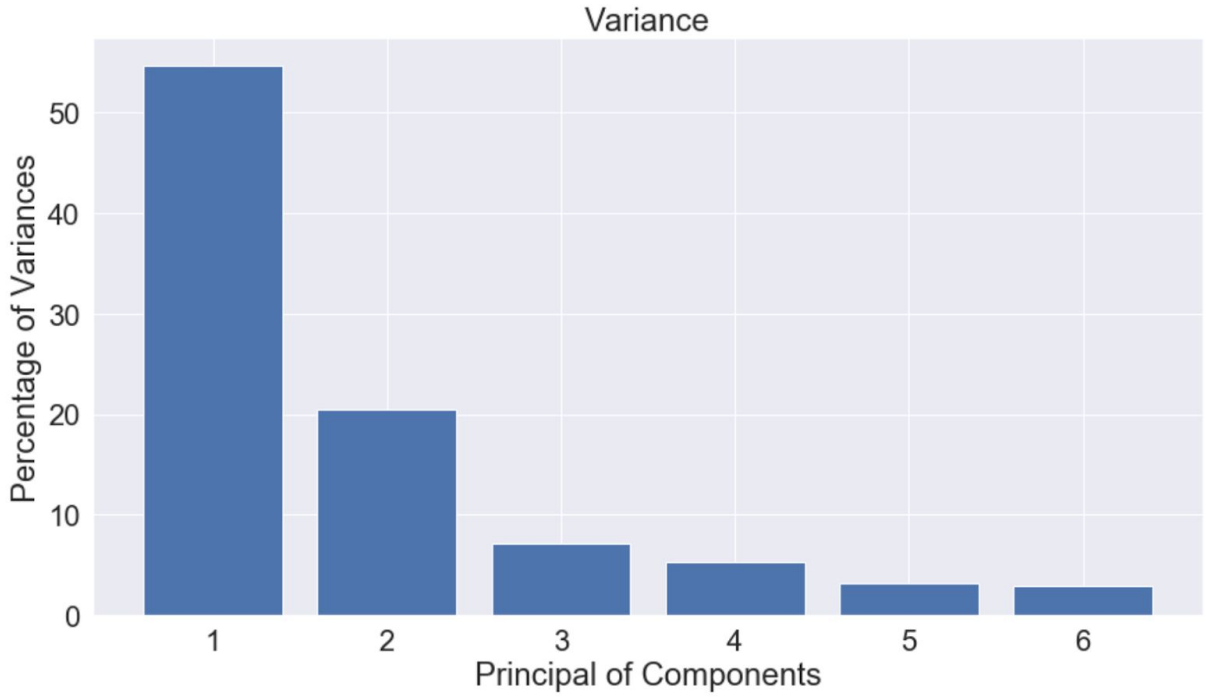
After filling in the missing values for the original fields in the dataset and creating new variables, we had 45 fields to begin with for the models. However, there might be some collinearity among these variables. We decided to reduce the dimensionality by performing the Principal Components Analysis (PCA) on the new dataset.

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables. It extracts variables that are linearly uncorrelated and combine them into a set of values of variables which is called principal components. The first principal component accounts for as much of the variability in the data as possible. And for each succeeding component, it accounts for as much of the remaining variability in the data as possible. PCA is a dimensionality reduction or data compress method. The goal of PCA is to reduce dimensionality, simply put, from a larger number of variables to a smaller number of non-dependent variables.

However, before we performed the PCA, we needed to perform the Z scale first. It is important to normalize the data because the original predictors can be on a different scale and can contribute significantly differently towards the variance, which has an impact on PCA. Below is how we do the normalization by Z scale. After we scaled the data, we can measure the "outlierness" of a record by looking at its distance from the origin.



In this project, we used the PCA function from sklearn.decomposition package. We put six as the number of our PCA output dimensions. These six dimensions consist of 91.62% of the total variance.



6. **Algorithms**

Method 1: Heuristic Function of z-scores

The assumption of method 1 is advantage of using heuristic function of z-scores
Steps of using z-scores to derive the fraud score:

1. Second z-scaling makes each PCA component equally important.
2. Square and sum up each record.
3. Compute the square root of the summations and rank the score descendingly.

The formula for Method 1 is shown as below:

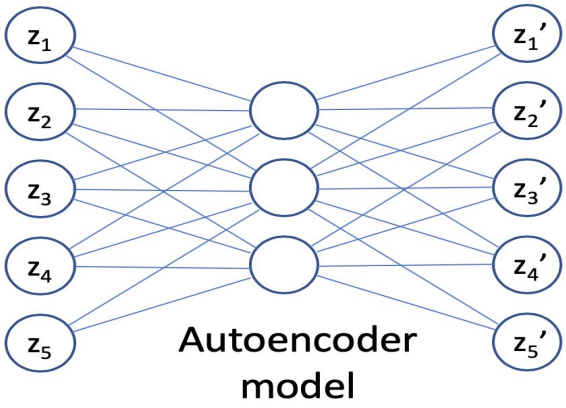
$$s_i = \left(\sum_k |z_k^i|^n \right)^{1/n}, \quad n \text{ anything}$$

An Euclidean score (n = 2) is chosen to use the square underline the anomaly from other records. When n = 1, abnormality is not as prominent as of n = 2. We are also interested in the distribution of the scores for Manhattan score (n = 3) and a similar distribution was derived.

Method 1 generates a score for each record, which is represented as s1 later in the report.

Method 2: Autoencoder

Used autoencoder as a model trained to output the original vector input. If the record has no anomaly, then the new output should be similar to the original one. Otherwise, it indicates there is something unusual.



Steps of using autoencoder algorithm to get the fraud score:

1. Train an autoencoder on the entire data set by using scaled PCA.
2. The model will learn as best as possible to reproduce the data records.
3. The reproduced records that are very different from the original records are unusual.
4. Measure the reproduction error as the formula below (here we use n=2):

$$s_i = \left(\sum_k |z'^i_k - z^i_k|^n \right)^{1/n}, \quad n \text{ anything}$$

Method 2 also generates a score for each record, which is represented as s2 later in the report.

Combining Two Scores: Quartile Binning

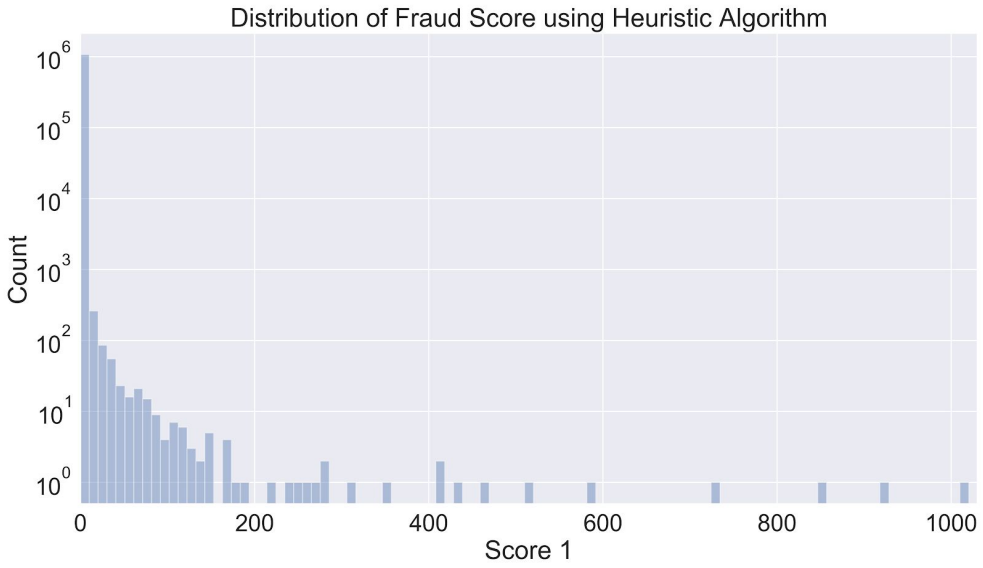
Now we have generated two fraud scores from two different methods. We decided to combine them together to get a more accurate final score. Since the original two fraud scores are on different scales, it is critical for us to first re-scale them. Here we used the quantile binning method. We replaced the score with the record's rank order after sorting by the score. Then we take the average of these two ranks to get the final score.

7. **Results**

7.1 Score Distributions

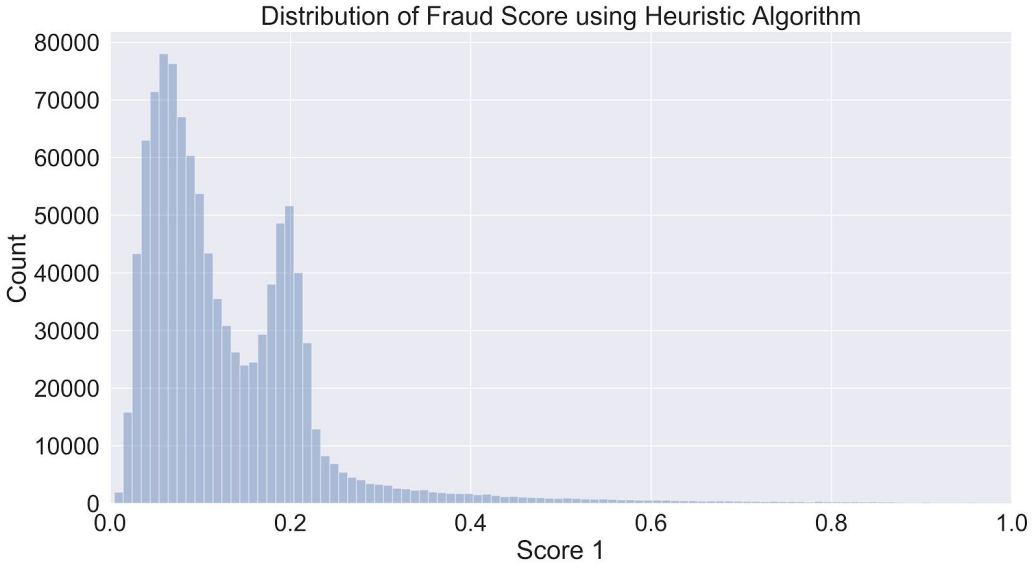
The following graphs show the distributions for s1, s2 and combined final score.

Distribution of s1

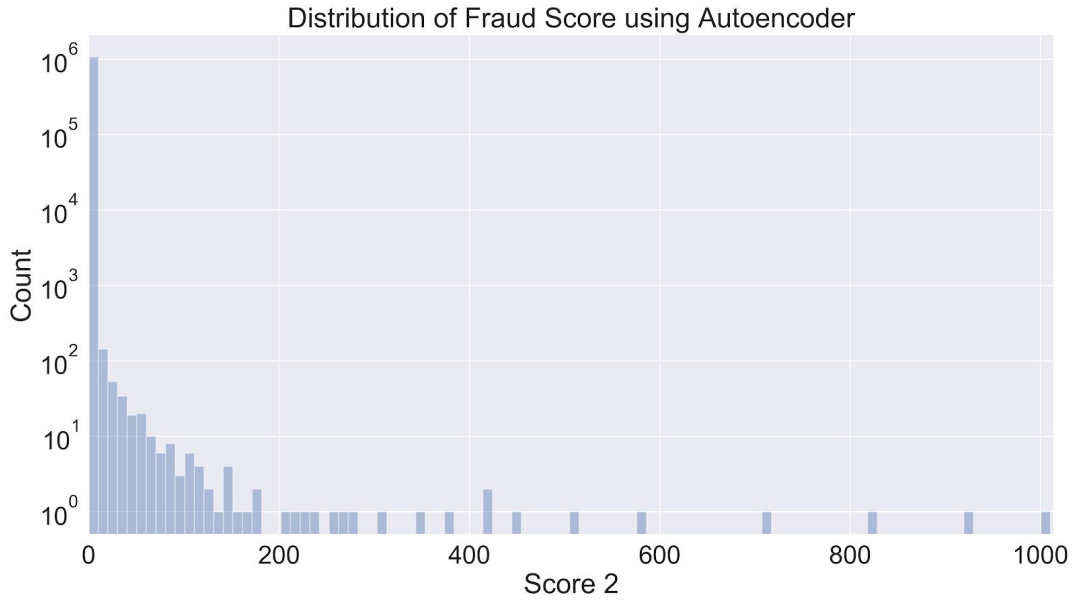


We found that more than 98.5% of s1 values are smaller than 1, so here we also drew the distribution of s1 when we set x limit between 0 and 1:

Distribution of s1 with xlim(0,1)

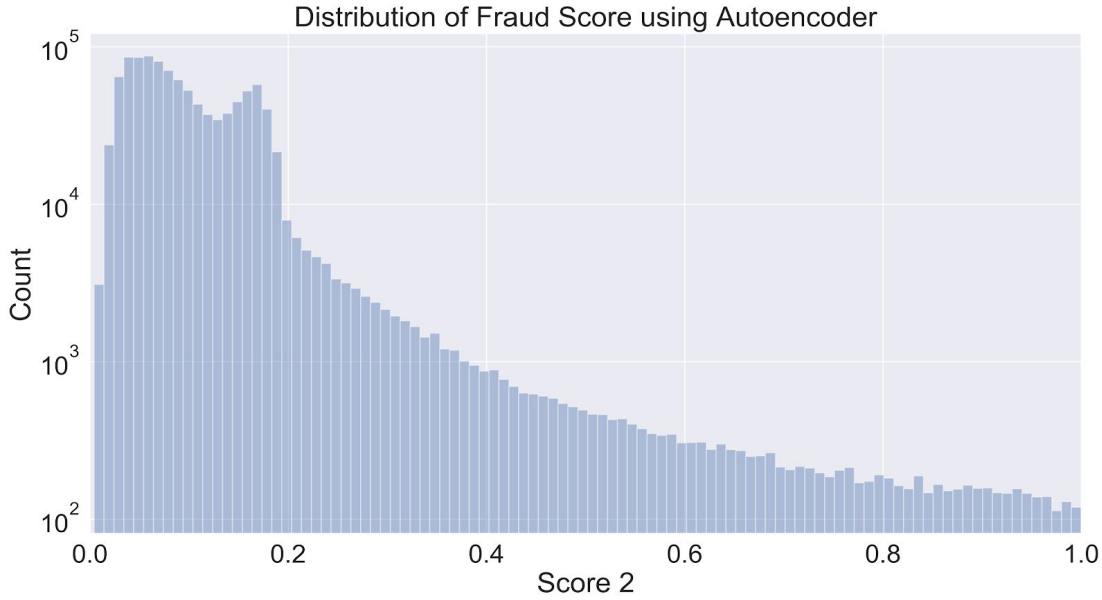


Distribution of s2

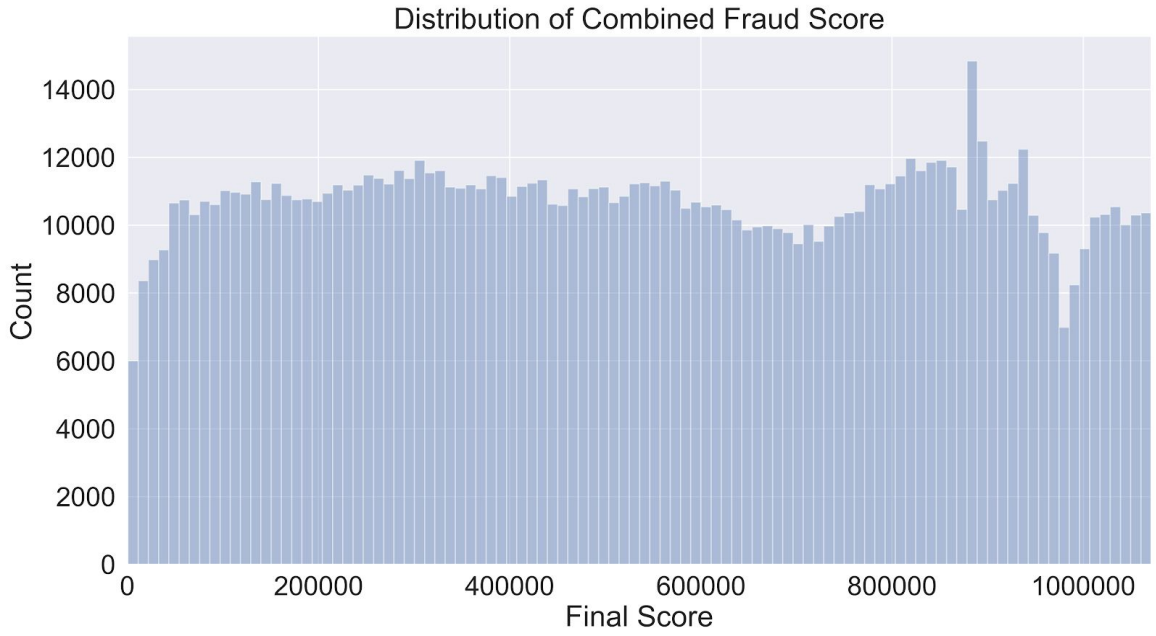


We found that more than 99% of s2 values are smaller than 1, so here we also drew the distribution of s2 when we set x limit between 0 and 1:

Distribution of s2 with xlim (0,1)



Distribution of Combined Score



After completing the above steps, a list of top ten records with the highest fraud score is generated. The records are 632816, 565392, 1067360, 917942, 585118, 585439, 85886, 585120, 67129, 565398, ranked by the combined score starting from the highest.

7.2 Top Ten Properties: Research & Explanations

Index	Record	Rank1	Rank2	Weighted Rank
0	632816	1070994	1070994	1070994.0
1	565392	1070993	1070993	1070993.0
2	1067360	1070992	1070992	1070992.0
3	917942	1070991	1070991	1070991.0
4	585118	1070990	1070990	1070990.0
5	585439	1070989	1070989	1070989.0
6	85886	1070988	1070988	1070988.0
7	67129	1070987	1070987	1070987.0
8	585120	1070986	1070986	1070986.0
9	565398	1070985	1070985	1070985.0

1. **RECORD: 632816**

Original data record

RECORD	632816	**AVLAND**	1.32E+06
BBLE	4018420001	**AVTOT**	1.32E+06
B	4	**EXLAND**	0
BLOCK	1842	**EXTOT**	0
LOT	1	**EXCD1**	NaN
EASEMENT	NaN	**STADDR**	86-55 BROADWAY
OWNER	864163 REALTY, LLC	**ZIP**	11373
BLDGCL	D9	**EXMPTCL**	NaN
TAXCLASS	2	**BLDFRONT**	1
LTFRONT	157	**BLDDEPTH**	1
LTDEPTH	95	**AVLAND2**	NaN
EXT	NaN	**AVTOT2**	NaN
STORIES	1	**EXLAND2**	NaN
FULLVAL	2.93E+06	**EXTOT2**	NaN
		EXCD2	NaN

When we looked at this address on the map, we found it is a five-stories luxury apartment. According to the information above, it does not make sense to have both BLDFRONT and BLDDEPTH equal to 1, while LTFRONT and LTDEPTH equal to 157 and 95 respectively. Therefore, when we used BLDFRONT and BLDDEPTH to calculate s2, we would get an s2 that equal to 1. As a result, r2, r5, r8 would be much larger than it should be. The value of s3 is also inaccurate since it equals to s2*stories, which renders r3, r6, r9 relatively large. Therefore, record 632816 is suspicious and we suggest further investigation.

2. **RECORD: 565392**

Original data record

RECORD	565392	**AVLAND**	1.95E+09
BBLE	3085900700	**AVTOT**	1.95E+09
B	3	**EXLAND**	1946836665
BLOCK	8590	**EXTOT**	1946836665
LOT	700	**EXCD1**	2231
EASEMENT	NaN	**STADDR**	FLATBUSH AVENUE
OWNER	U S GOVERNMENT OWNRD	**ZIP**	NaN
BLDGCL	V9	**EXMPTCL**	X1
TAXCLASS	4	**BLDFRONT**	NaN
LTFRONT	117	**BLDDEPTH**	NaN
LTDEPTH	108	**AVLAND2**	8.48E+08
EXT	NaN	**AVTOT2**	8.48E+08
STORIES	NaN	**EXLAND2**	8.48E+08
FULLVAL	4.33E+09	**EXTOT2**	8.48E+08
		EXCD2	NaN

Filled-in data record

RECORD	565392	**AVLAND**	1.95E+09
BBLE	3085900700	**AVTOT**	1.95E+09
B	3	**EXLAND**	1.95E+09
BLOCK	8590	**EXTOT**	1.95E+09
LOT	700	**EXCD1**	2231
EASEMENT	NaN	**STADDR**	FLATBUSH AVENUE
OWNER	U S GOVERNMENT OWNRD	**ZIP**	11234
BLDGCL	V9	**EXMPTCL**	X1
TAXCLASS	4	**BLDFRONT**	25
LTFRONT	117	**BLDDEPTH**	59.5
LTDEPTH	108	**AVLAND2**	8.48E+08
EXT	NaN	**AVTOT2**	8.48E+08
STORIES	1	**EXLAND2**	8.48E+08
FULLVAL	4.33E+09	**EXTOT2**	8.48E+08
		EXCD2	NaN

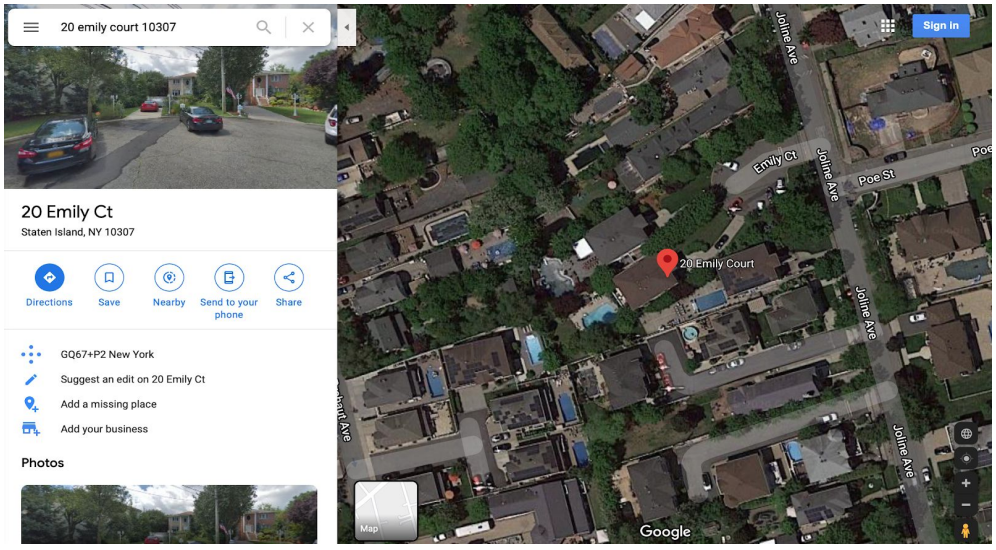
The property with a record number of 565392 is a US Government-Owned Property with an extremely high market value (third highest market value in this dataset) and low LTFRONT and LTDEPTH. There are also no Stories, BLDFRONT, and BLDDEPTH and we filled in the missing value using the median of other available data. In our created values, since we filled in stories with value 1, which drove up the r6 [AVLAND/(BLDFRONT*BLDDEPTH*STORIES)] and caused the fraud score to be extremely high. We assume that this is an empty lot that has not been used for construction, thus, this property with a high anomaly score may not necessarily refer to fraud.

3. RECORD: 1067360

Original Data Record

RECORD	1067360	**AVLAND**	2.88E+04
BBLE	5078530085	**AVTOT**	5.02E+04
B	5	**EXLAND**	0
BLOCK	7853	**EXTOT**	0
LOT	85	**EXCD1**	NaN
EASEMENT	NaN	**STADDR**	20 EMILY COURT
OWNER	NaN	**ZIP**	10307
BLDGCL	B2	**EXMPTCL**	NaN
TAXCLASS	1	**BLDFRONT**	36
LTFRONT	1	**BLDDEPTH**	45
LTDEPTH	1	**AVLAND2**	NaN
EXT	NaN	**AVTOT2**	NaN
STORIES	2	**EXLAND2**	NaN
FULLVAL	8.36E+05	**EXTOT2**	NaN
		EXCD2	NaN

Google Map



For record number 1067360, it is a residential building with 2 stories. And the information matches with the numbers on Zillow for this property (https://www.zillow.com/homedetails/20-Emily-Ct-Staten-Island-NY-10307/58579273_zpid/). The value of LTFRONT and LTDEPTH are both 1, which makes the denominator smaller and causes a higher variables score ($r_1 - r_3$) and a high fraud score. In addition, the value of AVTOT is unusually much lower than FULLVAL. As the trading information on Zillow, this property was sold at $234,000 in 1999. The value of this property is much higher than $28,800 (AVTOT). As for a residential property, the value of the frontage lot does not seem reasonable. We think this is the main reason that caused the high fraud score. Therefore, record 1067360 is suspicious and we suggest a further investigation.

4. RECORD: 917942

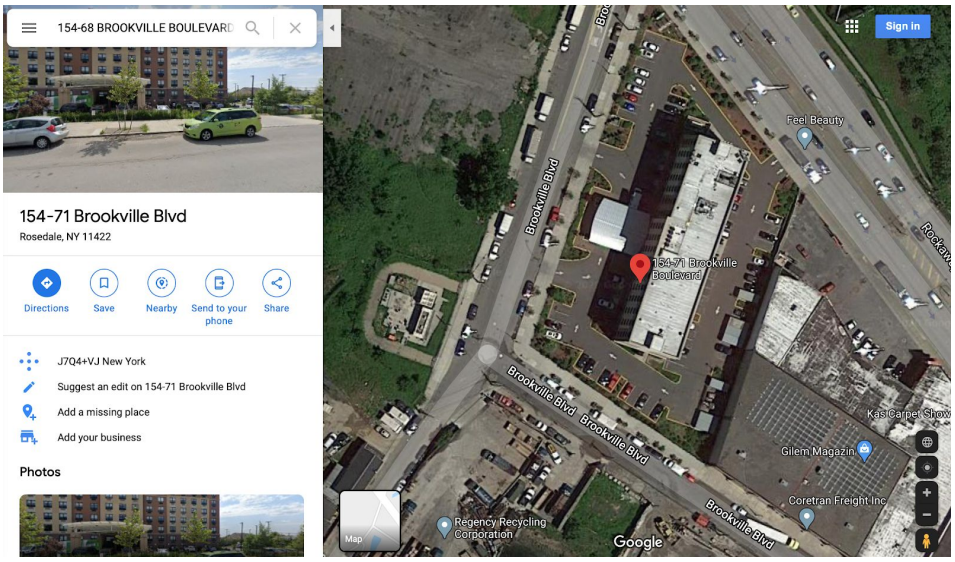
Original Data Record

RECORD	917942	**AVLAND**	1.79E+09
BBLE	4142600001	**AVTOT**	4.67E+09
B	4	**EXLAND**	1.79E+09
BLOCK	14260	**EXTOT**	4.67E+09
LOT	1	**EXCD1**	2198
EASEMENT	NaN	**STADDR**	154-68 BROOKVILLE BOULEVARD
OWNER	LOGAN PROPERTY, INC.	**ZIP**	11422
BLDGCL	T1	**EXMPTCL**	X4
TAXCLASS	4	**BLDFRONT**	0
LTFRONT	4910	**BLDDEPTH**	0
LTDEPTH	0	**AVLAND2**	1.64E+09
EXT	NaN	**AVTOT2**	4.50E+09
STORIES	3	**EXLAND2**	1.64E+09
FULLVAL	3.74E+08	**EXTOT2**	4.50E+09
		EXCD2	NaN

Filled-in Data Record

RECORD	917942	**AVLAND**	1.79E+09
BBLE	4142600001	**AVTOT**	4.67E+09
B	4	**EXLAND**	1.79E+09
BLOCK	14260	**EXTOT**	4.67E+09
LOT	1	**EXCD1**	2198
EASEMENT	NaN	**STADDR**	154-68 BROOKVILLE BOULEVARD
OWNER	LOGAN PROPERTY, INC.	**ZIP**	11422
BLDGCL	T1	**EXMPTCL**	X4
TAXCLASS	4	**BLDFRONT**	40
LTFRONT	4910	**BLDDEPTH**	40
LTDEPTH	100	**AVLAND2**	1.64E+09
EXT	NaN	**AVTOT2**	4.50E+09
STORIES	3	**EXLAND2**	1.64E+09
FULLVAL	3.74E+08	**EXTOT2**	4.50E+09
		EXCD2	NaN

Google Map



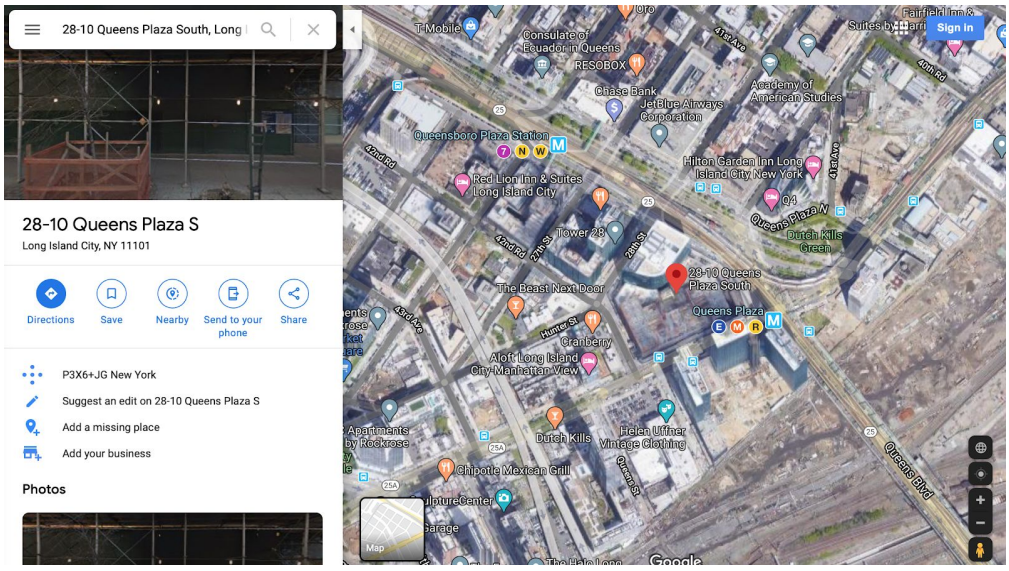
For record number 91742, as in profile, is an independent building with 3 stories, owned by LOGAN PROPERTY, INC. In the original data, three crucial variables (LTDEPTH, BLDFRONT, BLDDEPTH) contain value 0. After filling in the value according to the method described above, LTDEPTH = 100 feet, BLDFRONT = 40 feet, BLDDEPTH = 40 feet, which is extremely low compared with LTFRONT value. This low number would cause the relatively high value to all of our 9 variables ($r_1 - r_9$) since the denominators are all involved with LTDEPTH, BLDFRONT, BLDDEPTH, and the numerators are the full value and other values of this property which are really high numbers. This would make the fraud score unusually high. The Google Map also shows there are 6 floors in this building which doesn't match with the record in our profile. Therefore, record 91742 is suspicious and requires further investigation.

5. RECORD: 585118

Original data record

RECORD	585118	**AVLAND**	1.55E+06
BBLE	4004200001	**AVTOT**	1.55E+06
B	4	**EXLAND**	0
BLOCK	420	**EXTOT**	0
LOT	1	**EXCD1**	NaN
EASEMENT	NaN	**STADDR**	28-10 QUEENS PLAZA SOUTH
OWNER	NEW YORK CITY ECONOMI	**ZIP**	11101
BLDGCL	O3	**EXMPTCL**	X1
TAXCLASS	4	**BLDFRONT**	1
LTFRONT	298	**BLDDEPTH**	1
LTDEPTH	402	**AVLAND2**	1.59E+06
EXT	NaN	**AVTOT2**	1.59E+06
STORIES	20	**EXLAND2**	NaN
FULLVAL	3.44E+06	**EXTOT2**	NaN
		EXCD2	NaN

Google Map



For record number 685118, there isn't any zero or missing value that we need to fill in. According to our profile, this property is a building with 20 stories, owned by NEW YORK CITY ECONOMI. As we checked the satellite through Google Map, it is an independent building located in Queens Plaza. Compared with our data, the values of LTFRONT and LTDEPTH are within a higher range but reasonable (Refer to LTFRONT and LTDEPTH distribution in DQR in appendix). However, the values of BLDFRONT, BLDDEPTH are 1, which are not reasonable for such a huge building. These low values will cause some ($r_2 = \frac{V_1}{S_2}$, $r_5 = \frac{V_2}{S_2}, r_8 = \frac{V_3}{S_2}, r_3, r_6, r_9$, $S_2 = BLDFRONT * BLDDEPTH$) of our 9 variables to be unusually high, which might be the reason for ramping up to a high combined score. In addition,

we also found some information against our profile. According to information from StreetEasy website, this building was built in 2017 with 27 stories (https://streeteasy.com/building/28_10-queens-plaza-south-long_island_city). Therefore, based on the investigation, record 585118 is suspicious and requires further investigation.

6. **RECORD: 585439**

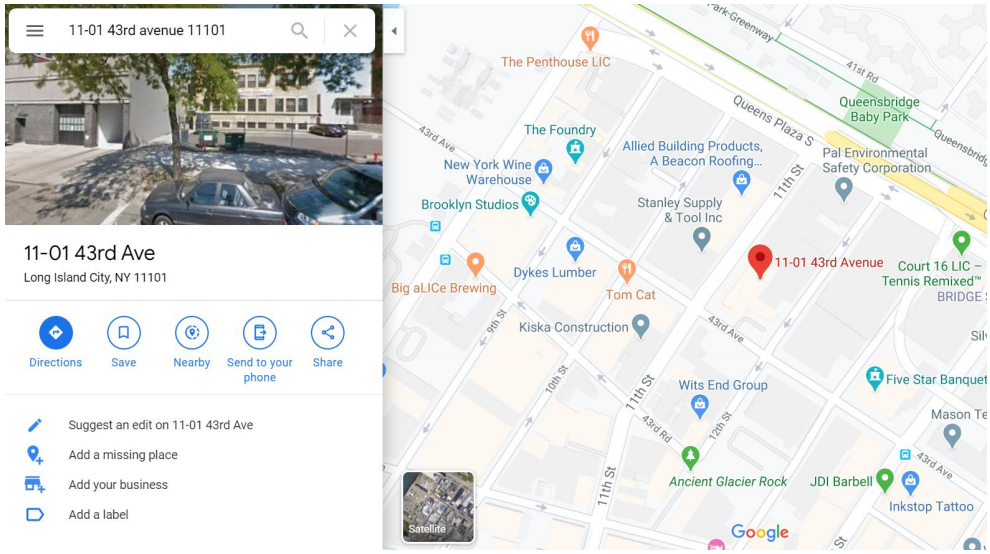
Original Data Record

RECORD	585439	**AVLAND**	2.52E+05
BBLE	4004590005	**AVTOT**	1.67E+06
B	4	**EXLAND**	0
BLOCK	459	**EXTOT**	1418400
LOT	5	**EXCD1**	1986
EASEMENT		**STADDR**	11-01 43 AVENUE
OWNER	11-01 43RD AVENUE REA	**ZIP**	11101
BLDGCL	H9	**EXMPTCL**	
TAXCLASS	4	**BLDFRONT**	1
LTFRONT	94	**BLDDEPTH**	1
LTDEPTH	165	**AVLAND2**	
EXT		**AVTOT2**	
STORIES	10	**EXLAND2**	
FULLVAL	3.71E+06	**EXTOT2**	
		EXCD2	

Filled-in Data Record

RECORD	585439	**EXTOT**	1418400
BBLE	4004590005	**EXCD1**	1986
B	4	**STADDR**	11-01 43 AVENUE
BLOCK	459	**ZIP**	11101
LOT	5	**AVLAND**	252000
EASEMENT		**AVTOT**	1670400
OWNER	11-01 43RD AVENUE REA	**FULLVAL**	3712000
BLDGCL	H9	**LTFRONT**	94
TAXCLASS	4	**LTDEPTH**	165
EXT		**BLDFRONT**	1
STORIES	10	**BLDDEPTH**	1
EXLAND	0		

Google Map



This record is categorized as "abnormal" and has a high combined total score mainly because the values of BLDFRONT and BLDDEPTH equal to 1. Large r2, r3, r5, r6, r8 and r9 are derived from these two small denominators.

Further investigation reveals that this property is a hotel built in 2011, the year the data was collected from the merchant. Looking into other records with BLDFRONT and BLDDEPTH equal to 1 from the top 10 list, we discovered that these properties were all constructed in or after 2011, suggesting that the BLDFRONT and BLDDEPTH were left as 1 because of uncertainty rather than potential fraud.

7. RECORD: 85886

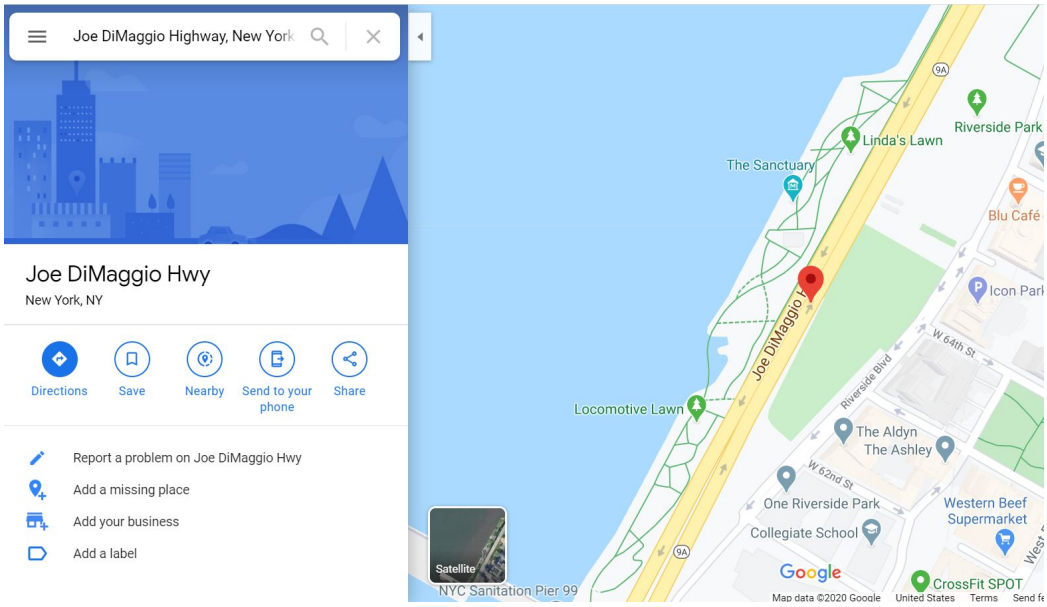
Original Data Record

RECORD	85886	**AVLAND**	3.15E+07
BBLE	1012540010	**AVTOT**	3.16E+07
B	1	**EXLAND**	31455000
BLOCK	1254	**EXTOT**	31596300
LOT	10	**EXCD1**	2231
EASEMENT		**STADDR**	JOE DIMAGGIO HIGHWAY
OWNER	PARKS AND RECREATION	**ZIP**	
BLDGCL	Q1	**EXMPTCL**	X1
TAXCLASS	4	**BLDFRONT**	8
LTFRONT	4000	**BLDDEPTH**	8
LTDEPTH	150	**AVLAND2**	2.81E+07
EXT		**AVTOT2**	2.83E+07
STORIES	1	**EXLAND2**	28134000
FULLVAL	7.02E+07	**EXTOT2**	28260180
		EXCD2	

Filled-in Data Record

RECORD	85886	**EXMPTCL**	X1
BBLE	1012540010	**AVLAND2**	28134000
B	1	**AVTOT2**	28260180
BLOCK	1254	**EXLAND2**	28134000
LOT	10	**EXTOT2**	28260180
OWNER	PARKS AND RECREATION	**EXCD2**	
BLDGCL	Q1	**AVLAND**	31455000
TAXCLASS	4	**AVTOT**	31596300
EXT		**FULLVAL**	70214000
STORIES	1	**LTFRONT**	4000
EXLAND	31455000	**LTDEPTH**	150
EXTOT	31596300	**BLDFRONT**	8
EXCD1	2231	**BLDDEPTH**	8
STADDR	JOE DIMAGGIO HIGHWAY		

Google Map



This property has unusually large AVLAND and AVTOT. Together with a relatively low BLDFRONT and BLDDEPTH, the r2 to r9 are distorted. The STADDR is not informative as Joe DiMaggio Hwy is a 5.2-mile highway and no further information can be traced from the owner, the NYC Park and Recreation Department. Since the property is owned by a government agency, the likelihood of fraud is small.

8. RECORD: 67129

Original data record

RECORD	67129	**AVLAND**	2.67E+09
BBLE	1011110001	**AVTOT**	2.77E+09
B	1	**EXLAND**	2668500000
BLOCK	1111	**EXTOT**	2767500000
LOT	1	**EXCD1**	2231
EASEMENT		**STADDR**	1000 5 AVENUE
OWNER	CULTURAL AFFAIRS	**ZIP**	10028
BLDGCL	Q1	**EXMPTCL**	X1
TAXCLASS	4	**BLDFRONT**	0
LTFRONT	840	**BLDDEPTH**	0
LTDEPTH	0	**AVLAND2**	2.37E+09
EXT	E	**AVTOT2**	2.47E+09
STORIES		**EXLAND2**	2.37E+09
FULLVAL	6.15E+09	**EXTOT2**	2.47E+09
		EXCD2	

Filled-in data record

RECORD	67129	**AVLAND**	2.67E+09
BBLE	1011110001	**AVTOT**	2.77E+09
B	1	**EXLAND**	2668500000
BLOCK	1111	**EXTOT**	2767500000
LOT	1	**EXCD1**	2231
EASEMENT		**STADDR**	1000 5 AVENUE
OWNER	CULTURAL AFFAIRS	**ZIP**	10028
BLDGCL	Q1	**EXMPTCL**	X1
TAXCLASS	4	**BLDFRONT**	38.5
LTFRONT	840	**BLDDEPTH**	96
LTDEPTH	102	**AVLAND2**	2.37E+09
EXT	E	**AVTOT2**	2.47E+09
STORIES	1	**EXLAND2**	2.37E+09
FULLVAL	6.15E+09	**EXTOT2**	2.47E+09
		EXCD2	

For record number 67129, it has 4 missing fields: STORIES, LTDEPTH, BLDFRONT and BLDDEPTH. The S_1, S_2 and S_3 are relatively small since this property has a large LTFRONT.

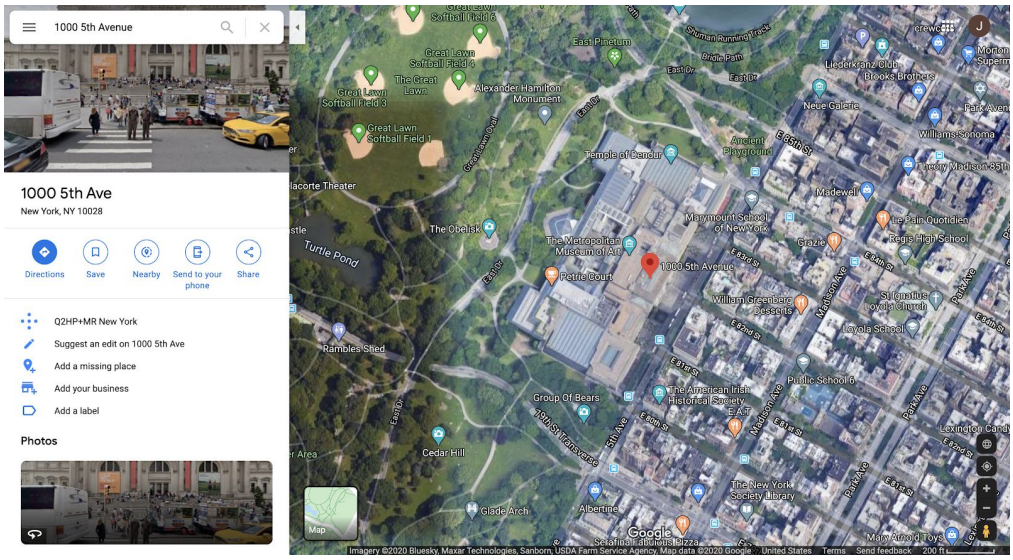
Furthermore, the FULLVAL, AVLAND and AVTOT are way higher than the average, thus r_1 to r_9 are extremely high.

Taken r1 as an example to illustrate the extent to which this record is comprised of 45 unusually high variables,

Name	Median	Record 67129
R1 Scaled by All	0.714	337.965
R1 Scaled by ZIP	0.920	123.724
R1 Scaled by TAXCLASS	0.672	304.790
R1 Scaled by B	0.849	191.679
R1 Scaled by ZIP3	0.920	123.724

According to Google Satellite, this property is the Metropolitan Museum of Art, a famous public building in New York. Further research is needed to uncover the reason for the abnormal record.

Google Map



9. RECORD: 585120

Original data record

RECORD	585120	**AVLAND**	9.68E+05
BBLE	4004200101	**AVTOT**	9.68E+05
B	4	**EXLAND**	0
BLOCK	420	**EXTOT**	0
LOT	101	**EXCD1**	NaN
EASEMENT	NaN	**STADDR**	28 STREET
OWNER	NaN	**ZIP**	10016
BLDGCL	O3	**EXMPTCL**	NaN
TAXCLASS	4	**BLDFRONT**	1
LTFRONT	139	**BLDDEPTH**	1
LTDEPTH	342	**AVLAND2**	9.75E+05
EXT	NaN	**AVTOT2**	9.75E+05
STORIES	20	**EXLAND2**	NaN
FULLVAL	2.15E+06	**EXTOT2**	NaN
		EXCD2	NaN

For record number 585120, there isn't any missing value or value 0 that we needed to fill in to create expert variables. Based on the dataset, the street address for this property is 28 STREET. This seems to be an incomplete address missing either house number of street name. Therefore, it is hard for us to identify which property this record refers to exactly. From the data in record, the values for BLDFRONT and BLDDEPTH are both 1, but the value for STORIES is 20. These values seem to contradict each other since a 20-story-building should not have such small building front or building depth. Therefore, we believe this record is considered unusual since it has a vague street address and contradictory values for data available.

10. RECORD: 565398

Original data record

RECORD	565398	**AVLAND**	1.04E+09
BBLE	3085910100	**AVTOT**	1.04E+09
B	3	**EXLAND**	1039898000
BLOCK	8591	**EXTOT**	1039898000
LOT	100	**EXCD1**	2191
EASEMENT		**STADDR**	FLATBUSH AVENUE
OWNER	DEPT OF GENERAL SERVI	**ZIP**	
BLDGCL	V9	**EXMPTCL**	X1
TAXCLASS	4	**BLDFRONT**	0
LTFRONT	466	**BLDDEPTH**	0
LTDEPTH	1009	**AVLAND2**	4.35E+08
EXT		**AVTOT2**	4.35E+08
STORIES		**EXLAND2**	4.35E+08
FULLVAL	2.31E+09	**EXTOT2**	4.35E+08
		EXCD2	

Filled-in data record

RECORD	565398	**AVLAND**	1.04E+09
BBLE	3085910100	**AVTOT**	1.04E+09
B	3	**EXLAND**	1039898000
BLOCK	8591	**EXTOT**	1039898000
LOT	100	**EXCD1**	2191
EASEMENT		**STADDR**	FLATBUSH AVENUE
OWNER	DEPT OF GENERAL SERVI	**ZIP**	11234
BLDGCL	V9	**EXMPTCL**	X1
TAXCLASS	4	**BLDFRONT**	25
LTFRONT	466	**BLDDEPTH**	59.5
LTDEPTH	1009	**AVLAND2**	4.35E+08
EXT		**AVTOT2**	4.35E+08
STORIES	1	**EXLAND2**	4.35E+08
FULLVAL	2.31E+09	**EXTOT2**	4.35E+08
		EXCD2	

For Record 565398, it is similar to the second record in the top 10 records, Record 565392. These two records have the same address on Flatbush Avenue, which is too ambiguous to locate the property address. The property is owned by the department of general service. The reason that this record has a high fraud score is also the same as Record 565392. Its FULLVAL, AVLAND and AVLAND are extremely high, and the filled-in values are relatively small, causing extremely high values for r1 to r9 variables. Overall, we considered this property most likely to be a property owned by a government agency, so it is unlikely to be a potential fraud.

8. Conclusion

In conclusion, the objective of this project is to select out anomaly records in the NY Property file and possible fraudulent ones among the anomalies. Our approach to frame this problem is to use two different models, heuristic algorithm and autoencoder, to obtain two different rankings for most potential fraudulent records and combine two rankings to get the final results. In order to do so, we first looked carefully at the data and generated a data quality report to validate the data. Then we determined how we wanted to fill in missing values/zero values with a logical approach for each field in the dataset. With all values filled, we then created 45 other variables, along with the original 9 variables we selected to use for our models. We utilized z-scaling and Principal Components Analysis to obtain our first anomaly rank and autoencoder for the second rank. Our approach helped us select the top ten properties that are most abnormal among all properties in the data set. We researched these ten properties to see if there are reasonable explanations behind their anomaly. With a conservative stand, we found that there are six out of the ten properties that we may want to look deeper into for potential property tax frauds.

If we had more time to dive deeper into this project, we would try out different methods for data cleaning and combination of aggregations for filling in missing/zero values. We noticed that using different methods for filling in missing/zero values could cause a slightly different final result. As for the second method, autoencoder, we would like to try different Python packages as we noticed that there are different codes for autoencoders. We would also like to consult with a domain expert if possible to see what variables and methods make the most sense in compliance with the industrial knowledge to develop a more realistic approach and final model at the end.

9. Appendix

1. Description of the data:

The dataset is the property valuation and assessment data of New York City. Data represent NYC properties assessments for purpose to calculate Property Tax, Grant eligible properties Exemptions and/or Abatements. It is provided by the department of Finance on NYC Open Data and covers only in the time period of November 2010. There are 32 columns and 1,070,994 records.

2. Summary Tables

2.1 Numerical Value

Field Name	Field Type	# unique values	# of records with value	% populated	# of records with value zero	Mean	Standard Deviation	Min	Max
LTFRONT	numerical	1297	1070994	100.00	169108	36.64	74.03	0	9999
LTDEPTH	numerical	1370	1070994	100.00	170128	88.86	76.40	0	9999
FULLVAL	numerical	109324	1070994	100.00	13007	874264.51	11582430.99	0	6150000000
AVLAND	numerical	70921	1070994	100.00	13009	85067.92	4057260.06	0	2668500000
AVTOT	numerical	112914	1070994	100.00	13007	227238.17	6877529.31	0	4668308947
EXLAND	numerical	33419	1070994	100.00	491699	36423.89	3981575.79	0	2668500000
EXTOT	numerical	64255	1070994	100.00	432572	91186.98	6508402.82	0	4668308947
BLDFRONT	numerical	612	1070994	100.00	228815	23.04	35.58	0	7575
BLDDEPTH	numerical	621	1070994	100.00	228853	39.92	42.71	0	9393
STORIES	numerical	112	1014730	94.75	0	5.01	8.37	1	119
AVLAND2	numerical	58592	282726	26.40	0	246235.72	6178962.56	3	2371005000
AVTOT2	numerical	111361	282732	26.40	0	713911.44	11652528.95	3	4501180002
EXLAND2	numerical	22196	87449	8.17	0	351235.68	10802212.67	1	2371005000
EXTOT2	numerical	48349	130828	12.22	0	656768.28	16072510.17	7	4501180002

2.2 Categorical Value

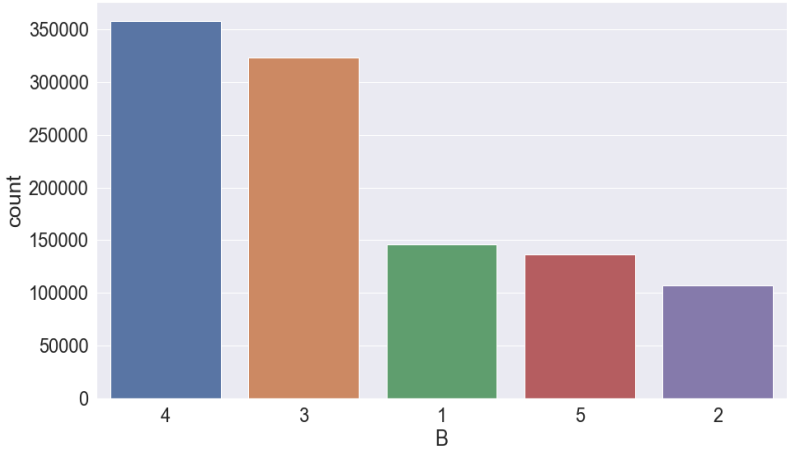
Field Name	Field Type	# unique values	# of records with value	% populated	most common field value
RECORD	categorical	1070994	1070994	100.00	n/a
BBLE	categorical	1070994	1070994	100.00	n/a
B	categorical	5	1070994	100.00	4
BLOCK	categorical	13984	1070994	100.00	3944
LOT	categorical	6366	1070994	100.00	1
EASEMENT	categorical	13	4636	0.43	E
OWNER	categorical	863347	1039249	97.04	PARKCHESTER PRESERVAT
BLDGCL	categorical	200	1070994	100.00	R4
TAXCLASS	categorical	11	1070994	100.00	1
EXT	categorical	4	354305	33.08	G
EXCD1	categorical	130	638488	59.62	1017
STADDR	categorical	839281	1070318	99.94	501 SURF AVENUE
ZIP	categorical	197	1041104	97.21	10314
EXMPTCL	categorical	15	15579	1.45	X1
EXCD2	categorical	61	92948	8.68	1017
PERIOD	categorical	1	1070994	100.00	FINAL
YEAR	date/time	1	1070994	100.00	2010/11
VALTYPE	categorical	1	1070994	100.00	AC-TR

3. Field Description

Field 1
Name: RECORD
Description: It is a unique integer label for each record.

Field 2
Name: BBLE
Description: Concatenation of AV_BORO, AV_BLOCK, AV_LOT, AV_EASEMENT. It's a unique combination for each property.

Field 3
Name: B
Description: Borough codes, from 1-5.



Field 4
Name: BLOCK
Description: Borough codes, from 1-5.

Top 10 common values

BLOCK	COUNT
3944	3888
16	3786
3943	3424
3938	2794
1171	2535
3937	2275
1833	1774
2450	1651
1047	1480
7279	1302

Field 5
Name: LOT
Description: The lot of the property.

Top 10 common values

LOT	COUNT
1	24367
20	12294
15	12171
12	12143
14	12074
16	12042
17	11982
18	11979
25	11949
21	11840

Field 6
Name: EASEMENT
Description: The right to cross/use the property land for special use.

Top 10 common values

EASEMENT	COUNT
E	4148
F	296
G	102
H	33
N	19
I	16
J	8
K	5
L	3
P	3

Field 7
Name: OWNER
Description: The owner of the property.

Top 10 common values

OWNER	COUNT
PARKCHESTER PRESERVAT	6020
PARKS AND RECREATION	4255
DCAS	2169
HOUSING PRESERVATION	1904
CITY OF NEW YORK	1450
DEPT OF ENVIRONMENTAL	1166
BOARD OF EDUCATION	1015
NEW YORK CITY HOUSING	1014
CNY/NYCTA	975
NYC HOUSING PARTNERSH	747

Field 8
Name: BLDGCL
Description: The building class.

Top 10 common values

BLDGCL	COUNT
R4	139879
A1	123369
A5	96984
B1	84208
B2	77598
C0	73111
B3	59240
A2	51130
A9	26177
B9	26133

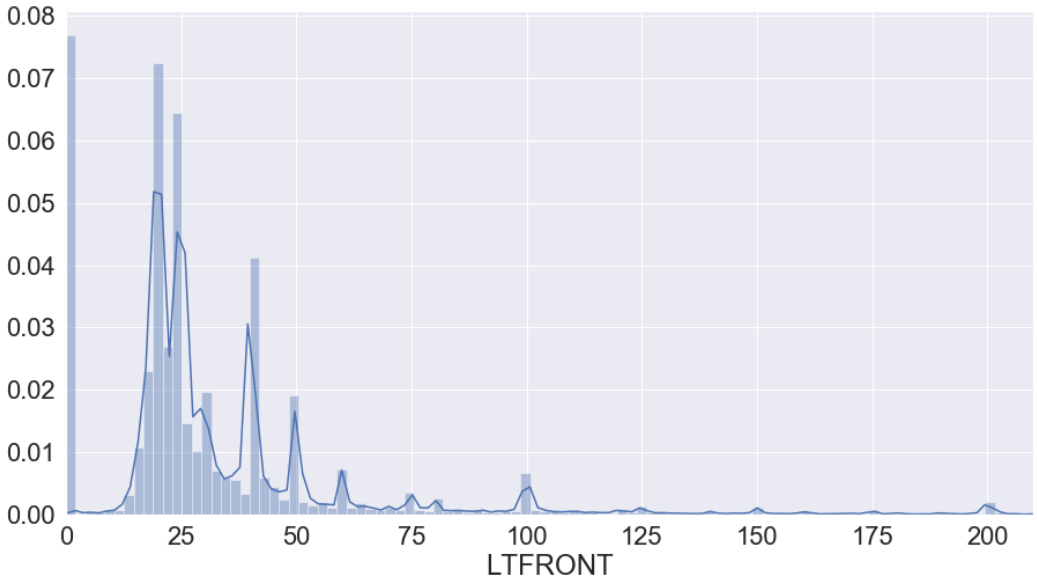
Field 9
Name: TAXCLASS
Description: The tax class.

Top 10 common values

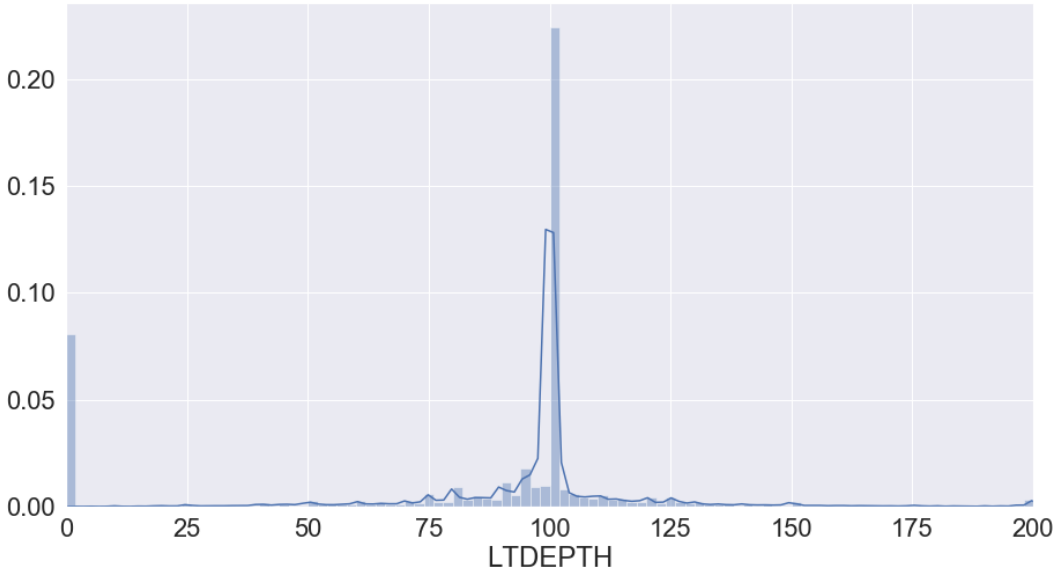
TAXCLASS	COUNT
1	660721
2	188612
4	104310
2A	40574
1B	24738
1A	21667
2B	13964
2C	10795
3	4638
1C	946

Field 10
Name: LTFRONT
Description: Lot width in feet.

Deleted outliers that have LTFRONT>210.
Data in the histogram is 98.67% populated among the records with values.



Field 11
Name: LTDEPTH
Description: Lot depth in feet.
Deleted outliers that have LTFRONT>200.
Data in the histogram is 97.62% populated among the records with values.



Field 12
Name: EXT
Description: Extension Indicator

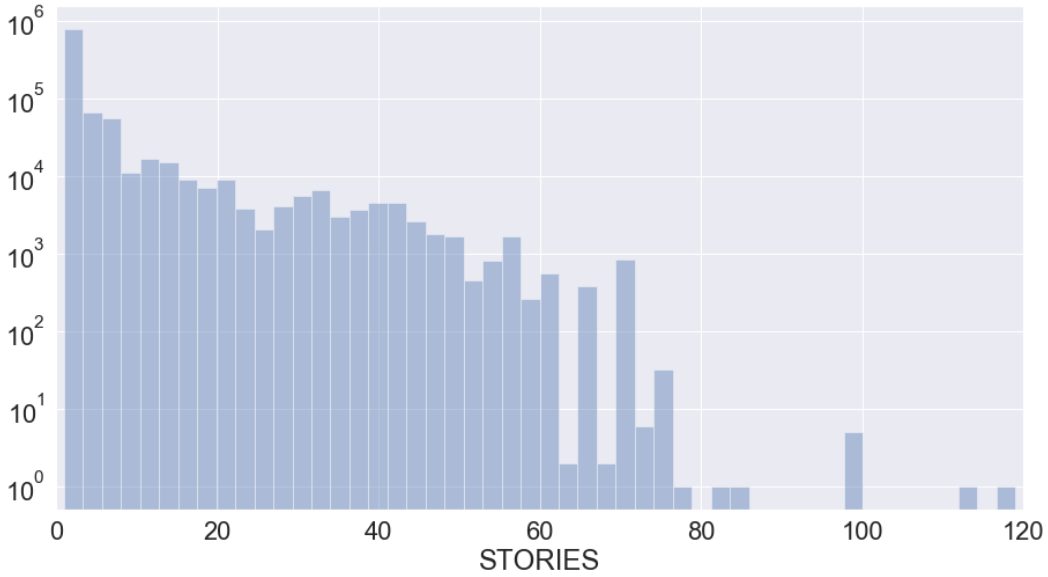
Top common values

EXT	COUNT
G	266970
E	49442
EG	37893

Field 13
Name: STORIES
Description: Number of Stories in Building

Didn't delete any outlier

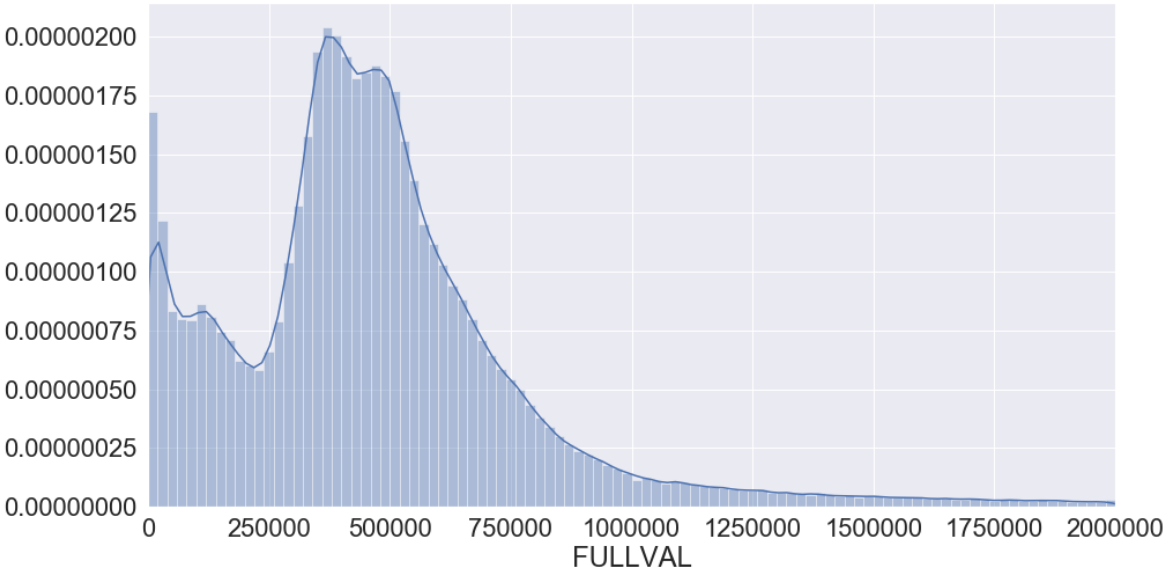
Data in the histogram is 100% populated among the records with values.



Field 14
Name: FULLVAL
Description: Market Value

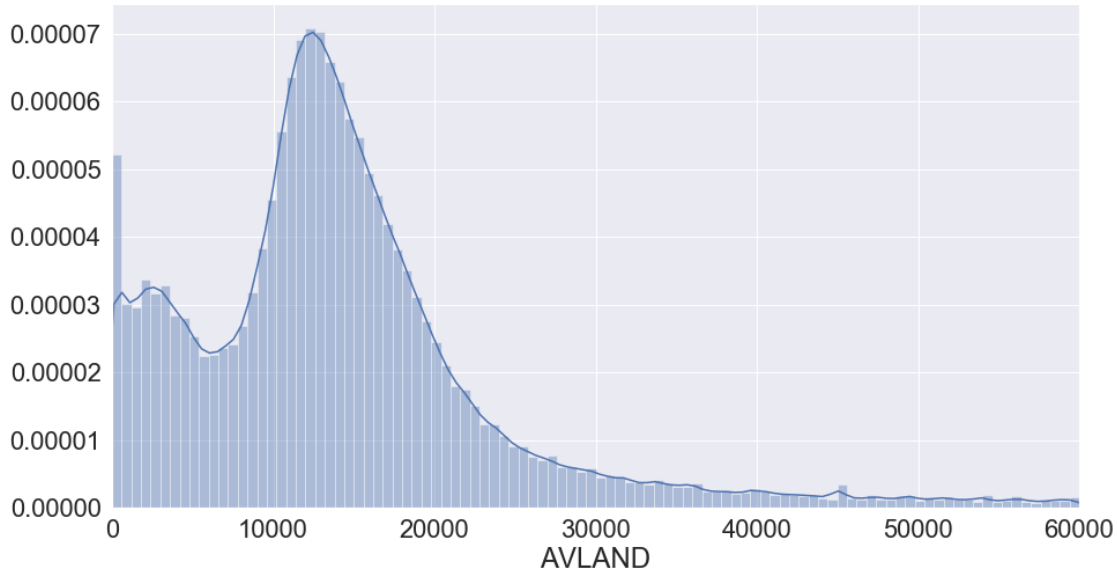
Deleted outliers that have fullval>2000000

Data in the histogram is 96.29% populated among the records with values.



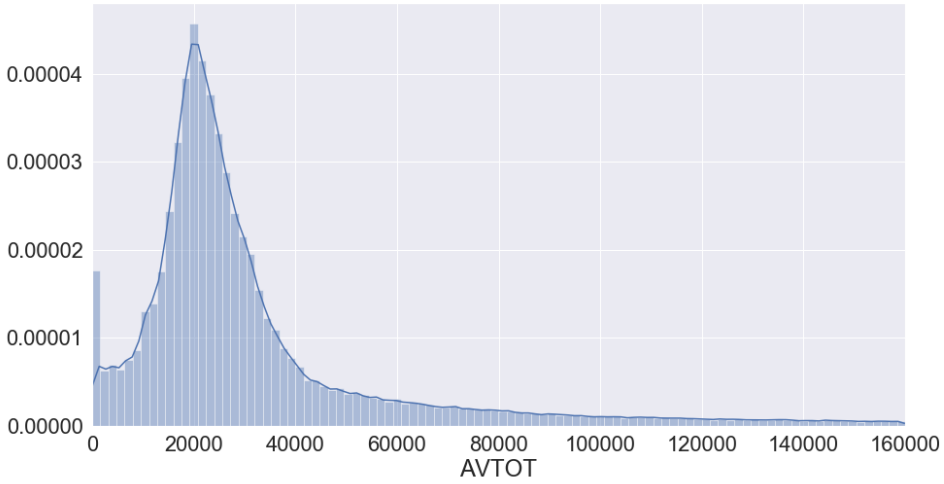
Field 15
Name: AVLAND
Description: Actual Land Value

Deleted outliers that have AVLAND>60000
Data in the histogram is 91.61% populated among the records with values.

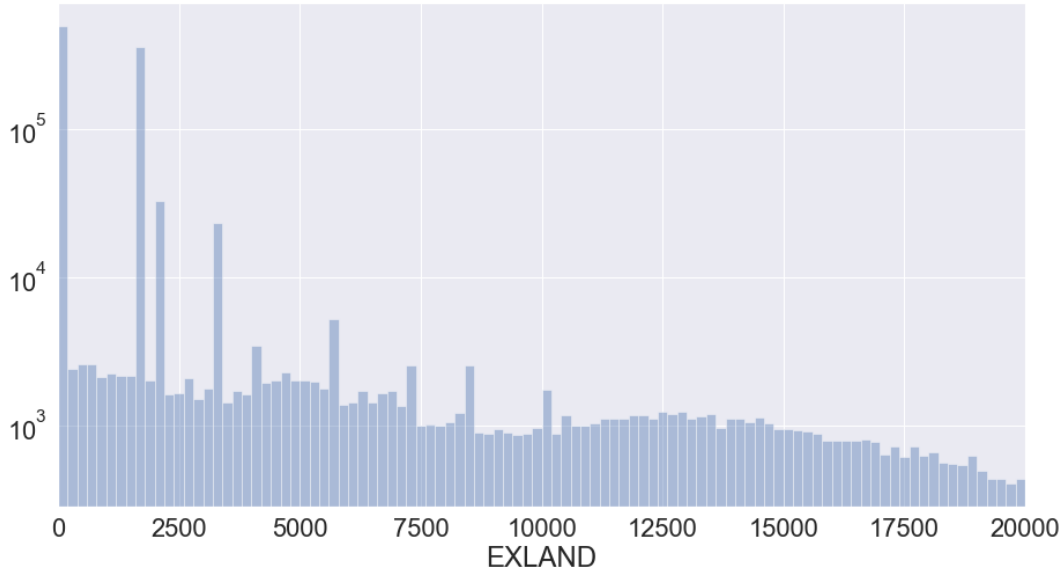


Field 16
Name: AVTOT
Description: Actual Total Value

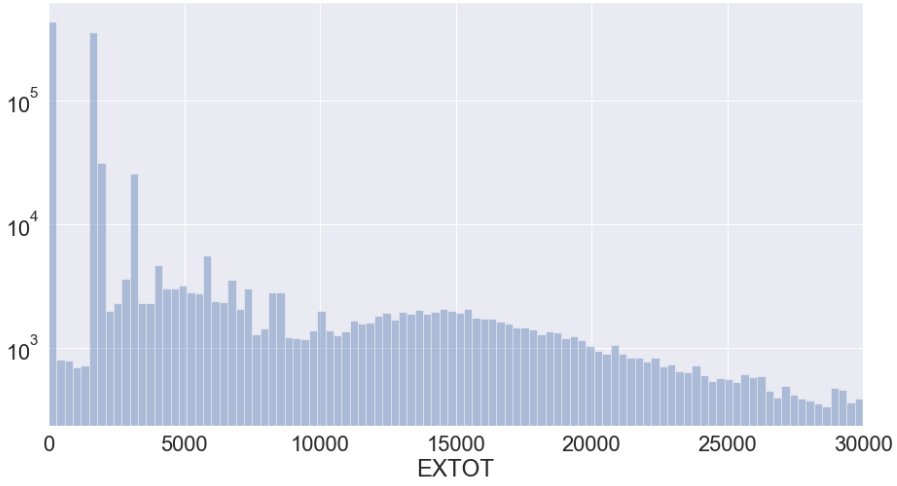
Deleted outliers that have AVTOT>160000.
Data in the histogram is 90.05% populated among the records with values.



Field 17
Name: EXLAND
Description: Actual Exempt Land Value
Deleted outliers that have EXLAND>20000
Data in the histogram is 96.82% populated among the records with values.



Field 18
Name: EXTOT
Description: Actual Exempt Land Total
Deleted outliers that have EXTOT>30000
Data in the histogram is 97.63% populated among the records with values.



Field 19
Name: EXCD1
Description: Exemption Code 1

Top 10 common values

EXCD1	COUNT
1017	425348
1010	49756
1015	31323
5113	23858
1920	17594
5110	16834
5114	14984
5111	10609
1021	6613
1986	4231

Field 20
Name: STADDR
Description: Street Address of the property

Top 10 common values

STADDR	COUNT
501 SURF AVENUE	902
330 EAST 38 STREET	817
322 WEST 57 STREET	720
155 WEST 68 STREET	671
20 WEST 64 STREET	657
1 IRVING PLACE	650
220 RIVERSIDE BOULEVARD	628
360 FURMAN STREET	599
200 EAST 66 STREET	585
30 WEST 63 STREET	562

Field 21
Name: ZIP
Description: zip code of the property

Top 10 common values

ZIP	COUNT
10314	24606
11234	20001
10312	18127
10462	16905
10306	16578
11236	15678
11385	14921
11229	12793
11211	12710
11207	12293

Field 22
Name: EXMPTCL
Description: Exemption Class of the property

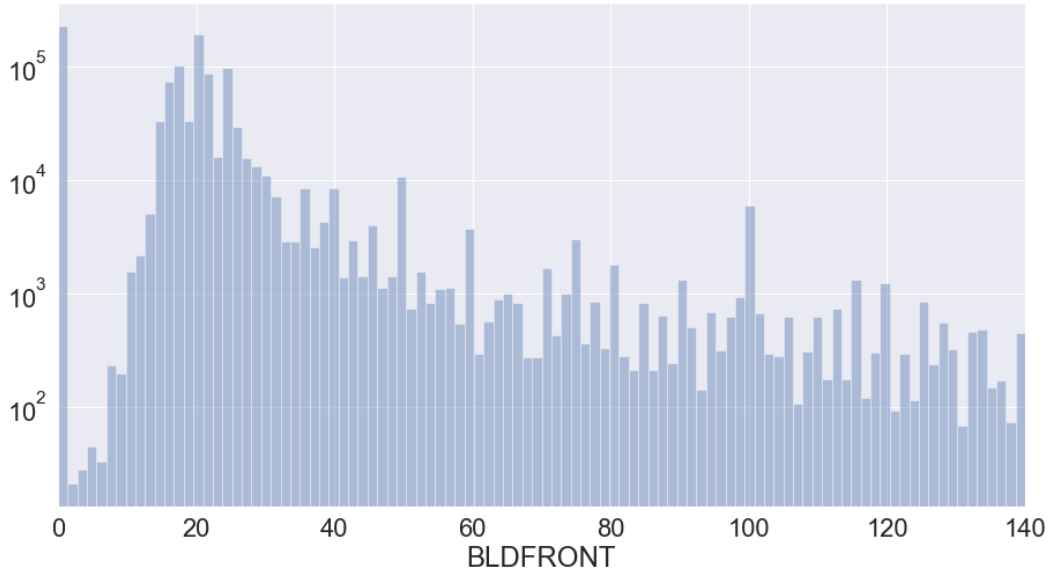
Top 10 common values

EXMPTCL	COUNT
X1	6912
X5	5208
X7	820
X2	770
X6	764
X4	441
X8	292
X3	259
X9	108
R4	1

Field 23
Name: BLDFRONT
Description: building width

Deleted outliers that have BLDFRONT>140

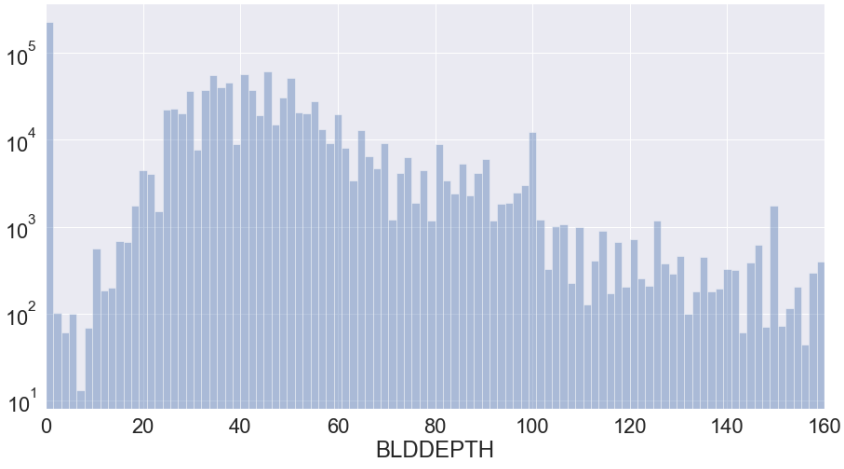
Data in the histogram is 98.39% populated among the records with values.



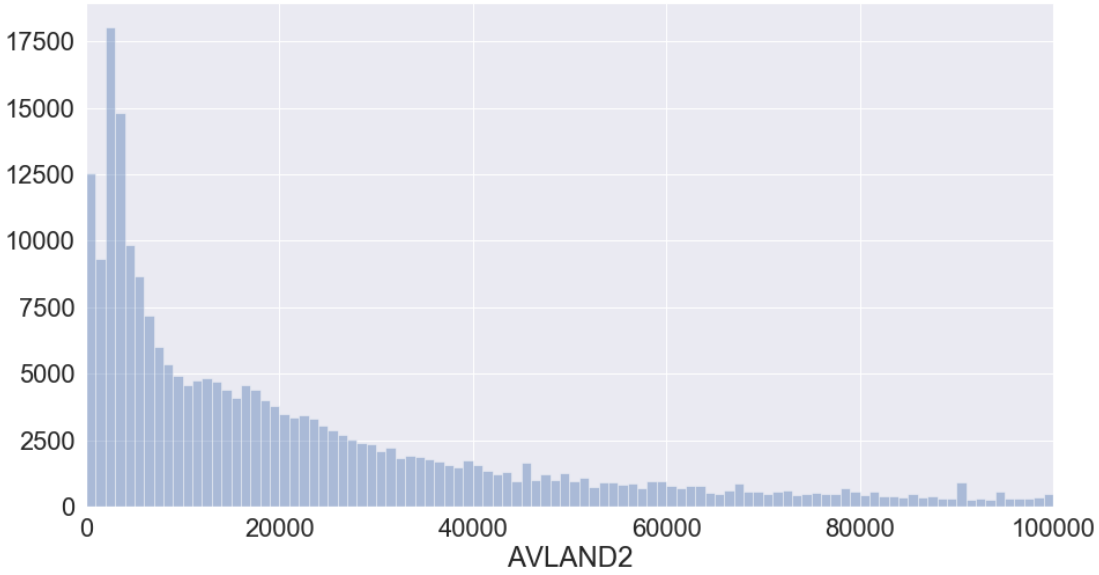
Field 24
Name: BLDDEPTH
Description: building depth

Deleted outliers that have BLDDEPTH>160

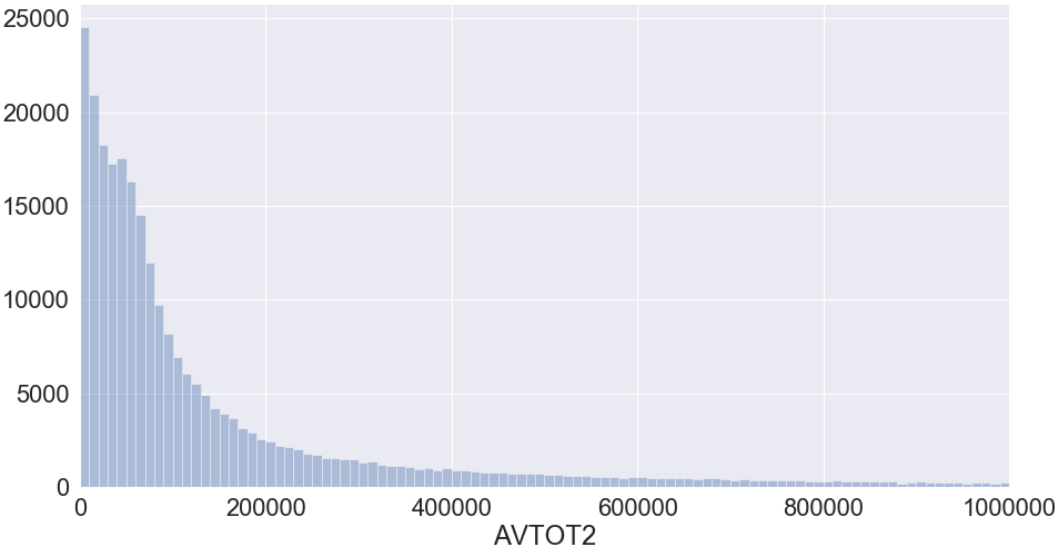
Data in the histogram is 98.91% populated among the records with values.



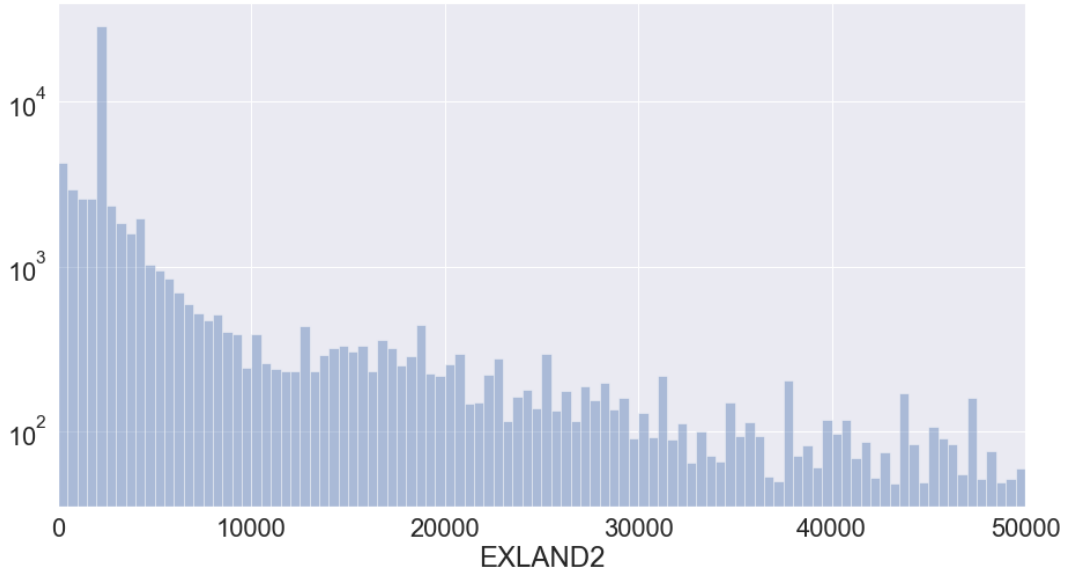
Field 25
Name: AVLAND2
Description: Transitional Land Value
Deleted outliers that have AVLAND2>100000
Data in the histogram is 81.33% populated among the records with values.



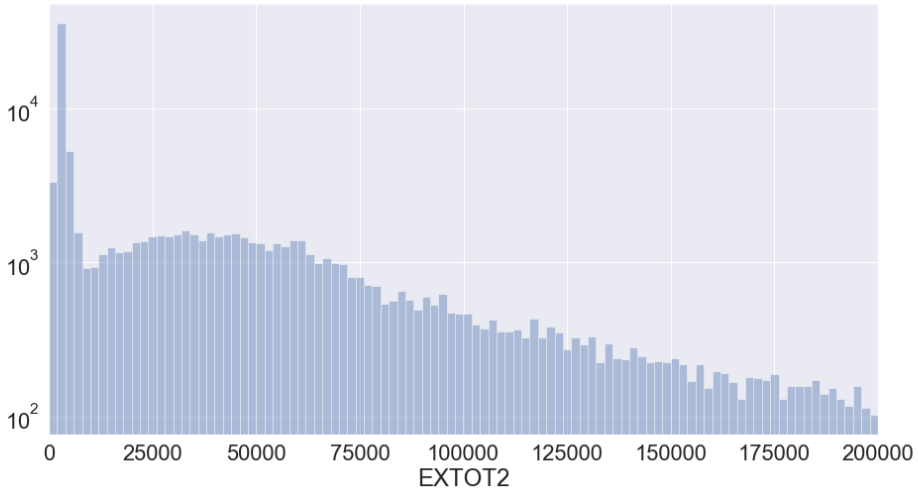
Field 26
Name: AVTOT2
Description: Transitional Total Value
Deleted outliers that have AVTOT2>1000000
Data in the histogram is 91.61% populated among the records with values.



Field 27
Name: EXLAND2
Description: Transitional Exemption Land Value
Deleted outliers that have EXLAND2>50000
Data in the histogram is 78.55% populated among the records with values.



Field 28
Name: EXTOT2
Description: Transitional Exemption Land Total
Deleted outliers that have EXTOT2>200000
Data in the histogram is 83.14% populated among the records with values.



Field 29
Name: EXCD2
Description: Exemption Code 2

Top 10 common values

EXCD2	COUNT
1017	65777
1015	12337
5112	6867
1019	3178
1920	2961
1200	881
1101	494
5129	227
1986	35
1022	31

Field 30
Name: PERIOD
Description: Assessment Period:
Only one unique record: Final

Field 31
Name: YEAR
Description: Assessment year:
Only one unique record: 2010/11 (November 2010)

Field 32
Name: VALTYPE
Description: Value type:
Only one unique record: AC-TR