

# **NEW YORK PROPERTY TAX FRAUD REPORT**

**By**

- 1. Ankur Ojha (3686240829)**
- 2. Aishwarya Joshi (8421803419)**
- 3. Santhoshini Jayachandran (8218065467)**
- 4. Aditya Chavan (8741411805)**

## **Table of Contents**

<b>Sr. No.</b>	<b>Title.</b>	<b>Pg. No.</b>
<b>1.</b>	<b>Executive Summary</b>	<b>1</b>
<b>2.</b>	<b>Description of Data</b>	<b>2</b>
<b>3.</b>	<b>Data Cleaning</b>	<b>4</b>
<b>4.</b>	<b>Variable Description</b>	<b>5</b>
<b>5.</b>	<b>Dimensionality Reduction</b>	<b>8</b>
<b>6.</b>	<b>Anomaly Detection</b>	<b>11</b>
<b>7.</b>	<b>Result</b>	<b>12</b>
<b>8.</b>	<b>Summary</b>	<b>12</b>
<b>9.</b>	<b>Case Study on top 10 properties having highest fraud score</b>	<b>14</b>
<b>10.</b>	<b>Appendix</b>	<b>19</b>

## 1. Executive Summary

This project aimed to build an unsupervised fraud model to analyze NYC property data for potential indications of fraud. The Property Valuation and Assessment Data of New York City was obtained from the NYC Open Data website and cleaned by filling missing values for nine key fields. Feature engineering was performed by creating 45 new variables and z-scaling the data to reduce dimensionality using principal component analysis (PCA) to 5 PCs.

Two anomaly detection algorithms (Autoencoder and PCA) were used to identify potential fraud cases, and the scores were combined to obtain a final fraud score. The top 100 records were explored, and a heatmap of the variable z-scores was used to identify which variables were driving the high score for these top properties.

In our analysis, we identified these reasons for the anomalies we observed in the data:

- The dollar value for building size is unusually high when the BLDFRONT and BLDDEPTH are unusually low. (Compared to other properties in the same borough, zip code, or tax class, which are typically found in the FULLVAL, AVLAND, and AVTOT fields.)
- Some properties have BLDFRONT and BLDDEPTH values that are unusually low compared to the lot size, which may indicate incorrect or incomplete building size data.
- There are usually high values for lot sizes for very small building front and depth.

Further research and inquiries are needed to verify whether the anomalies were caused by human error, plausible explanations, or fraudulent activities. Overall, the fraud model provided valuable insights into potential fraudulent activities, highlighting the importance of regular monitoring and review of property data to detect any suspicious activity.

## 2. Description of Data

The dataset (New York Property Valuation and Assessment) was provided by NYC Department of Finance. This Dataset has 32 fields and 1070994 records. The dataset represents NYC property assessments for the purpose of calculating Property Tax and Grant eligible properties Exemptions and/or Abatements. Data is collected and entered into the system by various City employees, like Property Exemption specialists, Property Assessors, ACRIS reporting, Department of Building reporting, etc.

### Summary Tables

#### Numerical Variables

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Mode	Min Value	Standard Deviation	Max Value
LTFRONT	1,070,994	100.00%	169,108	1297	0	0	9,999	74.03284
LTDEPTH	1,070,994	100.00%	170,128	1370	100	0	9,999	76.39628
STORIES	1,014,730	94.75%	0	111	2	1	119	8.365707
FULLVAL	1,070,994	100.00%	13,007	109,324	0	0	6.15E+09	11582431
AVLAND	1,070,994	100.00%	13,009	70,921	0	0	2.67E+09	4057260
AVTOT	1,070,994	100.00%	13,007	112,914	0	0	4.67E+09	6877529
EXLAND	1,070,994	100.00%	491,699	33,419	0	0	2.67E+09	3981576
EXTOT	1,070,994	100.00%	432,572	64,255	0	0	4.67E+09	6508403
BLDFRONT	1,070,994	100.00%	228,815	612	0	0	7,575	35.5797
BLDDEPTH	1,070,994	100.00%	228,853	621	0	0	9,393	42.70715
AVLAND2	282726	26.40%	0	58591	2408	3	2.37E+09	6178963
AVTOT2	282732	26.40%	0	111360	750	3	4.5E+09	11652529
EXLAND2	87449	8.17%	0	22195	2090	1	2.37E+09	10802213
EXTOT2	130828	12.22%	0	48348	2090	7	4.5E+09	16072510

## Categorical Variables

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common Value
RECORD	1,070,994	100.00%	0	1,070,994	1
BBLE	1,070,994	100.00%	0	1,070,994	1000010101
BORO	1,070,994	100.00%	0	5	4
BLOCK	1,070,994	100.00%	0	13,984	3944
LOT	1,070,994	100.00%	0	6,366	1
EASEMENT	4,636	0.43%	0	12	E
OWNER	1,039,249	97.04%	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100.00%	0	200	R4
TAXCLASS	1,070,994	100.00%	0	11	1
EXT	354,305	33.08%	0	3	G
EXCD1	638,488	59.62%	0	129	1017
STADDR	1,070,318	99.94%	0	839,280	501 SURF AVENUE
ZIP	1,041,104	97.21%	0	196	10314
EXMPTCL	15,579	1.45%	0	14	X1
EXCD2	92,948	8.68%	0	60	1017
PERIOD	1,070,994	100.00%	0	1	FINAL
YEAR	1,070,994	100.00%	0	1	2010/11
VALTYPE	1,070,994	100.00%	0	1	AC-TR

### 3. Data Cleaning

#### Exclusion Logic Description

For the column Owner, rows where "OWNER" contains 'NYC', 'NEW', 'YORK', 'DEPT', 'STATE OF NEW YORK', 'PUBLIC SERV', 'BOARD', 'GOVT OWNED', 'CNY', 'PRESERVATION', 'PARKS', 'PARK', 'GOVERNMENT', 'NATIONAL', 'DEPARTMENT', 'CITY', 'N.Y.C.', 'N.Y.', 'N.Y.C', 'N.Y', 'YORK CITY', 'NYS', 'NYS DEPT', 'NEW YORK CITY' ignoring case were removed. About 8,800 were removed

Then later top 20 count row was identified, and public properties was removed again  
About 34,985 rows were removed.

#### Imputation Logic Description

This section provides a clear and concise description of the logic used to impute missing values for the fields required to calculate the variables. The description explains how missing values were identified and what strategy was used to fill in these missing values. The logic should be reasonable and consider any relevant information about the dataset or domain knowledge.

Only 9 fields are considered for logical imputation: FULLVAL, AVLAND, AVTOT, ZIP, STORIES, LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH.

1. ZIP: Fill in missing ZIP values with the mode ZIP for the tax
2. STORIES: First created a new column named FAR this is the Floor Area Ratio. This is the ratio of building volume to lot size. This was chosen as every county has a set of rules for floor to area ratio. Using non-zero entries mean FAR is calculated for every zip code. Then this is used to fill the missing number of stories.
3. FULLVAL: Fill in missing FULLVAL values by calculating the average FULLVAL for properties with the same building class (TAXCLASS) and in the same zip code (ZIP).
4. AVLAND: Fill in missing AVLAND values by calculating the average AVLAND for properties with the same building class (TAXCLASS) and in the same zip code (ZIP).
5. AVTOT: Fill in missing AVTOT values by calculating the average AVTOT for properties with the same building class (TAXCLASS) and in the same zip code (ZIP).
6. LTFRONT: Fill in missing LTFRONT values by taking the median LTFRONT value for properties with the same building class (TAXCLASS) and on the same block (BLOCK).
7. LTDEPTH: Fill in missing LTDEPTH values by taking the median LTDEPTH value for properties with the same building class (TAXCLASS) and on the same block (BLOCK).
8. BLDFRONT: Fill in missing BLDFRONT values by taking the median BLDFRONT value for properties with the same building class (TAXCLASS) and on the same block (BLOCK).
9. BLDDEPTH: Fill in missing BLDDEPTH values by taking the median BLDDEPTH value for properties with the same building class (TAXCLASS) and on the same block (BLOCK)

## 4. Variables Description

This section provides a clear and concise description of all calculated variables, including the logic for why each variable measures some kind of unusualness that is of interest. The description explains show each variable was calculated and what it measures. Variable are created are limited as it is unsupervised. Inverse of ratio variable is taken so the outlier is easily identified with larger values.

Variable Name	Description
lotarea	The product of the LTFRONT and LTDEPTH fields, which gives the total area of the lot in square feet.
bldarea	The product of the BLDFRONT and BLDDEPTH fields, which gives the total area of the building in square feet.
bldvol	The product of bldarea and STORIES, which gives the total volume of the building.
r1	The ratio of FULLVAL to lotarea, which gives the price per square foot for the land.
r2	The ratio of FULLVAL to bldarea, which gives the price per square foot for the building.
r3	The ratio of FULLVAL to bldvol, which gives the price per cubic foot for the building.
r4	The ratio of AVLAND to lotarea, which gives the assessed value per square foot for the land.
r5	The ratio of AVTOT to lotarea, which gives the assessed value per square foot for the total property.
r6	The ratio of FULLVAL to AVLAND, which gives the relative price of the land compared to the total property value.
r7	The ratio of FULLVAL to AVTOT, which gives the relative price of the total property compared to the assessed value.
r8	the ratio of the total assessed value of the property (V3) to the area of the building on the property
r9	r9 is a variable that represents the ratio of the total property value v3 to the total area of the property s3
r1_inv	The inverse of r1, which flags unusually small values of the land price per square foot.
r2_inv	The inverse of r2, which flags unusually small values of the building price per square foot.
r3_inv	The inverse of r3, which flags unusually small values of the building price per cubic foot.
r4_inv	The inverse of r4, which flags unusually small values of the assessed value per square foot for the land.

r5_inv	The inverse of r5, which flags unusually small values of the assessed value per square foot for the total property.
r6_inv	The inverse of r6, which flags unusually large values of the relative price of the land compared to the total property value.
r7_inv	The inverse of r7, which flags unusually large values of the relative price of the total property compared to the assessed value.
r8_inv	r8_inv refers to the inverse of r8, this variable can be used to identify unusually small \$ values as outliers.
r9_inv	r9_inv refers to the inverse of r9, this variable can be used to identify unusually small \$ values as outliers.
VRnorm	The normalized version of r8, with the mean value set to 1.
Value_ratio	The maximum value of VRnorm and its inverse, which flags unusually large or small values of the ratio of the total property value to the sum of the assessed values for the land and total property.
FAR	Floor to Area Ratio (bldvol/lotarea)
r1_zip5	r1_zip5: Ratio of FULLVAL to lot area, which gives the price per square foot for the property. Grouped by zip5 class then averaged.
r2_zip5	r2_zip5: Ratio of total building area (BLDAREA) to lot area, which gives the building density per unit land area. Grouped by zip5 class then averaged
r3_zip5	r3_zip5: Ratio of building area of units built before 1940 (BLDAREA_LT40) to lot area, which gives the density of pre-1940 buildings per unit land area. Grouped by zip5 class then averaged.
r4_zip5	r4_zip5: Ratio of land area of property that is zoned for commercial use (LANDAREA_COMR) to lot area, which gives the density of commercial zoned land per unit land area. Grouped by zip5 class then averaged.
r5_zip5	r5_zip5: Ratio of land area of property that is zoned for residential use (LANDAREA_RES) to lot area, which gives the density of residential zoned land per unit land area. Grouped by zip5 class then averaged.
r6_zip5	r6_zip5: Ratio of building area of property that is zoned for commercial use (BLDAREA_COMR) to lot area, which gives the density of commercial buildings per unit land area. Grouped by zip5 class then averaged.
r7_zip5	r7_zip5: Ratio of building area of property that is zoned for residential use (BLDAREA_RES) to lot area, which gives the density of residential buildings per unit land area. Grouped by zip5 class then averaged.
r8_zip5	r8_zip5: Ratio of total property value (FULLVAL) to total building area (BLDAREA), which gives the price per unit building area. Grouped by zip5 class then averaged.
r9_zip5	r9_zip5: Ratio of total property value (FULLVAL) to product of total building area and number of stories (BLDAREA*STORIES), which gives the price per unit of building area and story. Grouped by zip5 class then averaged
r1inv_zip5	The inverse of ratio of FULLVAL to Lotarea, which gives the price per square foot for the land. Grouped by zip5 class then averaged.



r2inv_zip5	The inverse of ratio of r2, grouped by zip5 class then averaged.
r3inv_zip5	r3inv_zip5: Inverse ratio of FULLVAL to AVLAND.
r4inv_zip5	Inverse ratio of FULLVAL to AVTOT.
r5inv_zip5	Inverse ratio of AVLAND to lot area.
r6inv_zip5	Inverse ratio of AVTOT to lot area.
r7inv_zip5	Inverse ratio of AVLAND to AVTOT.
r8inv_zip5	Inverse ratio of FULLVAL to the building area.
r9inv_zip5	Inverse ratio of AVTOT to the building area.
r1_taxclass	The ratio of FULLVAL to lotarea, which gives the price per square foot for the land. Grouped by tax class then averaged.
r2_taxclass	The ratio of FULLVAL to bldarea, which gives the price per square foot for the building. Grouped by tax class then averaged.
r3_taxclass	The ratio of FULLVAL to bldvol, which gives the price per cubic foot for the building. Grouped by tax class then averaged.
r4_taxclass	The ratio of AVLAND to lotarea, which gives the assessed value per square foot for the land. Grouped by tax class then averaged.
r5_taxclass	The ratio of AVTOT to lotarea, which gives the assessed value per square foot for the total property. Grouped by tax class then averaged.
r6_taxclass	The ratio of FULLVAL to AVLAND, which gives the relative price of the land compared to the total property value. Grouped by tax class then averaged.
r7_taxclass	The ratio of FULLVAL to AVTOT, which gives the relative price of the total property compared to the assessed value. Grouped by tax class then averaged.
r8_taxclass	The ratio of the total assessed value of the property (V3) to the area of the building on the property. Grouped by tax class then averaged.
r9_taxclass	It is a variable that represents the ratio of the total property value v3 to the total area of the property s3. Grouped by tax class then averaged.
r1inv_taxclass	r1inv_zip5 is the inverse of the ratio of r1, Grouped by tax class and averaged.
r2inv_taxclass	r2inv_zip5 is the inverse of the ratio of FULLVAL to lot area, indicating price per square foot for land, grouped by tax class and averaged.
r3inv_taxclass	r1inv_zip5 is the inverse of the ratio of FULLVAL to lot area, indicating price per square foot for land, grouped by tax class and averaged.
r4inv_taxclass	r1inv_zip5 is the inverse of the ratio of FULLVAL to lot area, indicating price per square foot for land, grouped by tax class and averaged.

r5inv_taxclass	refers to the inverse of the ratio of the building area to the lot area, which gives the price per square foot for the building
r6inv_taxclass	The inverse of ratio of V2 (AVLAND) to S2 (building area), which gives the price per square foot for the land. Grouped by TAXCLASS then averaged.
r7inv_taxclass	The inverse of the ratio of AVLAND to the building square footage, grouped by tax class and averaged.
r8inv_taxclass	The inverse of ratio of V3 to S2, representing the price per cubic foot of the building, grouped by tax class then averaged.
r9inv_taxclass	The inverse of the ratio of AVTOT to building area, grouped by tax class code, and then averaged.

## 5. Dimensionality Reduction

Dimensionality reduction refers to the process of decreasing the number of variables in a dataset. This technique finds application in various domains like statistics, data analysis, speech, and image processing, where large datasets with high dimensions are difficult to handle using traditional algorithms due to the curse of dimensionality. By reducing dimensions, dimensionality reduction facilitates faster algorithm runtimes, efficient storage of data, elimination of feature correlation, and enables easy visualization of data by plotting 2D or 3D datasets. The two methods used for dimensionality reduction in this work are explained as principal component analysis (PCA) and autoencoder.

### Z – Scaling

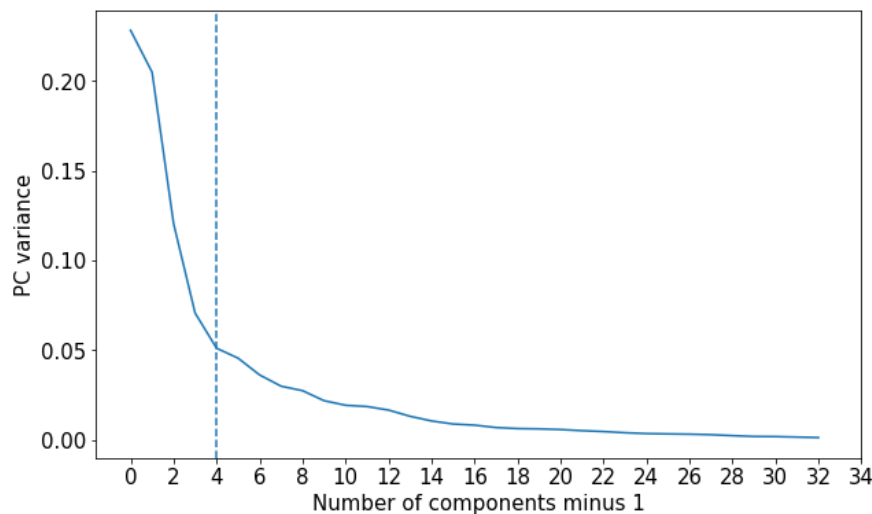
To prepare our dataset for dimensionality reduction, we start by applying the z-scaling method, which involves a linear transformation of the dataset to ensure that each feature (or variable) has a mean value of zero and a standard deviation value of one. Mathematically, if we consider a dataset  $X \in \mathbb{R}^m \times n$ , where there are  $m$  data points and  $n$  features each, then we perform the following transformation:

$$z_i = (x_i - \bar{x})/\sigma_i, \forall i \in \{1, \dots, n\}$$

The main advantage of standardizing a dataset is that every dimension of the dataset is centered around zero with the same scaling, and this helps identify outlier values using a simple measure of distance from the origin. In Figures 1a and 1b, we can see the results of standardizing a sample dataset that contains 10,000 values. We observe that the standardized dataset is distributed around zero, with values between -1 and 1 standard deviations. However, it's important to note that the histogram of the standardized dataset may not be perfectly centered around zero due to outliers that shift the majority of the data towards more positive or negative values.

### Principal Component Analysis

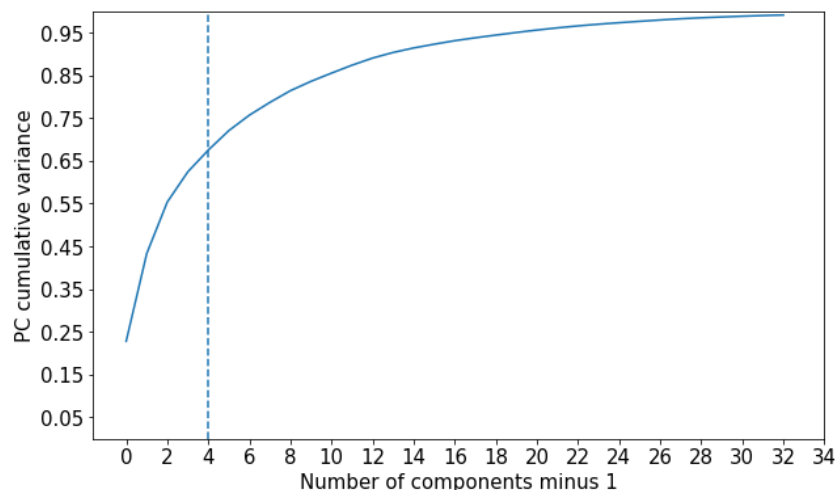
PCA is a popular technique for dimensionality reduction that involves transforming the original dataset to create a new set of features that captures the maximum variance in the data. The transformation is performed in such a way that the new features, or principal components, are ordered according to their variance, and those features with low variance are believed to have high linear correlation with other features. By choosing the top  $k$  features that contain a desired percentage of the total variance, PCA reduces the dimensionality of the dataset while retaining the most important information. Therefore, PCA is a powerful tool for extracting meaningful insights from large datasets with many features.



For the above graph, the x-axis represents the number of components minus one, while the y-axis represents the proportion of total variance explained by each component. The first principal component always explains the most variance, with subsequent components explaining progressively less variance.

The plot shows a steep drop-off in variance explained between the first few principal components, indicating that these components capture the most important information in the data. As we move further to the right on the plot, the proportion of variance explained by each additional component decreases, suggesting that these components may be less important for describing the variation in the data.

The vertical dotted line drawn at  $x=4$  in the plot indicates the number of components where the explained variance ratio exceeds 0.05, which is often used as a rule of thumb for selecting the number of components to keep in PCA. In this case, it suggests that keeping the first 4 principal components may be a reasonable choice for retaining most of the important information in the data while reducing the dimensionality of the dataset.



The graph shows the proportion of total variance in the original data explained by including increasing numbers of principal components. It starts with the first principal component, which by definition

explains the most variance, and progressively adds subsequent components until all components are included.

The steepness of the curve represents the rate at which the variance is explained, with steeper sections indicating that adding more components captures more variance. In general, a steep initial curve followed by a more gradual curve suggests that the first few components are important for describing the variation in the data. The vertical line drawn at  $x=4$  shows the point at which 99% of the variance in the data is explained and can be used to guide the selection of the number of components to retain for further analysis.

## **Autoencoder**

Autoencoder is a type of neural network or statistical model used for learning data encodings. Its architecture consists of an input layer, an output layer, and one or more hidden layers that compress the input data to a lower-dimensional space before reconstructing it back to the output layer. By doing so, autoencoders can effectively perform dimensionality reduction and remove noise or outliers in the data. This makes them useful for various applications, such as image processing or fraud detection. Autoencoders are trained to learn the best possible encoding of the input data and are optimized to minimize the difference between the input and output data.

In this project, we trained a neural network model using the `MLPRegressor` class from `scikit-learn` library. The model was trained on a standardized dataset obtained after performing PCA (Principal Component Analysis) on the original dataset. The `MLPRegressor` class was configured to have a single hidden layer with 3 neurons and a logistic activation function. The maximum number of iterations was set to 50, and a random seed of 1 was used to ensure reproducibility. The model was trained using the `fit()` method on the PCA-transformed dataset. The goal of this experiment was to explore the effectiveness of using neural networks to model the relationship between the principal components obtained from PCA and to evaluate the potential benefits of this approach for reducing the dimensionality of the dataset.

## 6. Anomaly Detection Algorithms

To calculate the fraud scores for each data point, we utilized the distance measure of the standardized output from both PCA and autoencoder. The distance measure, also known as the norm of a vector, is determined by a parameter  $p$  that defines the type of distance metric used. The most commonly used values for  $p$  are  $p=1$ ,  $2$ , and  $\infty$ , with  $p=2$  being used in this study. We calculated the norm of the vector for each data point and used this as the basis for assigning fraud scores. All three  $p$  values produced similar results, but we chose to present the results for  $p=2$  in our simulations.

### Score 1

The first score (S1) represents the 2-norm of the standardized PCA output, calculated as  $S1 = \|Pz\|_2$ . Since PCA aims to maximize the variance of the data, any outlier values would contribute to the variance and thus be included in the dimensionality reduction performed by PCA. Therefore, such values can be detected through a distance metric in the standardized dataset.

$$s_i^1 = \left( \sum_k |PCz_k^i|^p \right)^{1/p}, \quad p \text{ anything}$$

### Score 2

The second score (S2) represents the 2-norm of the difference between the autoencoder input,  $Pz$  (standardized PCA output), and the autoencoder output,  $Pa$ , calculated as  $S2 = \|Pz - Pa\|_2$ . The autoencoder is designed to reduce the dimensionality of the data in its hidden layers while reproducing the input in its output. As a result, outlier values in the input are lost in the output, making it possible to detect them using a distance metric between the input and output.

$$s_i^2 = \left( \sum_k \overset{\text{Autoencoder output}}{|PCz_k'^i - PCz_k^i|^p} \right)^{1/p}, \quad p \text{ anything}$$

### Final Score

The final score was obtained through a weighted average of the rank orders of two scores - score1 and score2. In Step 1, the rank orders of both scores were calculated using the ``rank()`` function in pandas, which assigns a rank to each value based on its position in the sorted list. In Step 2, the final score was obtained by taking the average of the two rank orders, with each rank order given equal weight (i.e., multiplied by 0.5). This produces a final score that reflects the combined relative performance of the two scores, where higher ranks correspond to higher scores. Thus, the final score is a weighted average of the rank orders of score1 and score2, with each rank order given equal weight.

## 7. Result

The fraud scores produced by both the autoencoder, and z-scores were found to be skewed to the right. This was expected because many, even after removing some of the government properties and parks many were left that have missing values, including missing building fronts and depths, and they have high property value with fewer stories. Upon closer analysis of the data, we found that some properties had exceptionally high or low values, while others had incomplete property data, such as missing values in the lot depth or lot front fields. And also, some of the properties whose lot size was very small, but their total value was very high. Additionally, the full value of the property per building area was found to be either excessively high or low in some cases (compared to other properties in the same borough, zip code, or tax class).

## 8. Summary

To examine the top properties with the highest fraud scores, we sort the records by the final fraud score in descending order and focus on the top 100 records. Additionally, a heatmap of the variable z-scores is used to identify which variables are driving the high score for these top properties. This information helped us better to understand the reasons behind the high fraud score and guide further investigation into potential fraudulent activities.

### **Steps followed for using fraud score after feature Engineering:**

**1. Data cleaning:** Clean the data by removing irrelevant information and handling missing data using imputation techniques. The missing values for nine key fields were filled.

**2. Feature engineering:** Create new variables that are designed to look for the kinds of anomalies or frauds you are interested in. In this case, variables that help to identify unusual property valuations were created.

**3. Dimensionality reduction:** Since the dataset has a high number of features, it's important to reduce the number of features to avoid overfitting. One way to do this is to remove correlations and reduce dimensionality using principal component analysis (PCA). Z-scale the data so that they are on the same footing before conducting PCA to reduce dimensionality of the data to 5 principal components (PCs). Z-scale the 5 PCs again to make each retained PC equally important.

**4. Anomaly detection:** Use two different anomaly detection algorithms to identify potential fraud cases. The first method looks for outliers in the final scaled PC space using a Minkowski distance from the origin. The second method involves building a simple autoencoder, and the fraud score is then the reproduction error.

**5. Combining scores:** Since we have two scores, we average them to obtain a final fraud score. Replace the score with its rank order, and then average the rank-ordered scores for the final score.

**6. Sort the data:** Sort the records by the final score and explore the top records to investigate potential fraud cases. A heat map of the variable z-scores can help identify which variables are driving the top scores.

To improve the fraud model and investigate potential property fraud further, we recommend taking the following steps:

**1. Explore additional data cleaning methods:** While we have used various methods to fill in missing values, we could explore additional techniques such as imputation methods to identify and clean the data more effectively.

**2. Seek expert opinions:** Subject matter experts in the field of fraud examination and real estate industry could provide valuable insights into the potential causes of the anomalies in the data and help form hypotheses regarding fraudulent activities.

**3. Verify anomalous data against third-party sources:** Conducting site visits or internet research could help verify the integrity of anomalous data and strengthen the fraud detection model.

**4. Incorporate additional data sources:** Incorporating additional data sources such as property ownership, crime rates, and average income in the ZIP codes' areas could provide a more comprehensive picture of the properties and owners, and ultimately improve the model's accuracy.



## 9. Case Study on Top 10 Properties that have highest Fraud Scores

Case Study on Top 10 Properties that have highest Fraud Scores

RECORD	917942	956520	658933	1059883	139726	665158	116647	684704	12076	333412
OWNER	LOGAN PROPERTY, INC.	TROMPETA RIZALINA	WAN CHIU CHEUNG		BRADHURST EQUITIES, L	ST JOHNS CEMETERY	MF ASSOCIATES OF NEW	W RUFERT	15 WORTH STREET PROPE	SPOONER ALSTON
LTFRONT	4910	25	25	5	4	1412	25	2	74	17
LTDEPTH	0	91	100	5	5	2532	75	2	150	85
STORIES	3	3	3			1	35		1	3
FULLVAL	374,019,883.00	348,200.00	776,000.00	-	-	29,355,000.00	161,000,000.00	-	2,610,000.00	9060
AVLAND	1,792,808,947.00	15,600.00	26,940.00	-	-	13,140,000.00	19,215,000.00	-	1,170,000.00	3874
AVTOT	4,668,308,947.00	20,892.00	46,560.00	-	-	13,209,750.00	72,450,000.00	-	1,174,500.00	4077
STADDR	154-68 BROOKVILLE BOULEVARD	12 ONEIDA AVENUE	54-76 83 STREET	SAGONA COURT	BRADHURST AVENUE	80-01A METROPOLITAN AVENUE	1849 2 AVENUE	69 STREET	170 WEST BROADWAY	37 MONROE STREET
ZIP	11422	10301	11373			11379	10128		10013	11238
BLDFRONT	0	1812	2500	0	0	12	70	0	5	4017
BLDDEPTH	0	5020	5600	0	0	18	456	0	5	42
r1	0.8	-0.1	0.2	205.0	256.3	-0.4	162.1	176.6	0.0	-0.4
r2	68.2	-0.5	-0.5	0.0	0.0	133.3	4.4	-0.4	102.3	-0.5
r3	47.4	-0.5	-0.5	-0.3	-0.3	278.8	-0.2	-0.4	214.1	-0.5
r4	42.2	-0.1	0.0	253.7	317.2	-0.1	147.5	32.9	1.4	-0.1
r5	896.6	-0.1	-0.1	0.2	0.2	162.8	1.6	-0.1	125.3	-0.1
r6	643.3	-0.1	-0.1	0.0	0.0	350.6	0.0	-0.1	269.7	-0.1
r7	37.8	-0.1	0.0	251.4	314.2	-0.1	191.5	11.3	0.4	-0.1
r8	937.3	-0.1	-0.1	0.2	0.2	65.7	2.4	-0.1	50.5	-0.1
r9	898.6	-0.1	-0.1	0.1	0.1	189.0	0.1	-0.1	145.2	-0.1
r1inv	-0.1	-0.1	-0.1	-0.1	-0.1	0.0	-0.1	-0.1	-0.1	0.1
r2inv	-0.1	27.0	19.9	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	20.4
r3inv	-0.2	14.0	11.2	-0.1	-0.1	-0.2	-0.1	-0.1	-0.2	11.4
r4inv	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
r5inv	-0.2	25.6	25.1	-0.2	-0.2	-0.2	-0.2	-0.1	-0.2	9.0
r6inv	-0.3	9.1	9.1	-0.3	-0.3	-0.3	-0.3	-0.2	-0.3	5.3
r7inv	-0.2	-0.1	-0.1	-0.2	-0.2	0.0	-0.2	-0.2	-0.2	0.0
r8inv	-0.1	41.4	38.2	-0.1	-0.1	-0.1	-0.1	0.1	-0.1	14.8
r9inv	-0.2	17.3	16.8	-0.2	-0.2	-0.2	-0.2	0.1	-0.2	10.3
r1 zip5	2.2	0.2	0.2	342.0	176.0	-0.5	112.1	202.7	-0.1	-0.5
r2 zip5	66.6	-0.5	-0.5	0.0	0.1	93.1	4.2	-0.4	120.7	-0.5
r3 zip5	31.4	-0.4	-0.4	-0.2	-0.1	136.4	0.1	-0.3	354.7	-0.4
r4 zip5	102.4	0.0	0.1	492.4	180.5	-0.1	80.1	67.3	0.6	-0.1
r5 zip5	635.1	-0.1	-0.1	0.3	0.6	239.3	2.1	-0.1	139.7	-0.1
r6 zip5	429.9	-0.1	-0.1	0.0	0.3	272.1	0.1	-0.1	245.2	-0.1
r7 zip5	146.2	-0.1	0.0	560.5	140.7	-0.1	88.2	26.0	0.1	-0.2
r8 zip5	767.9	-0.1	-0.1	0.8	0.5	172.5	3.1	-0.1	59.4	-0.1
r9 zip5	631.2	-0.1	-0.1	0.2	0.3	243.7	0.4	-0.1	192.4	-0.1
r1inv zip5	-0.1	-0.1	-0.1	-0.1	-0.1	0.2	-0.1	-0.1	-0.1	0.1
r2inv zip5	-0.1	26.5	15.9	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	25.9
r3inv zip5	-0.1	7.8	8.3	-0.1	-0.1	-0.1	-0.1	0.0	-0.1	13.1
r4inv zip5	-0.3	-0.3	-0.3	-0.3	-0.3	-0.1	-0.3	-0.3	-0.3	-0.2
r5inv zip5	-0.2	23.3	15.3	-0.2	-0.2	-0.2	-0.2	0.1	-0.2	4.4
r6inv zip5	-0.2	7.5	5.3	-0.2	-0.2	-0.2	-0.2	0.2	-0.2	2.7
r7inv zip5	-0.2	-0.2	-0.2	-0.2	-0.2	0.1	-0.2	-0.2	-0.2	0.1
r8inv zip5	-0.1	30.1	27.7	-0.1	-0.1	-0.1	-0.1	0.2	-0.1	15.6
r9inv zip5	-0.2	11.0	12.5	-0.1	-0.1	-0.2	-0.2	0.5	-0.2	10.6
r1 taxclass	0.5	-0.1	0.1	142.3	178.0	-0.3	347.5	383.3	0.0	-0.3
r2 taxclass	47.6	-0.3	-0.3	0.1	0.1	92.9	11.2	0.0	71.3	-0.3
r3 taxclass	46.6	-0.3	-0.3	-0.1	-0.1	273.4	2.0	0.0	210.0	-0.3
r4 taxclass	20.0	0.1	0.3	121.6	152.1	-0.3	329.6	387.6	0.4	-0.3
r5 taxclass	670.2	-0.2	-0.2	0.0	0.0	121.6	6.9	0.0	93.5	-0.2
r6 taxclass	514.7	-0.2	-0.2	-0.1	-0.1	280.5	1.2	0.0	215.7	-0.2
r7 taxclass	21.7	0.0	0.3	145.4	181.8	-0.3	344.5	435.8	0.0	-0.3
r8 taxclass	766.1	-0.2	-0.2	0.0	0.0	53.6	6.8	0.0	41.1	-0.2
r9 taxclass	755.9	-0.2	-0.2	-0.1	-0.1	158.9	1.2	0.0	122.0	-0.2
r1inv taxclass	-0.2	0.0	-0.1	-0.2	-0.2	-0.1	-0.2	-0.2	-0.2	7.1
r2inv taxclass	-0.1	762.7	562.8	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	286.7
r3inv taxclass	-0.1	775.1	619.5	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	198.6
r4inv taxclass	-0.5	-0.2	-0.3	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.3
r5inv taxclass	-0.4	441.4	433.7	-0.4	-0.4	-0.4	-0.4	-0.3	-0.4	19.9
r6inv taxclass	-0.5	348.3	346.0	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	15.0
r7inv taxclass	-0.4	-0.1	-0.2	-0.4	-0.4	-0.2	-0.4	-0.4	-0.3	3.4
r8inv taxclass	-0.2	472.6	436.1	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	305.7
r9inv taxclass	-0.4	422.8	409.8	-0.4	-0.4	-0.4	-0.4	-0.3	-0.4	255.7
value ratio	3.7	-0.1	-0.1	0.3	0.3	0.4	0.2	-0.1	0.4	0.4

Record	917942	956520	658933	1059883	139726	665158	116647	684704	12076	333412
max of all z score variable	937.3	775.1	619.5	560.5	317.2	350.6	347.5	435.8	354.7	305.7

Field with strange value	AVTOT and AVLAND unreasonably high	BLDFRONT, BLDDEPTH unusually high compared to Lot size	BLDFRONT, BLDDEPTH unusually high compared to Lot size	LTDEPTH and LTFRONT is very low and building size is 0	LTDEPTH and LTDEPTH is very low	Dollar value for building size is unusually high (BLDFRONT, BLDDEPTH unusually LOW)	Dollar value for lot size is unusually high (Lot size is small as compared to Building size.)	Lot size (LTDEPTH & LTFRONT) unusually low and also BLDDEPTH & BLDFRONT is 0	BLDFRONT, BLDDEPTH unusually LOW compared to Lot size	BLDFRONT is unusually large (very large)
--------------------------	------------------------------------	--	--	--	---------------------------------	---	---	--	---	--

Some examples of properties that have high fraud scores:

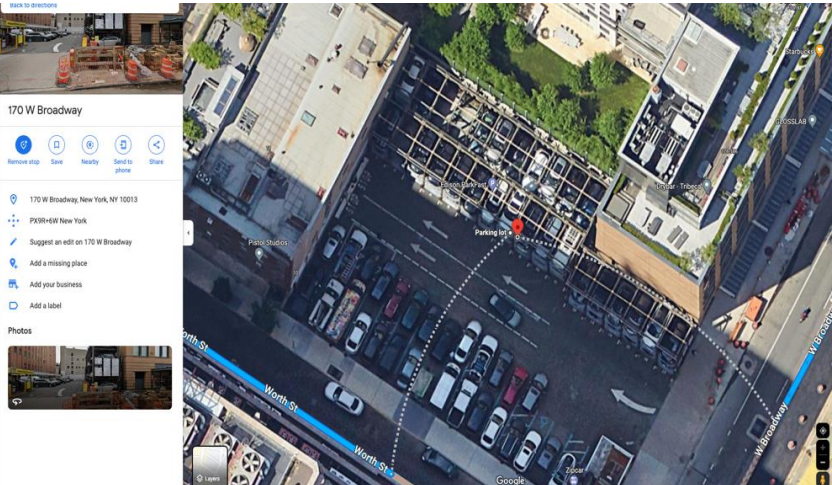
BLDFRONT is unusually large as compared to another dimension, and the FULLVAL, AVLAND, AVTOT is very low, for the residential property which indicate there might be some fraud.

RECORD	333412
BBLE	3019850059
BORO	3
BLOCK	1985
LOT	59
EASEMENT	
OWNER	SPOONER ALSTON
LTFRONT	17
LTDEPTH	85
STORIES	3
FULLVAL	9060
AVLAND	3874
AVTOT	4077
STADDR	37 MONROE STREET
ZIP	11238
BLDFRONT	4017
BLDDEPTH	42



BLDFRONT, BLDDEPTH unusually LOW as compared to Lot size, for this property.

RECORD	12076
OWNER	15 WORTH STREET PROPE
LTFRONT	74
LTDEPTH	150
STORIES	1
FULLVAL	2,610,000.00
AVLAND	1,170,000.00
AVTOT	1,174,500.00
EXLAND	-
STADDR	170 WEST BROADWAY
ZIP	10013
BLDFRONT	5
BLDDEPTH	5





Dollar value for lot size is unusually high (Lot size is small as compared to Building size.)

RECORD	116647
BLOCK	1541
LOT	21
OWNER	MF ASSOCIATES OF NEW
LTFRONT	25
LTDEPTH	75
STORIES	35
FULLVAL	161,000,000.00
AVLAND	19,215,000.00
AVTOT	72,450,000.00
STADDR	1849 2 AVENUE
ZIP	10128
BLDFRONT	70
BLDDEPTH	456

Directions Save Nearby Send to phone Share

1849 2nd Ave, New York, NY 10128

Suggest an edit on 1849 2nd Ave

Add a missing place

Add your business

Add a label

Photos

Information about address from net : This is a business registration address for five companies. These are some of the names: Dahlgren, Lauren and Rite Aid. Mf Associates of New York LLC owns this real estate property. 1986 is the year the property was built. The property is 37 years old, which is 69 years younger than the average age of a building in New York of 106 years. Estimated value for the property \$106,415,550. The property features 968,264 sqft of living area. The size of the land lot is 92,927 sqft. The building has 35 floors. (

<https://clustrmaps.com/a/2gf1mv/>)

Dollar value for building size is unusually high (BLDFRONT, BLDDEPTH unusually LOW)

RECORD	665158
OWNER	ST JOHNS CEMETERY
LTFRONT	1412
LTDEPTH	2532
STORIES	1
FULLVAL	29,355,000.00
AVLAND	13,140,000.00
AVTOT	13,209,750.00
STADDR	80-01A METROPOLITAN AVENUE
ZIP	11379
BLDFRONT	12
BLDDEPTH	18

## BLDFRONT, BLDDEPTH unusually high compared to Lot size

RECORD	658933
OWNER	WAN CHIU CHEUNG
LTFRONT	25
LTDEPTH	100
STORIES	3
FULLVAL	776,000.00
AVLAND	26,940.00
AVTOT	46,560.00
STADDR	54-76 83 STREET
ZIP	11373
BLDFRONT	2500
BLDEPTH	5600

54-76 83rd St  
Building

Directions Save Nearby Send to phone Share

54-76 83rd St, Queens, NY 11373


Suggest an edit on 54-76 83rd St

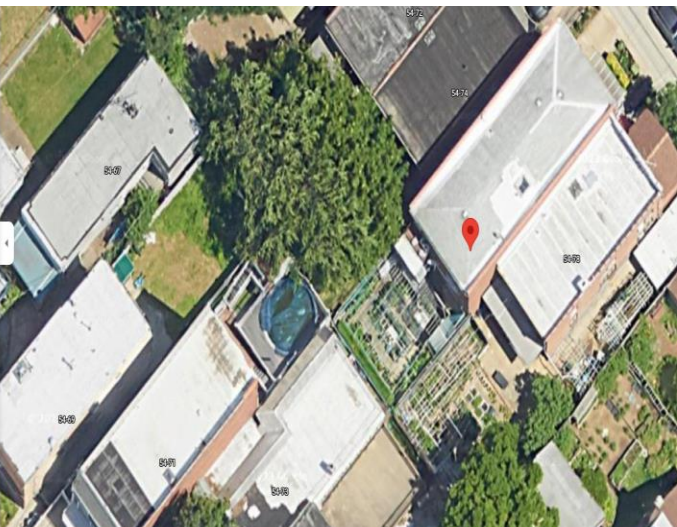
Add a missing place

Add your business

Add a label

Photos





## Observations after investigation of five fraud cases:

In our analysis, we identified these reasons for the anomalies we observed in the data:

- The dollar value for building size is unusually high when the BLDFRONT and BLDDEPTH are unusually low. (Compared to other properties in the same borough, zip code, or tax class, which are typically found in the FULLVAL, AVLAND, and AVTOT fields.)
- Some properties have BLDFRONT and BLDDEPTH values that are unusually low compared to the lot size, which may indicate incorrect or incomplete building size data.
- There are usually high values for lot sizes for very small building front and depth.

Record NO. 170125 : This property has a two-story building and its full value is about \$468million . This results in a very high value per building volume, which led us to investigate.

RECORD	OWNER	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	STADDR	ZIP	BLDFRONT
170125	BRONX V A MEDICAL CEN	587	994	2	468100000	9945000	210645000	110 WEST KING	10468	117

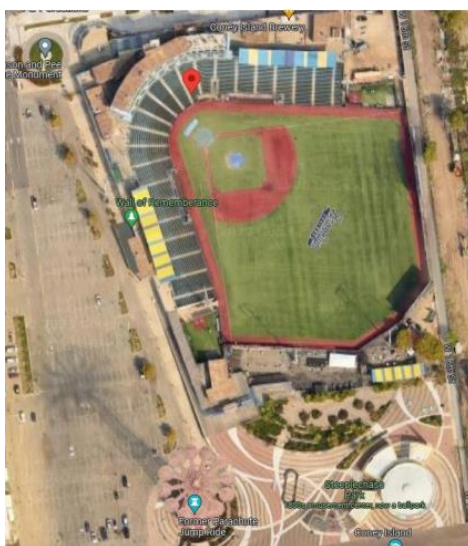


This investigation led us to identify that building had more than two story.

Record No. 935158: The front and depth of this building is 1, which is too small as compared to its values \$10.4 million,136 lot front,132 depth and 8 stories. Probably because the old data wasnot updated since the building was built in 2012.

RECORD	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP
935158	RICH-NICH REALTY,LLC	D3	2	136	132	1	1	8	1040000	236250	468000	10301

Record number: 504002 Lot Depth: 9999 Lot Front: 7960 Building Depth: 360 Building Front: 360 FAR:  $129,600/79,592,040=0.001628$  Full Value: 65,220,000 St. Address: 1904 Surf Avenue, Brooklyn, NY I created a variable called FAR (Floor to area ratio). Most counties have fixed regulations for this variable. The cost of the property was too high for FAR. Later after googling the streets address, I found out that it is a seaside baseball stadium with a park which increased the value of the property.





## Appendix

### Data Quality Report

#### 1. Data Description

The dataset (New York Property Valuation and Assessment) was **provided by NYC Department of Finance**. This Dataset has **32 fields** and **1070994 records**. The dataset represents NYC property assessments for the purpose of calculating Property Tax and Grant eligible properties Exemptions and/or Abatements. Data is collected and entered into the system by various City employees, like Property Exemption specialists, Property Assessors, ACRIS reporting, Department of Building reporting, etc.

#### Numerical Variables:

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Mode	Min Value	Standard Deviation	Max Value
LTFRONT	1,070,994	100.00%	169,108	1297	0	0	9,999	74.03284
LTDEPTH	1,070,994	100.00%	170,128	1370	100	0	9,999	76.39628
STORIES	1,014,730	94.75%	0	111	2	1	119	8.365707
FULLVAL	1,070,994	100.00%	13,007	109,324	0	0	6.15E+09	11582431
AVLAND	1,070,994	100.00%	13,009	70,921	0	0	2.67E+09	4057260
AVTOT	1,070,994	100.00%	13,007	112,914	0	0	4.67E+09	6877529
EXLAND	1,070,994	100.00%	491,699	33,419	0	0	2.67E+09	3981576
EXTOT	1,070,994	100.00%	432,572	64,255	0	0	4.67E+09	6508403
BLDFRONT	1,070,994	100.00%	228,815	612	0	0	7,575	35.5797
BLDDEPTH	1,070,994	100.00%	228,853	621	0	0	9,393	42.70715
AVLAND2	282726	26.40%	0	58591	2408	3	2.37E+09	6178963
AVTOT2	282732	26.40%	0	111360	750	3	4.5E+09	11652529

EXLAND2	87449	8.17%	0	22195	2090	1	2.37E+09	10802213
EXTOT2	130828	12.22%	0	48348	2090	7	4.5E+09	16072510

### Categorical Variables:

Field Name	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common Value
RECORD	1,070,994	100.00%	0	1,070,994	1
BBLE	1,070,994	100.00%	0	1,070,994	1000010101
BORO	1,070,994	100.00%	0	5	4
BLOCK	1,070,994	100.00%	0	13,984	3944
LOT	1,070,994	100.00%	0	6,366	1
EASEMENT	4,636	0.43%	0	12	E
OWNER	1,039,249	97.04%	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100.00%	0	200	R4
TAXCLASS	1,070,994	100.00%	0	11	1
EXT	354,305	33.08%	0	3	G
EXCD1	638,488	59.62%	0	129	1017
STADDR	1,070,318	99.94%	0	839,280	501 SURF AVENUE
ZIP	1,041,104	97.21%	0	196	10314
EXMPTCL	15,579	1.45%	0	14	X1
EXCD2	92,948	8.68%	0	60	1017
PERIOD	1,070,994	100.00%	0	1	FINAL
YEAR	1,070,994	100.00%	0	1	2010/11
VALTYPE	1,070,994	100.00%	0	1	AC-TR

## 2. Visualization of Each Field

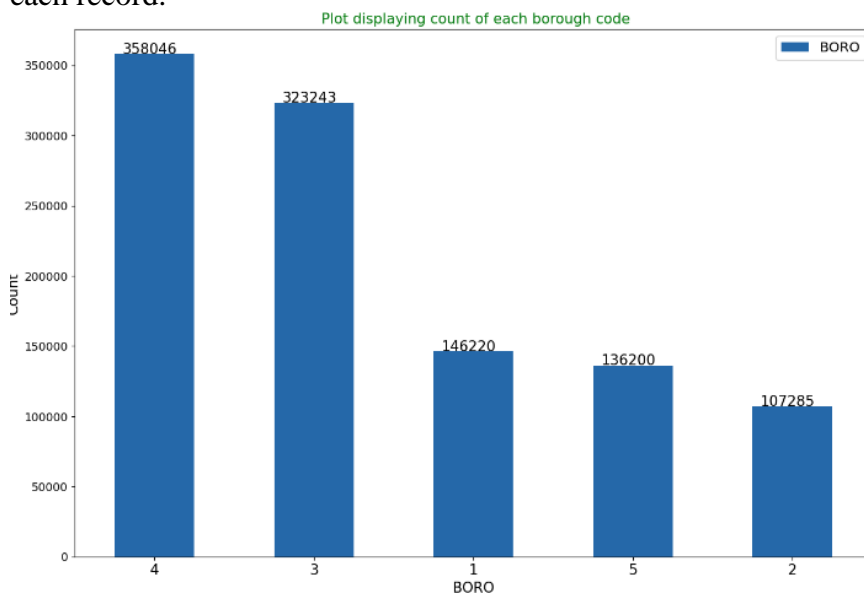
The Line charts, KDE and bar charts represented for each categorical and numerical field give some high-level information of fields of New York Property Valuation and Assessment Data.

1. **Field Name:** RECORD

**Description:** A categorical field containing unique positive integers for each record from 1 to 10,70994.

2. **Field Name:** BBLE

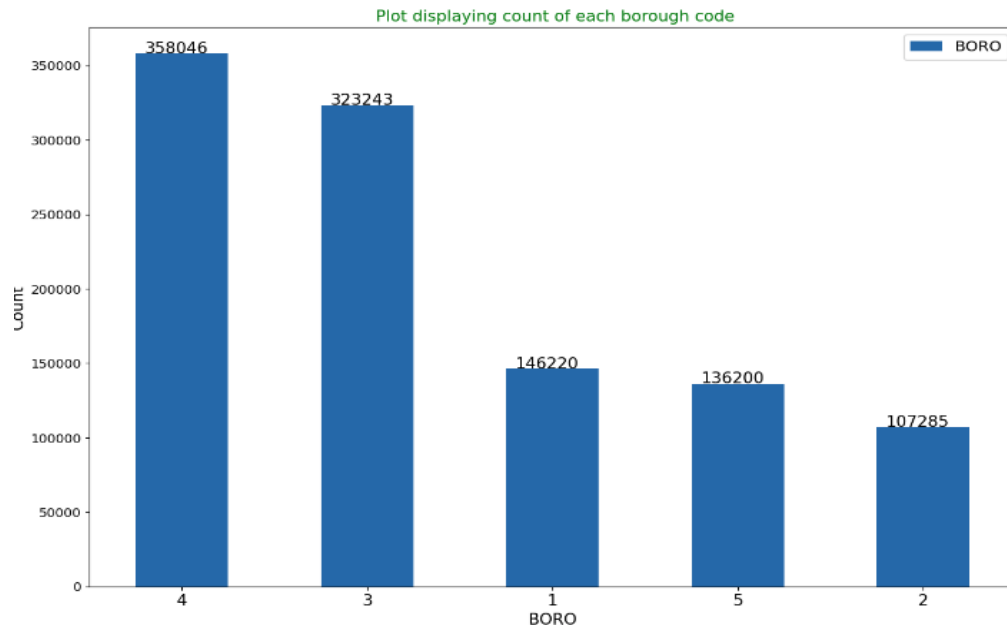
**Description:** This field is the concatenation of BORO, BLOCK, LOT, and EASEMENT fields, giving a unique number to each record. There are 10,70994 unique values of BBLE which is one unique value for each record.



3. **Field Name:** BORO

**Description:** This field represents Borough codes which are as follows: (1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island)  
The bar chart below shows the count for each borough code in descending order of their count.

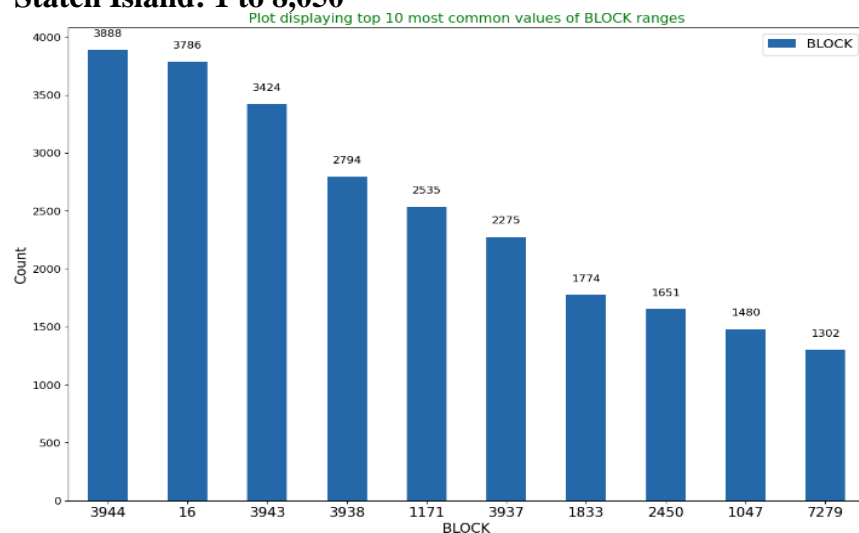




#### 4. Field Name: BLOCK

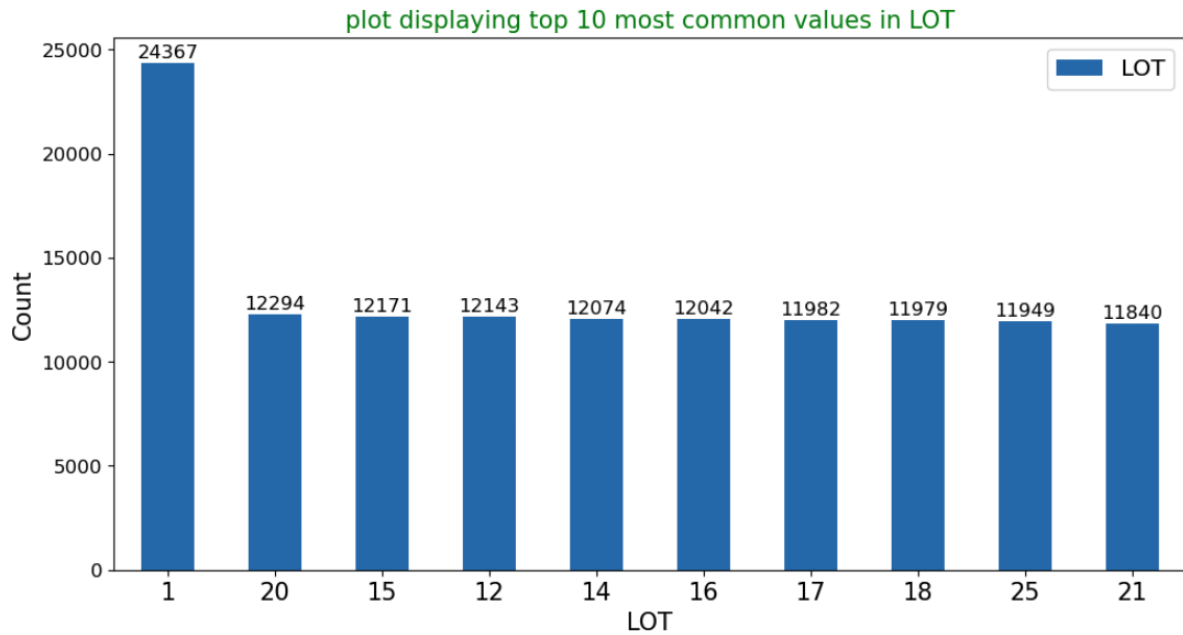
**Description:** This field represents the valid block ranges by borough codes. This histogram plot shows the count of 10 most common values for Block ranges.

- **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**



#### 5. Field Name: LOT

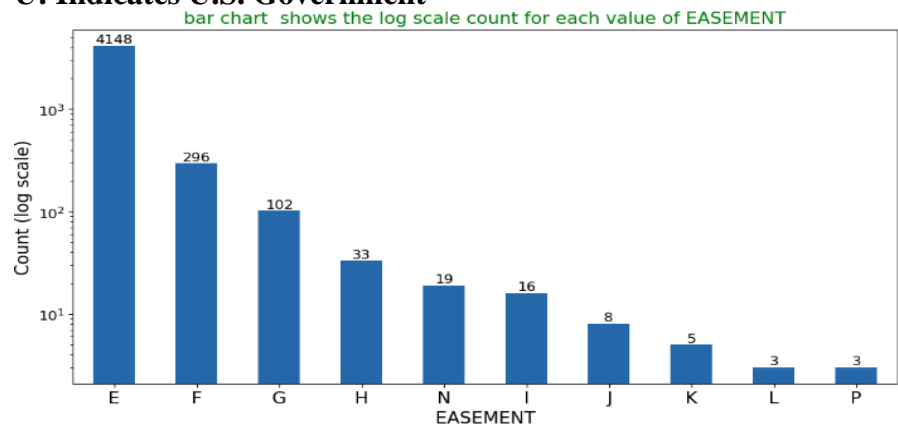
**Description:** This field represents unique codes withing borough codes and ranges.



6. **Field Name:** EASEMENT

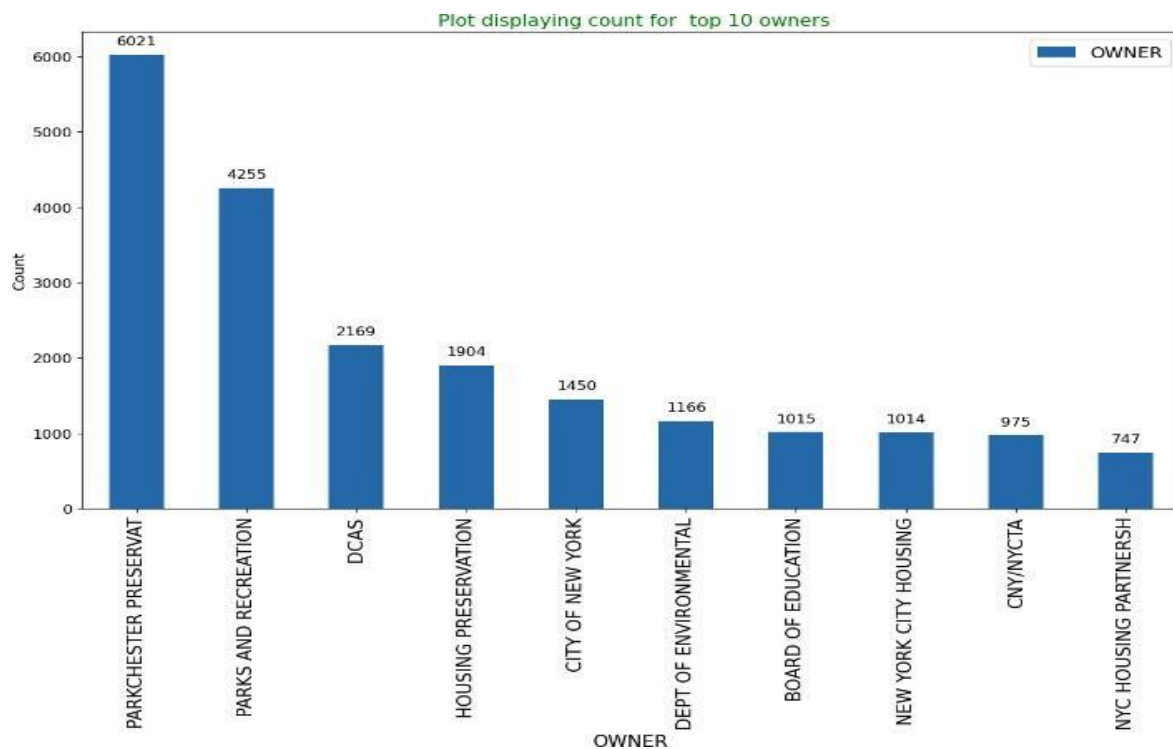
**Description:** This field represents a legal right to use another's land for a specific purpose.

- **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 through M** Are duplicates of 'E'
- **N:** Indicates Non-Transit Easement
- **P:** Indicates Piers
- **R:** Indicates Railroads
- **S:** Indicates Street
- **U:** Indicates U.S. Government

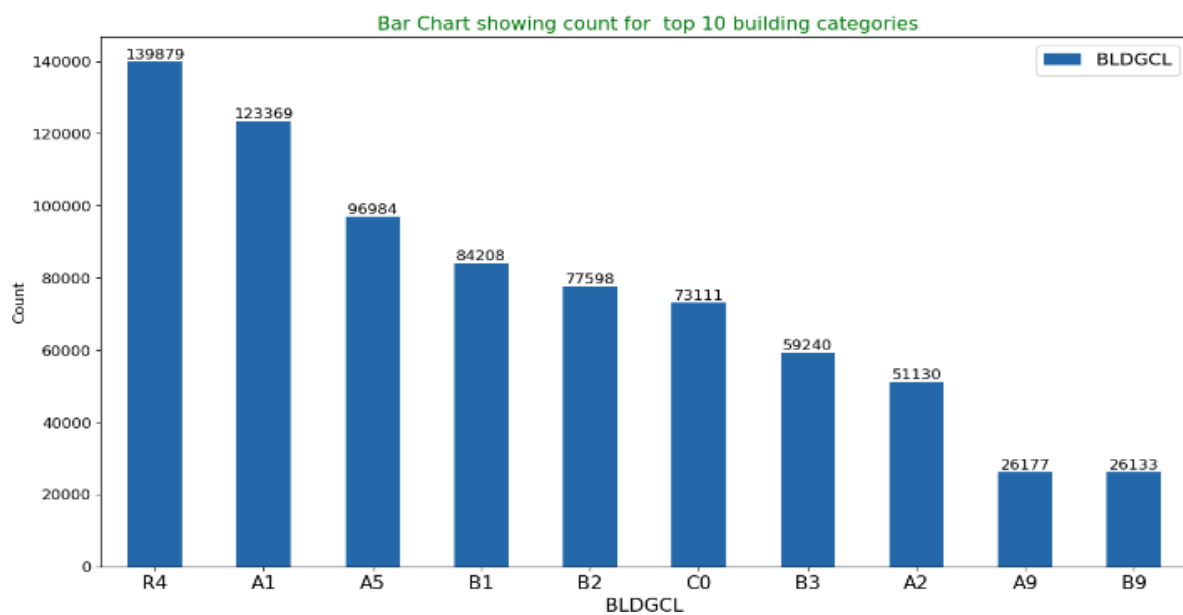


7. **Field Name:** OWNER

**Description:** This field represents the property owner's name, plot displaying count for top 10 owners.

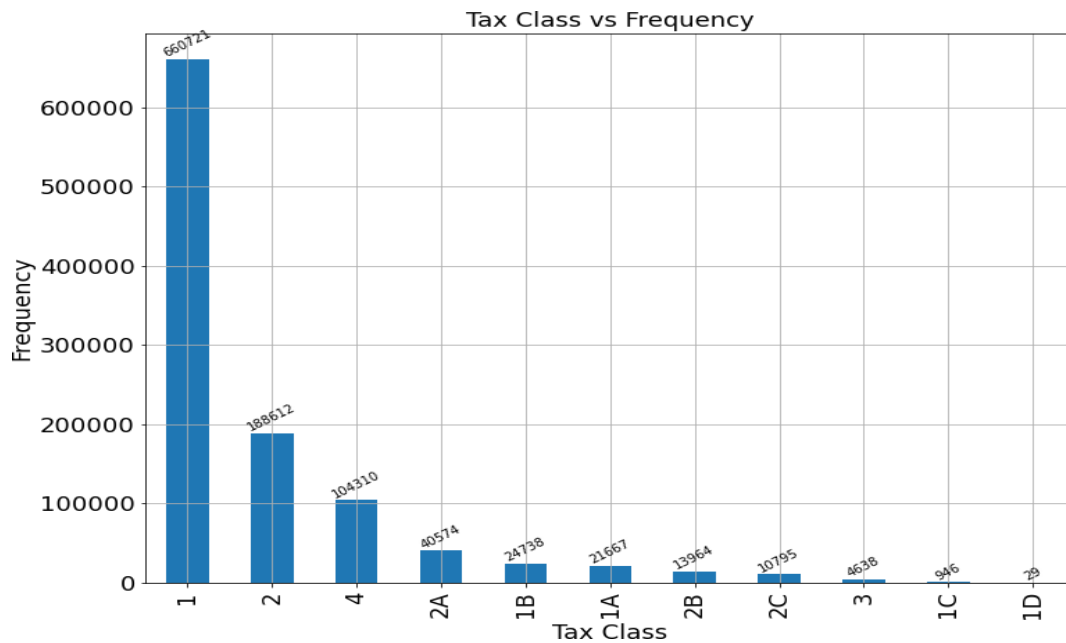


8. **Field Name: BLDGCL**  
**Description:** This field indicates the building class, bar chart shows the count of the top 10 building class



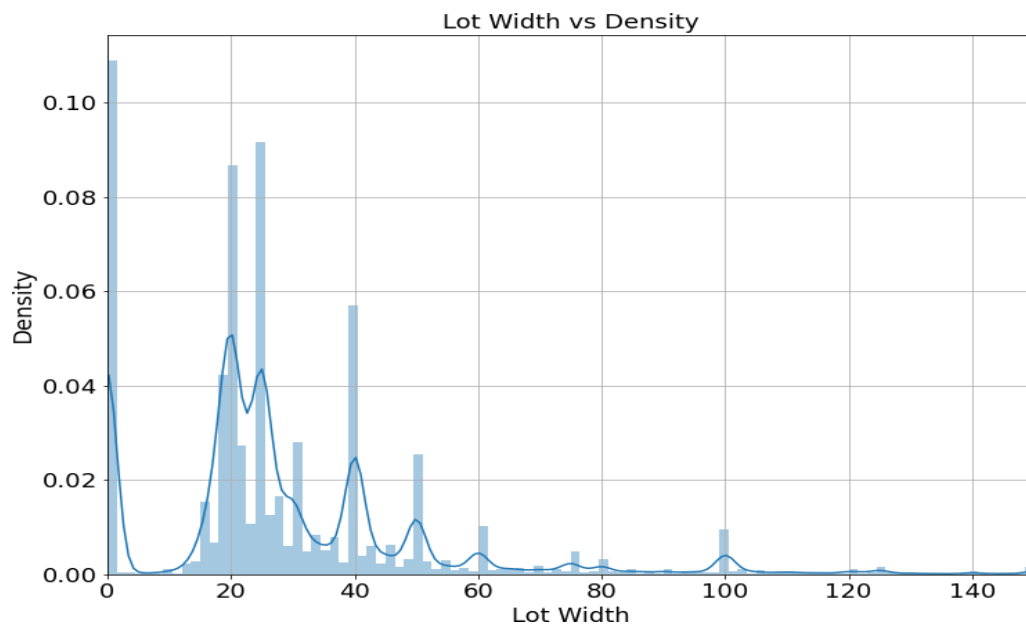
9. **Field name: TAXCLASS**  
**Description:** This field represents the tax class of the property  
 1. 1 - 3 Unit Residence

2. Apartments, 2A = 4, 5, or 6 Units
3. Utilities
4. All Others



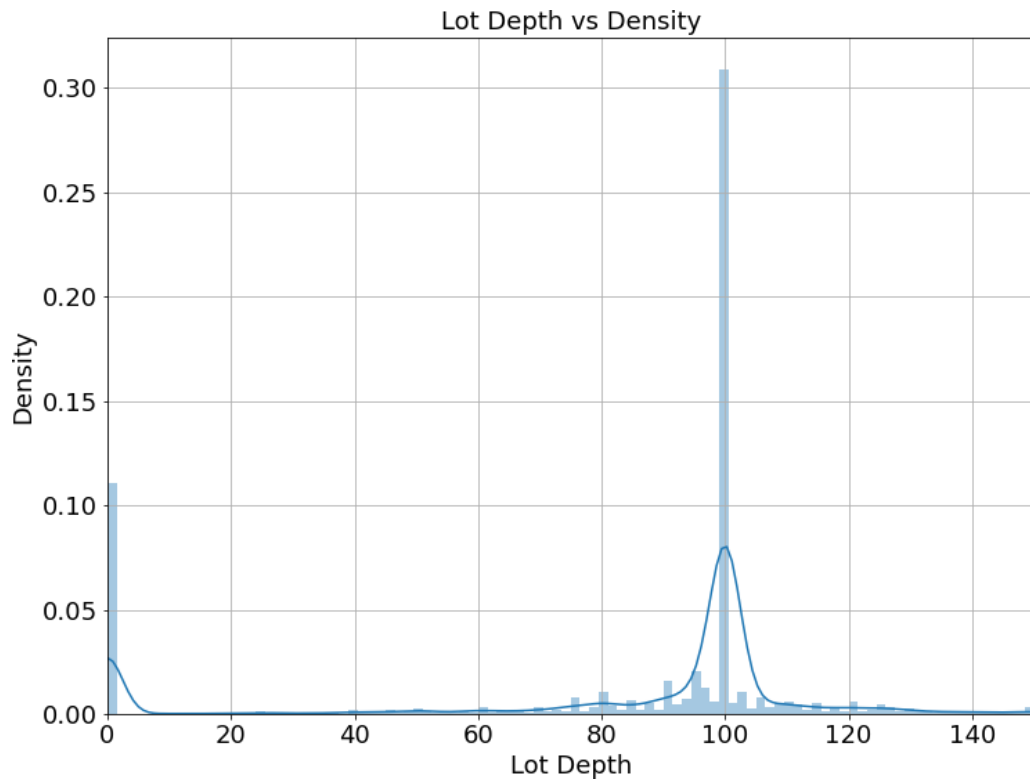
**10. Field Name: LTFRONT**

**Description:** It is the lot width. The below figure uses a logarithmic scale, and the lot width is limited to 150..

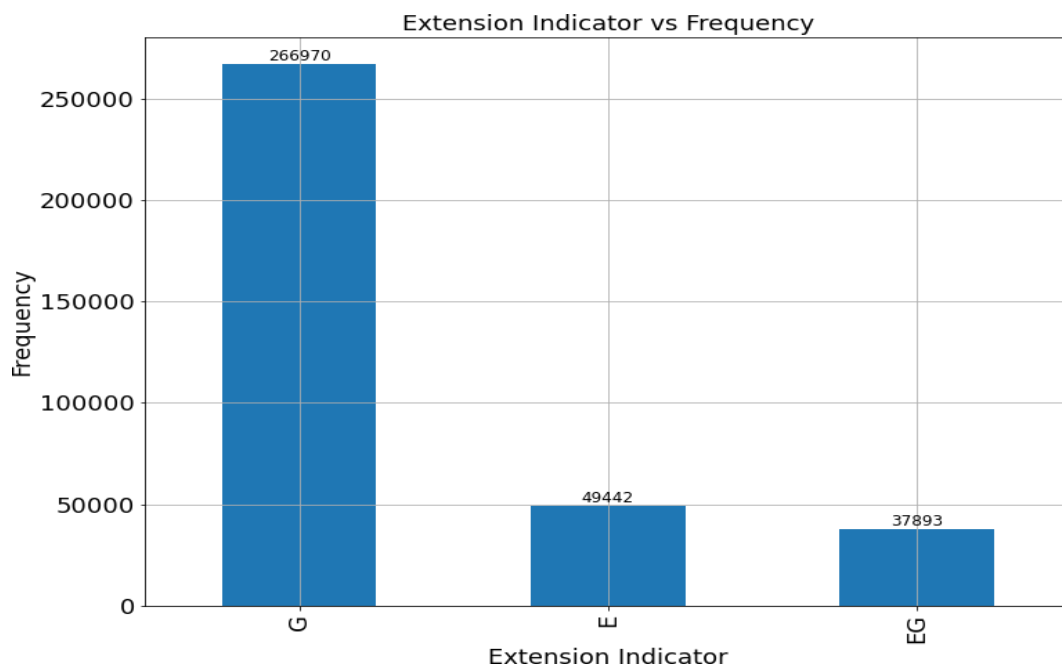


**11. Field name: LTDEPTH: It is the Lot Depth.**

**Description:** The below figure uses a logarithmic scale and the Lot Depth is limited to 150.

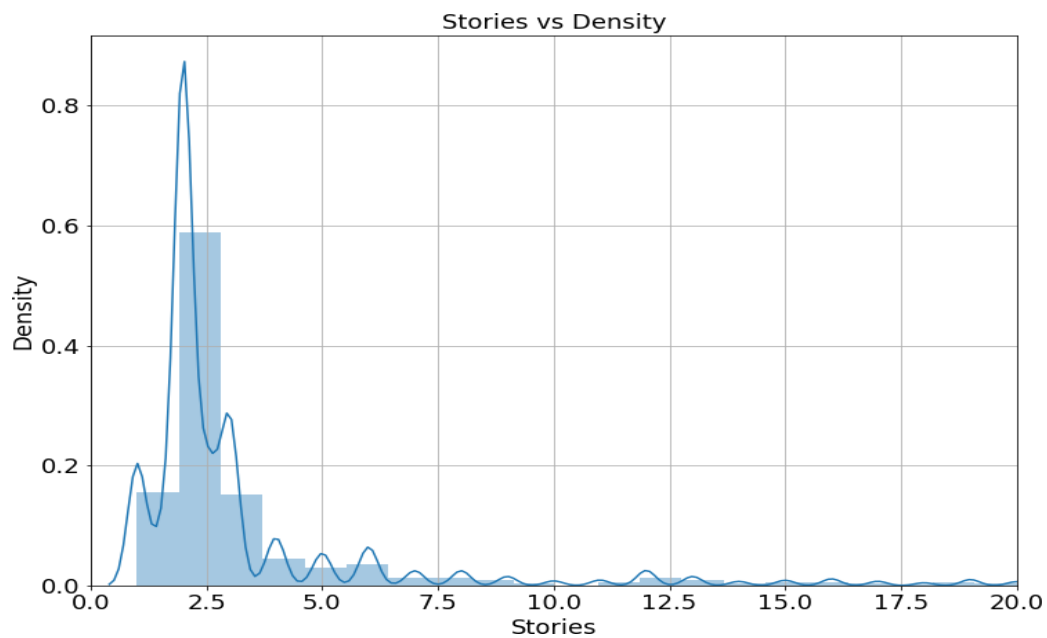


12. **Field Name:** EXT  
**Description:** Extension Indicator



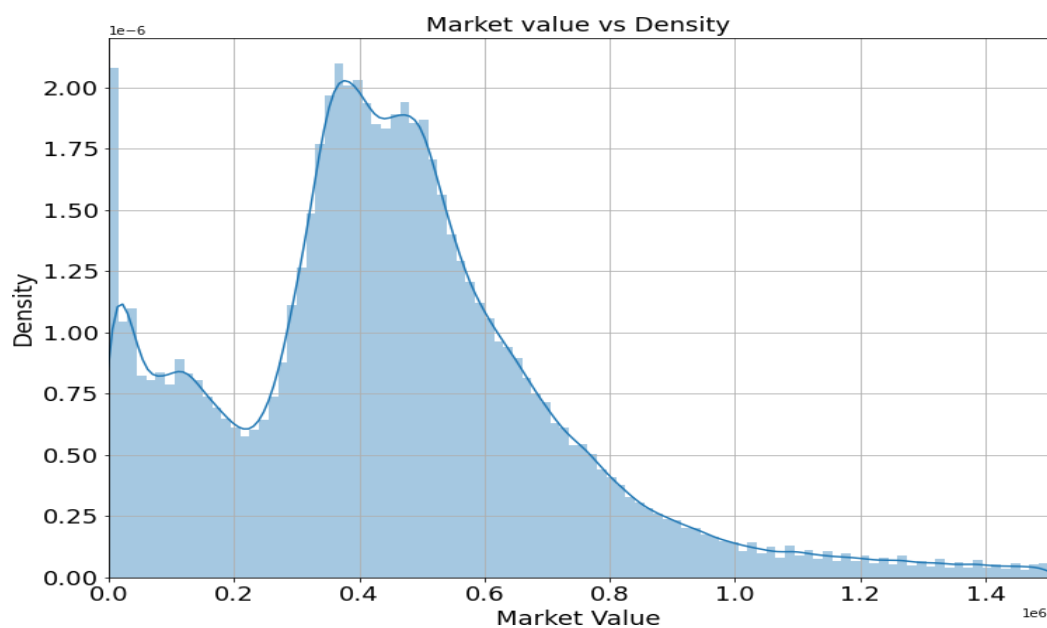
13. **Field Name:** Stories

**Description:** Number of Stories in Building.



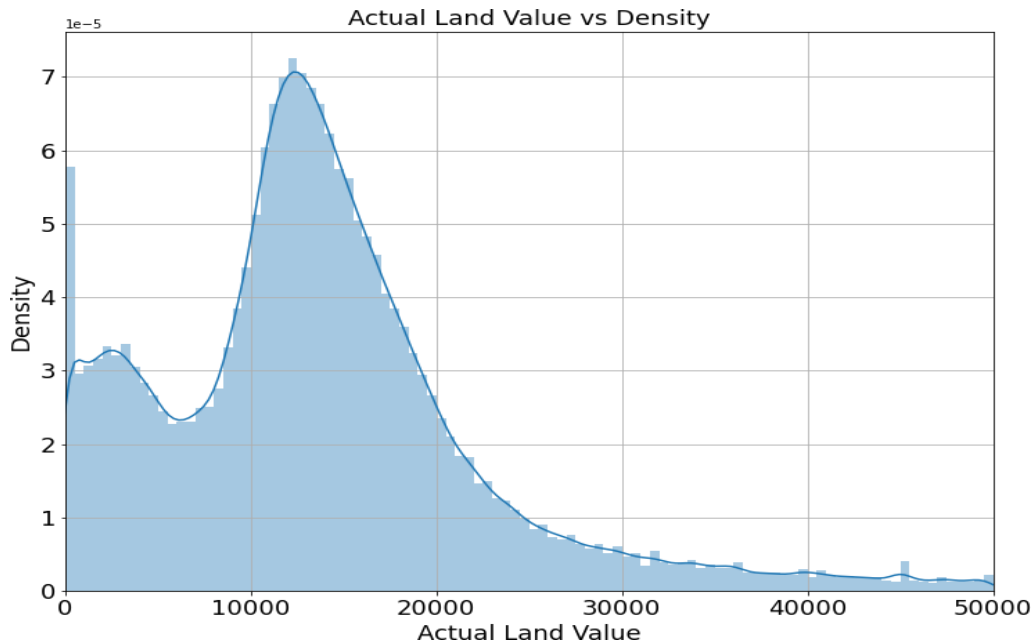
14. **Field Name:** FULLVAL

**Description:** The below figure uses a logarithmic scale and displays market value.



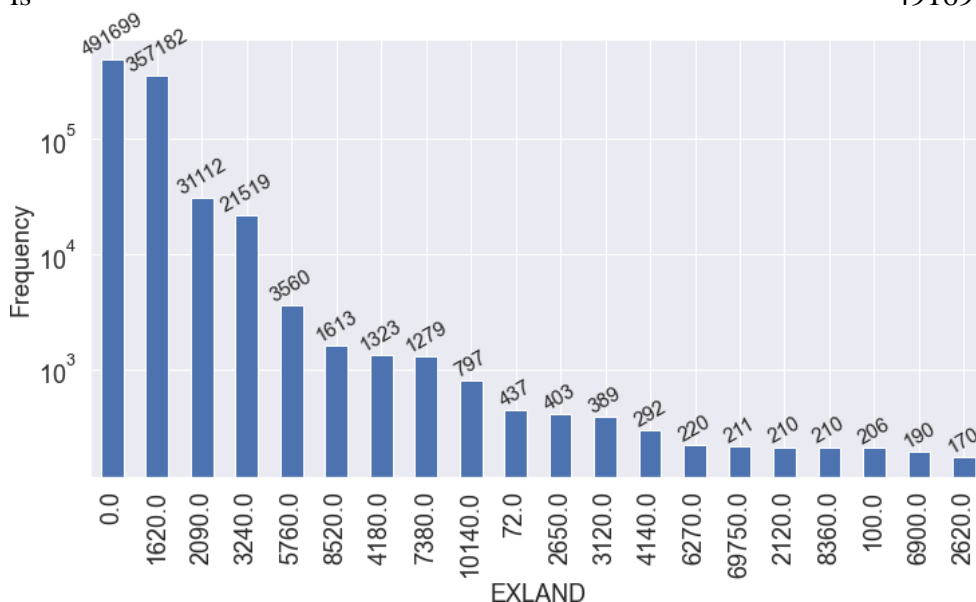
15. **Field Name:** AVLAND Actual Land Value

**Description:** This field represents the actual land value The distribution displays up to a value of 50,000.



16. **Field Name:** EXLAND

**Description:** Merch Zip: The field “EXLAND” refers to the Actual Exempt Land Value. The distribution indicates the top 20 most common Actual Exempt Land Value. The most common value of the EXTOT is 0.0, the count is 491699.

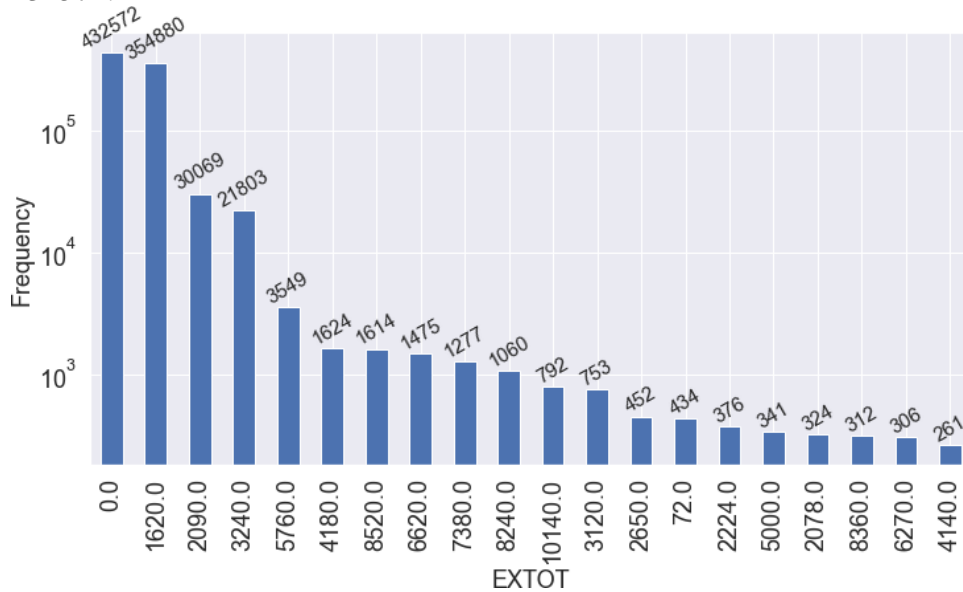


17. **Field Name:** EXTOT

**Description:** Merch Zip: The field “EXTOT” refers to the Actual Exempt Land Total. The distribution indicates the top 20 most common Actual Exempt Land Total used. The most common value of the EXTOT is 0.0, the

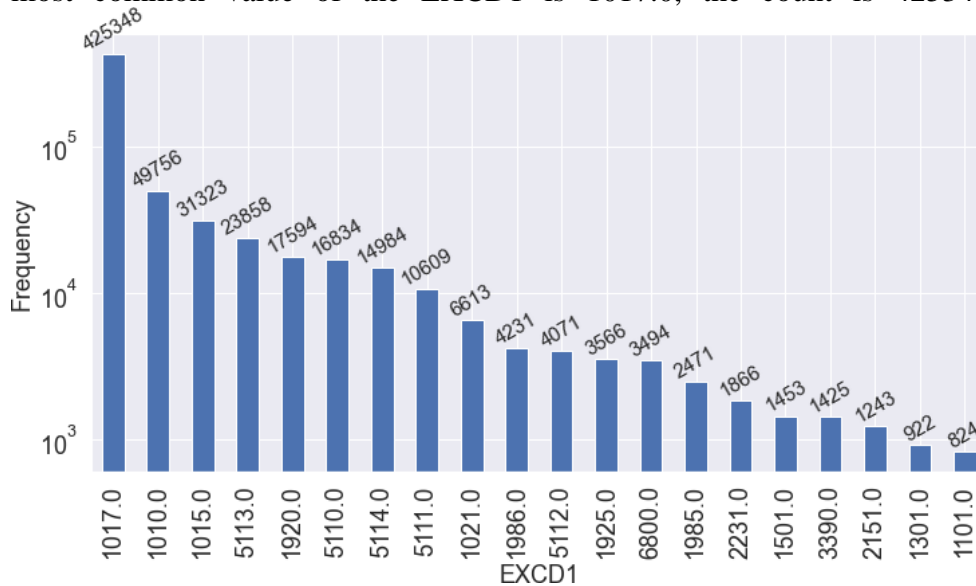
count  
432572.

is



18. **Field Name:** EXCD1

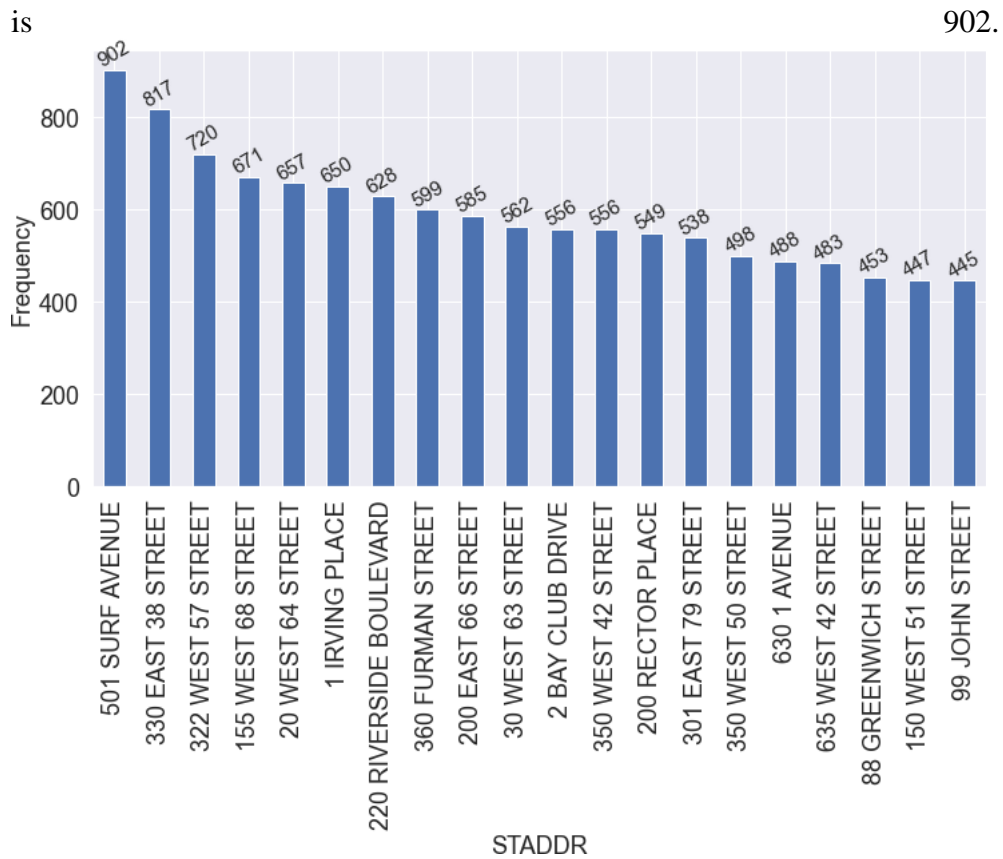
**Description:** The field “EXCD1” refers to the Exemption Code 1. The distribution indicates the top 20 most common Exemption Code 1 used. The most common value of the EXCD1 is 1017.0, the count is 425348.



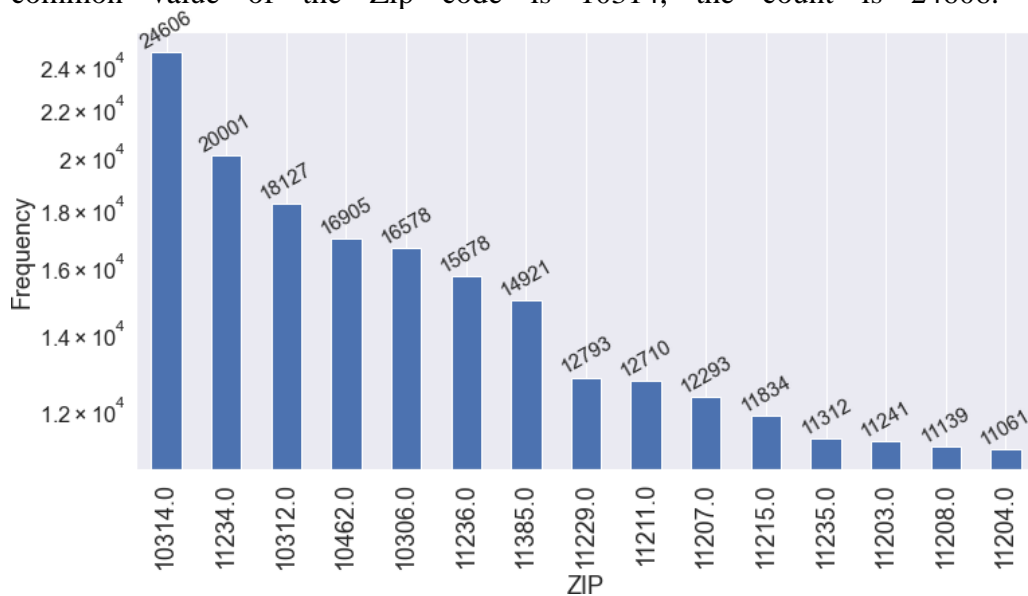
19. **Field Name:** STADDR

**Description:** The field “STADDR” refers to the Street Address of the property. The distribution indicates the top 20 most common STARDDR used. The most common value of the STADDR is 501 Surf Avenue, the count



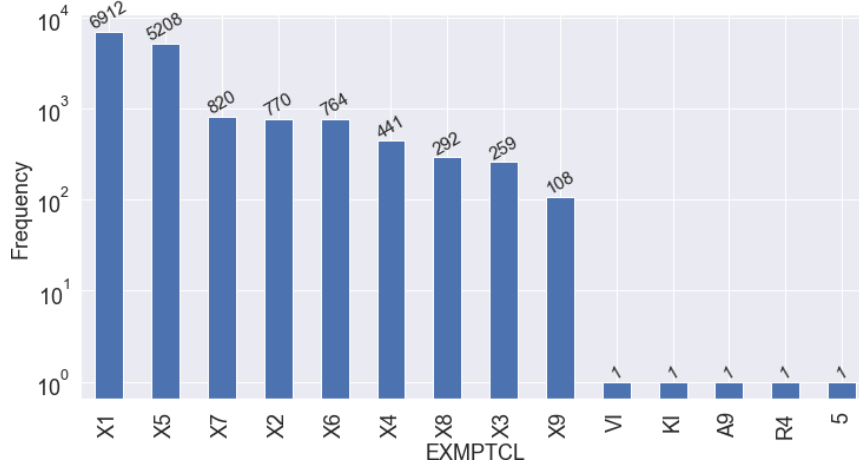


20. **Field Name:** ZIP  
**Description:** The field “ZIP” refers to the Zip code of the property. The distribution indicates the top 15 most common Zip code used. The most common value of the Zip code is 10314, the count is 24606.



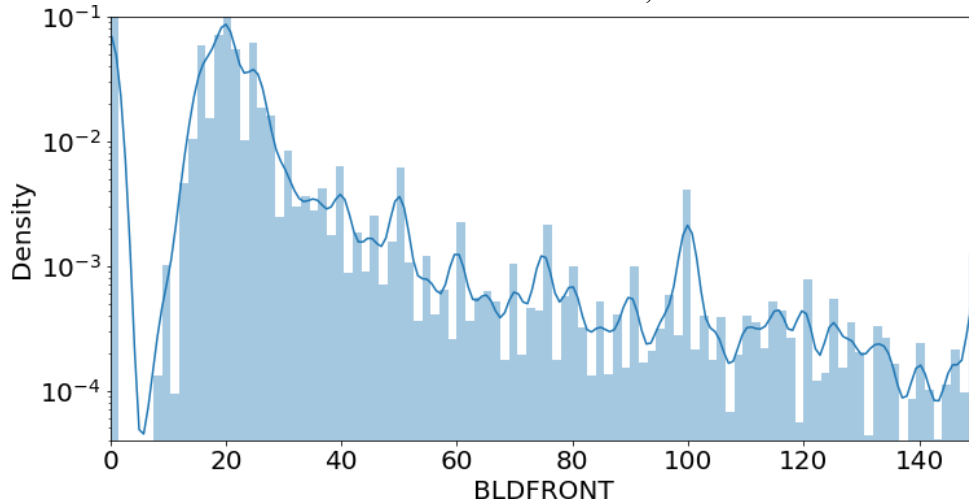
21. **Field Name:** EXMPTCL

**Description:** Merch Zip: The field “EXMPTCL” refers to the Exemption Class. The distribution indicates the 14 EXMPTCL used. The most common value of the EXMPTCL is X1, the count is 6912.



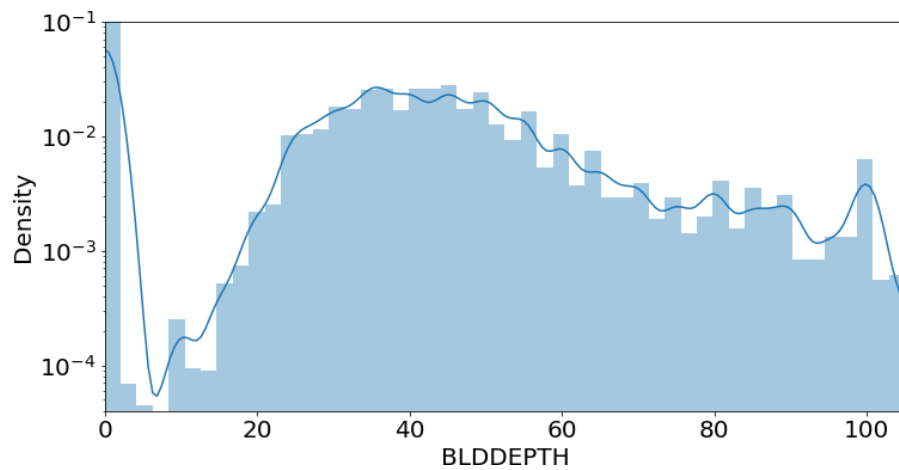
22. **Field Name:** BLDFRONT

**Description:** The field “BLDFRONT” refers to the building Width of the property. The distribution displays up to a Building Width of 150. The most common value of the BLDFRONT” is 0, the count is 228815.

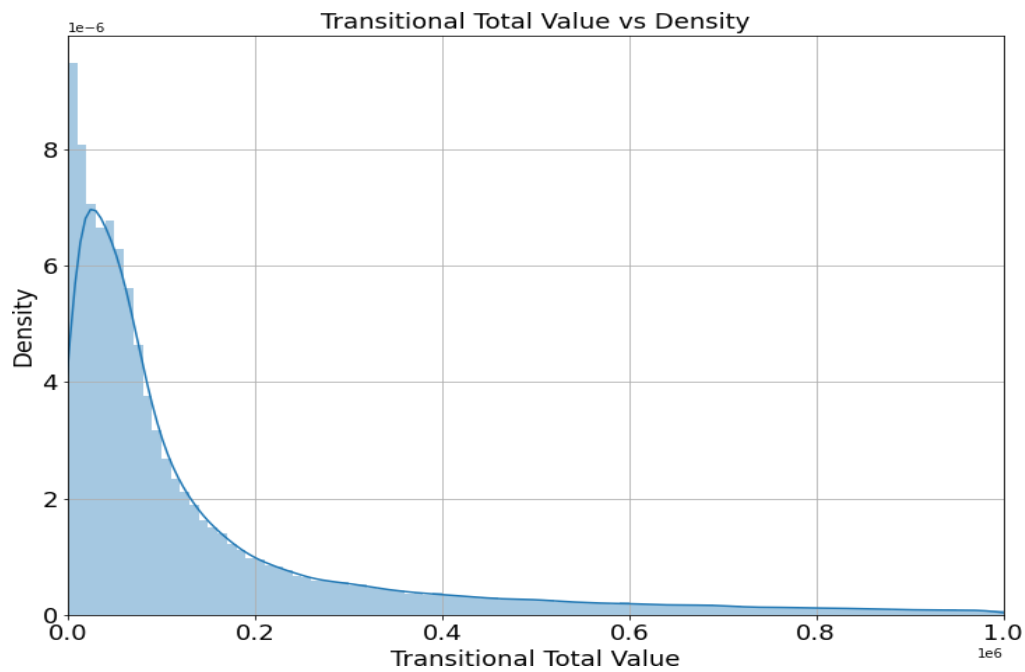


23. **Field Name:** BLDDEPTH

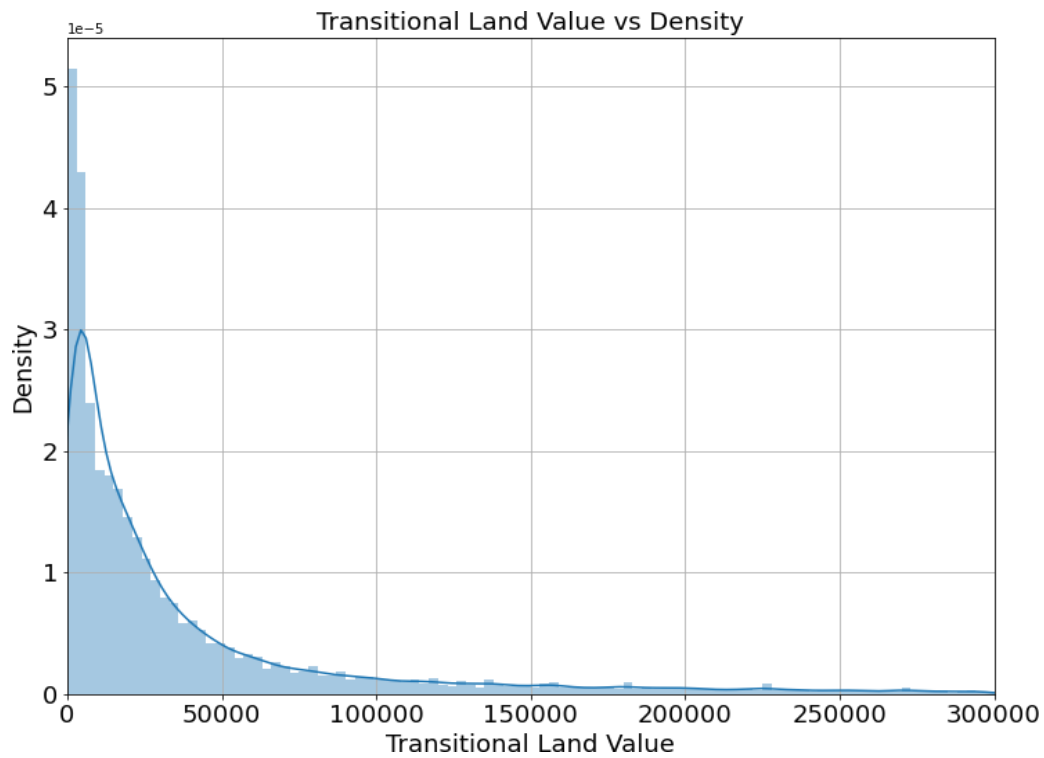
**Description:** The field “BLDDEPTH” refers to the building depth. The distribution displays Density up to a Building Depth of 100. The most common value of the Zip code is 0, the count is 228853.



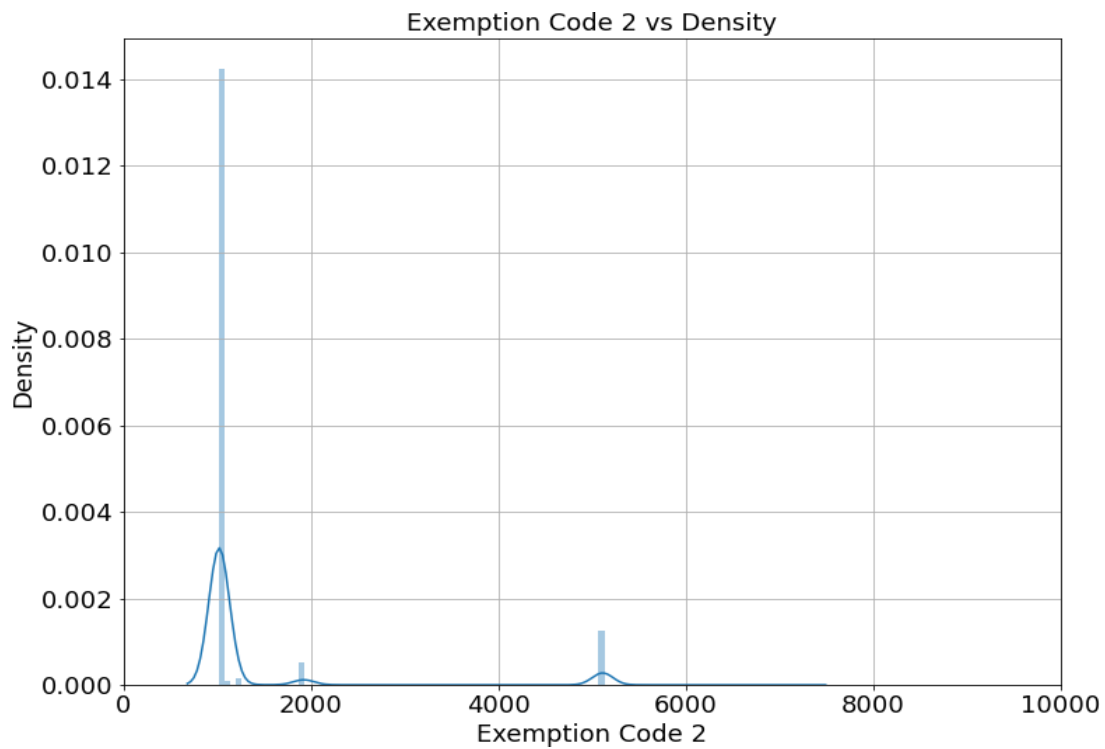
24. **Field Name:** AVLAND2  
**Description:** This field represents the transactional land value



25. **Field Name:** AVTOT2  
**Description:** This field represents the Transitional Total Value

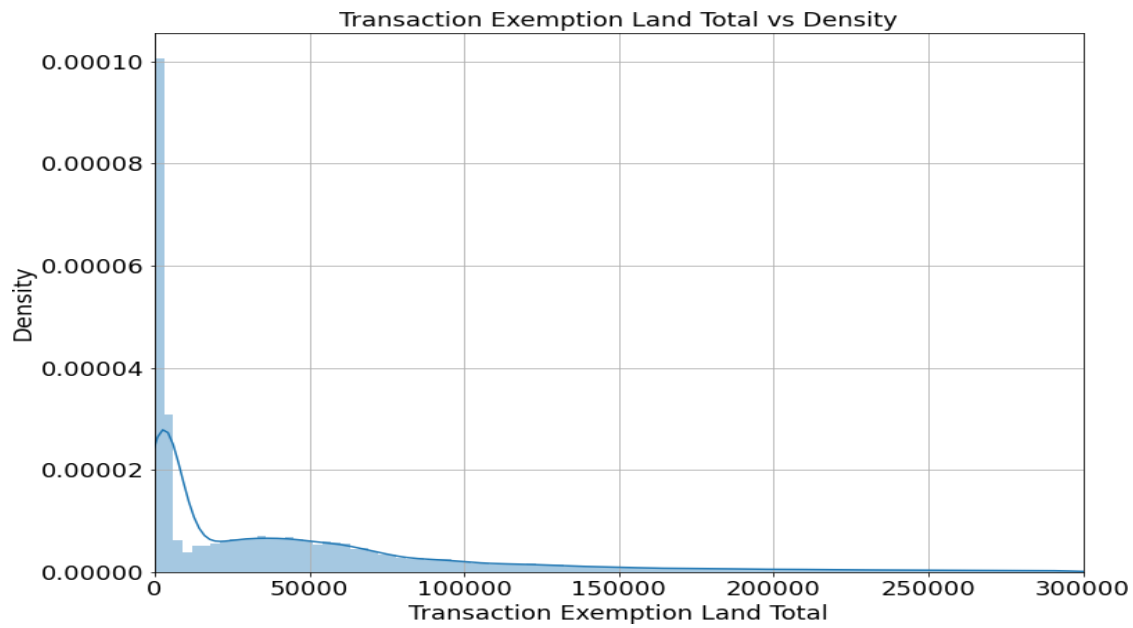


26. **Field Name:** EXLAND2  
**Description:** This field represents the Transitional Exemption Land Value.

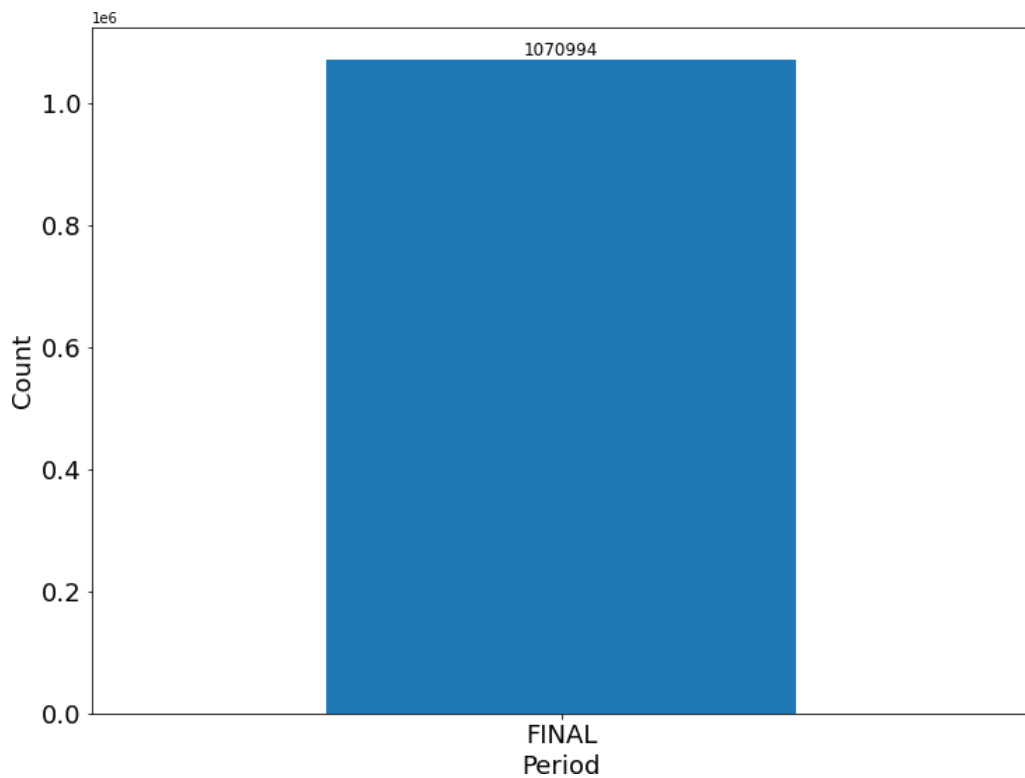


27. **Field Name:** EXTOT2

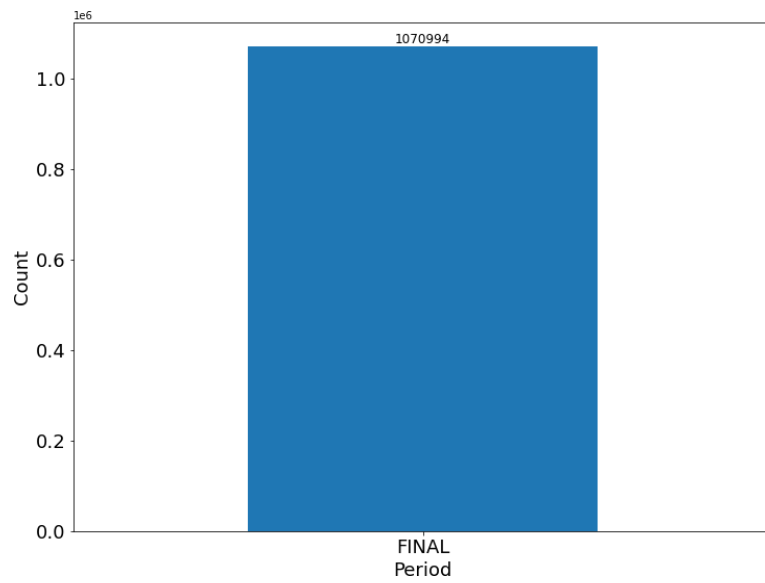
**Description:** This field represents the Transitional Exemption Land Total.



28. **Field Name:** PERIOD  
**Description:** Has only 1 Value



29. **Field Name:** YEAR  
**Description:** Has only 1 value.



**30. Field Name : VALTYPE**

**Description :** Has only one Value.

