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

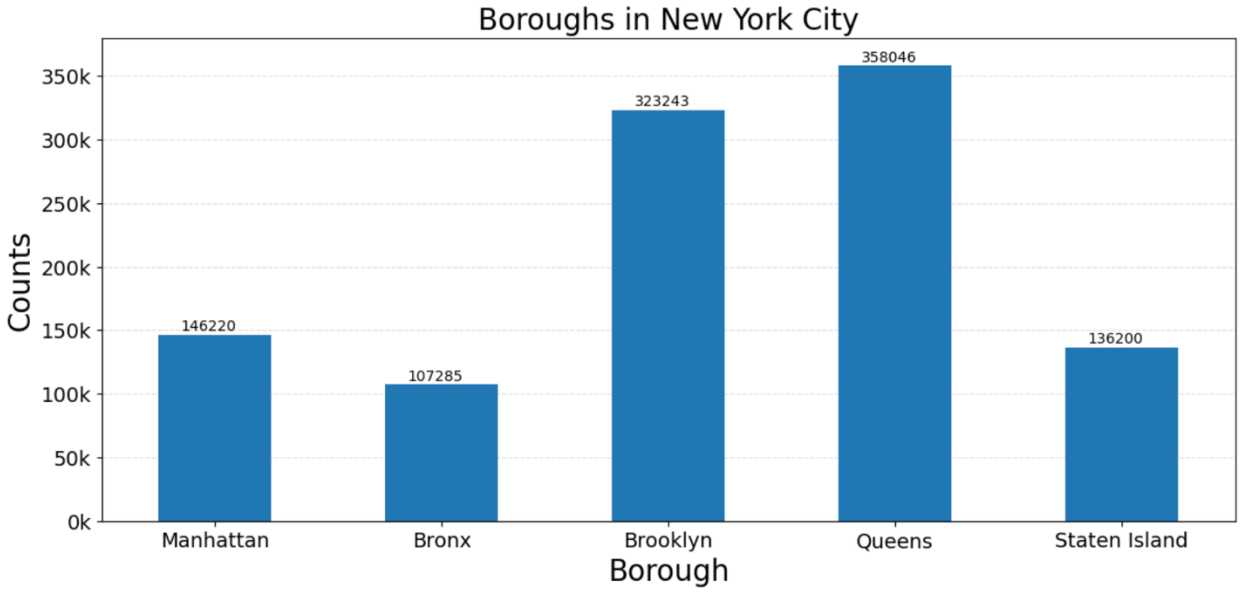
   It also features several categorical fields including 'RECORD', 'BBLE', 'BORO', 'BLOCK', 'LOT', 'EASEMENT', 'OWNER', 'BLDGCL', 'TAXCLASS', 'EXT', 'EXCD1', 'STADDR', 'ZIP', 'EXMPTCL', 'EXCD2', 'PERIOD', 'YEAR', and 'VALTYPE'.

	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
    a. 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.

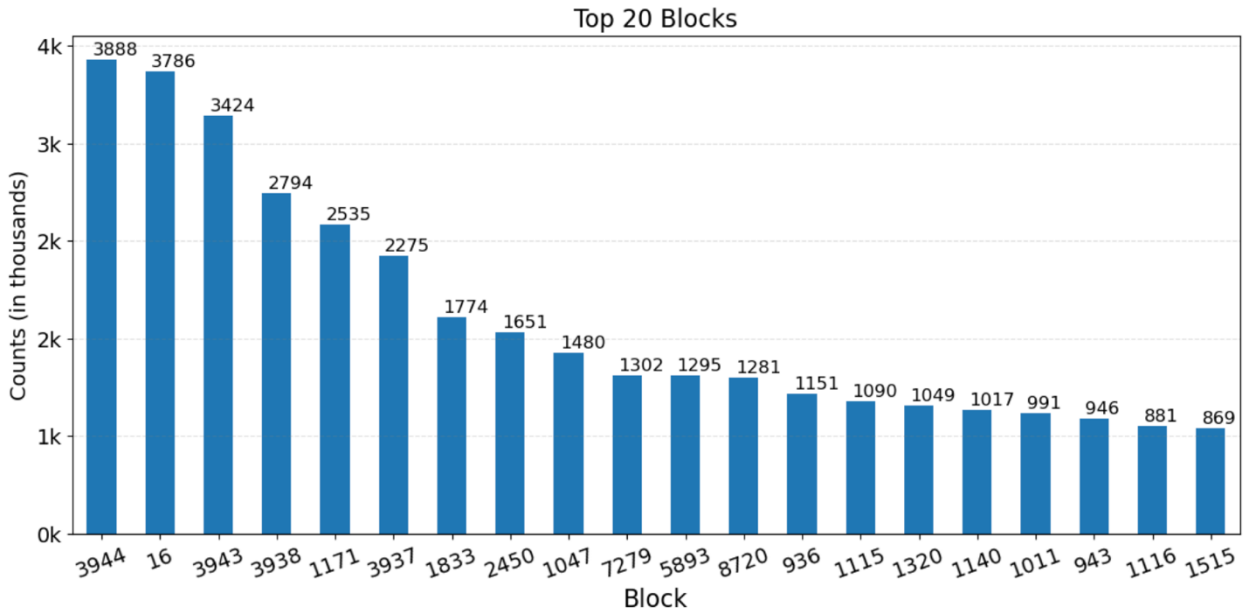
    b. Filed Name: BBLE
    Description: The BBLE field in the dataset serves as a unique identifier for each property, combining the borough, block, lot, and easement code, which is critical for tracking and managing property-related information across New York City's extensive real estate database.

    c. Filed Name: BORO



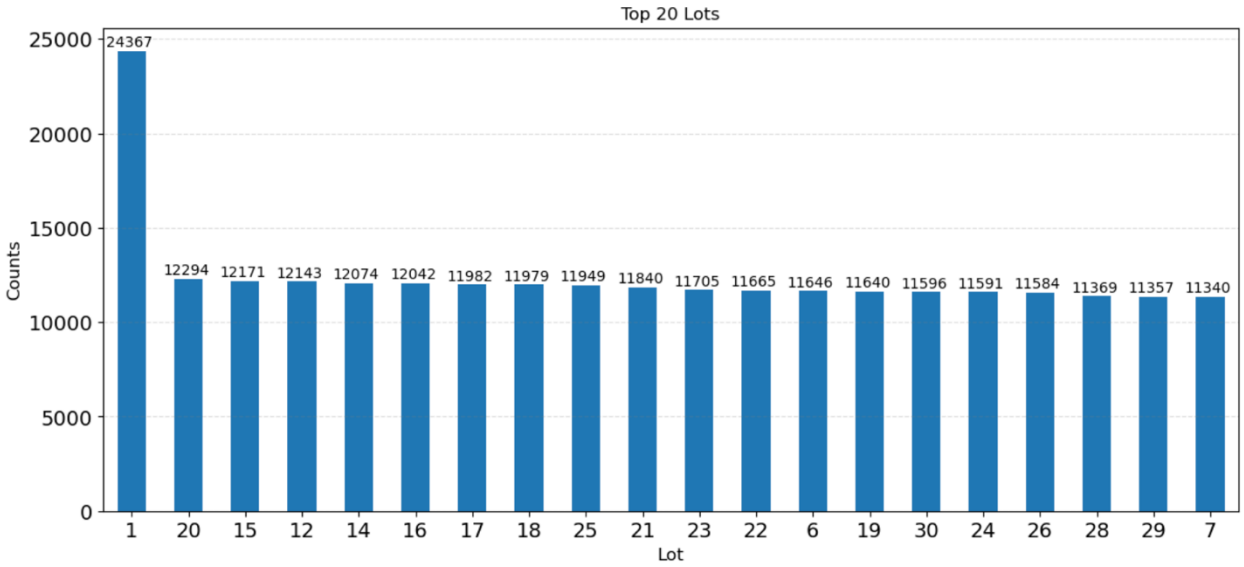
Description: BORO represent the five boroughs of New York City. Queens has the highest number of properties with a count of 358,046, followed by Brooklyn with 323,243 properties. Manhattan, despite its prominence, has fewer properties recorded at 146,220. The Bronx and Staten Island have 107,285 and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

d. Filed Name: BLOCK



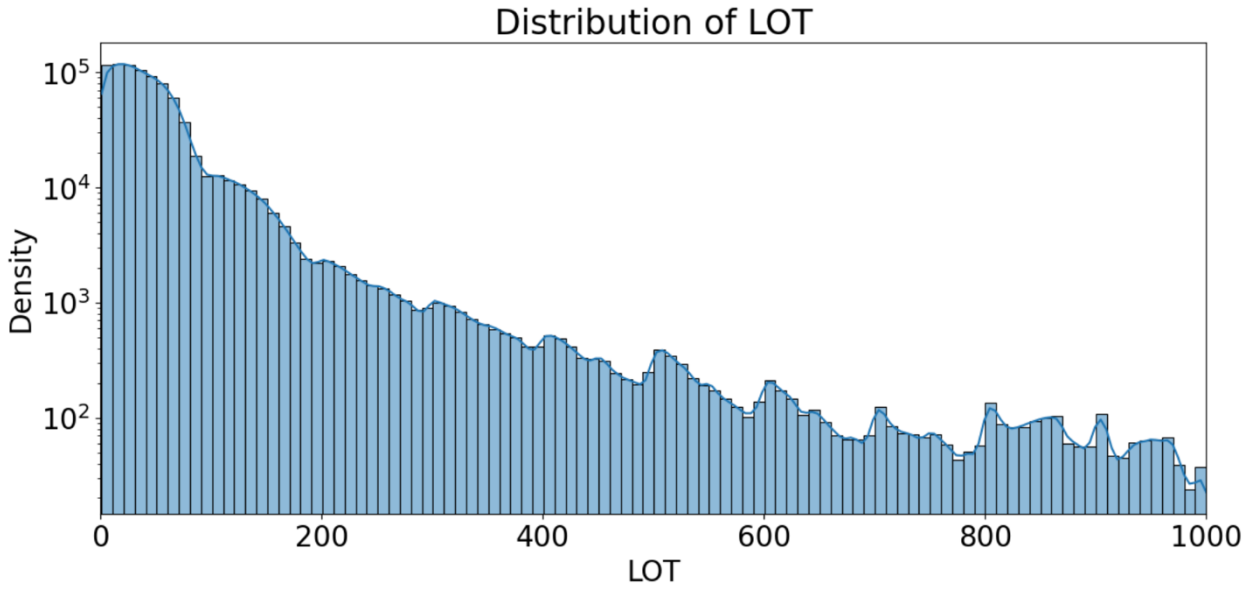
Description: The BLOCK variable in the dataset represents specific city blocks within New York City, serving as a numeric identifier that groups properties into defined segments. Each borough has its unique range of valid block numbers, for instance, Manhattan has block numbers ranging from 1 to 2255, 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.

e. 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.

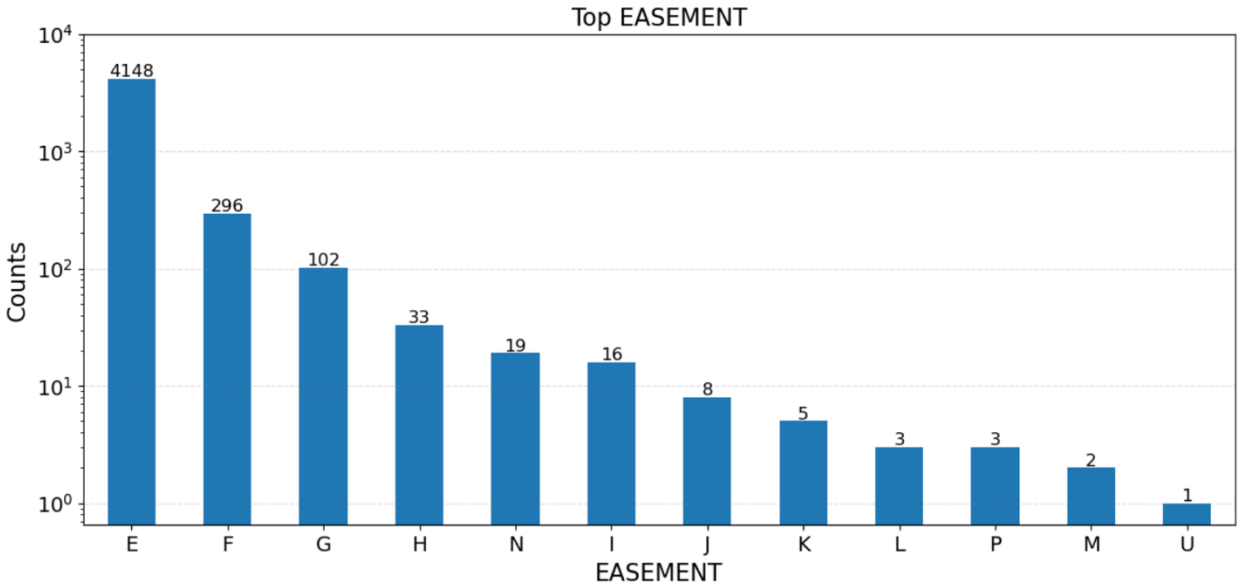
The top 20 lots in the dataset show a wide range of property counts, with the most populated lot having 24,367 properties and the least among the top 20 holding 8,869, indicating significant variation in lot usage and development density across the city.



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

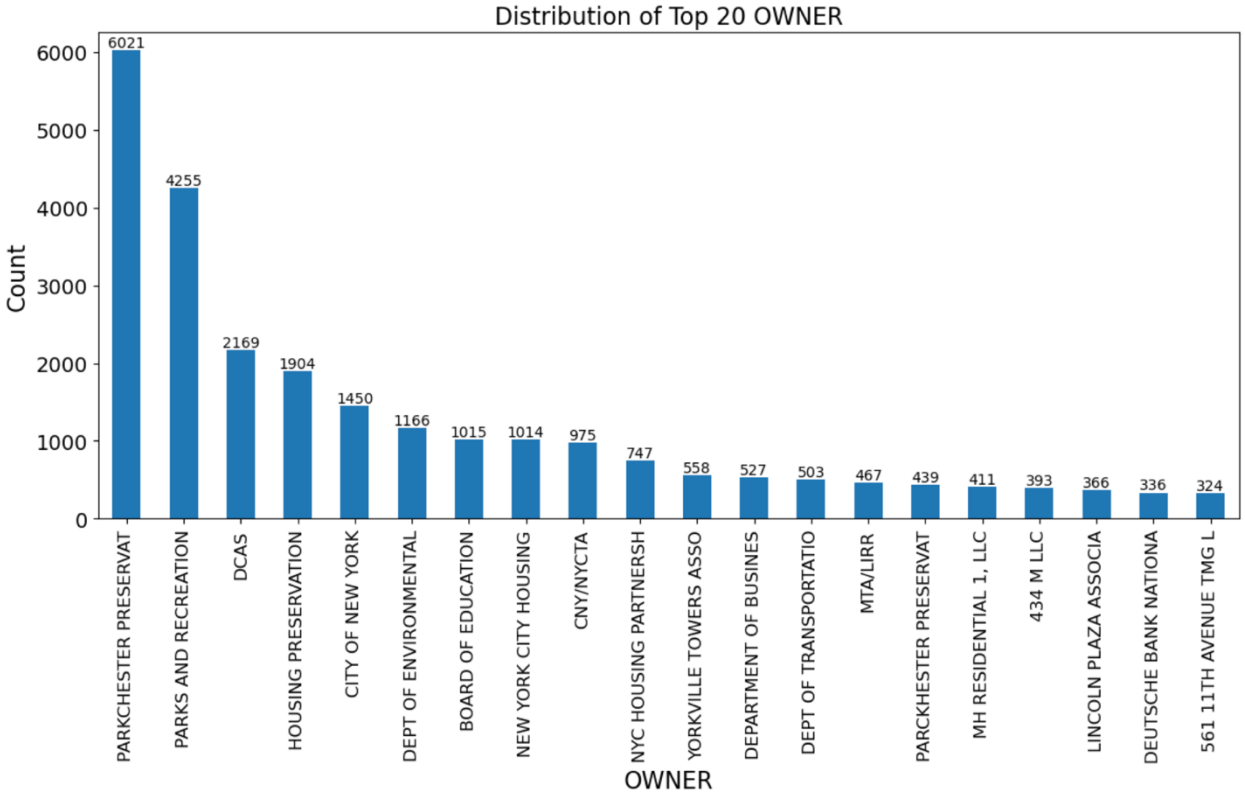
commonly registered, there are notable concentrations of property records around higher lot numbers, possibly reflecting larger developments or subdivisions within certain blocks.

f. Filed Name: EASEMENT



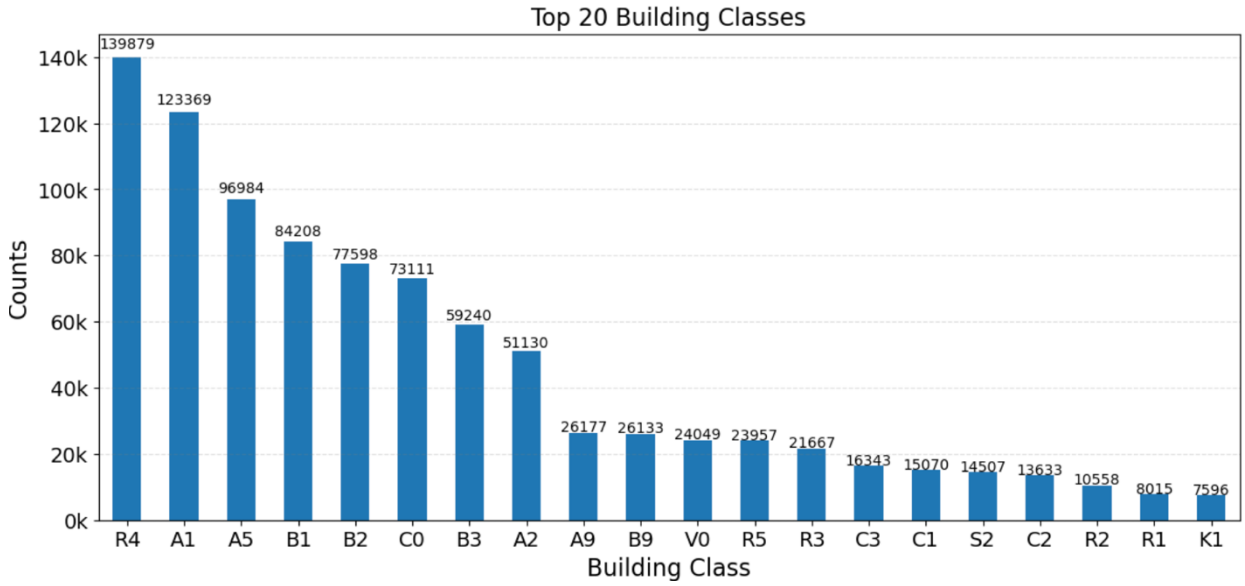
Description: The EASEMENT variable in the dataset categorizes properties based on specific rights or restrictions associated with the property's use, such as air rights, land access, or governmental use. This categorical variable includes types like 'A' for Air Easement, 'E' for Land Easement, and 'U' 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.

g. Filed Name: OWNER



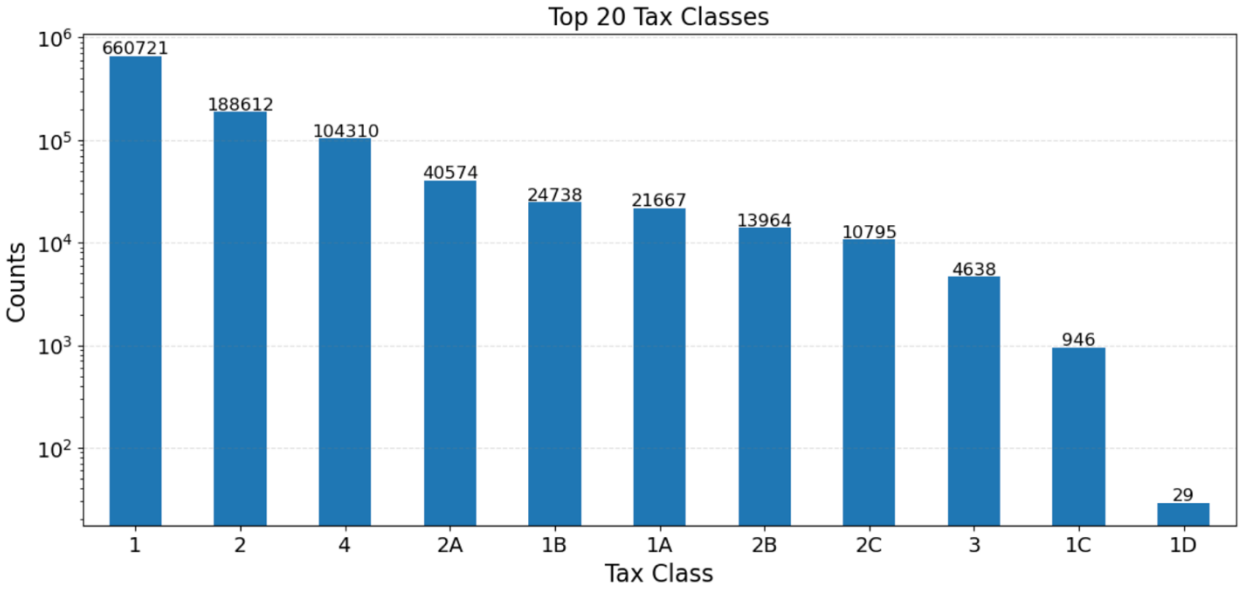
Description: The OWNER variable in the dataset specifies the name of the entity or individual that holds title to the property. This categorical variable is significant as it reflects ownership diversity across the dataset, ranging from public organizations like "PARKCHESTER PRESERVAT" and "NYC HOUSING 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.

h. Filed Name: BLDGCL



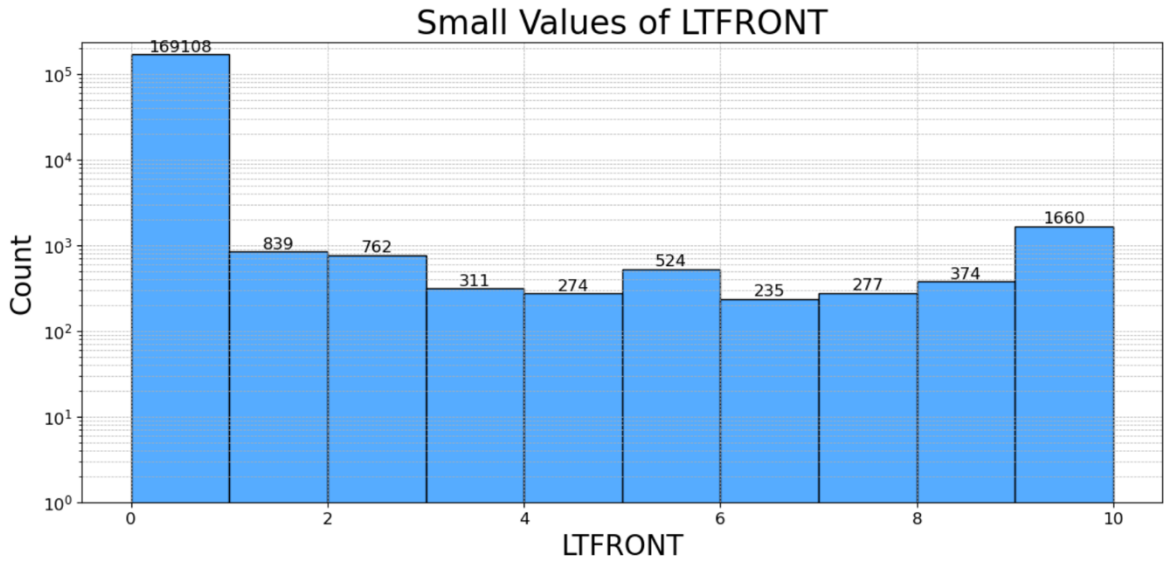
Description: The BLDGCL (Building Class) variable in the dataset categorizes properties based on their type and use, such as residential, commercial, or mixed-use buildings. Each class is represented by a code, such as R4 for residential condominiums, A1 for one-family dwellings, and C0 for walk-up apartments, 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.

i. Filed Name: TAXCLASS



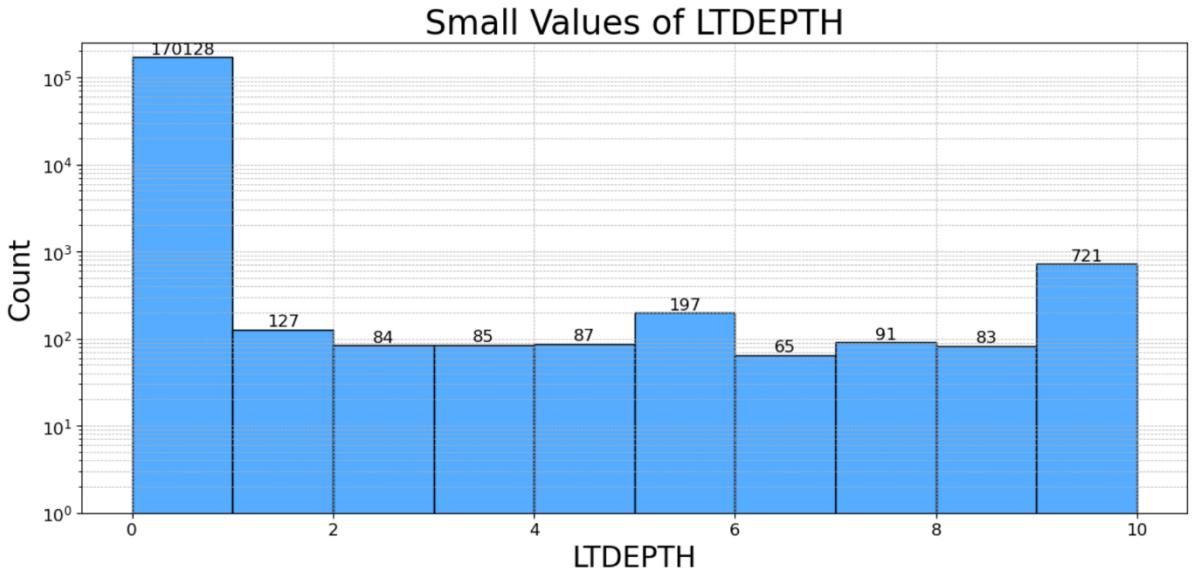
Description: The TAXCLASS variable categorizes properties based on their primary use and potential tax liability, significantly impacting how properties are assessed for tax purposes. The distribution of property types across various tax classes shows a wide range, with Class 1 (1-3 unit residences) 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.

j. Filed Name: LTFRONT



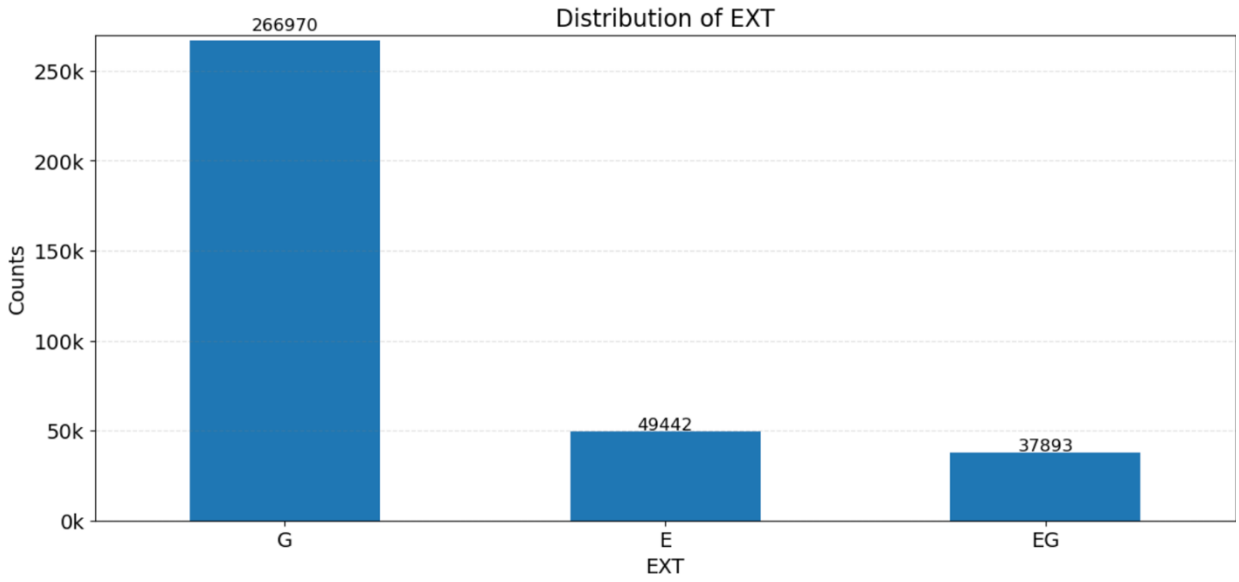
Description: The LTFRONT variable represents the width of the lot facing the street, measured in feet, and is crucial for understanding property layout and assessing property value. Analysis of both boxplot and distribution plot reveals that most LTFRONT values are concentrated within 10 feet, indicating 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.

k. Filed Name: LTDEPTH



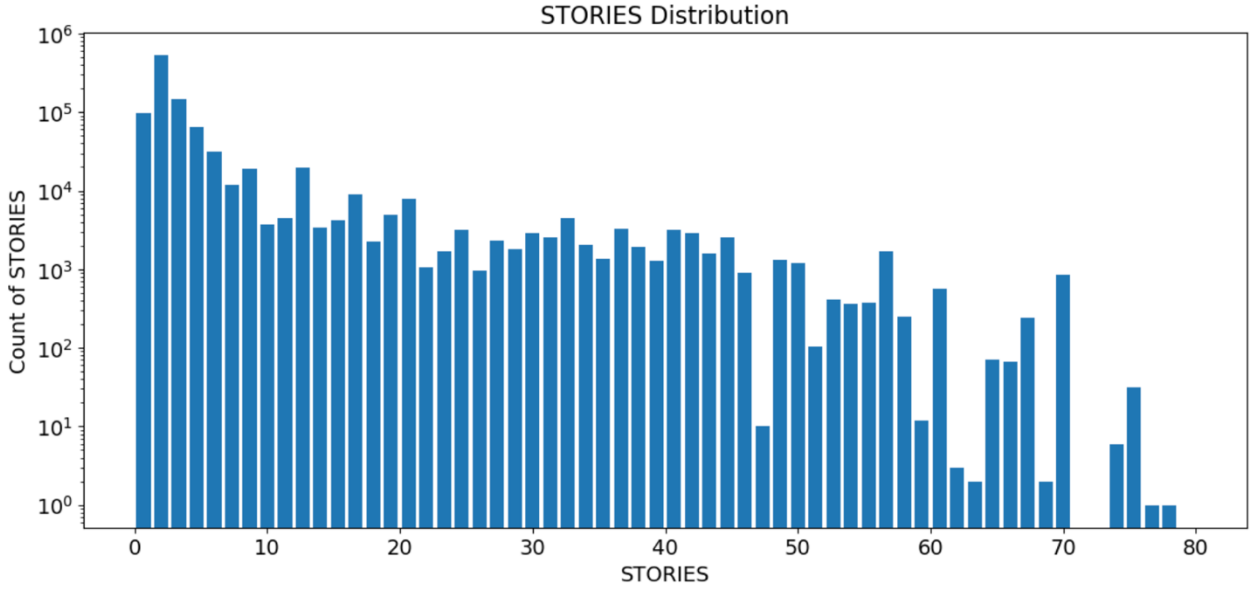
Description: The LTDEPTH variable measures the depth of a property lot in feet, from the street front to the back of the lot. Analysis of both boxplot and distribution plot shows a strong concentration of values within 10 feet, highlighting a common characteristic in densely built urban environments 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.

l. Filed Name: EXT



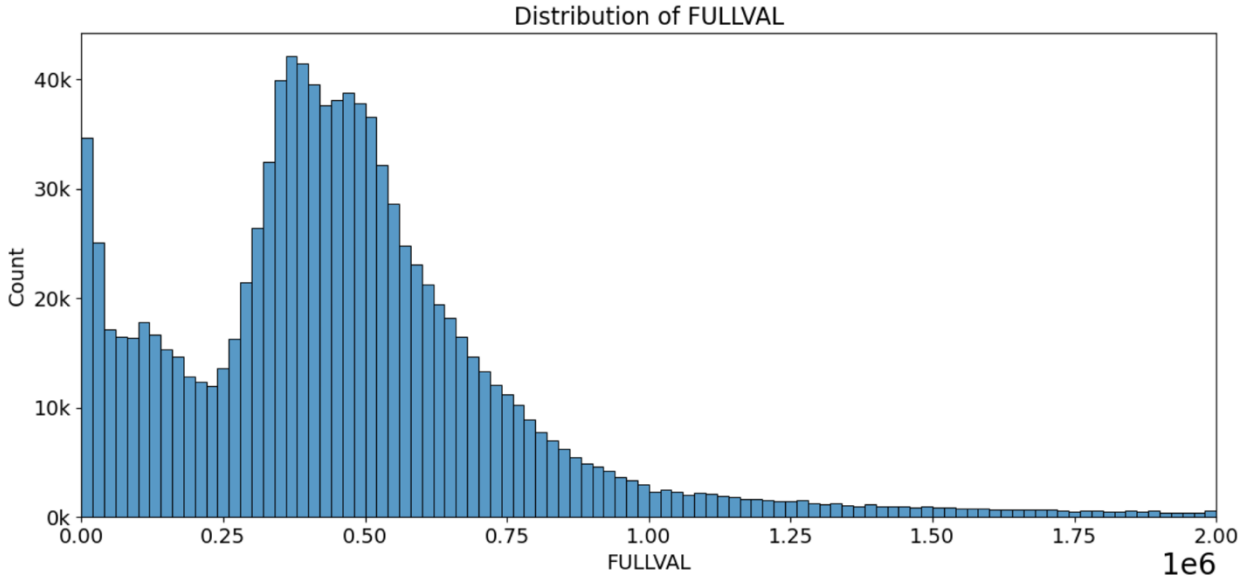
Description: The EXT variable, known as the Extension Indicator, categorizes properties based on whether they have an extension, using the codes: 'G' for no extension, 'E' for some extension, and 'EG' for extensive extension. The distribution shows a significant majority of properties without extensions (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.

m. Filed Name: STORIES



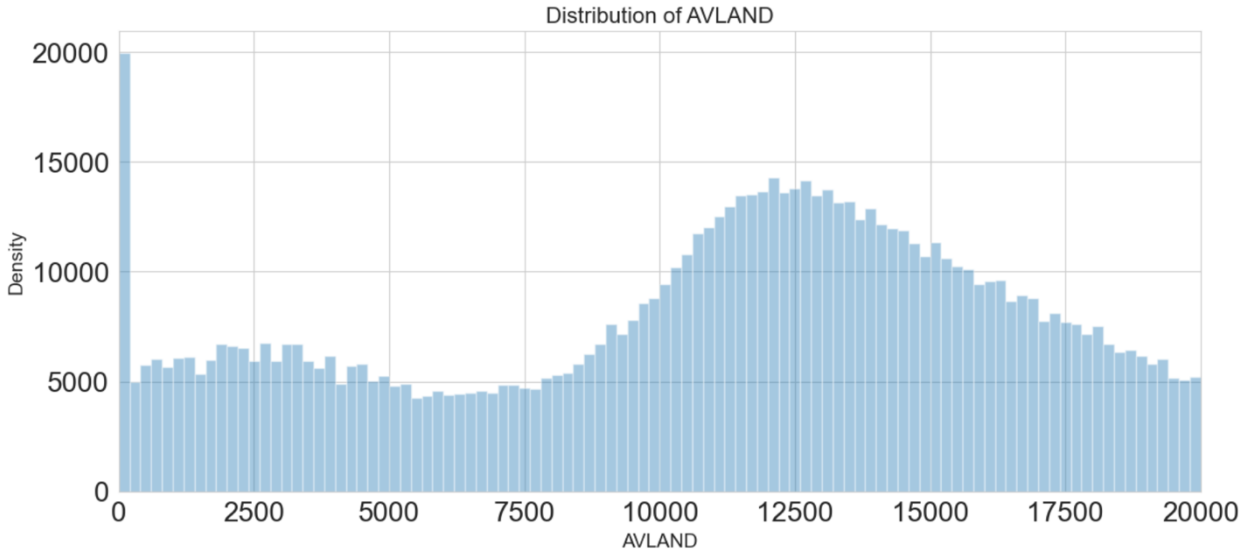
Description: The STORIES variable represents the number of stories in a building, as recorded in the dataset. The distribution demonstrates a broad range of building heights, with a notable decrease in frequency as the number of stories increases, suggesting that taller buildings are less common in the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

n. Filed Name: FULLVAL



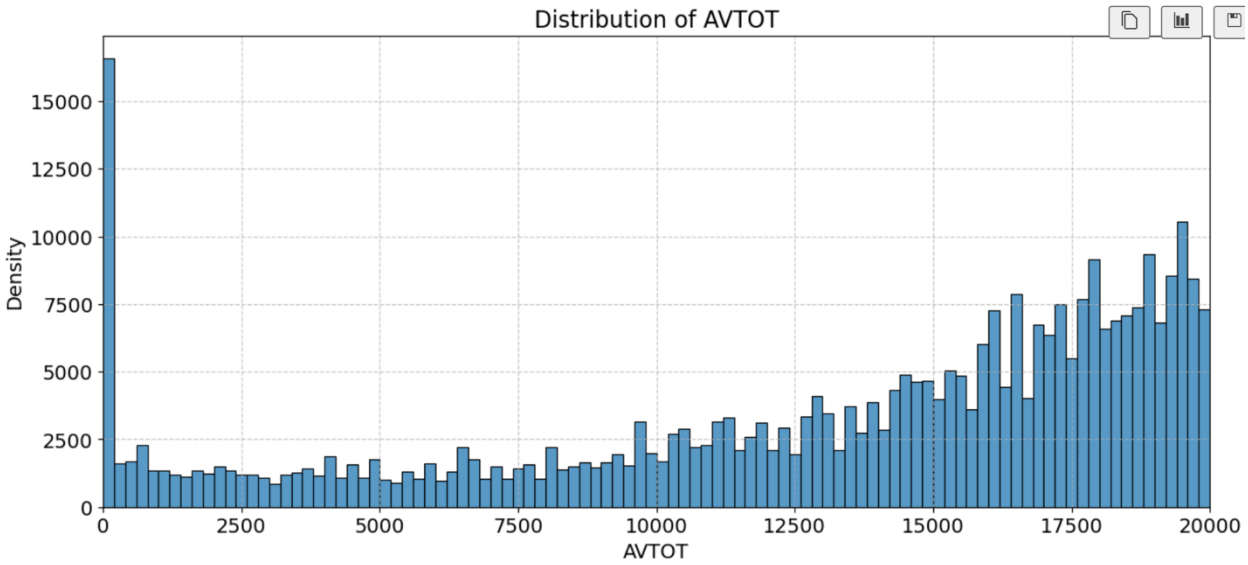
Description: The FULLVAL variable represents the market value of properties as assessed. The distribution shows a right-skewed pattern where most properties are valued under $1 million, which is typical for the dataset, indicating a higher concentration of lower-valued properties. Initial examination 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.

o. Filed Name: AVLAND



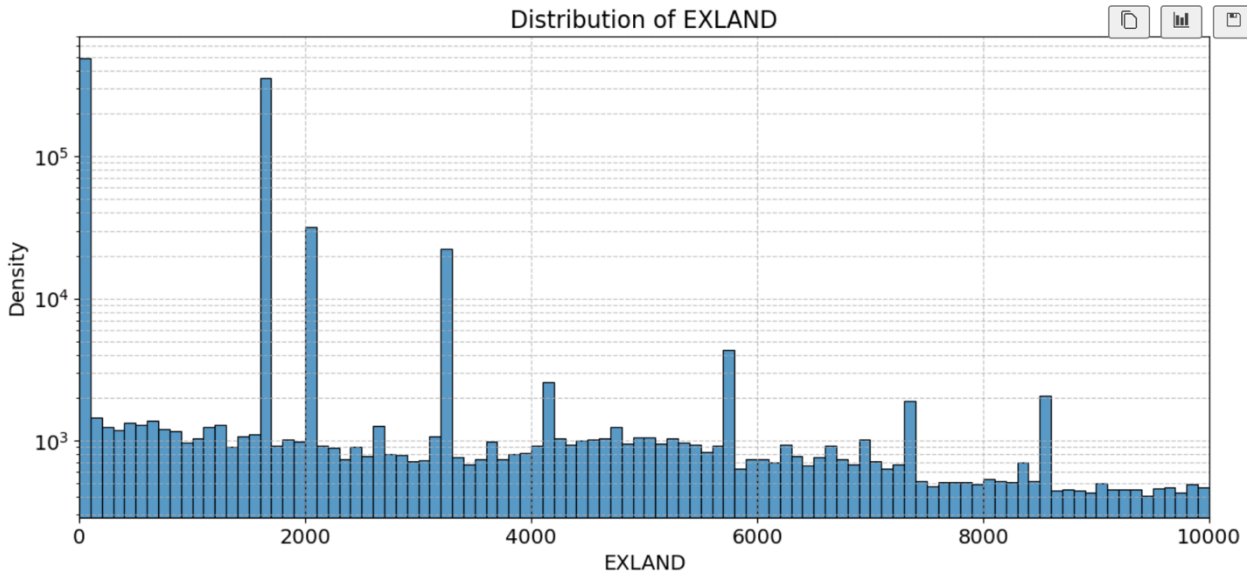
Description: The variable "AVLAND" represents the Actual Land Value of properties in the dataset. It measures the assessed value assigned to the land component of a property, crucial for determining property taxes. The distribution of "AVLAND" shows a peak around values under $20,000, indicating a concentration 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.

p. Filed Name: AVTOT



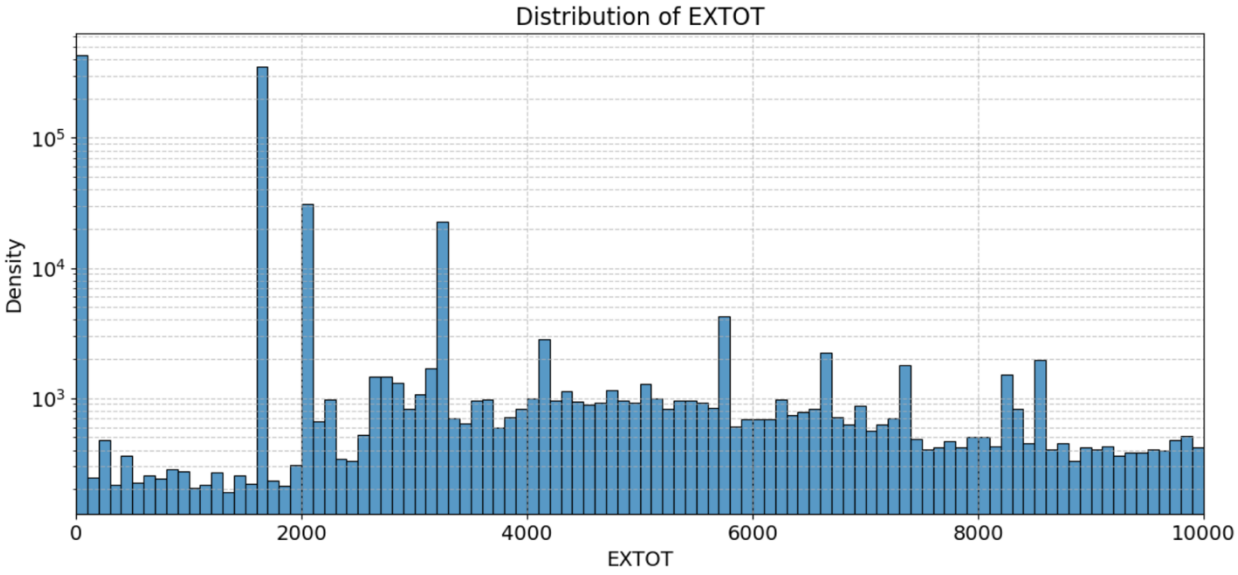
Description: The variable AVTOT, representing the Actual Total Value, shows a broad range of values across the dataset. Initial analysis with a boxplot revealed that the majority of data points are concentrated below $100,000, prompting a narrower focus in subsequent visualizations. Upon closer examination 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.

q. Filed Name: EXLAND



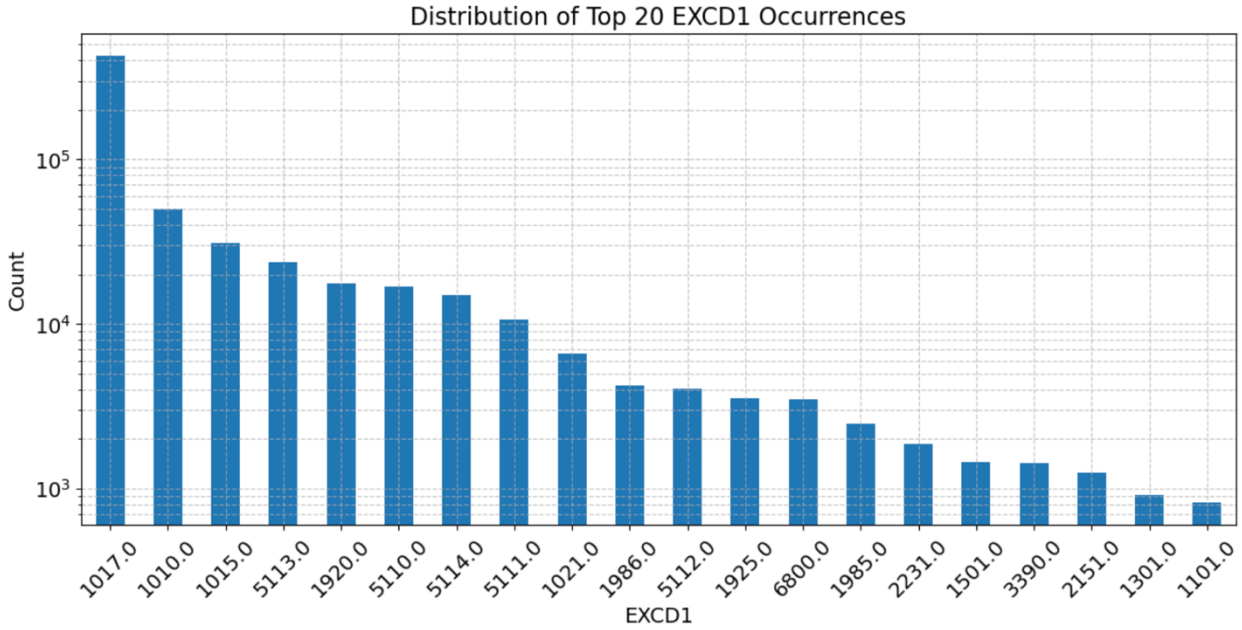
Description: EXLAND, representing "Actual Exempt Land Value," measures the value of land exempt from property taxes under various programs. Analyzing the distribution of EXLAND values reveals a dense concentration of data points within the first 10,000 units. This analysis was supported by a boxplot 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.

r. Filed Name: EXTOT



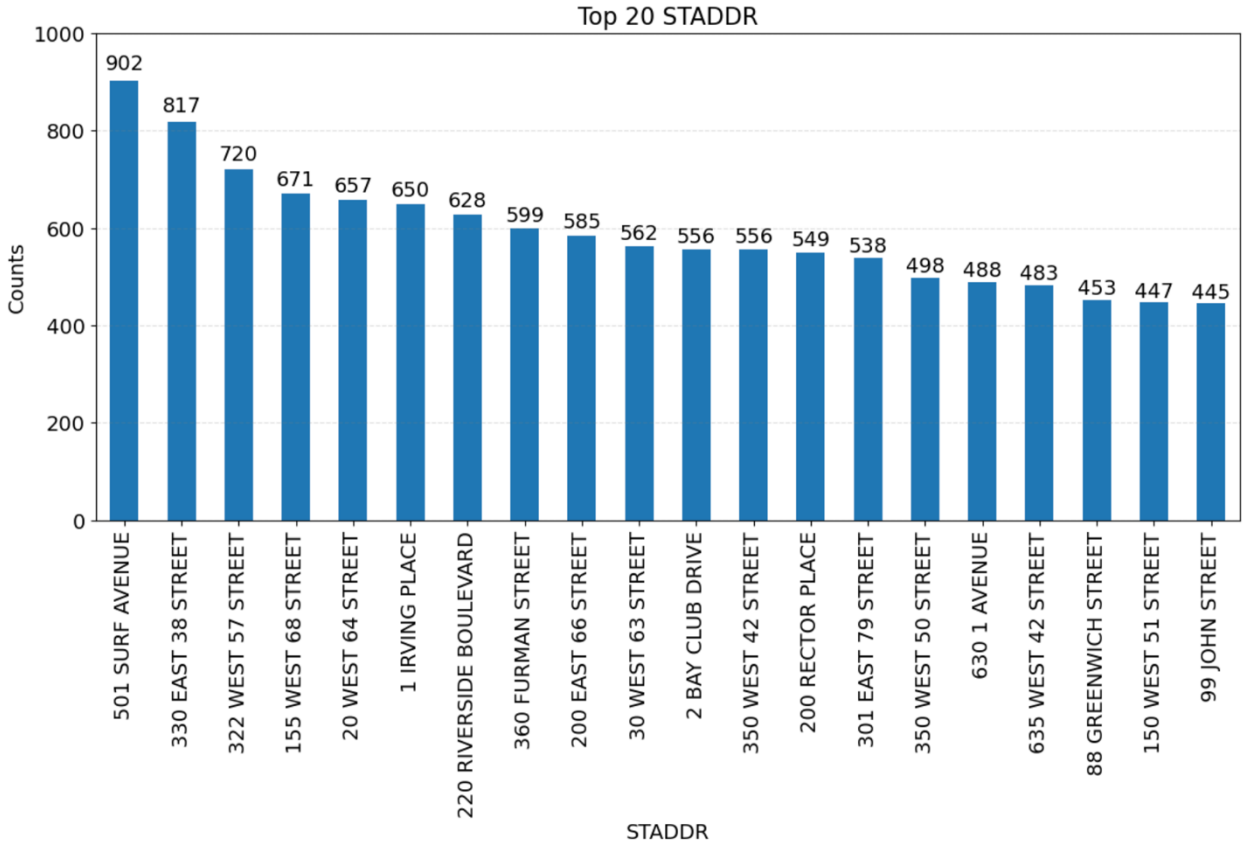
Description: EXTOT represents "Actual Exempt Land Total," indicating the total value of land exempt from taxes. A detailed analysis of EXTOT values highlighted a significant concentration of entries below 10,000, as observed from the boxplot. This informed the decision to limit the x-axis to 10,000 in 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.

s. Filed Name: EXCD1



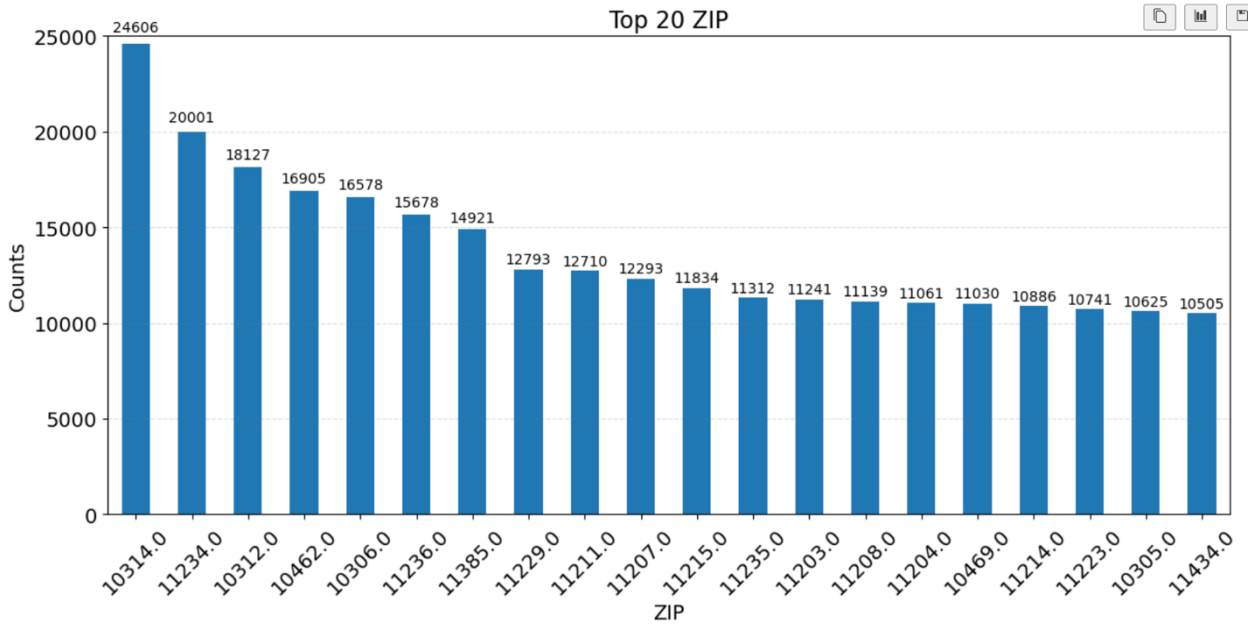
Description: EXCD1 represents "Exemption Code 1," which categorizes various types of property tax exemptions applicable to a property. The histogram displays the distribution of the top 20 most frequent EXCD1 codes. The y-axis is on a logarithmic scale to better visualize the frequency disparity between 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.

t. Filed Name: STADDR



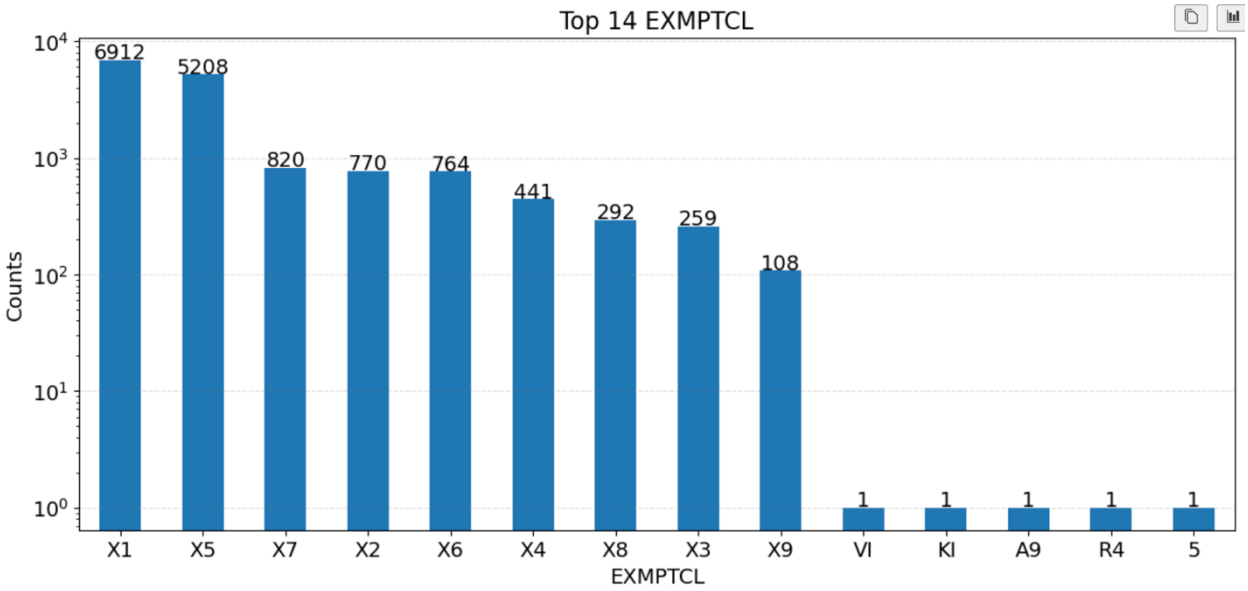
Description: STADDR refers to the "Street Address" of properties. The bar chart above illustrates the distribution of the top 20 most frequent addresses in the dataset. The address at '501 Surf Avenue' appears most frequently, with a count of 902 occurrences, indicating a high concentration of recorded 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.

u. Filed Name: ZIP



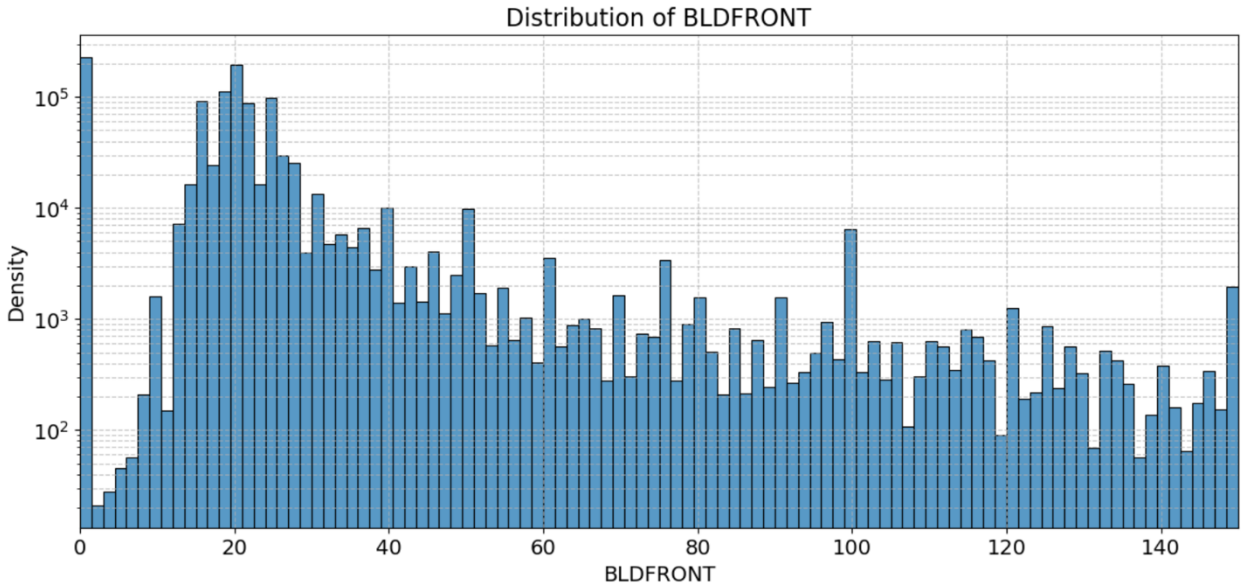
Description: The "ZIP" variable represents the zip codes associated with properties in the dataset. This bar chart highlights the distribution of the top 20 most frequent zip codes. The highest frequency is observed in the zip code 10314 with 24,606 occurrences, indicating a significant concentration 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.

v. Filed Name: EXMPTCL



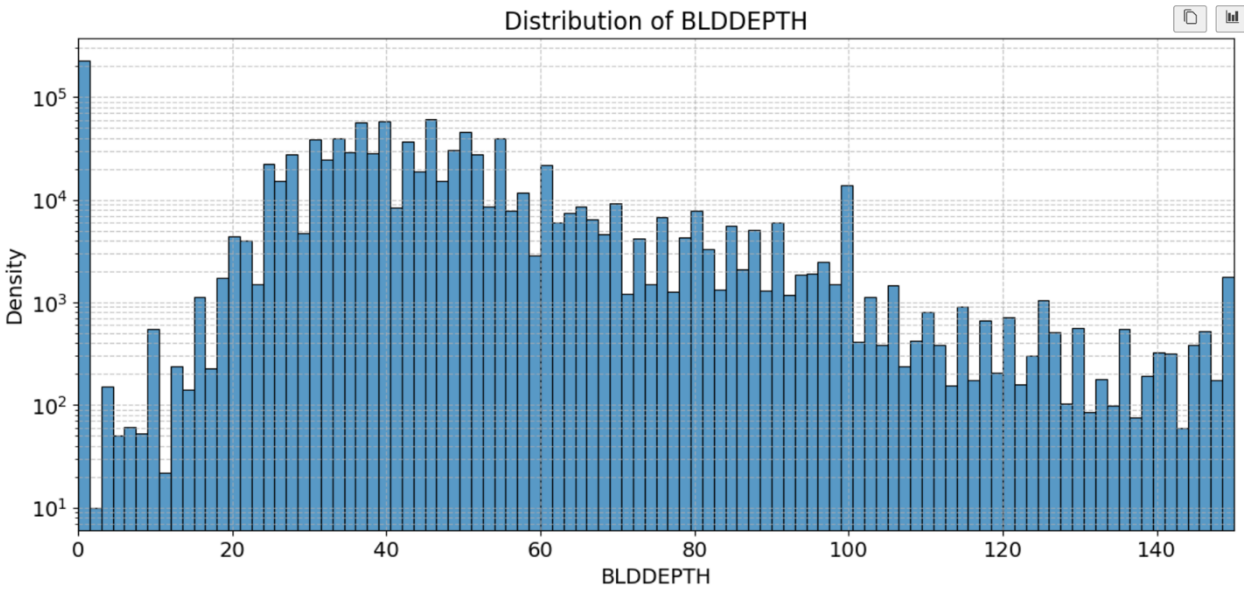
Description: The "EXMPTCL" variable represents the exemption classes associated with properties. This bar chart displays the distribution of the top 14 exemption classes among the properties in the dataset. The exemption class "X1" appears most frequently with 6,912 instances, indicating significant 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.

w. Filed Name: BLDFRONT



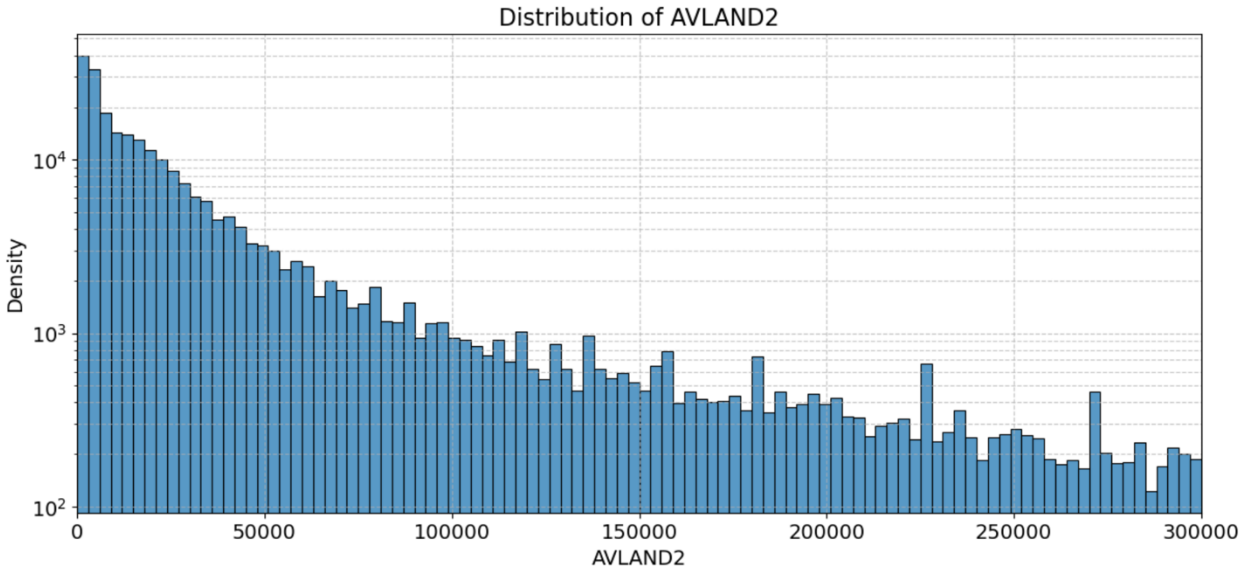
Description: The "BLDFRONT" variable represents building width in feet and shows a varied distribution. The distribution is right-skewed, with a high frequency of smaller values, peaking around 20 feet, and a gradual decrease in frequency as the building frontage increases, reflecting a common urban 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.

x. Filed Name: BLDDEPTH



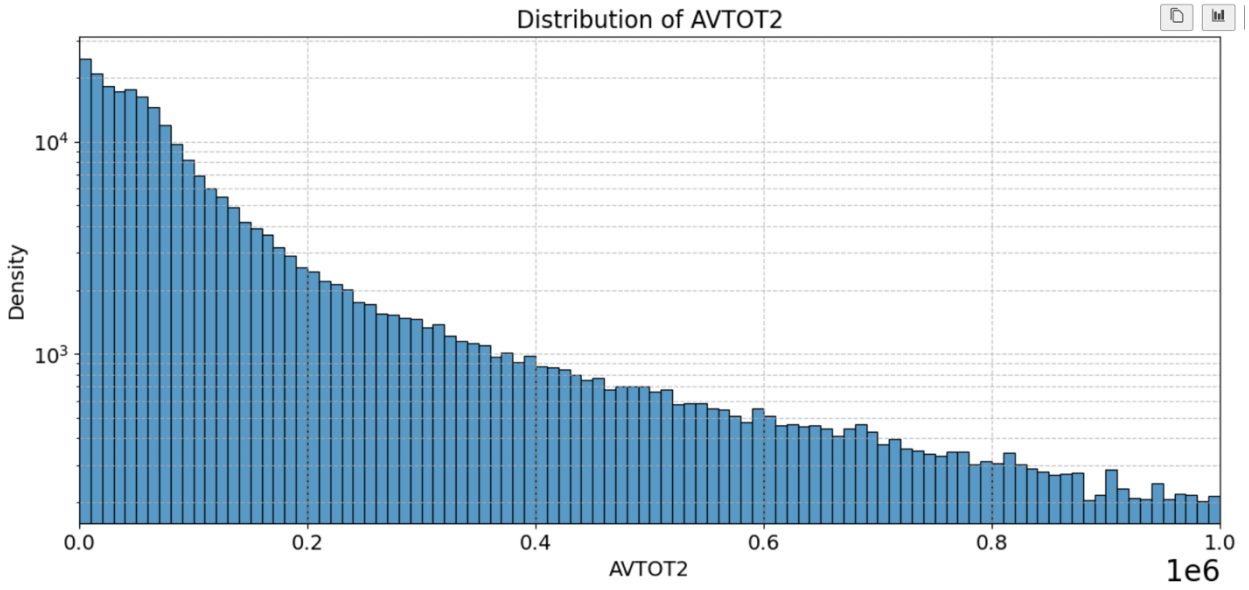
Description: The "BLDDEPTH" variable indicates the depth of buildings measured in feet. The distribution exhibits a relatively uniform distribution with multiple peaks, predominantly in the range of 20 to 100 feet. This suggests variability in building depths, likely reflecting a mix of property types 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.

y. Filed Name: AVLAND2



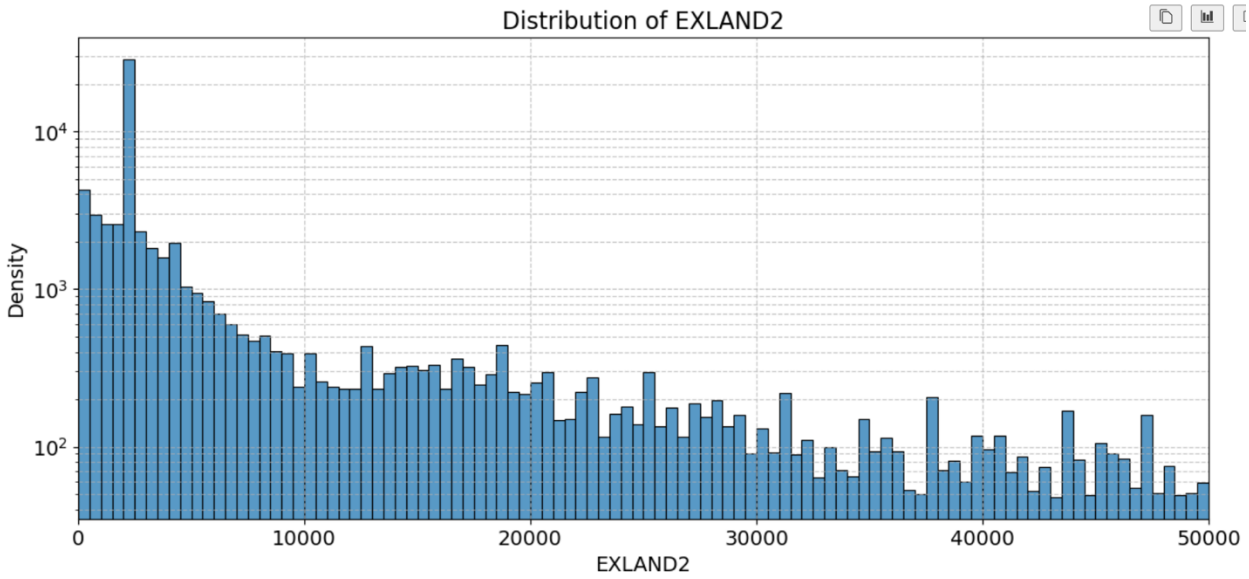
Description: The "AVLAND2" variable represents the transitional land value. Analysis of the boxplot indicates that the majority of values are concentrated within $300,000. The histogram presented here limits the x-axis to $300,000 and applies a logarithmic scale to the y-axis to enhance the visualization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

z. Filed Name: AVTOT2



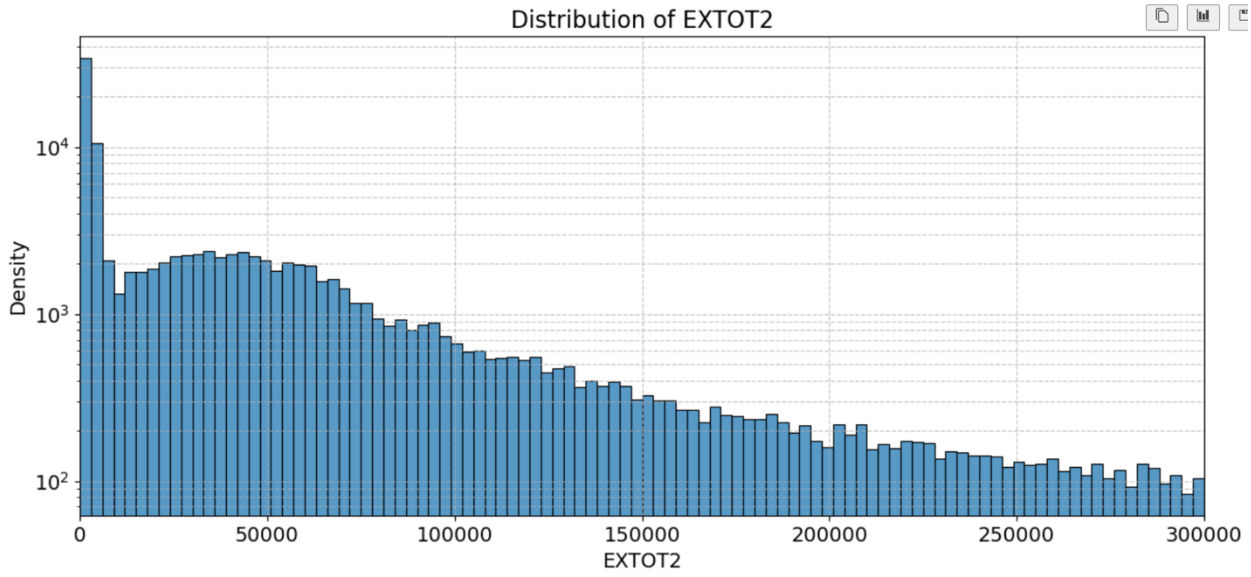
Description: The "AVTOT2" variable represents the transitional total value of properties. The distribution shows most properties have lower transitional land values, with a sharp decrease in frequency as values rise, indicating fewer high-valued lands. Examination of the boxplot data revealed a significant 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.

aa. 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.

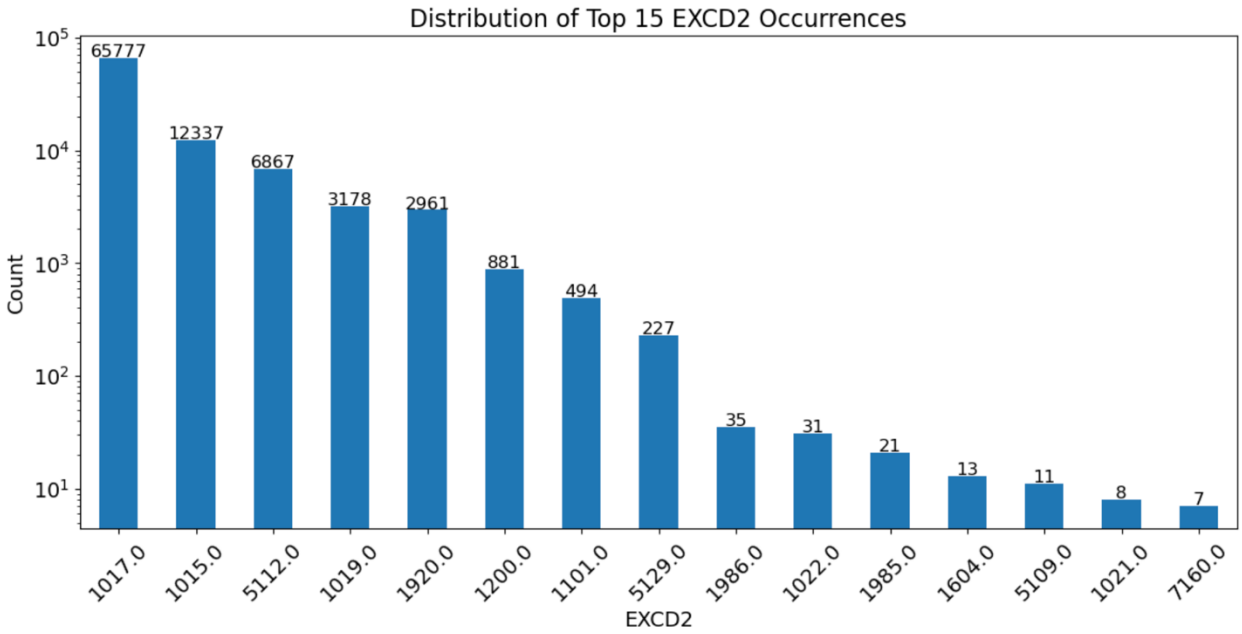
bb. Filed Name: EXTOT2



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.

cc. 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.