Unsupervised Tax Fraud Detection

1. Executive Summary

The project aims to build an unsupervised tax fraud detection model with the dataset**,** Property Valuation and Assessment Data, and find people misrepresenting their property characteristics and underpaying tax. For the project, data cleaning and imputation are conducted. Then, we think about the project goal to build variables that can better estimate a property's value and suggest abnormal properties. 93 new variables are created in this process based on the original field, and the less relevant fields are dropped. Given the large number of features in the dataset, PCA is implemented to reduce dimensionality and 5 main PCs are kept. After the dimensionality reduction, to find the anomality in the dataset, two scoring methods are implemented. Two additional columns are created to store the rank of each record in the dataset when the dataset is sorted in a descending order according to the two scores respectively. Then, the final fraud score is computed and the records with a high fraud score are considered as properties with high probability of having a fraud, which will be passed to experts for analysis and feedback so that the model can be further improved.

2. Description of Data

Basic Description

The dataset, Property Valuation and Assessment Data, is retrieved from NYC open data platform. It is formed with real estate assessment property data, which can be used for property tax fraud detection. It has 1,070,994 rows and 32 fields. 19 out of 32 fields have no null data, and 13 fields have null values.

Field Summary Table

1. Numerical Table

	Field Name	# Records Have Values	% Populated	# Zeros	Min	Max	Most Common
0	LTFRONT	1,070,994	100.00%	169,108	0	9,999	0
1	LTDEPTH	1,070,994	100.00%	170,128	0	9,999	100
2	STORIES	1,014,730	94.75%	0	1	119	2
3	FULLVAL	1,070,994	100.00%	13,007	0	6,150,000,000	0
4	AVLAND	1,070,994	100.00%	13,009	0	2,668,500,000	0
5	AVTOT	1,070,994	100.00%	13,007	0	4,668,308,947	0
6	EXLAND	1,070,994	100.00%	491,699	0	2,668,500,000	0
7	EXTOT	1,070,994	100.00%	432,572	0	4,668,308,947	0
8	BLDFRONT	1,070,994	100.00%	228,815	0	7,575	0
9	BLDDEPTH	1,070,994	100.00%	228,853	0	9,393	0
10	AVLAND2	282,726	26.40%	0	3	2,371,005,000	2,408
11	AVTOT2	282,732	26.40%	0	3	4,501,180,002	750
12	EXLAND2	87,449	8.17%	0	1	2,371,005,000	2,090
13	EXTOT2	130,828	12.22%	0	7	4,501,180,002	2,090

2. Categorical Table

	Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	RECORD	1,070,994	100.00%	0	1,070,994	N/A
1	BBLE	1,070,994	100.00%	0	1,070,994	N/A
2	BORO	1,070,994	100.00%	0	5	4
3	BLOCK	1,070,994	100.00%	0	13,984	3944
4	LOT	1,070,994	100.00%	0	6,366	1
5	EASEMENT	4,636	0.43%	0	12	E
6	OWNER	1,039,249	97.04%	0	863,347	PARKCHESTER PRESERVAT
7	BLDGCL	1,070,994	100.00%	0	200	R4
8	TAXCLASS	1,070,994	100.00%	0	11	1
9	EXT	354,305	33.08%	0	3	G
10	EXCD1	638,488	59.62%	0	129	1017
11	STADDR	1,070,318	99.94%	0	839,280	501 SURF AVENUE
12	ZIP	1,041,104	97.21%	0	196	10314
13	EXMPTCL	15,579	1.45%	0	14	X1
14	EXCD2	92,948	8.68%	0	60	1017
15	PERIOD	1,070,994	100.00%	0	1	FINAL
16	YEAR	1,070,994	100.00%	0	1	2010/11
17	VALTYPE	1,070,994	100.00%	0	1	AC-TR

3. Data Cleaning and Imputation Logic

Exclusions

There are many properties that we really aren't interested in. Since we are looking for private owners that commit tax fraud, and many of the records in the dataset are government-owned properties, we create a remove list that contains 20 most seen property owners. The properties with an owner in the remove list are regarded as benign properties and are dropped. After we remove the exclusions, we proceed to fill in the missing values.

Imputations

Zip

There are 21,537 missing values for the column *Zip* in the dataset. We assume that the dataset is already sorted by zip value, and we impute if the neighbors of the row with missing zip value have the same zip value. For those that are still not filled using the logic above, we impute them with the previous zip value.

AVTOT, AVLAND, FULLVAL

We first calculate the mean values for *AVTOT, AVLAND, FULLVAL* within different taxclasses, avoiding the records with zeros, and then we substitute the inappropriate values in these fields with their averages within different taxclasses.

STORIES

We first calculate the mean values for *STORIES* within different taxclasses, and then we substitute the decent values accordingly.

LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT

We first calculate the mean values for *LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT* within different taxclasses, and then we substitute the decent values accordingly.

4. Variable Creation

Description of Variable	# Variables Created
r1-r9 variables: Normalize the monetary value fields *FULLVAL, AVLAND*, *AVTOT* by the field *Lot Size, Building Size* and *Building Volume* respective. This will create 9 more variables. And then take the inverse of these variables to create 9 more variables. By narrowing the distribution, we make the extreme values more likely to stand out, and thus it is easier to cluster based on these variables.	18
Grouped average Variables: Average the 18 *r1-r9* variables grouping by *Zip5, Taxclass, Zip3*, *Boro*. Divide each of the 18 ratio variables by the four scale factors from these groupings. By narrowing the distribution, we make the extreme values more likely to stand out, and thus it is easier to cluster based on these variables.	72
Value Measure Ratio Variables: Comparing the 3 monetary value fields: FULLVAL / (AVLAND+AVTOT), normalize to the mean, and get the max value of (VRnorm, 1/VRnorm). By narrowing the distribution, we make the extreme values more likely to stand out, and thus it is easier to cluster based on these variables.	1
Exemption Dummy Variables: Set the value to be 1 if exemption value equals actual value, e.g. EXLAND=AVLAND, AVTOT=EXTOT. Otherwise, set the value to be 0. If they are not equal, it is likely that the someone is lying.	2

The logic for the new variable creation is attached below.

```
data['land_equal']=np.nan
data['tot_equal']=np.nan
for i in range(len(data)):
    if data['AVLAND'][i]==data['EXLAND'][i]:
        data['land_equal'][i]=1
    else:
        data['land_equal'][i]=0
    if  data['AVTOT'][i]==data['EXTOT'][i]:
        data['tot_equal'][i]=1
    else:
        data['tot_equal'][i]=0
```

5. Dimensionality Reduction

After variable creation, 93 new variables exist in the dataset. Given the large dimensionality in the dataset, dimensionality reduction is introduced to solve this problem. Since we do not have a dependent variable, we z-scale all the features by subtracting the mean value of each feature for each record and dividing the result by the standard deviation of that feature. Then we implement a PCA, plot the cumulative variance and decide the optimal n_components for PCA should be 5 based on the change of gradient of the graph. At last we redo the PCA with n_components set to be 5, keep the top PCs, and z-scale again so that each retained PC is equally important.

6. Anomaly Detection Algorithm

Since fraud is unusual, we believe outliers can be regarded as fraud and thus we wish to find the outliers. Although there are multiple ways to find outliers, we generally use two method and combine the results to decide whether a datapoint should be considered as a fraud.

The first method is using Zscores to detect outliers. After we complete the PCA with n_components equal to 5 and z-scale again, for each record we add up the value of the z-scaled

variables, which are also known as Zscores, without letting them cancel each other out by introducing the Minkowski distance. The scoring formula is:

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

where z_n^i is the n-th Zscore for the i-th record, p is the pre-determined power for the Minkowski distance, and s_i is the score for i-th record.

The second method is autoencoder, which is a model trained to output the original vector input. It can help to detect anomaly because during training, the autoencoder algorithm adjusts itself to minimize the error on the entire set of records, and it learns how to reproduce the main bulk of records fairly well. The records that are not reproduced well can be regarded as the outliers we are looking for. A neural net is used for this project. And the fraud score is the reconstruction error with a formula:

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

where $z_n'^i$ is the n-th output of the autoencoder for the i-th record, z_n^i is the original value for the i-th record, p is the pre-determined power for the Minkowski distance, and s_i is the score for i-th record.

Then we create two columns, score1 rank and score2 rank, that store the rank of each record based on its computed score from method 1 and method 2 respectively. The final fraud score for each record is the average of score1 rank and score2 rank.

7. Result

After we create the final fraud score for each record, the dataset is sorted in a descending order based on the fraud score. The records with a high fraud score are considered as properties with high probability of having a fraud. To examine the results of the unsupervised model, we look at the z-scaled variables of the top records, as Zscores reflect many standard deviations from the

mean and thus we can immediately see which variables have unusual values. In addition, we make heatmaps of z-scaled variables to see which variables are driving the high scores and further investigate some properties we are interested in, which are listed below.

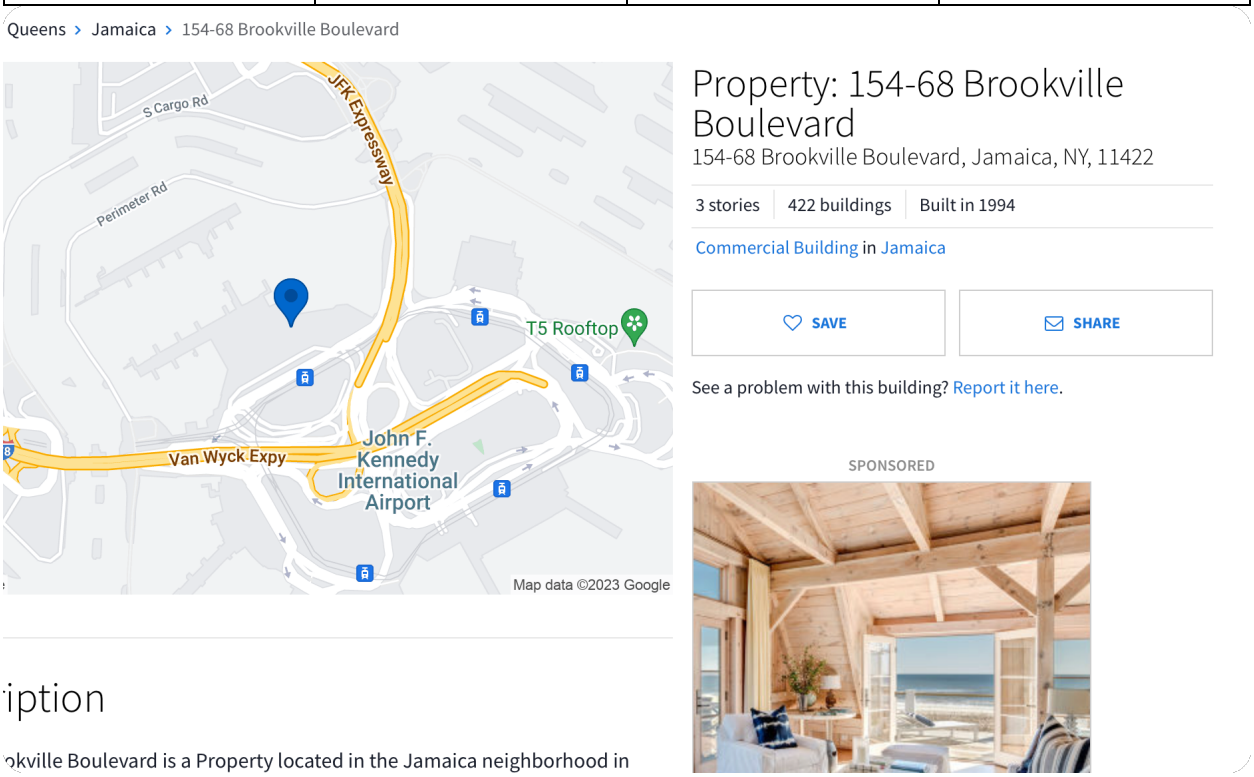
Five Interesting Case Studies

1. Unusual NY Property 1

Record 917942

LOGAN PROPERTY, INC.

OWNER	LOGAN PROPERTY, INC.	LTFRONT	4910
ADDRESS	154-68 BROOKVILLE BOULEVARD	LTDEPTH	0
FULLVAL	374,019,883	BLDFRONT	0
AVLAND	1,792,808,947	BLDDEPTH	0
AVTOT	4,668,308,947	STORIES	3



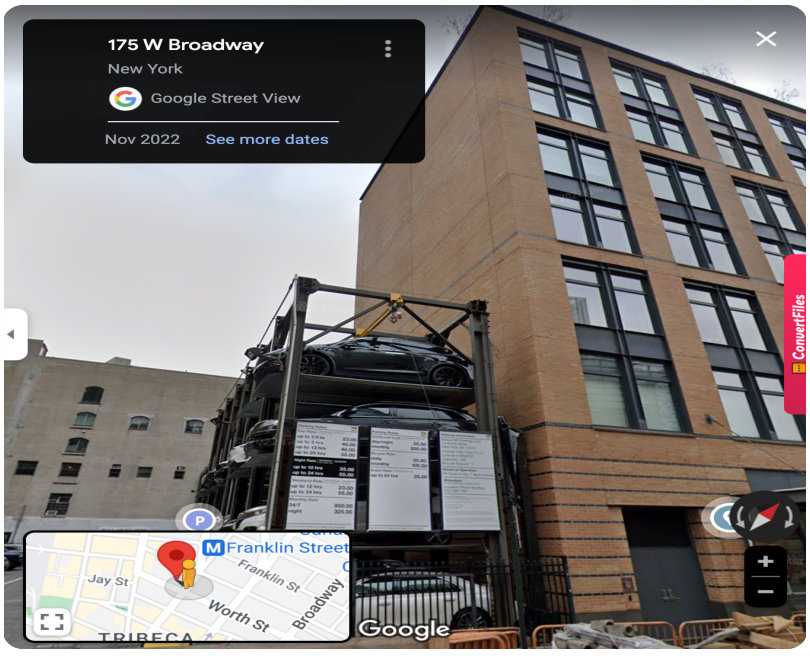
The property is quite large with a building size of 14,745,183 square feet, and has non-zero values for LTFEPTH, BLDFRONT and BLDDEPTH, indicating that it is not an empty lot or undeveloped land.

The high Z-scores for variables such as BLDFRONT, BLDDEPTH, FULLVAL, AVLAND, and AVTOT suggest that the assessed values for these variables are significantly higher than the average or expected values for properties in the area. This could be due to various factors such as the size, location, condition, or intended use of the property.

2. Unusual NY Property 2

Record 12076

OWNER	15 WORTH STREET PROPE	LTFRONT	74
ADDRESS	170 WEST BROADWAY	LTDEPTH	150
FULLVAL	2,610,000	BLDFRONT	5
AVLAND	1,170,000	BLDDEPTH	5
AVTOT	1,174,500	STORIES	1



170 WEST BROADWAY is a well-known street located in the Tribeca neighborhood of New York City, known for its historic buildings and trendy urban vibe. Instead of having one story, the building has 6 stores, resulting in extreme values in r3 and r6.

3. Unusual NY Property 3

Record 665158

OWNER	BREEZY POINT COOPERAT	LTFRONT	2,798
ADDRESS	217-02 BREEZY POINT BLVD217-02 BREEZY POINT BLVD	LTDEPTH	997
FULLVAL	273,000,000	BLDFRONT	30
AVLAND	10,920,000	BLDDEPTH	40
AVTOT	16,380,000	STORIES	1

Located in the Breezy Point neighborhood in Queens, 217-02 Breezy Point Boulevard is a house with high r2 and r3 ratings, but a relatively low r6 rating, which suggests a potential issue

with FULLVAL. Propertyshark.com estimates the assessed value of the property to be around $18,566,231, much lower than $273,000,000.

4. Unusual NY Property 4

Record 658933

OWNER	WAN CHIU CHEUNG	LTFRONT	25
ADDRESS	54-76 83 STREET	LTDEPTH	100
FULLVAL	776,000	BLDFRONT	2500
AVLAND	26,940	BLDDEPTH	5600
AVTOT	46,560	STORIES	3

C	CD	CE	CF	CG	CH	CI	CJ	CK
axcla	2inv_taxcla	3inv_taxcla	4inv_taxcla	5inv_taxcla	6inv_taxcla	7inv_taxcla	8inv_taxcla	9inv_taxcla
0	0	0	-1	0	-1	0	0	0
0	0	0	-1	0	-1	0	0	0
0	0	0	-1	0	-1	0	0	0
0	0	0	-1	0	-1	0	0	0
0	0	0	0	0	-1	0	0	0
0	0	0	-1	0	-1	0	0	0
0	0	0	0	0	-1	0	0	0
0	763	775	0	441	348	0	473	423
0	0	0	-1	0	-1	0	0	0
0	0	0	-1	0	-1	0	0	0
0	0	0	-1	0	-1	0	0	0
0	0	0	-1	0	-1	0	0	0
0	563	620	0	434	346	0	436	410
0	0	0	-1	0	-1	0	0	0

It is an apartment with 3 stores. The BLDFRONT and BLDDEPTH are considerably large, resulting negative r2 and r3. In addition, it has extreme value for r2inv_taxclass, r3inv_taxclass, r5inv_taxclass, r6inv_taxclass, r8inv_taxclass, r9inv_taxclass have extreme values. All of them are relevant to S2, which is BLDFRONT * BLDDEPTH.

Source: https://www.redfin.com/NY/Flushing/5476-83rd-St-11373/home/20904778

5. Unusual NY Property 5

Record: 106681

OWNER	79TH REALTY LLC	LTFRONT	25
ADDRESS	350 EAST 79 STREET	LTDEPTH	100
FULLVAL	114,000,000	BLDFRONT	25
AVLAND	33,750,000	BLDDEPTH	100
AVTOT	51,300,000	STORIES	44

3	86	44	2	194	36	2	102	22	1

Sophisticated and elegant, The Lucerne has a large selection of family-sized homes of up to 3- and 4-bedrooms and duplexes — a rarity in Manhattan. The layouts flow beautifully from room to room, with nine-foot ceilings and enough private spaces to accommodate everyone. For such a large building, the LTFRONT and LTFEPTH are relatively small, resulting a large z score for all the rs, especially r1, r4, r7. Because r1,r4, and r7 are related to LTFRONT and LTDEPTH
Source: https://streeteasy.com/building/the-lucerne

6. Unusual NY Property 6

Record: 95995

OWNER	BERGAMINI, JENNIFER BERGAMINI, JENNIFER	LTFRONT	197
ADDRESS	724 1 AVENUE	LTDEPTH	378
FULLVAL	17,354,800	BLDFRONT	15
AVLAND	7,785,000	BLDDEPTH	20
AVTOT	7,809,660	STORIES	1

The building has more than 20 stories instead of one. It has a high r3 and r6, which are indicators of something wrong with stories.

Source: https://www.propertyshark.com/mason/Property/21906/724-1-Ave-New-York-NY-10017/

8. Summary

This project aims to build an unsupervised tax fraud detection model with the dataset**,** Property Valuation and Assessment Data, to find people misrepresenting their property characteristics and underpaying tax. Since we are looking for private owners that commit tax fraud, and many of the records in the dataset are government-owned properties, we create a remove list that contains 20 most seen property owners. The properties with an owner in the remove list are regarded as benign properties and are dropped. In addition to exclusion removal, missing values are also imputed. Then, we think about the project goal to build variables that can better estimate

a property's value and suggest abnormal properties. We decide to focus on measurements like dollars per square foot for the land and dollars per building volume. In this case, 93 new variables are created based on the original field, and the less relevant fields are dropped.

Given the large number of features in the dataset, PCA is implemented to reduce dimensionality. We z-scale all the features by subtracting the mean value of each feature for each record and dividing the result by the standard deviation of that feature. Then we implement a PCA, plot the cumulative variance and decide the optimal n_components for PCA should be 5 based on the change of gradient of the graph. At last, we redo the PCA with n_components set to be 5, keep the top PCs, and z-scale again so that each retained PC is equally important.

After the dimensionality reduction, to find the anomality in the dataset, two scoring methods, Zscores and the distance between the output of an autoencoder with a neural net and the original value, are implemented. Two additional columns are created to store the rank of each record in the dataset when the dataset is sorted in a descending order according to the two scores respectively. Then, the final fraud score is computed and the records with a high fraud score are considered as properties with high probability of having a fraud. In addition, we make heatmaps of z-scaled variables to see which variables are driving the high scores and further investigate some properties we are interested in. Still, this is the initial edition of the unsupervised model, and results should be passed to experts for analysis. After we receive their feedback, we can adjust the model by improving exclusions, creating better variables, and modifying the parameters and hyperparameters of the supervised model and scoring method, including the value of n_components for the PCA, the power we use for the Minkowski distance, and the shape of the neural net. After modification, we can send the new results to experts and ask their feedback again. By several iterations of model improvements, we can finalize our model and deploy it in real life business setting.

Appendix – Data Quality Report

The dataset has 32 different fields. The description of and the visualization of each field can be found below.

1. *Record*

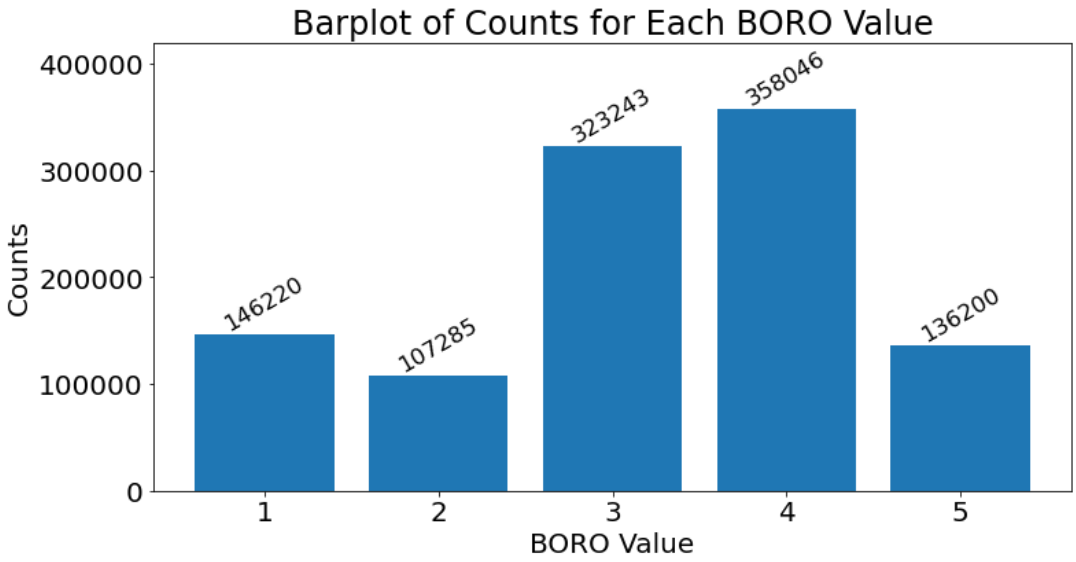
The field *Record* can be regarded as the index. It is ordinal unique positive integer for each record, from 1 to 1,070,994.

2. *BBLE*

The field *BBLE* is the file key. It is concated by fields *BORO, BLOCK, LOT* and *EASEMENT* code. It has 1,070,994 unique records and no missing value.

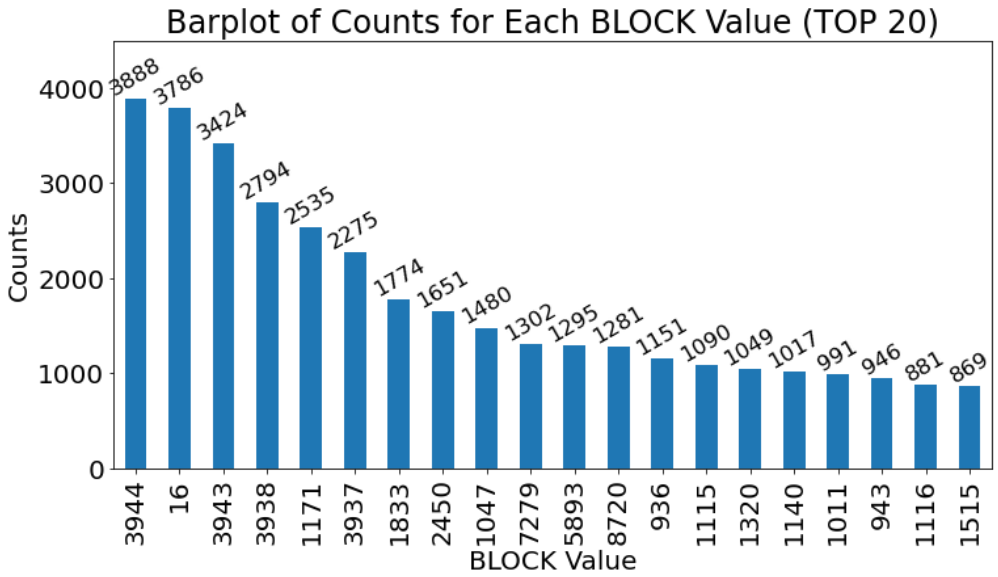
3. *BORO*

The field *BORO* contains the information of the Borough the datapoint is located in. It is a categorical variable with 5 unique positive integers as values, where 1 stands for Manhattan, 2 stands for Bronx, 3 stands for Brooklyn, 4 stands for Queens, and 5 stands for Staten Island. It has 1,070,994 records and no missing value. The visualization for the number of counts for the 5 values is attached below.



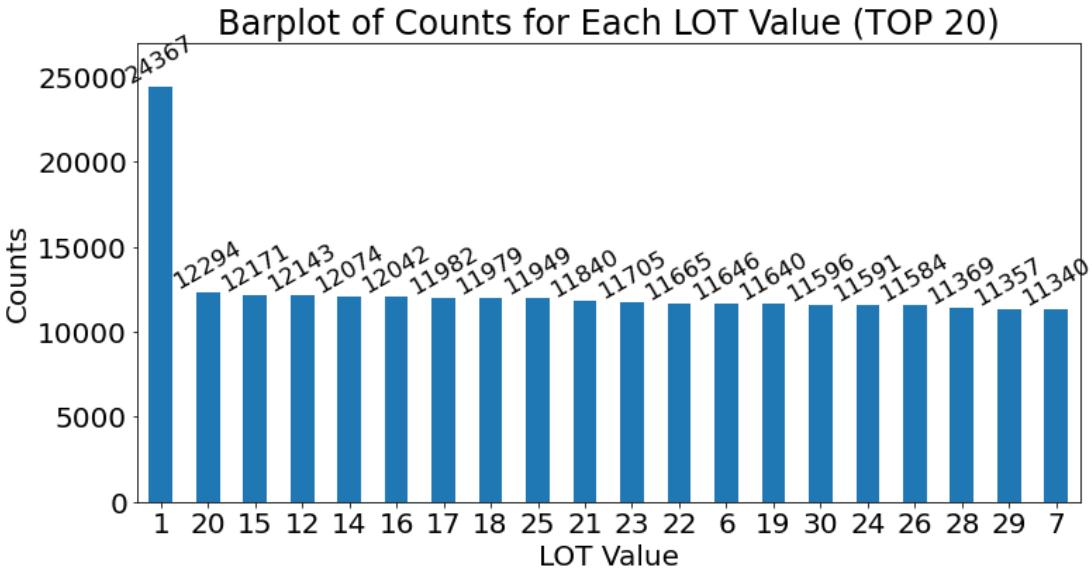
4. *BLOCK*

The field *BLOCK* contains the block number of the datapoint is assigned to. It is a categorical variable with 5 different ranges depending on the value of the field *BORO*. For datapoints with a *BORO* value of 1, the range is 1 to 2,255. For datapoints with a *BORO* value of 2, the range is 2,260 to 5,958. For datapoints with a *BORO* value of 3, the range is 1 to 8,955. For datapoints with a *BORO* value of 4, the range is 1 to 16,350. For datapoints with a *BORO* value of 5, the range is 1 to 8,050. It has 1,070,994 records and no missing values. The visualization for the number of counts for the top 20 most seen values is attached below.



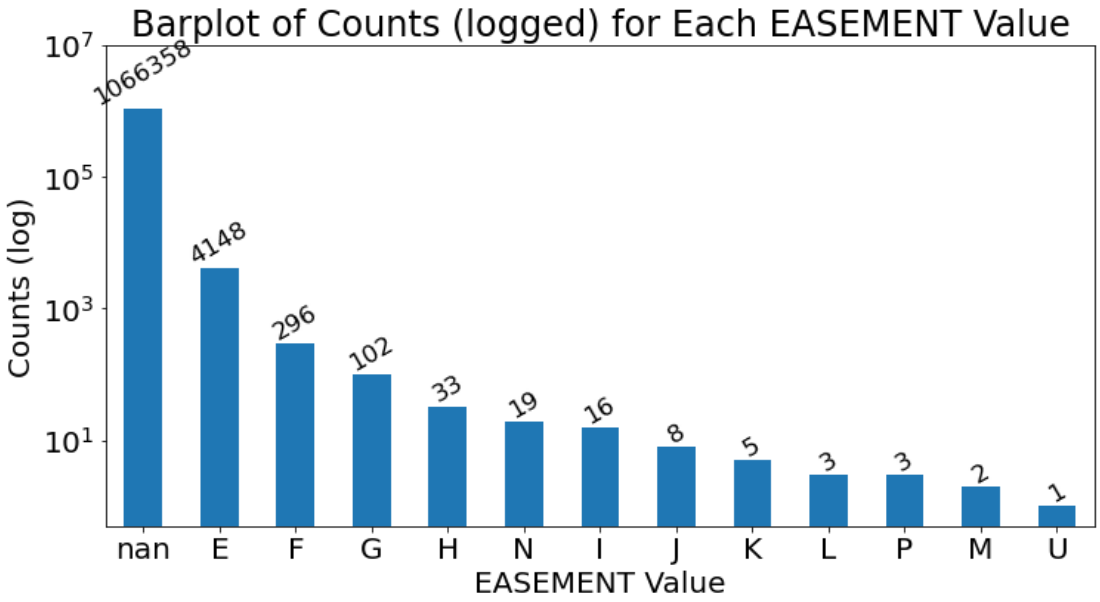
5. *LOT*

The field *LOT* contains the unique number within the corresponding block or boro. It has 6,366 unique values, 1,070,994 records and no missing values. The visualization of the counts for the top 20 most seen values is attached below.



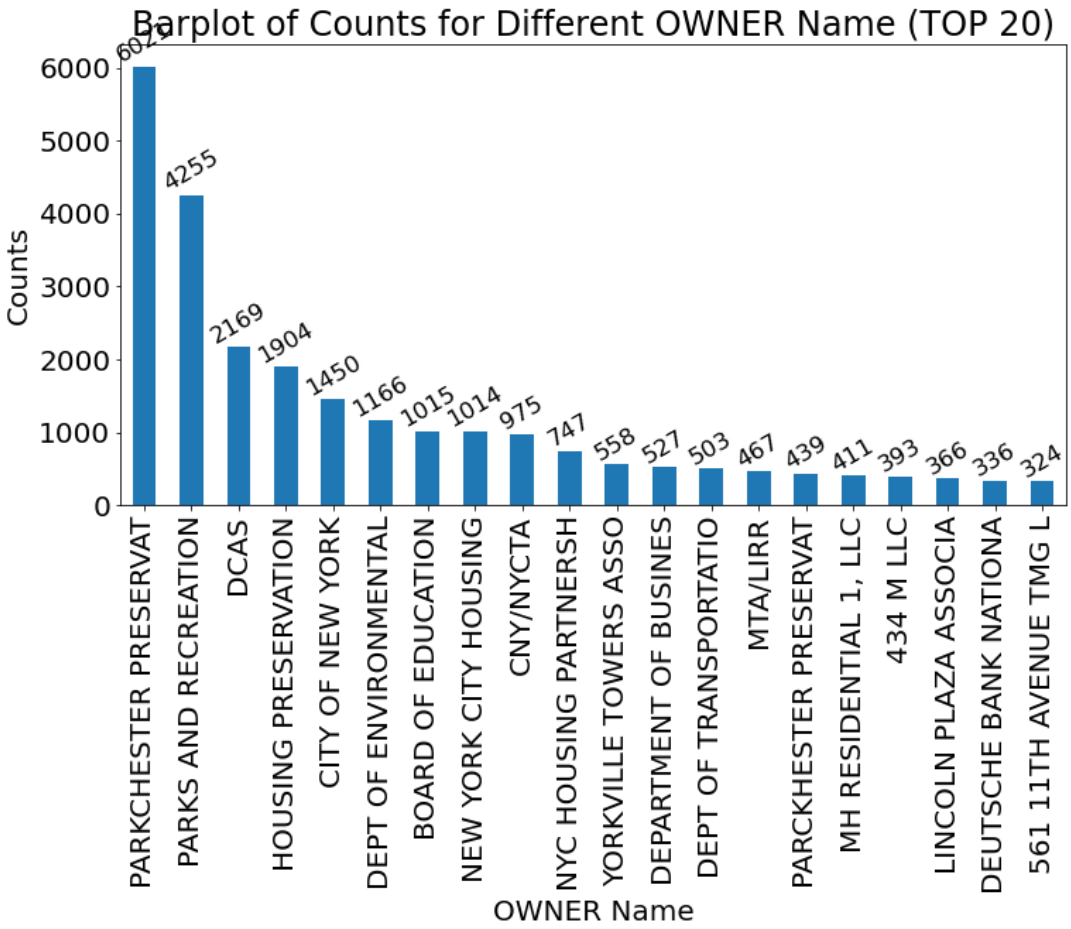
6. *EASEMENT*

The field *EASEMENT* is used to describe the easement of each datapoint. It is a categorical variable with 13 unique values including nan value, which means the datapoint has no easement according to the documentation. It has 4636 records and 1066358 nan value. The visualization for the number of counts for the top 20 most seen values is attached below.



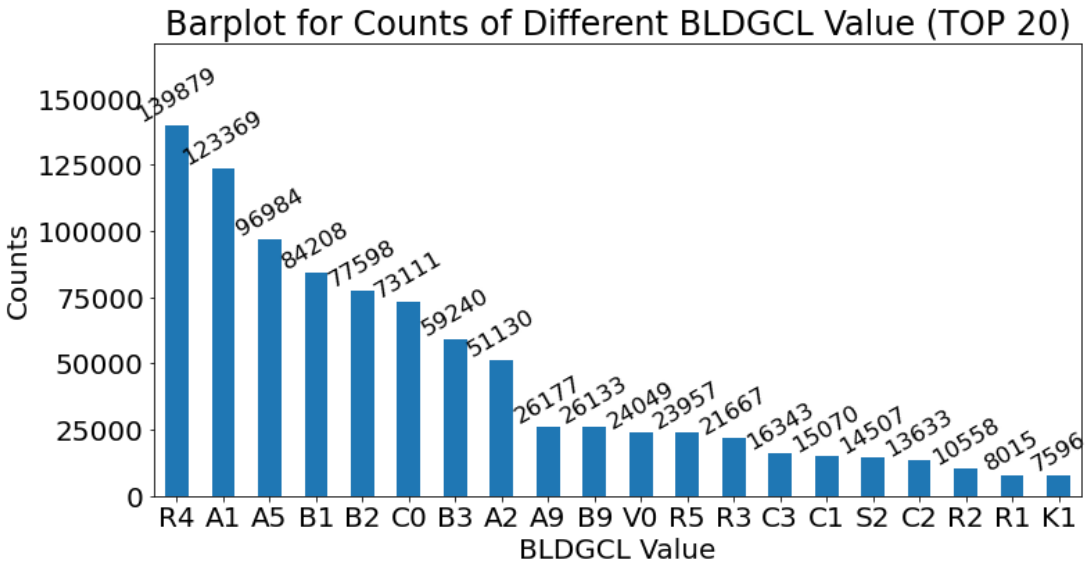
7. *OWNER*

The field *OWNER* stores the name of the owner. It is a categorical variable with 863348 unique values, 1039249 records and 31745 missing values. The visualization for the counts for top 20 most seen values is attached below.



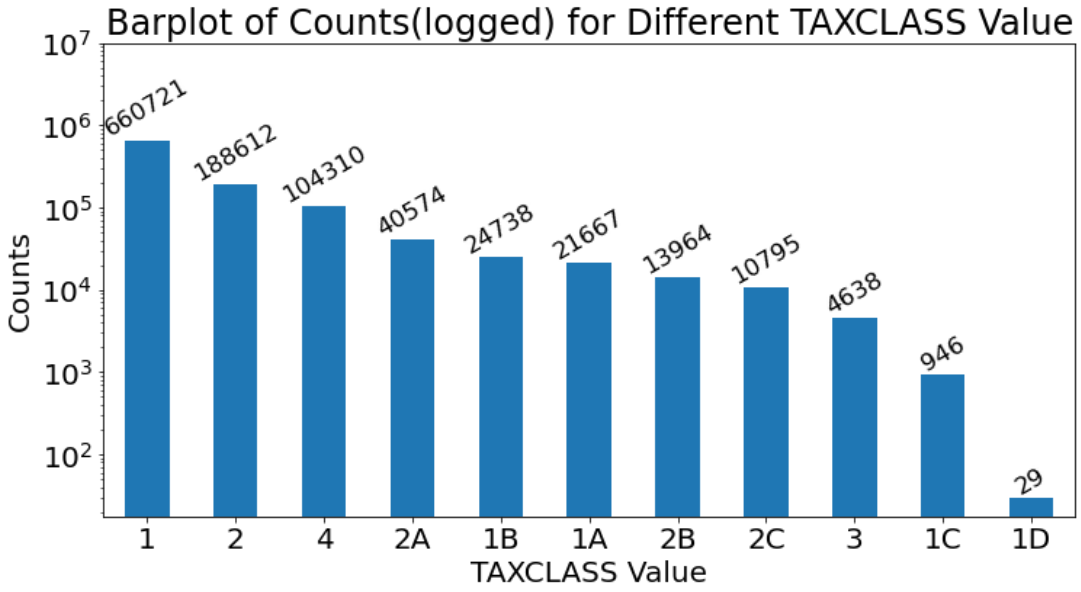
8. *BLDGCL*

The field *BLDGCL* stores the building class for each datapoint. It is a categorical variable with 200 unique values, 1070994 records and no missing value. The visualization for the counts of the top 20 most seen values is attached below.



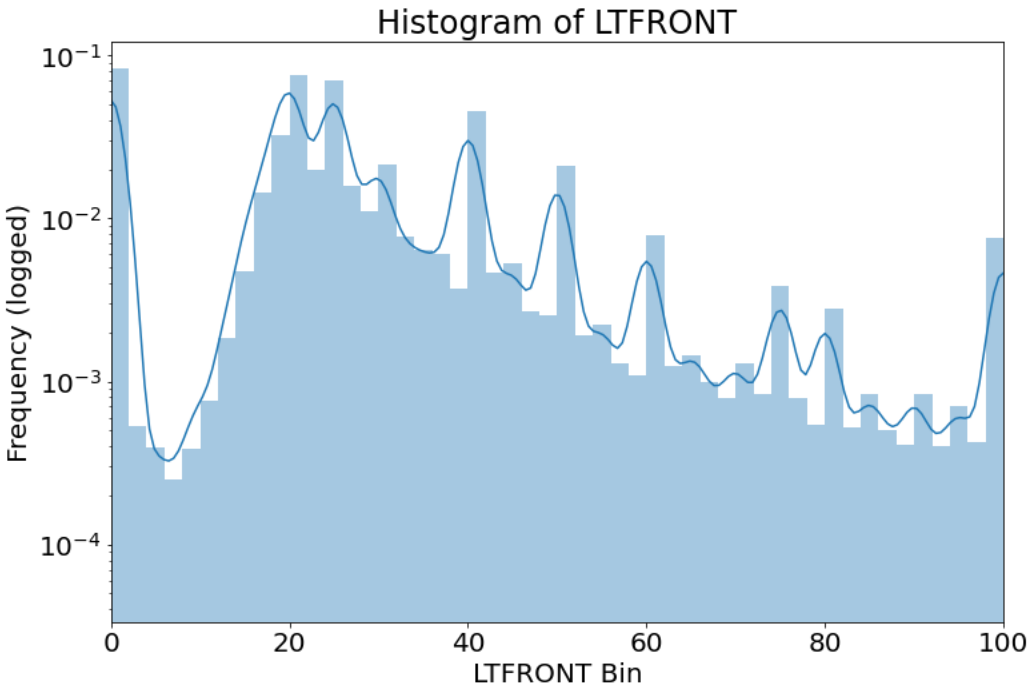
9. *TAXCLASS*

The field *TAXCLASS* stores the tax class for each datapoint. It is a categorical variable with 11 unique values, 1070994 records and no missing value. The visualization for the counts of different values is attached below.



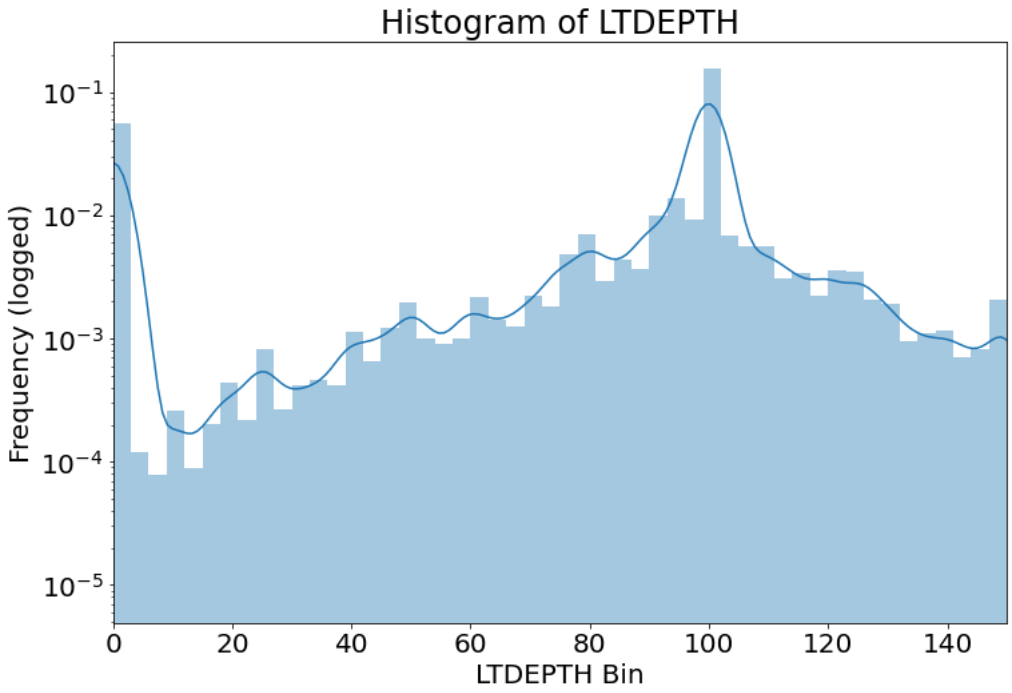
10. *LTFRONT*

The field *LTFRONT* stores the lot width of each datapoint. It is a numerical variable with 1,070,994 records and no missing values. The value of the field ranges from 0 to 9999, and 95% of it are located in the range 0-100. The histogram of *LTFRONT* with a value less than or equal to 100 is attached below.



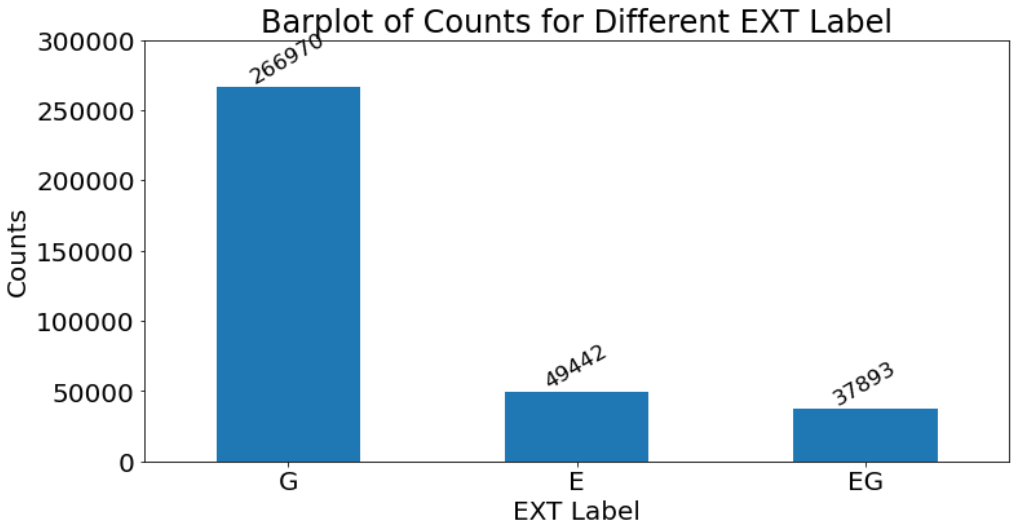
11. *LTDEP*

The field *LTDEP* stores the lot depth of each datapoint. It is a numerical variable with 1,070,994 records and no missing values. The value of the field ranges from 0 to 9999, and 96% of it are located in the range 0-150. The histogram of *LTDEP* with a value less than or equal to 150 is attached below.



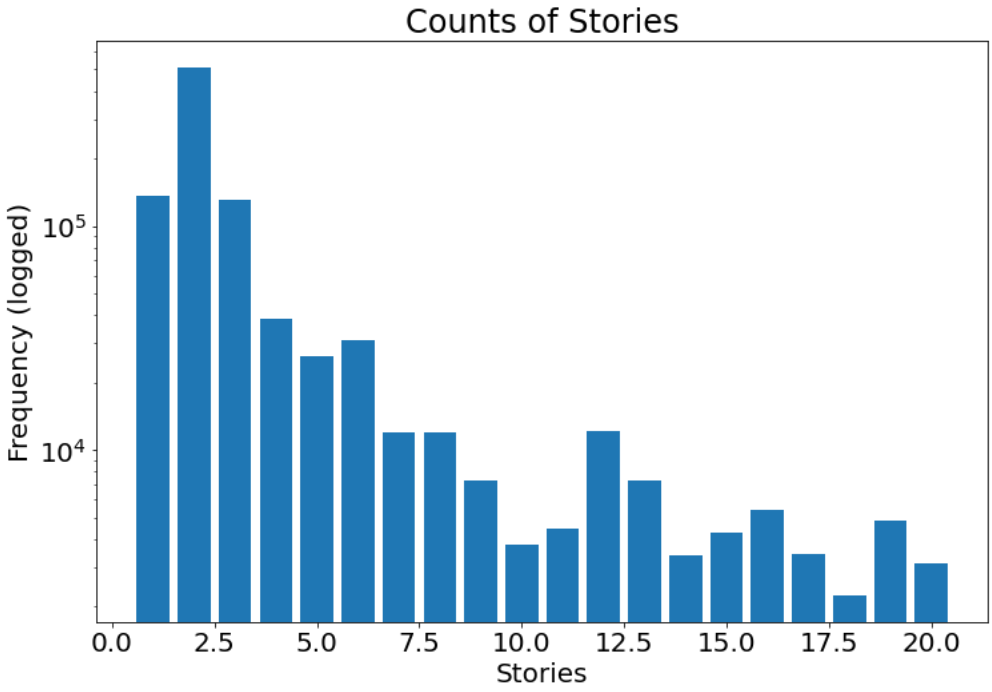
12. *EXT*

The field *EXT* stores the extension indicator for each datapoint. It is a categorical variable with 3 unique values, including “E”, “G” and “EG”, 354,305 records and 716,689 missing values. The value “E” stands for extension, the value “G” stands for garage, and the value “EG” stands for extension and garage. The visualization of the counts for different values is attached below.



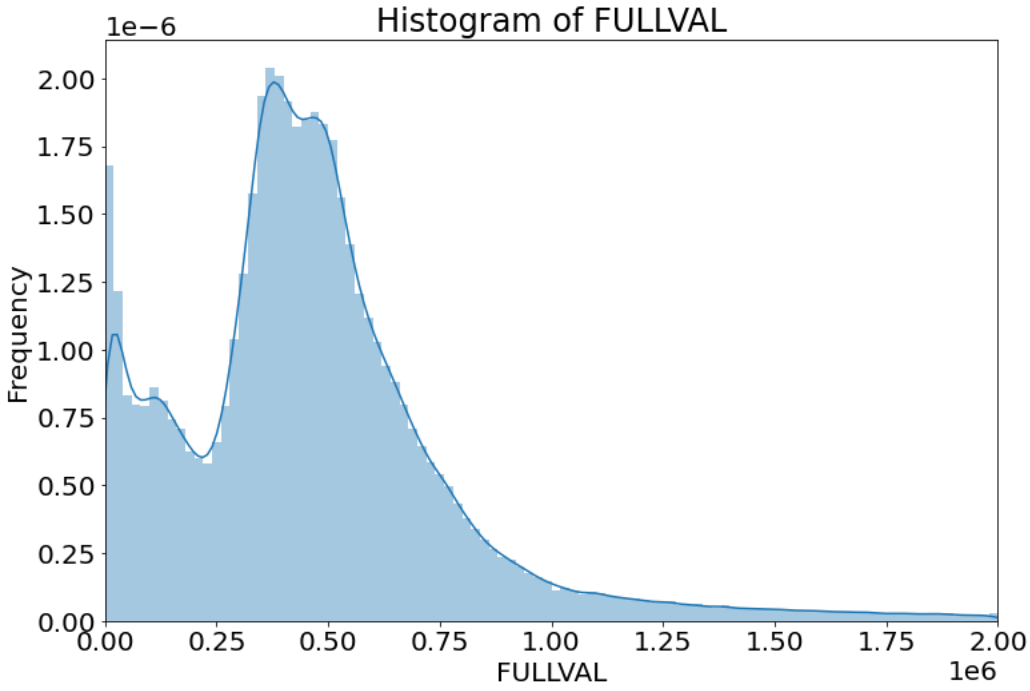
13. STORIES

The field *STORIES* stores the number of stories in building for each datapoint. It is a numerical variable with 1,014,730 records and 56,264 missing values. It ranges from 1 to 119, with 94.5% of its values locate in the range 1-20. The visualization of the counts for different values of *STORIE* with a value less than or equal to 20 is attached below.



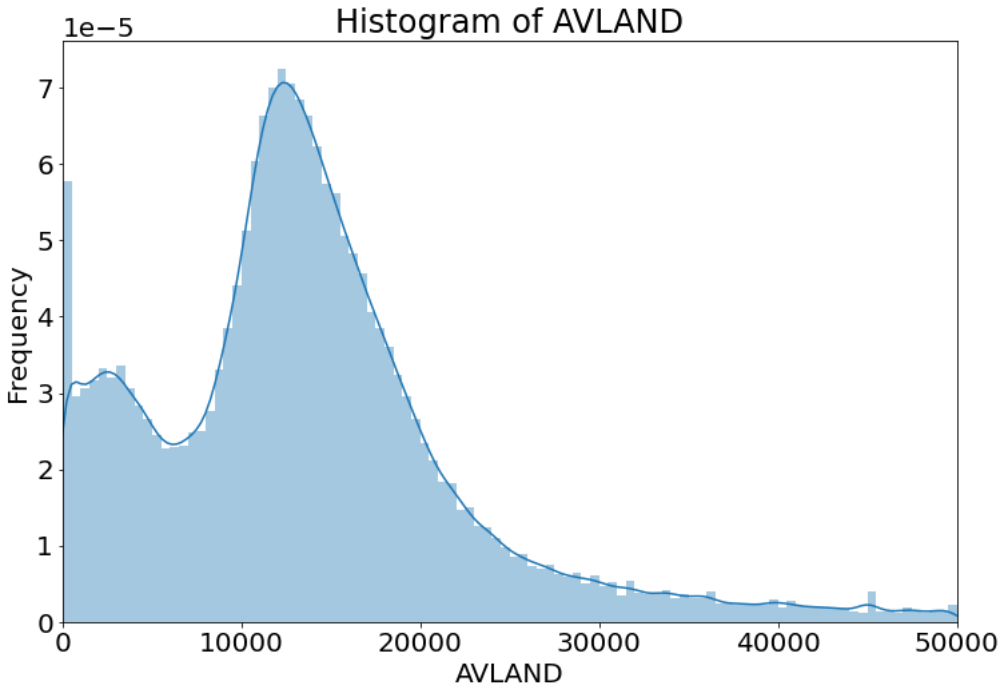
14. FULLVAL

The field *FULLVAL* stores Market Value of each datapoint. It is a numerical variable with 1,070,994 records and no missing values. It ranges from 0 to 6,150,000,000, with 96.3% of its values locate in the range 0 - 2,000,000. The histogram of *FULLVAL* with a value less than or equal 2,000,000 is attached below.



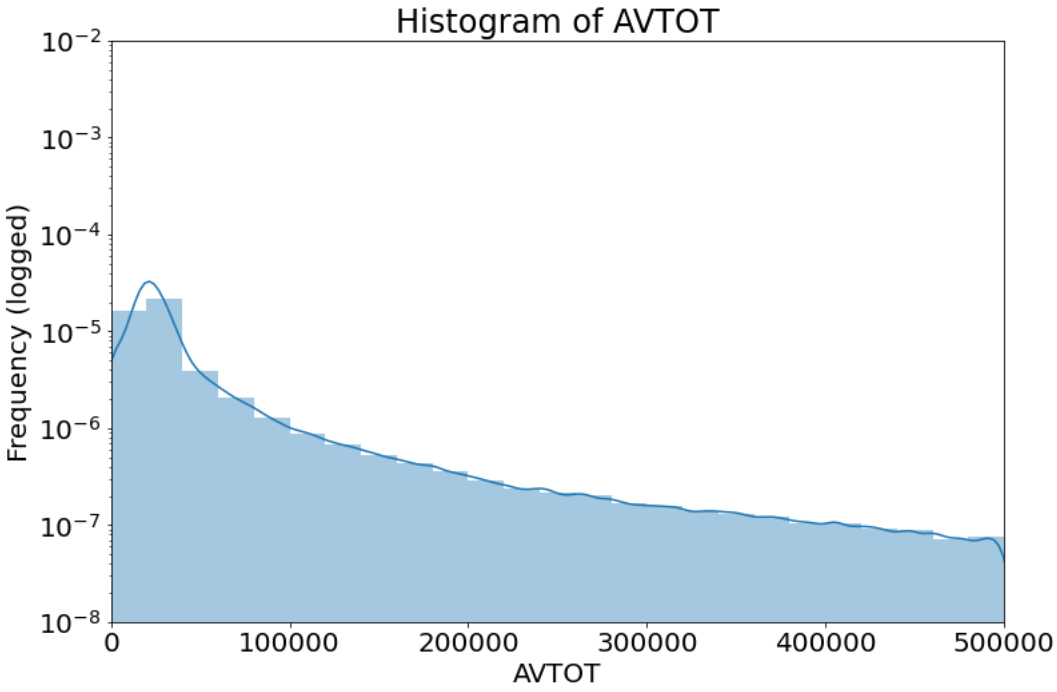
15. AVLAND

The field *AVLAND* stores the actual land value of each datapoint. It is a numerical variable with 1,070,994 records and no missing values. It ranges from 0 to 2,668,500,000, with 90% of its values locate in the range 0 to 50,000. The rest is evenly distributed in the range of 50,000 to 2,668,500,000. The histogram of *AVLAND* with a value less than or equal 50,000 is attached below.



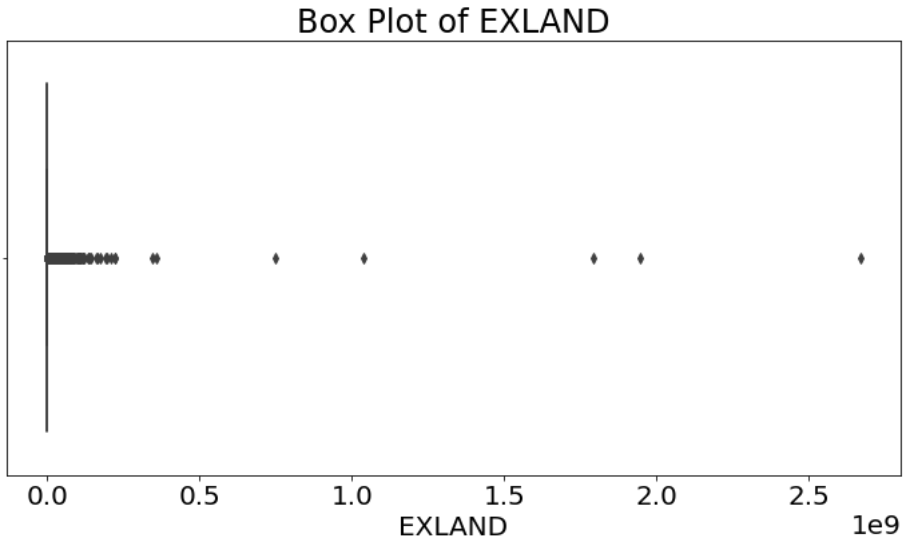
16. *AVTOT*

The field *AVTOT* stores the actual total value of each datapoint. It is a numerical variable with 1,070,994 records and no missing values. It ranges from 0 to 4,668,308,947, with 95% of its values locate in the range 0 to 500,000. The histogram of *AVTOT* with a value less than or equal 50,000 is attached below.



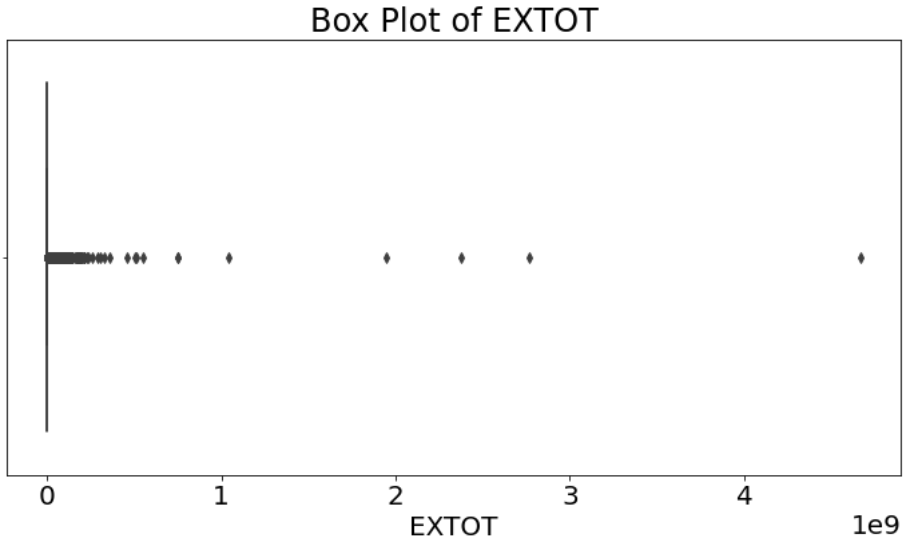
17. *EXLAND*

The field *EXLAND* stores the actual exempt land value of each datapoint. It is a numerical variable with 1,070,994 records and no missing values. It ranges from 0 to 2,668,500,000, and 85% of its values locate in the range 0 to 3,000. Among those values, 46% of its values equal 0, and 33.4% of its values equal 1620. The box plot of *EXLAND* is attached below.



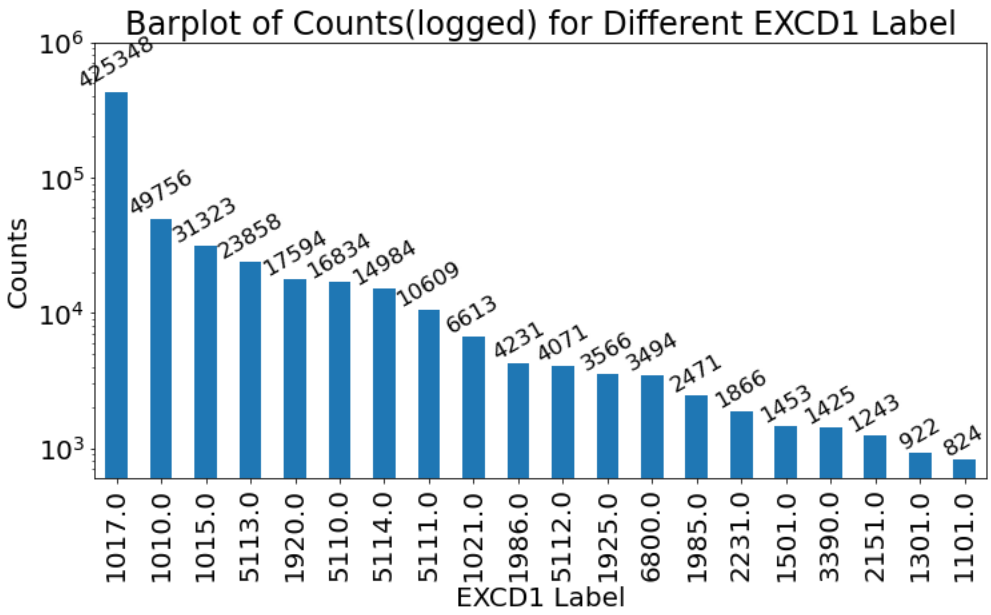
18. *EXTOT*

The field *EXTOT* stores the actual exempt land total of each datapoint. It is a numerical variable with 1,070,994 records and no missing values. It ranges from 0 to 4,668,308,947, and 85% of its values locate in the range 0 to 10,000. Among those values, 40% of its values equal 0, and 33% of its values equal 1620. The box plot of *EXTOT* is attached below.



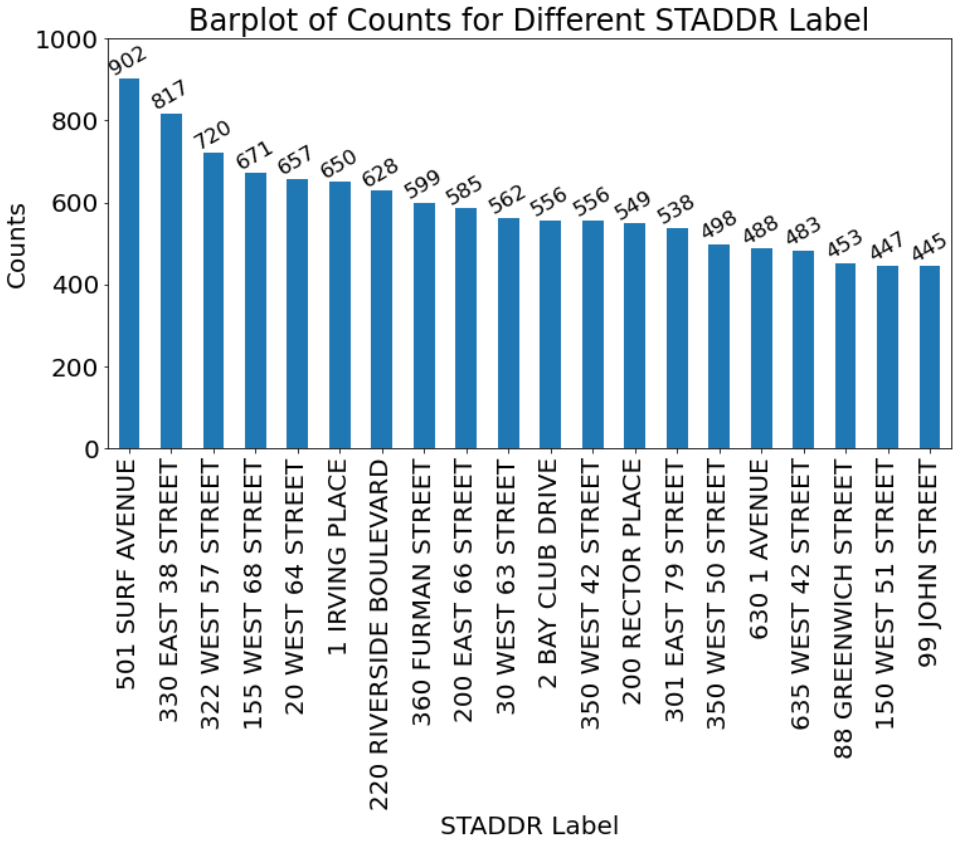
19. *EXCD1*

The field *EXCD1* stores the exemption code 1 of each datapoint. It is a categorical variable with 1,070,994 records and 432,507 missing values. The visualization for the counts for top 20 most seen values is attached below.



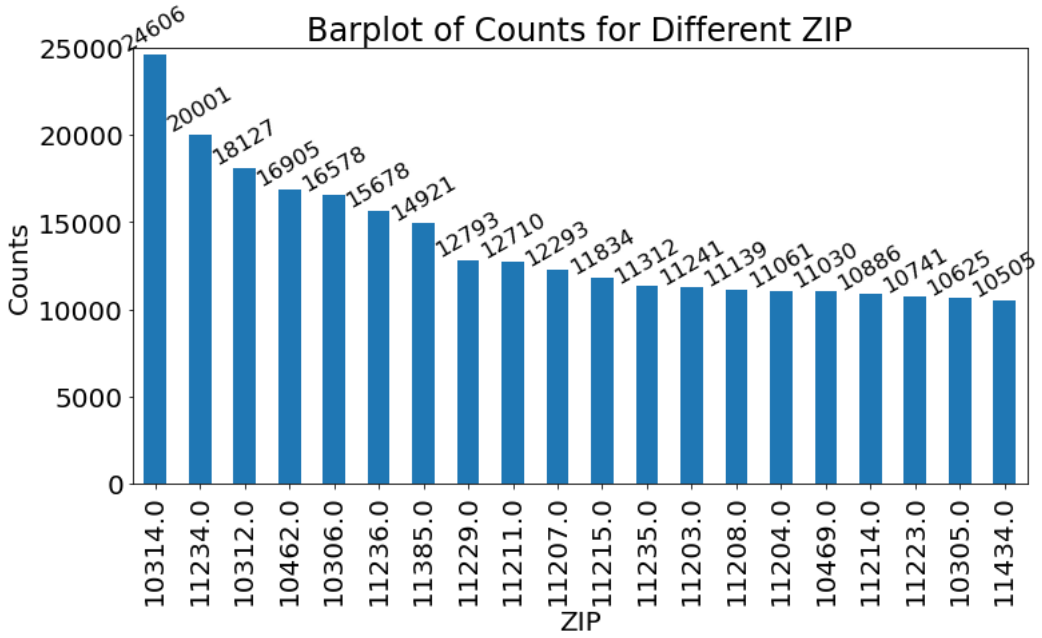
20. *STADDR*

The field *STADDR* stores the street address of each datapoint. It is a categorical variable with 1,070,994 records and 676 missing values. The visualization for the counts for top 20 most seen values is attached below.



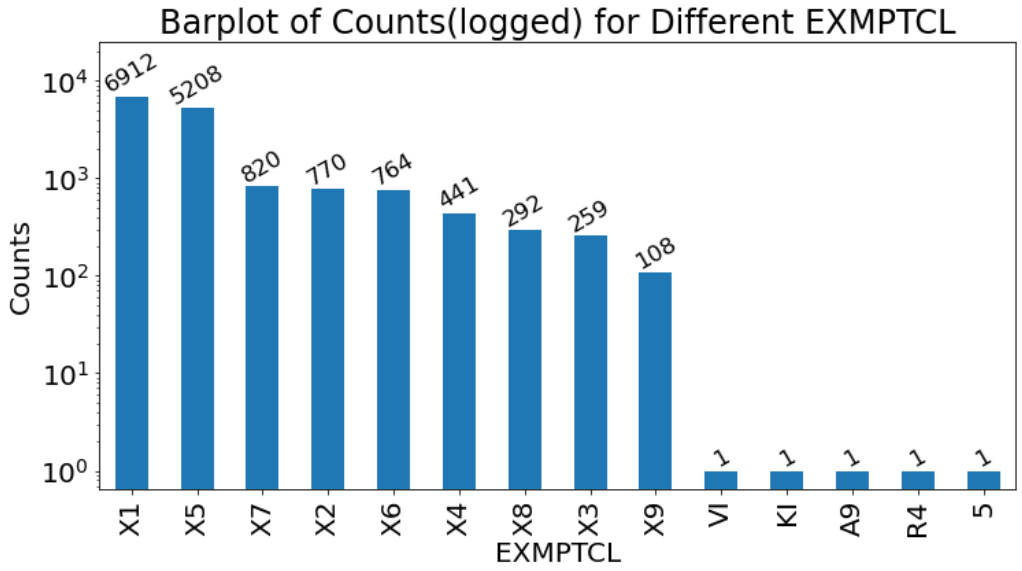
21. *ZIP*

The field *ZIP* stores the zip code of each datapoint. It is a categorical variable with 1,070,994 records and 29,890 missing values. There are 196 unique zips. The visualization for the counts for top 20 most seen values is attached below.



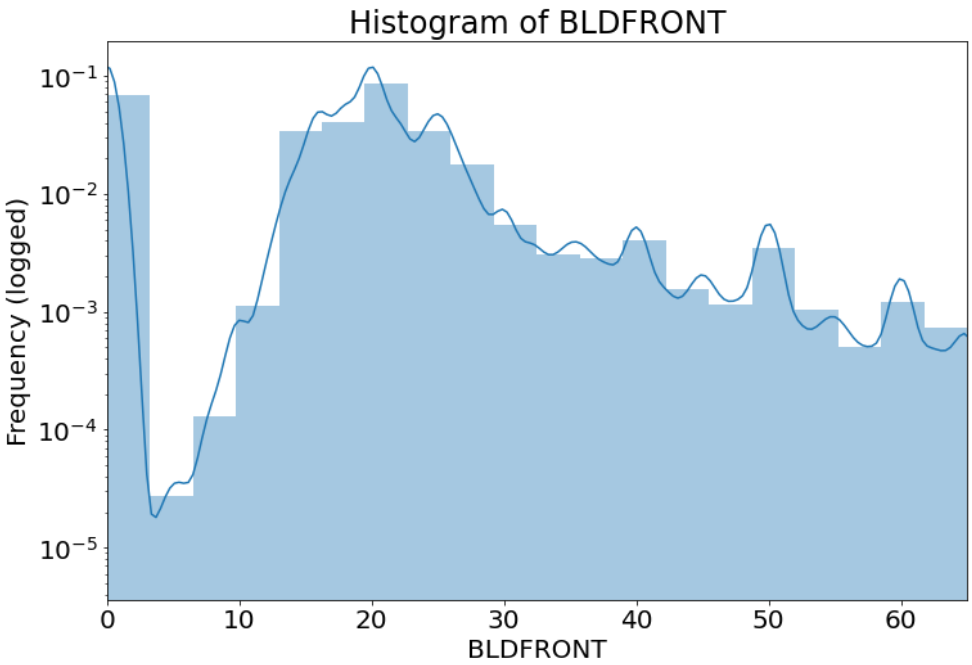
22. *EXMPTCL*

The field *EXMPTCL* stores the exemption class of each datapoint. It is a categorical variable with 1,070,994 records and 1,055,415 missing values. The visualization for the counts for different EXMPTCLs is attached below.



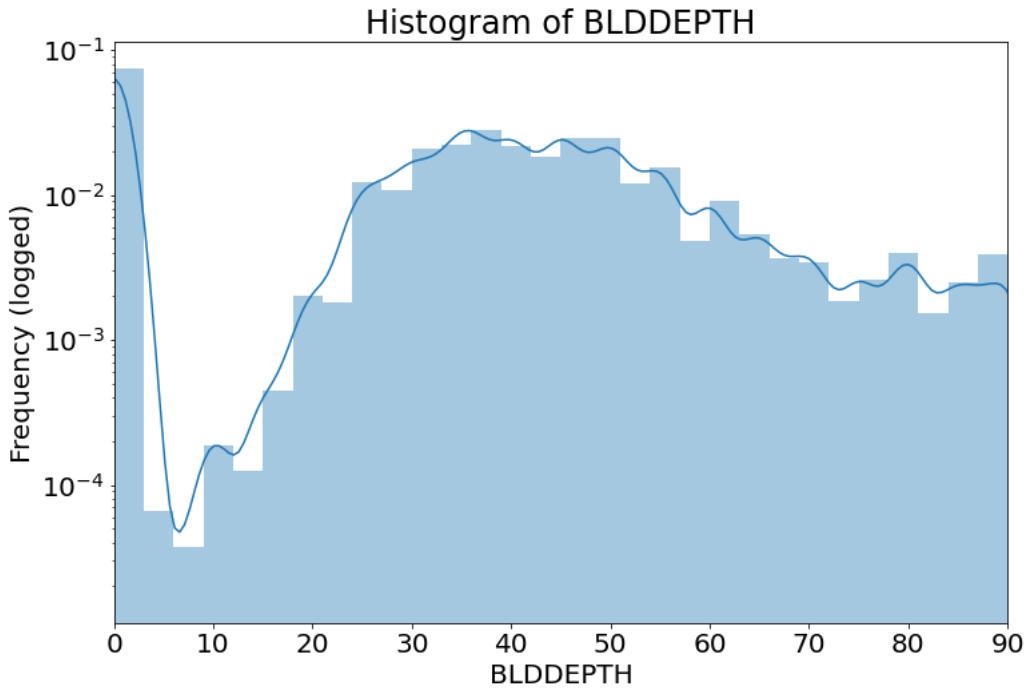
23. *BLDFRONT*

The field *BLDFRONT* stores the building width of each datapoint. It is a numerical variable with 1,070,994 records and no missing values. It ranges from 0 to 7,575, with 95% of its values locate in the range 0 to 65. The histogram of BLDFRONT with a value less than or equal 65 is attached below.



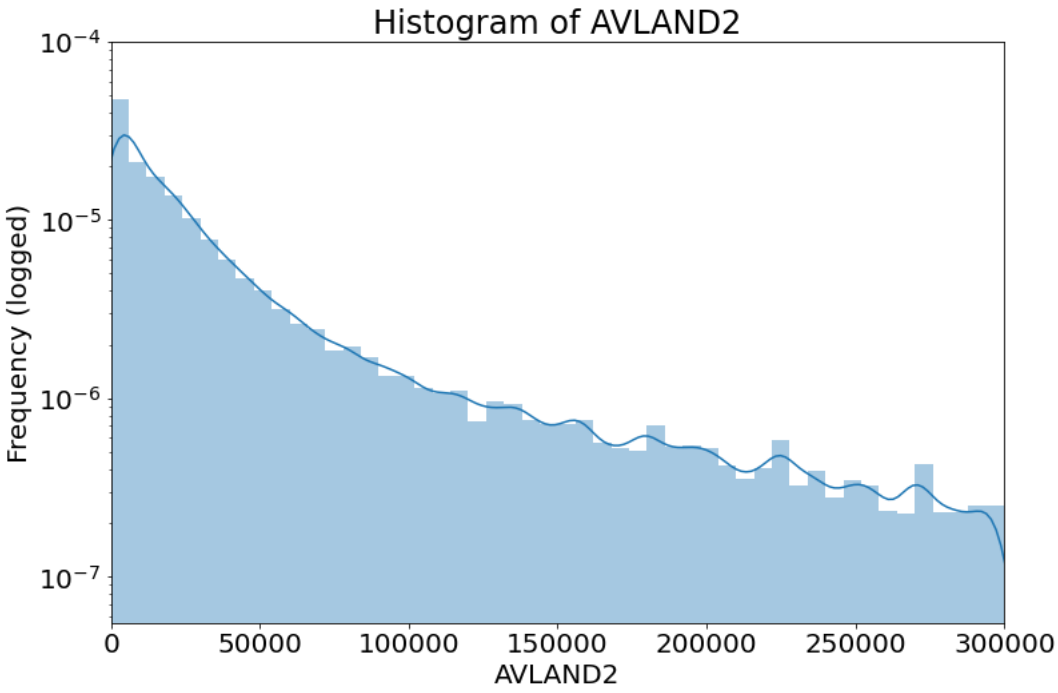
24. *BLDDEPTH*

The field *BLDFRONT* stores the building depth of each datapoint. It is a numerical variable with 1,070,994 records and no missing values. It ranges from 0 to 9,393, with 95% of its values locate in the range 0 to 90. The histogram of *BLDDEPTH* with a value less than or equal 90 is attached below.



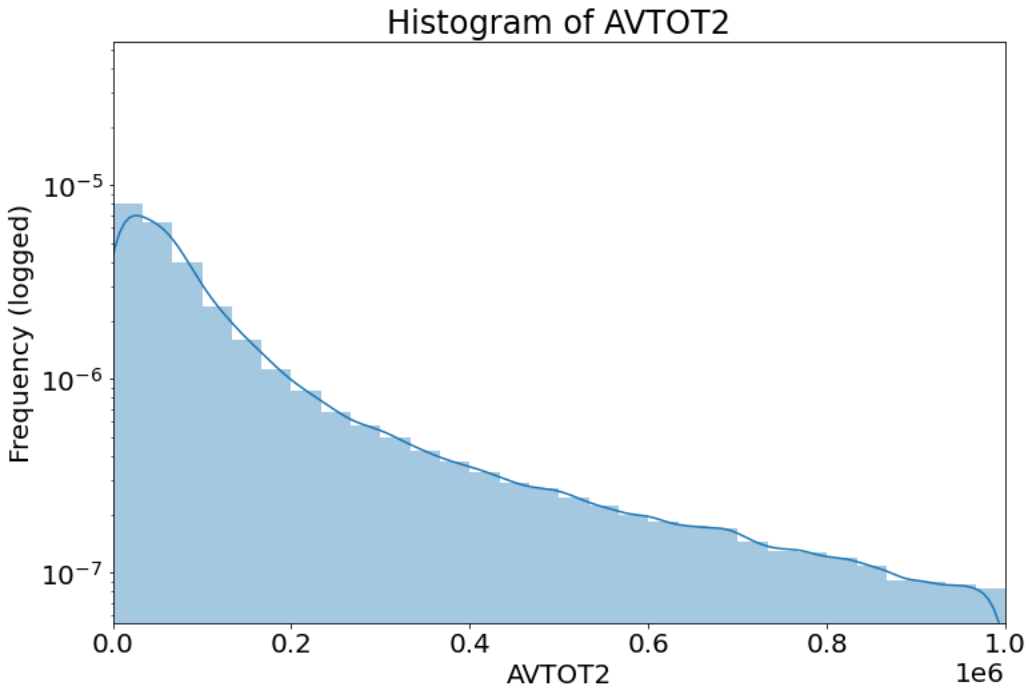
25. *AVLAND2*

The field *AVLAND2* stores the transitional land value of each datapoint. It is a numerical variable with 1,070,994 records and 288,278 missing values. It ranges from 0 to 2,371,005,000, with 91% is of its values locate in the range 0 to 300,000. The histogram of *AVLAND2* with a value less than or equal 300,000 is attached below.



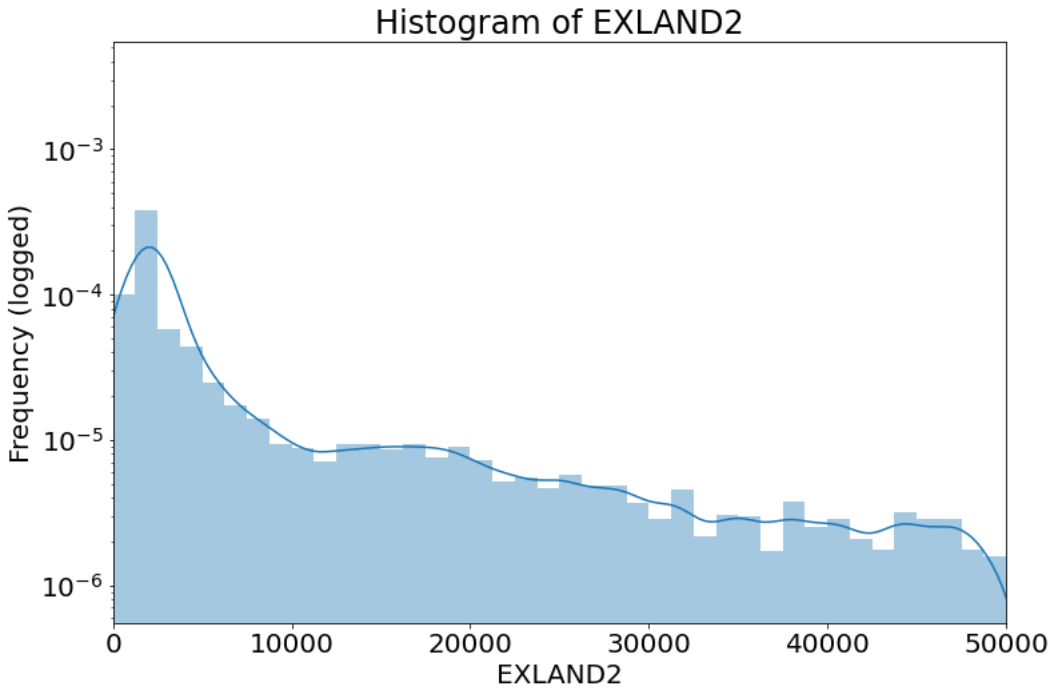
26. *AVTOT2*

The field AVTOT2 stores the transitional total value of each datapoint. It is a numerical variable with 1,070,994 records and 788,262 missing values. It ranges from 0 to 4,501,180,002, with 92% is of its values locate in the range 0 to 1,000,000. The histogram of *AVTOT2* with a val ue less than or equal 1,000,000 is attached below.



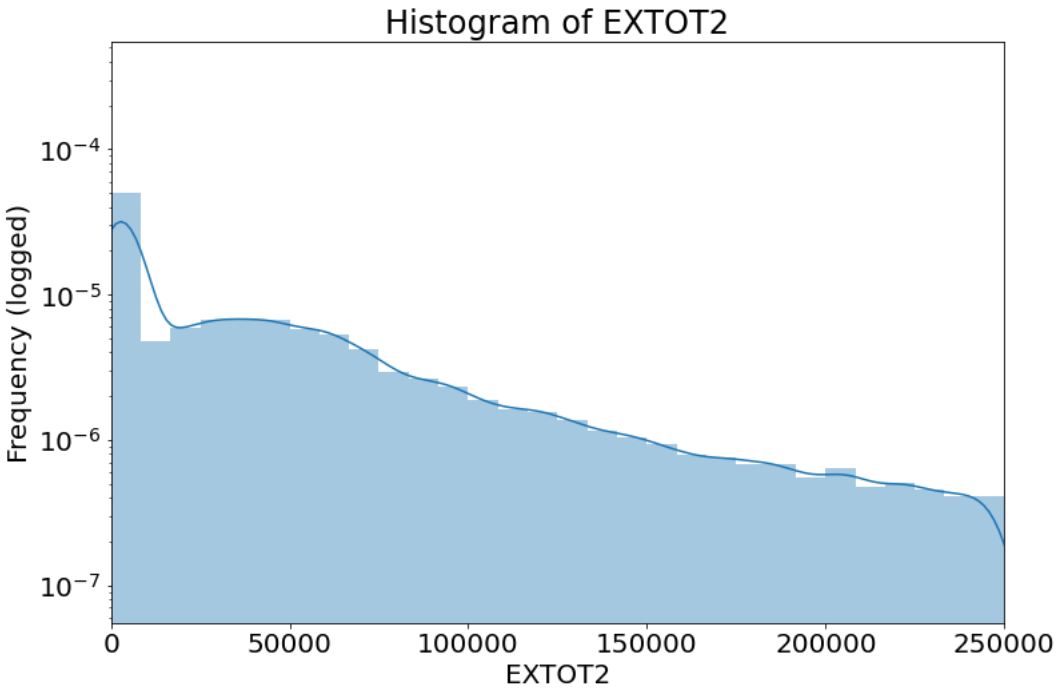
27. *EXLAND2*

The field *EXLAND2* stores the transitional exemption land value of each datapoint. It is a numerical variable with 1,070,994 records and 983,545 missing values. It ranges from 0 to 2,371,005,000, with 78% is of its values locate in the range 0 to 50,000. The histogram of *EXLAND2* with a value less than or equal 1,000,000 is attached below.



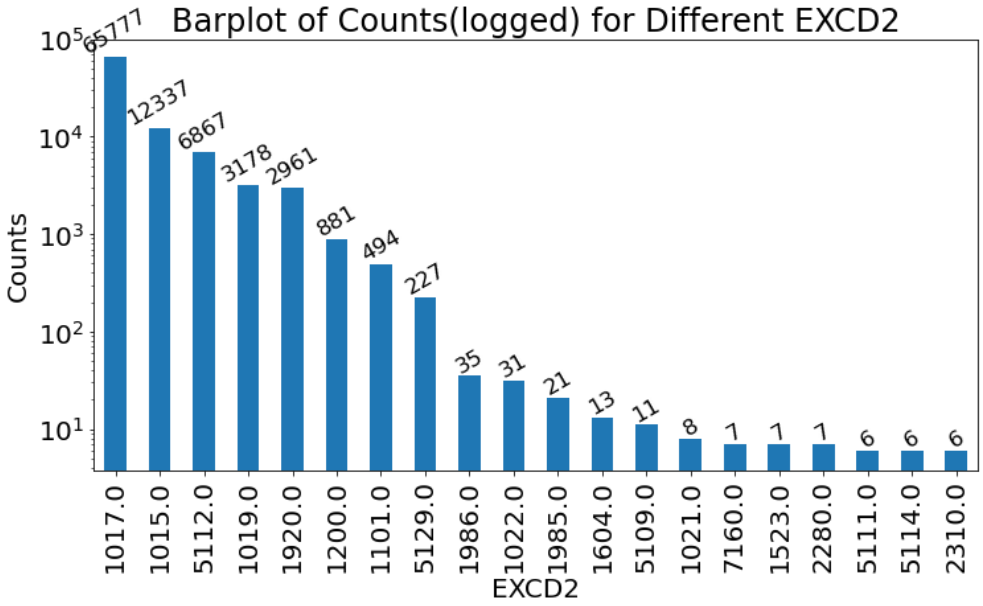
28. *EXTOT2*

The field *EXTOT2* stores the transitional total value of each datapoint. It is a numerical variable with 1,070,994 records and 940,166 missing values. It ranges from 0 to 4,501,180,002, with 85% is of its values locate in the range 0 to250,000. The histogram of *EXLAND2* with a value less than or equal 1,000,000 is attached below.



29. *EXCD2*

The field *EXCD2* stores the exemption code 2 of each datapoint. It is a categorical variable with 1,070,994 records and 978,046 missing values. The visualization for the counts for top 20 most seen values is attached below.



30. *PERIOD*

The field *PERIOD* stores the assessment period of each datapoint. It is a categorical variable with 1,070,994 records and no missing values. It has all identical values of ‘final’.

31. *YEAR*

The field *YEAR* stores the assessment year of each datapoint. It is a categorical variable with 1,070,994 records and no missing values. It has all identical values of ‘2010/11’.

32. *VALTYPE*

The field *VALTYPE* stores the filter field when the data was pulled. It is a categorical variable with 1,070,994 records and no missing values. It has all identical values of ‘AC-TR’.