Shuhan Wang

Fraud Analytics

NY Property Project Report

I. Executive Summary

This project builds unsupervised anomaly detection algorithms to detect tax fraud based on the NY property data. The steps include a data quality report that provide statistical summary of each variable and data cleaning process, in which we fill in missing values and implemented data exclusions. The next steps are variable creation and dimensionality reduction to determine the final number of chosen variables, and anomaly detection algorithms that described score methods used to measure fraudulent property data.

II. Data Description

The NY property data contains data about NY properties. The department of finance uses property data to calculate property tax bills every year. Specifically, each record contains information about each property, such as address, owners, property characteristics etc. The dataset is in the year **2010-2011** and has **32 fields** of NY property information and **1,070,994 records**.

2. Summary Tables

(1) Numerical Table

Field Name	% Populated	Min	Max	Mean	Stdev	# Zero
LTFRONT	100.00	0.0	9999.00	36.64	74.03	169108
LTDEPTH	100.00	0.0	9999.00	88.86	76.40	170128
STORIES	94.75	1.0	119.00	5.01	8.37	0
FULLVAL	100.00	0.0	6150000000.00	875264.51	11582430.99	13007
AVLAND	100.00	0.0	2668500000.00	85067.92	4057260.06	13009
AVTOT	100.00	0.0	4668308947.00	227238.17	6877529.31	13007
EXLAND	100.00	0.0	2668500000.00	36423.89	3981575.79	491699
EXTOT	100.00	0.0	4668308947.00	91186.98	6508402.82	432572
BLDFRONT	100.00	0.0	7575.00	23.04	35.58	228815
BLDDEPTH	100.00	0.0	9393.00	39.92	42.71	228853
AVLAND2	26.40	3.0	2371005000.00	246235.72	6178962.56	0
AVTOT2	26.40	3.0	4501180002.00	713911.44	11652528.95	0

EXLAND2	8.17	1.0	2371005000.00	351235.68	10802212.67	0
EXTOT2	12.22	7.0	4501180002.00	656768.28	16072510.17	0

(2) Categorical Table

Field Name	% Populated	# Blank	# Zeros	# Unique Values	Most Common Value
RECORD	100.00	0	0	1070994	1
BBLE	100.00	0	0	1070994	1000010101
BORO	100.00	0	0	5	4
BLOCK	100.00	0	0	13984	3944
LOT	100.00	0	0	6366	1
EASEMENT	0.43	1066358	0	12	E
OWNER	97.04	31745	0	863347	PARKCHESTER PRESERVAT
BLDGCL	100.00	0	0	200	R4
TAXCLASS	100.00	0	0	11	1
EXT	33.08	716689	0	3	G
EXCD1	59.62	432506	0	129	1017.0
STADDR	99.94	676	0	839280	501 SURF AVENUE
ZIP	97.21	29890	0	196	10314.0
EXMPTCL	1.45	1055415	0	14	X1
EXCD2	8.68	978046	0	60	1017.0
PERIOD	100.00	0	0	1	FINAL
YEAR	100.00	0	0	1	2010/11
VALTYPE	100.00	0	0	1	AC-TR

III. Data Cleaning

In the data cleaning stage, our data imputation method of the NY property dataset involves removing exclusions and fill in the missing values, in which we substitute absent data with alternative data by looking at related fields.

To remove the exclusions, we look at the "**Owner**" field and remove the rows/properties owned by governments, leaving only private property owners because we are only interested in private owners committing tax fraud. We removed 24478 rows with government properties.

For the missing value in the field **ZIP** (Total missing 21537), if ZIP value before and after the missing value are the same, we fill in the missing value with the value before and after that. We do this because we want the filled in values to be as normal/appropriate as possible. This helped us fill in 11423 missing values in the ZIP field, leaving 10114 unfilled missing ZIP values. For these remaining missing values, we fill in the missing values with the record above it.

For the null values in **FULLVAL, AVLAND,** and **AVTOT,** we replace the missing values with the mean value grouped by taxclass.

For the missing values in **STORIES** (Total missing values is 43684), we also fill in the missing values by taking the averages of values grouped by tax class.

Lastly, for the missing data values in **LTFRONT**, LTDEPTH, BLDDEPTH, BLDFRONT, we values like 0 and 1 with NA, because values like 0 and 1 are invalid values for these fields and also by replacing them with NA we don't count them when calculating the average. Similar to STORIES, we replace these missing values with NA with the average grouped by TAXCLASS.

IV. Variable Creation

In the Variable Creation stage after cleaning the data related to New York properties for the purpose of detecting tax fraud, I created **58 baseline variables** and **18 additional variables** based on the listed characteristics of the NY properties. Creating variables based on the original variables given in the dataset helps simulating tax fraud situations specific to properties in New York. By combining certain variables, we can also mimic patterns and fraudulent behaviors commonly observed in actual tax fraud cases.

The first variable is the **Size Variables**, which include Lot size, building area, and building volume variables. Lot size is obtained from multiplying the original variables LTFRONT and LTDEPTH, Building area is obtained from BLDFRONT and BLDDEPTH, and Building Volume is obtained from Lot size * STORIES.

I created **Price Ratio Variables**, which are r1,r2,r3,r4,r5,r6,r7,r8,r9. These 9 variables are made from normalizing each of the FULLVAL, AVLAND, AVTOT by each of the ltsize, bldsize, bldvol. By normalizing these variables, we can use ratio to identify if properties are reporting inflated or deflated property values. Then I created another 9 variables that are inverse of Price Ratio to capture the overall relationship between price and sizes. These inverse variables will detect properties that are overvalued and undervalued compared to other similar properties in the area. It can compare properties across neighborhoods, cities, and regions.

The next two variables are **Ratio by Zip** and **Ratio by Tax Class**, and each of them have 18 variables. Ratio by Zip look at property values and benchmark based on each geographical area

rather on the individual levels and Ratio by Tax Class calculates the relationship between variables within a specific tax class.

The last baseline variable is **Value Ratio Variable**, which is calculated as the ratio of FULLVAL(market value) to the sum of AVTOT (actual total value) and AVLAND (actual land value) to provide an estimation of the accuracy of assessed property values.

In addition, I added one more variable - **Ratio by Stories** Variable. The grouped averages of 18 variables (r1, r2, r3, r4, r5, r6, r7, r8, r9, r1inv, r2inv, r3inv, r4inv, r5inv, r6inv, r7inv, r8inv, r9inv) by Stories enable us to analyze and compare property values within specific building structures, providing insights into the relationship between these variables based on the number of stories.

Below is the table providing the description and number of each variable I created.

Description of Variables	# Variable Created
Size Variables: Lot size, building area, and building volume variables. Obtained from original variables LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, STORIES	3
Price Ratio Variables: Each of the 3 value fields(FULLVAL, AVLAND, AVTOT) normalized by each of the 3 sizes (ltsize, bldsize, bldvol). This gives us 9 variables (r1,r2,r3,r4,r5,r6,r7,r8,r9). These variables are created to identify undervalued properties. Higher price ratio or lower ratio could suggest the property own is reporting inflated or deflated property values to reduce property taxes.	9
Inverse of Price Ratio Variables: 1 divided by the sum of each normalized 9 variables and epsilon (0.01). This gives us 9 inverse variables: r1inv,r2inv,r3inv,r4inv,r5inv,r6inv,r7inv,r8inv,r9inv. Inverse of Price Ratio Variables are useful because it serves as a single variable that capture the overall relationship between price and sizes. These inverse variables will detect properties that are overvalued and undervalued compared to other similar properties in the area. It can compare properties across neighborhoods, cities, and regions.	9
Ratio by Zip Variables: Grouped averages of 18 variables(r1, r2, r3, r4, r5, r6, r7, r8, r9, r1inv, r2inv, r3inv, r4inv, r5inv, r6inv, r7inv, r8inv, r9inv) by Zip Codes. Grouping variables by zipcodes allow us to assess	18

property values within a particular geographic area. This helps us to identify anomalies that are not apparent on the individual level.	
Ratio by Tax Class Variables: Grouped averages of 18 variables(r1,r2,r3,r4,r5,r6,r7,r8,r9, r1inv,r2inv,r3inv,r4inv,r5inv,r6inv,r7inv,r8inv,r9inv) by Tax Class. Ratio by Tax Class Variables help us compare different properties in the same tax class and help us detect which properties are overvalued or undervalued in that tax class. It also helps to compare similar property values across different tax class. Lastly, Ratio by Zip variables helps monitoring changes in property values over time for a tax class.	18
Value Ratio Variable: FULLVAL divided by the sum of AVLAND and AVTOT. Value Ratio Variable helps estimate the accuracy of assessed property value. It compares the market value to the actual land value and actual total value. If the Value Ratio variable is lower than expected, it may indicate that property owners try to undervalue the property values to result in lower tax payments.	1
Ratio by Stories Variables (New Variable): Grouped averages of 18 variables(r1,r2,r3,r4,r5,r6,r7,r8,r9, r1inv,r2inv,r3inv,r4inv,r5inv,r6inv,r7inv,r8inv,r9inv) by Stories. Ratio by stories variables helps us compare different property values and their relationship within each particular stories of building structures.	18

V. Dimensionality Reduction

Performing dimensionality reduction on the NY property data variables help preventing curse of dimensionality, which is the challenge that arises when working with high-dimensional data. For unsupervised fraud detection model on NY property data, we firstly do z-scaling on all the variables. Z-scaling is important before performing PCA (Principal Component Analysis) because it equalizes variable scales, mitigates the impact of outliers, and preserves relative distances. Standardizing the variables we have in the dataset makes sure all variables have equal contribution, prevent outliers that could potentially skew results, and improves the effectiveness of PCA.

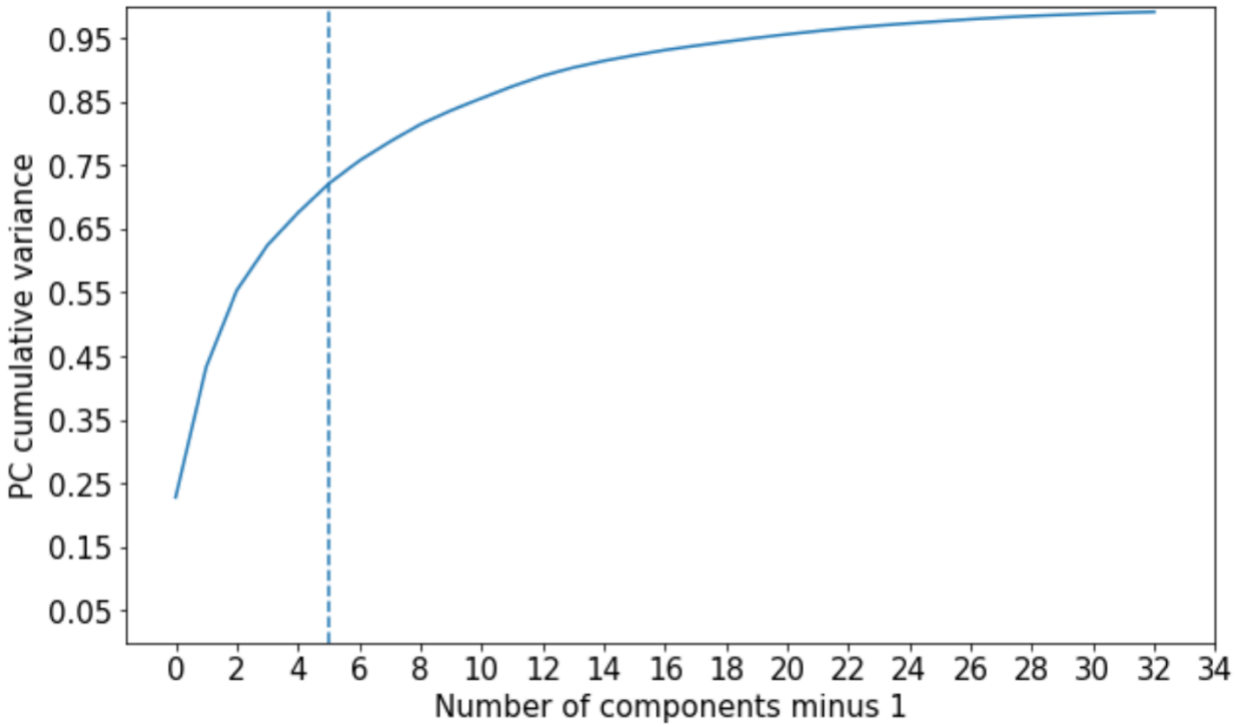
After z-scaling all the variables, I performed PCA on the variables, which is an unsupervised linear method that reduces dimensions and remove linear correlations. I set the PCA parameter to be

n_components = .99, svd_solver = 'full',

which produces a cumulative variance plot shown below showing how much of the total variance in the data we can retain by keeping some number of PCS. Based on the graph, I decide keep the

first **5 PCs** to keep about **70%** of all variances. This is because that we want to z-scale the variables again to maintain consistent scaling and interpretability. We want all the dimensions to be equally important for the Minkowski Distance.

Cumulative Variance Plot



VI. Anomaly Detection Algorithms

There are two score methods used to build anomaly detection algorithms and detect outliers and unusual records.

The first score method is **zscores outlier** to scale and center all the dimensional, then distance to the origin. The formula of zscore outlier involves taking the absolute values of the z-scaled data raised to the power of "p1," summing them row-wise, and then raising the sum to the reciprocal of "p1." p1 denotes the power of the Minkowski distances, ranging from 1 to 4. Higher scores indicate a higher likelihood of fraud. We detected that the record with the highest score has a score of **690.0544**.

The second score method is **autoencoder**, which is nontrivial functional mapping of a record back to itself. If a record has large error, it indicates an unusual record. Autoencoder is calculated as the difference (error) between the original input vector and the model output vector, giving us

the fraud score. An autoencoder is the error or Minkowski distance between the original record and the output of encoder. I chose Neural Network as the nonlinear model for autoencoder.

Finally, I combined the first score from zscores outlier and second score autoencoder by first doing the rank order replacement for each of them and take the average of these two scores as the final score.

VI. Results

Changes in choices from baseline	Top 100 (.01%)	Top 1,000 (.1%)	Top 10,000 (1%)
baseline: p1 = 2, p2 = 2, 4PCs	-	-	-
p1 = 1, p2 = 4	**92**	**90.4**	**96.57**
3PCs	**81**	**64.9**	**53.16**
p1 = 3	**94**	**97.7**	**98.68**
p1 = 1	**93**	**96.3**	**97.67**
p2 = 3	**96**	**96.6**	**99.03**
p2 = 4	**95**	**96.2**	**98.57**
5PCs	**86**	**77.4**	**69.93**
don't zscale PCs	**88**	**84.7**	**87.25**
p1 = 1, p2 = 3	**92**	**77.6**	**97.00**
2PCs	**61**	**67.0**	**61.47**
Neural Network (hidden_layer_sizes=(4,4), max_iter = 40)	**92**	**95.6**	**95.91**
Neural Network (hidden_layer_sizes = (2,2), max_iter = 20)	**80**	**75.7**	**83.35**
p1 = 4	**93**	**94.2**	**97.84**
p2 = 1	**94**	**93.6**	**97.69**

High Level Observations

1. The first observation after tuning the parameters from the baseline is that when changing P1 and P2 (Minkowski Distance) values does not have a significant difference on the percent common score. After I changed the values of P1 and P2 in various experiment within range from 1 to 4, the percent common score stays very high within 95 to 100 for Top 10,000 records. For top 1,000 records and top 100 records, the score also fluctuates from 90 to 100.

2.The second observation is that top records are very sensitive to the change in PCs. When I changed PC value from baseline value 4 to 2,3, and 5, the percent common score drops

significantly. The lowest one is PC = 3, which the score for Top 10,000 records dropped to 53.16. For PC = 2, the score also dropped significantly to 60s.

3. The third observation is that top records has moderate sensitivity to whether we zscale PCs. In the experiment, I chose to not zscale PCs and observed how the Top 100, 1000, and 10000 records react. It turned out that the percent common score fluctuate in 80s, which is relatively stable.

4. I also tried to tune the parameters for autoencoders (Neural Network) by changing the hidden_layer_sizes and max_iter values. After I tune the default hidden_layer_sizes = 3 to (2,2) and (4,4) and max_iter = 40 and 20, I saw that the top records respond moderately to these changes in parameters. For hidden_layer_sizes = (4,4) and max_iter = 40, the percent common score stays in 90s, and for hidden_layer_sizes = (2,2) and max_iter = 20, the percent common score stays around 80.

IX. Summary

In summary, to build an anomaly detection algorithm on the tax fraud for the NY property, it took stages beginning with providing statistical summary of each of **32** fields with **1,070,994** rows in the given dataset. Presenting a statistical summary for the categorical and numerical variables and visualizing them in form of bar chart and histograms is an essential step before starting the model building because it helps me to understand the distribution of each variable in the dataset, which can provide insights into the nature of the data. Data cleaning helps to identify data quality issues. By removing exclusions, such as government properties, the dataset becomes more focused on the specific target of private owners committing tax fraud. Filling in missing values with appropriate substitutes enhances the integrity of the dataset, allowing for more accurate analysis and insights. Replacing invalid values and outliers with NA or appropriate averages ensures consistency and avoids distorting statistical measures.

Following data cleaning is the variable creation

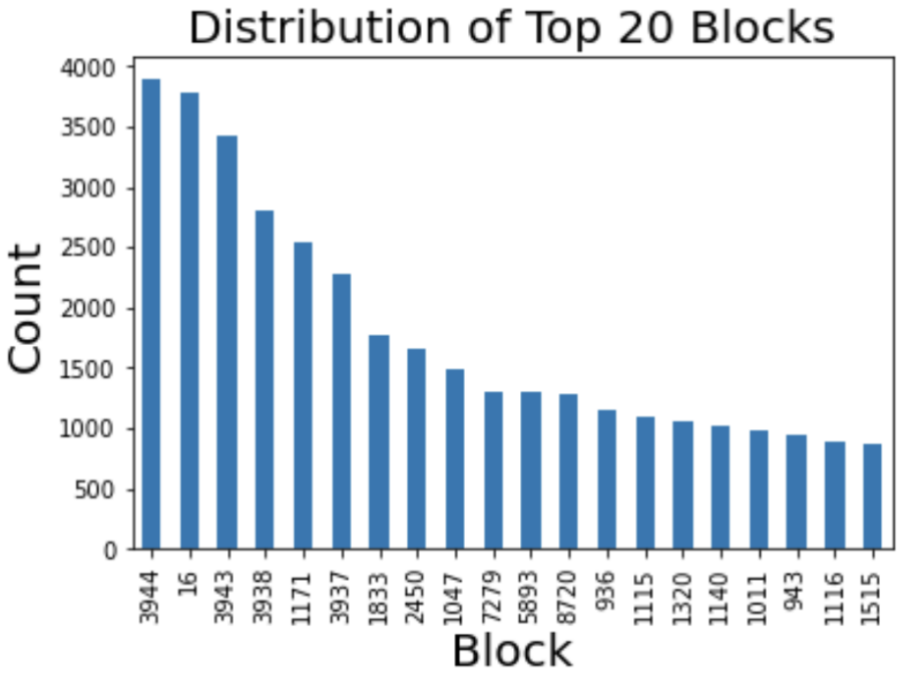
X. Appendix

(1) Field Name: **RECORD**
Description: Record field is categorical and tracks the number of rows/records of NY property data in the dataset, using ordinal unique positive number from **1 to 1,070,994**.

(2) Field Name: **BBLE**
Description: BBLE field is categorical and indicate file key values. Key BBLE = Boro, Block, Lot, and Easement code. I didn't create a graph to visualize this field because every value is unique.

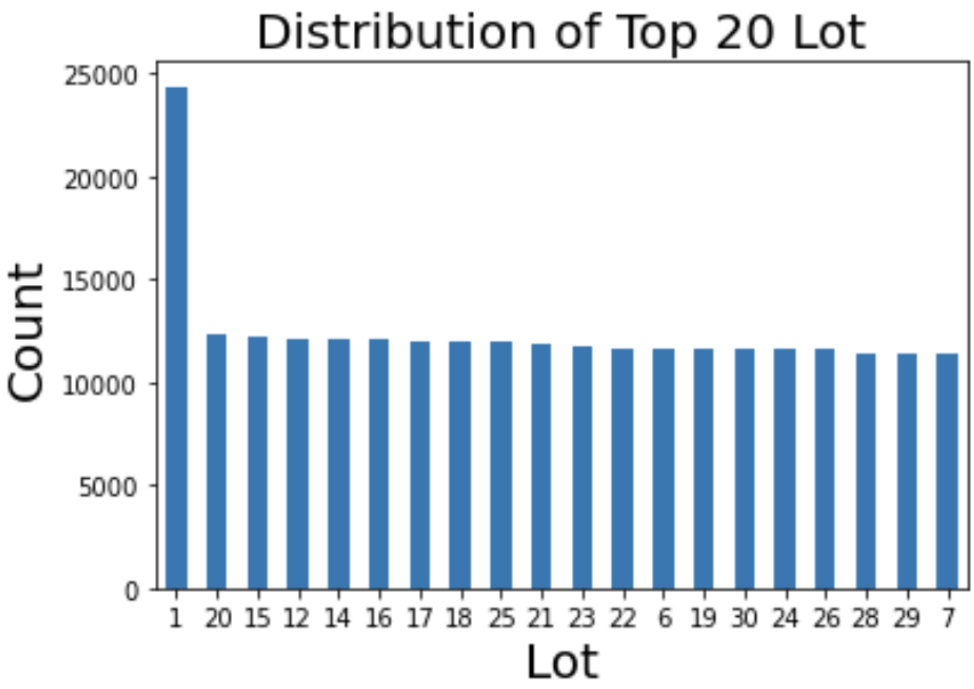
(3) Field Name: **BLOCK**

Description: BLOCK field is categorical and means the valid block range by BORO. When BORO is Manhattan, Block ranges from 1 to 2255. When BORO is Bronx, block ranges from 2260 to 5958. When BORO is Brooklyn, block ranges from 1 to 8955. When BORO is Queens, block ranges from 1 to 16350. When BORO is Staten Island, block ranges from 1 to 8050.



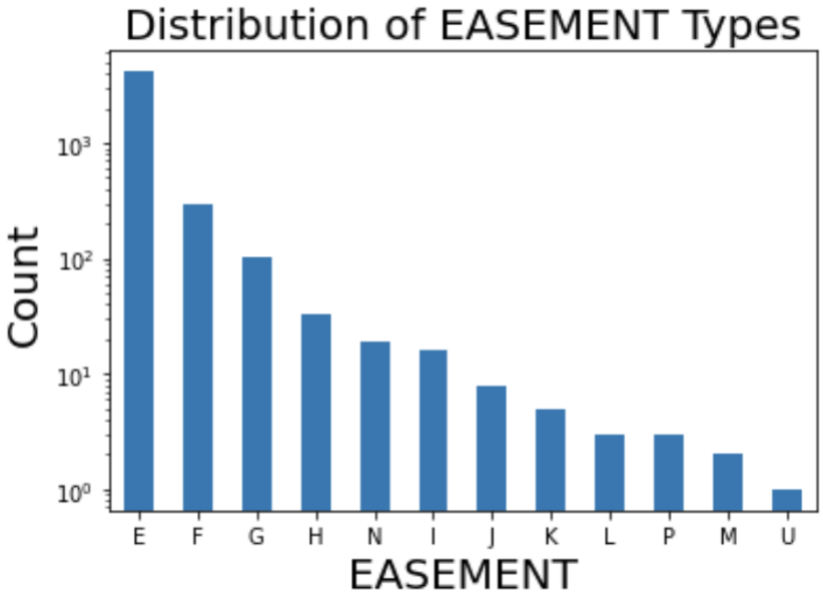
(4) Field Name: **LOT**

Description: Lot field of the NY property data. The value with the highest frequency is **1**, which has a total count of **24367**.



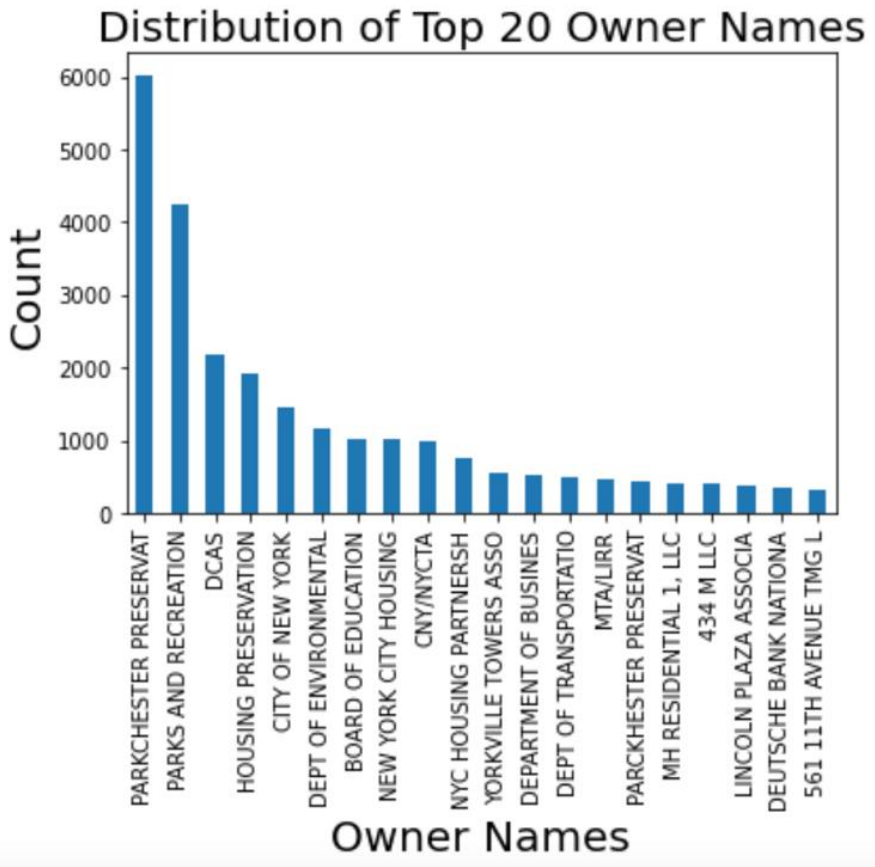
(5) Field Name: **EASEMENT**

Description: Easement field. It is a categorical field. Below displays a bar chart of the top 20 values. The most common easement is E with a total count of **4148**. When easement is bank, it means there is no easement for a property. Easement values meaning: 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



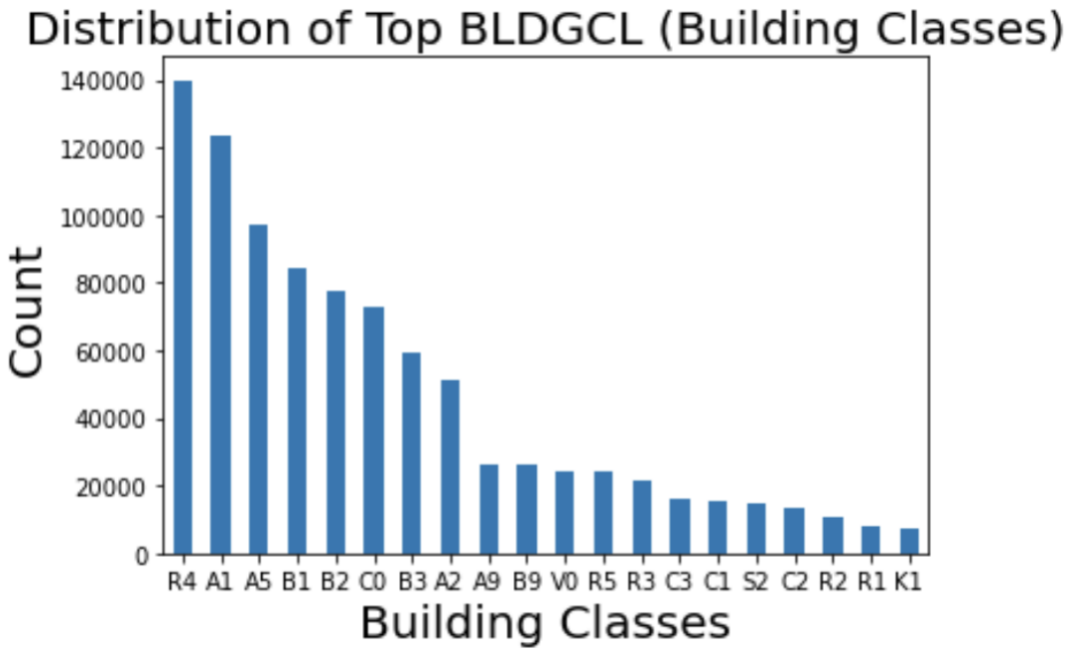
(6) Field Name: **OWNER**

Description: Owner Names field of the NY property data. The owner name with the highest frequency is **Parkchester Preservat**, which appeared **6021** times.



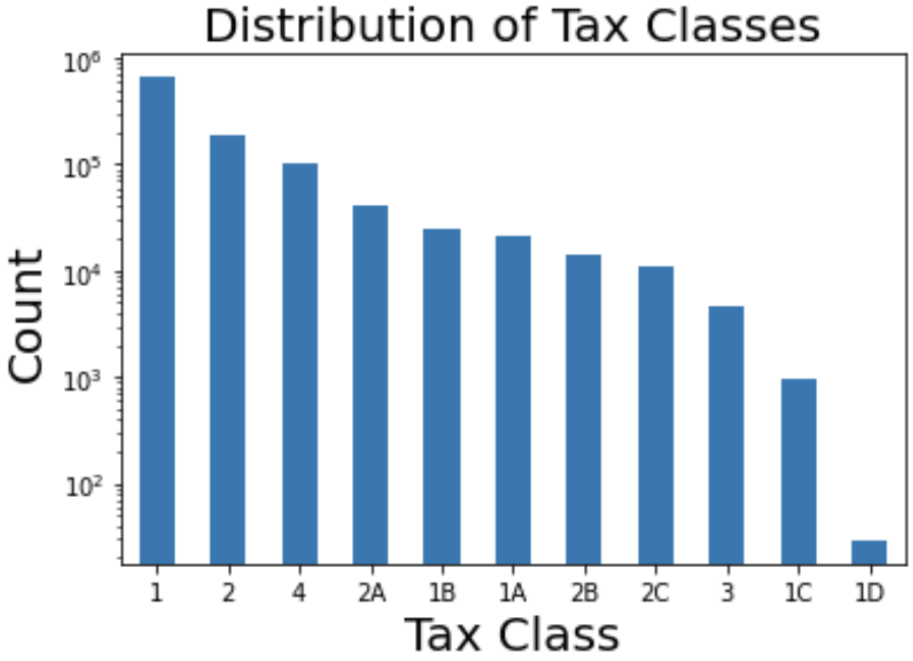
(7) Field Name: **BLDGCL**

Description: Building Class field, which is categorical. The building class with the highest frequency is **R4**, which appeared **139879** times.



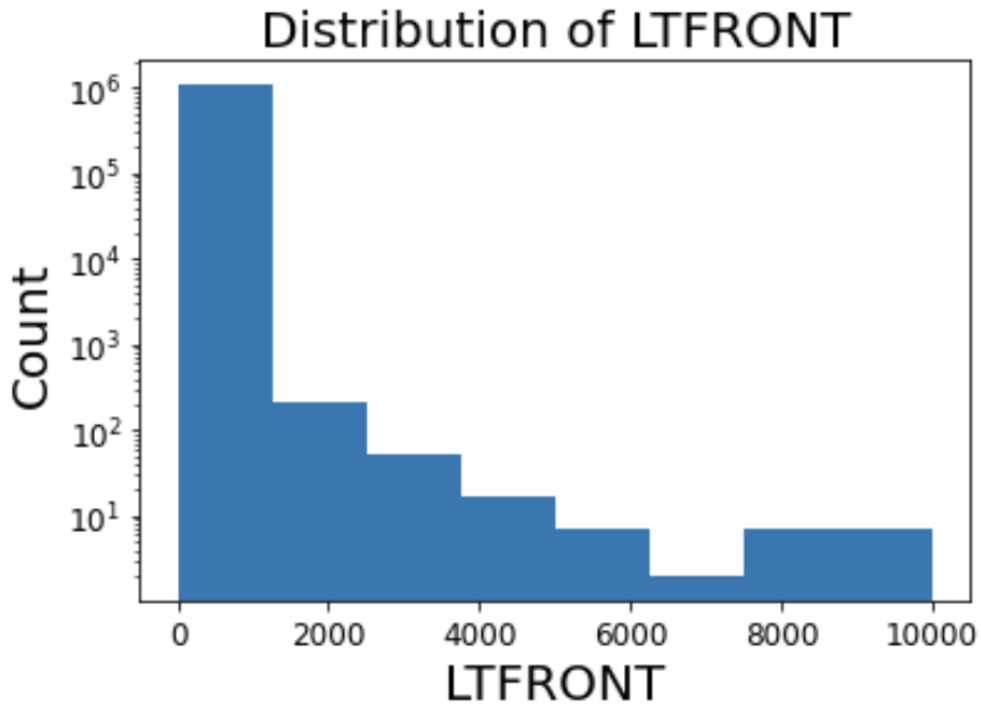
(8) Field Name: **TAXCLASS**

Description: Tax class field of the NY property data, which is categorical. The type of tax class with the highest frequency is **1** and it appeared **660721** times. 1 means 1 - 3 Unit Residence, 2 means Apartments, 2A means 4, 5, or 6 Units, 3 means Utilities, 4 means All Others.

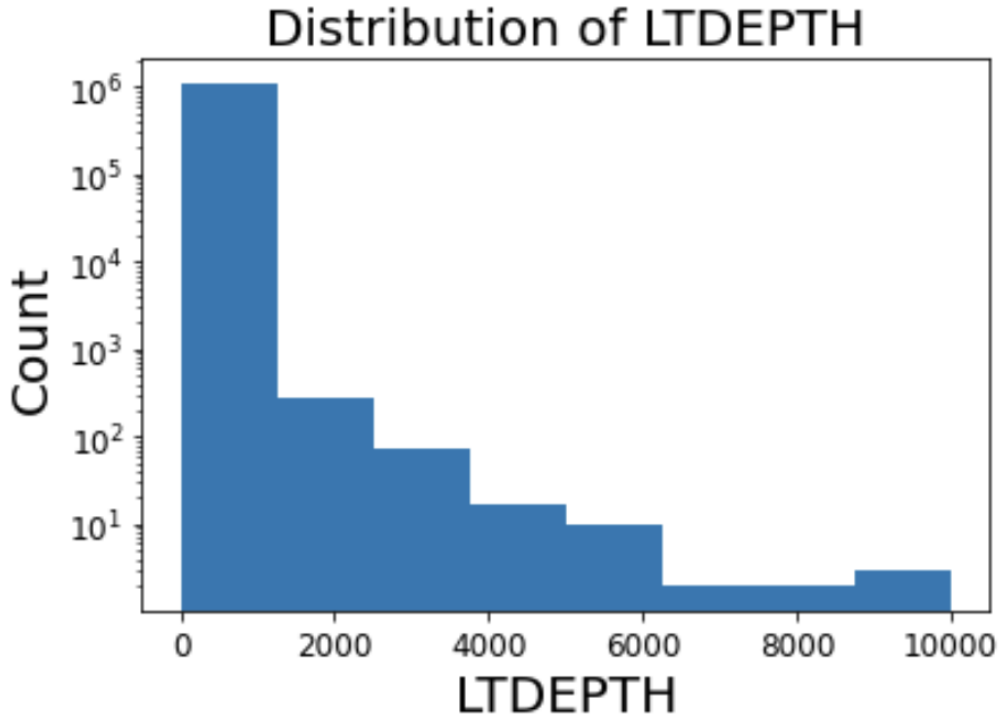


(9) Field Name: **LTFRONT**

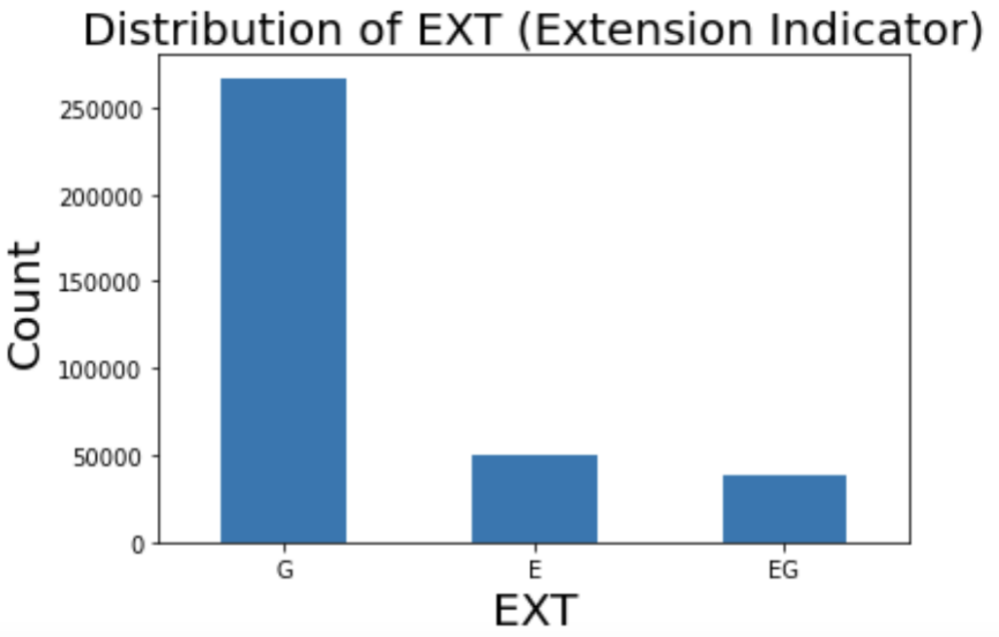
Description: LTFRONT is the lot width field of the NY property data. The most frequent lot width is **0**, which appeared **169108** times.



(10) Field Name: **LTDEPTH**
Description: Lot depth field of the NY property data.

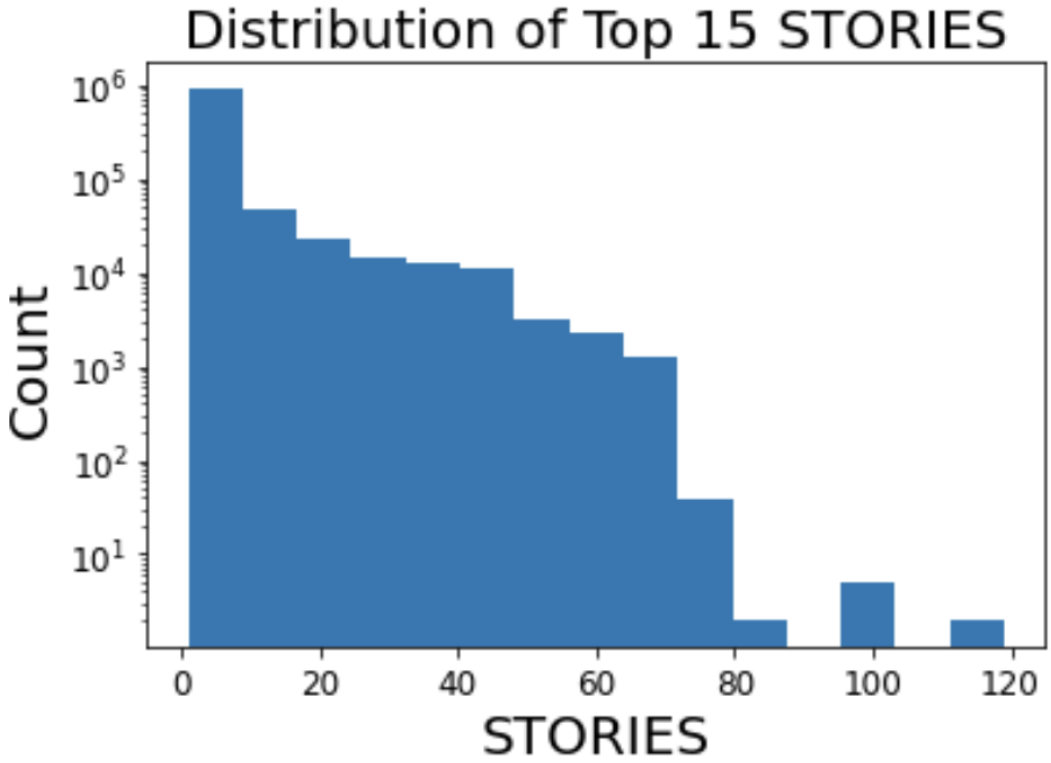


(11) Field Name: **EXT**
Description: EXT is the Extension Indicator field of the NY property data, which is categorical.



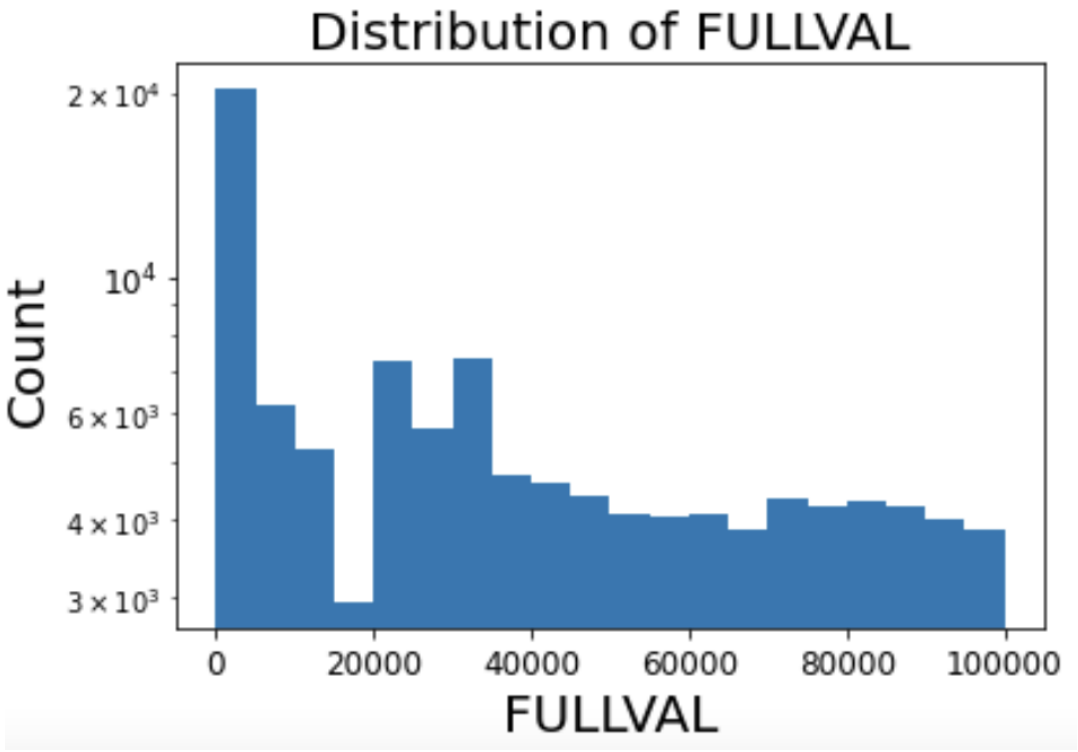
(12) Field Name: **STORIES**

Description: STORIES field is the number of stories in building, which is numerical. Below is a histogram of STORIES field.



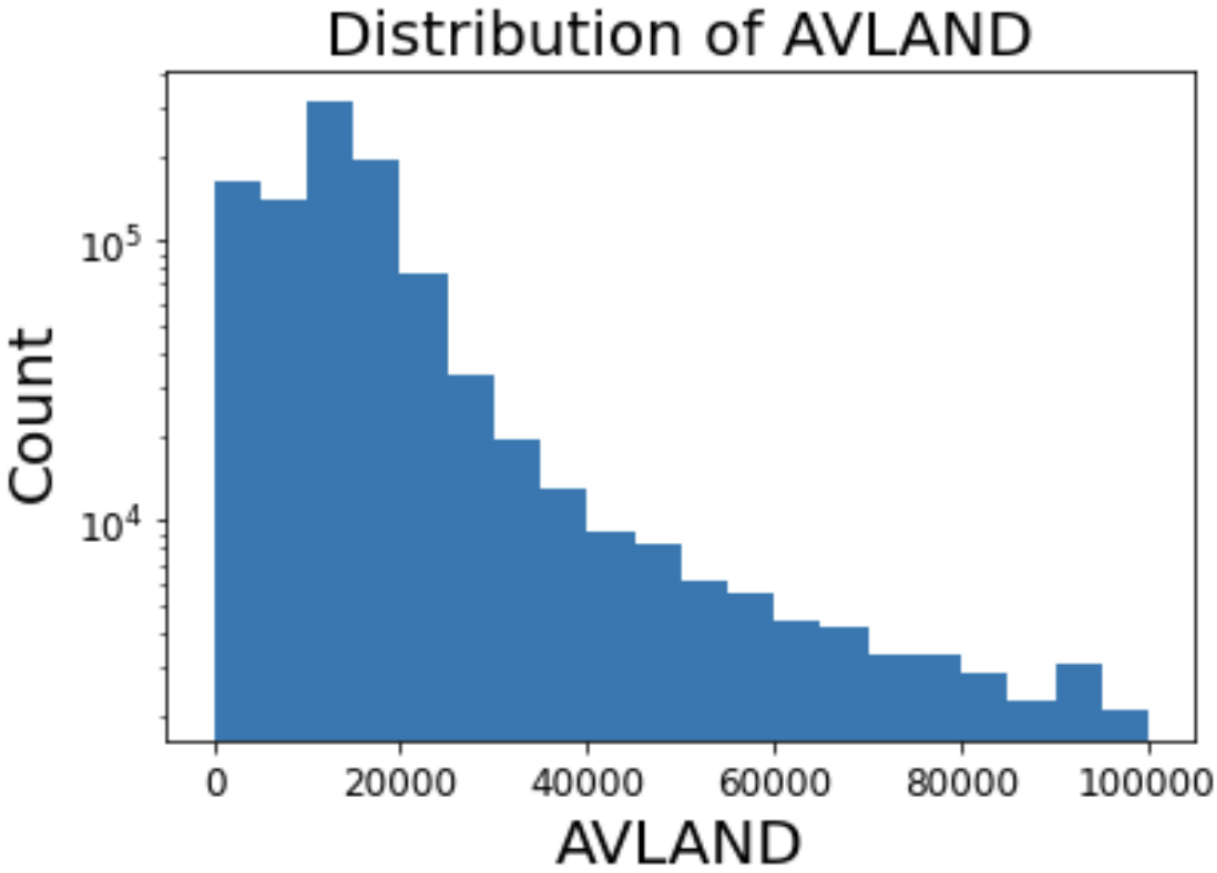
(13) Field Name: **FULLVAL**

Description: FULLVAL is the market value field of the NY property data, which is numerical.

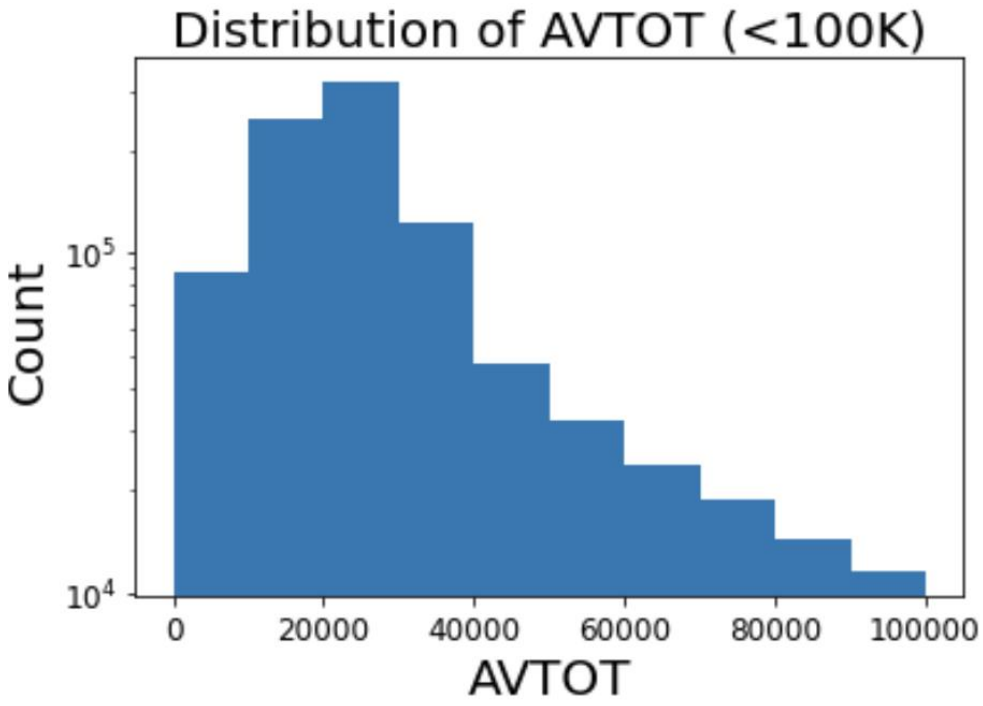


(14) Field Name: **AVLAND**

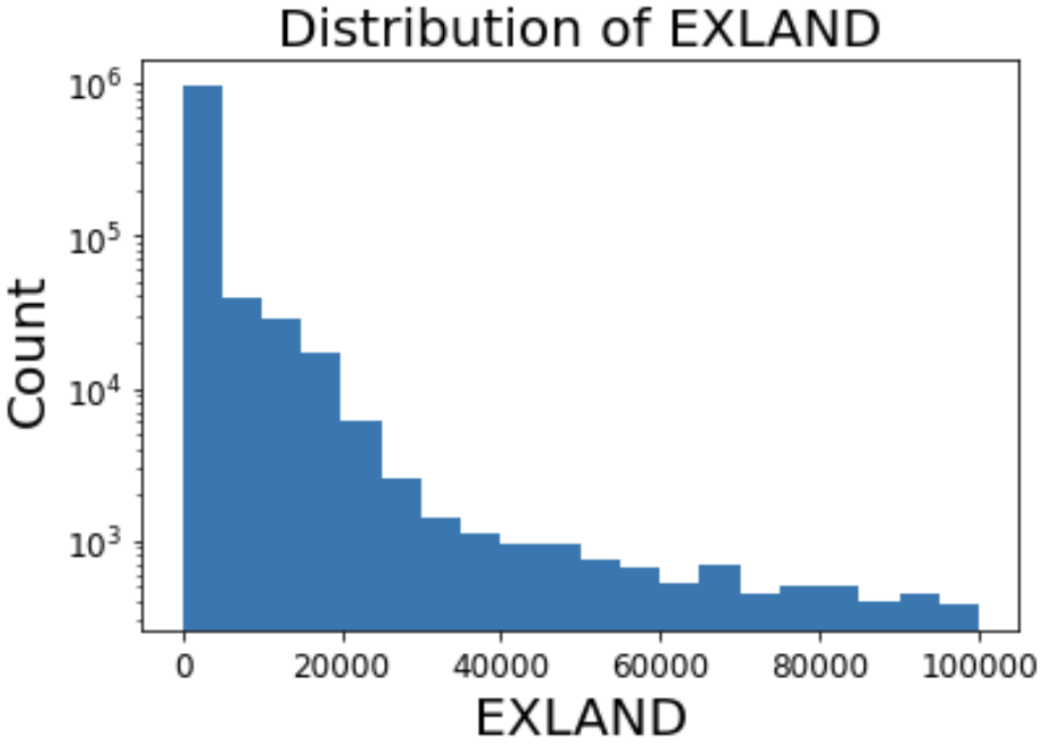
Description: AVLAND is the Actual Land Value field, which is numerical.



(15) Field Name: **AVTOT**
Description: AVTOT is the Actual Total Value field, which is numerical.

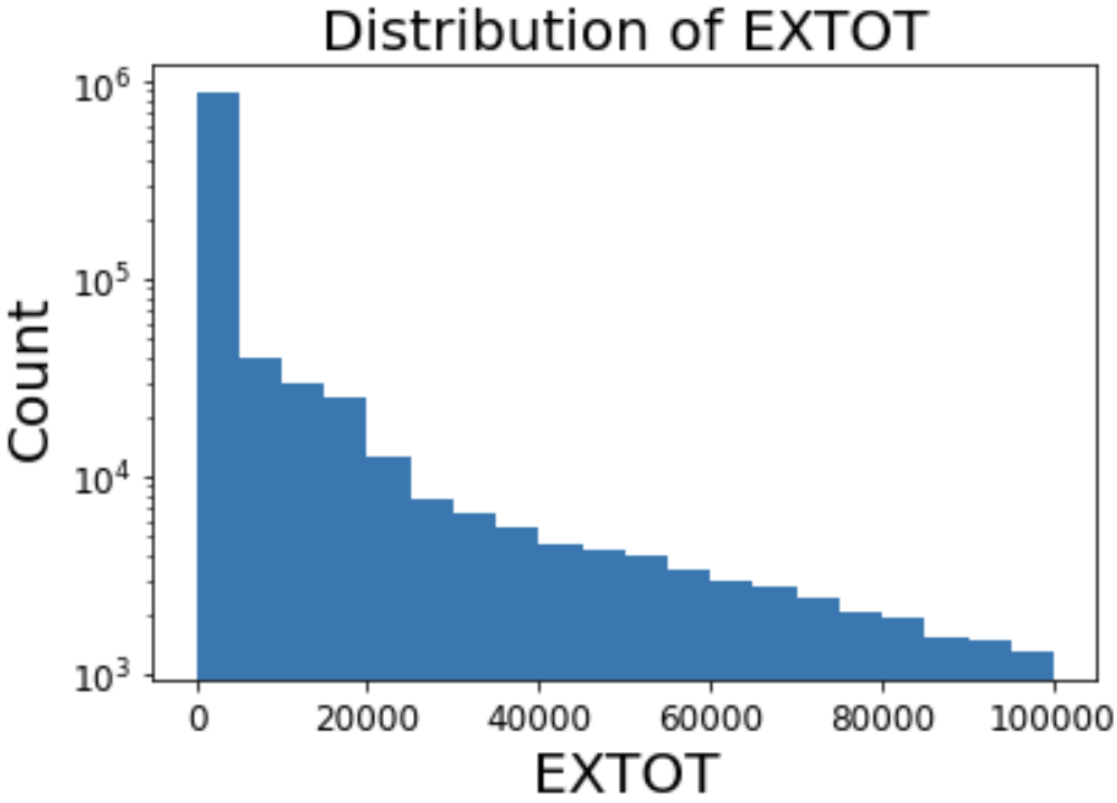


(16) Field Name: **EXLAND**
Description: Actual Exempt Land Value field, which is numerical.



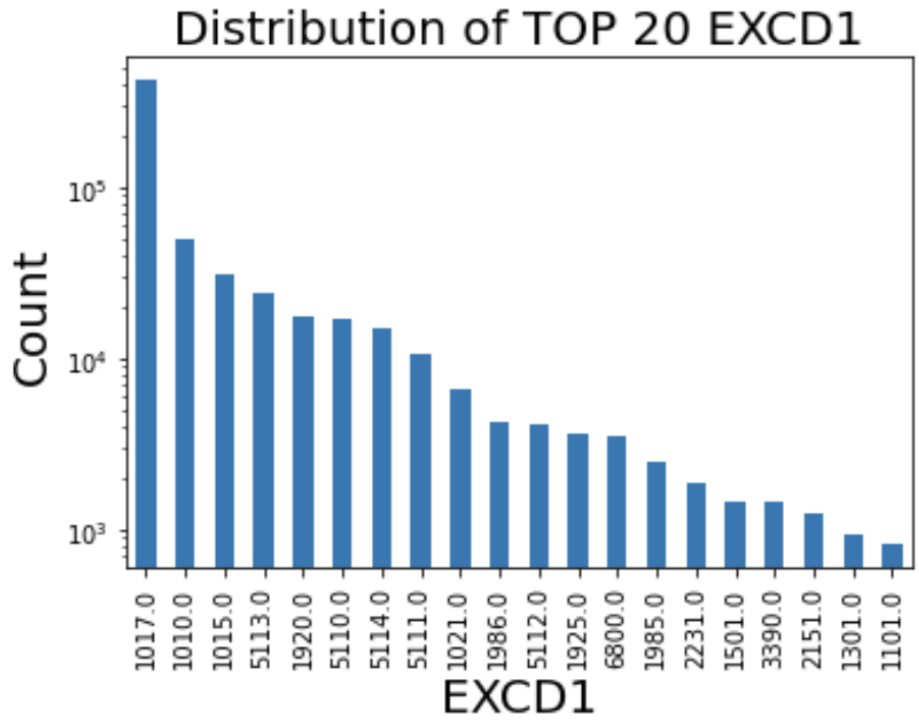
(17) Field Name: **EXTOT**

Description: Actual Exempt Land Total field, which is numerical.

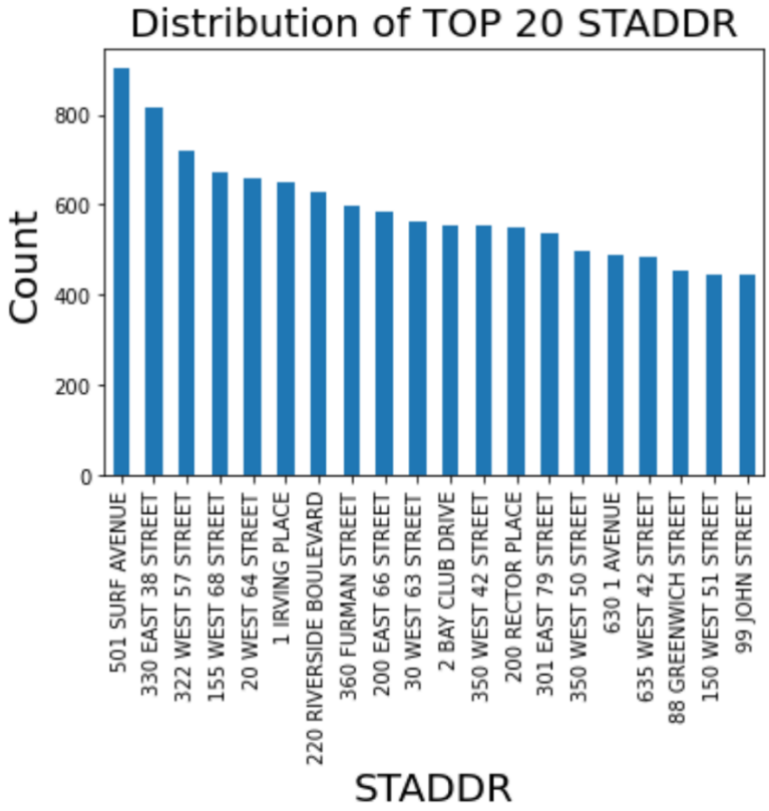


(18) Field Name: **EXCD1**

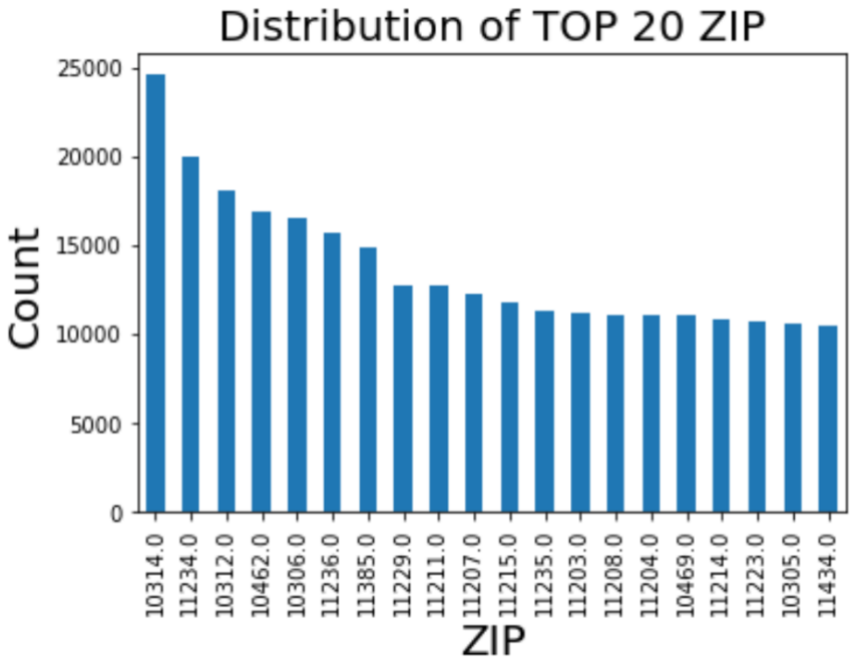
Description: EXCD1 is the Exemption Code 1 field, which is categorical.



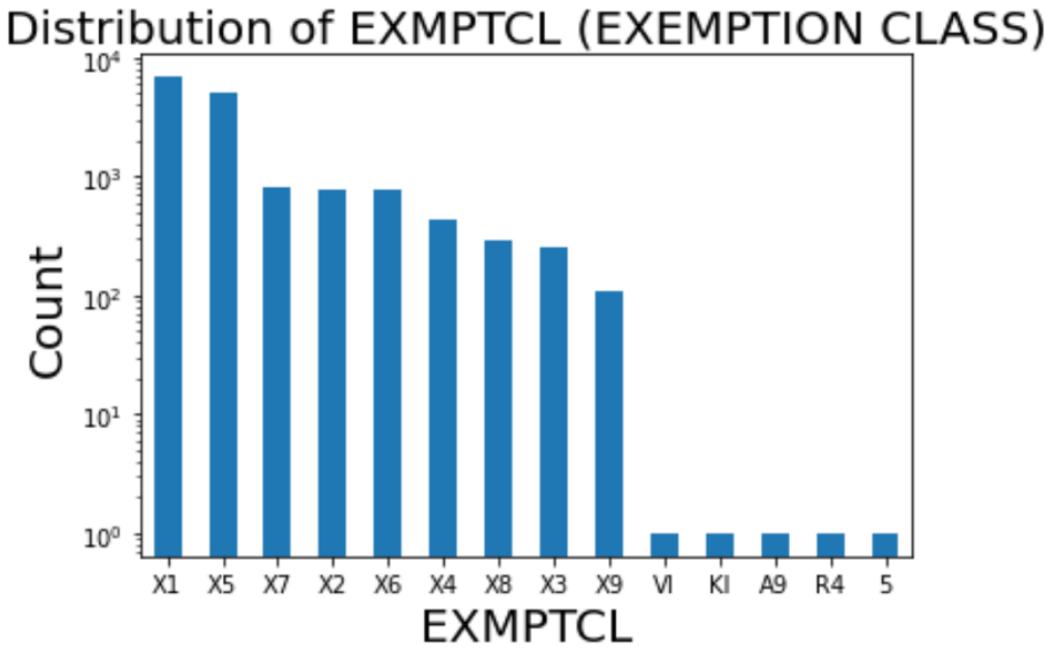
(19) Field Name: **STADDR**
Description: Street Address field, which is categorical.



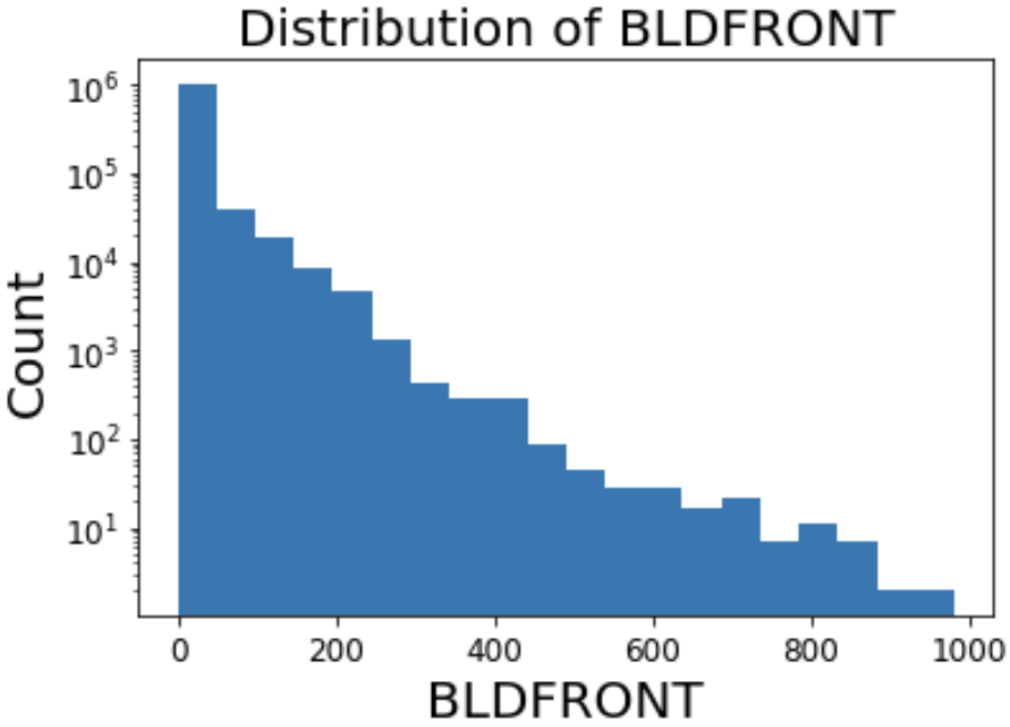
(20) Field Name: **ZIP**
Description: ZIP is the zip code field, which is categorical.



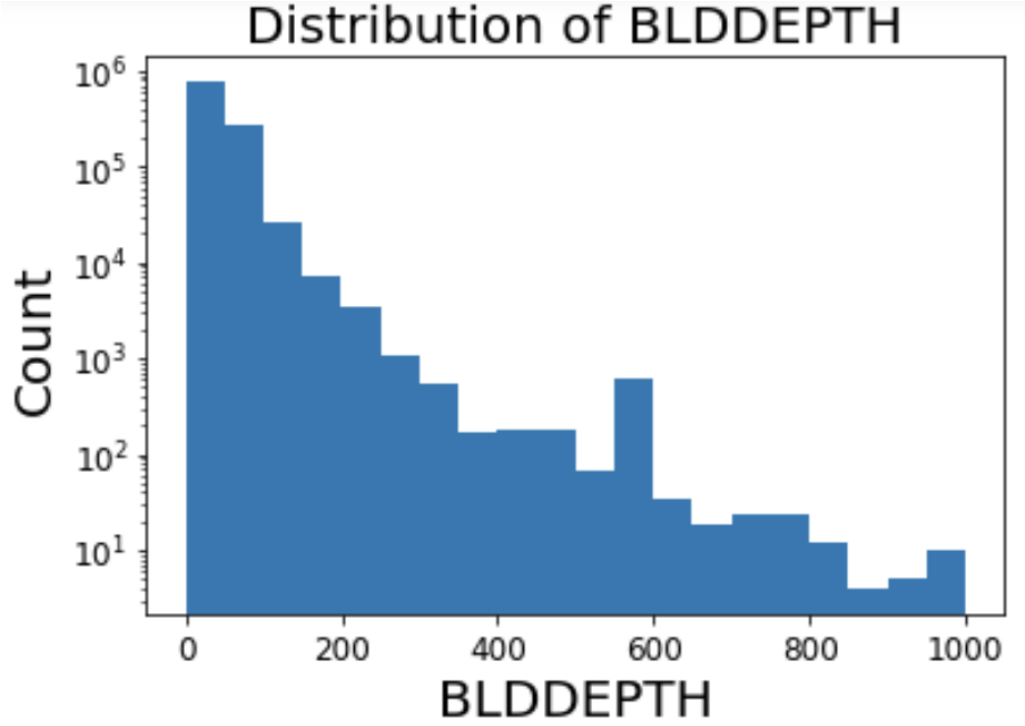
(21) Field Name: **EXMPTCL**
Description: EXMPTCL is the Exemption class field, which is categorical.



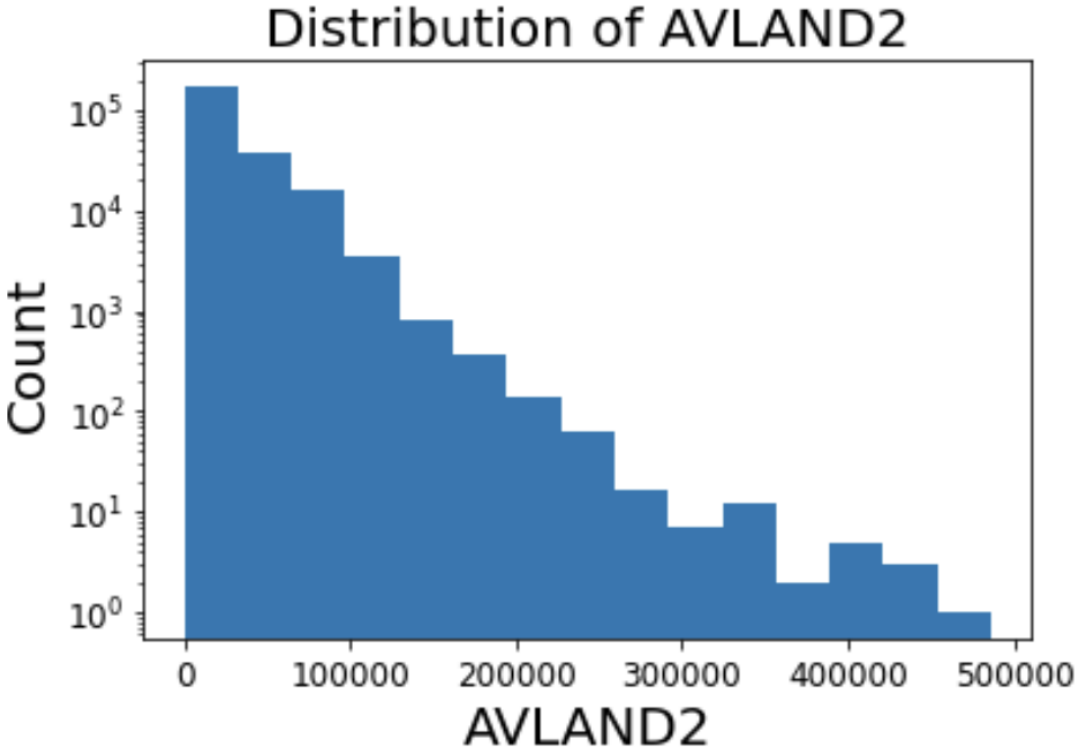
(22) Field Name: **BLDFRONT**
Description: BLDFRONT is the building width field



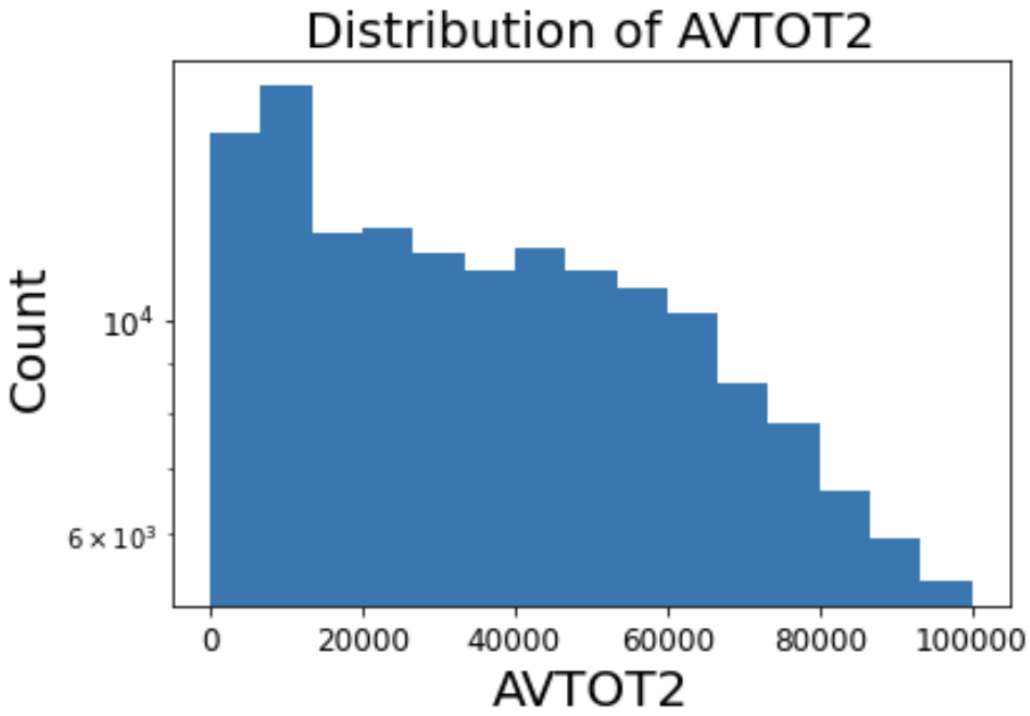
(23) Field Name: **BLDDEPTH**
Description: BLDDEPTH is the building depth field.



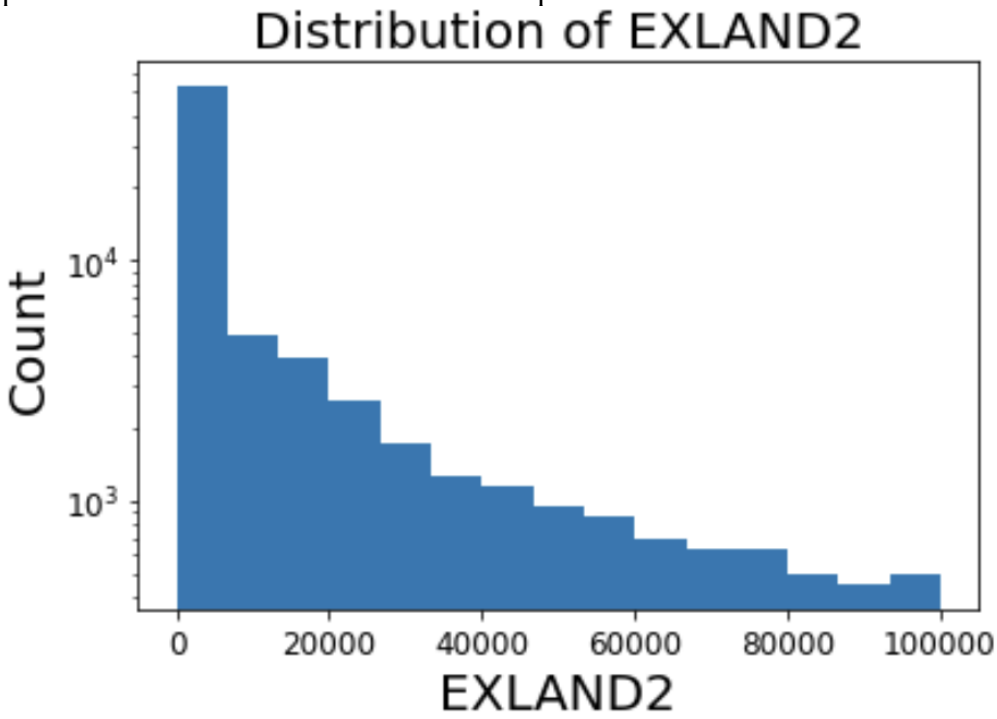
(24) Field Name: **AVLAND2**
Description: AVLAND2 is the Transitional Land Value field.



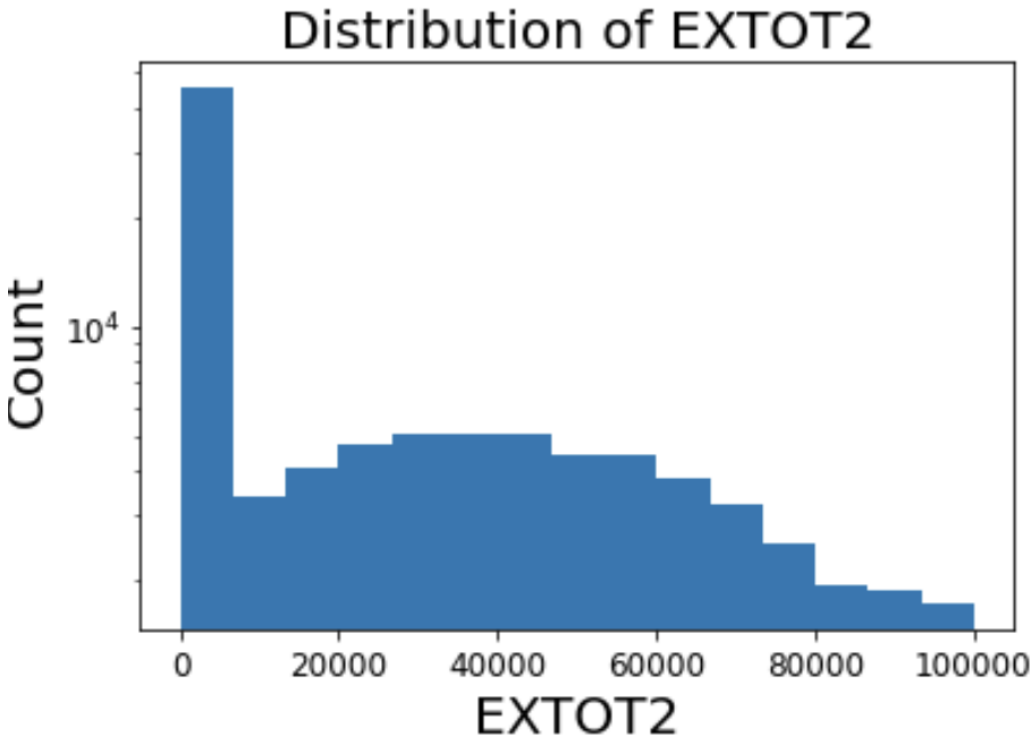
(25) Field Name: **AVTOT2**
Description: AVTOT2 is the transitional total value field.



(26) Field Name: **EXLAND2**
Description: EXLAND2 is the transitional exemption land field.

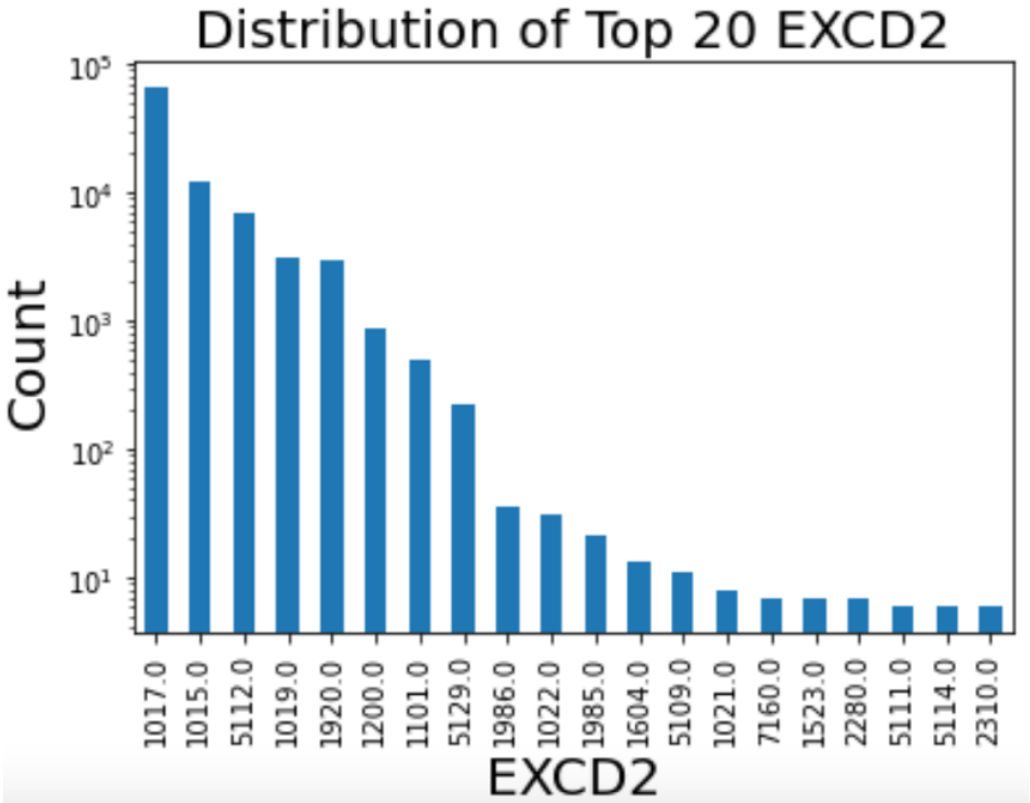


(27) Field Name: **EXTOT2**
Description: EXTOT2 is the transitional exemption land total field



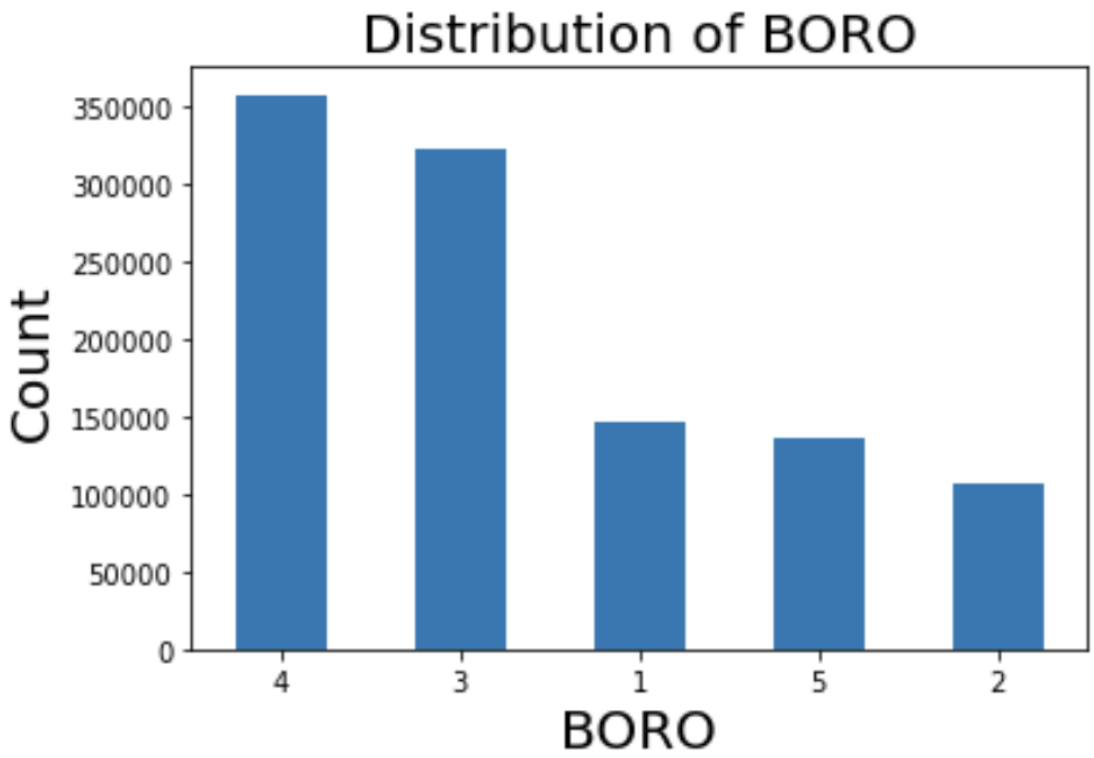
(28) Field Name: **EXCD2**

Description: EXCD2 is the exemption code 2 field.



(29) Field Name: **BORO**

Description: BORO is the borough field, which is categorical. 1=Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island



(30) Field Name: **PERIOD**

Description: Period is the assessment period field. There is no graph to visualize this field because it only has one type of value, which is **'FINAL'** and appeared **1070994** times.

(31) Field Name: **YEAR**

Description: Year is the assessment year field. There is no graph to visualize this field because it only has one type of value, which is **'2010/11'** and appeared **1070994** times.

(32) Field Name: **VALTYPE**

Description: Valtype is the market type field. There is no graph to visualize this field because it only has one type of value, which is **'AC-TR'** and appeared **1070994** times.