MGTA 463 Fraud Analytics Project 2

New York Property Tax Record Fraud Detection

Joshua Chen

June 9th, 2024

Table of Contents

Executive Summary

Identifying potential property tax fraud is crucial as it ensures fairness in the taxation system, safeguards vital municipal revenues needed for public services, and reduces administrative and legal challenges. By proactively addressing discrepancies, New York City can maintain public trust and prevent significant revenue losses, ensuring an equitable distribution of tax burdens among property owners.

In our analysis of New York City's real estate data, I thoroughly examined over one million property records to detect inconsistencies in tax assessments. I enriched the dataset by creating additional variables that emphasize property values and sizes, essential for identifying discrepancies. By analyzing these metrics, I was able to assess the likelihood of incorrect tax assessments for each property.

The analysis culminated in a ranked list of properties, prioritized by the potential risk of tax fraud. The top-ranked properties were identified as high priorities for further investigation by tax professionals. This streamlined approach allows for effective detection and management of tax assessment issues, facilitating prompt and accurate resolution.

Data Description

The dataset used for the analysis within this report is a public dataset produced by the NYC Department of Finance (DOF) and provided by NYC Open Data. It consists of property valuations and assessments on NYC real estate, and was intended for calculating property taxes and granting eligible properties exemptions and/or abatements during NYC's 2010/2011 fiscal year.

Although it was originally created in September 2011, the dataset was updated more recently in September 2018. The dataset contains a total of 1,070,994 property records and 32 data fields. Within the 32 data fields, there are 14 numerical data fields and 18 categorical data fields. Summary tables for the numerical and categorical data fields can be seen in Tables 1 and 2, respectively.

Data Summary

Table 1 (numerical fields)

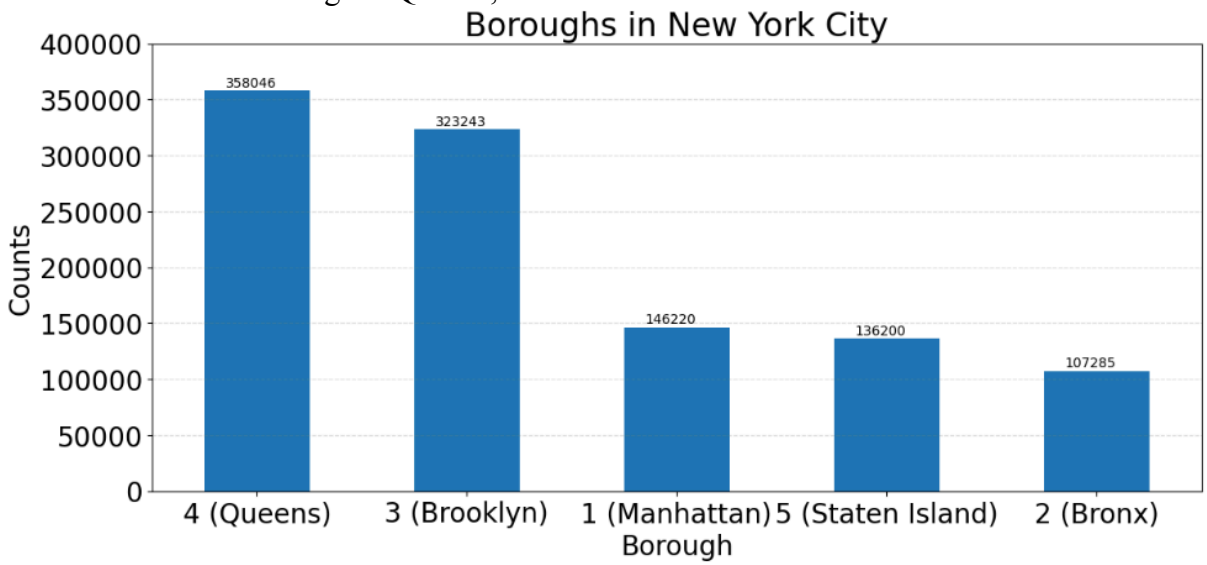
Field Name	# records that have a value	% populated	Most Common	# records with value 0	Min	Max	Mean	Standard Deviation
LTFRONT	1,070,994	100.0%	0	169108	0	9,999	36.64	74.03
LTDEPTH	1,070,994	100.0%	100	170128	0	9,999	88.86	76.40
STORIES	1,014,730	94.7%	2	0	1	119	5.01	8.37
FULLVAL	1,070,994	100.0%	0	13007	0	6,150,000,000	874,264.51	11,582,425.58
AVLAND	1,070,994	100.0%	0	13009	0	2,668,500,000	85,067.92	4,057,258.16
AVTOT	1,070,994	100.0%	0	13007	0	4,668,308,947	227,238.17	6,877,526.09
EXLAND	1,070,994	100.0%	0	491699	0	2,668,500,000	36,423.89	3,981,573.93
EXTOT	1,070,994	100.0%	0	432572	0	4,668,308,947	91,186.98	6,508,399.78
BLDFRONT	1,070,994	100.0%	0	228815	0	7,575	23.04	35.58
BLDDEPTH	1,070,994	100.0%	0	228853	0	9,393	39.92	42.71
AVLAND2	282,726	26.4%	2,408	0	3	2,371,005,000	246,235.71	6,178,951.64
AVTOT2	282,732	26.4%	750	0	3	4,501,180,002	713,911.44	11,652,508.34
EXLAND2	87,449	8.2%	2,090	0	1	2,371,005,000	351,235.68	10,802,150.91
EXTOT2	130,828	12.2%	2,090	0	7	4,501,180,002	656,768.28	16072448.75

Table 2 (categorical fields)

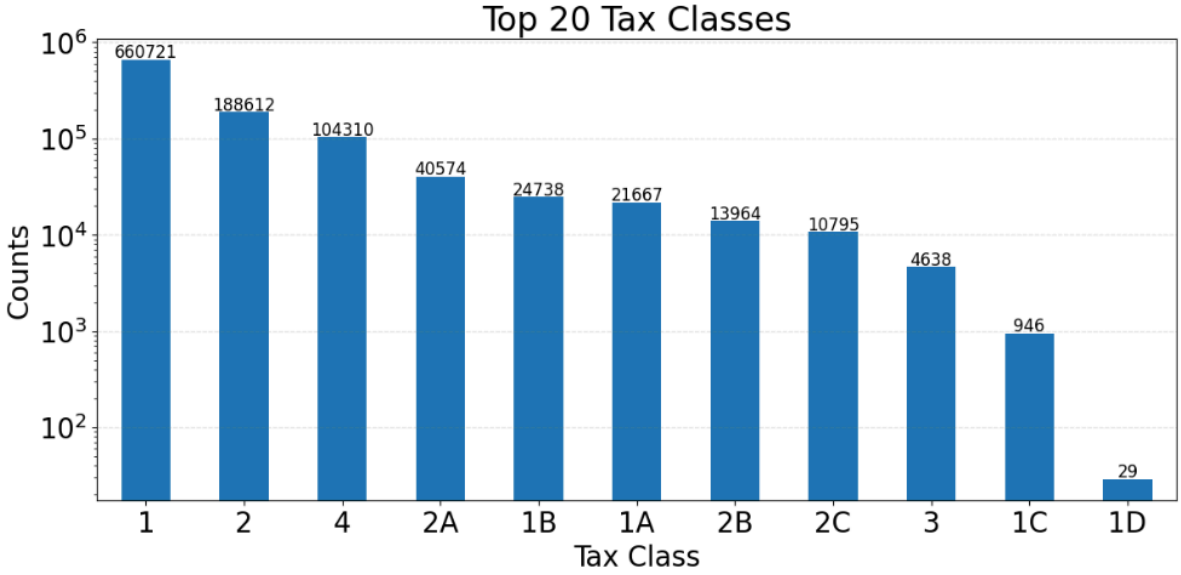
Field Name	# records that have a value	% populated	# unique values	Most common value
RECORD	1,070,994	100.0%	1,070,994	1
BBLE	1,070,994	100.0%	1,070,994	1,000,010,101
BORO	1,070,994	100.0%	5	4
BLOCK	1,070,994	100.0%	13,984	3,944
LOT	1,070,994	100.0%	6,366	1
EASEMENT	4,636	0.4%	12	E
OWNER	1,039,249	97.0%	863,347	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100.0%	200	R4
TAXCLASS	1,070,994	100.0%	11	1
EXT	354,305	33.1%	3	G
EXCD1	638,488	59.6%	129	1,017
STADDR	1,070,318	99.9%	839,280	501 SURF AVENUE
ZIP	1,041,104	97.2%	196	10,314
EXMPTCL	15,579	1.5%	14	X1
EXCD2	92,948	8.7%	60	1,017
PERIOD	1,070,994	100.0%	1	FINAL
YEAR	1,070,994	100.0%	1	2010/11
VALTYPE	1,070,994	100.0%	1	AC-TR

Data Exploration

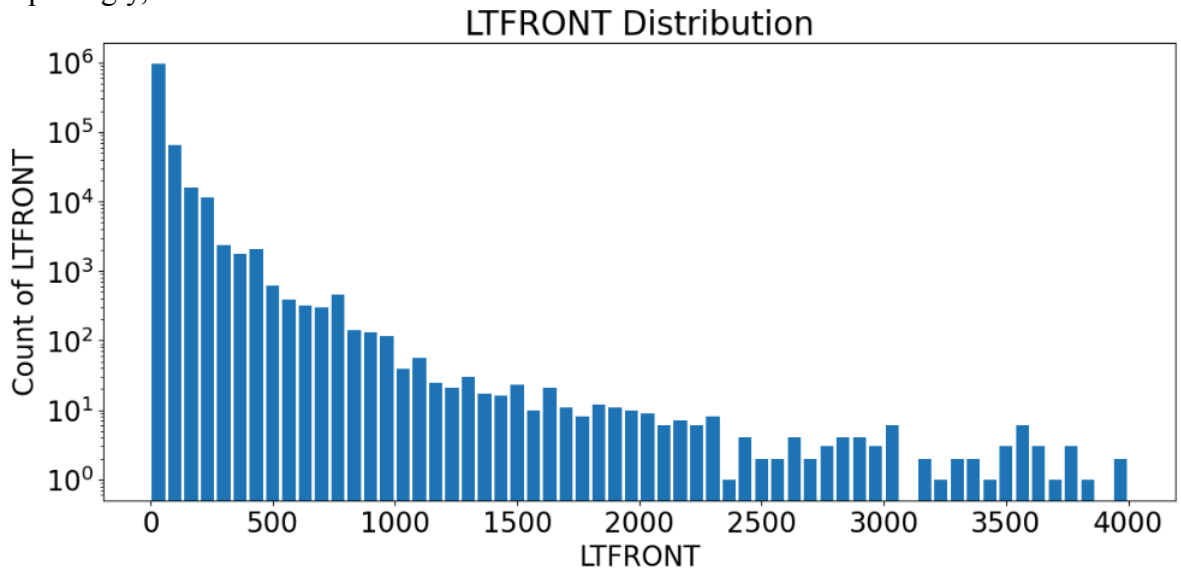
Field Name: BORO
Description: Borough Codes. The plot shows the distribution of values across the 5 boroughs. The most common borough is Queens, with a total count of 358046.



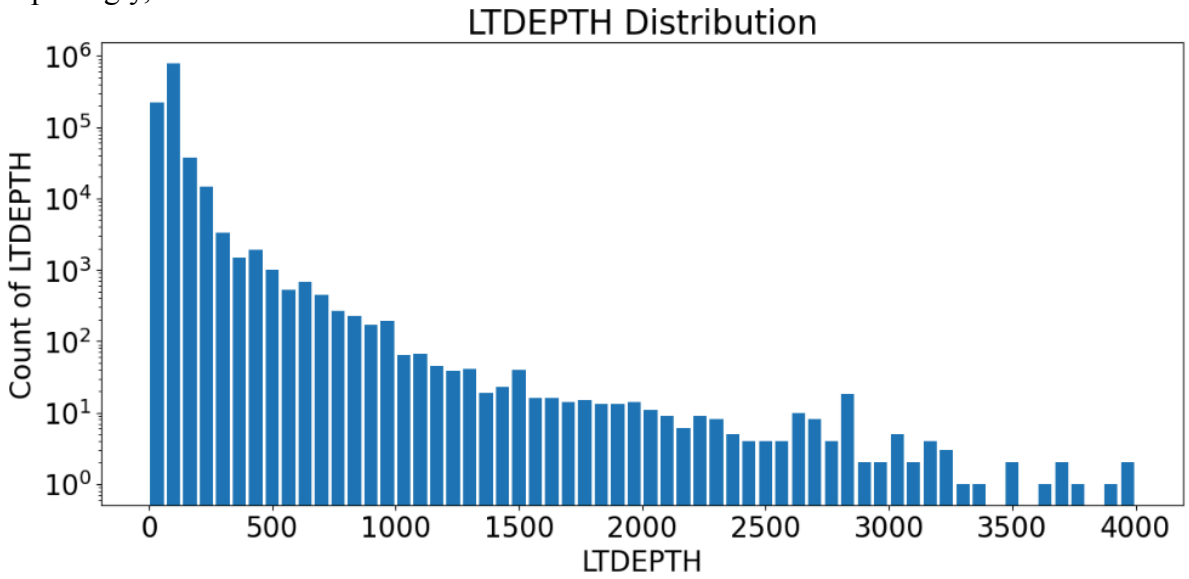
Field Name: TAXCLASS
Description: Current Tax Class Code. 1, 1A, 1B, 1C, 1D = 1-3 Unit Resident. 2, 2A, 2B, 2C = Apartments. 3 = Utilities. 4 = All Others. The plot below shows the distribution of the Taxclass Field. The most common taxclass is 1, with a total count of 660721.



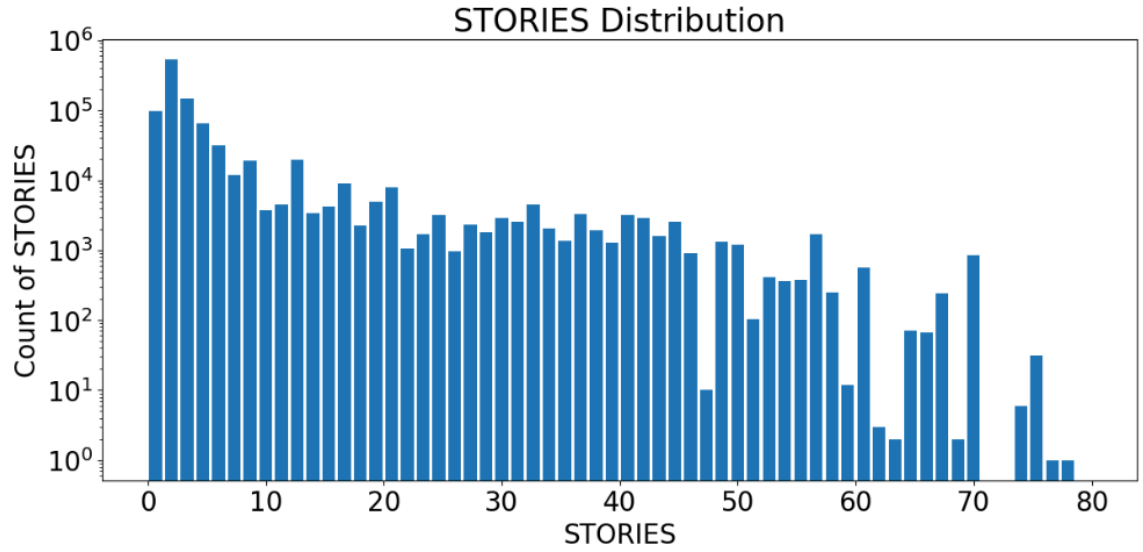
Field Name: LTFRONT
Description: Lot Width. The plot shows the histogram of LTFRONT values from 0 - 4000. Surprisingly, there are a lot of values that have 0 width.



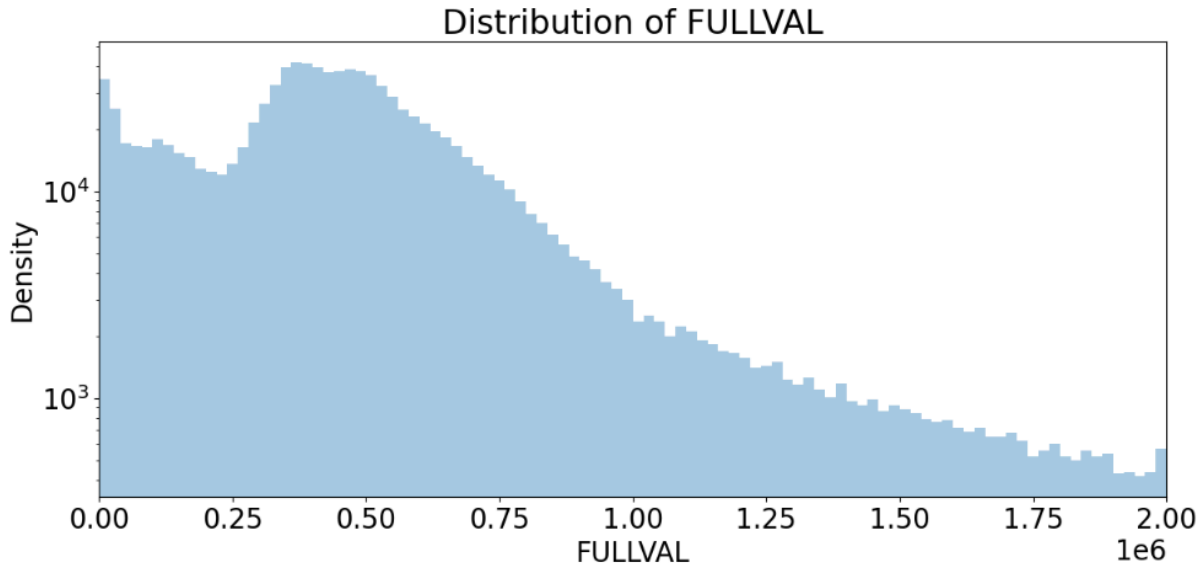
Field Name: LTDEPTH
Description: Lot Depth. The plot shows the histogram of LTDEPTH values from 0 - 4000. Surprisingly, there are a lot of values that have 0 width.



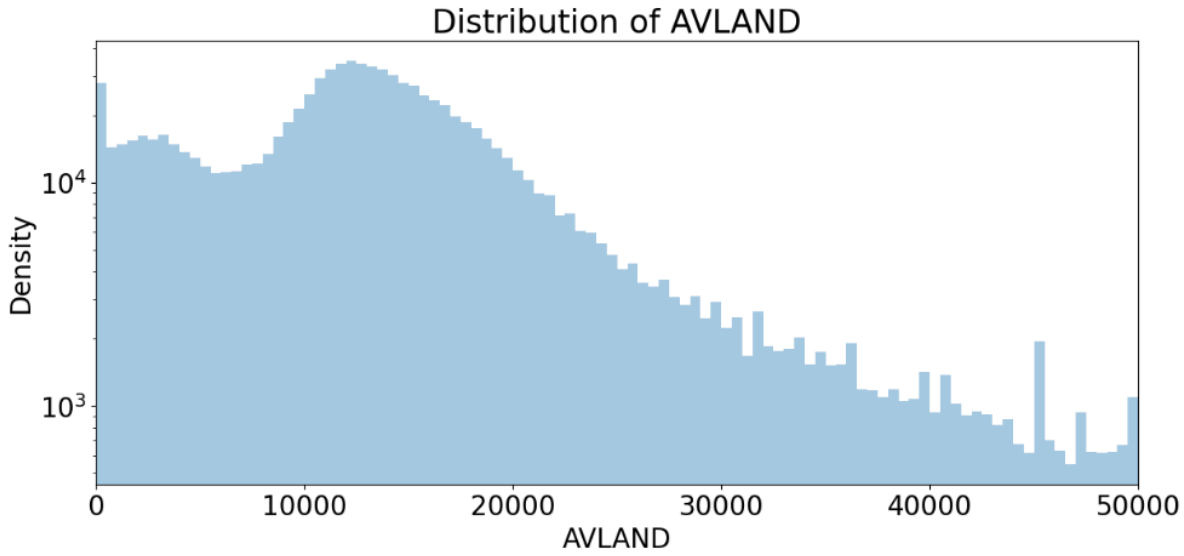
Field Name: STORIES
Description: Number of stories in the building. The plot shows the histogram of stories up to 80.



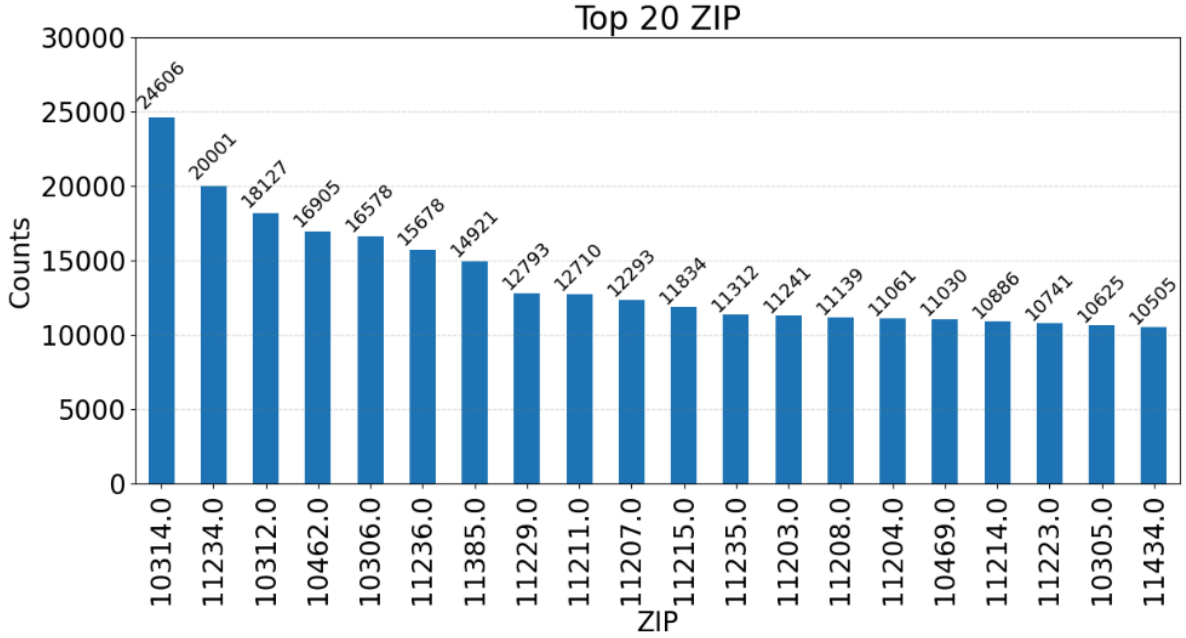
Field Name: FULLVAL
Description: Market Value. The plot shows the histogram of FULLVAL up to 2000000. The distribution appears to be bimodal. Most FULLVAL value seems to be around 500000. Surprisingly, a lot of the property seems to have 0 FULLVAL.



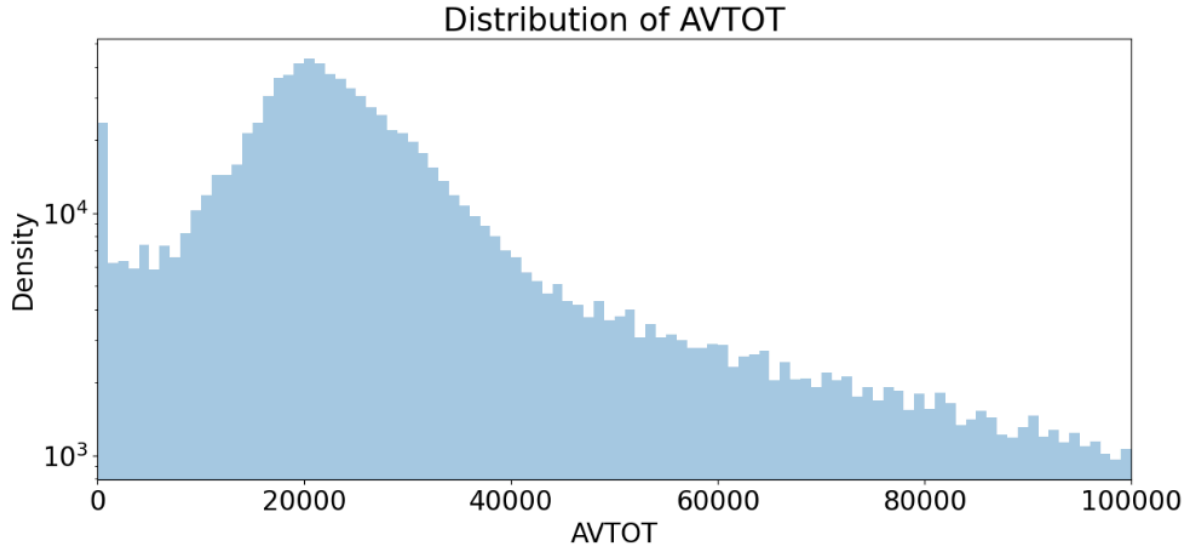
Field Name: AVLAND
Description: Actual Value of Land. The plot below shows the histogram of AVLAND, where values greater than 50000 removed. Most AVLAND value seems to be around 15000. Surprisingly, a lot of the property seems to have 0 AVLAND.



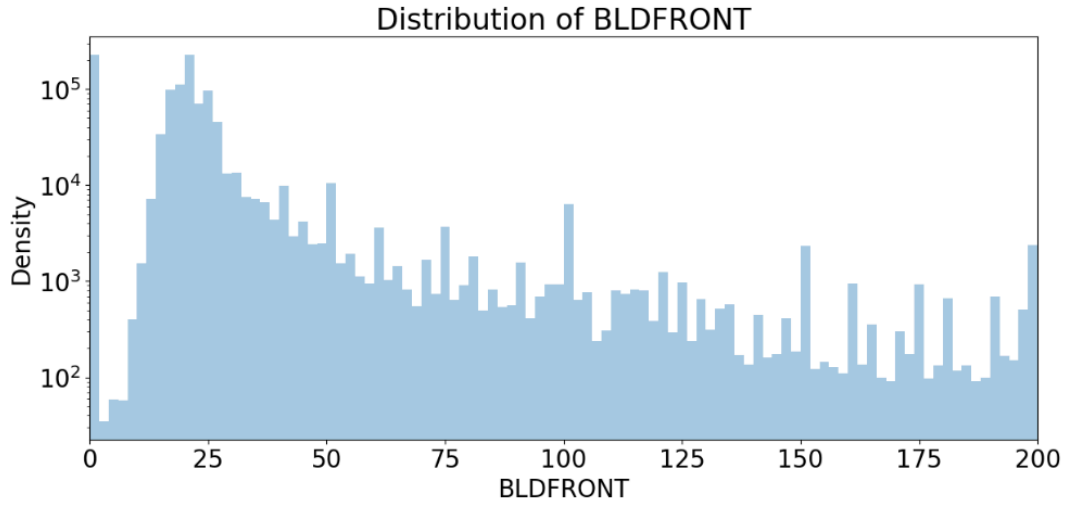
Field Name: ZIP
Description: Postal ZIP code of the property. The plot below shows the distribution of the top 20 ZIP Values. The most common appears to be 10314, with a total count of 24606.



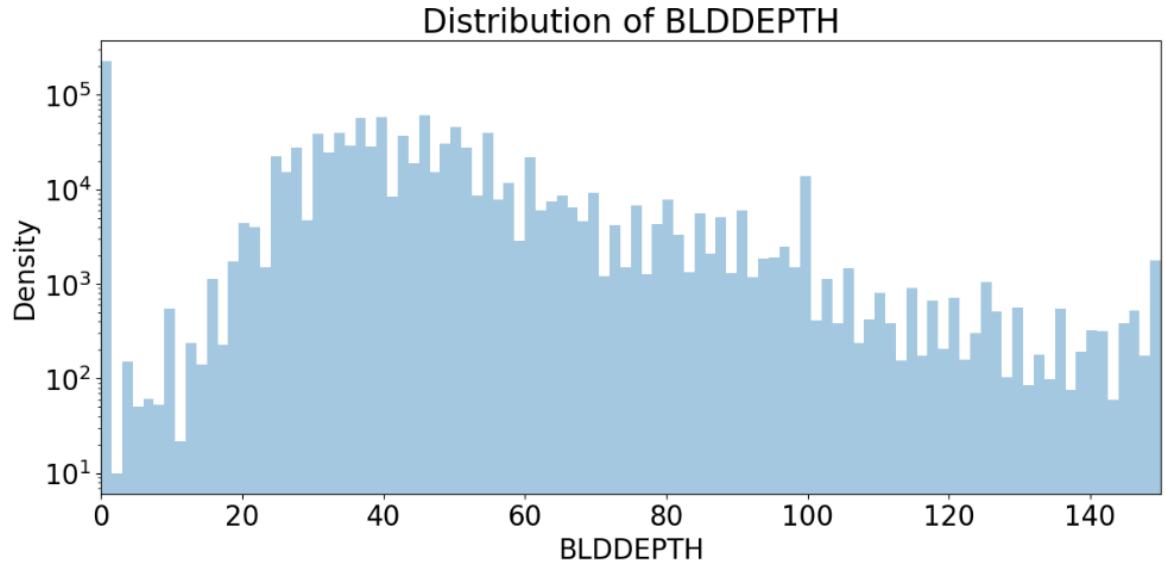
Field Name: AVTOT
Description: Actual Total Value. The plot shows the histogram of AVTOT, with values greater than 100000 removed. Most AVTOT value seems to be around 20000. Surprisingly, a lot of the property seems to have 0 AVTOT.



Field Name: BLDFRONT
Description: Building Width. The plot below shows the histogram of BLDFRONT up to 200. Surprisingly, a lot of the values seems to be 0.



Field Name: BLDDEPTH
Description: Building Depth. The plot below shows the histogram of BLDDEPTH up to 150. Surprisingly, a lot of the values seems to be 0.



Data Cleaning

A majority of the data fields had missing values. We therefore needed to use data imputation techniques before fully analyzing the dataset for potential fraud cases. The following data fields were assessed and cleaned: FULLVAL, AVLAND, AVTOT, ZIP, STORIES, LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH. In addition, stakeholders have told us that there are many properties that they are not interested in. In the subsections below, I will first describe the records that we will be excluding and then describe the imputations for each variables listed above.

Data Exclusion

Per the stakeholder's instructions, they are not interested in government owned properties and private owners with strange records.

These are the properties removed:

- Removed records with EASEMENT type as government
    - Removed 1 record
- Removed records with OWNER having the words: DEPT, DEPARTMENT, UNITED STATES, GOVERNMENT, GOVT, and CEMETERY as these properties are government owned. Also added some owners that are likely linked to the government
    - Removed 26500 records

The data exclusion step removed 26501 records in total, lowering the size of the dataset to 1044493 total records.

Data Imputation

FULLVAL

- FULLVAL have **10025 records** with value of either missing or zero. A value of zero is treated as missing because all properties should have a value.
- I grouped all records by the combination of TAXCLASS, BORO and BLDDCL, and calculated the average of FULLVAL for each group, and replaced the missing values in each group with the average of the group.
    - This fills 2718 records with 7307 still missing
- I then grouped all records by the combination of TAXCLASS and BORO and calculated the average of FULLVAL for each group, and replaced the missing values in each group with the average of the group.
    - This fillls 6921 records with 386 stillmissing
- I then grouped all records by TAXCLASS and calculated the average of FULLVAL for each group, and replaced the missing values in each group with the average of the group
    - This fills the remaining 386, successfully imputing all missing values

AVLAND

- AVLAND have **10027 records** with value of either missing or zero. A value of zero is treated as missing because all properties should have a value.
- I grouped all records by the combination of TAXCLASS, BORO and BLDDCL, and calculated the average of AVLAND for each group, and replaced the missing values in each group with the average of the group.
    - This fills 2720 records with 7307 still missing
- I then grouped all records by the combination of TAXCLASS and BORO and calculated the average of AVLAND for each group, and replaced the missing values in each group with the average of the group.
    - This fillls 6921 records with 386 stillmissing
- I then grouped all records by TAXCLASS and calculated the average of AVLAND for each group, and replaced the missing values in each group with the average of the group
    - This fills the remaining 386, successfully imputing all missing values

AVTOT

- AVTOT have **10025 records** with value of either missing or zero. A value of zero is treated as missing because all properties should have a value.
- I grouped all records by the combination of TAXCLASS, BORO and BLDDCL, and calculated the average of AVTOT for each group, and replaced the missing values in each group with the average of the group.
    - This fills 2718 records with 7307 still missing
- I then grouped all records by the combination of TAXCLASS and BORO and calculated the average of AVTOT for each group, and replaced the missing values in each group with the average of the group.
    - This fillls 6921 records with 386 stillmissing
- I then grouped all records by TAXCLASS and calculated the average of AVTOT for each group, and replaced the missing values in each group with the average of the group
    - This fills the remaining 386, successfully imputing all missing values

ZIP

- ZIP have **20431 records** with value of either missing or zero.
- The first step is to concatenate the STADDR and BORO columns into a new column and get a dictionary mapping the concatenated column with the ZIP. I then mapped the dictionary to fill in the missing values in ZIP.
    - This fills 2832 records with 17599 still missing
- The next step assumes that the data is already sorted by the ZIP. so if a ZIP is missing and the before and after ZIP are the same, I filled the ZIP with that value.
    - This fills 9491 records with 8108 still missing
- The next steps fills in the remaining missing ZIP with the previous record's ZIP
    - This fills the remaining 8108 records, successfully imputing all missing values

STORIES

- STORIES have **42030 records** with value of either missing or zero. A value of zero is treated as missing because all properties should have a value.
- I grouped all records by the combination of BORO and BLDDCL, and calculated the average of STORIES for each group, and replaced the missing values in each group with the average of the group.
    - This fills 4108 records with 37922 still missing
- I then grouped all records by TAXCLASS and calculated the average of STORIES for each group, and replaced the missing values in each group with the average of the group
    - This fills the remaining 37922 records, successfully imputing all missing values

LTFRONT

- LTFRONT have **160565 records** with value of either missing or zero. A value of zero is treated as missing because all properties should have a value.
- I grouped all records by the combination of TAXCLASS and BORO, and calculated the average of LTFRONT for each group, and replaced the missing values in each group with the average of the group.
    - This fills 160563 records with 2 still missing
- I then grouped all records by TAXCLASS and calculated the average of LTFRONT for each group, and replaced the missing values in each group with the average of the group
    - This fills the remaining 2 records, successfully imputing all missing values

LTDEPTH

- LTDEPTH have **161656 records** with value of either missing or zero. A value of zero is treated as missing because all properties should have a value.
- I grouped all records by the combination of TAXCLASS and BORO, and calculated the average of LTDEPTH for each group, and replaced the missing values in each group with the average of the group.
    - This fills 161654 records with 2 still missing
- I then grouped all records by TAXCLASS and calculated the average of LTDEPTH for each group, and replaced the missing values in each group with the average of the group
    - This fills the remaining 2 records, successfully imputing all missing values

BLDFRONT

- BLDFRONT have **206851 records** with value of either missing or zero. A value of zero is treated as missing because all properties should have a value.
- I grouped all records by the combination of TAXCLASS, BORO and BLDGCL, and calculated the average of BLDFRONT for each group, and replaced the missing values in each group with the average of the group.
    - This fills 188179 records with 18672 still missing
- I then grouped all records by TAXCLASS and BORO and calculated the average of BLDFRONT for each group, and replaced the missing values in each group with the average of the group
    - This fills 2954 records with 15718 still missing
- I then grouped all records by TAXCLASS and calculated the average of BLDFRONT for each group, and replaced the missing values in each group with the average of the group
    - This fills the remaining 15718 records, successfully imputing all missing values

BLDEPTH

- BLDEPTH have **206886 records** with value of either missing or zero. A value of zero is treated as missing because all properties should have a value.
- I grouped all records by the combination of TAXCLASS, BORO and BLDGCL, and calculated the average of BLDEPTH for each group, and replaced the missing values in each group with the average of the group.
    - This fills 191082 records with 15804 still missing
- I then grouped all records by TAXCLASS and BORO and calculated the average of BLDEPTH for each group, and replaced the missing values in each group with the average of the group
    - This fills 2974 records with 12830 still missing
- I then grouped all records by TAXCLASS and calculated the average of BLDEPTH for each group, and replaced the missing values in each group with the average of the group
    - This fills the remaining 12830 records, successfully imputing all missing values

Variables Creation

To evaluate property value in relation to size, I first created three variables. I wanted to consider size from all angles so the first variable created is the area of the lot. It is important to consider the size of the entire lot as it may affect the value of the property. To calculate the area of the lot, I multiplied the lot frontage size with the lot depth size. I then focused on building size to get the building area and the building volume. Bigger and higher buildings are expected to have higher property valuations. To find the building area, I multiplied BLDFRONT with BLDDEPTH, building volume was calculated by multiplying the building area with the number of stores that the building has.

The formula is shown below:

LTSIZE (area of the lot) = LTFRONT x LTDEPTH
BLDSIZE (building size) = BLDFRONT x BLDDEPTH
BLDVOL (building volume) = BLDSIZE x STORIES

These three measurements should capture the size of the property well. With these variables created, I moved on to use these variables to get a normalized figure for the assessment values.

The thress assessment values (FULLVAL, AVLAND, AVTOT) were then normalized by each of the three size variables mentioned above (LTSIZE, BLDSIZE, BLDVOL). These set of three variable groups were then used to create nine variables, where the three assessment values were normalized by diving the three sizes.

The variables are listed below:

R1 = FULLVAL / LTSIZE
R2 = FULLVAL / BLDSIZE
R3 = FULLVAL / BLDVOL
R4 = AVLAND / LTSIZE
R5 = AVLAND / BLDSIZE
R6 = AVLAND / BLDVOL
R7 = AVTOT / LTSIZE
R8 = AVTOT / BLDSIZE
R9 = AVTOT / BLDVOL

I then calculated the inverse of the 9 values created above and only kept the maximum of the original or inverse since I only care about values either very large or very small.

The last step is creating grouped average of these 9 variables, grouping by ZIP and TAXCLASS, I then standardized the variables above by dividing R1-R9 with the 2 grouped averages, creating 18 more new variables.

Conceptually, it is shown as below:

$$R1 / \langle R1 \rangle_{TAXCLASS} , R2 / \langle R2 \rangle_{TAXCLASS} \ldots R8 / \langle R8 \rangle_{ZIP} , R9/\langle R9 \rangle_{ZIP}$$

Where $\langle R\# \rangle_{Variable}$ is the average of R# grouped by the Variable.

Lastly, I created 2 more variables comparing the value ratio and size ratio. The value ratio is calculated as FULLVAL/ (AVLAND+AVTOT) and took the maximum value of the original or the inverse asI want to find the lower outliers. The size ratio is calculated as BLDSIZE / (LTSIZE +1).

Creation Table

Description	# Variables Created	Cumulative Variables
LTSIZE LTFRONT * LTDEPTH, The area of the lot.	1	1
BLDSIZE BLDFRONT * BLDEPTH , The size of the building	1	2
BLDVOL BLDSIZE * STORIES, The volume of the building	1	3
Standardized Assessment Values The maximum of FULLVAL, AVLAND, AVTOT divided by LTSIZE, BLDSIZE, BLDVOL or its inverse. This ratio helps pinpoint properties where the tax values relative to their size are unusually high or low, potentially flagging them for further investigation for tax fraud	9	12
Grouped Averages Divide each of the Standardized Assessment Values variables by the grouped averages of either ZIP or TAXCLASS. This adjustment allows for a more accurate comparison between properties, aiding in the detection of outliers or anomalies that may indicate potential tax fraud.	18	30
Value Ratio Maximum of FULLVAL / (AVLAND + AVTOT) or its inverse. This ratio highlights properties where the total value might not align well with the land and total assessed values, potentially flagging inconsistencies or errors in tax assessments	1	31
Size Ratio BLDSIZE / (LTSIZE + 1). This metric helps identify properties where the building size might disproportionately dominate the lot size, potentially indicating unusual property features	1	32

Dimensionality Reduction

Overview

Dimensionality reduction is crucial in data analysis and machine learning to simplify the dataset by reducing the number of variables. It enhances model performance, reduces computational demands, and improves data visualization. Reducing dimensions also helps address the curse of dimensionality, which can adversely affect models when handling high-dimensional data.

First Z - Scale

Initially, I standardized the dataset to ensure all features contribute equally to the analysis, which is vital because it eliminates any bias introduced by varying scales of data features. This step involves adjusting each feature to have a mean of zero and a standard deviation of one. The transformation is as follows:

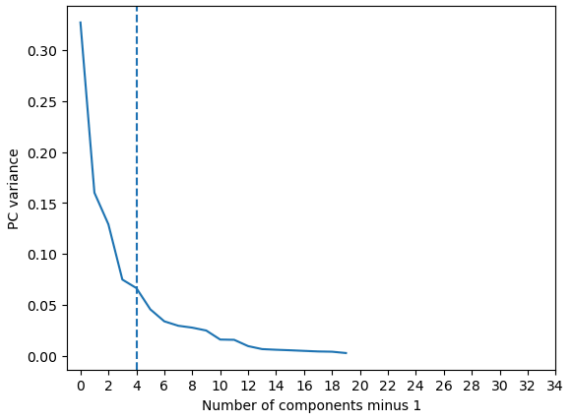
$$zi = (xi - \bar{x}) / \sigma i \ , \ \forall i \in \{1, \dots n\}$$

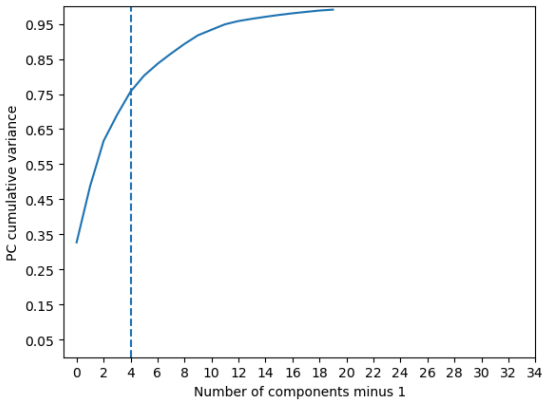
Standardizing the data is critical before applying PCA, as PCA is sensitive to the variances of the initial variables, and non-standardized data can skew the analysis towards high-variance variables. The outcome of standardizing the dataset is that every dimension of the dataset is clustered around 0 with the same scaling, and thus a simple measure of distance from the origin would yield any outlier values.

Principal Component Analysis

Following the first Z-scale, I applied PCA to reduce the dimensionality of the dataset. Principal component analysis (PCA) is a dimensionality reduction method that performs a transformation on the dataset and extracts a set of features to create a low dimensional dataset. The transformation projects the original dataset to the axes (i.e. principal components) for those features that have the maximum variance. Additionally, the features of the transformed dataset are ordered according to variance and therefore those features with small variance are considered to have a high linear correlation with other features. Hence, dimensionality reduction is performed by choosing the first k features of the resulting dataset that contain a desirable percentage of the total variance. This helps in focusing the analysis on the most significant features, which are the principal components that account for the most variance in the dataset.

From the scree plot (right figure) and the percent of variance explained plot (left figure), I determined that five principal components were optimal to capture the essential variability of the data while simplifying the dataset significantly.





Second Z - Scale

After PCA transformation, I applied a second round of z-scale standardization to the principal components. This additional standardization is beneficial for normalizing the new features, especially useful for the subsequent machine learning algorithms mentioned in the section below. This step ensures that the principal components are normalized, facilitating their use in downstream processes and maintaining consistency across analyses.

Rationale for Transformations

These transformations collectively streamline the dataset, ensuring the analysis focuses on the most informative attributes and preparing the data for efficient, high-quality analysis and modeling. By standardizing the data before and after PCA, I effectively manage feature scales and enhance the interpretability and applicability of the reduced dataset in subsequent analytical procedures.

Unsupervised Algorithm

Algorithm 1 - Z-Scores and Principal Components

Why Z-Scores and Principal Components?

Z-scores anchor data of different scales to the same standard normal distribution with a mean of zero and standard deviation of one. Outliers usually are extreme in some dimensions. In this way, some dimensions of outliers will be likely to take relatively extreme values on the standard normal curve. To measure the general distance of dimensions from normality, I needed to leverage the average deviation of all dimensions and highlight extreme deviations in certain dimensions by combining the Z-scores of the principal components that remain in the dataset using distance from origin point in n-dimensional space. When n was large enough, I was able to highlight extreme deviation. Considering leveraging other features, I specified n as 2 to calculate distance under 2-dimensional space. This enabled me to be able to identify records that displayed higher abnormality across the most important dimensions in the data by observing whether records showed large distance from the origin point.

Z - Scores Formula

The formula of the first fraud score was:

$$\text{score} = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

Where x_i represents the i-th element in the dataset, p is the parameter that determines the norm (in this case p = 2, making this an L2 norm), n is the number of elements in the dataset

The resulting positive number reflected the distance of a record from the origin. The higher the number is, the higher possibility that the record was an outlier.

Algorithm 2 - Autoencoder

What is an Autoencoder and Why did I use it?

An autoencoder is considered a statistical model or a type of neural network that can be trained to learn data encodings. The architecture of autoencoders allows them to start learning the data encodings in the input layer where it is subsequently reconstructed and eventually sent to the output layer for analysis. Between the input and output layers are the hidden layers of the autoencoder where the data is compressed to a lower-dimensional space and forwarded to the output layer. This process of dimensionality reduction and data reconstruction offers noise reduction and outlier detection in the dataset.

By going through the process of reducing dimensions and restoring data, an autoencoder keeps the main patterns and features of the data input and removes the noises that cause deviation from normal estimation.

If a record was abnormal, it would not follow common patterns, so the restored value of an abnormal record would be quite different from the original record. I calculated how far a restored record was from its original value, and the records with the largest deviation will likely be outliers.

Autoencoder Scoring Formula

The second score is calculated using a neural network model implemented as an autoencoder. This model, set up with three neurons in a single hidden layer and a logistic activation function, is trained on the z-standardized principal component data above, functioning both as input and output to learn efficient data codings. The training aims to enable the network to reconstruct the input data, whereupon it predicts the principal component scores post-training. The reconstruction error is then determined by the discrepancy between the predicted and actual data points. This error is processed by squaring each component, summing these squared errors for each observation, and then taking the square root, essentially computing the Euclidean distance from the original data points. The formula is shown as follows:

$$\text{score2} = \left(\sum_{i=1}^{n} |e_i|^2\right)^{1/2}$$

Where e_i represents the error for the i-th dimension in your dataset, calculated as the difference between the predicted and actual values for that dimension and n is the number of dimensions in the data

This scoring method, by focusing on the magnitude of reconstruction errors, helps identify data points where the model's representation significantly deviates from the actual data, pointing to potential outliers or anomalies in the dataset.

The positive fraud score reflects the distance of one record from the reproduced dataset to the original dataset. The higher the number is, the higher possibility that the record is an outlier.

Final Score - Leveraging Both Algorithms

Why not use the two original scores?

Using only one of the original scores (from the previous algorithms) to identify anomalies or outliers in the dataset might not capture the full complexity and nuances of the data. Each score measures different characteristics: Algorithm 1 might focus on the magnitude of the deviations within the principal components, providing insights into variance from the norm based on a

standardized PCA framework. On the other hand, Algorithm 2 evaluates the reconstruction error from an autoencoder, highlighting data points where the neural network model fails to accurately predict or reconstruct the original data. Relying on just one score could lead to biases inherent in the respective model's sensitivity or specificity, potentially missing critical outliers or falsely flagging normal data points as anomalous. Combining the two scores helps mitigate these issues by leveraging the strengths of both analytical approaches and providing a more balanced view of the data.

Final Score Calculation

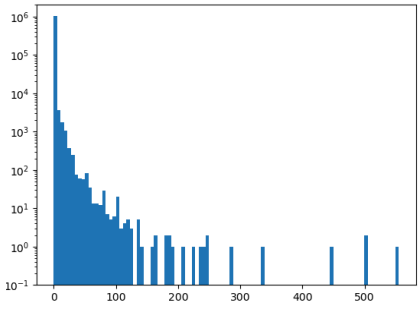
The final score is calculated by averaging the ranks of Algorithm 1 and Algorithm 2, assigning an equal weight of 0.5 to each. This is represented by the formula:

$$\text{Final score} = 0.5 * \text{score 1 rank} + 0.5 * \text{score 2 rank}$$

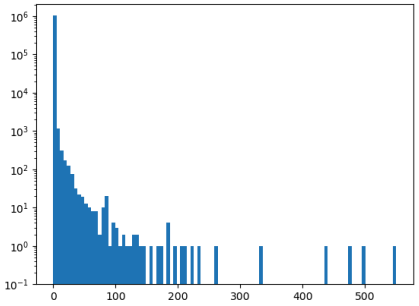
This method of averaging ranks instead of direct values accounts for the relative performance of each data point within each scoring system, ensuring that the final score reflects a balanced consideration of both models. The ranks are used because they normalize the distribution scales of the scores, which might differ significantly, especially if one score is typically higher or spreads more widely than the other. By sorting these combined scores in descending order and focusing on the top entries, you effectively highlight the most significant outliers or anomalies in your dataset. This final combined score offers a robust metric that integrates diverse analytical perspectives, enhancing the reliability and comprehensiveness of anomaly detection in your unsupervised learning model.

Results

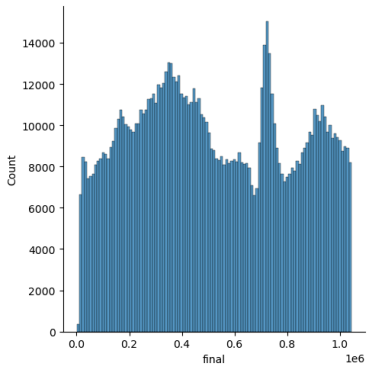
Distributions of Scores



Distribution of raw score 1



Distribution of raw score 2



Distribution of final ranked score

As seen in the distributions above, most of the entries have scores that are less than 200, with some having max scores of greater than 300. Records with max scores of greater than 300 are properties that are likely to contain fraudulent information.

How to Use Final Score

To determine properties that likely contain fraudulent information using the final score, I sorted the final scores from highest to lowest, the higher the score, the more likely that the property had fraudulent information.

This then allowed me to check which properties and the record number of the properties. From my analysis, the top 5 highest ranked properties were shown below

	RECORD	BBLE	BORO	BLOCK	LOT	EASEMENT	OWNER
956519	956520	5006590012	5	659	12	NaN	TROMPETA RIZALINA
658932	658933	4029060054	4	2906	54	NaN	WAN CHIU CHEUNG
632815	632816	4018420001	4	1842	1	NaN	864163 REALTY, LLC
1067359	1067360	5078530085	5	7853	85	NaN	NaN
1067000	1067001	5078120132	5	7812	132	NaN	DRANOVSKY, VLADIMIR

Examining Properties

Now equipped with the record numbers, addresses, and other relevant details (such as the number of stories, building size, and lot size) of the properties, I can enhance the verification process by directly examining these properties on Google Maps. This approach allows me to visually inspect each property and assess if there are any discrepancies or potential fraudulent entries based on their physical attributes and reported data. This method provides a practical, real-world check that complements the analytical insights gained from our data-driven model, ensuring a more comprehensive investigation of potential tax fraud.

5 Interesting Properties

Upon closer examination of the properties with the highest final rank scores, it appears that many of these entries exhibit extreme anomalies in their reported dimensions or total values, often showing figures as low as 0 or 1. These unusual values are likely due to misentries or data errors.

Among the properties flagged in the top 50 highest final score, five stood out as not fitting this pattern of extreme values. These properties displayed more typical and plausible characteristics, yet still ranked high on our fraud detection scores. Below are the 5 properties:

1st Unusual Property

Record	658933	Taxclass	1
BLDFRONT	2500	STADDR	54-76 83 Street
BLDDEPTH	5600	STORIES	3
LTFRONT	25	AVTOT	46560
LTDEPTH	100	AVLAND	26940

This property is located on 54-76 83 Street in Queens and appears to be a personal residential property.

When I dig the values created for the fraud algorithm, I found these anomalies:
r2_taxclass: 401.43, **r3_taxclass**: 456.59, **r5_taxclass**: 379.02, **r6_taxclass**: 497.22, **r8_taxclass**: 315.14, **r9_taxclass**: 467.77.

These variables were created to account for the building dimensions grouped by their taxclass. These extremely high values suggest that this property is extremely unusual compared to the other properties in the same tax class.

This is supported from the picture from Google Maps, it appears that the BLDFRONT and BLDDEPTH (Buidling dimensions) are way bigger than it should be.

2nd Unusual Property

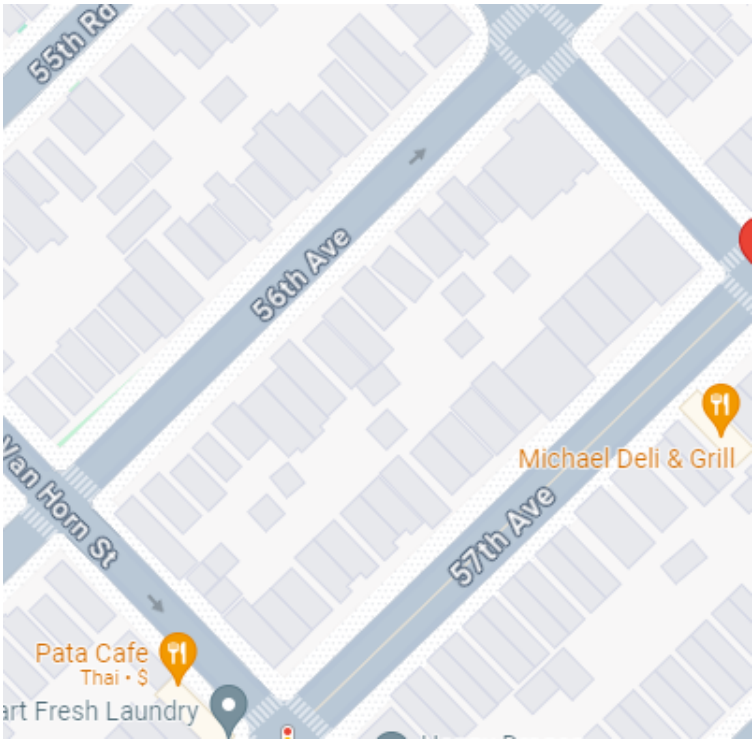
Record	649717	Taxclass	4
BLDFRONT	0	STADDR	57 Avenue
BLDDEPTH	0	STORIES	NaN
LTFRONT	51	AVTOT	10
LTDEPTH	940	AVLAND	10

This property is located on 57 Avenue in Queens and appears to be in a personal residential property area (does not have exact building number).

When I dig the values created for the fraud algorithm, I found these anomalies:
r1_zip5: 286.95, **r4_zip5**: 295.69, **r7_zip5**: 304.43, **r1**: 90.48, **r2**: 36.49, **r4**: 89.98, **r7**: 163.21

These variables were created to account for the property value and lotsize, and also when grouped by the surrounding zip code. These extremely high values suggest that this property is extremely unusual compared to the other properties in the same zip code.

This is supported from the picture from Google Maps, although the exact building number was not given, from the view, there does not appear to be any building that could potentially have a lot depth of 940 feet. Furthermore, the records show a building dimension of 0, null value for number for of stories, and only 10 for both land values.



3rd Unusual Property

Record	980276	Taxclass	1
BLDFRONT	15	STADDR	160 WILCOX STREET
BLDDEPTH	30	Stories	1
LTFRONT	278	AVTOT	92
LTDEPTH	190	AVLAND	89

This property is located on 160 Wilcox Street in Staten Island and looks to be a personal residential property.

When I dig the values created for the fraud algorithm, I found these anomalies:
r1_taxclass: 229.48, **r4_taxclass**: 225.04, **r7_taxclass**: 357.28, **r7**: 67.07

These variables were created to account for the property value and lotsize, and also when grouped by the tax class. These extremely high values suggest that this property is extremely unusual compared to the other properties in the same tax class.

Surprisingly, from the picture from Google Maps, there does not appear to be an error in the lot size and building dimension. However, the actual land value and total value are definitely too low and should be investigated for under reporting.

4th unusual property

Record	333412	Taxclass	2B
BLDFRONT	4017	STADDR	37 Monroe Street
BLDDEPTH	42	Stories	3
LTFRONT	17	AVTOT	4077
LTDEPTH	85	AVLAND	3874

This property is located on 37 Monroe Street in Brooklyn and is an residential apartment.

When I dig the values created for the fraud algorithm, I found these anomalies:
r2_taxclass: 251.49, **r3_taxclass**: 253.22, **r9_taxclass**: 65.06

These variables were created to account for the property value and building size/volume, and also when grouped by the tax class. These extremely high values suggest that this property is extremely unusual compared to the other properties in the same tax class.

From the picture from Google Maps, it appears that the BLDFRONT is way bigger than it should be. Furthermore, there appears to be 4 stories but was only reported as 3. Lastly, the reported values for the property seems to be too low.

5th Unusual Property

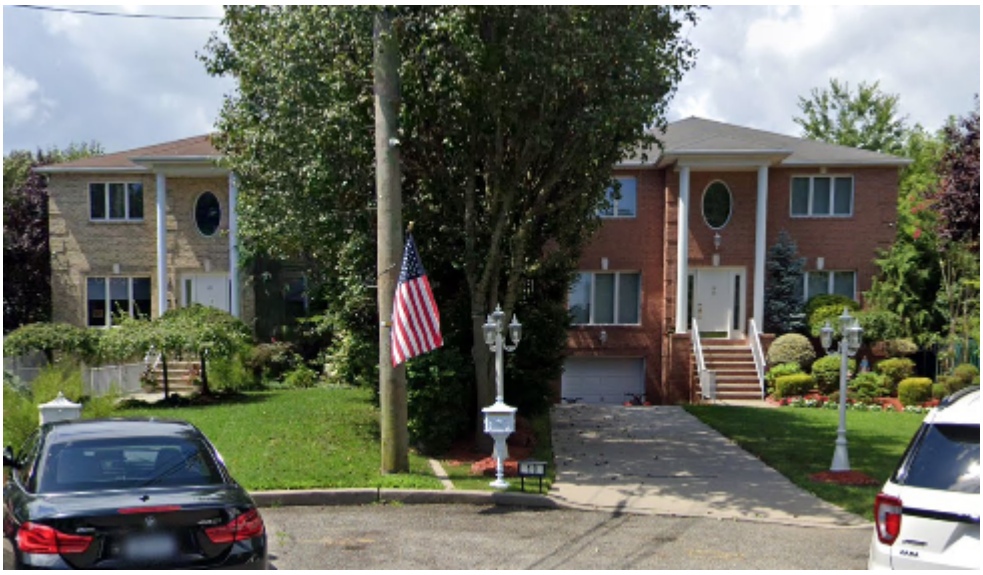
Record	1067360	Taxclass	2B1
BLDFRONT	36	STADDR	20 Emily Court
BLDDEPTH	45	Stories	2
LTFRONT	1	AVTOT	50160
LTDEPTH	1	AVLAND	28800

This property is located on 37 Monroe Street in Brooklyn and is an residential apartment.

When I dig the values created for the fraud algorithm, I found these anomalies:
r1_taxclass: 651.03, **r4_taxclass**: 659.15, **r7_taxclass**: 604.23, **r1_zip5**: 60.44, **r4_zip5**: 150.30, **r7_zip5**: 96.50.

These variables were created to account for the property value and lot size, and also when grouped by the tax class or zip code. These extremely high values suggest that this property is extremely unusual compared to the other properties in the same tax class and also compared to the same properties in the zip code.

This is supported from the picture from Google Maps, it appears that the LTFRONT and LTDEPTH (lot dimensions) are way smaller than it should be.



Conclusions

In this project, a comprehensive analysis of NYC property assessment values was conducted to identify potential fraudulent records. Utilizing sophisticated data analysis techniques, including Principal Component Analysis (PCA) and a neural network autoencoder, I derived two distinct fraud scores for each record. These scores were then combined to generate a single, consolidated fraud score rank, highlighting properties with the highest likelihood of inaccuracies or fraud.

While investigating the top 50 records with the highest fraud scores, it was noted that many properties exhibited unusually low dimension sizes (such as building depth and lot width) and minimal total values, often as low as 0 or 1. This recurring trend of minimal dimensions and values suggested potential data entry errors or systematic issues which warrant further investigation. However, among these, five properties stood out not because of low dimensional or value metrics but due to other factors that still flagged them as high-risk in our fraud detection models. These five cases, which didn't follow the common trend of minimal metrics, was scrutinized further and analyzed in the results section above.

Moving forward, based on client feedback and the outcomes of our initial findings, several improvements can be made to enhance the accuracy and effectiveness of the fraud detection process:

1. **Algorithm Refinement:** Adjust the parameters of the PCA and autoencoder models based on expert feedback to better capture subtle nuances and complexities in the data that may indicate fraud.
2. **Feedback Integration:** Continuously incorporate feedback from property tax experts into the analysis process, using their insights to potentially create new variables that will likely enhance the detection capabilities of our models.
3. **Anomaly Investigation:** Systematically investigate the causes behind the prevalence of very low property dimensions and values to determine if these are due to systematic reporting errors or fraudulent attempts to minimize tax liabilities.

These steps will not only help address specific issues identified in the current analysis but also build a more robust and reliable system for ongoing fraud detection in property assessments. This proactive approach will enhance the credibility of the assessment process and ensure fair and accurate property taxation.

APPENDIX - DATA QUALITY REPORT

1. Data Description

Data Name: Property Valuation and Assessment Data

Data Source: Open data of NYC government

Data URL: https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8

No. of Fields: 32

No. of Records: 1,070,994

2. Summary

2.a) Numerical

Field Name	# records that have a value	% populated	Most Common	# records with value 0	Min	Max	Mean	Standard Deviation
LTFRONT	1,070,994	100.0%	0	169108	0	9,999	36.64	74.03
LTDEPTH	1,070,994	100.0%	100	170128	0	9,999	88.86	76.40
STORIES	1,014,730	94.7%	2	0	1	119	5.01	8.37
FULLVAL	1,070,994	100.0%	0	13007	0	6,150,000,000	874,264.51	11,582,425.58
AVLAND	1,070,994	100.0%	0	13009	0	2,668,500,000	85,067.92	4,057,258.16
AVTOT	1,070,994	100.0%	0	13007	0	4,668,308,947	227,238.17	6,877,526.09
EXLAND	1,070,994	100.0%	0	491699	0	2,668,500,000	36,423.89	3,981,573.93
EXTOT	1,070,994	100.0%	0	432572	0	4,668,308,947	91,186.98	6,508,399.78
BLDFRONT	1,070,994	100.0%	0	228815	0	7,575	23.04	35.58
BLDDEPTH	1,070,994	100.0%	0	228853	0	9,393	39.92	42.71
AVLAND2	282,726	26.4%	2,408	0	3	2,371,005,000	246,235.71	6,178,951.64
AVTOT2	282,732	26.4%	750	0	3	4,501,180,002	713,911.44	11,652,508.34
EXLAND2	87,449	8.2%	2,090	0	1	2,371,005,000	351,235.68	10,802,150.91
EXTOT2	130,828	12.2%	2,090	0	7	4,501,180,002	656,768.28	16072448.75

2.b) Categorical

Field Name	# records that have a value	% populated	# unique values	Most common value
RECORD	1,070,994	100.0%	1,070,994	1
BBLE	1,070,994	100.0%	1,070,994	1,000,010,101
BORO	1,070,994	100.0%	5	4
BLOCK	1,070,994	100.0%	13,984	3,944
LOT	1,070,994	100.0%	6,366	1
EASEMENT	4,636	0.4%	12	E
OWNER	1,039,249	97.0%	863,347	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100.0%	200	R4
TAXCLASS	1,070,994	100.0%	11	1
EXT	354,305	33.1%	3	G
EXCD1	638,488	59.6%	129	1,017
STADDR	1,070,318	99.9%	839,280	501 SURF AVENUE
ZIP	1,041,104	97.2%	196	10,314
EXMPTCL	15,579	1.5%	14	X1
EXCD2	92,948	8.7%	60	1,017
PERIOD	1,070,994	100.0%	1	FINAL
YEAR	1,070,994	100.0%	1	2010/11
VALTYPE	1,070,994	100.0%	1	AC-TR

3. DATA FIELD EXPLORATION

Field 1

Field Name: RECORD

Description: Ordinal unique positive integer for each record, from 1 to 1070994

Field 2

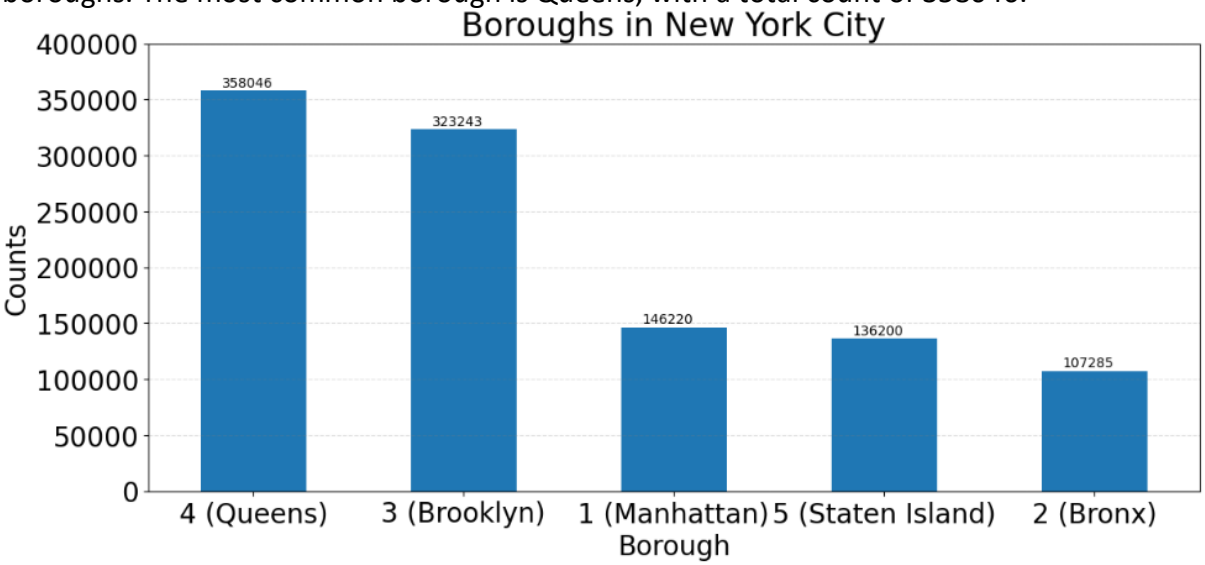
Field Name: BBLE

Description: Concatenation of Borough code (BORO), block code (BLOCK), LOT, and Easement Code (EASEMENT). Used as file key

Field 3

Field Name: BORO

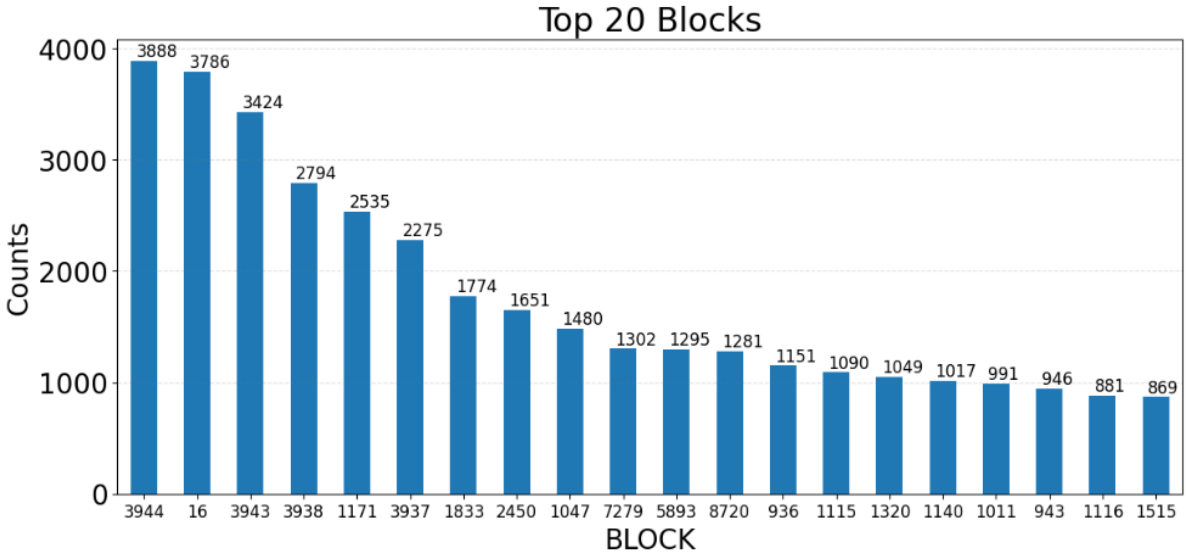
Description: Borough Codes. The plot shows the distribution of values across the 5 boroughs. The most common borough is Queens, with a total count of 358046.



Field 4

Field Name: BLOCK

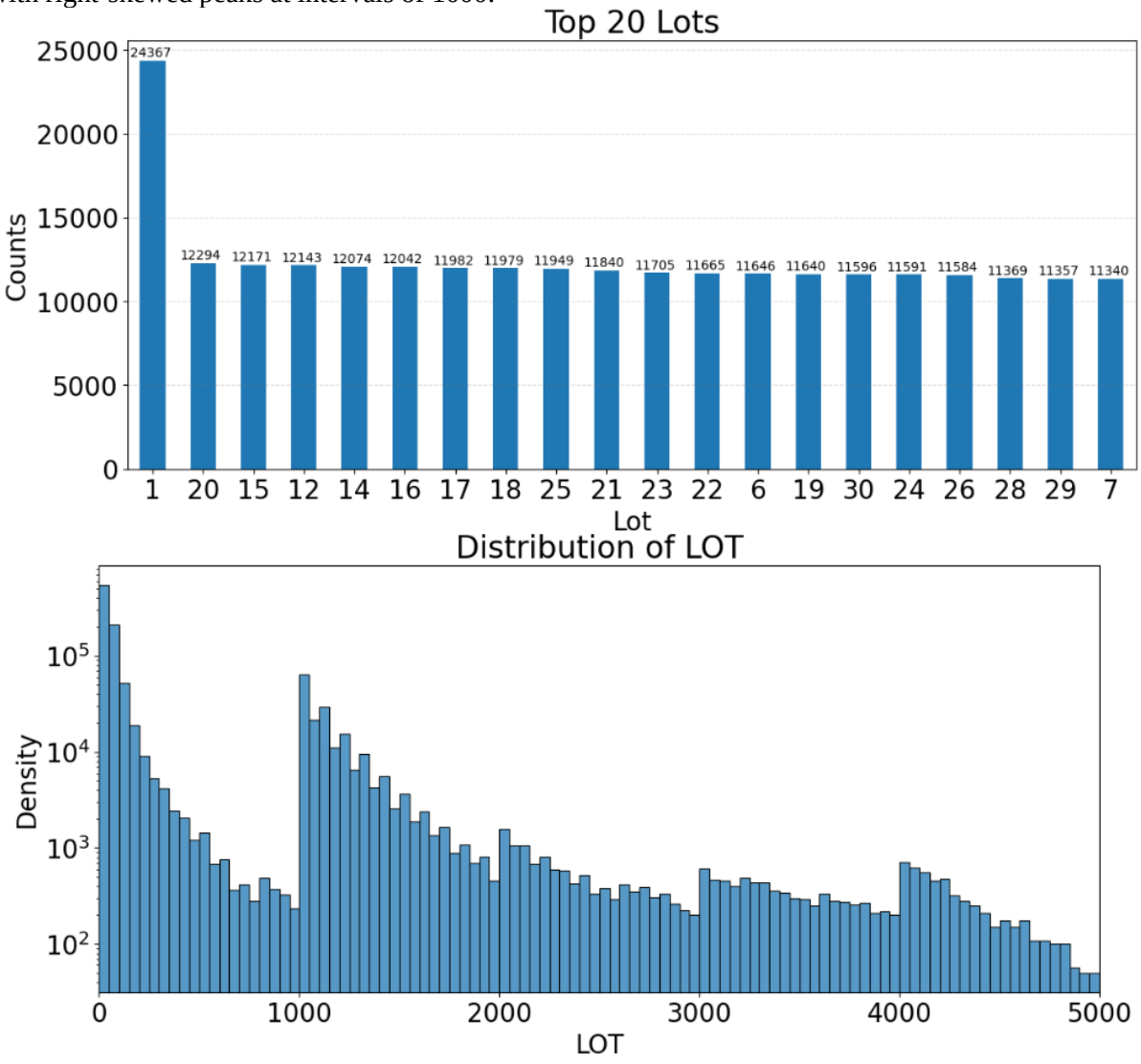
Description: Valid block ranges by borough codes. Below is the histogram of the top 20 most common blocks in the data. The most common block is 3944, with a count of 3888.



Field 5

Field Name: LOT

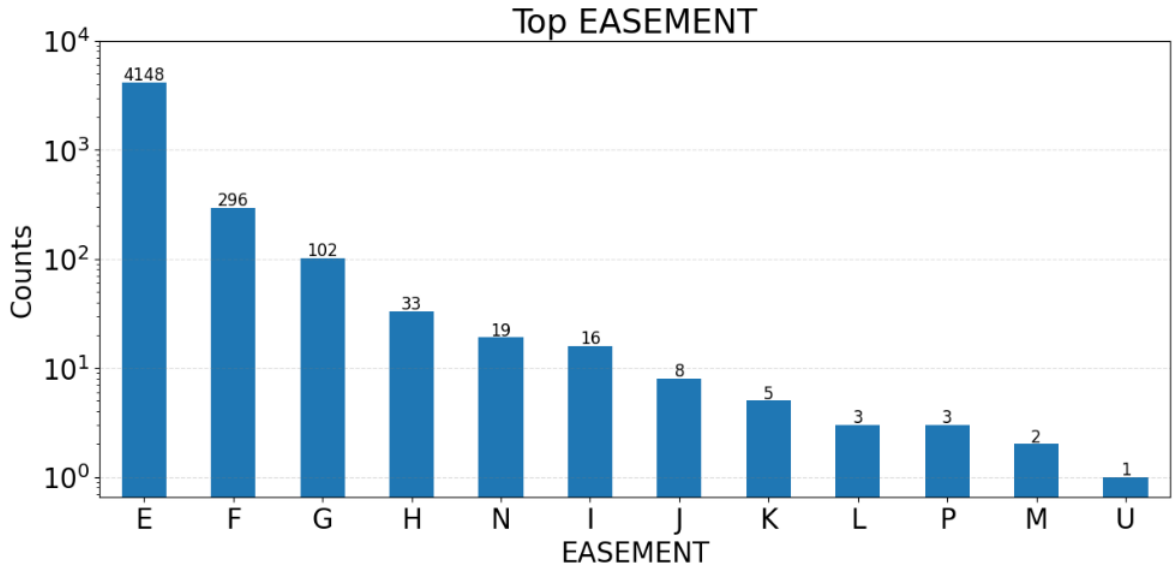
Description: Unknown field. The first plot shows the top 20 most common lot values in the data. The most common lot is 1, with a total count of 24367. The second plot shows the distribution of LOT values up to 5000. Interestingly, it appears to be a multi-model distribution with right-skewed peaks at intervals of 1000.



Field 6

Field Name: EASEMENT

Description: A field that is used to describe easement. The plot below shows the distribution of EASEMENT values. The most common is E, with a total count of 4148.

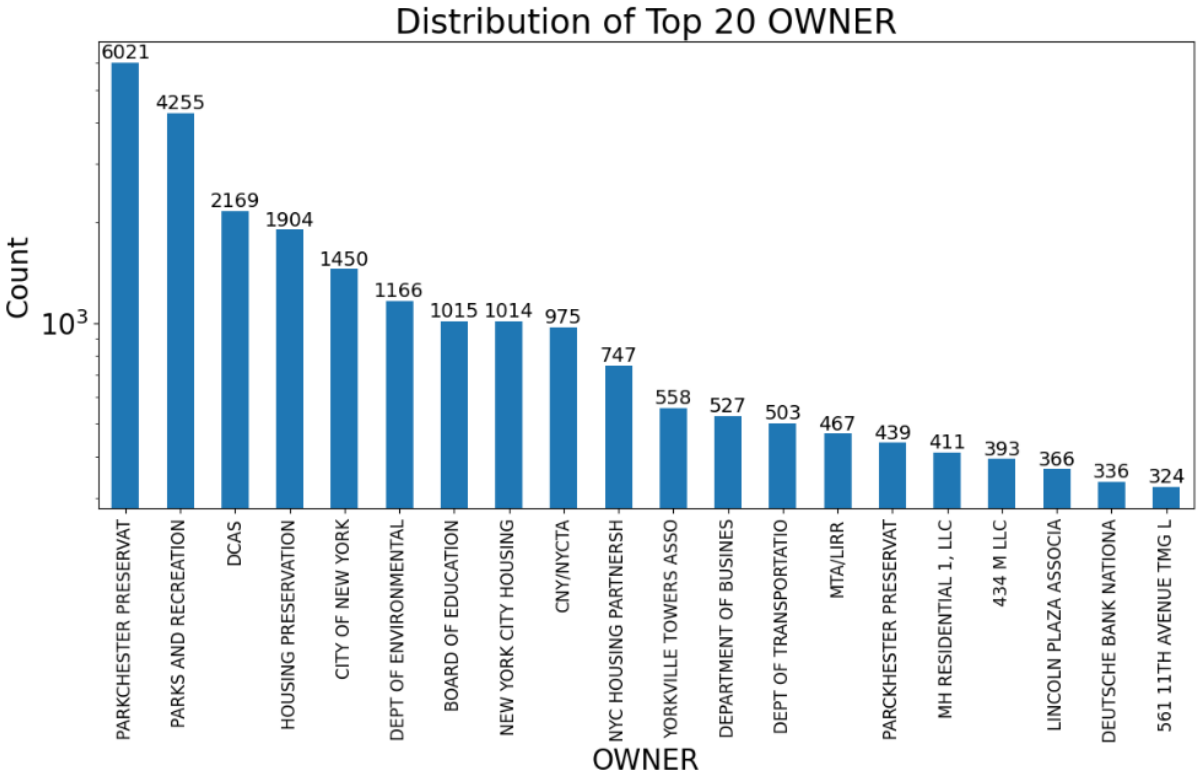


E, F, G, H I, J, K, L, M = Land Easement. **N** = Non-Transit Easement. **P** = Pier. **U** = U.S. Government

Field 7

Field Name: OWNER

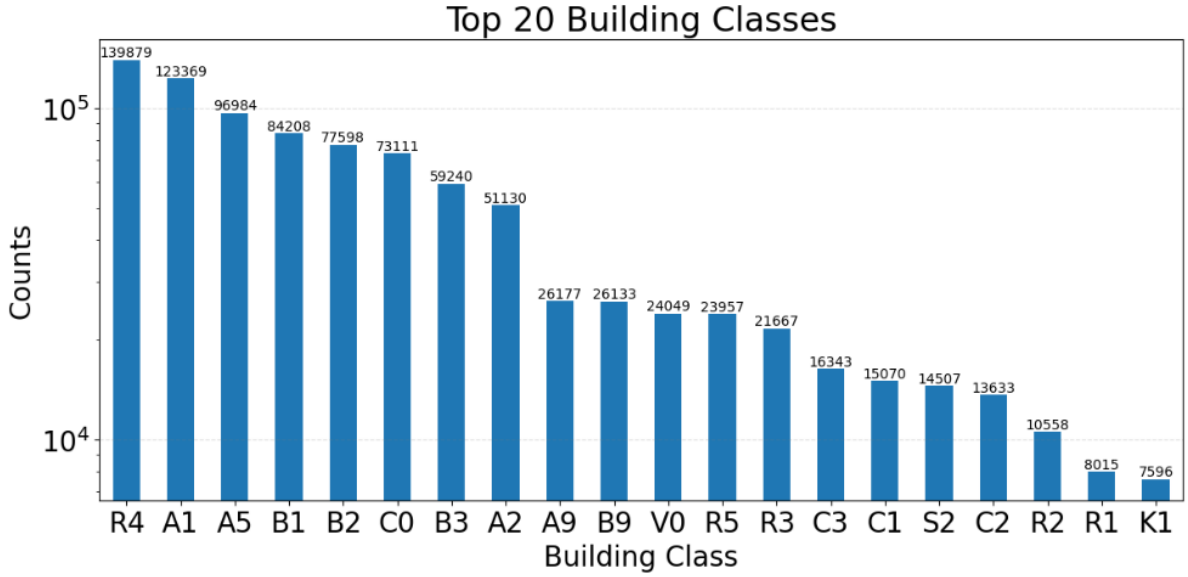
Description: Owner's Name. The plot below shows the distribution of the top 20 most common owner name. The most common is Parkchester Preservat, with a total count of 6021.



Field 8

Field Name: BLDGCL

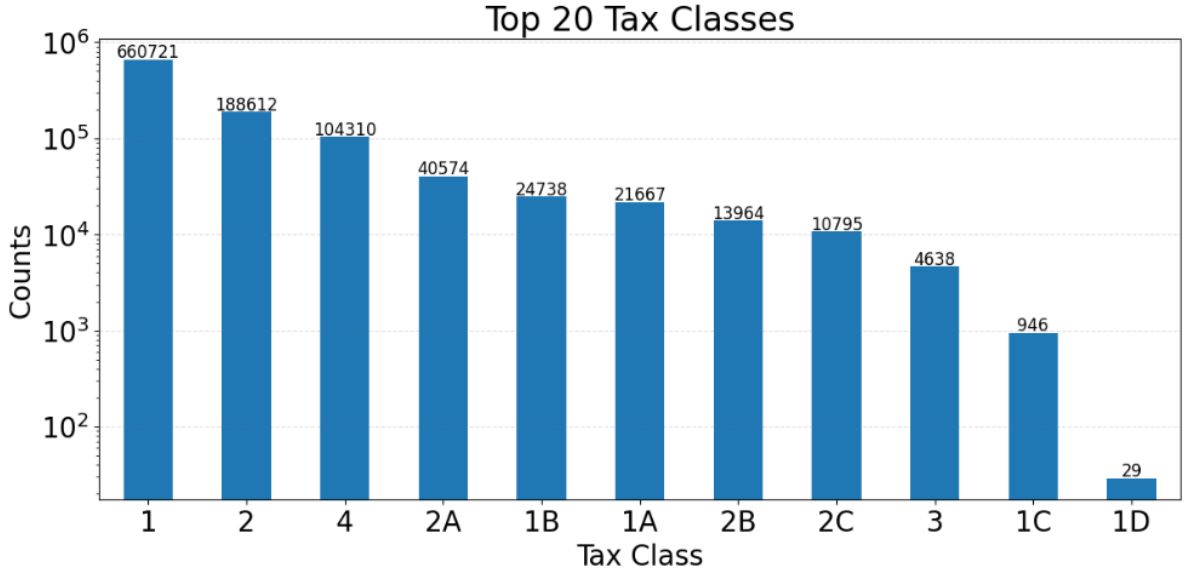
Description: Building class where position 1 is an alphabet & position 2 is a number. The plot below shows the distribution of the top 20 most common building class. The most common is R4, with a total count of 139879.



Field 9

Field Name: TAXCLASS

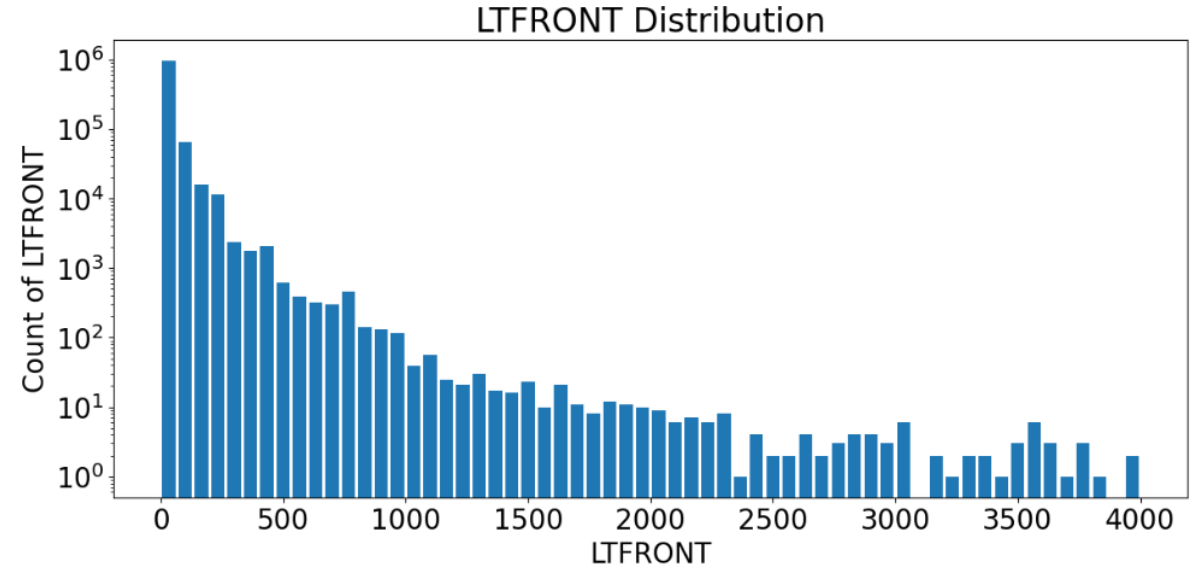
Description: Current Tax Class Code. 1, 1A, 1B, 1C, 1D = 1-3 Unit Resident. 2, 2A, 2B, 2C = Apartments. 3 = Utilities. 4 = All Others. The plot below shows the distribution of the Taxclass Field. The most common taxclass is 1, with a total count of 660721.

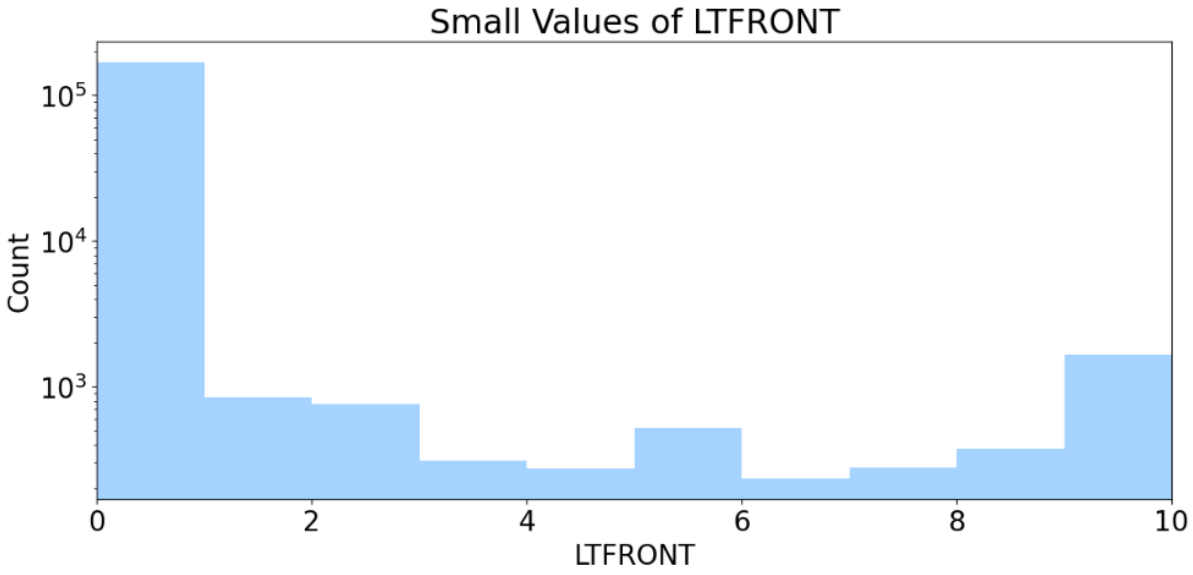


Field 10

Field Name: LTFRONT

Description: Lot Width. The first plot shows the histogram of LTFRONT values from 0 - 4000. The second plot shows the histogram of LTFRONT from values of 0-10. Surprisingly, there are a lot of values that have 0 width.

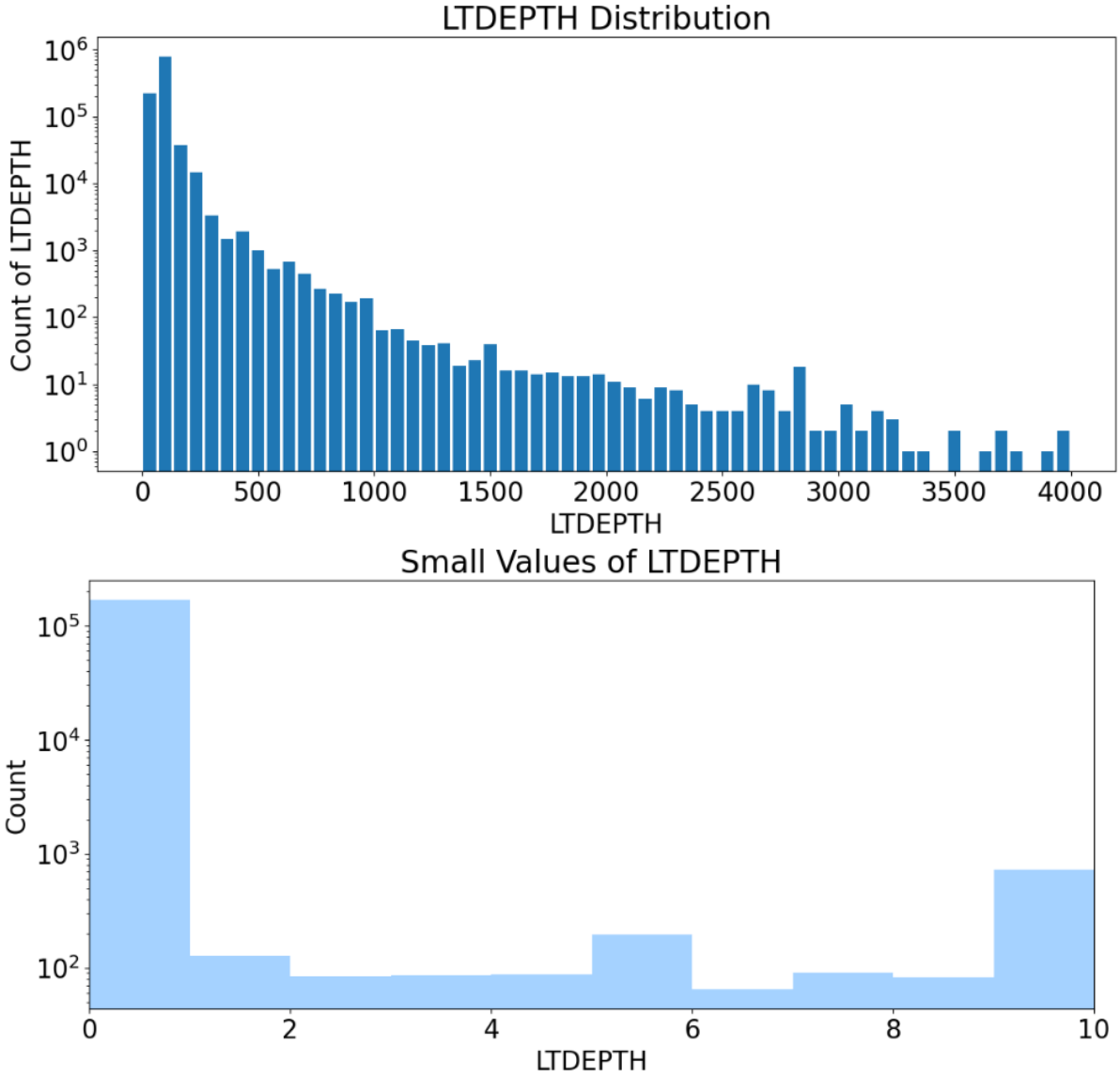




Field 11

Field Name: LTDEPTH

Description: Lot Depth. The first plot shows the histogram of LTDEPTH values from 0 - 4000. The second plot shows the histogram of LTDEPTH from values of 0-10. Surprisingly, there are a lot of values that have 0 width.

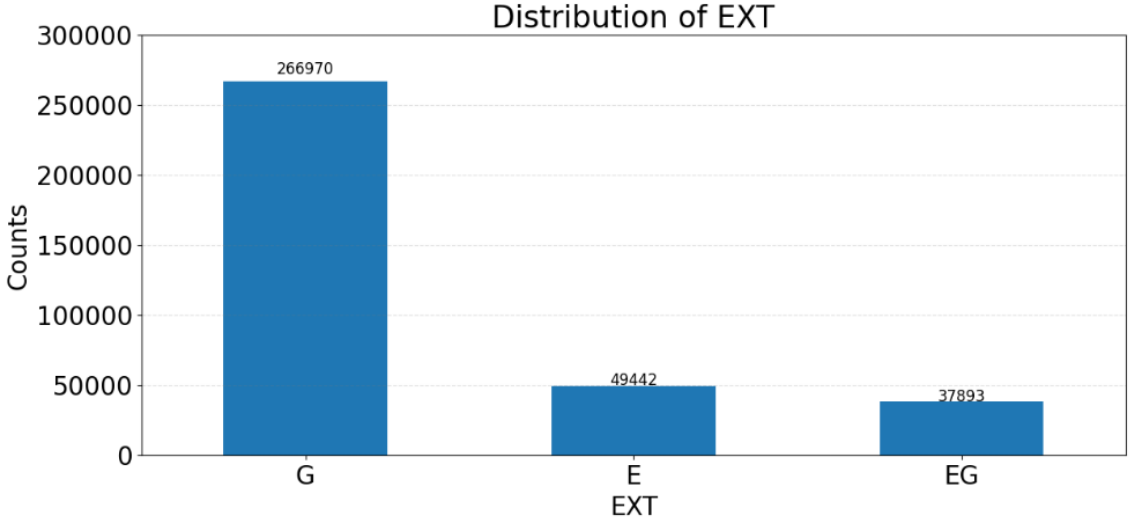


Note: The distributions of LTDEPTH and LTFRONT are very similar, which makes sense as a larger LTDEPTH likely means a larger LTFRONT.

Field 12

Field Name: EXT

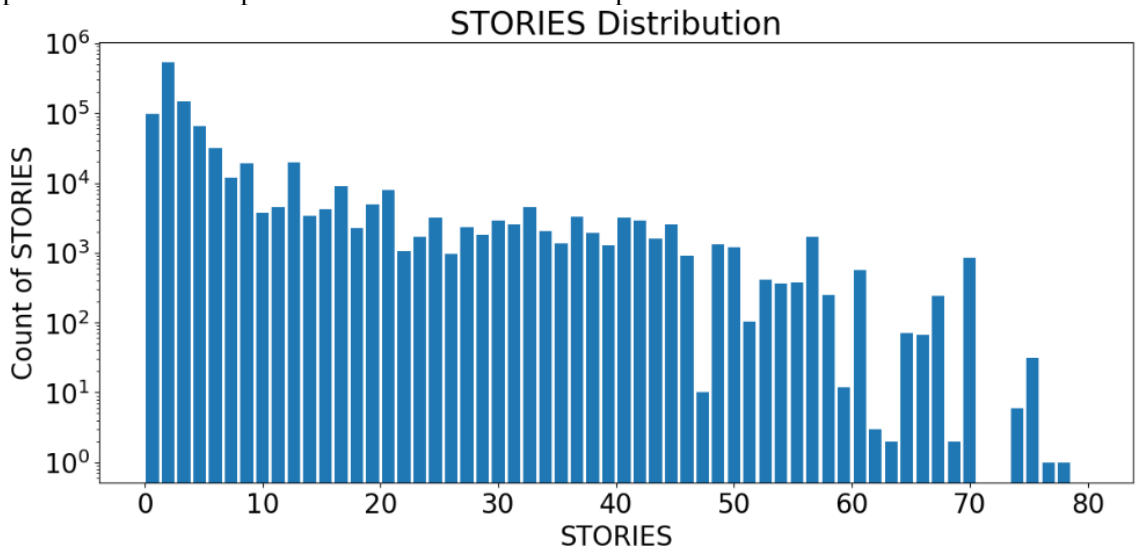
Description: Extension Indicator. The plot below shows the distribution of the values. The most common value is G, with a total count of 266970.

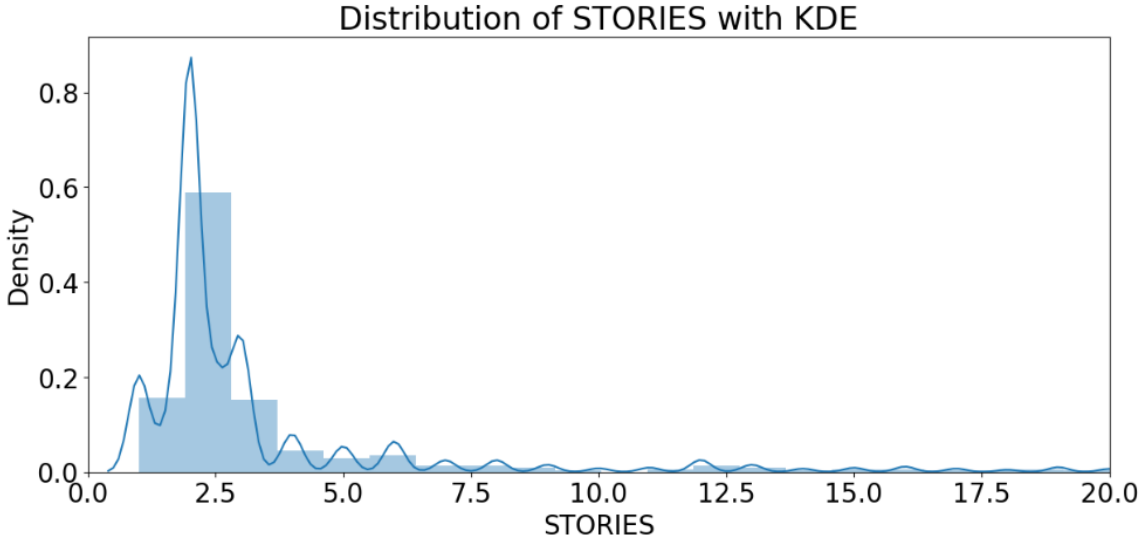


Field 13

Field Name: STORIES

Description: Number of stories in the building. The first plot shows the histogram of stories up to 80. The second plot shows the KDE of stories up to 20.

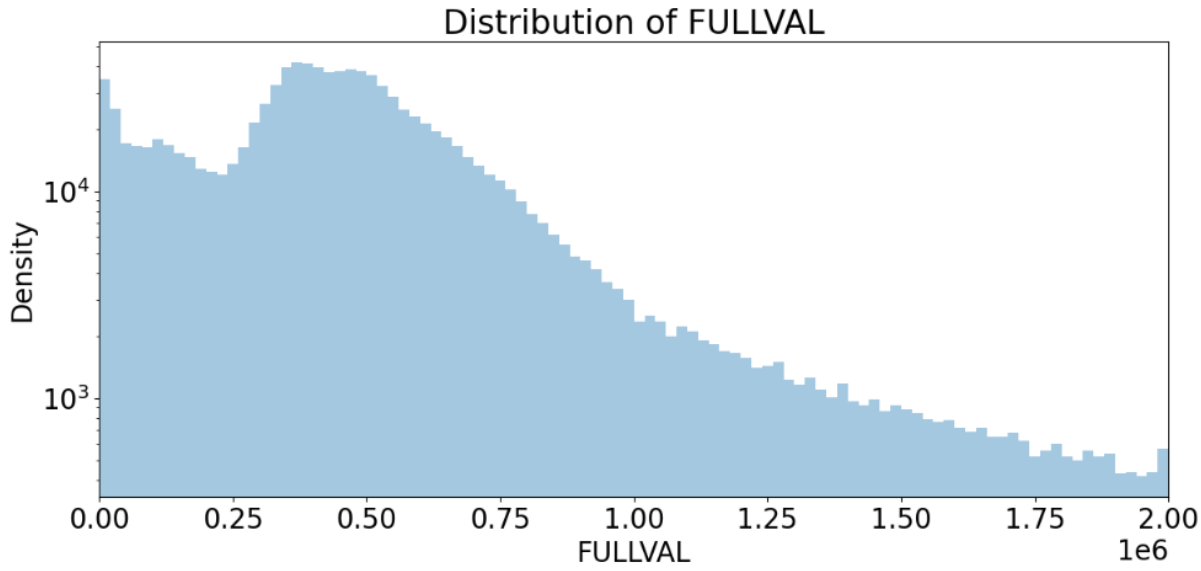




Field 14

Field Name: FULLVAL

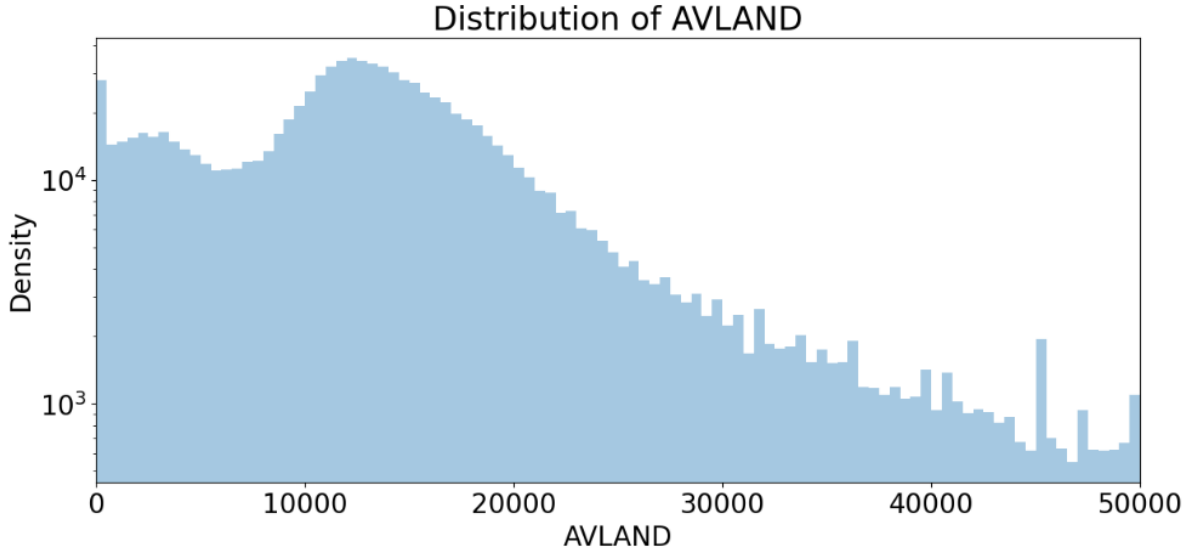
Description: Market Value. The plot shows the histogram of FULLVAL up to 2000000. The distribution appears to be bimodal. Most FULLVAL value seems to be around 500000. Surprisingly, a lot of the property seems to have 0 FULLVAL.



Field 15

Field Name: AVLAND

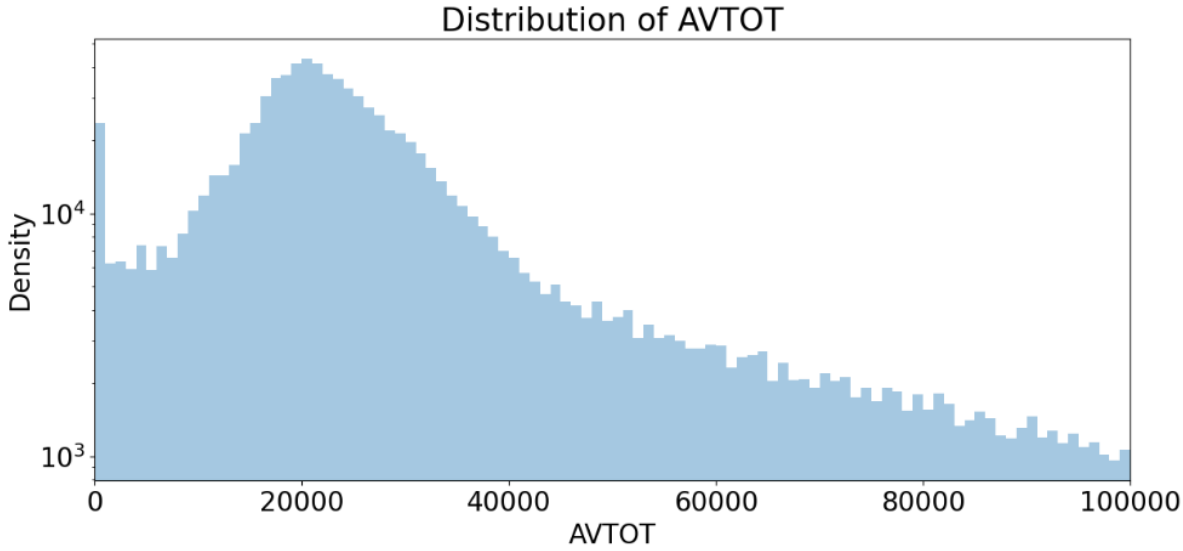
Description: Actual Value of Land. The plot below shows the histogram of AVLAND, where values greater than 50000 removed. Most AVLAND value seems to be around 15000. Surprisingly, a lot of the property seems to have 0 AVLAND.



Field 16

Field Name: AVTOT

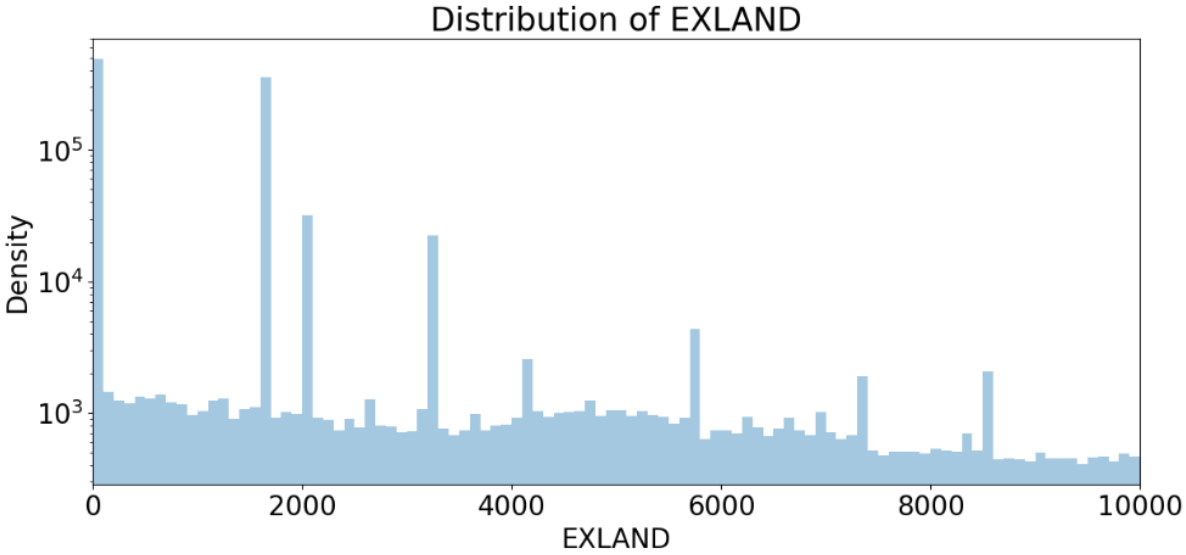
Description: Actual Total Value. The plot shows the histogram of AVTOT, with values greater than 100000 removed. Most AVTOT value seems to be around 20000. Surprisingly, a lot of the property seems to have 0 AVTOT.



Field 17

Field Name: EXLAND

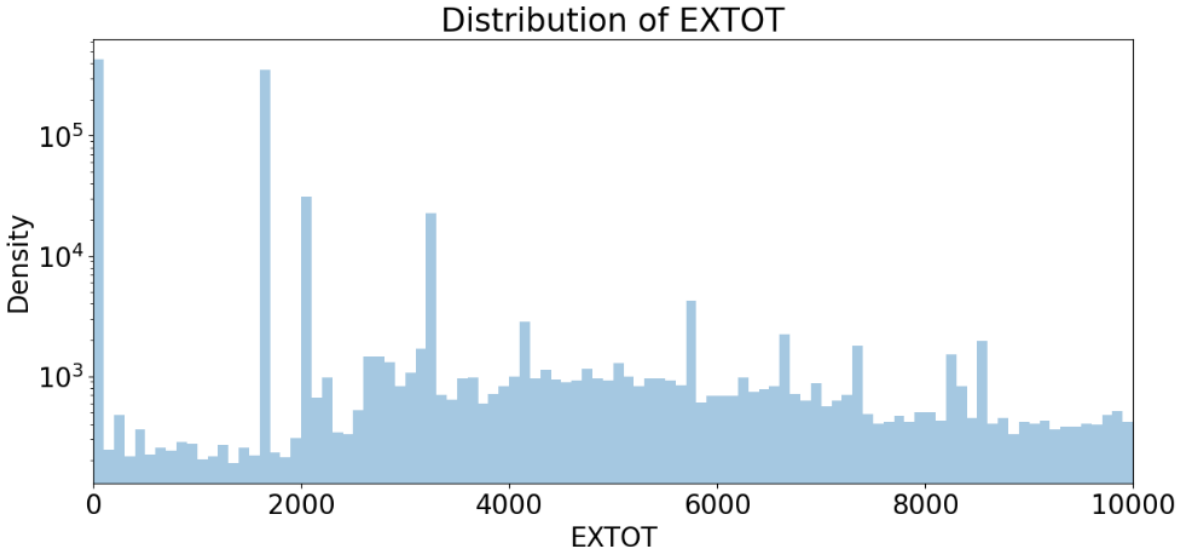
Description: Actual Exempt Land Value. The plot below shows the histrogarm of EXLAND, with values greater than 10000 removed. The distribution appears to be uniform, with discrete peaks around 0, 1750, 2000, 2600, 3250, 5800, 7500, and 8500.



Field 18

Field Name: EXTOT

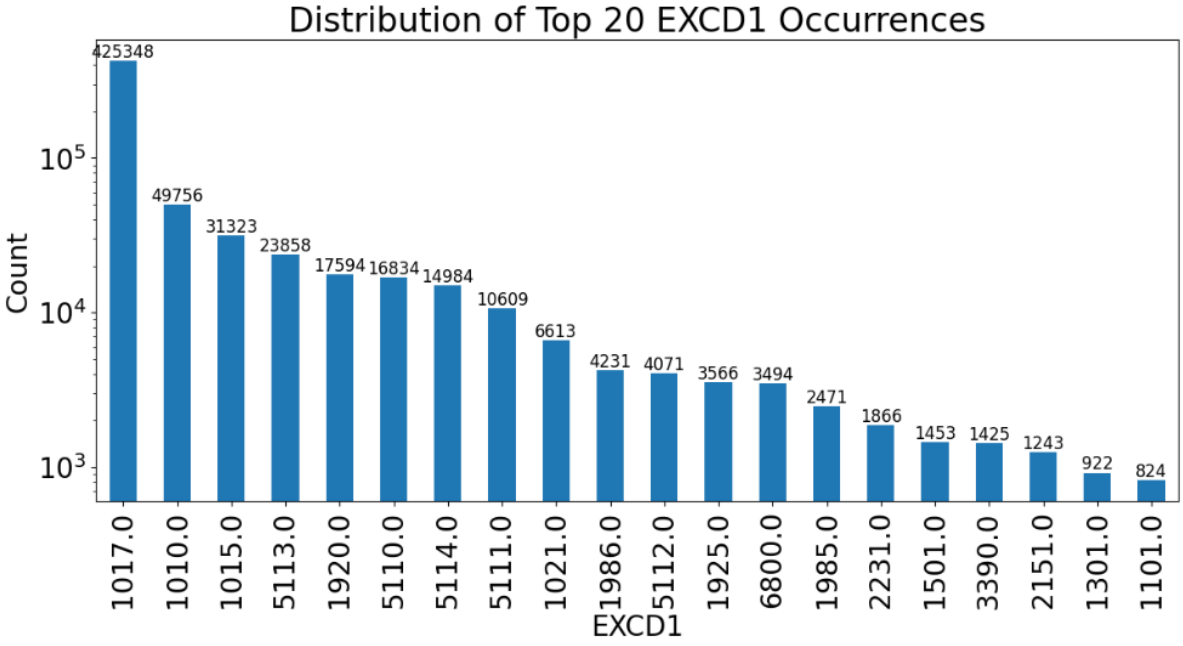
Description: Actual Exempt Land Total. The plot below shows the histrogarm of EXLAND, with values greater than 10000 removed. The distribution appears to be a somewhat uniform distribution, with discrete peaks around 0, 1750, 2000, 2600, 3250, 5800, 7500, and 8500. The peaks are very similar to the peaks in EXLAND.



Field 19

Field Name: EXCD1

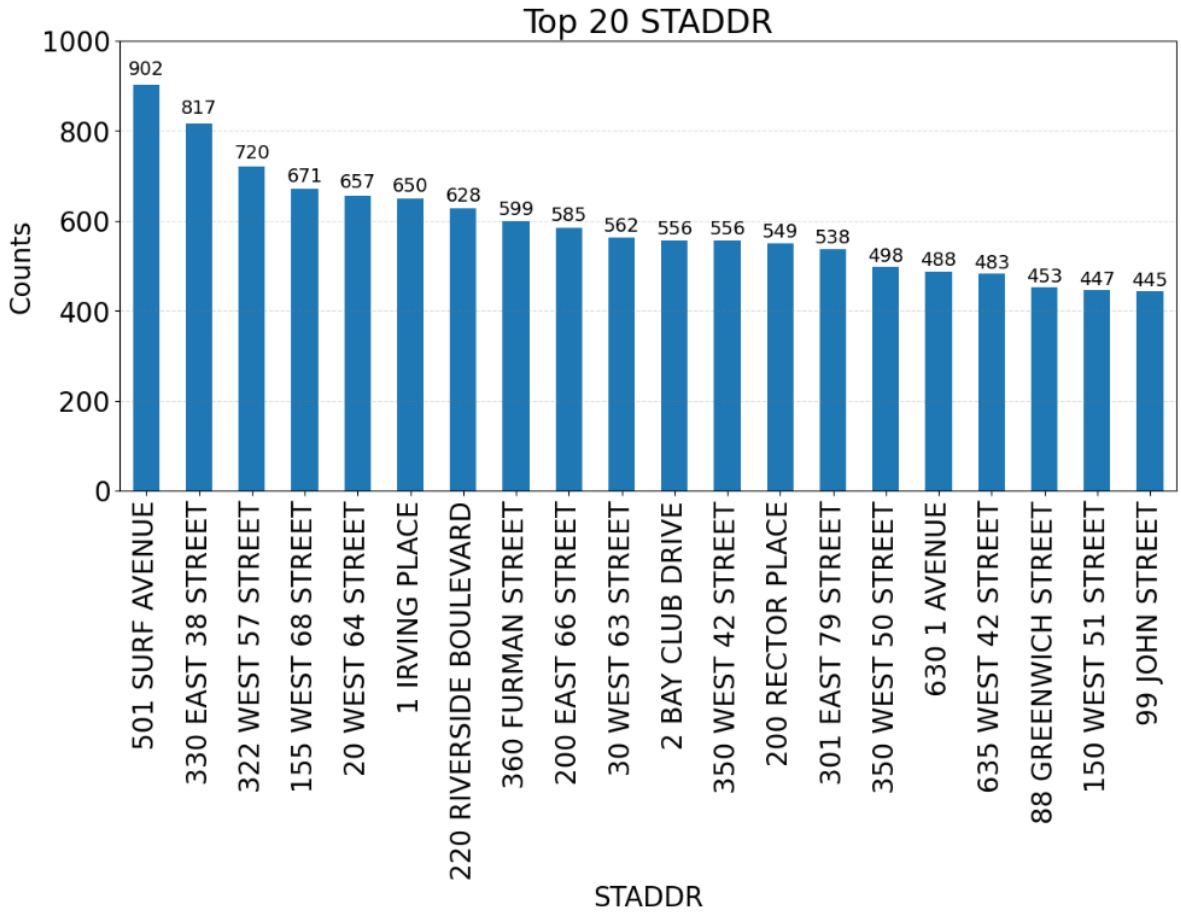
Description: Exemption Code 1. The plot below shows the distribution of the top 20 EXCD1 Values. The most common appears to be 1017, with a total count of 42534.



Field 20

Field Name: STADDR

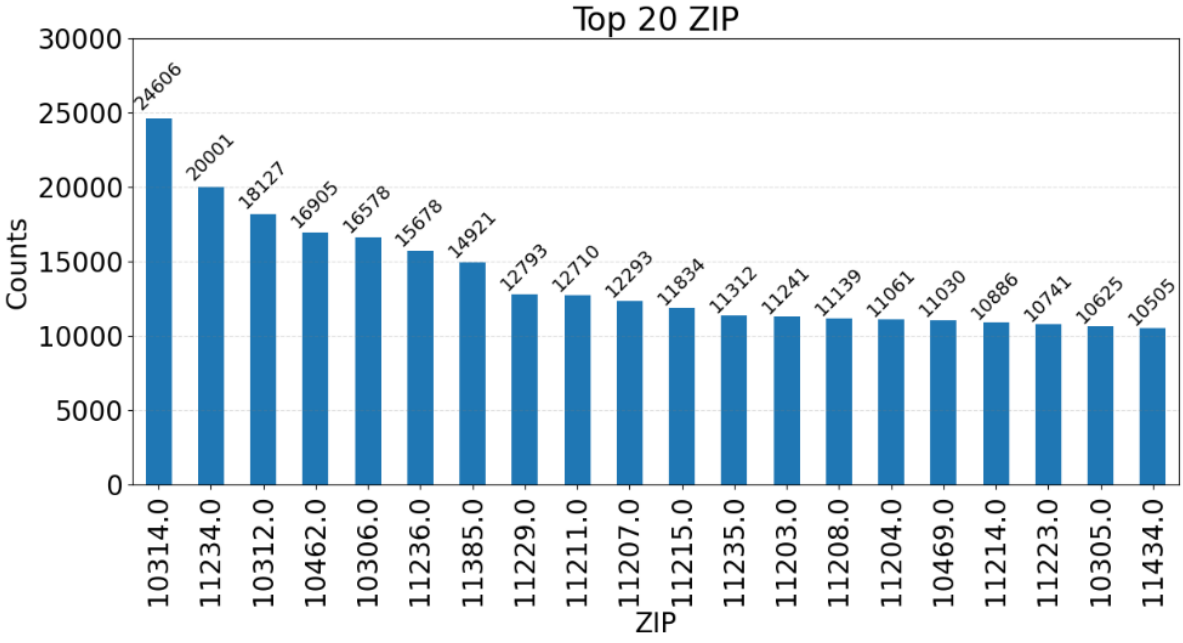
Description: Street Address of property. The plot below shows the distribution of the top 20 STADDR Values. The most common appears to be 501 Surf Avenue, with a total count of 902.



Field 21

Field Name: ZIP

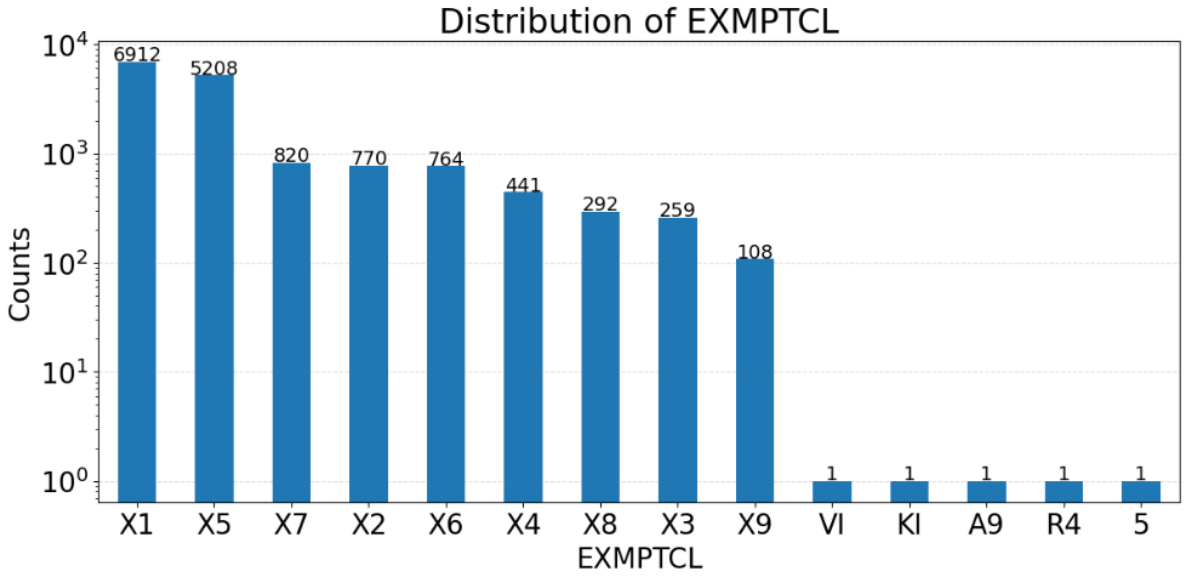
Description: Postal ZIP code of the property. The plot below shows the distribution of the top 20 ZIP Values. The most common appears to be 10314, with a total count of 24606.



Field 22

Field Name: EXMPTCL

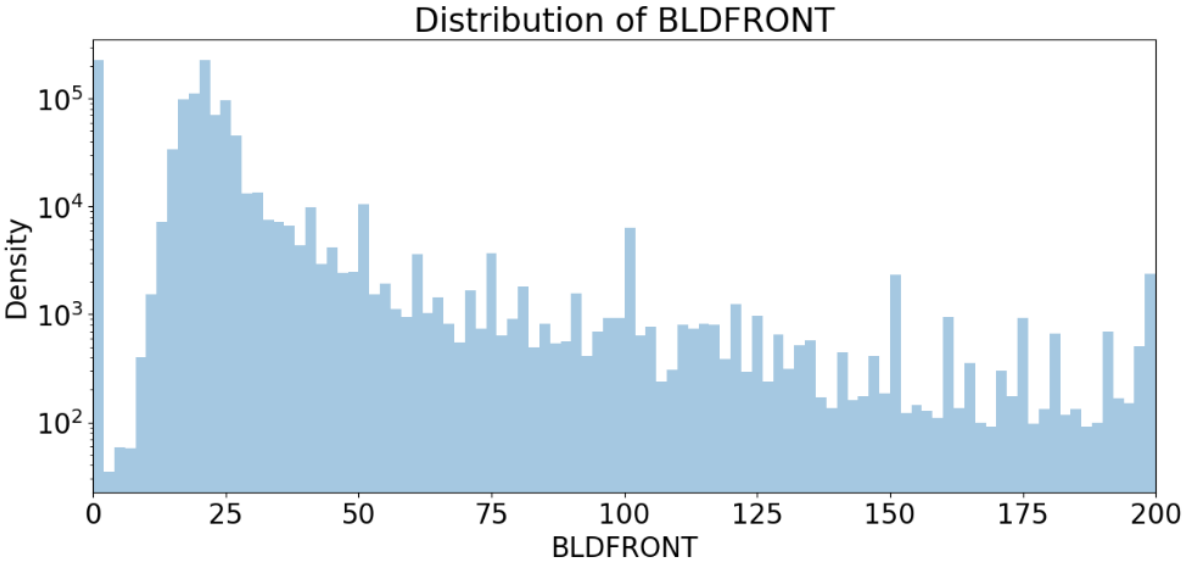
Description: Exemption Class. The plot below shows the distribution of all values in EXMPTCL.

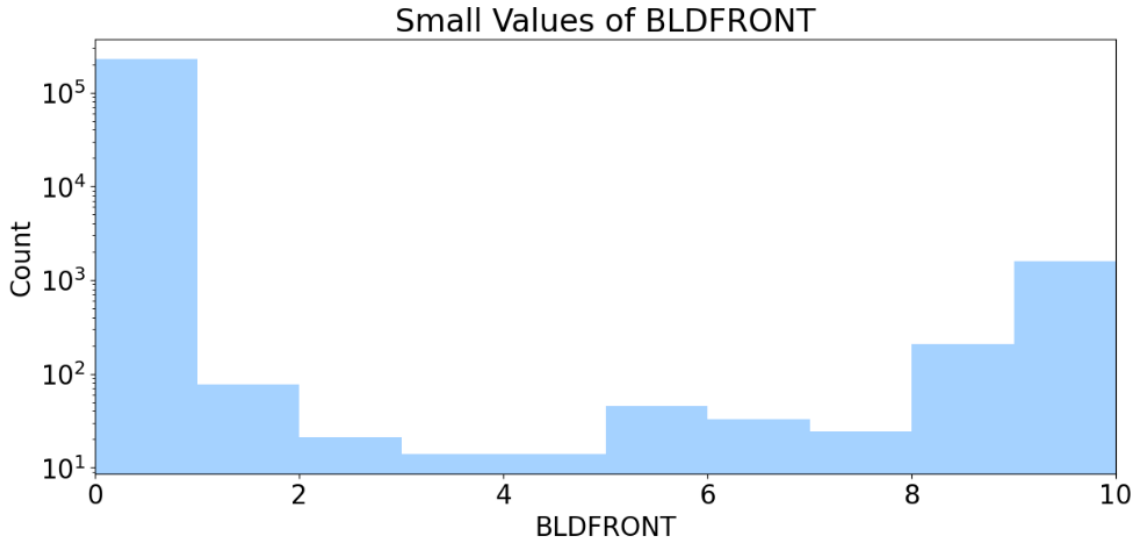


Field 23

Field Name: BLDFRONT

Description: Building Width. The plot below shows the histogram of BLDFRONT up to 200. The second plot shows the histogram of BLDFRONT up to 10, surprisingly, a lot of the values seems to be 0.

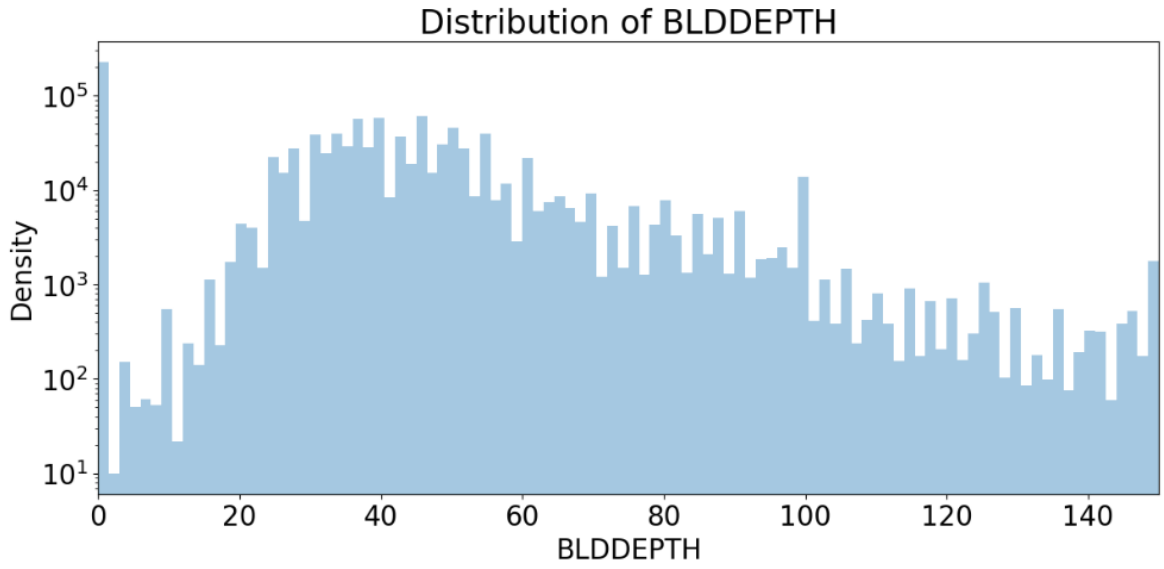


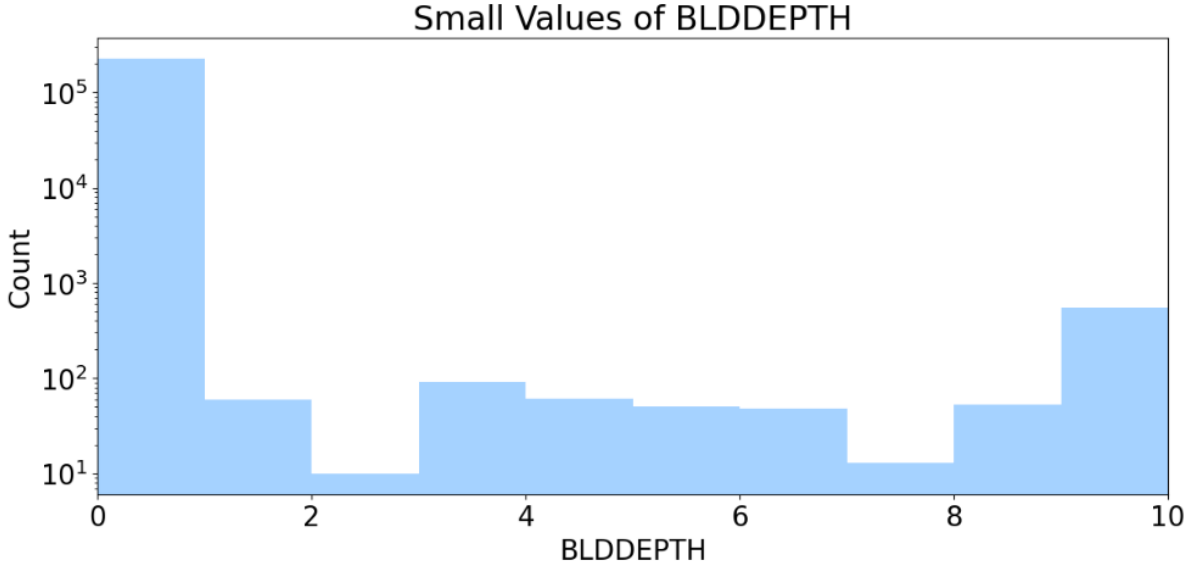


Field 24

Field Name: BLDDEPTH

Description: Building Depth. The plot below shows the histogram of BLDDEPTH up to 150. The second plot shows the histogram of BLDDEPTH up to 10, surprisingly, a lot of the values seems to be 0.

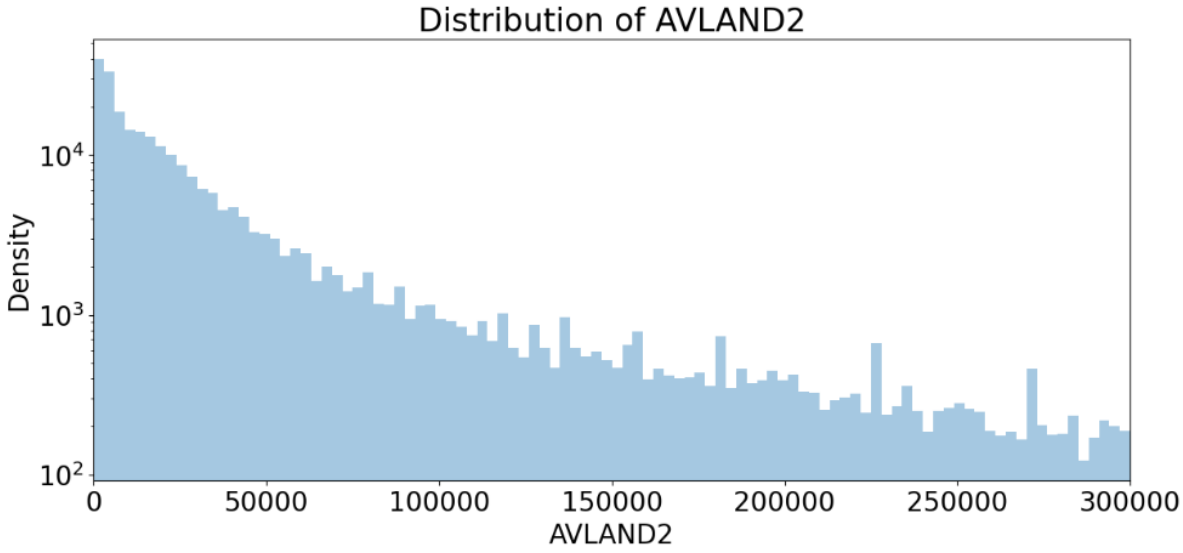




Field 25

Field Name: AVLAND2

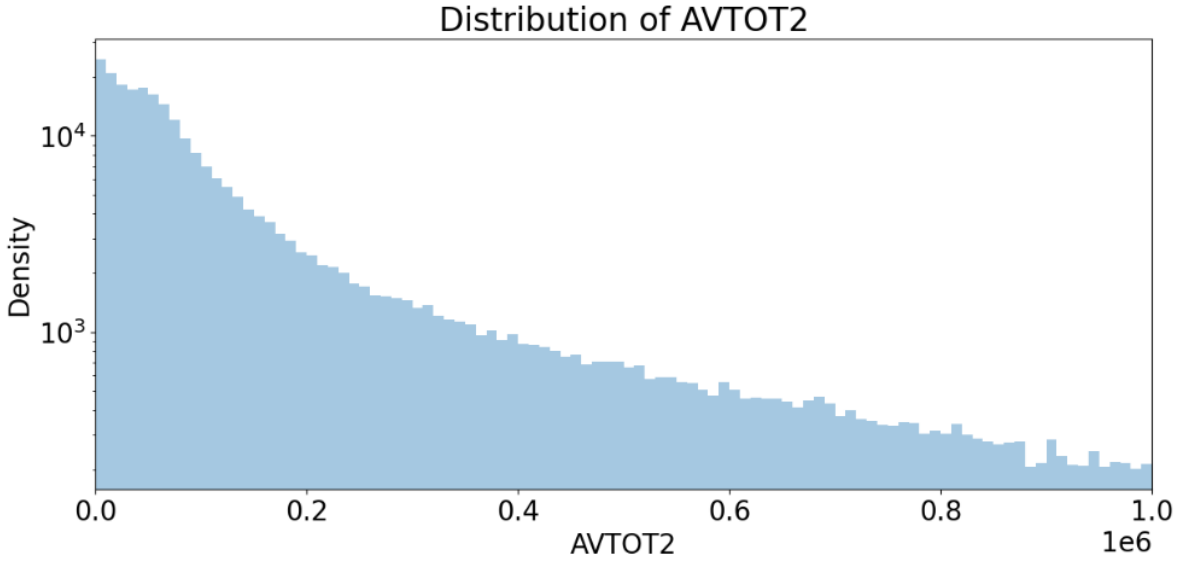
Description: Transitional Land value. The plot below shows the distribution of AVLAND2 up to 300000. It appears to be skewed right.



Field 26

Field Name: AVTOT2

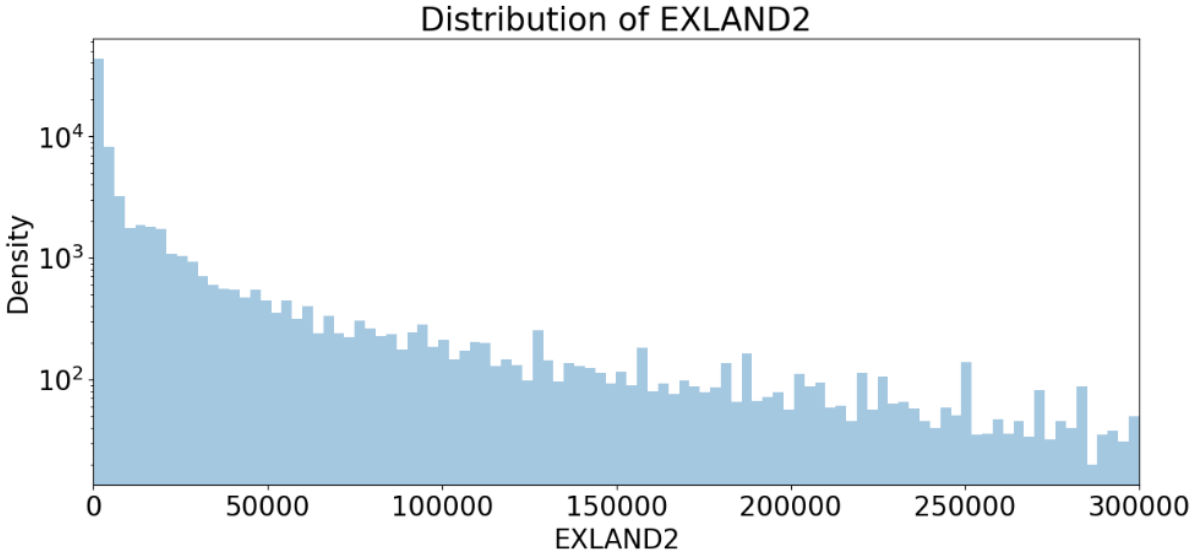
Description: Transitional Total Value. The plot below shows the distribution of AVTOT2 up to 1000000. It appears to be skewed right.

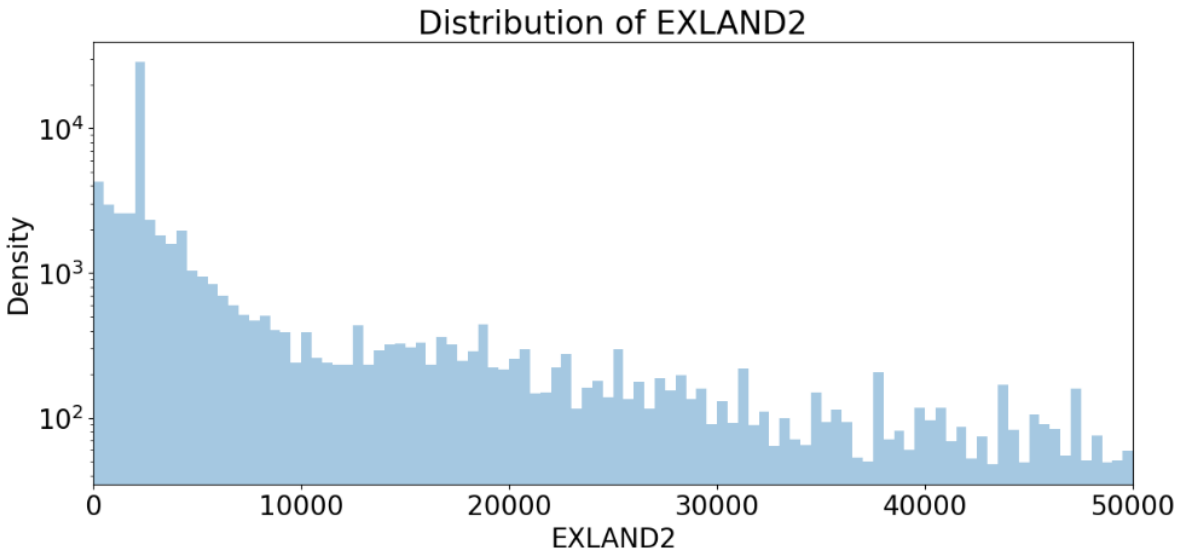


Field 27

Field Name: EXLAND2

Description: Transitional exemption land value. The first plot below shows the distribution of EXLAND2 up to 300000. The second plot shows the distribution of EXLAND2 up to 50000. There appears to be a peak around 3000. The distributions are both skewed right.

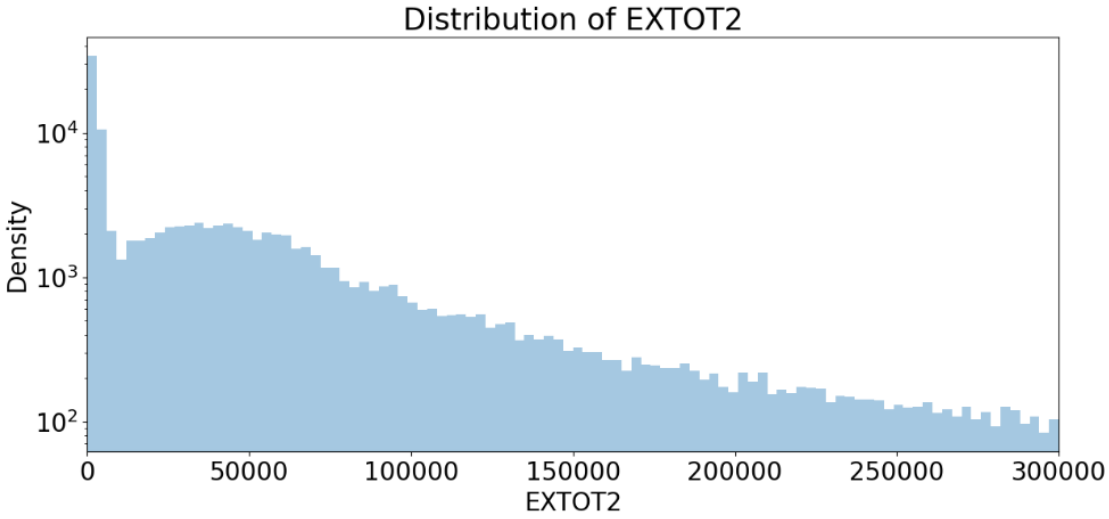


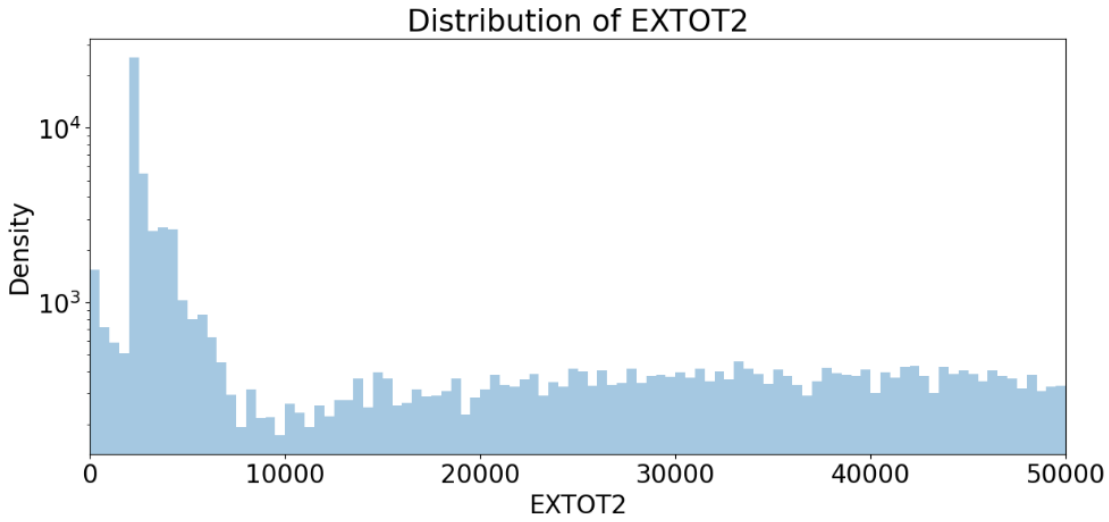


Field 28

Field Name: EXTOT2

Description: Transitional exemption land total. The first plot below shows the distribution of EXTOT2 up to 300000. The second plot shows the distribution of EXTOT2 up to 50000. There appears to be a peak around 3000, close to the peak of EXLAND2. The distributions are both skewed right.

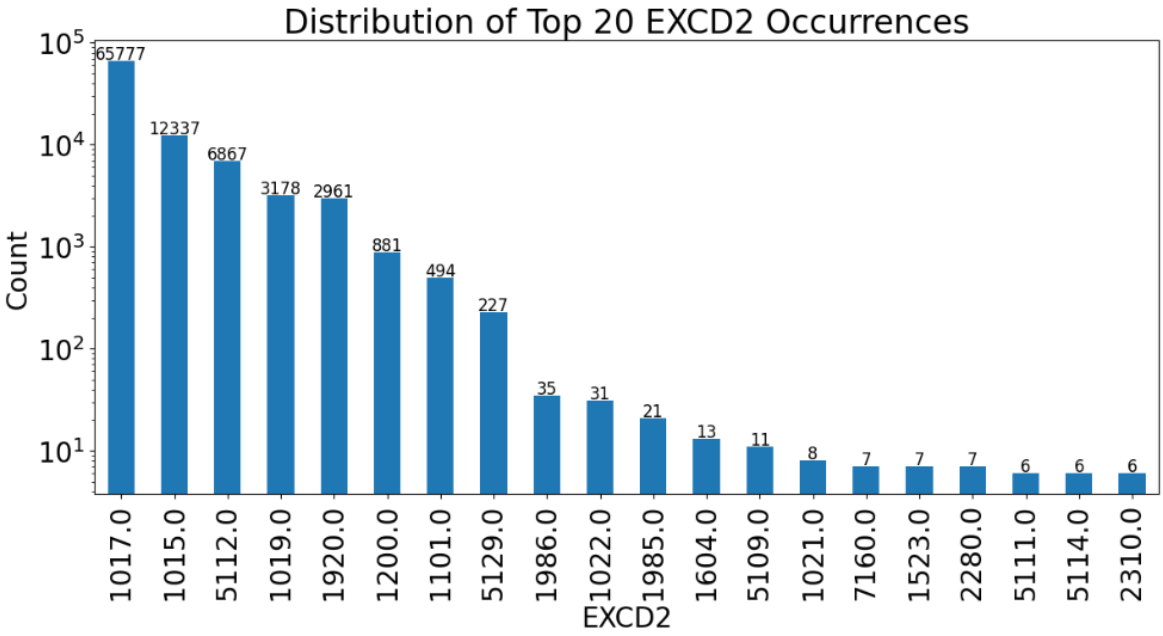




Field 29

Field Name: EXCD2

Description: Exemption Code 2. The plot below shows the distribution of the top 20 EXCD2 Values. The most common appears to be 1017, with a total count of 65777. This is also the most common value of Exemption Code 1 (EXCD1).



Field 30

Field Name: Period

Description: Assessment Period. There is only one value in this field, which is '**FINAL**.'

Field 31

Field Name: Year

Description: Assessment Year. There is only one value in this field, which is '**2010/11**'

Field 32

Field Name: Valtype

Description: There is only one value in this field, which is **AC-TR**.