MGTA 463 Project 2 Final Report

- ❖ Executive Summary
- ❖ Part I: Description of the data
- ❖ Part II: Data Cleaning
- ❖ Part III: Variable Creation
- ❖ Part IV: Dimensionality Reduction
- ❖ Part V: Anomaly Detection Algorithms
- ❖ Part VI: Results
- ❖ Part VII: Summary
- ❖ Appendix - Data Quality Report

Executive Summary

In this project, we addressed the issue of identifying property tax fraud in New York City using anomaly detection techniques. The dataset comprised over one million property records, including various numeric and categorical fields. Our approach involved data cleaning, variable creation, dimensionality reduction, and the implementation of two anomaly detection algorithms. By ranking the records based on these algorithms, we identified the most suspicious property records.

Our findings were consolidated into several output files for detailed investigation, which highlighted numerous anomalies primarily due to null values in key fields. The output files enable us to investigate the anomalous property records. Our current approach involves matching these records with online data to identify erroneous information that could indicate tax fraud. This approach and the results should be further confirmed by experts in the field.

Part I - Description of the data

- Data Overview

The dataset records Property Valuation and Assessment data in New York City. The data encompasses 32 detailed description fields of each property in New York City yet without a label. The 32 fields are built by 14 numeric fields and 18 categorical fields. The data consist of 10,70,994 property records.

- Summary Tables

Numeric Fields Table

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

Categorical Fields Table

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

- Critical Fields and Label's Distribution Plot

Field Name: BBLE

Description: BBLE stands for Borough, Block, Lot, and Easement. It is a combination of digits representing each of these elements to uniquely identify a property.

# of Unique Values: 1,070,994

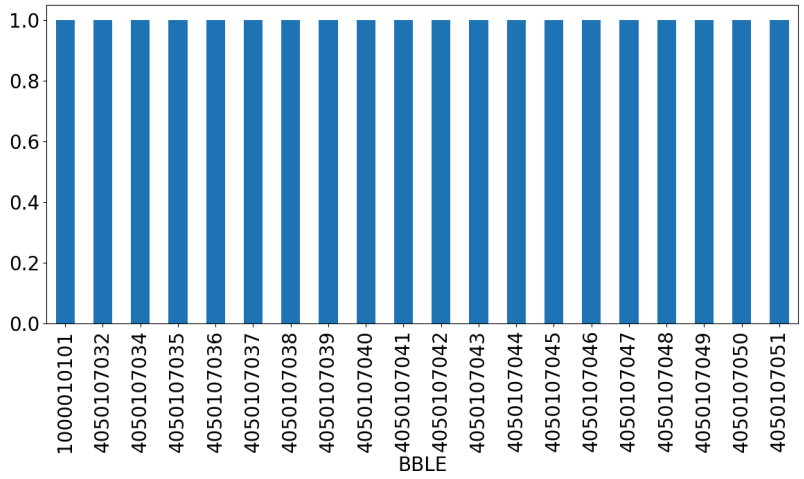
Null Values: 0

Observation: One unique BBLE for each property.

Naming Rule:

1st Digit: Boro; 2 to 6th Digit: Block; 7 to 10th Digit: Lot; Last: Easement

Ex: 1(Boro)08050(Block)0038(LOT)E(EASEMENT)



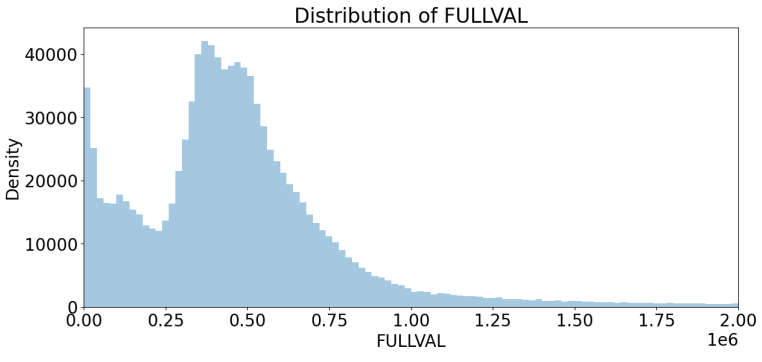
Field Name: FULLVAL

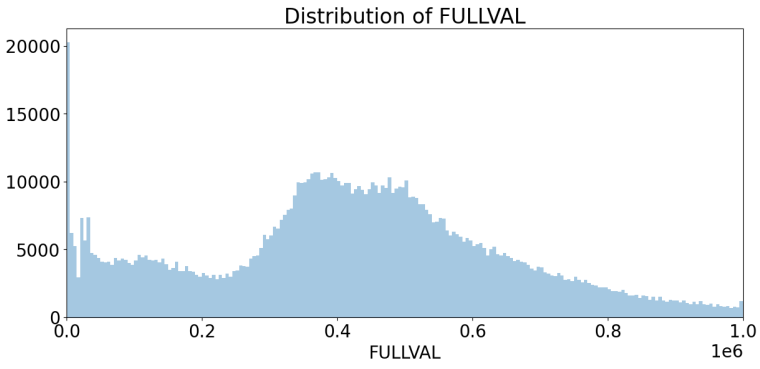
Description: Market value of the property

# of Unique Values: 109,324

Null Values: 0

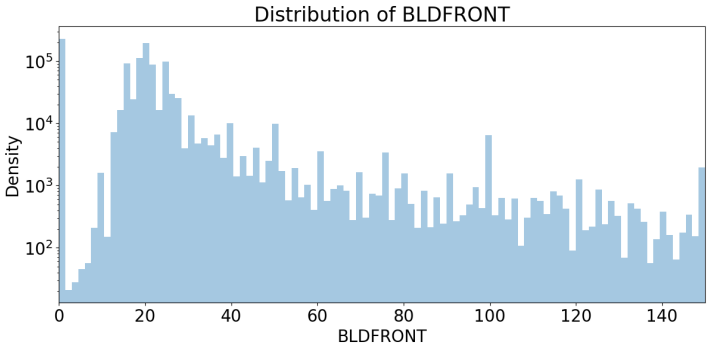
Observation: This field is bimodal distributed. The reason for this distribution is that 13,007 properties recorded as $0 market value. In the nature of this field, most of the values are valued around $30,000 to 50,000 dollars. Values more than $35,621,557 are considered outliers, with 1,919 properties.

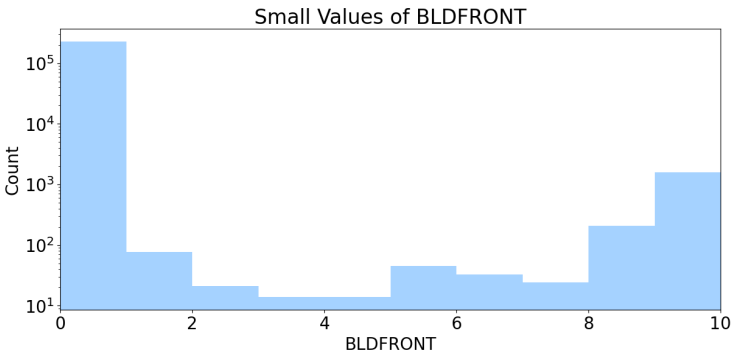




Field Name: BLDFRONT

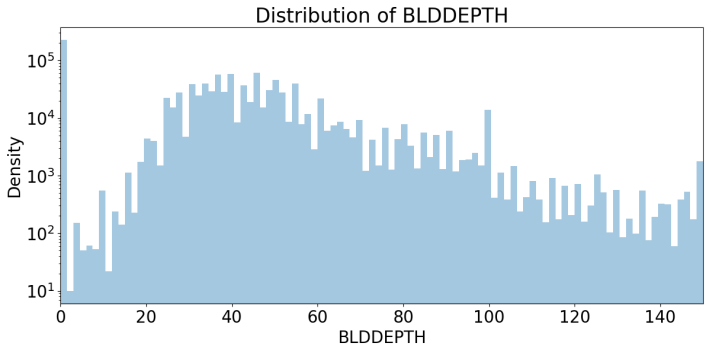
Description: The building width of the property
# of Unique Values: 612
Null Values: 0
Observation: The distribution of this field is highly right-skewed, but noting that there are 228,815 properties with value of 0, forming a bimodal-like distribution. Neglecting records of 0s, the distribution is concentrated around 20. Values higher than 129.78 are considered as outliers, with a total of 18,922 properties.

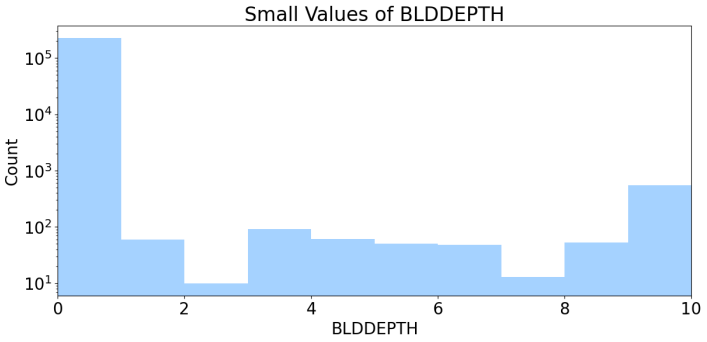




Field Name: BLDDEPTH

Description: The building depth of the property
# of Unique Values: 621
Null Values: 0
Observation: The distribution of this field is highly right-skewed, but noting that there are 228,853 properties with value of 0, forming a bimodal-like distribution. Neglecting records of 0s, the distribution is concentrated around 30 to 60. Values higher than 168.04 are considered as outliers, with a total of 10,647 properties.





Part II - Data Cleaning

The data includes 1,070,994 records of property data, with 32 fields. In the data cleaning process, we excluded government property records that are irrelevant to our objectives, then imputed the null values with distinct methodologies for each field. Our model will not consider all the fields in the data, so we will solely focus the data cleaning procedure on the fields that are implemented in the further process.

- Data Exclusions

The detection model should be focusing on the property of citizens. Hence, for this part, we aimed to exclude government property records. Our concept of defining government property is based on two steps. First, remove property record with ‘EASEMENT’ type as ‘government’. Then, remove property records that recorded their owner name that include ['DEPT ', 'DEPARTMENT', 'UNITED STATES','GOVERNMENT',' GOVT ', 'CEMETERY']. Additionally, we intuitively removed the top 20 property owners in terms of count that are likely to be government facilities. Actual removal process as following:

Field Name: Easement

Records removed: 1

Description: Remove the records with easement type as ‘government’.

Total records: 1,070,994 - 1 = **1,070,993**

Field Name: OWNER

Records removed: 26,500

Description: Remove the records that the OWNER field includes ['DEPT ', 'DEPARTMENT', 'UNITED STATES','GOVERNMENT',' GOVT ', 'CEMETERY'] and the property that is owned by the top 20 owners in terms of property count.

Total records: 1,070,993 - 26,500 = **1,044,493**

- Null Values Imputation for the fields

Field: ZIP

# of null values: 20,431

1) There are **20,431** null values for ZIP. The following steps set forth the imputation process for ZIP.
2) Trace back the property's zip code that retains information of borough and address, which are columns 'BORO' and 'STADDR'. This action took care of 2,832 records, leaving **17,599** null values remaining.
3) Assume that data is already sorted by zip. For every null zip code, if the previous and the next zip codes are the same, we fill in the zip with that value. This action took care of 9,491 records, leaving **8,108** null values remaining.
4) For the remaining null values, we decided to fill in with the previous record's zip code. This action winds down the number of missing values in ZIP to **0**.

Field: FULLVAL

# of null values or values with 0: 10,025

1) There are **10,025** null values for FULLVAL. The following steps set forth the imputation process for FULLVAL.
2) Recover the property's FULLVAL with the average FULLVAL of each specific group of 'TAXCLASS', 'BORO', and 'BLDGCL. This action took care of 2,718 records, leaving **7,307** null values remaining.
3) Recover the property's FULLVAL with the average FULLVAL of each specific group of 'TAXCLASS', and 'BORO'. This action took care of 6,921 records, leaving **386** null values remaining.
4) For the remaining null values, we decided to fill in with the average FULLVAL of each specific 'TAXCLASS'. This action winds down the number of missing values in FULLVAL to **0**.

Field: AVLAND

# of null values or values with 0: 10,027

1) There are **10,027** null values for AVLAND. The following steps set forth the imputation process for AVLAND.
2) Recover the property's AVLAND with the average AVLAND of each specific group of 'TAXCLASS', 'BORO', and 'BLDGCL. This action took care of 2,720 records, leaving **7,307** null values remaining.
3) Recover the property's AVLAND with the average AVLAND of each specific group of 'TAXCLASS', and 'BORO'. This action took care of 6,921 records, leaving **386** null values remaining.
4) For the remaining null values, we decided to fill in with the average AVLAND of each specific 'TAXCLASS'. This action winds down the number of missing values in AVLAND to **0**.

Field: AVTOT

# of null values or values with 0: 10,025

1) There are **10,025** null values for AVTOT. The following steps set forth the imputation process for AVTOT.
2) Recover the property's AVTOT with the average AVTOT of each specific group of 'TAXCLASS', 'BORO', and 'BLDGCL. This action took care of 2,718 records, leaving **7,307** null values remaining.
3) Recover the property's AVTOT with the average AVTOT of each specific group of 'TAXCLASS', and 'BORO'. This action took care of 6,921 records, leaving **386** null values remaining.
4) For the remaining null values, we decided to fill in with the average AVTOT of each specific 'TAXCLASS'. This action winds down the number of missing values in AVTOT to **0**.

Field: LTFRONT

# of null values(transferred from 0 and 1s): 160,565

1) There are **160,565** null values for LTFRONT. The following steps set forth the imputation process for LTFRONT.
2) Recover the property's LTFRONT with the median LTFRONT of each specific group of 'TAXCLASS' and 'BORO'. This action took care of 160,563 records, leaving **2** null values remaining.
3) Recover the property's LTFRONT with the average LTFRONT of each specific 'TAXCLASS'. This action winds down the number of missing values in LTFRONT to **0**.

Part III - Variable creation

First, we created 3 area metrics for lot area, building area, and building size:
ltsize = LTFRONT * LTDEPTH
bldsize = BLDFRONT * BLDDEPTH
bldvol = bldsize * STORIES
Variable Creation Concept:

To identify anomalies in property tax fraud, we focused on the property's value, location, and category. We created variables by dividing property value by property area. Additionally, we examined the differences in these variables across different zip codes and tax classes. To ensure accuracy, we developed another variable to verify that the ratio of value to area remains consistent, addressing instances where submitted values and property areas may not align with the actual figures.

Description	# of Field Created	Cumulative Fields
Field Name: [‘r1’, ‘r2’, ‘r3’, ‘r4’, ‘r5’, ‘r6’, ‘r7’, ‘r8’, ‘r9’] To calculate the price per square foot for land and buildings, we divided three value metrics [‘FULLVAL’, ‘AVLAND’, ‘AVTOT’] by three area metrics [‘ltsize’, ‘bldsize’, ‘bldvol’]. Then, to identify outliers that are too small in the data, we created nine features that are the inverse of r1 through r9. We will only consider the larger value between the original data and its inverse, as our goal is to identify the outliers.	9	9
Field Name: ['rn_zip5'], where n ranges from 1 to 9. We created nine additional features based on the nine price metrics, divided by the average value of the price metrics for each 'ZIP' code.	9	18
Field Name: ['rn_taxclass'], where n ranges from 1 to 9. We created nine additional features based on the nine price metrics, divided by the average value of the price metrics for each 'TAXCLASS'.	9	27
Field Name: ['value_ratio'] To check for falsely inputted records, we divided the full value of the property by the sum of the actual land value and the actual total value. We then divided the result by the average property value to identify potential false records. Normally, the data should have a value close to 0. Using the same approach as before, we considered the larger value between the result and its inverse to identify outliers.	1	28
Field Name: [‘size_ratio’] The field is to compare the size of building and lot size. It is the ratio of building size(bldsize) and lot size(ltsize).	1	29

Part IV - Dimensionality Reduction

- Why Dimensionality Reduction

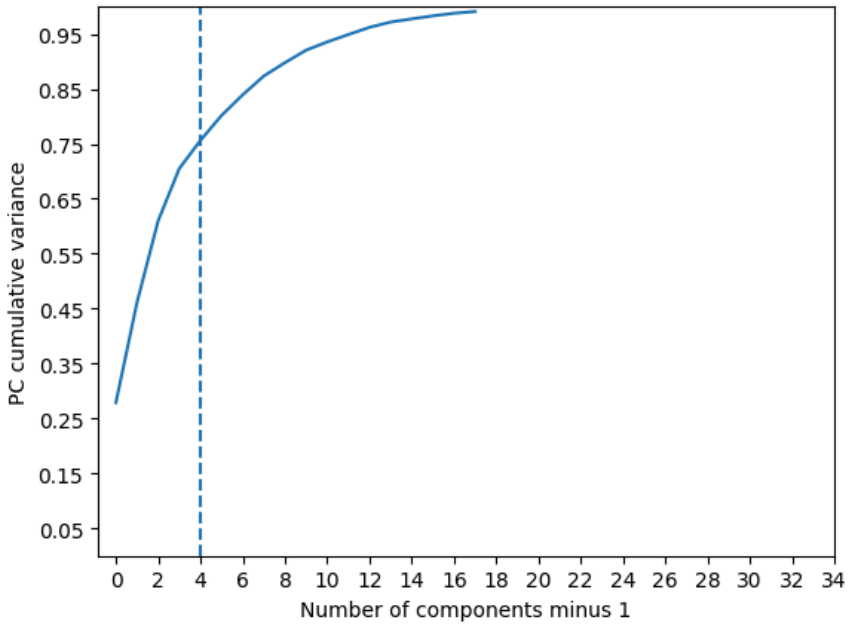
In the last section, we developed 29 expert variables based on our experience. However, the creation of these features largely involves combinations of duplicated fields, such as FULLVAL, TOTVAL, and AVLAND. As a result, it is likely that the expert variables are highly correlated. To reduce correlation between the variables and conduct feature selection, we use Principal Component Analysis (PCA) as the standard method to address this issue.

- What is Principal Component Analysis?

Principal Component Analysis (PCA) identifies the directions of maximum variation in the data and aligns the coordinate system with these directions. The principal components (PCs) form a new, rotated coordinate system, with each PC being a linear combination of the original variables. These PCs are orthogonal to each other and are ranked based on the amount of variance they capture in the data.

- Data Transformation Sequence

In this project, our initial process to reduce dimensionality is to z-scale the 29 expert variables, making all the variables within the same scale. Then, we use Principal Component Analysis to rotate the coordinate system, avoiding correlation between the features and also select the top PCs based on Scree Plot. We selected the top 5 PCs that could explain approximately 75% of the variation in the original features. Finally, we z-scaled the 5 PCs once again to ensure that they are on the same scale. This process of z-scale, PCA, then z-scale is crucial for our observation since our scoring metrics take each PCs on the same scale. Different scaling will put more weight on specific PCs, biasing the scoring metric to PCs with larger units. The detailed explanation of the scoring metrics will be explained in the next section.



Part V - Anomaly Detection Algorithms

Recall from the previous part, we acquired the top 5 PCs after the process of z-scale, PCA, then z-scale. The 5 PCs are now mostly non-correlated, similarity scaled, and the mean of the features are centered at 0.

1. Detection Algorithm 1 - Z Scores

With the dimension reduction process, the values of each field display how unusual that record is to that specific dimension. We defined them as "z scores". We created a method based on the concept of Minkowski Distance to add up the "z scores" without canceling out one another.

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

Si = Score for record i
p = Selection for Minkowski Distance
n = Dimensions

By adding up the z scores, the sum will be the total distance away from the center point, which is the "outlierness". The larger the outlierness, the more suspicious the property record is in terms of fraud.

2. Detection Algorithm 2 - AutoEncoder

An AutoEncoder is a model trained to output the original vector input. There is no restriction on which model to implement, but the suggestion is to use a more simpler model. We will use Neural Net for our Auto Encoder with 1 layer 3 nodes. After the model is trained, the difference between the original input vector and the model output vector is the record's fraud score.

The difference between actual "z-score" of the features and the "z-score" outputted from the trained AutoEncoder can be a great metric to measure outlierness since AutoEncoder can reproduce the normal values of the records. The difference can provide explicit meaning to look for anomalies in the data.

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

Si = Score for record i
n = Dimensions,
p = Selection for Minkowski Distance
z′ = AutoEncoder zscore
z = Reocrd acutal zscore

3. Final Scoring Metrics

We used two detection algorithms to score potential anomalies:
Score 1: Based on the sum of Z-scores for each feature.
Score 2: Based on the error between the actual Z-scores and the Z-scores by an AutoEncoder.

Since these two algorithms operate on different scales, we decided to rank the records according to each score. Each record receives:
Rank 1: Based on detection algorithm 1.
Rank 2: Based on detection algorithm 2.

To effectively identify property tax fraud, we select the maximum of the two ranks for each record. This approach allows us to consider anomalies detected by either algorithm, thus maximizing the potential to capture fraudulent activities.

Part VI - Results

- How to implement final fraud scores?

We created two ranks: "score 1 rank" based on z-score outliers and "score 2 rank" from the AutoEncoder results. As mentioned earlier, we determined the final score by selecting the maximum value of these two ranks. By sorting the records by this final score, we generated a file, NY_top_with_zs.xlsx, which contains the top 10,000 suspicious property records. This file also includes the original non-PCA fields for a better observation of the actual values.

However, we found that many anomalies were due to null values in lot size and building size. To address this, we generated two additional files: NY_top_sizes_ne_0.xlsx, excluding records with a lot size of 0, and NY_top_lotsize_ne_0.xlsx, excluding records with null building sizes.

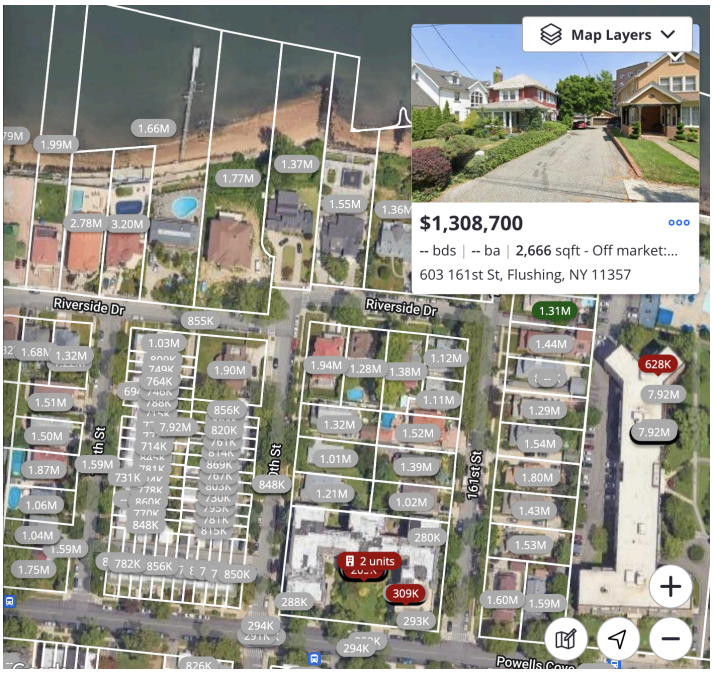
- How to examine suspicious properties?

The data is currently sorted by fraud score, allowing us to investigate each record individually. Our focus is on the "z-score" of 29 expert variables. The "z-score" represents the value of the variables, calculated as the Minkowski Distance from the average. By identifying the outlier z-scores in each record, we aim to understand why a particular property record is considered anomalous.

- Detection Tips

Outlier value	Original Fields to Investigate
r1, r4, r7	Lot Size
r2, r5, r8	Building Size
r3, r6, r9	Stories of Building
r1, r2, r3	FULLVAL
r4, r5, r6	AVLAND
r7, r8, r9	AVTOT
Value Ratio	All Value Variables
Size Ratio	All Area Variables

- **Five Unusal NY Properties**

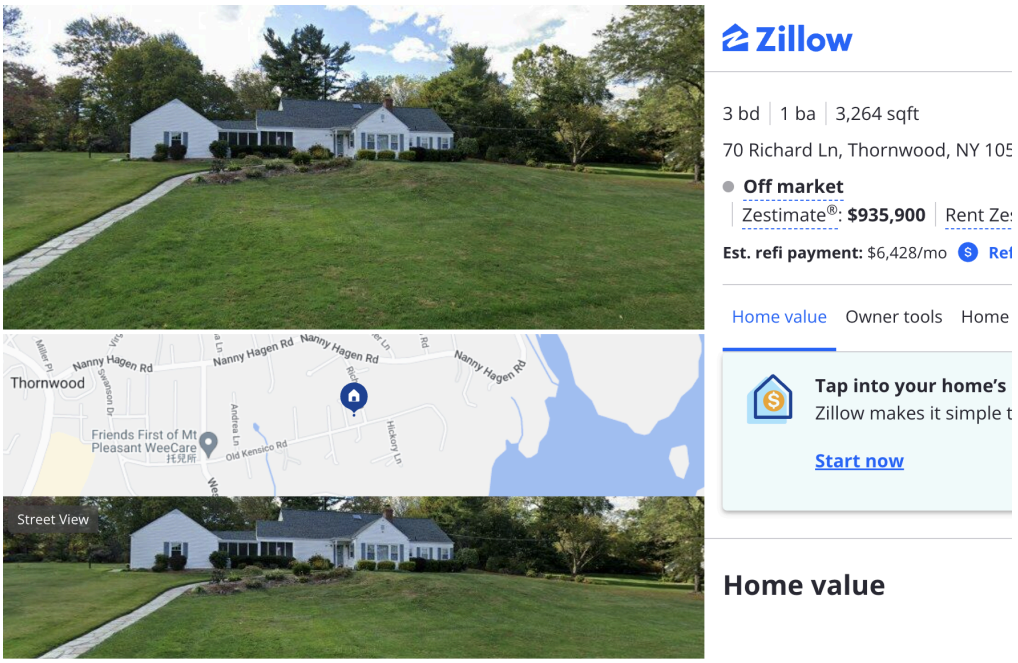


Record: 855940
Owner: DECKER STEVEN
Address: 6-01 161 STREET
Zip Code: 11357
Anomaly Features: r2_zip5, r3_zip5, r5_zip5, r6_zip5, r8_zip5, r9_zip5
Anomaly Fields:

- FULLVAL: $3,130,000
- AVLAND: $69,230
- AVTOT: $70,583
- BLDFRONT: 0
- BLDDEPTH: 0

Doubts:

We doubted this property due to its outlier values r2, r3, r5, r6, r8, r9 in the zip code 11357. This phenomenon occurs when building size or building stories are likely to be different from other buildings in zip code 11357. From the original fields, we can observe that BLDFRONT and BLDDEPTH are 0. However, from the image, there is a building built on the property. Additionally, the value of FULLVAL, AVLAND, and AVTOT are listed highly suspicious, where FULLVAL is too high and AVLAND, AVTOT are too low. To prove the inspection, the market price is currently $1,308,700, according to Zillow, instead of $3,130,000.



GREENTREE ESTATES IN
Address: 70 RICHARD LANE
Zip Code: 10314
Anomaly Features:
r1, r4, r7, r1_zip5, r4_zip5, r7_zip5, r1_taxclass, r4_taxclass, r7_taxclass
Anomaly Fields:

- FULLVAL: $5,710
- AVLAND: $117
- AVTOT: $120
- BLDFRONT: 13
- BLDDEPTH: 30
- LTFRONT: 184 (Should be accurate)
- LTDEPTH: 138 (Should be accurate)

Doubts:

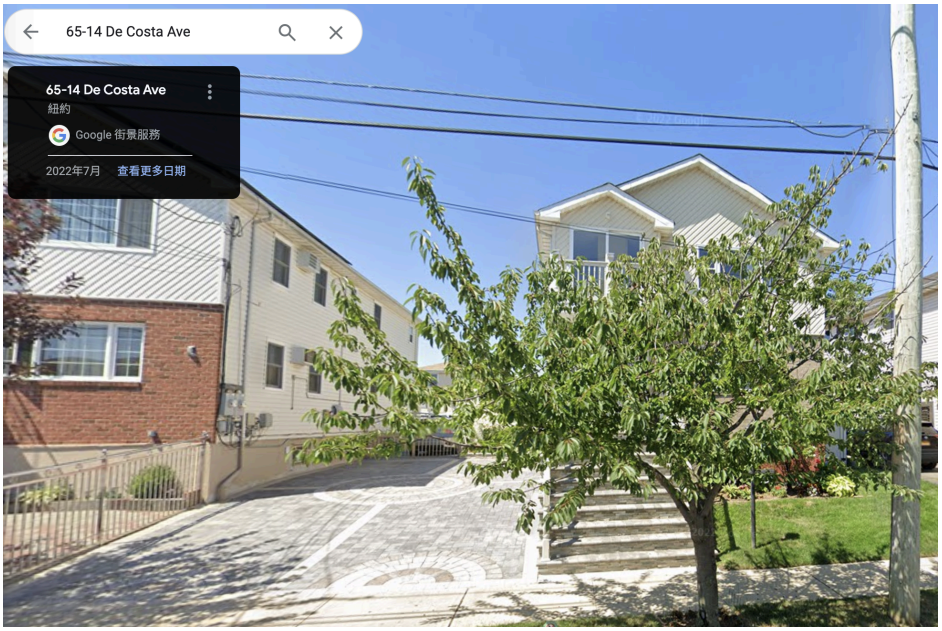
Obviously, The FULLVAL, AVLAND, and AVTOT values are significantly inaccurate compared to the actual value of the property. The current sale listing on Zillow confirms that the property's value was not entered correctly. The anomalies in rows r1, r4, and r7 are due to erroneous entries for the building size, though the lot size appears to be accurate.

Record: 106681
Owner: 79TH REALTY LLC
Address: 350 EAST 79 STREET
Zip Code: 10075
Anomaly Features:
r1, r4, r7, r4_zip5, r7_zip5, r4_taxclass, r7_taxclass
Anomaly Fields:

- LTFRONT: 25
- LTDEPTH: 100
- BLDFRONT: 25
- BLDDEPTH: 100
- FULLVAL: $114,000,000

Doubts:

This is a real estate enterprise’s corporate building that consists of 44 floors. The lot area and building area are not compelling for us to believe that these are accurate values of the property information. Not to mention the property images from google maps are nowhere close to the property’s area data.

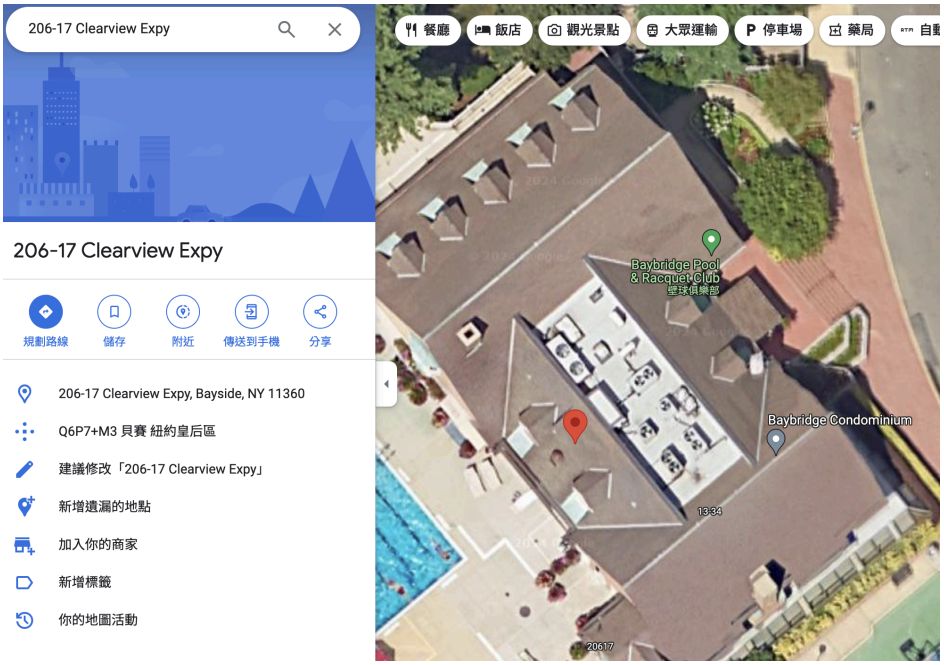


Record: 927414
Owner: ROBINSON-HUGGINS PATR
Address: 65-14 DE COSTA AVENUE
Zip Code: 11692
Anomaly Features:
r1, r4, r7, r1_zip5, r4_zip5, r7_zip5, r1_taxclass, r4_taxclass, r7_taxclass
Anomaly Fields:

- LTFRONT: 4,318
- LTDEPTH: 100

Doubts:

The anomalies in the record indicate that there is an issue with the lot size of the property. The image of the property does not seem to encompass 4,318 ft. Hence, the record is mistakenly inputted and should be altered.



Record: 735653
Owner: BAYBRIDGE HOA
Address: 206-17 CLEARVIEW EXPRESSWAY
Zip Code: 11692
Anomaly Features:
r1, r4, r7, r1_zip5, r4_zip5, r7_zip5, r1_taxclass, r4_taxclass, r7_taxclass
Anomaly Fields:

- FULLVAL: $ 2,160
- AVLAND: $ 495
- AVTOT: $ 972

Doubts:

This is a club that has 318 ft * 174 ft of lot size. However, the three values of the record do not match the value in reality. The owner could misinterpret the unit of the dollar value. This record of the property should be further fixed.

.

Part VII - Summary

In this project, we aimed to identify property tax fraud within New York City's extensive property dataset. The steps taken included:

1. Data Cleaning: We removed irrelevant government property records and imputed null values using targeted methodologies.
2. Variable Creation: We generated new variables to capture property value, area metrics, and ratios that could indicate anomalies.
3. Dimensionality Reduction: Using Principal Component Analysis (PCA), we reduced the dimensionality of the data to five principal components, ensuring they were orthogonal and similarly scaled.
4. Anomaly Detection Algorithms: We implemented two algorithms:
   Algorithm 1: Calculated a score based on the sum of Z-scores for each feature.
   Algorithm 2: Used an AutoEncoder to measure the error between actual Z-scores and predicted Z-scores.
   We ranked the records based on these scores and selected the maximum rank from the two algorithms to identify anomalies.
   The results were summarized in the NY_top_with_zs.xlsx file, containing the top 10,000 suspicious property records. Additionally, due to a significant number of anomalies arising from null values in lot and building sizes, we produced two supplementary files: NY_top_sizes_ne_0.xlsx and NY_top_lotsize_ne_0.xlsx.

To adjust the algorithm with expert feedback, one can modify the variables and exclusions based on new insights. This could involve refining the imputation methods, adjusting the criteria for outlier detection, or incorporating additional data fields. The overall methodology remains robust, ensuring comprehensive detection of property tax fraud.

Appendix - Data Quality Report

1. Data Description

The dataset records Property Valuation and Assessment data in New York City. The data encompasses 32 detailed description fields of each property in New York City yet without a label. The 32 fields are built by 14 numeric fields and 18 categorical fields. The data consist of 10,70,994 property records.

2. Summary Tables

Numeric Fields Table

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

Categorical Fields Table

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

3. Visualization of Each Field

Categorical:

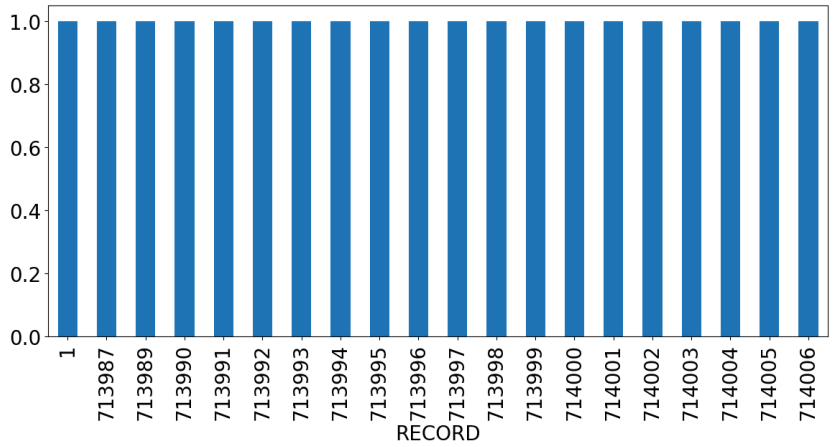
1) **Field Name: Record**

Description: The record number for each property.

# of Unique Values: 1,070,994

Null Values: 0

Observation: One unique record number for each property.



2) **Field Name: BBLE**

Description: BBLE stands for Borough, Block, Lot, and Easement. It is a combination of digits representing each of these elements to uniquely identify a property.

# of Unique Values: 1,070,994

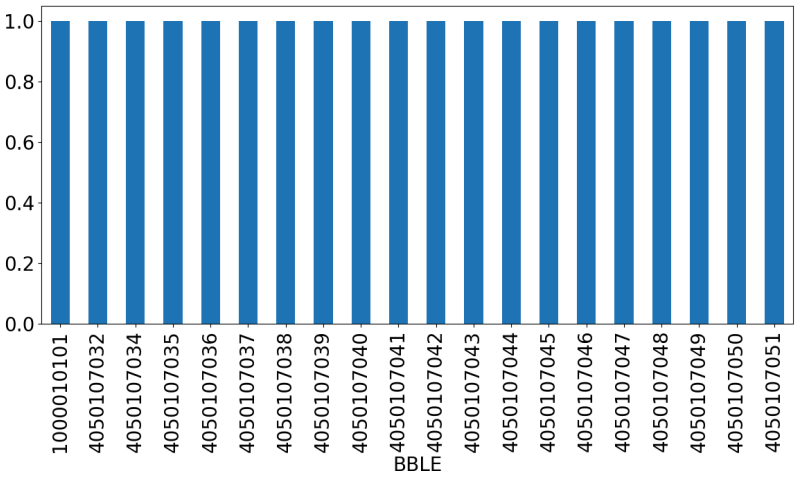
Null Values: 0

Observation: One unique BBLE for each property.

Naming Rule:

1st Digit: Boro; 2 to 6th Digit: Block; 7 to 10th Digit: Lot; Last: Easement

Ex: 1(Boro)08050(Block)0038(LOT)E(EASEMENT)



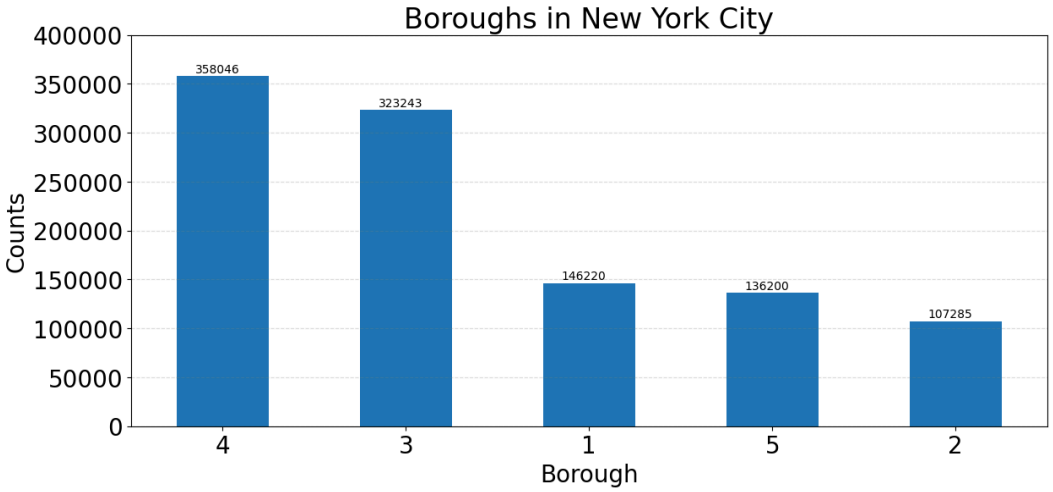
3) **Field Name: BORO**

Description: Boroughs in New York City. Digits represent each borough. 1=Manhattan; 2 = Bronx; 3 = Brooklyn; 4 = Queens; 5 = Staten Island

# of Unique Values: 5

Null Values: 0

Observation: Brooklyn and Queens encompass most of the properties in New York City while Manhattan, Bronx, and Staten Island have fewer.



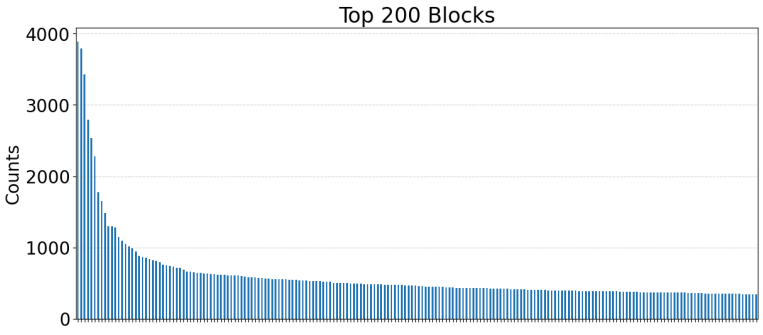
4) **Field Name: BLOCK**

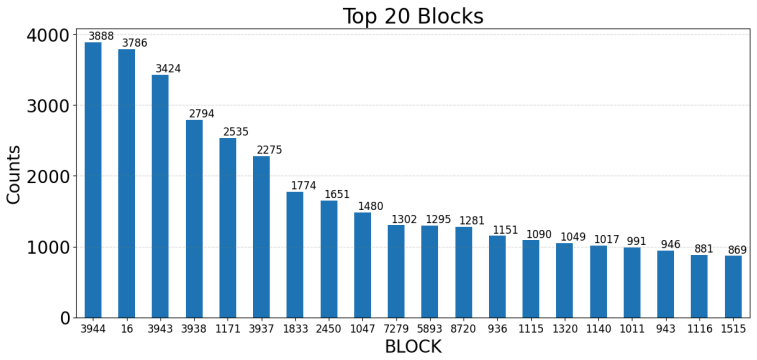
Description: Valid block ranges by borough

# of Unique Values: 13,894

Null Values: 0

Observation: There are 13,894 unique block numbers. The distribution of the block (left plot) is highly right-skewed, with most of the blocks encompassing less than 500 properties. The right plot displays the top 20 blocks in terms of count. The block with the most properties is 3944 with 3,888 blocks, 2nd is Block 16 with 3,796 properties, 3rd is Block 3943 with 3,424 properties.





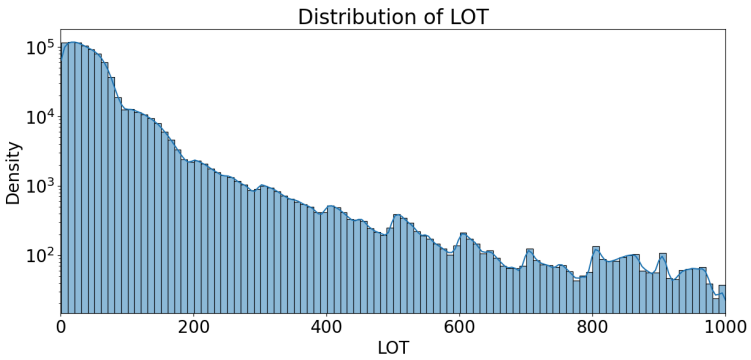
5) Field Name: LOT

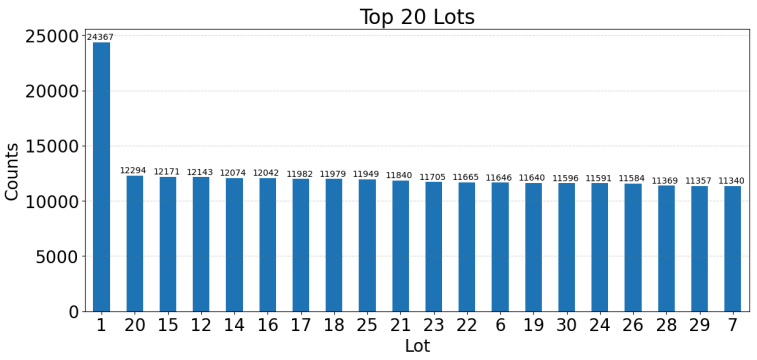
Description: Valid Lot ranges by borough

# of Unique Values: 6,366

Null Values: 0

Observation: There are 6,366 unique lot numbers. The distribution of the lot (left plot) is highly right-skewed. The largest lot in terms of size is lot 1, with 24,367 properties. The nature of this field is that this field ranges from 1 to 9,978 and decreases as the lot number increases.





6) Field Name: EASEMENT

Description: Representations of different easement types.

Space = No Easement A = Air Easement

B = Non-Air Rights E = Land Easement

N = Non-Transit Easement

P = Pier R = Railroad

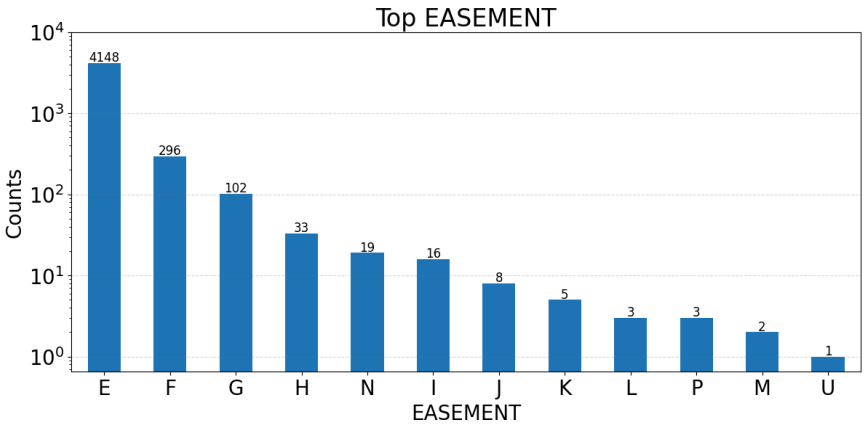
S = Street U = U.S. Government

F Thru M = duplicates of E

# of Unique Values: 13

Null Values: 9,258

Observation: Most of the properties are without easement, with 9,258 properties that are null records. For properties with easement, most of the data are with land easements.



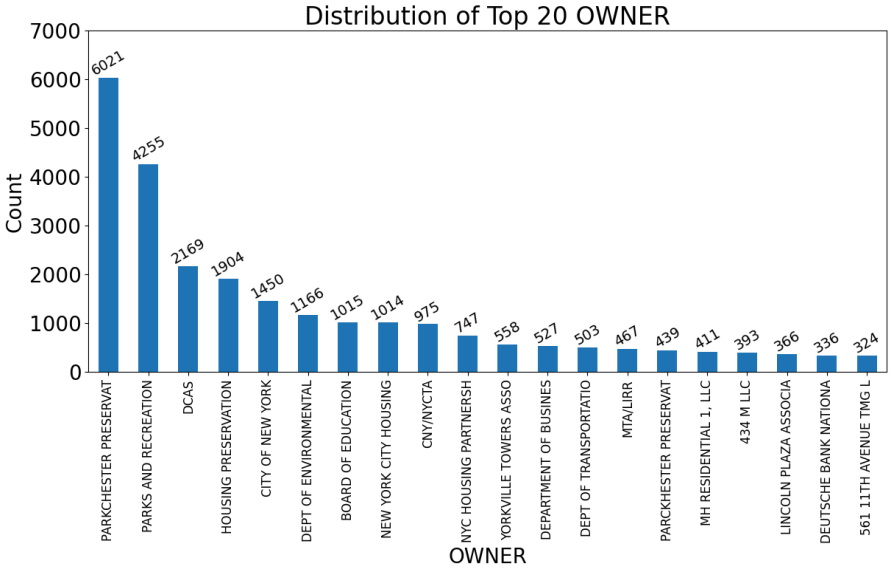
7) **Field Name: OWNER**

Description: Name of the property owner

# of Unique Values: 863,348

Null Values: 31,745

Observation: The Owner distribution is highly right-skewed since most of the inhibitors do not have more than 1 property. This plot shows the top 20 owners in terms of number of property owned. The owner with the most properties in New York City is Parkchester Preservat, with 6,021 properties, then it's Parks and Recreation, with 4,255 properties.



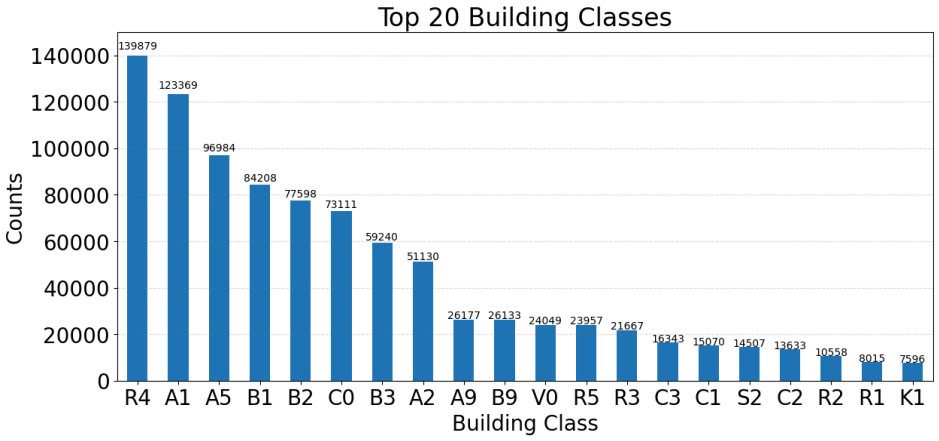
8) **Field Name: BLDCL**

Description: Building Class, a shorthand description that can provide a potential investor with information about its age, location, amenities, condition, rental rates, and sales price.

# of Unique Values: 863,348

Null Values: 0

Observation: The class of buildings distribution is highly right-skewed. Most of the buildings are of R4, A1, A5, B1, B2, C0, with R4 as the main class.



9) **Field Name: TAXCLASS**

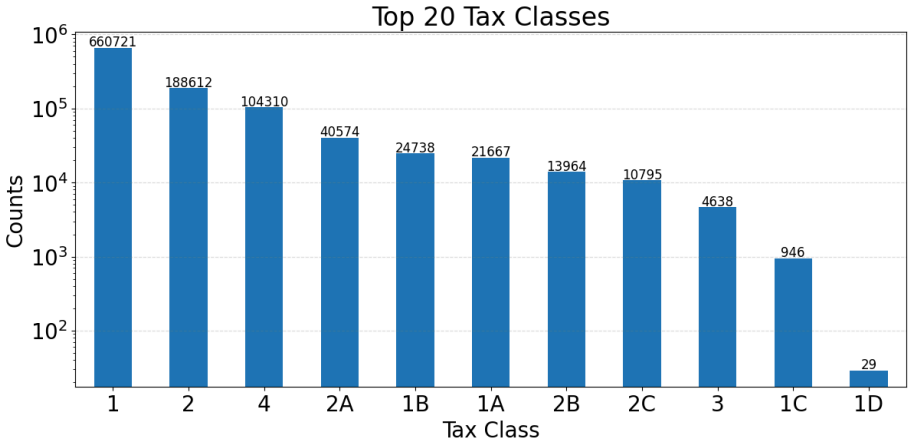
Description: The tax class of the property.

1= 1 to 3 unit residence; 2 = Apartments; 2A = 4, 5, or 6 Units;
3 = Utilities; 4 = All Others

# of Unique Values: 11

Null Values: 0

Observation: The tax class is composed of Class 1 and Class 2. The higher the Class number is the lower the property count it encompasses.



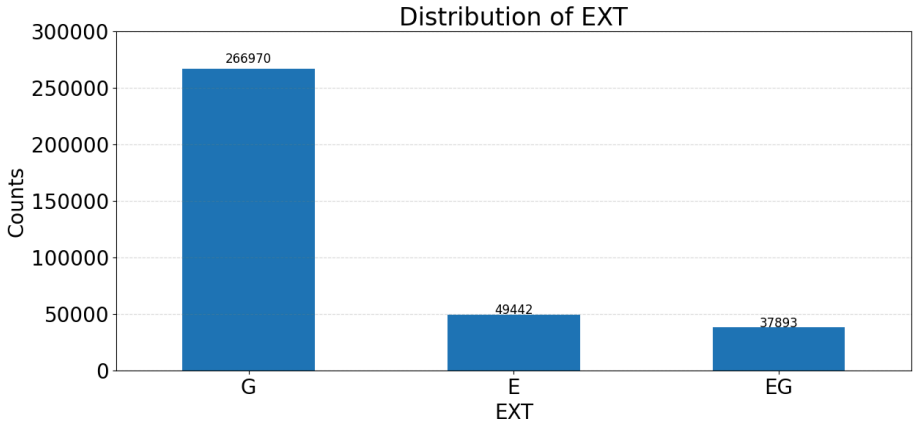
10) **Field Name: EXT**

Description: Extension Indicator

# of Unique Values: 4

Null Values: 716,689

Observation: Category G consists of 266,970 properties. Category E consists of 49,442 properties. Category EG consists of 37,893 properties.



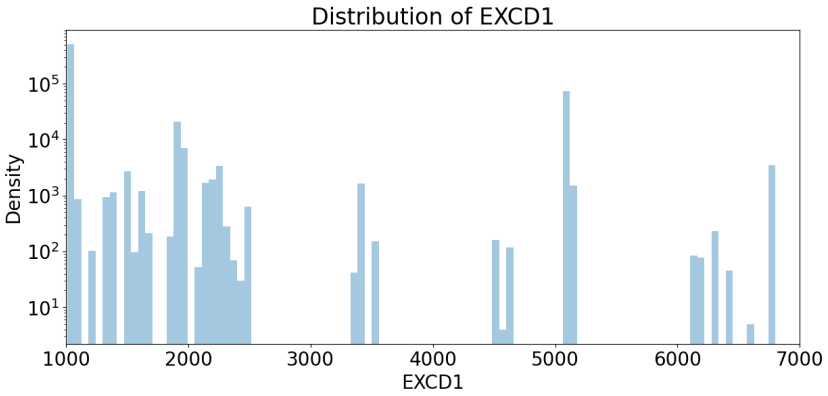
11) Field Name: EXCD1

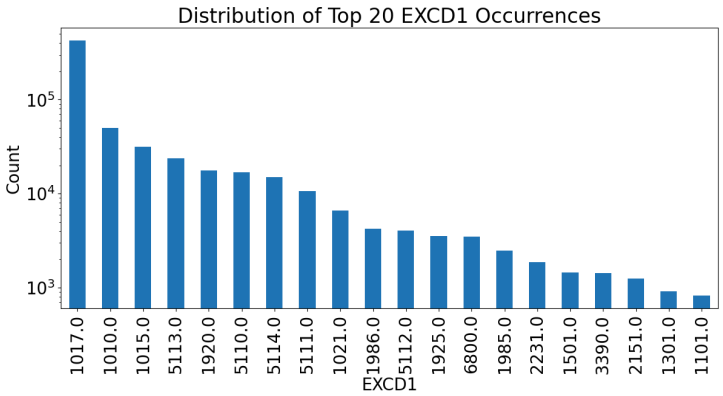
Description: Exemption code 1 of the property

# of Unique Values: 130

Null Values: 432,506

Observation: The top 4 EXCD1 with the most properties are: 1017: 425,348; 1010: 49,756; 1015: 31,323; 5113: 23,858; 1920: 17,594.





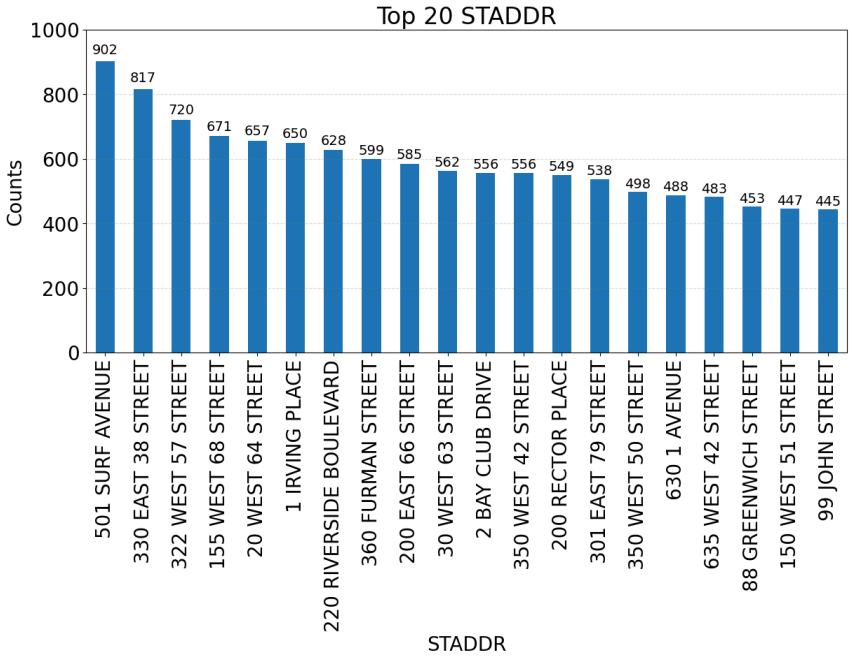
12) Field Name: STADDR

Description: Street Address of the property

# of Unique Values: 839,281

Null Values: 676

Observation: With multi-story buildings or apartments, properties have identical addresses. However, most of the properties tend to have unique addresses.



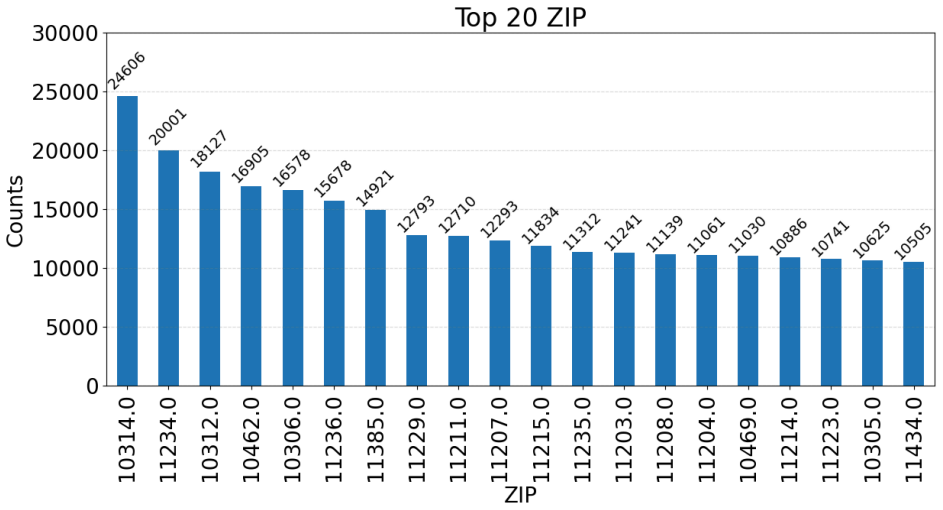
13) Field Name: ZIP

Description: Zip code of the property

# of Unique Values: 197

Null Values: 29,890

Observation: Top 3 Zip code: Zip code 10314 includes 24,6060 properties; Zip code 11234 includes 20,001 properties; Zip code 10312 includes 18,127 properties



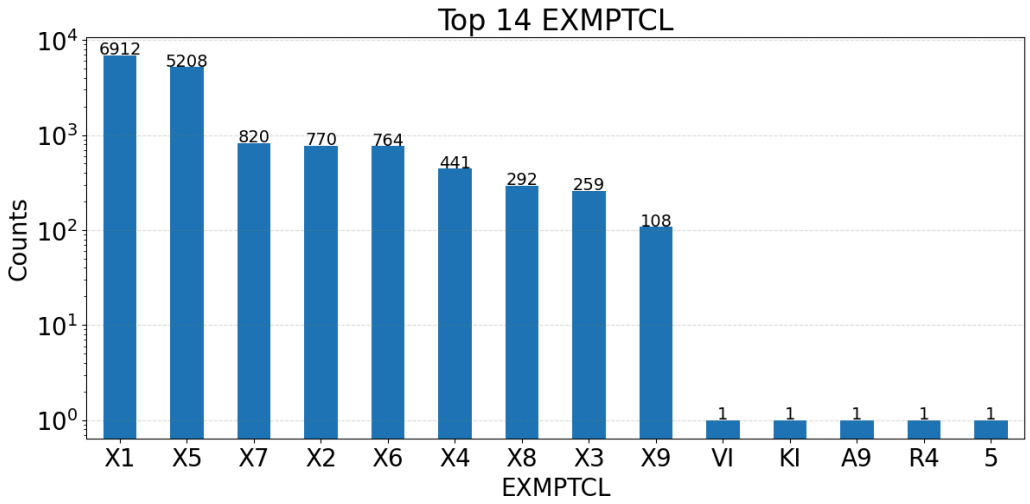
14) Field Name: EXMPTCL

Description: Exemption class of the property

# of Unique Values: 15

Null Values: 1,055,415

Observation: Most of the class started with initial X, with X1 and X5 as the most properties' exemption class.



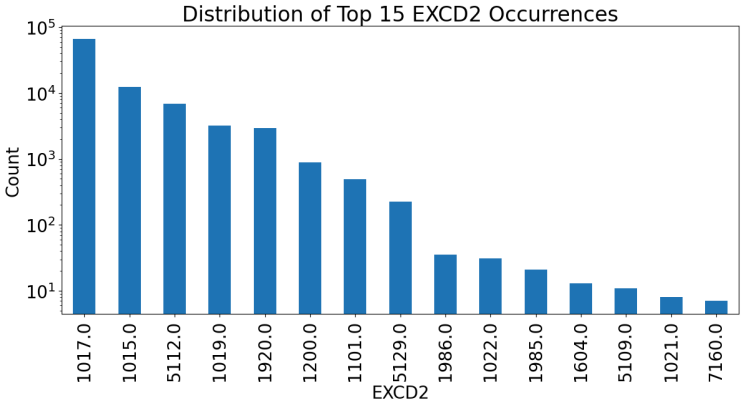
15) Field Name: EXCD2

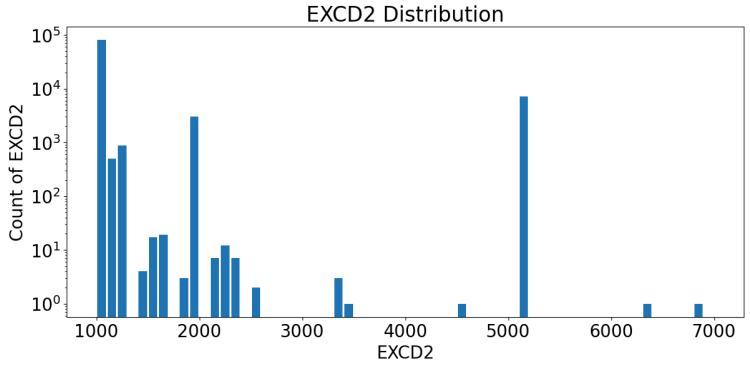
Description: Exemption class 2 of the property

# of Unique Values: 61

Null Values: 978,046

Observation: Properties mostly match to the EXCD2 interval based on characteristics.The top 3 EXCD2 values based on property count: 1017: 65,777; 1015: 12,337; 5112: 6867.





16) Field Name: PERIOD

Description: Assessment period

# of Unique Values: 1

Null Values: 0

Observation: All the data are with the same value: ‘FINAL PERIOD’. No plot is needed for this field.

17) Field Name: YEAR

Description: Assessment year

# of Unique Values: 1

Null Values: 0

Observation: All the data are with the same value: ‘2010/11 YEAR’. No plot is needed for this field.

18) Field Name: VALTYPE

Description: Valuation type

# of Unique Values: 1

Null Values: 0

Observation: All the data are with the same value: ‘AC-TR VALTYPE’. No plot is needed for this field.

Numeric:

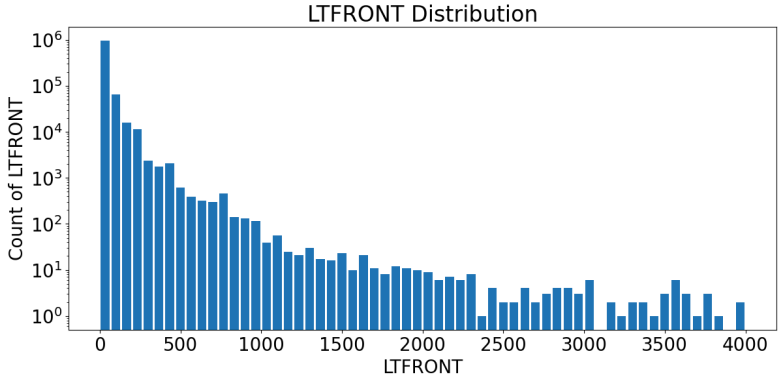
19) **Field Name: LTFRONT**

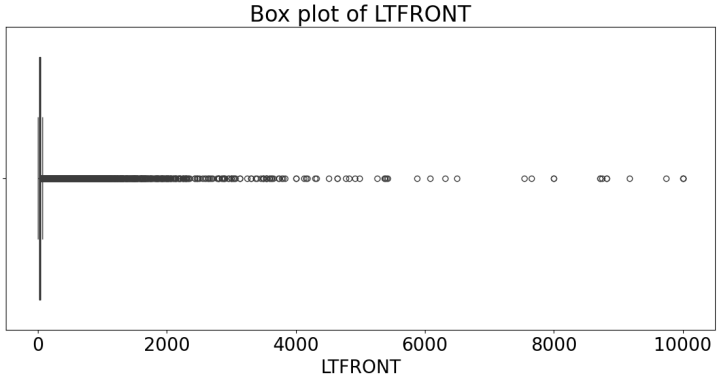
Description: Lot width

# of Unique Values: 1,297

Null Values: 0

Observation: Field ranging from 0~9,999. According to the top distribution plot, the data is highly right-skewed. Most of the properties are with relatively small lot width. If we look closer at the distribution of small width sizes(bottom figures), the data contains a large number of properties with 0 lot width. Values larger than 258.74 are considered as outliers, with a count of 9,722.





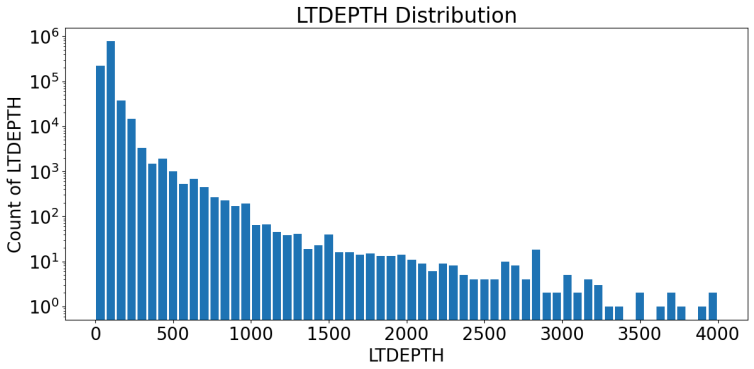
20) Field Name: LTDEPTH

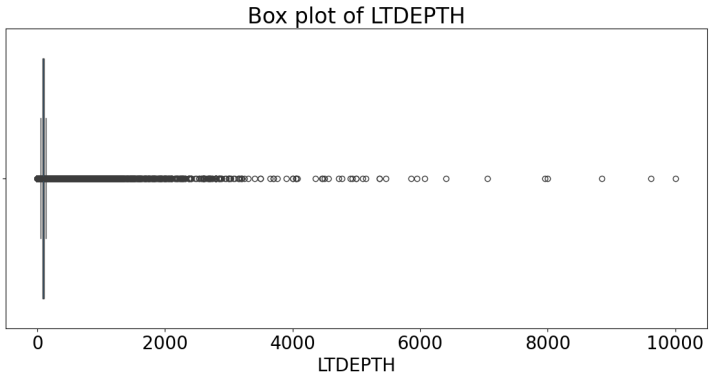
Description: Lot depth

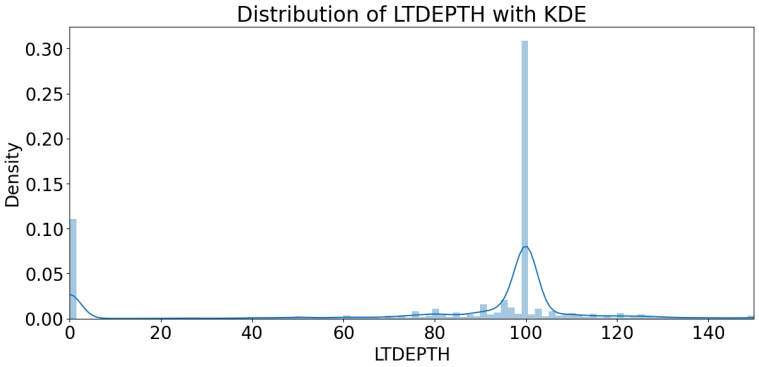
# of Unique Values: 1,370

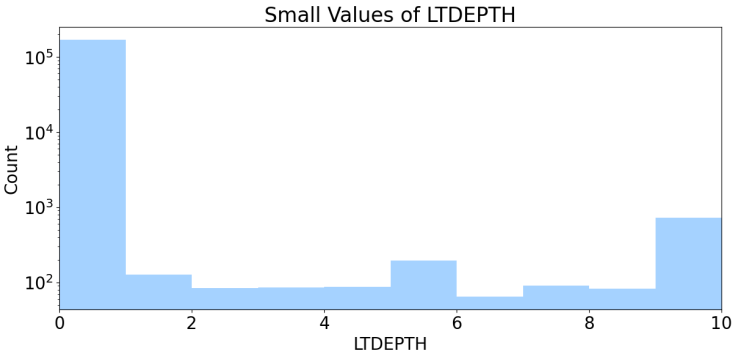
Null Values: 0

Observation: Field ranging from 0~9,999. According to the top distribution plot, the data is highly right-skewed. Most of the properties are with relatively small lot width. If we look closer at the distribution of small depth sizes(bottom figures), various properties are with 100 lot width. Values larger than 318.05 are considered as outliers, with a count of 8,347.









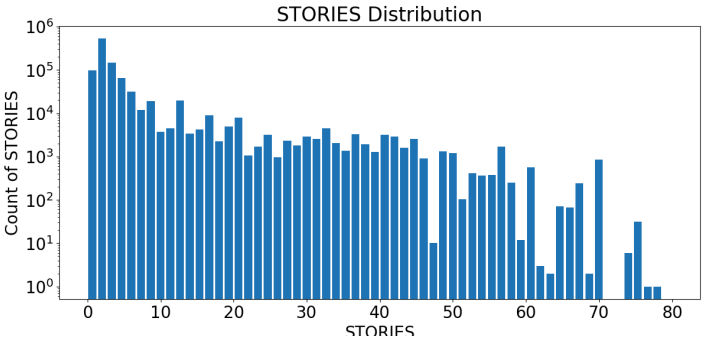
21) Field Name: STORIES

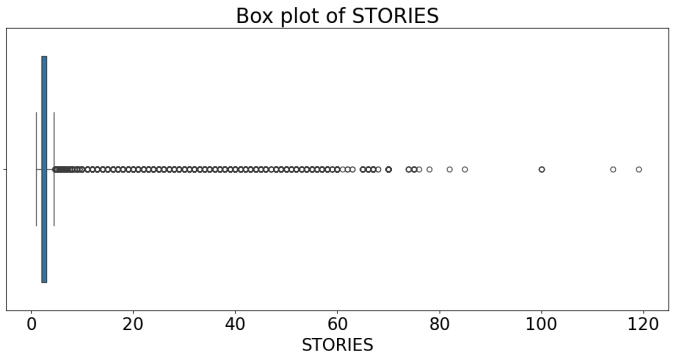
Description: Properties story

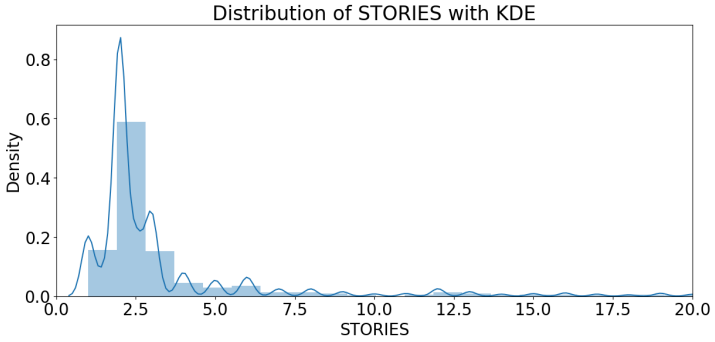
# of Unique Values: 112

Null Values: 56,264

Observation: Values range from 1 to 119 stories. The distribution is right-skewed, meaning that fewer properties are with high stories. From the bottom figure, we can observe that most of the properties are 1 to 3 stories high. Story values higher than 30 are considered as outliers, encompassing 35,785 properties.







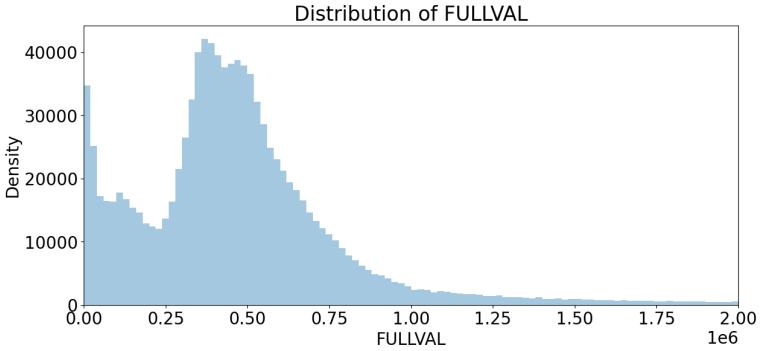
22) Field Name: FULLVAL

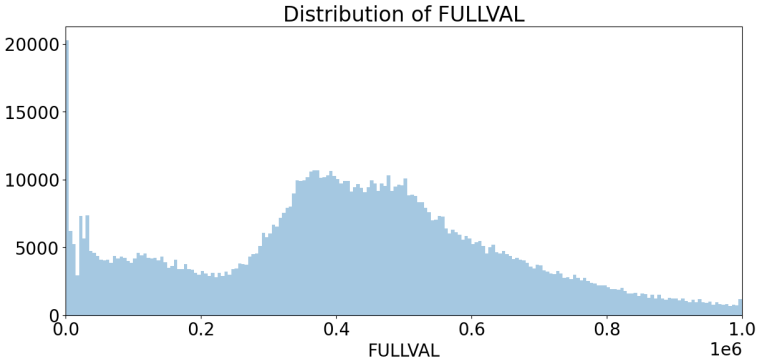
Description: Market value of the property

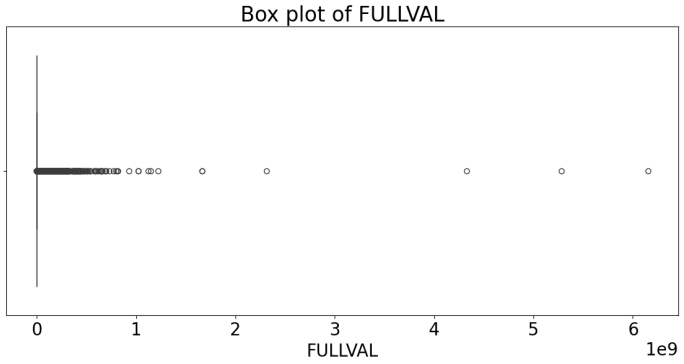
# of Unique Values: 109,324

Null Values: 0

Observation: This field is bimodal distributed. The reason for this distribution is that 13,007 properties recorded as $0 market value. In the nature of this field, most of the values are valued around $30,000 to 50,000 dollars. Values more than $35,621,557 are considered outliers, with 1,919 properties.







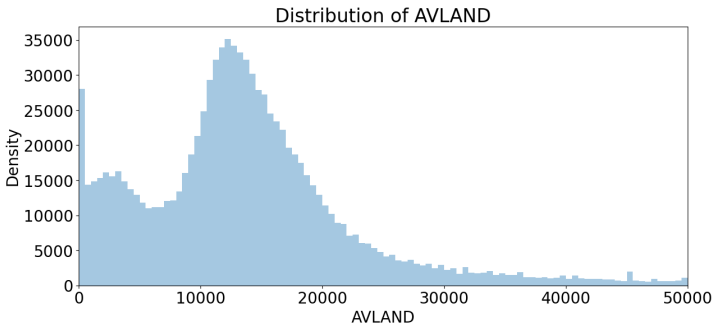
23) Field Name: AVLAND

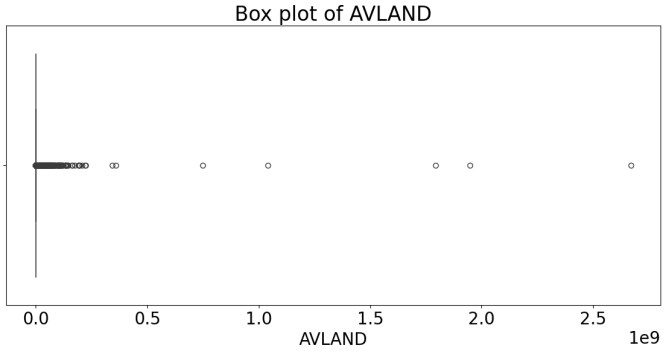
Description: Actual land value of the property

# of Unique Values: 70,921

Null Values: 0

Observation: This field is bimodal distributed. The reason for this distribution is that 13,009 properties recorded as $0 land value. In the nature of this field, most of the values are valued around $10,000 to 20,000 dollars. Values more than $12,256,848 are considered outliers, with 799 properties.





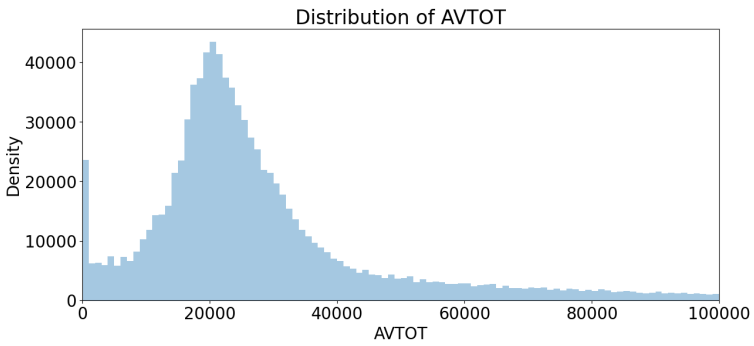
24) Field Name: AVTOT

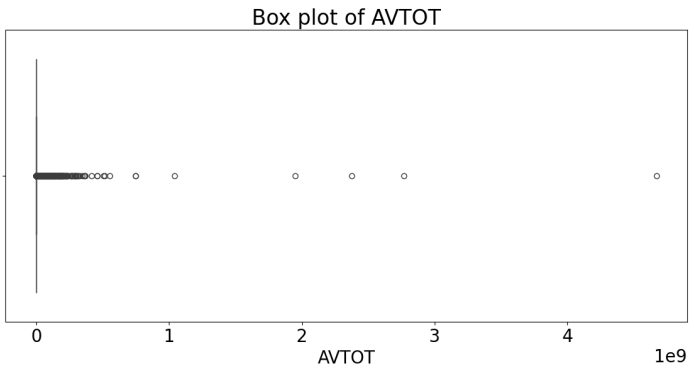
Description: Actual total value of the property

# of Unique Values: 112,914

Null Values: 0

Observation: This field is bimodal distributed. The reason for this distribution is that 13,007 properties recorded as $0 total value. In the nature of this field, most of the values are valued around $20,000 dollars. Values more than $20,859,826 are considered outliers, with 1,429 properties.





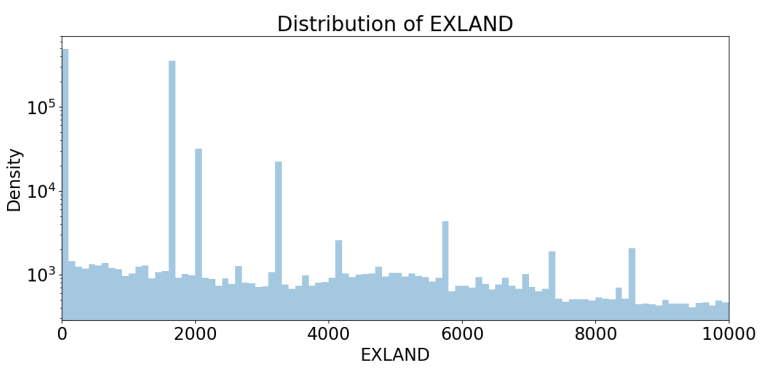
25) Field Name: EXLAND

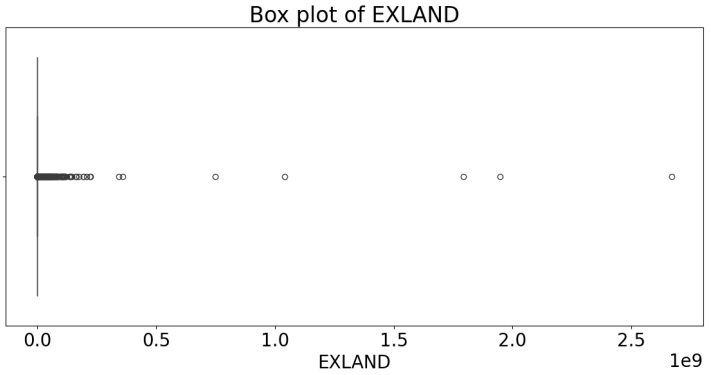
Description: Actual exempt land value of the property

# of Unique Values: 33,419

Null Values: 0

Observation: This field is multi-modal distributed. The top 4 values with the most properties are: $0: 491,699; $1,620 357,182; $2,090: 31,112; $3240: 21,519. Values more than $11,981,151 are considered outliers, with 394 properties.





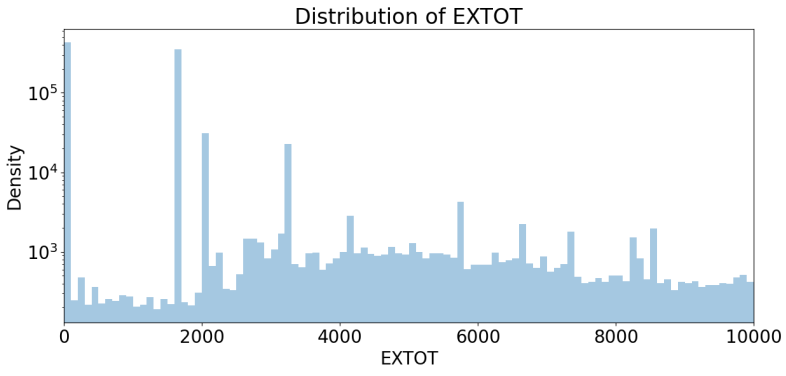
26) Field Name: EXTOT

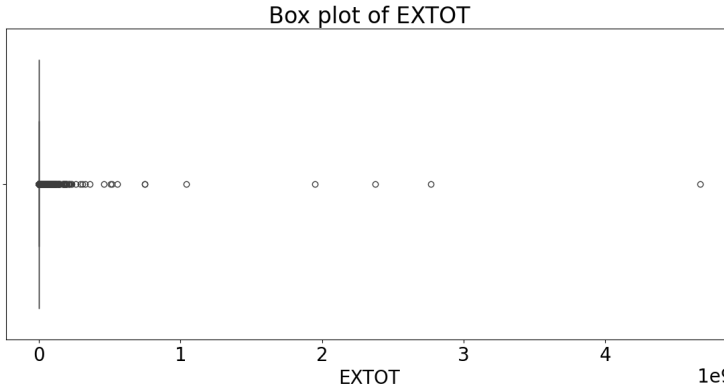
Description: Actual exempt total value of the property

# of Unique Values: 64,255

Null Values: 0

Observation: This field is multi-modal distributed. The top 4 values with the most properties are: $0: 432,572; $1,620 354,880; $2,090: 30,069; $3240: 21,803. Values more than $19616395 are considered outliers, with 669 properties.





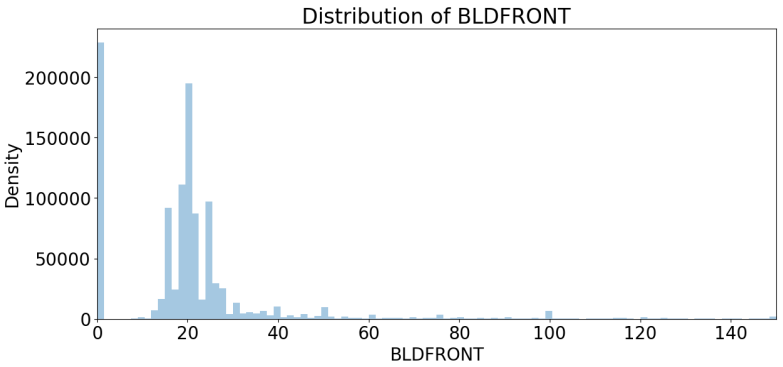
27) Field Name: BLDFRONT

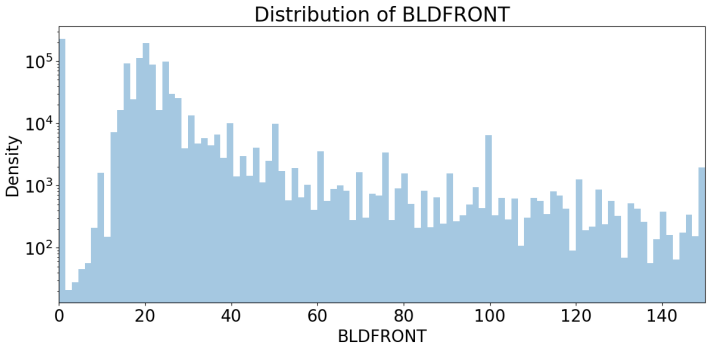
Description: The building width of the property

# of Unique Values: 612

Null Values: 0

Observation: The distribution of this field is highly right-skewed, but noting that there are 228,815 properties with value of 0, forming a bimodal-like distribution. Neglecting records of 0s, the distribution is concentrated around 20. Values higher than 129.78 are considered as outliers, with a total of 18,922 properties.





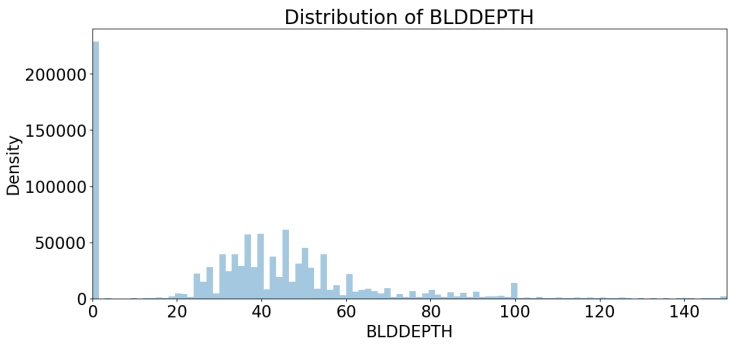
28) Field Name: BLDDEPTH

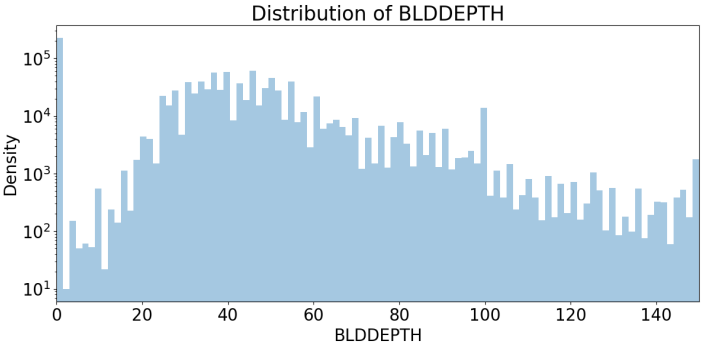
Description: The building depth of the property

# of Unique Values: 621

Null Values: 0

Observation: The distribution of this field is highly right-skewed, but noting that there are 228,853 properties with value of 0, forming a bimodal-like distribution. Neglecting records of 0s, the distribution is concentrated around 30 to 60. Values higher than 168.04 are considered as outliers, with a total of 10,647 properties.





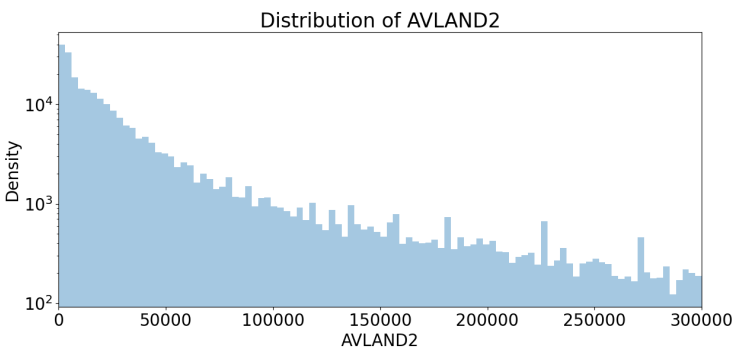
29) Field Name: AVLAND2

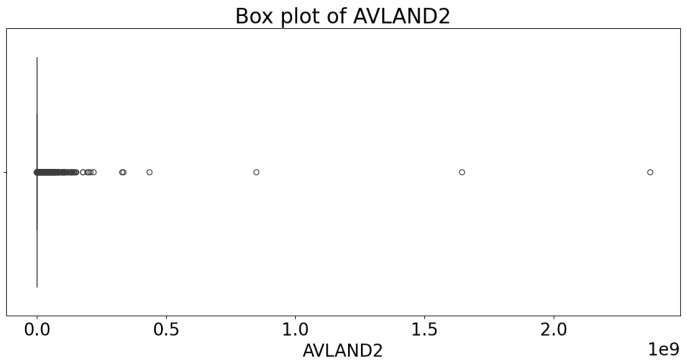
Description: The transitional land value of the property

# of Unique Values: 58,592

Null Values: 788,268

Observation: The distribution of this field is right-skewed, meaning that most of the properties are distributed of relatively low values. Properties worth more than 18,783,123 are considered outliers, with a total of 425 properties.





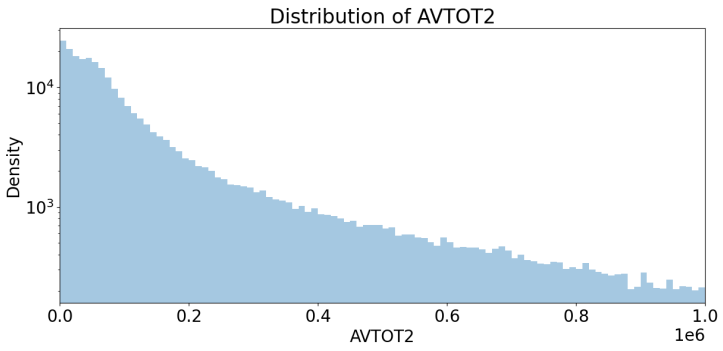
30) Field Name: AVTOT2

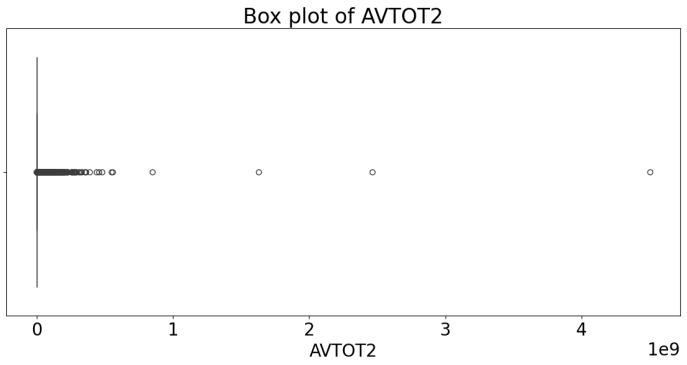
Description: The transitional total value of the property

# of Unique Values: 111,361

Null Values: 788,268

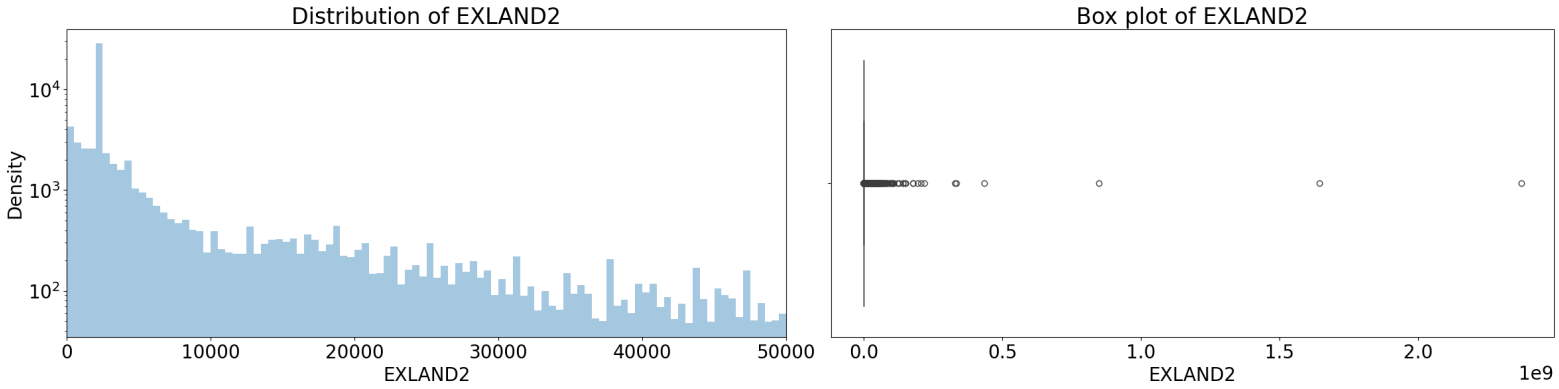
Observation: The distribution of this field is right-skewed, meaning that most of the properties are distributed of relatively low values. Properties worth more than $35,671,498 are considered outliers, with a total of 689 properties.





31) Field Name: EXLAND2

Description: The transitional exemption land value of the property
# of Unique Values: 22,196
Null Values: 983,545
Observation: The peak of the distribution is $2,090, with 26,393 properties. Otherwise, the rest of the field distribution is right-skewed. Properties with EXLAND2 higher than $32,757,873 are considered outliers, with a total of 105 properties.



32) Field Name: EXTOT2

Description: The transitional exemption total value of the property
# of Unique Values: 48,349
Null Values: 940,166
Observation: The peak of the distribution is $2,090, with 24,739 properties. Otherwise, the rest of the field distribution is right-skewed. Properties with EXTOT2 higher than $48,874,298 are considered outliers, with a total of 183 properties.