NYC PROPERTY FRAUD DETECTION

GROUP 3

New York City Property Tax Assessment

Fraud Analysis

Professor: Stephen Coggeshall

February 22, 2018

Team 3

Justin Ramirez
Roselyn Byrd
Rachel Feldman
Jane Eshagpoor
Linglan Mei
Hanci Wang
Nikhil Gupta

Table of Contents

Executive Summary

The purpose of this project is to analyze and identify potential fraud regarding the tax assessment for properties in the City of New York Property Valuation and Assessment Data using unsupervised machine learning methods. We used R and Python to generate the code for our diverse analysis, which included Z-Scaling, Principal Component Analysis (PCA) and designing an Autoencoder. The data set is comprised of over 1 million records of property recorded with owner’s name, address of property, block, borough, lot measurements, building measurements, zip-code as well as market value, assessed total value, and assessed value for the land.

This project involved several processes to clean the raw data to improve analysis. We used the three main valuation variables, full market value (FULLVAL), assessed value of land (AVLAND), and assessed total value (AVTOT), to construct expert variables incorporating the measurement variables: lot frontage, depth measurements, building frontage, depth measurements, and stories. Other variables were created using these expert variables to help us explain the maximum variation in our data.

Using 2 different algorithms, we calculated 2 fraud scores for each record, finally combining the scores to arrive at a final score used to sort the records. The top 10 records are enumerated in the Results & Insights section of this report.

- On average, properties with fraud potential usually have a significantly higher market value, assessed total value and land value compared to other properties.
- Upon analyzing the TAXCLASS of the top 100 high score records, we found that 63% of the properties belong to TAXCLASS 4 while in the whole dataset it’s less than 10%. Referring to definition, TAXCLASS 4 means “all commercial and industrial properties” and “all other”. This could indicate that fraud is being conducted by business owners of commercial establishments to reduce the property tax burden on their businesses.
- Most top scoring records have large house agencies, real estate companies and government entities as owners. Very few of these properties belong to single households

Data Description

The City of New York Property Valuation and Assessment Data file is a publicly available dataset posted by the Department of Finance on the City of New York Open Data website[1]. The dataset consists of records of more than a million properties across City of New York and information on their sizes, values, owner, building classes, tax classes, market values and assessed values, etc. The dataset contains a total of 1,048,575 records (rows) and 30 variables (columns). Among the variables, 13 are categorical, 14 are numeric, 2 are string, and 1 is a date variable. All records are from 2010/2011 tax year.

Some of the more important fields/variables used in our analysis are described below. These and all other fields are discussed in greater detail in the Data Quality Report provided as Appendix-I.

RECORD

RECORD is a nominal, categorical variable that is used as the unique identifier of each property record. There are a total of 1,048,575 unique values for RECORD with no missing values.

BBLE

BBLE is a nominal, categorical variable that is 11 characters long. These characters are a concatenation of the BORO (1st character), BLOCK (next 5 characters), LOT (next 4 characters), and EASEMENT (last character). It has 1,048,575 unique values with no missing values. This field was used to create a new variable 'BORO' during our analysis.

BLOCK

BLOCK is a numerical variable that categorizes each record into its block number within its BORO. The five BOROs and their corresponding BLOCK ranges are Manhattan (1 to 2,255), Bronx (2,260 to 5,958), Brooklyn (1 to 8,955), Queens (1 to 16,350), and Staten Island (1 to 8,050).[2] There were 13,949 unique values and no entries were missing.

[1] https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8
[2] Data Dictionary

LOT

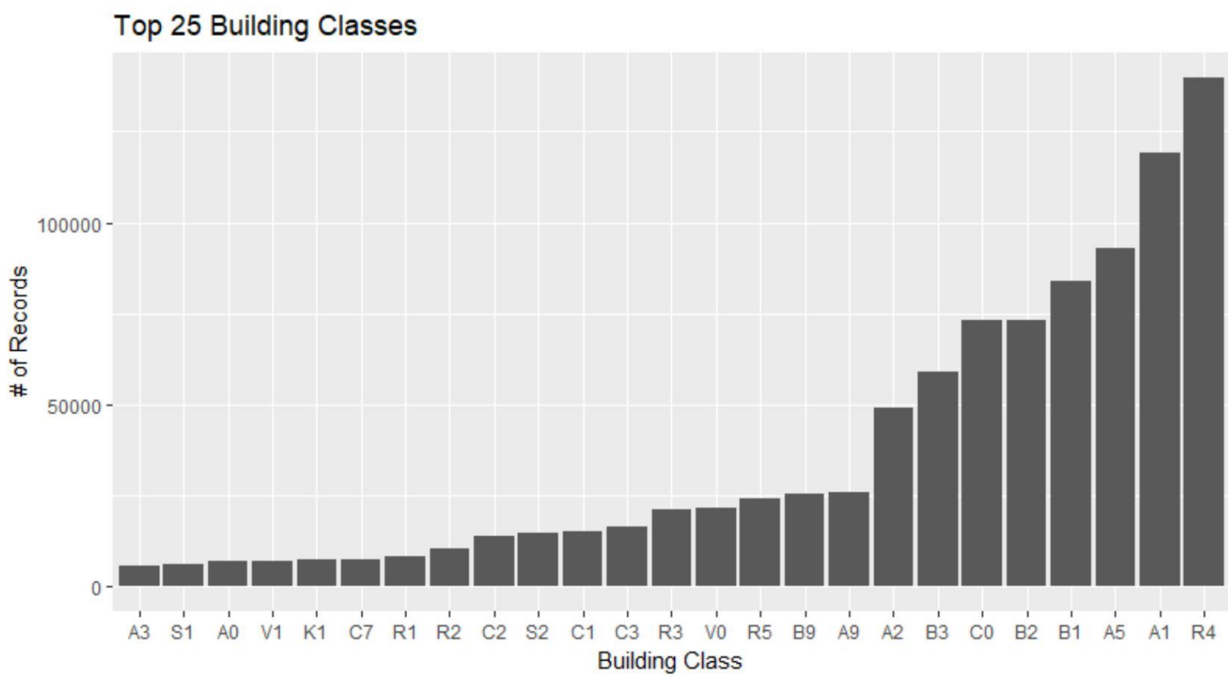
LOT is a nominal, categorical variable that represents the unique lot number for a property record within each BORO and BLOCK. It has 6,366 unique values and there were no missing values.

ZIP

ZIP is a 5-digit nominal, categorical variable that represents the zip code of property. ZIP has 197 unique values and 2.51% records are missing a ZIP value.

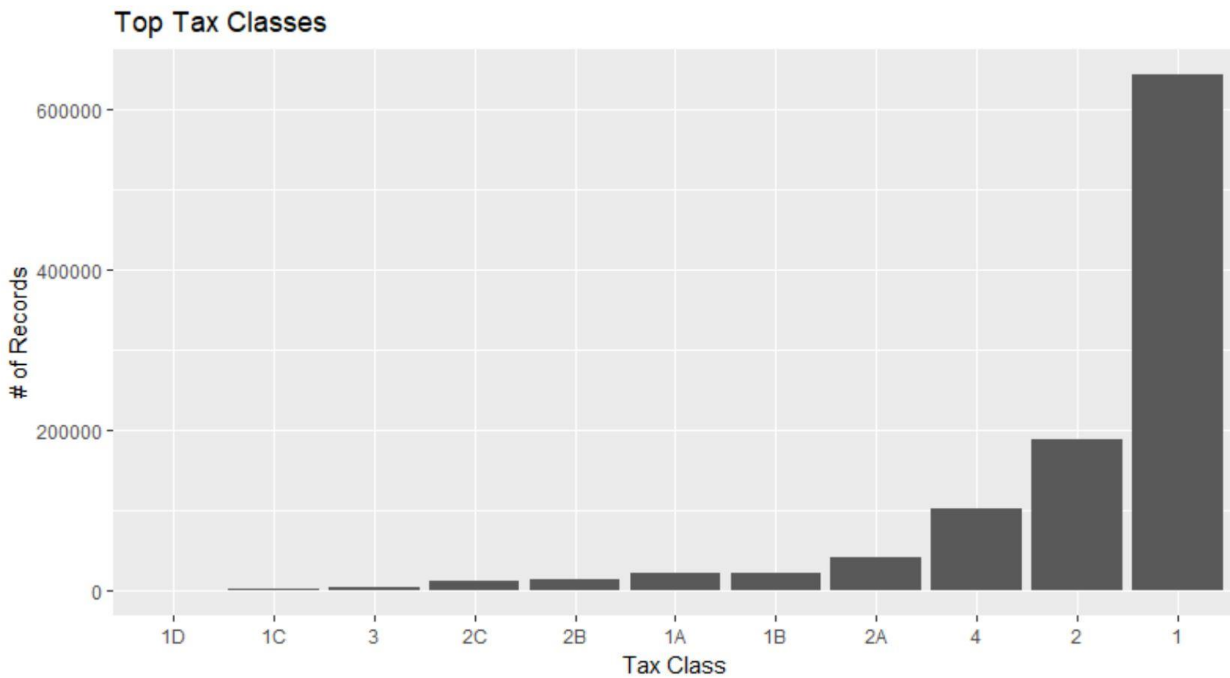
BLDGCL

BLDGCL (Building Class) is a 2-digit categorical variable. Position one is an alpha character and position two is a numeric. This is used to describe the class of the building and is directly correlated with tax class. There are 200 unique values for this field and there are no missing values. The top Building Classes are shown below:



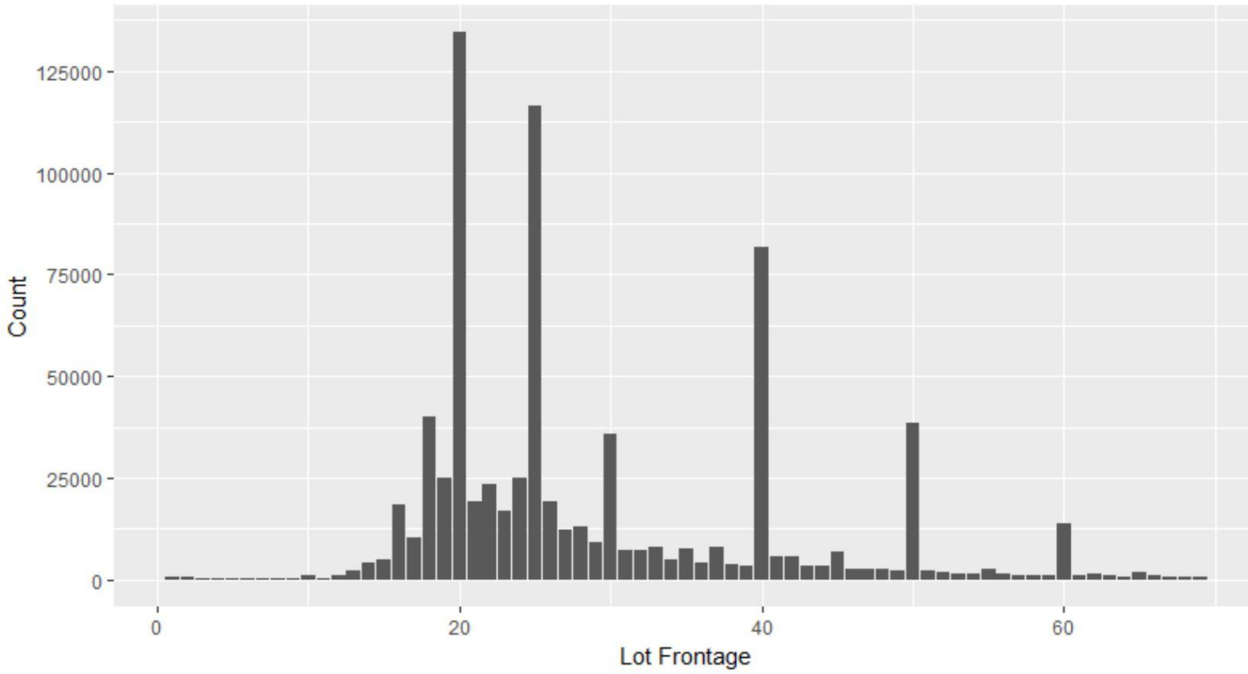
TAXCLASS

TAXCLASS is a categorical variable that represents the current property tax class code. It is directly correlated with BLDGCL field. There are 11 unique values, which are described in more detail in Appendix-I and there are no missing values. The top Tax Classes are shown below:



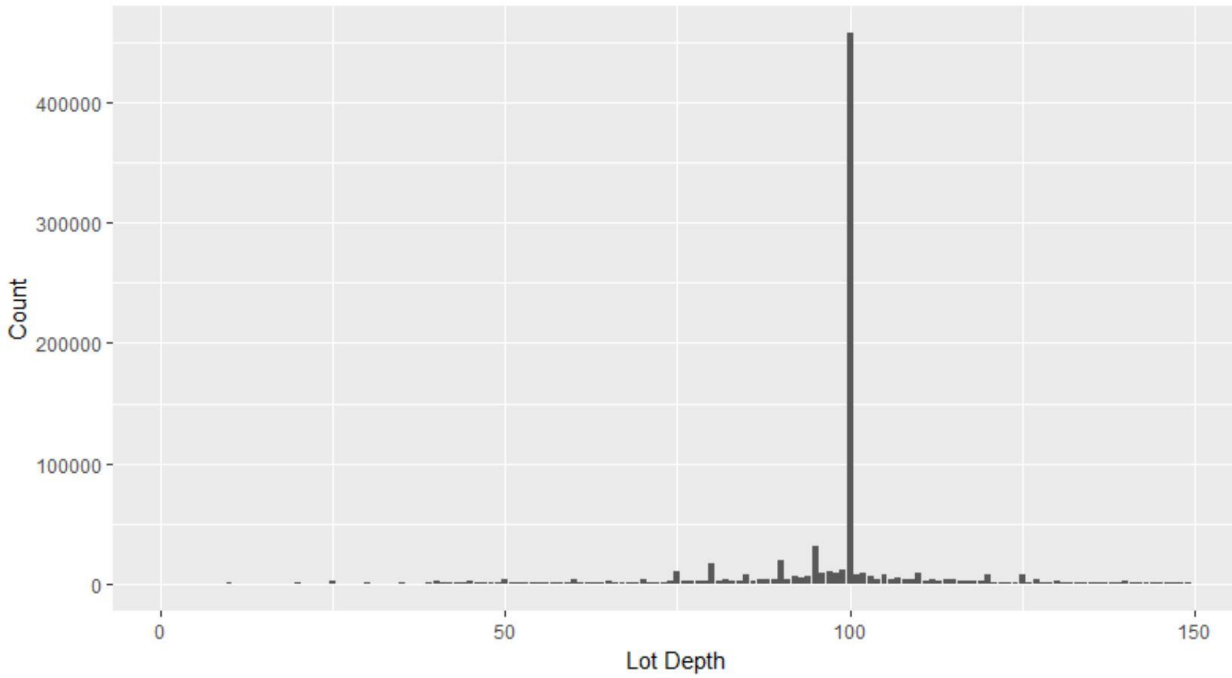
LTFRONT

LTFRONT (Lot Front) is a numeric variable that represents the length of the front of the lot in feet. There are 1,277 unique variables that range from 0 to 9,999. There are no missing values, however, 168,867 records have LTFRONT value of 0. We have considered these 0 values as missing in our analysis. The LTFRONT distribution is shown below:



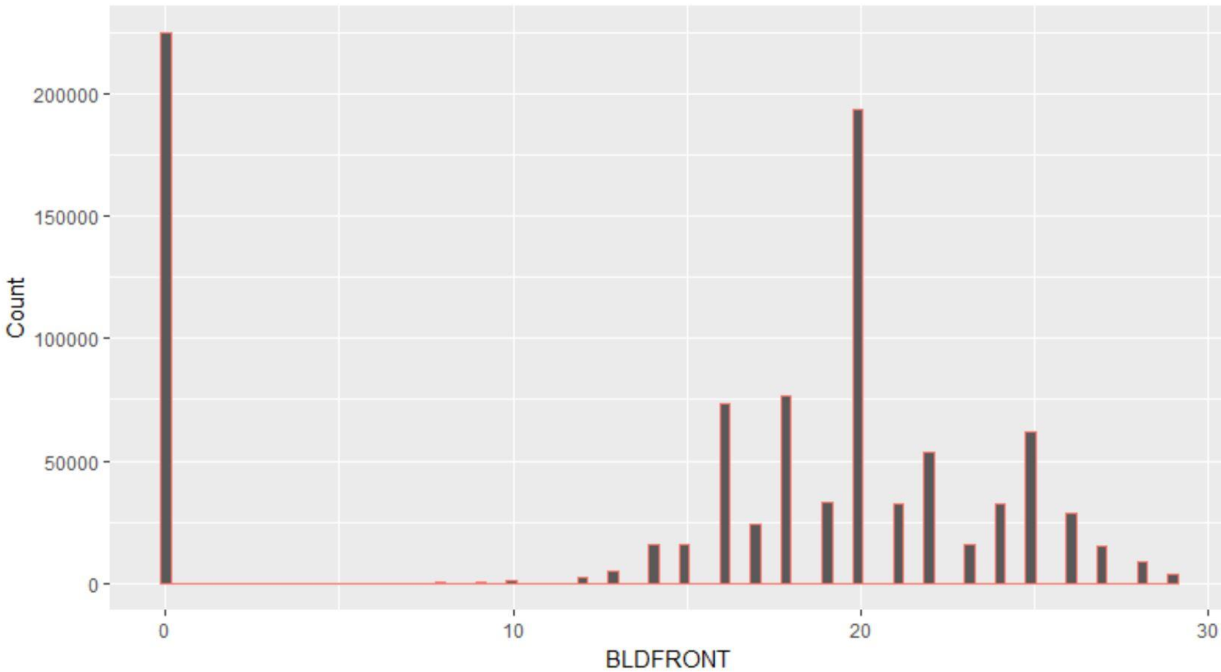
LTDEPTH

LTDEPTH (Lot Depth) is a numeric variable that represents the width/depth of the lot in feet. There are 1,336 unique values that range from 0 to 9,999. There are no missing values, however, 168,999 records have a LTDEPTH value of 0. We have considered these 0 values as missing in our analysis. The LTDEPTH distribution is shown below:



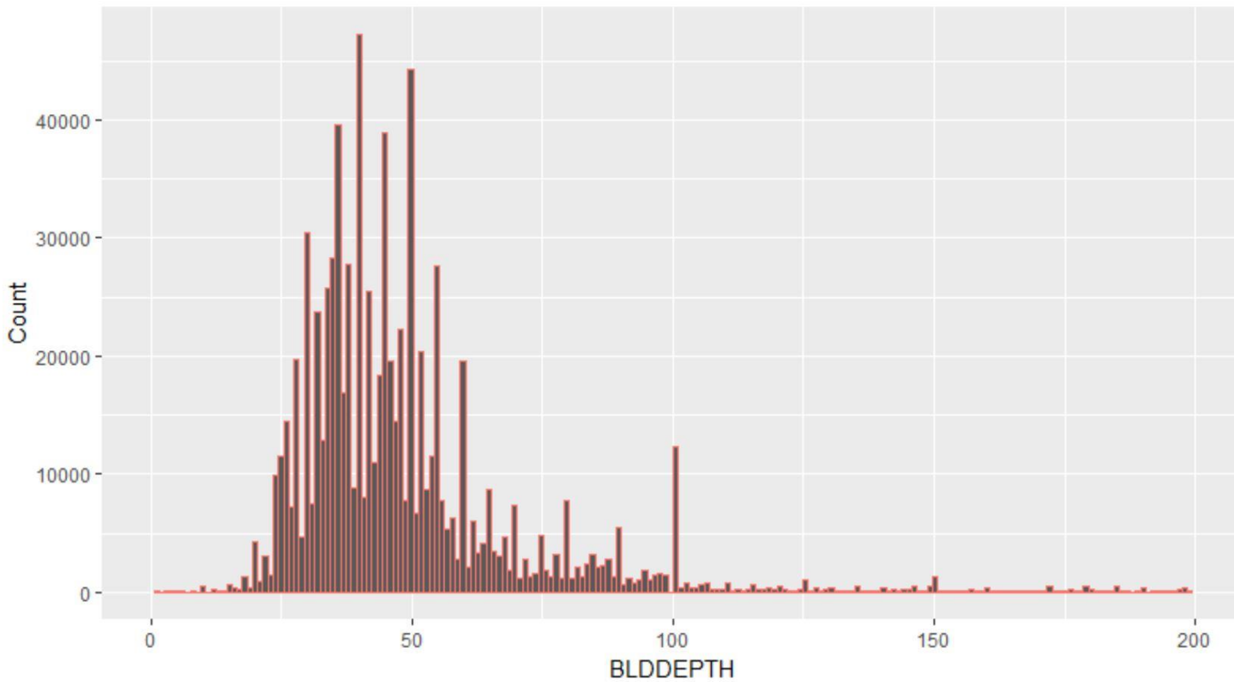
BLDFRONT

BLDFRONT (Building Front) is a numeric variable that represents the length of front of a building in feet. This field has 610 unique values that range from 0 to 7,575 feet (almost 1.5 miles). There are no missing values, however, there are 224,661 records that have BLDFRONT value of 0. We have considered these 0 values as missing in our analysis. The distribution of the field values is shown below:



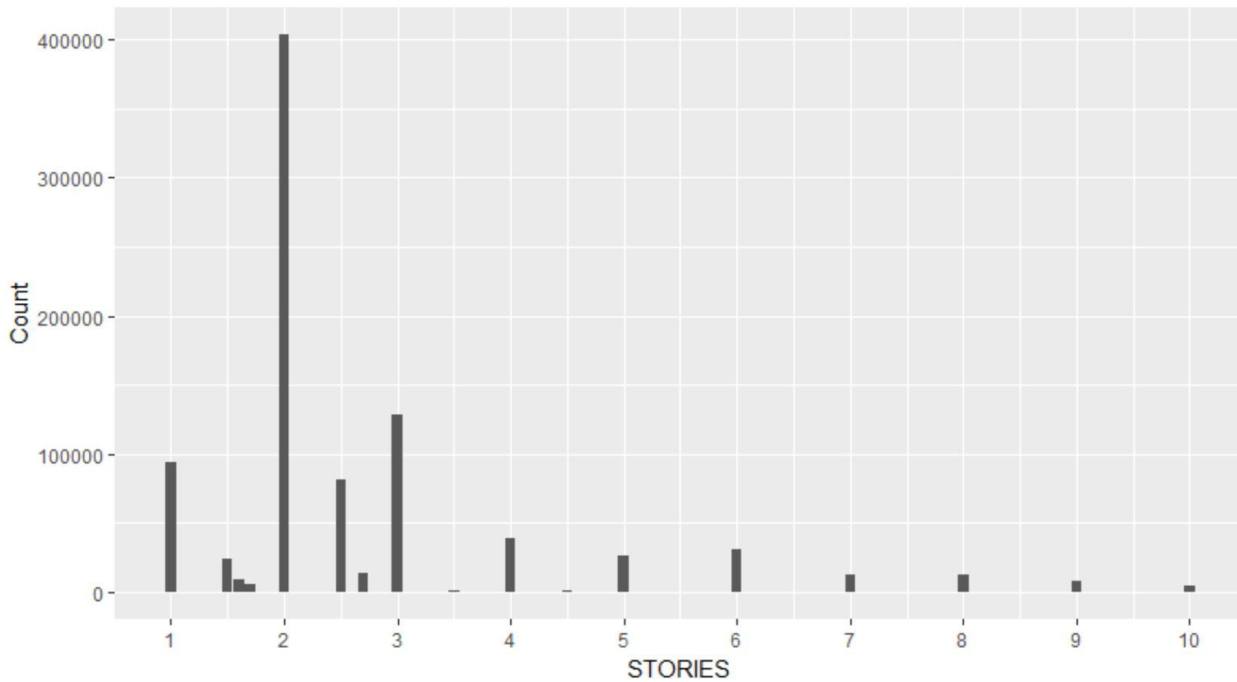
BLDDEPTH

BLDDEPTH (Building Depth) is a numeric variable that represents the width/depth of building in feet. There are 620 unique variables that range from 0 to 9,393 feet (almost 2 miles). There are no missing values, however, 224,699 records have a BLDDEPTH value of 0. We have considered these 0 values as missing in our analysis. The distribution of the field values is shown below:



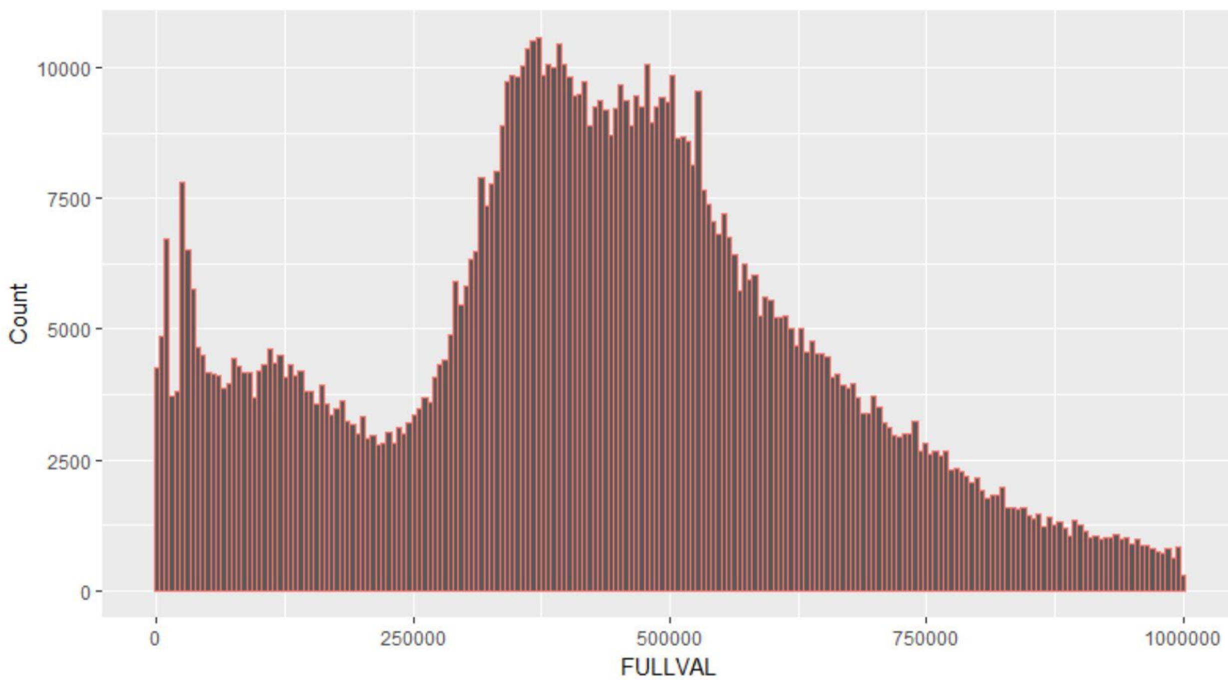
STORIES

STORIES is a numeric variable that represents the number of stories (floors) in the building. There are 112 unique variables that range from 1 to 119. There are 52,142 missing fields for this field. The distribution of the field values is shown below:



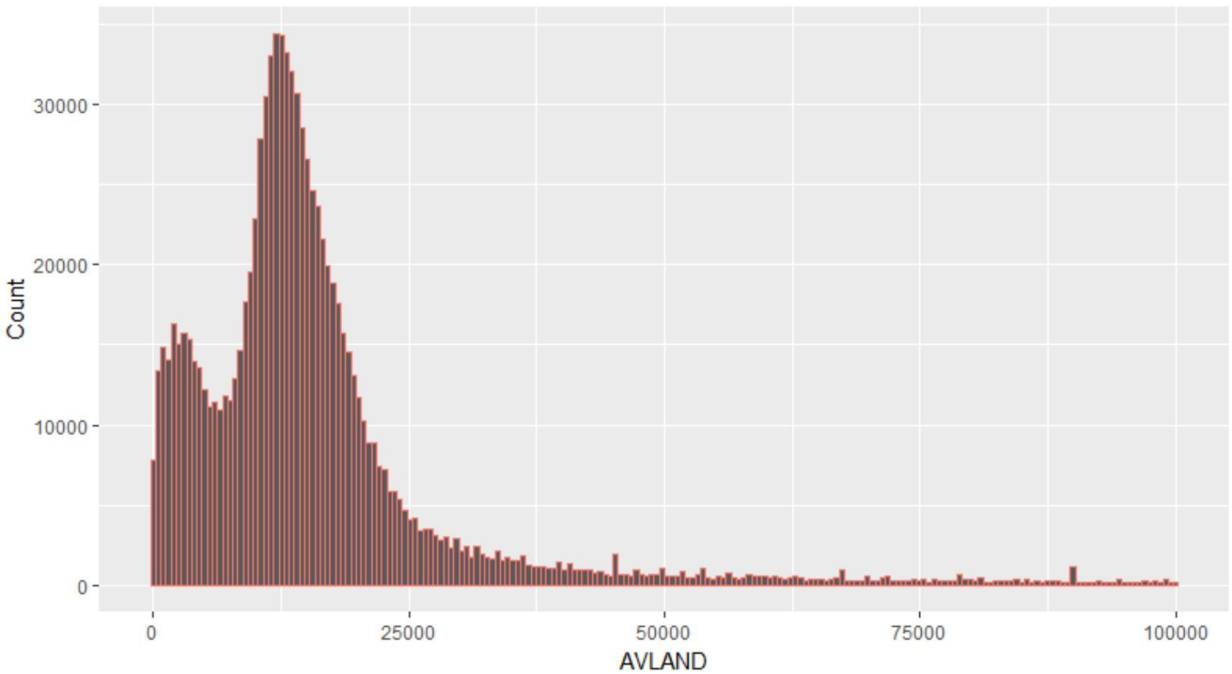
FULLVAL

FULLVAL is a numeric variable that represents the total market value of the property. There are 108,277 unique values for this field ranging from 0 to 6,150,000,000. There are no missing values, however there were 12,762 values that have a FULLVAL value of 0. We have not considered these 0 values as missing in our analysis. The distribution of the field values is shown below:



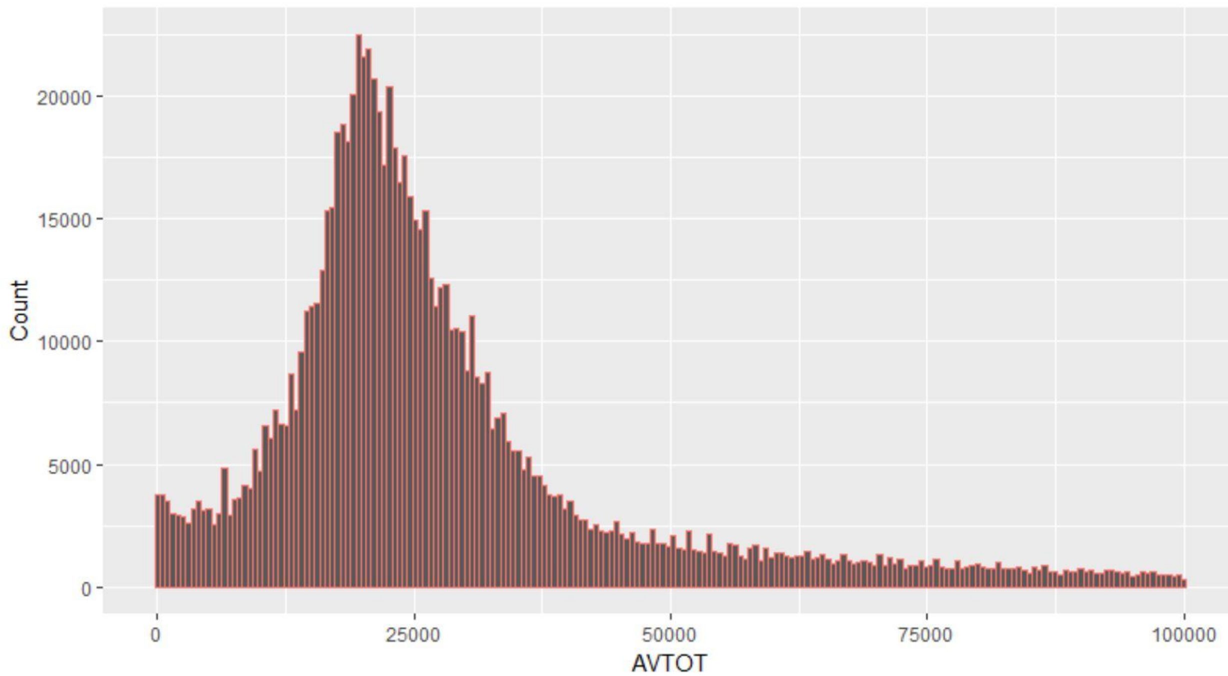
AVLAND

AVLAND is a numeric variable that represents the total assessed value of the land. There are 70529 unique values in this field ranging from 0 to 2,700,000,000. There are no missing values, however there are 12,764 instances where the AVLAND value equals 0. We have not considered these 0 values as missing in our analysis. The distribution of the field values is shown below:



AVTOT

AVTOT is a numeric variable that represents the total assessed value of property (land plus building). There are 112,294 unique values ranging from 0 to 4,700,000,000. There are no missing values, however there are 12,762 records which have an AVTOT value of 0. We have not considered these 0 values as missing in our analysis. The distribution of the field values is shown below:



Data Cleaning

Prior to our analysis, we identified the key fields to be utilized and proceeded to identify rules to populate these fields if they contained missing values. We further analyzed each field to identify outliers, if any. The following sections discuss our methodology for handling outliers and filling in the missing values.

Outliers

We retained all outlier values for all fields. We believe that since our analysis is to look for anomalies in the data, the outliers become extremely critical as they could indicate an anomaly.

Missing Values

The critical fields identified for our analysis were ZIP, STORIES, LTDEPTH, LTFRONT, BLDFRONT and BLDDEPTH. This section describes the methodology we followed to fully populate these fields.

STORIES

To fill in the missing values for this field, we calculated the average value of STORIES under each Building Class (BLDGCL). We chose BLDGCL as the field to be used to populate STORIES as Building Class is representative of the building type (2 family, condo, warehouse etc.) a record belongs to and this leads to a more precise estimation of the number of stories in a particular record.

ZIP

To populate this field, we first selected all unique combinations of BORO, BLOCK and LOT in our dataset. For each of these unique combinations, we discovered that there were maximum of 2 ZIP per combination. We then proceeded to randomly select 1 ZIP value out of these 2 to give us a minimum 50% probability of identifying the ZIP correctly.

Post this process, we were still left with around 24000 missing values for ZIP. We then selected the unique combinations of BORO & BLOCK in our dataset essentially broadening our search for the correct ZIP value for a record. We followed the same steps as with our previous categorization. However, this time, we had a maximum of 5 unique values within each unique combination of BORO and BLOCK giving us a minimum probability of 25% for identifying the correct ZIP.

Post this, we were left with only 1600 missing ZIP values which we populated following the exact same steps as above but this time randomly choosing a ZIP value from the unique ZIP values within a BORO. Through this method, we could fully populate the ZIP field.

LTFRONT & LTDEPTH

We grouped the records by building class, because this metric was 100% populated based on the original raw data given to us. We then calculated an average of LTFRONT and LTDEPTH under each building class and inserted this average value into the missing data fields (fields with LTFRONT & LTDEPTH values of 0). We believe that building class is the best predictor for LTFRONT & LTDEPTH since all records from the same building class should have similar Lot Front and Lot Depth values.

Note: Before we calculated the average value for each building class we removed all zero values to find a true representative measure to input.

BLDFRONT & BLDDEPTH

We grouped the records by BLDGCL (building class), because this metric was 100% populated based on the original raw data that was given to us. We then calculated an average of BLDFRONT and BLDDEPTH under each building class and inserted this average value into the missing data fields (fields with BLDFRONT & BLDDEPTH values of 0). We believe that building class is the best predictor for BLDFRONT & BLDDEPTH since all records from the same building class should have similar Building Front and Building Depth values.

However, for some Building Classes, we were not able to calculate a non-zero average value for either BLDFRONT or BLDDEPTH. For these records, we categorized the records by the TAXCLASS and computed the average for BLDFRONT and BLDDEPTH under each Tax Class. We used these average values to fill in the missing values for records specific to a Tax Class.

Note: Before we calculated the average value for each building class and tax class, we removed all zero values to find a true representative measure to input.

Feature Engineering

Once the data cleaning step was completed, we moved on to feature engineering to develop new variables based on the interaction among the existing variables and help improve our fraud model performance. This is a critical step in our analysis and one in which we depended on expert domain knowledge to isolate key information from our variables.

Expert Variables

Property tax fraud can often be found in the assessment and classification process, allowing an owner to pay less taxes on a property whose value is significantly less than the expected market value of a similar property given the location and type of building and land. Based on the discussions with the business team, we narrowed down our focus to 3 fields i.e. FULLVAL, AVLAND and AVTOT. All these were in dollar terms and since in real estate it makes more sense to talk in dollar values per sq. ft., we decided to scale these 3 fields by area of the lot and building and the volume of the building. For this, we created 3 new variables:

Variable Name	Formula	Description
LotArea	LTFRONT*LTDEPTH	Area of the Lot in sq.ft.
BLDArea	BLDFRONT*BLDDEPTH	Area of the Building in sq.ft.
BLDVol	BLDArea*STORIES	A measure for the Building Volume

Using the above 3 variables, we created 9 more expert variables by dividing each of our 3 main fields of focus i.e. FULLVAL (full market value), AVLAND (assessed value land), and AVTOT (assessed value total) by each of these 3 expert variables i.e. LotArea, BLDArea and BLDVol. The 9 new expert variables are all listed and defined in the table below:

Variable Name	Formula	Description
FULLVAL1	FULLVAL/ LotArea	Full Market Value of a property defined per sq.ft. of Lot Area, Building Area and per unit of Building Volume
FULLVAL2	FULLVAL /BLDArea	
FULLVAL3	FULLVAL / BLDVol	
AVLAND1	AVLAND / LotArea	Assessed Value of Land for a property defined

AVLAND2	AVLAND /BLDArea	per sq.ft. of Lot Area, Building Area and per unit of Building Volume
AVLAND3	AVLAND / BLDVol	
AVTOT1	AVTOT / LotArea	Assessed Total Value of a property defined per sq.ft. of Lot Area, Building Area and per unit of Building Volume
AVTOT2	AVTOT /BLDArea	
AVTOT3	AVTOT / BLDVol	

Other Variables

We then proceeded to use these 9 new variables in Table 2 to create more features using other relevant fields such as ZIP, BORO etc. To accomplish this, we divided each of our Expert Variables B by the average value for that expert variable within a particular category of the relevant fields such as ZIP, BORO etc. This is explained in more detail in the following sub-sections.

Other Variables: ZIP

We divided each of our 9 expert variables by the average of that expert variable within a given zip code using ZIP (5-digit zip-code) field. These 9 new variables provide us a scaled measure for each of our 9 expert variables scaled by zip code.

Other Variables: BORO

We divided each of our 9 expert variables by the average of that expert variable within a given Borough using the BORO field. These 9 new variables provide us a scaled measure for each of our 9 expert variables scaled by BORO.

Other Variables: BLDGCL

We divided each of our 9 expert variables by the average of that expert variable within a given Building Class using the BLDGCL field. These 9 new variables provide us a scaled measure for each of our 9 expert variables scaled by BLDGCL. Due to there being some building classes where the average value of some of the expert variables was 0, we decided to remove these 9 variables from the next steps.

Other Variables: TAXCLASS

We divided each of our 9 expert variables by the average of that expert variable within a given Tax Class using the TAXCLASS field. These 9 new variables provide us a scaled measure for each of our 9 expert variables scaled by TAXCLASS.

Other Variables: BLOCK

We divided each of our 9 expert variables by the average of that expert variable within a given Block using the BLOCK field. These 9 new variables provide us a scaled measure for each of our 9 expert variables scaled by BLOCK. Due to there being some blocks where the average value of some of the expert variables was 0, we decided to remove these 9 variables from the next steps.

Other Variables: ZIP3

For scaling our expert variables over a broader geography, we decided to use a new field ZIP3 characterized by the first 3 digits of the ZIP field. We divided each of our 9 expert variables by the average of that expert variable within a given 3-digit Zip Code using the ZIP3 field. These 9 new variables provide us a scaled measure for each of our 9 expert variables scaled by ZIP3.

Other Variables: STORIESCat

To create variables providing information for records having same structure (in terms of number of stories), we created a new variable to classify all the records into Stories Category: 0-2 stories, 2-4 stories, 4-7 stories, and 7+ stories. These categories were based on the distribution of the STORIES field. We divided each of our 9 expert variables by the average of that expert variable within a given category of Stories using the STORIESCat field. These 9 new variables provide us a scaled measure for each of our 9 expert variables scaled by the stories category of the property.

The complete list of Other Variables is provided below:

Variable Name	**Formula**	**Description**
FULLVAL4	FULLVAL1 / Avg. (FULLVAL1 | ZIP)	Expert Variables scaled

Variable Name	Formula	Description
FULLVAL5	FULLVAL2 / Avg. (FULLVAL2 | ZIP)	by the average of the specific expert variables across a given ZIP code
FULLVAL6	FULLVAL3 / Avg. (FULLVAL3 | ZIP)	
AVLAND4	AVLAND1 / Avg. (AVLAND1 | ZIP)	
AVLAND5	AVLAND2 / Avg. (AVLAND2 | ZIP)	
AVLAND6	AVLAND3 / Avg. (AVLAND3 | ZIP)	
AVTOT4	AVTOT1 / Avg. (AVTOT1 | ZIP)	
AVTOT5	AVTOT2 / Avg. (AVTOT2 | ZIP)	
AVTOT6	AVTOT3 / Avg. (AVTOT3 | ZIP)	
FULLVAL7	FULLVAL1 / Avg. (FULLVAL1 | BORO)	Expert Variables scaled by the average of the specific expert variables across a given Borough
FULLVAL8	FULLVAL2 / Avg. (FULLVAL2 | BORO)	
FULLVAL9	FULLVAL3 / Avg. (FULLVAL3 | BORO)	
AVLAND7	AVLAND1 / Avg. (AVLAND1 | BORO)	
AVLAND8	AVLAND2 / Avg. (AVLAND2 | BORO)	
AVLAND9	AVLAND3 / Avg. (AVLAND3 | BORO)	
AVTOT7	AVTOT1 / Avg. (AVTOT1 | BORO)	
AVTOT8	AVTOT2 / Avg. (AVTOT2 | BORO)	
AVTOT9	AVTOT3 / Avg. (AVTOT3 | BORO)	
FULLVAL10	FULLVAL1 / Avg. (FULLVAL1 | BLDGCL)	Expert Variables scaled by the average of the
FULLVAL11	FULLVAL2 / Avg. (FULLVAL2 | BLDGCL)	

Variable Name	Formula	Description
FULLVAL12	FULLVAL3 / Avg. (FULLVAL3 | BLDGCL)	specific expert variables across a given Building Class
AVLAND10	AVLAND1 / Avg. (AVLAND1 | BLDGCL)	
AVLAND11	AVLAND2 / Avg. (AVLAND2 | BLDGCL)	
AVLAND12	AVLAND3 / Avg. (AVLAND3 | BLDGCL)	
AVTOT10	AVTOT1 / Avg. (AVTOT1 | BLDGCL)	
AVTOT11	AVTOT2 / Avg. (AVTOT2 | BLDGCL)	
AVTOT12	AVTOT3 / Avg. (AVTOT3 | BLDGCL)	
FULLVAL13	FULLVAL1 / Avg. (FULLVAL1 | TAXCLASS)	Expert Variables scaled by the average of the specific expert variables across a given Tax Class
FULLVAL14	FULLVAL2 / Avg. (FULLVAL2 | TAXCLASS)	
FULLVAL15	FULLVAL3 / Avg. (FULLVAL3 | TAXCLASS)	
AVLAND13	AVLAND1 / Avg. (AVLAND1 | TAXCLASS)	
AVLAND14	AVLAND2 / Avg. (AVLAND2 | TAXCLASS)	
AVLAND15	AVLAND3 / Avg. (AVLAND3 | TAXCLASS)	
AVTOT13	AVTOT1 / Avg. (AVTOT1 | TAXCLASS)	
AVTOT14	AVTOT2 / Avg. (AVTOT2 | TAXCLASS)	
AVTOT15	AVTOT3 / Avg. (AVTOT3 | TAXCLASS)	
FULLVAL16	FULLVAL1 / Avg. (FULLVAL1 | BLOCK)	Expert Variables scaled by the average of the specific expert variables across a given Block
FULLVAL17	FULLVAL2 / Avg. (FULLVAL2 | BLOCK)	
FULLVAL18	FULLVAL3 / Avg. (FULLVAL3 | BLOCK)	

<table>
<tr><th>Variable Name</th><th>Formula</th><th>Description</th></tr>
<tr><td>AVLAND16</td><td>AVLAND1 / Avg. (AVLAND1 | BLOCK)</td><td rowspan="6"></td></tr>
<tr><td>AVLAND17</td><td>AVLAND2 / Avg. (AVLAND2 | BLOCK)</td></tr>
<tr><td>AVLAND18</td><td>AVLAND3 / Avg. (AVLAND3 | BLOCK)</td></tr>
<tr><td>AVTOT16</td><td>AVTOT1 / Avg. (AVTOT1 | BLOCK)</td></tr>
<tr><td>AVTOT17</td><td>AVTOT2 / Avg. (AVTOT2 | BLOCK)</td></tr>
<tr><td>AVTOT18</td><td>AVTOT3 / Avg. (AVTOT3 | BLOCK)</td></tr>
<tr><td>FULLVAL19</td><td>FULLVAL1 / Avg. (FULLVAL1 | ZIP3)</td><td rowspan="9">Expert Variables scaled by the average of the specific expert variables across a given ZIP3 code</td></tr>
<tr><td>FULLVAL20</td><td>FULLVAL2 / Avg. (FULLVAL2 | ZIP3)</td></tr>
<tr><td>FULLVAL21</td><td>FULLVAL3 / Avg. (FULLVAL3 | ZIP3)</td></tr>
<tr><td>AVLAND19</td><td>AVLAND1 / Avg. (AVLAND1 | ZIP3)</td></tr>
<tr><td>AVLAND20</td><td>AVLAND2 / Avg. (AVLAND2 | ZIP3)</td></tr>
<tr><td>AVLAND21</td><td>AVLAND3 / Avg. (AVLAND3 | ZIP3)</td></tr>
<tr><td>AVTOT19</td><td>AVTOT1 / Avg. (AVTOT1 | ZIP3)</td></tr>
<tr><td>AVTOT20</td><td>AVTOT2 / Avg. (AVTOT2 | ZIP3)</td></tr>
<tr><td>AVTOT21</td><td>AVTOT3 / Avg. (AVTOT3 | ZIP3)</td></tr>
<tr><td>FULLVAL22</td><td>FULLVAL1 / Avg. (FULLVAL1 | STORIESCat)</td><td rowspan="4">Expert Variables scaled by the average of the specific expert variables across a given Stories Category</td></tr>
<tr><td>FULLVAL23</td><td>FULLVAL2 / Avg. (FULLVAL2 | STORIESCat)</td></tr>
<tr><td>FULLVAL24</td><td>FULLVAL3 / Avg. (FULLVAL3 | STORIESCat)</td></tr>
<tr><td>AVLAND22</td><td>AVLAND1 / Avg. (AVLAND1 | STORIESCat)</td></tr>
</table>

Variable Name	Formula	Description
AVLAND23	AVLAND2 / Avg. (AVLAND2 | STORIESCat)	
AVLAND24	AVLAND3 / Avg. (AVLAND3 | STORIESCat)	
AVTOT22	AVTOT1 / Avg. (AVTOT1 | STORIESCat)	
AVTOT23	AVTOT2 / Avg. (AVTOT2 | STORIESCat)	
AVTOT24	AVTOT3 / Avg. (AVTOT3 | STORIESCat)	

Algorithm Design & Implementation

Our aim is to build a fraud detecting machine learning model which could provide us a score for each of our records. Based on this score, we shall arrive at a measure to detect anomalies in our dataset and flag such records. To accomplish this, we followed the common approach of z-scaling, PCA and z-scaling again to prepare our dataset for input to our fraud detecting algorithm. As an input for this step, we retained 54 variables from our expanded dataset from the preceding step. These included the 9 expert variables and 45 other variables (excluding those for BLOCK and BLDGCL as explained earlier). On the algorithm front, we utilized 2 different approaches to calculate scores for each record. Finally, we combined these 2 scores (post scaling them through Quantile-Binning) to arrive at a final score for each record. Based on this final score, we classified a record as an anomaly or not. This process is explained in detail in the following sections.

Common Approach

Z-Scaling

Since our variables were on different scales, we needed to bring them all to a common scale prior to further analysis. For this, we z-scaled them which is subtracting the mean of the variable column from the variable while dividing by the standard deviation of that column. This way, all variables were on the same scale, or scale-less.

$$z = \frac{x - \mu}{\sigma}$$

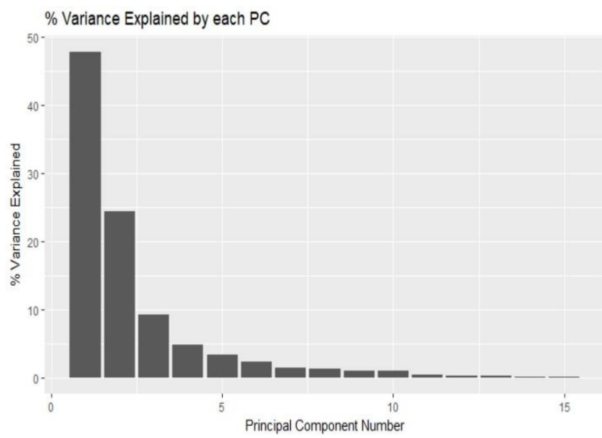
Principal Component Analysis

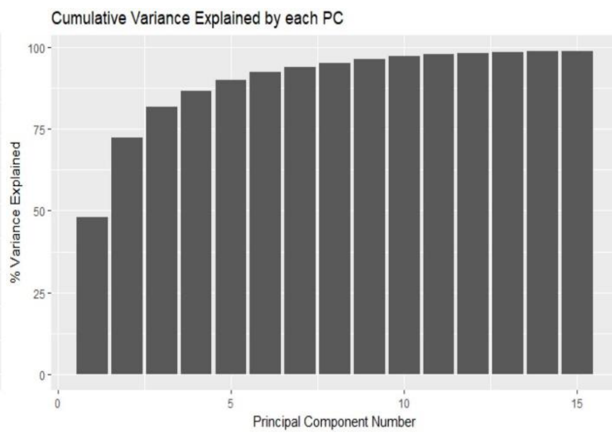
The Principal Component Analysis (PCA) is a linear technique used to emphasize variation and identify strong directional patterns in the data under review. This analysis uses an orthogonal transformation to convert the correlated expert variables into a set of linearly uncorrelated variables creating the principal components of any data set.

- PCA is useful for limiting dimensionality and reducing correlation between variables

- However, due to PCA essentially being a linear technique, it may perform poorly in identifying the principal components if non-linear interactions exist between the variables

Based on the PCA, we finalized 10 PCs for further analysis. These PCs were selected based on their ability to explain the maximum variance in our dataset. Graphs depicting the amount of variance explained per principal component and the cumulative variance explained are provided below:





As may be seen from the above graphs, the top 10 PCs explain ~97% of the variance in our data.

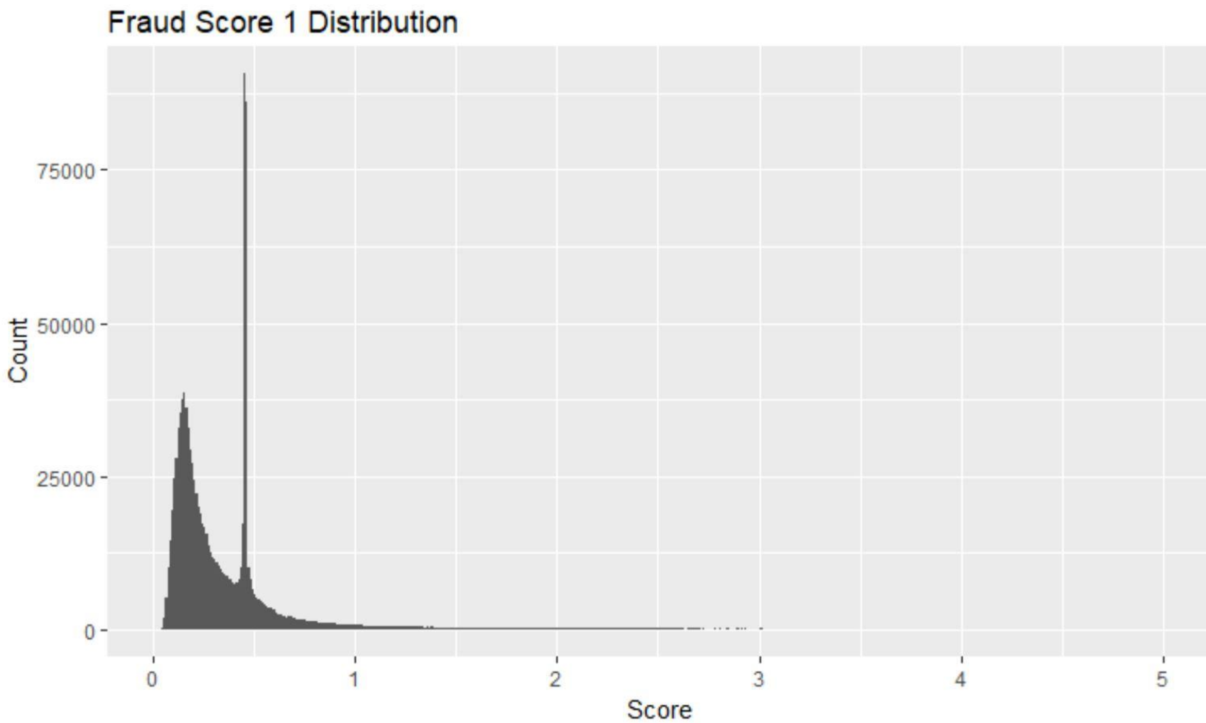
Z-Scaling Again

Once the PCs were selected and the corresponding PC values for each record extracted, we again z-scaled these 10 PCs to bring them to the same scale. This is required since during the Principal Component Analysis, the PCs were calculated through linear combinations of various variables which changed their scale and they need to be brought to same scale prior to proceeding with the next steps. This step converts each column into a distribution with an expected value of 0 and standard deviation of 1.

Algorithm 1: Mahalanobis Distance

Post the above steps, we were in possession of a dataset with the 10 PCs and the scaled values of those PCs for each of our records. To calculate our first fraud score, we decided to calculate the Mahalanobis distance for each record. This algorithm calculates the Euclidean distance of each record from the origin which is simply the root of the sum of the squared values of the 10 PCs for each record. Since the expected value for each PC for each record is 0 (as we z-scaled the PCs), we expect the Mahalanobis distance for each record to also have an expected value of 0.

After this calculation, we ranked the records by distance from the center. Essentially, ranking records by greatest to smallest distance away from the center (the fraud score) provided us the records which were located furthest away from the origin thus indicating an anomaly. A distribution of the fraud score 1 is provided below. As may be seen, it is a right-skewed distribution which is expected as we expect most of the records to lie close to the origin.



Algorithm 2: Autoencoder

For our second algorithm, we decided to build an Autoencoder whose function is two-fold:

- It tries to explain the input dataset through reduced number of variables. This is akin to the Principal Component Analysis. However, one advantage the Autoencoder holds over the PCA is that it also tries to learn any non-linear interactions between the variables. It accomplishes this through creation of hidden layers (user defined)
- It tries to reproduce the original dataset by using these reduced number of variables (we could also have built an Autoencoder which instead of reducing the dimensionality would have exploded it)

After selecting the PCs and z scaling them, the dataset obtained earlier was used to train our Autoencoder with 3 hidden layers where we directed the Autoencoder to learn patterns in our data by reducing the number of variables first from 10 to 7, then from 7 to 5 and then again expand these 5 to 7 variables taking it finally back to 10.

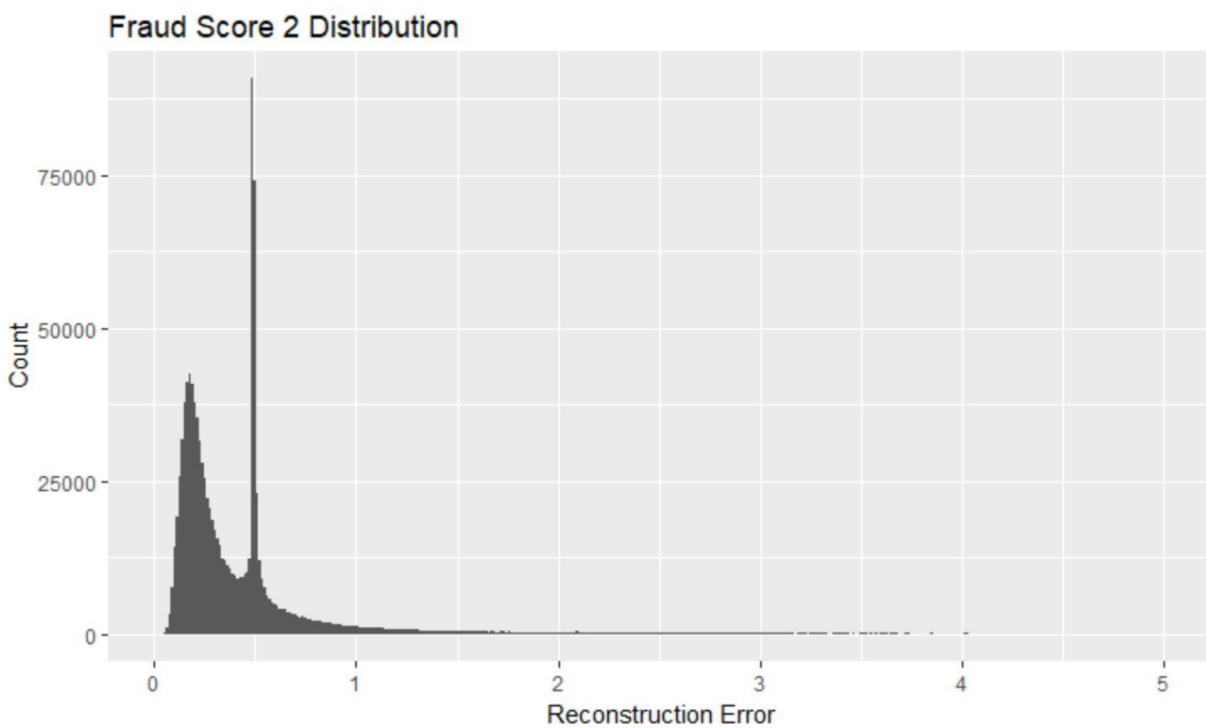
After training our Autoencoder, we used it to predict the values of the dataset. We were looking for records which the Autoencoder found difficult to reproduce correctly. To obtain these, for each record, we calculated the Reconstruction Error using the following formula (like the Mahalanobis distance calculated earlier):

$$S = \left(\Sigma_i \left| z_i - z_i' \right|^n\right)^{1/n}$$

Here, z_i is the value of the i^{th} PC in the original dataset and z'_i is the value of the i^{th} PC in the reconstructed dataset. We chose n=2 to calculate the Euclidean distances.

The above calculation provided us our Fraud Score 2 for each record of the dataset. The higher the score, the higher the difficulty our Autoencoder faced in reconstructing the record indicating an anomaly.

We ranked the records according to the score. The score 2 distribution is provided below. As may be seen, it is right-skewed as expected with majority of the records having a reconstruction error close to 0.



Combining Algorithms

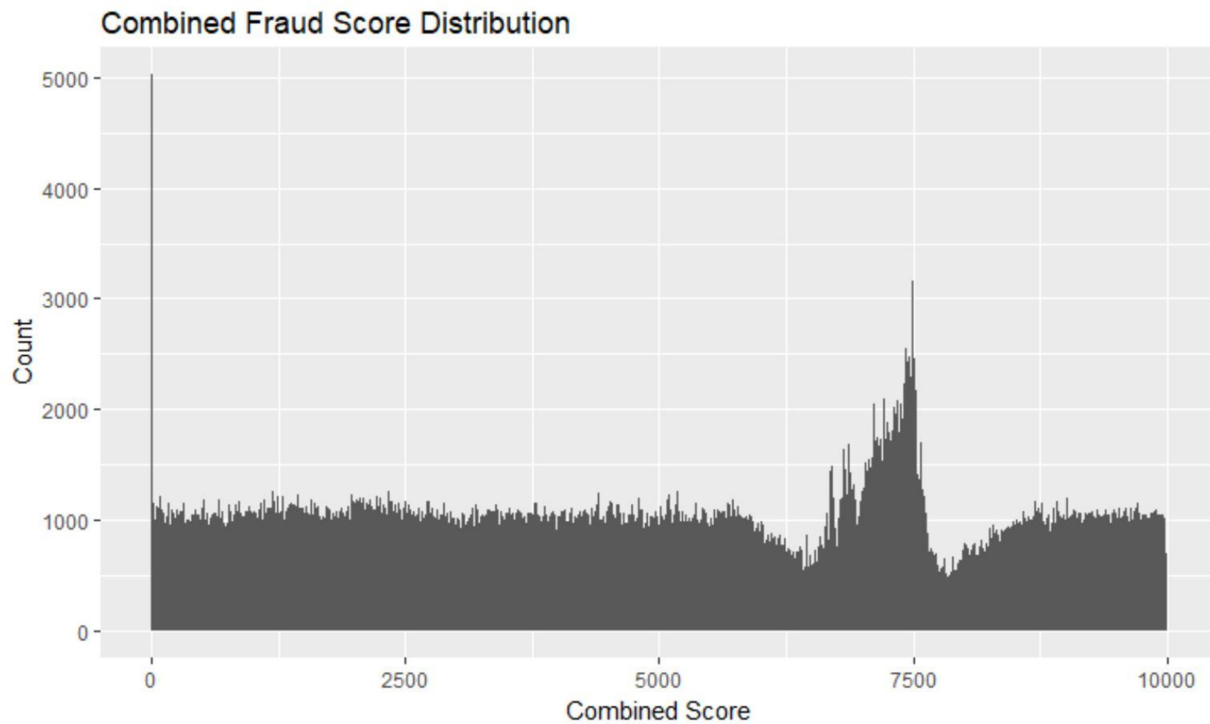
Our intent from the beginning was to detect records in the dataset which were anomalous. Once we had obtained the 2 scores for each record through our algorithms, the next step was to combine these 2 scores to arrive at a final score for ranking the records to have a common measure by which to detect anomalies. To achieve this, we first had to bring both the scores to a common scale to combine them. For this, we used the concept of binning with total number of bins equal to 10,000.

Binning

We ranked the records according to their individual fraud scores, first by score 1 and then by score 2 from highest to lowest. In both the cases, we began from the top record (one with the highest score) and assigned the first 104 records (top 0.01%) the highest bin number i.e. 10,000. The next 104 records were given the bin number 9,999. We followed the same step and kept on assigning the bin numbers till we reached the end of the dataset. Finally, we had 2 bin numbers for each record. These can be same or different depending on the record's ranking under each of the fraud score. We then replaced the fraud scores for each record under both the algorithms by their corresponding bin numbers.

Final Fraud Score

By following the binning procedure, we brought both the fraud scores on a common scale. Top 0.01% records (ranked in descending order of scores) under both the algorithms had the same bin number. To combine the 2 bin numbers to arrive at a combined score, we assigned equal weights to the 2 scores (as we believe both the algorithms are equivalent in nature) and computed the weighted average of the 2 bin numbers for each record. The records were then ranked according to this combined score and the top 50 records were analyzed to find out unusual records. Our findings are detailed in the next section. Provided below is a distribution of the combined scores for each record.



Combined Fraud Score Distribution
Count
5000
4000
3000
2000
1000
0
0
2500
5000
7500
10000
Combined Score

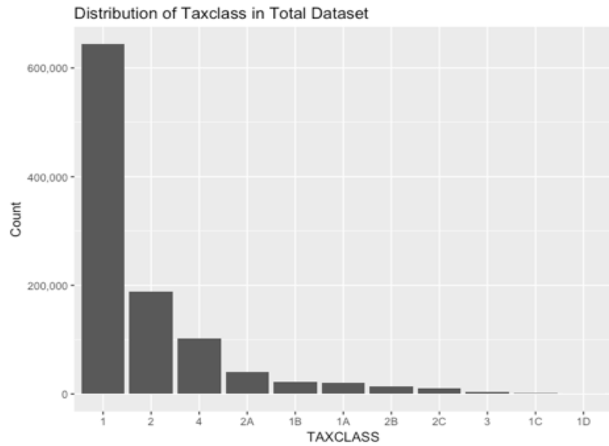
Results & Insights

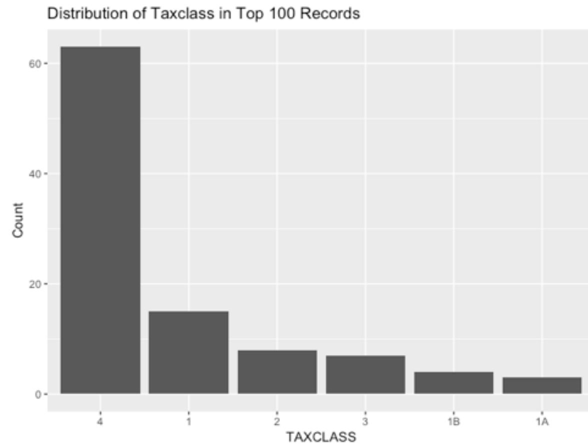
We sorted the records with respect to the combined score and then fraud score 1 so that we can order the records in the same bin. The chart below is an overview of comparison between the total dataset and the 100 records with highest combined score in the first bin.

	Entire Dataset			Top 100 Records		
	Mean	St.Dev	Median	Mean	St.Dev	Median
FULLVAL	880487.7	11702927	446000	236691452	791180854	33850000
AVLAND	85995.03	4100755	13646	106164041	387455570	10012500
AVTOT	230758.2	6951206	25339	149210667	578779917	13756275
LTFRONT	36.17	73.73	25	797.2	1627.51	104.82
LTDEPTH	88.28	75.48	100	515.79	891.32	130.47
STORIES	5.06	8.43	2	7.5	13.01	2
BLDFRONT	23.02	35.79	20	31.84	38.05	20
BLDDEPTH	40.07	43.04	39	49.55	71.28	20
LOTAREA	8922.3	154918.3	3000	1231394	4276796	12468
BLDAREA	4319.81	99744.46	1080	3595	9827.74	463
BLDVOL	61293.36	2319537	2268	90734.22	389808	520

- FULLVAL, AVLAND and AVTOT, the three main variables that we focus on and LTFRONT, LTDEPTH and LotArea all have significantly higher mean, median and standard deviation values in the top 100 potential fraud records compared to the total dataset. Based on this fact, the buildings with high fraud score are big and have high market value as well as assessed value.
- Moreover, mean, standard deviation and median are also visibly higher for variables like STORIES, BLDFRONT, BLDDEPTH and BLD_VOLUME for top 100 records. This finding also tells us that these records represent the larger buildings.
- When we looked at the TAXCLASS of the top 100 high score records, we found that 63% of the properties belong to TAXCLASS 4 while in the whole dataset it’s less than 10%. Referring to definition, TAXCLASS 4 means “all commercial and industrial properties” and “all other”. This could indicate that fraud is being conducted by business

owners of commercial establishments to reduce the property tax burden on their businesses.





- When we looked at the owners of the high score properties, we find that most of the owners are large house agencies, real estate companies and government entities. Very few of these properties belong to single households.

Top 10 Records

To extract the top 10 scoring records, we first sorted the records by the Combined Score and then for records having the same combined score, we sorted them according to the record's score 1 arrived at through the first algorithm. We excluded government-owned properties and examined the remaining records one by one. We selected the most suspicious 10 records and the field values that were filled during the data cleaning process are marked in red in the below table.

RECORD	FULLVAL	AVLAND	AVTOT	LTFRONT	LTDEPTH
5393	$2,930,000	$1,318,500	$1,318,500	157	95
24586	$3,712,000	$252,000	$1,670,400	94	165
977471	$3,443,400	$1,549,530	$1,549,530	298	402
750447	$251,989	$1,001	$8,934	1	1
787892	$2,151,600	$968,220	$968,220	139	342
83457	$100	$45	$45	1	1
597387	$138,000,000	$11,025,000	$62,100,000	39	50
824497	$266,000	$10,395	$119,700	9	242
638884	$625,000	$56,250	$281,250	43	50
432852	$114,000,000	$33,750,000	$51,300,000	25	100

RECORD	BLDFRONT	BLDDEPTH	OWNER	TAXCLASS	BLDGCL
5393	1	1	864163 REALTY, LLC	2	D9
24586	1	1	11-01 43RD AVENUE REA	4	H9
977471	1	1	NEW YORK CITY	4	O3
750447	25	46	OH, LAURA E	1A	R3
787892	1	1	NA	4	O3
83457	10	18	NA	3	U1
597387	39	50	BERKOWTIZ, ULWT LOUIS	4	W6
824497	38	80	EMC MORTGAGE CORP.	3	U7
638884	1	1	JAMES T MORIATES	2	D6
432852	25	100	79TH REALTY LLC	2	D8

RECORD	STORIES	ZIP	LotArea	BLDArea	BLDVol
5393	1	11373	14915	1	1
24586	10	11101	15510	1	10
977471	20	11101	119796	1	20
750447	1	11364	1	1163	1163
787892	20	11101	47538	1	20
83457	1.33	11215	1	175	233
597387	2	10010	1950	1950	3900
824497	1.33	11691	2178	3040	4043
638884	9	11432	2150	1	9
432852	44	10075	2500	2500	110000

Record Number	Finding
5393	This property has an incredibly high market value compared to other similar buildings. The building frontage and building depth are both 1, which is very suspicious. Further, google map indicates that the property has 7 floors rather than 1 floor as reported.
24586	This property is a 10-story hotel. Its building depth and building frontage, however, are both 1 feet, which is unreasonable. Further, it does not have a valid owner name. Its exemption value is greater than zero, which should not be because its exemption type is missing
977471	This property is an extremely large office building. It has a lot front of 298 feet, a lot depth of 402 and 20 stories. The market value of this building, however, is way too low in contrast with its volume. The

	average value for buildings with 20 stories is $2.95 mil with an average LTFRONT and LTDEPTH equal to 144 and 155 feet respectively. This property is much bigger but the market value ($3.44 mil) does not correspond to its size.
750447	The assessed land value is only $1,001. Also, the last name of the owner is "Oh". Both indicate potential fraud. The record shows that the building has 1 floor but google map shows that the building has 3 stories
787892	This property record doesn't have an owner name or precise address. Its market value $2.15 mil is very low relative to its size. The property has 20 stories and average market value for properties this size is $2.95 mil
83457	According to google map, this property seems to be a warehouse along Hamilton Avenue. It has a full value of only $100 and a land value of only $45. Additionally, there is no data for number of stories and ownership identification. All these indications make the property extremely suspicious.
597387	This property is a 2-story small building located on the cross of Lexington Avenue and East 49th Street. However, the full value of this property is more than $138 million, which is very unusual. Also, this record does not have an exemption class, but has been given a tax exemption.
824497	This property has an abnormally high lot depth and no floor information. Additionally, Building Front of the property is greater than the Lot Front for the property (BLDFRONT = 38 feet while LTFRONT = 9 feet)
638884	This record has building front and depth of only 1 foot. There are no buildings on 178th street with 9 stories. Also, the building front and depth measure only 1 foot. Google Map shows that the building at this address has maximum of 2 stories
432852	The full value of this property is $114 mil, which is extremely high

	compared to its building volume. Average market value for properties with 44 stories is $1.87 mil

Conclusion

Beginning with a raw dataset of the NYC property assessment records, we carried out the data visualization and cleaning steps before our main analysis using machine learning methods. We used both a heuristic algorithm (calculating the Mahalanobis distance) and the autoencoder to identify 10 records (listed in the previous section) as the most likely candidates for fraud in the New York Property Data Set. We picked these 10 as they had the highest combined fraud scores from the algorithms **after removing all government owned properties**.

While it is not certain that these properties are fraudulent, the high combined fraud score indicates that of all these properties/records give the highest indication of potential fraud and we would benefit from further investigation of these records. Further, majority of the top scoring records are commercial establishments indicating that incidence of fraud may be higher among business owners of such commercial properties rather than working individuals.

We found it interesting that several government properties were flagged as anomalies with high fraud scores, but chose to exclude them as we believe that Govt. properties are sometimes exempt from property taxes and it would be more fruitful to focus on other than Govt. properties.

We believe that the algorithms combined with our carefully calculated expert variables were very effective in isolating the most likely candidates for fraud. Given more time, we would like to spend it in creating more expert as well as other variables to make our algorithms more precise.

APPENDIX-I

Data Quality Report

Dataset File Name: NY property 1 million.xlsx
Source: Department of Finance, New York City
Number of Records: 1,048,575
Number of Fields: 30
Year of recording Data: 2011
Year of Analysis: 2018

Summary Table

Categorical Variables

Variable	Description	#Distinct Value	Percent populated
RECORD	Record Number	1048575	100%
BBLE	Concatenation of BORO, BLOCK, LOT and EASEMENT	1048575	100%
BORO	BORO Codes	5	100%
BLOCK	Block Codes	13949	100%
LOT	Unique Number within BORO/BLOCK	6366	100%
EASEMENT	Easement	12	0.39%
OWNER	Owner	847055	97.04%
BLDGCL	Building Class	200	100%
TAXCLASS	Current Property Tax Class Code	11	100%
STADDR	Street Address	820638	99.94%
ZIP	Zip Code	197	97.49%
EXMPTCL	Exemption Class	14	1.43%
PERIOD	Period	1	100%
YEAR	Year	1	100%
VALTYPE	Value Type	1	100%

Numerical Variables

Variable	Description	Mean	Standard Deviation	Min	Max	Percent Populated
LTFRONT	Lot Frontage	36.17	73.73	0	9999	100%
LTDEPTH	Lot Depth	75.48	88.28	0	9999	100%
STORIES	Stories	5.06	8.43	1	119	95.02%
FULLVAL	Total Market Value	8.805e+05	1.170e+07	0	6.150e+09	100%
AVLAND	Assessed Value of Land	8.600e+04	4.101e+06	0	2.668e+09	100%
AVTOT	Assessed Total Value	2.308e+05	6.951e+06	0	4.668e+09	100%
EXLAND	Exemption Value of Land	3.681e+04	4.024e+06	0	2.668e+09	100%
EXTOT	Total Exemption Value	9.254e+04	6.578e+06	0	4.668e+09	100%
EXCD1	EXCD1	1604	1388.132	1010	7170	59.38%
BLDFRONT	Building Frontage	23.02	35.78	0	7575	100%
BLDDEPTH	Building Depth	40.07	43.04	0	9393	100%
AVLAND2	Assessed Value of Land	2.464e+05	6.199e+06	3	2.371e+09	26.80%
AVTOT2	Assessed Total Value	7.161e+05	1.169e+07	3	4.501e+09	26.80%
EXLAND2	Exemption Value of Land	3.518e+05	1.085e+07	1	2.371e+09	20.20%
EXTOT2	Exemption Total Value 2	6.581e+05	1.613e+07	7	4.501e+09	16.44%
EXCD2	EXCD2	1372	1105.49	1011	7160	19.84%

Field-wise Analysis

Field Name	RECORD
Field Type	Numeric
% Field Populated	100%
% of “NA” Values	0%
No. of Unique Values	1048575
Minimum Value	1
Maximum Value	1048575
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Unique Identifier for each record

Field Name	BBLE
Field Type	Character
% Field Populated	100%
% of “NA” Values	0%
No. of Unique Values	1048575
Minimum Value	-
Maximum Value	-
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Concatenation of BORO, BLOCK, LOT, EASEMENT Unique Identifier for each record

Field Name	BLOCK
Field Type	Numeric
% Field Populated	100%
% of "NA" Values	0%
No. of Unique Values	13949
Minimum Value	-
Maximum Value	-
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Block to which the property belongs

Top 25 Blocks

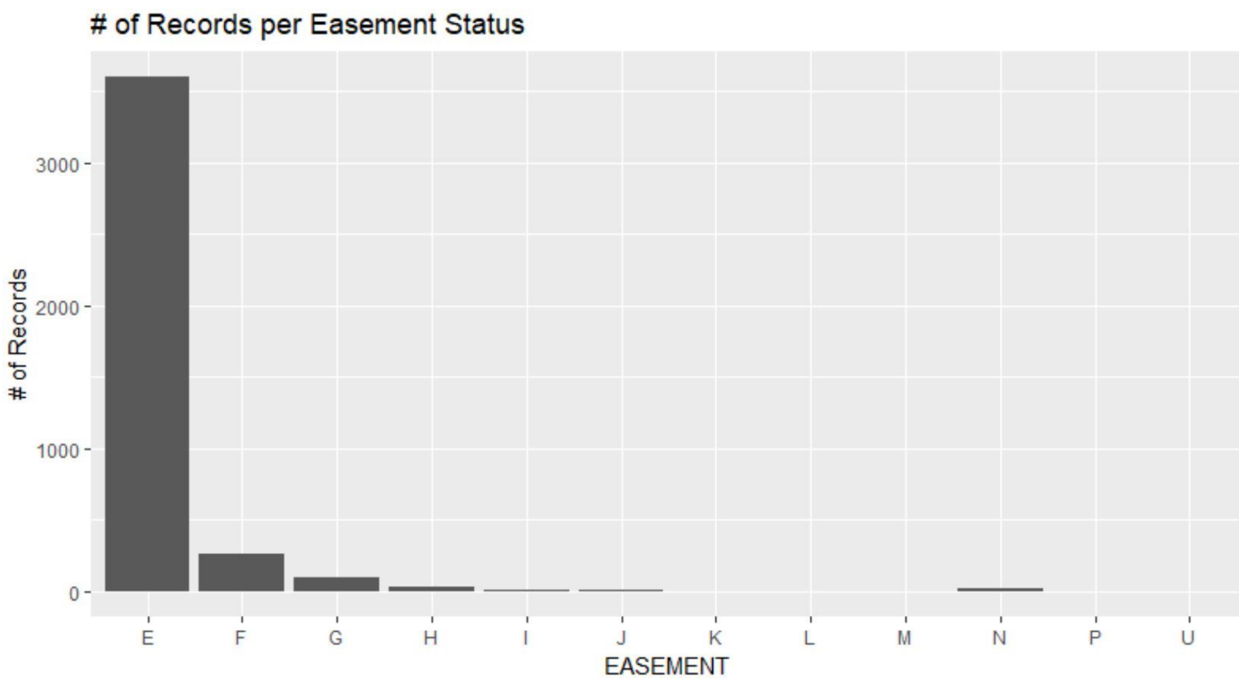
Block	# of Records	% of Field
3944	3888	0.37
16	3786	0.36
3943	3424	0.33
3938	2794	0.27
1171	2535	0.24
3937	2275	0.22
1833	1774	0.17
2450	1651	0.16
1047	1480	0.14
7279	1302	0.12
5893	1295	0.12
8720	1281	0.12
936	1151	0.11
1115	1090	0.10
1320	1049	0.10
1140	1017	0.10
1011	991	0.09
943	946	0.09
1116	881	0.08
1515	869	0.08
3432	853	0.08
1537	842	0.08
1040	821	0.08
870	809	0.08
1536	796	0.08

Field Name	LOT
Field Type	Numeric
% Field Populated	100%
% of “NA” Values	0%
No. of Unique Values	6366
Minimum Value	1
Maximum Value	9978
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Unique LOT number within each BORO/BLOCK

Top 25 Lot numbers

Lot #	**# of Records**	**% of Field**
1	23570	2.25
20	12045	1.15
15	11904	1.14
12	11894	1.13
14	11864	1.13
16	11810	1.13
18	11763	1.12
17	11728	1.12
25	11692	1.12
21	11593	1.11
23	11469	1.09
22	11462	1.09
6	11418	1.09
19	11408	1.09
24	11392	1.09
26	11390	1.09
30	11354	1.08
28	11170	1.07
29	11149	1.06
27	11107	1.06
13	11086	1.06
7	11070	1.06
10	10876	1.04
9	10872	1.04
11	10773	1.03

Field Name	EASEMENT
Field Type	Character
% Field Populated	0.38%
% of "NA" Values	99.61%
No. of Unique Values	13
Minimum Value	-
Maximum Value	-
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Field used to describe Easement Status for the property



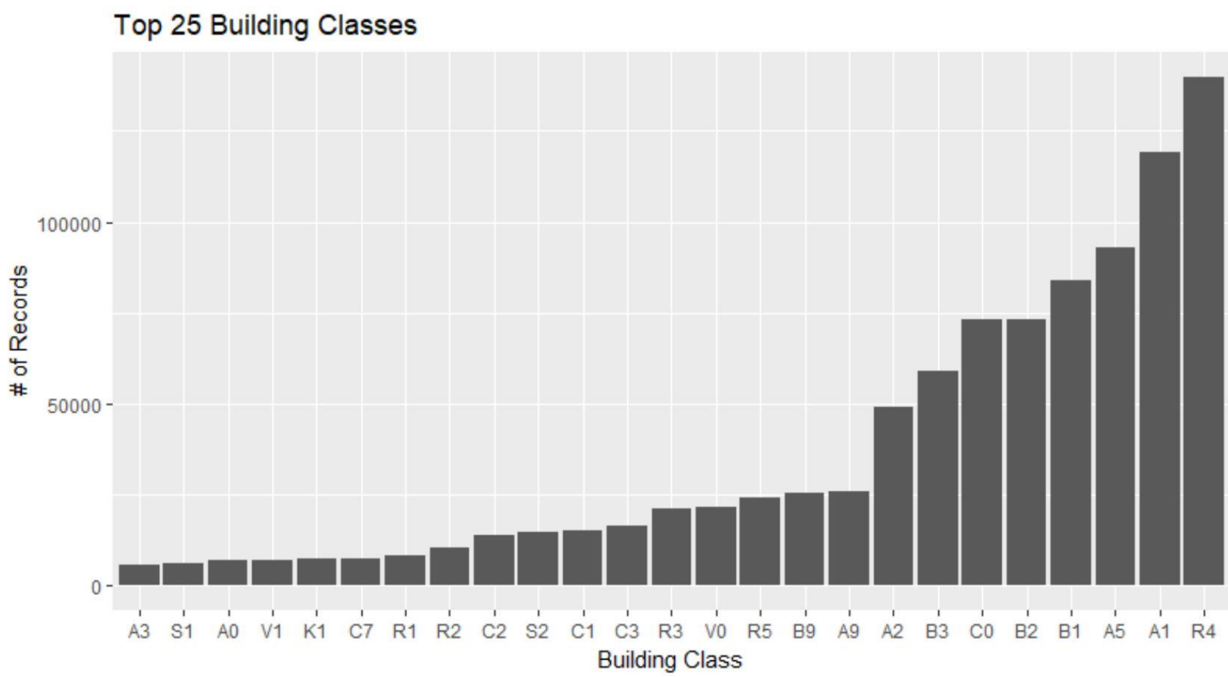
**Removed NAs*

Field Name	OWNER
Field Type	Character
% Field Populated	97.03%
% of “NA” Values	2.96%
No. of Unique Values	847055
Minimum Value	-
Maximum Value	-
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Name of the Owner under whom the property is listed

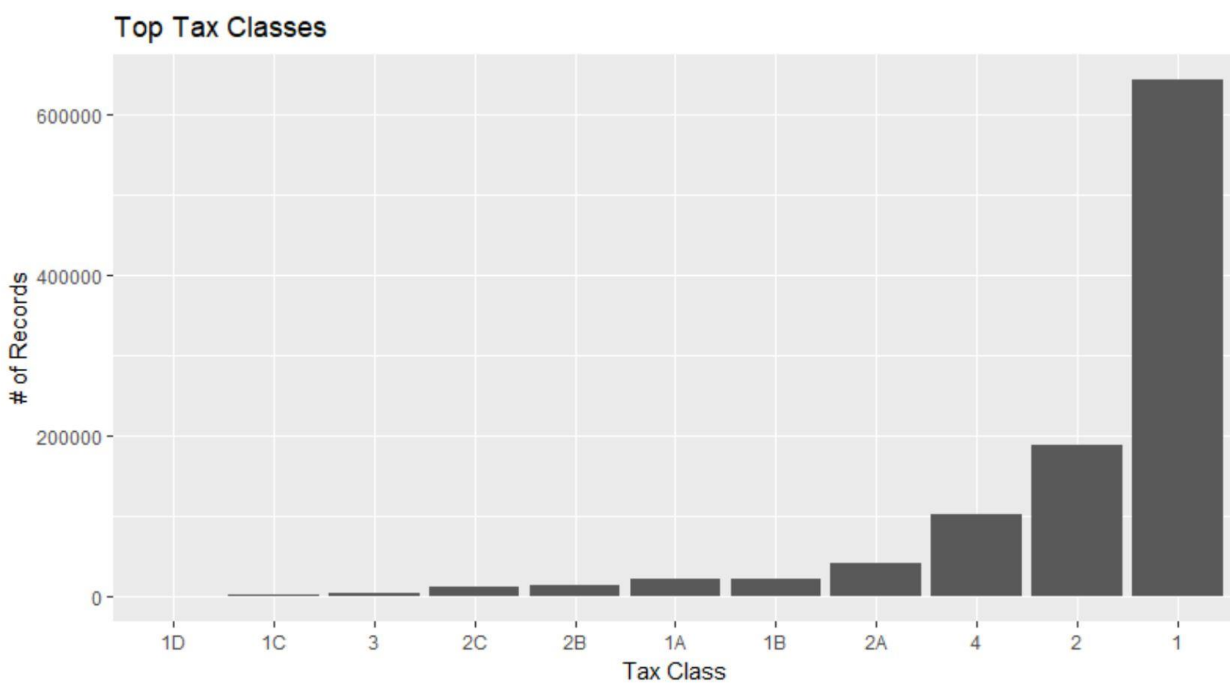
Top 25 Owners

Owner	**# of Properties**	**% of field**
PARKCHESTER PRESERVAT	6021	0.57
PARKS AND RECREATION	3358	0.32
DCAS	2053	0.20
HOUSING PRESERVATION	1900	0.18
CITY OF NEW YORK	1189	0.11
NEW YORK CITY HOUSING	1014	0.10
BOARD OF EDUCATION	1003	0.10
CNY/NYCTA	975	0.09
NYC HOUSING PARTNERSH	747	0.07
DEPT OF ENVIRONMENTAL	644	0.06
YORKVILLE TOWERS ASSO	558	0.05
DEPARTMENT OF BUSINES	526	0.05
DEPT OF TRANSPORTATIO	484	0.05
MTA/LIRR	467	0.04
PARCKHESTER PRESERVAT	439	0.04
MH RESIDENTIAL 1, LLC	411	0.04
434 M LLC	393	0.04
LINCOLN PLAZA ASSOCIA	366	0.03
DEUTSCHE BANK NATIONA	333	0.03
561 11TH AVENUE TMG L	324	0.03
CPW TOWERS	314	0.03
OCEAN SHELL LLC	314	0.03
DORCHESTER ASSOCIATES	313	0.03
PM PARTNERS	301	0.03
99 JOHN ST.,LLC	296	0.03

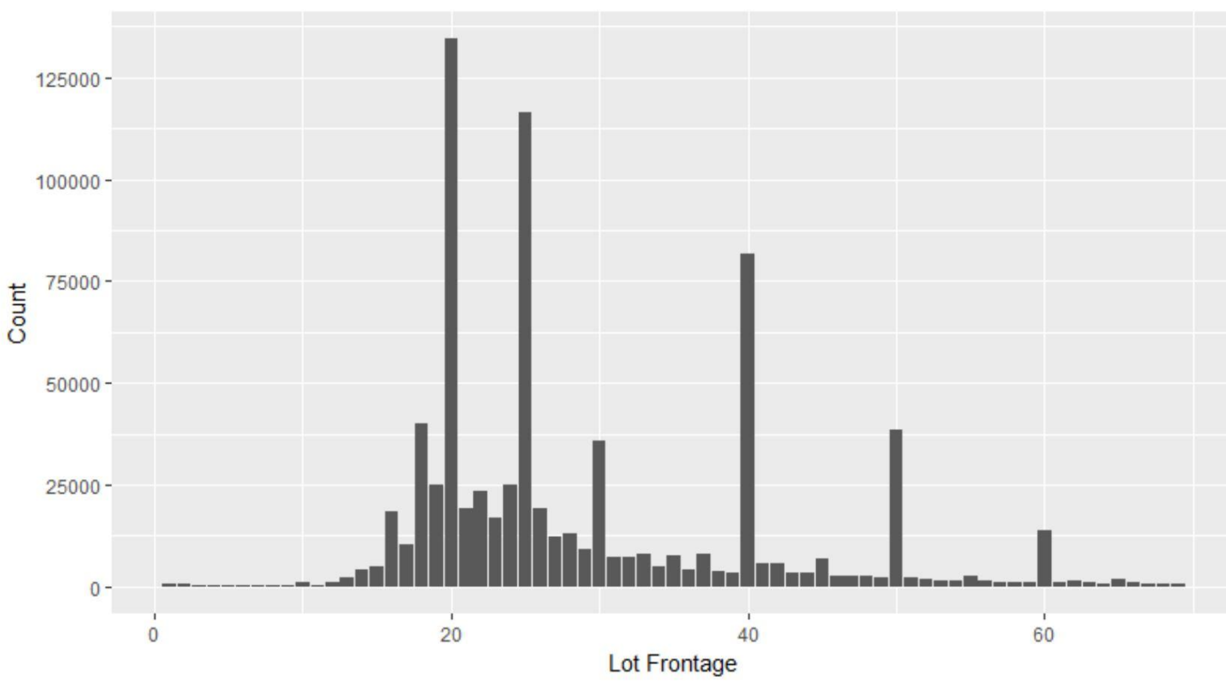
Field Name	BLDGCL
Field Type	Character
% Field Populated	100%
% of "NA" Values	0%
No. of Unique Values	200
Minimum Value	-
Maximum Value	-
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Building Class There is direct correlation between Building Class and Tax Class



Field Name	TAXCLASS
Field Type	Character
% Field Populated	100%
% of “NA” Values	0%
No. of Unique Values	11
Minimum Value	-
Maximum Value	-
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Tax Class There is direct correlation between Building Class and Tax Class

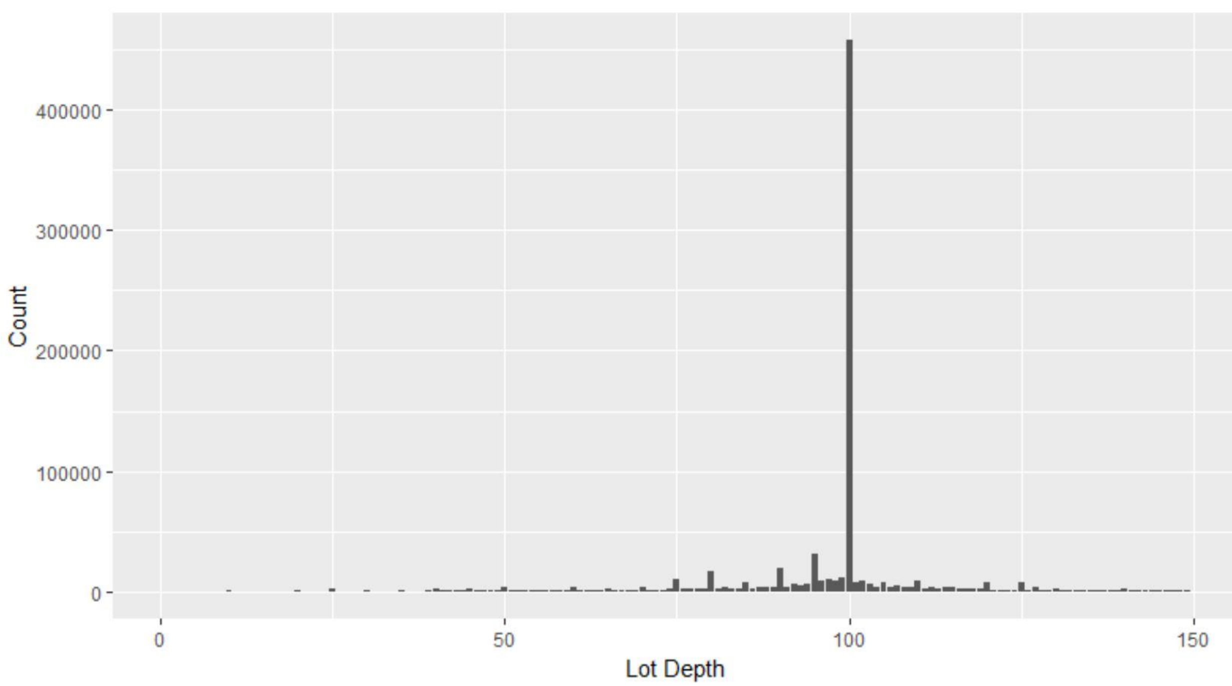


Field Name	LTFRONT
Field Type	Numeric
% Field Populated	100%
% of "NA" Values	0%
No. of Unique Values	1277
Records with Value 0	168867
Minimum Value	0 feet
Maximum Value	9999 feet
Mean	36.17 feet
1st Quartile	19 feet
Median	25 feet
3rd Quartile	40 feet
Std. Dev	73.73 feet
Description	Lot Frontage in Feet
Comments	For Graph below, • Removed records with value 0 • # of records with value 0: 168867 (16.10% of field) • Lot Frontage limited to 70 feet or below (covers 91.02% of field)

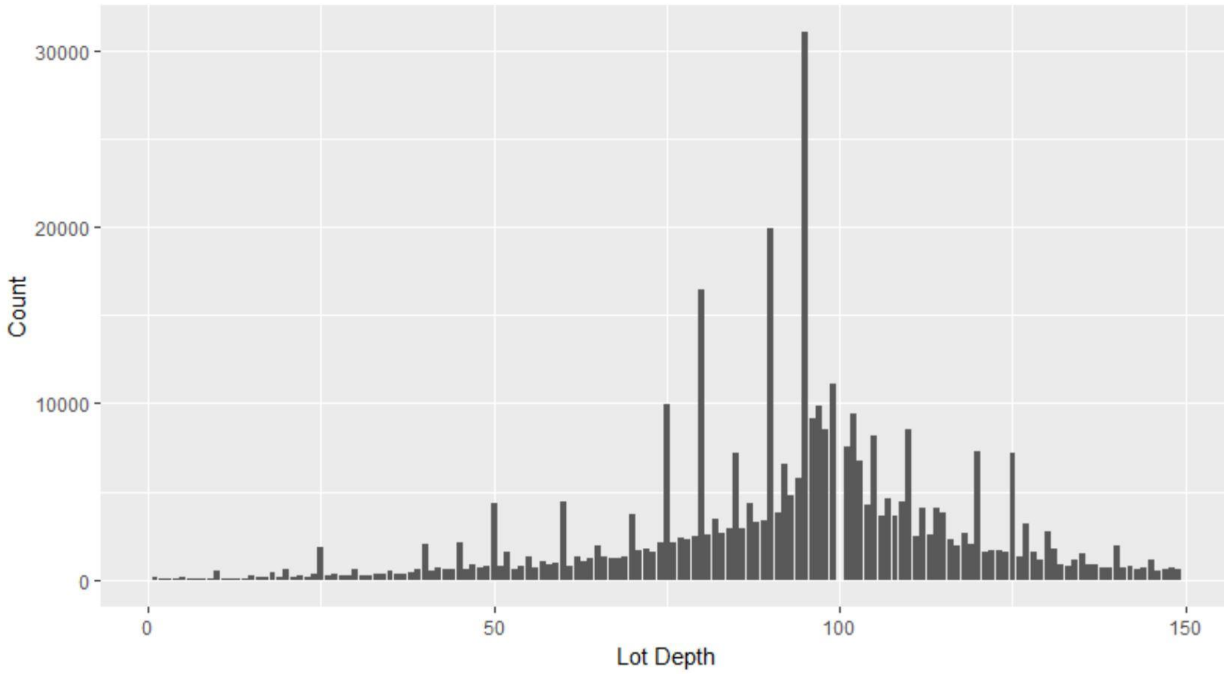


Field Name	LOTDEPTH
Field Type	Numeric
% Field Populated	100%
% of "NA" Values	0%
No. of Unique Values	1346
Records with Value 0	169888
Minimum Value	0 feet
Maximum Value	9999 feet
Mean	88.28 feet
1st Quartile	80 feet
Median	100 feet
3rd Quartile	100 feet
Std. Dev	75.47 feet
Description	Lot Depth in Feet
Comments	For Graph below, • Removed records with value 0 • Graph shown with and without records with Lot Depth of 100 feet

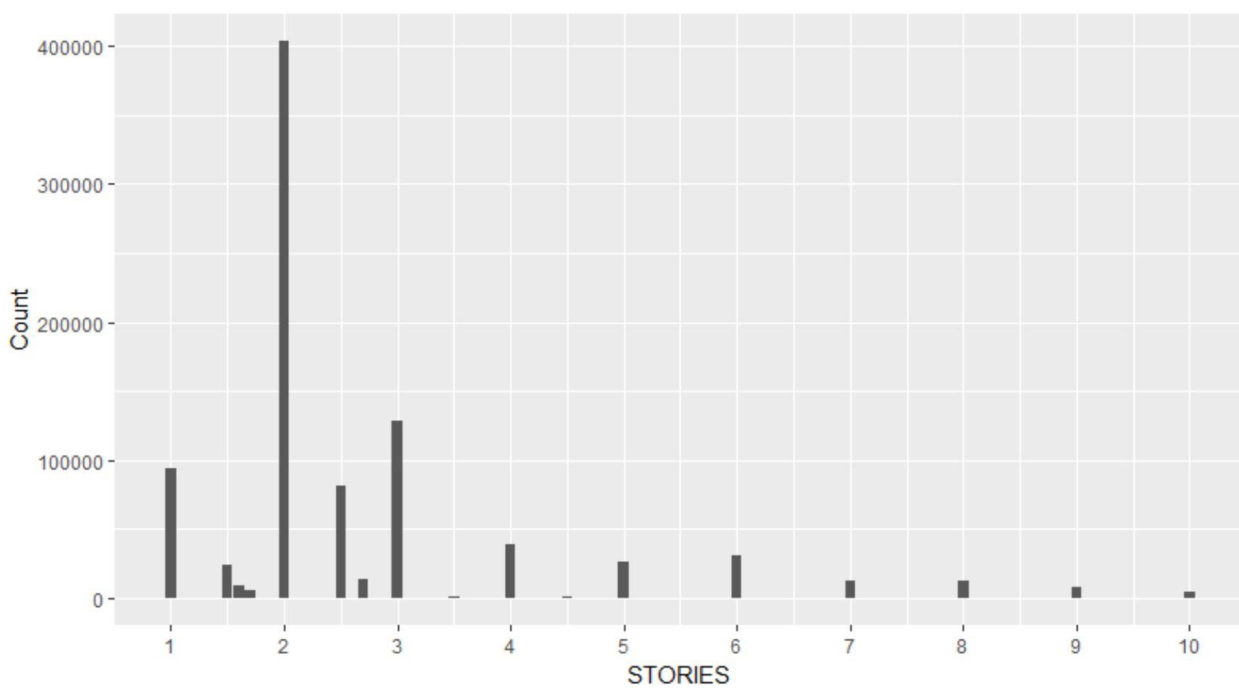
With Lot Depth 100 feet



Without Lot Depth 100 feet

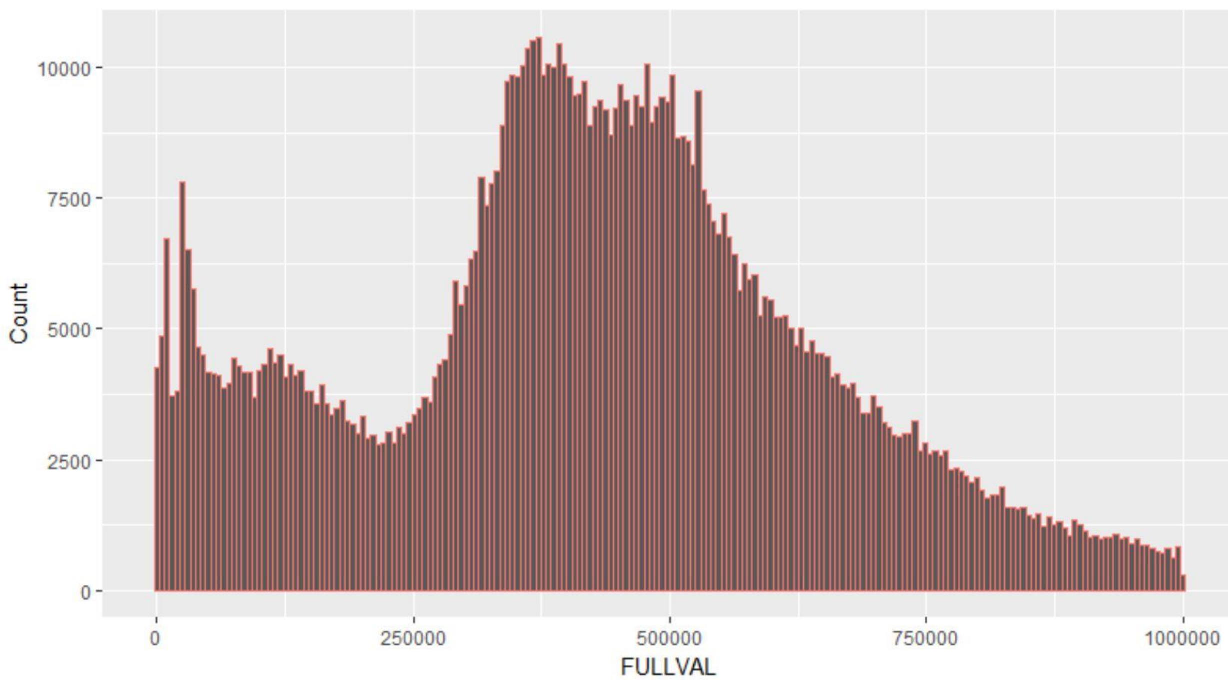


Field Name	STORIES
Field Type	Numeric
% Field Populated	95.02%
% of “NA” Values	4.97%
No. of Unique Values	112
Records with Value 0	0
Minimum Value	1
Maximum Value	119
Mean	5.06
1st Quartile	2
Median	2
3rd Quartile	3
Std. Dev	8.43
Description	Number of Stories in a Property
Comments	For Graph below, • Limited to records with Stories <= 10 (covers 85% of field)

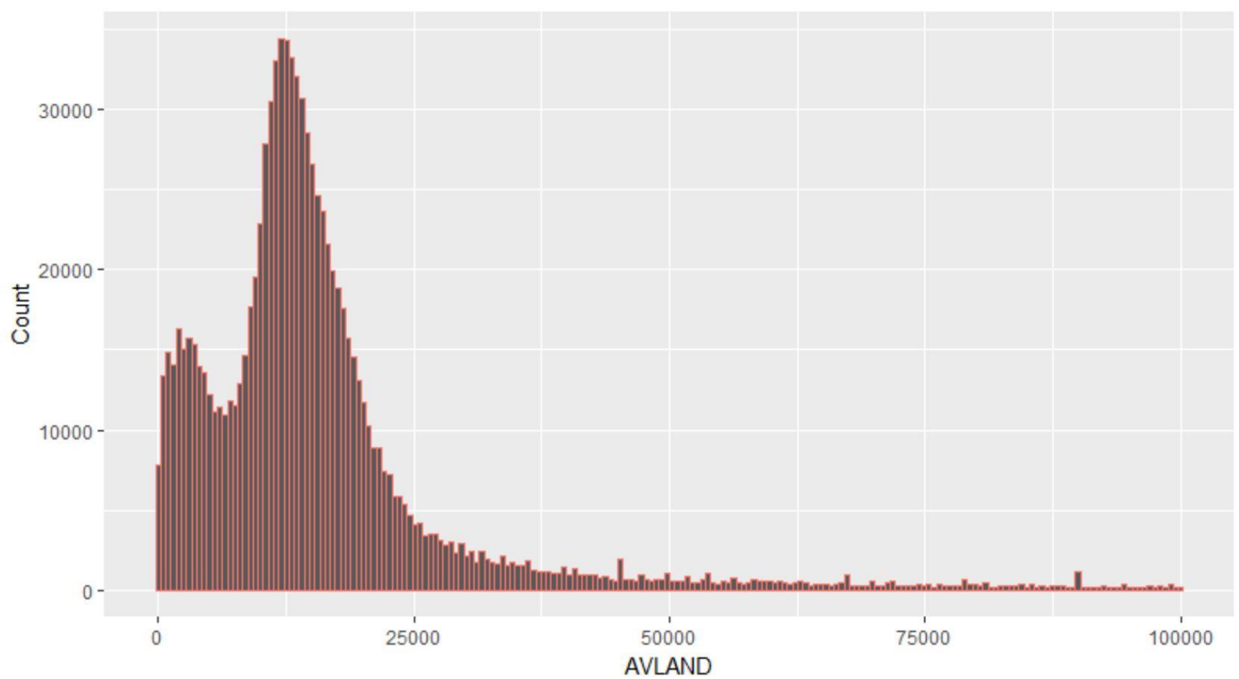


Field Name	FULLVAL
Field Type	Numeric
% Field Populated	100%
Number of "NA" Values	0%
No. of Unique Values	108277
Records with Value 0	12762
Minimum Value	0
Maximum Value	6150000000
Mean	880488
1st Quartile	303000
Median	446000
3rd Quartile	619000
Std. Dev	11702927
Description	Current Year's Total Market Value
Comments	For Graph below, • Removed all 0 value records • Limited to FULLVAL < 1,000,000 (covers 90% of field)

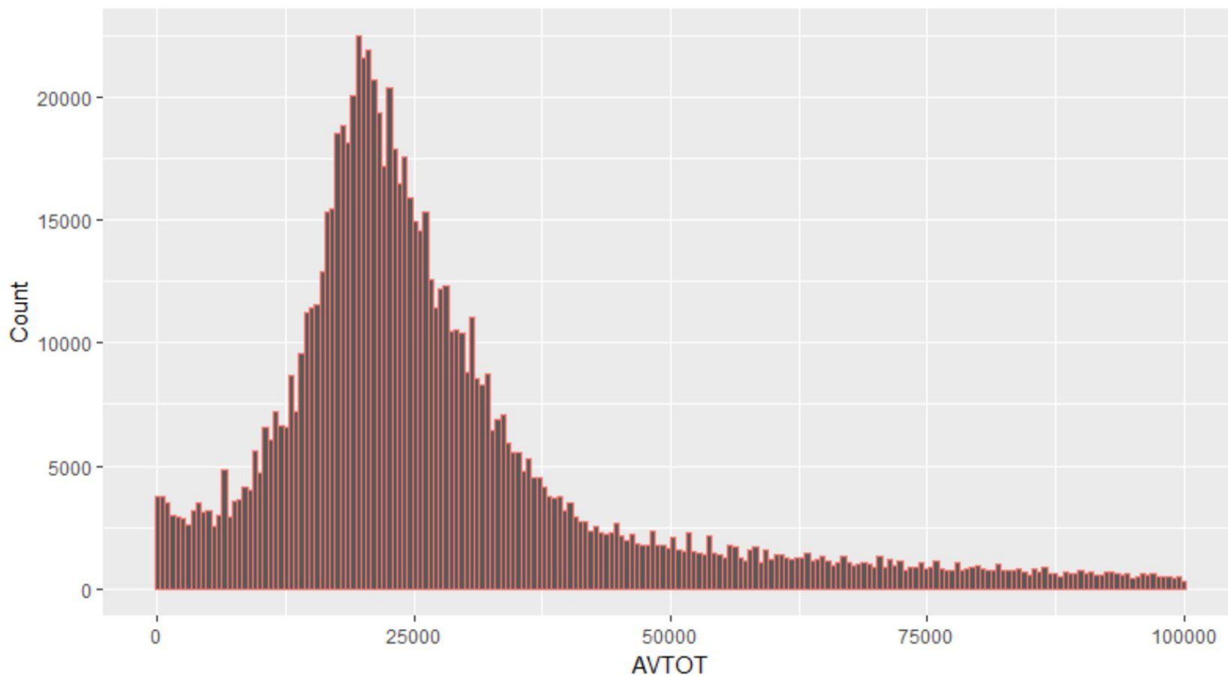
**



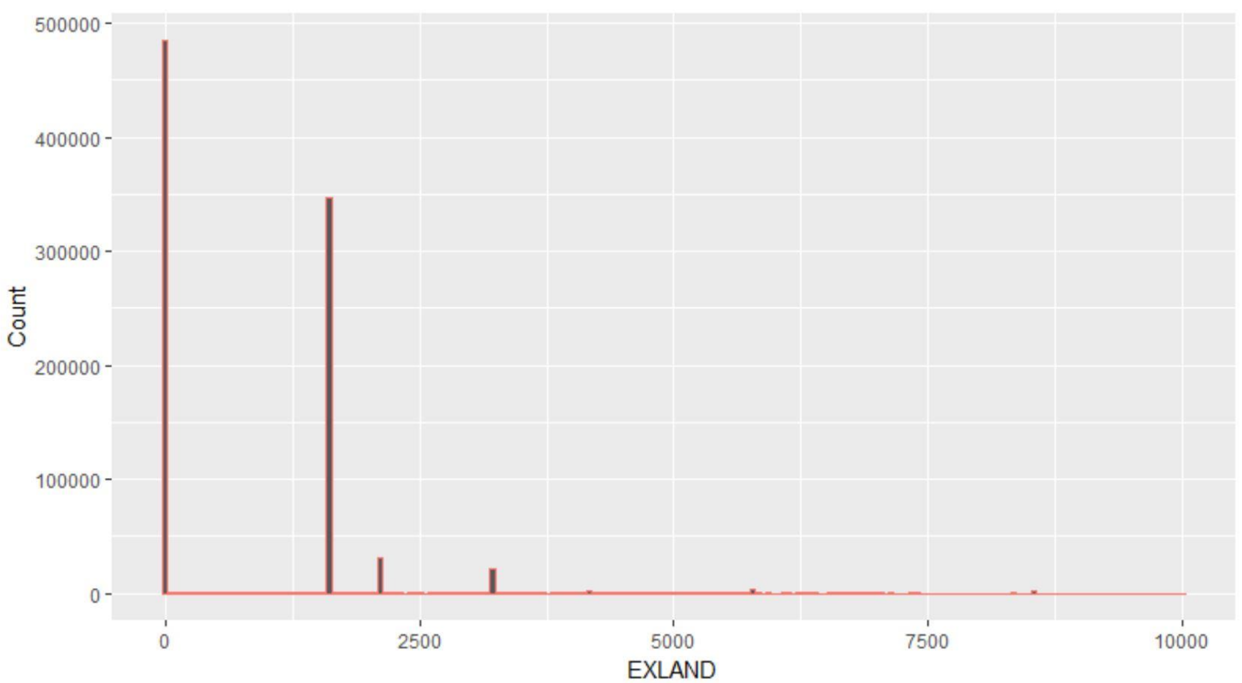
Field Name	AVLAND
Field Type	Numeric
% Field Populated	100%
Number of "NA" Values	0%
No. of Unique Values	70529
Records with Value 0	12764
Minimum Value	0
Maximum Value	2668500000
Mean	85995
1st Quartile	9160
Median	13646
3rd Quartile	19706
Std. Dev	4100755
Description	Current Year's Assessed Value of Land
Comments	Graph below, • Removed records with value 0 • Limited to AVLAND less than 100,000 (covers 92.73% of field)



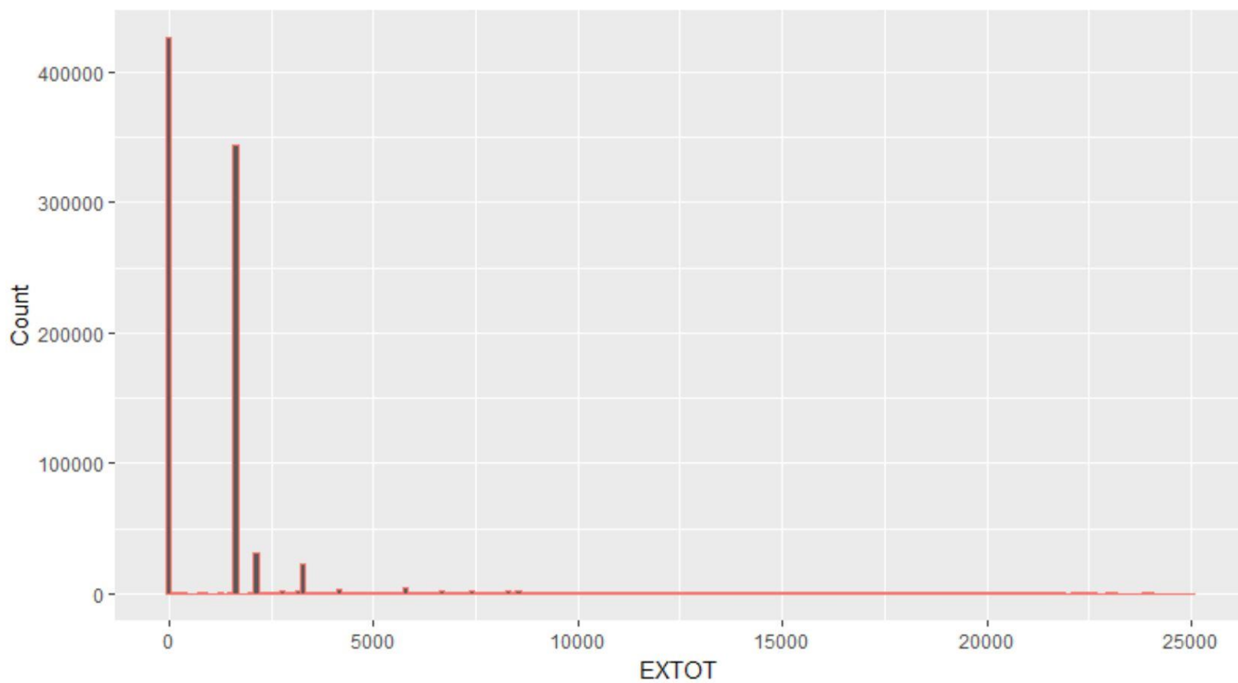
Field Name	AVTOT
Field Type	Numeric
% Field Populated	100%
Number of "NA" Values	0%
No. of Unique Values	112294
Records with Value 0	12762
Minimum Value	0
Maximum Value	4668308947
Mean	230758
1st Quartile	18385
Median	25339
3rd Quartile	46095
Std. Dev	6951206
Description	Current Year's Total Assessed Value
Comments	Graph below, • Removed records with value 0 • Limited to AVTOT < 100,000 (covers 84.62% of field)



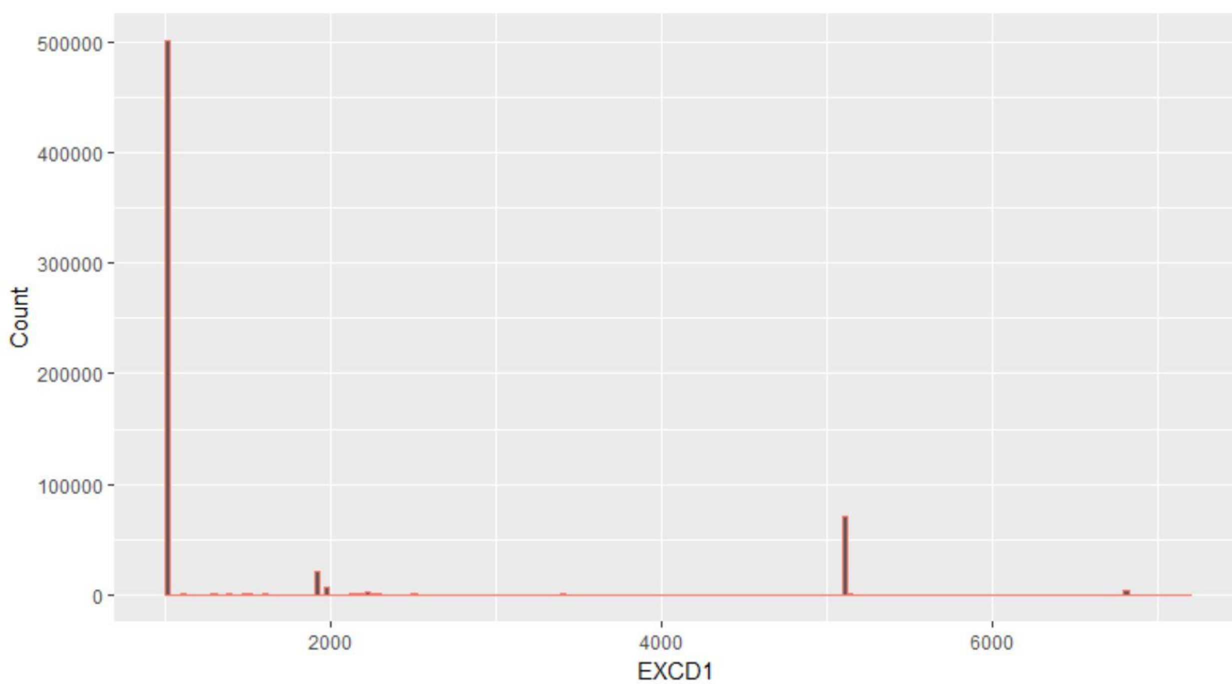
Field Name	EXLAND
Field Type	Numeric
% Field Populated	100%
Number of "NA" Values	0%
No. of Unique Values	33186
Records with Value 0	484224
Minimum Value	0
Maximum Value	2668500000
Mean	36812
1st Quartile	0
Median	1620
3rd Quartile	1620
Std. Dev	4024330
Description	Current Year's Exempt Value of Land
Comments	Graph below, • 46% records with 0 value • 33% records with 1620 value



Field Name	EXTOT
Field Type	Numeric
% Field Populated	100%
Number of “NA” Values	0%
No. of Unique Values	63805
Records with Value 0	425999
Minimum Value	0
Maximum Value	4668308947
Mean	92544
1st Quartile	0
Median	1620
3rd Quartile	2090
Std. Dev	6578281
Description	Current Year’s Exempt Value Total
Comments	Graph below, • 40% records have 0 value • 33% records have 1620 value



Field Name	EXCD1
Field Type	Numeric
% Field Populated	59.37%
Number of "NA" Values	40.62%
No. of Unique Values	130
Records with Value 0	0
Minimum Value	1010
Maximum Value	7170
Mean	1604
1st Quartile	1017
Median	1017
3rd Quartile	1017
Std. Dev	1388.13
Description	-
Comments	Graph below, • 40% records missing • 39.5% records have value 1017 • 2.27% records have value 5113



Field Name	STADDR
Field Type	Character
% Field Populated	99.94%
% of “NA” Values	0.06%
No. of Unique Values	820638
Records with Value 0	-
Minimum Value	-
Maximum Value	-
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Street Address of the Property

Top 25 Street Addresses

Street Address	**# of Records**	**% of Field**
501 SURF AVENUE	902	0.09
330 EAST 38 STREET	817	0.08
322 WEST 57 STREET	720	0.07
155 WEST 68 STREET	671	0.06
20 WEST 64 STREET	657	0.06
1 IRVING PLACE	650	0.06
220 RIVERSIDE BOULEVARD	628	0.06
360 FURMAN STREET	599	0.06
200 EAST 66 STREET	585	0.06
30 WEST 63 STREET	562	0.05
2 BAY CLUB DRIVE	556	0.05
350 WEST 42 STREET	556	0.05
200 RECTOR PLACE	549	0.05
301 EAST 79 STREET	538	0.05
350 WEST 50 STREET	498	0.05
630 1 AVENUE	488	0.05
635 WEST 42 STREET	483	0.05
88 GREENWICH STREET	453	0.04
150 WEST 51 STREET	447	0.04
99 JOHN STREET	445	0.04
25 CENTRAL PARK WEST	441	0.04
138-35 ELDER AVENUE	437	0.04
1623 3 AVENUE	434	0.04
1 BAY CLUB DRIVE	427	0.04
5 EAST 22 STREET	426	0.04

Field Name	ZIP
Field Type	Numeric
% Field Populated	97.48%
Number of "NA" Values	2.51%
No. of Unique Values	197
Records with Value 0	0
Minimum Value	-
Maximum Value	-
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	Zip Code corresponding to property

Top 25 ZIP Codes

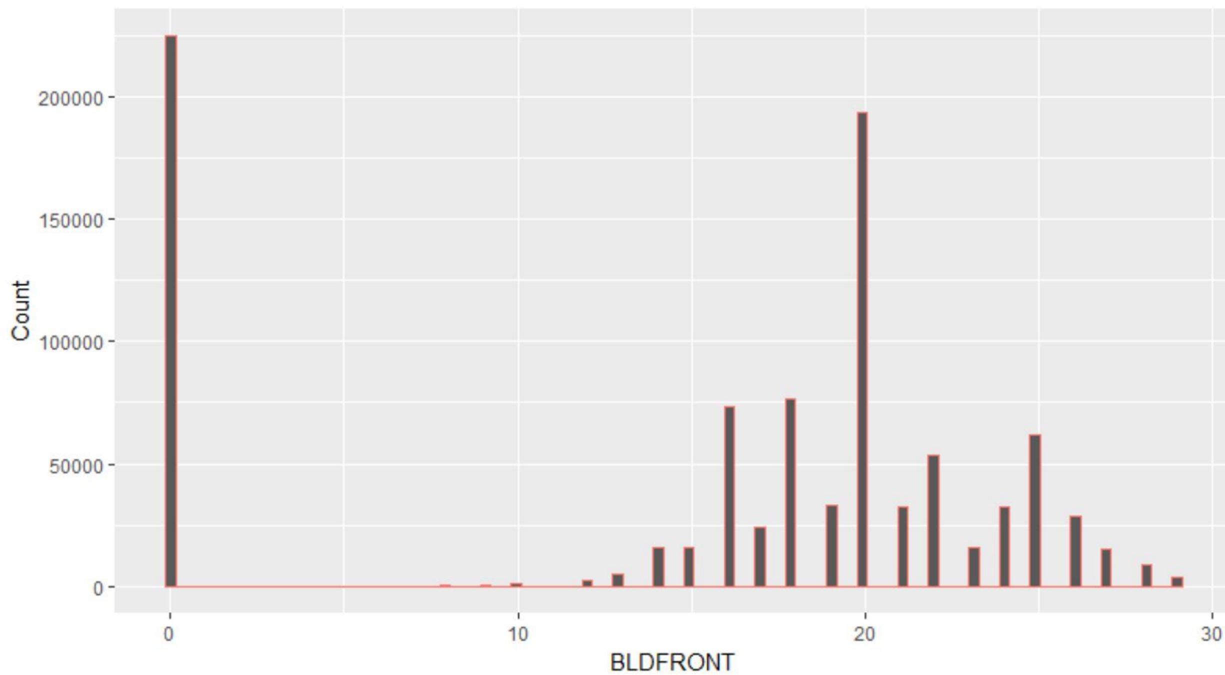
ZIP Code	# of Records	% of Field
10314	24605	2.35
11234	20001	1.91
10462	16905	1.61
10306	16576	1.58
11236	15678	1.50
11385	14921	1.42
11229	12793	1.22
11211	12710	1.21
10312	12634	1.20
11207	12293	1.17
11215	11834	1.13
11235	11312	1.08
11203	11241	1.07
11208	11139	1.06
11204	11061	1.05
10469	11030	1.05
11214	10886	1.04
11223	10741	1.02
10305	10624	1.01
11434	10505	1.00
11355	10492	1.00
11219	10300	0.98
11357	9851	0.94
11413	9784	0.93
11373	9779	0.93

Field Name	EXMPTCL
Field Type	Character
% Field Populated	1.42%
Number of "NA" Values	98.57%
No. of Unique Values	15
Records with Value 0	-
Minimum Value	-
Maximum Value	-
Mean	-
1st Quartile	-
Median	-
3rd Quartile	-
Std. Dev	-
Description	-

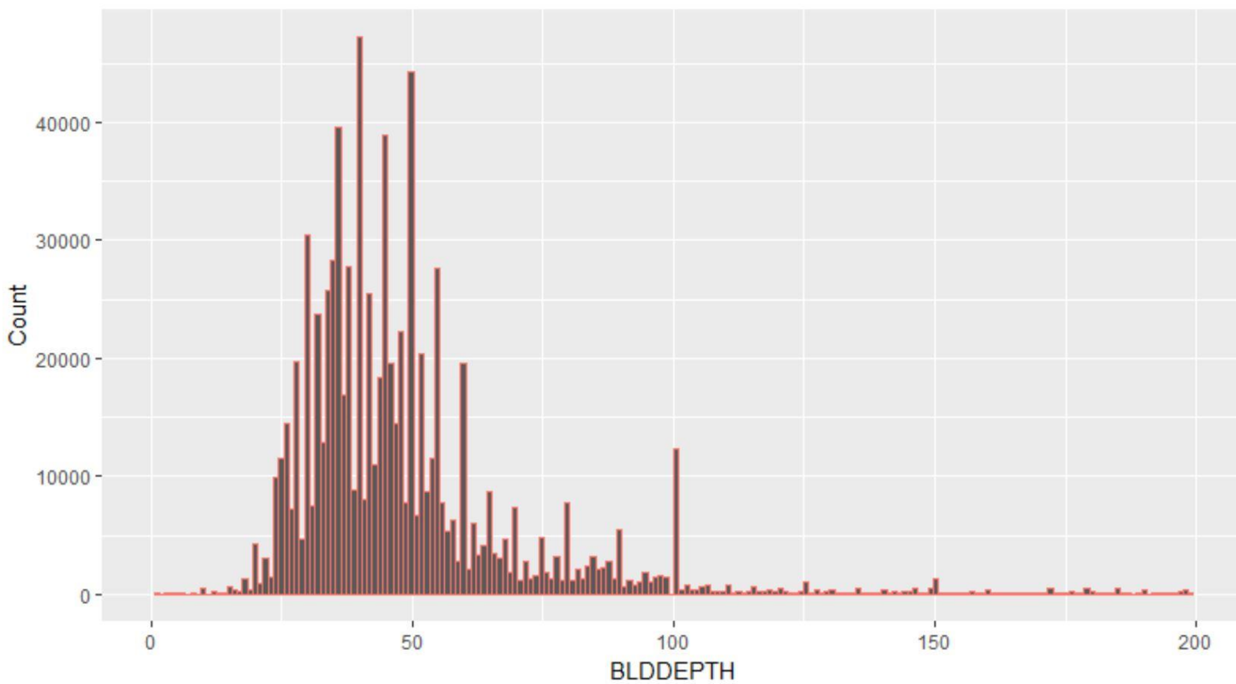
Top 10 EXMPTCL values

EXMPTCL	**# of Records**
X1	6494
X5	5158
X7	818
X6	760
X2	665
X4	438
X8	289
X3	260
X9	105
5	1

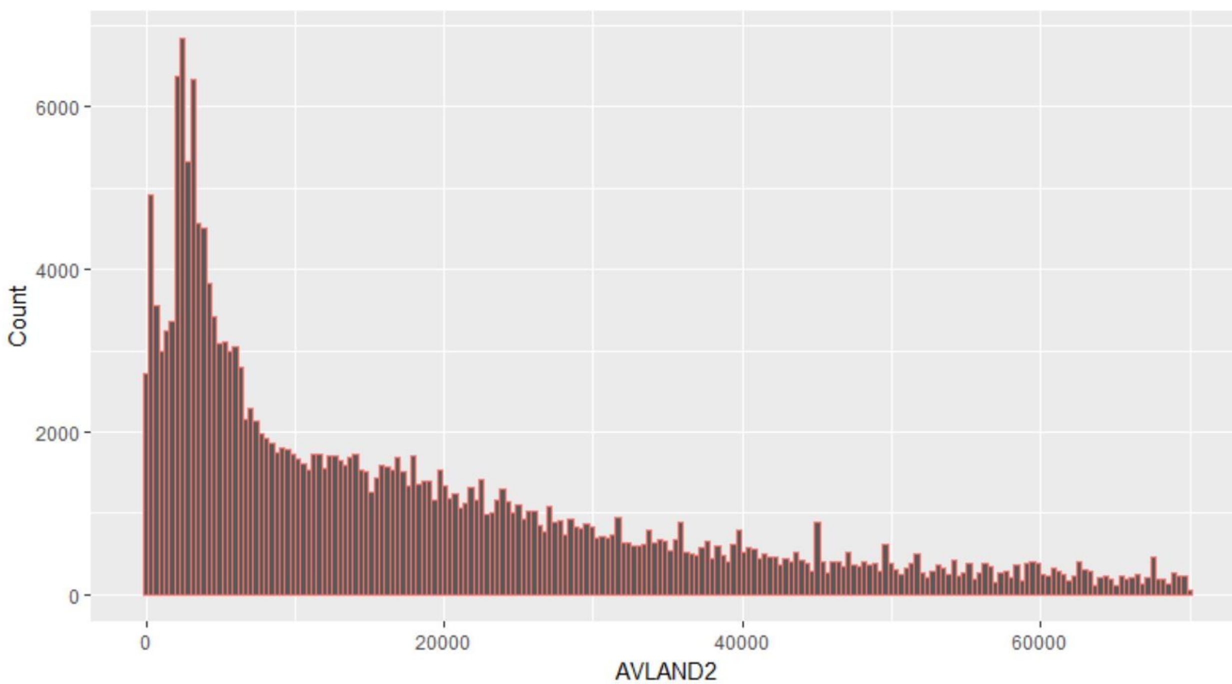
Field Name	BLDFRONT
Field Type	Numeric
% Field Populated	100%
Number of "NA" Values	0%
No. of Unique Values	610
Records with Value 0	224661
Minimum Value	0
Maximum Value	7575 feet
Mean	23.02 feet
1st Quartile	15 feet
Median	20 feet
3rd Quartile	24 feet
Std. Dev	35.78 feet
Description	Building Frontage in Feet
Comments	• 21% records have 0 Building Frontage • 18% records have 20 Building Frontage



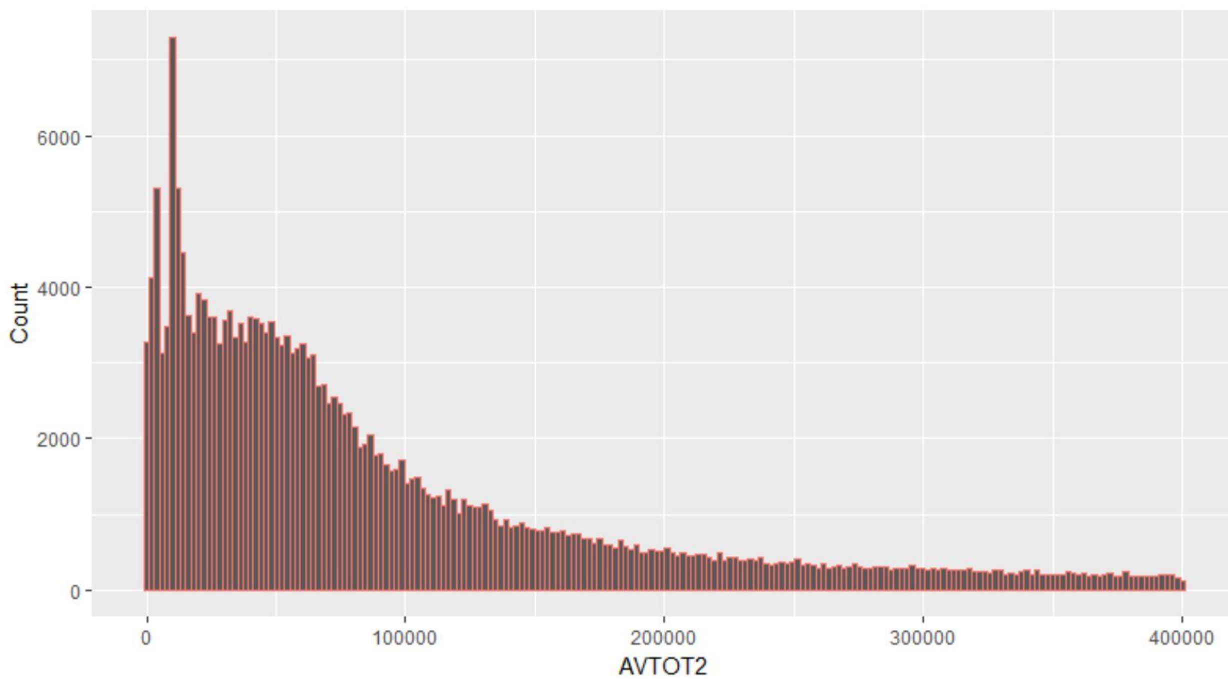
Field Name	BLDDEPTH
Field Type	Numeric
% Field Populated	100%
Number of “NA” Values	0%
No. of Unique Values	620
Records with Value 0	224699
Minimum Value	0
Maximum Value	9393 feet
Mean	40.07 feet
1st Quartile	26 feet
Median	39 feet
3rd Quartile	51 feet
Std. Dev	43.03 feet
Description	Building Depth in Feet
Comments	For Graph below, • Removed records with 0 Building Depth (21.42% of field) • Limited to BLDDEPTH <200 (covers 77.96% of field)



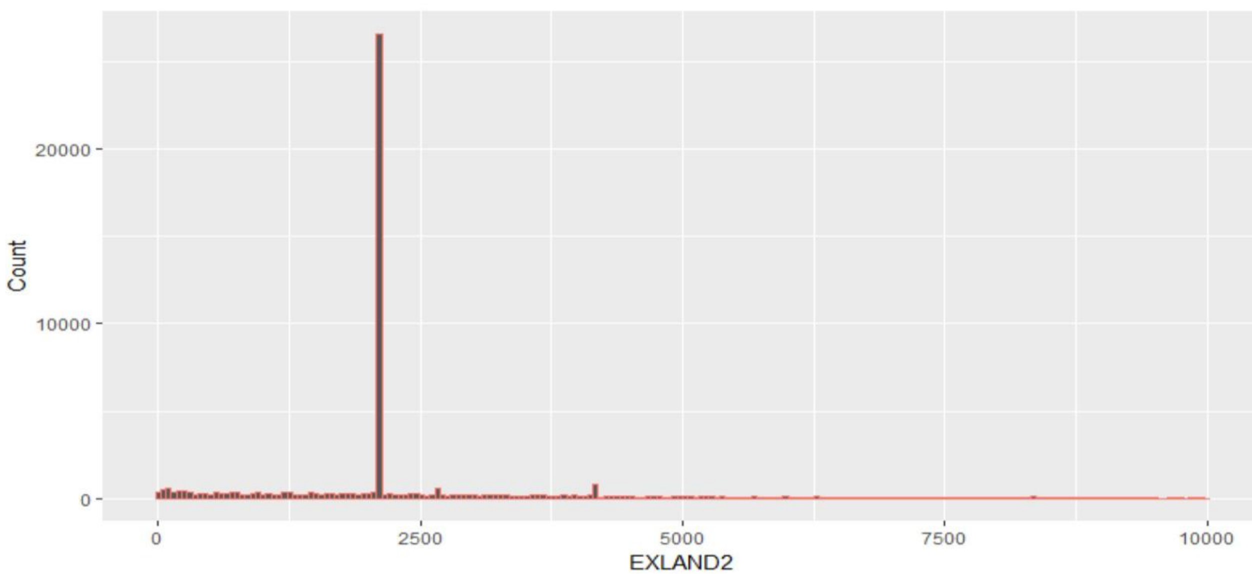
Field Name	AVLAND2
Field Type	Numeric
% Field Populated	26.8%
Number of “NA” Values	73.2%
No. of Unique Values	58170
Records with Value 0	0
Minimum Value	3
Maximum Value	2371005000
Mean	246365
1st Quartile	5705
Median	20059
3rd Quartile	62339
Std. Dev	6199390
Description	-
Comment	Graph below, • Limited to records with AVLAND2<70000 (covers 20.56% of field)



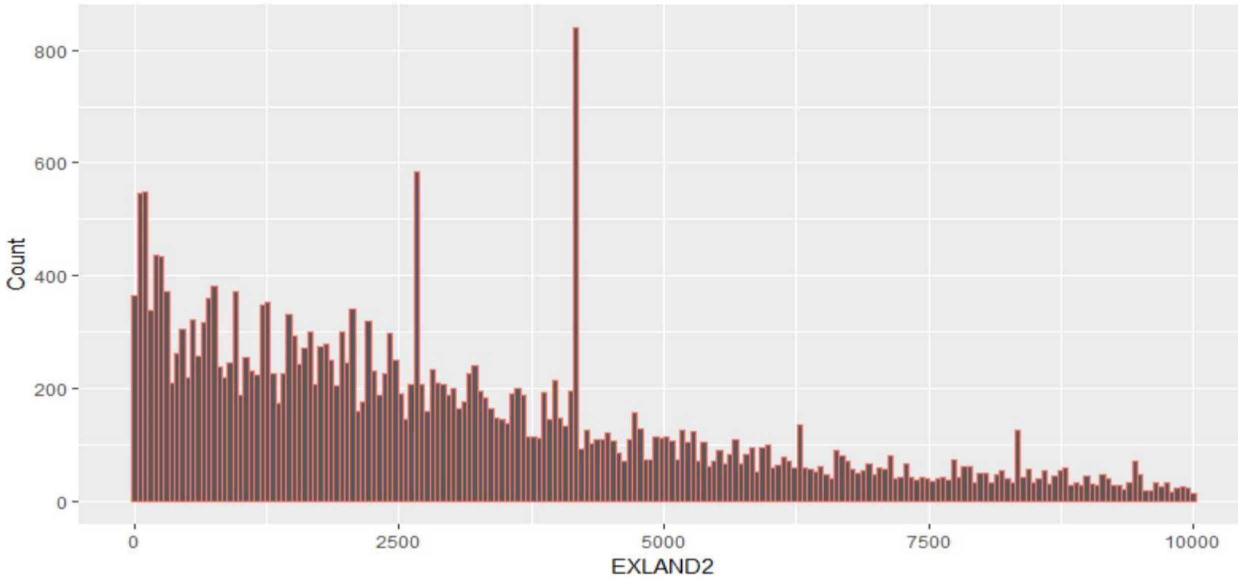
Field Name	AVTOT2
Field Type	Numeric
% Field Populated	26.8%
Number of "NA" Values	73.2%
No. of Unique Values	110891
Records with Value 0	0
Minimum Value	3
Maximum Value	4501180002
Mean	716079
1st Quartile	34014
Median	80010
3rd Quartile	240792
Std. Dev	11690165
Description	-
Comment	Graph below, • Limited to AVTOT2 < 400000 (covers 22.03% of field)



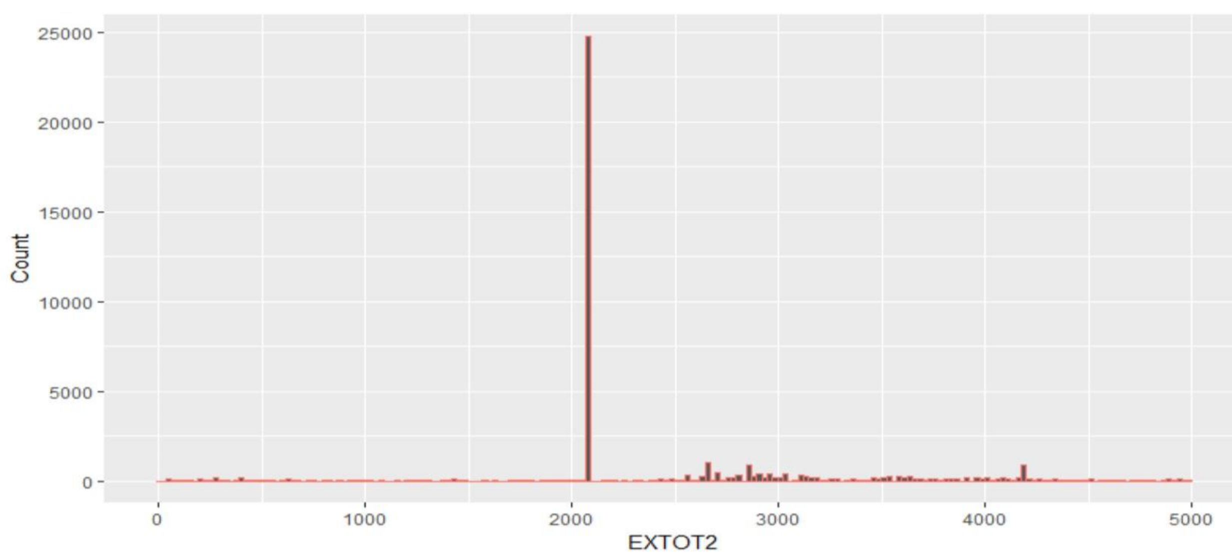
Field Name	EXLAND2
Field Type	Numeric
% Field Populated	8.26%
Number of "NA" Values	91.73%
No. of Unique Values	21997
Records with Value 0	0
Minimum Value	1
Maximum Value	2371005000
Mean	351802
1st Quartile	2090
Median	3053
3rd Quartile	31419
Std. Dev	10852484
Description	-
Comments	• 2.5% (~26000 records) have the value '2090'



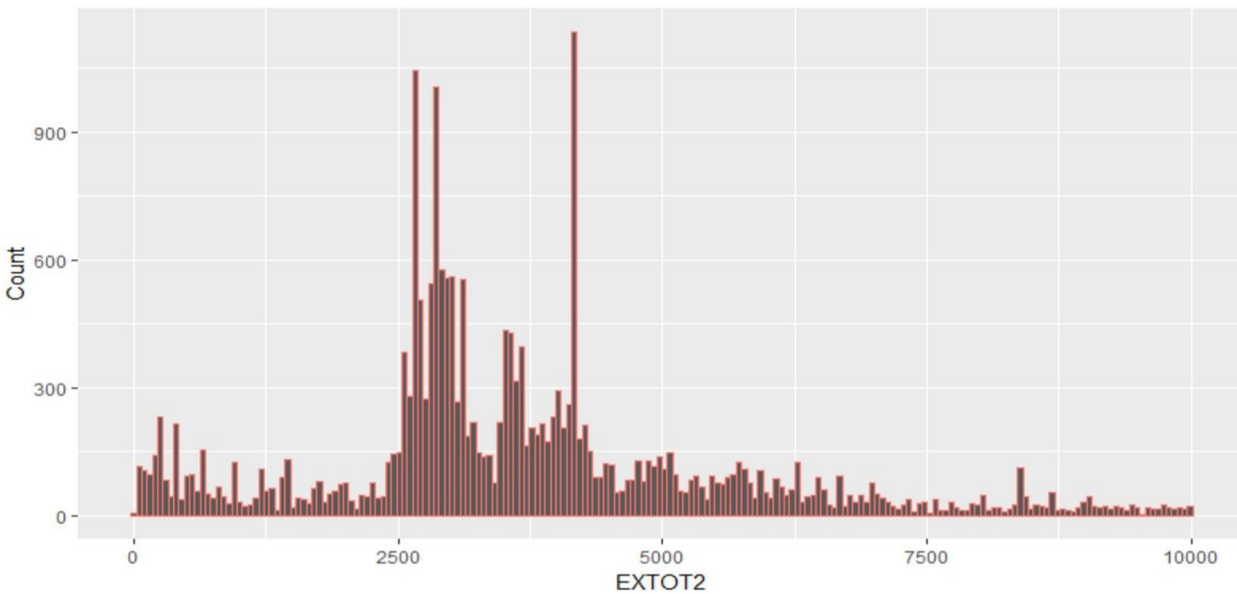
After removing Records with 2090 value



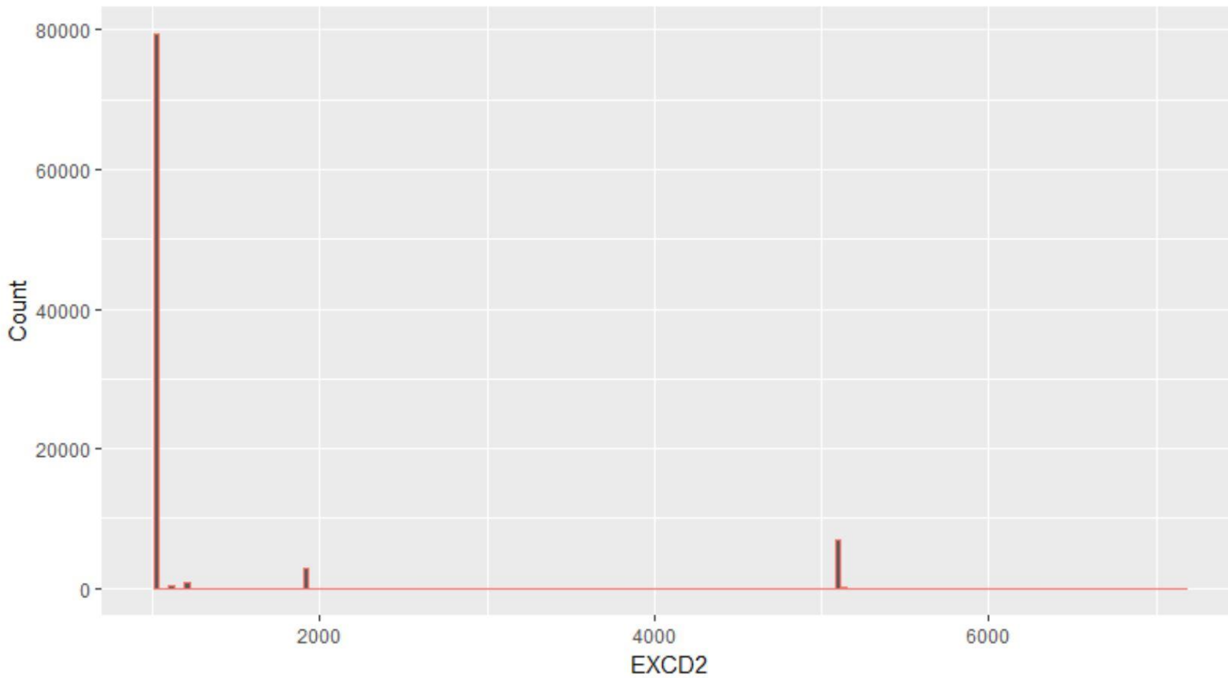
Field Name	EXTOT2
Field Type	Numeric
% Field Populated	12.39%
Number of “NA” Values	87.61%
No. of Unique Values	48107
Records with Value 0	0
Minimum Value	7
Maximum Value	4501180002
Mean	658115
1st Quartile	2889
Median	37116
3rd Quartile	106629
Std. Dev	16129808
Description	-
Comments	• 2.3% (~24739 records) have the value ‘2090’



After removing records with value ‘2090’



Field Name	EXTOT2
Field Type	Numeric
% Field Populated	8.67%
Number of "NA" Values	91.32%
No. of Unique Values	61
Records with Value 0	0
Minimum Value	1011
Maximum Value	7160
Mean	1372
1st Quartile	1017
Median	1017
3rd Quartile	1017
Std. Dev	1105
Description	-
Comments	• 6.12% records have the value '1017' • 1.14% have the value '1015' • 0.65% have the value '5112'


80000
60000
40000
20000
0
Count
2000
4000
6000
EXCD2

Field Name	PERIOD
Field Type	Character
% Field Populated	100%
Number of "NA" Values	0%
No. of Unique Values	1
Description	-
Comments	All records have same value "Final"

Field Name	YEAR
Field Type	Character
% Field Populated	100%
Number of "NA" Values	0%
No. of Unique Values	1
Description	-
Comments	All records have same value "2010/11"

Field Name	VALTYPE
Field Type	Character
% Field Populated	100%
Number of "NA" Values	0%
No. of Unique Values	1
Description	-
Comments	All records have same value "AC-TR"