Data Quality Report

1. Data description

The Department of Finance collects Property Valuation and Assessment Data every year to determine the value of properties within the city for the purpose of calculating property tax bills. Each record in this dataset represents a single property within the city and includes information such as the property's location, size, assessed value, and any exemptions or abatements. The dataset includes a collection of 1,070,994 records. There are in total 32 fields, including both numeric and categorical ones. This dataset will be used to identify patterns and trends in these properties to detect tax fraud and to develop a fraud detection algorithm that can accurately identify and prevent it.

2. Summary Tables

a. Numerical Table

Field	% Populated	Min	Max	Mean	Standard Deviation	%Zero
LTFRONT	100%	0	9,999	36.64	74.03	15.79%
LTDEPTH	100%	0	9,999	88.86	76.40	15.89%
STORIES	94.75%	1	119	5	8.37	0%
FULLVAL	100%	0	6,150,000,000	874,265	11,582,431	1.21%
AVLAND	100%	0	2,668,500,000	85,067.92	4,057,260	1.21%
AVTOT	100%	0	4,668,308,947	227,238	6,877,529	1.21%
EXLAND	100%	0	2,668,500,000	36,423.89	3,981,576	45.91%
EXTOT	100%	0	4,668,308,974	91,187	6,508,403	40.39%
BLDFRONT	100%	0	7,575	23.04	35.58	21.36%
BLDDEPTH	100%	0	9,393	39.92	42.71	21.37%
AVALAND2	26.4%	3	2,371,005,000	246,236	6,178,963	0%
AVTOT2	26.4%	3	4,501,180,000	713,911	11,652,530	0%
EXLAND2	8.17%	1	2,371,005,000	351,236	10,802,21	0%
EXTOT2	12.22%	7	4,501,180,000	656,768	16,072,510	0%

b. Categorical table

Field	% Populated	# Blank	# Zeros	# Unique Values	Most Common Value
RECORD	100%	0	0	1,070,994	N/A
BBLE	100%	0	0	1,070,994	N/A
BORO	100%	0	0	5	4
BLOCK	100%	0	0	13,984	3944
LOT	100%	0	0	6,366	1
EASEMENT	0.43%	1,066,358	0	12	E

OWNER	97.04%	31,745	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	100%	0	0	200	R4
TAXCLASS	100%	0	0	11	1
EXT	33.08%	716,689	0	3	G
EXCD1	59.62%	432,506	0	129	1017.0
STADDR	99.94%	676	0	839,280	501 SURF AVENUE
ZIP	97.21%	29,890	0	196	10314.0
EXMPTCL	1.45%	1,055,415	0	14	X1
EXCD2	8.68%	978,046	0	60	1017.0
PERIOD	100%	0	0	1	2010/11
YEAR	100%	0	0	1	FINAL
VALTYPE	100%	0	0	1	AC-TR

3. Data Visualization

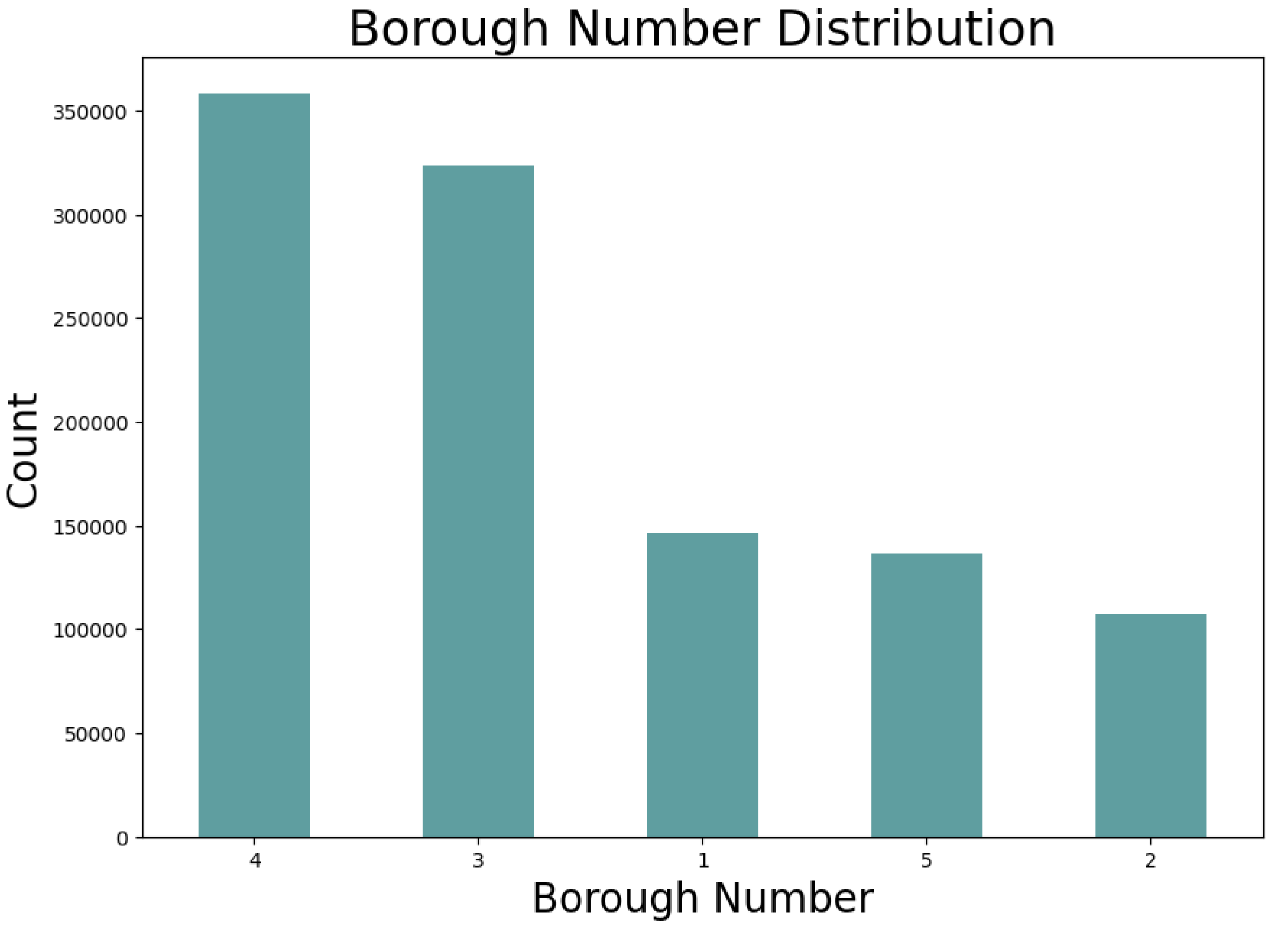
a. **RECORD**
Record numbers assigns a unique number to each property. They are ordinal numbers from 1 to 1,070,994. Each row is one unique positive number.

b. **BBLE**
This field assigns a unique file key number to each property, identifying the location of the property (key BBLE = Boro, Block, Lot, and Easement Code)

c. **BORO**
The borough number refers to a numerical code assigned to a specific borough or district within a city, commonly used to identify the location of a property. In this dataset, the number indicates the following:

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

The most common borough number is 4 (Queens) with more than 350,000 records. This chart shows the 5 borough numbers with its distribution.

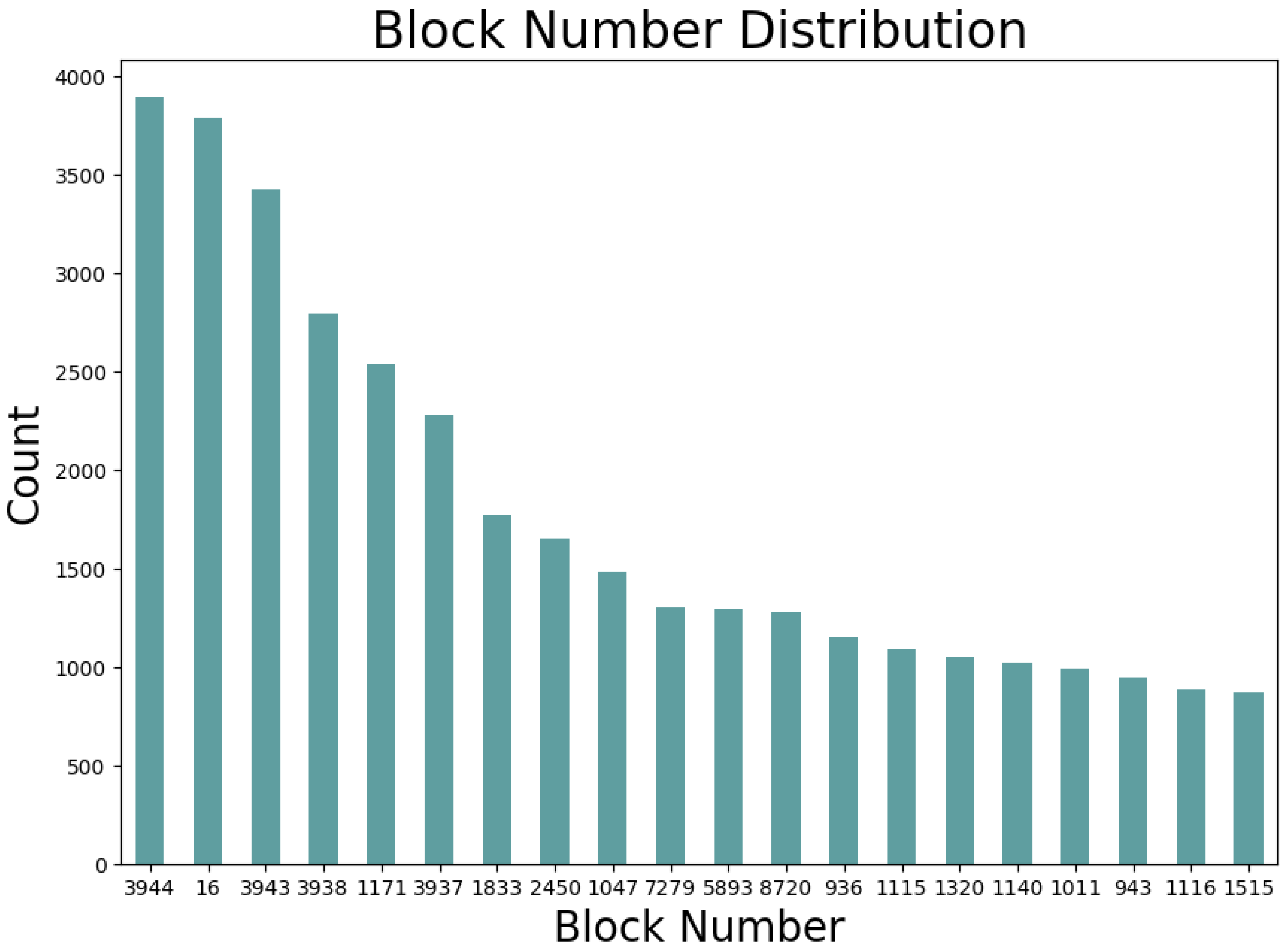


d. **BLOCK**

A block number refers to a group of properties that share a common boundary. By providing valid block ranges by borough numbers, the dataset ensures consistency. In this dataset, block number is assigned as below:

Manhattan: 1 to 2255
Bronx: 2260 to 5958
Brooklyn: 1 to 8955
Queens: 1 to 16350
Staten Island: 1 to 8050

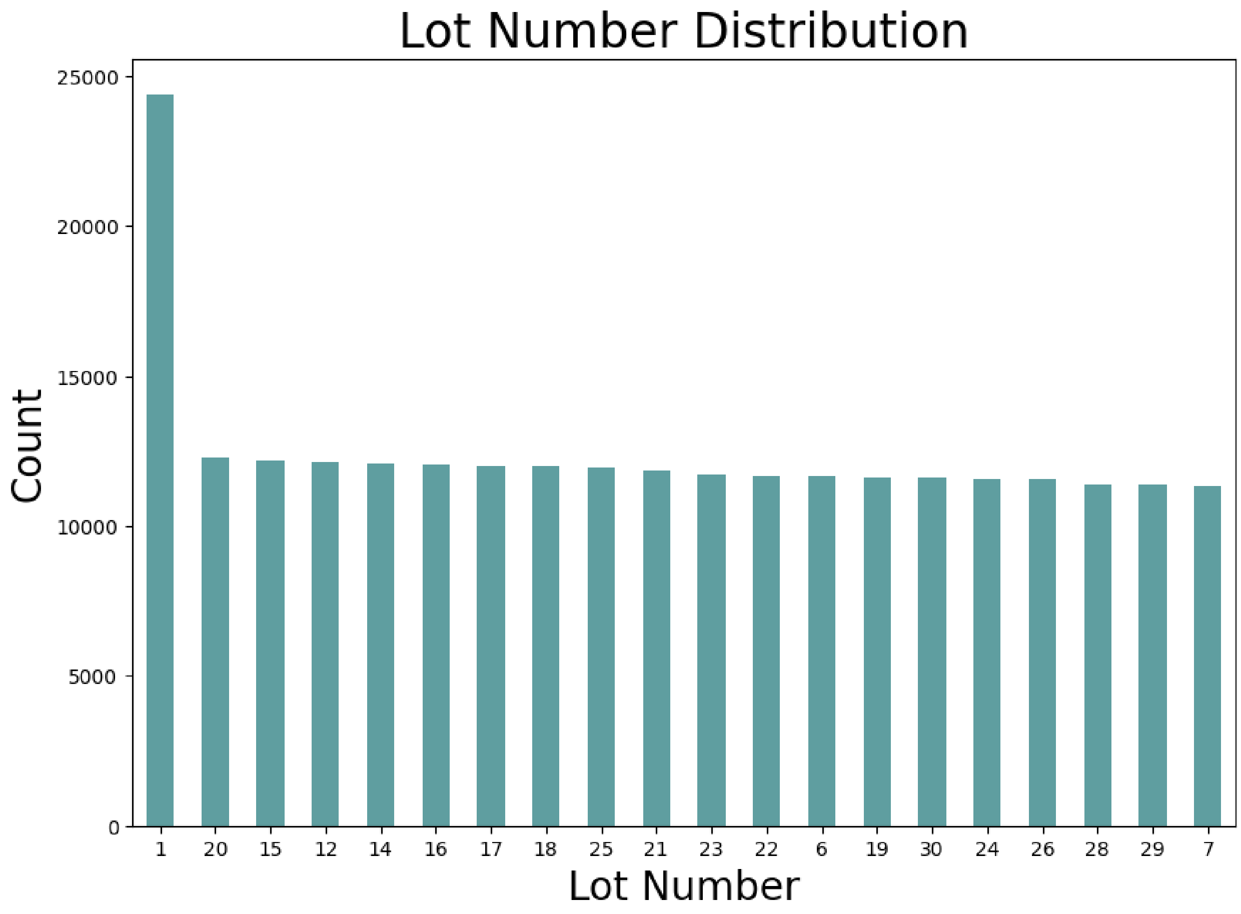
3394 (Bronx) is the most common value, following by 16 (Manhattan). This chart shows the 20 blocks with the highest count, among a total of 13,984 unique block numbers

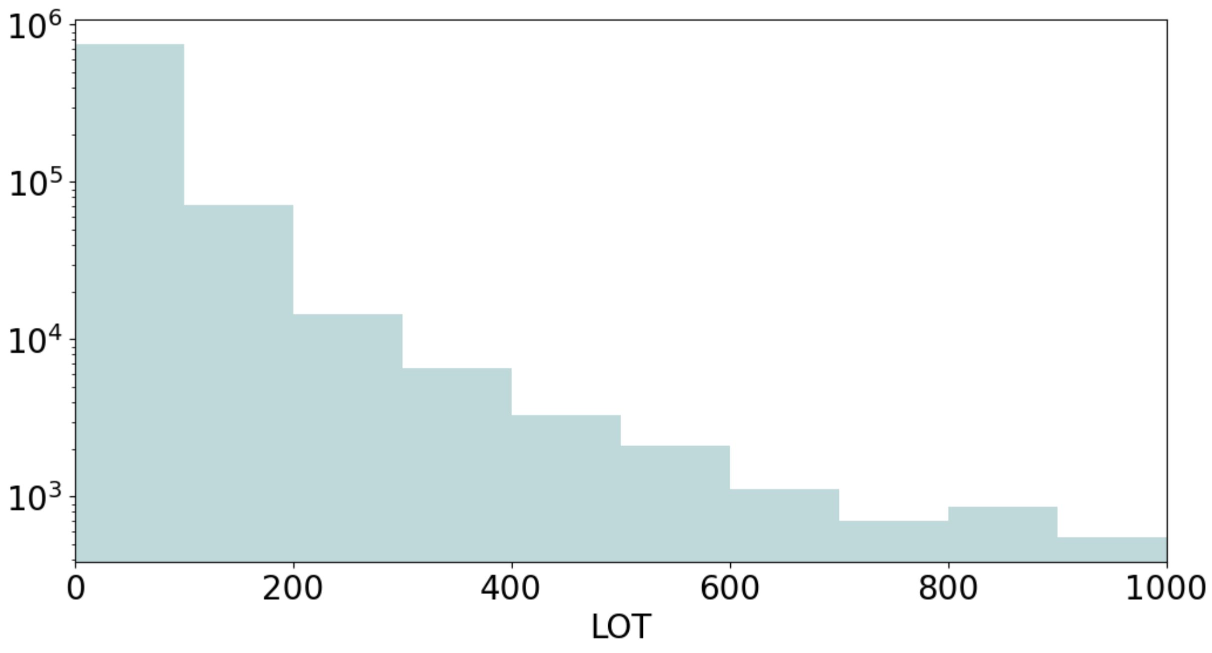


e. **LOT**

Lot number is used in conjunction with block number and borough numbers to create a unique identifier for each property.

The most common value is 1 with almost 25,000 records. The chart shows the 20 most common values of lot numbers, among a total of 6,366 unique ones. Following is the histogram showing the distribution of the first 1000 lots.



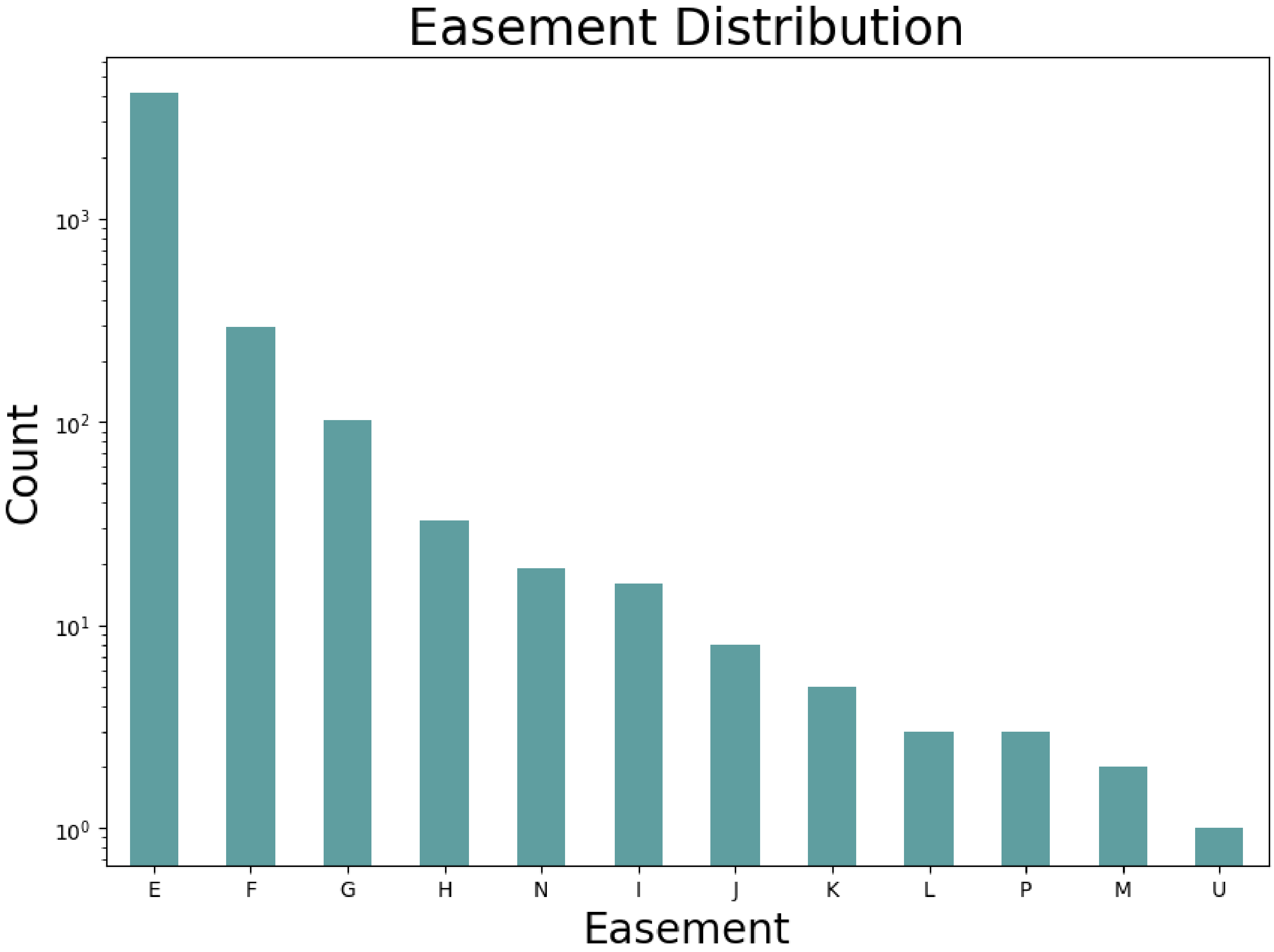


f. **EASEMENT**

An easement is a legal right granted to a person or entity to use or access a portion of another person's property for a specific purpose. In this dataset, easement is understood as following:

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

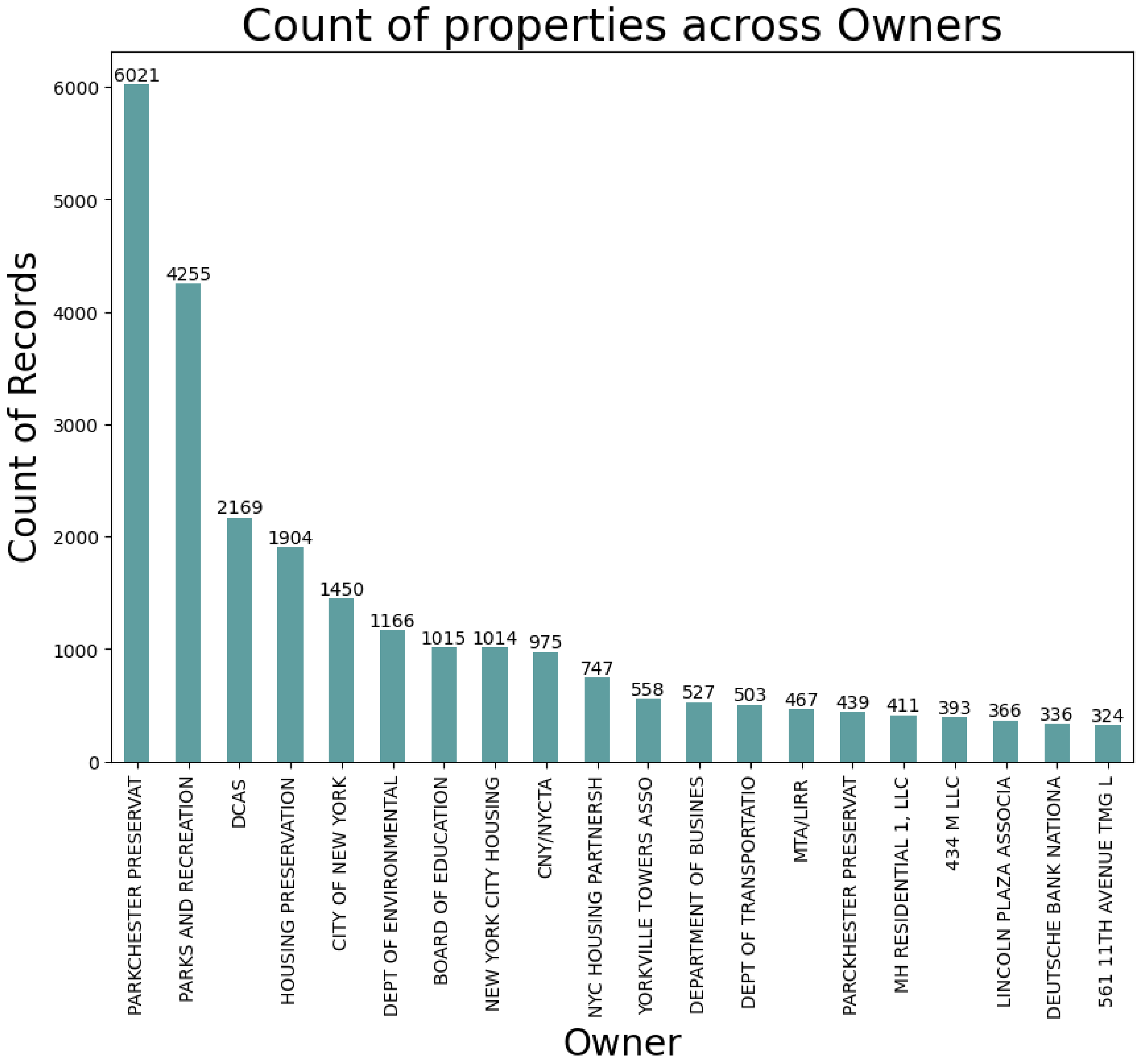
E (Land Easement) is the most common value with the highest count while U has the lowest count (US. Government). This chart shows the distribution of the 12 easements.



g. **OWNER**

This field indicates the owners of the properties. It could be used in combination with other fields to detect potential tax fraud. The most common value is PARKCHESTER PRESERVAT with around 6,000 properties

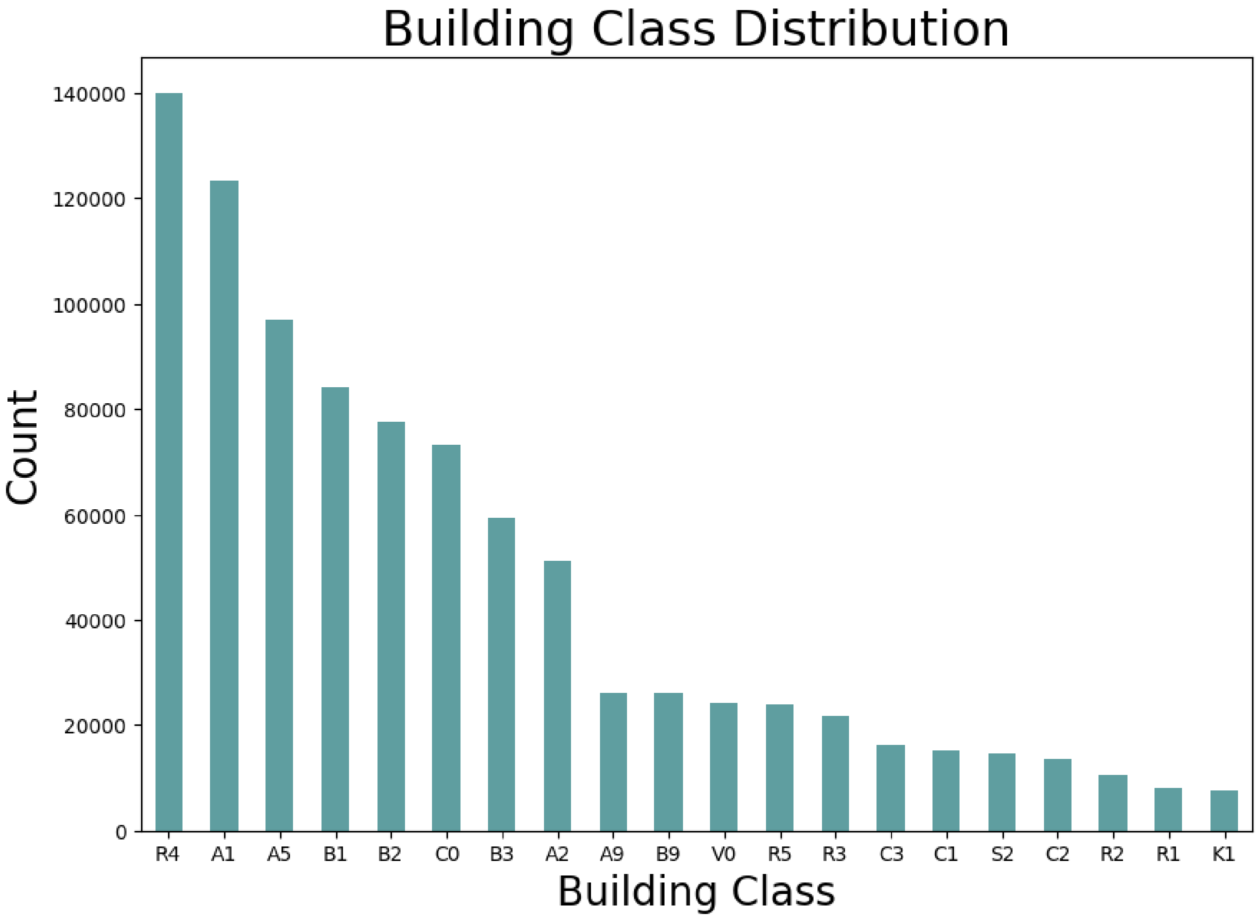
The chart shows the 20 most common values of property owners with the count of properties, among a total of 883,348 unique ones.



h. **BLDGCL**

This field indicates the building class, a classification system to categorize property based on their physical and functional characteristics. The most common value is R4 with around 14,000 properties

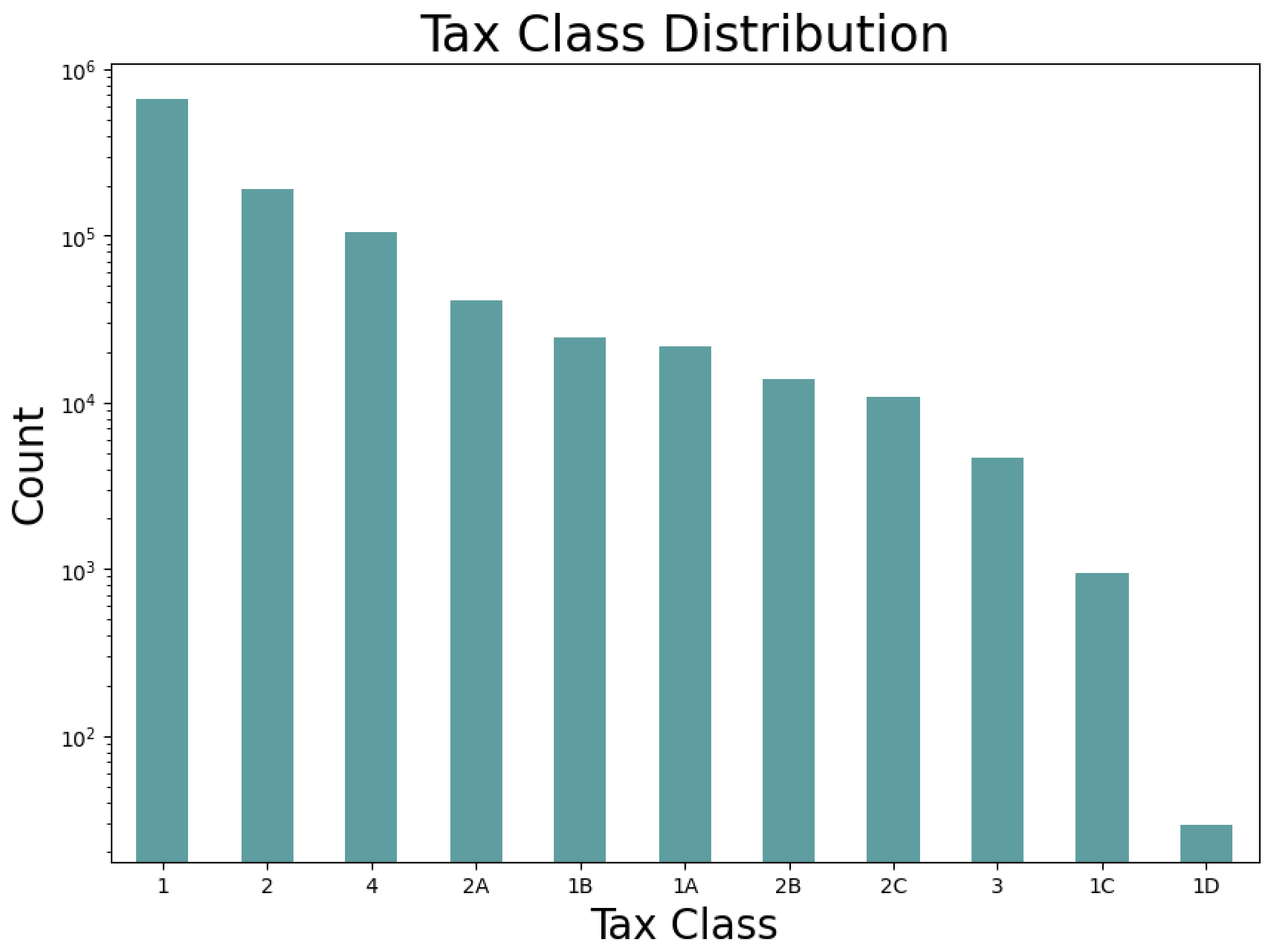
The chart shows the 20 most common values of building classes, among a total of 200 unique ones.



i. **TAXCLASS**
This field refers to a classification system used by local governments to determine the tax rate and tax liability for each property. In this dataset, tax class is assigned as below:

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

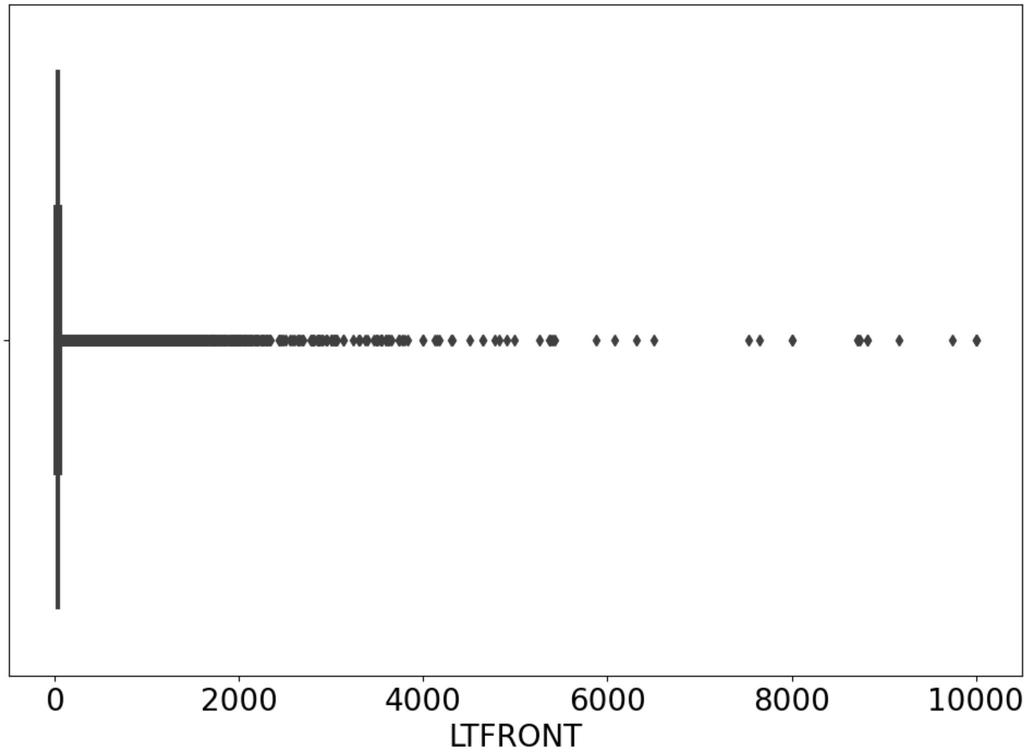
This chart shows the distribution of 11 tax classes, with 1 (1-3 Unit Residence) being the most common value.

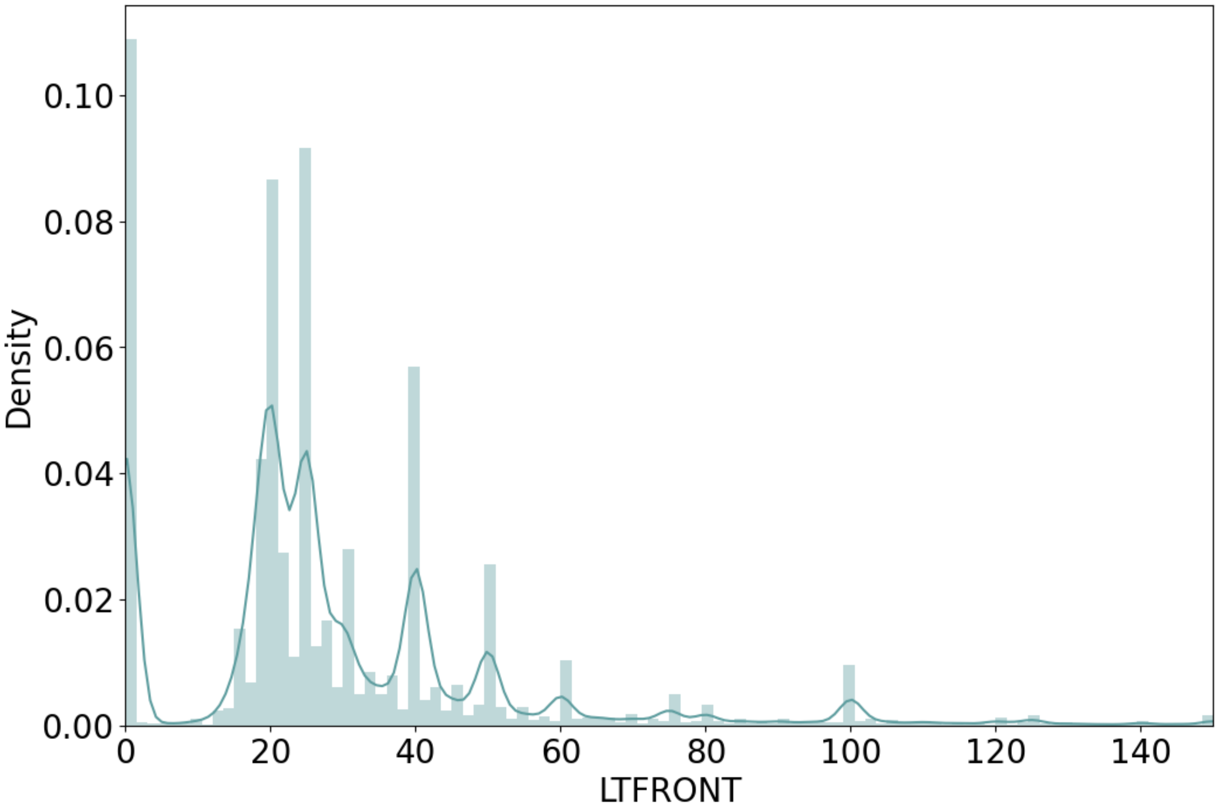


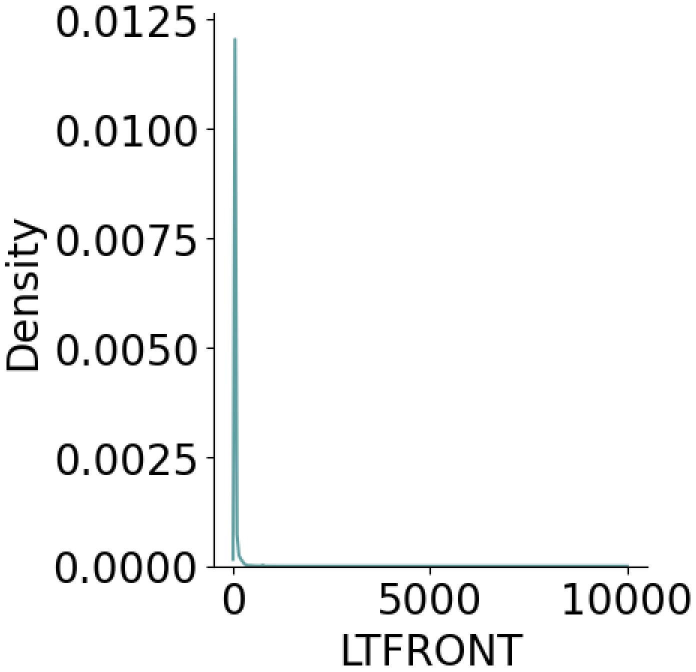
j. **LTFRONT**

This field refers to lot width, the distance between the side property lines of land. The minimum width is 0 and the maximum width is 9,999, with a mean of 36.64, and standard deviation of 74.03. Below are the plots that show the distribution of lot widths across this dataset.

According to the density graph, lot widths at 0 and the range between 20 and 30 are the densest, especially around 25. After that peak, the density of lot widths goes down as the lot width get bigger.


0
2000
4000
6000
8000
10000
LTFRONT

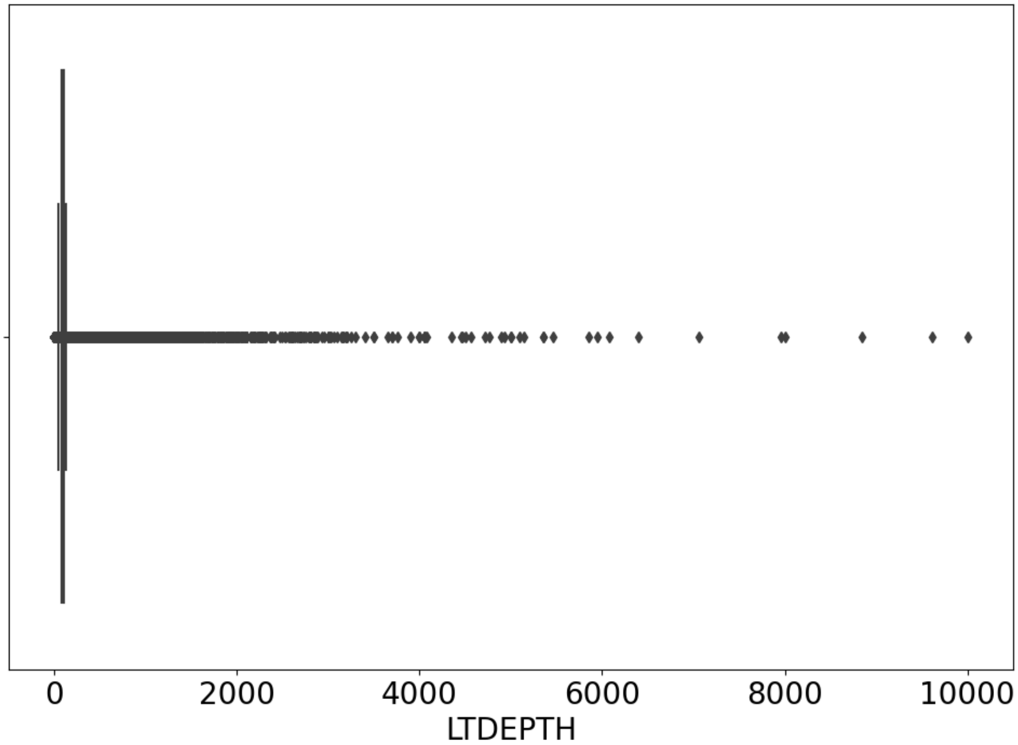

0.10
0.08
0.06
0.04
0.02
0.00
Density
0
20
40
60
80
100
120
140
LTFRONT

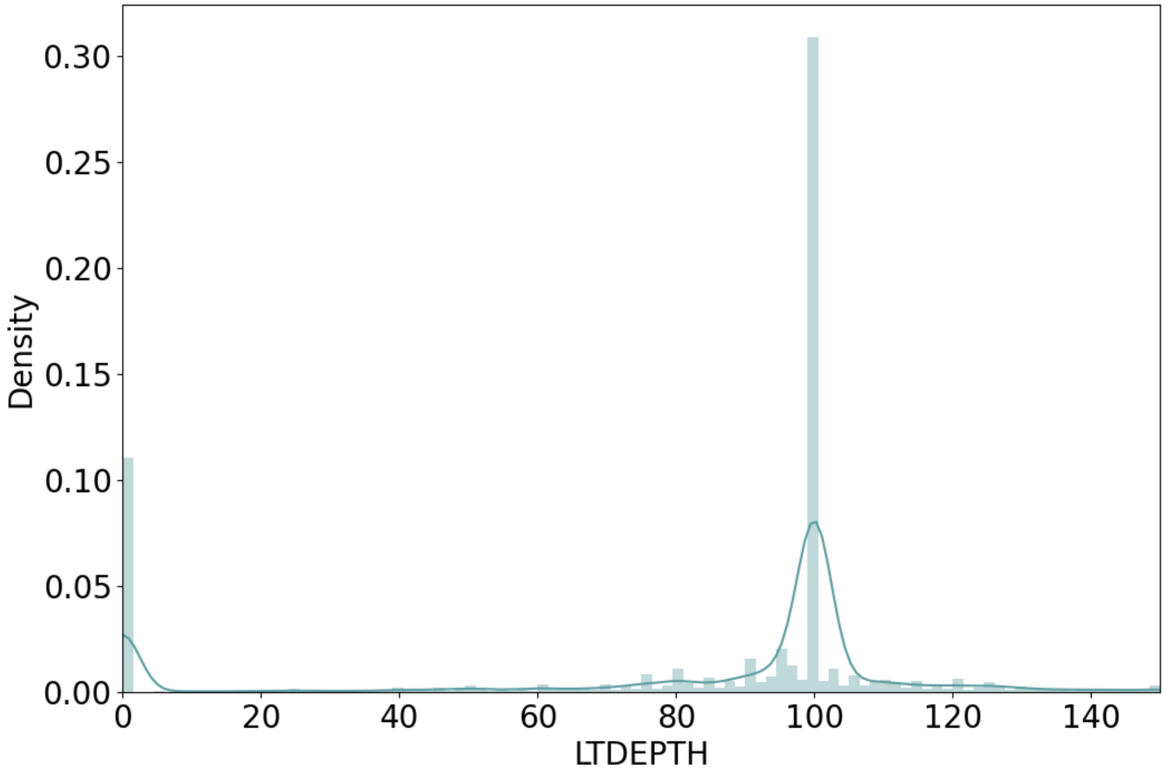


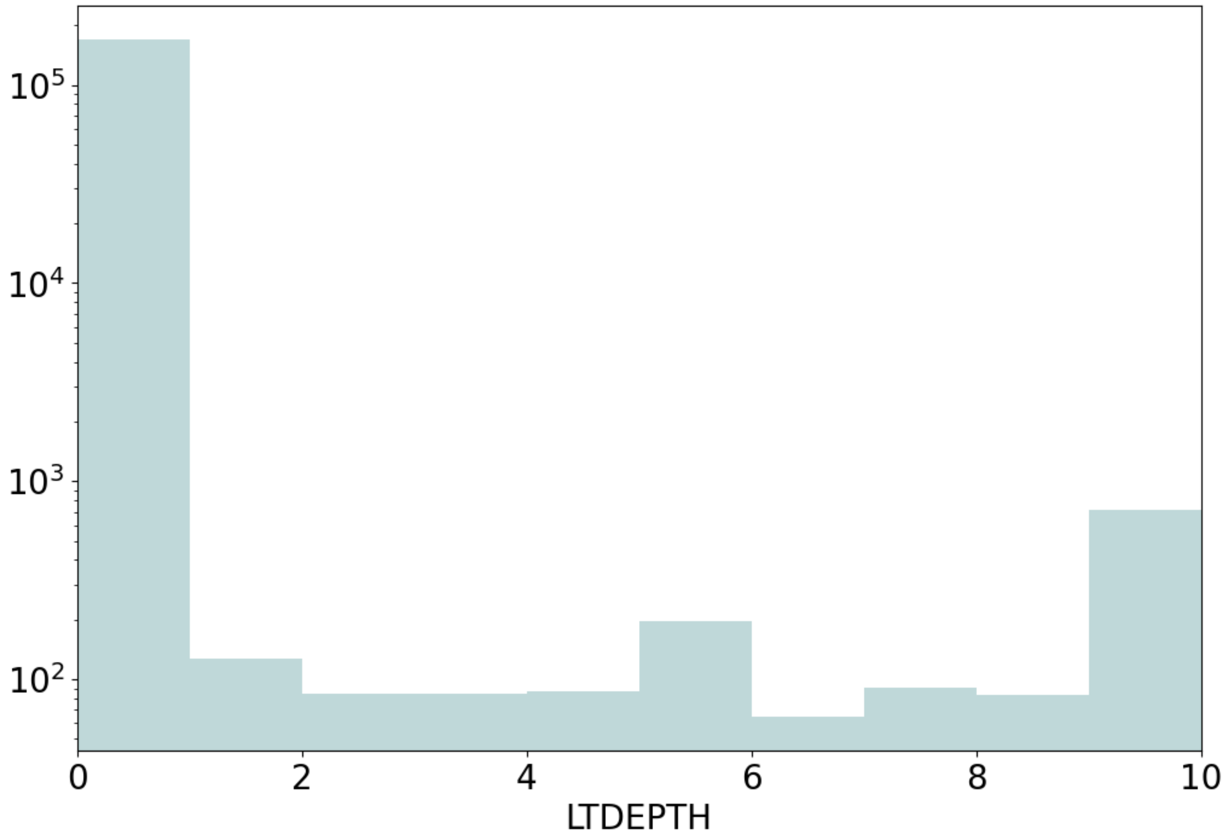
k. **LTDEPTH**

This field refers to lot depth, the distance between the front and rear property lines of land. The minimum depth is 0 and the maximum is 9,999, with a mean of 88.86, and standard deviation of 76.40. Below are the plots that show the distribution of lot depths across this dataset.

According to the density graph, lot widths at 0 and 100 are the densest. 0 does not mean that The area from 0 to 60 has really low frequency of properties. It starts picking up from 60, peaks at 100, and goes down drastically after 100.


0
2000
4000
6000
8000
10000
LTDEPTH

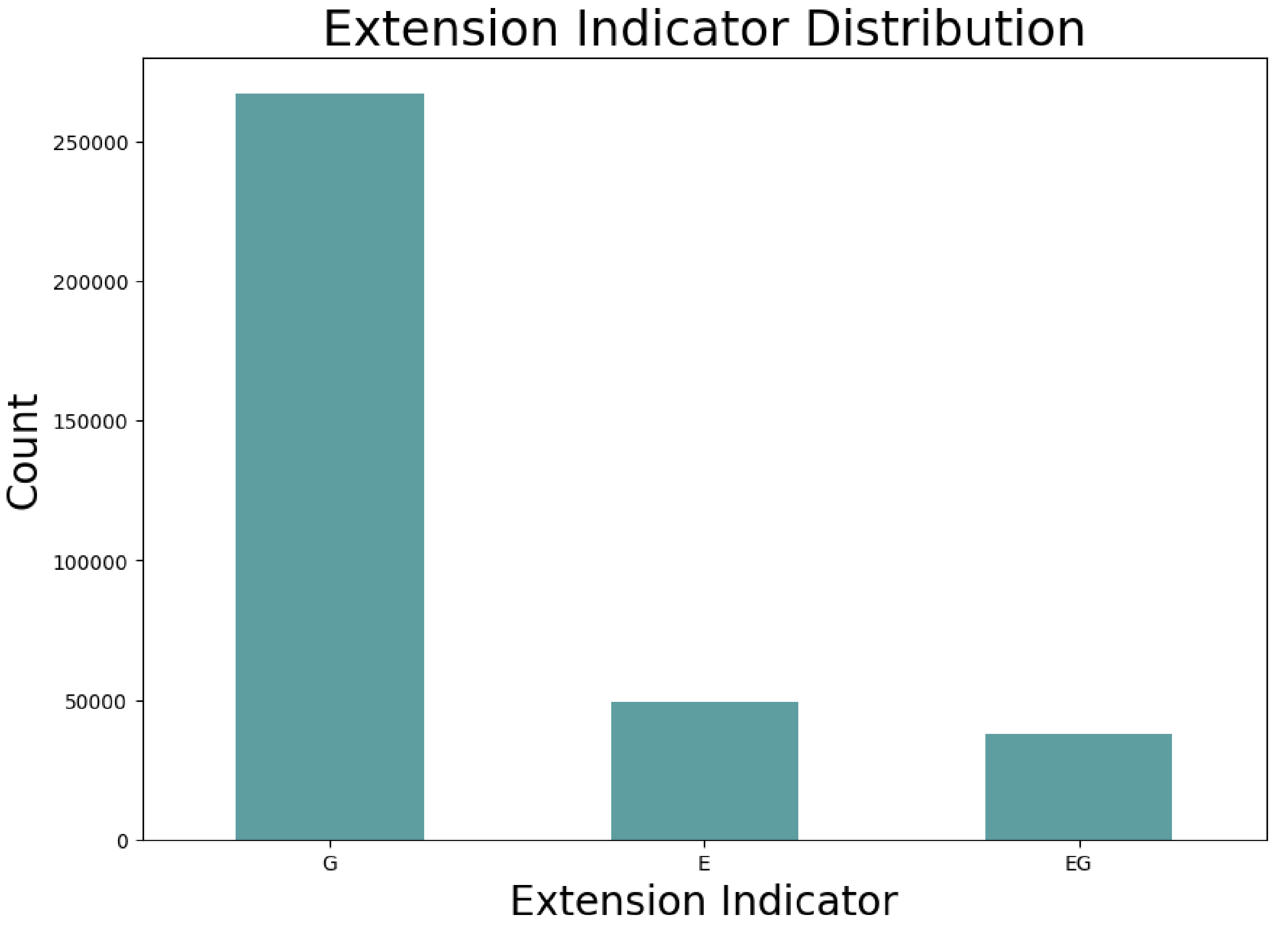

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0
20
40
60
80
100
120
140
LTDEPTH



l. **EXT**

This field refers to extension indicator, indicating whether the property has undergone any modifications or expansions beyond its original construction.

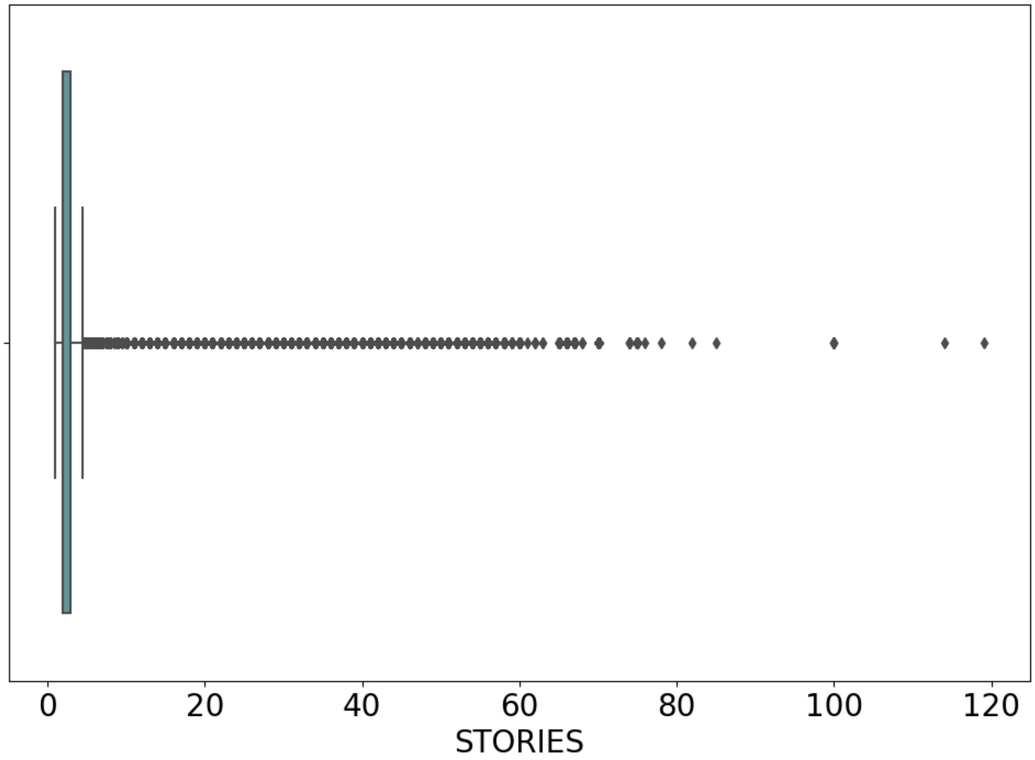
The most common extension indicator is G with more than 250,000 records. This chart shows the 3 indicators with its distribution.

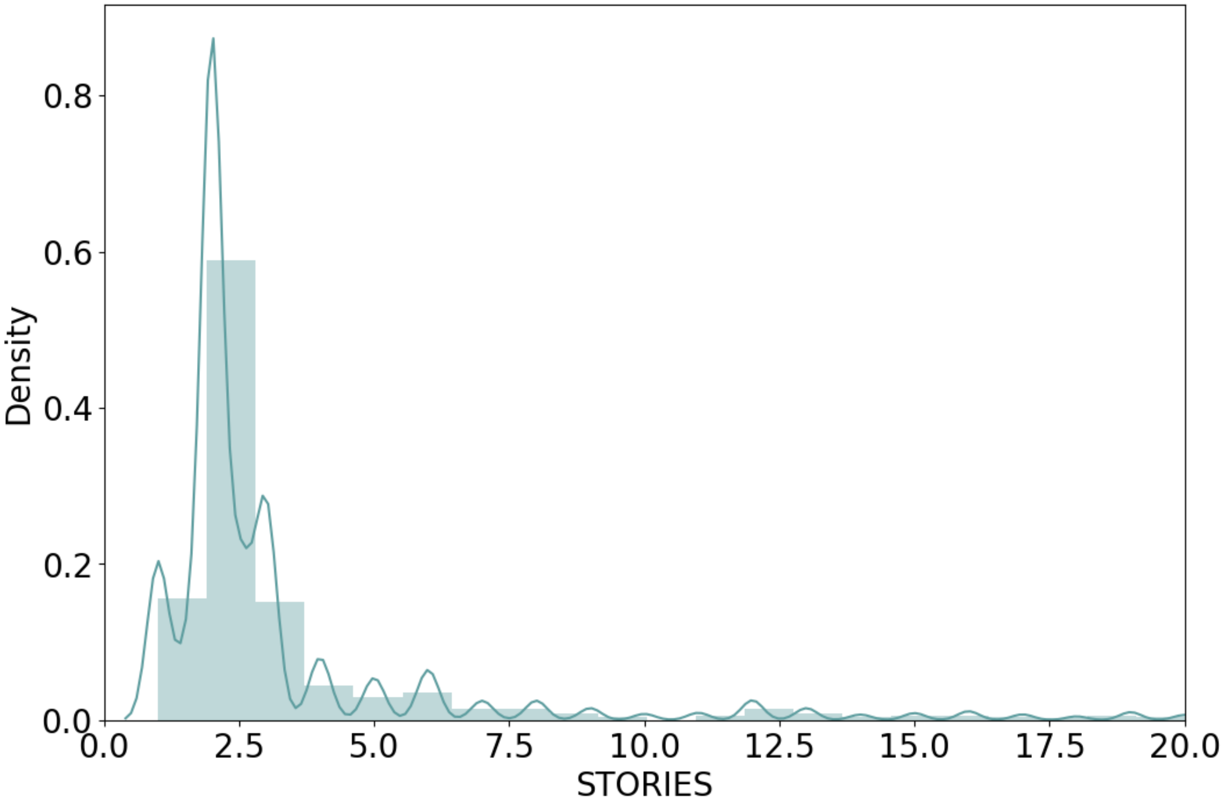


m. **STORIES**

This field refers to the number of stories in the building. The minimum number of stories is 1 and the maximum is 119, with a mean of 5, and standard deviation of 8. Below are the plots that show the distribution of stories across this dataset.

According to the density graph, number of stories peaks around 2.5 and goes down right after. The higher the number of stories, the lower the frequency.



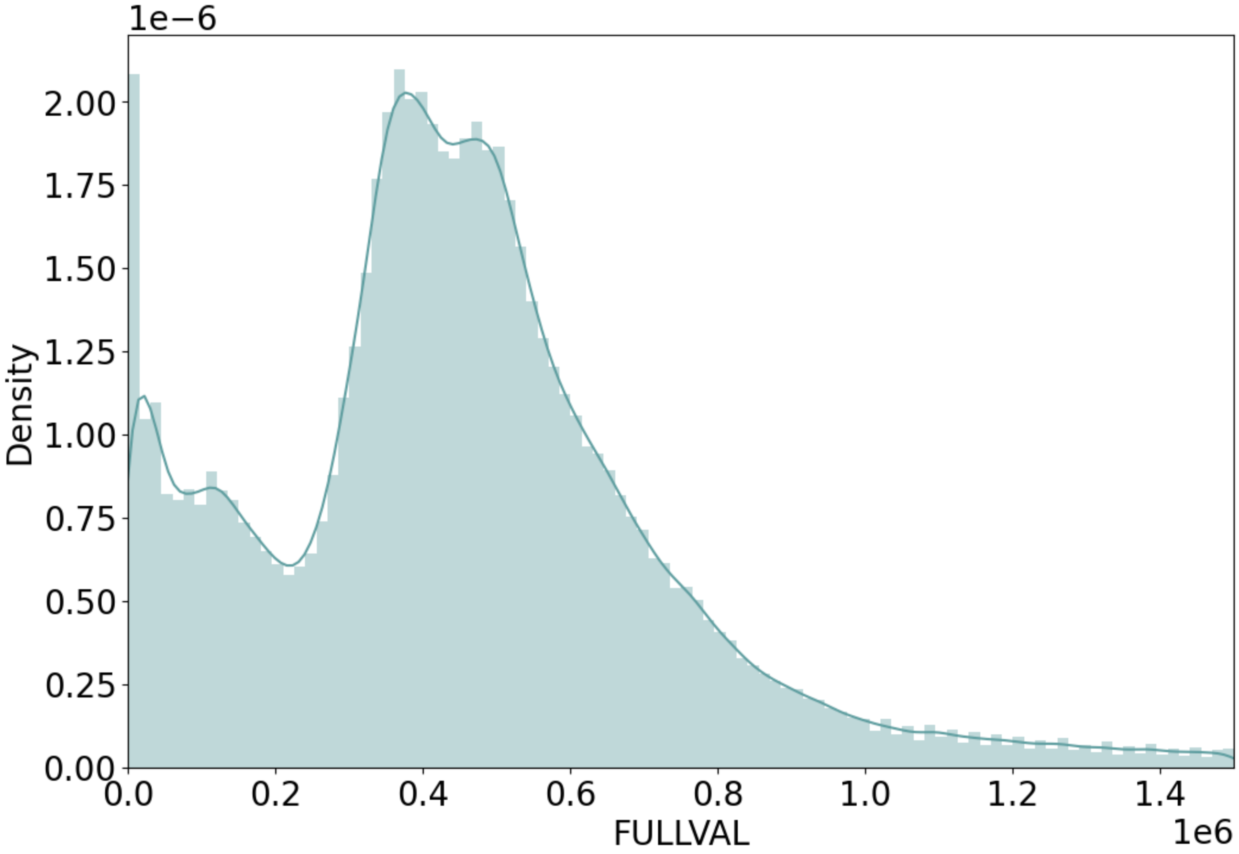


n. **FULLVAL**

This field refers to market value of the property. The minimum market value is $0 and the maximum is $6,150,000,000, with a mean of $874,264.51, and standard deviation of

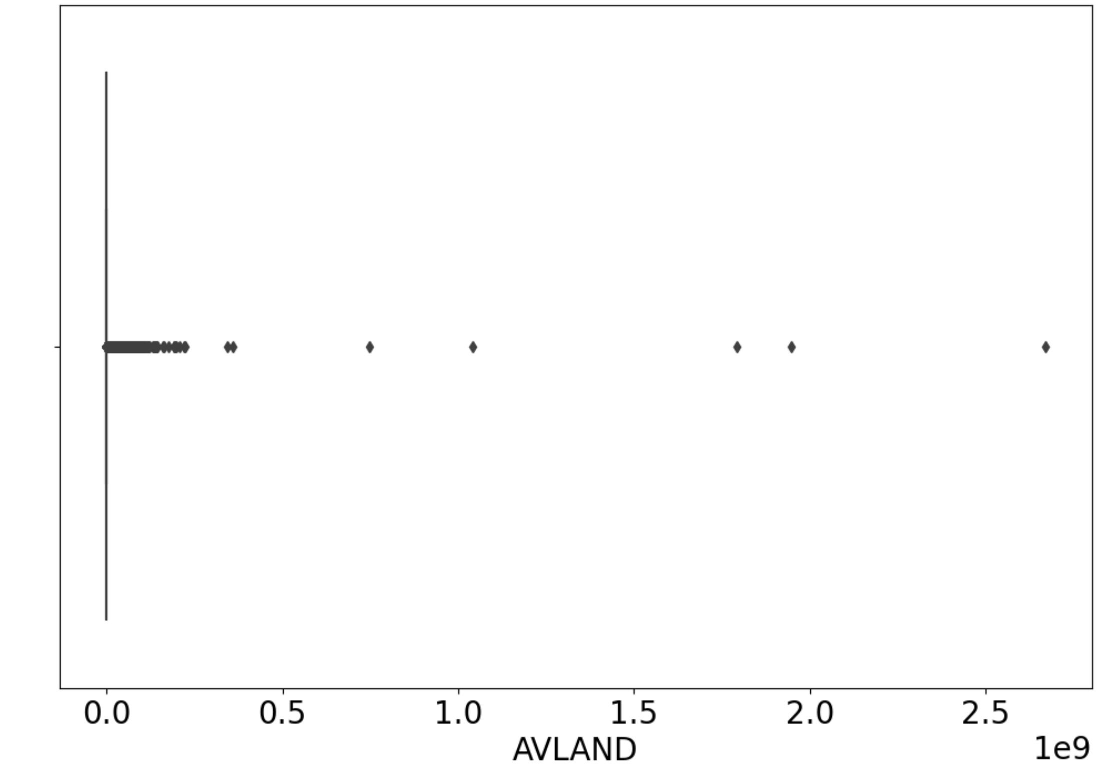
$11,582,430. Below are the plots that show the distribution of the market value across this dataset.

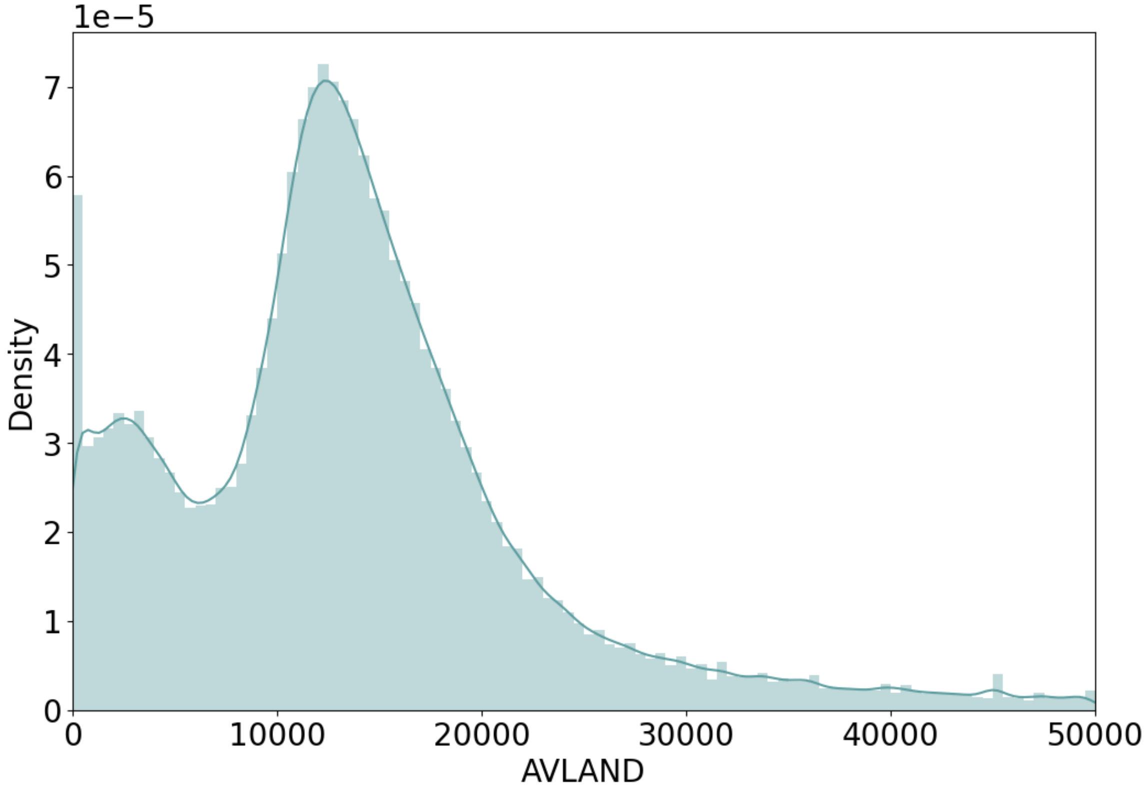
According to the density graph, market value peaks around $400,000 and decreases more after $500,000. The higher the market value, the lower the frequency.



o. **AVLAND**
This field refers to the actual land value. The minimum is $0 and the maximum is $2,668,500,000, with a mean of $85,067.92, and standard deviation of $4,057,260. Below are the plots that show the distribution of the actual land value across this dataset.

According to the density graph, actual land value peaks around $18,000 and decreases drastically as the value gets higher. There are a few extreme outliers in the box plot.

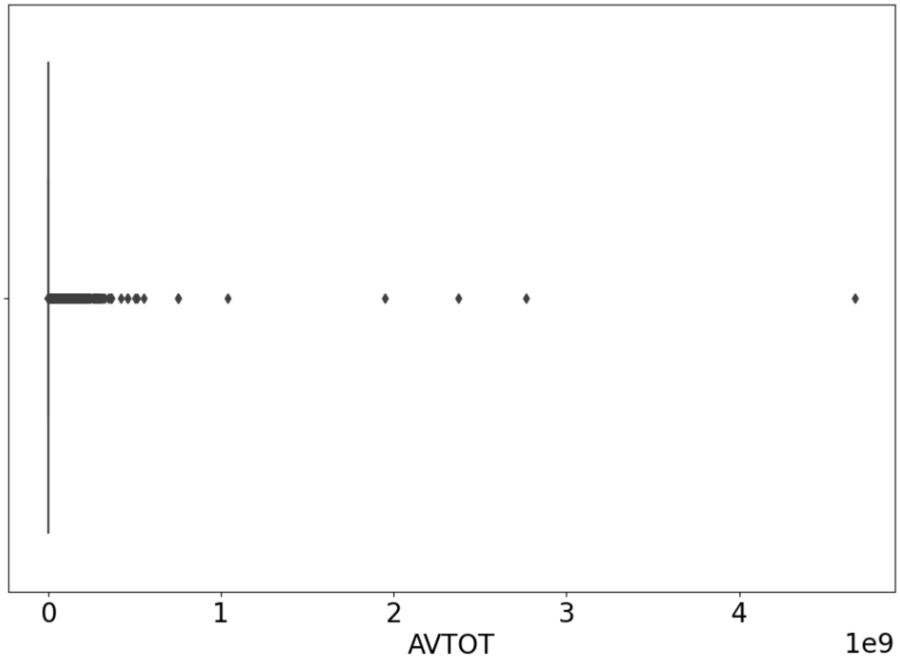

0.0
0.5
1.0
1.5
2.0
2.5
1e9
AVLAND

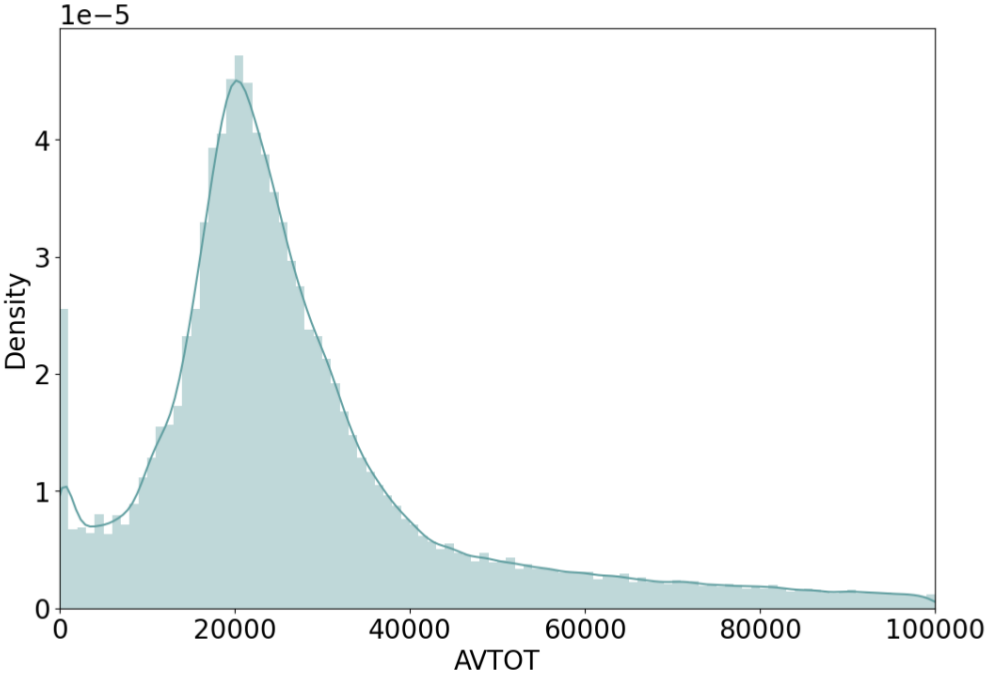

1e−5
7
6
5
4
3
2
1
0
Density
0
10000
20000
30000
40000
50000
AVLAND

p. **AVTOT**

This field refers to the actual total value. The minimum is $0 and the maximum is $4,668,309,000, with a mean of $227,238.17, and standard deviation of $6,877,529. Below are the plots that show the distribution of the actual total value across this dataset.

According to the density graph, actual total value peaks around $20,000 and decreases drastically as the value gets higher. There are a few extreme outliers in the box plot.

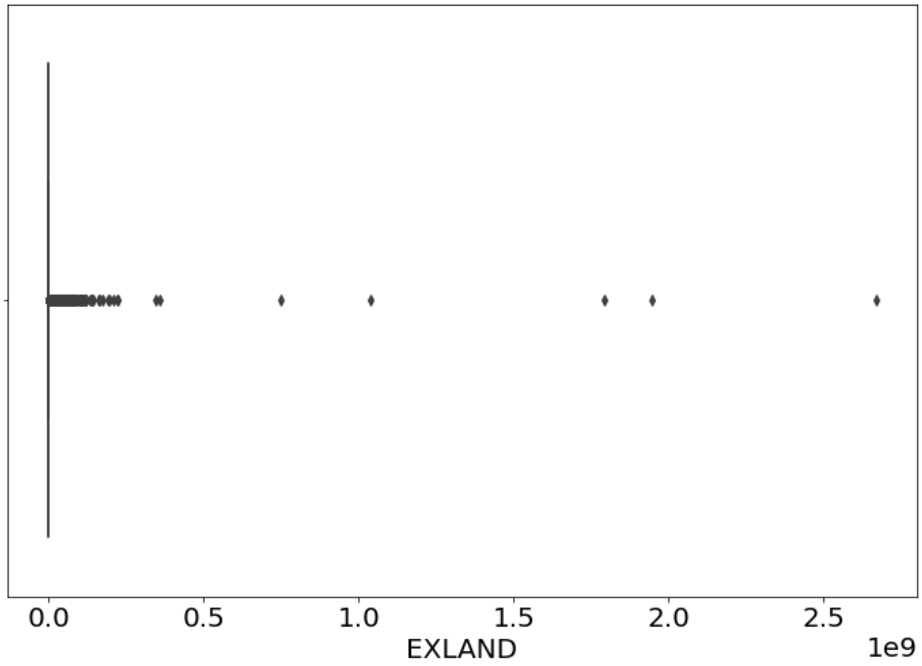


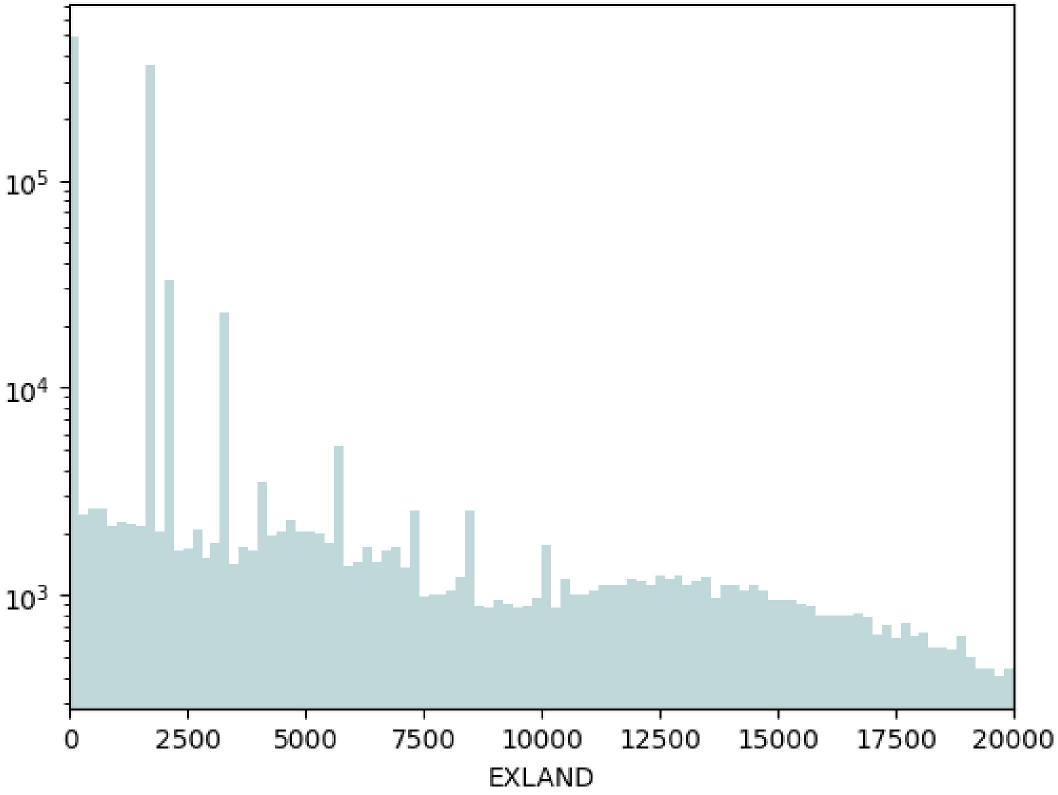


q. **EXLAND**

r. This field refers to the actual exempt land value. The minimum is $0 and the maximum is $2,668,500,000, with a mean of $36,423.89, and standard deviation of $3,981,576. Below are the plots that show the distribution of the actual exempt land value across this dataset.

Most properties have the actual total value at 0 and around $1600-1700. The properties at 0 might be missing values.

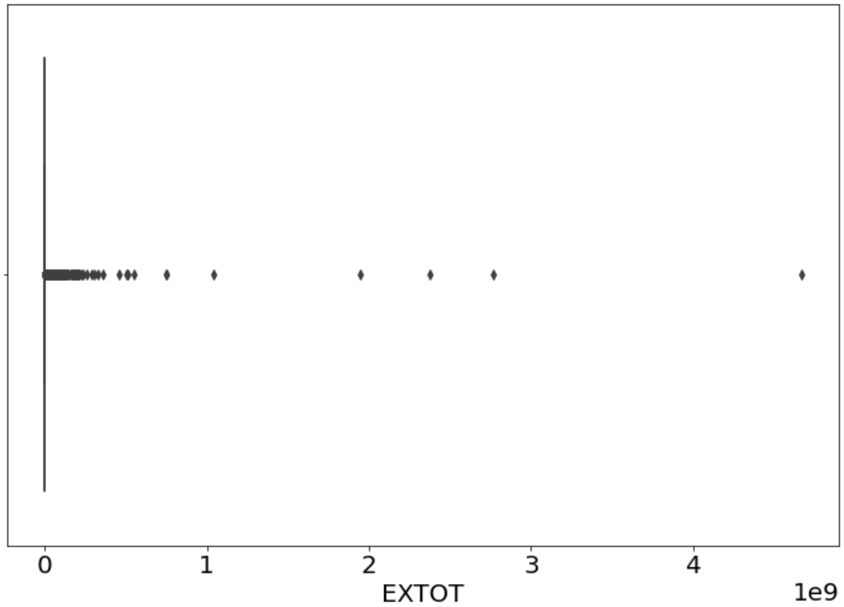


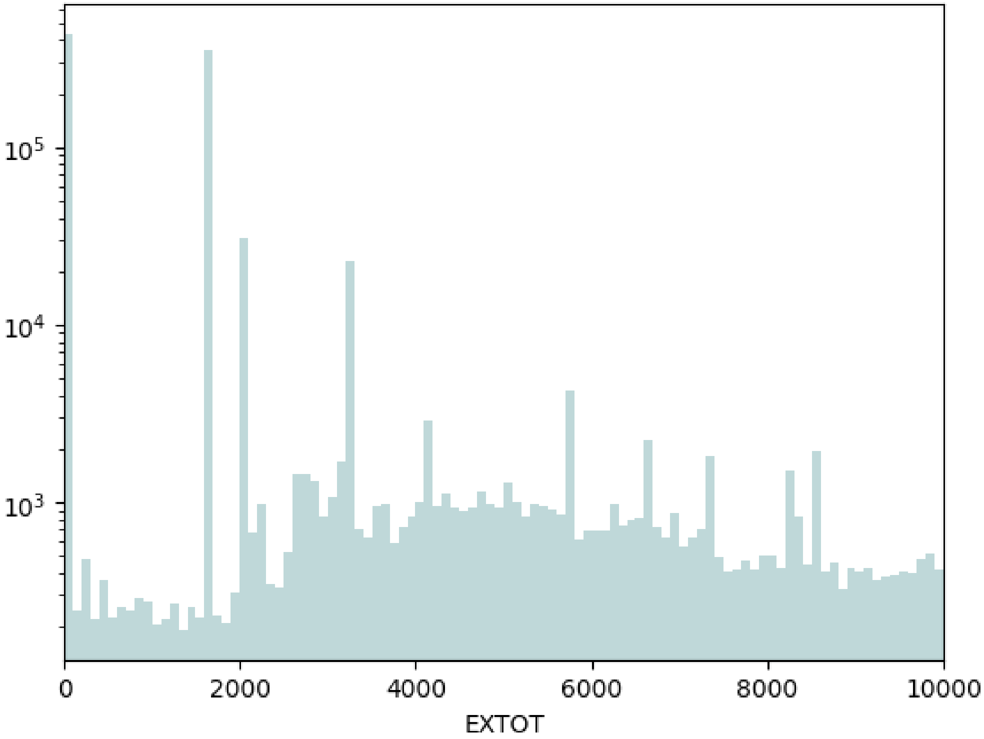


s. **EXTOT**

This field refers to the actual exempt land total. The minimum is $0 and the maximum is $4,668,309,000, with a mean of $91,186.98, and standard deviation of $6,508,403. Below are the plots that show the distribution of the actual exempt land total across this dataset.

Most properties have the actual total value at 0 and around $1600-1700, same as actual exempt land value. The properties at 0 might be missing values.

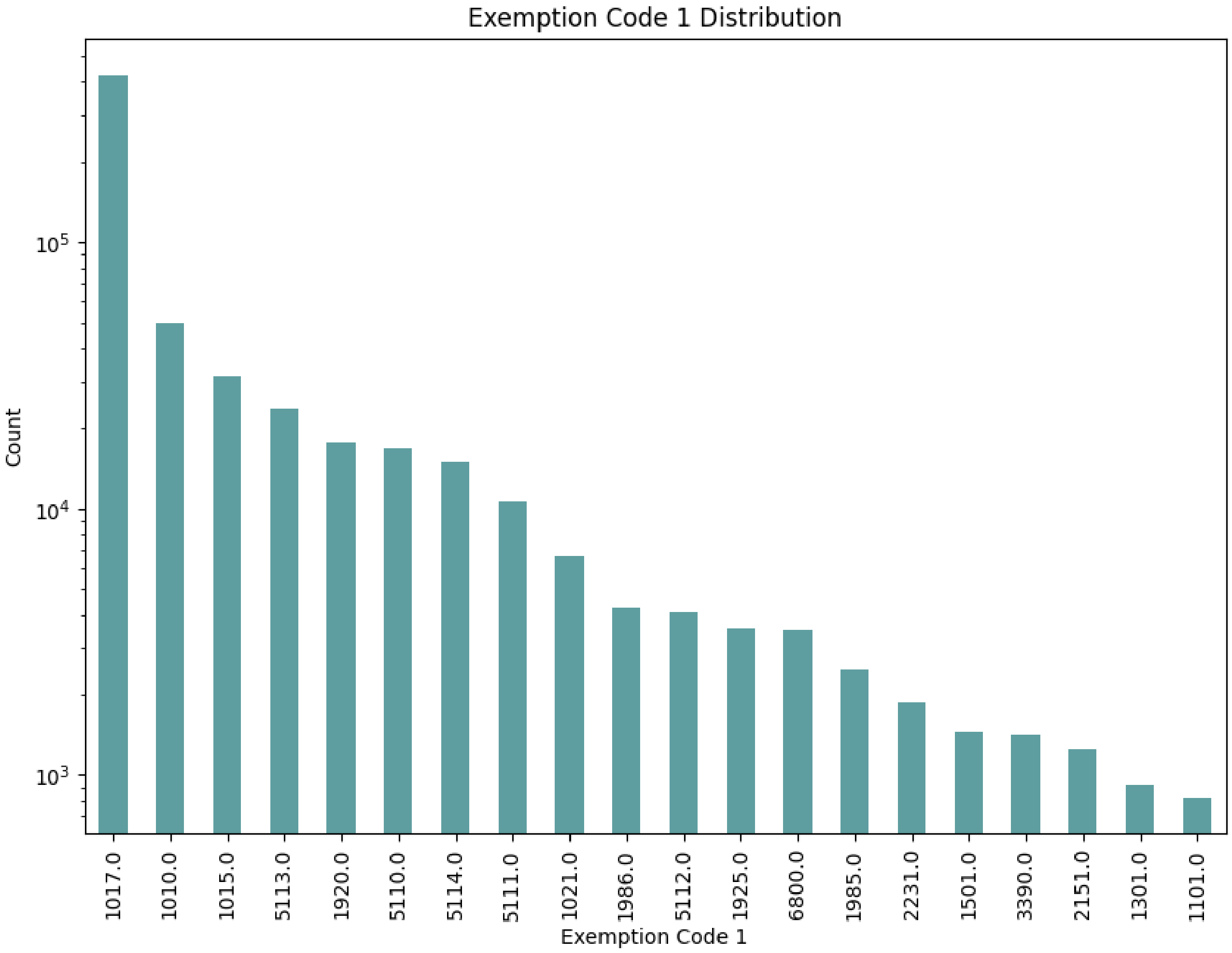




t. **EXCD1**

This field indicates exemption code 1. The most common value is 1017.0 with more than 400,000 properties.

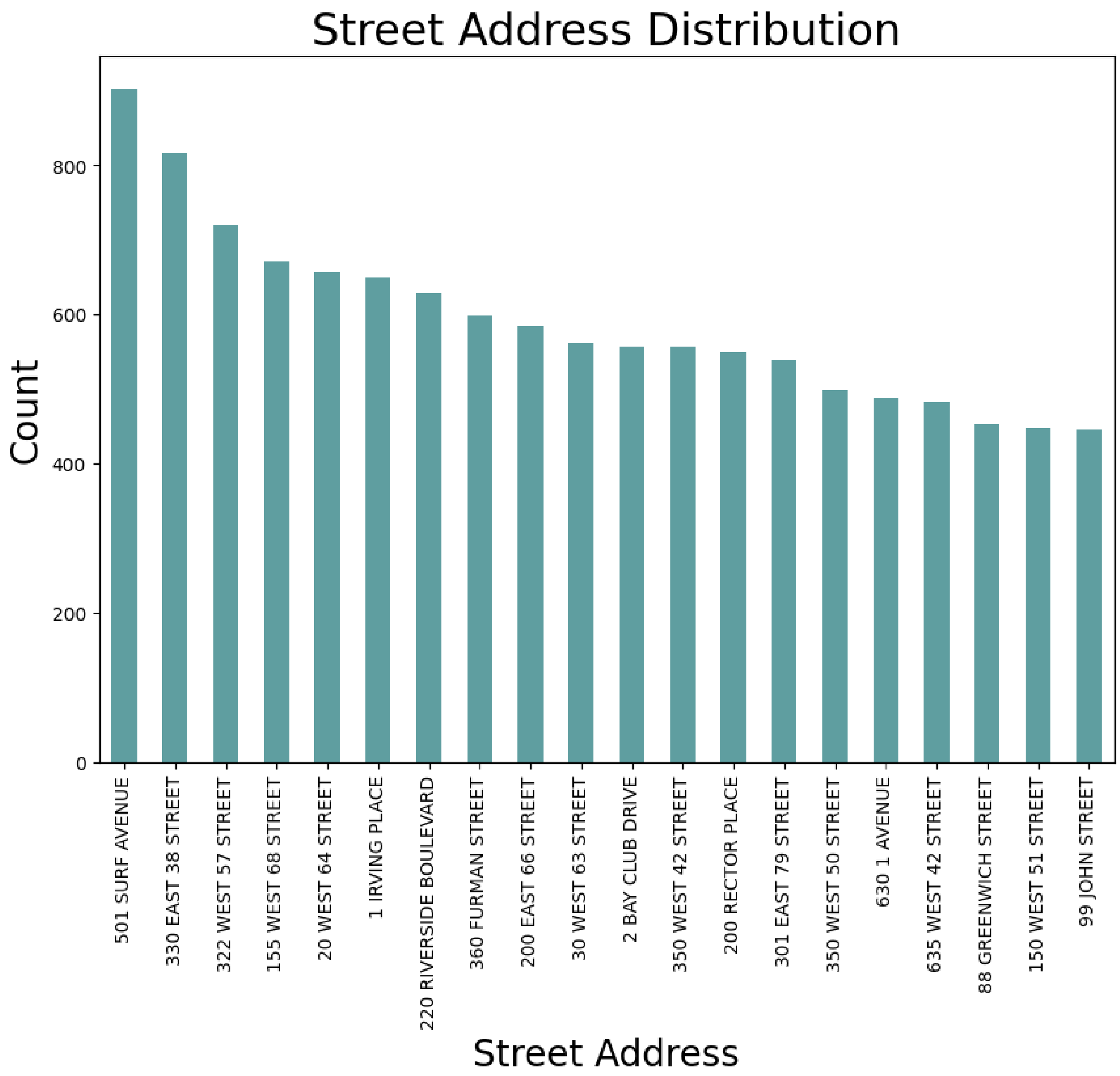
The chart shows the 20 most common values of exemption code 1 with the count of properties, among a total of 129 unique ones.



u. **STADDR**

This field indicates street address. The most common value is 501 SURF AVENUE with more than 800 properties.

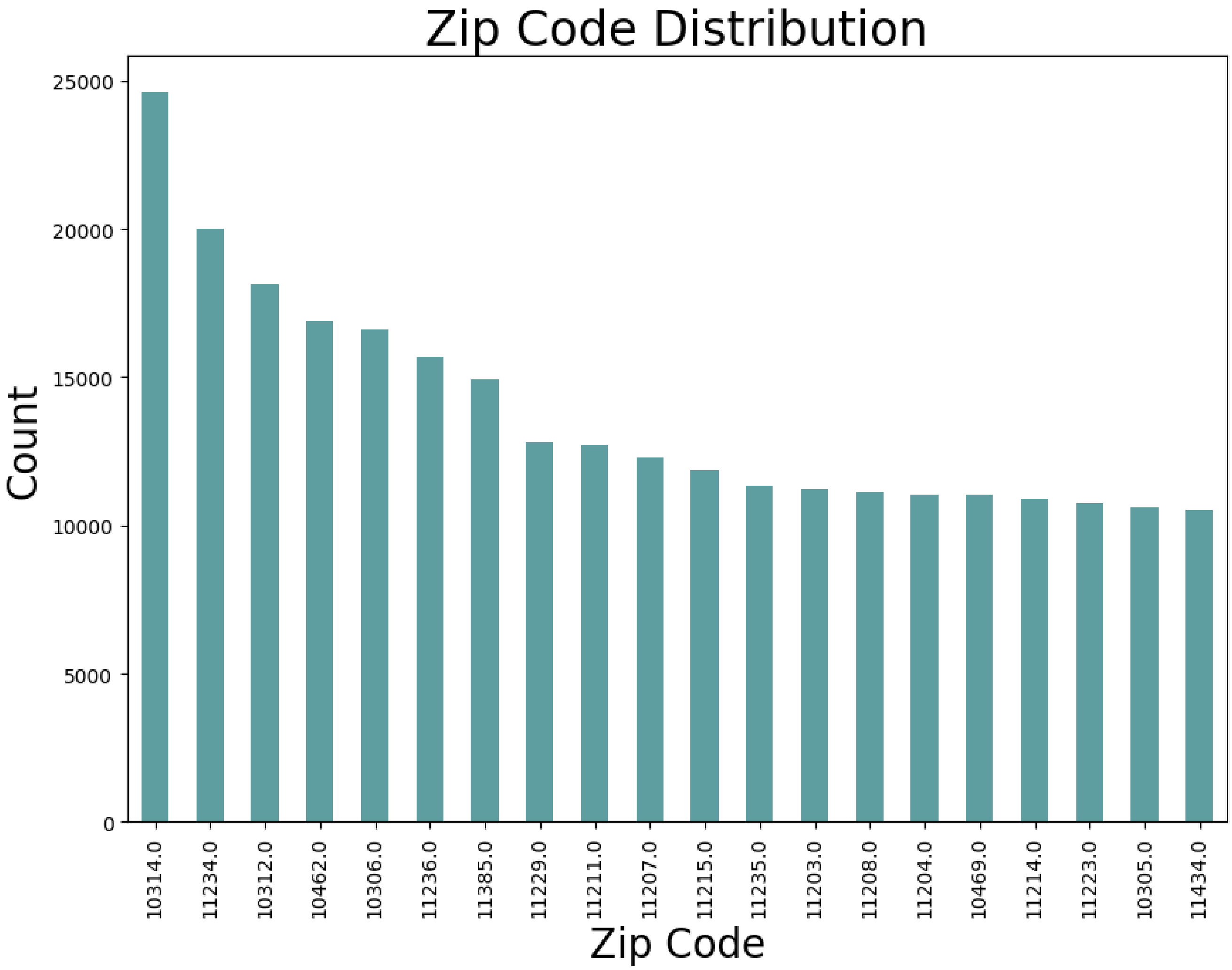
The chart shows the 20 most common values of street address 1 with the count of properties, among a total of 839,281 unique ones.



v. **ZIP**

This field indicates zip code. The most common value is 10314 with almost 25,000 properties.

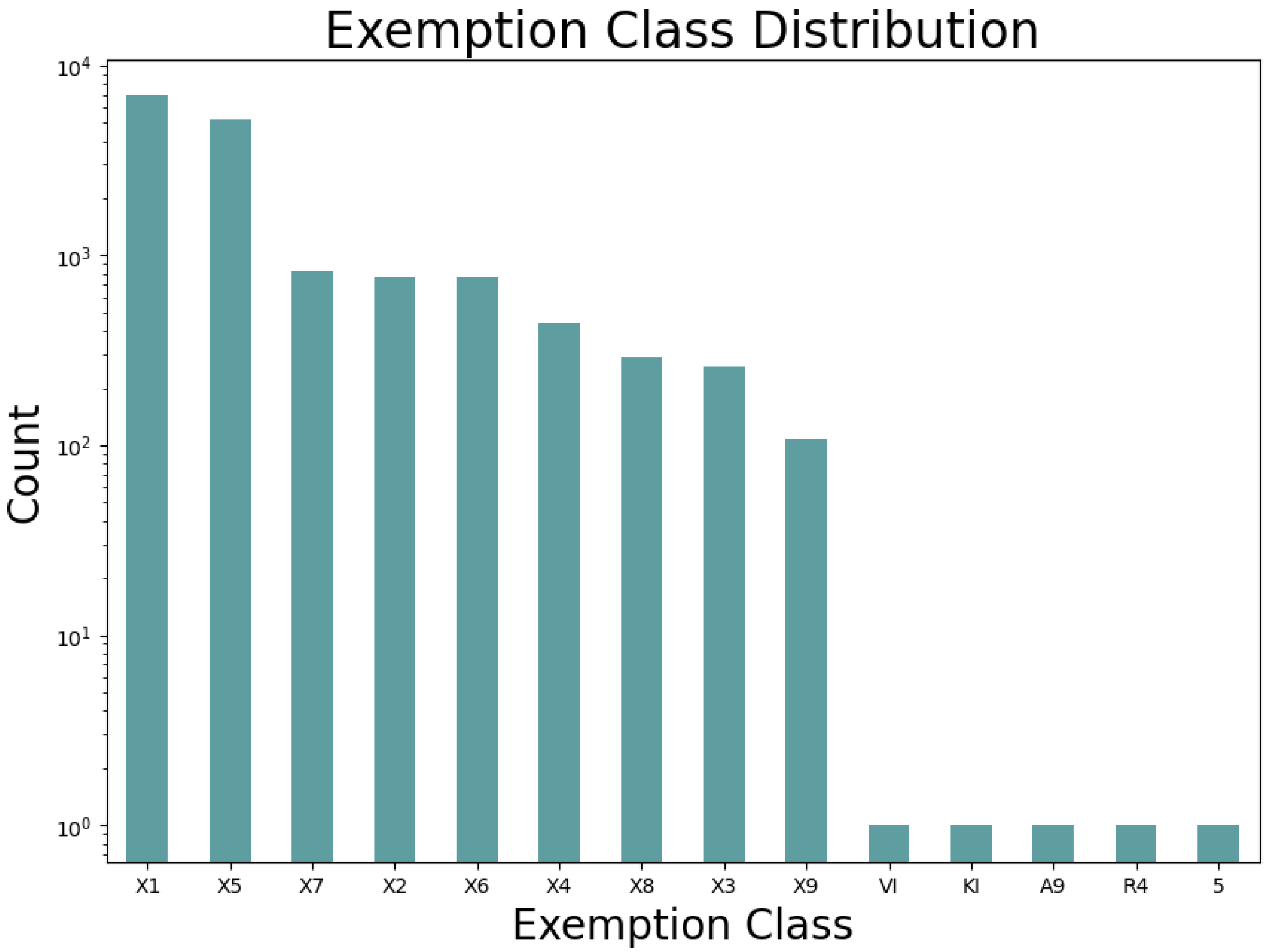
The chart shows the 20 most common values of street address with the count of properties, among a total of 197 unique ones.



w. **EXMPTCL**

This field indicates exemption class. The most common value is X1 with almost 7000 properties.

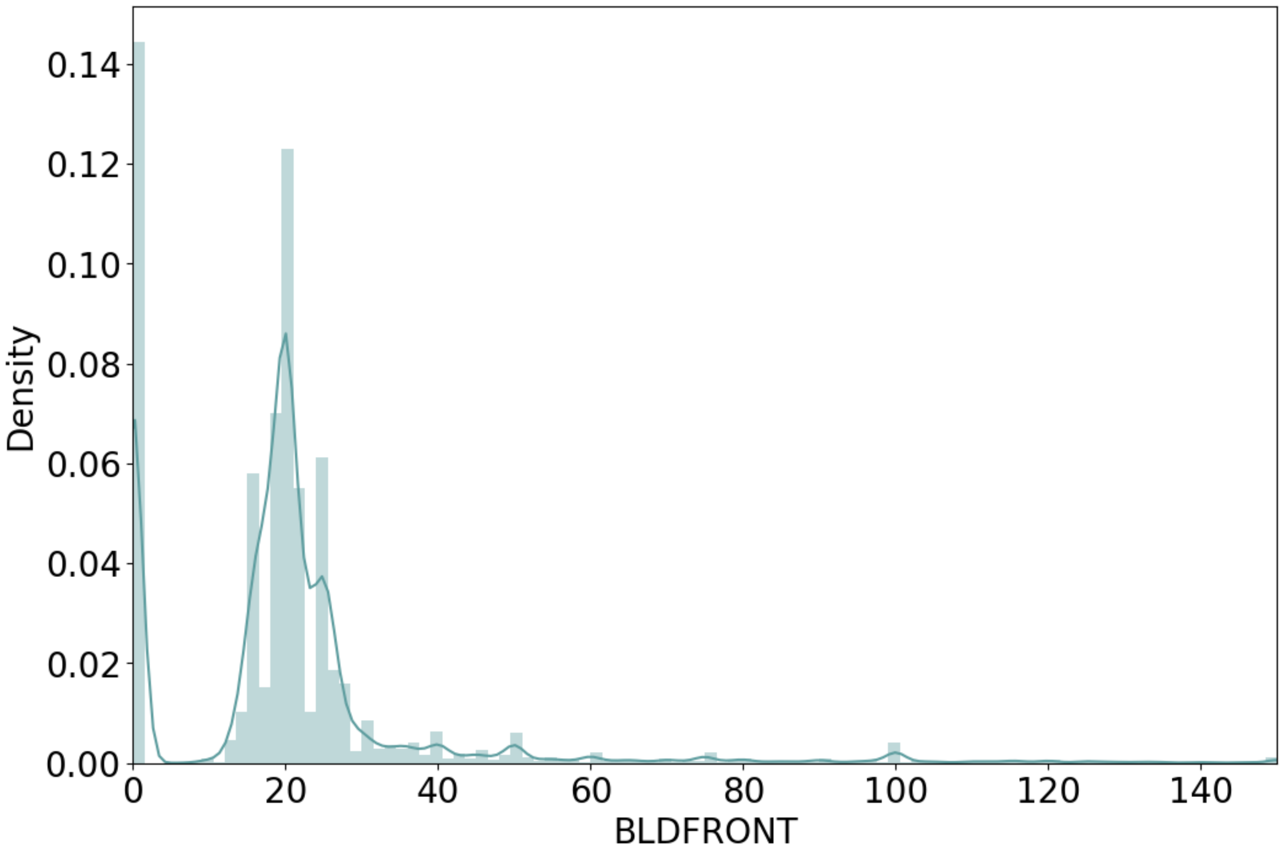
The chart shows 15 values of exemption class with the count of properties.



x. **BLDFRONT**

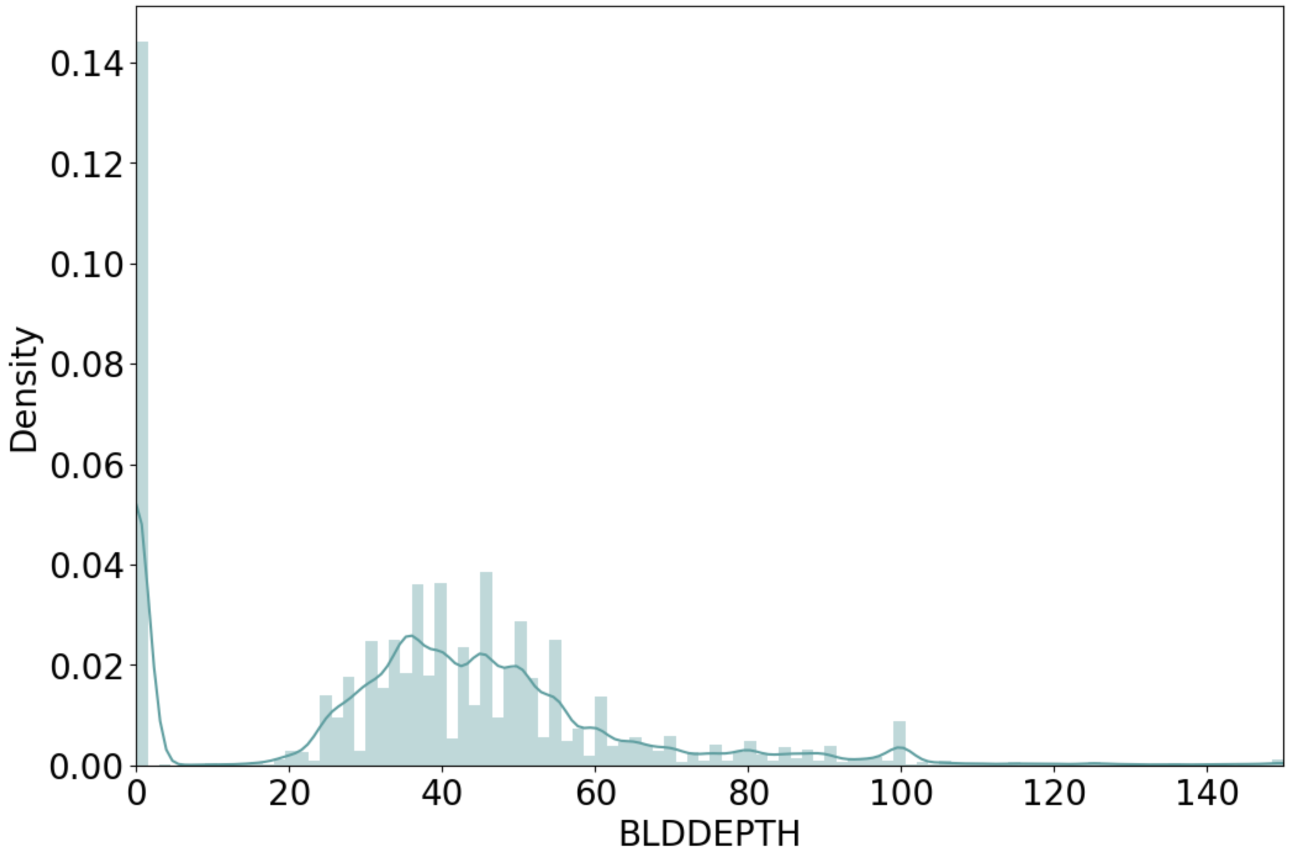
This field refers to building width. The minimum is 0 and the maximum is 7,575, with a mean of 23.04, and standard deviation of 35.58. Below is the density plot that shows the distribution of building widths across this dataset.

Most properties have building widths at 0 and around 20 (15-25).



y. **BLDDEPTH**
This field refers to building depth. The minimum is 0 and the maximum is 9,373, with a mean of 39.92, and standard deviation of 42.71. Below is the density plot that shows the distribution of building depths across this dataset.

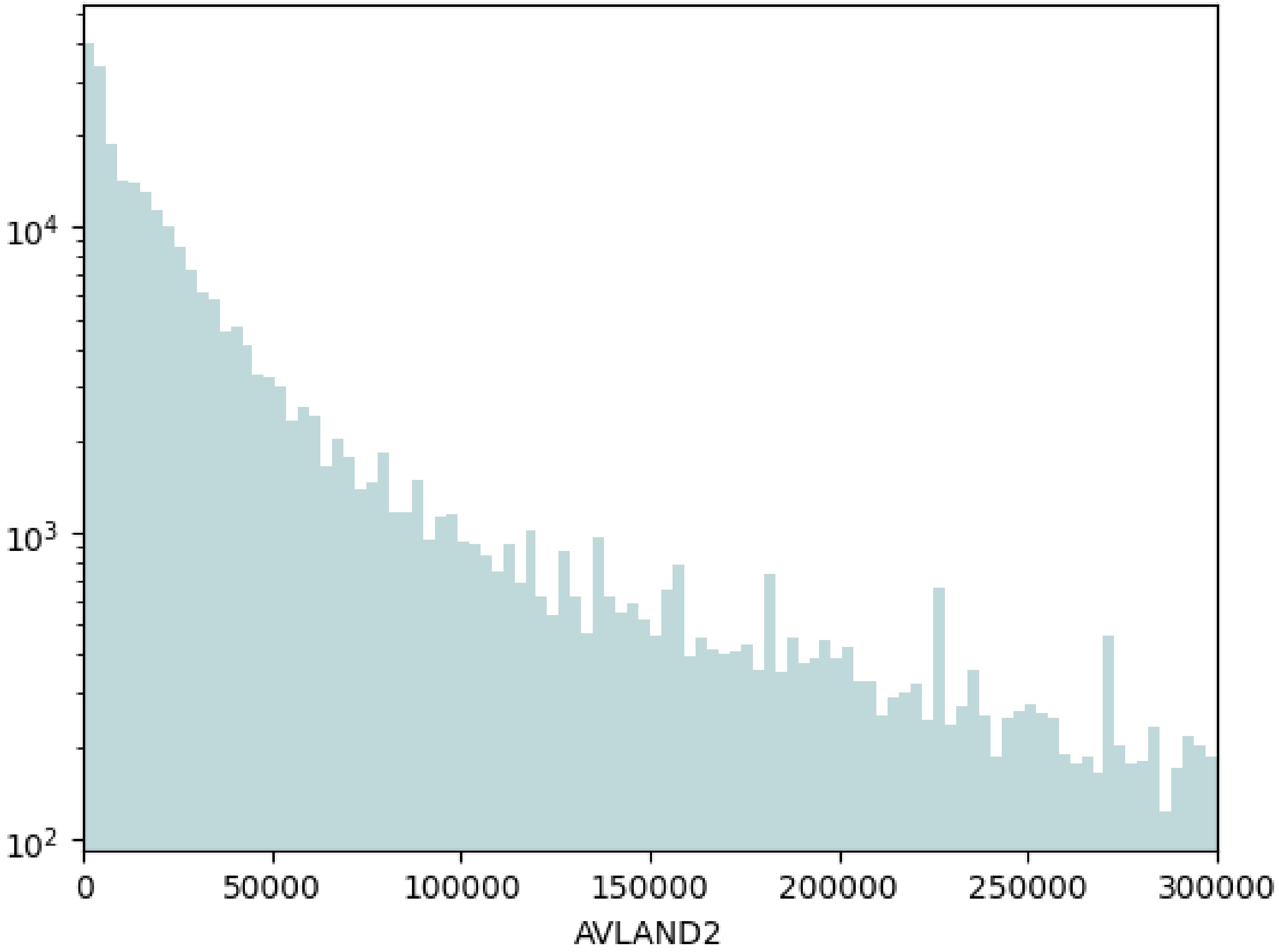
Most properties have building widths at 0 and the range between 20 and 100.



z. **AVALAND2**

This field refers to transitional land value. The minimum is $3 and the maximum is $2,371,005,000, with a mean of $246,235.72, and standard deviation of $6,178,963. Below is the plot that shows the distribution of transitional land value across this dataset.

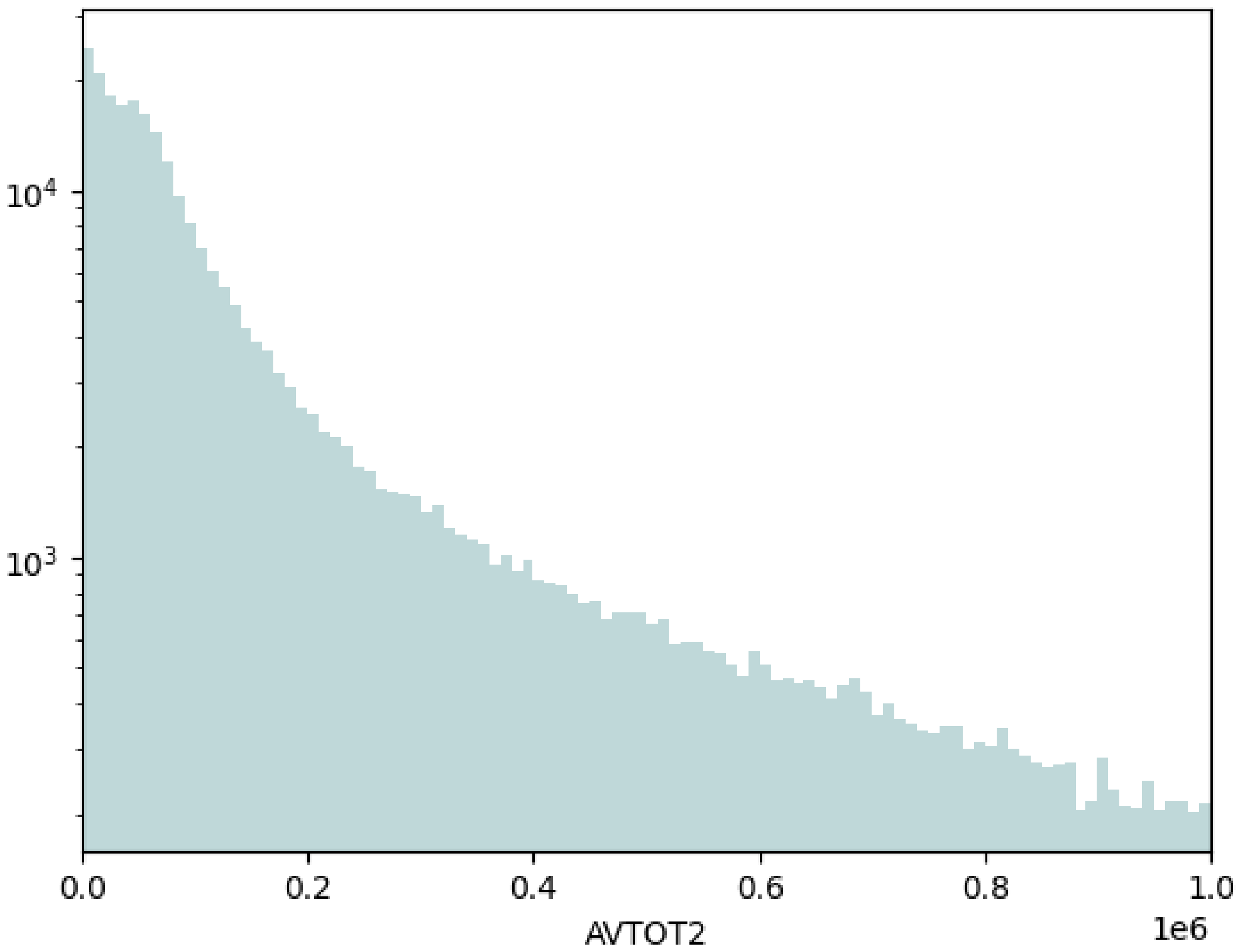
Properties distributes the most when the transitional land value is small. As the value gets higher, there are fewer properties.



aa. **AVATOT2**

This field refers to transitional total value. The minimum is $3 and the maximum is $4,501,180,000, with a mean of $713,911.44, and standard deviation of $11,652,530. Below is the plot that shows the distribution of transitional total value across this dataset.

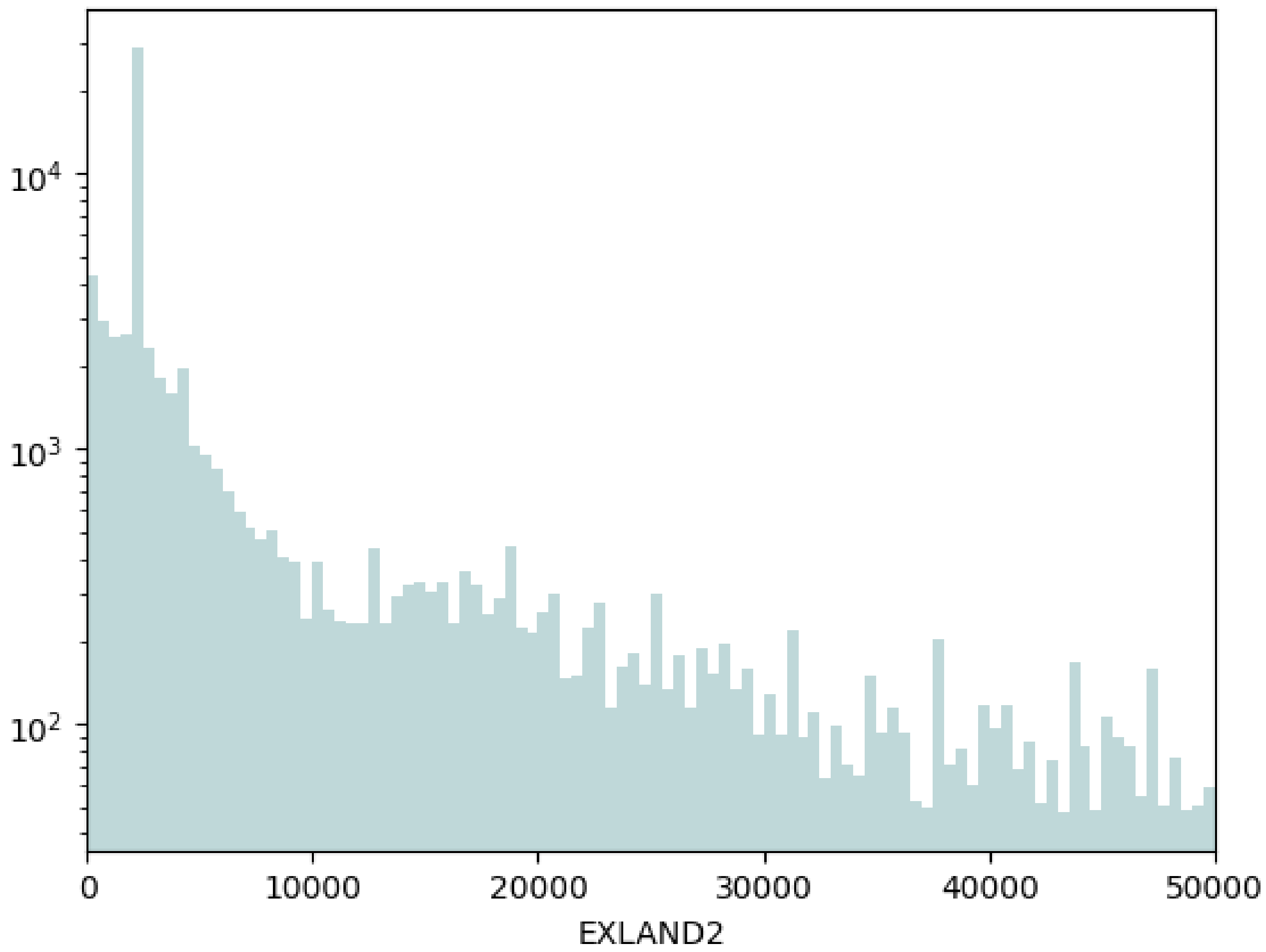
Properties distributes the most when the transitional total value is small. As the value gets higher, there are fewer properties.



bb. **EXLAND2**

This field refers to transitional exemption land value. The minimum is $1 and the maximum is $2,371,005,000, with a mean of $351,235.68, and standard deviation of $10,802,21. Below is the plot that shows the distribution of transitional exemption land value across this dataset.

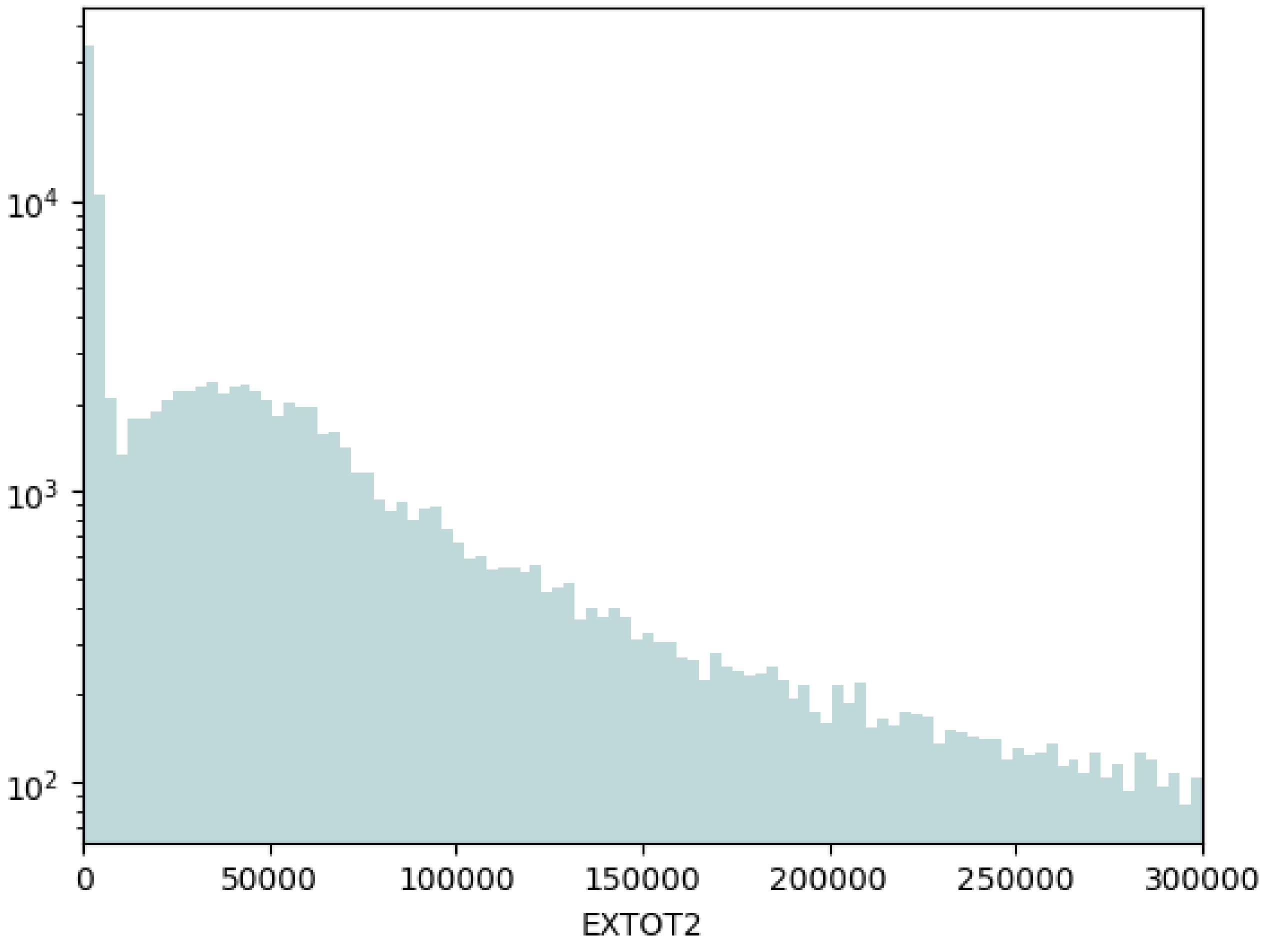
Properties distributes the most when the transitional exemption land value is small, especially around 2000. As the value gets higher, there are fewer properties.



cc. **EXTOT2**

This field refers to transitional exemption land total. The minimum is $7 and the maximum is $4,501,180,000, with a mean of $656,768.28, and standard deviation of $16,072,510. Below is the plot that shows the distribution of transitional exemption land total across this dataset.

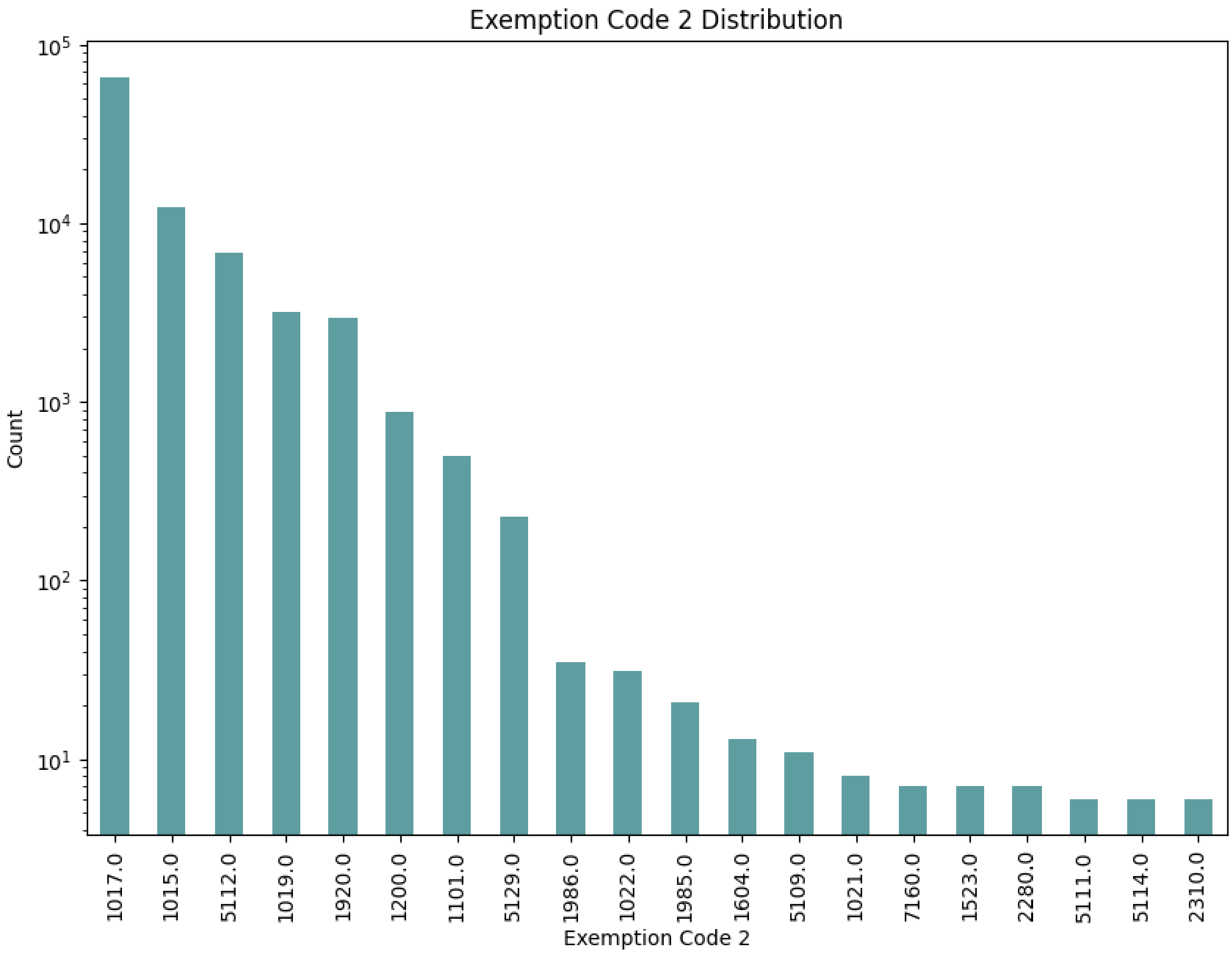
Properties distributes the most when the transitional exemption land total is small, especially around $7-1000. As the value gets higher, there are fewer properties.



dd. **EXCD2**

This field indicates exemption code 2. The most common value is 1017.0.

The chart shows the 20 most common values of exemption code 1 with the count of properties, among a total of 60 unique ones.



ee. **PERIOD**

This field indicates the assessment period when data was created. There is only 1 value for every property.

ff. **YEAR**

This field indicates the assessment period. There is only 1 value for every property.

gg. **VALTYPE**

This field is identical for all records, indicating the file identity, not a field.