Project Report
Tax Fraud Detection for NY Property Data

Abhisheik Jadhav
A69027702

Table of Contents:

1. Executive Summary:

The New York Property Tax Fraud Detection project aimed to address the pressing issue of property tax fraud within New York City's extensive property database. Tax fraud can manifest through underreporting property values, misclassifying property types, or manipulating property characteristics to reduce tax liabilities. These fraudulent activities result in significant revenue losses for the city and create an unfair tax burden on compliant property owners. The primary objective of this project was to develop a robust system to identify potentially fraudulent property records, assisting city authorities in ensuring tax compliance and fairness.

Our approach combined data cleaning, variable creation, dimensionality reduction using PCA, and advanced anomaly detection algorithms. The project successfully identified numerous properties with suspicious characteristics that could indicate potential tax fraud. Key findings include:

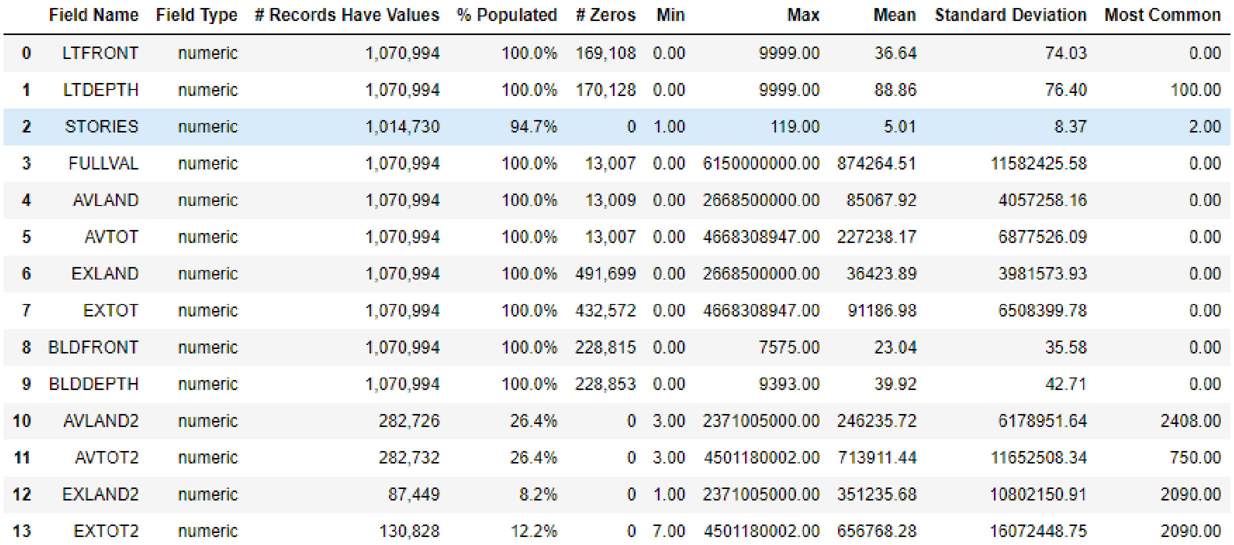
High Fraud Scores: Several properties exhibited high fraud scores based on a combination of PCA transformations and neural network reconstruction errors.
Inconsistent Valuation Metrics: Properties were found with full market values, assessed land values, and total assessed values that did not align with their physical characteristics or reported usage.
Unusual Size and Value Ratios: Anomalies in size and value ratios were detected, suggesting discrepancies between reported property dimensions and their valuations.
Significant Irregularities: Detailed examination of the top flagged properties revealed substantial irregularities, warranting further investigation by city authorities.

2. Description of the Data

The dataset "Property Valuation and Assessment Data," comprises **1,070,994 records** detailing real estate assessment property data for the **year 2010/11**. It encompasses **32 fields** and is updated annually by the Department of Finance. This dataset provides an extensive collection of information related to property assessments, evaluations, and condensed roll data within the city government category. Key fields include property ID, assessment value, property type, location, and other relevant property-specific information.

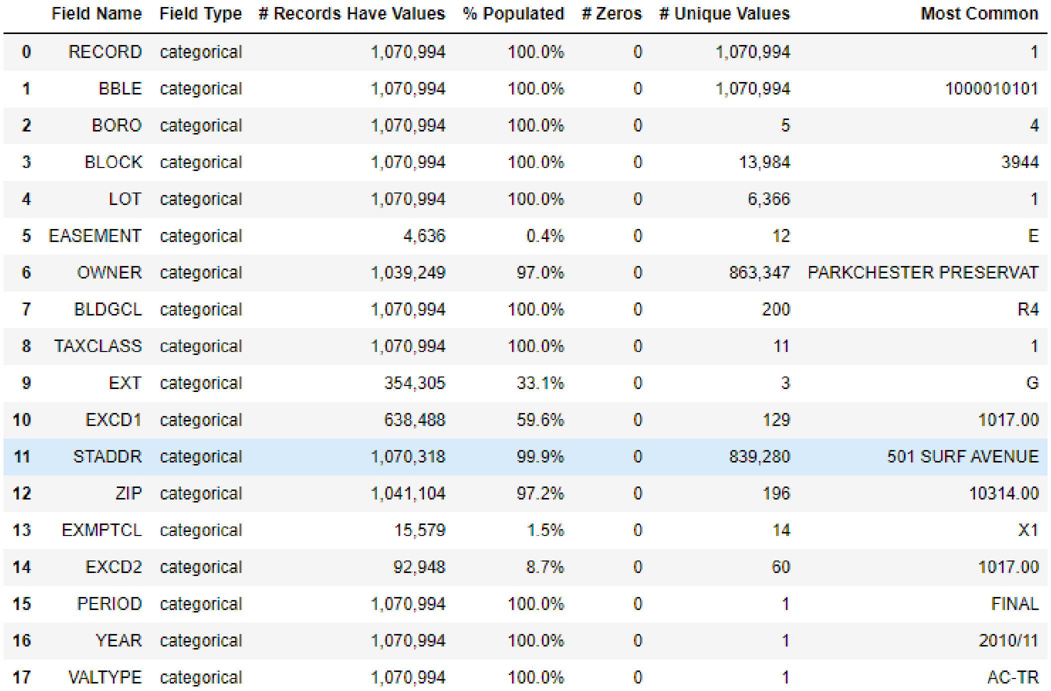
Statistic Tables:

Numeric Fields Table



	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	LTFRONT	numeric	1,070,994	100.0%	169,108	0.00	9999.00	36.64	74.03	0.00
1	LTDEPTH	numeric	1,070,994	100.0%	170,128	0.00	9999.00	88.86	76.40	100.00
2	STORIES	numeric	1,014,730	94.7%	0	1.00	119.00	5.01	8.37	2.00
3	FULLVAL	numeric	1,070,994	100.0%	13,007	0.00	6150000000.00	874264.51	11582425.58	0.00
4	AVLAND	numeric	1,070,994	100.0%	13,009	0.00	2668500000.00	85067.92	4057258.16	0.00
5	AVTOT	numeric	1,070,994	100.0%	13,007	0.00	4668308947.00	227238.17	6877526.09	0.00
6	EXLAND	numeric	1,070,994	100.0%	491,699	0.00	2668500000.00	36423.89	3981573.93	0.00
7	EXTOT	numeric	1,070,994	100.0%	432,572	0.00	4668308947.00	91186.98	6508399.78	0.00
8	BLDFRONT	numeric	1,070,994	100.0%	228,815	0.00	7575.00	23.04	35.58	0.00
9	BLDDEPTH	numeric	1,070,994	100.0%	228,853	0.00	9393.00	39.92	42.71	0.00
10	AVLAND2	numeric	282,726	26.4%	0	3.00	2371005000.00	246235.72	6178951.64	2408.00
11	AVTOT2	numeric	282,732	26.4%	0	3.00	4501180002.00	713911.44	11652508.34	750.00
12	EXLAND2	numeric	87,449	8.2%	0	1.00	2371005000.00	351235.68	10802150.91	2090.00
13	EXTOT2	numeric	130,828	12.2%	0	7.00	4501180002.00	656768.28	16072448.75	2090.00

Categorical Fields Table

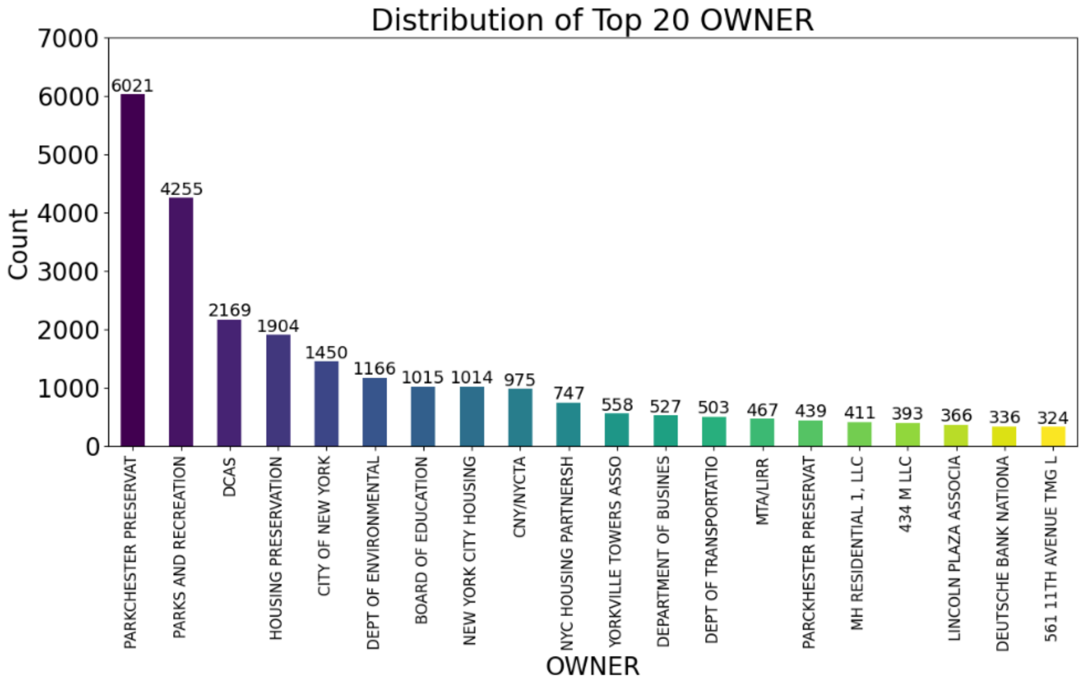


	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	RECORD	categorical	1,070,994	100.0%	0	1,070,994	1
1	BBLE	categorical	1,070,994	100.0%	0	1,070,994	1000010101
2	BORO	categorical	1,070,994	100.0%	0	5	4
3	BLOCK	categorical	1,070,994	100.0%	0	13,984	3944
4	LOT	categorical	1,070,994	100.0%	0	6,366	1
5	EASEMENT	categorical	4,636	0.4%	0	12	E
6	OWNER	categorical	1,039,249	97.0%	0	863,347	PARKCHESTER PRESERVAT
7	BLDGCL	categorical	1,070,994	100.0%	0	200	R4
8	TAXCLASS	categorical	1,070,994	100.0%	0	11	1
9	EXT	categorical	354,305	33.1%	0	3	G
10	EXCD1	categorical	638,488	59.6%	0	129	1017.00
11	STADDR	categorical	1,070,318	99.9%	0	839,280	501 SURF AVENUE
12	ZIP	categorical	1,041,104	97.2%	0	196	10314.00
13	EXMPTCL	categorical	15,579	1.5%	0	14	X1
14	EXCD2	categorical	92,948	8.7%	0	60	1017.00
15	PERIOD	categorical	1,070,994	100.0%	0	1	FINAL
16	YEAR	categorical	1,070,994	100.0%	0	1	2010/11
17	VALTYPE	categorical	1,070,994	100.0%	0	1	AC-TR

Some Important Fields:

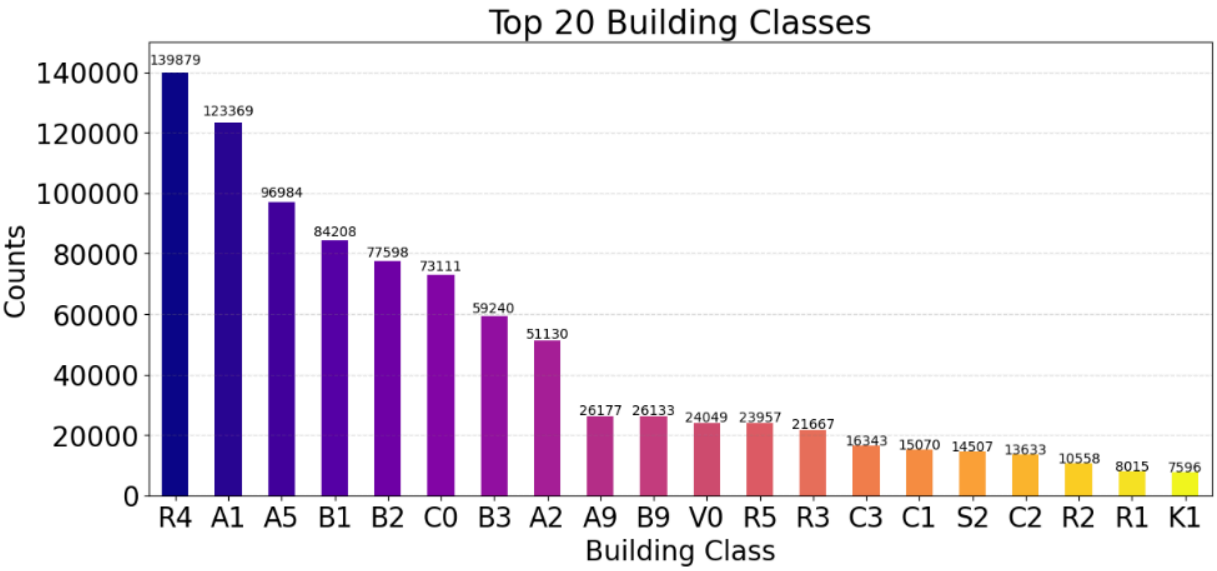
OWNER:

The "OWNER" field, capturing categorical data representing owner names, is prevalent in 97.0% of the 1,070,994 records, with no zero entries. It encompasses a wide diversity of 863,347 unique owners, with "PARKCHESTER PRESERVAT" being the most frequent. A visual representation highlights the top 20 owners for further analysis.



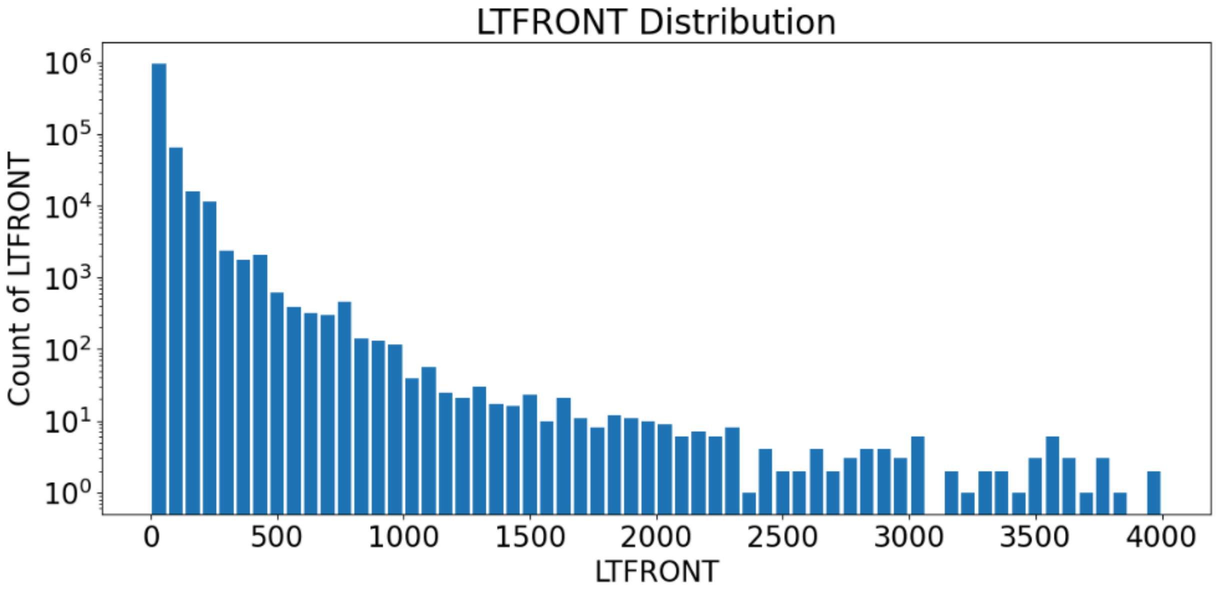
BLDGCL:

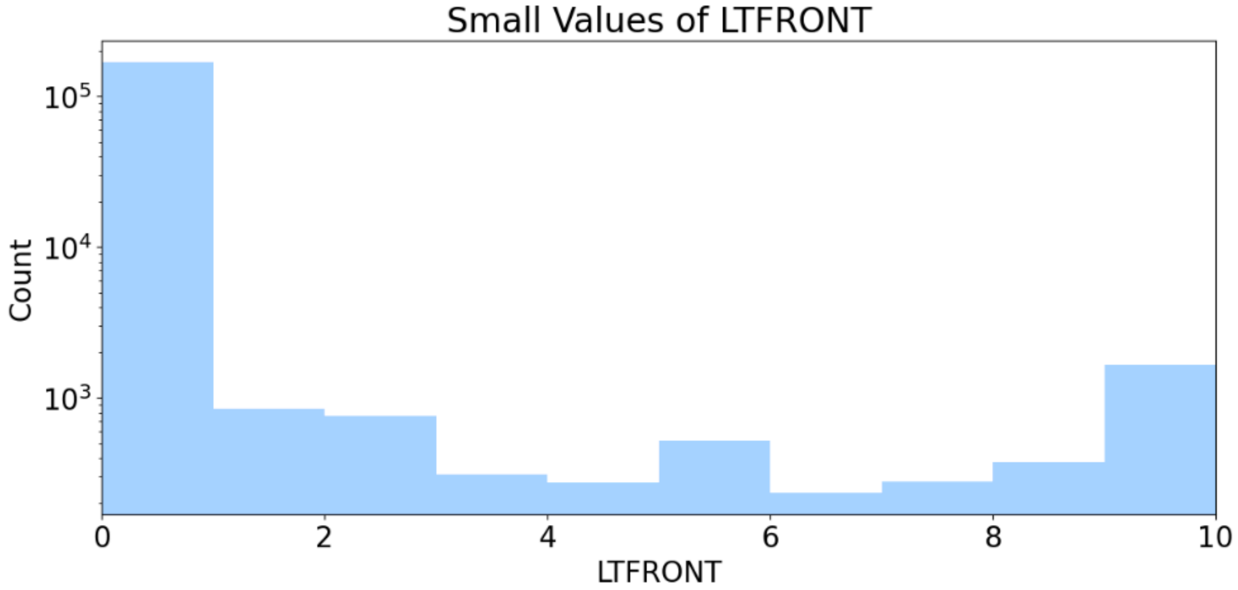
The "BLDGCL" field, categorical and present in all 1,070,994 records without any zeros, represents building classes. With 200 unique categories, "R4" emerges as the most common class. A visual overview displays the top 20 building classes.



LTFRONT:

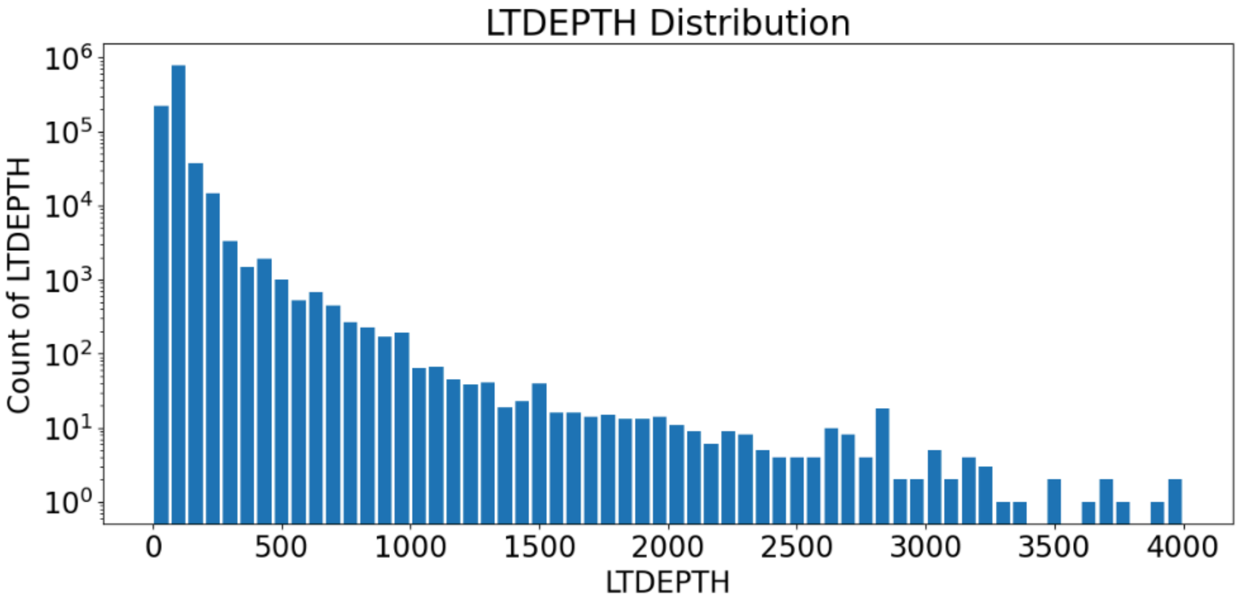
The "LTFRONT" field, a numeric attribute, is fully populated across 1,070,994 records. It denotes lot width, ranging from 0.00 to 9999.00, with an average width of 36.64 and a median of 74.03. Visualizations include distributions of widths and a focused one on smaller values.

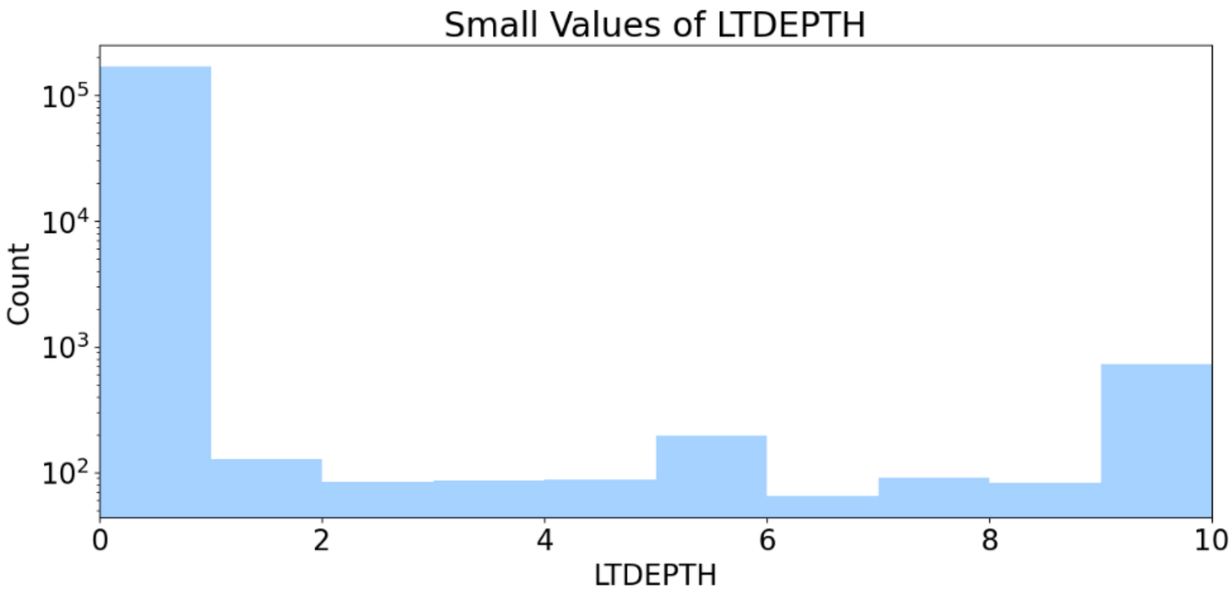




LTDEPTH:

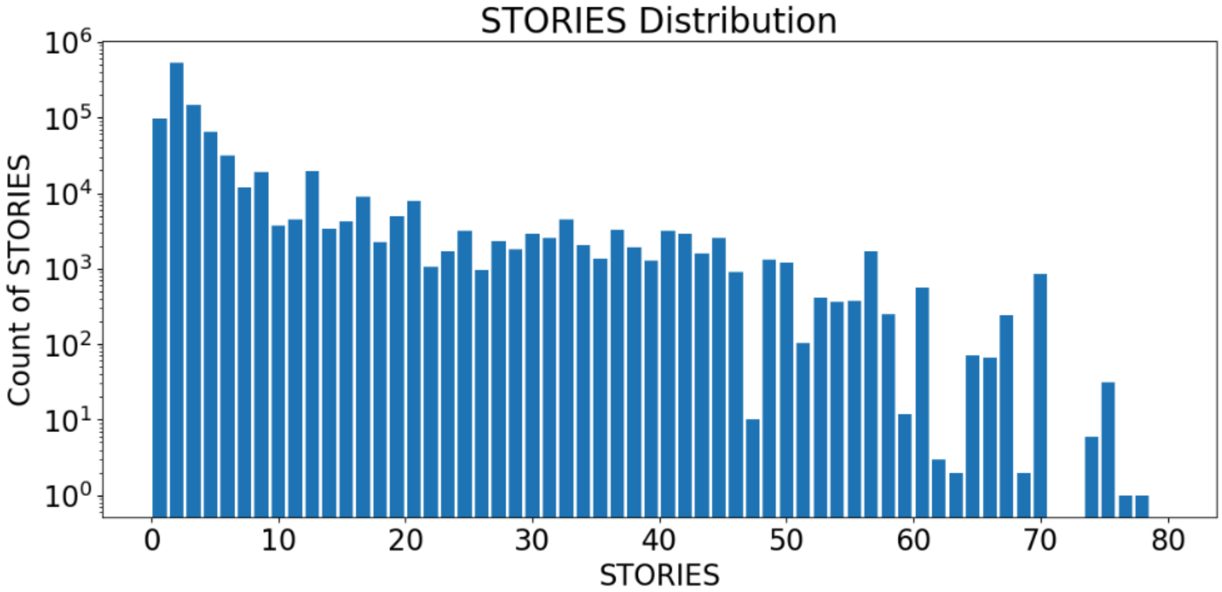
The "LTDEPTH" field, a numeric attribute, is fully populated across 1,070,994 records. It represents lot depth, ranging from 0.00 to 9999.00, with an average depth of 88.86 and a median of 76.40. Visualizations encompass distributions of depths and focus on the distribution of smaller values.





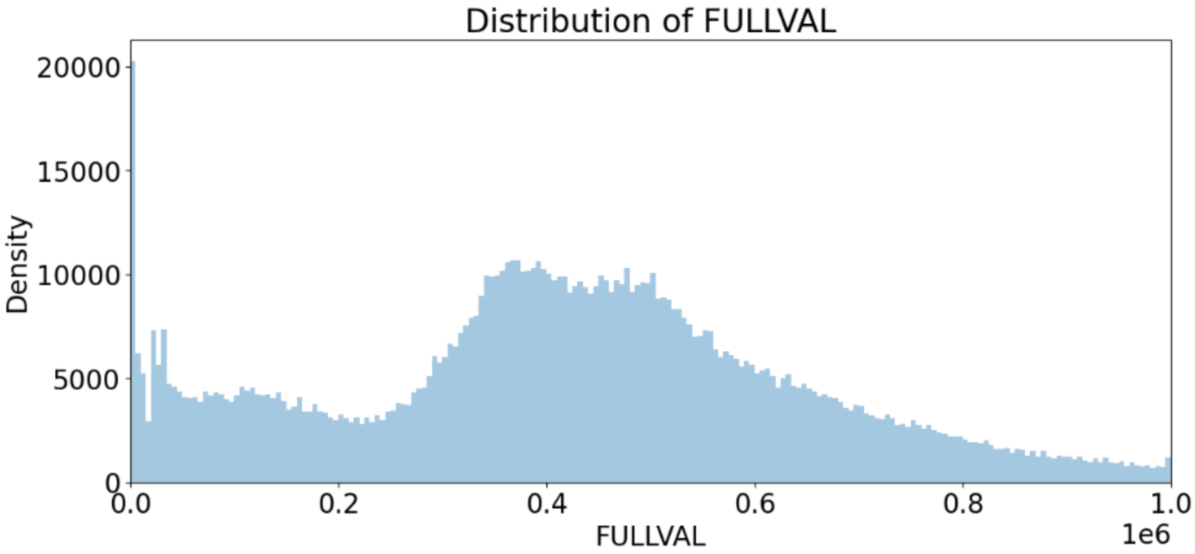
STORIES:

The "STORIES" field, a numeric attribute, is present in 94.7% of the 1,014,730 records, indicating the number of stories in a building. With values ranging from 1.00 to 119.00, the average number of stories is 5.01, with a median of 8.37. Visualizations encompass the distribution of stories within buildings.



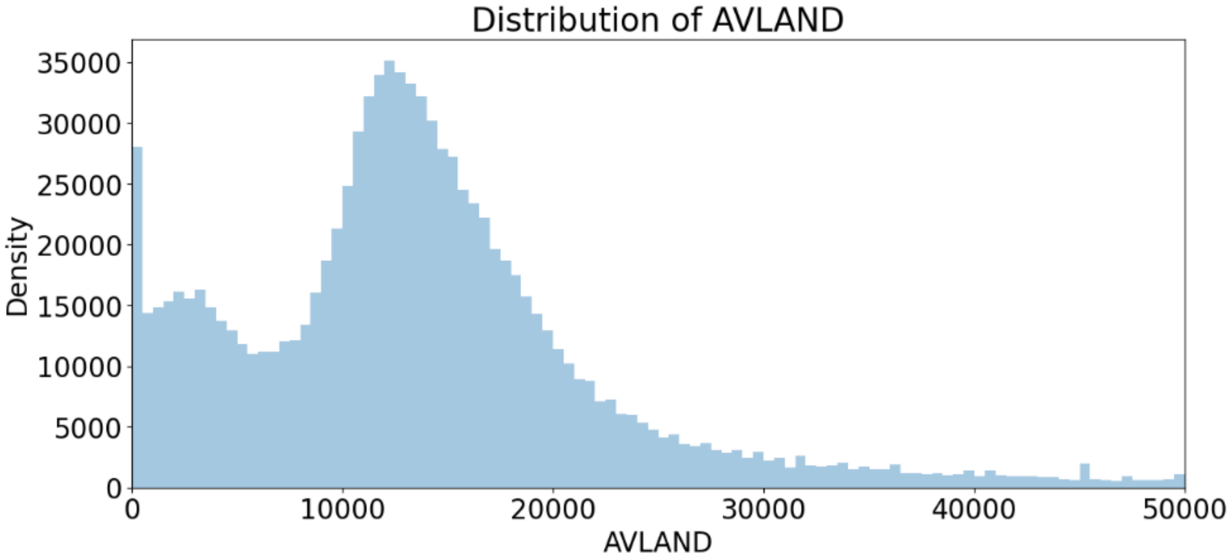
FULLVAL:

The "FULLVAL" field, a numeric attribute, is fully populated across 1,070,994 records, representing the market value. The values range from 0.00 to 6,150,000,000.00, with an average value of $874,264.51 and a median of $11,582,425.58. Visualizations showcase the distribution of market values (FULLVAL).



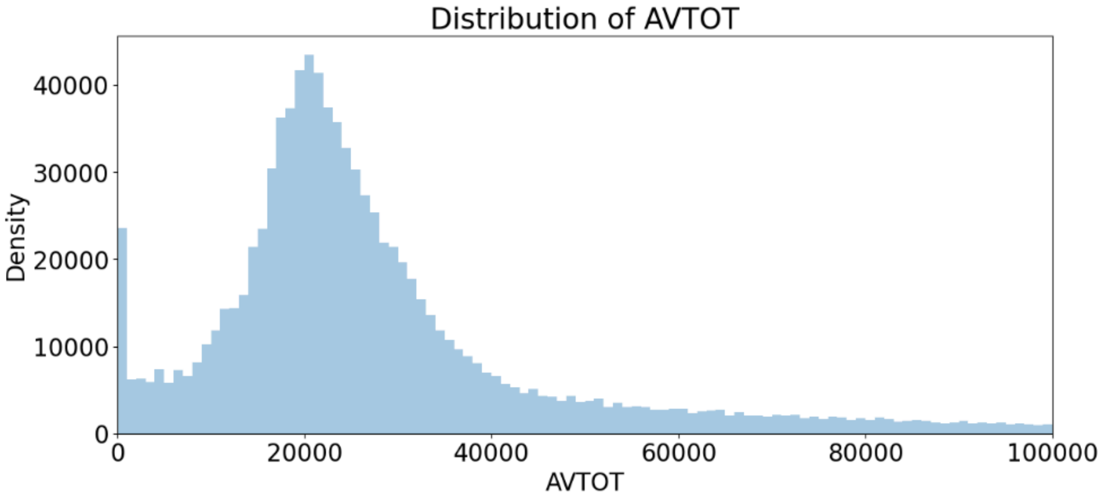
AVLAND:

The "AVLAND" field, a numeric attribute present in all 1,070,994 records, represents the actual land value. Values range from 0.00 to 2,668,500,000.00, with an average value of $85,067.92 and a median of $4,057,258.16. Visualizations illustrate the distribution of actual land values (AVLAND).



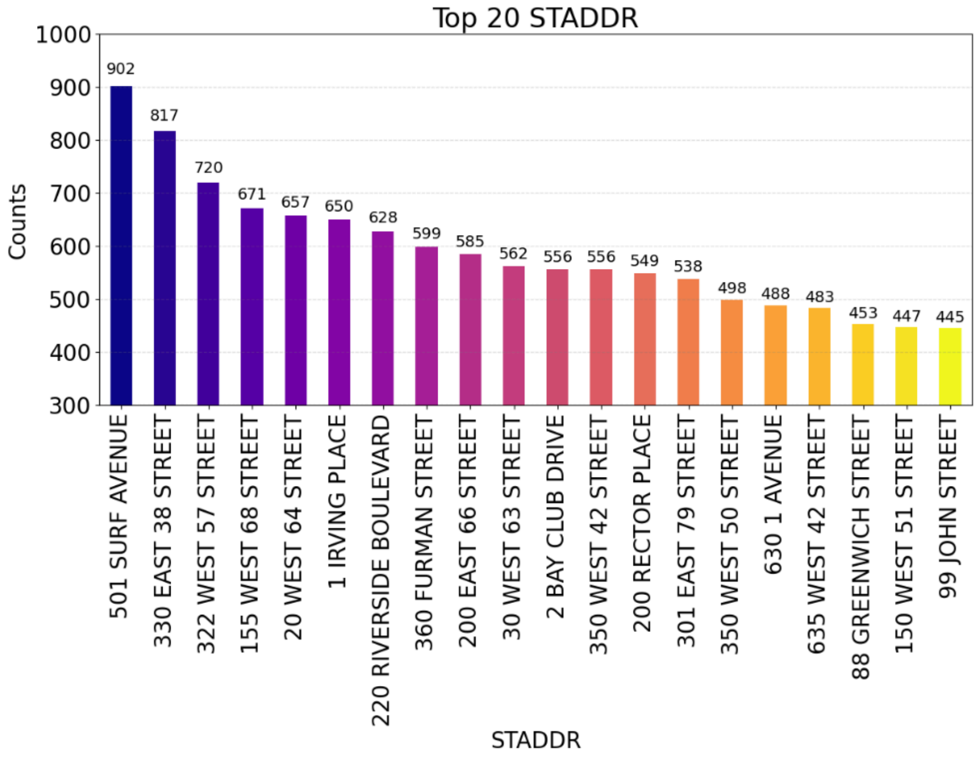
AVTOT:

The "AVTOT" field, a numeric attribute present in all 1,070,994 records, represents the actual total value. Values range from 0.00 to 4,668,308,947.00, with an average value of $227,238.17 and a median of $6,877,526.09. Visualizations illustrate the distribution of actual total values (AVTOT).



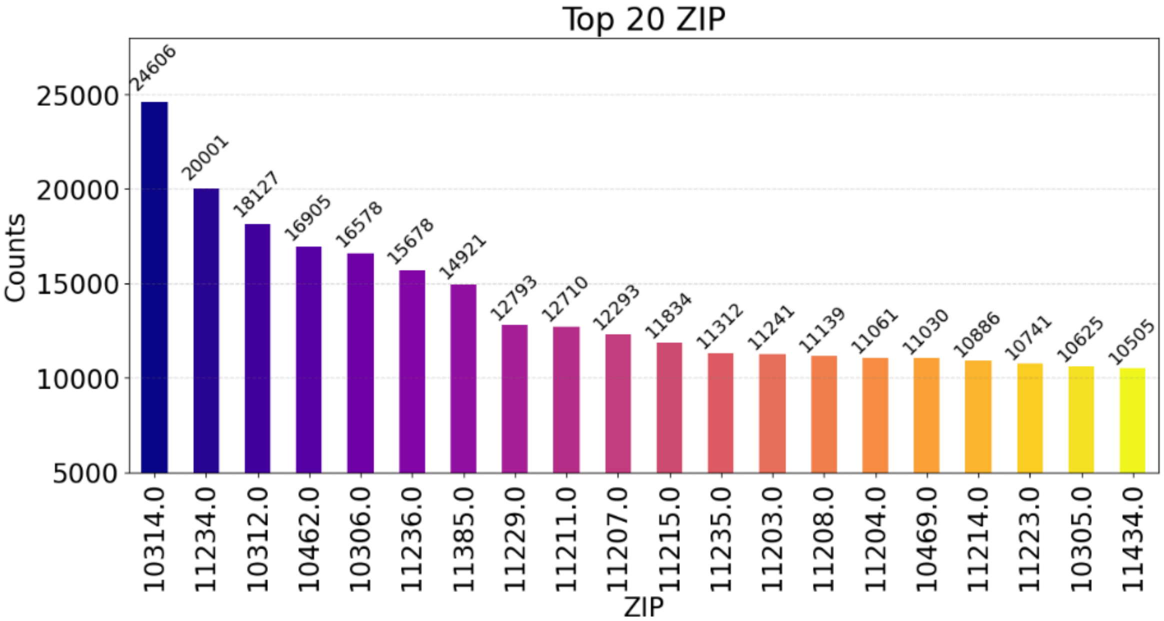
STADDR:

The "STADDR" field, a categorical attribute, is present in 99.9% of the 1,070,318 records, representing street addresses. With 839,280 unique addresses, the top 20 occurrences are visualized for analysis.



ZIP:

The "ZIP" field, a categorical attribute, is present in 97.2% of the 1,041,104 records, indicating zip codes. With 196 unique zip codes, the top 20 occurrences are visualized for analysis.



3. Data Cleaning

In this section, we have performed **data cleaning and variable creation**. Our objective was to refine the dataset, generate informative variables, and **avoid the removal of outliers** to facilitate a thorough analysis of potential fraud indicators.

1. Logic and Results of Exclusions:

A total of **26,501 property records** were excluded from the dataset based on stakeholder guidance to focus on private ownership. These exclusions primarily targeted government-owned properties, public entities, and cemeteries using criteria such as easement types, owner names indicative of government ownership, and frequent owners representing public institutions. **The rationale behind this is that government entities are unlikely to be involved in tax fraud, so removing these properties helps streamline the analysis**. This refined dataset ensures that the analysis targets private property owners, increasing the relevance and accuracy of subsequent fraud detection efforts. After these exclusions, imputation techniques were applied to fill in missing values, ensuring a complete and robust dataset for analysis.

2. Logic and Results of Field Imputations:

Imputation for Zip codes

- The dataset initially had **20,431** missing ZIP codes.
- A new column was created by concatenating street addresses and boroughs to create a unique identifier for each address-borough combination.
- Missing ZIP codes were filled by mapping the unique address-borough combinations to their corresponding ZIP codes. This step filled in **2,832** ZIP codes.
- Remaining missing ZIP codes were further filled if the ZIP codes before and after a missing entry were the same. This filled in an additional **9,491** ZIP codes.

- For the remaining missing ZIP codes, the ZIP code of the previous record was used, which filled the remaining **8,108** missing ZIP codes, resulting in no missing ZIP codes in the dataset.

Imputation for FULLVAL, AVLAND, and AVTOT:

- Missing values and zeros were initially replaced with NaN.
- **FULLVAL**
  - Initial Statistics: Missing/Zero Values: **10,025**
  - Step 1: Filling with Mean Values (Grouped by TAXCLASS, BORO, and BLDGCL)
    - Results: Reduced missing values to **7,307**.
  - Step 2: Filling with Mean Values (Grouped by TAXCLASS and BORO)
    - Results: Further reduced missing values to **386**.
  - Step 3: Filling with Mean Values (Grouped by TAXCLASS)
    - Results: All missing values were filled.
  - Final Result: No missing values in FULLVAL.

- **AVLAND**
  - Initial Statistics: Missing/Zero Values: **10,027**
  - Step 1: Filling with Mean Values (Grouped by TAXCLASS, BORO, and BLDGCL)
    - Results: Reduced missing values to **7,307**.
  - Step 2: Filling with Mean Values (Grouped by TAXCLASS and BORO)
    - Results: Further reduced missing values to **386**.
  - Step 3: Filling with Mean Values (Grouped by TAXCLASS)
    - Results: All missing values were filled.
  - Final Result: No missing values in AVLAND.

- **AVTOT**
  - Initial Statistics: Missing/Zero Values: **10,025**
  - Step 1: Filling with Mean Values (Grouped by TAXCLASS, BORO, and BLDGCL)
    - Results: Reduced missing values to **7,307**.

  - Step 2: Filling with Mean Values (Grouped by TAXCLASS and BORO)
    - Results: Further reduced missing values to **386**.
  - Step 3: Filling with Mean Values (Grouped by TAXCLASS)
  - Results: All missing values were filled.
  - Final Result: No missing values in AVTOT

Imputation for STORIES:

- Initial Missing Values: The STORIES column had **42,030** missing values.
- Step 1: Filling with Mode (Grouped by BORO and BLDGCL)
  - Missing values were filled with the mode of STORIES for each combination of BORO and BLDGCL.
  - Results: Substantial reduction in missing values. **37922** remaining.
- Step 2: Filling with Mean (Grouped by TAXCLASS)
  - Remaining missing values were filled with the mean of STORIES grouped by TAXCLASS
  - Results: All missing values were filled.
- Final Result: No missing values in the STORIES column.

Imputation for LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT with averages by TAXCLASS:

- **LTFRONT**
  - Initial Missing and Zero Values: Identified **160,565** zero values.
  - Transformation Steps:
    - Replace zeros with NaN.
    - Fill NaN values with the mean of LTFRONT grouped by TAXCLASS and BORO.
    - Remaining NaN values filled with the mean of LTFRONT grouped by TAXCLASS.
  - Result: No missing values in LTFRONT.

- **LTDEPTH**
  - Initial Missing and Zero Values: Identified **161,656** zero values.
  - Transformation Steps:
    - Replace zeros with NaN.

        - Fill NaN values with the mean of LTDEPTH grouped by TAXCLASS and BORO.
        - Remaining NaN values filled with the mean of LTDEPTH grouped by TAXCLASS.
    - Result: No missing values in LTDEPTH.

- **BLDFRONT**
    - Initial Values: Initially no missing values, but zeros are considered invalid.
    - Transformation Steps:
        - Replace zeros with NaN.
        - Fill NaN values with the mean of BLDFRONT grouped by TAXCLASS, BORO, and BLDGCL.
        - Further fill NaN values with the mean of BLDFRONT grouped by TAXCLASS and BORO.
        - Final step involves filling NaN values with the mean of BLDFRONT grouped by TAXCLASS.
    - Result: No missing values in BLDFRONT.

- **BLDDEPTH**
    - Initial Values: Initially no missing values, but zeros are considered invalid.
    - Transformation Steps:
        - Replace zeros with NaN.
        - Fill NaN values with the mean of BLDDEPTH grouped by TAXCLASS, BORO, and BLDGCL.
        - Further fill NaN values with the mean of BLDDEPTH grouped by TAXCLASS and BORO.
        - Final step involves filling NaN values with the mean of BLDDEPTH grouped by TAXCLASS.
    - Result: No missing values in BLDDEPTH.

- **ZIP Column:**
    - Converted ZIP column from float to string.

- Created a new column **zip3** to store the first three digits of the ZIP codes.

4. Variable Creation:

Description of what kind of anomalies we're looking for.

In this tax fraud detection project, we are primarily looking for anomalies that indicate potential underreporting or misrepresentation of property data. Specifically, we are targeting properties with unusual financial or physical characteristics that deviate significantly from the norm. Examples include properties with discrepancies between reported and actual usage, unexpected patterns in property valuation or tax assessments, and inconsistencies in ownership or transaction records. These anomalies suggest possible tax evasion or fraud, warranting further investigation.

Logic for all created Variables:

r1 to r9:

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

These variables represent the ratios of each of the three $ value fields (FULLVAL, AVLAND, AVTOT) standardized by each of the three sizes (lotarea, bldarea, bldvol). Each ratio provides insight into how the property values relate to their corresponding sizes.

r1_zip5 to r9_zip5:

These variables are similar to r1 to r9 but are grouped by ZIP codes. They provide information on how the ratios of property values to sizes compare to the average ratios within each ZIP code. This helps identify outliers within each ZIP code area.

r1_taxclass to r9_taxclass:
Similar to r1 to r9, but these variables are grouped by TAXCLASS. They provide insights into how the ratios of property values to sizes compare to the average ratios within each tax class category. This helps identify outliers within each tax class.

$$\frac{r_1}{< r_1 >_g}, \quad \frac{r_2}{< r_2 >_g}, \quad \frac{r_3}{< r_3 >_g}, \quad \ldots \quad \frac{r_9}{< r_9 >_g} \qquad g = \text{ZIP, TAXCLASS}$$

value_ratio:
This variable calculates the ratio of FULLVAL to the sum of AVLAND and AVTOT. It normalizes the ratio so that the mean is 1. The max of this normalized ratio or its inverse is taken, indicating whether the value ratio is unusually large or small. This variable helps assess how appropriately the three value fields relate to each other.

$$\text{Value ratio VR} = \text{FULLVAL}/(\text{AVLAND} + \text{AVTOT}).$$

size_ratio:
This variable compares the size of the building to the size of the lot. It checks if the building size is larger than the lot size, which could indicate potential data issues or anomalies.

$$\text{size ratio} = \frac{\text{building size}}{\text{lot size}} = \frac{\text{BLDFRONT x BLDDEPTH}}{\text{LTFRONT x LTDEPTH}}$$

<u>4. List of all created Variables:</u>

- **r1 to r9:** Ratios of property values (FULLVAL, AVLAND, AVTOT) standardized by sizes (lotarea, bldarea, bldvol).
- **r1_zip5 to r9_zip5:** Ratios grouped by ZIP codes, comparing property values to sizes against ZIP code averages.

- **r1_taxclass to r9_taxclass:** Ratios grouped by TAXCLASS, comparing property values to sizes against tax class averages.
- **value_ratio:** Ratio of FULLVAL to the sum of AVLAND and AVTOT, normalized to assess appropriateness of value fields' relation.
- **size_ratio:** Comparison of building size to lot size, identifying potential data anomalies.

Table of Created Variables:

Sr. No.	Variable
1.	r1
2.	r2
3.	r3
4.	r4
5.	r5
6.	r6
7.	r7
8.	r8
9.	r9
10.	r1_zip5
11.	r2_zip5
12.	r3_zip5
13.	r4_zip5
14.	r5_zip5
15.	r6_zip5
16.	r7_zip5

17.	r8_zip5
18.	r9_zip5
19.	r1_taxclass
20.	r2_taxclass
21.	r3_taxclass
22.	r4_taxclass
23.	r5_taxclass
24.	r6_taxclass
25.	r7_taxclass
26.	r8_taxclass
27.	r9_taxclass
28.	value_ratio
29.	size_ratio

5. Dimensionality Reduction:

Dimensionality reduction is the process of reducing the number of variables under consideration, simplifying the dataset while preserving its essential information. In the context of tax fraud detection in New York property data, dimensionality reduction is crucial for improving the efficiency and accuracy of anomaly detection algorithms.

Why is Dimensionality Reduction Important?

High Dimensionality Issues: The property data contains numerous variables, many of which may be correlated. High dimensionality can lead to overfitting, making it difficult to identify true patterns indicative of fraud.

Computational Efficiency: Reducing the number of dimensions decreases computational requirements, allowing for quicker analysis and real-time detection of fraudulent activities.

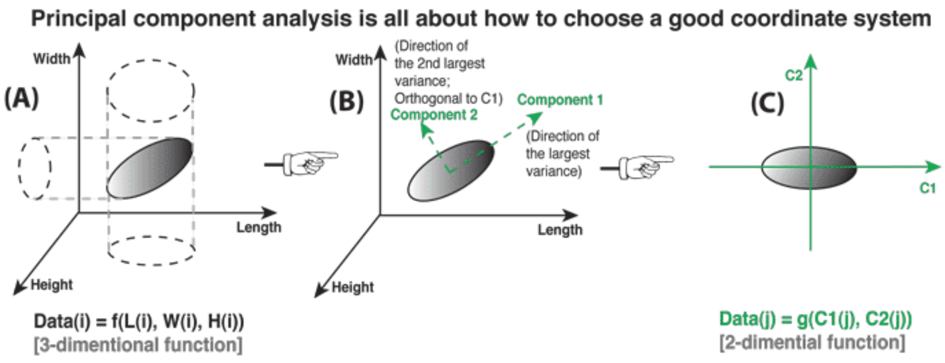
Improved Anomaly Detection: By focusing on the most significant features, the detection algorithms can more effectively identify unusual patterns that may indicate tax fraud.

How is it Implemented in our case?

Z-Scaling: Standardizing the data to have a mean of zero and a standard deviation of one, ensuring each variable contributes equally to the subsequent analysis.

Principal Component Analysis (PCA): Transforming the dataset into a set of orthogonal components that explain the maximum variance, and selecting the top components that capture the most significant features.

Post-PCA Z-Scaling: Ensuring that the retained principal components are equally important by standardizing them again.


Principal component analysis is all about how to choose a good coordinate system
(A)
Width
Length
Height
Data(i) = f(L(i), W(i), H(i))
[3-dimentional function]
(B)
Width
(Direction of the 2nd largest variance; Orthogonal to C1)
Component 2
Component 1
(Direction of the largest variance)
Length
Height
(C)
C2
C1
Data(j) = g(C1(j), C2(j))
[2-dimential function]

Steps Involved:

1) Initial Z-Scaling: Standardizing the data to have a mean of zero and a standard deviation of one.
2) Complete PCA: Performing PCA to identify the number of principal components that explain 99% of the variance.
3) Select Top Components: Redoing PCA to retain the top principal components (e.g., top five) based on the cumulative variance plot.
4) Post-PCA Z-Scaling: Standardize the principal components to ensure equal importance.

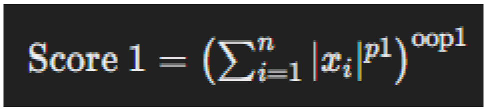
Dimensionality reduction through PCA and z-scaling is essential. These techniques streamline the dataset, reduce computational complexity, and improve the robustness and accuracy of fraud detection algorithms. By focusing on the most significant features, the project can more effectively identify fraudulent activities, making these transformations a vital component of the data analysis process.

6. Two Score Algorithms

Description and Formulas

In unsupervised fraud detection, two primary algorithms are often used to detect anomalies:

Minkowski Distance-based Score (Score 1): This method calculates the distance of each data point from the origin in the scaled principal component (PC) space. The formula for the Minkowski distance is given by:



$$\text{Score 1} = \left(\sum_{i=1}^{n} |x_i|^{p1}\right)^{oop1}$$

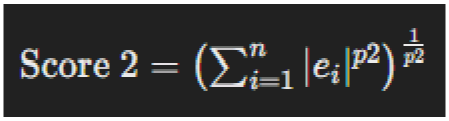
Where,

xi: The i-th component of the data point in the standardized principal component space.

pi: The power to which each absolute value of the principal component is raised.

oop1: The power to which the sum of the powered absolute values is raised.

Autoencoder-based Reconstruction Error (Score 2): This method involves training an autoencoder neural network to reconstruct the input data. The fraud score is the error between the original and reconstructed data points. The formula for the reconstruction error is:



$$\text{Score 2} = \left(\sum_{i=1}^{n} |e_i|^{p2}\right)^{\frac{1}{p2}}$$

Where,

ei is the reconstruction error for the i-th principal component.

p2 is the power parameter.

1/p2 is the reciprocal of p2, which we denote as oop2.

Why These Formulas Make Good Fraud Scores

Score 1 (Minkowski Distance-based Score):

- This score identifies outliers by measuring the distance from the origin in a transformed space where the most significant variance directions are considered.
- It is sensitive to both small and large deviations, making it effective in detecting anomalies that deviate significantly from the norm.
- The use of PCA reduces dimensionality and removes correlations, making the distance calculation more robust and meaningful.

Score 2 (Autoencoder-based Reconstruction Error):

- Autoencoders are trained to minimize reconstruction error for normal data patterns. Hence, anomalies, which deviate from normal patterns, will have higher reconstruction errors.
- This method captures non-linear relationships in the data, providing a powerful way to detect complex anomalies that may not be evident in the original feature space.
- It complements the Minkowski distance score by focusing on the difficulty of recreating unusual data patterns.

How to Scale and Combine

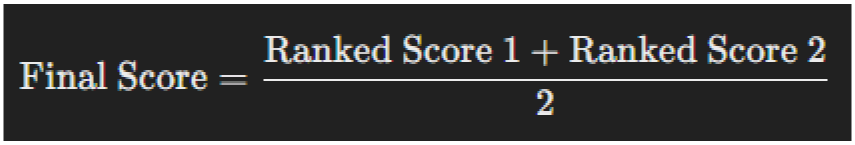
To combine the two scores into a single fraud score:

1. Rank Transformation:

First, transform each score into rank order. This step is essential to ensure that both scores contribute equally, irrespective of their scale.

2. Averaging the Ranks:

Average the rank-ordered scores to get a final combined score.Average the rank-ordered scores to get a final combined score.



$$\text{Final Score} = \frac{\text{Ranked Score 1} + \text{Ranked Score 2}}{2}$$

7. Results:

Using the Fraud Scores:

Combining and Using the Scores:

The two anomaly scores (Score 1 and Score 2) were calculated for each property.
Score 1 was based on the Minkowski distance in the principal component space.
Score 2 was based on the reconstruction error from the autoencoder.
Each score was then transformed into a rank order to ensure comparability, and these ranks were averaged to produce a final combined fraud score.

Excel Sheet Compilation:

An Excel sheet was created containing all the original fields from the property dataset along with the final combined fraud scores.
This sheet allows easy sorting and filtering to identify properties with the highest likelihood of tax fraud.

Examining Properties:

Investigative Process:

Reviewing High Fraud Scores:

Properties were examined by sorting the dataset by the final fraud score in descending order.
The top-ranked properties, which have the highest scores, were flagged for further investigation.

Field Analysis and Validation:

For each flagged property, the individual scores (Score 1 and Score 2) and their component values were reviewed to identify which specific fields or behaviors contributed to the high score.

Cross-Referencing with External Data:

Google Maps was used to visually inspect properties that appeared suspicious based on their fraud scores.

Unusual patterns such as discrepancies between reported and actual property usage, signs of underreporting, or other anomalies were noted.

Google Maps Inspection involved checking the physical state(Lot area, total area, number of stories, building length and width) of the property against reported data.

Example of Five Unusual cases:

1)

RECORD: 700779
OWNER NAME: DEMAREST, ARTHUR
Building Class: V0
FULLVAL: 93,288
AVLAND: 10
AVTOT: 10
Address: 11 AVENUE
ZIP: 11357

Big Problem: Wrong Address (It's a street or freeway address)

- The data provided for this RECORD shows highly unusual property values.

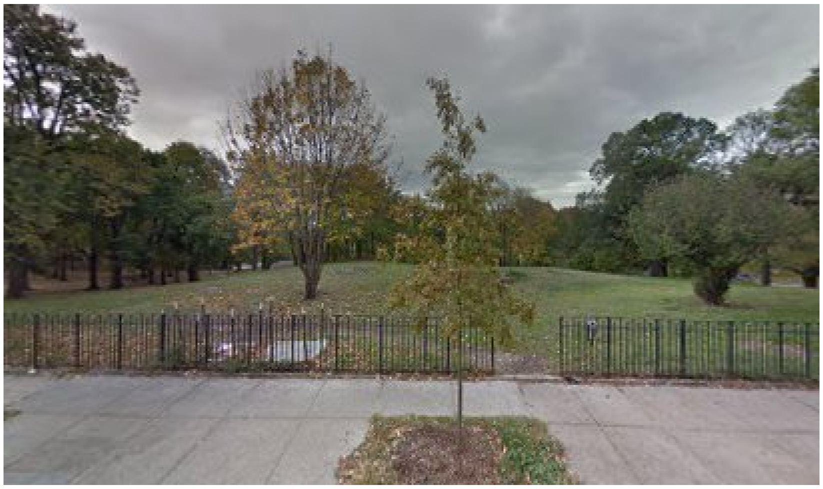
- The AVLAND and AVTOT are significantly low ($10 each), leading to several ratio calculations being zero(r1, r2, r4, r5, r7, and r8). The non-zero ratios involving building volume (r3, r6, r9) suggest some value correlation.
- The ZIP code and tax class ratios further highlight this property as an outlier with unusually low assessments in certain categories.
- Additionally, the incorrect address ("11 AVENUE 11357") raises concerns about data accuracy and entry errors.
- This requires further investigation.

2)
RECORD: 536544
OWNER NAME: JEANTY JOSEPH P
LTFRONT: 15
LTDEPTH: 97
Building Class: V0
FULLVAL: $146,960
AVLAND: $10
AVTOT: $10
Address: 1952 TROY AVENUE
ZIP: 11234

Big Problem: The actual property frontage (LTFRONT) appears smaller when viewed on Google Maps compared to the recorded value.

- The data provided for this record shows highly unusual property values.
- The AVLAND and AVTOT are significantly low ($10 each), leading to several ratio calculations being zero. The non-zero ratios involving building volume (r3, r6, r9) suggest some value correlation.
- The ZIP code and tax class ratios further highlight this property as an outlier with unusually low assessments in certain categories.
- Additionally, the discrepancy in the recorded lot frontage indicates potential errors in the lot dimensions.
- This requires further investigation.

3)
RECORD: 161711
OWNER NAME: DEIRA REALTY CORP.
LTFRONT: 140
LTDEPTH: 0
Building Class: V1
FULLVAL: $1,050
AVLAND: $473
AVTOT: $473
Address: EAST 174 STREET
ZIP: 10457



Big Problem: LTDEPTH not mentioned.

- The data provided for this record shows highly unusual property values.
- The AVLAND and AVTOT are significantly low ($473 each), leading to several ratio calculations being zero or negative.
- The non-zero ratios involving building volume (r3, r6, r9) suggest some value correlation.
- Zero building dimensions (BLDFRONT, BLDDEPTH)
- Additionally, the discrepancy in the recorded lot dimensions, with LTDEPTH being 0 and LTFRONT at 140, indicates potential errors in the lot dimensions. This requires further investigation to ensure accurate and complete data collection and property assessment.

4)
RECORD: 848747
OWNER NAME: 128 STREET CORP.
LTFRONT: 9170
LTDEPTH: 118
Building Class: V0
FULLVAL: $285,000
AVLAND: $17,100
AVTOT: $17,100
Address: 110-41 FARMERS BOULEVARD
ZIP: 11412

Big Problem: The high LTFRONT value.

- The data provided for this record shows highly unusual property values.
- The AVLAND and AVTOT are significantly low ($17,100 each), leading to several ratio calculations being zero or negative.
- The non-zero ratios involving building volume (r3, r6, r9) suggest some value correlation.
- Additionally, the discrepancy in the recorded lot dimensions, with an unusually high LTFRONT of 9170, indicates potential errors in the lot dimensions. This requires further investigation to ensure accurate and complete data collection and property assessment.

5)
RECORD: 167787
OWNER NAME: JACKSON, BELGICA
LTFRONT: 66
LTDEPTH: 25
Building Class: V0
FULLVAL: $125,025
AVLAND: $10
AVTOT: $10
Address: HONEYWELL AVENUE
ZIP: Not mentioned

A random closest house picture from HONEYWELL AVENUE:

Big Problem: The significantly low AVLAND and AVTOT values, and ZIP code not mentioned.

- The data provided for this record shows highly unusual property values.
- The AVLAND and AVTOT are significantly low ($10 each), leading to several ratio calculations being zero or negative.
- The non-zero ratios involving building volume (r3, r6, r9) suggest some value correlation.
- Additionally, the absence of a ZIP code indicates incomplete address information. These discrepancies require further investigation to ensure accurate and complete data collection and property assessment.

8. Summary

The New York Property Tax Fraud Detection project involved several critical steps to identify potential tax fraud in NYC's property assessment data. We started with a thorough data cleaning process to ensure all relevant fields were accurate and complete. Next, we created new variables to enhance the analysis and conducted dimensionality reduction using Principal Component Analysis (PCA) to efficiently handle the high-dimensional data.

The core of our fraud detection relied on two scoring algorithms. The first algorithm calculated a fraud score based on the Minkowski distance in the PCA-transformed data, while the second algorithm calculated a score based on reconstruction errors from PCA. These scores were then combined to generate a final fraud score for each property.

Key results include:

High Fraud Scores: Properties with high fraud scores were flagged for further investigation. These scores indicated significant deviations from typical property characteristics.
Inconsistent Valuation Metrics: Several properties had valuation metrics that did not align with their physical characteristics or reported usage, indicating possible misreporting.
Unusual Size and Value Ratios: We identified anomalies in the size and value ratios of properties, suggesting discrepancies that could indicate fraud.
Significant Irregularities: Detailed case studies of the top flagged properties revealed substantial irregularities, justifying further investigation.

To use the scores, stakeholders can refer to the Excel sheet containing all property fields and their corresponding fraud scores. High scores indicate properties that require closer scrutiny. For flagged properties, investigating fields with unusually high r-values and cross-referencing these properties using tools like Google Maps can help verify any discrepancies.

How to Improve Based on Client Feedback

The algorithm can be refined based on expert feedback by modifying parameters in the scoring algorithms or excluding certain variables that may not significantly contribute to fraud detection. For continuous improvement, it is essential to:

1. **Adjust Parameters:** Fine-tune the parameters of the anomaly detection algorithms to better capture fraudulent patterns as more data and insights become available.
2. **Exclude Irrelevant Variables:** Remove variables that do not provide meaningful information or contribute to false positives, based on expert analysis.
3. **Incorporate New Data:** Regularly update the model with new data to capture evolving patterns of fraud.
4. **Client Feedback Loop:** Establish a feedback loop with city authorities and other stakeholders to continuously improve the model's accuracy and relevance.

By implementing these adjustments, the model can remain robust and adaptive, ensuring ongoing effectiveness in detecting potential tax fraud in New York City's property data.

9. Appendix

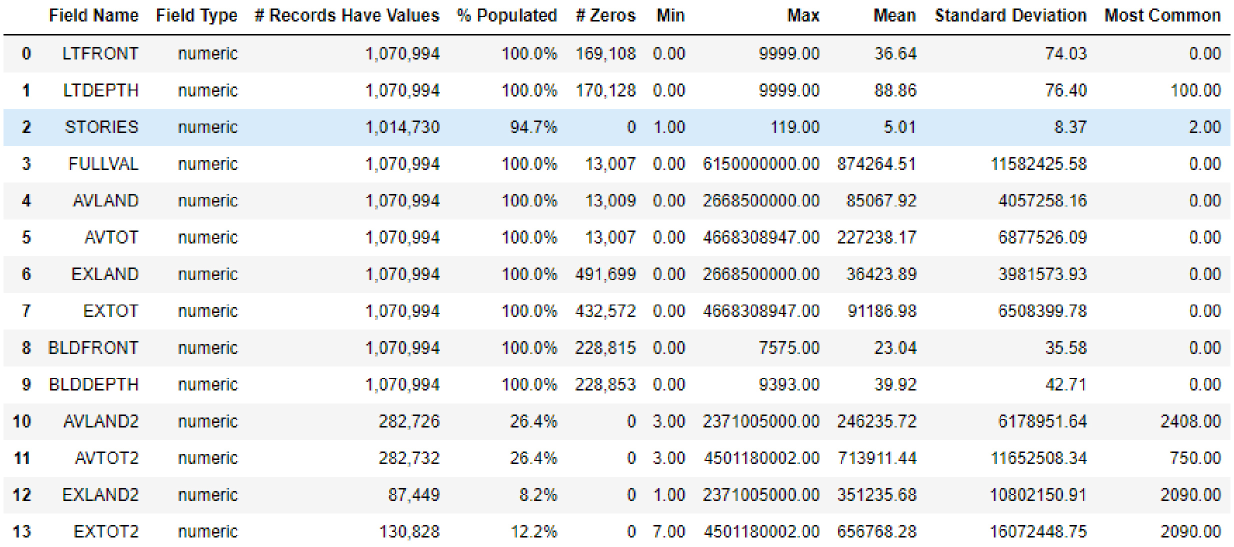
Data Quality Report

1. Data Description

The dataset "Property Valuation and Assessment Data," comprises **1,070,994 records** detailing real estate assessment property data for the **year 2010/11**. It encompasses **32 fields** and is updated annually by the Department of Finance. This dataset provides an extensive collection of information related to property assessments, evaluations, and condensed roll data within the city government category. Key fields include property ID, assessment value, property type, location, and other relevant property-specific information.

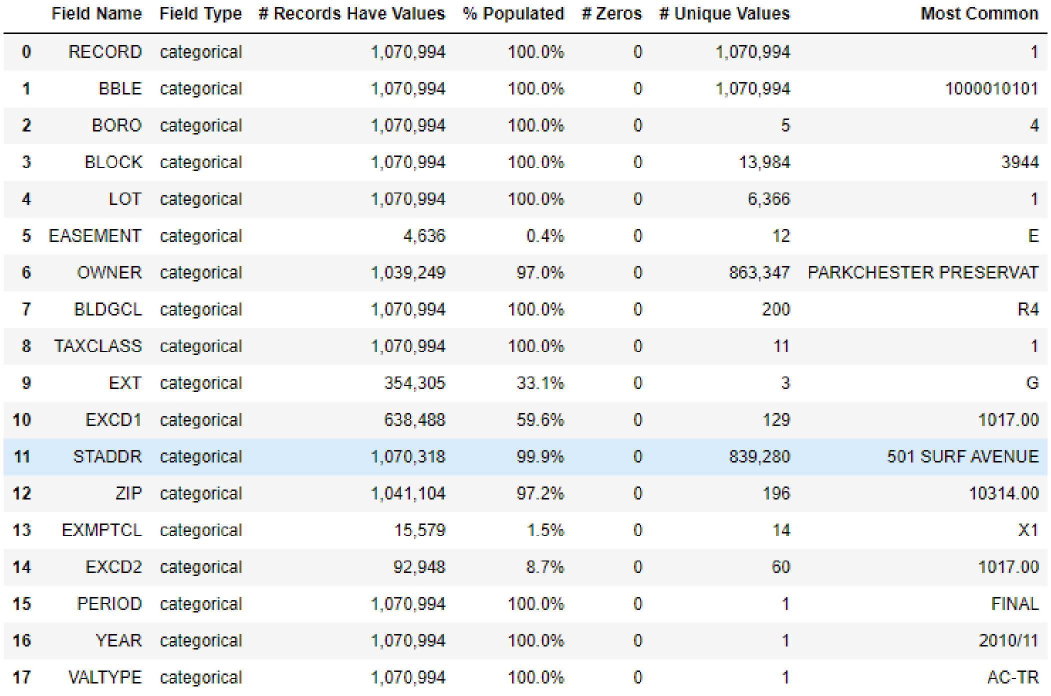
2. Summary Tables:

Numeric:



	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	LTFRONT	numeric	1,070,994	100.0%	169,108	0.00	9999.00	36.64	74.03	0.00
1	LTDEPTH	numeric	1,070,994	100.0%	170,128	0.00	9999.00	88.86	76.40	100.00
2	STORIES	numeric	1,014,730	94.7%	0	1.00	119.00	5.01	8.37	2.00
3	FULLVAL	numeric	1,070,994	100.0%	13,007	0.00	6150000000.00	874264.51	11582425.58	0.00
4	AVLAND	numeric	1,070,994	100.0%	13,009	0.00	2668500000.00	85067.92	4057258.16	0.00
5	AVTOT	numeric	1,070,994	100.0%	13,007	0.00	4668308947.00	227238.17	6877526.09	0.00
6	EXLAND	numeric	1,070,994	100.0%	491,699	0.00	2668500000.00	36423.89	3981573.93	0.00
7	EXTOT	numeric	1,070,994	100.0%	432,572	0.00	4668308947.00	91186.98	6508399.78	0.00
8	BLDFRONT	numeric	1,070,994	100.0%	228,815	0.00	7575.00	23.04	35.58	0.00
9	BLDDEPTH	numeric	1,070,994	100.0%	228,853	0.00	9393.00	39.92	42.71	0.00
10	AVLAND2	numeric	282,726	26.4%	0	3.00	2371005000.00	246235.72	6178951.64	2408.00
11	AVTOT2	numeric	282,732	26.4%	0	3.00	4501180002.00	713911.44	11652508.34	750.00
12	EXLAND2	numeric	87,449	8.2%	0	1.00	2371005000.00	351235.68	10802150.91	2090.00
13	EXTOT2	numeric	130,828	12.2%	0	7.00	4501180002.00	656768.28	16072448.75	2090.00

Categorical:



	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	RECORD	categorical	1,070,994	100.0%	0	1,070,994	1
1	BBLE	categorical	1,070,994	100.0%	0	1,070,994	1000010101
2	BORO	categorical	1,070,994	100.0%	0	5	4
3	BLOCK	categorical	1,070,994	100.0%	0	13,984	3944
4	LOT	categorical	1,070,994	100.0%	0	6,366	1
5	EASEMENT	categorical	4,636	0.4%	0	12	E
6	OWNER	categorical	1,039,249	97.0%	0	863,347	PARKCHESTER PRESERVAT
7	BLDGCL	categorical	1,070,994	100.0%	0	200	R4
8	TAXCLASS	categorical	1,070,994	100.0%	0	11	1
9	EXT	categorical	354,305	33.1%	0	3	G
10	EXCD1	categorical	638,488	59.6%	0	129	1017.00
11	STADDR	categorical	1,070,318	99.9%	0	839,280	501 SURF AVENUE
12	ZIP	categorical	1,041,104	97.2%	0	196	10314.00
13	EXMPTCL	categorical	15,579	1.5%	0	14	X1
14	EXCD2	categorical	92,948	8.7%	0	60	1017.00
15	PERIOD	categorical	1,070,994	100.0%	0	1	FINAL
16	YEAR	categorical	1,070,994	100.0%	0	1	2010/11
17	VALTYPE	categorical	1,070,994	100.0%	0	1	AC-TR

3. Fields:

RECORD:

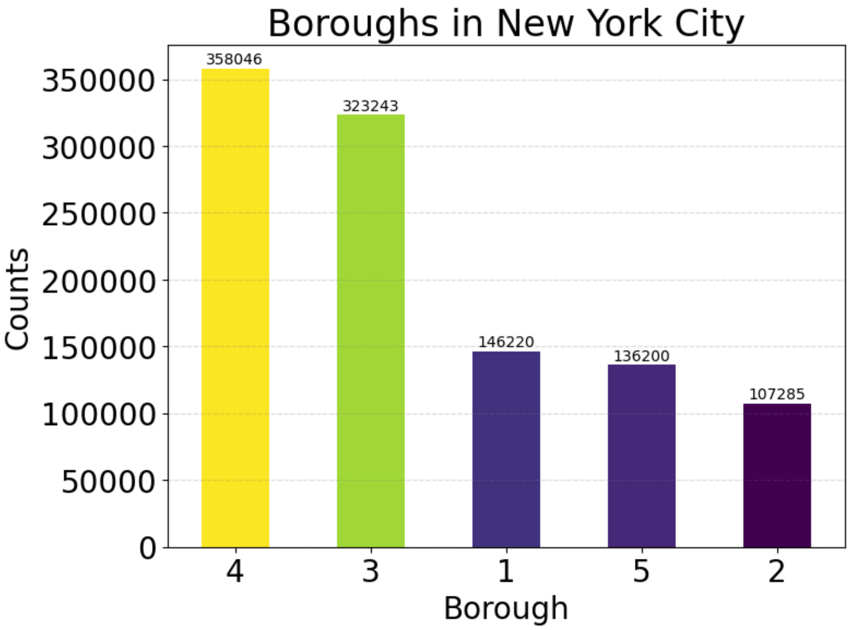
The "RECORD" field is a categorical field that uniquely identifies each of the 1,070,994 records in the dataset, with 100% population and no duplicate values.

BBLE:

The "BBLE" field is a categorical field that uniquely identifies each record as a file key, with all 1,070,994 records populated and no duplicate values.

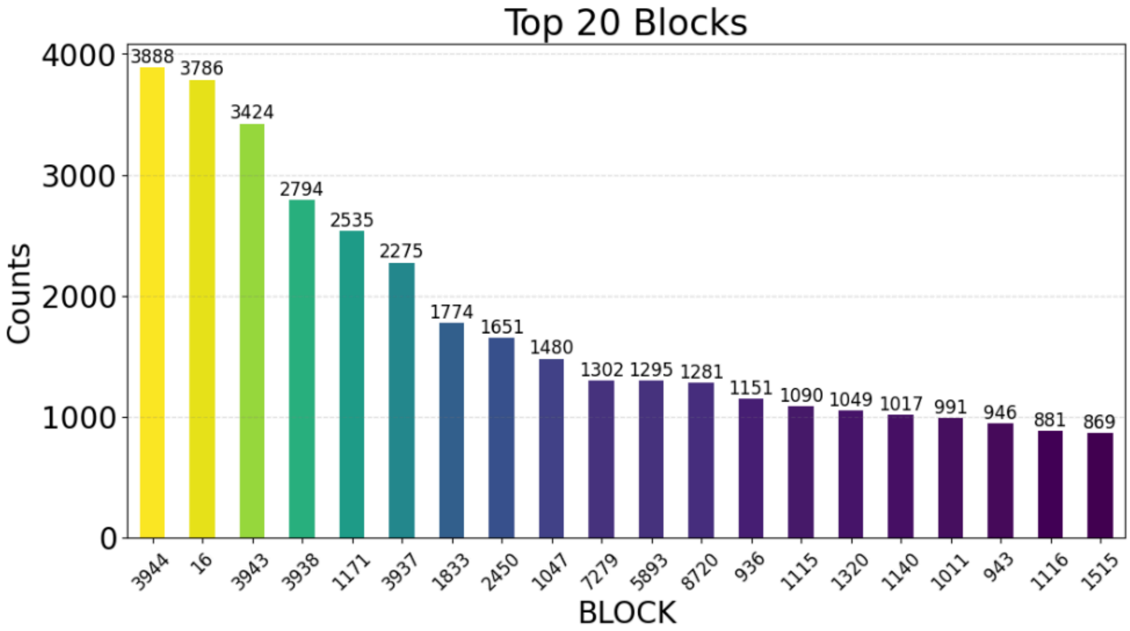
BORO:

The "BORO" field is a categorical field representing the boroughs of NYC, with all 1,070,994 records populated and no zeros. There are 5 unique values, with the most common being "4." The following graph shows the visualization of boroughs in NYC.



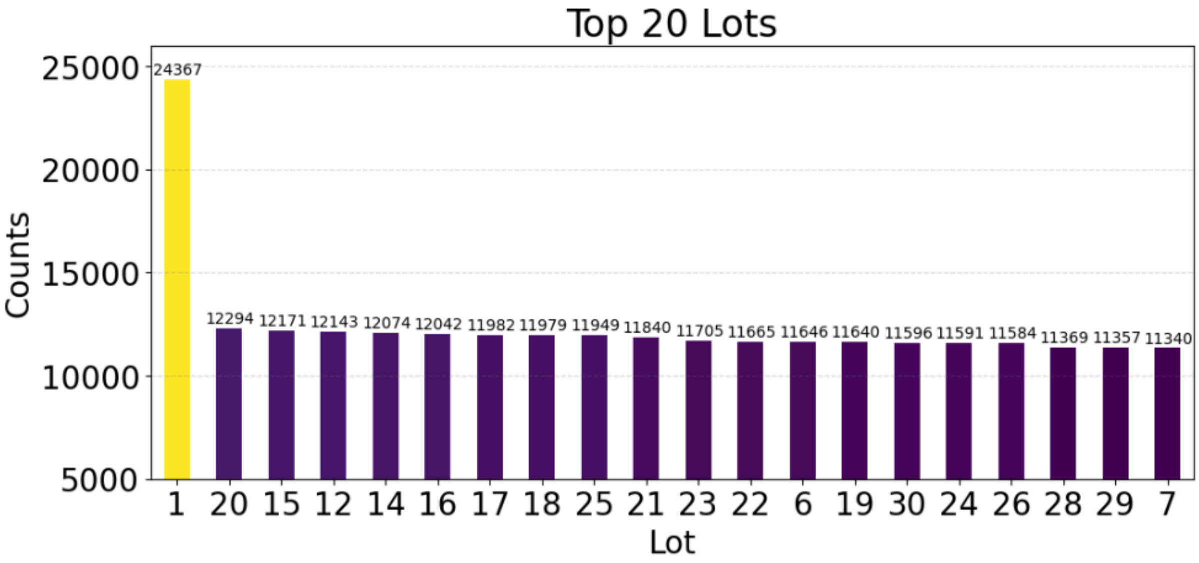
BLOCK:

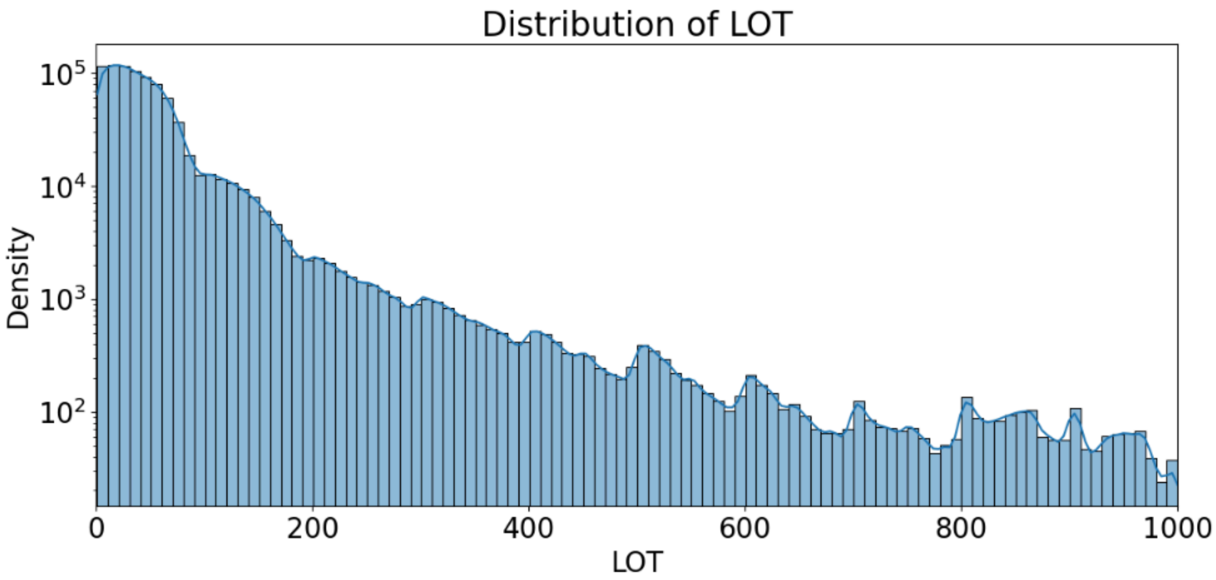
The "BLOCK" field is a categorical field representing valid block ranges by borough, with all 1,070,994 records populated and no zeros. There are 13,984 unique values, with the most common being "3944." The following graph visualizes the top 20 blocks.



LOT:

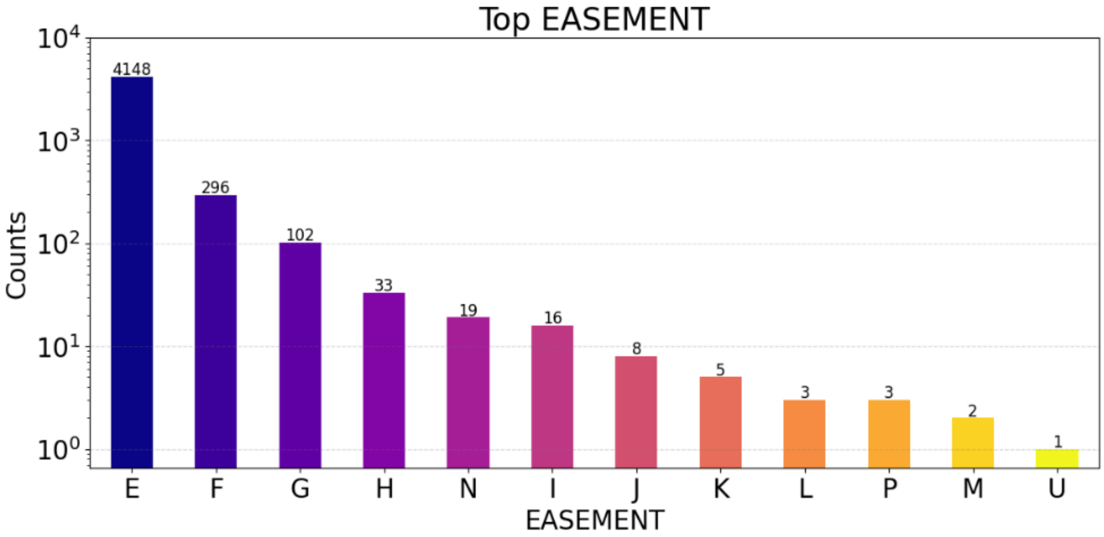
The "LOT" field, a categorical identifier with 100% coverage across 1,070,994 records and zero missing values, features 6,366 distinct categories. Notably, the most prevalent category is "1." Additionally, visualizations depicting the top 20 lots and the distribution of lots across the dataset are available for analysis.




Distribution of LOT
10^5
10^4
10^3
10^2
Density
0
200
400
600
800
1000
LOT

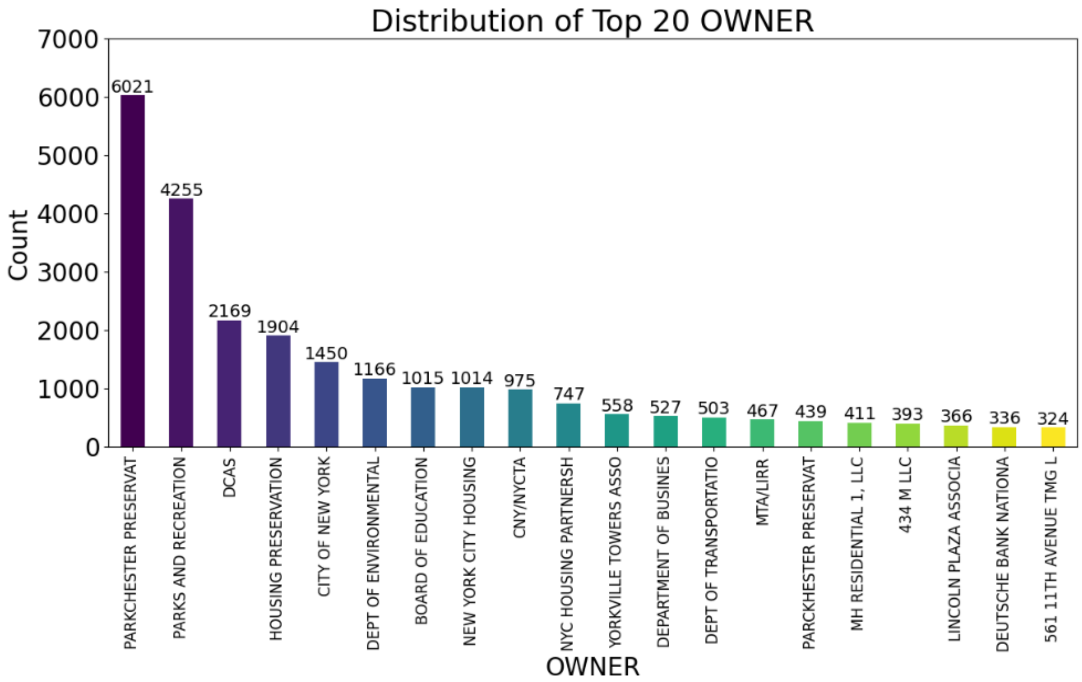
EASEMENT:

The "EASEMENT" field is categorical, present in 0.4% of the 1,070,994 records, with no instances of zero values. It comprises 12 unique categories, with "E" being the most common. An accompanying visualization showcases the distribution of the top easements within the dataset.



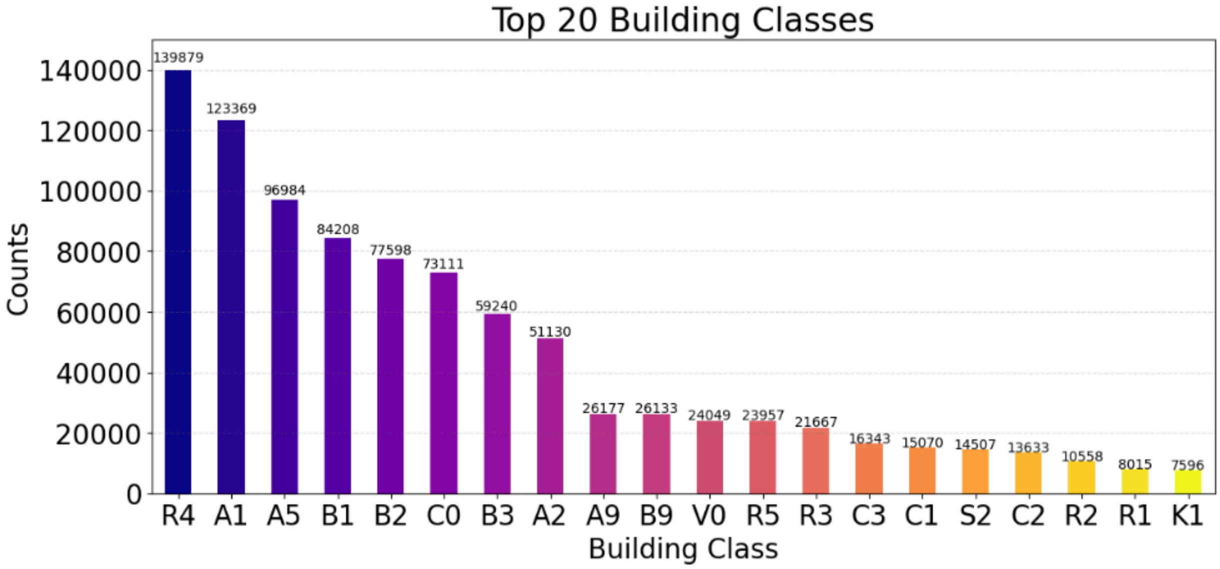
OWNER:

The "OWNER" field, capturing categorical data representing owner names, is prevalent in 97.0% of the 1,070,994 records, with no zero entries. It encompasses a wide diversity of 863,347 unique owners, with "PARKCHESTER PRESERVAT" being the most frequent. A visual representation highlights the top 20 owners for further analysis.



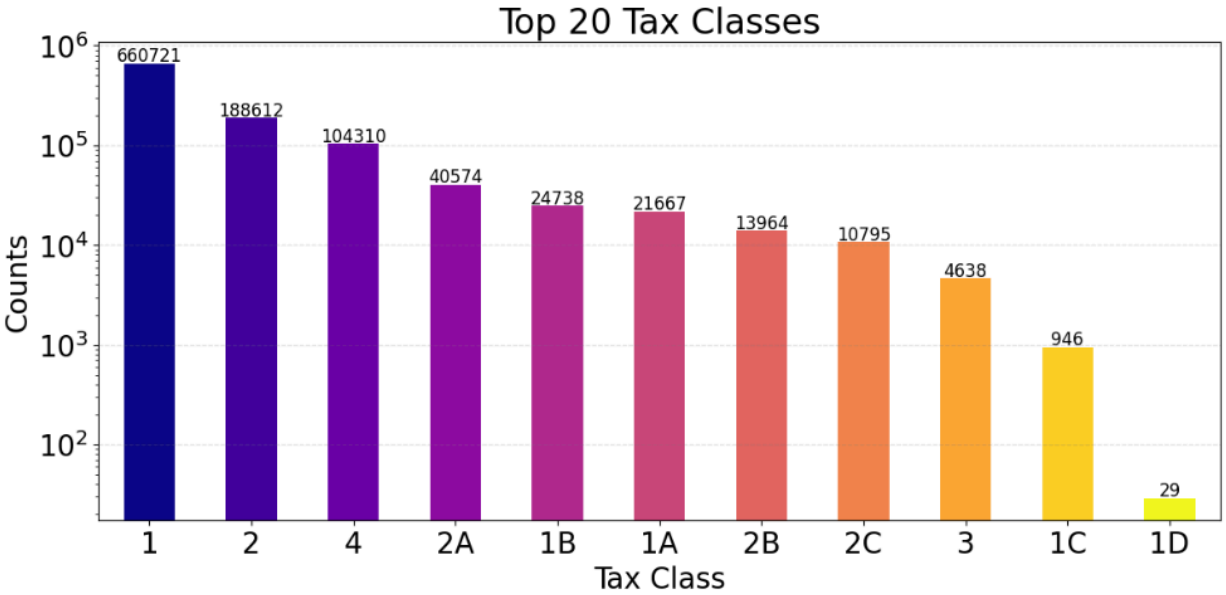
BLDGCL:

The "BLDGCL" field, categorical and present in all 1,070,994 records without any zeros, represents building classes. With 200 unique categories, "R4" emerges as the most common class. A visual overview displays the top 20 building classes.



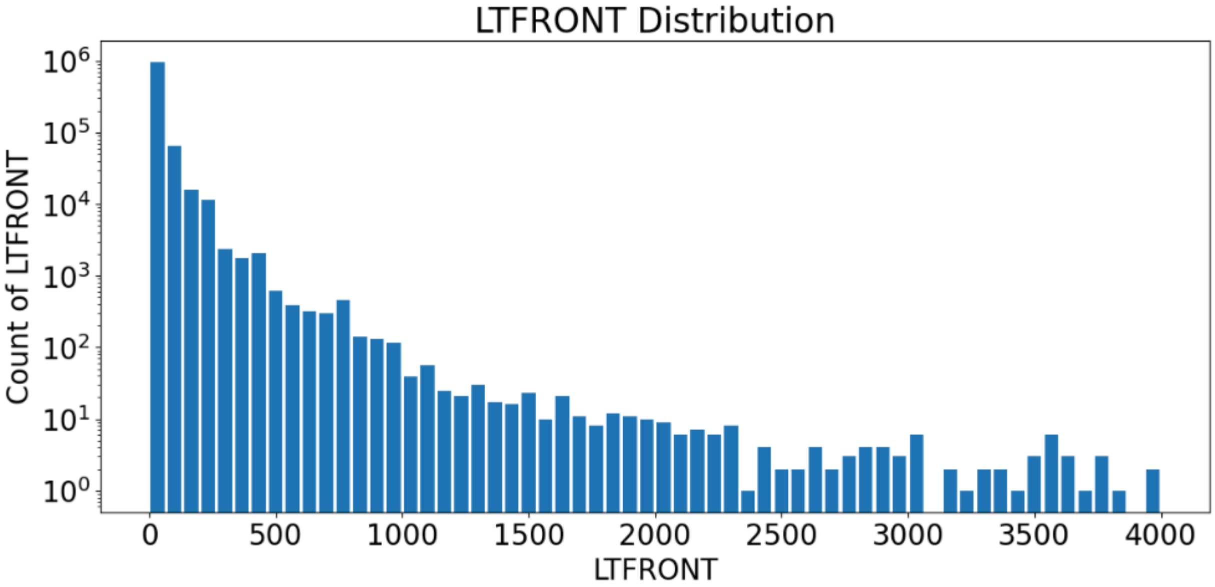
TAXCLASS:

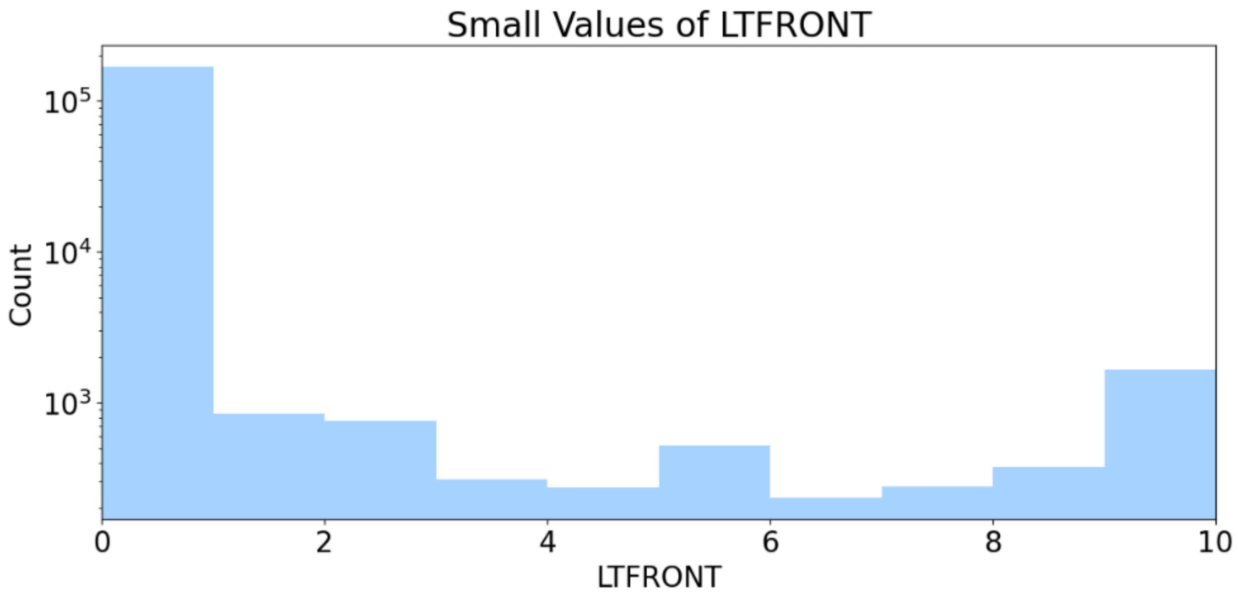
The "TAXCLASS" field, categorical and fully populated across all 1,070,994 records without any zeros, categorizes tax classes. With 11 unique classes, "1" is the most prevalent. A visual representation illustrates the distribution of the top 20 tax classes.



LTFRONT:

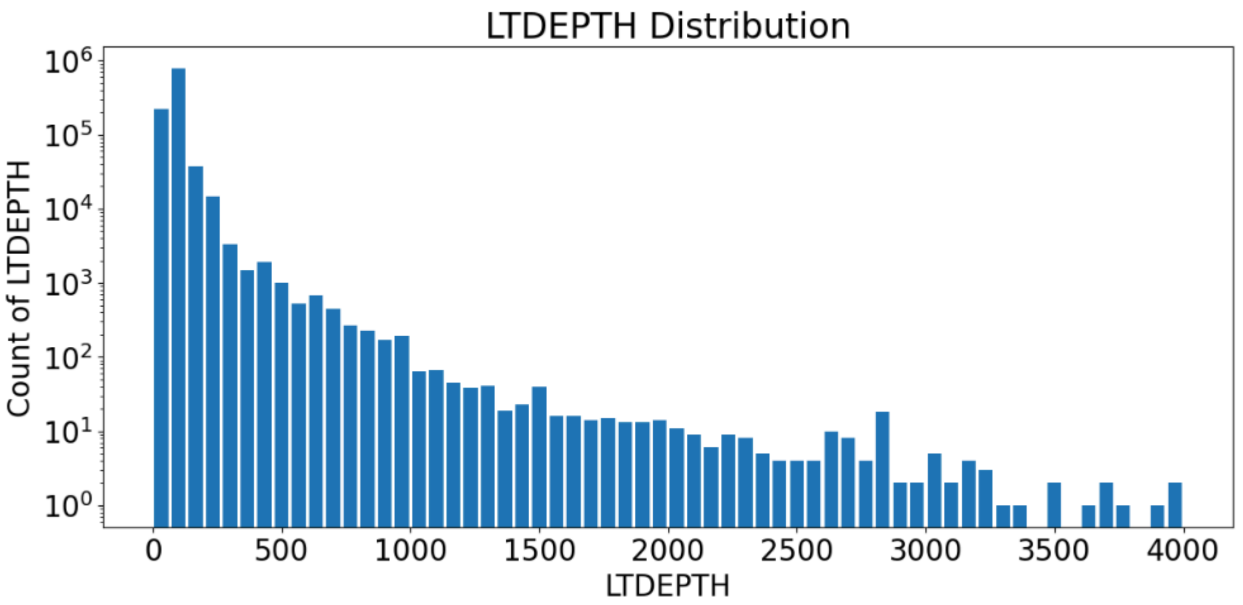
The "LTFRONT" field, a numeric attribute, is fully populated across 1,070,994 records. It denotes lot width, ranging from 0.00 to 9999.00, with an average width of 36.64 and a median of 74.03. Visualizations include distributions of widths and a focused one on smaller values.

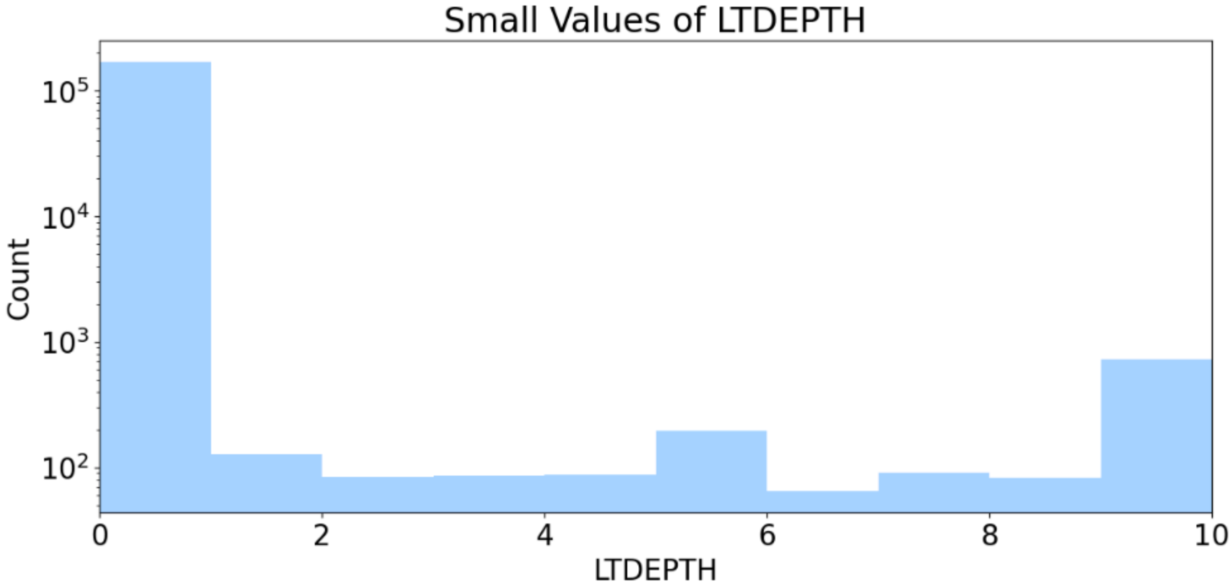




LTDEPTH:

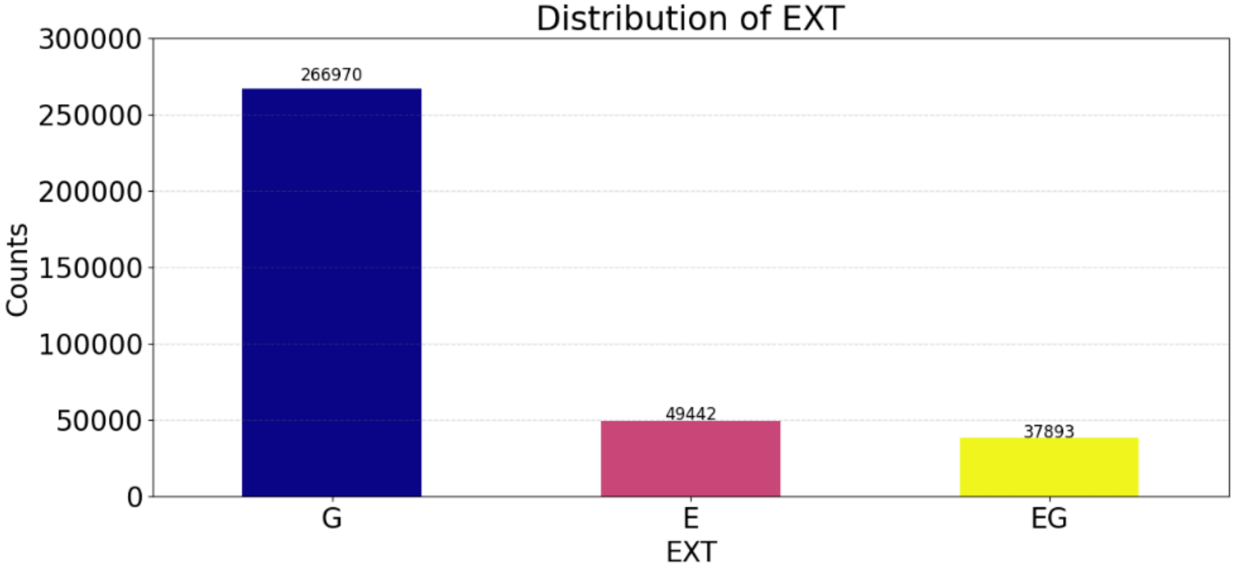
The "LTDEPTH" field, a numeric attribute, is fully populated across 1,070,994 records. It represents lot depth, ranging from 0.00 to 9999.00, with an average depth of 88.86 and a median of 76.40. Visualizations encompass distributions of depths and focus on the distribution of smaller values.





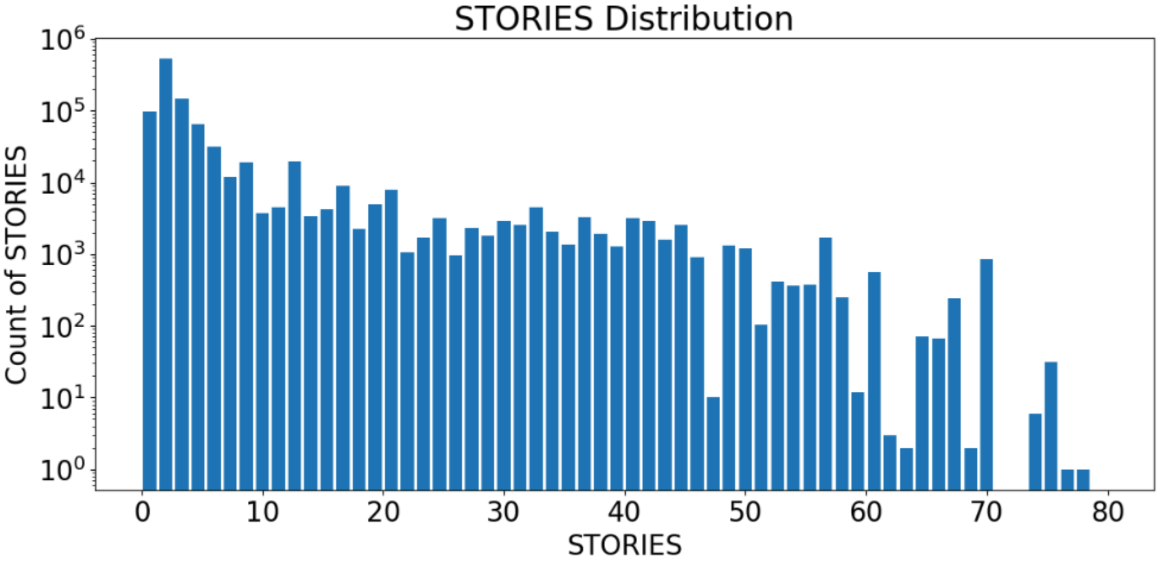
EXT:

The "EXT" field, categorical with 33.1% coverage across 354,305 records, denotes extension indicators with three unique categories. The visualization includes a distribution analysis of these extension indicators.



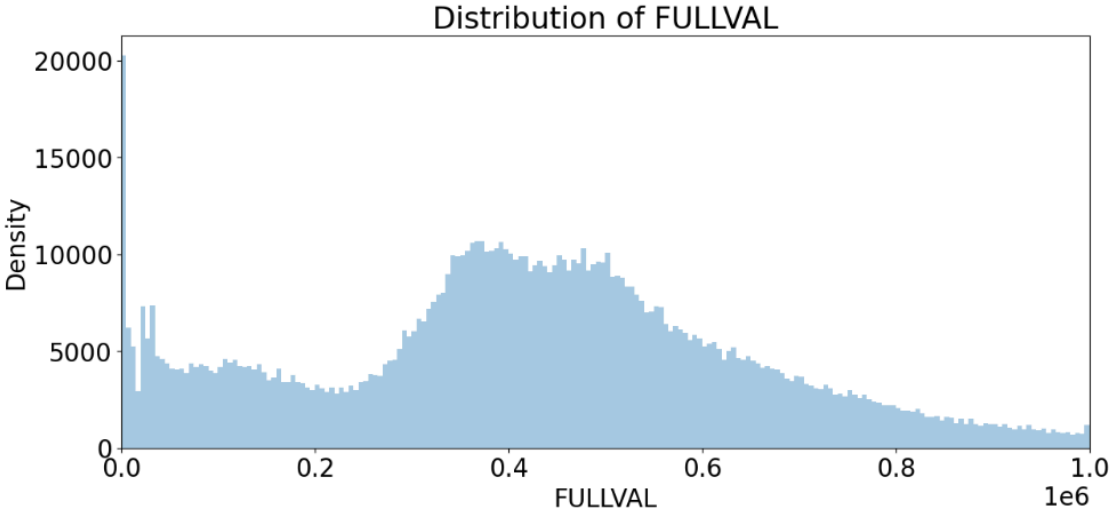
STORIES:

The "STORIES" field, a numeric attribute, is present in 94.7% of the 1,014,730 records, indicating the number of stories in a building. With values ranging from 1.00 to 119.00, the average number of stories is 5.01, with a median of 8.37. Visualizations encompass the distribution of stories within buildings.



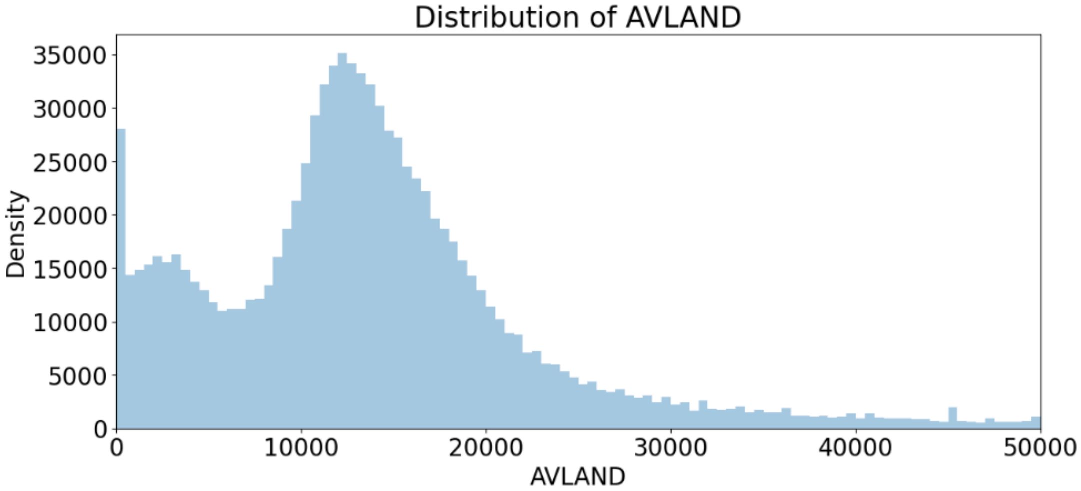
FULLVAL:

The "FULLVAL" field, a numeric attribute, is fully populated across 1,070,994 records, representing the market value. The values range from 0.00 to 6,150,000,000.00, with an average value of $874,264.51 and a median of $11,582,425.58. Visualizations showcase the distribution of market values (FULLVAL).



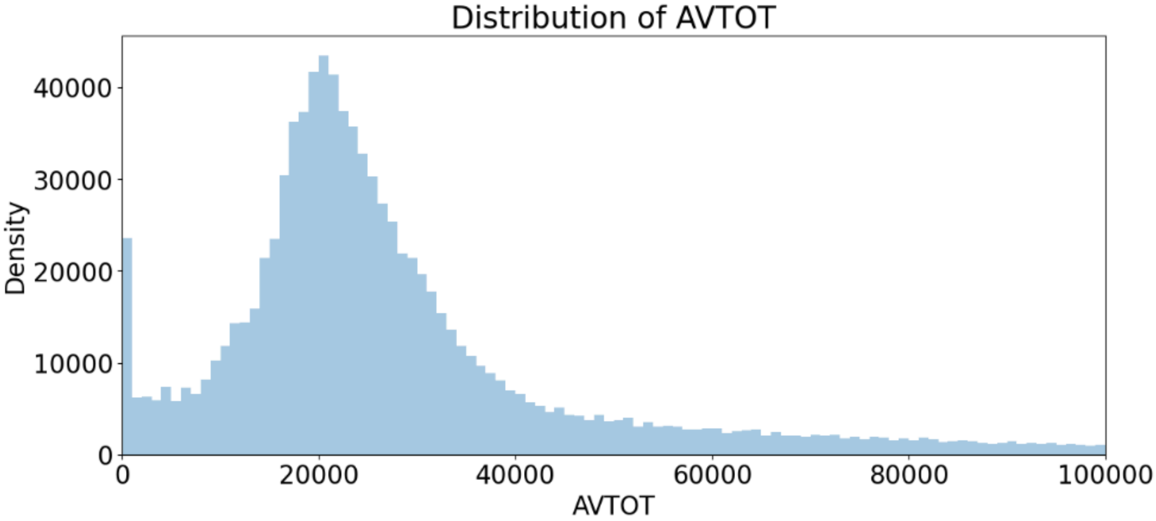
AVLAND:

The "AVLAND" field, a numeric attribute present in all 1,070,994 records, represents the actual land value. Values range from 0.00 to 2,668,500,000.00, with an average value of $85,067.92 and a median of $4,057,258.16. Visualizations illustrate the distribution of actual land values (AVLAND).



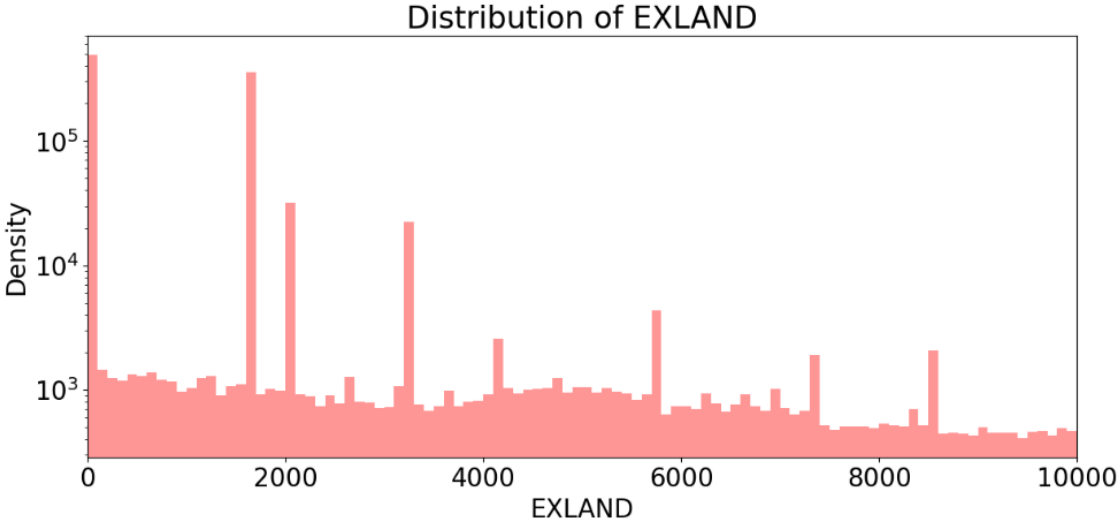
AVTOT:

The "AVTOT" field, a numeric attribute present in all 1,070,994 records, represents the actual total value. Values range from 0.00 to 4,668,308,947.00, with an average value of $227,238.17 and a median of $6,877,526.09. Visualizations illustrate the distribution of actual total values (AVTOT).



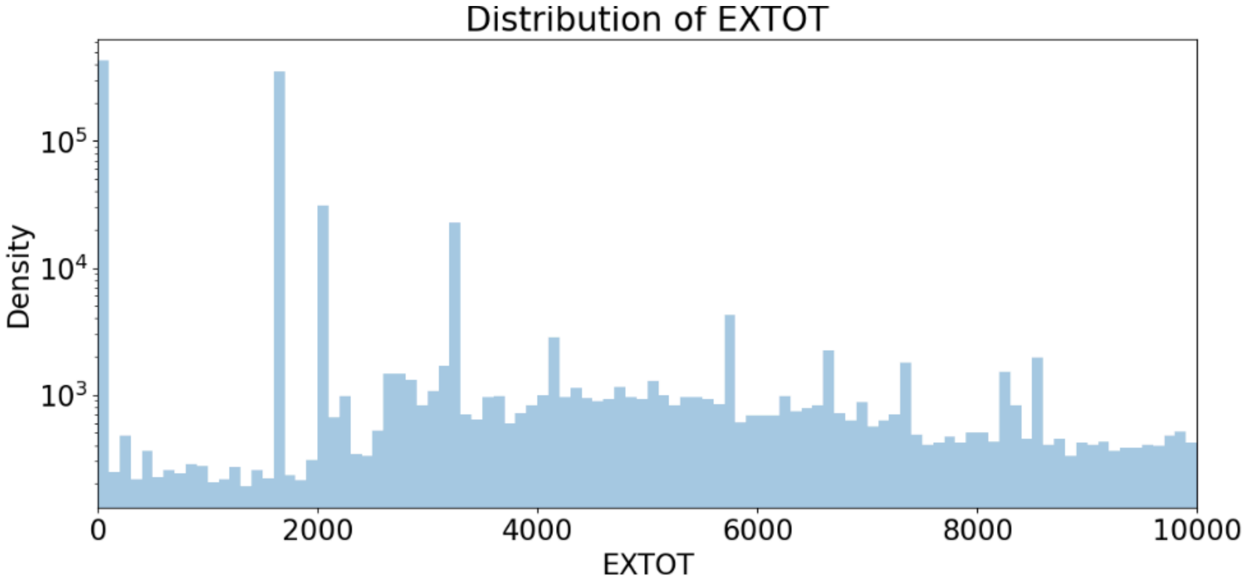
EXLAND:

The "EXLAND" field, a numeric attribute present in all 1,070,994 records, represents the actual exempt land value. Values range from 0.00 to 2,668,500,000.00, with an average value of $36,423.89 and a median of $3,981,573.93. Visualizations illustrate the distribution of actual exempt land values (EXLAND).



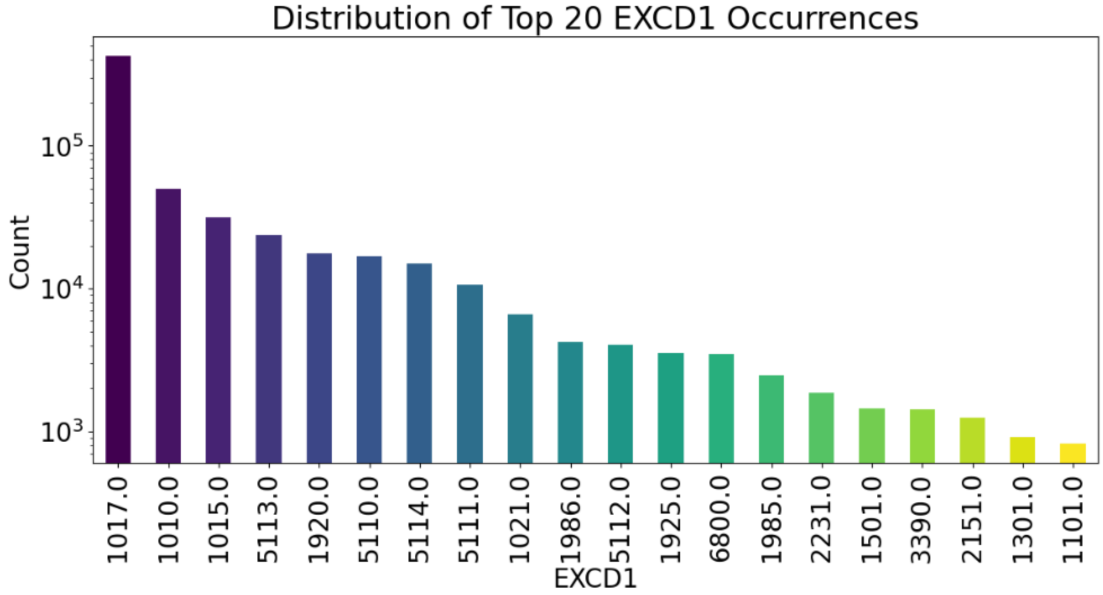
EXTOT:

The "EXTOT" field, a numeric attribute present in all 1,070,994 records, denotes the actual exempt land total. Values range from 0.00 to 4,668,308,947.00, with an average value of $91,186.98 and a median of $6,508,399.78. Visualizations depict the distribution of actual exempt land totals (EXTOT).



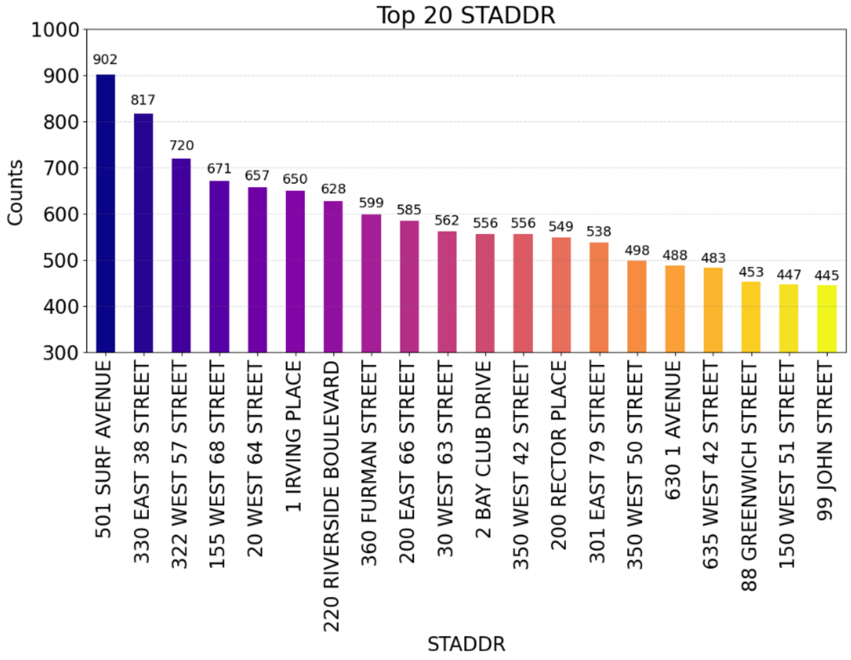
EXCD1:

The "EXCD1" field, a categorical attribute present in 59.6% of the 638,488 records, denotes exemption codes. There are 129 unique codes, with the top 20 occurrences visualized for analysis.



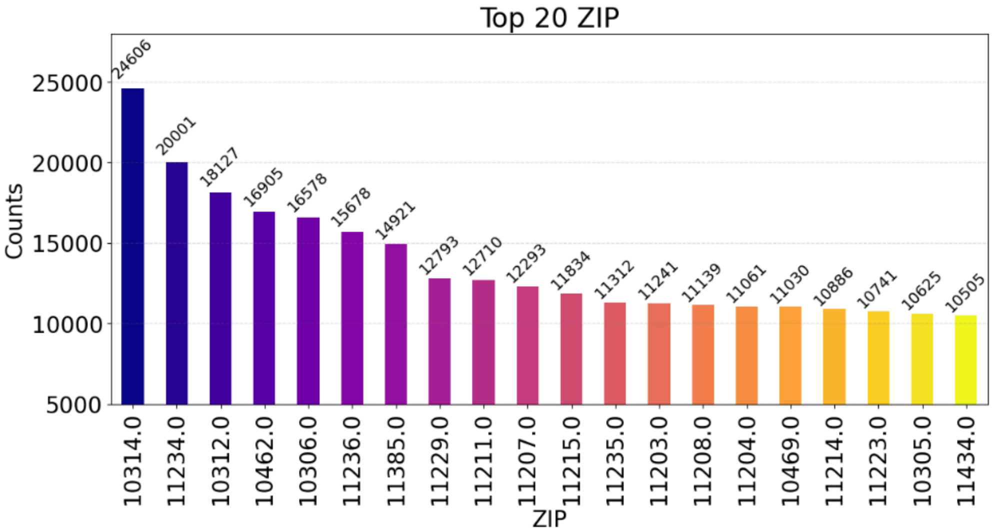
STADDR:

The "STADDR" field, a categorical attribute, is present in 99.9% of the 1,070,318records, representing street addresses. With 839,280 unique addresses, the top 20 occurrences are visualized for analysis.



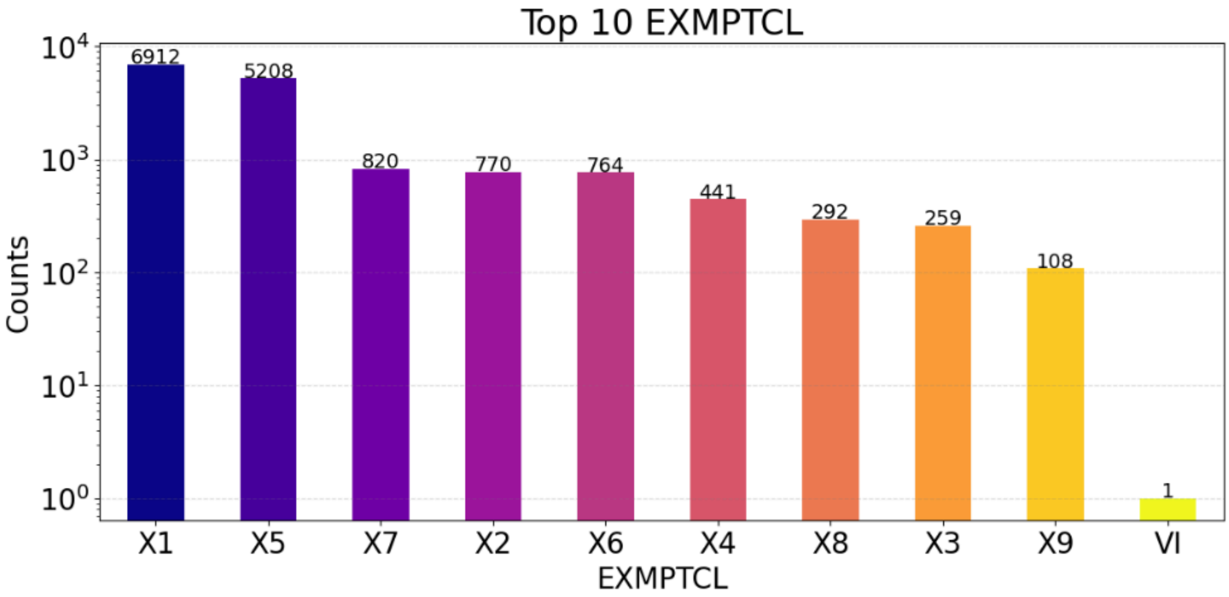
ZIP:

The "ZIP" field, a categorical attribute, is present in 97.2% of the 1,041,104 records, indicating zip codes. With 196 unique zip codes, the top 20 occurrences are visualized for analysis.



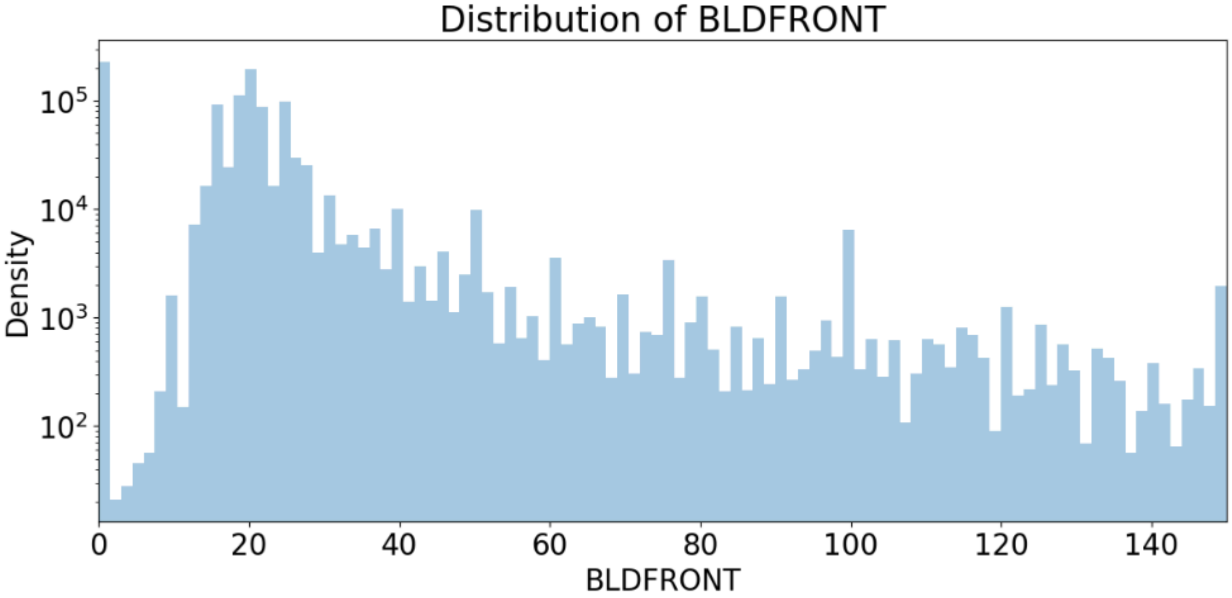
EXMPTCL:

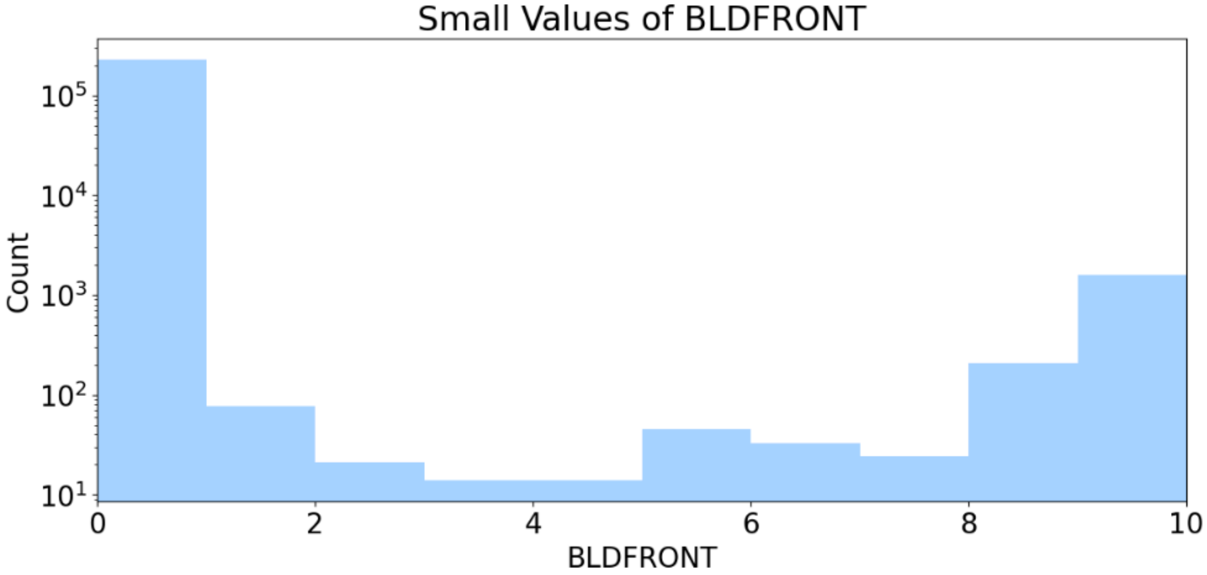
The "EXMPTCL" field, a categorical attribute, is present in 1.5% of the 15,579 records, representing exemption classes. With 14 unique classes, The visual representation showcases the top 10 occurrences for the "EXMPTCL" field.



BLDFRONT:

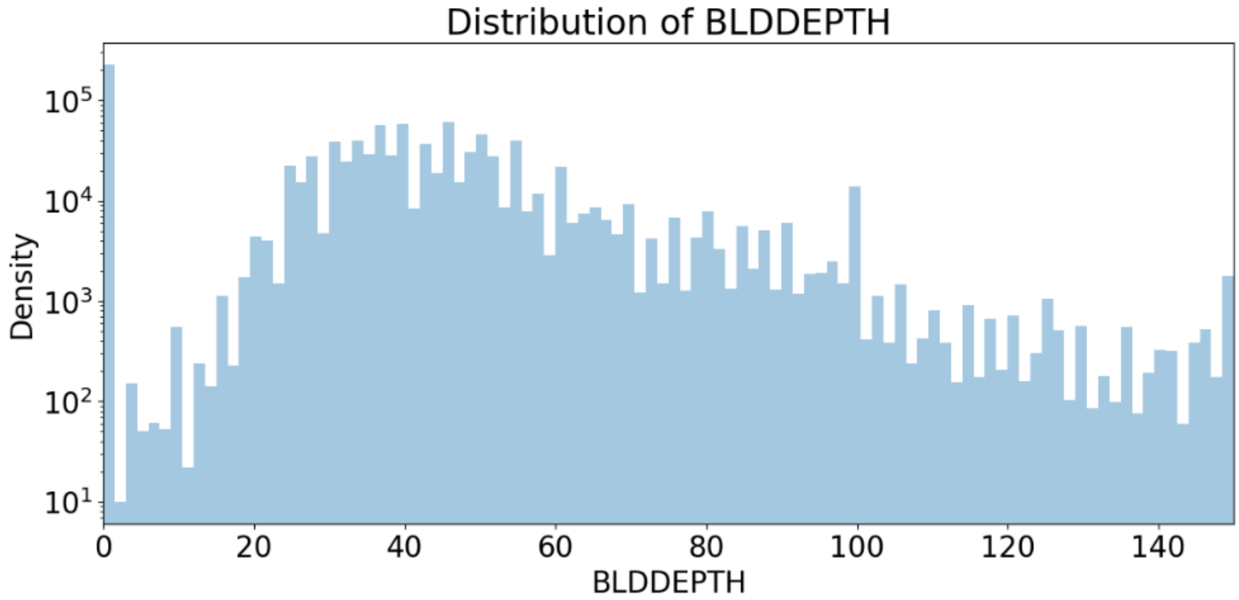
The "BLDFRONT" field, a numeric attribute, is fully populated across 1,070,994 records, indicating the building width. Ranging from 0.00 to 7575.00, with an average width of 23.04 and a median of 35.58, visualizations encompass distributions of building widths (BLDFRONT) and focus on the distribution of smaller values.

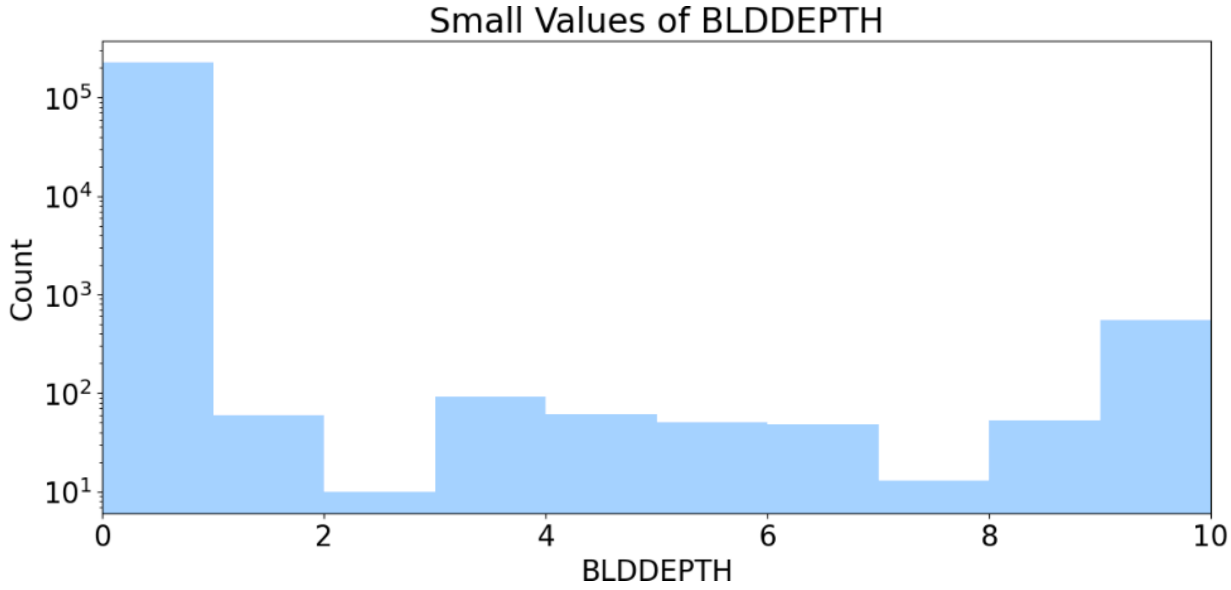




BLDDEPTH:

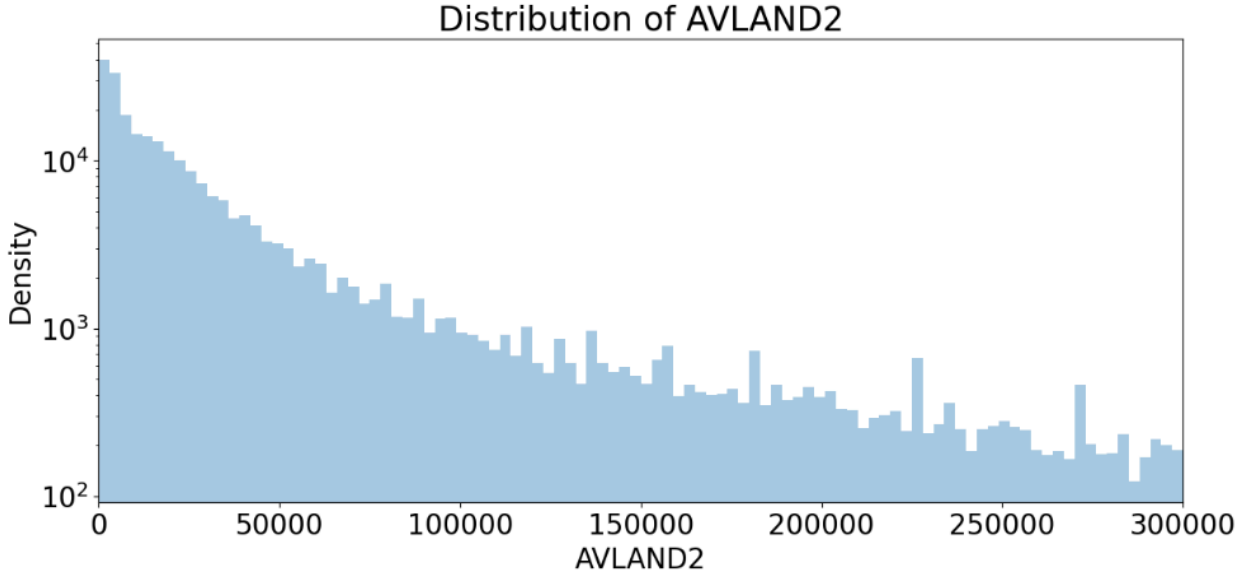
The "BLDDEPTH" field, a numeric attribute, is fully populated across all 1,070,994 records, representing building depth. Ranging from 0.00 to 9393.00, with an average depth of 39.92 and a median of 42.71, visualizations showcase the distribution of building depths (BLDDEPTH) and focus on smaller value distributions.





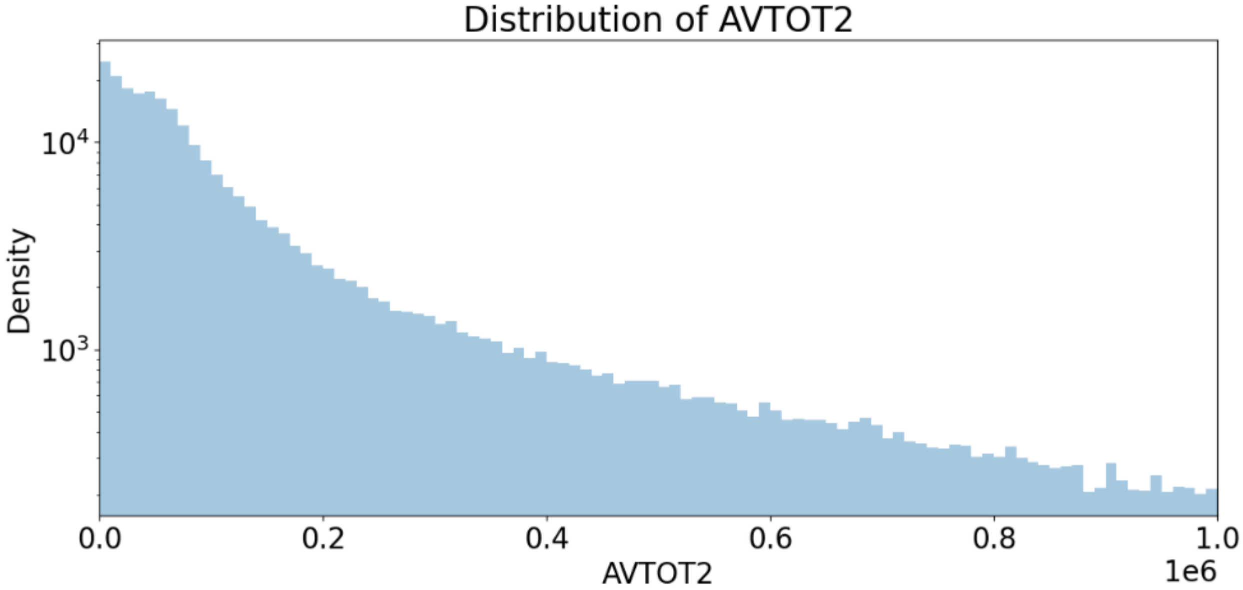
AVLAND2:

The "AVLAND2" field, a numeric attribute present in 26.4% of the 282,726 records, signifies transitional land values. Ranging from 3.00 to 2,371,005,000.00, with an average value of $246,235.72 and a median of $6,178,951.64, the visualizations depict the distribution of transitional land values (AVLAND2).



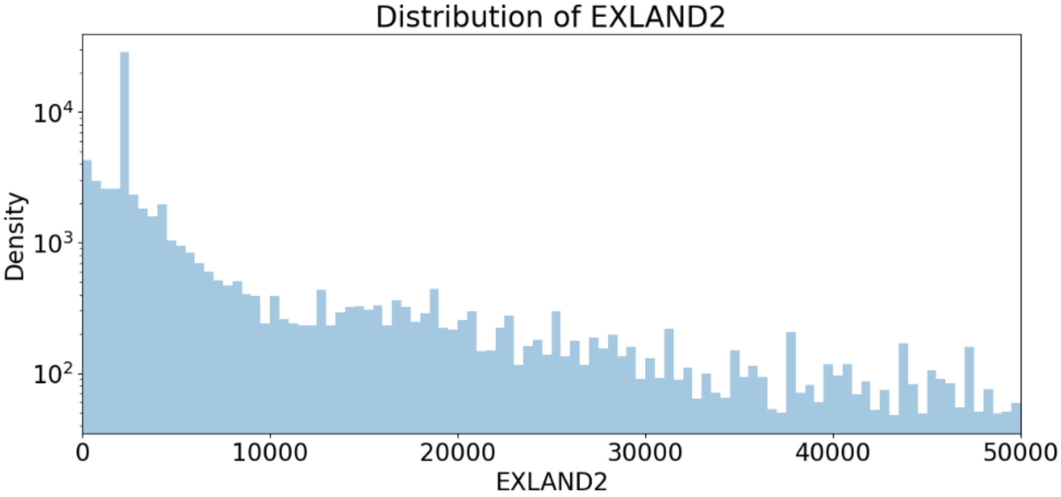
AVTOT2:

The "AVTOT2" field, a numeric attribute present in 26.4% of the 282,732 records, represents transitional total values. Ranging from 3.00 to 4,501,180,002.00, with an average value of $713,911.44 and a median of $11,652,508.34, visualizations illustrate the distribution of transitional total values (AVTOT2).



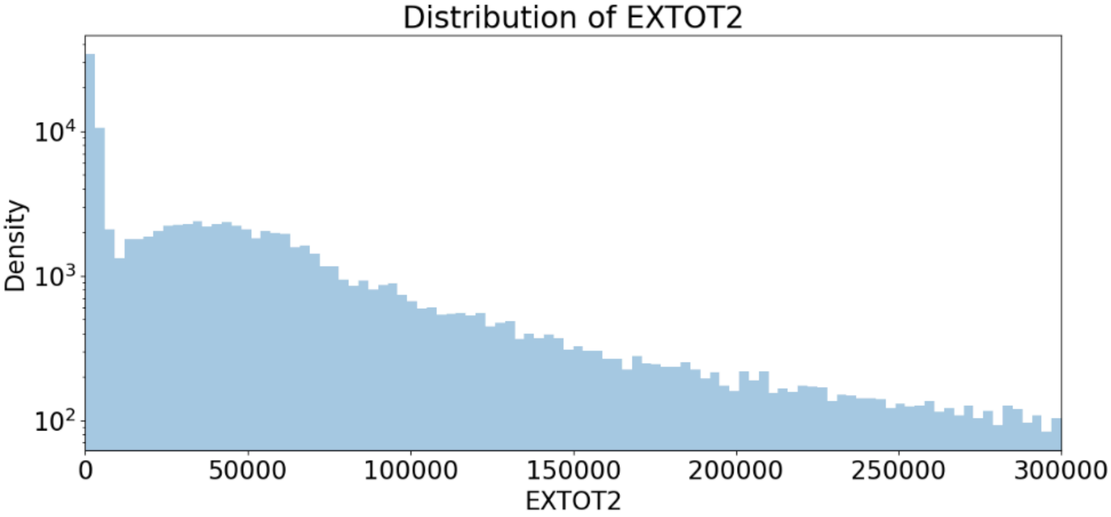
EXLAND2:

The "EXLAND2" field, a numeric attribute present in 8.2% of the 87,449 records, signifies transitional exemption land values. Ranging from 1.00 to 2,371,005,000.00, with an average value of $351,235.68 and a median of $10,802,150.91, visualizations illustrate the distribution of transitional exemption land values (EXLAND2).



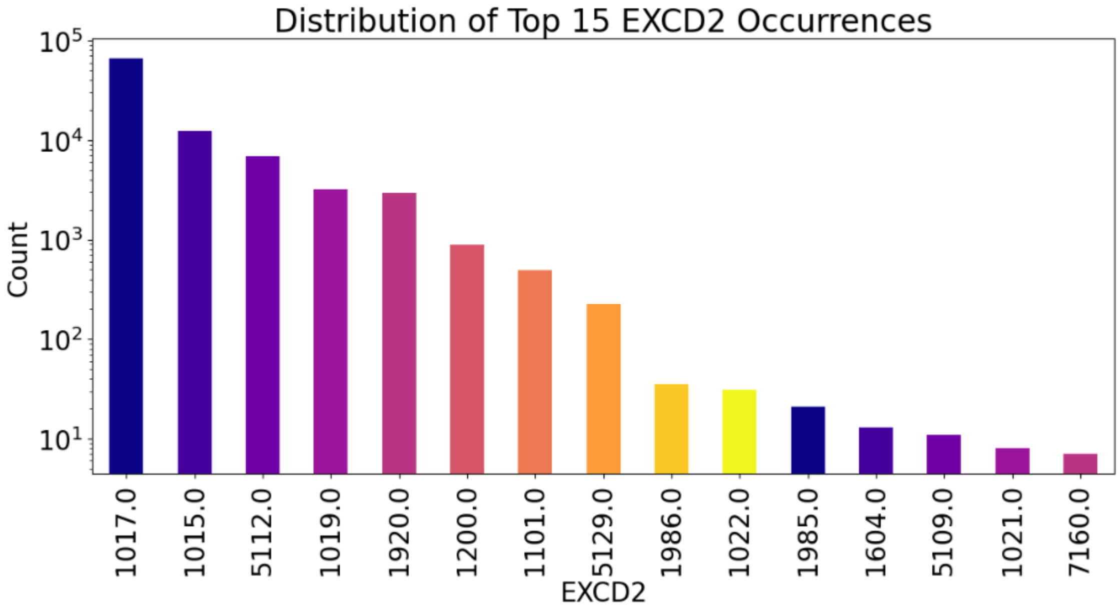
EXTOT2:

The "EXTOT2" field, a numeric attribute present in 12.2% of the 130,828 records, represents transitional exemption land totals. Ranging from 7.00 to 4,501,180,002.00, with an average value of $656,768.28 and a median of $16,072,448.75, visualizations illustrate the distribution of transitional exemption land totals (EXTOT2).



EXCD2:

The "EXCD2" field, a categorical attribute present in 8.7% of the 92,948 records, represents exemption codes (second). With 60 unique codes, the top 15 occurrences are visualized for analysis.



PERIOD:

The "PERIOD" field, a categorical attribute present in all 1,070,994 records, denotes the assessment period. With only one unique value, "FINAL," visualizations may not be necessary due to the lack of variation.

YEAR:

The "YEAR" field, a categorical attribute present in all 1,070,994 records, indicates the assessment year. With only one unique value, "2010/11," visualizations may not be necessary due to the lack of variation.

VALTYPE:

The "VALTYPE" field, a categorical attribute present in all 1,070,994 records, denotes the valuation type. With only one unique value, "AC-TR," visualizations may not be necessary due to the lack of variation.