Tax Fraud Detection for New York Property Data

Project Report

Reuben Chatterjee
PID - A59026323

Table of Contents

1. Executive Summary

The New York Property Tax Fraud Detection project aimed to address the significant issue of property tax fraud within New York City's extensive property database. Property tax fraud can manifest in various forms, such as underreporting property values, misclassifying property types, or manipulating property characteristics to reduce tax liabilities. Such fraudulent activities not only result in substantial revenue losses for the city but also create an unfair tax burden on honest property owners. The primary objective of this project was to develop a robust system capable of identifying potentially fraudulent property records, thereby aiding city authorities in ensuring tax compliance and fairness.

The project successfully identified numerous properties with suspicious characteristics that could indicate potential tax fraud. By employing a combination of data cleaning, variable creation, dimensionality reduction, and advanced anomaly detection algorithms, we were able to flag properties with inconsistent valuation metrics and unusual property characteristics. The key findings include:

1. **High Fraud Scores:** Several properties exhibited high fraud scores based on the combined results of Isolation Forest and Local Outlier Factor algorithms.
2. **Inconsistent Valuation Metrics:** Properties were found with full market values, assessed land values, and total assessed values that did not align with their physical characteristics or reported usage.
3. **Unusual Size and Value Ratios:** The analysis revealed anomalies in size and value ratios, indicating discrepancies between reported property dimensions and their valuations.
4. **Significant Irregularities:** Detailed case studies of the top flagged properties highlighted substantial irregularities that warrant further investigation by city authorities.

2. Description of Data

The Dataset is **Property Valuation and Assessment Data,** covering various attributes of properties in the dataset. It includes a total of **1,070,994 records** and **32 fields**. It covers property assessments for the **year 2010/11**. This dataset provides insights into property characteristics and valuations, which could be used for real estate analysis, tax assessments, and urban planning. The dataset includes specific code definitions and ranges for certain fields such as boroughs (Manhattan, Bronx, Brooklyn, Queens, Staten Island), building classes (1-1 to 1-3 unit residence, apartments, utilities, all others), and easement types (Non-Air Rights, Transit Easement, etc.)

Summary Tables

Numeric Fields Table :

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
LTFRONT	numeric	1070994	100.0%	169108	0	9999	36.63530141	74.03	0
LTDEPTH	numeric	1070994	100.0%	170128	0	9999	88.861594	76.4	100
STORIES	numeric	1014730	94.7%	0	1	119	5.006917899	8.37	2
FULLVAL	numeric	1070994	100.0%	13007	0	6150000000	874264.5054	11582425.58	0
AVLAND	numeric	1070994	100.0%	13009	0	2668500000	85067.91867	4057258.16	0
AVTOT	numeric	1070994	100.0%	13007	0	4668308947	227238.1687	6877526.09	0
EXLAND	numeric	1070994	100.0%	491699	0	2668500000	36423.89069	3981573.93	0
EXTOT	numeric	1070994	100.0%	432572	0	4668308947	91186.98168	6508399.78	0
BLDFRONT	numeric	1070994	100.0%	228815	0	7575	23.042769961	35.58	0
BLDDEPTH	numeric	1070994	100.0%	228853	0	9393	39.922836617	42.71	0
AVLAND2	numeric	282726	26.4%	0	3	2371005000	246235.7193	6178951.64	2408
AVTOT2	numeric	282732	26.4%	0	3	4501180002	713911.4362	11652508.34	750
EXLAND2	numeric	87449	8.2%	0	1	2371005000	351235.6843	10802150.91	2090
EXTOT2	numeric	130828	12.2%	0	7	4501180002	656768.2819	16072448.75	2090

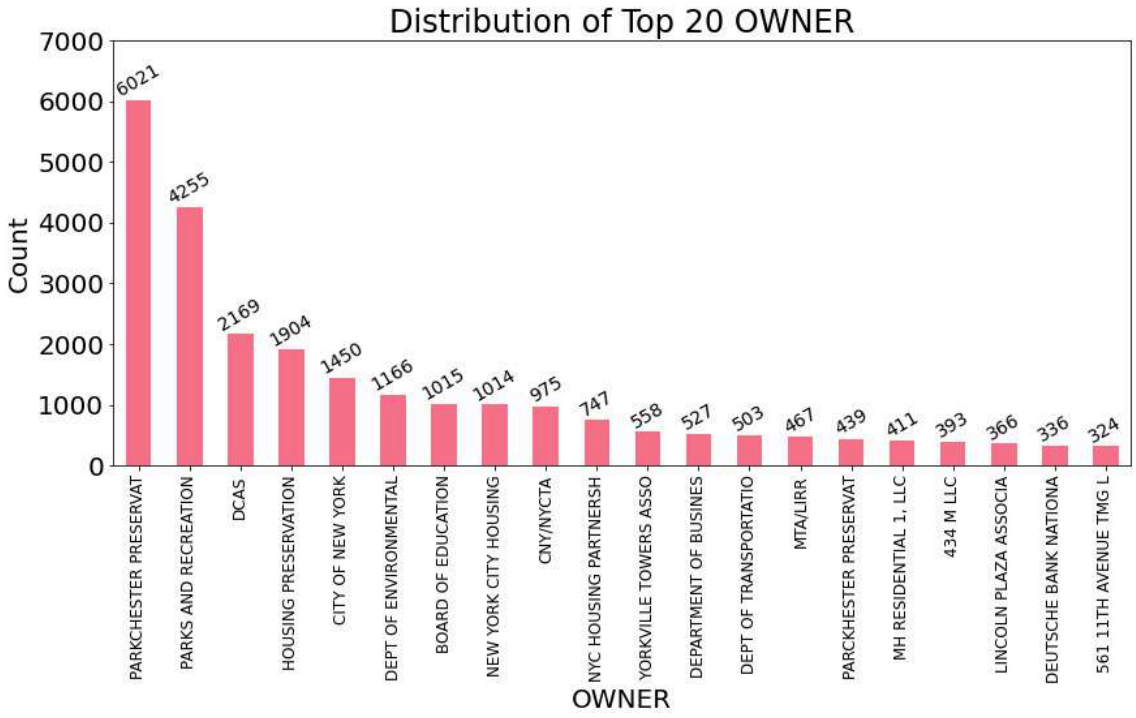
Categorical Fields Table :

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
RECORD	categorical	1070994	100.0%	0	1070994	1
BBLE	categorical	1070994	100.0%	0	1070994	1000010101
BORO	categorical	1070994	100.0%	0	5	4
BLOCK	categorical	1070994	100.0%	0	13984	3944
LOT	categorical	1070994	100.0%	0	6366	1
EASEMENT	categorical	4636	0.4%	0	12	E
OWNER	categorical	1039249	97.0%	0	863347	PARKCHESTER PRESERVAT
BLDGCL	categorical	1070994	100.0%	0	200	R4
TAXCLASS	categorical	1070994	100.0%	0	11	1
EXT	categorical	354305	33.1%	0	3	G
EXCD1	categorical	638488	59.6%	0	129	1017
STADDR	categorical	1070318	99.9%	0	839280	501 SURF AVENUE
ZIP	categorical	1041104	97.2%	0	196	10314
EXMPTCL	categorical	15579	1.5%	0	14	X1
EXCD2	categorical	92948	8.7%	0	60	1017
PERIOD	categorical	1070994	100.0%	0	1	FINAL
YEAR	categorical	1070994	100.0%	0	1	2010/11
VALTYPE	categorical	1070994	100.0%	0	1	AC-TR

Some notable fields and their distributions:

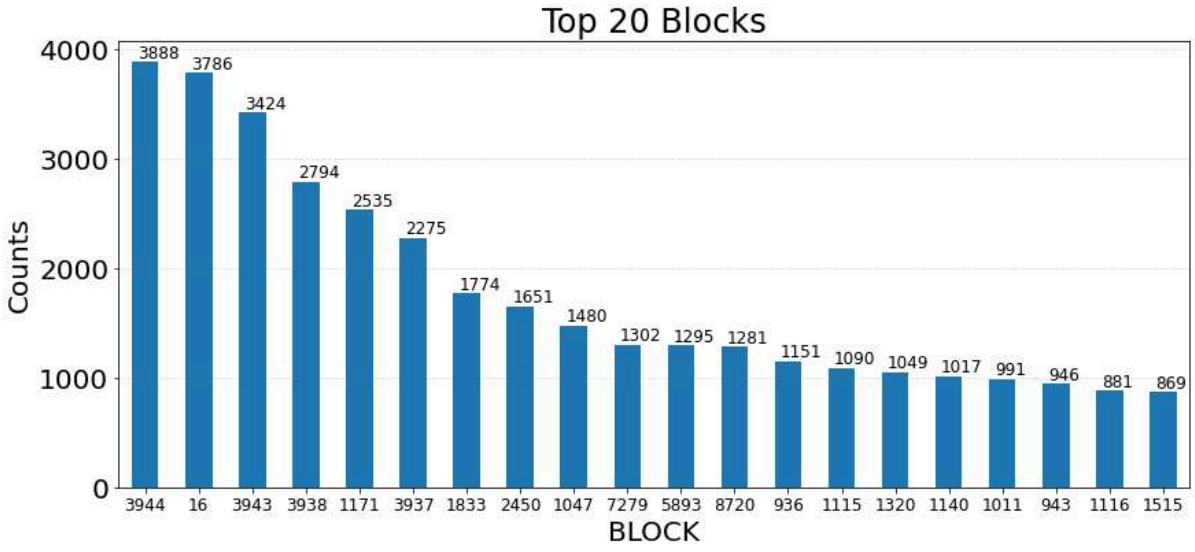
1. **Field Name : OWNER**

   Description: Name of the property owner. This field is 97.03% populated



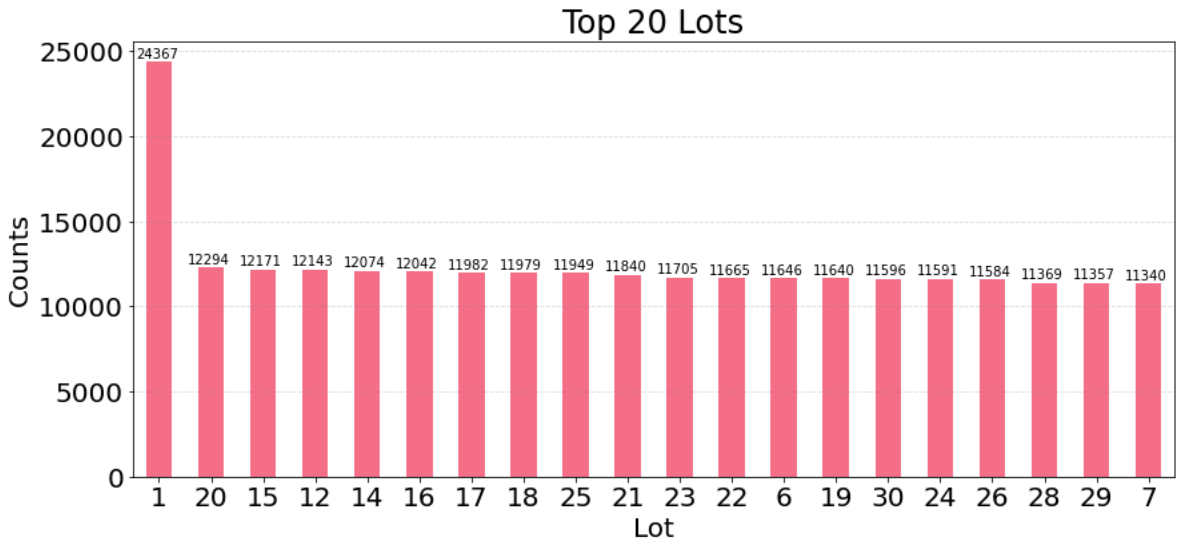
2. **Field Name : BLOCK**

   Description: Block number within the borough, indicating a specific area.



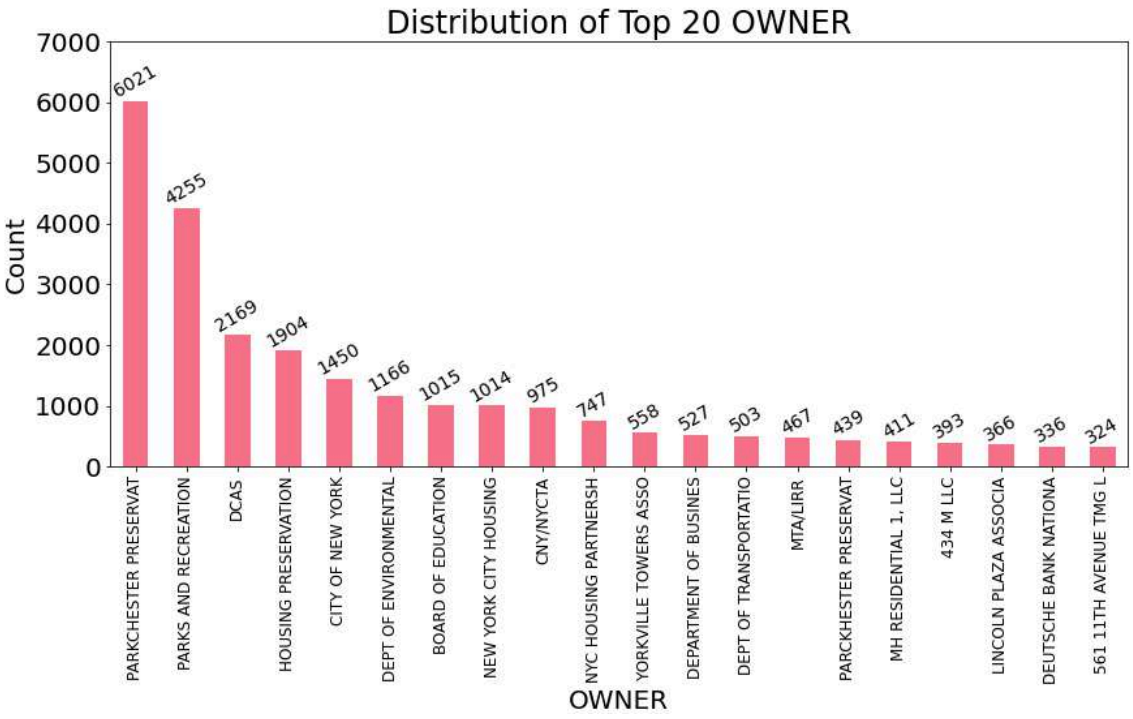
3. **Field Name : LOT**

Description: Lot number within the block, specifying a particular parcel of land.



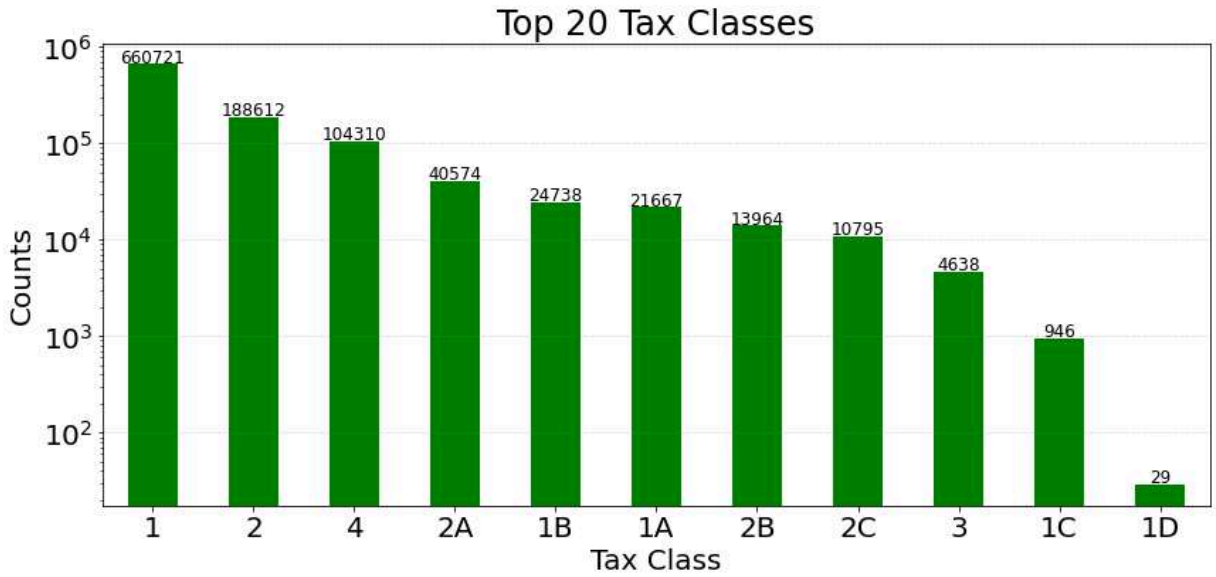
4. **Field Name : OWNER**

Description: Name of the property owner. This field is 97.03% populated



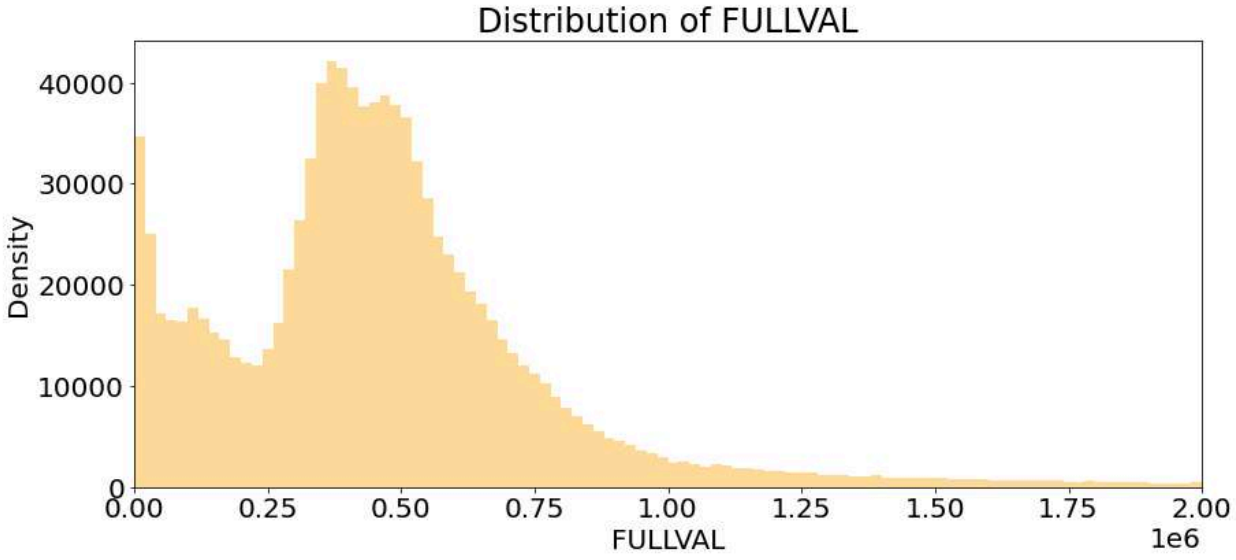
5. **Field Name : TAXCLASS**

Description: Tax class of the property, which affects its tax rate (e.g., 1 = 1-3 Unit Residence, 2 = Apartments).



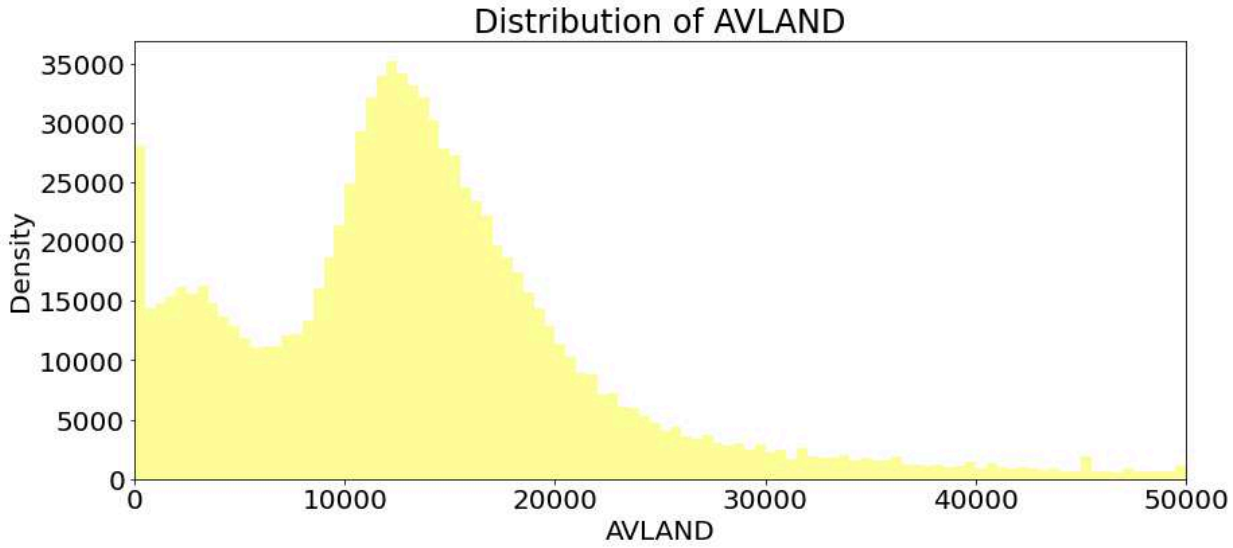
6. **Field Name : FULLVAL**

Description: Full market value of the property. Appears to have a bimodal distribution.



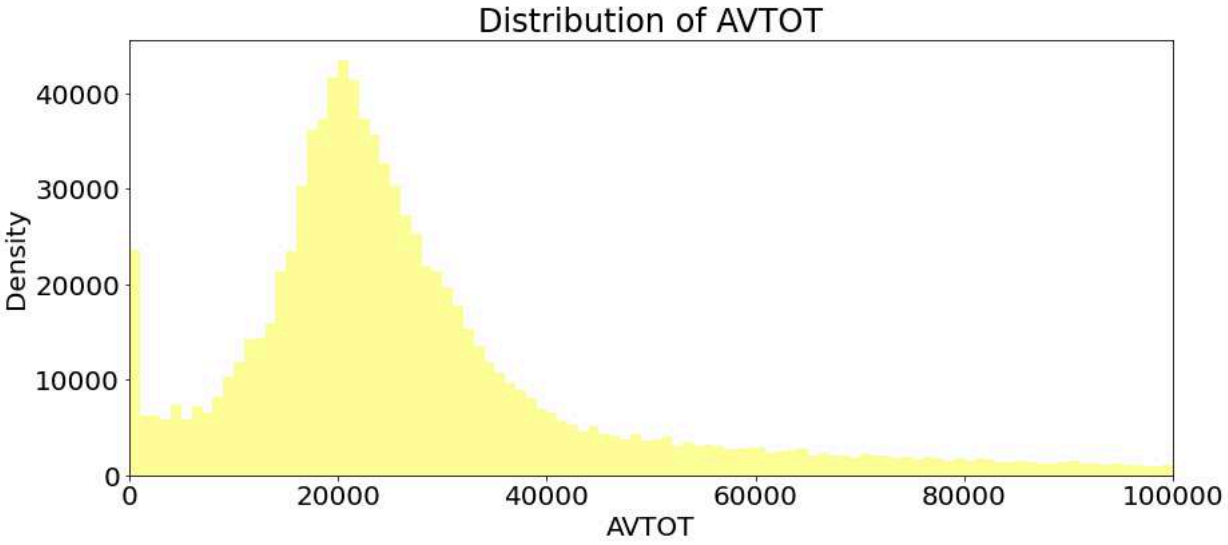
7. **Field Name : AVLAND**

Description: Assessed value of the land portion of the property.



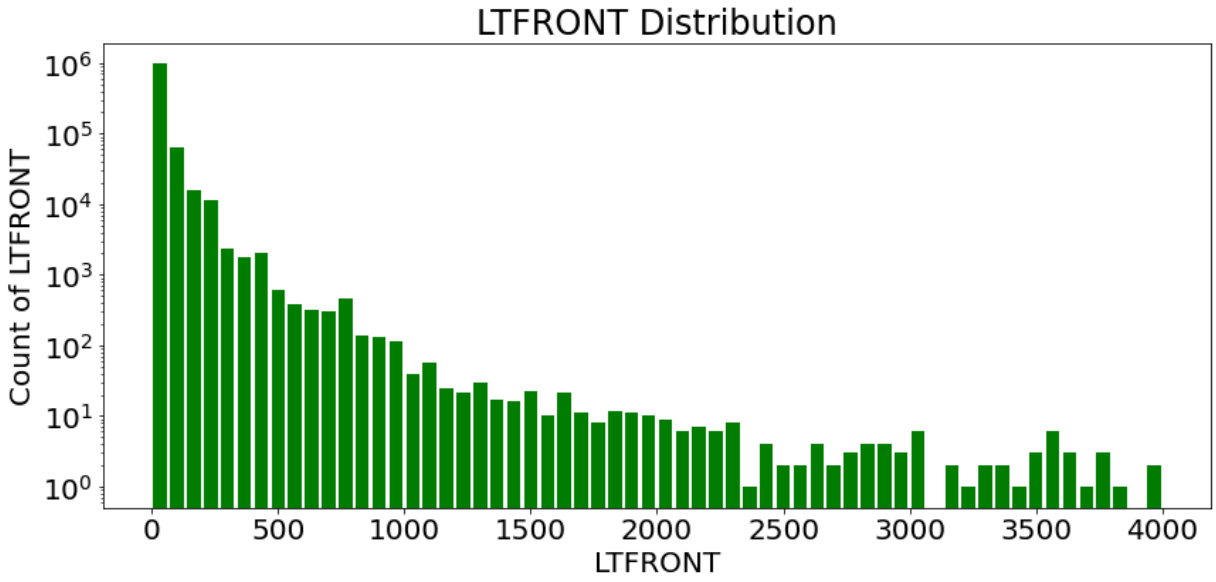
8. Field Name : AVTOT

Description: Total assessed value of the property, including land and improvements.

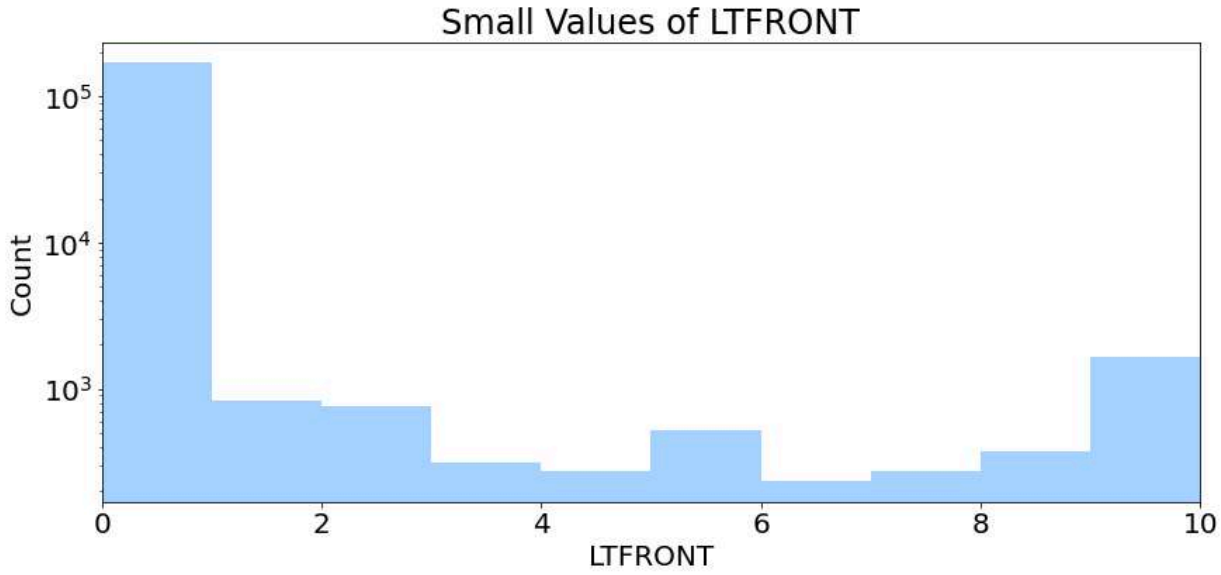


9. **Field Name : LTFRONT**

Description: Width of the lot's front in feet.

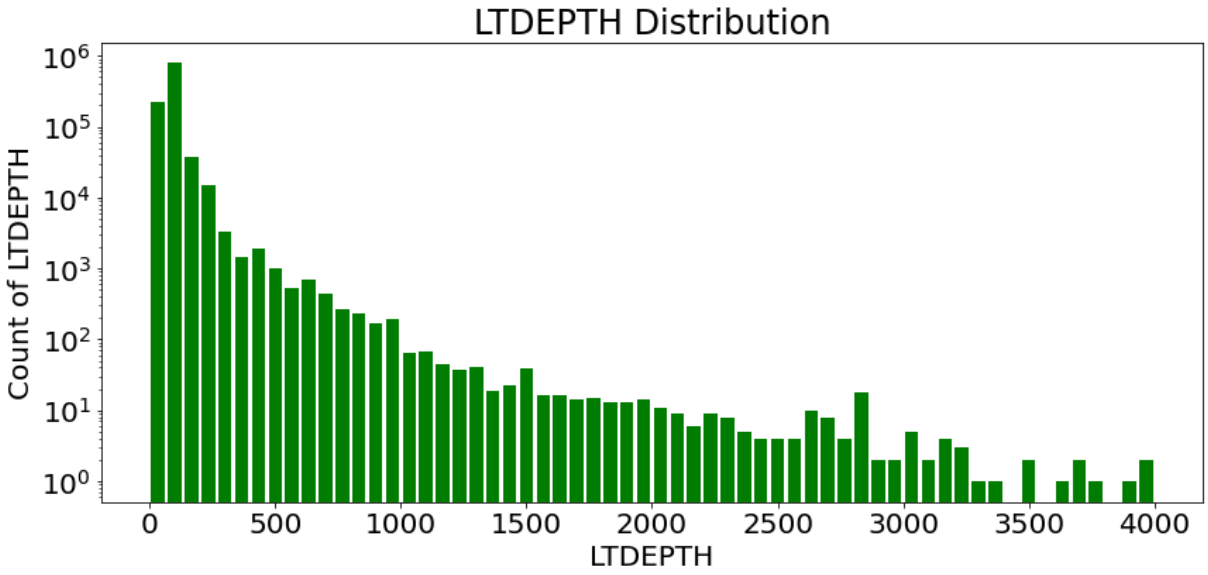


Additionally, there are numerous small values of LTFRONT, which are visualized here.

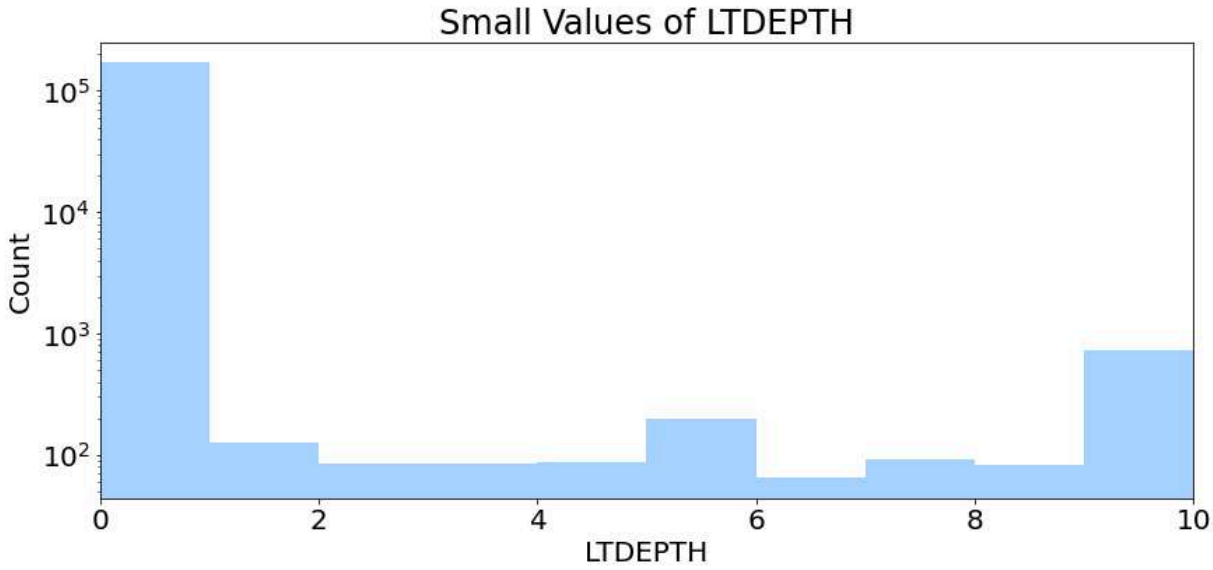


10. Field Name : LTDEPTH

Description: Depth of the lot in feet.



Additionally, there are numerous small values of LTDEPTH, which are visualized here.



3. Data Cleaning

Data Cleaning

In this section, I describe the process of data cleaning and imputation performed on the NY Property dataset to handle missing and zero values. The objective was to ensure that all fields have reasonable and consistent values for further analysis.

- Initially, I identified properties owned by government entities. These properties were marked by specific keywords in the OWNER field.
- A total of 12,349 records were identified as government properties.

These records were deemed irrelevant for the analysis as we are looking for property tax fraud; and government owned properties do not make sense to be included. Hence they were removed from the dataset, reducing the dataset from its original size to exclude these 12,349 records.

ZIP Code Imputation:

- The dataset had 20,431 records with missing ZIP codes.
- 364 records had missing street addresses (STADDR), while the BORO field had no missing values
- I created a new column staddr_boro by concatenating STADDR and BORO.
- Next, I mapped ZIP codes based on unique combinations of staddr_boro.
- Then I filled in the missing ZIP codes using this mapping, which filled in 2,832 records.
- For remaining missing ZIP codes, I iterated through adjacent records and filled ZIP codes based on neighboring values. This filled an additional 9,491 records.
- Finally, the remaining 8,108 missing ZIP codes were filled using the previous record's ZIP code.

.

Thus, all missing ZIP codes were successfully filled, resulting in zero missing ZIP codes.

FULLVAL, AVLAND, and AVTOT Imputation

The imputation process for these three fields was identical and produced the same outputs.

- I identified 10,025 records with either missing or zero values in the FULLVAL and AVTOT fields, while the AVLAND field had 10,027.
- I grouped records by TAXCLASS, BORO, and BLDGCL and calculated the group average to fill missing values. This filled 2,718 records, leaving 7,307 remaining.
- So I then grouped by TAXCLASS and BORO, and filled missing values with group averages This filled 6,921 records, leaving 386 remaining.
- Finally, I grouped by TAXCLASS and filled in the remaining missing values. All records now had reasonable values for FULLVAL.

STORIES Imputation

- The STORIES field had 42,030 missing values.
- I calculated the mode of STORIES within groups defined by BORO and BLDGCL, filling 3,108 missing values.
- Grouped by TAXCLASS and filled remaining missing values using group averages. This completed the imputation for all STORIES values.

LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH Imputation

- Values of zero and one for these fields were considered invalid and treated as missing (replaced with NaN).
- I grouped by TAXCLASS, BORO, and BLDGCL, and filled in missing values using group averages.
- Then I further grouped by TAXCLASS alone and completed the imputation process.

4. Variable Creation

The variable creation process involves generating new metrics from existing data to identify properties that exhibit unusual characteristics, which may suggest potential anomalies or outliers.

New Variables Created:

Size Variables

- Lot Area (lt_size): Product of LTFONT and LTDEPTH.
- Building Area (bld_size): Product of BLDFRONT and BLDDEPTH.
- Building Volume (bld_vol): Product of bld_size and STORIES.

Value Ratios
I calculate nine primary value ratios, where each value field (FULLVAL, AVLAND, AVTOT) is divided by each of the three size variables:

- r1 = FULLVAL / lt_size
- r2 = FULLVAL / bld_size
- r3 = FULLVAL / bld_vol
- r4 = AVLAND / lt_size
- r5 = AVLAND / bld_size
- r6 = AVLAND / bld_vol
- r7 = AVTOT / lt_size
- r8 = AVTOT / bld_size
- r9 = AVTOT / bld_vol

Inverse Value Ratios
To identify unusually small values, we also calculate the inverse of these ratios:

- r1_inv = 1 / r1
- r2_inv = 1 / r2

- r3_inv = 1 / r3
- r4_inv = 1 / r4
- r5_inv = 1 / r5
- r6_inv = 1 / r6
- r7_inv = 1 / r7
- r8_inv = 1 / r8
- r9_inv = 1 / r9

For each pair of ratios (r and r_inv), we keep only the larger value because we are interested in both unusually large and small values.

Grouped Averages

We calculate the grouped averages of these nine ratios by grouping records by ZIP and TAXCLASS:

- Grouped by ZIP: r1_zip5, r2_zip5, r3_zip5, r4_zip5, r5_zip5, r6_zip5, r7_zip5, r8_zip5, r9_zip5
- Grouped by TAXCLASS: r1_taxclass, r2_taxclass, r3_taxclass, r4_taxclass, r5_taxclass, r6_taxclass, r7_taxclass, r8_taxclass, r9_taxclass

Standardized Variables

Each of the nine primary value ratios is then standardized by dividing by the two scale factors from these groupings:

- Standardized by ZIP: r1 / r1_zip5, r2 / r2_zip5, etc.
- Standardized by TAXCLASS: r1 / r1_taxclass, r2 / r2_taxclass, etc.

Additional Variables

- Value Ratio (value_ratio): Indicates how well the three value fields relate. Calculated as FULLVAL / (AVLAND + AVTOT), then normalized to a mean of 1.
- Size Ratio (size_ratio): Compares building size to lot size. Calculated as BLDFRONT * BLDDEPTH / (LTFONT * LTDEPTH).

The table of the 29 new created variables:

Sr. No.	Variable	Description
1	r1	FULLVAL / lt_size (or r1_inv i.e 1 / r1)
2	r2	FULLVAL / bld_size (or r2_inv i.e 1 / r2)
3	r3	FULLVAL / bld_vol (or r3_inv i.e 1 / r3)
4	r4	AVLAND / lt_size (or r4_inv i.e 1 / r4)
5	r5	AVLAND / bld_size (or r5_inv i.e 1 / r5)
6	r6	VLAND / bld_vol (or r6_inv i.e 1 / r6)
7	r7	AVTOT / lt_size (or r7_inv i.e 1 / r7)
8	r8	AVTOT / bld_size (or r8_inv i.e 1 / r8)
9	r9	AVTOT / bld_vol (or r9_inv i.e 1 / r9)
10	r1_zip5	Grouped average of r1 when grouped by zip
11	r2_zip5	Grouped average of r2 when grouped by zip
12	r3_zip5	Grouped average of r3 when grouped by zip
13	r4_zip5	Grouped average of r4 when grouped by zip
14	r5_zip5	Grouped average of r5 when grouped by zip
15	r6_zip5	Grouped average of r6 when grouped by zip
16	r7_zip5	Grouped average of r7 when grouped by zip
17	r8_zip5	Grouped average of r8 when grouped by zip
18	r9_zip5	Grouped average of r9 when grouped by zip
9	r1_taxclass	Grouped average of r1 when grouped by taxclass
20	r2_taxclass	Grouped average of r2 when grouped by taxclass
21	r3_taxclass	Grouped average of r3 when grouped by taxclass
22	r4_taxclass	Grouped average of r4 when grouped by taxclass
23	r5_taxclass	Grouped average of r5 when grouped by taxclass

24	r6_taxclass	Grouped average of r6 when grouped by taxclass
25	r7_taxclass	Grouped average of r7 when grouped by taxclass
26	r8_taxclass	Grouped average of r8 when grouped by taxclass
27	r9_taxclass	Grouped average of r9 when grouped by taxclass
28	value_ratio	FULLVAL / (AVLAND + AVTOT)
29	size_ratio	BLDDEPTH / (LTFONT * LTDEPTH)

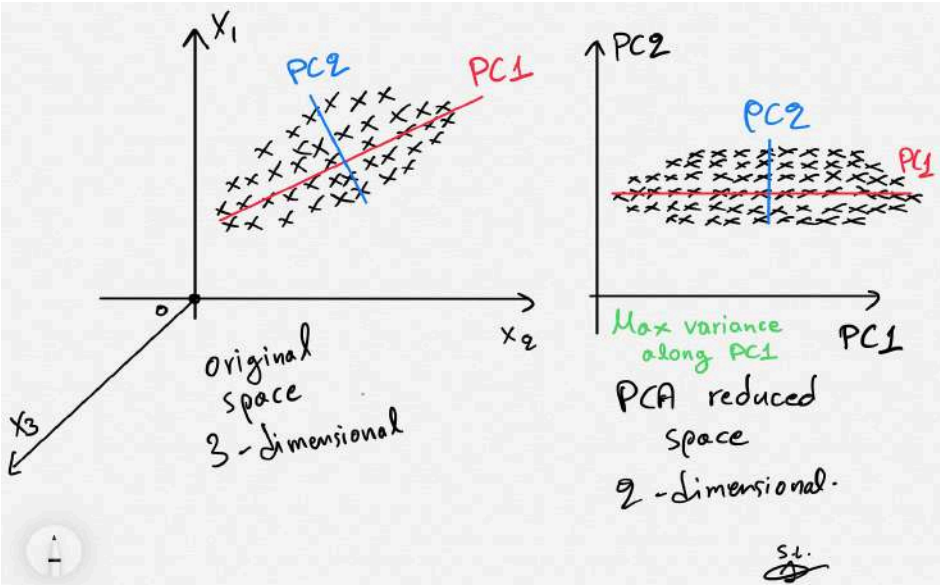
5. Dimensionality Reduction

Dimensionality reduction was performed using Principal Component Analysis (PCA) to simplify the dataset while preserving essential characteristics. The steps involved were:

1. **Standardization:**
   All numerical features in the dataset were standardized to have a mean of 0 and a standard deviation of 1. This step is crucial because PCA is sensitive to the variances of the initial variables. Standardizing ensures that each feature contributes equally to the analysis.

2. **PCA Application:**
   PCA was applied to the standardized data. The goal was to reduce the dimensionality while retaining as much variance as possible. The number of principal components was chosen based on the cumulative explained variance.

   Initially, the explained variance ratio was examined to determine the number of components that capture a significant portion of the total variance. It was found that retaining 10 principal components explained approximately 90% of the variance.



3. **Transformation:**
   The original dataset was transformed into a new set of dimensions represented by the principal components. This reduced feature set captures the majority of the variability in the data with fewer dimensions.

4. **Explained Variance:**
   The cumulative explained variance plot was analyzed to ensure that the chosen number of components was appropriate. This plot helps in visualizing how much of the total variance is captured by each principal component.

   A new DataFrame was created to store these principal components for subsequent analysis.

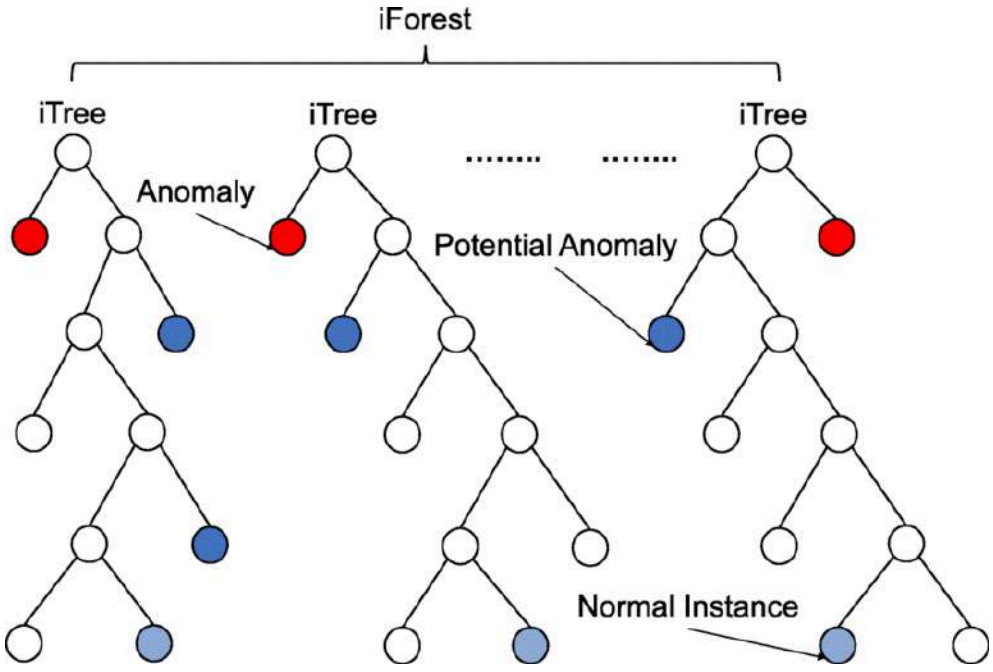
By reducing the dimensionality of the dataset, the computational complexity was decreased, and the efficiency of the anomaly detection algorithms was improved. This step was crucial for handling the high-dimensional data effectively and ensuring that the subsequent analysis was both robust and interpretable.

6. Anomaly Detection Algorithms

Two anomaly detection algorithms were employed to identify potentially fraudulent property records: Isolation Forest and Local Outlier Factor (LOF). These methods were chosen for their effectiveness in detecting anomalies in high-dimensional datasets.

Isolation Forest:

The Isolation Forest algorithm isolates observations by randomly selecting a feature and then a split value between the maximum and minimum values of the selected feature. The idea is that outliers are few and different, making them easier to isolate.
The algorithm creates an ensemble of isolation trees (iTrees) for the data and calculates the anomaly score for each data point based on the path length to isolate the point.



Working of Isolation Forest

Implementation:

The Isolation Forest model was trained on the principal components obtained from the PCA. The contamination parameter, which specifies the expected proportion of outliers in the dataset, was set to 10%.

Anomaly scores were computed for each record, indicating the likelihood of being an outlier. Records with higher anomaly scores are more likely to be fraudulent.

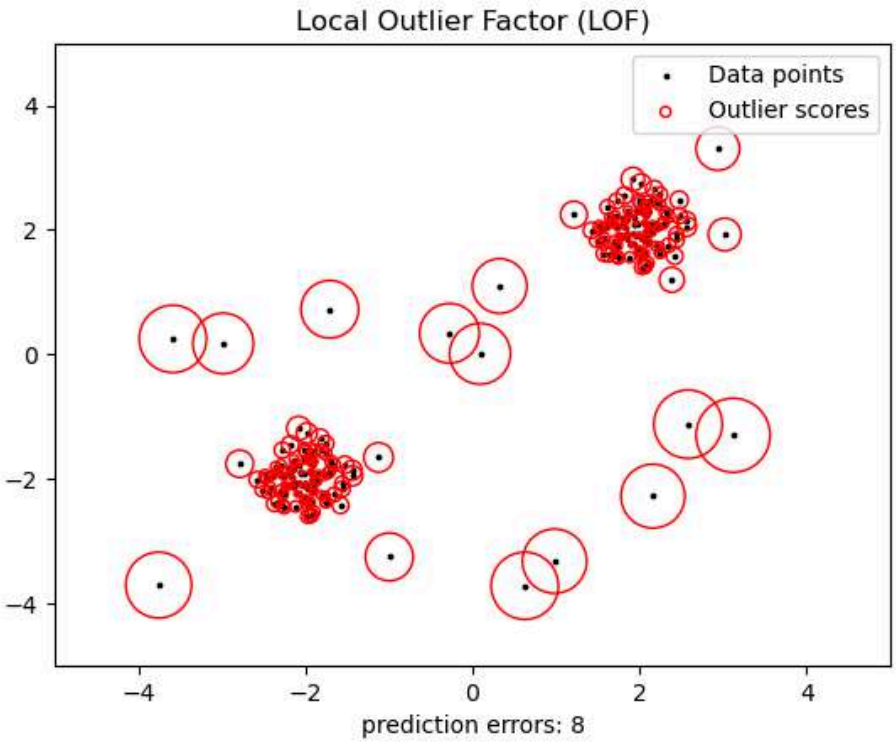
Anomaly Scores:

The decision function gives the anomaly score, with negative values indicating anomalies. The predict method labels the points as -1 for anomalies and 1 for normal points.

Local Outlier Factor (LOF)

The LOF algorithm measures the local density deviation of a given data point with respect to its neighbors. Points with significantly lower density compared to their neighbors are considered outliers.

LOF calculates the Local Reachability Density (LRD) of each data point and compares it with the LRD of its neighbors to compute the LOF score.



Working of Local Outlier Factor

Implementation:

The LOF algorithm was applied to the principal components with the contamination parameter set to 10%. The number of neighbors parameter was set to 20, which specifies how many neighbors to consider for computing the local density.

The fit_predict method computes the outlier scores and returns labels for each data point, where -1 indicates an anomaly.

Outlier Scores:
The LOF scores are negative for anomalies. Similar to Isolation Forest, the predict method labels the points, with -1 indicating anomalies and 1 indicating normal points.

Additionally,
To enhance the accuracy and robustness of anomaly detection, we could combine the results of the Isolation Forest and Local Outlier Factor (LOF) algorithms. This would possibly leverage the strengths of both methods.

We can achieve this by:

1. Standardization of Scores:
   The anomaly scores from both algorithms could be standardized to a common scale to ensure comparability.We may consider using z-score normalization for this.

2. Combining the Scores:
   After standardizing the scores, they can be combined using a simple average. Alternatively, weighted averaging can be used if the contribution of each algorithm needs to be adjusted based on their respective strengths.

3. Thresholding and Labeling:
   A threshold may also beapplied to the combined scores to label data points as anomalies or normal points. This threshold can be tuned based on the desired sensitivity and specificity of the anomaly detection.

By scaling and combining the anomaly scores from Isolation Forest and LOF, we can enhance the robustness and accuracy of the anomaly detection process. This dual-algorithm approach ensures that both global and local anomalies are effectively detected, providing a comprehensive analysis of potential tax fraud in New York City's property data.

7. Results

The anomaly detection results were analyzed to identify the top properties with the highest fraud scores. The process involved:

Scoring: Each property received a combined score based on the Isolation Forest and LOF results.
Top Properties: The top properties with the highest scores were examined in detail, cross-referencing the detected anomalies with known property characteristics and public records.

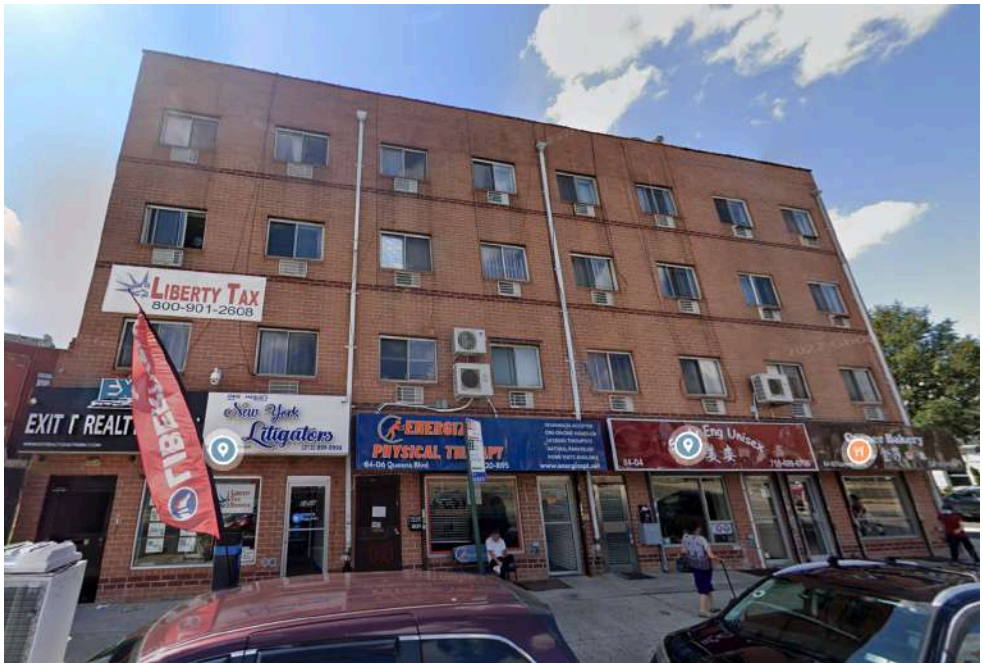
We format the results excel file to better identify the anomalies as follows:

Next, I looked at records with their ‘r’ values to be greater than 10 standard deviations, as this indicates anomalies. Having noted down which r values are higher for the given record, I looked at the corresponding fields and checked for either incorrect entry or absence of data, cross checking with publicly available satellite images of the properties using their street address and zip code.

Five interesting records were selected to illustrate the types of anomalies detected. These included properties with inconsistent valuation metrics, unusual size and value ratios, and discrepancies in reported characteristics.

1. **Record 970**

Owner	Jia Ping Li	**LTFRONT**	20
Address	84-06 QUEENS BOULEVARD, 11373	**LTDEPTH**	53
FULLVAL	773,000	**BLDFRONT**	0
AVLAND	191,250	**BLDDEPTH**	0
AVTOT	347,850	**STORIES**	4



This property looks like it is an Apartment Complex, not a private home belonging to an individual, as the Owner name suggests. Next, it is classified into tax class ‘S3’ which means Individuals that are 62 Years of Age and Older with Specific Income Limitations. Moreover, the LTFRONT and LTDEPTH values of 20 and 53 respectively seem to be too small for the entire complex. The FULLVAL, AVLAND and AVTOT values also do not seem to add up. The tax class values show higher than usual deviations which makes sense as the tax class into which ot is classified shows income limitations, but the AVLAND and AVTOT values do not comply with these findings.

2. Record 877

Owner	Teofil Budanik	**LTFRONT**	93
Address	28 BROOME STREET, 11222	**LTDEPTH**	42
FULLVAL	821,000	**BLDFRONT**	0
AVLAND	173,799	**BLDDEPTH**	0
AVTOT	315,684	**STORIES**	-

This property too looks like it is an Apartment Complex, not a private home belonging to an individual, as the Owner name suggests. Moreover, the LTFRONT and LTDEPTH values of 93 and 42 respectively seem to be too small for the entire complex. The BLDFRONT and BLDDEPTH values are mentioned as 0, which is again wrong, and the number of stories is not specified for this record, but it is clearly a 4 storied complex.

3. Record 590

Owner	Mohammed Haque	**LTFRONT**	35
Address	11-04 34 AVENUE, 11106	**LTDEPTH**	59
FULLVAL	669,000	**BLDFRONT**	0
AVLAND	6,420	**BLDDEPTH**	0
AVTOT	27,360	**STORIES**	4

This property also looks like it is an Apartment Complex, not a private home belonging to an individual, as the Owner name suggests. The number of stories in the data mentions that it is 4 stories, but we can observe that it is 3 stories. The AVLAND and AVTOT values are particularly low for a complex. The r4_taxclass, r5_taxclass, r8_taxclass and r9_taxclass values are very high, which makes sense as they are related to AVLAND and AVTOT. This is why this record might have a high fraud score.

4. Record 530

Owner	Edward Buchholz	**LTFRONT**	36
Address	38-11 LITTLE NECK PARKWAY, 11363	**LTDEPTH**	203
FULLVAL	604,000	**BLDFRONT**	0
AVLAND	9,022	**BLDDEPTH**	0
AVTOT	10,812	**STORIES**	1

This property looks like a private home belonging to an individual, but is wrongly classified as having 1 storey only. It clearly has 2 stories. Moreover, the LTFRONT value of 36 seems to be too small for the entire lot. The BLDGFRONT and BLDGDEPTH are wrongly mentioned as 0, which is why the r2_taxclass, r3_taxclass, r5_taxclass, r6_taxclass, r8_taxclass, and r9_taxclass are unusually high, giving this record a potentially high fraud score.

5. Record 266

Owner	Billy Williams	**LTFRONT**	136
Address	380 BEACH 86 STREET, 11693	**LTDEPTH**	118
FULLVAL	613000	**BLDFRONT**	0
AVLAND	16381	**BLDDEPTH**	0
AVTOT	19090	**STORIES**	-

This property looks like a private home belonging to an individual, as the Owner name suggests. The LTFRONT and LTDEPTH values are too high and do not seem to match the streetview image of the house. The r2_taxclass, r3_taxclass, r5_taxclass, r6_taxclass, r8_taxclass, and r9_taxclass are unusually high, probably due to the low AVLAND and AVTOT values, giving this record a potentially high fraud score.

7. Summary

Data Description:
I began by examining the Property Valuation and Assessment Data, which contains 1,070,994 records and 32 fields. The dataset includes various property characteristics and valuations for the year 2010/11, covering boroughs such as Manhattan, Bronx, Brooklyn, Queens, and Staten Island.

Data Cleaning:
Data cleaning involved handling missing and zero values to ensure consistency. Properties owned by government entities were removed, as they are irrelevant for tax fraud analysis. Missing ZIP codes were imputed using a combination of street addresses and adjacent records. Missing or zero values in fields like FULLVAL, AVLAND, and AVTOT were filled using group averages based on TAXCLASS, BORO, and BLDGCL. Similar imputation strategies were applied to other fields such as STORIES, LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH.

Variable Creation:
New variables were created to identify unusual characteristics in properties. These included size variables (lot area, building area, building volume) and value ratios (ratios of FULLVAL, AVLAND, and AVTOT to the size variables). Inverse value ratios were also calculated to detect unusually small values. Grouped averages and standardized variables were derived based on ZIP and TAXCLASS to enhance anomaly detection.

Dimensionality Reduction:
Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while retaining essential characteristics. Numerical features were standardized, and PCA was used to transform the data into a new set of dimensions represented by 10 principal components, capturing approximately 90% of the variance. This reduction improved the efficiency of the anomaly detection algorithms.

Anomaly Detection Algorithms:
Two anomaly detection algorithms, Isolation Forest and Local Outlier Factor (LOF), were employed. Isolation Forest isolates observations by randomly selecting features and split values, creating an ensemble of isolation trees. LOF measures the local density deviation of data points with respect to their neighbors. Anomaly scores from both algorithms were standardized and combined to enhance detection accuracy.

Results:
The combined anomaly scores were used to identify properties with the highest fraud scores. Detailed case studies of five interesting records highlighted significant irregularities, such as inconsistent valuation metrics, unusual size and value ratios, and discrepancies in reported characteristics.

1. High Fraud Scores: Several properties exhibited high fraud scores based on the combined results of Isolation Forest and LOF algorithms.
2. Inconsistent Valuation Metrics: Properties were found with full market values, assessed land values, and total assessed values that did not align with their physical characteristics or reported usage.
3. Unusual Size and Value Ratios: The analysis revealed anomalies in size and value ratios, indicating discrepancies between reported property dimensions and their valuations.
4. Significant Irregularities: Detailed case studies of the top flagged properties highlighted substantial irregularities that warrant further investigation by city authorities.

Improvement Based on Client Feedback-

1. Refining Anomaly Detection Thresholds:
   We can adjust the contamination parameter and threshold values for both Isolation Forest and LOF based on feedback from domain experts to better capture genuine anomalies without flagging too many false positives.

2. Incorporating Additional Data Sources:
   We could also enhance the dataset by incorporating additional data sources such as recent property sales records, construction permits, and historical property tax records to provide a more comprehensive view and improve anomaly detection accuracy.

3. Improving Variable Creation:
   Creation of more nuanced variables by considering additional factors like property age, recent renovations, and neighborhood crime rates can help in identifying more subtle anomalies related to property valuations.

4. Ensemble Learning:
   Exploring advanced ensemble learning techniques (such as stacking, bagging, or boosting) to combine the strengths of multiple anomaly detection algorithms and further improve the robustness of the detection system.

5. Continuous Feedback Loop:
   We can establish a continuous feedback loop with city authorities and domain experts to iteratively refine the model. Regular updates based on feedback and new data will ensure the model remains accurate and relevant.

6. User-Friendly Interface:
   Lastly, we can also develop a user-friendly interface for city authorities to interact with the anomaly detection system. Visualizations and dashboards can help in better understanding the flagged properties and making informed decisions.

7. Appendix (DQR)

Data Quality Report

1. Data Description

The Dataset is **Property Valuation and Assessment Data,** covering various attributes of properties in the dataset (). It includes a total of **1,070,994 records** and **32 fields**. It covers property assessments for the **year 2010/11**. This dataset provides insights into property characteristics and valuations, which could be used for real estate analysis, tax assessments, and urban planning. The dataset includes specific code definitions and ranges for certain fields such as boroughs (Manhattan, Bronx, Brooklyn, Queens, Staten Island), building classes (1-1 to 1-3 unit residence, apartments, utilities, all others), and easement types (Non-Air Rights, Transit Easement, etc.).

2. Summary Tables

Numeric Fields Table :

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
LTFRONT	numeric	1070994	100.0%	169108	0	9999	36.63530141	74.03	0
LTDEPTH	numeric	1070994	100.0%	170128	0	9999	88.861594	76.4	100
STORIES	numeric	1014730	94.7%	0	1	119	5.006917899	8.37	2
FULLVAL	numeric	1070994	100.0%	13007	0	6150000000	874264.5054	11582425.58	0
AVLAND	numeric	1070994	100.0%	13009	0	2668500000	85067.91867	4057258.16	0
AVTOT	numeric	1070994	100.0%	13007	0	4668308947	227238.1687	6877526.09	0
EXLAND	numeric	1070994	100.0%	491699	0	2668500000	36423.89069	3981573.93	0
EXTOT	numeric	1070994	100.0%	432572	0	4668308947	91186.98168	6508399.78	0

BLDFRONT	numeric	1070994	100.0%	228815	0	7575	23.04276961	35.58	0
BLDDEPTH	numeric	1070994	100.0%	228853	0	9393	39.92283617	42.71	0
AVLAND2	numeric	282726	26.4%	0	3	2371005000	246235.7193	6178951.64	2408
AVTOT2	numeric	282732	26.4%	0	3	4501180002	713911.4362	11652508.34	750
EXLAND2	numeric	87449	8.2%	0	1	2371005000	351235.6843	10802150.91	2090
EXTOT2	numeric	130828	12.2%	0	7	4501180002	656768.2819	16072448.75	2090

Categorical Fields Table :

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
RECORD	categorical	1070994	100.0%	0	1070994	1
BBLE	categorical	1070994	100.0%	0	1070994	1000010101
BORO	categorical	1070994	100.0%	0	5	4
BLOCK	categorical	1070994	100.0%	0	13984	3944
LOT	categorical	1070994	100.0%	0	6366	1
EASEMENT	categorical	4636	0.4%	0	12	E
OWNER	categorical	1039249	97.0%	0	863347	PARKCHESTER PRESERVAT
BLDGCL	categorical	1070994	100.0%	0	200	R4
TAXCLASS	categorical	1070994	100.0%	0	11	1
EXT	categorical	354305	33.1%	0	3	G
EXCD1	categorical	638488	59.6%	0	129	1017
STADDR	categorical	1070318	99.9%	0	839280	501 SURF AVENUE
ZIP	categorical	1041104	97.2%	0	196	10314
EXMPTCL	categorical	15579	1.5%	0	14	X1
EXCD2	categorical	92948	8.7%	0	60	1017
PERIOD	categorical	1070994	100.0%	0	1	FINAL
YEAR	categorical	1070994	100.0%	0	1	2010/11
VALTYPE	categorical	1070994	100.0%	0	1	AC-TR

3. Visualization of Each Field

1. Field Name : Record

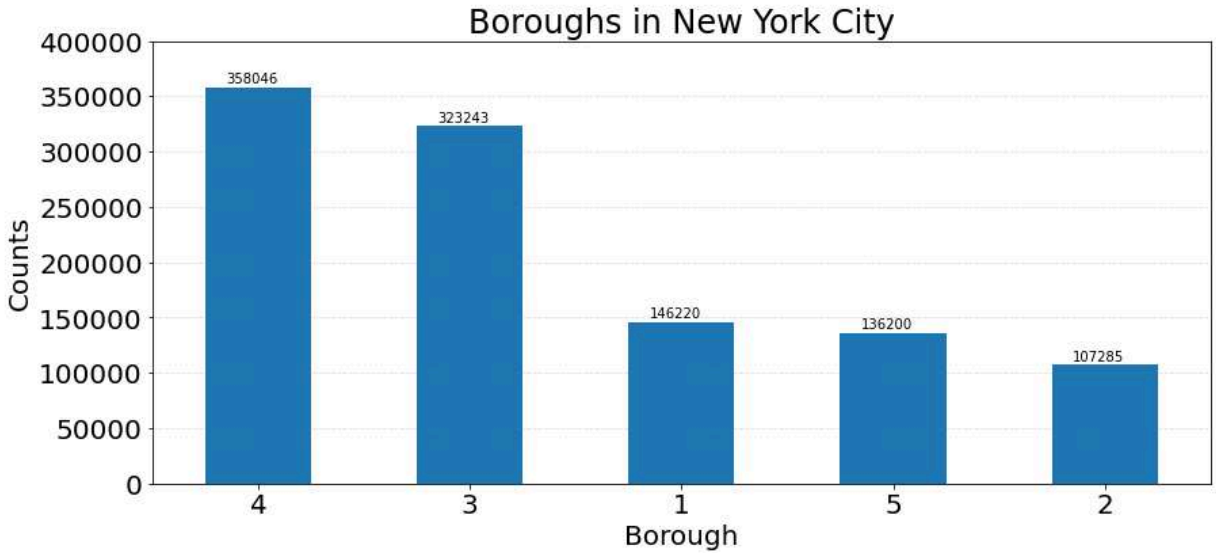
Description: Unique, positive record numbers from 1 to 1,070,994 depicting the unique properties in the city of New York. (It does not make sense to visualize all 97,852 record numbers.)

2. Field Name : BBLE

Description: Seems to be an abbreviation for Boro, Block, Lot and Easement. There are 1,070,994 unique BBLE values , which makes sense as no two properties can have the same BBLE combinations.

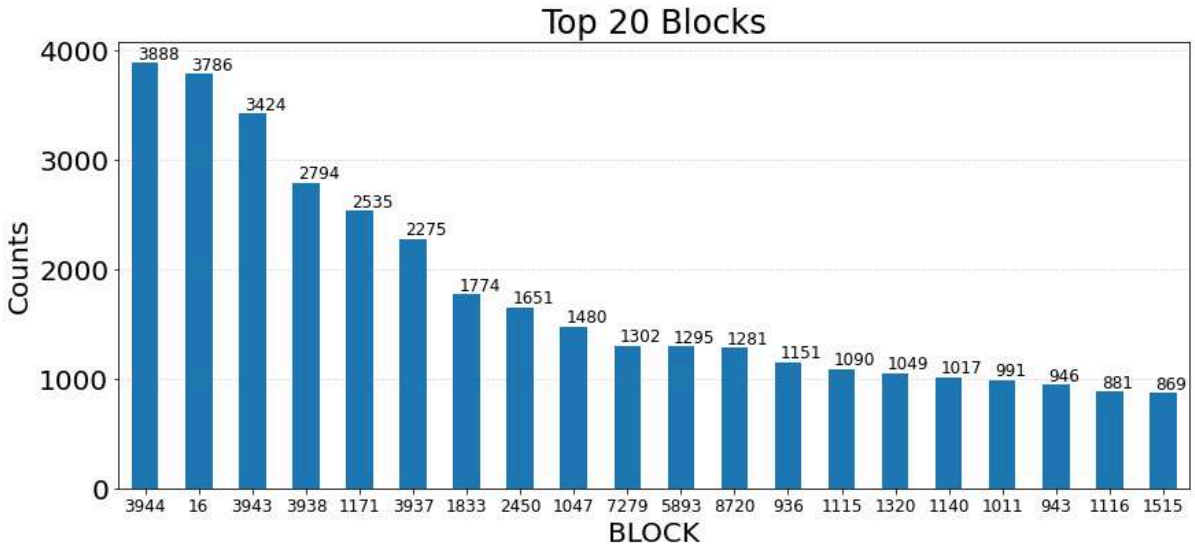
3. Field Name : BORO

Description: Code representing the borough where the property is located (1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island).



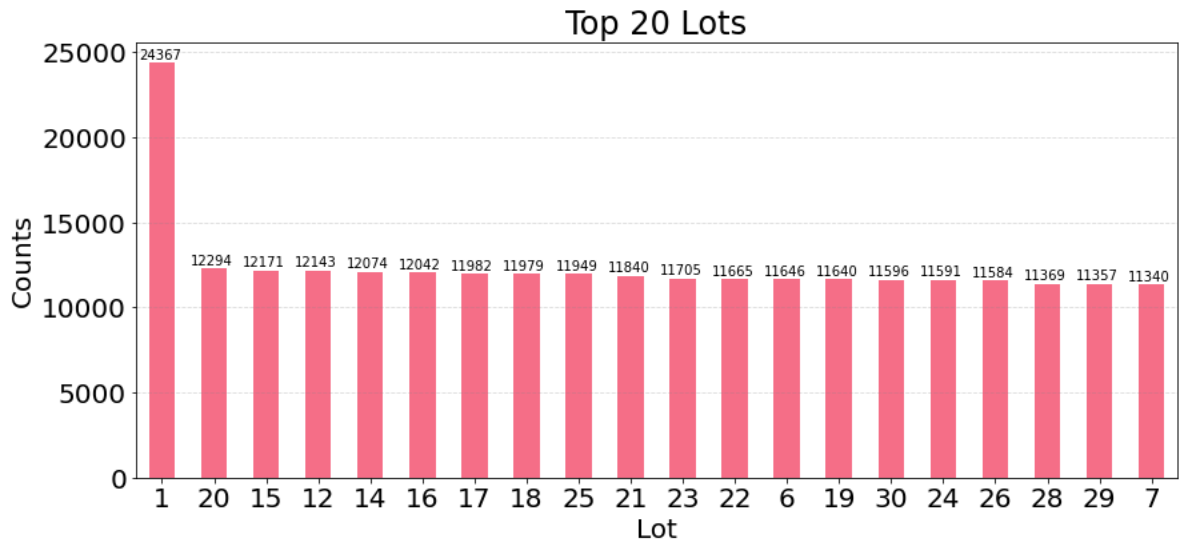
4. **Field Name : BLOCK**

Description: Block number within the borough, indicating a specific area.



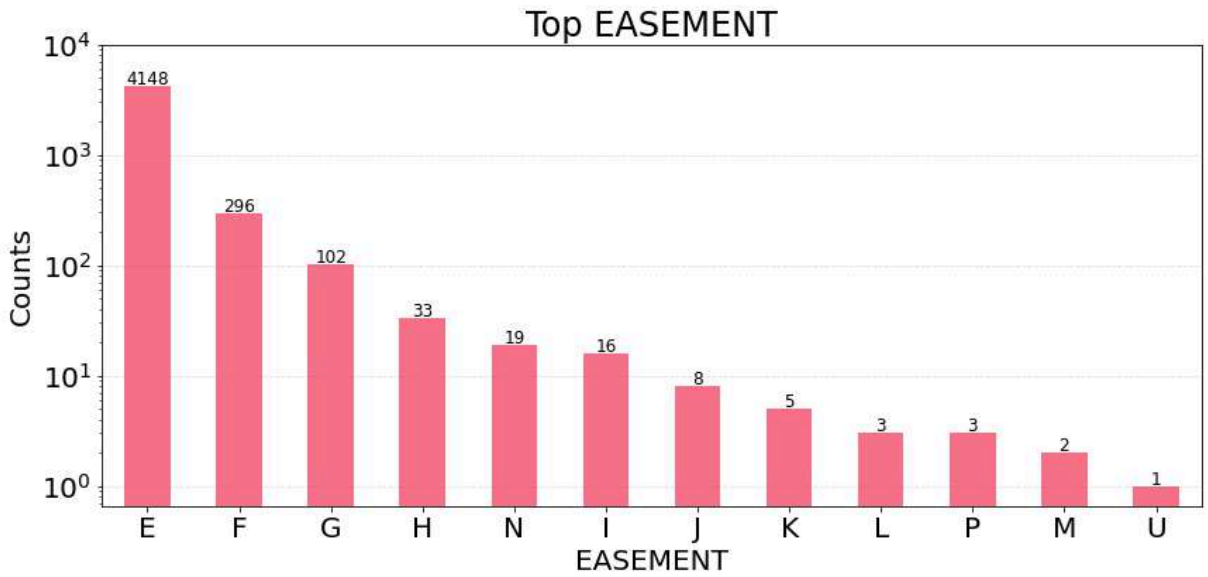
5. **Field Name : LOT**

Description: Lot number within the block, specifying a particular parcel of land.



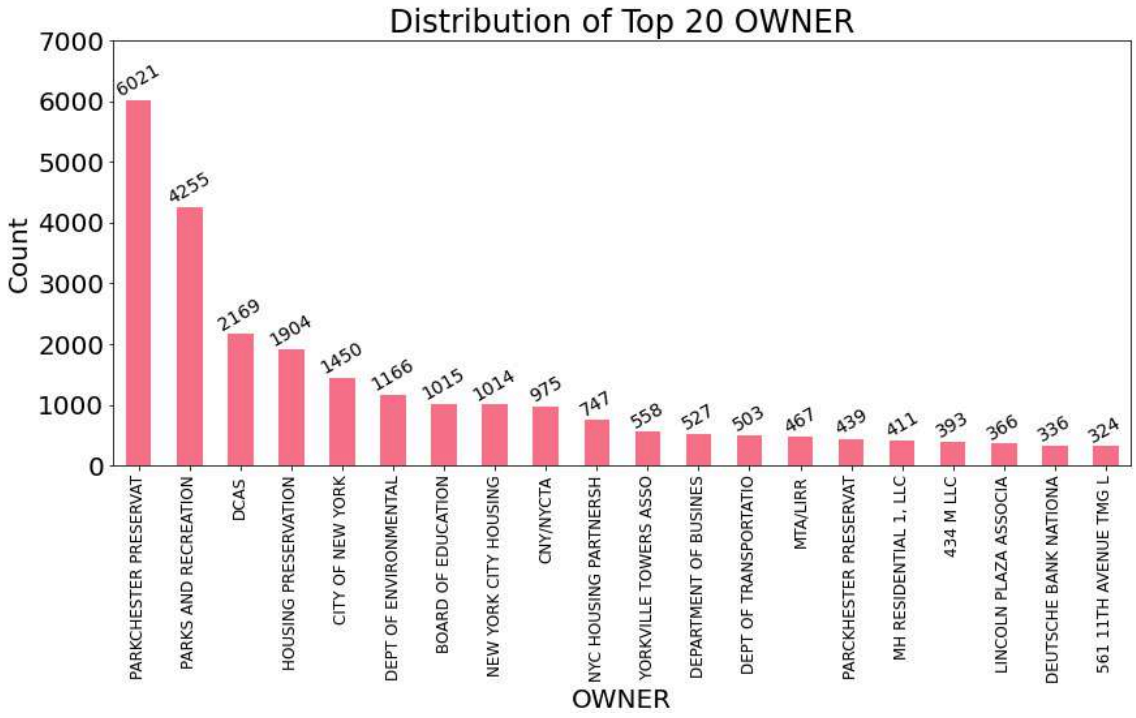
6. **Field Name : EASEMENT**

Description: Information about easements on the property (e.g., A = Air Easement, L = Land Easement). Only 43.2% of this field is populated.



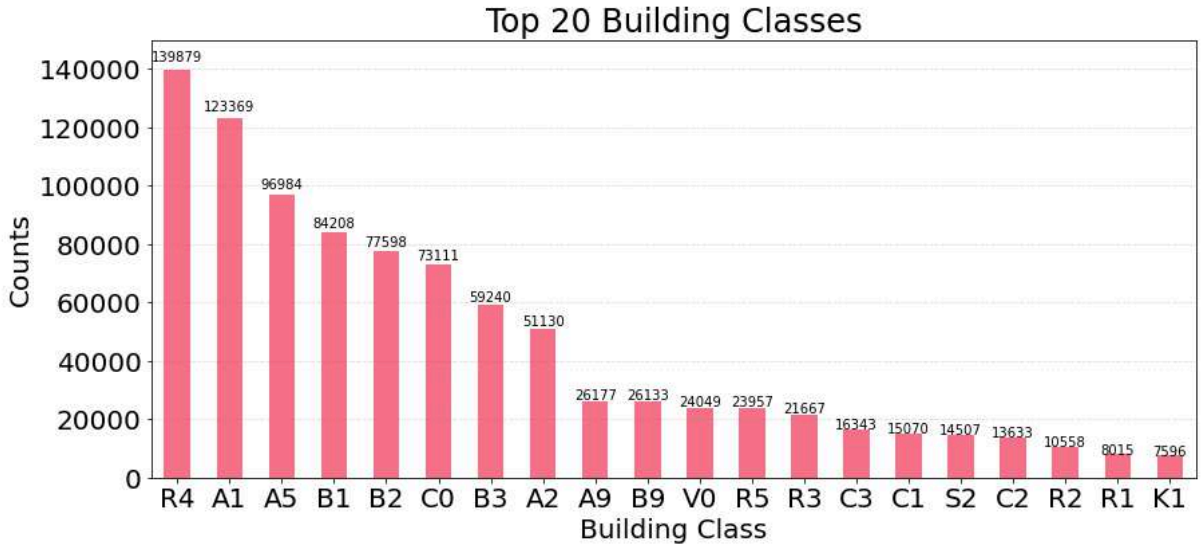
7. **Field Name : OWNER**

Description: Name of the property owner. This field is 97.03% populated



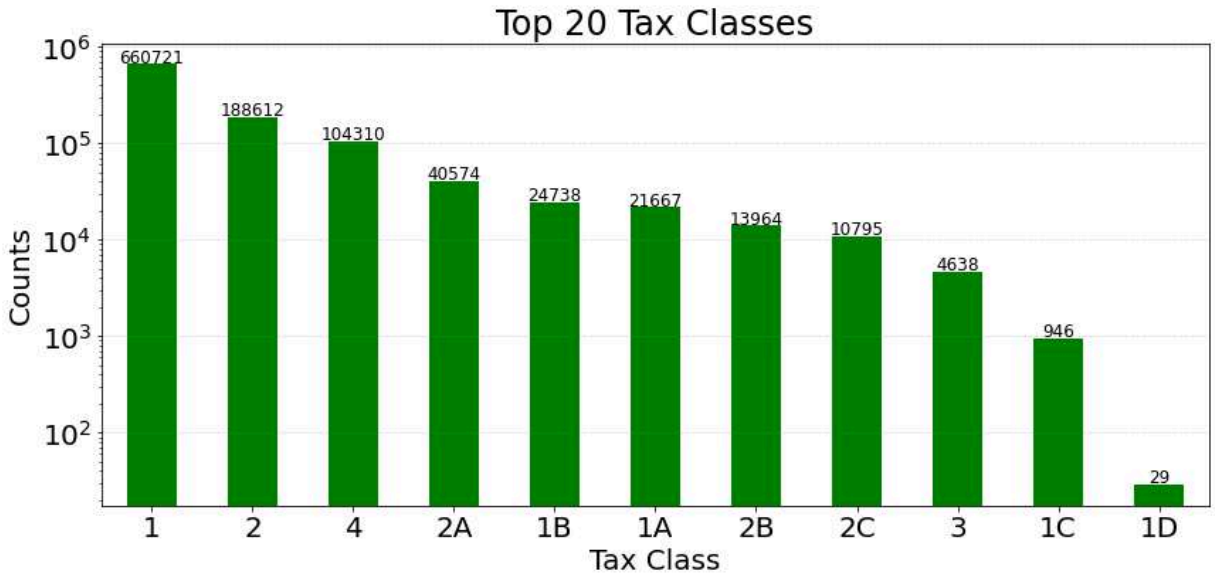
8. **Field Name : BLDGCL**

Description: Building class code, indicating the type of building (e.g., 1 = 1-3 Unit Residence, 2 = Apartments).



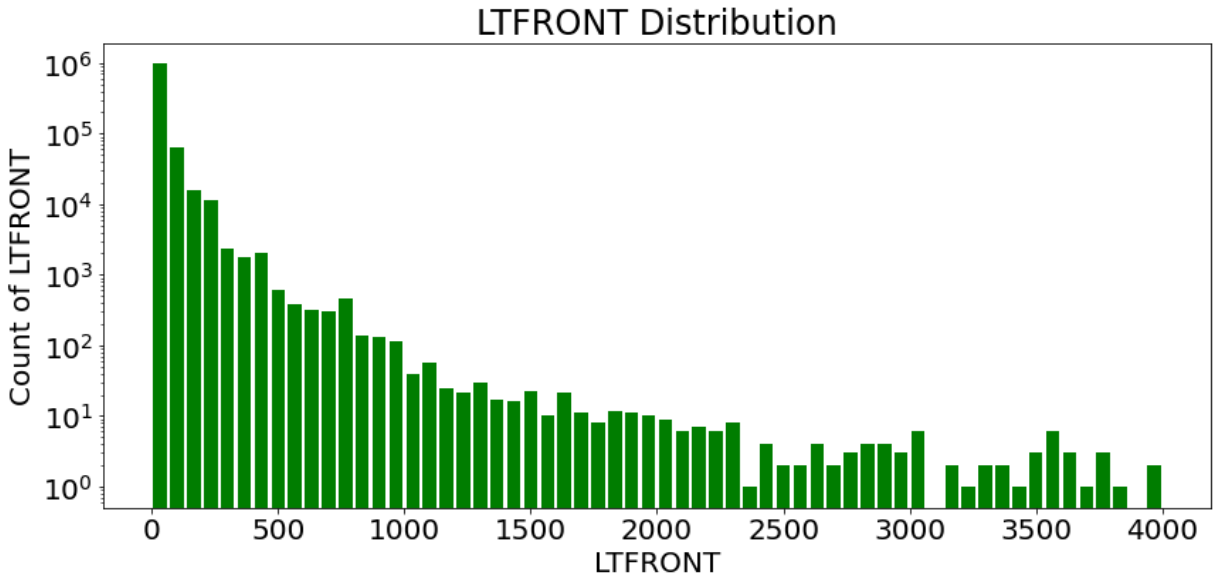
9. **Field Name : TAXCLASS**

Description: Tax class of the property, which affects its tax rate (e.g., 1 = 1-3 Unit Residence, 2 = Apartments).

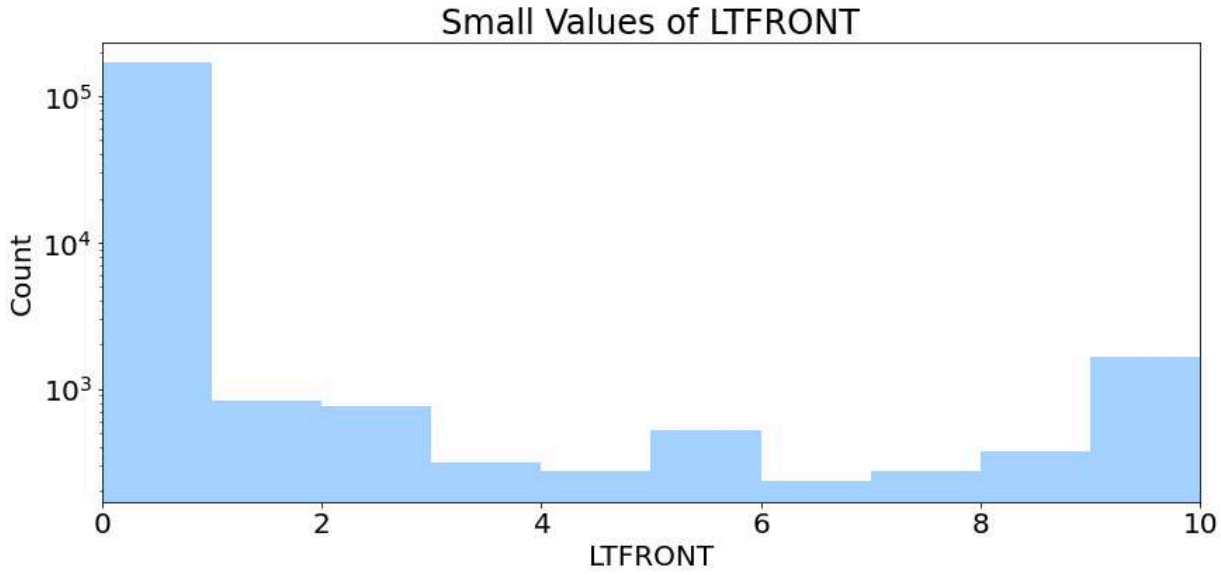


10. Field Name : LTFRONT

Description: Width of the lot's front in feet.

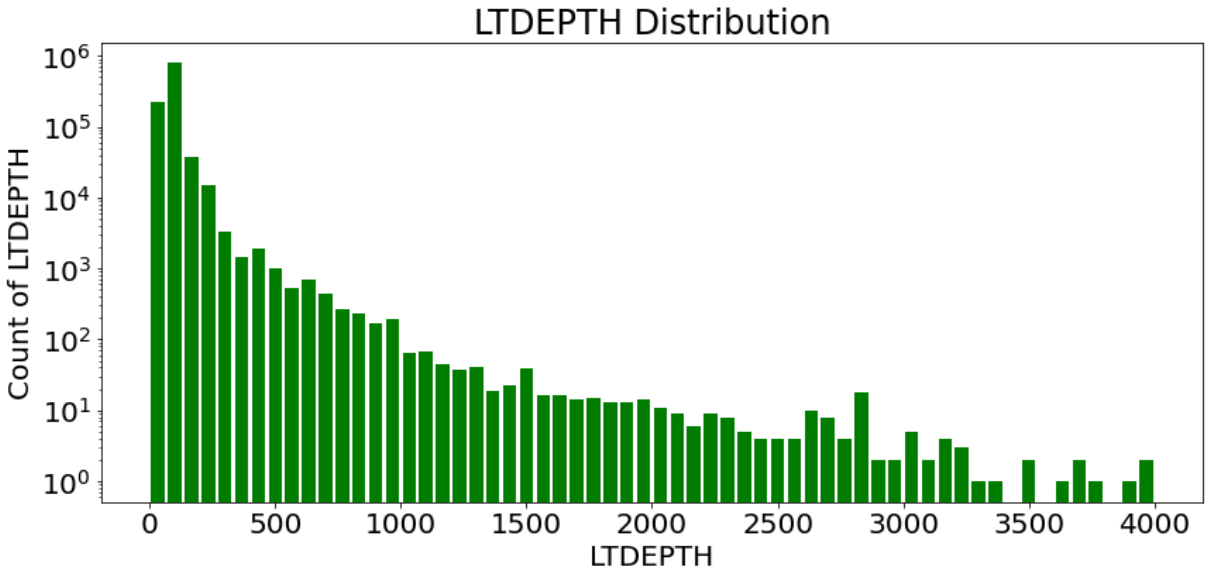


Additionally, there are numerous small values of LTFRONT, which are visualized here.

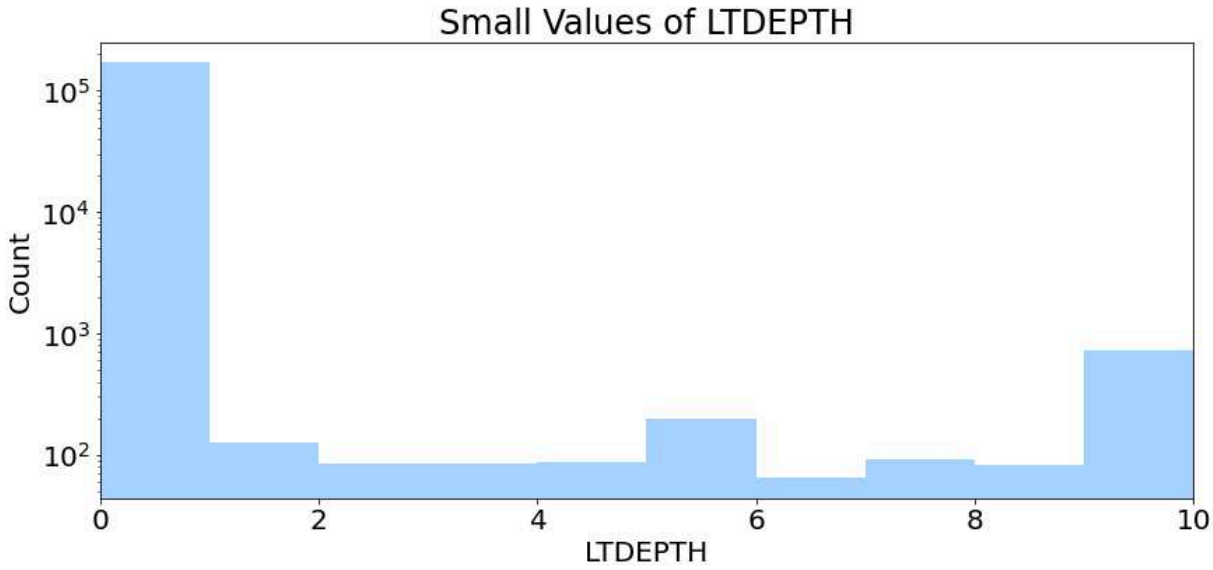


11. Field Name : LTDEPTH

Description: Depth of the lot in feet.

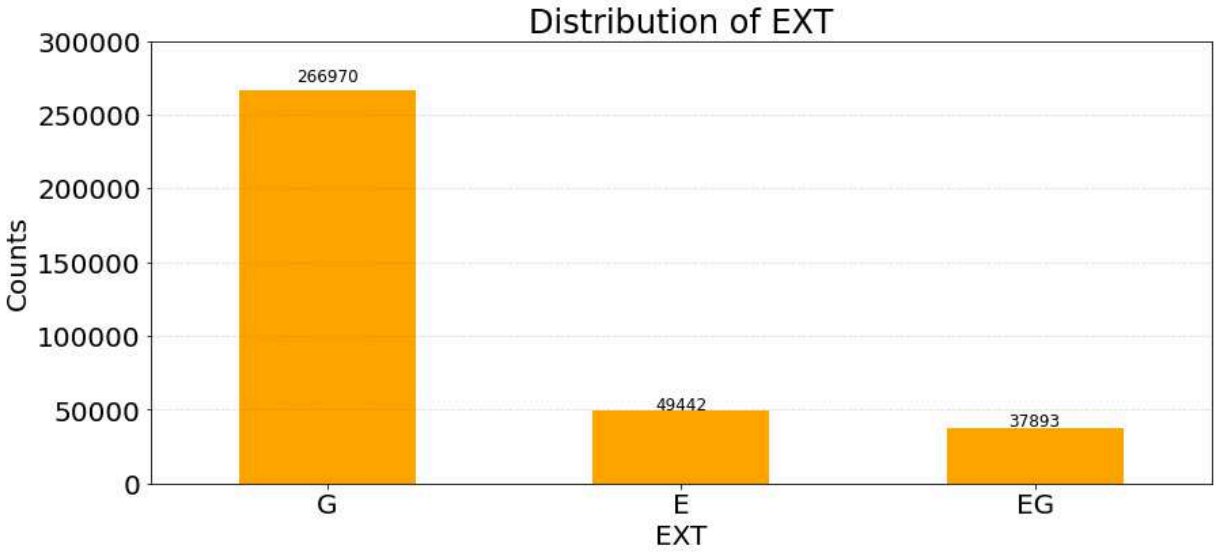


Additionally, there are numerous small values of LTDEPTH, which are visualized here.



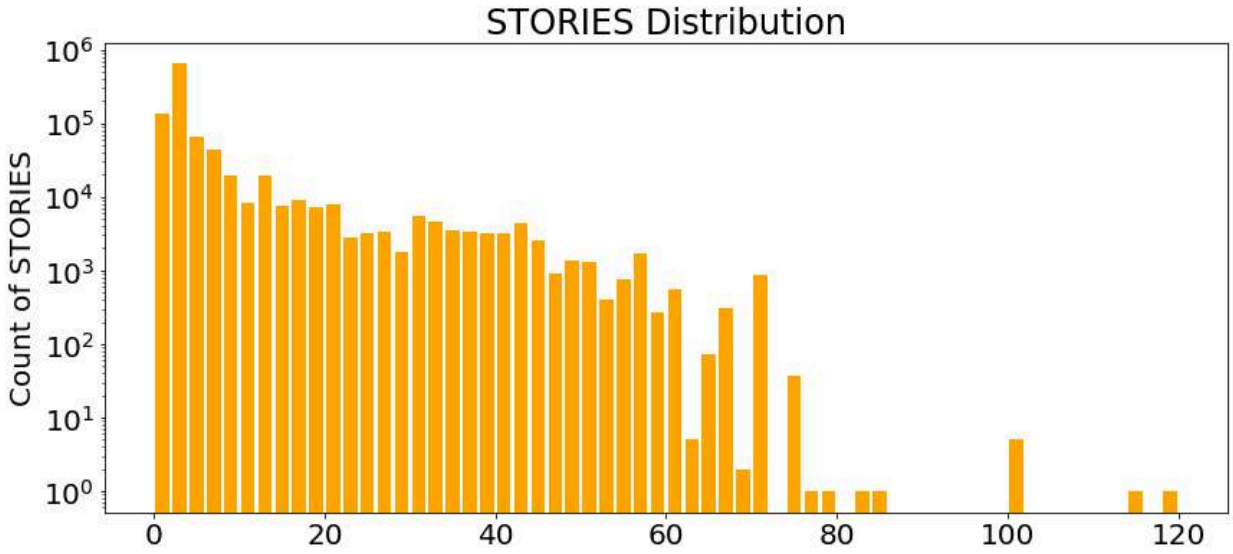
12. Field Name : EXT

Description: Extension identifier, providing additional detail about the property. Only 33.08% of this field is populated.



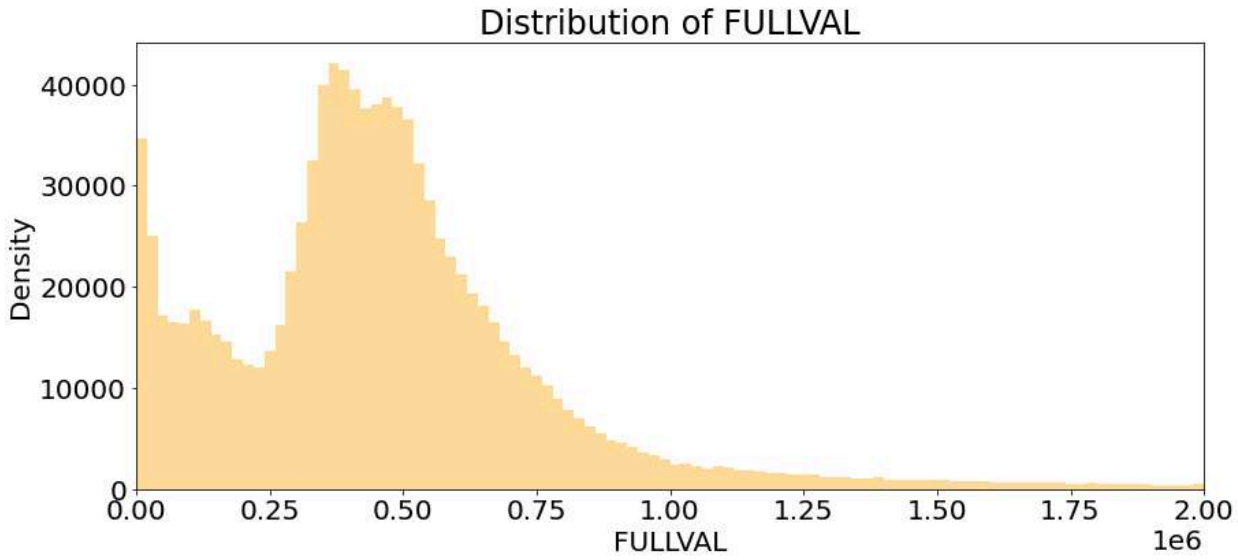
13. Field Name : STORIES

Description: Number of stories in the building. 99.4% of this field is populated.



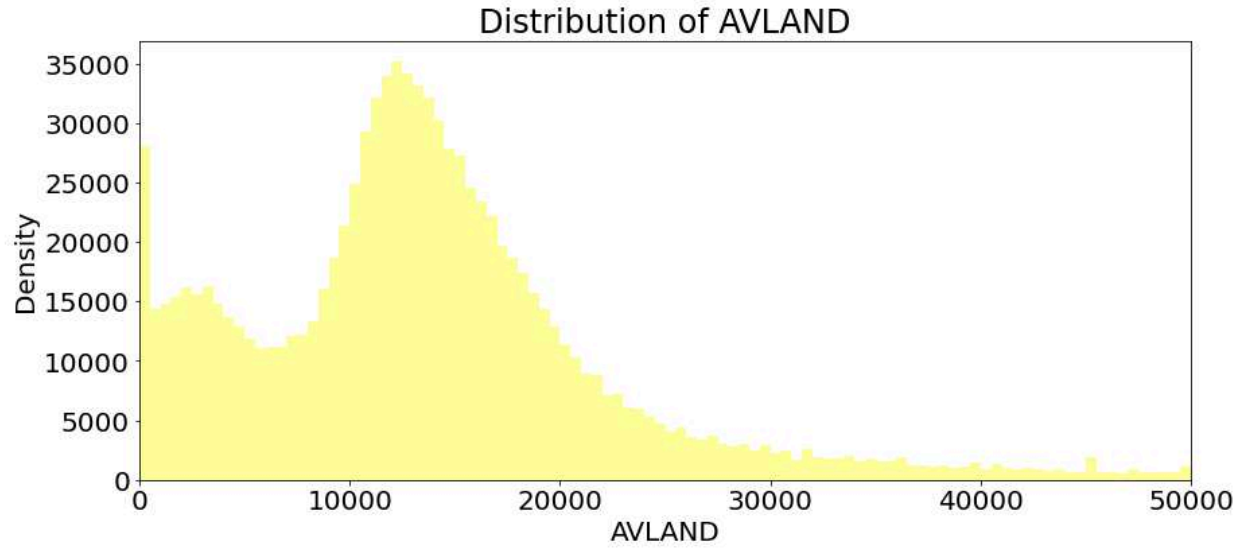
14. Field Name : FULLVAL

Description: Full market value of the property. Appears to have a bimodal distribution.



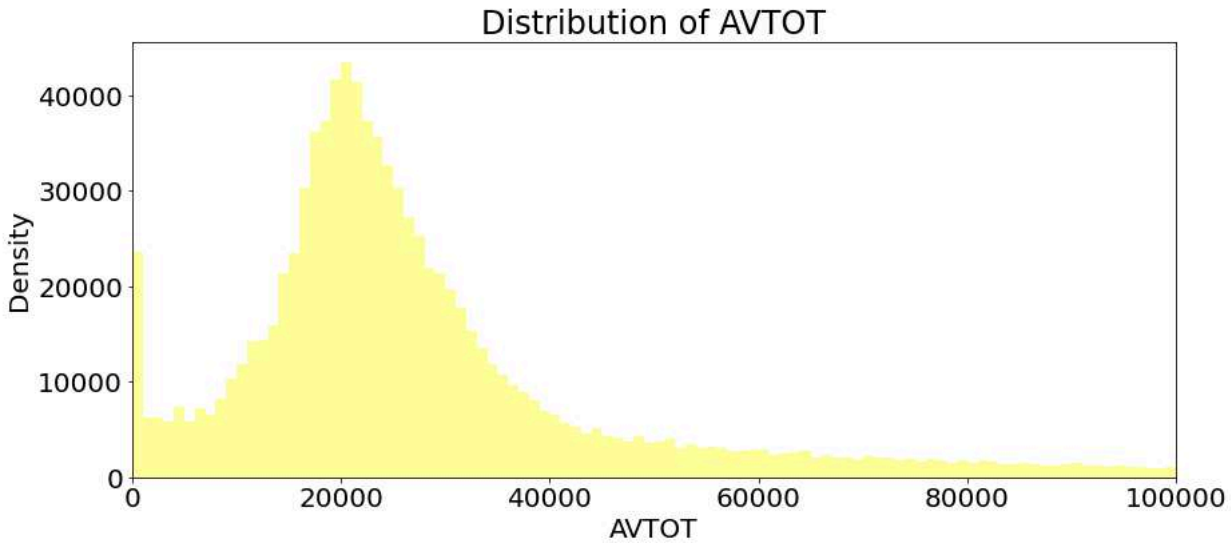
15. Field Name : AVLAND

Description: Assessed value of the land portion of the property.



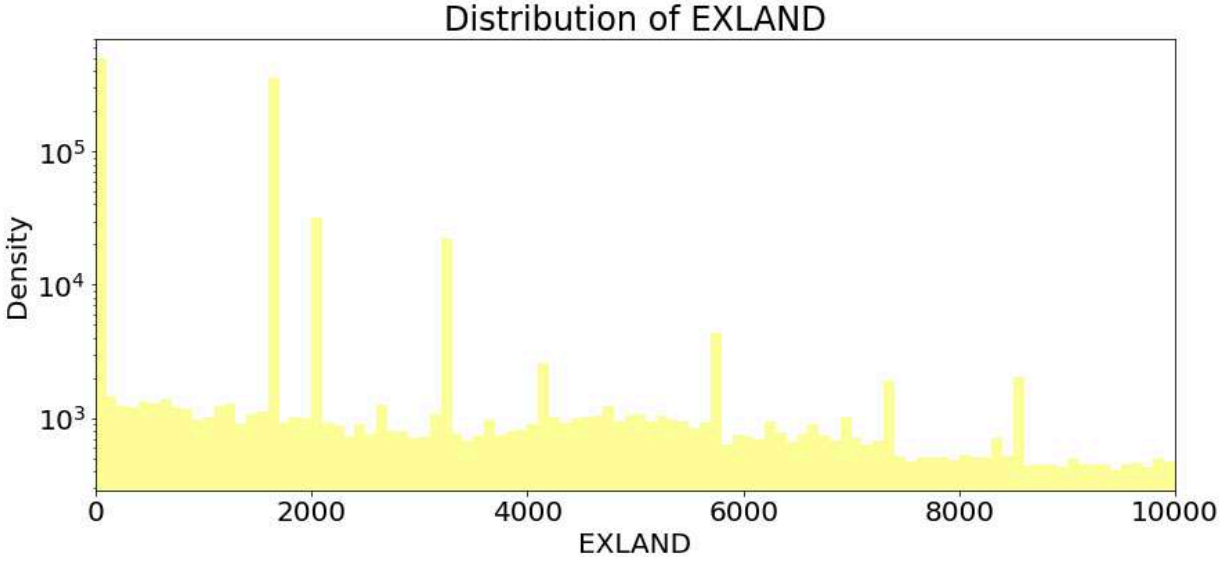
16. Field Name : AVTOT

Description: Total assessed value of the property, including land and improvements.



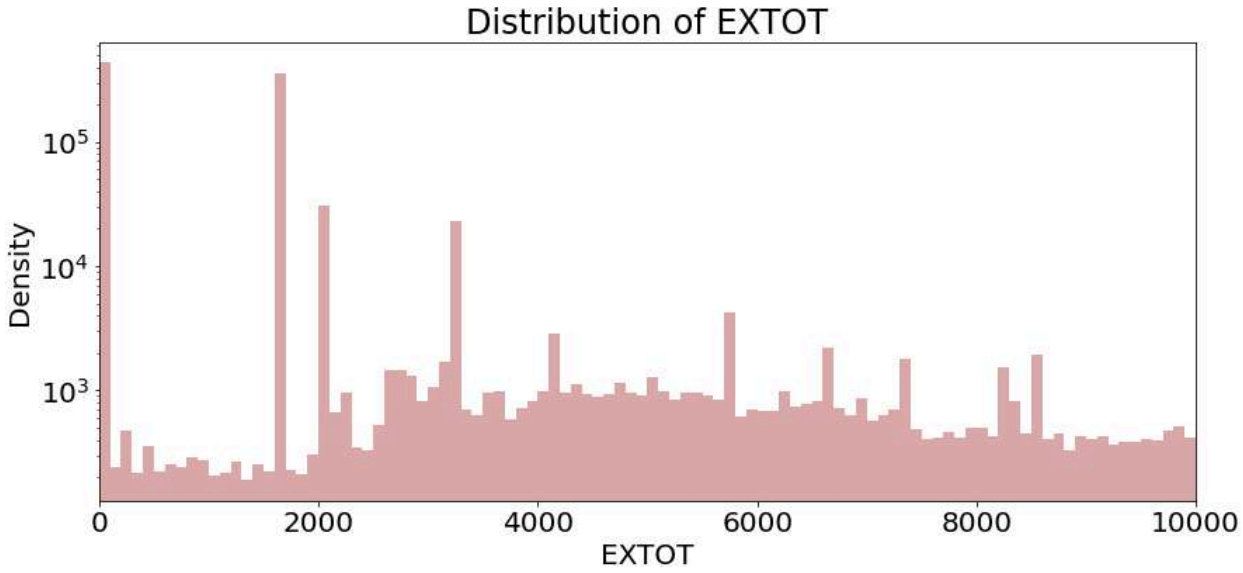
17. Field Name : EXLAND

Description: Exempt land value, indicating the portion of land value that is exempt from taxes.



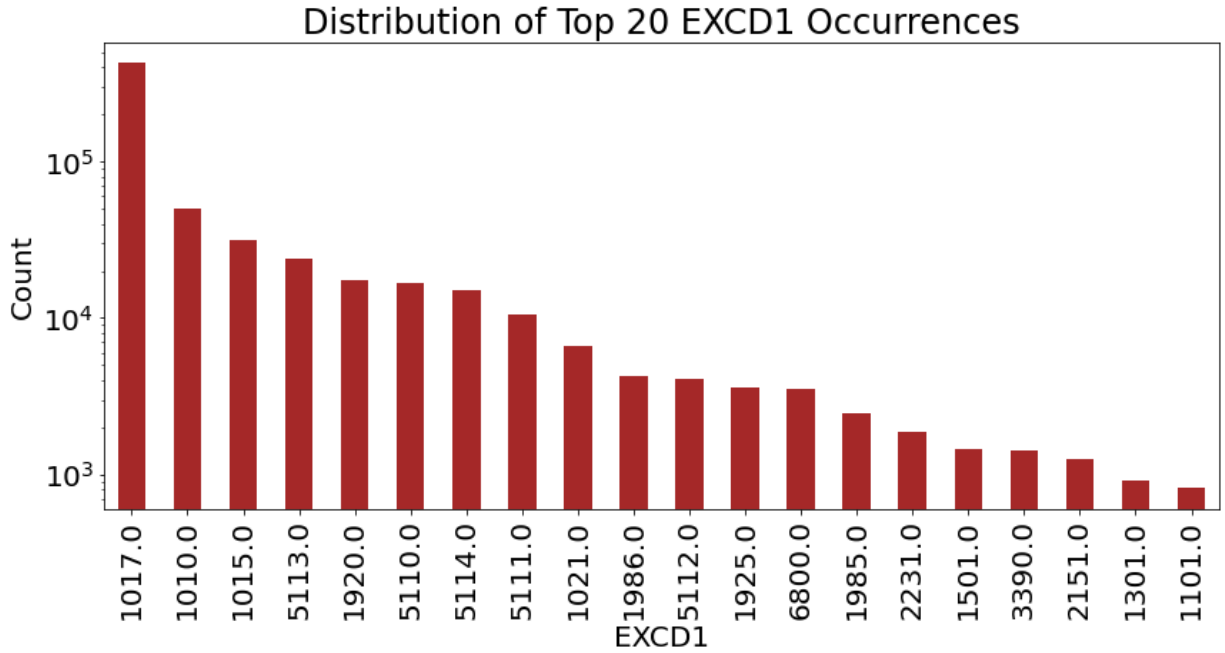
18. Field Name : EXTOT

Description: Total exempt value, including land and improvements.



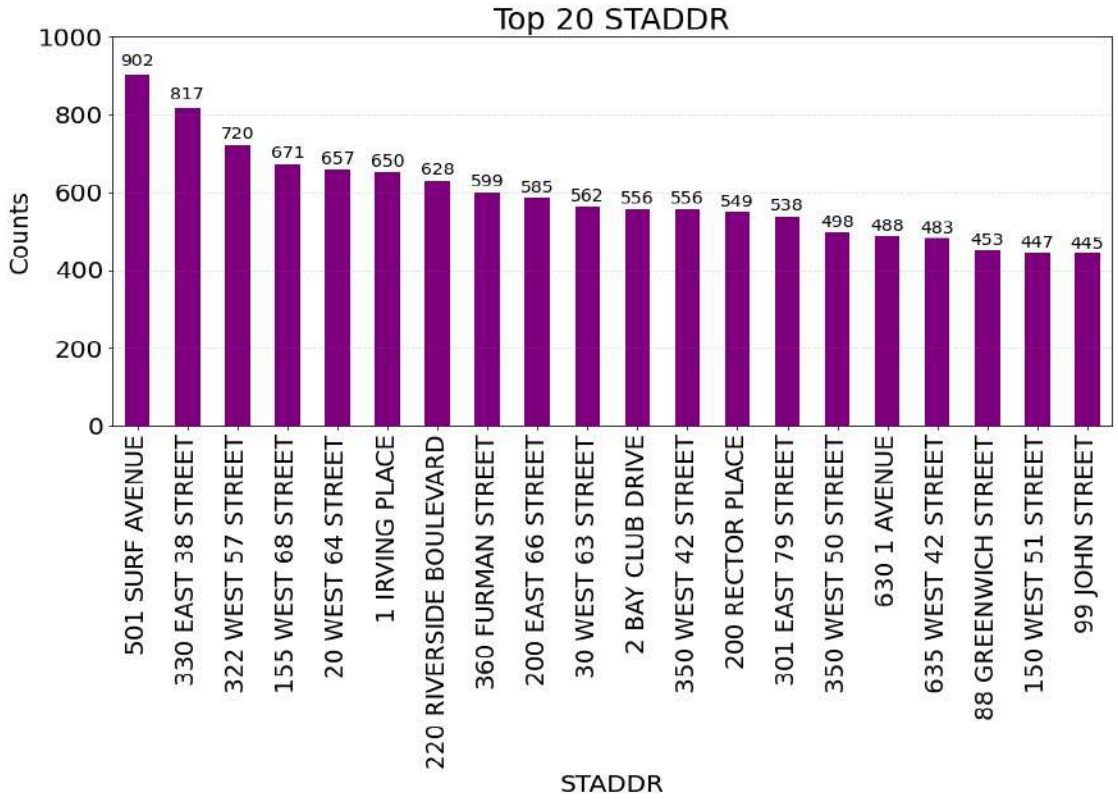
19. Field Name : EXCD1

Description: Exemption code 1, providing specific details on the type of exemption.
59.6% of this field is populated.



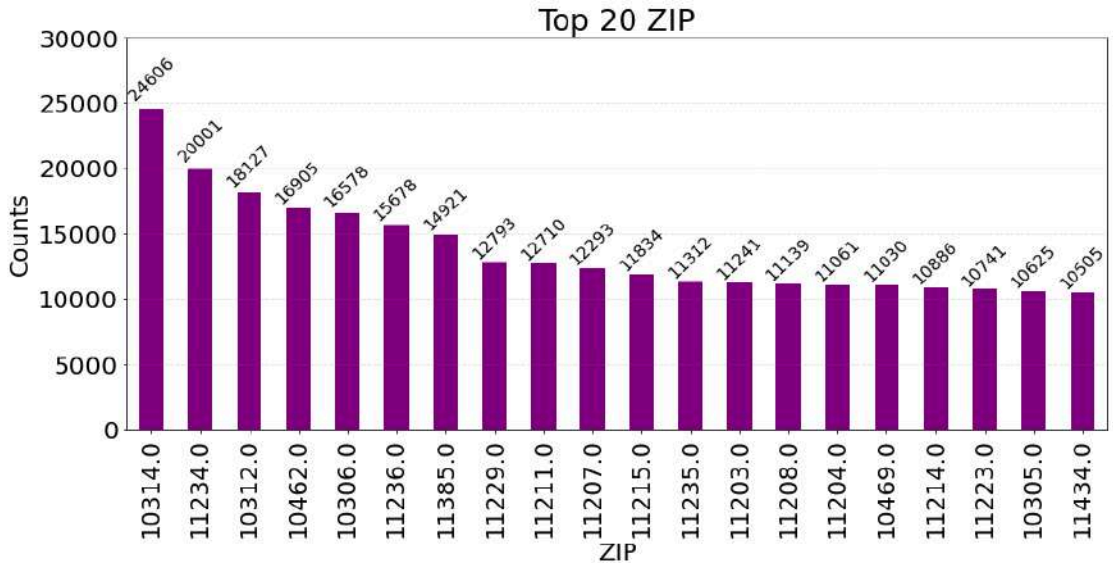
20. Field Name : STADDR

Description: Street address of the property. 99.93% of this field is populated.



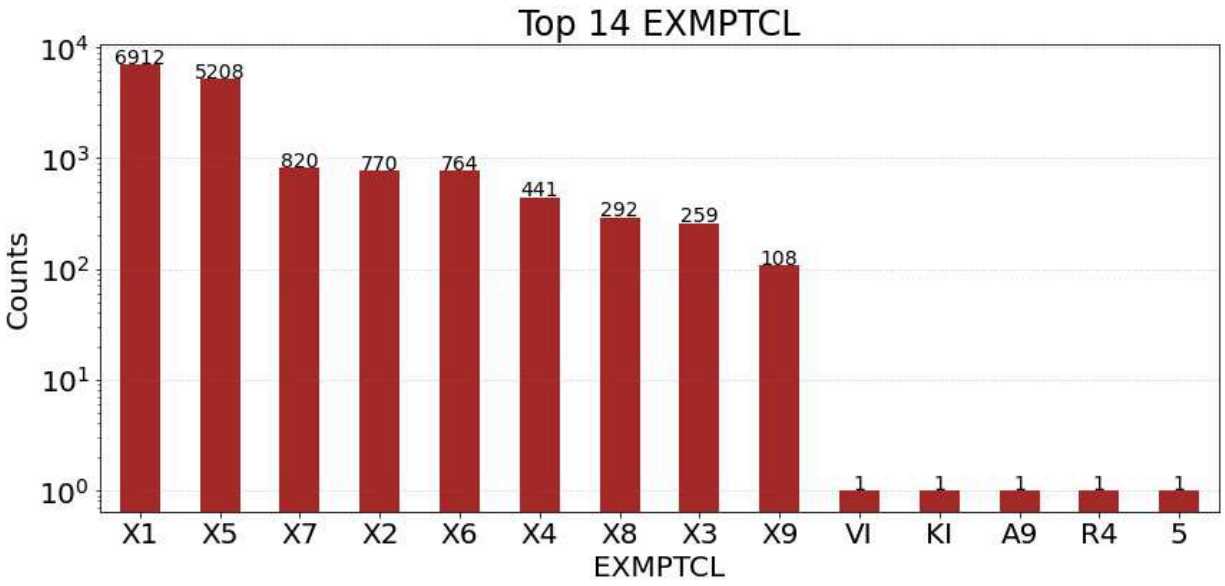
21. Field Name : ZIP

Description: Street address of the property. 97.2% of this field is populated.



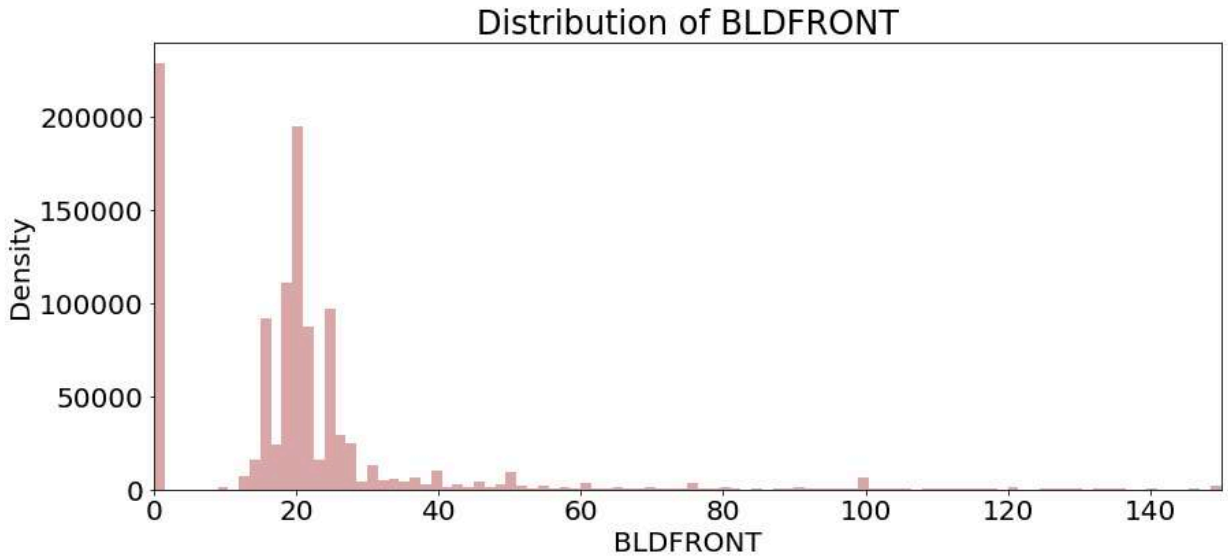
22. Field Name : EXMPTCL

Description: Exemption class, specifying the category of exemption. Only 1.45% of this field is populated.

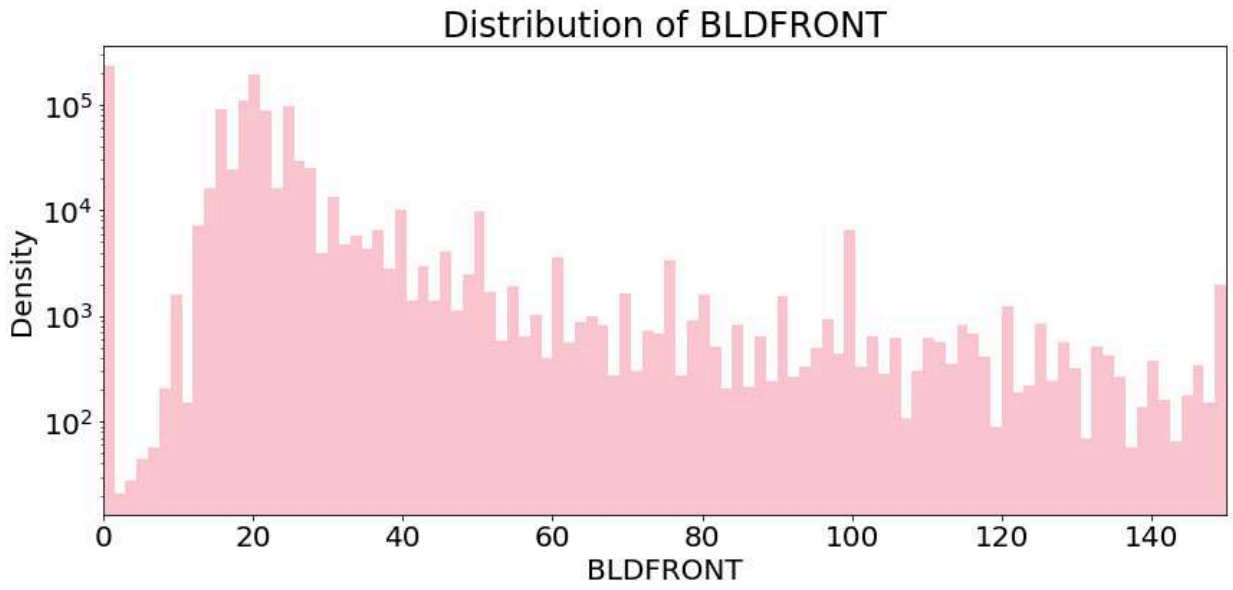


23. Field Name : BLDFRONT

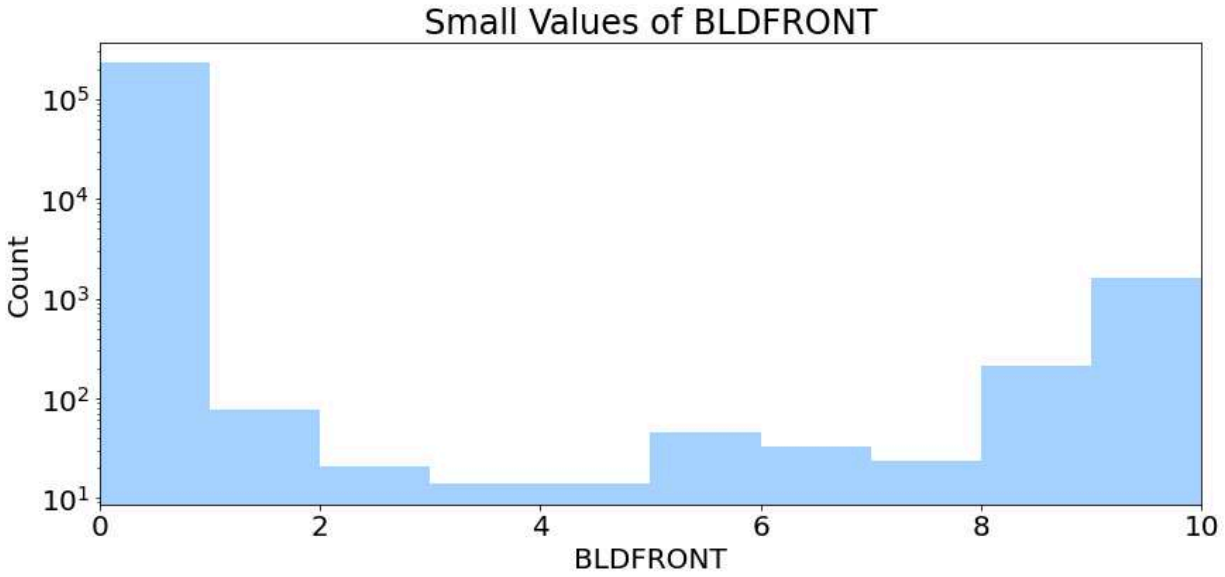
Description: Width of the building's front in feet. Only 1.45% of this field is populated.



Looking at the most relevant range for BLDFRONT:

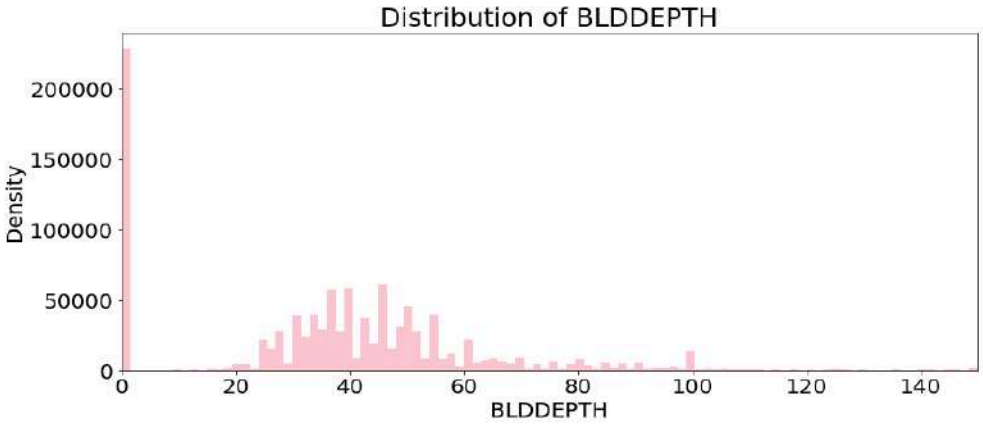


Looking at the small values:

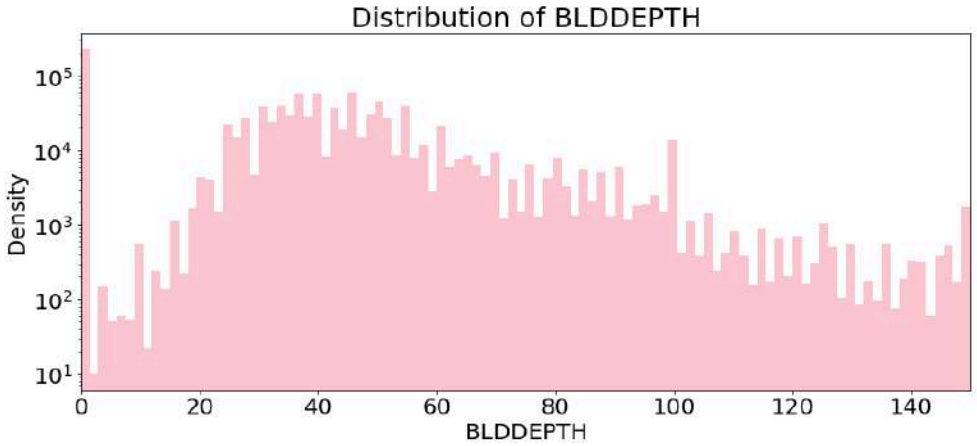


24. Field Name : BLDDEPTH

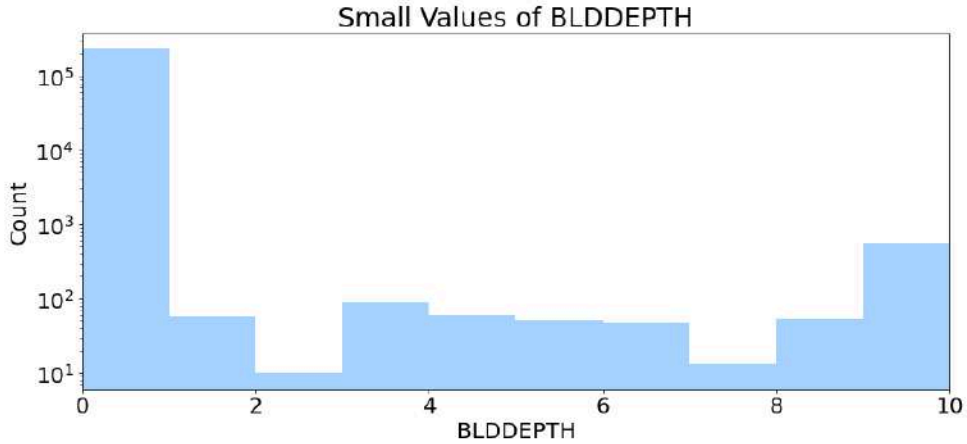
Description: Depth of the building in feet.



Looking at the most relevant range for BLDFRONT:

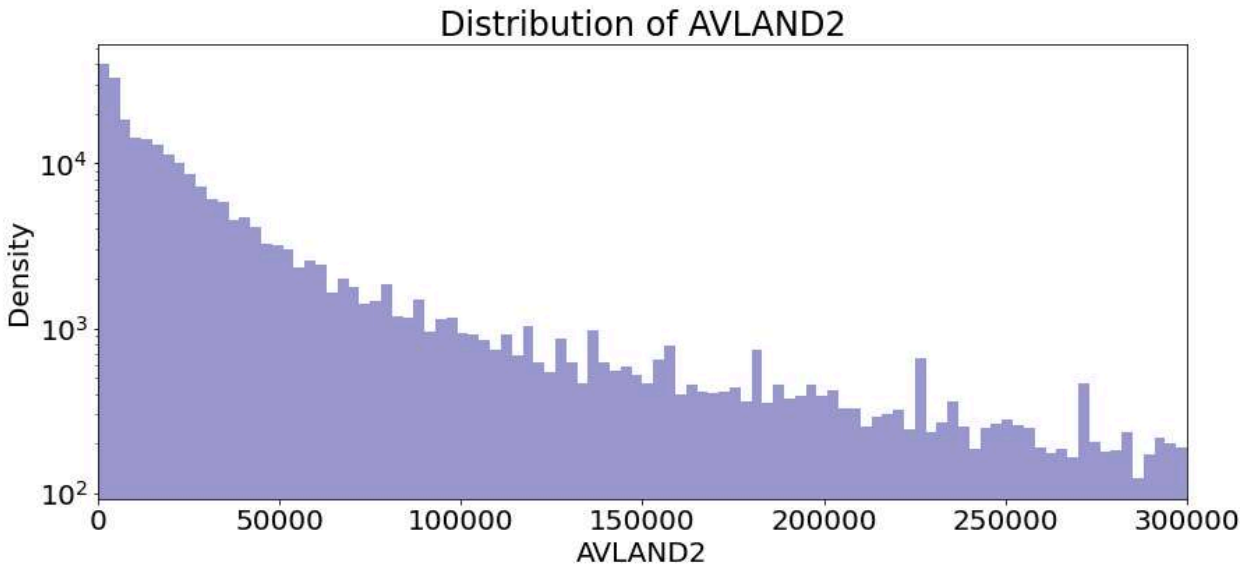


Looking at the small values:



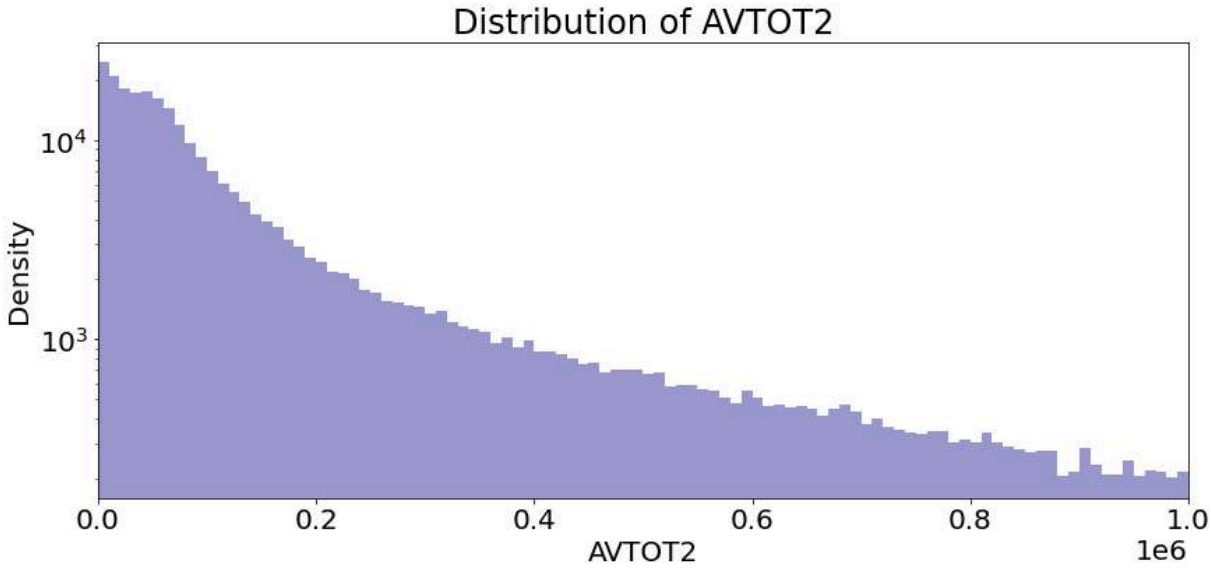
25. Field Name : AVLAND2

Description: Second assessed value of the land, for additional valuation scenarios. Only 26.39% of this field is populated.



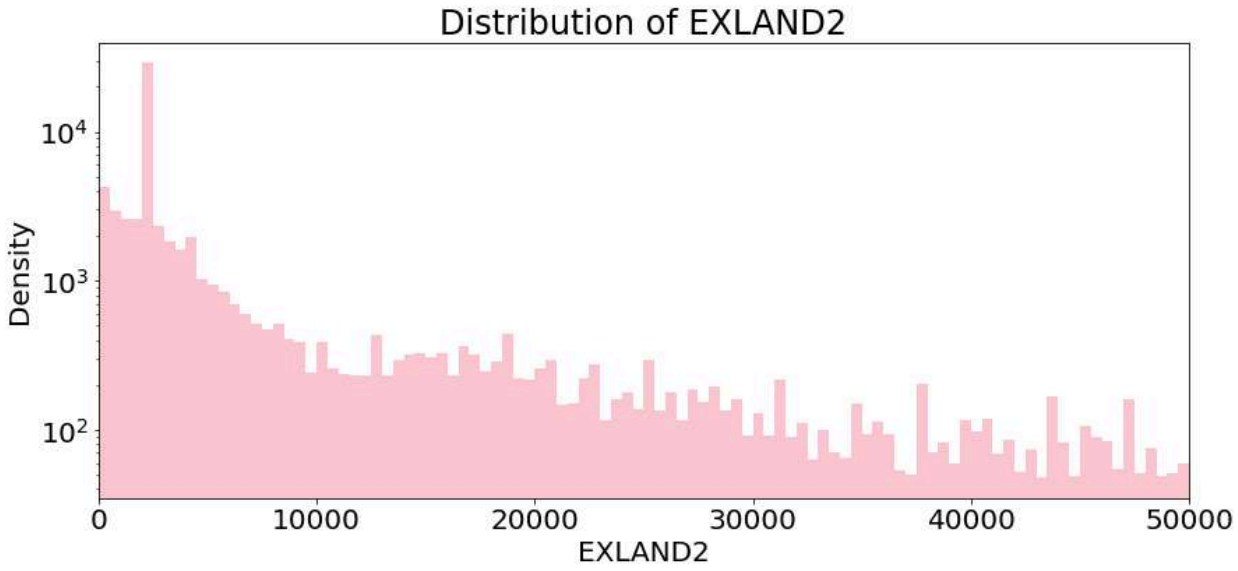
26. Field Name : AVTOT2

Description: Second total exempt value, for additional valuation scenarios. Only 26.39% of this field is populated.



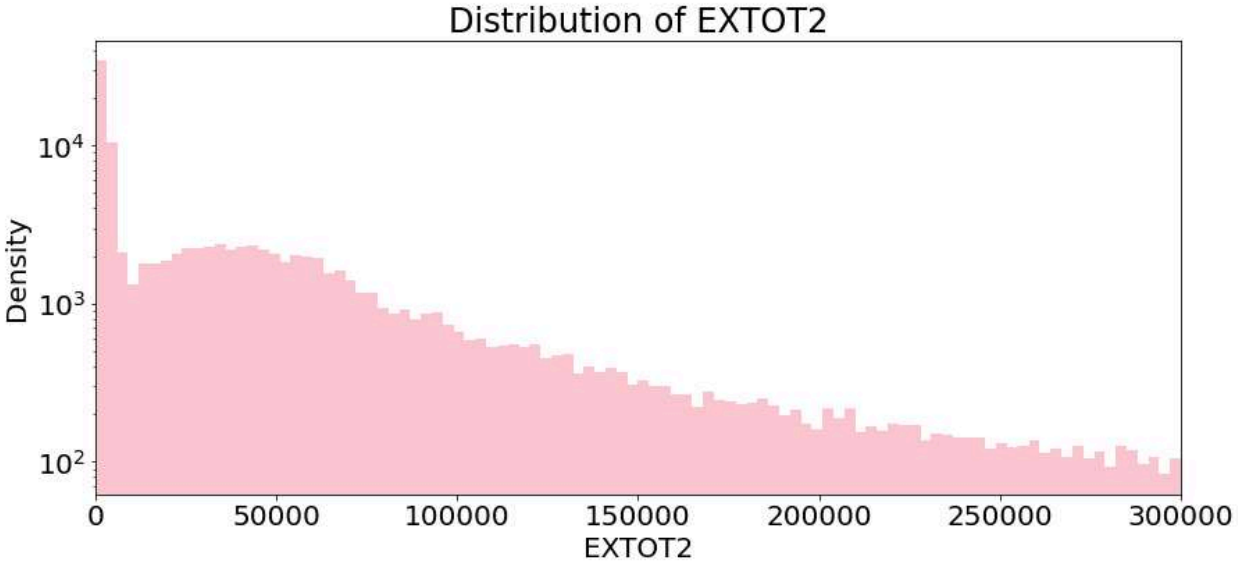
27. Field Name : EXLAND2

Description: Second exempt land value, for additional valuation scenarios. Only 8.16% of this field is populated.



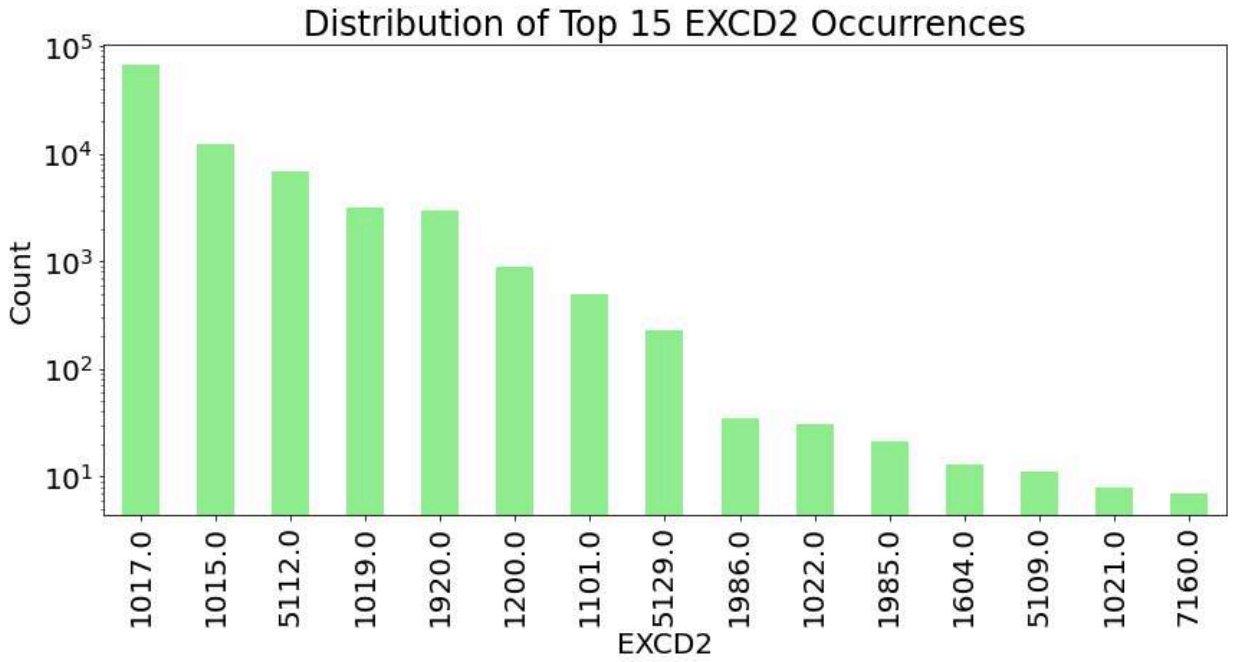
28. Field Name : EXTOT2

Description: Second total exempt value, for additional valuation scenarios. Only 12.21% of this field is populated.

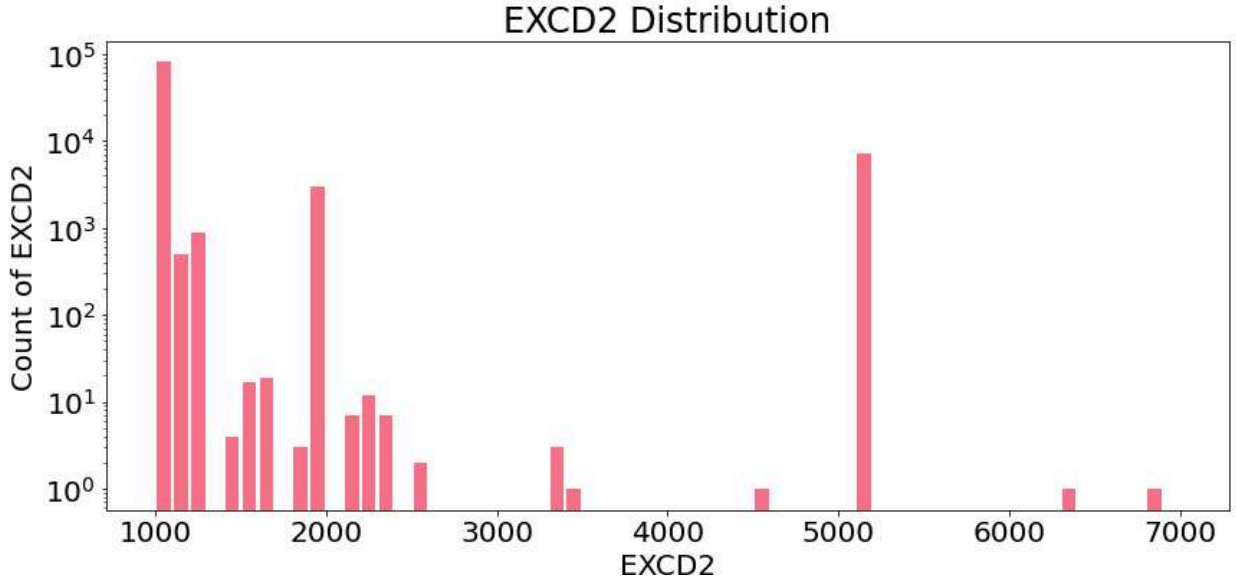


29. Field Name : EXTOT2

Description: Exemption code 2, providing further details on additional exemptions. Only 8.67% of this field is populated.



Distribution of EXCD2:



30. Field Name : PERIOD

Description: Time period of the assessment, indicating the specific timeframe covered. There is only 1 unique value- “Final”.

31. Field Name : Year

Description: Year of the assessment. There is only 1 unique value- 2010/11.

32. Field Name : VALTYPE

Description: Type of valuation, describing the nature of the property valuation (e.g., final, temporary). There is only 1 unique value- “AC-TR”