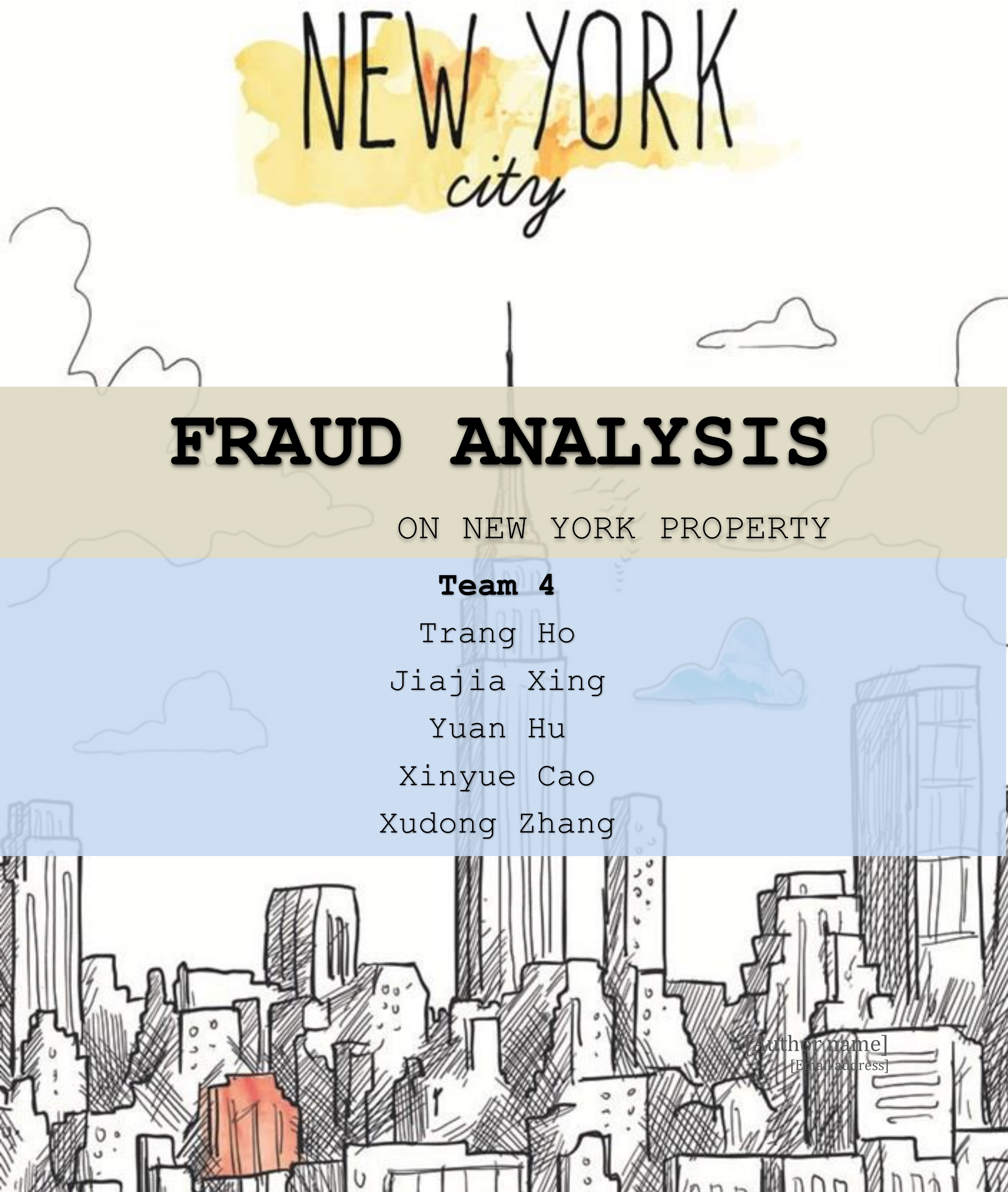


FRAUD ANALYSIS

ON NEW YORK PROPERTY

Team 4

Trang Ho

Jiajia Xing

Yuan Hu

Xinyue Cao

Xudong Zhang

[Author name]
[Email address]

TABLE OF CONTENT

1. Executive Summary

This report provides an analysis and evaluation of The City of New York Property Valuation and Assessment Data ("NY Property Data") for the purpose of detecting potential fraud using unsupervised machine learning methods. The dataset, prepared by the NYC Department of Finance, can be downloaded directly from the NYC OpenData Website.

The general flow of our process is from data description, data cleaning and data preparation of expert variables, standardization and dimensionality reduction, application of fraud algorithm, calculation of fraud scores, and finally identification of potential fraud and significant results.

We started with evaluation of the dataset's quality. The results of this procedure are included in the Data Quality Report in the Appendix. We first selected the important variables that would be used in our creation of expert variables and our construction of fraud algorithms. We concluded that FULLVAL, AVTOT and AVLAND would be the three key base variables. There were also other 12 important variables (along with their derivatives) that could be used in our analysis.

Realizing that the dataset contains many missing and zero values, which will not provide meaningful insights for the information recorded, we decide to fill those values with the average or reasonable values based on our judgements. In this data preparation stage, we also generated 58 expert variables based on the previously identified important variables.

We used Z-scale to normalize the data before performing PCA. We picked the Top 10 principal components that explained 93.9 percent of the overall variance. In building fraud algorithm, we used Mahalanobis distance as one measurement of fraud score. In addition, AutoEncoder was also applied to this case, in which we summed the all the reconstruction errors for each data point as its second fraud score. The overlap percentage between these two algorithm results is about 86 percent. We then used quantile binning to scale these two sets of scores and select the overlap Top 0.1 percent record for further investigation. Besides, we also implemented individual analysis of the Top 10 records.

The result showed that the most potentially fraudulent records have higher total market value, assessed total value and assessed land value compared to the other records in the dataset. We also came to the conclusion that the properties with highest fraud scores are mostly non-residential entities and the large proportion of suspicious properties are located in Staten Islands.

For future improvement of the results, we suggest that the use of other algorithms such as k-Means Clustering should be included to enhance the results of the analysis.

2. Data Description

2.1 Summary of Dataset

File name: NY Property Data.xlsx
Data source: Link <https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8>
Data size: 1,048,575 records
Number of fields: 30 (16 categorical and 14 numerical variables including the **BORO** field and excluding the first **RECORD** field as it can be considered as index column)
Time: September 2nd, 2011

The *NY Property Data* from the NYC Department of Finance (DOF) is an open government data that can be downloaded directly on *NYC OpenData* website. According to the DOF, the purpose of this data is to calculate property tax, grant eligible properties exemptions and/or abatements. The data is collected and entered into the system by various city employee including property assessors, property exemption specialists, ACRIS reporting, Department of Building reporting, etc. The data provides information for NYC properties such as owners, stories, zip code, building class, tax class, lot depth, lot length, building front, building depth, etc. Important numbers comprising total market value of the property, current assessed land value, and current assessed total value are also reported in this data. *NY Property Data* was created in September 2nd, 2011 and updated recently in December 22nd, 2017. There are a total of 1,048,575 records given with 30 categorical and numerical fields.

Following are some descriptions of the variables we consider to be the most important and the reasons we choose those variables. The complete Data Quality Report can be found in Appendix.

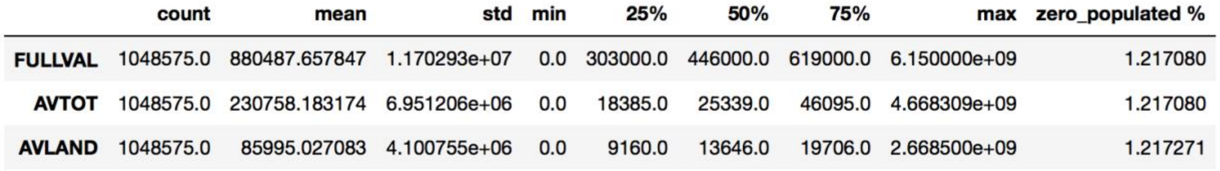
2.2 Important Variables

There are 15 variables in the dataset that we deemed important in our analysis of potential fraud in *NY Property Data*. We will use 3 variables (**FULLVAL**, **AVTOT**, **AVLAND**) as key base variables and the other 12 variables (with their derivatives) as complement variables to derive our expert variables.

2.2.1 Key Base Variables

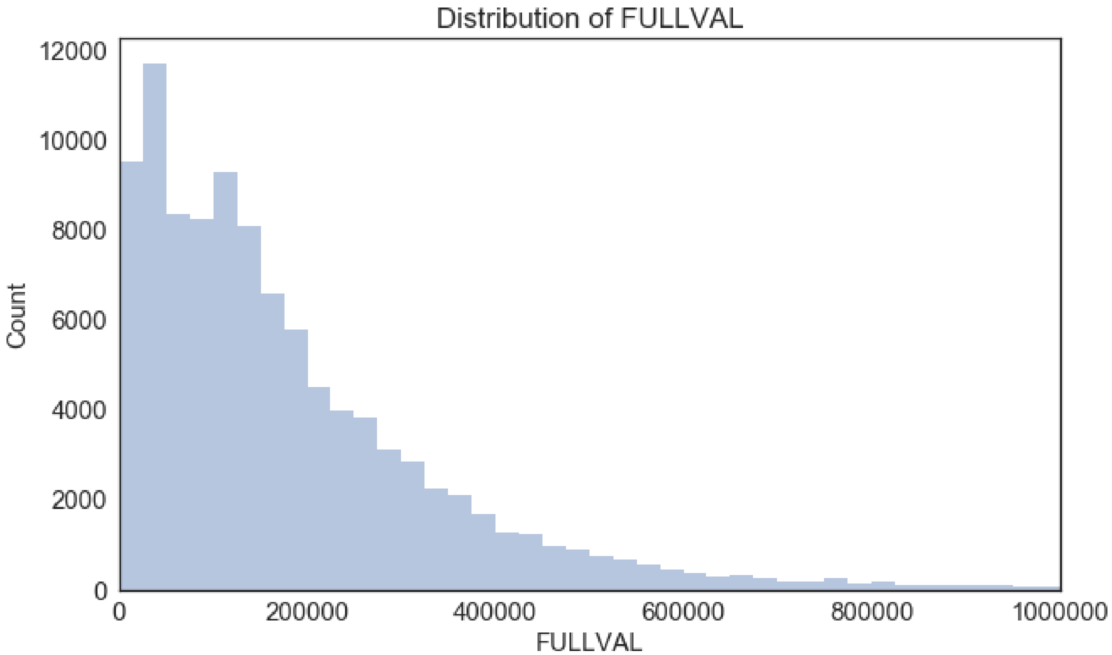
We decided to use the three monetary variables including **FULLVAL**, **AVTOT**, and **AVLAND** as key base variables to create competitive expert variables and build fraud algorithms for the *NY Property Data*.

These three variables are all numerical variables. Below is a summary of statistics for the three variables.

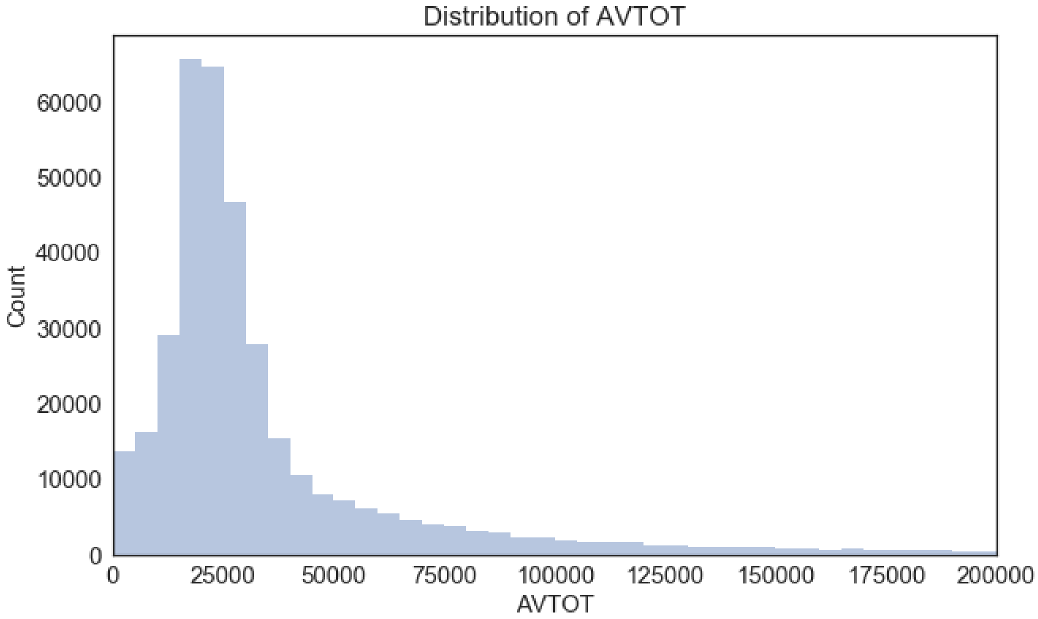


	count	mean	std	min	25%	50%	75%	max	zero_populated %
FULLVAL	1048575.0	880487.657847	1.170293e+07	0.0	303000.0	446000.0	619000.0	6.150000e+09	1.217080
AVTOT	1048575.0	230758.183174	6.951206e+06	0.0	18385.0	25339.0	46095.0	4.668309e+09	1.217080
AVLAND	1048575.0	85995.027083	4.100755e+06	0.0	9160.0	13646.0	19706.0	2.668500e+09	1.217271

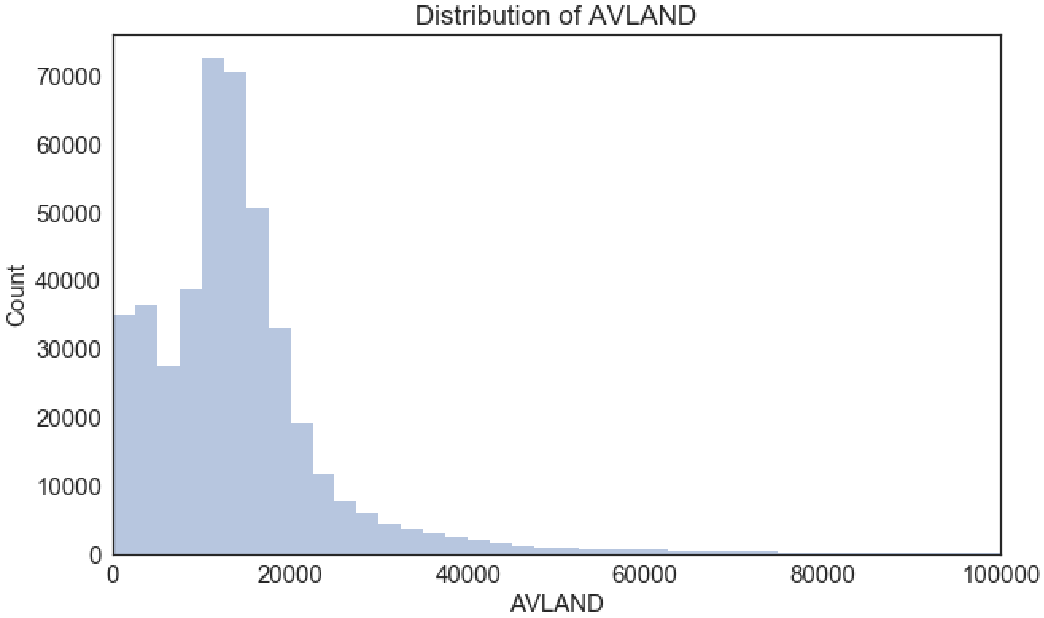
FULLVAL is the total market value (full value) of the property. The NYC Department of Finance stated that the total market value of property is determined by the property's tax class, size and location[1]. This variable seems to have the most meaning based on its relations with other important variables.



AVTOT is the current assessed total value. According to the NYC Department of Finance, this value is calculated depending on the percentage of total market value. This percentage is known as the Level of Assessment or Assessment Ratio, which is based on the property’s tax class. Because this value is derived from the total market value of the property, we decided to take it as a key base variable.



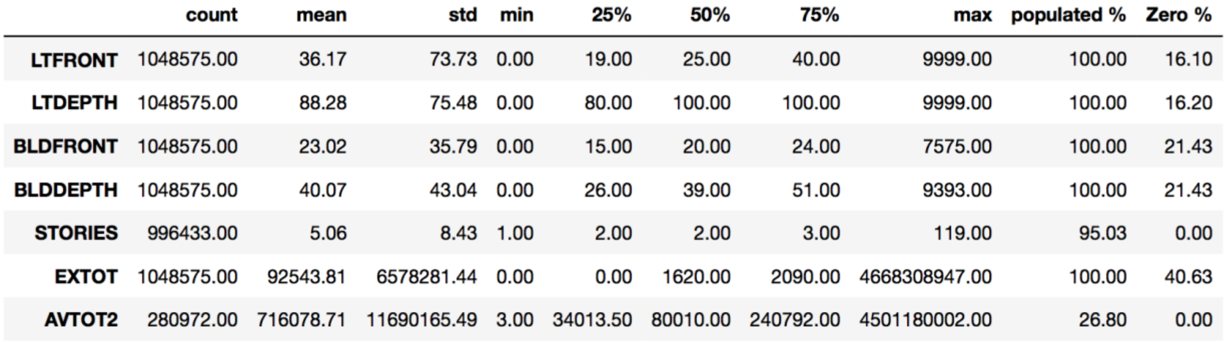
AVLAND is the current assessed land value. Similar to **AVTOT**, this value has a significant correlation with the total market value of property, so we chose it as a key base variable.



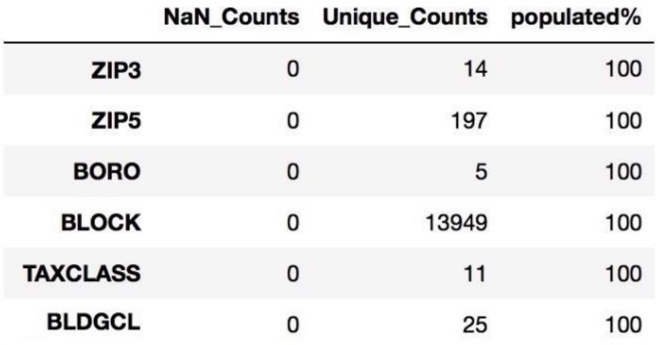
2.2.2. Other Important Variables

Besides the three variables mentioned above, we worked on 12 variables and their derivatives to create expert variables and build fraud algorithms for the NY Property Data. These important variables are **LTFRONT**, **LTDEPTH**, **BLDFRONT**, **BLDDEPTH**, **STORIES**, **ZIP** (**ZIP3** and **ZIP5**), **BBLE** (**BORO**), **BLOCK**, **EXTOT**, **AVTOT2**, **TAXCLASS**, **BLDGCL**. Among these variables, **LTFRONT**, **LTDEPTH**, **BLDFRONT**, **BLDDEPTH**, **STORIES**, **EXTOT**, **AVTOT2** are numerical variables, while the rest are categorical variables.

Below is a summary of statistics for the above variables.

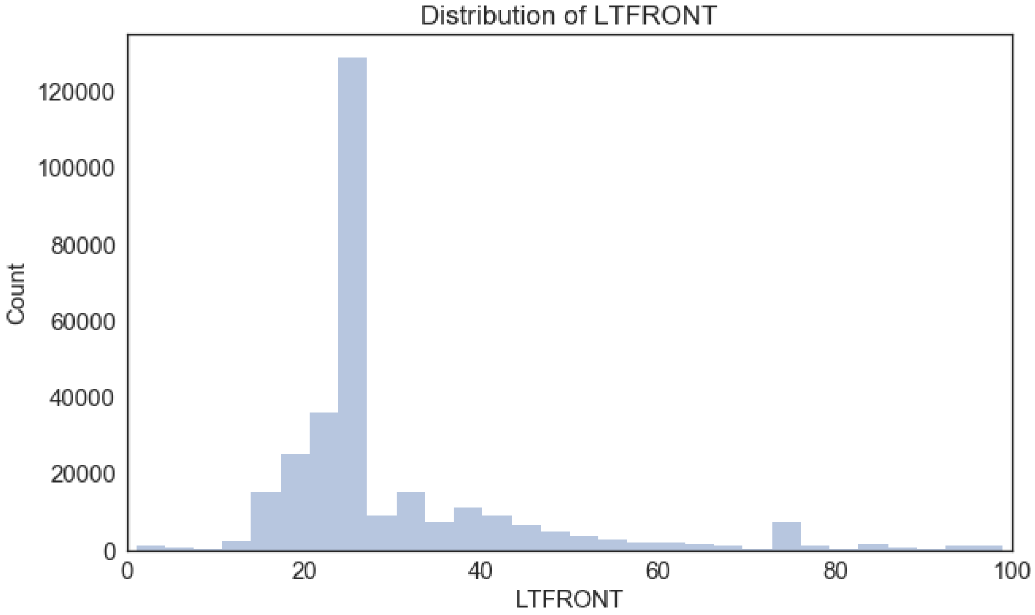


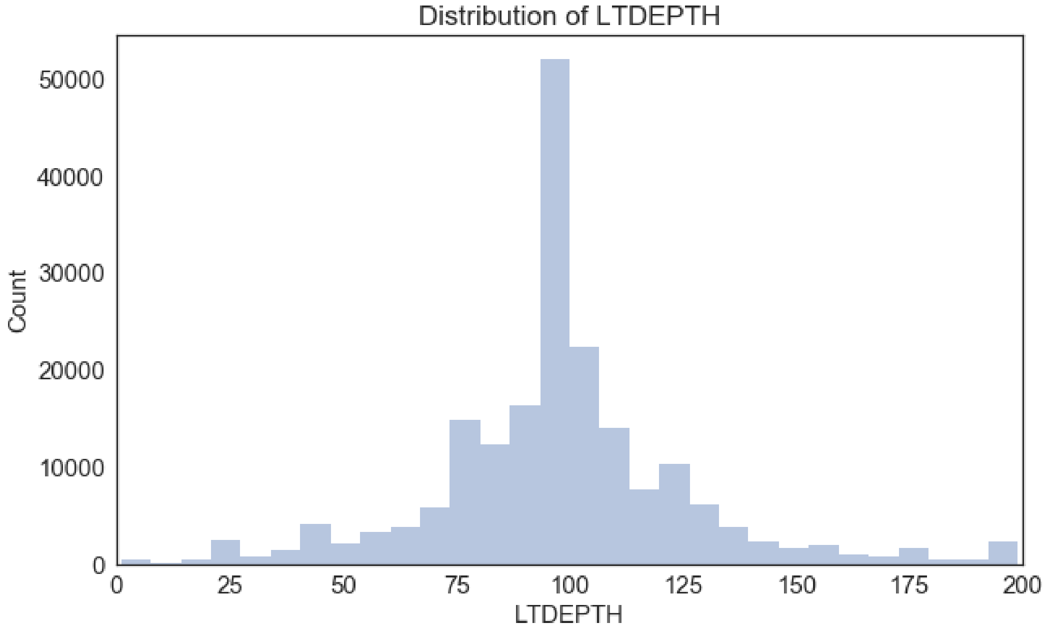
	count	mean	std	min	25%	50%	75%	max	populated %	Zero %
LTFRONT	1048575.00	36.17	73.73	0.00	19.00	25.00	40.00	9999.00	100.00	16.10
LTDEPTH	1048575.00	88.28	75.48	0.00	80.00	100.00	100.00	9999.00	100.00	16.20
BLDFRONT	1048575.00	23.02	35.79	0.00	15.00	20.00	24.00	7575.00	100.00	21.43
BLDDEPTH	1048575.00	40.07	43.04	0.00	26.00	39.00	51.00	9393.00	100.00	21.43
STORIES	996433.00	5.06	8.43	1.00	2.00	2.00	3.00	119.00	95.03	0.00
EXTOT	1048575.00	92543.81	6578281.44	0.00	0.00	1620.00	2090.00	4668308947.00	100.00	40.63
AVTOT2	280972.00	716078.71	11690165.49	3.00	34013.50	80010.00	240792.00	4501180002.00	26.80	0.00



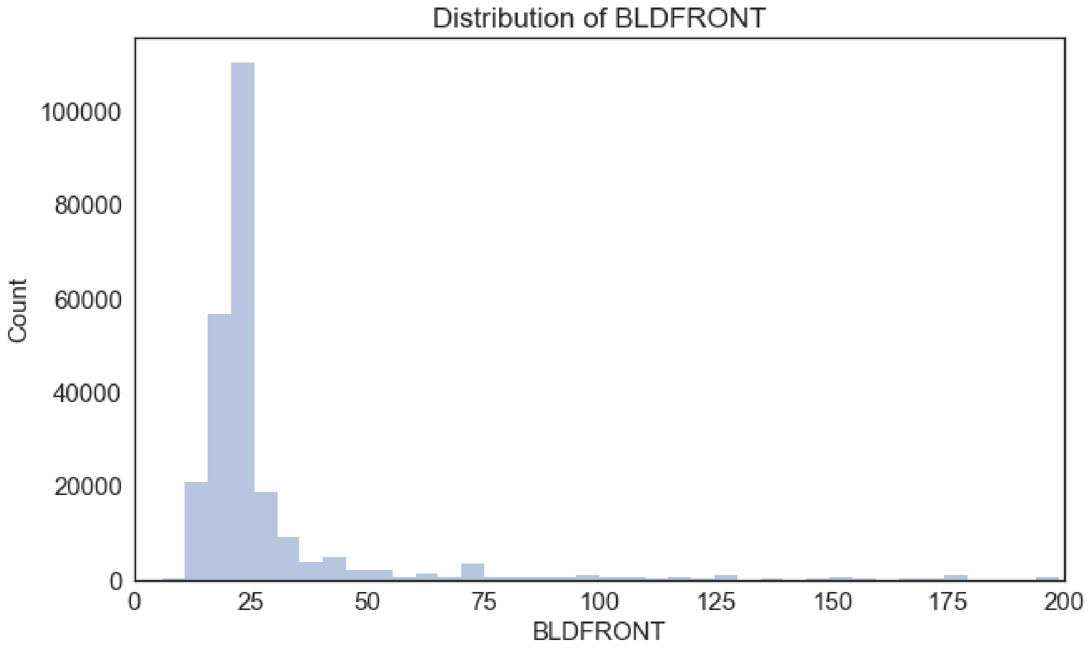
	NaN_Counts	Unique_Counts	populated%
ZIP3	0	14	100
ZIP5	0	197	100
BORO	0	5	100
BLOCK	0	13949	100
TAXCLASS	0	11	100
BLDGCL	0	25	100

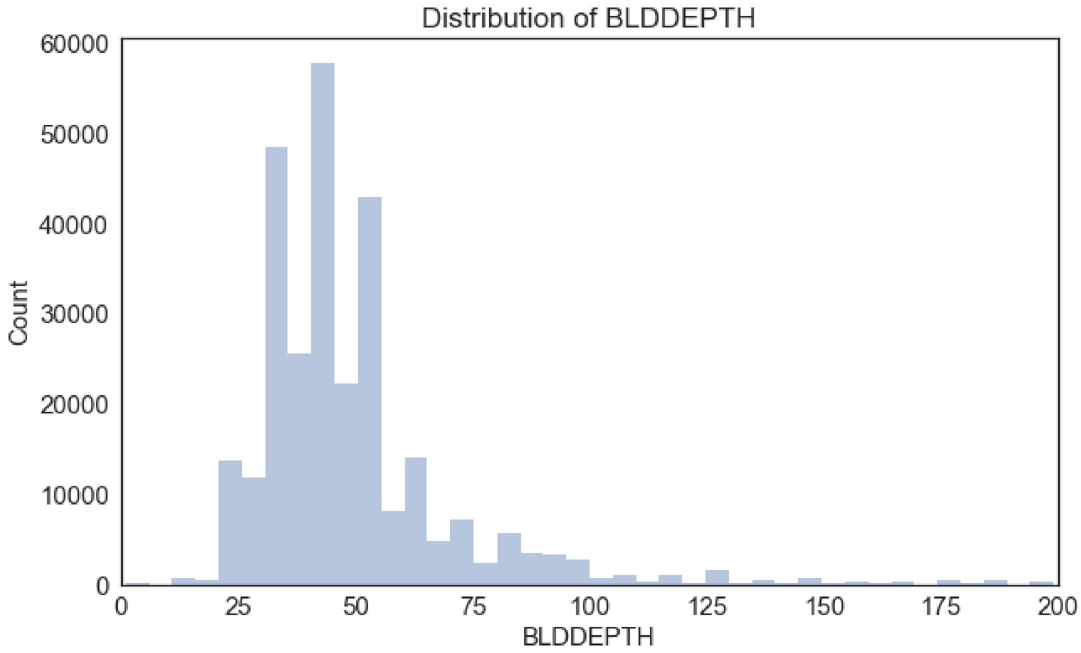
LTFRONT is the lot frontage in feet of a property while **LTDEPTH** is the lot depth in feet of the same property. **LTFRONT** has 1277 unique values ranging from 0 to 9999. No missing values exist. There are 168,867 records of 0 in **LTFRONT**. **LTDEPTH** has 1336 unique values ranging from 0 to 9999. No missing values exist. There are 169,888 records of 0 in **LTDEPTH**. A **LTFRONT** or **LTDEPTH** of 0 may indicate missing value. As these two variables can be multiplied together to determine the size of the lot, which then is meaningful in deriving the total market value of the data, we decided to select these two variables as important.



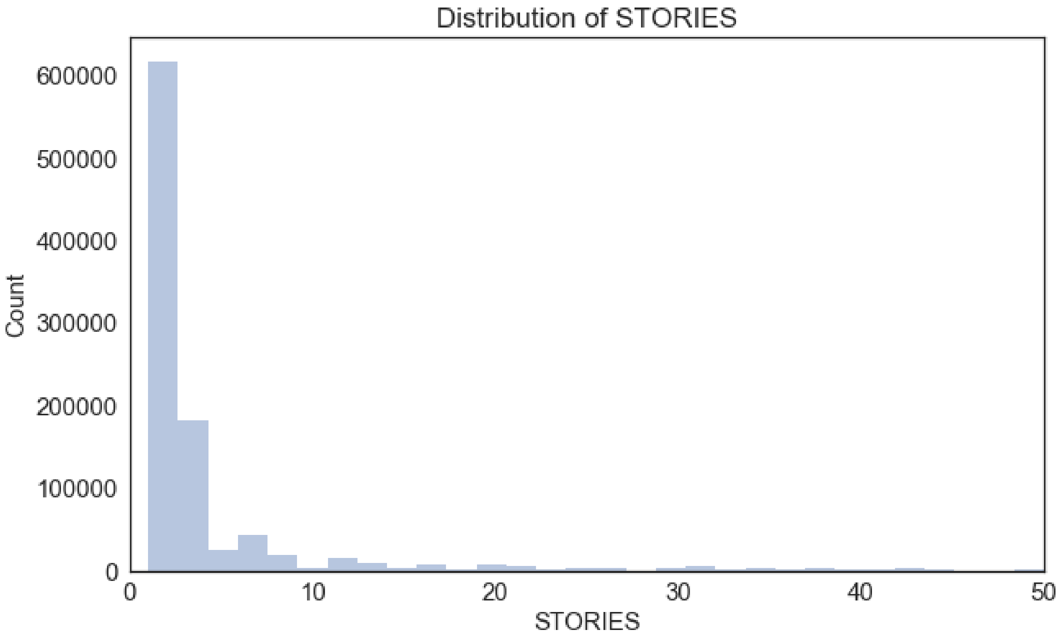


Similar to the two numerical variables above, **BLDFRONT** and **BLDDEPTH** are the building frontage and building depth in feet of a property. The product of these variables give us the total area of the building. **BLDFRONT** has 610 unique values ranging from 0 to 7575. No missing values exist. However, there are 224,661 records of **BLDFRONT** with value 0, which could be in fact missing values. **BLDDEPTH** has 620 unique values ranging from 0 to 9393. No missing values exist. However, there are 224,699 records of **BLDDEPTH** with value 0, which could be in fact missing values.

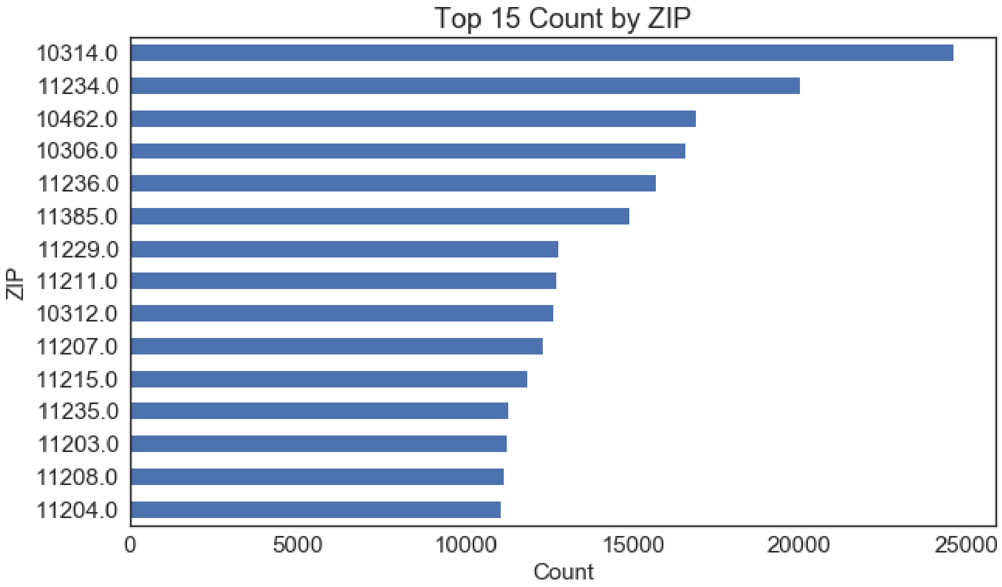


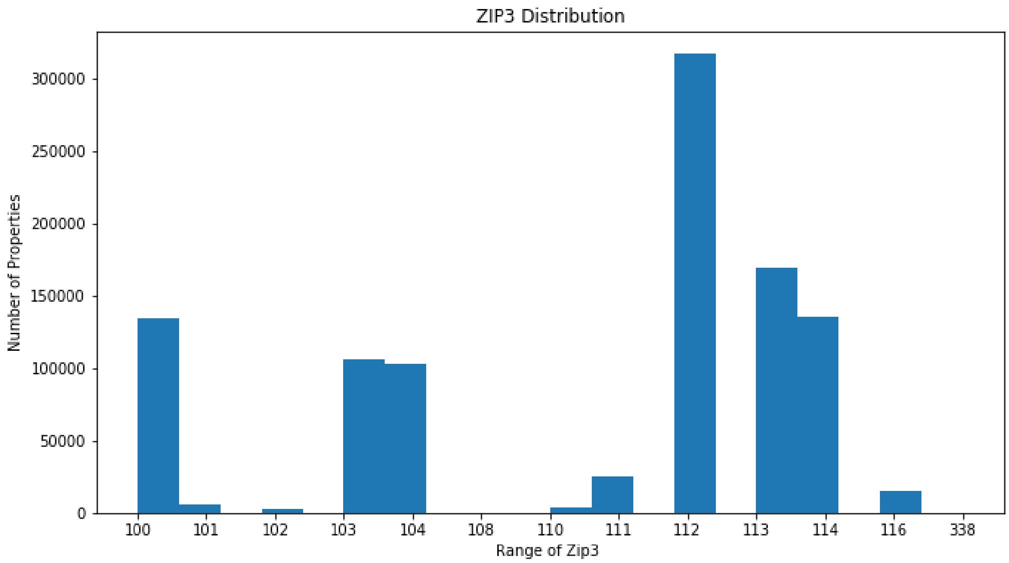


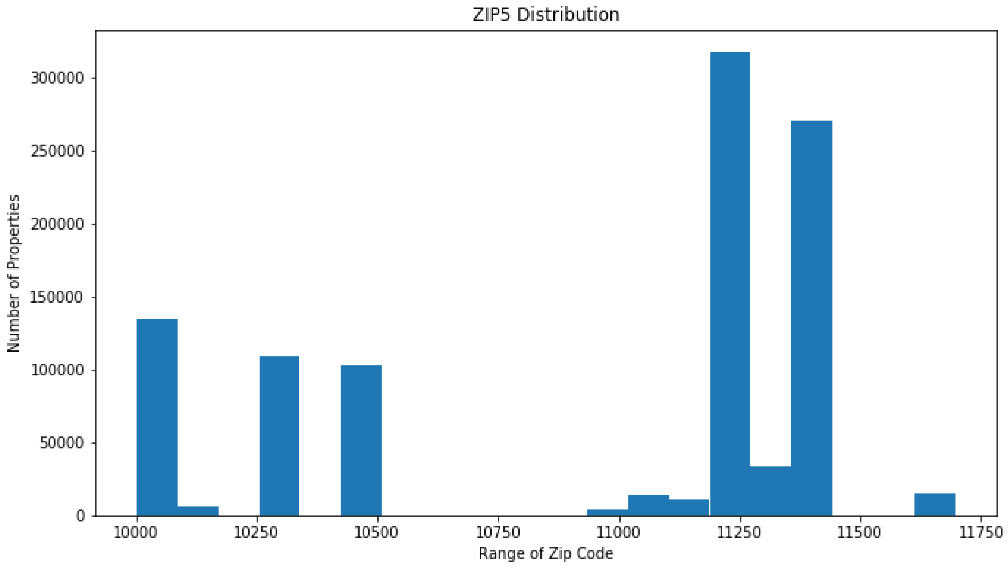
STORIES, which is the number of floors of the building, can be multiplied with the total area of the building to get the building's volume. **STORIES** has 112 unique values ranging from 1 to 119. There are 52,142 missing values in the **STORIES** field. **BLDFRONT**, **BLDDEPTH** and **STORIES** are important variables that connects to the total market value of the property.



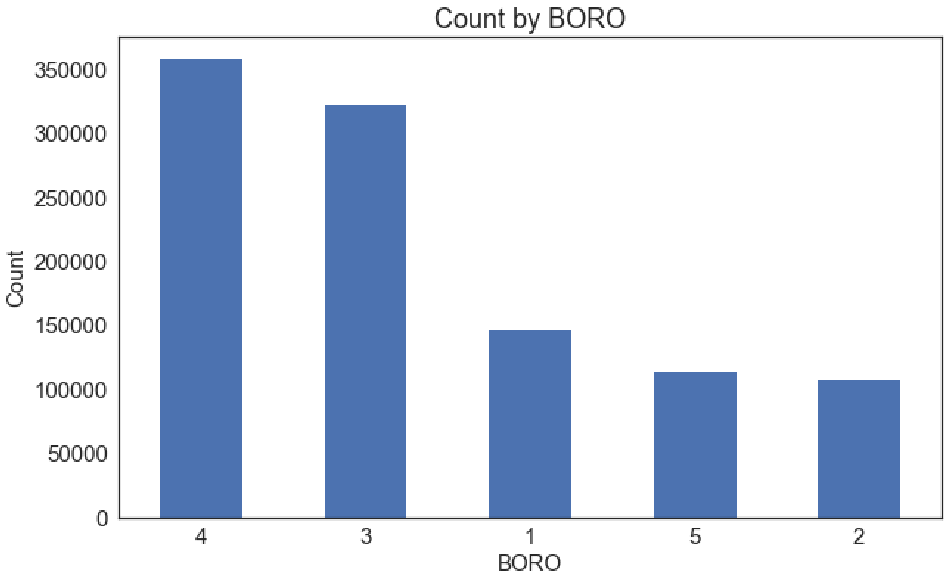
The categorical variables related to location such as **ZIP**, **BBLE** and **BLOCK** also help us determine the total market value of the property, according to the NYC Department of Finance. **ZIP** is the postal zip code of a property. **ZIP** has 197 unique values and 26,356 missing values. There are three obvious anomaly records with **ZIP** of 33803, which should be in Florida. We decided to create derivative variables including the original 5-digit zip codes as **ZIP5** and take the first three digits of the zip codes as **ZIP3** for our grouping purposes later.

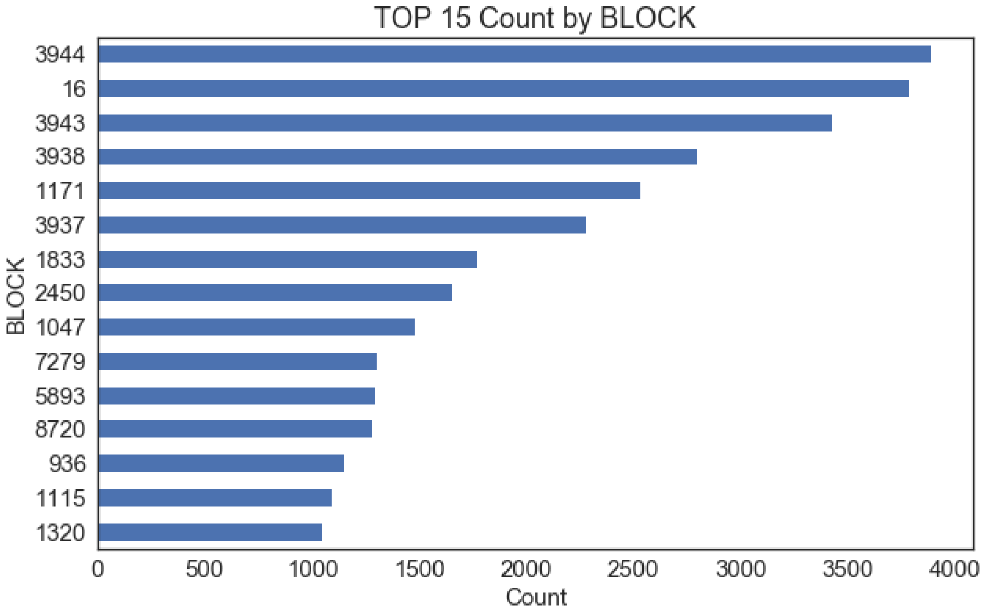




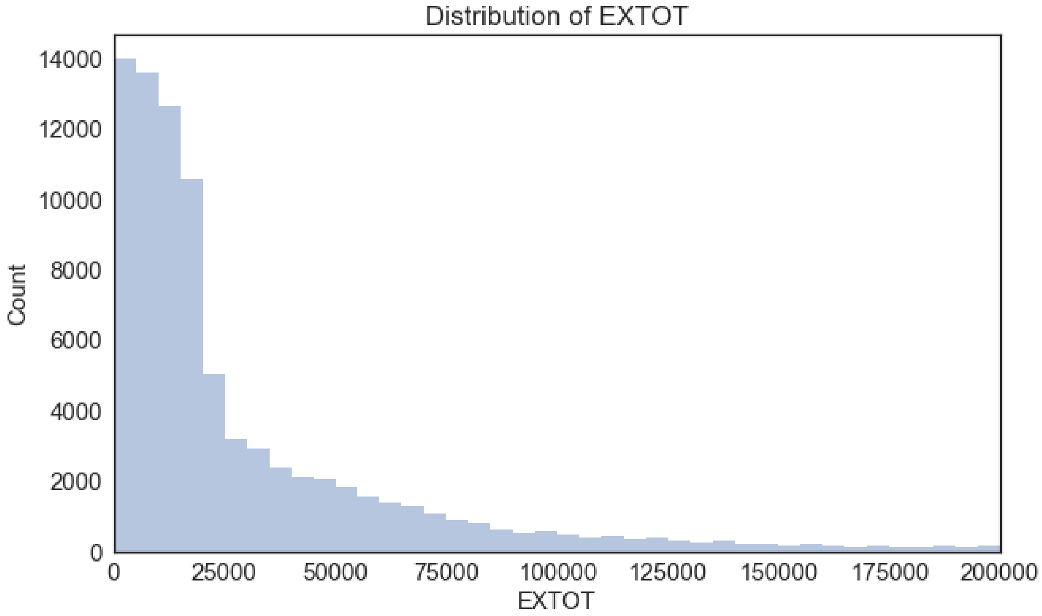


BBLE is just a unique number for a property that concatenate **BORO**, **BLOCK**, **LOT**, and **EASEMENT** fields. However, in order to derive meaningful results from this variable, we decided to extract the first letter of **BBLE** field and only use **BORO**, which is the Borough codes for five areas in NYC including Manhattan, Bronx, Brooklyn, Queens, and Staten Islands. **BLOCK** is the valid block ranges by Borough codes. It has 13949 unique values, ranging from 1 to 16350 with no missing values. The combination of **BORO** and **BLOCK** will allow us to fill in missing data and perform proper grouping in our analysis. Therefore, they are important variables.

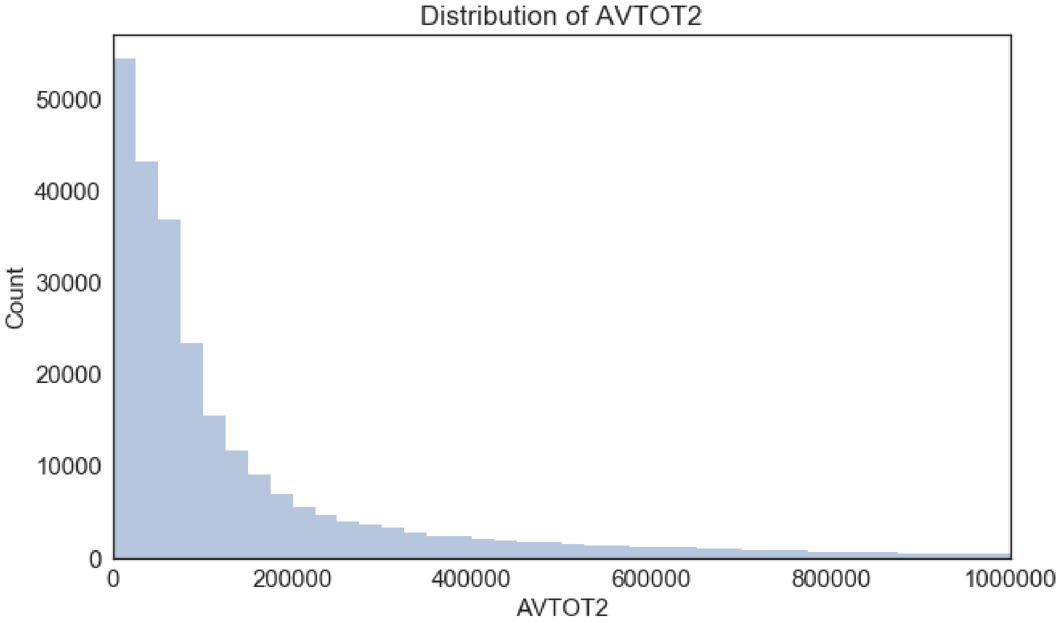




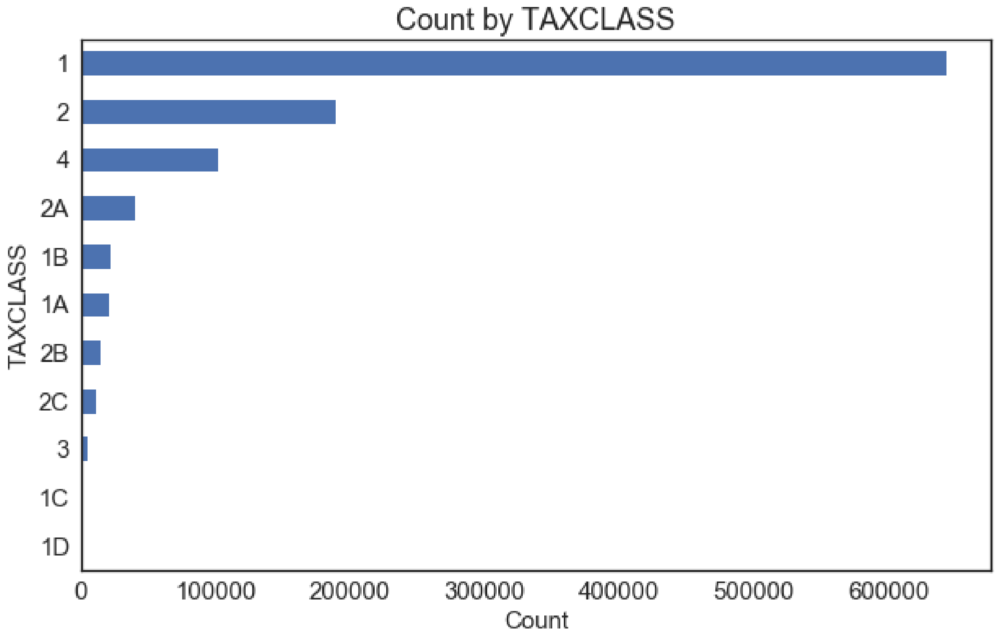
EXTOT is the current actual exempt total value of a property. By its name, the variable shows its strong correlation to the total market value of a property. Therefore, we believe that this variable will provide us insight into some potential fraud in this data relating to exemption.

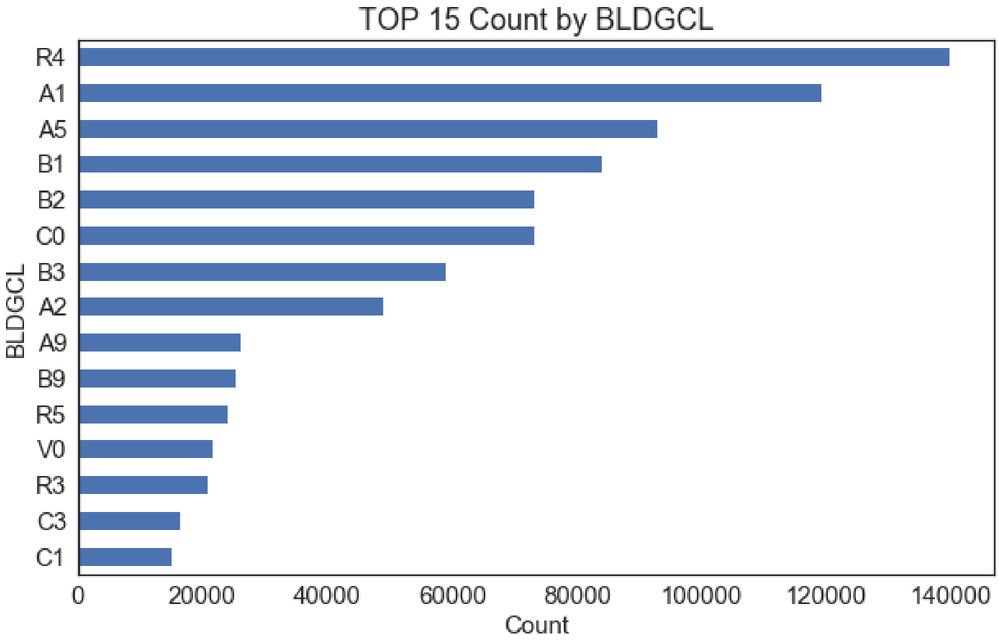


AVTOT2 is the assessed total value of the property. It could be the updated assessed value compared to **AVTOT**, which is the first assessed total value of the property. We believe that the comparison of **AVTOT** and **AVTOT2** will bring us meaningful results.



TAXCLASS and **BLDGCL** are extremely important in deriving total market value of property, according to the NYC Department of Finance[2.] **TAXCLASS** is the current property tax class code according to NY State's classifications. It has 11 unique levels – "1", "1A", "1B", "1C", "1D", "2", "2A", "2B", "2C", "3", and "4". No missing values exist. This variable has a strong correlation with **BLDGCL** field, which is the building class of the property. **BLDGCL** has 200 unique levels. Each level has 2 digits – the first digit is a character from A to Z, the second digit is a number from 0 to 9. To avoid redundant levels in this categorical variable and to deal with rare levels, we can simply use the first digit of **BLDGCL** in the model building process. If the building class is known, the tax class can be generated and vice versa.




TOP 15 Count by BLDGCL
BLDGCL
R4
A1
A5
B1
B2
C0
B3
A2
A9
B9
R5
V0
R3
C3
C1
0
20000
40000
60000
80000
100000
120000
140000
Count

3. Data Preparation

3.1 Imputation of Missing Values and Zero Values

Missing (null) values and zero values in data are common phenomena in real-world data. Handling them is a required step to reduce bias and to produce powerful models during data exploration and preparation. Before creating expert variables, we replaced missing values and zero values with reasonable substituted values.

- For zero values in key base fields **FULLVAL**, **AVTOT**, **AVLAND**, we imputed them with the corresponding *average value of data subset, which includes a range of non-zero values to 97 percentiles values* to minimize the influence of outliers.

FIELDS	MISSING DATA POPULATED	SUBSTITUTED VALUE
FULLVAL	1.217 %	485698.19
AVTOT	1.217 %	55217.61
AVLAND	1.217 %	20478.79

- For zero values in fields **LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH,** we imputed them with the corresponding ***average value of data subset, which includes a range of non-zero values to 97 percentiles values*** again to decrease the impact of outliers.

FIELDS	MISSING DATA POPULATED	SUBSTITUTED VALUE
LTFRONT	16.2%	34.08
LTDEPTH	16.2%	97.77
BLDFRONT	21.4%	23.70
BLDDEPTH	21.4%	45.71

- For field **STORIES,** we filled in the missing values with the ***average STORIES in their own BLDGCL* (Building Class)**.

FIELDS	MISSING DATA POPULATED	SUBSTITUTED VALUE
STORIES	4.97%	ASSIGN MEAN VALUE OF ITS BLDGCL GROUP

- For field **ZIP**, we found that if two properties have identical first 6 digits of **BBLE**, they would share the same zip code. Therefore, we ***assigned the missing zip code of one property with the zip code of another property that have the same first 6 digits of BBLE.***

FIELDS	MISSING DATA POPULATED	SUBSTITUTED VALUE
ZIP	2.51 %	ASSIGN VALUE FROM THE PREPERTY THAT HAS SAME FITST 6 DIGITS OF BBLE

- For field **AVTOT2**, the missing values indicate those properties haven't gone through the second assessment, so we ***replaced missing value with AVTOT of the same record***.

FIELDS	MISSING DATA POPULATED	SUBSTITUTED VALUE
AVTOT2	73.2 %	ASSIGN VALUE OF AVTOT FROM THE SAME RECORD

3.2 Expert Variables

To construct the fraud detection model, we set 3 monetary variables (**FULLVAL**, **AVTOT** and **AVLAND**) scaled by 3 levels (Lot Area, Building Area and Building Volume) and then grouped by 5 entities (**BORO**, **TAXCLASS**, **ZIP3**, **ZIP5** and **BLDGCL**). Besides, we created 4 reasonable expert variables. Finally, we got 58 expert variables in total.

Variable	**Description**
VAL_LOT	Total value per unit of lot area
VAL_BLD	Total value per unit of building area
VAL_VOL	Total value per unit of building volume
TOT_LOT	Assessed value per unit of lot area
TOT_BLD	Assessed value per unit of building area
TOT_VOL	Assessed value per unit of building volume
LAND_LOT	Assessed value of land per unit of lot area
LAND_BLD	Assessed value of land per unit of building area
LAND_VOL	Assessed value of land per unit of building volume
VAL_LOT_TAX	Ratio of VAL_LOT and Average VAL_LOT of building grouped by Tax Class
VAL_BLD_TAX	Ratio of VAL_BLD and Average VAL_BLD of building grouped by Tax Class
VAL_VOL_TAX	Ratio of VAL_VOL and Average VAL_VOL of building grouped by Tax Class
TOT_LOT_TAX	Ratio of TOT_LOT and Average TOT_LOT of building grouped by Tax Class
TOT_BLD_TAX	Ratio of TOT_BLD and Average TOT_BLD of building grouped by Tax Class
TOT_VOL_TAX	Ratio of TOT_VOL and Average TOT_VOL of building grouped by Tax Class
LAND_LOT_TAX	Ratio of LAND_LOT and Average LAND_LOT of building grouped by Tax Class
LAND_BLD_TAX	Ratio of LAND_BLD and Average LAND_BLD of building grouped by Tax Class
LAND_VOL_TAX	Ratio of LAND_VOL and Average LAND_VOL of building grouped by Tax Class

VAL_LOT_BLDGCL	Ratio of VAL_LOT and Average VAL_LOT of building grouped by Building Class
VAL_BLD_BLDGCL	Ratio of VAL_BLD and Average VAL_BLD of building grouped by Building Class
VAL_VOL_BLDGCL	Ratio of VAL_VOL and Average VAL_VOL of building grouped by Building Class
TOT_LOT_BLDGCL	Ratio of TOT_LOT and Average TOT_LOT of building grouped by Building Class
TOT_BLD_BLDGCL	Ratio of TOT_BLD and Average TOT_BLD of building grouped by Building Class
TOT_VOL_BLDGCL	Ratio of TOT_VOL and Average TOT_VOL of building grouped by Building Class
$LAND\_LOT\_BLDGCL$	Ratio of LAND_LOT and Average LAND_LOT of building grouped by Building Class
$LAND\_BLD\_BLDGCL$	Ratio of LAND_BLD and Average LAND_BLD of building grouped by Building Class
$LAND\_VOL\_BLDGCL$	Ratio of LAND_VOL and Average LAND_VOL of building grouped by Building Class
VAL_LOT_BORO	Ratio of VAL_LOT and Average VAL_LOT of building grouped by Borough
VAL_BLD_BORO	Ratio of VAL_BLD and Average VAL_BLD of building grouped by Borough
VAL_VOL_BORO	Ratio of VAL_VOL and Average VAL_VOL of building grouped by Borough
TOT_LOT_BORO	Ratio of TOT_LOT and Average TOT_LOT of building grouped by Borough
TOT_BLD_BORO	Ratio of TOT_BLD and Average TOT_BLD of building grouped by Borough
TOT_VOL_BORO	Ratio of TOT_VOL and Average TOT_VOL of building grouped by Borough
$LAND\_LOT\_BORO$	Ratio of LAND_LOT and Average LAND_LOT of building grouped by Borough
$LAND\_BLD\_BORO$	Ratio of LAND_BLD and Average LAND_BLD of building grouped by Borough
$LAND\_VOL\_BORO$	Ratio of LAND_VOL and Average LAND_VOL of building grouped by Borough
VAL_LOT_ZIP5	Ratio of VAL_LOT and Average VAL_LOT of building grouped by Zip Code
VAL_BLD_ZIP5	Ratio of VAL_BLD and Average VAL_BLD of building grouped by Zip Code
VAL_VOL_ZIP5	Ratio of VAL_VOL and Average VAL_VOL of building grouped by Zip Code
TOT_LOT_ZIP5	Ratio of TOT_LOT and Average TOT_LOT of building grouped by Zip Code
TOT_BLD_ZIP5	Ratio of TOT_BLD and Average TOT_BLD of building grouped by Zip Code
TOT_VOL_ZIP5	Ratio of TOT_VOL and Average TOT_VOL of building grouped by Zip Code
$LAND\_LOT\_ZIP5$	Ratio of LAND_LOT and Average LAND_LOT of building grouped by Zip Code
$LAND\_BLD\_ZIP5$	Ratio of LAND_BLD and Average LAND_BLD of building grouped by Zip

	Code
$LAND\_VOL\_ZIP5$	Ratio of LAND_VOL and Average LAND_VOL of building grouped by Zip Code
VAL_LOT_ZIP3	Ratio of VAL_LOT and Average VAL_LOT of building grouped by first 3 digits of Zip
VAL_BLD_ZIP3	Ratio of VAL_BLD and Average VAL_BLD of building grouped by first 3 digits of Zip
VAL_VOL_ZIP3	Ratio of VAL_VOL and Average VAL_VOL of building grouped by first 3 digits of Zip
TOT_LOT_ZIP3	Ratio of TOT_LOT and Average TOT_LOT of building grouped by first 3 digits of Zip
TOT_BLD_ZIP3	Ratio of TOT_BLD and Average TOT_BLD of building grouped by first 3 digits of Zip
TOT_VOL_ZIP3	Ratio of TOT_VOL and Average TOT_VOL of building grouped by first 3 digits of Zip
$LAND\_LOT\_ZIP3$	Ratio of LAND_LOT and Average LAND_LOT of building grouped by first 3 digits of Zip
$LAND\_BLD\_ZIP3$	Ratio of LAND_BLD and Average LAND_BLD of building grouped by first 3 digits of Zip
$LAND\_VOL\_ZIP3$	Ratio of LAND_VOL and Average LAND_VOL of building grouped by first 3 digits of Zip
$EASEMENT$	Whether the property has an easement type or not
$FULL\_ASES$	Ratio of total value and assessed value
EX_ASES	Ratio of exempt value and assessed value
VAL_INC	The percentage increase of the property value

3.2.1 Creating New Variables

- Property Size Measurement

To calculate the unit value, the property size could be measured in 3 levels. The product of **LTFRONT** and **LTDEPTH** indicates the lot area of the property. The product of **BLDFRONT** and **BLDDEPTH** indicates the building area of the property. The product of building area and **STORIES** indicates the building volume of the property.

- Lot Area(**LOTAREA**)= **LTFRONT** * **LTDEPTH**
- Building Area(**BLDAREA**)= **BLDFRONT** * **BLDDEPTH**
- Building Volume (**BLDVOL**)= **BLDAREA*** **STORIES**

- Monetary Variables Scaled by Size

  Unit value was calculated by **FULLVAL**, **AVTOT** and **AVLAND** divided by **LOTAREA**, **BLDAREA** and **BLDVOL** accordingly.

  - $VAL\_LOT = \frac{FULLVAL}{LOTAREA}$ indicates value per unit of lot area
  - $VAL\_BLD = \frac{FULLVAL}{BLDAREA}$ indicates value per unit of building area
  - $VAL\_VOL = \frac{FULLVAL}{BLDVOL}$ indicates value per unit of building volume
  - $TOT\_LOT = \frac{AVTOT}{LOTAREA}$ indicates assessed value per unit of lot area
  - $TOT\_BLD = \frac{AVTOT}{BLDAREA}$ indicates assessed value per unit of building area
  - $TOT\_VOL = \frac{AVTOT}{BLDVOL}$ indicates assessed value per unit of building volume
  - $LAND\_VOL = \frac{AVLAND}{LOTAREA}$ indicates assessed value of land per unit of lot area
  - $LAND\_BLD = \frac{AVLAND}{BLDAREA}$ indicates assessed value of land per unit of building area
  - $LAND\_VOL = \frac{AVLAND}{BLDVOL}$ indicates assessed value of land per unit of building volume

3.2.2 Grouping numerical variables by different categorical variables

By performing group-by transformation of numerical variables, we can see how much one property is deviated from other properties within the same group. The first step was to calculate the average of each group, then divide each value by its corresponding group average.

CATEGORICAL VARIABLES	UNIQUE COUNT
TAX CLASS	11
BUILDING CLASS	25
BORO	5
ZIP5	196
ZIP3	14

- **TAXCLASS**
  Tax class defines the different classes of property that might result in different tax rates. For example, tax class 1 is for 1-3 unit residences while tax class 3 is for utilities.

$$ie.VAL\_LOT\_TAX = \frac{VAL\_LOT}{Avg(VAL\_LOT)\ of\ TAXCLASS}$$

- **BLDGCL**
  Building class defines what kind of building the property belongs to. For example, Class D Buildings are all elevator apartments and D8 is the luxury type; Class L Buildings are all loft buildings and L8 is with Retail Stores. There are 200 unique type of building class. To get more general comparison, we only keep first digit of building class. Under this circumstance, there are 25 unique levels.

$$ie.VAL\_LOT\_BLDGCL = \frac{VAL\_LOT}{Avg(VAL\_LOT)\ of\ BLDGCL}$$

- **BORO**
  The first digit of **BBLE**, **BORO** indicates the Borough code of the property.

  1 = MANHATTAN
  2 = BRONX
  3 = BROOKL YN
  4 = QUEENS
  5 = STATEN ISLAND

$$ie.VAL\_LOT\_BORO = \frac{VAL\_LOT}{Avg(VAL\_LOT)\ of\ BORO}$$

- **ZIP5**
  The zip code of the property.

$$ie.VAL\_LOT\_ZIP5\ = \frac{VAL\_LOT}{Avg(VAL\_LOT)\ of\ ZIP5}$$

- **ZIP3**
  The first 3 digits of the zip code of the property. The first digit designates the general area of the country with numbers starting lower in the east and increasing as you move west. The next two digits referred to one of the 450+ Sectional Center Facilities (SCFs) in the US.

$$ie.VAL\_LOT\_ZIP3 = \frac{VAL\_LOT}{Avg(VAL\_LOT)\ of\ ZIP3}$$

3.2.3 Other Expert Variables

- $FULL\_ASES = \frac{Full\ Value}{Assessed\ Total\ Value}$ is the ratio of market value and assessed value.

- $EX\_ASES = \frac{Exepmt\ Total\ Value}{Assessed\ Total\ Value}$ indicates what proportion of property's value has been reduced.

- $VAL\_INC = \frac{AVTOT2 - AVTOT}{AVTOT}$: indicates the percentage increase of the property value from first assessment and second assessment. The unusual increasing percentage contributes to detect abnormal records.

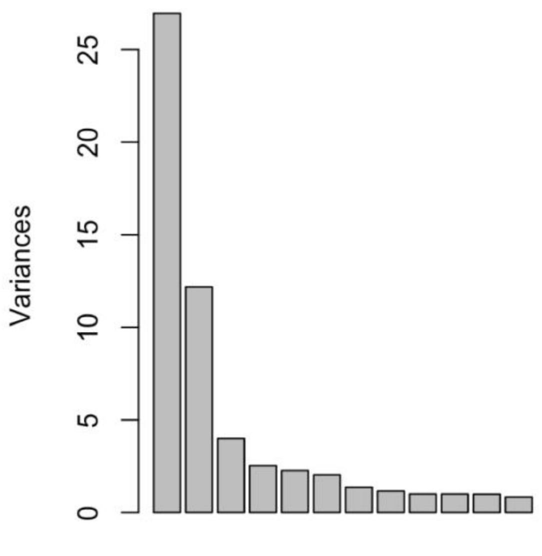
4. Building Fraud Algorithm

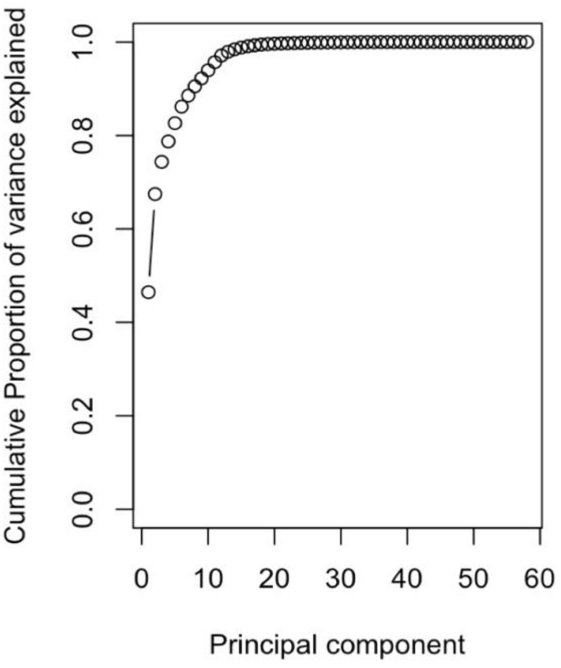
4.1 Standardization and Dimensionality Reduction

After we substituted missing values and generated all the expert variables through the methods mentioned above, we proceeded to the standardization of expert variables and dimensionality reduction to calculate fraud scores of each record for further analysis.

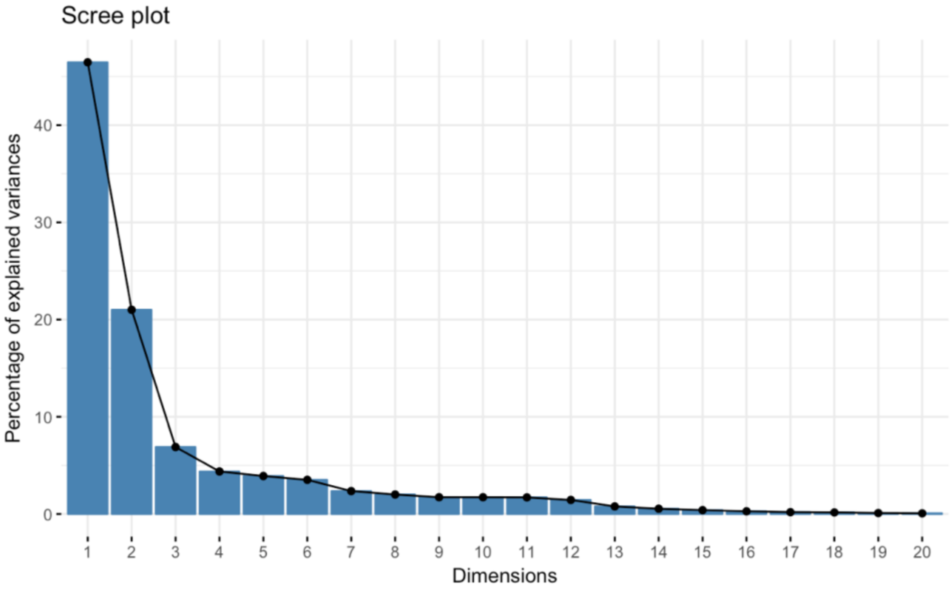
We performed principal components analysis using the **prcomp(data, scale = TRUE, center = TRUE)** function in R to find the dominant directions in the data. This is a built-in function that is used to perform PCA in R. By setting **scale = TRUE** and **center = TRUE**, we scaled the variables to a normal distribution with mean of 0 and standard deviation of 1 to realize standardization of variables.

Then we divided the variance explained of each principal components(PCs) by the total variance of all PCs to compute the percentage of explained variance of each PC. By this way we can sum the percentage to plot a cumulative proportion of variance explained to determine how many PCs and which PCs to keep. We want to reduce the dimensionality and keep a limited number of variables that most represent the dataset. The left plot describes the value of variances of each PC starting from the largest on the left to the smallest on the right. The right plot tells us that PC1 to PC10 explained 93.97611% of all 58 PCs' variance.

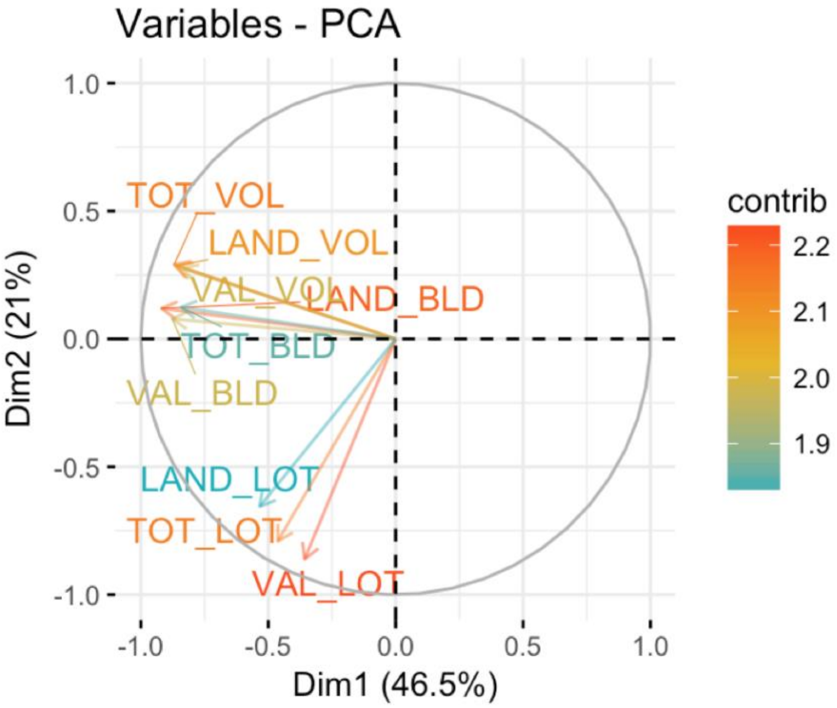




We made the following scree plot using the **fviz_eig()** function in the “factoextra” package of R to plot the percentage of explained variance for 20 variables directly in a clearer way.



We used **fviz_pca_var()** function in R to have a clear picture of what variables are contributing to the first two principal components (PC1 and PC2). We observed that nine out of the above twelve variables are contributing significantly to PC1 and PC2, meaning that they explained a large proportion of variance.

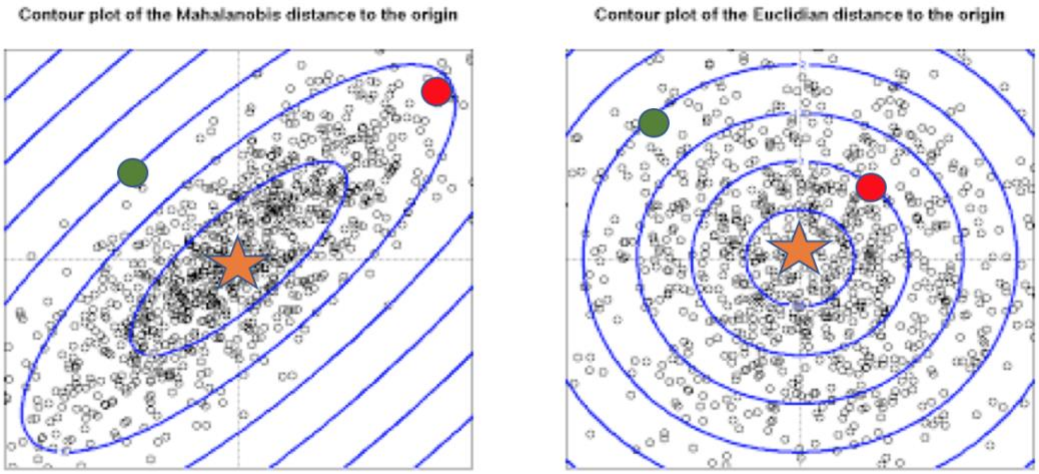


Then, we used two different methods to calculate fraud scores of the standardized data with 20 PCs based on the previous analysis including:

- Heuristic algorithm: Mahalanobis distance
- AutoEncoder: Reconstruction error

4.2 Heuristic Algorithm

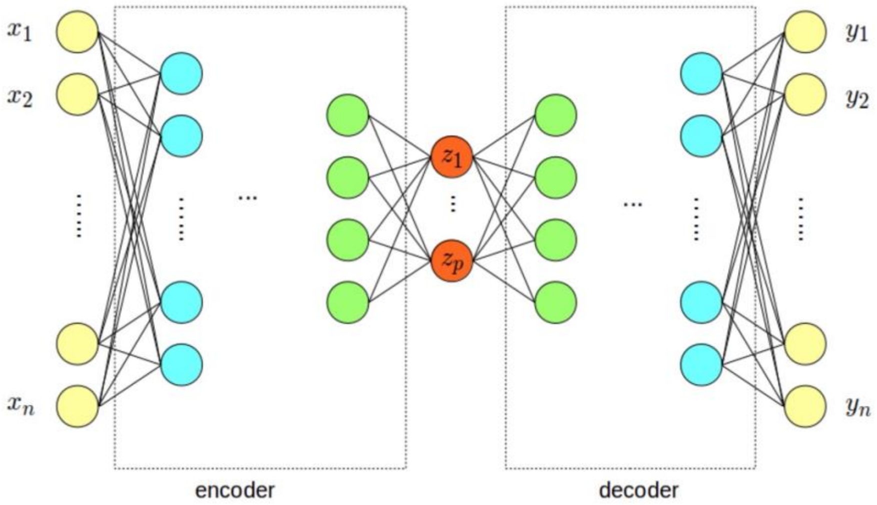
Mahalanobis distance is a measure of the distance between a point and a distribution. The Mahalanobis distance takes into account the different scales and correlations. It draws equal contours by scaling based on the correlations and different standard deviations. Along each PC axis, Mahalanobis distance measures the number of standard deviation from the point to the mean of the distribution.



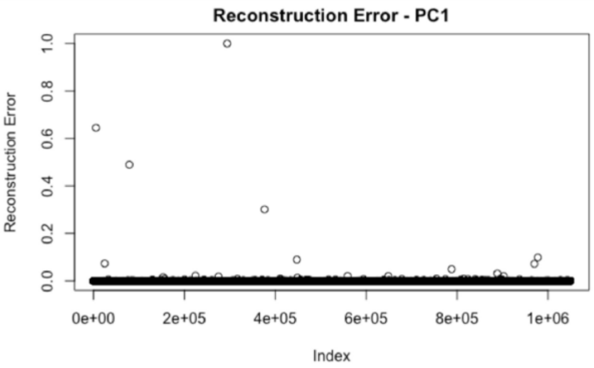
Our algorithm was to use this distance as a fraud score for each record. We used **mahalanobis()** function in R to calculate the distance, thus the fraud score.

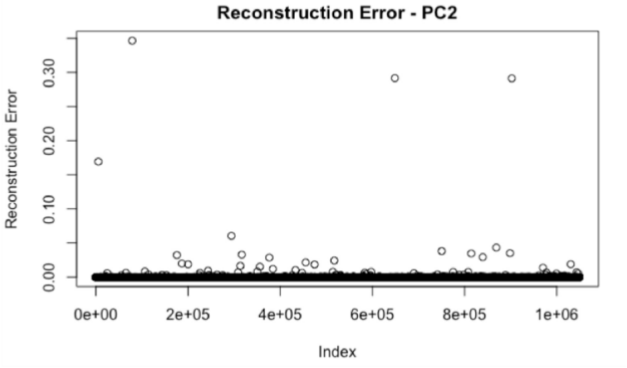
4.3 AutoEncoder

AutoEncoder can train the model to reproduce the original data. The aim of an AutoEncoder is to learn a representation (encoding) for a set of data, typically for the purpose of dimensionality reduction.



We tried to AutoEncoder PCs by using the “h2o” package in R. We used **h2o.deeplearning()** function with the parameter **Autoncoder = TRUE**. This function could AutoEncoder the dataset with all PCs by building a deep neural network. Then we used **h2o.anaomaly()** function which was designed to detect anomalies. The principal was to reconstruct the dataset using the reduced set of features(10 PCs) and calculated the mean squared error. With the parameter **per_feature = TRUE**, we could get individual squared reconstruction of each features. By plotting the reconstruction error for different PCs, we found out that the abnormal values might indicate fraud. Below are some examples of two PCs in an easy-to-read way after a log transformation.





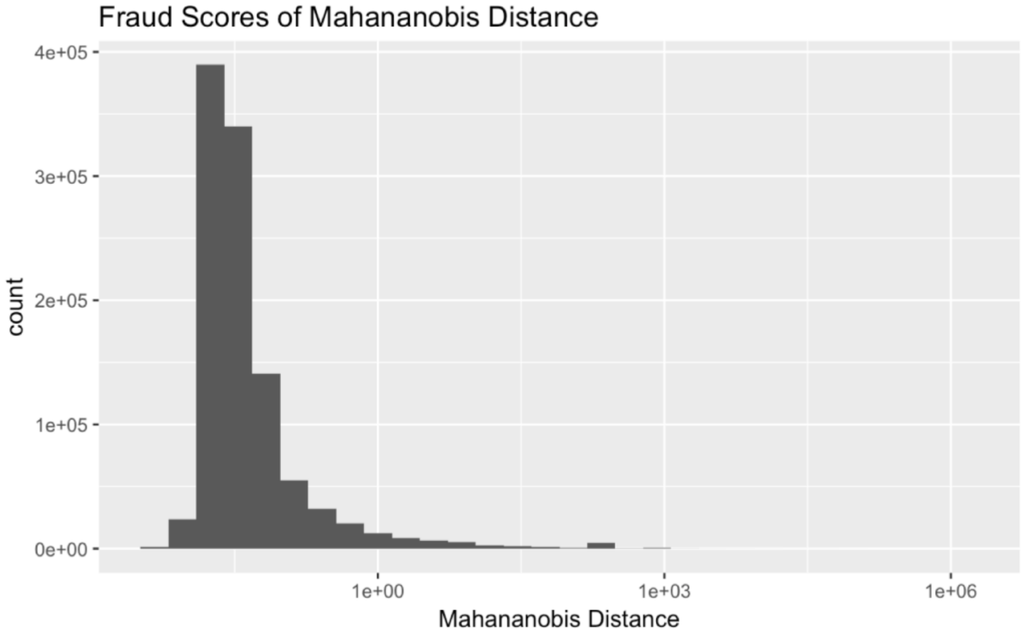
Finally, we summed up the value of reconstruction errors of 10 PCs to get a single score (fraud score) for each record.

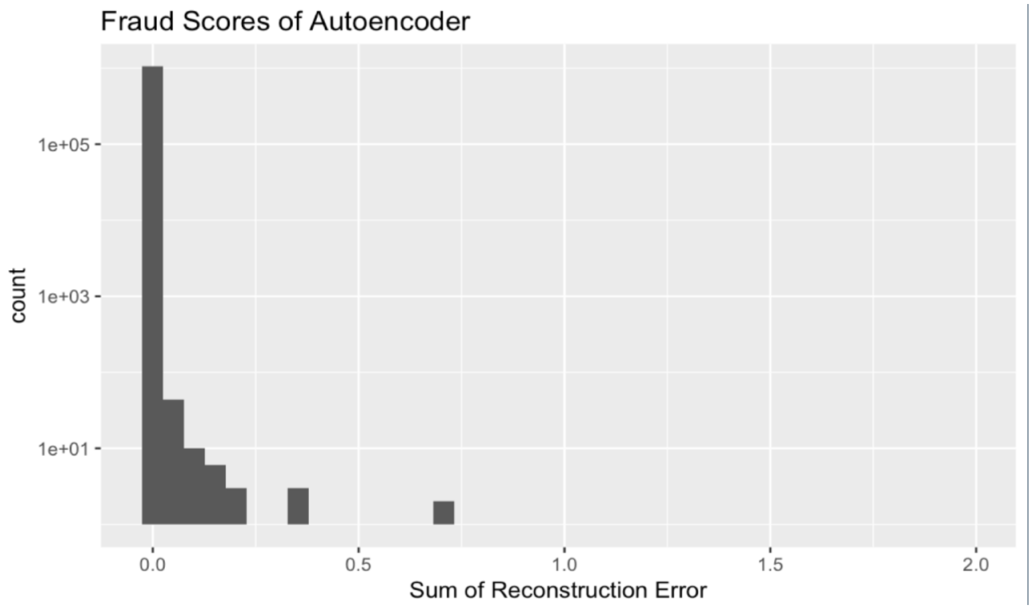
5. Results

5.1 Fraud score distribution

Each record of the dataset was assigned two separate fraud scores using different algorithms. The distributions of fraud score are right-skewed meaning that the majority of records has low scores while a small proportion of records has high scores.

Below are the distributions of fraud scores for both algorithms.





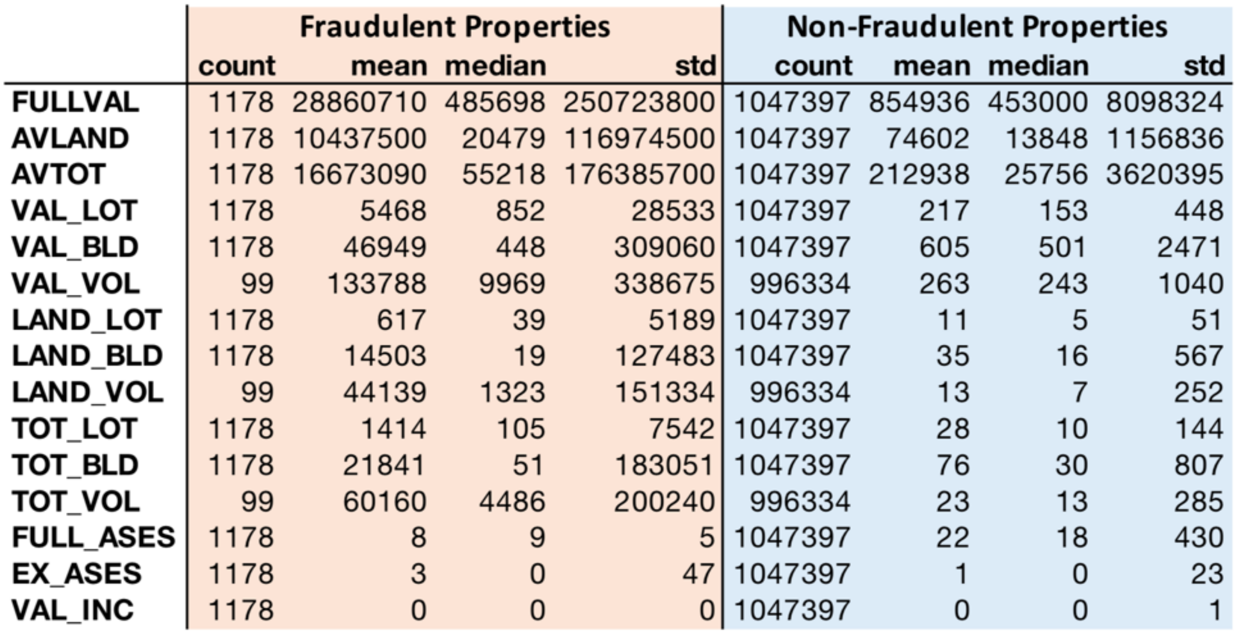
5.2 Fraud Score Integration

For each of the two scores, we first sorted the records by score and then applied quantile binning. Quantile binning discretized variables into equal-sized buckets based on sample quantiles. For each score, we used 0.1 percent quantile as the threshold and as a result we got 1000 bins. Each record in the highest bins was assigned the value of 1000. Each record in the second highest bin was assigned the value of 999. This procedure was repeated until each record in the last bins was assigned the value of 1.

By adding two scaled value, each record had a new integrated score. We sorted the record again by the new scores and selected top 0.1% high score records (1178 records in total as anomaly for further analysis.

5.3 Significant Insights

5.3.1 Statistical Results



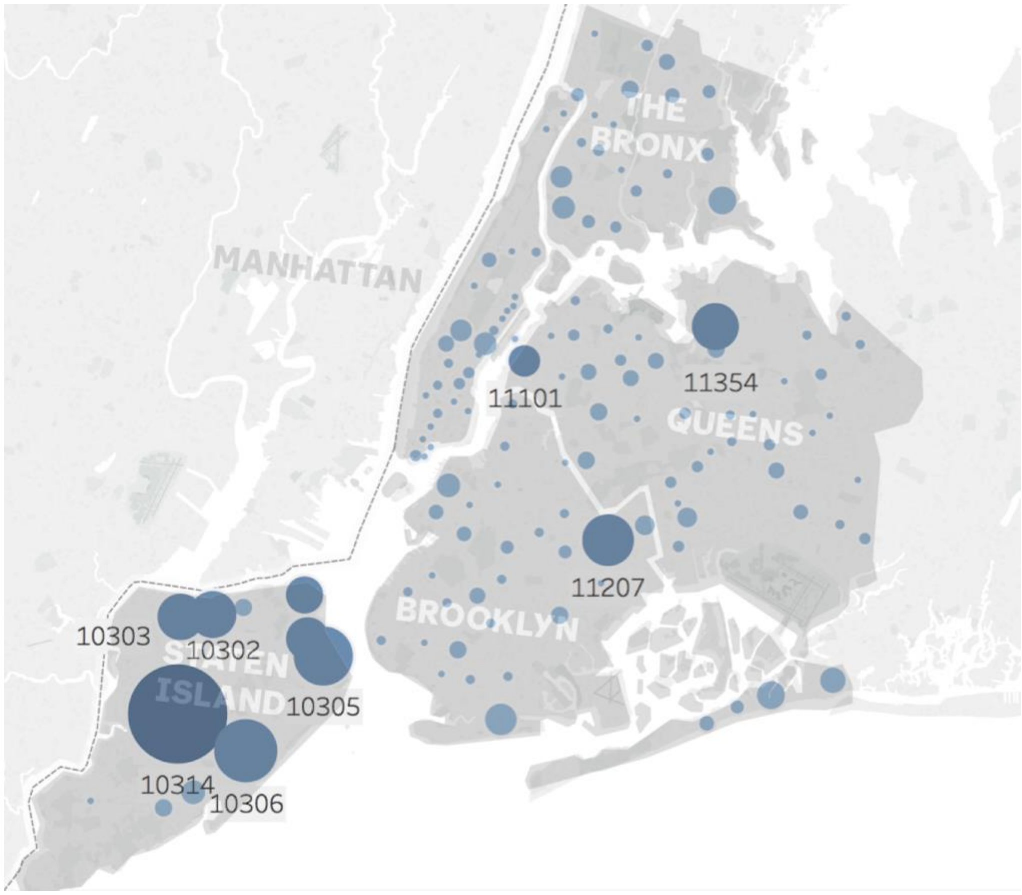
	Fraudulent Properties				Non-Fraudulent Properties			
	count	mean	median	std	count	mean	median	std
FULLVAL	1178	28860710	485698	250723800	1047397	854936	453000	8098324
AVLAND	1178	10437500	20479	116974500	1047397	74602	13848	1156836
AVTOT	1178	16673090	55218	176385700	1047397	212938	25756	3620395
VAL_LOT	1178	5468	852	28533	1047397	217	153	448
VAL_BLD	1178	46949	448	309060	1047397	605	501	2471
VAL_VOL	99	133788	9969	338675	996334	263	243	1040
LAND_LOT	1178	617	39	5189	1047397	11	5	51
LAND_BLD	1178	14503	19	127483	1047397	35	16	567
LAND_VOL	99	44139	1323	151334	996334	13	7	252
TOT_LOT	1178	1414	105	7542	1047397	28	10	144
TOT_BLD	1178	21841	51	183051	1047397	76	30	807
TOT_VOL	99	60160	4486	200240	996334	23	13	285
FULL_ASES	1178	8	9	5	1047397	22	18	430
EX_ASES	1178	3	0	47	1047397	1	0	23
VAL_INC	1178	0	0	0	1047397	0	0	1

From the result, we could see that potentially fraudulent properties tend to have significant higher mean, median and standard deviation of FULLVAL, AVLAND, AVTOT, VAL_LOT, VAL_BLD, VAL_VOL, LAND_LOT, LAND_BLD, LAND_VOL, TOT_LOT, TOT_BLD, TOT_VOL compared to the non-fraudulent properties. However, for FULL_ASES, the ratio of total value and assessed value of properties, fraudulent properties got much lower values, which might indicate that they were undervalued intentionally so that the owners could pay fewer taxes than what they should pay.

5.3.2 Categorical Analysis

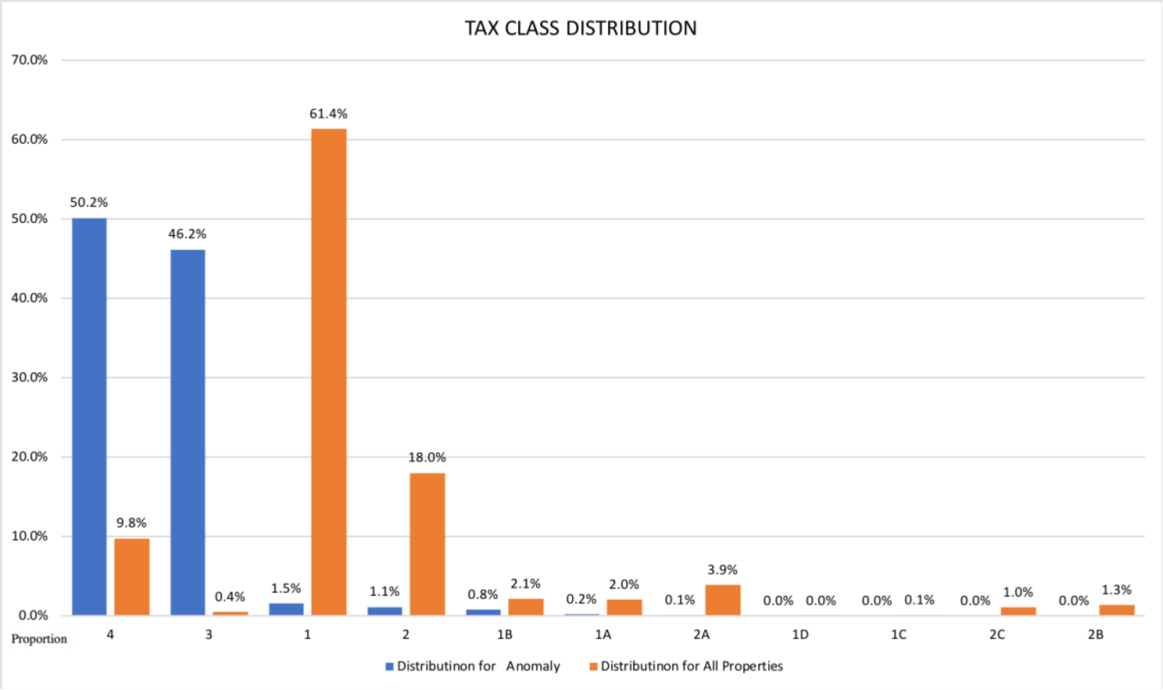
- **Zip code and Borough**

We generated a geographic map based on zip code. Most abnormal records locate in STATEN ISLAND which is Borough 5. Compared to the proportion of Borough 5 in full dataset which only takes 10.8%, their high proportion(30.83%) in abnormal records is very unusual. Properties in these areas, especially those with zip codes of 10314, 10306, 10305, tend to be classified as fraudulent entities.



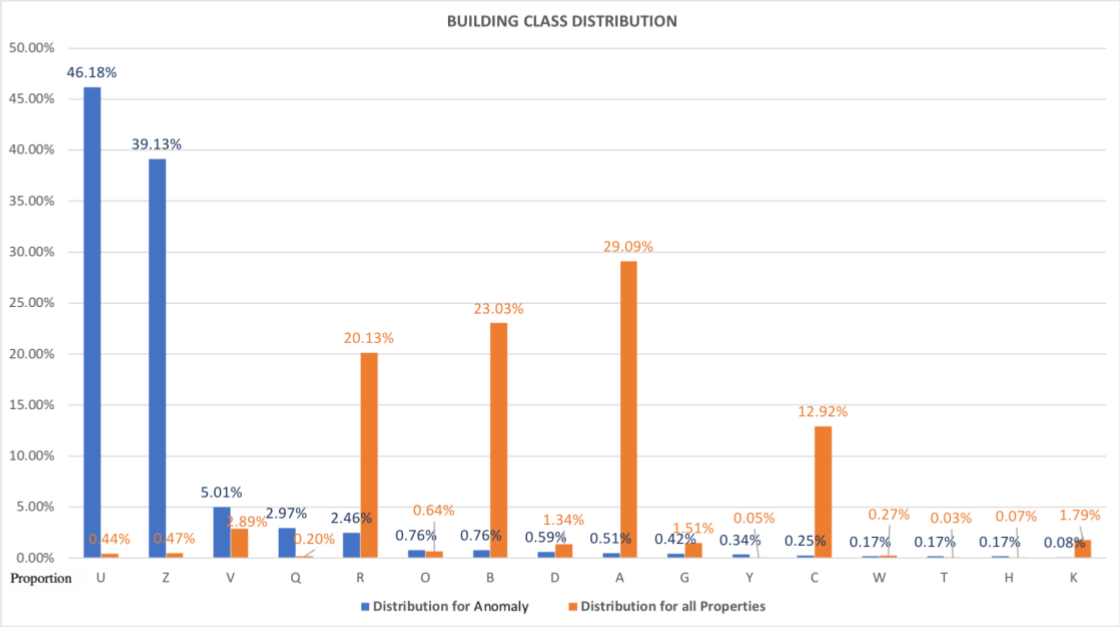
** Color shows details about ZIP. Size shows sum of records. The Marks are labeled by ZIP*

- **Tax Class**



97 % records of high scores belongs to Tax Class 4 and Tax Class 3. According to NY government website[3], Class 4 refers to all commercial and industrial properties, Class 3 refers to most utility property. Class 1 and Class 2 are all residential property. Residential properties might have more stable and regulated market value. Commercial properties could have more variation.

- **Building Class**



Building class U and Z account for a large proportion of abnormal records. Class U Buildings are all Utility Bureau Properties, and Class Z are miscellaneous properties, such as lakes in Central Park and foreign government. These properties might follow different rules of measuring values, and for that reason they don't conform to regular anomaly detection algorithms. It is also possible that fraud might happen in the valuation of these properties.

- **Owner**

We ranked the top owners of records. Excluding government agency and public authority, we listed top 3 commercial owner or private owner.

1. Ocean View Estates:
All properties in Ocean View Estates listed online are in Huntington Beach, CA. Not a single property is in New York. This estate should be paid attention to in later investigation.

2. Buckeye Pipeline
The 15 properties owned by BUCKEYE PIPELINE are all labeled as abnormal records. The building class of these properties are all U9 which includes miscellaneous buildings like private improvements on city lands and in public places. Two records miss data of **STADDR**.

3. Haven Builders, Inc.
2 properties owned by HAVEN BUILDERS, INC are both labeled as abnormal records. Two records both have extreme small **AVTOT** compared to **FULLVAL**. It's very unusual because the correlation of **AVTOT** and **FULLVAL** of whole dataset is 0.77 which means those two variables are strongly related.

- **Easement**

80.7 percent of records with high scores contains Easement element, while Easement only contributes to 0.38 percent of total records. This means that most of the potentially fraudulent properties has Easement status, which can be a red flag area that needs further analysis.

5.3.4 Top 10 Records

In order to take a close look at each record, we decided to filter out overlapped records for each algorithm based on quantile binning and to analyze the top 10 records. The method we used to find top 10 records was to assign two rank numbers for each overlapped record based on fraud scores from both algorithms with the highest score ranked first. Then we summed up two rank numbers for each record and sorted the sum to find the first 10 records as top 10 fraud score records for further analysis.

1. **Record 5393**
   According to its **STADDR**, we can see this property belongs to an apartment building with 6 floors. But record shows it only has 1 stories, building front and building depth are both 1 which is obviously inconsistent with our finding.

2. **Record 78804**
   This property belongs to US government. It has large gap between first assessed value and second assessed value. **AVTOT/AVLAND/EXTOT/EXLAND** all decreased from 2 billion to 0.8 billion. This large discrepancy need further investigation.

3. **Record 294061**
   This property is The Metropolitan Museum of Art! It owns the greatest value in fields of **FULLVAL**, **AVLAND**, **EXLAND** which makes this property be abnormal. We think these great values are fair and reasonable since the MET is one of the greatest museum in the world. This property is considered as false positive.

4. **Record 435535**
   This property is a two-floor house owned by private. It has unusual value of Full value divided by building volume with respect to **BORO**, **TAXCLASS**.

5. **Record 376243**
   This property has largest value of **AVTOT/EXTOT/AVTOT2/EXTOT2**. According to street address, we found this property is Holiday Inn Hotel near New York JFK Airport. The fraction between **FULLVAL** and **AVTOT** is extreme small compared to other properties. The owner LOGAN PROPERTY, INC. is a family owned and operated company. Considered the size of this company, it's very unusual for them to possess this property

6. **Record: 902256**
   This property is a residential building on Bell Avenue. It achieves abnormally very high values in all lot-area-related variables including **VAL_LOT**, **LAND_LOT** and **TOT_LOT** that are caused by **LTFRONT** and **LTDEPTH** both equal to 1.

7. **Record 648675**
   This property is a residential building on Vleigh Place. It achieves abnormally very high values in all lot-area-related variables including **VAL_LOT**, **LAND_LOT** and **TOT_LOT** that are caused by **LTFRONT** and **LTDEPTH** both equal to 1.

8. **Record 977471**
   This property is an office building on 28-10 Queens Plaza South. It achieves abnormally very high values in all building-area-related variables including **VAL_BLD**, **VAL_VOL**,

LAND_BLD, **LAND_VOL**, **TOT_BLD**, **TOT_VOL** that are caused by **BLDFRONT** and **BLDDEPTH** both equal to 1.

9. **Record 447396**
   This property refers to a vacant land owned by Department of General Services on Flatbush Avenue. It achieves very high values in **FULLVAL**, **AVLAND** and **AVTOT**. Besides, it achieves abnormal very high values in building-area-related variables including **VAL_BLD**, **LAND_BLD** and **TOT_BLD** that are caused by relatively small **BLDAREA** compared with large **LOTAREA**.

10. **Record 970081**
    This property refers to a park on Joe Dimaggio Highway. It achieves abnormally very high values in building-area-related variables including **VAL_BLD**, **VAL_VOL**, **LAND_BLD**, **LAND_VOL**, **TOT_BLD**, **TOT_VOL** that are caused by very small values of **BLDFRONT** and **BLDDEPTH**.

6. Conclusion

Our analysis shows that fraudulent properties are more likely to have remarkably higher numbers such as total market value, assessed value of land, assessed total value, etc. compared to non-fraudulent entities. They also tend to have Easement status. Lower ratio of total value and assessed value of properties implied that the properties were intentionally devalued for the tax benefits of the owner.

The geographical representation of zip code indicates that the STATEN ISLAND (especially in areas with zip codes 10314, 10306 and 10305) has the highest probability of having fraudulent activities involving real estate.

Investigation of tax class and building class signifies that a large proportion of fraud records are properties of either government entities or public authorities. One possible reason to explain this result is that government properties may be appraised using different criteria compared to other properties. Thus, higher scores of those properties are expected.

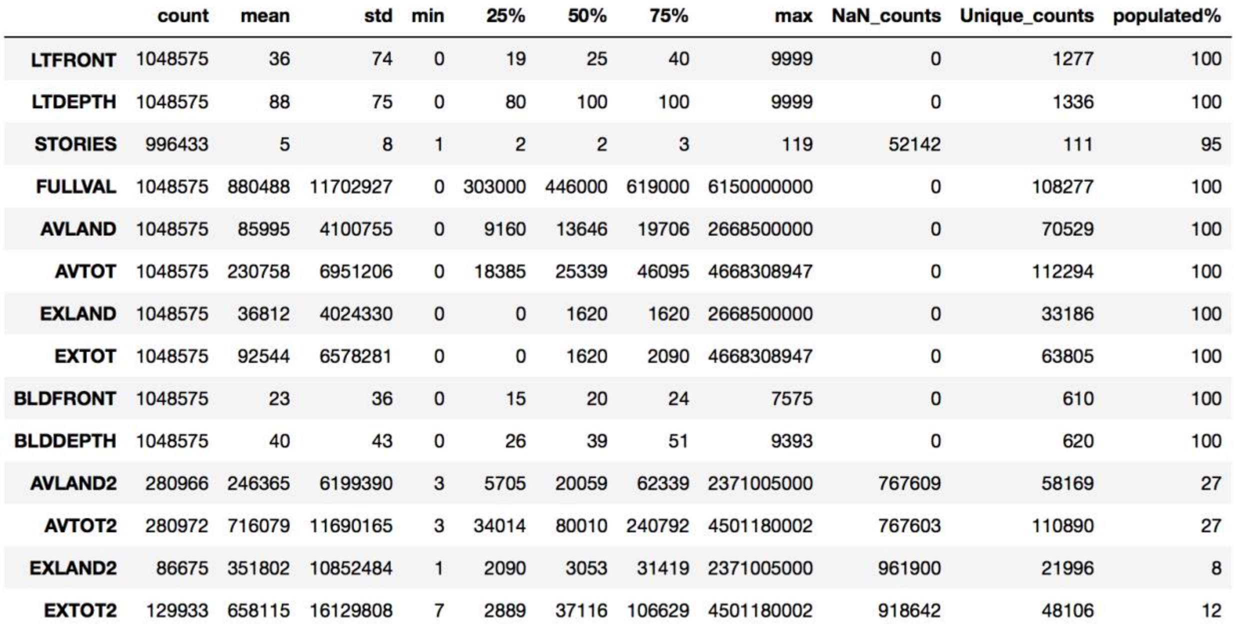
Ocean View Estates, Buckeye Pipeline and Haven Builders, Inc. are the three owners that stand out to us as abnormal. Besides, the analysis of top 10 records shows that fraudulent records tend to have significant lower or higher value in some features such as total market value, lot frontage and lot depth of the property.

In order to derive better results, further research should be done in order to create a comprehensive mechanism of assigning fraud scores to different properties according to their tax and building classes. Other algorithms such as k-Means Clustering can be performed to compare the results of our report with the findings generated from those algorithms.

7. Appendix: Data Quality Report on NY Property Data

1. Summary Statistics for Numerical Variables:

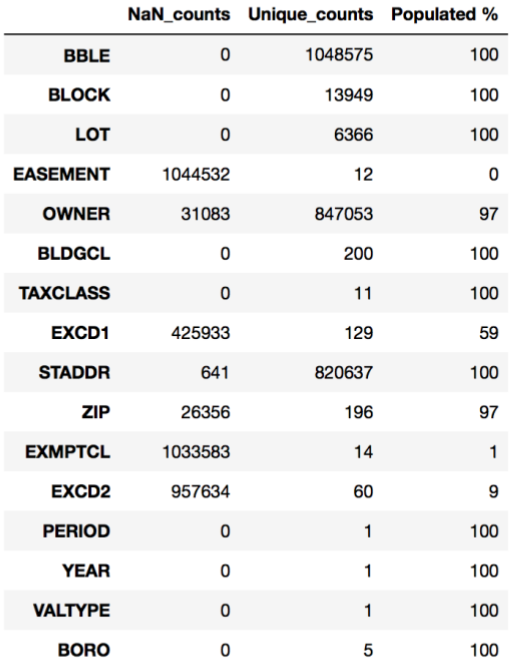
For 14 numerical variables in the dataset, we summarize some mathematical statistics such as mean, median (which is the 50th percentile), standard deviation, minimum, maximum, the first and third percentile, number of NA, number of unique values, and percentage of populated records for that variable. A summary table is provided below to demonstrate those characteristics.



	count	mean	std	min	25%	50%	75%	max	NaN_counts	Unique_counts	populated%
LTFRONT	1048575	36	74	0	19	25	40	9999	0	1277	100
LTDEPTH	1048575	88	75	0	80	100	100	9999	0	1336	100
STORIES	996433	5	8	1	2	2	3	119	52142	111	95
FULLVAL	1048575	880488	11702927	0	303000	446000	619000	6150000000	0	108277	100
AVLAND	1048575	85995	4100755	0	9160	13646	19706	2668500000	0	70529	100
AVTOT	1048575	230758	6951206	0	18385	25339	46095	4668308947	0	112294	100
EXLAND	1048575	36812	4024330	0	0	1620	1620	2668500000	0	33186	100
EXTOT	1048575	92544	6578281	0	0	1620	2090	4668308947	0	63805	100
BLDFRONT	1048575	23	36	0	15	20	24	7575	0	610	100
BLDDEPTH	1048575	40	43	0	26	39	51	9393	0	620	100
AVLAND2	280966	246365	6199390	3	5705	20059	62339	2371005000	767609	58169	27
AVTOT2	280972	716079	11690165	3	34014	80010	240792	4501180002	767603	110890	27
EXLAND2	86675	351802	10852484	1	2090	3053	31419	2371005000	961900	21996	8
EXTOT2	129933	658115	16129808	7	2889	37116	106629	4501180002	918642	48106	12

2. Summary Statistics for Categorical Variables:

For 15 categorical variables in the dataset and our newly created “BORO” field, we summarize the mathematical statistics for the total of 16 categorical variables: Number of null values, number of unique values and percentage of populated records for that variable.



	NaN_counts	Unique_counts	Populated %
BBLE	0	1048575	100
BLOCK	0	13949	100
LOT	0	6366	100
EASEMENT	1044532	12	0
OWNER	31083	847053	97
BLDGCL	0	200	100
TAXCLASS	0	11	100
EXCD1	425933	129	59
STADDR	641	820637	100
ZIP	26356	196	97
EXMPTCL	1033583	14	1
EXCD2	957634	60	9
PERIOD	0	1	100
YEAR	0	1	100
VALTYPE	0	1	100
BORO	0	5	100

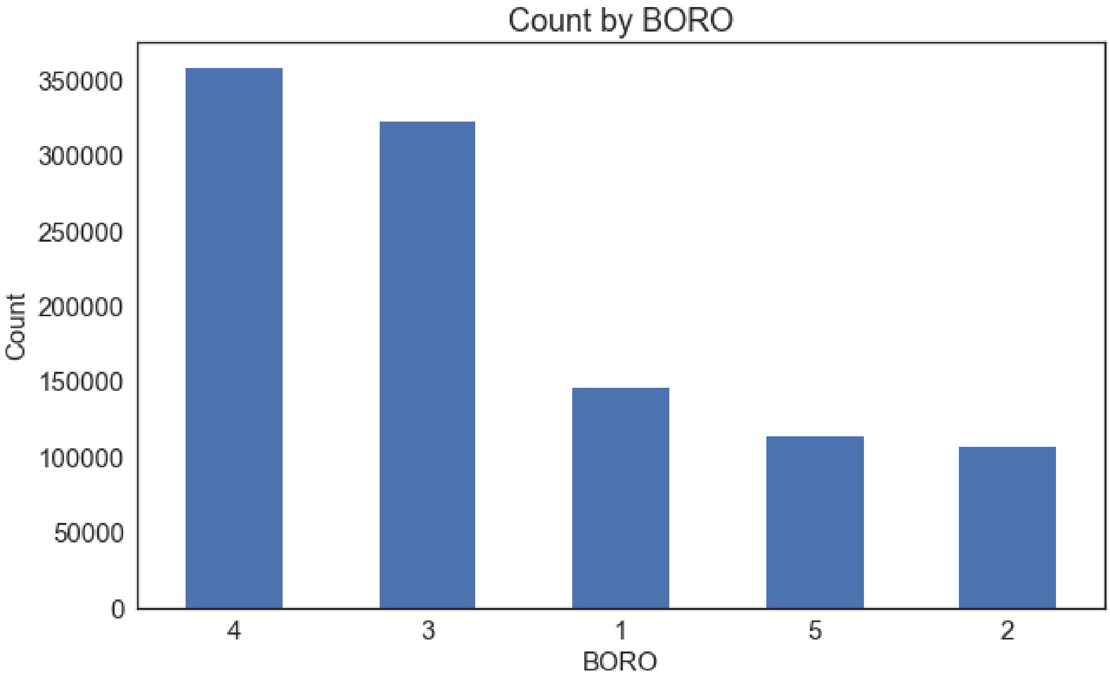
2. Information for Each Field

Below is the general information for the 30 fields (including the newly created “BORO”) in the dataset. Each field is demonstrated by “Field Name”, “Type” and “Descriptions”.

“RECORD” is considered as an index and does not provide any information regarding the quality of the data. Therefore, it is not included in this section.

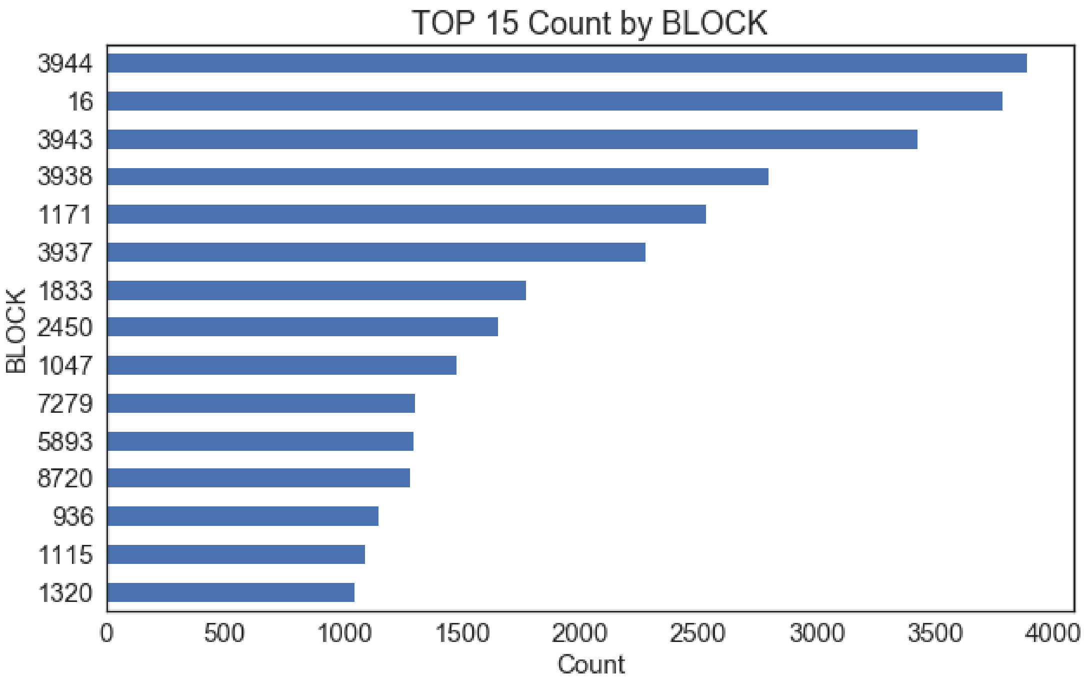
The first field of the data, “BBLE”, does not provide insightful information about the quality of the data because it is the concatenation of “BORO”, “BLOCK”, “LOT”, and “EASEMENT” fields. Therefore, in order to assess the quality of “BBLE” field, we analyze the four fields mentioned above in details and provide descriptions of each field below. we extract the first letter of “BBLE” field and add a new column named “BORO”.

Field Name	**Type**	**Description**
BORO	Categorical	Borough Codes 1 = MANHATTAN 2 = BRONX 3 = BROOKLYN 4 = QUEENS 5 = STATEN ISLAND

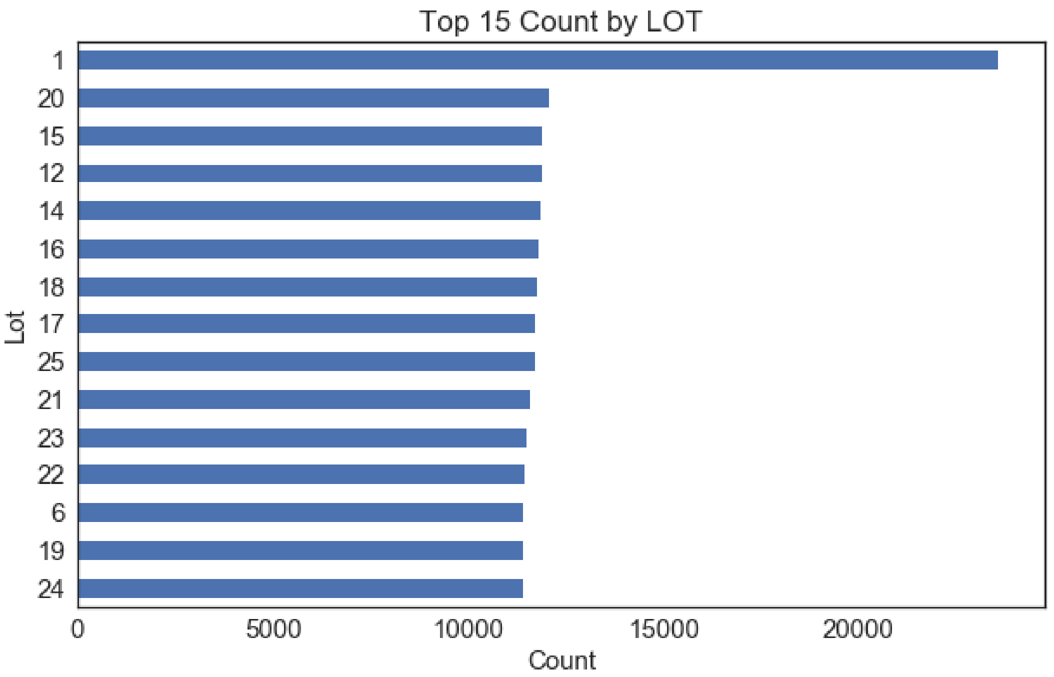


Field Name	Type	Description

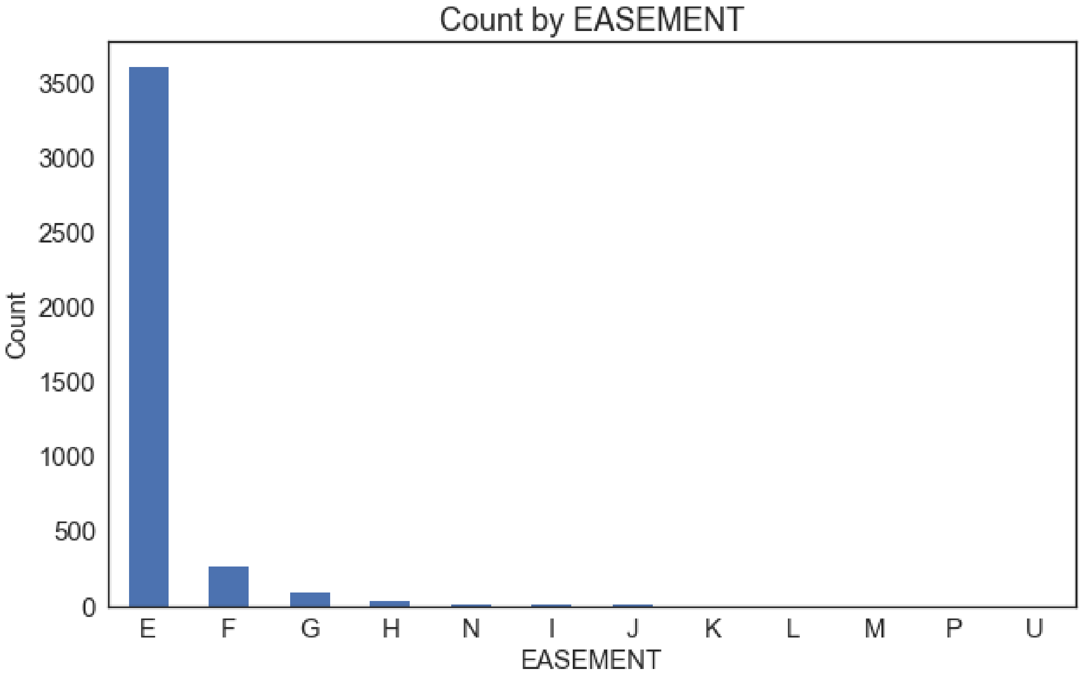
BLOCK	Categorical	Valid Block Ranges by Borough
		MANHATTAN = 1 TO 2,255
		BRONX = 2,260 TO 5,958
		BROOKLYN = 1 TO 8,955
		QUEENS = 1 TO 16,350
		STATEN ISLAND = 1 TO 8,050



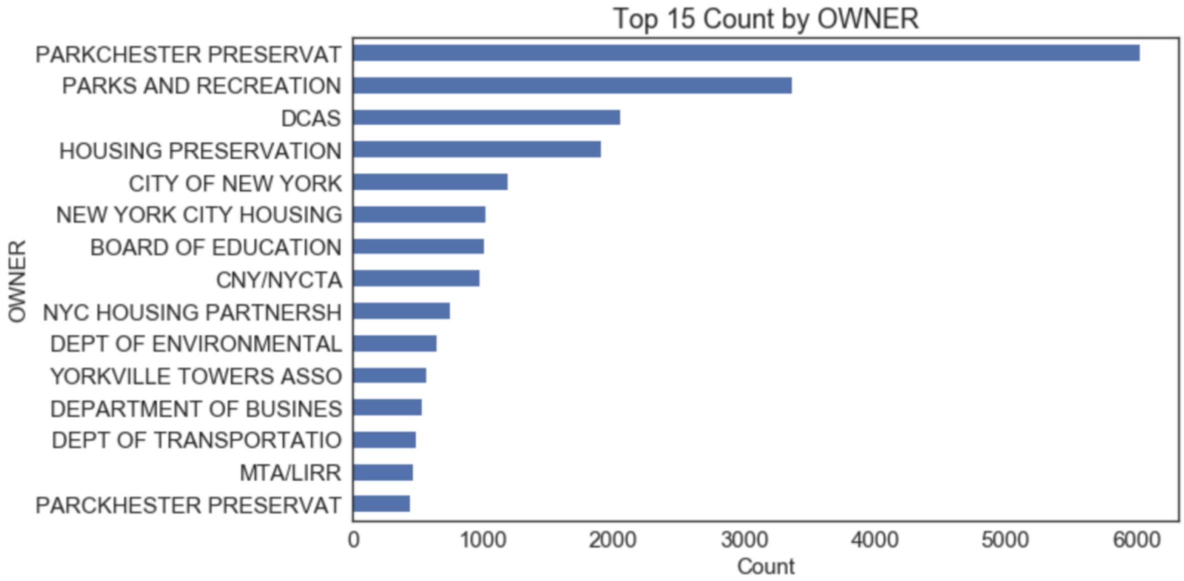
Field Name	Type	Description
LOT	Categorical	Unique Number Within Borough and Block



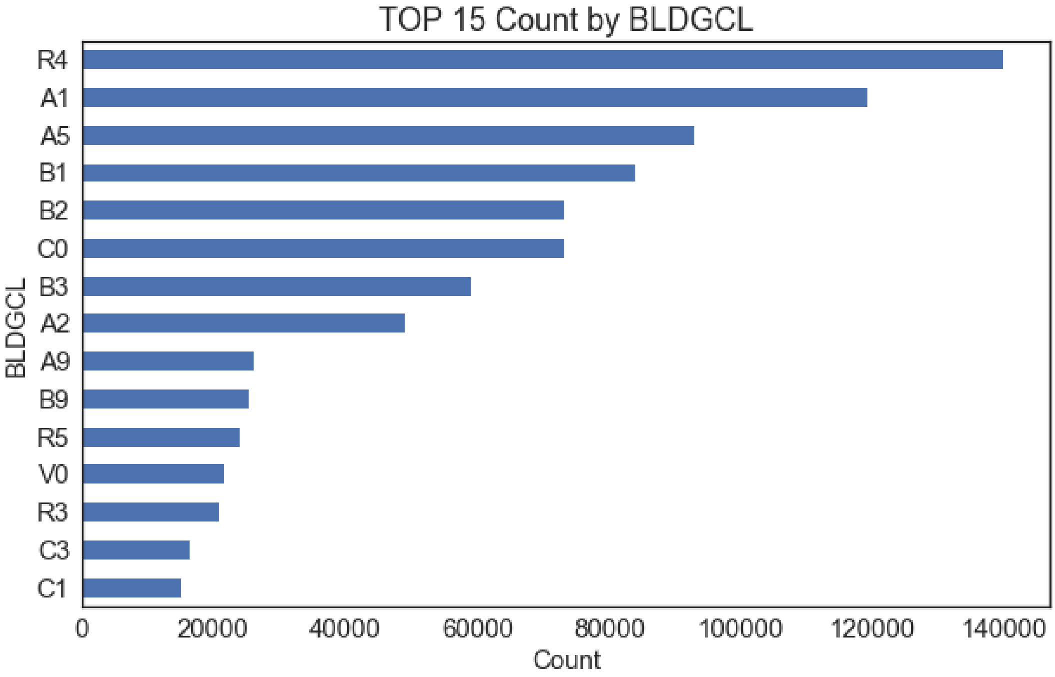
Field Name	Type	Description
EASEMENT	Categorical	Describe Easement Status of The Property SPACE = the Lot Has No Easement. 'A' = the Portion of the Lot that Has an Air Easement 'B' = Non-Air Rights. 'E' = the Portion of the Lot that Has a Land Easement 'F' through 'M' = Duplicates of 'E' 'N' = Non-Transit Easement 'P' = Piers.



Field Name	Type	Description
OWNER	Categorical	The Owner’s Name

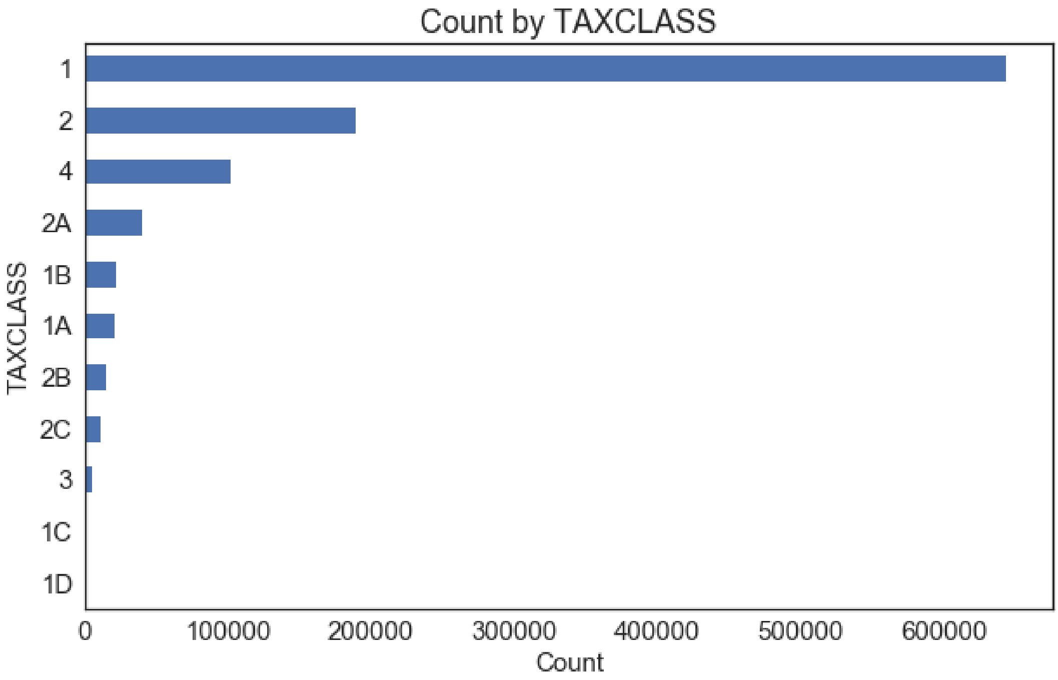


Field Name	**Type**	**Description**
BLDGCL	Categorical	Building Class. This field has a direct correlation with Tax Class field. If the Building Class is known, the Tax Class can be generated.

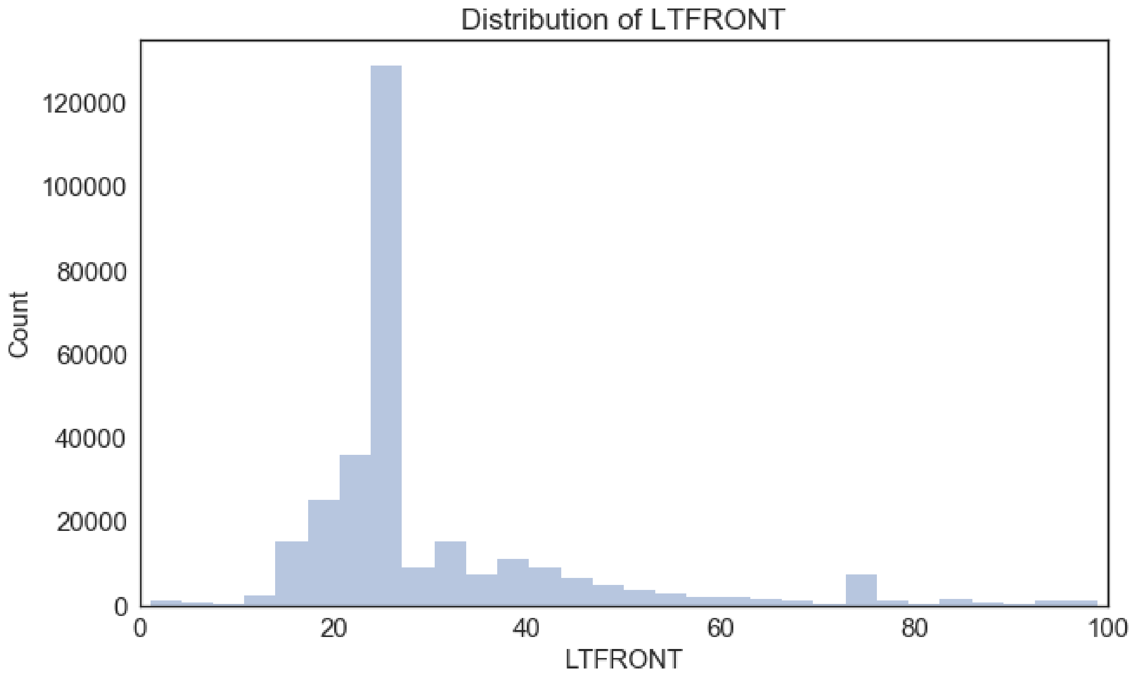


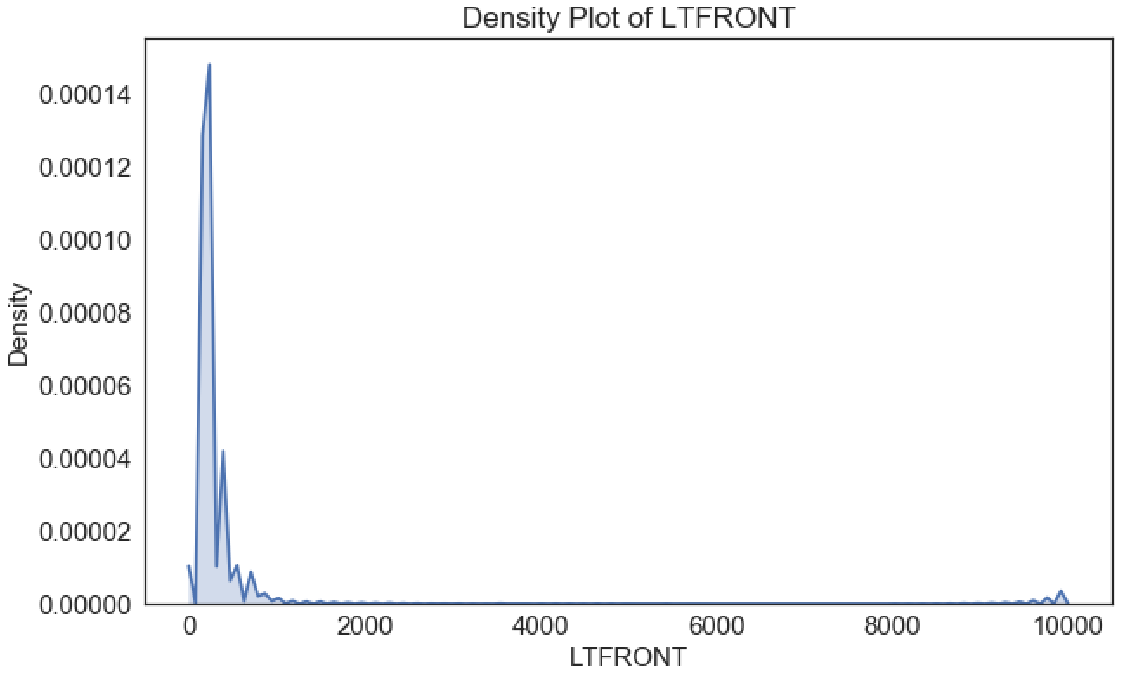
Field Name	**Type**	**Description**

TAXCLASS	Categorical	Current Property Tax Class Code (NYS Classification). This field has a direct correlation with Building Class field. If the Building Class is known, the Tax Class can be generated.

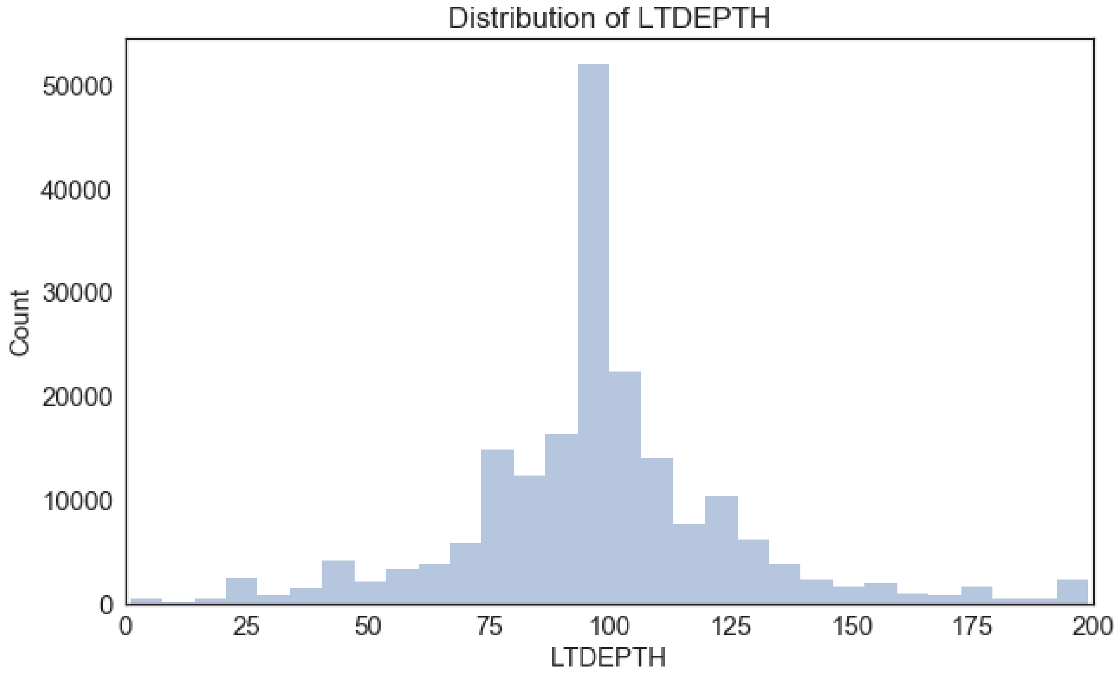


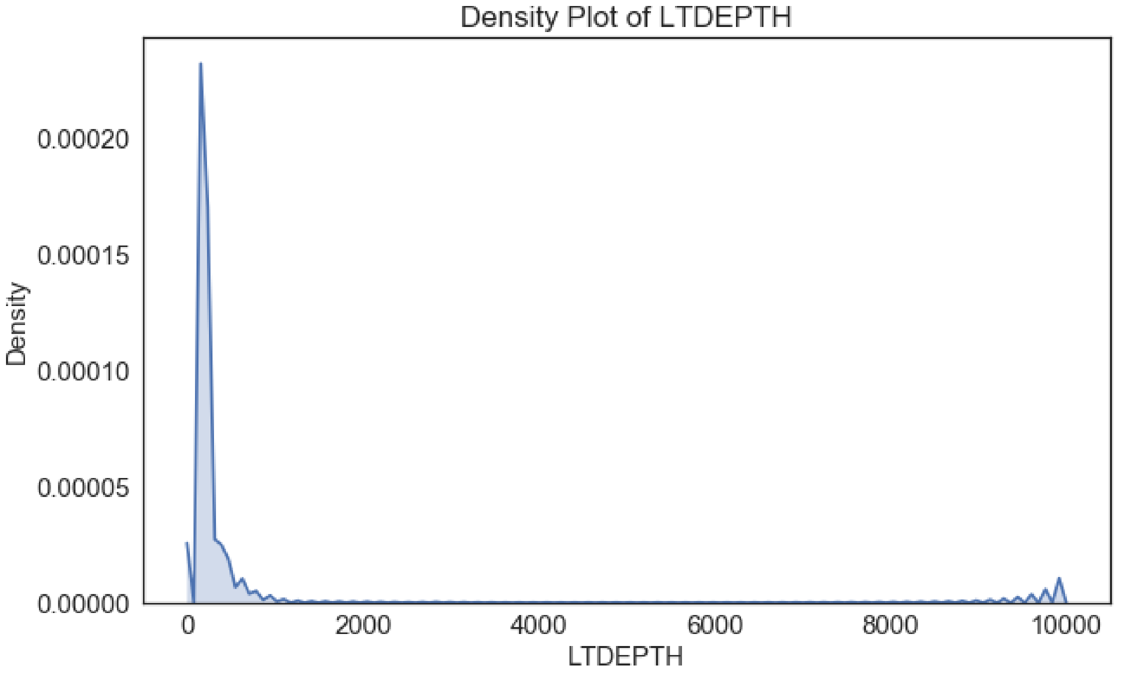
Field Name	**Type**	**Description**
LTFRONT	Numerical	Lot Frontage in Feet





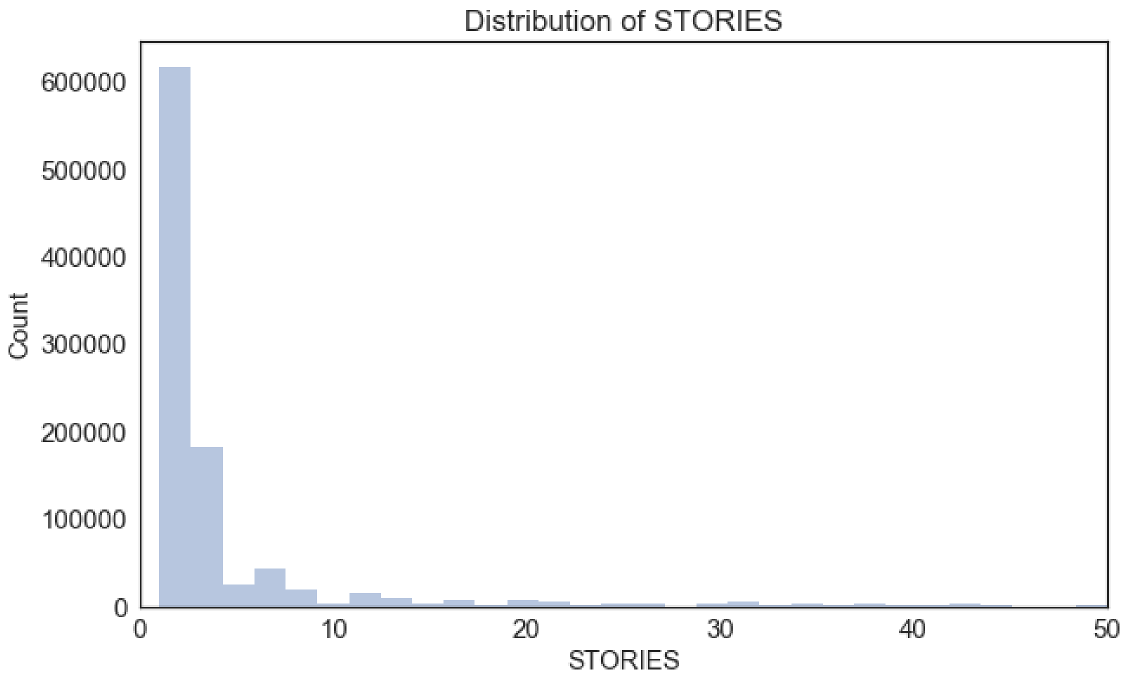
Field Name	**Type**	**Description**
LTDEPTH	Numerical	Lot Depth in Feet

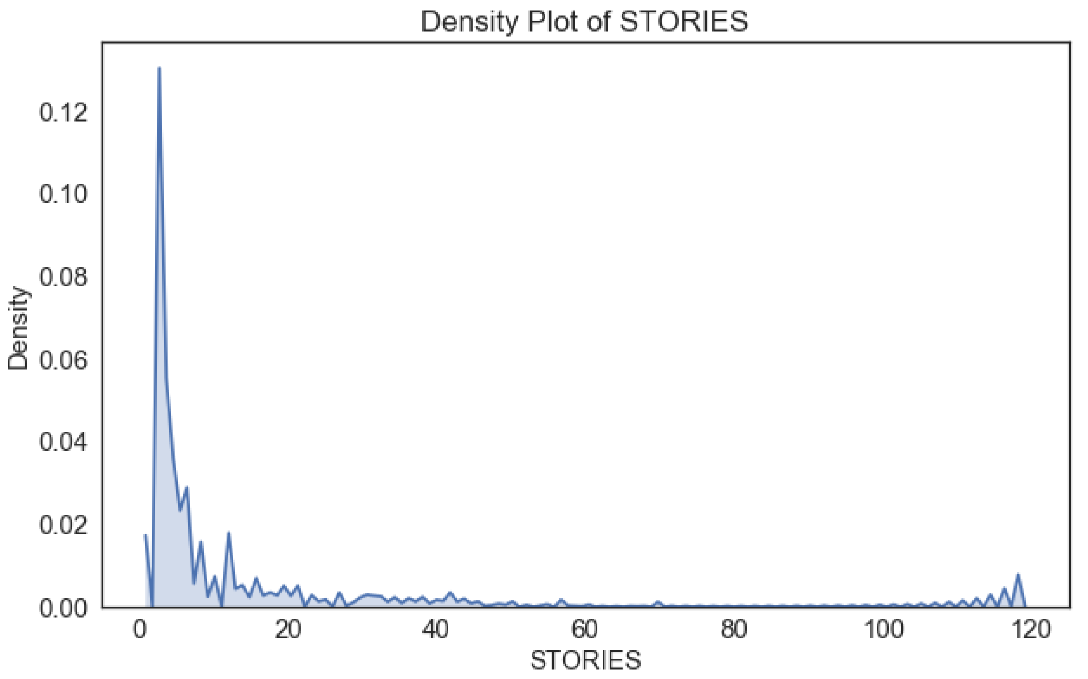




Field Name	Type	Description

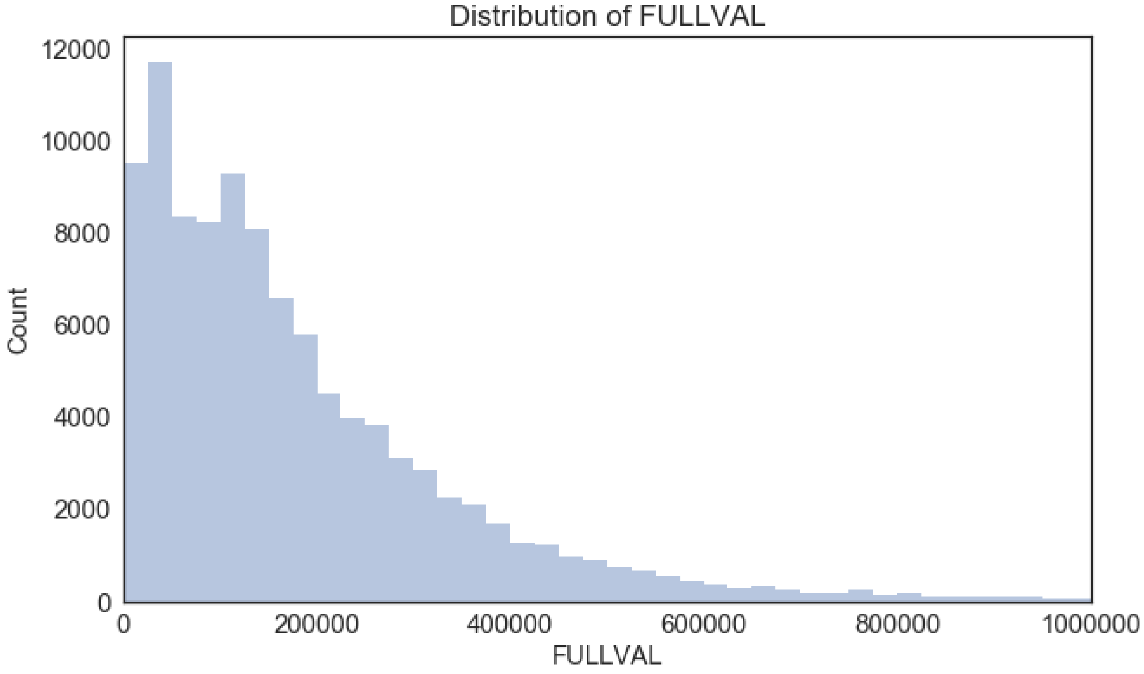
STORIES	Numerical	The Number of Stories for the Building (# of Floors)

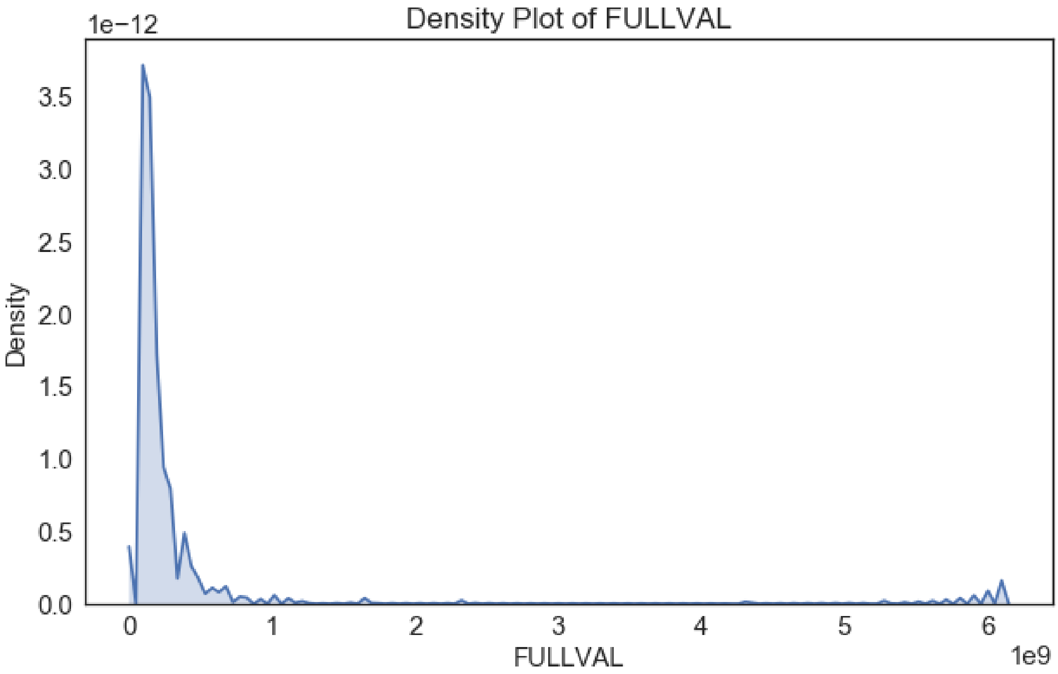




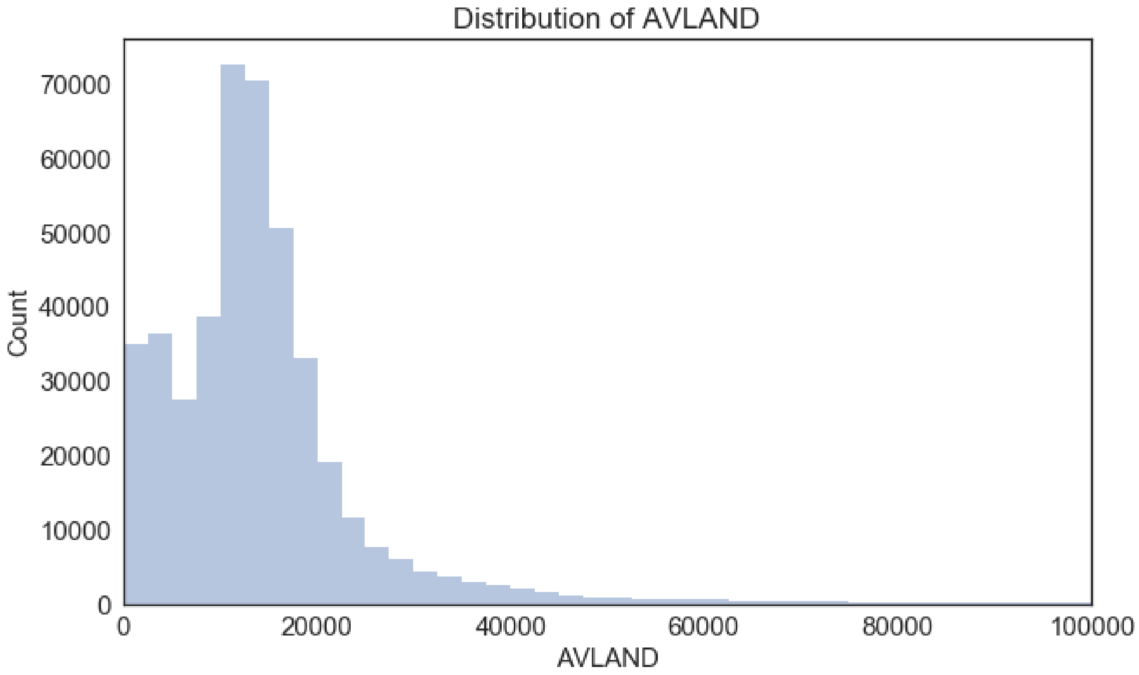
Field Name	**Type**	**Description**

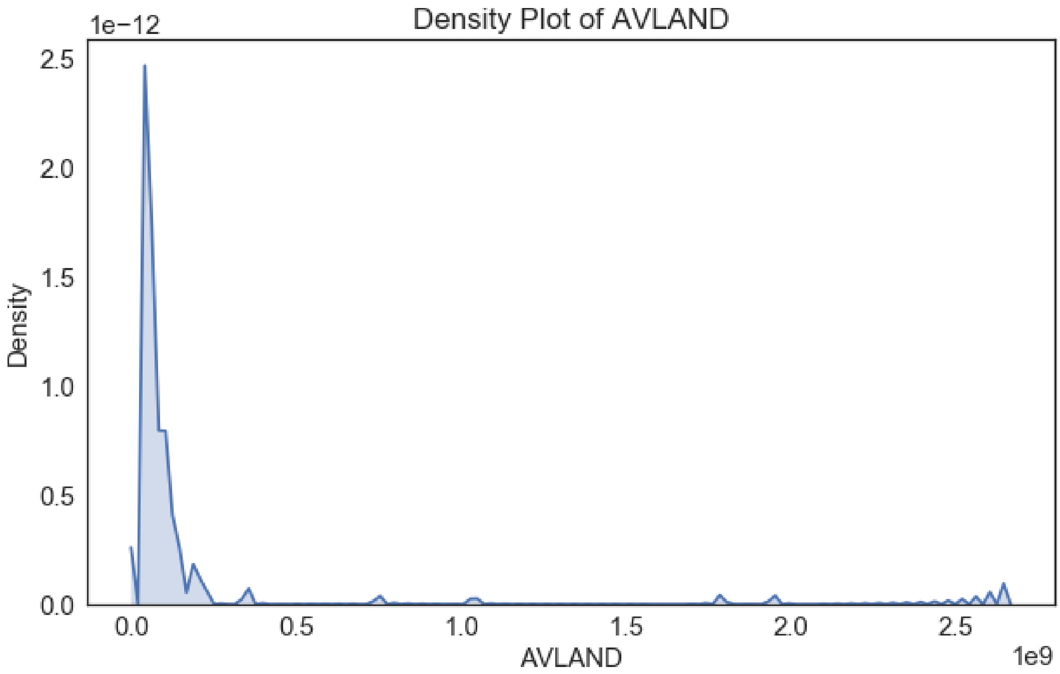
FULLVAL	Numerical	Total Market Value (Full Value) of the Property





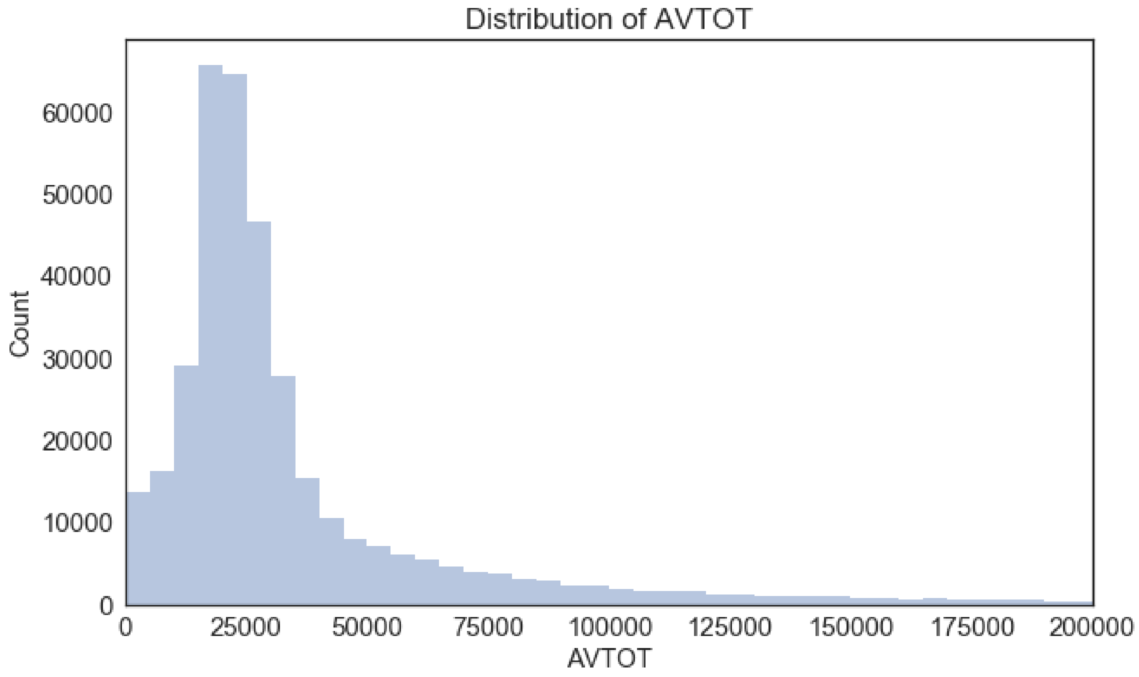
Field Name	**Type**	**Description**
AVLAND	Numerical	Current Assessed Land Value of the Property

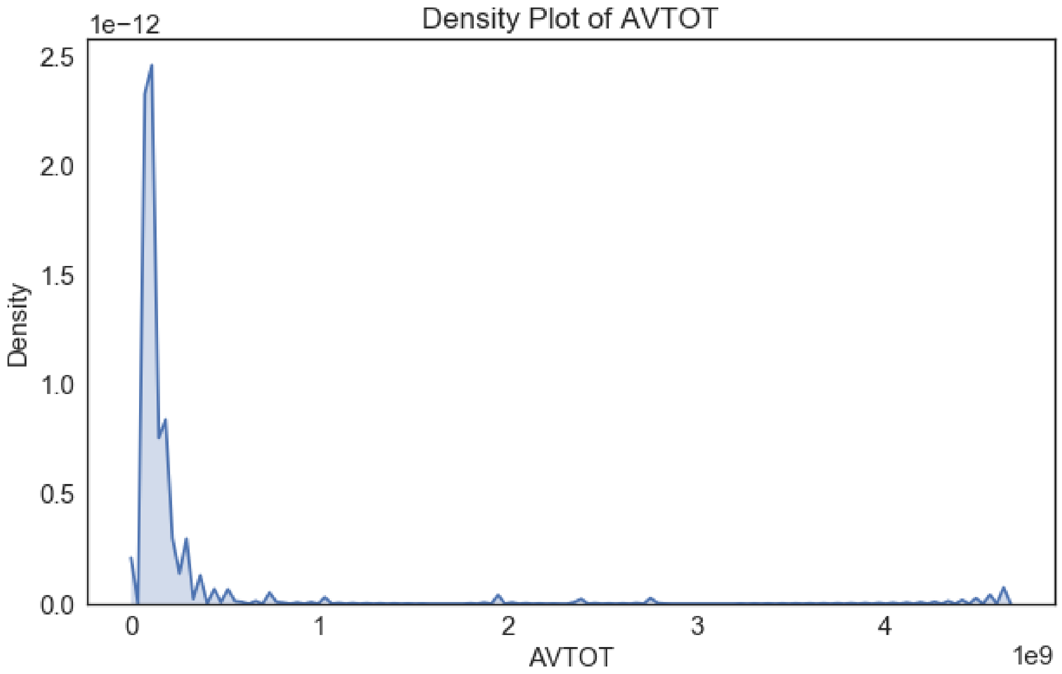




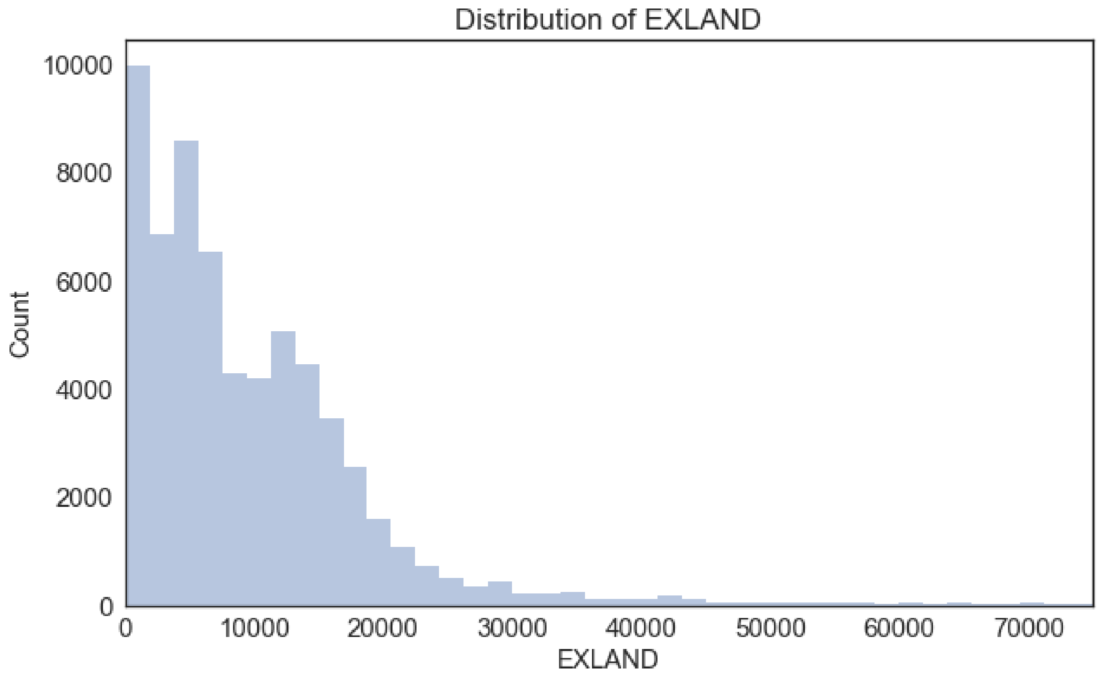
Field Name	**Type**	**Description**

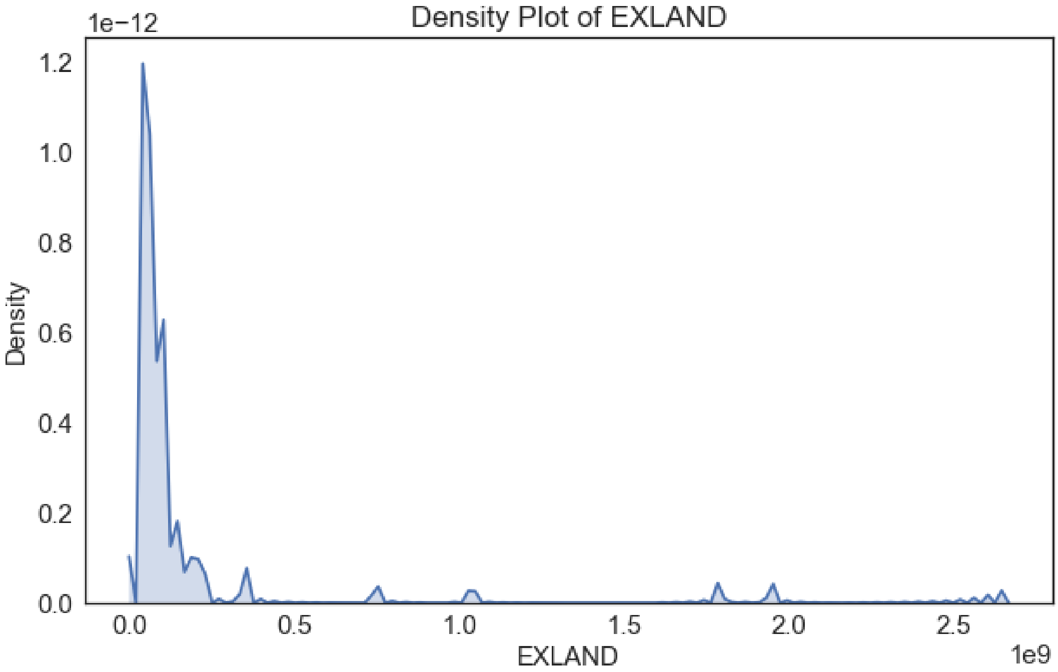
AVTOT	Numerical	Current Assessed Total Value of the Property



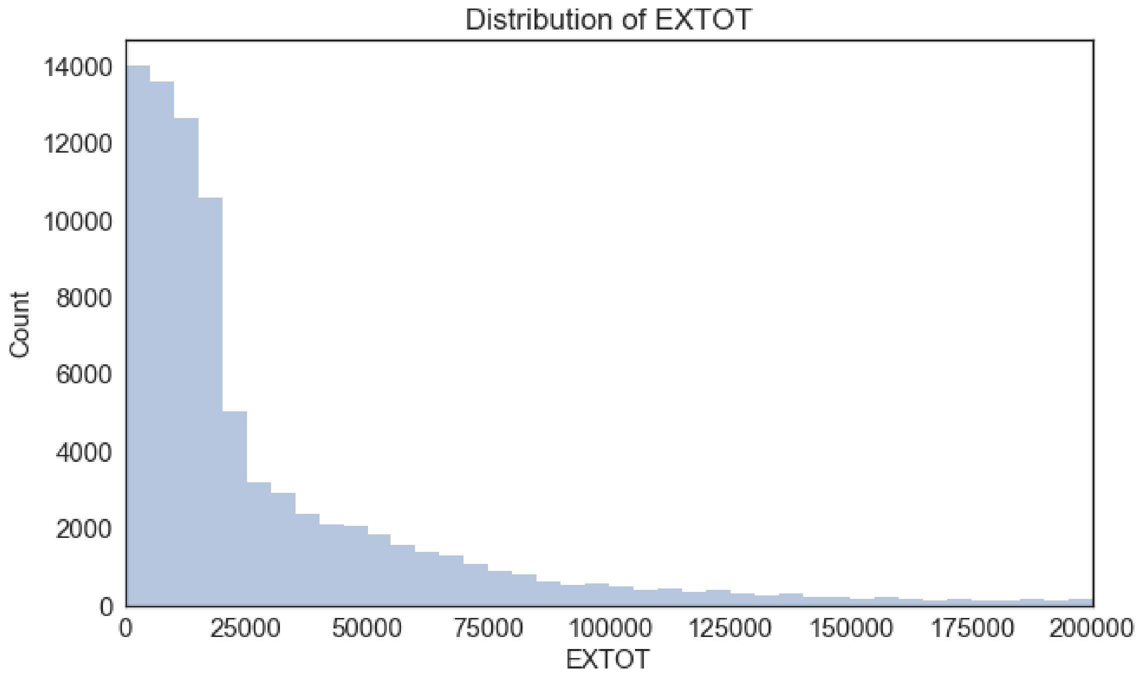


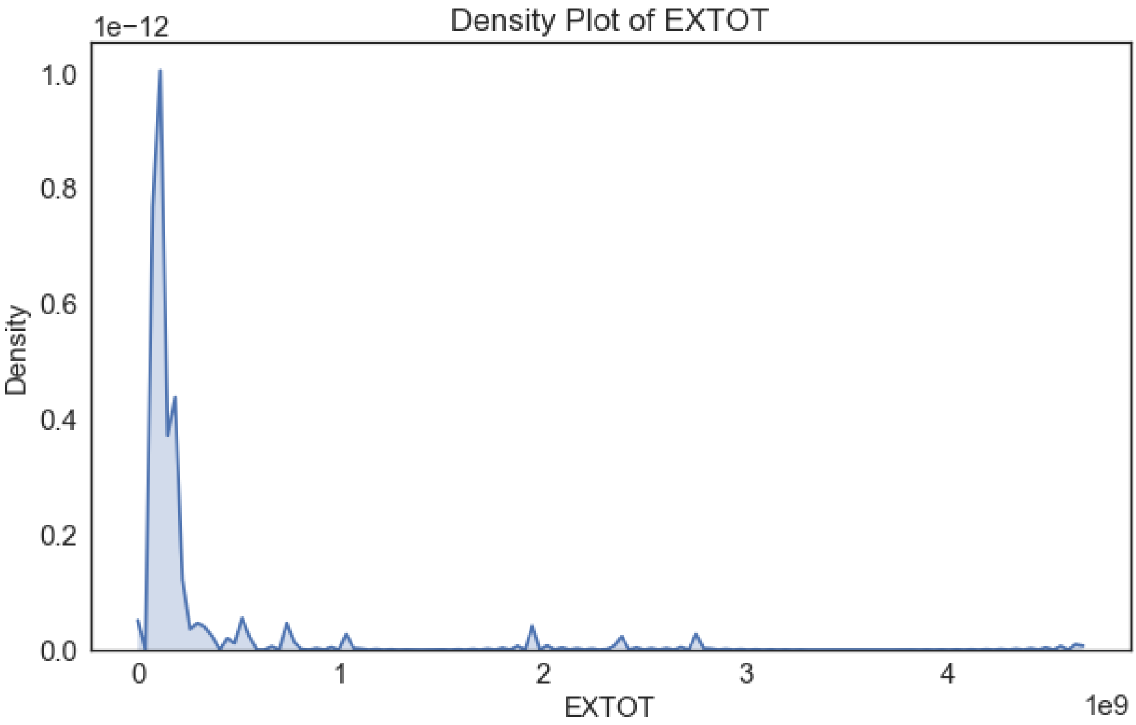
Field Name	Type	Description
EXLAND	Numerical	Current Exempt Land Value of the Property





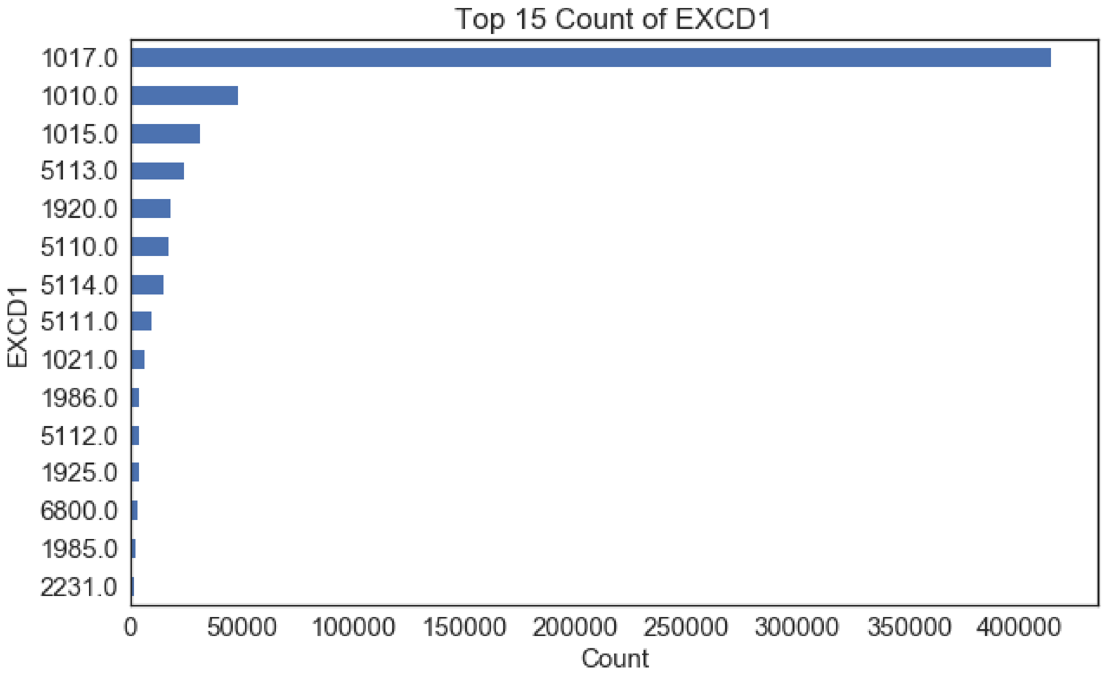
Field Name	**Type**	**Description**
EXTOT	Numerical	Current Exempt Total Value of the Property



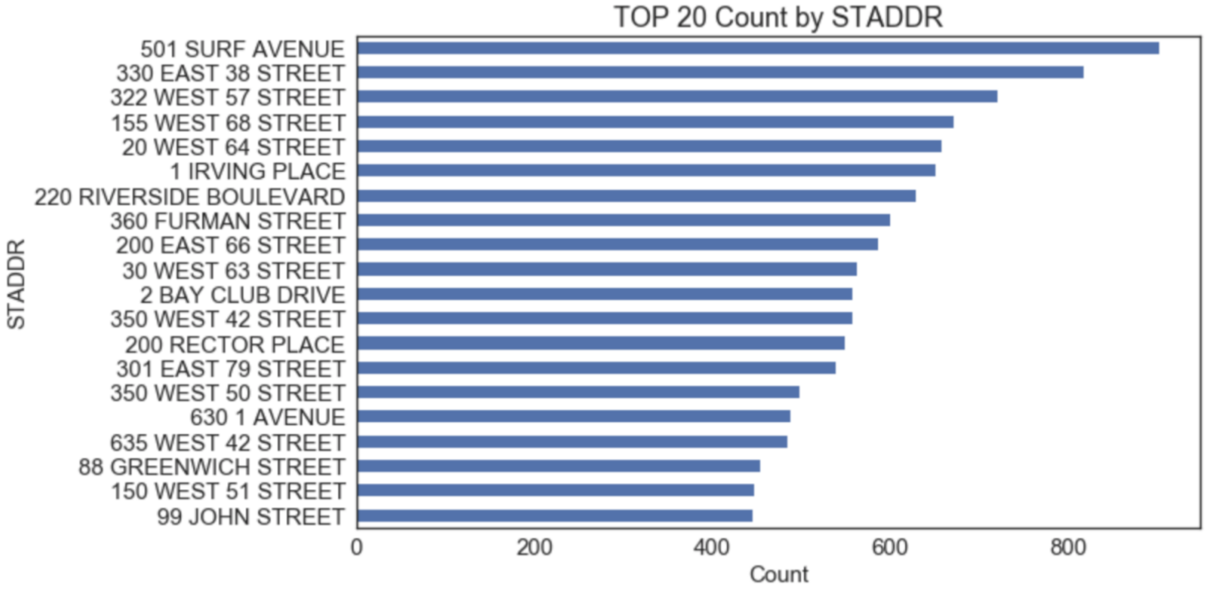


Field Name	Type	Description

EXCD1	Categorical	Exemption Code 1

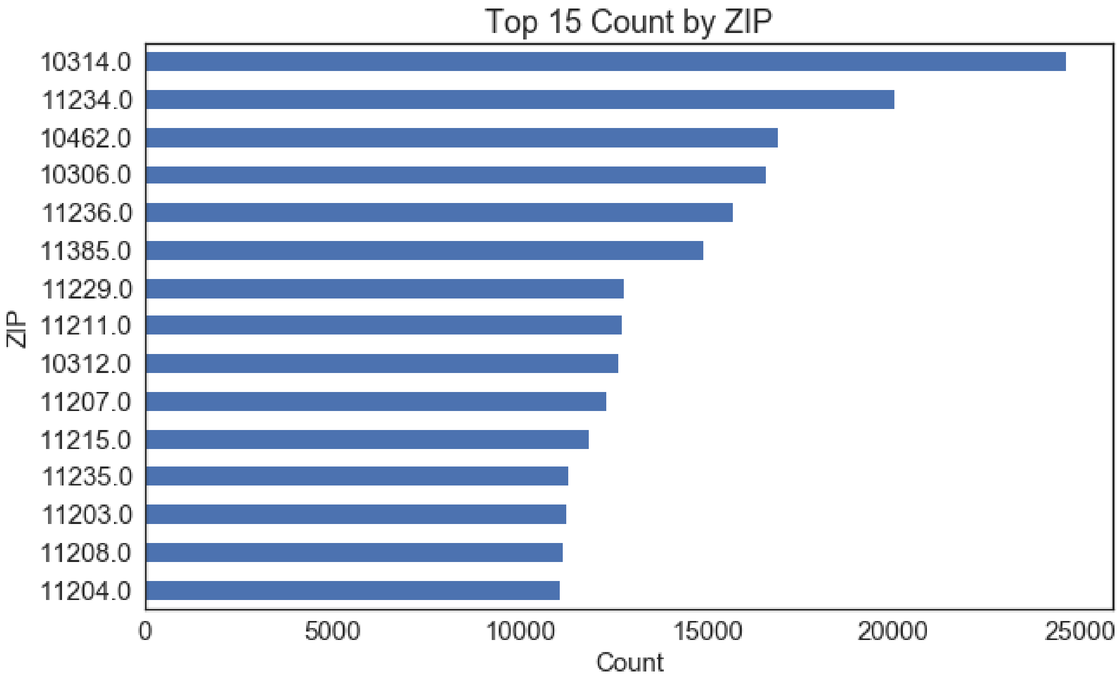


Field Name	**Type**	**Description**
STADDR	Categorical	Street Address for the Property

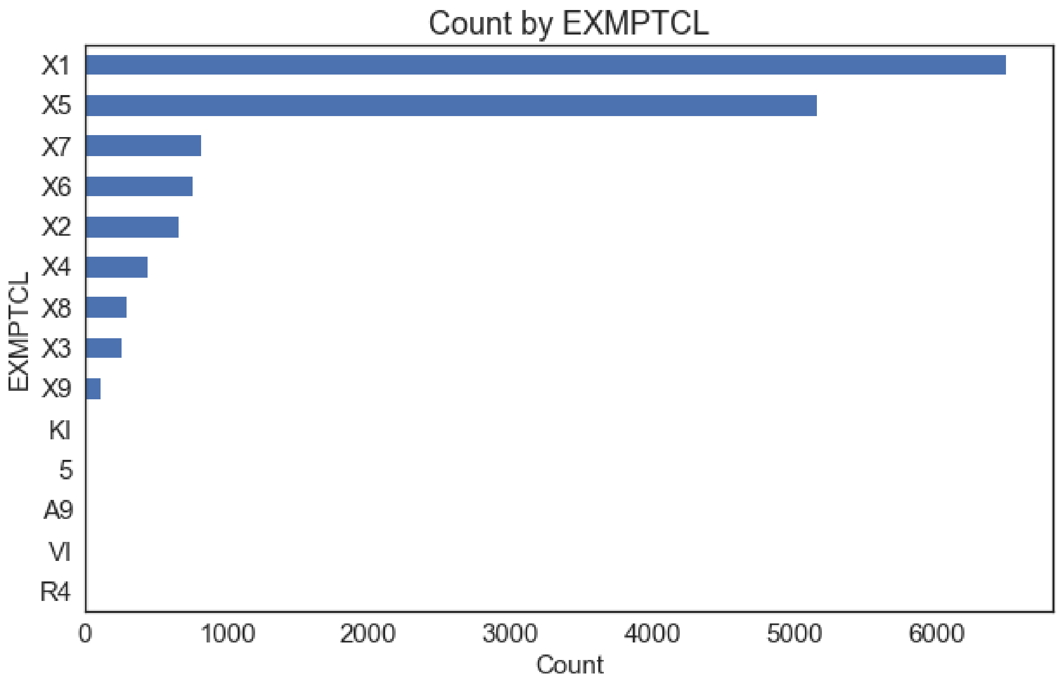


Field Name	**Type**	**Description**

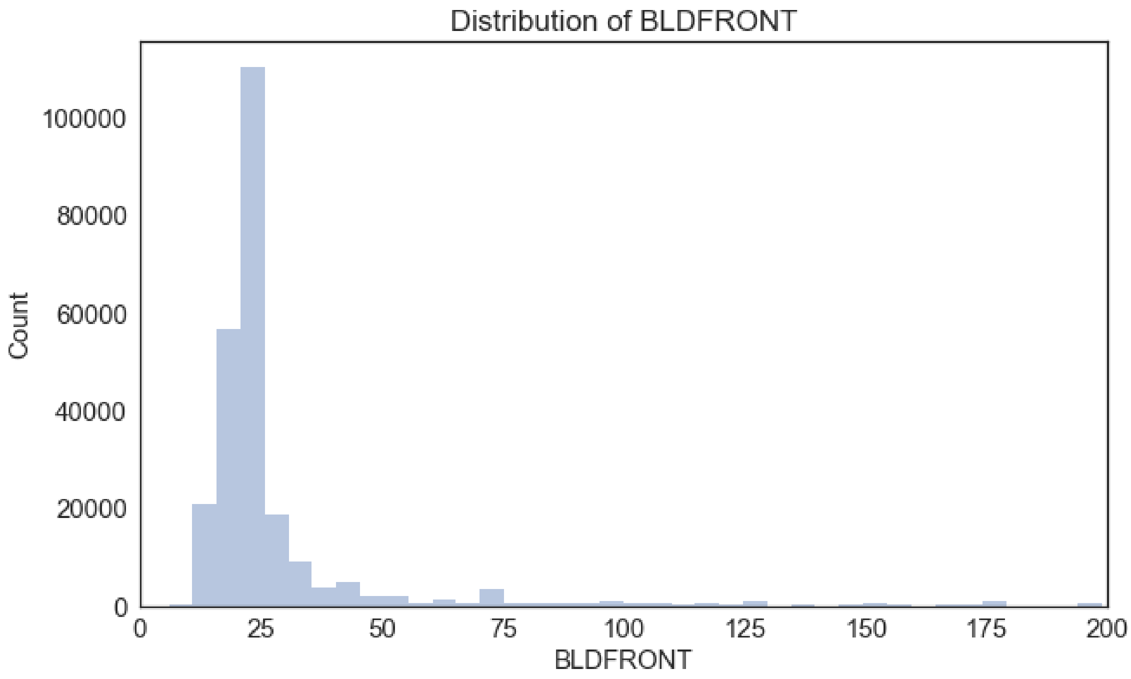
ZIP	Categorical	Postal Zip Code of the Property

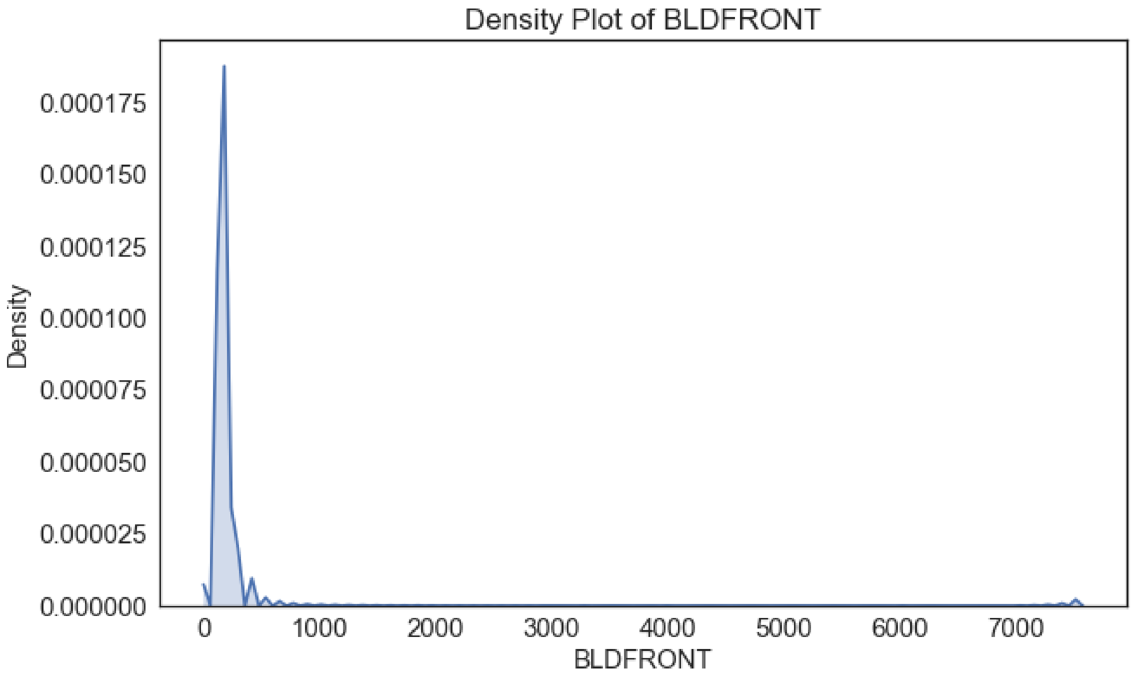


Field Name	**Type**	**Description**
EXMPTCL	Categorical	Exempt Class Used for Fully Exempt Properties Only

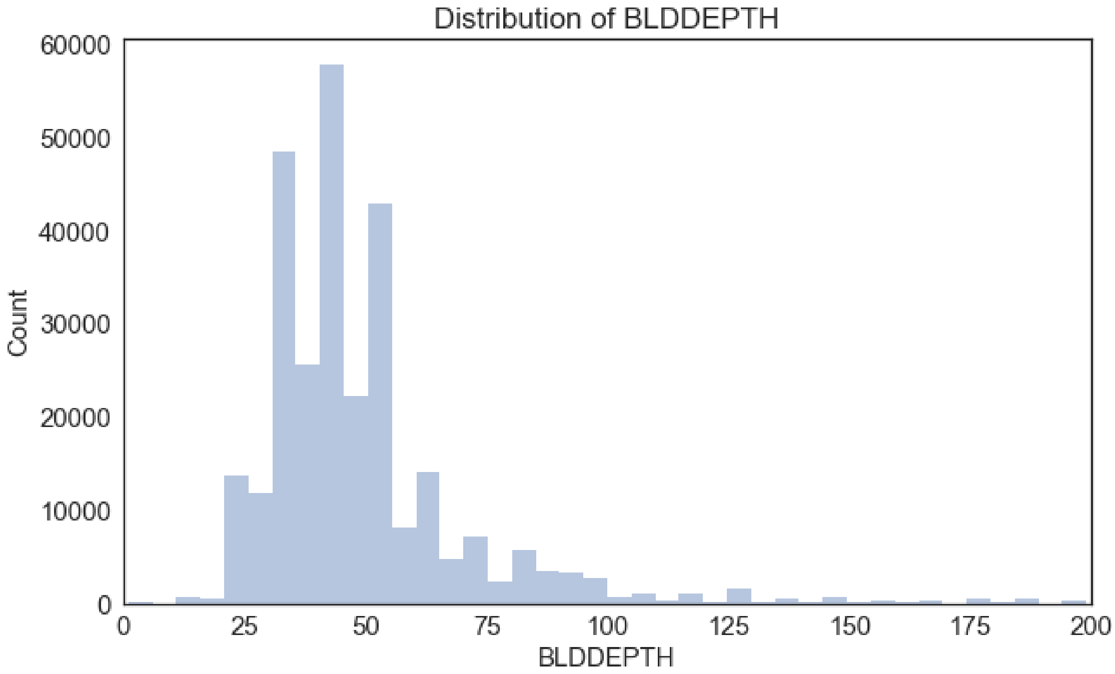


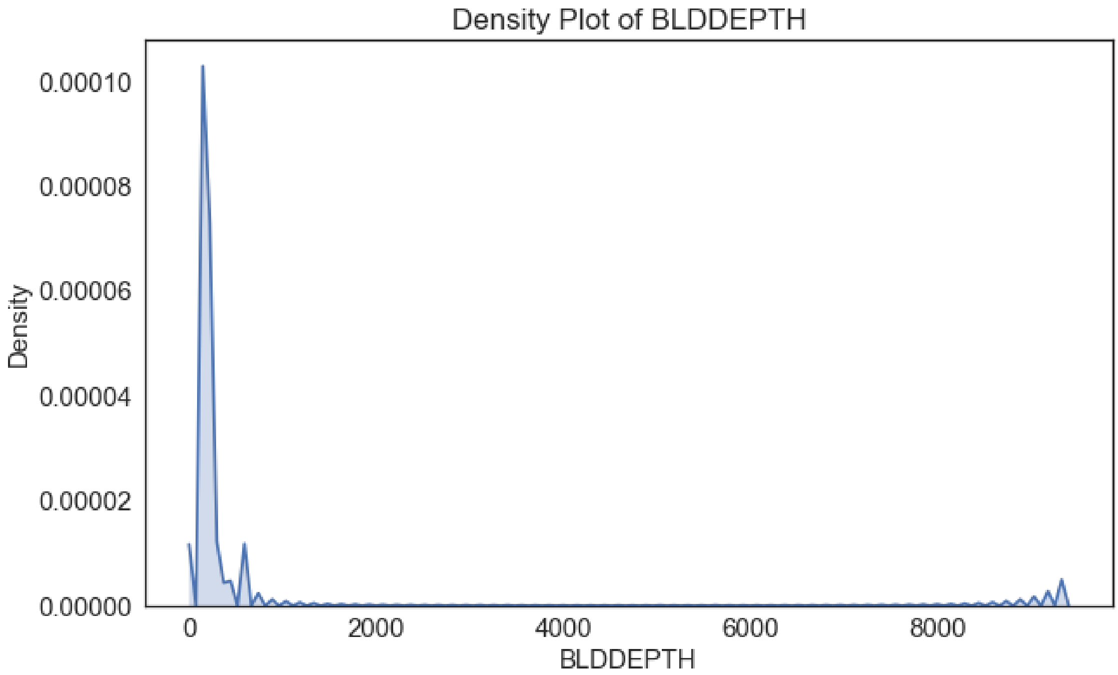
Field Name	Type	Description
BLDFRONT	Numerical	Building Frontage in Feet



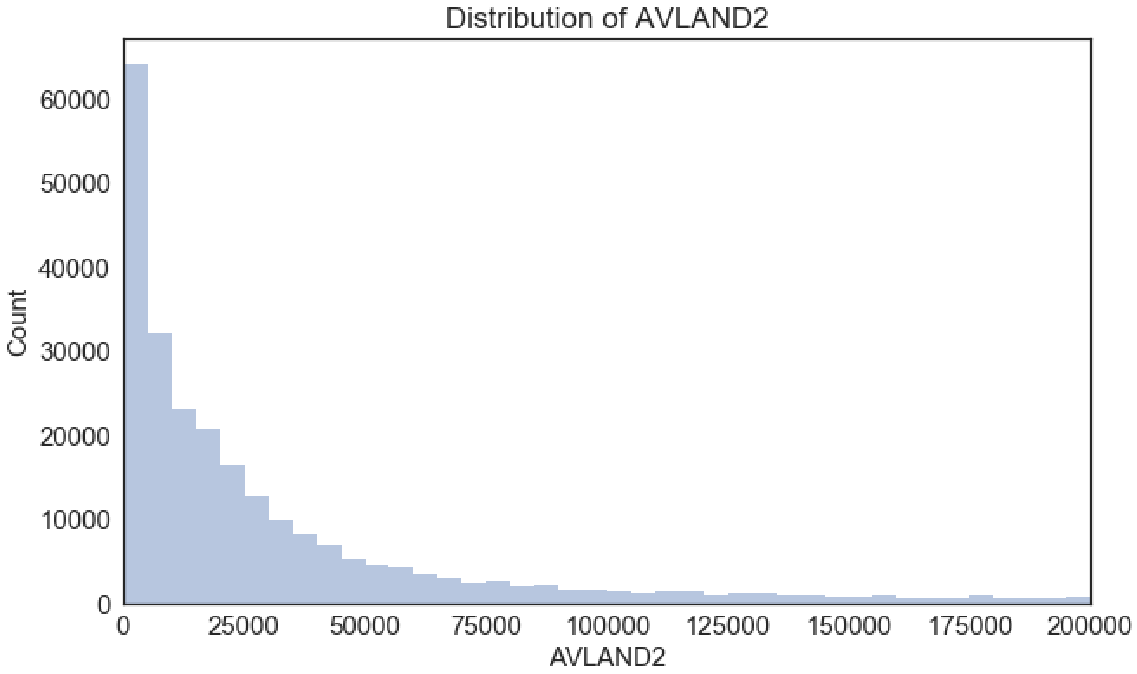


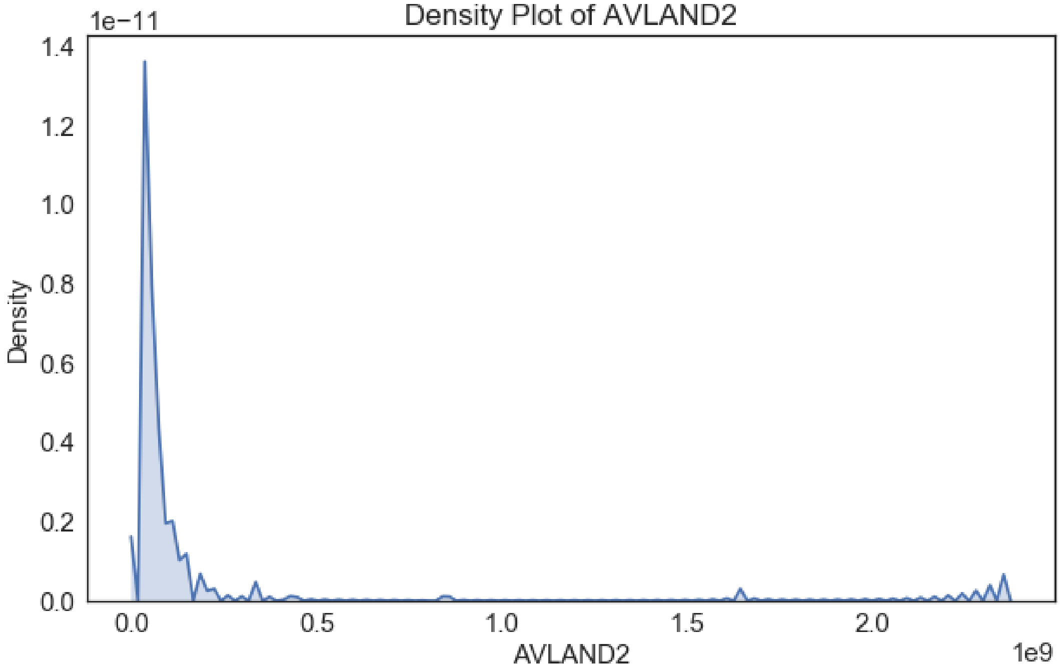
Field Name	**Type**	**Description**
BLDDEPTH	Numerical	Building Depth in Feet



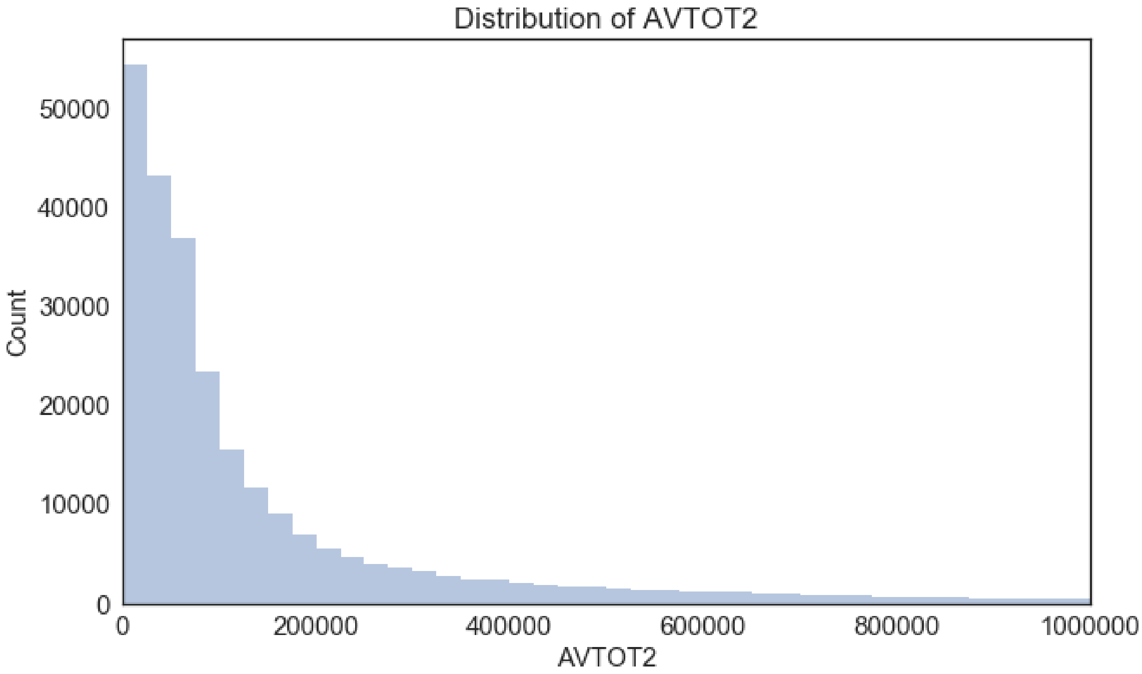


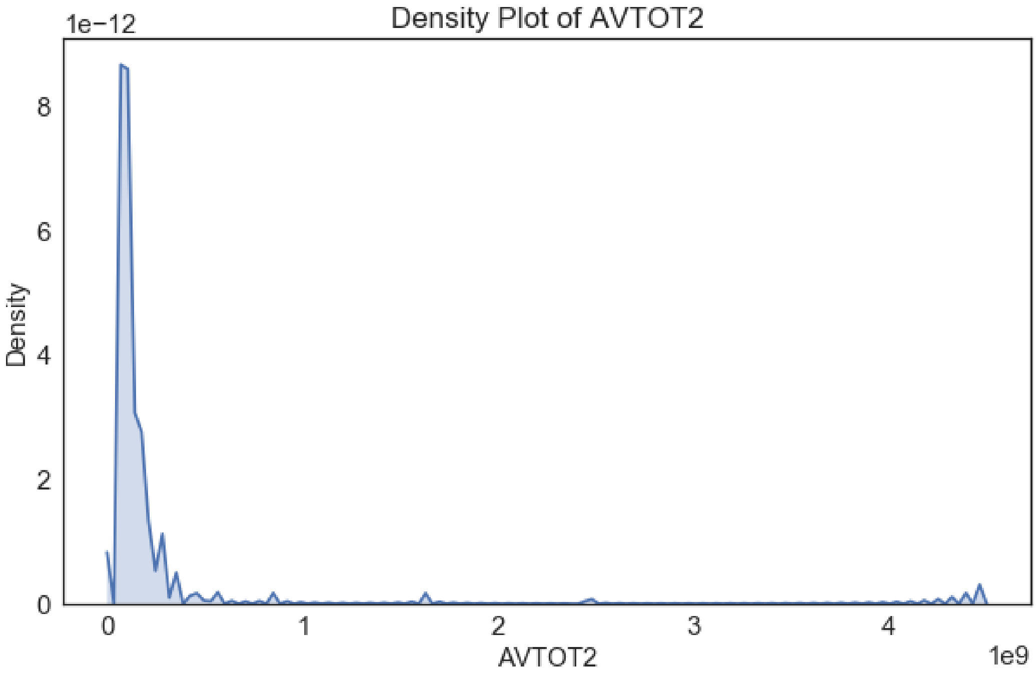
Field Name	Type	Description
AVLAND2	Numerical	Current Assessed Land Value 2 of the Property



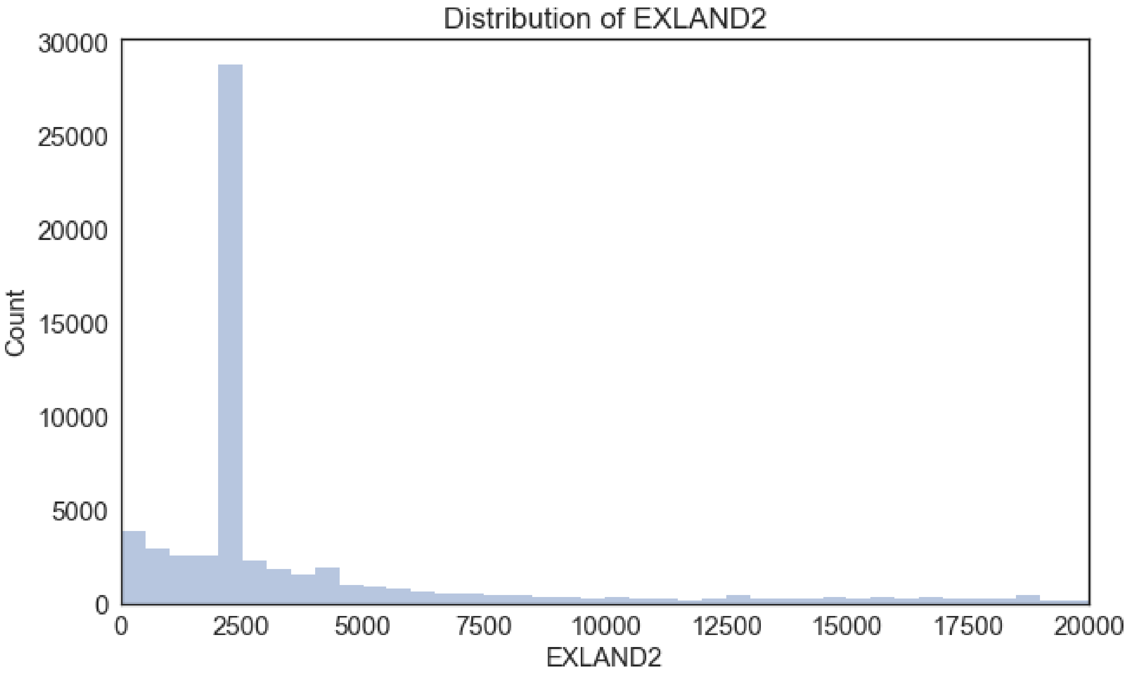


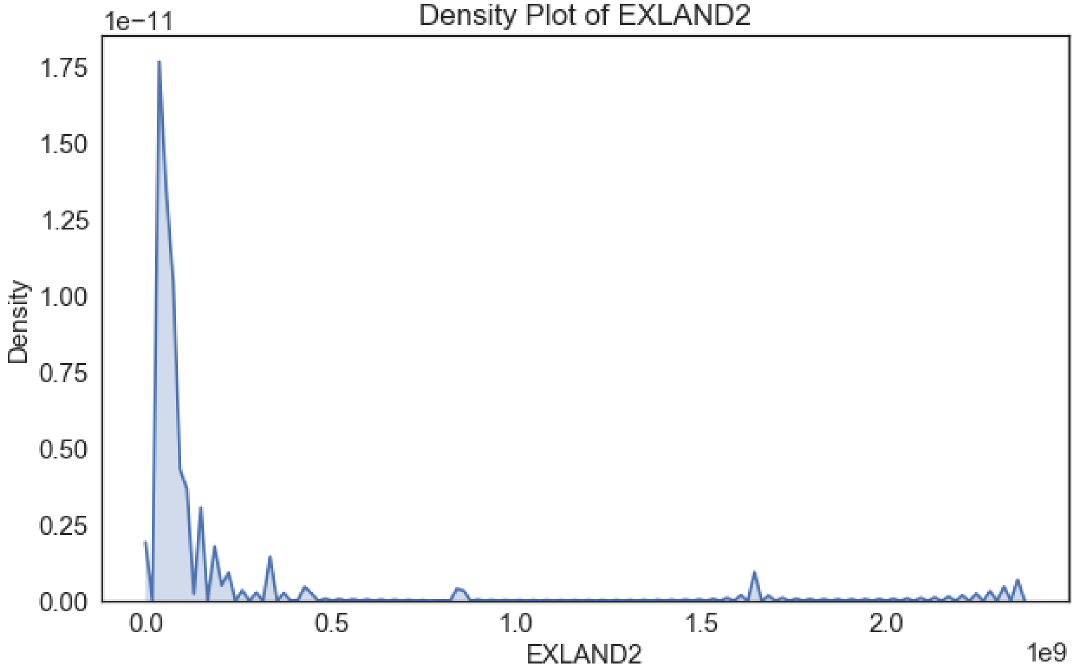
Field Name	Type	Description
AVTOT2	Numerical	Current Assessed Total Value 2 of the Property



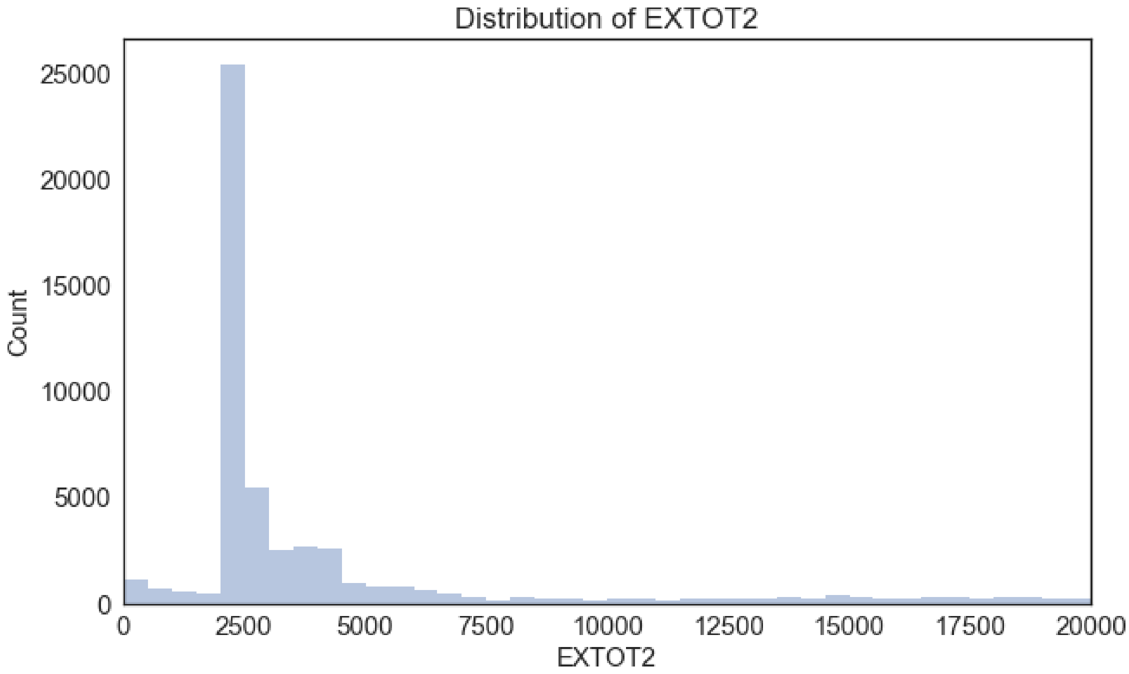


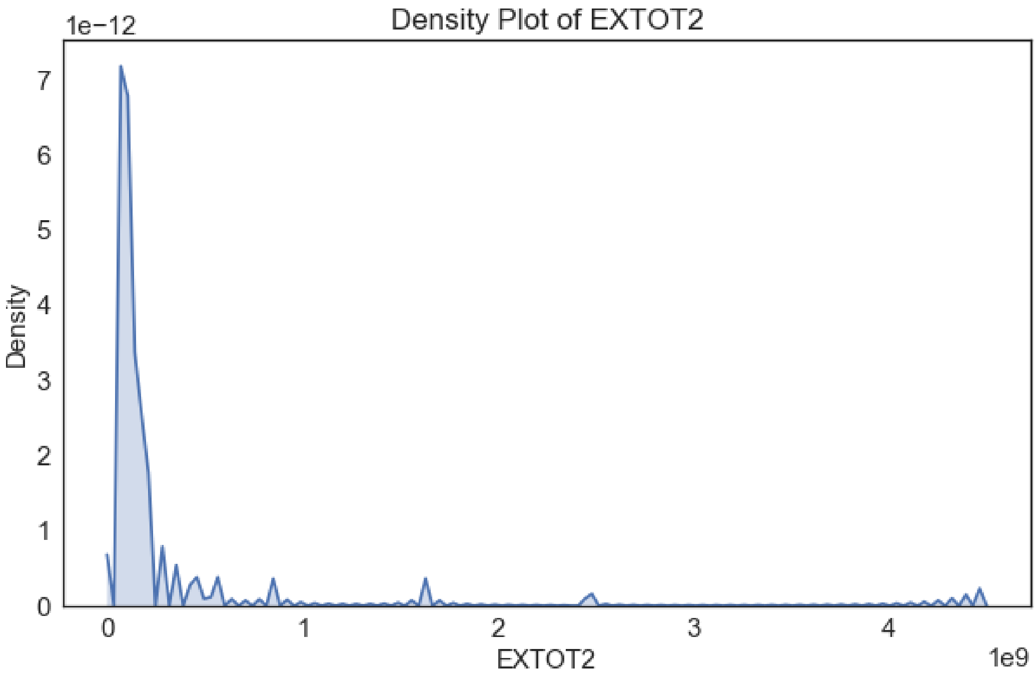
Field Name	Type	Description
EXLAND2	Numerical	Current Exempt Land Value 2 of the Property



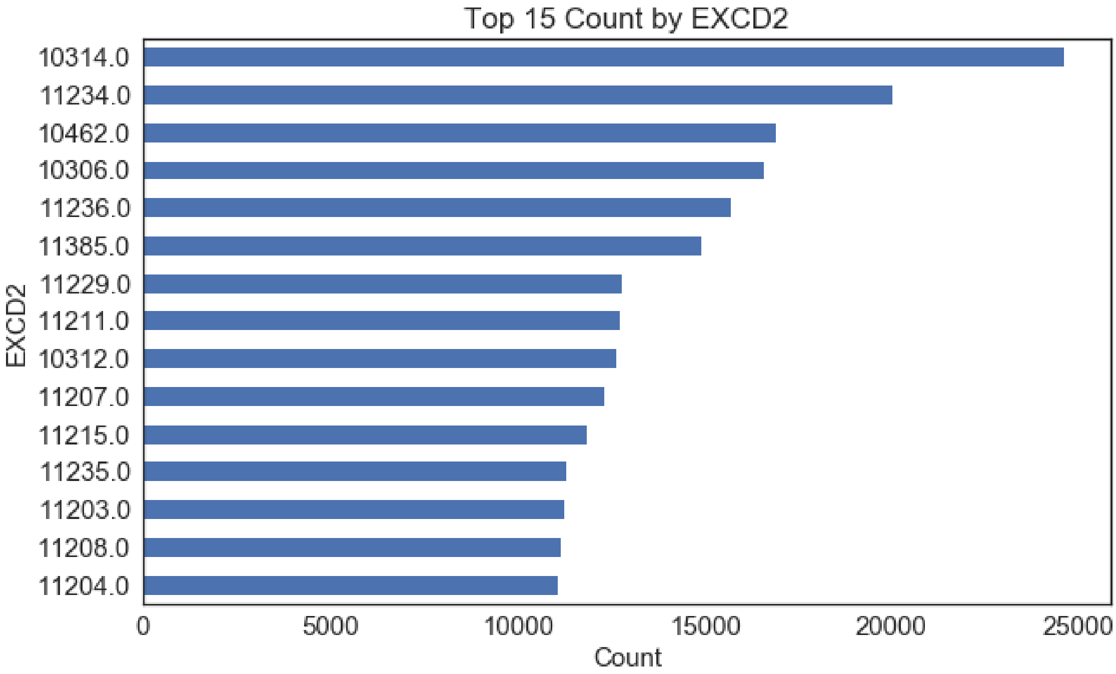


Field Name	**Type**	**Description**
EXTOT2	Numerical	Current Exempt Total Value 2 of the Property





Field Name	Type	Description
EXCD2	Categorical	Exemption Code 2



Field Name	**Type**	**Description**
BBLE	Categorical	Includes 10 or 11 digits. It is the concatenation of "BORO" code (1 digit), "BLOCK" code (5 digit), "LOT" code (4 digit) and "EASEMENT" code (1 digit if exists)
PERIOD	Categorical	Identical values ('FINAL') for all records. No missing value
YEAR	Categorical	Identical values ('2010/11') for all records. No missing value
VALTYPE	Categorical	Identical values ('AC-TR') for all records. No missing value

8. References

[1]http://www1.nyc.gov/site/finance/taxes/property-determining-your-market-value.page

[2]http://www1.nyc.gov/site/finance/taxes/property-determining-your-assessed-value.page

[3]http://www1.nyc.gov/site/finance/taxes/definitions-of-property-assessment-terms.page