Homework 9 - Project 2 Informal Report

Executive summary

The project involved analyzing a comprehensive dataset of real estate properties in New York City, released by the Department of Finance. This dataset encompassed essential property details like borough, block, and lot information from 2010/11, and included 1,070,994 records across 32 fields. The analysis covered numerical fields like LTFRONT, LTDEPTH, STORIES, FULLVAL, among others, along with categorical fields like RECORD, BBLE, BORO, BLOCK, LOT, and many more.

The outcomes of the analysis uncovered vital insights into data distribution and prevalence. For example, the analysis identified that some fields were fully populated while others had a varied population percentage. In addition, it noted the occurrence of zeros, min-max values, mean, standard deviation, and most common instances in numerical fields. In categorical fields, the percentage of populated records, the number of zeros, the count of unique values, and the most common instances were summarized. The results from the data quality analysis have laid a strong foundation for further data exploration, modeling, and inference in the real estate sector, thereby guiding businesses in making strategic decisions.

Data Description

1. Overview of data

The dataset appears to be a comprehensive record of real estate properties located in New York City, published by the **Department of Finance**. The data comprises various essential details about each property, including a unique identification number, location information such as borough, block, and lot for the **year 2010/11**. This data can be beneficial for businesses involved in the real estate industry to gain insights into property trends and make informed decisions related to investments, acquisitions, and sales. The dataset contains **1,070,994 records** across **32 fields**.

Akansha Lalwani (A59019733)

2. Statistics Tables

Numeric Fields Table

Field Name	# Records Have Values	% Populated	% Zeros	Min	Max	Mean	Standard Deviation	Most Common
LTFRONT	1,070,994	100.00%	15.79 %	0	9,999	36.63	74.03	0
LTDEPTH	1,070,994	100.00%	15.89 %	0	9,999	88.86	76.4	100
STORIES	1,014,730	94.75%	0.00%	1	119	5.00	8.37	2
FULLVAL	1,070,994	100.00%	1.21%	0	6,150,000,000	874,264.50	11,582,425.58	0
AVLAND	1,070,994	100.00%	1.21%	0	2,668,500,000	85,067.91	4,057,258.16	0
AVTOT	1,070,994	100.00%	1.21%	0	4,668,308,947	227,238.16	6,877,526.09	0
EXLAND	1,070,994	100.00%	45.91 %	0	2,668,500,000	36,423.89	3,981,573.93	0
EXTOT	1,070,994	100.00%	40.39 %	0	4,668,308,947	91,186.98	6,508,399.78	0
BLDFRONT	1,070,994	100.00%	21.36 %	0	7,575	23.04	35.58	0
BLDDEPTH	1,070,994	100.00%	21.37 %	0	9,393	39.92	42.71	0
AVLAND2	282,726	26.40%	0.00%	3	2,371,005,000	246,235.71	6,178,951.64	2,408
AVTOT2	282,732	26.40%	0.00%	3	4,501,180,002	713,911.43	11,652,508.34	750
EXLAND2	87,449	8.17%	0.00%	1	2,371,005,000	351,235.68	10,802,150.91	2,090
EXTOT2	130,828	12.22%	0.00%	7	4,501,180,002	656,768.28	16,072,448.75	2,090

Categorical Fields Table

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
RECORD	1,070,994	100.00%	0	1,070,994	1
BBLE	1,070,994	100.00%	0	1,070,994	1000010101
BORO	1,070,994	100.00%	0	5	4
BLOCK	1,070,994	100.00%	0	13,984	3944
LOT	1,070,994	100.00%	0	6,366	1
EASEMENT	4,636	0.43%	0	12	E
OWNER	1,039,249	97.04%	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100.00%	0	200	R4
TAXCLASS	1,070,994	100.00%	0	11	1
EXT	354,305	33.08%	0	3	G
EXCD1	638,488	59.62%	0	129	1017
STADDR	1,070,318	99.94%	0	839,280	501 SURF AVENUE
ZIP	1,041,104	97.21%	0	196	10314
EXMPTCL	15,579	1.45%	0	14	X1
EXCD2	92,948	8.68%	0	60	1,017
PERIOD	1,070,994	100.00%	0	1	FINAL
YEAR	1,070,994	100.00%	0	1	2010/11
VALTYPE	1,070,994	100.00%	0	1	AC-TR

Data Cleaning

1. The logic and results of the exclusions

The data cleaning part primarily focuses on excluding certain properties from the analysis which are not deemed useful for the fraud detection model.

Here's a breakdown of the logic and results of the exclusions from the code:

A. Removing records with easement type as government (U):

An easement refers to the right to use someone else's land for a specific purpose. For example, utilities often have easements on properties to run power lines. Records with easement type U are removed from the data, as these properties are typically government-owned or related and may not accurately reflect market property values. Only one record is removed in this step.

B. Removing records related to government and cemetery:

The code then removes any properties that have a name associated with government departments or cemeteries. The logic behind this is that such properties, owned by government agencies or designated as cemeteries, usually have a distinct valuation pattern that might skew the anomaly detection for general properties. A list of keywords associated with government entities and cemeteries is defined (like 'DEPT', 'DEPARTMENT', 'UNITED STATES', 'GOVERNMENT', 'GOVT', 'CEMETERY'), and the `OWNER` names are scanned for these keywords. All such properties are removed from the analysis.

It also ensures that the term 'STORES' is not included in an owner's name, presumably to avoid excluding commercial properties that might include terms like 'DEPT' in their name but are not government-owned.

The code first identifies a list of unique property owners and then goes through this list to find names that match the government or cemetery-related keywords. These matched names are then added to a `remove_list`.

By performing these exclusions, the resulting dataset is expected to consist primarily of privately owned properties, which are the main focus for fraud detection in this case. We removed 26501 records in total.

2. The logic and results for every field imputation

Field	Method of Imputation	Original Number of Missing Values	Filled Missing Values
ZIP	1. Mapping by 'staddr_boro' 2. Filling with before and after ZIPs 3. Filling with previous ZIP	20431	1. 2832 2. 9491 3. 8108
FULLVAL	1. Group by ['TAXCLASS', 'BORO', 'BLDGCL'] and fillna with mean 2. Group by ['TAXCLASS', 'BORO'] and fillna with mean 3. Group by ['TAXCLASS'] and fillna with mean	10025	1. 2718 2. 6647 3. 386
AVLAND	1. Group by ['TAXCLASS', 'BORO', 'BLDGCL'] and fillna with mean 2. Group by ['TAXCLASS', 'BORO'] and fillna with mean 3. Group by ['TAXCLASS'] and fillna with mean	10027	1. 2720 2. 6641 3. 386
AVTOT	1. Group by ['TAXCLASS', 'BORO', 'BLDGCL'] and fillna with mean 2. Group by ['TAXCLASS', 'BORO'] and fillna with mean 3. Group by ['TAXCLASS'] and fillna with mean	10025	1. 2718 2. 6647 3. 386
STORIES	1. Fill with modes grouped by ['BORO', 'BLDGCL'] 2. Group by ['TAXCLASS'] and fillna with mean	42030	1. 4108 2. 37922
LTFRONT	Groupby ['TAXCLASS', 'BORO'], fill with mean. If any nulls remain, group by ['TAXCLASS'], fill with mean.	160565	160563
LTDEPTH	Groupby ['TAXCLASS', 'BORO'], fill with mean. If any nulls remain, group by ['TAXCLASS'], fill with mean.	161656	161654
BLDFRONT	Groupby ['TAXCLASS', 'BORO', 'BLDGCL'], fill with mean. If any nulls remain, group by ['TAXCLASS', 'BORO'], then ['TAXCLASS'], fill with mean.	0 (only zero values treated as missing)	All zero values replaced
BLDDEPTH	Groupby ['TAXCLASS', 'BORO', 'BLDGCL'], fill with mean. If any nulls remain, group by ['TAXCLASS', 'BORO'], then ['TAXCLASS'], fill with mean.	0 (only zero values treated as missing)	All zero values replaced

Variable Creation

Description what, why, how

- First, we generated three foundational variables that embody the dimensions of the properties: lot size, building size, and building volume. These variables will serve as the foundation for creating further insightful variables.
- We then evaluated the property's worth per unit size, taking into account three valuation types: FULLVAL, AVTOT, and AVLAND. These, in combination with the three size measures, led to the formation of nine unique variables.
- We also calculated the inverse of these nine variables, giving us another set of valuable data points.
- The first nine variables aim to pinpoint properties with unusually high valuations, whereas their inverse counterparts help us spot properties with suspiciously low valuations. Both sets are integral to identifying potential fraudulent property activities.
- Two additional variables were introduced to further enhance our fraud detection capabilities. The 'value ratio' denotes the property's price per unit volume, aiding in identifying properties with anomalously low or high values. The 'size ratio', on the other hand, compares the building size to the lot size, indicating potential irregularities when a building size exceeds its lot size.
- Lastly, to prepare for Principal Component Analysis (PCA), we scaled down the values using the z-scale method. This standardization is essential for PCA, which aims for dimensionality reduction.
- We then preserved these freshly minted variables for further use in the PCA process.

Description why these are likely useful variables

- These variables are likely useful as they can reveal patterns and relationships not immediately evident in the raw data. By transforming and standardizing the data, these variables can help in identifying trends and anomalies, facilitating the development of more accurate and nuanced predictive models. They also help to incorporate domain knowledge and context, thereby increasing the interpretability of the results.
- This process of variable creation not only prepares our dataset for further analysis but also unveils a host of new dimensions for potential insights and red flags, particularly in terms of property fraud detection.
- Each of these variables offers a different perspective and helps uncover distinct aspects of property values, making them invaluable tools in our real estate analysis.

Table showing variable types and #

Variable Group	Variable Names	Algebraic Form	Description
Ratio Variables	r1	FULLVAL / ltsize	Ratios of three property value variables to three size variables
	r2	FULLVAL / bldsize	
	r3	FULLVAL / bldvol	
	r4	AVLAND / ltsize	
	r5	AVLAND / bldsize	
	r6	AVLAND / bldvol	
	r7	AVTOT / ltsize	
	r8	AVTOT / bldsize	
	r9	AVTOT / bldvol	
Inverse Ratio Variables	r1inv to r9inv	1/(ri + epsilon) where i is in [1, 9]	Inverse of the nine ratio variables (r1 - r9)
Standardized Ratio Variables (grouped by ZIP)	r1_zip5 to r9_zip5, r1inv_zip5 to r9inv_zip5	ri / mean(ri) for each ZIP where i is in [1, 9] and also in [r1inv, r9inv]	Nine ratio variables and their inverses standardized by ZIP group
Standardized Ratio Variables (grouped by TAXCLASS)	r1_taxclass to r9_taxclass, r1inv_taxclass to r9inv_taxclass	ri / mean(ri) for each TAXCLASS where i is in [1, 9] and also in [r1inv, r9inv]	Nine ratio variables and their inverses standardized by TAXCLASS group
Additional Variables	value_ratio	max(FULLVAL / (AVLAND + AVTOT) / mean(FULLVAL / (AVLAND + AVTOT)), 1 / (FULLVAL / (AVLAND + AVTOT) / mean(FULLVAL / (AVLAND + AVTOT)))	Indicator of how well the three value fields (FULLVAL, AVLAND, AVTOT) relate to each other
	size_ratio	bldsize / ltsize	Ratio of building size to lot size

Dimensionality Reduction

Dimensionality reduction, such as Principal Component Analysis (PCA), is a method used to simplify high-dimensional datasets while retaining as much relevant information as possible. The aim is to combat the curse of dimensionality, where data becomes sparse in higher dimensions, making learning difficult and leading to overfitting.

In this case, PCA has been used to reduce the original dataset to 4 principal components. This is achieved by transforming the original variables to a new set of variables, which are linear combinations of the original variables.

To compress the data while keeping the essential patterns and structures, we utilized Principal Component Analysis (PCA) on 32 variables. PCA essentially creates a new set of variables that are linear combinations of the original ones but are uncorrelated with each other. The overall form and size of the dataset are retained; only the standard deviation of each variable changes.

To ensure a fair representation of all variables in the data, we standardized the variables using a z-scaling process before applying PCA. This action neutralizes the effect of differing variable scales, allowing PCA to extract meaningful patterns without being biased towards larger-scale variables.

The PCA process starts once the variables have been z-scaled. Analyzing the scree plot and variation plot helped us decide to keep the top 5 Principal Components (PCs). These 5 PCs cover approximately 80% of the dataset's variability.

Post PCA transformation, we applied z-scaling again to the transformed variables. This step maintains consistency across the dataset, enabling us to perform further analyses and interpretations more effectively.

Here are the steps for dimensionality reduction using PCA:

1. Standardize the dataset using z-scaling to ensure all variables are on the same scale and can be compared directly. This removes the influence of differing variable scales.

2. Compute the covariance or correlation matrix to understand the relationships between variables.

3. Find the eigenvalues and eigenvectors. The eigenvalues indicate the variance explained by each principal component, and the eigenvectors point to the direction of the principal components.

4. Sort the eigenvalues and select the top-k eigenvectors to capture the maximum variance in the dataset.

5. Transform the original dataset into a lower-dimensional space using the selected eigenvectors.

This process creates a set of uncorrelated variables (the principal components) that retain most of the original dataset's variance. The final number of components kept depends on the desired level of

reduction and the variance explained by each component. After the PCA transformation, further z-scaling is applied to the new components to maintain consistency.

Z-scaling is an essential data preprocessing step before applying statistical techniques like PCA. It standardizes all variables to have a mean of zero and a standard deviation of one, ensuring all variables are directly comparable. This standardization removes the influence of differing variable scales, thus preventing variables with larger ranges from overpowering the analysis, ensuring a fair assessment of each variable's importance.

Anomaly Detection Algorithms

Anomaly detection is the process of identifying rare events or observations that significantly deviate from the majority of the data. Such anomalies could indicate problems like credit card fraud or manufacturing defects. Two unsupervised fraud scores are calculated in this case:

1. A score based on Minkowski distance, a generalized metric form of Euclidean and Manhattan distances, which measures the distance between two points in a multi-dimensional space.

2. A score based on the reconstruction error of an autoencoder, a type of artificial neural network that learns efficient codings of input data. In this context, it learns a compressed representation of the PCA-transformed data, and the difference between the original and the reconstructed data is used as an anomaly score.

High anomaly scores suggest potential anomalies. Keep in mind that the importance assigned to the Principal Components, the powers used for Minkowski distances, and the number of iterations for training the autoencoder can all impact the accuracy of the anomaly detection, and should be adjusted based on the specific use case.

1. Z-SCORE OUTLIER:

The anomaly detection begins with a group of expert variables. To prepare this data for analysis, we first standardize it using Z-scaling. This ensures all variables are comparable. We then apply PCA to remove correlations and reduce the data's dimensionality while preserving vital information. We only retain a few principal components for further analysis.

After the PCA, we do another round of Z-scaling to maintain equality among the principal components. Now each variable's value on every record represents its anomaly level in that dimension. These are transformed into Z-scores, which show how unusual a record is. To accumulate these Z-scores, we use the Minkowski distance to the origin. This results in a total anomaly score for each record. Higher scores signify a higher likelihood of the data point being an anomaly or fraudulent.

FORMULA:

The Z-score outliers anomaly detection algorithm uses the Minkowski distance formula to calculate the fraud score. The formula measures the total anomaly level for each data point.

$$s_i = \left(\sum_n |z_n^i|^p \right)^{1/p}$$

Score for record i

2. **AUTOENCODER:**

We also use an autoencoder, a type of neural network, for anomaly detection. The autoencoder maps each record to itself. Once trained, the autoencoder reconstructs both normal and potentially fraudulent transactions. The difference or error margin between the original and reconstructed data serves as a fraud score. Genuine records are reconstructed efficiently, but fraudulent ones will have a larger error margin, making it easier to detect anomalies or potentially fraudulent behavior.

FORMULA:

$$D(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^q \right)^{1/q}$$

We use a combination of PCA, Minkowski distances, and an autoencoder-based approach to identify anomalies. This approach gives us a numerical measure of the anomaly scores.

To find the best fraud detection model, we average both scores, which ensures a balance between true positives and false positives in the final output.

SCORE BINNING:

After both scores are calculated, we use quantile binning to scale and combine the data. This step is crucial for integrating outputs from multiple models or variables. It divides the data into bins with approximately equal observations or proportions. This strategy minimizes the influence of outliers and extreme values, enhancing the reliability and robustness of the analysis.

Results

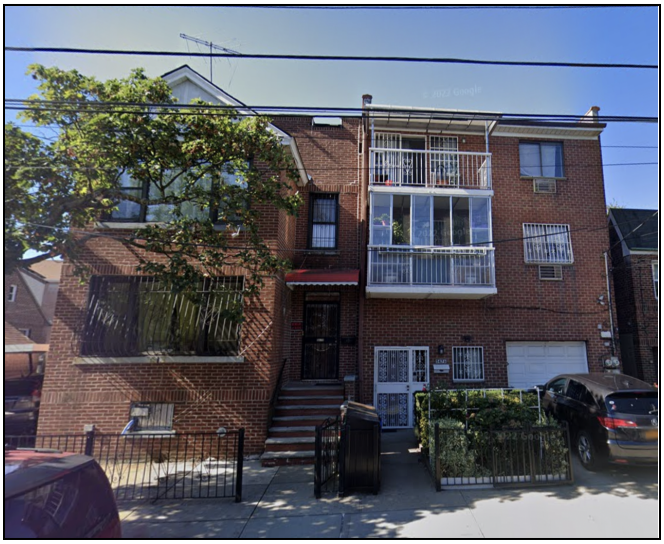
Once the final scores were computed, properties were sorted in decreasing order, giving priority to those with high inference values. Each property with significant inference was then cross-checked manually for added accuracy. In a further step, I used Google Earth to visually validate the actual properties. Interestingly, many of the properties I reviewed showed inconsistencies or discrepancies when compared to the data on record. Consequently, the properties flagged as potentially fraudulent were reported for further investigation.

1. **Property Record: 658933**

This property is a residential building located at 54-76 83 STREET. This property is owned by an individual, WAN CHIU CHEUNG. The initial examination of this record does not reveal any evident missing values, making it appear normal at first glance. However, this record has been marked as unusual, which necessitates a more thorough review of the variables.

The record includes data such as LOT FRONTAGE (LTFRONT) and LOT DEPTH (LTDEPTH), both of which are typically expected to be larger than the building dimensions (BLDFRONT and BLDDEPTH). However, in this case, we observe an anomaly; the dimensions of the building are larger than those of the lot. This discrepancy suggests a possible data entry error or potentially fraudulent activity, as such an occurrence is practically impossible.

Field	Value
OWNER	Wan Chiu Cheung
LTFRONT	25.0 ft
LTDEPTH	100.0 ft
STORIES	3
FULLVAL	$776,000.00
AVLAND	$26,940.00
AVTOT	$46,560.00
STADDR	54-76 83 Street
BLDFRONT	2,500 ft
BLDDEPTH	5,600 ft



Moreover, upon diving deeper into the value-to-size ratios (r8inv and r9inv), alongside the corresponding values grouped under zip code (r2inv_zip5, r3inv_zip, r5inv_zip, etc.), and tax class categories (r2inv_taxclass, r3inv_taxclass), we find these figures to be flagged as unusual. The value of the size_ratio variable, in particular, is strikingly high for this property.

r8inv: 2.919099 x 10^2
r8inv_zip5: 5.148627 x 10^3
r8_taxclass: 2.072328 x 10^-8

r9inv: 8.274232 x 10^2
r9inv_zip5: 5.125755 x 10^3
r9_taxclass: 2.072055 x 10^-8

Additionally, another aspect of this record that raises concerns is the extraordinarily high monetary values per tax class and zip code categories. Such unusually high figures call into question the legitimacy of the stated property values and underline the necessity for further investigation.

2. **Property Record: 1053859**

The property in question is located on DISCALA LANE and is classified as a residential building. Several irregularities have been identified in the associated record from the dataset, which include both missing and zero values. Most notably, the building dimensions fields, namely BLDFRONT and BLDDEPTH, are recorded as zero, which is atypical.

Field	Value
LTFRONT	1000
LTDEPTH	1000
STORIES	2
FULLVAL	450
AVLAND	203
AVTOT	203
STADDR	DISCALA LANE
BLDFRONT	0
BLDDEPTH	0

These inconsistencies are reflected in the heatmap analysis, revealing how they affect the record. Variables such as r1inv, r4inv, r7inv, and their respective groupings under zip5 and taxclass appear to be irregular.

To illustrate:
r1inv: 1818.181818
r4inv: 3300.330033
r7inv: 3300.330033

On closer inspection, it is clear that while the building dimensions (BLDFRONT, BLDDEPTH) are recorded as zero, the lot dimensions (LTFRONT, LTDEPTH) are recorded as non-zero, alongside a likely incorrect STORIES value. This inconsistency could stem from a simple data entry error, or it may suggest potential fraudulent activity. Hence, this property has been flagged for potential fraud due to these oddities.

Moreover, the value-to-size ratio's inverse is unusually high, further emphasizing this record as an anomaly. This irregularity can be attributed to the property's low dollar values, causing these ratios to skyrocket. On revisiting the original fields, it's clear that the monetary values – FULLVAL, AVLAND, and AVTOT – are seemingly underreported for this property. Therefore, questions arise about the legitimacy of this property record due to these unusual $ values.

3. Property Record: 1067360

This property is situated at 20 Emily Court and is referenced by the record number 1067360. Intriguingly, there's no owner indicated in the property record. However, using Google Street View, we could infer that this is likely a private residential property.

Field	Value
LTFRONT	1
LTDEPTH	1
STORIES	2
FULLVAL	836,000
AVLAND	28,800
AVTOT	50,160
STADDR	20 EMILY COURT
BLDFRONT	36
BLDDEPTH	45

The z-scores of certain variables such as R1 (V1/S1), R4 (V2/S1), and R7 (V3/S1) exhibit significant deviation. These are calculated based on the lot front (LTFRONT) and depth (LTDEPTH)

These ratios have unusually high z-scores, which could be due to the recorded lot front and depth, both being listed as '1'. It's unusual for a residential property to have such small dimensions, and this impacts the z-scores for these ratios, making them very high when compared with other properties of the same tax class and in the same zip code.

Adding to the anomaly is the size_ratio, which is calculated by dividing the building size (front and depth) by the lot size. In this case, the building's front and depth are 36 and 45, respectively, yielding a size_ratio significantly higher than 1, suggesting that the building's dimensions exceed the lot's dimensions - a physical impossibility.

This property's lot depth and lot front values are the main factors driving the variation in these variables. This could be a simple data entry error, or it could potentially be a sign of fraudulent activity. To ensure accuracy and maintain the integrity of our property database, a more thorough investigation is needed for this property.

4. **Property Record: 956520**

This record pertains to a privately owned residential property registered under the name of Trompeta Rizalina. On initial examination, there are several anomalies that raise suspicion. The first noticeable discrepancy is the exceptionally high ratio of building size to lot size. Detailed analysis shows that the documented frontage and depth measurements of the building significantly exceed those of the lot, a situation which is architecturally impossible.

Field	Value
OWNER	TROMPETA RIZALINA
LTFRONT	91
LTDEPTH	3
STORIES	25
FULLVAL	$348,200
AVLAND	$15,600
AVTOT	$20,892
STADDR	12 ONEIDA AVENUE
BLDFRONT	1812
BLDDEPTH	5020

Further compounding suspicions about this property are the unusually high ratios of property value to size. The inverses of the ratios R2 = V1/S2, R3 = V1/S3, R5 = V2/S2, R6 = V2/S3, R7 = V3/S2, and R8 = V3/S3 have extremely high z-scores for this property. This indicates that the reported values for this property are considerably low in comparison to the building front and depth values, as categorized under Tax Class 1.

These anomalies suggest two possible issues. First, the reported property prices may be fraudulent, intentionally undervalued to circumvent proper tax payments. Second, the recorded property size could be inflated, creating a distorted perception of the property's value. Due to these potential issues, further investigation into this property is warranted.

5. **Property Record: 293330**

The property under investigation is located at 95 Prospect Park West. Owned by the City of NY Parks, this mansion is identified in the record as having 2 stories. However, upon visual inspection, it's clear that the actual number of stories surpasses this number.

Field	Value
OWNER	CITY OF NY PARKS
LTFRONT	526
LTDEPTH	250
STORIES	2
FULLVAL	270500800
AVLAND	115650000
AVTOT	121725360
STADDR	95 PROSPECT PARK WEST
BLDFRONT	0
BLDDEPTH	0

Several discrepancies were noted in this property's record. Missing values were found, mainly concerning tax exemption fields. Most notably, the 'BLDFRONT' and 'BLDFRONT' fields, representing the building's frontage and depth respectively, were recorded as zero. This is atypical and affects the integrity of the record.

The heatmap analysis further indicates significant anomalies in several variables, namely 'r2', 'r3', 'r5', 'r6', 'r2inv', 'r3inv', 'r5inv', and 'r6inv', including their grouped values under 'taxclass'. The values for these variables appear as:

r2: 2,705,008,000,000.0
r3: 901,669,333,333.33
r5: 1,156,500,000,000.0

r2inv: 0
r3inv: 0
r5inv: 0

Upon examining the property's initial record, we discovered that while the lot dimensions ('LTFRONT', 'LTDEPTH') were non-zero, the building dimensions ('BLDFRONT', 'BLDDEPTH') were indeed recorded as zero. In conjunction with the incorrect 'STORIES' value, this discrepancy could be due to a data entry error or potentially suggest fraudulent activity.

Summary

This project aimed to identify potential tax fraud cases within a dataset of one million New York property records. These records included 32 fields such as property characteristics, owner details, address, and assessed tax values. An integral part of the project's initial phase was thorough data cleaning to address missing values and rectify any inconsistencies or errors in the data. Missing values were tackled using suitable imputation techniques, ensuring a robust dataset for further analysis.

To manage the high-dimensional nature of the dataset, we applied Principal Component Analysis (PCA), reducing the complexity of the 32 variables while preserving important information. This step of dimensionality reduction allowed for a more streamlined analysis and facilitated efficient anomaly detection for tax fraud identification.

Next, we calculated a fraud score for each record using the Minkowski distance. This score measured the degree of deviation of each property from the general pattern observed in the dataset, helping pinpoint potential tax fraud instances. Each property was evaluated against the overall statistical distribution of the dataset, enabling us to identify outliers or suspicious property values indicative of potential fraud.

Lastly, we employed quantile binning to standardize and integrate the fraud scores. This process transformed the continuous fraud scores into discrete bins based on their distribution, putting the scores on a common scale. The standardization and equal distribution of these scores across bins made it easier to compare and merge results from different models or variables, thereby enhancing the efficacy of the overall analysis.

Appendix

Data Quality Report

1. Data Description

The dataset appears to be a comprehensive record of real estate properties located in New York City, published by the **Department of Finance**. The data comprises various essential details about each property, including a unique identification number, location information such as borough, block, and lot for the **year 2010/11**. This data can be beneficial for businesses involved in the real estate industry to gain insights into property trends and make informed decisions related to investments, acquisitions, and sales. The dataset contains **1,070,994 records** across **32 fields**.

2. Statistics Tables

Numeric Fields Table

Field Name	# Records Have Values	% Populated	% Zeros	Min	Max	Mean	Standard Deviation	Most Common
LTFRONT	1,070,994	100.00%	15.79%	0	9,999	36.63	74.03	0
LTDEPTH	1,070,994	100.00%	15.89%	0	9,999	88.86	76.4	100
STORIES	1,014,730	94.75%	0.00%	1	119	5.00	8.37	2
FULLVAL	1,070,994	100.00%	1.21%	0	6,150,000,000	874,264.50	11,582,425.58	0
AVLAND	1,070,994	100.00%	1.21%	0	2,668,500,000	85,067.91	4,057,258.16	0
AVTOT	1,070,994	100.00%	1.21%	0	4,668,308,947	227,238.16	6,877,526.09	0
EXLAND	1,070,994	100.00%	45.91%	0	2,668,500,000	36,423.89	3,981,573.93	0
EXTOT	1,070,994	100.00%	40.39%	0	4,668,308,947	91,186.98	6,508,399.78	0
BLDFRONT	1,070,994	100.00%	21.36%	0	7,575	23.04	35.58	0
BLDDEPTH	1,070,994	100.00%	21.37%	0	9,393	39.92	42.71	0
AVLAND2	282,726	26.40%	0.00%	3	2,371,005,000	246,235.71	6,178,951.64	2,408

AVTOT2	282,732	26.40%	0.00%	3	4,501,180,002	713,911.43	11,652,508.34	750
EXLAND2	87,449	8.17%	0.00%	1	2,371,005,000	351,235.68	10,802,150.91	2,090
EXTOT2	130,828	12.22%	0.00%	7	4,501,180,002	656,768.28	16,072,448.75	2,090

Categorical Fields Table

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
RECORD	1,070,994	100.00%	0	1,070,994	1
BBLE	1,070,994	100.00%	0	1,070,994	1000010101
BORO	1,070,994	100.00%	0	5	4
BLOCK	1,070,994	100.00%	0	13,984	3944
LOT	1,070,994	100.00%	0	6,366	1
EASEMENT	4,636	0.43%	0	12	E
OWNER	1,039,249	97.04%	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100.00%	0	200	R4
TAXCLASS	1,070,994	100.00%	0	11	1
EXT	354,305	33.08%	0	3	G
EXCD1	638,488	59.62%	0	129	1017
STADDR	1,070,318	99.94%	0	839,280	501 SURF AVENUE
ZIP	1,041,104	97.21%	0	196	10314
EXMPTCL	15,579	1.45%	0	14	X1
EXCD2	92,948	8.68%	0	60	1,017
PERIOD	1,070,994	100.00%	0	1	FINAL
YEAR	1,070,994	100.00%	0	1	2010/11
VALTYPE	1,070,994	100.00%	0	1	AC-TR

3. Visualization of each Field

1) Field Name: RECORD

Description: The RECORD column is a unique identifier for each property record in the dataset. It contains no zeros and has 100% populated values, with a total of 1,070,994 unique values, which signifies the total number of records in the dataset.

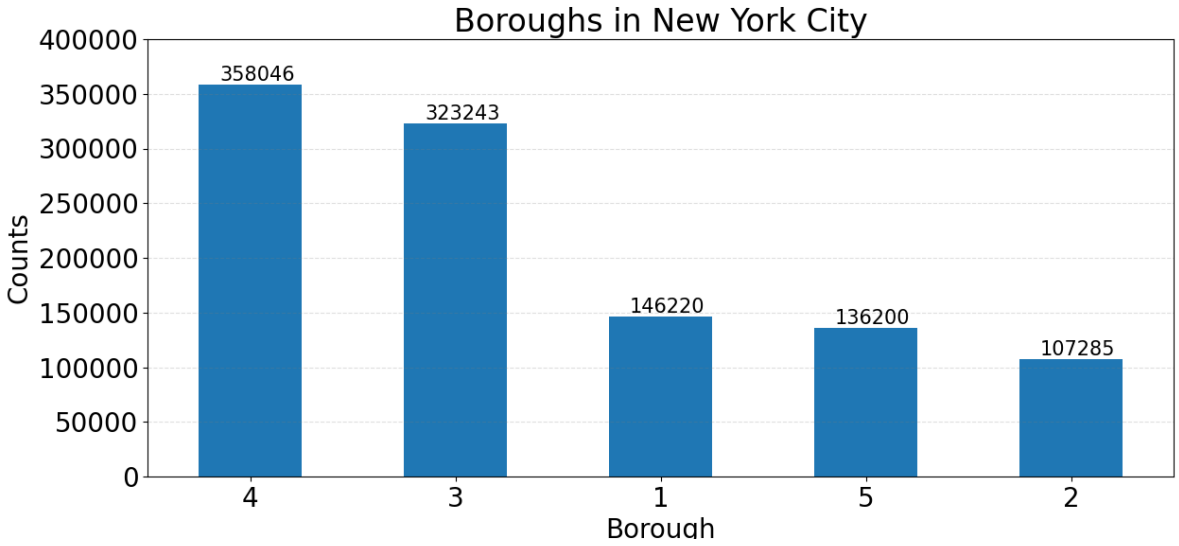
2) Field Name: BBLE

Description: The BBLE column is a categorical column that stands for "borough, block, and lot" and is a unique identifier for each property in New York City. The first digit of the BBLE represents the borough followed by the block number and the lot number. This column is used to link the property information with other datasets.

3) Field Name: BORO

Description: The BORO column contains the borough code where the property is located. There are 5 borough codes and they are assigned like 1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island. This column is used for identifying the location of the property within New York City.

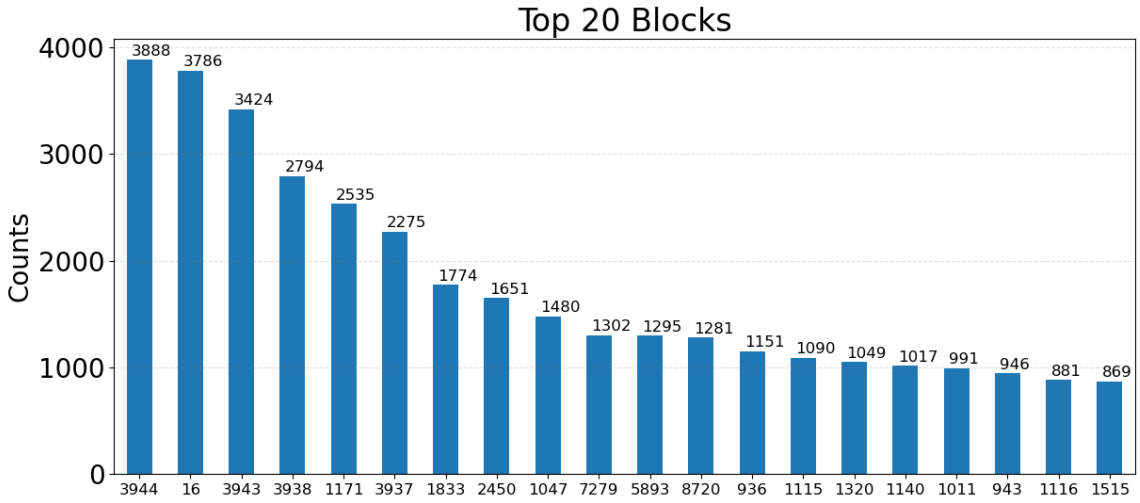
Visualization:



4) Field Name: BLOCK

Description: The BLOCK column in the dataset represents the tax block for each property, a subdivision of a borough used for identifying a group of properties. Each tax block has a unique identification number, and there are a total of 13,984 distinct tax blocks in the dataset.

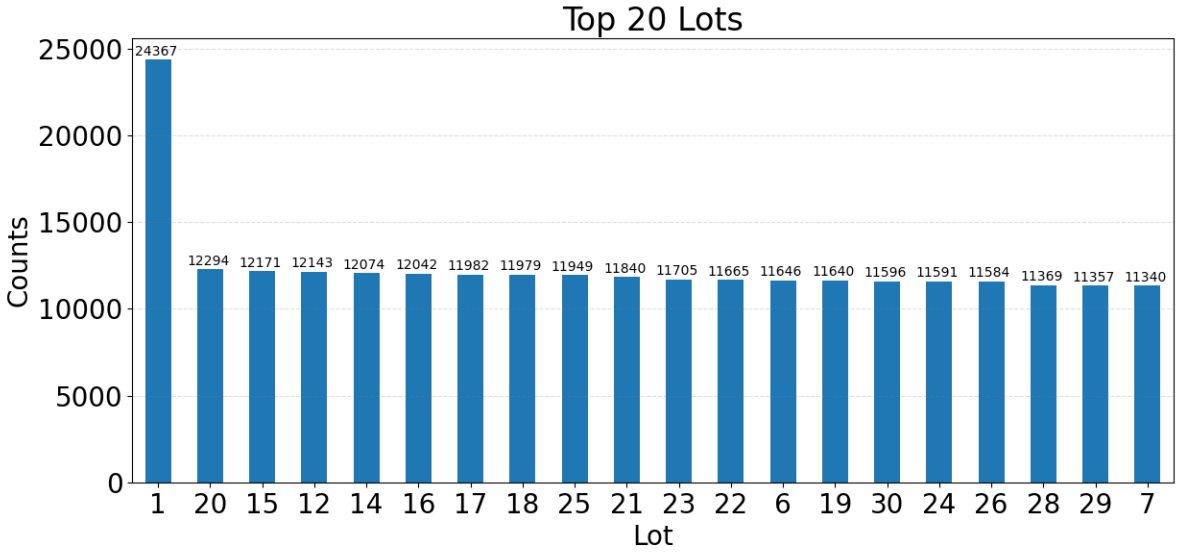
Visualization:

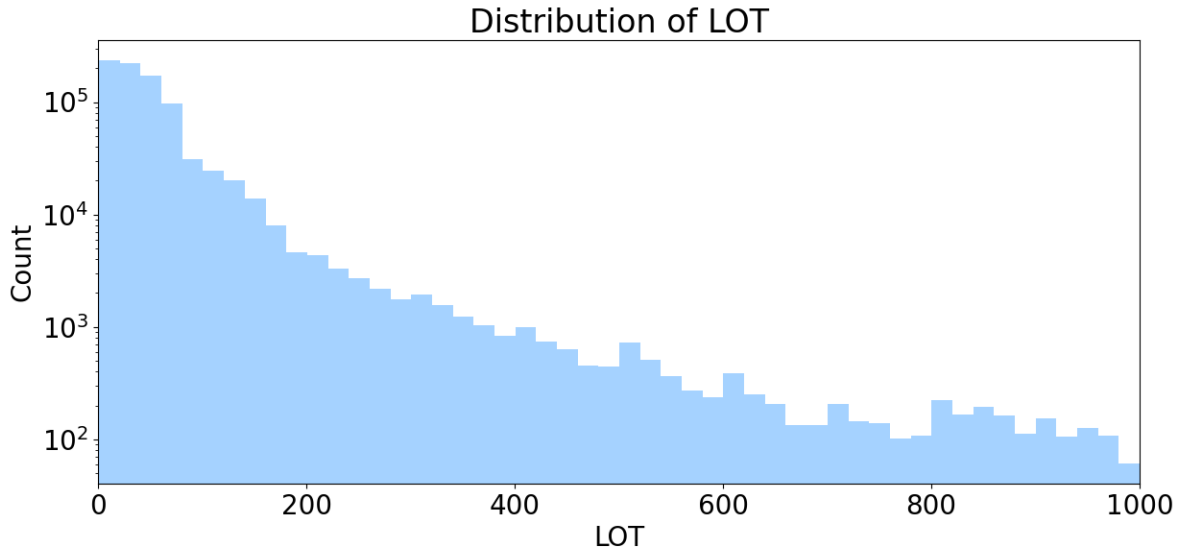


5) Field Name: LOT

Description: The LOT column refers to the tax lot number assigned to the property by the New York City Department of Finance. It is a unique identifier for each tax lot within a block and borough. The LOT number is used for determining ownership, assessed value, and tax liability of each property.

Visualization:

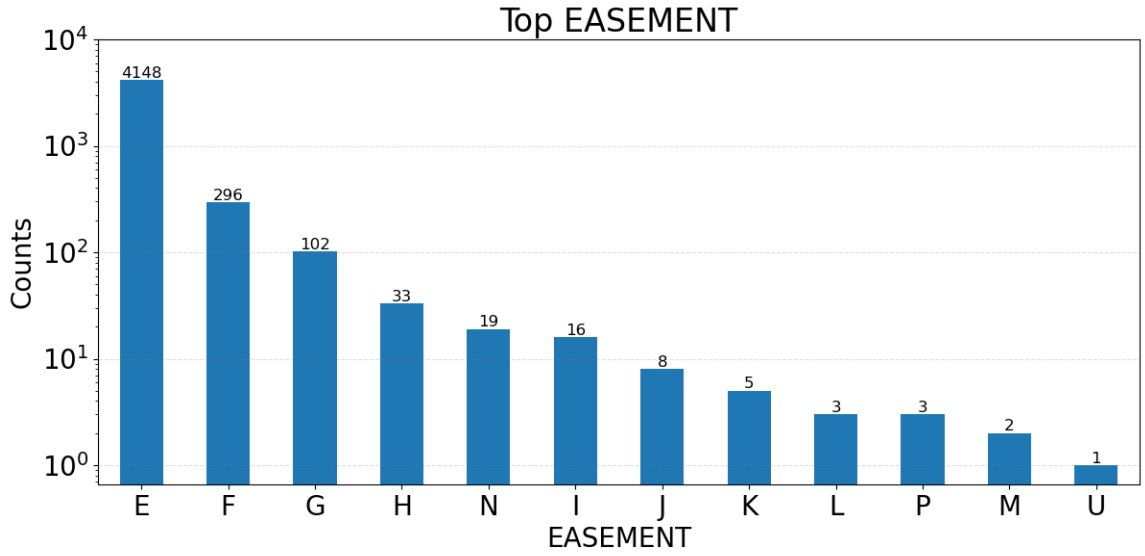




6) Field Name: EASEMENT

Description: The EASEMENT column contains information about any easement associated with a property. This column has 0.43% populated values, indicating that easements are present for only a small percentage of the properties in the dataset. The column has 12 unique values, with the most common value being "E".

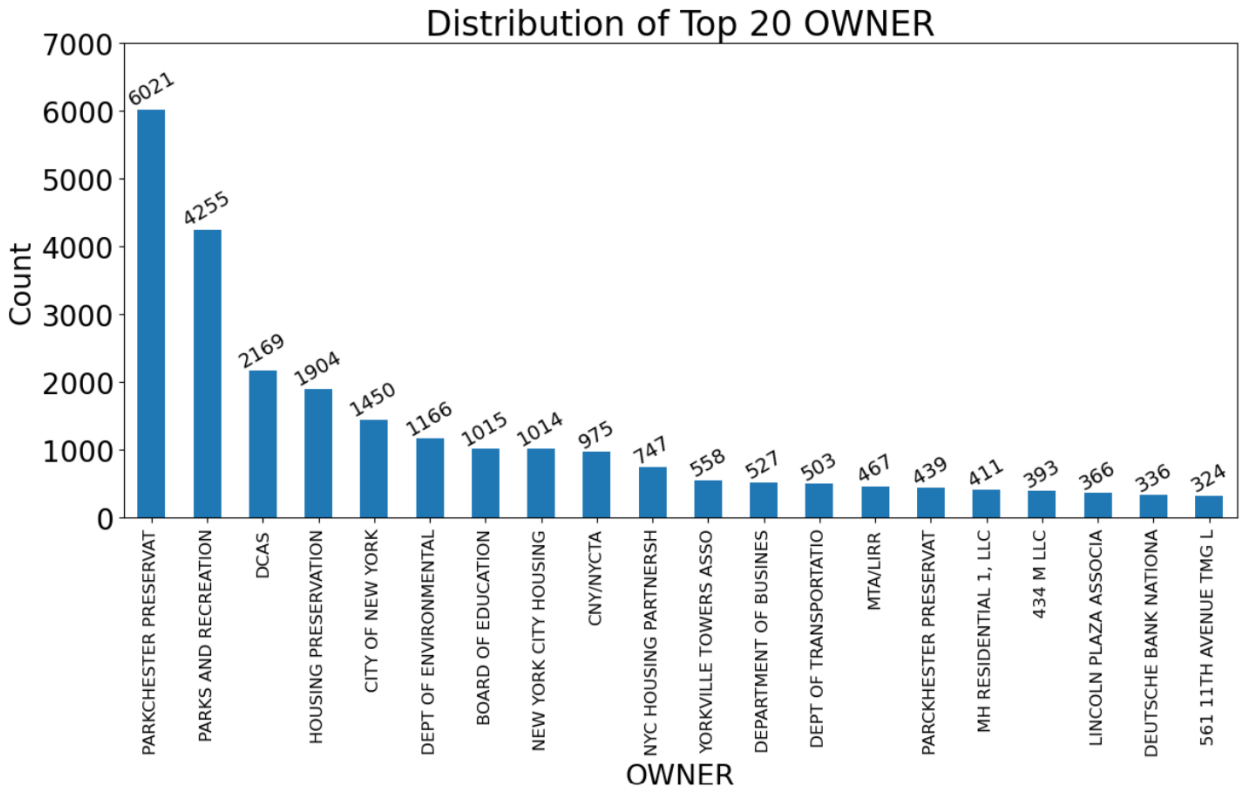
Visualization:



7) **Field Name: OWNER**

Description: The OWNER column in the dataset contains the name of the property owner. It has 1,039,249 records and 97.04% populated. It has 863,347 unique values, and the most common owner is "PARKCHESTER PRESERVAT".

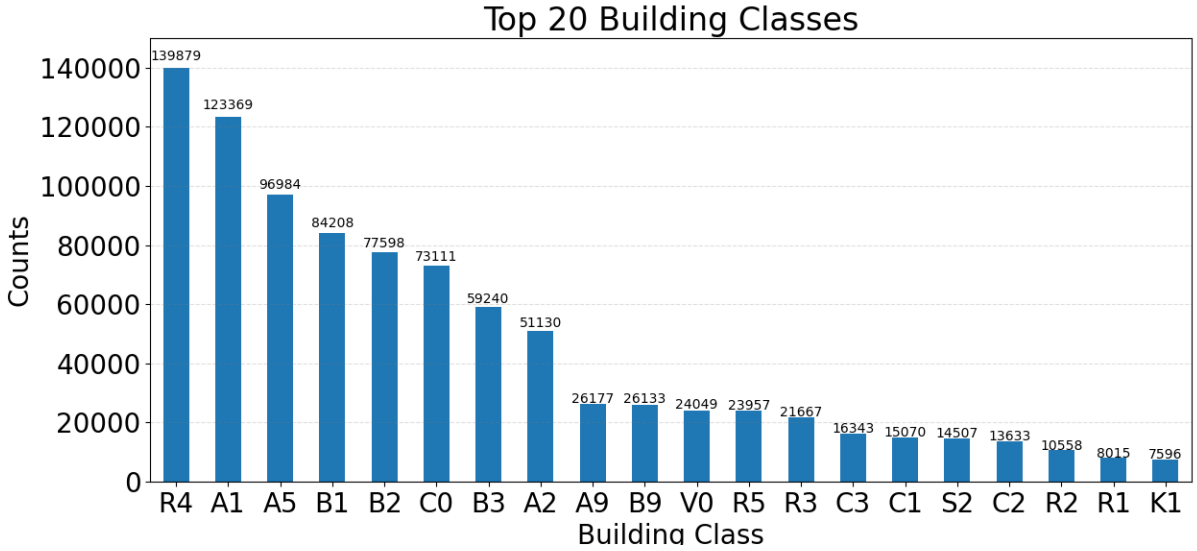
Visualization:



8) **Field Name: BLDGCL**

Description: The BLDGCL column represents the Building Class Code, which is a code used to classify properties based on the type of building they are. The code consists of a two to four-character alpha-numeric characters.

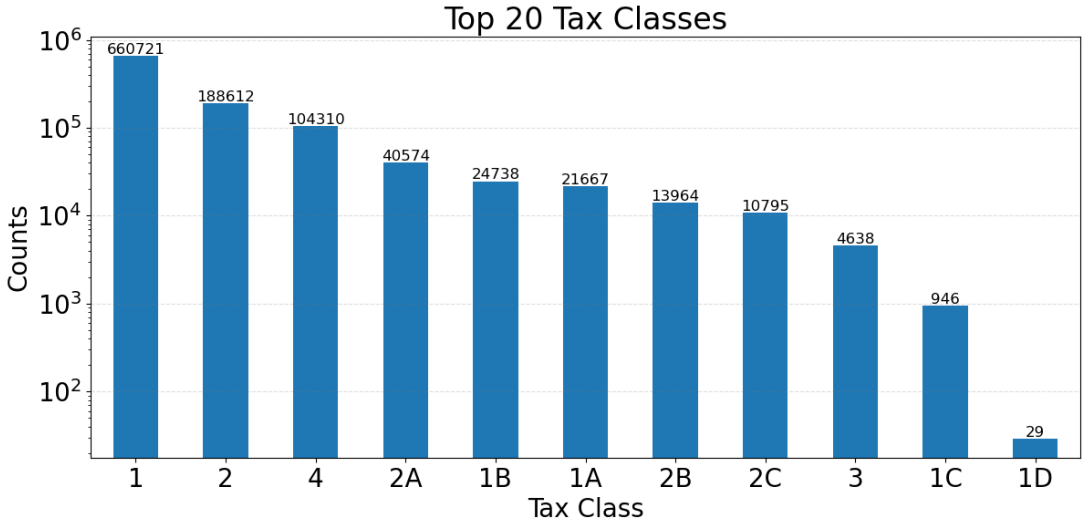
Visualization:



9) Field Name: TAXCLASS

Description: The TAXCLASS column is a categorical column that indicates the tax class of each property. There are 11 unique tax classes represented as integers from 1 to 4, and then 6, 7, 8, and 9, and letters like 'A', 'B', and 'C'.

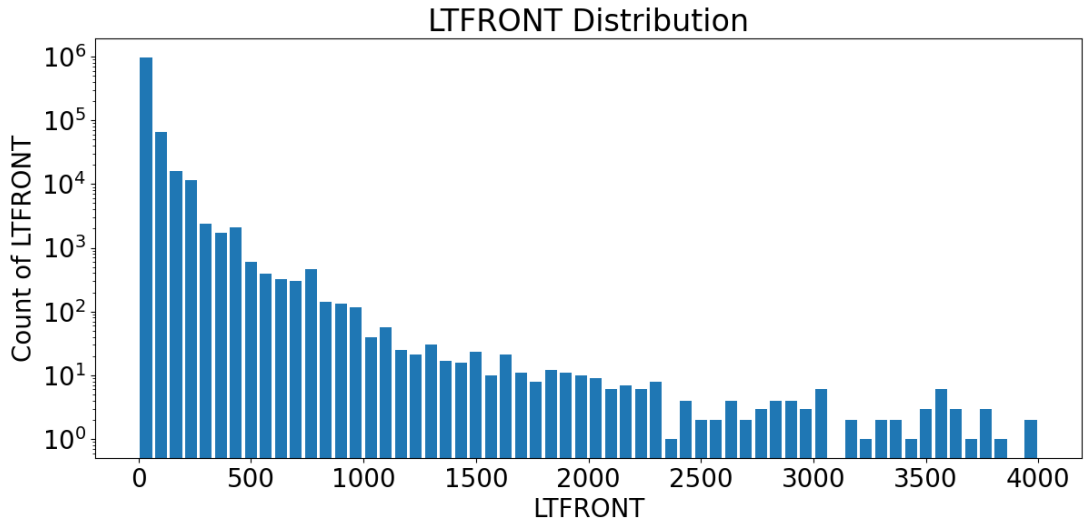
Visualization:

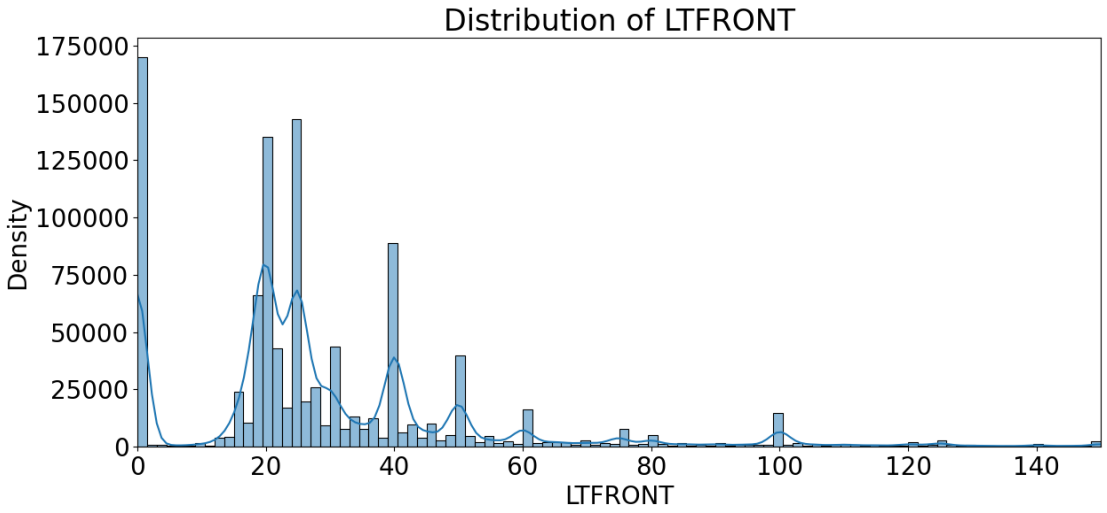


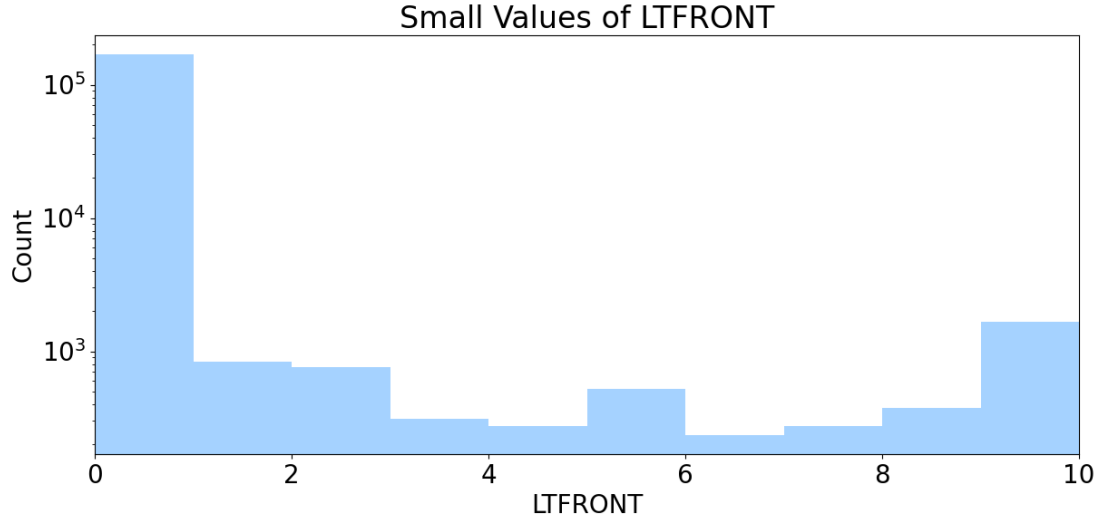
10) Field Name: LTFRONT

Description: The LTFRONT column refers to the lot frontage, the linear feet of the street front of the property. There are 1,070,994 records with values in this column, and it is 100% populated. The minimum value is zero, the maximum value is 9,999, with 15.79% of the values being zero. The mean value is 36.63 feet, and the standard deviation is 74.03 feet.

Visualization:



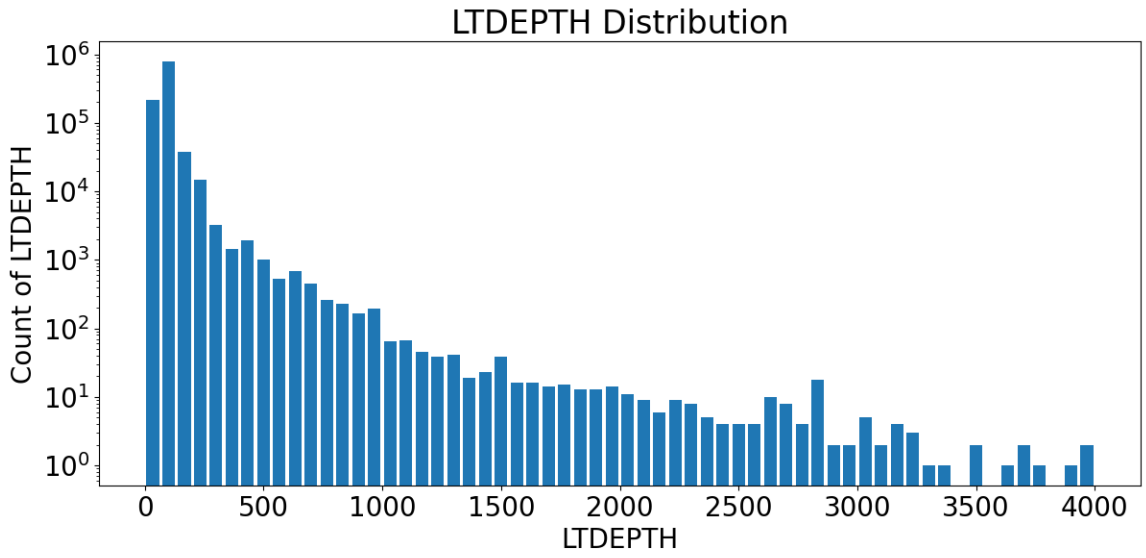


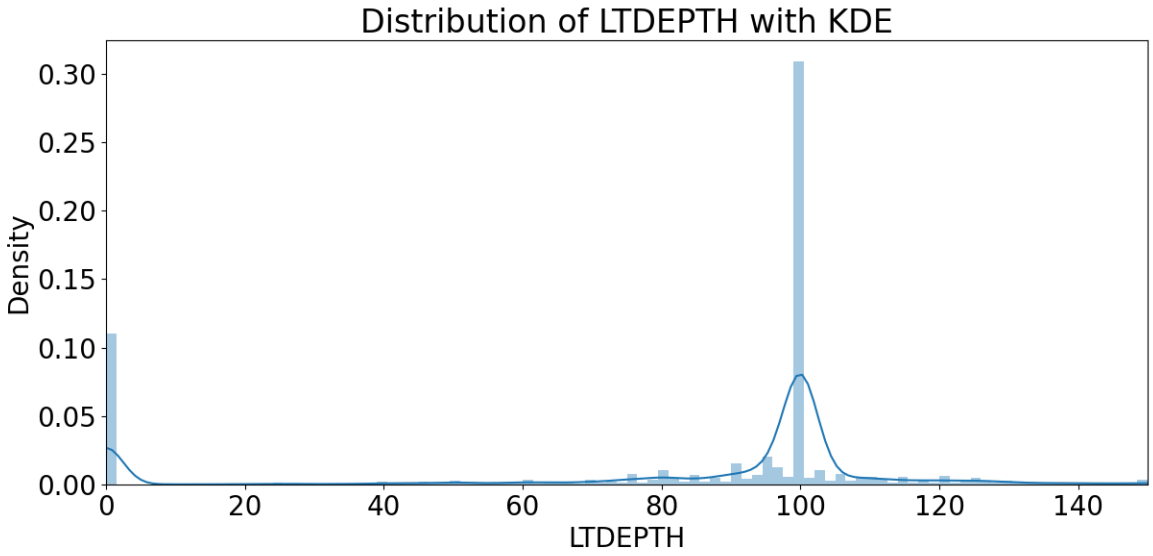


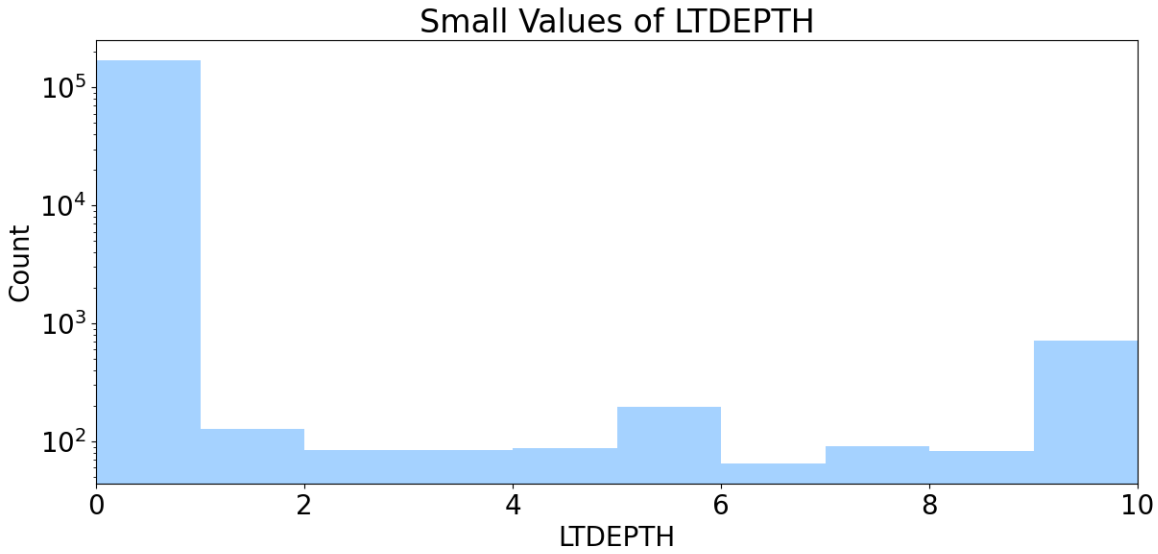
11) Field Name: LTDEPTH

Description: The LTDEPTH column represents the depth of the lot in feet. It has 1,070,994 records with values, and 100% of the records are populated. The minimum value is 0 feet, the maximum value is 9,999 feet, with 15.89% of the values being zero.. The mean depth of the lot is 88.86 feet, and the standard deviation is 76.4 feet.

Visualization:



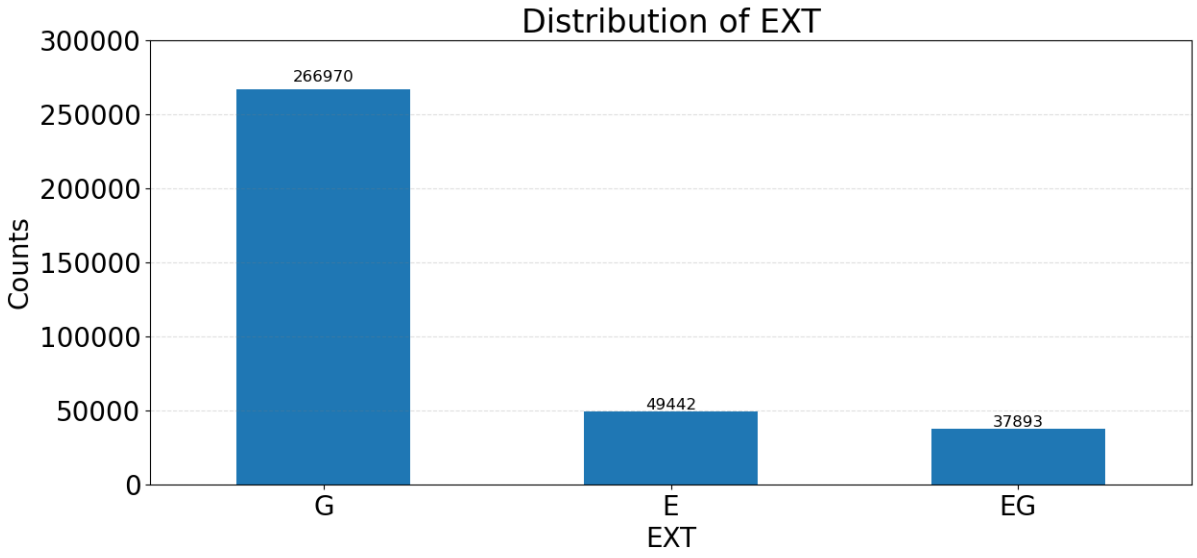




12) Field Name: EXT

Description: The EXT column in the given dataset has 354,305 records with 33.08% populated values. The column represents the Extension code of the building or the property. The unique values in this column are 'G', 'E', and ' ', where 'G' denotes a garage, 'E' denotes an extension, and ' ' denotes the absence of a garage or an extension.

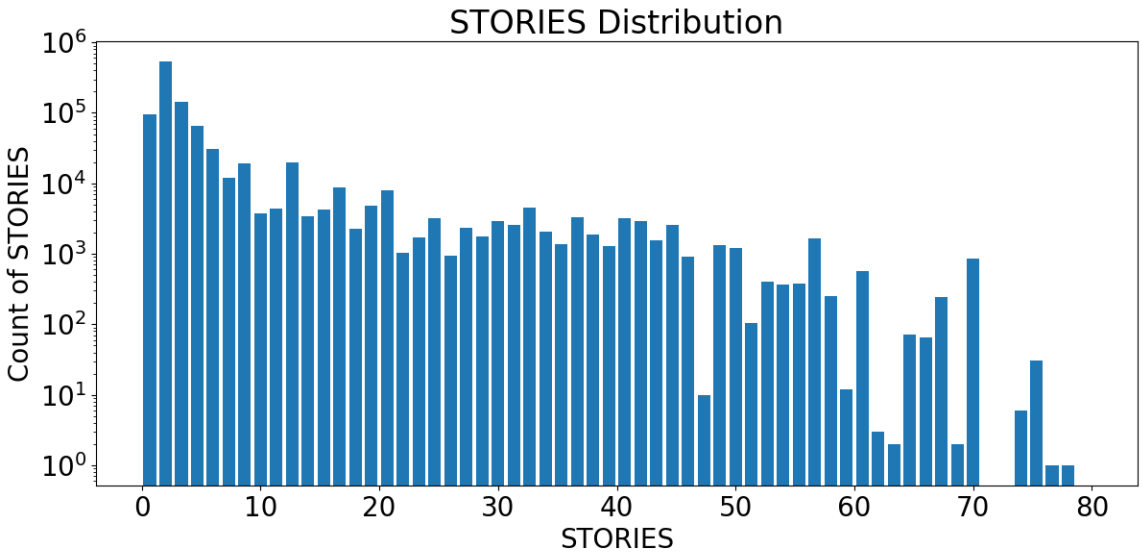
Visualization:

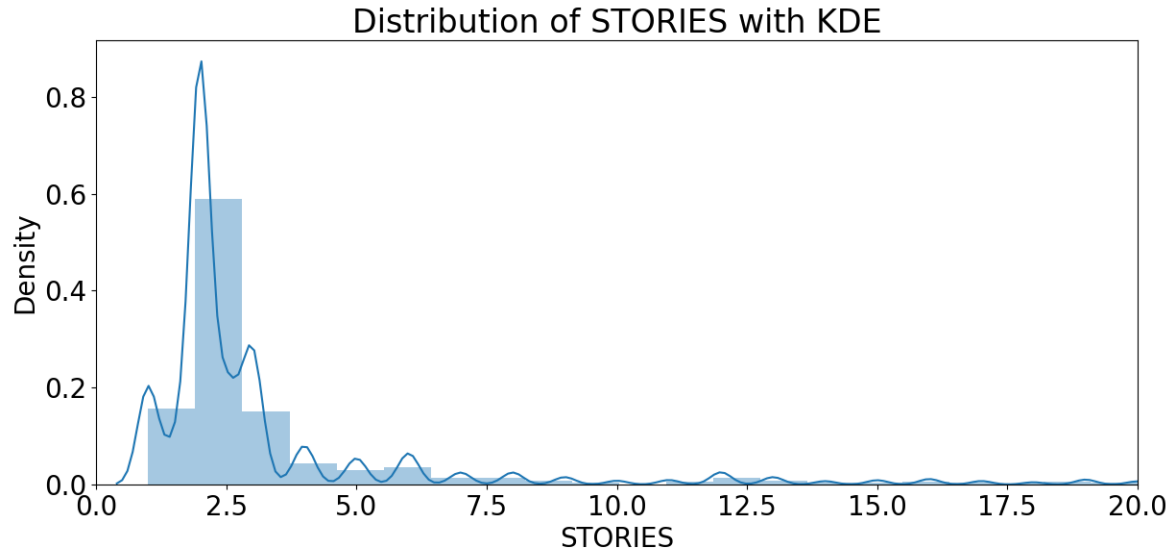


13) Field Name: STORIES

Description: The STORIES column in the dataset contains the number of stories in the building that the property is located in. It is a numerical column with a range of values from 1 to 119. This column has 1,014,730 records with values, and it is 94.75% populated. The most common value in this column is 2, with a mean of 5.00 and a standard deviation of 8.37.

Visualization:

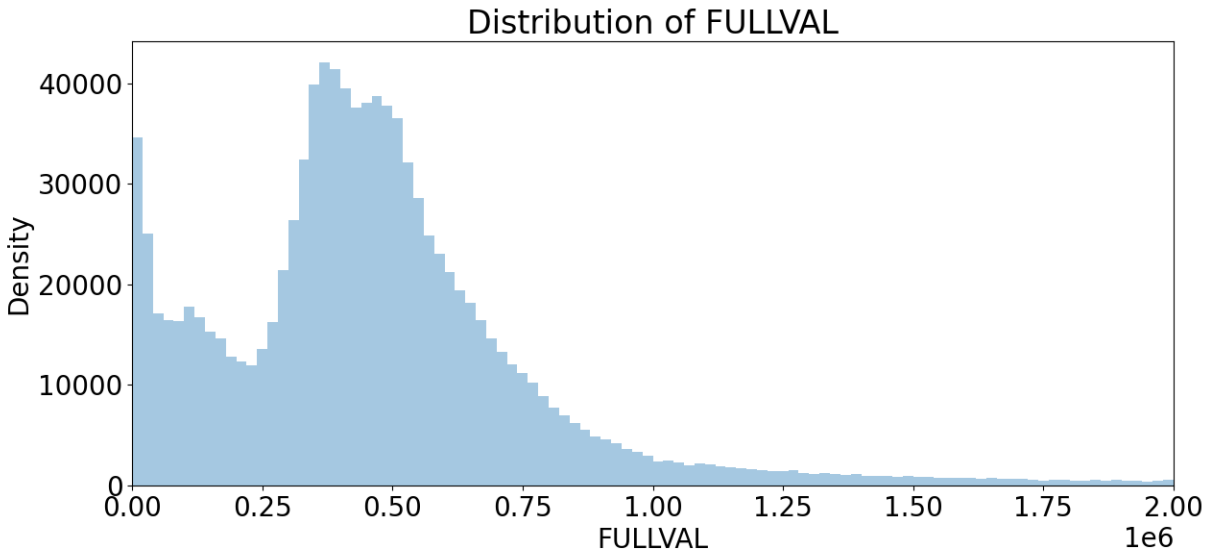


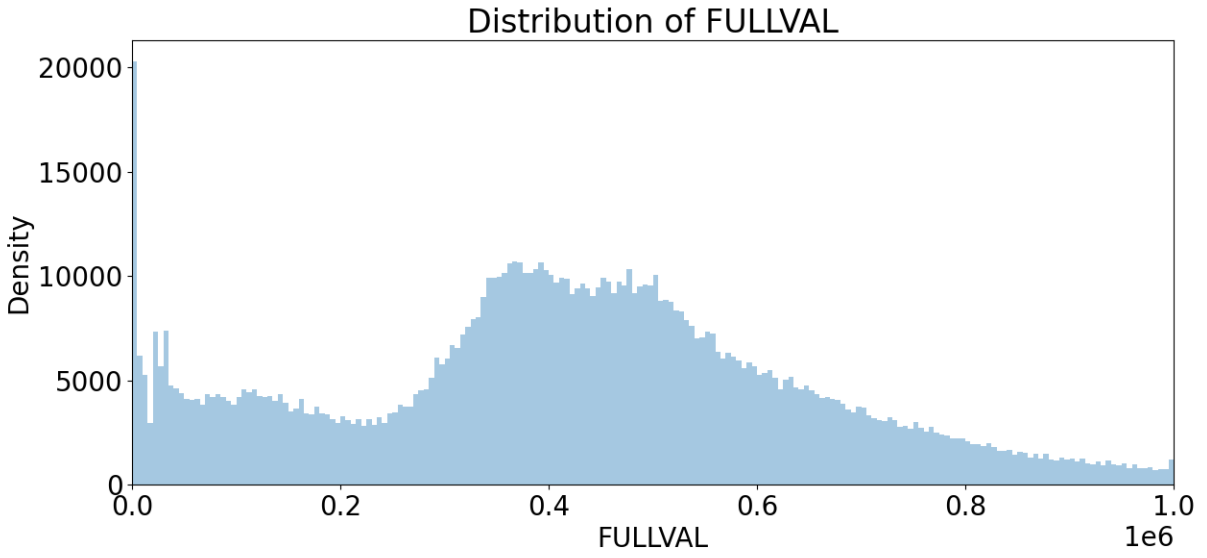


14) Field Name: FULLVAL

Description: The FULLVAL column contains the assessed value of the property for tax purposes. It is expressed in dollars and has a mean value of $874,264.50 and standard deviation of $11,582,425.58. It has a minimum value of 0 and a maximum value of $6,150,000,000.

Visualization:

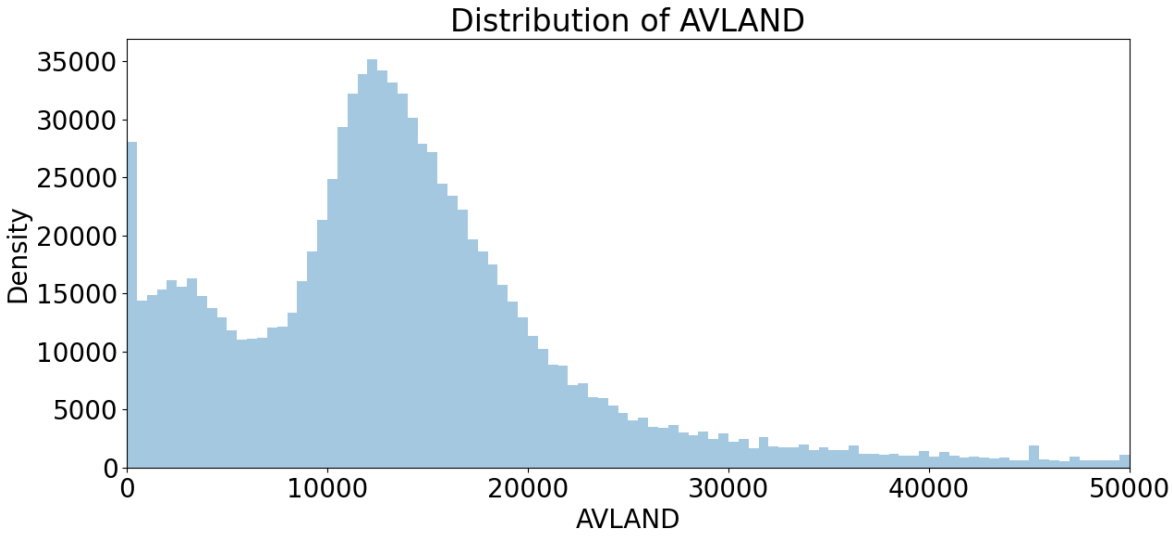




15) Field Name: AVLAND

Description: The AVLAND column represents the assessed value of land for the tax lot. The values range from 0 dollars to a maximum value of 2,668,500,000 dollars. The most common value in this column is also 0 dollars signifying there are many tax lots that do not have any assessed value for land.

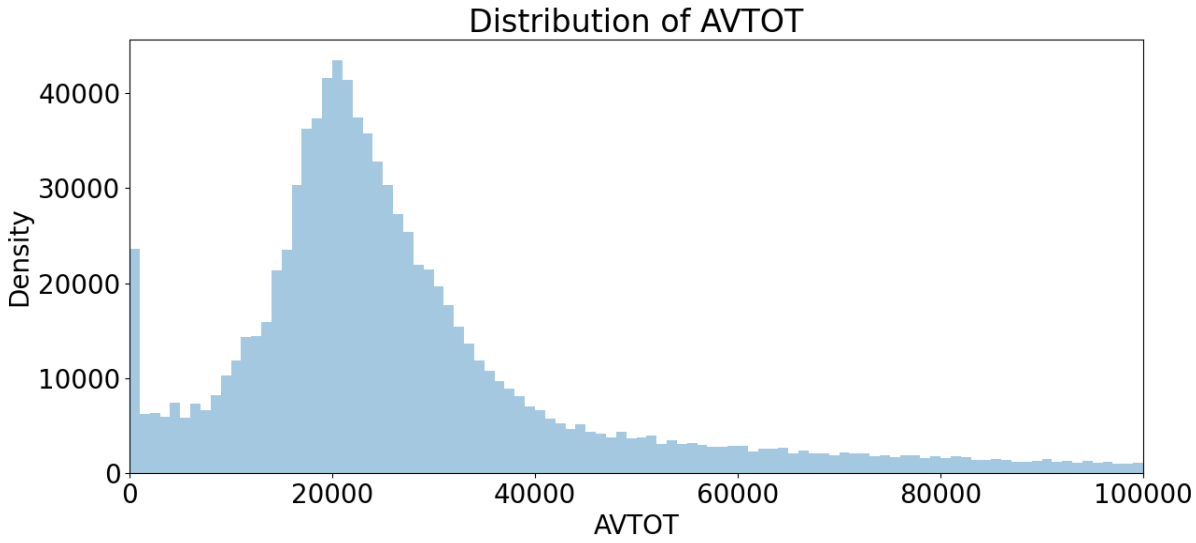
Visualization:



16) Field Name: AVTOT

Description: AVTOT column contains the assessed total value for land and building for each property. The column has 1,070,994 records, and all of them have value. The values range from 0 to 4,668,308,947, with 1.21% values being zero.

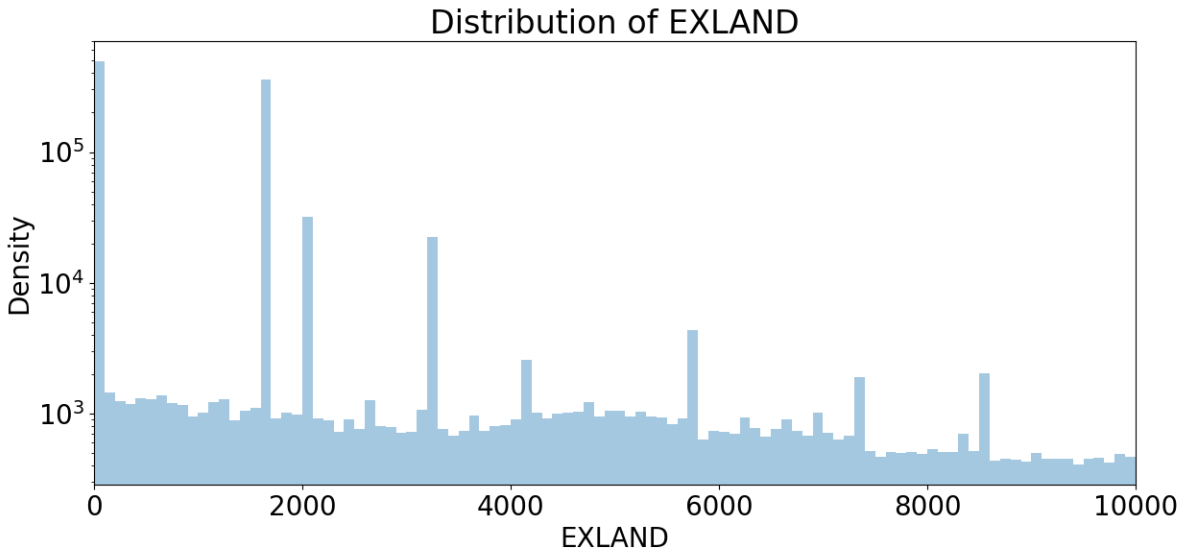
Visualization:



17) Field Name: EXLAND

Description: The EXLAND column contains the assessed value of the land excluding any buildings or structures on it. The mean value in the column is 36,423.89, and the standard deviation is 3,981,573.93.

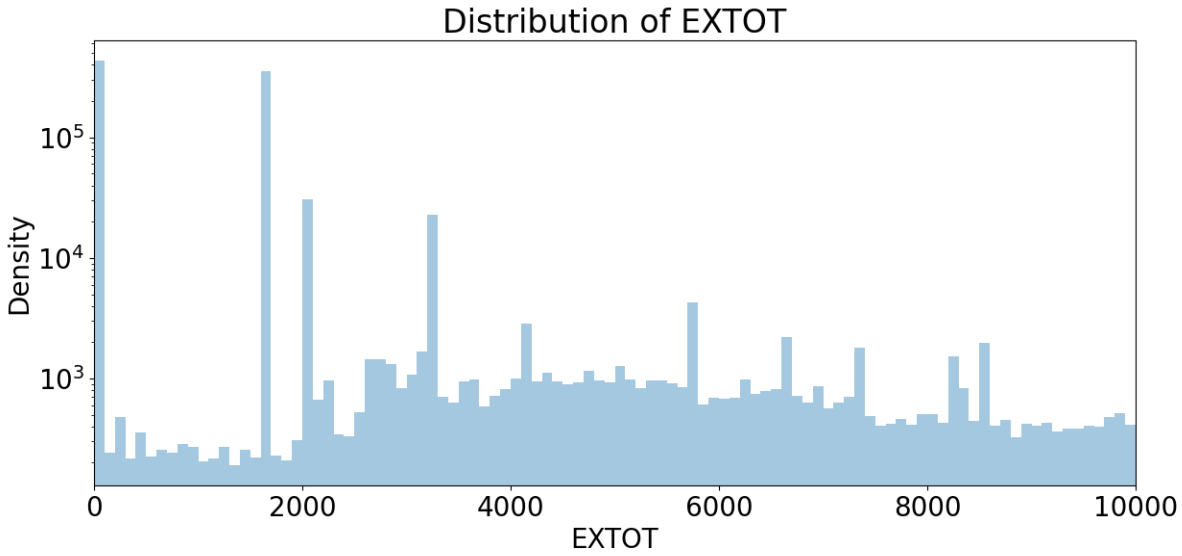
Visualization:



18) Field Name: EXTOT

Description: The EXTOT column contains the total estimated market value of the building's exterior lot area. It has 1,070,994 records, with 40.39% of them having zero values. The mean and standard deviation are 91,186.98 and 6,508,399.78, respectively.

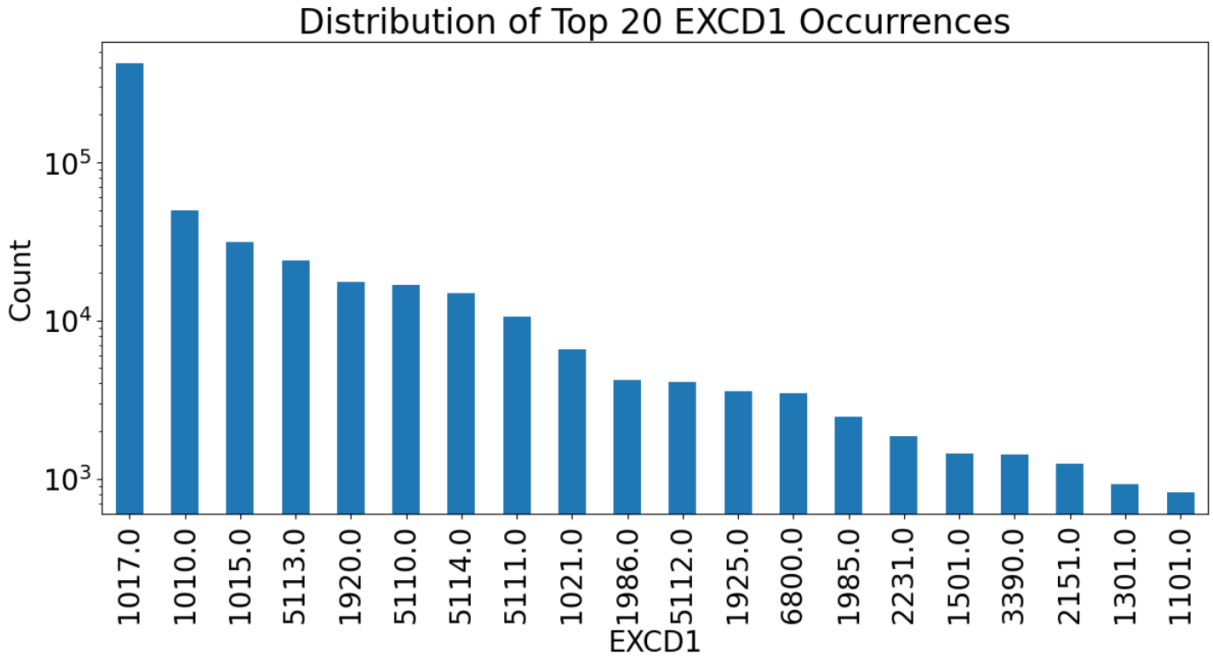
Visualization:

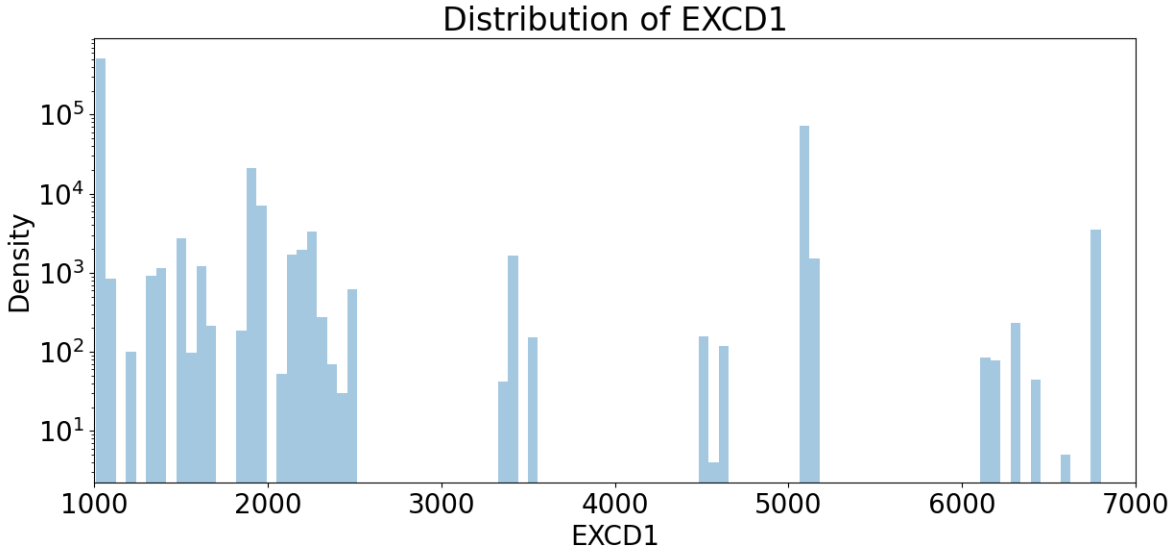


19) Field Name: EXCD1

Description: The EXCD1 column refers to the first exemption code, which is a numeric code that indicates the reason for a property's exemption from taxation. This column has a relatively low level of completeness, with only 45.48% of the records having non-zero values.

Visualization:

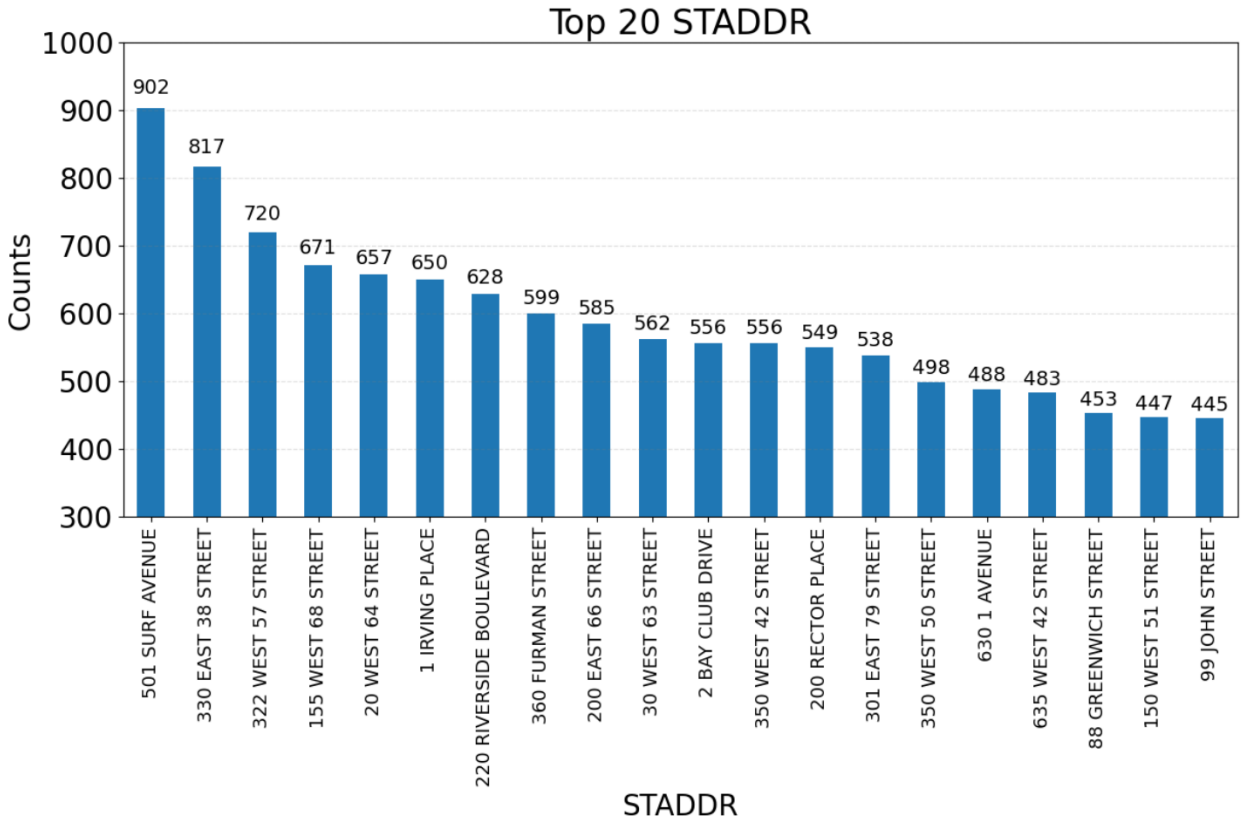




20) Field Name: STADDR

Description: The STADDR column likely represents the street address of each property. Majority of the values are populated with the most common value being “501 SURF AVENUE” appearing 902 times.

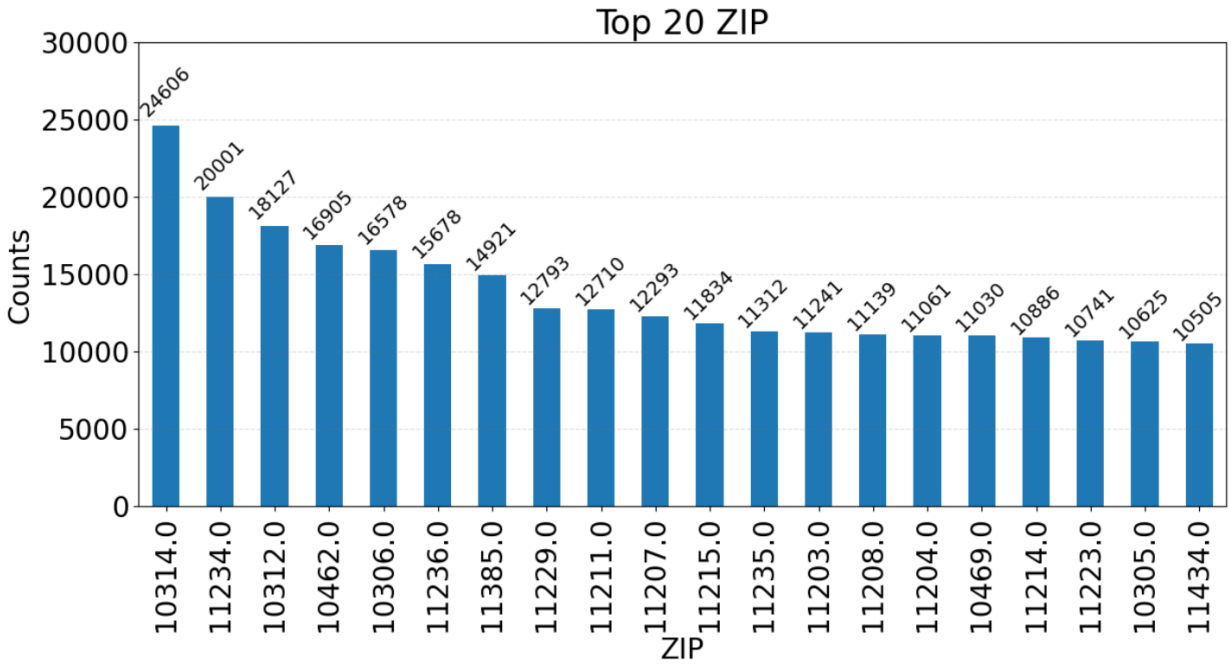
Visualization:



21) Field Name: ZIP

Description: The ZIP column in the dataset contains the postal code associated with each property in New York City. The most common zip code is 10314 with a count of 24606.

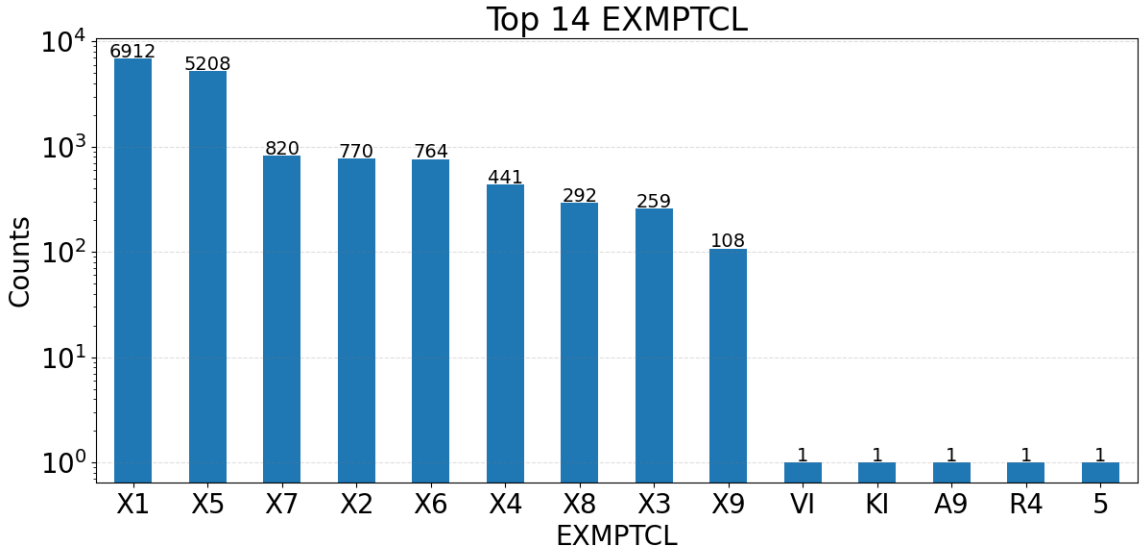
Visualization:



22) Field Name: EXMPTCL

Description: The EXMPTCL column in the dataset refers to the exempt class of the property. The values represent a two-digit code that specifies the type of exemption, such as "1A" for a property owned by a government entity, "2A" for a property used for religious purposes and so on. The most common value is X1 with a total count of 6912.

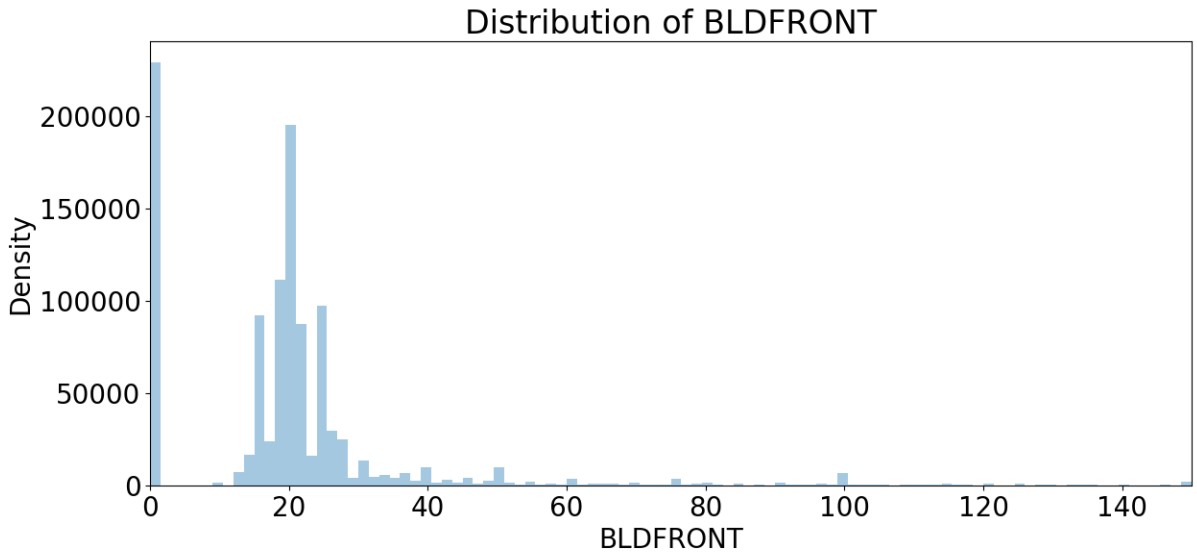
Visualization:

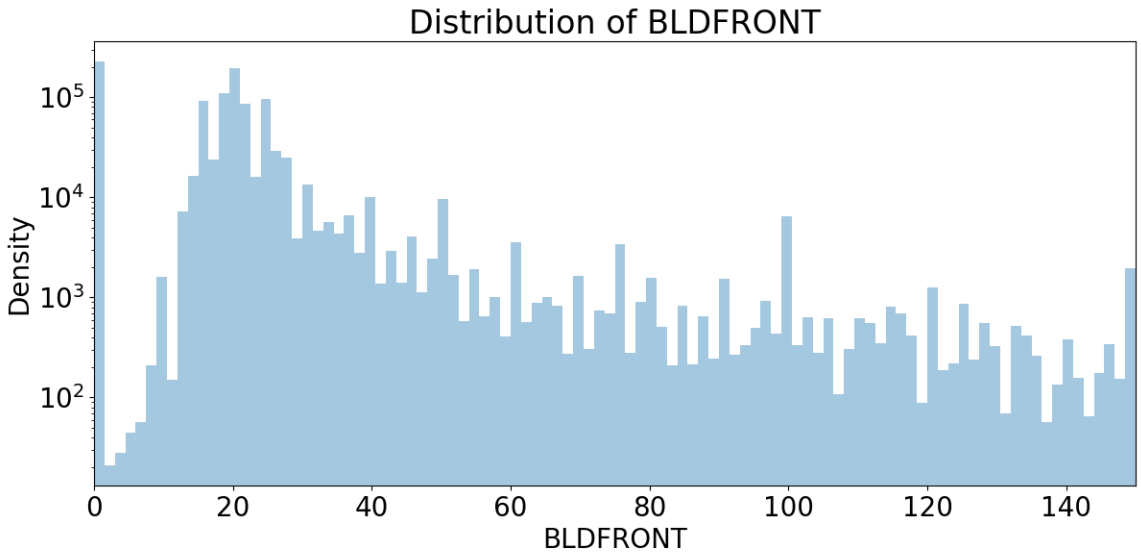


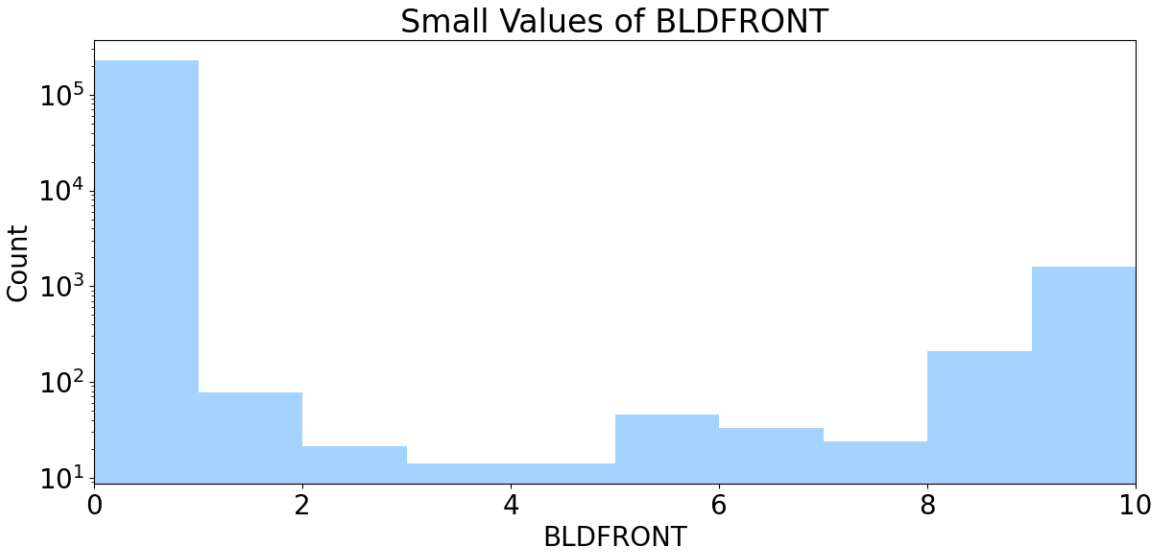
23) Field Name: BLDFRONT

Description: The BLDFRONT column in the dataset contains the number of feet of the building facing the street. It is a continuous numerical column ranging from 0 to 7,575 feet with 1,070,994 records, and 21.36% of zeros. The mean value is 23.04 feet, and the standard deviation is 35.58 feet.

Visualization:



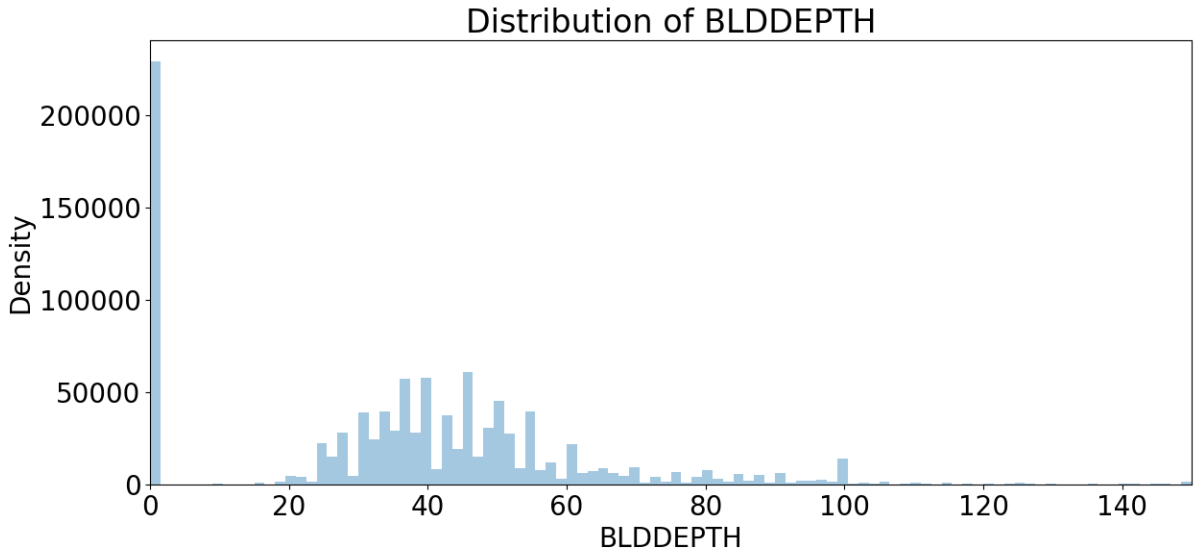


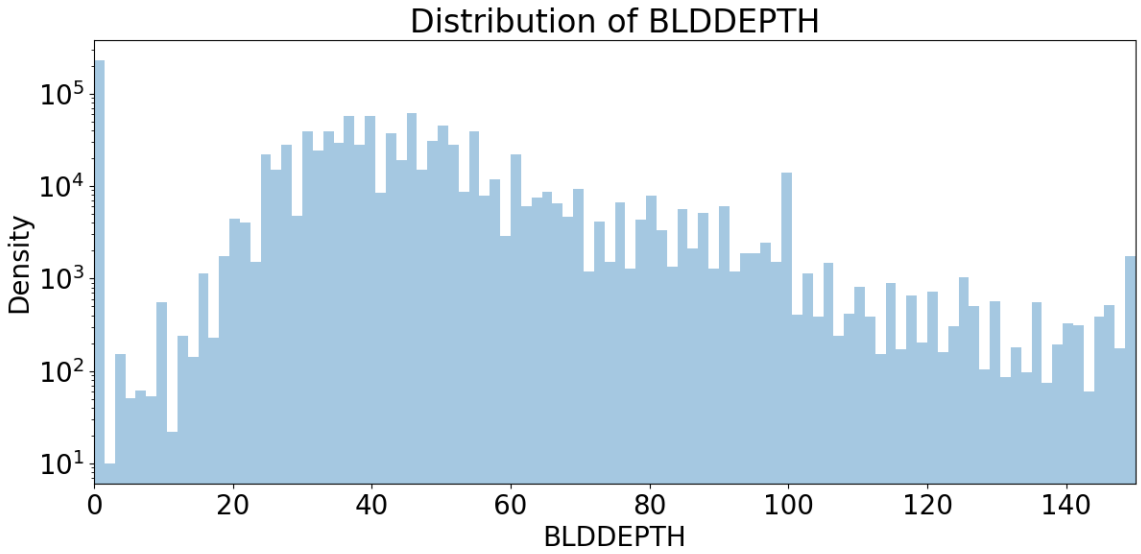


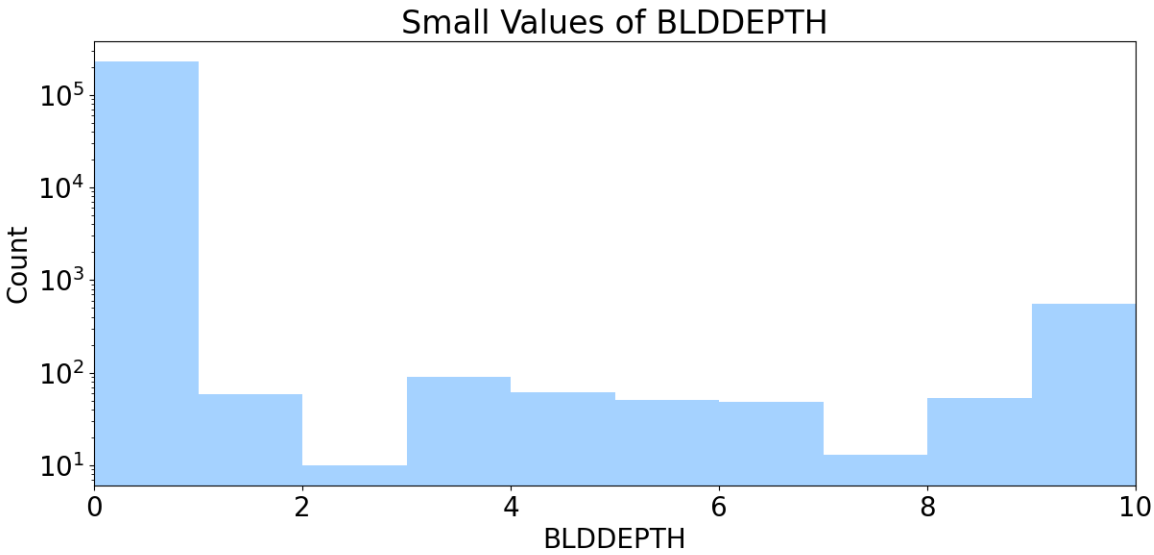
24) Field Name: BLDDEPTH

Description: The BLDDEPTH column contains the building depth in feet. The data in this column is 100% populated, and it has a mean value of 39.92 feet and a standard deviation of 42.71 feet. The minimum value in this column is 0 feet, which suggests that some records have missing values or an unusual entry.

Visualization:



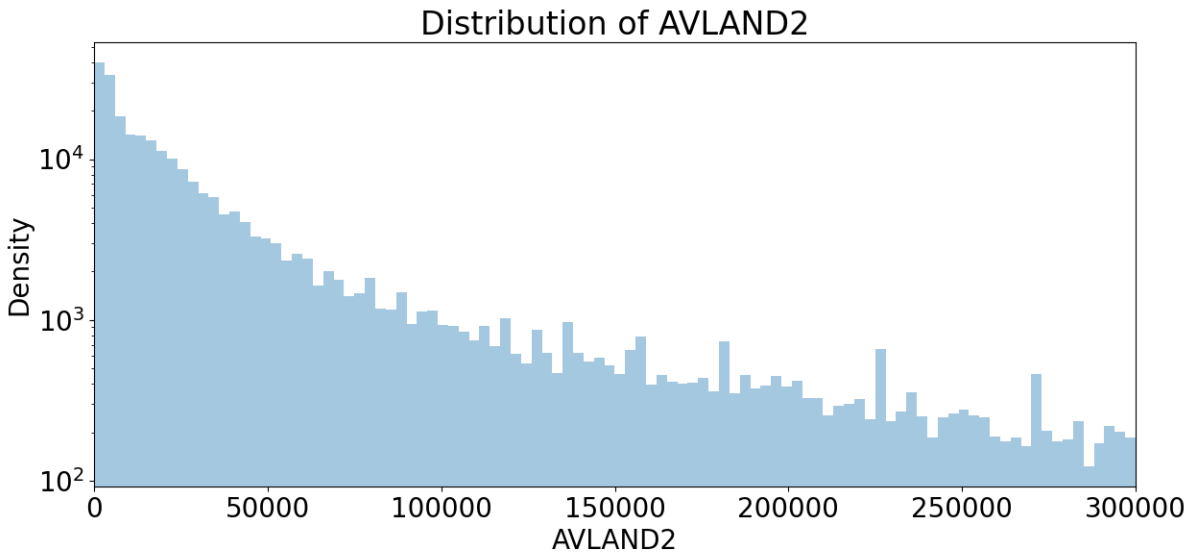




25) Field Name: AVLAND2

Description: The AVLAND2 column contains the assessed land value of the tax lot from the second assessment roll. It has 282,726 records with 26.40% populated and a minimum value of 3 and a maximum value of 2,371,005,000. The mean and standard deviation of the column are 246,235.71 and 6,178,951.64, respectively.

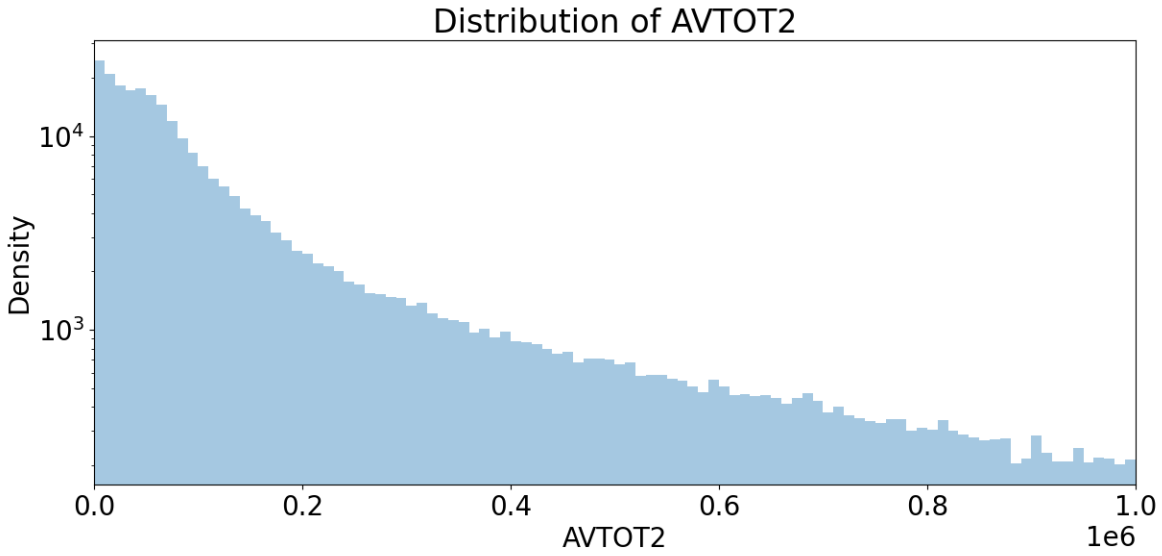
Visualization:



26) Field Name: AVTOT2

Description: The AVTOT2 column refer to the tentative assessed total value for properties with final assessment rolls. It has 282,732 records, with a population percentage of 26.40%. The minimum value is 3, and the maximum value is 4,501,180,002, with a mean value of 713,911.43 and a standard deviation of 11,652,508.34.

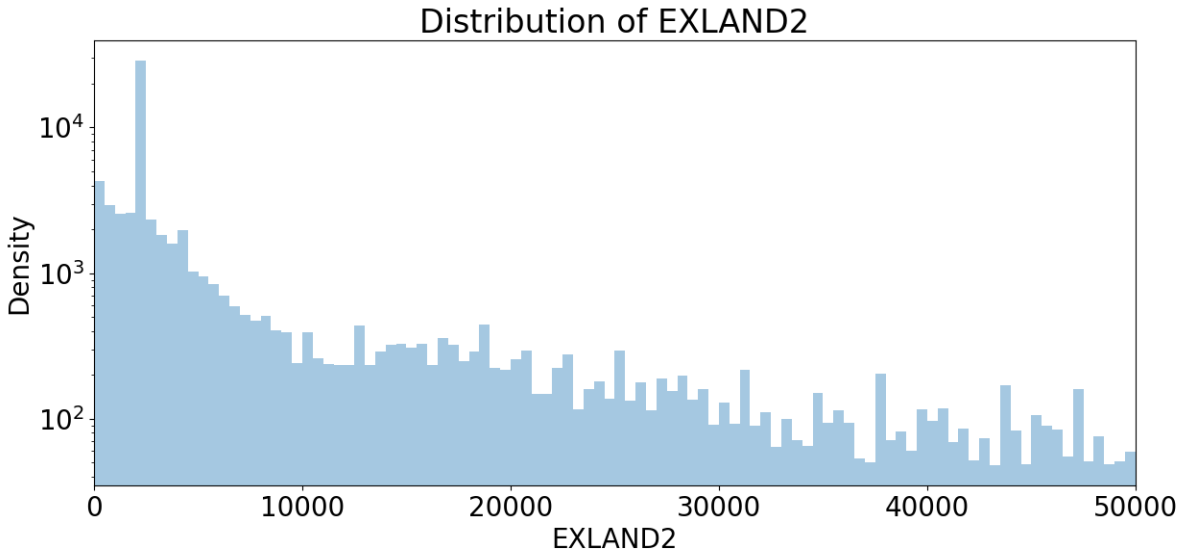
Visualization:



27) Field Name: EXLAND2

Description: The EXLAND2 column refers to the assessed value of the land excluding exempt amounts for properties with multiple units, condominiums, and cooperatives. It has 87,449 records, with a data completeness percentage of 8.17% and a maximum value of 2,371,005,000. The average assessed value is $351,235.68, and the standard deviation is $10,802,150.91.

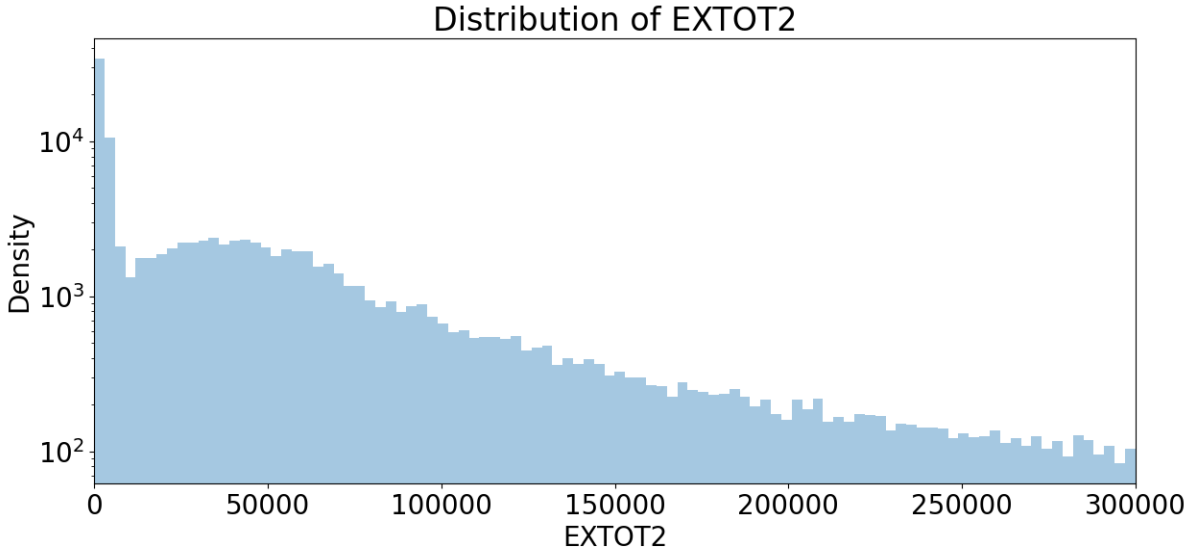
Visualization:



28) Field Name: EXTOT2

Description: The EXTOT2 column contains the assessed value of the property's exterior total area in the second assessment roll. It contains 130,828 non-null values and has a data population of 12.22%. The values are expressed in US dollars and range from 7 to 4,501,180,002.

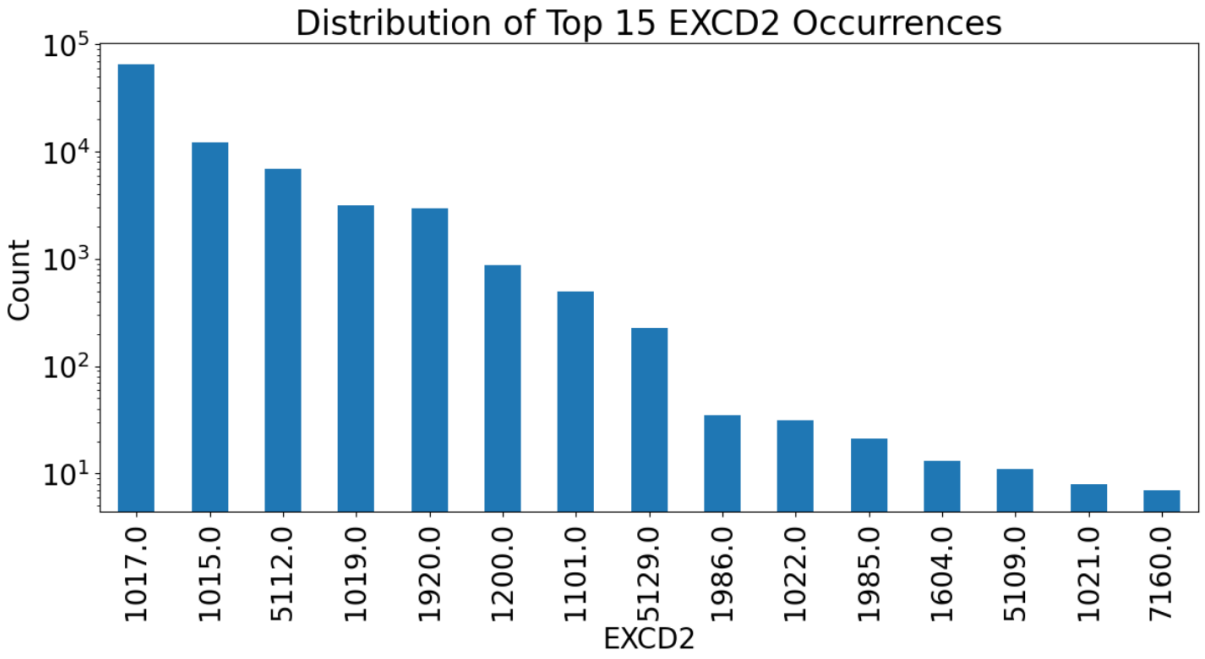
Visualization:

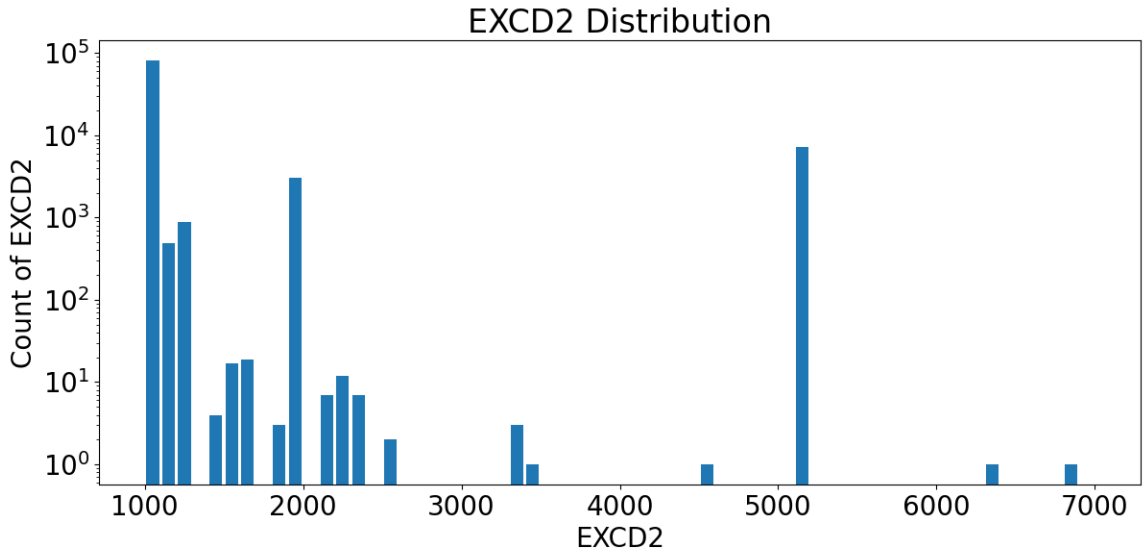


29) Field Name: EXCD2

Description: The EXCD2 column contains the second extended code description, which provides additional information on the tax class of the property. It is a numerical column that ranges from 101 to 7170 and has 166 unique values. It has 226,837 non-null values and 844,157 null values.

Visualization:





30) Field Name: PERIOD

Description: The PERIOD column indicates the fiscal period for which the property's assessed value is being reported. It has only one unique value, which is "FINAL".

31) Field Name: YEAR

Description: The YEAR column contains information about the fiscal year that the property was assessed. This column has 100% populated values and 0 zero values. It has only one unique value, which is "2010/11".

32) Field Name: VALTYPE

Description: The VALTYPE column contains information about the type of value associated with each property record. This column is fully populated and there are no zero values in this column. It has only one unique value, which is "AC-TR" which suggests that all the property records are based on the actual sale transactions and not on assessed values.