Data Quality Report

Sheetal Srivastava

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.

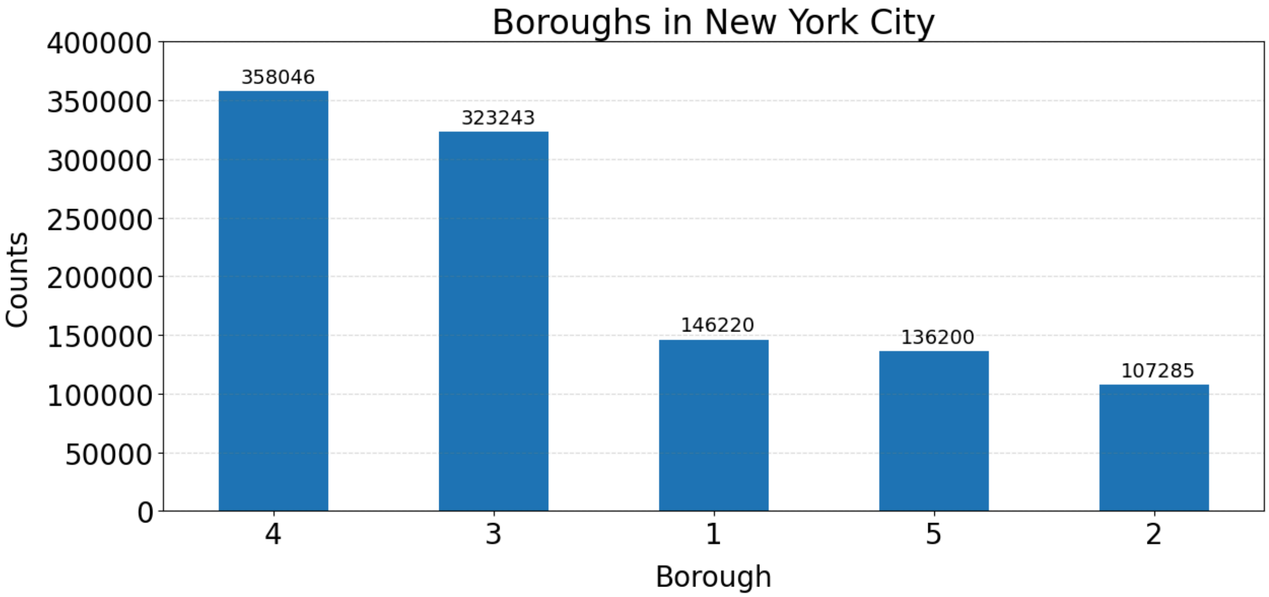
b. Field Name : **BBLE**

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**

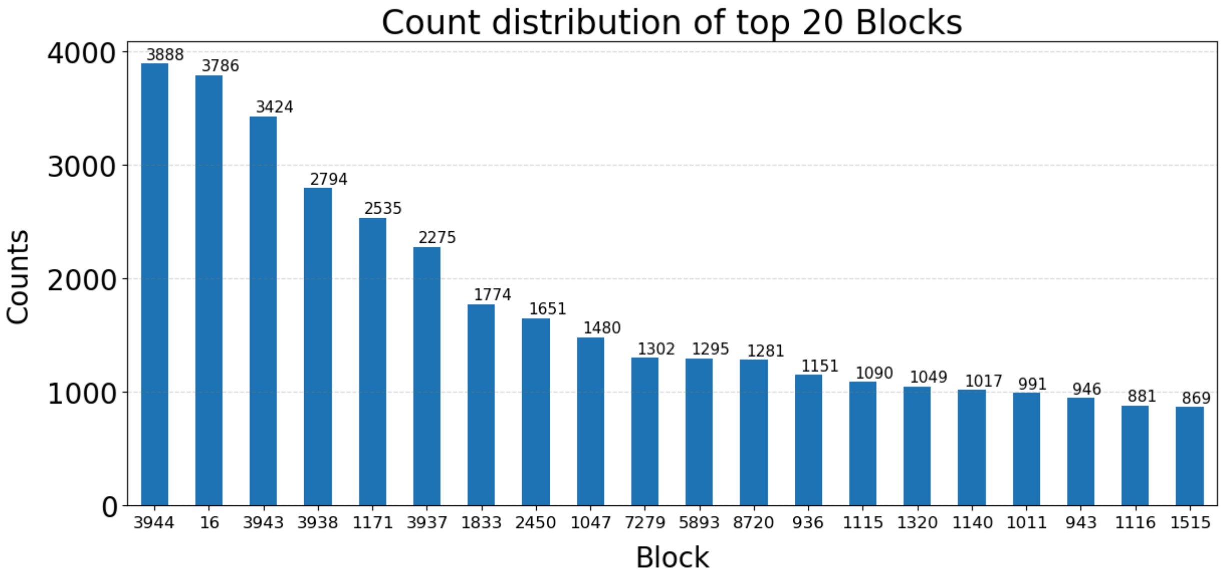
Description : This categorical field represents the borough. It has 5 unique values which are listed below. The plot below shows the count distribution of each category. From the plot we can see that Queens (4) has the highest number of properties listed in the dataset.

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



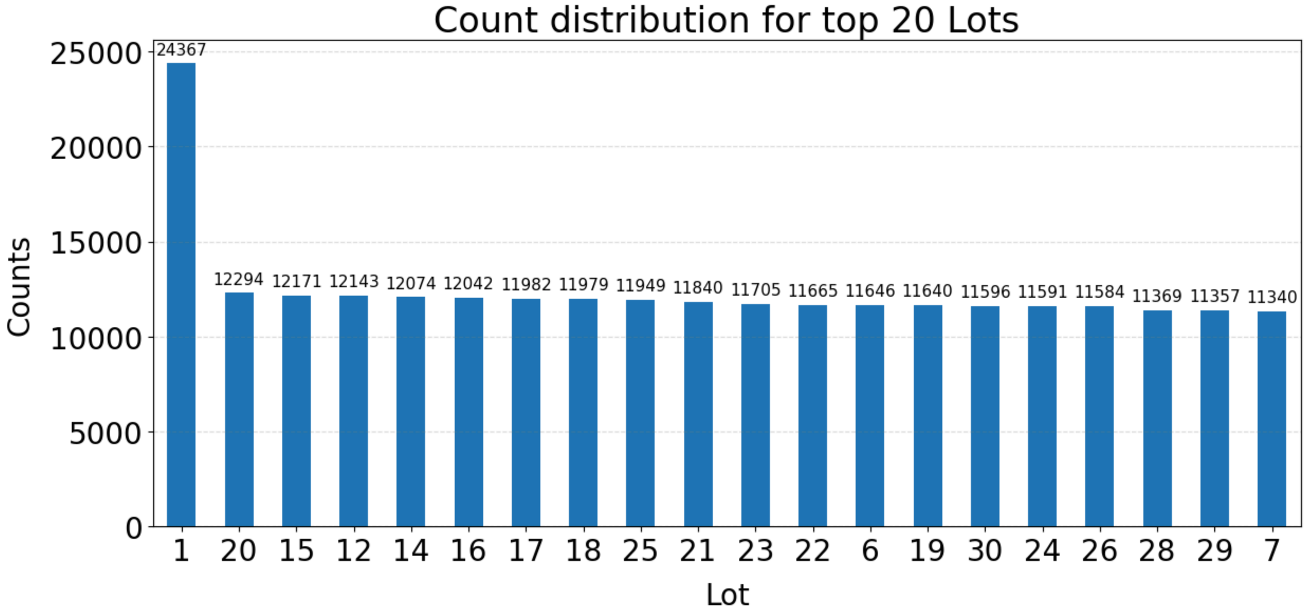
d. Field Name : **BLOCK**

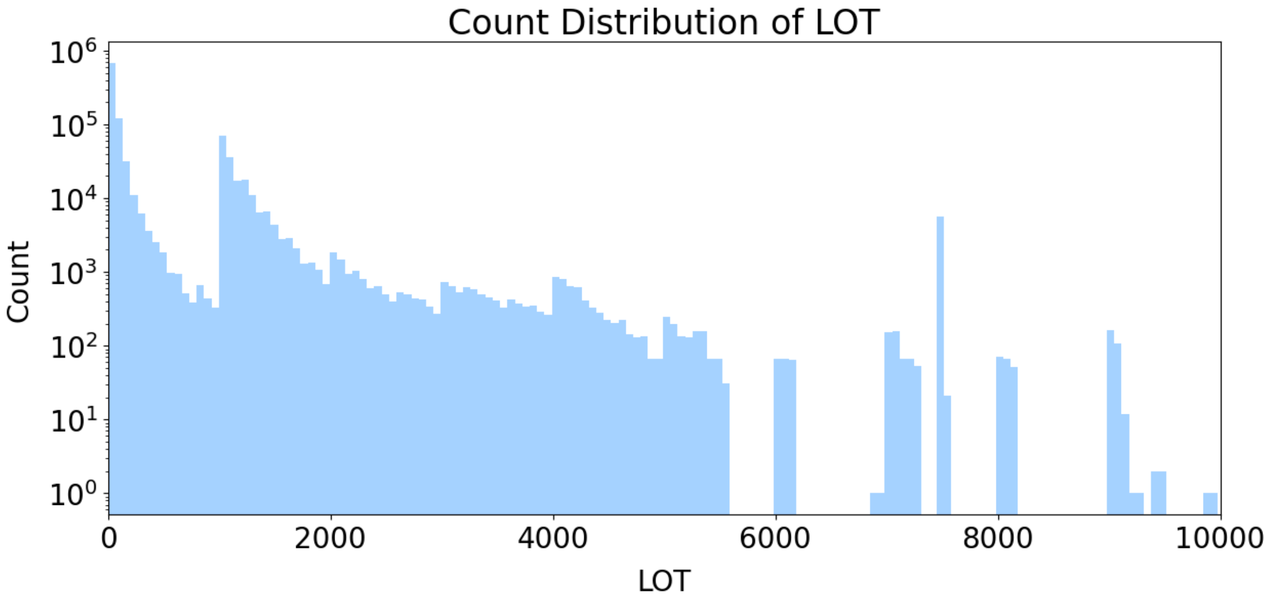
Description : This is the valid block range by borough. This categorical field has 13984 distinct values. We see that block 3944 has the highest number of properties listed.



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.

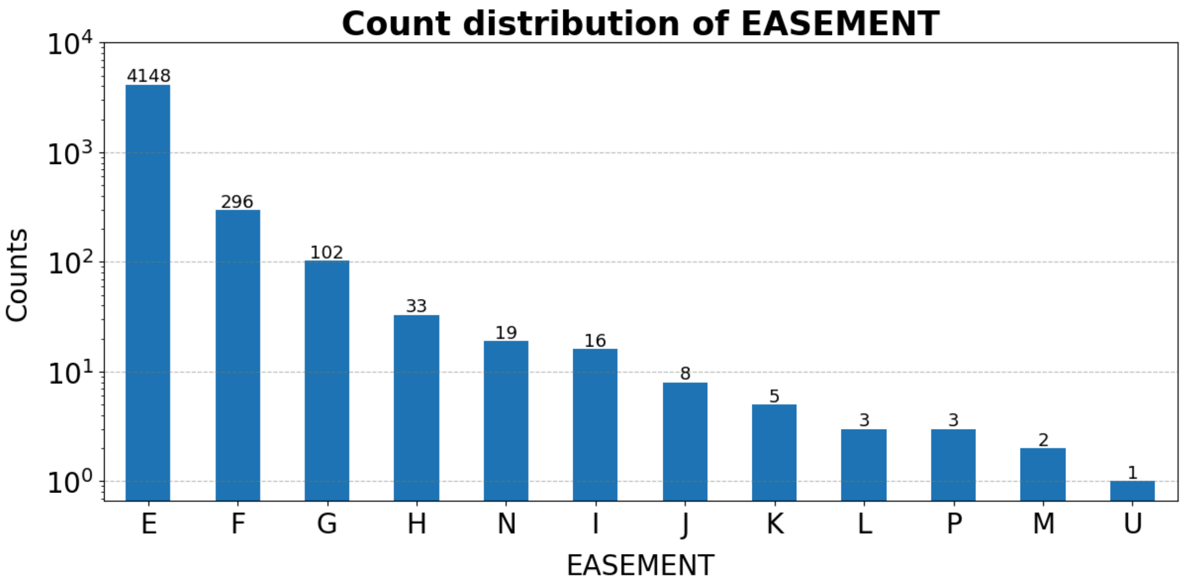




f. Field Name : **EASEMENT**

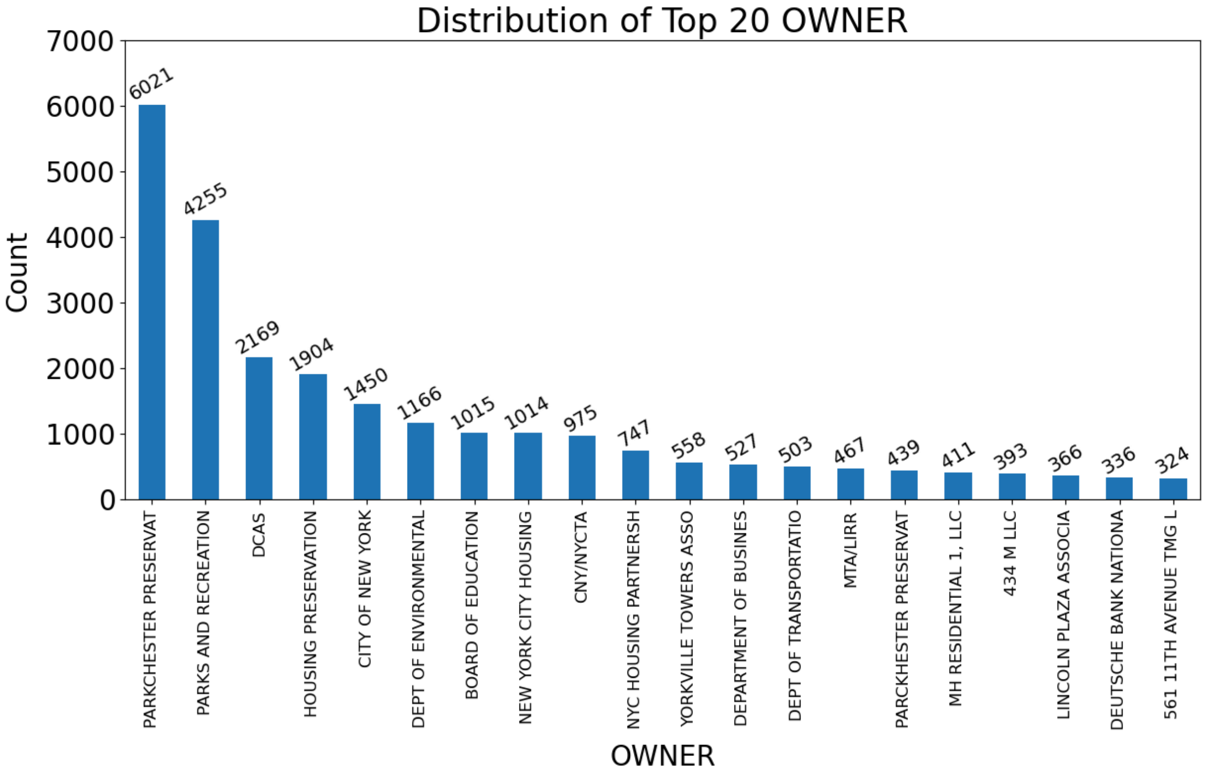
Description : This field specifies the easement rights the property holds. This is a categorical field that contains either a space (converted to null) or a single letter indicating the easement type. This field has 13 unique values, or 12 unique values if we do not count the first one (i.e. space) :

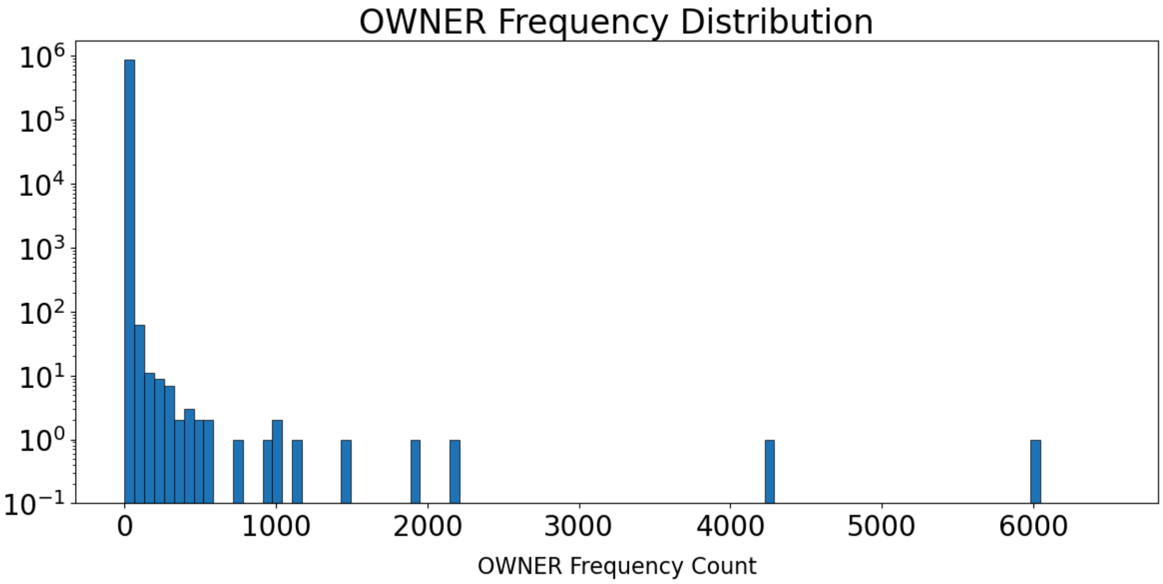
- E = Land Easement
- F, G, H, I, J, K, L, M = are duplicates of E
- N = Non-Transit Easement
- P = Pier
- U = U.S. Government



g. Field Name : **OWNER**

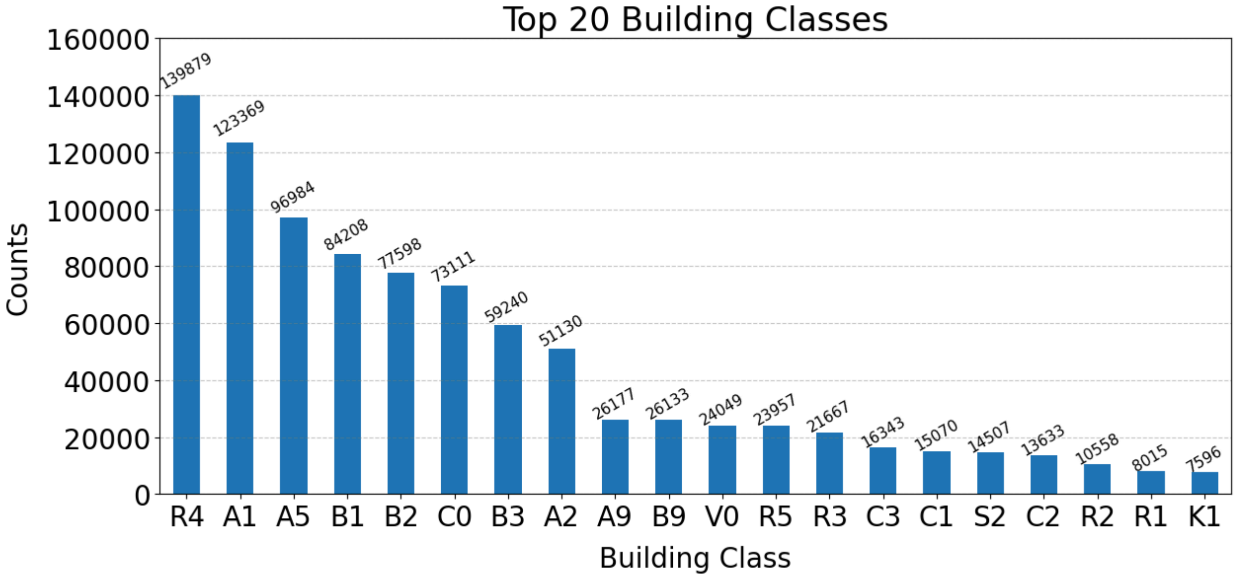
Description : Name of the property owner. This field has 863348 distinct values with about 97% of the records populated with values. The most common owner is “PARKCHESTER PRESERVAT” with 6021 properties. The first plot shows the top 20 owners along with their names. The second plot shows the frequency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

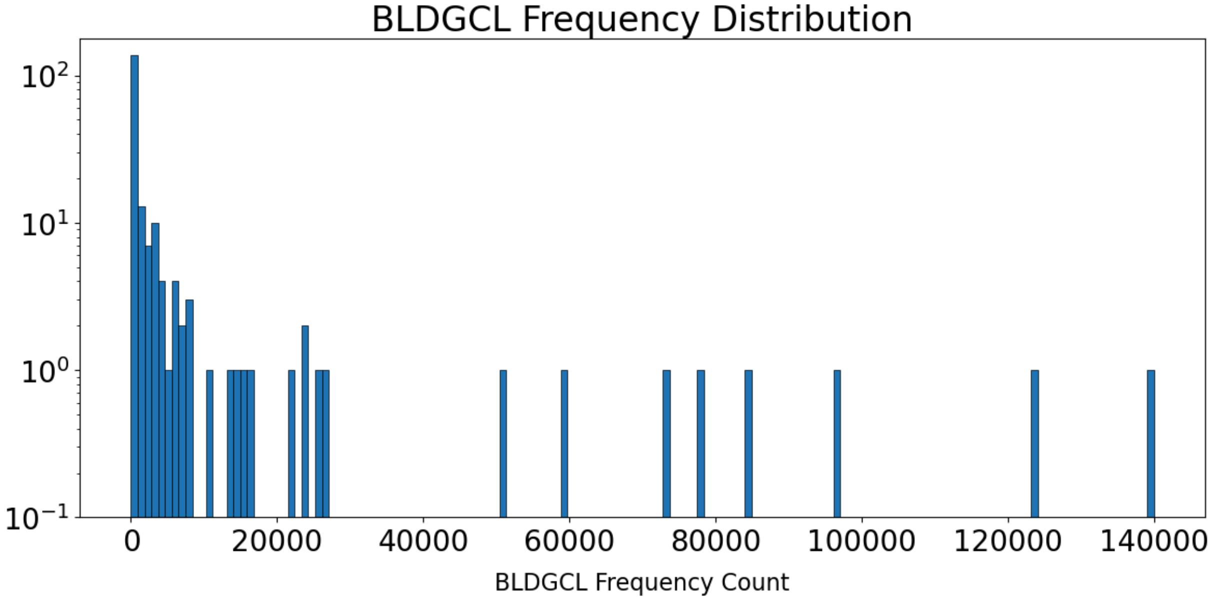




h. Field Name : **BLDGCL**

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.

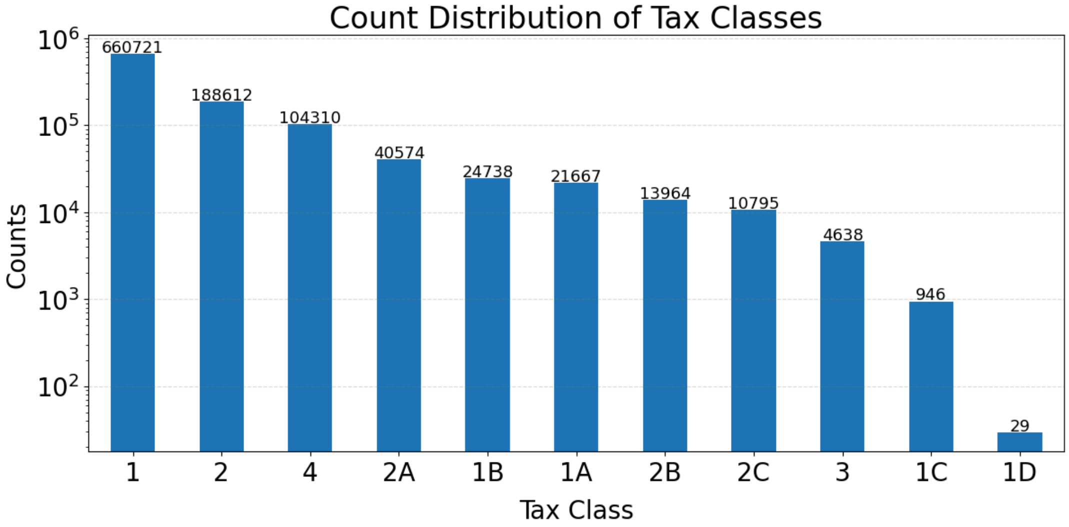




i. Field Name : **TAXCLASS**

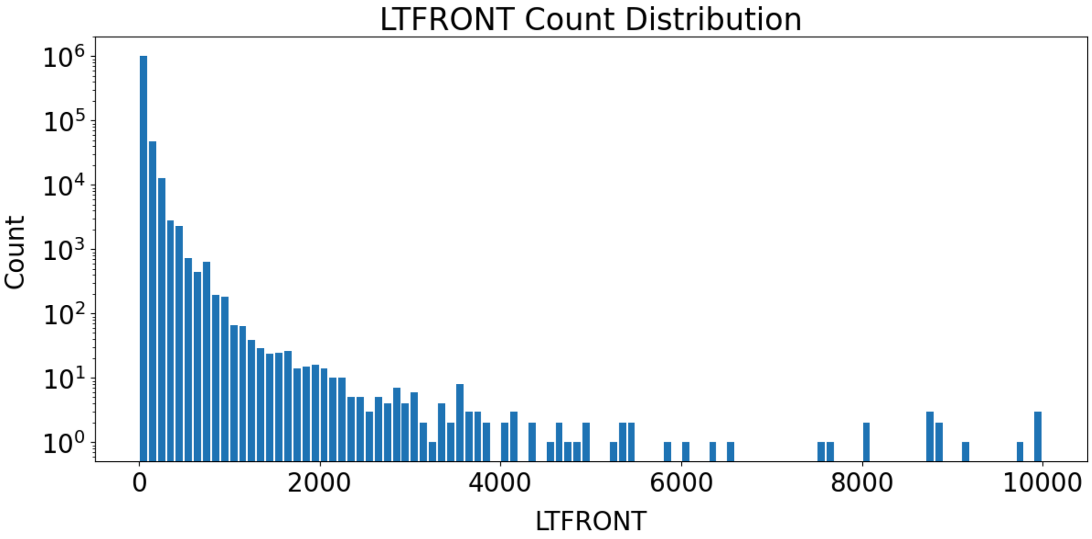
Description : Tax Class. This categorical field indicates the type of tax category the property falls in. This field has 11 distinct classes with the most common class being “1”. Following is the description of some of the classes :

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

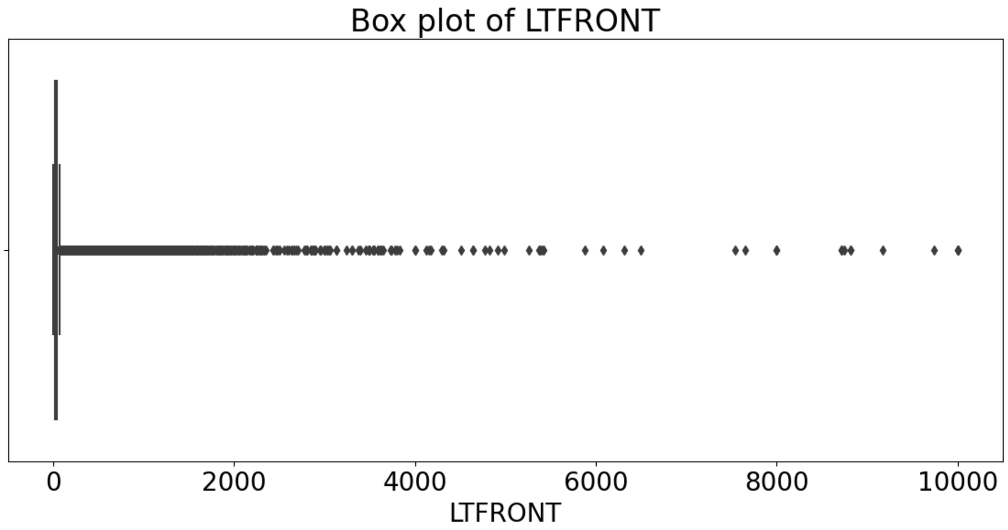


j. Field Name : **LTFRONT**

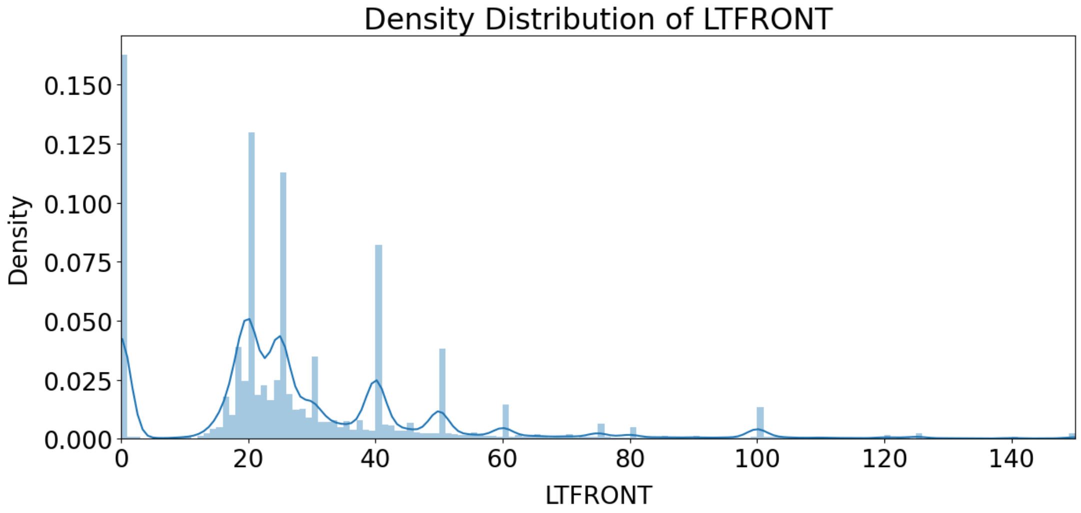
Description : Lot width. This numerical field has values ranging from 0 to 9999, with a mean value of 36.64 and close to 16% of the records are 0.



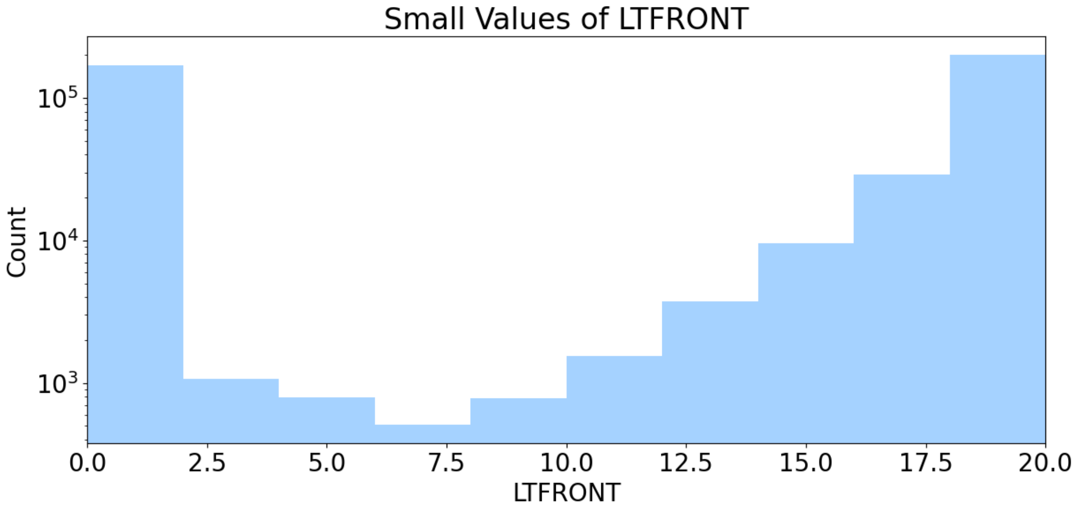
LTFRONT values above 6000 are low in occurrence and hence are possibly outliers. The values are more concentrated around the smaller range (0-50).



Box plot of LTFRONT



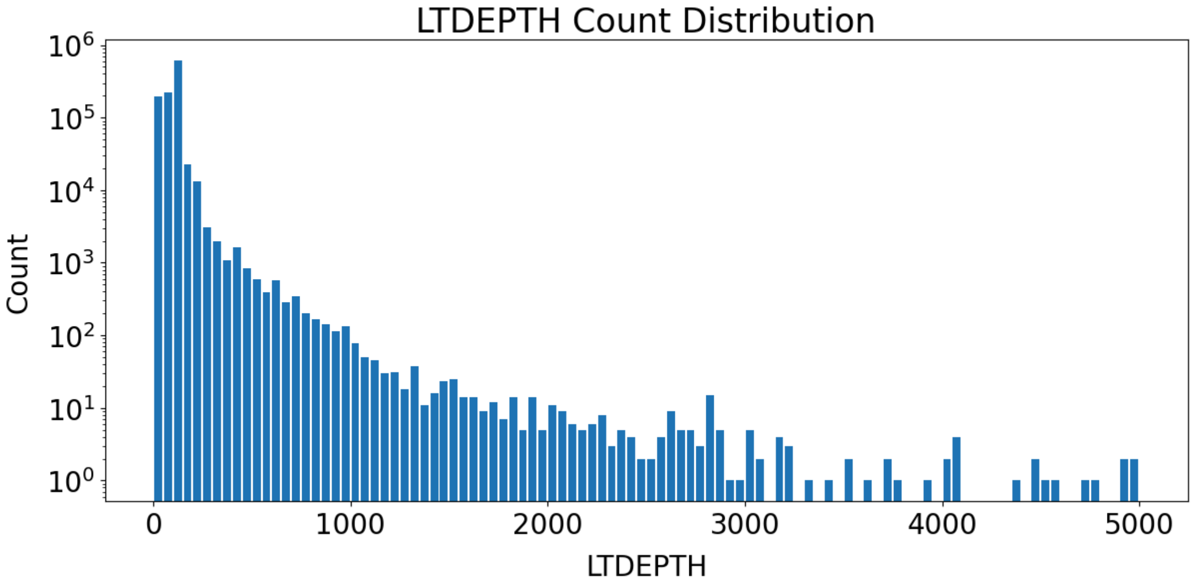
Density Distribution of LTFRONT

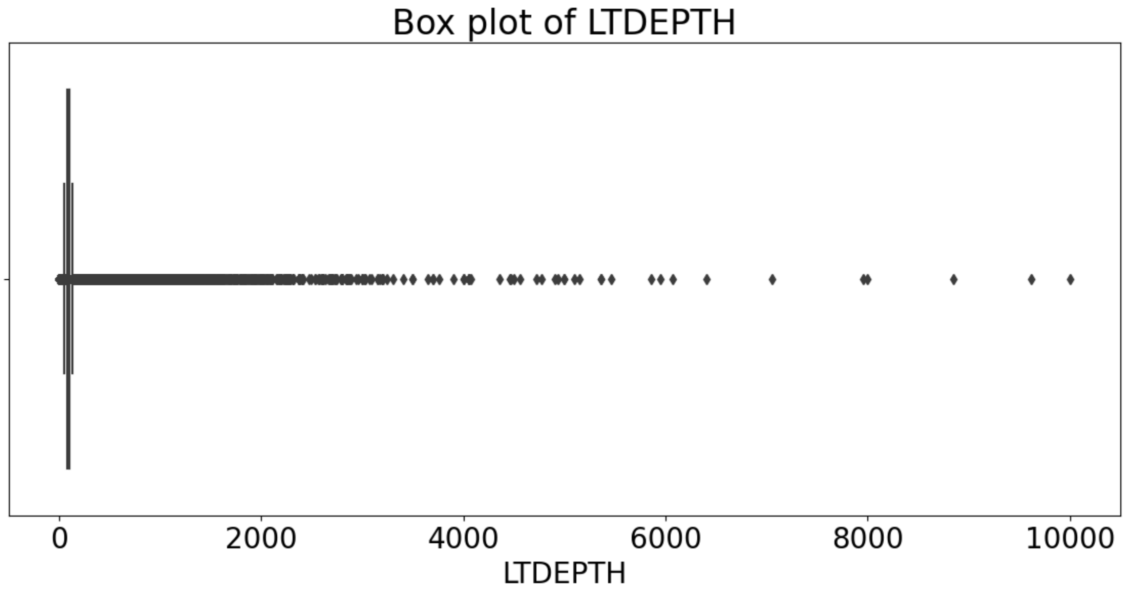


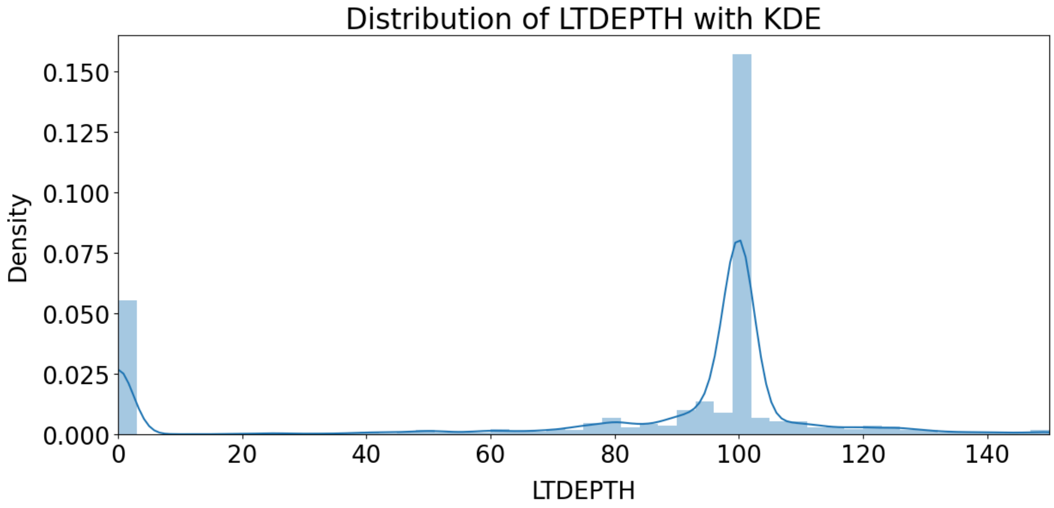
Small Values of LTFRONT

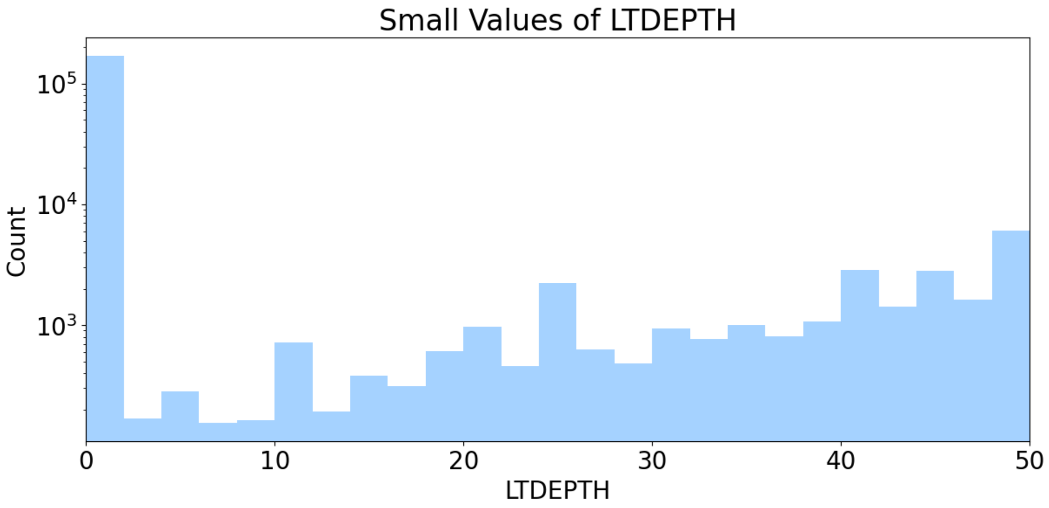
k. Field Name : **LTDEPTH**

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.



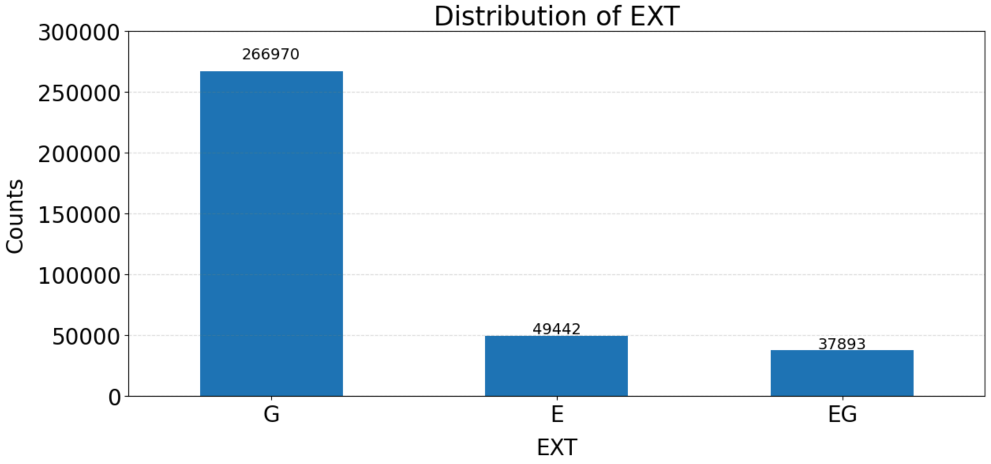






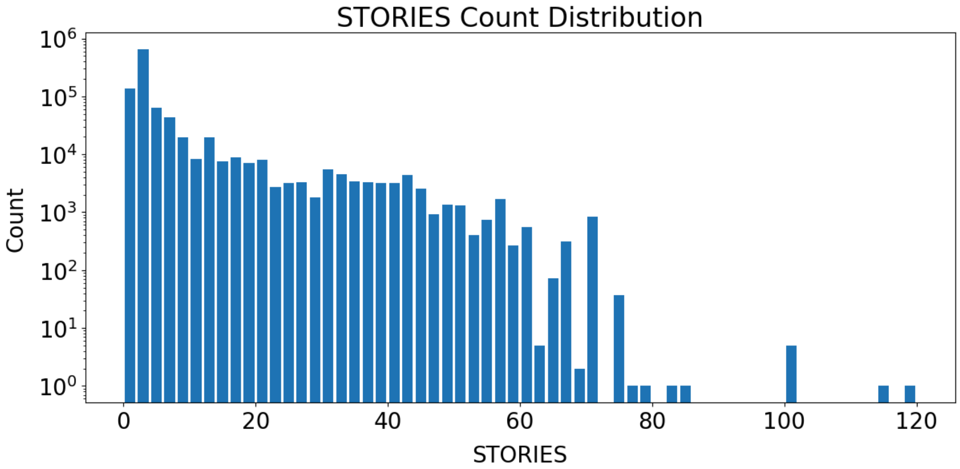
I. Field Name : **EXT**

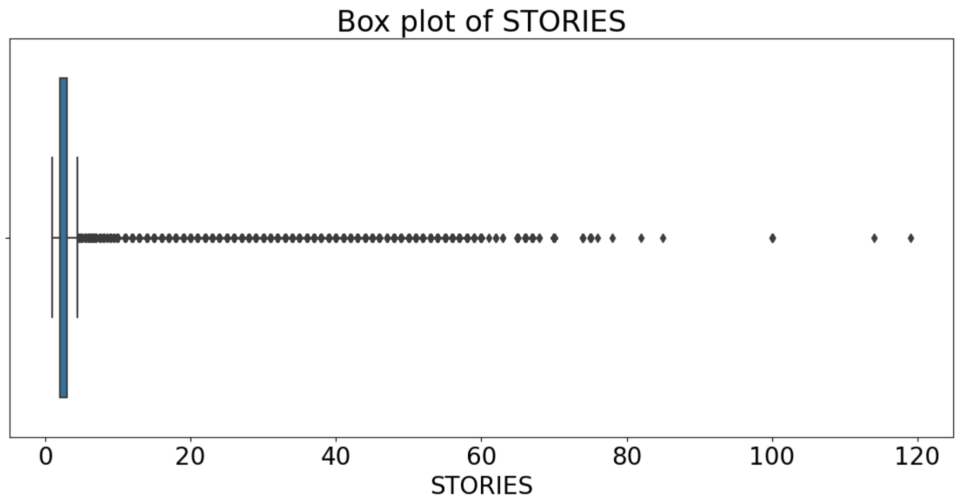
Description : Extension Indicator. This categorical field has 3 unique values with the value “G” being the most common.

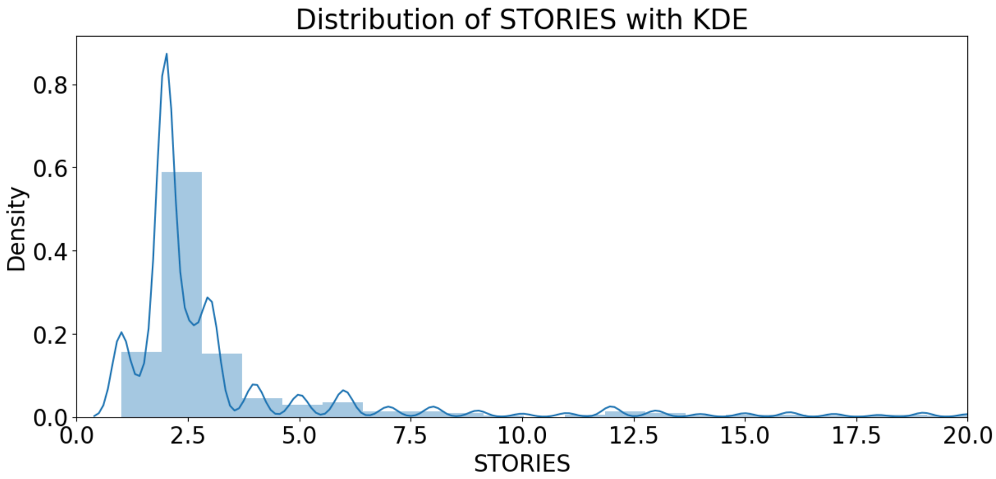


m. Field Name : **STORIES**

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.

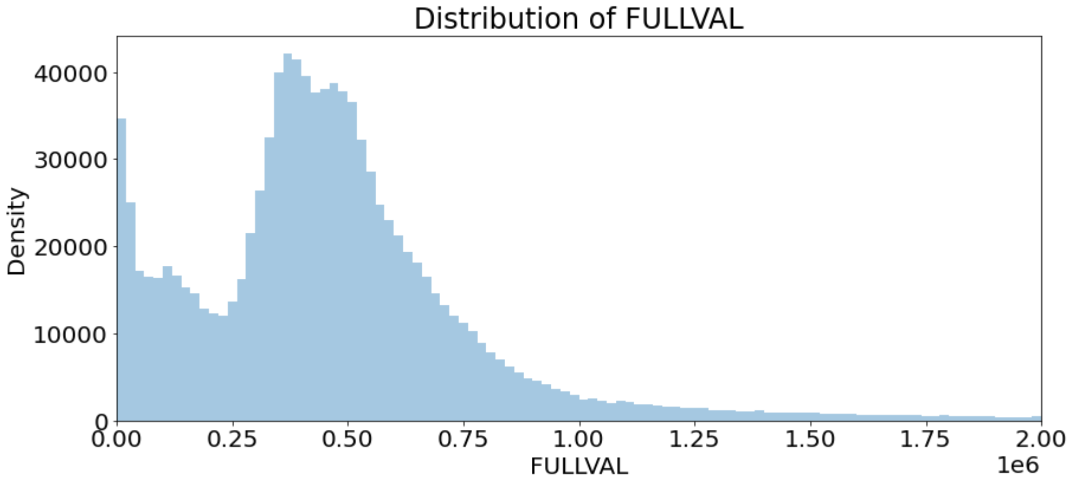


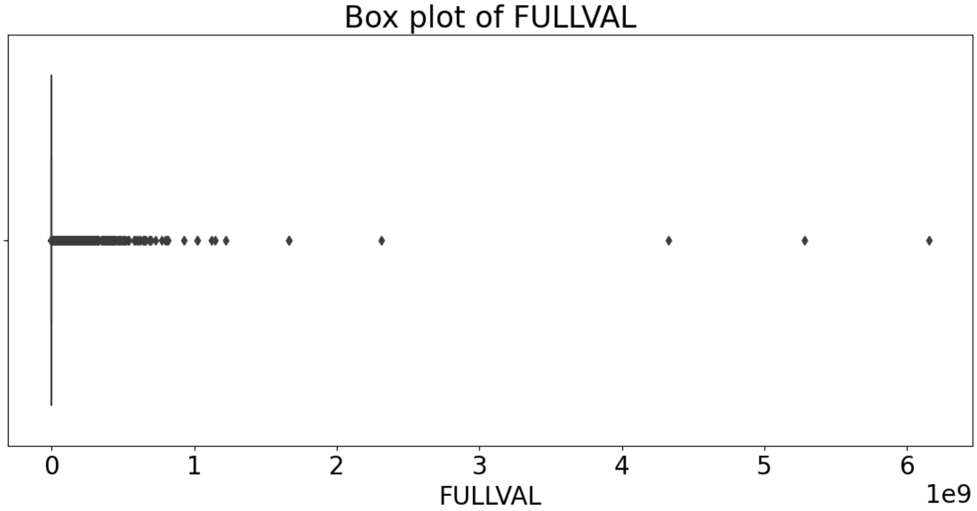


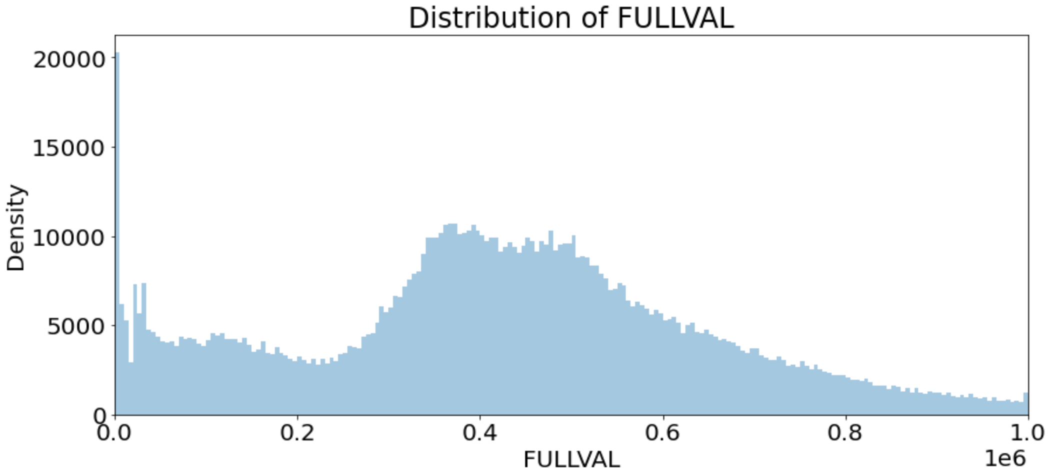


n. Field Name : **FULLVAL**

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.

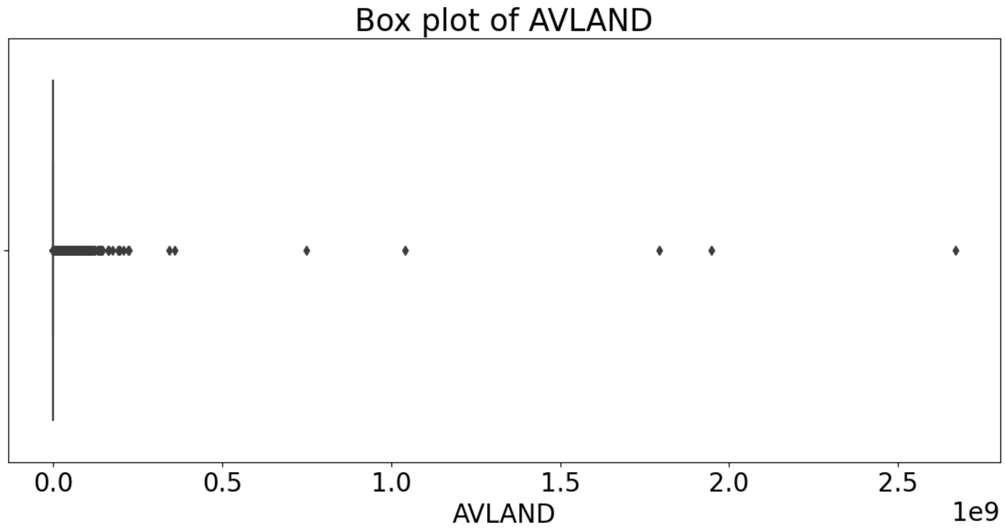


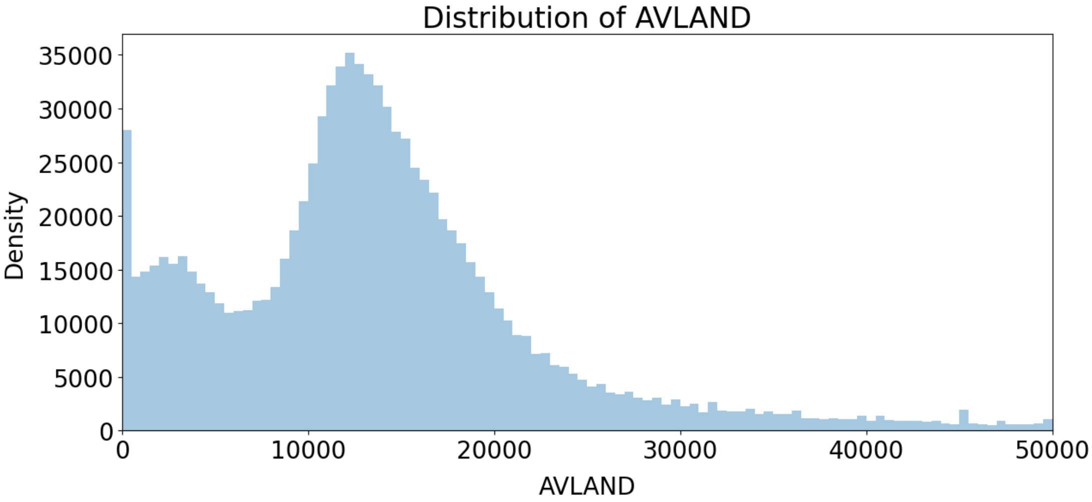




o. Field Name : **AVLAND**

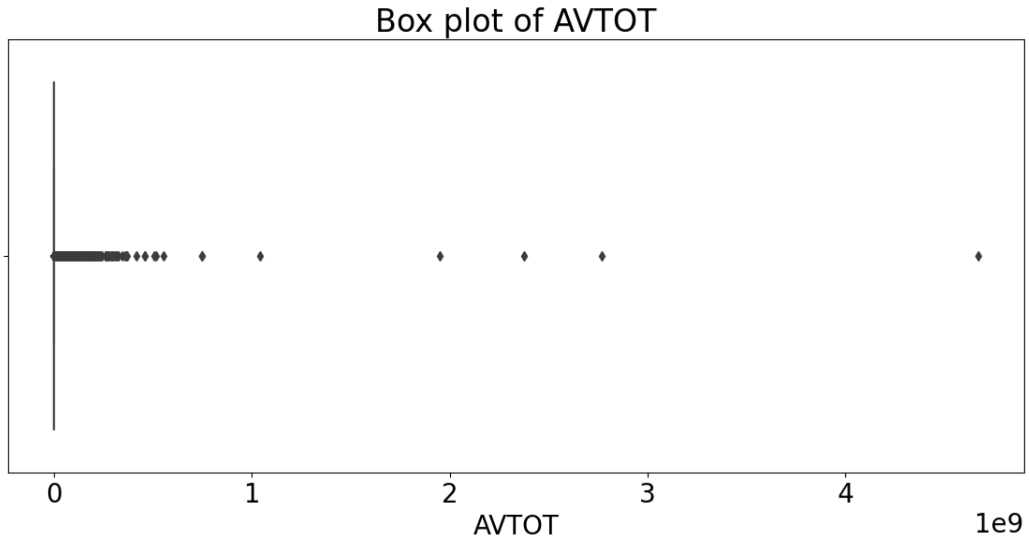
Description : Actual Land Value. This numerical field with values ranging from 0 to 2.6x10^9 with a mean value of 85067.92 and mode 0. This field has 70921 distinct values. Values above 4x10^8 are outliers.

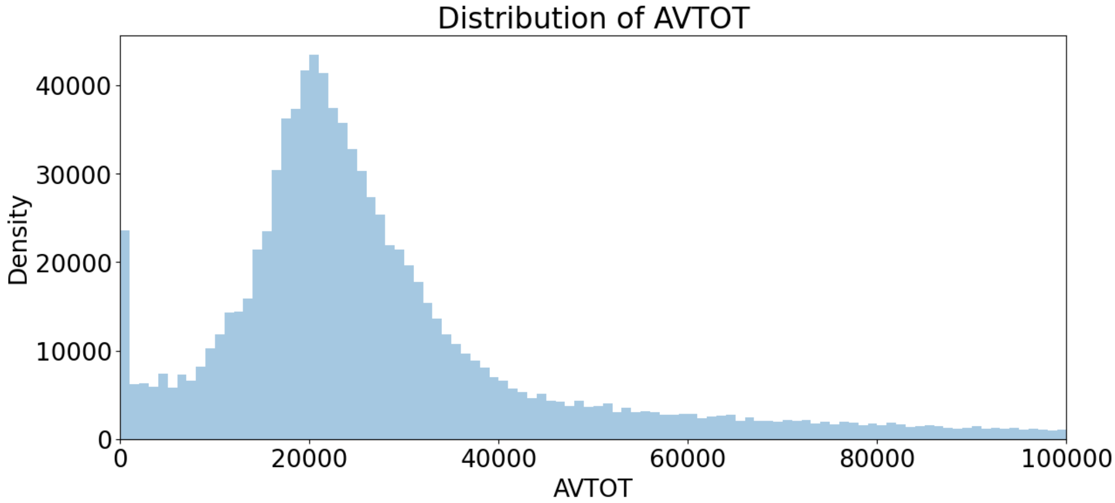




p. Field Name : **AVTOT**

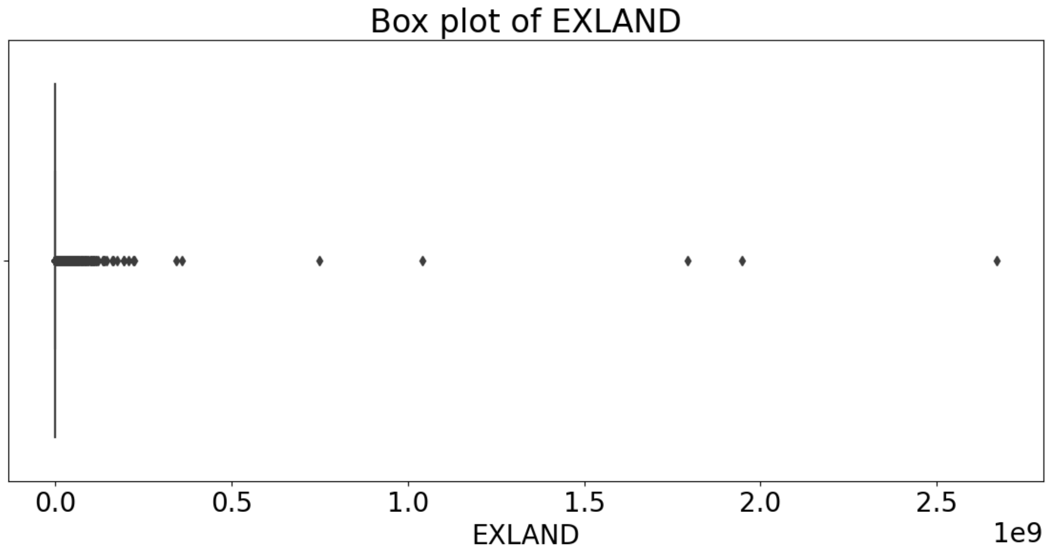
Description : Actual Total Value. This numerical field has values ranging from 0 to 4.6x10^9 with a mean of 227238.17 and mode 0. The boxplot shows that 1x10^9 are outliers. This field has 112914 unique values.

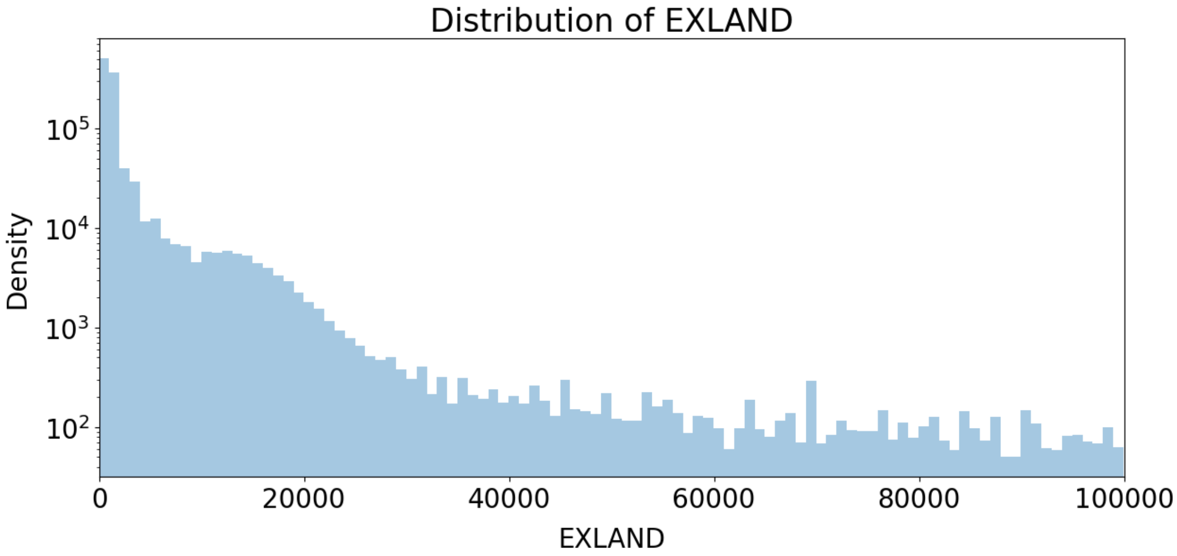




q. Field Name : **EXLAND**

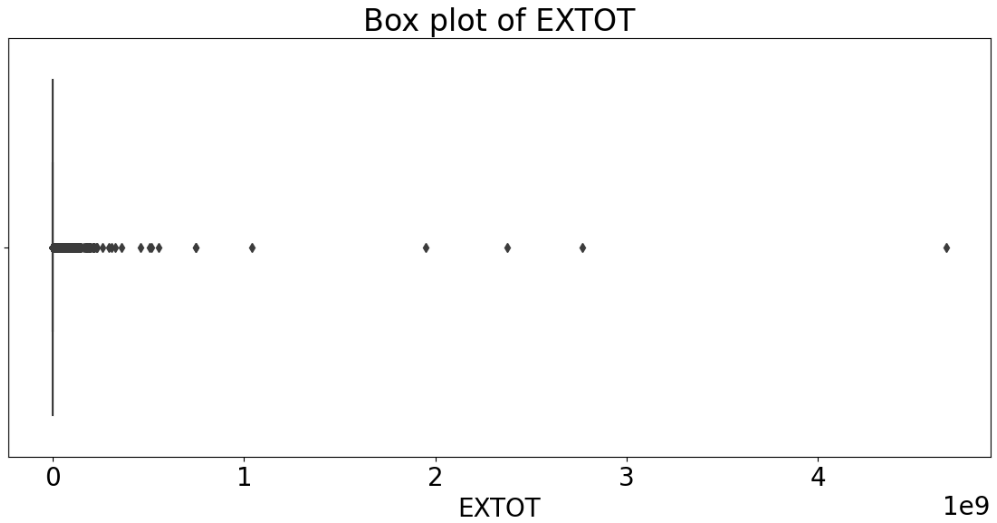
Description : Actual Exempt Land Value. This numerical field has values ranging from 0 to 2.6x10^9 with a mean value of 36423.89 and mode 0. This field has 33419 unique values. Values above 5x10^8 are outliers.

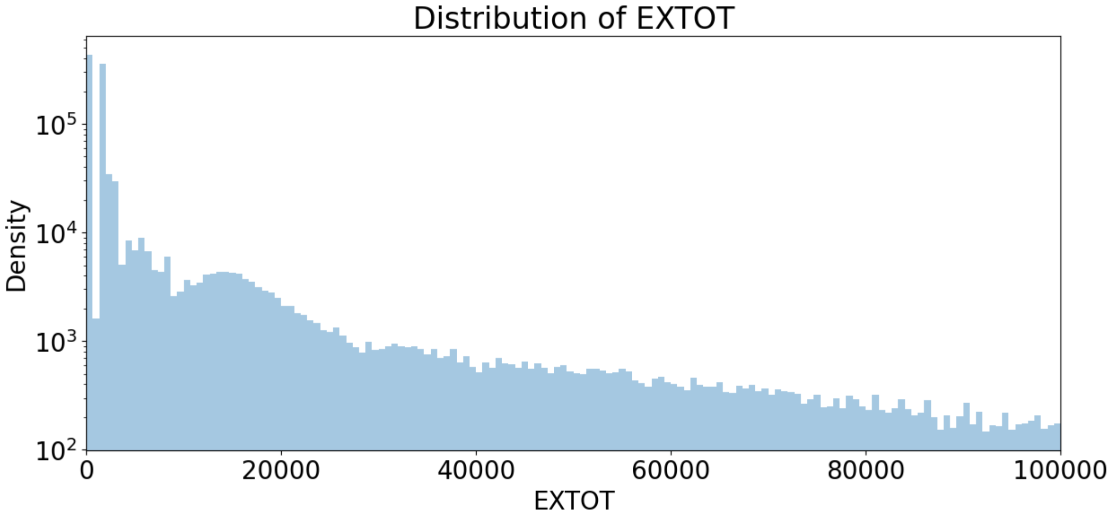




r. Field Name : **EXTOT**

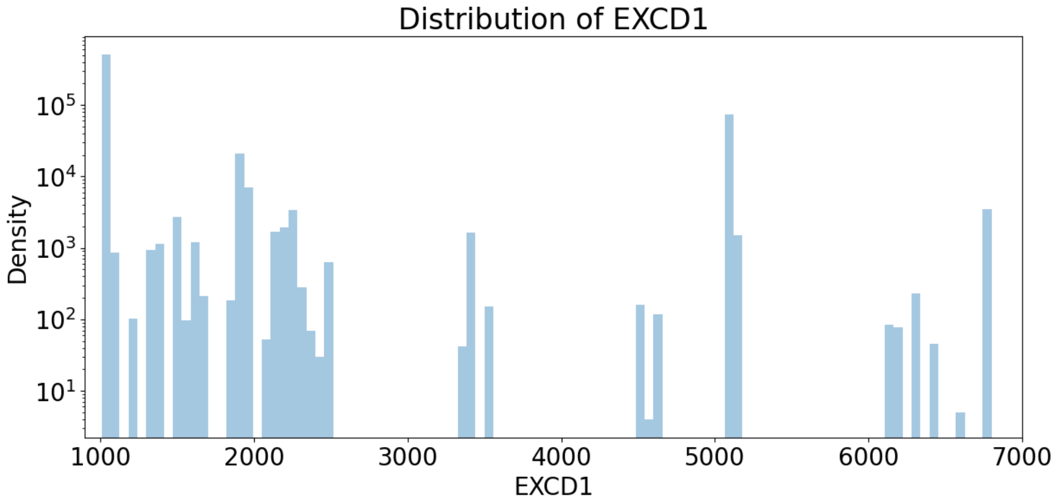
Description : Actual Exempt Land Total. This numerical field has values ranging from 0 to 4.6x109 with a mean of 91186.98 and mode of 0. This field has 64255 unique values. Values above 1x10^9 are outliers.

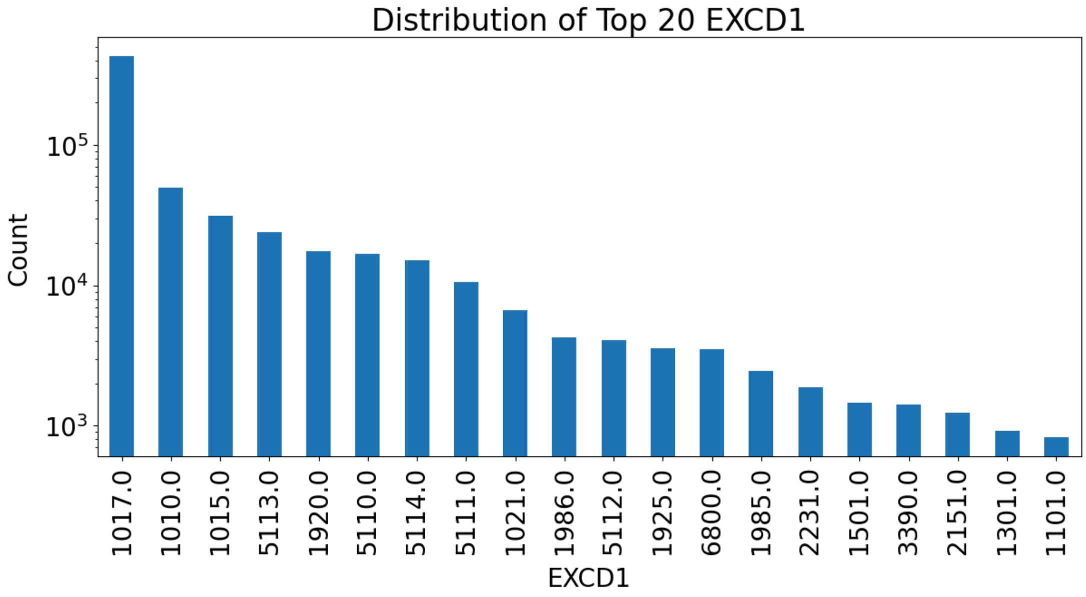


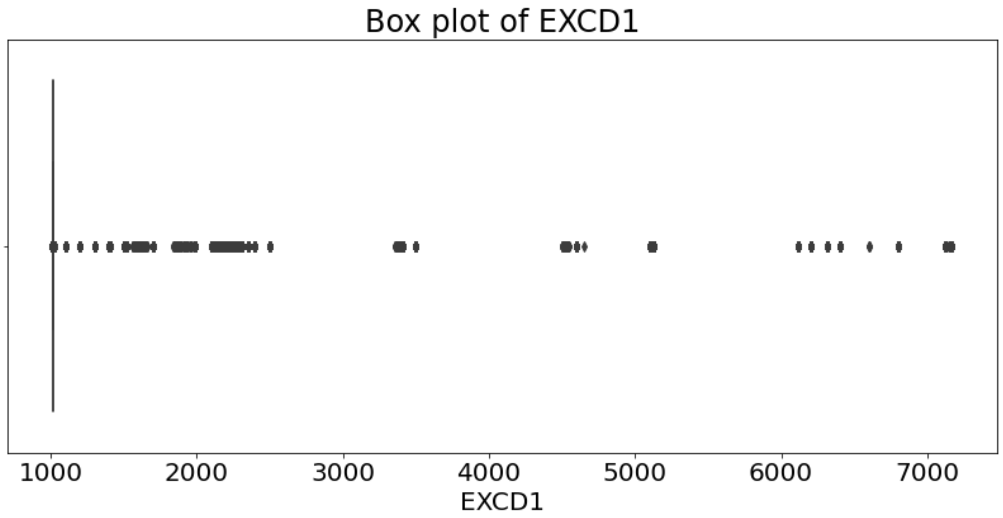


s. Field Name : **EXCD1**

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.

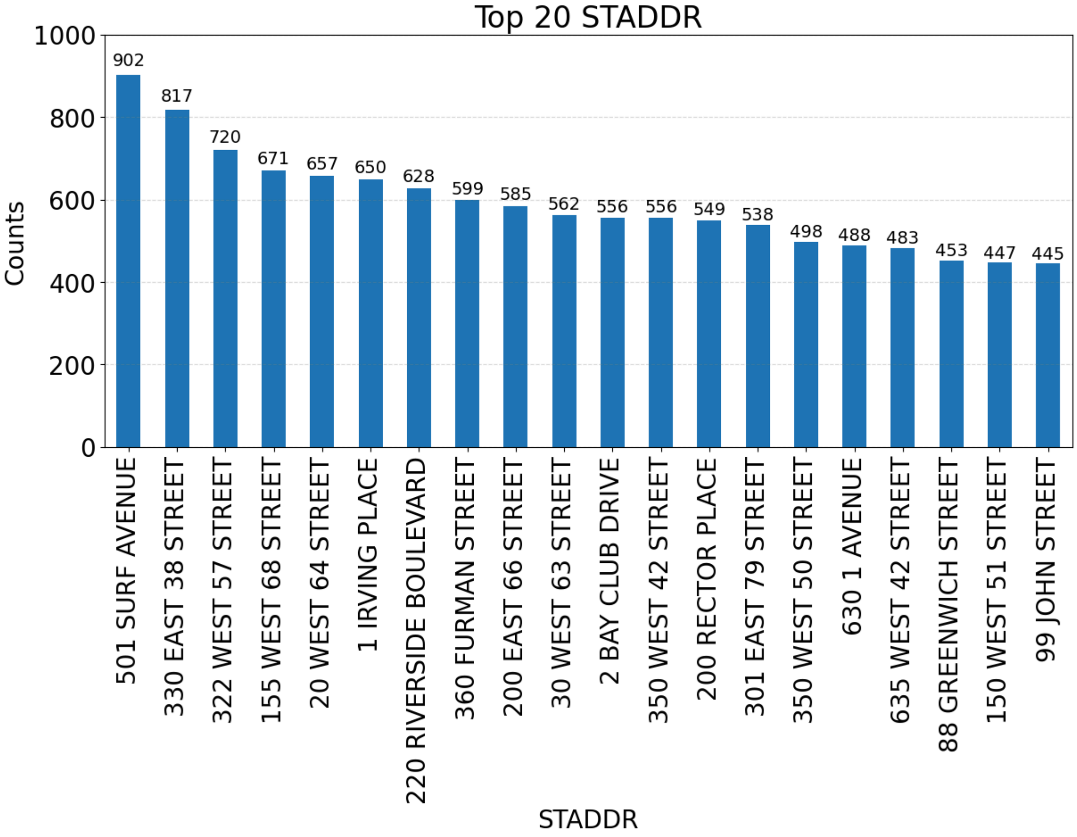

Distribution of EXCD1
Density
10^5
10^4
10^3
10^2
10^1
1000
2000
3000
4000
5000
6000
7000
EXCD1

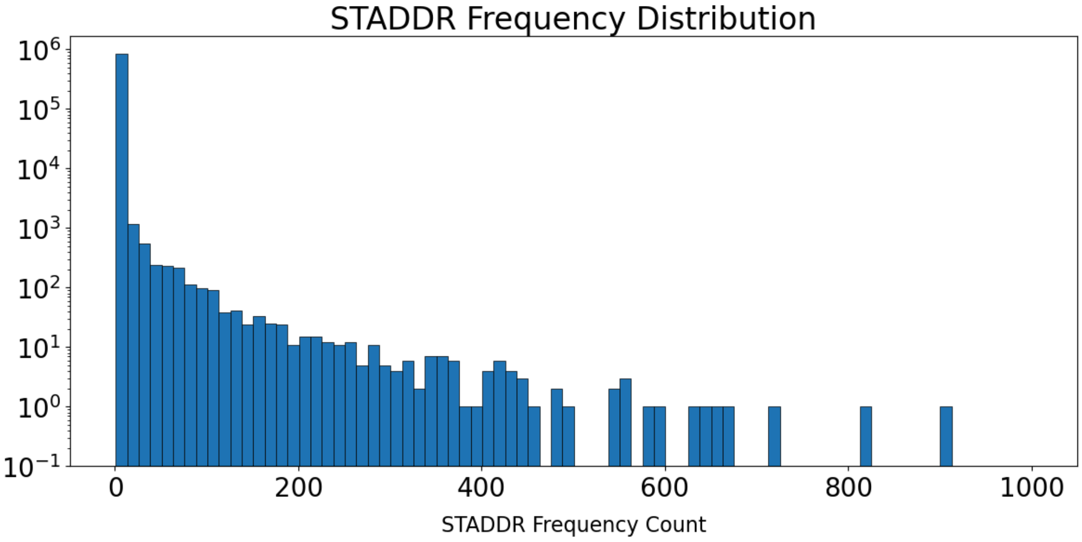

Distribution of Top 20 EXCD1
Count
10^5
10^4
10^3
1017.0
1010.0
1015.0
5113.0
1920.0
5110.0
5114.0
5111.0
1021.0
1986.0
5112.0
1925.0
6800.0
1985.0
2231.0
1501.0
3390.0
2151.0
1301.0
1101.0
EXCD1


Box plot of EXCD1
1000
2000
3000
4000
5000
6000
7000
EXCD1

t. Field Name : **STADDR**

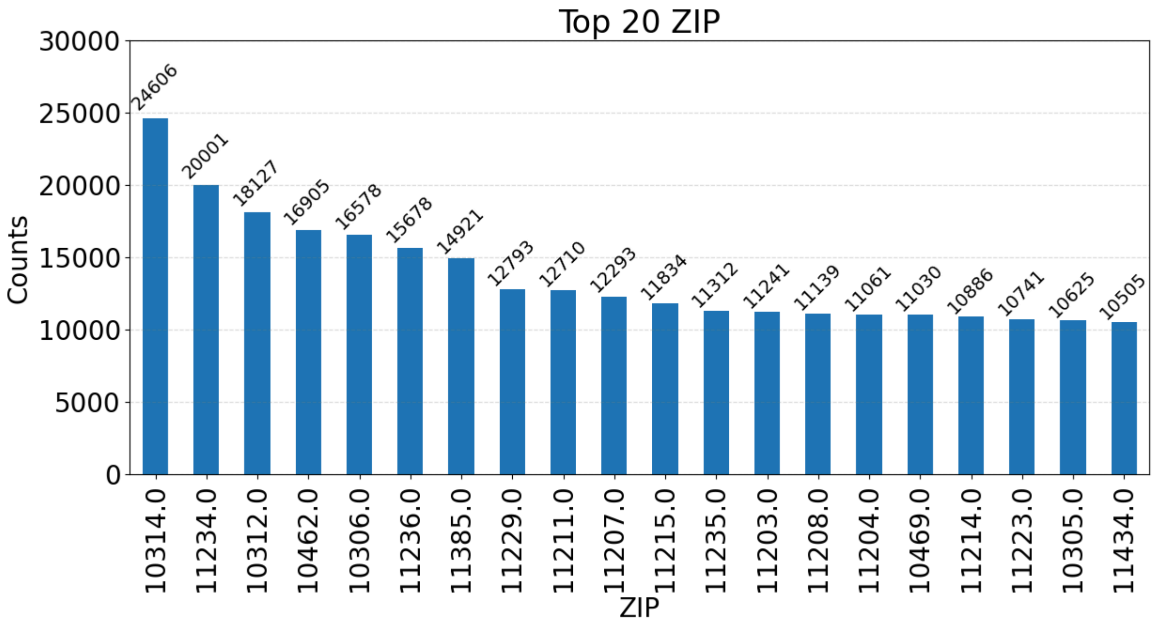
Description : Street Address. This categorical field has 839280 unique values with “501 SURF AVENUE” being the most common one, occurring about 900 times. Studying the frequency distribution plot we can see that most addresses occur only 1-50 times, but a few of them occur about 400-900 times in the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





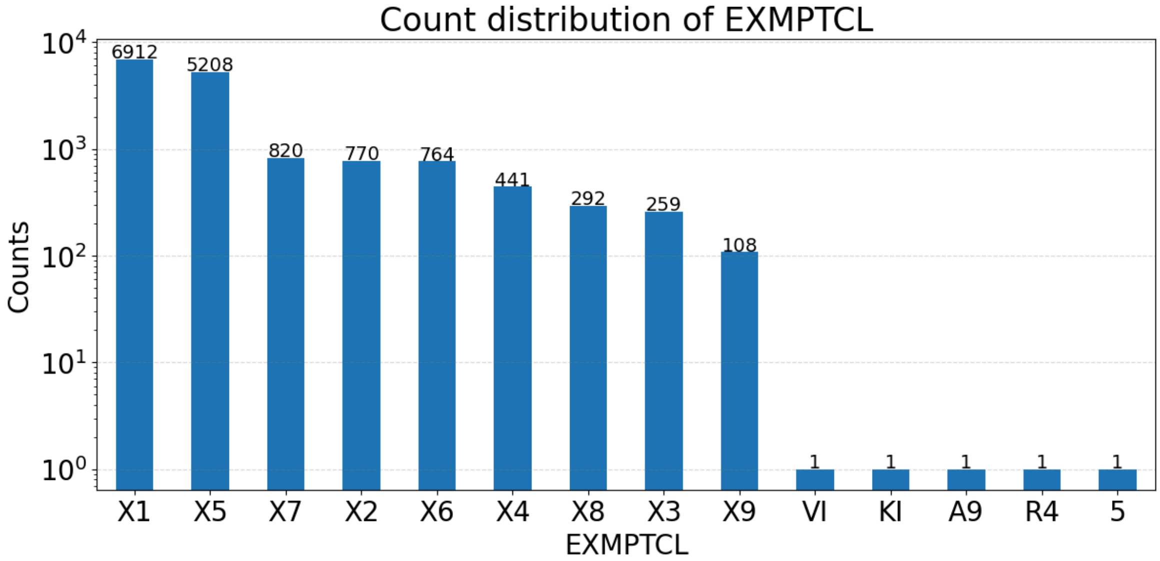
u. Field Name : **ZIP**

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



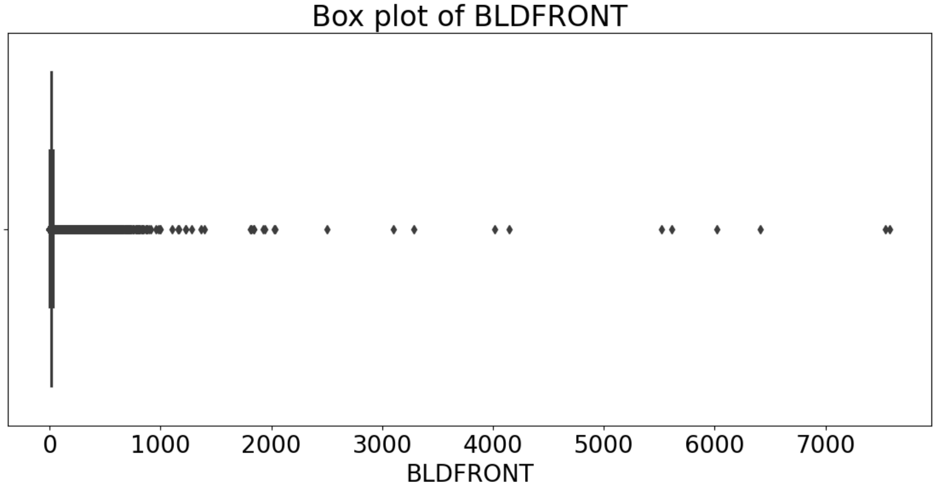
v. Field Name : **EXMPTCL**

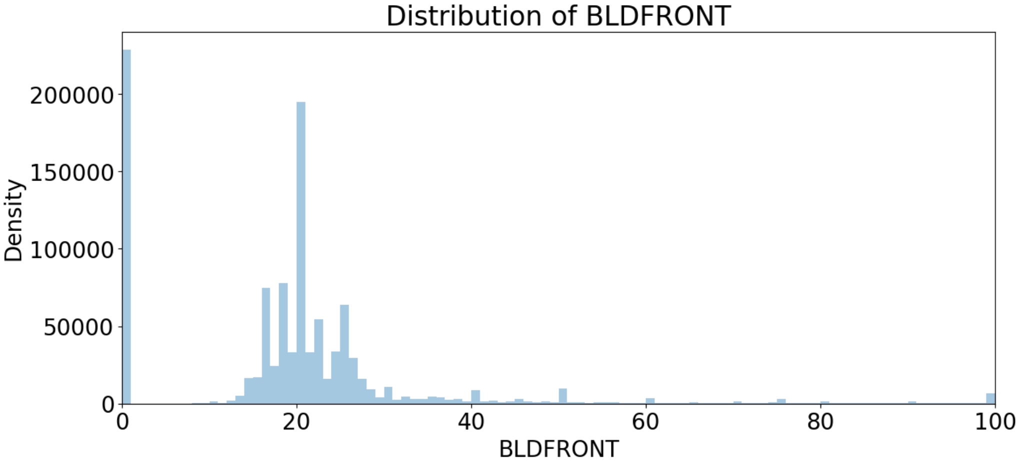
Description : Exemption Class. This categorical variable has 14 unique values with the most common value being “X1” occurring 6912 times.

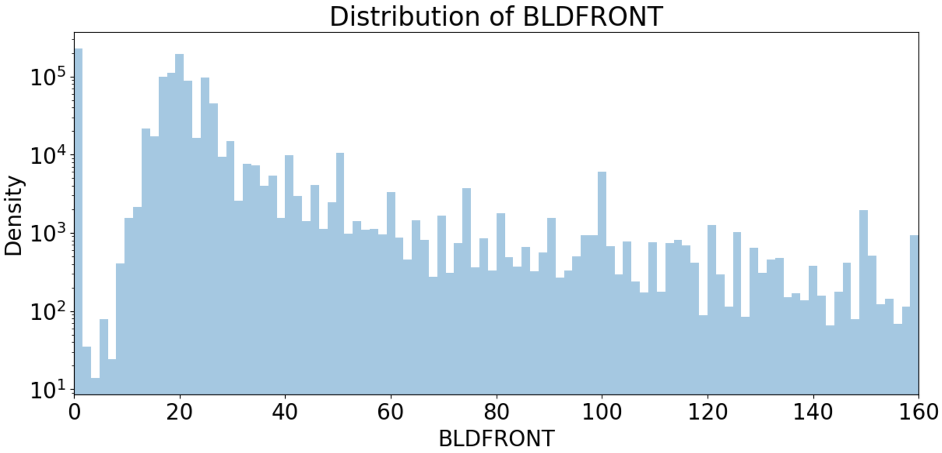


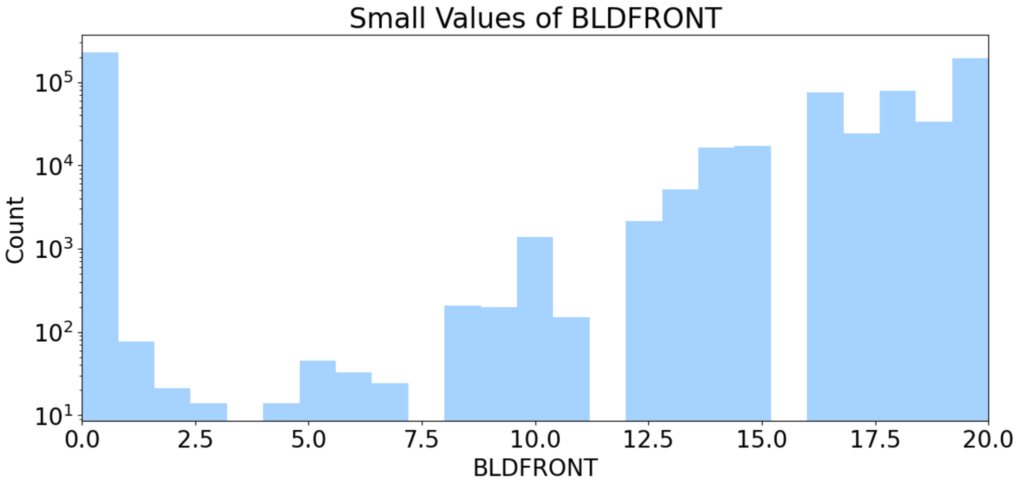
w. Field Name : **BLDFRONT**

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.



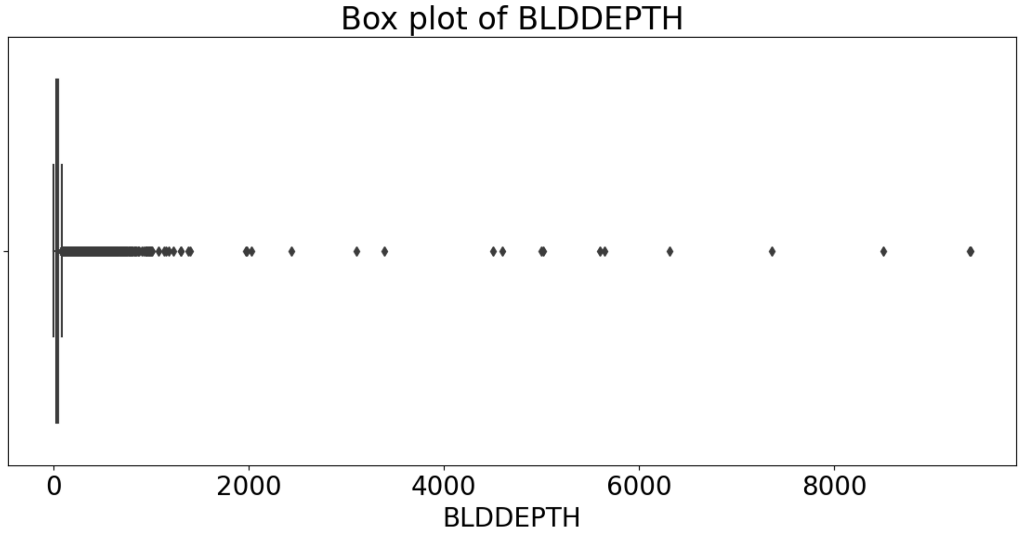


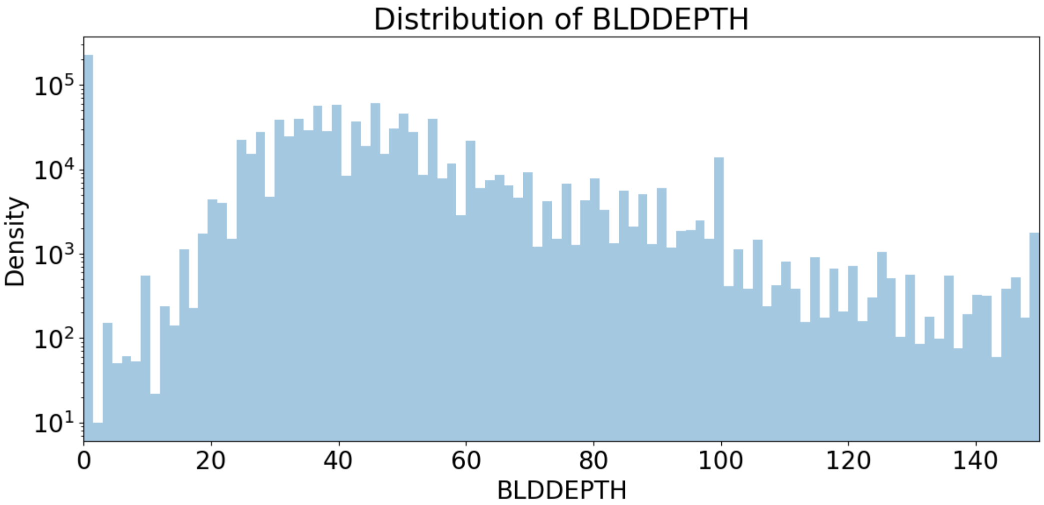


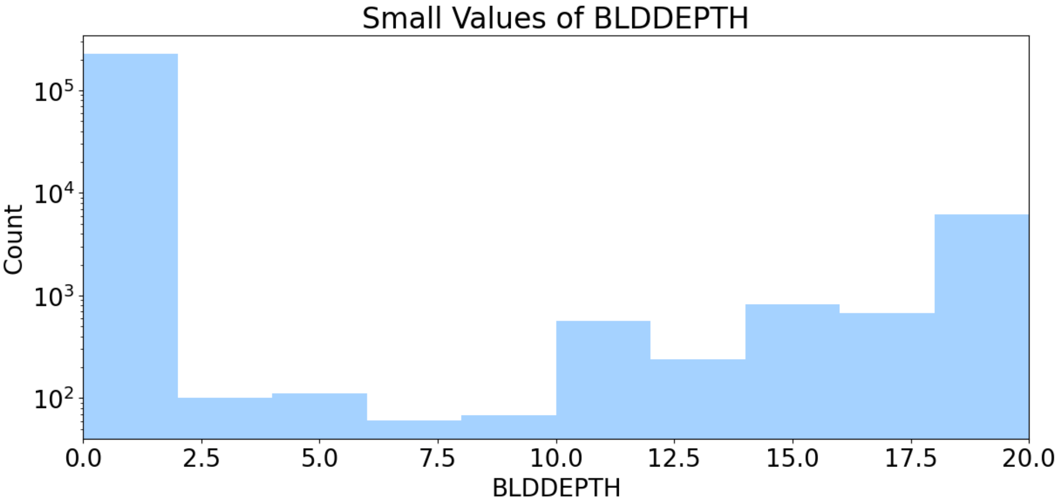


x. Field Name : **BLDDEPTH**

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.

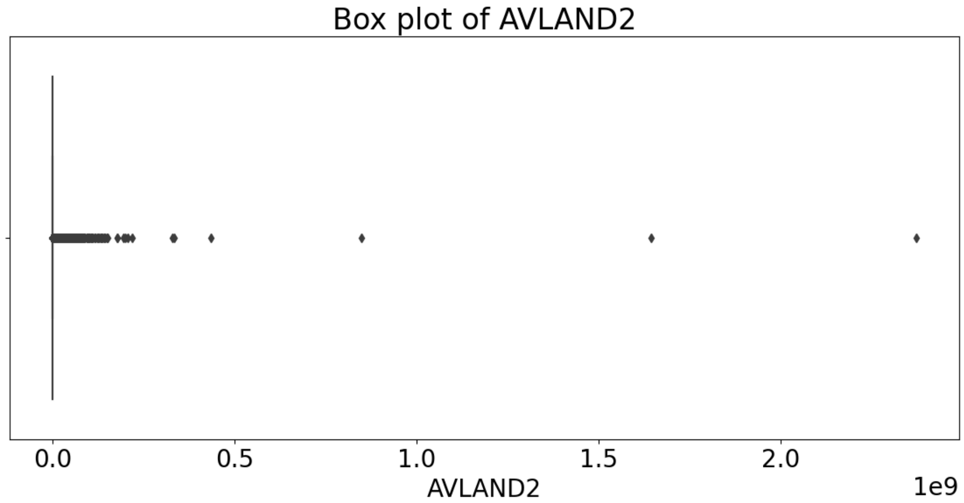


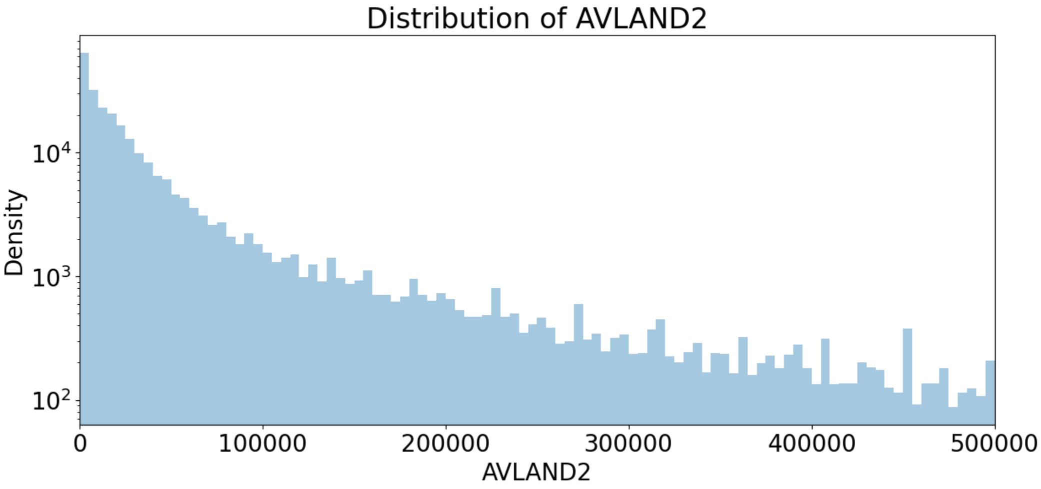




y. Field Name : **AVLAND2**

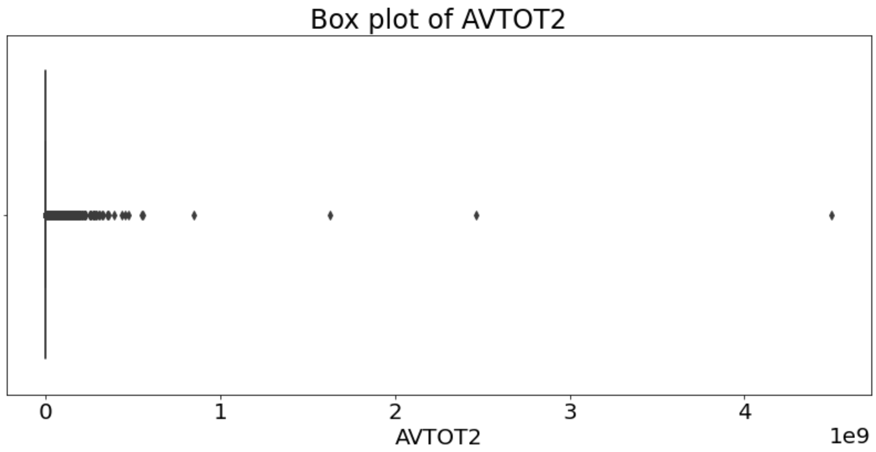
Description : Transitional Land Value. This numerical field value ranges from 3 to 2.3x10^9 with a mean value of 246235.72 and mode 2408. This field has 58592 distinct values. The boxplot below shows values above 5x108 are outliers.

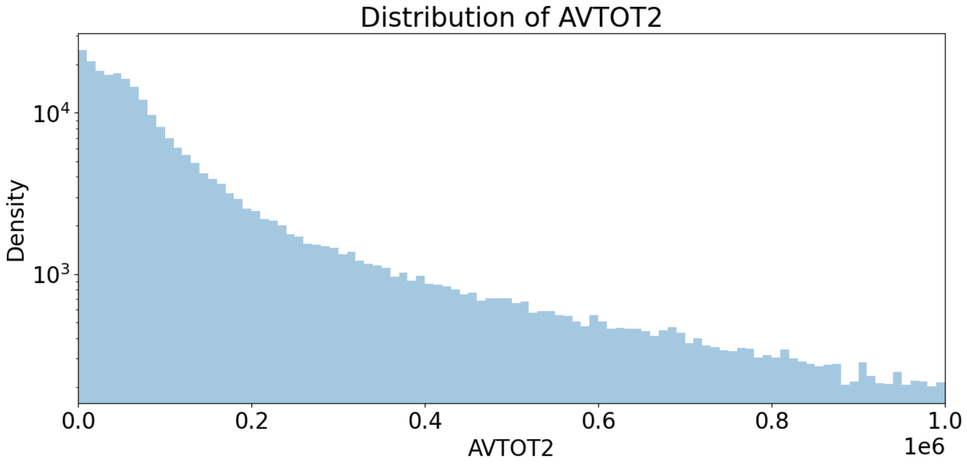




z. Field Name : **AVTOT2**

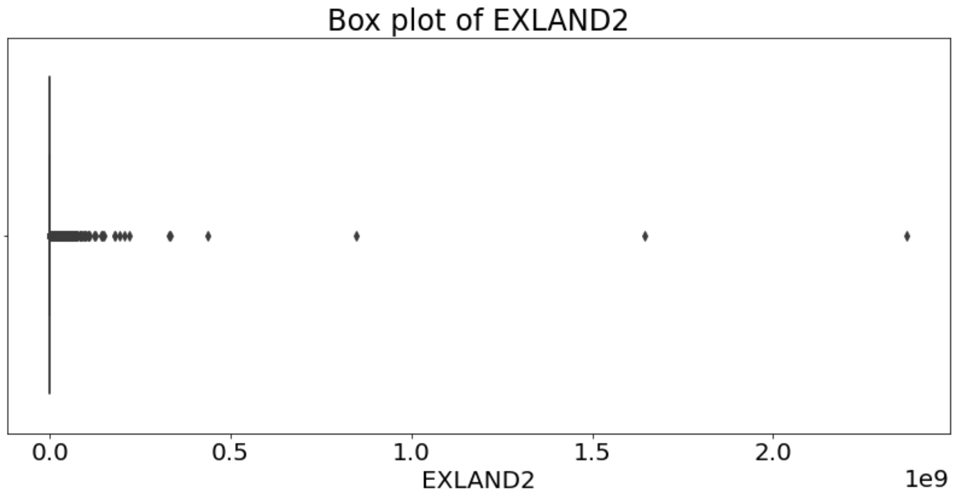
Description : Transitional Total Value. This numerical field value ranges from 3 to 4.5x10^9 with a mean value 713911.44 and mode 750.

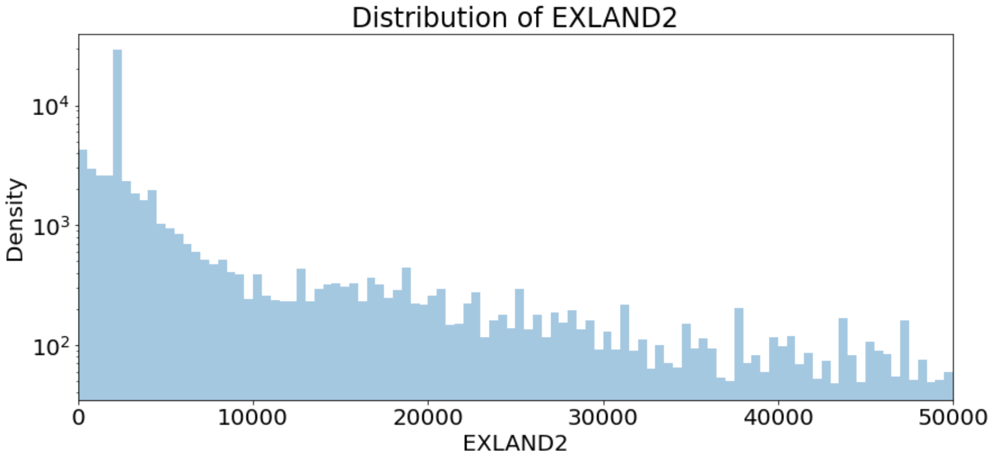




aa. Field Name : **EXLAND2**

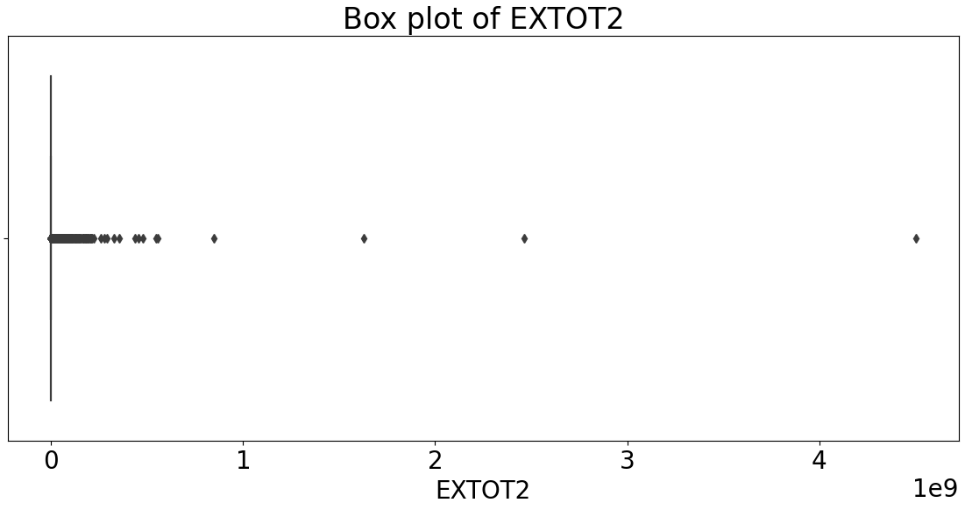
Description : Transitional Exemption Land Value. This numerical field value ranges from 1 to 2.3x10^9 with a mean value of 351235.68 and mode 2090.

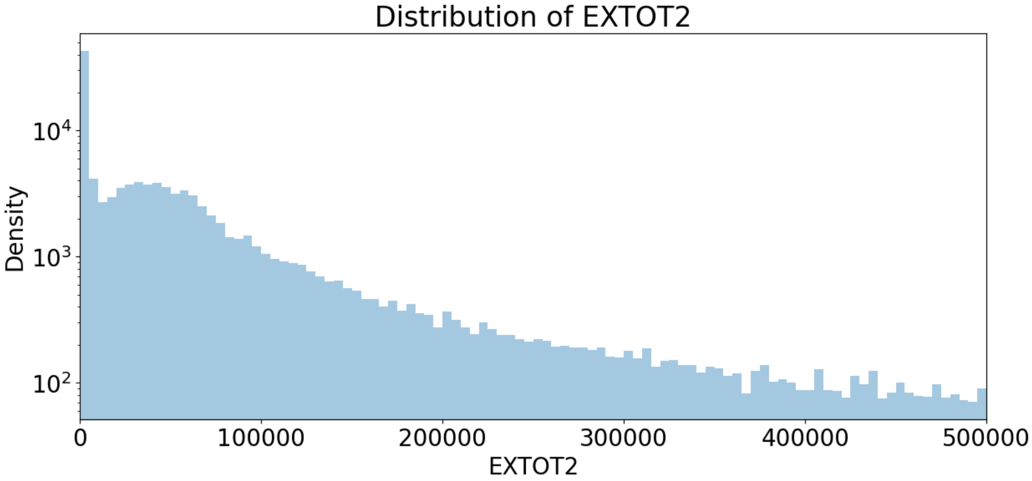




bb. Field Name : **EXTOT2**

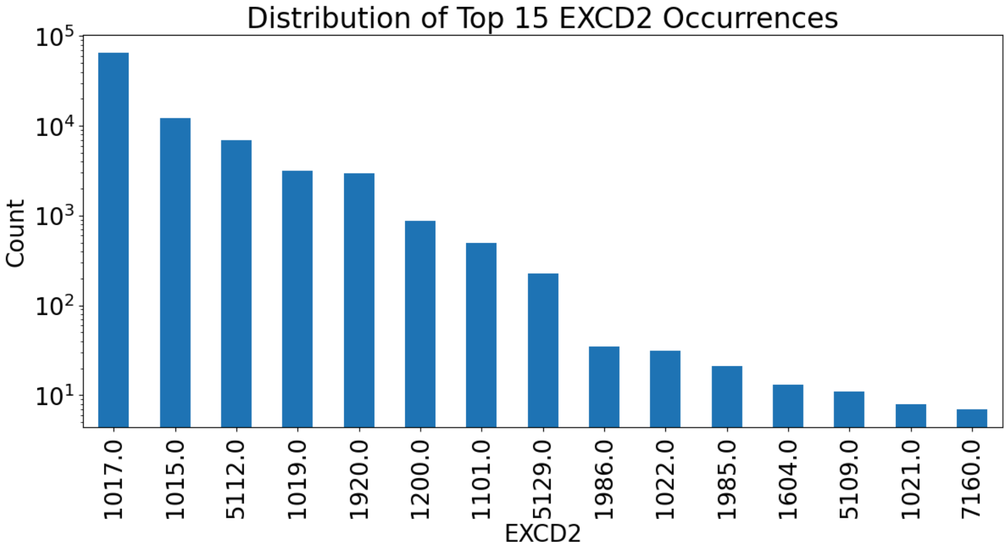
Description : Transitional Exemption Land Total. This numerical field ranges from 7 to 4.5x10^9 with a mean of 656768.28 and mode 2090.

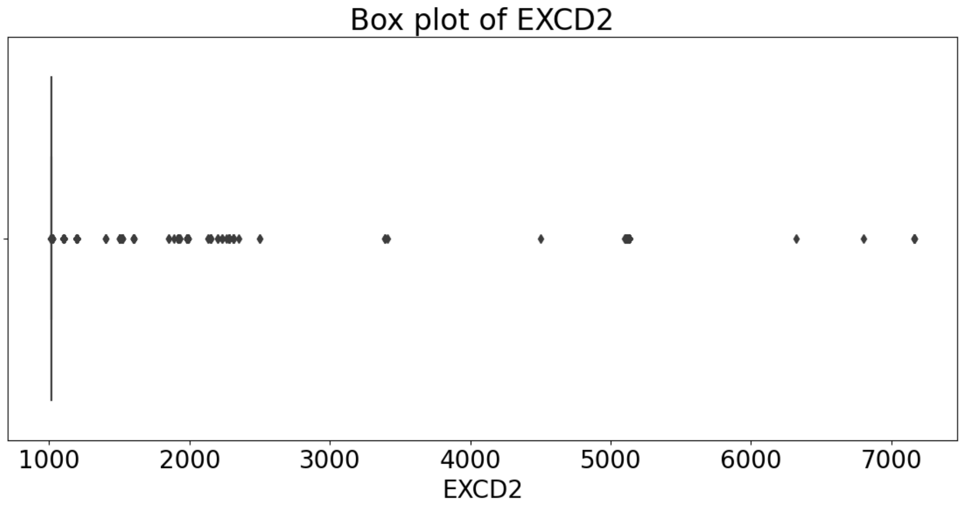


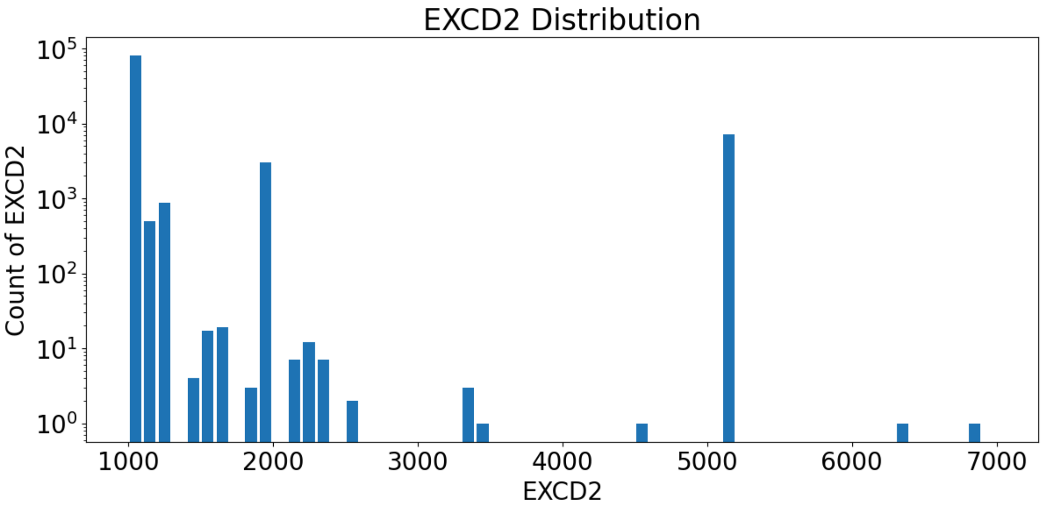


cc. Field Name : **EXCD2**

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


Distribution of Top 15 EXCD2 Occurrences


Box plot of EXCD2


EXCD2 Distribution

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”.