# Fraud Detection in NYC Real Estate

Team 2
Christina Acojedo
Mrinal Gupta
Athanasios Rompokos
Guyane Valian
Min Zhu

Project 1 Report
DSO562 | Spring 2020
February 13, 2020

# Table of Contents

Table of Contents ..... 2

Executive Summary ..... 3

Description of Data ..... 4

Data Cleaning ..... 11

Variable Creation ..... 17

Dimensionality Reduction ..... 19

Algorithms ..... 24

Results ..... 25

Conclusions ..... 30

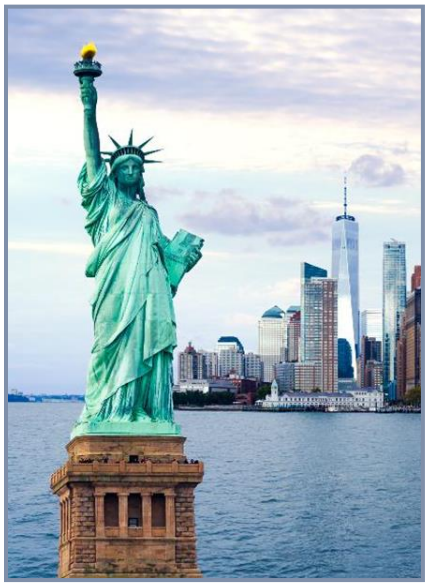
Appendix: Data Quality Report (DQR) ..... 31

1 Data Overview ..... 31

2 Summary Tables ..... 32

3 Data Field Exploration ..... 34

# Executive Summary



This report analyzes New York City's (NYC's) real estate data to specifically identify property tax fraud. The main indicators of property tax fraud were property tax assessments that were too high or too low. Given a property dataset of 1,070,994 records and 32 data fields, we first described, visualized, and filled in missing values for each variable. Second, 45 additional variables were created in order to create the most accurate algorithm. Next, we used dimensionality reduction techniques to refine our dataset. Finally, we used (principal component analysis (PCA) and an autoencoder) to obtain two separate fraud scores. The scores were combined and then ranked to get a final fraud score.

Each of the 32 variables in the dataset were analyzed. For each variable, the following values were calculated and summarized in a table: mean, standard deviation, min, max, number of populated values, percent populated, number unique values, and number of records equal to zero. We also provided a histogram or table for each field to get a better feel for its distribution. Many of the variables had a right skewed distribution. To determine if property tax assessments were too high or too low, we focused on property value variables (FULLVAL, AVLAND, AVTOT), property size variables (LTFRONT, LTDEPTH, BLDDEPTH, BLDFRONT, STORIES), and a location variable (ZIP). In analyzing each significant variable, we realized there were missing values or records with zero values. For each variable, we created rules to fill in these ineffective records based on its distribution and definition. Since many distributions were right-skewed, medians were used instead of means for these rules.

Now that the significant variables were identified and cleaned, we created 45 new variables in order to create the most effective model. We focused on size and created three variables for the lot area, building area, and building volume. These variables helped us assess property values given the size of each property. Each of these property value variables were normalized by the three size variables to create nine variables. Next, we calculated grouped averages of these nine variables. The five groups were zip5, zip3, tax class, borough, and all (no group). Finally, for each group, we calculated the average for each of the nine variables, which resulted in our final forty-five variables.

We then used dimensionality reduction to refine our 45 variables and built two different algorithms to get two different fraud scores. The first model used PCA and the second model used an autoencoder. The PCA performs an orthogonal transformation that changes values linearly uncorrelated variables. Furthermore, we trained a neural network autoencoder with three layers, an input layer of dimension eight (equal to the dataset dimension after performing PCA), a hidden layer of five nodes (for further dimensionality reduction) and an output layer of eight nodes to get our second fraud score. The two models were then combined into one final fraud score using their harmonic mean. Next, the final results were ranked from highest fraud score to lowest fraud scores. Finally, we further analyzed the top ten fraud scores to determine which variable looked unusual. The list has to be further analyzed by property tax experts to determine if the assessments are actually fraudulent or not.

# Description of Data

The dataset used for the analysis within this report is a public dataset produced by the NYC Department of Finance (DOF) and provided by NYC Open Data at https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8. It consists of property valuations and assessments on NYC real estate, and was intended for calculating property taxes and granting eligible properties exemptions and/or abatements during NYC's 2010/2011 fiscal year. Although it was originally created in September 2011, the dataset was updated more recently in September 2018.

The dataset contains a total of 1,070,994 property records and 32 data fields. Within the 32 data fields, there are 14 numerical data fields and 18 categorical data fields. Summary tables for the numerical and categorical data fields can be seen in Tables 1 and 2, respectively.

| Data Field | Num Records w/ a Value | Percent Populated | Num Unique Values | Num Records w/ a Value of Zero | Mean | Standard Deviation | Min | Max |
|---|---|---|---|---|---|---|---|---|
| **LTFRONT** | 1,070,994 | 100% | 1,297 | 169,108 | 37 ft | 74 ft | 0 ft | 9,999 ft |
| **LTDEPTH** | 1,070,994 | 100% | 1,370 | 170,128 | 89 ft | 76 ft | 0 ft | 9,999 ft |
| **STORIES** | 1,014,730 | 94.7% | 111 | 0 | 5.0 | 8 | 1 | 119 |
| **FULLVAL** | 1,070,994 | 100% | 109,324 | 13,007 | $874,265 | $11,582,431 | $0 | $6,150,000,000 |
| **AVLAND** | 1,070,994 | 100% | 70,921 | 13,009 | $85,068 | $4,057,260 | $0 | $2,668,500,000 |
| **AVTOT** | 1,070,994 | 100% | 112,914 | 13,007 | $227,238 | $6,877,529 | $0 | $4,668,308,947 |
| **EXLAND** | 1,070,994 | 100% | 33,419 | 491,699 | $36,424 | $3,981,576 | $0 | $2,668,500,000 |
| **EXTOT** | 1,070,994 | 100% | 64,255 | 432,572 | $91,187 | $6,508,403 | $0 | $4,668,308,947 |
| **BLDFRONT** | 1,070,994 | 100% | 612 | 228,815 | 23 ft | 35 ft | 0 ft | 7,575 ft |
| **BLDDEPTH** | 1,070,994 | 100% | 621 | 228,853 | 39 ft | 42 ft | 0 ft | 9,393 ft |
| **AVLAND2** | 282,726 | 26.4% | 58,591 | 0 | $246,236 | $6,178,963 | $3 | $2,371,005,000 |
| **AVTOT2** | 282,732 | 26.4% | 111,360 | 0 | $713,911 | $11,652,529 | $3 | $4,501,180,002 |
| **EXLAND2** | 87,449 | 8.2% | 22,195 | 0 | $351,236 | $10,802,213 | $1 | $2,371,005,000 |
| **EXTOT2** | 130,828 | 12.2% | 48,348 | 0 | $656,769 | $16,072,510 | $7 | $4,501,180,002 |

**Table 1. Summary Table for the Numerical Data Fields.**

| Data Field | Num Records w/ a Value | Percent Populated | Num Unique Values | Most Common Value |
|---|---|---|---|---|
| **RECORD** | 1,070,994 | 100% | 1,070,994 | N/A |
| **BBLE** | 1,070,994 | 100% | 1,070,994 | N/A |
| **B** | 1,070,994 | 100% | 5 | 4 |
| **BLOCK** | 1,070,994 | 100% | 13,984 | 3944 |
| **LOT** | 1,070,994 | 100% | 6,366 | 1 |
| **EASEMENT** | 4,636 | 0.43% | 12 | E |
| **OWNER** | 1,039,249 | 97.04% | 863,346 | PARKCHESTER PRESERVAT |
| **BLDGCL** | 1,070,994 | 100% | 200 | R4 |
| **TAXCLASS** | 1,070,994 | 100% | 11 | 1 |
| **EXT** | 354,305 | 33.08% | 3 | G |
| **EXCD1** | 638,488 | 59.62% | 129 | 1017 |
| **STADDR** | 1,070,318 | 99.94% | 839,280 | 501 SURF AVENUE |
| **ZIP** | 1,041,104 | 97.21% | 196 | 10314 |
| **EXMPTCL** | 15,579 | 1.45% | 14 | X1 |
| **EXCD2** | 92,948 | 8.68% | 60 | 1017 |
| **PERIOD** | 1,070,994 | 100% | 1 | FINAL |
| **YEAR** | 1,070,994 | 100% | 1 | 2010/11 |
| **VALTYPE** | 1,070,994 | 100% | 1 | AC-TR |

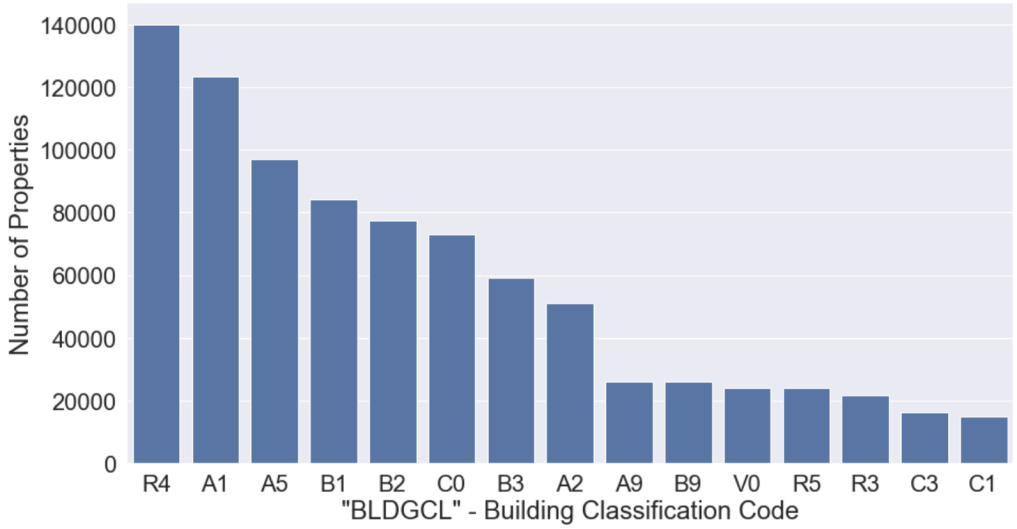
**Table 2. Summary Table for the Categorical Data Fields.**

Amongst the 32 data fields, we determined the most critical data fields in detecting potential fraud cases were those relating to the property record's location (B (borough), ZIP, and BLOCK), dimensions (LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, and STORIES), tax classification codes (BLDGCL and TAXCLASS), and monetary values (FULLVAL, AVLAND, and AVTOT). Some of the relevant depictions amongst the critical data fields are provided below in the order in which they appear. For a full description of all the data fields in the dataset, please see Appendix for the Data Quality Report (DQR).

## BLDGCL

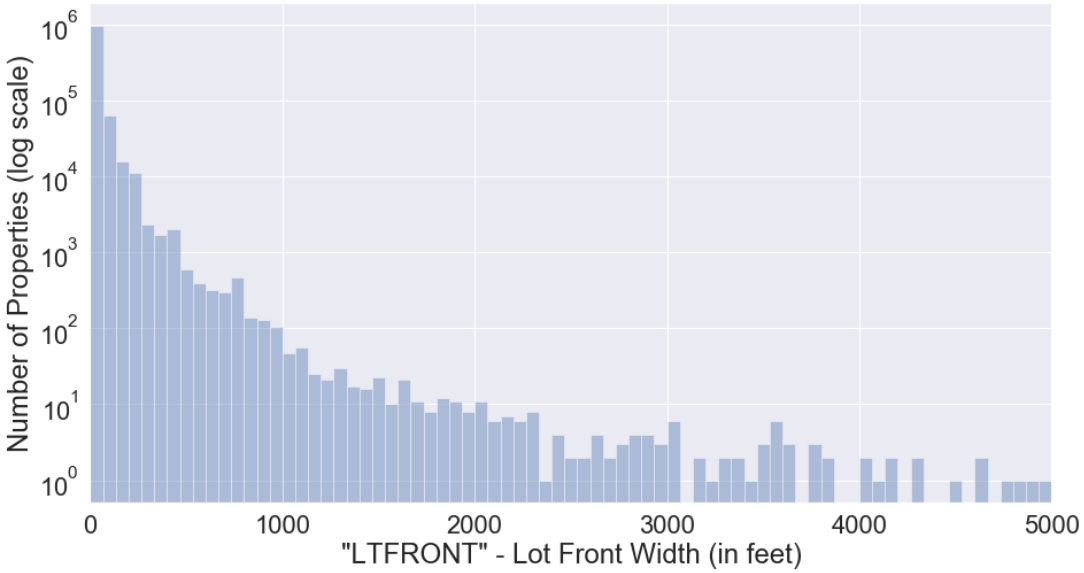
A 2-character length alphanumeric data field containing the NYC building classification code. It is important to note that there is a direct correlation between the building classification code and the tax classification code. The first character in the building classification code is a letter and the second

character is a number. All records in the dataset contain a building classification code. The bar chart below provides the top 15 building classification codes with the most property records listed in the dataset. Of the 200 unique values in this data field, R4 is the most common code.



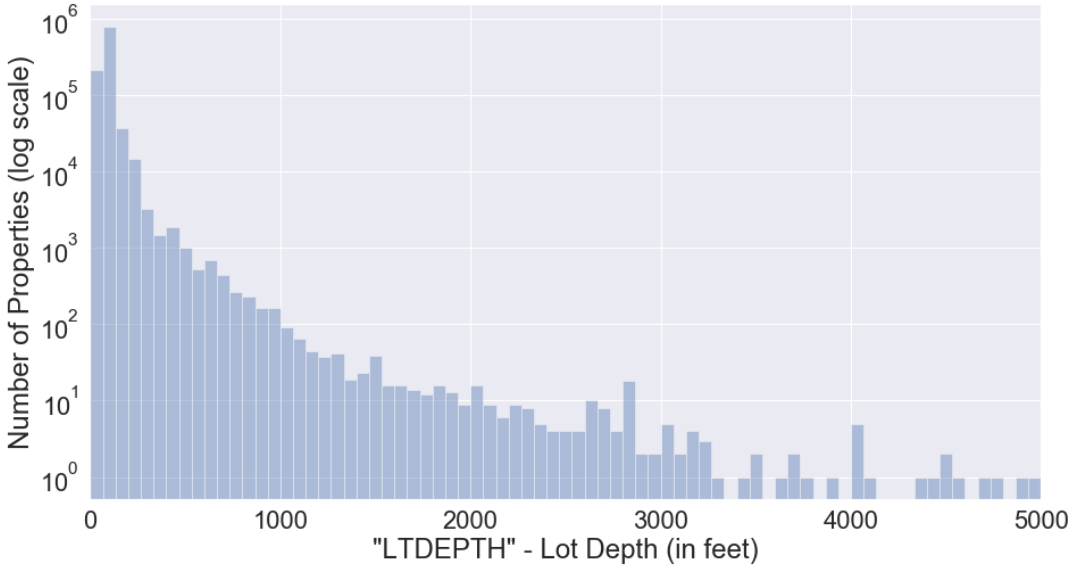

## LTFRONT

An integer data field containing the measurement of the property lot's front width in feet. All property records in the dataset contain a front width lot measurement. The distribution plot below shows the front width lot measurements up to 5,000 feet for the property records in the dataset. The widest measurement for the front width is 9,999 feet. The most common measurement is zero feet with 169,108 property records having this erroneous value. Excluding those property records with a measurement of zero feet, a measurement of 20 feet is the next most common value. Lastly, a majority of the property records have a measurement of 50 feet or less.



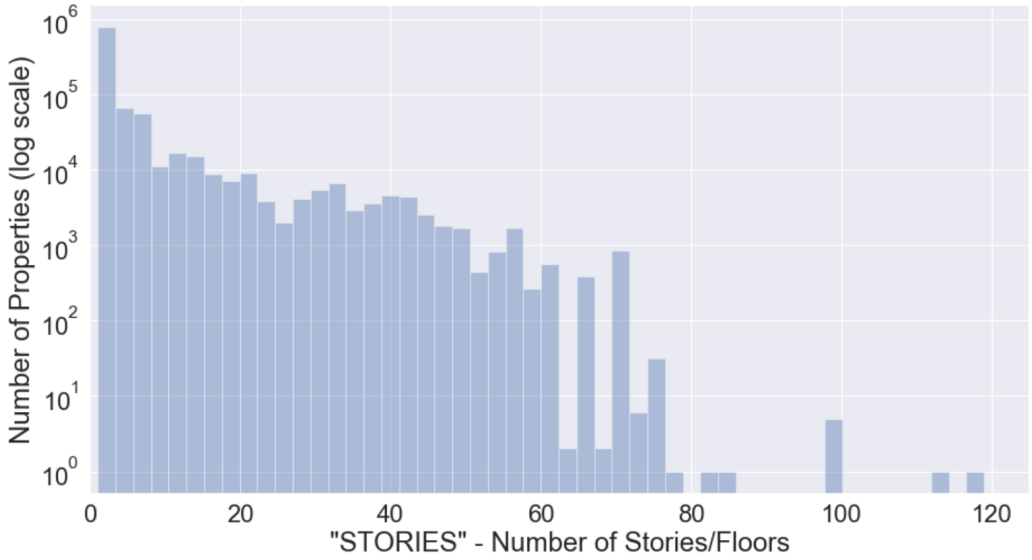
## LTDEPTH

An integer data field containing the depth of the lot measured in feet. All property records in the dataset contain a measurement for the property lot's depth. The distribution plot below shows the property lot depth measurements up to 5,000 feet for the property records in the dataset. The longest measurement for a property lot's depth is 9,999 feet. The most common property lot depth is 100 feet and a majority of the property records have a measurement of 102 feet or less. The second most common property lot depth is zero feet with 170,128 property records having this erroneous value.



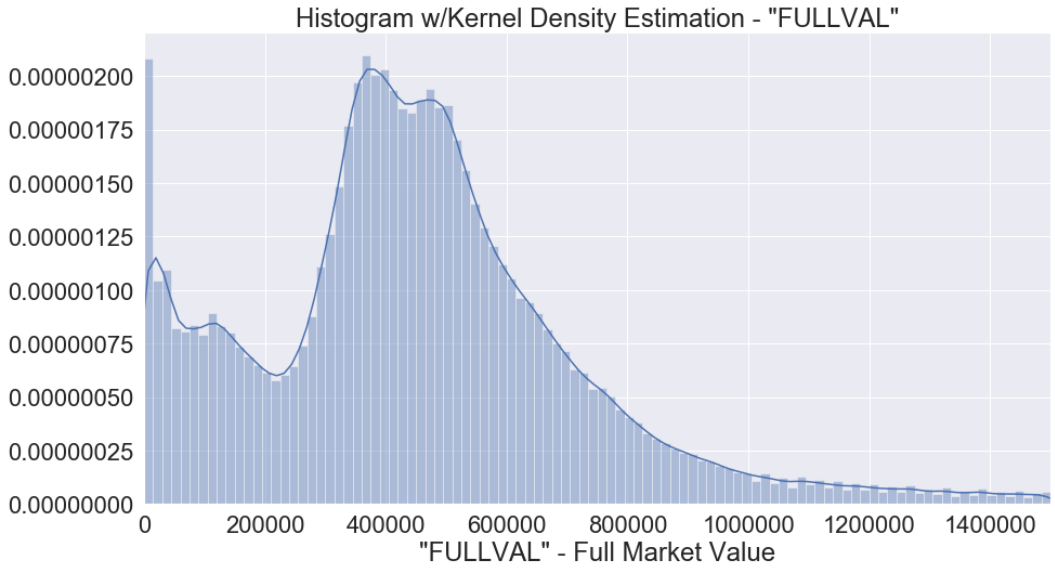

## STORIES

A numerical data field containing the number of stories (i.e. floors) in the building. This data field is 94% populated and has 56,264 property records with no value for the number of stories listed. The distribution below shows all property records with a value for this data field. The most common number of stories is two, and the highest number of stories for a building seen in the dataset is 119. A majority of the property records have buildings with 12 stories or less.



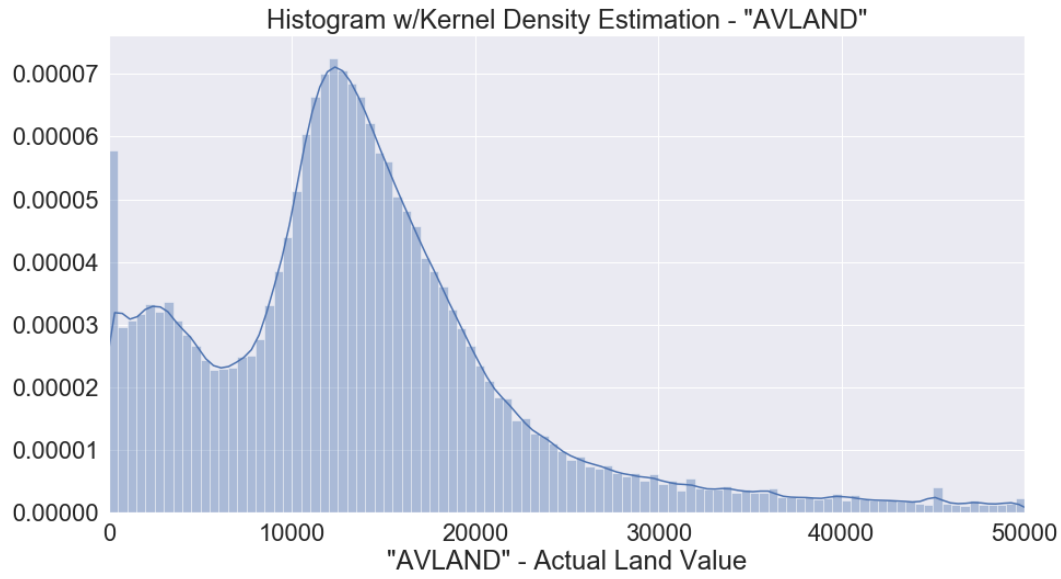
## FULLVAL

A numerical data field containing the full market value of the property. All property records in the dataset have a full market value listed and the range is from $0 to $6.15 billion. The graph below is a histogram of the full market value data field for the property records listed in the dataset having up to a value of $1,500,000.



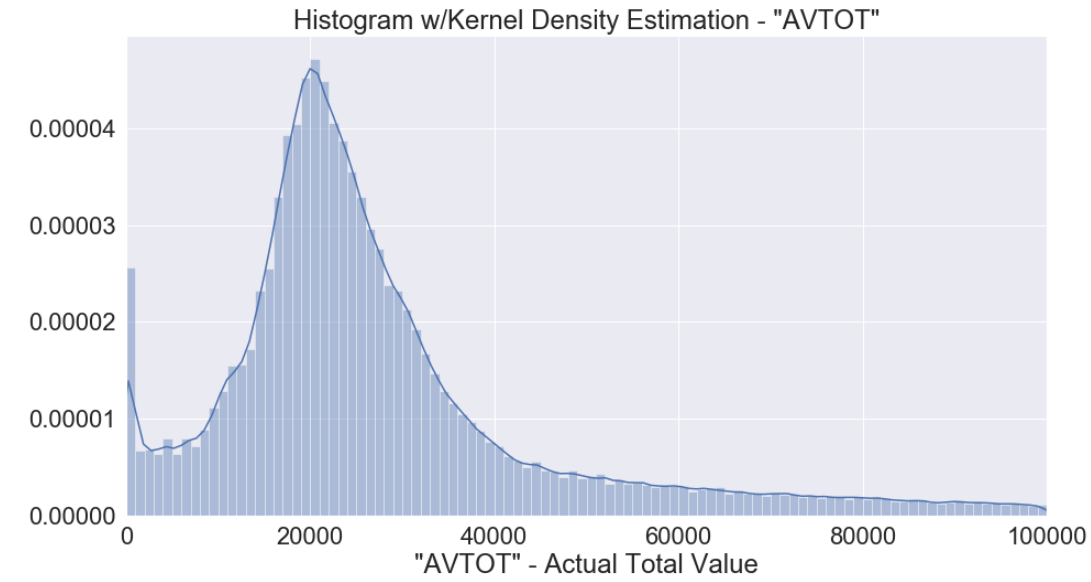

## AVLAND

A numerical data field containing the actual value of the land on the property. All records in the dataset contain a value for this data field and the range is from $0 to $2.6685 billion. The graph below is a histogram of the actual land value data field for the property records listed in the dataset having up to a value of $50,000.



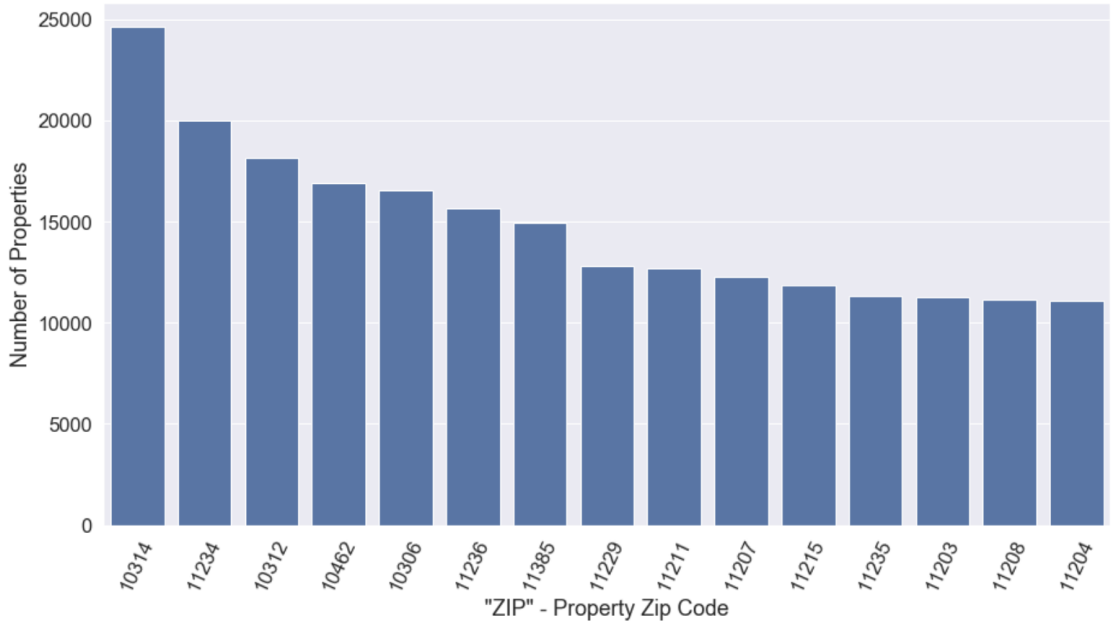
## AVTOT

A numerical data field containing the actual total value of the property. All records in the dataset contain a value for this data field and the range is from $0 to $4.6683 billion. The graph below is a histogram of the actual total value data field for the property records listed in the dataset having up to a value of $100,000.



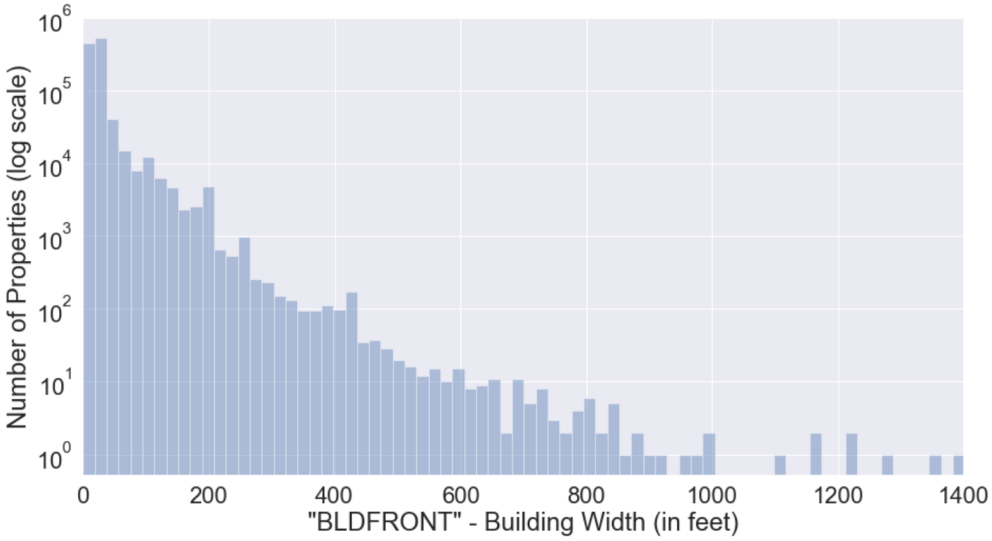

## ZIP

A categorical data field containing the zip code for the property record. This data field is 97% populated with 29,890 property records missing a value. The bar chart below shows the top 15 zip codes with the most property records in the dataset. There are 196 different zip codes in total and there are three property records containing a zip code (33803) for the state of Florida.



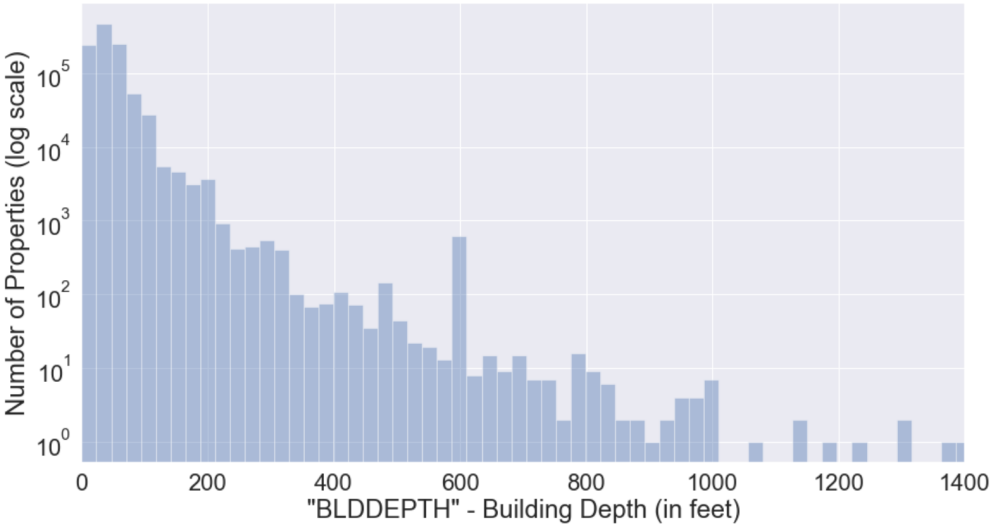
## BLDFRONT

An integer data field containing the front width of the building measured in feet. All property records in the dataset have a value for this data field. The distribution plot below shows the building front width data field up to 1,400 feet for the property records in the dataset. The most common measurement was zero feet with 228,815 property records having this erroneous value. Aside from a measurement of zero feet, 20 feet is the most common width. A measurement of 7,575 feet is the longest building width in the dataset and a majority of the buildings have a width of 26 feet or less.




## BLDDEPTH

An integer data field containing the depth of the building measured in feet. All property records in the dataset have a value for this data field. The distribution plot below shows the building depth data field up to 1,400 feet for the property records in the dataset. The most common measurement was zero feet with 228,853 property records having this erroneous value. Aside from a measurement of zero feet, 40 feet is the most common depth. A measurement of 9,393 feet is the longest depth of a building in the dataset and a majority of the buildings have a depth of 60 feet or less.



# Data Cleaning

A majority of the critical data fields had missing or erroneous values. We therefore needed to use data imputation techniques before fully analyzing the dataset for potential fraud cases. The following data fields were assessed and cleaned for missing or erroneous values: FULLVAL, AVLAND, AVTOT, ZIP, STORIES, LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH. A description of the data imputation conducted for each of these data fields is discussed in the subsections below.

## ZIP

The ZIP data field had with 29,890 property records with a missing value. All of the subsequent data fields necessitated the use of the ZIP data field while assessing and cleaning the data. The ZIP data field was therefore the very first data field where we conducted our data imputation methods. Since this was such an important data field for the data imputation on the dataset as a whole, we worked to methodically and granularly fill the missing values based on location related data fields from the dataset. This meant using the B (borough), BLOCK, and STADDR (street address) data fields.

To use the STADDR data field, we first copied the data field into a temporary column (e.g. STADDR_temp), then we modified the values such that the address number was removed and only the street name remained. When doing this, it was important to understand that addresses such as "1 STREET" or "2 STREET" were reduced to "STREET" and left a potential for causing inaccuracies. Nevertheless, we attempted to reduce the risk of inaccuracies when using this data field by grouping the data with the B and BLOCK data fields. Once the address numbers were removed from the temporary column, we then replaced the empty values with a dummy value (e.g. "empty address") so that we could properly group and aggregate the data.

With our newly formed temporary column for the STADDR data field, we then used the following logic to replace the empty values for the ZIP data field:

**STEP 1** → We filled the missing values with the most common value after grouping the dataset first by B, then by BLOCK, and then by the temporary column for the STADDR data field.

**STEP 2** → Since there were missing values that still remained after Step 1, we then filled the missing values with the most common value after grouping the dataset first by B, then by BLOCK.

**STEP 3** → Since there were missing values that still remained after Step 2, we then filled the missing values with the most common value after grouping the dataset first by B, then by the temporary column for the STADDR data field.

**STEP 4** → Since there were missing values that still remained after Step 3, we then filled the missing values with the most common value after grouping the dataset first by B, then by BLOCK.

**STEP 5** → Since there were missing values that still remained after Step 4, we then filled the missing values with the most common value after grouping the dataset simply by B. At this point, the entire ZIP data field was filled.

## STORIES

The STORIES data field had 56,264 property records with no value for the number of stories listed in the property record. To fill the missing values, we used the following logic:

**STEP 1** → We grouped the dataset first by ZIP, then by BLDGCL, and then by BLOCK. If that grouping had five or more values, then we used the median of that grouping to fill in any missing values within that grouping. If that grouping did not have five or more values, then we made no changes to the missing values.

**STEP 2** → Since there were missing values that still remained after Step 1, we grouped the dataset first by ZIP and then by BLDGCL. If that grouping had five or more values, then we used the median of that grouping to fill in any missing values within that grouping. If that grouping did not have five or more values, then we made no changes to the missing values.

**STEP 3** → Since there were missing values that still remained after Step 2, we grouped the dataset first by ZIP and then by TAXCLASS. If that grouping had five or more values, then we used the median of that grouping to fill in any missing values within that grouping. If that grouping did not have five or more values, then we made no changes to the missing values.

**STEP 4** → Since there were missing values that still remained after Step 3, we grouped the dataset first by B (borough) and then by BLDGCL. If that grouping had five or more values, then we used the median of that grouping to fill in any missing values within that grouping. If that grouping did not have five or more values, then we made no changes to the missing values.

**STEP 5** → At this point, we tried a variety of other methods to reduce the number of missing values for STORIES, but we found our other attempts to be unsuccessful. We therefore opted to reduce the minimum number of required values from five to three. When doing so, we only found success when grouping by TAXCLASS and we were only able to reduce the number of missing values by 4,635 property records with 24,736 property records still missing a value. Regardless, we chose to move forward with this logic and so after grouping the dataset by TAXCLASS, we filled the missing values with the median of that grouping if there were at least three or more values. If that grouping did not have three or more values, then we made no changes to the missing values.

**STEP 6** → After having no more success in using a requirement for a minimum number of records in a grouping before replacing a missing value, we made the determination to simply use the most common value after grouping by BLDGCL. For this step, we therefore simply grouped the dataset by BLDGCL and replaced the missing values with the most common value of that grouping.

**STEP 7** → Since there were missing values that still remained after Step 6, we grouped the dataset by TAXCLASS and replaced the missing values with the most common value of that grouping. At this point, the entire STORIES data field was filled.

## LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH

The LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH data fields do not have any empty values (i.e. NaN values), but there are zero values entered as inputs that we considered erroneous for a large quantity of property records as seen in Table 3 below.

| Data Field | Number of Zero Values |
|---|---|
| **LTFRONT** | 169,108 |
| **LTDEPTH** | 170,128 |
| **BLDFRONT** | 228,815 |
| **BLDDEPTH** | 228,853 |

**Table 3. Number of Property Records with Zero Values for Property Dimension Data.**

In replacing the zero values for each of these data fields, we used the same methodology. Our reasoning for using the groupings seen in the steps described below are based on the idea that the types of buildings associated with the building classification codes (BLDGCL) are of similar size within the same localized area (e.g. a zip code). As an example, we will walk through the steps we took to replace the zero values in the LTFRONT data field.

**STEP 1** → We replaced the zeros with the most common value after grouping the dataset first by ZIP and then by BLDGCL.

**STEP 2** → Since there were 74,917 zero values that still remained after Step 1, we then broadened our scope and replaced the zero values with the most common value after simply grouping the dataset by ZIP. At this point, all zero values were replaced in the data field.

## FULLVAL, AVLAND, and AVTOT

The FULLVAL, AVLAND, and AVTOT data fields do not have any empty values (i.e. NaN values), but there are zero values entered as inputs that we considered erroneous for some of the property records as seen in the Table 4 below.

| Data Field | Number of Zero Values |
|---|---|
| **FULLVAL** | 13,007 |
| **AVLAND** | 13,009 |
| **AVTOT** | 13,007 |

**Table 4. Number of Property Records with Zero Values for Property Monetary Value Data.**

In replacing the zero values for each of these data fields, we used the same methodology. Our reasoning for using the groupings seen in the steps described below are based on the idea that the property values are better classified according to their tax classification code (TAXCLASS), building classification code (BLDGCL), and zip code (ZIP). Additionally, our data imputation technique involved using the median value of the groupings due to the fact that the distributions are skewed for each of these data fields. As an example, we will walk through the steps we took to replace the zero values in the FULLVAL data field.

**STEP 1** → We replaced the zeros with the median value after grouping the dataset first by TAXCLASS, then by BLDGCL, and then by ZIP. This step replaced approximately 19% of the zero values in the FULLVAL data field.

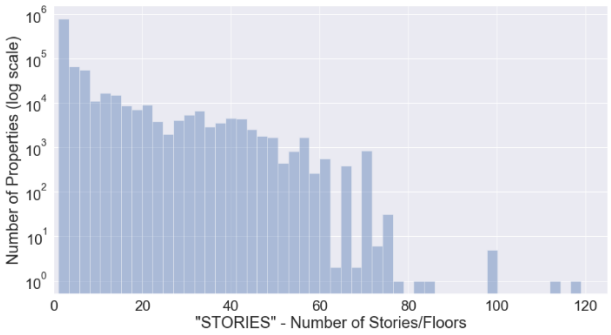
**STEP 2** → Since there were zero values that still remained after Step 1, we replaced the zeros with the median after grouping the dataset first by TAXCLASS and then by ZIP.

**STEP 3** → Since there were zero values that still remained after Step 2, we decided to alter our methodology and replaced the zeros with the median after grouping the dataset simply by ZIP. This step replaced the majority of the remaining zero values.

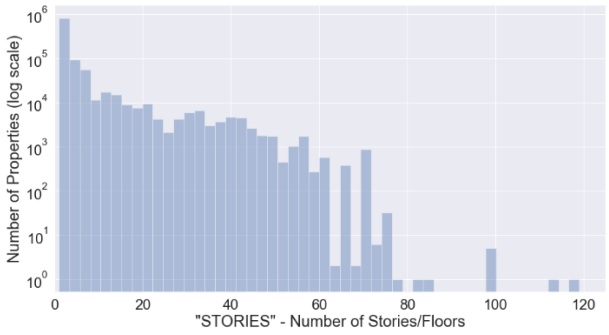
**STEP 4** → Since there were zero values that still remained after Step 3, we replaced the zeros with the median after simply grouping the dataset by TAXCLASS. At this point, all zero values were replaced in the data field.

## Data Imputation Analysis

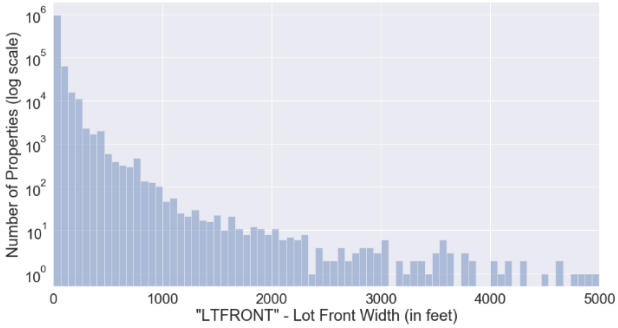
After replacing all the missing or zero values in the numerical data fields for FULLVAL, AVLAND, AVTOT, STORIES, LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH, we verified that the data imputation had a neutral effect on the distribution of the data. The following figures therefore demonstrate the neutral effect of the data imputation by depicting minimal changes from before and after the data imputation. The column of graphs on the left show the distributions of the data before the data imputation, and the column of graphs on the right show the distributions of the data after the data imputation.



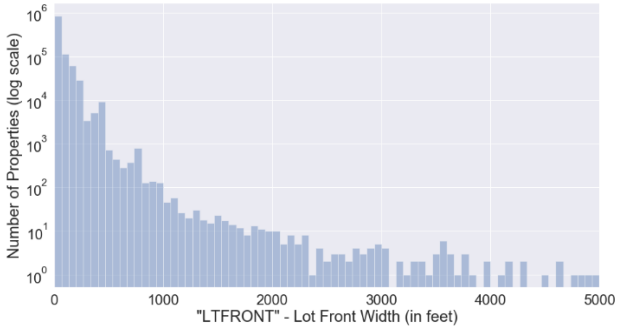

**STORIES w/ 56,264 Missing Values**



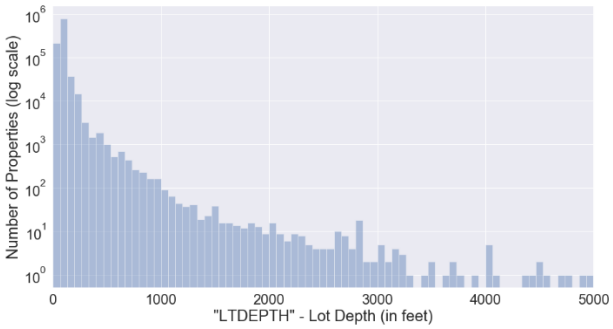

**STORIES Fully Filled**



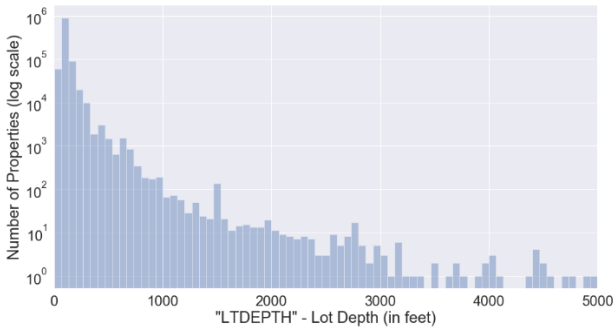

**LTFRONT w/ 169,108 Zero Values**



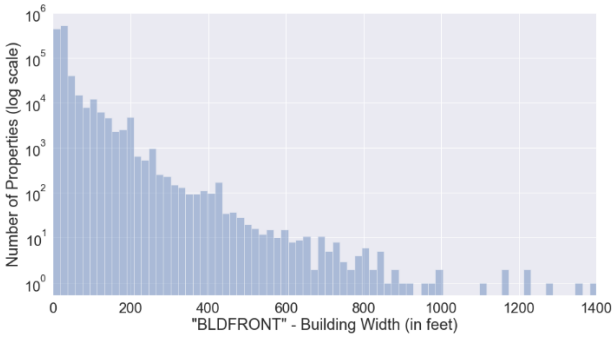

**LTFRONT w/ No Zero Values**



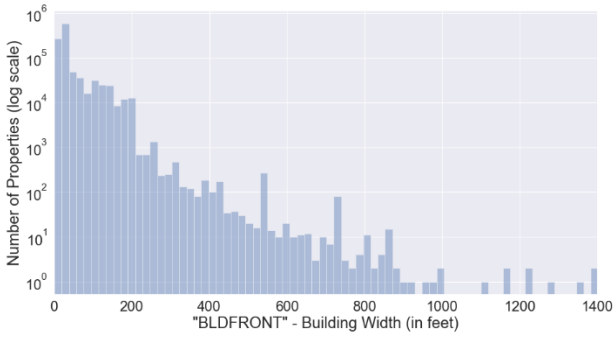

**LTDEPTH w/ 170,128 Zero Values**



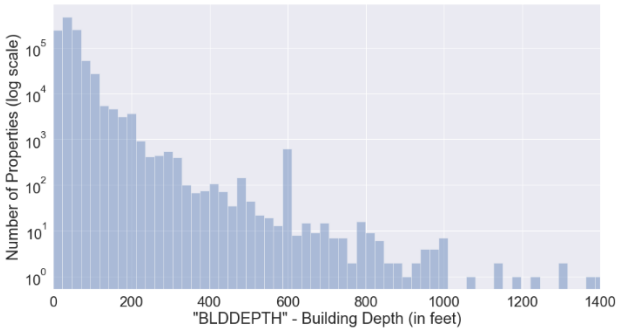

**LTDEPTH w/ No Zero Values**



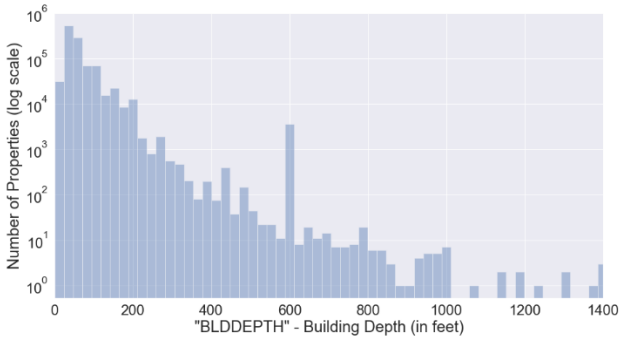

**BLDFRONT w/ 228,815 Zero Values**



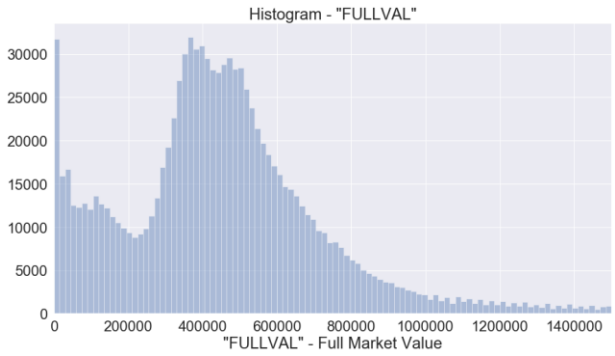

**BLDFRONT w/ No Zero Values**



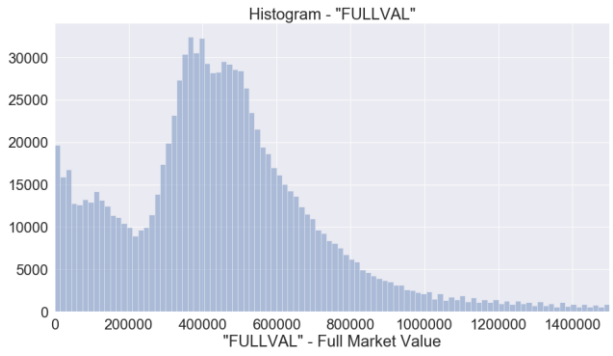

**BLDDEPTH w/ 228,853 Zero Values**



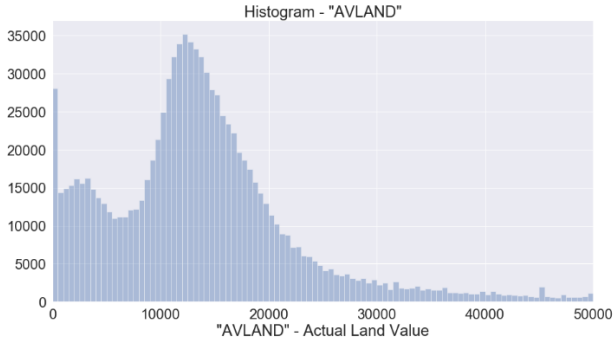

**BLDDEPTH w/ No Zero Values**



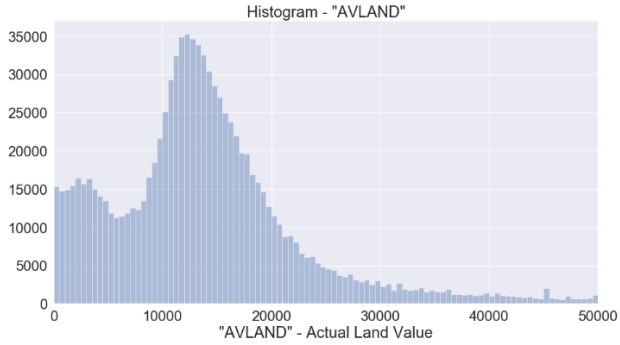

**FULLVAL w/ 13,007 Zero Values**



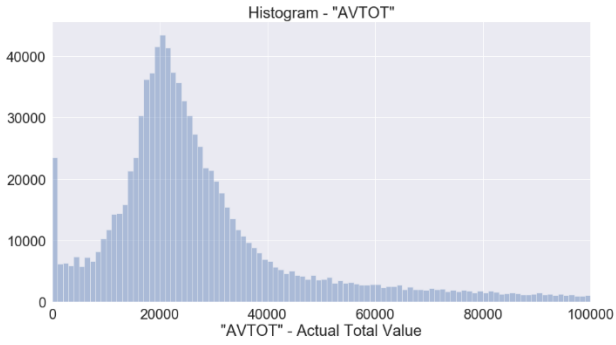

**FULLVAL w/ No Zero Values**



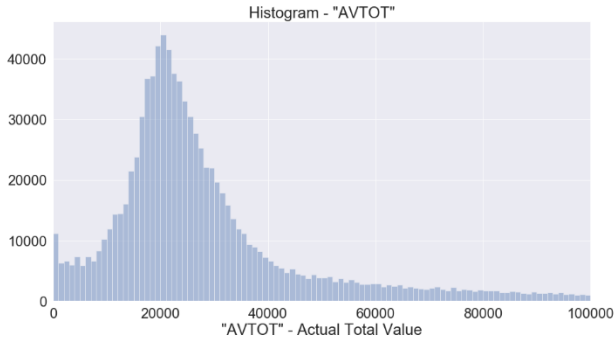

**AVLAND w/ 13,009 Zero Values**




**AVLAND w/ No Zero Values**




**AVLAND w/ 13,007 Zero Values**




**AVTOT w/ No Zero Values**

# Variable Creation

Since this is an unsupervised fraud problem, we do not have a dependent variable to guide us. The task is to find patterns in the data without having this relationship. The dataset provided to us from the City of New York provided a few different property dollar values as well as measurement variables, among others. We decided to focus on these variables to determine if the assessed value was either too low or too high. A valuation was deemed fraudulent if the assessment value fell as an outlier after considering size, location, type of property/building, etc. Variable creation is a significant part of this process of creating an accurate model to predict fraud. In total, 45 variables were created, then dimensionality reduction was used to refine the variables.

To evaluate property value in relation to size, we first created three variables. We wanted to consider size from all different angles to be able to get the most well-rounded model. The first variable is to calculate the area of the lot ($Area_{LOT}$). It is important to not only consider the size of the building on the property, but also the size of the entire lot as well, as this may affect value. In addition, a property record lot may not have any buildings on it – it may be a vacant lot. To calculate the lot area ($Area_{LOT}$), we multiplied the lot frontage size (LTFRONT) with the lot depth size (LTDEPTH). We then focused on building size specifically to get the building area ($Area_{BLD}$) and the building volume ($Vol_{BLD}$). Both of these variables can change property values. Bigger and higher buildings are expected to have higher property valuations. To find the building area ($Area_{BLD}$), the building frontage size (BLDFRONT) was multiplied with the building depth size (BLDDEPTH). Furthermore, building volume ($Vol_{BLD}$) was calculated by multiplying the building area ($Area_{BLD}$) with the number of stories that the building had. Below are the formulas for the three new variables:

$$S_1 = Area_{LOT} = LTFRONT \times LTDEPTH$$

$$S_2 = Area_{BLD} = BLDFRONT \times BLDDEPTH$$

$$S_3 = Vol_{BLD} = S_2 \times STORIES$$

Overall, these three measurement variables capture size well, given the data that was provided. Now the size variables were created, we moved on to using these values to get a normalized figure for the assessment values.

The three assessment values were then normalized by each of the three sizes mentioned above (lot area ($Area_{LOT}$), building area ($Area_{BLD}$), and building volume ($Vol_{BLD}$)). The three assessment values which were given in the dataset were FULLVAL, AVLAND, and AVTOT. FULLVAL is the market value of the property, AVLAND is the assessed value of the property, and AVOTOT is the assessed total value of the property. These three variables are below:

$$V_1 = FULLAL \qquad V_2 = AVLAND \qquad V_3 = AVTOT$$

These set of three variable groups were used to then create nine variables. The three value figures were normalized by the three sizes:

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

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

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

Next, we calculated grouped averages of these nine variables which were grouped by ZIP5, ZIP3, TAXCLASS, B (borough), and ALL (no group). This would divide the variables by location and type of building (embedded into TAXCLASS) which will help assess whether a property assessment value is too high or too low. ZIP3 is a shortened version of ZIP5 to get a more general location as there are many unique zip codes in each B (borough). Each of these groups was labelled as $g_n$, where the range of $n$ was $1$ to $5$. Finally, for each group $g$, we calculated the average for each $r_i$, which resulted in the final 45 variables.

For each $g_{1\to5}$:

$$\frac{r_1}{<r_1>_{g_1}}, \quad \frac{r_2}{<r_2>_{g_1}}, \quad \frac{r_3}{<r_3>_{g_1}}, \quad \dots \quad \frac{r_7}{<r_7>_{g_5}}, \quad \frac{r_8}{<r_8>_{g_5}}, \quad \frac{r_9}{<r_9>_{g_5}}$$

These 45 variables were our final variables. In the next section, we discuss dimensionality reduction and how we refine this set of 45 variables.

# Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables (i.e. features) in a dataset. It is widely used in statistics, data analysis, image processing, and speech processing areas, among others, where the abundance of high dimensional data renders conventional algorithms ineffective due to the curse of dimensionality. Dimensionality reduction offers faster algorithm runtimes, allows data to be stored in reduced storage space, removes correlation between features, and allows visualization of data since 2-D or 3-D datasets can be easily plotted. In the following sections, we describe the two methods for dimensionality reduction used in this work, namely principal component analysis (PCA) and autoencoder.

## $z$-scaling

As an initial step before applying dimensionality reduction, we standardize our dataset using the $z$-scaling method. The $z$-scaling method conducts a linear transformation of the dataset that results in every feature (i.e. variable) of the dataset to have a mean value of zero and a standard deviation value of one. More formally, if the dataset is $X \in R^{m \times n}$ (where there are $m$ data points with $n$ features each), then the transformation is performed as follows:

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

The outcome of standardizing a dataset is that every dimension of the dataset is clustered around $0$ with the same scaling, and thus a simple measure of distance from the origin would yield any outlier values. In Figures 1a and 1b we can see the standardization results for a sample dataset with $10{,}000$ values. More specifically, we observe that the standardized dataset is clustered around $0$ between $-1$ and $1$ standard deviations. It should be noted that the standardized histogram is not quite aligned around $0$ due to the outliers on the negative side that shift the majority of the data samples towards more positive values.

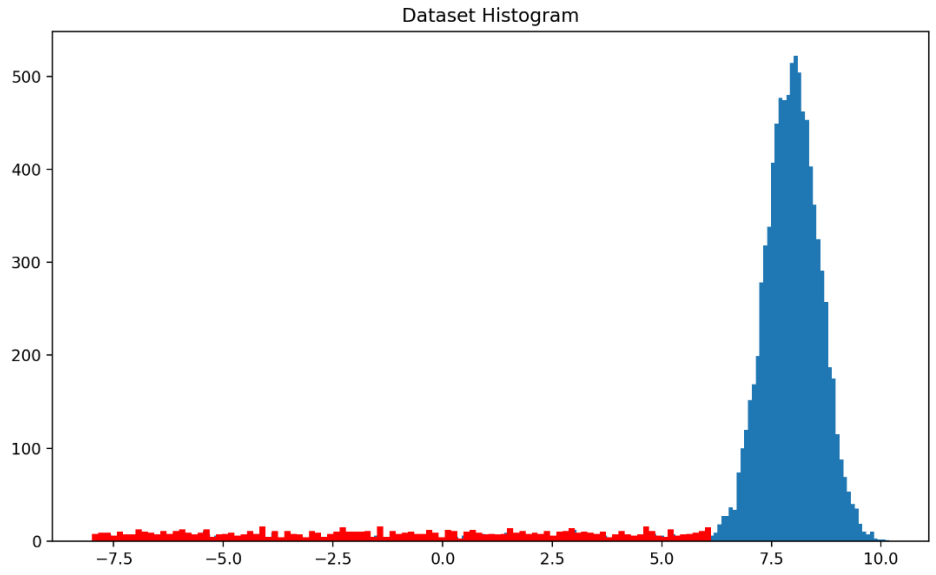



**Figure 1a. Histogram of Sample Dataset**

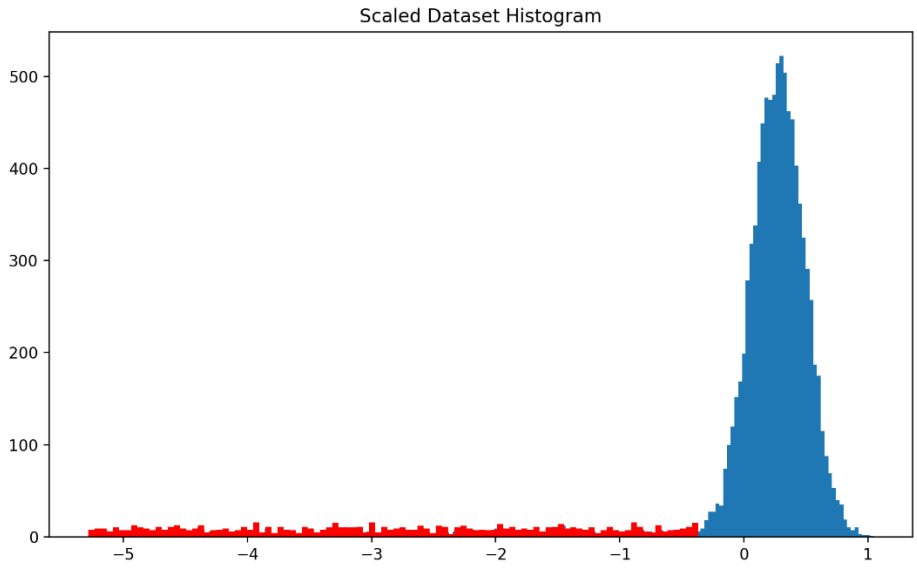



**Figure 1b. Histogram of Standardized Sample Dataset**

## Principal Component Analysis (PCA)

Principal component analysis (PCA) is a dimensionality reduction method that performs a transformation on the dataset and extracts a set of features to create a low dimensional dataset. The transformation projects the original dataset to the axes (i.e. principal components) for those features that have the maximum variance. Additionally, the features of the transformed dataset are ordered according to variance and therefore those features with small variance are considered to have a high linear correlation with other features. Hence, dimensionality reduction is performed by choosing the first $k$ features of the resulting dataset that contain a desirable percentage of the total variance. An illustration of PCA is shown in Figure 2.

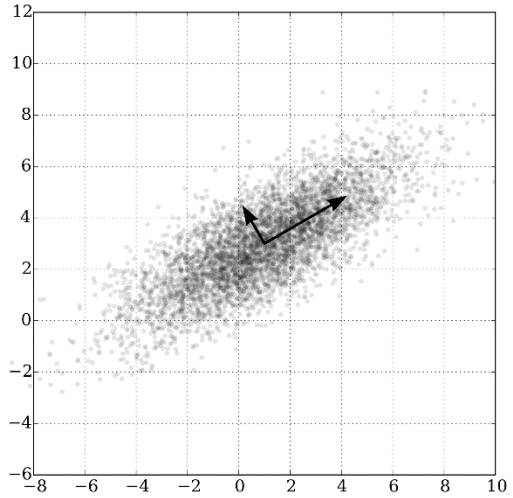



**Figure 2. The cluster of data in the 2-D plane and the 2 axes with max. variance. (Wikipedia)**

The percentage of variance explained by increasing order of principal components of our dataset is shown in Figure 3.

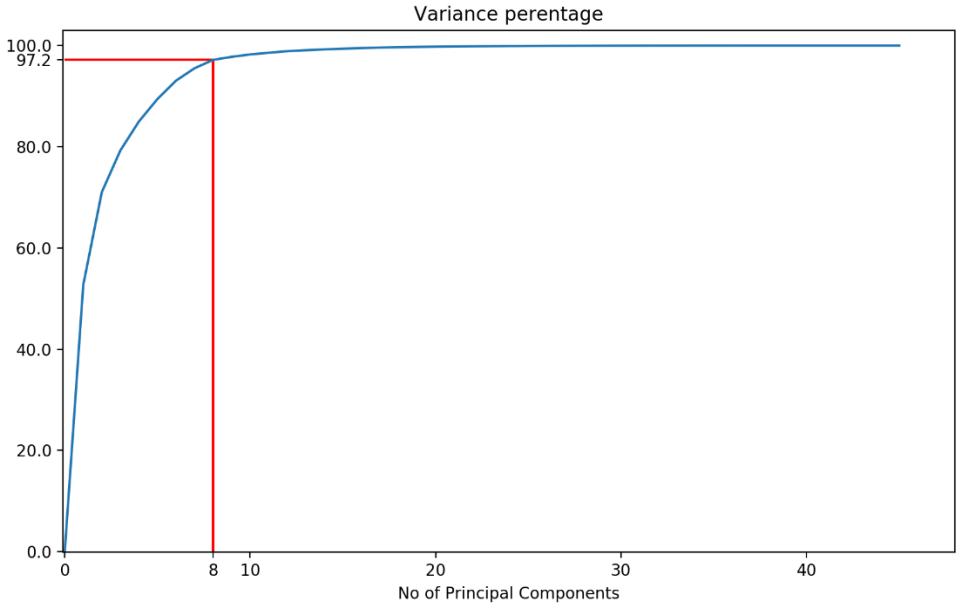



**Figure 3. Percentage of Variance Explained.**

We notice in the figure that $97.2\%$ of the total variance of the original dataset is explained by the first eight principal components.

Another common technique to determine the number of principal components to use is the scree plot, which is a line plot of the variance (eigenvalues of principal components). The scree plot for the eight principal components derived from the PCA analysis on our dataset is presented in Figure 4.

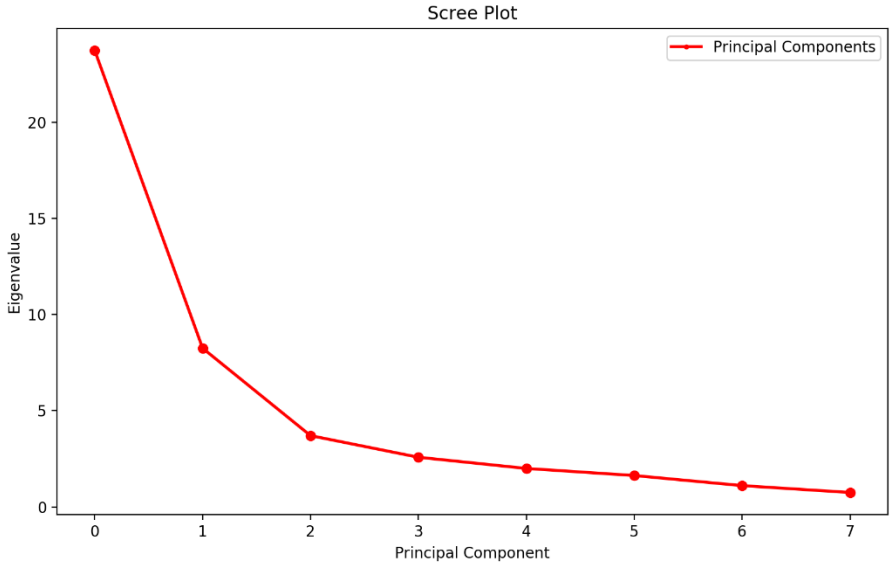



**Figure 4: Scree plot of 8 principal components.**

We observe in Figure 4 that the variance (i.e. eigenvalue) of the last principal component is almost negligible. In this work, we present the results of a PCA analysis with eight principal components and thus we achieved dimensionality reduction of $45 - 8 = 37$ features. It is worth noting that we performed additional simulations with six and ten principal components and yielded the same results in each of the three cases. Finally, we standardized the output of the PCA process so that each feature vector in the dataset has zero mean and unit standard deviation.

## Autoencoder

An autoencoder is considered a statistical model or a type of neural network that can be trained to learn data encodings. The architecture of autoencoders allows them to start learning the data encodings in the input layer where it is subsequently reconstructed and eventually sent to the output layer for analysis. Between the input and output layers are the hidden layers of the autoencoder where the data is compressed to a lower-dimensional space and forwarded to the output layer. This process of dimensionality reduction and data reconstruction offers noise reduction and outlier detection in the dataset. Thus, autoencoders are widely used for image processing or fraud (anomaly) detection.

In this work, we trained an autoencoder with three layers, an input layer of eight nodes to capture the eight principal components resulting after performing PCA, a hidden layer of five nodes (for further dimensionality reduction) and an output layer of eight nodes to reconstruct the input (Figure 5). The input of the autoencoder was the standardized dataset that was produced in the PCA step.

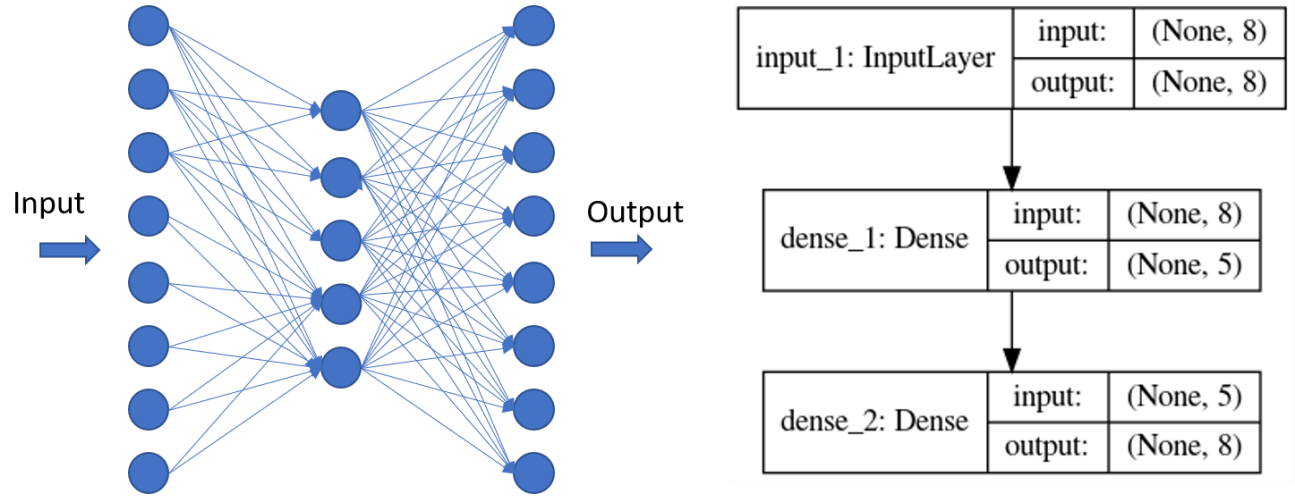



**Figure 5. Autoencoder architecture used in simulations.**

The autoencoder was trained for a different number of nodes in the hidden layer for 15 epochs (training with all data points) using mean squared error as the loss function. In Figure 6 we can see that the loss function converges to a final value after four epochs for each number of nodes. Thus, in our simulations we used five nodes in the hidden layer and trained for four epochs for brevity.

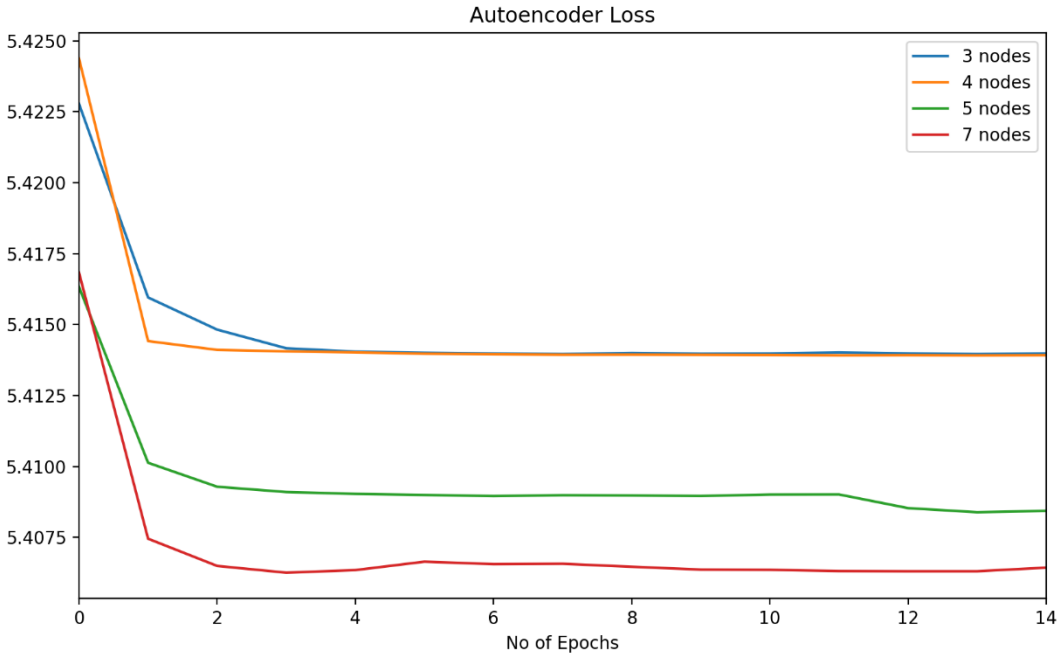



**Figure 6. Autoencoder loss for 3, 4, 5 & 7 nodes in the hidden layer.**

In the next section, we describe the calculation of the outlier scores based on the output of the two methods described, namely the PCA and autoencoder methods.

# Algorithms

To define the fraud scores of each data entry, we used the distance measure for each data point in the standardized output of the PCA algorithm and the autoencoder. The measure of distance of a vector from the origin or the norm of a vector $x \in R^n$ is defined as $\|x\|_p = \left(\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$. Parameter $p$ defines a different distance metric and as $p \to \infty$, then $\|x\|_p \to \max\{x_i\}$, namely the norm of a vector converges to its maximum element as $p$ is increasing. The most commonly used $p$ values are $p = 1, 2, \infty$. In our simulations we used all these three $p$ values and produced the same fraud results. We therefore opted to present the results for $p = 2$.

## Score 1

Score 1 ($S_1$) is the 2-norm of the standardized PCA output ($P_z$), $S_1 = \|P_z\|_2$. Any outliers in the original dataset would indicate high variance and PCA maximizes variance by projecting in the principal components. Hence, an outlier value in a dimension would be included in the dimensionality reduction performed by PCA and could be easily detected by a distance metric in the standardized dataset.

## Score 2

Score 2 ($S_2$) is the 2-norm of the distance (difference) of the autoencoder input $P_z$ (standardized PCA output) and autoencoder output $P_a$, $S_2 = \|P_z - P_a\|_2$. The autoencoder is trained to reproduce the input in its output while performing dimensionality reduction in the hidden layer. The dimensionality reduction results in loss of the outlier values in the output and thus a distance metric of the difference of the input and output yields those values.

## Final Score

For the final score $S_F$, Score 1 ($S_1$) and Score 2 ($S_2$) were combined with the harmonic mean, namely $S_F = 2\frac{S_1 S_2}{S_1 + S_2}$. The harmonic mean of the two scores tends to be larger when both scores are large enough and decreases if at least one of the two scores is smaller in value than the other, thereby making it is robust to inconsistencies between the scores. Thus, the final score is higher if both $S_1$ and $S_2$ are high in order and lower otherwise. In additional simulations, we used other combining methods such as the average or the maximum of the two scores and the results were almost identical. We concluded that the harmonic mean would be the optimal combining method for the aforementioned reasons.

# Results

The histogram of Score 1 is shown in Figure 7 and is shown with a logarithmic y-axis. In reviewing the results, we can clearly see the outlier values.

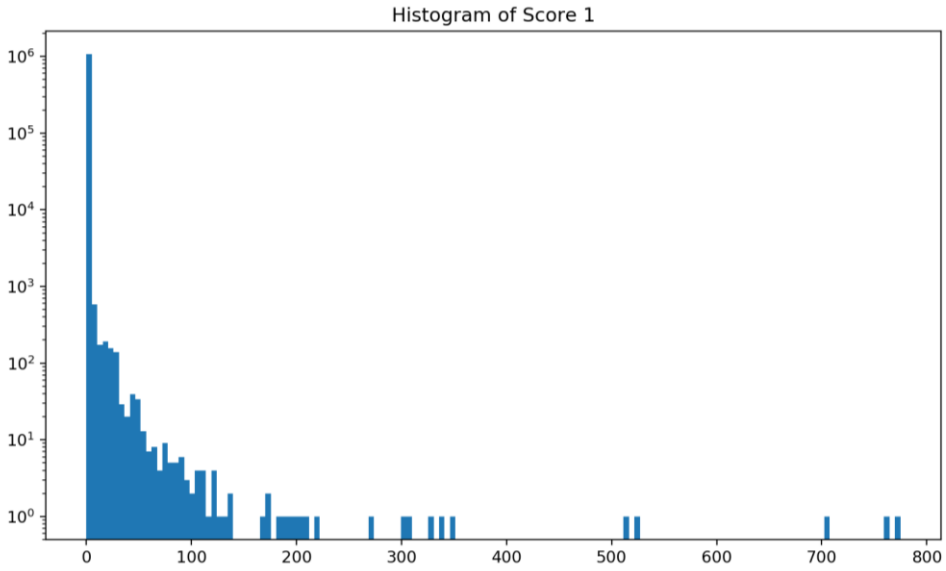



**Figure 7. Score 1 Histogram.**

The histogram of Score 2 is shown in Figure 8. Similar to the histogram of Score 1, the y-axis is in logarithmic scale and we can easily identify outlier values in Score 2. Moreover, there is a difference in the order of the two scores. The maximum value of Score 1 ($S_1$) is on the order of approximately $800$ whereas one of the Score 2 ($S_2$) values is on the order of approximately $3{,}500$ as seen along the x-axis. The latter remark suggests that the two scores should be standardized to be on the same order before being combined to a final score where we can then conduct a proper analysis and determine which property records are potentially fraudulent.

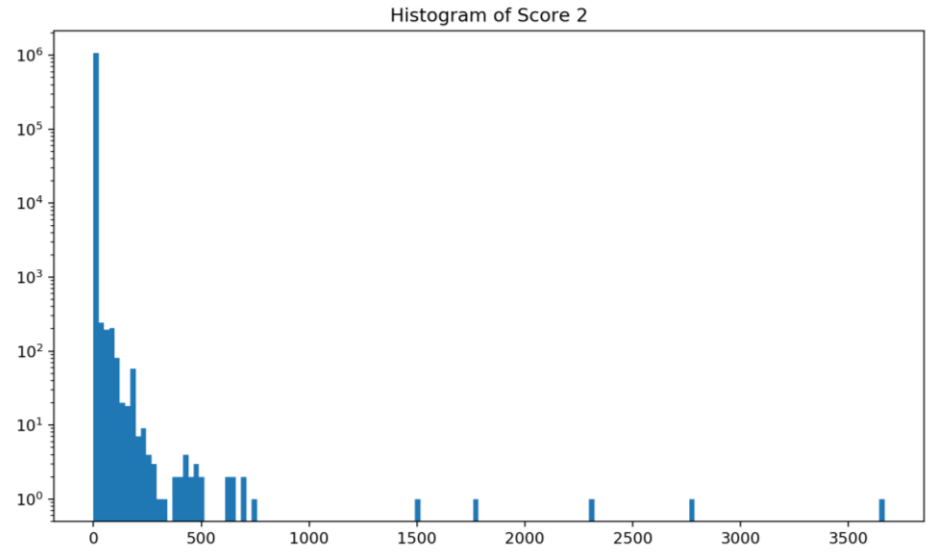



**Figure 8. Score 2 Histogram.**

## Final Score

The histogram of the final score is shown in Figure 9. Similar to Score 1 and Score 2, a logarithmic y-axis was used in the histogram. Again, it is clear that some data points are outliers. To have a better understanding of the number of fraudulent data points to examine, we plotted the final score of the maximum 150 data points (Figure 10).

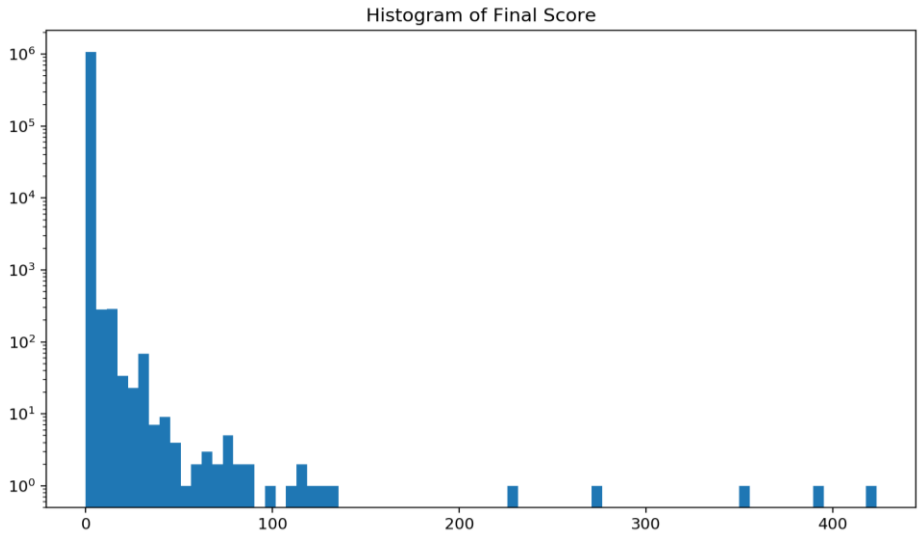



**Figure 9. Final Score Histogram.**

In Figure 10 we observe that the final score decreased significantly after the 10 maximum data points; hence we examined the top 10 entries in the original dataset (red line indicates the threshold). Finally, for a better illustration of the outlier values, we present the scatter plot of normalized Score 1 and Score 2 and point out these values with red in Figure 11.

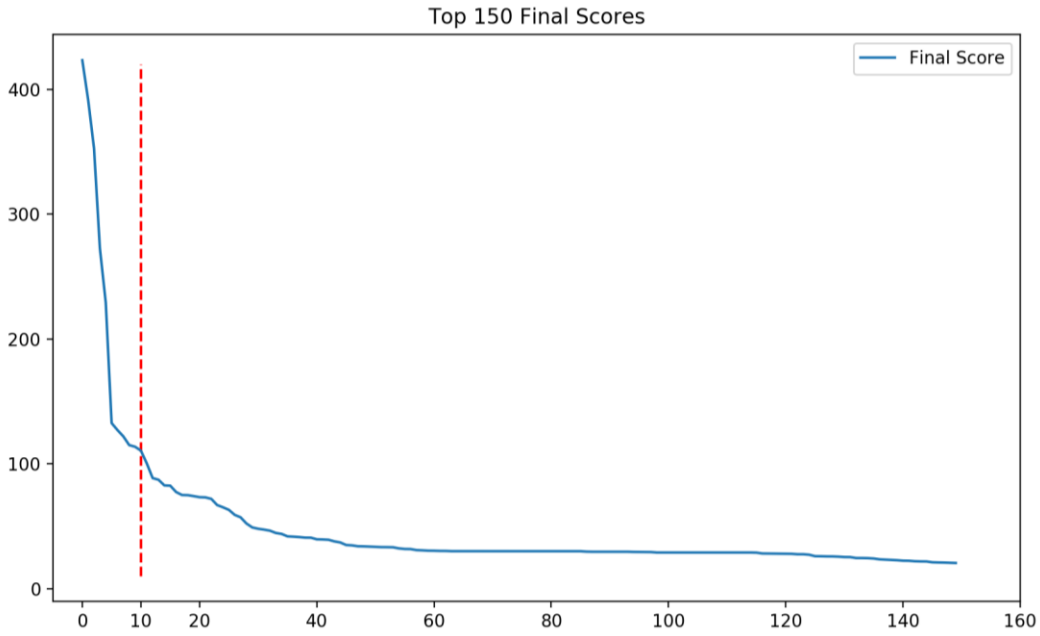



**Figure 10. Final Score of top 150 data points.**

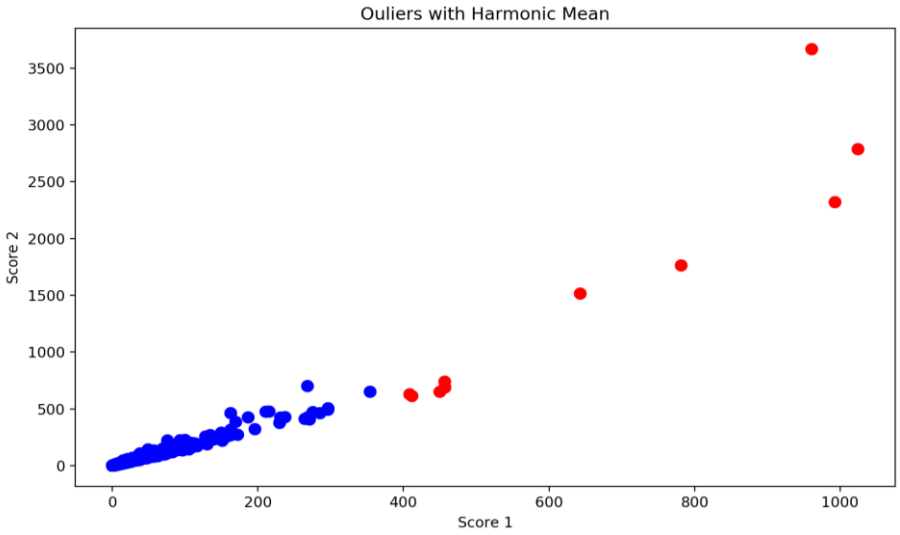



**Figure 11. Scatter plot of normalized scores.**

From the scatter plot we notice that the outliers have both a Score 1 and Score 2 with extreme values. This explains why other methods of combining the two scores (max, or average) produced identical results.

A summary depiction of the whole process to detect fraudulent values that was followed in this work is presented below in Figure 12.

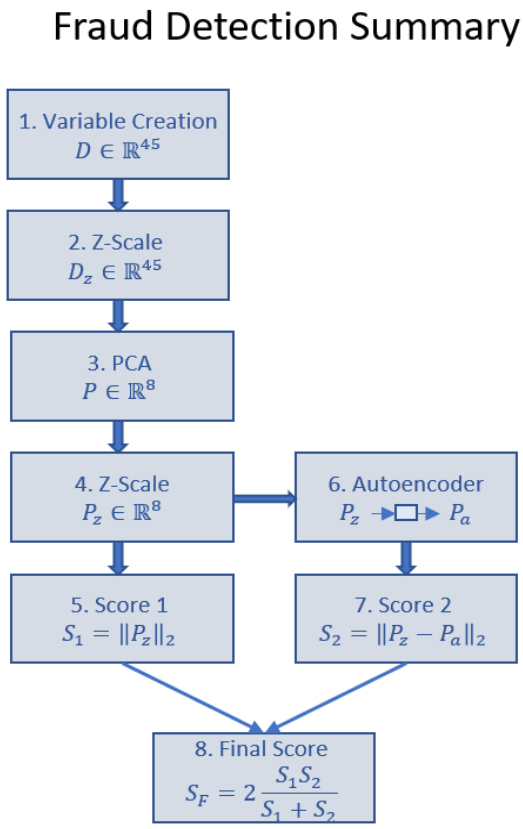



**Figure 12. Summary.**

After combining the two fraud scores into one final fraud scores, we analyzed the top ten properties that had the highest fraud scores. We looked at the overall data points for each property to highlight why the property resulted in a high fraud score. Some variables had strangely low or strangely high values - for example, for LTFRONT and LTDEPTH. These could represent a fraudulent record, clerical error, etc. Further investigation and discussion will be needed will our NYC property expert counterparts to determine if these records are indeed fraudulent or not. Also, we looked at the z scaled variables for the top ten records with the highest fraud scores to see which variables had unusual variables. Our analysis for each of the top ten records are below. Many are owned by government entities.

| RECORD | B | BLOCK | OWNER | BLDG CL | TAX CLASS | LT FRONT | LT DEPTH | STORIES | FULL VAL | AVLAND | AVTOT | STADDR | ZIP | BLD FRONT | BLD DEPTH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **565392** | 3 | 8590 | U S GOVERNMENT OWNRD | V9 | 4 | 117 | 108 | 1 | 4326303700 | 1946836665 | 1946836665 | FLATBUSH AVENUE | 11234 | 20 | 40 |
| **632816** | 4 | 1842 | 864163 REALTY, LLC | D9 | 2 | 157 | 95 | 1 | 2930000 | 1318500 | 1318500 | 86-55 BROADWAY | 11373 | 1 | 1 |
| **917942** | 4 | 14260 | LOGAN PROPERTY, INC. | T1 | 4 | 4910 | 100 | 3 | 374019883 | 1792808947 | 4668308947 | 154-68 BROOKVILLE BOULEVARD | 11422 | 20 | 28 |
| **1067360** | 5 | 7853 | | B2 | 1 | 1 | 1 | 2 | 836000 | 28800 | 50160 | 20 EMILY COURT | 10307 | 36 | 45 |
| **565398** | 3 | 8591 | DEPT OF GENERAL SERVI | V9 | 4 | 466 | 1009 | 1 | 2310884200 | 1039897890 | 1039897890 | FLATBUSH AVENUE | 11234 | 20 | 40 |
| **585118** | 4 | 420 | NEW YORK CITY ECONOMI | O3 | 4 | 298 | 402 | 20 | 3443400 | 1549530 | 1549530 | 28-10 QUEENS PLAZA SOUTH | 11101 | 1 | 1 |
| **565390** | 3 | 8590 | PARKS AND RECREATION | Q1 | 4 | 143 | 128 | 1 | 223680000 | 99697500 | 100656000 | 3000 FLATBUSH AVENUE | 11234 | 4 | 42 |
| **585439** | 4 | 459 | 11-01 43RD AVENUE REA | H9 | 4 | 94 | 165 | 10 | 3712000 | 252000 | 1670400 | 11-01 43 AVENUE | 11101 | 1 | 1 |
| **85886** | 1 | 1254 | PARKS AND RECREATION | Q1 | 4 | 4000 | 150 | 1 | 70214000 | 31455000 | 31596300 | JOE DIMAGGIO HIGHWAY | 10027 | 8 | 8 |
| **248665** | 2 | 5650 | PARKS AND RECREATION | Q1 | 4 | 600 | 4000 | 6 | 190000000 | 79200000 | 85500000 | PELHAM BAY PARK | 10462 | 20 | 20 |

## Top 10 Results

1. **Record No. 565392**. This property was owned by the US government and did not have an exact street address so we could look up the exact location to investigate further from this perspective. The building code and size variables corresponded to a small vacant lot. It had a much higher FULLVALL value than AVLAND and AVTOT. Overall, most of the z scores had high in absolute value across all variables in the zip3 group.

2. **Record No. 63281**. BLDFRONT AND BLTDEPTH were both one foot for this property. This could represent a fraudulent record, clerical error, etc. Further investigation and discussion will be needed to determine if this property has a fraudulent property assessment or not. FULLVALL/BLDVOL and AVLAND/BLDAREA, and AVTOT/BLDVOL in the tax class group have higher z scores (in absolute value) relative to other variables.

3. **Record No. 917942**. This property had a very high LTFRONT value compared to LTDEPTH, BLDFRONT, and BLDDEPTH. In regards to z scores (in absolute value), AVTOT/BLDAREA in the zip 3 group had higher values. Currently, the property looks to be a hotel, but this may not have been the case when the data was originally collected.

4. **Record No. 1067360**. LTFRONT and LTDEPTH were both one foot for this property. This could represent a fraudulent record, clerical error, etc. Further investigation and discussion will be needed to determine if this property has a fraudulent property assessment or not. There was also no owner listed for this record and per the building class looks to be a two family dwelling. FULLVALL/LOT AREA, AVLAND/LOTAREA, and AVLAND/LOTAREA in the tax class group had relatively high z scores (in absolute value) than the other variables.

5. **Record No. 565398**. This property is owned by the Department of General Services and looks to be a park. The street address did not have a specific address number before the street, so we could not pinpoint the location. AVLAND/BLDVOL in the zip3 group as well as AVLAND/BLDVOL in the B group has relatively high scores (in absolute value).

6. **Record No. 585118**. BLDFRONT and BLTDEPTH were both one foot for this property. This could represent a fraudulent record, clerical error, etc. Further investigation and discussion will be needed to determine if this property has a fraudulent property assessment or not. AVLAND/BLDAREA in group B and FULLVAL/BLDAREA in zip3 group had higher z scores (in absolute value).

7. **Record No. 565390**. This property was owned by the Parks and Recreation Department. It has a very low BLDFRONT at 4ft, but BLDEPTH at 42ft, LTFRONT at 143ft, and LTDEPTH at 146 ft. AVALND/BLDVOL in the zip3 group as well as FULLVAL/BDLVOL in the all group had higher z scores (in absolute value).

8. **Record No. 585439**. BLDFRONT AND BLTDEPTH variables had values of one for this property . This could represent a fraudulent record, clerical error, etc. Further investigation and discussion will be needed to determine if this property has a fraudulent property assessment or not. FULLVALL/BLDAREA in the B group and FULLVAL/BLDAREA in the zip3 group had higher z score values (in absolute value).

9. **Record No. 85886**. This property is Pelham Bay Park which is the largest public park in NYC at 2722 acres (three times the size of Central Park). FULLVALL was much higher than AVLAND and AVTOT. In terms of z scores (in absolute value), FULLVALL/BLDAREA in the zip5 group and AVLAND/BLDVOL in the zip5 group had higher values.

10. **Record No. 248665**. This property is owned by the Parks and Recreation Department. It has small building measurement but has bigger lot measurements (BLDFRONT and BLDDEPTH at 8ft with LTFRONT at 4000 ft and LTDEPTH at 150ft). It also has a relatively large lot area at 2,400,000 feet squared. The original dataset had zero values for the property record's BLDFRONT and BLDDEPTH data fields, but we replaced both of those values with a value of 20 feet during our data imputation. Also, FULLVAL/BLDVOL in the zip3 and B groups had higher z score values (in absolute value).

# Conclusions

A comprehensive analysis of NYC property assessment values was performed to determine possible fraudulent records. Principal Component Analysis and a neural network autoencoder were used to obtain two fraud scores and then the scores were combined into one final fraud score. The top ten records were further analyzed using z scaled values to determine which variables looked unusual. Following-up on these results with property tax experts will help determine if these records are actually fraudulent. Many of the top 10 records included properties owned by the government. As part of our future steps, this is an area to further look into to see why this trend exists.

In addition, some of the variables such as LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH has extreme values such as 1. Further investigation would be needed to check if these are fraudulent values, administrative errors, a fill-in value for missing data in the original dataset, or some other result.

As mentioned above, since many of the top 10 properties were government owned, we also looked at the top 30 properties with the highest fraud scores and excluded government owned properties. The revised version of the top 10 results are listed below. As a future step, we would further analyze this set of properties.

| RECORD | OWNER |
|---|---|
| **632816** | 864163 REALTY, LLC |
| **917942** | LOGAN PROPERTY, INC. |
| **1067360** | |
| **585439** | 11-01 43RD AVENUE REA |
| **750816** | M FLAUM |
| **935158** | RICH-NICH REALTY,LLC |
| **1067001** | DRANOVSKY, VLADIMIR |
| **920628** | PLUCHENIK, YAAKOV |
| **67128** | CULTURAL AFFAIRS |
| **776305** | TONY CHEN |

# Appendix: Data Quality Report (DQR)

## 1 Data Overview

**Description**: The data analysis contained in this report is from public data on New York City (NYC) real estate consisting of property valuations and assessments. The purpose of this data is for calculating property taxes and granting eligible properties exemptions and/or abatements. The NYC Department of Finance (DOF) collects this information for each fiscal year.

**Data Source**: The NYC DOF provided the data to NYC OpenData (https://opendata.cityofnewyork.us/) and can be found at https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8.

**Data Time Period**: NYC 2010/2011 fiscal year (begins on July 1st of the calendar year and ends on June 30th of the following calendar year).

**Number of Data Fields**: 32

**Number of Records**: 1,070,994

**Name of Data File**: "NY property data.csv"

**Size of Data File**: 170.906 MB

# 2 Summary Tables

The tables below provide summary information for the numerical and categorical data found in the dataset. Within the 32 data fields there are 14 numerical data fields and 18 categorical data fields. Section 2.1 contains the summary table for the numerical data fields and Section 2.2 contains the summary table for the categorical data fields.

## 2.1 Numerical Summary Table

| Data Field | Num Records w/ a Value | Percent Populated | Num Unique Values | Num Records w/ a Value of Zero | Mean | Standard Deviation | Min | Max |
|---|---|---|---|---|---|---|---|---|
| **LTFRONT** | 1,070,994 | 100% | 1,297 | 169,108 | 37 ft | 74 ft | 0 ft | 9,999 ft |
| **LTDEPTH** | 1,070,994 | 100% | 1,370 | 170,128 | 89 ft | 76 ft | 0 ft | 9,999 ft |
| **STORIES** | 1,014,730 | 94.7% | 111 | 0 | 5.0 | 8 | 1 | 119 |
| **FULLVAL** | 1,070,994 | 100% | 109,324 | 13,007 | $874,265 | $11,582,431 | $0 | $6,150,000,000 |
| **AVLAND** | 1,070,994 | 100% | 70,921 | 13,009 | $85,068 | $4,057,260 | $0 | $2,668,500,000 |
| **AVTOT** | 1,070,994 | 100% | 112,914 | 13,007 | $227,238 | $6,877,529 | $0 | $4,668,308,947 |
| **EXLAND** | 1,070,994 | 100% | 33,419 | 491,699 | $36,424 | $3,981,576 | $0 | $2,668,500,000 |
| **EXTOT** | 1,070,994 | 100% | 64,255 | 432,572 | $91,187 | $6,508,403 | $0 | $4,668,308,947 |
| **BLDFRONT** | 1,070,994 | 100% | 612 | 228,815 | 23 ft | 35 ft | 0 ft | 7,575 ft |
| **BLDDEPTH** | 1,070,994 | 100% | 621 | 228,853 | 39 ft | 42 ft | 0 ft | 9,393 ft |
| **AVLAND2** | 282,726 | 26.4% | 58,591 | 0 | $246,236 | $6,178,963 | $3 | $2,371,005,000 |
| **AVTOT2** | 282,732 | 26.4% | 111,360 | 0 | $713,911 | $11,652,529 | $3 | $4,501,180,002 |
| **EXLAND2** | 87,449 | 8.2% | 22,195 | 0 | $351,236 | $10,802,213 | $1 | $2,371,005,000 |
| **EXTOT2** | 130,828 | 12.2% | 48,348 | 0 | $656,769 | $16,072,510 | $7 | $4,501,180,002 |

## 2.2 Categorical Summary Table

| Data Field | Num Records w/ a Value | Percent Populated | Num Unique Values | Most Common Value |
|---|---|---|---|---|
| **RECORD** | 1,070,994 | 100% | 1,070,994 | N/A |
| **BBLE** | 1,070,994 | 100% | 1,070,994 | N/A |
| **B** | 1,070,994 | 100% | 5 | 4 |
| **BLOCK** | 1,070,994 | 100% | 13,984 | 3944 |
| **LOT** | 1,070,994 | 100% | 6,366 | 1 |
| **EASEMENT** | 4,636 | 0.43% | 12 | E |
| **OWNER** | 1,039,249 | 97.04% | 863,346 | PARKCHESTER PRESERVAT |
| **BLDGCL** | 1,070,994 | 100% | 200 | R4 |
| **TAXCLASS** | 1,070,994 | 100% | 11 | 1 |
| **EXT** | 354,305 | 33.08% | 3 | G |
| **EXCD1** | 638,488 | 59.62% | 129 | 1017 |
| **STADDR** | 1,070,318 | 99.94% | 839,280 | 501 SURF AVENUE |
| **ZIP** | 1,041,104 | 97.21% | 196 | 10314 |
| **EXMPTCL** | 15,579 | 1.45% | 14 | X1 |
| **EXCD2** | 92,948 | 8.68% | 60 | 1017 |
| **PERIOD** | 1,070,994 | 100% | 1 | FINAL |
| **YEAR** | 1,070,994 | 100% | 1 | 2010/11 |
| **VALTYPE** | 1,070,994 | 100% | 1 | AC-TR |

# 3 Data Field Exploration

Each subsection below provides additional detailed information about each data field within the dataset. The data fields are listed in the order in which they appear.

## 3.1 Field 1: RECORD

**Description**: A categorical data field containing an integer representing the unique record number identifier from 1 to 1,070,994. All records in the dataset contain a record number.

## 3.2 Field 2: BBLE

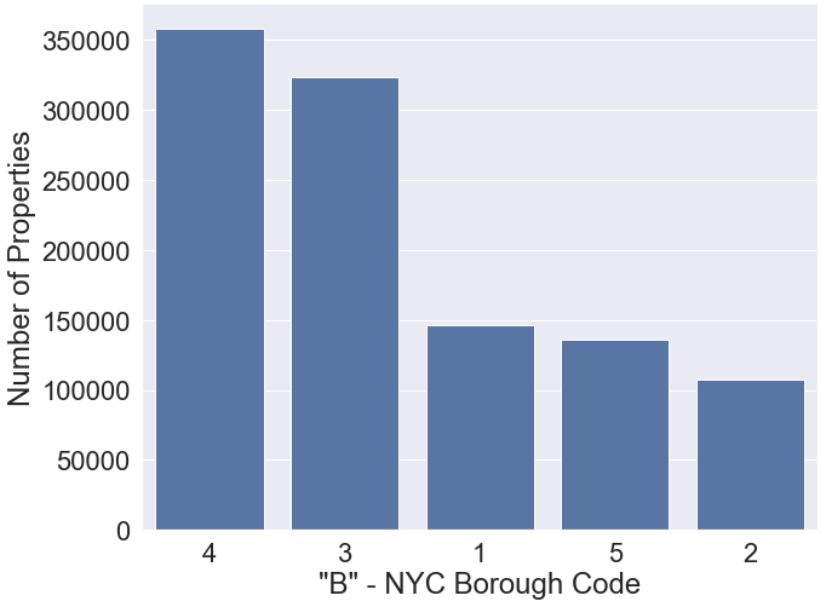
**Description**: A categorical data field containing a unique 11-digit value that is resultant from a concatenation of the "B" field (1-digit), "BLOCK" field (5-digits), "LOT" field (4-digits), and "EASEMENT" field (1-character). All records in the dataset contain a BBLE value.

## 3.3 Field 3: B

**Description**: A categorical data field containing integer values for NYC's borough codes as annotated below. All records in the dataset contain a borough code.

| Borough Code | Borough Name |
|---|---|
| **1** | Manhattan |
| **2** | Bronx |
| **3** | Brooklyn |
| **4** | Queens |
| **5** | Staten Island |

The bar chart listed below provides the total number of property records per borough as found in the dataset. Queens has the greatest number of property records, while the Bronx has the least number of property records.

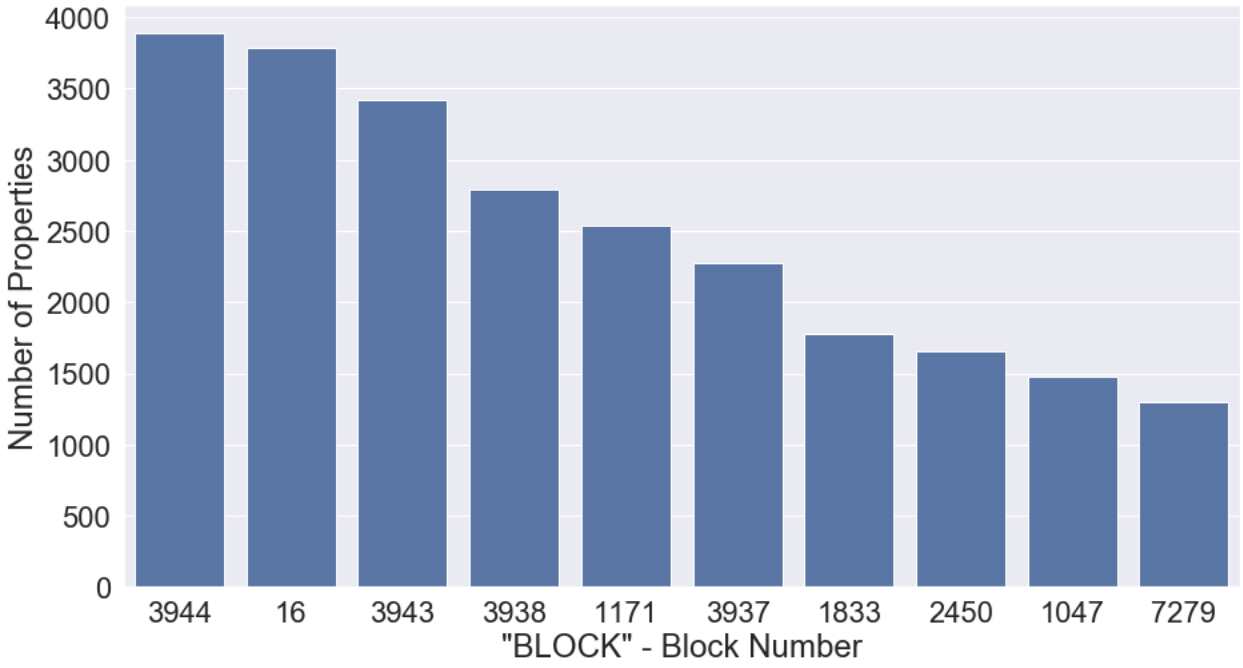


## 3.4 Field 4: BLOCK

**Description**: A categorical data field containing integer values for the block number within the borough for that particular property record. All records in the dataset contain a block number and the boroughs have the following valid block ranges:

| Borough Name | Valid Block Range |
|---|---|
| **Manhattan** | 1 to 2,255 |
| **Bronx** | 2,260 to 5,958 |
| **Brooklyn** | 1 to 8,955 |
| **Queens** | 1 to 16,350 |
| **Staten Island** | 1 to 8,050 |

The bar chart below provides the top 10 block numbers with the most property records in the dataset. It is important to note that since the values for the block ranges overlap, property records from different boroughs can have the same block number.



## 3.5 Field 5: LOT

**Description**: A categorical data field containing integer values for the unique lot number within the block of the borough for that particular property record. All records in the dataset contain a lot number.

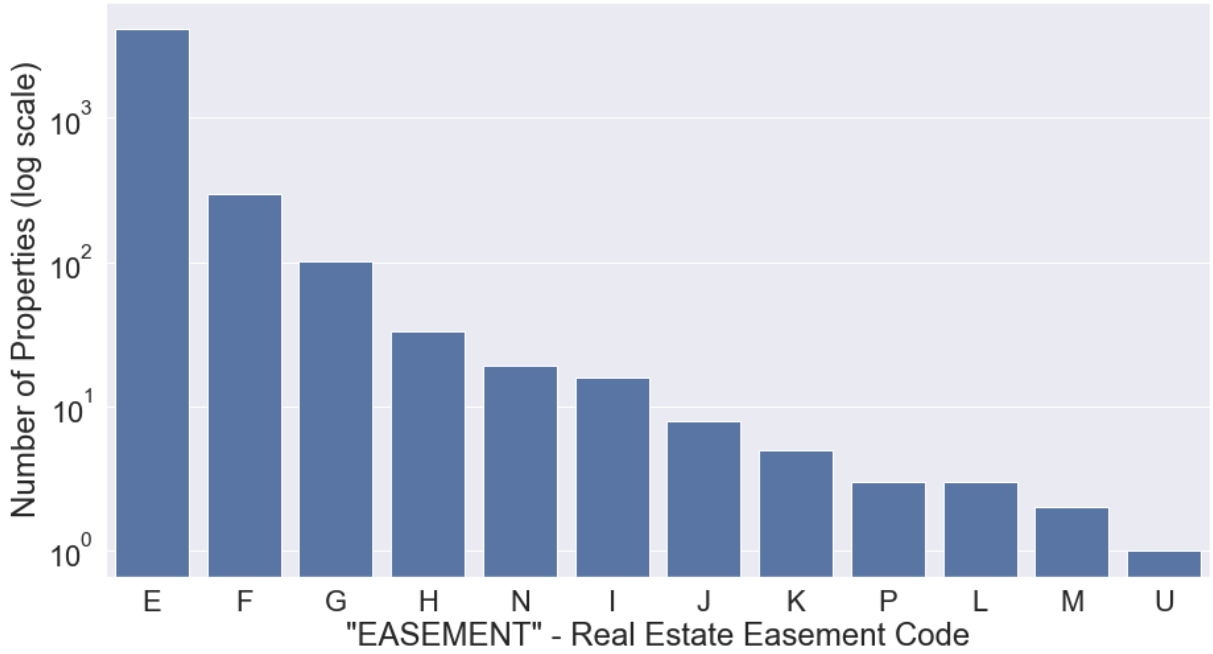
The bar chart below provides the top 10 lot numbers with the most property records in the dataset. It is important to note that since lot numbers amongst the various blocks and boroughs can overlap, property records from different blocks and boroughs can have the same lot number.

## 3.6 Field 6: EASEMENT

**Description**: A categorical data field containing a letter code for any real estate easements authorized for that particular property record. This data field is sparsely filled with only 4,636 records in the dataset containing an easement code. The easement letter codes are described as follows:

| Easement Code | Description |
|---|---|
| **SPACE** | Indicates the lot has no Easement |
| **'A'** | Indicates the portion of the Lot that has an Air Easement |
| **'B'** | Indicates Non-Air Rights |
| **'E'** | Indicates the portion of the lot that has a Land Easement |
| **'F' THRU 'M'** | Duplicates of 'E' |
| **'N'** | Indicates Non-Transit Easement |
| **'P'** | Indicates Piers |
| **'R'** | Indicates Railroads |
| **'S'** | Indicates Street |
| **'U'** | Indicates U.S. Government |

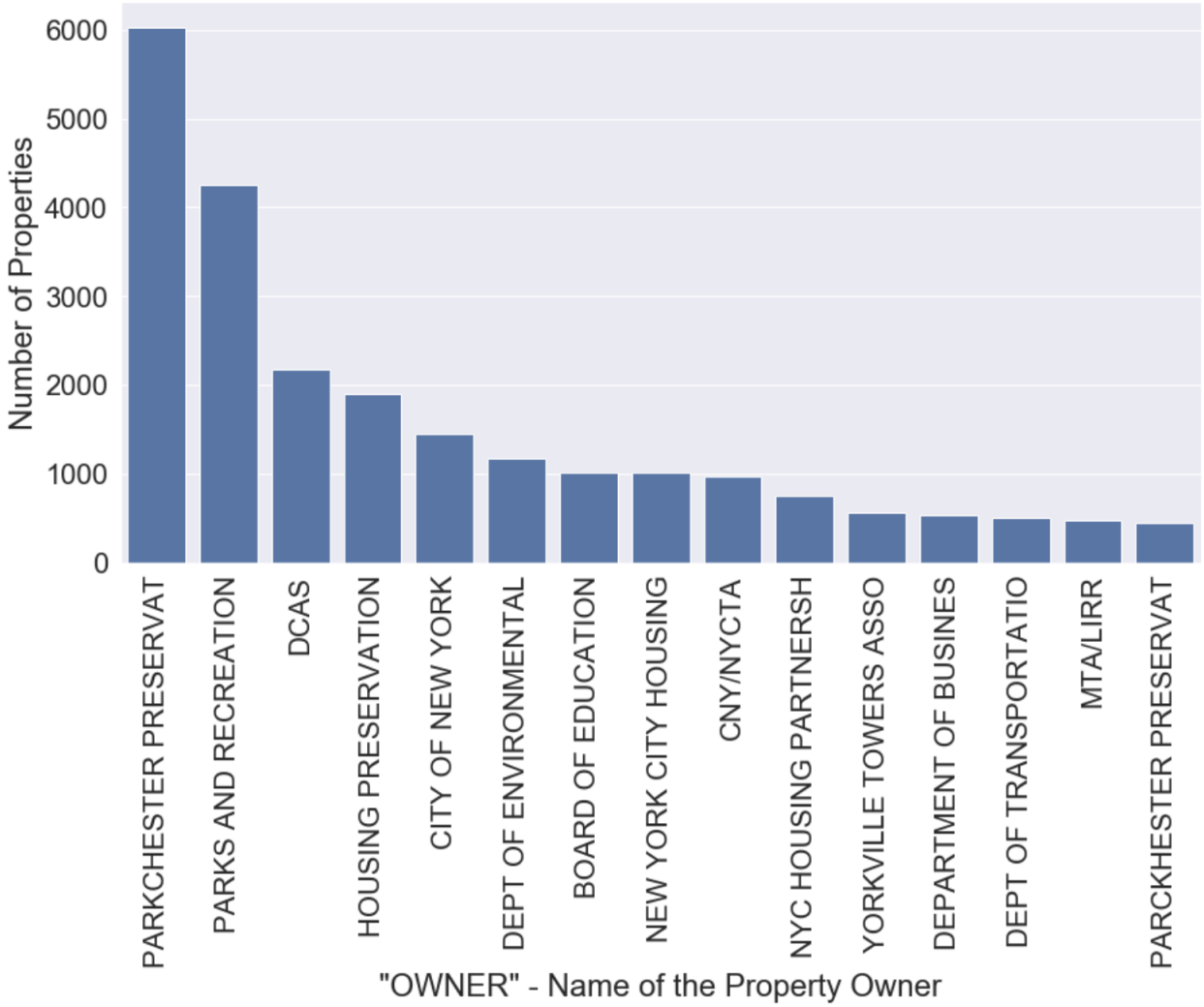
The bar chart below provides the number of property records for each easement code in the dataset. For those property records that have an easement code, easement code “E” was the most common value and it indicates that the property record has a Land Easement.



## 3.7 Field 7: OWNER

**Description**: A categorical data field containing the name of the owner for that particular property record. This data field is 97% populated in the dataset and there are 31,745 property records without an owner listed.

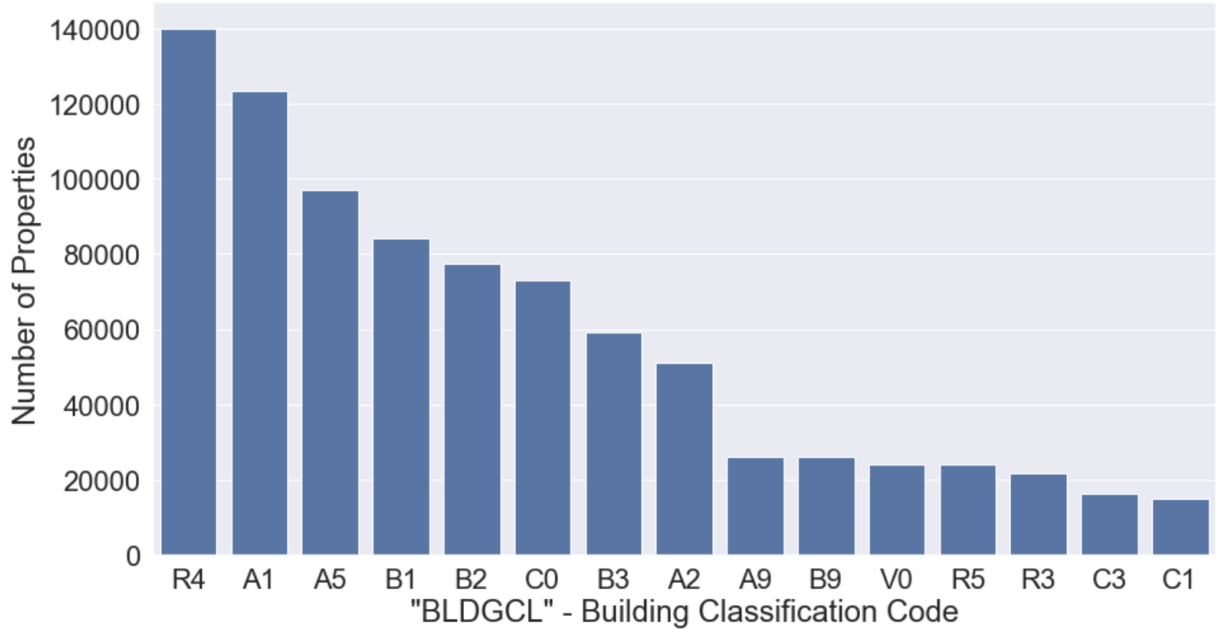
The bar chart below provides the top 15 owners with the most property records listed in the dataset. Of the 863,346 property owners listed in the dataset, Parkchester Preservation Management has the most property records. However, as we can see in the chart below, “PARKCHESTER PRESERVAT” has two different bars or calculations and so there may be some variances or discrepancies in the names of the owners and/or the data entry for this data field.



## 3.8 Field 8: BLDGCL

**Description**: A 2-character length alphanumeric data field containing the building classification code where there is a direct correlation between the building classification code and the tax classification code. The first character in the building classification code is a letter and the second character is a number. All records in the dataset contain a building classification code.

The bar chart below provides the top 15 building classification codes with the most property records listed in the dataset. Of the 200 unique values in this data field, R4 is the most common code.




## 3.9 Field 9: TAXCLASS

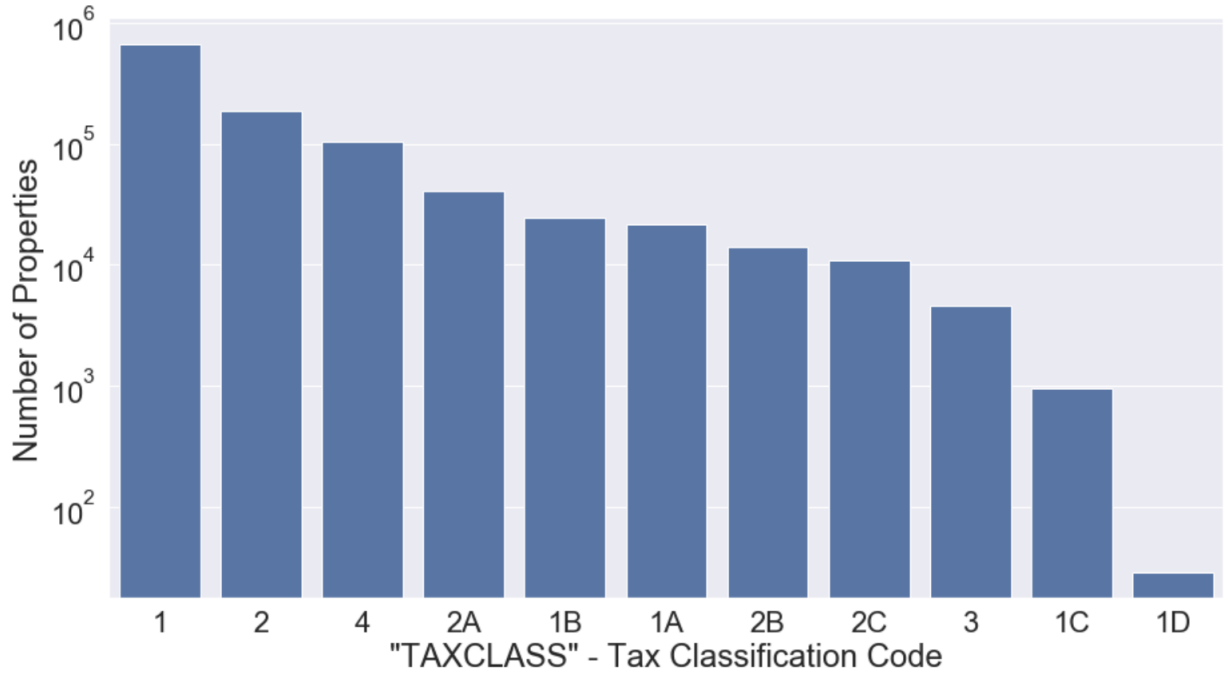
**Description**: A 2-character length alphanumeric data field containing the tax classification code based on knowledge of the building classification code (reference Section 3.8 for information related to the building classification code). All records in the dataset contain a tax classification code and the valid tax classification codes are as follows:

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

Additionally, the first character of the tax classification code is assigned based on the following building classification codes:

| Tax Classification Code | Building Classification Code |
|---|---|
| **1** | A0 - A9, B1 - B9, C0, G0, R3, R6, R7, S0 - S2, V0, V2, V3, Z0 |
| **2** | C1 - C9, D0 - D9, R0, R1, R2, R4, R8, R9, S3, S4, S5, S9 |
| **3** | U1 - U2, U4 - U9 |
| **4** | All others |

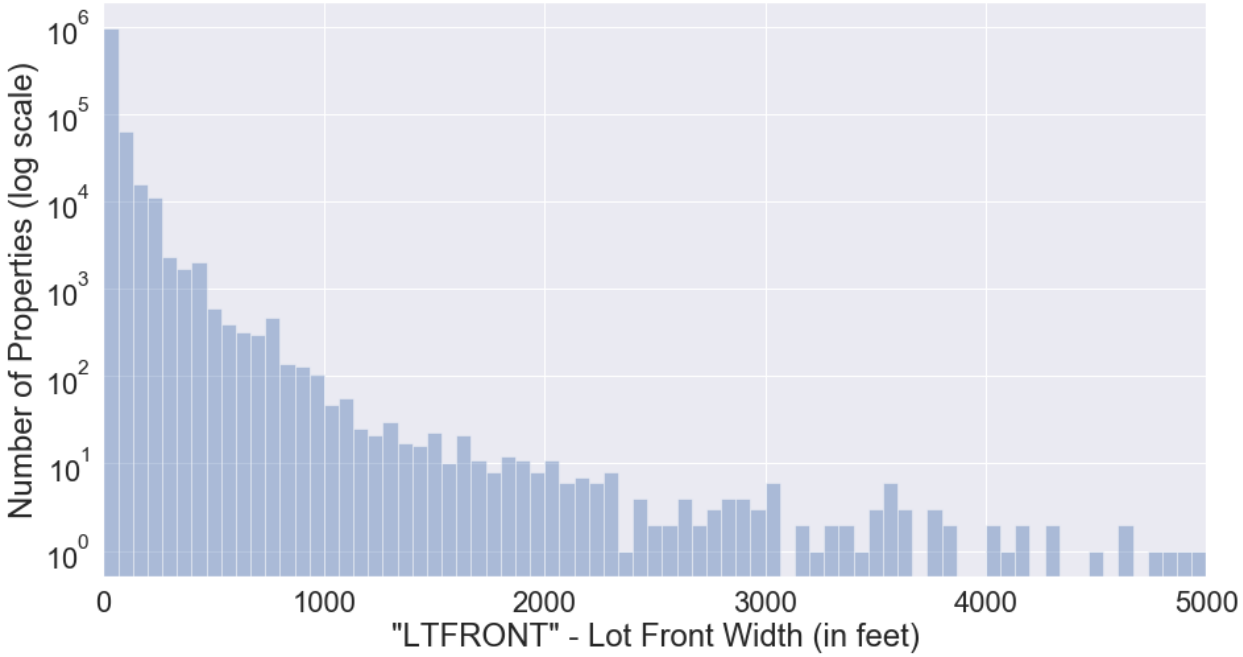
The bar chart below provides all tax classifcation codes listed in the dataset and the number of property records associated with each of the tax classifcation codes. Tax classifcation code 1 is the most common value for this data field.



## 3.10 Field 10: LTFRONT

**Description**: An integer data field containing the front width of the lot measured in feet. All property records in the dataset contain a front width lot measurement.

The distribution plot below shows the front width lot measurements up to 5,000 feet for the property records in the dataset. The widest measurement for the front width is 9,999 feet. Excluding those property records with a front width lot measurement of zero feet, a measurement of 20 feet for a property lot's front width has the most property records.



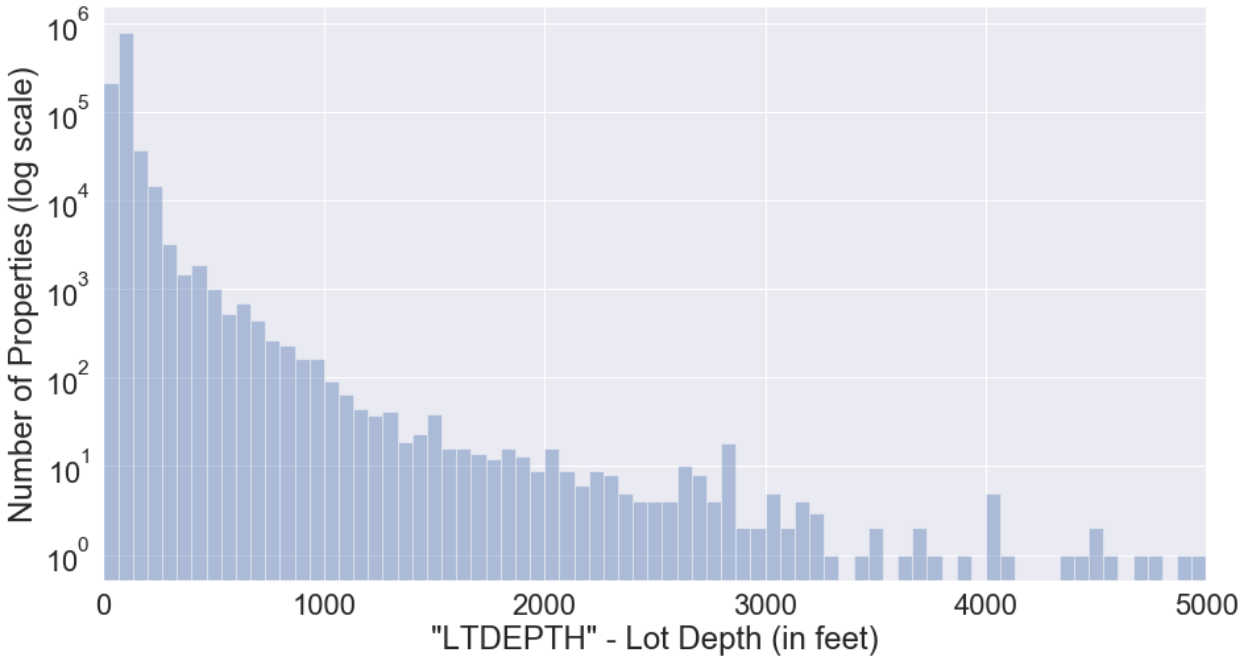

For a better understanding of the most common LTFRONT values, the table below is in descending order from the most to least frequent measurement for the top 10 most frequent values.

| LTFRONT Measurement | Number of Properties |
|---|---|
| **0** | 169,108 |
| **20** | 135,178 |
| **25** | 117,306 |
| **40** | 85,389 |
| **18** | 40,668 |
| **50** | 39,897 |
| **30** | 36,340 |
| **24** | 25,710 |
| **19** | 25,381 |
| **22** | 23,420 |

## 3.11 Field 11: LTDEPTH

**Description**: An integer data field containing the depth of the lot measured in feet. All property records in the dataset contain a measurement for the property lot's depth.

The distribution plot below shows the lot depth measurements up to 5,000 feet for the property records in the dataset. The longest measurement for a property lot's depth is 9,999 feet. The most common property lot depth is 100 feet.




For a better understanding of the most common LTDEPTH values, the table below is in descending order from the most to least frequent measurement for the top 10 most frequent values.

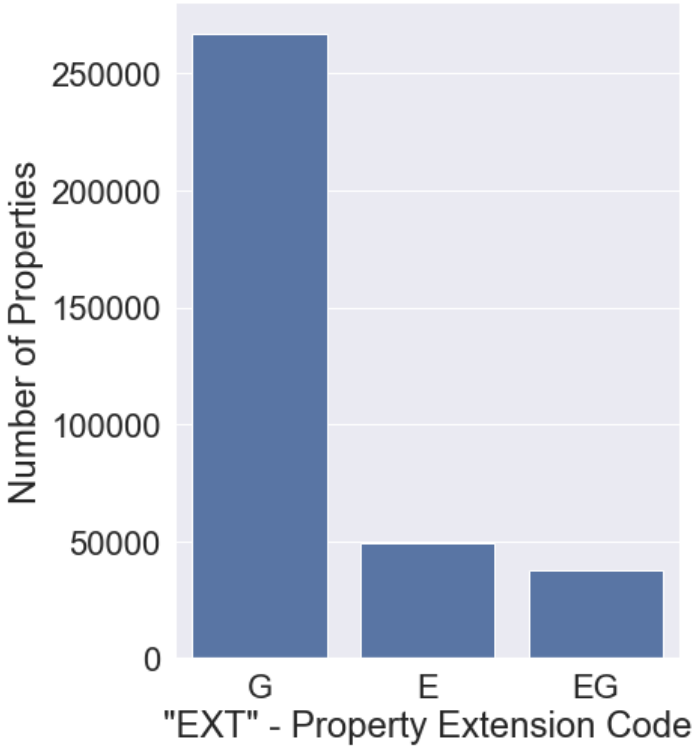
| LTDEPTH Measurement | Number of Properties |
|---|---|
| **100** | 464,541 |
| **0** | 170,128 |
| **95** | 31,612 |
| **90** | 20,294 |
| **80** | 16,671 |
| **99** | 11,390 |
| **97** | 10,227 |
| **75** | 10,161 |
| **102** | 9,607 |
| **96** | 9,329 |

## 3.12 Field 12: EXT

**Description**: A categorical data field containing a letter code for the type of extension that may be associated with that particular property record. This data field is only 33% populated with 354,305 property records containing an extension code. The letter codes are representative as follows:

| Extension Code | Description |
|---|---|
| **'E'** | EXTENSION |
| **'G'** | GARAGE |
| **'EG'** | EXTENSION AND GARAGE |

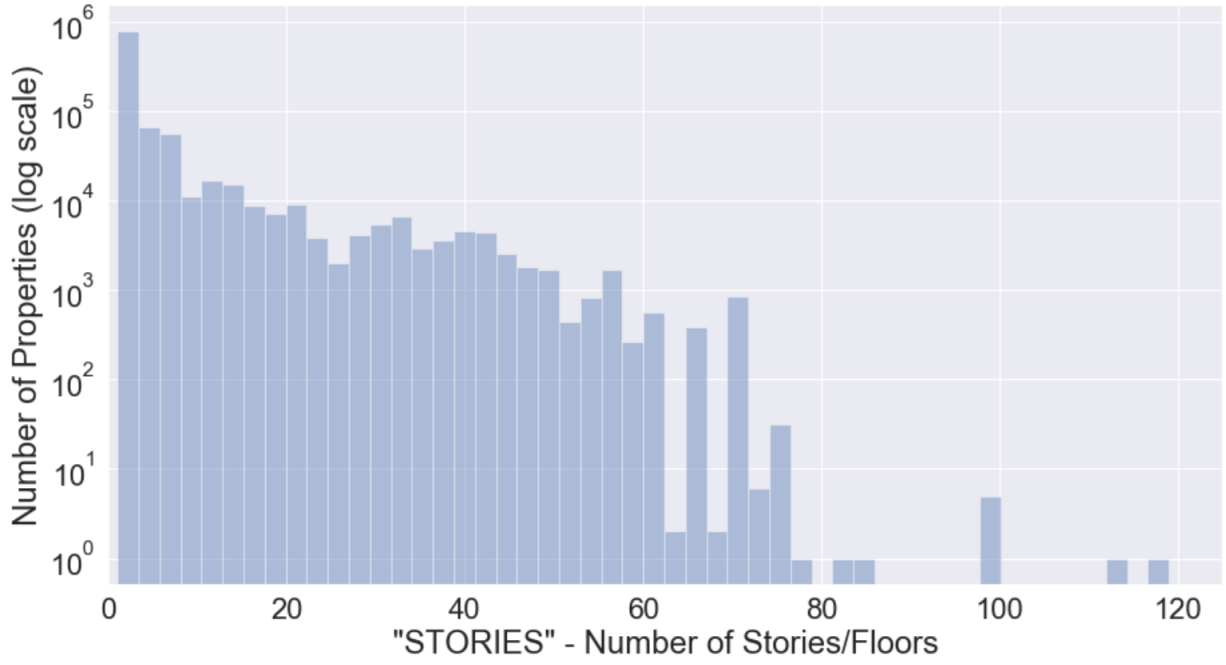
The bar chart below lists all extension code values listed in the dataset. When a property records has an extension code, the most common value is “G” to annotate that the property has a garage.



## 3.13 Field 13: STORIES

**Description**: A numerical data field containing the number of stories (i.e. floors) in the building. This data field is 94% populated and has 56,264 property records with no value for the number of stories listed.

The distribution shows the number of stories up to 119 stories for the property records in the dataset. The most common number of stories is two, and the highest number of stories for a building seen in the dataset is 119.



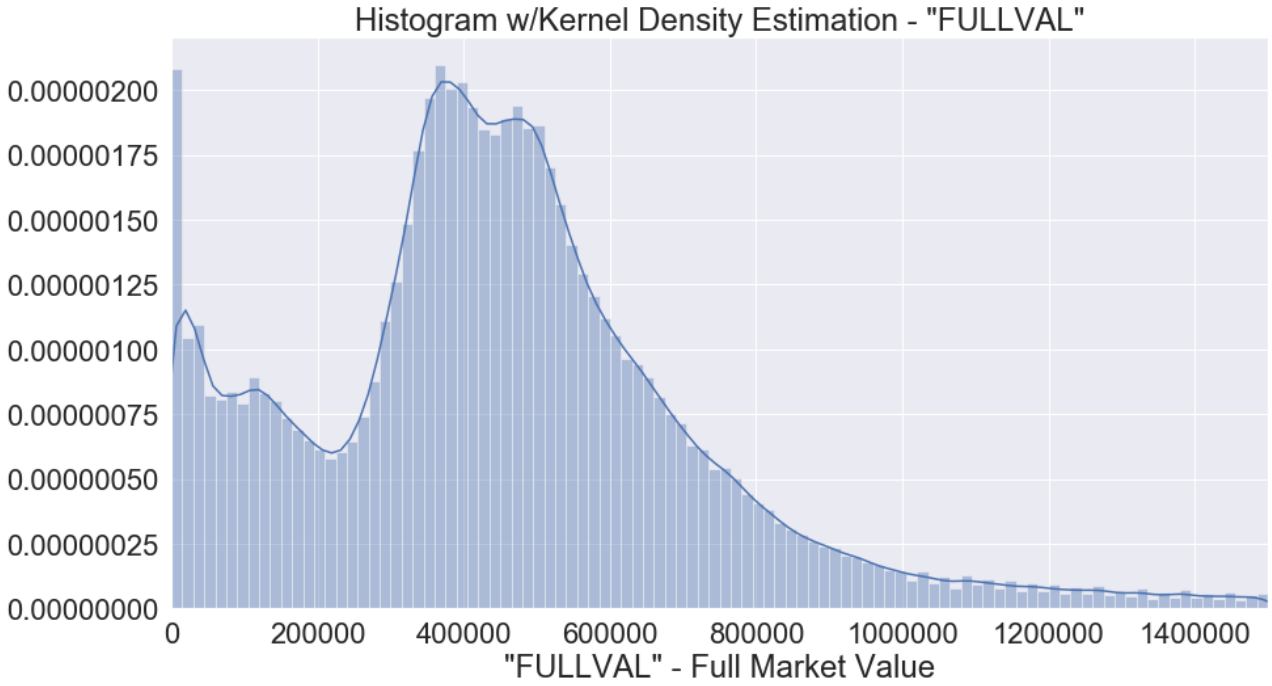

For a better understanding of the most common STORIES values, the table below is in descending order from the most to least frequent number of stories/floors for the top 10 values.

| Number of STORIES | Number of Properties |
|---|---|
| **2.0** | 415,092 |
| **3.0** | 130,127 |
| **1.0** | 96,706 |
| **2.5** | 82,292 |
| **4.0** | 38,342 |
| **6.0** | 30,936 |
| **5.0** | 25,971 |
| **1.5** | 24,770 |
| **2.7** | 13,595 |
| **12.0** | 12,198 |

## 3.14 Field 14: FULLVAL

**Description**: A numerical data field containing the full market value of the property. All property records in the dataset have a full market value listed and the range is from $0 to $6.15 billion.

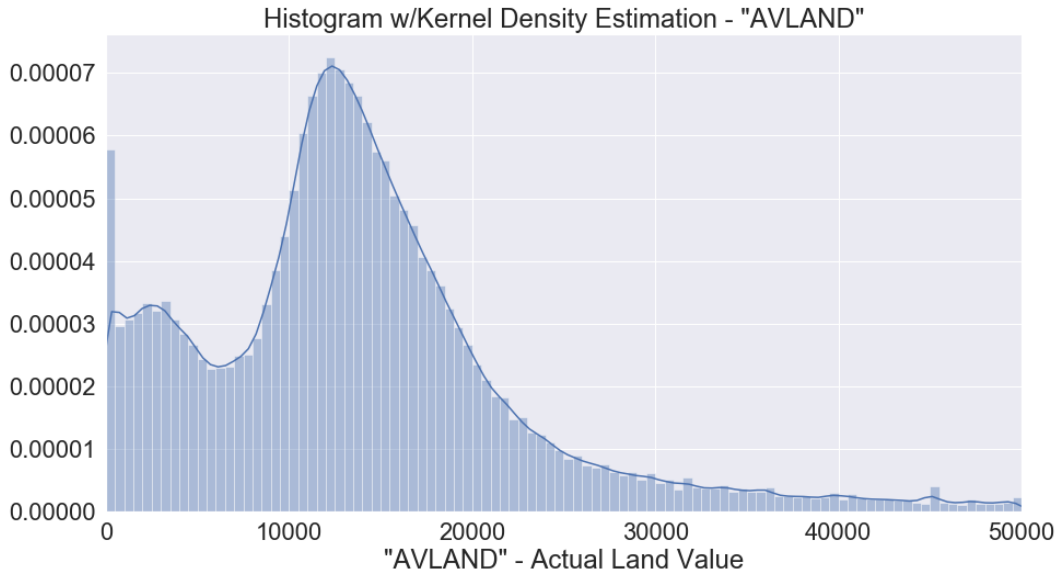
The graph below is a histogram of the full market value data field for the property records listed in the dataset having up to a value of $1,500,000.



## 3.15 Field 15: AVLAND

**Description**: A numerical data field containing the actual value of the land. All records in the dataset contain a value for this data field and the range is from $0 to $2.6685 billion.

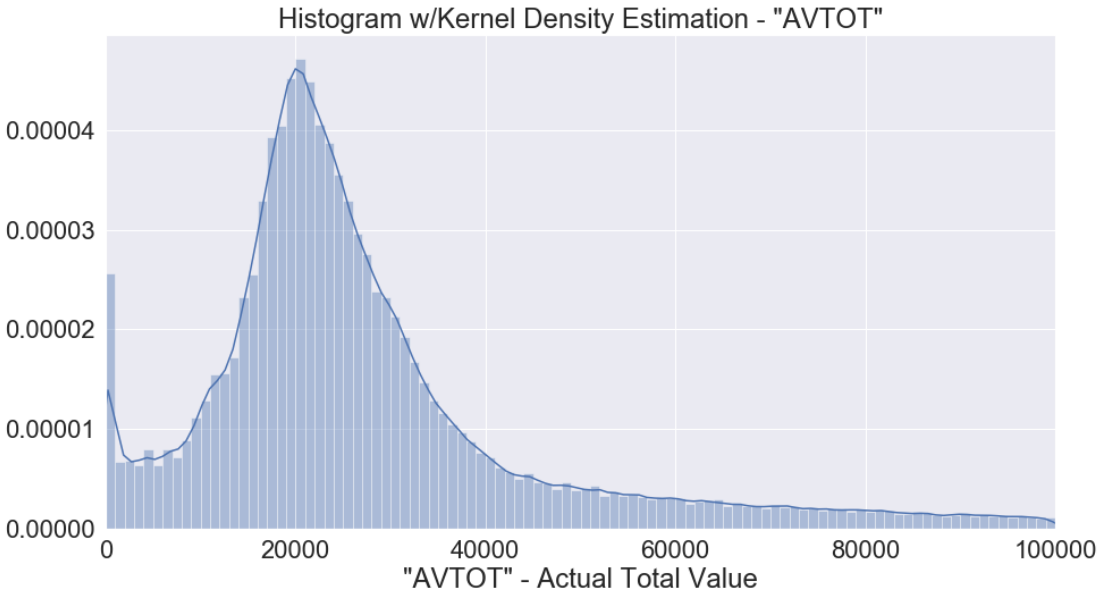
The graph below is a histogram of the actual land value data field for the property records listed in the dataset having up to a value of $50,000.




## 3.16 Field 16: AVTOT

**Description**: A numerical data field containing the actual total value of the property. All records in the dataset contain a value for this data field and the range is from $0 to $4.6683 billion.

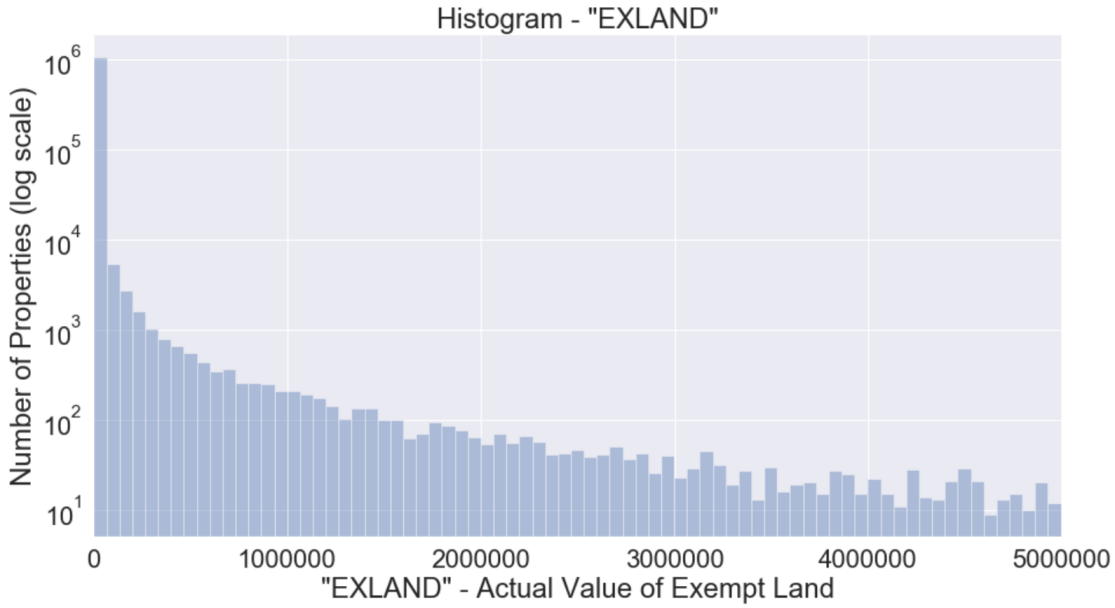
The graph below is a histogram of the actual total value data field for the property records listed in the dataset having up to a value of $100,000.



## 3.17 Field 17: EXLAND

**Description**: A numerical data field containing the actual value of the exempt land. All records in the dataset contain a value for this data field and the range is from $0 to $2.6685 billion.

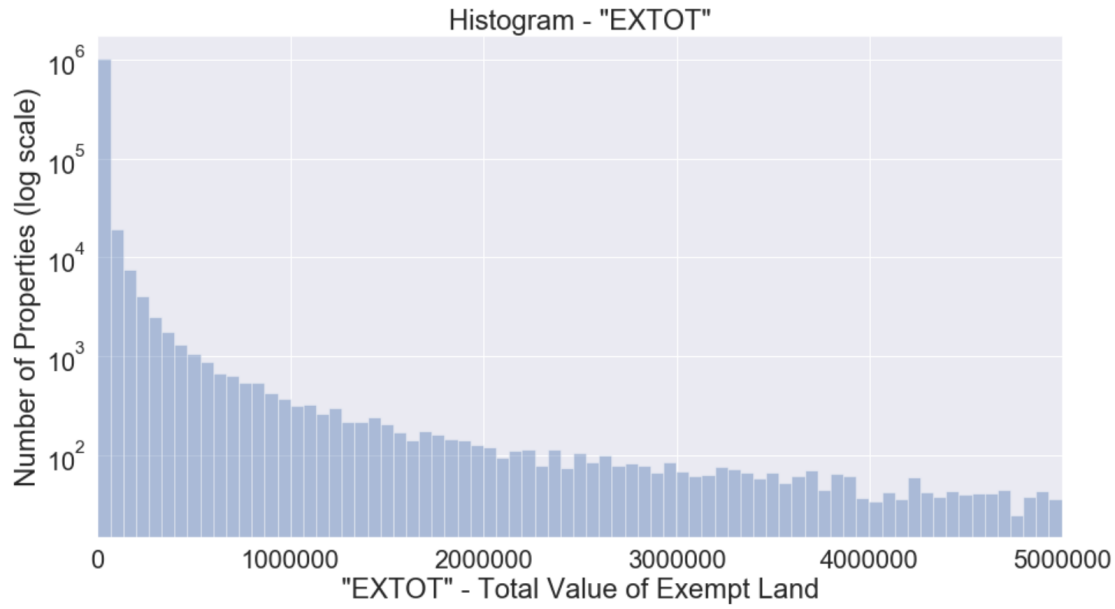
The graph below is a histogram of the actual value of exempt land data field for the property records listed in the dataset having up to a value of $5,000,000.




## 3.18 Field 18: EXTOT

**Description**: A numerical data field containing the actual total value of the exempt land. All records in the dataset have a value for this data field and the range is from $0 to $4.6683 billion.

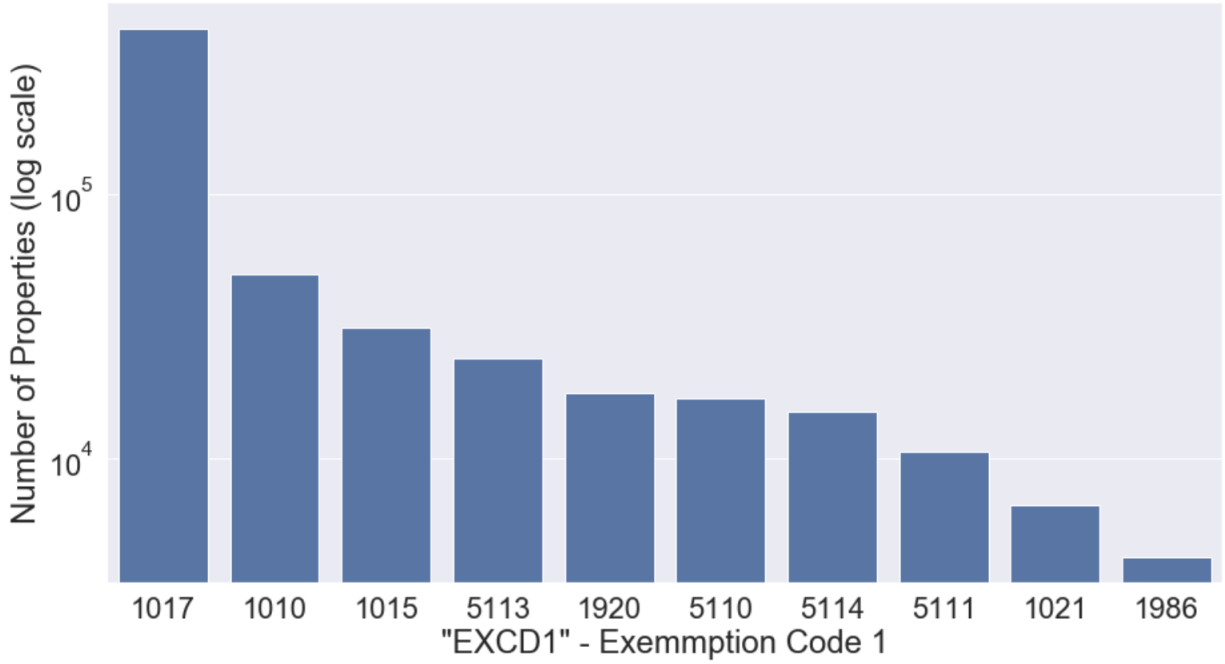
The graph below is a histogram of the actual total value of the exempt land data field for the property records listed in the dataset having up to a value of $5,000,000.



## 3.19 Field 19: EXCD1

**Description**: A categorical data field containing exemption code 1 for the property. This data field is 59% populated with 638,488 property records having a value.

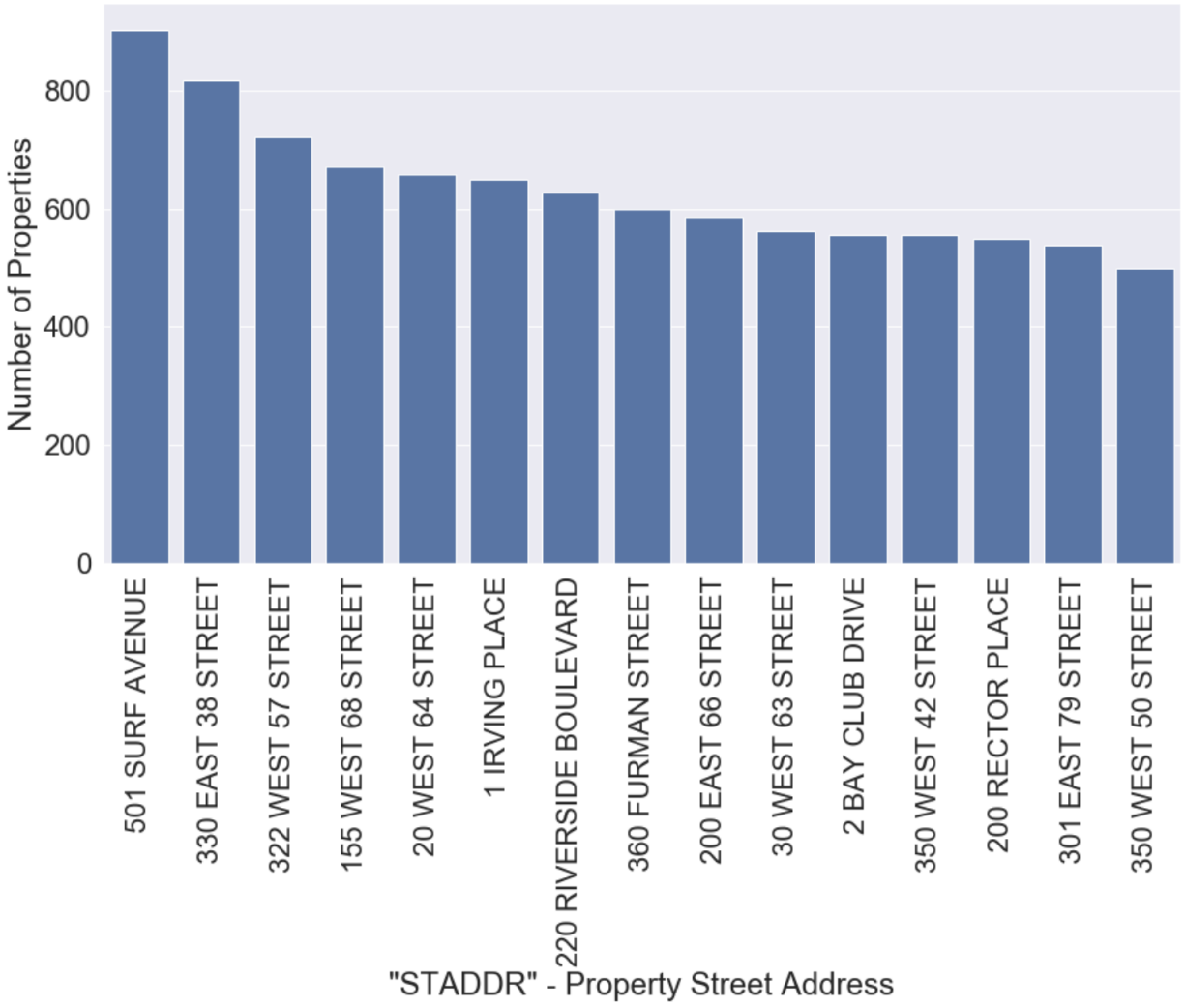
The bar chart below provides the top 10 exemption codes with the most property records. There is a total of 129 unique values for these codes and 1017 is the most common code.



## 3.20 Field 20: STADDR

**Description**: A categorical data field containing the street address of the property. This data field is 99% populated with only 676 property records missing a value.

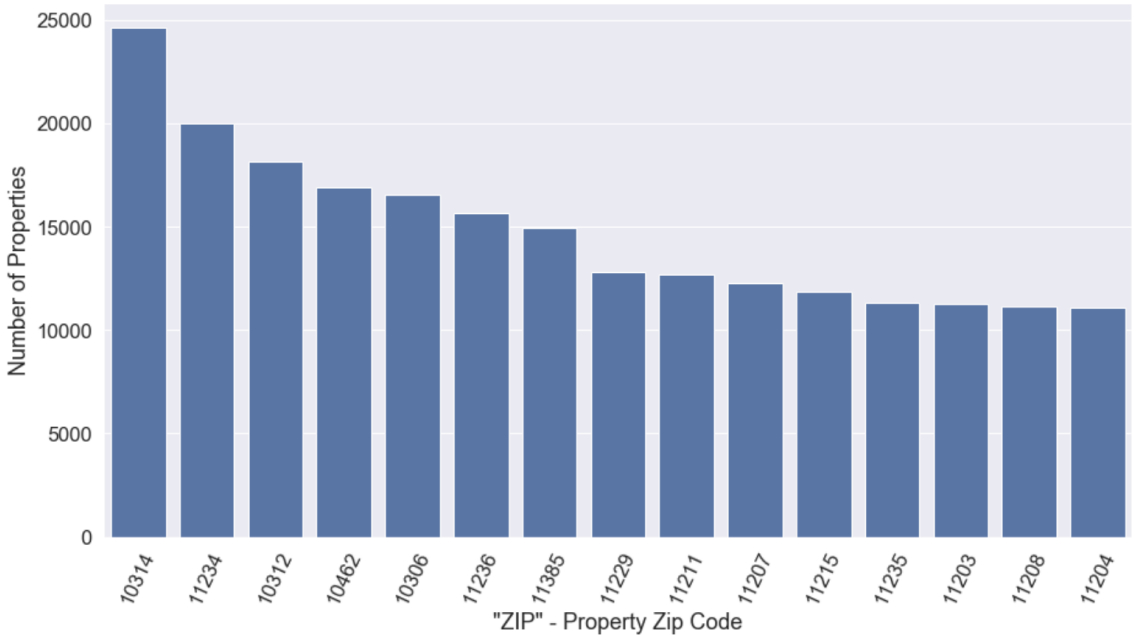
The bar chart below provides the top 15 street addresses with the most property records in the dataset.



## 3.21 Field 21: ZIP

**Description**: A categorical data field containing the zip code for the property. This data field is 97% populated with only 29,890 property records missing a value.

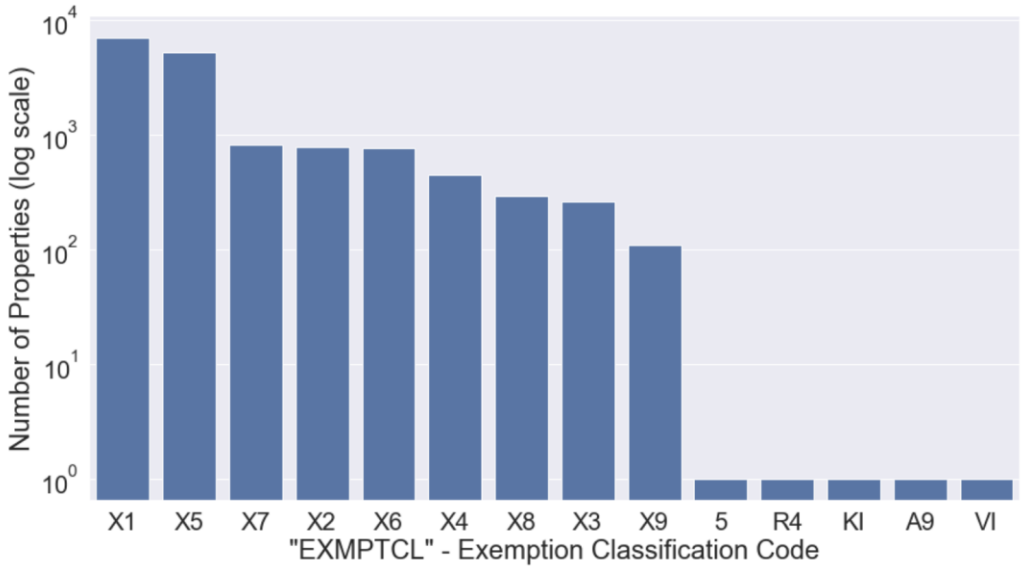
The bar chart below provides the top 15 zip codes with the most property records in the dataset. There are a total of 196 different zip codes in the dataset.




## 3.22 Field 22: EXMPTCL

**Description**: A 2-character alphanumeric data field containing the exemption classification code that is only used for fully exempt properties. This is a sparsely filled data field with only 15,579 property records containing a value.

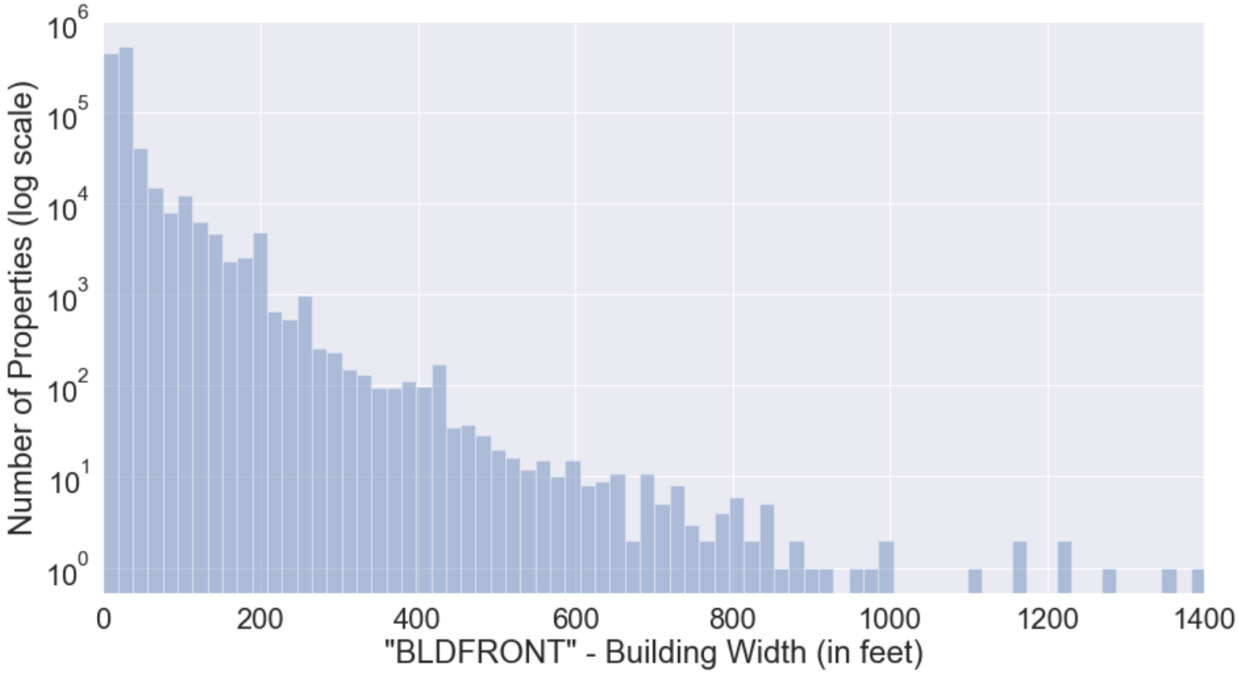
The bar chart below provides all exemption classification codes seen in the dataset.



## 3.23 Field 23: BLDFRONT

**Description**: An integer data field containing the front width of the building measured in feet. All property records in the dataset have a value for this data field.

The distribution plot below shows the building front width data field up to 1,400 feet for the property records in the dataset. Aside from a measurement of zero feet, 20 feet is the most common width. A measurement of 7,575 feet is the longest building width in the dataset.



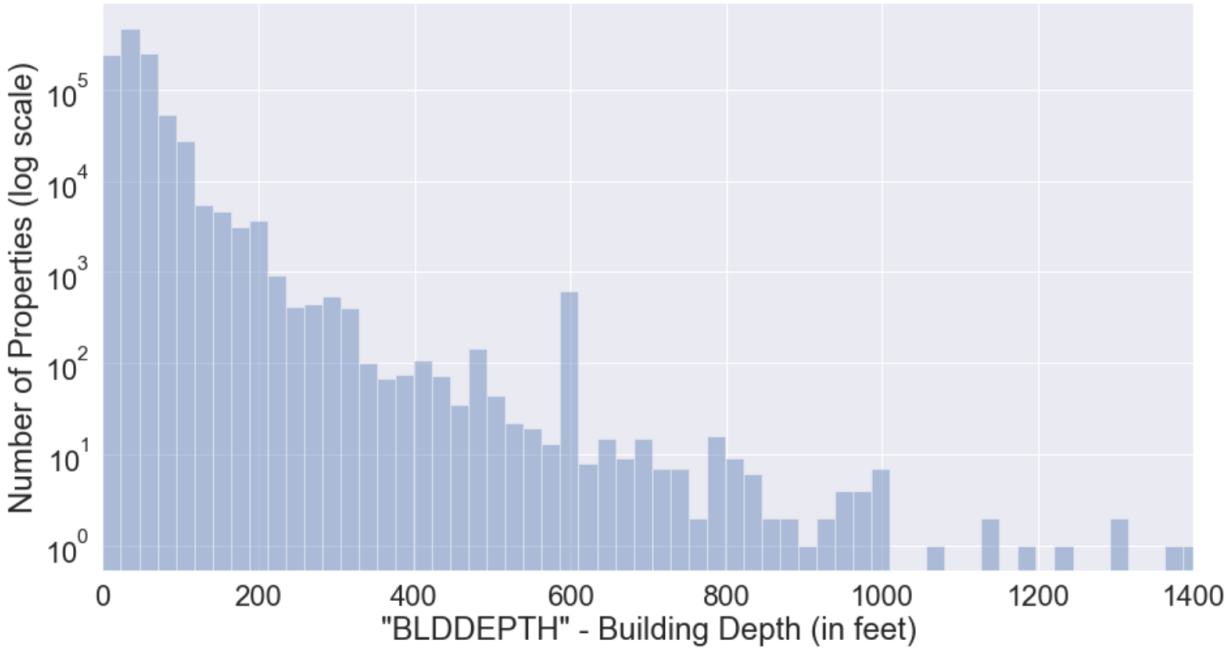

For a better understanding of the most common BLDFRONT values, the table below is in descending order from the most to least measurements for the top 10 most frequent values.

| BLDFRONT Measurement | Number of Properties |
|---|---|
| **0** | 228,815 |
| **20** | 195,101 |
| **18** | 77,705 |
| **16** | 74,687 |
| **25** | 63,684 |
| **22** | 54,297 |
| **24** | 33,486 |
| **19** | 33,383 |
| **21** | 32,904 |
| **26** | 29,445 |

## 3.24 Field 24: BLDDEPTH

**Description**: An integer data field containing the depth of the building measured in feet. All property records in the dataset have a value for this data field.

The distribution plot below shows the building depth data field up to 1,400 feet for the property records in the dataset. Aside from the measurement of zero feet, 40 feet is the most common depth and 9,393 feet is the longest depth of a building in the dataset.



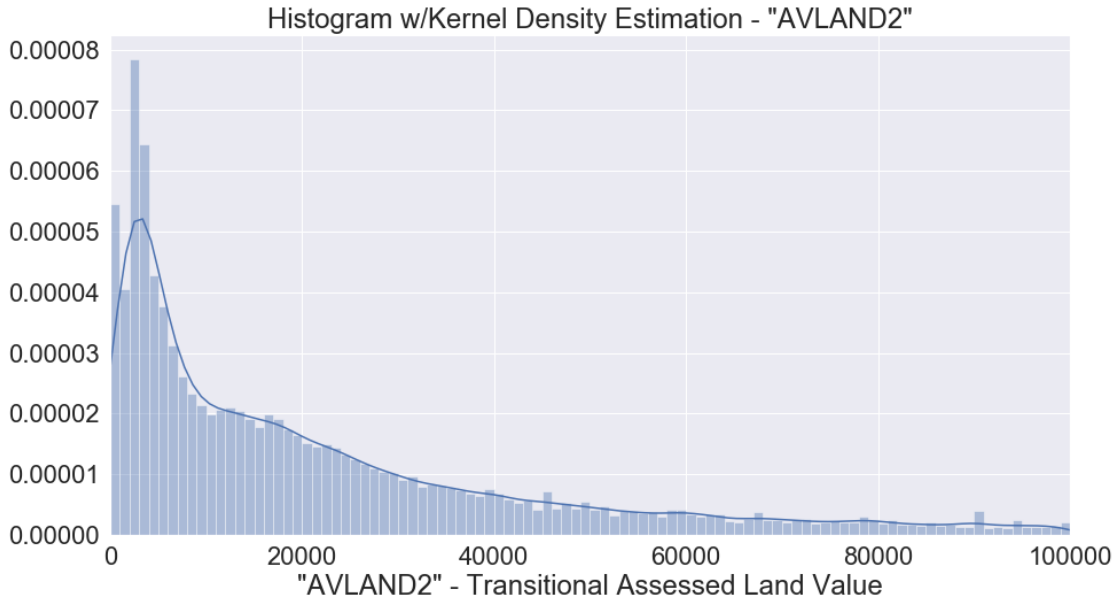

For a better understanding of the most common BLDDEPTH values, the table below is in descending order from the most to least measurements for the top 10 most frequent values.

| BLDDEPTH Measurement | Number of Properties |
|---|---|
| **0** | 228,853 |
| **40** | 48,775 |
| **50** | 45,358 |
| **45** | 40,670 |
| **36** | 40,109 |
| **30** | 31,553 |
| **35** | 29,054 |
| **38** | 28,202 |
| **55** | 27,830 |
| **42** | 26,248 |

## 3.25 Field 25: AVLAND2

**Description**: A numerical data field containing the transitional assessed land value. Only 26% of the property records have a value for this data field and the range of values is from $3 to $2.371 billion.

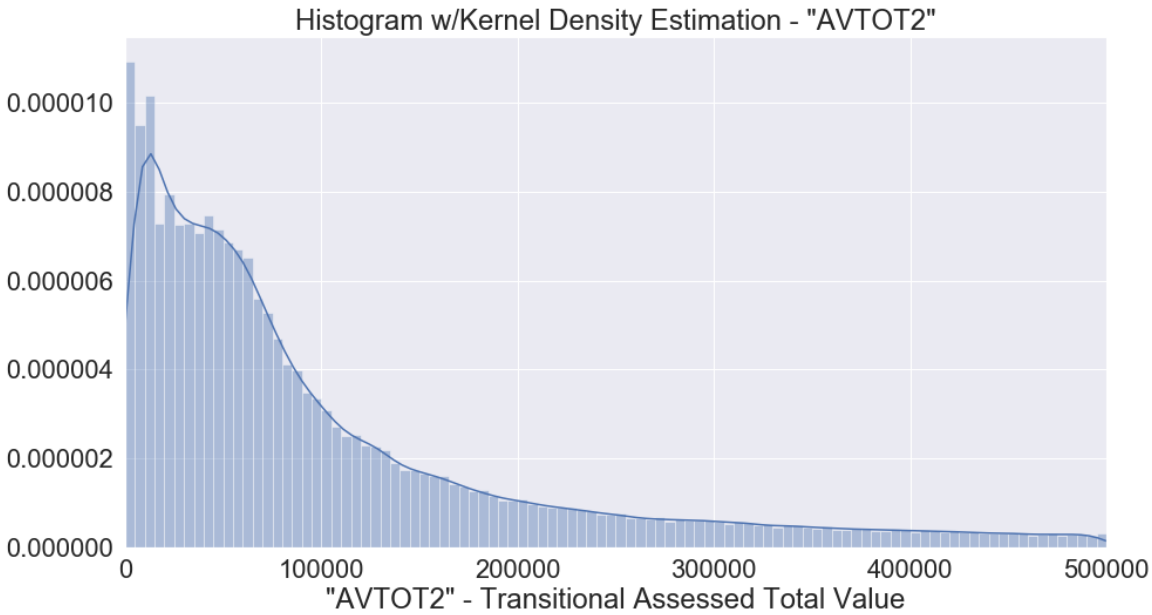
The graph below is a histogram of the transitional assessed land value data field for the property records listed in the dataset having a value up to $100,000.




## 3.26 Field 26: AVTOT2

**Description**: A numerical data field containing the transitional assessed total value. Only 26% of the property records have a value for this data field and the range of values is from $3 to $4.50118 billion.

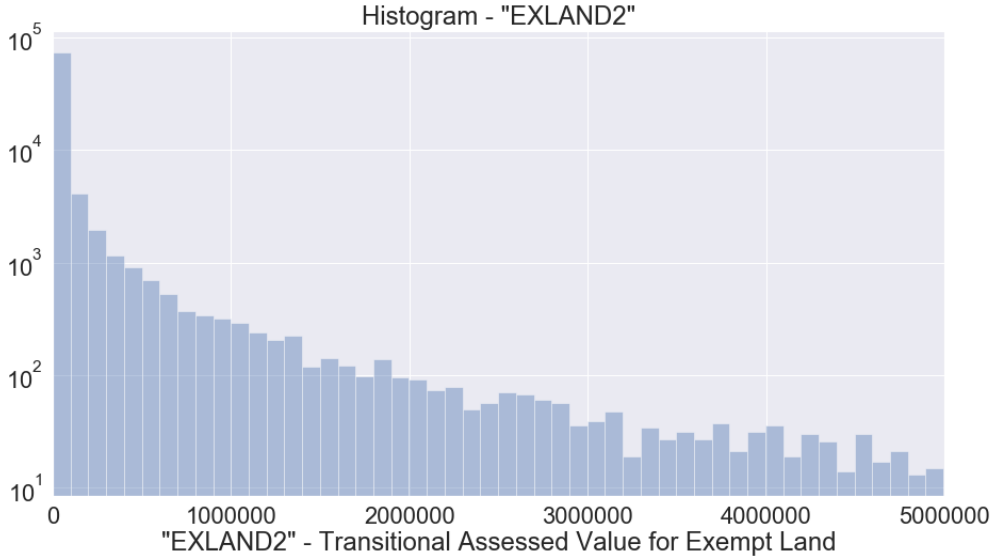
The graph below is a histogram for the transitional assessed total value data field for the property records listed in the dataset having a value up to $500,000.



## 3.27 Field 27: EXLAND2

**Description**: A numerical data field containing the transitional assessed value for exempt land. This data field is sparsely filled with only being 8% populated and 87,449 property records with a value. The range of values is from $1 to $2.371 billion.

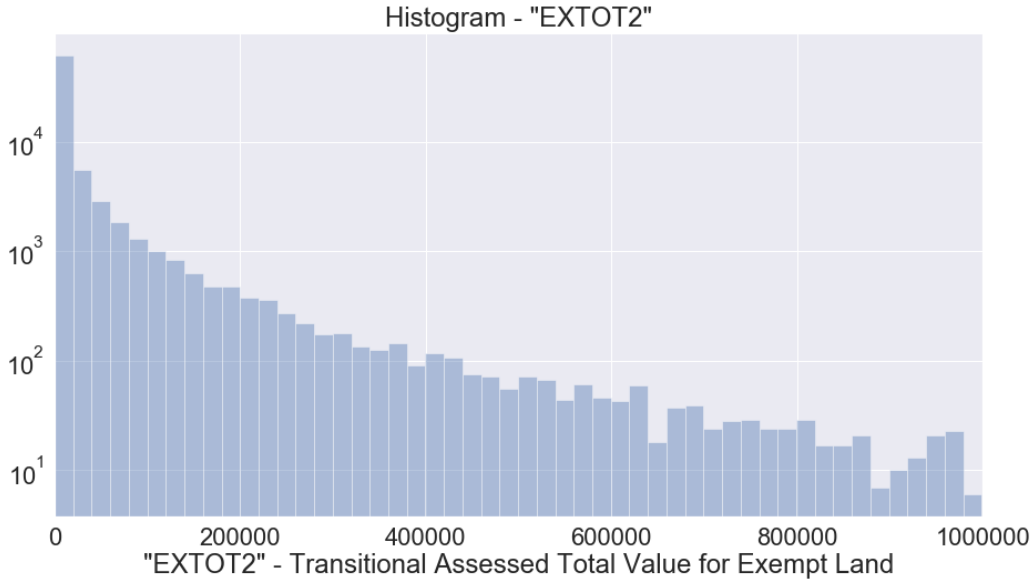
The graph below is a histogram of the transitional assessed value for exempt land data field for the property records in the dataset having a value up to $5,000,000.




## 3.28 Field 28: EXTOT2

**Description**: A numerical data field containing the transitional assessed total value for exempt land. This data field is sparsely filled with only being 12% populated and 130,828 property records with a value. The range of values is from $7 to $4.50118 billion.

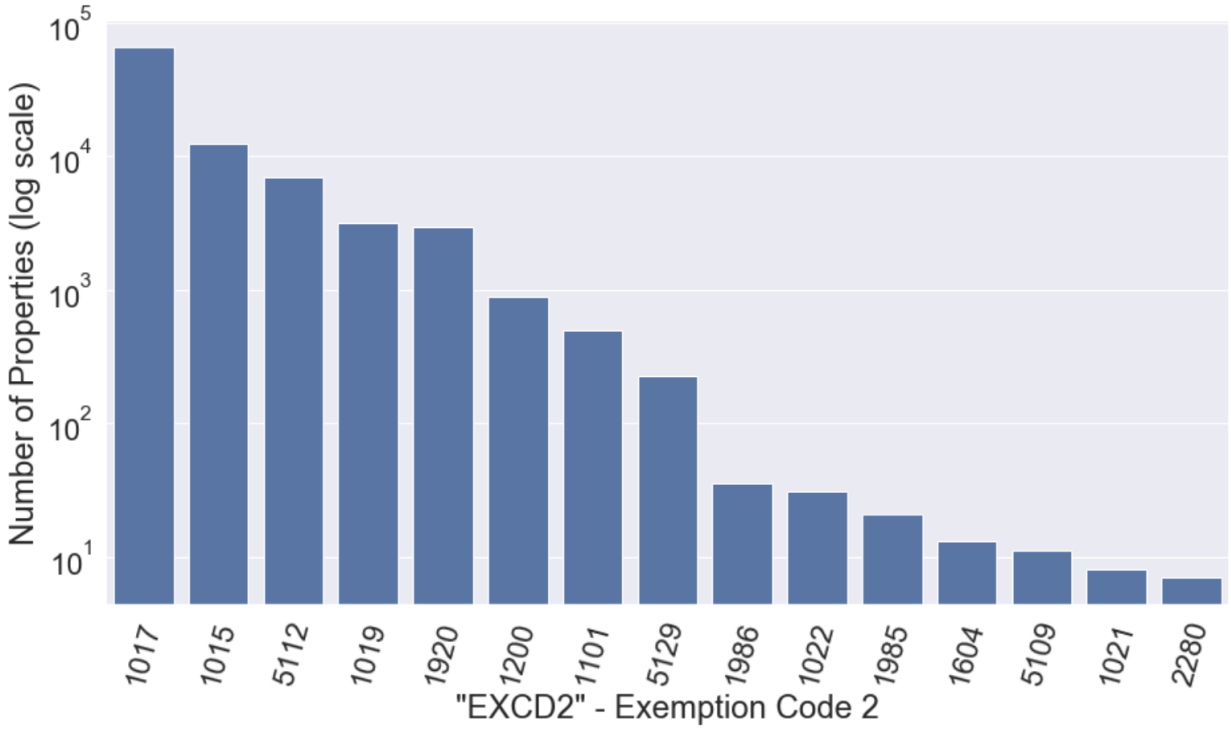
The graph below is a histogram of the transitional assessed total value for exempt land data field for the property records in the dataset having a value up to $1,000,000.



## 3.29 Field 29: EXCD2

**Description**: A categorical data field containing exemption code 2 for the property. This data field is sparsely filled with only being 8% populated and 92,948 property records having a value.

The bar chart below provides the top 15 exemption codes with the most property records in the dataset. There is a total of 60 unique values for this data field.




## 3.30 Field 30: PERIOD

**Description**: A data field containing the assessment period when the file was created as described in the Access Database provided by NYC Open Data; however, all records in the dataset simply contain the same value of "FINAL" in the data field. At this time and since all values are the same, the data field appears to have little to no value to the dataset.

## 3.31 Field 31: YEAR

**Description**: A data field containing the assessment year for the property based on the NYC's fiscal year (begins on July 1st of the calendar year and ends on the June 30th of the following calendar year). All records in the dataset contain the same value of "2010/11" in the data field. At this time and since all values are the same, the value of the data field appears to simply be a means to verify that all records are within the same fiscal year.

## 3.32 Field 32: VALTYPE

**Description**: A categorical data field with no description provided by the data owners (NYC Open Data) or the NYC Department of Finance; however, "VALTYPE" is often associated with the "value type" of a particular data field. This data field appears to have little to no value to the dataset at this time, and all records in the dataset contain the same value of "AC-TR" in the data field.