Project 2 Report

MGTA 463

RAN JI

Table of Contents

Executive Summary

The primary business problem addressed by this project is the detection of anomalies and inconsistencies in property valuations across New York City. These inaccuracies are crucial to resolve as they directly impact the fairness and accuracy of property tax assessments. Inaccurate valuations can lead to uneven tax burdens, potentially harming property owners and the city's fiscal health.

This project involved a comprehensive analysis of the New York Property Data, which includes over one million property records managed by the Department of Finance. The main objective was to identify potential anomalies and inconsistencies in property valuations, which are crucial for accurate real estate assessments and the subsequent calculation of property tax liabilities across New York City. Utilizing advanced data analysis techniques, we streamlined the dataset to highlight unusual records that could indicate errors or potential fraud.

The results of this analysis have significant implications for improving the accuracy and fairness of property tax assessments. By pinpointing specific anomalies and patterns of inconsistency, we provided actionable insights that can help refine assessment processes and enhance fiscal policies. The project not only aids in better fiscal management but also supports the city's ongoing efforts to ensure transparency and fairness in tax assessments, ultimately benefiting all stakeholders involved in New York City's real estate market.

Description of the Data

- Overview of the Data:
  - The dataset is titled **New York Property Data**, which contains comprehensive property valuation and assessment information. The data was collected by the Department of Finance and encompasses **1,070,994 records across 32 fields**, which include both categorical and numerical data types. This dataset is essential for the annual real estate assessment process, which ultimately determines property tax liabilities for various properties within New York City.
- Data Description (See Figures below):
  - The dataset includes numerical fields such as 'LTFRONT', 'LTDEPTH', 'STORIES', 'FULLVAL', 'AVLAND', 'AVTOT', 'EXLAND', 'EXTOT', 'BLDFRONT', 'BLDDEPTH', 'AVLAND2', 'AVTOT2', 'EXLAND2', and 'EXTOT2'.

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	LTFRONT	numeric	1070994	100.0%	169108	0.00	9999.00	36.64	74.03	0.00
1	LTDEPTH	numeric	1070994	100.0%	170128	0.00	9999.00	88.86	76.40	100.00
2	STORIES	numeric	1014730	94.7%	0	1.00	119.00	5.01	8.37	2.00
3	FULLVAL	numeric	1070994	100.0%	13007	0.00	6150000000.00	874264.51	11582425.58	0.00
4	AVLAND	numeric	1070994	100.0%	13009	0.00	2668500000.00	85067.92	4057258.16	0.00
5	AVTOT	numeric	1070994	100.0%	13007	0.00	4668308947.00	227238.17	6877526.09	0.00
6	EXLAND	numeric	1070994	100.0%	491699	0.00	2668500000.00	36423.89	3981573.93	0.00
7	EXTOT	numeric	1070994	100.0%	432572	0.00	4668308947.00	91186.98	6508399.78	0.00
8	BLDFRONT	numeric	1070994	100.0%	228815	0.00	7575.00	23.04	35.58	0.00
9	BLDDEPTH	numeric	1070994	100.0%	228853	0.00	9393.00	39.92	42.71	0.00
10	AVLAND2	numeric	282726	26.4%	0	3.00	2371005000.00	246235.72	6178951.64	2408.00
11	AVTOT2	numeric	282732	26.4%	0	3.00	4501180002.00	713911.44	11652508.34	750.00
12	EXLAND2	numeric	87449	8.2%	0	1.00	2371005000.00	351235.68	10802150.91	2090.00
13	EXTOT2	numeric	130828	12.2%	0	7.00	4501180002.00	656768.28	16072448.75	2090.00

  - It also features several categorical fields including 'RECORD', 'BBLE', 'BORO', 'BLOCK', 'LOT', 'EASEMENT', 'OWNER', 'BLDGCL', 'TAXCLASS', 'EXT', 'EXCD1', 'STADDR', 'ZIP', 'EXMPTCL', 'EXCD2', 'PERIOD', 'YEAR', and 'VAL

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	RECORD	categorical	1070994	100.0%	0	1070994	1
1	BBLE	categorical	1070994	100.0%	0	1070994	1000010101
2	BORO	categorical	1070994	100.0%	0	5	4
3	BLOCK	categorical	1070994	100.0%	0	13984	3944
4	LOT	categorical	1070994	100.0%	0	6366	1
5	EASEMENT	categorical	4636	0.4%	0	12	E
6	OWNER	categorical	1039249	97.0%	0	863347	PARKCHESTER PRESERVAT
7	BLDGCL	categorical	1070994	100.0%	0	200	R4
8	TAXCLASS	categorical	1070994	100.0%	0	11	1
9	EXT	categorical	354305	33.1%	0	3	G
10	EXCD1	categorical	638488	59.6%	0	129	1017.00
11	STADDR	categorical	1070318	99.9%	0	839280	501 SURF AVENUE
12	ZIP	categorical	1041104	97.2%	0	196	10314.00
13	EXMPTCL	categorical	15579	1.5%	0	14	X1
14	EXCD2	categorical	92948	8.7%	0	60	1017.00
15	PERIOD	categorical	1070994	100.0%	0	1	FINAL
16	YEAR	categorical	1070994	100.0%	0	1	2010/11
17	VALTYPE	categorical	1070994	100.0%	0	1	AC-TR

  - Source: New York Property Data collected by the Department of Finance.
  - Fields: 32 fields, including both numerical and categorical data.
  - Records: 1,070,994 records.

- Purpose: To determine property tax liabilities for various properties within New York City

- Important Field Distributions
  - STORIES (See Figure below)
    - The STORIES variable represents the number of stories in a building, as recorded in the dataset. The distribution demonstrates a broad range of building heights, with a notable decrease in frequency as the number of stories increases, suggesting that taller buildings are less common in the dataset. Buildings with fewer stories are more prevalent, indicating a higher frequency of low-rise constructions.

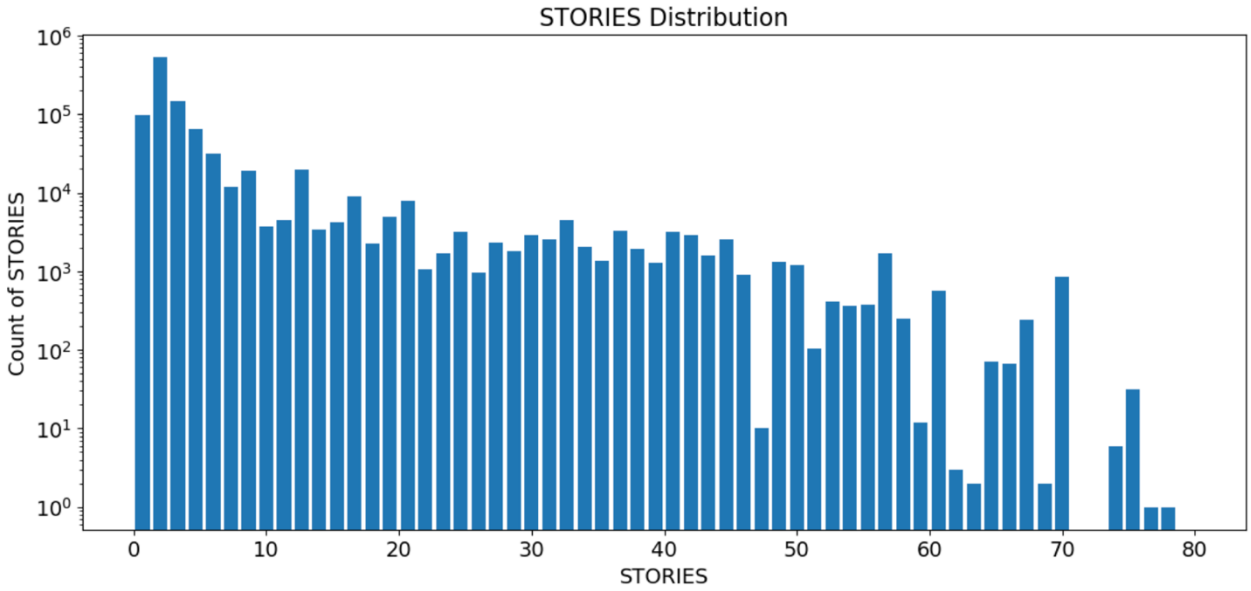
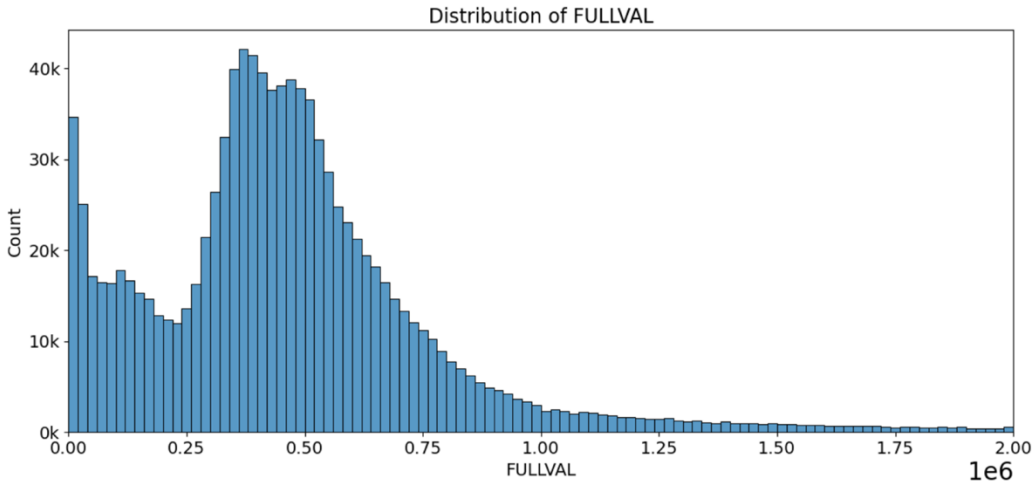


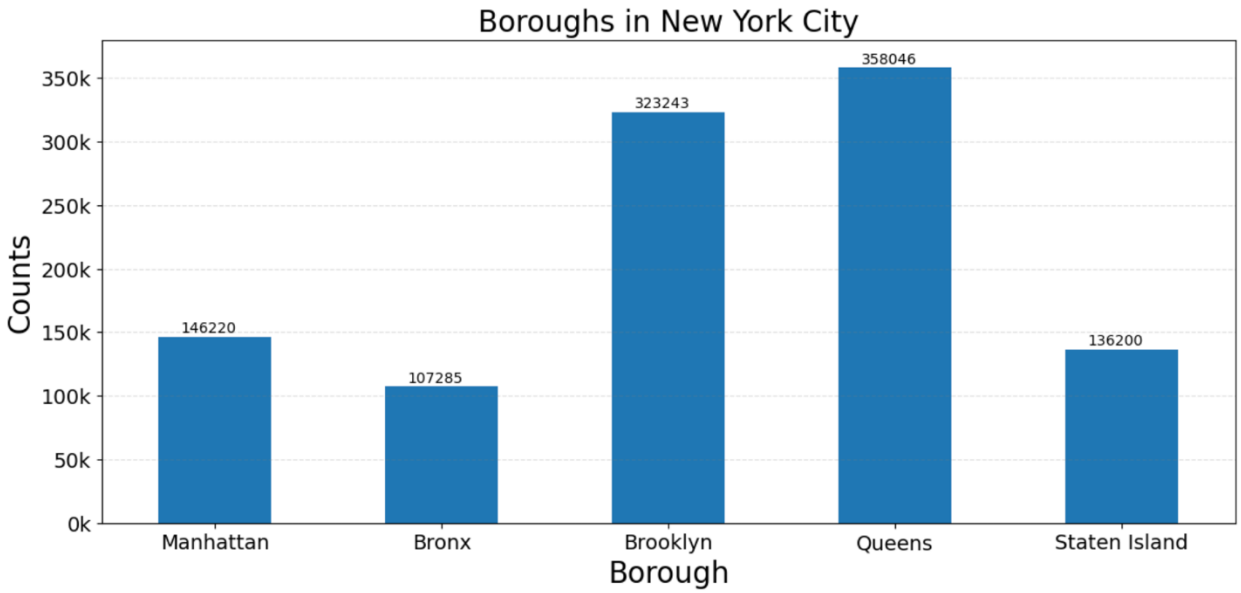
Figure 2

- FULLVAL (See Figure below)



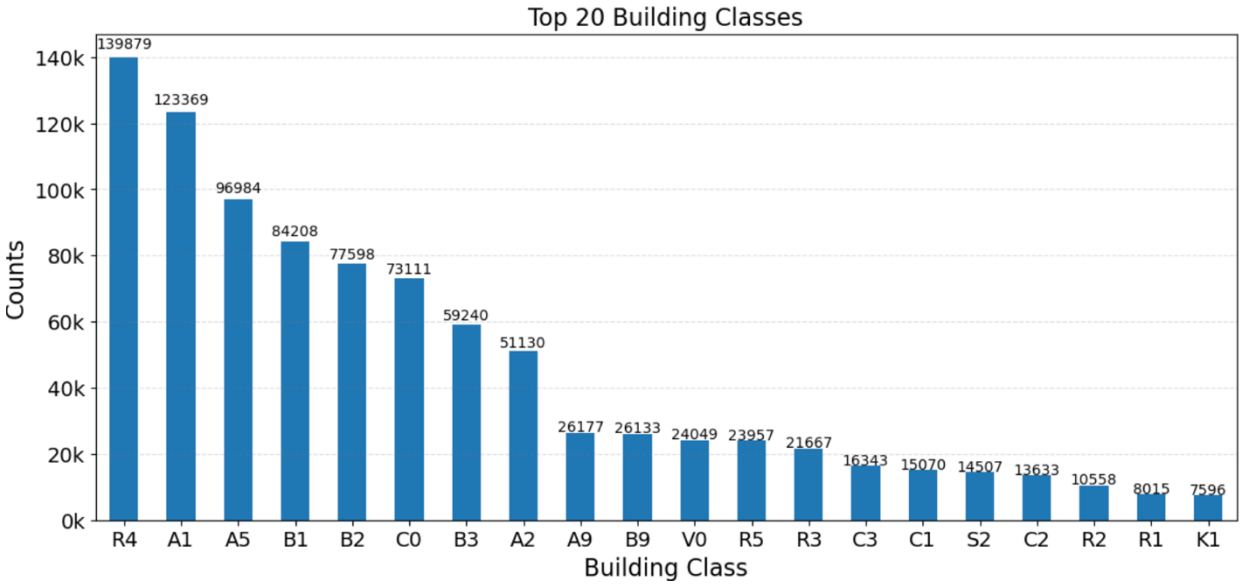
- Description: The FULLVAL variable represents the market value of properties as assessed. The distribution shows a right-skewed pattern where most properties are valued under $1 million, which is typical for the dataset, indicating a higher concentration of lower-valued properties. Initial examination of a boxplot reveals that the bulk of data points cluster below $2 million. Consequently, the distribution's x-axis is limited to $2 million to highlight the area where the majority of values lie, providing a clearer view of the distribution characteristics.

- BORO (See Figure below)



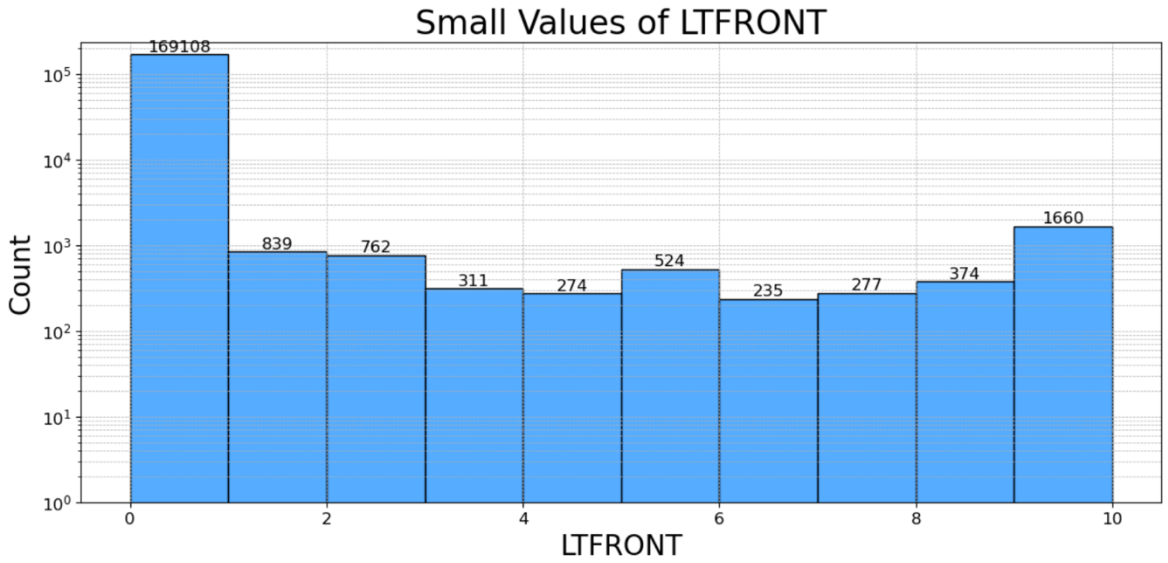
- BORO represent the five boroughs of New York City. Queens has the highest number of properties with a count of 358,046, followed by Brooklyn with 323,243 properties. Manhattan, despite its prominence, has fewer properties recorded at 146,220. The Bronx and Staten Island have 107,285 and 136,200 properties respectively, indicating a varied distribution of real estate across the boroughs.

- BLDGCL (See Figure below)



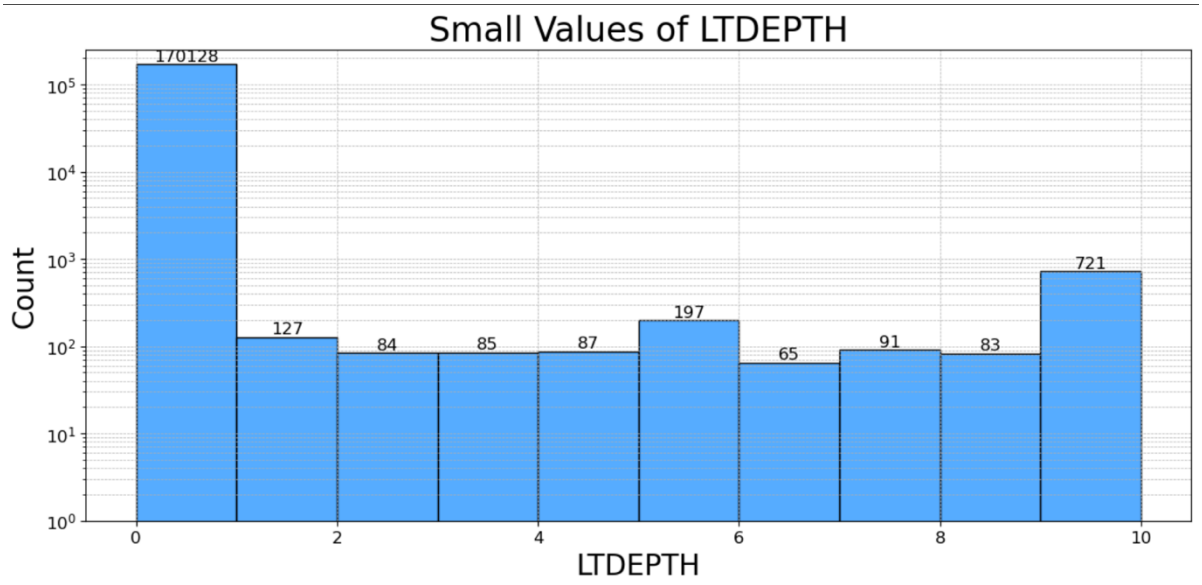
- The BLDGCL (Building Class) variable in the dataset categorizes properties based on their type and use, such as residential, commercial, or mixed-use buildings. Each class is represented by a code, such as R4 for residential condominiums, A1 for one-family dwellings, and C0 for walk-up apartments, among others. The distribution across the top 20 building classes highlights the diversity of building types within New York City, with the highest counts found in classes that typically represent densely populated residential areas.

- LTFRONT



  - The LTFRONT variable represents the width of the lot facing the street, measured in feet, and is crucial for understanding property layout and assessing property value. Analysis of both boxplot and distribution plot reveals that most LTFRONT values are concentrated within 10 feet, indicating a common urban property characteristic where lots have smaller street-facing dimensions. This is typical in densely built areas where space is at a premium. The distribution shows a skew towards smaller lot frontages, with a marked decline in occurrence as lot width increases beyond 10 feet. This visualization uses a logarithmic scale on the y-axis to better display the frequency of smaller values, enhancing the visual interpretation of data spread and concentration, with the x-axis limited to 10 feet to focus on the most common property widths.

- LTDEPTH



- The LTDEPTH variable measures the depth of a property lot in feet, from the street front to the back of the lot. Analysis of both boxplot and distribution plot shows a strong concentration of values within 10 feet, highlighting a common characteristic in densely built urban environments where space is maximized. The distribution, skewed toward smaller lot depths, reflects a notable range in property sizes within New York City. This visualization specifically limits the x-axis to 10 feet to focus on the most prevalent measurements and employs a logarithmic scale on the y-axis to clearly illustrate the frequency of smaller lot depths.

Data Cleaning

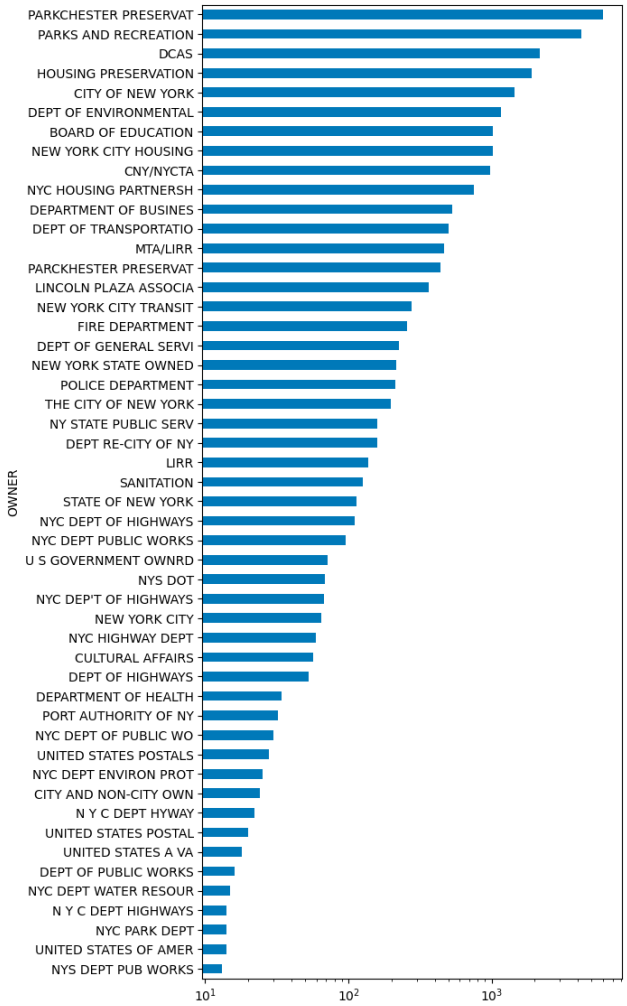
Describe Cleaning and Variables

1. Motivation of exclusions:
    a. In the initial stages of this project, a meticulous process of exclusions was essential to refine the dataset and ensure its suitability for analyzing non-governmental property transactions. This section describes the motivations and steps involved in excluding certain records from the New York Property Data, which initially contained a specific number of entries (numrecords_orig).
2. Remove Exclusions
    a. Initially, the dataset consists of a specific number of records (numrecords_orig). This count represents the total dataset size before any exclusions are applied.
    b. Removing Government Easements
        i. The first exclusion step targets records with an "EASEMENT" marked as "U", indicating government use. These records are removed because they are not relevant to the analysis focusing on non-governmental property transactions. The difference between the original record count and the count after this removal provides the number of records excluded in this step, which is stored and printed as numremoved. The number of records removed is 1.
    c. Identifying Owners for Exclusion
        i. A list of keywords (gov_list = ['DEPT ', 'DEPARTMENT', 'UNITED STATES','GOVERNMENT',' GOVT ', 'CEMETERY']) that typically signify government or cemetery ownership is compiled. These keywords are used to filter out owners from the dataset. The script checks each owner name against this list, excluding names that contain any of the keywords but ensuring names with 'STORES' are not incorrectly removed to avoid excluding commercial entities erroneously. Here, the total owner number before removing is 863347.
    d. Analyzing Frequent Owners
        i. The script identifies the top 20 most frequently occurring owner names in the dataset. This step aims to uncover any large entities or frequently appearing names that might skew the analysis if not excluded. Additional names known to represent government entities are manually appended to the exclusion list such as ‘THE CITY OF NEW YORK’ , ‘NYS URBAN DEVELOPMENT’…
    e. Refining and Applying Exclusions
        i. Further refinement of the owner exclusion list is undertaken by manually reviewing and removing any names that should not be excluded, such as those mistakenly identified or irrelevant to government or cemetery

associations. The comprehensive exclusion list is then applied to the dataset, removing all records associated with these owners.

f. For now, we have removed numbers of records are 26501.

g. After applying the exclusions, a bar chart is generated to visually inspect the frequencies of the remaining owner names. This visual confirmation helps ensure that no significant entities that could bias the analysis remain. The final count of records removed through these exclusions is calculated by comparing the adjusted dataset size to the original, confirming the extent of data reduction.



h. After thoroughly executing these exclusion steps, the dataset is significantly refined, ensuring that it primarily includes records pertinent to the analysis. By meticulously removing entries associated with government-owned properties and cemeteries, we have enhanced the dataset's relevance and quality for detecting anomalies in non-governmental property transactions. This careful preparation sets a robust foundation for the subsequent phases of data cleaning, including

field imputation and detailed anomaly detection analysis. This meticulous approach ensures that the analysis will be based on the most relevant and accurate data available, maximizing the potential for insightful and actionable findings.

3. Motivation of imputation logic:
    a.
4. Fill in missing ZIP
    a. Initial Identification of Missing ZIP Codes
        i. Initially, the process begins with identifying the total number of missing ZIP codes in the dataset. Using a condition to find null values in the 'ZIP' column, it is determined that there are 20,431 missing ZIP codes. This step establishes the baseline for the imputation work that follows.
    b. Verification of Related Column Integrity
        i. Before proceeding with ZIP code imputation, it's essential to check the integrity of related columns that will be used in constructing unique identifiers. It's found that there are no missing values in the 'BORO' column, but there are 364 missing entries in the 'STADDR' column. Ensuring minimal missing data in these supporting columns is crucial for the accuracy of the subsequent imputation steps.
    c. Creation of Composite Address Identifier
        i. To facilitate a more accurate ZIP code mapping, a new column named 'staddr_boro' is created by concatenating the 'STADDR' (street address) and 'BORO' (borough) columns. This composite identifier uniquely represents each address across boroughs, which aids in associating missing ZIP codes with their correct locations based on address data.
    d. Initial ZIP Code Imputation Using Address Mapping
        i. A dictionary is constructed to map these unique 'staddr_boro' identifiers to known ZIP codes, capturing the relationships where available. This dictionary is then used to fill in missing ZIP codes by mapping back to the 'staddr_boro' in the data. This method successfully imputes 2,832 ZIP codes, reducing the total number of missing ZIPs to 17,599.
    e. Sequential ZIP Code Consistency Check
        i. Leveraging the orderliness of the dataset, a method is employed where missing ZIP codes are filled based on the consistency of neighboring ZIP values. If a ZIP code is absent and the ZIP codes of the immediate preceding and following records are the same, that consistent ZIP code is used to fill the gap. This step fills in 9,491 more ZIP codes, bringing the number of still-missing ZIPs down to 8,108.
    f. Final ZIP Code Filling Using Previous Record Values

i. For the remaining missing ZIP codes, a straightforward approach is adopted. Each missing ZIP is filled with the ZIP code from the previous record, assuming geographical proximity implies similar ZIPs. This process fills all the remaining 8,108 missing ZIPs, resulting in no missing ZIP codes left in the dataset.

g. Data Cleanup Post-Imputation
    i. After all missing ZIP codes are imputed, the temporary 'staddr_boro' column, used solely for aiding the imputation process, is removed from the dataset.

h. The process of filling missing ZIP codes in the dataset has been successfully completed, reducing the initial 20,431 missing entries to zero through a structured approach involving address mapping, sequential consistency checks, and direct carryovers. This has significantly enhanced the dataset's completeness and accuracy, ensuring it is well-prepared for further analysis or modeling

5. FULLVAL, AVLAND, AVTOT
    a. FULLVAL
        i. Initially, the number of properties listed with a FULLVAL of zero was identified to be 10,025. These entries potentially represent incomplete or unassessed properties.
        ii. To facilitate imputation, all zero values in the FULLVAL field were converted to NaN (null values), allowing for more standardized filling methods. After this conversion, the total number of missing values in FULLVAL remained 10,025, as no previously missing values existed in the dataset.
        iii. The dataset was grouped by TAXCLASS, BORO, and BLDGCL (building class), and the missing values in FULLVAL were filled using the mean FULLVAL of each group. After this step, the number of missing values was reduced to 7,307, indicating that not all groups had sufficient data to compute a mean.
        iv. To address the remaining missing values, a less granular grouping was used, this time only by TAXCLASS and BORO. This reduced the number of missing values further to 386, showing an improvement but still leaving some entries unfilled.
        v. The last step involved grouping by TAXCLASS alone, ensuring that all properties at least had a fallback mean value based on their tax classification. This step successfully filled all remaining missing values, resulting in zero missing values in the FULLVAL field.
    b. AVLAND

i. Initially, the dataset was examined to identify how many properties had an AVLAND value of zero, indicating potentially unassessed or incorrectly recorded entries. The total count of properties with a zero value was found to be 10,027.
ii. To facilitate imputation, all zero values in the AVLAND field were replaced with NaN (null values). This conversion was necessary to standardize missing values and prepare for data imputation. After replacing zeros with NaNs, the total number of missing values in AVLAND remained at 10,027.
iii. The first imputation attempt involved grouping the data by TAXCLASS, BORO, and BLDGCL (building class). Within each group, missing AVLAND values were filled using the mean AVLAND value of the group. This step aimed to preserve the valuation characteristics that are specific to the property type and location. However, after this step, 7,307 values remained unfilled, indicating incomplete groups or those without sufficient data to calculate a mean.
iv. To address the still-missing values, a less granular approach was employed by grouping the properties only by TAXCLASS and BORO. The missing values were again filled using the mean of these new groups. This reduced the number of missing values to 386, showing that broader group averages could provide a fallback for more properties.
v. In the final imputation step, the dataset was grouped solely by TAXCLASS. This grouping ensured that every missing AVLAND value had at least a tax class-specific mean value for filling, regardless of borough or building class distinctions. This step successfully filled all remaining missing values, bringing the total number of missing values in the AVLAND field to 0.
vi. This detailed, tiered approach to imputing missing values in the AVLAND field ensures that each property's land valuation is estimated as accurately as possible with the available data. By progressively broadening the grouping criteria, the imputation process effectively minimized the number of properties with undefined actual land values, enhancing the dataset's overall quality and usability for further analysis.

c. AVTOT
i. The process starts by identifying properties in the dataset where AVTOT (total assessed value) is recorded as zero. A total of 10,025 properties initially had a zero value, indicating incomplete or potentially incorrect assessments.
ii. All zero values in the AVTOT field were replaced with NaN to standardize the handling of missing data. This conversion facilitates more uniform

data imputation strategies. After this step, there were 10,025 missing values in the AVTOT field.

iii. The initial imputation strategy grouped properties by TAXCLASS, BORO, and BLDGCL (building class). The mean AVTOT of each group was used to fill missing values. This approach respects the property classification and location, ensuring that the imputed values are as realistic as possible. However, after this step, 7,307 values remained unfilled, indicating that some groups lacked enough data to calculate a meaningful average.

iv. To further reduce the number of missing values, the dataset was grouped by just TAXCLASS and BORO. The missing AVTOT values were again filled using the mean of these broader groups. This step reduced the number of missing values to 386, demonstrating that a less detailed grouping could still provide useful imputation values for many properties.

v. The last imputation step involved grouping the properties solely by TAXCLASS. Every missing AVTOT value was filled using the mean AVTOT of the respective tax class. This final step successfully filled all remaining missing values, resulting in zero missing values in the AVTOT field.

vi. This hierarchical, step-by-step imputation process has ensured that all missing or incorrect entries in the AVTOT field were addressed comprehensively.

6. Fill in the missing STORIES
    a. The first step was to assess the extent of missing data in the STORIES field. It was found that 42,030 entries lacked this information, representing a considerable portion of the dataset that required attention for accurate property analysis.
    b. To address these missing values, the most common number of stories (mode) for buildings was calculated within groups defined by their BORO and BLDGCL (building class). This grouping strategy was chosen because building heights can vary significantly across different boroughs and types of buildings. The mode represents the most frequently occurring value in the dataset, making it a realistic choice for filling in missing STORIES. After applying this imputation method, the number of missing entries was reduced to 37,922.
    c. For the remaining missing values, a second imputation step was employed, grouping buildings by TAXCLASS. Within each tax class group, missing STORIES values were filled using the mean number of stories. This step assumes that properties within the same tax class will have similar structural characteristics, including the number of floors. This method successfully filled all remaining missing values, reducing the count of missing STORIES to 0.

d. After completing the imputation steps, a verification showed that there were no remaining missing values in the STORIES field. The dataset's head was displayed to ensure that the STORIES values were appropriately filled and to confirm the overall integrity and consistency of the data following the imputation process.
e. This methodical approach to filling in missing STORIES ensured that each property's record was completed in a manner consistent with its characteristics and those of similar properties.

7. Fill in LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT with averages by TAXCLASS
    a. Firstly, it was identified that LTFRONT, LTDEPTH, BLDDEPTH, and BLDFRONT fields had zero values, which are invalid for dimensional measurements. These zero values needed to be treated as missing data to accurately represent property dimensions.
    b. LTFRONT
        i. Initial Missing Values: Initially, 160,565 entries for LTFRONT were identified as missing or zero and converted to NaNs for accurate processing.
        ii. Mean values for LTFRONT were calculated and applied within each group formed by TAXCLASS and BORO. This reduced the NaN count significantly, but some values remained NaN, particularly in groups lacking sufficient data.
        iii. Action: A broader grouping by TAXCLASS alone was used to fill remaining NaNs using the mean of this group. This step successfully filled all remaining missing values, reducing the NaN count to zero.
    c. LTDEPTH
        i. There were 161,656 missing or zero values for LTDEPTH.
        ii. The dataset was grouped similarly by TAXCLASS and BORO, and means were applied. Some entries still lacked data due to insufficient group data.
        iii. Remaining NaNs were filled using the mean values calculated from grouping by TAXCLASS alone.
        iv. All missing values were successfully filled, with the final NaN count at zero.
    d. BLDFRONT
        i. Initially, there were no missing values identified explicitly before the grouping imputation.
        ii. Avenues included grouping by combinations of TAXCLASS, BORO, and BLDGCL, then by TAXCLASS and BORO, and finally by TAXCLASS alone, applying mean values accordingly.
        iii. Each step confirmed no remaining NaNs, effectively handling any potential missing data thoroughly through the grouping strategies.
    e. BLDDEPTH

i. Similar to BLDFRONT, no missing values were mentioned before processing.
ii. The same tiered grouping strategy as BLDFRONT was applied.
iii. Post-imputation checks confirmed no missing values, indicating a comprehensive coverage and filling of any potential gaps.

f. This detailed and hierarchical imputation process for building and lot dimensions ensured that all properties in the dataset had realistic and consistent dimensional values.

8. Conversion of ZIP Code Data Type:
    a. The ZIP field, originally stored as a float (which could lead to incorrect ZIP code formats due to rounding or dropping of leading zeros), is converted to a string format. This conversion is crucial for maintaining the accuracy of ZIP codes, particularly when they are used in analyses that require precise geographical identification.
    b. The conversion is done using the astype(str) method, which transforms each ZIP code entry into a string, preserving all characters and avoiding any data loss that comes from numerical representation.
9. Extraction of the First Three Digits of ZIP Codes
    a. A new column zip3 is created to store only the first three digits of each ZIP code. The first three digits of a ZIP code are often used to broadly categorize geographical areas, making this reduction useful for analyses that require a more general location indicator without the need for precise ZIP code data. This is achieved by slicing the first three characters from the ZIP string.
10. Summary
    a. All columns representing dimensional data (like LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT) had zero values converted to NaNs, recognizing these as missing data and preparing them for accurate imputation.
    b. Missing data for dimensions and other important property characteristics were imputed using a hierarchical strategy based on TAXCLASS and sometimes additional characteristics like BORO or BLDGCL.
    c. The ZIP codes were normalized by converting them into string format to preserve leading zeros and extracting the first three digits for broader geographical categorization, useful in macro-level analyses.

Variable Creation

1. In the New York Property Data analysis, our primary objective is to identify potential anomalies that may suggest inaccuracies or fraudulent activities in property valuations. These anomalies are critical to detecting as they can significantly impact the accuracy of property tax assessments. To achieve this, we developed a series of derived variables that enable a detailed examination of property values in relation to their physical dimensions and assessed values.

2. Creation of Derived Variables
    a. Lot and Building Calculations:
        i. Lot Size (ltsize): We calculate the total area of the property by multiplying the lot frontage (LTFRONT) by the lot depth (LTDEPTH). A small constant (epsilon) is added to avoid division by zero in subsequent calculations.
        ii. Building Size (bldsize): The building area is obtained by multiplying the building front (BLDFRONT) with the building depth (BLDDEPTH), also incorporating epsilon to handle zeros effectively.
        iii. Building Volume (bldvol): By multiplying the building area by the number of stories (STORIES), we estimate the total volume of the building, adjusting for potential zero values with epsilon.
    b. Ratio Variables:
        i. We compute nine ratios (r1 to r9) that relate the property's financial values (FULLVAL, AVLAND, AVTOT) to its physical dimensions (lot size, building size, building volume). These ratios help highlight discrepancies between the property's assessed value and its physical size, which are often indicators of erroneous data or potential manipulation.
    c. Scaling and Inversion of Ratios:
        i. Each ratio is scaled by its median across the dataset to normalize the values, making comparisons more consistent.
        ii. To enhance the detection of anomalies, especially those values close to zero, we invert each ratio. This transformation elevates small values into large outliers, making them easier to identify.
    d. Maximum Value Selection:
        i. For each property, we select the maximum value between the original and the inverted ratio. This approach ensures that both high and low extremes are captured, highlighting all potential outliers.
    e. Group Normalization:
        i. The variables are further refined by standardizing them within logical groups defined by ZIP code and tax class. This involves calculating the

mean of each ratio within these groups and creating new variables that express each property's ratios relative to the averages of its group. This step is crucial for spotting anomalies not just on a city-wide scale but within localized or categorically similar properties.

3. Detailed logic for all variables
    a. Creation of Derived Variables
        i. Lot Size Calculation (ltsize): Computes the product of lot frontage (LTFRONT) and lot depth (LTDEPTH) to get the total lot area, adding a small number (epsilon) to avoid division by zero in subsequent calculations.
        ii. Building Size Calculation (bldsize): Calculates the building area by multiplying building front (BLDFRONT) by building depth (BLDDEPTH), also adding epsilon.
        iii. Building Volume Calculation (bldvol): Multiplies the building area (bldsize) by the number of stories (STORIES) to estimate the total volume of the building, including epsilon.
    b. Ratio Variables
        i. R1-R9 represent different ratios used to compare property values to physical dimensions:
            1. r1, r2, r3: These ratios relate the total property value (FULLVAL) to the lot size, building size, and building volume, respectively.
            2. r4, r5, r6: Similar to the above, but using land value (AVLAND) instead.
            3. r7, r8, r9: Use the total property assessment value (AVTOT) for comparison.
    c. Scaling of Ratio Variables
        i. Each of the nine ratio variables is then scaled by dividing by the median of that variable, normalizing them to a consistent scale and making them easier to compare across the dataset.
    d. Inversion of Ratio Variables
        i. To facilitate the detection of very low outliers (which are close to zero and thus not many standard deviations below the mean), the inverse of each ratio variable is calculated. This transformation turns small values into large outliers, which can be more easily detected.
    e. Selection of Max Values
        i. For each property, the maximum value between the original and the inverse ratio is retained. This method ensures that both extremely high and extremely low values are captured as potential outliers.
    f. Removal of Inverse Columns

        i. Once the maximum values are determined, the inverse columns are dropped since they are no longer needed, simplifying the dataset.
    g. Group Normalization
        i. Additional derived variables are standardized by grouping by logical categories such as ZIP code and tax class. This step involves:
        ii. Calculating the mean of each ratio variable for each group (zip5_mean and taxclass_mean).
        iii. Creating new variables for each ratio that represent the ratio divided by the group mean, allowing for comparisons of each property against its group's typical values.
    h. Additional Derived Variables
        i. value_ratio: A new variable combining the full property value with land and total values, normalized across the dataset.
        ii. Handling of Extremely Low Values: For value_ratio, where the variable is below one, the inverse is used if it is larger, ensuring that low outliers are emphasized.
    i. Summary
        i. This thorough approach ensures that each property's values are not only compared to its physical dimensions but also to its peers within the same ZIP code or tax class, enhancing the detection of anomalies. The final dataset consists of these refined variables, now ready for unsupervised modeling to identify potentially fraudulent or unusual properties. This preprocessing prepares the data effectively by highlighting extremes in property characteristics relative to assessed values and physical dimensions, crucial for spotting anomalies in a real estate dataset.
4. Motivation and Business Impact
    a. The motivation behind these comprehensive calculations and transformations is to prepare the dataset for unsupervised modeling that can effectively identify outliers. These outliers might represent potential fraud, errors in data entry, or unusual but legitimate property characteristics that require further investigation. By comparing each property not only to city-wide norms but also to localized standards, we can pinpoint anomalies with greater precision and relevance.
    b. This detailed approach to variable creation and analysis ensures that the city's finance department can better understand the landscape of property valuations and take informed steps to address discrepancies. The results of this analysis will ultimately lead to more accurate property tax assessments, ensuring fairness and efficiency in tax collection, which is pivotal for the city’s financial health and the trust of its residents.

5. Variable list:

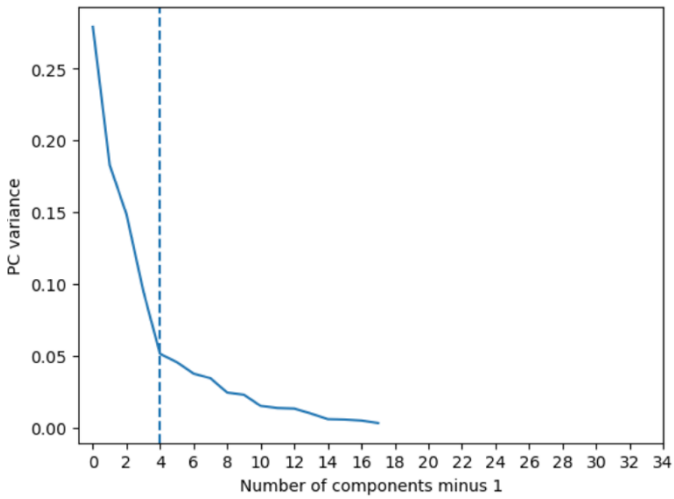
Description	# Variables_Created
Ratios based on property dimensions and valuation metrics R1-r9: These ratios compare different valuation metrics (FULLVAL, AVLAND, AVTOT) to property size dimensions (ltsize, bldsize, bldvol).	9
Normalized ratios by ZIP code r1_zip5 to r9_zip5: These variables are standardized versions of r1 to r9 ratios, normalized by the average values for their respective groups defined by ZIP code (ZIP)	9
Normalized ratios by tax class r1_taxclass to r9_taxclass: These variables are standardized versions of r1 to r9 ratios, normalized by the average values for their respective groups defined by tax classification (TAXCLASS).	9
Value ratio A composite metric that compares the total property value (FULLVAL) to the sum of the land (AVLAND) and total assessed values (AVTOT). This ratio is then normalized across the dataset to identify properties whose valuation ratios significantly deviate from the norm.	1
Size ratio This ratio measures the building size relative to the lot size, providing a metric of how much of the lot is covered by the building. This can indicate properties that are unusually developed relative to their lot size, which might be of interest in certain urban planning or zoning analyses.	1

Dimensionality Reduction

- Background and Setup
  - In order to reduce dimension, we implemented the Principal Component Analysis (PCA) to perform on a dataset of New York property data. The initial setup includes importing necessary libraries such as Scikit-learn for PCA, Pandas for data manipulation, and Matplotlib and Seaborn for visualization. PCA is intended to reduce the dimensionality of the dataset while retaining as much variance as possible.
- Steps Performed in PCA
  - Data Standardization
    - Prior to applying PCA, the data is standardized to ensure that each feature contributes equally to the analysis. This is critical because PCA is sensitive to the variances of the initial variables.
  - Applying PCA to Retain 99% Variance
    - A PCA object is configured to retain 99% of the variance in the dataset, ensuring that the most significant features are captured.

```
# do a complete PCA and look at the scree and cumulative variance plots
pca = PCA(n_components = .99, svd_solver = 'full')
```

    - This step determines the number of components needed to capture 99% of the variance without arbitrarily choosing the number of principal components.
  - Variance Analysis
    - The variance ratio for each principal component is plotted to evaluate how much variance each component accounts for.

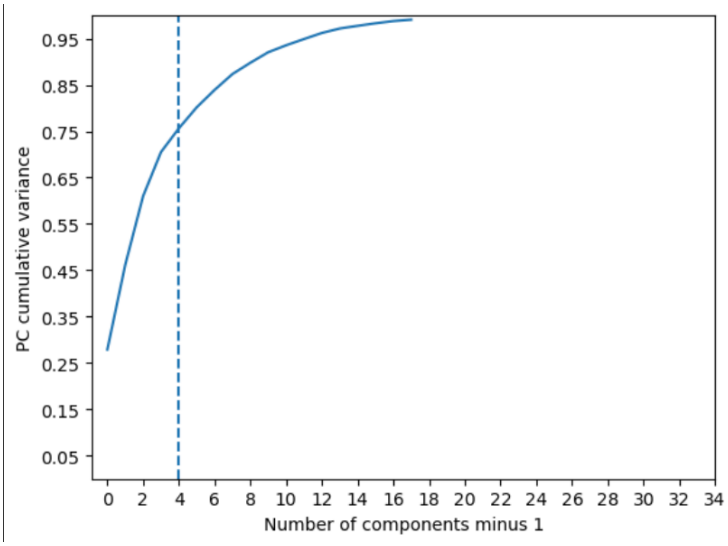


      - Steep Decline: The first principal component explains the most variance, and there is a significant drop in the amount of variance explained by subsequent components. This steep decline indicates

that a few components capture most of the information in the dataset.

- Key Components: The dashed blue line in the graph marks the first four principal components. It suggests that these four components capture the vast majority of the variance. This insight implies that for most analytical purposes, retaining just these first four components would be sufficient.

- Axes Interpretation:
  - Horizontal Axis: Represents the number of components, adjusted by minus one for clarity and consistency in presentation, which is common in PCA visualizations.
  - Vertical Axis: Shows the proportion of the dataset's total variance that each component explains.
- This type of visualization is typically used to decide the number of principal components to retain in a PCA model. In this case, the sharp drop after the fourth component suggests that retaining more components might not add significant value to the analysis. Thus, it could be practical to limit the model to these four principal components to reduce model complexity while retaining most of the critical information from the dataset.
- A cumulative variance plot is also generated to visually assess the total variance captured as more components are added.



- The cumulative variance plot provides a visual assessment of the variance captured as additional principal components are incorporated into the PCA model.

- Rapid Gain: The plot shows a sharp increase in explained variance with the initial components, with about 75% of the variance captured by the first four components, indicated by the blue dashed line.
- Diminishing Returns: Beyond the fourth component, the rate of variance accumulation slows significantly, suggesting that additional components contribute less to capturing new information.
- The cumulative variance plot is essential for determining the optimal number of principal components. The observed plateau beyond the initial few components suggests that retaining more than four components yields minimal additional benefits. This analysis helps in making efficient decisions about the number of components to retain, ensuring the PCA model is both concise and informative while capturing the majority of useful variance with minimal complexity.

- Data Transformation
  - The data is transformed using the PCA model to project it onto the principal components.

```
%%time
# now redo the PCA but just keep the top few PCs
data_zs = data_zs_save.copy()
pca = PCA(n_components = 5, svd_solver = 'full')
princ_comps = pca.fit_transform(data_zs)
pca.n_components_
```

```
CPU times: user 9.49 s, sys: 692 ms, total: 10.2 s
Wall time: 3.08 s

5
```

```
print(np.cumsum(pca.explained_variance_ratio_))
```

```
[0.27895806 0.46160261 0.61006368 0.70507372 0.75630657]
```

```
data_pca = pd.DataFrame(princ_comps, columns = ['PC' + str(i) for i in range(1, pca.n_components_+1)])
data_pca.shape
```

```
(1044493, 5)
```

- After transforming the data, a second z-scaling is applied to the principal components to make them equally important. This step is akin to standardizing the principal components, making later distance calculations (such as Mahalanobis distance) more effective.

- Results
    - The PCA effectively reduced the dimensionality of the dataset to a smaller number of principal components while retaining significant variance. This reduction simplifies subsequent analyses, such as anomaly detection, by focusing on the most informative features of the data. The second z-scaling step ensures that all retained principal components are equally weighted in any further analysis, enhancing the detection capabilities for outliers or anomalies.
- Summary
    - The PCA process implemented in this project is thorough and robust, aiming to maintain as much information as possible while reducing the dataset's complexity. The steps taken, from standardization through to the transformation and second z-scaling, are well-justified to ensure the effectiveness of the dimensionality reduction for subsequent anomaly detection tasks.

Anomaly Detection Algorithms

- Overview
  - In this project, two scoring methods are utilized to detect anomalies in the dataset. Each method scores each observation based on its likelihood of being an anomaly, with higher scores indicating higher likelihoods.
- Method 1: Minkowski Distance-Based Score
  - The first method for anomaly detection uses a Minkowski distance-based approach. The Minkowski distance is a generalized metric that can be adjusted according to the parameter p, allowing it to represent different types of distances, such as Euclidean (p=2) or Manhattan (p=).
  - Equation:

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

    - Where x_i are the elements of the transformed and standardized PCA component scores for each data point.
  - This score calculates the Euclidean distance of each data point from the origin in the transformed PCA space, effectively measuring the "outlyingness" based on the PCA components.
- Method 2: Autoencoder Error Score
  - The second scoring method involves an autoencoder neural network that is trained to minimize the reconstruction error of the input data. An autoencoder learns to compress (encode) the data into a lower-dimensional space and then reconstruct (decode) it back to the original space.
  - The autoencoder is trained with a simple architecture suitable for capturing the underlying patterns without overfitting to outliers. The anomaly score for each instance is computed as the reconstruction error:

$$\left(\sum_{i=1}^{n} |y_i - \hat{y}_i|^p\right)^{1/p}$$

  - Where y_i are the original data points, y_i_hat are the reconstructed data points from the autoencoder, and p is set to 2.
- Combining Scores:
  - Scores from both methods are combined to enhance the robustness of the anomaly detection. This combination helps in mitigating the weaknesses of individual scoring methods.
  - Combination Method:

        - Scores are combined using a simple average, though other methods like weighted average, maximum, or minimum could also be considered based on specific use cases or performance considerations.
    - This approach ensures that the final score benefits from the strengths of both the distance-based and reconstruction error-based methods, providing a more reliable indicator of anomalies in the dataset. This combined scoring method is particularly effective in scenarios where anomalies may not strictly be the farthest points in the dataset but are unusual in terms of their deviation from typical data patterns captured by the autoencoder.

Results

- In unsupervised fraud detection models, especially those dealing with large datasets with complex interactions like property data, results are typically assessed through a combination of statistical measures and visualizations. This method allows us to discern patterns, identify anomalies, and focus on specific properties or records that display unusual characteristics. Here's a step-by-step approach on how to examine these results, illustrated with the NY property data analysis.
- Step 1: Variable Calculation and Ratio Analysis
    - The model first recalculates certain variables and ratios from the data:
        - Variables like V1, V2 and V3 are derived from key property attributes (FULLVAL, AVLAND, and AVTOT).
        - Composite variables like S1, S2 and S3 are computed using dimensional attributes of properties to create new metrics that can be more informative about potential anomalies.

V_1 = FULLVAL
V_2 = AVLAND
V_3 = AVTOT

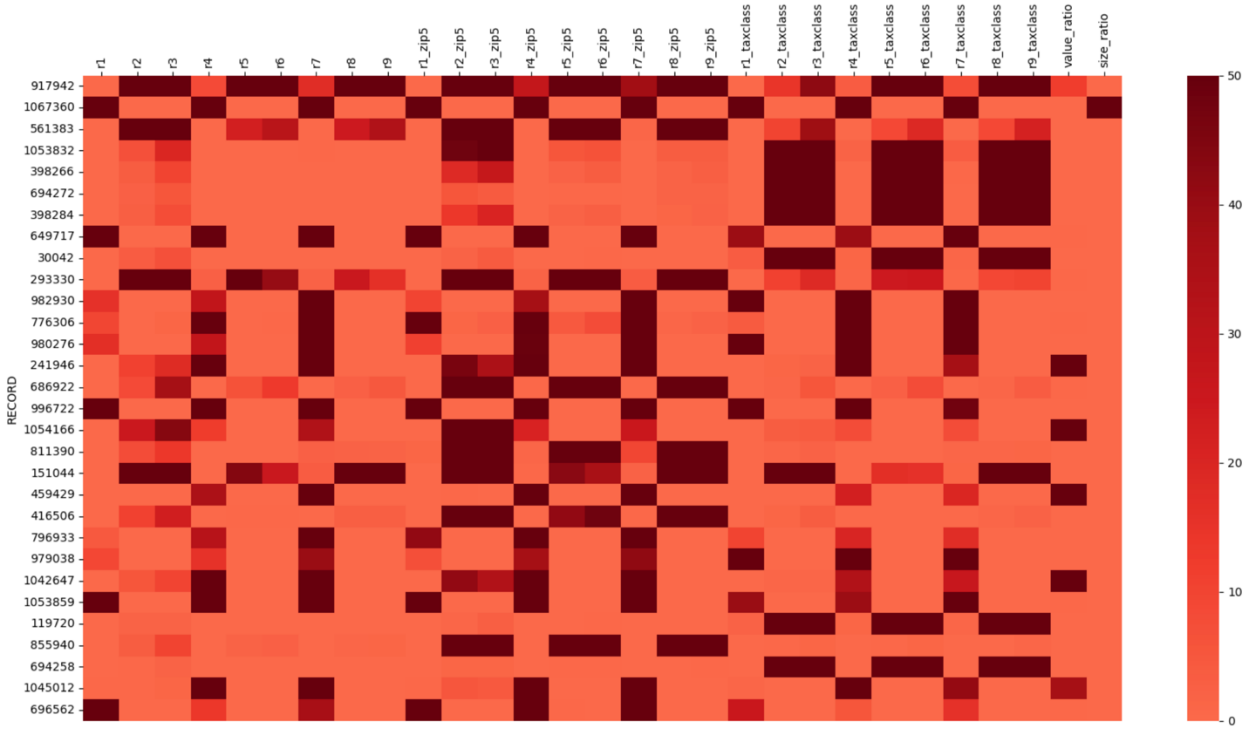
S_1 = LTFRONT * LTDEPTH
S_2 = BLDFRONT * BLDDEPTH
$S_3 = S_2$ * STORIES

- Step 2: Ratio Comparisons
    - For each property record, several ratios are computed to assess the proportionality and relative scaling of the property values:

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

    - Ratios help in understanding how the property values relate to the size and other dimensional attributes. These ratios are especially useful to identify outliers where the property values do not seem proportionate to their physical dimensions.
- Step 3: Z-Score Standardization
    - The variables and ratios are standardized using Z-scores, which measure how many standard deviations an element is from the mean. This standardization is critical in identifying how unusual a value is within the context of the dataset.

- Step 4: Heatmap Visualization
    - Heatmaps are generated to visualize the standardized scores (Z-scores) of the variables across the top records. Heatmaps provide a color-coded method of quickly identifying high and low values across multiple variables and records.
    - The heat intensity in a heatmap corresponds to the degree of standard deviation from the mean, with extreme values highlighted in contrasting colors (see figure below)



- Step 5: Focused Investigation Based on Heatmaps
    - Specific patterns observed in the heatmap can guide further investigation. For example, a consistently high Z-score in certain variables across multiple records might suggest systematic issues or anomalies that warrant deeper analysis.
    - Individual records showing extreme values in unexpected variables can be flagged for case-by-case review to understand the anomaly's nature.
- Step 6: Contextual Group Analysis
    - Properties are also grouped by relevant categories such as ZIP code or tax class to calculate average scores for these groups. This helps in identifying whether an anomaly is truly unique to a property or indicative of broader regional or categorical trends.

- Example Case Study Approach 1:
  - **Record: 776306**

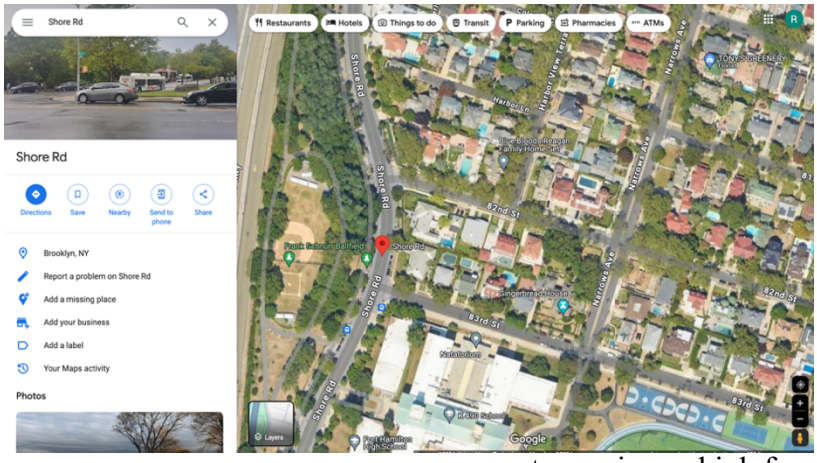
Owner	Tony Chen	LTFRONT	6
Address	SHORE ROAD	LTDEPTH	1
FULLVAL	0	BLDFRONT	0
AVLAND	0	BLDDEPTH	0
AVTOT	0	STORIES	1
BLDGCL	Q9 – miscellaneous outdoor recreation facility		

  - **Anomalies:**

variables	R4	R7	R1_zip5	R4_zip5	R7_zip5	r4_taxclass	r7_taxclass
z-score	181.45	165.73	79.34	423.66	292.50	79.00	72.20

  - **Missing values:**
    ZIP, EXMPTCL, EXCD1, AVLAND2, AVTOT2 , EXLAND2, EXTOT2, EXCD2

  - **For record 776306**, owned by Tony Chen and located at Shore Road, several anomalies and missing values are evident. The property details show that FULLVAL, AVLAND, AVTOT, BLDFRONT, and BLDDEPTH are all listed as 0, which is highly unusual and suggests incorrect data entry. Additionally, building class is Q9 – miscellaneous outdoor recreation facility, but this place is in a residential area. Additionally, the disproportionate dimensions with LTFONT at 6 and LTDEPTH at 1 are abnormal. The high z-scores for variables like R4, R7, R1_zip5, R4_zip5 , r7_zip5, r4_taxclass, r7_taxclass indicate that these values are extreme outliers, further suggesting errors and which could be the reason why this property receives a high fraud score. Moreover, critical information such as ZIP, EXMPTCL, and EXCD1 is missing.



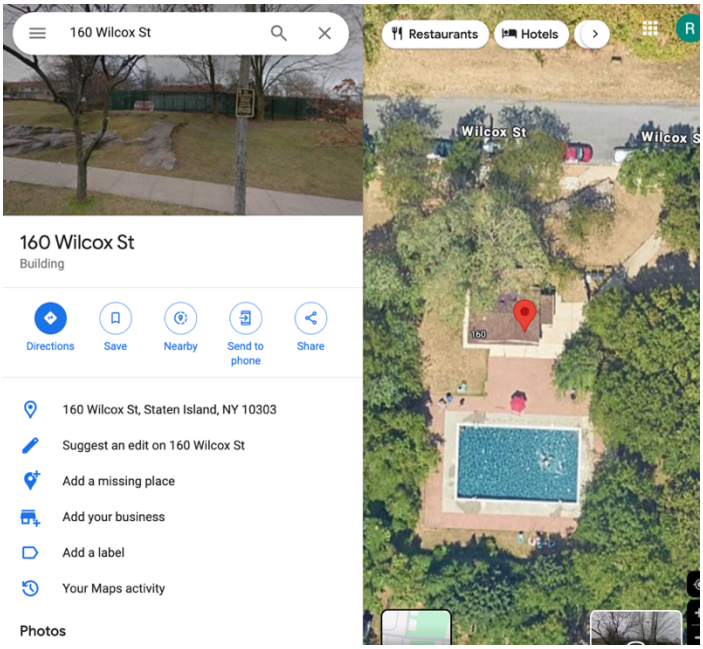
The map image of the property shows it is in a built-up residential area, contradicting the recorded values of 0. This evidence points to significant inaccuracies in the property record that require thorough review and correction.

- Example Case Study Approach 2:
  - Record: 980276

Owner	Woodmont West HOA Inc	LTFRONT	279
Address	160 Wilcox Street	LTDEPTH	190
FULLVAL	3670	BLDFRONT	15
AVLAND	89	BLDDEPTH	30
AVTOT	92	STORIES	1
BLDGCL	Z0 – Tennis court, pool, shed, etc.		

  - Anomalies:

variables	R4	R7	R1_zip5	R4_zip5	R7_zip5	R1_taxclass	R4_taxclass	r7_taxclass
z-score	27.95	67.07	10.98	49.29	216.09	216.10	225.05	344.28



  - Missing values: EXLAND, EXTOT, EXCD1, EXLAND2, EXTOT2, EXCD2
  - For record 980276, owned by Woodmont West HOA Inc. and located at 160 Wilcox Street, there are several notable anomalies and missing values. The property has FULLVAL, AVLAND, AVTOT values that are significantly lower than expected, with scores of 3670, 89, and 92 respectively, which are unusual for a property of its size and location. The high z-scores for variables such as R4, R7, R1_zip5, R4_zip5, R7_zip5, R1_taxclass, R4_taxclass, r7_taxclass indicate extreme outliers, suggesting potential errors in the data. Moreover, critical information such as EXLAND, EXTOT, and EXCD1 is missing, which further complicates accurate assessment. The map image shows the property in a developed residential area, contradicting the low values recorded. This evidence highlights significant inaccuracies in the property record, necessitating a thorough review and correction to ensure accurate property assessment and records.

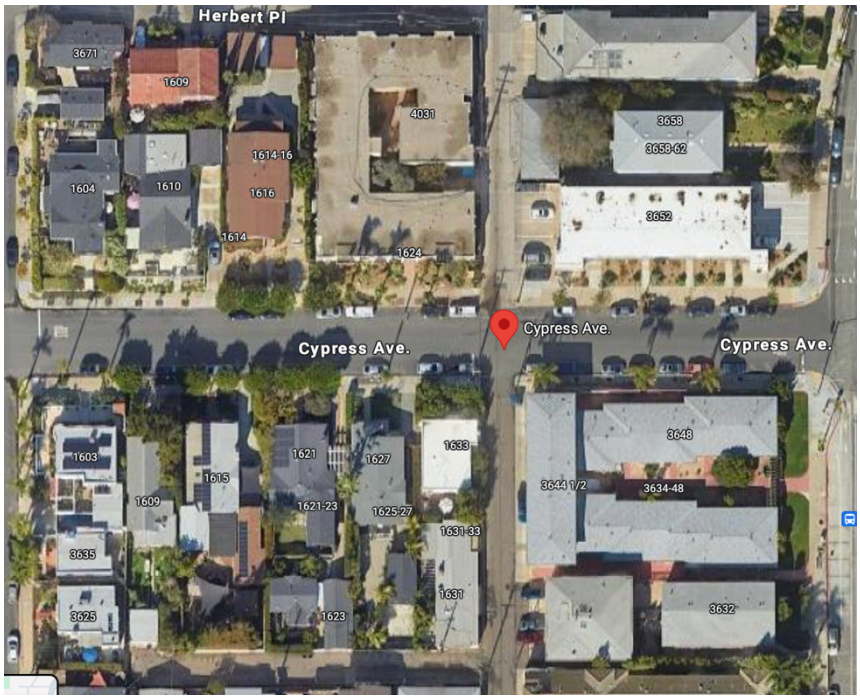
- Example Case Study Approach 3:
  - Record: 686922

Owner	WILLIAM J FREED	LTFRONT	437
Address	CYPRESS AVENUE	LTDEPTH	421
FULLVAL	25100000	BLDFRONT	0
AVLAND	11295000	BLDDEPTH	0
AVTOT	11295000	STORIES	NaN
BLDGCL	Z8 - cemetery		

  - Anomalies:

variables	R3	R2_zip5	R3_zip5	R5_zip5	R6_zip5	R8_zip5	R9_zip5
z-score	36.87	145.08	215.87	229.74	300.27	148.53	211.37

  - Missing values:
    EASEMENT, EXT, STORIES, EXCD2



  - For property record 980276 at Cypress Avenue, several anomalies suggest potential fraud. The property has unusually large lot dimensions (LTFRONT 437, LTDEPTH 421) and very high values (FULLVAL 25,100,000; AVLAND and AVTOT 11,295,000). However, both building frontage and depth are recorded as 0, and the number of stories is missing, which is highly improbable for such a valuable property. The high z-scores for several variables (R3, R2_zip5, R3_zip5, etc.) indicate these values are extreme outliers compared to other properties, suggesting potential errors or data manipulation. Missing critical information, such as easement, extension, and additional exemption codes, further raises suspicion. The map image of a built-up area contradicts the zero building dimensions, indicating potential fraudulent activity. Additionally, the property is classified as Z8 - cemetery, but this place is in a residential area and no sign of cementery.

- Example Case Study Approach 4:
  - Record: 1045012

Owner	LINDA VITALONE	LTFRONT	726
Address	ERIKA LOOP	LTDEPTH	1174
FULLVAL	2300000	BLDFRONT	0
AVLAND	129	BLDDEPTH	0
AVTOT	129	STORIES	NaN
BLDGCL	V0 – zoned residential; not Manhattan		

  - Anomalies:

variables	R4	R7	R4_zip5	R7_zip5	R4_taxclass	R7_taxclass
z-score	98.42	172.43	163.12	133.68	63.71	41.01

  - Missing values:
    EASEMENT, EXT, STORIES, ZIP, AVLAND2, AVTOT2, EXLAND2, EXTOT2, EXCD2

  - For property record 1045012 at Erika Loop, several anomalies suggest potential fraud. The property has extremely large lot dimensions (LTFRONT 726, LTDEPTH 1174) and a high FULLVAL of 2,300,000, yet the building frontage and depth are both recorded as 0, and the number of stories is missing, which is unrealistic for such a valuable property. The high z-scores for variables like R4, R7, and others indicate extreme outliers, suggesting inaccuracies. Additionally, critical information such as easement, extension, stories, zip code, and additional assessed values is missing. The map image of a developed area contradicts the zero building dimensions. These discrepancies strongly suggest inaccuracies in the property record, indicating potential fraudulent activity.

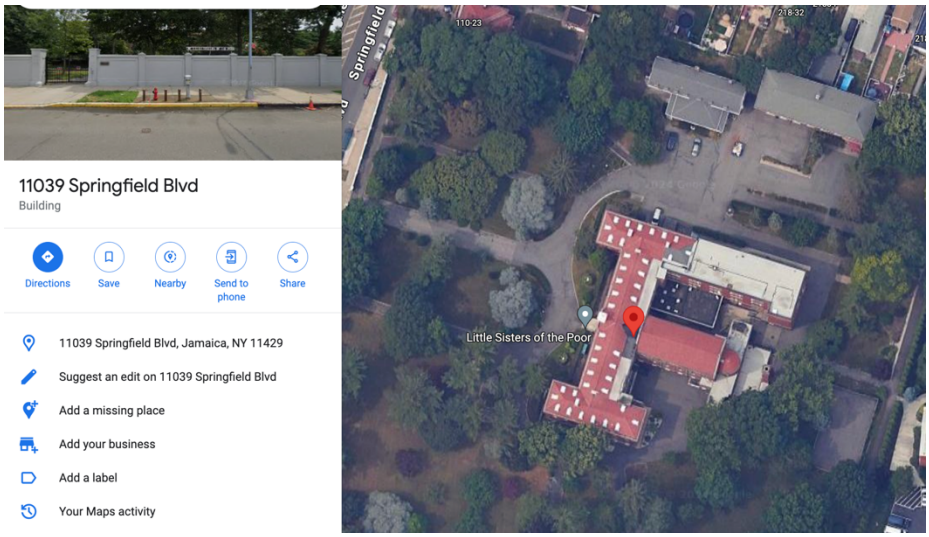
- Example Case Study Approach 5:
  - Record: 855940

Owner	LITTLE SISTERS OF THE POOR	LTFRONT	402
Address	110-39 SPRINGFIELD BLVD	LTDEPTH	545
FULLVAL	10400000	BLDFRONT	0
AVLAND	3330000	BLDDEPTH	0
AVTOT	4680000	STORIES	NaN
BLDGCL	M4 - convent		

  - Anomalies:

variables	R2_zip5	R3_zip5	R5_zip5	R6_zip5	R8_zip5	R9_zip5
z-score	96.97	110.65	122.51	135.37	116.39	119.78

  - Missing values:
    EASEMENT, EXT, STORIES, EXMPTCL, EXCD2



  - For property record 866940 at 110-39 Springfield Blvd, several anomalies suggest potential fraud. The property has large lot dimensions (LTFRONT 402, LTDEPTH 545) and a high FULLVAL of 10,400,000, with AVLAND and AVTOT valued at 3,330,000 and 4,680,000 respectively. However, both building frontage (BLDFRONT) and depth (BLDDEPTH) are recorded as 0, and the number of stories is missing, which is highly improbable for such a high-value property. High z-scores for multiple variables (R2_zip5, R3_zip5, etc.) indicate these values are extreme outliers. Additionally, critical information such as easement, extension, stories, exemption codes, and additional assessed values is missing. The map image shows a substantial building, contradicting the zero building dimensions. These discrepancies strongly suggest inaccuracies in the property record, indicating potentially fraudulent activity

Summary

In this project, we conducted a detailed analysis of the New York Property Data, a dataset comprising over one million records provided by the Department of Finance. Our primary aim was to identify anomalies within property valuations and assessments, key components that influence the city's property tax liabilities. By leveraging statistical and machine learning techniques, we reduced the complexity of the data, enabling us to effectively identify and focus on records that exhibited unusual or inconsistent patterns compared to the norm.

The analysis yielded several key insights, including the identification of specific properties with irregular valuation metrics, which could potentially indicate erroneous assessments or fraudulent activities. These findings are crucial for the city's finance department as they provide a basis for revisiting and refining assessment protocols, thereby ensuring more accurate and fair property tax evaluations.

To enhance the effectiveness of the anomaly detection process, the algorithm can be adjusted based on expert feedback. This involves recalibrating the weight and influence of certain variables that are found to be more indicative of anomalies, as well as excluding records that, upon expert review, are deemed to be outliers due to legitimate reasons such as unique property features or clerical errors in data entry. Incorporating expert feedback helps in fine-tuning the model to reduce false positives and improve its overall predictive accuracy.

Overall, the project offers valuable methodologies and tools that can be utilized by municipal authorities to maintain the integrity and accuracy of property valuations. By continuously integrating feedback and making necessary adjustments, the model remains robust and adaptable to changing data characteristics and external conditions, ensuring its long-term utility and relevance in urban financial management and planning.

Appendix

Data Quality Report

1. Data Description

   The dataset is titled **New York Property Data**, which contains comprehensive property valuation and assessment information. The data was collected by the Department of Finance and encompasses **1,070,994 records across 32 fields**, which include both categorical and numerical data types. This dataset is essential for the annual real estate assessment process, which ultimately determines property tax liabilities for various properties within New York City.

2. Summary Tables

   The dataset includes numerical fields such as 'LTFRONT', 'LTDEPTH', 'STORIES', 'FULLVAL', 'AVLAND', 'AVTOT', 'EXLAND', 'EXTOT', 'BLDFRONT', 'BLDDEPTH', 'AVLAND2', 'AVTOT2', 'EXLAND2', and 'EXTOT2'.

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	LTFRONT	numeric	1070994	100.0%	169108	0.00	9999.00	36.64	74.03	0.00
1	LTDEPTH	numeric	1070994	100.0%	170128	0.00	9999.00	88.86	76.40	100.00
2	STORIES	numeric	1014730	94.7%	0	1.00	119.00	5.01	8.37	2.00
3	FULLVAL	numeric	1070994	100.0%	13007	0.00	6150000000.00	874264.51	11582425.58	0.00
4	AVLAND	numeric	1070994	100.0%	13009	0.00	2668500000.00	85067.92	4057258.16	0.00
5	AVTOT	numeric	1070994	100.0%	13007	0.00	4668308947.00	227238.17	6877526.09	0.00
6	EXLAND	numeric	1070994	100.0%	491699	0.00	2668500000.00	36423.89	3981573.93	0.00
7	EXTOT	numeric	1070994	100.0%	432572	0.00	4668308947.00	91186.98	6508399.78	0.00
8	BLDFRONT	numeric	1070994	100.0%	228815	0.00	7575.00	23.04	35.58	0.00
9	BLDDEPTH	numeric	1070994	100.0%	228853	0.00	9393.00	39.92	42.71	0.00
10	AVLAND2	numeric	282726	26.4%	0	3.00	2371005000.00	246235.72	6178951.64	2408.00
11	AVTOT2	numeric	282732	26.4%	0	3.00	4501180002.00	713911.44	11652508.34	750.00
12	EXLAND2	numeric	87449	8.2%	0	1.00	2371005000.00	351235.68	10802150.91	2090.00
13	EXTOT2	numeric	130828	12.2%	0	7.00	4501180002.00	656768.28	16072448.75	2090.00

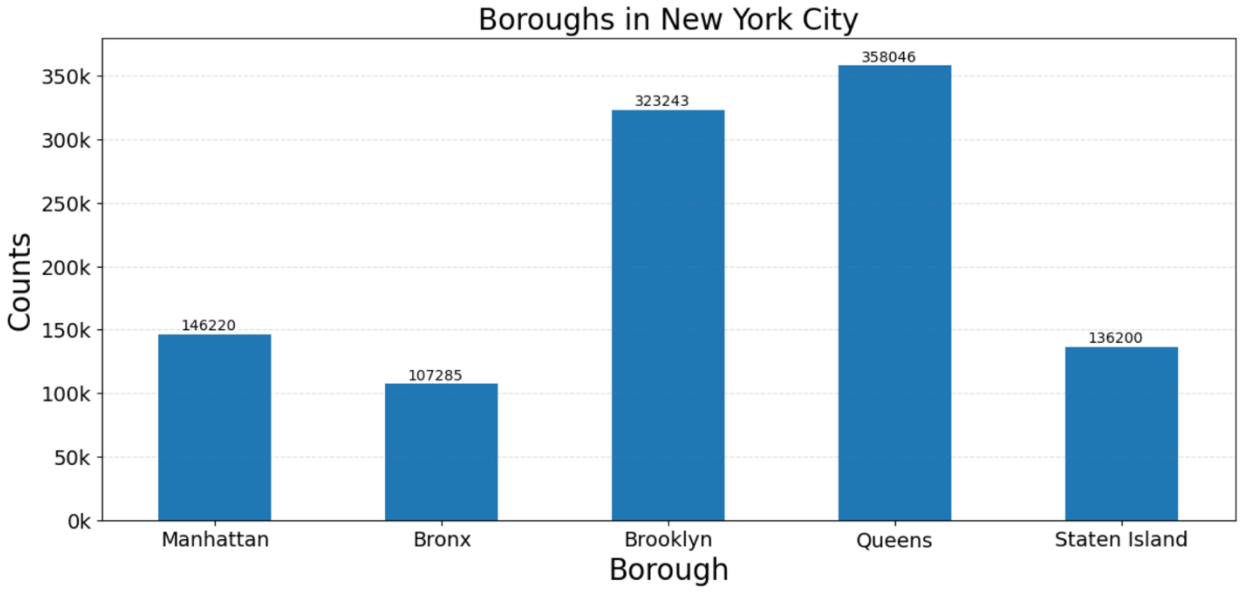
   It also features several categorical fields including 'RECORD', 'BBLE', 'BORO', 'BLOCK', 'LOT', 'EASEMENT', 'OWNER', 'BLDGCL', 'TAXCLASS', 'EXT', 'EXCD1', 'STADDR', 'ZIP', 'EXMPTCL', 'EXCD2', 'PERIOD', 'YEAR', and 'VALTYPE'.

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	RECORD	categorical	1070994	100.0%	0	1070994	1
1	BBLE	categorical	1070994	100.0%	0	1070994	1000010101
2	BORO	categorical	1070994	100.0%	0	5	4
3	BLOCK	categorical	1070994	100.0%	0	13984	3944
4	LOT	categorical	1070994	100.0%	0	6366	1
5	EASEMENT	categorical	4636	0.4%	0	12	E
6	OWNER	categorical	1039249	97.0%	0	863347	PARKCHESTER PRESERVAT
7	BLDGCL	categorical	1070994	100.0%	0	200	R4
8	TAXCLASS	categorical	1070994	100.0%	0	11	1
9	EXT	categorical	354305	33.1%	0	3	G
10	EXCD1	categorical	638488	59.6%	0	129	1017.00
11	STADDR	categorical	1070318	99.9%	0	839280	501 SURF AVENUE
12	ZIP	categorical	1041104	97.2%	0	196	10314.00
13	EXMPTCL	categorical	15579	1.5%	0	14	X1
14	EXCD2	categorical	92948	8.7%	0	60	1017.00
15	PERIOD	categorical	1070994	100.0%	0	1	FINAL
16	YEAR	categorical	1070994	100.0%	0	1	2010/11
17	VALTYPE	categorical	1070994	100.0%	0	1	AC-TR

3. Visualization of Each Field
    - Field Name: RECORD
      Description: The RECORD field in the dataset uniquely identifies each entry, ensuring that all 1,070,994 records are distinctly cataloged without any missing values, facilitating straightforward data retrieval and management.

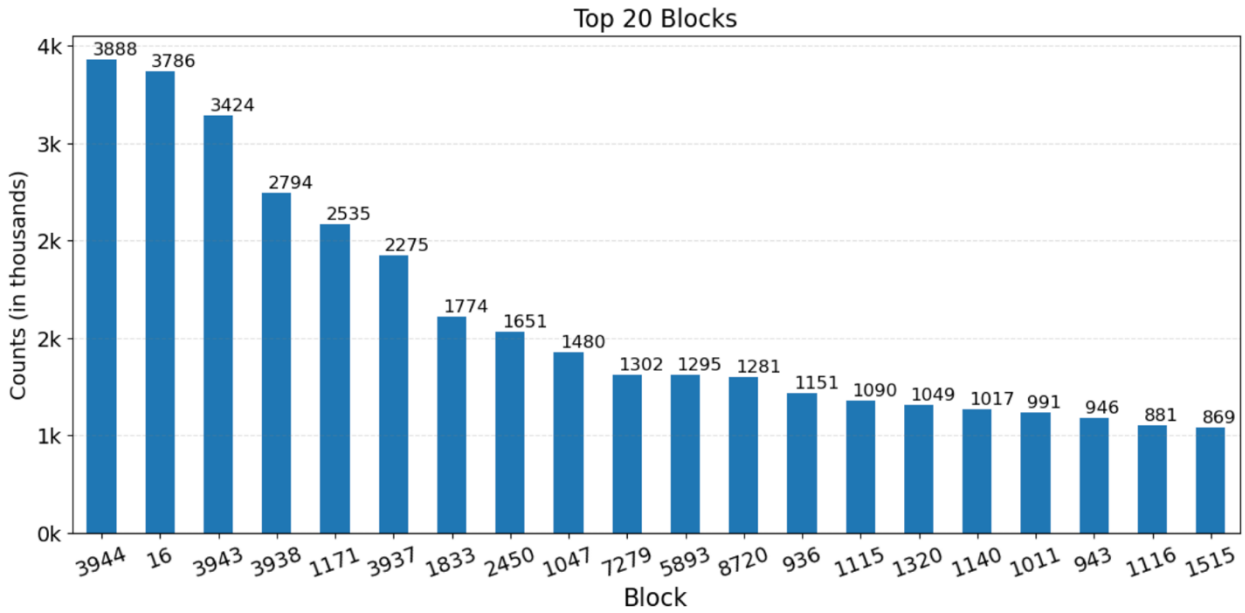
    - Filed Name: BBLE
      Description: The BBLE field in the dataset serves as a unique identifier for each property, combining the borough, block, lot, and easement code, which is critical for tracking and managing property-related information across New York City's extensive real estate database.

    - Filed Name: BORO



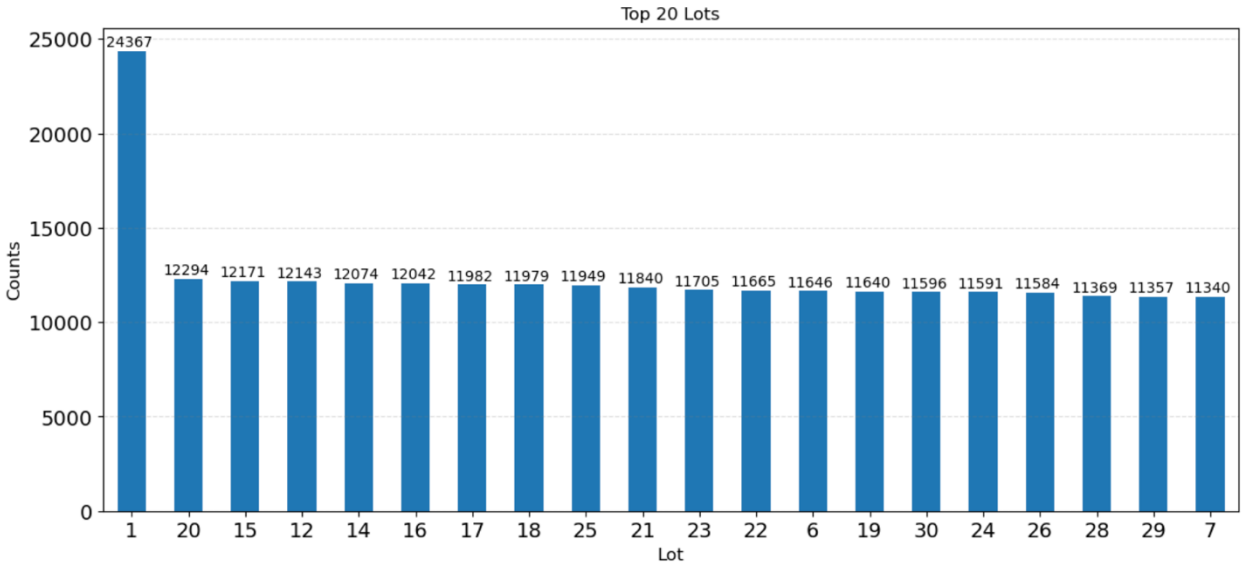
      Description: BORO represent the five boroughs of New York City. Queens has the highest number of properties with a count of 358,046, followed by Brooklyn with 323,243 properties. Manhattan, despite its prominence, has fewer properties recorded at 146,220. The Bronx and Staten Island have 107,285 and 136,200 properties respectively, indicating a varied distribution of real estate across the boroughs.

- Filed Name: BLOCK



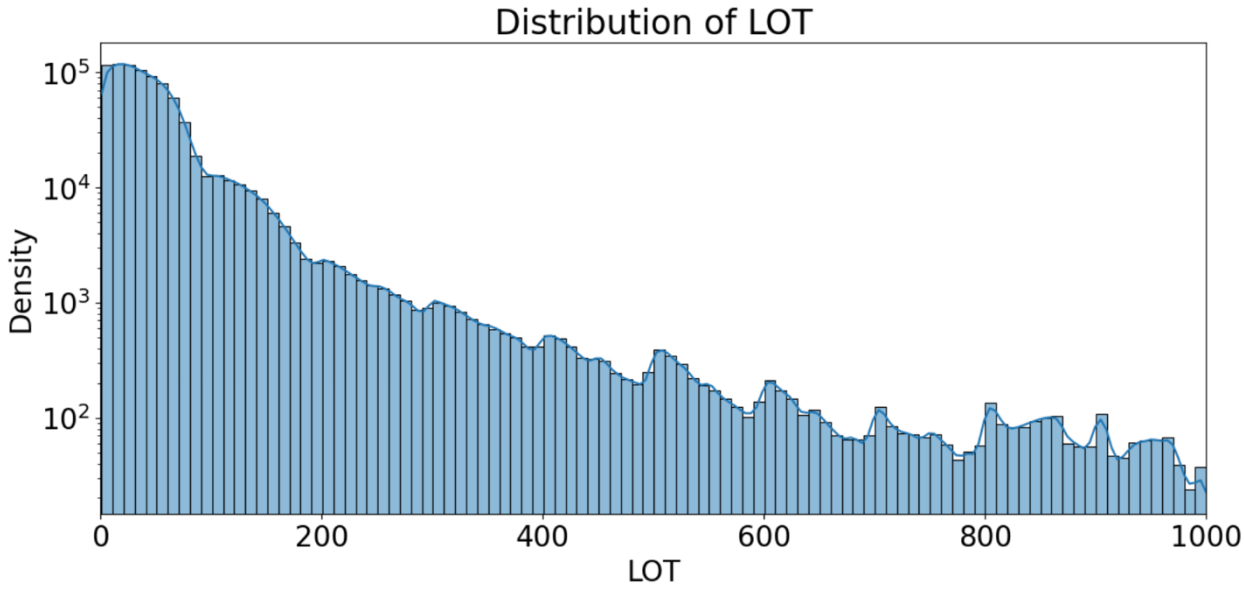
Description: The BLOCK variable in the dataset represents specific city blocks within New York City, serving as a numeric identifier that groups properties into defined segments. Each borough has its unique range of valid block numbers, for instance, Manhattan has block numbers ranging from 1 to 2255, and Queens from 1 to 16350. The variable is essential for geographical and administrative classification, helping to localize properties precisely within the vast urban landscape of the city.

- Filed Name: LOT



Description: The LOT variable in the dataset identifies specific lots within a block, representing smaller divisions of land within each property's block designation. Each lot number corresponds to a unique plot of land within its borough and block.

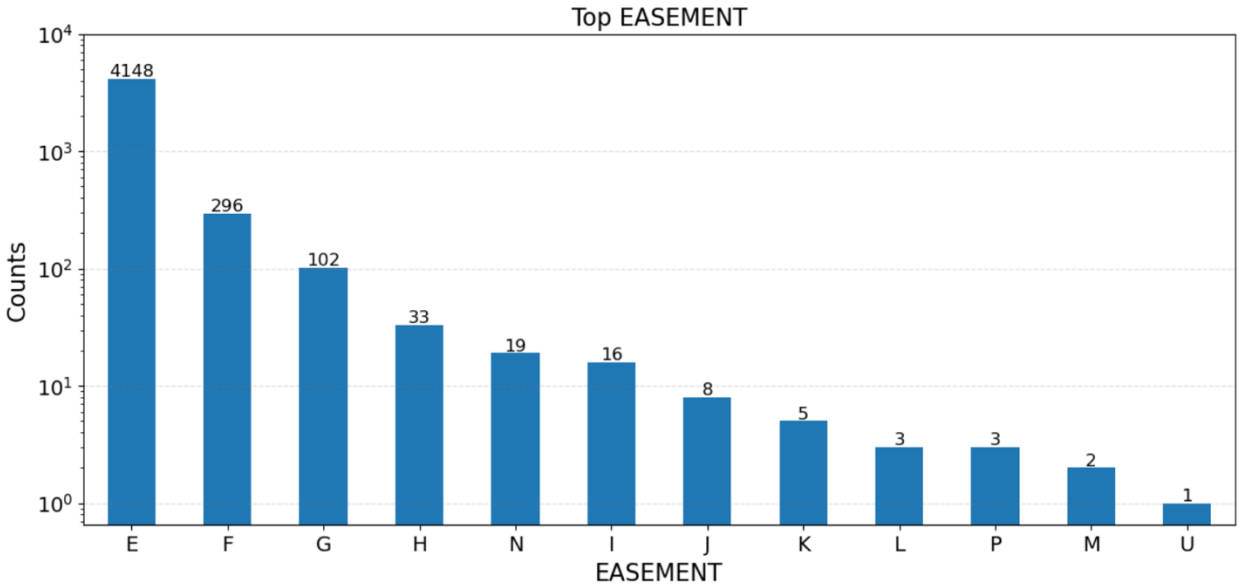
The top 20 lots in the dataset show a wide range of property counts, with the most populated lot having 24,367 properties and the least among the top 20 holding 8,869, indicating significant variation in lot usage and development density across the city.



The distribution of the LOT variable shows a skewed pattern with a high frequency of lower numbered lots, gradually tapering off but with periodic spikes across the range. This suggests that while smaller lot numbers are more

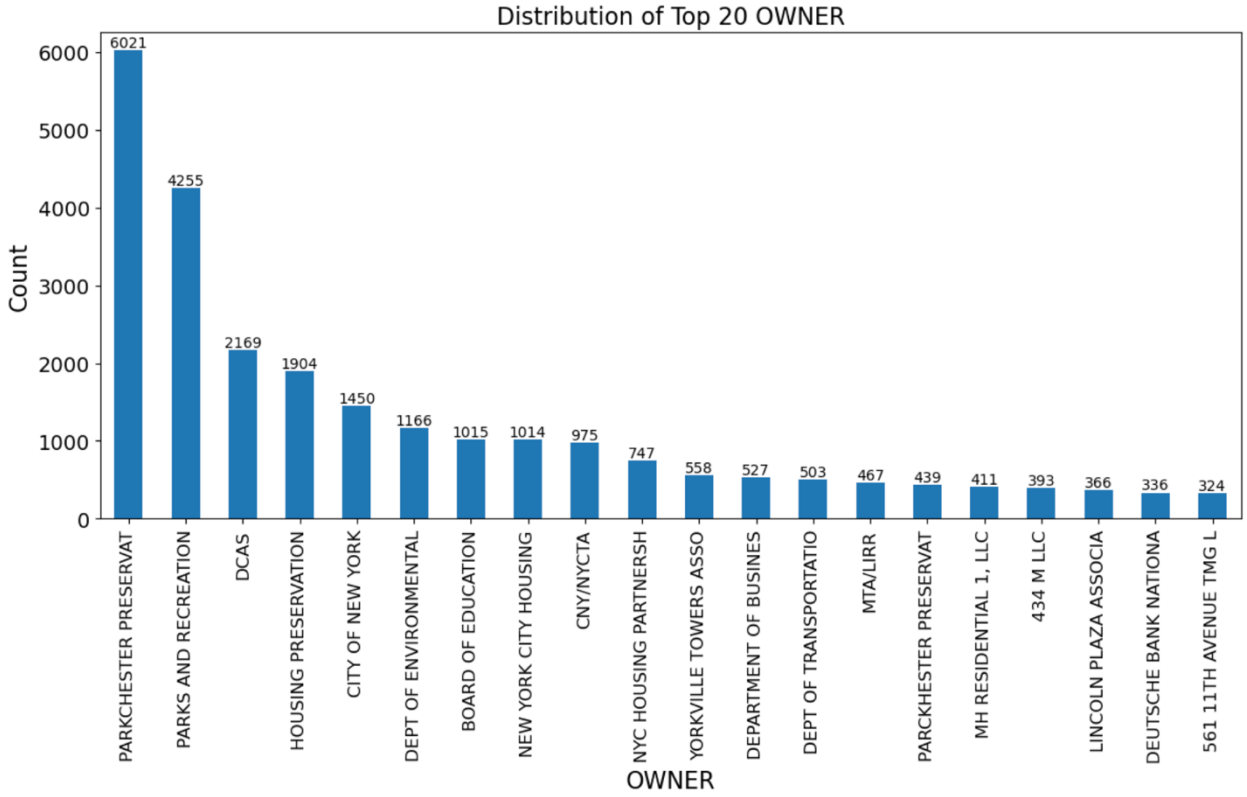
commonly registered, there are notable concentrations of property records around higher lot numbers, possibly reflecting larger developments or subdivisions within certain blocks.

- Filed Name: EASEMENT



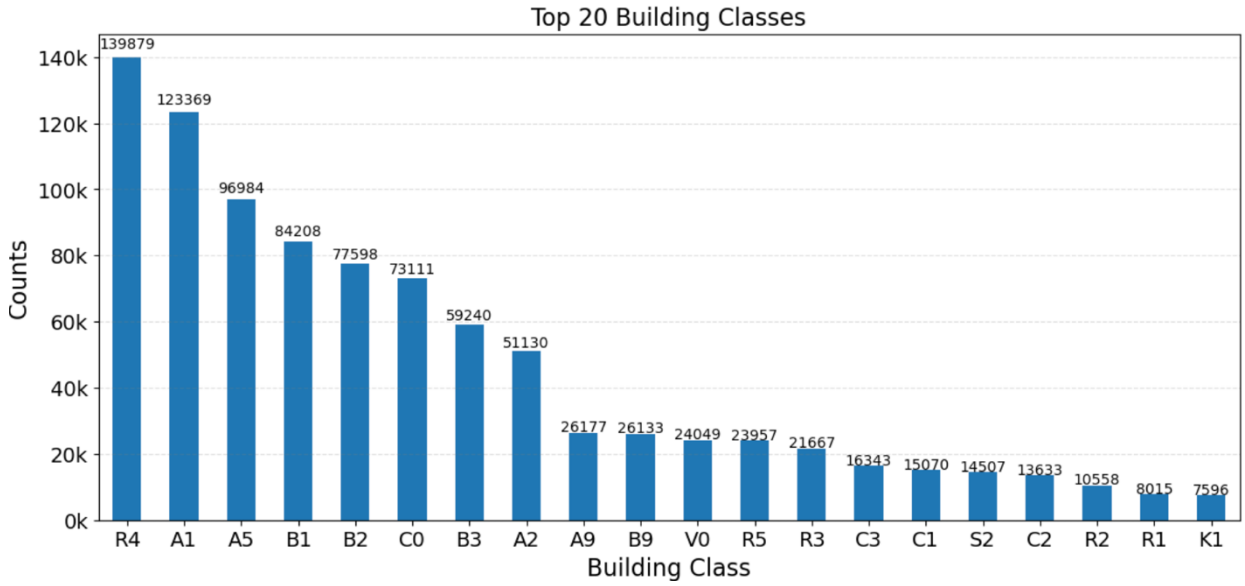
Description: The EASEMENT variable in the dataset categorizes properties based on specific rights or restrictions associated with the property's use, such as air rights, land access, or governmental use. This categorical variable includes types like 'A' for Air Easement, 'E' for Land Easement, and 'U' for properties owned by the U.S. Government, among others, each providing insights into the unique legal and physical characteristics of the lots in question.

- Filed Name: OWNER



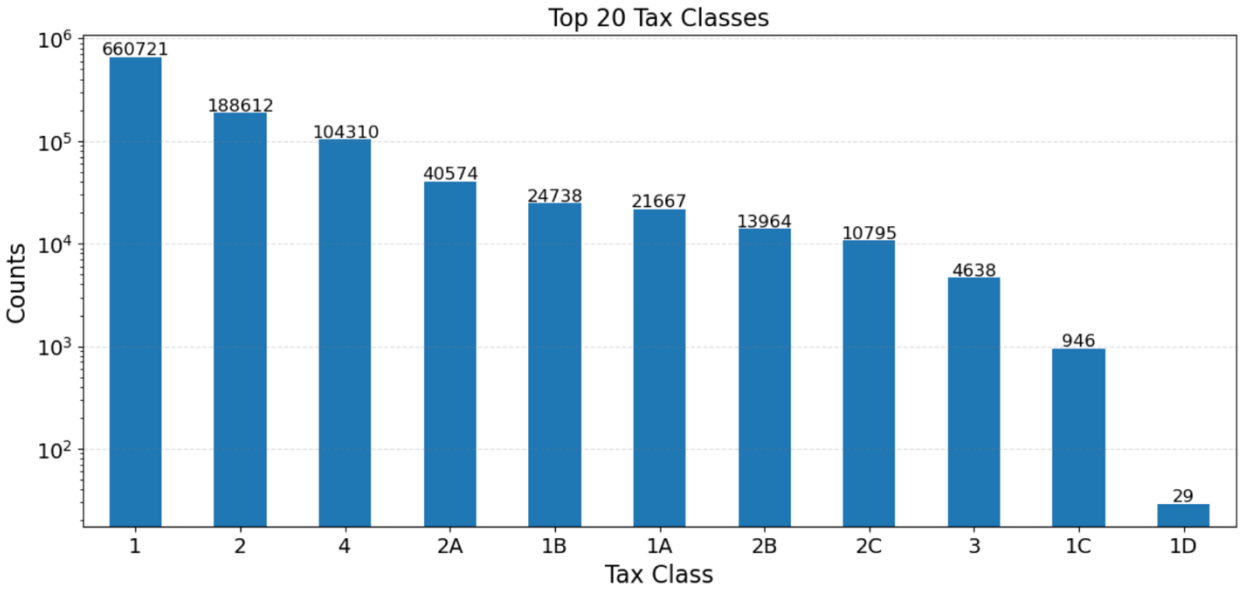
Description: The OWNER variable in the dataset specifies the name of the entity or individual that holds title to the property. This categorical variable is significant as it reflects ownership diversity across the dataset, ranging from public organizations like "PARKCHESTER PRESERVAT" and "NYC HOUSING PARTNERSH" to private entities and individuals. The distribution highlights key stakeholders in New York City's property market, with the largest counts of properties under the management of major public and private housing, educational, and governmental institutions.

- Filed Name: BLDGCL



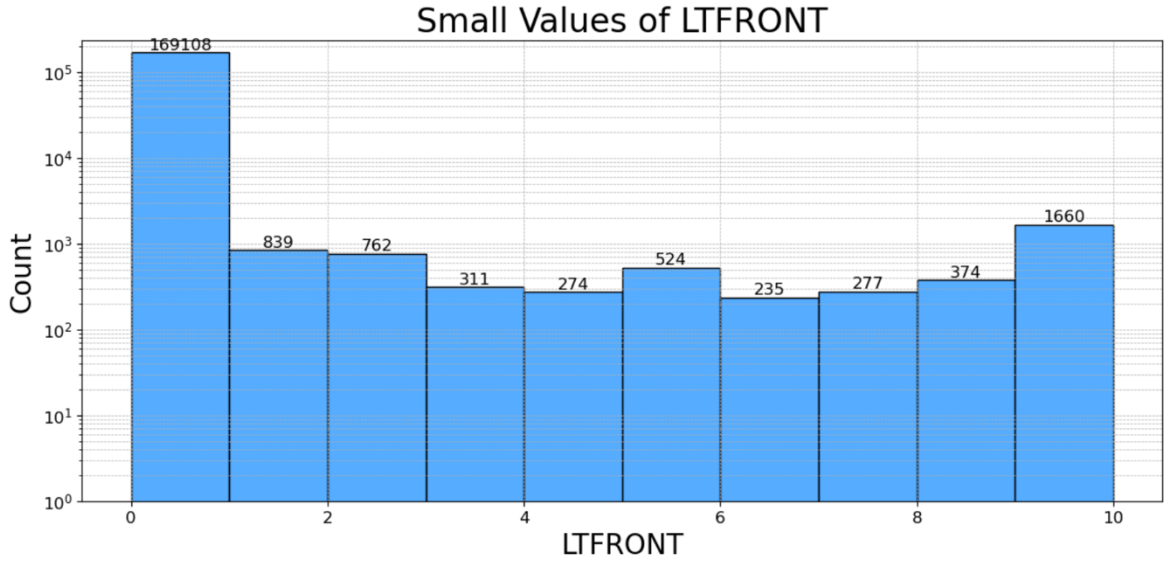
Description: The BLDGCL (Building Class) variable in the dataset categorizes properties based on their type and use, such as residential, commercial, or mixed-use buildings. Each class is represented by a code, such as R4 for residential condominiums, A1 for one-family dwellings, and C0 for walk-up apartments, among others. The distribution across the top 20 building classes highlights the diversity of building types within New York City, with the highest counts found in classes that typically represent densely populated residential areas.

- Filed Name: TAXCLASS



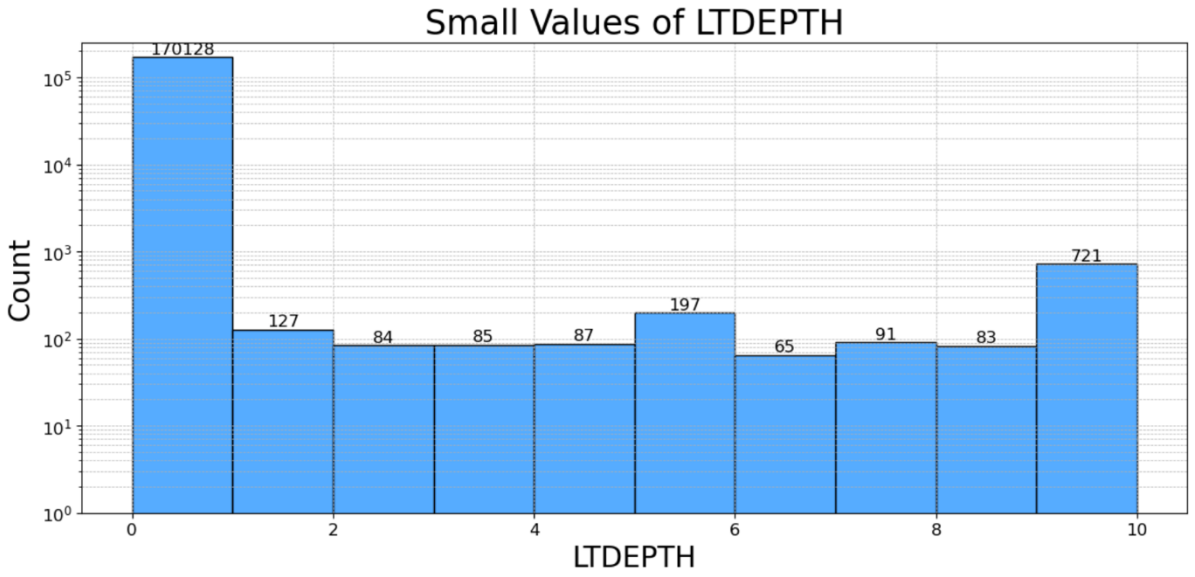
Description: The TAXCLASS variable categorizes properties based on their primary use and potential tax liability, significantly impacting how properties are assessed for tax purposes. The distribution of property types across various tax classes shows a wide range, with Class 1 (1-3 unit residences) containing the majority of properties at 660,721, and Class 2 (apartments) following with 188,612 properties. Other classes represent smaller segments, including utilities and specialized housing units, indicating varied tax assessments across the city's diverse property landscape.

- Filed Name: LTFRONT



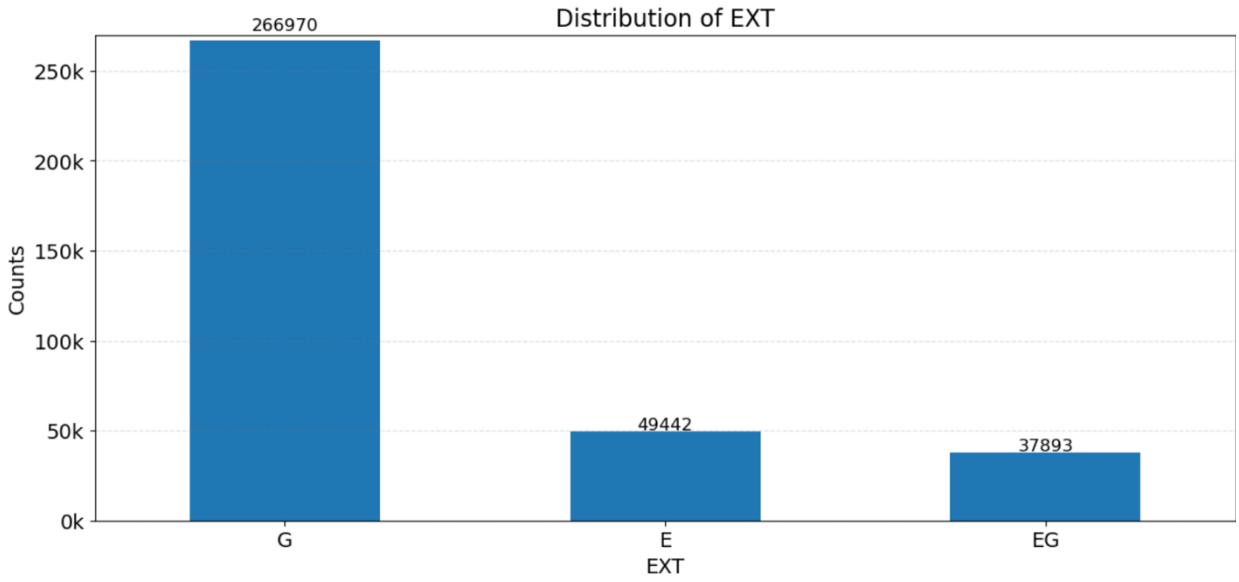
Description: The LTFRONT variable represents the width of the lot facing the street, measured in feet, and is crucial for understanding property layout and assessing property value. Analysis of both boxplot and distribution plot reveals that most LTFRONT values are concentrated within 10 feet, indicating a common urban property characteristic where lots have smaller street-facing dimensions. This is typical in densely built areas where space is at a premium. The distribution shows a skew towards smaller lot frontages, with a marked decline in occurrence as lot width increases beyond 10 feet. This visualization uses a logarithmic scale on the y-axis to better display the frequency of smaller values, enhancing the visual interpretation of data spread and concentration, with the x-axis limited to 10 feet to focus on the most common property widths.

- Filed Name: LTDEPTH



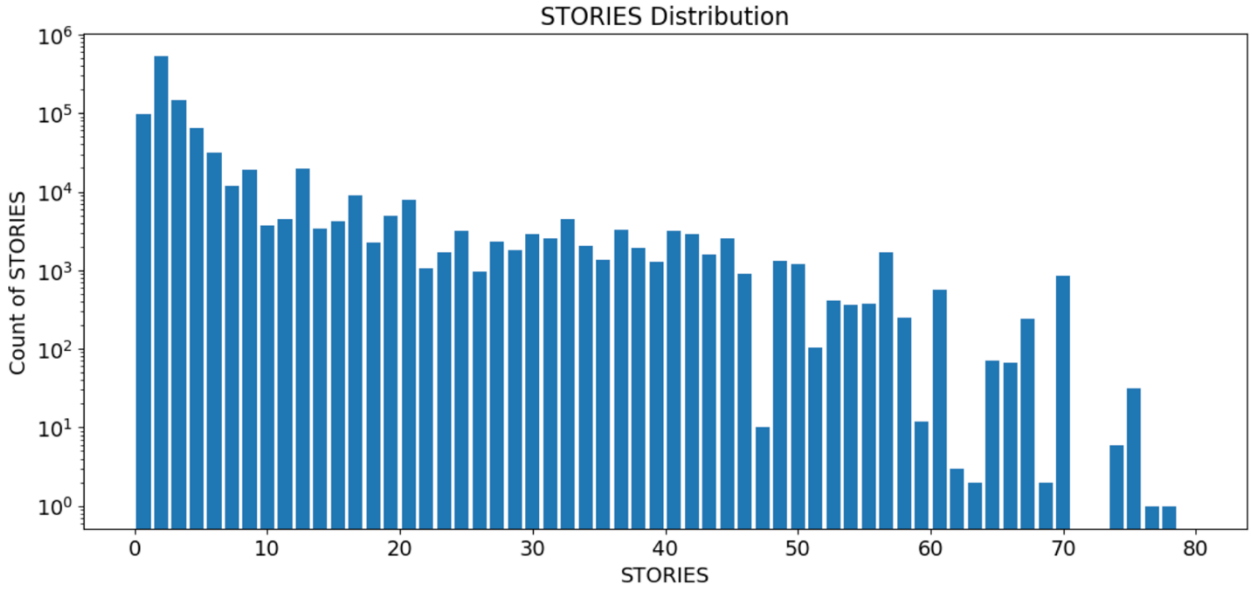
Description: The LTDEPTH variable measures the depth of a property lot in feet, from the street front to the back of the lot. Analysis of both boxplot and distribution plot shows a strong concentration of values within 10 feet, highlighting a common characteristic in densely built urban environments where space is maximized. The distribution, skewed toward smaller lot depths, reflects a notable range in property sizes within New York City. This visualization specifically limits the x-axis to 10 feet to focus on the most prevalent measurements and employs a logarithmic scale on the y-axis to clearly illustrate the frequency of smaller lot depths.

- Filed Name: EXT



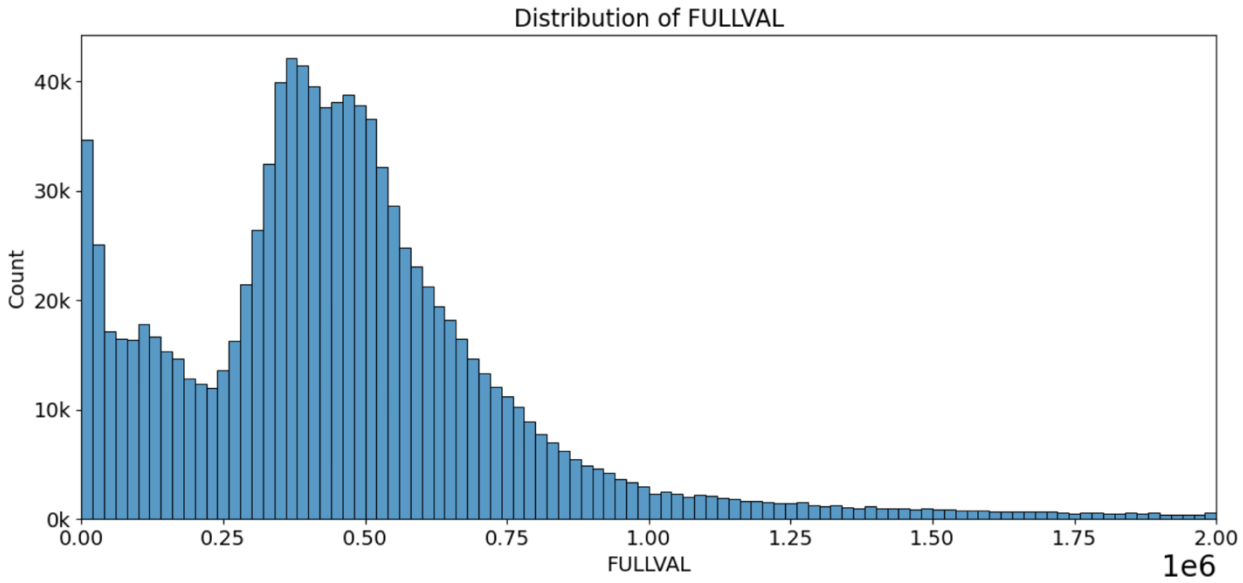
Description: The EXT variable, known as the Extension Indicator, categorizes properties based on whether they have an extension, using the codes: 'G' for no extension, 'E' for some extension, and 'EG' for extensive extension. The distribution shows a significant majority of properties without extensions (G), followed by a smaller proportion with some extension (E), and even fewer properties that have extensive extensions (EG). This indicates that extensions are less common in the property dataset.

- Filed Name: STORIES



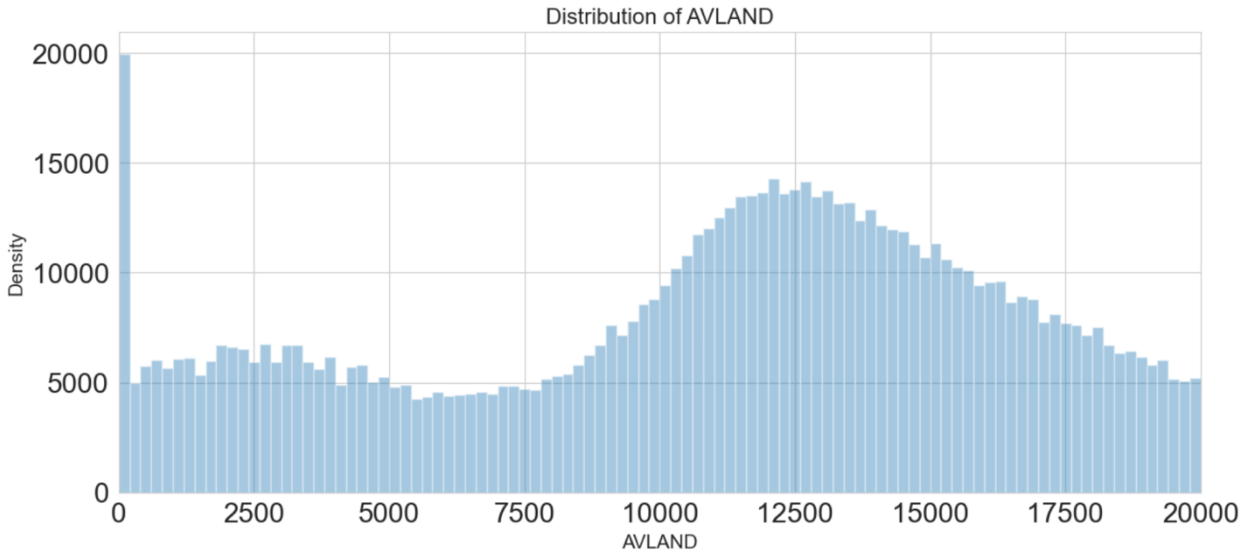
Description: The STORIES variable represents the number of stories in a building, as recorded in the dataset. The distribution demonstrates a broad range of building heights, with a notable decrease in frequency as the number of stories increases, suggesting that taller buildings are less common in the dataset. Buildings with fewer stories are more prevalent, indicating a higher frequency of low-rise constructions.

- Filed Name: FULLVAL



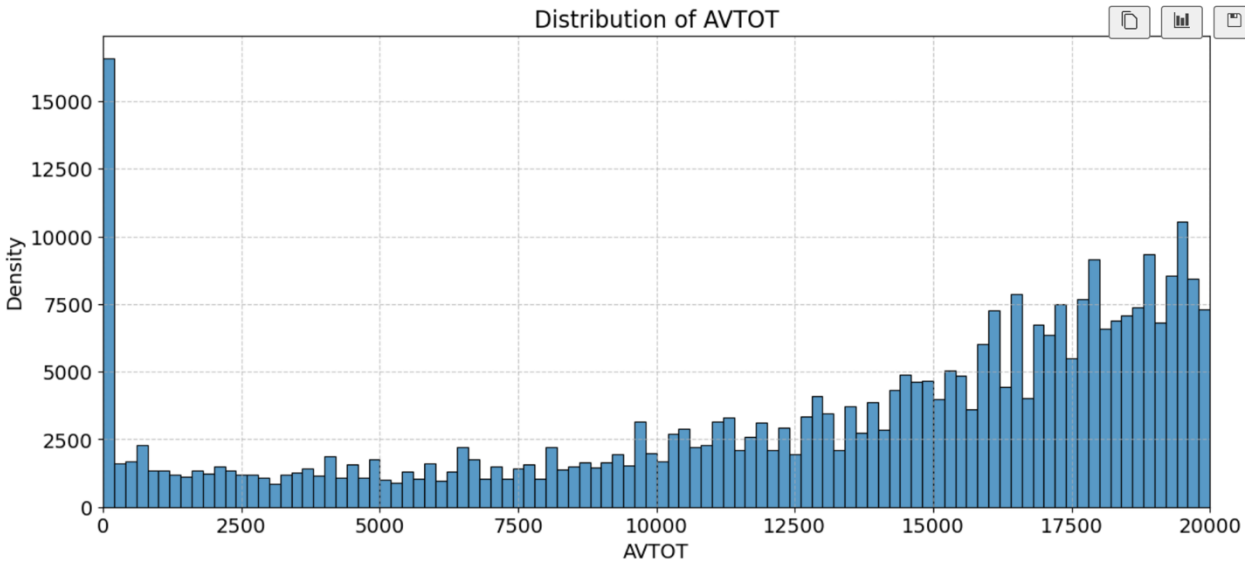
Description: The FULLVAL variable represents the market value of properties as assessed. The distribution shows a right-skewed pattern where most properties are valued under $1 million, which is typical for the dataset, indicating a higher concentration of lower-valued properties. Initial examination of a boxplot reveals that the bulk of data points cluster below $2 million. Consequently, the distribution's x-axis is limited to $2 million to highlight the area where the majority of values lie, providing a clearer view of the distribution characteristics.

- Filed Name: AVLAND



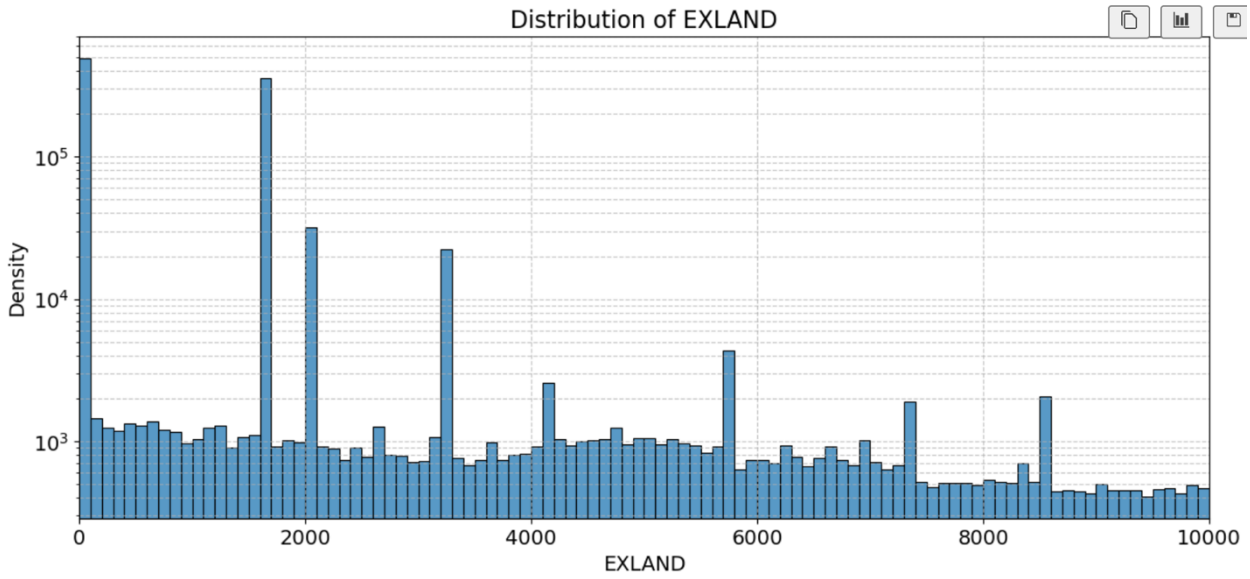
Description: The variable "AVLAND" represents the Actual Land Value of properties in the dataset. It measures the assessed value assigned to the land component of a property, crucial for determining property taxes. The distribution of "AVLAND" shows a peak around values under $20,000, indicating a concentration of land assessments within this range, as most values are concentrated below $20,000 according to initial boxplot analysis. The x-axis is therefore limited to $20,000 to focus on the primary range of values, and the density quickly diminishes as values increase, highlighting that higher land values are less common in the dataset.

- Filed Name: AVTOT



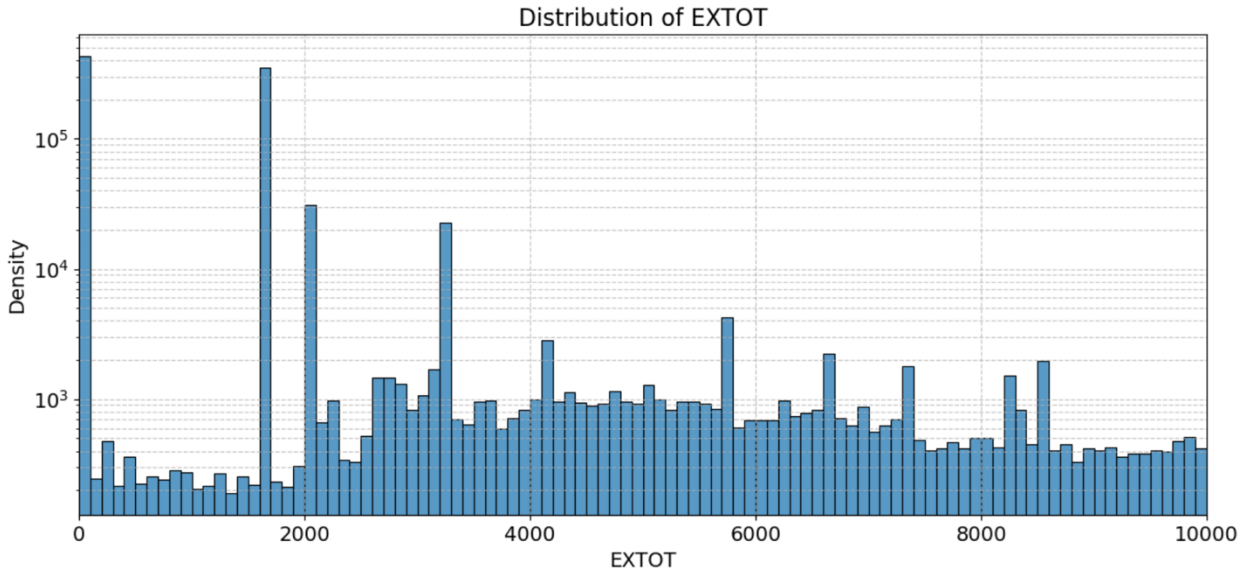
Description: The variable AVTOT, representing the Actual Total Value, shows a broad range of values across the dataset. Initial analysis with a boxplot revealed that the majority of data points are concentrated below $100,000, prompting a narrower focus in subsequent visualizations. Upon closer examination with histograms constrained to this threshold, a significant peak in density was observed around $20,000, leading to a decision to limit the x-axis to this value in the final visualization. The overall distribution shows a gradual increase in density from zero, peaks in the mid-range, and then tapers off, indicating fewer properties with higher values within this subset.

- Filed Name: EXLAND



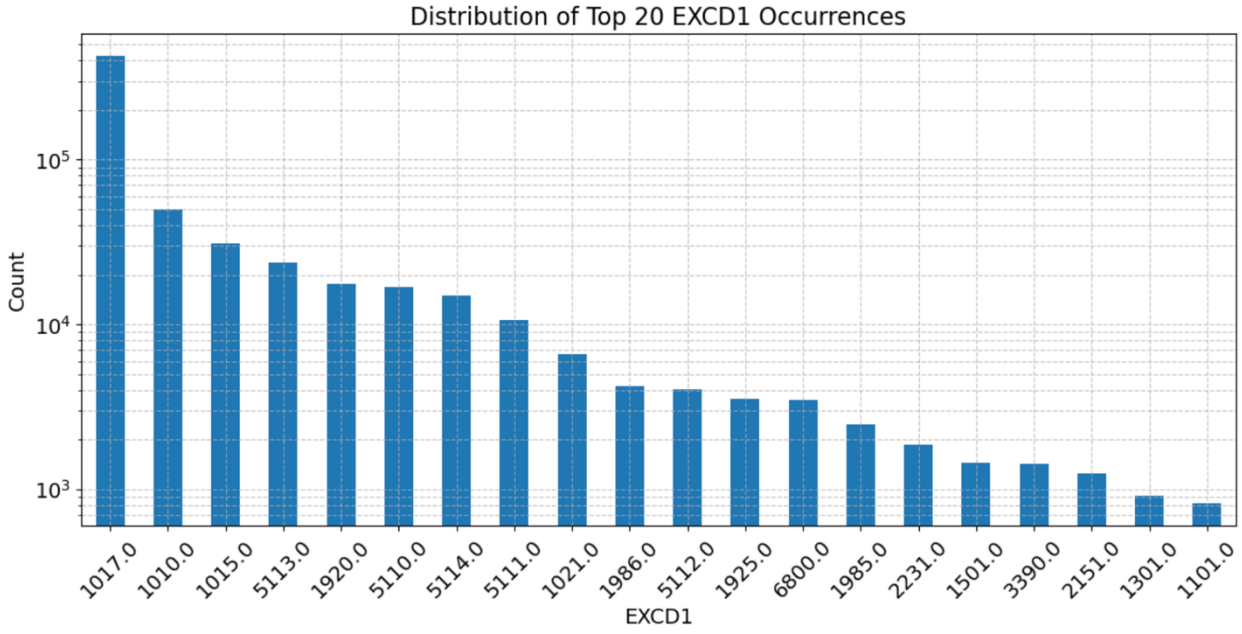
Description: EXLAND, representing "Actual Exempt Land Value," measures the value of land exempt from property taxes under various programs. Analyzing the distribution of EXLAND values reveals a dense concentration of data points within the first 10,000 units. This analysis was supported by a boxplot observation indicating that the majority of values are tightly clustered in this range. Consequently, the histogram was plotted with an x-axis limit of 10,000 to focus on the main body of the data, showing several pronounced peaks in frequency for values at lower ranges, reflecting specific exempt land valuation groups within the dataset.

- Filed Name: EXTOT



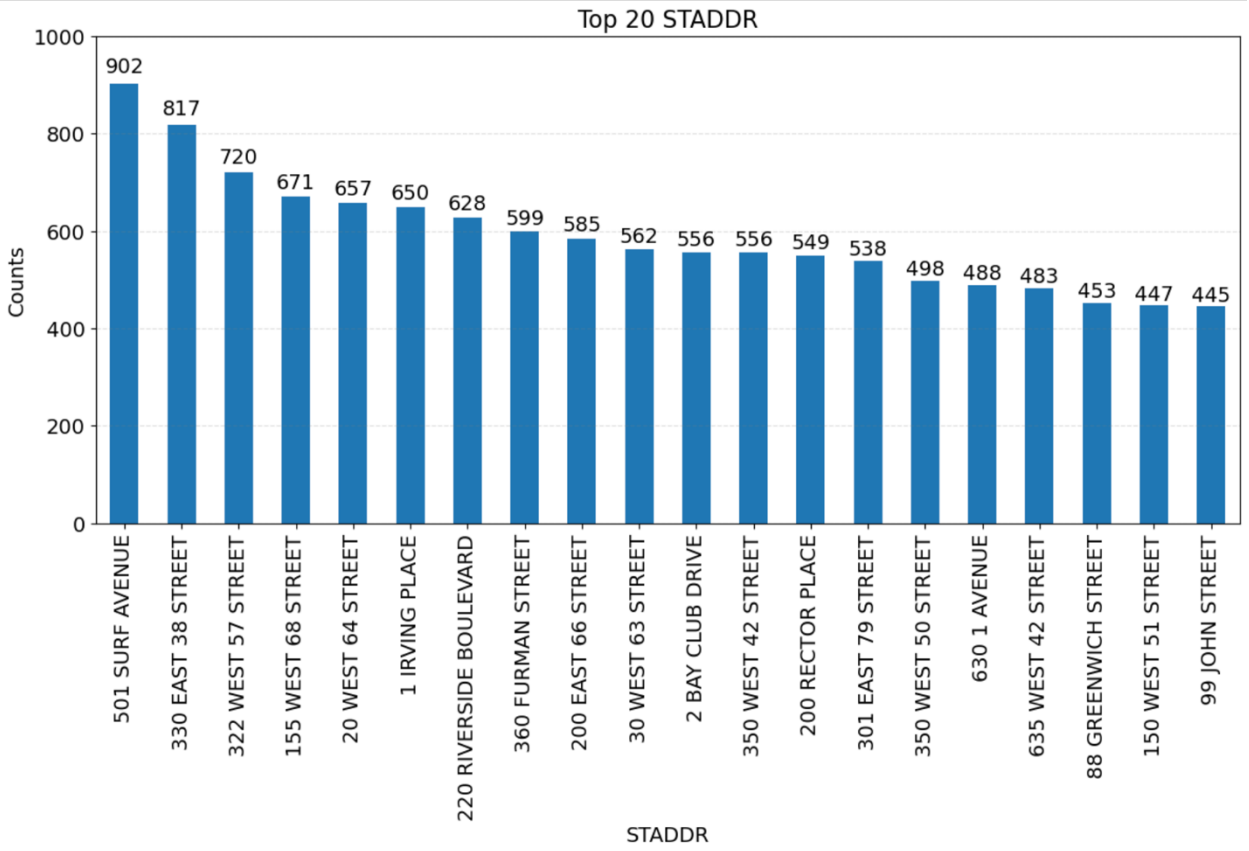
Description: EXTOT represents "Actual Exempt Land Total," indicating the total value of land exempt from taxes. A detailed analysis of EXTOT values highlighted a significant concentration of entries below 10,000, as observed from the boxplot. This informed the decision to limit the x-axis to 10,000 in the histogram to better visualize the data distribution. The histogram shows multiple prominent spikes, particularly around lower values, depicting frequent exemption categories within the dataset.

- Filed Name: EXCD1



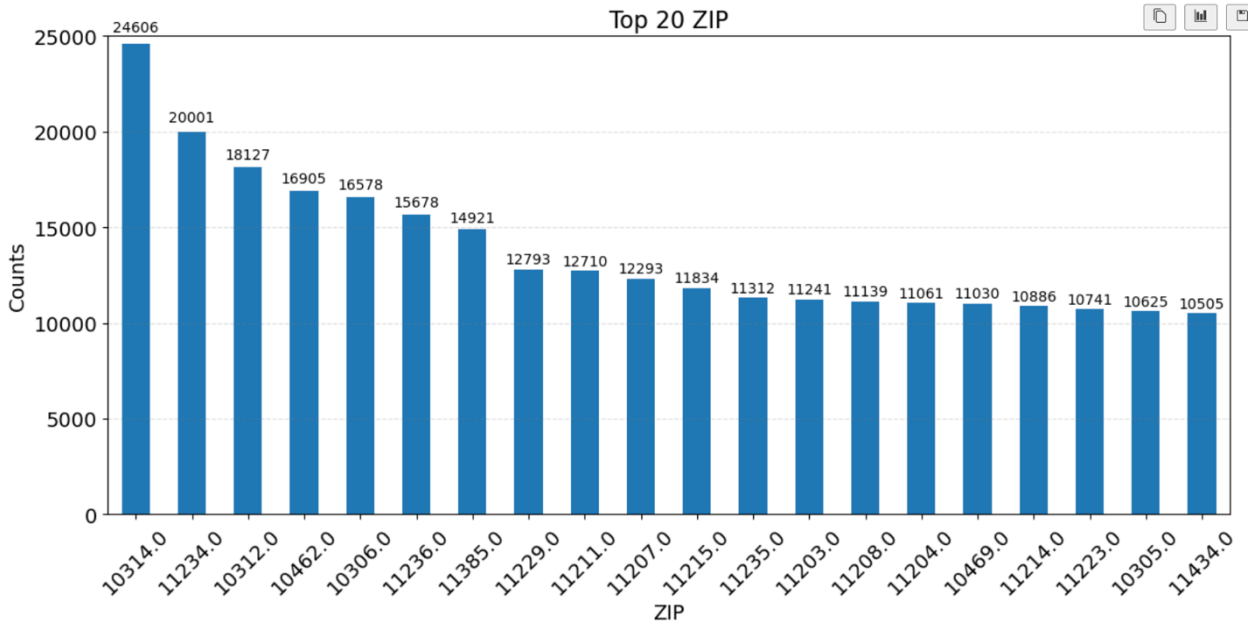
Description: EXCD1 represents "Exemption Code 1," which categorizes various types of property tax exemptions applicable to a property. The histogram displays the distribution of the top 20 most frequent EXCD1 codes. The y-axis is on a logarithmic scale to better visualize the frequency disparity between codes. The most common exemption code, 1017.0, significantly outnumbers the others, indicating it is the most prevalent type of exemption granted. The distribution reveals a sharp decline in frequency from the most common codes to less frequent ones, highlighting the concentration of specific exemption types in the dataset.

- Filed Name: STADDR



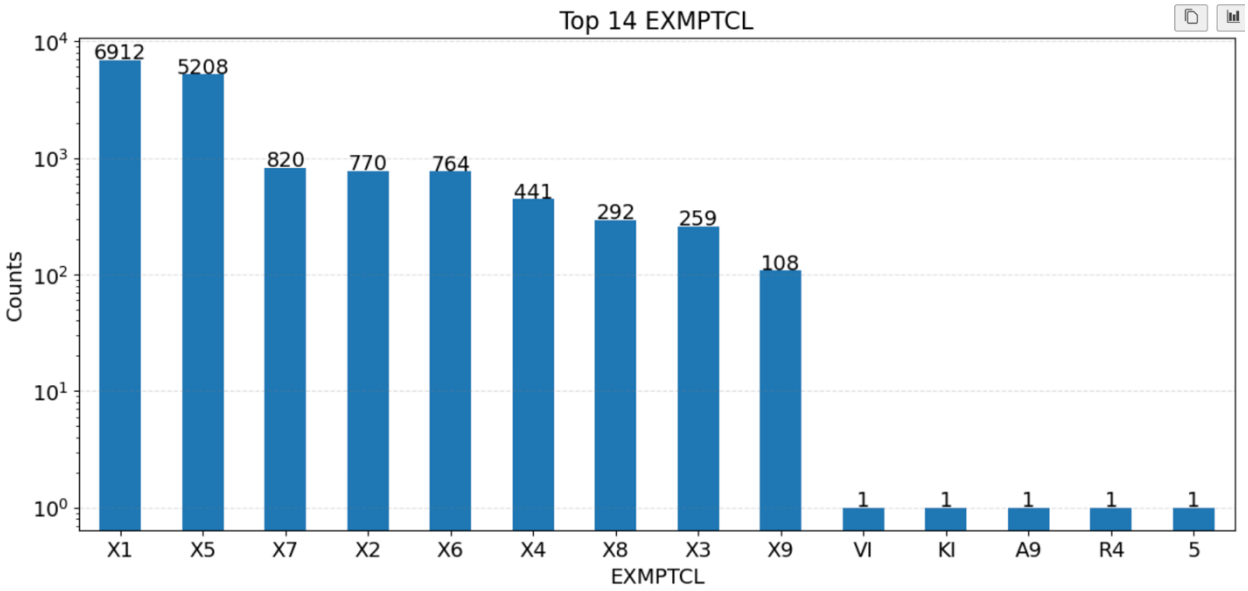
Description: STADDR refers to the "Street Address" of properties. The bar chart above illustrates the distribution of the top 20 most frequent addresses in the dataset. The address at '501 Surf Avenue' appears most frequently, with a count of 902 occurrences, indicating a high concentration of recorded activities or transactions at this location. Each subsequent address shows a gradual decrease in frequency, with '91 John Street' still having a significant count of 445.

- Filed Name: ZIP



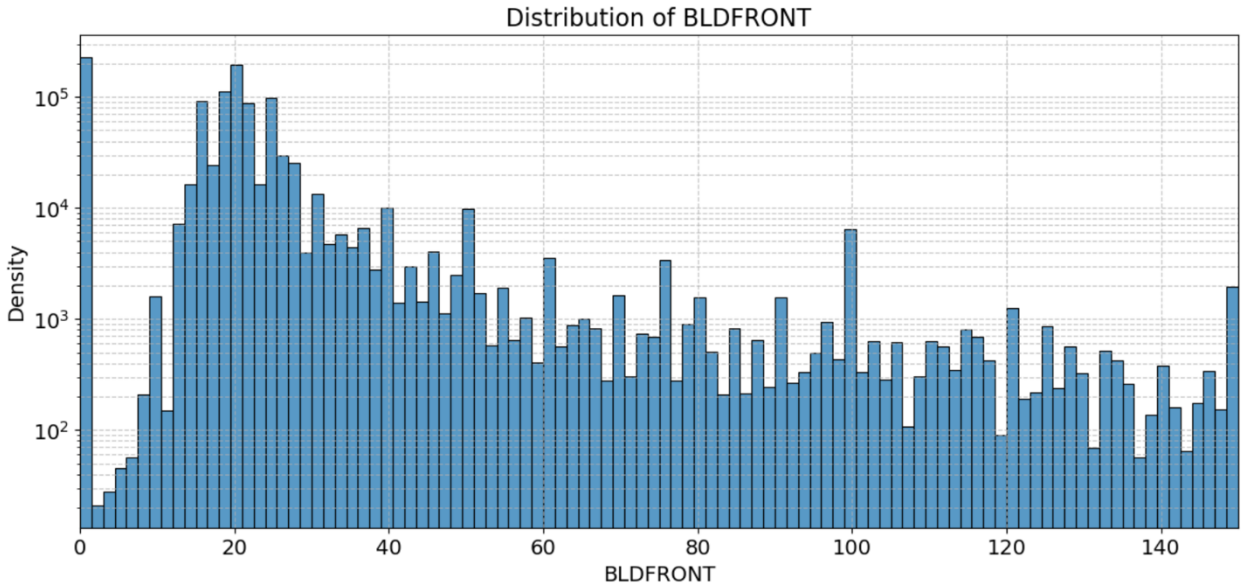
Description: The "ZIP" variable represents the zip codes associated with properties in the dataset. This bar chart highlights the distribution of the top 20 most frequent zip codes. The highest frequency is observed in the zip code 10314 with 24,606 occurrences, indicating a significant concentration of property records in this area. This is followed by zip codes 11234 and 20001, showing high activity levels as well. The visualization provides a clear picture of which areas, based on zip codes, have the highest number of property-related records, useful for regional analysis and targeted decision-making in urban planning or real estate investment.

- Filed Name: EXMPTCL



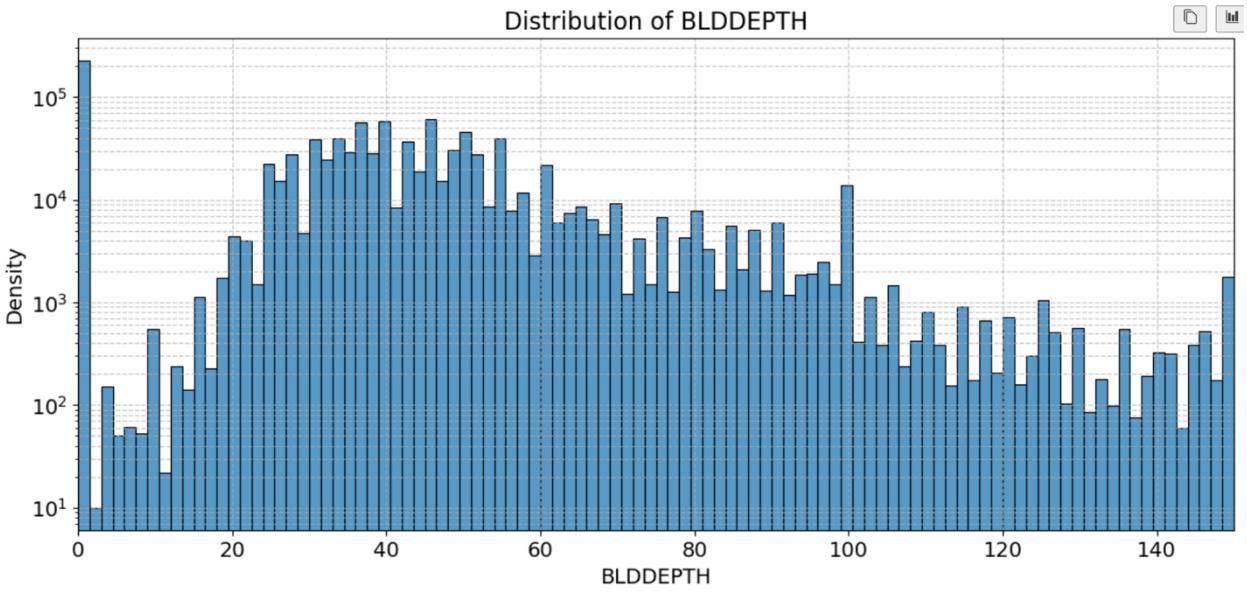
Description: The "EXMPTCL" variable represents the exemption classes associated with properties. This bar chart displays the distribution of the top 14 exemption classes among the properties in the dataset. The exemption class "X1" appears most frequently with 6,912 instances, indicating significant property exemptions under this category, followed by "X5" and "X7". The chart illustrates a steep decline in occurrence after the initial few classes, with several exemption classes such as "VI", "KI", "A9", "R4", and "5" appearing only once, suggesting these exemptions are much less common or very specific in their application.

- Filed Name: BLDFRONT



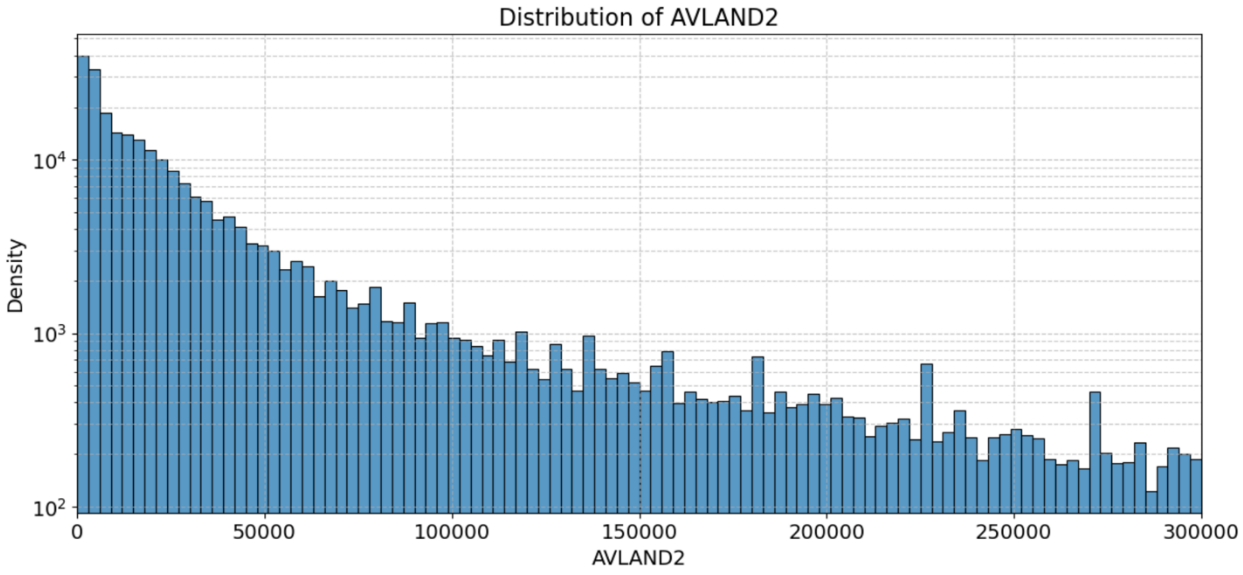
Description: The "BLDFRONT" variable represents building width in feet and shows a varied distribution. The distribution is right-skewed, with a high frequency of smaller values, peaking around 20 feet, and a gradual decrease in frequency as the building frontage increases, reflecting a common urban property layout where smaller frontages are more prevalent. An analysis of the boxplot data reveals that the majority of building widths are concentrated within 150 feet. The histogram limits the x-axis to 150 feet to focus on common building widths. The y-axis is logarithmically scaled to enhance visibility of frequency distribution across different widths.

- Filed Name: BLDDEPTH



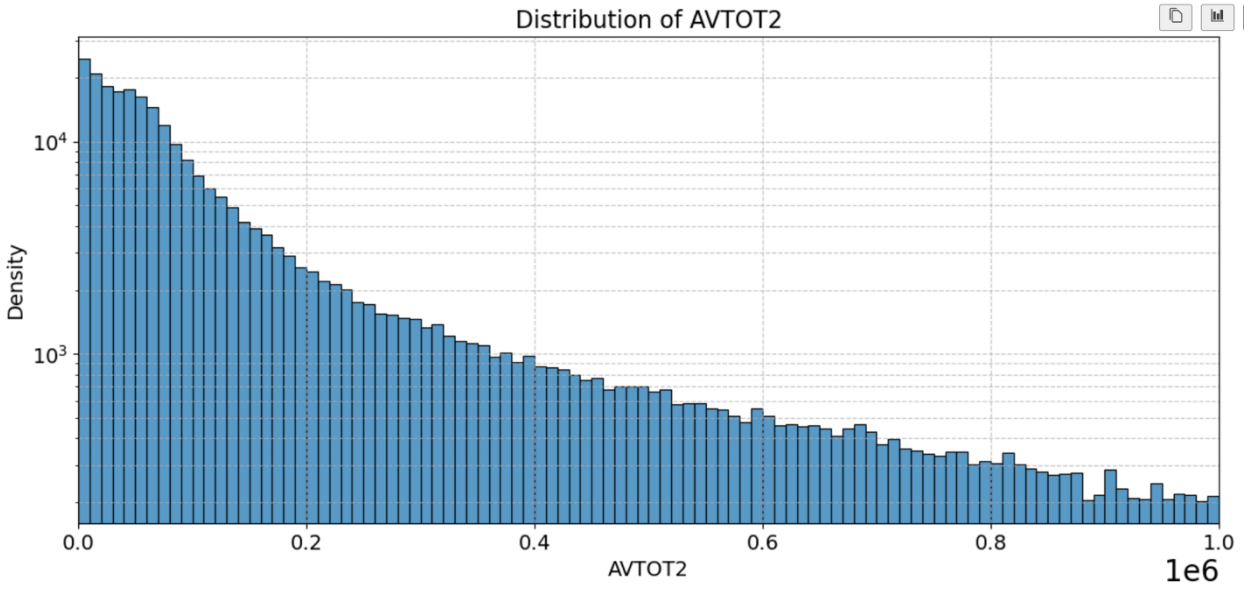
Description: The "BLDDEPTH" variable indicates the depth of buildings measured in feet. The distribution exhibits a relatively uniform distribution with multiple peaks, predominantly in the range of 20 to 100 feet. This suggests variability in building depths, likely reflecting a mix of property types and zoning regulations within an urban setting. Analysis of the boxplot data reveals a concentration of building depths primarily within 150 feet. The histogram visualizes this by limiting the x-axis to 150 feet and applying a logarithmic scale to the y-axis.

- Filed Name: AVLAND2



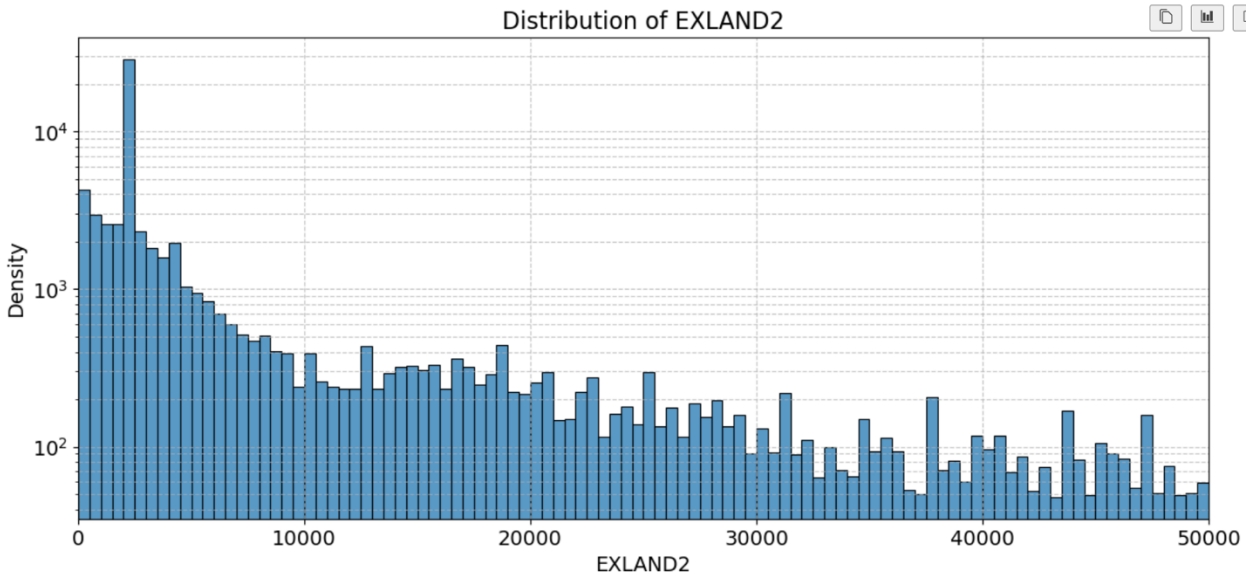
Description: The "AVLAND2" variable represents the transitional land value. Analysis of the boxplot indicates that the majority of values are concentrated within $300,000. The histogram presented here limits the x-axis to $300,000 and applies a logarithmic scale to the y-axis to enhance the visualization of data distribution across a wide range of values.

- Filed Name: AVTOT2



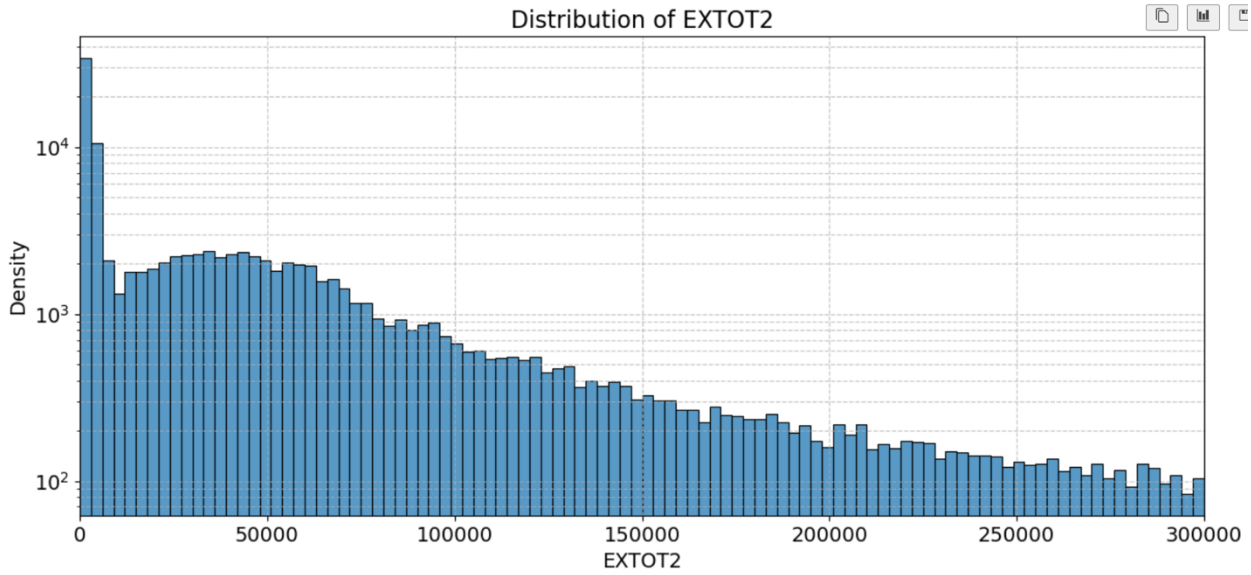
Description: The "AVTOT2" variable represents the transitional total value of properties. The distribution shows most properties have lower transitional land values, with a sharp decrease in frequency as values rise, indicating fewer high-valued lands. Examination of the boxplot data revealed a significant concentration of values within $1,000,000. The presented histogram limits the x-axis to $1,000,000 and employs a logarithmic transformation on the y-axis to effectively display the distribution of values, ensuring clarity in visualizing both the density of common value ranges and the tails extending towards higher values. This adjustment provides a detailed view into the distribution, highlighting the density peaks and variances across the spectrum of transitional total values.

- Filed Name: EXLAND2



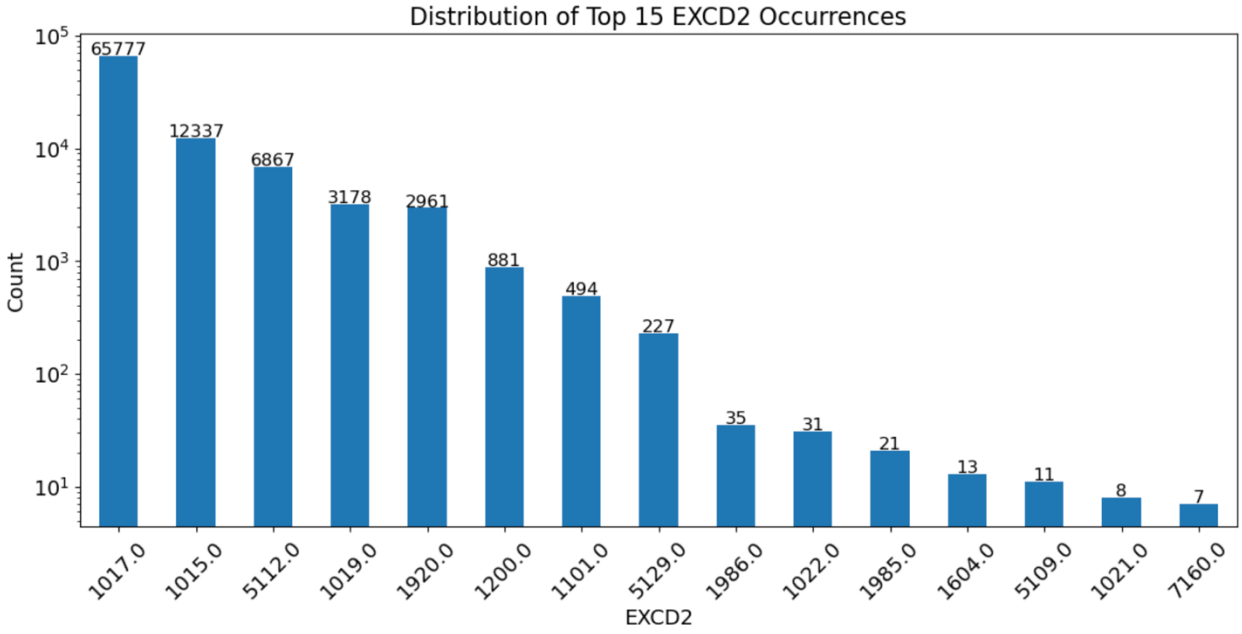
Description The "EXLAND2" variable reflects the transitional exemption land value of properties. The distribution illustrates that most properties possess relatively modest exempt land values, predominantly clustered below $10,000, with occurrences gradually tapering off at higher values. An examination of the boxplot indicated that a significant proportion of the data concentrates within $50,000. The histogram restricts the x-axis to $50,000 and uses a logarithmic scale on the y-axis to more effectively display the distribution.

- Filed Name: EXTOT2



Description: The "EXTOT2" variable represents the transitional exemption land total of properties. The distribution highlights a significant concentration of properties with relatively low total exempt values, primarily peaking below $50,000, with a gradual decline observed towards higher values up to $300,000. Analysis of the boxplot indicated that the majority of the data is concentrated within $300,000. The histogram visualizes this distribution with an x-axis limited to $300,000 and employs a logarithmic scale on the y-axis.

- Filed Name: EXCD2



Description: The variable "EXCD2" corresponds to the second exemption code assigned to properties. This histogram shows the distribution of the top 15 occurrences of EXCD2, displaying a rapidly declining frequency of these codes. The highest occurrence is for code 1017.0, indicating it's the most common exemption, followed by a significant drop to the next most frequent codes, demonstrating a skewed distribution towards a few specific exemptions.