Appendix

I. Executive Summary

The aim of this report is to provide an overview of the machine learning model developed to detect data fraud in New York Real Estate Dealing. With the increasing amount of real estate dealing, the records' input accuracy becomes something crucial. The risk of fraudulent data can harm the record and make it unusable for future use. In response, the use of machine learning algorithms has become an effective approach to detect and prevent fraud data in real-time. The model is trained on a large dataset of real estate deal transactions containing both fraudulent and non-fraudulent records. The dataset is preprocessed to remove any outliers, and fill in missing data, resulting in a more precise dataset for modeling. We then created relevant variable for modeling use. We used a combination of Z-scale normalization and principal component analysis (PCA) to reduce the dimensionality of our dataset. In the end we are able to use heatmap to identify abnormality in records, tested the accuracy and proven the model is accurate and great for reference use, which can potentially save businesses and consumers from financial losses.

II. Data Description

The dataset consists of **1,070,994 data records** of property valuation and assessment data collected by the city government department of finance, comprising **32 independent fields (14 numeric, 18 categorical)**.

1. Summary Tables

(1) Categorical Table

Field Name	# Records Have Values	% Populated	# Zeros	# Blanks	# Unique Values	Most Common Value
RECORD	1070994	100.00%	0	0	1070994	1
BBLE	1070994	100.00%	0	0	1070994	1000010101
BORO	1070994	100.00%	0	0	5	4
BLOCK	1070994	100.00%	0	0	13984	3944
LOT	1070994	100.00%	0	0	6366	1
EASEMENT	4636	0.43%	0	1066358	12	E
OWNER	1039249	97.04%	0	31745	863347	PARKCHESTER PRESERVAT
BLDGCL	1070994	100.00%	0	0	200	R4
TAXCLASS	1070994	100.00%	0	0	11	1
EXT	354305	33.08%	0	716689	3	G
EXCD1	638488	59.62%	0	432506	129	1017
STADDR	1070318	99.94%	0	676	839280	501 SURF AVENUE
ZIP	1041104	97.21%	0	29890	196	10314
EXMPTCL	15579	1.45%	0	1055415	14	X1
EXCD2	92948	8.68%	0	978046	60	1017
PERIOD	1070994	100.00%	0	0	1	FINAL
YEAR	1070994	100.00%	0	0	1	2010/11
VALTYPE	1070994	100.00%	0	0	1	AC-TR

(2) Numerical Table

Field Name	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Stdev	Most Common Value
LTFRONT	1070994	100.00%	169108	0	9999	36.635301	74.03284	0
LTDEPTH	1070994	100.00%	170128	0	9999	88.861594	76.39628	100
STORIES	1014730	94.75%	0	1	119	5.006918	8.365707	2
FULLVAL	1070994	100.00%	13007	0	6150000000	874264.5054	11582430	0
AVLAND	1070994	100.00%	13009	0	2668500000	85067.91867	4057260	0
AVTOT	1070994	100.00%	13007	0	4668309000	227238.1687	6877529	0
EXLAND	1070994	100.00%	491699	0	2668500000	36423.89069	3981576	0
EXTOT	1070994	100.00%	432572	0	4668309000	91186.98168	6508403	0
BLDFRONT	1070994	100.00%	228815	0	7575	23.04277	35.5797	0
BLDDEPTH	1070994	100.00%	228853	0	9393	39.922836	42.70715	0
AVLAND2	282726	26.40%	0	3	2371005000	246235.7193	6178963	2408
AVTOT2	282732	26.40%	0	3	4501180000	713911.4362	11652530	750
EXLAND2	87449	8.17%	0	1	2371005000	351235.6843	10802210	2090
EXTOT2	130828	12.22%	0	7	4501180000	656768.2819	16072510	2090

III. Data Cleaning

Remove Irrelevant Records:

In order to cleanse the data, our initial step involved eliminating cert ain records that were irrelevant to our analysis. We removed 24,478 data records. These included benign properties and government-owned propertie s, which were not pertinent to the changes we sought to examine. We iden tified the most frequently occurring property owners and disregarded the top 20 owners from our list. Additionally, we removed extra owners for e xclusion. As a result, the final dataset contained a reduced number of r ecords compared to the original dataset.

Fill in missing ZIP:

We identified 21,537 missing zip records. Because the data is sorted by zip, when zip is missing, before and after zips are the same, we filled in the zip with that value. With that we have 10,114 zip records to fill in. We then filled in these zips with the previvious record's zip.

AVTOT, AVLAND, FULLVAL:
We grouped the value by taxclass, applied statistical smoothing on these columns. We replaced 'NaN' with np.nan to properly represent missing values. Calculated the avages and filled in the missing value with the calculated value.

STORIES:
There're 43,684 missing values in STORIES, we also applied the statistical smoothing method to stories. Calculated the mean stories for each tax class group and computed the na value with the mean value.

LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT:
Because these 4 fileds to not have NAs, we just replaced the 0s. Replace 0 and 1's by NAs so they are not counted in calculating mean. Similar to the step with AVTOT, we calculated groupwise average. Used the statistical smoothing method and replaced 0s with mean values. In the end we converted ZIP to a string rather than a float. Now the data is ready for next step, the variable creation.

IV. Variable Creation

1. Imputation Logic

To improve the accuracy and interpretability of the data, a statistical smoothing technique was applied to each numeric field, while grouping the data by tax class. The specific numeric fields that were included in this process are AVTOT, AVLAND, FULLVAL, STORIES, LTFRONT, LTDEPTH, BLDDEPTH, and BLDFRONT. By grouping the data based on the tax class attribute and applying statistical smoothing, the dataset was refined to reduce noise, variability, and outliers, resulting in a more reliable and meaningful representation of the underlying data.

2. Description of variables:

Varibale Name	Variable Description	# Variables Created
size variables: ltsize, bldsie, bldvol	lot area: multiply 'LTFRONT' by 'LTDEPTH' ; building size: multiply 'BLDFRONT' by 'BLDDEPTH'; bulding volume: multiply 'bldsize' by 'STORIES'	3
Value ratios: r1-r9	Divide each of the 3 $ value fields in {FULLVAL, AVLAND, AVTOT} by each of the property size metrics in {lotsize, bldsize, bldvol}	9
Inverse value ratios : r1inv-r9inv	The inverse of the above value ratios	9
relative ratios grouped by ZIP and tax class: r1_zip5 - r9_zip5, r1inv_zip5-r9inv_zip5, r1_taxtclass-r9_taxclass,r1inv_taxclass-r9inv_taxclass	compute the average of 9 value ratios variables (r1-r9) and inverse value ratios (r1inv-r9inv) by grouping by {ZIP,tax class},and then we applied statistical smoothing on the averages of each group, and divide the above value ratios variables by the grouped averages	36
comparison value ratio	compare the 3$ value measures: : FULLVAL/(AVLAND+AVTOT)	1
exempt ratio:'exempt_ratio'	calculated by: data['EXLAND']/data['EXTOT']	1
smoothed exempt ratio by ZIP	calulated the average of exempt ratio grouping by zip, then applied statistical smoothing on the average value. then we divide the exempt ratio by the average of each zip	1
Total Number of Variables	**60**	

Entities: *BBLE, BORO, BLOCK, LOT, EASEMENT, OWNER, BLDGCL, TAXCLASS , LTFRONT, LTDEPTH , EXT, STORIES, FULLVAL, AVLAND , AVTOT , EXLAND, EXTOT, EXCD1, STADDR, ZIP, EXMPTCL, BLDFRONT, BLDDEPTH, AVLAND2, AVTOT2 , EXLAND2, EXTOT2, EXCD2, PERIOD, YEAR, VALTYPE*

a. Code improvement

To further reduce the variability and noise caused by different number of builiding in different group of zip code and taxclass, we applied statstistical smoothing on existing code to improve the credibility of the relative ratio grouped by zip and taxlclass.

b. New variables creation

Here is the detailed description of newly created variables:

exempt_ratio: calculated by EXAND/EXTOT

This feature aims to use the ratio of EXAND and EXTOT to detect abnormalities.Typically, the Actual Exempt Land Total is the total area of the land that is exempt from taxes, while the Actual Exempt Land Value is the value of that exempt land. These two factors should have a direct relationship with each other since the value of the exempt land should increase with its size.

However, if there is fraud involved, the relationship between Actual Exempt Land Total and Actual Exempt Land Value may not follow this expected pattern.

For instance, if a property owner claims an exemption for a larger area of land than they actually own or if the value of the exempt land is overestimated, the relationship between the Actual Exempt Land Total and Actual Exempt Land Value may be inconsistent with what is expected.

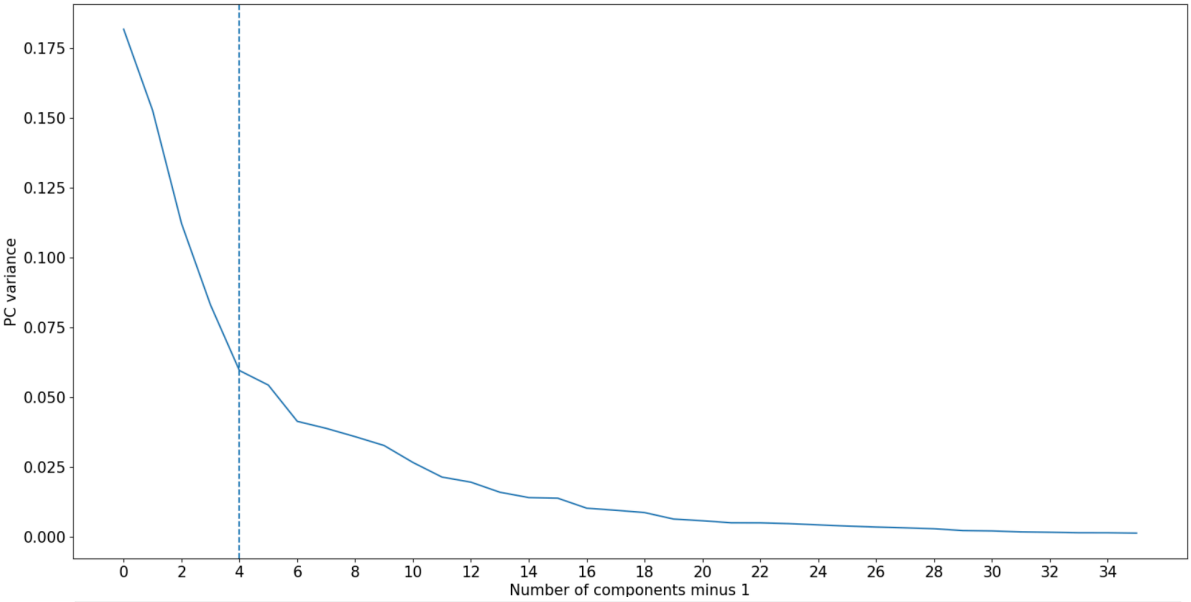
smoothed_exempt_ratio_zip5: Then we applied statistical smoothing on the average of each ZIP group, and divide the exempt_ratio variables by the grouped average. By doing this we analyzed how each variable varies across different geographic regions and make the variables more credible.

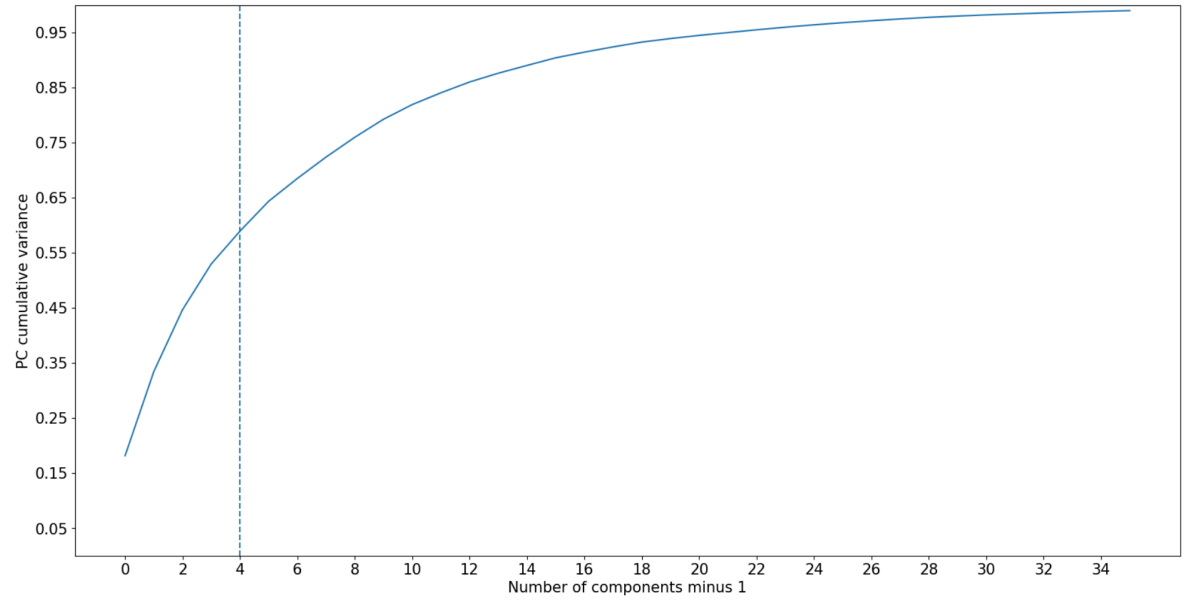
V. Dimensionality Reduction

We us a combination of Z-scale normalization and principal component analysis (PCA) to reduce the dimensionality of our dataset.

First, we apply Z-scale normalization to standardize the variables in our dataset to have zero mean and unit variance, noting that each dimension's scaled value is a measure of unusualness. This ensure that variables with extreme values and different ranges do not disturb our analysis and that all variables are equally important.

Next, we then use PCA to identify the most important components that explain the dataset. PCA help us identify patterns and correlations among variables and reduce the number of variables in our dataset while preserving the most important information. We select the top principal components that explain the most variance in the data.





For instance, from the graphs above, we select n_components = 4 as our base line model. The 4 principal components cover around 60% of the variance cum ulatively.

Finally, we apply Z-scale normalization again to the reduced dataset to tra nsform the columns into the form where standard deviation is 1 and mean is 0. This ensures that all variables are equally important and we therefore p repare the data for further analysis, such as calculating our fraud scores.

VI. Anomaly Detection Algorithms

a. Anomaly Detection Algorithms

In this section, we will discuss two score methods that can be used to detect anomalies, and explain how these scores can be combined to improve the accuracy of the detection process.

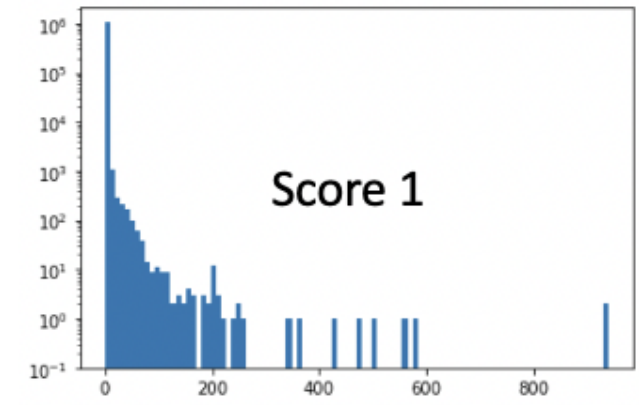
The first score method is the z-score outlier score. After dimension reduction, the new reduced dimensions are mostly uncorrelated, similarly scaled, and centered around a mean of zero. This allows us to calculate the z-scores for each record, which explicitly shows how unusual that record is in that dimension. We use the Minkowski distance to the origin to add up these z-scores on each record, without letting them cancel each other out. Finally, we calculate the distance to the origin for each point after these steps, which gives us the first fraud score. The equation for the score is below.

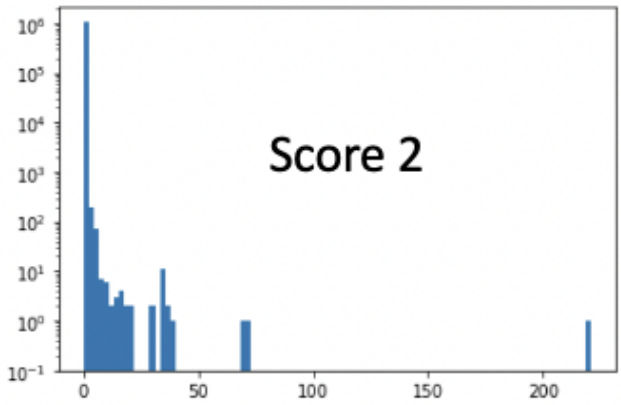
$$s_i = \left(\sum_n |z_n^i|^p \right)^{1/p}$$

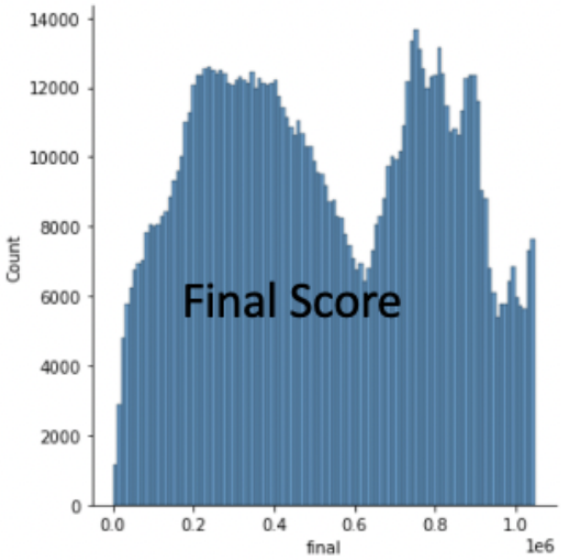
The second score method involves using an autoencoder, which is a model trained to output the original vector input. It is a functional mapping of a record back to itself and does a good job reproducing the "normal" records. However, some records are not reproduced well, and these are the unusual records that we are looking for. The error from an autoencoder is a good measure of the unusualness of a record. We use a neural network as our autoencoder, and after the model is trained, we calculate the difference (error) between the original input vector and the model output vector, which gives us the second fraud score. The equation for the score is below.

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

To combine these two scores, we use average rank orders to get the final score. we sort each of the score by its rank and use the average rank of the two score as our final fraud score. We can see that the distributions of Scores 1 and 2 are typical shapes for fraud distributions, while a ranking distribution is more uniform. Therefore, we use the average ranking of our final score to get a more accurate and robust detection process.







In summary, combining the z-score outlier score with the autoencoder score using average rank orders can improve the accuracy of anomaly detection algorithms, particularly in fraud detection scenarios. By using these two score methods together, we can identify unusual records and improve the overall security and integrity of our data.

b. Experiments on Sensitivity of Results to Different Parameters

Additionally, we tried different combinations of power of p in the Minkowski distance formula and number of principle components to see the sensitivity of the results to various choices.

1. Table with experiments across variations

Changes in choice from baseline	Top 100 (.01%)	Top 1,000 (.1%)	Top 10,000 (1%)
baseline: p1=p2=2, 4 PCs			
p1=1	98.0	96.9	97.1
p2=1	98.0	98.5	96.1
5PCs	91.0	57.5	76.4
3PCs	92.0	77.9	64.5
p1=3	96.0	98.8	98.1
p1=1.5, p2=2.5	99.0	98.3	98.0
p1=4,p2=4	93.0	96.7	96.0
Don't Z Scale 5PCs	94.0	52.8	76.8
Average over variations	94.0	84.7	84.8

According to our analysis, we found that the sensitivity level of the algorithm is comparable for both the top 1000 and top 10000 records. However, we observed that decisions related to principal components have a more significant impact on the sensitivity of the top records when compared to the top 1000 and top 10000 records. Additionally, we found that the accuracy of the autoencoder is minimally sensitive to algorithm choices, as is the z-score outlier score. Finally, we noticed that the highest records are less sensitive to algorithm choices compared to the other records.

VII. Results

Logic:

To identify anomalies or potential fraudulent records, we utilize a heatmap that highlights extremely high scores within the data. Once an abnormality is detected, we zoom in on the specific record to scrutinize the factors that contributed to the high score. We formulate hypotheses and leverage online resources such as pictures and listing prices to validate our findings. We have documented five interesting cases that we have identified and thoroughly examined.

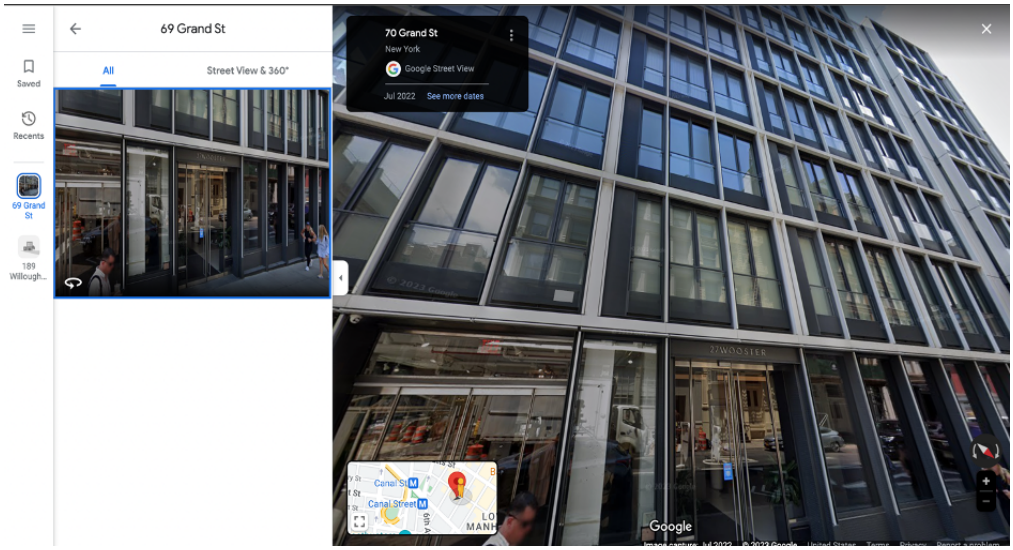
Case 1:

Property record: 14979

RECORD	14979
BBLE	1002280030
BORO	1
BLOCK	228
LOT	30
EASEMENT	NaN
OWNER	ENJAY ASSOCIATES
BLDGCL	G6
TAXCLASS	4
LTFRONT	114
LTDEPTH	80
EXT	NaN
STORIES	1.0
FULLVAL	2680000.0
AVLAND	1201500.0
AVTOT	1206000.0
EXLAND	0.0
EXTOT	0.0
EXCD1	NaN
STADDR	69 GRAND STREET
ZIP	10013.0
EXMPTCL	NaN
BLDFRONT	8
BLDDEPTH	6

AVLAND2	1281800.0
AVTOT2	1287400.0
EXLAND2	NaN
EXTOT2	NaN
EXCD2	NaN
PERIOD	FINAL
YEAR	2010/11
VALTYPE	AC-TR
score1 rank	1046484.0
score2 rank	1046494.0
final	1046489.0

The picture of actual property:



1. This property has only 1 story recorded in the data. However, the actual building has multiple stories (8)
2. The r2, r3,r5,r6, r9 and related variables grouped by zip and tax(r1_zip_mean, r5_zip_mean, r1_zip5, r5_zip5, r5_taxclass) are high.
3. This may be a result of the inaccurate record of number of stories or because of inaccurate BLDFRONT (8 is inappropriate for such a tall building) or BLDDEPTH (6 is inappropriate). We can look at the calculates of these unusual variables:

```
data['ltsize'] = data['LTFRONT'] * data['LTDEPTH']
data['bldsize'] = data['BLDFRONT'] * data['BLDDEPTH']
data['bldvol'] = data['bldsize'] * data['STORIES'] data['r1']
          = data['FULLVAL'] / data['ltsize']
```

```
data['r2'] = data['FULLVAL'] / data['bldsize']
data['r3'] = data['FULLVAL'] / data['bldvol']
data['r5'] = data['AVLAND'] / data['bldsize']
data['r6'] = data['AVLAND'] / data['bldvol']
data['r9'] = data['AVTOT'] / data['bldvol']
```

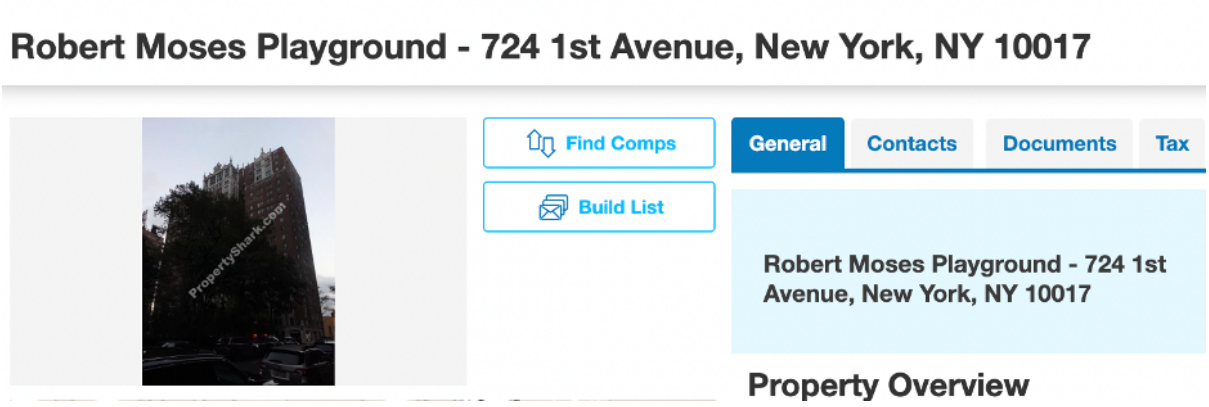
4. The most common factor in the calculation of variables listed a bove that differentiate them from the rest of variables can be the value of STORIES, BLDFRONT, BLDDEPTH.
5. Further investigation is required for this property

Case 2:
Property record: 95995

The data 95995 has abnormality, extremely high in r2, r3, r5, r6, r8, r9.
Here is the equations for the calculations:

```
data['ltsize'] = data['LTFRONT'] * data['LTDEPTH']
data['bldsize'] = data['BLDFRONT'] * data['BLDDEPTH']
data['bldvol'] = data['bldsize'] * data['STORIES'] data['r1'] = data
['FULLVAL'] / data['ltsize']
data['r2'] = data['FULLVAL'] / data['bldsize']
data['r3'] = data['FULLVAL'] / data['bldvol']
data['r4'] = data['AVLAND'] / data['ltsize']
data['r5'] = data['AVLAND'] / data['bldsize']
data['r6'] = data['AVLAND'] / data['bldvol']
data['r7'] = data['AVTOT'] / data['ltsize']
data['r8'] = data['AVTOT'] / data['bldsize']
data['r9'] = data['AVTOT'] / data['bldvol']
```

In close examination of the equations and data, we see that AVTOT and AVLAN D is in reasonable range and abnormality is likely due to faulty BLDFRONT, BLDDEPTH, or Stories. We then investigated with other online record to veri fy that hypothesis.

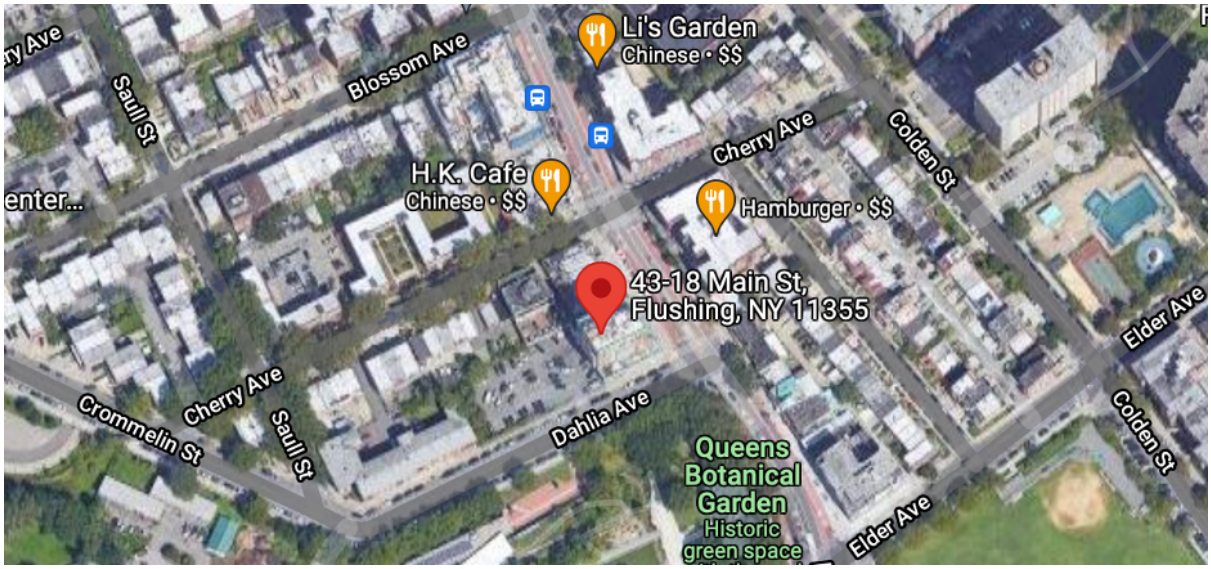


d

Case 3:
Property Record: 718883
Owner: GARDEN VIEW LTD
Address: 43-18 MAIN STREET

Ratio	inv	Inv_zip5	Inv_zip3	Inv_taxclass	Inv_boro
R1	209	231	269	289	296
R2	261	269	389	226	436
R3	376	316	711	169	777

The full value unit lot size, unit building size, and unit building volume are notably low here. After validating with the variable creation equations, I conclude the value of 12 for full values of this property might be fraudulent. There might be other faulty records that required further investigation.

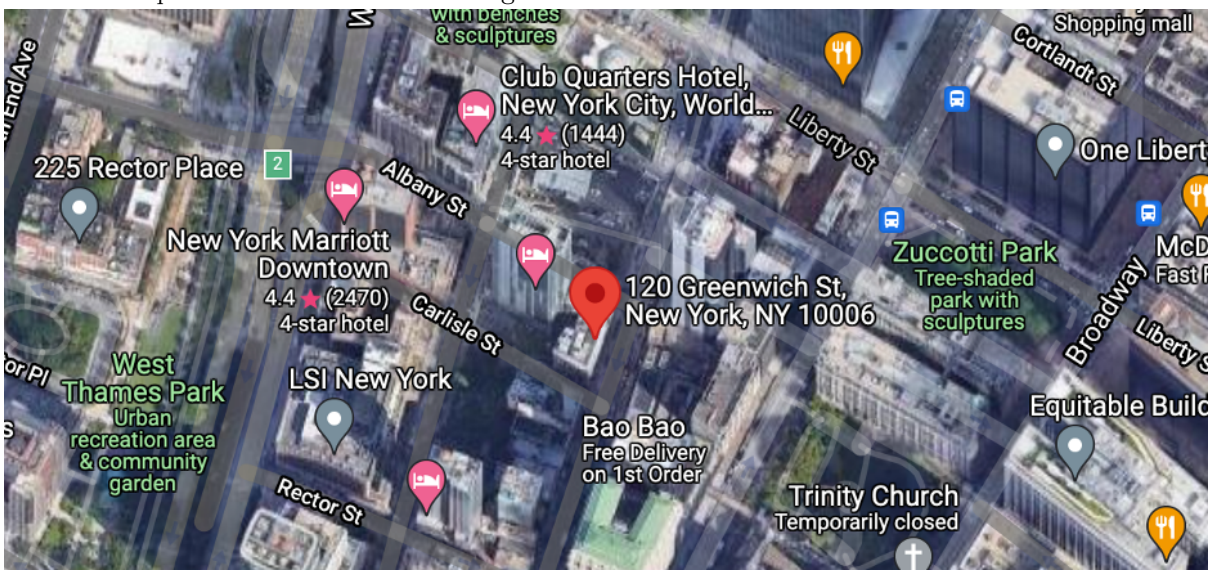


Case 4:
Property Record: 7034
Owner: HSIA, JONATHAN
Address: 120 GREENWICH STREET

Ratio	R1inv	R2inv	R3inv	R4inv	R5inv	R6inv	R8inv	R9inv

inv	140	174	233	59	211	160	108	211

The property characteristics are exceptionally low here. I researched the property and found out it is a condo in Manhattan financial district. The property value is lower than its neighboring properties, in particular, the full values, actual value of land and actual total value per unit. Since the pattern is consistent on all value fields, we think the record of full values of this property might be fraudulent. There might be other faulty records that required further investigation.



Case 5:

Property Record: 86946

The record 86946 has abnormality, extremely high in r1,r2,r4,r7, it has a final score of 1046491.

The values of R1-R9 are accordingly:

Here are the equations for the calculations:

1. data['ltsize'] = data['LTFRONT'] * data['LTDEPTH']
2. data['bldsize'] = data['BLDFRONT'] * data['BLDDEPTH']
3. data['bldvol'] = data['bldsize'] * data['STORIES']
4. data['r1'] = data['FULLVAL'] / data['ltsize']
5. data['r2'] = data['FULLVAL'] / data['bldsize']
6. data['r3'] = data['FULLVAL'] / data['bldvol']
7. data['r4'] = data['AVLAND'] / data['ltsize']
8. data['r5'] = data['AVLAND'] / data['bldsize']
9. data['r6'] = data['AVLAND'] / data['bldvol']
10. data['r7'] = data['AVTOT'] / data['ltsize']
11. data['r8'] = data['AVTOT'] / data['bldsize']
12. data['r9'] = data['AVTOT'] / data['bldvol']

In close examination of the equations and data, we see abnormality is likel y due to faulty LFTRONT, LIDEPTH, and FULLVALUE.

LFTRONT:28
LIDEPTH:150
BLDFRONT:25
BLDDEPTH :115
FULLVAL: 275000000

712 FIFTH AVENUE L P

We see that LIDEPTH should be much bigger values. Since the building here h as 52 stories, which way exceed the mean of stories, which is 5, it should have very large LIDEPTH compared to other buildings. However, its LIDEPTH value of 150, given the mean of 88 and the common value of 100, is not that big as its stories are way more than the common value. There might be other faulty records that need to be further investigated.

VIII. Summary

The report provides a detailed analysis of an unsupervised machine learning algorithm's performance in identifying fraudulent NY property cases in term s of data cleaning, variable creation, dimensionality reduction and anomaly detection. Since we do not have fraud labels for the data, we can not use s tandard metrics to measure the classification model. Thus, we need to talk with domain experts to deeply examine the performance of our unsupervised m odel. Specifically, We built the model using as much guidance from the exp erts as possible particularly in the process of building variables and sett ing exclusions. Then, we scored all records with the fraud algorithm, sorte

d the records by score and looked at the top records, trying to see if the se records made sense and aligned with what we expected. If so, we tweaked the algorithm and processes to ensure the experts will find the results int riguing. Next, we provided the experts with a list of the top few hundred r ecords to examine, starting from the top and providing feedback on what the y find noticeable.

Based on their feedback, we further modify the model by enhancing exclusion s and variables. After updating the model, we provide another list of top r ecords to the experts and seek additional feedback. We continue this iterat ive process until the experts indicate satisfaction with the findings. Typi cally, two to three iterations are required, although sometimes no iteratio ns are needed. The extent of iteration depends mostly on the quality of our initial variable selection and how well we heed the experts' advice.

However, the report also highlights some limitations of the unsupervised al gorithm, such as its reliance on the quality and completeness of the input data. Therefore, it is important to talk with domain experts and adjust bas ed on their feedback that better refine the model's performance. To adjust the model with expert feedback, the report recommends modifying variables s uch as the thresholds for suspicious behavior and exclusions such as specif ic fields of property. By incorporating expert knowledge into the algorith m, it is possible to improve its accuracy and efficiency in detecting fraud ulent cases. For example, one way to adjust an unsupervised algorithm with expert feedback is to modify variables. This involves tweaking the algorith m's parameters to better match the desired outcome. For example, if the al gorithm is clustering data points and the expert feedback suggests that the clusters are not distinct enough, the variables can be adjusted to increase the distance between clusters.
Additionally, we also try to adjust the algorithm through exclusions. Somet imes certain data points can be excluded from the algorithm to better align with the expert's knowledge. For example, if the expert has knowledge that certain data points do not fit the desired outcome, those data points can b e excluded. By modifying variables and exclusions, the algorithm can be bet ter tuned and have better overall performance.

IX. Appendix

Data Quality Report

1. Data Description

The dataset consists of **1,070,994 data records** of property valuation and assessment data collected by the city government department of finance, comprising **32 independent fields (14 numeric, 18 categorical)**.

2. Summary Tables

(1) Categorical Table

Field Name	# Records Have Values	% Populated	# Zeros	# Blanks	# Unique Values	Most Common Value
RECORD	1070994	100.00%	0	0	1070994	1
BBLE	1070994	100.00%	0	0	1070994	1000010101
BORO	1070994	100.00%	0	0	5	4
BLOCK	1070994	100.00%	0	0	13984	3944
LOT	1070994	100.00%	0	0	6366	1
EASEMENT	4636	0.43%	0	1066358	12	E
OWNER	1039249	97.04%	0	31745	863347	PARKCHESTER PRESERVAT
BLDGCL	1070994	100.00%	0	0	200	R4
TAXCLASS	1070994	100.00%	0	0	11	1
EXT	354305	33.08%	0	716689	3	G
EXCD1	638488	59.62%	0	432506	129	1017
STADDR	1070318	99.94%	0	676	839280	501 SURF AVENUE
ZIP	1041104	97.21%	0	29890	196	10314
EXMPTCL	15579	1.45%	0	1055415	14	X1
EXCD2	92948	8.68%	0	978046	60	1017
PERIOD	1070994	100.00%	0	0	1	FINAL
YEAR	1070994	100.00%	0	0	1	2010/11
VALTYPE	1070994	100.00%	0	0	1	AC-TR

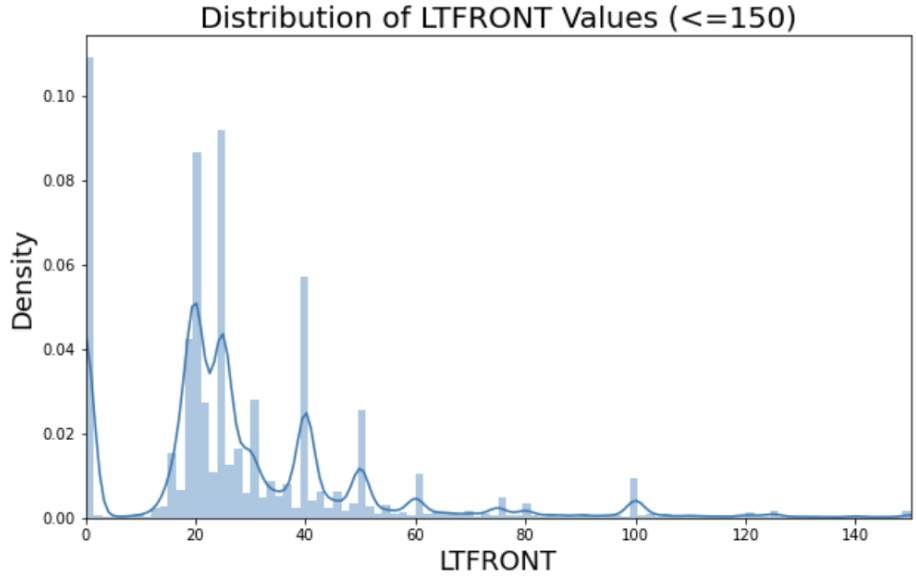
(2) Numerical Table

Field Name	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Stdev	Most Common Value
LTFRONT	1070994	100.00%	169108	0	9999	36.635301	74.03284	0
LTDEPTH	1070994	100.00%	170128	0	9999	88.861594	76.39628	100
STORIES	1014730	94.75%	0	1	119	5.006918	8.365707	2
FULLVAL	1070994	100.00%	13007	0	6150000000	874264.5054	11582430	0
AVLAND	1070994	100.00%	13009	0	2668500000	85067.91867	4057260	0
AVTOT	1070994	100.00%	13007	0	4668309000	227238.1687	6877529	0
EXLAND	1070994	100.00%	491699	0	2668500000	36423.89069	3981576	0
EXTOT	1070994	100.00%	432572	0	4668309000	91186.98168	6508403	0
BLDFRONT	1070994	100.00%	228815	0	7575	23.04277	35.5797	0
BLDDEPTH	1070994	100.00%	228853	0	9393	39.922836	42.70715	0
AVLAND2	282726	26.40%	0	3	2371005000	246235.7193	6178963	2408
AVTOT2	282732	26.40%	0	3	4501180000	713911.4362	11652530	750
EXLAND2	87449	8.17%	0	1	2371005000	351235.6843	10802210	2090
EXTOT2	130828	12.22%	0	7	4501180000	656768.2819	16072510	2090

3. Visualization of Each Field

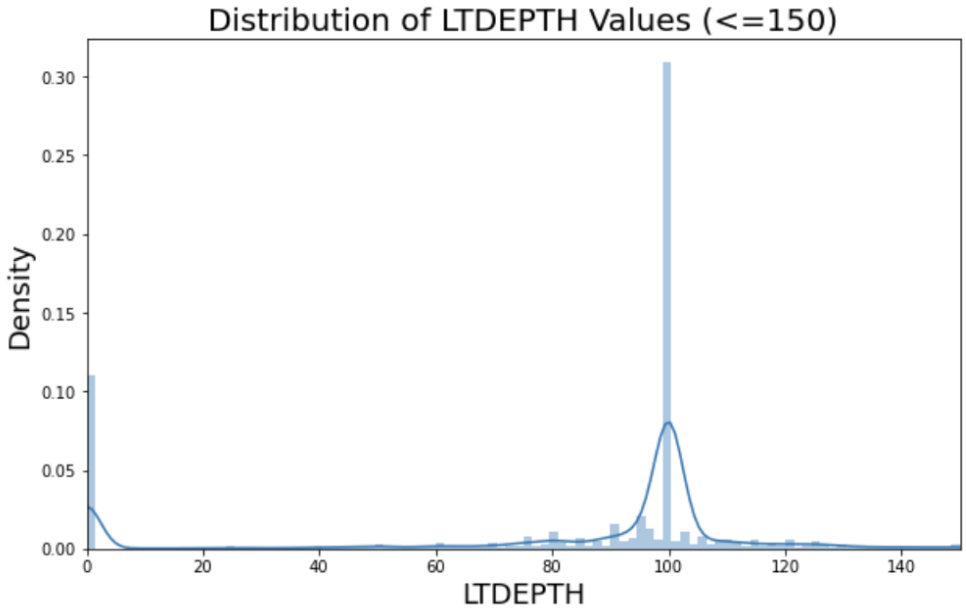
1) Field Name: LTFRONT

Description: a numeric field, the lot width. The histogram shows the distribution of lot width with the range between 0 to 150. The most common value is 0, count of which is 169108. The mean amount for all the transactions is 36.63. No missing value.



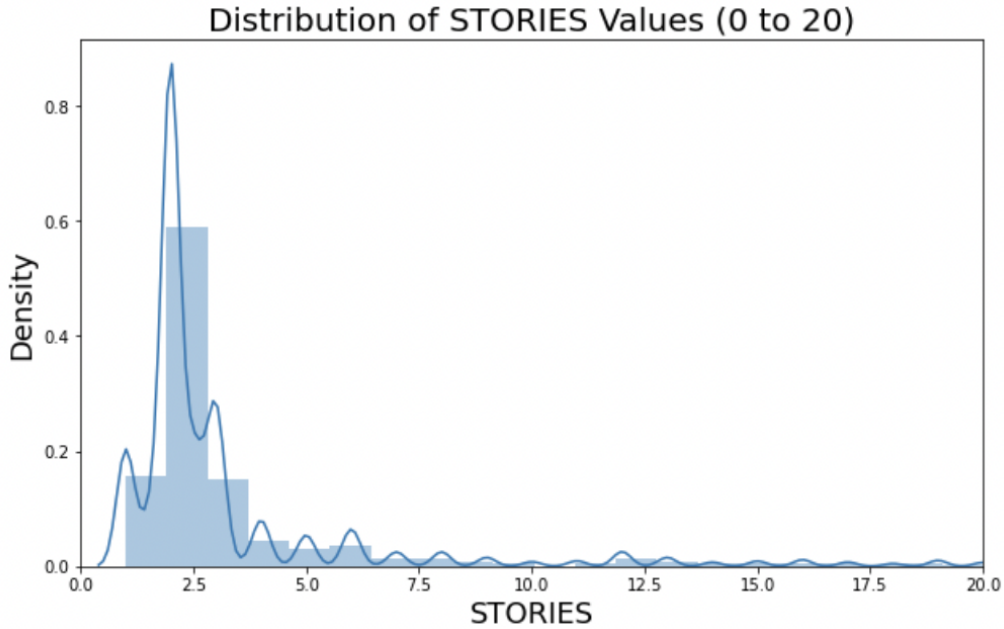
2) Field Name: LTDEPTH

Description: a numeric field, the lot depth. The histogram shows the distribution of lot width with the range between 0 to 150. The most common value is 100, count of which is 464,541. The mean amount for all the transactions is 88.86. No missing value.



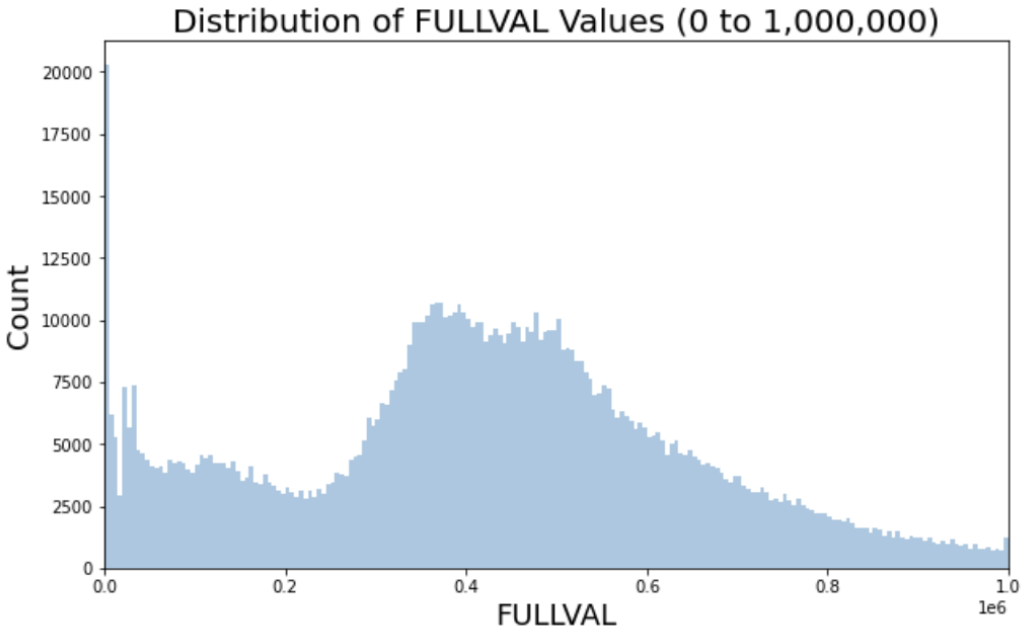
3) Field Name: STORIES

Description: a numeric field, the Number of Stories in Building. The histogram shows the distribution of lot width with the range between 0 to 20. The most common value is 2.0, count of which is 415092. The mean amount for all the transactions is 5.01. There are 56,264 missin g values.



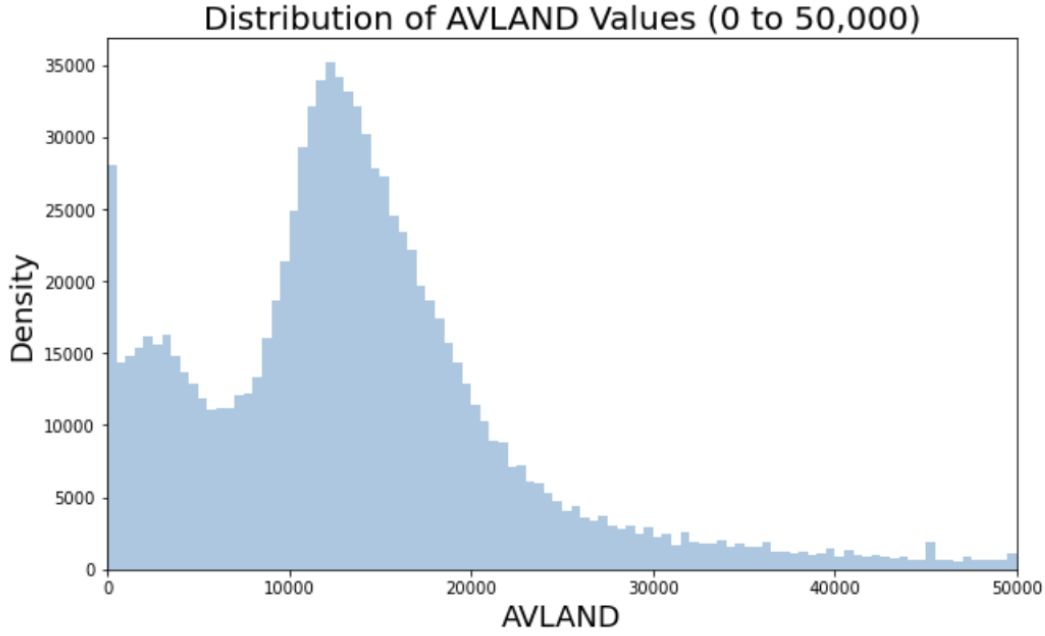
4) Field Name: FULLVAL

Description: a numeric field, the Market Value. The histogram shows t he distribution of lot width with the range between 0 to 1,000,000. T he most common value is 0, count of which is 13,007. The mean amount for all the transactions is 874264.51. No missing value.



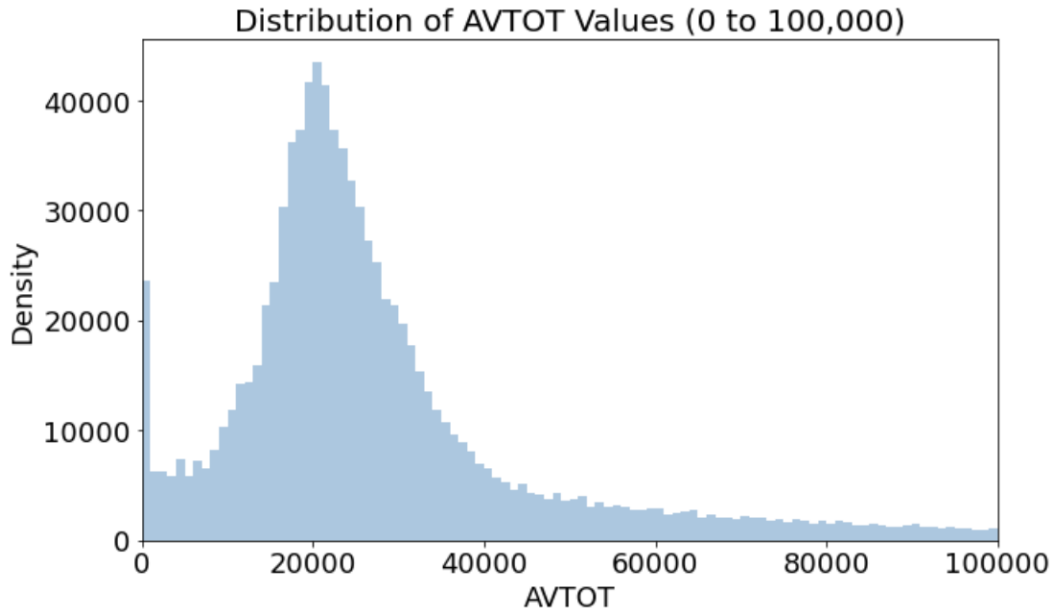
5) Field Name: AVLAND

Description: a numeric field, the Actual Land Value. The histogram shows the distribution of lot width with the range between 0 to 50,000. The most common value is 0, count of which is 13,009. The mean amount for all the transactions is 85067.92. No missing value.



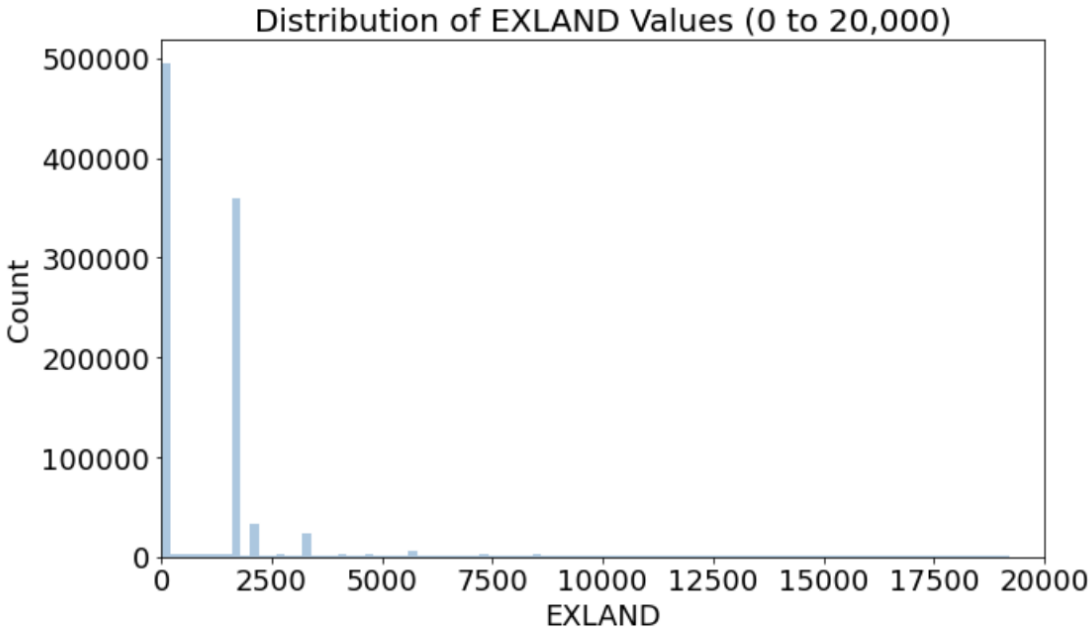
6) Field Name: AVTOT

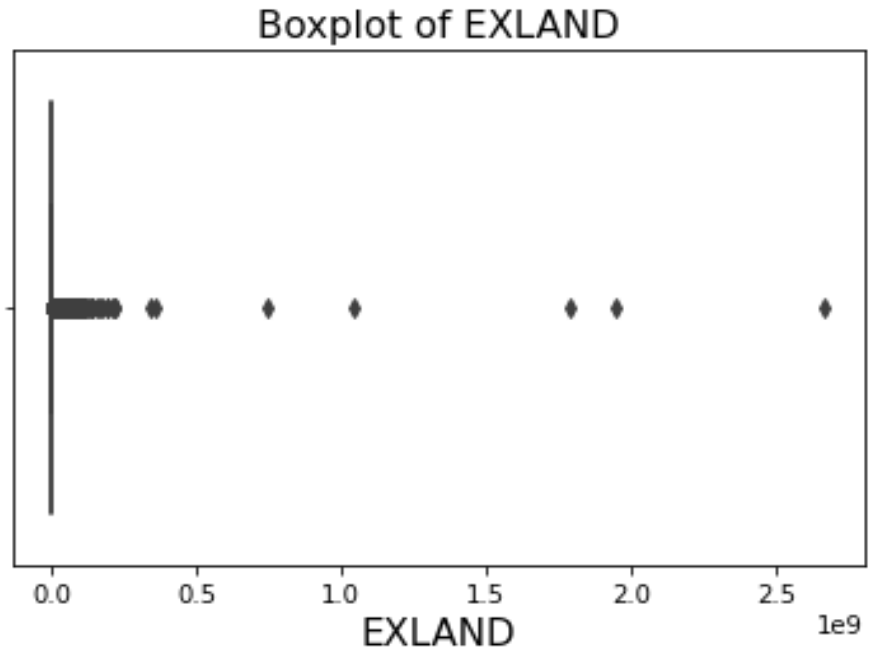
Description: a numeric field, the Actual Total Value. The histogram shows the distribution of lot width with the range between 0 to 100,000. The most common value is 0, count of which is 13,007. The mean amount for all the transactions is 227238.17. No missing value.



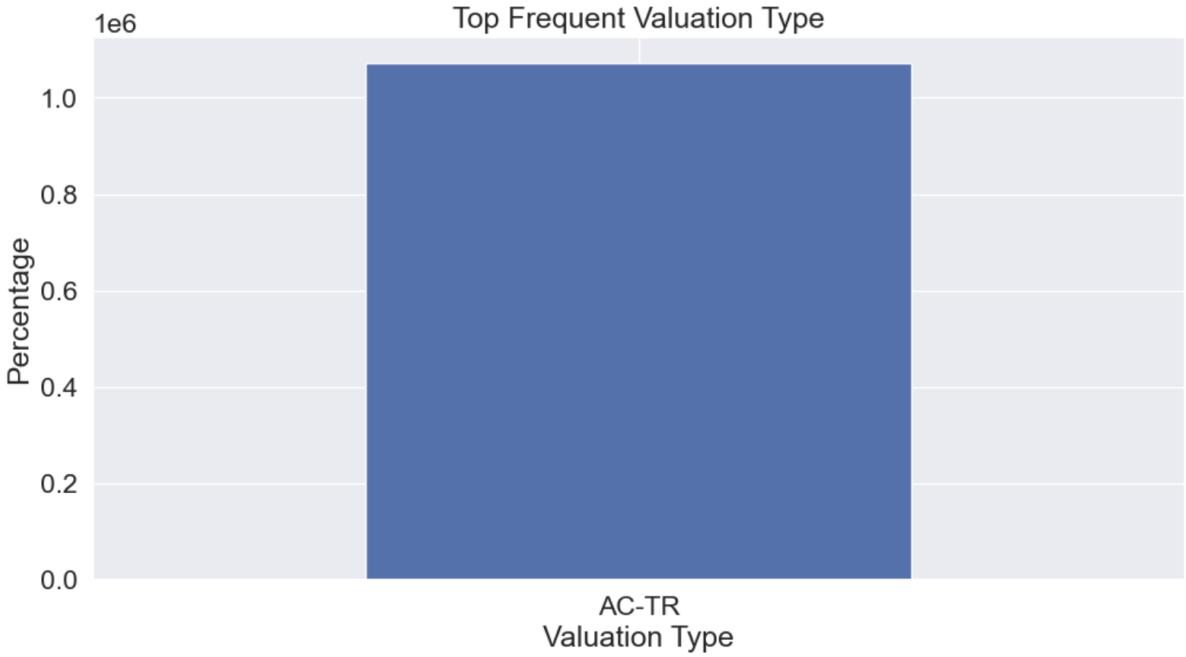
7) Field Name: EXLAND

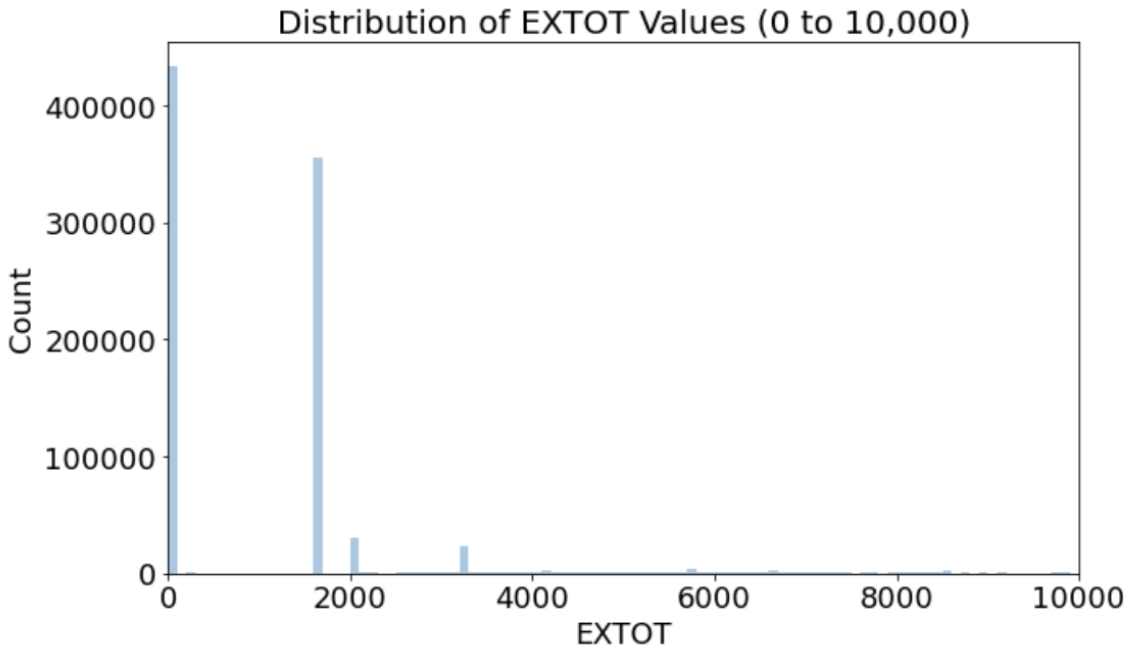
Description: a numeric field, the Actual Exempt Land Value. The histo gram shows the distribution of lot width with the range between 0 to 20,000. The most common value is 0, count of which is 491,699. The me an amount for all the transactions is 36423.89. No missing value.

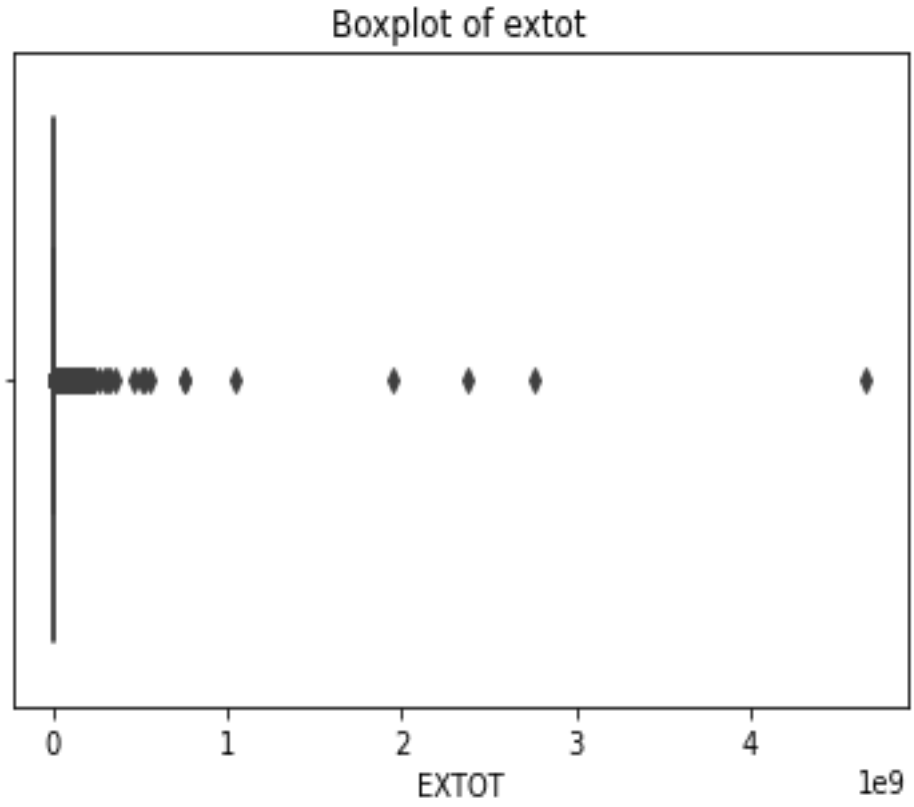




8) Field Name: EXTOT
Description: a numeric field, the Actual Exempt Land Total. The histogram shows the distribution of lot width with the range between 0 to 10,000. The most common value is 0, count of which is 432,572. The mean amount for all the transactions is 91186.98. No missing value.

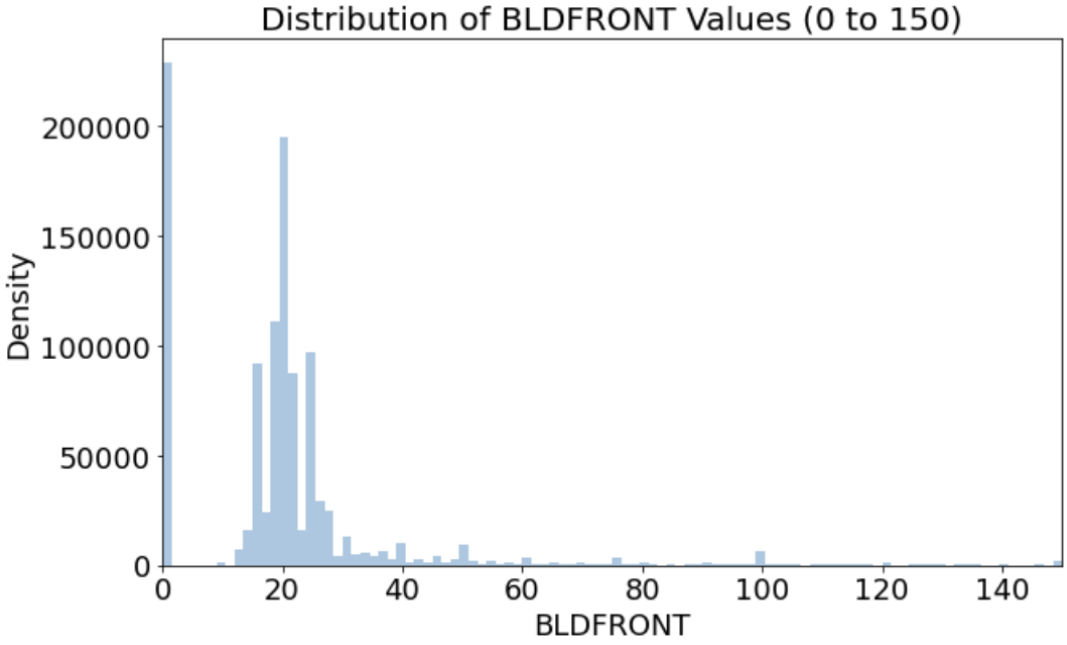

Top Frequent Valuation Type
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Percentage
AC-TR
Valuation Type


Distribution of EXTOT Values (0 to 10,000)
400000
300000
200000
100000
0
Count
0
2000
4000
6000
8000
10000
EXTOT



9) Field Name: BLDFRONT

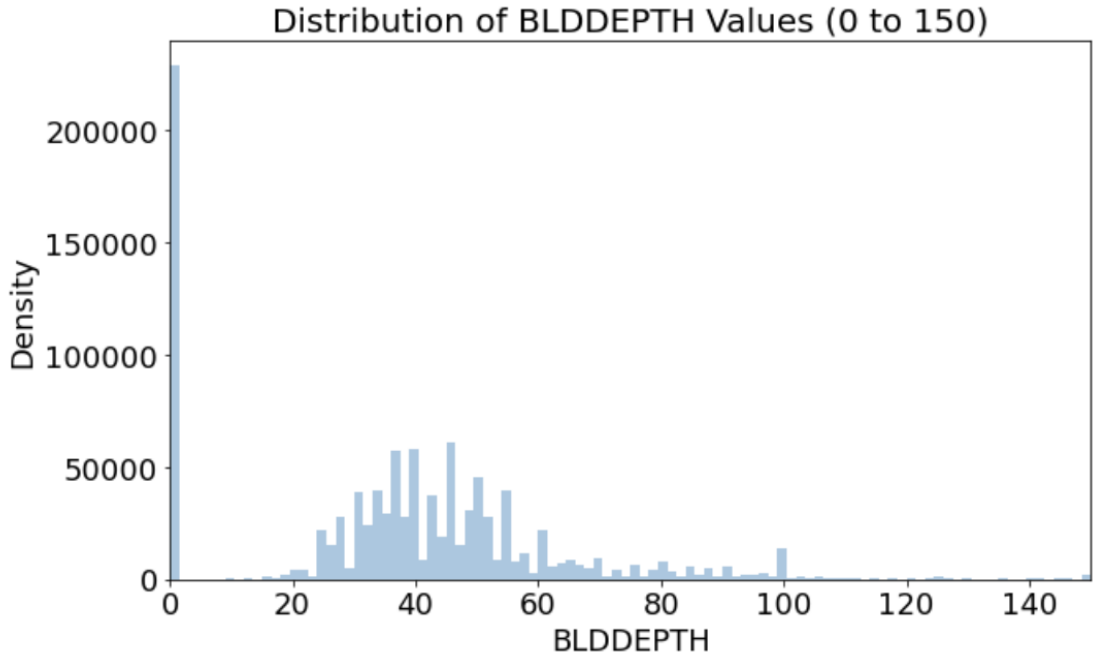
Description: a numeric field, the Building Width. The histogram shows the distribution of lot width with the range between 0 to 150. The most common value is 0, count of which is 228,815. The mean amount for all the transactions is 23.04. No missing value.



10) Field Name: BLDDEPTH

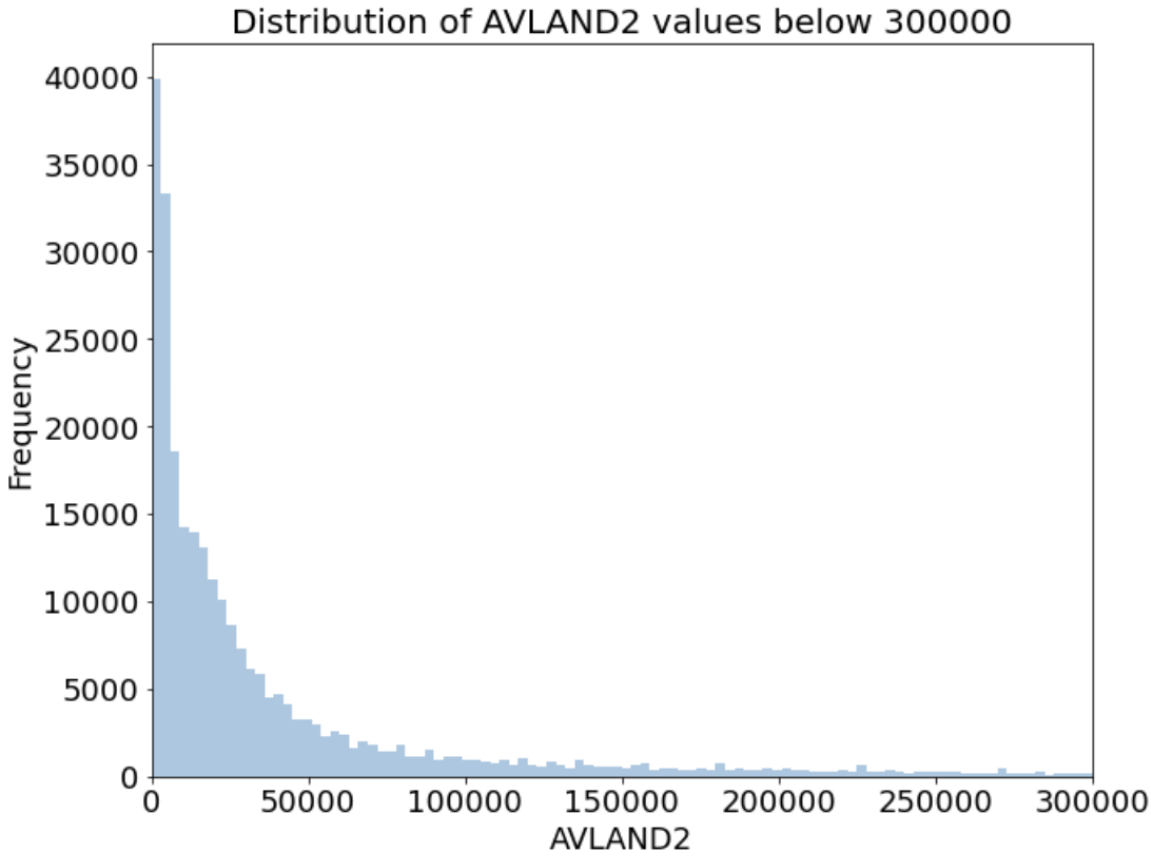
Description: a numeric field, the Building Depth. The histogram shows the distribution of lot width with the range between 0 to 150. The mo

st common value is 0, count of which is 228,853. The mean amount for a ll the transactions is 39.92. No missing value.



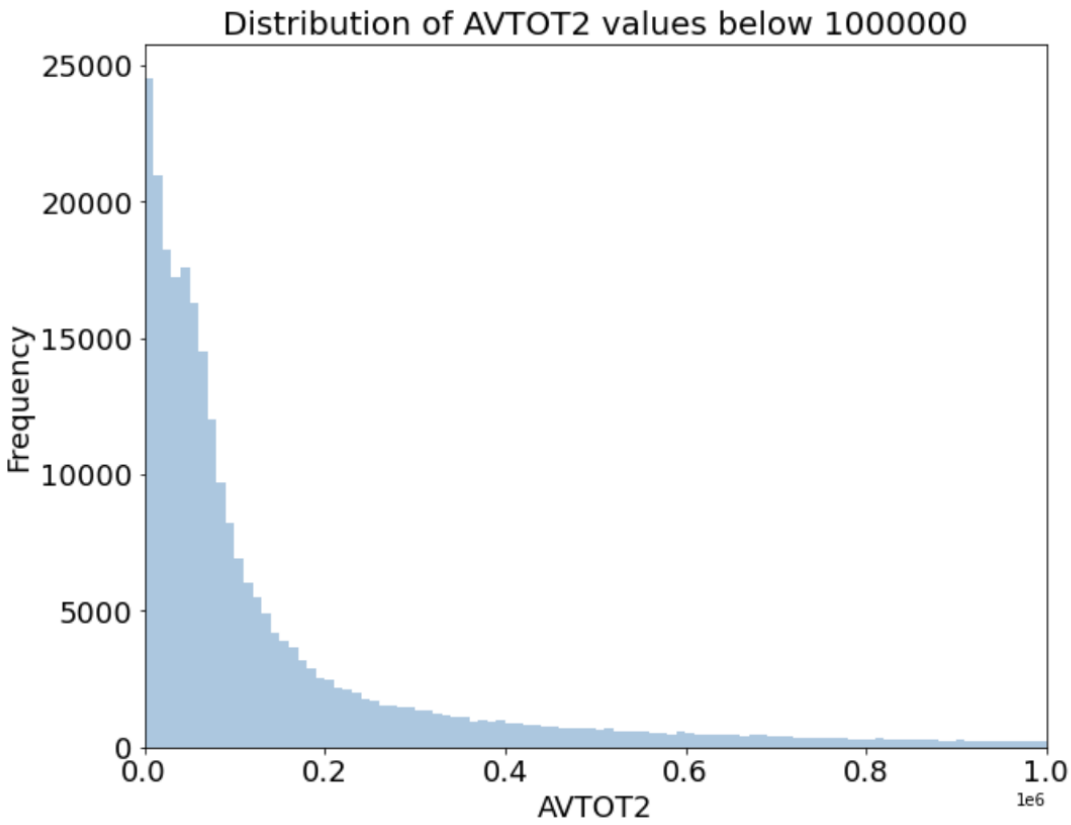
11)Field Name: AVLAND2

Description: a numeric field, the Transitional Land Value. The histog ram shows the distribution of lot width with the range between 0 to 3 00,000. The most common value is 2408.0, count of which is 767. The m ean amount for all the transactions is 246235.7193. There are 788,268 missing values.



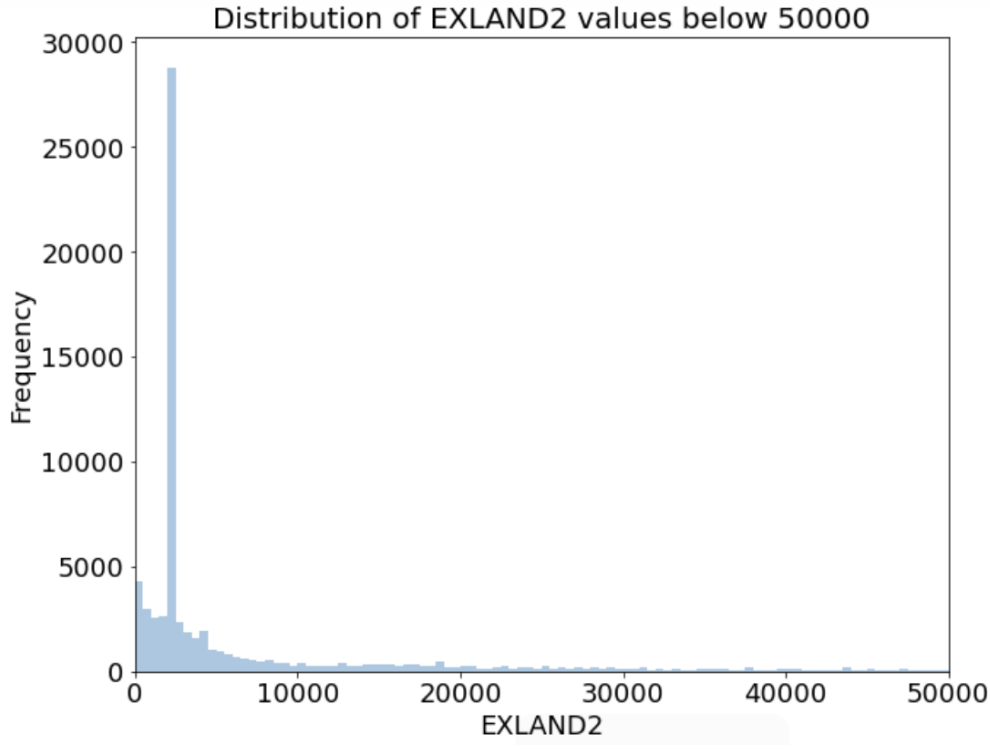
12) Field Name: AVTOT2

Description: a numeric field, the Transitional Total Value. The histogram shows the distribution of lot width with the range between 0 to 1,000,000. The most common value is 750, count of which is 656. The mean amount for all the transactions is 713911.44. There are 788,262 missing values.



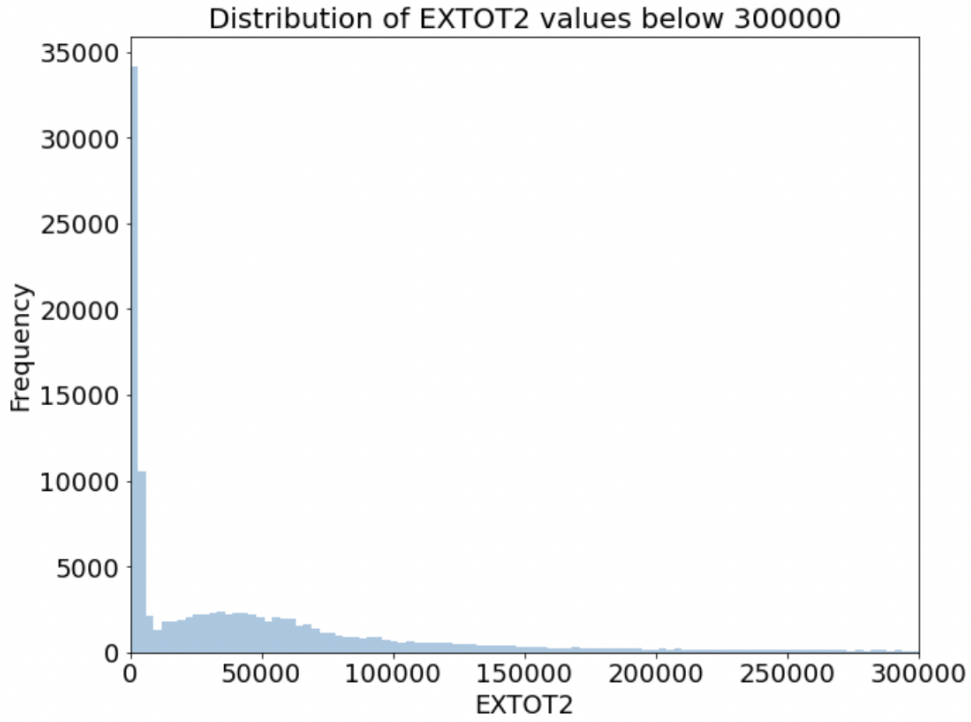
13) Field Name: EXLAND2

Description: a numeric field, the Transitional Exemption Land Value. The histogram shows the distribution of lot width with the range between 0 to 50,000. The most common value is 2090, count of which is 26,393. The mean amount for all the transactions is 351235.68. There are 983,545 missing values.



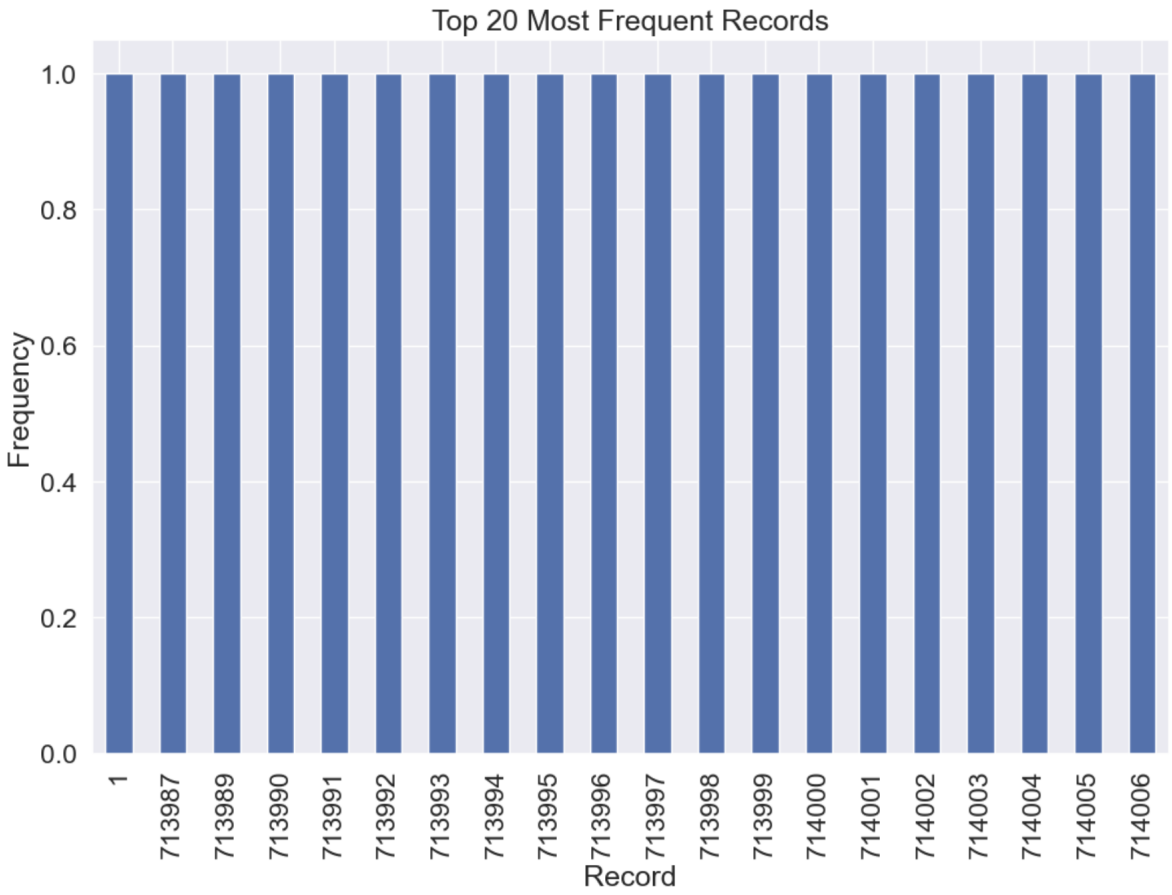
14) Field Name: EXTOT2

Description: a numeric field, the Transitional Exemption Land Total. The histogram shows the distribution of lot width with the range between 0 to 300,000. The most common value is 2090, count of which is 24,739. The mean amount for all the transactions is 656768.28. There are 940,166 missing values.



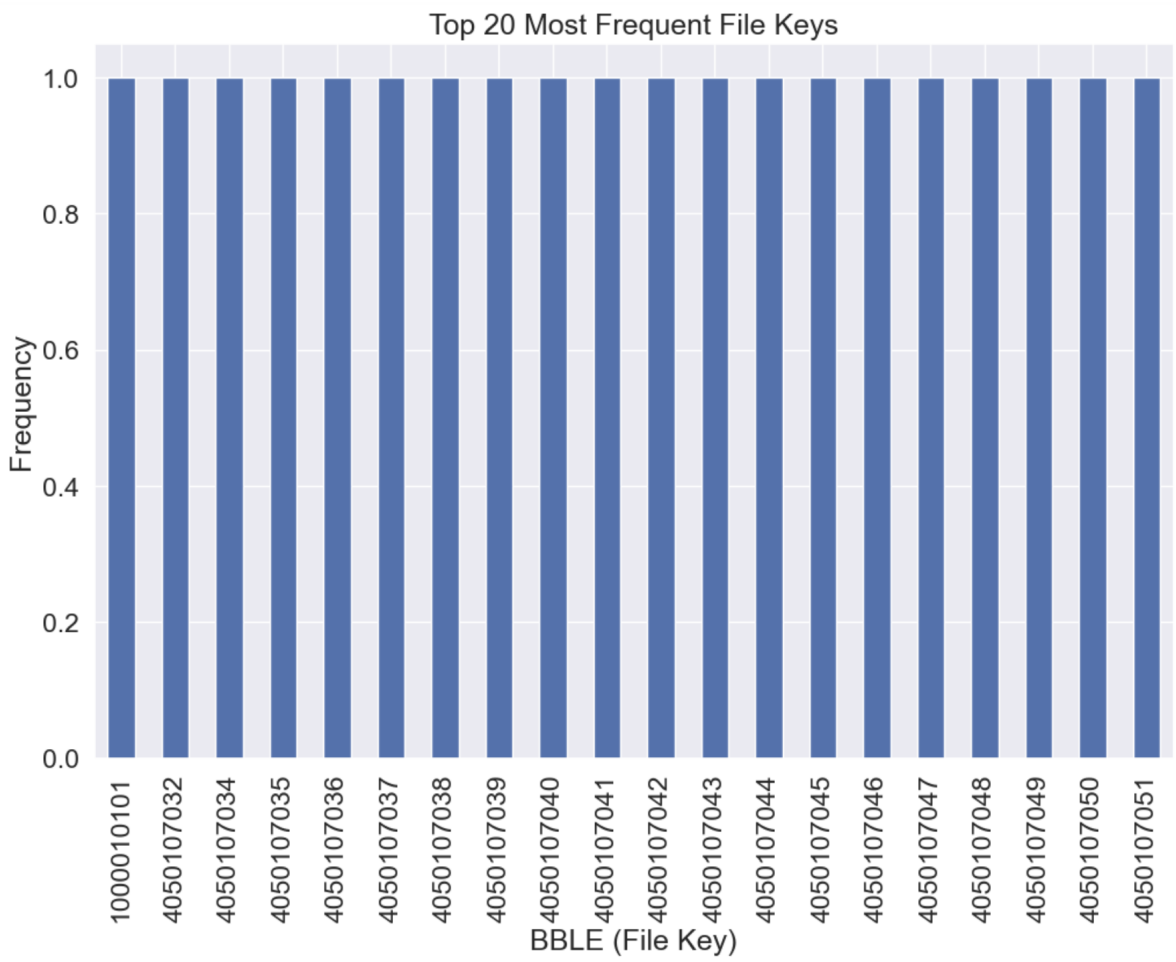
15) Field Name:RECORD

Description: a categorical field, a unique identifier for each record in the dataset. There are no zeros or missing values in this field. There are 1,070,994 unique values.



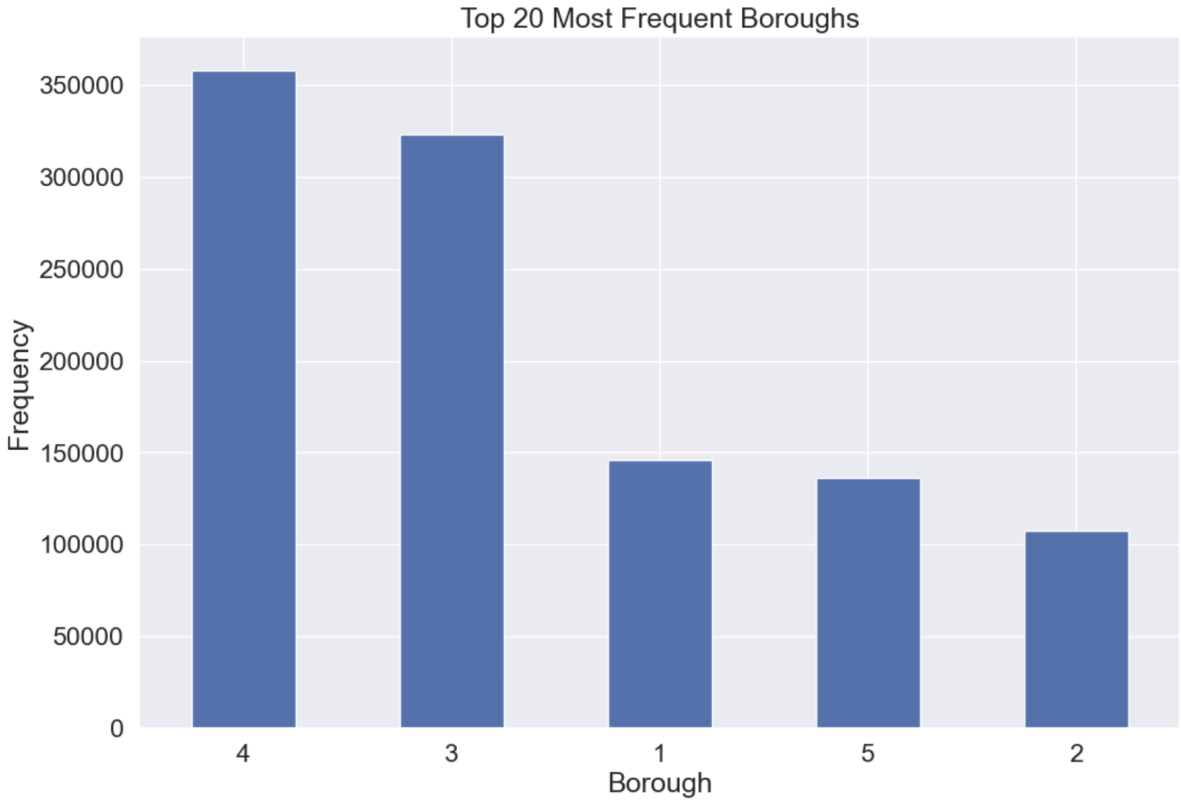
16) Field Name: BBLE

Description: a categorical field, a unique identifier for each property(Borough, Block, and Lot number) in New York City. There are no zeros in this field and there are 1,070,994 unique values.



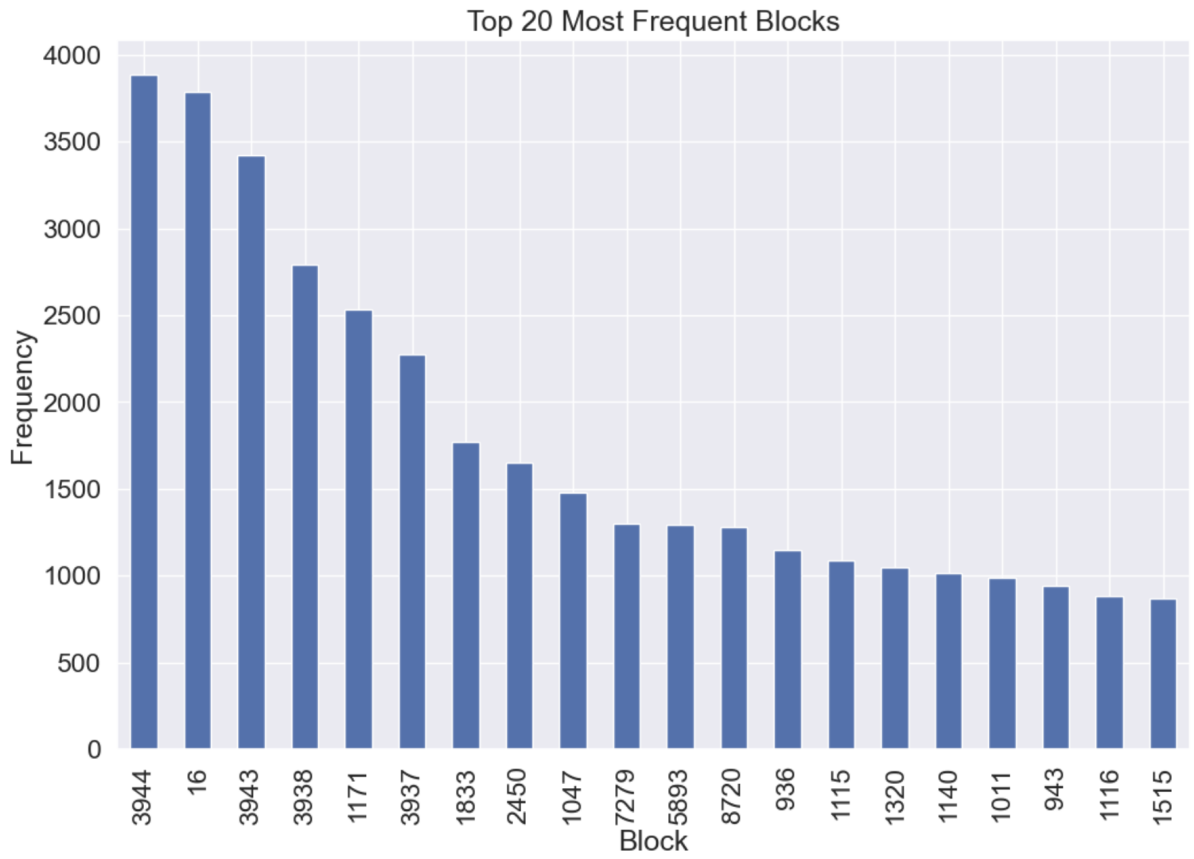
17) Field Name: BORO

Description: a categorical field, the borough (or county) in which the property is located. The histogram shows the frequency of top 20 field values of boroughs. The most common value is 2090, count of which is 24,739. There are no zeros in this field and there are 5 unique values (1=Manhattan, 2=Bronx, 3=Brooklyn, 4=Queens, 5=Staten Island).



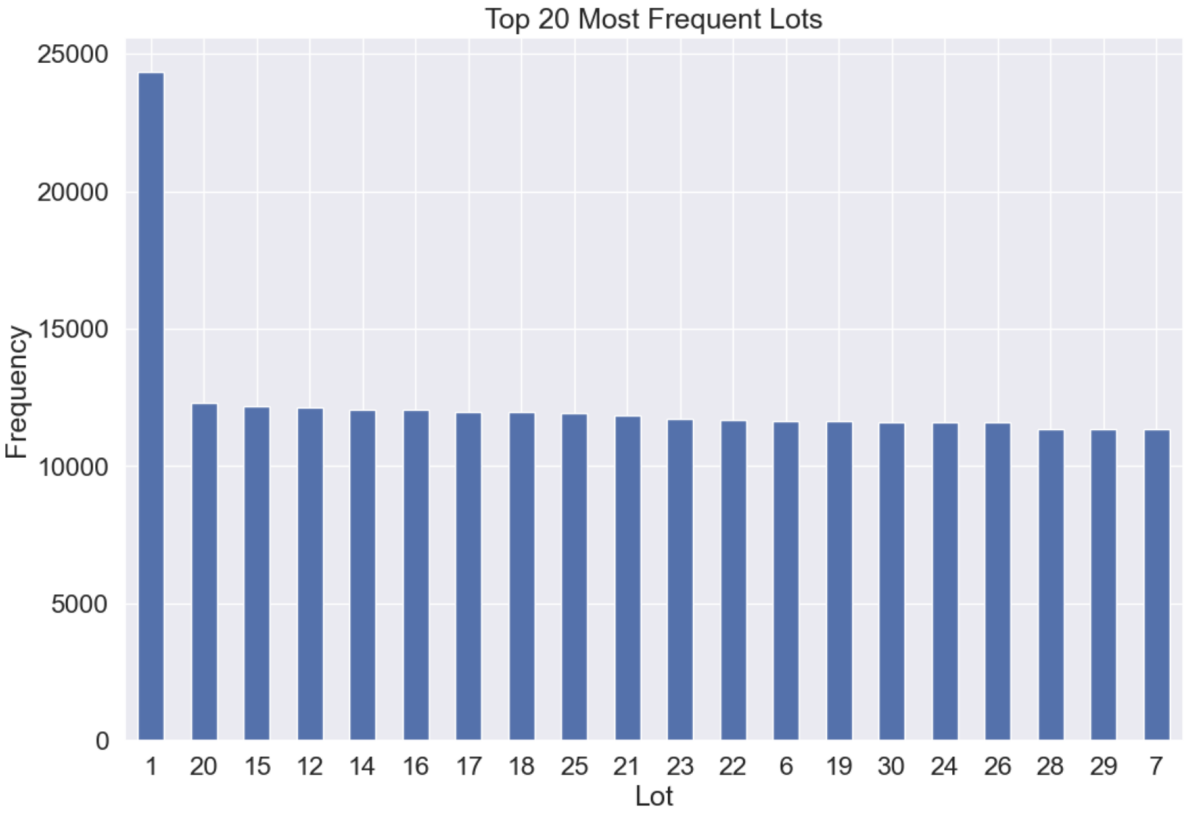
18)Field Name: BLOCK

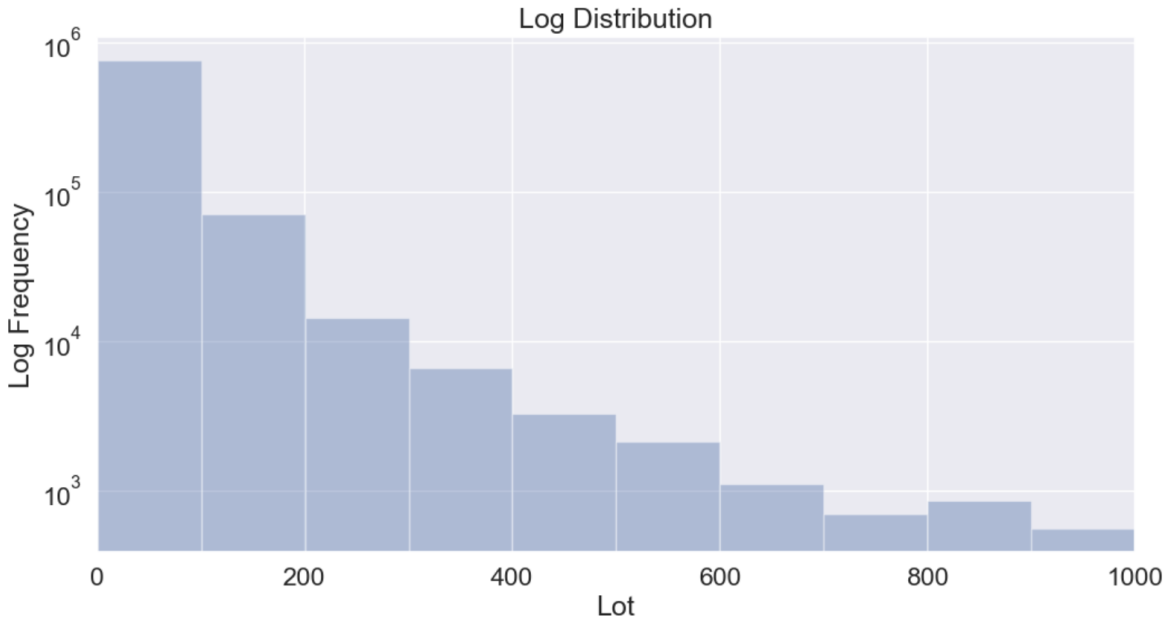
Description: a categorical field, the tax block in which the property is located. The histogram shows the frequency of top 20 field values of blocks. The most common value is 3944, count of which is 3888. The re are no zeros in this field and there are 13,984 unique values.



19) Field Name: LOT

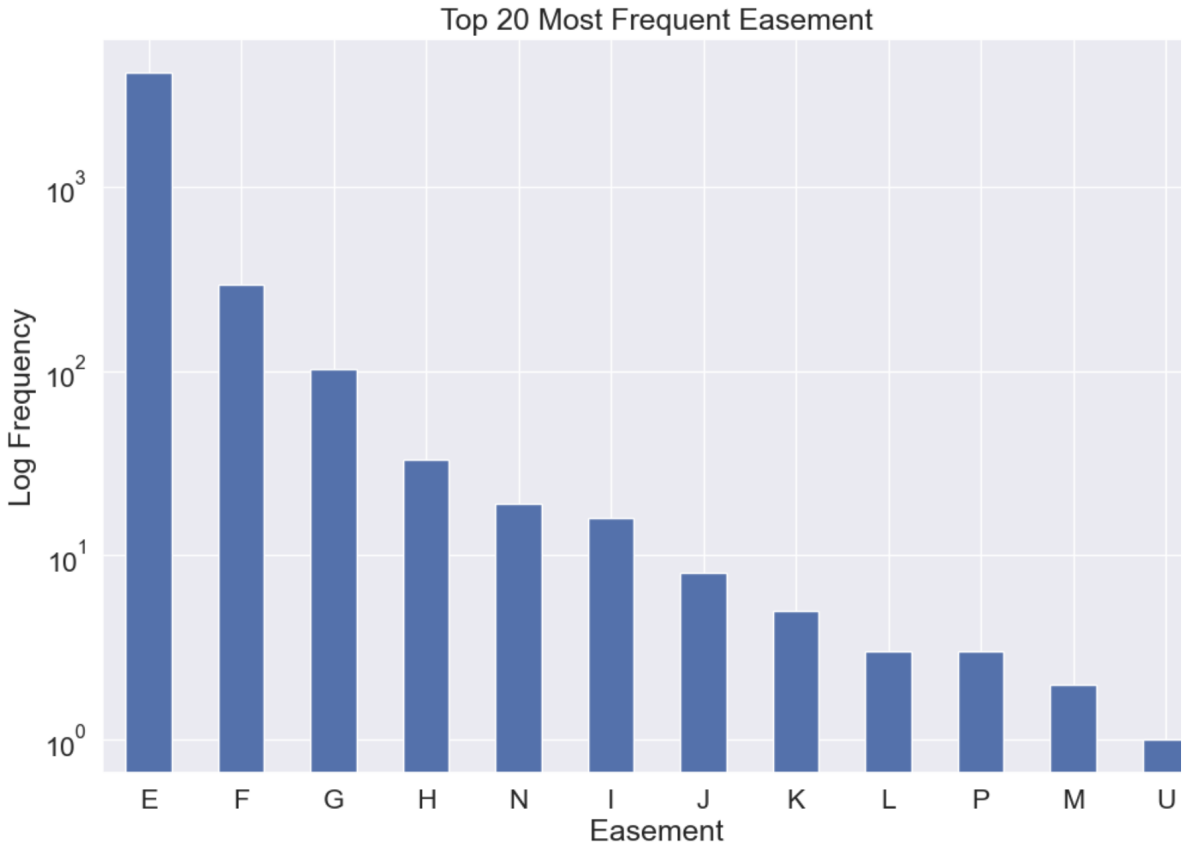
Description: a categorical field, the tax lot in which the property is located. The first histogram shows the frequency of top 20 field values of lots. The most common value is 1, count of which is 24367. There are no zeros in this field and there are 6,366 unique values. The second histogram sets the y axis scale of the plot to a logarithmic scale and only looks at the value in the 'LOT' column is less than or equal to 1000.





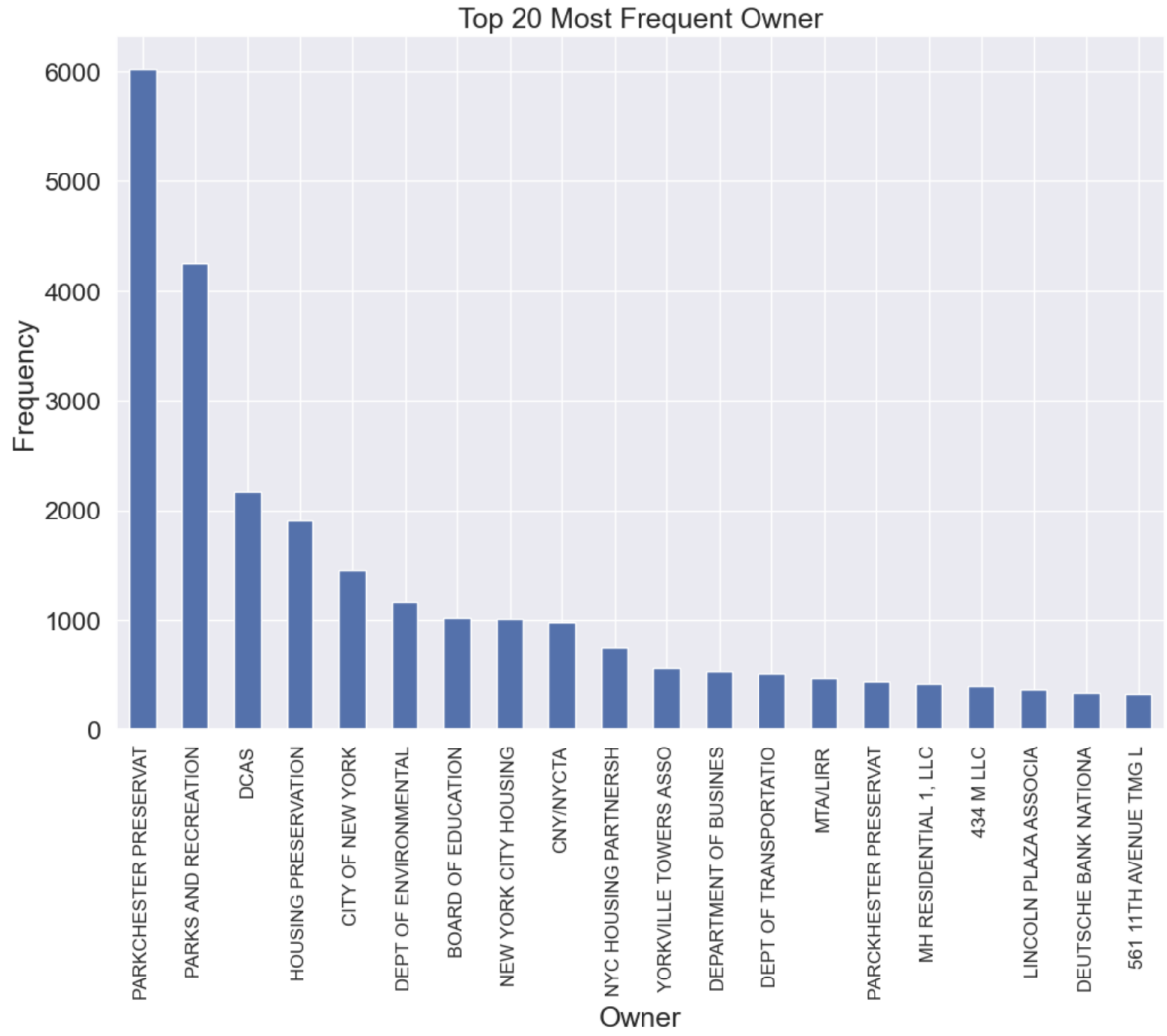
20) Field Name: EASEMENT

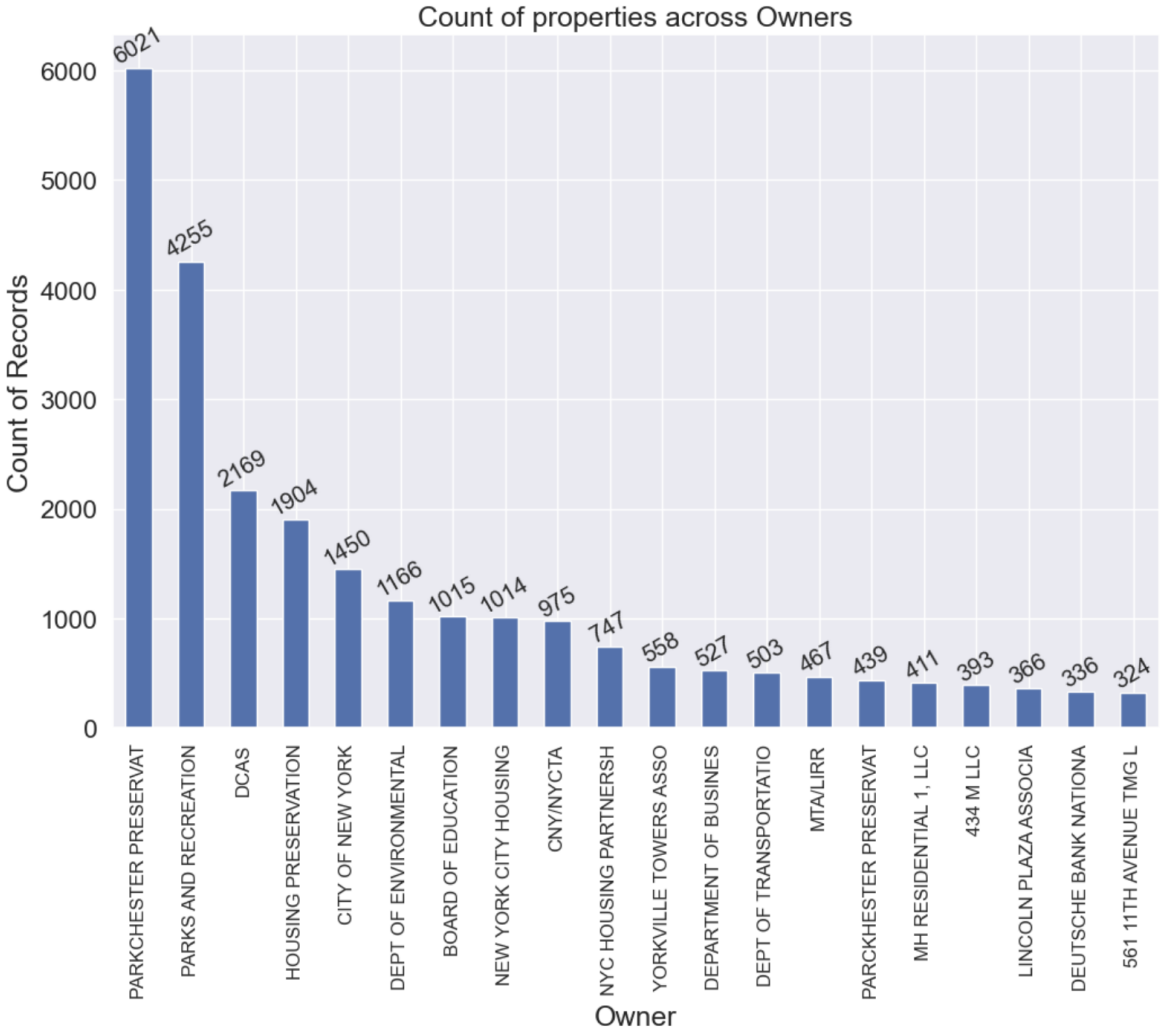
Description: a categorical field, a type of property easement, which is a right of use over the property of another. The histogram shows the log frequency of top 20 field values of EASEMENT. The most common value is E, count of which is 4148 . There are 1,062,358 zeros in this field, indicating that almost all properties in the dataset do not have an easement. There are 12 unique values.



21) Field Name: OWNER

Description: a categorical field, the name of the property owner. There are 30,745 zeros in this field, indicating that about 3% of the properties do not have an owner listed. The histogram shows the frequency of top 20 field values of OWNER. The most common value is PARKCHESTER PRESERVAT, count of which is 6021. There are 863,347 unique values.


Top 20 Most Frequent Owner
Frequency
6000
5000
4000
3000
2000
1000
0
PARKCHESTER PRESERVAT
PARKS AND RECREATION
DCAS
HOUSING PRESERVATION
CITY OF NEW YORK
DEPT OF ENVIRONMENTAL
BOARD OF EDUCATION
NEW YORK CITY HOUSING
CNY/NYCTA
NYC HOUSING PARTNERSH
YORKVILLE TOWERS ASSO
DEPARTMENT OF BUSINES
DEPT OF TRANSPORTATIO
MTA/LIRR
PARCKHESTER PRESERVAT
MH RESIDENTIAL 1, LLC
434 M LLC
LINCOLN PLAZA ASSOCIA
DEUTSCHE BANK NATIONA
561 11TH AVENUE TMG L
Owner



22) Field Name: BLDGCL

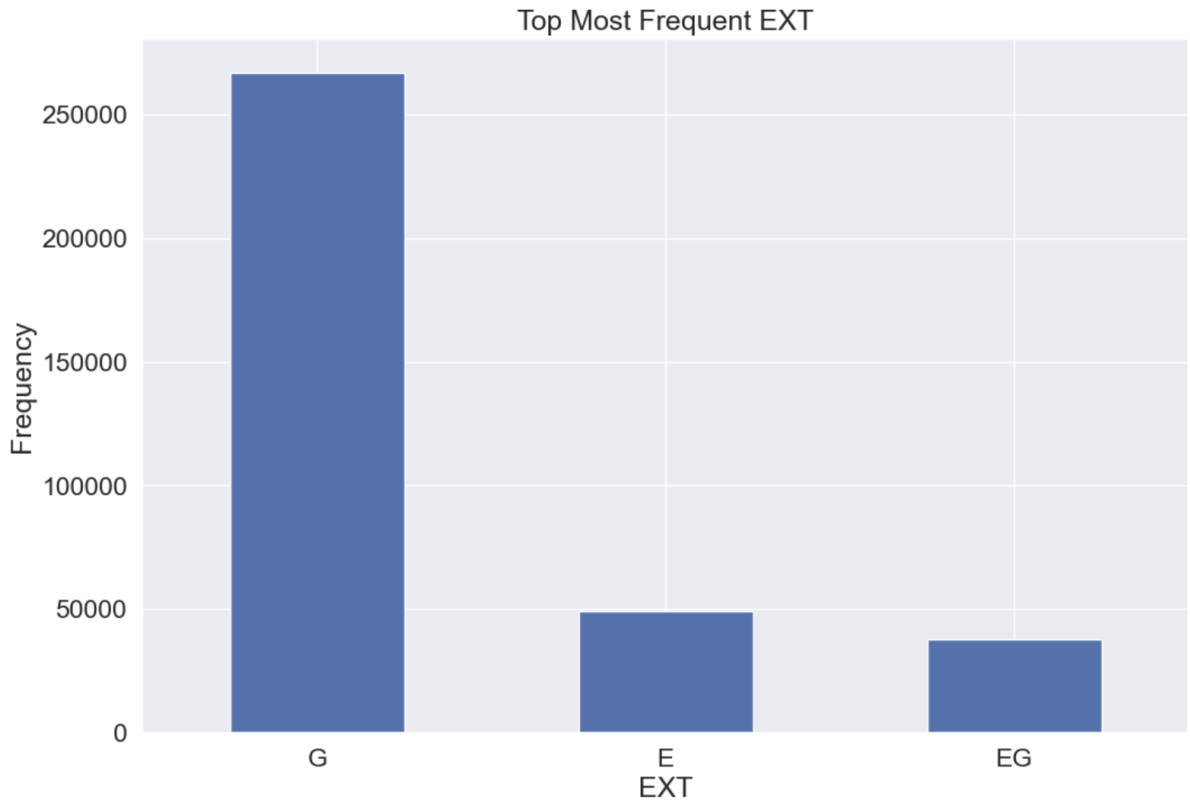
Description: a categorical field, the building class category code, which describes the type of building. The histogram shows the frequency of top 20 field values of Building Class. The most common value is R4, count of which is 139879. There are no zeros in this field and there are 200 unique values.

23) Field Name: TAXCLASS

Description: a categorical field, the property tax class code, which determines the tax rate for the property. The histogram shows the frequency of top 20 field values of Tax Class. The most common value is 1, count of which is 660721. There are no zeros in this field and there are 11 unique values.

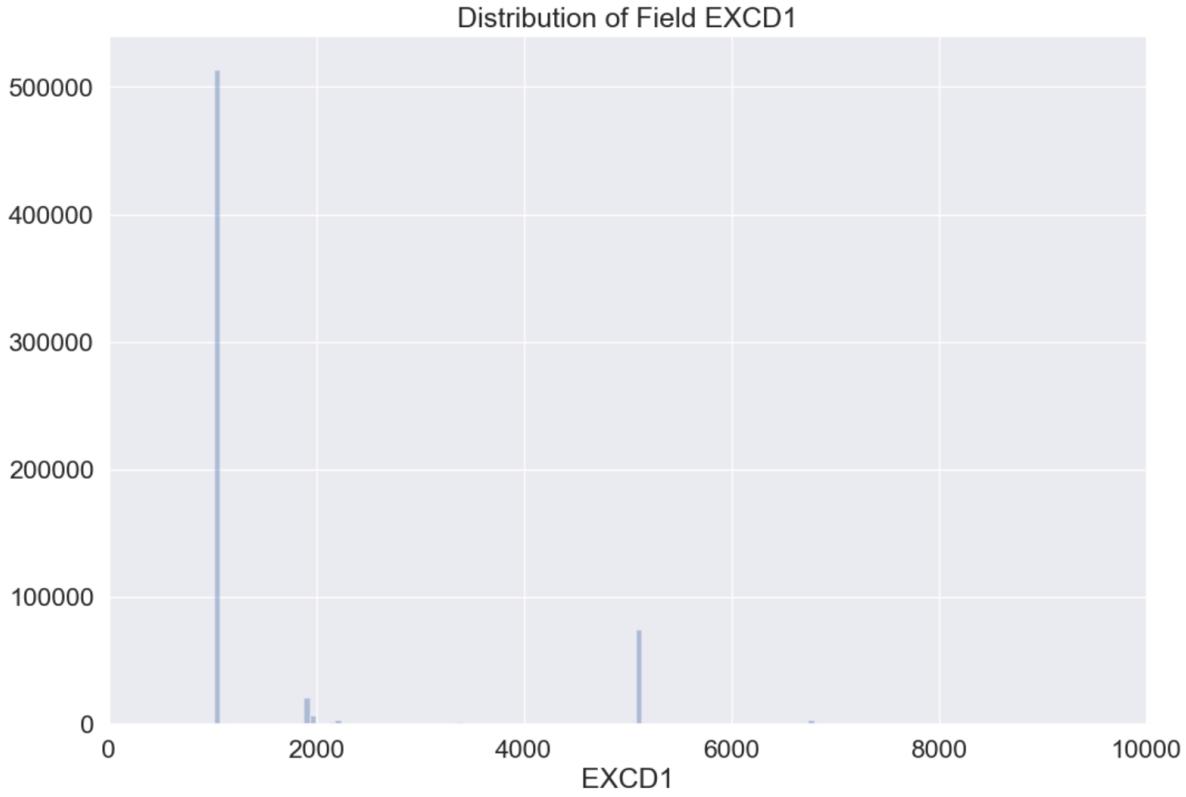
24) Field Name: EXT

Description: a categorical field, a code indicating the type of extension to the building, such as a garage or a porch. The histogram shows the frequency of top 3 field values of EXT. The most common value is G, count of which is 266970. There are 716,689 zeros in this field, indicating that about two-thirds of the properties do not have an extension. There are 3 unique values(G,E,EG).



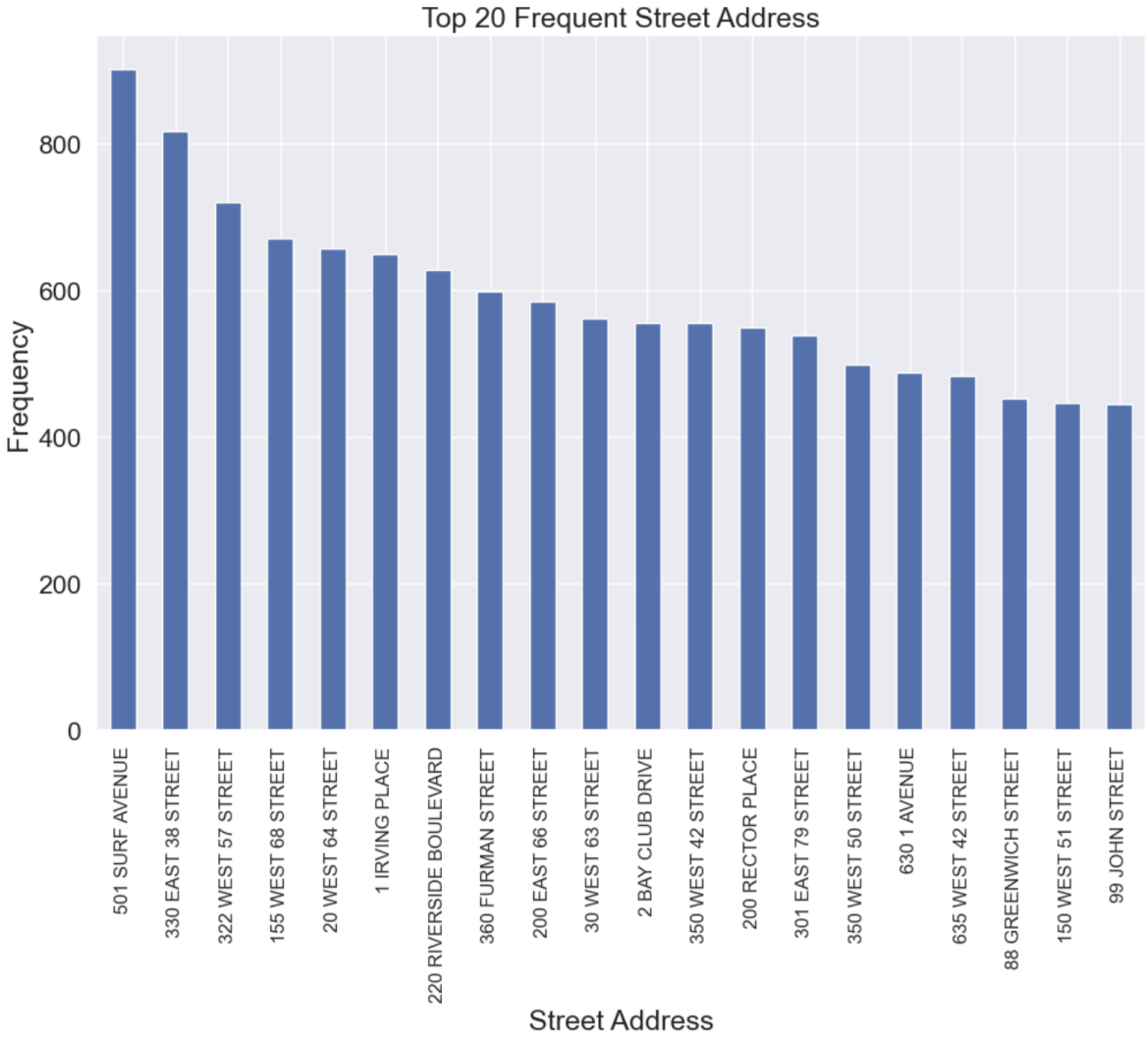
25) Field Name: EXCD1

Description: a categorical field, a code indicating a type of exemption or abatement from property taxes. There are 431,506 zeros in this field, indicating that about 40% of the properties do not have an exemption code listed. There are 129 unique values. The most common value is 1017.



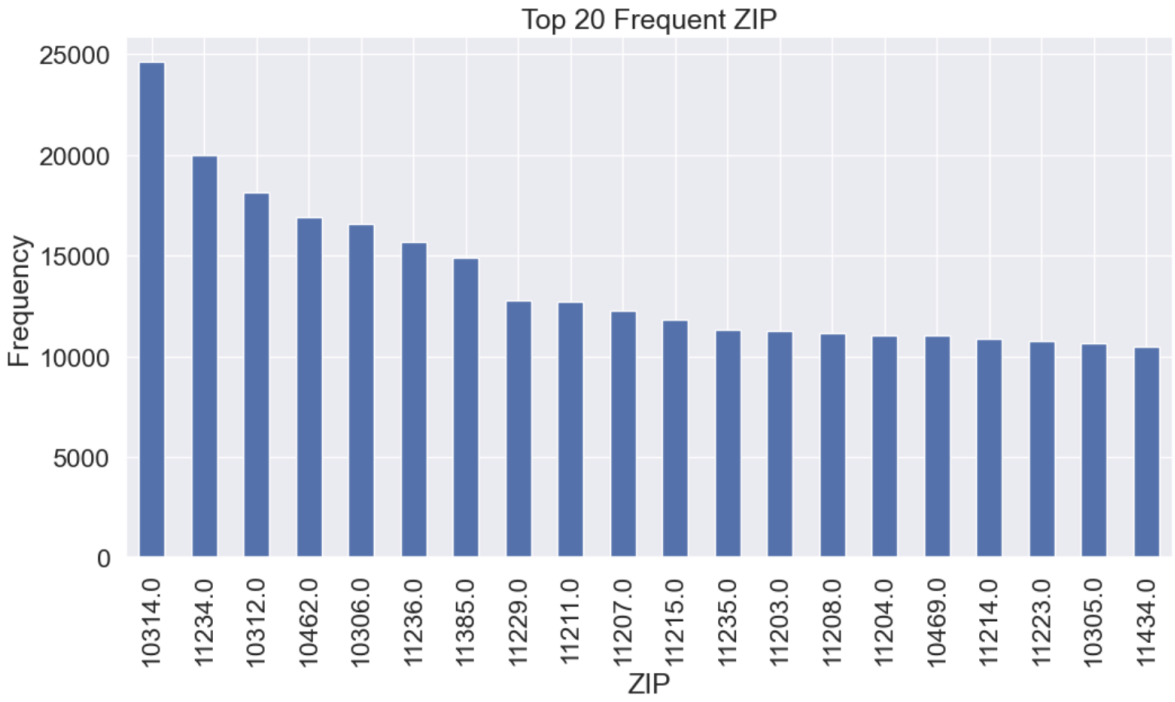
26) Field Name:STADDR

Description: a categorical field, the street address of the property. The histogram shows the frequency of top 20 field values of Street address. The most common value is 501 SURF AVENUE, count of which is 902. There are 676 zeros in this field, indicating that less than 0.1% of the properties do not have an address listed. There are 839,280 unique values.



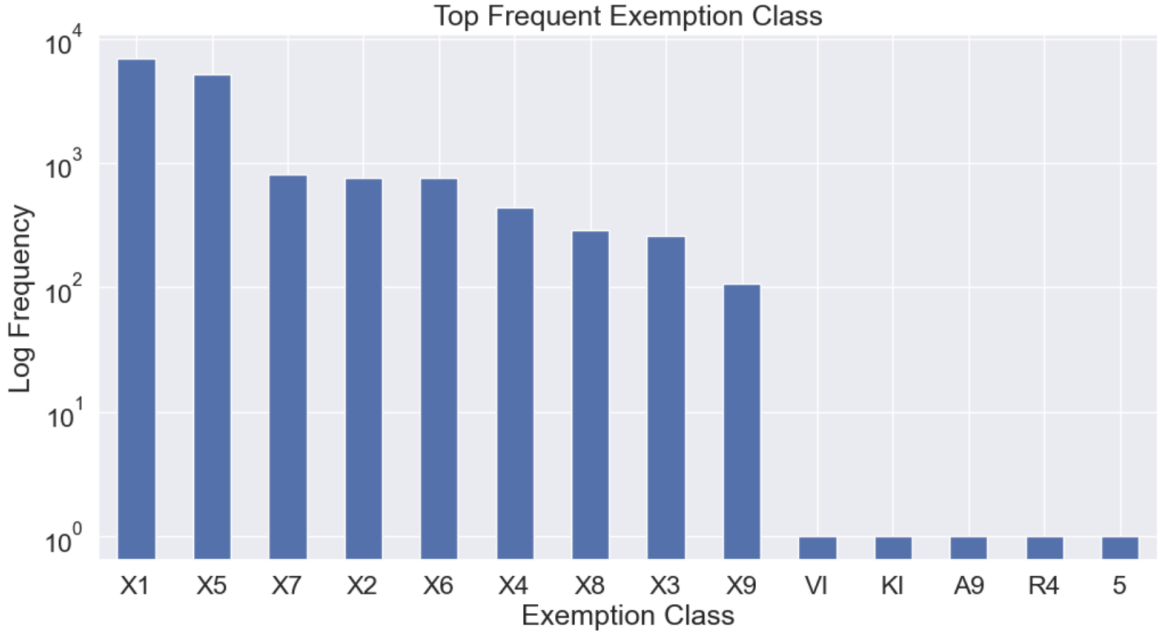
27) Field Name: ZIP

Description: a categorical field, the zip code of the property. The histogram shows the frequency of top 20 field values of ZIP. The most common value is 10314.0count of which is 24606. There are 29,890 zeros in this field, indicating that about 3% of the properties do not have a zip code listed. There are 196 unique values.



28) Field Name: EXMPTCL

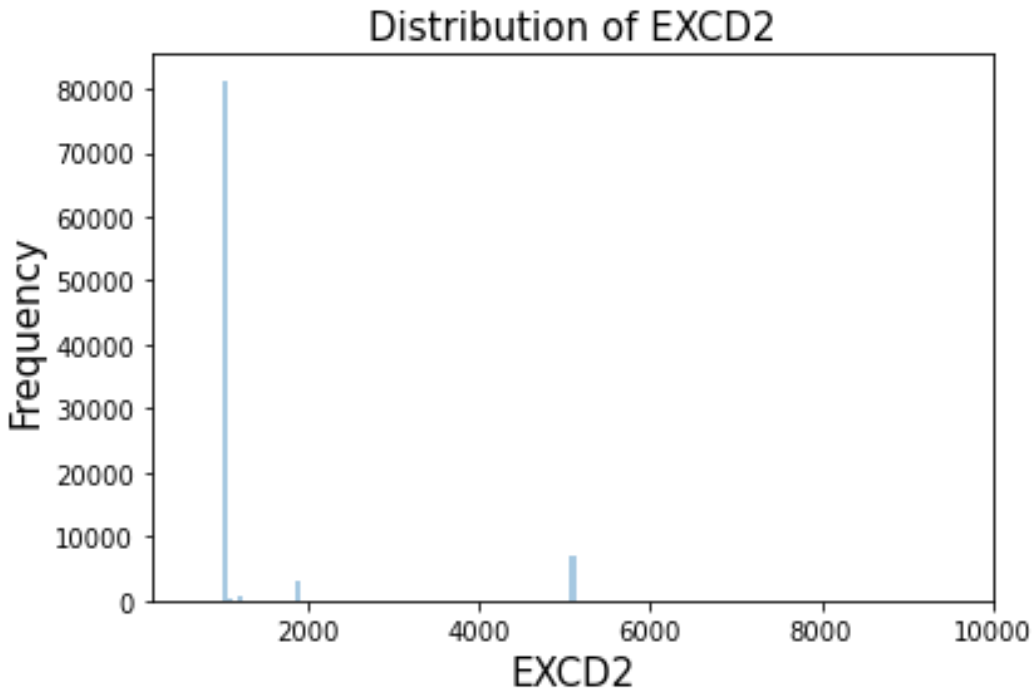
Description: a categorical field, a code indicating a type of exemption from property taxes, such as for religious or charitable organizations. The histogram shows the frequency of top 20 field values of Exemption Class. The most common value is X1, count of which is 6912. There are 1,055,415 zeros in this field, indicating that almost all properties do not have an exemption code listed. There are 14 unique values.



29) Field Name: EXCD2

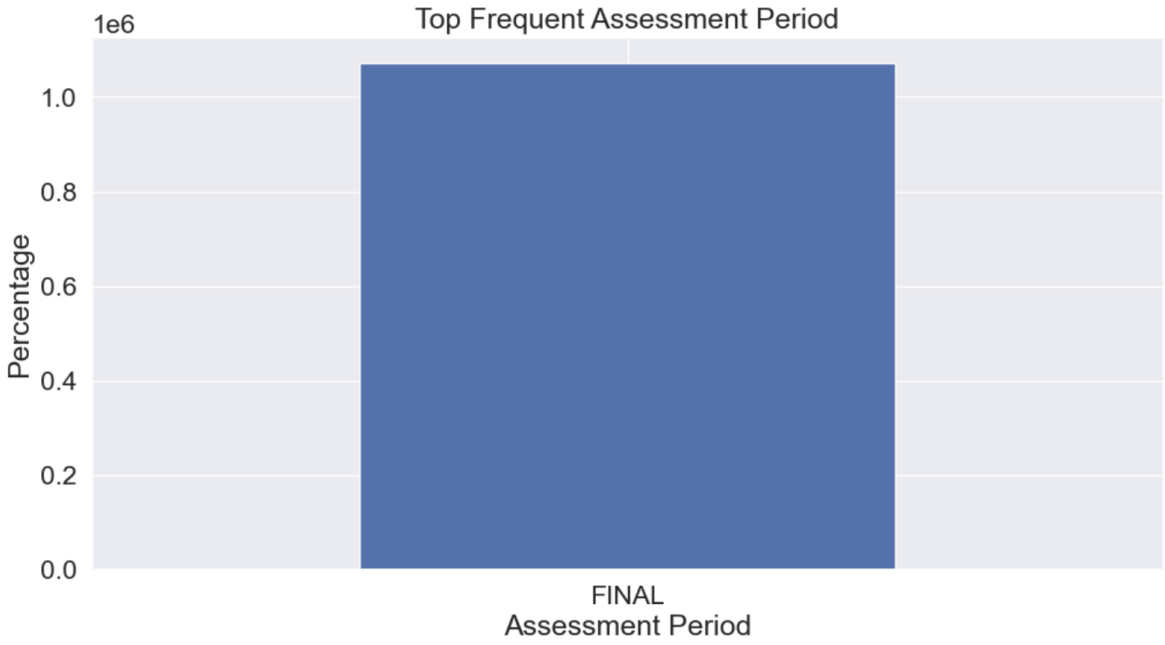
Description: a categorical field, a second code indicating a type of exemption or abatement from property taxes. There are 941,046 zeros in this field, indicating that about 88% of the properties do not have

a second exemption code listed. There are 60 unique values. The most common values is 1017.



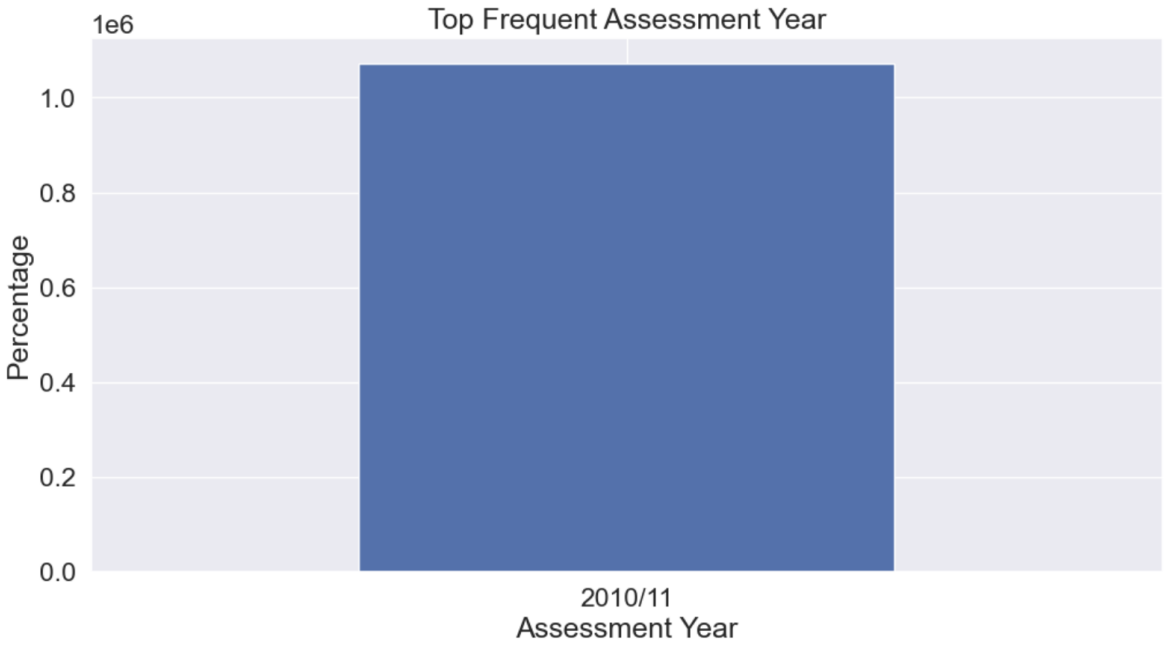
30) Field Name: PERIOD

Description: a categorical field, the time period for which the property assessment data is valid. The histogram shows the frequency of top field values of PERIOD. The most common value is FINAL, count of which is 1070994. There are no zeros in this field and there is only one unique value ("FINAL").



31) Field Name: YEAR

Description: a categorical field, representing the year of the data. The histogram shows the frequency of top field values of Year. The most common value is 2010/11, count of which is 1070994.



32) Field Name: VALTYPE

Description: a categorical field, representing the valuation type of the data (e.g., AC-TR, AC-EXT). The histogram shows the frequency of top field values of VALTYPE. The most common value is AC-TR, count of which is 1070994