NEW YORK
PROPERTY FRAUD IDENTIFICATION

Nate Errez

Min Kim

Sayaka Kuwayama

Gerardo Plascencia

Yao Wan

Date: April 27th, 2021

Professor Stephen Coggeshall

Project Advisor

Table of Contents

Rady | UC San Diego School of Management

New York Property Data Executive Summary

Nate Errez, Min Kim, Sayaka Kuwayama, Gerardo Plascencia, Yao Wan

---

Overview

We were hired by the City of New York to build a model for determining if any of the properties on their records were out of the ordinary and potential candidates for further fraud investigation. City officials had reason to believe that there might be property tax fraud occurring in their city but needed more than a hunch to go off of.

This problem was not one that could be solved by manually looking through the records. The New York property data has over one million records with 32 fields with many missing values. We built an algorithm which would determine which records were strange or unusual and then sort all the records by their unusualness rank or fraud score.

Expert Variables

The first and most important step of the project was building expert variables that would help identify potential fraudulent records. Using field calculations, we determined each property's lot size, building size and volume, and then created ratios of these with the full property value, assessed value of the land and assessed total value. These ratios were then compared to the mean values within each of the zip codes (3 and 5 digits), tax classes, and Boroughs.

Models

Two unsupervised models were used to determine which records were unusual, Z-Score outliers and Autoencoder error. Both models gave an indication of how far away from the others each record is. Each model assigned every record with a score and then a final score was determined by averaging each of the original scores. Both models provided very similar results.

Results

The final results of the modeling left us with a rank ordered list of records from most strange to least strange. Our preliminary investigations showed that many of these records were strange because of likely data entry errors, but further investigation by the City of New York is required.

1 Data Description

1.1 File Description

The “NY Property Data.csv”, is a property valuation and assessment dataset created around November in 2010, representing the New York City properties assessments for the purpose to calculate property tax, grant eligible property exemptions and/or abatements. Data is collected and entered into the system by various city employees - property assessors, property exemption specialists, ACRIS reporting, Department of Building reporting, etc. It is made up of 32 fields* and 1,070,994 records gathered from the NYC open source data.

*The details of each field can be found in the DQR located at the very end of this report

1.2 Summary Statistics

The summary statistics tables are developed according to the type of each field, whether it is numerical or categorical. Of the 32 fields, 14 fields were numerical and 18 fields categorical.

Table 1.2.1 Summary Statistics of Numerical Fields

Field Name	Field Type	# Records	% Populated	Mean	Standard Deviation	Min	Max	# Zeros
LTFRONT	Numeric	1070994	100.00%	36.64	74.03	0	9999	169108
LTDEPTH	Numeric	1070994	100.00%	88.86	76.40	0	9999	170128
STORIES	Numeric	1014730	94.75%	5.01	8.37	1	119	0
FULLVAL	Numeric	1070994	100.00%	874264.51	11582430.99	0	6150000000	13007
AVLAND	Numeric	1070994	100.00%	85067.92	4057260.06	0	2668500000	13009
AVTOT	Numeric	1070994	100.00%	227238.17	6877529.31	0	4668309000	13007
EXLAND	Numeric	1070994	100.00%	36423.89	3981575.79	0	2668500000	491699
EXTOT	Numeric	1070994	100.00%	91186.98	6508402.82	0	4668309000	432572
BLDFRONT	Numeric	1070994	100.00%	23.04	35.58	0	7575	228815
BLDDEPTH	Numeric	1070994	100.00%	39.92	42.71	0	9393	228853
AVLAND2	Numeric	282726	26.40%	246235.72	6178962.56	3	2371005000	0
AVTOT2	Numeric	282732	26.40%	713911.44	11652528.95	3	4501180000	0
EXLAND2	Numeric	87449	8.17%	351235.68	10802212.67	1	2371005000	0
EXTOT2	Numeric	130828	12.22%	656768.28	16072510.17	7	4501180000	0

Table 1.2.2 Summary Statistics of Categorical Fields

Field Name	Field Type	# Records	% Populated	# Unique	Most Common
RECORD	Categorical	1070994	100.00%	1070994	N/A
BBLE	Categorical	1070994	100.00%	1070994	N/A
B	Categorical	1070994	100.00%	5	4
BLOCK	Categorical	1070994	100.00%	13984	3944
LOT	Categorical	1070994	100.00%	6366	1
EASEMENT	Categorical	4636	0.43%	13	E
OWNER	Categorical	1039249	97.04%	863348	PARKCHESTER PRESERVAT
BLDGCL	Categorical	1070994	100.00%	200	R4
TAXCLASS	Categorical	1070994	100.00%	11	1
EXT	Categorical	354305	33.08%	4	G
EXCD1	Categorical	638488	59.62%	130	1017
STADDR	Categorical	1070318	99.94%	839281	501 SURF AVENUE
ZIP	Categorical	1041104	97.21%	197	10314
EXMPTCL	Categorical	15579	1.45%	15	X1
EXCD2	Categorical	92948	8.68%	61	1017
PERIOD	Categorical	1070994	100.00%	1	FINAL
YEAR	Datetime	1070994	100.00%	1	2010/11
VALTYPE	Categorical	1070994	100.00%	1	AC-TR

Overlooking the summary statistics table, we are able to see some distributions that will later play important roles in creating variables that will become valuable in running our algorithms and detecting property fraud. For the numerical table, the % of populated columns lets us visualize the amount of missing values that will need to be substituted in the process of producing variables. The mean and the sd can be used to compare the severity of the fraud committed once results are gained through running our algorithms.

2 Data Cleaning

Before the creation of variables that will be used in our algorithms to search for property fraud, we first cleaned our data by excluding some records that will not be necessary in producing variables. Then, we filled in missing values for chosen fields that will be used to create variables, a step known as imputation. Within this problem, because we are looking for outliers, we did not implement any outlier treatment.

2.1 Exclusions

To remove records that we are not interested in, we mainly looked at the field 'OWNER', which represents the people, group, company, etc., that possesses each property. Specifically, we eliminated property records that belong to the state, city or the federal government as we believed that these groups would not have the motivation to commit property fraud as they are under the nation. Also, removing these records would resolve any problem that may occur by the records skewing the statistics and values of variables.

Within the records that are owned by governmental groups, records that held typos or different versions of the same owner were also discarded, removing 24,168 records as a result.

Table 2.1 Excluded Organizations from the OWNERS field

```
['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',
 'DEPARTMENT OF BUSINES',
 'DEPT OF TRANSPORTATIO',
 'MTA/LIRR',
 'PARCKHESTER PRESERVAT',
 'MH RESIDENTIAL 1, LLC',
 'LINCOLN PLAZA ASSOCIA',
 'UNITED STATES OF AMER',
 'U S GOVERNMENT OWNRD',
 'THE CITY OF NEW YORK',
 'NYS URBAN DEVELOPMENT',
 'NYS DEPT OF ENVIRONME',
 'CULTURAL AFFAIRS',
 'DEPT OF GENERAL SERVI',
 'DEPT RE-CITY OF NY']
```

2.2 Imputation

In order to produce sufficient fraud scores for all records through our algorithms, we would have to fill in values for missing fields. The replacements for missing values would have to be developed by us using what is already given. So, it would be crucial that we create them to be as appropriate and consistent with the existing data.

For our data, rather than looking at all fields to fill in missing data, we only looked at FULLVAL (Total Market Value of the Land), AVLAND (Assessed Land Value), AVTOT (Assessed Total Value), ZIP (Postal Zip Code of the Property) , STORIES (Number of Stories for the Building/# of Floors), LTFRONT (Lot Frontage in Feet), LTDEPTH (Lot Depth in Feet), BLDFRONT (Building Frontage in Feet) and BLDDEPTH (Building Depth in Feet) as these fields were fundamental in creating our variables that will be used later for algorithms.

The values representing monetary and characteristics were considered differently in the process. Amongst FULLVAL, AVLAND and AVTOT, approximately 1% were missing values, shown with an NA meaning a blank value or with the value of 0. To fill up all of these possibilities, we filled in the data by taking an average of the missing value record's TAXCLASS.

In occupying the empty 21,772 values for the field 'ZIP', if the zip of the record that is before and after the missing record are equal, then we replaced it with that zip code. This process would narrow down the missing zips to 10,400. For the ones that are left, we replaced it with the zip that is equal to the record just above it.

For the field of STORIES, missing values equal to 43,968 records, which is about 5% of the field. None of the missing STORIES had the value zero. To keep the consistency amongst all fields, we also used TAXCLASS to group by. One understanding for this specific field was that even though STORIES represented the number of stories for a building, there are values that are not integers, implying that there may be parts to a building of different numbers of stories.

Lastly, about 200,000 records were missing amongst LOT (LTFRONT, LTDEPTH) and BUILDING (BLDFRONT, BLDDEPTH) fields. Unlike the STORIES field, these did not have blank, empty values, but all had 0s that we needed to replace. In addition to the 0s, we also took records with 1s into account as 1 could also be seen as an inappropriate value amongst the rest of the data. For all 4 of these fields, we first replaced the 0's and 1's with NAs so that they aren't looked at in calculating the mean. Then, using groupby on TAXCLASS, we got the groupwise average for each field. In accordance with the TAXCLASS, missing values for the fields were replaced with the averages.

3 Variable Creation

3.1 Motivation

While it may be easy to simply look for outliers in the data fields, we are only interested in finding "meaningful" outliers in detecting fraud. To achieve this goal, we need to build the right space to look for the outliers, and this involves two steps. The first is to create expert variables, and the second is to reduce dimensionality as much as possible. In this section, we will address how we went about building the expert variables.

3.2 Implementation

In order to create expert variables, we first considered general principles regarding properties' values. In particular, the principles we noted are the followings:

1. The bigger the property (i.e., land or/and building), the more expensive
2. Location has a large influence on a property's value
3. Property type can have a notable influence on a property's value

Based on the first principle, we created the expert variables of the followings:

- **Value** per ft2 for **lot area** - *Value/ (LTFRONT* LTDEPTH)*
- **Value** per ft2 for **building area** - *Value/ (BLDFRONT* BLDDEPTH)*
- **Value** per ft3 for **building volume** - *Value/(BLDFRONT* BLDDEPTH* STORIES)*

Since we are looking for unusual valuations in the fields of FULLVAL (total market value of the land), AVLAND (assessed land value), and AVTOT (assessed total value), we used these three values to substitute in the above equations. This process led to 9 new variables (i.e., 3 Value fields* 3 size metrics) to be created for each of the property records. Please see below for the visual representation of this process.

$$r_1 = \frac{V_1}{S_1} \qquad r_4 = \frac{V_2}{S_1} \qquad r_7 = \frac{V_3}{S_1}$$

$$r_2 = \frac{V_1}{S_2} \qquad r_5 = \frac{V_2}{S_2} \qquad r_8 = \frac{V_3}{S_2}$$

$$r_3 = \frac{V_1}{S_3} \qquad r_6 = \frac{V_2}{S_3} \qquad r_9 = \frac{V_3}{S_3}$$

V1: FULLVAL, V2: AVLAND, V3: AVTOT, S1: Lot area (LTFRONT LTDEPTH), S2: Building Area (BLDRONT* BLDEPTH), S3: Building volume (BLDRONT* BLDEPTH* STORIES)

*r1: FULLVAL per square foot for land area, r2: FULLVAL per square foot for building area, r3: FULLVAL per cubic foot for building volume, r4: AVLAND per square foot for lot area, r5: AVLAND per square foot for building area, r6: AVLAND per cubic foot for building volume, r7: AVTOT per square foot for lot area, r8: AVTOT per square foot for building area, r9: AVTOT per cubic foot for building volume

Furthermore, in order to account for the second and third principles of property valuation, namely the influence of the location and property type on the property value, we normalized the newly created expert variables by comparing to what is typical for the location and type of each property. When normalizing with respect to the location, we used the fields of the zip code with five digits, zip code with three digits, and Borough code independently. For normalizing with respect to the property type, we used the tax-class field. This process created 9 variables per normalization, and we ended with 45 variables (i.e., 9 variables from the original r1 to r9 and 36 normalized variables) to utilize in the fraud algorithms.

4 Dimensionality Reduction

4.1 Motivation

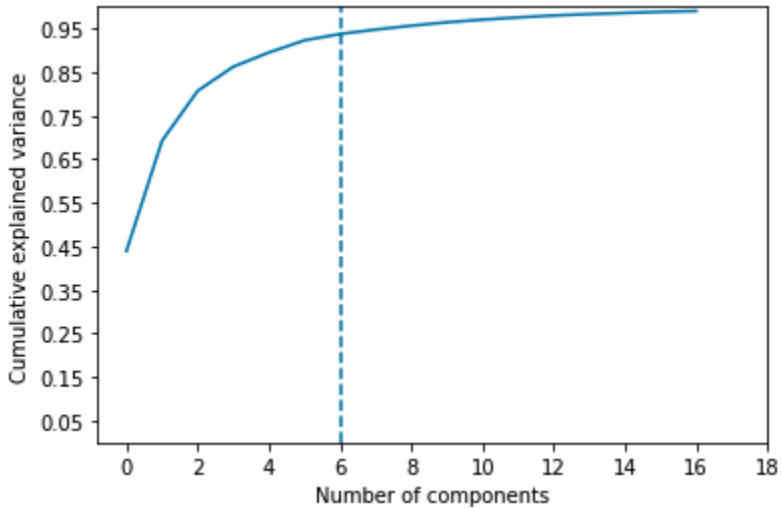
As briefly touched upon in the preceding section, dimensionality reduction is the second step necessary to build the right space to look for the outliers. This process is critical as unsupervised learning models suffer to produce good outcomes when dealing with high dimensional data. This section of the report addresses how we reduced the dimensionality of the data space by utilizing Principal Component Analysis (PCA).

PCA is a method that allows us to reduce dimensionality by transforming a large set of variables into a smaller set. Each PC is a linear combination of the original variables, and the PCs are found in order of the degree of variance spread in their direction. Many of the PCs will have a small magnitude which we can forego, and in this way, we can represent the data in a reduced dimension.

4.2 Implementation

Before we proceeded with PCA, we first z-scaled the 45 expert variables to prepare for dimensionality reduction. This process ensures that all the variables carry the same importance by normalizing the distances in the current space.

Then, we computed most of the PCs and produced the scree plot which showcases the cumulative explained variance as the number of PCs increases. Since the length of the PCs is proportional to the data variance in that PCs' directions, we can look at where the increase in the cumulative variance explained by the PCs stops growing to gauge the eigenvalue decay of the PCs.



The plot demonstrates that the increase in the cumulative variance explained by the PCs starts decreasing around 6. In other words, after the 6th PC, the magnitude of the PCs becomes fairly negligible. Based on this result, we decided to keep 6 PCs and forego all the others. As an outcome, we now represented the records in the data in a new reduced space, and as a byproduct, we also removed the multicollinearity problem of the highly correlated expert variables.

At the end, we again z-scaled these 6 PCs so that they would have the same importance when fed into the fraud algorithms.

5 Model Algorithm

Two models were built create the anomaly scores:

5.1 Model 1: Z-Score Outliers

In the first model, the fraud score is determined by the Euclidean distance between the six-dimensional record and the origin. By using the Z-scaled principal components, this method measures how far away one record is from the average. The squared Euclidean distance is calculated by taking each scaled principal component to the power of two, then sum the results up. Then we can get the Euclidean distance by taking the square root of the sum. The function is shown below where "PCz" denotes Z-scaled principal component, "i" denotes the record number, and "k" denotes the dimension.

$$s_{1i} = \left(\sum_{k}^{6} |PCz_{ik}|^{2} \right)^{1/2}$$

This method can better capture records with extreme values than simply summing the principal components. A record with a higher score may have more extreme values, and thus is more abnormal than a record with a lower score.

5.2 Model 2: Autoencoder Error

In the second model, an autoencoder was used to first encode the values to a lower dimensional representation then decode them back to try to reproduce the original values. The autoencoder is expected to minimize the reconstruction error for most records, and the fraud score is the difference between the original values and the autoencoder outputs.

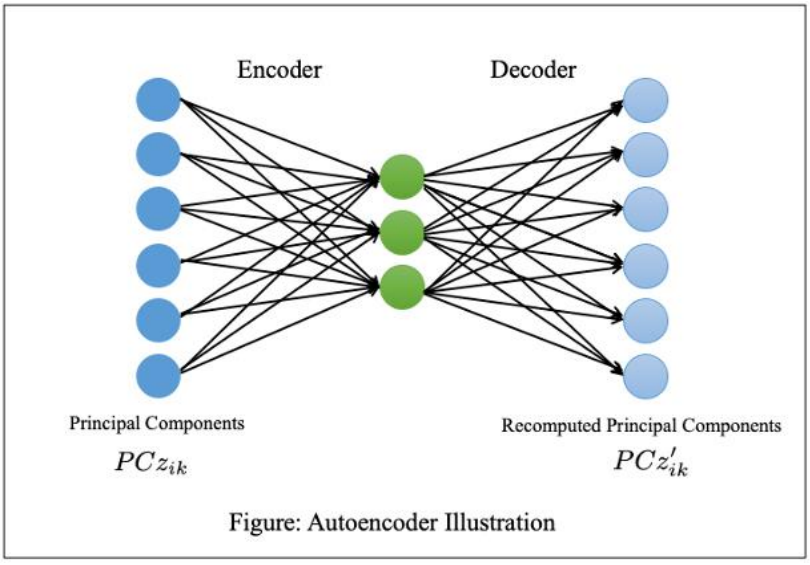


Figure: Autoencoder Illustration

We compiled the autoencoding model using the Keras library in Python. The auto-encoding model was trained using Z-scaled principal components with optimizer “adam” and loss function “mean squared error”. The model only runs for 3 epochs, but it is enough to identify abnormal records because the normal records will have much smaller errors than abnormal records even when auto-encoded by a relatively weak model. The encoder takes the 6-dimensional principal components as input and then compresses it to three dimensions using a dense neural net with activation function “tanh”. The reconstruction uses another dense neural net with “tanh” and takes the encoded representation and decodes it back to a 6-dimensional record. The autoencoder error of each principal component is calculated by subtracting the original value from the reproduced value. And finally, the second fraud score is calculated by taking the Euclidean distance with power of 2. The function is shown below where “PCz” denotes Z-Scaled principal component, “PCz’” denotes the recomputed Z-Scaled principal component, “i” denotes the record number, and “k” denotes the dimension.

$$s_{2i} = \left(\sum_{k}^{6} |PCz'_{ik} - PCz_{ik}|^2 \right)^{1/2}$$

5.3 Final Score

The final anomaly score combined the first and second models. Instead of using fraud scores calculated by models, we generated a rank order for all records based on the value of each fraud score to put them into the same scale. The rank orders range from 1 to 1046826, and a record with higher fraud score will get a higher rank order. Assuming both models are equally important to our final evaluation, the final anomaly score is calculated by taking the average of two rank orders. The properties with higher anomaly score (e.g. 1046826) are more suspicious than properties with lower anomaly score.

6 Results

Record # 684704
This property has extreme values in the R1, 4, & 7 variables which means that something is off about the lot size. The lot size and depth are both listed as being 2 feet, leaving a lot size of 4 square ft. The building front is reportedly 0 ft and the building depth is 0 ft, which is likely the result of missing values. The street address listed is also incomplete, '69 Street', and is likely missing the house number. The tax class is given as 1B, residential vacant land, so it would be odd to have a 4-story building on the property. The owner is listed as a 'W Rufert.'

Record # 14979
This record has extreme values for all of the expert variables but especially those associated with the building size and volume. This property has a listed building front of 8 ft and a depth of 6 ft, leading to a building size of 48 sq ft. The property is listed as having 1 story giving a volume of 48 cubic ft. Looking at the building on Google Maps we can clearly see that this is wrong. The property is also classified in tax class 4 "All others", when it appears that it should be in tax class 2 which is for apartment buildings.

Record # 1065870
This record has high values in the variables associated with building size and building volume. Although the lot front and lot depth for this property is large, building front, building depth, building size and number of stories are 0. This property is owned by PEOPLE OF THE ST OF N, and is located in HYLAN BOULEVARD, but there is no zip-code. It is part of the 1B tax-class, which means it is categorized as residential vacant land. This would explain why there are no values for building front, building depth, building size, and stories, and why the respective z-scaled values are extreme.

Record # 1059883
This property has extremely high R1, R4, R7 values, which indicates that the lot area used is unusual. On the original data, lot front (LTFRONT) and lot depth (LTDEPTH) were both recorded as 5 feet which is abnormal for a property in tax class 4. Additionally, values are missing from variable STORIES, FULLVAL, AVLAND, AVTOT, BLDFRONT, and BLDDEPTH for this property, which could also be a reason why the pipeline gave high anomaly scores.

Record # 973912
This property has extremely high R2, R3, R5, R6, R8, R9, which indicates that the building area and building volume are unusual. The property has a building front of 2 feet and building depth of 12 feet which is relatively low given the total market value (FULLVAL) of 3090000, the assessed land value (AVLAND) of 1309500, and the assessed total value (AVTOT) of 1390500

dollars. The owner listed was “THE PORT OFNY & NJ " which is a government agency, which might be a reason for the extreme values.

Record # 151044

The valuation of this property is more than 1.66 billion dollars, while the building frontage and building depth of this property are both 0 ft. These zero values indicate that the data entry for the building frontage and depth of this property was missing. Although the mean value of the building frontage and depth (62.08 ft and 86.30 ft, respectively) were utilized in calculating the z-scores, all the scores still turned out to be extremely large. However, searching for this property on Google Maps revealed that this property is Yankee Stadium, which explains why this property has such a high valuation. This high valuation contributed to driving the z-scores high, and thus this property was classified as unusual by the algorithm.

Record # 330291

This property has high z-scores for the variables involving R2, R3, R5, R6, R8, and R9 indicating that there is some anomaly in the building area.
The building frontage and depth for this property are respectively 6 ft and 8ft, and these values are extremely small for a building with the valuation of $2,772,746.
Looking at the property on Google Maps, we can observe that the building is easily larger than 6 ft by 8 ft, and it is likely that there was a miss-entry for this property’s building frontage and depth.

Record # 917942

This property belongs to a Logan Property, INC. It is shown to be 3 stories high. The lot front is 4910 feet, but the lot depth equals to . Along with the lot depth, the building front and building depth also equals 0. The valuation of the property is shown as $374M, which is very odd to have a building that is non existent according to the building front and depth cost way above the average valuation of property making this record anomalous.

Record # 638993

This property is located on Horace Harding Expressway, owned by Alexander’s of Regopa.
The valuation of the property is set to be 2.3M. Despite the value being set at a higher price than the average, the lot front is 379 feet with the lot depth being 205 and 5, 6 feet for building depth and building front respectively. The lot front and the lot depth are multiple times above the average, but the building depth and front are multiple times below the average which estimates that the valuation of the property shouldn’t be as highly priced making this record anomalous.

Record # 60469

This property belongs to 42/9 Residential LLC and is located in 360 West 43 Street. The record shows the Loot Front and Lot depth is 20 and 50 respectively. The issue is that the building depth and building front is 200 and 210 respectively. It does not make much sense for the building size to be larger than the lot size. The extreme z-score values for R1, R4, and R7 make sense since they are associated with lot size. This lot size is too small for a property of such value and with respect to building size. Further investigation would be needed to determine if this was a data entry error or case of fraud.

7 Summary and Conclusion

In this project, we created machine learning models to help identify fraudulent property tax records by detecting unusual records. Before building these unsupervised models, we sought to understand the basic characteristics of the New York property data. We wrote out the definition of each variable and performed Exploratory Data Analysis (EDA) on every variable to create a Data Quality Report (DQR). This served as a check to make sure the data is correct and helped us familiarize ourselves with the data.

After understanding the data and confirming it was correct, we created 45 expert variables based on research and expert recommendations. We cleaned the data by removing unnecessary records and filling in null values present in fields that were used to create these expert variables. We determined each property's lot size, building size and building volume, and then created 9 ratios. Using these nine ratios, we compared the values to the means within each three and five digit zip code, tax class and Borough. Each ratio was compared to the mean in the four groups, totaling 36 variables measuring the abnormality of each record.

In order to reduce dimensionality and multicollinearity in these variables, Principal Component Analysis was used to reduce the 45 variables down to 6 principle components. The modeling consisted of two separate strategies. The first model combines the z-scores with a heuristic algorithm and looks for extremes. The second model was measuring the error of an autoencoder trained on the reduced dimensionality data. There were no substantial differences in model performance between the z-scale outlier model and the autoencoder error. A final score was determined by taking the average of the scores from the two models.

The final score allowed us to sort the records from the most unusual records to the least unusual. Seeing as our expertise is not in the domain of property tax fraud, we used abnormality as a potential indicator of fraud. In our preliminary investigation, we determined that these abnormalities could simply be caused by data entry error and that further investigation by domain experts is necessary.

These fraud algorithms worked well and helped identify properties that are anomalous, and further investigation will confirm whether these unusual records are cases of fraud. However, there is room for improvement. Only 45 variables were used due to our limited field knowledge and it is not uncommon for 100 or more to be created. Doing so would require more time investment in speaking to property tax fraud experts. We also believe that our methodology for imputing missing values and cleaning the data could be improved (specifics). We learned that the most important part of any data analytics project is designing an approach to obtain our desired outcome, then understanding and exploring the data and then engineering expert variables that can be used to indicate our feature of interest. In our case, this was property tax fraud. Many

analysts tend to become infatuated with the most advanced tools and algorithms, rather than properly designing a solution to the problem at hand.

Appendix: Data Quality Report

I. Data Description

Dataset Name: Property Valuation and Assessment Data

Dataset Source: NYC Open Data, Department of Finance (DOF)

Dataset Purpose: Data represent NYC properties assessments for purpose to calculate Property Tax, Grant eligible properties Exemptions and/or Abatements.

Time Period: November 17, 2010

Number of Fields: 32

Number of Records: 1,070,994

II. Data Summary

i. Numeric Fields:

Field Name	Field Type	# Records	% Populated	Mean	Standard Deviation	Min	Max	# Zeros
LTFRONT	Numeric	1070994	100.00%	36.64	74.03	0	9999	169108
LTDEPTH	Numeric	1070994	100.00%	88.86	76.40	0	9999	170128
STORIES	Numeric	1014730	94.75%	5.01	8.37	1	119	0
FULLVAL	Numeric	1070994	100.00%	874264.51	11582430.99	0	6150000000	13007
AVLAND	Numeric	1070994	100.00%	85067.92	4057260.06	0	2668500000	13009
AVTOT	Numeric	1070994	100.00%	227238.17	6877529.31	0	4668309000	13007
EXLAND	Numeric	1070994	100.00%	36423.89	3981575.79	0	2668500000	491699
EXTOT	Numeric	1070994	100.00%	91186.98	6508402.82	0	4668309000	432572
BLDFRONT	Numeric	1070994	100.00%	23.04	35.58	0	7575	228815
BLDDEPTH	Numeric	1070994	100.00%	39.92	42.71	0	9393	228853
AVLAND2	Numeric	282726	26.40%	246235.72	6178962.56	3	2371005000	0
AVTOT2	Numeric	282732	26.40%	713911.44	11652528.95	3	4501180000	0
EXLAND2	Numeric	87449	8.17%	351235.68	10802212.67	1	2371005000	0
EXTOT2	Numeric	130828	12.22%	656768.28	16072510.17	7	4501180000	0

ii. Categorical Fields:

Field Name	Field Type	# Records	% Populated	# Unique	Most Common
RECORD	Categorical	1070994	100.00%	1070994	N/A
BBLE	Categorical	1070994	100.00%	1070994	N/A
B	Categorical	1070994	100.00%	5	4
BLOCK	Categorical	1070994	100.00%	13984	3944
LOT	Categorical	1070994	100.00%	6366	1
EASEMENT	Categorical	4636	0.43%	13	E
OWNER	Categorical	1039249	97.04%	863348	PARKCHESTER PRESERVAT
BLDGCL	Categorical	1070994	100.00%	200	R4
TAXCLASS	Categorical	1070994	100.00%	11	1
EXT	Categorical	354305	33.08%	4	G
EXCD1	Categorical	638488	59.62%	130	1017
STADDR	Categorical	1070318	99.94%	839281	501 SURF AVENUE
ZIP	Categorical	1041104	97.21%	197	10314
EXMPTCL	Categorical	15579	1.45%	15	X1
EXCD2	Categorical	92948	8.68%	61	1017
PERIOD	Categorical	1070994	100.00%	1	FINAL
YEAR	Datetime	1070994	100.00%	1	2010/11
VALTYPE	Categorical	1070994	100.00%	1	AC-TR

III. Data Field Exploration

i. RECORD

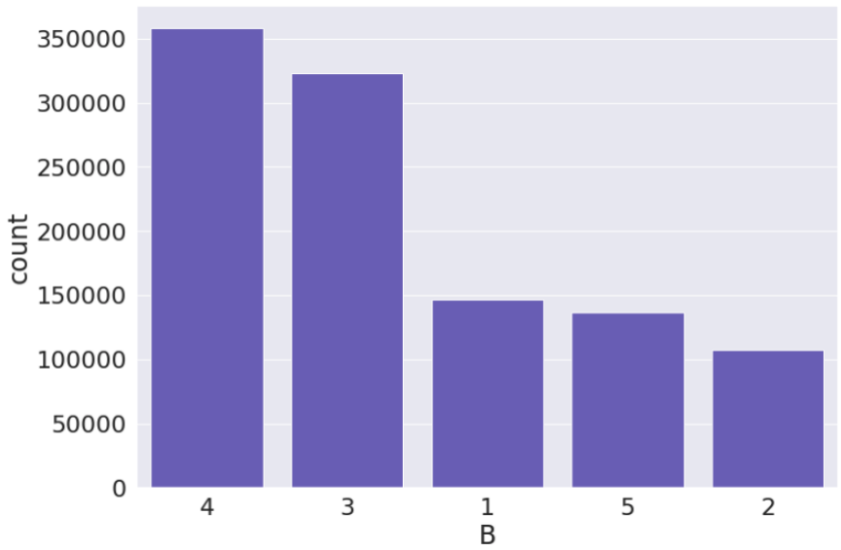
The unique identifier of each entry in the data. It consists of numbers from 1 to 1070994.

ii. BBLE

The concatenation of borough code, block code and lot. It consists of 10-digit numbers.

iii. B

Borough codes. The distribution plot:



iv. BLOCK

Valid block ranges by borough:

MANHATTAN 1 TO 2,255,

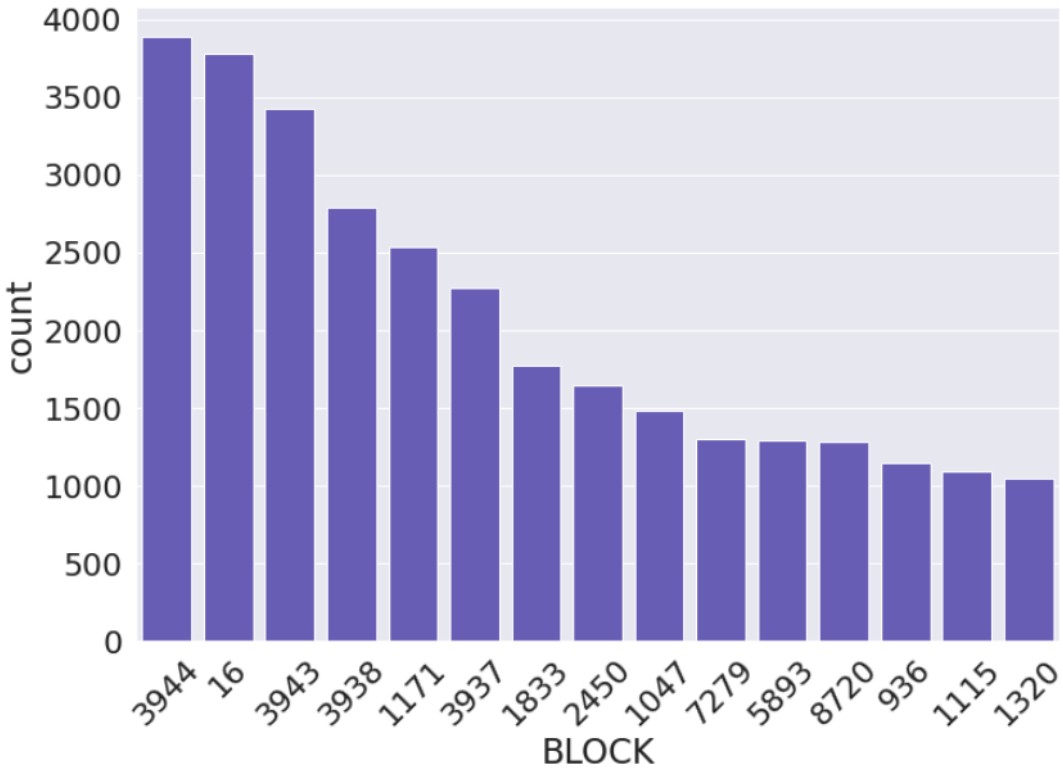
BRONX 2,260 TO 5,958,

BROOKLYN 1 TO 8,955,

QUEENS 1 TO 16,350,

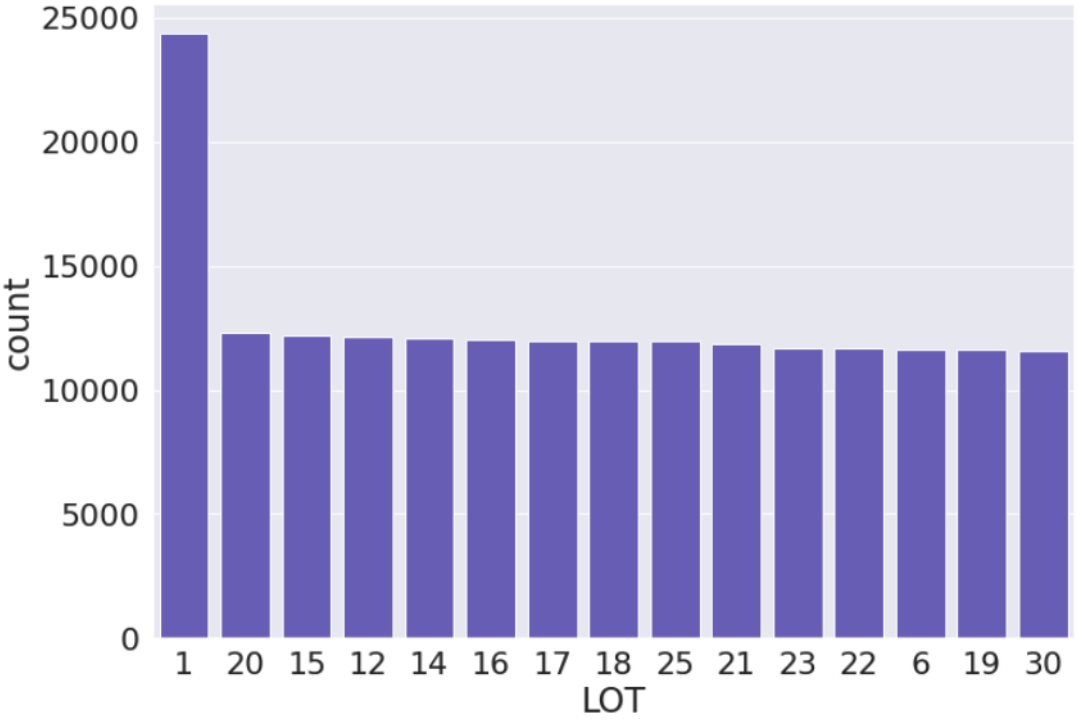
STATEN ISLAND 1 TO 8,050.

The distribution of top 15 field values:



v. LOT

The unique number within a borough or block. The distribution of top 15 values:



vi. EASEMENT

The field that is used to describe easement:

SPACE Indicates the lot has no Easement.

'A' Indicates the portion of the Lot that has an Air Easement

'B' Indicates Non-Air Rights.

'E' Indicates the portion of the lot that has a Land Easement

'F' THRU 'M' Are duplicates of 'E'.

'N' Indicates Non-Transit Easement

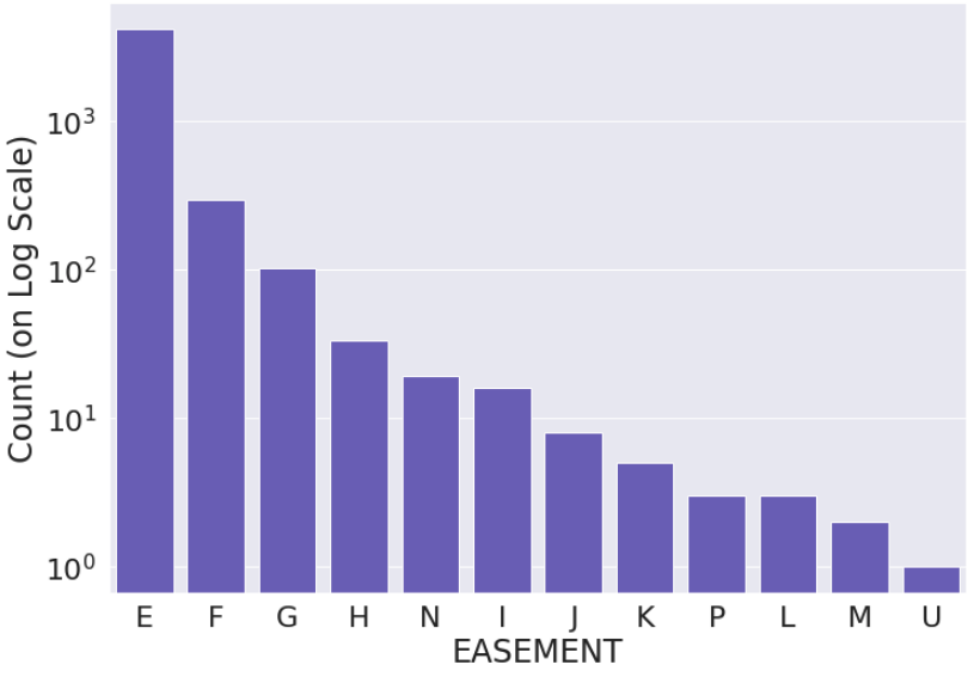
'P' Indicates Piers.

'R' Indicates Railroads.

'S' Indicates Street

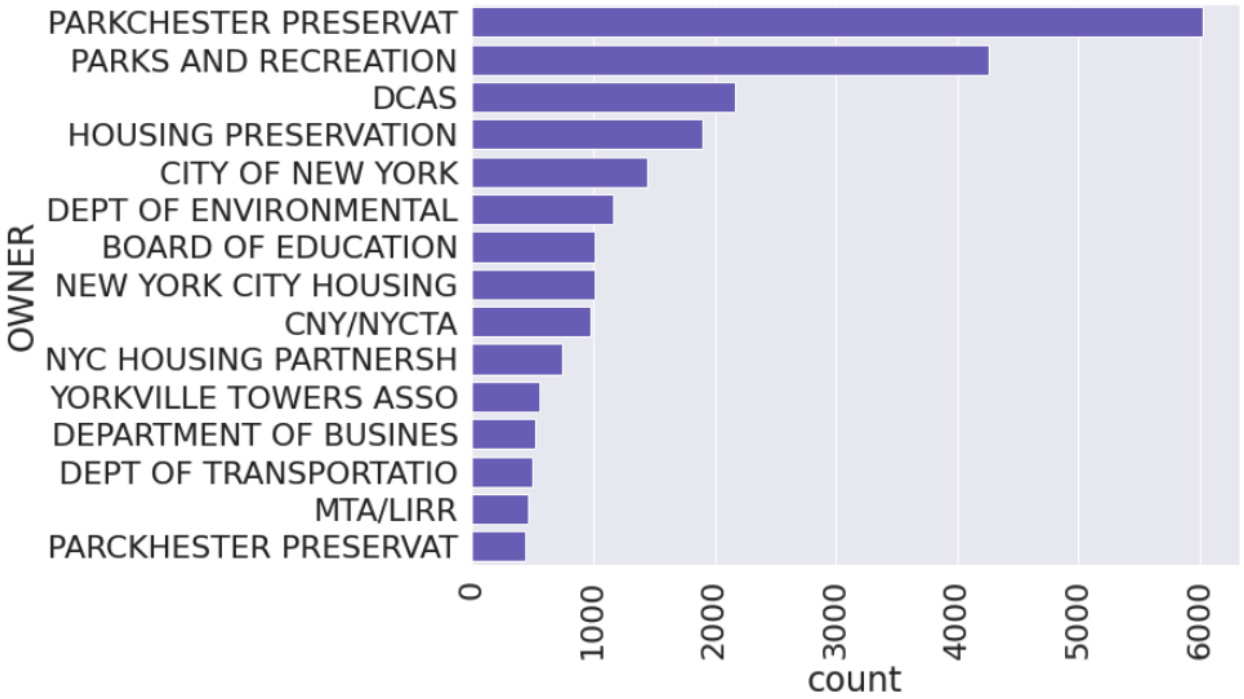
'U' Indicates U.S. Government

The distribution on log scale:



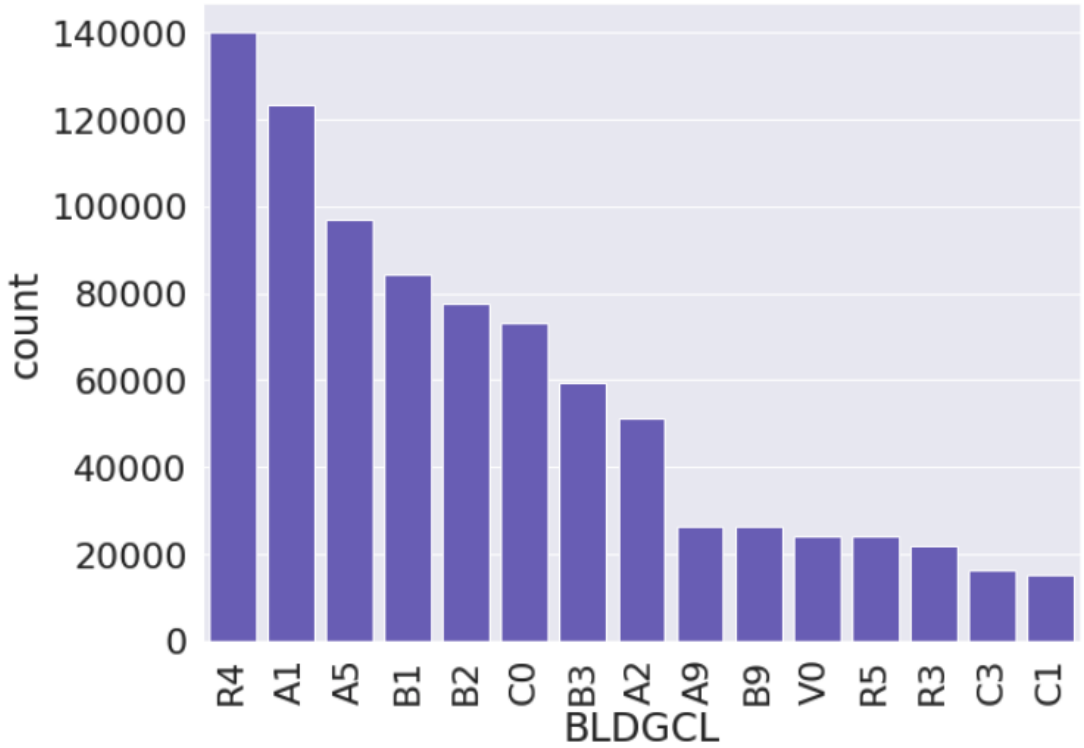
vii. OWNER

Owner's Name. The distribution of top 15 field values:



viii. BLDGCL

Building class. The distribution of top 15 field values:

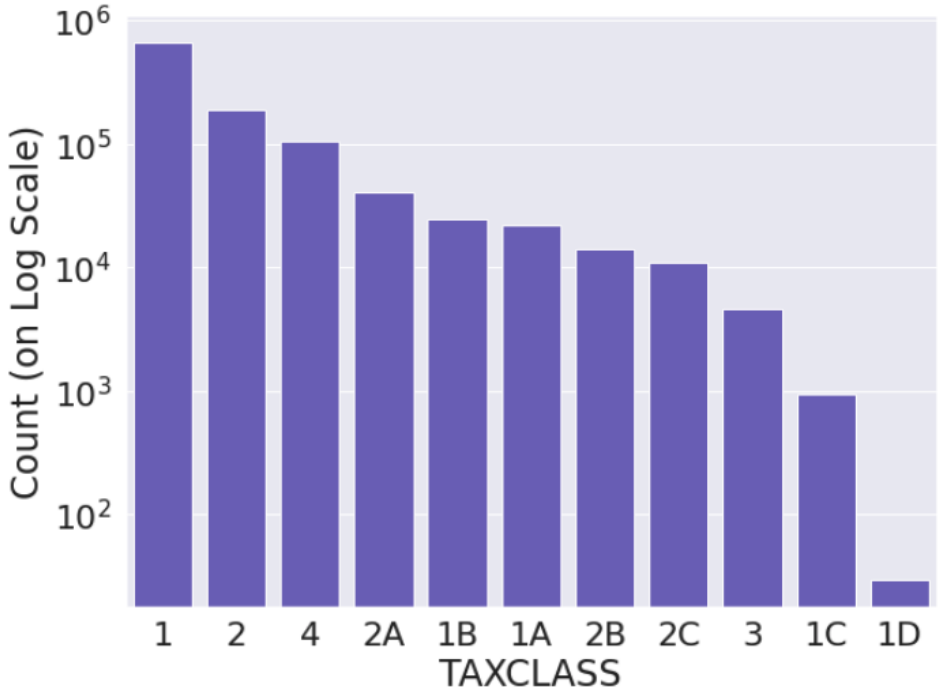


ix. TAXCLASS

Current Property Tax Class Code (NYS Classification):

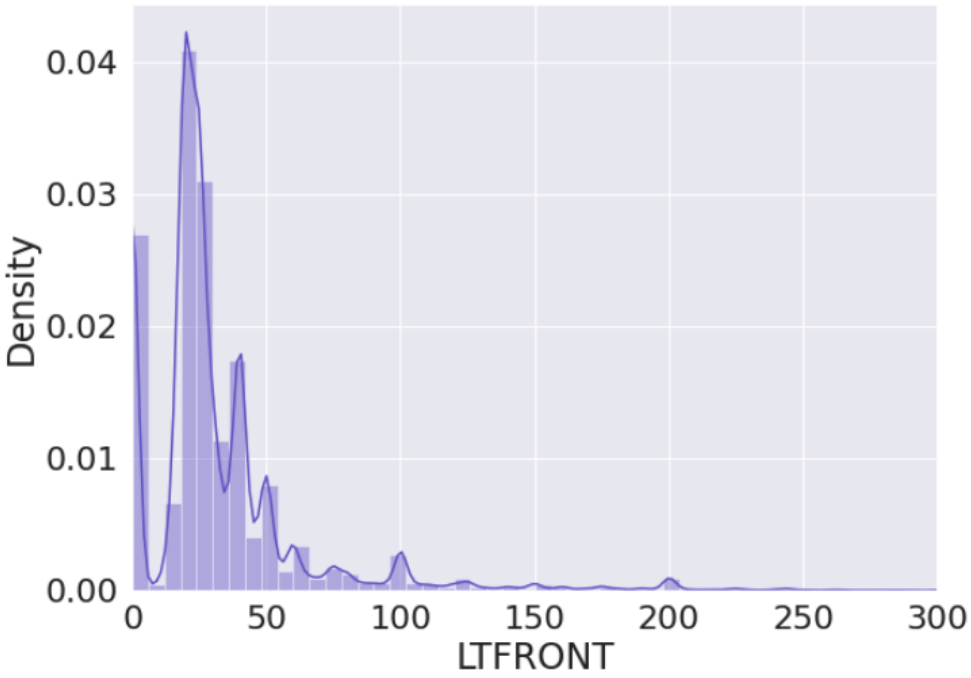
TAX CLASS	UNITS
TAX CLASS 1	1-3 UNIT RESIDENCES
TAX CLASS 1A	1-3 STORY CONDOMINIUMS
TAX CLASS 1B	RESIDENTIAL VACANT LAND
TAX CLASS 1C	1-3 UNIT CONDOMINUMS
TAX CLASS 1D	SELECT BUNGALOW COLONIES
TAX CLASS 2	APARTMENTS
TAX CLASS 2A	APARTMENTS WITH 4-6 UNITS
TAX CLASS 2B	APARTMENTS WITH 7-10 UNITS
TAX CLASS 2C	COOPS/CONDOS WITH 2-10 UNITS
TAX CLASS 3	UTILITIES (EXCEPT CEILING RR)
TAX CLASS 4A	UTILITIES - CEILING RAILROADS
TAX CLASS 4	ALL OTHERS

The distribution plot on log scale:



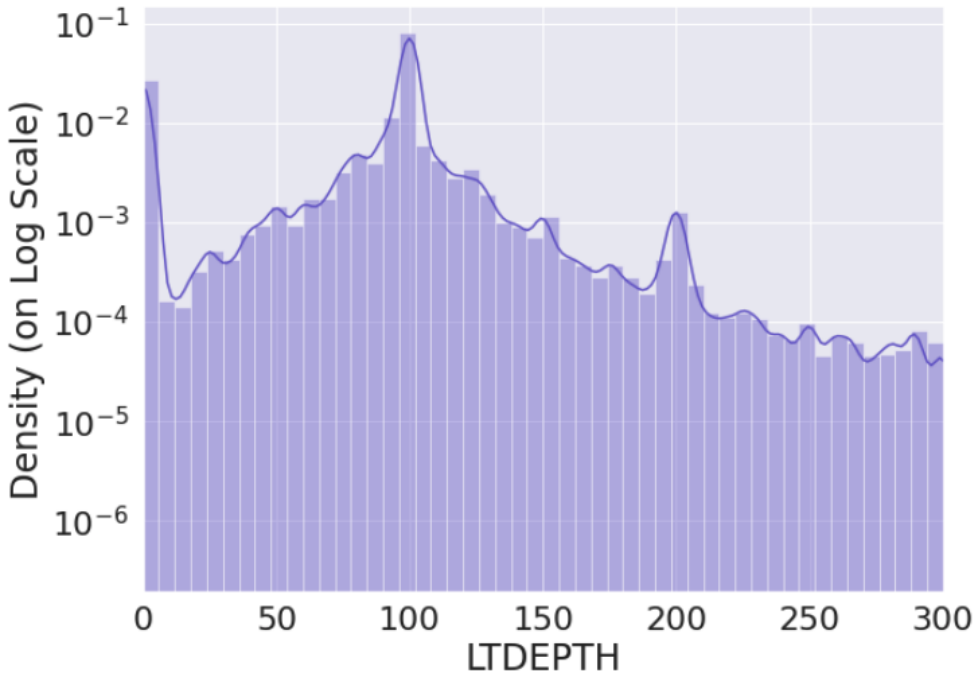
x. LTFRONT

Lot Frontage in feet. Distribution of 99.30% of data, excluding outliers greater than 300:



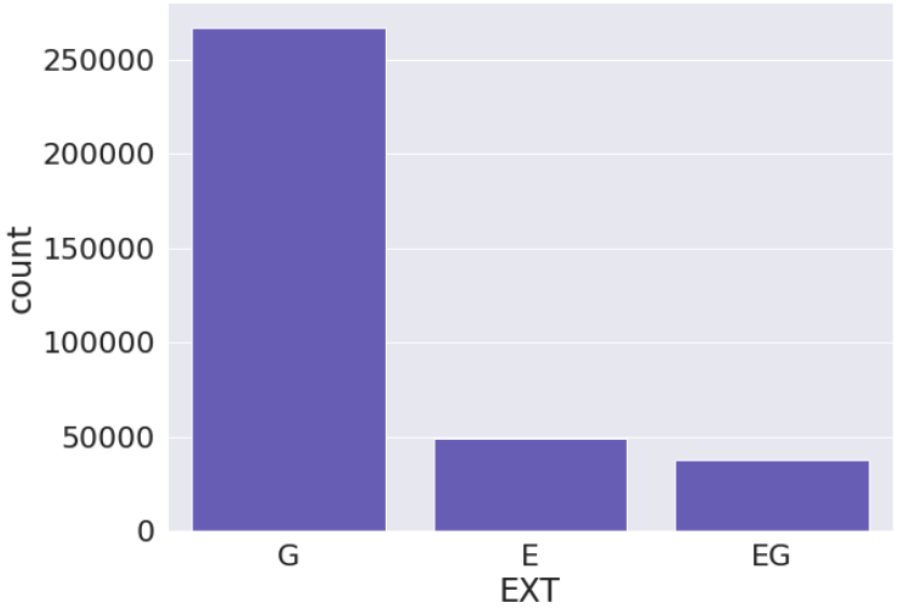
xi. LTDEPTH

Lot depth in feet. Distribution excluding outliers greater than 300 (99.17% of data) in log scale:



xii. EXT

Extension: 'E' = EXTENSION; 'G' = GARAGE; 'EG' = EXTENSION AND GARAGE.

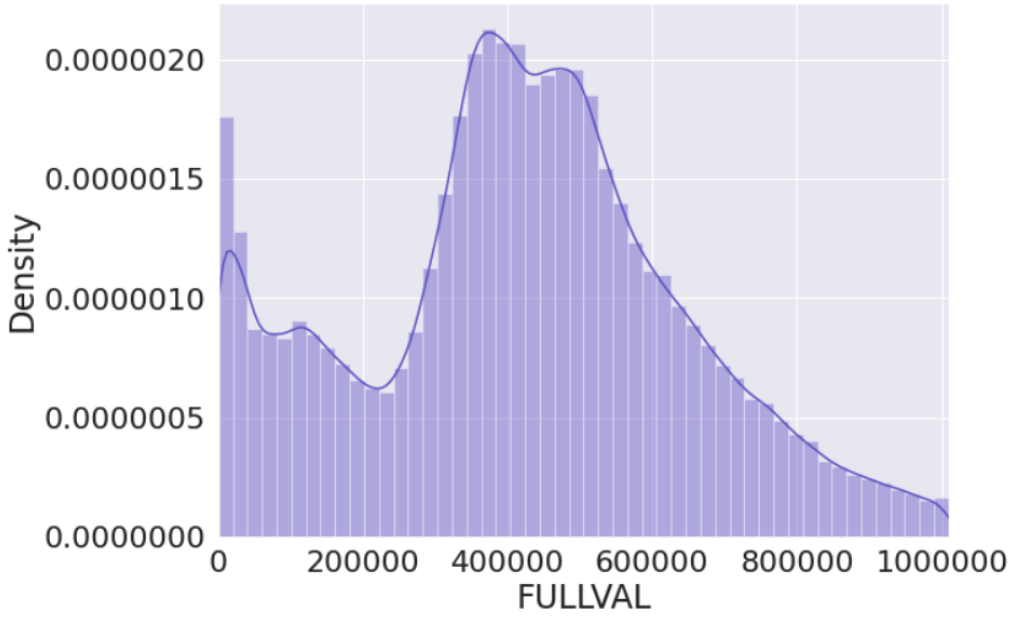


xiii. STORIES

The number of stories for the building. Distribution of 87.86% of data excluding outliers greater than 15:

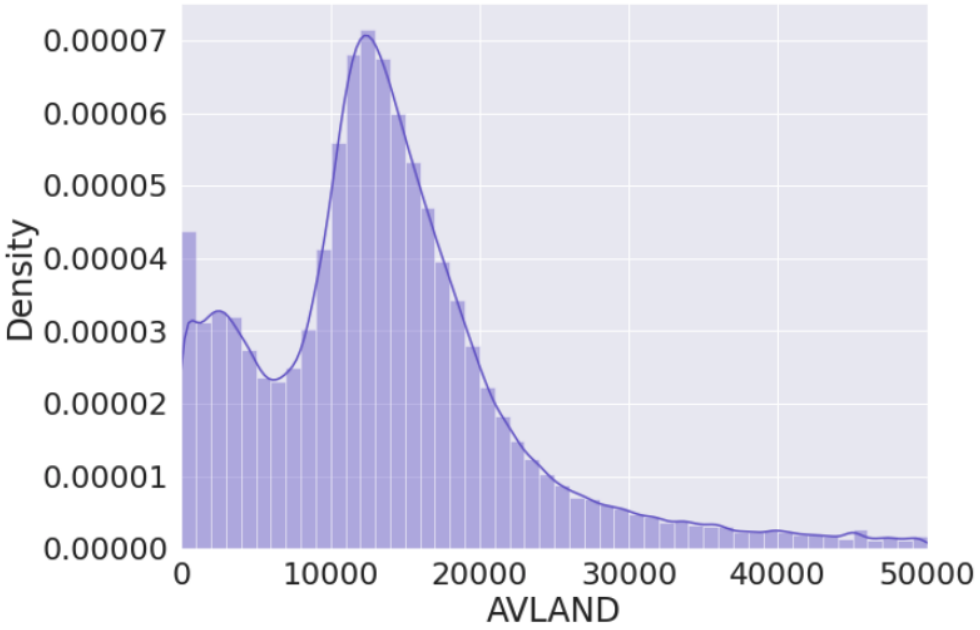
xiv. FULLVAL

Total market value of the land. Distribution of 91.48% of the data, excluding outliers greater than 1010000:



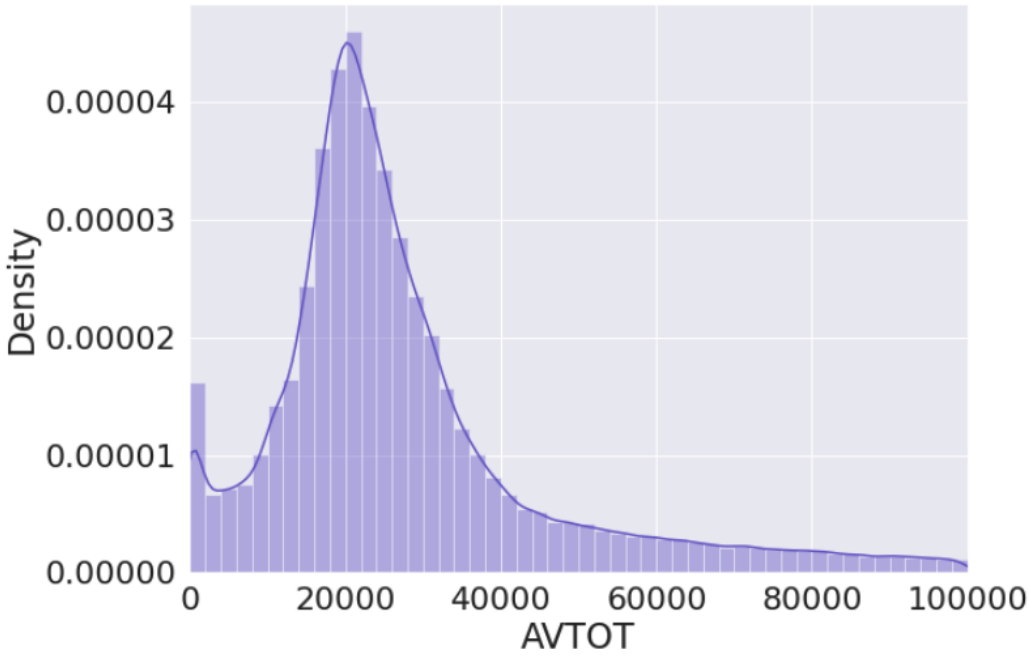
xv. AVLAND

Assessed value of the land. Distribution of 90.53% of data, excluding outliers greater than 50000:



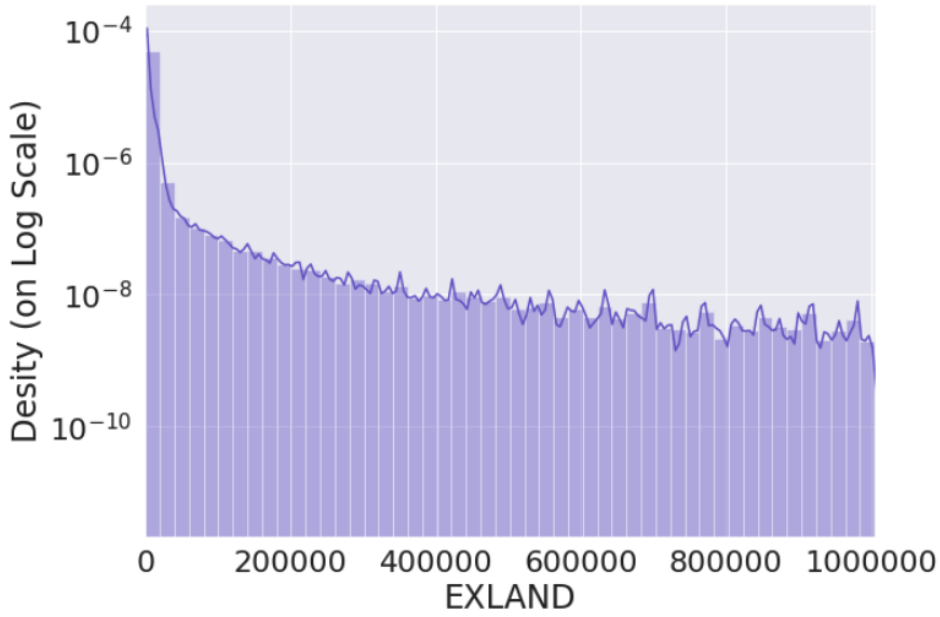
xvi. AVTOT

Assessed total value. Distribution of 86.05% of data, excluding outliers greater than 100000:



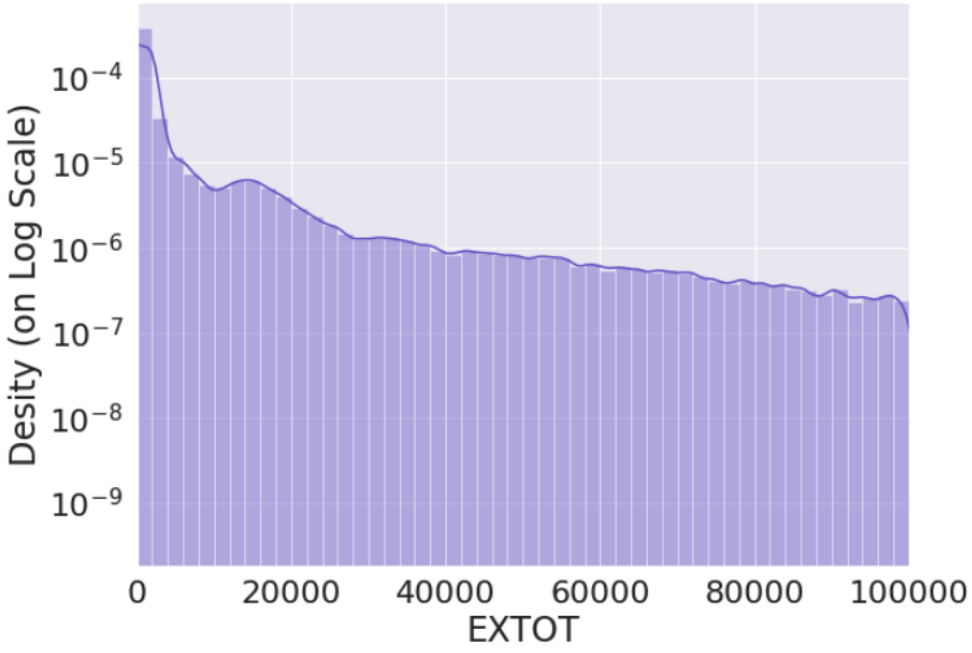
xvii. EXLAND

Exempt land value. Distribution of 99.63% of data, excluding outliers greater than 1000000 in log scale:



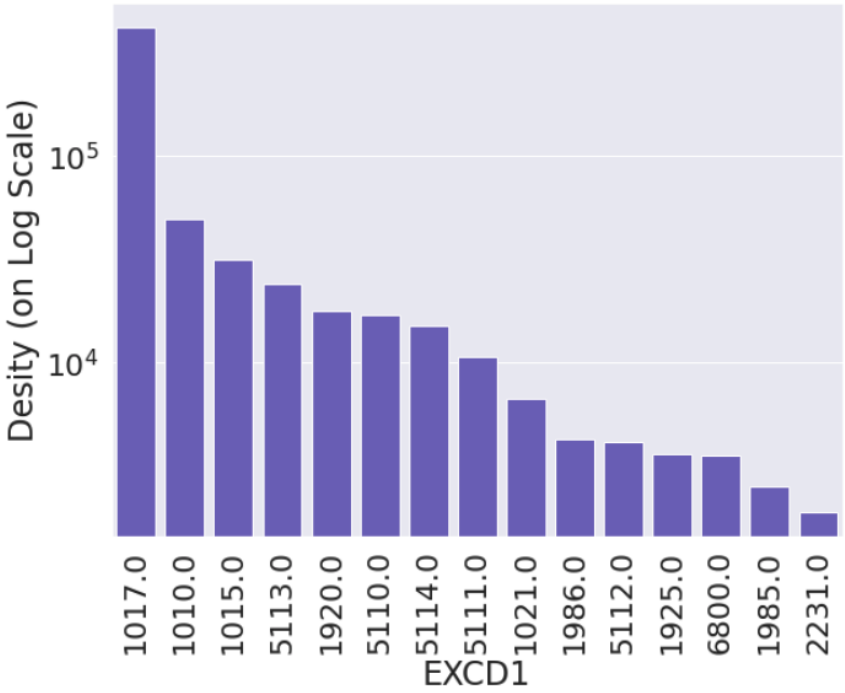
xviii. EXTOT

Exempt total value. Distribution of 96.45% of data, excluding outliers greater than 100000 in log scale:



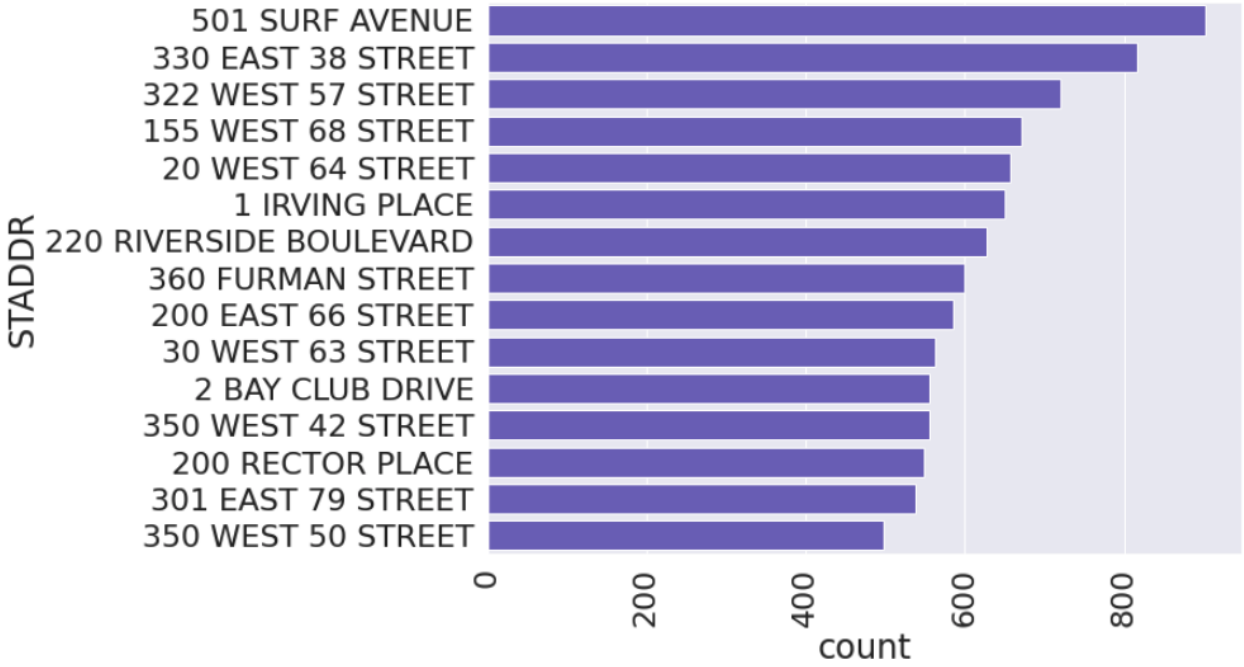
xix. EXCD1

Old exempt property restored date. The distribution of top 15 field values in log scale:



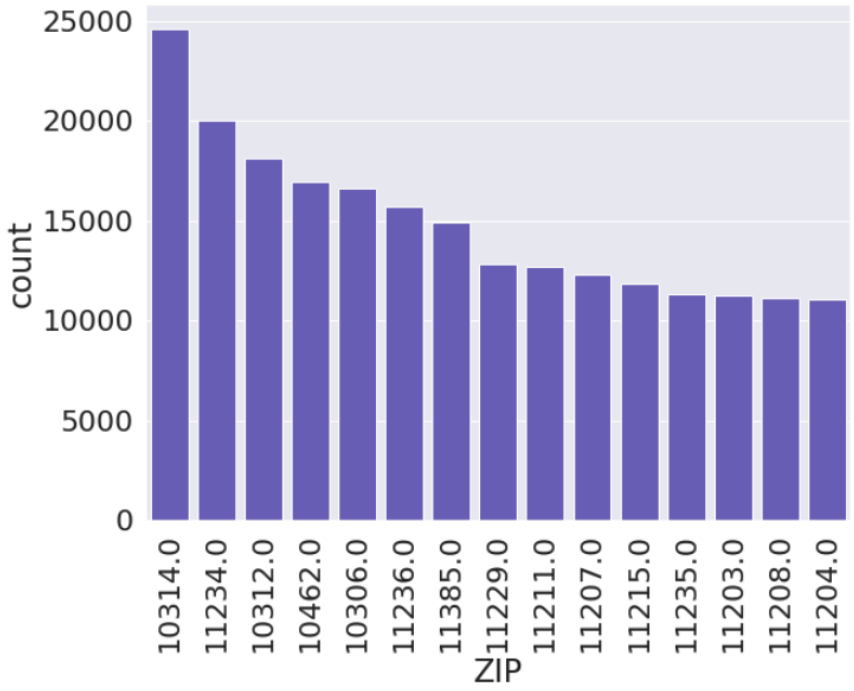
xx. STADDR

Street address of the property. The distribution of top 15 field values:



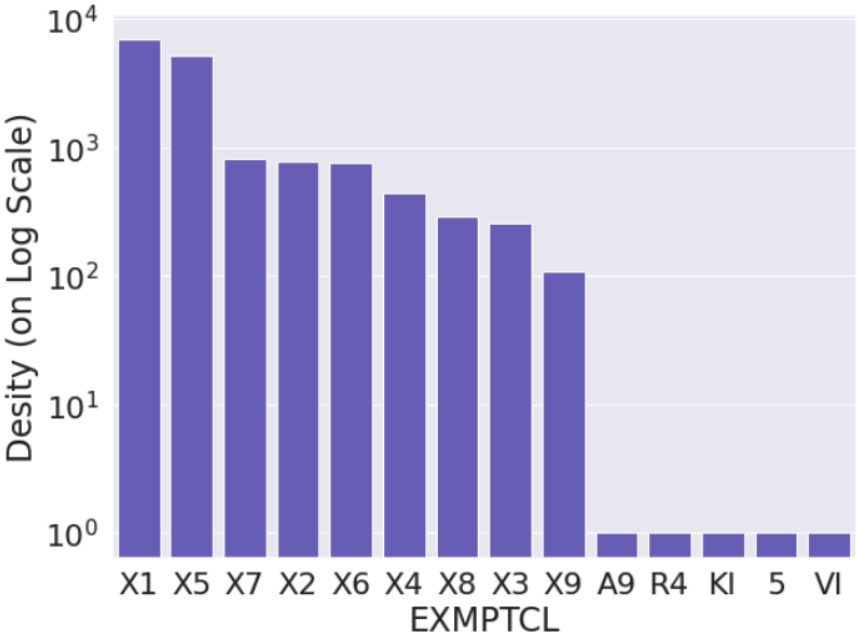
xxi. ZIP

Postal zip code of the property. The distribution of top 15 field values:



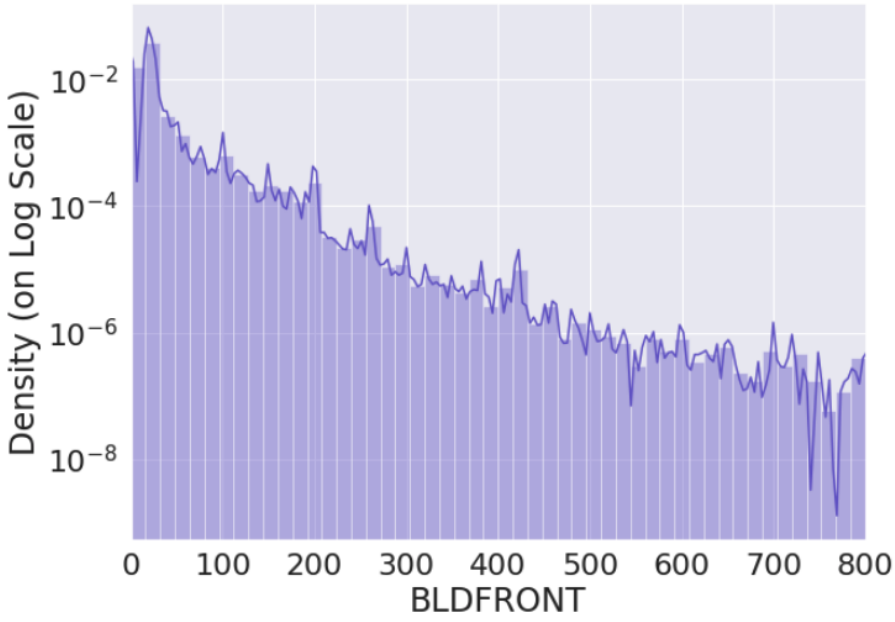
xxii. EXMPTCL

Exempt class used for fully exempt properties. The distribution of top 15 field values in log scale:



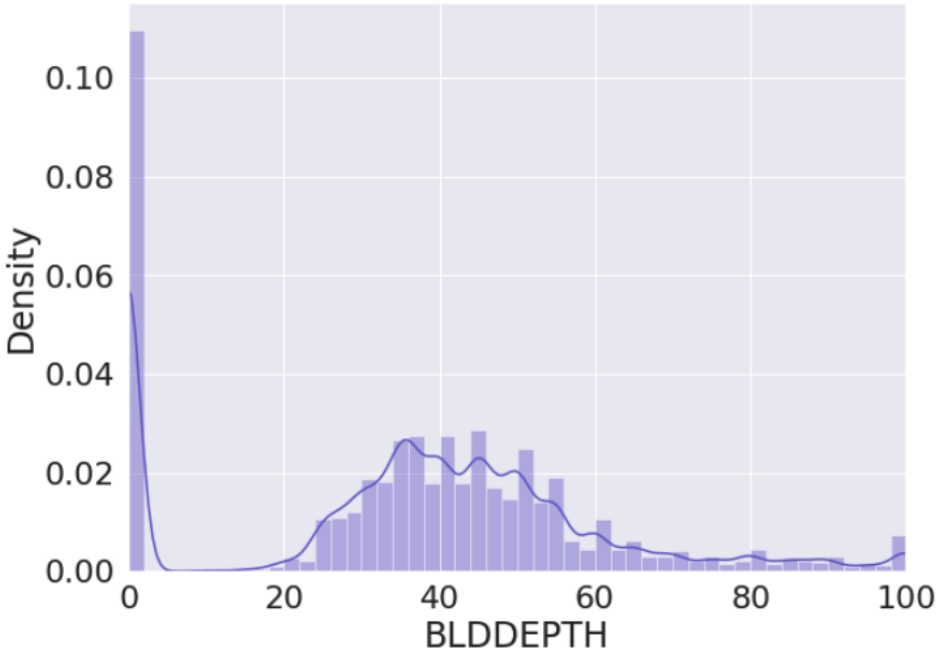
xxiii. BLDFRONT

Building frontage in feet. Distribution of 99.99% of data, excluding outliers greater than 800 in log scale:



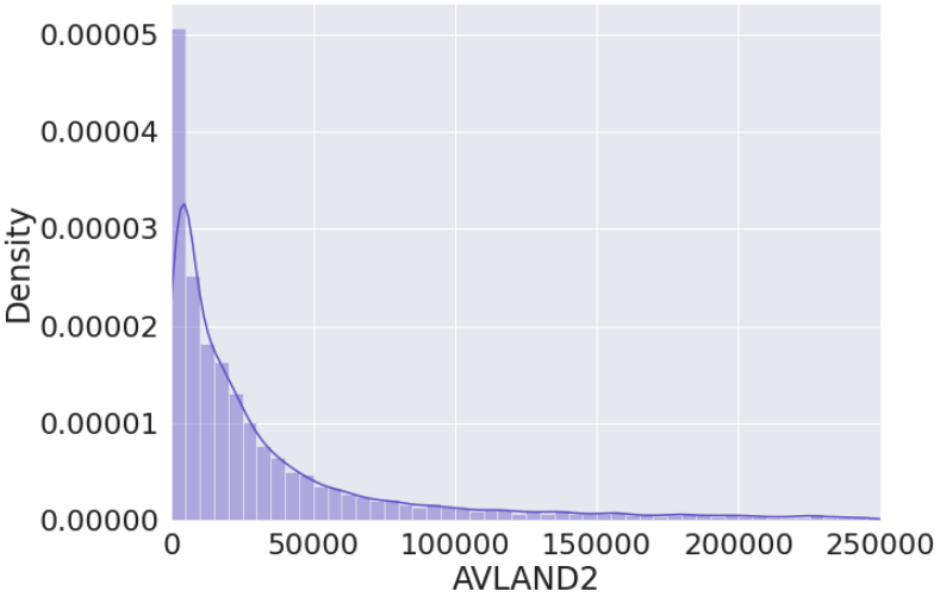
xxiv. BLDDEPTH

Building depth in feet. Distribution of 97.40% of data, excluding outliers greater than 100:



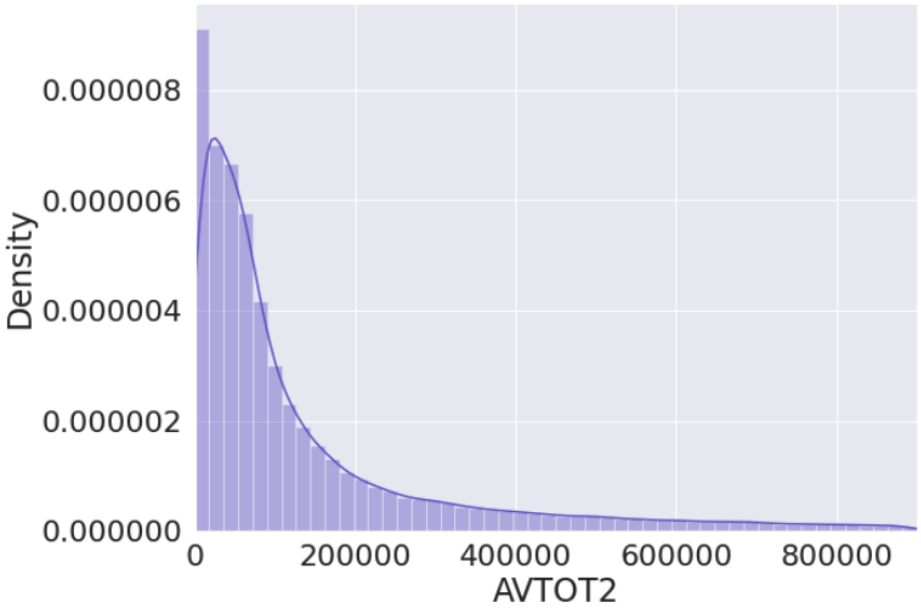
xxv. AVLAND2

New assessed value of land. Distribution of 90.22% of data, excluding outliers greater than 250000:



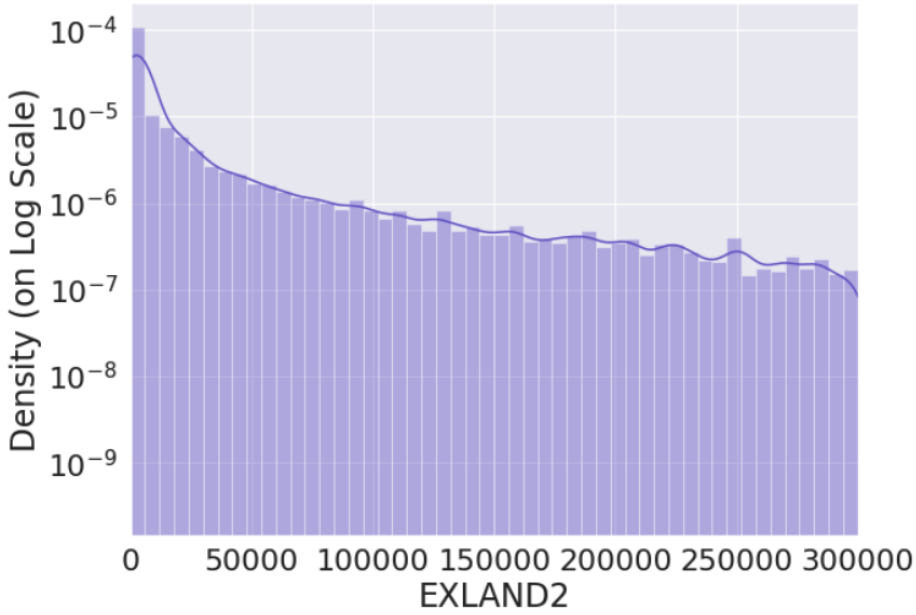
xxvi. AVTOT2

New total assessed value of the property. Distribution of 90.83% of data, excluding outliers greater than 900000:



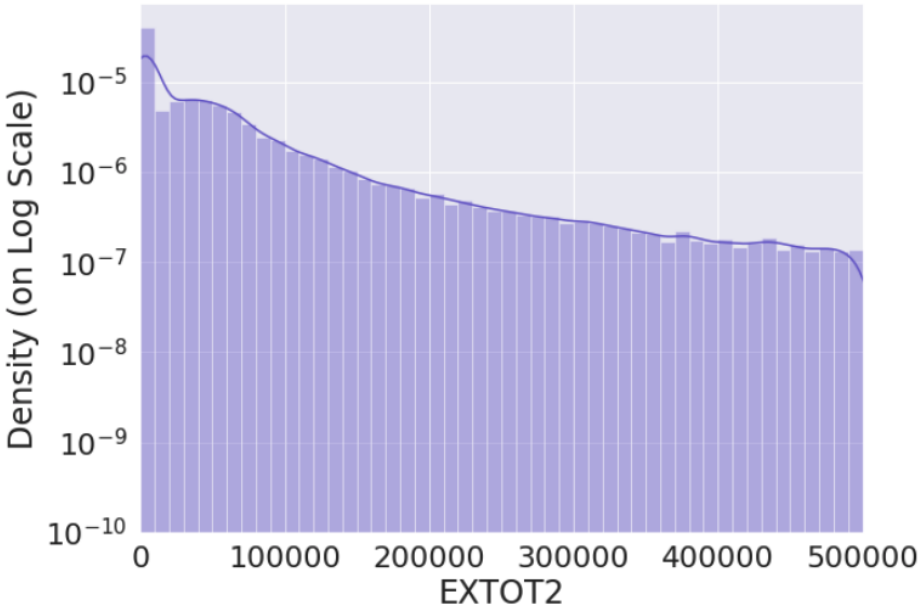
xxvii. EXLAND2

New exempt land value. Distribution of 90.91% of data, excluding outliers greater than 300000, on a log scale:



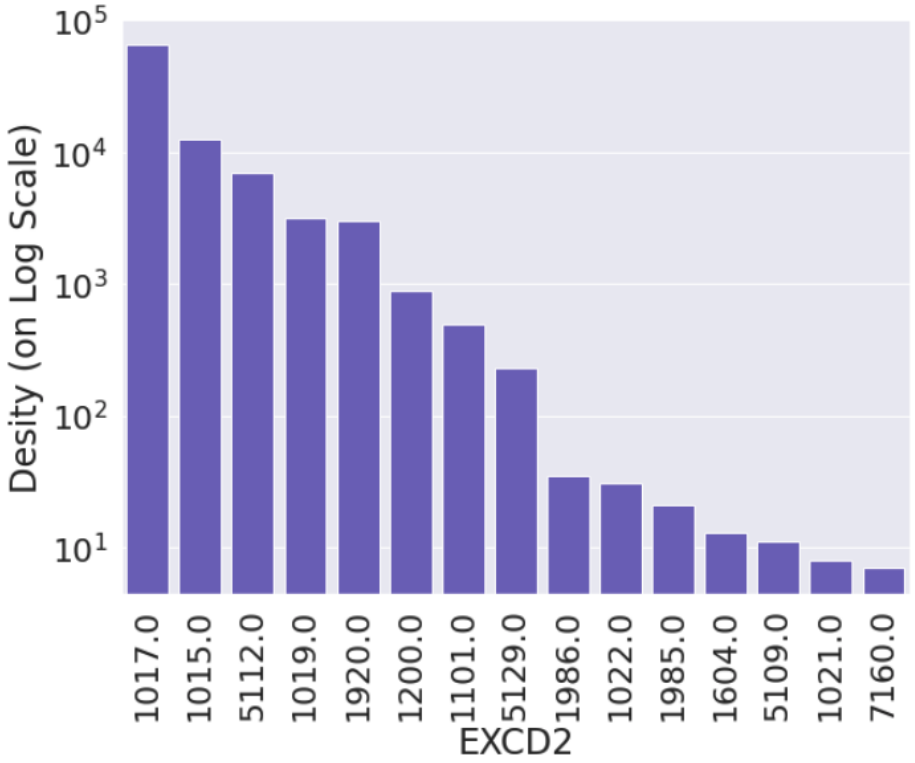
xxviii. EXTOT2

New exempt total value of the property. Distribution of 90.01% of data, excluding outliers greater than 500000, on a log scale:



xxix. EXCD2

New exempt property restored date. The distribution of top 15 field values on a log scale:



xxx. PERIOD

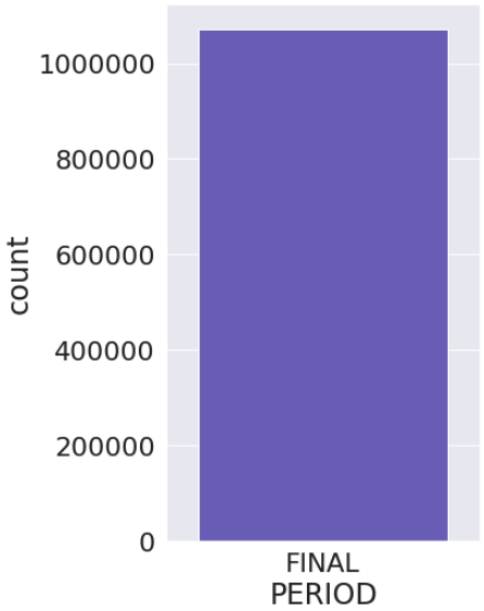
Indicator for the change of period of file:

'0' = TENTATIVE,

'C' = CHANGE BY NOTICE,

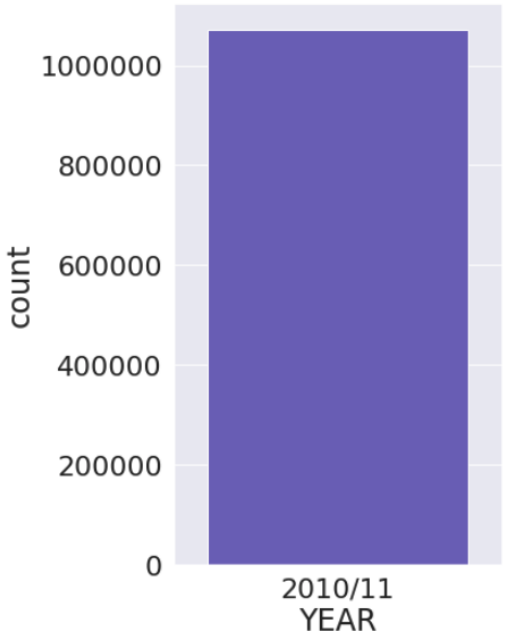
'1' = FINAL.

All records have the same value for PERIOD:



xxxi. YEAR

Year and month of the file. All records have the same value for YEAR:



xxxii. VALTYPE

Identifier of where data was extracted. All records have the same value for VALTYPE: