Informal Business Report

Sheetal Srivastava

1. Executive Summary

This report presents the findings from the analysis of NY Property dataset which consists of real estate property data for the city of New York provided by the Department of Finance, spanning over a couple of years 2010-11. The objective of the analysis is to build an unsupervised model that can flag unusual property records in the given dataset of about a million records, with the intention of flagging potential tax fraud cases for further manual investigation.

The analysis involved several steps, including data cleaning, imputations, variable creation, dimensionality reduction, and the calculation of anomaly scores. The results of the analysis is a list of properties ranked based on these anomaly scores in descending order, identifying the top most unusual properties for further investigation. The analysis provides a valuable tool for the client in their efforts to detect underpayment of taxes resulting from misrepresented property characteristics and enables targeted scrutiny of the flagged properties. On manual investigation of some of the top unusual properties we find that while some of these properties may be flagged due to rarity of occurrence of certain characteristics (for example very high market value for a small upscale shopping center) there are some properties that do require further investigation and could be potential tax fraud records.

2. Data Description

The dataset is **NY Property Data**, which contains property valuation and assessment data for the city of New York provided by the Department of Finance. The data consists of real estate assessment property data coming from the New York city government. The data is updated annually and spans over the time period of a couple of years, 2010/11. There are **32 fields** and **1070994 records**.

a. Numeric Fields Table

	Field Name	Field Type	# Records Have Values	% Populated	% Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	LTFRONT	numeric	1070994	100.00%	15.79%	0	9999	36.64	74.03	0
1	LTDEPTH	numeric	1070994	100.00%	15.89%	0	9999	88.86	76.40	100
2	STORIES	numeric	1014730	94.75%	0.00%	1	119	5.01	8.37	2
3	FULLVAL	numeric	1070994	100.00%	1.21%	0	6150000000	874,264.51	11,582,425.58	0
4	AVLAND	numeric	1070994	100.00%	1.21%	0	2668500000	85,067.92	4,057,258.16	0
5	AVTOT	numeric	1070994	100.00%	1.21%	0	4668308947	227,238.17	6,877,526.09	0
6	EXLAND	numeric	1070994	100.00%	45.91%	0	2668500000	36,423.89	3,981,573.93	0
7	EXTOT	numeric	1070994	100.00%	40.39%	0	4668308947	91,186.98	6,508,399.78	0
8	BLDFRONT	numeric	1070994	100.00%	21.36%	0	7575	23.04	35.58	0
9	BLDDEPTH	numeric	1070994	100.00%	21.37%	0	9393	39.92	42.71	0
10	AVLAND2	numeric	282726	26.40%	0.00%	3	2371005000	246,235.72	6,178,951.64	2408
11	AVTOT2	numeric	282732	26.40%	0.00%	3	4501180002	713,911.44	11,652,508.34	750
12	EXLAND2	numeric	87449	8.17%	0.00%	1	2371005000	351,235.68	10,802,150.91	2090
13	EXTOT2	numeric	130828	12.22%	0.00%	7	4501180002	656,768.28	16,072,448.75	2090

b. Categorical Fields Table

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	RECORD	categorical	1070994	100.00%	0	1070994	1
1	BBLE	categorical	1070994	100.00%	0	1070994	1000010101
2	BORO	categorical	1070994	100.00%	0	5	4
3	BLOCK	categorical	1070994	100.00%	0	13984	3944
4	LOT	categorical	1070994	100.00%	0	6366	1
5	EASEMENT	categorical	4636	0.43%	0	12	E
6	OWNER	categorical	1039249	97.04%	0	863347	PARKCHESTER PRESERVAT
7	BLDGCL	categorical	1070994	100.00%	0	200	R4
8	TAXCLASS	categorical	1070994	100.00%	0	11	1
9	EXT	categorical	354305	33.08%	0	3	G
10	EXCD1	categorical	638488	59.62%	0	129	1017
11	STADDR	categorical	1070318	99.94%	0	839280	501 SURF AVENUE
12	ZIP	categorical	1041104	97.21%	0	196	10314
13	EXMPTCL	categorical	15579	1.45%	0	14	X1
14	EXCD2	categorical	92948	8.68%	0	60	1017
15	PERIOD	categorical	1070994	100.00%	0	1	FINAL
16	YEAR	categorical	1070994	100.00%	0	1	2010/11
17	VALTYPE	categorical	1070994	100.00%	0	1	AC-TR

3. Data Cleaning

- **Exclusions Logic**

- OWNER : Identify and exclude some properties owned by government bodies, public institutions, and cemeteries. We are not interested in these estates for our analysis since these entities often have different tax and property rules that may not align with the analysis. This removes about 25,000 property records.

- FULLVAL, AVLAND, AVTOT : Exclude values that are zero. For the real estate dataset, if FULLVAL (full market value), AVLAND (assessed land value), and AVTOT (total assessed value) are zero, this suggests some sort of error in the data collection process, and does not give us any information on the property. The values for these variables should be strictly greater than zero.

- LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH : Exclude values equal to zero or greater than 10,000. Again, zero values or extremely large values for these parameters are likely to be errors, since they represent the physical dimensions of the property.

- STORIES : Exclude values that are either zero or null. This is important because the number of stories is a key factor affecting property value. This value cannot be zero as the minimum number of stories any property has is one.

- TAXCLASS : Different tax classes have different assessment rules, so mixing properties from different classes could distort the analysis. Excluding certain tax classes helps to ensure that the analysis is comparing like with like.

- r1 to r9 : Exclude values greater than 1e9. These ratios (r1 to r9) are calculated for standardization and comparison across different groupings. If a ratio exceeds 1e9, it suggests a strong deviation from the mean, which could lead to distortion.

- FULLVAL / AVLAND / AVTOT - per square foot : Exclude values greater than $10,000. Property value per square foot is a common measure of estate valuation and values greater than $10,000 are likely to be outliers that could distort the analysis. These could reflect as an extremely high-value property, but could also be erroneous or spurious data points.

- EASEMENT : If the property has an easement, it means that someone else has a right to use the property for a specific purpose, which could potentially affect the property's value and the owner's ability to use or modify the property. If many properties in the dataset have easements, it might be useful to exclude these from a general analysis of property values.

- **Imputations Logic**

- ZIP : Zip codes are a valuable attribute of a property. Since we have 20431 missing values for this variable we follow a 3-step approach for the imputation:
    i. Concatenate the 'staddr' and 'boro' columns into a new 'staddr_boro' column to replace missing zip values. This fills about 2832 values.
    ii. Assume the data is already sorted by zip. If a zip is missing, and the before and after zips are the same, fill in the zip with that value. This fills 9491 values.
    iii. For the remaining missing zips, just fill in with the previous record's zip.

- FULLVAL, AVLAND, and AVTOT : The code fills in missing or zero values in these fields with the median value of the non-missing or non-zero values in the same tax class. The idea behind this is that properties in the same tax class should have similar characteristics and, consequently, similar market and assessed values. Imputing with the median (instead of the mean, for example) helps mitigate the impact of outliers.

- STORIES : The missing or zero values are imputed with the median number of stories for the buildings in the same tax class. The idea behind this is similar to the imputation for FULLVAL, AVLAND, and AVTOT. The number of stories in a building can have a significant impact on its value, and thus it's important to have this variable available for all properties.

- LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH : Similar to the imputations discussed above, these fields are imputed based on the median values in the same tax class. These physical dimensions of the lot / building could have a significant effect on property value, and thus it's important to impute missing values to have a complete dataset.

4. Variable Creation

A) Variable Creation Logic

- ZIP3, ZIP5 : These new variables represent the 3-digit and the 5-digit ZIP code, respectively. They provide a way to analyze property data at different geographic levels. The 3-digit ZIP code corresponds to a larger area than the 5-digit code.
- LOTAREA, BLDAREA, BLDVOL : These new variables are created to give an overall description of the dimensions of the lot and building. They are calculated by multiplying frontage and depth measurements. These variables could potentially capture the interactions between the dimensionality variables.
- FULLVAL_per_sqft, AVLAND_per_sqft, AVTOT_per_sqft : These new variables represent the value per square foot and would give a more normalized measure of the property value. This is extremely helpful for comparisons between properties of different sizes.
- AllZip, AllTaxClass, AllBoro : These new variables are aggregated values that provide the count of all properties within the same ZIP code, tax class, and borough, respectively. They are intermediate variables used in normalization and creation of other new variables (r1 to r9).
- r1 to r9 : These are variables created to hold ratios of various property value measurements (FULLVAL, AVLAND, AVTOT) to the average measurements of all properties within the same group (ZIP5, ZIP3, tax class, and borough). These ratios provide a comparison of a given property's value to the average values in its group and could be useful for identifying properties that are significantly over or under-valued compared to their peers.
- r1_all to r9_all : These variables are ratios of the various property value measurements to the average measurements of all properties in the entire dataset. They provide a comparison of a given property's value to the average values across all properties and could be useful for identifying properties that are significantly over or under-valued on a city-wide basis.
- R1 to R9 : These variables are essentially the same as r1 to r9, except that instead of dividing by the mean, we are dividing by the total count of properties in the group. This gives us the average property value measurements per property, allowing for a more straightforward comparison between individual properties and the average.

- R1_all through R9_all : These variables are the same as R1 through R9, except they are relative to the entire dataset instead of their respective groups.
- Value_ratio : Ratio of FULLVAL to (AVLAND + AVTOT).
- Size_ratio : Ratio of building size to lot size.

B) List of new Variables

Field Name / Category	Description	# Variables Created
zip3, zip5	Cleaned, 3-digit zip code and 4-digit zip code.	2
all_zip	Total count of all properties within the same ZIP code.	1
ltsize	Lot size, area of the lot.	1
bldsize	Building size, area of the lot.	1
bldvol	Building volume	1
r1 to r9	Ratio of all 3 property values wrt the 3 size variables (above)	9
r1inv to r9inv	Inverse of r1 through r9 ratios	9
FULLVAL_per_sqft	Market Value per square foot.	1
AVLAND_per_sqft	Actual Land Value per square foot.	1
AVTOT_per_sqft	Actual Total Value per square foot.	1
value_ratio	Ratio of FULLVAL to (AVLAND + AVTOT).	1
size_ratio	Ratio of building size to lot size.	1
r1_zip5 to r9_zip5	Ratios r1 through r9 grouped by zip code category.	9
r1inv_zip5 to r9inv_zip5	Ratios r1inv through r9inv grouped by zip code categories.	9
r1_taxclass to r9_taxclass	Ratios r1 through r9 grouped by taxclass code.	9
r1inv_taxclass to r9inv_taxclass	Ratios r1inv through r9inv grouped by taxclass code.	9
R1 to R9	Ratios r1 through r9 grouped by total count of properties in the category.	9
R1inv to R9inv	Inverse of ratios R1 through R9.	9

Total of 83 new variables created with some extra intermediate variables.

5. Dimensionality Reduction

The dimensionality reduction step in the analysis employed Principal Component Analysis (PCA) and z-scaling methods to reduce the number of variables while retaining the essential information in the dataset. Dimensionality reduction is crucial in data analysis to address the issue of high-dimensional data and to improve computational efficiency.

a) Z-scaling

The z-scaling method, standardizes the variables by subtracting the mean and dividing by the standard deviation. This transformation ensures that all variables are on the same scale, preventing the dominance of certain variables due to their larger magnitudes. The formula used for z-scaling is:

$z = (x - \mu) / \sigma$

Where:

- z is the z-score
- x is the original value of the variable
- μ is the mean of the variable
- σ is the standard deviation of the variable

b) Principal Component Analysis (PCA)

PCA is used to extract the most important components or dimensions from the dataset. It achieves this by transforming the original variables into a new set of uncorrelated variables called principal components. PCA identifies the principal components, which are linear combinations of the original variables that capture the most significant variance in the dataset. These principal components are orthogonal to each other, allowing for the reduction of dimensions while retaining as much information as possible.

The first step in PCA involves computing the covariance matrix of the z-scaled dataset. Then, PCA calculates the eigenvectors and eigenvalues of this covariance matrix. The eigenvectors represent the principal components, and the corresponding eigenvalues indicate the amount of variance explained by each component.

To determine the number of components to retain, a scree plot and cumulative variance plot were generated. The scree plot displays the

explained variance ratio for each component, while the cumulative variance plot illustrates the cumulative variance explained by the components. These plots aid in deciding the appropriate number of components to retain, balancing the reduction in dimensions with the retention of information.

Optionally, after retaining the PCs, a second z-scaling step was applied to make all the retained PCs equally important. This optional z-scaling procedure ensures that each retained PC contributes equally to the subsequent analysis, particularly when a small number of PCs are selected for further investigation. Once the number of components is determined, PCA is rerun with the specified number of components. The dataset is transformed into a new set of variables called principal components.

6. Anomaly Detection Algorithms

This section focuses on the implementation of two anomaly detection algorithms: the z-scale score and the auto-encoder score. These algorithms were utilized to calculate anomaly scores for each record in the dataset, allowing the identification of unusual properties that may indicate potential tax fraud.

1. Score 1: Z-scale

The first algorithm, the z-scale score, measures the deviation of each record from the mean in the standardized dataset. It calculates the sum of the absolute values of the z-scaled variables raised to a specified power (p) where p could be any integer. Changing the value of p here does not change the result and is hence not significant here. The formula for the z-scale score is:

$$s_i = (\Sigma_k \mid PCz^i_k \mid^p)^{1/p}$$

As we can see, this z-score is calculated using the Minkowski distance formula. This algorithm captures the overall deviation of each record from the mean, enabling the detection of outliers or unusual patterns in the dataset.

2. **Score 2: Auto-encoder**

   The second algorithm, the auto-encoder score, employs a neural network-based approach. An autoencoder is trained on the z-scaled dataset to reconstruct the original input. The difference between the reconstructed output and the original input (reconstruction error) is used as the anomaly score. The autoencoder learns to encode the information and then decode it back to reconstruct the input. The formula for the auto-encoder score is:

   $$s_i = (\Sigma_k \mid PCz`^i_k - PCz^i_k \mid ^p)^{1/p}$$

   The auto-encoder score focuses on capturing the information loss during the reconstruction process, identifying records that are difficult to reconstruct accurately.

3. **Final Score**

   To combine these two scores and obtain a final anomaly score, the scores from both algorithms are scaled and averaged. The scaling ensures that both scores are on the same scale and equally contribute to the final score. The formula for the final score is:

   final_score = 0.5 * (score1 rank + score2 rank)

   Here, the ranks of the scores are used instead of the scores themselves to ensure consistent scaling across different datasets. The records are then sorted based on the final score in descending order, enabling the identification of the top records with the highest anomaly scores.

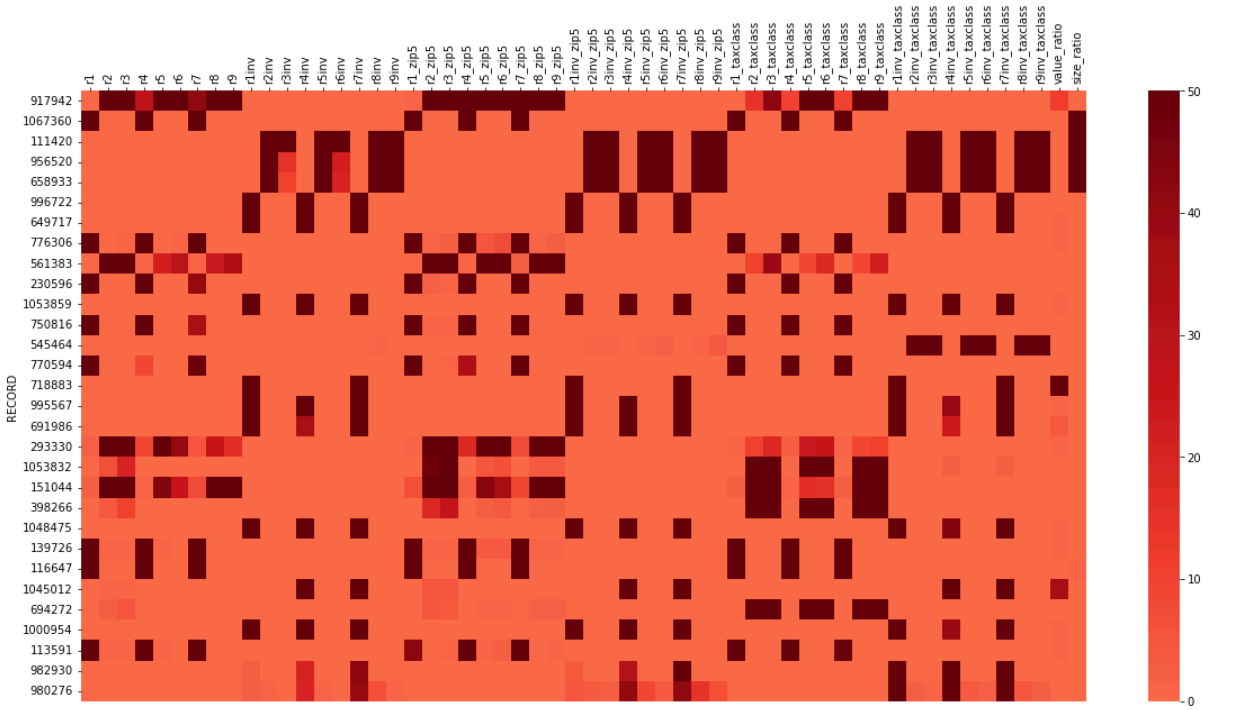
   By using the z-scale score and auto-encoder score, this analysis provides a comprehensive approach to detecting unusual properties within the dataset. The combination of these algorithms enhances the accuracy of flagging potential tax fraud cases, enabling the client to prioritize further investigation on the top-ranked records with the highest anomaly scores.

7. Results

Since we don't have any fraud labels or a dependent variable we cannot use standard measures for building classification models. Hence we build our own unsupervised model that calculates an anomaly score, sorts by this score and gives us the top most records that are most likely to be anomalous.

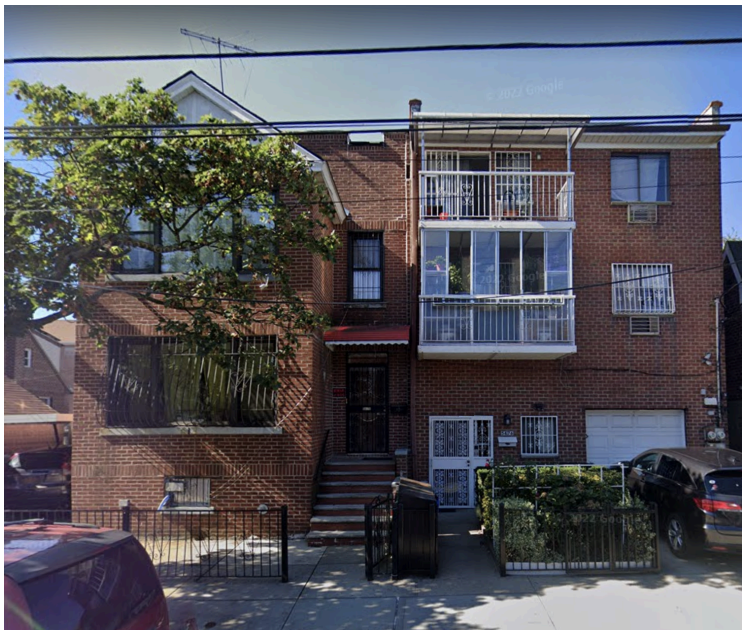
The results were examined using several steps to gain insights into the flagged properties.

- Firstly, the z-scaled variables of the top records were analyzed. These z-scores indicate the number of standard deviations the record deviates from the mean, providing a measure of abnormality.
- To gain further understanding, a heat map of the z-scaled variables was plotted. This visualization helped identify the specific variables contributing to the high anomaly scores.
- To narrow down the focus of the investigation, the 18 core variables were considered, and a heat map was generated for this subset.
- Subsequently, a manual examination of the most unusual properties was conducted, one by one, with focus on the unusual variables seen in the heatmap. Sample heat map shown below.
- Looking at the google street view image of the property can also help paint a clearer picture of the property and support the investigation.



Five Unusual Properties

1) Record 658933 : 54-76 83 STREET



RECORD	658933
BBLE	4029060054
BORO	4
BLOCK	2906
LOT	54
EASEMENT	NaN
OWNER	WAN CHIU CHEUNG
BLDGCL	C0
TAXCLASS	1
LTFRONT	25
LTDEPTH	100
EXT	NaN
STORIES	3.0
FULLVAL	776000.0
AVLAND	26940.0
AVTOT	46560.0
EXLAND	1620.0
EXTOT	1620.0
EXCD1	1017.0
STADDR	54-76 83 STREET
ZIP	11373.0
EXMPTCL	NaN
BLDFRONT	2500
BLDDEPTH	5600
AVLAND2	NaN
AVTOT2	NaN
EXLAND2	NaN
EXTOT2	NaN
EXCD2	NaN
PERIOD	FINAL
YEAR	2010/11
VALTYPE	AC-TR

- The property has very small lot sizes when compared to the building sizes. As we can see from the record details above the BLDFRONT and BLDDEPTH values are 2500 and 5600 respectively, whereas the LTFRONT and LTDEPTH values are 25 and 100. Since the lot size cannot be smaller than the building dimensions this record is anomalous.
- The r8_inv value for this property is unusually large. This suggests that the ratio of the actual total value of the property with respect to the building size is smaller than usual.
- The inverse ratios grouped by zip codes are also unusually large, especially r8inv_zip5 and r9inv_zip5 values. This again corroborates the previous finding that the building size and volume is larger than expected when compared to the actual total property value.

2) Record 111420 : 1438 3 AVENUE

RECORD	111420
BBLE	1015101092
BORO	1
BLOCK	1510
LOT	1092
EASEMENT	NaN
OWNER	BOXWOOD FLTD PARNTERS
BLDGCL	R4
TAXCLASS	2
LTFRONT	75
LTDEPTH	93
EXT	NaN
STORIES	31.0
FULLVAL	296508.0
AVLAND	22896.0
AVTOT	133429.0
EXLAND	0.0
EXTOT	0.0
EXCD1	NaN
STADDR	1438 3 AVENUE
ZIP	10028.0
EXMPTCL	NaN
BLDFRONT	7575
BLDDEPTH	9393
AVLAND2	22896.0
AVTOT2	146183.0
EXLAND2	NaN
EXTOT2	NaN
EXCD2	NaN
PERIOD	FINAL
YEAR	2010/11
VALTYPE	AC-TR

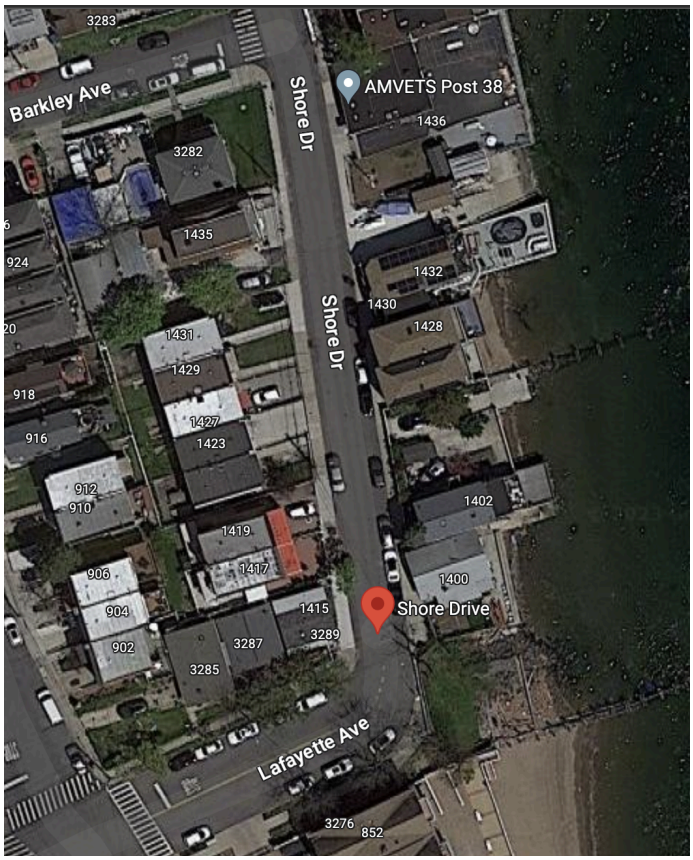
- This property has exceptionally small lot sizes when compared to the building sizes. The BLDFRONT and BLDDEPTH values are 7575 and 9393 respectively, whereas the LTFRONT and LTDEPTH values are 75 and 93. Since the lot size cannot be smaller than the building dimensions this record is anomalous. Based on the number it looks like the values are arbitrary or erroneous and it is likely to not be a fraudulent record.
- The variable size_ratio is extremely large for this record which is due to the same reason mentioned above since this value is the ratio of building size wrt lot size.
- The r2inv and r3inv values for this record is also extremely high suggesting that the FULLVAL or the market value for this property is too low for a building of this size / volume.
- The inverse ratios of r2, r3, r5, r8 grouped by tax class are also pretty large for this property suggesting that the property values (market value, actual land value, and actual total value) are too low for a property of this size in this particular tax class. The tax class for this property is 2 which indicates that this is a residential property, more specifically apartments.

3) Record 151044 : <u>1 EAST 161 STREET</u>

RECORD	151044
BBLE	2024930001
BORO	2
BLOCK	2493
LOT	1
EASEMENT	NaN
OWNER	NaN
BLDGCL	Q6
TAXCLASS	4
LTFRONT	798
LTDEPTH	611
EXT	NaN
STORIES	6.0
FULLVAL	1663775000.0
AVLAND	78750000.0
AVTOT	748698750.0
EXLAND	78750000.0
EXTOT	748698750.0
EXCD1	2500.0
STADDR	1 EAST 161 STREET
ZIP	10451.0
EXMPTCL	NaN
BLDFRONT	0
BLDDEPTH	0
AVLAND2	NaN
AVTOT2	NaN
EXLAND2	NaN
EXTOT2	NaN
EXCD2	NaN
PERIOD	FINAL
YEAR	2010/11
VALTYPE	AC-TR

- Based on the street view and the building class of the property we can conclude that this is the “Yankee Stadium” which is an outdoor recreation facility. [BLDGCL = Q6 which implies the building is one of the following: STADIUM, RACE TRACK, BASEBALL FIELD]
- Given that this is a huge outdoor stadium, the lot sizes LTFRONT and LTDEPTH values of 798 and 611 respectively are too low for this property.
- The building sizes BLDFRONT and BLDDEPTH are both zero which is also incorrect given that the number of stories for the building is 6.
- The r2 and r3 ratio values are too large for this property indicating that the market value of the property is way too high given the building size. This follows from the aforementioned point since the building dimensions here are zero which seems to be an error.
- Overall while this record is unusual, the reason behind that seems to just be that this type of property is rare in occurrence and not because it is a suspicious record that needs further investigation.

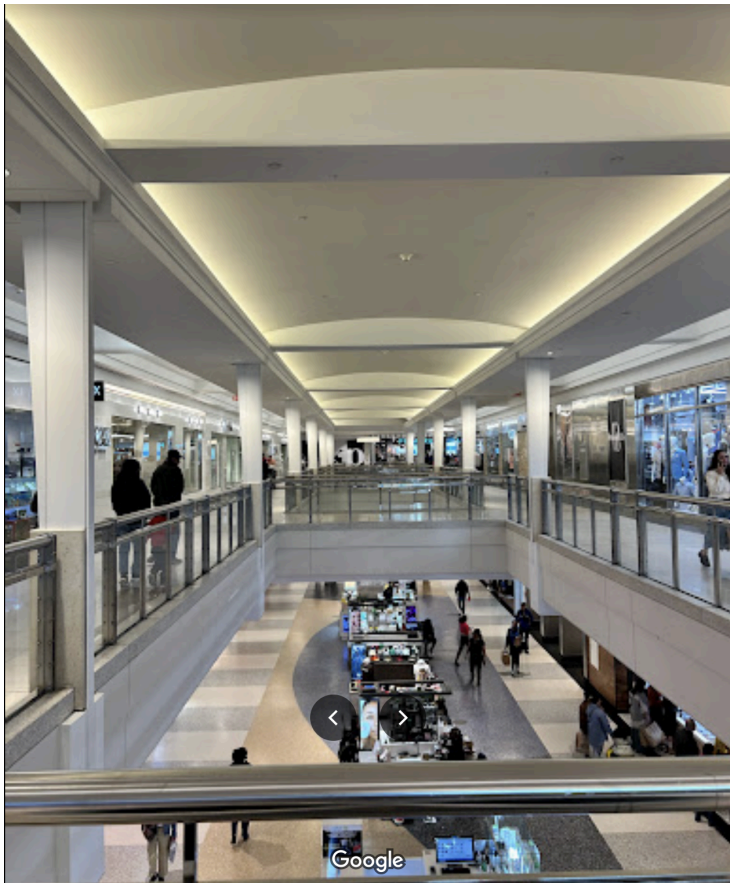
4) Record 241946 : SHORE DRIVE



RECORD	241946
BBLE	2054670100
BORO	2
BLOCK	5467
LOT	100
EASEMENT	NaN
OWNER	RUFFALO ENTERPRISES
BLDGCL	V0
TAXCLASS	1B
LTFRONT	503
LTDEPTH	999
EXT	NaN
STORIES	NaN
FULLVAL	31236727.0
AVLAND	134.0
AVTOT	134.0
EXLAND	134.0
EXTOT	134.0
EXCD1	3390.0
STADDR	SHORE DRIVE
ZIP	NaN
EXMPTCL	X2
BLDFRONT	0
BLDDEPTH	0
AVLAND2	NaN
AVTOT2	NaN
EXLAND2	NaN
EXTOT2	NaN
EXCD2	NaN
PERIOD	FINAL
YEAR	2010/11
VALTYPE	AC-TR

- The building classification code of this record is BLDGCL = V0 which denotes a vacant land that is “ZONED RESIDENTIAL; NOT MANHATTAN”. Based on this information and the street view it looks like this property is an empty lot on this street. This would explain why the BLDFRONT and BLDDEPTH values are zero.
- There is no detailed street address for this property and the lot number (LOT = 100) seems a little strange since the street view does not show any lot 100 and the lot values are all 1400 and above for Shore Drive.
- The actual land value and the actual total value for this property is way too low when compared to the FULLVALL or the full market value. The value_ratio variable is way too large for this property because of this reason.
- The AVLAND, AVTOT, EXLAND, EXTOT values are all exactly the same which also suggests that these values are either arbitrary or erroneous. While this may not be concerning by itself but combined with the previous points, this indicates that this record needs further investigation.

5) Record 561383 : 5120 AVENUE U



RECORD	561383
BBLE	3084700055
BORO	3
BLOCK	8470
LOT	55
EASEMENT	NaN
OWNER	YILDIZ HOLDING A.S.
BLDGCL	K6
TAXCLASS	4
LTFRONT	930
LTDEPTH	650
EXT	NaN
STORIES	2.0
FULLVAL	258000000.0
AVLAND	40590000.0
AVTOT	116100000.0
EXLAND	0.0
EXTOT	13326371.0
EXCD1	1985.0
STADDR	5120 AVENUE U
ZIP	11234.0
EXMPTCL	NaN
BLDFRONT	0
BLDDEPTH	0
AVLAND2	40590000.0
AVTOT2	118079991.0
EXLAND2	NaN
EXTOT2	13326371.0
EXCD2	NaN
PERIOD	FINAL
YEAR	2010/11
VALTYPE	AC-TR

- The building class code for this property is K6 which falls under the store buildings category, more specifically referring to “SHOPPING CENTER WITH OR WITHOUT PARKING”. The street view corroborates this information and hence we can conclude that this record refers to a specific shop in the Kings Plaza Shopping Center shown above.
- The values for both the building dimensions are zero here, which clashes with the value for the number of stories for this record which is 2.
- The following ration values for this property are extremely high : r2_zip, r3_zip, r5_zip, r6_zip, r8_zip, and r9_zip. This follows from the previous observation since all of these six ration are with respect to the building dimensions and grouped by the zip code. This zip code primarily contains high-end shopping outlets and therefore the market values for these places is typically high but the building sizes are also generally bigger than usual.
- Overall this record was in the top scorer list because of the building dimensions being zero while the market value was high. However this seems to be more of an error than a suspicious record and can probably be ignored.

8. Summary

In conclusion, this business report has outlined the findings and results of the analysis performed on the NY Property dataset, aimed at identifying unusual property records that may indicate potential tax fraud cases. The analysis involved a series of steps, including data cleaning, exclusions, imputations, variable creation, dimensionality reduction, and the calculation of anomaly scores. The output of the analysis is a ranked list of properties based on their anomaly scores, highlighting the most unusual properties for further investigation.

Upon manual investigation of a subset of the top unusual properties, it was determined that while some of these properties were flagged due to the rarity of certain characteristics, such as exceptionally high market value for a small upscale shopping center, there were indeed properties that warranted further scrutiny and could potentially be tax fraud records.

Based on client feedback, there are several ways to improve the analysis in the future. Firstly, refining the anomaly scoring method by considering additional factors and incorporating domain knowledge could enhance the accuracy of flagging unusual properties. Additionally, conducting more thorough investigations on the flagged properties by gathering additional data or utilizing external resources could provide a more comprehensive understanding of potential tax fraud cases.

Overall, this analysis serves as a valuable tool for the client's efforts in detecting tax underpayments resulting from misrepresented property characteristics. By leveraging the insights gained from this analysis and incorporating client feedback, future iterations can further enhance the accuracy and effectiveness of identifying and addressing tax fraud in the real estate sector.

9. Appendix

Data Quality Report

---

1. Data Description

The dataset is **NY Property Data**, which contains **property valuation and assessment data** for the city of New York provided by the Department of Finance. The data consists of real estate assessment property data coming from the New York city government. The data is updated annually and spans over the time period of a couple of years, **2010/11**. There are **32 fields** and **1070994 records**.

2. Summary Tables

a. Numeric Fields Table

	Field Name	Field Type	# Records Have Values	% Populated	% Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	LTFRONT	numeric	1070994	100.00%	15.79%	0	9999	36.64	74.03	0
1	LTDEPTH	numeric	1070994	100.00%	15.89%	0	9999	88.86	76.40	100
2	STORIES	numeric	1014730	94.75%	0.00%	1	119	5.01	8.37	2
3	FULLVAL	numeric	1070994	100.00%	1.21%	0	6150000000	874,264.51	11,582,425.58	0
4	AVLAND	numeric	1070994	100.00%	1.21%	0	2668500000	85,067.92	4,057,258.16	0
5	AVTOT	numeric	1070994	100.00%	1.21%	0	4668308947	227,238.17	6,877,526.09	0
6	EXLAND	numeric	1070994	100.00%	45.91%	0	2668500000	36,423.89	3,981,573.93	0
7	EXTOT	numeric	1070994	100.00%	40.39%	0	4668308947	91,186.98	6,508,399.78	0
8	BLDFRONT	numeric	1070994	100.00%	21.36%	0	7575	23.04	35.58	0
9	BLDDEPTH	numeric	1070994	100.00%	21.37%	0	9393	39.92	42.71	0
10	AVLAND2	numeric	282726	26.40%	0.00%	3	2371005000	246,235.72	6,178,951.64	2408
11	AVTOT2	numeric	282732	26.40%	0.00%	3	4501180002	713,911.44	11,652,508.34	750
12	EXLAND2	numeric	87449	8.17%	0.00%	1	2371005000	351,235.68	10,802,150.91	2090
13	EXTOT2	numeric	130828	12.22%	0.00%	7	4501180002	656,768.28	16,072,448.75	2090

b. Categorical Fields Table

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	RECORD	categorical	1070994	100.00%	0	1070994	1
1	BBLE	categorical	1070994	100.00%	0	1070994	1000010101
2	BORO	categorical	1070994	100.00%	0	5	4
3	BLOCK	categorical	1070994	100.00%	0	13984	3944
4	LOT	categorical	1070994	100.00%	0	6366	1
5	EASEMENT	categorical	4636	0.43%	0	12	E
6	OWNER	categorical	1039249	97.04%	0	863347	PARKCHESTER PRESERVAT
7	BLDGCL	categorical	1070994	100.00%	0	200	R4
8	TAXCLASS	categorical	1070994	100.00%	0	11	1
9	EXT	categorical	354305	33.08%	0	3	G
10	EXCD1	categorical	638488	59.62%	0	129	1017
11	STADDR	categorical	1070318	99.94%	0	839280	501 SURF AVENUE
12	ZIP	categorical	1041104	97.21%	0	196	10314
13	EXMPTCL	categorical	15579	1.45%	0	14	X1
14	EXCD2	categorical	92948	8.68%	0	60	1017
15	PERIOD	categorical	1070994	100.00%	0	1	FINAL
16	YEAR	categorical	1070994	100.00%	0	1	2010/11
17	VALTYPE	categorical	1070994	100.00%	0	1	AC-TR

3. Visualization of Each Field

a. Field Name : **RECORD**

Description : Ordinal unique positive integer for each property record, from 1 to 1070994.

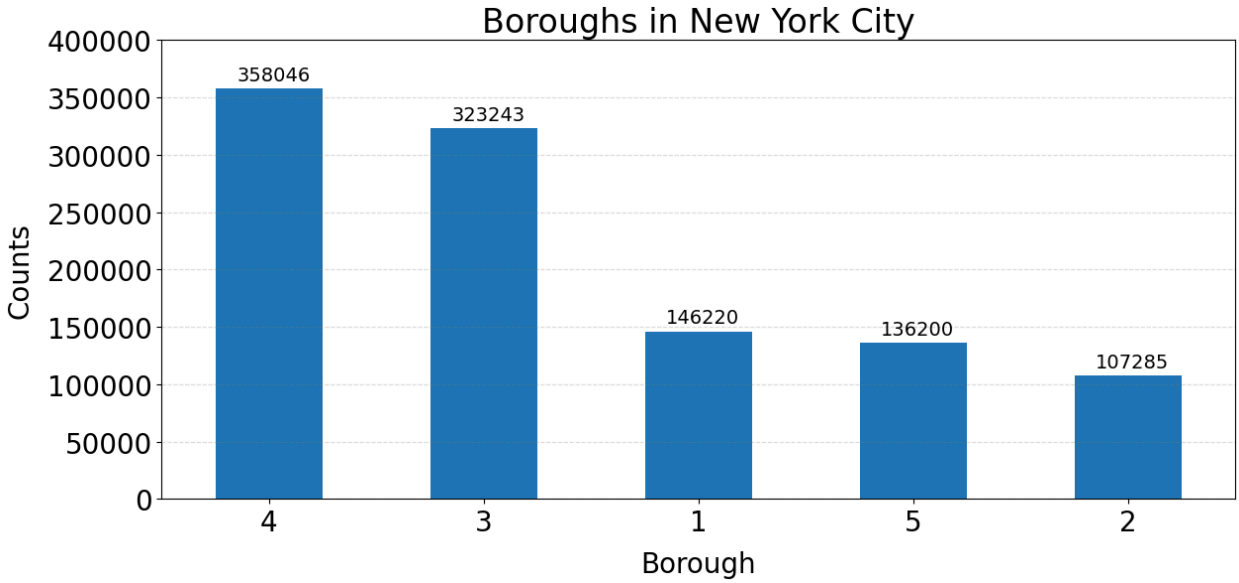
b. Field Name : **BBLE**

Description : This is the file key. BBLE stands for - Boro, Block, Lot and Easement code. This field has 1070994 unique values.

c. Field Name : **BORO**

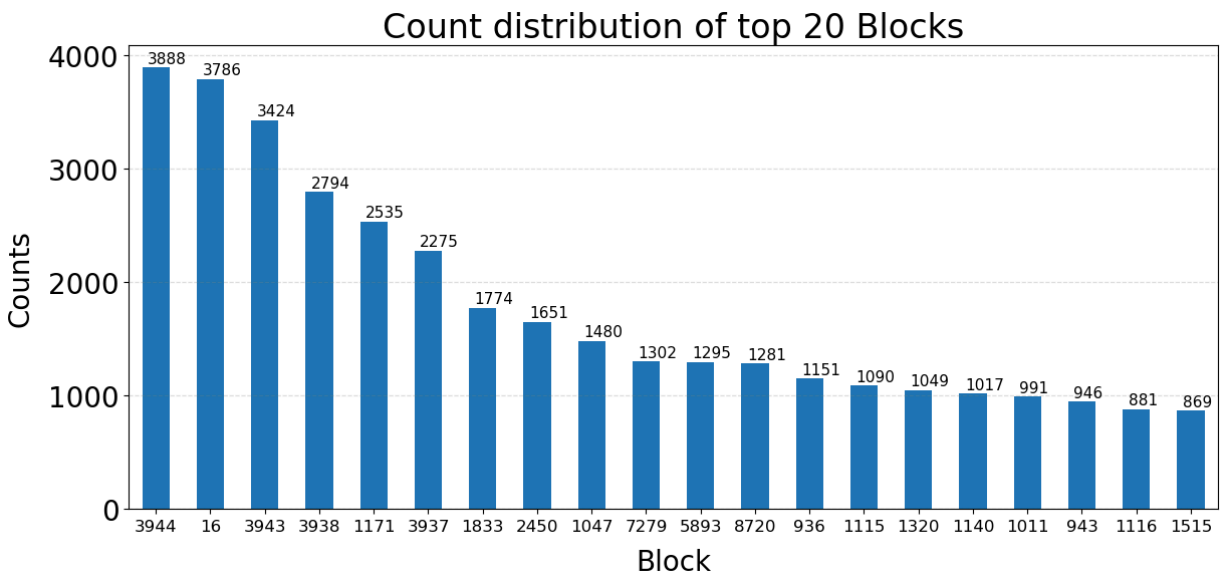
Description : This categorical field represents the borough. It has 5 unique values which are listed below. The plot below shows the count distribution of each category. From the plot we can see that Queens (4) has the highest number of properties listed in the dataset.

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



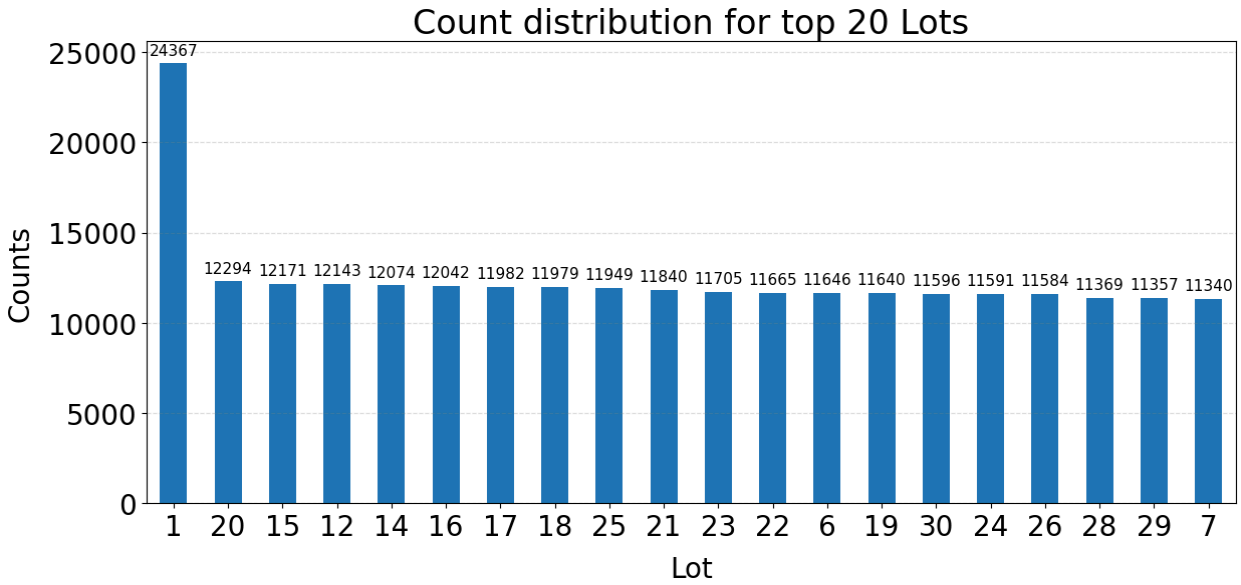
d. Field Name : **BLOCK**

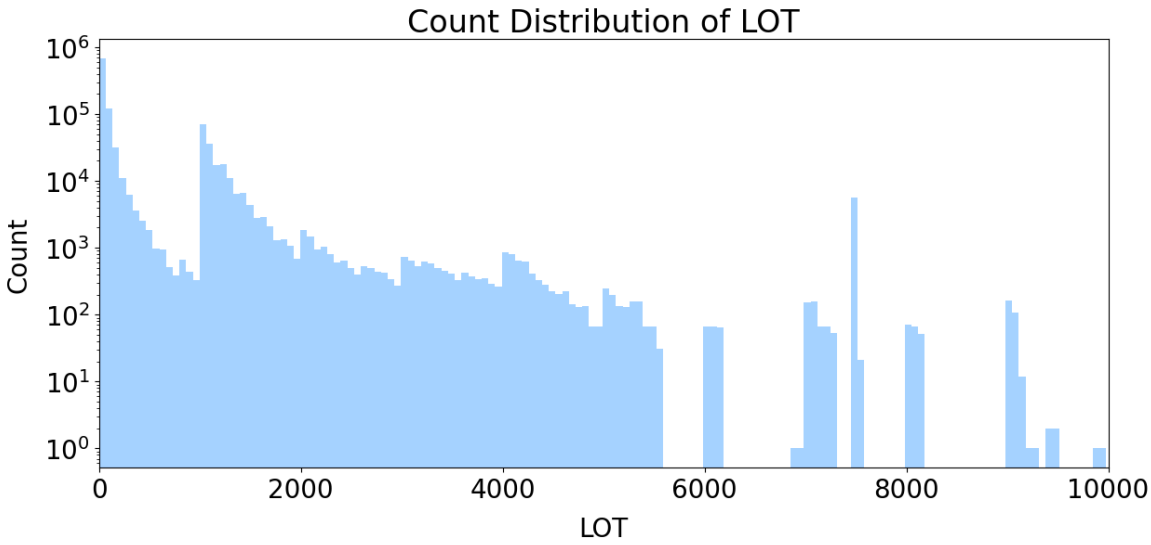
Description : This is the valid block range by borough. This categorical field has 13984 distinct values. We see that block 3944 has the highest number of properties listed.



e. Field Name : **LOT**

Description : This is the property lot. This categorical field has 6366 unique values. Lot 1 is the most common one with 24367 properties. The first plot shows the count distribution of the top 20 lots on a regular scale while the second plot shows the count distribution of all lots on a log scale. It is easier to observe the distribution on a log scale for this field.

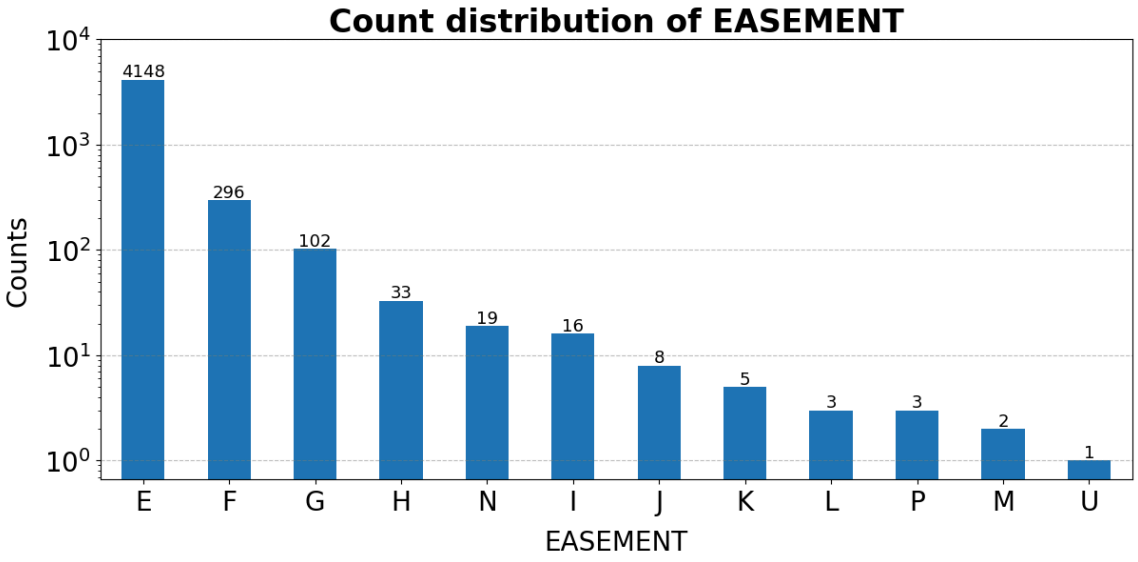




f. Field Name : **EASEMENT**

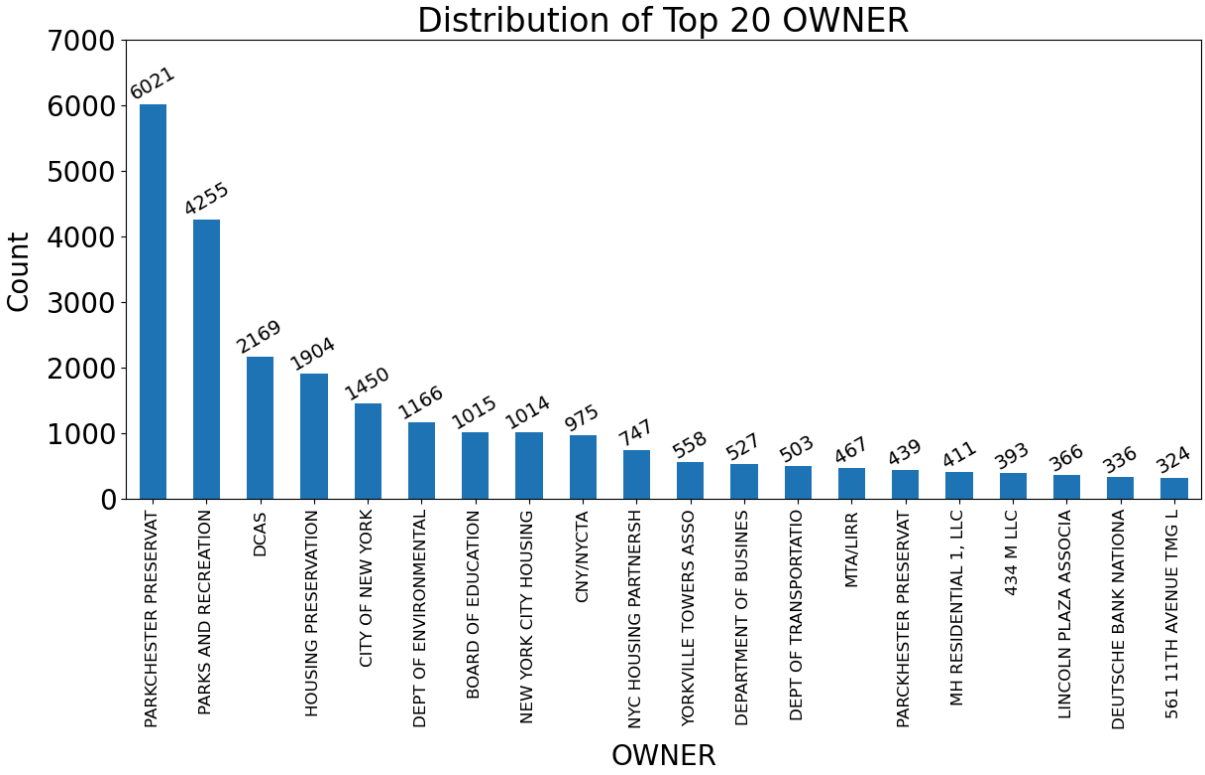
Description : This field specifies the easement rights the property holds. This is a categorical field that contains either a space (converted to null) or a single letter indicating the easement type. This field has 13 unique values, or 12 unique values if we do not count the first one (i.e. space) :

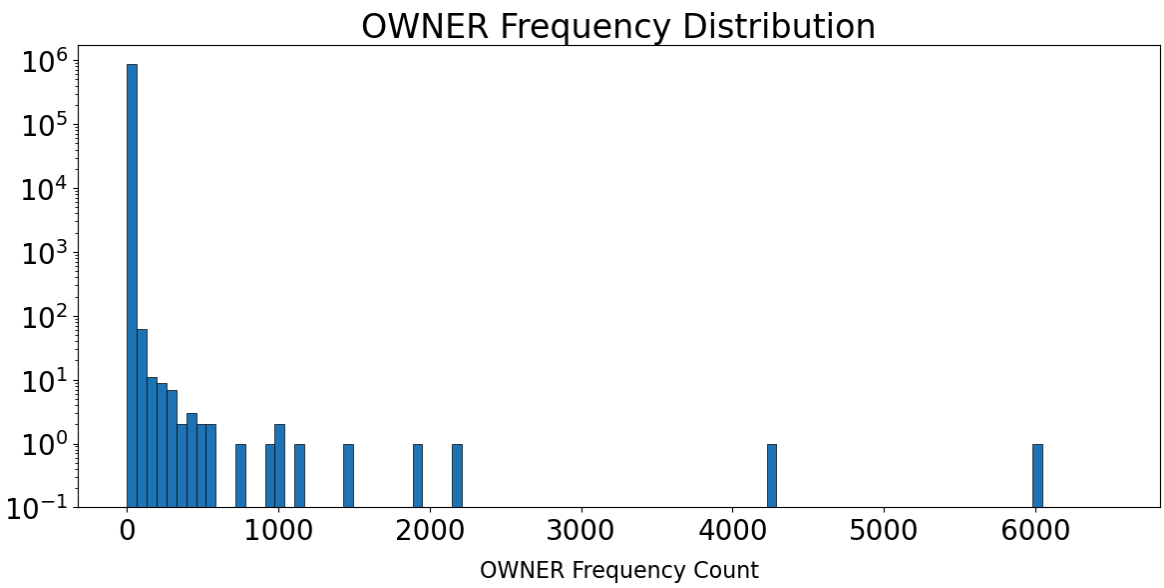
- E = Land Easement
- F, G, H, I, J, K, L, M = are duplicates of E
- N = Non-Transit Easement
- P = Pier
- U = U.S. Government



g. Field Name : **OWNER**

Description : Name of the property owner. This field has 863348 distinct values with about 97% of the records populated with values. The most common owner is “PARKCHESTER PRESERVAT” with 6021 properties. The first plot shows the top 20 owners along with their names. The second plot shows the frequency of occurrences of an owner's name. Most owners own a single property, and only a handful that own multiple.

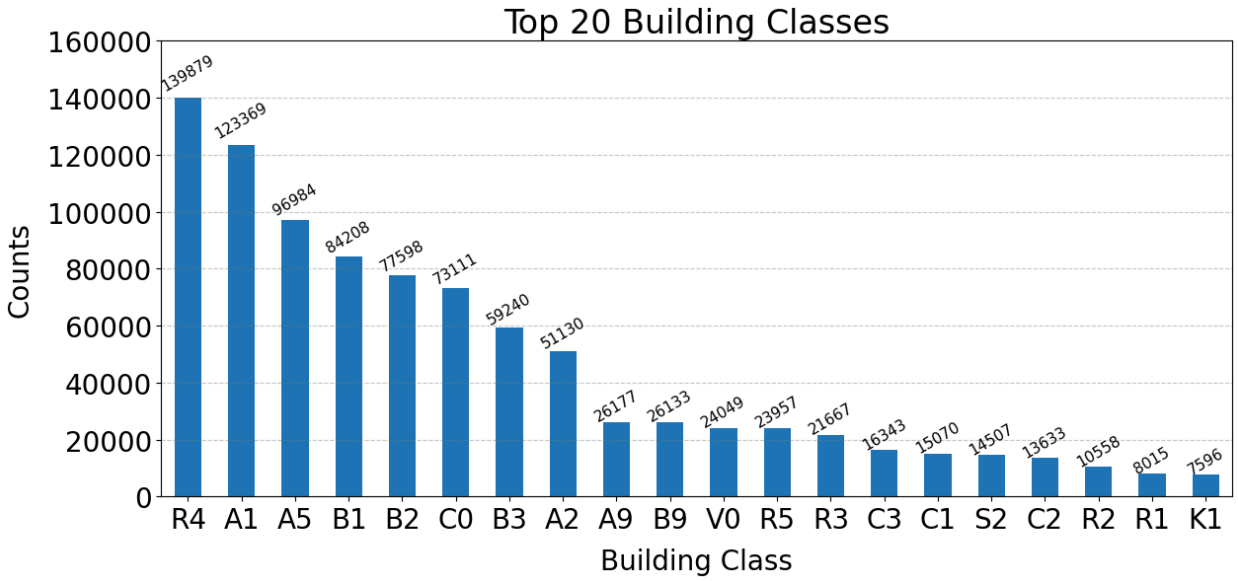


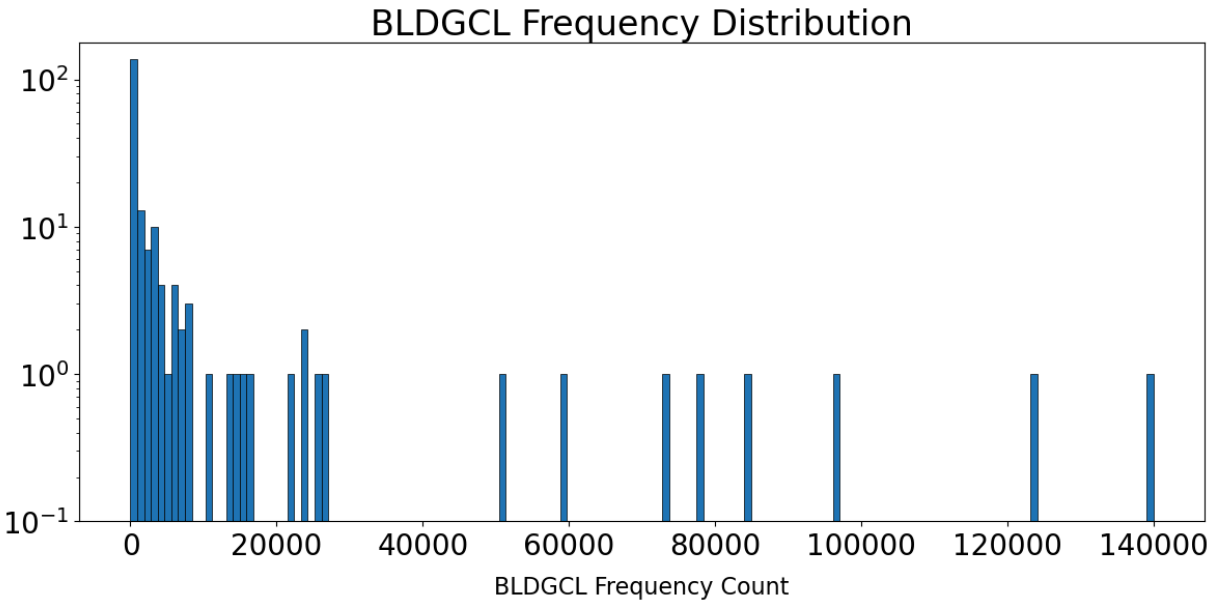


h. Field Name : **BLDGCL**

Description : Building Class. This categorical field indicates the general condition and quality of a building. This field has 200 distinct values with the most common being “R4”.

The first plot shows the top 20 most common building classes. The second plot shows the frequency distribution of building classes based on the number of times they occur in the dataset. Most building class types occur up to 10k types with only a couple of them occurring more than 100k times.

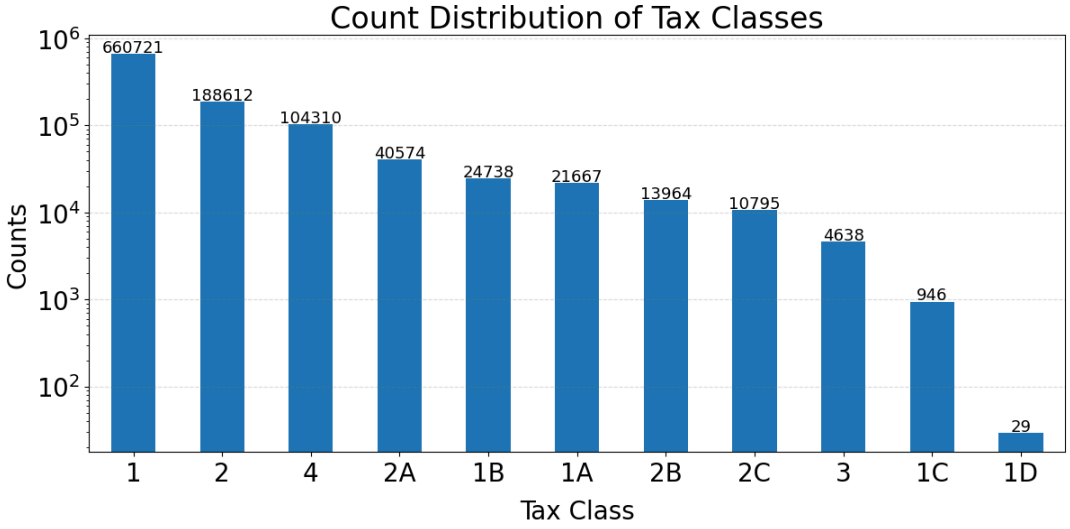




i. Field Name : **TAXCLASS**

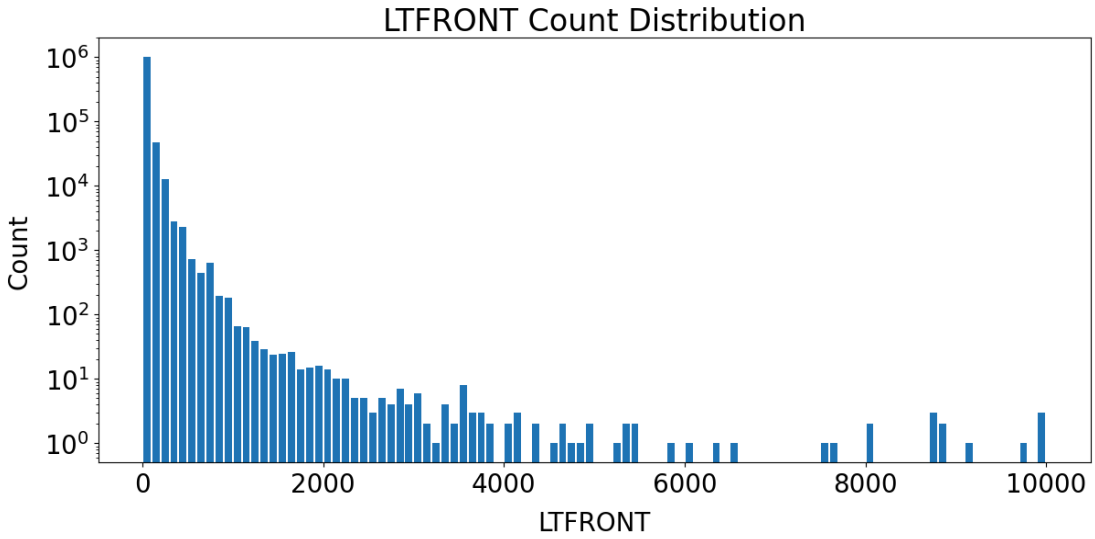
Description : Tax Class. This categorical field indicates the type of tax category the property falls in. This field has 11 distinct classes with the most common class being “1”. Following is the description of some of the classes :

- 1 = 1 - 3 Unit Residence
- 2 = Apartments; 2A = 4, 5, or 6 Unit apartments
- 3 = Utilities
- 4 = All Others

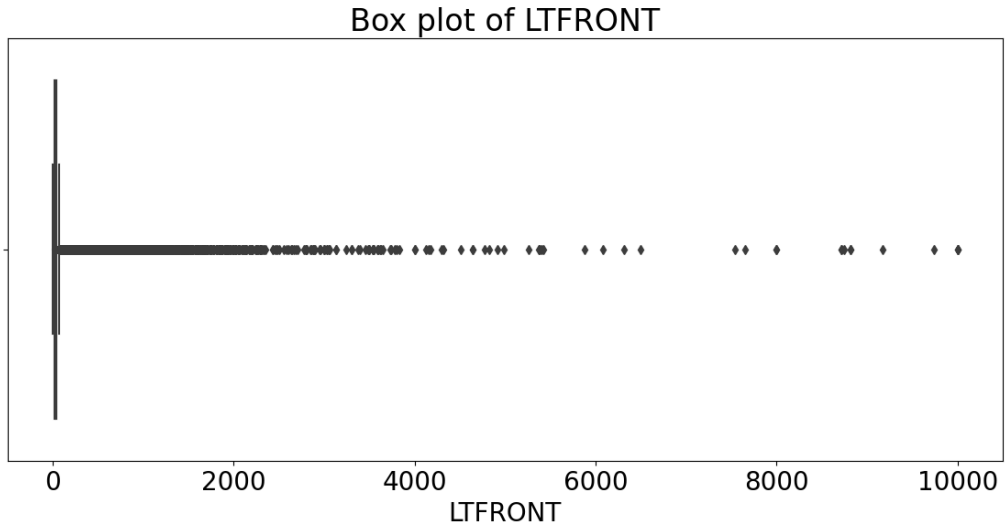


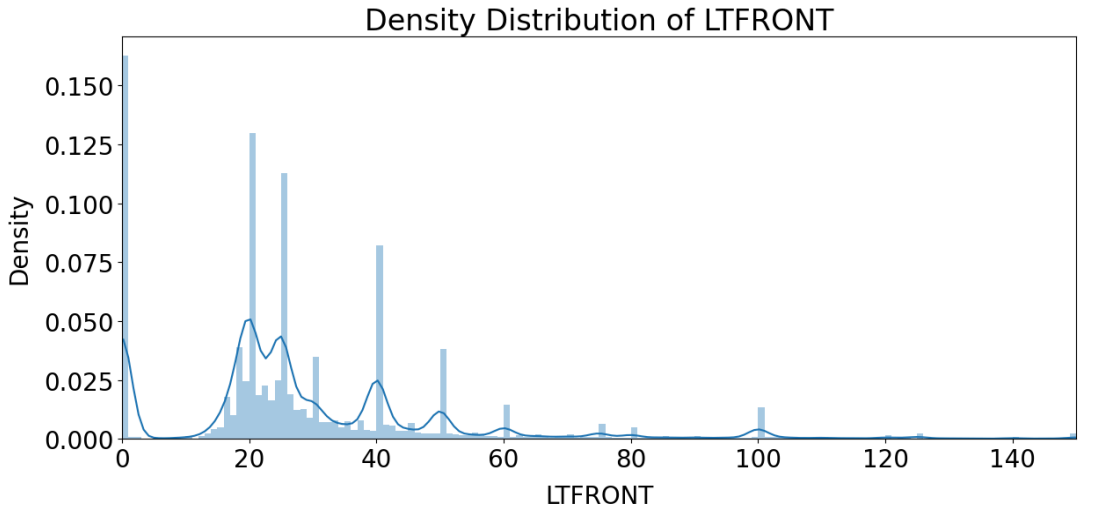
j. Field Name : **LTFRONT**

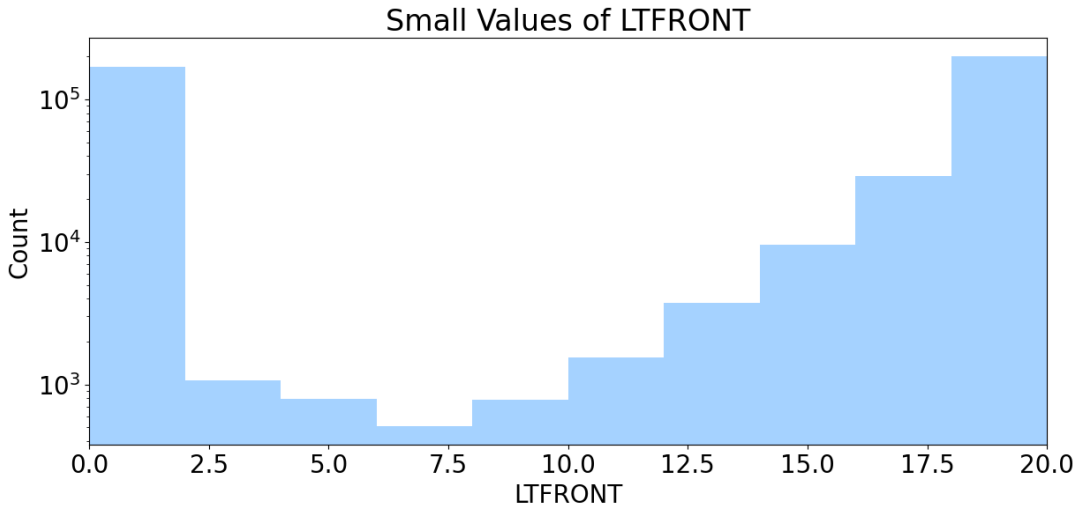
Description : Lot width. This numerical field has values ranging from 0 to 9999, with a mean value of 36.64 and close to 16% of the records are 0.



LTFRONT values above 6000 are low in occurrence and hence are possibly outliers. The values are more concentrated around the smaller range (0-50).

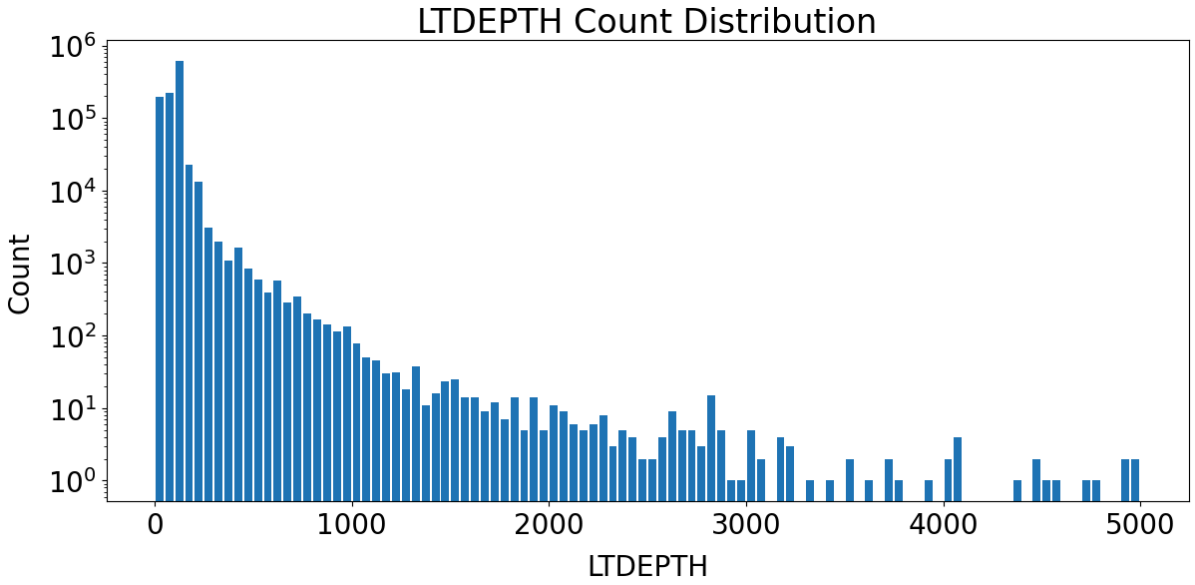


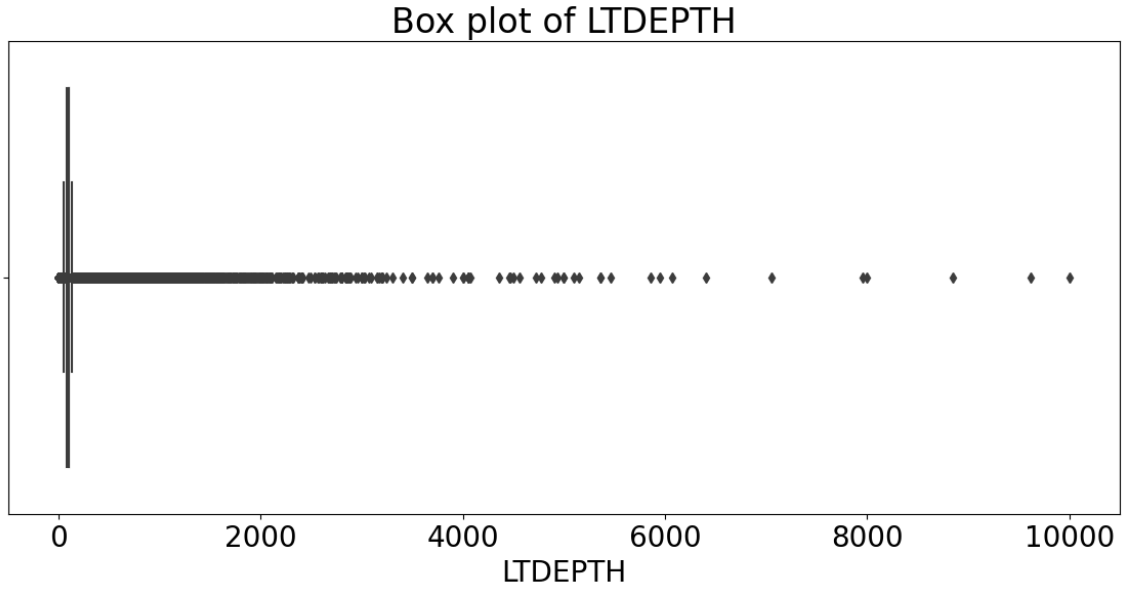


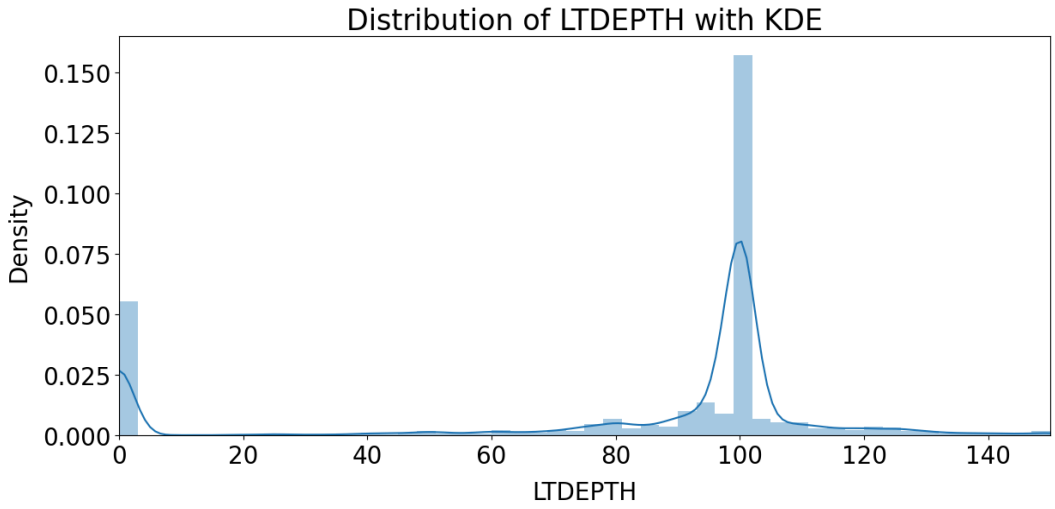


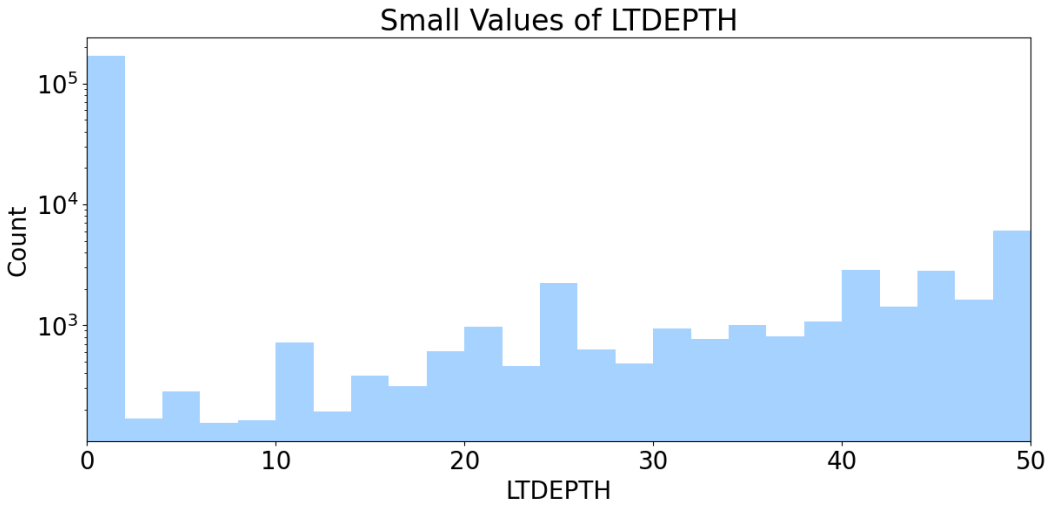
k. Field Name : **LTDEPTH**

Description : Lot Depth. This numerical field has values ranging from 0 to 9999, with a mean value of 88.86 and close to 16% of the records are 0. This field has 1370 unique values with the most common value being 100. The first plot shows the distribution of lot depth values up to 5k. The boxplot shows that values above 8k are infrequent and possibly outliers.


LTDEPTH Count Distribution
Count
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
1000
2000
3000
4000
5000
LTDEPTH

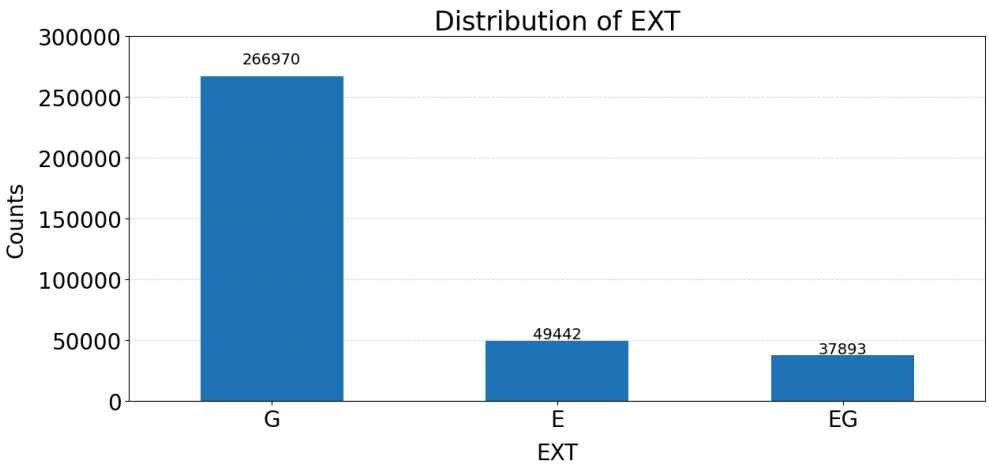

Box plot of LTDEPTH
0
2000
4000
6000
8000
10000
LTDEPTH





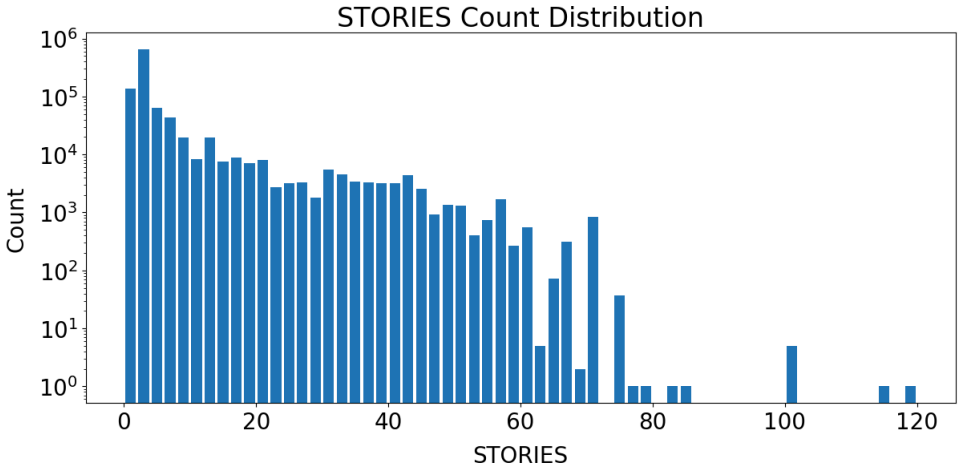
I. Field Name : **EXT**

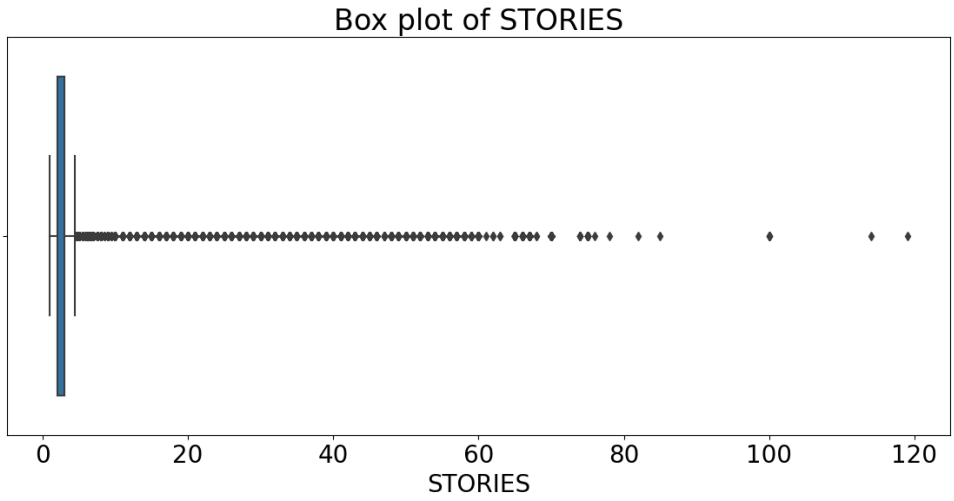
Description : Extension Indicator. This categorical field has 3 unique values with the value “G” being the most common.

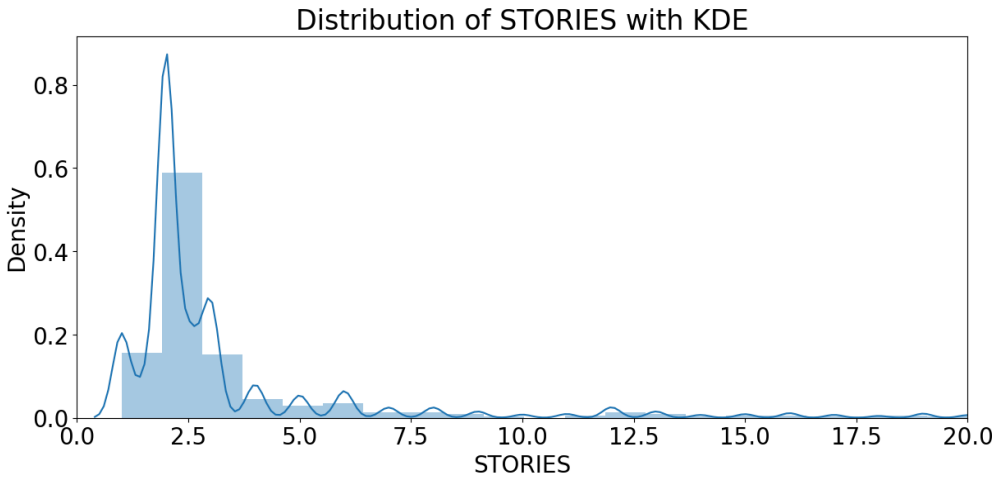


m. Field Name : **STORIES**

Description : This is the number of stories in the building. This numerical field has values ranging from 1 to 119 with a mean value of 5 and mode 2. The boxplot shows that values above 100 are rare / outliers.

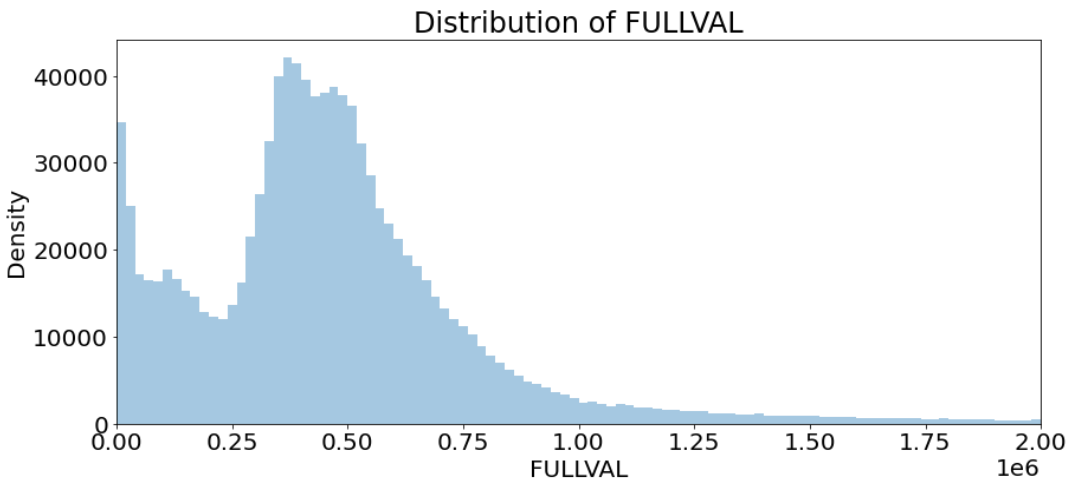


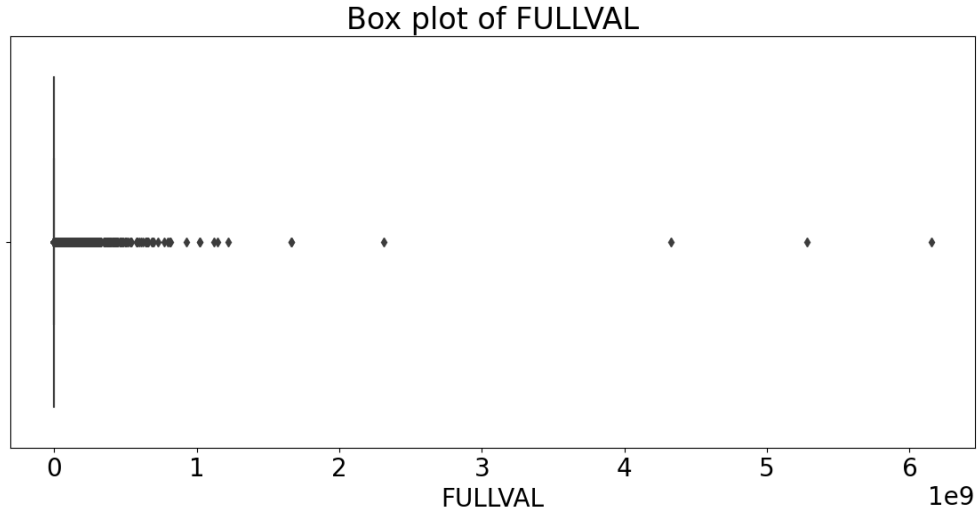


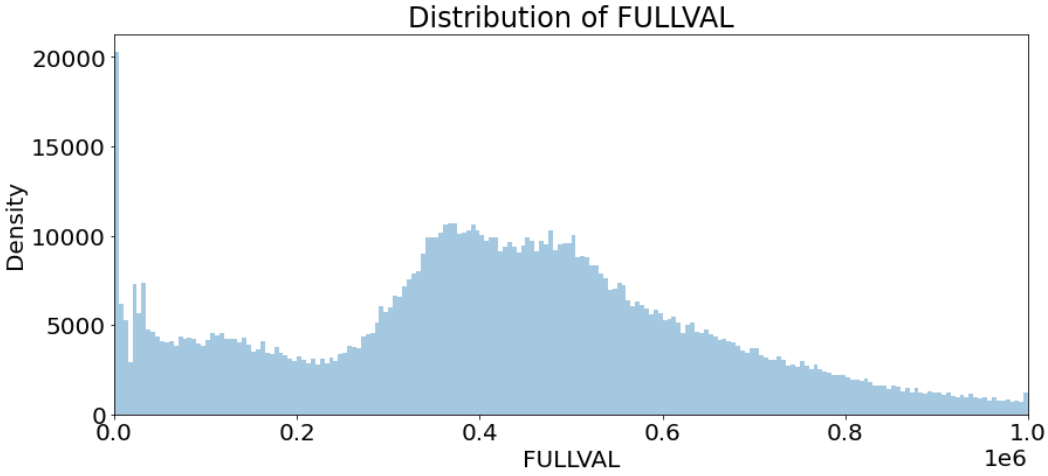


n. Field Name : **FULLVAL**

Description : Market Value. This is a numerical field with values ranging from 0 to 6150000000 with a mean value of 874264.51 and mode of 0. The boxplot shows that values above $2x10^9$ are outliers.

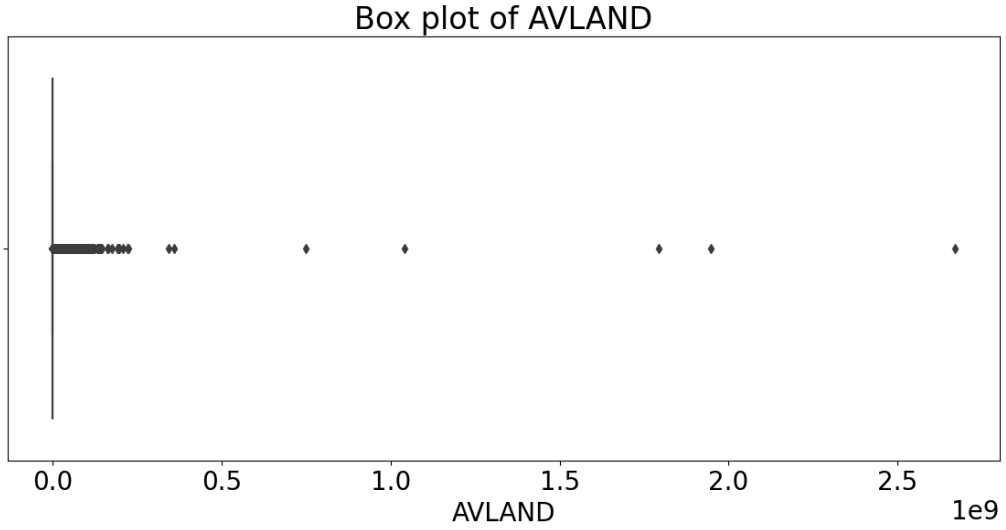

Distribution of FULLVAL
Density
40000
30000
20000
10000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
FULLVAL
1e6

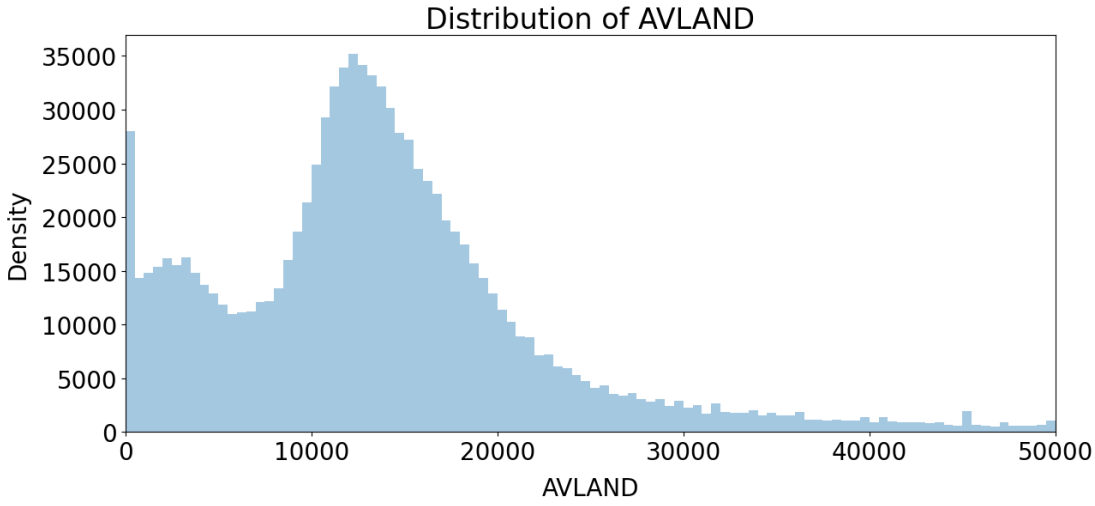

Box plot of FULLVAL
0
1
2
3
4
5
6
FULLVAL
1e9


Distribution of FULLVAL
Density
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0
FULLVAL
1e6

o. Field Name : **AVLAND**

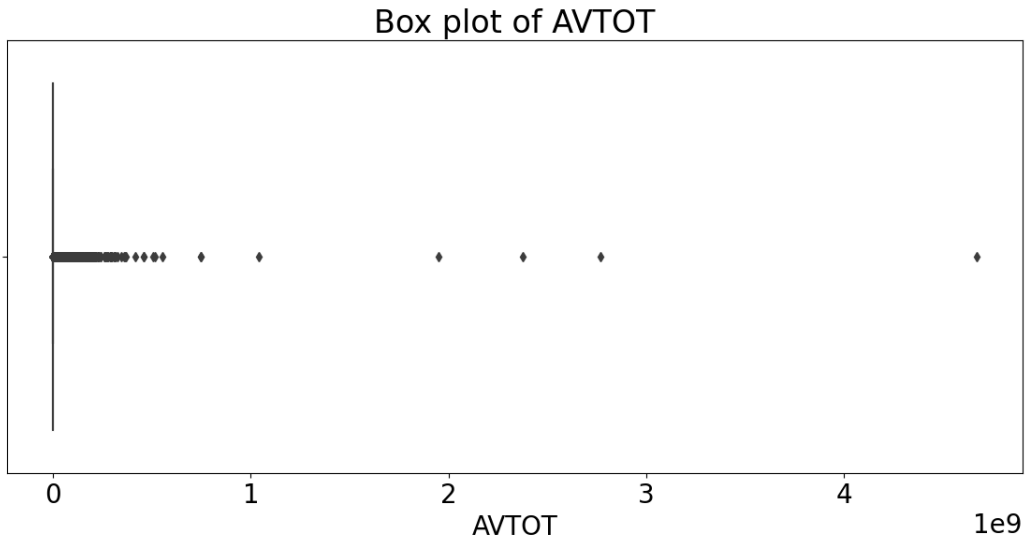
Description : Actual Land Value. This numerical field with values ranging from 0 to 2.6×10^9 with a mean value of 85067.92 and mode 0. This field has 70921 distinct values. Values above 4×10^8 are outliers.

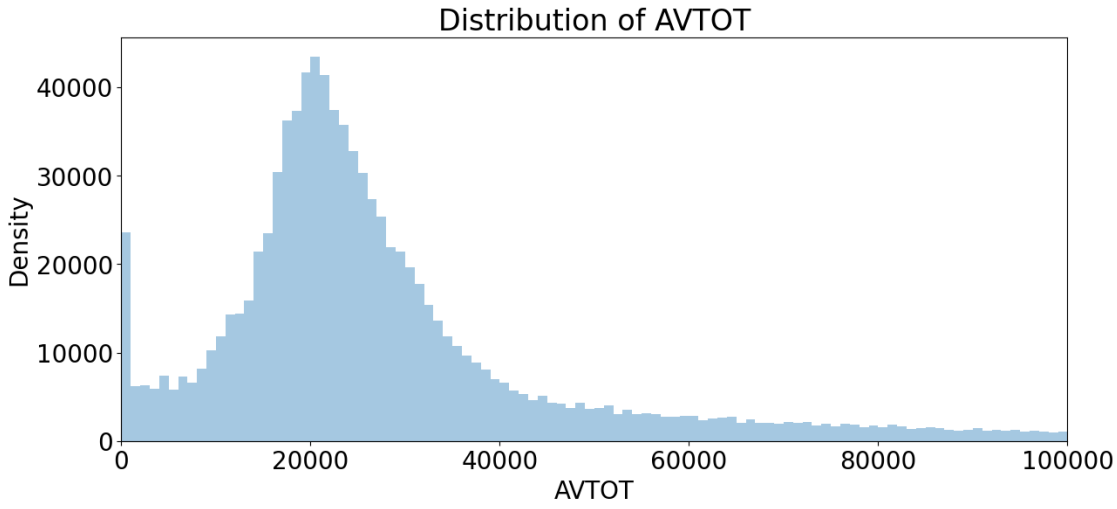




p. Field Name : **AVTOT**

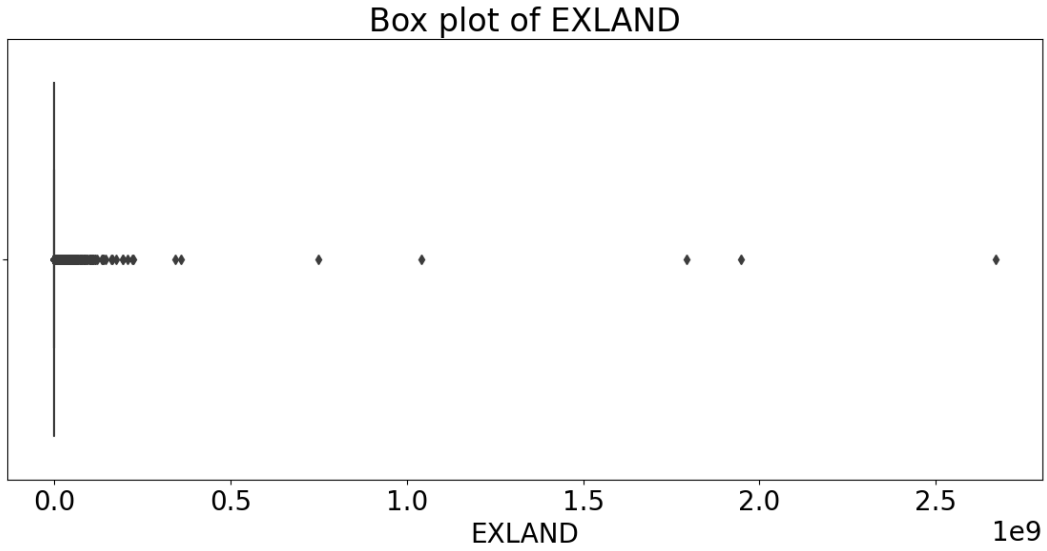
Description : Actual Total Value. This numerical field has values ranging from 0 to 4.6×10^9 with a mean of 227238.17 and mode 0. The boxplot shows that 1×10^9 are outliers. This field has 112914 unique values.

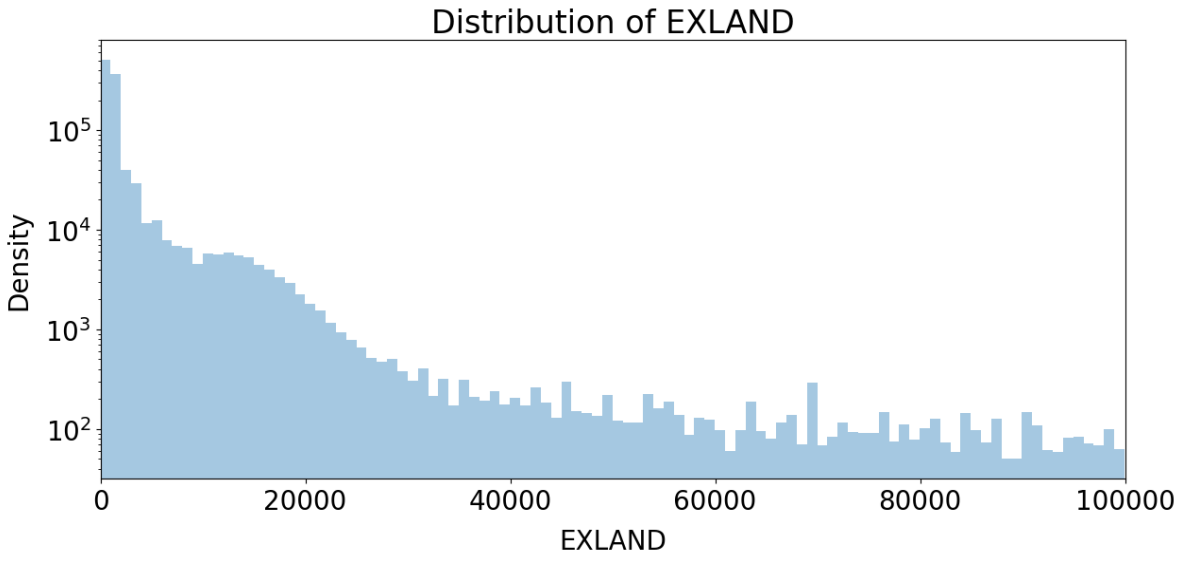




q. Field Name : **EXLAND**

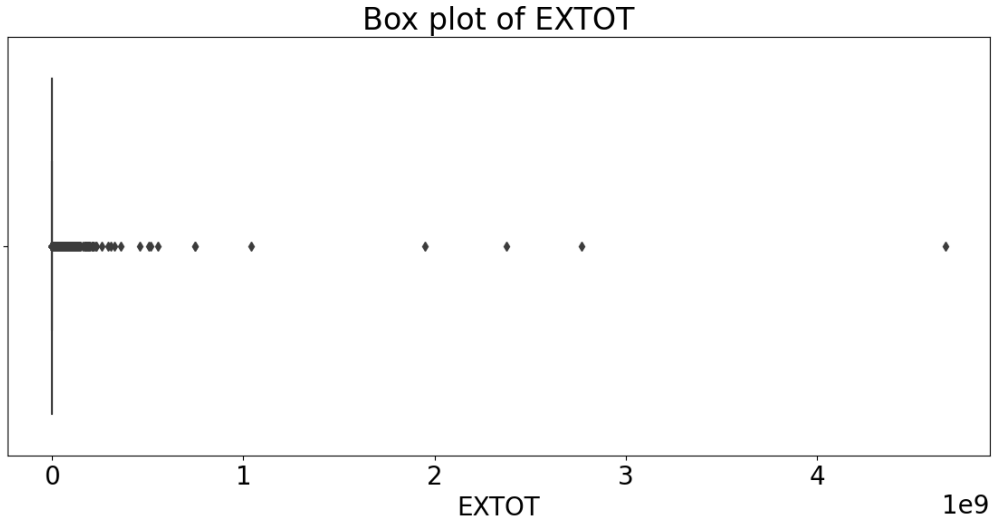
Description : Actual Exempt Land Value. This numerical field has values ranging from 0 to 2.6×10^9 with a mean value of 36423.89 and mode 0. This field has 33419 unique values. Values above 5×10^8 are outliers.

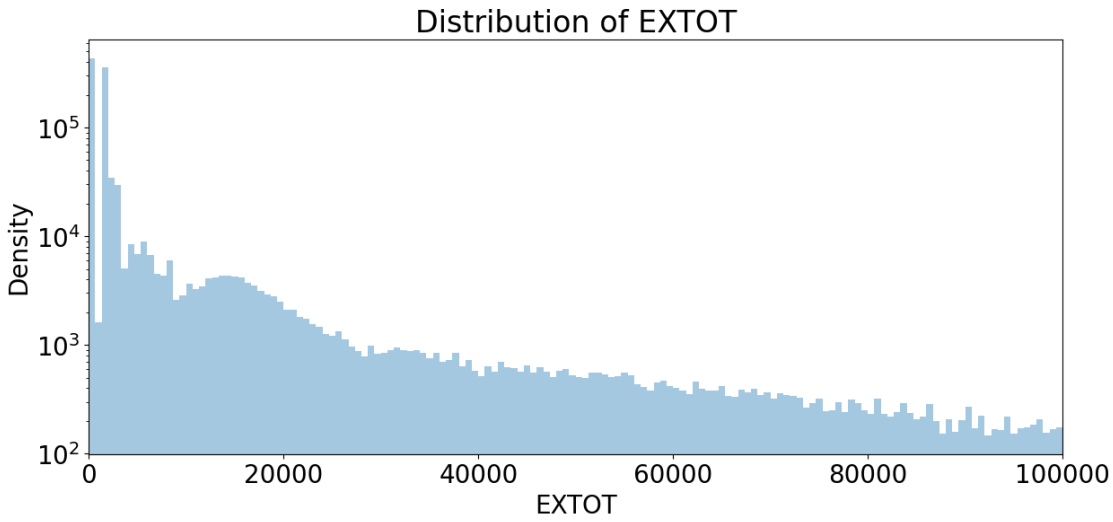




r. Field Name : **EXTOT**

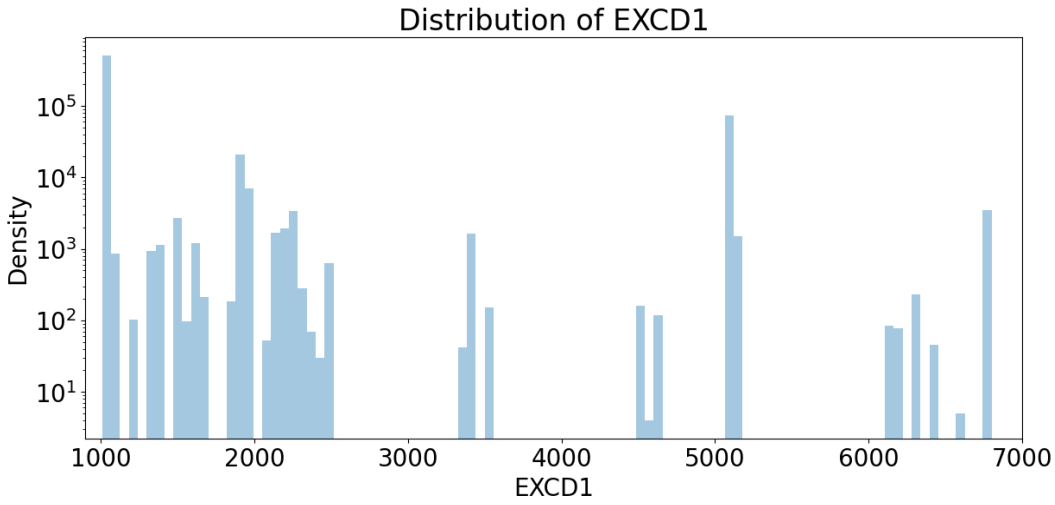
Description : Actual Exempt Land Total. This numerical field has values ranging from 0 to 4.6x109 with a mean of 91186.98 and mode of 0. This field has 64255 unique values. Values above $1x10^9$ are outliers.

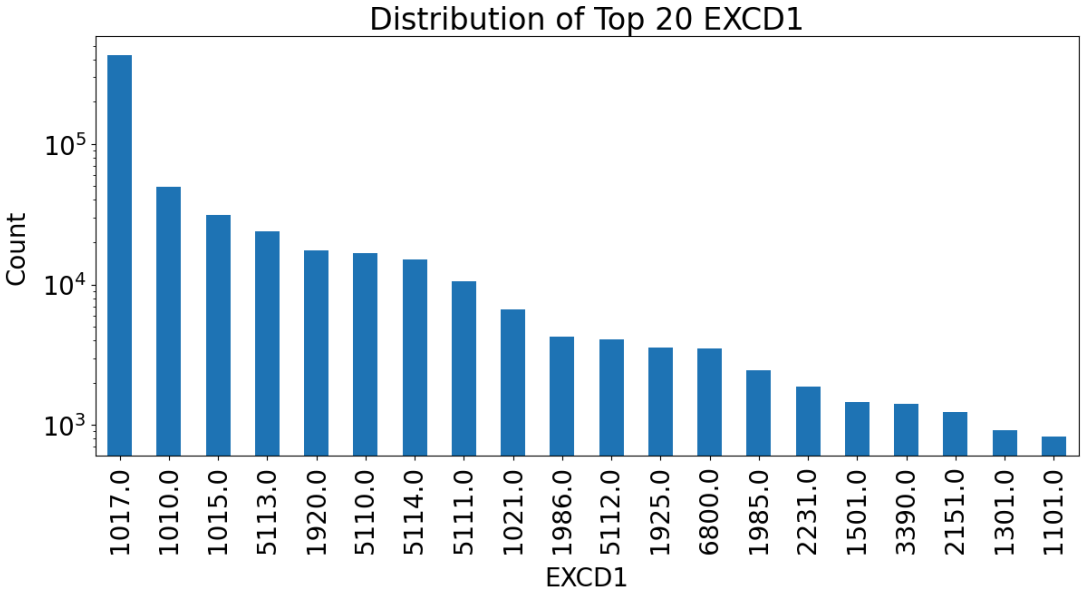


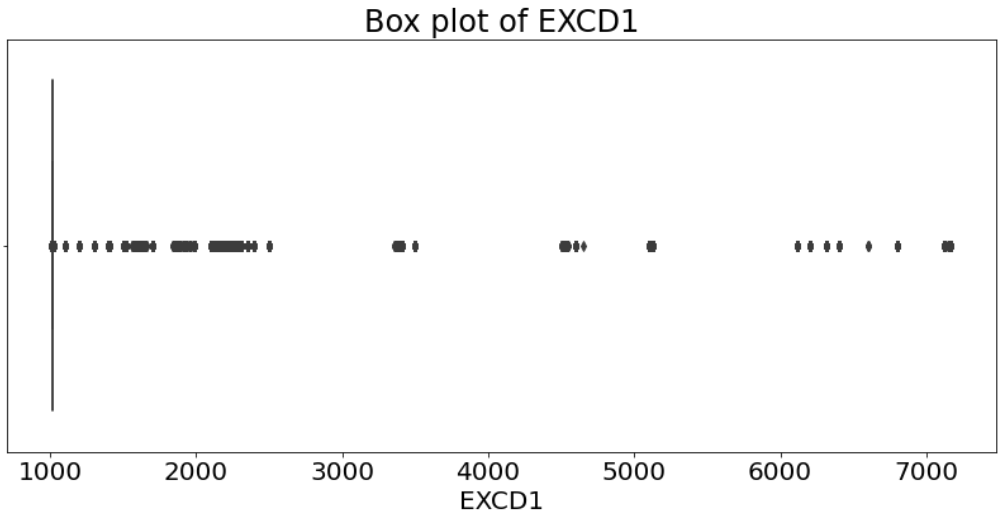


s. Field Name : **EXCD1**

Description : Exemption Code 1. This categorical field has 130 unique values with almost 60% values being 0 and “1017” being the most code. The values of the code range from 1000 to 7000 with a highly discontinuous distribution as seen in the count distribution plot below.

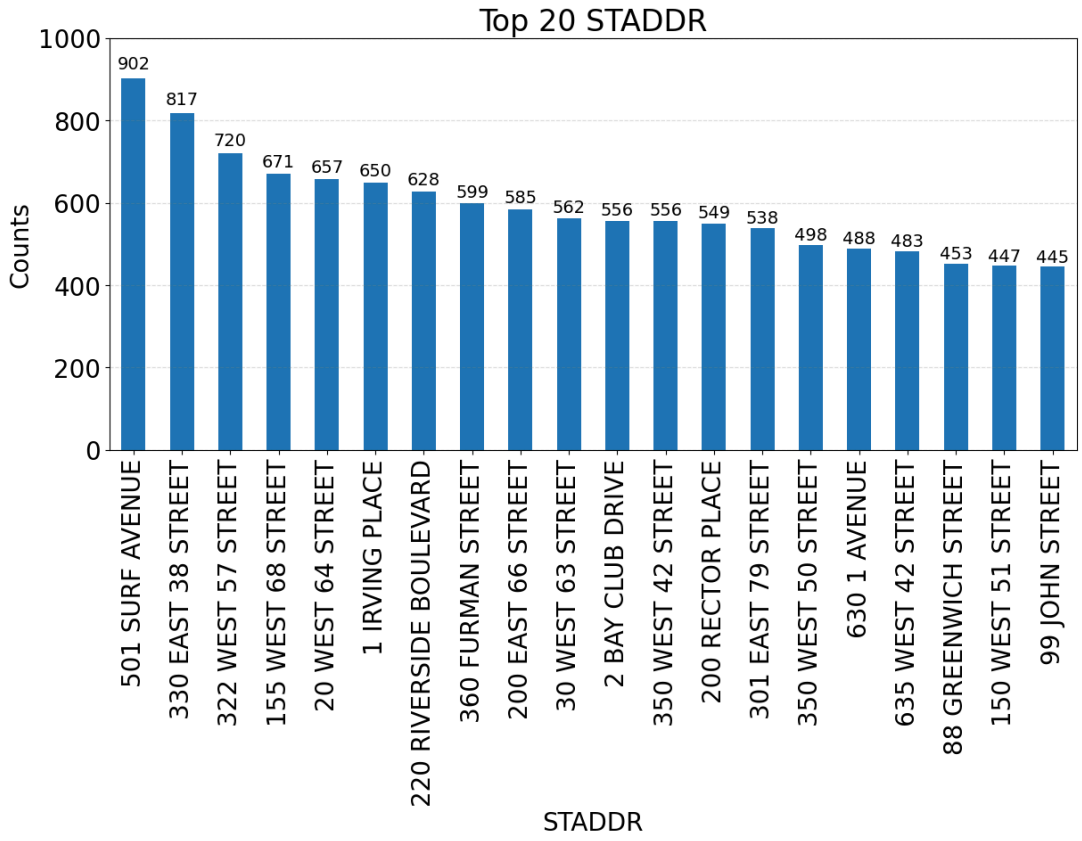

Distribution of EXCD1
Density
10^5
10^4
10^3
10^2
10^1
1000
2000
3000
4000
5000
6000
7000
EXCD1

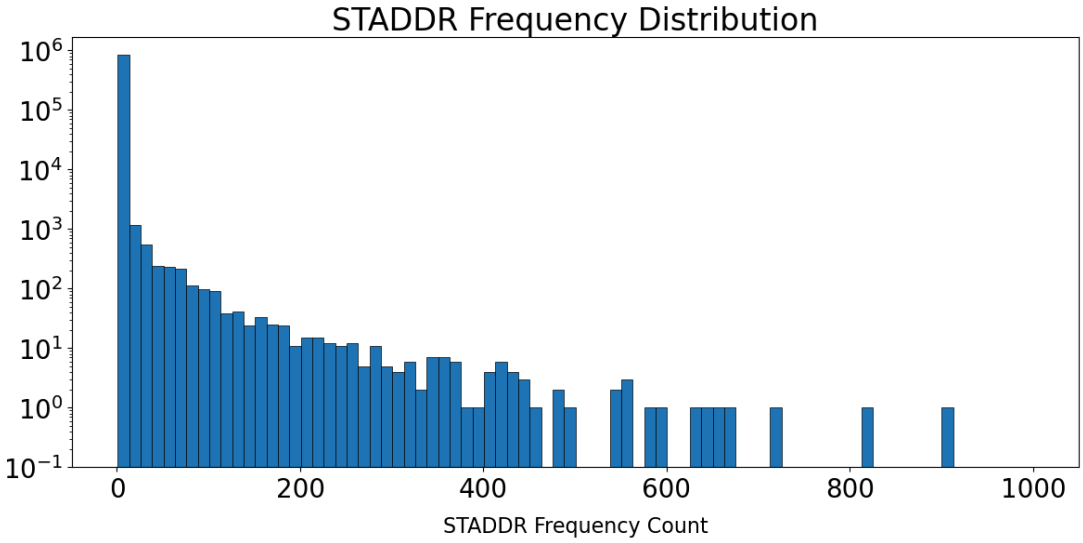

Distribution of Top 20 EXCD1
Count
10^5
10^4
10^3
1017.0
1010.0
1015.0
5113.0
1920.0
5110.0
5114.0
5111.0
1021.0
1986.0
5112.0
1925.0
6800.0
1985.0
2231.0
1501.0
3390.0
2151.0
1301.0
1101.0
EXCD1



t. Field Name : **STADDR**

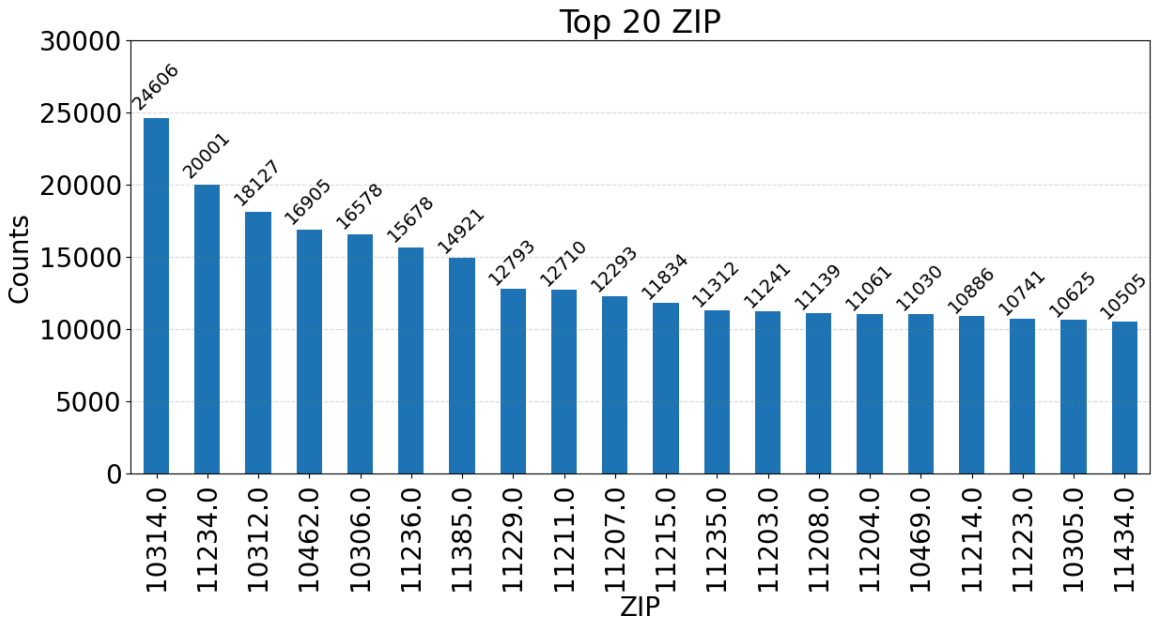
Description : Street Address. This categorical field has 839280 unique values with “501 SURF AVENUE” being the most common one, occurring about 900 times. Studying the frequency distribution plot we can see that most addresses occur only 1-50 times, but a few of them occur about 400-900 times in the dataset.





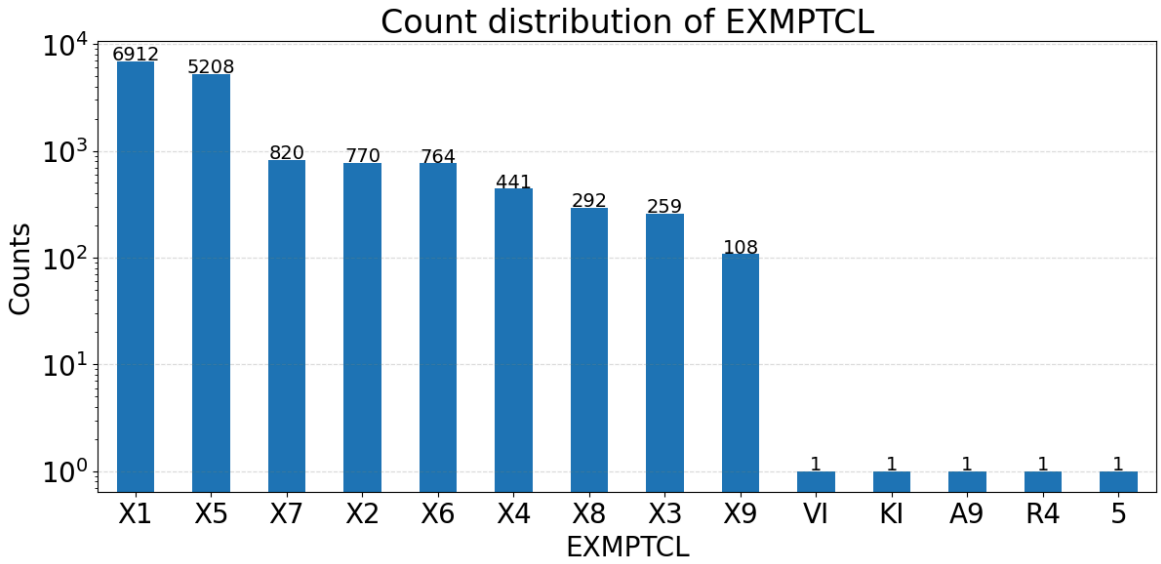
u. Field Name : **ZIP**

Description : Zip code. This categorical field has 196 unique values with “10314” being the most common value occurring 24606 times.



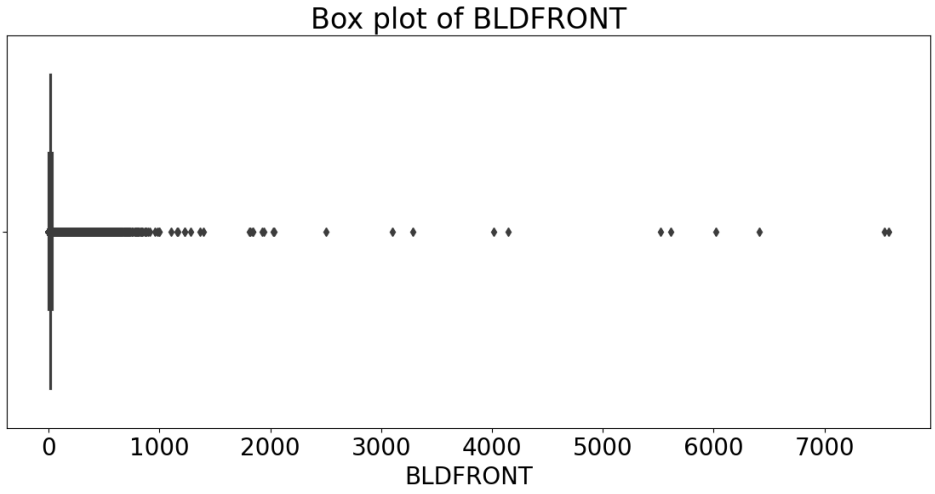
v. Field Name : **EXMPTCL**

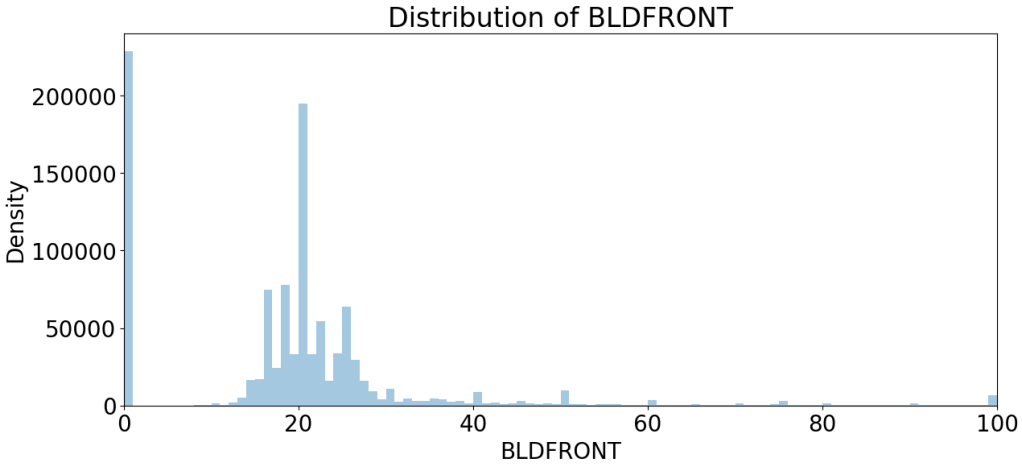
Description : Exemption Class. This categorical variable has 14 unique values with the most common value being “X1” occurring 6912 times.

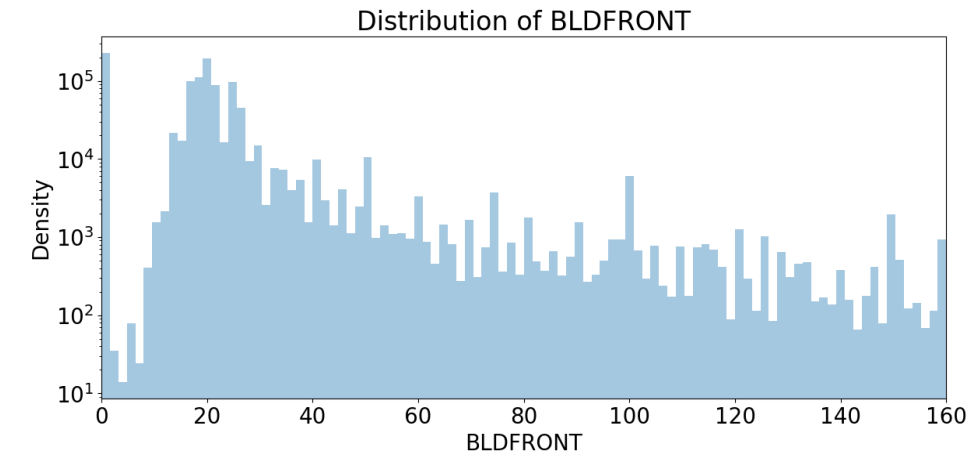


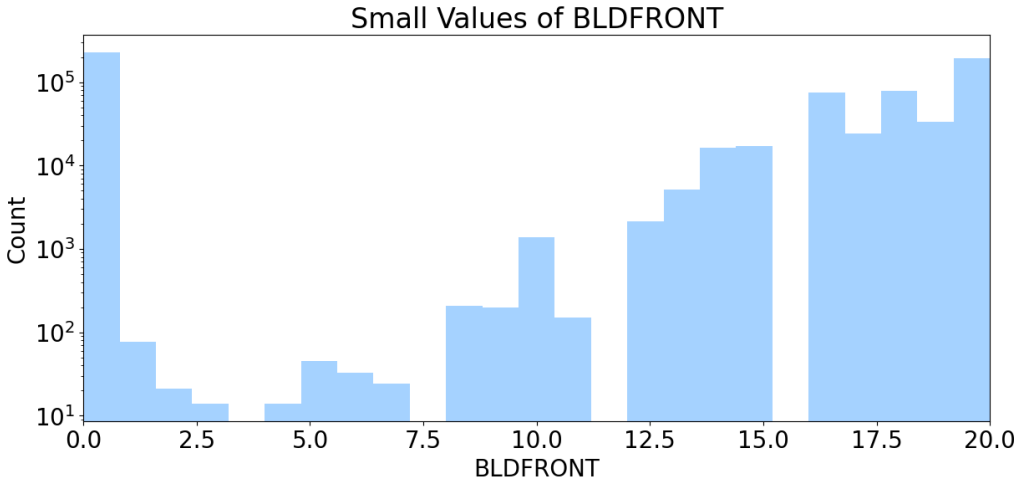
w. Field Name : **BLDFRONT**

Description : Building Width. This numerical field has values ranging from 0 to 7575 with a mean value of 23 and mode 0. This field has 612 distinct values. The values above 2500 are outliers.



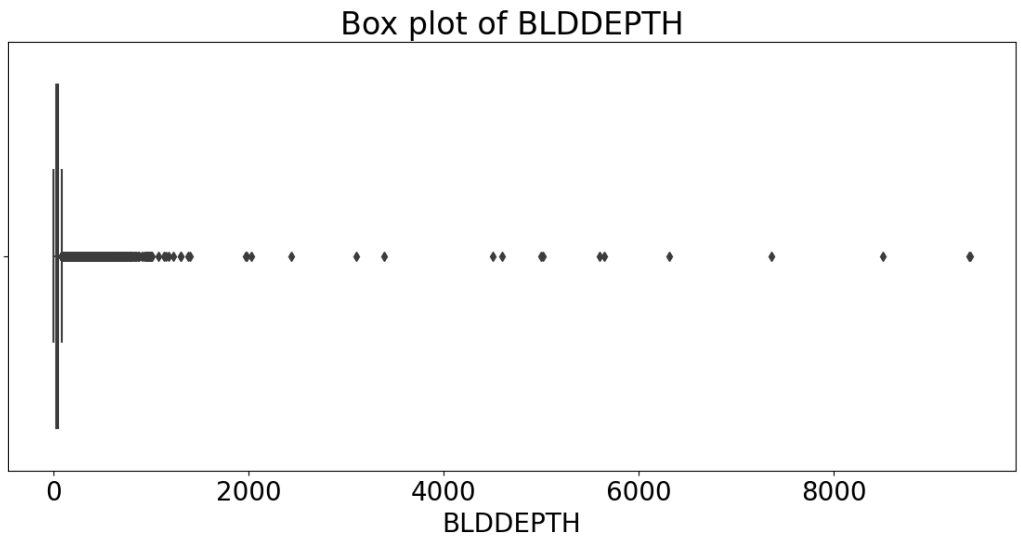


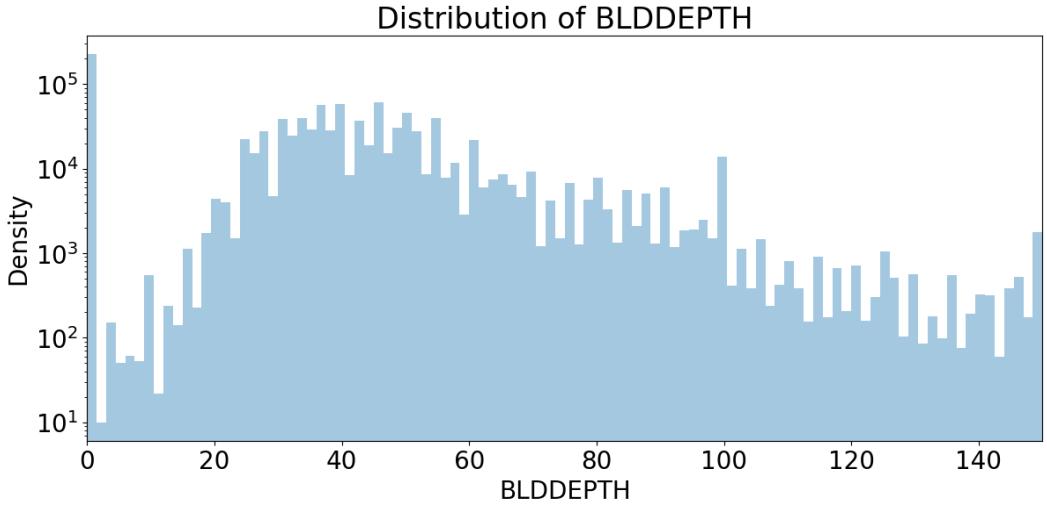


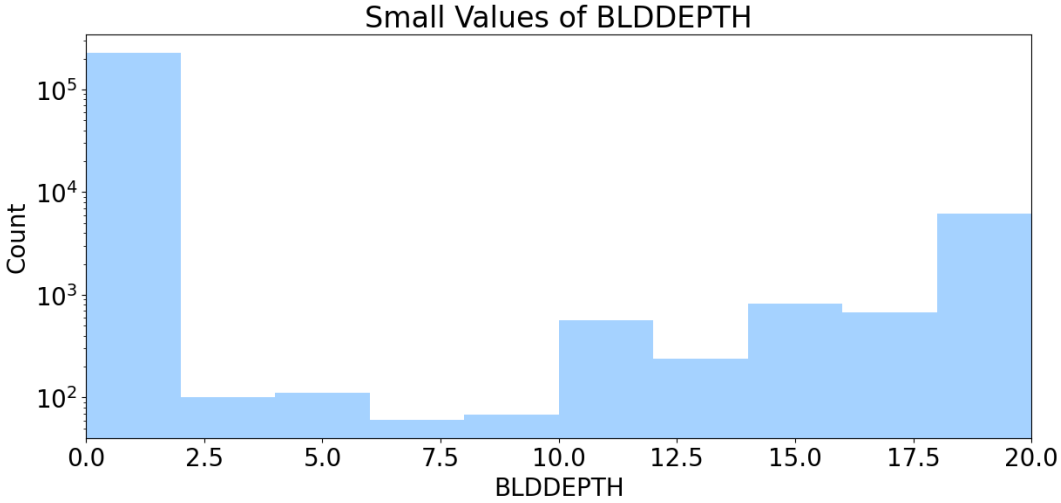


x. Field Name : **BLDDEPTH**

Description : Building Depth. This numerical field ranges from 0 to 9393 with a mean of about 40 with mode of 0. Values above 3000 are outliers.

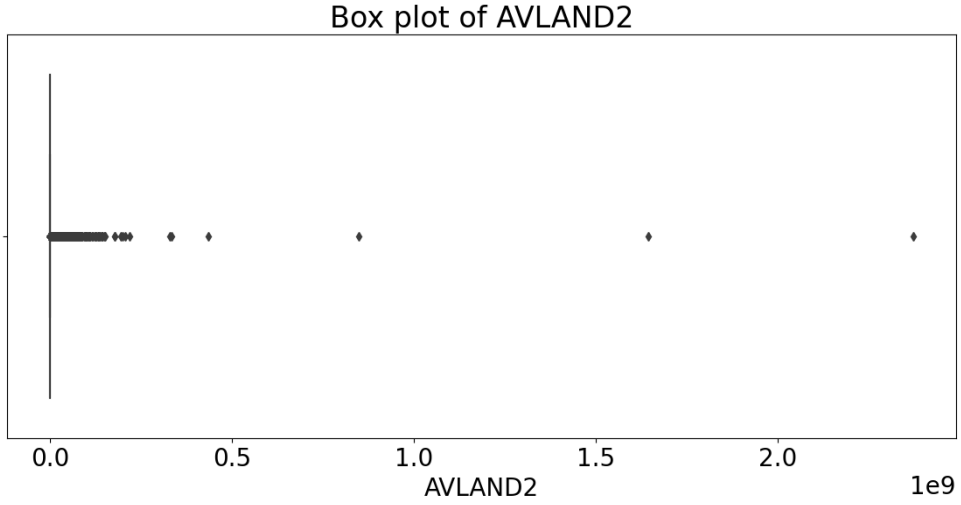

Box plot of BLDDEPTH
0
2000
4000
6000
8000
BLDDEPTH

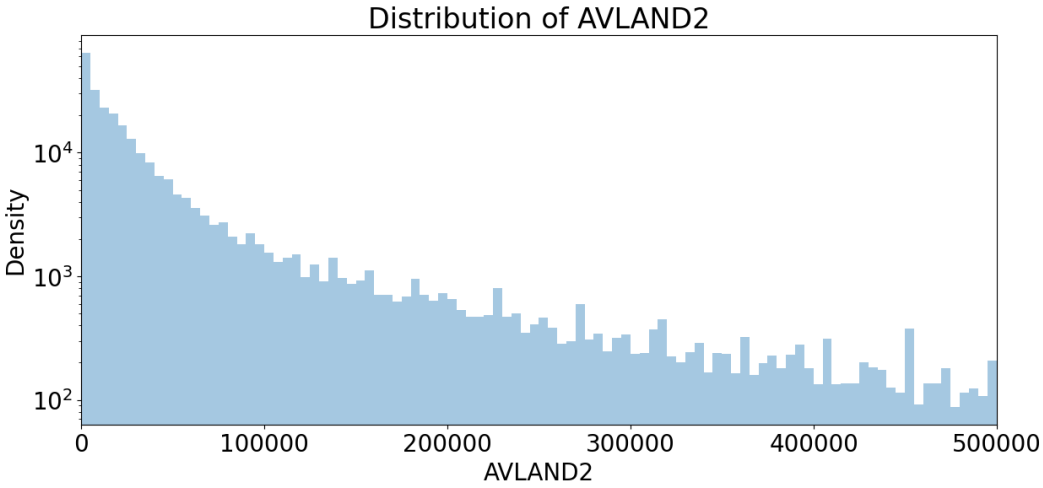

Distribution of BLDDEPTH
Density
10^5
10^4
10^3
10^2
10^1
0
20
40
60
80
100
120
140
BLDDEPTH


Small Values of BLDDEPTH
Count
10^5
10^4
10^3
10^2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
BLDDEPTH

y. Field Name : **AVLAND2**

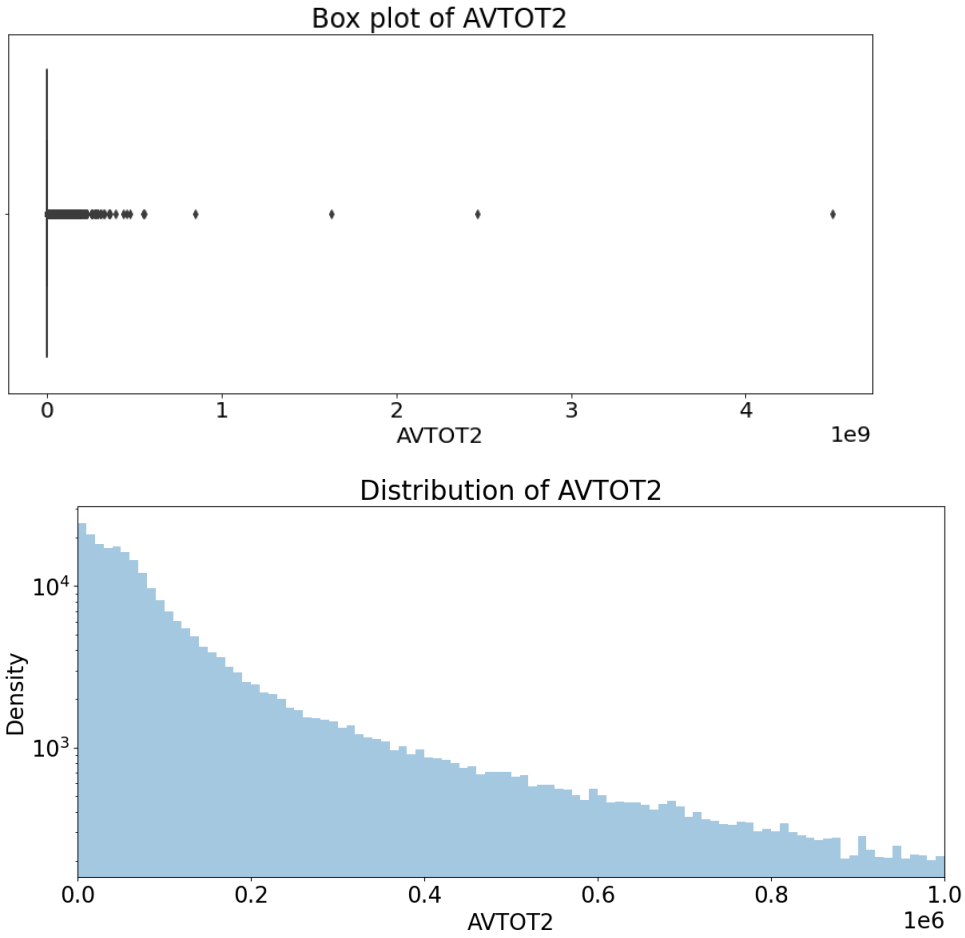
Description : Transitional Land Value. This numerical field value ranges from 3 to $2.3x10^9$ with a mean value of 246235.72 and mode 2408. This field has 58592 distinct values. The boxplot below shows values above 5x108 are outliers.





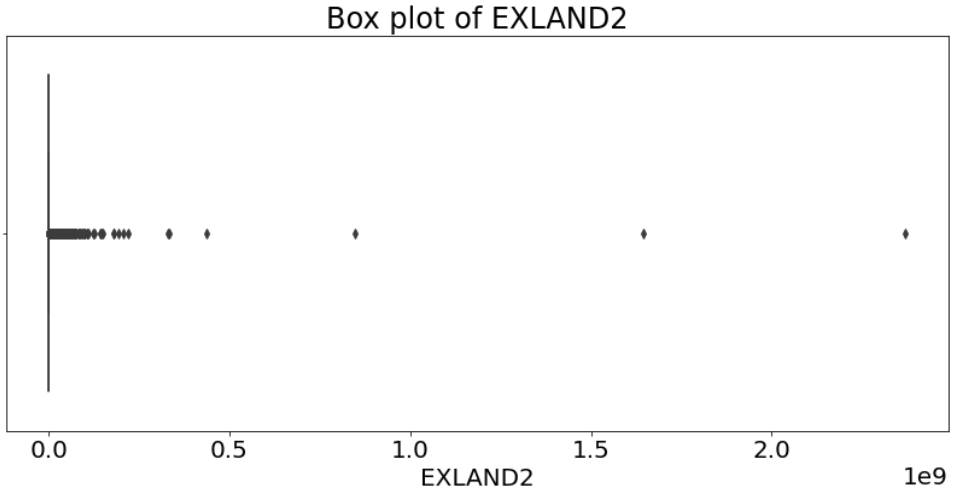
z. Field Name : **AVTOT2**

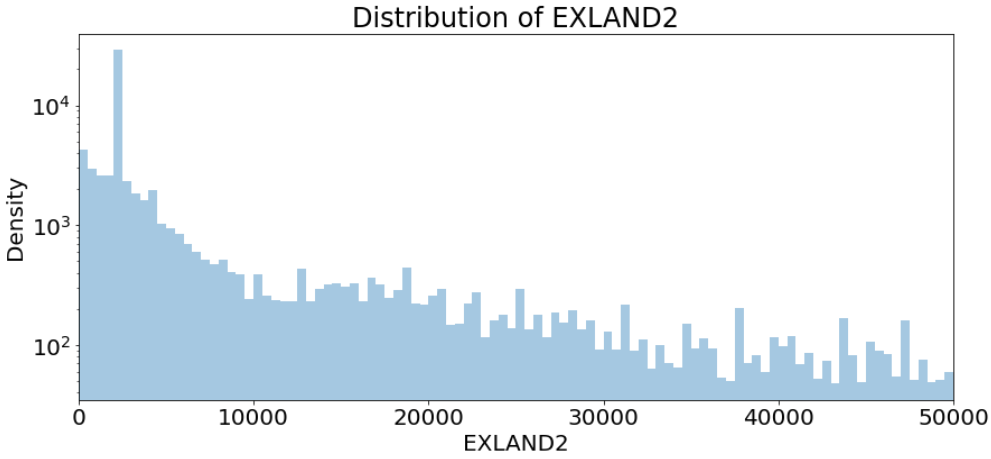
Description : Transitional Total Value. This numerical field value ranges from 3 to $4.5x10^9$ with a mean value 713911.44 and mode 750.



aa. Field Name : **EXLAND2**

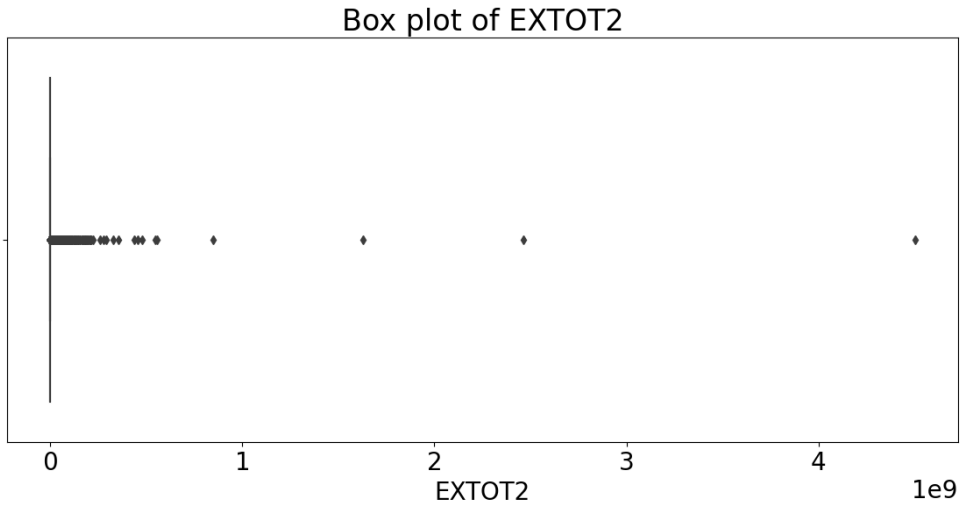
Description : Transitional Exemption Land Value. This numerical field value ranges from 1 to 2.3×10^9 with a mean value of 351235.68 and mode 2090.

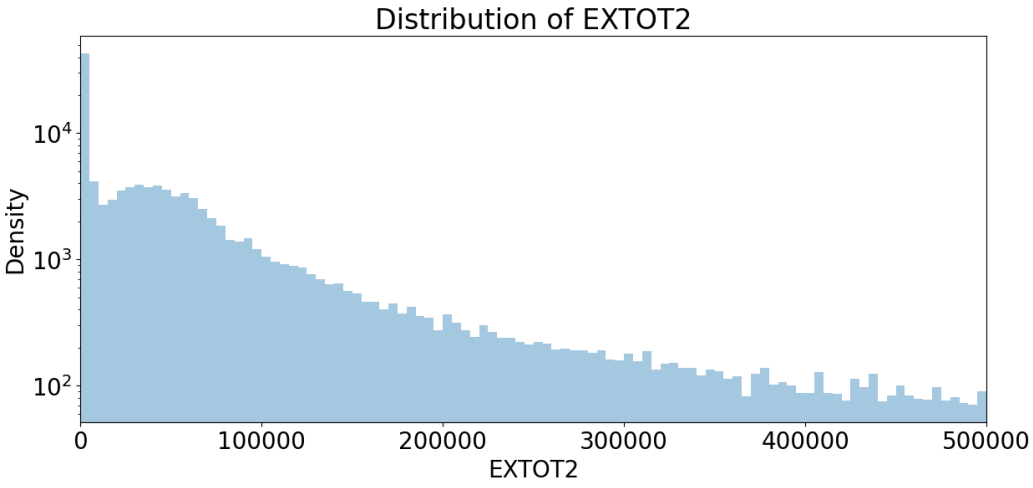




bb. Field Name : **EXTOT2**

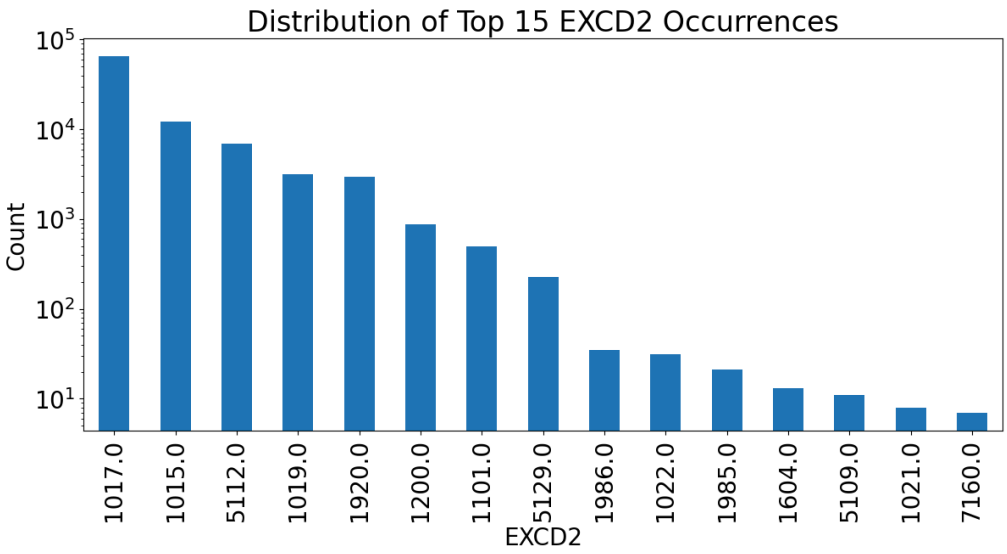
Description : Transitional Exemption Land Total. This numerical field ranges from 7 to 4.5×10^9 with a mean of 656768.28 and mode 2090.

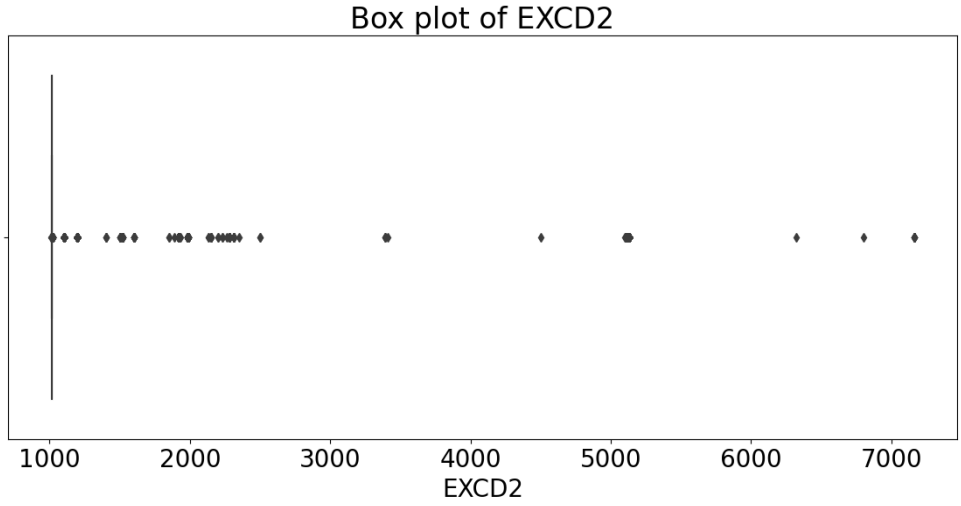


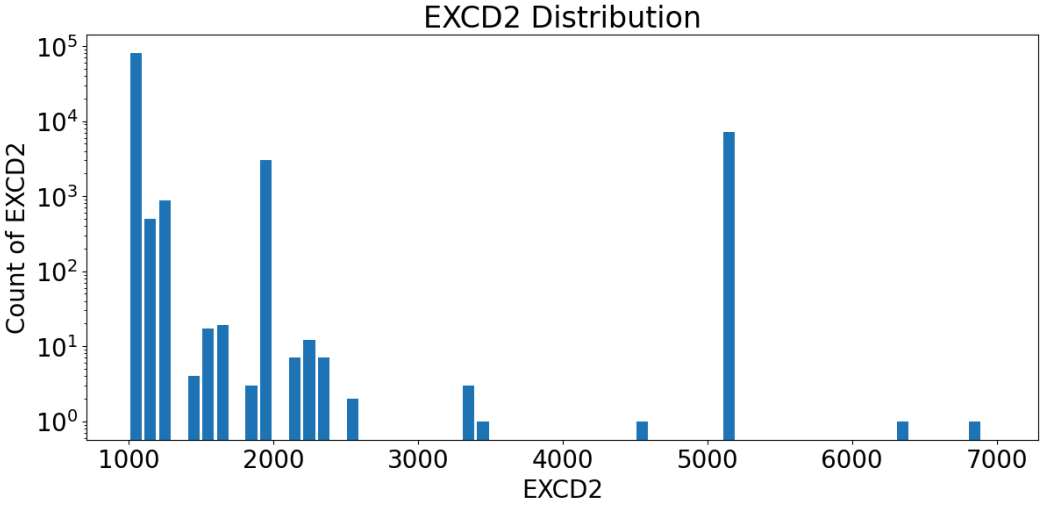


cc. Field Name : **EXCD2**

Description : Exemption Code 2. This categorical field has 60 unique values with “1017” being most common.


Distribution of Top 15 EXCD2 Occurrences
Count
EXCD2
1017.0
1015.0
5112.0
1019.0
1920.0
1200.0
1101.0
5129.0
1986.0
1022.0
1985.0
1604.0
5109.0
1021.0
7160.0


Box plot of EXCD2
EXCD2


EXCD2 Distribution
Count of EXCD2
EXCD2

dd. Field Name : **PERIOD**

Description : Assessment Period. The categorical field has one unique value “FINAL”.

ee. Field Name : **YEAR**

Description : Assessment Year. This categorical field has only one unique value “2010/11”.

ff. Field Name : **VALTYPE**

Description : Value Type. This categorical field has only unique value “AC-TR”.