Project 2

REPORT ON NY PROPERTY

Duyen Tran

MGTA 463 – Fraud Analytics

Table of Contents

Executive Summary

New York City, with its dynamic and highly valuable real estate market, faces significant challenges related to property fraud. Fraudulent activities in the property sector can take many forms, including underreporting property values, illegal occupancy, and deceptive trxansactions. These activities have far-reaching consequences, impacting the city’s revenue, market stability, legal compliance, public safety, and consumer protection.

Anomaly detection is a technique used in data analysis to identify unusual patterns or observations that do not conform to the expected behavior or norm. These anomalies, also known as outliers, can represent critical incidents, such as technical glitches, fraud, or other significant deviations that require attention.

We undertook an unsupervised learning project aimed at identifying potential fraudulent activities in New York City's property records. Our primary goal was to detect anomalies in property values relative to their characteristics, such as lot size, building size, and assessed values. We engaged closely with the client to understand their concerns and specific requirements, ensuring our approach was tailored to their needs. Using various anomaly detection algorithms, we identified numerous properties with atypical valuations. These findings provide a solid foundation for further investigation and validation. Moving forward, we recommend an iterative refinement process based on client feedback to enhance the accuracy and relevance of the detection model. This process involves adjusting the variables and exclusion criteria to better capture relevant patterns and incorporating domain-specific insights to fine-tune the detection algorithms.

I. Data Description

1. Data overview

The dataset contains property and tax assessment information in New York. This dataset comprises 1,070,994 records and 32 fields, covering comprehensive details on property assessments, exemptions, and other relevant characteristics. It includes various identifiers, measurements, valuations, and classification codes related to the properties:

- **Record:** An identifier or record number for each entry. This is a unique number assigned to each property record, from 1 to 1,070,994.
- **BBLE:** Borough-Block-Lot-Easement identifier
- **BLOCK:** The block number within the borough, is used to identify a specific area or block.
- **LOT:** The lot number within the borough
- **EASEMENT:** A code or indicator related to property easements
- **Owner:** An identifier for the property owner
- **BLDGCL:** Building class, which classifies the type of building.
- **TAXCLASS**: Tax class code.
- **LTFRONT:** Lot frontage, representing the width of the lot at the front, measured in feet.
- **LTDEPTH:** Lot depth, representing the depth of the lot, measured in feet.
- **EXT:** An extension code.
- **STORIES:** Number of stories (floors) in the building, indicating the vertical size of the structure.
- **FULLVAL**: Full market value of the property, representing the estimated market value for taxation purposes.
- **AVLAND / AVLAND2:** Assessed value of the land, representing the value of the land portion of the property for tax assessment purposes.
- **AVTOT / AVTOT2**: Assessed total value, including both land and improvements (buildings, structures) on the property.
- **EXTOT / EXTOT2:** Extension of the total value.
- **EXPAND / EXLAND2:** Represents the exempt portion of the land values.
- **EXCD / EXCD 2:** Represent the exemption code.
- **STADDR:** Street address of the property, providing the physical location address.
- **ZIP:** ZIP code of the property location, indicating the postal code for the area.
- **EXMPTCL:** Exemption class, indicating if the property has any tax exemptions and the type of exemption.
- **BLDFRONT:** Building frontage, representing the width of the building at the front, measured in feet.
- **BLDDEPTH:** Building depth, representing the depth of the building, measured in feet.
- **PERIOD:** A period indicator, serving as a unique code to denote the ‘Final’ period with 1,070,994 counts.
- **VALTYPE:** Valuation type, indicating the method or type of property valuation used.
- **YEAR:** Year of the data or assessment, indicating when the data was recorded or the assessment took place. This dataset covers the timeframe of 2010/11.

2. Summary Tables

Table 1: Numeric Fields Table

Field Name	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Std Dev	Most Common Value
LTFRONT	1,070,994	100.00%	0	0.00	9,999	36.64	74,03	0.00
LTDEPTH	1,070,994	100.00%	0	0.00	9,999	88.96	76.40	100.00
STORIES	1,014,730	94,70%	56,264	1.00	119.00	6.01	8.37	2.00
FULLVAL	1,070,994	100.00%	0	0.00	6,150,000,000.00	874,264.51	11,582,425.58	0.00
AVLAND	1,070,994	100.00%	0	0.00	2,668,500,000.00	85,067.92	4,057,258.16	0.00
AVLAND2	282726	26.40%	788,268	3.00	2,371,005,000.00	246235.72	6,178,951.64	2,408.00
AVTOT	1,070,994	100.00%	0	0.00	4,668,308,947.00	227,238.17	6,877,526.09	0.00
AVTOT2	282,732	26.40%	788,262	3.00	4,501,180,002.00	713,911.44	11,652,508.34	750.00
EXLAND	1,070,994	100.00%	0	0.00	2,668,500,000.00	36,423.89	3,981,573.93	0.00
EXLAND2	87,449	8.20%	983545	1.00	2,371,005,000.00	351,235.68	10,802,150.91	2,090.00
EXTOT	1,070,994	100.00%	0	0.00	4,668,308,947.00	91,186.98	6,508,399.78	0.00
EXTOT2	130,828	12.20%		7.00	4,501,180,002.00	656,768.28	16,072,448.75	2,090.00
BLDFRONT	1,070,994	100.00%	0	0.00	7,575.00	23.04	35.58	0.00
BLDDEPTH	1,070,994	100.00%	0	0.00	9,393.00	39.92	42.71	0.00

Table 2: Categorical Fields Table

Field Name	# Records with Values	% Populated	# Zeros	# Unique Values	Most Common Value
RECORD	1,070,994	100.00%	0	1,070,994	1
BBLE	1,070,994	100.00%	0	10.070,994	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	4.00%	1,066,358	13	E
EXCD	638,488	60.00%	432,506	129	1017
EXCD2	92,948	9.00%	978,046	60	1017
OWNER	1,039,249	97.00%	31,745	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	33.00%	716,689	3	G
STADDR	1,070,318	99.99%	676	839,280	501 SURF AVENUE
ZIP	1,041,104	97.20%	29,890	196	10314
EXMPTCL	15,579	1.50%	1,055,415	14	X1
PERIOD	1,070,994	100.00%	0	1	FINAL
VALTYPE	1,070,994	100.00%	0	1	AC-TR
YEAR	1,070,994	100.00%	0	1	2010/11

3. Distributions

Bellow we show some visualization of important fields of the data:

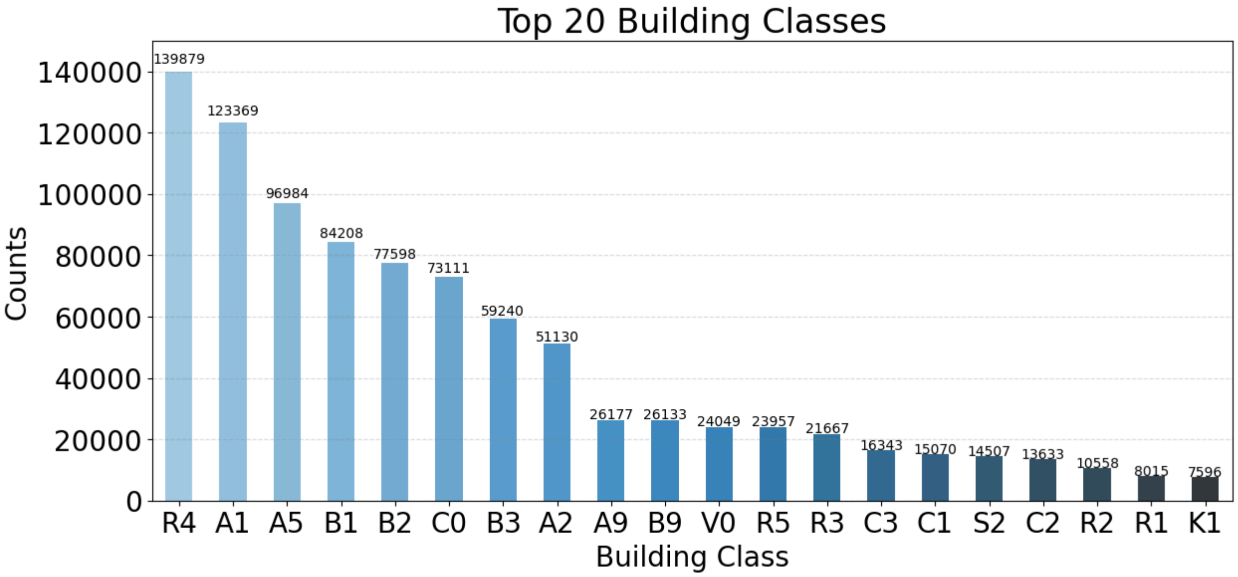


Figure 1: Distribution of top 20 building classes

Figure 1 represents the distribution of the top 20 building classes by the number of property records with the x-axis representing the different building class codes and the y-axis representing the count of property records for each building class. **R4** is the most common building class with **139,879 counts**, followed by **A1** with **123,369 counts**, suggesting they might represent common types of buildings such as residential homes or apartment complexes. There is a clear decrease in the number of property records as we move from the most common building classes (R4, A1) to the less common ones (R1, K1).

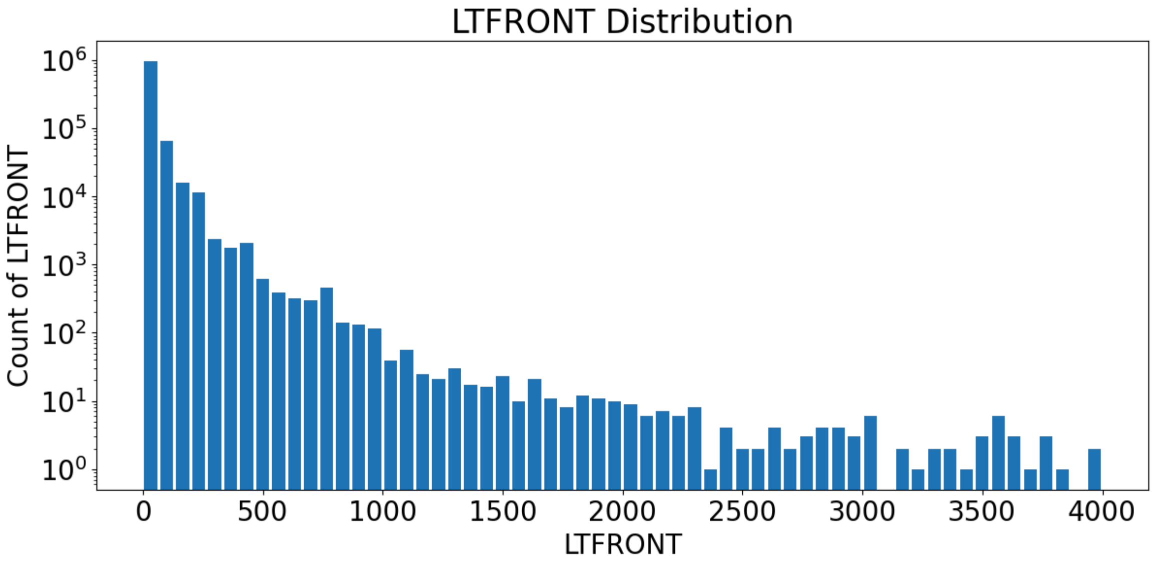


Figure 2: Distribution of LTFRONT

The distribution represents the distribution of LTFRONT values in the dataset with the x-axis representing the LTFRONT values, which are measurements of lot frontage and the y-axis represents the count of occurrences for each LTFRONT value. The majority of properties have small lot frontage. This is evident from the tall bars at the lower end of the LTFRONT scales. As the LTFRONT increases, the frequency of properties with those frontages smaller lots are more common, and larger lots are less frequent. By understanding the distribution of lot frontages, we can inform urban planning and zoning regulations. Areas with predominantly small lot frontages might be residential zones with single-family homes, while larger lot frontages might indicate commercial or industrial areas.

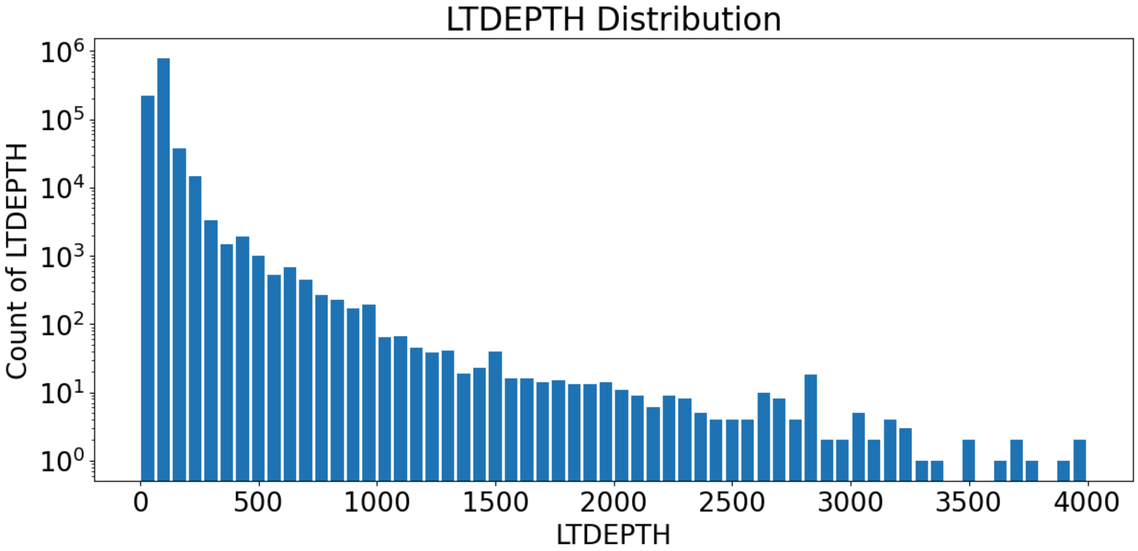


Figure 3: Distribution of LTDEPTH

Figure 3 shows the distribution of LTDEPTH values in the dataset, with the y-axis representing the count of occurrences for the LTDEPTH value, and the x-axis representing the LTDEPTH values. Same as LTFRONT, the majority of LTDEPTH values are concentrated at the lower end of the scale (e.g. 0 to 500 feet), as indicated by the tail bars at the beginning of the histogram. This suggests that many properties have small lot depths. There is a long tail extending to the right, indicating that there are some properties with very large lot depths, though these are relatively rare. As LTDEPTH increases, the frequency of occurrences drops significantly, which is a common pattern in property data where smaller lot dimensions are more prevalent.

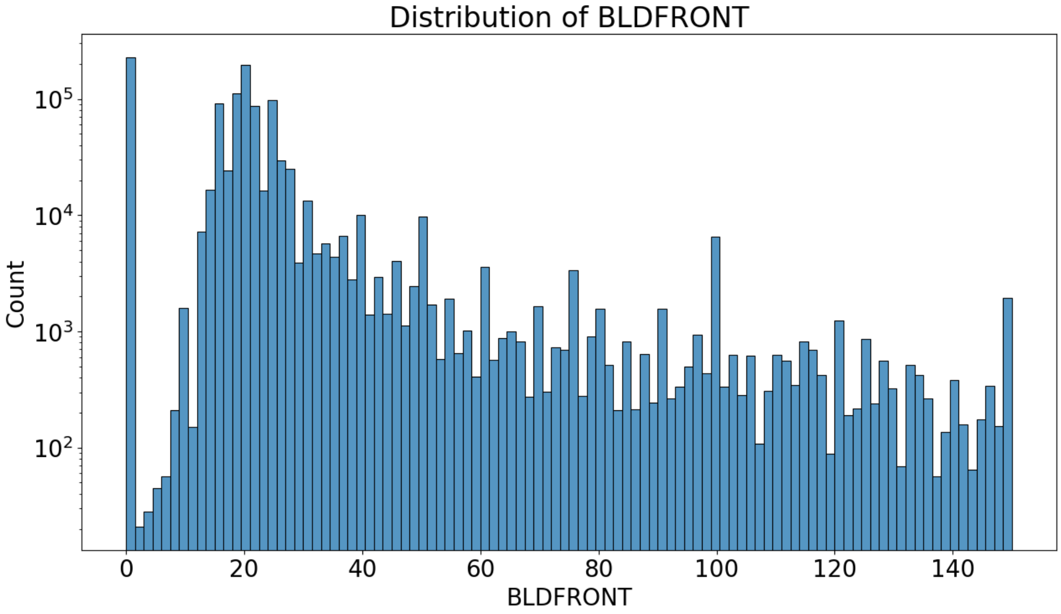


Figure 4: Distribution of BLDFRONT

The histogram displays the distribution of BLDFRONT values, with the x-axis representing the BLDFRONT values and the y-axis representing the count of occurrences on a logarithmic scale. The highest frequency is observed at very low BLDFRONT values, particularly around 0. This suggests that a significant number of buildings have very small or zero frontage. The majority of BLDFRONT values fall within the range of 10 to 50. Within this range, there are several peaks, indicating common frontage dimensions for many buildings. Values above 100 are less frequent but still present. The frequency continues to decrease as the BLDFRONT value increases, showing that large building frontages are less common.

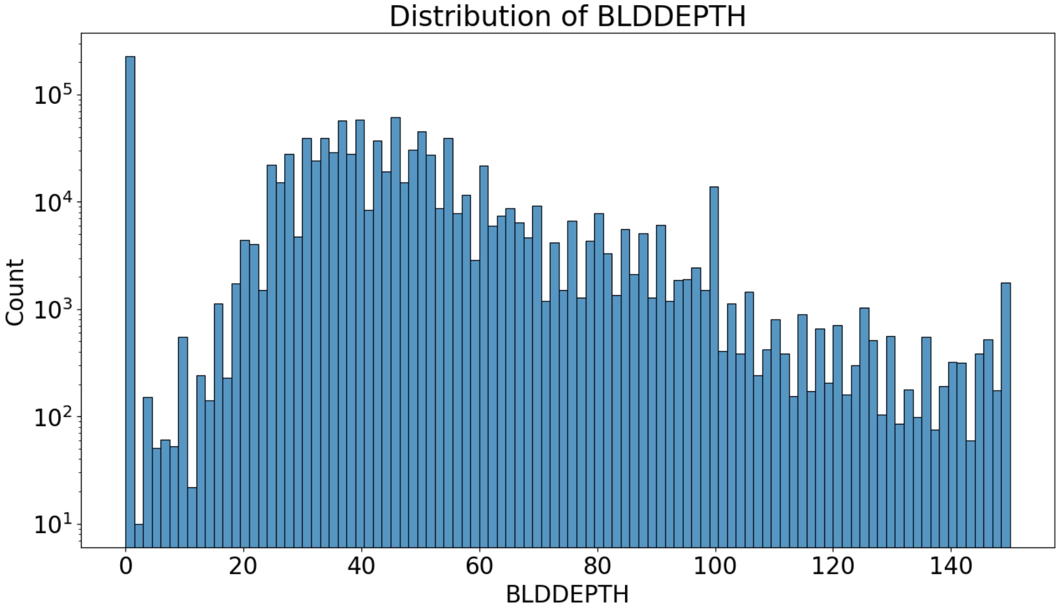


Figure 5: Distribution of BLDDEPTH

Figure 5 displays the distribution of BLDDEPTH values, highlighting the predominance of smaller depths and the decreasing frequency of larger depths. Similar to BLDFRONT, the highest frequency is observed at very low BLDDEPTH values, particularly around 0. This indicates that a significant number of buildings have very small or zero depth, The majority of BLDDEPTH values fall within the range of 20 to 60. This range shows a consistently high frequency of buildings, indicating standard building depths in this range. The distribution is right-skewed, with a long tail extending towards higher BLDDEPTH values. This indicates that while smaller depths are very common, larger depths are progressively less frequent.

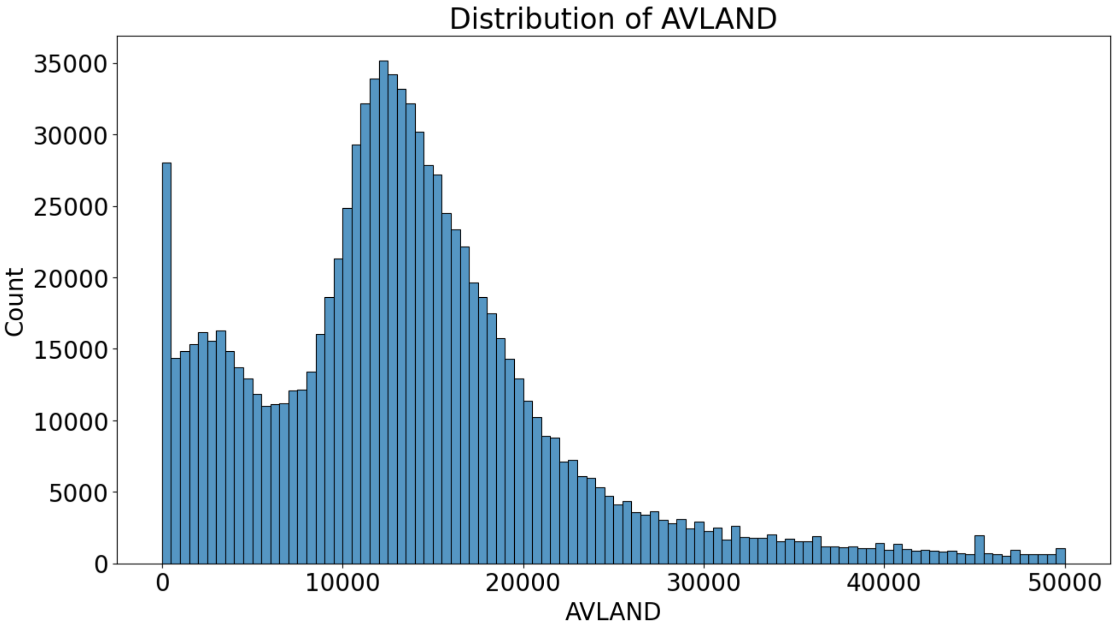


Figure 6: Distribution of AVLAND

The histogram provides a comprehensive view of the distribution of AVLAND values of the data. There is a significant peak at the lower end of the AVLAND values, indicating that most properties have lower assessed land values. The plot peaks around the $10,000 to $15,000 range, indicating a high concentration of properties with land values within this range. There is a notable count at the zero mark, similar to the AVTOT distribution, which could indicate properties that have no assessed land value, possibly due to exemptions or data entry errors. As the assessed land value increases, the frequency of occurrences generally decreases. This is a common pattern in property data where lower-valued land is more prevalent. This information is crucial for understanding the overall property characteristics and can guide further investigation into specific property types, real estate market analysis, and data modeling.

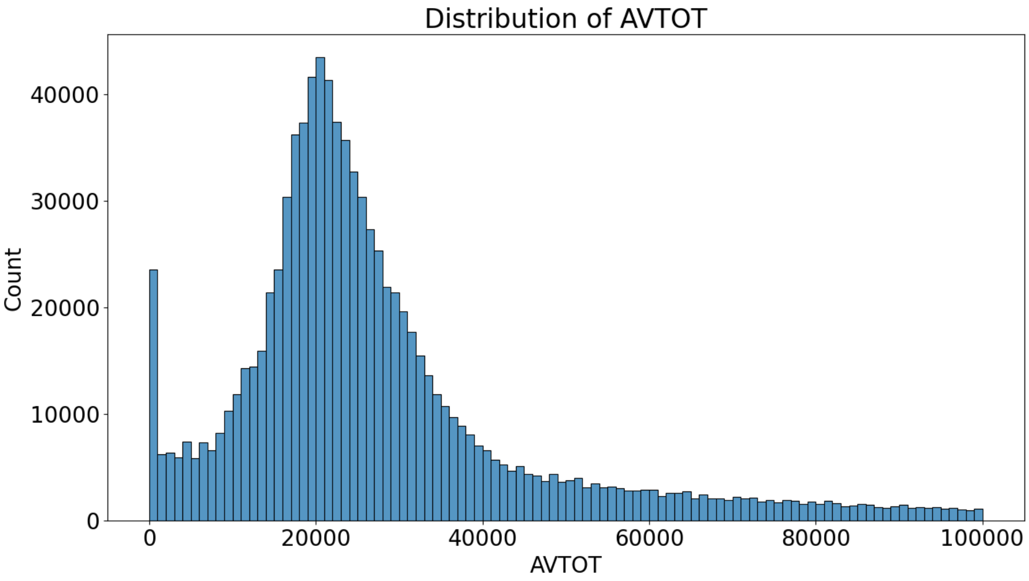


Figure 7: Distribution of AVTOT

Figure 7 shows the distribution of AVTOT of the data. There is a significant peak at the lower end of the AVTOT values, indicating that most properties have lower assessed total values. The high concentration of around $20,000 suggests that many properties fall into the lower to mid-value range. The spike at zero could indicate properties that are exempt from assessment, unoccupied land, or errors in data entry. As the assessed total value increases, the frequency of occurrences generally decreases which shows a long tail extending to the right, indicating that there are some properties with much higher assessed total values, though these are relatively rare. This is a common pattern in property data where lower-valued properties are more prevalent.

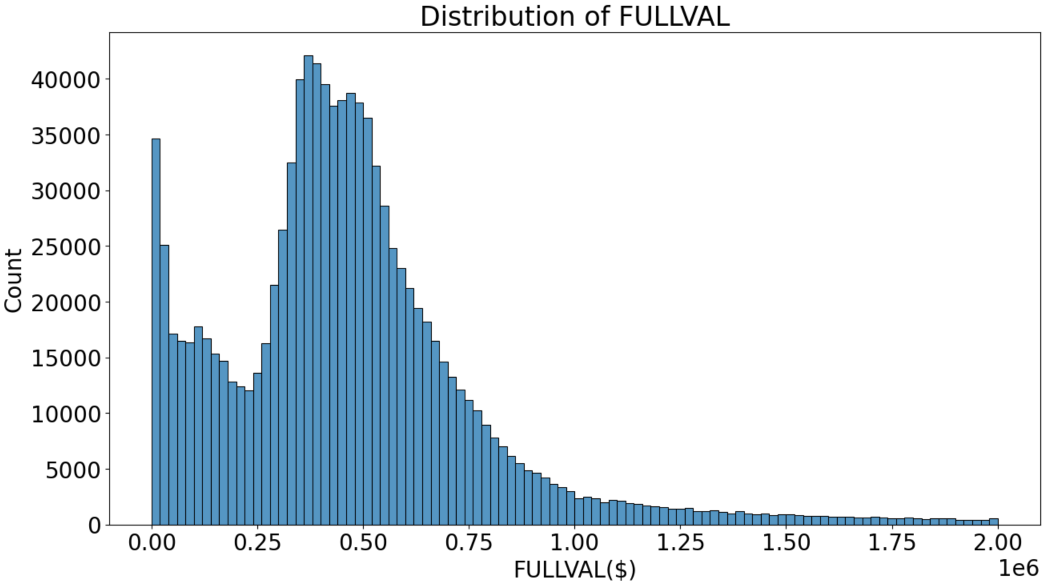


Figure 8: Distribution of FULLVAL

Figure 8 represents the distribution of FULLVAL of the data. The distribution shows two distinct peaks: one around $250,000 and another around $500,000. This suggests two common value ranges for properties, which might be typical for residential homes, smaller commercial buildings, or properties in less expensive areas. There is a significant count of properties with a full market value close to zero, which might indicate tax-exempt properties or data anomalies. As the market value increases, the frequency of occurrences generally decreases. This is a common pattern in property data where lower-valued properties are more prevalent.

II. Data Cleaning

1. Exclusion

Overview:

When conducting fraud detection analysis on property records, it is often beneficial to remove government-owned properties from the dataset. Government-owned properties are typically not subject to the same market forces and valuation practices as privately-owned properties, which can skew the analysis and lead to misleading results. This exclusion is necessary for several reasons that enhance the accuracy and relevance of the analysis.

- **Different Valuation Practices:**
  - Government-owned properties are often valued based on criteria that differ from market-based valuations. Including these properties can distort the analysis and make it challenging to detect anomalies in market-based valuations accurately.
- **Lack of Fraud Incentives:**

  - ○ Government entities generally do not engage in property fraud for financial gain. Excluding these properties allows the analysis to focus on cases with a higher likelihood of fraud, thus making the detection process more efficient and relevant.

- **Data Consistency:**
  - ○ For accurate fraud detection, a homogeneous dataset is crucial. Government-owned properties introduce inconsistencies due to their unique valuation and ownership circumstances. A more homogeneous dataset leads to better model performance and more reliable anomaly detection.
- **Regulatory and Administrative Differences:**
  - ○ Government-owned properties are governed by different regulatory and administrative frameworks, which can introduce additional variables that complicate the analysis. Removing these properties simplifies the analysis by reducing the number of extraneous factors.

By removing government-owned properties, we ensure that our fraud detection analysis is focused, accurate, and relevant to market-based property values. This approach enhances the integrity and effectiveness of detecting potential fraudulent activities in New York City's property market. As per our stakeholder's request, the goal is to focus on private owners and exclude government-owned properties from our analysis. To achieve this, we identified and removed properties with indicators of government ownership.

Decision:

Excluded all properties that might belong to the government or a cemetery, focusing the analysis on private properties.

2. Imputation

Imputation of missing data is a critical step in the analysis of property records for detecting fraudulent activities. The primary motivations for imputing variables are to preserve data integrity, enhance model performance, improve analysis quality, maintain robustness, and support regulatory compliance. Preserving data integrity ensures that the dataset is complete and accurate, reducing bias from missing values and leading to more reliable findings. Enhancing model performance is crucial as machine learning models require complete data to function effectively.

By imputing missing values, we maintain consistency and the predictive power of these models. Improving analysis quality involves enabling a comprehensive analysis by including all records, which maximizes data utilization and provides a fuller picture of the property market.

Maintaining robustness means making the analysis resilient to missing data, reducing the likelihood of significant impacts from incomplete records, and minimizing errors. Supporting regulatory compliance ensures that datasets meet regulatory requirements for completeness, and provides a transparent approach to handling missing data, which is important for audits and regulatory scrutiny.

Imputation for each field:

- **Zip**

Initial condition: There are 19,823 properties with missing **“ZIP”**

Strategy	**# Filled in**	**#Remaining Missing**
Use the STADDR field to fill in the most appropriate ZIP for that STADDR	2,906	16,917
Concatenate the "BORO" and "STADDR" into a new column for each street address and borough combination then use the new column to fill in the most appropriate Zip for that new field	3,187	16,636
Use the BLOCK field to fill in the most appropriate ZIP for that BLOCK	15,401	1,235
Fill in missing values based on the previous row	1,235	0

- **FULLVAL**

Initial condition: There are 9,905 missing **FULLVAL** value

Strategy	**# Filled in**	**# Remaining Missing**
Group all records by the combination of "TAXCLASS", "BORO", and "BLDGCL", calculate the average of FULLVAL for each group, and replace the missing value in the group with the average of the group	2,648	7,257
Group all records by the combination of TAXCLASS and BORO, calculate the average of FULLVAL for each group, and replace the missing value in the group with the average of the group	6,878	379
Group all records by TAXCLASS, calculate the average of FULLVAL for each group, and replace the missing value in the group with the average of the group	379	0

- **AVLAND**

Initial condition: There are 9,907 missing values of AVLAND

Strategy	**# Filled in**	**# Remaining Missing**

Group all records by the combination of “TAXCLASS”, “BORO”, and “BLDGCL”, calculate the average of AVLAND for each group, and replace the missing value in the group with the average of the group.	2,650	7,257
Group all records by the combination of TAXCLASS and BORO, calculate the average of AVLAND for each group, and replace the missing value in the group with the average of the group	6,878	379
Group all records by TAXCLASS, calculate the average of AVLAND for each group, and replace the missing value in the group with the average of the group	379	0

- **AVTOT**

Initial condition: There are 9,905 missing values of AVTOT

Strategy	**# Filled in**	**# Remaining Missing**
Group all records by the combination of “TAXCLASS”, “BORO”, and “BLDGCL”, calculate the average of AVTOT for each group, and replace the missing value in the group with the average of the group	2,648	7,257
Group all records by the combination of TAXCLASS and BORO, calculate the average of AVTOT for each group, and replace the missing value in the group with the average of the group.	6,878	379
Group all records by TAXCLASS, calculate the average of AVTOT for each group, and replace the missing value in the group with the average of the group	379	0

- **STORIES**

Initial condition: There are 41,027 missing STORIES values

Strategy	**# Filled in**	**# Remaining Missing**
Group all records by the combination of BORO and BLDGCL, and find the most appropriate STORIES for that combination	4,054	36,976

Group all records by TAXCLASS, calculate the average of STORIES for each group, and replace the missing value in the group with the average of the group. (It’s noticeable that the stories will float with heavy decimals after taking the average of the group, so we convert all these data to integers)	36,976	0

- **Impute for LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT**

These fields do not have NAs but we think 0 and 1 are invalid values for these fields so we just need to replace 0s and 1s as NaN

- **LTFRONT**

Initial conditions: There are 160,420 missing values

Strategy	# Filled in	# Remaining Missing
Group all records by the combination of TAXCLASS and BORO, calculate the average of LTFRONT for each group, and replace the missing value in the group with the average of the group	160,418	2
Group all records by the TAXCLASS, calculate the average of LTFRONT for each group, and replace the missing value in the group with the average of the group	2	0

- **LTDEPTH**

Initial conditions: There are 160,017 missing values

Strategy	# Filled in	# Remaining Missing
Group all records by the combination of TAXCLASS and BORO, calculate the average of LTDEPTH for each group, and replace the missing value in the group with the average of the group	160,017	2
Group all records by the TAXCLASS, calculate the average of LTDEPTH for each group, and replace the missing value in the group with the average of the group	2	0

- **BLDDEPTH**

Initial conditions: There are 205,209 missing values.

Strategy	# Filled in	# Remaining Missing
Group all records by the combination of “TAXCLASS”, “BORO”, and “BLDGCL”, calculate the average of BLDDEPTH for each group, and replace the missing value in the group with the average of the group.	187,097	18,112
Group all records by the combination of TAXCLASS and BORO, calculate the average of BLDDEPTH for each group, and replace the missing value in the group with the average of the group.	2,755	15,357
Group all records by TAXCLASS, calculate the average of BLDDEPTH for each group, and replace the missing value in the group with the average of the group	15,357	0

- **BLDFRONT**

Initial conditions: There are 205,223 missing values.

Strategy	# Filled in	# Remaining Missing
Group all records by the combination of “TAXCLASS”, “BORO”, and “BLDGCL”, calculate the average of BLDFRONT for each group, and replace the missing value in the group with the average of the group.	189,917	15,306
Group all records by the combination of TAXCLASS and BORO, calculate the average of BLDFRONT for each group, and replace the missing value in the group with the average of the group.	2,784	12,522
Group all records by TAXCLASS, calculate the average of BLDFRONT for each group, and replace the missing value in the group with the average of the group	12,522	0

III. Variable Creation

1. Type of Anomalies

In the context of identifying fraudulent properties in New York City, we are looking for anomalies that significantly deviate from typical patterns in property records. These anomalies may indicate potential fraudulent activities such as tax evasion, money laundering, or illegal occupancy. By creating specific variables during the analysis, we can detect these unusual patterns and outliers.

- **Lot Area Anomalies -** Inconsistent Lot Size Reporting:

Description: Properties where the reported lot size significantly deviates from historical records or neighboring properties.

Motivation: To detect misreporting or changes in property boundaries without proper documentation.

- **Building Area Anomalies -** Unusual Building-to-Lot Size Ratio:

Description: Properties with a building area that is unusually large or small compared to the lot size.

Motivation: To identify properties with potential unapproved extensions or illegal constructions.

- **Actual Assessment Anomalies -** Discrepancies Between Market Value and Assessed Value:

Description: Properties where the assessed value significantly deviates from the market value, either too high or too low.

*Motivation***:** To uncover potential errors in assessment or deliberate undervaluation/overvaluation for tax purposes.

- **Building Class Anomalies -** Inconsistent Building Class Records:

Description: Properties with building class records that do not match their actual use or structure.

Motivation: To detect misclassifications that could lead to improper taxation or zoning violations.

- **Number of Stories Anomalies -** Mismatch Between Reported and Actual Number of Stories:

Description: Properties where the reported number of stories does not match the physical structure.

Motivation: To identify illegal additions or misreporting of building structure.

- **Lot Shape and Usability Anomalies** - Unusual Lot Shapes Affecting Usability:

Description: Properties with irregular lot shapes that significantly impact usability and value but are not reflected in the assessment.

Motivation: To ensure assessments accurately reflect the functional value of the property.

- **Unusual Property Improvements** - Discrepancies in Reported Property Improvements:

Description: Properties with significant improvements or renovations that are not reflected in the assessment records.

Motivation: To detect unreported improvements that should increase property value and tax assessments.

- **Building Class Violations -** Properties Not Complying with Building Class Regulations:

Description: Properties that are being used or have structures that do not comply with current building class regulations.

Motivation: To identify and rectify building class violations that could pose legal and safety risks.

2. Variables Table

To effectively detect these anomalies, we need to select and analyze variables that capture key aspects of property records. These variables should help in identifying unusual patterns and deviations from the norm. The motivation for choosing specific variables is based on their relevance to detecting the types of anomalies listed above.

There are 3 size variables created:

1. **Total Area of the lot**: Calculated as the product of lot front and lot depth
2. **Total Area of the building's footprint**: Calculated as the product of the building front and building depth
3. **Total Volume of the building**: Calculated as the product of building size and number of stories

From these 3 size variables, we can create the variable list in the table below:

Description	**# Variables Created**	**# Cumulate Variables**	**Why this variable**
Value per unit of lot area/building area/building volume The highest value of the ratio of the property's full market and value to the lot size/building size/building volume and its inverse, respectively	3	3	To measure the market value density per unit of lot area/building area/building volume, which can be useful in evaluating how much market value is derived from the lot area, the building itself, and the property's three-dimensional space
Assessed land value per unit of lot area/building area/building volume The highest value of the ratio of the assessed land value to the lot size/building size/building volume and its inverse, respectively	3	6	**Assessed land value per unit of lot area** To determine the assessed land value density per unit of lot area, which helps to compare the land valuation across different properties and regions **Assessed land value per unit of building area** To analyze the relationship between assessed land value and the building size, which can indicate how much land value is attributed relative to the building footprint. **Assessed land value per unit of building volume** To examine the assessed land value per unit of building volume, providing a three-dimensional perspective of land valuation.
Assessed total value per unit of lot area/building area/building volume The highest value of the ratio of the assessed total value to the lot size/building size/building volume and its inverse, respectively	3	9	**Assessed total value per unit of lot area** To measure the overall assessed value density per unit of lot area. This is important for understanding the total value assessment efficiency relative to the lot size. **Assessed total value per unit of building area** To evaluate the total assessed value density per unit of building area. This helps in comparing the overall

			valuation efficiency of different properties. **Assessed total value per unit of building volume** To analyze the total assessed value per unit of building volume, providing a comprehensive view of the property's value assessment.
Standardized values Standardized values of Value per unit of lot area/building area/building volume by the same Zipcode	3	12	Standardizing these ratios by the mean values within the same ZIP code helps in normalizing the data to local market conditions. This allows for better comparison across properties within the same geographic area, identifying outliers, and understanding relative performance within each ZIP code.
Standardized values Standardized values of Assessed land value per unit of lot area/building area/building volume by the same Zipcode	3	15	
Standardized values Standardized values of Assessed total value per unit of lot area/building area/building volume by the same Zipcode	3	18	
Standardized values Standardized values of Assessed land value per unit of lot area/building area/building volume by the same Tax Class	3	21	Standardizing these ratios by the mean values within the same TAXCLASS helps in normalizing the data to similar types of properties. TAXCLASS typically groups properties by their usage type (e.g., residential, commercial), so this standardization allows for comparison across similar property types, identifying outliers and understanding relative performance within each class.
Standardized values Standardized values of Value per unit of lot area/building area/building volume by the same class		21	
Standardized values Standardized values of Assessed total value per unit of lot area/building area/building volume by the same Tax Class	3	24	
Measure of the property's full market value relative to its assessed values Standardized and adjusted to highlight low outliers by using the inverse when the ratio is less than 1.	1	25	This helps in focusing on properties that might be undervalued or have other unique characteristics affecting their market assessment.

Building value density The ratio of the building's footprint area to the lot size	1	26	To evaluate how much value is attributed to the building itself per unit area. High building value density can signify high-quality construction or desirable building features.
Lot to Building Area Ratio The ratio of the Total lot area to the Total building area	1	27	To understand the value assigned to the land per unit area. High land value density can indicate prime real estate locations or areas with high development potential.
High market value A flag indicating if the full market value is above the 90th percentile.	1	28	To easily identify properties with exceptionally high market value. These high-value properties might require special attention or separate analysis due to their significant market impact.
High building density indicator A flag indicating if the building size to lot size ratio is above 0.70.	1	29	To highlight properties where the building footprint occupies a significant portion of the lot. High building density can affect various factors such as property taxes, zoning considerations, and potential for redevelopment.
Lot building interaction The product of lot size and building size.	1	30	To capture the combined effect of lot and building sizes, which might be relevant for understanding the overall scale of the property. Interaction terms can reveal relationships that aren't apparent when considering each variable separately.
Market-assessed value interaction The product of full market value and assessed total value.	1	31	To explore the relationship between the market value and the assessed value of the property. This interaction term can help identify properties where the market perception of value significantly differs from the assessed value, indicating potential mispricing or assessment errors.

IV. Dimensionality reduction

Principal Component Analysis (PCA) is a powerful statistical technique used in data analysis and machine learning to reduce the dimensionality of a dataset. The primary goal of PCA is to transform a large set of variables into a smaller set that still contains most of the information in the large set. This reduction helps in simplifying the dataset, making it easier to visualize and analyze, and can improve the performance of machine learning models by reducing noise and computational complexity.

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that new variables are uncorrelated and most of the information within the initial variables is squeezed and compressed into the first components. So the idea is that 20-dimensional data gives us 20 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second component and so on, until having some thing like shown in the plot below:

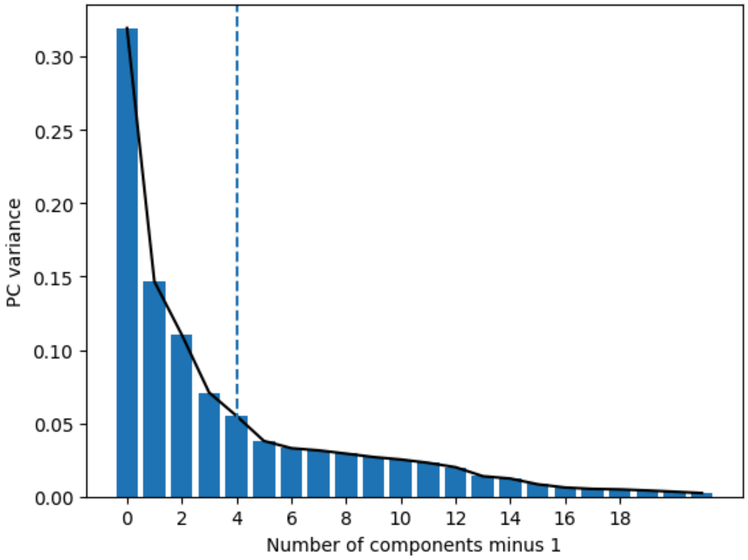


Figure 9: Percentage of Variance of each by PC

Organizing information in principal component this way will allow us to reduce dimensionality without losing much information, and this by discarding the components with low information and considering the remaining components as our new variables. An important thing to realize is that the principal components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables.

1. Dimensionality reduction process

a. Initial Standardization

The aim of this step is to standardize the range of the continuous variables initial variables so that each one of them contributes equally to the analysis.

More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges (for example, a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1), which will lead to biased results. So, transforming the data to comparable scales can prevent this problem.

Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation}$$

Once the standardization is done, all the variables will be transformed to the same scale.

b. PCA

- Covariance Matrix Computaion

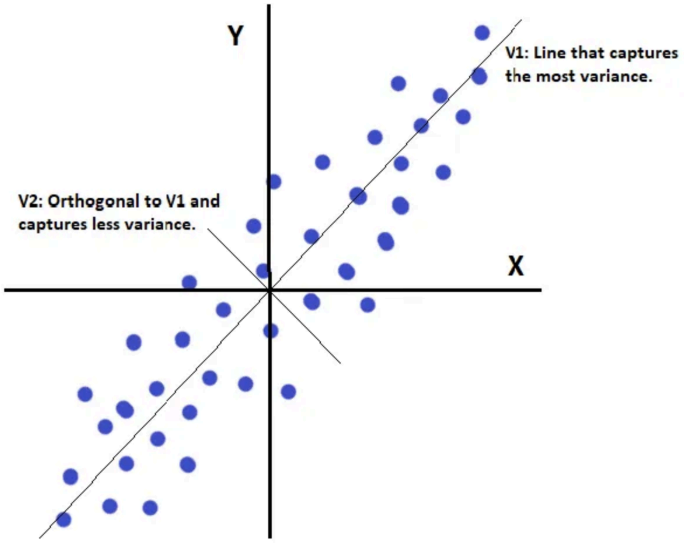
The covariance matrix is a square matrix that shows the covariance (a measure of how much two random variables change together) between each pair of variables in the dataset. The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So in order to identify these correlations, we compute the covariance matrix.

- **Eigen Decomposition**

Eigen Decomposition involves breaking down a square matrix into its eigenvalues and eigenvectors. This decomposition helps in understanding the structure of the matrix and in performing various transformations. Here is what they mean:

Eigenvalue: An eigenvalue (λ) represents a scalar that indicates how much variance is explained by the corresponding eigenvector. In PCA, eigenvalues quantify the importance of each principal component. They are always non-negative, and the eigenvalue corresponding to a principal component measures the proportion of the total variance in the data explained by that component.

Eigenvector: An eigenvector (v) is a vector associated with an eigenvalue. In PCA, eigenvectors represent the directions along which the data varies the most. Each eigenvector points in a specific direction in the feature space and corresponds to a principal component. Eigenvectors are typically normalized, meaning their length is 1.


Y
X
V1: Line that captures the most variance.
V2: Orthogonal to V1 and captures less variance.

We calculated the eigenvalues and eigenvectors of the covariance matrix. Eigenvectors represent the directions of the principal components, and eigenvalues represent the magnitude of variance in these directions. Eigen decomposition breaks down the covariance matrix into its eigenvalues and eigenvectors. The eigenvalues indicate the amount of variance carried by each eigenvector, and the eigenvectors determine the direction of the new feature space.

- **Performing PCA**

High-dimensional data can be complex and noisy, making it difficult to identify patterns. PCA transforms the original variables into a smaller set of uncorrelated variables (principal components) that capture the most significant variance in the data. This simplification helps in detecting anomalies by focusing on the main sources of variation.

The goal of this step is initialzied PCA to retain the top 5 principal components. Selecting the top 5 principal components is a strategic decision aimed at achieving a balance between reducing the dimensionality of the data and retaining the majority of its variance. The selection is made by sorting the eigenvalues in descending order and choosing the corresponding eigenvectors. The number of components is determined by the amount of variance we want to retain. In this case, the top 5 components were selected because they capture around 80% of the total variance, which is considered substantial for the analysis.

c. Second Standardization of Principal Components

After selecting the principal components, we standardized them again to ensure they contribute equally to subsequent analysis steps. This step makes all retained principal components equally important and helps in calculating distances, such as Minkowski distance, similar to Mahalanobis distance. This is particularly useful when only a small number of principal components are retained.

V. Anomaly Detection Algorithms

Anomaly detection algorithms for detecting unusual properties are computational methods used to identify properties with values or characteristics that deviate significantly from what is expected or typical. These algorithms analyze various property attributes, such as size, price per square foot, location, and property type, to find outliers that may indicate potential errors, fraud, or other significant discrepancies in property data.

1. Distance-based methods

Minkowski distance is a distance measured between two points in N-dimensional space. It is basically a generalization of the Euclidean distance and the Manhattan distance. It is widely used in the field of Machine learning, especially in the concept to find the optimal correlation or classification of data.

To identify potential fraudulent activities in property records, we implemented a distance-based anomaly detection method. This method relies on measuring the distance of each record from typical values, identifying those that deviate significantly as anomalies. We calculated a composite score for each record based on the Minkowski distance, specifically the Euclidean distance. This score measures the overall deviation of a record from typical values.

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

The method capture the overall deviation of a record from the mean in a multi-dimensional space defined by z-scores of principal components. This approach ensures that all deviations, regardless of direction, are considered, providing a comprehensive measure of how unusual each record is compared to the typical data pattern. The flexibility of the parameter p allows for adjusting the sensitivity of the distance calculation, ensuring effective identification of both small and large anomalies. Reasonable choices for the parameter p are 1, 2, or 3, depending on the desired sensitivity to deviations.

Records were ranked based on their distance scores. Higher scores indicate a higher likelihood of being anomalies. The top-ranked records with the highest scores were selected for further investigation, providing a targeted approach to identifying potential anomalies in property valuations. Furthermore, z-scaling the principal components ensures that each variable contributes equally to the analysis, preventing bias from variables with larger scales and leading to a balanced and fair assessment of anomalies.

This distance-based method effectively identifies records with unusual patterns, prioritizing them for further review and investigation. By measuring how far each record deviates from typical values, we can detect anomalies that may indicate fraudulent activities.

2. Autoencoder Reproduction Error

An autoencoder is a type of artificial neural network used to learn efficient codings of unlabeled data. It is an unsupervised learning algorithm that applies backpropagation, setting the target values to be equal to the inputs. Autoencoders are used for various tasks, such as dimensionality reduction, feature learning, and anomaly detection.

To identify potential fraudulent activities in property records, we implemented an autoencoder-based anomaly detection method. This method leverages the reconstruction capabilities of an autoencoder to flag records that deviate significantly from normal patterns.

- **Model Configuration:**
  - ○ We used a Multi-Layer Perceptron (MLP) regressor as an autoencoder. The model had a single hidden layer with 3 nodes and used the logistic activation function.
  - ○ The model was trained for 50 iterations, which was sufficient to detect unusual records without overfitting.
- **Training the Autoencoder:** The autoencoder was trained using the PCA-transformed, z-scaled data. The objective was to minimize the reconstruction error, ensuring the autoencoder learned the typical patterns in the data.
- **Reconstruction and Error Calculation:**
  - ○ After training, all property records were passed through the autoencoder to obtain reconstructed versions.
  - ○ The reconstruction error for each record was calculated by comparing the original data with the reconstructed data. This error indicates how well the autoencoder could reproduce the record based on the patterns it learned.
- **Anomaly Scoring:**
  - ○ The reconstruction error served as the anomaly score. Higher reconstruction errors indicated records that deviated from the learned patterns, suggesting potential anomalies.
  - ○ The anomaly score was calculated as the Euclidean distance of the reconstruction error vector:

$$\text{Anomaly Score} = \left(\sum_{i=1}^{n} |\text{error}_i|^2 \right)^{\frac{1}{2}}$$

In short, the autoencoder method employs a type of neural network designed to learn the internal patterns and structures within the dataset. By attempting to reconstruct the input data, the autoencoder generates a reconstruction error, which serves as an anomaly score. Higher reconstruction errors indicate records that deviate significantly from the learned patterns, highlighting potential anomalies. The ability of autoencoders to learn data patterns and their sensitivity to deviations makes them effective in identifying unusual records. Moreover, their robustness and adaptability allow them to be fine-tuned for different types of data and anomalies, enhancing their utility in detecting potential fraudulent activities in property records

VI. Results

To create a comprehensive and robust anomaly detection system, we combined the two independent anomaly scores derived from the distance-based method and the autoencoder reconstruction error method. This combination leverages the strengths of both approaches, providing a more reliable detection capability.

The final anomaly score for each record was calculated as the weighted average of the two ranked scores:

$$\text{Final Score} = (\textbf{weight} \times \textbf{score1 rank}) + ((1 - \textbf{weight}) \times \textbf{score2 rank})$$

Combining these two methods enhances the robustness of the fraud detection system. The distance-based method excels in identifying outliers based on statistical deviations, while the autoencoder method is effective in detecting complex, non-linear anomalies. Together, they provide comprehensive coverage, ensuring that different types of anomalies, whether based on individual variable deviations or complex pattern deviations, are effectively detected. This combined approach improves the overall accuracy and reliability of the fraud detection system by leveraging the strengths of both methods. The integration of scores from both methods offers a more comprehensive assessment of anomalies, leading to a more reliable identification of potential potential anomalies in the property records and providing a solid foundation for further investigation.

The top-ranked records, based on the final score, were flagged for several notable cases:

1. **Record: 1049911**

Owner: KENILWORTH HOLDINGS L

Address: 1927 Arthur Kill Rd, Staten Island, NY 10312

Building Class	B9	**LTDEPTH**	193	**FULLVAL**	560,000	**r4_taxclass**	105
Tax Class	1	**BLDFRONT**	13	**AVLAND**	25	**r5_taxclass**	51
Stores	3	**BLDDEPTH**	55	**AVTOT**	20,356	**r6_taxclass**	70
LTFRONT	23						

With a B9 building class, this property is categorized as a miscellaneous two-family dwelling. The provided data and image reveal notable characteristics that is relevant for fraud detection algorithms:

The assessed land value (AVLAND) appears unusually low, leading to significantly high standardized ratios (r4_taxclass, r5_taxclass, r6_taxclass). These ratios compare the AVLAND to the lot size, building size, and building volume, respectively, and are all substantially above the average for properties within the same TAXCLASS. This suggests that the property has exceptionally high assessed land value densities relative to its dimensions.

Given these characteristics, the property stands out as an outlier within its TAXCLASS. This could indicate unique valuation characteristics, potentially due to special location advantages, distinctive property features, or possible assessment discrepancies.

In summary, the unusually high assessed land value densities relative to the lot size, building size, and building volume warrant further investigation.

2. Record: 658933

Owner: WAN CHIU CHEUNG

Address: 54-76 83 St Unit 3, Elmhurst, NY 11373

Building Class	C0	**FULLVAL**	776000	**r2**	26	**r2_taxclass**	564
Tax Class	1	**AVLAND**	26940	**r3**	11	**r3_taxclass**	607
Stores	3	**AVTOT**	46560	**r5**	34	**r5_taxclass**	550
LTFRONT	25	**BLDFRONT**	2500	**r8**	43	**r8_taxclass**	510
LTDEPTH	100	**BLDDEPTH**	5600	**r9**	25	**r9_taxclass**	607

This property exhibits several unusual characteristics that warrant further investigation:

- The building front (2500) and building depth (5600) are exceptionally high compared to the lot size. These dimensions indicate an unusually large building relative to the lot area, suggesting potential data inaccuracies or extraordinary property features.
- The AVLAND of 26,940 is very low relative to both the lot and building areas, indicating potential undervaluation or unique valuation factors.
- High standardized values for ratios such as r2_taxclass, r3_taxclass. R5_taxclass, r8_taxclass, r9_taxclass indicate that this property is an outlier within its TAXCLASS. These ratios suggest unique valuation characteristics or possible inconsistencies in the property's assessed values compared to similar properties within the same TAXCLASS.

Given these anomalies, the property warrants further investigation to understand the reaåsons behind these unusual characteristics, whether they stem from unique property features, location advantages, or potential assessment errors.

3. Record: 333412

Owner: SPOONER ALSTON

Address: 37 Monroe St, Brooklyn, NY 11238

Building Class	C5	**FULLVAL**	9,060	**r2**	26	**r2_taxclass**	341
Tax Class	2B	**AVLAND**	3,874	**r3**	11	**r3_taxclass**	324
Stores	3	**AVTOT**	4,077	**r2_zip5**	41		
LTFRONT	17	**BLDFRONT**	4,017	**r3_zip5**	25		
LTDEPTH	85	**BLDDEPTH**	42				

This property exhibits several unusual characteristics that warrant further investigation:

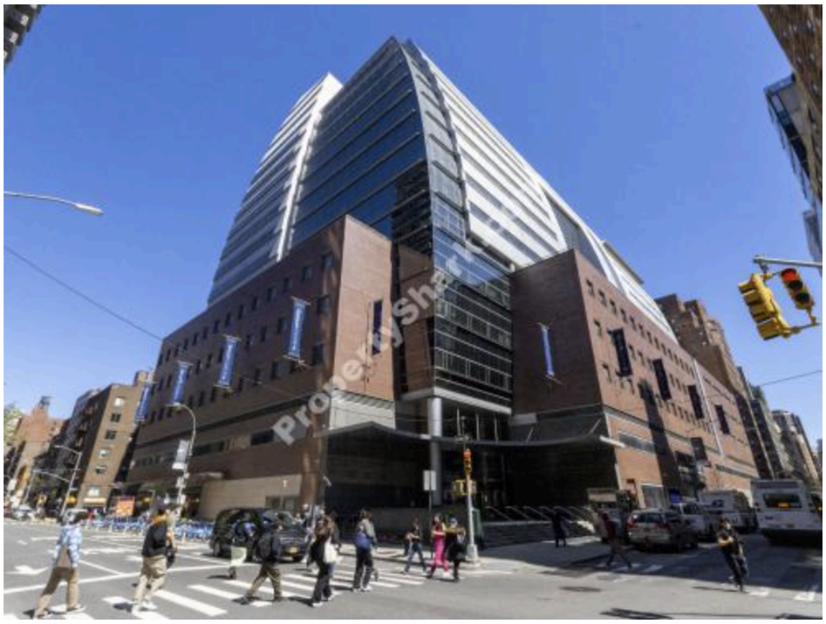
- This building class refers to converted dwellings or rooming houses, which are buildings originally designed for one purpose but later converted into multi-unit residential properties. This conversion could lead to incorrect or outdated information about building dimensions, especially the building front (BLDFRONT). The building front (4,017) is disproportionately high compared to the building depth, lot front, and lot depth, which may indicate potential data inaccuracies or unique property features.
- The high standardized values for r2 and r3 within its TAXCLASS suggest that the property is a significant outlier, indicating unique valuation characteristics or potential inconsistencies in the assessment.

4. **Record number: 47984**

Owner: BERKOWITZ, ULWT LOUIS

Address: 49 Lexington Avenue, New York, NY 10010

Building Class	W6	**FULLVAL**	138,000,000	**r2**	26	**r2_taxclass**	123
Tax Class	4	**AVLAND**	11,025,000	**r3**	11	**r3_taxclass**	92
Stories	2	**AVTOT**	62,100,000	**r5**	34	**r5_taxclass**	237
LTFRONT	39	**BLDFRONT**	39	**r8**	43	**r8_taxclass**	325
LTDEPTH	50	**BLDDEPTH**	50	**r9**	25	**r9_taxclass**	196



- This property has multiple stories (14) instead of having 2 stories
- Furthermore, the ratios (r2, r3, r5, r8, r9) and their standardized values indicate the property is significantly above average within its TAXCLASS, making it a notable outlier. The initial incorrect reporting of 2 stories would drastically alter these calculations, underscoring the importance of accurate data.
- The most common factor in calculating the variables listed above that differentiates them from the rest of the variables used is the value for the number of stories in the building. This strongly suggests that the value for the number of stories is driving most of the variation in the variables

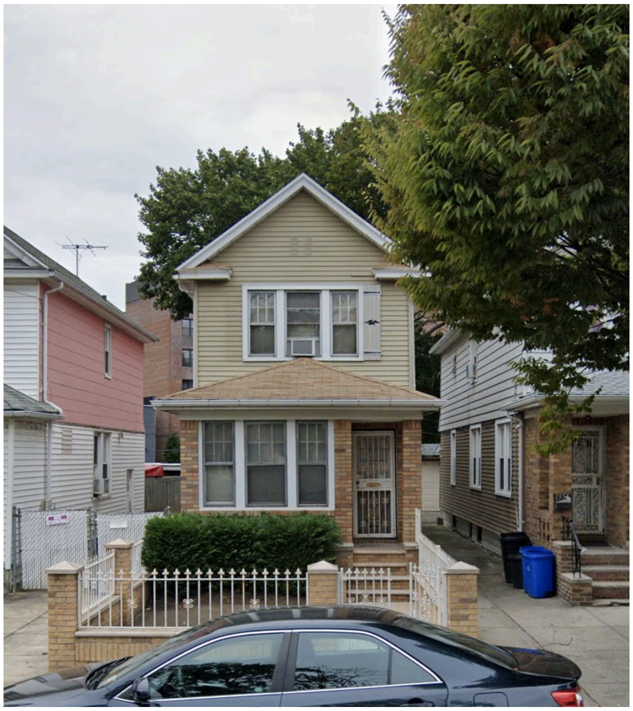
Given these discrepancies and the impact of the number of stories on the valuation, further investigation is required to verify the accuracy of the property information and understand the reasons behind these unusual characteristics. This includes verifying the number of stories and reassessing the building and lot dimensions to ensure accurate property valuation.

5. Record 536544

Owner: JEANTY JOSEPH P

Address: 1952 Troy Ave, Brooklyn, NY 11234

Building Class	V0	**FULLVAL**	1,469,960	**r5**	33	**r5_zip5**	81
Tax Class	1B	**AVLAND**	10	**r6**	16	**r6_taxclass**	92
Stories	NA	**AVTOT**	10	**r7**	21	**r7_taxclass**	58
LTFRONT	15	**BLDFRONT**	0	**r8**	55	**r8_taxclass**	92
LTDEPTH	97	**BLDDEPTH**	0	**r9**	33	**r9_taxclass**	91



Given the building area is zero, the ratios involving the building size and volume are not directly applicable. However, there are some characteristics showing that this property is unusual:

- As we can see from the photo, this is an existing property, so it's impossible to have building area is zero. Despite the building area being zero, the ratios involving the building size and volume are high, suggesting potential anomalies or special circumstances regarding its valuation.
- The property has unusually high ratios (r5 to r9) when standardized by TAXCLASS, indicating that the assessed land and total values relative to the building size and volume are significantly higher than average within its TAXCLASS
- Both the assessed land value (AVLAND) and the assessed total value (AVTOT) are unusually low. Despite this, the high ratios (r5 to r9) indicate that the property has high assessed value densities relative to the non-existent building dimensions, suggesting potential anomalies

Given these characteristics, further investigation is required to understand the reasons behind these anomalies. This includes verifying the accuracy of the building dimensions, reassessing the property for any unique features or location advantages, and checking for potential assessment errors.

Summary

In this project, we developed and implemented an unsupervised learning approach to detect anomalies in New York City property records. The primary objective was to identify unusual patterns and deviations in property valuations that could indicate potential fraudulent activities.

The project began with an extensive data cleaning and standardization phase. This step was crucial to ensure the accuracy and reliability of the subsequent analysis. We removed irrelevant records, such as government-owned properties, which do not follow typical market valuation practices and could distort the analysis. Missing values were addressed through imputation methods to maintain the integrity and completeness of the dataset.

Following the data cleaning phase, we created a series of relevant variables designed to highlight unusual property valuations. These variables included ratios such as the full market value to lot size, the assessed land value to building size, and others. These ratios were selected to capture various aspects of property value density and assessment efficiency, facilitating the detection of outliers and anomalies.

To manage the high dimensionality of the dataset and focus on the most significant patterns, we employed Principal Component Analysis (PCA). PCA allowed us to reduce the dimensionality of the dataset while retaining the most critical variance information. This step simplified the data, making it easier to identify anomalies by transforming it into a set of uncorrelated principal components.

After reducing the dimensionality, we applied two primary anomaly detection methods: a distance-based approach and an autoencoder method. The distance-based approach involved calculating the Minkowski distance of each record from the origin in the multi-dimensional space defined by the z-scores of the principal components. This method provided a measure of how unusual each record was compared to the norm.

In parallel, we used an autoencoder, a type of neural network designed to learn efficient codings of the data. The autoencoder was trained to reconstruct the input data, and the reconstruction error was used as an anomaly score. Higher reconstruction errors indicated records that deviated significantly from the learned patterns, suggesting potential anomalies.

To enhance the robustness of our detection system, we combined the scores from both methods. This combination leveraged the strengths of each approach, providing a more comprehensive assessment of potential anomalies. The final scores highlighted several properties with atypical valuations, warranting further investigation.

The results of this project provide a solid foundation for identifying potential fraudulent activities in the property market. Additionally, ongoing collaboration with stakeholders will be essential to refine the model continuously. By incorporating feedback and insights from domain experts, the model can be adjusted to reflect the latest trends and nuances in the property market, ensuring it remains effective in detecting fraudulent activities.

Overall, this project demonstrates the potential of unsupervised learning approaches in fraud detection and highlights the importance of data preparation, dimensionality reduction, and robust anomaly detection methods. By continuously refining the model and incorporating expert feedback, we can enhance the accuracy and effectiveness of fraud detection in New York City's property records.

Appendix: Data Quality Report

1. Data description

The dataset contains property and tax assessment information in New York. This dataset comprises 1,070,994 records and 32 fields, covering comprehensive details on property assessments, exemptions, and other relevant characteristics. It includes various identifiers, measurements, valuations, and classification codes related to the properties.

2. Summary Tables

Table 1: Numeric Fields Table

Field Name	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Std Dev	Most Common Value
LTFRONT	1,070,994	100.00%	0	0.00	9,999	36.64	74,03	0.00
LTDEPTH	1,070,994	100.00%	0	0.00	9,999	88.96	76.40	100.00
STORIES	1,014,730	94,70%	56,264	1.00	119.00	6.01	8.37	2.00
FULLVAL	1,070,994	100.00%	0	0.00	6,150,000,000.00	874,264.51	11,582,425.58	0.00
AVLAND	1,070,994	100.00%	0	0.00	2,668,500,000.00	85,067.92	4,057,258.16	0.00
AVLAND2	282726	26.40%	788,268	3.00	2,371,005,000.00	246235.72	6,178,951.64	2,408.00
AVTOT	1,070,994	100.00%	0	0.00	4,668,308,947.00	227,238.17	6,877,526.09	0.00
AVTOT2	282,732	26.40%	788,262	3.00	4,501,180,002.00	713,911.44	11,652,508.34	750.00
EXLAND	1,070,994	100.00%	0	0.00	2,668,500,000.00	36,423.89	3,981,573.93	0.00
EXLAND2	87,449	8.20%	983545	1.00	2,371,005,000.00	351,235.68	10,802,150.91	2,090.00
EXTOT	1,070,994	100.00%	0	0.00	4,668,308,947.00	91,186.98	6,508,399.78	0.00
EXTOT2	130,828	12.20%		7.00	4,501,180,002.00	656,768.28	16,072,448.75	2,090.00
BLDFRONT	1,070,994	100.00%	0	0.00	7,575.00	23.04	35.58	0.00
BLDDEPTH	1,070,994	100.00%	0	0.00	9,393.00	39.92	42.71	0.00

Table 2: Categorical Fields Table

Field Name	**# Records with Values**	**% Populated**	**# Zeros**	**# Unique Values**	**Most Common Value**
RECORD	1,070,994	100.00%	0	1,070,994	1
BBLE	1,070,994	100.00%	0	10.070,994	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	4.00%	1,066,358	13	E
EXCD	638,488	60.00%	432,506	129	1017
EXCD2	92,948	9.00%	978,046	60	1017
OWNER	1,039,249	97.00%	31,745	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	33.00%	716,689	3	G
STADDR	1,070,318	99.99%	676	839,280	501 SURF AVENUE
ZIP	1,041,104	97.20%	29,890	196	10314
EXMPTCL	15,579	1.50%	1,055,415	14	X1
PERIOD	1,070,994	100.00%	0	1	FINAL
VALTYPE	1,070,994	100.00%	0	1	AC-TR
YEAR	1,070,994	100.00%	0	1	2010/11

3. Visualization of Each Field

1) Record

Description: An identifier or record number for each entry. This is a unique number assigned to each property record, from 1 to 1,070,994.

2) BBLE

Description: Borough-Block-Lot-Easement identifier, which is a unique identifier used in the property database to locate and identify a property within a specific borough.

3) BORO

Description: The borough number, indicates the specific borough within the city.

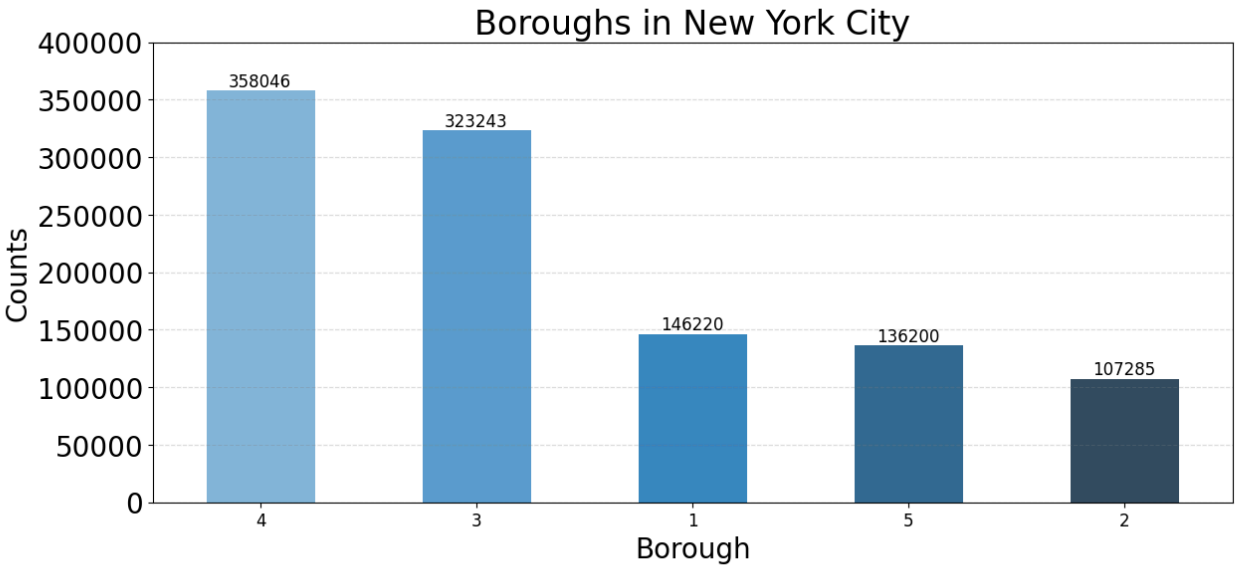


Figure 1: Distribution of Borough

Figure 1 shows the distribution of property records across the different boroughs in New York City with the x-axis representing the boroughs in New York City, and the y-axis representing the count of property records in each borough. Borough 4 has the highest number of property records suggesting it may be the most populated or densely developed borough in terms of property data, with 358,046 counts, followed by borough 3 with 323,243 counts.

4) BLOCK

Description: The block number within the borough, is used to identify a specific area or block.

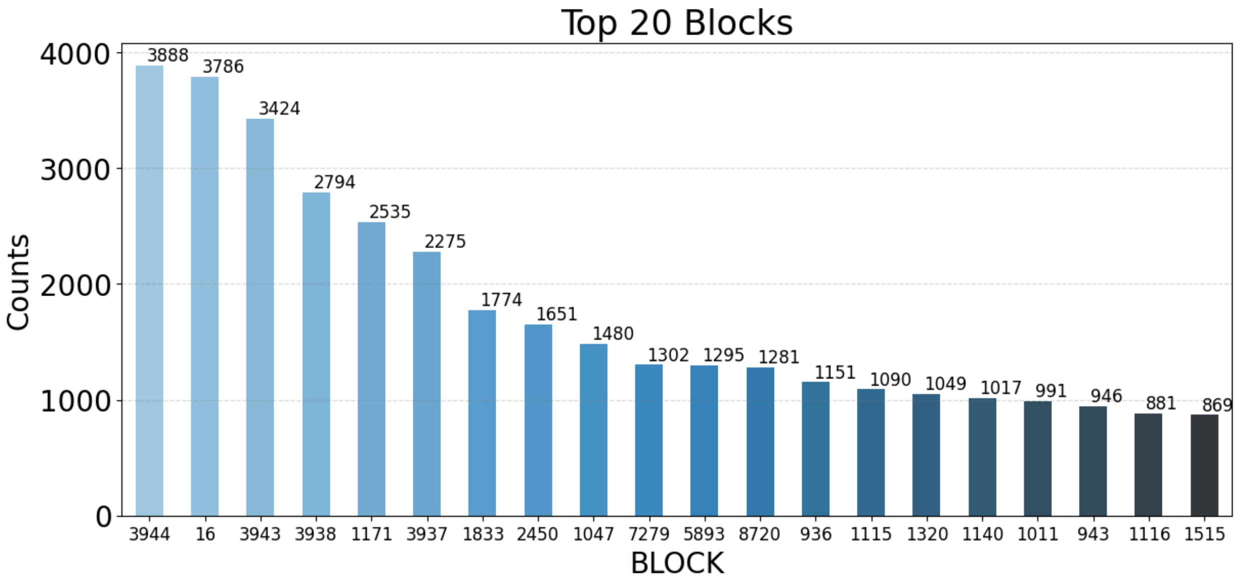


Figure 2: Distribution of top 20 Blocks

Figure 2 represents the top 20 blocks with the highest concentration of property records. The distribution of counts suggests that a few blocks have significantly more property records compared to others. **Block 3944** has the most property records, with **3,888 counts,** indicating it is likely a highly populated or densely developed block, followed by **block 16** with **3,786 records**. This plot can be useful for understanding property distribution and identifying areas with high property density for urban planning, real estate analysis and resource allocation.

5) LOT

Description: The lot number within the borough

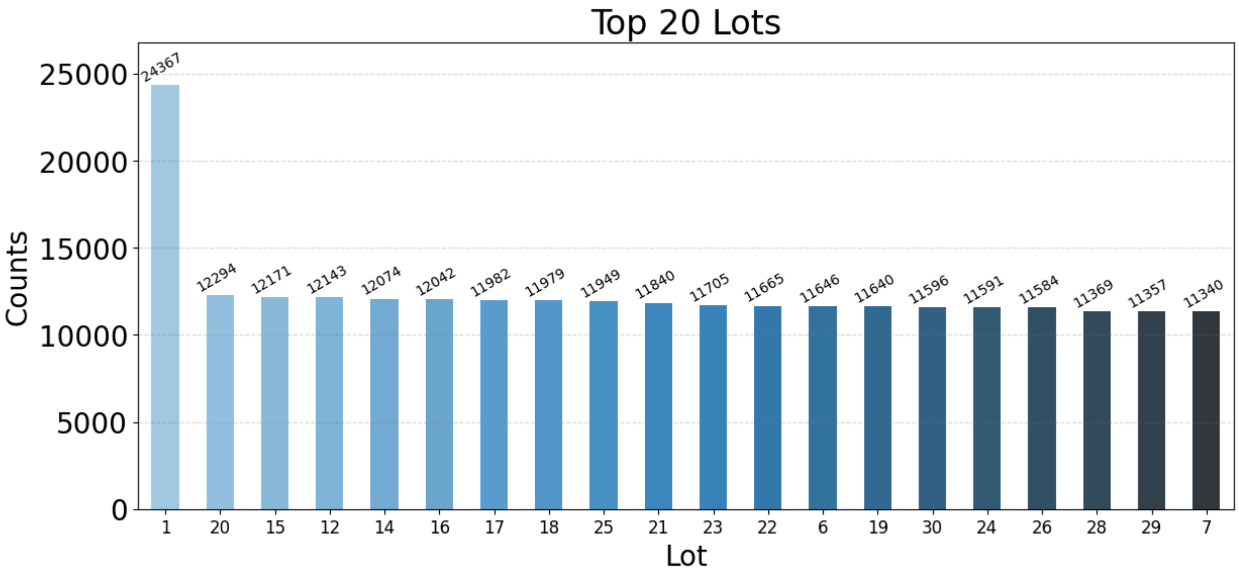


Figure 3: Distribution of top 20 Lots

Figure 3 represents the top 20 lots with the highest number of property records which has an x-axis representing the lot number and a y-axis representing the count of property records in each lot. **Lot 1** stands out with a substantially higher number of property records compared to other lots, indicating it might be a large lot or a common lot number assigned to many properties. This might warrant further investigation to understand why this lot has so many records compared to others. The counts for the other top 19 lots are relatively close, ranging from around 11,340 to 12,294 records. Overall, this distribution suggests that a few specific lot numbers are associated with a large number of properties, which could be due to various reasons such as zoning regulations, lot size, or administrative practices.

6) EASEMENT

Description: A code or indicator related to property easements, which are legal rights to use a portion of the property for a specific purpose.

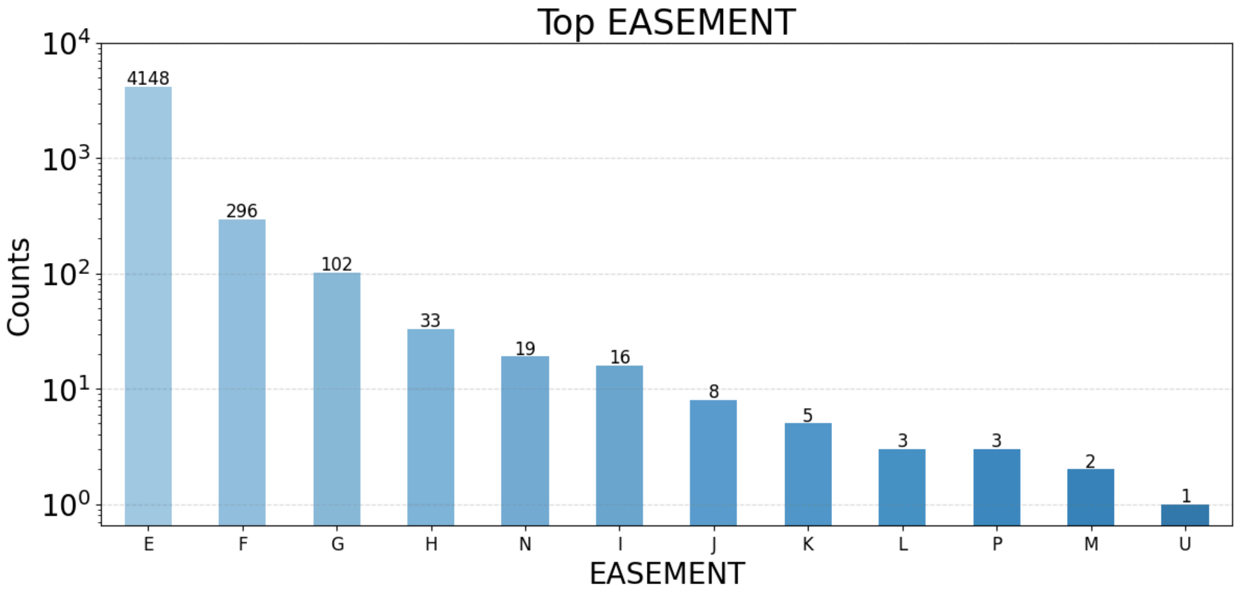


Figure 4: Distribution of top Easement

Figure 4 represents the count of various easement types, showcasing the top easement codes in the dataset with the x-axis representing the different easement codes and the y-axis representing the count of property records for each easement code. The easement **code E** is by far the most common, with **4,148 counts**, suggesting it is a frequently applied easement type within the dataset. There is a steep drop-off from E to the next (F), indicating that the easement code E is significantly more prevalent. The distribution also shows a few easement codes with moderately high counts (F, G), followed by a long tail of codes with much lower counts.

7) Owner

Description: An identifier for the property owner

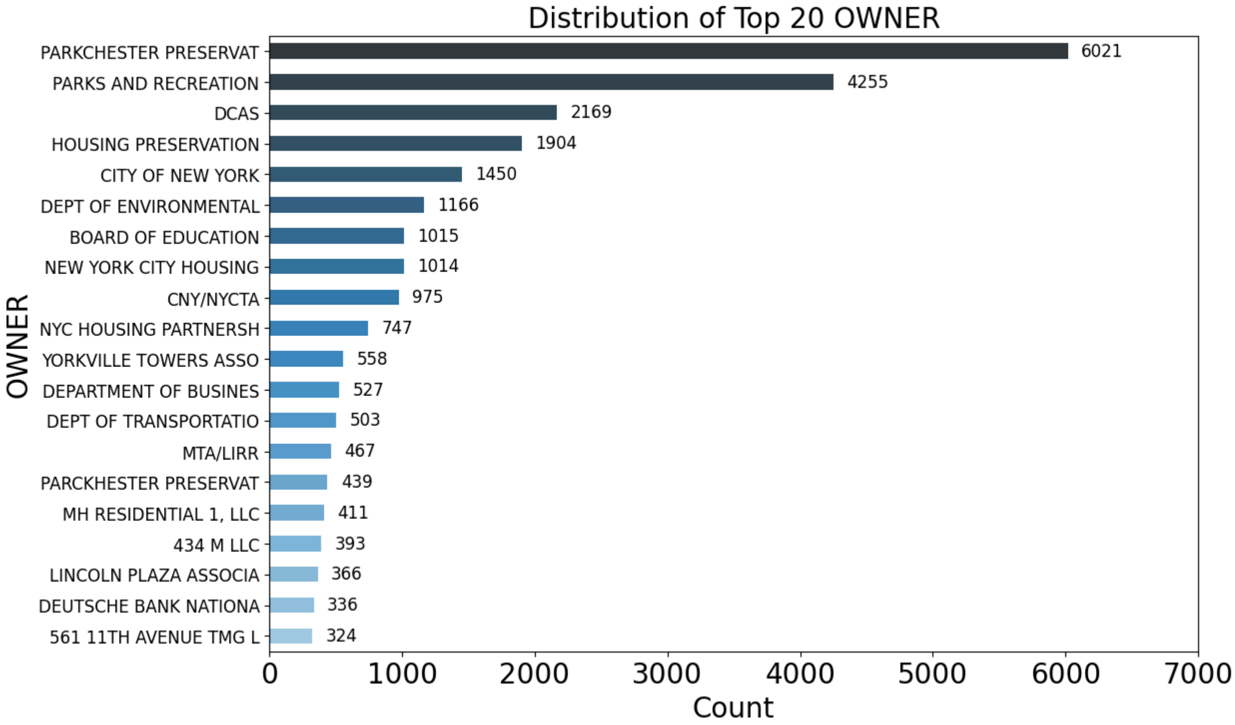


Figure 5: Distribution of top 20 owner

Figure 5 shows the top 20 property owners by the number of properties they own with the y-axis representing the names of the owners and the x-axis representing the count of the property owned by each owner. **PARKCHESTER PRESERVAT** is the highest property owner with **6,021 properties**, followed by **PARKS AND RECREATION** with **4,255 properties**. These owners are likely large organizations, government agencies, or institutions with significant property holdings. The plot shows a clear drop-off in the number of properties owned as we move down the list of top owners. From this plot, we can do further analysis such as: Investigating the types of properties owned by these top owners to understand their property portfolios better or analyzing geographic distribution to see if these owners have properties concentrated in specific areas.

8) BLDGCL

Description: Building class, which classifies the type of building.

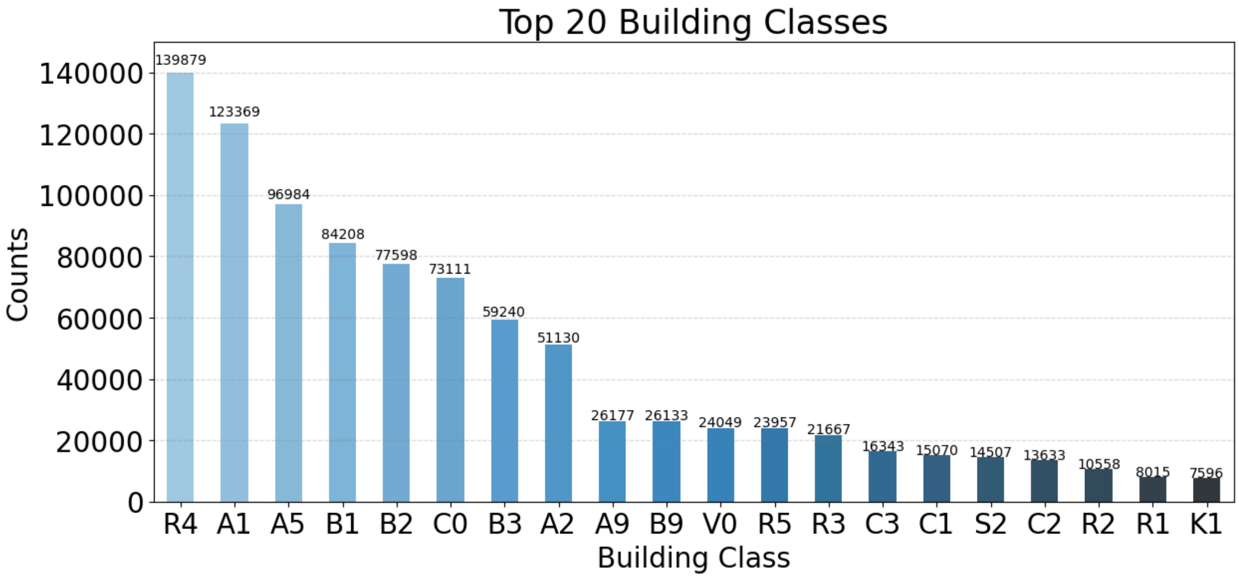


Figure 6: Distribution of top 20 building classes

Figure 6 represents the distribution of the top 20 building classes by the number of property records with the x-axis representing the different building class codes and y-axis representing the count of property records for each building class. **R4** is the most common building class with **139,879 counts**, followed by **A1** with **123,369 counts**, suggesting they might represent common types of buildings such as residential homes or apartment complexes. There is a clear decrease in the number of property records as we move from the most common building classes (R4, A1) to the less common ones (R1, K1)

9) **TAXCLASS**

Description: Tax class code, indicating the property tax classification, which affects the tax rate applied to the property.

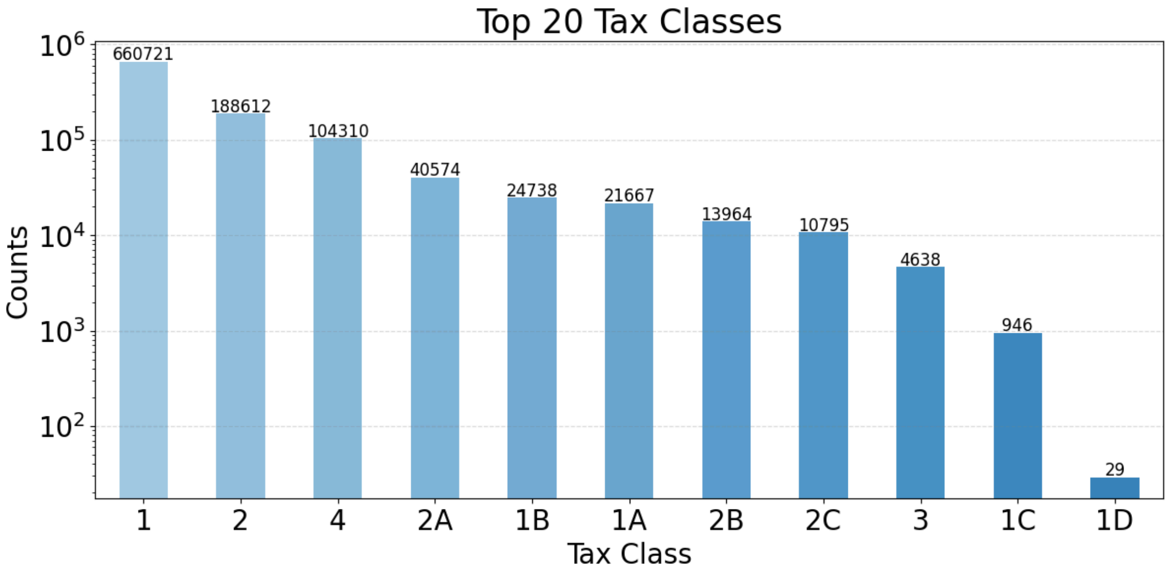


Figure 7: Distribution of the top 20 Tax Classes

Figure 7 shows the distribution of the tip 20 tax classes by the number of property records with the x-axis representing the different tax class codes and the y-axis representing the count of property records for each tax classes. **Tax class 1** is by far the most common with **660,721 counts**, representing that a majority of the properties fall under this classification. Tax **Class 2** also has a significant number of records with **188,612 counts** but much fewer than Tax Class 1. The remaining tax classes have considerably lower counts with a steep drop-off from the top two classes to the rest.

10) **LTFRONT**

Description: Lot frontage, representing the width of the lot at the front, measured in feet.

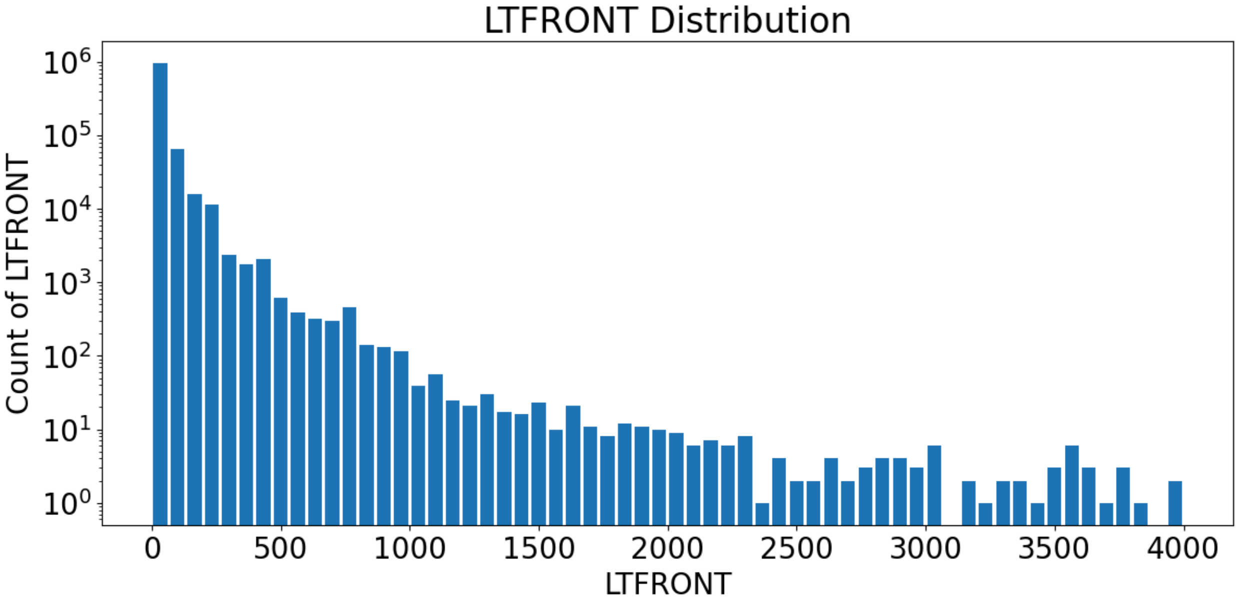


Figure 7: Distribution of LTFRONT

The distribution represents the distribution of LTFRONT values in the dataset with the x-axis representing the LTFRONT values, which are measurements of lot frontage and the y-axis represents the count of occurrences for each LTFRONT value. The majority of properties have small lot frontage. This is evident from the tall bars at the lower end of the LTFRONT scales. As the LTFRONT increases, the frequency of properties with those frontages smaller lots are more common, and larger lots are less frequent. By understanding the distribution of lot frontages, we can inform urban planning and zoning regulations. Areas with predominantly small lot frontages might be residential zones with single-family homes, while larger lot frontages might indicate commercial or industrial areas.

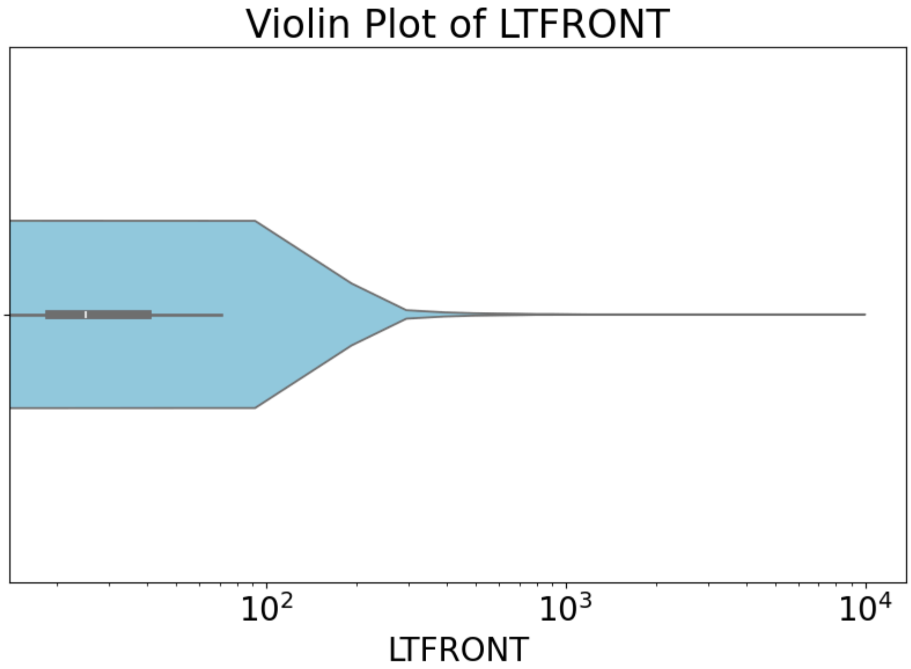


Figure 8: Violin Plot of LTFRONT

The plot shows a density estimate of the LTFRONT values with the width of the violin representing the density of the data at different values. The plot is wider at lower values and narrows significantly as the values increase, indicating that most data points are concentrated at lower LTFRONT values, which is consistent with Figure 7 showing a large number of small values. The long right tail of the plot indicates that there are some properties with very large lot frontage, extending up to 10,000 feet. However, these are relatively rare.

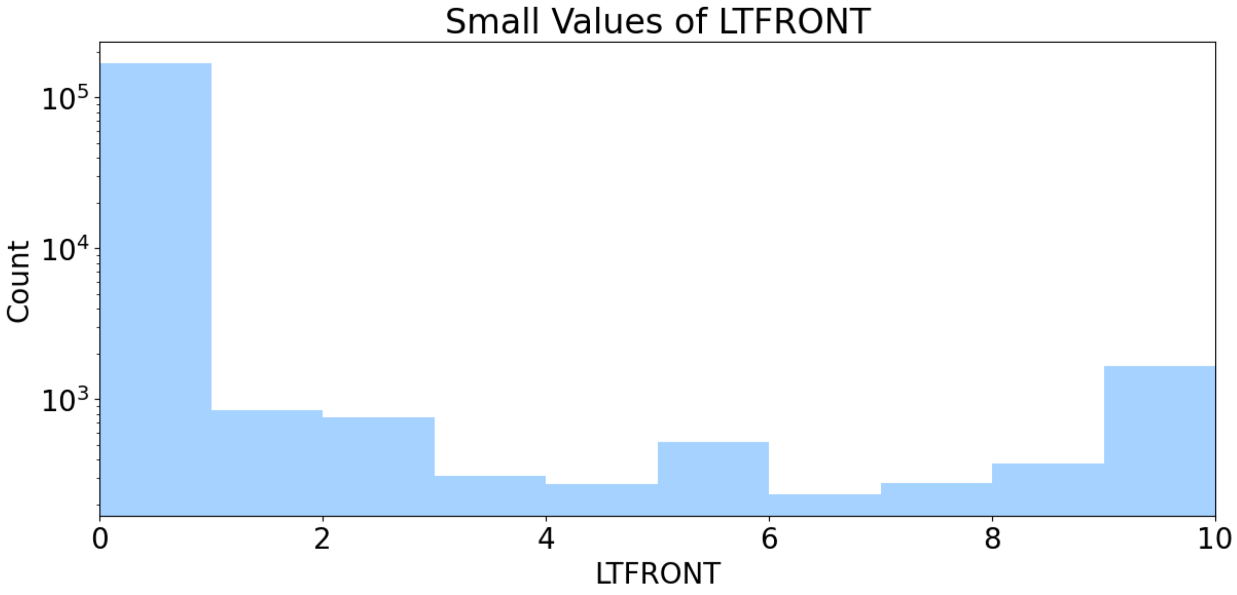


Figure 9: Distribution of Small Values of LTFRONT.

Figure 9 shows the distribution of small values of LTFRONT, focusing on the range from 0 to 10 feet. The majority of the LTFRONT values are concentrated at the lower end of the range (0 to 1 foot), as indicated by the tall bar at the beginning of the histogram. This suggests that many properties have lot frontages in this very small range.

11) **LTDEPTH**

Description: Lot depth, representing the depth of the lot, measured in feet.

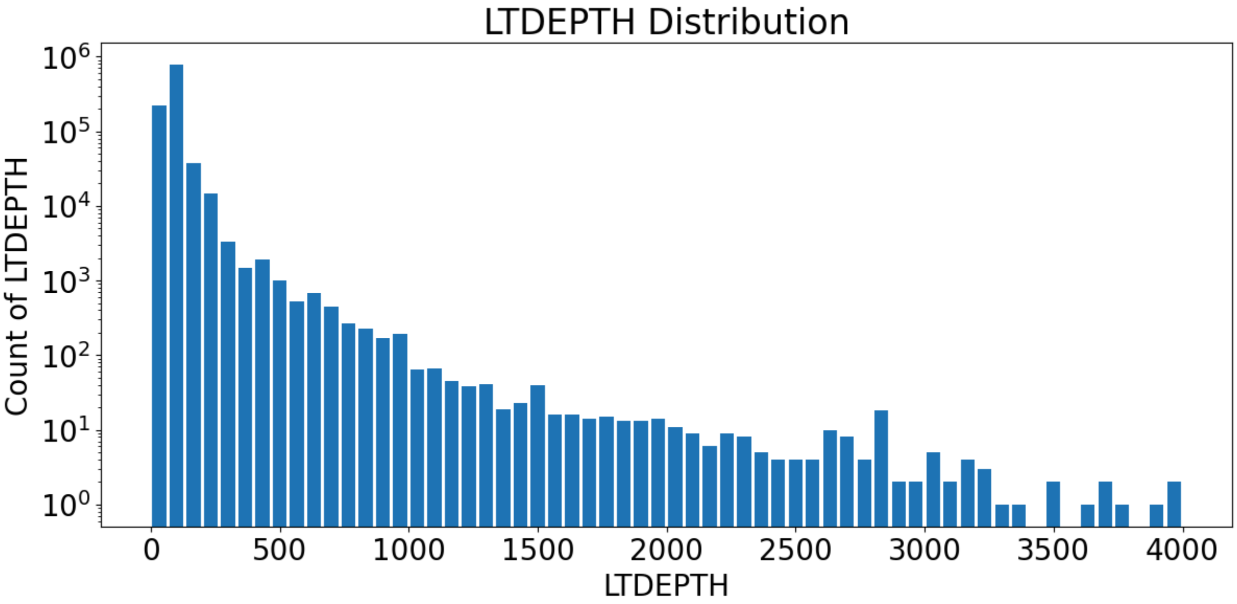


Figure 10: Distribution of LTDEPTH

Figure 10 shows the distribution of LTDEPTH values in the dataset, with the y-axis representing the count of occurrences for the LTDEPTH value, and the x-axis representing the LTDEPTH values. Same as LTFRONT, the majority of LTDEPTH values are concentrated at the lower end of the scale (e.g. 0 to 500 feet), as indicated by the tail bars at the beginning of the histogram. This suggests that many properties have small lot depths. There is a long tail extending to the right, indicating that there are some properties with very large lot depths, though these are relatively rare. As LTDEPTH increases, the frequency of occurrences drops significantly, which is a common pattern in property data where smaller lot dimensions are more prevalent.

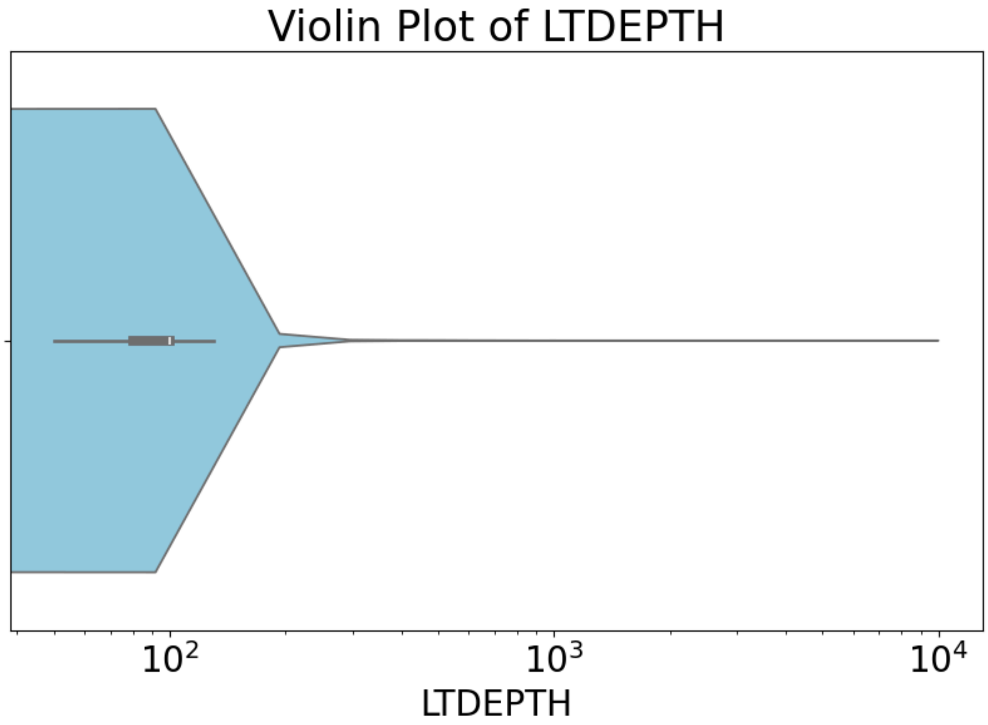


Figure 11: Violin Plot of LTDEPTH

Figure 11 provides a clear visual representation of the distribution, highlighting the concentration of data at lower values and the presence of outliers with very large lot depths. This plot is crucial for understanding the overall property characteristics and making informed decisions in urban planning, real estate analysis, and data modeling.

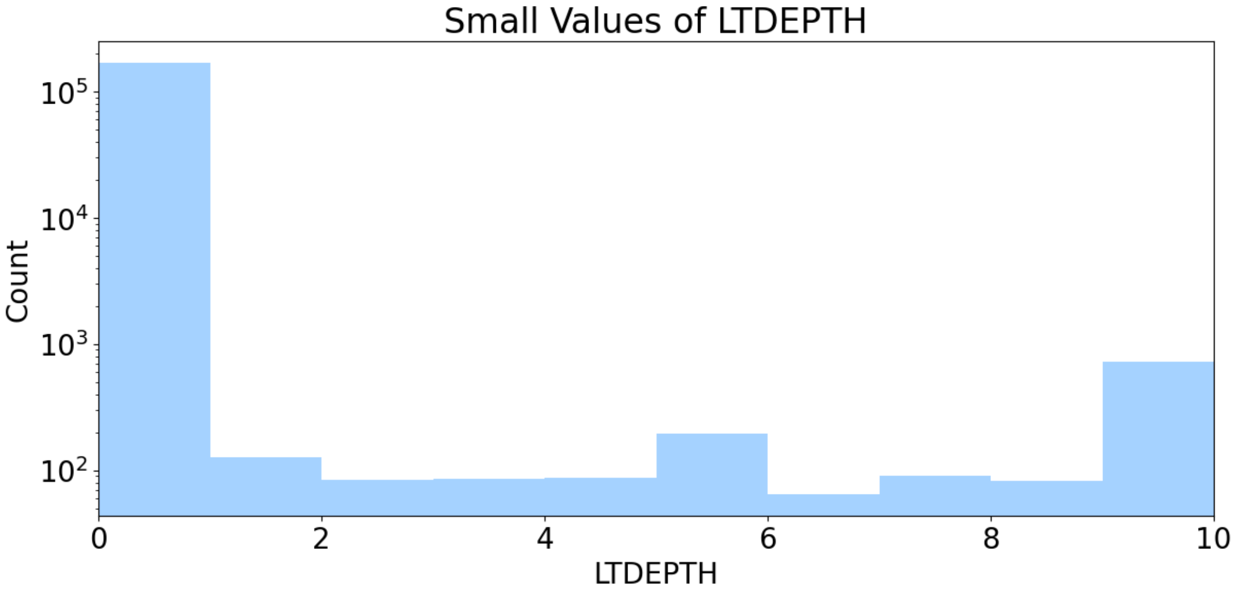


Figure 12: Distribution of small values of LTDEPTH

The histograms provide a focused view of the distribution of small LTDEPTH values, showing the same states as the two plots above that a large number of properties have extremely small lot depths

12) EXT

Description: An extension code.

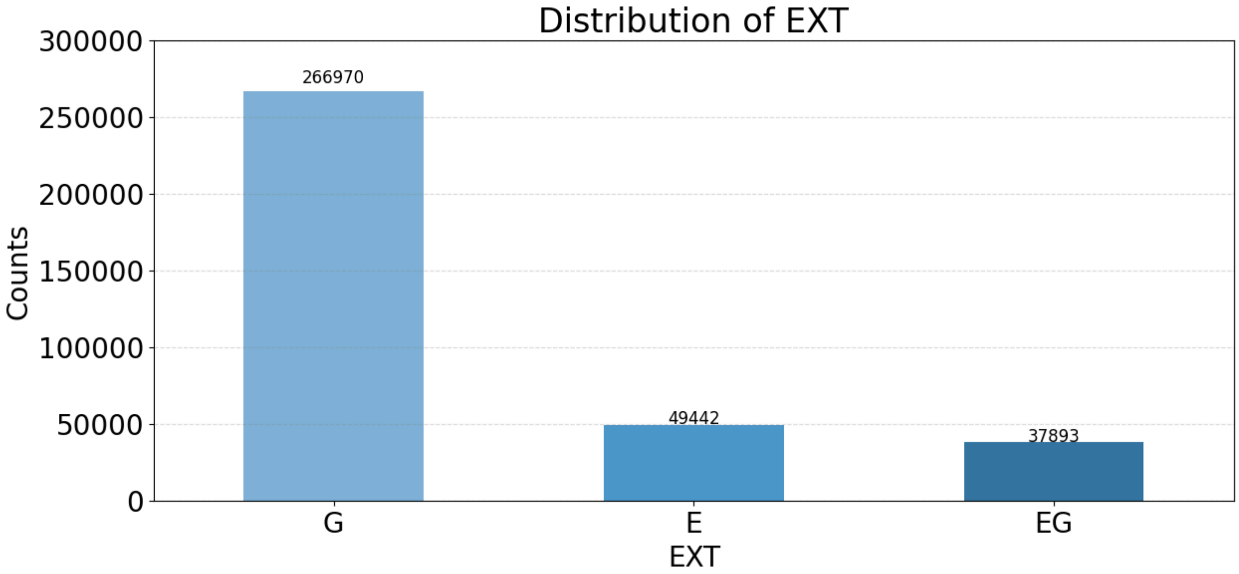


Figure 13: Distribution of EXT

Figure 13 illustrates the distribution of EXT values, showing that the majority of properties fall under category **G with 266,970 properties**, which could be indicative of a common property characteristic or status within this dataset. Categories E and EG have significantly fewer records compared to G, suggesting that these categories are less common.

13) **STORIES**

Description: Number of stories (floors) in the building, indicating the vertical size of the structure.

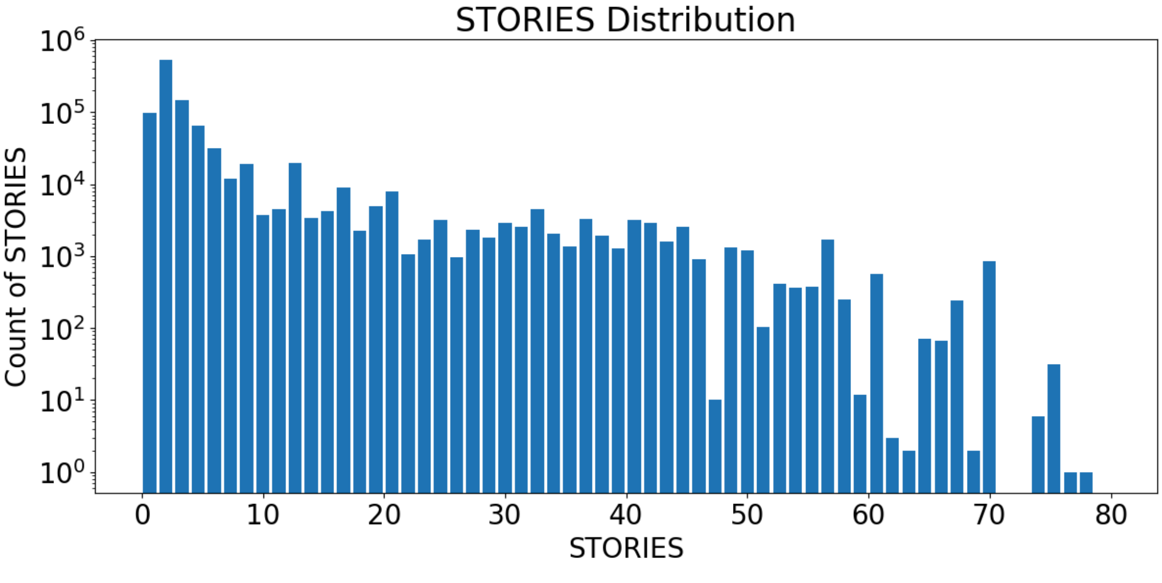


Figure 14: Distribution of STORIES

Figure 14 demonstrates the distribution of STORIES values with the x-axis representing the number of stories in the building, ranging from 0 to 80, and the y-axis representing the count of occurrences for each number of stories. There is a significant peak at the lower end of the STORIES values, indicating that most buildings have fewer stories. The highest counts are seen around 1 to 10 stories. As the number of stories increases, the frequency of occurrences generally decreases, which is a common pattern in building data where shorter buildings are more prevalent. There are some buildings with very high story counts (50 to 80 stories), but these are relatively rare, as indicated by the shorter bars at the higher end of the scale.

14) **FULLVAL**

Description: Full market value of the property, representing the estimated market value for taxation purposes.

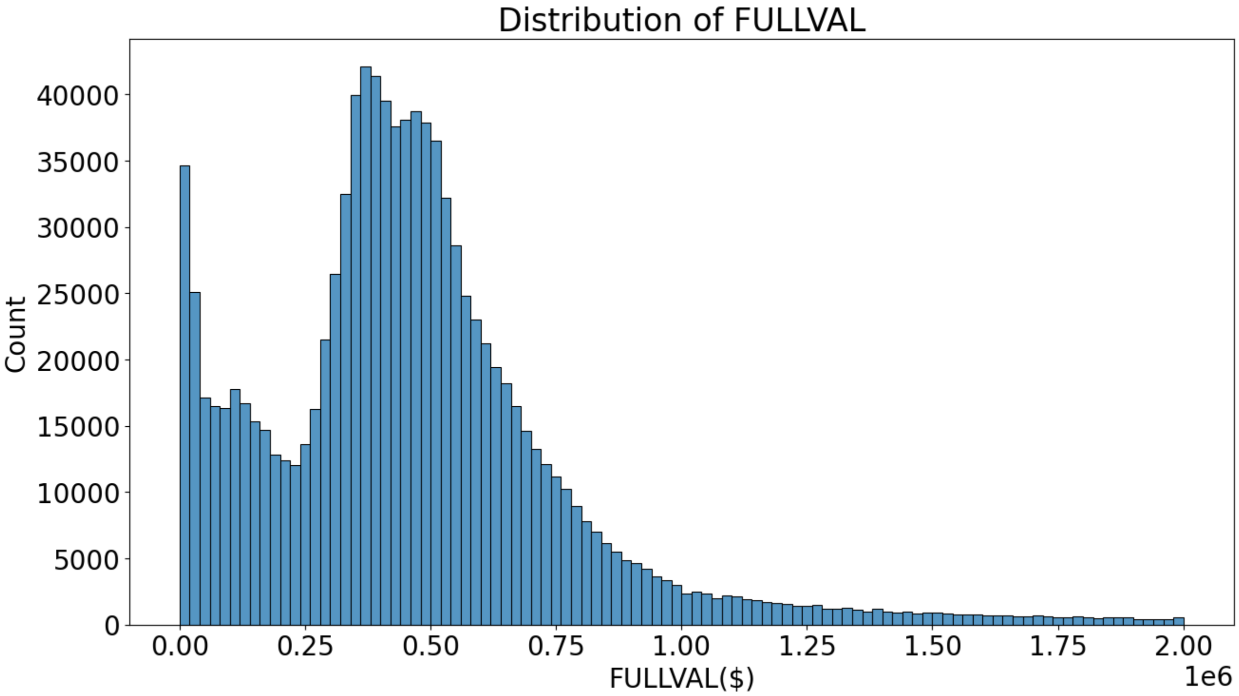


Figure 15: Distribution of FULLVAL

This histogram represents the distribution of FULLVAL of the data. The distribution shows two distinct peaks: one around $250,000 and another around $500,000. This suggests two common value ranges for properties, which might be typical for residential homes, smaller commercial buildings, or properties in less expensive areas. There is a significant count of properties with a full market value close to zero, which might indicate tax-exempt properties or data anomalies. As the market value increases, the frequency of occurrences generally decreases. This is a common pattern in property data where lower-valued properties are more prevalent.

15) **AVLAND / AVLAND2**

Description: Assessed value of the land, representing the value of the land portion of the property for tax assessment purposes.

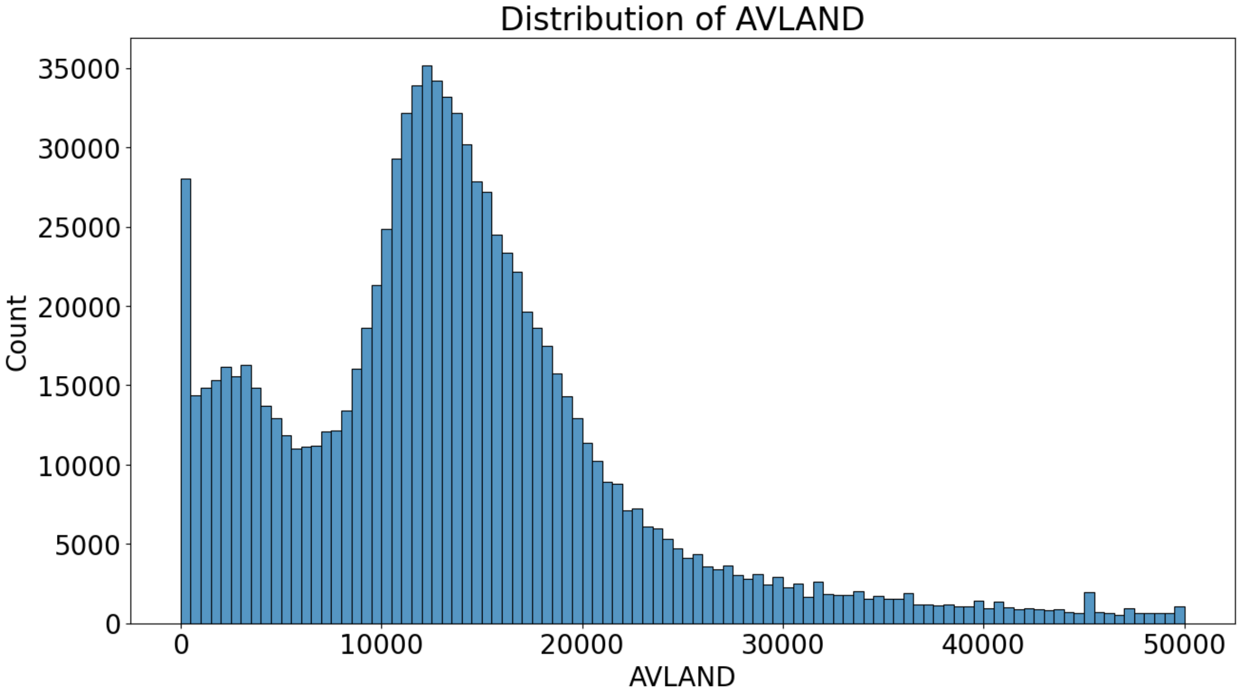


Figure 16: Distribution of AVLAND

The histogram provides a comprehensive view of the distribution of AVLAND values of the data. There is a significant peak at the lower end of the AVLAND values, indicating that most properties have lower assessed land values. The plot peaks around the $10,000 to $15,000 range, indicating a high concentration of properties with land values within this range. There is a notable count at the zero mark, similar to the AVTOT distribution, which could indicate properties that have no assessed land value, possibly due to exemptions or data entry errors. As the assessed land value increases, the frequency of occurrences generally decreases. This is a common pattern in property data where lower-valued land is more prevalent. This information is crucial for understanding the overall property characteristics and can guide further investigation into specific property types, real estate market analysis, and data modeling

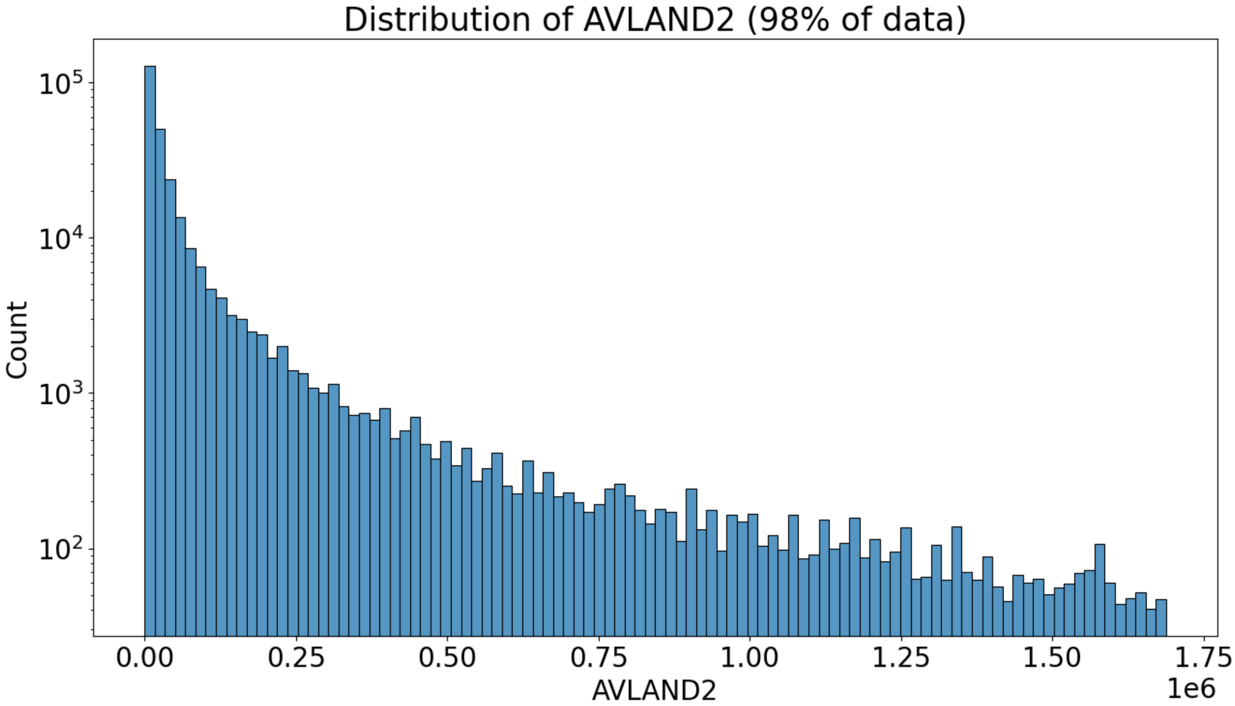


Figure 17: Distribution of AVLAND2

Figure 17 represents the distribution of AVLAND2 (Assessed Value of Land, possibly an alternative or additional assessment) for 98% of the data. There is a significant peak at the lower end of the AVLAND2 values, indicating that most properties have lower assessed land values. The highest counts are seen between approximately \$0 and \$250,000. As the assessed land value increases, the frequency of occurrences generally decreases. This is a common pattern in property data where lower-valued land is more prevalent.

16) **AVTOT / AVTOT2**

Description: Assessed total value, including both land and improvements (buildings, structures) on the property.

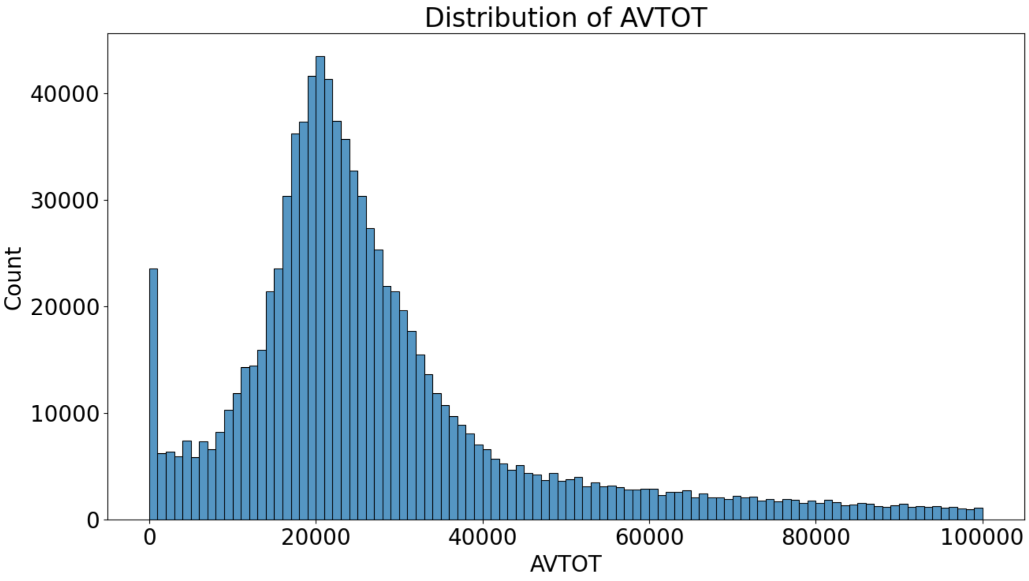


Figure 18: Distribution of AVTOT

Figure 18 shows the distribution of AVTOT of the data. There is a significant peak at the lower end of the AVTOT values, indicating that most properties have lower assessed total values. The high concentration of around $20,000 suggests that many properties fall into the lower to mid-value range. The spike at zero could indicate properties that are exempt from assessment, unoccupied land, or errors in data entry. As the assessed total value increases, the frequency of occurrences generally decreases which shows a long tail extending to the right, indicating that there are some properties with much higher assessed total values, though these are relatively rare. This is a common pattern in property data where lower-valued properties are more prevalent.

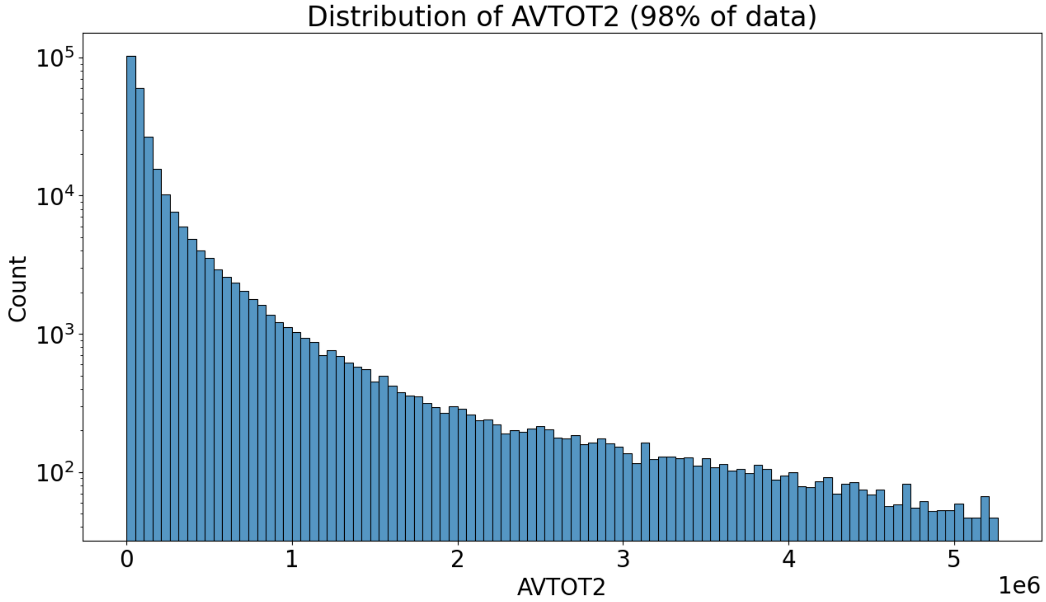


Figure 19: Distribution of AVTOT2

Figure 19 shows the distribution of AVTOT2 for 98% of the data. There is a pronounced peak at the lower end of the value range, indicating that the majority of properties have lower assessed total values. Specifically, the highest concentration is around 0 to 0.5 million dollars. As the value of AVTOT2 increases, the frequency of occurrences decreases. This suggests that higher total assessed values are less common. The distribution has a long tail extending towards higher values, indicating that while less frequent, there are properties with significantly higher total assessed values of up to 5 million dollars.

17) **EXTOT / EXTOT2**

Description: Extension of the total value, possibly an adjusted total value including various factors or adjustments.

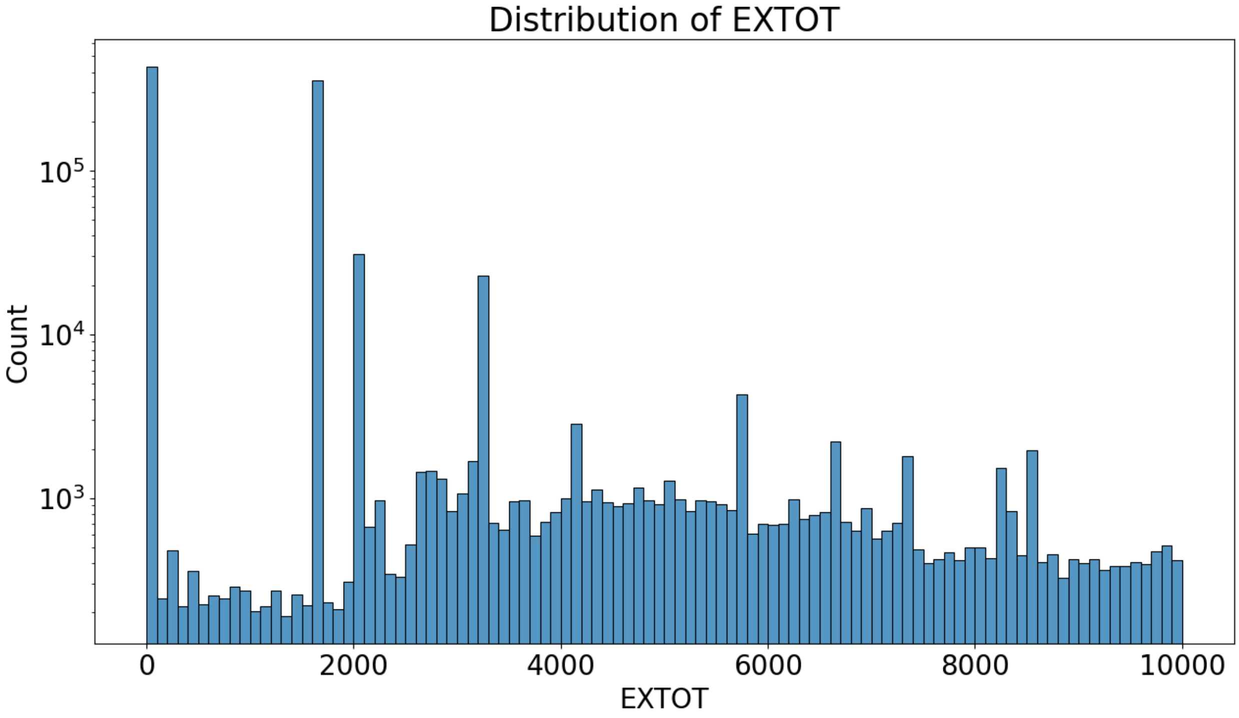


Figure 20: Distribution of EXTOT

The histogram of EXTOT values shows a distribution with a significant count of properties with lower EXTOT values, particularly close to zero. This indicates that a large number of properties have minimal to no total exempt amount. The plot exhibits several distinct peaks at regular intervals (around 2000, 4000, 6000, etc.). These peaks suggest common total exempt amounts applied to groups of properties.

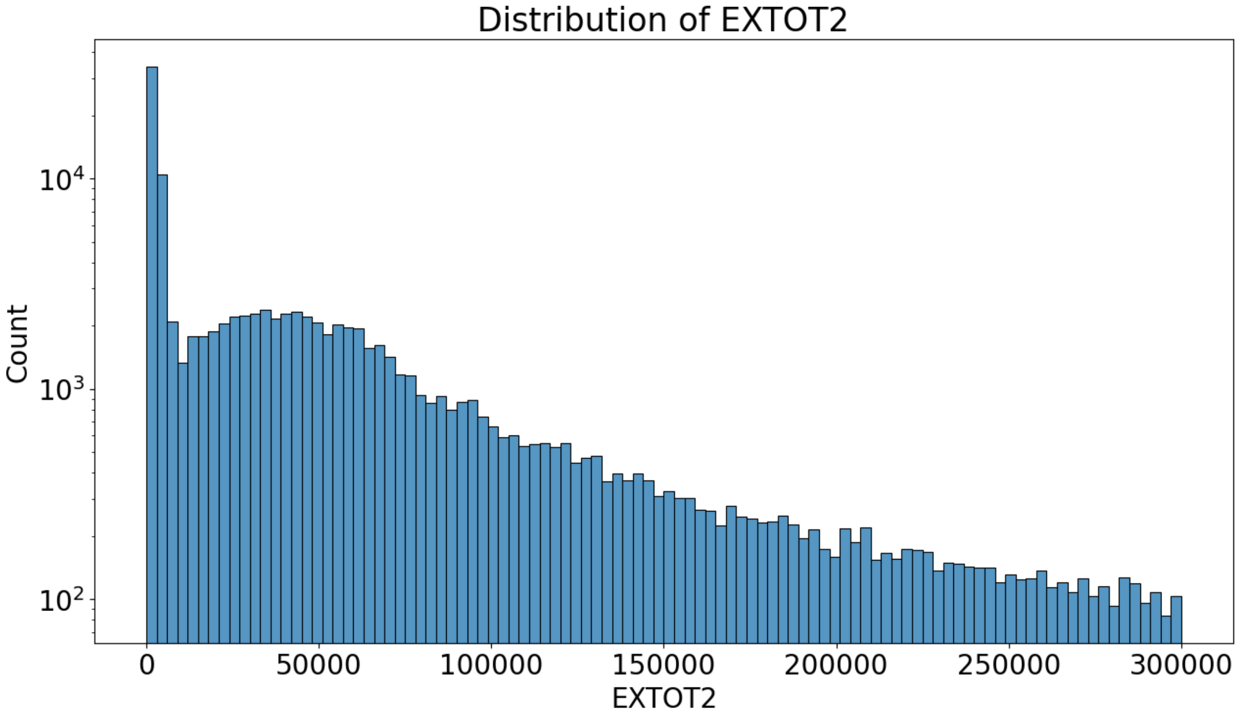


Figure 21: Distribution of EXTOT2

The histogram of EXTOT2 values reveals a significant initial peak at lower values and a gradual decline as the values increase. The initial peak suggests that a significant portion of properties have little to no total exempt amount. This could be due to specific exemption policies or a large number of properties not qualifying for exemptions. The decreasing trend indicates that properties with higher exempt amounts are less common. This could reflect the distribution of property types or the criteria for exemptions.

18) **EXPAND / EXLAND2**

Description: Represents the exempt portion of the land values

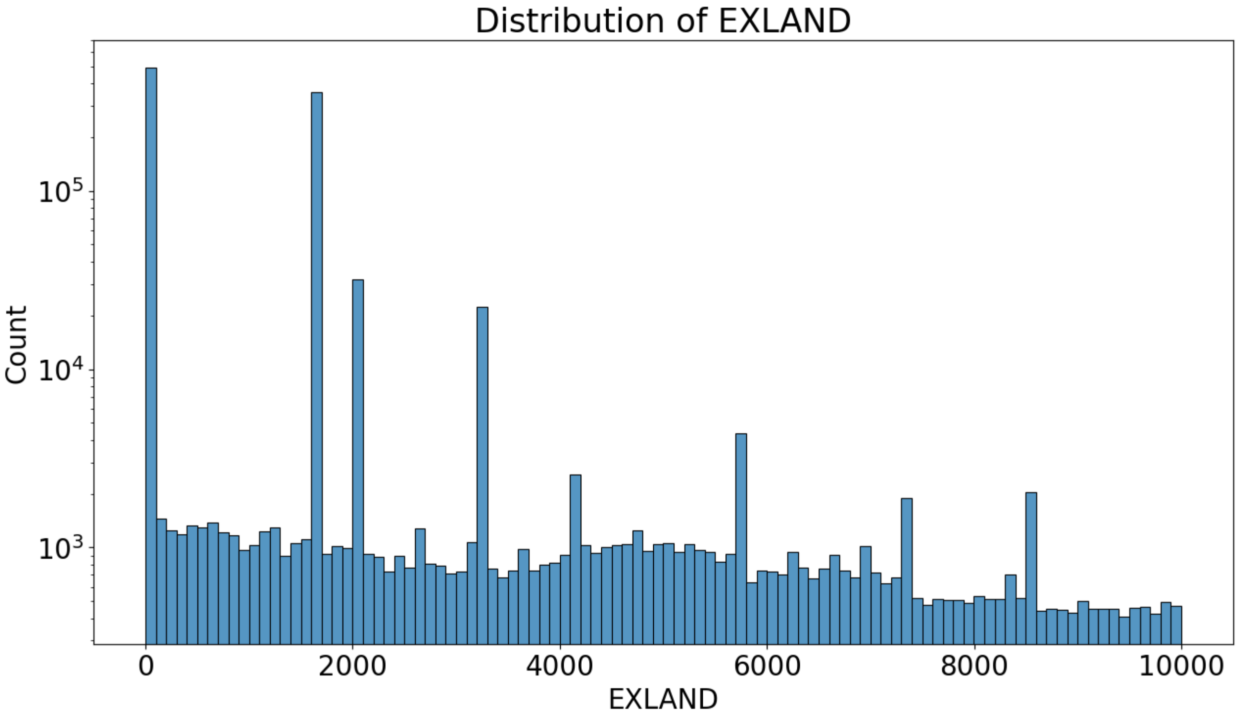


Figure 22: Distribution of EXLAND

The histogram shows the distribution of EXLAND with the x-axis representing the exempt portion of the land value (EXLAND), with values ranging from 0 to 10,000, and the y-axis representing the count of occurrences for each range of EXLAND values. There is a significant count of properties with an exempt land value of zero. This suggests that many properties do not have an exempt portion or that the exempt portion is recorded as zero. The distribution shows multiple distinct peaks at intervals (e.g., around 2000, 4000, 6000, and 8000). These peaks might indicate common values for exemptions, suggesting that there are standard exempt land values applied to certain groups of properties, possibly due to specific exemption policies or regulations.

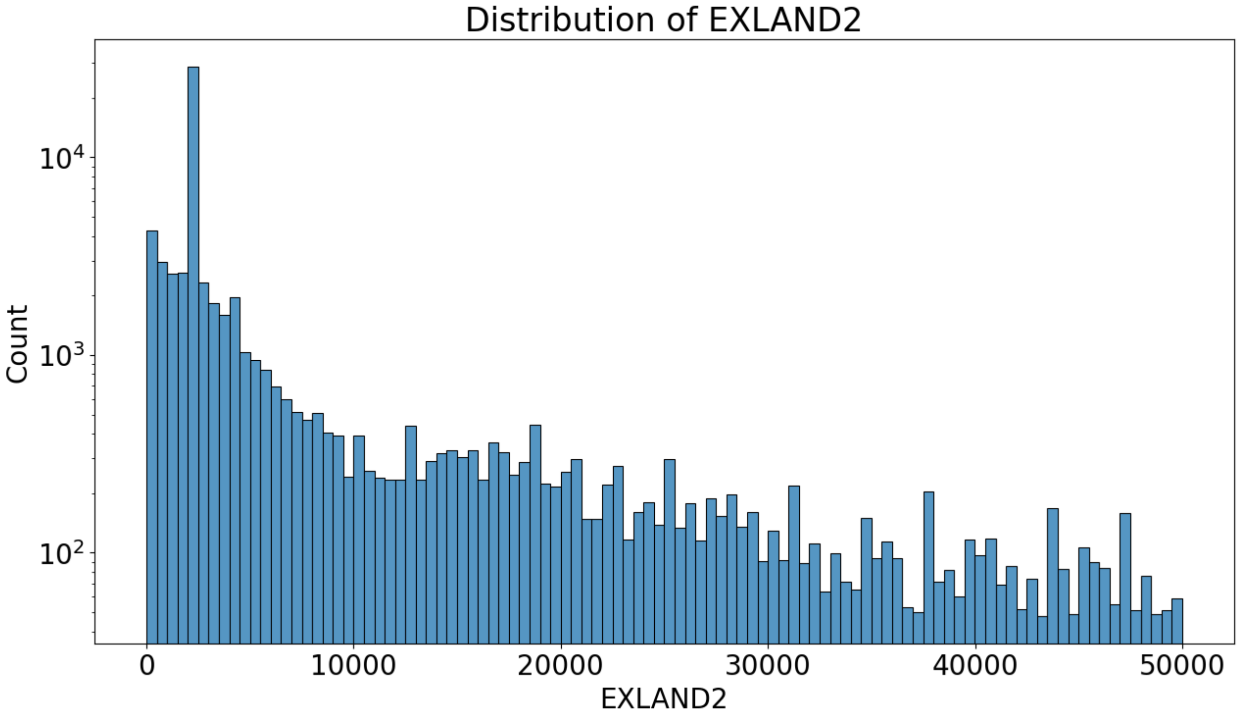


Figure 23: Distribution of EXLAND2

The histogram of EXLAND2 values shows a distribution with a significant initial peak at lower values and multiple distinct peaks at higher values. There is a significant count of properties with lower EXLAND2 values, particularly close to zero. This indicates that a large number of properties have minimal to no secondary land value exemption. The plot exhibits several distinct peaks, particularly around 10,000 and at intervals up to 50,000. These peaks suggest common values for secondary exemptions.

19) EXCD / EXCD 2

Description: Represent the exemption code, which specifies the type of tax exemption applied to the property.

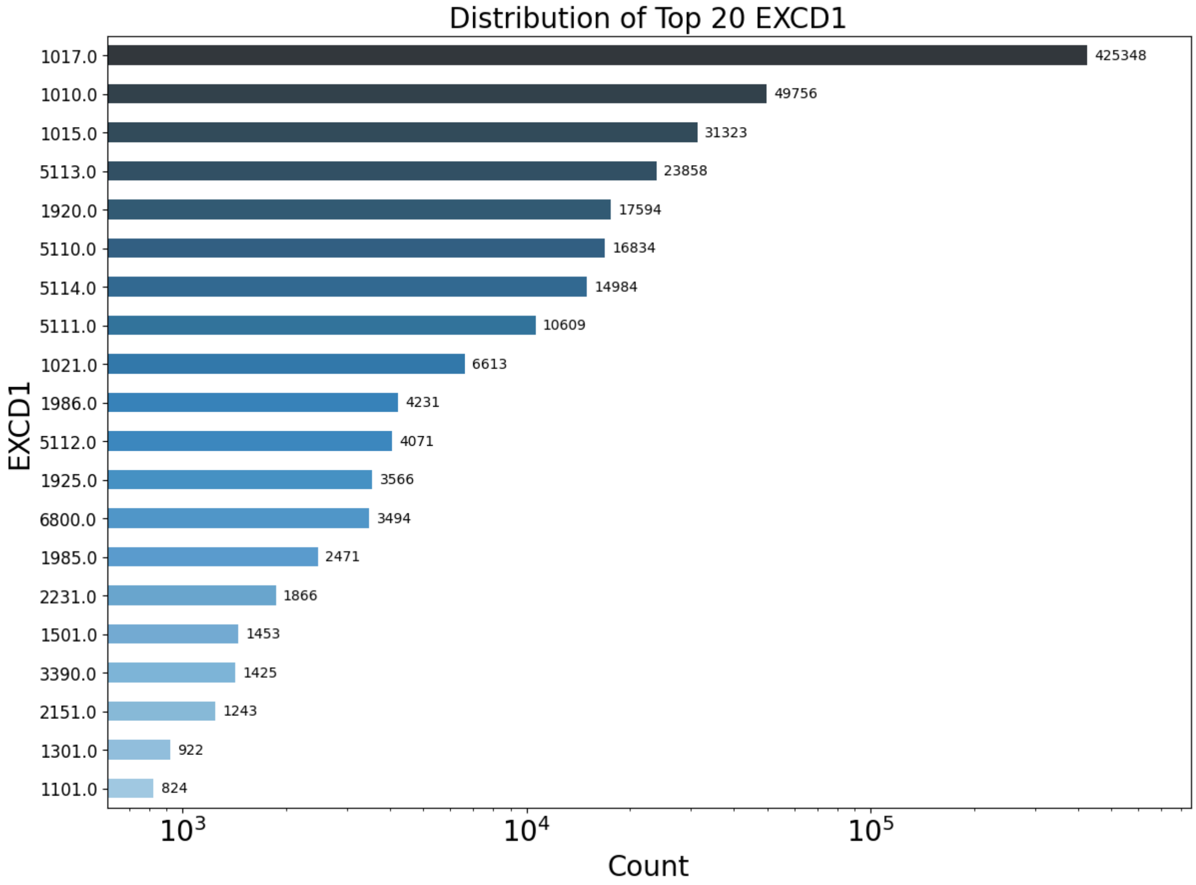


Figure 24: Distribution of top 20 EXCD1

The bar chart shows the distribution of the top 20 occurrences of EXCD1 values. with the x-axis representing the count of occurrences on a logarithmic scale, and the y-axis representing the EXCD1 value.

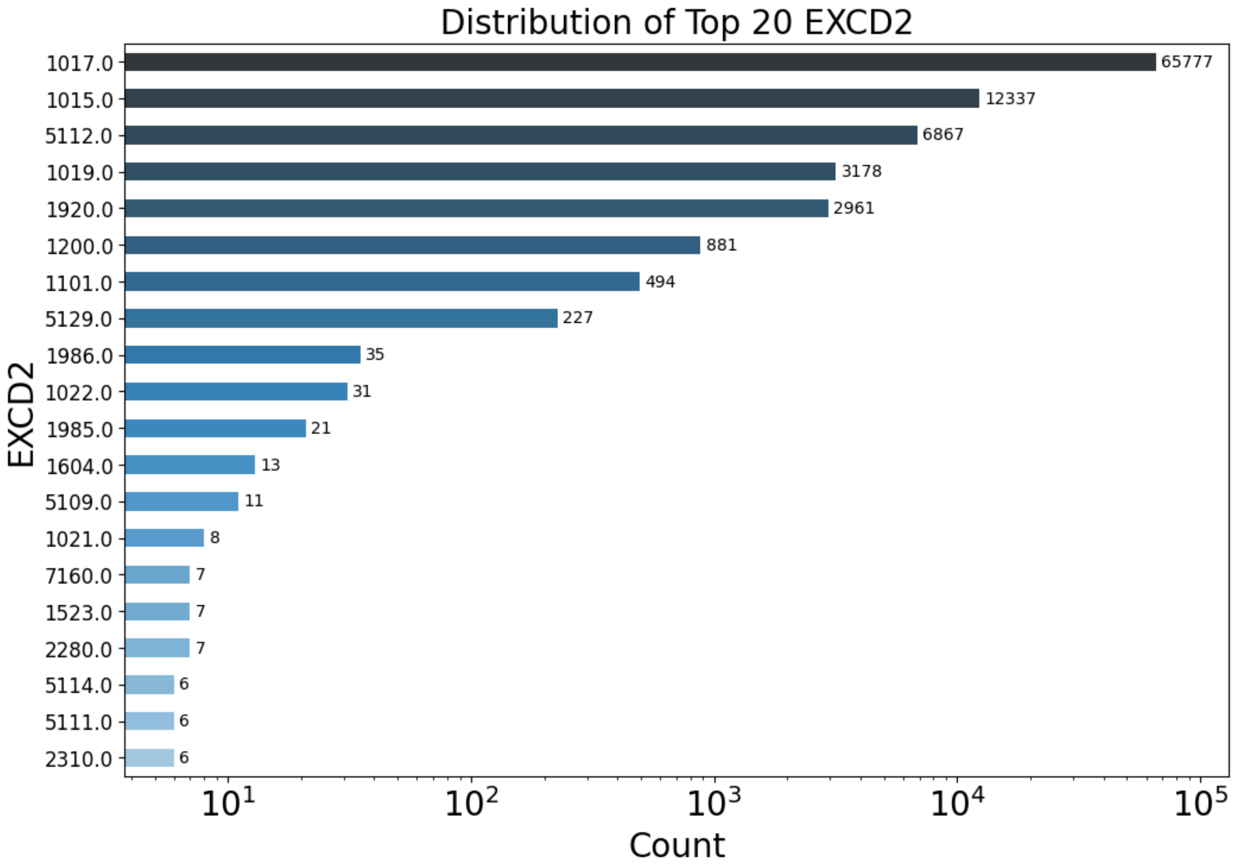


Figure 25: Distribution of top 20 EXCD2

Same as EXCD1, figure 25 illustrates the distribution of the top 20 occurrences of EXCD2 values with x-axis representing the count of occurrences on a logarithmic scale, and the y-axis representing the EXCD2 value. There is a significant drop from the highest occurrence (1017.0) with **425,348 counts** to the second highest (1015.0) with **49,756 counts**, showing that 1017.0 is an outlier or represents a very dominant category. The least common EXCD2 values in the top 20 are 5114.0, 5111.0, and 2310.0 each with 6 occurrences, which is significantly less common compared to the highest occurrence.

20) **STADDR**

Description: Street address of the property, providing the physical location address.

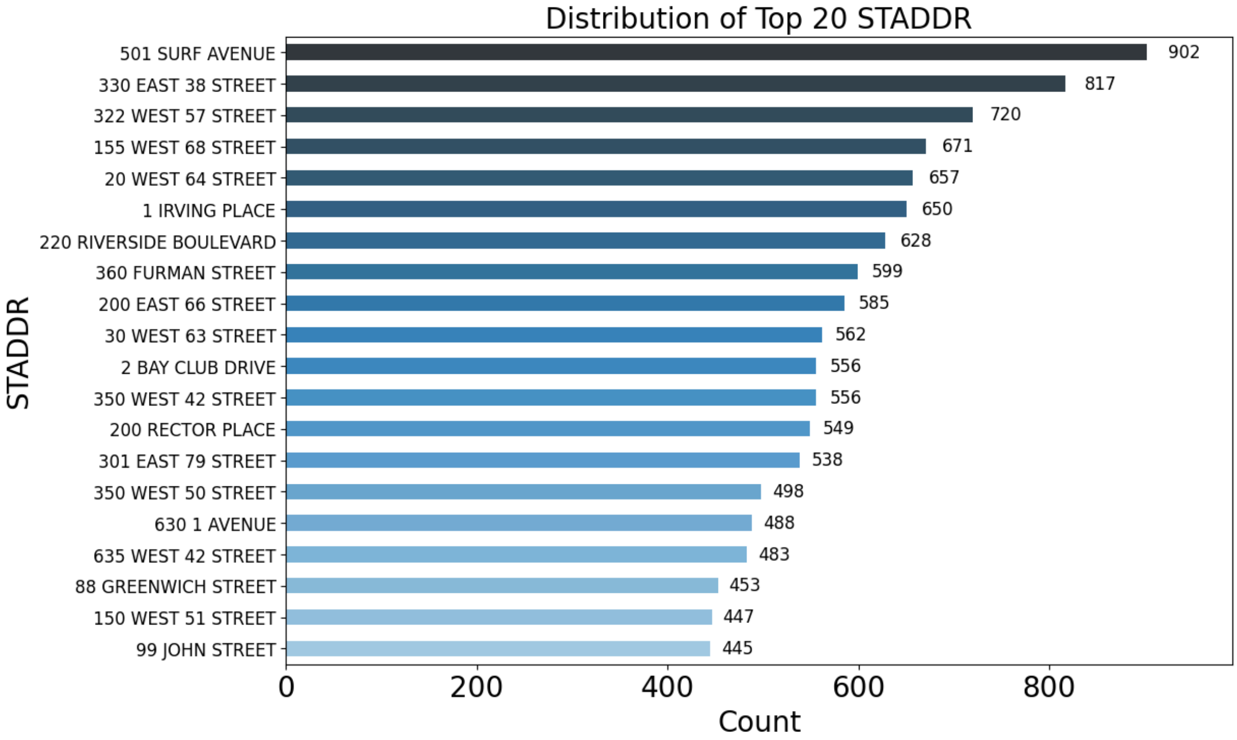


Figure 26: Distribution of top 20 STADDR

Figure 26 shows the distribution of the top 20 occurrences of STADDR values, which has an x-axis representing the count of occurrences and a y-axis representing the STADDR values. **501 SURF AVENUE** has the highest occurrence with 902 instances, indicating it is the most frequently listed address in the dataset. The next highest occurrences are 330 EAST 38 STREET with 817 instances and 322 WEST 57 STREET with 720 instances. This distribution can help identify which addresses are most common, potentially indicating areas of higher property concentration or significance in the dataset.

21) **ZIP**

Description: ZIP code of the property location, indicating the postal code for the area.

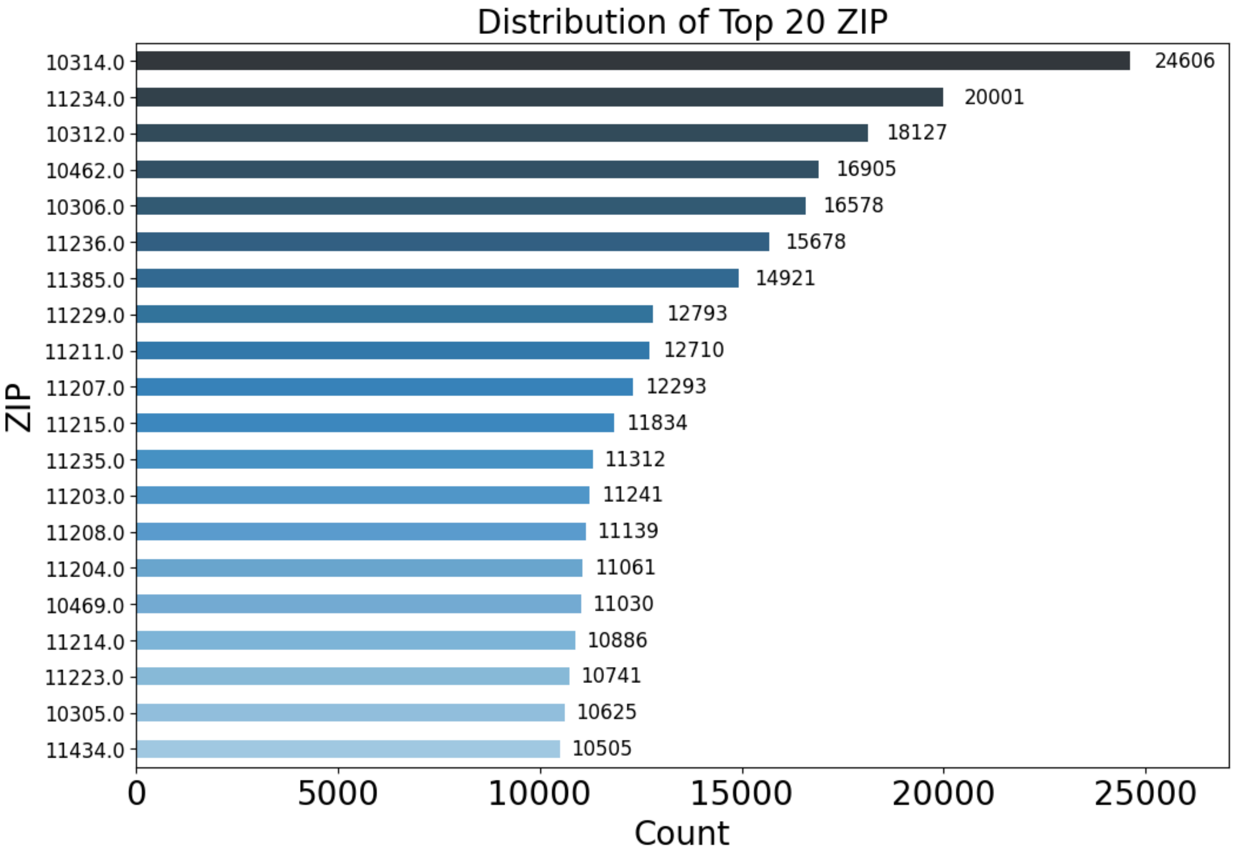


Figure 27: Distribution of the top 20 ZIP

The bar chart depicts the distribution of the top 20 ZIP codes, with the x-axis representing the count of occurrences and the y-axis representing the ZIP codes.**10314** has the highest occurrence with **24,606 instances**, making it the most frequently listed ZIP code in the dataset. The next highest occurrences are 11234 with 20,001 instances and 10312 with 18,127 instances.

22) **EXMPTCL**

Description: Exemption class, indicating if the property has any tax exemptions and the type of exemption.

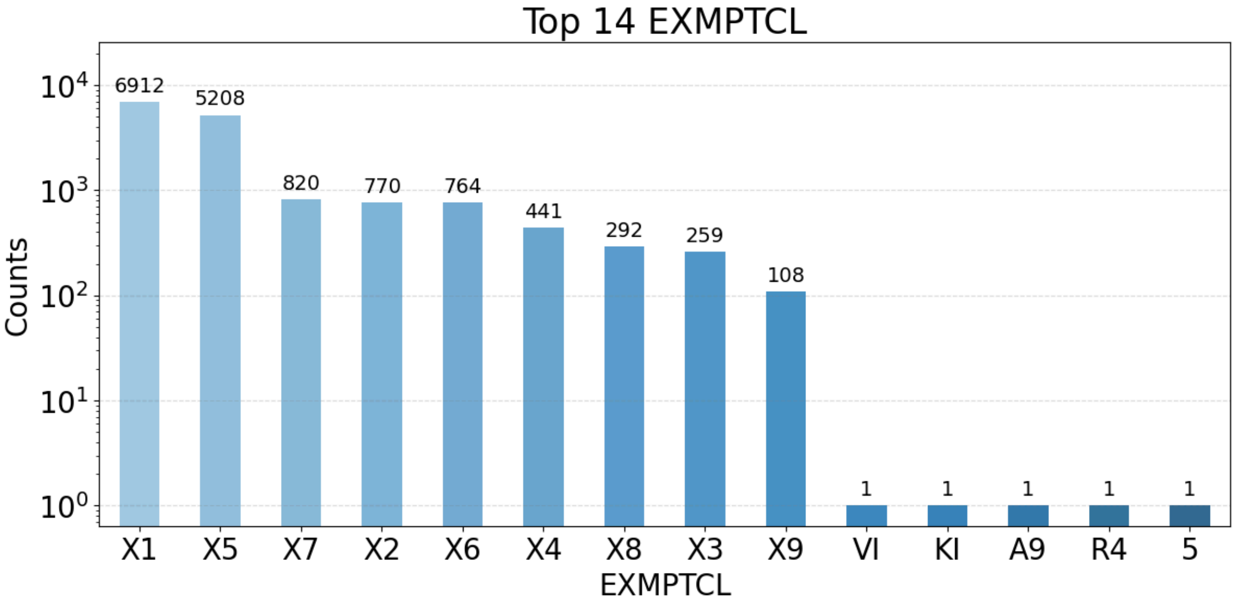


Figure 28: Distribution of top 14 EXMPTCL

Figure 28 illustrates the distribution of the top 14 EXMPTCL categories, highlighting the most and least frequent ones within the dataset, with the x-axis representing the EXMPTCL categories and the y-axis representing the count of occurrences on a logarithmic scale. X1 is the most frequent category, with 6912 counts followed by X5 with 5208 counts. These two categories dominate the dataset compared to the other categories. X4, X8, X3, and X9 have lower counts, ranging from 108 to 441. These categories are less common but still have notable occurrences.

23) **BLDFRONT**

Description: Building frontage, representing the width of the building at the front, measured in feet.

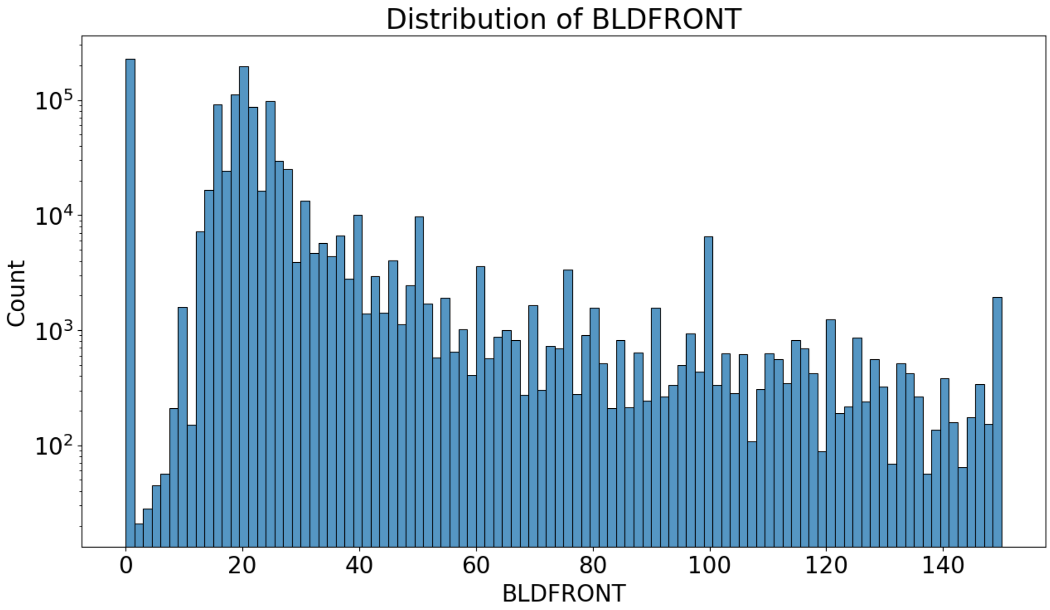


Figure 29: Distribution of BLDFRONT

The histogram displays the distribution of BLDFRONT values, with the x-axis representing the BLDFRONT values and the y-axis representing the count of occurrences on a logarithmic scale. The highest frequency is observed at very low BLDFRONT values, particularly around 0. This suggests that a significant number of buildings have very small or zero frontage. The majority of BLDFRONT values fall within the range of 10 to 50. Within this range, there are several peaks, indicating common frontage dimensions for many buildings. Values above 100 are less frequent but still present. The frequency continues to decrease as the BLDFRONT value increases, showing that large building frontages are less common.

24) **BLDDEPTH**

Description: Building depth, representing the depth of the building, measured in feet

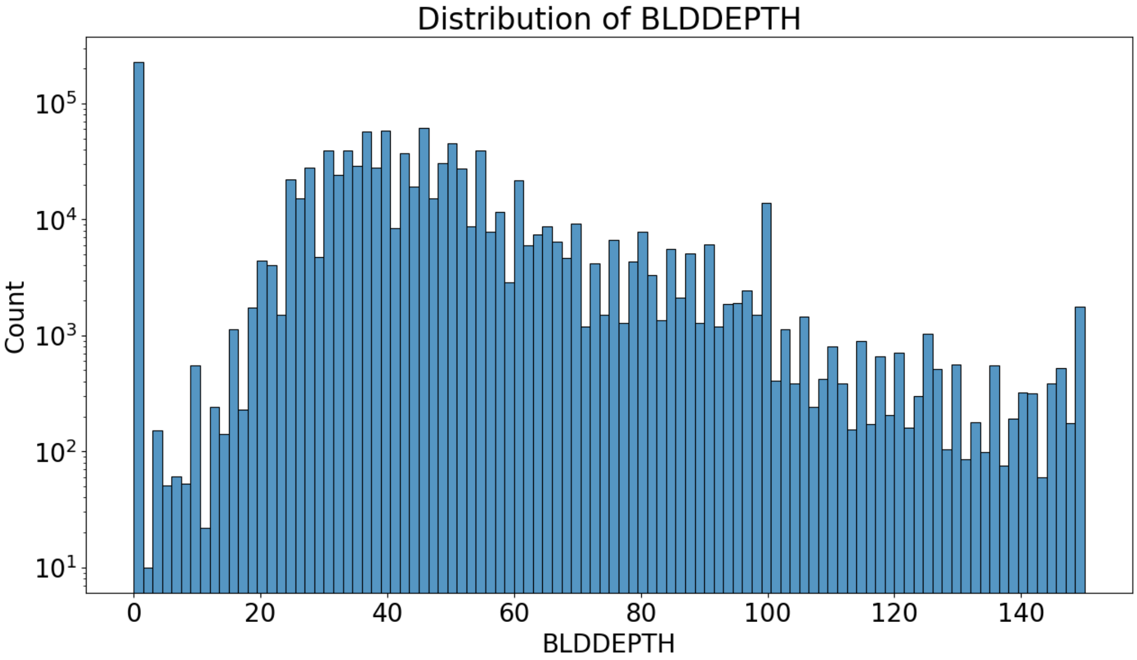


Figure 30: Distribution of BLDDEPTH

Figure 30 displays the distribution of BLDDEPTH values, highlighting the predominance of smaller depths and the decreasing frequency of larger depths. Similar to BLDFRONT, the highest frequency is observed at very low BLDDEPTH values, particularly around 0. This indicates that a significant number of buildings have very small or zero depth, The majority of BLDDEPTH values fall within the range of 20 to 60. This range shows a consistently high frequency of buildings, indicating standard building depths in this range. The distribution is right-skewed, with a long tail extending towards higher BLDDEPTH values. This indicates that while smaller depths are very common, larger depths are progressively less frequent.

25) **PERIOD**

Description: A period indicator, serving as a unique code to denote the 'Final' period with 1,070,994 counts

26) **VALTYPE**

Description: Valuation type, indicating the method or type of property valuation used. There is only 1 value of VALTYPE as AC-TR with 1,070,994 counts in this dataset

27) **YEAR**

Description: Year of the data or assessment, indicating when the data was recorded or the assessment took place. This dataset covers the timeframe of 2010/11.