Table of Contents

Executive Summary

Tax evasion is the act of avoiding one's true tax payments by deceiving the government. Property tax evasion is among the most common forms of tax evasion, especially in New York City, where tax rates are high and the city is dense with properties. Property tax fraud occurs at all levels of property ownership, ranging from the large real estate firms, to the individual property owners. Identifying and flagging properties with anomalous information, typically through manipulation of their real valuation or assessment, is a necessary task.

Our analysis will focus on identifying likely property tax fraud for New York City, through two fraud algorithms: a heuristic algorithm and an autoencoder. We will identify unusual values of variables, related to attributes of the property, that raise a cause for concern. The goal of this project is to assign each non-government property a score of fraudulence, identifying the principal one hundred most likely fraudulent records.

The report details the complete architecture, analysis, and determination processes of the unsupervised fraud models, which comprises seven sections:

1. Description of Data - original variables examined and summarised
2. Data Cleaning - excluded data and imputation
3. Variable Creation - manipulated variables to create expert variables
4. Dimensionality Reduction - removed correlations and performed linear transformation
5. Fraud Model Algorithms - established unsupervised machine learning model
6. Results - detailed description of results
7. Summary and Conclusions - summarized results and discussed future improvements

We used two different methods in building our unsupervised fraud model. The first being a heuristic function of z-scores, with a Minkowsky distance of p = 2, and the second being an autoencoder reproduction error, also with a Minkowsky distance of p = 2. After the construction and implementation of these two algorithms, we combined the two final scores by the average rank order, which became our final ranking system for which records were quantified by their possibility of fraudulence in descending order.

In addition, we discussed the implications of the inconsistencies in the New York City Property data, as well as our own methods in cleaning the dataset and assembling the unsupervised model. We also recognized future improvements that would make for a stronger analysis of the fraudulence of properties.

1 Data Description

1.1 File Description

The “NY property data.csv” is a property valuation and assessment dataset that represents the New York City property information. It stores 1,070,994 records of New York City properties of multiple characteristics and owners. It covers 32 fields, both numerical and categorical. A detailed description of the fields can be found in this report’s appendix. The dataset comes from NYC Open Data, a partnership between the Mayor’s Office of Data Analytics (MODA) and the Department of Information Technology and Telecommunications (DoITT). The dataset was last updated on November 17, 2010.

Table 1.1: File Description

Dataset Name	Property Valuation and Assessment Data
Dataset Purpose	Represent NYC properties assessments for the purpose to calculate Property Tax, Grant eligible properties Exemptions, and/or Abatements.
Data Source	NYC Open Data
Time Period	November 17, 2010
# of Fields	32
# of Records	1,070,994

1.2 Summary Statistics Table

In the dataset, 14 of the 32 fields are treated as numeric fields, while 16 of the 32 are treated as categorical fields. The data type of the remaining fields, EXCD1 and EXCD2, are unclear. Not enough information is provided about what these fields represent. EXCD1 is listed in both categorical and numeric summary tables, while EXCD2 is listed only in the categorical summary table. The field “VALTYPE”, also lacks information regarding what it represents. All records have the same VALTYPE value: (AC-TR).

Table 1.2.1: Summary Statistics of Categorical Fields

Field Name	# Of Records	% Populated	Unique Values	Most Common Field Value
RECORD	1070994	100%	1070994	N/A
BBLE	1070994	100%	1070994	N/A
B	1070994	100%	5	4
BLOCK	1070994	100%	13984	3944
LOT	1070994	100%	6366	1
EASEMENT	4636	.43%	13	E
OWNER	1039249	97.04%	863348	PARKCHESTER PRESERVAT
BLDGCL	1070994	100%	200	R4
TAXCLASS	1070994	100%	11	1
EXT	354305	33.08%	4	G
ZIP	1041104	97.21%	197	10314
EXCD1	638488	59.62%	130	1017
STADDR	1070318	99.84%	839281	501 SURF AVENUE
EXMPTCL	15579	1.45%	15	X1
EXCD2	92948	8.68%	61	1017
PERIOD	1070994	100%	1	FINAL
YEAR	1070994	100%	1	2010/11

VALTYPE	1070994	100%	1	AC-TR

Table 1.2.2: Summary Statistics of Numeric Fields

Field Name	# Of Records	Percent Populated	Unique Values	Mean	Standard Deviation	Min. Value	Max. Value	# Zeros
LTFRONT	1070994	100%	1297	36.64	74.03	0	9999	169108
LTDEPTH	1070994	100%	1370	88.86	76.4	0	9999	170128
STORIES	1014730	94.75%	112	5.01	8.37	1	119	0
FULLVAL	1070994	100%	109324	874264.51	11582430.99	0	6150000000	13007
AVLAND	1070994	100%	70921	85067.92	4057260.06	0	2668500000	13009
AVTOT	1070994	100%	112914	227238.17	6877529.31	0	4668308947	13007
EXLAND	1070994	100%	33419	36423.89	3981575.79	0	2668500000	491699
EXTOT	1070994	100%	64255	91186.98	6508402.82	0	4668308947	432572
EXCD1	638488	59.62%	130	1602.01	1384.23	1010	7170	0
BLDFRONT	1070994	100%	612	23.04	35.58	0	7575	228815
BLDDEPTH	1070994	100%	621	39.92	42.71	0	9393	228853
AVLAND2	282726	26.40%	58592	246235.7	6178962.56	3	2371005000	0

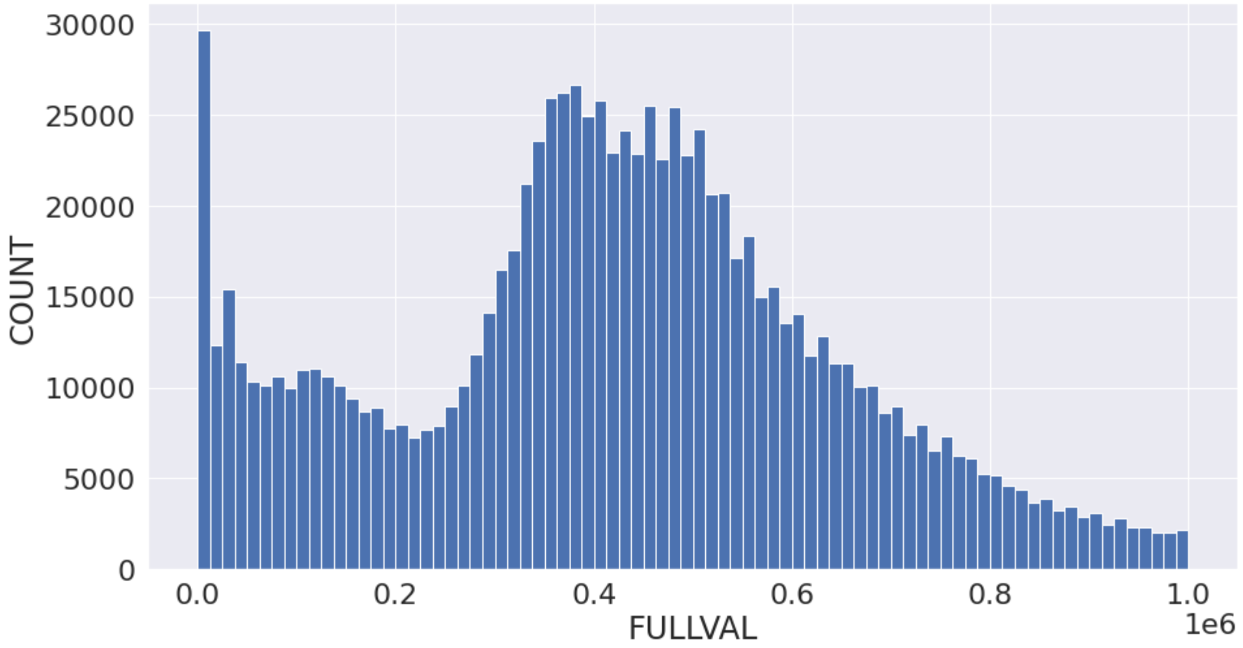
				2				
AVTOT2	282732	26.40%	111361	713911.44	11652528.95	3	4501180002	0
EXLAND2	87449	8.17%	22196	351235.68	10802212.67	1	2371005000	0
EXTOT2	130828	12.23%	48349	656768.28	16072510.17	7	4501180002	0

1.3 Field Examples

1.3.1 Field “FULLVAL”

Table 1.3.1: FULLVAL

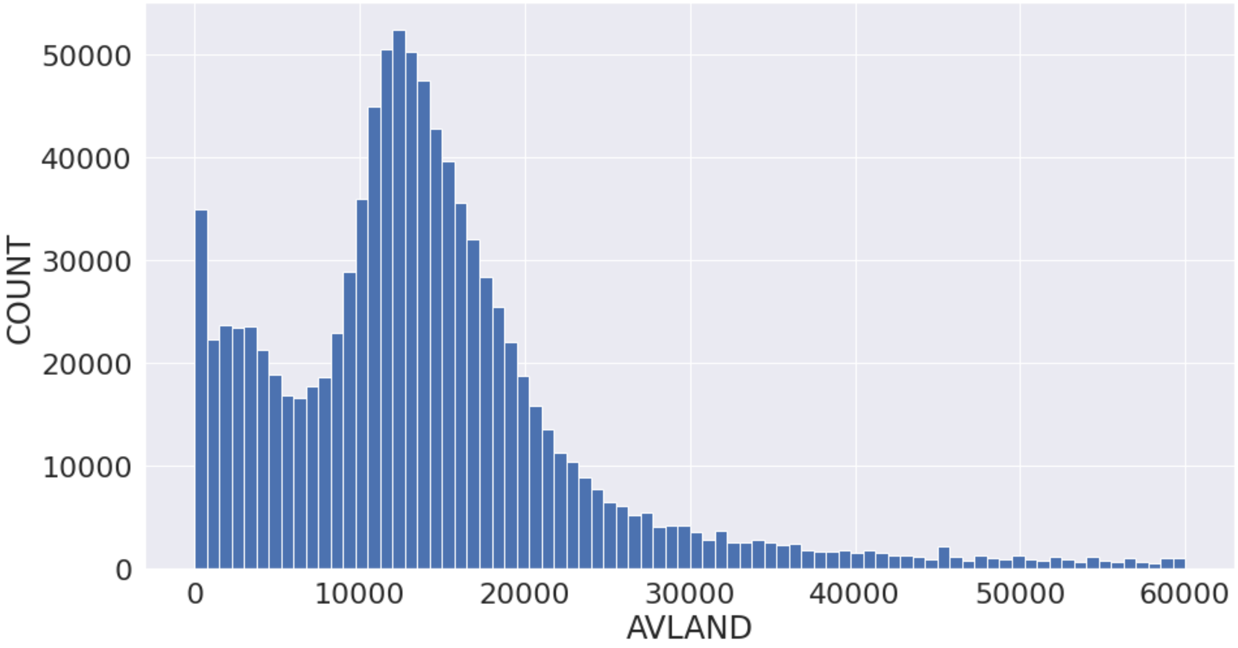
Name: FULLVAL
Description: Total market value of the land.
Exclude outliers > 1000000, data in histogram is 91.35 % populated.



1.3.2 Field “AVLAND”

Table 1.3.2: AVLAND

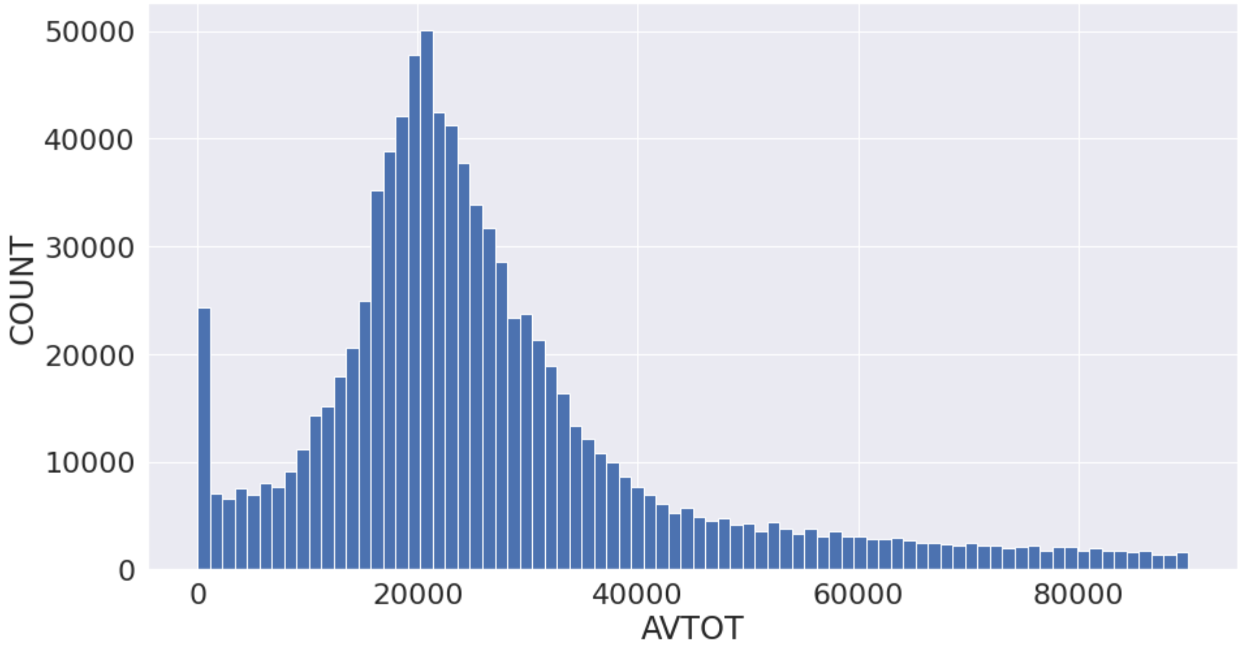
Name: AVLAND
Description: Assessed land value.
Exclude outliers > 60000, data in histogram is 91.61 % populated.



1.3.3 Field “AVTOT”

Table 1.3.3: AVTOT

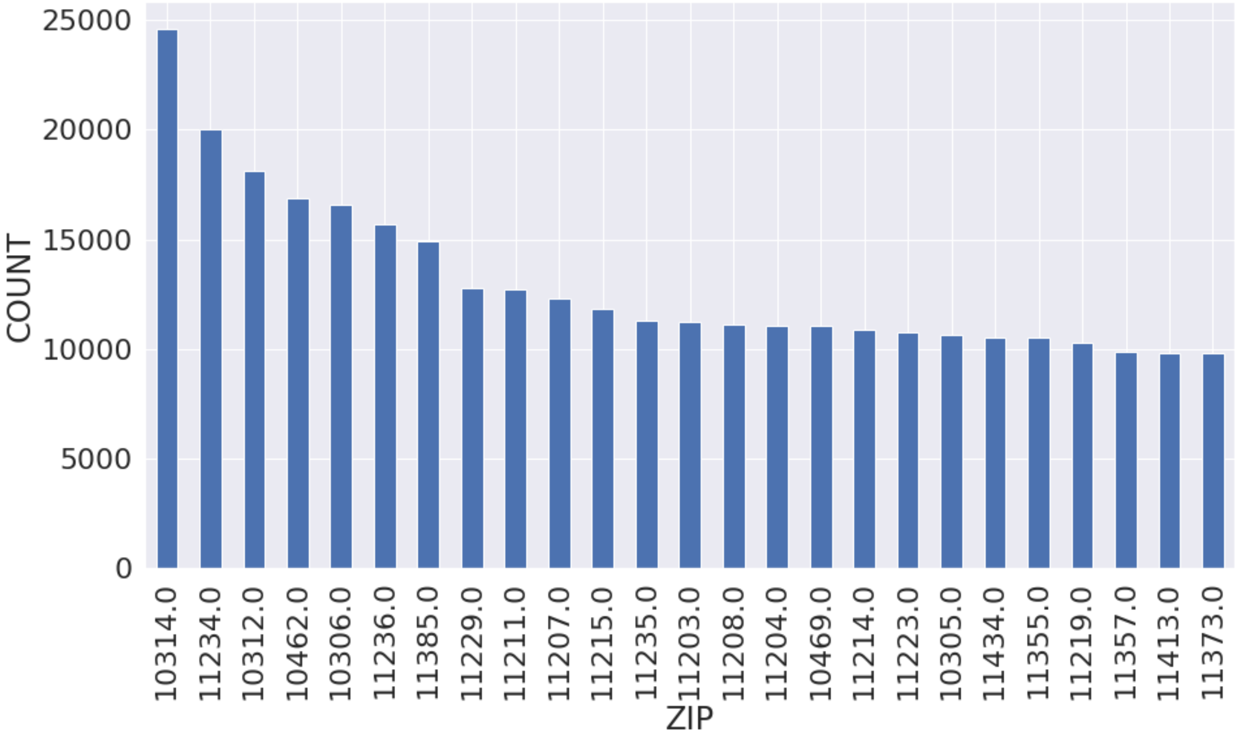
Name: AVTOT
Description: Assessed total value.
Exclude outliers > 90000, data in histogram is 84.98 % populated.



1.3.4 Field “ZIP”

Table 1.3.4: ZIP

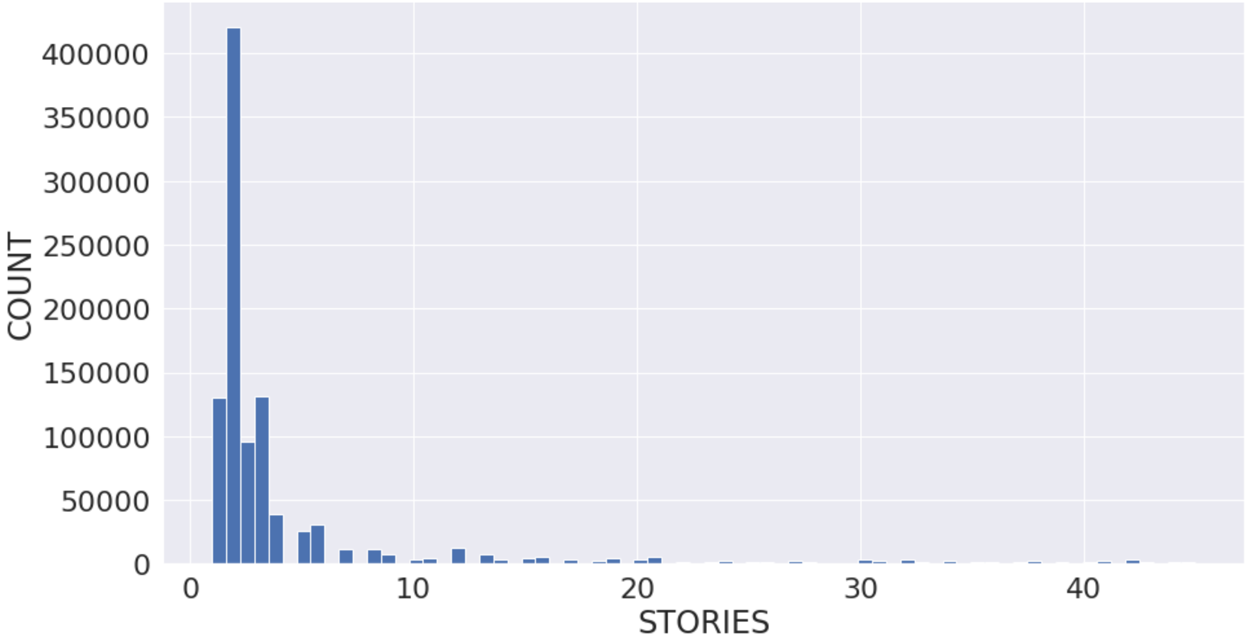
Name: ZIP
Description: Zip code of the property



1.3.5 Field “STORIES”

Table 1.3.5: STORIES

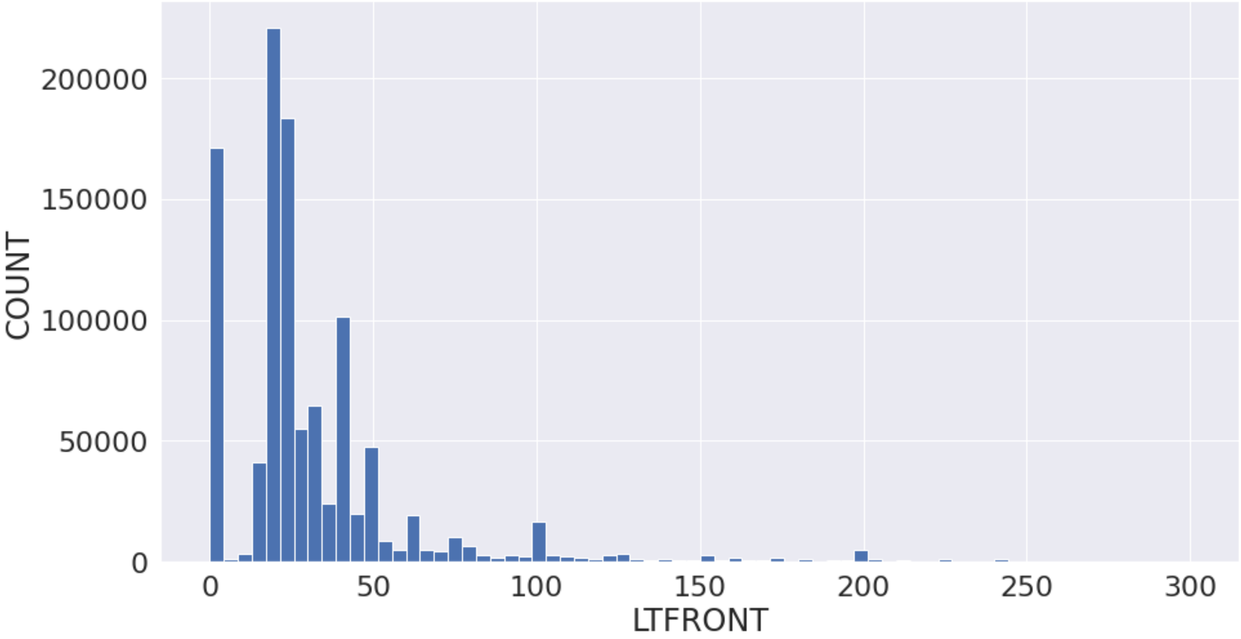
Name: STORIES
Description: The number of stories for the building (# of Floors).
Exclude outliers > 45, data in histogram is 99.16 % populated.



1.3.6: Field “LTFRONT”

Table 1.3.6: LTFRONT

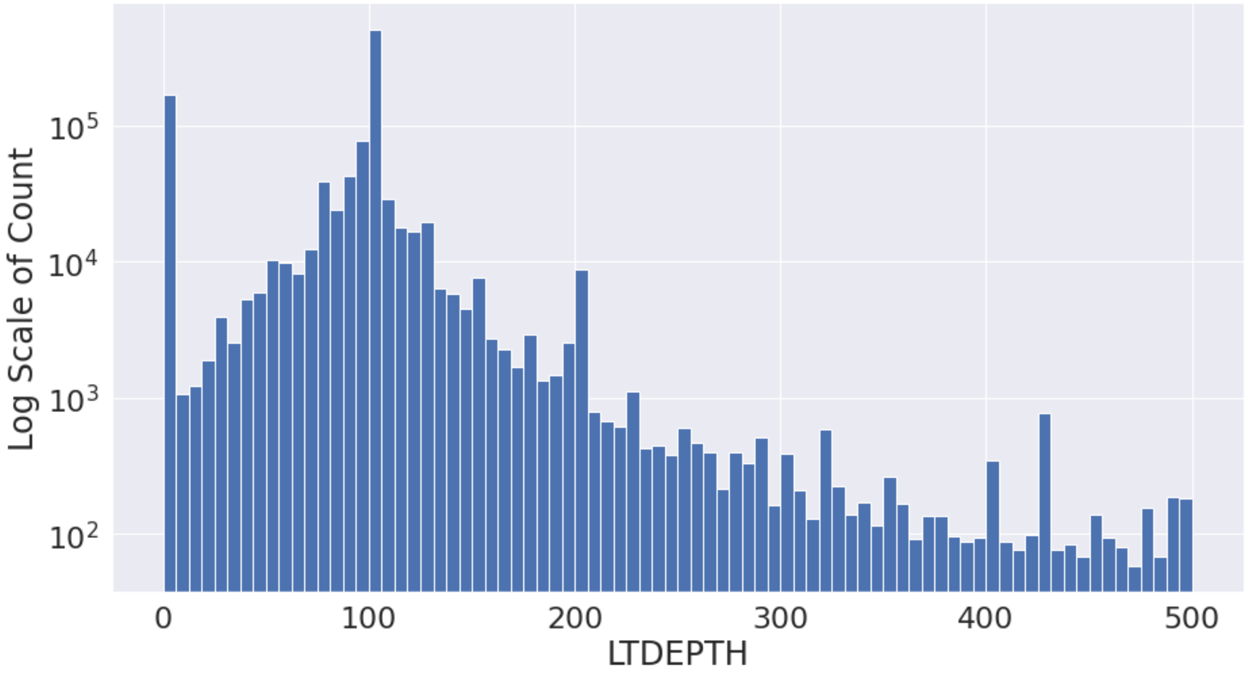
Name: LTFRONT
Description: Lot frontage in feet.
Exclude outliers > 300, data in histogram is 99.30 % populated.



1.3.7: Field “LTDEPTH”

Table 1.3.7: LTDEPTH

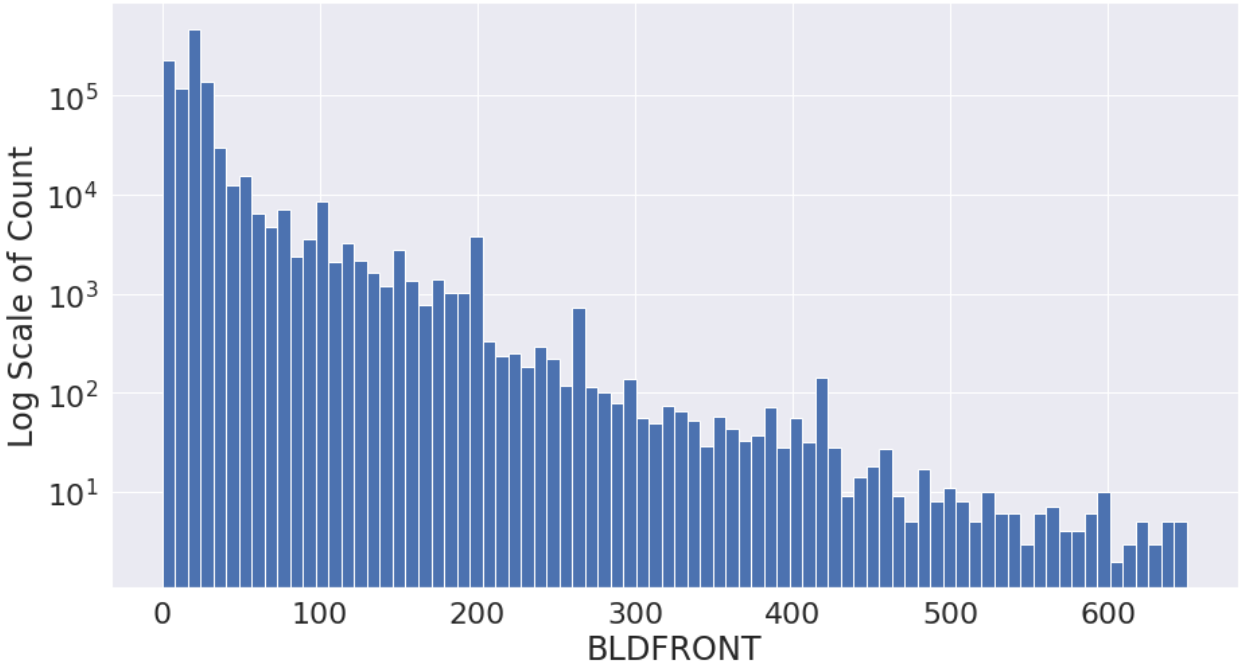
Name: LTDEPTH
Description: Lot depth in feet.
Exclude outliers > 500, data in histogram is 99.68 % populated.



1.3.8 Field “BLDFRONT”

Table 1.3.8: BLDFRONT

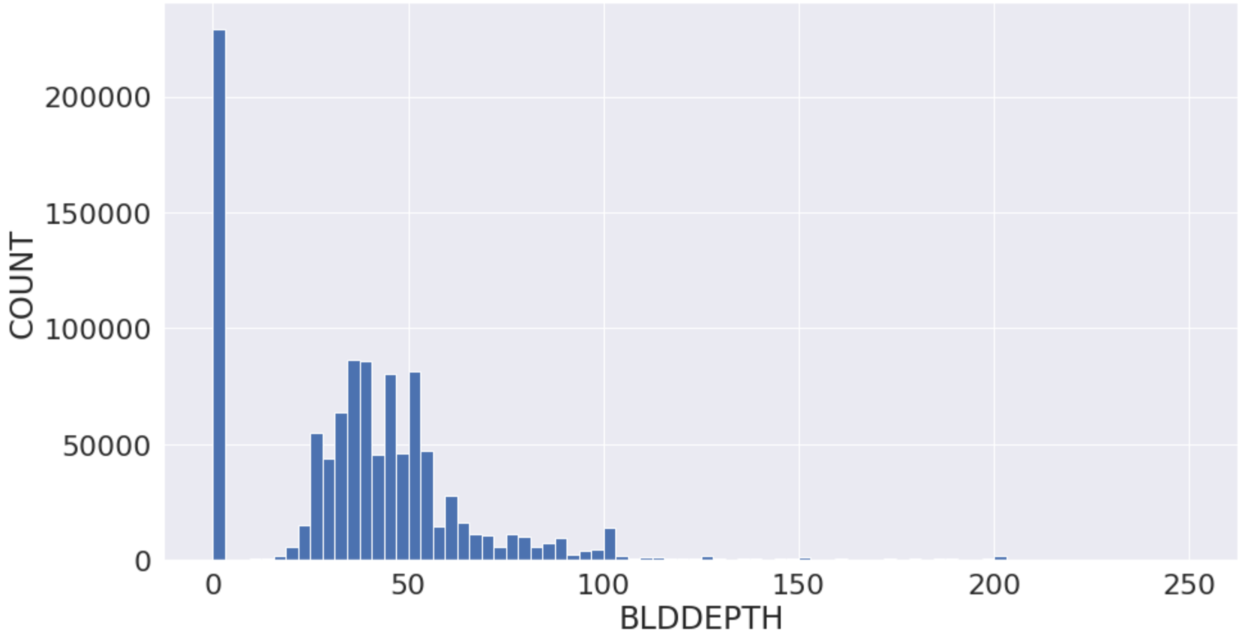
Name: BLDFRONT
Description: Building Frontage in feet.
Exclude outliers > 650, data in histogram is 99.99 % populated.



1.3.9 Field “BLDDEPTH”

Table 1.3.9: BLDDEPTH

Name: BLDDEPTH
Description: Lot depth in feet.
Exclude outliers > 300, data in histogram 99.72 % populated.



2 Data cleaning

Prior to creation of the expert variables, cleaning the data was necessary. Because we are concerned with identifying potential property tax fraud, we are not interested in properties owned by city, state, or federal governments. For this reason, we have excluded these properties before performing any variable calculations. Including them would skew the statistics and significance of each variable. We removed all records with the following owners: 'PARKCHESTER PRESERVAT', 'PARKS AND RECREATION', 'DCAS', 'HOUSING PRESERVATION', 'CITY OF NEW YORK', 'DEPT OF ENVIRONMENTAL', 'BOARD OF EDUCATION', 'NEW YORK CITY HOUSING', 'CNY/NYCTA', 'NYC HOUSING PARTNERSH', 'DEPARTMENT OF BUSINES', 'DEPT OF TRANSPORTATIO', 'MTA/LIRR', 'PARCKHESTER PRESERVAT', 'MH RESIDENTIAL 1, LLC', 'LINCOLN PLAZA ASSOCIA', 'UNITED STATES OF AMER', 'U S GOVERNMENT OWNRD', 'THE CITY OF NEW YORK', 'NYS URBAN DEVELOPMENT', 'NYS DEPT OF ENVIRONME', 'CULTURAL AFFAIRS', 'DEPT OF GENERAL SERVI', 'DEPT RE-CITY OF NY'.

In addition, we filled in the missing field values of all variables necessary for Principal Component Analysis. For the unsupervised fraud model to accurately score all non-government records, even those with some missing fields, it is critical that we fill in missing values with neutral values so that these values do not set off any unnecessary alarm. We filled these missing fields with the most typical value for that field, for that record. Also, it is important to note that we only filled missing values for fields that will be used to create variables, which are: FULLVAL, AVLAND, AVTOT, ZIP, STORIES, LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH.

2.1 Filling in missing ZIP

Of the dataset, 21,772 records are missing the ZIP variable (after removing exclusions). We observed that the records are for the most part sorted by zip code. In filling the missing zip codes, we adopted the following procedure: if the zip on both the record before and after the record with the missing zip code are the same, replace the missing value with that zip code (this reduces the original 21,7772 records with missing zips to 10,400). We then replaced all missing zip codes from the remaining records with the zip code from the record above.

2.2 Filling in missing FULLVAL, AVLAND, AVTOT

About 13,000 records are missing these property value amounts, we resolved this issue by filling in the missing values with the average by that record’s TAXCLASS.

2.3 Filling in missing STORIES

43,968 records are missing the number of stories in the building, we resolved this issue by filling in the missing values with the average by that record’s TAXCLASS.

2.4 Filling in missing LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH

About 200,000 records are missing some or all values related to lot and building size, approximately 20% of the records. We followed the following procedure to handle the missing values of all these fields: Replace the 0’s and 1’s by NAs so they are not counted in calculating the mean. Then group by TAXCLASS and calculate the groupwise average of each of the missing lot and building size variables. Lastly, impute the resulting values for each of the fields.

3 Variable Creation.

To estimate a property's value effectively we need to understand how the size of the property correlates with dollar value, price per square foot. Similarly, for location and property type. Dollars per building volume is an important metric for high-rise buildings. Keeping average values as a baseline and comparing the records will be helpful in determining unusual properties. A total of 45 variables were created based on existing dollar value and size combinations to better understand the records.

We start by creating 3 sizes for the properties:

- **S1 = lotarea** = LTFRONT * LTDEPTH
  Area of the lot i.e. how big the front of the lot is times how deep the lot is.
- **S2 = bldarea** = BLDFRONT * BLDDEPTH
  Footprint of the building i.e how large the building is.
- **S3 = bldvol** = S2 * STORIES
  Volume of building i.e. footprint multiplied by the number of stories.

Next, we created 9 variables, each of the three dollar value fields divided by each of the three sizes. (3 * 3 = 9)

We take **V1** = FULLVAL , **V2** = AVLAND, **V3** =AVTOT and for each record we append 9 ratios.
All valuations divided by three sizes:

1. **r1** = V1/S1
   Market value of land per area of the lot.
2. **r2** = V1/S2
   Market value of land per footprint of the building.
3. **r3** = V1/S3
   Market value of land per volume of the building.
4. **r4** = V2/S1
   Assessed land value per area of the lot.
5. **r5** = V2/S2
   Assessed land value per footprint of the building.
6. **r6** = V2/S3
   Assessed land value per volume of the building.
7. **r7** = V3/S1
   Assessed total value per area of the lot.
8. **r8** = V3/S2
   Assessed total value per footprint of the building.
9. **r9** = V3/S3
   Assessed total value per volume of the building.

Now, we calculate the dollar per square foot by groups like geographical region i.e. ‘zip3’, ‘zip5’, borough, and including logical grouping, such as the tax class the property belongs to. This is to understand what the value ratios for these neighbouring areas are. For example, dollars per square foot is different for Manhattan compared to The Bronx. We do this by normalizing our variables with the above four variables. (9 * 4 = 36)

Separately grouping records by these 4 groups: ‘zip5’, ‘zip3’, ‘TAXCLASS’, ‘BOROUGH’. And for each group g, we calculate $< r_i >_g$, the average of each ratio r_i for each group g and then, append that to 36 additional variables.
$r_1/ < r_1 >_g$, $r_2/ < r_2 >_g$, $r_3/ < r_3 >_g$,... $r_9/ < r_9 >_g$ and g = 1,2,3,4.
For example : **r1_zip5 =** $r_1/ < r_1 >_{zip5}$ i.e. Market value of land per area of lot per its average value in ‘zip5’.

1. **r1_zip5**
2. **r2_zip5**
3. **r3_zip5**
4. **r4_zip5**
5. **r5_zip5**
6. **r6_zip5**
7. **r7_zip5**
8. **r8_zip5**
9. **r9_zip5**
10. **r1_zip3**
11. **r2_zip3**
12. **r3_zip3**
13. **r4_zip3**
14. **rS_zip3**
15. **r6_zip3**
16. **r7_zip3**
17. **r8_zip3**
18. **r9_zip3**
19. **r1_taxclass**
20. **r2_taxclass**
21. **r3_taxclass**
22. **r4_taxclass**
23. **rS_taxclass**
24. **r6_taxclass**
25. **r7_taxclass**
26. **r8_taxclass**
27. **r9_taxclass**
28. **r1_boro**
29. **r2_boro**
30. **r3_boro**
31. **r4_boro**
32. **rS_boro**
33. **r6_boro**
34. **r7_boro**
35. **r8_boro**
36. **r9_boro**

4 Dimensionality Reduction.

After building the 45 expert variables that quantify signals for the fraud algorithms, we first z-scaled each variable to prepare for dimensionality reduction.

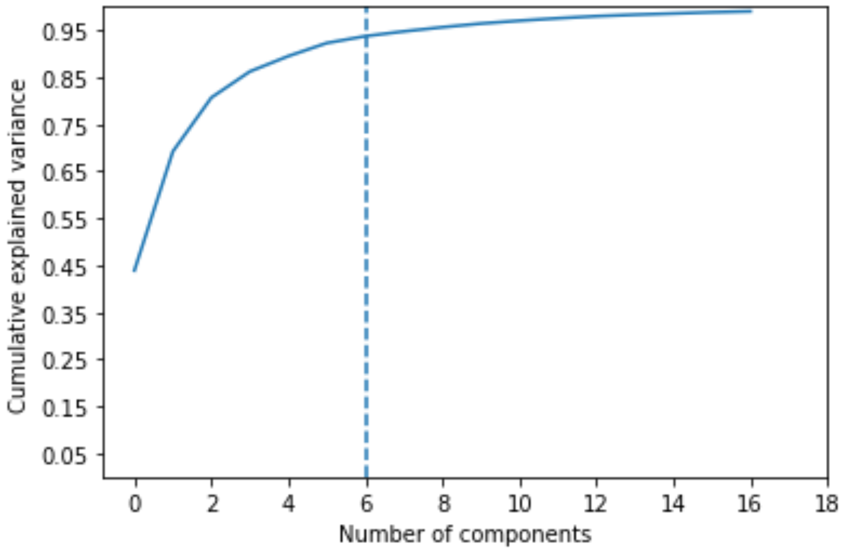
4.1 Z Scale Variables

From a dataframe of only those 45 variables, with each record in place, we z-scaled each column, subtracting each data point mean from the data point, and then dividing it by the data point's standard deviation.

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

4.2 Reduce Dimensions via PCA

After z-scaling the 45 variables, we performed principal component analysis. In this step, we maintained 99% of the components running the exact full singular value decomposition to select the components by postprocessing. A scree plot, depicting the Number of Components on the x-axis and their Cumulative Explained Variance on the y-axis, is exhibited below.



We then selected the top six principal components, based on their respective eigenvalues and the plot above (depicted by the vertical dotted line), as a benchmark to continue on with our analysis.

$$PC_1 = \Sigma_i a_i z_i$$

$$PC_2 = \Sigma_i b_i z_i$$

...

4.3 Z-scale Principal Components

Once the top six principal components had been selected, we once again z-scaled the reduced variables, subtracting the data point mean from the data point, and dividing by the data point’s standard deviation.

$$PCz_i = \frac{PC_i - \mu_{PC_i}}{\sigma_{PC_i}}$$

5 Fraud Model Algorithms

Now the data is ready for model training. First, we generated two different fraud scores by two different methods with the six principal components. The first score, Score 1, is based on the z-score and the second score, Score 2, is calculated from the autoencoder reproduction error. Then, all records were assigned a respective rank based on each score, and the average of the two rankings was used as the Final Score.

5.1 Score1: z score outliers

The first fraud algorithm model, Score 1, is based on distance of z-score. First, We generated 45 variables based on domain knowledge and reduced the dimensionality of those variables to obtain principal components. We then used the top six principal components. We calculated the Euclidean distance of the standardized PCs for each record. In other words, it is the Mahalanobis distance of the original PCs. The larger this distance is, the more extreme the record is than the average.

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

5.2 Score2

The other fraud algorithm model, Score 2, is based on an autoencoder. An autoencoder is a neural network that consists of three components: encoder, code and decoder. The encoder compresses the input and generates the code, and the decoder decodes this code and reconstructs the input only from the code.

For computation, we used the same six PCs as Score 1 and chose Keras API for computation. We calculated the Euclidean distance between original input and the output as the fraud score.

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

5.3 Final Score

Final Score is constructed from the average ranking between Score 1 and Score 2. We ranked the records according to each of the two scores in order from largest to smallest and took the average ranking as the final score.

6 Results

6.1 Top 5 Records

Record #1: 917942

This property is owned by Logan Property, Inc. in Queens, near JFK International Airport. The size of the front lot is listed as 4910 feet but the lot depth was missing and after cleaning was imputed with 124.44 feet. The building front and depth was missing as well and was imputed with a building front of 62.08 feet, building depth of 86.29 feet, and 3 stories. The full market value is listed at $374,019,883, meanwhile the assessed value of the land is $1,792,808,947 and the assessed total value is $4,668,308,947. The property has a building classification of T1 which corresponds to airports, airfields, or terminals.

This property was flagged as an anomaly due to its very enormous assessed values (both land and total). Many of our variables are sensitive to these assessed values and as a result this property was given the top fraud ranking, especially when grouped by zip code and borough. It’s market value of $374 million also contributed to its high fraud ranking, but the rank was largely in part due to its assessed value. It is interesting to note that while this property has the classification of T1 it appears to be a Holiday Inn in 2021 - the records were last updated in 2018. This is very plausible, as hotels are always looking to acquire land near airports, and ownership may have changed. Regardless, this record needs to be reevaluated.

Record #2: 684704

The address for the property is 69th Street (in what we are assuming to be Queens). It is of building class V0, meaning it is zoned residential, not in Manhattan. The owner of this property is said to be “W Rufert”' and the lot front and depth is 2 feet by 2 feet respectively, which seems unlikely. However, the building front and depth values were missing and were imputed with a building front of 39.5 feet, building depth of 82.67 feet, and 4 stories. The full market value of the property is $373,839.87. The assessed value of the land and total property is $9,185.39 and $9,189.08, respectively.

Quite a bit about this record seems to have been flagged as an outlier by the algorithms. To begin with, the property’s lot size, 2 by 2, paired with the market value of the property, resulted in many of our variables being very large. In addition, the building volume and low assessed value of the property resulted in the property being labeled as an outlier, especially within its borough.

Record #3: 1065870

Similar to record 2, this property is also zoned residential (building class V0), but this property is in Staten Island and has a lot front and lot depth of 2891 and 1488 feet, respectively. The building front, depth, and stories were missing and were imputed with a front of 39.5 feet, depth of 82.67 feet, and 4 stories. The owner is said to be "The People of the St of N." The full market value of this property is $290,174,603.00. The assessed value of the land is $17,410,476.00, which is also equal to the total assessed value of the property - determining the building on the property to be assessed at $0.

The fraud algorithm ranked this property high because of its small building area and volume relative to its market and assessed values. In particular, the large assessed value of the land on this property led to a very high fraud ranking - especially when grouped geographically (by zip code or borough). Further, because the area of the lot is large, when it normalizes market value, the values of the variables are actually below the mean, contributing to the algorithm's high rank.

Record #4: 1059883

This property's address is Sagona Court, in Staten Island, which is a small (what looks to be) residential culd-a-saque. The property is in building class Z7 (Easement) and has an easement value of 'E' which indicates a portion of the lot that has a Land Easement. The lot front and lot depth are 5 and 5 feet respectively, which seem unlikely, and additionally, the building front, depth, and stories were missing and were imputed with a front of 62.1 feet, depth of 86.3 feet, and 5.5 stories. The full market value is $2,772,746.70, the assessed value of the land is $444,996.05, and the total assessed value is $1,294,236.95.

This is now the second record (record 2 as well) that has a lot area less than the building area. The variables, when normalized by this tiny lot area, skew the fraud rank. When evaluating the variables that are normalized by building area and volume, they are not major outliers.

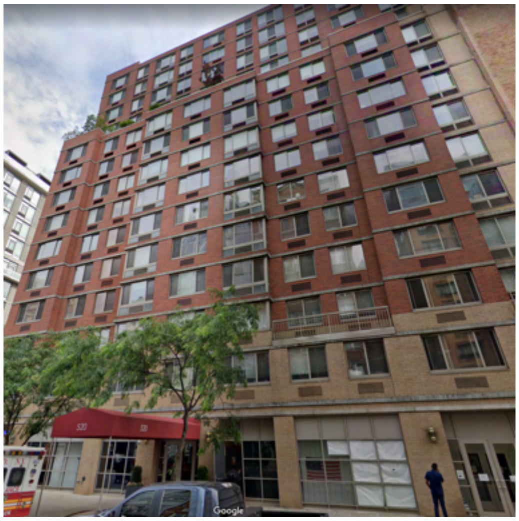
Record #5: 151044

This property is listed in building class Q6 (Stadium, Racetrack, Baseball Field) with a lot front and depth of 798 and 611 feet, respectively. It does not have an owner listed, but this property seems to be Yankee Stadium (via Google Maps). It has a full market value of $1,663,775,000.00 and has a total assessment value of $748,698,750.00 (Land assessment value of $78,750,000). The building front, depth, and stories were missing and were imputed with a building front and building depth of 62.08 feet and 86.30 feet, respectively, and 6 stories.

This record does not need much investigation, as it is clear that given the building area/volume and the market/assessed value of this property, the algorithm would identify this record as an outlier. There are not many 1.66 billion dollar properties in the Bronx, let alone with that small of building area and volume.

6.2 5 Other Intriguing Records

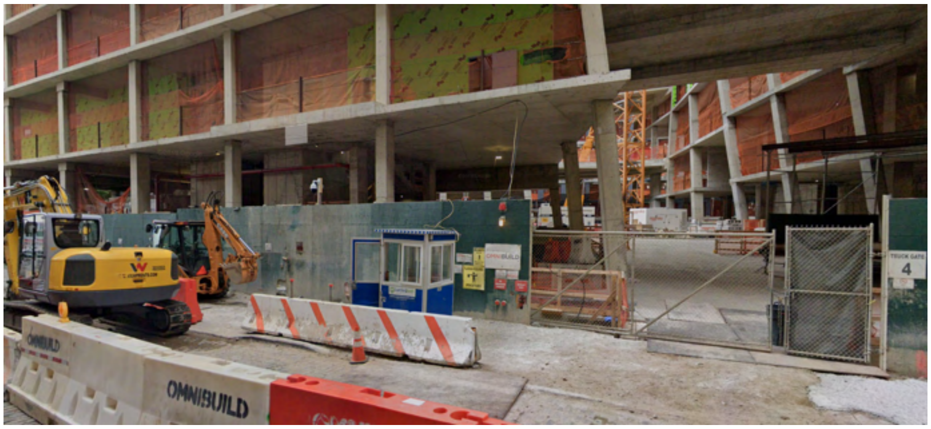
Record #6: 33751



This property has an address of 520 W 23rd St in lower Manhattan and the owner is Frank Guidara. The property is of building class D4 (Elevator Cooperatives), meaning it is an elevator building with coops, rather than condominiums. The lot front is 122 feet, lot depth is 98 feet, and the property is listed to have 15 stories. The building front is listed as 8 feet and the building depth is listed as 10 feet. The full market value is $14,400,000.00. The assessed value of the land is $540,000.00 and the total assessed value of the property is $6,480,000.00.

Similar to a few of the records above, the algorithm assigned the record a high rank due to a very small building area and volume and high value. When the market value and assessed values were normalized by building area and volume, the algorithm identified this record as an outlier - regardless of geography and tax class. It is clear from the picture that the building on this property is much larger than 8 by 10 feet.

Record #7: 33564

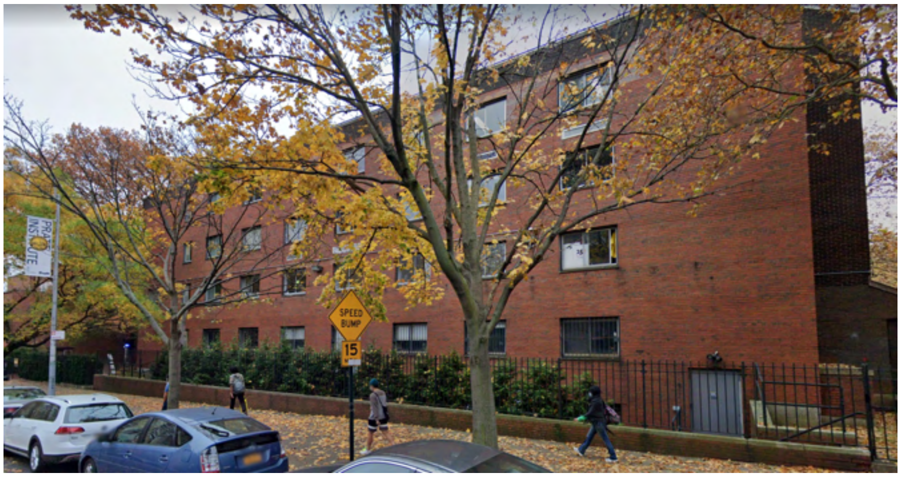


This property is located at 501 West 17 Street in lower Manhattan and is owned by HLP Properties, LLC. The property has a building class of G6 meaning it is classified as a licensed parking lot. The lot front and lot depth are 380 and 184 feet, respectively, and the one story building on the property is said to have a building front of 8 feet and building depth of 12 feet. The full market value is $8,005,000.00 and the total assessed value of the property is $3,602,250.00. Further investigation shows the assessed value of the land to be $3,600,000.00, meaning the building on the property has an assessed value of $2,250.

We found this to be an interesting record, because at initial glance it seems to be another scenario where the building dimensions are not true. However, this turned out to not be the case. At the time the dataset was last updated (2018), it looks very plausible that this property could have actually been a parking lot. This neighborhood - lower central Manhattan along the Hudson - has been a popular place for development in New York City this century, so it does not come as a surprise that a company could have come along and bought this property.

Regardless, the fraud algorithm, which does not take into account building classification, has assigned this a high rank because of the property's very small building area and volume (which was likely just a small building for the parking attendant). Given its location in lower Manhattan, when grouped with other properties in the area, a small building on land valued/assessed in the multiple millions is destined to be an outlier.

Record #8: 330291



This property is located at 166 Willoughby Avenue in Brooklyn and is owned by Pratt Institute. This property has a building classification of W3, meaning it is a school or academy. The property has a lot front of 200 feet, lot depth of 698 feet, and a full market value of $13,500,000.00. The building on the property has 4 stories and is listed to have a building front and building depth of 6 and 8 feet respectively. The land value is assessed to be $1,764,000 and the total assessed value is $6,075,000.

Fraud algorithms flagged this property because of its incredibly small building area and volume. When these metrics normalize market value and assessed values, the resulting score is much greater than the mean even when grouping by geography and tax class.

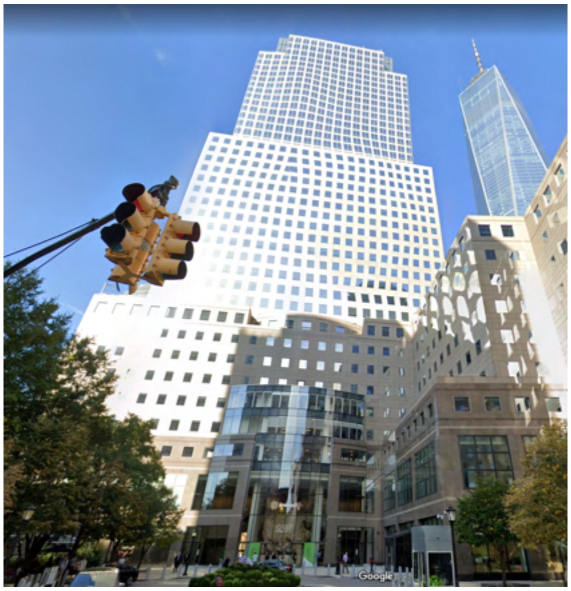
We are intrigued by this record because the owner is the Pratt Institute, a private university. The picture above clearly shows the building is of larger dimensions than 6 by 8.

Record #9: 56136

This property is located at 1111 Avenue Of The Americas in Manhattan and is owned by One Bryant Park. This property has a building classification of O4 which corresponds to an office (only) building with 20 or more stories. The lot front and lot depth of the property are 201 and 437 feet respectively, and the property is listed to have 55 stories - which checks out with its O4 classification. The full market value is $1,120,000,000.00 and the total assessed value is $504,000,000.00, while the assessed value of the land is $99,000,000.00. The building front and depth were missing and were imputed with a front listed of 62.08 feet and building depth of 86.30 feet.

We chose this record because we recognized the address and wanted to further investigate why our algorithm gave the Bank of America Tower a high fraud rank. We determined that this property received a high rank because of its small building area and volume. Very similar to many of the other records in this section, the BoA Tower is assigned a high fraud rank because of its incredible $1.12 billion market value and high assessment value while its building dimensions are listed to be very small.

Record #10: 428



This property is located at 225 Liberty Street in lower Manhattan and is owned by WFP Tower B CO LP. Similar to the property above, it also has a building classification of O4 (office building with 20+ stories), but its building front and building depth are 125 feet for both values. The lot front is listed as 489 feet and the lot depth is listed at 217 feet. The full market value of this property is $417,000,000.00. The total assessed value of the property is $187,650,000.00 and the assessed value of the land $76,950,000.00.

While the building dimensions for this highrise are greater than those listed for the BOA Tower (record #9), these dimensions are still too small and still cause the algorithm to identify this record as an outlier. The high fraud rank is caused by a high valuation and assessment, but low building area volume. Given the picture above, even the listed lot size looks as if may be false.

7 Summary and Conclusions

For this project we were tasked with identifying potential fraudulent property valuations, evaluations, and assessments in New York City property data. The data we used came directly from the local government of New York City and represented over 1,000,000 NYC property assessments for the purpose of calculating property tax, determining grant eligible properties, exemptions, and abatements. The data was collected and entered into the system by various city employees, such as property assessors, property exemption specialists, ACRIS reporting, department of building reporting, and many others.

The first step in building our fraud algorithms was exploring the data to determine which fields would be relevant to us during this project. To do this, we constructed a Data Quality Report (DQR) and analyzed every field in the dataset. Our focus during this process was placed on understanding the importance and feasibility of each field in determining the valuation of a property. We determined that some fields, such as ‘EXTOT2’, ‘EXCD2’, and ‘VALTYPE’ among others, were not going to provide us much information in determining fraudulent properties, whereas other fields did not have enough records to be of use to us. From this process we determined that to best identify fraudulent records, we needed to find unusual valuations on the fields ‘FULLVAL’, ‘AVLAND’, and ‘AVTOT’.

The next step, once we fully understood the dataset, was to clean the fields that we would be using moving forward in the construction of our fraud algorithms. The highlight of this process included: removing all properties with specific owners given to us by the client and filling in all zip codes and tax classes for the records.

With the data cleaned, we could now begin building our special, *expert*, variables. The goal of these expert variables is to best describe the underlying relationship between the various fields in a record. These variables should quantify the fraudulent characteristics of each record. Odd (very high or low) values of these variables will be able to be used to identify outlier properties with certain characteristics that stand out from similar properties. We created a total of 45 new variables to be used in our algorithms - all of which were focused on highlighting abnormalities in ‘FULLVAL’, ‘AVLAND’, and ‘AVTOT’. To build these 45 variables, we consulted with a NYC tax *guru*. Through his guidance we normalized ‘FULLVAL’, ‘AVLAND’, and ‘AVTOT’ by lot area, building area, and building volume. Then grouped the records by full zip code, the first 3 digits of the zip code, tax class, and borough.

Once they were built, the next process was to remove all correlations between the variables to find the lowest dimensionality possible through Principal Component Analysis. Due to varying units, it was imperative we z-scaled our data before and after the PCA. We could now begin building our fraud algorithms. It is important to note that the dataset does not contain any labeled fraudulent records, meaning that we do not have any known instances of fraud in the data. Thus, our fraud models will be unsupervised in attempting to identify anomalies.

We built our fraud algorithm using two methods to find outliers and then took the average of both methods to determine the final fraud rank for each record. The first method we used was very simple. We calculated the Minkowsley distance (p=2) to find z-score outliers and the fraud score was just the Minkowskey distance of each z-score. The second method was slightly more complex, as we used an autoencoder on the scaled principal components and the fraud score assigned was the reconstruction error. We then scaled these two fraud scores and calculated each record's final fraud score by taking the average of its two scaled scores. Now that every record had a Final Score, we began investigating the properties with the highest fraud scores.

Through our investigation of the records with the top fraud scores, we found a few underlying themes that our algorithm highlighted. If we were to continue with this project we would be sure to build upon and resolve these discrepancies:

1. Our algorithm seems to have done a decent job identifying instances in which the building area or lot area is suspicious given the valuation and assessment of the property. That being said, we believe it is reasonable to assume that many of the records we investigated did not actually have the values of building front/depth, lot front/depth, or valuation/assessment that the expert variables used for calculations. The manner in which we cleaned our data and filled in zeros in these columns likely resulted in this. In relation to building front/depth, filling in zeros with the average value based on tax group can create awkward instances, such as what happened with Yankee Stadium (Record #5 above).

2. We are also concerned with the implications of typing errors. Such as what may have happened with the Pratt Institute (record #8) or record #10. If there is a way to make sure that the input values make common sense before submitting it to the database, we would be able to see a reduction in "outliers" that were mistyped.

3. Improper addresses and missing information makes it difficult to validate the property and explore what exactly is being misrepresented. Many of the top ranked records had only a street name and were missing the majority of their information. This makes these records arduous to understand and relies on our own abilities to intuitively fill in the missing information during data cleaning - which we have already experienced is tasking to do and may still lead to inaccurate information.

Appendix

Data Quality Report

Description

Dataset Name: Property Valuation and Assessment Data
Dataset Purpose: Data represents NYC properties assessments for the purpose to calculate Property Tax, Grant eligible properties Exemptions, and/or Abatements.
Data Source: NYC Open Data
Time Period: November 17, 2010
Number of Fields: 32
Number of Records: 1,070,994

Summary Tables

Numeric Fields

Field Name	# Of Records	Percent Populated	Unique Values	Mean	Standard Deviation	Minimum Value	Maximum Value	# Zeros
LTFRONT	1070994	100%	1297	36.64	74.03	0	9999	169108
LTDEPTH	1070994	100%	1370	88.86	76.4	0	9999	170128
STORIES	1014730	94.75%	112	5.01	8.37	1	119	0
FULLVAL	1070994	100%	109324	874264.51	11582430.99	0	6150000000	13007
AVLAND	1070994	100%	70921	85067.92	4057260.06	0	2668500000	13009
AVTOT	1070994	100%	112914	227238.17	6877529.31	0	4668308947	13007
EXLAND	1070994	100%	33419	36423.89	3981575.79	0	2668500000	491699
EXTOT	1070994	100%	64255	91186.98	6508402.82	0	4668308947	432572
EXCD1	638488	59.62%	130	1602.01	1384.23	1010	7170	0
BLDFRONT	1070994	100%	612	23.04	35.58	0	7575	228815

BLDDEPTH	1070994	100%	621	39.92	42.71	0	9393	228853
AVLAND2	282726	26.40%	58592	246235.72	6178962.56	3	2371005000	0
AVTOT2	282732	26.40%	111361	713911.44	11652528.95	3	4501180002	0
EXLAND2	87449	8.17%	22196	351235.68	10802212.67	1	2371005000	0
EXTOT2	130828	12.23%	48349	656768.28	16072510.17	7	4501180002	0

Categorical Fields

Field Name	# Of Records	% Populated	Unique Values	Most Common Field Value
RECORD	1070994	100%	1070994	N/A
BBLE	1070994	100%	1070994	N/A
B	1070994	100%	5	4
BLOCK	1070994	100%	13984	3944
LOT	1070994	100%	6366	1
EASEMENT	4636	.43%	13	E
OWNER	1039249	97.04%	863348	PARKCHESTER PRESERVAT
BLDGCL	1070994	100%	200	R4
TAXCLASS	1070994	100%	11	1
EXT	354305	33.08%	4	G
ZIP	1041104	97.21%	197	10314
EXCD1	638488	59.62%	130	1017
STADDR	1070318	99.84%	839281	501 SURF AVENUE
EXMPTCL	15579	1.45%	15	X1
EXCD2	92948	8.68%	61	1017
PERIOD	1070994	100%	1	FINAL
YEAR	1070994	100%	1	2010/11

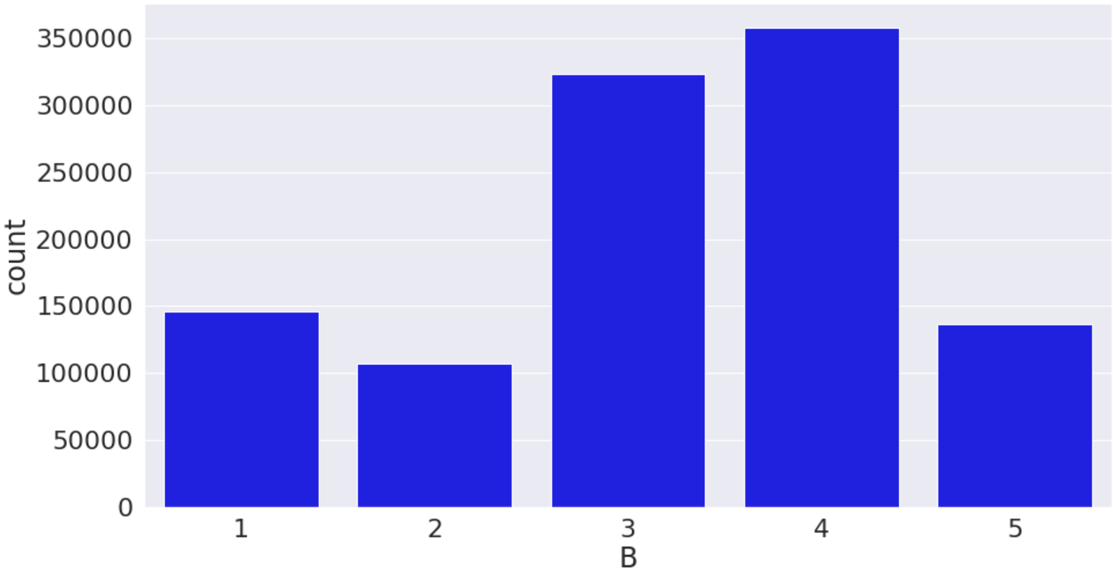
VALTYPE	1070994	100%	1	AC-TR

Data Field Exploration

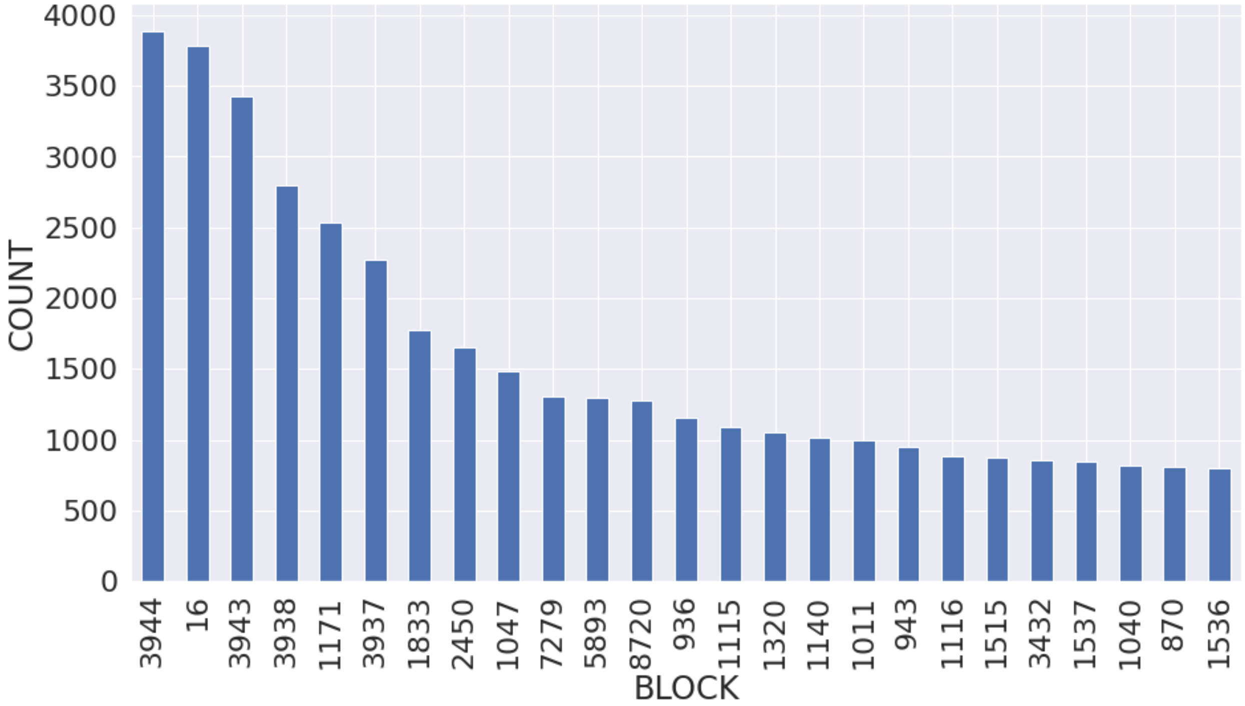
Field 1:
Name: Record
Description: Unique identifier of each entry in the dataset.

Field 2:
Name: BBLE
Description: Concatenation of borough code, block code, lot, and easement classification.

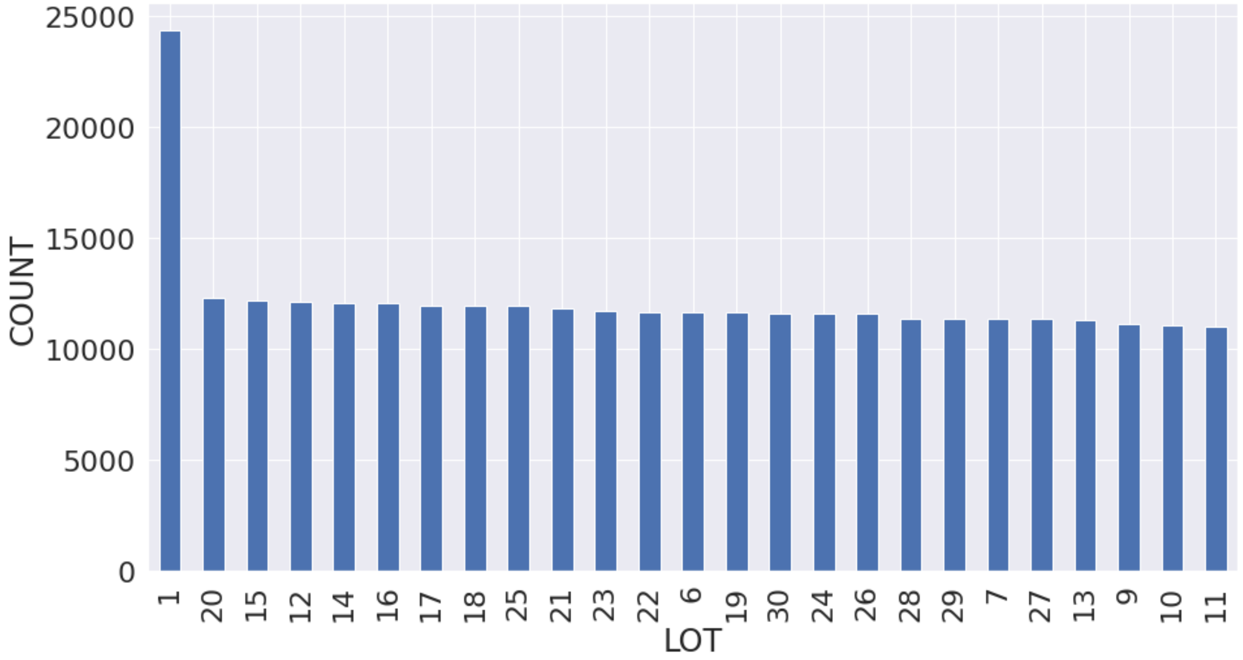
Field 3:
Name: B
Description: Borough code.



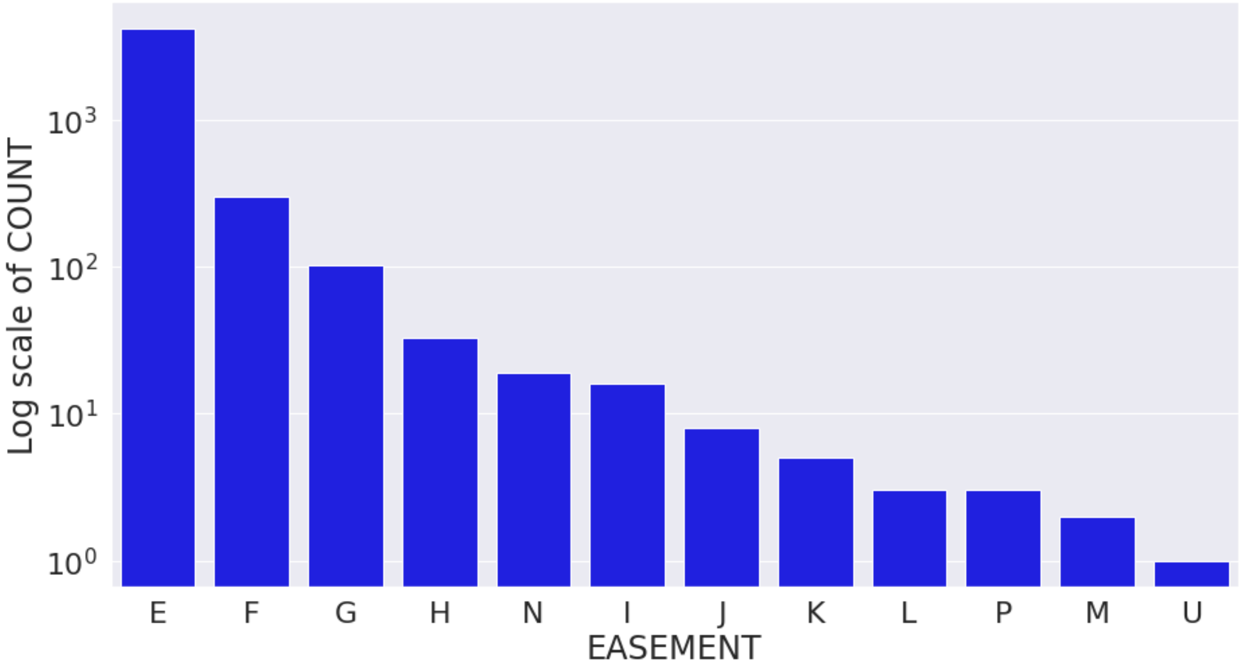
Field 4:
Name: BLOCK
Description: Valid block ranges by borough. Manhattan 1 to 2255, Bronx 2260 to 5958, Brooklyn 1 to 8955, Queens 1 to 16350, Staten Island 1 to 8050.



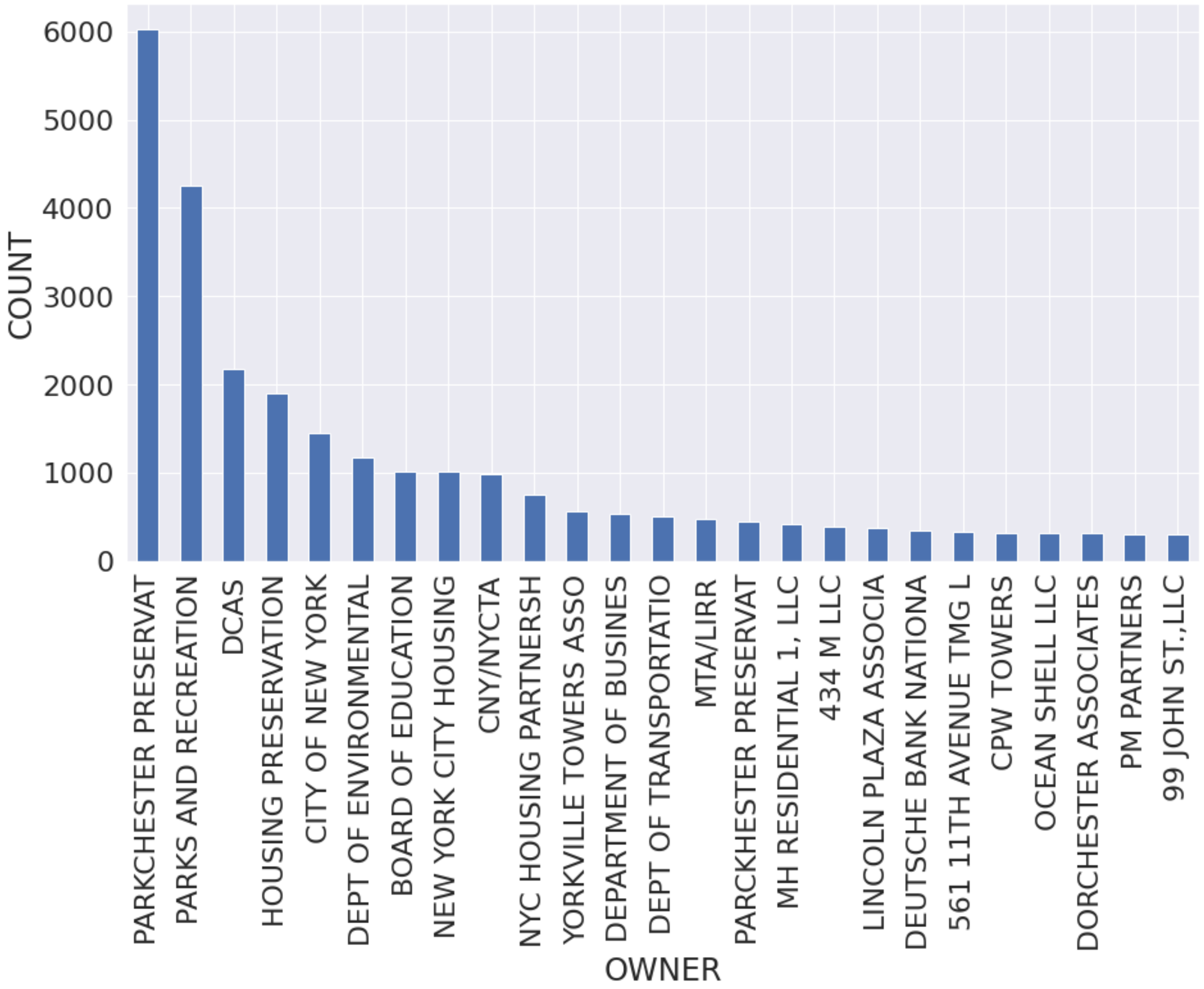
Field 5:
Name: LOT
Description: Unique number within borough/block.



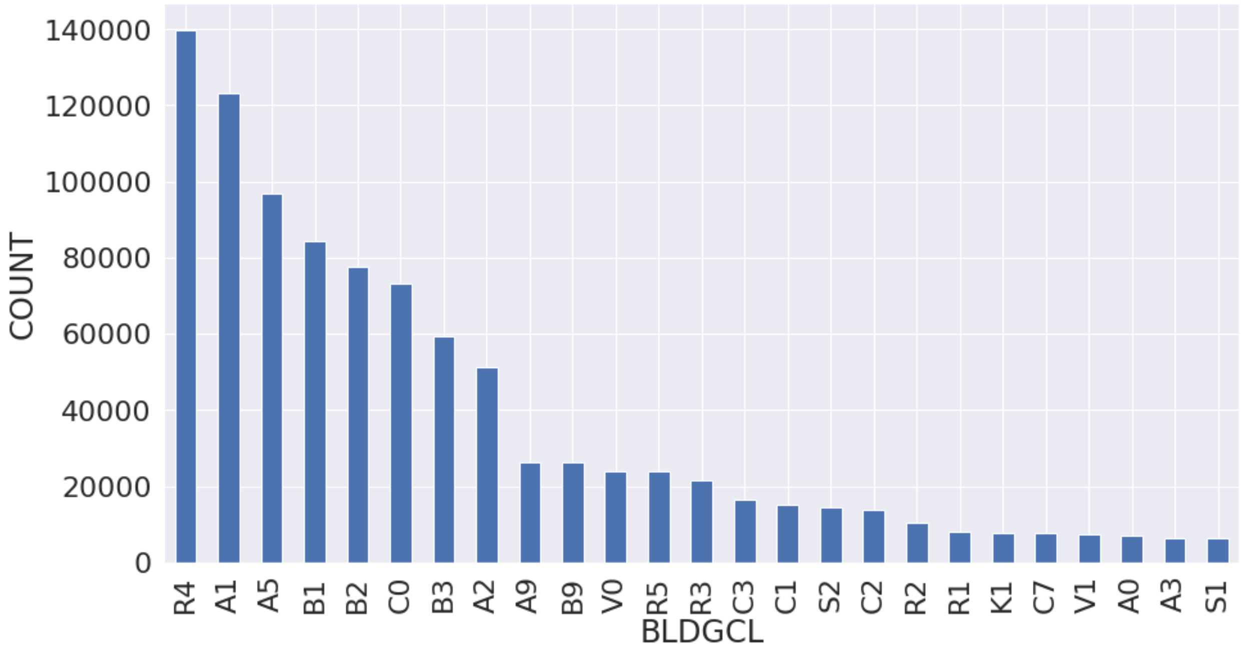
Field 6:
Name: EASEMENT
Description: Field that is used to describe easement.
'A' Indicates the portion of the Lot that has an Air Easement.
'B' Indicates Non-Air Rights.
'E' Indicates the portion of the lot that has a Land Easement.
'F' THRU 'M' Are duplicates of 'E'.
'N' Indicates Non-Transit Easement.
'P' Indicates Piers.
'R' Indicates Railroads.
'S' Indicates Street.
'U' Indicates U.S. Government.



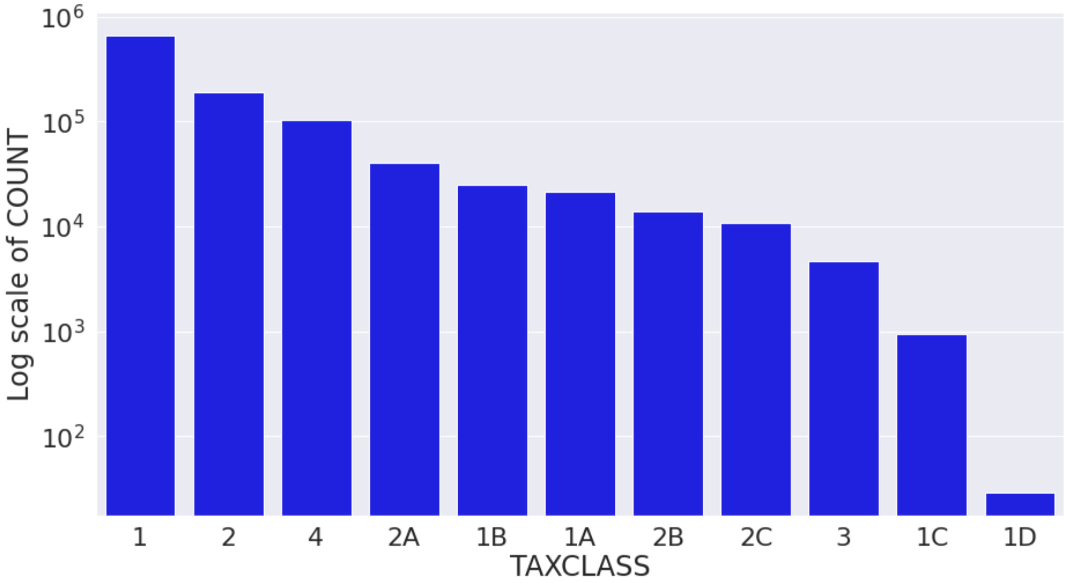
Field 7:
Name: OWNER
Description: Owner’s name.



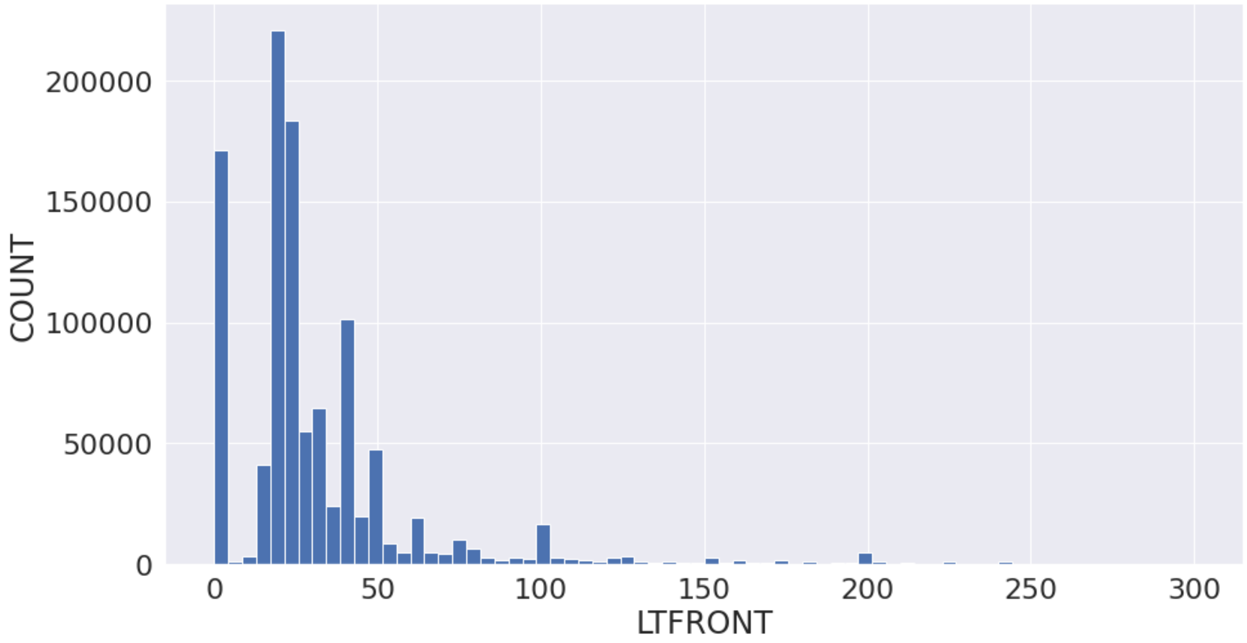
Field 8:
Name: BLDGCL
Description: Building Class.



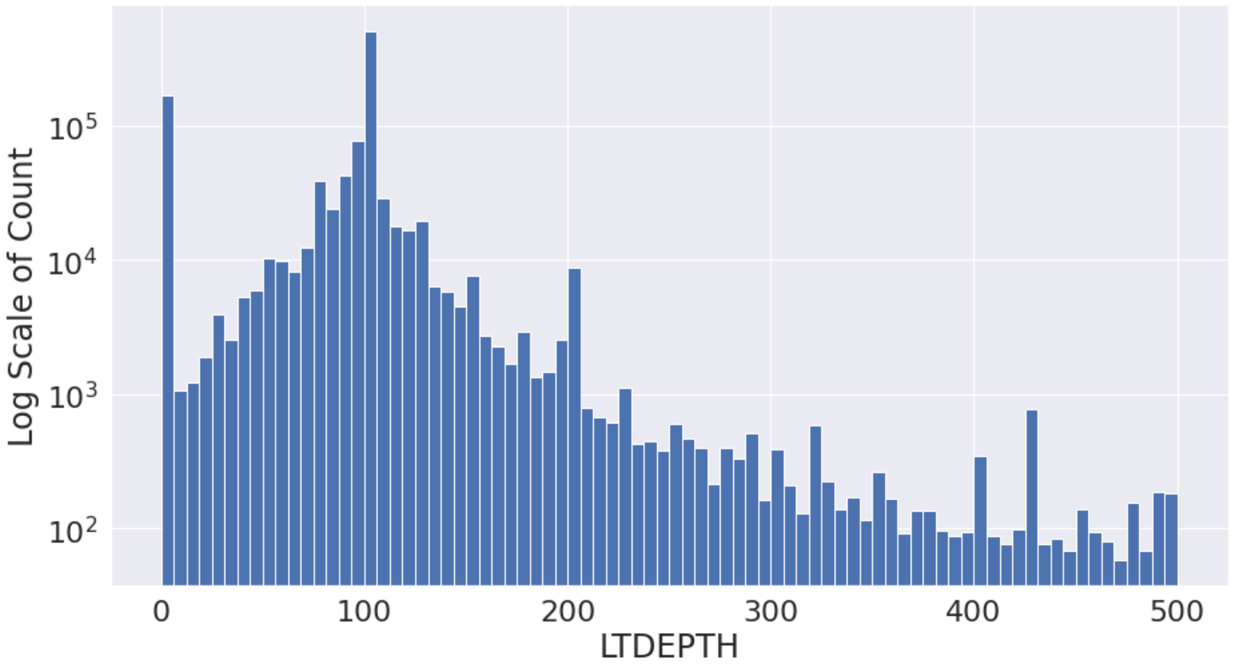
Field 9:
Name: TAXCLASS
Description: Current Property Tax Class Code (NYS Classification).
Tax Class 1 = 1-3 Unit Residences
Tax Class 1A = 1-3 Story Condominiums
Tax Class 1B = Residential Vacant Land
Tax Class 1C = 1-3 Unit Condominums
Tax Class 1D = Select Bungalow Colonies
Tax Class 2 = Apartments
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



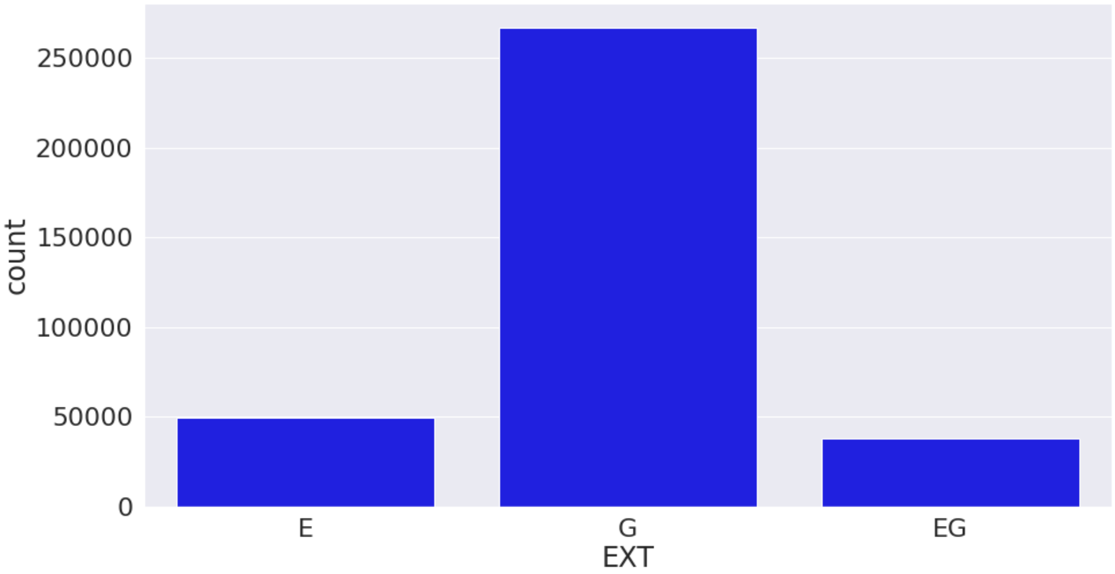
Field 10:
Name: LTFRONT
Description: Lot frontage in feet.
Exclude outliers > 300, data in histogram is 99.30 % populated.



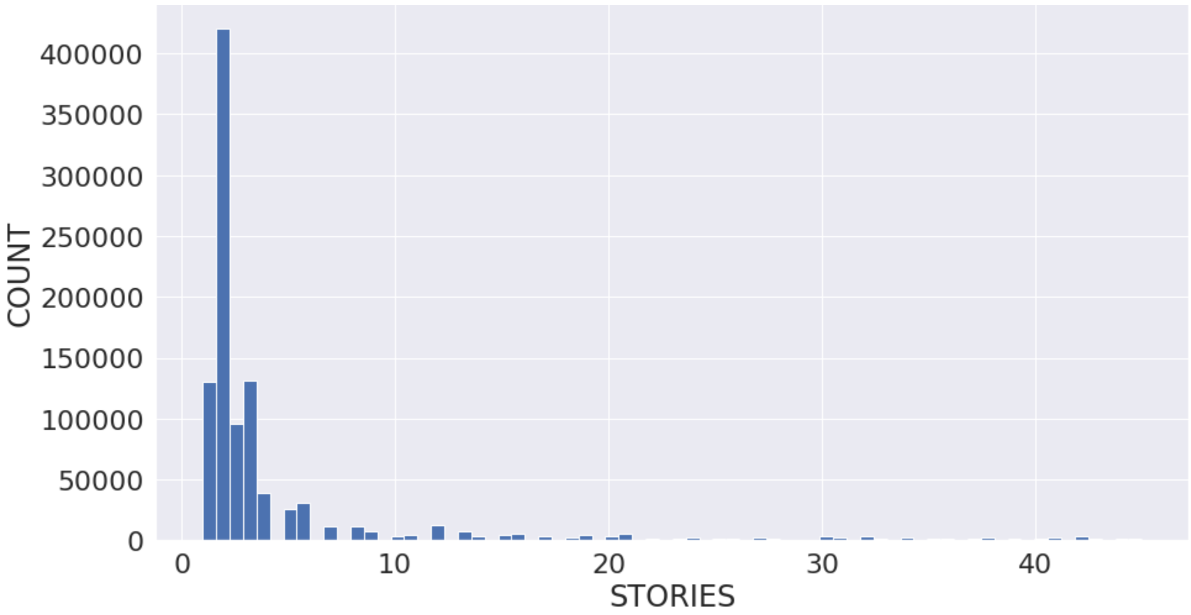
Field 11
Name: LTDEPTH
Description: Lot depth in feet.
Exclude outliers > 500, data in histogram is 99.68 % populated.



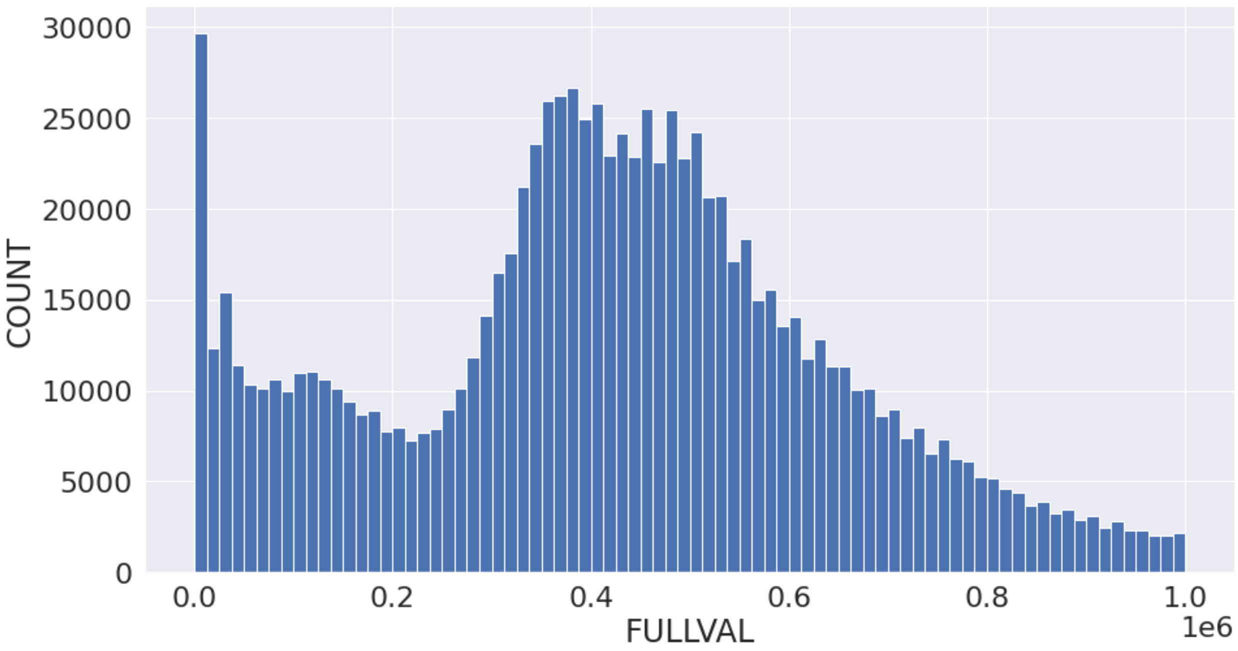
Field 12
Name: EXT
Description: Extension.
'E' = Extension
'G' = Garage
'EG'' = Extension And Garage



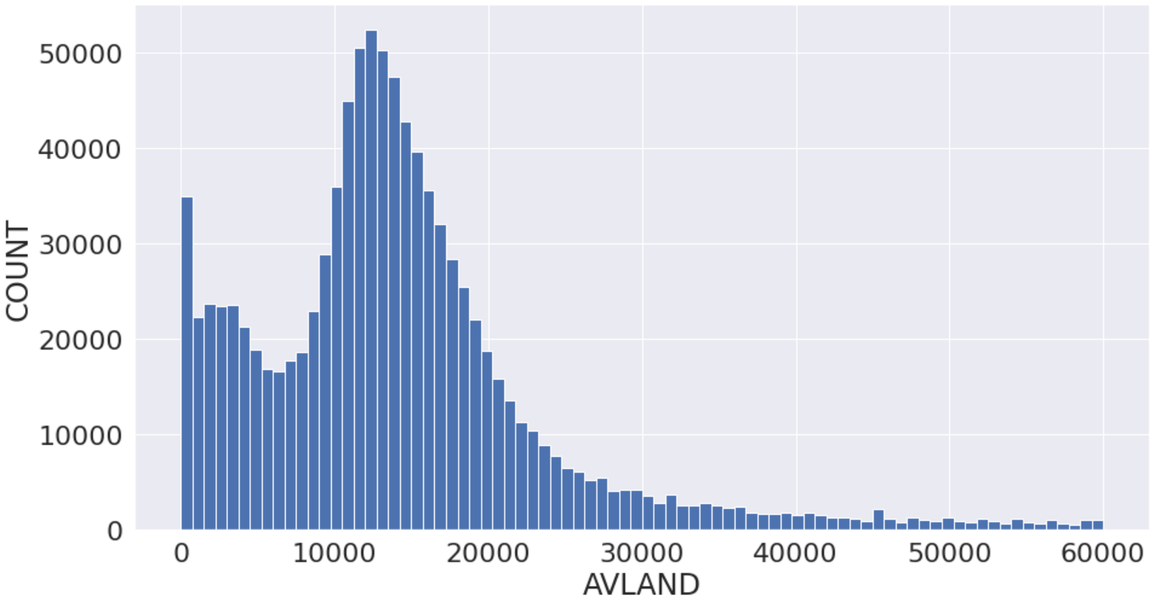
Field 13
Name: STORIES
Description: The number of stories for the building (# of Floors).
Exclude outliers > 45, data in histogram is 99.16 % populated.



Field 14
Name: FULLVAL
Description: Total market value of the land.
Exclude outliers > 1000000, data in histogram is 91.35 % populated.



Field 15
Name: AVLAND
Description: Assessed land value.
Exclude outliers > 60000, data in histogram is 91.61 % populated

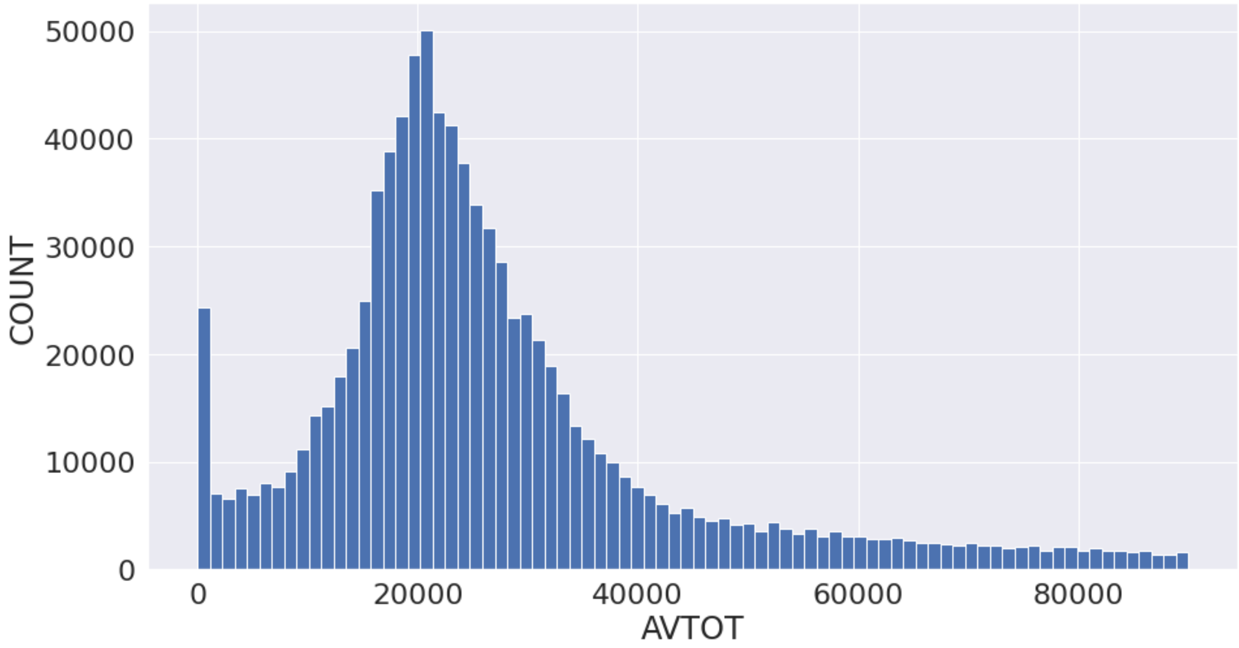


Field 16

Name: AVTOT

Description: Assessed total value.

Exclude outliers > 90000, data in histogram is 84.98 % populated.

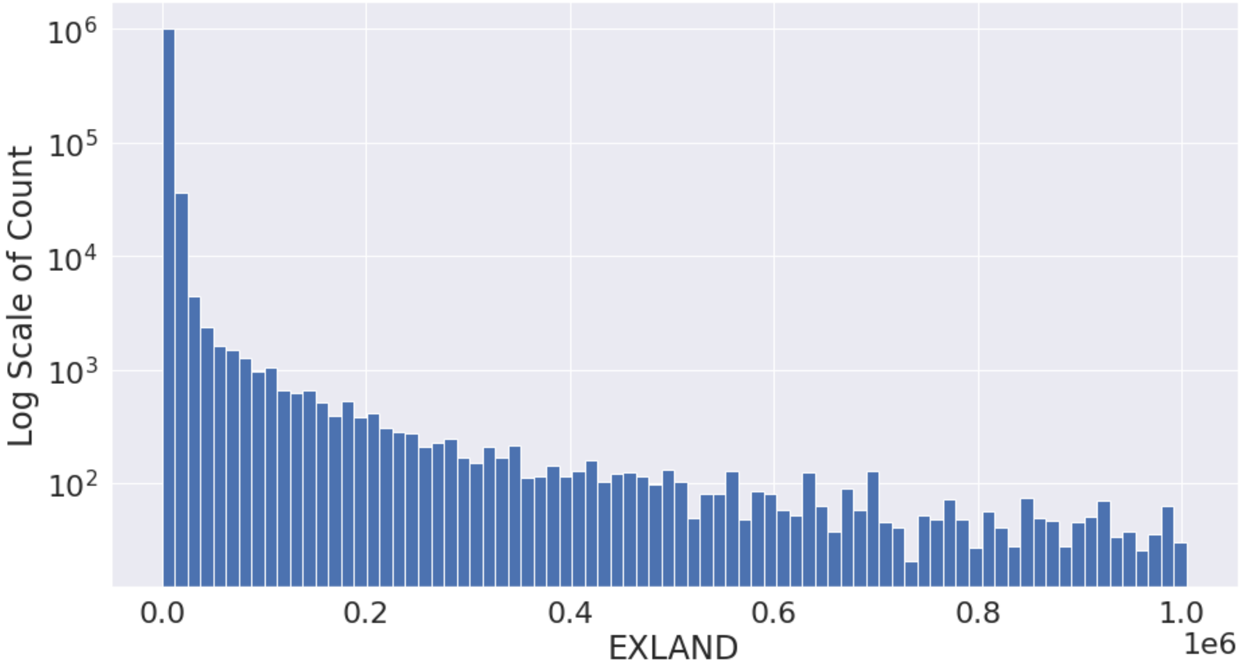


Field 17

Name: EXLAND

Description: Exempt land value.

Exclude outliers > 100500, data in histogram is 99.62% populated.

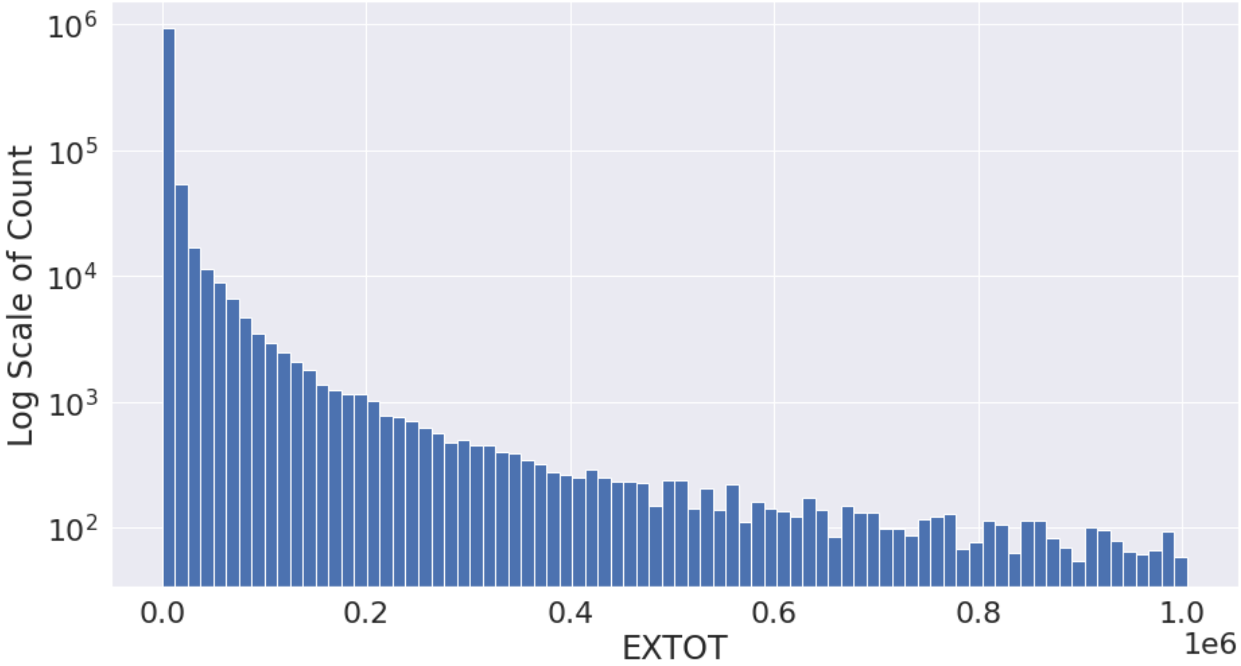


Field 18

Name: EXTOT

Description: Exempt total value.

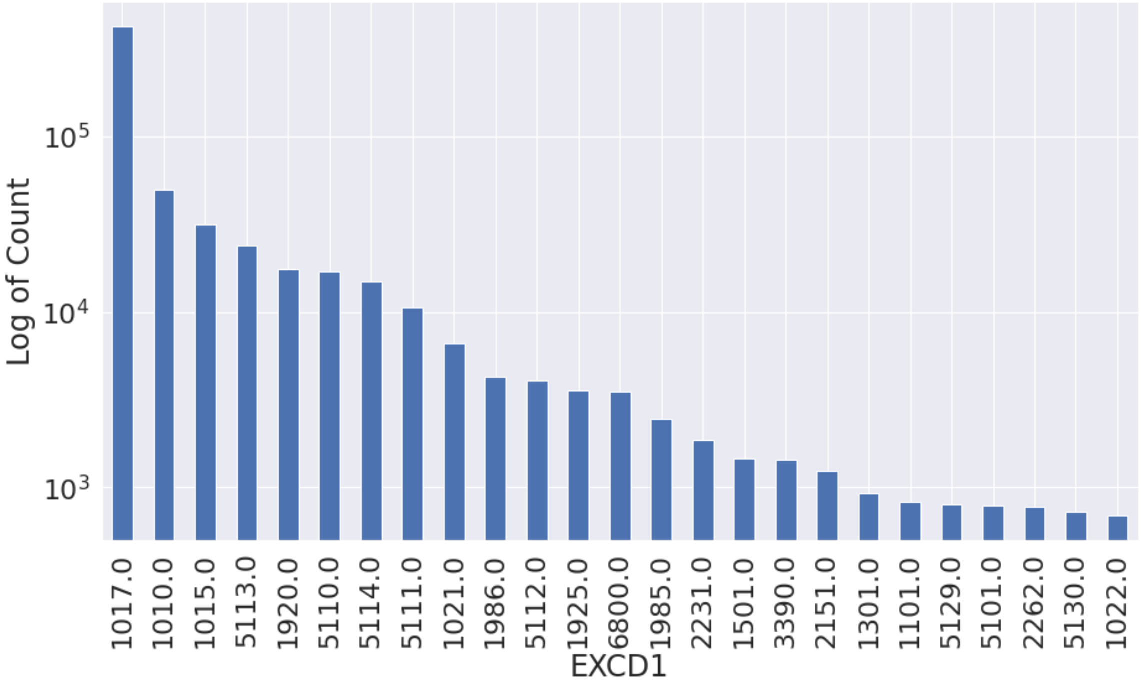
Exclude outliers > 100500, data in histogram is 99.16 % populated.



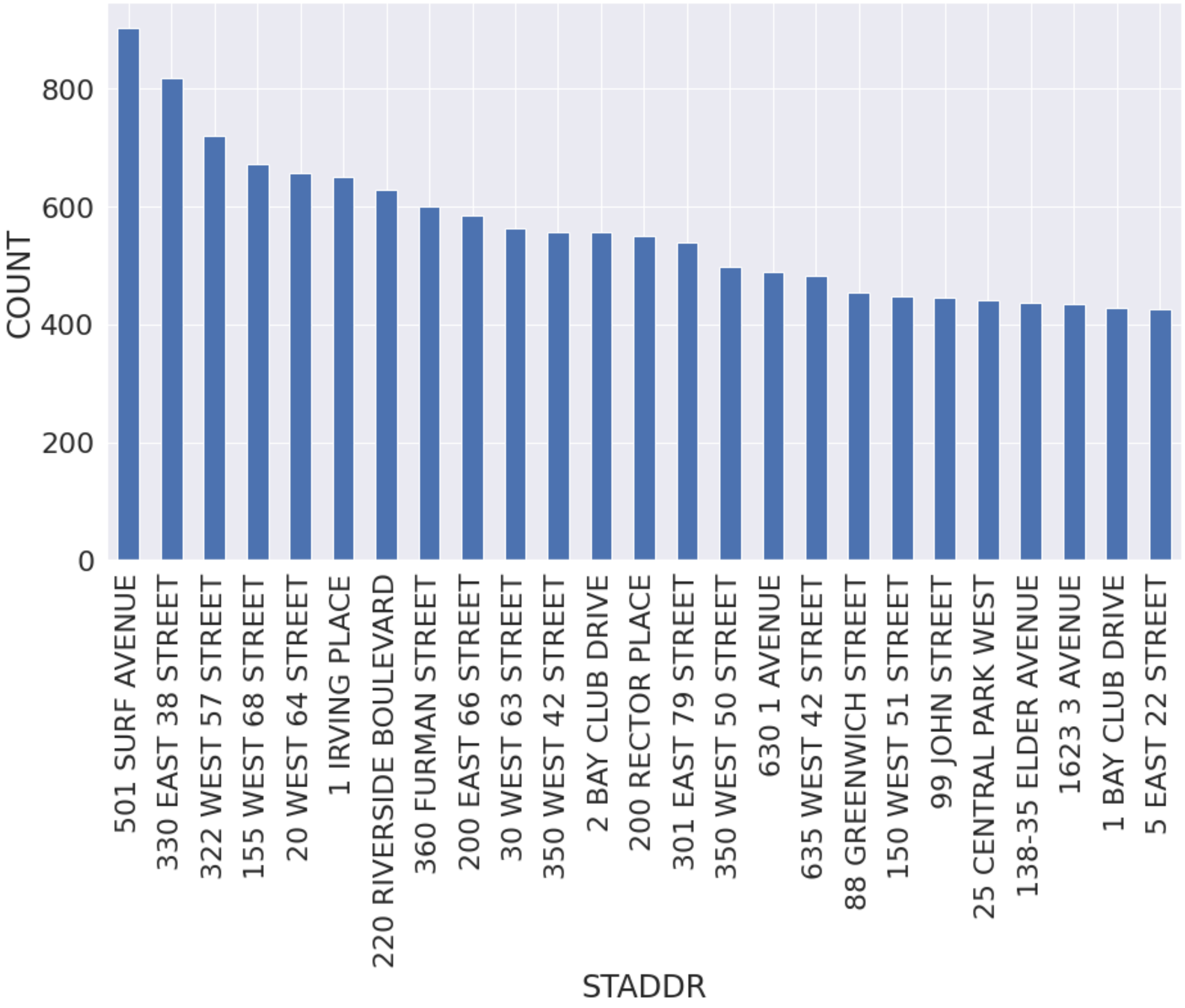
Field 19

Name: EXCD1

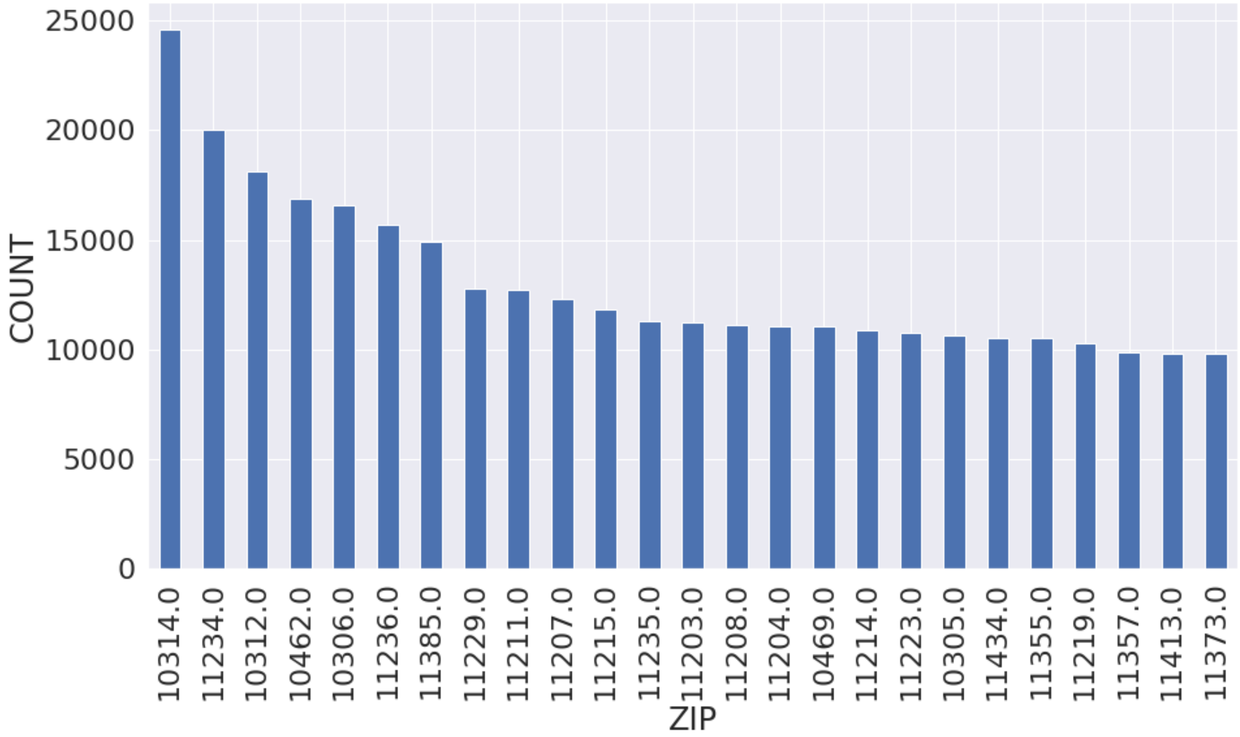
Description: (Not enough information on this field).



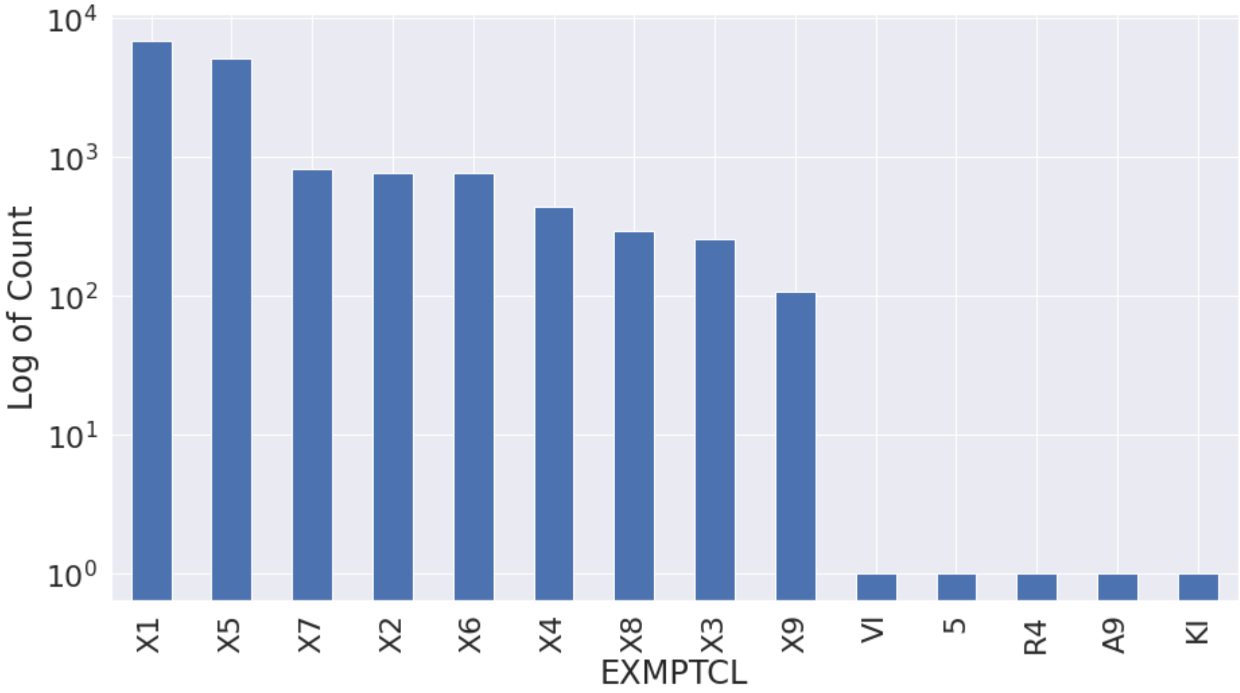
Field 20:
Name: STADDR
Description: Street name for the property.



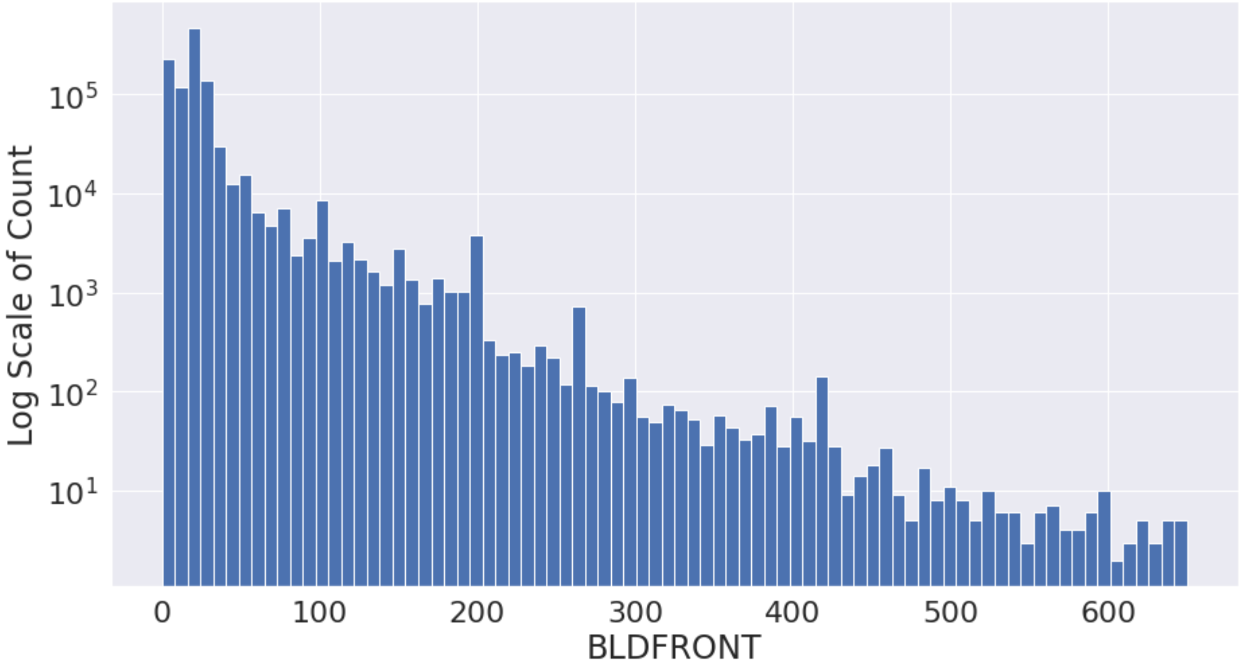
Field 21:
Name: ZIP
Description: Zip code of the property



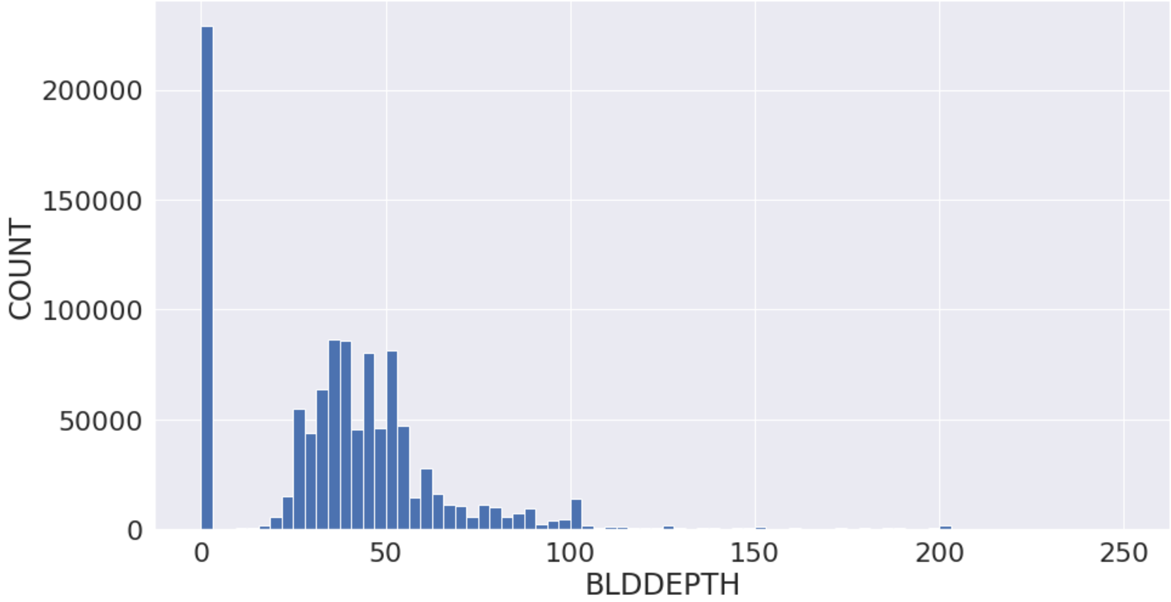
Field 22:
Name: EXMPTCL
Description: Exempt Class used for fully exempt properties only.



Field 23:
Name: BLDFRONT
Description: Building Frontage in feet.
Exclude outliers > 650, data in histogram is 99.99 % populated.



Field 24:
Name: BLDDEPTH
Description: Lot depth in feet.
Exclude outliers > 300, data in histogram 99.72 % populated.

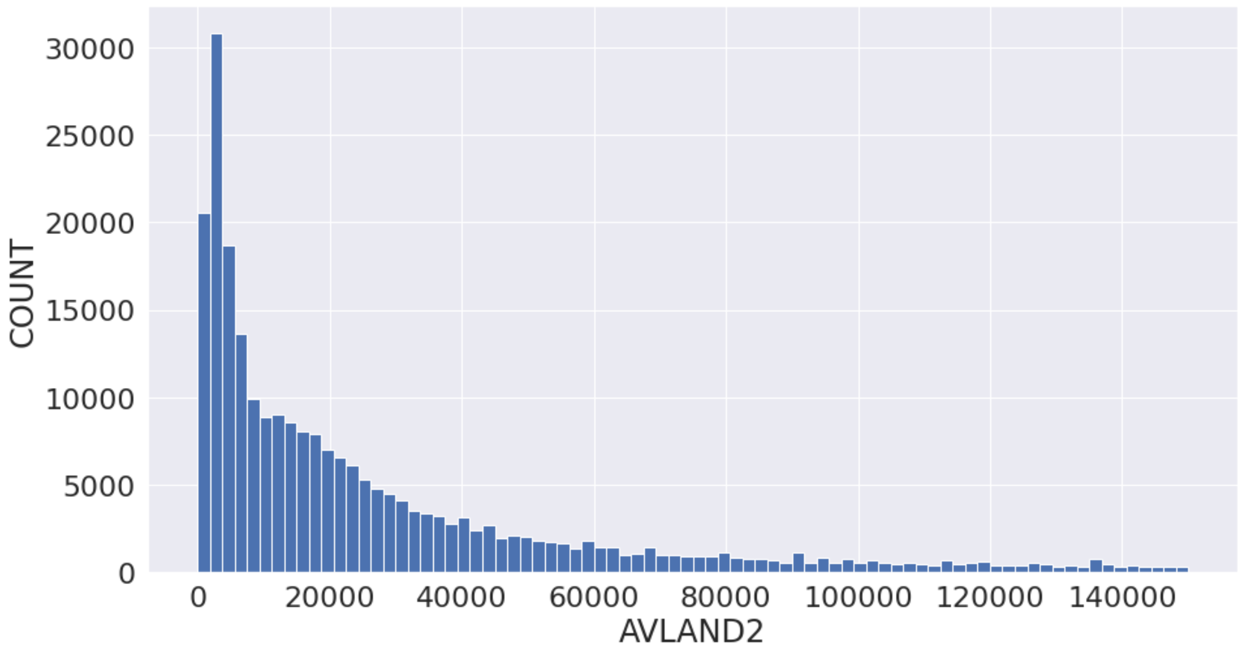


Field 25:

Name: AVLAND2

Description: New market value of land.

Exclude outliers > 150000, data in histogram is 85.64 % populated.

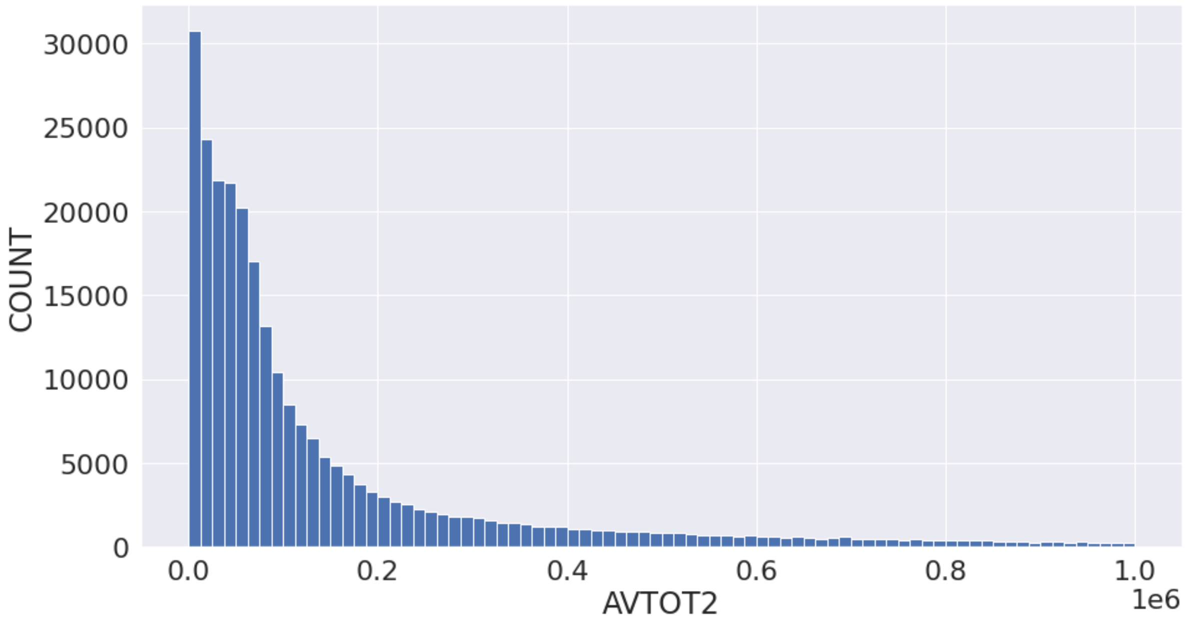


Field 26:

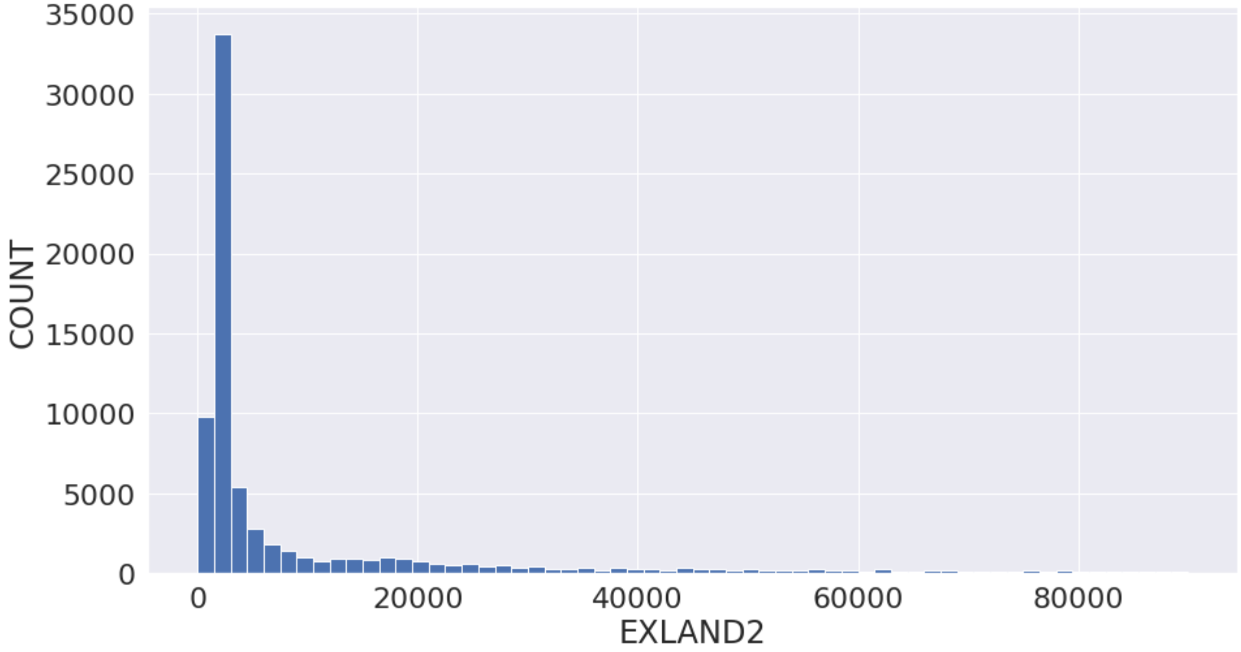
Name: AVTOT2

Description: New total market value.

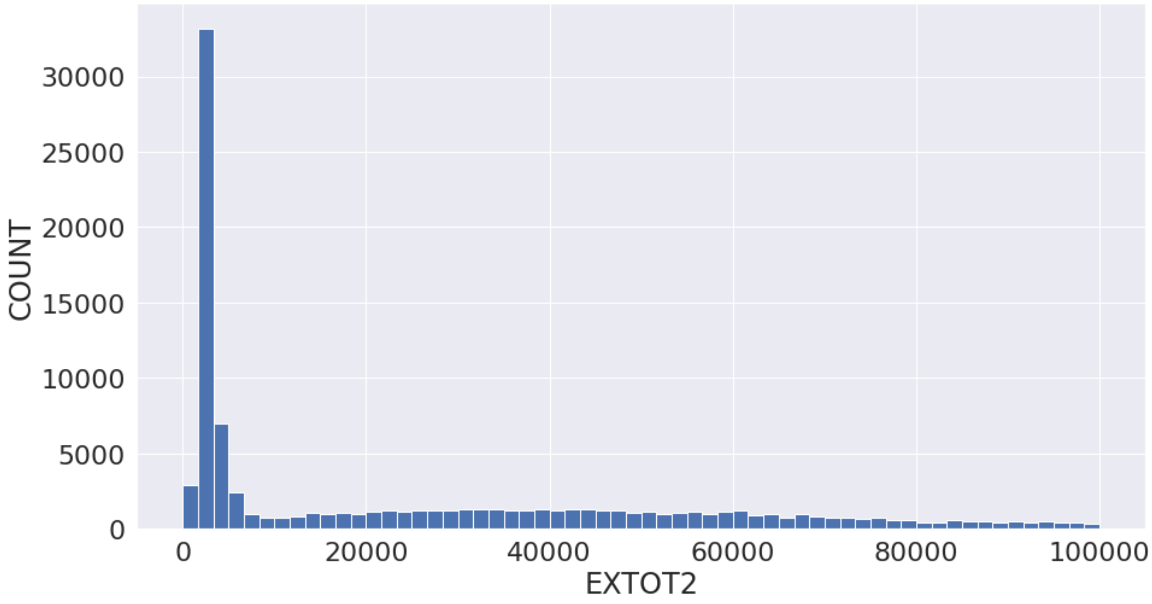
Exclude outliers > 1000000, data in histogram is 91.61 % populated.



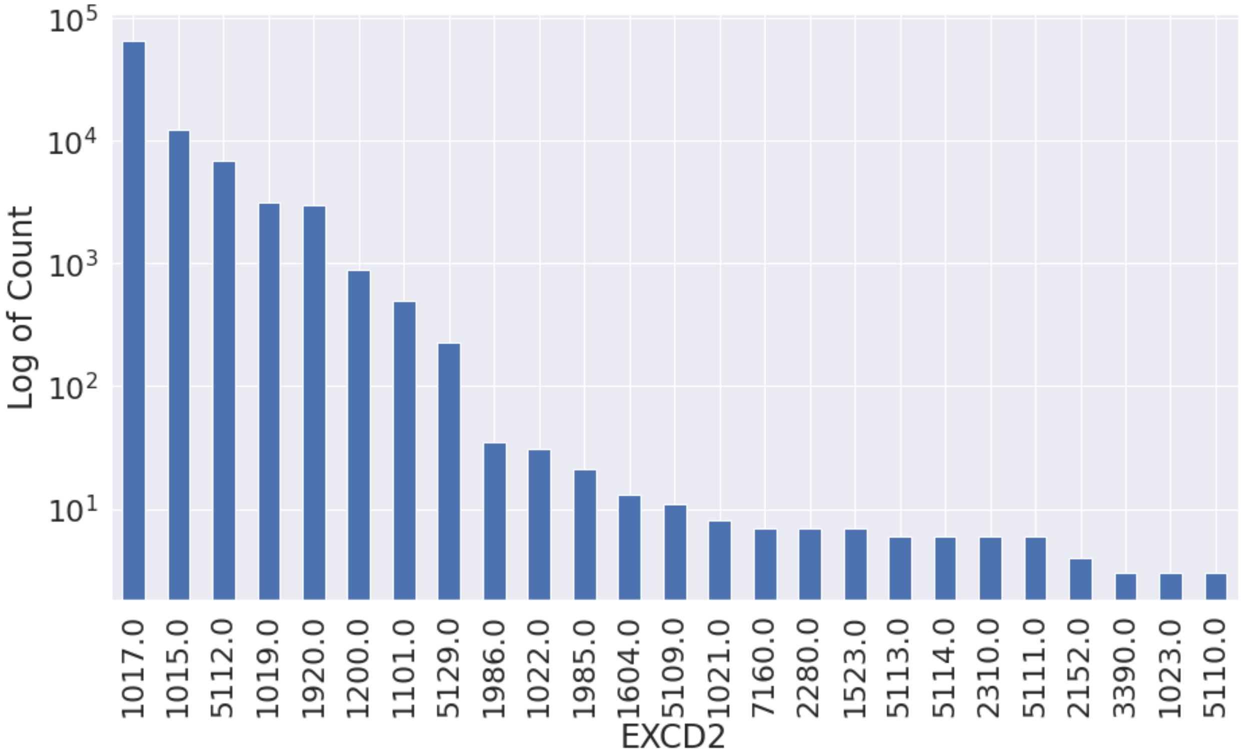
Field 27:
Name: EXLAND2
Description: New exempt land value.
Exclude outliers > 90000, data in histogram is 83.21 % populated.



Field 28:
Name: EXTOT2
Description: New exempt total value.
Exclude outliers > 100000, data in histogram is 73.92%% populated.



Field 29:
Name: EXCD2
Description: (Not enough information on this field).



Field 30:
Name: PERIOD
Description: Indicator for the Change Period of the File. All records have the same value (FINAL).

Field 31:
Name: YEAR
Description: Month and year when the data was most recently revised. All records have the same value (2010/11).

Field 32:
Name: VALTYPE
Description: Not enough information on this field. All record have the same value (AC-TR).