DSO 562 Project 1

USC Marshall
School of Business

NY Property Data Fraud Report

UNDER THE GUIDANCE OF

Professor Stephen Coggeshall

2/13/2020

JENNY SHANG |

JENNY WANG |

LINGROU WANG |

RORY WANG |

Table of Contents

Part I. Executive Summary

This report is a summary on the analysis of New York property data extracted from the New York City government's Department of Finance. The dataset contains 1,070,994 real estate property records with 32 attributes, and it is collected for the purpose of calculating property tax, grant eligibility, exemptions and/or abatements. The objective of the analysis is to identify the top 10 most abnormal records within the dataset for further investigation of potentially fraudulent activity.

Since there is no prior knowledge nor expert indication of whether a record is fraudulent, an unsupervised learning method is utilized to identify the abnormal records. After thorough examination of the data across all 32 fields, missing values are populated using averages computed from records that share the same characteristics in one or many of the fields. Next, 45 expert variables are created and z-scaled to normalize the result. After reducing dimensionality to eight dimensions using Principal Component Analysis (PCA), 2 methods are used to calculate the fraud score for each record. The first method outputs the fraud score by utilizing a heuristic function to sum the normalized PCA values across the 8 dimensions. Records with higher values are perceived to be more abnormal. The second method produces the fraud score by first training an autoencoder, and then computing the difference between the actual value and the value predicted by autoencoder. Similar to the first method, records with higher values are perceived to be more abnormal. The final fraud score is a weighted combination of the two scores produced using each method. Records are rank ordered based on this final fraud score in descending order, where the top 10 highest scores are selected for investigation. This complete process is summarized in Figure 1.1 below.

Figure 1.1 - Overall Process of Project

The top 10 records with the highest final fraud scores are detailed in the Results section of the report. After examining each record, it can be concluded that while some abnormalities are due to the property being government-owned, many still require further investigation.

The following sections provide a detailed description of the data, and deep dive into the various steps taken in the analysis.

Part II. Description of Data

Overview of Data

The property valuation and assessment data are shared by the department of finance of the New York City government with an annual update. The main purpose of this data is to help the city government to calculate property tax, grant eligible property exemptions and or abatements. The data file is accessed from NYC OpenData online with an assessment year of 2010/11. There are 1,070,994 records and 32 fields in this data file, and a specific data quality report is conducted with details below. The report starts with a summary of numerical and categorical variables, followed with detailed individual analysis, and then concluded with questions for data preparation.

Summary Tables

There is one variable that is the unique identifier of the property, RECORD. In addition, there is a concatenated variable BBLE that contains a combination of information from the dataset. Details will be discussed in section II. Detailed Variable Visualization starting on page 3.

Table 1.1 - Numerical Variables Summary Table

Variable Name	Min	Mean	Max	Standard Deviation	# of Records with Value	% populated	Records with value of 0
LTFRONT	0.00E+00	3.66E+01	1.00E+04	7.40E+01	1,070,994	100%	169,108
LTDEPTH	0.00E+00	8.89E+01	1.00E+04	7.64E+01	1,070,994	100%	170,128
STORIES	1.00E+00	5.01E+00	1.19E+02	8.37E+00	1,014,730	98%	0 (56,264 Null)
FULLVAL	0.00E+00	8.74E+05	6.15E+09	1.16E+07	1,070,994	100%	13,007
AVLAND	0.00E+00	8.51E+04	2.67E+09	4.06E+06	1,070,994	100%	13,009
AVTOT	0.00E+00	2.27E+05	4.67E+09	6.88E+06	1,070,994	100%	13,007
EXLAND	0.00E+00	3.64E+04	2.67E+09	3.98E+06	1,070,994	100%	491,699
EXTOT	0.00E+00	9.12E+04	4.67E+09	6.51E+06	1,070,994	100%	432,572
BLDFRONT	0.00E+00	2.30E+01	7.58E+03	3.56E+01	1,070,994	100%	228,815
BLDDEPTH	0.00E+00	3.99E+01	9.39E+03	4.27E+01	1,070,994	100%	228,853
AVLAND2	3.00E+00	2.46E+05	2.37E+09	6.18E+06	282,726	26%	0
AVTOT2	3.00E+00	7.14E+05	4.50E+09	1.17E+07	282,732	26%	0
EXLAND2	1.00E+00	3.51E+05	2.37E+09	1.08E+07	87,449	8%	0
EXTOT2	7.00E+00	6.57E+05	4.50E+09	1.61E+07	130,828	12%	0

Table 1.2 – Categorical and Date/Time Variables Summary Table

Variable Name	Variable Type	Most Frequent Value	# of Records with Value	% populated	# of Unique Values	Records with value of 0
BBLE	Categorical	Unique Identifier	1,070,994	100%	1,070,994	0
B	Categorical	4	1,070,994	100%	5	0
BLOCK	Categorical	3944	1,070,994	100%	13,984	0
LOT	Categorical	1	1,070,994	100%	6,366	0
EXCD1	Categorical	1017	638,488	60%	130	0
EASEMENT	Categorical	E	4,636	0%	13	0
OWNER	Categorical	PARKCHESTER PRESERVAT	1,039,249	97%	863,347	0
EXCD2	Categorical	1017	92,948	9%	61	0
BLDGCL	Categorical	R4	1,070,994	100%	200	0
TAX-CLASS	Categorical	1	1,070,994	100%	11	0
EXT	Categorical	G	354,305	33%	3	0
STADDR	Categorical	501 SURF AVENUE	1,070,318	100%	839,281	0
ZIP	Categorical	10314	1,041,104	97%	197	0
EXMPTCL	Categorical	X1	15,579	1%	15	0
VALTYPE	Categorical	AC-TR	1,070,994	100%	1	0
PERIOD	Date / Time	FINAL	1,070,994	100%	1	0
YEAR	Date / Time	2010/11	1,070,994	100%	1	0

Valuable Visualization Graphs

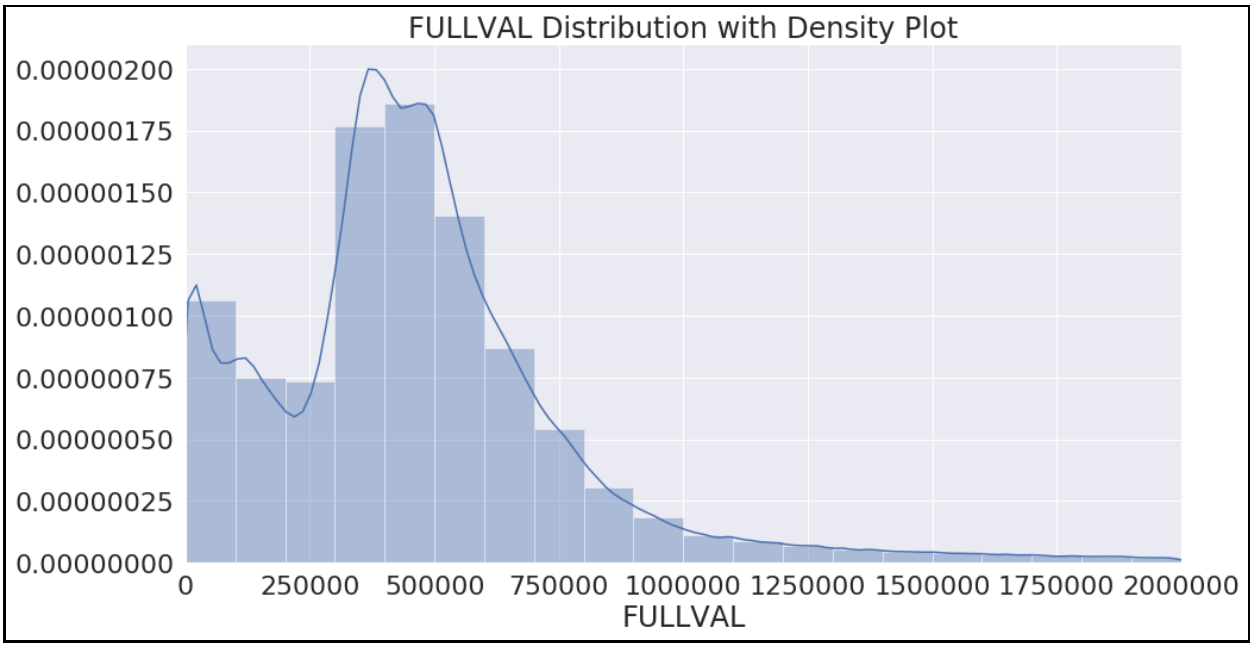


Figure 1.2 FULLVAL Distribution Graph

This graph suggested that there is a noticeable portion of properties with full value at zero, and the majority of properties are around $250,000 to $750,000. It will be essential to investigate and fill in the zero-valued properties appropriately. Other similar value fields of properties like AVTOT and AVLAND have a similar distribution.

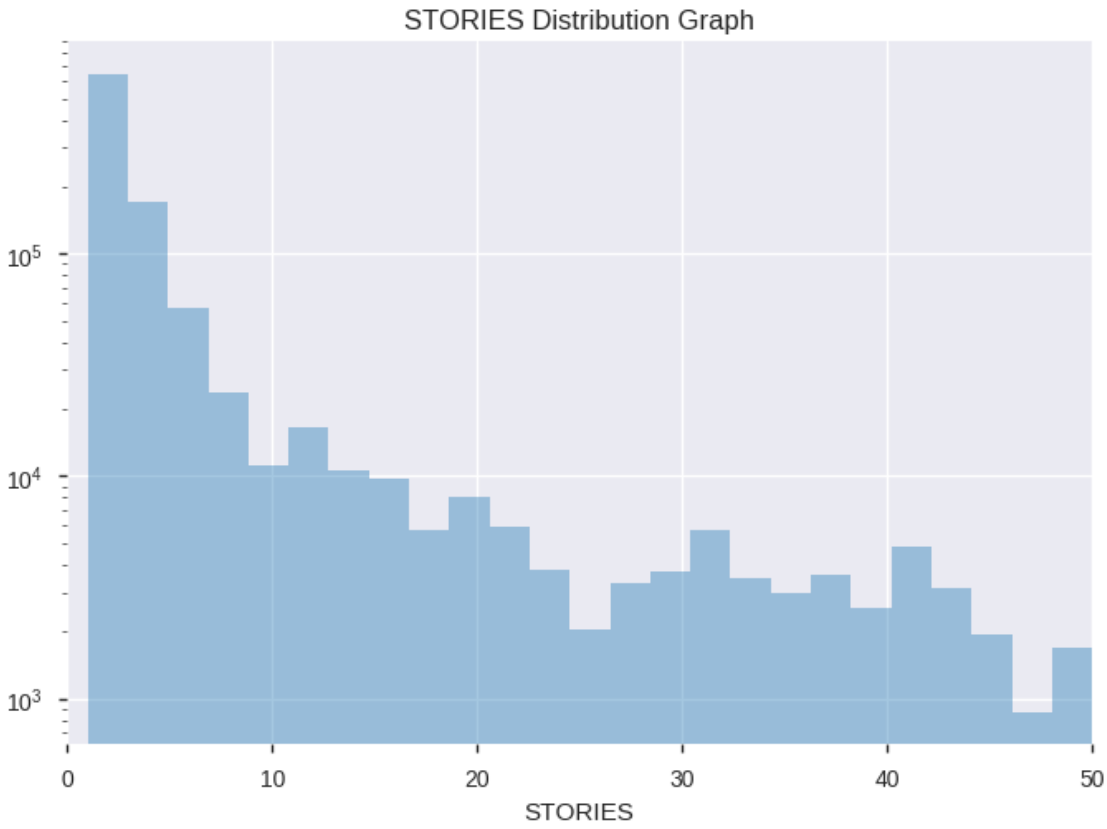


Figure 1.3 STORIES Distribution Graph

The graph above indicated that most properties have less than 10 floors, while some extreme properties will have over 20 floors. This information will be useful when filling in null values on the STORIES variable.

The graph below shows the distribution of the categorical variable, B, which is the borough code of a specific property. The dataset has more properties that are located on borough code four and three compared with other borough codes. Since all property within the dataset has a borough code, it is reasonable to consider using borough code as a grouping level when designing rules to fill in missing values.

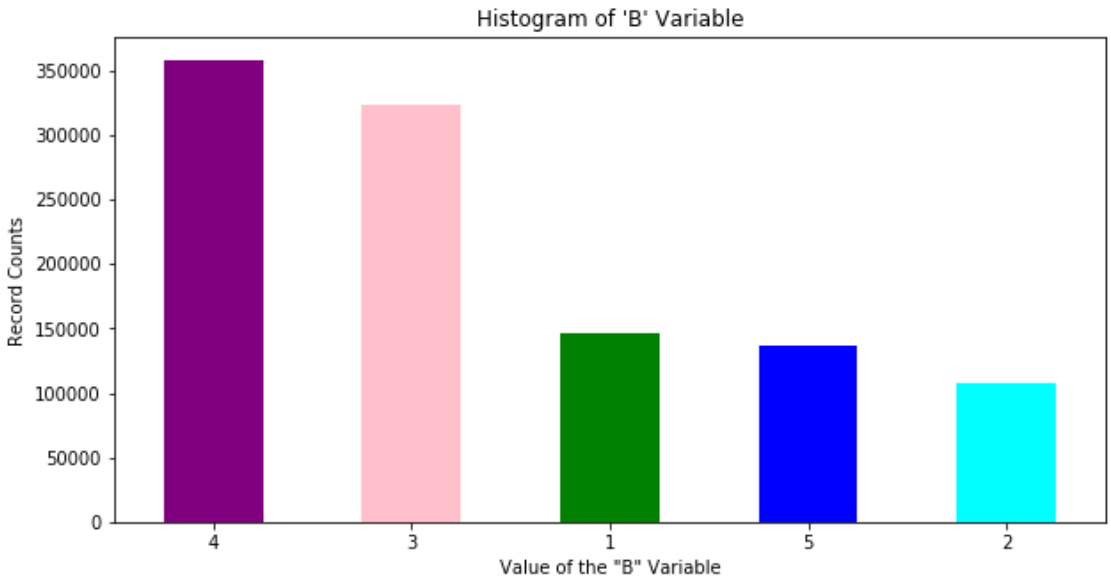


Figure 1.4 Borough Code Distribution Graph

ZIP Value	Count of Records	% of Total Valid Records
10314	24,606	2.4%
11234	20,001	1.9%
10312	18,127	1.7%
10462	16,905	1.6%
10306	16,578	1.6%
11236	15,678	1.5%
11385	14,921	1.4%
11229	12,793	1.2%
11211	12,710	1.2%
11207	12,293	1.2%

Table 1.3 Most Frequent Zip Codes

More graphs and data analysis are available in the data quality report in Appendix 1-1.

Part III. Data Cleaning

A standardized ground rule is finalized with three levels of filling null values:

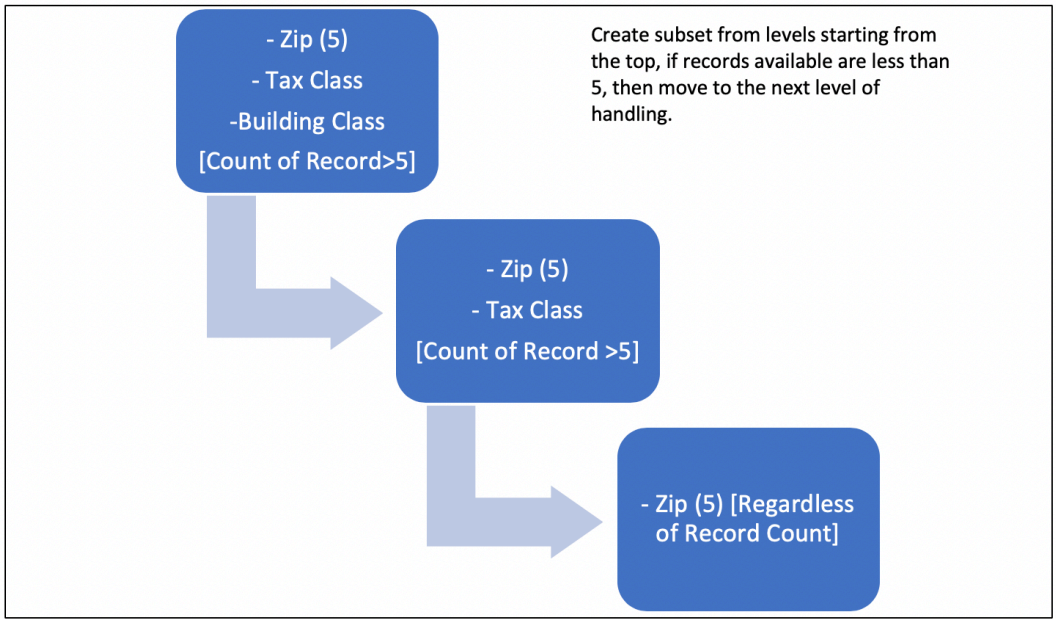


Figure 1.5 Overall Logic on Filling Null Values

The general process of filling in missing values in numerical variables has three layers with a bottom-up approach. For each layer, the variable will be grouped by different categories, in layer one, for example, a median value of the target variable will be used to fill null values if the count of records in that specific grouping is more than five. If the count is less than five, then the next level of grouping will be adopted.

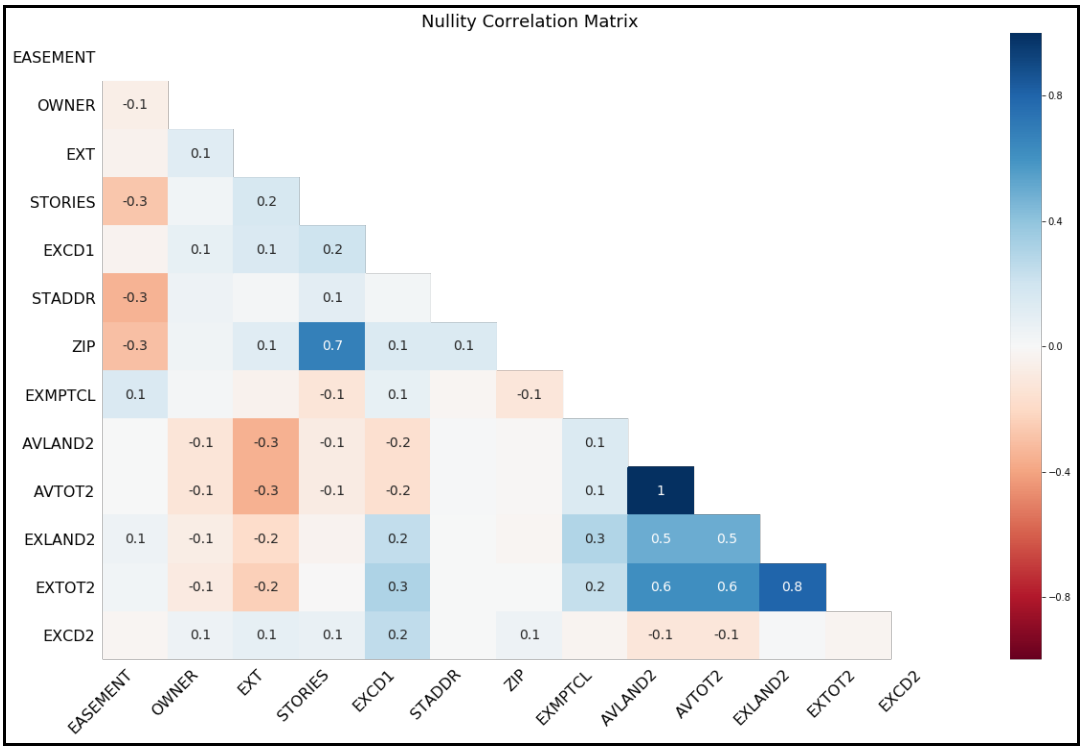


Figure 1.6 Nullity Correlation Matrix

One exception to the general process is for STORIES because it seems less effective to use the same hierarchy, because if STORIES is at null, ZIP is also very likely to be null. Therefore, a median value of STORIES with grouping initially at Borough, LOT and TAXCLASS combination level will be used to determine the substitute value. If the number of records on each subset is less than 5, LOT will be eliminated and only Borough and TAXCLASS will be used to estimate null values. If the subset is still not optimal, the overall median of STORIES by Borough code will be applied to the rest of the null records.

An overall table of filling null value progress of the nine variables is listed below:

Variable Name	Overall Null Value	Level 1 Null value Filled	Level 2 Null Value Filled	Level 3 Null Value Filled
AVLAND	13,009	2,652 (20%)	6,469 (62% of remaining null)	3,888 (100% of remaining null)
AVTOT	13,007	2,650 (20%)	6,469 (62% of remaining null)	3,888 (100% of remaining null)
BLDFRONT	228,815	153,178 (67%)	44,411 (59% of remaining null)	31,226 (100% of remaining null)*
BLDDEPTH	228,853	153,212 (67%)	44,513 (59% of remaining null)	31,128 (100% of remaining null)*
FULLVAL	13,007	487 (4%)	8,969 (72% of remaining null)	3,551 (100% of remaining null)
LTFRONT	169,108	152,893 (90%)	13,108 (81% of remaining null)	3,107 (100% of remaining null)
LTDEPTH	170,128	152,175 (89%)	14,792 (82% of remaining null)	3,161 (100% of remaining null)
STORIES	56,264	25,381 (45%)	8,566 (27% of remaining null)	22,317 (100% of remaining null)
ZIP	Adopted forward fill method.			

Table 1.4 Filling Null Value Progress Table

** Since there were still 191 null values remaining after level 3, these null values were filled using the median value of properties within the same borough.*

Part IV. Variable Creation

Before reducing dimensionality and utilizing different algorithms to look for anomalies in each property, relevant variables must be created. Since the goal is to find when the assessed value of the property is too high, the variables relevant to that value will be used to create new variables to perform the analysis. An important aspect of the value of the property is the size of the property. Therefore, three new variables were created to assess the size of each property. Nine variables were created to act as ratios of different sizes and values, and then to normalize these nine variables, the mean five different groups were applied onto these nine variables to finally produce a total of 45 variables. Further detail of this process is described below.
To create a total of 45 variables, the first task was to create three new size variables: Lot Area, Building Area, and Building Volume. The formulas for the new variables are given below, and are created from using already existing variables that have had their missing values completely filled.

LTAREA = LTFRNT * LTDEPTH
BLDAREA = BLDFRNT * BLDDEPTH
BLDVOL = BLDAREA * STORIES

After creating these three new variables, different symbols were assigned to the value and size variables in the existing data. These new variables would be later used to create the ratios necessary to the 45 variables. Below, S_1, S_2, and S_3 are the three new size variables that were created earlier. V_1, V_2, and V_3 are all pre-existing variables that indicate the different values that a property could have and that are significant to the algorithms that will be developed later on.

$$V_1 = \text{FULLVAL} \qquad S_1 = \text{LOTAREA}$$
$$V_2 = \text{AVLAND} \qquad S_2 = \text{BLDAREA}$$
$$V_3 = \text{AVTOT} \qquad S_3 = \text{BLDVOL}$$

Using these six variables, nine ratios were then created, as shown in the image below. V_1, V_2, and V_3 are the numerators and S_1, S_2, and S_3 are the denominators. Through this the values were all normalized by the different sizes, and nine unit-value variables were created. These nine variables are based on Lot Area, Building Area, and Building Volume, and will be the basis of the 45 variables that will later be used for Principal Component Analysis and the Autoencoder.

$$r_1 = \frac{V_1}{S_1} \qquad r_4 = \frac{V_2}{S_1} \qquad r_7 = \frac{V_3}{S_1}$$
$$r_2 = \frac{V_1}{S_2} \qquad r_5 = \frac{V_2}{S_2} \qquad r_8 = \frac{V_3}{S_2}$$
$$r_3 = \frac{V_1}{S_3} \qquad r_6 = \frac{V_2}{S_3} \qquad r_9 = \frac{V_3}{S_3}$$

In the next step, the nine unit-value variables were grouped by five scale groups: ZIP3, ZIP5, BOROUGH, TAXCLASS, and ALL (all of the data). For each group g, the average was calculated for the respective r_i, as shown in the image below. From this calculation 45 averages were created $< r_i >_g$ (i = 1,2,3,...9) from the five scale groups (g = 1,2,3,4,5).

$$\frac{r_1}{< r_1 >_g}, \quad \frac{r_2}{< r_2 >_g}, \quad \frac{r_3}{< r_3 >_g}, \quad \cdots \quad \frac{r_9}{< r_9 >_g}$$

Finally, a list of 45 variables was created. The names of the variables and the method that they were grouped by are listed below.

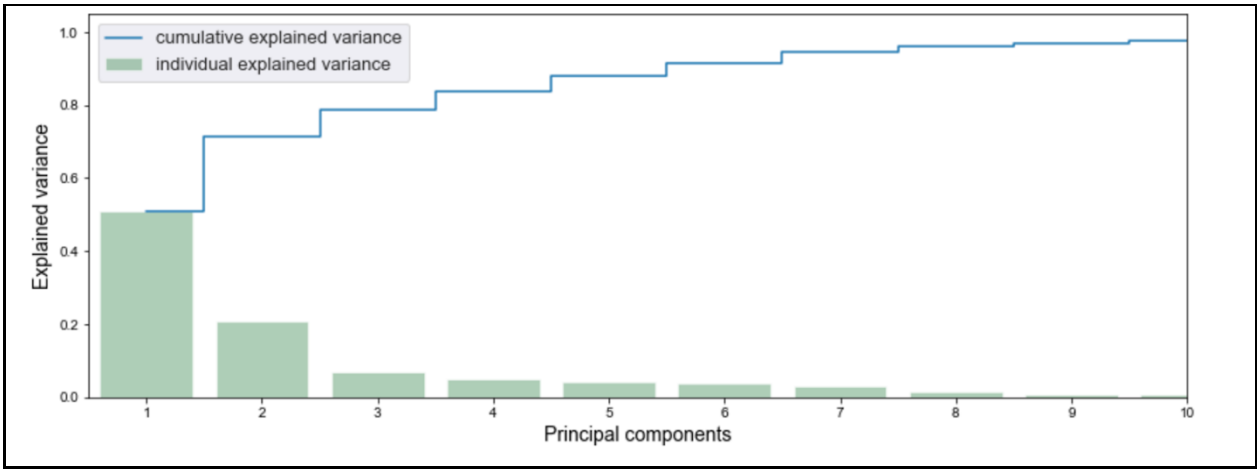
Grouped by: ZIP3	*Grouped by: ZIP5*	*Grouped by: Borough*	*Grouped by: Tax Class*	*Grouped by: All*
r_1ZIP3	r_1ZIP5	r_1B	r_1TAXCLASS	r_1r_1
r_2ZIP3	r_2ZIP5	r_2B	r_2TAXCLASS	r_2r_2
r_3ZIP3	r_3ZIP5	r_3B	r_3TAXCLASS	r_3r_3
r_4ZIP3	r_4ZIP5	r_4B	r_4TAXCLASS	r_4r_4
r_5ZIP3	r_5ZIP5	r_5B	r_5TAXCLASS	r_5r_5
r_6ZIP3	r_6ZIP5	r_6B	r_6TAXCLASS	r_6r_6
r_7ZIP3	r_7ZIP5	r_7B	r_7TAXCLASS	r_7r_7
r_8ZIP3	r_8ZIP5	r_8B	r_8TAXCLASS	r_8r_8
r_9ZIP3	r_9ZIP5	r_9B	r_9TAXCLASS	r_9r_9

Table 1.5 Overview of 45 variables Created

Part V. Dimensionality Reduction

After creating 45 variables, the next step is to perform Principal Component Analysis (PCA). PCA is an advanced technique to extract important independent and uncorrelated principal features from a large set of variables in a dataset. Before building fraud scores for each property record, PCA is used to reduce dimensionality and to remove correlation between features. To avoid the different scales and variances of original variables, z-scaling was performed beforehand to normalize variables in order to then perform PCA.

It's expected that there are overlaps and correlations between all 45 variables that were created, and PCA creates independent principal components that are a linear combination of the original variables, so the maximum variance can be extracted from variables with no correlation. This helps to reduce the number of variables and to transform them into a small number of features that contains the most important information. The top principal component (PC) captures the most variance in data, and the following PC's capture the remaining variance. In order to get the most important part of the data, only the top 8 PC's were kept and the remaining PC's with less influence were dropped. The top 8 PC's together explain 96.1% of the variance, whereas the remaining 37 PC's only explain 3.9% in total.



Graph 1.6 Principal Components Visual

PC	Cumulative Variance %
1	50.9
2	71.6
3	78.6
4	83.6
5	87.9
6	91.6
7	94.5
8	96.1

Table 1.6 Principle Component vs. Cumulative Variance Explained

PCA not only created PC's to capture most of the variance, it also satisfies the orthogonality condition, as each PC represents an eigenvector perpendicular to other PCs in the eigenspace. By creating a new orthogonal coordinate system, PCA helps remove the correlation of features.

Part VI. Explanation of Algorithms

Fraud Score 1

The first fraud score is calculated by standardizing the values in the matrix after dimension reduction and computing a sum of absolute standardized values across the dimensions according to a Heuristic function (see Figure 1.7 below). After reducing the dataset to focus on eight dimensions that explain 96.1% of the variance, the output is a matrix that consists of 1,074,994 rows (records) and 8 columns (dimensions). This matrix is then standardized again by taking the difference of each value and the average for the column, and dividing this difference by the standard deviation of that dimension. This second standardization is important because it ensures that the fraud score is not influenced by large or small values that are characteristic to a particular dimension. In other words, it places equal importance on all dimensions. The resulting values will reflect how much a particular value deviates from the average.

To capture the impact of all dimensions, the fraud score for a particular record takes the sum of the squared absolute value of each dimension, and square roots this sum to maintain the same unit of account. By taking the absolute value of each record, it eliminates the concern of negative values and positive values canceling out each other. The fraud score is then rank ordered in descending order where the record with the largest value would be the most abnormal record. This method produces a reasonable fraud score because records with large fraud scores have one or more dimensions where the value is relatively extreme and largely deviates from the average of the whole dataset.

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

Figure 1.7 Heuristic Function of the z-scores

Fraud Score 2

The second fraud score is calculated by first standardizing the values in the post-dimension reduction matrix, then training an autoencoder with the goal of replicating the data and finally computing the sum of the absolute value of the reconstruction errors (see Figure 1.8 below). The reconstruction error is the difference between the original standardized value and the output value according to the autoencoder.

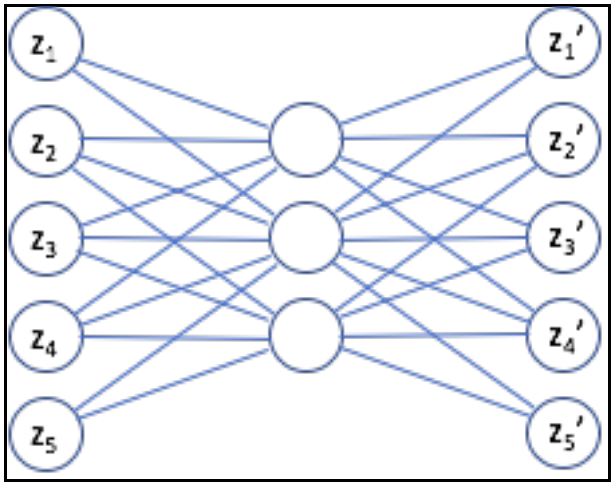


Figure 1.8 Autoencoder Visualization

The autoencoder is designed using the keras package in Python and takes the data table with all records and the eight standardized dimensions as training data. Specified parameters of the autoencoder include an input size, which is the number of dimensions in the training set (in this case 8), and the number of dimensions for each of the hidden layers. Using pre-built functions in the keras package and running on an epoch of 10, the autoencoder learns from each epoch and tries to reproduce the output as best as possible.

The fraud score for each record is taken as the Euclidean distance where the absolute difference between the original value and the value produced by the autoencoder is squared and summed across the dimensions (see Figure 1.9 below). This value is then square rooted to maintain the unit of account. Similar to Fraud Score 1, the most abnormal record would be the one with the largest value. This is due to the assertion that the value is so extreme compared to other records such that the autoencoder is not able to reproduce it well.

$$s_i = \left(\sum_k |z_k'^i - z_k^i|^n \right)^{1/n}, \quad n \text{ anything}$$

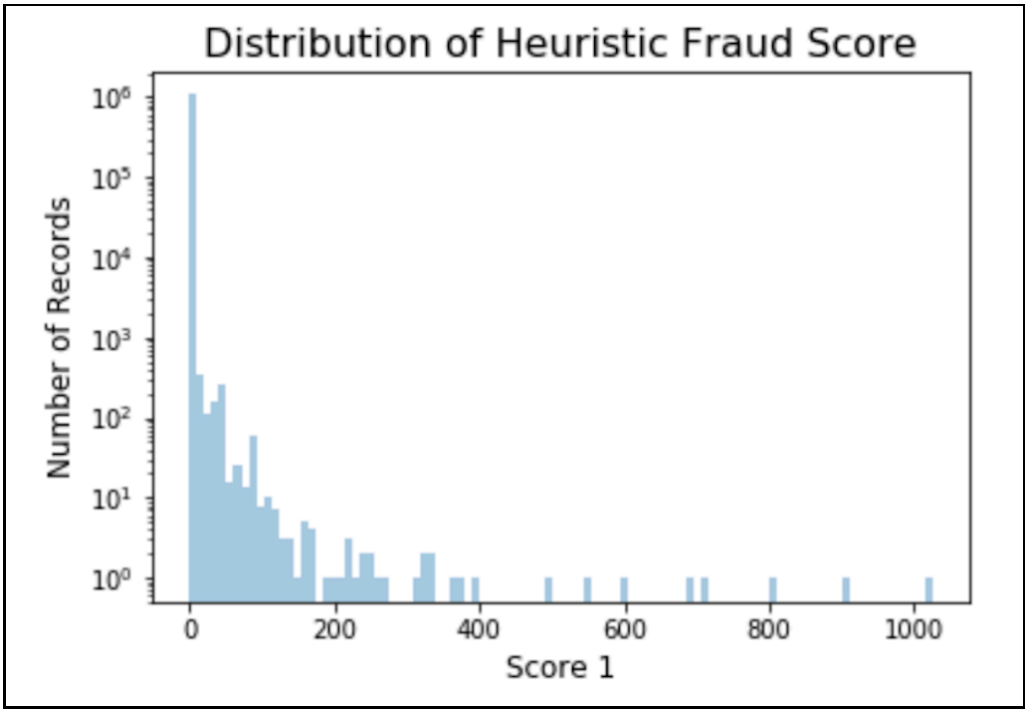
Figure 1.9 Euclidean distance calculation

Final Fraud Score

The final fraud score is the sum of the weighted average of Fraud Score 1 and Fraud Score 2. Since there is no specific preference in either of the scores, the weight is selected to be 0.5 for each. Once the final scores are calculated, the records are rank ordered in descending order where the record with highest score is perceived to be the most abnormal.

Part VII. Results

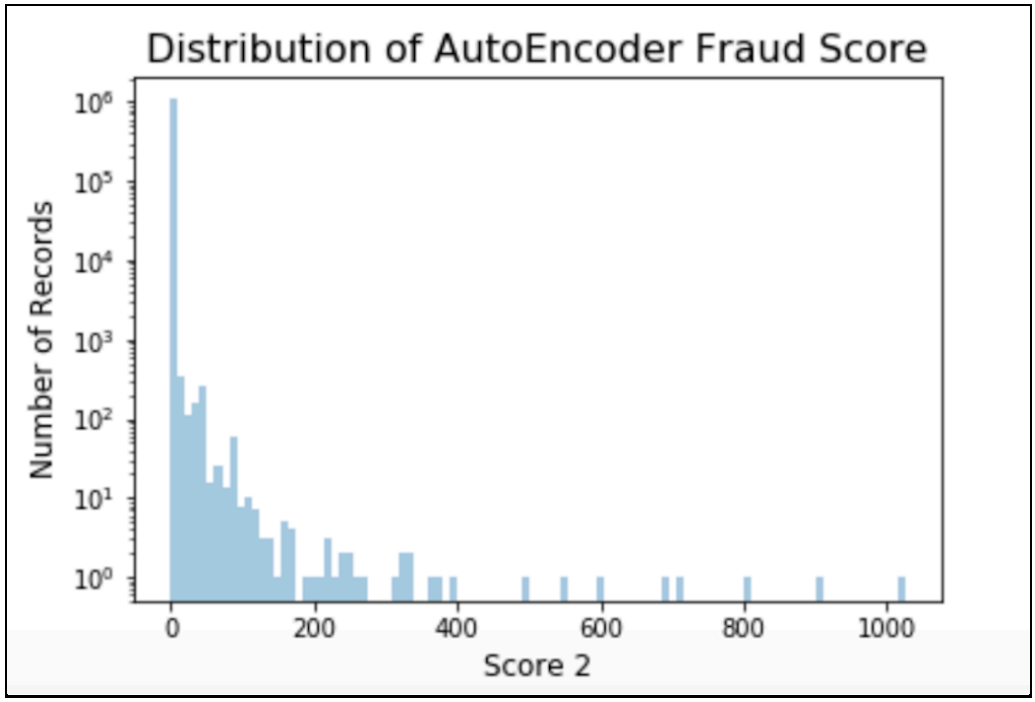
Distribution Graphs



Graph 1.10 Distribution of Heuristic Score Visual

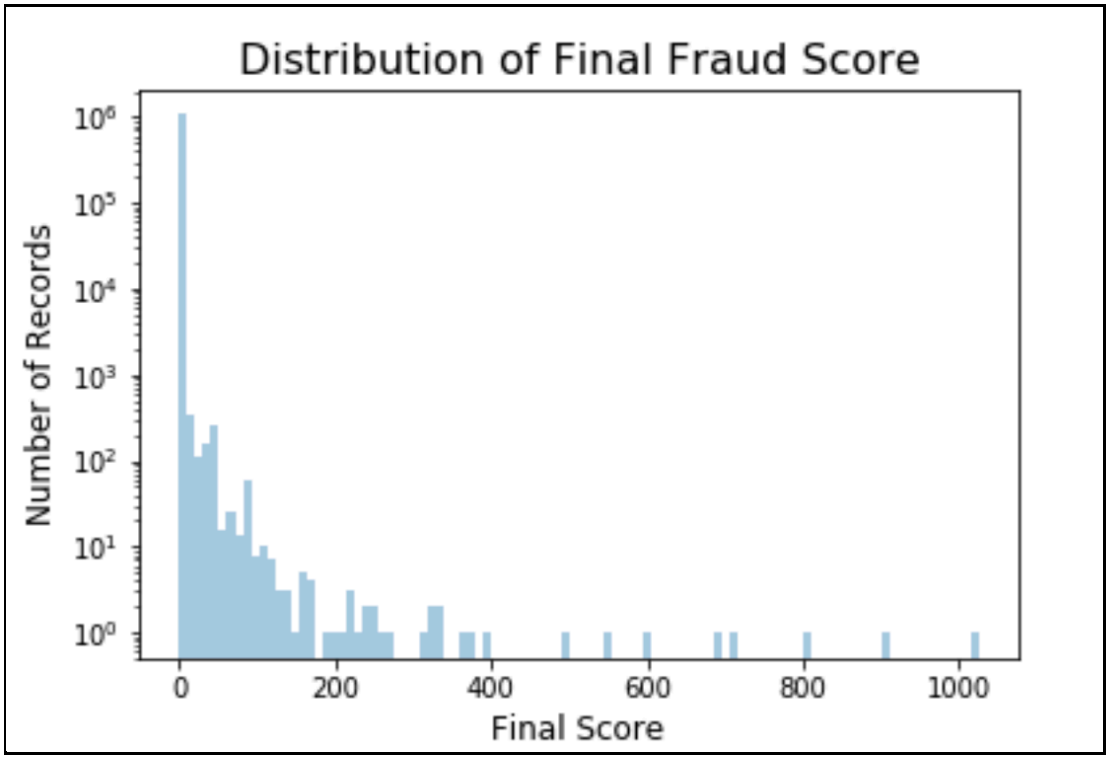
The graph below shows the distribution of fraud scores generated from the Heuristic function. The distribution is slightly right-skewed with a long tail.

The next graph shows the distribution of fraud scores generated from the Autoencoder function. The distribution is slightly right-skewed with a long tail.



Graph 1.11 Distribution of Autoencoder Score Visual

Combining all the fraud scores gained from the Heuristic model and Autoencoder model, and divided by 2, we got the final fraud scores. The distribution of final fraud scores are plotted as below:



Graph 1.12 Distribution of Final Fraud Score Visual

Normal & Abnormal Record Comparison

In order to understand more about the abnormal record, we compare the mean, median and standard deviation of normal with top 10 abnormal records.

	Normal Record			Top 10 Abnormal Record		
	Mean	Median	STD	Mean	Median	STD
FULLVAL	878,682	448,000	11,588,086	709,468,510	3,577,700	1,459,394,634
AVTOT	229,916	25,642	6,883,191	769,220,079	1,609,965	1,518,357,297
AVLAND	85,773	13,767	4,057,582	481,511,973	1,434,015	800,567,944
LTFRONT	61	30	104	1,018	148	1,829
LTDEPTH	114	100	84	237	129	301
BLDFRONT	40	22	48	16	14	15
BLDDEPTH	60	47	56	28	22	28
STORIES	5	2	8	6	2	8

Table 1.7 Normal vs. Abnormal Record

1. Based on the table above, we found that the top 10 abnormal records have extremely high FULLVAL, AVTOT and AVLAND. This indicates that potential fraud properties tend to have larger land area and are reported as high market value and assessed value.

2. From the table above, we also found that the top 10 abnormal records have higher LTFRONT and LTDEPTH, but have lower BLDFRONT and BLDDEPTH. This suggests that potential fraud properties tend to be reported with high lot frontage and lot depth, but with low building front and building depth.

Record #	Fraud Score 1	Fraud Score 2	Final Fraud Score	Highest Z-Score Variable	Highest Z-Score	Second Highest Z-Score Variable	Second Highest Z-Score
632816	1024.18	1023.58	1023.88	r_6TAXCLASS	1031.5	r_3TAXCLASS	1027.2
565392	905.73	903.97	904.85	r_4B	900.1	r_4ZIP3	878.1
1067360	801.69	799.65	800.67	r_7TAXCLASS	772.6	r_1ZIP5	688.9
917942	709.92	710.21	710.06	r_8ZIP3	632.1	r_8r_8	509.1
565398	692.90	692.36	692.63	r_6ZIP3	522.0	r_6B	498.6
585118	600.48	600.67	600.57	r_5r_5	532.9	r_5B	461.4
85886	544.70	543.51	544.10	r_3B	482.2	r_3ZIP3	478.9
585439	497.37	497.14	497.26	r_8r_8	437.2	r_2r_2	436.7
750816	396.48	394.41	395.44	r_1TAXCLASS	359.1	r_7TAXCLASS	330.5
585120	375.18	375.36	375.27	r_5r_5	333.0	r_5B	288.3

Table 1.8 Highest Z-Scores for Top 10 Abnormal Records

The top 10 records are investigated in groups by Special Public Use Properties, Government Owned, Hotel and Other categories. Detailed investigations on each property are available on the next page.

Property Investigation

A. Special Public Use Property Abnormalities:

The records below have extremely unusual value given the TAXCLASS they are in. Investigations are done by first finding out how extreme each property has in the most unusual expert variable, and appropriate online research is conducted carefully.

1. *Record: 632816, Owner: 864163 REALTY, LLC, TAXCLASS: 2*
   Address: 86-55 BROADWAY

The assessed land value of the property is $1,318,500 compared with the average value of $127,045 (10.38 times). The building front and depth measurements are extremely small at 1 foot each. Considering the extremely high land value and small building floor area of 1 square foot, this property stood out as an abnormal entry where values across all dimensions are significantly different from average. In particular, its ratio of land value to building area is extremely abnormal. One possible explanation for the high land value is that the property is situated right beside a subway station along the major road Queens Blvd. The building is a mixed-use property where it is split with commercial offices on the ground floor and residential suites above. Thus, it may be difficult for individual residents to input the proper BLDFRONT and BLDDEPTH. However, this property is worth additional investigation for fraud since the location is likely to house a private business - Star Paradise School.

Resource: https://childcarecenter.us/provider_detail/star_paradise_inc_elmhurst_ny

2. *Record: 1067360, Owner: No Owner Information, TAXCLASS: 1*
   Address: 20 EMILY COURT

The actual total land value per lot area of the property seems to be abnormal compared to similar records. Although the assessed total value of $50,160 is not as extremely different from the average value of $24,356, the lot area size is extremely small at 1 square foot, which inflated the expert variable ratio to be extremely high and stood out as an unusual record. Given the Zip, the full value to lot area ratio seems to be abnormal than similar records.

This location is a private dwelling and should be investigated for potential fraud. A screenshot of the property is provided in Appendix A1.1.

Resource: https://www.redfin.com/NY/Staten-Island/20-Emily-Ct-10307/home/56057273

3. *Record: 750816, Owner:M FLAUM, TAXCLASS: 1*
   Address: VLEIGH PLACE

The $836,000 full property value is considerably different than the group average of $242,138 (3.5 times average). Also,its assessed total value is $4179.5, while the mean value is at $14,567 (71% less than mean). With the extremely small lot area of only 1 square foot, it resulted in a significant abnormal ratio of full property value to lot area, as well as assessed total value to lot area. It might be a playground according to a vague Google search. However, this record should be investigated further since playgrounds are typically owned by NYC Parks.

Resource: https://www.yelp.com/biz/vleigh-place-playground-queens
https://www.nycgovparks.org/parks/vleigh-playground/map

B. Government Owned Property Abnormality:

The records below are highly likely to be owned by the government, which could be possibly why they are recognized as abnormal records by the model. Because both records have a significantly higher than average assessed land value, they stood out as extreme abnormal records.

Record #	Owner	Stress Address	Assessed Land Value	Investigation
565392	U S GOVERNMENT OWNED	FLATBUSH AVENUE BR4, ZIP3R4	$1,946,840,000 (7235 times higher than average)	Government Owned
565398	DEPT OF GENERAL SERVICE	FLATBUSH AVENUE BR6,ZIP3R6	$1,039,900,000 (3,865 times higher than average)	Government Owned

Table 1.9 Government Owned Abnormal Property List

After investigation, both of the properties seem to be government owned and share the same first three digit of the zip code, as they also have the same address and have similar owner fields related to the U.S. government. After confirming the lot area of both properties are appropriate, the assessed land value seems to be the main reason for the two records to seem abnormal. It might be that government owned properties tend to have higher than average assessed land value.

C. Hotel as Abnormal Properties:

The two records below turned out to be hotels after investigation. One hotel seems to have a much higher assessed value, while the other one is misrepresenting the building area.

1. *Record: 917942, Owner: LOGAN PROPERTY, INC.*
   Address: 154-68 BROOKVILLE BOULEVARD

Although the building area seems normal, comparing the assessed total value to building area ratio is abnormal. The AVTOT is $4,668,308,947, compared with an overall average of $107,302 (43,506 times higher than mean) in the same zip code area and 20,304 times higher than average value of $229,916. The location is the Holiday Inn JFK hotel. This property has high AVLAND and AVTOT values probably because the property is right beside the JFK airport, which is a major hub for international and domestic flights.

Resource:https://www.ihg.com/holidayinn/hotels/us/en/jamaica/nycka/hoteldetail?cm_mmc=GoogleMaps-_-HI-_-US-_-NYCKA

2. *Record: 585439, Owner: 11-01 43RD AVENUE REA*
   Address: 11-01 43 AVENUE

The full property value of $3,712,000 is higher than the mean value of $ 952,297 (3.9 times), while the assessed total value is $ 1,670,400 (6.2 times higher than mean). The building area of this property is only 1 square foot (with building front and depth at 1 foot), which inflated the ratio of full assessed value to building area as well as the ratio of assessed total value to building area to be extremely high. The location is the Z NYC Hotel. It's not possible for a hotel to have only 1 square feet and thus is worth further investigation.

Resource: https://www.zhotelny.com/

D. Other Abnormal Properties in Borough Level:

The rest of abnormal records are investigated below:

1. *Record: 585118, Owner:NEW YORK CITY ECONOMI*
   Address: 28-10 QUEENS PLAZA SOUTH

Given the building front and depth at 1 foot, the building area is extremely small at 1 square foot. The assessed land value of $1,549,530 is already 29.6 times higher than the mean value at the same borough ($52,413), and 18 times the overall average land value ($85,772). Both numerator and denominator are extremely abnormal, which is why the models identify it as unusual. The location now should be a commercial building according to an online search. It is almost impossible for a luxurious commercial rental property to have a building area of 1 square foot, and thus is worth further investigation.

Resource: https://newyorkyimby.com/category/28-10-queens-plaza-south

2. *Record: 85886, Owner:PARKS AND RECREATION*
   Address: JOE DIMAGGIO HIGHWAY

The building area is much lower than average across the borough (64 sqft compared with average at 394,364 sqft), which might be due to the low building front and depth (8 feet each) with only one story on record. The assessed land value of the property is $31,455,000, which is extremely high compared to the $337,995 borough average (93 times the average value).

Since the owner is Parks and Recreation, the location is likely to be a park (e.g. Riverside Park). Since the park covers a relatively large piece of land in Manhattan, this may be the reason for the high assessed land value. The small building front and depth may be attributed to small public infrastructure built in the park such as public restrooms. Therefore, this property would be considered as less suspicious.

Resource: https://www.loc.gov/resource/hhh.ny2021.photos

3. *Record: 585120, Owner:No Owner Information*
   Address: 28 STREET

The location only has a building length and building depth of 1 foot each, resulting in a way below average building area of 1 square foot compared to borough average at 1,995 square feet. In addition, the property has an assessed value of land at $968,220, which is 18.5 times more than borough average of $52,413, and 11.3 times higher than overall land value at $85,773. Since both numerator and denominator are significantly different, it is reasonable for this record to be identified by the model.

Although the property did not have a specific address, it is suspected that the property should be a rental building or apartment high rise building since it claims to have 20 floors through vague online search. This property requires further investigation.
Resource: https://streeteasy.com/building/41_17-28-street-long_island_city
https://www.tower28lic.com/

Part VIII. Conclusion

In conclusion, through this process the top ten properties that are the most likely to have anomalies were found. Through first doing an exploratory analysis of all of the variables, the most relevant variables were found and the variables that still had missing fields were found. Then, the missing fields were filled in through grouping each variable by different fields, such as ZIP Code, Tax Class, and Building Class, and then filling in the null values with the median value of those groups.

Afterwards, 45 variables were created based on the size and value characteristics. Ratios of the different sizes and values were made and were then normalized based on the average five different categorical groups. These variables were then the basis of the Principal Component Analysis and the algorithms used to create the fraud scores. Before conducting PCA, the variables were first z-scaled to normalize them. Then PCA was used to reduce dimensionality and correlation, and the top eight principal components could be used to explain most of the variance. Finally, two algorithms were used to calculate two different fraud scores. The first was a heuristic function of the z-scores and the second was an autoencoder. Then these two scores were weighted to find a combined score which was then used to find the top ten properties that contained anomalies.

These top ten properties were then investigated, and classified into different types of abnormalities. Some were abnormal due to being many standard deviations away from their tax class average, and some were anomalies potentially because the properties were owned by government entities. The remaining properties were flagged as suspicious, and are left for further analysis in the future.

If there was more time to spend on this report, there would be further investigations made on these top ten properties to see what other properties their owners had and what may be causing this anomaly. Additionally, through this report only the values that were zero were filled in with the median value based on groupings. With more time, other values such as having a 1 for Lot Front and Lot Depth would also be filled in with a different value, as it is not possible for a building to have a Lot Front of Depth of one.

Additionally, further investigation could be done on construction reports and transactions relating to these properties as there can also be anomalies relating to those areas. Properties may change because they have not been constructed yet or still have various transactions that have yet to be completed. Finally, further analysis would be done on the property with the Record number 1067360, as according to Redfin it appears to be a parking lot, but looking at actual images of the property it is very clearly a house (refer to image below).

Appendix

Figure A1.11 Google Image of 20 Emily Court

Data Quality Report is Available on Next Page.

Appendix Report 1-1

USC Marshall
School of Business

DSO 562 Data Quality Report

UNDER THE GUIDANCE OF

Professor Stephen Coggeshall

1/22/2020

JENNY SHANG |
JENNY WANG |
LINGROU WANG |
RORY WANG |

Table of Contents

Part IX. Introduction

The property valuation and assessment data are shared by department of finance of the New York city government with an annual update. The main purpose of this data is to help the city government to calculate property tax, grant eligible properties exemptions and or abatements. The data file is accessed from NYC OpenData online with an assessment year of 2010/11. There are 1,070,994 records and 32 fields in this data file, and a specific data quality report is conducted with details below. The report starts with a summary of numerical and categorical variables, followed with detailed individual analysis, and then concluded with questions for data preparation.

Part X. Summary Tables

There is one variable that is the unique identifier of the property, RECORD. In addition, there is a concatenated variable BBLE that contains a combination of information from the dataset. Details will be discussed in section II. Detailed Variable Visualization starting from page 3.

Table 1.1 - Numerical Variables Summary Table

Variable Name	Min	Mean	Max	Standard Deviation	# of Records with Value	% populated	Records with value of 0
LTFRONT	0.00E+00	3.66E+01	1.00E+04	7.40E+01	1,070,994	100%	169,108
LTDEPTH	0.00E+00	8.89E+01	1.00E+04	7.64E+01	1,070,994	100%	170,128
STORIES	1.00E+00	5.01E+00	1.19E+02	8.37E+00	1,014,730	98%	0 (56,264 Null)
FULLVAL	0.00E+00	8.74E+05	6.15E+09	1.16E+07	1,070,994	100%	13,007
AVLAND	0.00E+00	8.51E+04	2.67E+09	4.06E+06	1,070,994	100%	13,009
AVTOT	0.00E+00	2.27E+05	4.67E+09	6.88E+06	1,070,994	100%	13,007
EXLAND	0.00E+00	3.64E+04	2.67E+09	3.98E+06	1,070,994	100%	491,699
EXTOT	0.00E+00	9.12E+04	4.67E+09	6.51E+06	1,070,994	100%	432,572
BLDFRONT	0.00E+00	2.30E+01	7.58E+03	3.56E+01	1,070,994	100%	228,815
BLDDEPTH	0.00E+00	3.99E+01	9.39E+03	4.27E+01	1,070,994	100%	228,853
AVLAND2	3.00E+00	2.46E+05	2.37E+09	6.18E+06	282,726	26%	0
AVTOT2	3.00E+00	7.14E+05	4.50E+09	1.17E+07	282,732	26%	0
EXLAND2	1.00E+00	3.51E+05	2.37E+09	1.08E+07	87,449	8%	0
EXTOT2	7.00E+00	6.57E+05	4.50E+09	1.61E+07	130,828	12%	0

Table 1.2 – Categorical and Date/Time Variables Summary Table

Variable Name	Variable Type	Most Frequent Value	# of Records with Value	% populated	# of Unique Values	Records with value of 0
BBLE	Categorical	Unique Identifier	1,070,994	100%	1,070,994	0
B	Categorical	4	1,070,994	100%	5	0
BLOCK	Categorical	3944	1,070,994	100%	13,984	0
LOT	Categorical	1	1,070,994	100%	6,366	0
EXCD1	Categorical	1017	638,488	60%	130	0
EASEMENT	Categorical	E	4,636	0%	13	0
OWNER	Categorical	PARKCHESTER PRESERVAT	1,039,249	97%	863,347	0
EXCD2	Categorical	1017	92,948	9%	61	0
BLDGCL	Categorical	R4	1,070,994	100%	200	0
TAX-CLASS	Categorical	1	1,070,994	100%	11	0
EXT	Categorical	G	354,305	33%	3	0
STADDR	Categorical	501 SURF AVENUE	1,070,318	100%	839,281	0
ZIP	Categorical	10314	1,041,104	97%	197	0
EXMPTCL	Categorical	X1	15,579	1%	15	0
VALTYPE	Categorical	AC-TR	1,070,994	100%	1	0
PERIOD	Date / Time	FINAL	1,070,994	100%	1	0
YEAR	Date / Time	2010/11	1,070,994	100%	1	0

Part XI. Detailed Variable Visualization

This section gives an overview of the variable and provides a picture to visualize the variable.

1. **Record** – This is a unique identifier of each property, does not have much visualization value as it does not have any repetitive values or trends.

2. **BBLE** – BBLE records the combined information of boro, block, lot and easement codes. This is a concatenate identifier that included combined geographic information of the property, and also does not have much visualization value because it contains all unique values.

3. **B** – BORO Codes with categories of where the property is located at. Borough, Block, and Lot (also called Borough/Block/Lot or BBL) is the parcel number system used to identify each unit of real estate in New York City for numerous city purposes (see visual in the next page).

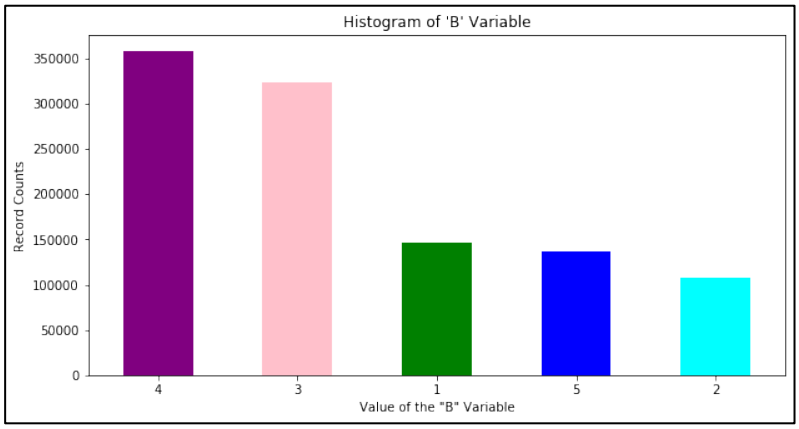


Figure 2.1 Histogram of 'B' Variable

4. **BLOCK** – Identifies the block that property is located at. Here is a table with the top 10 frequent values that covers 2.33% of all 1,070,994 records.

Block Value	Number of Record Counts
3944	3,888
16	3,786
3943	3,424
3938	2,794
1171	2,535
3937	2,275
1833	1,774
2450	1,651
1047	1,480
7279	1,302

Table 2.1 Top 10 Block Value List

5. **LOT** – Identifies the lot that property is located at, which is a type of index recording unique number of lot. Here is a table with the top 10 frequent values that covers 12.4% of all 1,070,994 records.

LOT Values	Number of Record Counts
1	24,367
20	12,294
15	12,171
12	12,143
14	12,074
16	12,042
17	11,982
18	11,979
25	11,949
21	11,840

Table 2.2 Top 10 LOT Value List

6. **EASEMENT** – Indicates whether or not this field is used to describe some public properties like street and railroads. Only 0.43% of the whole dataset has values in this field. The table below shows the distribution of 12 non-NA values of this variable, and y-axis is the log of counts of records.

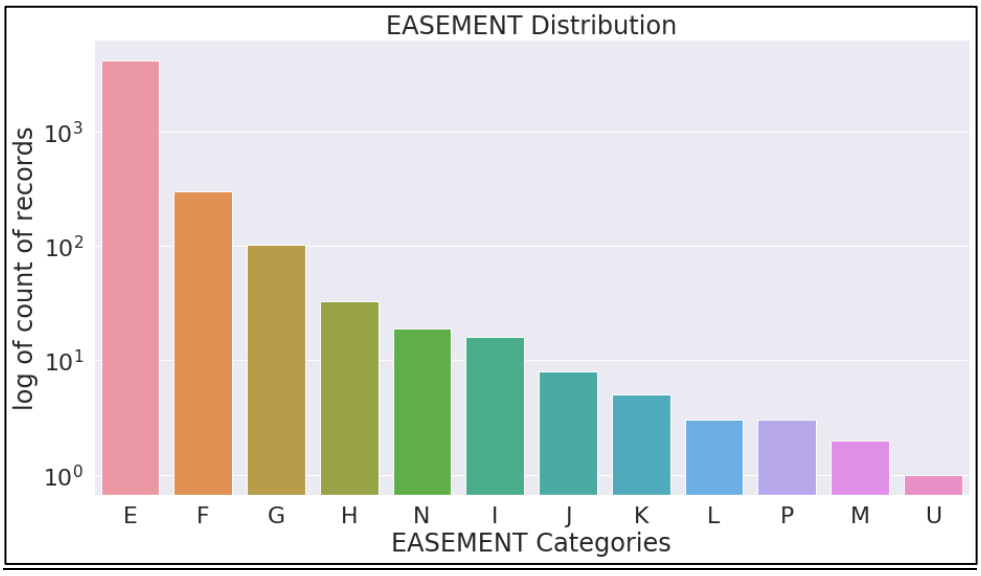


Figure 2.2 EASEMENT Distribution Graph

7. **OWNER** – Denotes the owner's name of the property. Here is a table with the top 10 frequent values that covers 2.0% of all 1,039,249 records.

OWNER Name	Count of Records
PARKCHESTER PRESERVAT	6020
PARKS AND RECREATION	4255
DCAS	2169
HOUSING PRESERVATION	1904
CITY OF NEW YORK	1450
DEPT OF ENVIRONMENTAL	1166
BOARD OF EDUCATION	1015
NEW YORK CITY HOUSING	1014
CNY/NYCTA	975
NYC HOUSING PARTNERSH	747

Table 2.3 Top 10 Owner Name List

8. **BLDGCL** – Building class of the property. There is a direct correlation between the Building Class and the Tax Class per data documentation from data source. Here is a table with the top 10 frequent values that covers 70.8% of all 1,070,994 records.

BLDGCL Values	Count of Records
R4	139,879
A1	123,369
A5	96,984
B1	84,208
B2	77,598
C0	73,111
B3	59,240
A2	51,130
A9	26,177
B9	26,133

Table 2.4 Top 10 BLDGCL Values List

9. **TAX-CLASS** – Current Property Tax Class Code (NYS Classification) of the property. Here is a table with the distributions of each categories with record counts in relation to percentage of the whole dataset.

TAXCLASS	Count of Records	% of Total Records
1	660,721	61.7%
2	188,612	17.6%
4	104,310	9.7%
2A	40,574	3.8%
1B	24,738	2.3%
1A	21,667	2.0%
2B	13,964	1.3%
2C	10,795	1.0%
3	4,638	0.4%
1C	946	0.1%
1D	29	0.0%

Table 2.4 TAXCLASS Values Distribution Overview

10. **LTFRONT** – Lot frontage in feet of the property. Below is a graph showing its distribution excluding 10,033 records (0.94% of total data points) with LTDEPTH greater than 250.

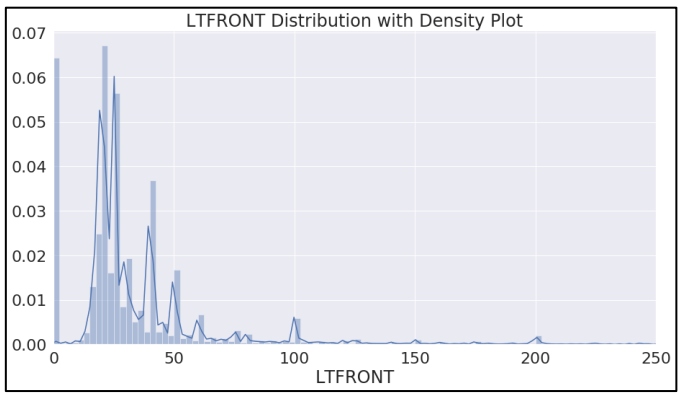


Figure 2.3 LTFRONT Distribution Graph

11. **LTDEPTH** – Lot depth in feet of the property. Below is a graph showing its distribution with density plot excluding 18,784 records (1.75% of total data points) with LTDEPTH greater than 200.

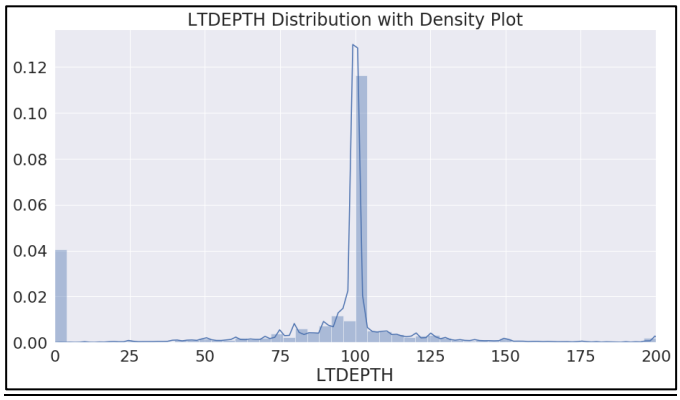


Figure 2.4 LTDEPTH Distribution Graph

12. **EXT** – Denotes if the property has extension or garage. Here is a table with the distributions of each categories with record counts in relation to percentage of the whole dataset.

EXT Values	Count of Records	% of EXT Records
G	266,970	75.4%
E	49,442	14.0%
EG	37,893	10.7%

Table 2.5 EXT Values Distribution Overview

13. **STORIES** – Indicates the number of stories for the property (# of Floors). Below is a graph of variable value distribution records with STORIES greater than 50.

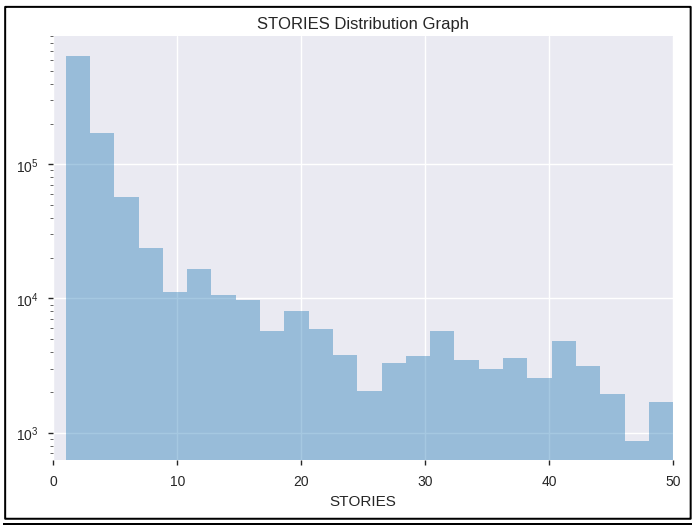


Figure 2.5 STORIES Distribution Graph

14. **FULLVAL** – If not zero, current year's total market value of the land. Here is a table of variable value distribution excluding 39,496 records (3.69% of all data points) with FULLVAL over 2,000,000. It's also noting that the variable has 13,007 records (1.21% of all data points) with a value of 0. Since the variable description specifically mentioned "if not zero", it would be critical to clarify what does zero represent.

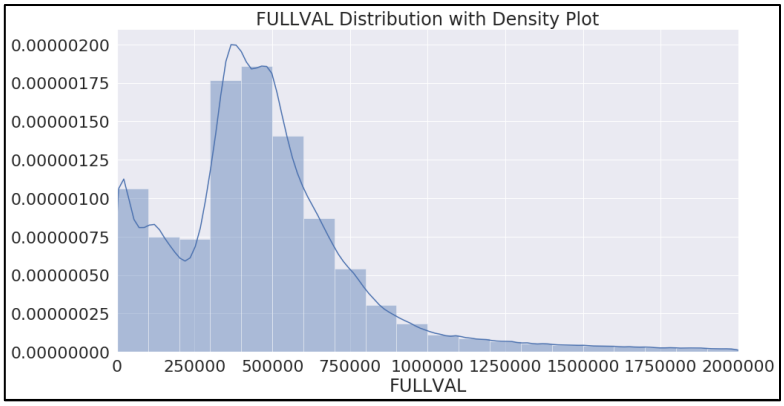


Figure 2.6 FULLVAL Distribution Graph

15. **AVLAND** – Actual land Value of the property of the assessment period. Here is a table of variable value distribution excluding 101,453 records (9.47% of all data points) with AVLAND over 50,000. It's also noting that the variable has 13,009 records (1.21% of all data points) that is at a value of 0. It will be essential to find out if values of zero AVLAND is truly zero or if it's missing data.

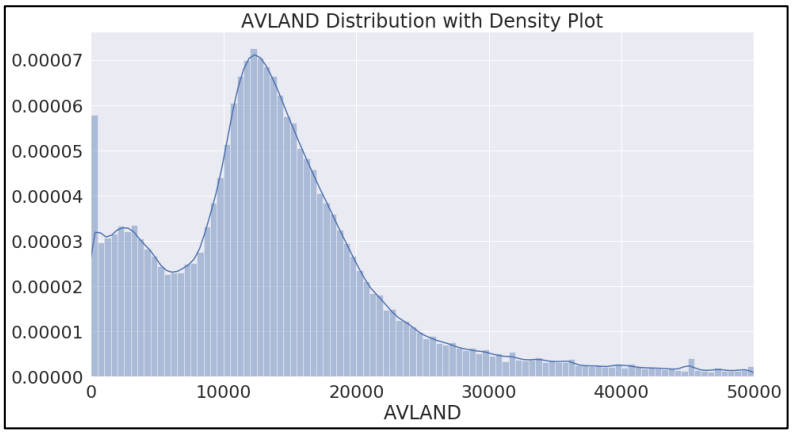


Figure 2.7 AVLAND Distribution Graph

16. **AVTOT** – Actual total Land Value of the property of the assessment period. Here is a table of variable value distribution excluding 149,378 records (13.94% of all data points) that has an AVTOT over 100,000. It's also noting that the variable has 13,007 records (1.21% of all data points) with value of 0. It will be essential to find out if values of zero AVTOT is truly zero or if it's missing data.

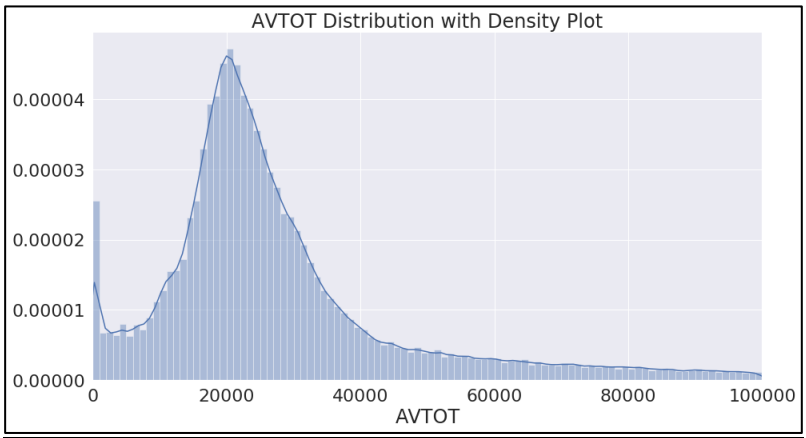


Figure 2.8 AVTOT Distribution Graph

17. **EXLAND** – Actual Exempt Land Value of the property of the assessment period. Here is a graph of variable value distribution excluding 25,281 records (2.36% of all data points) that has an EXLAND over 30,000. It's also worth noting that the variable has 491,699 records (45.91% of all data points) with value of 0. It will be essential to find out if values of zero EXLAND is truly zero or if it's missing data.

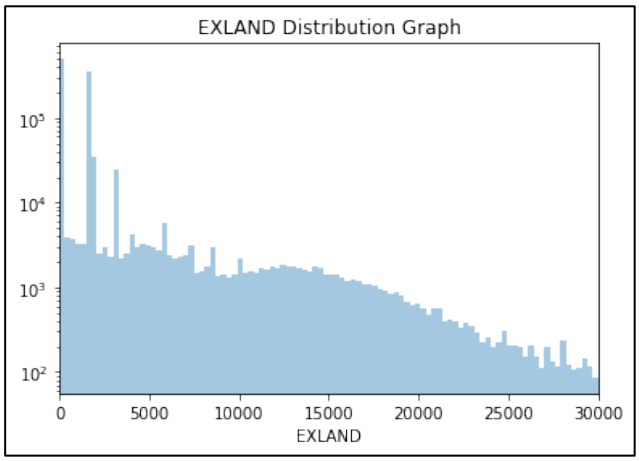


Figure 2.9 EXLAND Distribution Graph

18. **EXTOT** – Actual exempt land total value of the property of the assessment period.
    a. Here is a graph of variable value distribution excluding 82,579 records (7.71% of all data points) that has an EXTOT over 30,000. The y-axis has been scaled to log of count of records to show trend.
    b. It's also noting that the variable has 432,572 records (40.39% of all data points) with value of 0. It will be essential to find out if values of zero EXTOT is truly zero or if it's missing data.

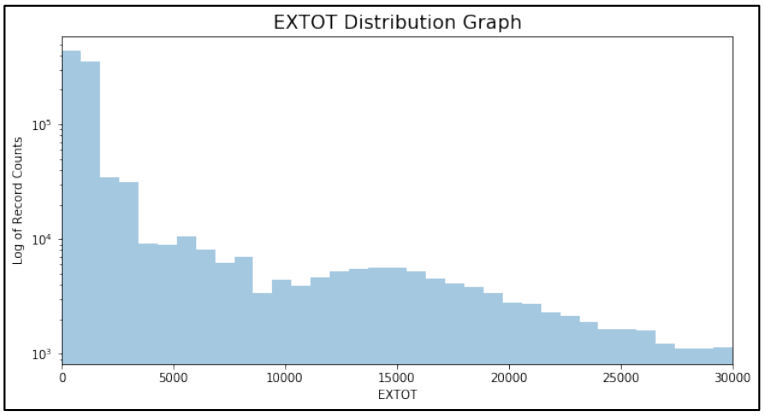


Figure 2.10 EXTOT Distribution Graph

19. **EXCD1** – Exemption code 1 of the property. Here is table with the distributions of top 10 categories with record counts in relation to percentage of the EXCD1 dataset.

EXCD1 Values	Count of Records	% of Total Valid Records
1017	425,348	66.6%
1010	49,756	7.8%
1015	31,323	4.9%
5113	23,858	3.7%
1920	17,594	2.8%
5110	16,834	2.6%
5114	14,984	2.3%
5111	10,609	1.7%
1021	6,613	1.0%
1986	4,231	0.7%

Table 2.6 EXCD1 Values Distribution Overview

20. **STADDR** – Street address of the property. Here is table with the distributions of top 10 STADDR values with record counts in relation to percentage of the STADDR dataset, which is reasonable to see the top 10 most frequent values only cover 0.63% of total valid records with a STADDR value.

STADDR Value	Count of Records	% of Total Valid Records
501 SURF AVENUE	902	0.08%
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 RIVING PLACE	650	0.06%
220 RIVERSIDE BOULEVARD	628	0.06%
360 FURMAN STREET	599	0.06%
200 EAST 66 STREET	585	0.05%
30 WEST 63 STREET	562	0.05%

Table 2.7 STADDR Values Distribution Overview

21. **ZIP** – Zip code of the property. Here is table with the distributions of top 10 ZIP values with record counts in relation to percentage of the ZIP dataset.

ZIP Value	Count of Records	% of Total Valid Records
10314	24,606	2.4%
11234	20,001	1.9%
10312	18,127	1.7%
10462	16,905	1.6%
10306	16,578	1.6%
11236	15,678	1.5%
11385	14,921	1.4%
11229	12,793	1.2%
11211	12,710	1.2%
11207	12,293	1.2%

Table 2.7 ZIP Values Distribution Overview

22. **EXMPTCL** – The exemption class of the property. Here is table with the distributions of each categories with record counts in relation to percentage of the EXMPTCL dataset. It's worth noticing that this distribution only consists of 1% of the total data points.

EXMPTCL Value	Count of Records	% of Total Valid Records
X1	6,912	44.4%
X5	5,208	33.4%
X7	820	5.3%
X2	770	4.9%
X6	764	4.9%
X4	441	2.8%
X8	292	1.9%
X3	259	1.7%
X9	108	0.7%
R4	1	0.0%
5	1	0.0%
KI	1	0.0%
VI	1	0.0%
A9	1	0.0%

Table 2.8 EXMPTCL Values Distribution Overview

23. **BLDFRONT** – Building width of the property. Here is a graph of variable value distribution excluding 19,807 records (1.85% of all data points) that has an BLDFRONT over 125. It's also noting that the variable has 228,815 records (21.36% of all data points) with value of 0. It will be essential to find out if values of zero AVTOT is truly zero or if it's missing data.

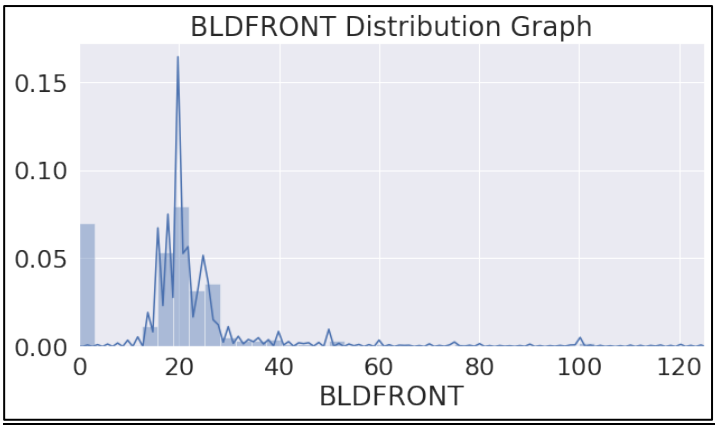


Figure 2.11 BLDFRONT Distribution Graph

24. **BLDDEPTH** – Building depth of the property. Here is a graph of variable value distribution excluding 2,950 records (0.28% of all data points) that has an BLDDEPTH over 125. It's also noting that the variable has 228,853 records (21.36% of all data points) with value of 0. It will be essential to find out if values of zero AVTOT is truly zero or if it's missing data.

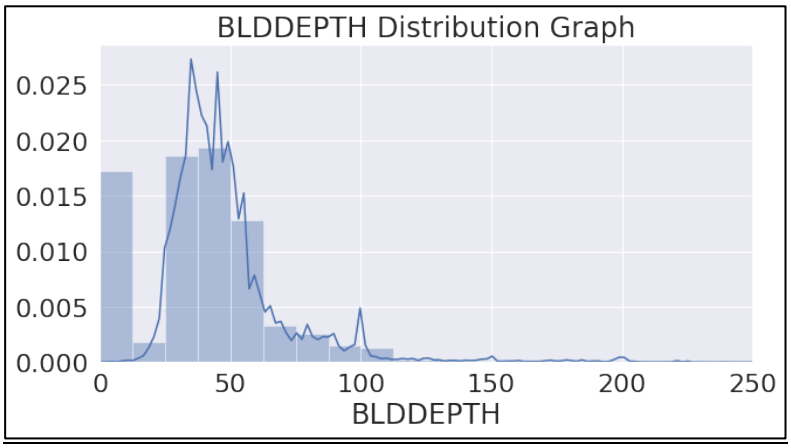


Figure 2.12 BLDDEPTH Distribution Graph

25. **AVLAND2** – Transitional land value of the property. Here is a graph of variable value distribution excluding 9,151 records (3.24% of all data points) that has an AVLAND2 over 100,000.

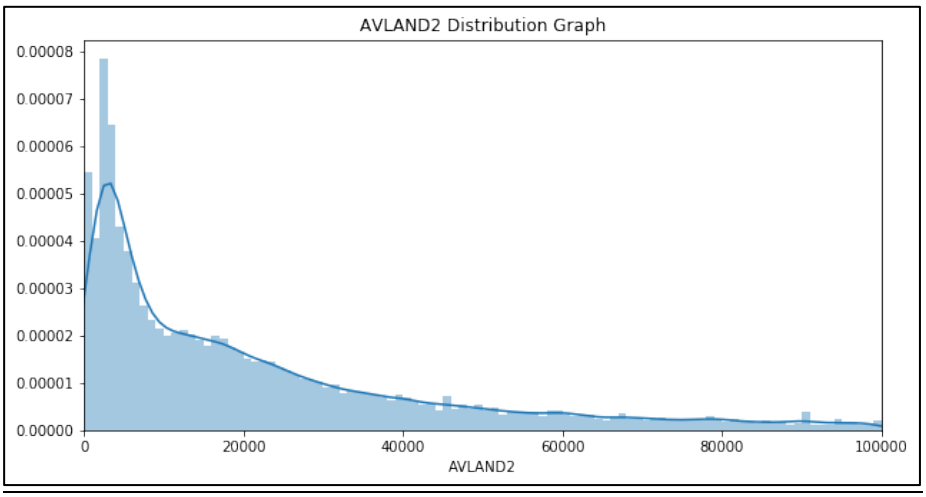


Figure 2.12 AVLAND2 Distribution Graph

26. **AVTOT2** – Transitional land value of the property. Here is a graph of variable value distribution excluding 13,306 records (4.71% of all data points) that has an AVTOT2 over 2,000,000, with count of records in a log scale.

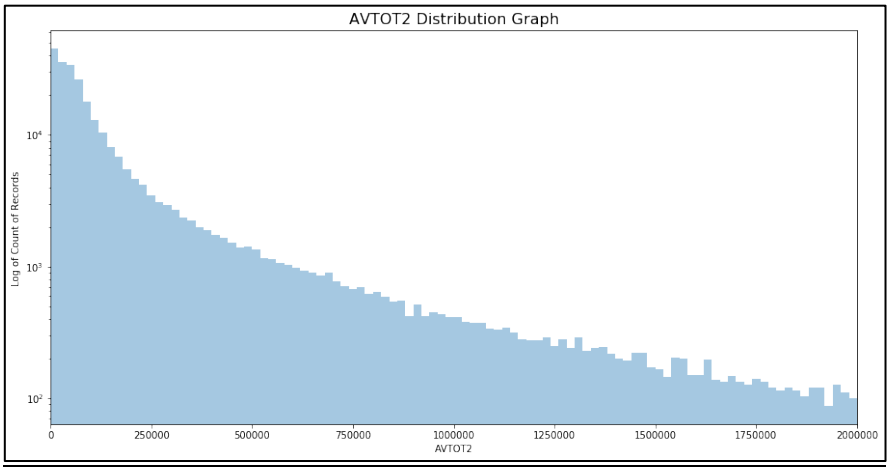


Figure 2.13 AVTOT2 Distribution Graph

27. **EXLAND2** – Transitional exemption land value of the property. Here is a graph of variable value distribution excluding 9,909 records (7.57% of all data points) that has an EXLAND2 over 200,000, with the y-axis of log of scale of record counts.

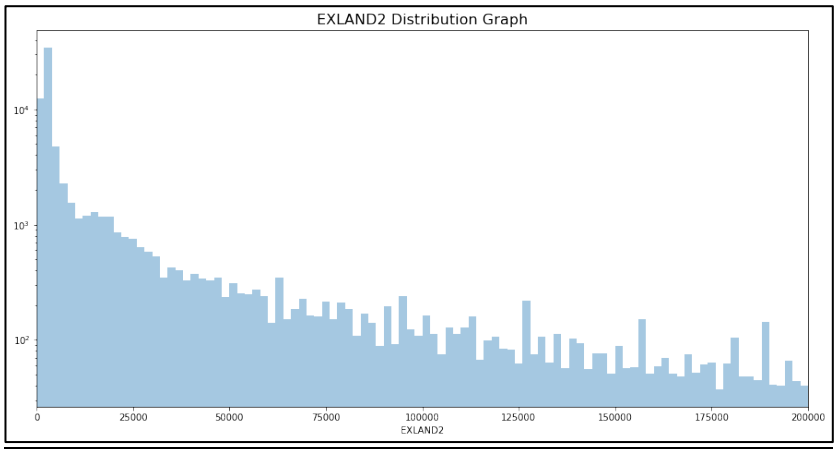


Figure 2.14 EXLAND2 Distribution Graph

28. **EXTOT2** – Transitional exemption land total value of the property. Here is a graph of variable value distribution excluding 17,459 records (13.35% of all data points) that has an EXTOT2 over 300,000, with the y-axis of log of scale of record counts.

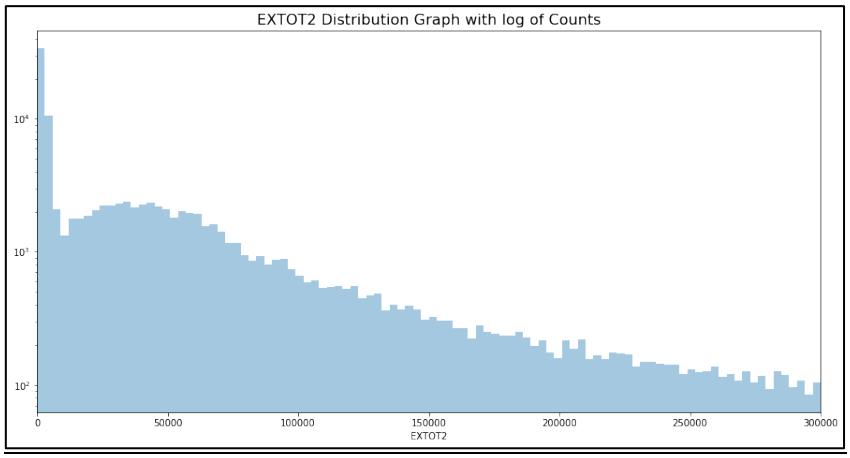


Figure 2.15 EXTOT2 Distribution Graph

29. **EXCD2** – The exemption code 2 of the property. Exemption code 1 of the property. Here is table with the distributions of top 10 categories with record counts in relation to percentage of the EXCD2 dataset.

EXCD2 Values	Count of Records	% of Total Valid Records
1017	65777	70.8%
1015	12337	13.3%
5112	6867	7.4%
1019	3178	3.4%
1920	2961	3.2%
1200	881	0.9%
1101	494	0.5%
5129	227	0.2%
1986	35	0.0%
1022	31	0.0%

Table 2.9 EXCD2 Values Distribution Overview

30. **PERIOD** – The period of record of this dataset. All 1,070,994 rows have one single value of "FINAL".

31. **YEAR** – The year of record of this dataset. All 1,070,994 rows have one single value of "2010/11".

32. **VALTYPE** – The dataset documentation did not have clear definition but might be the valuation type of the property. All 1,070,994 rows have one single value of "AC-TR".

Part XII. Questions for Further Clarification:

There are several questions that need clarification from either data source representative or subject matter expert:

1. **Value of Zeros:**
    a. For variables like AVLAND and AVTOT, the meaning of zero needs to be clarified on whether zero means a value of zero or missing data.
2. **Description of Variables:**
    a. According to the documentation provided, variables like AVTOT and VALTYPE are not clarified. In order to prepare data for efficient analysis, it is critical to understand the business problem and the variables that are available for analysis.
3. **Values for EXCD1 and EXCD2:**
    a. The most frequent value of both EXCD1 and EXCD2 are both 1017. This could be a possible area for clarification on the two variables and might provide ideas for database collection process improvement in the future.
4. **Version of Data / Documentation:**
    a. Per most recent version of the Property Valuation and Assessment Data on NYC Open Data (2019), the documentation included more variable explanations that was not found in original documentation provided by Professor Coggeshall that needs to confirm with subject matter expert or data source owner.
    b. Some data explanations like AVTOT used the documentation with the latest version, while variable VALTYPE still did not have any updated description and was described to the author's best knowledge.

Part XIII. Reference

Property Valuation and Assessment Data: Nyc Open Data
Department of Finance - https://data.cityofnewyork.us/City-Government/Property-Valuation-and-Assessment-Data/yjxr-fw8i. Accessed on 1/21/2019.