NY Property Project Report

Jiayi (Lily) Hu, Chaiyapuk (KK) Titinanapun

Index

1. Executive Summary

This project aims to investigate the potential property tax fraud at the city of New York. The main challenge of this project is that there is no data for previous tax fraud available. Yet, the city of New York suspect that property tax fraud may happen unnoticeably by fraudsters misrepresenting their property characteristics.

In order to investigate such a case, we used unsupervised learning method with main focus on the ratio between property values and sizes. After data cleaning, we create variables and define potential tax fraud as the outliers in terms of value per size, both overall value and the value compared to the group of property (taxclass, zipcode). We use Principal Component Analysis (PCA) to reduce dimensions and use the result from PCA to calculate the anomaly score, which is higher as the property become an outlier. Finally, we sort and filter the top 1,000 records based on the anomaly score and investigate those abnormal records to find the potential frauds. The investigation mainly shows properties with abnormally low value or abnormally big size, which needs to be investigated further to confirm whether they are wrong inputs or frauds.

2. Description of Data

This dataset is about property data of New York City, which comes from NYC Open Data. This dataset contains data about each property's owner, class, size, actual values, transitional values, etc. The duration of events covered in this dataset is within the month of November, 2010. There are 32 fields (14 are numerical and 18 are categorical) and 1,070,994 records. There is no field acts as the label of fraud or abnormality.

The following table shows the basic summary of the 32 fields.

Table 1. Field Summary Table

Categorical Fields					
Field Name	**% Populated**	**# Blanks**	**# Zeros**	**# Unique Values**	**Most Common**
RECORD	100.00%	0	0	1,070,994	1
BBLE	100.00%	0	0	1,070,994	1000010101
BORO	100.00%	0	0	5	4
BLOCK	100.00%	0	0	13,984	3944
LOT	100.00%	0	0	6,366	1
EASEMENT	0.43%	1,066,358	0	12	E
OWNER	97.04%	31,745	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	100.00%	0	0	200	R4
TAXCLASS	100.00%	0	0	11	1
EXT	33.08%	716,689	0	3	G
EXCD1	59.62%	432,506	0	129	1017
STADDR	99.94%	676	0	839,280	501 SURF AVENUE
ZIP	97.21%	29,890	0	196	10314
EXMPTCL	1.45%	1,055,415	0	14	X1
EXCD2	8.68%	978,046	0	60	1017
PERIOD	100.00%	0	0	1	FINAL
YEAR	100.00%	0	0	1	2010/11
VALTYPE	100.00%	0	0	1	AC-TR

Numerical Fields								
Field Name	**% Populated**	**# Blanks**	**# Zeros**	**Min**	**Max**	**Mean**	**Stdev**	**Most Common**
LTFRONT	100.00%	0	169,108	0	9,999	36.64	74.03	0
LTDEPTH	100.00%	0	170,128	0	9,999	88.86	76.40	100
STORIES	94.75%	56,264	0	1	119	5.01	8.37	2
FULLVAL	100.00%	0	13,007	0	6,150,000,000	874,264.51	11,582,430.00	0
AVLAND	100.00%	0	13,009	0	2,668,500,000	85,067.92	4,057,260.00	0
AVTOT	100.00%	0	13,007	0	4,668,309,000	227,238.17	6,877,529.00	0
EXLAND	100.00%	0	491,699	0	2,668,500,000	36,423.89	3,981,576.00	0
EXTOT	100.00%	0	432,572	0	4,668,309,000	91,186.98	6,508,403.00	0
BLDFRONT	100.00%	0	228,815	0	7,575	23.04	35.58	0
BLDDEPTH	100.00%	0	228,853	0	9,393	39.92	42.71	0
AVLAND2	26.40%	788,268	0	3	2,371,005,000	246,235.72	6,178,963.00	2,408
AVTOT2	26.40%	788,262	0	3	4,501,180,000	713,911.44	11,652,530.00	750
EXLAND2	8.17%	983,545	0	1	2,371,005,000	351,235.68	10,802,210.00	2,090
EXTOT2	12.22%	940,166	0	7	4,501,180,000	656,768.28	16,072,510.00	2,090

3. Data Cleaning

3.1 Data Exclusions

The goal of this project is to detect abnormal property records in the New York City. However, some benign properties may have data with extreme deviation from the mean due to their special characteristics. Those kinds of properties usually belong to governmental owners. Therefore, we organized a list of governmental owners and removed 24,478 property records that we are not interested in according to the removal list. After data exclusion, there are 1,046,516 records left.

Here are the property OWNERs in the removal list:

Table 2. Owners on the remove_list

OWNERs on the remove_list			
1	PARKCHESTER PRESERVAT	18	U S GOVERNMENT OWNRD
2	PARKS AND RECREATION	19	THE CITY OF NEW YORK
3	DCAS	20	NYS URBAN DEVELOPMENT
4	HOUSING PRESERVATION	21	NYS DEPT OF ENVIRONME
5	CITY OF NEW YORK	22	CULTURAL AFFAIRS
6	DEPT OF ENVIRONMENTAL	23	DEPT OF GENERAL SERVI
7	BOARD OF EDUCATION	24	DEPT RE-CITY OF NY
8	NEW YORK CITY HOUSING	25	NY STATE PUBLIC WORKS
9	CNY/NYCTA	26	NYC DEPT OF HIGHWAYS
10	NYC HOUSING PARTNERSH	27	NYC DEP'T OF HIGHWAYS
11	DEPARTMENT OF BUSINES	28	CITY WIDE ADMINISTRAT
12	DEPT OF TRANSPORTATIO	29	DEPT OF PUBLIC WORKS
13	MTA/LIRR	30	NEW YORK CITY
14	PARCKHESTER PRESERVAT	31	THE PORT OFNY & NJ
15	MH RESIDENTIAL 1, LLC	32	NYC DEPT OF PUB WORKS
16	LINCOLN PLAZA ASSOCIA	33	NEW YORK STATE DEPART
17	UNITED STATES OF AMER	34	CITY AND NON-CITY OWN

3.2 Data Imputation

There are 13 fields with missing, and 9 fields with a large amount of 0 and 1, which is highly possible to be frivolous values. However, not all the fields must be cleaned for us to make predictions. The following table illustrates the fields that we applied data cleaning and the imputation logics when clean the data.

Table 3. Data Imputation

Fields	Issue	Imputation Logic
ZIP	Missing value	If the before and after zips of that missing ZIP are the same, then fill in the missing ZIP with that value; if not, just fill in the missing value with the previous record's zip
FULLVAL	0	The average of FULLVAL according to TAXCLASS
AVLAND	0	The average of AVLAND according to TAXCLASS
AVTOT	0	The average of AVTOT according to TAXCLASS
STORIES	Missing value	The average of STORIES according to TAXCLASS

LTFRONT	0, 1	The average of LTFRONT according to TAXCLASS
LTDEPTH	0, 1	The average of LTDEPTH according to TAXCLASS
BLDFRONT	0, 1	The average of BLDFRONT according to TAXCLASS
BLDDEPTH	0, 1	The average of BLDDEPTH according to TAXCLASS
AVLAND2	Missing value	The average of all AVLAND2 values[1]
AVTOT2	Missing value	The average of all AVTOT2 values[1]

4. Variables

We first created two new fields as the preparation for creating variables:

- FULLVAL2: FULLVAL2 = AVLAND2 + AVTOT2 (market price based on transactional price)
- zip3: The first 3 digit of ZIP

The project aims to find out the properties with abnormal data which might affect tax payments. Size and price (value) are the two most important elements when calculating property taxes, and location and tax class also matter. For the property size, we have the data of lot front and depth, building front and depth, and stories. Therefore, we can know the area of lot floor area, building floor area, and building total area according to the data we have. For the price, we have actual total and land value, transactional total and land value, and market value (actual and transactional). We also have ZIP and zip3 to group the properties in specific locations and we have the TAXCLASS field.

Therefore, according to the goal and fields we have, here are the 97 variables that we created to make further predictions:

Table 4. Descriptions of Variables

Description of variables	# Variables created
Value per Size (Square foot): The ratio of actual values of the property (FULLVAL, AVLAND, AVTOT) divided by calculated variables for property size as follows: - lotarea = LTFRONT * LTDEPTH (Lot floor area) - bldarea = BLDFRONT * BLDDEPTH (Building floor area) - bldvol = bldarea * STORIES (Building total area) The objective is to identify unusual patterns in building prices, primarily determined by the price per square foot. The variables will primarily focus on detecting outliers with significant values.	9
Inverse of Value per Size (Square foot): The variables represent the inverse of 'Value per Size'. These kinds of variables aim to capture the unusual patterns in building prices just as 'Value per Size'. By inverting the value, these variables can detect outliers with small values.	9

[1] AVLAND2 and AVTOT2 have too many missing values (about 73% are missing values), we can't get the average of these two fields according to TAXCLASS. Therefore, we use the average of all values to clean these two fields.

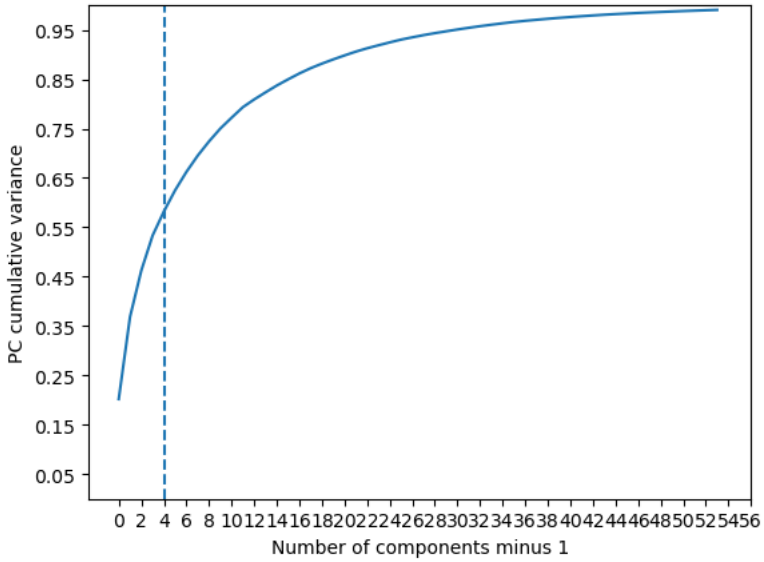
Relative Value per Size (Square foot): The relative value of Value per Size / Inverse of Value per Size over the representative average ratio of the property group (group by ZIP, zip3, or TAXCLASS). The goal is to capture the abnormal records among a specific group of properties.	54
Relative Value Ratio: The ratio between FULLVAL and AVLAND+AVTOT. The goal is to test how well FULLVAL = AVLAND+AVTOT is and to identify abnormal records resulting from the relationship between these values	1
Ratio of Actual / Transitional Price: The ratio between the actual values (FULLVAL, AVLAND, AVTOT) and the transitional values (FULLVAL2, AVLAND2, AVTOT2). These variables are designed to help detect whether there are some manipulations on the properties' market prices. The variables will identify the outliers with large values.	3
Inverse of Ratio of Actual / Transitional Price: The variables represent the inverse of 'Ratio of Actual / Transitional Price'. These variables are designed to capture outliers with small values in potential properties' market prices manipulations.	3
Relative Ratio of Actual / Transitional Price: The relative value of Ratio of Actual / Transitional Price & Inverse of Ratio of Actual / Transitional Price over the representative average ratio of the property group (group by ZIP, zip3, or TAXCLASS). The goal is to capture the abnormal records among a specific group of properties.	18
Total # of variables	**97**

5. Dimensionality Reduction

The dataset doesn't provide any field that can be considered as label, so we have to use an unsupervised learning method. Here, we used Principal Component Analysis (PCA) to conduct dimensionality reduction and make preparation for further prediction.

First, we standardize all the variables using Z-scaling. The reason of having this step is because PCA is sensitive to the scale of variables. By conducting Z-scaling, we force all the variables to have a mean of 0 and standard deviation of 1, that ensures all the variables have equal importance in PCA. Besides, Z-scaling makes all the variables to be at the same scale, so that the principal components (PC) will not be dominated by a small number of variables with high variance.

Second, we need to decide how many PCs we want to keep. PCA will find the dominant directions in the data and rotate the coordinate system along the directions, starting with the direction with the largest variance. In other word, the variance of a PC is always smaller than the previous one. The lower the variance, the less representative a PC is. Therefore, we set the n_components = 0.99 (which means we keep enough PCs to explain 99% of the variance in the original data), and visualize the cumulative variance. We don't want to have too many PCs because it will increase model complexity and cause overfitting. Therefore, according to the cumulative variance plot, we decided to use 5 PCs.



After the number of PCs is decided, we redid PCA to just keep the top 5 PCs. We used Z-scaling again on those 5 PCs to make all the 5PCs equally important when making predictions. This Z-scaling step is optional, based on whether you believe your PCs have the same importance or not.

6. Anomaly Detection Algorithms

Z-scaled PCs are used to calculate the anomaly scores. We have designed two separate anomaly detection algorithms to help detect abnormal records, they are Z-scores outlier algorithm and autoencoder.

(1). Z-score outlier

By using Z-scaled PCs to calculate the anomaly scores, we make all the PCs centered and similarly scaled, which makes them equally important for the Minkowski distance. Therefore, we can discover the outliers from calculating the distance between a specific value of a variable to the origin.

Here is the Z-scores outlier algorithm illustrates how to calculate the score for record i (s_i) using its values of different variables (z_n^i):

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

Where i: the i_th record;

n: the n_th PCs;

p_1: the power for calculating distance (reasonable range is from 1~3)

(2). Autoencoder

An autoencoder is a model designed to output the original vector input after training. Usually, the outputs from an autoencoder model should be close to the original inputs unless the records are outliers. Therefore, by detecting the output of which record is apparently different from the input (i.e. finding the large errors), we can figure out the anomalies.

In this project, we choose to use a neural network as the autoencoder model. We train it with all the records and use it to predict the outputs, with one hidden layer and three nodes using logistic activation function. We calculate the distance between input and output using Minkowski distance, which represented by the following algorithm:

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

Where $z_n'^i$: the predicted value (output) of the n_th PCs of the i_th record
z_n^i: the value (input) of the n_th PCs of the i_th record
$\sum_n |z_n'^i - z_n^i|$: the error of the autoencoder for the i_th record
p_2: the power for calculating distance (reasonable range is from 1~3)

(3) Combine scores

The final anomaly score will be the combination of the values given by the Z-score outlier and autoencoder. To ensure that we can compare the two scores in the same scale, we use the rank order of each score to calculate the final anomaly score. Because our goal is to sort out the abnormal records to be investigated, the exact values of the anomaly score do not matter.

Here is how we combine the two scores into the final score:

$$Score_{final} = (w_1 \times Z\_score\ outlier\ ranking) + (w_2 \times autoencoder\ ranking)$$

Where w_1 and w_2 are the weights of each score ($w_1 + w_2 = 1$).

(4) Decision on parameters and observations

The parameters we choose to calculate the final anomaly score are shown below:

$p_1 = 2$, $p_2 = 2$, #PCs = 5, Z-scaled PCs,
$w_1 = 0.5, w_2 = 0.5$,
hidden_layer_sizes = (3) (1 hidden layer with 3 neurons in neural network)

We also explore the model sensitivity by adjusting different parameters and compare the final lists according to sorted anomaly scores. Here is the experimental results:

Table 5. Experimental results on changing parameters

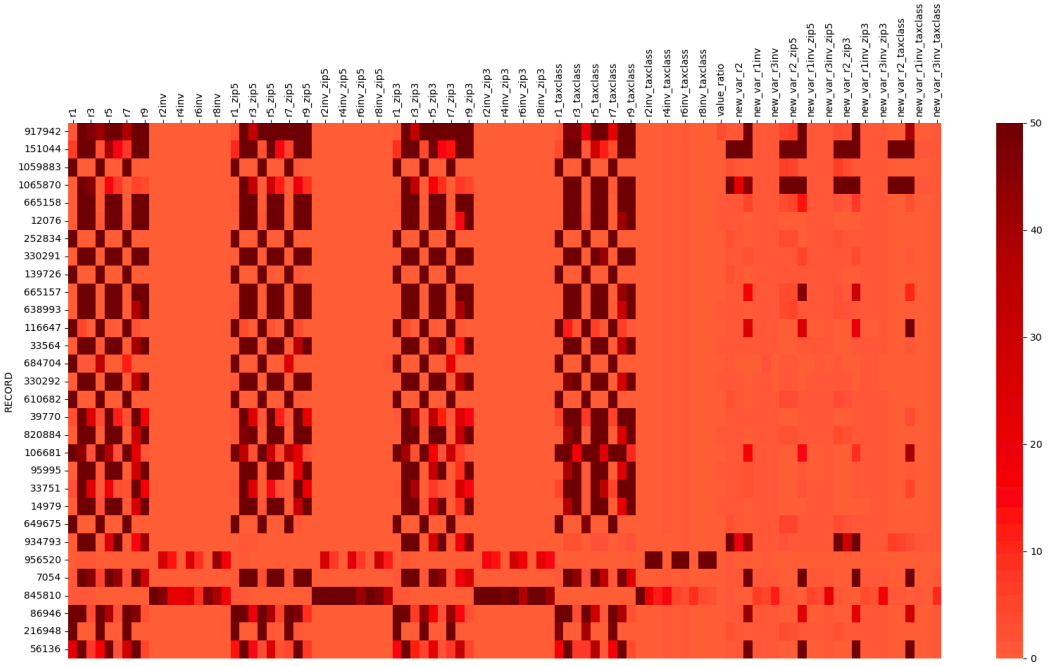
Changes in choices from baseline	Top 100 (.01%)	Top 1,000 (.1%)	Top 10,000 (1%)
Baseline: p1 = 2, p2 = 2, 5PCs, Z-scored PCs, score1_score2 weight = (0.5,0.5), NN_hidden_layer_sizes = (3)	-	-	-
p1 = 3	100.00	98.30	97.18
p1 = 4	99.00	97.30	95.85
p2 = 3	100.00	98.40	96.03
p2 = 4	99.00	98.00	94.78
NN_hidden_layer_sizes = (5)	95.00	77.10	60.82
NN_hidden_layer_sizes = (1)	93.00	74.30	67.35
NN_hidden_layer_sizes = (4)	96.00	88.50	83.12
score1_score2 weight = (0.3,0.7)	99.00	93.80	88.75

score1_score2 weight = (0.1,0.9)	97.00	87.70	77.77
score1_score2 weight = (0.7,0.3)	98.00	95.30	88.78
score1_score2 weight = (0.9,0.1)	94.00	84.30	77.04
p1 = 1, p2 = 1	98.00	97.20	91.50
4 PCS	91.00	82.70	68.17
don't Z-scored PCs	92.00	83.90	79.24
average over variations	96.50	89.77	83.31

The followings are our observations:
(1) The choices of powers in the Minkowski distance measures don't make much difference.
(2) The sizes of the Neural Network layers are among the most sensitive variations to algorithm choices.
(3) The weights of score1 and score2 have moderate sensitivity to algorithm choices, the sensitivity is mostly based on the gap between the weights of the two scores, and whether score1 weighs more than score2 doesn't make much difference.
(4) The number of PCs is among the most sensitive variations to algorithm choices.
(5) Z-score of PCs also has an apparent influence on the sensitivity of algorithm choices.
(6) The higher the score of a record, the less sensitive it is to the algorithm's choices.

7. Results

The combined score can be used to sort records from the most abnormal one to the least. The sorted records can be used to select cases which should be investigated further in order to examine potential fraud. We sort the final anomaly score from high to low, and merge it back into the original dataset. We choose the top 1,000 records and filter the values of every variable in those 1,000 records for further exploration. A heatmap is used to help visualize what variables are driving the high scores. We then explore the top records manually to find out the exact reasons why those records are abnormal.

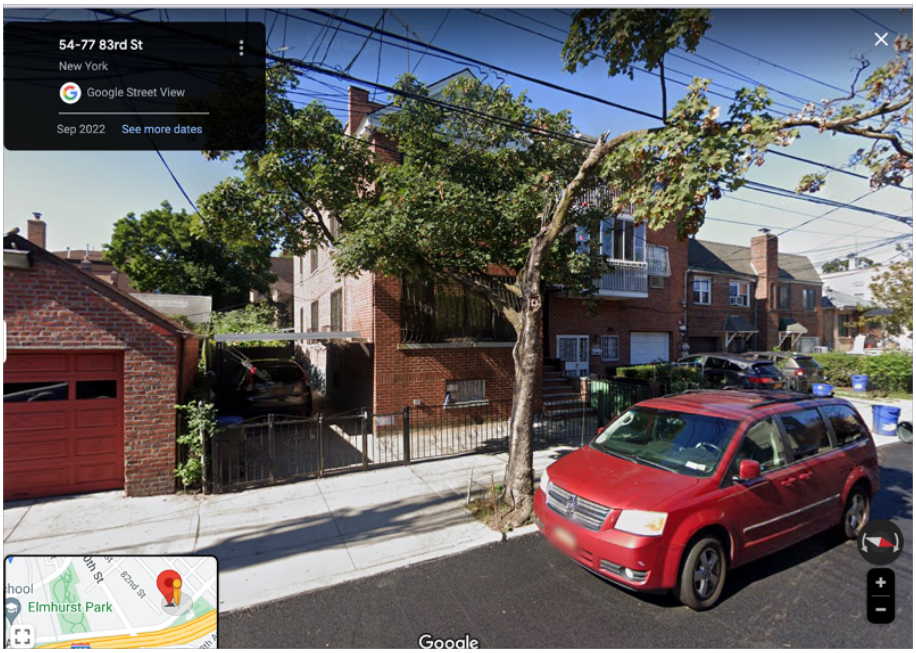


7.1 Five Case Studies

Here are the observations of 5 case studies:

Sample 1:

Record No. 658933



OWNER	WAN CHIU CHEUNG	LTFRONT	25
ADDRESS	54-76 83 STREET, 11373	LTDEPTH	100
FULLVAL	776,000	BLDFRONT	2,500
AVLAND	26,940	BLDDEPTH	5,600
AVTOT	46,560	STORIES	3

This property is suspicious mainly because its BLDFRONT and BLDDEPTH are too big. According to the view from Google Maps, this building only has 3 stories, but the building front and building depth are about 100 and 56 times the lot front and lot depth respectively, which seems impossible. From the fraud algorithms, I found the following variables with unusual values:

r8inv	38.2131;	r2inv_zip3	31.5000;
r8inv_zip3	48.1896;	r2inv_taxclass	562.7630;
r3inv_taxclass	619.5108;	r5inv_taxclass	433.6783;
r6inv_taxclass	346.0420;	r8inv_taxclass	436.1320;
r9inv_taxclass	409.7530		

According to the definitions of those variables, we can find out that the unusually large building size is the main reason making this property suspicious. Those values of r{2,3,5,6,8,9}inv_taxclass are extremely high, which means the price per unit according to the building size and the full building area is far lower than the normal price of that tax class. Therefore, it is reasonable for us to list this property as an abnormal or fraudulent record based on the investigation.

Sample 2:
Record No. 33751

OWNER	GUIDARA, FRANK	LTFRONT	122
ADDRESS	520 WEST 23 STREET, 10011	LTDEPTH	98
FULLVAL	14,400,000	BLDFRONT	8
AVLAND	540,000	BLDDEPTH	10
AVTOT	6,480,000	STORIES	15

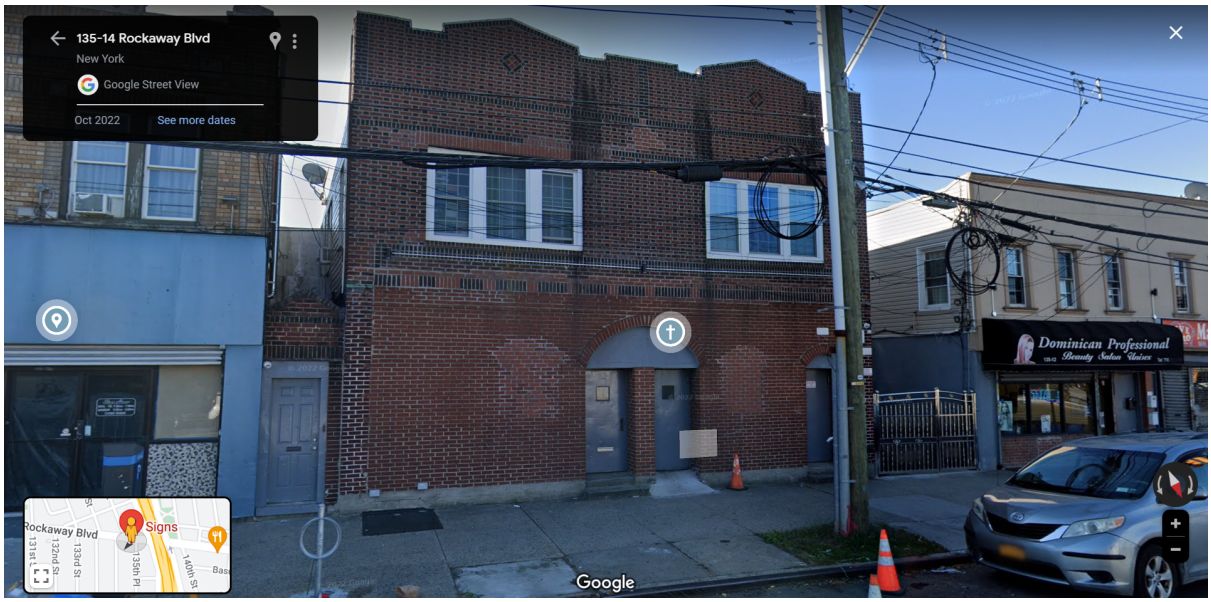
This property is suspicious mainly because its BLDFRONT and BLDDEPTH are too small. According to the view from Google Maps, the building front should be roughly the same size as the lot front, but the lot front in the data is 15 times bigger. From the fraud algorithms, I found the following variables with unusual values:

r2	176.68;	r2_zip5	115.67;
r3	24.16;	r3_zip5	22.89;
r5	18.01;	r5_zip5	15.93;
r8	87.03;	r8_zip5	75.72;

r9	16.64;	r9_zip5	20.82;
r2_zip3	160.92;	r3_taxclass	194.07;
r3_zip3	39.70;	r5_taxclass	78.57;
r8_zip3	27.05;	r6_taxclass	34.43;
r9_zip3	14.64;	r8_taxclass	247.96;
r2_taxclass	409.64;	r9_taxclass	117.82

According to the definitions of those variables, we can find out that the unusually small building size is the main reason making this property suspicious. Those values of r{2,3,5,6,8,9} and their derivatives are extremely high, which means the price per unit according to the building size and the full building area is far higher than the normal price. Therefore, it is reasonable for us to list this property as an abnormal or fraudulent record based on the investigation.

Sample 3:
Record No. 871048



OWNER	VALERIE SHAKESPEARE	LTFRONT	0
ADDRESS	135-16 ROCKAWAY BOULEVARD, 11420	LTDEPTH	99
FULLVAL	210	BLDFRONT	0
AVLAND	95	BLDDEPTH	0
AVTOT	95	STORIES	-

This property is suspicious mainly because its values and its sizes are too small. According to the view from Google Maps, the building should be combined together with 135-14 ROCKAWAY BOULEVARD. Yet, both of its values and its sizes have been recorded on the file (with abnormally low values). From the fraud algorithms, these are some examples of unusual values:

r1inv_zip5	175.05;	r2inv_zip3	154.91;
r2inv_zip5	143.41;	r3inv_zip3	147.93;
r3inv_zip5	142.72;	r1inv	32.00;
r1inv_zip3	139.97;	r2inv	26.48

According to the definitions of those variables, we can find out that the unusually small values are the main reason making this property suspicious. Those values of r{1,2,3}inv and their derivatives are extremely high, which means the price per unit according to the building size and the full building area is far lower than the normal price. Therefore, it is reasonable for us to list this property as an abnormal or fraudulent record based on the investigation.

Sample 4:
Record No. 519314/519315

OWNER	CHEN, XU XIAO	LTFRONT	0
ADDRESS	2378 EAST 14 STREET, 11229	LTDEPTH	0
FULLVAL	116	BLDFRONT	0
AVLAND	52	BLDDEPTH	0
AVTOT	52	STORIES	3

This property is suspicious mainly because its values are too small. Although the building sizes are missing, the FULLVAL of $116 is still suspicious. From the fraud algorithms, these are some examples of unusual values:

r1inv	53.71;	r2inv_zip5	98.26;
r2inv	41.21;	r7inv_zip5	60.10;
r7inv	36.46;	r1inv_zip3	68.28;
r1inv_zip5	105.09;	r2inv_zip3	66.87

According to the definitions of those variables, we can find out that the unusually small values are the main reason making this property suspicious. Those values of r{1,2,7}inv and their derivatives are extremely high, which means the price per unit according to the building size and the full building area is far lower than the normal price. Therefore, it is reasonable for us to list this property as an abnormal or fraudulent record based on the investigation.

Sample 5:
Record No. 333412

OWNER	SPOONER ALSTON	LTFRONT	17
ADDRESS	37 MONROE STREET, 11238	LTDEPTH	85
FULLVAL	9,060	BLDFRONT	4017
AVLAND	3,874	BLDDEPTH	42
AVTOT	4,077	STORIES	3

This property is suspicious mainly because its values are too small and the building front is bigger than the land front. Although the building sizes are small (assuming that building front data is incorrect), the FULLVAL of $9,060 is still suspicious. From the fraud algorithms, these are some examples of unusual values:

r2inv_taxclass	286.72;	new_var_r3inv_ taxclass	38.59;
r3inv_taxclass	198.63;	r2inv	20.45;
r8inv_taxclass	305.72;	r2inv_zip5	25.87;
r9inv_taxclass	255.71;	r2inv_zip3	33.20

According to the definitions of those variables, we can find out that the unusually small values are the main reason making this property suspicious. Those values of r{2,3,8,9}inv_taxclass are extremely high, which means the price per unit according to the building size and the full building area is far lower than the normal price. Therefore, it is reasonable for us to list this property as an abnormal or fraudulent record based on the investigation.

8. Summary

The project is designed for building an unsupervised machine learning model to detect abnormal records of properties in New York City. The outcome of this project can provide insights on property tax management.

The dataset used for this project has 32 fields (14 are numerical and 18 are categorical) and 1,070,994 records. After exploring data quality, we find there are 13 fields with missing, and 9 fields with frivolous values. The dataset also has governmental owners whose properties are benign but with extreme values in some fields. Therefore, we remove those benign records that may confuse the model and clean the data by taking the average according to groups before variable generation.

We believe size and price are the most important features for property tax, and price per unit area and the ratio of actual/transactional value are the most crucial variables for making predictions. Based on this logic, we generated 97 variables and put them into the PCA model after scaling. We keep the top 5 PCs, Z-scale them, and use those Z-scaled PCs to calculate the anomaly scores using two different algorithms: Z-score outlier and autoencoder. The final anomaly score of each record is the combination of the Z-score outlier and autoencoder's rank orders. Finally, we select the top 1,000 records with the highest final anomaly scores and explore them using heatmap and case studies.

According to the exploration, we find that the abnormal records have some shared issues. The following table illustrates the issues and how we adjust the model to improve performance:

Table 6. Issues in the abnormal records and suggested adjustments

Issues in the values	Adjustment
The building front and depth are larger than the lot front and depth. We believe many of them are data entry errors.	Create variables for the ratio between building front/depth and lot front/depth to detect extreme values.
Some lot front/depth = 1 are due to the irregular shape of the land.	Explore the data more carefully and use specialized calculations when making variables for lands with irregular shapes. We might need APIs to link to maps or geographic information systems to help us detect irregularly shaped lands without finding them out manually.
Some records have extremely large differences in building and/or lot front and depth.	Create variables for the ratio between the building front and building depth, and the lot front and lot depth.

9. Appendix

Data Quality Report

1. Data Description

This data is about property data in the New York City. This dataset contains data about each property's owner, class, size, actual values, transitional values, etc. The duration of events covered in this dataset is within November 2010. There are **32 fields** (14 are numerical and 18 are categorical) and **1,070,994 records**.

2. Summary Tables

(1) Numerical Table

Field Name	% Populated	# Blanks	# Zeros	Min	Max	Mean	Stdev	Most Common
LTFRONT	100.00%	**0**	169,108	0	9,999	36.64	74.03	0
LTDEPTH	100.00%	0	170,128	0	9,999	88.86	76.40	100
STORIES	94.75%	56,264	0	1	119	5.01	8.37	2
FULLVAL	100.00%	0	13,007	0	6,150,000,000	874,264.51	11,582,430.00	0
AVLAND	100.00%	0	13,009	0	2,668,500,000	85,067.92	4,057,260.00	0
AVTOT	100.00%	0	13,007	0	4,668,309,000	227,238.17	6,877,529.00	0
EXLAND	100.00%	0	491,699	0	2,668,500,000	36,423.89	3,981,576.00	0
EXTOT	100.00%	0	432,572	0	4,668,309,000	91,186.98	6,508,403.00	0
BLDFRONT	100.00%	0	228,815	0	7,575	23.04	35.58	0
BLDDEPTH	100.00%	0	228,853	0	9,393	39.92	42.71	0
AVLAND2	26.40%	788,268	0	3	2,371,005,000	246,235.72	6,178,963.00	2,408
AVTOT2	26.40%	788,262	0	3	4,501,180,000	713,911.44	11,652,530.00	750
EXLAND2	8.17%	983,545	0	1	2,371,005,000	351,235.68	10,802,210.00	2,090
EXTOT2	12.22%	940,166	0	7	4,501,180,000	656,768.28	16,072,510.00	2,090

(2) Categorical Table

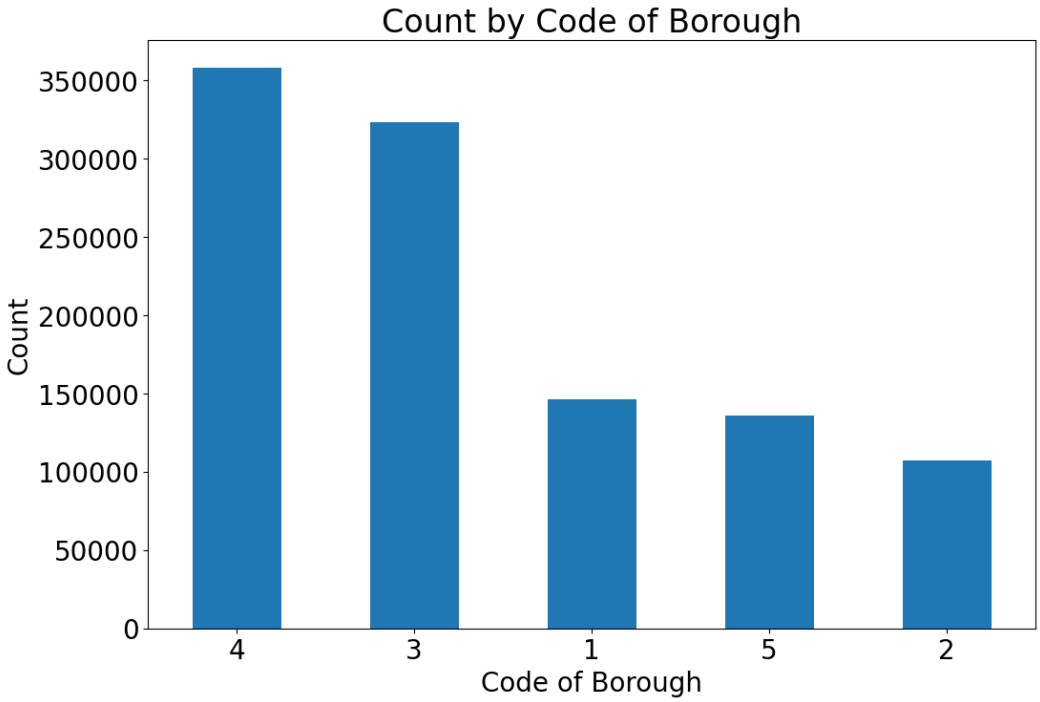
Field Name	% Populated	# Blanks	# Zeros	# Unique Values	Most Common
RECORD	100.00%	0	0	1,070,994	1
BBLE	100.00%	0	0	1,070,994	1000010101
BORO	100.00%	0	0	5	4
BLOCK	100.00%	0	0	13,984	3944
LOT	100.00%	0	0	6,366	1
EASEMENT	0.43%	1,066,358	0	12	E
OWNER	97.04%	31,745	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	100.00%	0	0	200	R4
TAXCLASS	100.00%	0	0	11	1
EXT	33.08%	716,689	0	3	G
EXCD1	59.62%	432,506	0	129	1017
STADDR	99.94%	676	0	839,280	501 SURF AVENUE
ZIP	97.21%	29,890	0	196	10314
EXMPTCL	1.45%	1,055,415	0	14	X1
EXCD2	8.68%	978,046	0	60	1017
PERIOD	100.00%	0	0	1	FINAL
YEAR	100.00%	0	0	1	2010/11
VALTYPE	100.00%	0	0	1	AC-TR

3. Visualization of Each Field

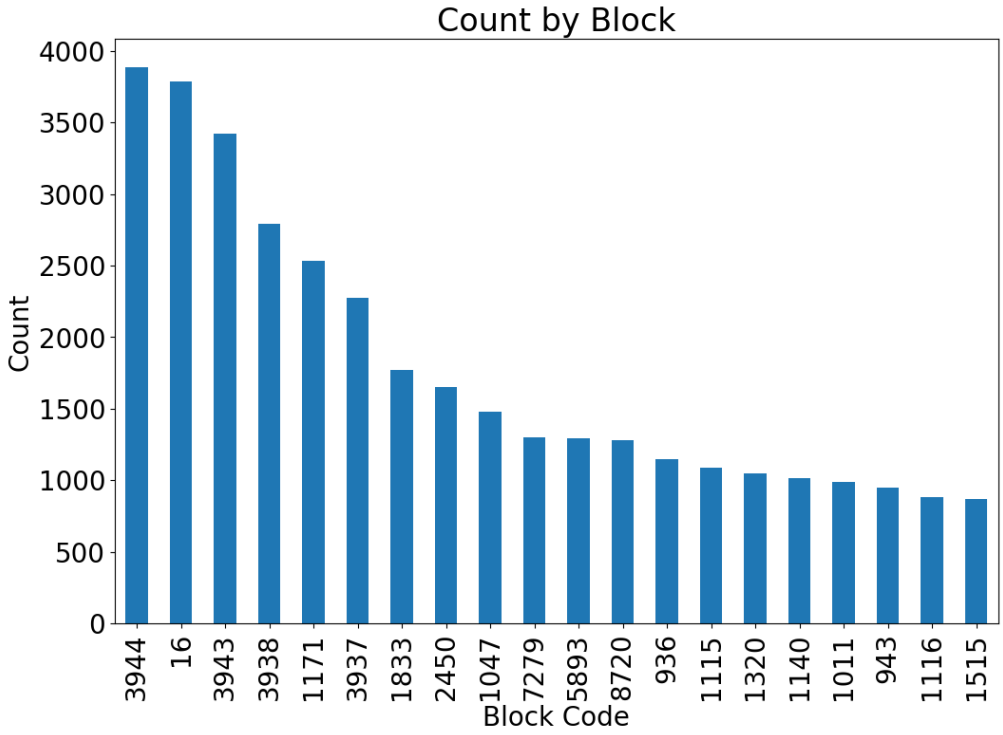
(1) Field Name: RECORD
Description: Number of Records. Ordinal unique positive integer for each record, from 1 to 1,070,994.

(2) Field Name: BBLE
Description: The unique file key of each property. Each BBLE is a combination of BORO, BLOCK, LOT and EASEMENT code.

(3) Field Name: BORO
Description: The code of the borough where the property is located in.
In this dataset, 1 = Manhattan; 2 = Bronx; 3 = Brooklyn; 4 = Queens; 5 = Staten Island.
This distribution shows the number of records in each borough. The majority of all the records are in borough 4 (Queens) and 3 (Brooklyn). Borough #4 (Queens) has the most property records in this dataset, the count of which is over 350,000.

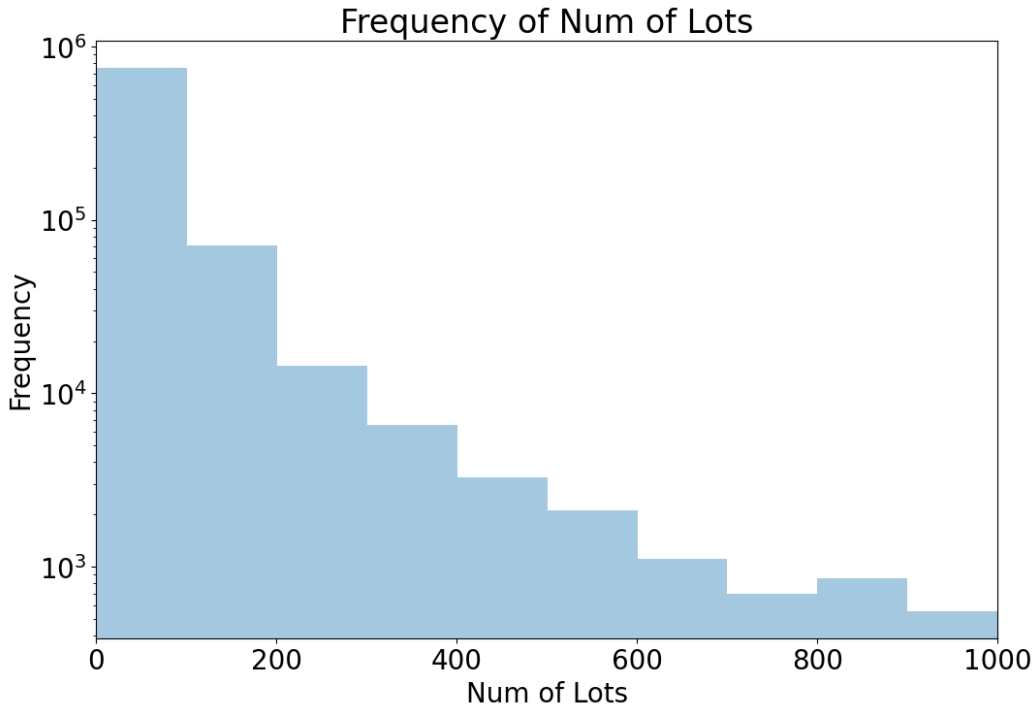


(4) Field Name: BLOCK
Description: Valid block code ranges by BORO. In this dataset, the block code follows this specific term: Manhattan: 1 to 2255; Bronx: 2260 to 5958; Brooklyn: 1 to 8955; Queens: 1 to 16350; Staten Island: 1 to 8050.
This distribution shows the top 20 values of this field. '3944' is the block code with most records, the count of which is about 3,900.



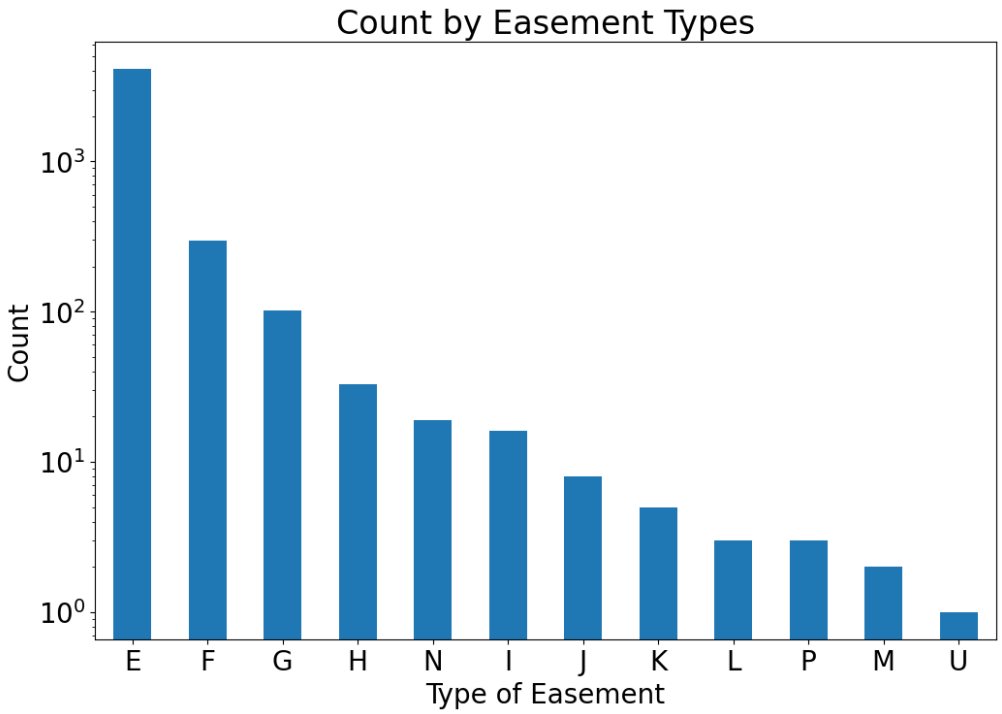
(5) Field Name: LOT

Description: The specific identity numbers of lots. This distribution shows the frequency of lot numbers, the most frequent range is from 0-100, the count of which is about 800,000.



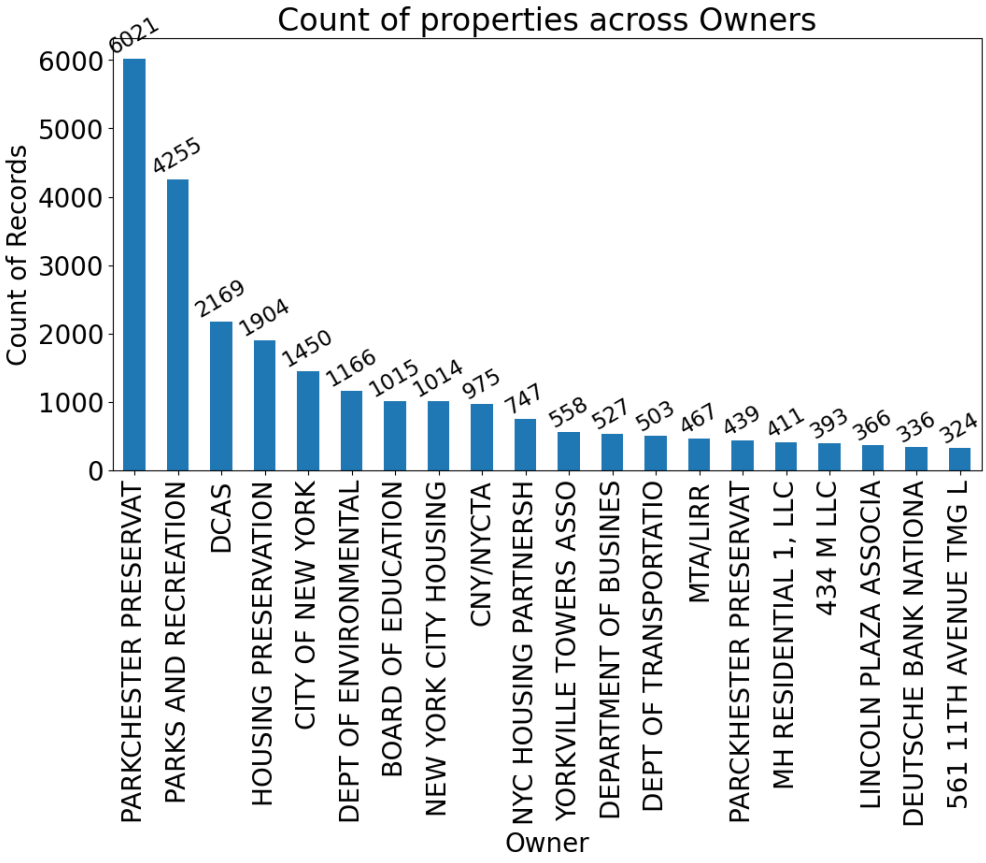
(6) Field Name: EASEMENT

Description: Types of easement of each property. The value of this field follows this specific term: Space = No Easement; A = Air Easement; B = Non-Air Rights; E~M = Land Easement; N = Non-Transit Easement; P = Pier; R = Railroad; S = Street; U = U.S. Government. The distribution shows the top 20 values of this field. 'E' (Land Easement) is the easement type with most records, the count of which is about 4,000.



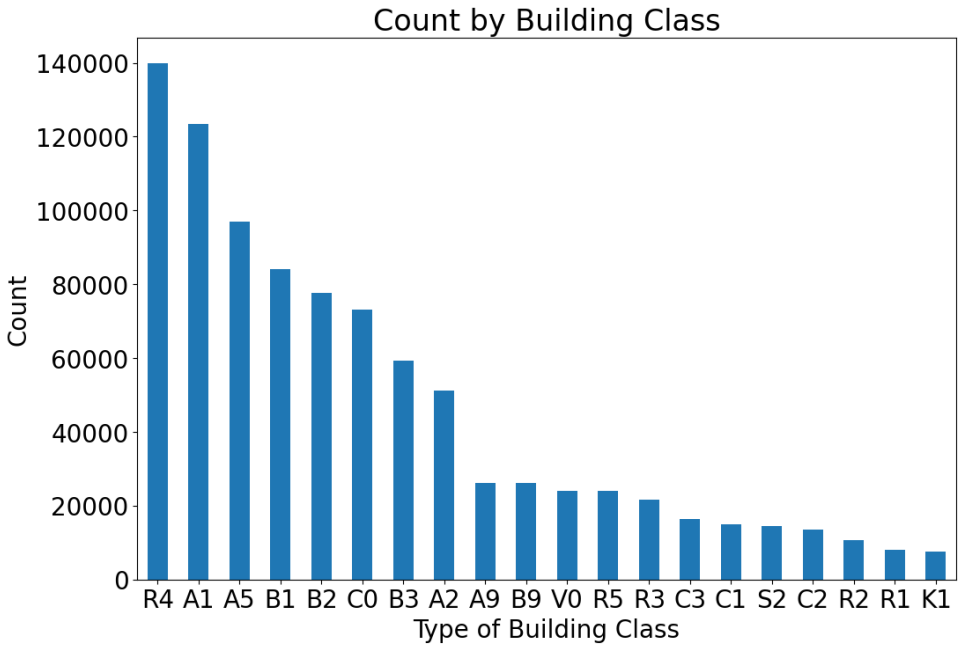
(7) Field Name: OWNER

Description: The owner's name of a specific property. The distribution shows the top 20 values of this field. 'PARKCHESTER PRESERVAT' is the owner with the largest number of properties in this dataset, whose number of property records is about 6,000.



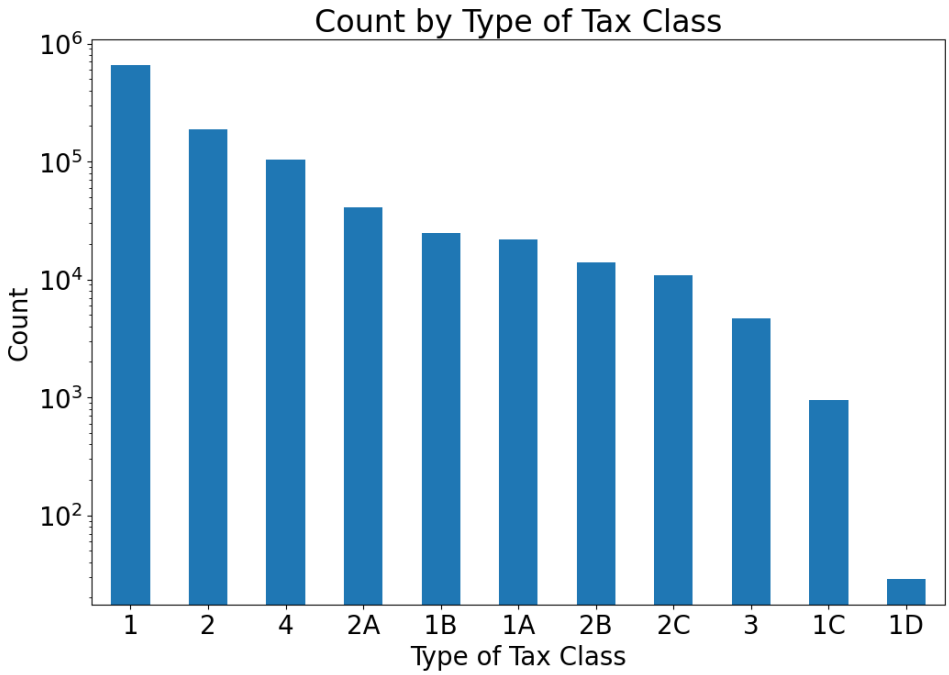
(8) Field Name: BLDGCL

Description: The building class of each property. The distribution shows the top 20 values of this field. 'R4' is the building class with most records in this dataset, the count of which is about 140,000.



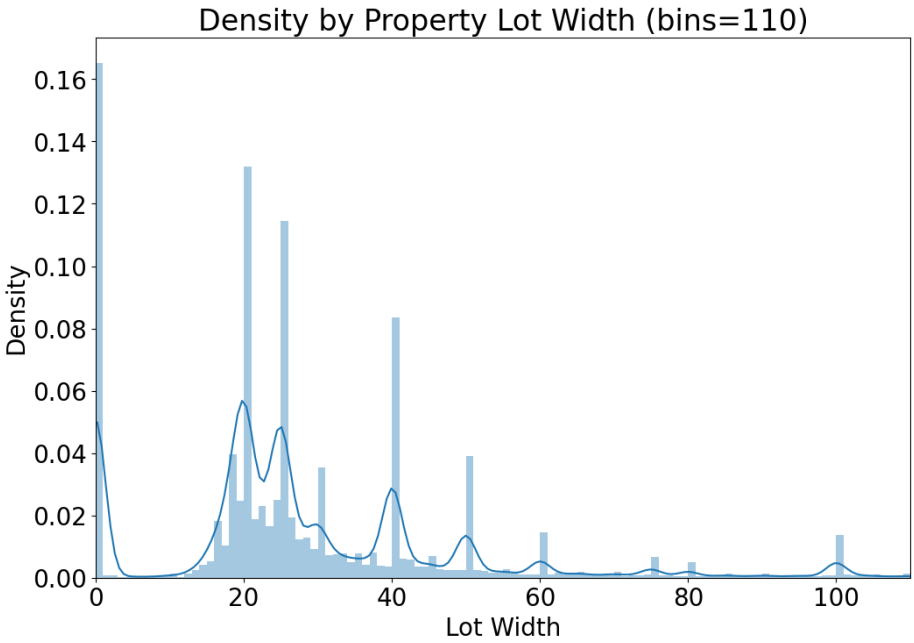
(9) Field Name: TAXCLASS

Description: The tax class of each property. The value of this field follows this specific term: 1 = 1~3 Unit Residence; 2 = Apartments; 2A = 4,5, or 6 units; 4 = All others. The distribution shows the top 20 values of this field. '1' (1~3 Unit Residence) is the tax class with most records in this dataset, the count of which is about 700,000.



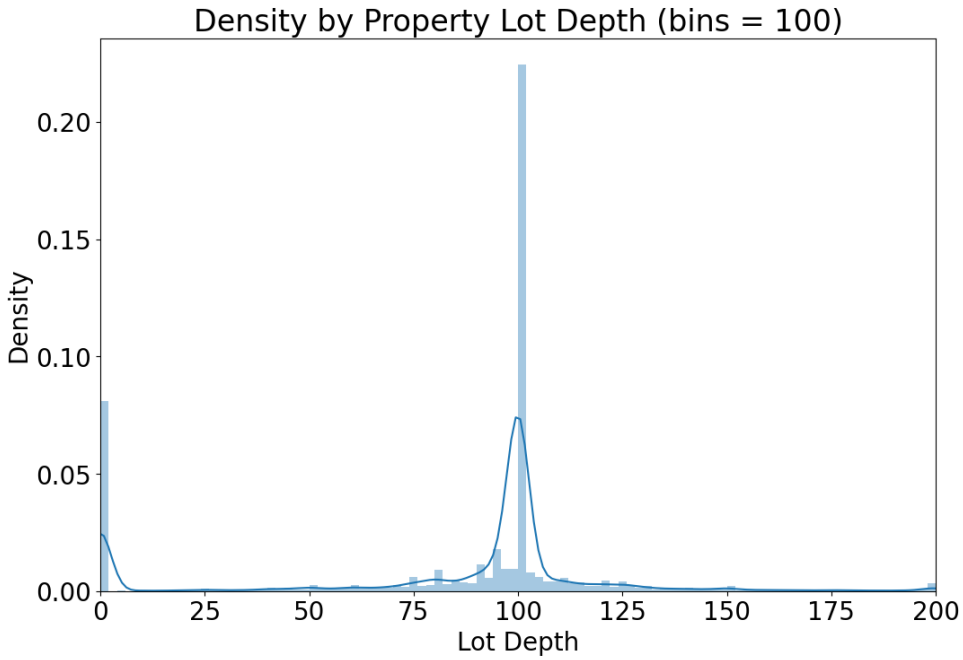
(10) Field Name: LTFRONT

Description: The lot width of the specific property. This distribution shows the frequency density according to the properties' lot width in this dataset. The range of 'Lot Width' is set to be 0~110 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that 0 lot width has the highest density, which is more than 0.16 (most probably caused by unclean data (value = 0)), and most of the properties have lot widths within the range of 15~60.



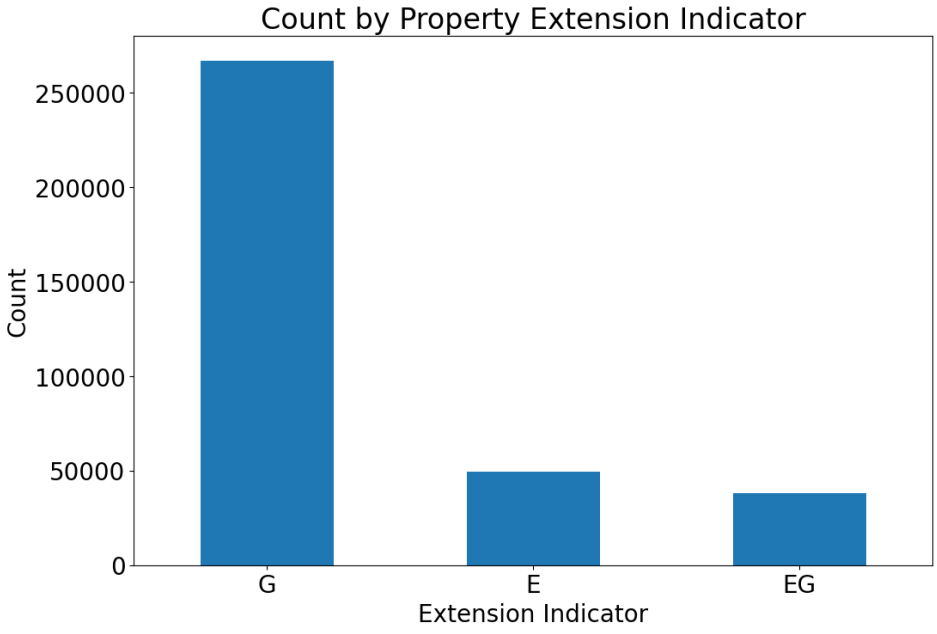
(11) Field Name: LTDEPTH

Description: The lot depth of the specific property. This distribution shows the frequency density according to the properties' lot depth in this dataset. The range of 'Lot Depth' is set to be 0~200 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that most of the properties have lot widths within the range of 75~125. The first bin (0~1) also has a high density, which may cause by unclean data (value = 0).



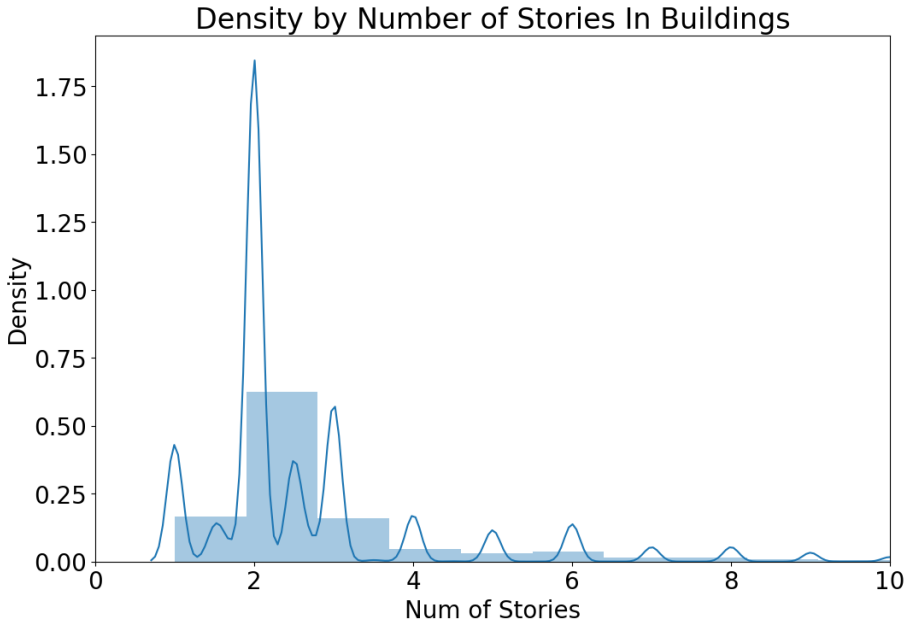
(12) Field Name: EXT

Description: The extension indicator of each property. The field only has 3 values: 'G', 'E', and 'EG'. The distribution shows the total number of each extension indicator in the dataset, the majority of which is 'G', the count of which is more than 260,000.



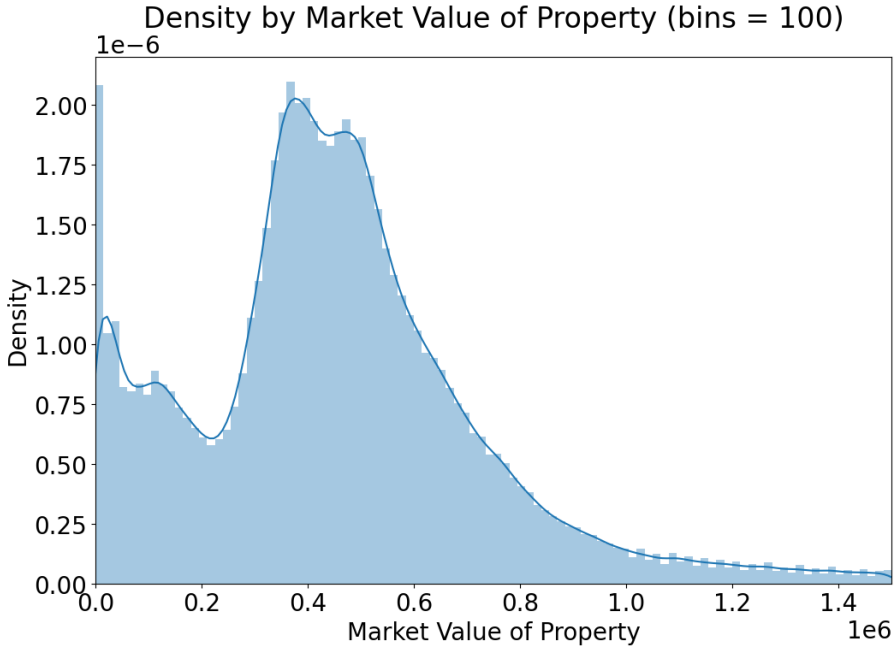
(13) Field Name: STORIES

Description: Number of stories in the buildings. The distribution shows the frequency density according to number of stories. The range of 'Number of Stories' is set to be 0~10 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that most of the properties have 1~3 stories in their buildings.



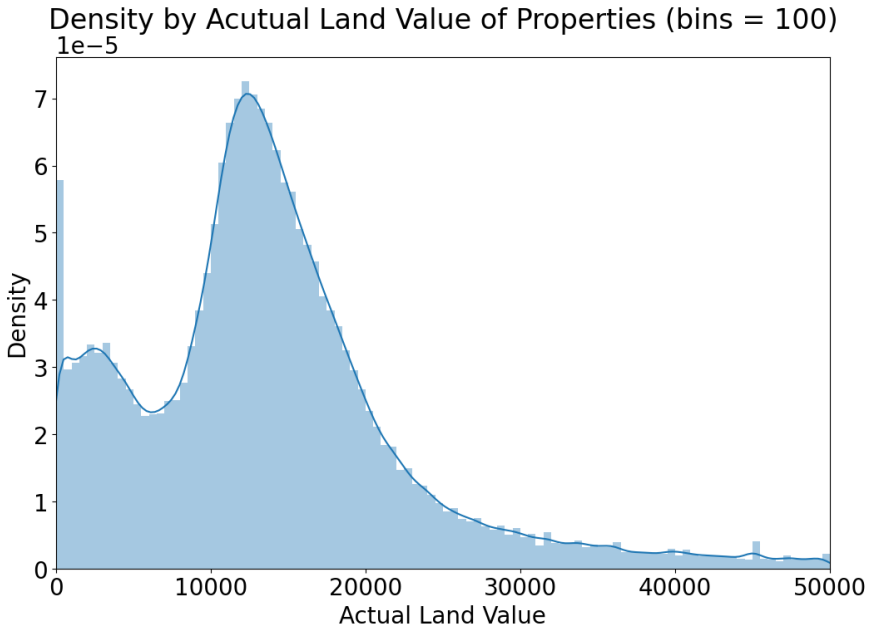
(14) Field Name: FULLVAL

Description: The market value of each property. The distribution shows the frequency density according to market value. The range of 'Market Value of Property' is set to be 0 ~ 1,500,000 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that majority of the properties have the market value below 1,000,000. The first bin (0 ~ 10,000) has a pretty high density and that may cause by unclean data (value = 0).



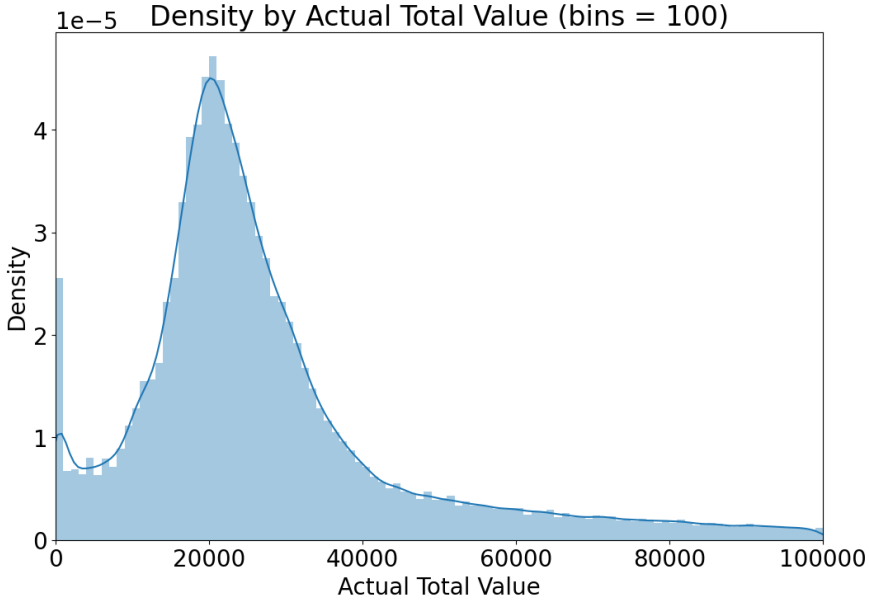
(15) Field Name: AVLAND

Description: Actual land value of a specific property. The distribution shows the frequency density according to actual land value. The range of 'Actual Land Value' is set to be 0 ~ 500,000 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that majority of the properties share the actual land value below 250,000. The first bin (0 ~ 5,000) has a pretty high density and that may cause by unclean data (value = 0).



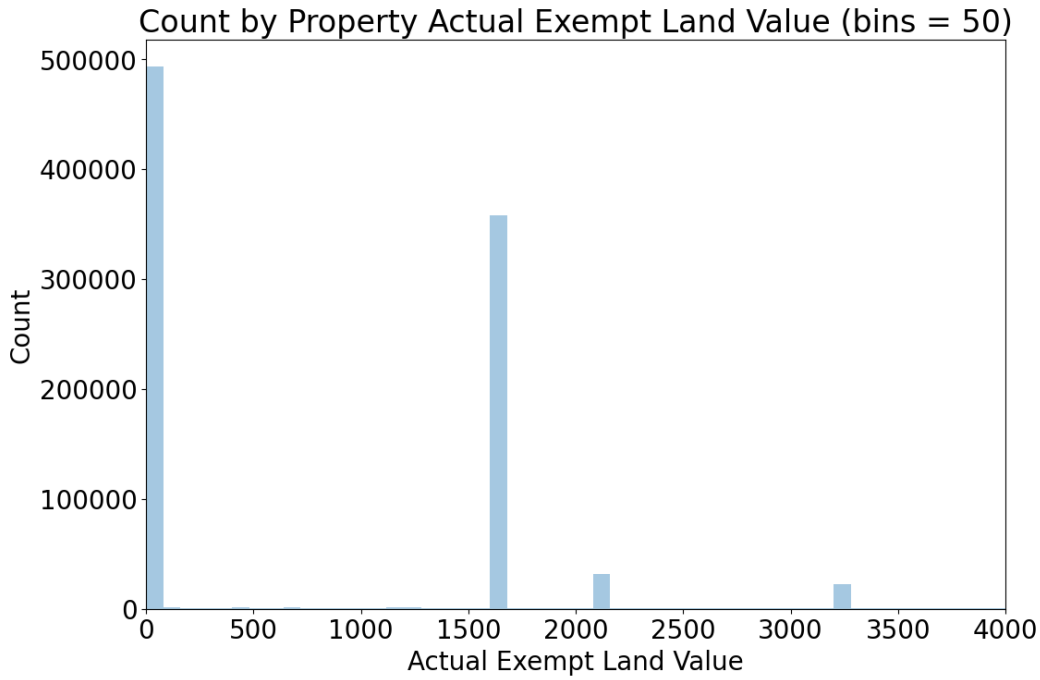
(16) Field Name: AVTOT

Description: The actual total value of a specific property. The distribution shows the frequency density according to actual total value. The range of 'Actual Land Value' is set to be 0 ~ 100,000 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that majority of the properties share the actual total value between 10,000~40,000. The first bin (0 ~ 1,000) has a pretty high density and that may cause by unclean data (value = 0).



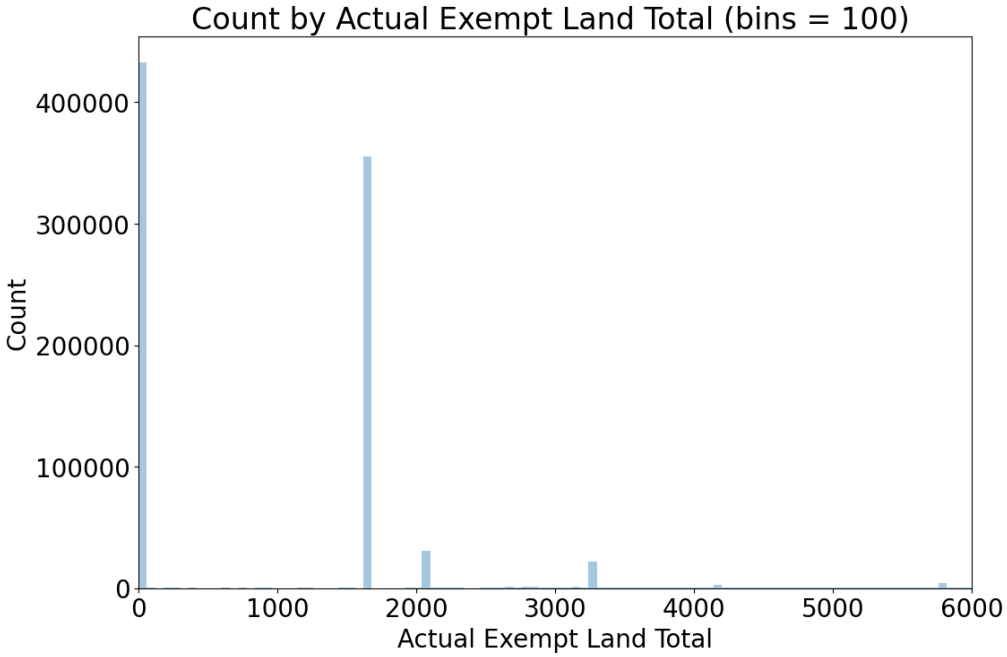
(17) Field Name: EXLAND

Description: Actual exempt land value of a specific property. The distribution shows the total number of records according to the actual exempt land value of each property. The range of 'Actual Exempt Land Value' is set to be 0 ~ 4,000 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. From the distribution, the first bin (0 ~ 80) has the highest number of records, but 491,699 of the records are valued at 0, which has a high probability to be missing values. Besides, most properties share the actual exempt land value of around 1600, the count of which is about 350,000.



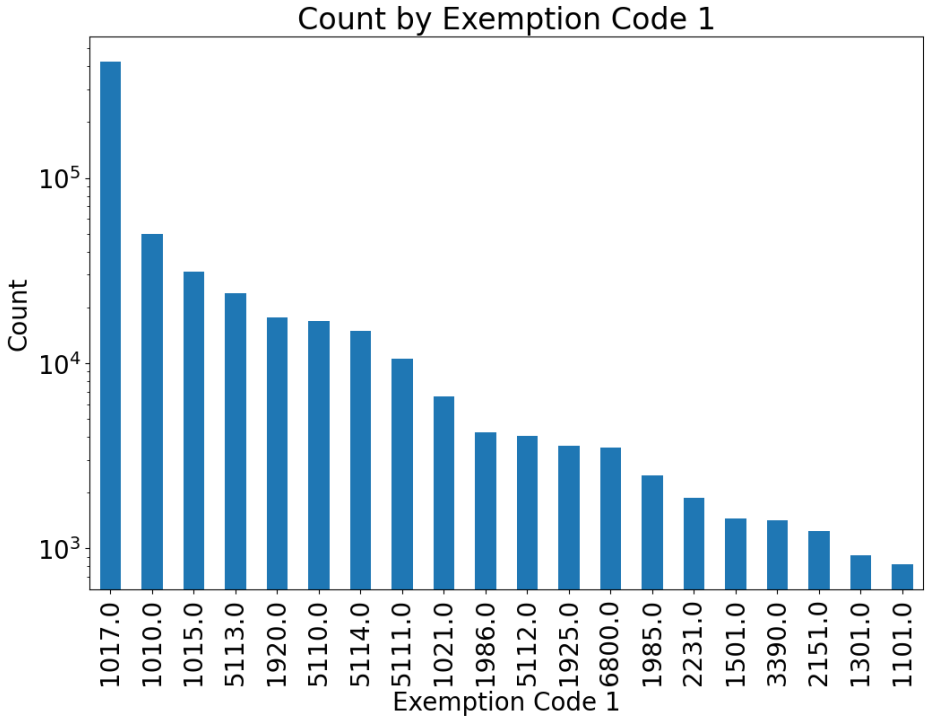
(18) Field Name: EXTOT

Description: The Actual exempt land total of a specific property. The distribution shows the total number of records according to the actual exempt land total of each property. The range of 'Actual Exempt Land Total' is set to be 0 ~ 6,000 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. From the distribution, the first bin (0 ~ 60) has the highest number of records, but 432,572 of the records are valued at 0, which has a high probability to be missing values. Besides, most properties share the actual exempt land value of around 1600, the count of which is about 350,000.



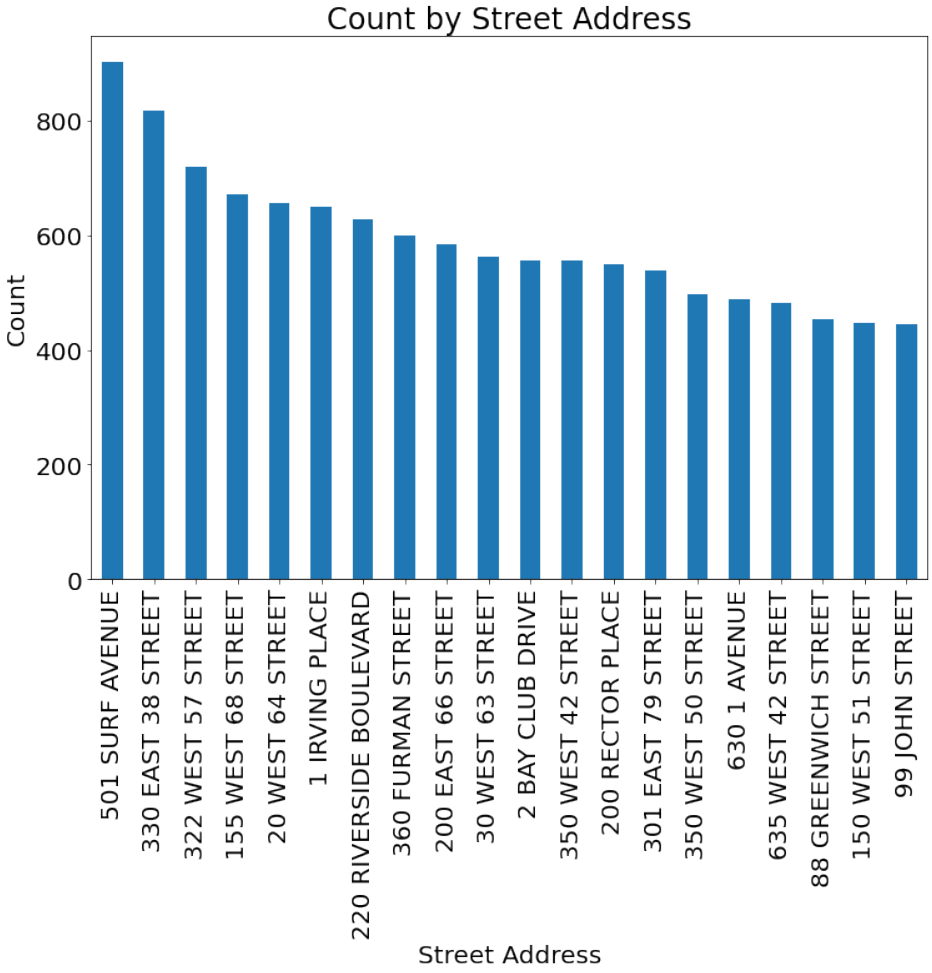
(19) Field Name: EXCD1

Description: The exemption code 1 of a specific property. The distribution shows the top 20 values of this field. '1017' is the exemption code 1 with most records in this dataset, the count of which is about 400,000.



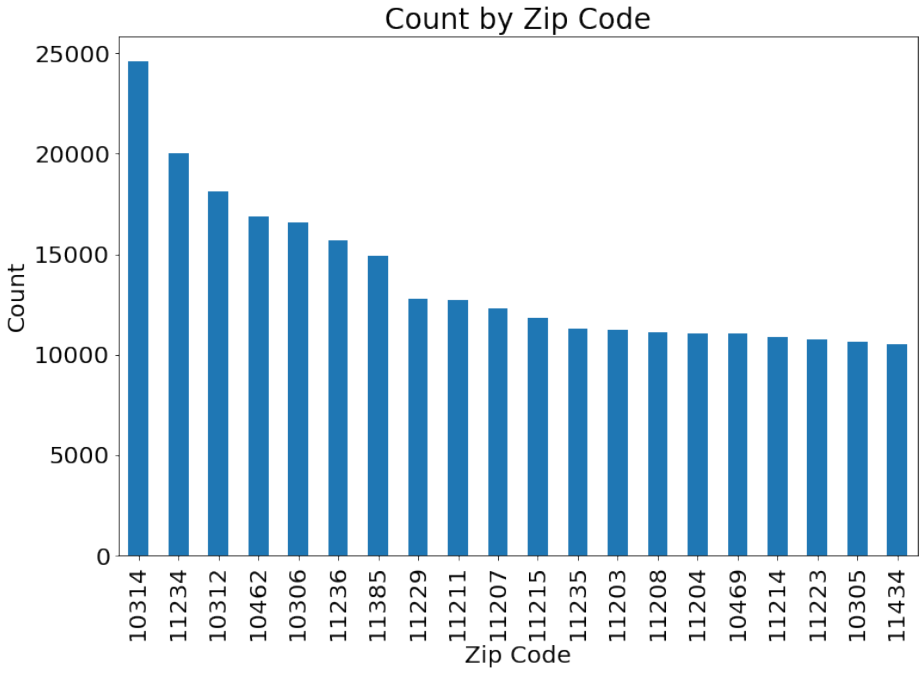
(20) Field Name: STADDR

Description: The street address of each property. The distribution shows the top 20 values of this field. The most common value is 501 SURF AVENUE, the count of which is about 900.



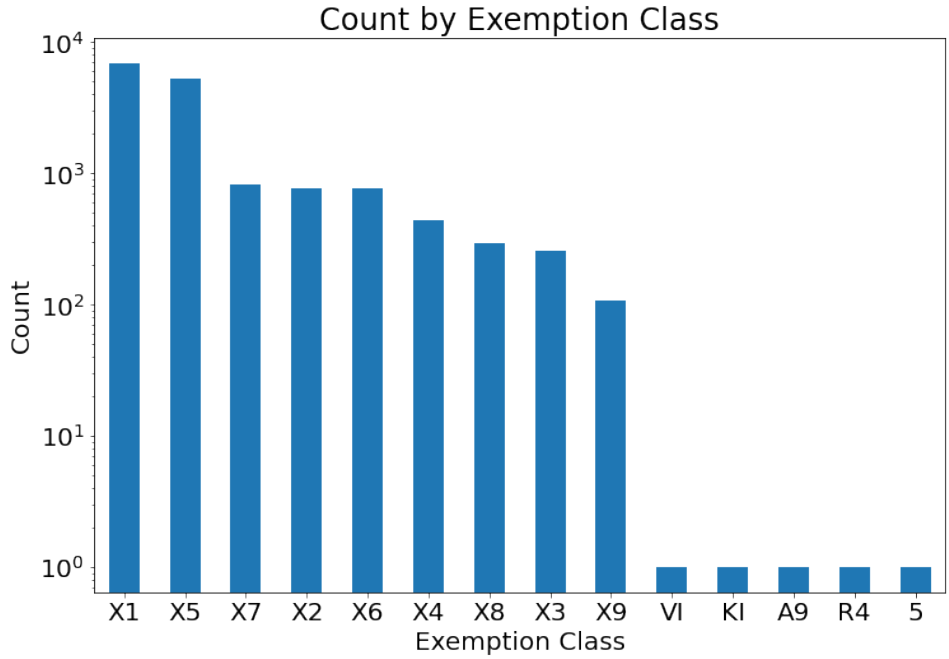
(21) Field Name: ZIP

Description: The zip code of each property. The distribution shows the top 20 values of this field. The most common value is 10314, the count of which is about 900.



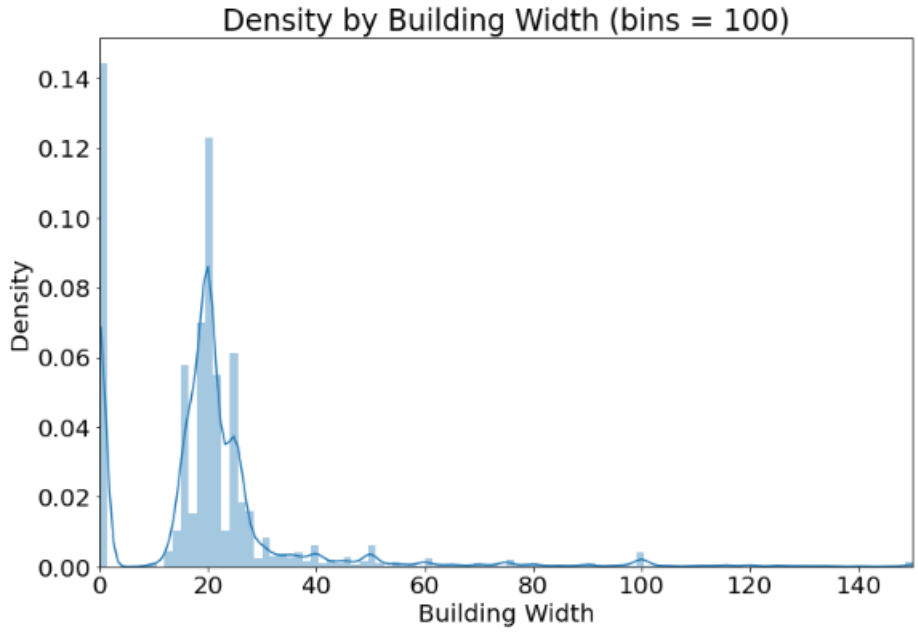
(22) Field Name: EXMPTCL

Description: The exemption class of each property. The distribution shows the top 20 values of this field. The most common value is 'X1', the count of which is about 7,000.



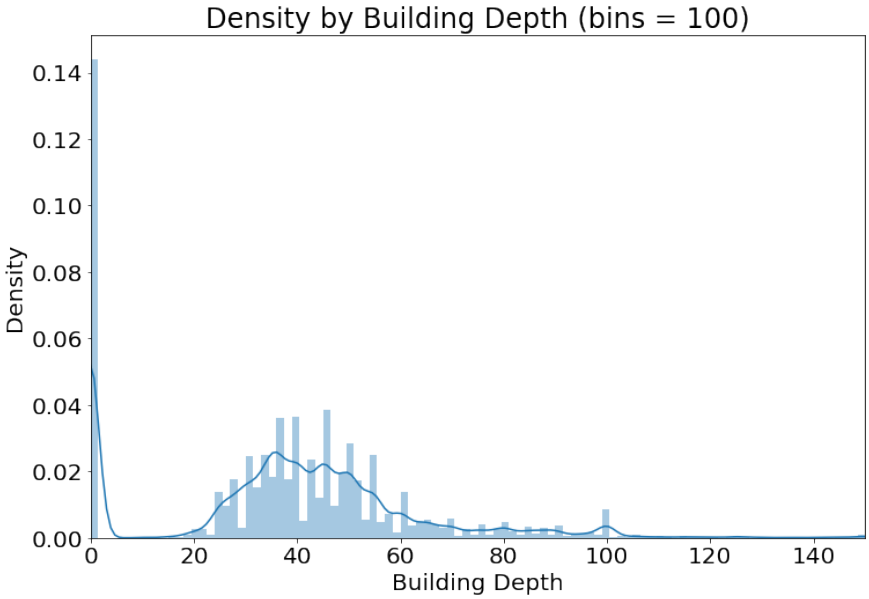
(23) Field Name: BLDFRONT

Description: The building width of a specific property. This distribution shows the frequency density according to the properties' building width in this dataset. The range of 'Building Width' is set to be 0~150 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. From the distribution, the first bin (0 ~ 1.5) has the highest number of records, but 228,815 of the records are valued at 0, which has a high probability to be missing values. Besides, the distribution indicates that most of the properties have building widths within the range of 10~50.



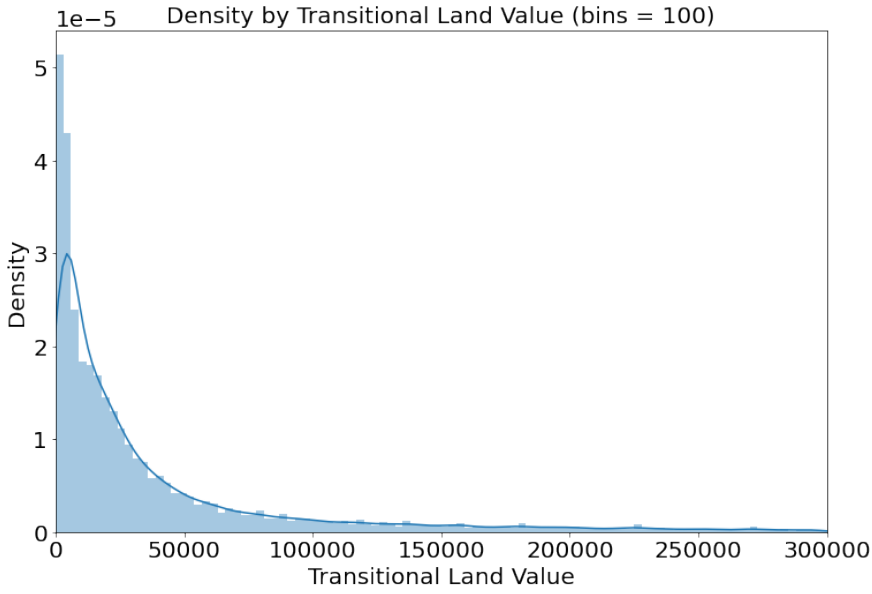
(24) Field Name: BLDDEPTH

Description: The building depth of a specific property. This distribution shows the frequency density according to the properties' building depth in this dataset. The range of 'Building Depth' is set to be 0~150 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. From the distribution, the first bin (0 ~ 1.5) has the highest number of records, but 228,853 of the records are valued at 0, which has a high probability to be missing values. Besides, the distribution indicates that most of the properties have building depths within the range of 20~100.



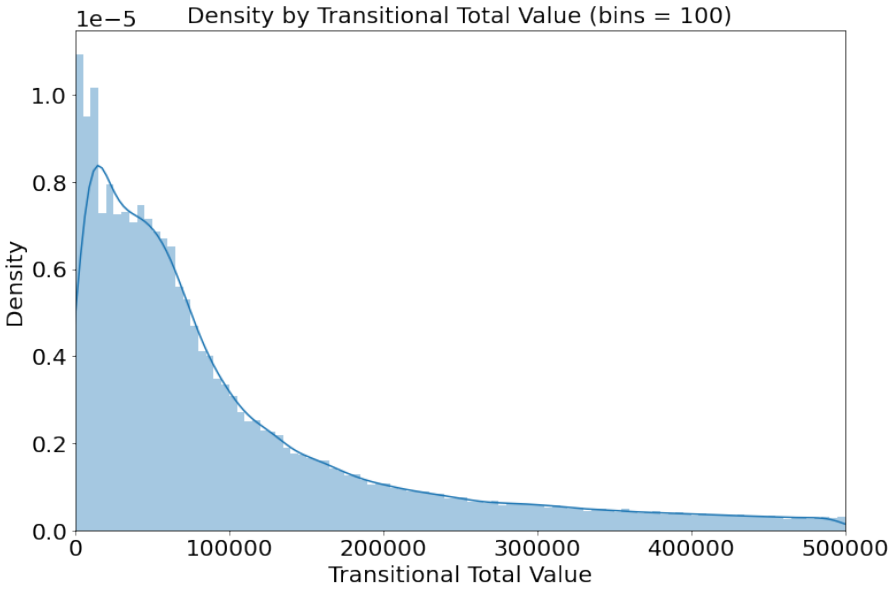
(25) Field Name: AVLAND2

Description: The transitional land value of a specific property. This distribution shows the frequency density according to the properties' transitional land value in this dataset. The range of 'Transitional Land Value' is set to be 0~300,000 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that most of the properties have transitional land value within the range of 0~50,000.



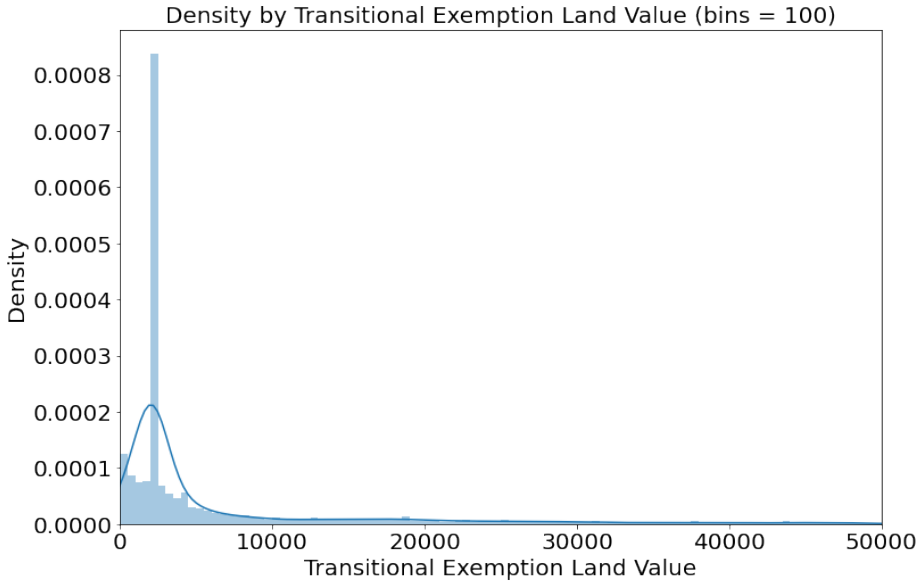
(26) Field Name: AVTOT2

Description: The transitional total value of a specific property. This distribution shows the frequency density according to the properties' transitional total value in this dataset. The range of 'Transitional Total Value' is set to be 0~500,000 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that most of the properties have transitional total value within the range of 0~200,000.



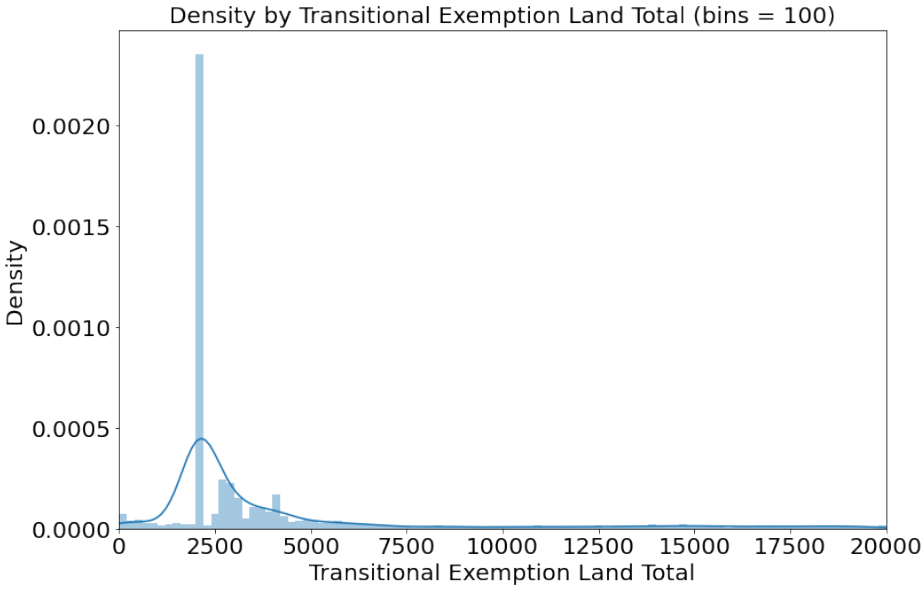
(27) Field Name: EXLAND2

Description: The transitional exemption land value of a specific property. This distribution shows the frequency density according to the properties' transitional exemption land value in this dataset. The range of 'Transitional Exemption Land Value' is set to be 0~50,000 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that most of the properties have transitional exemption land value within the range of 0~5,000.



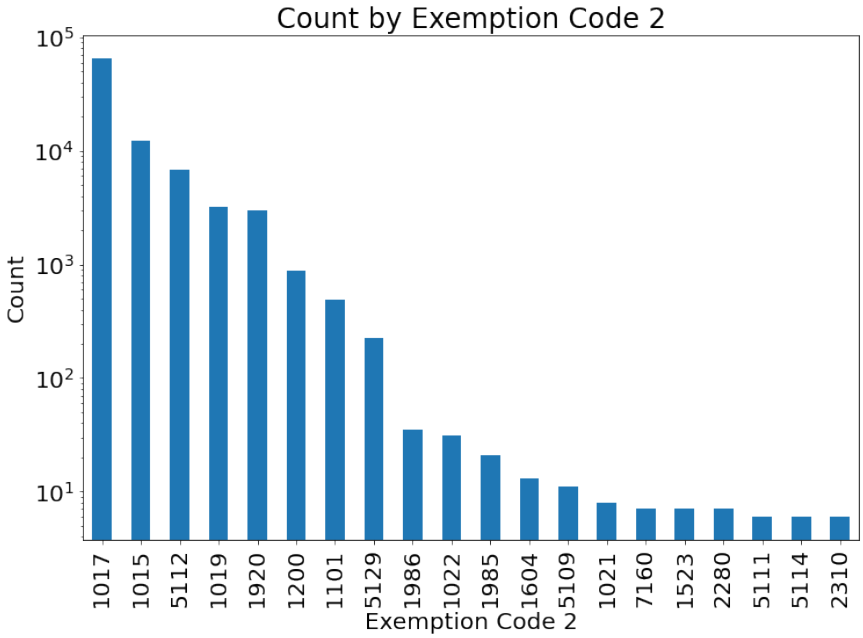
(28) Field Name: EXTOT2

Description: The transitional exemption and total of each property. This distribution shows the frequency density according to the properties' transitional exemption land total in this dataset. The range of 'Transitional Exemption Land Total' is set to be 0~20,000 because the vast majority of values are concentrated in this range and the range makes the visualization more readable. The distribution indicates that most of the properties have transitional exemption and total within the range of 0~5,000.



(29) Field Name: EXCD2

Description: The Exemption Code 2 of each property. The distribution shows the top 20 values of this field. The most common value is 1017, the count of which is about 70,000.



(30) Field Name: PERIOD
Description: The assessment period of each property. All the values of this field are FINAL in this dataset.

(31) Field Name: YEAR
Description: The assessment year of each property. All the values of this field are 2010/11 in this dataset.

(32) Field Name: VALTYPE
Description: The value type of each property. All the values of this field are AC-TR in this dataset.