1. Executive Summary:

The objective of this project is to construct an unsupervised model to detect tax fraud using the Property Valuation and Assessment Data dataset, which can identify individuals who may be misrepresenting their property characteristics and underpaying taxes. To accomplish this, data cleaning and imputation are performed. Subsequently, the team endeavors to develop variables that can provide better estimates of a property's value and detect abnormal properties. To achieve this, 91 novel variables are created by leveraging the original field, and less important features are discarded. Since the dataset contains many characteristics, PCA is employed to reduce dimensionality, and only the five most significant principal components are retained. Two scoring methods are employed to identify anomalies in the dataset after dimensionality reduction. Two additional columns are created to store the ranking of each record in the dataset, based on its respective score, when sorted in descending order. Then, the final fraud score is computed based on the equal weight of the two scores, and the properties are sorted by their fraud score from the highest to the lowest. 1,000 Properties with a high probability of fraud are identified based on their high fraud scores. These properties are passed to experts for analysis and feedback to enhance the model's effectiveness.

2. Description of Data

The dataset, Property Valuation and Assessment Data, 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 Name	Field Type	# Records Have Va	% Populated	# Zeros	Min	Max	Mean	Stv	Most Common
0	LTFRONT	numeric	1,070,994.00	100.00%	169,108.00	0	9,999.00	36.64	74.03	0
1	LTDEPTH	numeric	1,070,994.00	100.00%	170,128.00	0	9,999.00	88.86	76.40	100
2	STORIES	numeric	1,014,730.00	94.75%	-	1	119.00	5.01	8.37	2
3	FULLVAL	numeric	1,070,994.00	100.00%	13,007.00	0	6,150,000,000.00	874,264.51	11,582,430.99	0
4	AVLAND	numeric	1,070,994.00	100.00%	13,009.00	0	2,668,500,000.00	85,067.92	4,057,260.06	0
5	AVTOT	numeric	1,070,994.00	100.00%	13,007.00	0	4,668,308,947.00	227,238.17	6,877,529.31	0
6	EXLAND	numeric	1,070,994.00	100.00%	491,699.00	0	2,668,500,000.00	36,423.89	3,981,575.79	0
7	EXTOT	numeric	1,070,994.00	100.00%	432,572.00	0	4,668,308,947.00	91,186.98	6,508,402.82	0
8	BLDFRONT	numeric	1,070,994.00	100.00%	228,815.00	0	7,575.00	23.04	35.58	0
9	BLDDEPTH	numeric	1,070,994.00	100.00%	228,853.00	0	9,393.00	39.92	42.71	0
10	AVLAND2	numeric	282,726.00	26.40%	-	3	2,371,005,000.00	246,235.72	6,178,962.56	2408
11	AVTOT2	numeric	282,732.00	26.40%	-	3	4,501,180,002.00	713,911.44	11,652,528.95	750
12	EXLAND2	numeric	87,449.00	8.17%	-	1	2,371,005,000.00	351,235.68	10,802,212.67	2090
13	EXTOT2	numeric	130,828.00	12.22%	-	7	4,501,180,002.00	656,768.28	16,072,510.17	2090

Table 1. Summary Statistics for Numeric Variables

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

Table 2. Summary Statistics for Categorical Variables

3. Data Cleaning

The Imputation Logistic process is used to ensure that the data is complete and ready for analysis. Filling in missing values can improve the accuracy and quality of the analysis, as missing data can lead to biased results. In addition, because we only focused on fraud properties, by eliminating the benign properties, we are more able to detect the fraud properties. The imputation and exclusion logic is as follows

1. Remove benign properties that we aren't interested in, 14478 rows of data are being eliminates
2. Fill in missing *Zip*:
   a) Assume the data is already sorted by zip. If a zip is missing, and the before and after zips are the same, fill in the zip with that value. 11423 zips are being filled
   b) For the rest, find the zip to the stress address using Nominatim API (New Imputation logic). 7931 zips are being filled
   c) For the rest, fill in with the previous record's zip.
3. Fill in missing *AVTOT, AVLAND, FULLVAL:*
   a) Calculate means for *AVTOT, AVLAND, FULLVAL* by taxclass, avoiding the records with zeros
   b) Substitute decent values for *AVTOT, AVLAND*, *FULLVAL* from averages by taxclass
4. Fill in missing *STORIES:*
   a) Calculate means for *STORIES* by taxclass
   b) Substitute decent values for *STORIES* from averages by taxclass
5. Fill in *LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT :*

a) Calculate means for *LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT* by taxclass
b) Substitute decent values for *LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT* from averages by taxclass

6. Build variable zip3, which contains the first three digits of zip code

4. Variables Description

91 variables are built from original fields. Among original 32 variables, FULLVAL, AVLAND, AVTOT, LTFRONT, LTDEPTH, BLDFONT, BLDDEPTH are used to build ratio variables to measure the abnormal properties. By normalizing the original value fields and scaling them based on different factors such as Tax Class, location, and property characteristics, the distribution of the values could be tightened, making it easier to identify any unusual or outlying values that may be present in the data. This can be useful for identifying potential errors or anomalies in the data and for gaining a better understanding of the overall patterns and trends in the data. Additionally, these variables can be used in further analysis or modeling to improve accuracy and reduce the impact of outliers on the results.

Description of variables	# Variables created
r1-r9 variables: each of the 3 $(*FULLVAL, AVLAND, AVTOT*) value fields normalized by each of these 3 sizes (Lot Size, Building size, and Building volume); then Take the inverses of all these 9 variables as new variables. By tightening the distribution, the unusual values are more likely to visualized.	18
Grouped average Variables: averaged 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 tightening the distribution, the unusual values are more likely to visualized.	72
Value Measure Ratio Variables: comparing the 3 $ value measures: FULLVAL/(AVLAND+AVTOT), normalize to the mean, and get the max value of (VRnorm, 1/VRnorm). By tightening the distribution, the unusual values are more likely to visualized.	1
Total Variable After Dropping	91

Table 3. Variable Table

5. Dimensionality Reduction

To deal with the high dimensionality of the dataset after creating 93 new variables, dimensionality reduction techniques were employed. Since there was no dependent variable, all features were z-scaled by subtracting the mean value of each feature for each record and dividing the result by the standard deviation of that feature, which ensures that all dimensions are scaled similarly. Principal Component Analysis (PCA) was then implemented to reduce the dimensionality further. The cumulative variance plot was used to determine the optimal number of components, and it was found that the optimal n_components for PCA should be 5 based on

the change of gradient of the graph. The PCA was then redone with n_components set to 5, and the top PCs were retained. The retained PCs were z-scaled again to ensure that each PC was equally important. Two scores were used to calculate the final fraud score. The first score, Score 1, is any function of these z-scaled PCs that looks for extremes. The second score, Score 2, was calculated by training an autoencoder on all the data to reproduce the z-scaled PC records. The fraud score was calculated as any measure of difference between the original input record and the autoencoder output record. The final score was obtained by combining the two scores using a weighted average rank order, where each score was weighted equally at 0.5.

6. Anomaly Detection Algorithms

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.

6.1 Z-scores outlier

The first method is using Zscores to detect outliers. The Z scale formula is

$$z_i = \frac{x_i - \mu_i}{\sigma_i},$$

After proper scaling, finding outliers is Easy. Then we complete the PCA with n_components equals to 5. PCA removes the correlations, allows dimensionality reduction, leaving the main information. Only the first few PCs are kept. Then, Z scaling is conducted again to make all the retained PCs equally important. Now, on each record the value of each variable explicitly shows how unusual that record is in that dimension. 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 power p can be anything, but we usually use p = 2, the Euclidean distance. This method is most likely to be linear.

6.2 Autoencoder

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:

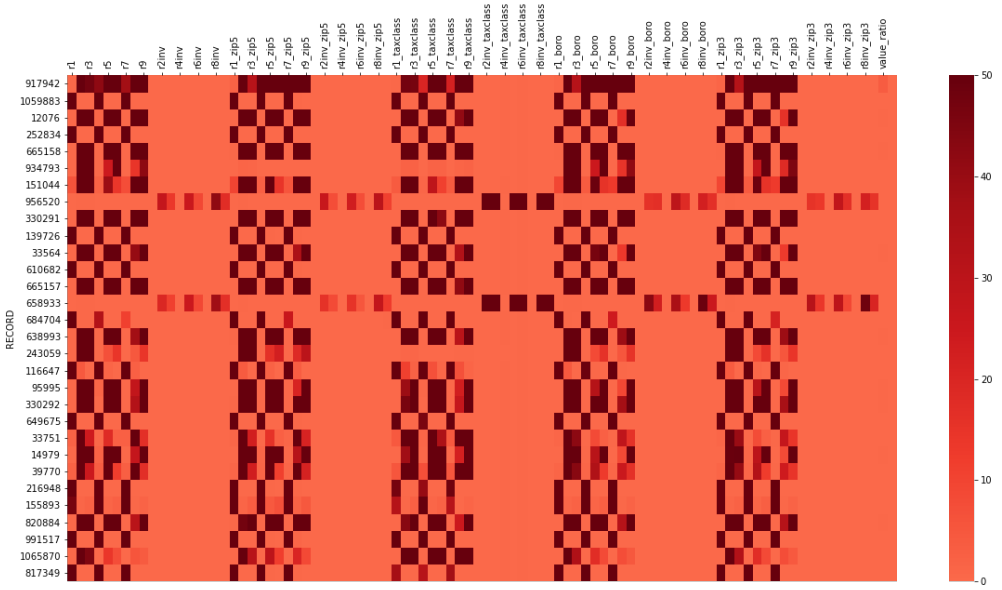
$$s_i = \left(\sum_n |z_n'^i - z_n^i|^p \right)^{1/p},$$

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. This method is generally nonlinear.

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. Results

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 6 properties on the list we are interested in, which are listed below.



Graph1. heatmaps of z-scaled variables

7.1 Unusual Property1-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

Table 4. Statistics of Property Record 658933

Graph 2. Graph of Property Record 658933

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

7.2 Unusual Property2- 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

Table 5. Statistics of Property Record 106681

Graph 3. Graph of Property Record 106681

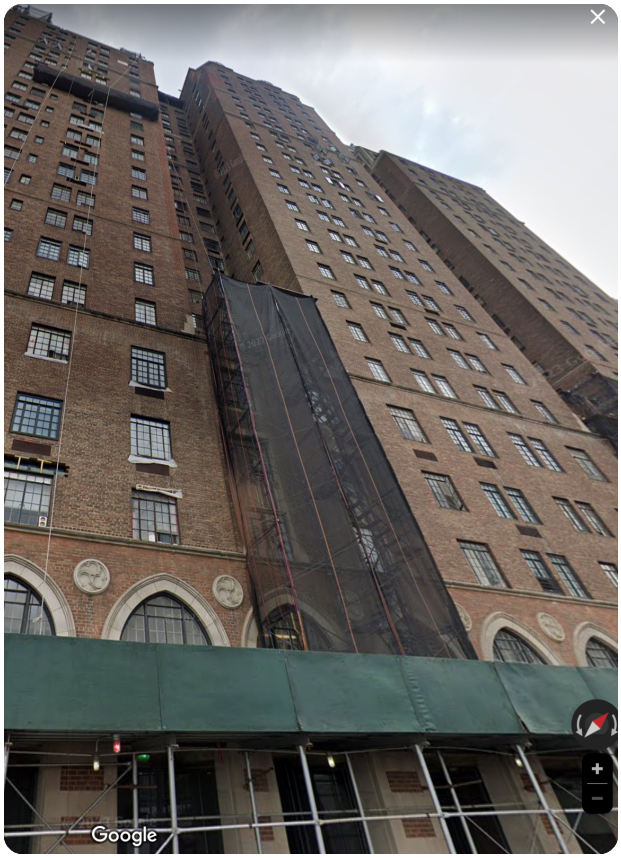
Sophisticated and elegant, The Lucerne has a large selection of family-sized homes of up to 3- and 4-bedrooms and duplexes — a rarity in Manhattan. The layouts flow beautifully from room

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

7.3 Unusual Property3- Record 95995

OWNER	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

Table 6. Statistics of Property Record 95995



Graph 4. Graph of Property Record 95995

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.

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

7.4 Unusual Property4-934793

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

Table 7. Statistics of Property Record 934793

Graph 5. Graph of Property Record 934793

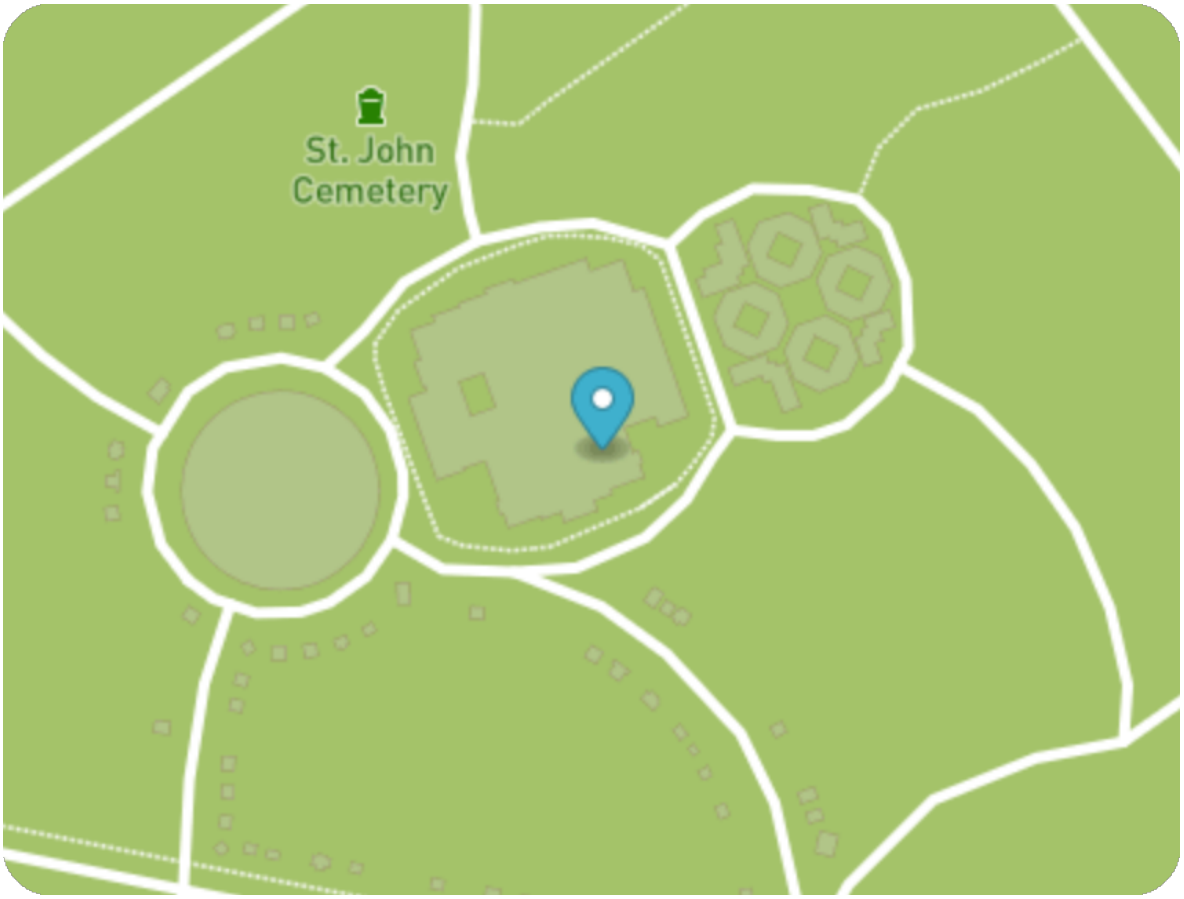
Located in the Breezy Point neighborhood in Queens, 217-02 Breezy Point Boulevard is a house with high r2 and r3 ratings, but a relatively low r6 rating, which suggests a potential issue with FULLVAL. Propertyshark.com estimates the assessed value of the property to be around $18,566,231, much lower than $273,000,000.

Source: https://streeteasy.com/building/217_02-breezy-point-boulevard-breezy_point

6.5 Unusual Property5-665158

OWNER	ST JOHNS CEMETERY	LTFRONT	1,412
ADDRESS	80-01A METROPOLITAN AVENUE	LTDEPTH	2,543
FULLVAL	29,355,000	BLDFRONT	12
AVLAND	13,140,000	BLDDEPTH	18
AVTOT	13,209,750	STORIES	6

Table 8. Statistics of Property Record 66518



Graph 6. Graph of Property Record 934793

The property is near St.john Cemetery. The property has relatively high r2,r3,r5,r6, indicating the number of story is abnormal and the BLDFRONT*BLDDEPTH is too large for specific LTFRONT and LTDEPTH.

Unusual Property6-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

Table 9. Statistics of Property Record 12076

Graph 6. Graph of Property Record 12076

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.

In general, though the abnormal properties are found by the calculated fraud scores, they have some characteristics in common. The stories are abnormal, the BLDFRONT*BLDDEPTH are too large for specific LTFRONT and LTDEPTH, or LTFRONT and LTDEPTH is too large for specific BLDFRONT*BLDDEPTH, etc. Thus, the fraud detection algorism is reliable, as it could correctly detect the abnormal properties.

8. Summary

The objective of this project is to develop an unsupervised tax fraud detection model using the Property Valuation and Assessment Data set. The goal is to identify private property owners who are misrepresenting their property characteristics and underpaying taxes. To achieve this, we first create a removal list of the 20 most commonly seen property owners, whose properties are regarded as benign and therefore dropped. Additionally, missing values are imputed.
Next, we focus on creating variables that can better estimate a property's value and identify abnormal properties. We concentrate on metrics such as dollars per square foot for the land and dollars per building volume, which result in 35 new variables created based on the original fields. Less relevant fields are discarded.

As the dataset has a large number of features, we implement Principal Component Analysis (PCA) to reduce dimensionality. We standardize all the features by subtracting the mean and dividing by the standard deviation. We plot the cumulative variance and decide on the optimal number of components for PCA to be 5 based on the change in gradient of the graph. We then redo the PCA with the top 5 components and standardize them again so that each retained component is equally important.

After reducing dimensionality, we use two scoring methods, Z-scores and the distance between the output of an autoencoder with a neural net and the original value, to detect anomalies in the dataset. We create two additional columns to store the rank of each record in the dataset when it is sorted in descending order based on the two scores. We then compute the final fraud score and consider 1000 records with high fraud scores as properties with a high probability of fraud. We also generate heatmaps of the standardized variables to identify which variables drive high scores and investigate the properties further.

However, this is just the initial version of the unsupervised model, and it is important to consult experts for analysis and feedback. Based on their feedback, we can adjust the model by improving exclusions, creating better variables, and modifying parameters and hyperparameters, such as the number of components for PCA, the power used for the Minkowski distance, and the shape of the neural net. Some improvement could be made. For intense, we could make a new variable called LTBLD ratio, which is calculated by lot size divided building size, as we could found from property investigation stage that the ratio is a good indicator of fraud property. We can then send the revised results to the experts for further feedback. By iteratively improving the model, we can finalize it and deploy it in a real-life business setting.

Appendix: DQR

Nature of the Dataset – NY Property data.csv

Description of Fields and Visualization

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

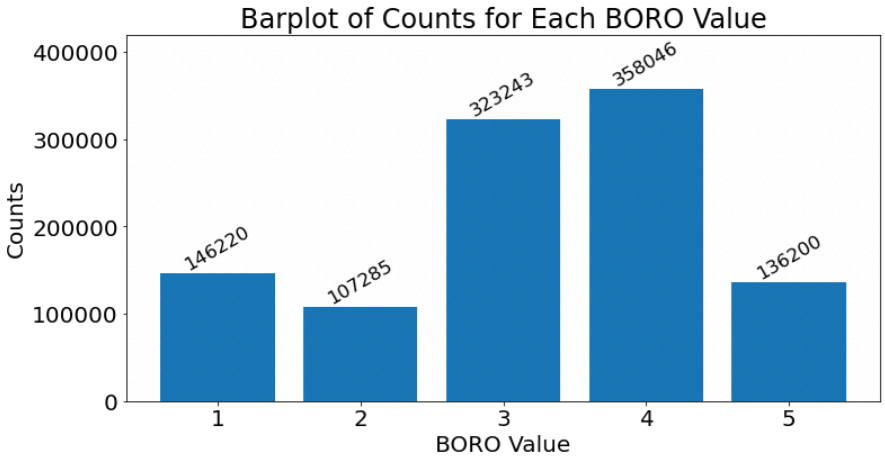
record, from 1 to 1,070,994.

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.

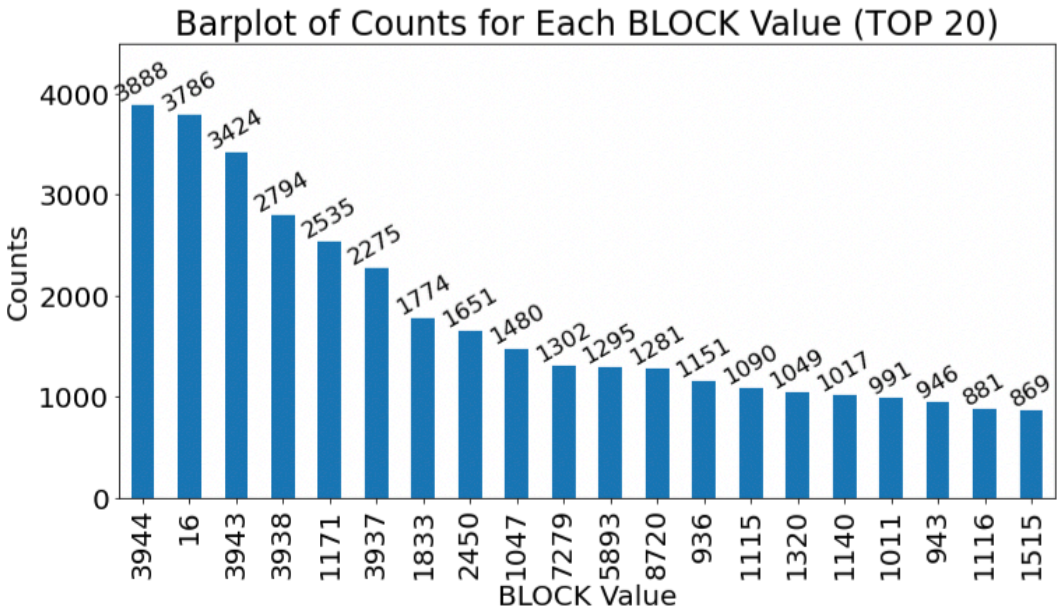
3. *BORO*

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.



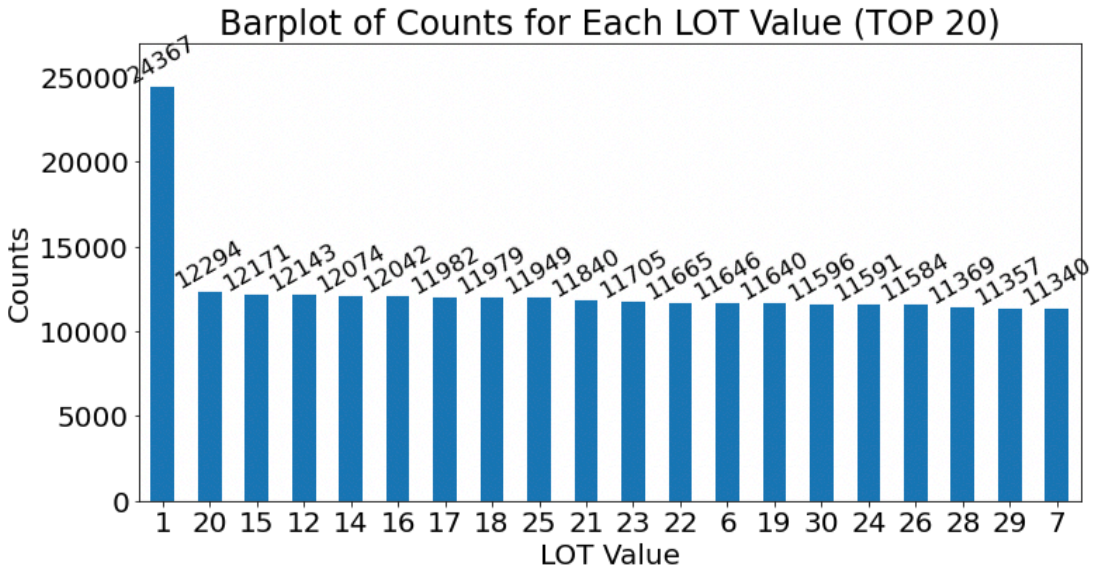
4. *BLOCK*

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.



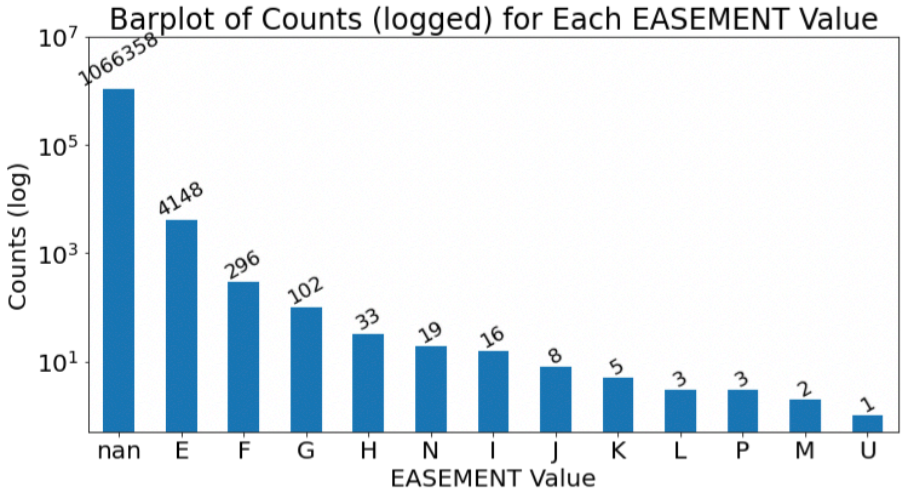
5. *LOT*

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.



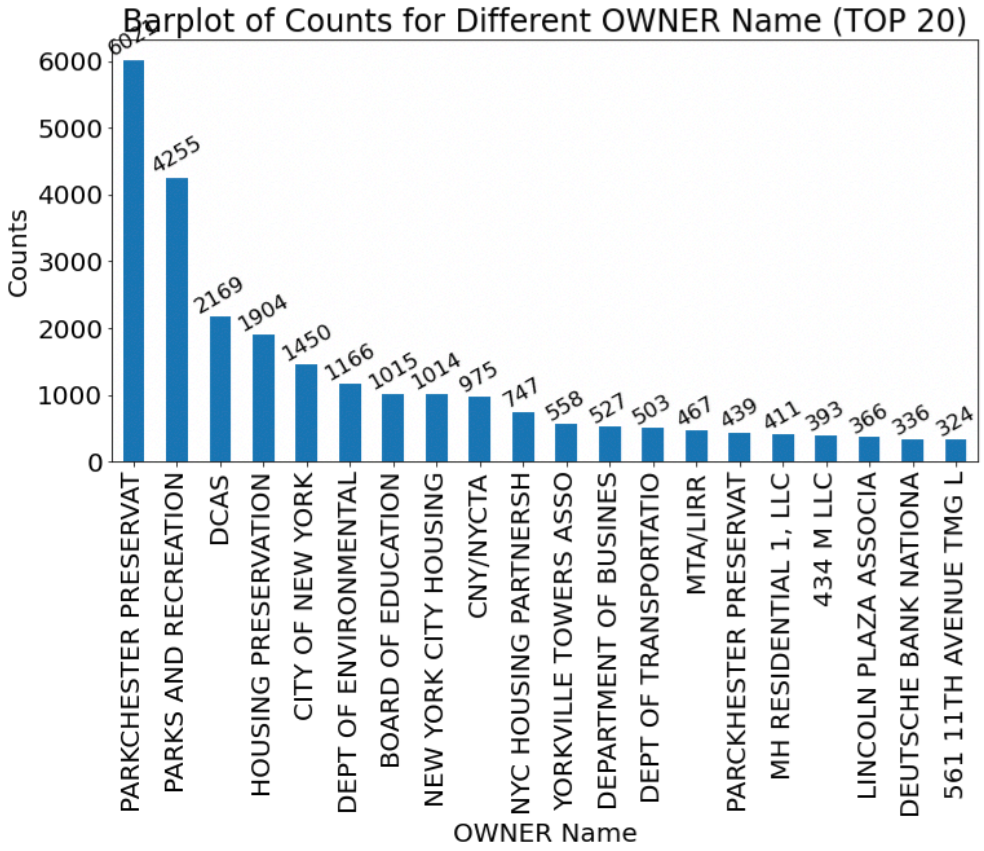
6. *EASEMENT*

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.



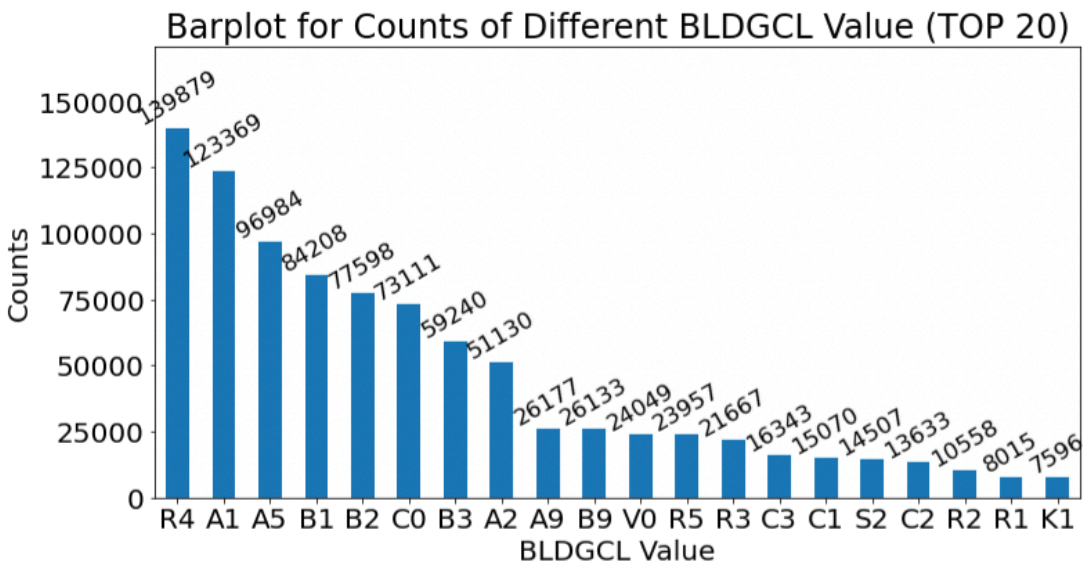
7. *OWNER*

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.



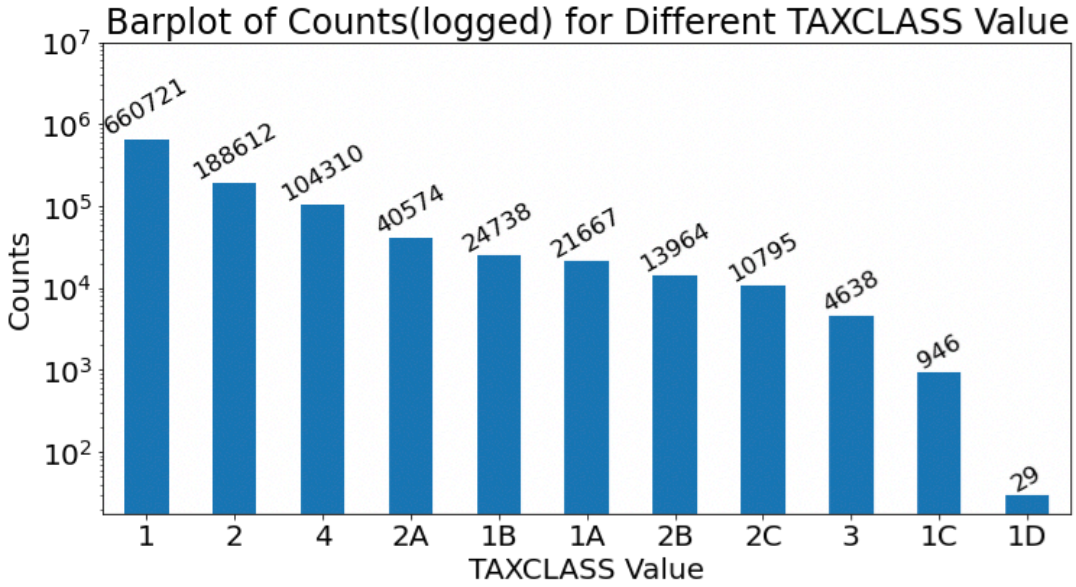
8. *BLDGCL*

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.



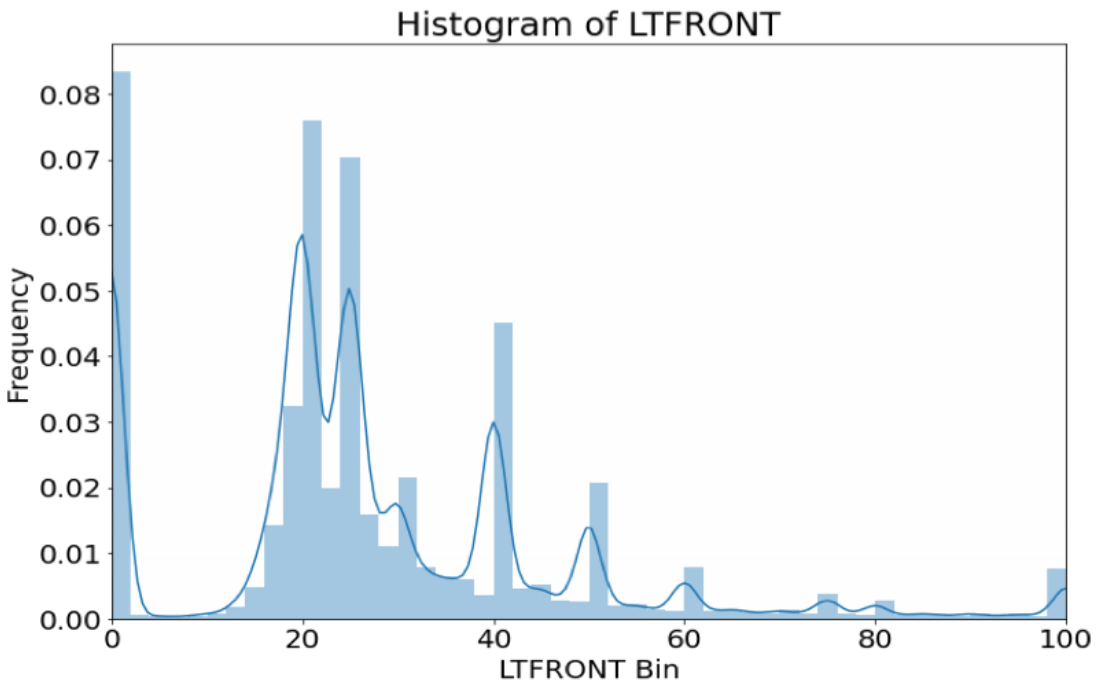
9. *TAXCLASS*

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.



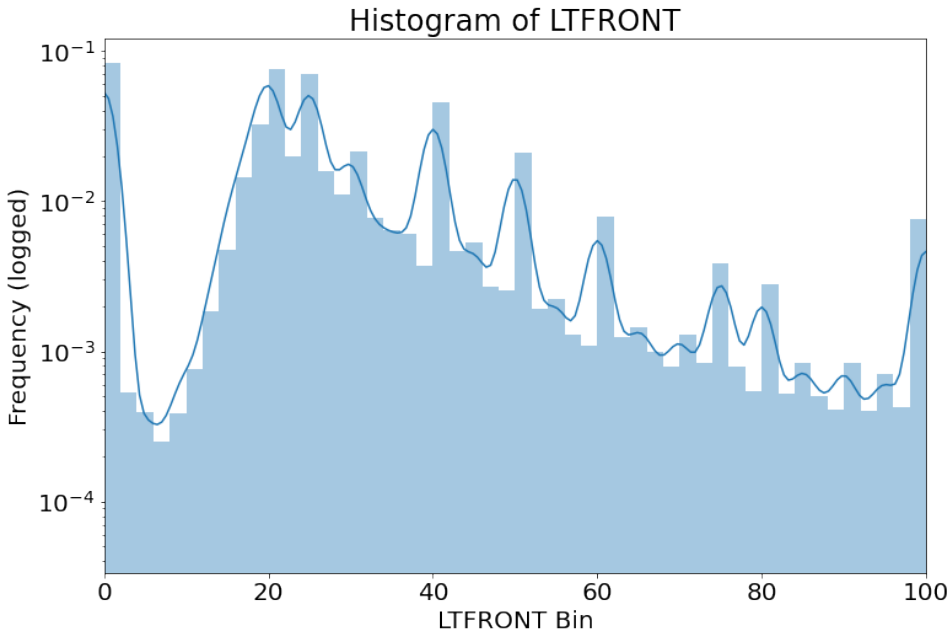
10. *LTFRONT*

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 in the range 0-100. The histogram of *LTFRONT* with a value less than or equal to 100 is attached below.



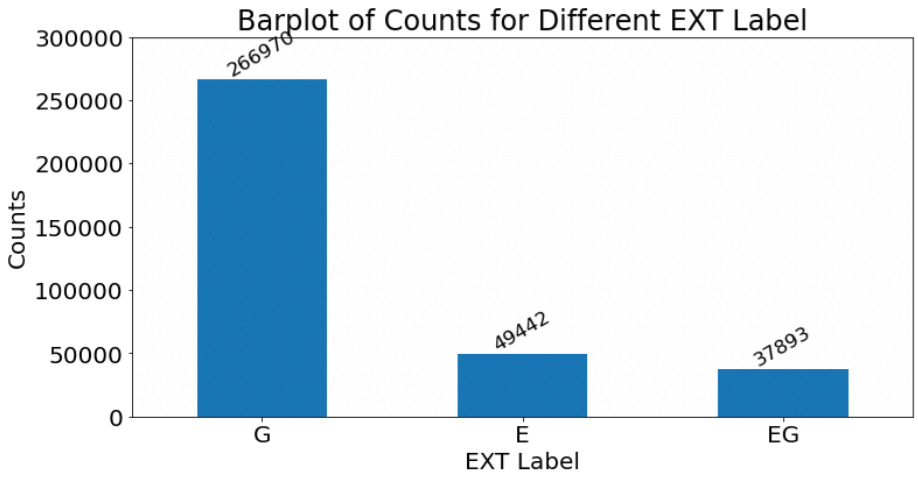
11. *LTDEP*

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.



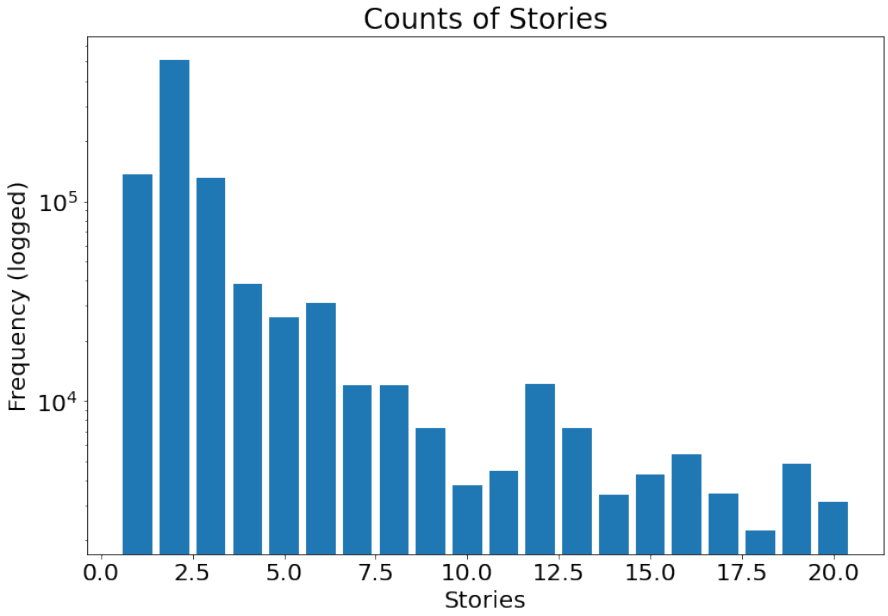
12. *EXT*

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.



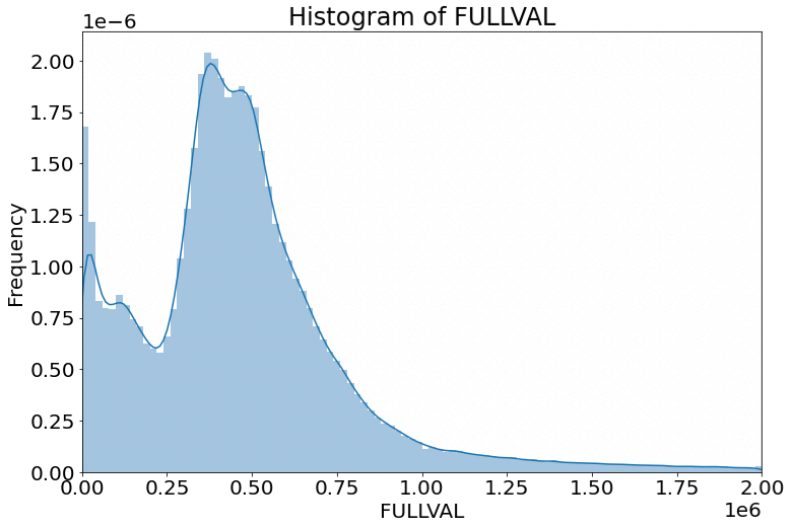
13. STORIES

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 histogram of *STORIE* with a value less than or equal to 20 is attached below.



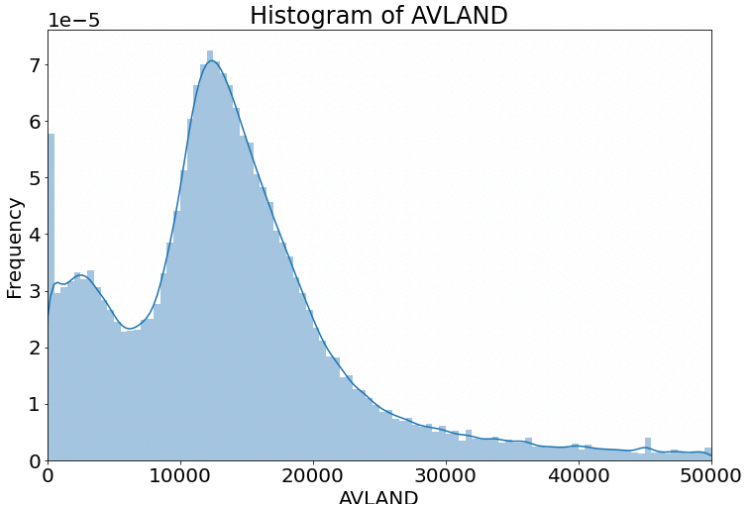
14. FULLVAL

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.



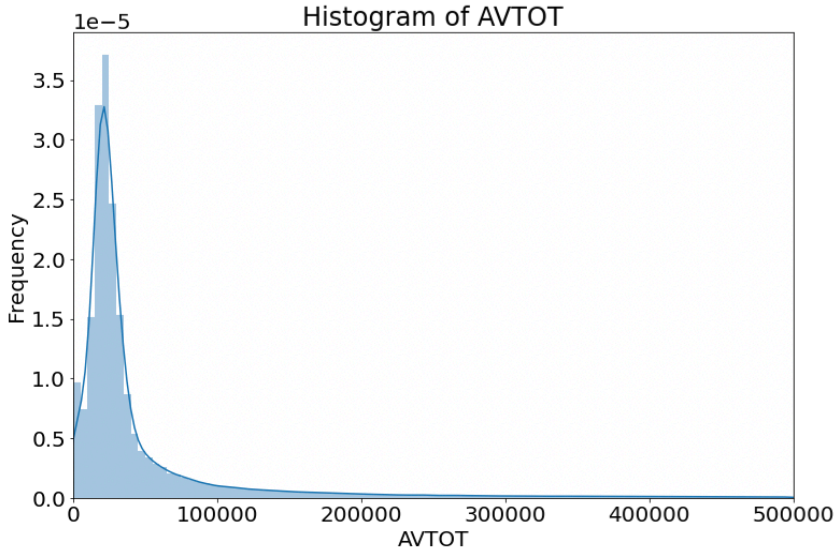
15. AVLAND

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.



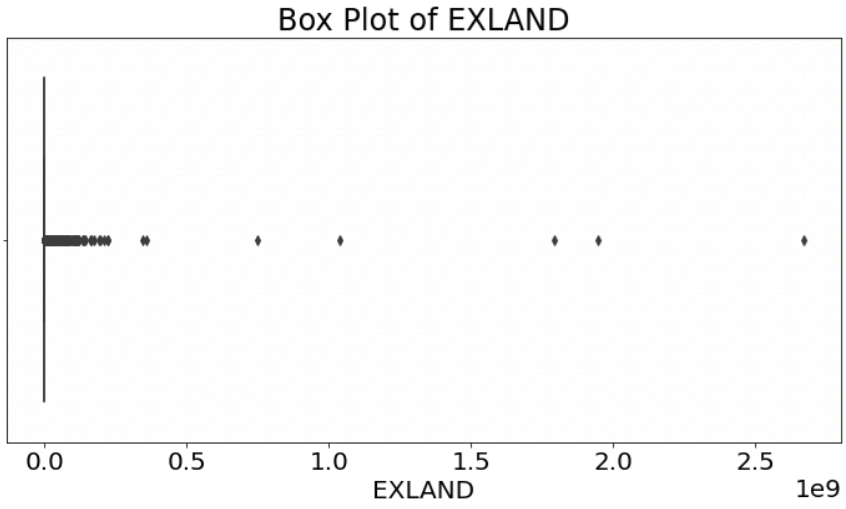
16. *AVTOT*

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.



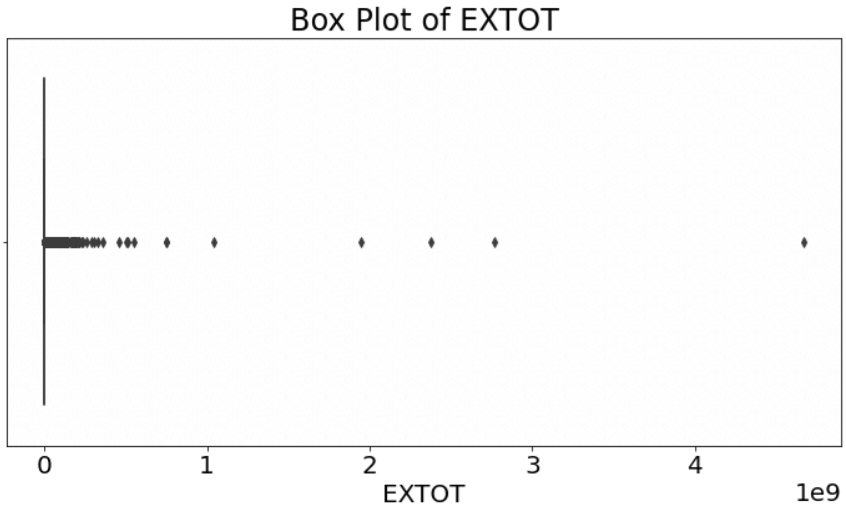
17. *EXLAND*

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.



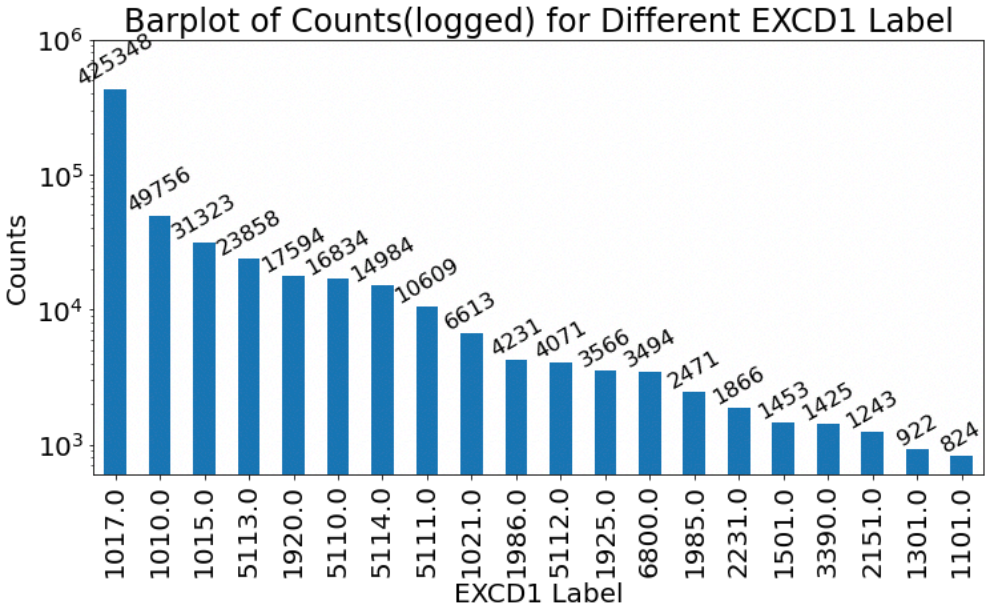
18. *EXTOT*

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.



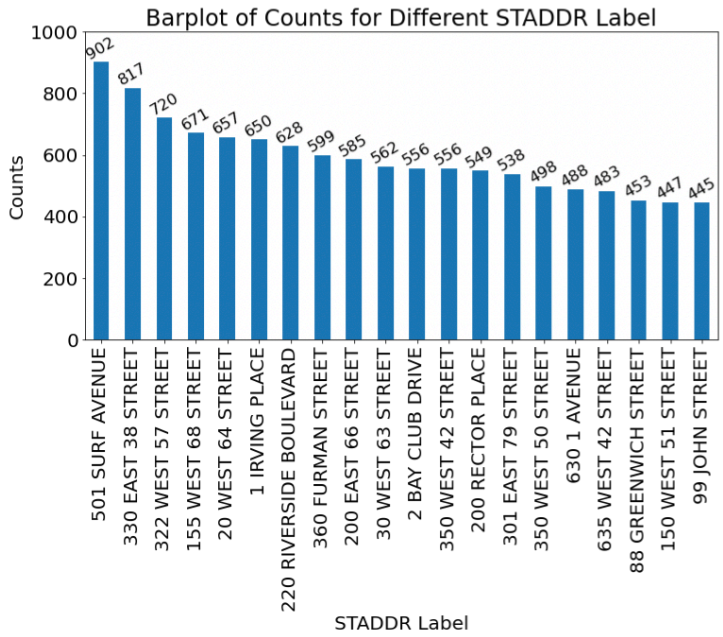
19. *EXCD1*

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.



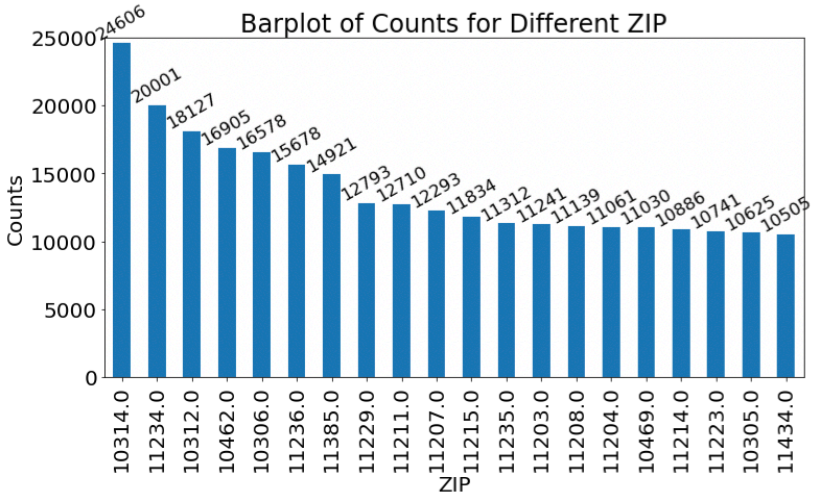
20. *STADDR*

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.



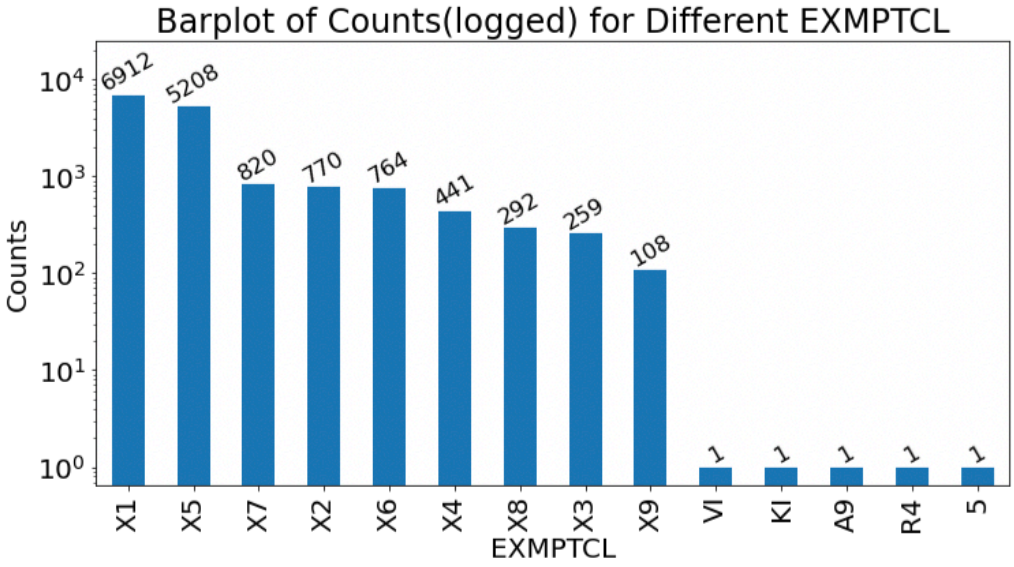
21. *ZIP*

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.



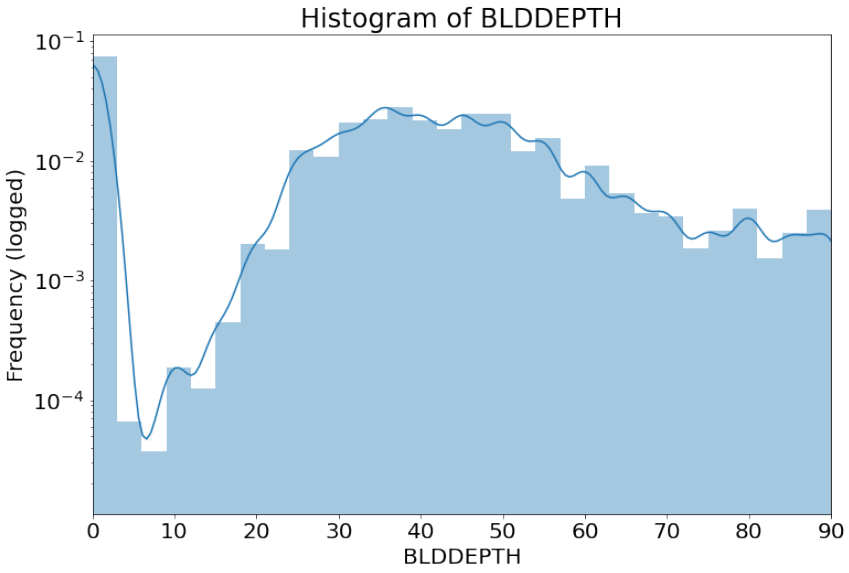
22. *EXMPTCL*

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.



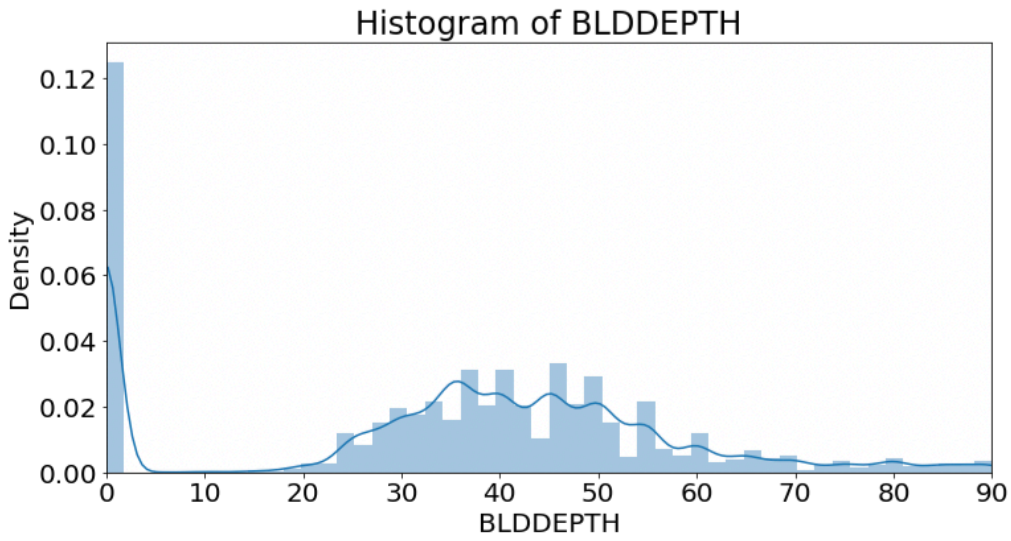
23. *BLDFRONT*

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.



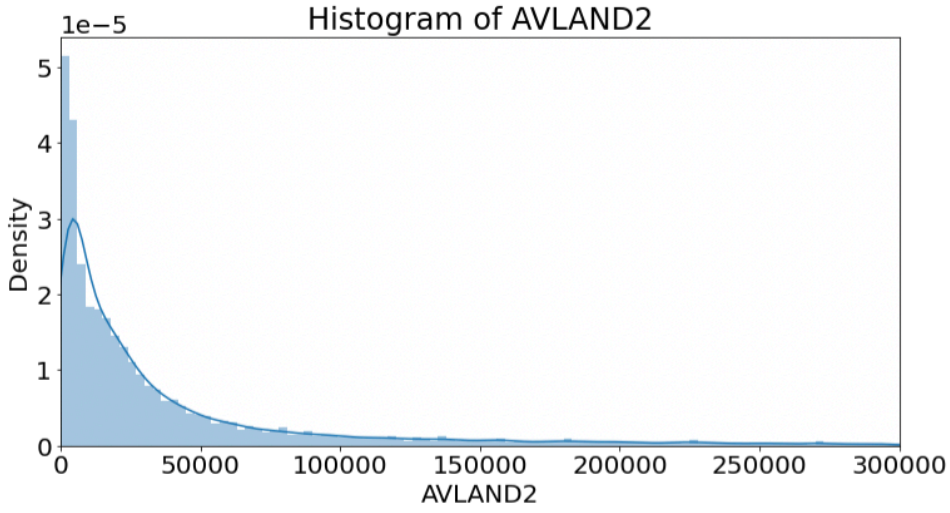
24. *BLDDEPTH*

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.



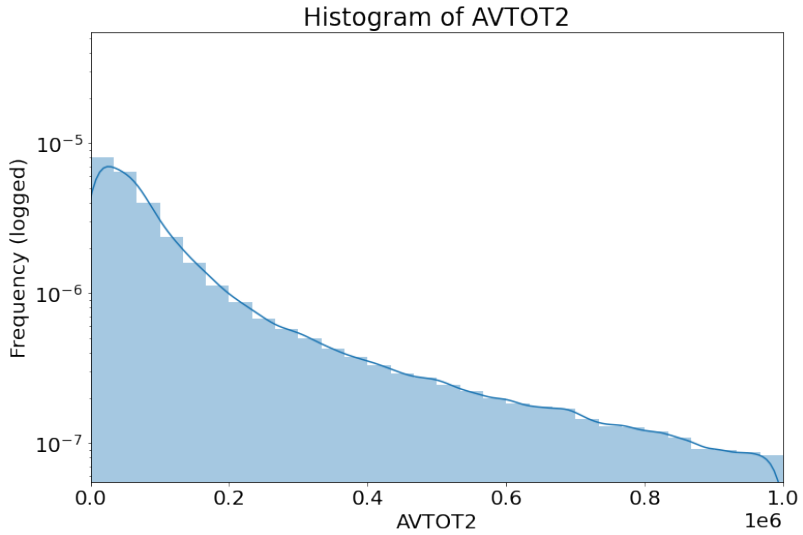
25. *AVLAND2*

The field *AVLAND2* stores the transitional land value of each datapoint. It is a numeric al variable with 1,070,994 records and 788,268 missing values. It ranges from 0 to 2,371,005,00 0, 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.



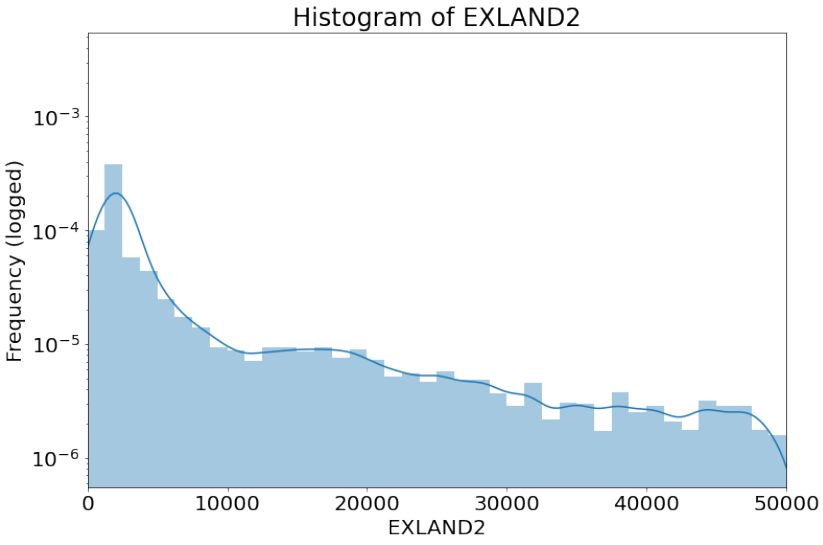
26. *AVTOT2*

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.



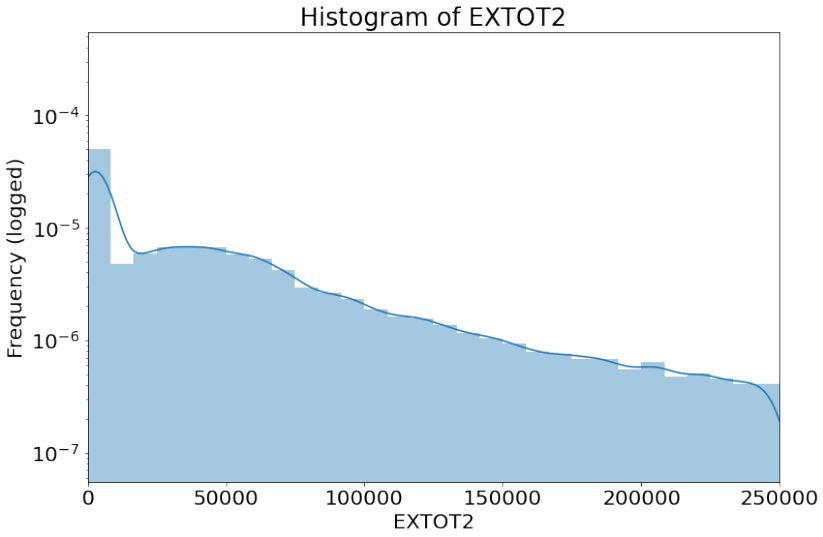
27. *EXLAND2*

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.



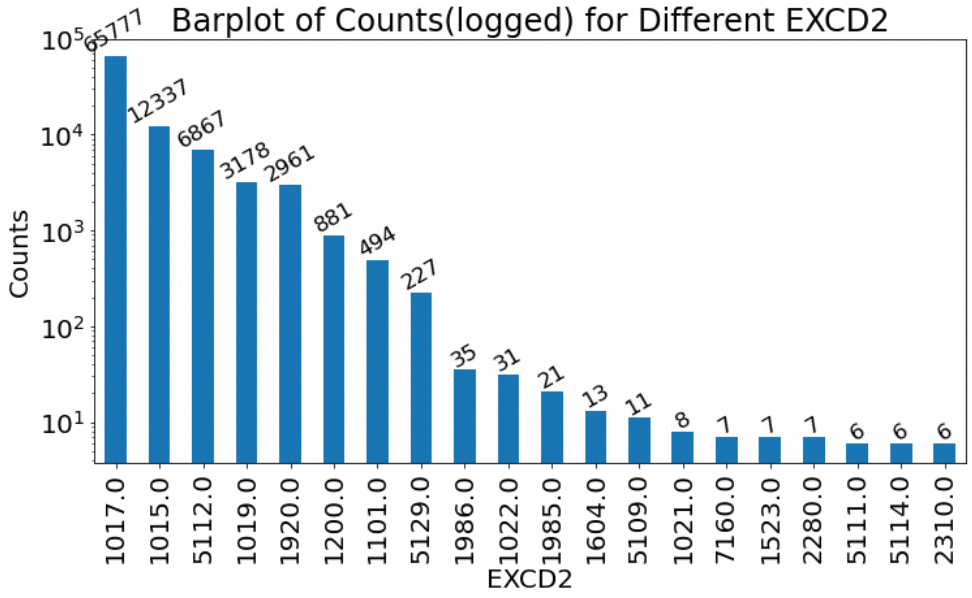
28. *EXTOT2*

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’.