Project 2 Informal Report

Cristian Barahona

1. Executive Summary

The following report provides a detailed description of the implementation of an unsupervised model for fraud detection. In particular, the business problem addressed is the detection of suspicious records on Property Valuation and Assessment Data from the city of New York. The main goal is to help identify anomalies in the property records, so they can be further investigated and determine whether these records are possibly fraud or not. The results of the model show that some of the top properties ranked by fraud score have unusual combinations of variables. When investigating in detail 5 of these properties, we were able to identify exactly what was wrong with the record, providing evidence from other sources to confirm this.

2. Data Description

The dataset is **Property Valuation and Assessment Data**, which contains real estate assessment property data in the fiscal **year 2010/11** (i.e., The Department of Finance values properties every year as one step in calculating property tax bills.). The data is a collection of **1,070,994 records** including **32 fields**.

The following tables show summarized statistics for the 14 numeric and 18 categorical fields in the dataset.

a. Numeric Fields Table

Field Name	# Records With Values	% Populated	% Zeros	Min	Max	Mean	Standard Deviation	Most Common
LTFRONT	1,070,994	100%	15.8%	0	9,999	37	74	0
LTDEPTH	1,070,994	100%	15.9%	0	9,999	89	76	100
STORIES	1,014,730	94.8%	0%	1	119	5	8	2
FULLVAL	1,070,994	100%	1.2%	0	6,150,000,000	874,265	11,582,426	0
AVLAND	1,070,994	100%	1.2%	0	2,668,500,000	85,068	4,057,258	0
AVTOT	1,070,994	100%	1.2%	0	4,668,308,947	227,238	6,877,526	0
EXLAND	1,070,994	100%	45.9%	0	2,668,500,000	36,424	3,981,574	0
EXTOT	1,070,994	100%	40.4%	0	4,668,308,947	91,187	6,508,400	0
BLDFRONT	1,070,994	100%	21.4%	0	7,575	23	36	0
BLDDEPTH	1,070,994	100%	21.4%	0	9,393	40	43	0
AVLAND2	282,726	26.4%	0%	3	2,371,005,000	246,236	6,178,952	2,408
AVTOT2	282,732	26.4%	0%	3	4,501,180,002	713,911	11,652,508	750
EXLAND2	87,449	8.2%	0%	1	2,371,005,000	351,236	10,802,151	2,090
EXTOT2	130,828	12.2%	0%	7	4,501,180,002	656,768	16,072,449	2,090

b. Categorical Fields Table

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
RECORD	1,070,994	100%	0	1,070,994	1
BBLE	1,070,994	100%	0	1,070,994	1000010101
BORO	1,070,994	100.0%	0	5	4
BLOCK	1,070,994	100%	0	13,984	3944
LOT	1,070,994	100%	0	6,366	1
EASEMENT	4,636	0%	0	12	E
OWNER	1,039,249	97%	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100%	0	200	R4
TAXCLASS	1,070,994	100%	0	11	1
EXT	354,305	33%	0	3	G
EXCD1	638,488	60%	0	129	1017
STADDR	1,070,318	100%	0	839,280	501 SURF AVENUE
ZIP	1,041,104	97%	0	196	10314
EXMPTCL	15,579	1%	0	14	X1
EXCD2	92,948	9%	0	60	1017
PERIOD	1,070,994	100%	0	1	FINAL
YEAR	1,070,994	100%	0	1	2010/11
VALTYPE	1,070,994	100%	0	1	AC-TR

3. Data Cleaning

In this and the following section, the main goals are to **clean the data** and **build the variables** for the unsupervised fraud (anomaly) detection model. To do so, we are going to first remove **exclusions**, which are records that we don't want our model to train on. Then, we will do **field imputation** to handle missing field values. Finally, we are going to **create new variables** that can be useful to solve the business problem - identifying unusual property values for the listed characteristics- and provide a **list of all the variables** with their description.

3.1. Exclusions

First, we are going to remove records that we don't want to include in our model. This includes mainly properties that are owned by the government, cemeteries, or other public properties such as parks. To do so, we do the following:

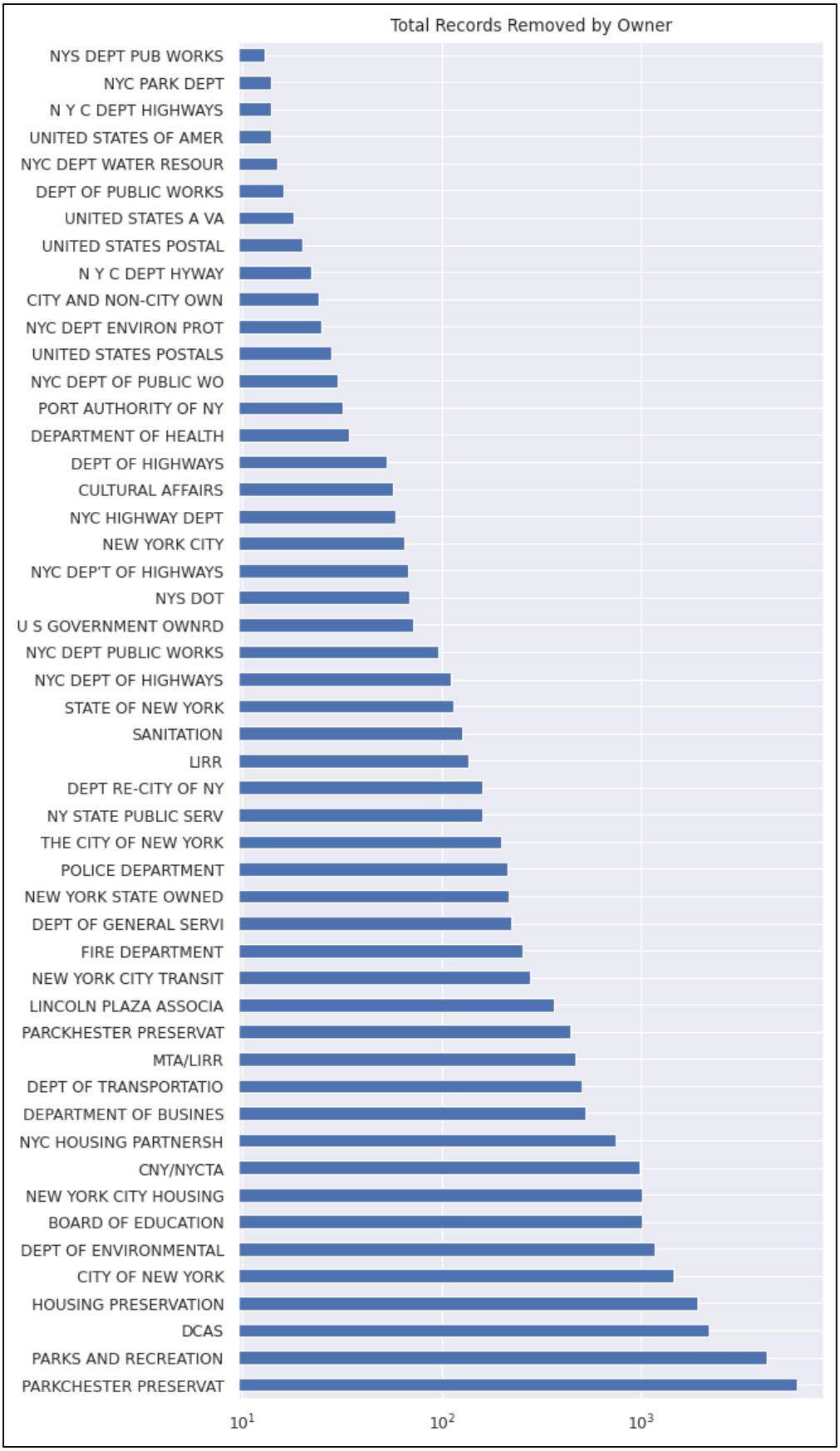
**Note that we are not removing outliers, as that is what we want to detect.*

i. Remove records with a U.S. Government **Easement** definition. (Easement = 'U')

ii. Remove all records with **Owner** name including the following words:
   - DEPT
   - DEPARTMENT
   - UNITED STATES
   - GOVERNMENT
   - GOVT
   - CEMETERY

iii. Look at the most common **Owner** names and remove all that appear to be government/public properties.

After this, a total of **26,501 records** are removed from the original dataset. The total number of records removed by **Owner** is shown in the following chart:


Total Records Removed by Owner
NYS DEPT PUB WORKS
NYC PARK DEPT
N Y C DEPT HIGHWAYS
UNITED STATES OF AMER
NYC DEPT WATER RESOUR
DEPT OF PUBLIC WORKS
UNITED STATES A VA
UNITED STATES POSTAL
N Y C DEPT HYWAY
CITY AND NON-CITY OWN
NYC DEPT ENVIRON PROT
UNITED STATES POSTALS
NYC DEPT OF PUBLIC WO
PORT AUTHORITY OF NY
DEPARTMENT OF HEALTH
DEPT OF HIGHWAYS
CULTURAL AFFAIRS
NYC HIGHWAY DEPT
NEW YORK CITY
NYC DEP'T OF HIGHWAYS
NYS DOT
U S GOVERNMENT OWNRD
NYC DEPT PUBLIC WORKS
NYC DEPT OF HIGHWAYS
STATE OF NEW YORK
SANITATION
LIRR
DEPT RE-CITY OF NY
NY STATE PUBLIC SERV
THE CITY OF NEW YORK
POLICE DEPARTMENT
NEW YORK STATE OWNED
DEPT OF GENERAL SERVI
FIRE DEPARTMENT
NEW YORK CITY TRANSIT
LINCOLN PLAZA ASSOCIA
PARCKHESTER PRESERVAT
MTA/LIRR
DEPT OF TRANSPORTATIO
DEPARTMENT OF BUSINES
NYC HOUSING PARTNERSH
CNY/NYCTA
NEW YORK CITY HOUSING
BOARD OF EDUCATION
DEPT OF ENVIRONMENTAL
CITY OF NEW YORK
HOUSING PRESERVATION
DCAS
PARKS AND RECREATION
PARKCHESTER PRESERVAT
10^1
10^2
10^3

3.2. Field Imputation

a. ZIP

There is a total of **20,431** records with the value of the field ZIP missing. To fill these missing values,

1. For each Street Address (STADDR) + Borough (BORO) combination, the corresponding Zip code was mapped based on the existing data. Then, the missing Zip codes that have a matching combination were mapped. This filled **2,832 values.**
2. Assuming the data is sorted by Zip, missing Zip records that have a matching Zip value in both the previous and next records were imputed that value. This step filled **9,491 values**.
3. For the remainder of the missing records, the value of the previous record was imputed. This filled the **8,108 records** remaining.

b. FULLVAL, AVLAND, AVTOT

The fields FULLVAL, AVLAND, and AVTOT follow the same imputation logic:

**Note that we are we are considering zero values as missing fields for these 3 fields*.

1. We group all records by the combination of **TAXCLASS**, **BORO** and **BLDCL**, calculate the **average of the imputed field** for each group, and replace the missing value in that group with the average of the group.
2. We then group all records by the combination of **TAXCLASS** and **BORO**, calculate the **average of the imputed field** for each group, and replace the missing value in that group with the average of the group.
3. We finally group all records by **TAXCLASS**, calculate the **average of the imputed field** for each group, and replace the missing value in that group with the average of the group.

For all the remaining fields with missing values, we follow a similar imputation process. The following table summarizes the process for all imputed fields:

Imputed Field	Grouped By	# Records Imputed
FULLVAL	TAXCLASS, BORO, BLDCL	2,718
	TAXCLASS, BORO	6,921
	TAXCLASS	386
AVLAND	TAXCLASS, BORO, BLDCL	2,720
	TAXCLASS, BORO	6,921
	TAXCLASS	386
AVTOT	TAXCLASS, BORO, BLDCL	2,718
	TAXCLASS, BORO	6,921
	TAXCLASS	386
STORIES	BORO, BLDCL	4,108
	TAXCLASS	37,922
LTFRONT	TAXCLASS BORO	160,563
	TAXCLASS	2
LTDEPTH	BORO, BLDCL	161,654
	TAXCLASS	2

4. Variable Creation

4.1. Variable Creation Logic

We are going to create variables that can be helpful to identify unusual property values. To do so, we are going to use some basic principles:

- Bigger properties (building & lot size) are more expensive. Therefore, price per square foot (building & lot) is a good, standardized measure.
- Location has a large influence on property value. Zip code is a good indicator of location.
- Property type (TAXCLASS) can also be an important factor.

With these principles in mind, we first create some size variables as follows:

- $Lot\ Area \rightarrow S_1 = LTFRONT * LTDEPTH$
- $Building\ Area \rightarrow S_2 = BLDFRONT * BLDDEPTH$
- $Building\ Volume \rightarrow S_3 = S_2 * STORIES$

Now, we consider the property value variables:

- $V_1 = FULLVAL$
- $V_2 = AVLAND$
- $V_3 = AVTOT$

And create **ratios** that combine any **value variable** with any **size variable:**

$$r_1 = \frac{V_1}{S_1} \qquad r_4 = \frac{V_2}{S_1} \qquad r_7 = \frac{V_3}{S_1}$$

$$r_2 = \frac{V_1}{S_2} \qquad r_5 = \frac{V_2}{S_2} \qquad r_8 = \frac{V_3}{S_2}$$

$$r_3 = \frac{V_1}{S_3} \qquad r_6 = \frac{V_2}{S_3} \qquad r_9 = \frac{V_3}{S_3}$$

We also consider the **inverse ratios** $\frac{1}{r_i}$, which gives us a total of **18 new variables**.

Finally, we also want to consider location and property type, so we group each of the 18 new variables into these 2 groups, and calculate $< r_i >_g$, the average of each ratio r_i for each group g.

Then, we calculate the additional **36 variables** as follows:

$$\frac{r_1}{< r_1 >_g}, \quad \frac{r_2}{< r_2 >_g}, \quad \frac{r_3}{< r_3 >_g}, \quad \cdots \quad \frac{r_9}{< r_9 >_g} \qquad g = 1, 2$$

Which results in **54 total** created variables.

Another possibly useful variables are **value and size ratios,** calculated as follows:

- $Value\ Ratio\ \rightarrow VR\ = \frac{FULLVAL}{(AVLAND+AVTOT)}$
- $Size\ Ratio\ \rightarrow SR\ = \frac{Building\ Size}{Lot\ Size} = \frac{BLDFRONT \times BLDEPTH}{LTFRONT \times LTDEPTH}$

*Note that for value ratio, we normalize the values and then take the max between the VR and its inverse, to include both unusually large & small ratios:

$$Value\ Ratio = \max\left(VR_{norm}, \frac{1}{VR_{norm}}\right)$$

4.2. Variables List

The following list shows all the created variables with their corresponding algebraic definitions, which were previously described.

Variable	Definition
r_1	V_1/S_1
r_2	V_1/S_2
r_3	V_1/S_3
r_4	V_2/S_1
r_5	V_2/S_2
r_6	V_2/S_3
r_7	V_3/S_1
r_8	V_3/S_2
r_9	V_3/S_3
r_1 inv	S_1/V_1
r_2 inv	S_1/V_2
r_3 inv	S_1/V_3
r_4 inv	S_2/V_1
r_5 inv	S_2/V_2
r_6 inv	S_2/V_3
r_7 inv	S_3/V_1
r_8 inv	S_3/V_2
r_9 inv	S_3/V_3
r_1 zip	$r_1/\langle r_1 \rangle_{zip}$
r_2 zip	$r_2/\langle r_2 \rangle_{zip}$
r_3 zip	$r_3/\langle r_3 \rangle_{zip}$
r_4 zip	$r_4/\langle r_4 \rangle_{zip}$
r_5 zip	$r_5/\langle r_5 \rangle_{zip}$
r_6 zip	$r_6/\langle r_6 \rangle_{zip}$
r_7 zip	$r_7/\langle r_7 \rangle_{zip}$

r_8 zip	$r_8/\langle r_8\rangle_{zip}$
r_9 zip	$r_9/\langle r_9\rangle_{zip}$
r_1 inv zip	$\frac{1}{r_1}\Big/\langle\frac{1}{r_1}\rangle_{zip}$
r_2 inv zip	$\frac{1}{r_2}\Big/\langle\frac{1}{r_2}\rangle_{zip}$
r_3 inv zip	$\frac{1}{r_3}\Big/\langle\frac{1}{r_3}\rangle_{zip}$
r_4 inv zip	$\frac{1}{r_4}\Big/\langle\frac{1}{r_4}\rangle_{zip}$
r_5 inv zip	$\frac{1}{r_5}\Big/\langle\frac{1}{r_5}\rangle_{zip}$
r_6 inv zip	$\frac{1}{r_6}\Big/\langle\frac{1}{r_6}\rangle_{zip}$
r_7 inv zip	$\frac{1}{r_7}\Big/\langle\frac{1}{r_7}\rangle_{zip}$
r_8 inv zip	$\frac{1}{r_8}\Big/\langle\frac{1}{r_8}\rangle_{zip}$
r_9 inv zip	$\frac{1}{r_9}\Big/\langle\frac{1}{r_9}\rangle_{zip}$
r_1 taxclass	$r_1/\langle r_1\rangle_{taxclass}$
r_2 taxclass	$r_2/\langle r_2\rangle_{taxclass}$
r_3 taxclass	$r_3/\langle r_3\rangle_{taxclass}$
r_4 taxclass	$r_4/\langle r_4\rangle_{taxclass}$
r_5 taxclass	$r_5/\langle r_5\rangle_{taxclass}$
r_6 taxclass	$r_6/\langle r_6\rangle_{taxclass}$
r_7 taxclass	$r_7/\langle r_7\rangle_{taxclass}$
r_8 taxclass	$r_8/\langle r_8\rangle_{taxclass}$
r_9 taxclass	$r_9/\langle r_9\rangle_{taxclass}$
r_1 inv taxclass	$\frac{1}{r_1}\Big/\langle\frac{1}{r_1}\rangle_{taxclass}$
r_2 inv taxclass	$\frac{1}{r_2}\Big/\langle\frac{1}{r_2}\rangle_{taxclass}$
r_3 inv taxclass	$\frac{1}{r_3}\Big/\langle\frac{1}{r_3}\rangle_{taxclass}$
r_4 inv taxclass	$\frac{1}{r_4}\Big/\langle\frac{1}{r_4}\rangle_{taxclass}$
r_5 inv taxclass	$\frac{1}{r_5}\Big/\langle\frac{1}{r_5}\rangle_{taxclass}$
r_6 inv taxclass	$\frac{1}{r_6}\Big/\langle\frac{1}{r_6}\rangle_{taxclass}$

r_7 inv taxclass	$\frac{1}{r_7} \Big/ \langle \frac{1}{r_7} \rangle_{taxclass}$
r_8 inv taxclass	$\frac{1}{r_8} \Big/ \langle \frac{1}{r_8} \rangle_{taxclass}$
r_9 inv taxclass	$\frac{1}{r_9} \Big/ \langle \frac{1}{r_9} \rangle_{taxclass}$
VR	$\max\left(VR_{norm}, \frac{1}{VR_{norm}}\right)$
SR	$\frac{BLDFRONT \times BLDEPTH}{LTFRONT \times LTDEPTH}$

5. Dimensionality Reduction

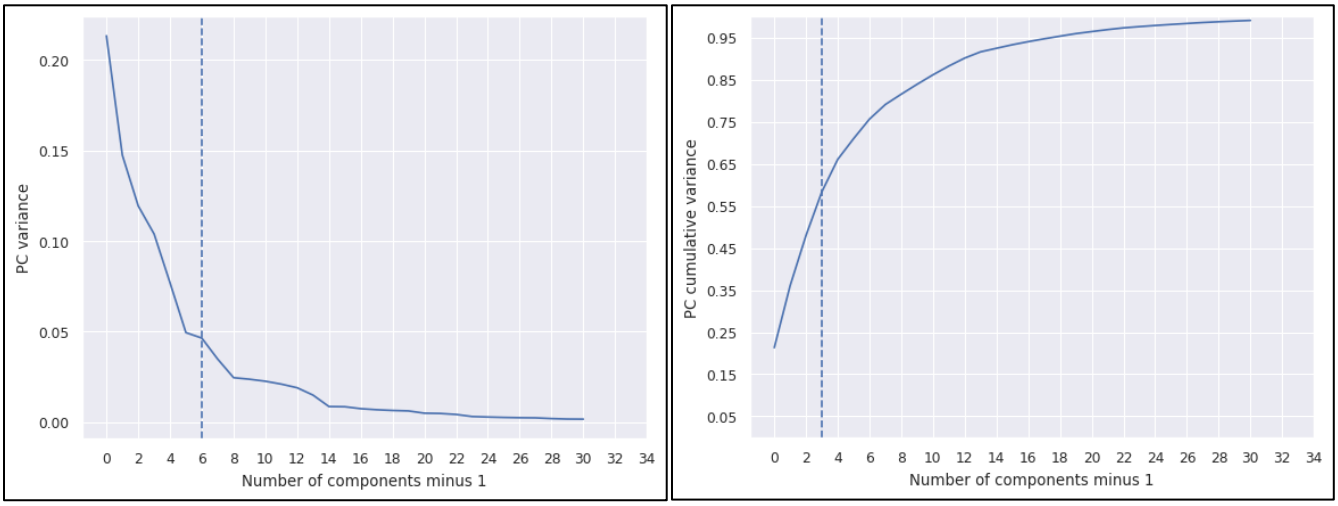
Once the variables are built, the next step is to **reduce dimensionality**. To do so, we're going to use **Principal Component Analysis (PCA),** which finds the dominant directions in the data and rotates the coordinate system along these directions. After the rotation, each formed **Principal Component (PC)** is a linear combination of all the original variables. Then, the PCA algorithm will rank the different PCs to see which ones account for the highest variance. This ranking where the PCs are sorted by magnitude is shown in the **scree plot.** Usually, we want to keep the first few PCs, trying to account for ~80 – 90% of the total variance. We select these Principal Components and discard the rest, reducing the dimensions to just a few. Another important effect of **PCA** is that it **removes correlation** by combining correlated variables into a single direction.

For the dimensionality reduction process, we will do the following:

1. Z scale variables.
2. Conduct PCA and select PCs.
3. Z scale transformed variables (PCs)

It is important to always perform **z scaling** before doing PCA, as the PCA algorithm looks for covariance which might be sensitive to different scaling in the variables. Z scaling is useful as it centers and scales the data. Furthermore, we will do another z scale after the PCA to make all the dimensions equally important for the **Minkowski distance**.

When performing PCA, we sort the PCs by the variance magnitude they account for. We keep a low number of PCs that account for most of the (cumulative) variance. The following scree and cumulative variance plots illustrate the variance explained by the principal components:



We observe from the cumulative variance plot that 4 principal components account for approximately 60% of the total variance. Since our main goal is to reduce dimensionality, we will select as few PCs as possible while maintaining a high level of variance explained. In this scenario, we'll select 4 PCs and perform another z scaling on these.

6. Anomaly Detection Algorithms

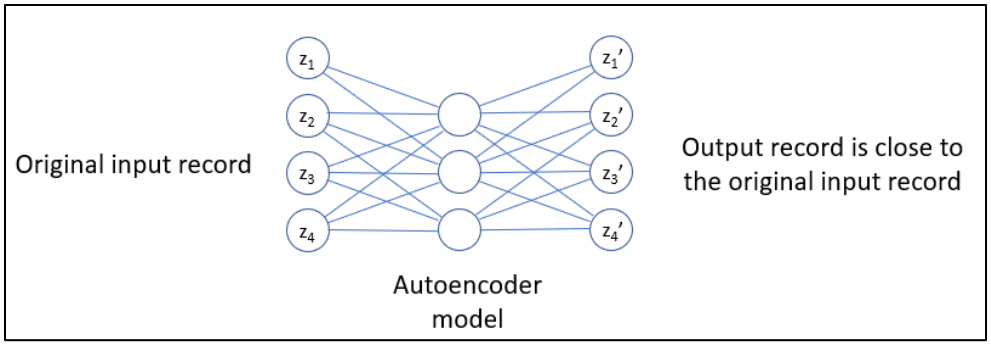
In this section, we will describe **two score algorithms** that will be used for anomaly detection. Both algorithms seek to detect outliers in the dataset.

The first method is to as fraud score the **Minkowski distance**, which will calculate the distance to the origin of each point, which will tell us explicitly the 'outlierness' in each of the variables. This score is calculated as follows:

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

where s_i represents the score for record i, and reasonable choices for power (p) are anywhere from 1 to 4.

The second method to score outliers is by using an **autoencoder**, a model trained to output the original vector input. This model is usually a neural network.



After the model is trained, the difference (error) between the original input vector and the model output vector is the fraud score for that record.

To implement an autoencoder we do the following:

- First get the data well prepared (z scale, PCA, reduce dimensions, z scale again). We do this the same way for both Methods 1 and 2.
- Train an autoencoder on the entire data set. The model will learn to reproduce the data records as well as possible, and will learn the nature of the bulk of the data.
- The records that aren't reproduced well are unusual records, which is what we're looking for.
- Therefore a measure of the reproduction error is a measure of unusualness for that record, and is thus a fraud score:

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

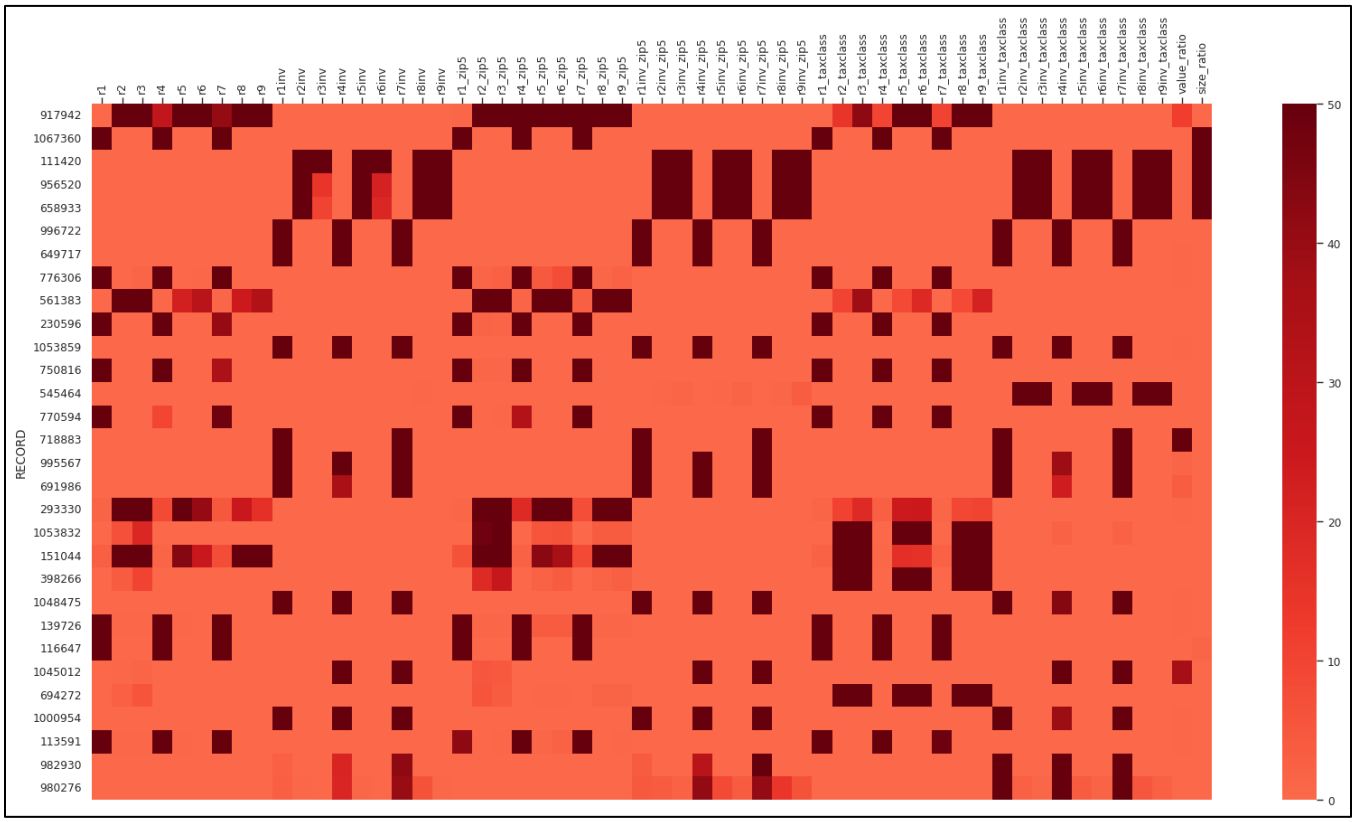
For this project, we will use **both fraud scores** by combining them using weighted average rank orders to get the **final score**.

7. Results

After obtaining a final score for all variables, we will first **sort all the records** based on the final score ranking. Then, we **select the top n** records. In this project, we will select the **top 1,000 records** based on their fraud score. When looking at our final output, we want to make sure that the original variables (not PCs) are present, and we have these variables **z scaled,** as we will look at unusually high or low values that explain the fraud score.

After selecting and scaling the top 1,000 records by fraud score, we export these records into an Excel spreadsheet, where we can easily analyze properties and sort/filter by relevant variables. Since we are looking for **unusual properties' ratios**, a heat map can also provide a good visual representation of which variables are driving the high scores:

**Note that both high and low values can be 'unusual'. To detect low values, we use the inverse ratios (r inv)*

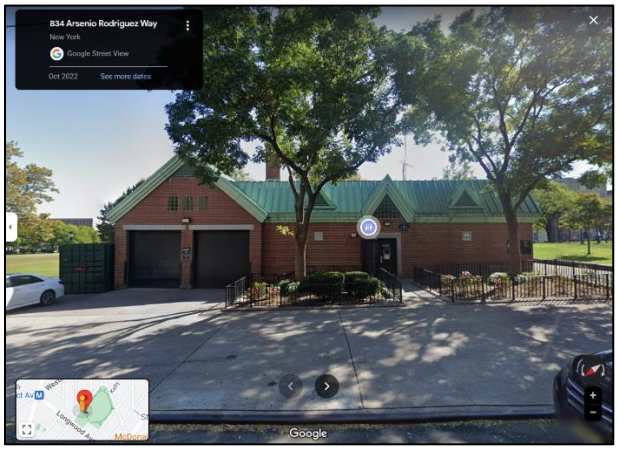


7.1. Property Investigation

Finally, we are going to use our export to **investigate NY properties** that were ranked high in our Fraud Detection model. We will select **5 properties** that have **'strange' records** and describe what is wrong with the record, along with supporting information/evidence to support it.

1. 810 DAWSON STREET

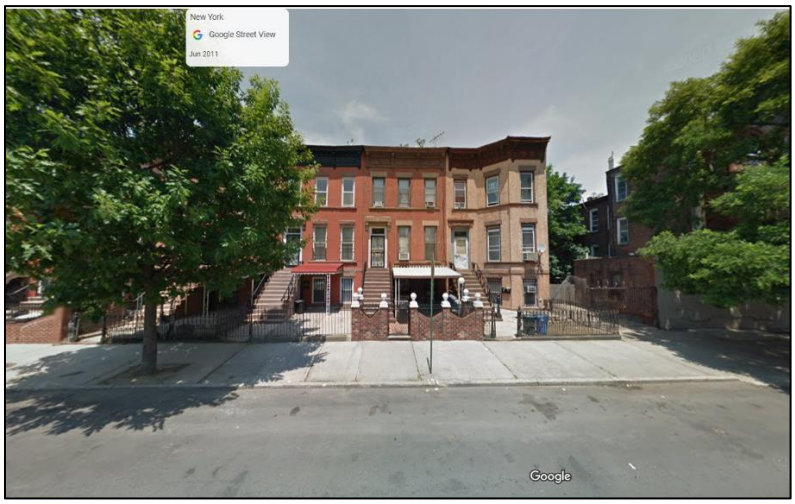
LOCATION		SIZE (FT.)		VALUE (USD)		
RECORD N°:	155893	LTFRONT:	4	FULLVAL:	$	3,080,000
OWNER:	ATTRACTIVE HOME, INC.	LTDEPTH:	31	AVLAND:	$	1,075,500
ADDRESS:	810 DAWSON STREET	BLDFRONT:	73	AVTOT:	$	1,386,000
ZIP:	10459	BLDDEPTH:	31			
UNUSUAL RATIOS:						
r1 = 23, r4 = 84, r7 = 60, r1_zip = 38, r2_zip = 158, r7_zip = 85						

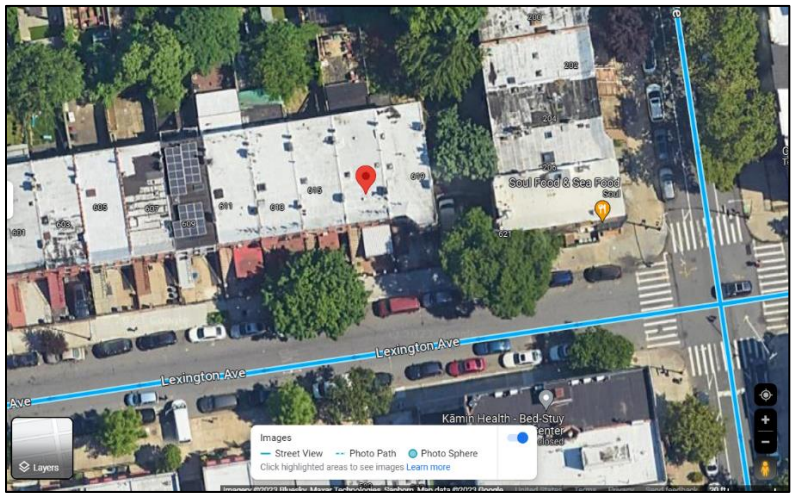


- High r1, r4, r7 values indicate that property has an **unusually low Lot Size (S1)**.
- These scores are even more unusual when taking into consideration location (zip).
- When looking at property information, we see that BLDFRONT is higher than LTFRONT, and LTFRONT is suspiciously low (4 ft.)
- This is confirmed by looking at images of the property, where we see that the Lot front is clearly larger than its listed value.

2. 617 LEXINGTON AVENUE

LOCATION		SIZE (FT.)		VALUE (USD)		
RECORD N°:	89347	LTFRONT:	200	FULLVAL:	$	429,000,000
OWNER:	DAI-ICHI LIFE INVESTM	LTDEPTH:	325	AVLAND:	$	38,250,000
ADDRESS:	617 LEXINGTON AVENUE	BLDFRONT:	0	AVTOT:	$	193,050,000
ZIP:	10022	BLDDEPTH:	0			
		STORIES:	57			
UNUSUAL RATIOS:						
r2 = 147, r8 = 41						

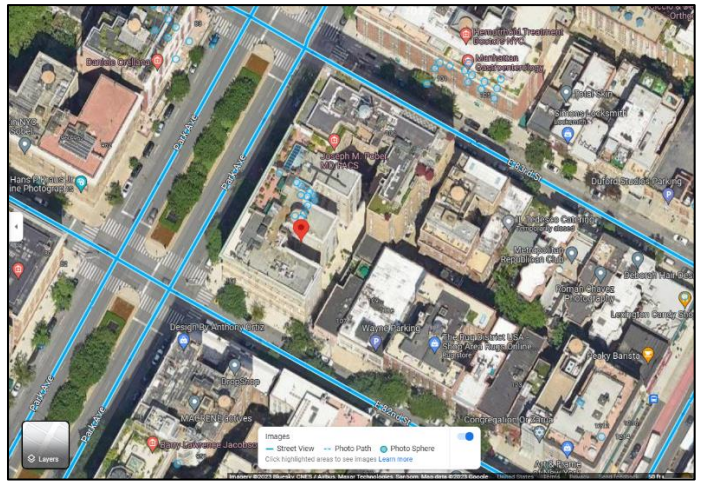




- High r2 and r8 values indicate that property has **unusually low Building Size (S2)**.
- When looking at property information, we see that both **BLDFRONT** and **BLDDEPTH** are **zero**.
- When looking at the property images, we observe that the building size is **higher than listed** (zero), and we also observe that the number of **stories** listed is incorrect (57).

3. 969 PARK AVENUE

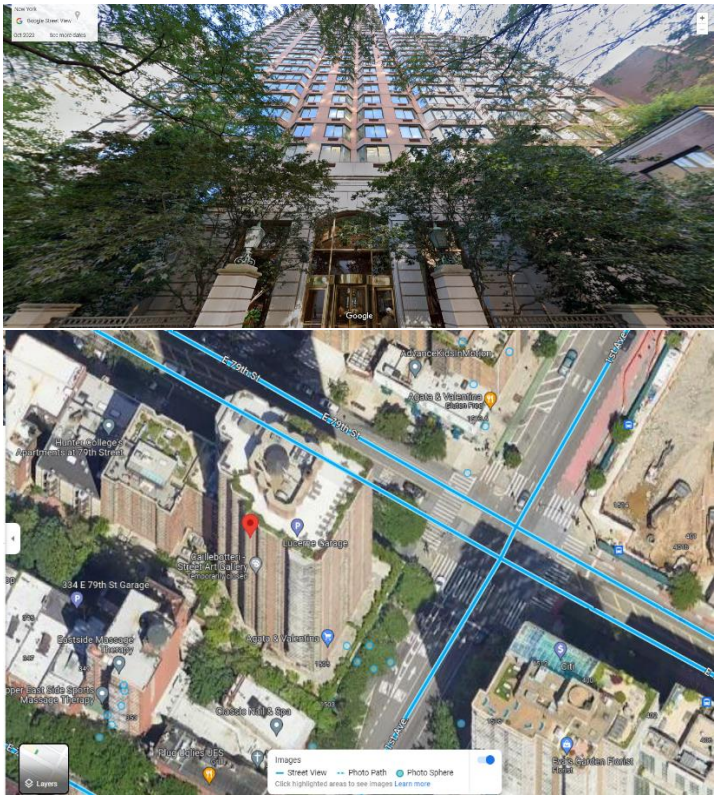
LOCATION		SIZE (FT.)		VALUE (USD)		
RECORD N°:	111426	LTFRONT:	175	FULLVAL:	$	20,400,000
OWNER:	969 PARK CORP	LTDEPTH:	193	AVLAND:	$	4,590,000
ADDRESS:	969 PARK AVENUE	BLDFRONT:	7,538	AVTOT:	$	9,180,000
ZIP:	10028	BLDDEPTH:	9,388			
		STORIES:	12			
UNUSUAL RATIOS:						
size_ratio = 171						



- Extremely high size ratio indicates that the property has an unusually high **building to lot size ratio**. This could mean that either the building size is too high, or lot size is too low (or both).
- When looking at property information, we see that both **BLDFRONT** and **BLDDEPTH** are extremely high (>7,000 ft), and higher than **LTFRONT** and **LTDEPTH** respectively.
- When looking at the property images, we corroborate that the building size is not as large as listed.

4. 350 EAST 79 STREET

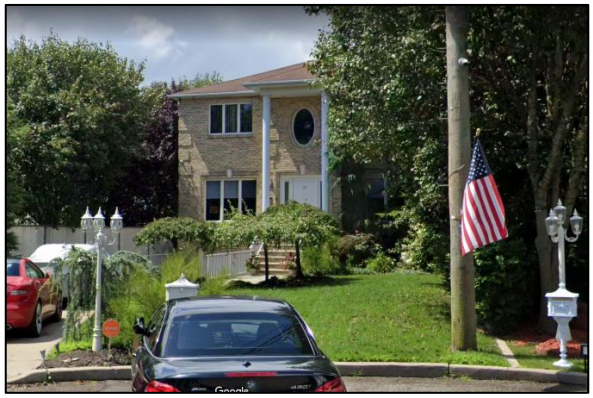
LOCATION		SIZE (FT.)		VALUE (USD)		
RECORD N°:	106681	LTFRONT:	25	FULLVAL:	$	114,000,000
OWNER:	79TH REALTY LLC	LTDEPTH:	100	AVLAND:	$	33,750,000
ADDRESS:	350 EAST 79 STREET	BLDFRONT:	25	AVTOT:	$	51,300,000
ZIP:	10075	BLDDEPTH:	100			
		STORIES:	44			
UNUSUAL RATIOS:						
r1 = 42, r4 = 131, r7 = 111						

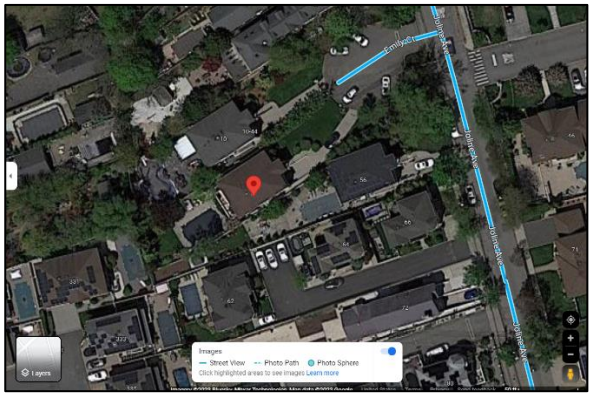


- High r1, r4, r7 values suggest that property has listed an **unusually low Lot Size (S1)**.
- When looking at property information, we see that **BLDFRONT** seems to be **too low** (25 ft.)
- This is confirmed by looking at images of the property, where we see that the building front is clearly larger than its listed value.

5. 20 EMILY COURT

LOCATION		SIZE (FT.)		VALUE (USD)		
RECORD N°:	1067360	LTFRONT:	1	FULLVAL:	$	836,000
OWNER:	N/A	LTDEPTH:	1	AVLAND:	$	28,800
ADDRESS:	20 EMILY COURT	BLDFRONT:	36	AVTOT:	$	50,160
ZIP:	10307	BLDDEPTH:	45			
		STORIES:	2			
UNUSUAL RATIOS:						
r1 = 777, r4 = 280, r7 = 270, r1_zip = 632, r2_zip = 416, r7_zip = 492, r1_taxclass = 602, r2_taxclass = 687, r7_taxclass = 731						





- All the unusual ratios suggest that the property has a listed lot size (S1) smaller than its actual value.
- When looking at property information, we see that both **LTFRONT** and **LTDEPTH** are listed as **1 ft,** even when the building size is larger than that.
- When looking at the property images, we corroborate that the property is a residential house, with **a lot size higher than listed**.

8. Summary

This report offers valuable insights into the effectiveness of implementing an unsupervised model in for property fraud detection.

In the first section, a description of the dataset used is provided, including summary tables for both categorical and numerical fields. The raw dataset included a total of 1,070,994 records across 32 fields. Since the original dataset included some missing values and records that we didn't want our model to train on, the data cleaning process handled both the exclusion of these records (government properties, cemeteries, etc.) and the imputation of missing values for the **ZIP**, **FULLVAL**, **AVLAND**, **AVTOT**, **STORIES**, **LTFRONT**, and **LTDEPTH** fields.

After cleaning the data, we built up variables that were helpful to identify unusual property records. Here, we used some basic principles of property valuation to create some **size (Si)** and **value (Vi)** variables. We used these size and value variables to create ratios, and then also included the inverse ratios and ratios grouped by **TAXCLASS** and **ZIP**. We also included two extra variables to for size and value ratios. A total of 56 variables were created in this step.

In the next section, we used a combination of scaling and PCA to apply dimensionality reduction. We selected the 4 principal components that explained the most variation in the data. After PCA, the 4 dimensions were again z scaled to center and standardize, which prepared the data for scoring calculation.

Once the principal components were selected and scaled, two different anomaly detection algorithms were used for fraud scoring. The first model used the **Minkowski** distance, which explicitly showed the 'outlierness' in each of the variables. The second model calculated fraud score using an **autoencoder,** which was trained to output the original input vector. Here, we used a measure of the reproduction error as fraud score, which means that the records that can't be reproduced accurately are possible outliers. We calculated a weighted average of both fraud scores (equally weighted), to get the final score.

Finally, we used the fraud scores from our model to rank the property records. We the investigated the top ranked records to understand what made them unusual. In particular, we selected 5 strange property records from the 1,000 highest fraud scores, and investigated these properties, providing evidence that confirmed the insights obtained from our model.

9. Annex: Data Quality Report

Data Quality Report

1. Data Description

The dataset is **Property Valuation and Assessment Data**, which contains real estate assessment property data in the **fiscal year 2010/11** (i.e., The Department of Finance values properties every year as one step in calculating property tax bills.). The data is a collection of **1,070,994** records including **32 fields**.

2. Summary Tables

Numeric Fields Table

Field Name	# Records With Values	% Populated	% Zeros	Min	Max	Mean	Standard Deviation	Most Common
LTFRONT	1,070,994	100%	15.8%	0	9,999	37	74	0
LTDEPTH	1,070,994	100%	15.9%	0	9,999	89	76	100
STORIES	1,014,730	94.8%	0%	1	119	5	8	2
FULLVAL	1,070,994	100%	1.2%	0	6,150,000,000	874,265	11,582,426	0
AVLAND	1,070,994	100%	1.2%	0	2,668,500,000	85,068	4,057,258	0
AVTOT	1,070,994	100%	1.2%	0	4,668,308,947	227,238	6,877,526	0
EXLAND	1,070,994	100%	45.9%	0	2,668,500,000	36,424	3,981,574	0
EXTOT	1,070,994	100%	40.4%	0	4,668,308,947	91,187	6,508,400	0
BLDFRONT	1,070,994	100%	21.4%	0	7,575	23	36	0
BLDDEPTH	1,070,994	100%	21.4%	0	9,393	40	43	0
AVLAND2	282,726	26.4%	0%	3	2,371,005,000	246,236	6,178,952	2,408
AVTOT2	282,732	26.4%	0%	3	4,501,180,002	713,911	11,652,508	750
EXLAND2	87,449	8.2%	0%	1	2,371,005,000	351,236	10,802,151	2,090
EXTOT2	130,828	12.2%	0%	7	4,501,180,002	656,768	16,072,449	2,090

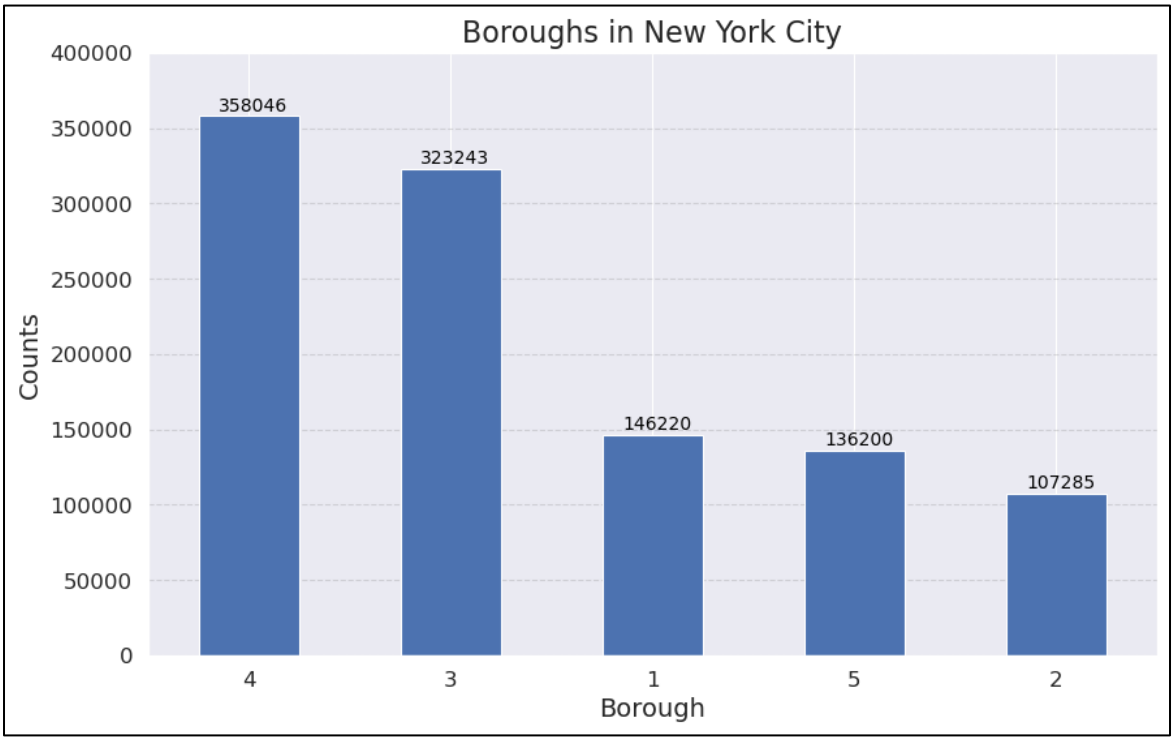
Categorical Fields Table

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
RECORD	1,070,994	100%	0	1,070,994	1
BBLE	1,070,994	100%	0	1,070,994	1000010101
BORO	1,070,994	100.0%	0	5	4
BLOCK	1,070,994	100%	0	13,984	3944
LOT	1,070,994	100%	0	6,366	1
EASEMENT	4,636	0%	0	12	E
OWNER	1,039,249	97%	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100%	0	200	R4
TAXCLASS	1,070,994	100%	0	11	1
EXT	354,305	33%	0	3	G
EXCD1	638,488	60%	0	129	1017
STADDR	1,070,318	100%	0	839,280	501 SURF AVENUE
ZIP	1,041,104	97%	0	196	10314
EXMPTCL	15,579	1%	0	14	X1
EXCD2	92,948	9%	0	60	1017
PERIOD	1,070,994	100%	0	1	FINAL
YEAR	1,070,994	100%	0	1	2010/11
VALTYPE	1,070,994	100%	0	1	AC-TR

3. Visualization of Each Field

1) **Field Name: RECORD**

Description: Ordinal unique positive integer for each property record, from 1 to 1,070,994.



2) **Field Name: BBLE**

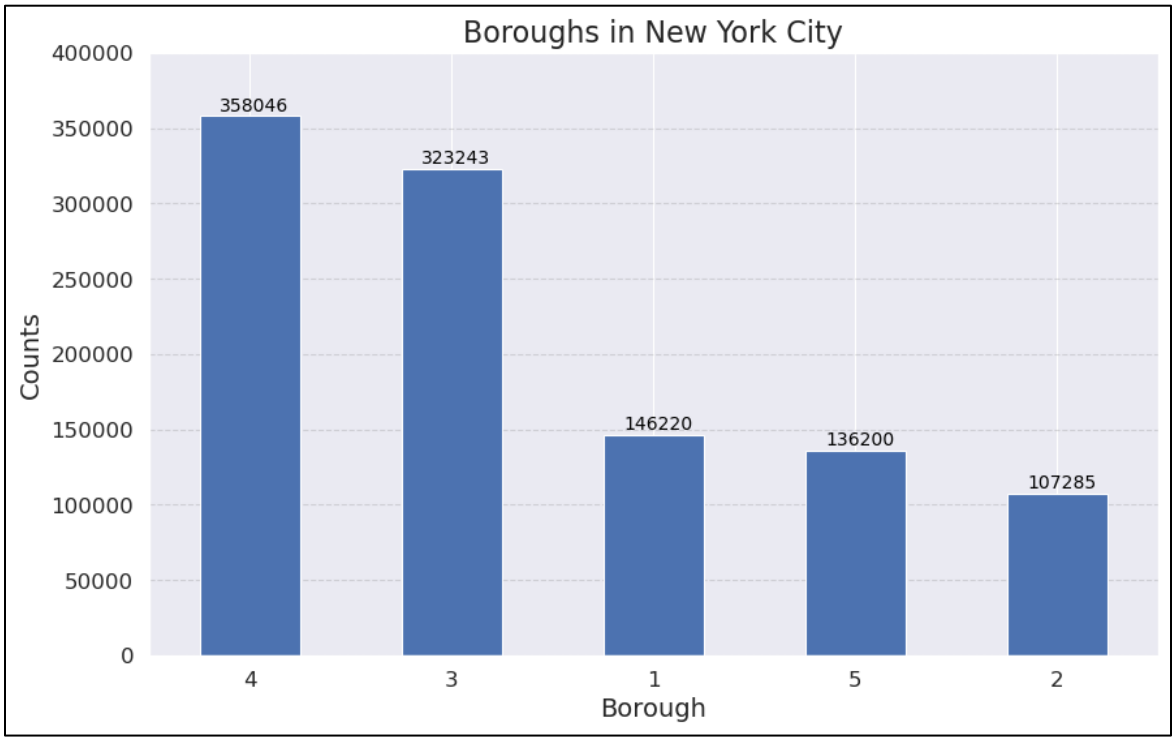
Description: Unique key value integer for each property record, formed from the fields BORO, Block, Lot, and Easement Code.

3) **Field Name: BORO**

Description: The property's Borough. Encoded as follows:

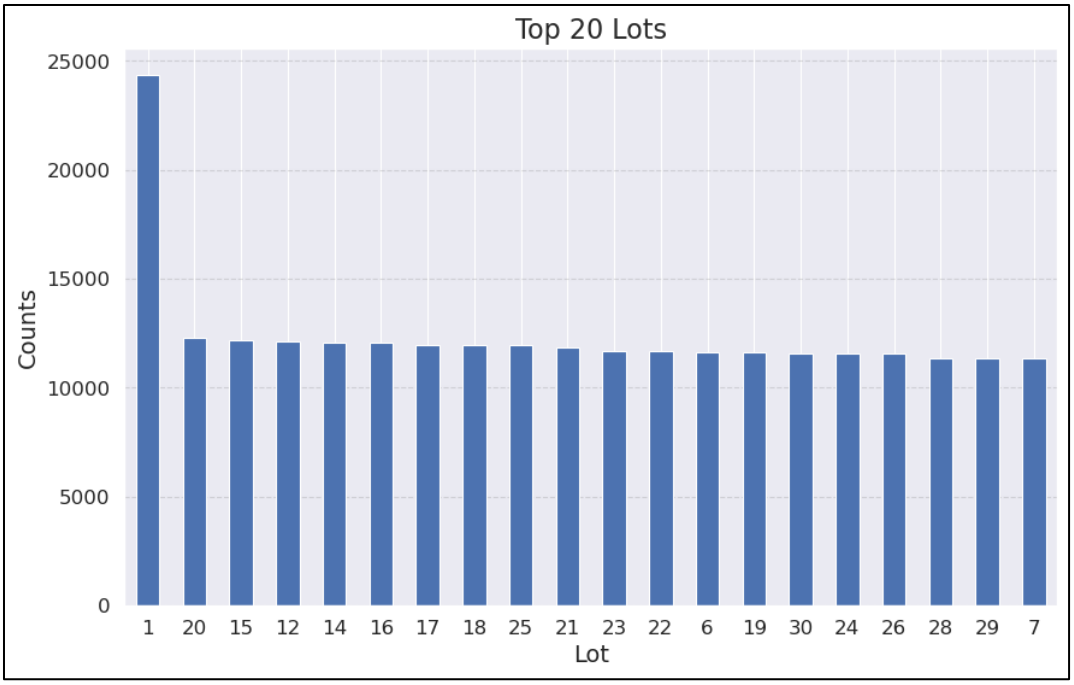
1 = Manhattan
2 = Bronx
3 = Brooklyn
4 = Queens
5 = Staten Island

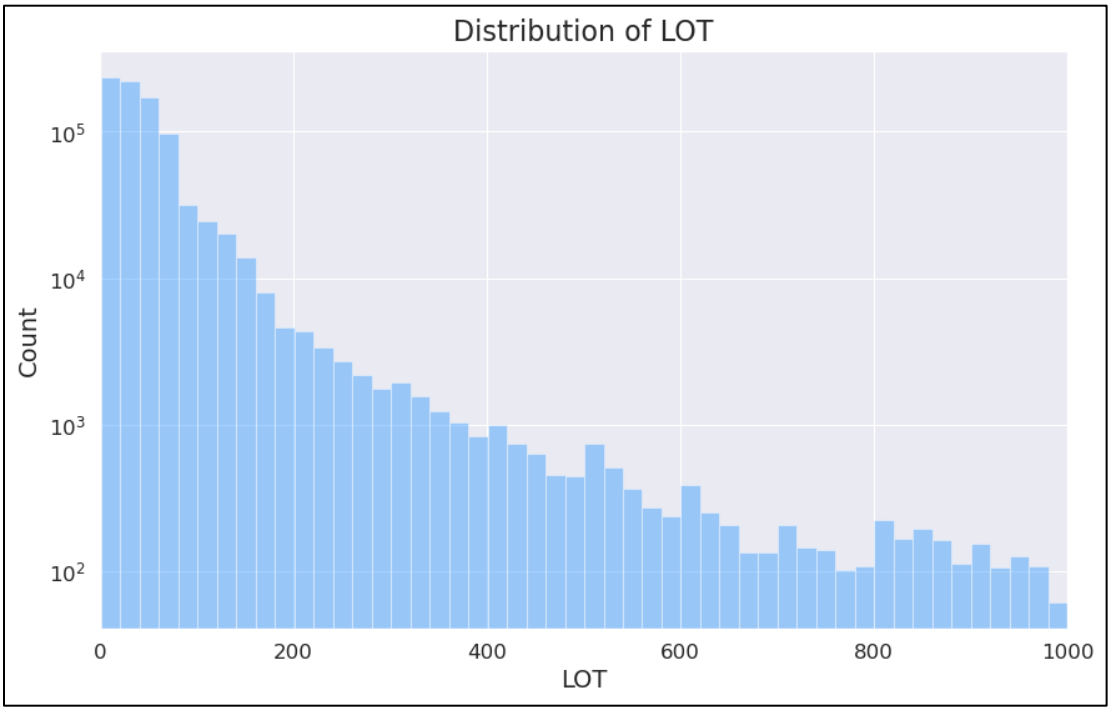
The histogram shows the distribution of property boroughs. The most common value is 4 (‘Queens’), with a total count of 358,046.



4) Field Name: LOT

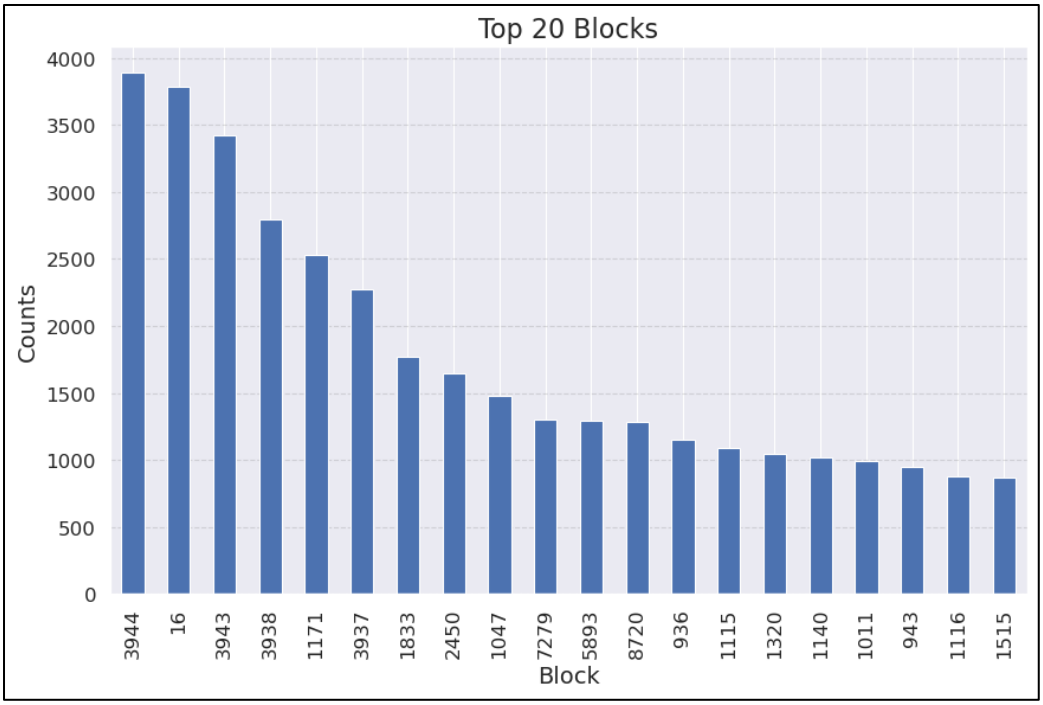
Description: Property lot number. There are a total of 6,366 unique lot numbers. The first chart shows the 20 most common values, and the second chart shows the distribution in logarithmic scale.
The most common lot number is ‘1’, with a total count of 24,367.





5) Field Name: BLOCK

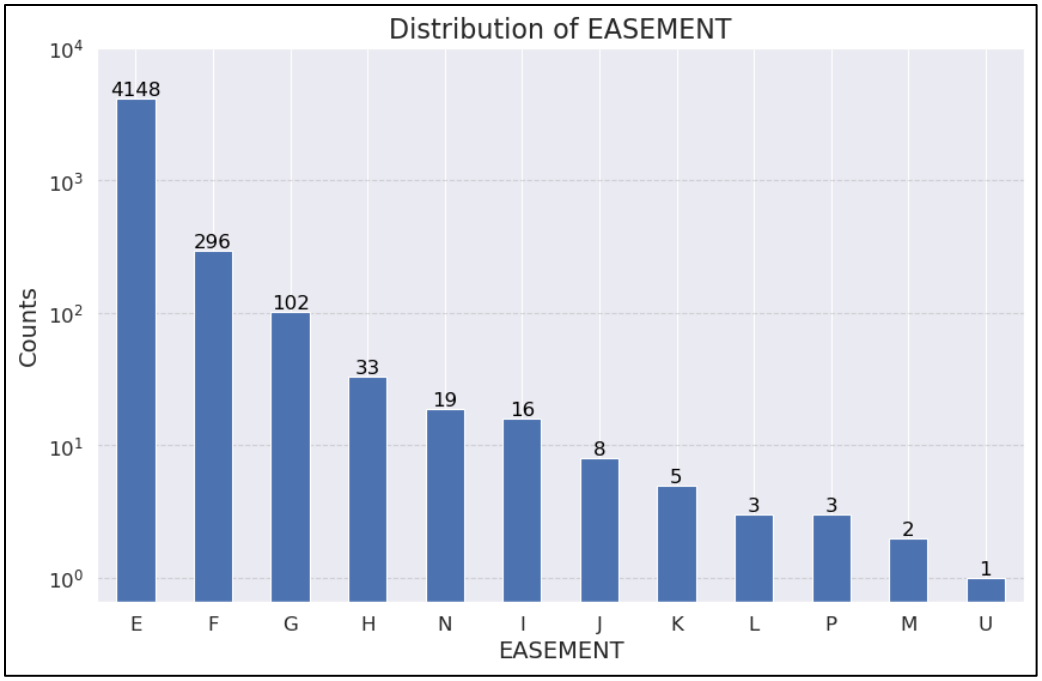
Description: Block number. There are a total of 13,984 unique block numbers. The distribution shows the top 20 values. The most common block number is '3944', with a total count of 3,888.



6) Field Name: EASEMENT

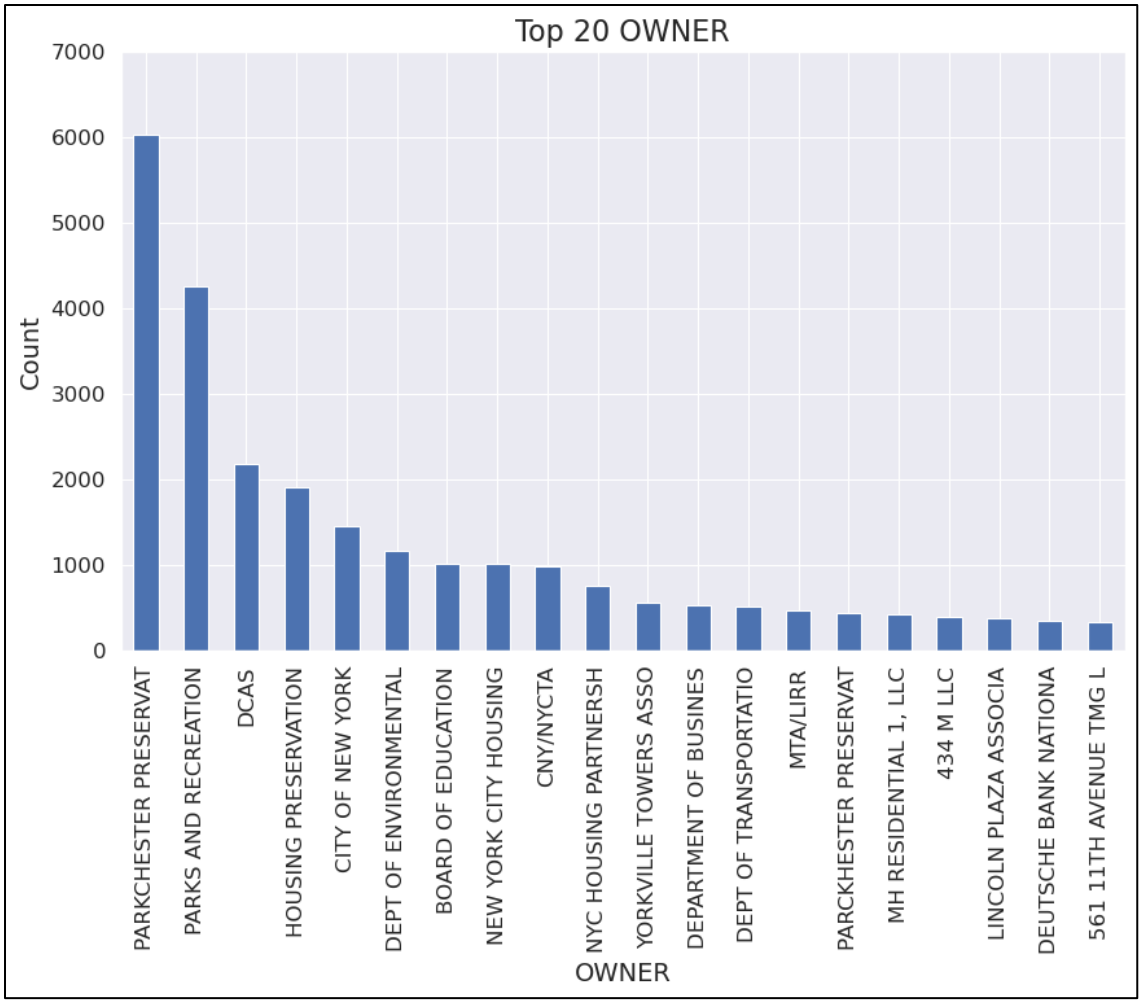
Description: Property's easement definition. There are a total of 13 unique values. The chart shows the distribution in logarithmic scale. The most common value is 'E', with a total count of 4,148.

Space = No Easement	A = Air Easement
B = Non-Air Rights	E = Land Easement
F Thru M are duplicates of E	N = Non-Transit Easement
P = Pier	R = Railroad
S = Street	U = U.S. Government



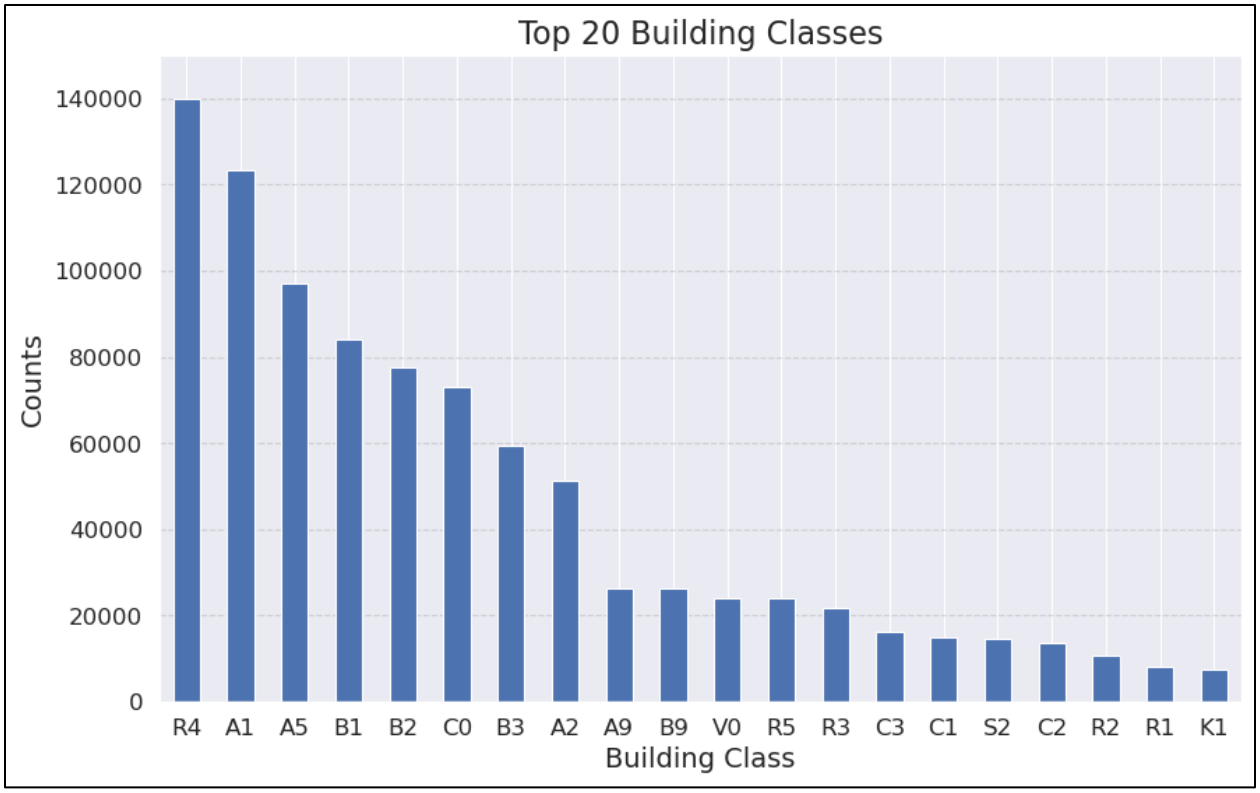
7) Field Name: OWNER

Description: Owner name. The distribution shows the top 20 field values of owner names. The most common name is 'PARKCHESTER PRESERVAT', with a total count of 6,021.



8) Field Name: BLDGCL

Description: Building class. There are a total of 200 unique building class values. The distribution shows the top 20 field values of building class. The most common value is ‘R4’, with a total count of 139,879.



9) Field Name: TAXCLASS

Description: Tax Class. There are a total of 11 unique tax class values, which are encoded as follows:

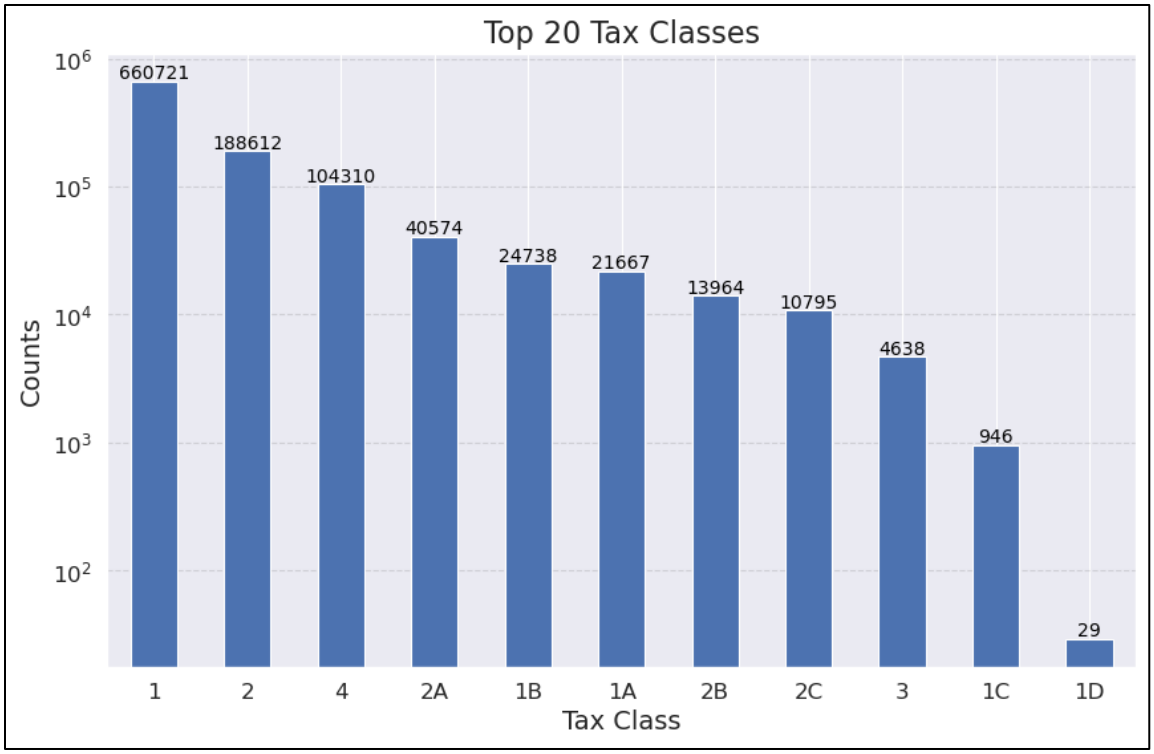
1 = 1 - 3 Unit Residence

2 = Apartments, 2A = 4, 5, or 6 Units

3 = Utilities

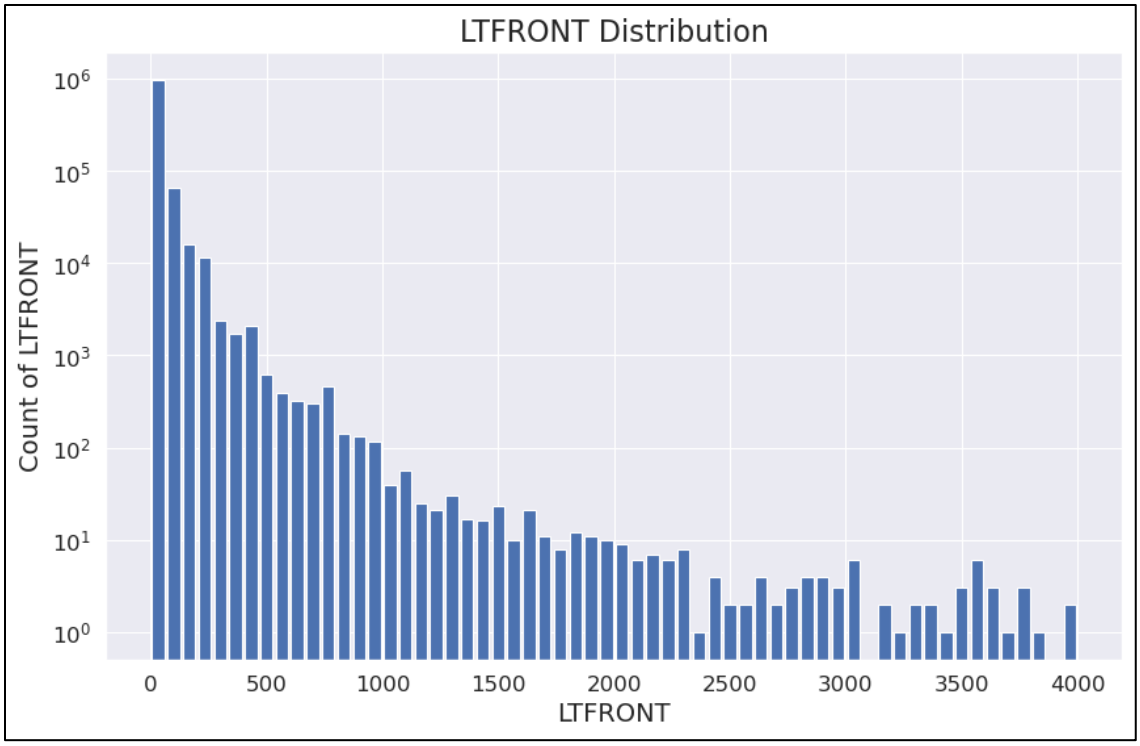
4 = All Others

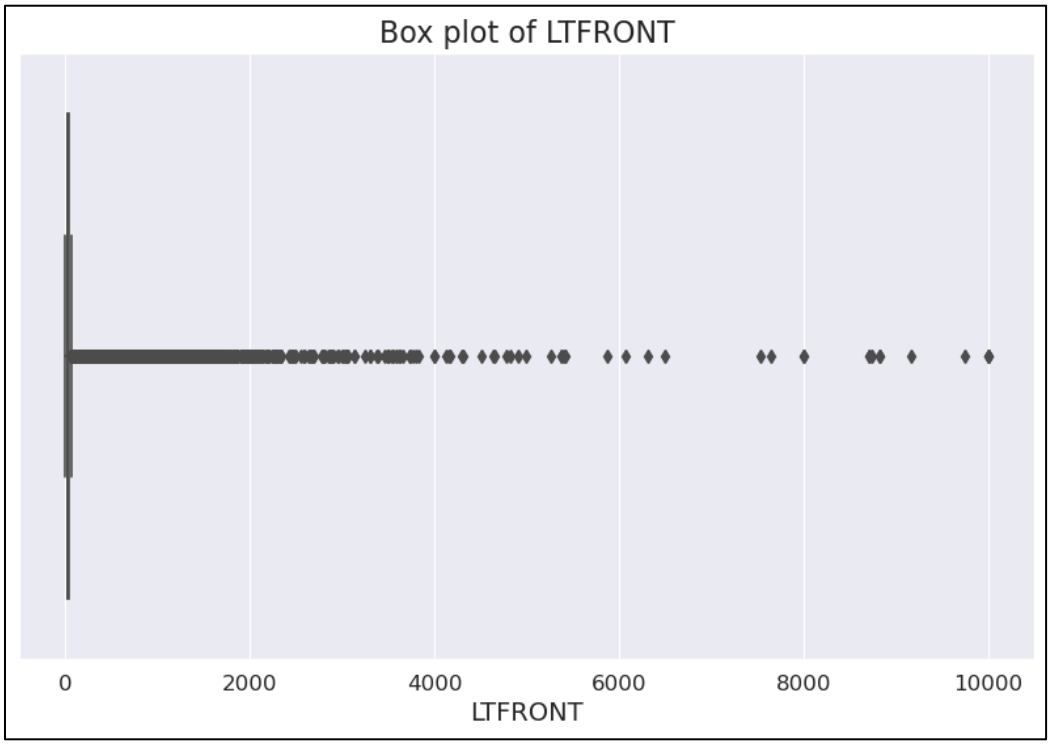
The most common tax class is '1', with a total count of 660,721.



10) Field Name: LTFRONT

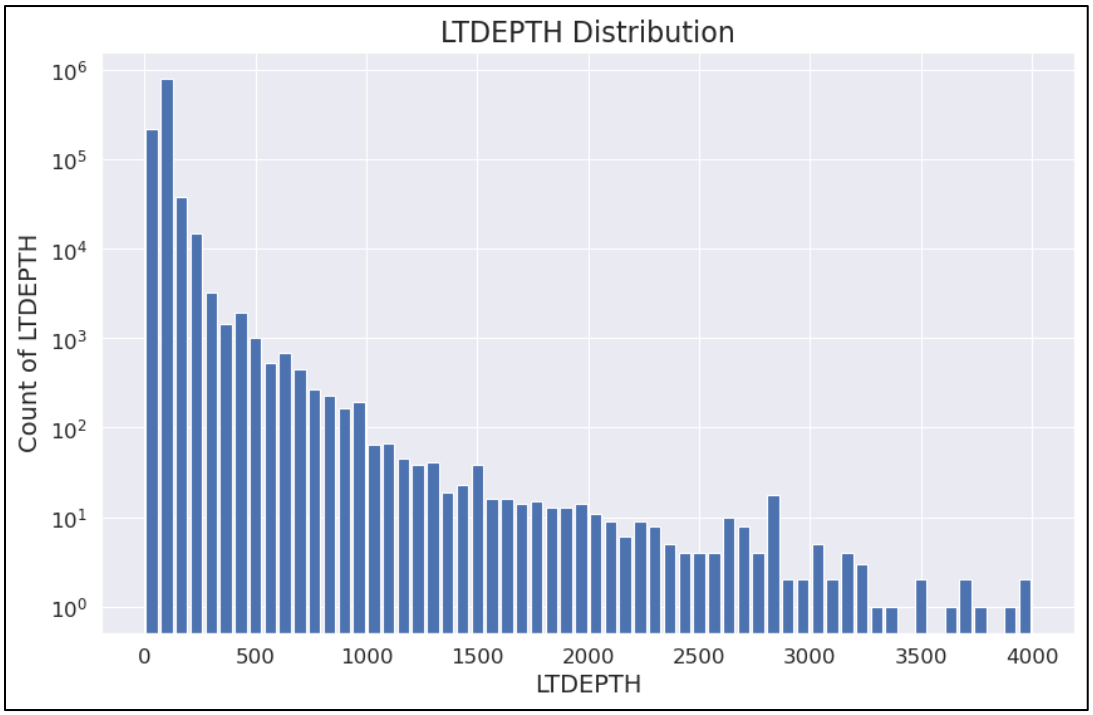
Description: Lot Width. The histogram shows the distribution of lot widths. The frequencies are shown in log scale, and only widths up to 4,000 are shown in the chart. The boxplot shows the entire distribution of the field. The maximum value is 9,999.

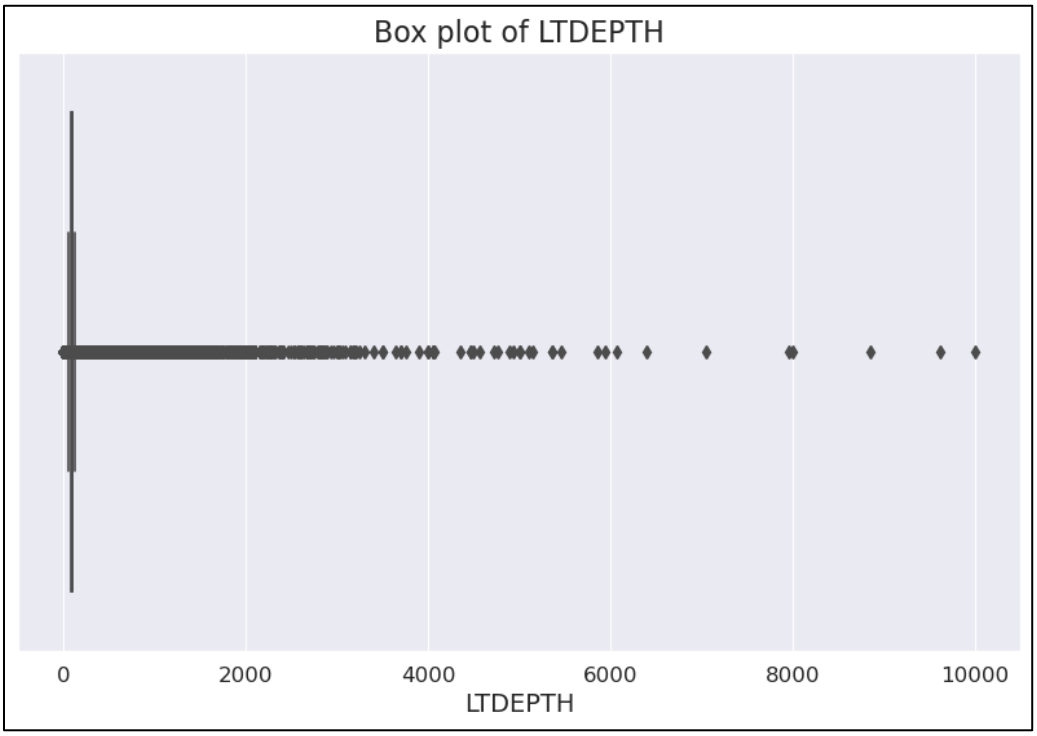




11) Field Name: LTDEPTH

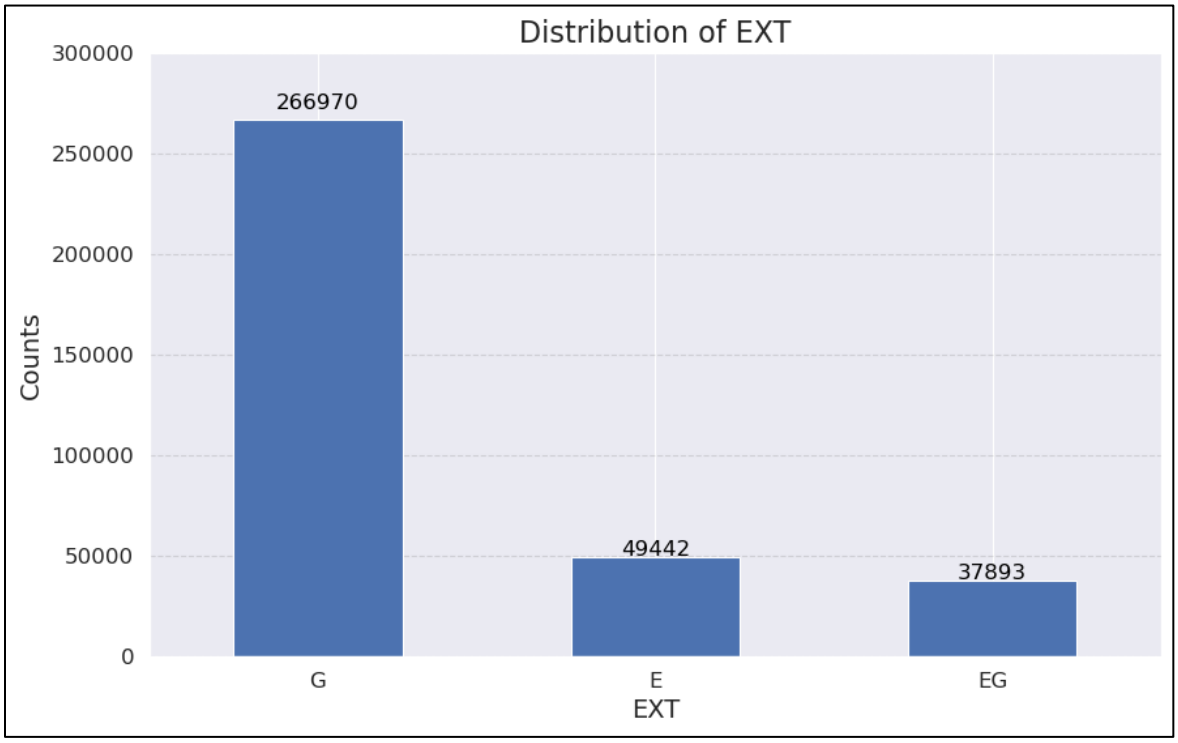
Description: Lot Depth. The histogram shows the distribution of lot depth. The frequencies are shown in log scale, and only widths up to 4,000 are shown in the chart. The boxplot shows the entire distribution of the field. The maximum value is 9,999.





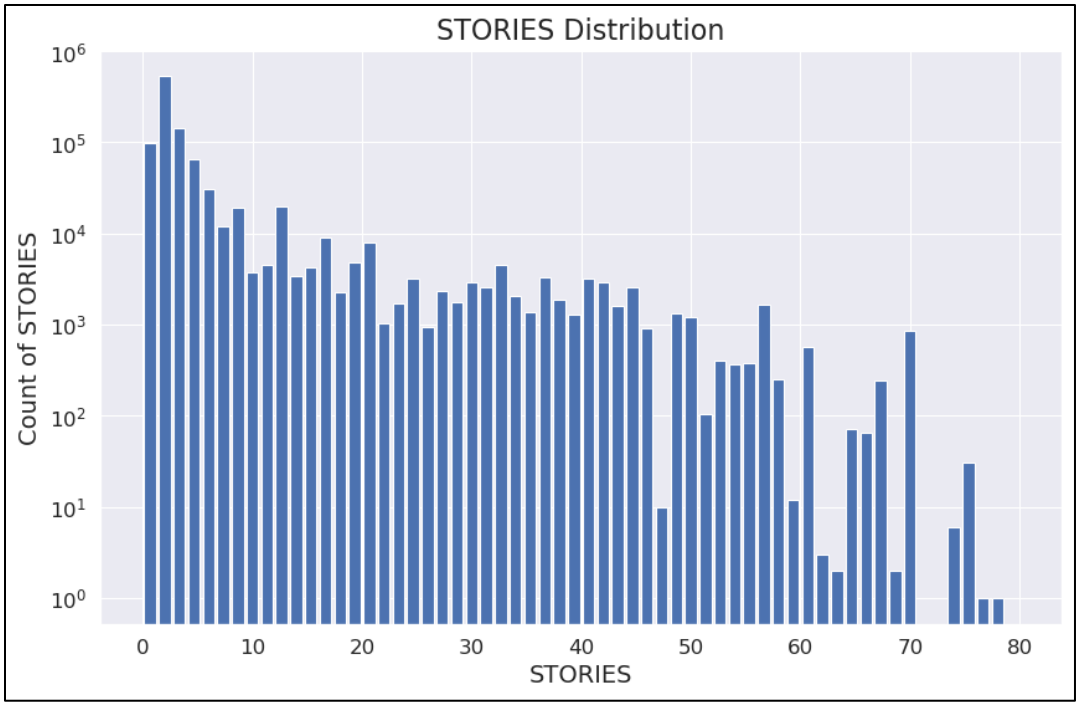
12) Field Name: EXT

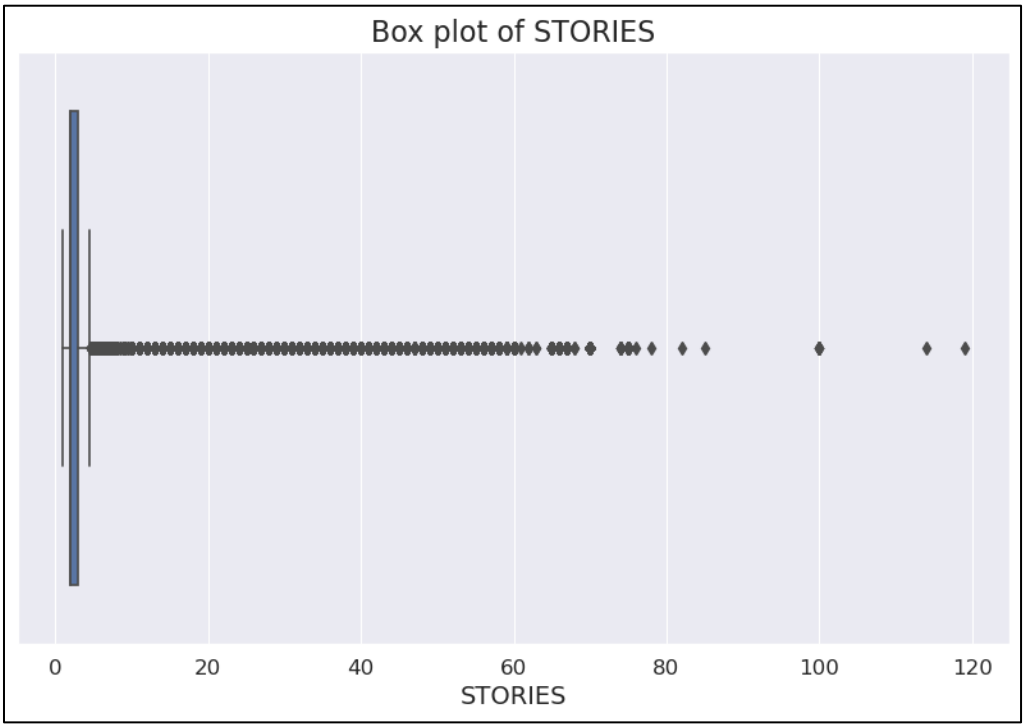
Description: Extension indicator. The chart shows the distribution of extension indicator. There are a total of 3 unique values for the field. The most common value is 'G', with a total count of 266,970.



13) Field Name: STORIES

Description: Number of stories in the building. There are a total of 112 unique values for this field. The histogram shows the distribution of the number of stories. The frequencies are shown in log scale, and only values up to 80 stories are shown in the chart. The boxplot shows the entire distribution of the field. The maximum value is 119.

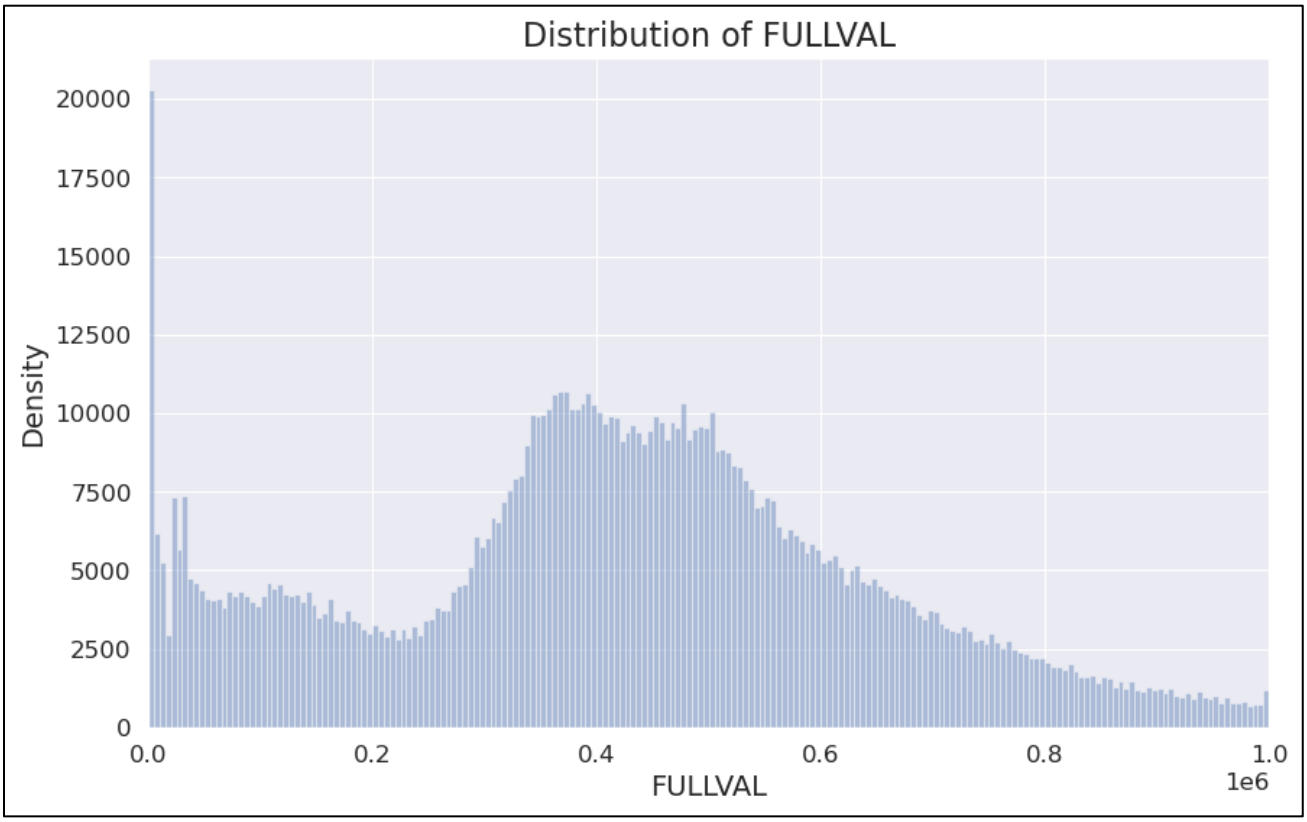


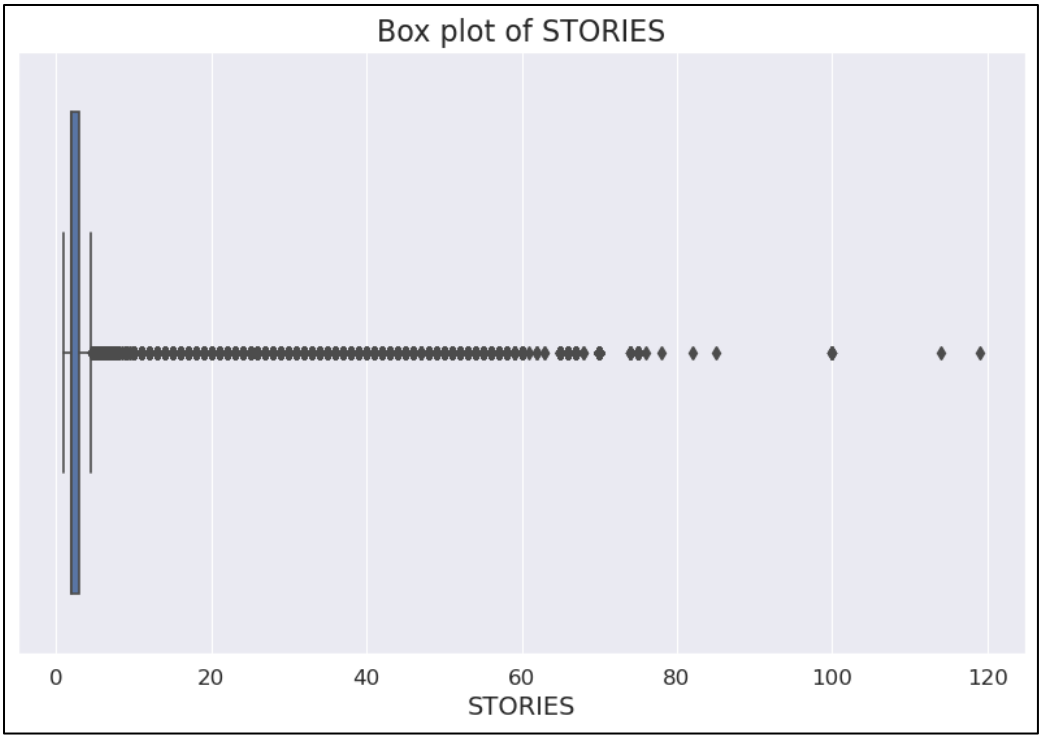


14) Field Name: FULLVAL

Description: Market value of the property. There is a total of 109,324 unique values for this field. The histogram shows the distribution of the market values. Only values up to 1,000,000 are shown in the

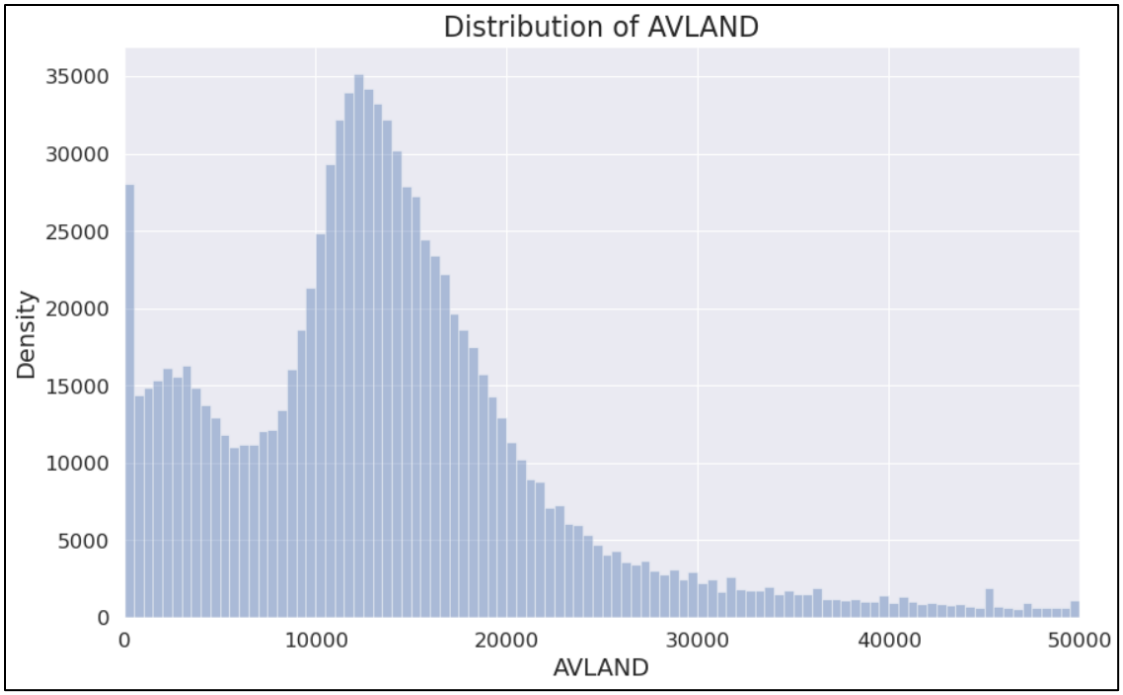
chart. The boxplot shows the entire distribution of the field. The most common value is ‘0’, with a total of 13,007 occurrences.

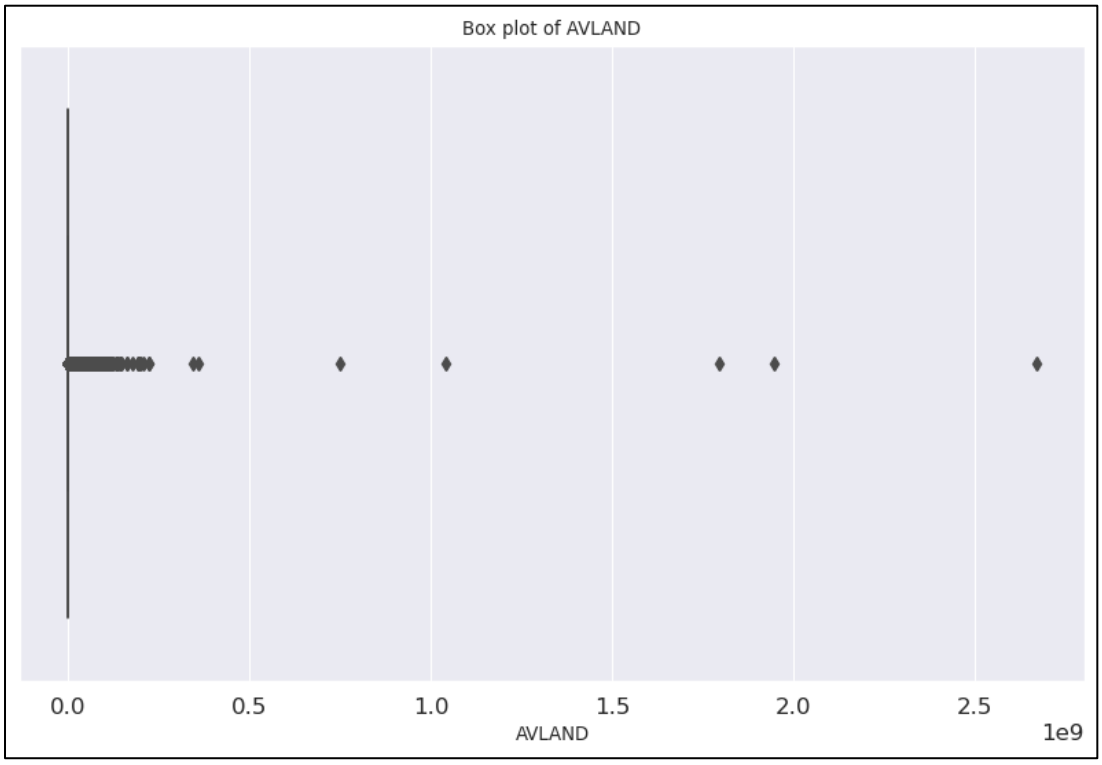




15) Field Name: AVLAND

Description: Actual land value. There is a total of 70,921 unique values for this field. The histogram shows the distribution of the actual land values. Only values up to 50,000 are shown in the chart. The boxplot shows the entire distribution of the field. The most common value is '0', with a total of 13,009 occurrences.

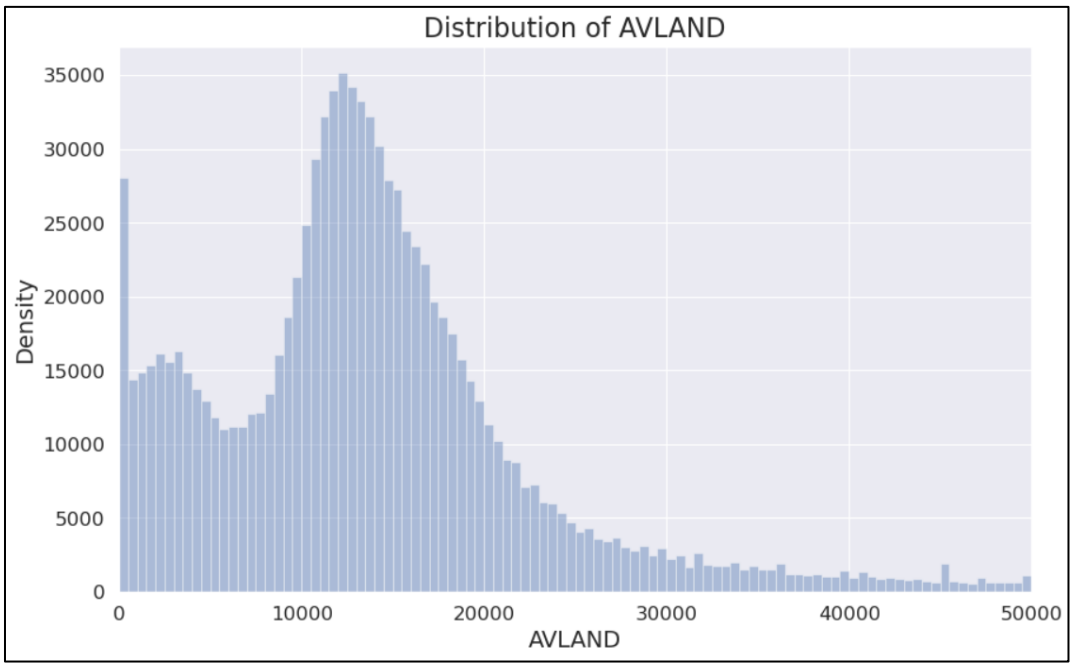




16) Field Name: AVTOT

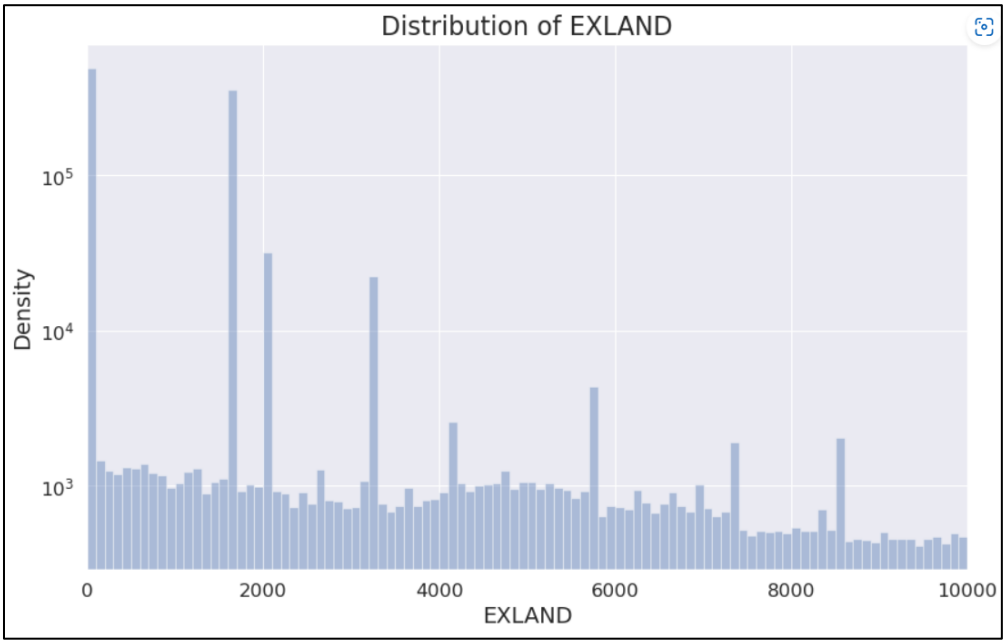
Description: Actual total value. There is a total of 112,914 unique values for this field. The histogram shows the distribution of the actual total values. Only values up to 100,000 are shown in the chart. The

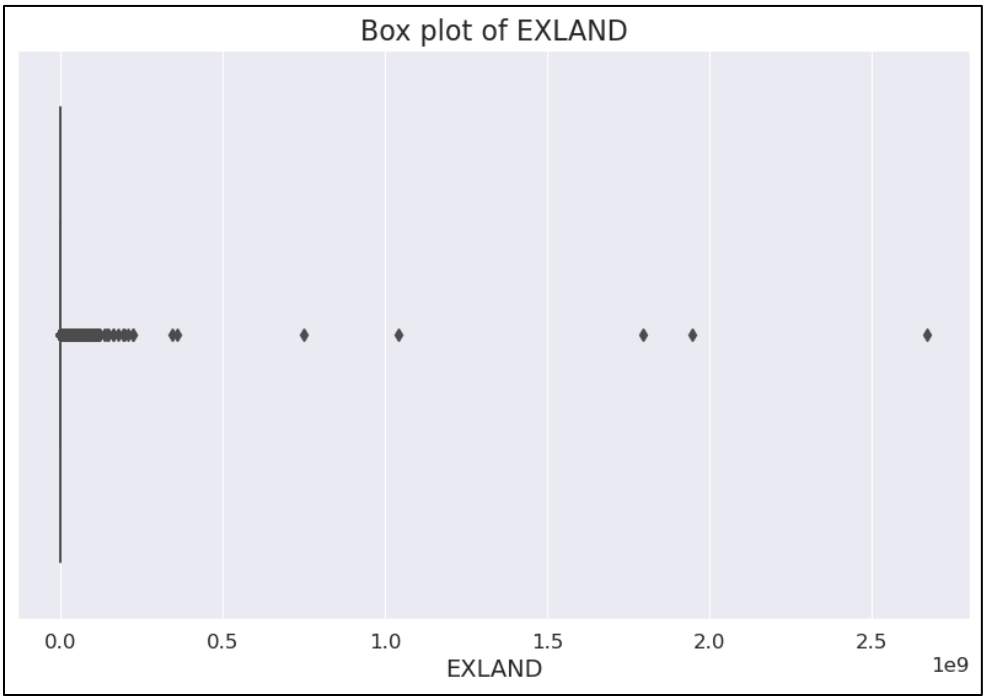
boxplot shows the entire distribution of the field. The most common value is ‘0’, with a total of 13,007 occurrences.



17) Field Name: EXLAND

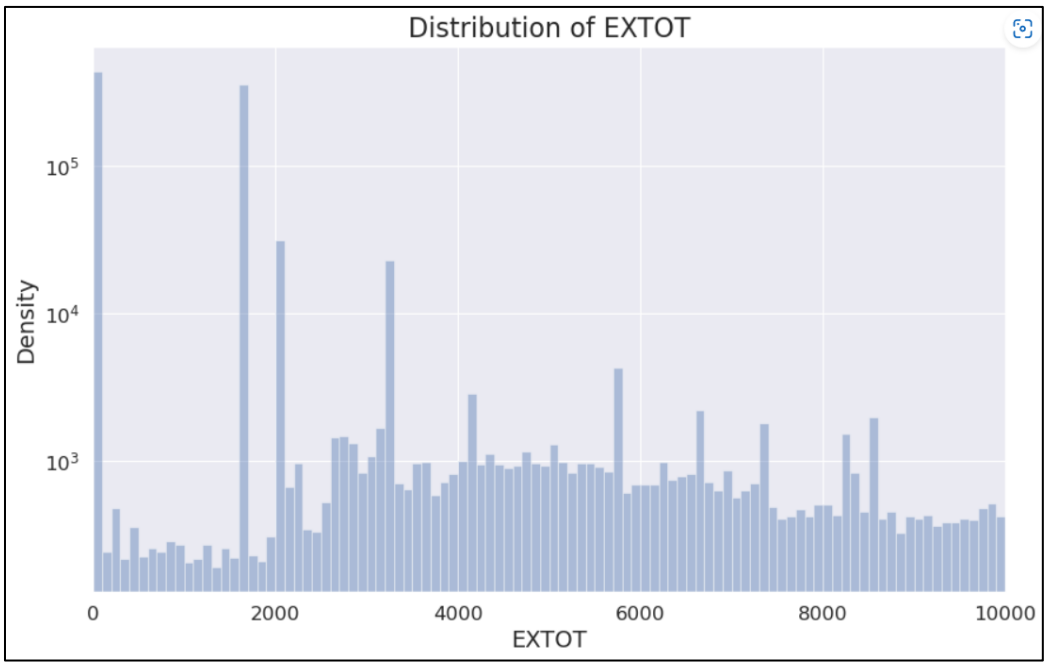
Description: Actual exempt land value. There is a total of 33,419 unique values for this field. The histogram shows the distribution of the actual exempt land values. Only values up to 10,000 are shown in the chart. The boxplot shows the entire distribution of the field. The most common value is '0', with a total of 491,699 occurrences.

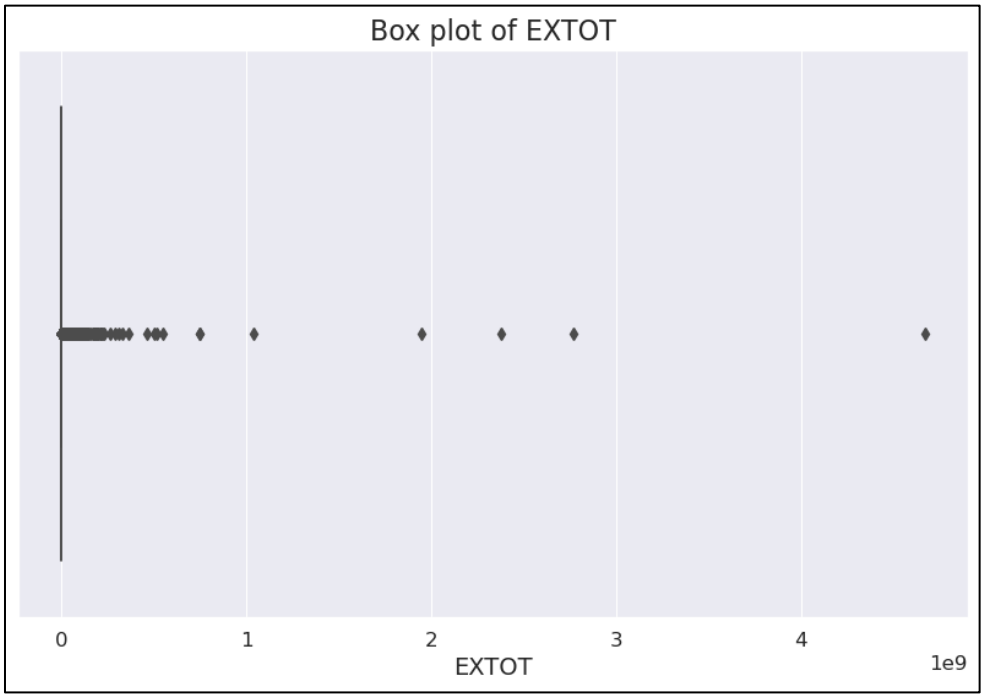




18) Field Name: EXTOT

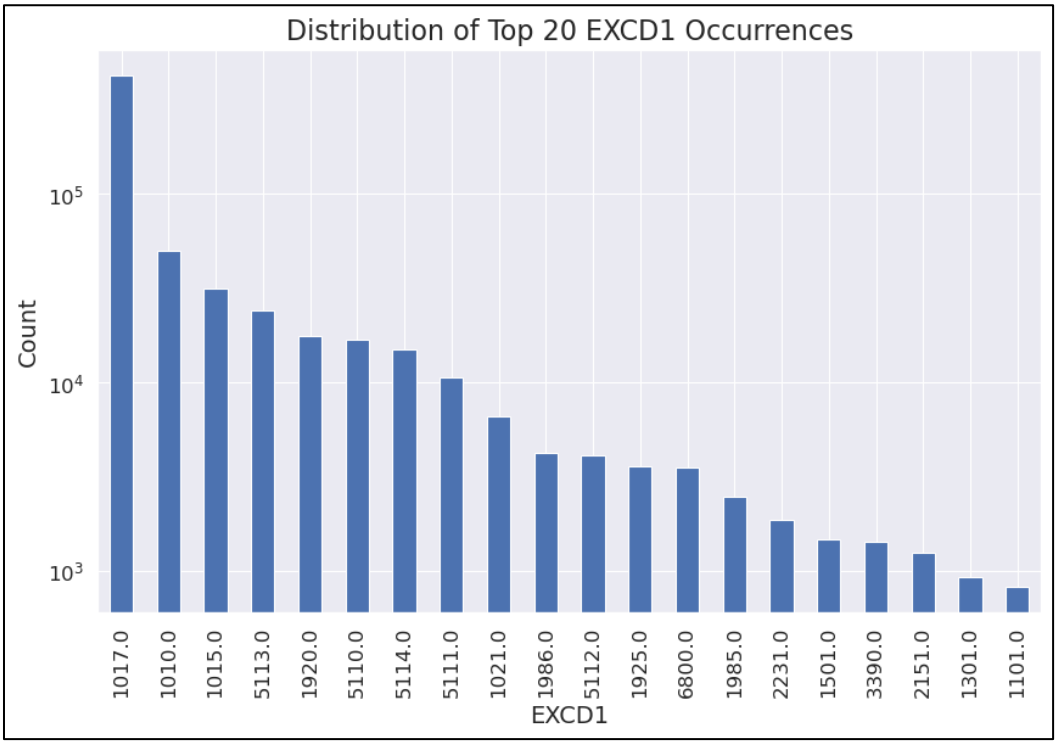
Description: Actual exempt land total. There is a total of 64,255 unique values for this field. The histogram shows the distribution of the actual exempt land values. Only values up to 10,000 are shown in the chart. The boxplot shows the entire distribution of the field. The most common value is '0', with a total of 432,572 occurrences.

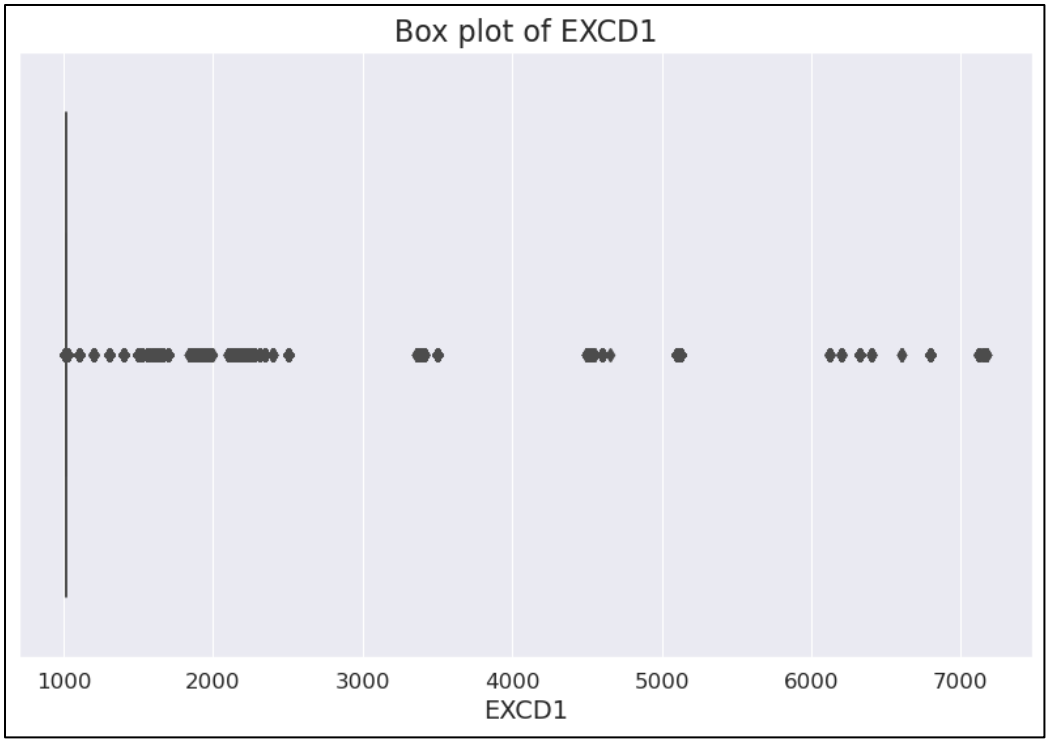




19) Field Name: EXCD1

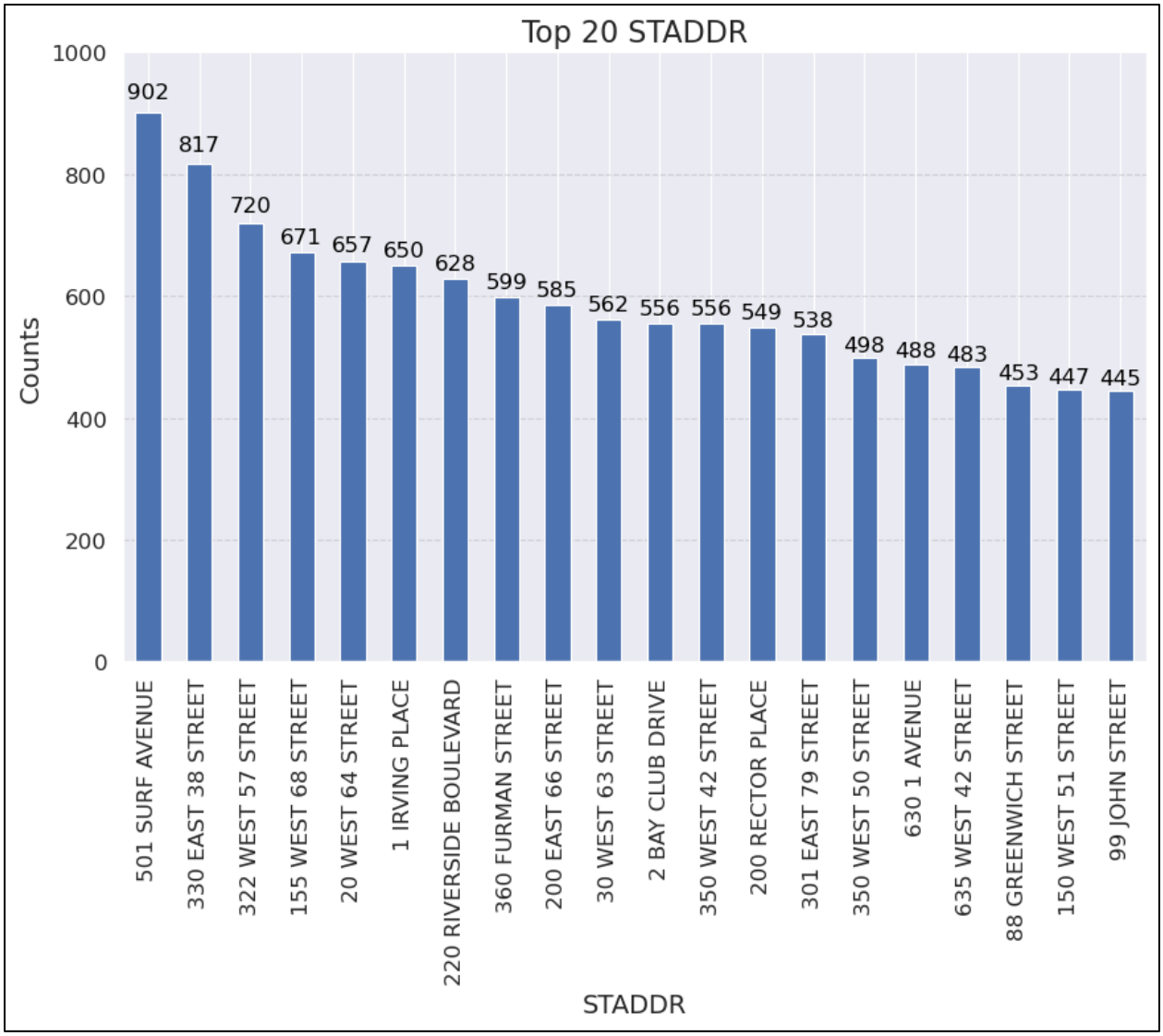
Description: Exemption code 1. There is a total of 130 unique values for this field. The chart shows the top 20 values for EXCD1. The boxplot shows the entire distribution of the field. The most common value is 1017, with a total of 425,348 occurrences.





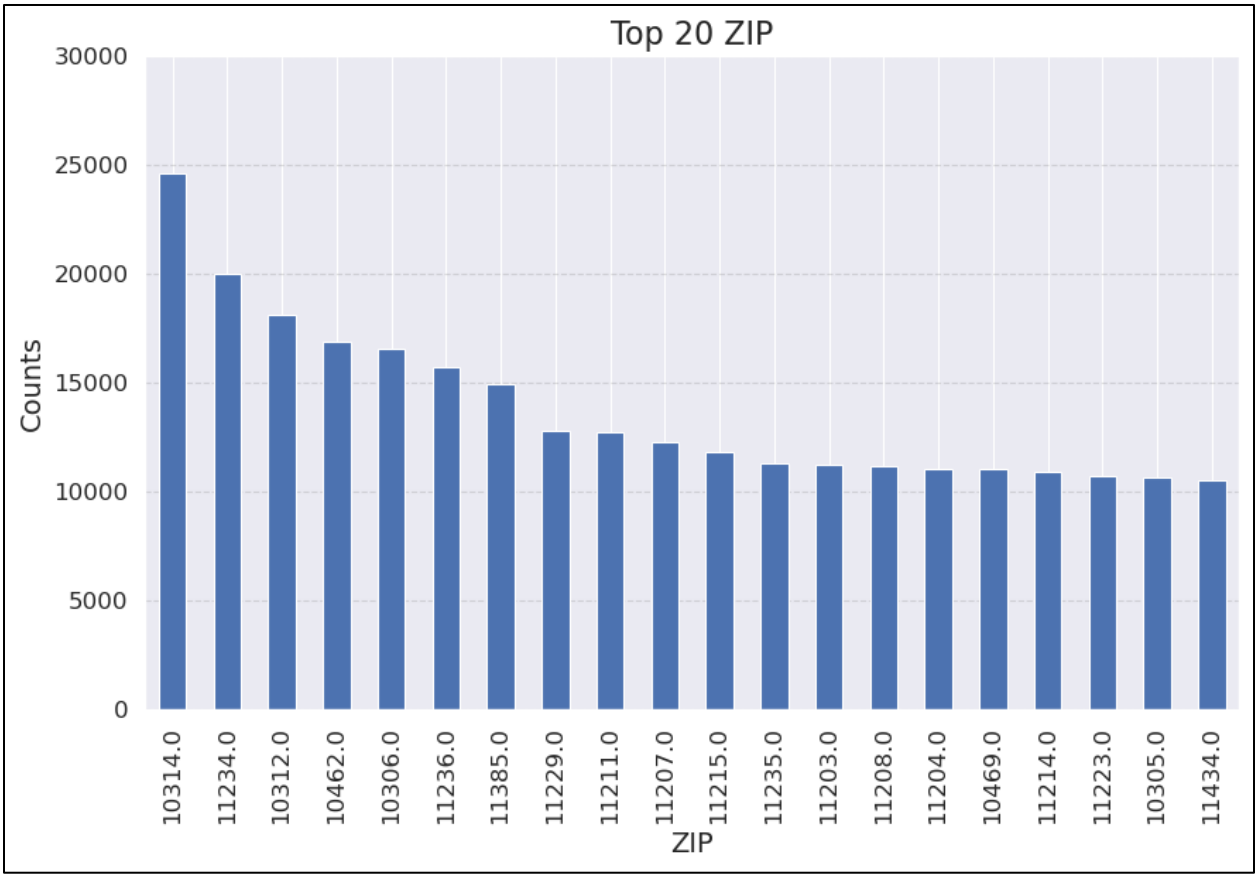
20) Field Name: STADDR

Description: Street Address. There is a total of 839,281 unique addresses. The distribution shows the top 20 field values. The most common address is ‘501 SURF AVENUE, with a total count of 902.



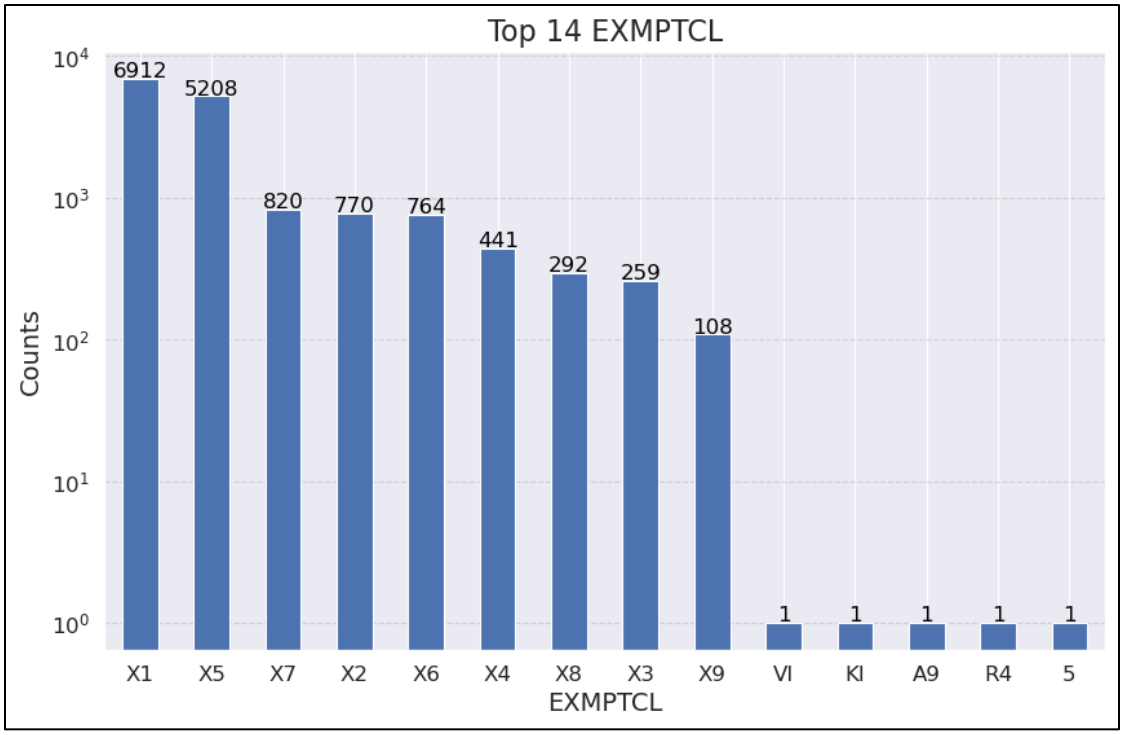
21) Field Name: ZIP

Description: Zip Code. There is a total of 197 unique zip codes. The distribution shows the top 20 field values. The most common zip code is '10314', with a total count of 24,606.



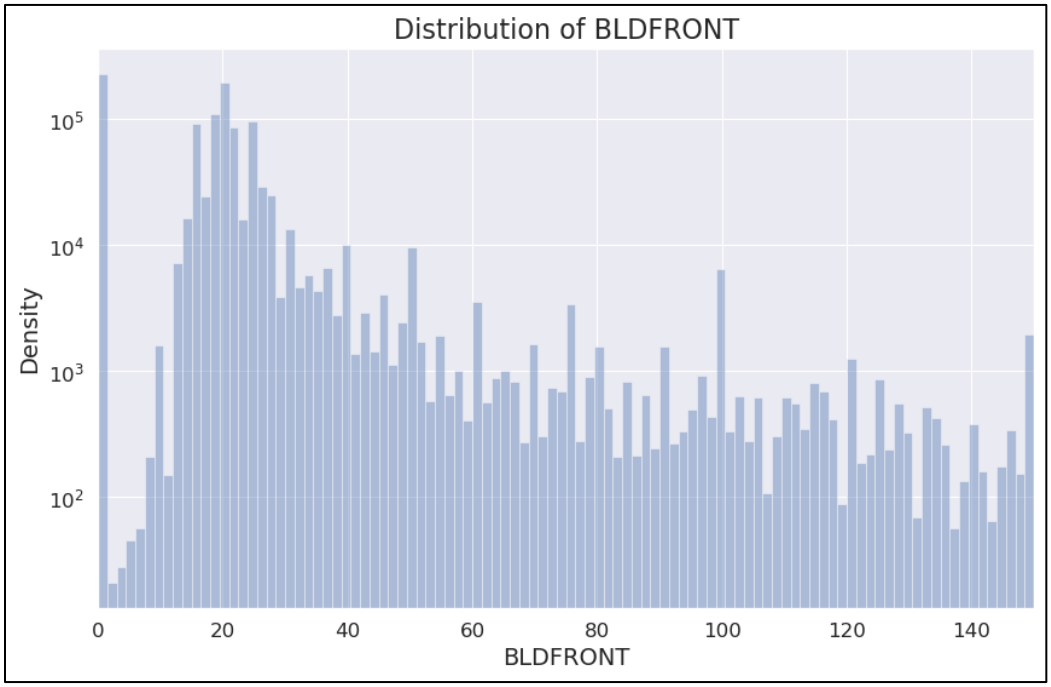
22) Field Name: EXMPTCL

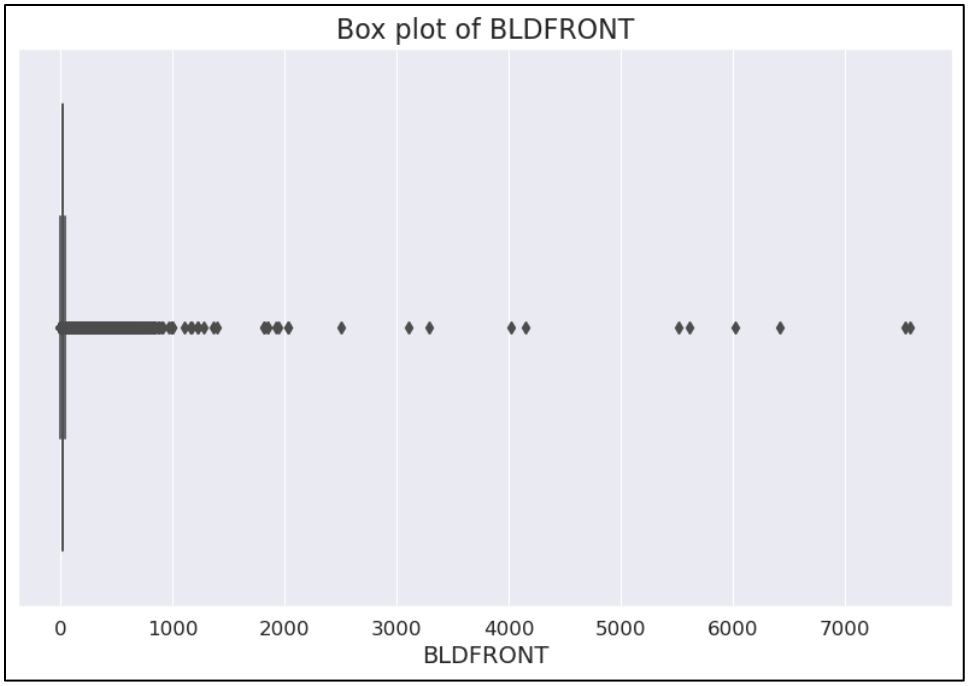
Description: Exemption Class. There is a total of 15 unique exemption classes. The chart shows the distribution in a log scale. The most common exemption class is 'X1', with a total count of 6,912.



23) Field Name: BLDFRONT

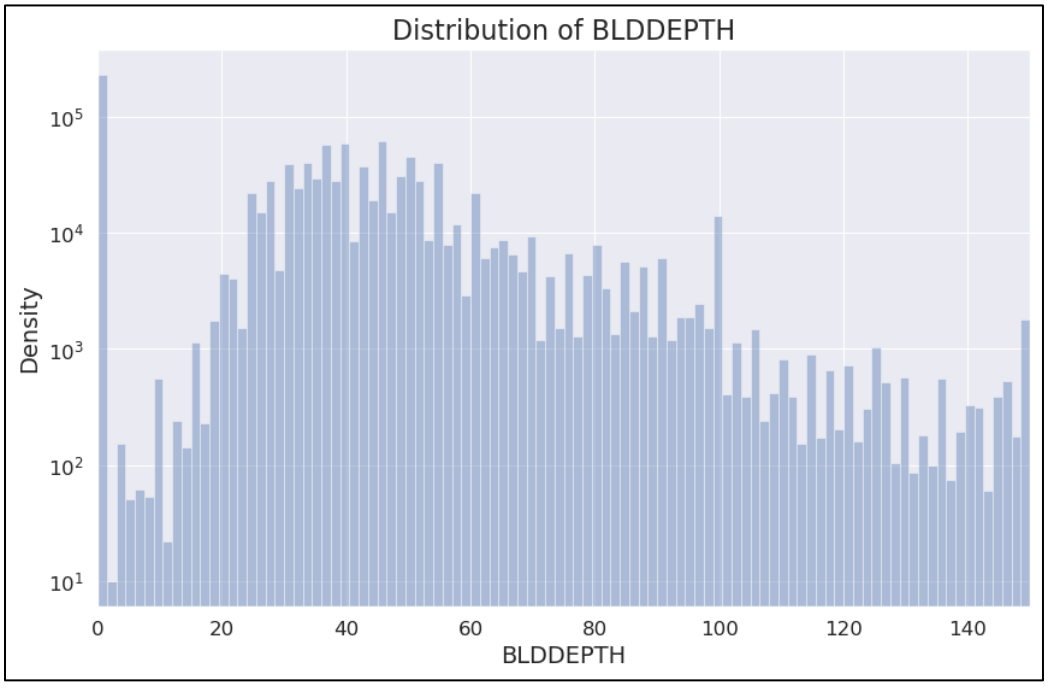
Description: Building Width. The histogram shows the distribution of building widths. The frequencies are shown in log scale, and only widths up to 150 are shown in the chart. The boxplot shows the entire distribution of the field.

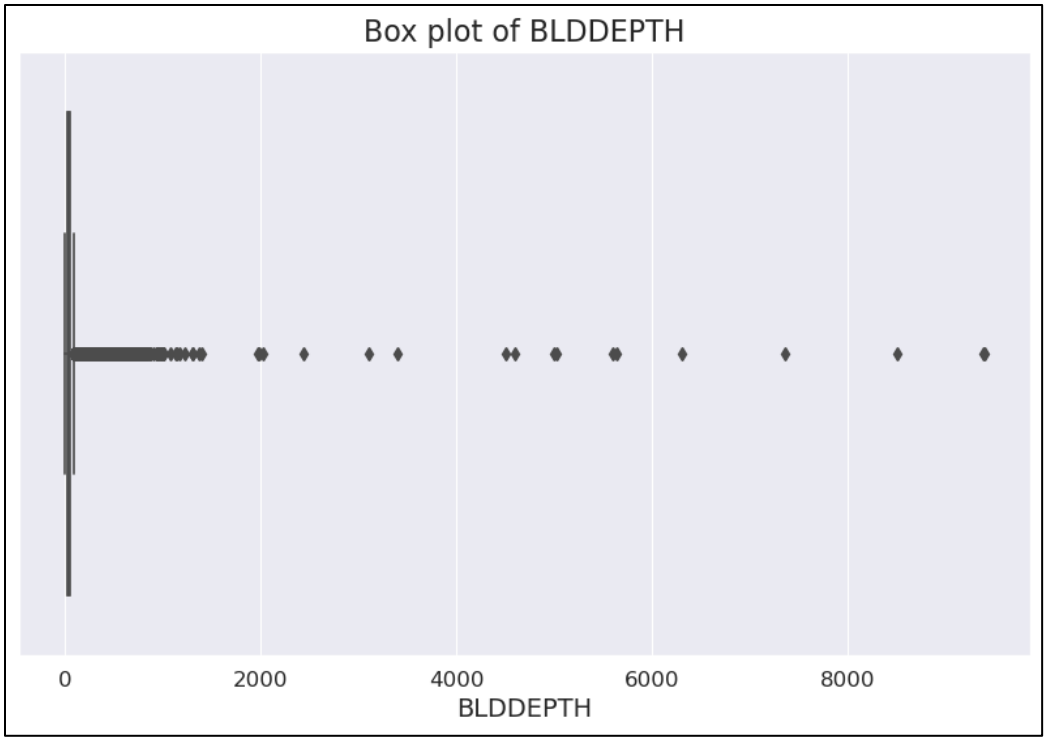




24) Field Name: BLDDEPTH

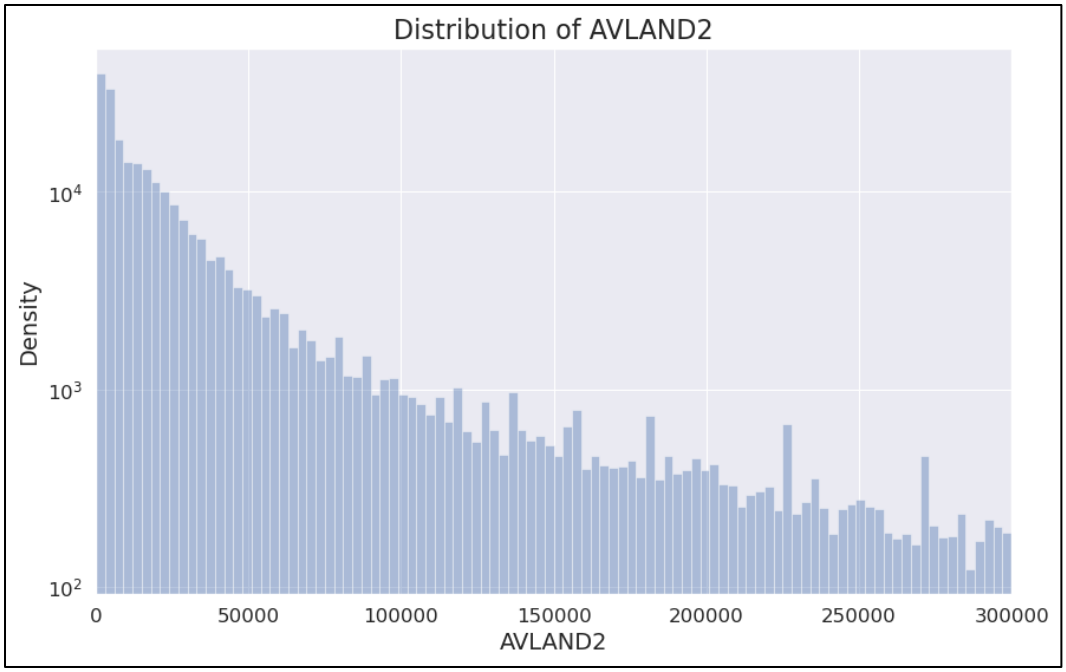
Description: Building Depth. The histogram shows the distribution of building depths. The frequencies are shown in log scale, and only depths up to 150 are shown in the chart. The boxplot shows the entire distribution of the field.

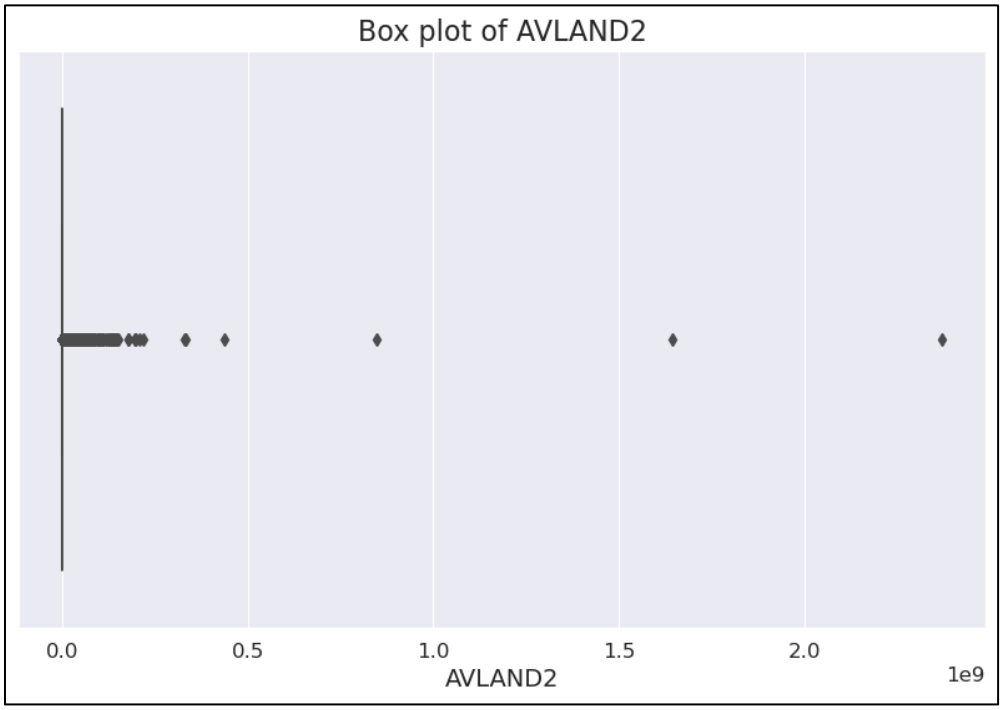




25) Field Name: AVLAND2

Description: Transitional land value. There is a total of 58,592 unique values for this field. The histogram shows the distribution of the actual exempt land values. Only values up to 300,000 are shown in the chart. The boxplot shows the entire distribution of the field. The most common value is '2,408', with a total of 767 occurrences.





26) Field Name: AVTOT2

Description: Transitional total value. There is a total of 111,361 unique values for this field. The histogram shows the distribution of the actual exempt land values. Only values up to 1,000,000 are

shown in the chart. The boxplot shows the entire distribution of the field. The most common value is ‘750’, with a total of 656 occurrences.