USC Marshall
School of Business

DSO 562 Fraud Analytics Project 1
Find Anomalies in the New York Property Data

Team 3
Session: 16237

Boyang Han, Leon Man, Ziqing (Juno) Wen, Dakota Wu, Chutong Yan, Yangzi Zhang,
February 13, 2020

TABLE OF CONTENTS

Executive Summary

This project is commissioned to demonstrate the process of using unsupervised machine learning techniques to detect anomalies that could potentially be property tax fraud. Using a public New York Property dataset acquired from the *NYC OpenData* website, the team developed two anomaly detection algorithms: 1) heuristic functions of z-scores and 2) autoencoder. The output from the two algorithms was combined to calculate the final anomaly scores. All records were sorted by the final scores. The 10 records listed below, having the largest final scores, were identified as the most anomalous records. The team then investigated why these records were flagged and whether they were false alarms or suspicious cases that were worthy of further inspection.

The top 10 anomalies sorted by extremeness are:
False alarms: record 632816, 85886, 67129, 585118, 245573, and 585439
Further investigation required: record 565392, 1067360, 917942, and 821853

1. Description of Data

The New York property dataset (“NY property data.csv”) stores property valuations collected annually by the Department of Finance of the city and is accessible to the public through the *NYC Open Data* website. It covers properties assessed in the year of 2010 to calculate property tax, grant eligible properties exemptions, and (or) abatements. The dataset contains 1,070,994 records and 32 fields.

1.1 Summary Tables

There are 14 numerical variables and 18 categorical variables. Their summary statistics are provided in Table 1.1.1 and 1.1.2 below.

Table 1.1.1: Categorical Fields

Field	%Populated	Unique Values	Most Common
RECORD	100	1070994	N/A
BBLE	100	1070994	N/A
B	100	5	4
BLOCK	100	13984	3944
LOT	100	6366	1
EASEMENT	43.29	13	E
OWNER	97.04	853347	PARKCHESTER PRESERVAT
BLDGCL	100	200	R4
TAXCLASS	100	11	1
EXT	33.08	4	G
STADDR	99,94	839281	501 SURF AVENUE
ZIP	97.21	197	10314.0
EXMPTCL	1.45	15	X1
PERIOD	100	1	Final
YEAR	100	1	2010/11
VALTYPE	100	1	AC-TR
EXCD1	12.22	48349	1017.0
EXCD2	8.68	61	1017.0

Table 1.1.2: Numerical Fields

Field	%Populated	Unique Values	No. 0s	SD	Mean
LTFRONT	100	1297	169108	**7.40E+01**	**3.66E+01**
LTDEPTH	100	1370	170128	**7.64E+01**	8.89E+01
STORIES	94.75	112	0	**8.37E+00**	5.01E+00
FULLVAL	100	109324	13007	**1.16E+07**	8.74E+05
AVLAND	100	70921	13009	**4.06E+06**	8.51E+04
AVTOT	100	112914	13007	**6.88E+06**	2.27E+05
EXLAND	97.21	33419	491699	**3.98E+06**	3.64E+04
EXTOT	100	64255	432572	**6.51E+06**	9.12E+04
EXCD1	59.62	130	0	**1.38E+03**	1.60E+03
BLDFRONT	100	612	228815	**3.56E+01**	2.30E+01
BLDDEPTH	100	621	228853	**4.27E+01**	4.00E+01
AVLAND2	26.40	58592	0	**6.18E+06**	2.46E+05
AVTOT2	26.40	111361	0	**1.17E+07**	7.14E+05
EXLAND2	8.17	22196	0	**1.08E+07**	3.51E+05

Among all the 32 fields, the team focused on the following 12 variables for the reasons specified below:

- *FULLVAL, AVLAND,* and *AVTOT*: represent property values, used to calculate property tax
- *LTFRONT, LTDEPTH, BLDFRONT,* and *BLDDEPTH*: represent property sizes, strongly related to property values
- *B, BLOCK,* and *ZIP*: represent location information, highly related to property values
- *BLDCLASS* and *TAXCLASS*: represent the property type, significantly affect tax calculation

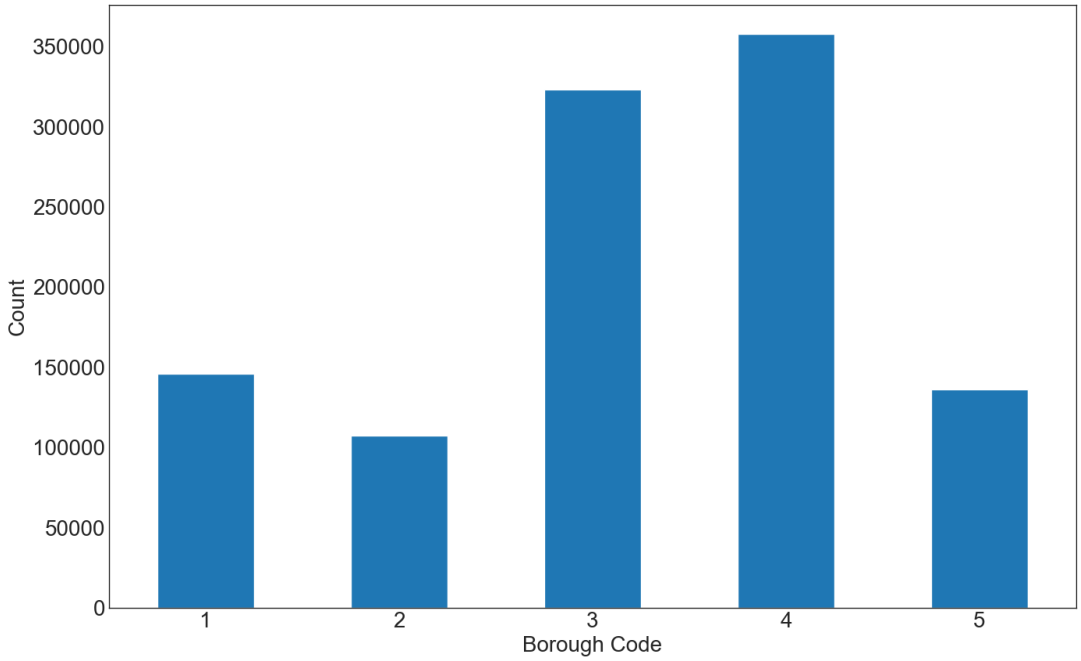
1.2 Distribution of Important Fields

- *B*: *B* is a categorical variable that represents five boroughs in New York. It ranges from 1 to 5 and no missing value is found in this field: 1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island (shown in Table 1.2.1). The chart (Figure 1.2.2) below shows the count for each value.

Table 1.2.1: Borough Code

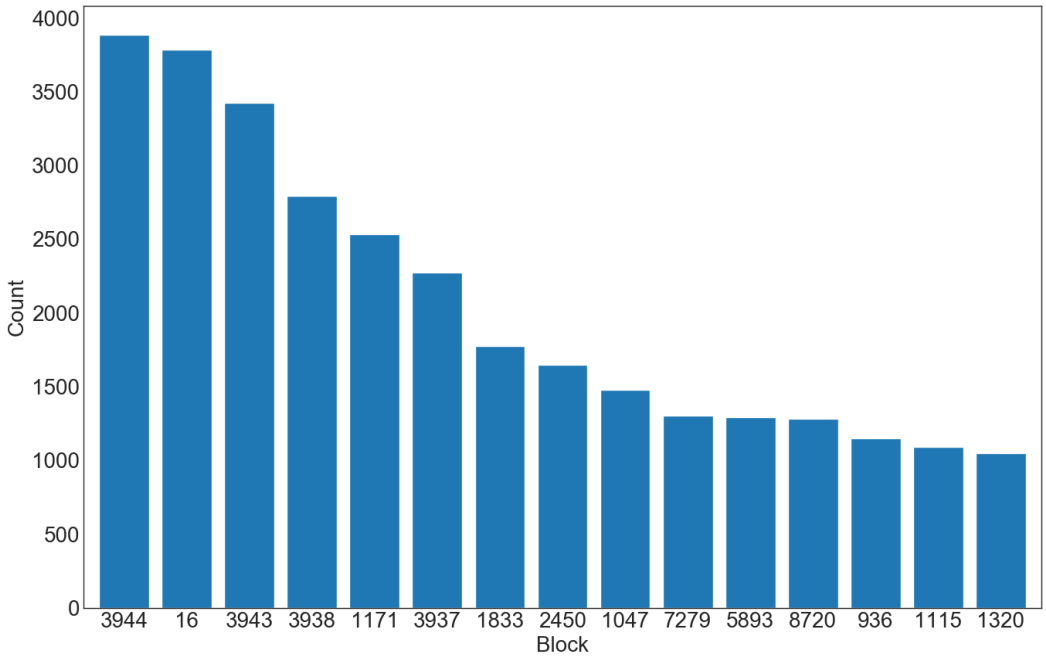
Code	Borough
1	Manhattan
2	Bronx
3	Brooklyn
4	Queens
5	Staten Island

Figure 1.2.2: Distribution of *B* Field



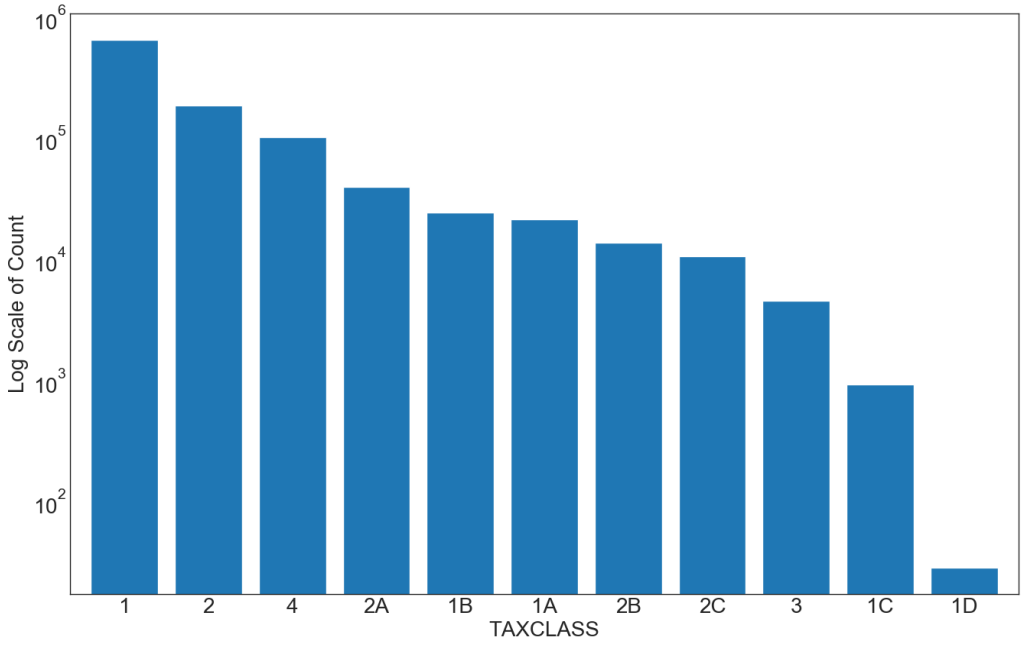
- ***BLOCK*: *BLOCK*** is a categorical variable showing the block number. The chart below shows the count for the 15 most common *BLOCK* values.

Figure 1.2.2: Distribution of *BLOCK* Field (15 Most Common Values)



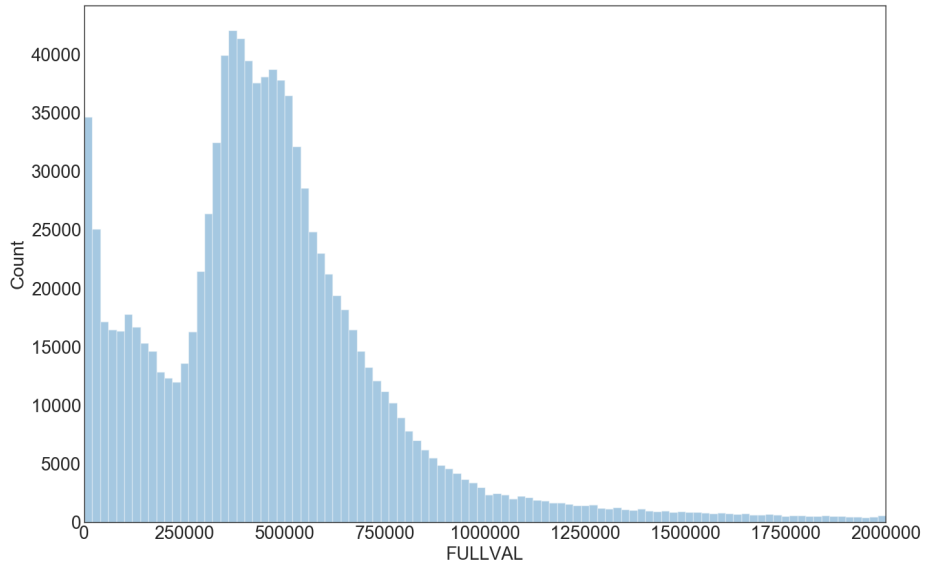
- ***TAXCLASS***: *TAXCLASS* is a categorical variable, which represents the property tax class code. It is directly correlated with *BLDGCL* (Building Class). The chart below shows the count of each value.

Figure 1.2.3: Distribution of *TAXCLASS* Field



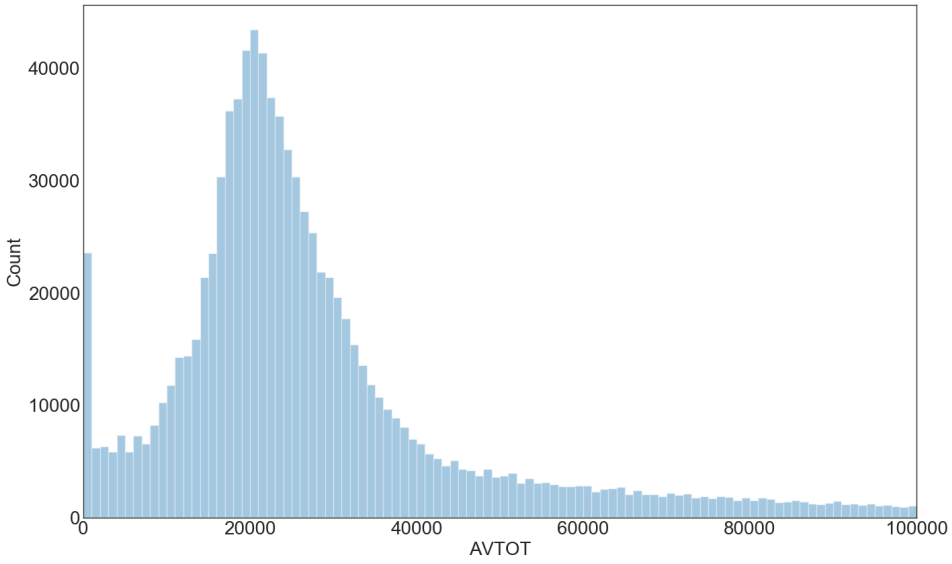
- ***FULLVAL****: FULLVAL* is a numerical variable that represents the total market value of a property. There are no missing values in this field. The following chart is the log scaled distribution of *FULLVAL* excluding values over 2 million.

Figure 1.2.4: Distribution of *FULLVAL* Field Excluding Values Over 2 Million



- ***AVTOT:*** *AVTOT* is a numerical variable that represents the actual assessed total value of a property. Typically, *AVTOT* is less than *FULLVAL* for the same property. The log scaled distribution of *AVTOT* excluding values over 1 million is shown below.

Figure 1.2.5: Distribution of *AVTOT* Field Excluding Values Over 1 Million



2. Data Cleaning

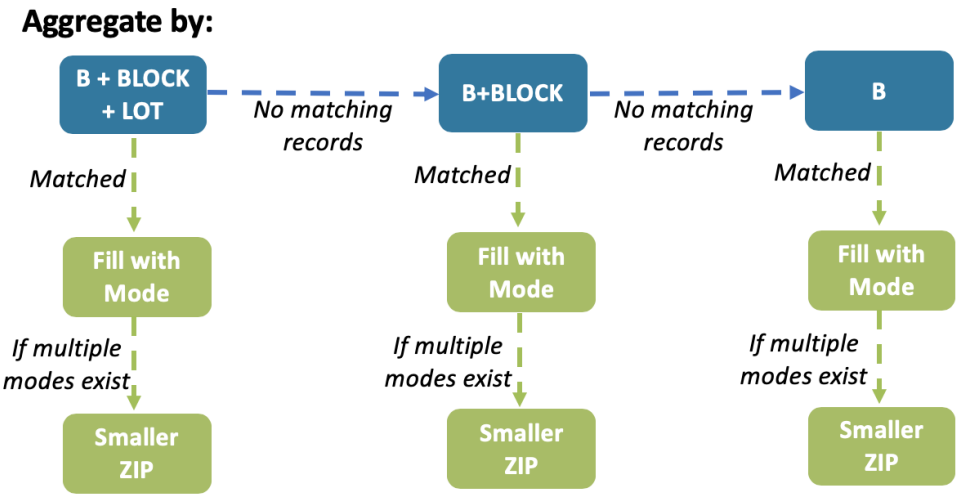
Among the key variables used for further analysis, *STORIES* and *ZIP* are not fully populated. Besides, *LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, FULLVAL, AVLAND,* and *AVTOT* have unreasonable zero values. Therefore, two major steps were taken to fill the missing values and unreasonable zeros.

2.1 Step 1

Fill in missing values in the *ZIP* field with the mode of the geographical subgroup:

The chart below demonstrates the process of Step 1.

Figure 2.1.1: Flowchart of Filling in *ZIP* Field Missing Values



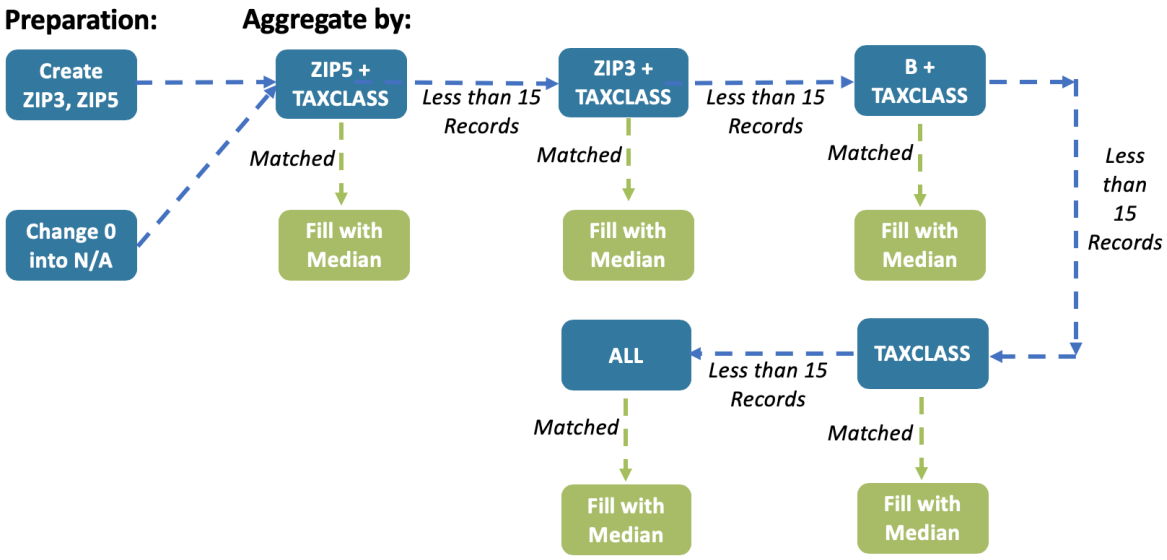
1. Aggregate the dataset by *B, BLOCK* and *LOT*, use the most frequent *ZIP* value of that group. If there are multiple modes in the group, use the *ZIP* value that is smaller in number.
2. If there is no *ZIP* value in the *B, BLOCK* and *LOT* group, aggregate the dataset by *B* and *BLOCK*, and use the most frequent *ZIP* value in the group. If there are multiple modes in the group, use the *ZIP* value that is smaller in number.
3. If there is no *ZIP* value in the *B* and *BLOCK* group, aggregate the dataset by *B* and use the most frequent *ZIP* value in the group. If there are multiple modes in the group, use the *ZIP* value that is smaller in number.

2.2 Step 2

Fill in missing values/unreasonable zeros in the *LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, FULLVAL, AVLAND, AVTOT,* and *STORIES* fields with the median of the subgroup:

The chart below demonstrates the process of Step 2.

Figure 2.2.1: Flowchart of Filling in Key Numerical Fields Missing Values



1. Create a new variable *ZIP3* by extracting the first 3 digits of the *ZIP* field. Rename the *ZIP* column as *ZIP5*.
2. Transform zero values of all columns to N/A.
3. Aggregate the dataset by *ZIP5* and *TAXCLASS*. If there are more than 15 records in the group, use the median value of the group.
4. Otherwise, aggregate the dataset by *ZIP3* and *TAXCLASS*, if there are more than 15 records in the group, use the median value of that group.
5. Otherwise, aggregate the dataset by *B* and *TAXCLASS*, if there are more than 15 records in the group, use the median value of that group.
6. Otherwise, aggregate the dataset by *TAXCLASS*, if there are more than 15 records in the group, use the median value of that group.
7. Otherwise, use the median of all values in that field.

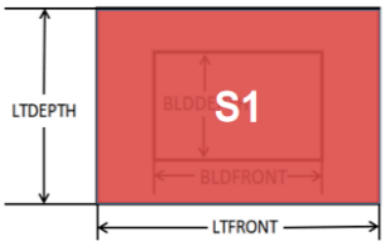
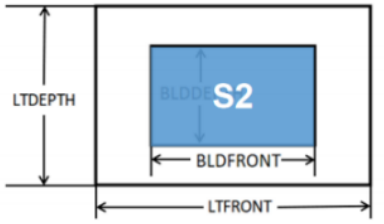
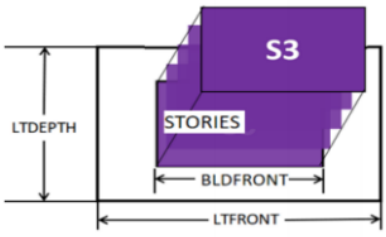
3. Variable Creation

Based on the exploratory analysis and careful consultation with experts, 45 expert variables were created. The logic and detailed steps are as follows.

3.1 Size Variables

When comparing properties across different types, value per square foot is a more reasonable metric than absolute total value. Therefore, the team calculated unit values by dividing property value by property size. Variables that involve values include *FULLVAL (renamed as V_1), AVLAND (renamed as V_2),* and *AVTOT (renamed as V_3)*. Three types of size variables were created using the formulas shown in Table 3.1.

Table 3.1.1: Size Variables

Graph Illustration	Formula
	$S_1 = LTFRONT\ x\ LTDEPTH$
	$S_2 = BLDFRONT\ x\ BLDDEPTH$
	$S_3 = S_2\ x\ STORIES$

Without further information on how each of the value variables was measured, it was hard to select the corresponding size and calculate the unit value. Therefore, all three value fields were divided by each of the size variables and nine new variables were created.

3.2 Nine Variables

Table 3.2.1: Nine Variables Creation

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

Locations and property types have a significant impact on property values, so the team believes that the value per square foot also varies across locations and tax classes. For example, $2,500 per square foot is extremely expensive for a property located in Inwood, Manhattan, but it is a totally normal unit price for properties in West Chelsea, Manhattan. A record should be marked as strange only if it is not within a reasonable range compared to the properties in that area. Similarly, properties in the same tax class should have comparable unit values.

3.3 45 Variables

To better assess how anomalous a property is, records were aggregated by three geographical subgroups, *ZIP5, ZIP3, and B.* In addition, records were also compared with their counterparts in the same *TAXCLASS* and in the whole New York City. After each of the five groupings, all unit values were transformed by dividing its original value by the group median, which measures how far a record is from the center of the group. A total of 45 variables were created (shown in Table 3.3).

Table 3.3.1: 45 Variables Creation

Key Variables				
Group by *ZIP5*	Group by *ZIP3*	Group by *TAXCLASS*	Group by *B*	Group by all
r1_ZIP5	*r1_ZIP3*	*r1_TAXCLASS*	*r1_B*	*r1_all*
r2_ZIP5	*r2_ZIP3*	*r2_TAXCLASS*	*r2_B*	*r2_all*
r3_ZIP5	*r3_ZIP3*	*r3_TAXCLASS*	*r3_B*	*r3_all*
r4_ZIP5	*r4_ZIP3*	*r4_TAXCLASS*	*r4_B*	*r4_all*
r5_ZIP5	*r5_ZIP3*	*r5_TAXCLASS*	*r5_B*	*r5_all*
r6_ZIP5	*r6_ZIP3*	*r6_TAXCLASS*	*r6_B*	*r6_all*
r7_ZIP5	*r7_ZIP3*	*r7_TAXCLASS*	*r7_B*	*r7_all*
r8_ZIP5	*r8_ZIP3*	*r8_TAXCLASS*	*r8_B*	*r8_all*
r9_ZIP5	*r9_ZIP3*	*r9_TAXCLASS*	*r9_B*	*r9_all*

4. Scaling and Dimensionality Reduction

4.1 Scaling Before Dimensionality Reduction

Why

The whole idea of detecting an anomaly, and from there identifying potential fraud, is based on calculating the distance of each observation from the center of the sample and catching the ones that are extremely far away. However, the features inherently have different scales. For example, a 100-unit difference in the *BLDDEPTH* variable, measured in feet, should be considered large given that the standard deviation of that variable is only 42.7. However, the same unit difference in the *FULLVAL* variable, measured in dollars, is relatively small since the standard deviation is above 11 million. To avoid bias towards dimensions of large units, all dimensions should be adjusted to the same footing. Therefore, all 45 features created in the last step were properly scaled before fitting into a model.

How

The z-scaling method was adopted to standardize the variables. Z-scores are calculated by subtracting the sample mean from each raw record and then dividing the difference by the sample standard deviation (see Formula 4.1.1). After the conversion, all dimensions are now measured in the unit of standard deviation and a record's distances to the center on different dimensions are readily comparable.

Formula 4.1.1: Z-scale

$$z_i = \frac{x_i - \mu_i}{\sigma_i}$$

4.2 Dimensionality Reduction

Why

When operating data in high-dimensional space, various phenomena have been observed that make machine learning particularly challenging. Such phenomena are usually referred to as the curse of dimensionality and needed to be addressed before building a model.

As dimensionality increases, data becomes sparse very quickly and all points become outliers. Especially in the anomaly detection realm, algorithms often rely on detecting areas where records share similar properties and catching those that fall out of clusters. However, in high dimensional space, all objects appear to be dissimilar, or in other words, closer to the edge rather than to the center. This makes it harder for models to identify the true outliers. And to make things more complicated, as dimensionality increases, so does noise. Data needed to tell true relationship from noise grows exponentially. In addition to the curse of dimensionality, when a data set has many variables, many of them tend to be correlated, meaning that the information they hold has great overlaps, and hence some are redundant. Besides, computing large data sets with high dimensions takes time.

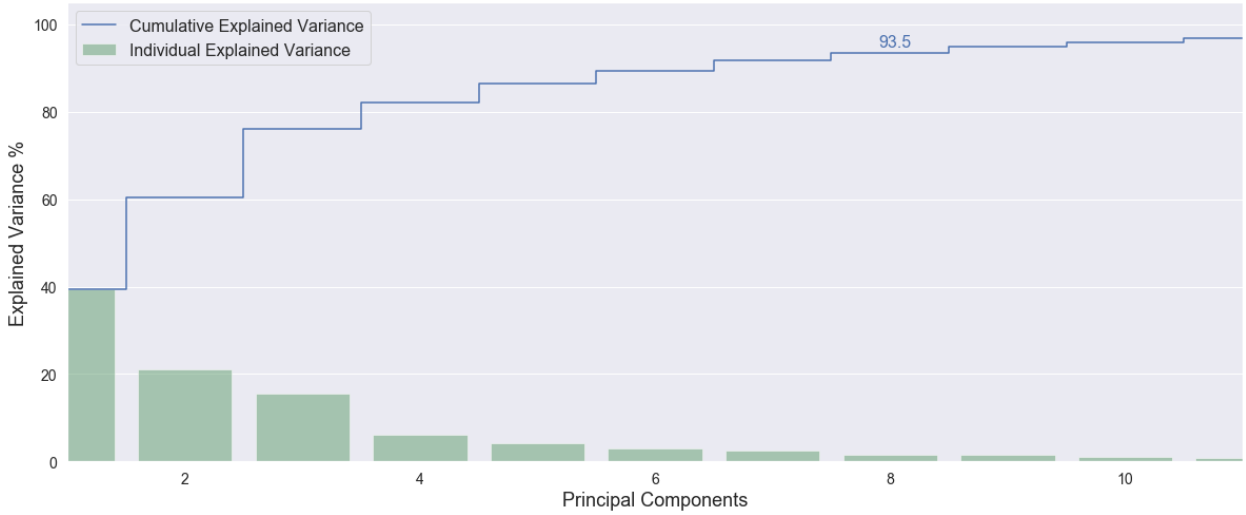
For those reasons, dimensionality reduction is a necessary step in machine learning that saves computation time and helps build a model that captures as much of the real pattern and as little of the noise in data as possible. It reduces the features under consideration by extracting a set of principal variables that hold the most substantial information.

How

As one of the most widely used dimensionality reduction methods in the machine learning community, Principal Component Analysis (PCA) was used to reduce the number of features in this project. PCA rotates and reconstructs the coordinate system repetitively to find low-dimensional representations of the data set that capture as much of the variation in data as possible. Principal Components (PCs) extracted by this process are ordered by the variance in their own dimension. To maximize variance, PCs are always orthogonal to each other, hence the linear correlations are completely removed.

The percentages of variance explained by each PC are plotted in Figure 4.2 below. The team decided to keep eight PCs for that they all together explained 93.5% – a fairly good amount of the variance in the original data. The remaining higher order PCs were discarded. The original records were then rewritten in terms of the PCs. Now, the dataset is down from 45 features to eight features.

Figure 4.2.1: Scree Plot of Principal Components



4.3 Scaling After Dimensionality Reduction

Since there was no obvious reason to believe why one PC is more important and should be weighted more than any others, all PCs were standardized using the z-scaling method so that all dimensions would be on the same footing. After the transformation, deviations from the center on all dimensions can now be equally valued in terms of standard deviation when determining outliers.

5. Anomaly Detection Algorithms

Two methods were deployed to detect anomalies in the NY Property dataset. The first one is the heuristic function of z-scores and the second is autoencoder. Both methods provide a score for each record, with larger numbers indicating greater extremeness. The output of the two methods was then ranked and combined to create the final anomaly score.

5.1 Method 1: Heuristic Function of the Z-scores (Linear Method)

The heuristic z-score method uses a fixed formula to calculate the distance of each data point to the centroid of the data space. This distance is used as the anomaly score to tell which records are outliers. By doing z-scaling after PCA, all PCs are now centered around zero and have standard deviations of one. Each record's distance to the origin directly shows how far away it is from other records, and hence is a good way to detect anomalies. Euclidean Distance technique was used to calculate such distances (see Formula 5.1.1 below). The calculated distances were named Anomaly Score 1. The greater the score is, the more anomalous a record is.

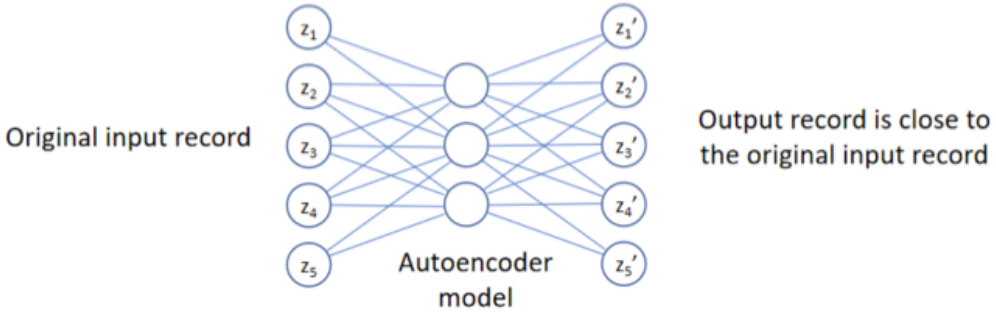
Formula 5.1.1: Anomaly Score 1

$$Anomaly\ Score\ 1\ = \sqrt{(Z_1)^2 + (Z_2)^2\ + (Z_3)^2 \ldots\ldots (Z_8)^2}$$

5.2 Method 2: Autoencoder (Non-Linear Method)

The second method is using an autoencoder, which is a special type of neural network (see Figure 5.2.1). An autoencoder is a model trained to reproduce the original input by studying the dominant pattern in the data. Because anomalies are outliers that don't follow the dominant pattern in the data well, the autoencoder would not do a good job reproducing an anomalous record and the resulting output would be relatively far from the input. Therefore, the distance between the input record and the output record is a good detector of anomalies.

Figure 5.2.1: Autoencoder Model



Similar to the first method, Euclidean Distance technique was used to calculate the distance between the input and output of the autoencoder (see Formula 5.2.2 below). Calculated distances were named Anomaly Score 2. The greater the score is, the more anomalous a record is.

Formula 5.2.2: Fraud Score 2

$$Anomaly\ Score\ 2 = \sqrt{(Z_{1,output} - Z_{1,input})^2 + (Z_{2,output} - Z_{2,input})^2 + \ldots\ldots + (Z_{8,output} - Z_{8,input})^2}$$

The Keras Python library was used to create the autoencoder. The hyperparameter setup of the autoencoder is documented in Appendix B. This autoencoder contains one hidden layer with five nodes.

5.3 Ensemble: Rank and Combine the Two Fraud Scores

The final anomaly score is an ensemble of the two scores output from the aforementioned algorithms. Because scores created using different models are on different scales, a one-unit difference in one score does not necessarily mean the same difference of extremeness in the other score. Thus, it is inappropriate to directly combine the numeric values of the scores (i.e. adding them together or taking the average). Instead, the team converted the scores to their relative ranks and then took the average of the ranks as the final score.

Firstly, all records were sorted by Anomaly Score 1 from the smallest to the largest. The record with the smallest score was given rank 1, while the record with the largest score was given rank 1,070,994. Then, the records were sorted and ranked once again by Anomaly Score 2. Finally, the average of two ranks was calculated for the final anomaly score (see Formula 5.3.1 below). For example, a record ranked at 8,000 by score 1 and 10,000 by score 2 would get a final score of 9,000. The larger the final score is, the more anomalous a record is. The 10 records with the highest final scores were sorted out for investigation.

Formula 5.3.1: Combine Two Anomaly Scores

$$Combined\ Anomaly\ Score_i = 0.5\ x\ Anomaly\ Score\ 1\ Rank_i + 0.5\ x\ \ Anomaly\ Score\ 2\ Rank_i$$

6. Results

6.1 Distribution of Anomaly Scores

The distributions of the anomaly score 1 and 2 are both skewed to the right, as shown below (see Figure 6.1.1 and 6.1.2). This finding complies with the intuition that anomalous cases are relatively rare compared to normal ones.

Figure 6.1.1: Anomaly Score 1 Distribution

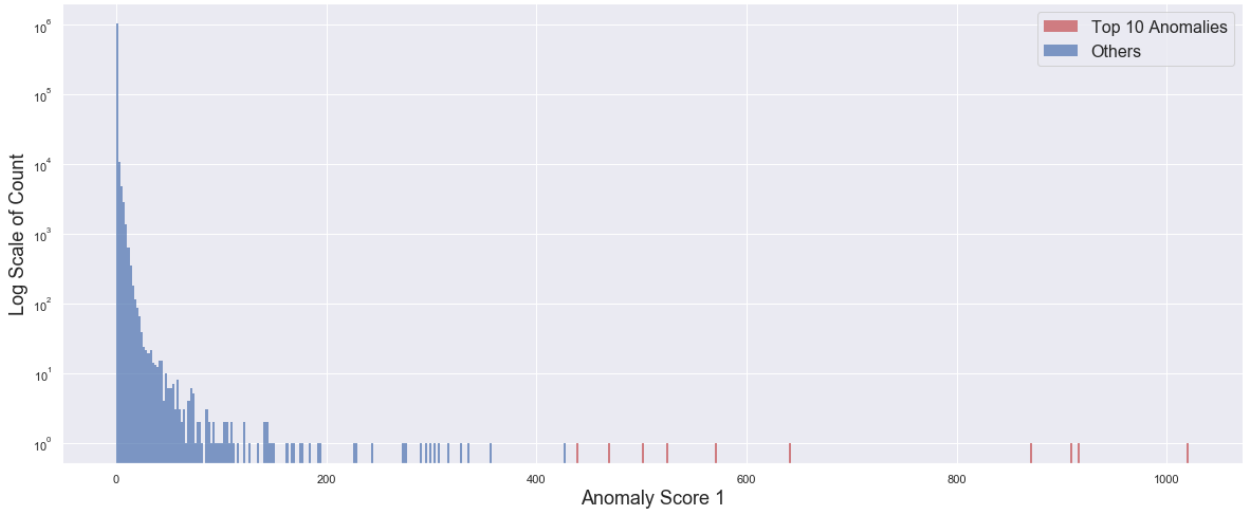
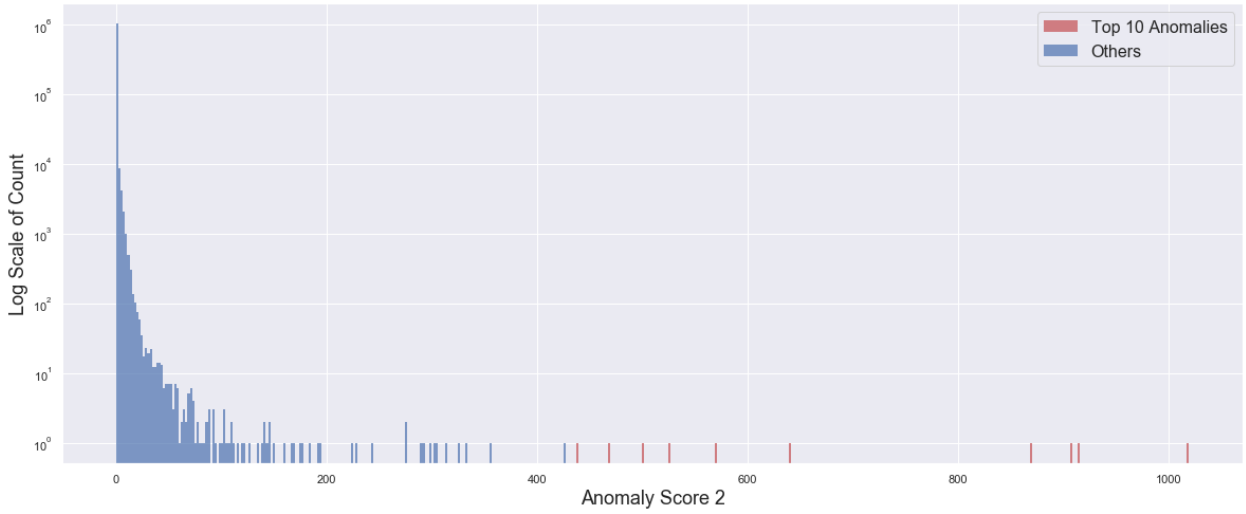
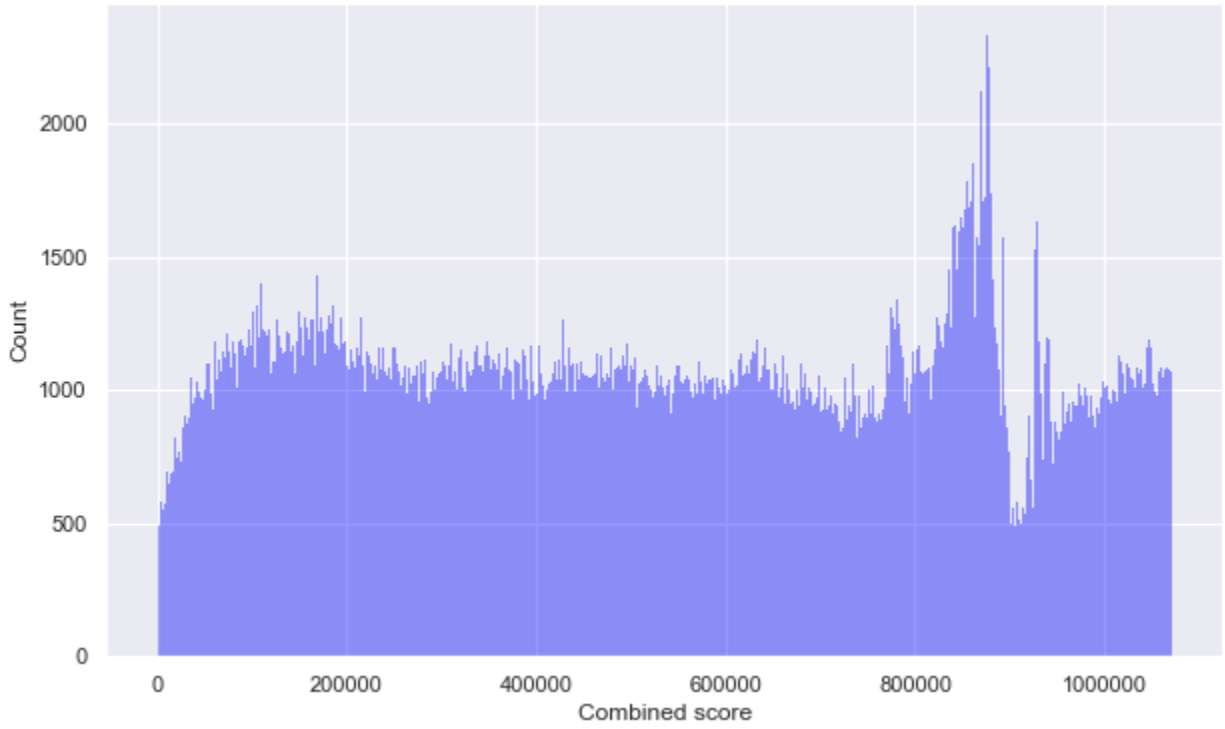


Figure 6.1.2: Anomaly Score 2 Distribution



The distribution of the final anomaly score is shown in Figure 6.1.3. It can be observed that the histogram is rather flat than rugged on the high score end, indicating that the two models identified the same set of anomalies and ranked them almost the same.

Figure 6.1.3: Final Anomaly Score Distribution



6.2 Summary of Top 10 Anomalies

The top 10 most anomalous records detected by the models are listed in Table 6.2.1 below. It's worth noting that the two models screened out the exact same ten records with the exact same order, which can be deemed as a reassurance that these records are truly outliers.

Table 6.2.1: Top 10 Anomalies Score Summarization

Record	Fraud Score 1	Fraud Score 2	Final Fraud Score
632816	1021.11	1020.27	1070994
565392	915.10	913.82	1070993
85886	908.77	907.89	1070992
67129	870.67	869.75	1070991
1067360	641.80	640.62	1070990
585118	571.28	570.71	1070989
245573	525.14	524.98	1070988
585439	500.50	499.73	1070987
917942	469.40	468.26	1070986
821853	439.76	438.71	1070985

The team further investigated the anomalies to understand 1) why they were flagged by the algorithms, and 2) whether they are false alarms or suspicious cases worthy of more inspection. Table 6.2.2 below listed the 45 z-scaled variables of the top 10 anomalies. All the fields with a z-score greater than 3 or less than -3 are highlighted.

Table 6.2.2: 45 Z-scaled Variables of the Top 10 Anomalies

	632816	565392	85886	67129	1067360	585118	245573	585439	917942	821853
r1_ZIP5_scaled	-0.07	52.35	0.12	50.54	231.82	-0.08	0.73	0.05	0.13	4.06
r1_ZIP3_scaled	-0.09	93.81	0.07	134.13	448.56	-0.15	-0.13	-0.09	0.2	9.09
r1_TAXCLASS_scaled	0.59	326.7	-0.02	68.39	267.46	-0.11	-0.08	0.09	0.59	368.2
r1_B_scaled	-0.08	91.08	0.07	140.7	435.52	-0.14	-0.13	-0.06	0.14	7.56
r1_all_scaled	-0.01	321.14	-0.09	67.17	784.43	-0.17	-0.15	0.03	0.52	18.57
r2_ZIP5_scaled	33.75	35	450.24	174.05	-0.03	373.39	377.3	402.52	2.47	-0.04
r2_ZIP3_scaled	74.56	86.33	467.1	708.47	-0.03	98.53	6.2	106.22	4.59	-0.05
r2_TAXCLASS_scaled	884.23	156.74	58.61	88.91	-0.01	184.01	8.4	198.36	8.31	-0.01
r2_B_scaled	86.15	90.53	487.51	739.44	-0.03	101.25	6.5	109.15	4.53	-0.05
r2_all_scaled	343.72	344.08	128.66	195.17	-0.01	403.96	18.41	435.48	18.22	-0.07
r3_ZIP5_scaled	6.75	5.2	894.23	49.7	-0.01	2.36	292.21	5.1	0.1	-0.01
r3_ZIP3_scaled	18.35	23.41	833.86	252.93	-0.01	1.48	1.73	3.21	0.38	-0.02
r3_TAXCLASS_scaled	1032.88	24.25	9.07	2.75	0	1.42	1.3	3.07	0.43	0
r3_B_scaled	19.23	21.56	834.2	253.04	-0.01	1.11	1.59	2.42	0.32	-0.02
r3_all_scaled	583.22	583.83	218.34	66.19	0	34.22	31.3	73.84	10.29	-0.05
r4_ZIP5_scaled	0.25	504.2	0.5	163.9	184.5	-0.01	1.96	0	23.35	39.42
r4_ZIP3_scaled	0.37	714.61	0.4	280.93	220.3	-0.03	0.08	-0.01	24.62	54.7
r4_TAXCLASS_scaled	2.38	513.77	0.07	103.78	215.59	-0.06	-0.02	-0.05	12.08	235.41
r4_B_scaled	0.43	699.6	0.42	297.17	215.68	0	0.07	0.02	20.89	46.44
r4_all_scaled	0.45	884.82	0.24	178.82	165.35	0.01	0.09	0.03	20.91	46.46
r5_ZIP5_scaled	191.54	161.11	537.36	279.61	-0.01	385.46	435.81	62.67	118.24	-0.01
r5_ZIP3_scaled	232.87	304.23	436.56	641.33	-0.01	295.2	17.31	48	145.13	-0.01
r5_TAXCLASS_scaled	997.97	95.3	35.48	52.13	0	111.88	5.11	18.19	53.93	0
r5_B_scaled	251.44	309.31	449.42	660.21	-0.01	295.5	17.6	48.05	142.45	-0.01
r5_all_scaled	448.71	449.18	167.25	245.7	-0.01	527.33	24.11	85.75	254.21	-0.01
r6_ZIP5_scaled	30.11	19.7	915.92	54.56	0	2.6	274.96	0.84	4.55	0
r6_ZIP3_scaled	59.52	83.7	856.42	251.61	-0.01	4.13	4.72	1.34	12.57	-0.01
r6_TAXCLASS_scaled	1034.46	13.09	4.87	1.43	0	0.77	0.7	0.25	2.47	0
r6_B_scaled	58.94	76.87	857.41	251.9	-0.01	3.46	4.33	1.12	11.13	-0.01
r6_all_scaled	633.83	634.49	236.26	69.41	-0.01	37.24	34.06	12.1	119.69	-0.01
r7_ZIP5_scaled	0.19	481.13	0.15	54.62	277.02	-0.05	0.81	0.23	56.79	39.06
r7_ZIP3_scaled	0.32	650.39	0.07	108.46	365.65	-0.07	0.03	0.26	66.44	62.89
r7_TAXCLASS_scaled	0.56	309.9	-0.02	64.87	331.32	-0.1	-0.07	0.09	19.01	349.25
r7_B_scaled	0.4	644.52	0.07	114.5	362.28	-0.04	0.03	0.51	54.64	51.71
r7_all_scaled	0.26	641.84	0.11	134.47	208.89	-0.06	0	0.34	39.51	37.39
r8_ZIP5_scaled	244.52	230.78	346.27	134.72	-0.02	378.64	290.17	408.18	433.69	-0.02
r8_ZIP3_scaled	295.71	360.73	213.44	323.73	-0.01	338.88	20.5	365.32	507.94	-0.02
r8_TAXCLASS_scaled	857.32	152.24	56.93	86.36	-0.01	178.73	8.16	192.67	224.36	-0.01
r8_B_scaled	316.82	354.78	208.97	316.95	-0.01	372.34	20.15	401.39	467.41	-0.02
r8_all_scaled	344.03	344.4	128.81	195.37	-0.01	404.32	18.48	435.86	507.55	-0.01
r9_ZIP5_scaled	61.55	45.08	900	50.02	-0.01	4.17	294.09	8.99	26.55	-0.01
r9_ZIP3_scaled	157.76	210.52	798.79	242.3	-0.01	9.79	12.17	21.11	92.42	-0.01
r9_TAXCLASS_scaled	1032.75	24.3	9.09	2.75	0	1.42	1.3	3.07	11.94	0
r9_B_scaled	162.55	195.39	802.79	243.51	-0.01	9.54	11.29	20.58	79.93	-0.01
r9_all_scaled	604.96	605.6	226.51	68.7	-0.01	35.54	32.51	76.63	297.48	-0.01

6.3 Analysis of Top 10 Anomalies

No. 1: Record 632816

Table 6.3.1: Record 632816 Feature Summary

632816	FULLVAL					AVLAND					AVTOT							
Lot Size	-0.07	-0.09	0.59	-0.08	-0.01	0.25	0.37	2.38	0.43	0.45	0.19	0.32	0.56	0.4	0.26	r1	r4	r7
Single Floor Area	33.75	74.56	884.23	86.15	343.72	191.54	232.87	997.97	251.44	448.71	244.52	295.71	857.32	316.82	344.03	r2	r5	r8
Total Floor Area	6.75	18.35	1032.88	19.23	583.22	30.11	59.52	1034.46	58.94	633.83	61.55	157.76	1032.75	162.55	604.96	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features:	r2, r3, r5, r6, r8, and r9
Anomaly Explanation:	Unusually small single floor area (*BLDFRONT x BLDDEPTH*) and total floor area (*BLDFRONT x BLDDEPTH x STORIES*)
Extreme Original Fields:	*BLDFRONT* =1, *BLDDEPTH* = 1
Property Address:	86-55 Broadway, NY 11373
Property Owner:	864163 REALTY, LLC
Building Class:	D9 - elevator apartments
Tax Class:	Class 2 - Residential properties, 4+ unit rental property, cooperatives and condominiums
Property Overview:	83-unit rental building located in Queens; construction completed in 2013[1]
Comment:	**False alarm.** *BLDFRONT* and *BLDDEPTH* were documented as 1 probably because the construction had not been completed by the time the data was collected.

No. 2: Record 565392

Table 6.3.2: Record 565392 Feature Summary

565392	FULLVAL					AVLAND					AVTOT							
Lot Size	52.35	93.81	326.7	91.08	321.14	504.2	714.61	513.77	699.6	884.82	481.13	650.39	309.9	644.52	641.84	r1	r4	r7
Single Floor Area	35	86.33	156.74	90.53	344.08	161.11	304.23	95.3	309.31	449.18	230.78	360.73	152.24	354.78	344.4	r2	r5	r8
Total Floor Area	5.2	23.41	24.25	21.56	583.83	19.7	83.7	13.09	76.87	634.49	45.08	210.52	24.3	195.39	605.6	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features:	r1, r2, r3, r4, r5, r6, r7, r8, and r9
Anomaly Explanation:	Extremely large market value and assessed land and total value (*FULLVAL, AVLAND, AVTOT*)
Extreme Original Fields:	*FULLVAL* over 4 billion, *AVLAND* nearly 2 billion, *AVTOT* nearly 2 billion
Property Address:	FLATBUSH AVENUE, originally missing ZIP, filled with 11234
Property Owner:	US government

[1] https://www.apartments.com/the-elm-east-elmhurst-ny/9p4eqz6/

Building Class:	V9 - miscellaneous vacant lots
Tax Class:	Class 4 - All other real property, including office buildings, factories, stores, hotels and lofts
Property Overview:	Without proper address and zip code, the exact property cannot be found.
Comment:	**Further investigation required.** Although this property is owned by the government who has little motivation to commit fraud, there is no convincing information from the data or from public records to explain the unusually high market value and assessed values.

No. 3: Record 85886

Table 6.3.3: Record 85886 Feature Summary

85886	FULLVAL					AVLAND					AVTOT							
Lot Size	0.12	0.07	-0.02	0.07	-0.09	0.5	0.4	0.07	0.42	0.24	0.15	0.07	-0.02	0.07	0.11	r1	r4	r7
Single Floor Area	450.2	467.1	58.61	487.5	128.66	537.36	436.56	35.48	449.42	167.25	346.27	213.44	56.93	208.97	128.81	r2	r5	r8
Total Floor Area	894.2	833.9	9.07	834.2	218.34	915.92	856.42	4.87	857.41	236.26	900	798.79	9.09	802.79	226.51	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features:	r2, r3, r5, r6, r8 and r9
Anomaly Explanation:	Unusually small floor area compared to lot size
Extreme Original Fields:	*BLDFRONT* = 8, *BLDDEPTH* = 8 vs. LTFRONT = 4000, LTDEPTH = 150
Property Address:	Joe Dimaggio Highway, originally missing ZIP, filled with 10023
Property Owner:	New York City Department of Parks and Recreation
Building Class:	Q1 - parks
Tax Class:	Class 4 - All other real property, including office buildings, factories, stores, hotels and lofts
Property Overview:	Based on the long and narrow lot shape, the location near Joe Dimaggio Highway, and its owner being the Department of Parks and Recreation, this property is very likely to be the Riverside Park in Manhattan.
Comment:	**False alarm**; reasonable lot shape and floor area for a park; government property.

No. 4: Record 67129

Table 6.3.4: Record 67129 Feature Summary

67129	FULLVAL					AVLAND					AVTOT							
Lot Size	50.54	134.1	68.39	140.7	67.17	163.9	280.93	103.78	297.17	178.82	54.62	108.46	64.87	114.5	134.47	r1	r4	r7
Single Floor Area	174.1	708.5	88.91	739.4	195.17	279.61	641.33	52.13	660.21	245.7	134.72	323.73	86.36	316.95	195.37	r2	r5	r8
Total Floor Area	49.7	252.9	2.75	253	66.19	54.56	251.61	1.43	251.9	69.41	50.02	242.3	2.75	243.51	68.7	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features:	r1, r2, r3, r4, r5, r6, r7, r8, and r9
Anomaly Explanation:	Extremely large market value and assessed land and total value
Extreme Original Fields:	*FULLVAL* over 6 billion, *AVLAND* over 2.6 billion, *AVTOT* over 2.7 billion; *LTDEPTH*, *BLDFRONT*, *BLDDEPTH* have values of 0 in the original dataset and were later imputed by the team
Property Address:	1000 5 Avenue, NY 10028
Property Owner:	New York City Department of Cultural Affairs
Building Class:	Q1 - parks
Tax Class:	Class 4 - All other real property, including office buildings, factories, stores, hotels and lofts
Property Overview:	This property is the Metropolitan Museum of Art, a worldly famous museum that houses over two million pieces of artwork.
Comment:	**False alarm**. As the largest art museum in the US, it is reasonable that the museum has a much higher property value than other properties in New York.

No. 5: Record 1067360

Table 6.3.5: Record 1067360 Feature Summary

10677360	FULLVAL					AVLAND					AVTOT							
Lot Size	231.8	448.6	267.46	435.5	784.43	184.5	220.3	215.59	215.68	165.35	277.02	365.65	331.32	362.28	208.89	r1	r4	r7
Single Floor Area	-0.03	-0.03	-0.01	-0.03	-0.01	-0.01	-0.01	0	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	r2	r5	r8
Total Floor Area	-0.01	-0.01	0	-0.01	0	0	-0.01	0	-0.01	-0.01	-0.01	-0.01	0	-0.01	-0.01	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features:	r1, r4, r7
Anomaly Explanation:	Unusually small lot size (*LTFRONT x LTDEPTH*)
Extreme Original Fields:	*LTFRONT* = 1, *LTDEPTH* = 1
Property Address:	20 Emily Court, NY 10307
Property Owner:	Missing
Building Class:	B2 - frame-built two-family dwellings
Tax Class:	Class 1 - Residential properties, one- to three- unit family homes
Property Overview:	2,850-sqft multiple occupancy home built in 1999, last sold for $234,000 in December 1999
Comment:	Potential entry error; **further investigation required**

No. 6: Record 585118

Table 6.3.6: Record 585118 Feature Summary

585118	FULLVAL					AVLAND					AVTOT							
Lot Size	-0.08	-0.15	-0.11	-0.14	-0.17	-0.01	-0.03	-0.06	0	0.01	-0.05	-0.07	-0.1	-0.04	-0.06	r1	r4	r7
Single Floor Area	373.4	98.53	184.01	101.3	403.96	385.46	295.2	111.88	295.5	527.33	378.64	338.88	178.73	372.34	404.32	r2	r5	r8
Total Floor Area	2.36	1.48	1.42	1.11	34.22	2.6	4.13	0.77	3.46	37.24	4.17	9.79	1.42	9.54	35.54	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features: r2, r5, r6, r8, r9

Anomaly Explanation: Unusually small single floor area

Extreme Original Fields: *BLDFRONT* = 1, *BLDDEPTH* = 1

Property Address: 28-10 Queens Plaza South, NY 11101

Property Owner: New York City Economic Development Corporation, a nonprofit organization contracted by the City of New York that serves as the City's primary entity for promoting and implementing economic development

Building Class: O3 - Office with only 7 - 19 stories

Tax Class: Class 4 - All other real property, including office buildings, factories, stores, hotels and lofts

Property Overview: City owned property leased to real estate company Tishman Speyer, who later started the construction of 1 and 3 Gotham Center on this site in 2016[2]

Comment: **False alarm**. *BLDFRONT* and *BLDDEPTH* were documented as 1 probably because the previous construction never finished due to the September 11 attacks and the new construction had not started by the time the data was collected. The city owns the property so there is little motivation to commit fraud.

No. 7: Record 245573

Table 6.3.7: Record 245573 Feature Summary

245573	FULLVAL					AVLAND					AVTOT							
Lot Size	0.73	-0.13	-0.08	-0.13	-0.15	1.96	0.08	-0.02	0.07	0.09	0.81	0.03	-0.07	0.03	0	r1	r4	r7
Single Floor Area	377.3	6.2	8.4	6.5	18.41	435.81	17.31	5.11	17.6	24.11	290.17	20.5	8.16	20.15	18.48	r2	r5	r8
Total Floor Area	292.2	1.73	1.3	1.59	31.3	274.96	4.72	0.7	4.33	34.06	294.09	12.17	1.3	11.29	32.51	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features: r2, r3, r5, r6, r8 and r9

Anomaly Explanation: Unusually small single and total floor area

Extreme Original Fields: *BLDFRONT* = 10, *BLDDEPTH* = 20

Property Address: Balcom Ave, originally missing ZIP, filled with 10462

[2] https://www.6sqft.com/construction-update-tishman-speyers-trio-of-long-island-city-rental-towers/

Property Owner:	New York City Department of Parks and Recreation
Building Class:	Q1 - parks
Tax Class:	Class 4 - All other real property, including office buildings, factories, stores, hotels and lofts
Property Overview:	Based on the block (5583) and lot (100) code, the location near Balcom Ave, and its owner being the Department of Parks and Recreation, this property is very likely to be part of the Trump Golf Links, a public golf course in the Bronx.
Comment:	**False alarm**; reasonable lot size and floor area for a golf course; government-owned property

No. 8: Record 585439

Table 6.3.8: Record 585439 Feature Summary

585439	FULLVAL					AVLAND					AVTOT							
Lot Size	0.05	-0.09	0.09	-0.06	0.03	0	-0.01	-0.05	0.02	0.03	0.23	0.26	0.09	0.51	0.34	r1	r4	r7
Single Floor Area	402.5	106.2	198.36	109.2	435.48	62.67	48	18.19	48.05	85.75	408.18	365.32	192.67	401.39	435.86	r2	r5	r8
Total Floor Area	5.1	3.21	3.07	2.42	73.84	0.84	1.34	0.25	1.12	12.1	8.99	21.11	3.07	20.58	76.63	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features:	r2, r3, r5, r6, r8 and r9
Anomaly Explanation:	Unusually small single and total floor area
Extreme Original Fields:	*BLDFRONT* = 1, *BLDDEPTH* = 1
Property Address:	11-01 43rd Avenue, NY 11101
Property Owner:	11-01 43rd Avenue Realty Corporation
Building Class:	H9 - various types of hotels that don't fit into any of the other H-class buildings
Tax Class:	Class 4 - All other real property, including office buildings, factories, stores, hotels and lofts
Property Overview:	Z NYC Hotel in Long Island; construction completed in 2011[3]
Comment:	**False alarm**. *BLDFRONT* and *BLDDEPTH* were documented as 1 probably because the construction had not been completed by the time the data was collected.

[3] https://therealdeal.com/2018/08/01/long-island-citys-z-nyc-hotel-sells-for-43m/

No. 9: Record 917942

Table 6.3.9: Record 917942 Feature Summary

917942	FULLVAL					AVLAND					AVTOT							
Lot Size	0.13	0.2	0.59	0.14	0.52	23.35	24.62	12.08	20.89	20.91	56.79	66.44	19.01	54.64	39.51	r1	r4	r7
Single Floor Area	2.47	4.59	8.31	4.53	18.22	118.24	145.13	53.93	142.45	254.21	433.69	507.94	224.36	467.41	507.55	r2	r5	r8
Total Floor Area	0.1	0.38	0.43	0.32	10.29	4.55	12.57	2.47	11.13	119.69	26.55	92.42	11.94	79.93	297.48	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features:	r2, r4, r5, r6, r7, r8 and r9
Anomaly Explanation:	Unusually high assessed land value and total value; extremely large lot size
Extreme Original Fields:	*AVLAND* = 1.8 billion, *AVTOT* = 4.7 billion, *LTFRONT* = 4910
Property Address:	154-68 Brookville Boulevard, NY 11422
Property Owner:	Logan Property, INC.
Building Class:	T1 - airport, airfield, terminal
Tax Class:	Class 4 - All other real property, including office buildings, factories, stores, hotels and lofts
Property Overview:	Very likely a school bus depot based on public record[4]
Comment:	**Further investigation required**. Although the unusually large lot size can be explained by its existence as a parking lot. There is no convincing information from the data or from public records to explain why the market value is extremely high.

No. 10: Record 821853

Table 6.3.9: Record 821853 Feature Summary

821853	FULLVAL					AVLAND					AVTOT							
Lot Size	4.06	9.09	368.2	7.56	18.57	39.42	54.7	235.41	46.44	46.46	39.06	62.89	349.25	51.71	37.39	r1	r4	r7
Single Floor Area	-0.04	-0.05	-0.01	-0.05	-0.07	-0.01	-0.01	0	-0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	r2	r5	r8
Total Floor Area	-0.01	-0.02	0	-0.02	-0.05	0	-0.01	0	-0.01	-0.01	-0.01	-0.01	0	-0.01	-0.01	r3	r6	r9
	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL	Z5	Z3	T	B	ALL			

Extreme Features:	r1, r4 and r7
Anomaly Explanation:	Unusually small lot size; single floor area larger than lot size
Extreme Original Fields:	*LTFRONT* = 2, *LTDEPTH* = 1, *BLDFRONT* = 22, *BLDDEPTH* = 48
Property Address:	87 drive, originally missing ZIP, filled with 11373
Property Owner:	CNY/NYCTA (New York City Transit Authority)
Building Class:	U7 - public ownership transportation utility properties
Tax Class:	Class 3 - Utility company equipment and special franchise property
Property Overview:	Without proper address and zip code, the exact property cannot be

[4] https://www1.nyc.gov/assets/planning/download/pdf/about/cpc/040309.pdf

found. But based on the building class and tax class, it could be a MTA bus stop, terminal or other transportation related building.

Comment: **Further investigation required**. Although this property is owned by a government authority who has little motivation to commit fraud, there is no convincing information from the data or from public records to explain why the lot size is so small or how the single floor area can possibly be larger than the lot size.

7. Conclusions

This report aims to explain the steps to build an unsupervised anomaly detection model that is reproducible for future use. The process of building the model involved the following key tasks:

1. Perform exploratory analysis to understand the quality of the dataset and the distribution of the variables
2. Fill in missing values of significant fields using proper imputation methods
3. Create expert variables based on research and expert suggestions
4. Perform z-scaling to standardize variables and Principal Component Analysis to reduce dimensionality
5. Apply the heuristic z-score function method and autoencoder to the final features respectively to calculate two anomaly scores for each record
6. Combine the two anomaly scores and identify the ten most anomalou records with the largest final anomaly scores
7. Investigate the anomalous records to find potential frand

After careful examination of the dataset and research on public records, four of the most anomalous records were flagged as suspicious cases that require a further inspection (record 565392, 1067360, 917942, and 821853). The remaining six records were determined to be reasonable, two of which were government-owned park or golf court that has large lots and relatively small building areas (record 85886 and 245573), one is the Metropolitan Museum of Art which was valued extremely high for good reasons (record 67129), and the other three have buildings under construction at the time the data was collected and therefore have unusually small floor area (record 632816 - rental property, 585118 - office building, 585439 - hotel).

Although the methodologies used for this project are solid, the implementation could be improved should there be more time. The following approaches are recommended to be explored:

- Consult with domain experts to understand how the data was collected and how each field was measured
- Include more variables that could be good indicators of tax fraud
- Use more reasonable techniques to fill in the missing values, such as incorporating Google API to retrieve more accurate zip codes.

Reference

Department of Finance. (2011, September 2). Property Valuation and Assessment Data: NYC Open Data. Retrieved January 20, 2020, from https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8

Appendix A. - Data Quality Report for New York Property Valuation and Assessment Data

1. Executive Summary

This report serves to provide a quality assessment of the New York Property dataset ('NY Property data.csv') and communicate the processes followed to measure the quality of the data. The dataset was collected by the Department of Finance of New York City and is accessible to the public through the NYC Open Data website. It is for New York properties valuation and assessment to calculate property tax, grant eligible properties exemptions, and (or) abatements (2011). In this data quality report, a descriptive data analysis was conducted, including data overview, variables summary tables, and visualizations for applicable fields.

2. Data Overview & Variable Tables

2.1. Data Overview

Data Name: Property Valuation and Assessment Data
Data Source: Department of Finance, City of New York (open source)
Time Period: 11/2010
Number of Fields: 32
Number of Records: 1,070,994

2.2. Numerical Fields Table

Field Name	# of Records w/ Value	% Populated	# of Unique Values	# of Records w/ Value 0	Mean	Standard Deviation	Min	Max
LTFRONT	1,070,994	100.00	1,297	169,108	36.64	74.03	0	9,999.00
LTDEPTH	1,070,994	100.00	1,370	170,128	88.86	76.40	0	9,999.00
STORIES	1,014,730	94.75	111	0	5.01	8.37	1	119.00
FULLVAL	1,070,994	100.00	109,324	13,007	874,264.51	11,582,430.99	0	6,150,000,000.00
AVLAND	1,070,994	100.00	70,921	13,009	85,067.92	4,057,260.06	0	2,668,500,000.00
AVTOT	1,070,994	100.00	112,914	13,007	227,238.17	6,877,529.31	0	4,668,308,947.00
EXLAND	1,070,994	100.00	33,419	491,699	36,423.89	3,981,575.79	0	2,668,500,000.00
EXTOT	1,070,994	100.00	64,255	432,572	91,186.98	6,508,402.82	0	4,668,308,947.00
BLDFRONT	1,070,994	100.00	612	228,815	23.04	35.58	0	7,575.00
BLDDEPTH	1,070,994	100.00	621	228,853	39.92	42.71	0	9,393.00
AVLAND2	282,726	26.40	58,591	0	246,235.72	6,178,962.56	3	2,371,005,000.00
AVTOT2	282,732	26.40	111,360	0	713,911.44	11,652,528.95	3	4,501,180,002.00

EXLAND2	87,449	8.17	22,195	0	351,235.68	10,802,212.67	1	2,371,005,000.00
EXTOT2	130,828	12.22	48,348	0	656,768.28	16,072,510.17	7	4,501,180,002.00

(Table 1)

2.3. Categorical Fields Table

Field Name	# of Records	% Populated	# of Unique Values	Most Common Field Value	Specific Variable Type
RECORD	1,070,994	100.00	1,070,994	n/a	Categorical
BBLE	1,070,994	100.00	1,070,994	n/a	Categorical
B	1,070,994	100.00	5	4	Categorical
BLOCK	1,070,994	100.00	13,984	3,944	Categorical
LOT	1,070,994	100.00	6,366	1	Categorical
EASEMENT	4,636	0.43	12	E	Categorical
OWNER	1,039,249	97.04	863,346	PARKCHESTER PRESERVAT	Categorical
BLDGCL	1,070,994	100.00	200	R4	Categorical
TAXCLASS	1,070,994	100.00	11	1	Categorical
EXT	354,305	33.08	3	G	Categorical
EXCD1	638,488	59.62	129	1,017	Categorical
STADDR	1,070,318	99.94	839,280	501 SURF AVENUE	Categorical
ZIP	1,041,104	97.21	196	10,314	Categorical
EXMPTCL	15,579	1.45	14	X1	Categorical
EXCD2	92,948	8.68	60	1,017	Categorical
PERIOD	1,070,994	100.00	1	FINAL	Categorical
YEAR	1,070,994	100.00	1	2010/11	Date/Time
VALTYPE	1,070,994	100.00	1	AC-TR	Categorical

(Table 2)

3. Field Explorations

3.1. Field 1

Field Name: RECORD
Field Type: categorical
Description: a unique integer label for each record, range from 1 to 1,070,994.

3.2. Field 2

Field Name: BBLE
Field Type: categorical
Description: a unique label for each record, a concatenation of B, BLOCK, LOT AND EASEMENT, max length is 11, alphanumeric format.

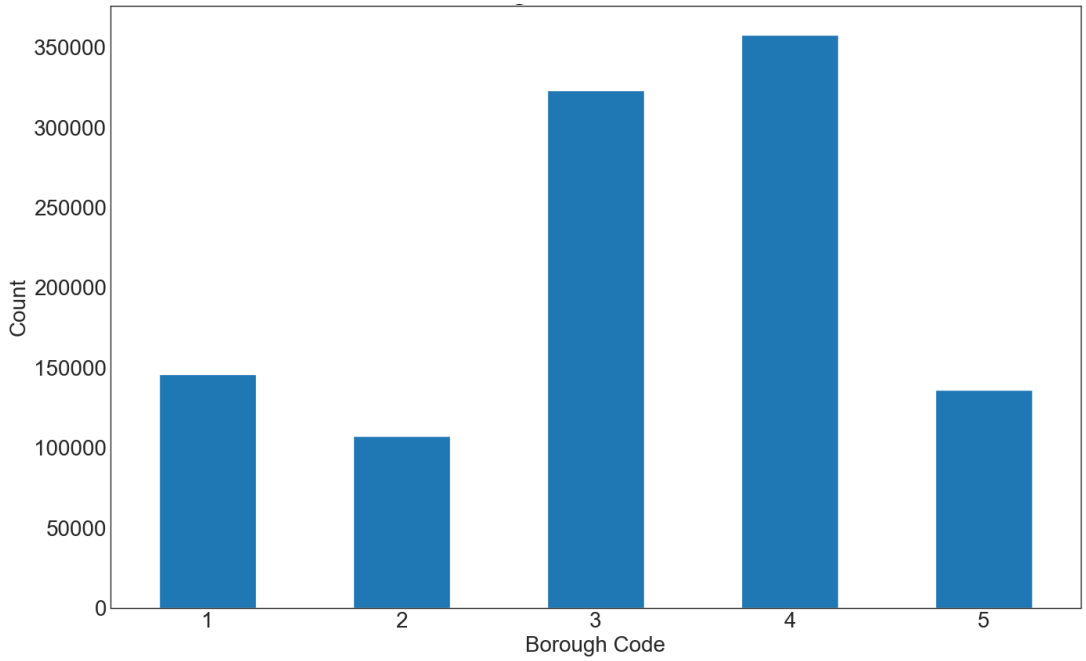
3.3. Field 3

Field Name: B

Field Type: Categorical

Description: borough codes (1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island). See *Figure 1*.

Figure 1: Borough Code Distribution



3.4. Field 4

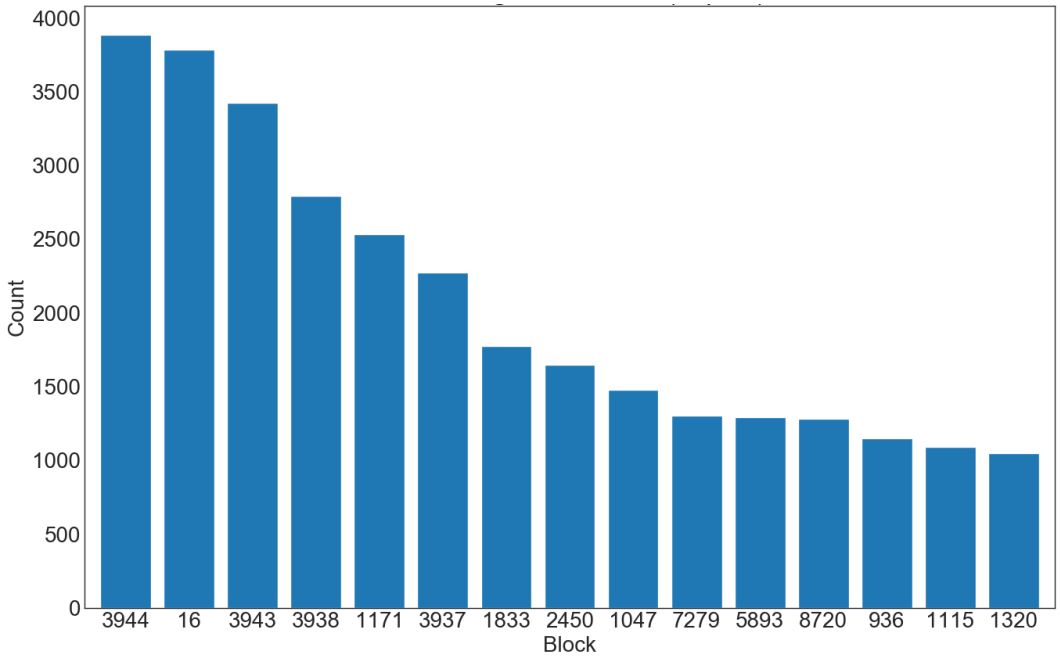
Field Name: BLOCK

Field Type: categorical

Description: valid block ranges by borough codes: Manhattan (1 to 2,255), Bronx (2,260 to 5,958), Brooklyn (1 to 8,955), Queens (1 to 16,350), Staten Island (1 to 8,050).

Ranking by property count, the top 15 block ranges were included in the following plot. See *Figure 2*.

Figure 2: Bloack Range Distribution (Top 15)

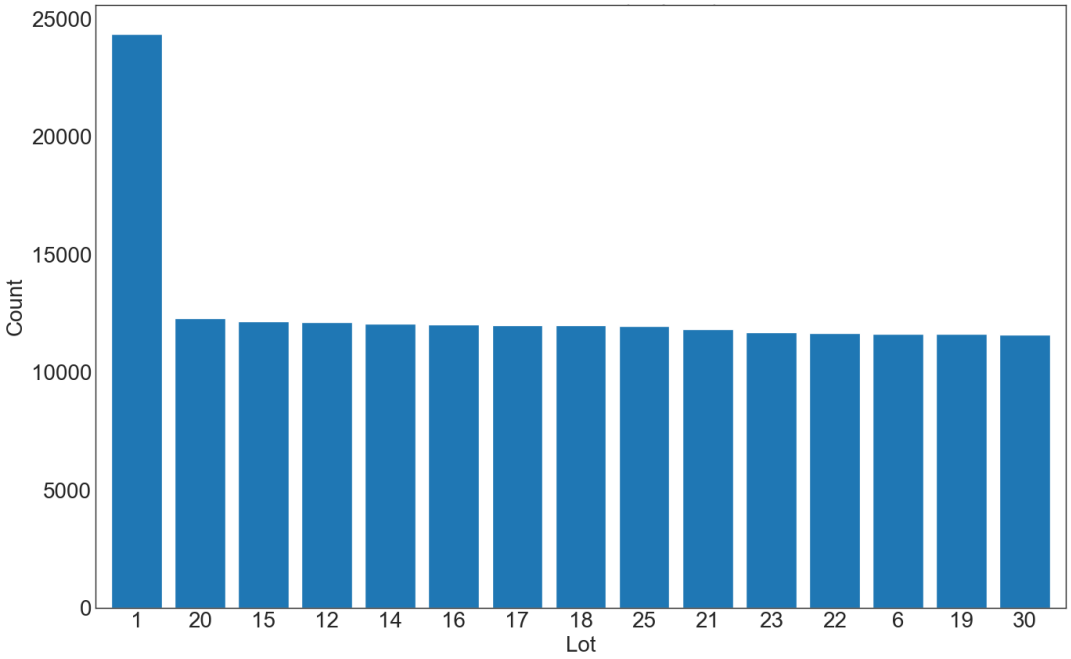


3.5. Field 5

Field Name: LOT
Field Type: categorical
Description: unique number within boro/block

Ranking by property count, the top 15 lots were included in the following plot. See *Figure 3*.

Figure 3 : Lot Distribution (Top 15)



3.6. Field 6

Field Name: EASEMENT

Field Type: categorical

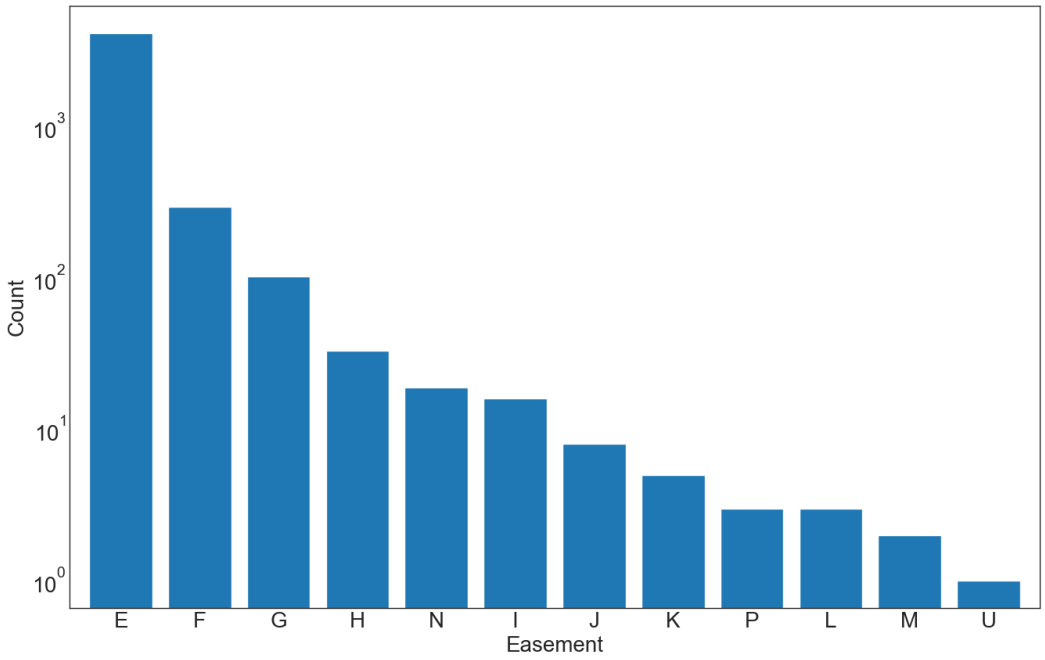
Description: a field that is used to describe easement with the following details (*Table 3*). See *Figure 4*.

SPACE Indicates the lot has no Easement.
'A' - Indicates the portion of the Lot that has an Air Easement
'B' - Indicates Non-Air Rights.
E' - Indicates the portion of the lot that has a Land Easement
'F' - THRU 'M' Are duplicates of 'E'.
'N' - Indicates Non-Transit Easement

'P' - Indicates Piers.
'R' - Indicates Railroads.
'S' - Indicates Street
'U' - Indicates U.S. Government

(Table 3)

Figure 4: Easement Distribution

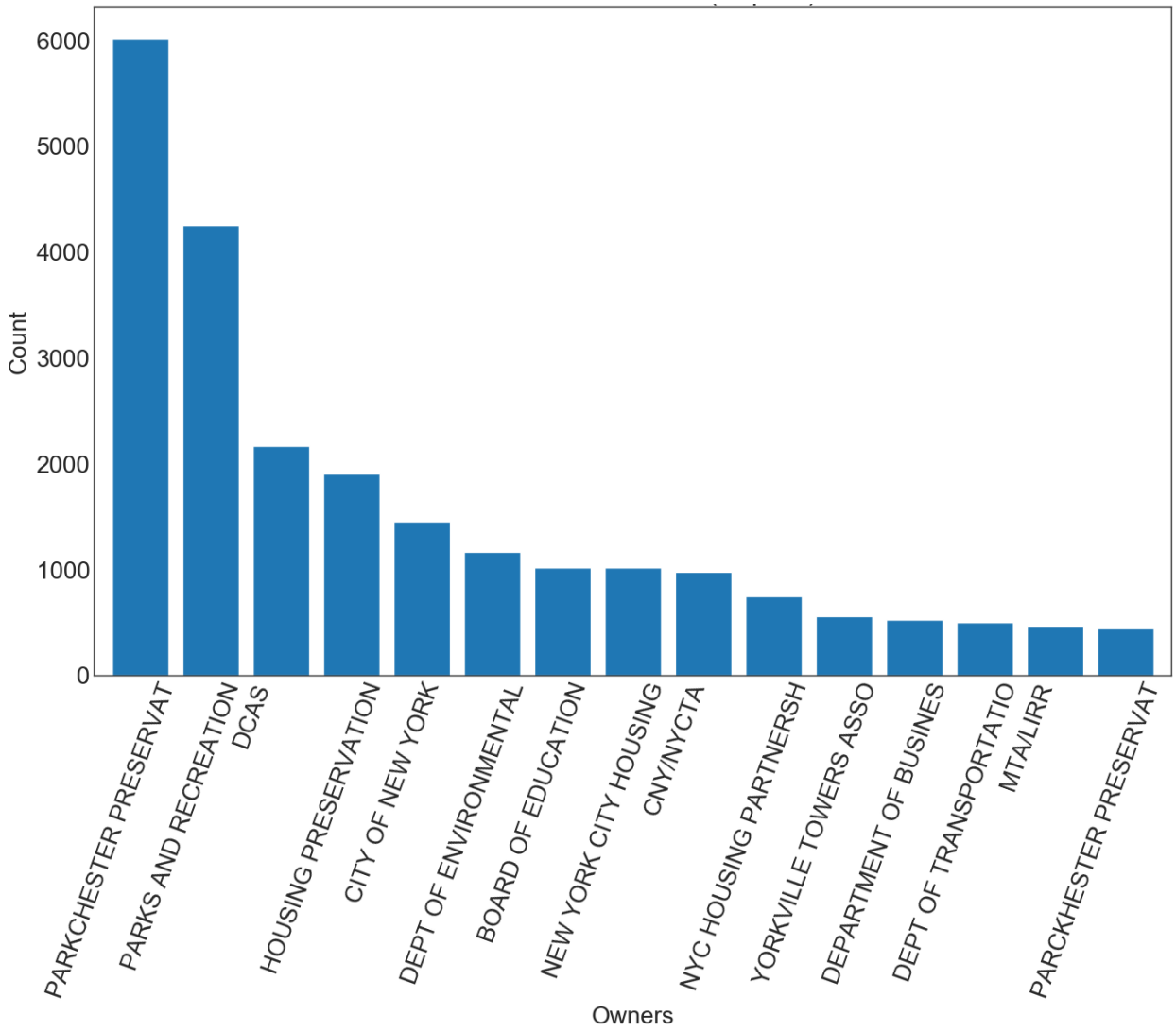


3.7. Field 7

Field Name: OWNER
Field Type: Categorical
Description: property owner's name

Ranking by property count, the top 15 owners' names were included in the following plot. See *Figure 5*.

Figure 5: Owners Distribution (Top 15)

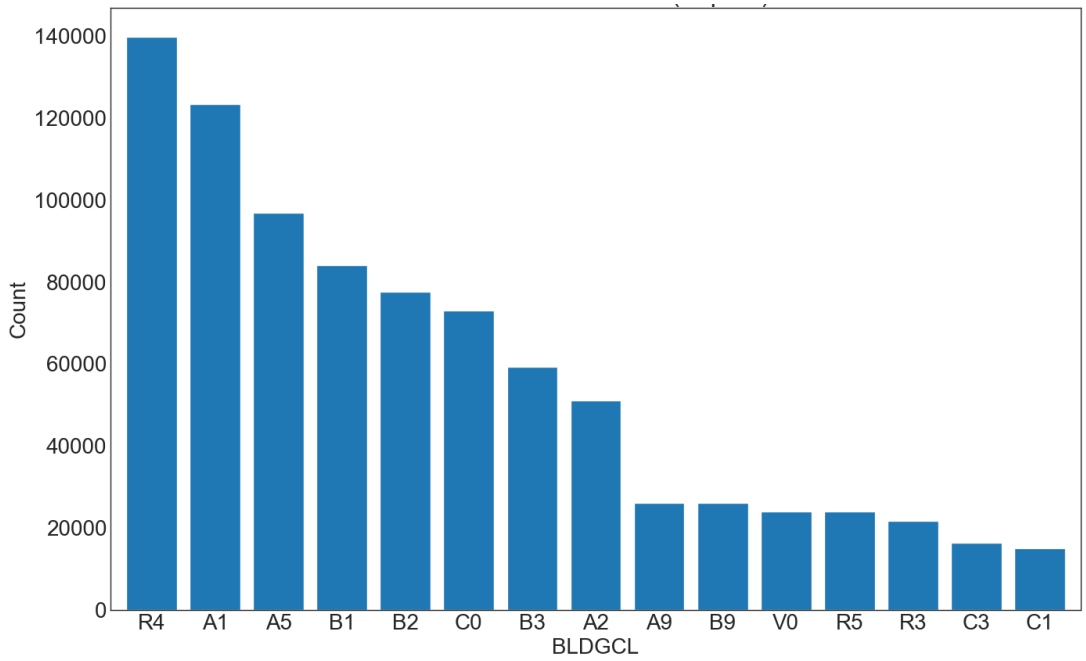


3.8. Field 8

Field Name: BLDGCL
Field Type: Categorical
Description: Building class

Ranking by property count, the top 15 building classes were included in the following plot. See *Figure 6*.

Figure 6: BLDGCL Distribution (Top 15)



3.9. Field 9

Field Name: TAXCLASS

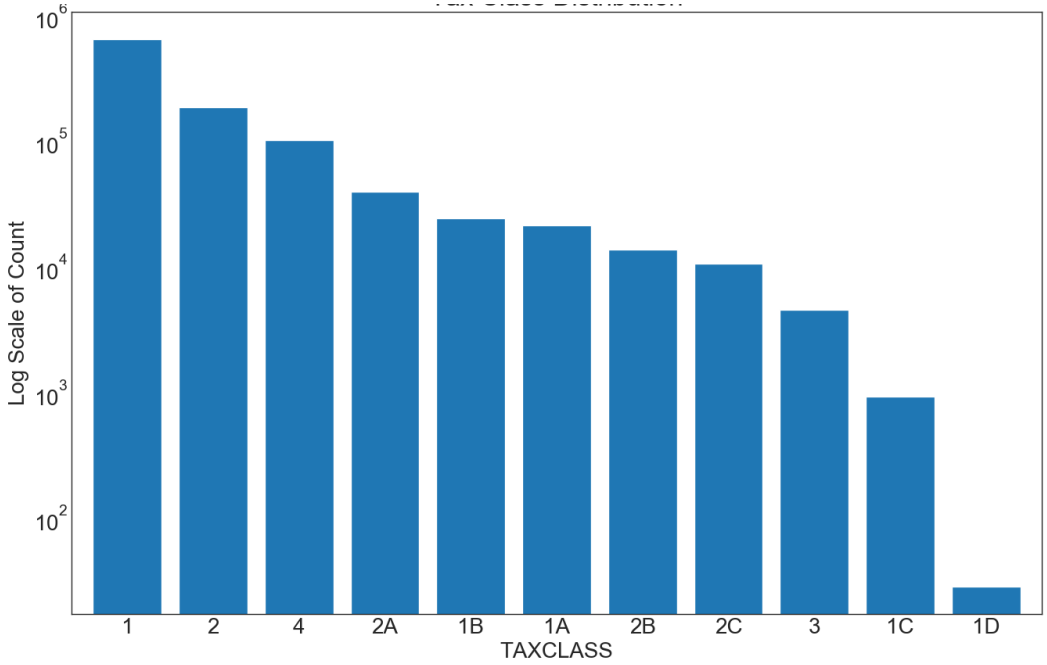
Field Type: Categorical

Description: a field indicates the tax class of a property. Plot see *Figure 7*. The details of the classes are shown as follows (*Table 4*):

'1' = 1 - 3 Unit Residence
'2' = Apartment, '2A' = 4, 5, or 6 units
'3' = Utilities
'4' = All others

(Table 4)

Figure 7: Tax Class Distribution

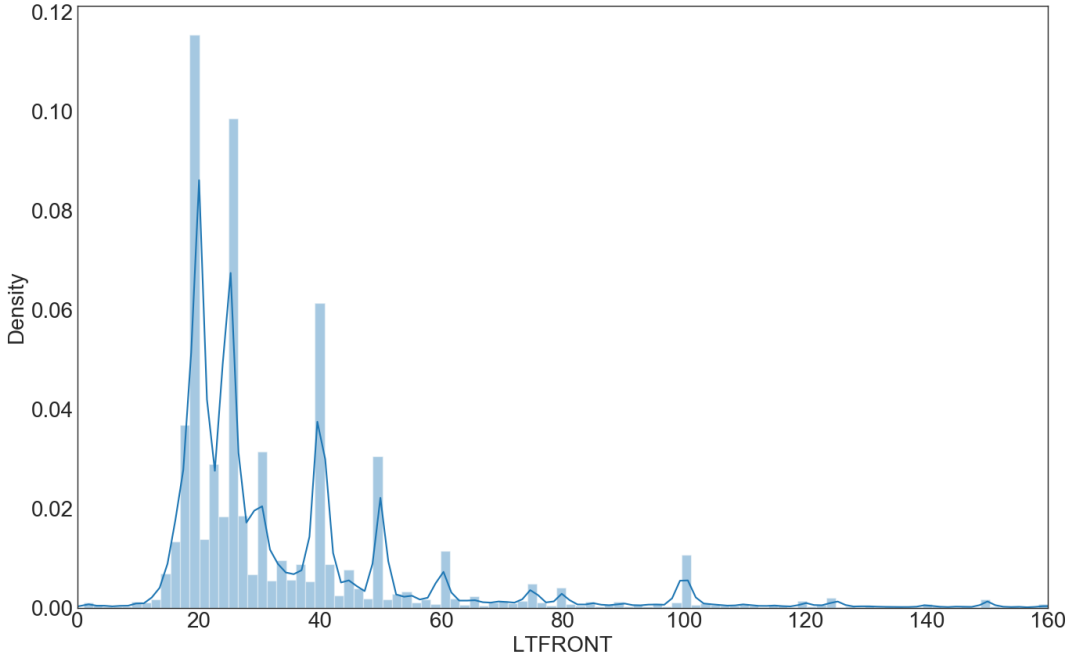


3.10. Field 10

Field Name: LTFRONT
Field Type: Numerical
Description:

Records with value 0 were removed from the plot below, considering that a lot width of zero has no practical meaning. Besides, records with lot widths over 160 were also omitted from the plot due to a small number of samples. See *Figure 8*.

Figure 8: Lot Width Distribution

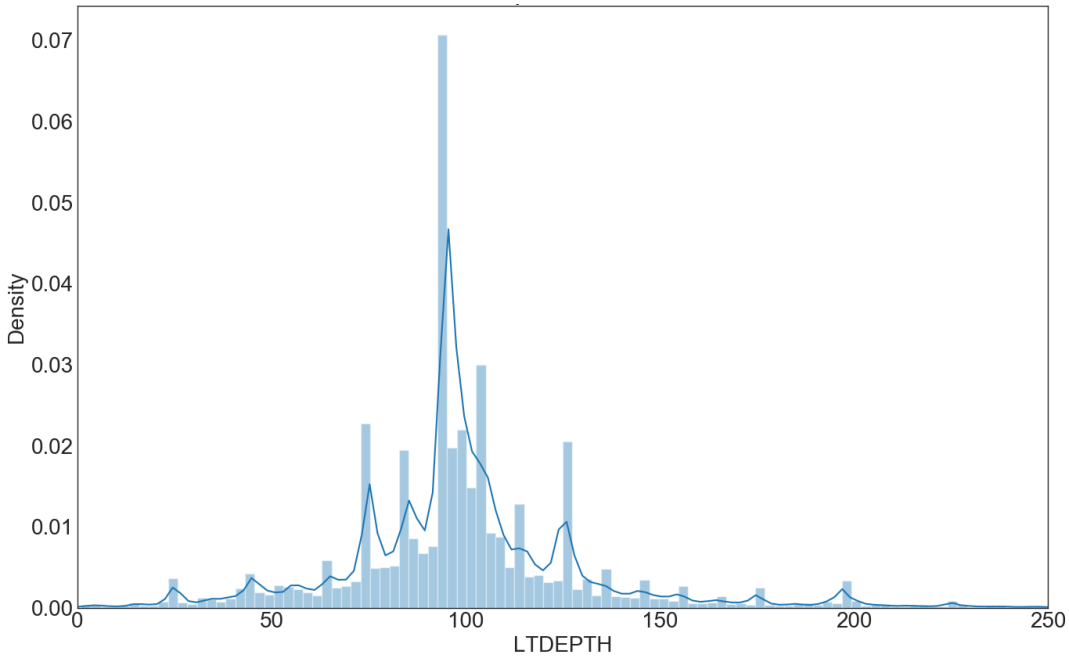


3.11. Field 11

Field Name: LTDEPTH
Field Type: Numerical
Description: Lot depth.

Records with value 0 were removed from the plot below, considering that lot width of zero has no practical meaning. Besides, records with lot depths over 250 were also omitted from the plot due to a low volume of data. See *Figure 9*.

Figure 9: Lot Depth Distribution



3.12. Field 12

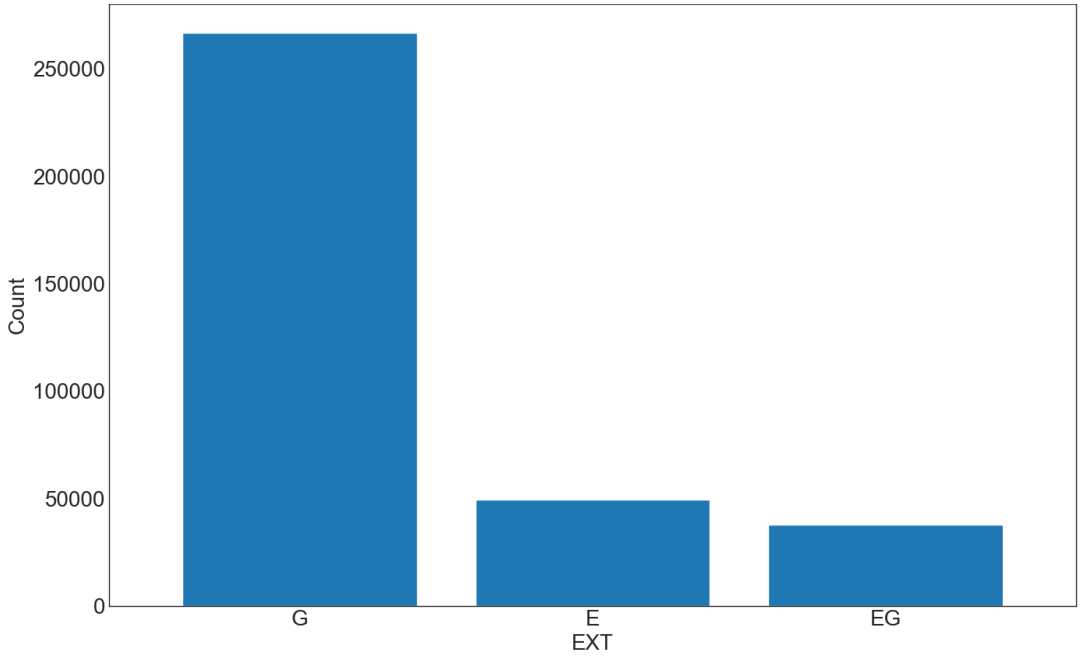
Field Name: EXT

Field Type: categorical

Description: extension indicator (‘E’ = extension, ‘G’ = garage, ‘EG’ = extension and garage).

See *Figure 10*.

Figure 10: Extension Distribution

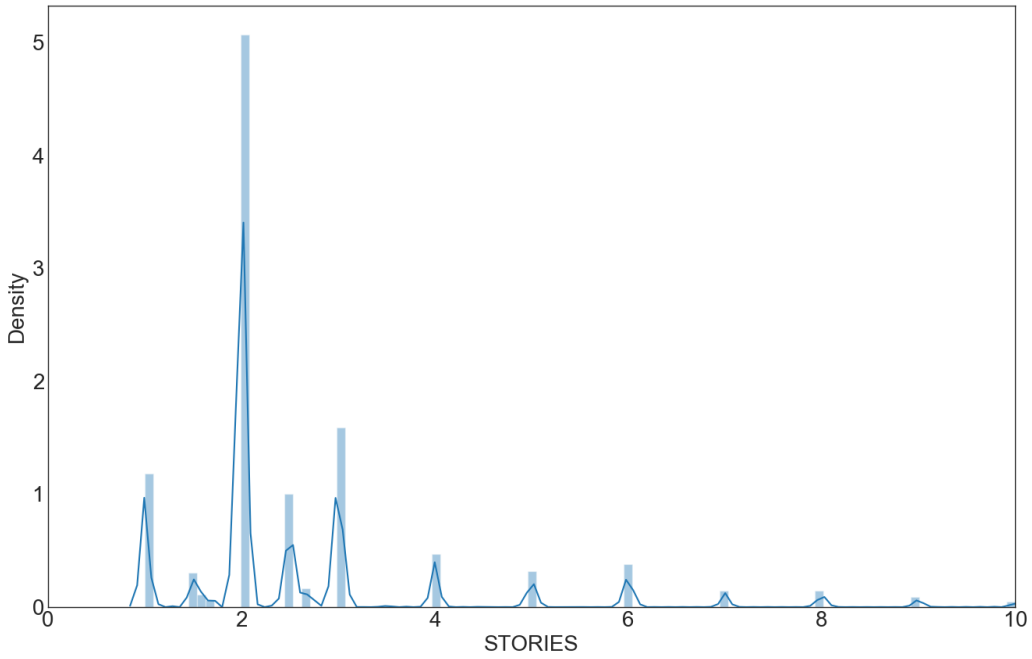


3.13. Field 13

Field Name: STORIES
Field Type: Numerical
Description: number of stories in the building.

Records with lot depths over 30 were omitted from the plot due to a low volume of data. See *Figure 11*.

Figure 11: Strories Distribution

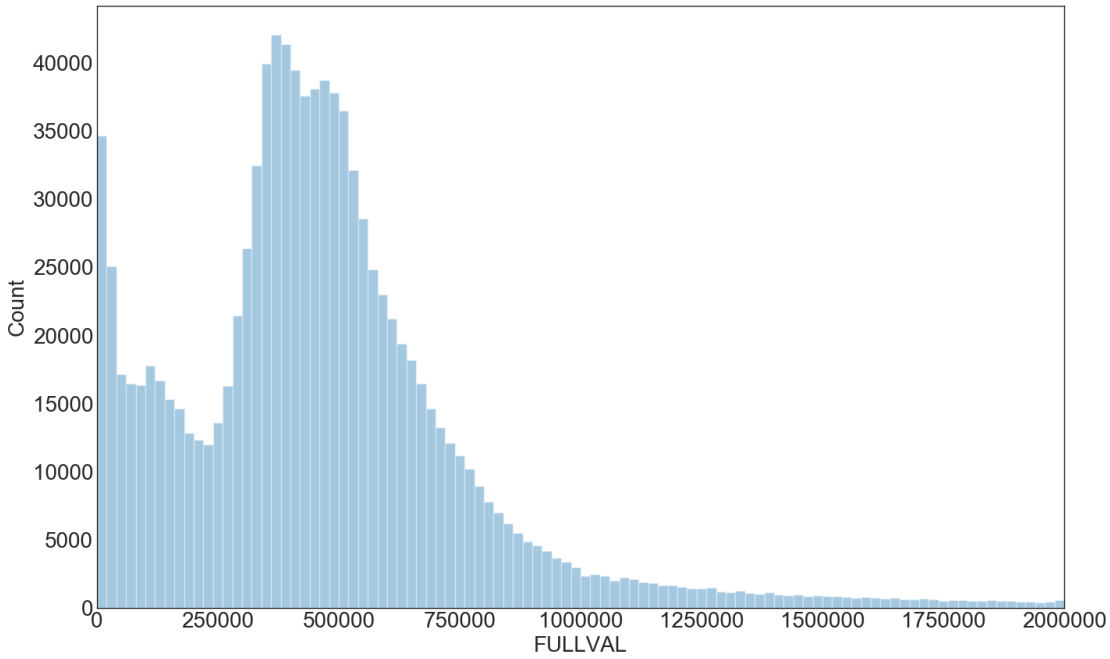


3.14. Field 14

Field Name: FULLVAL
Field Type: Numerical
Description: property market value.

Records with market values over 2 million were omitted from the plot due to a low volume of data. See *Figure 12*.

Figure 12: Market Value Distribution

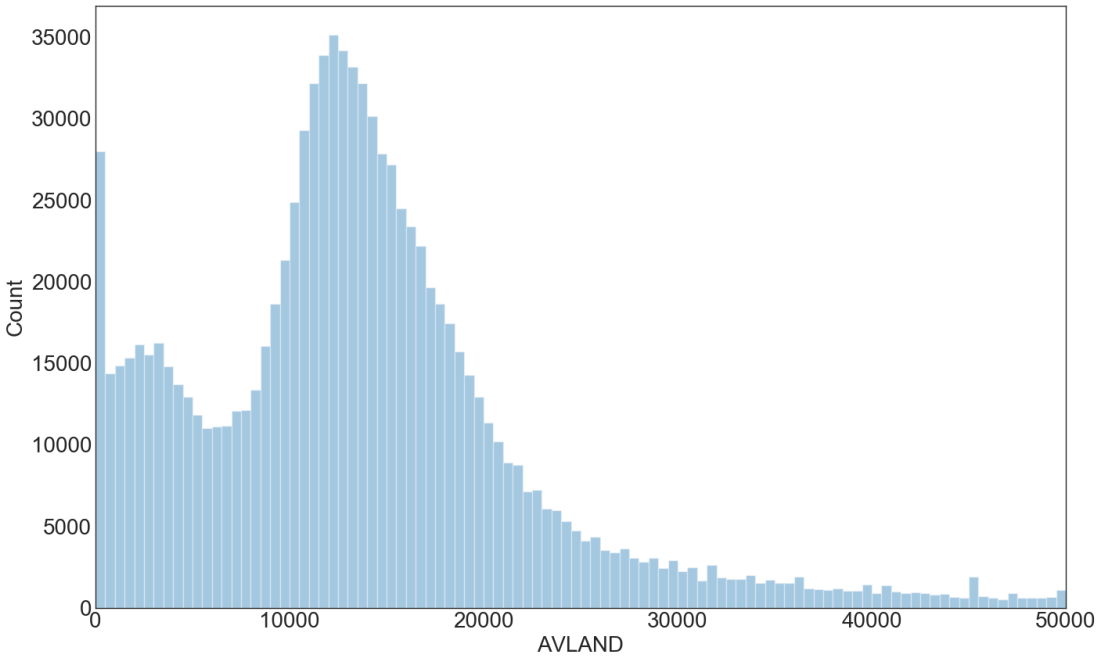


3.15. Field 15

Field Name: AVLAND
Field Type: Numerical
Description: actual land value.

Records with land values over 50,000 were omitted from the plot due to a lower volume of data. See *Figure 13*.

Figure 13: Actual Land Value Distribution

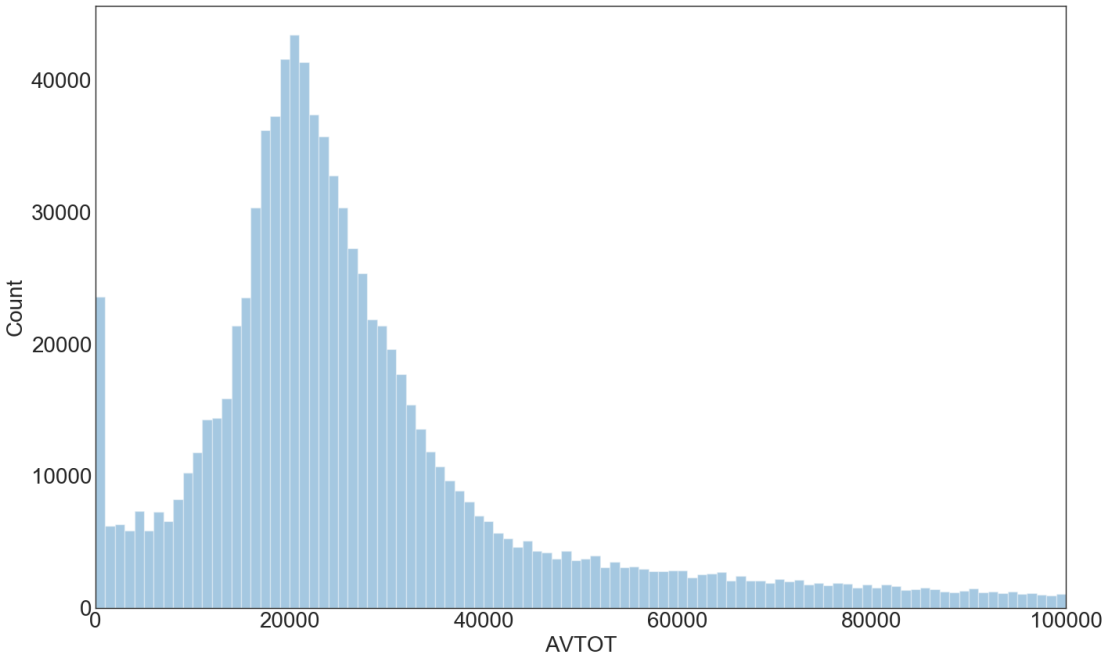


3.16. Field 16

Field Name: AVTOT
Field Type: Numerical
Description: actual total value.

Records with total values over 100,000 were omitted from the plot due to a low volume of samples. See *Figure 14*.

Figure 14: Actual Total Value Distribution

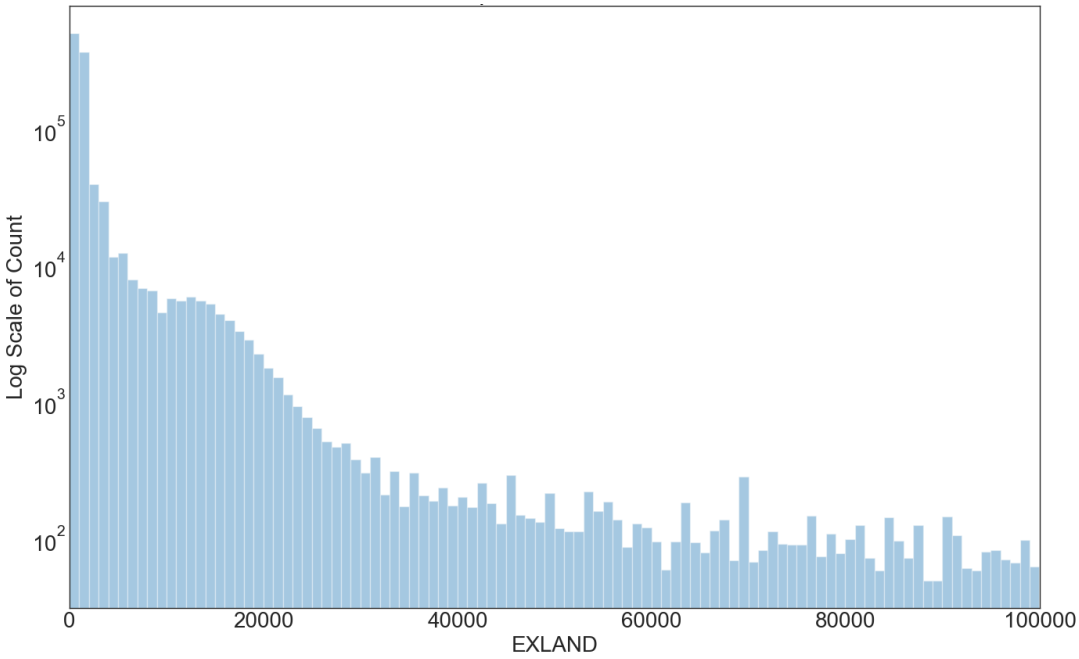


3.17. Field 17

Field Name: EXLAND
Field Type: Numerical
Description: actual exempt land value.

Records with actual exempt land values over 100,000 were omitted from the plot due to a low volume of samples. See *Figure 15*.

Figure 15: Actual Exempt Land Value Distribution

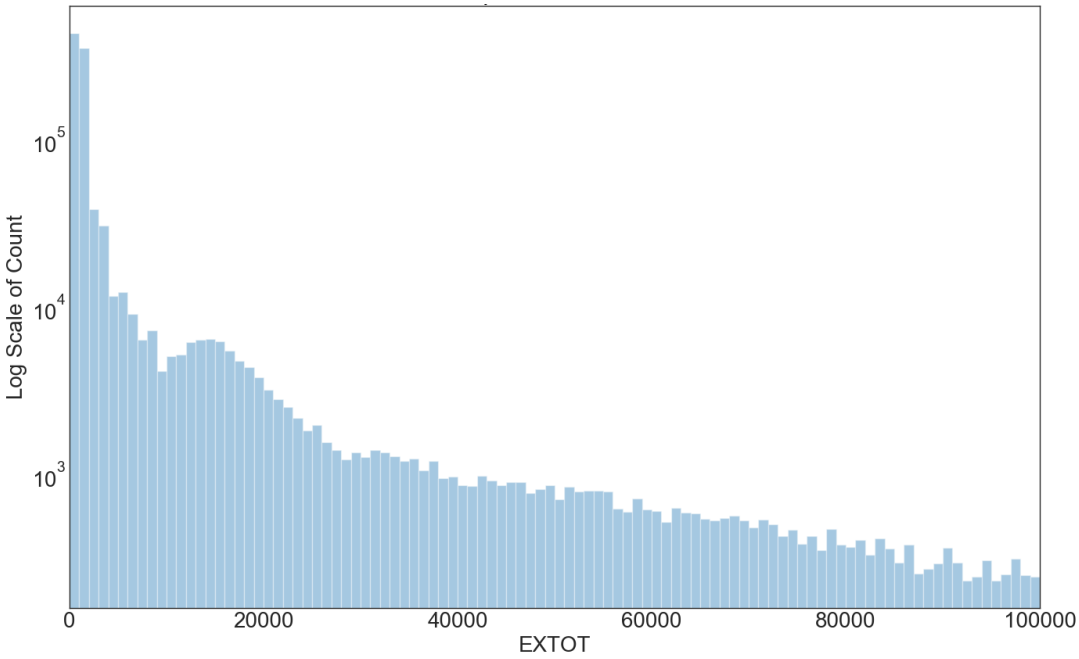


3.18. Field 18

Field Name: EXTOT
Field Type: Numerical
Description: actual exempt land total.

Records with actual exempt land totals over 100,000 were omitted from the plot due to a low volume of samples. See *Figure 16*.

Figure 16: Actual Exempt Land Total Distribution

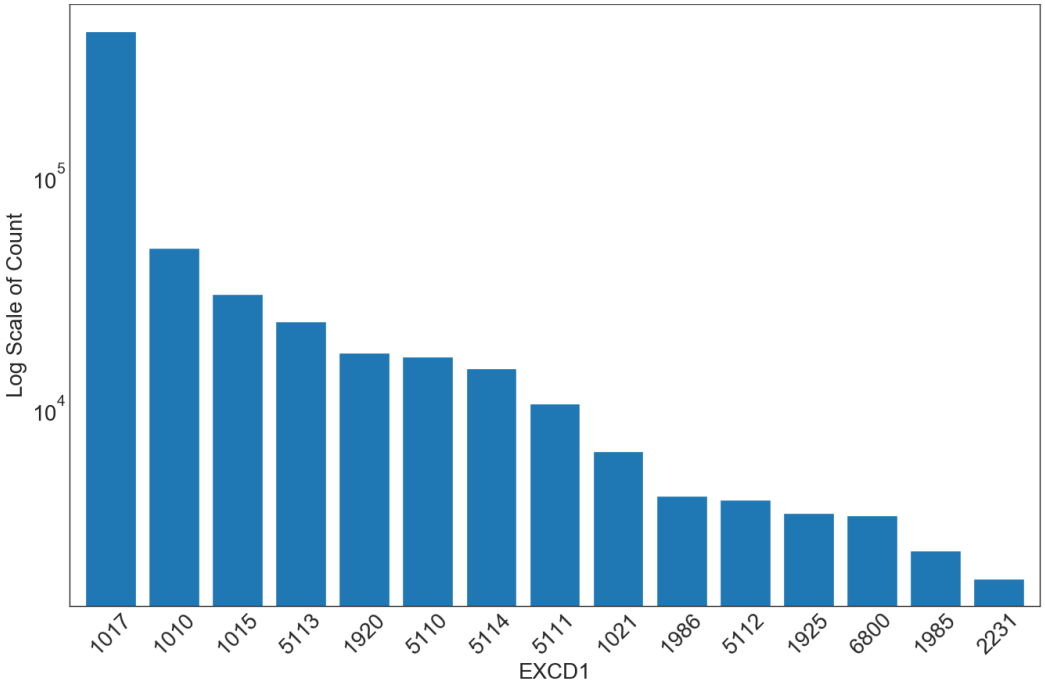


3.19. Field 19

Field Name: EXCD1
Field Type: Categorical
Description: exemption code 1.

Ranking by property count, the top 15 codes under exemption code 1 were included in the following plot. See *Figure 17.*

Figure 17: Exempt Code 1 Distribution (Top 15)

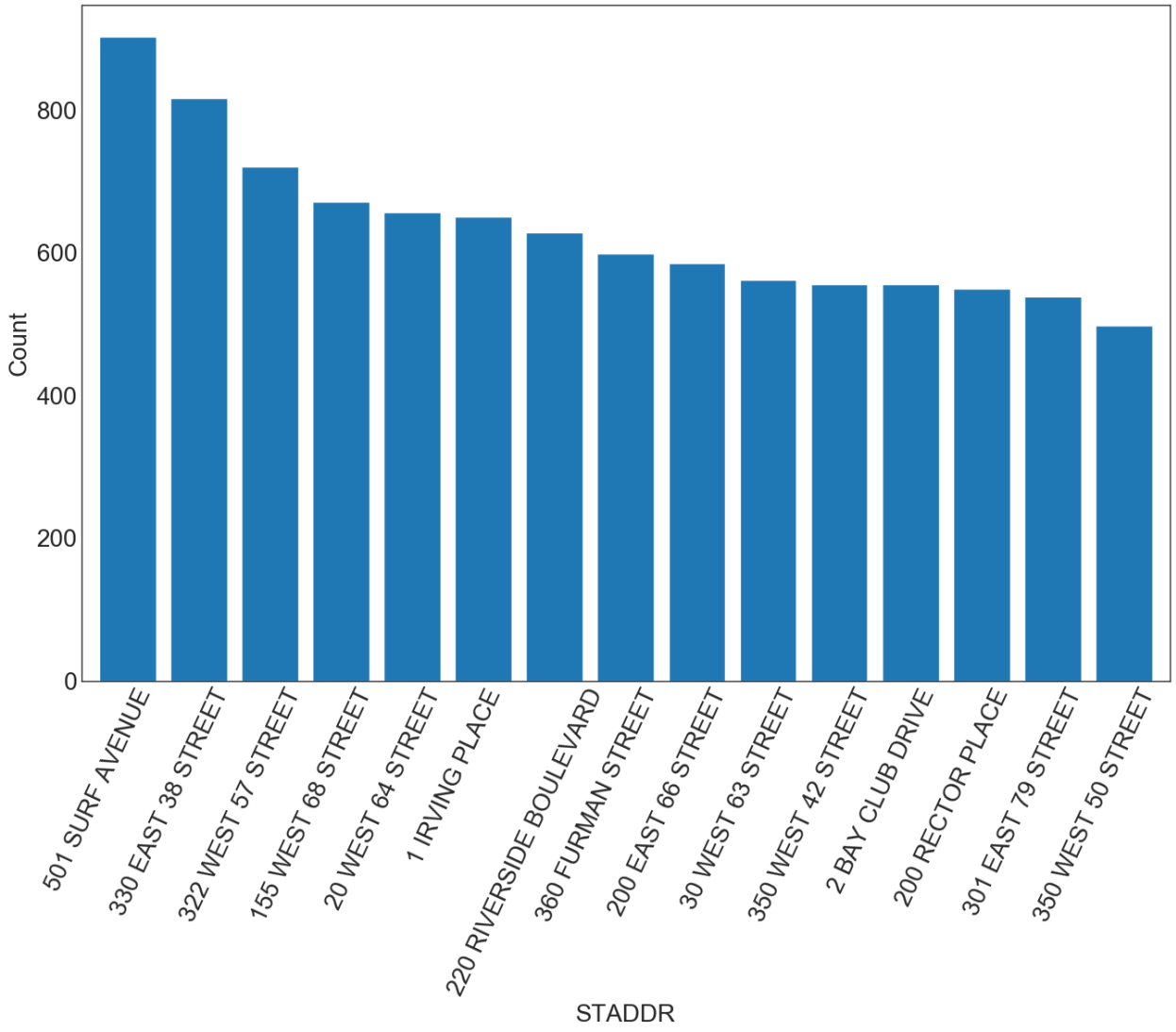


3.20. Field 20

Field Name: STADDR
Field Type: Categorical
Description: street address.

Ranking by property count, the top 15 street addresses were included in the following plot. See *Figure 18*.

Figure 18: Street Address Distribution (Top 15)

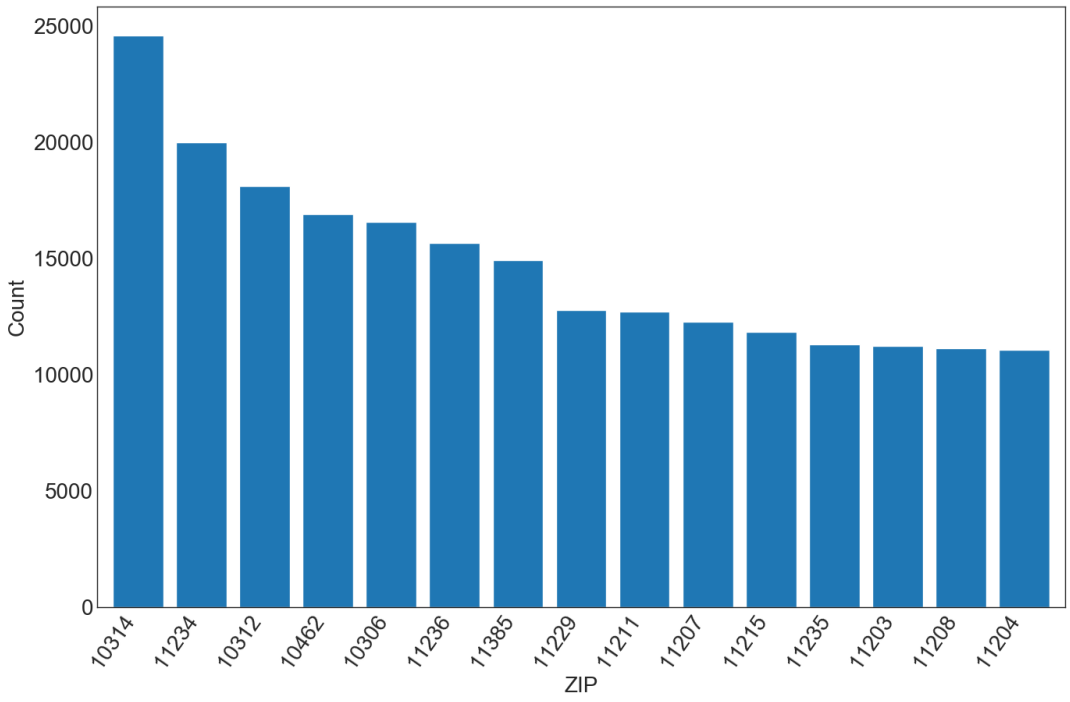


3.21. Field 21

Field Name: ZIP
Field Type: Categorical
Description: zip code of the property.

Ranking by property count, the top 15 zip codes were included in the following plot. See *Figure 19*.

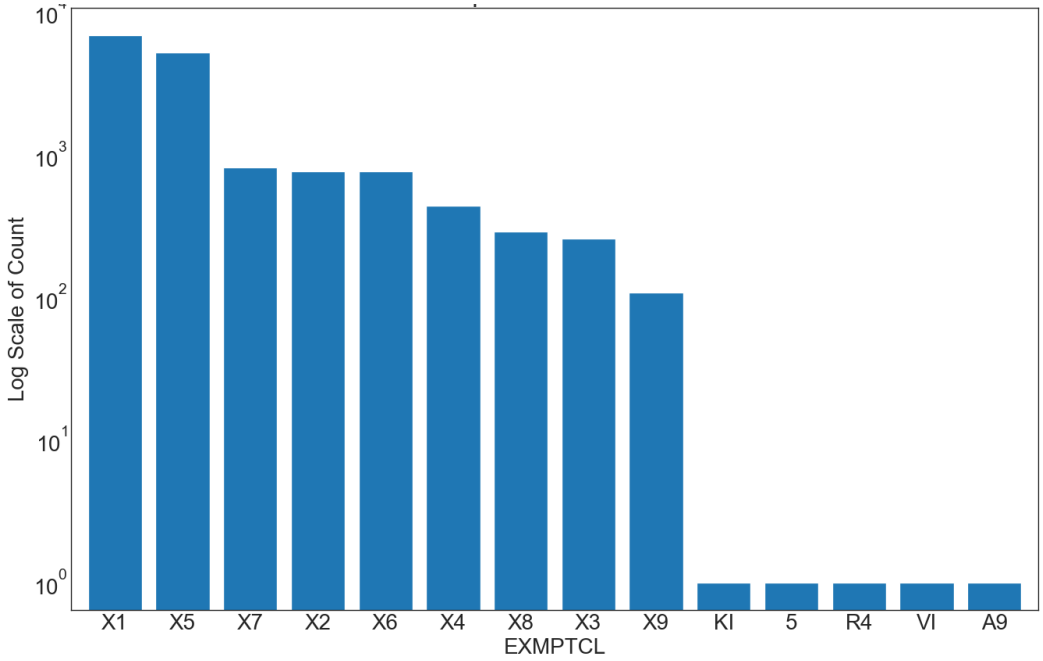
Figure 19: Zip Code Distribution (Top 15)



3.22. Field 22

Field Name: EXMPTCL
Field Type: Categorical
Description: exempt class. See *Figure 20*.

Figure 20: Exempt Class Distribution

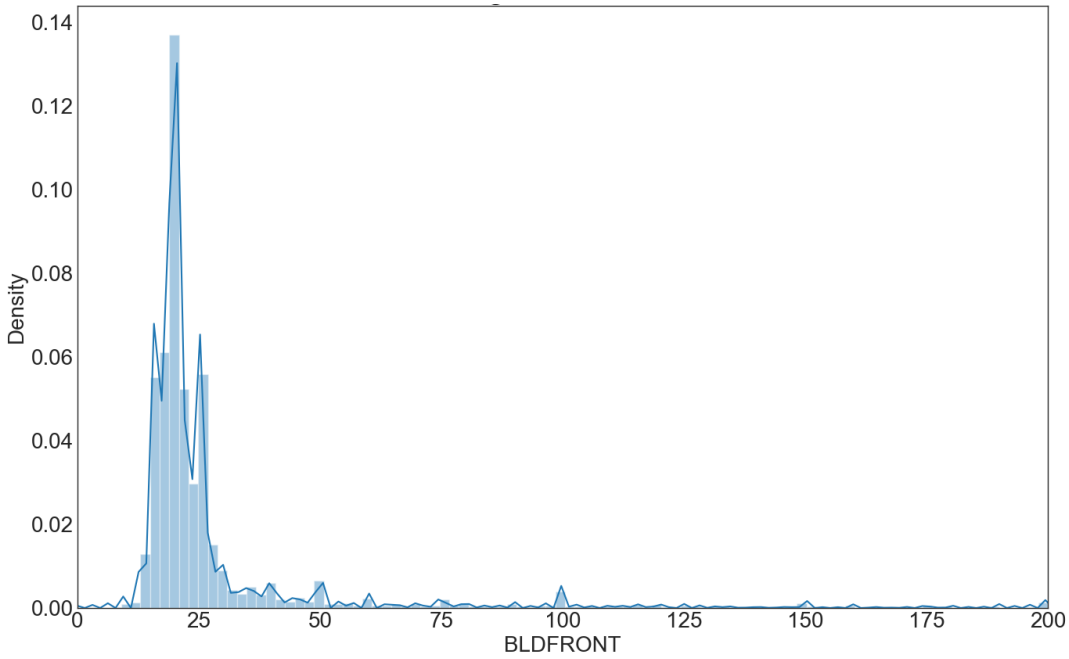


3.23. Field 23

Field Name: BLDFRONT
Field Type: Numerical
Description: building width.

Records with value 0 were removed from the plot below, considering that 0 building width has no practical meanings. Besides, records with building width over 200 were omitted from the plot due to a low volume of samples. See *Figure 21*.

Figure 21: Buidling Width Distribution

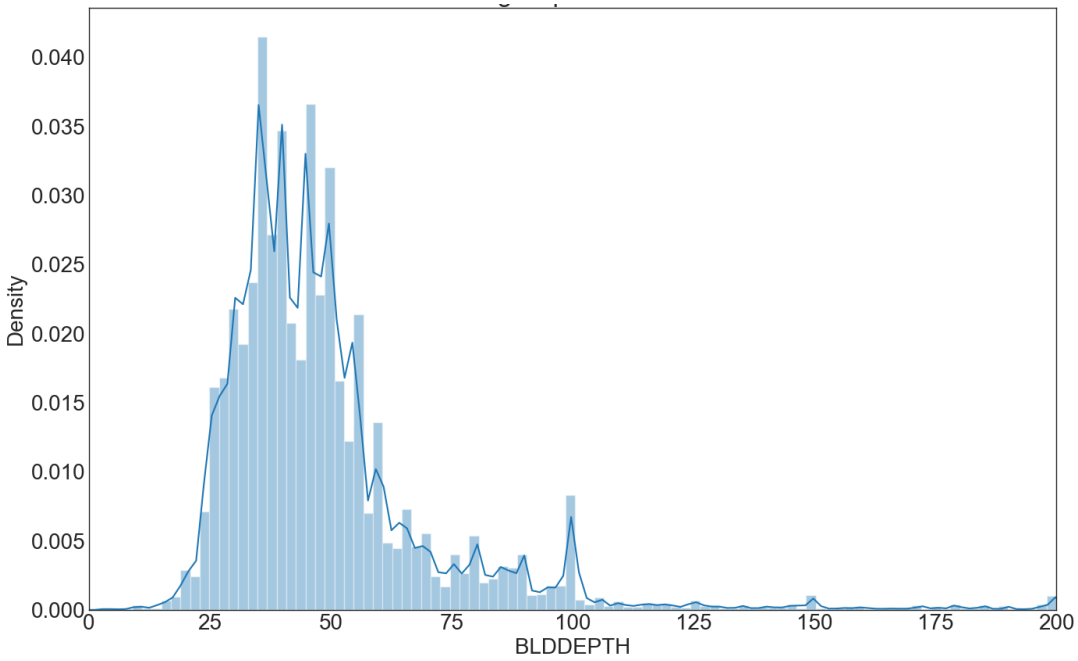


3.24. Field 24

Field Name: BLDDEPTH
Field Type: Numerical
Description: building depth.

Records with value 0 were removed from the plot below, considering that 0 building depth has no practical meanings. Besides, records with building depth over 200 were omitted from the plot due to a low volume of samples. See *Figure 22*.

Figure 22: Buidling Depth Distribution

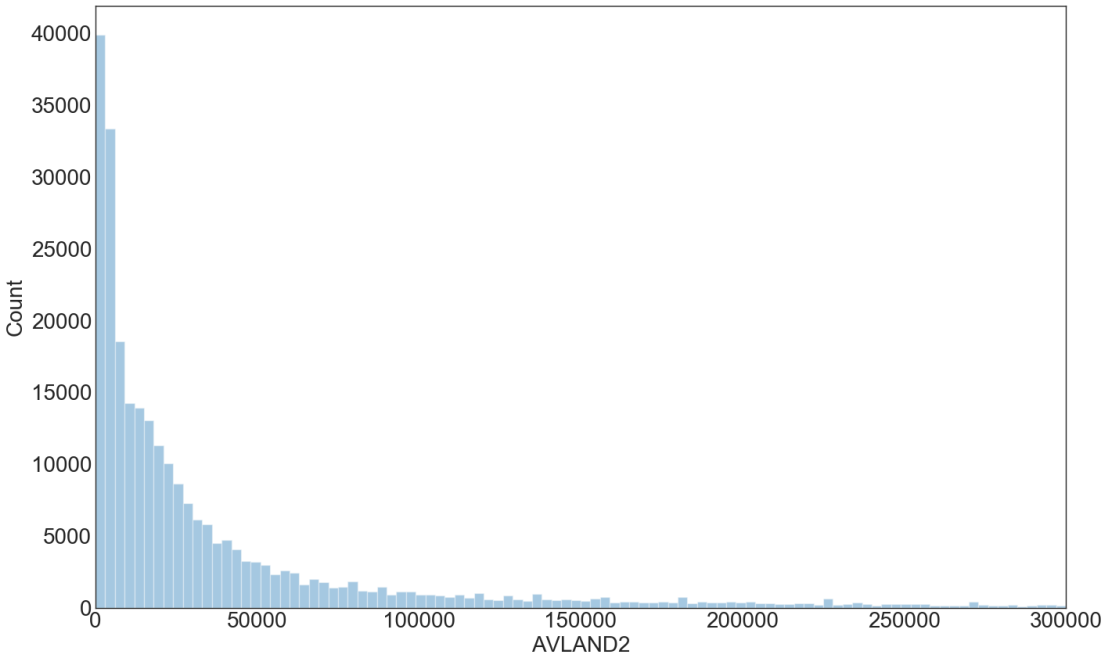


3.25. Field 25

Field Name: AVLAND2
Field Type: Numerical
Description: transitional land value.

Records with transitional land values over 300,000 were omitted from the plot due to low volume of samples. See *Figure 23*.

Figure 23: Transitional Land Value Distribution

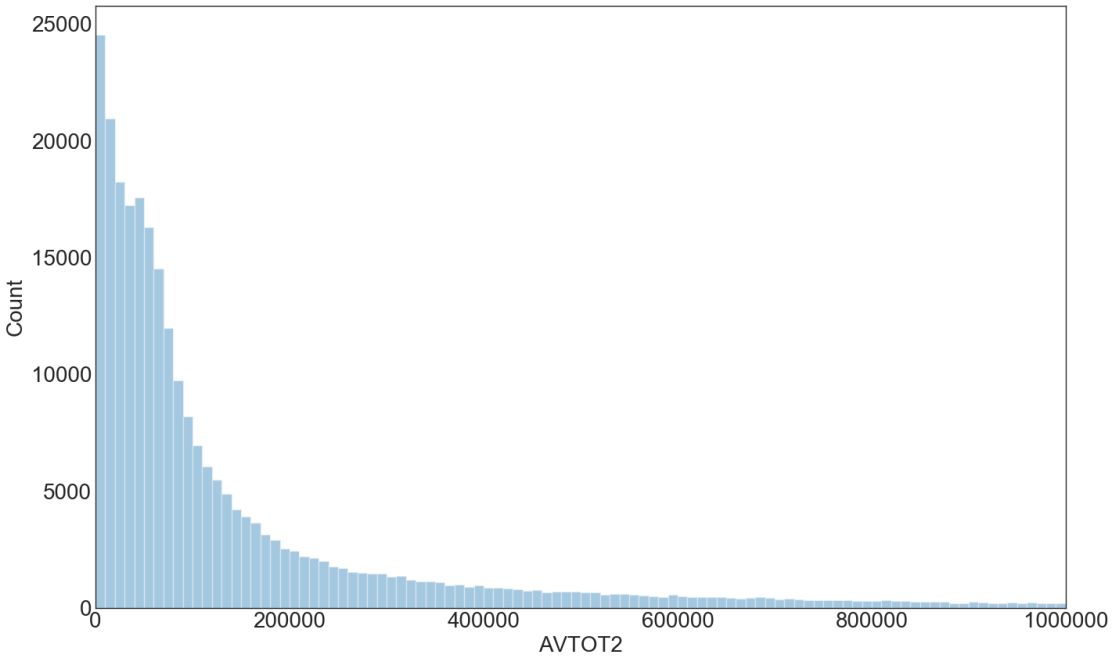


3.26. Field 26

Field Name: AVTOT2
Field Type: Numerical
Description: transitional total value.

Records with transitional total values over 300,000 were omitted from the plot due to low volume of samples. See *Figure 24*.

Figure 24: Transitional Total Value Distribution

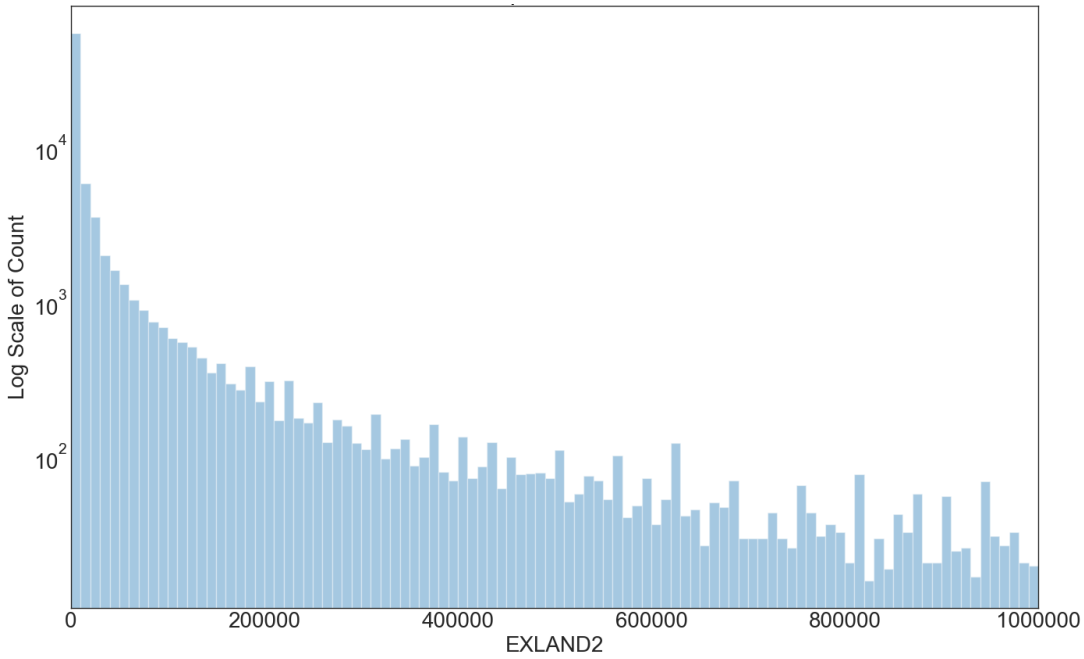


3.27. Field 27

Field Name: EXLAND2
Field Type: Numerical
Description: transitional exempt land value.

Records with actual exempt land values over 1,000,000 were omitted from the plot due to a low volume of samples. See *Figure 25*.

Figure 25: Transitionl Exempt Land Value Distribution

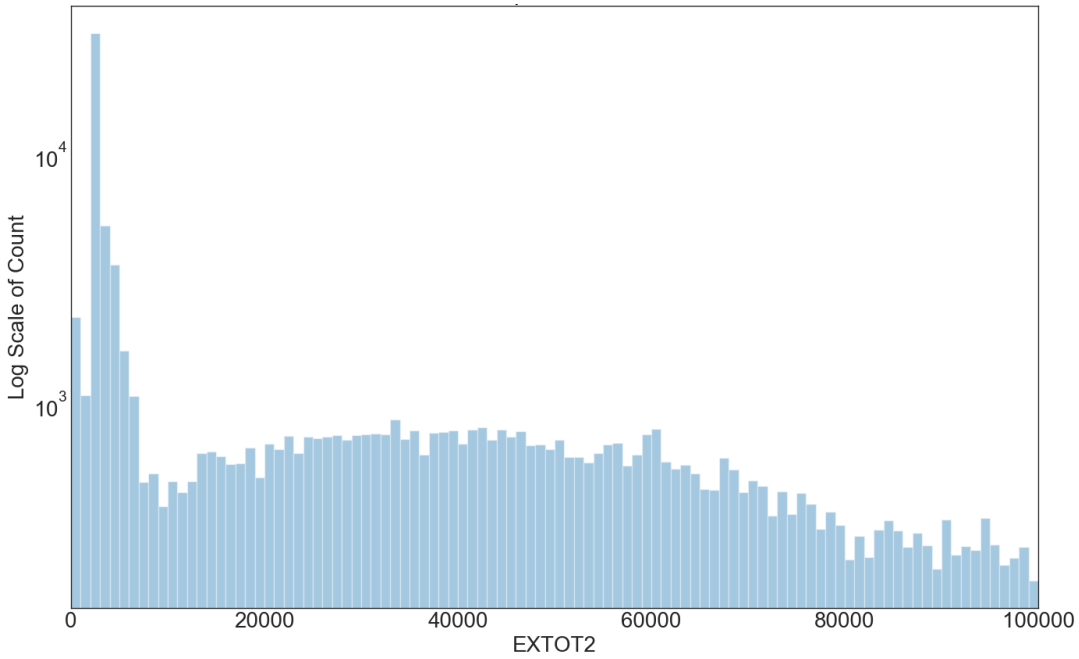


3.28. Field 28

Field Name: EXTOT2
Field Type: Numerical
Description: transitional exempt land total.

Records with transitional exempt land totals over 100,000 were omitted from the plot due to a low volume of samples. See *Figure 26*.

Figure 26: Transitional Expemt Land Total Distribution

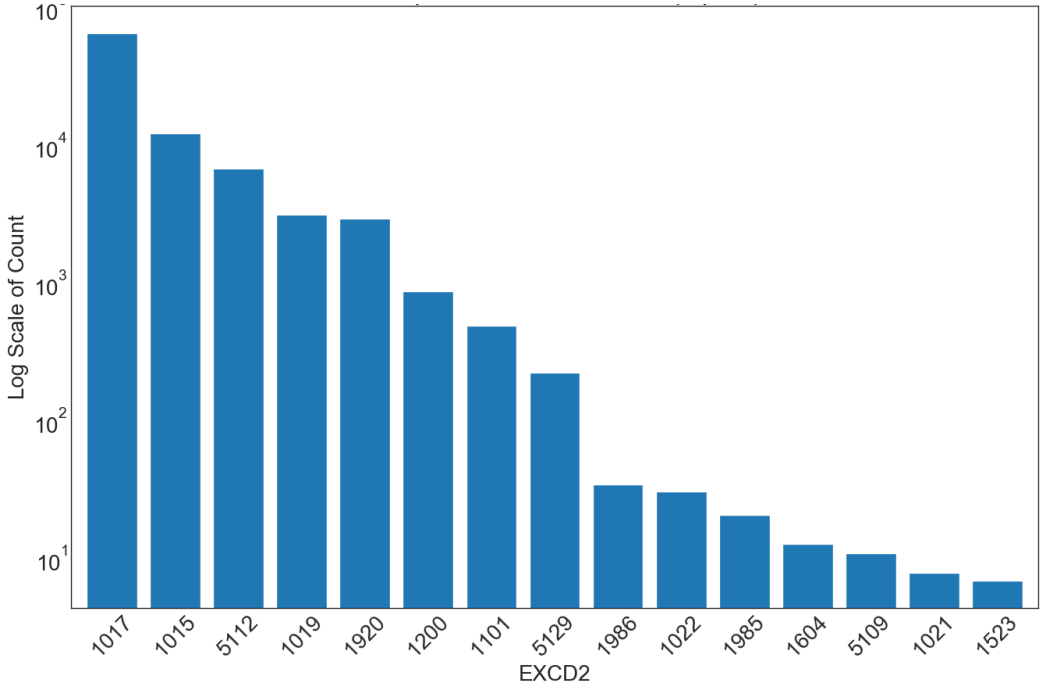


3.29. Field 29

Field Name: EXCD2
Field Type: Categorical
Description: exemption code 2.

Ranking by property count, the top 15 codes under exemption code 2 were included in the following plot. See *Figure 27*.

Figure 27: Expemt Code 2 Distribution (Top 15)



3.30. Field 30

Field Name: PERIOD

Field Type: Categorical

Description: assessment period when the file was created. There is only one value - 'Final' for all records and no missing value.

3.31. Field 31

Field Name: YEAR

Field Type: Categorical

Description: assessment year in which all records of properies were accessed. There is only one value - '2010/11' for all records and no missing value.

3.32. Field 32

Field Name: VALTYPE

Field Type: Categorical

Description: There is only one value - 'AC-TR' for all records and no missing value.

Appendix B. - The Hyperparameters of the Autoencoder

Function	Parameter Name	Description	Value
keras.models.Model.fit	epochs	How many times the entire dataset is passed forward and backward through the whole Autoencoder	100
	batch_size	The Number of samples per gradient update	1024
	shuffle	Whether to shuffle the training data before each epoch	True
keras.layers.Dense (for encoding)	activation	Method used to encode the input data into hidden layer	‘relu’
keras.layers.Dense (for decoding)	activation	Method used to decode encoded data in the hidden layer to output record	‘sigmoid’
keras.models.Model.compile	loss	Function used to evaluate model performance and select the best parameters	‘mean_squared_error’
	optimizer	Method used to change the attributes of the neural network such as weights and learning rate in order to reduce the losses.	‘adam’