New York Real Estate Project

FEB 22, 2018

Authored by:
Gyan Prakash, Yufei Wang, Wei Tang, Alok Abhishek, Weichen Zhang, Pratyush Shankar

Table of Contents

Executive Summary

As per a study conducted by Tax Justice Network, the US economy loses about $189 billion every year due to tax evasion.[1] Studying tax fraud therefore is very important in better understanding the drivers of tax fraud and ways to combat them.

This report specifically focuses on gaining insights into property tax fraud in New York City (NYC) using unsupervised machine learning techniques.[2] Over 1 million records corresponding to NYC tax data (2010-11) were used to analyze anomalies, develop insights and rank each property with fraud score(s). The following diagram illustrates the overall process.

For this project, fraud detection was done using weighted average of outlier detection using z-scores and autoencoder. Such a score was used to flag anomalous properties, which were then scrutinized manually for potential tax fraud. Based on the fraud score, a sizeable number of parks and government buildings were singled out. These records were excluded because such properties are large, low-story buildings on relatively large tracts of land; this makes them highly valuable but doesn't qualify them for tax fraud analysis purposes.

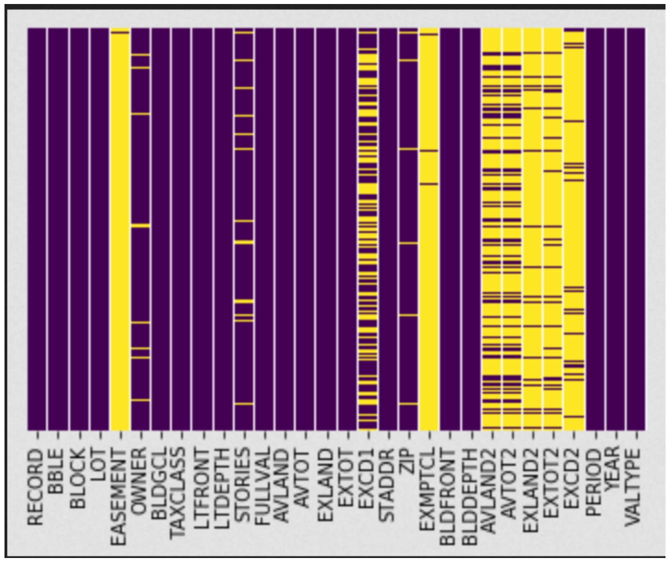
As for other candidates for tax evasion, the fraud algorithm flagged small buildings with very high value and large buildings with very low value as potential cases of fraud. Closer analysis of these records reveals that the full value of some of these properties is exceptionally high/low and they do not have complete building data (lot depth, lot front, etc., being missing fields). For some other records, the full value of the property per building area is either excessively high or low. A lot of these properties are owned by real estate firms and it can be inferred that either some of the properties they own have distinct characteristics as compared to an average property or they may be exploiting loopholes in tax property laws.

Description of Data

The New York City Department of Finance values properties in NYC every year to calculate property taxes. Their report provides property tax data such as market and assessed values, exemptions, abatements, etc., for the assessment year, 2010-11. The information is listed by categories such as borough, tax class, building type and so on. There are 1048575 records with 30 columns each – 14 of which are categorical variables and 16 numerical variables. The following table lists out descriptions for some important variables.

Abbreviation	Description
LTFRONT	Lot frontage in feet
LTDEPTH	Lot depth in feet
FULLVAL	Total market value of property
AVLAND	Market value of the land
AVTOT	Total market value
EXLAND	Exempt land value
EXTOT	Exempt total value
EXCD1	Exempt condo value
BLDFRONT	Building frontage in feet
BLDDEPTH	Building depth in feet
AVLAND2	2nd market value of the land
AVTOT2	2nd total market value
EXLAND2	2nd exempt land value
EXTOT2	2nd exempt total value
EXCD2	2nd exempt condo value
BLDGCL	Building class

A reality associated with real world data is the issue of missing values. The following heatmap helps visualize missing values in the data set. Yellow regions represent missing values.



Descriptive statistics of the overall data has been shown below.

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
RECORD	int64	1,048,575								100.00%	
BLOCK	int64	1,048,575	4,708.87	3,699.55	1.00	1,534.00	3,944.00	6,797.00	16,350	100.00%	13,949
LOT	int64	1,048,575	370.09	860.54	1.00	23.00	49.00	146.00	9,978	100.00%	6,366
LTFRONT	int64	1,048,575	36.17	73.73	0.00	19.00	25.00	40.00	9,999	100.00%	1,277
LTDEPTH	int64	1,048,575	88.28	75.48	0.00	80.00	100.00	100.00	9,999	100.00%	1,336
STORIES	float64	996,433	5.06	8.43	1.00	2.00	2.00	3.00	119	95.03%	111
FULLVAL	int64	1,048,575	880,487.66	11,702,930.00	0.00	303,000.00	446,000.00	619,000.00	6,150,000,000	100.00%	108,277
AVLAND	int64	1,048,575	85,995.03	4,100,755.00	0.00	9,160.00	13,646.00	19,706.00	2,668,500,000	100.00%	70,529
AVTOT	int64	1,048,575	230,758.18	6,951,206.00	0.00	18,385.00	25,339.00	46,095.00	4,668,309,000	100.00%	112,294
EXLAND	int64	1,048,575	36,811.79	4,024,330.00	0.00	0.00	1,620.00	1,620.00	2,668,500,000	100.00%	33,186
EXTOT	int64	1,048,575	92,543.81	6,578,281.00	0.00	0.00	1,620.00	2,090.00	4,668,309,000	100.00%	63,805
EXCD1	float64	622,642	1,604.50	1,388.13	1,010.00	1,017.00	1,017.00	1,017.00	7,170	59.38%	129
ZIP	float64	1,022,219	10,935.32	526.58	10,001.00	10,453.00	11,215.00	11,364.00	33,803	97.49%	196
BLDFRONT	int64	1,048,575	23.02	35.79	0.00	15.00	20.00	24.00	7,575	100.00%	610
BLDDEPTH	int64	1,048,575	40.07	43.04	0.00	26.00	39.00	51.00	9,393	100.00%	620
AVLAND2	float64	280,966	246,365.48	6,199,390.00	3.00	5,705.00	20,059.00	62,338.75	2,371,005,000	26.80%	58,169
AVTOT2	float64	280,972	716,078.71	11,690,170.00	3.00	34,013.50	80,010.00	240,792.00	4,501,180,000	26.80%	110,890
EXLAND2	float64	86,675	351,802.21	10,852,480.00	1.00	2,090.00	3,053.00	31,419.00	2,371,005,000	8.27%	21,996
EXTOT2	float64	129,933	658,114.78	16,129,810.00	7.00	2,889.00	37,116.00	106,629.00	4,501,180,000	12.39%	48,106
EXCD2	float64	90,941	1,371.66	1,105.49	1,011.00	1,017.00	1,017.00	1,017.00	7,160	8.67%	60
EASEMENT	object	4,043								0.39%	12
OWNER	object	1,017,492								97.04%	847,053
BLDGCL	object	1,048,575								100.00%	200
TAXCLASS	object	1,048,575								100.00%	11
STADDR	object	1,047,934								99.94%	820,637
EXMPTCL	object	14,992								1.43%	14
PERIOD	object	1,048,575								100.00%	1
YEAR	object	1,048,575								100.00%	1
VALTYPE	object	1,048,575								100.00%	1

Data Cleaning

Data Imputation:

❖ ZIP

Cells with missing values were filled with 10935 since, that number represents the average of all non-missing ZIP values.

❖ LTFRONT, LOTDEPTH, BLDFRONT, BLDEFPTH

Missing and zero values for the above variables were replaced by their averages i.e., 40, 100, 30 and 50, respectively.

❖ FULLVAL, AVLAND, AVTOT

Missing and zero values for the above variables were replaced by their averages rounded to the closest thousands i.e., 880000, 86000 and 230000, respectively.

❖ EXLAND, EXTOT

Filling in missing values and replace zero values by 1620.
Reason: 1620 is the mode for EXLAND and EXTOT columns. 33.1% EXLAND value is 1620; 32.8% value of EXTOT is 1620.

❖ STORIES

Records with zero story values were replaced by average number of stories for the ZIP in which the building was located. A table with average number of stories per ZIP has been provided below for reference.

ZIP	STORIES	ZIP	STORIES	ZIP	STORIES	ZIP	STORIES
10001	11	10302	2	11201	11	11362	2
10002	6	10303	2	11203	2	11363	2
10003	10	10304	2	11204	2	11364	2
10004	36	10305	2	11205	4	11365	2
10005	33	10306	2	11206	4	11366	2
10006	32	10308	2	11207	3	11367	3
10007	14	10310	2	11208	3	11368	3
10009	6	10312	2	11209	3	11369	2
10010	21	10314	2	11210	3	11370	2
10011	10	10451	4	11211	6	11372	3
10012	6	10452	3	11212	2	11373	3

10013	8	10453	3	11213	3	11374	6
10014	9	10454	3	11214	3	11375	4
10016	25	10455	3	11215	4	11377	2
10017	30	10456	3	11216	3	11378	2
10018	22	10457	4	11217	4	11379	2
10019	33	10458	3	11218	3	11385	2
10020	48	10459	3	11219	3	11411	2
10021	24	10460	3	11220	3	11412	2
10022	25	10461	2	11221	3	11413	2
10023	26	10462	8	11222	3	11414	2
10024	10	10463	5	11223	2	11415	4
10025	14	10464	2	11224	10	11416	2
10026	10	10465	2	11225	3	11417	2
10027	5	10466	2	11226	3	11418	3
10028	15	10467	3	11228	2	11419	2
10029	6	10468	4	11229	2	11420	2
10030	6	10469	2	11230	3	11421	2
10031	5	10470	2	11231	3	11422	2
10032	6	10471	3	11232	3	11423	2
10033	5	10472	2	11233	3	11426	2
10034	5	10473	3	11234	2	11427	2
10035	7	10474	2	11235	4	11428	2
10036	34	10475	2	11236	2	11429	2
10037	4	10803	3	11237	3	11432	3
10038	19	11001	2	11238	4	11433	2
10039	8	11004	2	11239	3	11434	2
10040	6	11040	2	11243	41	11435	3
10044	16	11101	6	11354	5	11436	2
10065	23	11102	6	11355	6	11691	3
10069	35	11103	3	11356	2	11692	2
10075	22	11104	3	11357	2	11693	3
10128	25	11105	2	11358	2	11694	3
10280	27	11106	3	11360	6	10935	4
10301	3	11109	18	11361	2		

Expert Variables

Variable	Name	Description
1	fv_la	Average FULLVAL per LOTAREA (LTFRONT*LTDEPTH)
2	vl_la	Average AVLAND per LOTAREA (LTFRONT*LTDEPTH)
3	vt_la	Average AVTOT per LOTAREA (LTFRONT*LTDEPTH)
4	xl_la	Average EXLAND per LOTAREA (LTFRONT*LTDEPTH)
5	xt_la	Average EXTOT per LOTAREA (LTFRONT*LTDEPTH)
6	fv_ba	Average FULLVAL per BLDAREA (BLDFRONT*BLDDEPTH)
7	vl_ba	Average AVLAND per BLDAREA (BLDFRONT*BLDDEPTH)
8	vt_ba	Average AVTOT per BLDAREA (BLDFRONT*BLDDEPTH)
9	xl_ba	Average EXLAND per BLDAREA (BLDFRONT*BLDDEPTH)
10	xt_ba	Average EXTOT per BLDAREA (BLDFRONT*BLDDEPTH)
11	fv_bv	Average FULLVAL per BLDVOL (BLDAREA*STORIES)
12	vl_bv	Average AVLAND per BLDVOL (BLDAREA*STORIES)
13	vt_bv	Average AVTOT per BLDVOL (BLDAREA*STORIES)
14	xl_bv	Average EXLAND per BLDVOL (BLDAREA*STORIES)
15	xt_bv	Average EXTOT per BLDVOL (BLDAREA*STORIES)
16	fv_la_z3	Ratio of fv_la and Average fv_la grouped by ZIP3
17	vl_la_z3	Ratio of vl_la and Average vl_la grouped by ZIP3
18	vt_la_z3	Ratio of vt_la and Average vt_la grouped by ZIP3
19	xl_la_z3	Ratio of xl_la and Average xl_la grouped by ZIP3
20	xt_la_z3	Ratio of xt_la and Average xt_la grouped by ZIP3
21	fv_ba_z3	Ratio of fv_ba and Average fv_ba grouped by ZIP3
22	vl_ba_z3	Ratio of vl_ba and Average vl_ba grouped by ZIP3
23	vt_ba_z3	Ratio of vt_ba and vt_ba Average grouped by ZIP3
24	xl_ba_z3	Ratio of xl_ba and Average xl_ba grouped by ZIP3
25	xt_ba_z3	Ratio of xt_ba and Average xt_ba grouped by ZIP3
26	fv_bv_z3	Ratio of fv_bv and Average fv_bv grouped by ZIP3
27	vl_bv_z3	Ratio of vl_bv and Average vl_bv grouped by ZIP3
28	vt_bv_z3	Ratio of vt_bv and Average vt_bv grouped by ZIP3
29	xl_bv_z3	Ratio of xl_bv and Average xl_bv grouped by ZIP3
30	xt_bv_z3	Ratio of xt_bv and Average xt_bv grouped by ZIP3

31	fv_la_z5	Ratio of fv_la and Average fv_la grouped by ZIP5
32	vl_la_z5	Ratio of vl_la and Average vl_la grouped by ZIP5
33	vt_la_z5	Ratio of vt_la and Average vt_la grouped by ZIP5
34	xl_la_z5	Ratio of xl_la and Average xl_la grouped by ZIP5
35	xt_la_z5	Ratio of xt_la and Average xt_la grouped by ZIP5
36	fv_ba_z5	Ratio of fv_ba and Average fv_ba grouped by ZIP5
37	vl_ba_z5	Ratio of vl_ba and Average vl_ba grouped by ZIP5
38	vt_ba_z5	Ratio of vt_ba and vt_ba Average grouped by ZIP5
39	xl_ba_z5	Ratio of xl_ba and Average xl_ba grouped by ZIP5
40	xt_ba_z5	Ratio of xt_ba and Average xt_ba grouped by ZIP5
41	fv_bv_z5	Ratio of fv_bv and Average fv_bv grouped by ZIP5
42	vl_bv_z5	Ratio of vl_bv and Average vl_bv grouped by ZIP5
43	vt_bv_z5	Ratio of vt_bv and Average vt_bv grouped by ZIP5
44	xl_bv_z5	Ratio of xl_bv and Average xl_bv grouped by ZIP5
45	xt_bv_z5	Ratio of xt_bv and Average xt_bv grouped by ZIP5
46	fv_la_tc	Ratio of fv_la and Average fv_la grouped by TAXCLASS
47	vl_la_tc	Ratio of vl_la and Average vl_la grouped by TAXCLASS
48	vt_la_tc	Ratio of vt_la and Average vt_la grouped by TAXCLASS
49	xl_la_tc	Ratio of xl_la and Average xl_la grouped by TAXCLASS
50	xt_la_tc	Ratio of xt_la and Average xt_la grouped by TAXCLASS
51	fv_ba_tc	Ratio of fv_ba and Average fv_ba grouped by TAXCLASS
52	vl_ba_tc	Ratio of vl_ba and Average vl_ba grouped by TAXCLASS
53	vt_ba_tc	Ratio of vt_ba and vt_ba Average grouped by TAXCLASS
54	xl_ba_tc	Ratio of xl_ba and Average xl_ba grouped by TAXCLASS
55	xt_ba_tc	Ratio of xt_ba and Average xt_ba grouped by TAXCLASS
56	fv_bv_tc	Ratio of fv_bv and Average fv_bv grouped by TAXCLASS
57	vl_bv_tc	Ratio of vl_bv and Average vl_bv grouped by TAXCLASS
58	vt_bv_tc	Ratio of vt_bv and Average vt_bv grouped by TAXCLASS
59	xl_bv_tc	Ratio of xl_bv and Average xl_bv grouped by TAXCLASS
60	xt_bv_tc	Ratio of xt_bv and Average xt_bv grouped by TAXCLASS
61	fv_la_bo	Ratio of fv_la and Average fv_la grouped by BOROUGH
62	vl_la_bo	Ratio of vl_la and Average vl_la grouped by BOROUGH

63	vt_la_bo	Ratio of vt_la and Average vt_la grouped by BOROUGH
64	xl_la_bo	Ratio of xl_la and Average xl_la grouped by BOROUGH
65	xt_la_bo	Ratio of xt_la and Average xt_la grouped by BOROUGH
66	fv_ba_bo	Ratio of fv_ba and Average fv_ba grouped by BOROUGH
67	vl_ba_bo	Ratio of vl_ba and Average vl_ba grouped by BOROUGH
68	vt_ba_bo	Ratio of vt_ba and vt_ba Average grouped by BOROUGH
69	xl_ba_bo	Ratio of xl_ba and Average xl_ba grouped by BOROUGH
70	xt_ba_bo	Ratio of xt_ba and Average xt_ba grouped by BOROUGH
71	fv_bv_bo	Ratio of fv_bv and Average fv_bv grouped by BOROUGH
72	vl_bv_bo	Ratio of vl_bv and Average vl_bv grouped by BOROUGH
73	vt_bv_bo	Ratio of vt_bv and Average vt_bv grouped by BOROUGH
74	xl_bv_bo	Ratio of xl_bv and Average xl_bv grouped by BOROUGH
75	xt_bv_bo	Ratio of xt_bv and Average xt_bv grouped by BOROUGH

Techniques

To start with, we examined the fields in the raw data and performed the following steps to set the stage for further analysis.

a. Preliminary exploratory analysis
b. Data cleaning, standardization
c. Transformation to create expert variables; creating such variables generally requires domain expertise
d. Encoding of categorical variables, creation of risk tables, z-scaling, other normalizations and outlier suppression techniques, nonlinear transformations such as taking log or binning, construction of ratios or products of fields

- **Feature Selection, Variable Reduction and Dimensionality Reduction**
  During this process, we focused on reducing the number of inputs to a model by considering which inputs are the most important to the model. Most modeling methods degrade when burdened with more inputs than are needed for a robust, stable model.

- **Model Validation**
  During this phase, the data is generally separated into multiple sets to ensure robustness of the model. It is a good modeling practice to reserve a set of data that is never used during training/testing; instead it is used for evaluation of the model on data that it has never seen before. Such holdout samples can be very useful to validate unsupervised learning models.

- **Boosting**
  It refers to an iterative procedure to create a series of weak models, where the final model is then a linear combination of the series of weak models. In theory, the next model in the series is trained on a weighted data set, where the records with the largest error are more heavily weighted.

- **Bagging**
  The term *bagging* comes from *bootstrap aggregation*. This is a technique to improve model stability and accuracy. It combines/aggregates the outcomes of many models, each having been built via bootstrap sampling.

Analysis

❖ We built 75 expert variables by various scaling processes based on five value fields, namely, FULLVAL, AVLAND, AVTOT, EXLAND and EXTOT. The purpose behind building such variables was to identify the ones which are strong predictors of fraud, by narrowing down to a smaller group of variables. In short, we built expert variables that quantify signals for various fraud modes.

$$\begin{array}{cc} & \sim 10^1 \\ \sim 10^6 & \begin{pmatrix} \text{Original} \\ \text{data} \end{pmatrix} \end{array} \longrightarrow \begin{array}{cc} & \sim 10^2 \\ \sim 10^6 & \begin{pmatrix} \text{Expert} \\ \text{variables} \\ \text{(x's)} \end{pmatrix} \end{array}$$

❖ Next, we z-scaled all variables for feature selection/dimensionality reduction.

$$\begin{array}{cc} & \sim 10^2 \\ \sim 10^6 & \begin{pmatrix} x1 & \dots & xn \\ \vdots & & \vdots \\ x1 & \dots & xn \end{pmatrix} \end{array} \longrightarrow \begin{array}{cc} & \sim 10^2 \\ \sim 10^6 & \begin{pmatrix} z1 & \dots & zn \\ \vdots & & \vdots \\ z1 & \dots & zn \end{pmatrix} \end{array} \qquad z_i = \frac{x_i - \mu_i}{\sigma_i}$$

❖ Principal Components Analysis (PCA) was performed on those 75 expert variables after which we decided to retain 8 PCs for further analysis. This way, we were able to lose dimensionality significantly without sacrificing variance explained.

$$\begin{array}{cc} & \sim 10^2 \\ \sim 10^6 & \begin{pmatrix} z1 & \dots & zn \\ \vdots & & \vdots \\ z1 & \dots & zn \end{pmatrix} \end{array} \longrightarrow \begin{array}{cc} & \sim 10^1 \\ \sim 10^6 & \begin{pmatrix} PC_1 & \dots & PC_n \\ \vdots & & \vdots \\ PC_1 & \dots & PC_n \end{pmatrix} \end{array} \qquad \begin{array}{c} PC_1 = \Sigma_i a_i z_i \\ PC_2 = \Sigma_i b_i z_i \\ \dots \end{array}$$

❖ The 8 PCs thus obtained were z-scaled again.

$$\begin{array}{cc} & \sim 10^1 \\ \sim 10^6 & \begin{pmatrix} PC_1 & \dots & PC_n \\ \vdots & & \vdots \\ PC_1 & \dots & PC_n \end{pmatrix} \end{array} \longrightarrow \begin{array}{cc} & \sim 10^1 \\ \sim 10^6 & \begin{pmatrix} z1 & \dots & zn \\ \vdots & & \vdots \\ z1 & \dots & zn \end{pmatrix} \end{array} \qquad z_i = \frac{PC_i - \mu_{PC_i}}{\sigma_{PC_i}}$$

❖ We used z-scores from the previous step to combine the model variables using a heuristic algorithm that used the sum of absolute z-scores, z-scores squared, maximum/minimum values, etc.

❖ Two fraud algorithms were built. The mathematical expressions for calculating the fraud scores corresponding to the two methods have been shown below.

- Outlier detection via z-scores

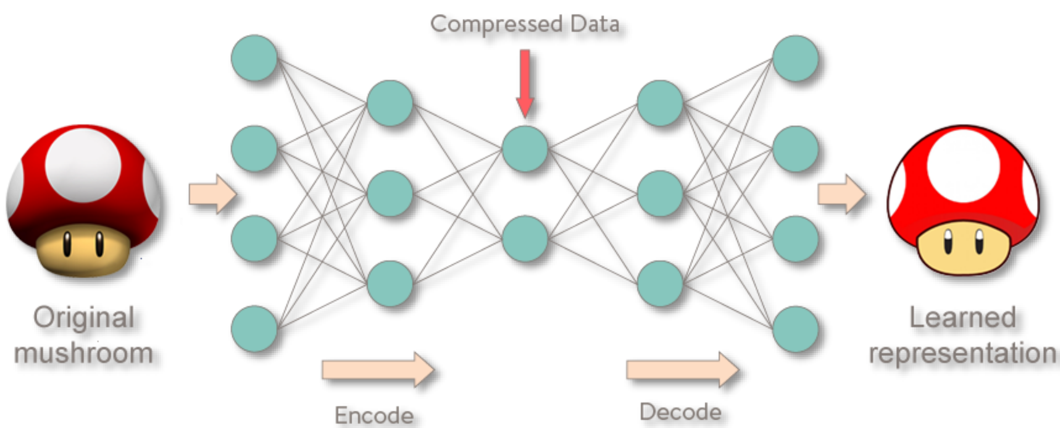
$$S = \left(\Sigma_i \left|z_i\right|^n\right)^{1/n}$$

- Autoencoder error

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

The fraud algorithms are constructed with the purpose of evaluating the fraud score. We do this by training the model to reproduce the original data. The reproduction error is a measure of the record's unusualness and thus, a fraud score.

For a more intuitive understanding of autoencoders, an illustration has been shown below.



The general expression for fraud score of a given record can be written as a function of z-scores.

- Outlier detection via z-scores

$$\text{Score for record i is } s_i = \left(\Sigma_k \left|z_k^i\right|^m\right)^{1/m}, \quad m \text{ anything}$$

- Autoencoder error

$$\text{Score for record i is } s_i = \left(\Sigma_k \left|{z'}_k^i - z_k^i\right|^m\right)^{1/m}, \quad m \text{ anything}$$

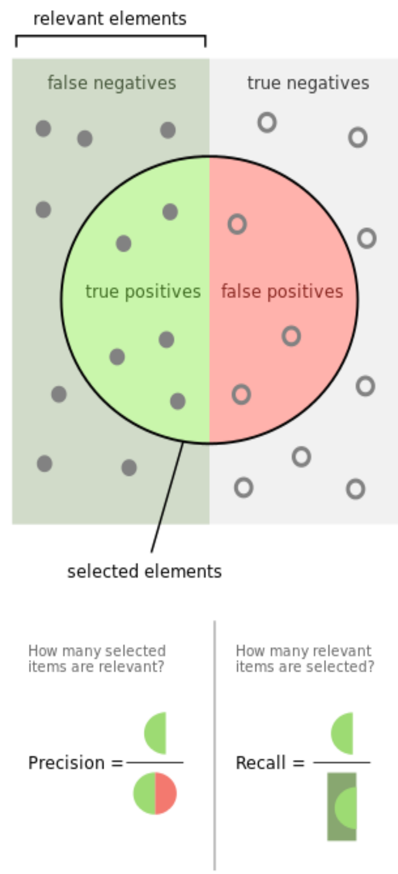
❖ In the final step, we combined the fraud scores via weighted quantile-scaling.

Examining Records: Precision versus Recall

Precision and recall are defined as under:

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

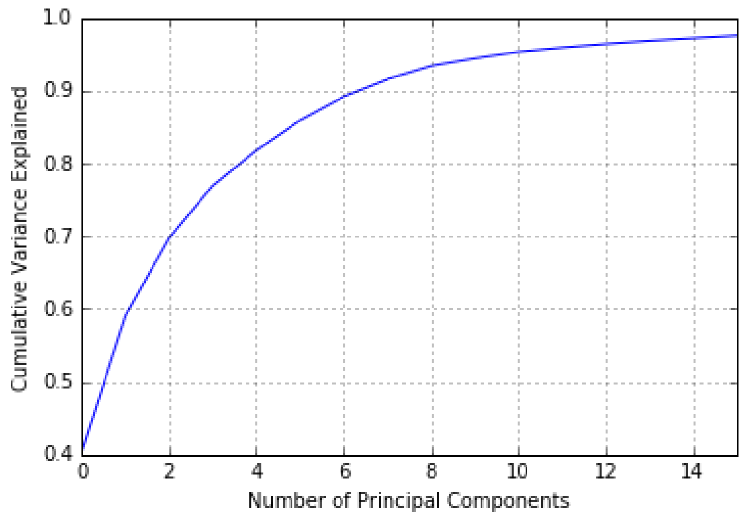


Precision measures the relevancy of obtained results. Recall, on the other hand, measures how many relevant results are returned. Both values can take values between 0 and 1.

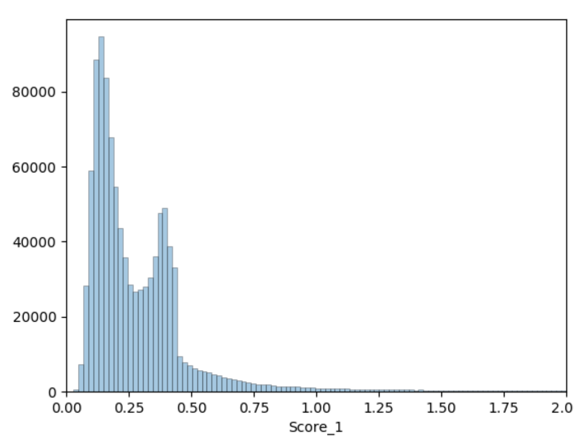
High recall but low precision implies a multitude of results, most of which have low or no relevancy. When precision is high, but recall is low, the converse happens – few returned results with very high relevancy. Ideally, we want high precision and high recall — many results, most of which are highly relevant. An illustration has been provided alongside for a better understanding of precision and recall.

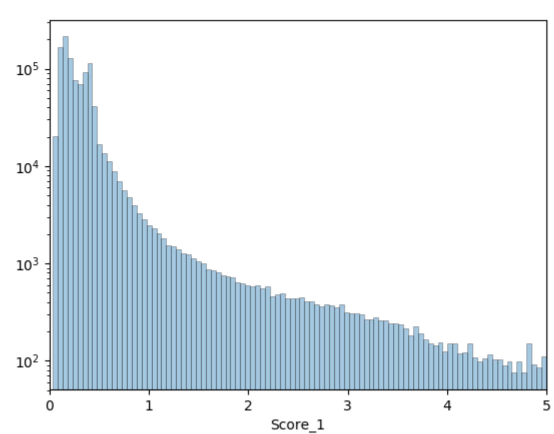
Results and Observations

Starting with over 75 variables, we reduced the number of variables to 8 through PCA (dimensionality reduction). These variables explained more than 93% of the variance in the data. A plot of number of principal components against cumulative variance explained (courtesy of PCA) has been shown below.

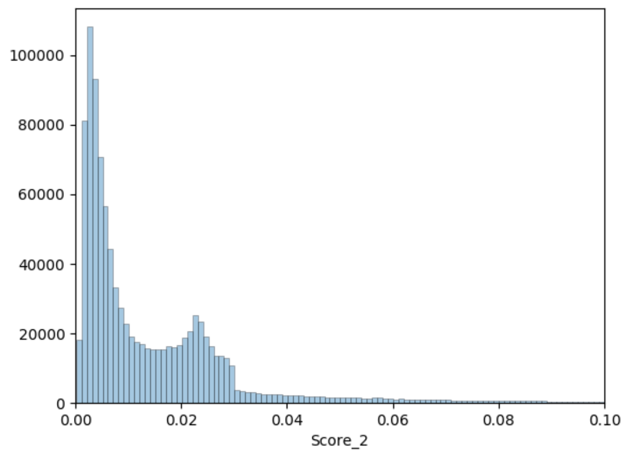


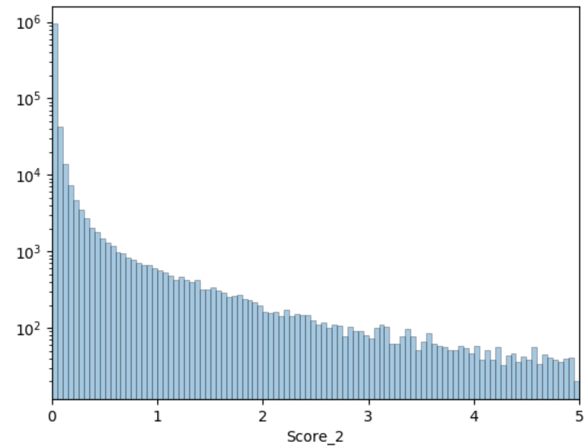
Fraud score calculated based on outlier detection via z-scores algorithm exhibited a right-skewed distribution (refer to the illustrations below). As expected, most of the records had low fraud scores and there were few outliers relative to the number of total records.



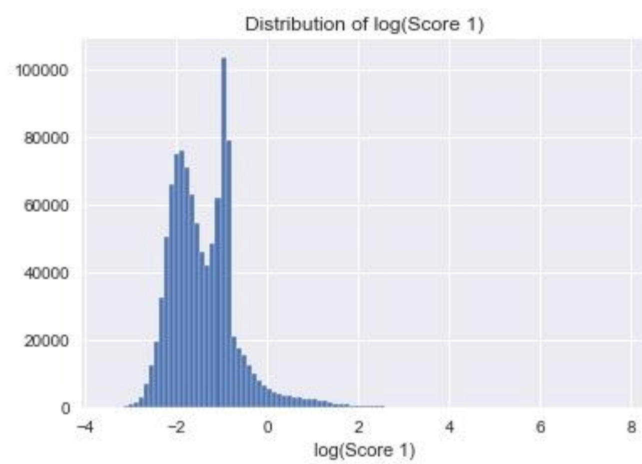


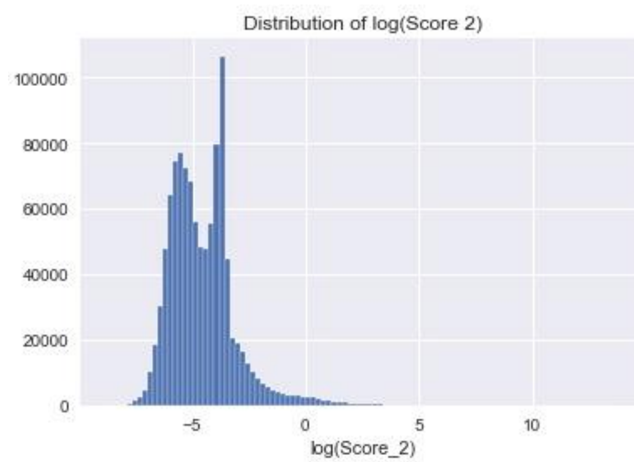
The distribution of fraud scores calculated using autoencoder algorithm also had a right skew (refer to the illustrations below). Once again, there were few records, relatively speaking, which had high fraud scores.





The log-scaled distributions of both scores show very similar patterns as can be seen from the following plots.





It is to be noted here that autoencoder uses Euclidean distance, whereas outlier detection algorithm relies on Manhattan distance, for calculating fraud scores. We used the two scores to create a weighted average fraud score. For weighted average score, we allotted 75% and 25% weights to scores obtained via autoencoder and outlier detection, respectively. The rationalization behind the weights being that after PCA, the dimensionality of the data was not a concern and our primary purpose was to focus on anomaly detection. Since, Euclidean distance represents the distance in hyperplane in between two points, we wanted to emphasize the abnormality using this feature and therefore, assigned higher weight to autoencoder fraud score.

We manually examined the records with high fraud scores. We filtered out the government properties and universities because although these properties are anomalies on several fronts, we know that government properties, parks and universities have very low risk of tax fraud. The top candidates for potential tax fraud have been listed below.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	STADDR	ZIP	BLDFRONT	BLDDEPTH
864163 REALTY, LLC	D9	2	157.00	95.00	1.00	$ 2,930,000.00	1,318,500.00	1,318,500.00	86-55 BROADWAY	11373	1	1
LOGAN PROPERTY, INC.	T1	4	4,910.00	-	3.00	$ 374,019,883.00	1,792,808,947.00	4,668,308,947.00	154-68 BROOKVILLE BOULEVARD	11422	0	0
RICH-NICH REALTY,LLC	D3	2	136.00	132.00	8.00	$ 1,040,000.00	236,250.00	468,000.00	224 RICHMOND TERRACE	10301	1	1
11-01 43RD AVENUE REA	H9	4	94.00	165.00	10.00	$ 3,712,000.00	252,000.00	1,670,400.00	11-01 43 AVENUE	11101	1	1
PLUCHENIK, YAAKOV	A1	1	91.00	100.00	2.00	$ 1,900,000.00	9,763.00	75,763.00	7-06 ELVIRA AVENUE	11691	1	1
HAVEN BUILDERS, INC.	B1	1	37.00	100.00	3.00	$ 1,356,000.00	15,408.00	79,248.00	91-25 75 STREET	11421	1	1
OH, LAURA E	R3	1A	1.00	1.00	1.00	$ 251,989.00	1,001.00	8,934.00	220-71 67 AVENUE	11364	0	0
WILLIAMSON-JOSEPH, DE	B2	1	19.00	83.00	2.00	$ 679,000.00	8,524.00	33,912.00	83 UTICA AVENUE	11213	1	1
MADAN M LACHMAN	C0	1	62.00	100.00	3.00	$ 1,088,000.00	8,518.00	44,698.00	84-15 101 AVENUE	11416	1	1
ATTRACTIVE HOME, INC.	Q1	4	4.00	31.00	1.00	$ 3,080,000.00	1,075,500.00	1,386,000.00	810 DAWSON STREET	10459	73	31
ARCHER, ALAN	B2	1	30.00	107.00	2.00	$ 590,000.00	8,291.00	32,328.00	1772 PACIFIC STREET	11233	1	1
JAMES T MORIATES	D6	2	43.00	50.00	9.00	$ 625,000.00	56,250.00	281,250.00	90-07 178 STREET	11432	1	1
DAVID R DOUGLAS	B2	1	19.00	107.00	2.00	$ 538,000.00	6,242.00	29,448.00	1760 PACIFIC STREET	11233	1	1
109 JAMAICA CORP.	A1	1	100.00	84.00	1.00	$ 697,600.00	21,437.00	21,437.00	1582 EAST 56 STREET	11234	1	1
HAVEN BUILDERS, INC.	B1	1	50.00	100.00	3.00	$ 975,000.00	17,814.00	55,254.00	91-15 75 STREET	11421	1	1
ENA SIMPSON	B2	1	19.00	83.00	2.00	$ 520,000.00	5,800.00	28,440.00	83 UTICA AVENUE	11213	1	1
PRATT INSTITUTE	Z9	4	60.00	540.00	1.00	$ 1,016,250.00	454,500.00	457,313.00	189 WILLOUGHBY AVENUE	11205	3	5
LBC IV, LLC	O6	4	30.00	99.00	1.00	$ 288,000.00	58,950.00	129,600.00	188-10 HILLSIDE AVENUE	11423	1	1
JAMAICA FIRST PARKING	Z2	4	350.00	292.00	1.00	$ 2,530,000.00	1,120,500.00	1,138,500.00	90-02 168 STREET	11432	4	10
VAN WAGNER COMMCATNSI	Z9	4	41.00	99.00	1.00	$ 356,000.00	112,500.00	160,200.00	330 EAST 126 STREET	10035	1	1
PETER ARIOLA	G0	1	60.00	94.00		$ 538,000.00	7,827.00	8,791.00	84-04 SUTTER AVENUE	11417	1	1
BROOKFIELD PROPERTIES	R5	4	-	-	54.00	$ 447,146,560.00	55,693,699.00	201,215,952.00	1 LIBERTY PLAZA	10006	0	0
BH HOTELS LLC	R5	4	-	-	46.00	$ 397,111,111.00	85,585,500.00	178,700,000.00	1335 AVENUE OF THE AMER	10019	0	0
AOL TIME WARNER REALT	R5	4	-	-	55.00	$ 436,000,000.00	36,616,950.00	196,200,000.00	10 COLUMBUS CIRCLE	10019	0	0
DROOPAD, BISHNU	B2	1	40.00	100.00	2.00	$ 476,000.00	12,036.00	20,388.00	130-35 125 STREET	11420	1	1
HOFFMAN JACOB	R3	1A	40.00	100.00	3.00	$ 386,729.00	6,530.00	22,894.00	1251 48 STREET	11219	1	1

Clearly, for some of these records, the full value is exceptionally high. Further scrutiny reveals that such records have no information about lot front, lot depth, etc. For some other records, the full value of the property per build area is either excessively high or low. Since, a lot of these properties are owned by real estate firms, we can infer that either the properties owned by them are significantly different from an average property or they might be exploiting loopholes (and/or committing potential tax fraud) in the property tax law.

Top 10 Candidates for Potential Fraud

1. This property only has a one-story building and its full value is about $3M. This results in a very high full value per unit building volume.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
864163 REALTY, LLC	D9	2	157.00	95.00	1.00	$ 2,930,000.00	1,318,500.00	1,318,500.00

2. The valuation of this property ($374M) seems unreasonably high. But since, lot depth is unavailable, value per unit building volume can't be calculated.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
LOGAN PROPERTY, INC.	T1	4	4,910.00	-	3.00	$ 374,019,883.00	1,792,808,947.00	4,668,308,947.00

3. The value of this property per unit building volume is too low.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
RICH-NICH REALTY,LLC	D3	2	136.00	132.00	8.00	$ 1,040,000.00	236,250.00	468,000.00

4. For this property, average land per unit building volume and full value per unit building volume seems very high.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
11-01 43RD AVENUE REA	H9	4	94.00	165.00	10.00	$ 3,712,000.00	252,000.00	1,670,400.00

5. Small lot front and only 2 stories in this building means that property has very high value per unit building volume.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
PLUCHENIK, YAAKOV	A1	1	91.00	100.00	2.00	$ 1,900,000.00	9,763.00	75,763.00

6. For this property, full value per unit building volume is very high.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
HAVEN BUILDERS, INC.	B1	1	37.00	100.00	3.00	$ 1,356,000.00	15,408.00	79,248.00

7. The reporting of specifications for this property may or may not have been intentional (as in it could also be a data entry error). Nevertheless, it results in an exceptionally high value for such a small building.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
OH, LAURA E	R3	1A	1.00	1.00	1.00	$ 251,989.00	1,001.00	8,934.00

8. For this property, the full value per unit building volume is very high.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
WILLIAMSON-JOSEPH, DE	B2	1	19.00	83.00	2.00	$ 679,000.00	8,524.00	33,912.00

9. With building front and building depth of 1 feet, this property has very high value per unit building volume.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
MADAN M LACHMAN	C0	1	62.00	100.00	3.00	$ 1,088,000.00	8,518.00	44,698.00

10. This property has very high assessed value of land per unit lot area, building area and building volume. The full value per unit lot area is also high.

OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
ATTRACTIVE HOME, INC.	Q1	4	4.00	31.00	1.00	$ 3,080,000.00	1,075,500.00	1,386,000.00

Conclusions

1. Data Analysis and Exploration

We started with exploratory analysis of the data which included descriptive analysis, visual analysis, correlation analysis, missing data analysis, etc. Next, we imputed the data, selected expert variables, created new variables to get data ready for modeling.

2. Variable Creation

During this step, we focused on reducing the number of inputs to a model by considering which inputs are the most important to the model. Most modeling methods degrade when presented with more inputs than is needed for a robust, stable model. We started with building 75 expert variables based on five value fields viz., FULLVAL, AVLAND, AVTOT, EXLAND and EXTOT.

3. Scaling

After z-scaling, we performed PCA and found that 8 PCs explained more than 93% of the variance. We chose those PCs for further analysis and this helped us minimize dimensionality. This was followed by z-scaling (again). After that, we combined model variables with a heuristic algorithm, that utilizes the sum of absolute z-scores, z-scores squared, maximum and minimum values, etc.

4. Unsupervised Learning

The algorithms were constructed with the purpose of assigning a fraud score to every record in the data set. We did this by training the model to reproduce the original data; the reproduction error being a measure of the record's unusualness and thus, a fraud score. Two approaches were used for calculating fraud scores – outlier detection via z-scores and autoencoder error.

5. Results

Fraud scores from both autoencoder and outlier detection via z-scores were right-skewed. The algorithm assigned high fraud scores to a lot of government properties, parks and universities, which is intuitive because these properties are generally much bigger in land area, have fewer stories and at the same time have very high property value. We filtered out these records and looked at remaining candidates for property tax fraud. Closer analysis of these records reveals that the full value of some of these properties is exceptionally high/low and they do not have complete building data (lot depth, lot front, etc., being missing fields). For some other records, the full value of the property per building area is either excessively high or low.

6. Scope for Improvement

There are several things that can be done to improve this model. Some have been listed below.

a. Improving data quality – There is a lot of missing data in original dataset. Although we imputed the data, the modelling techniques cannot make up for the missing data. The inference also cannot be derived fully for outliers missing a lot of fields. Cleaner data will allow for more a better final model.
b. Domain expertise – The final model can be improved by inputs from experts. If we point out outliers to experts, then we will know why these records are being flagged as potential fraud candidates and we can think of creating new attributes for properties to better capture information and improve the accuracy of our model.
c. Augmenting data with more information – If we can get more information about property owners, crime rates, average income in the ZIP codes' areas, etc., it may be useful in improving the model further.
d. Comparison with other cities/areas – It's a good idea to look at how fraud detection is applied in other areas and see what new ideas can be incorporated in this model. Also, if property tax fraud is applied in some other countries/cities, talking to subject matter expert from those areas will help in uncovering some new ideas and attributes which can further improve the model.

Appendix

Data Quality Report: New York Real Estate Data Set

1. Introduction

The New York City Department of Finance values properties in NYC every year to calculate property taxes. Their report provides property tax data such as market and assessed values, exemptions, abatements, etc., for the assessment year, 2010-11. The information is listed by categories such as borough, tax class, building type and so on. There are 1048575 records with 30 columns each – 14 of which are categorical variables and 16 numerical variables. The following table lists out descriptions for some important variables.

2. Dataset Description

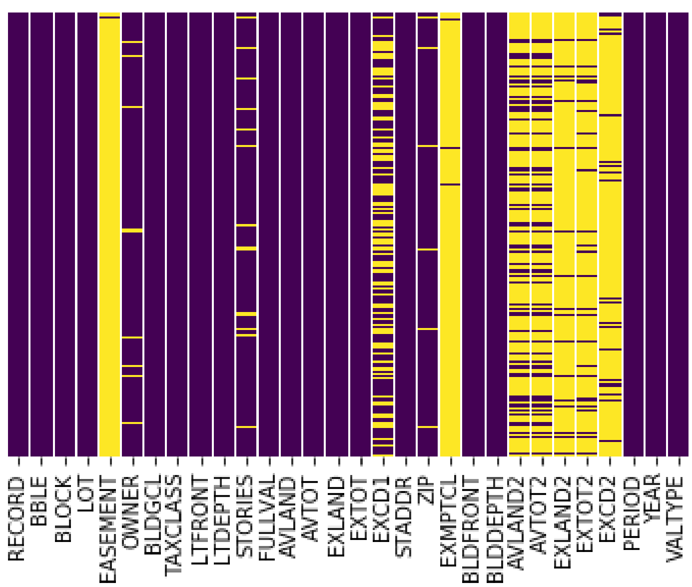
Acronyms

Abbreviation	Description
LTFRONT	Lot frontage in feet
LTDEPTH	Lot depth in feet
FULLVAL	Total market value of property
AVLAND	Market value of the land
AVTOT	Total market value
EXLAND	Exempt land value
EXTOT	Exempt total value
EXCD1	Exempt condo value
BLDFRONT	Building frontage in feet
BLDDEPTH	Building depth in feet
AVLAND2	2nd Market value of the land
AVTOT2	2nd Total market value
EXLAND2	Transitional Exempt land value
EXTOT2	2nd Exempt total value
EXCD2	2nd Exempt condo value
BLDGCL	Building class

Summary table

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
RECORD	int64	1,048,575								100.00%	
BLOCK	int64	1,048,575	4,708.87	3,699.55	1.00	1,534.00	3,944.00	6,797.00	16,350	100.00%	13,949
LOT	int64	1,048,575	370.09	860.54	1.00	23.00	49.00	146.00	9,978	100.00%	6,366
LTFRONT	int64	1,048,575	36.17	73.73	0.00	19.00	25.00	40.00	9,999	100.00%	1,277
LTDEPTH	int64	1,048,575	88.28	75.48	0.00	80.00	100.00	100.00	9,999	100.00%	1,336
STORIES	float64	996,433	5.06	8.43	1.00	2.00	2.00	3.00	119	95.03%	111
FULLVAL	int64	1,048,575	880,487.66	11,702,930.00	0.00	303,000.00	446,000.00	619,000.00	6,150,000,000	100.00%	108,277
AVLAND	int64	1,048,575	85,995.03	4,100,755.00	0.00	9,160.00	13,646.00	19,706.00	2,668,500,000	100.00%	70,529
AVTOT	int64	1,048,575	230,758.18	6,951,206.00	0.00	18,385.00	25,339.00	46,095.00	4,668,309,000	100.00%	112,294
EXLAND	int64	1,048,575	36,811.79	4,024,330.00	0.00	0.00	1,620.00	1,620.00	2,668,500,000	100.00%	33,186
EXTOT	int64	1,048,575	92,543.81	6,578,281.00	0.00	0.00	1,620.00	2,090.00	4,668,309,000	100.00%	63,805
EXCD1	float64	622,642	1,604.50	1,388.13	1,010.00	1,017.00	1,017.00	1,017.00	7,170	59.38%	129
ZIP	float64	1,022,219	10,935.32	526.58	10,001.00	10,453.00	11,215.00	11,364.00	33,803	97.49%	196
BLDFRONT	int64	1,048,575	23.02	35.79	0.00	15.00	20.00	24.00	7,575	100.00%	610
BLDDEPTH	int64	1,048,575	40.07	43.04	0.00	26.00	39.00	51.00	9,393	100.00%	620
AVLAND2	float64	280,966	246,365.48	6,199,390.00	3.00	5,705.00	20,059.00	62,338.75	2,371,005,000	26.80%	58,169
AVTOT2	float64	280,972	716,078.71	11,690,170.00	3.00	34,013.50	80,010.00	240,792.00	4,501,180,000	26.80%	110,890
EXLAND2	float64	86,675	351,802.21	10,852,480.00	1.00	2,090.00	3,053.00	31,419.00	2,371,005,000	8.27%	21,996
EXTOT2	float64	129,933	658,114.78	16,129,810.00	7.00	2,889.00	37,116.00	106,629.00	4,501,180,000	12.39%	48,106
EXCD2	float64	90,941	1,371.66	1,105.49	1,011.00	1,017.00	1,017.00	1,017.00	7,160	8.67%	60
EASEMENT	object	4,043								0.39%	12
OWNER	object	1,017,492								97.04%	847,053
BLDGCL	object	1,048,575								100.00%	200
TAXCLASS	object	1,048,575								100.00%	11
STADDR	object	1,047,934								99.94%	820,637
EXMPTCL	object	14,992								1.43%	14
PERIOD	object	1,048,575								100.00%	1
YEAR	object	1,048,575								100.00%	1
VALTYPE	object	1,048,575								100.00%	1

Heat map of missing values



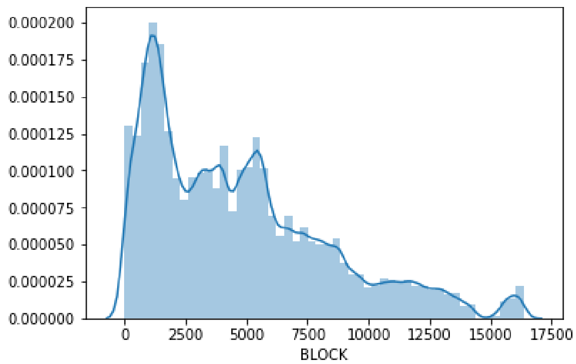
3. Numerical Data Analysis

BLOCK – Block number

Block ID cannot effectively show the real number of properties in a specific block.

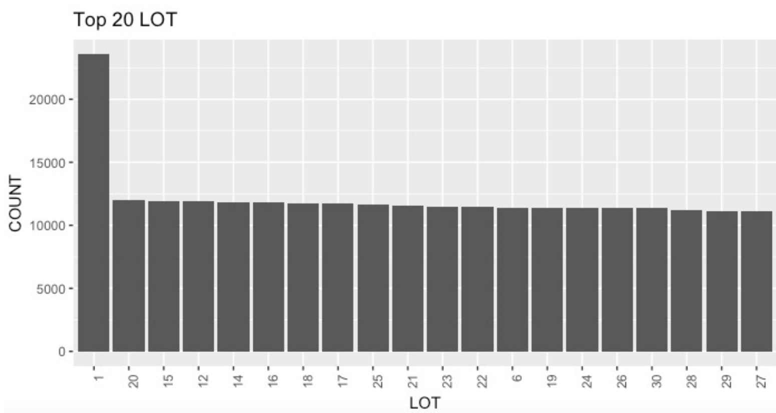
Block number to area mapping:
Manhattan - 1 to 2,255
Bronx - 2,260 to 5,958
Brooklyn - 1 to 8,955
Queens - 1 to 16,350
Staten Island - 1 to 8,050

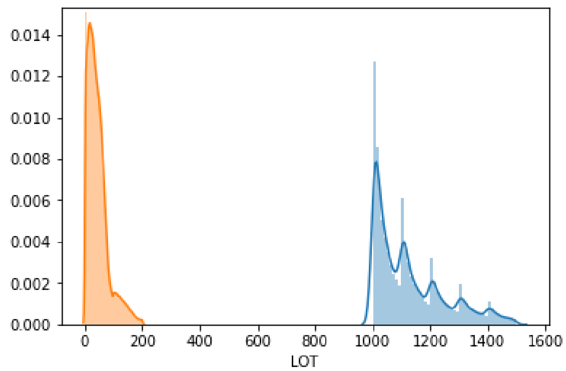
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
BLOCK	int64	1,048,575	4,708.87	3,699.55	1.00	1,534.00	3,944.00	6,797.00	16,350	100.00%	13,949

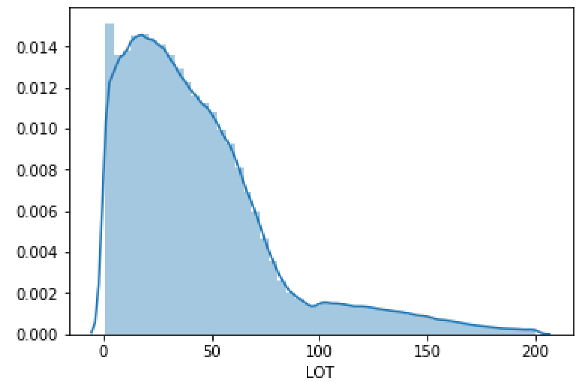


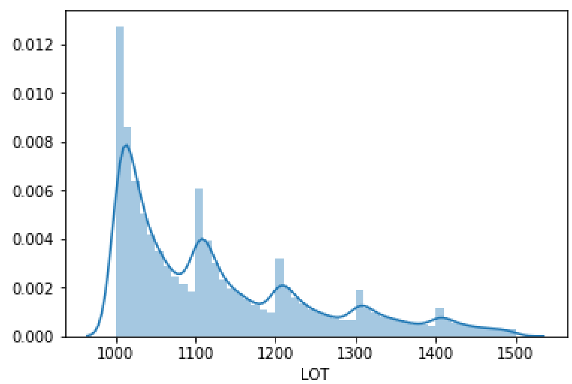
<u>LOT</u> – Lot number within block
Every record has a lot ID but some records share the same lot ID. Lot ID value of 1 has the highest frequency in this dataset, but that may not be the lot with highest number of properties as lot ID is only unique within a block.

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
LOT	int64	1,048,575	370.09	860.54	1.00	23.00	49.00	146.00	9,978	100.00%	6,366



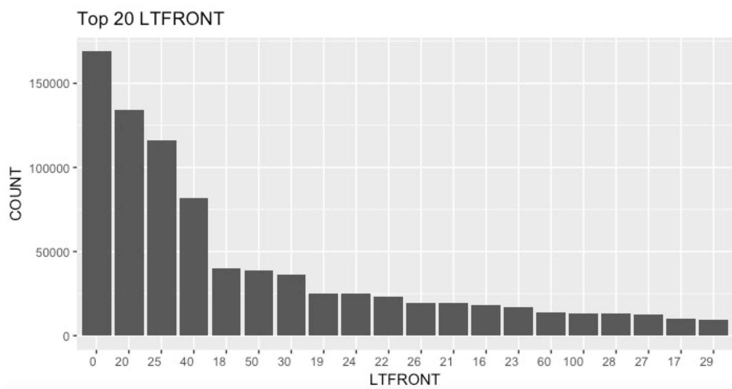

0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0 200 400 600 800 1000 1200 1400 1600
LOT

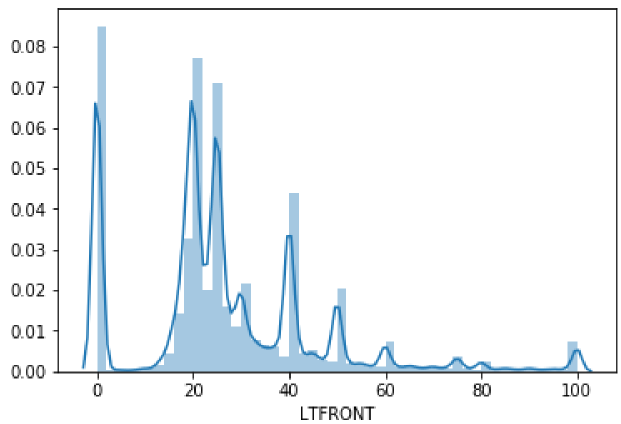

0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0 50 100 150 200
LOT


0.012
0.010
0.008
0.006
0.004
0.002
0.000
1000 1100 1200 1300 1400 1500
LOT

LTFRONT – Lot Width

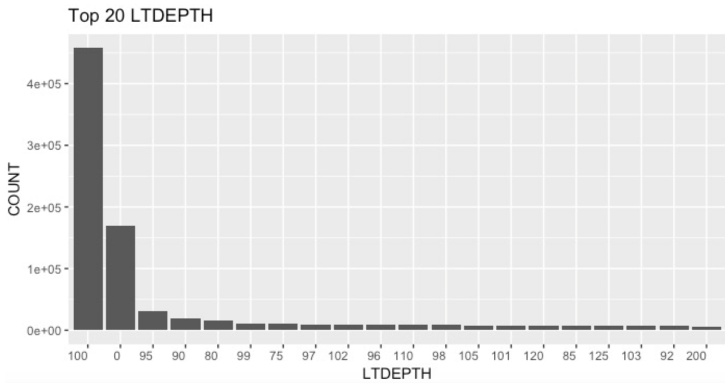
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
LTFRONT	int64	1,048,575	36.17	73.73	0.00	19.00	25.00	40.00	9,999	100.00%	1,277

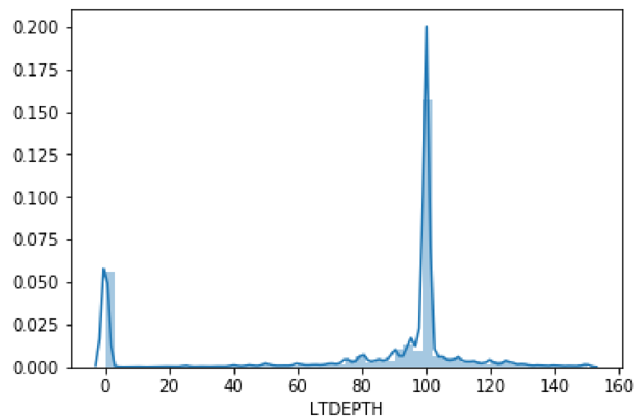




LTDEPTH – Lot Depth

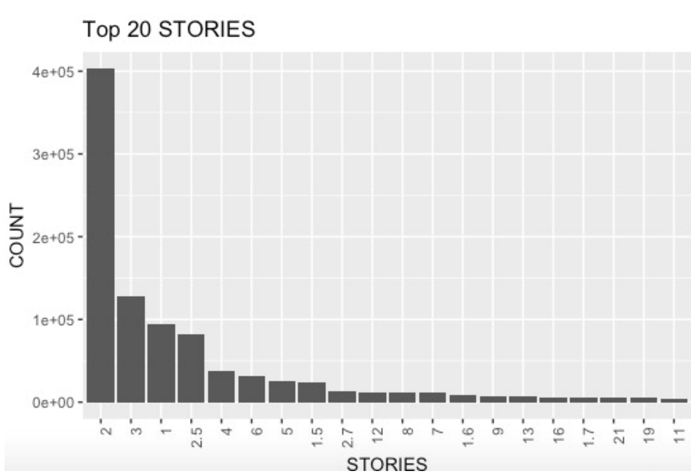
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
LTDEPTH	int64	1,048,575	88.28	75.48	0.00	80.00	100.00	100.00	9,999	100.00%	1,336

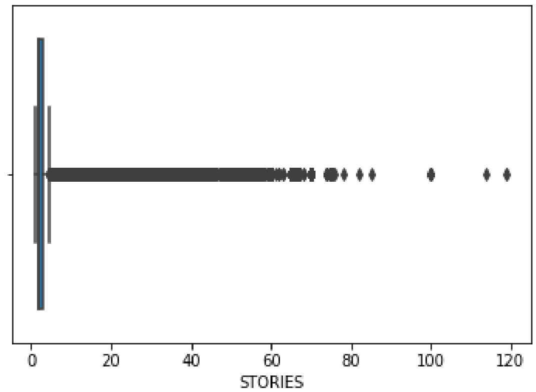


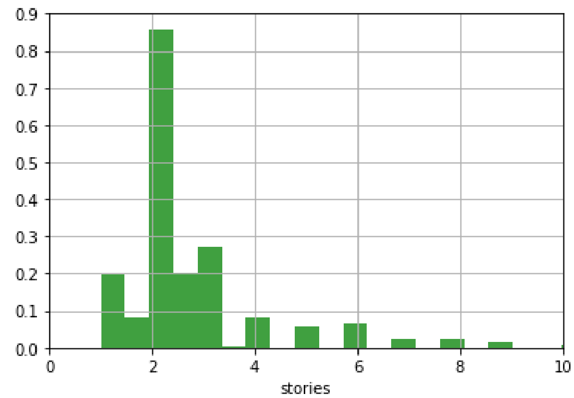


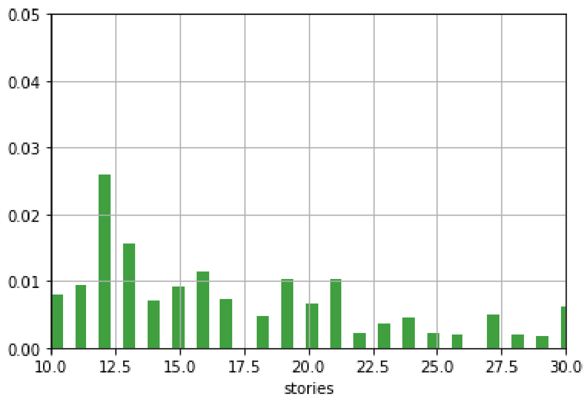
STORIES – Number of stories in the building

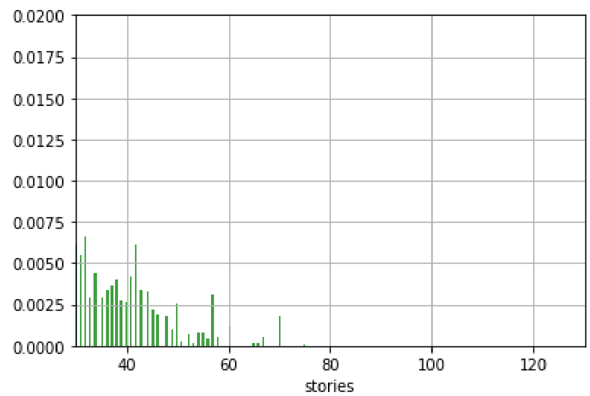
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
STORIES	float64	996,433	5.06	8.43	1.00	2.00	2.00	3.00	119	95.03%	111





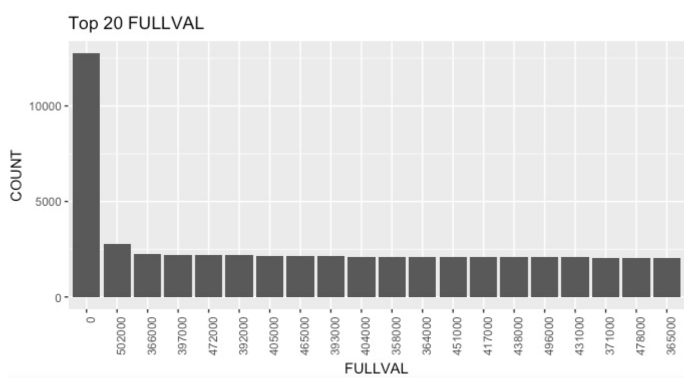


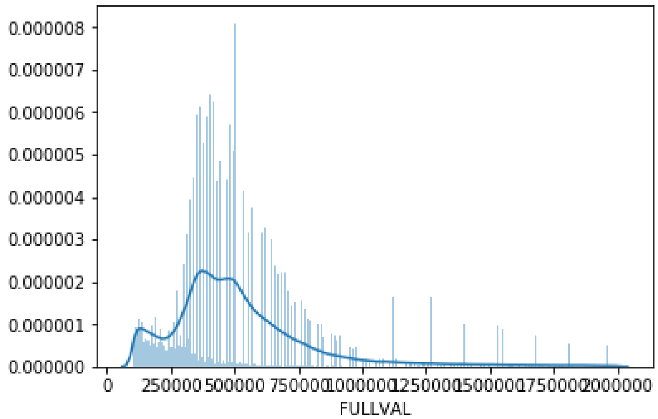




FULLVAL – Market Value

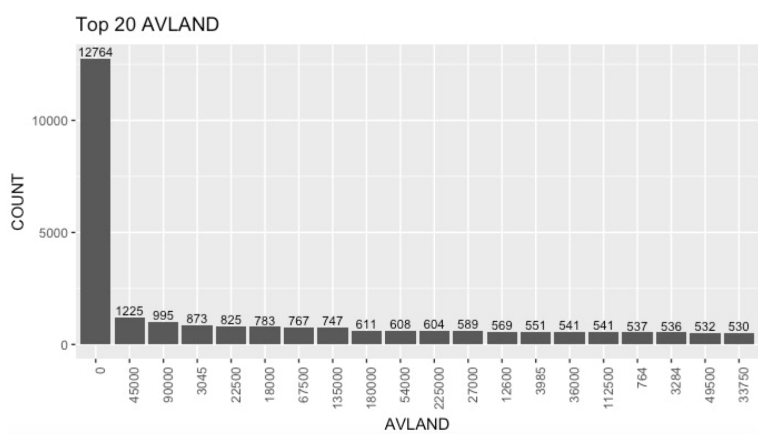
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
FULLVAL	int64	1,048,575	880,487.66	11,702,930.00	0.00	303,000.00	446,000.00	619,000.00	6,150,000,000	100.00%	108,277

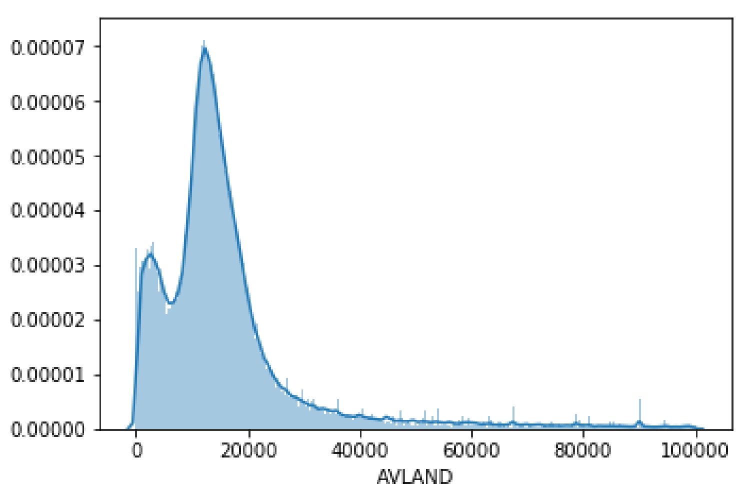




AVLAND – Actual Land Value

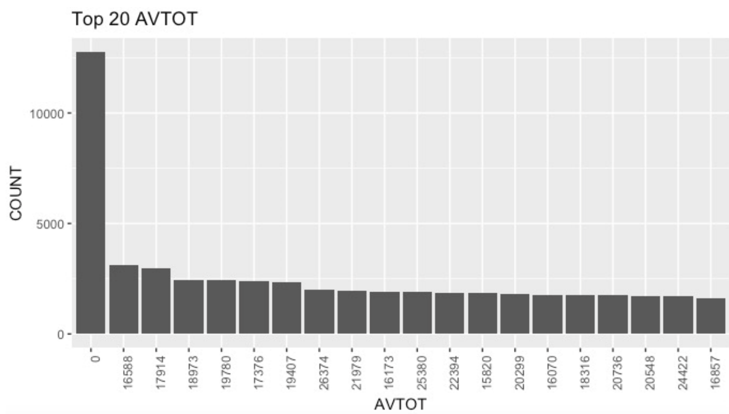
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
AVLAND	int64	1,048,575	85,995.03	4,100,755.00	0.00	9,160.00	13,646.00	19,706.00	2,668,500,000	100.00%	70,529

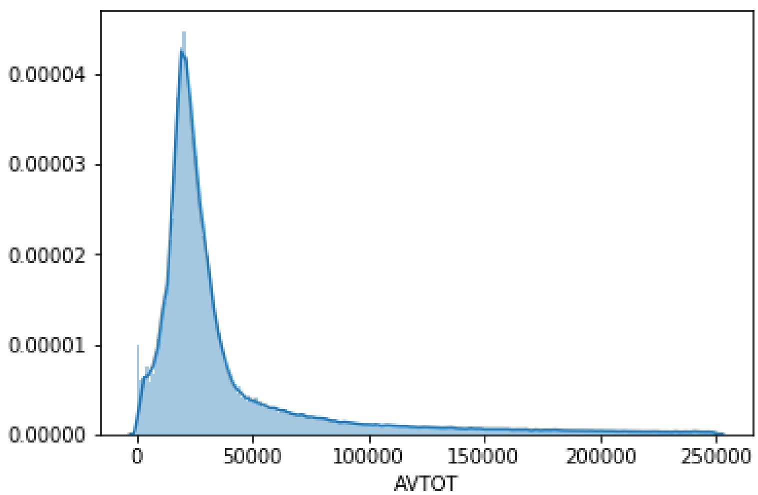




AVTOT – Actual Total Value

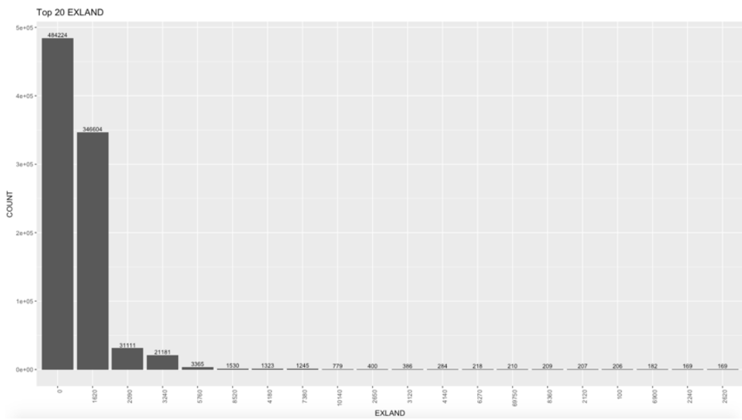
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
AVTOT	int64	1,048,575	230,758.18	6,951,206.00	0.00	18,385.00	25,339.00	46,095.00	4,668,309,000	100.00%	112,294

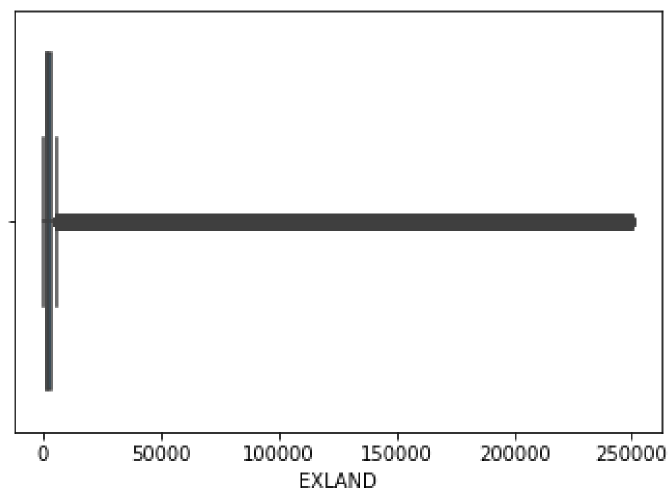


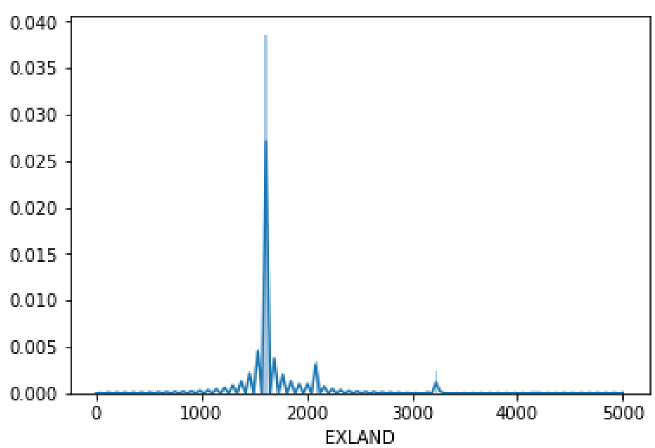


EXLAND – Actual Exempt Land Value

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXLAND	int64	1,048,575	36,811.79	4,024,330.00	0.00	0.00	1,620.00	1,620.00	2,668,500,000	100.00%	33,186

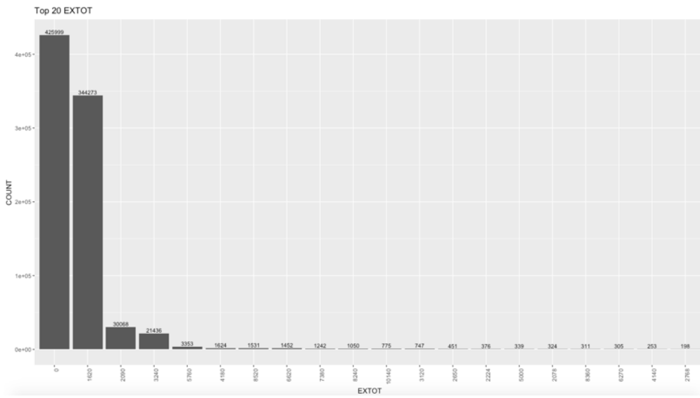


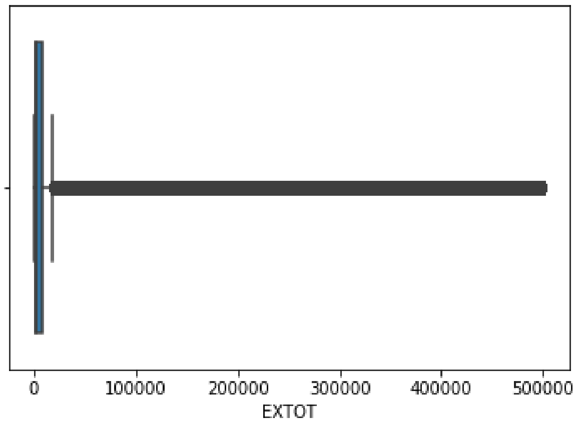


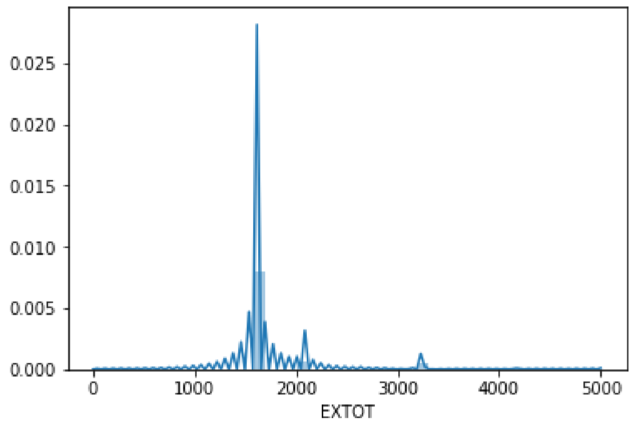


EXTOT – Actual Exempt Land Total

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXTOT	int64	1,048,575	92,543.81	6,578,281.00	0.00	0.00	1,620.00	2,090.00	4,668,309,000	100.00%	63,805

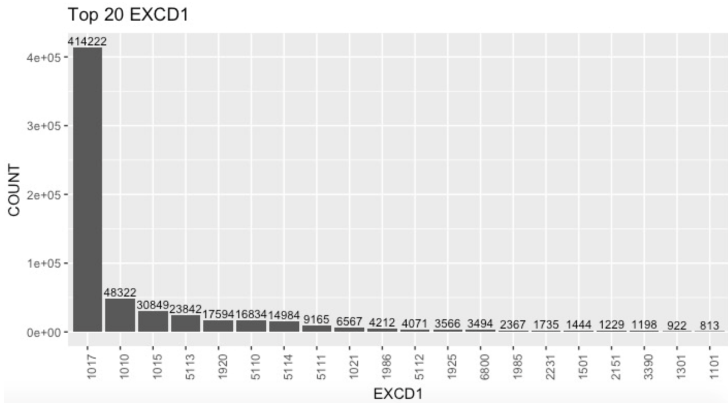


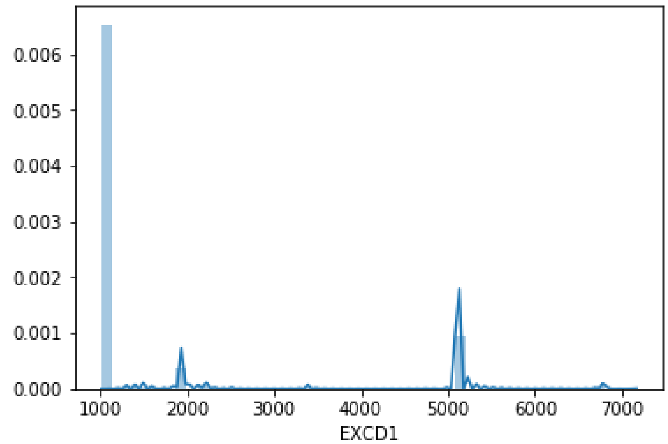




EXCD1 – Exemption Code 1

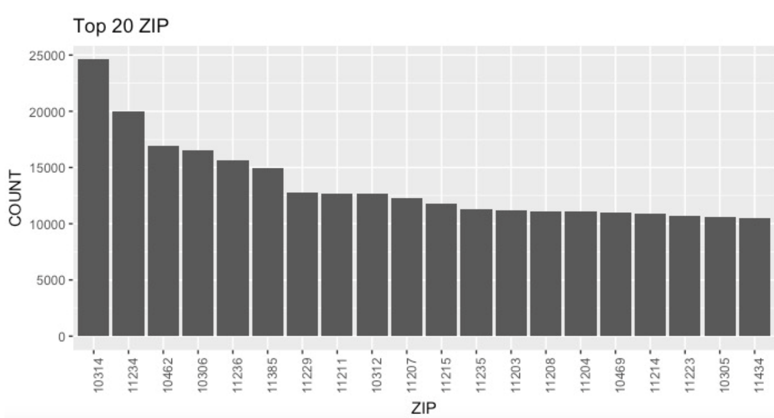
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXCD1	float64	622,642	1,604.50	1,388.13	1,010.00	1,017.00	1,017.00	1,017.00	7,170	59.38%	129

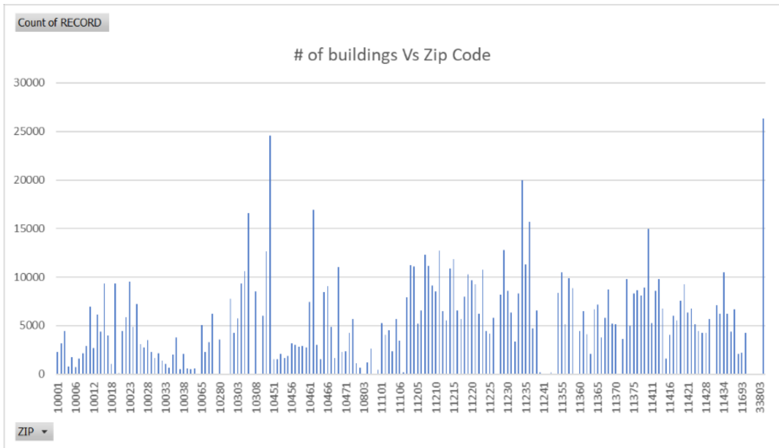




ZIP – ZIP Code

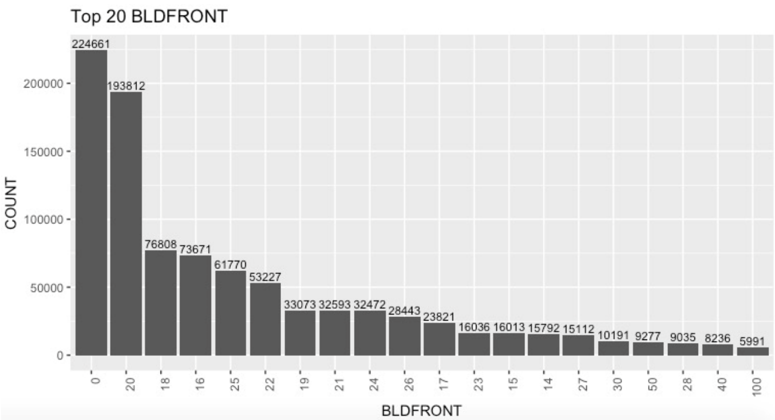
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
ZIP	float64	1,022,219	10,935.32	526.58	10,001.00	10,453.00	11,215.00	11,364.00	33,803	97.49%	196

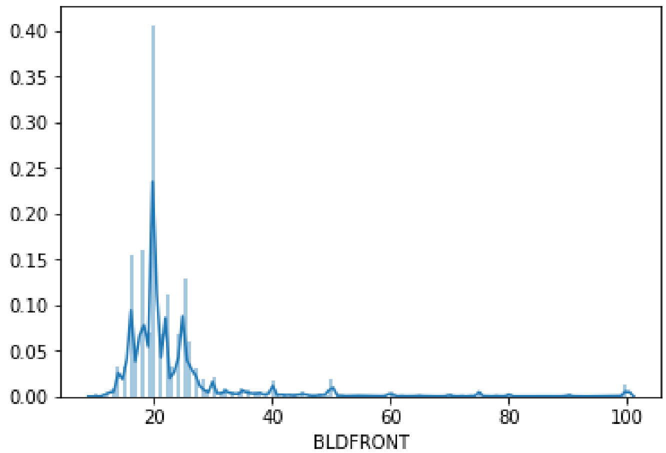




BLDFRONT – Building Width

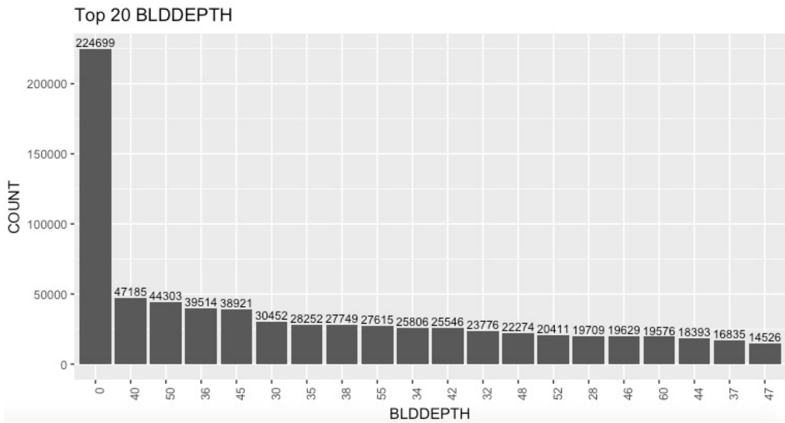
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
BLDFRONT	int64	1,048,575	23.02	35.79	0.00	15.00	20.00	24.00	7,575	100.00%	610

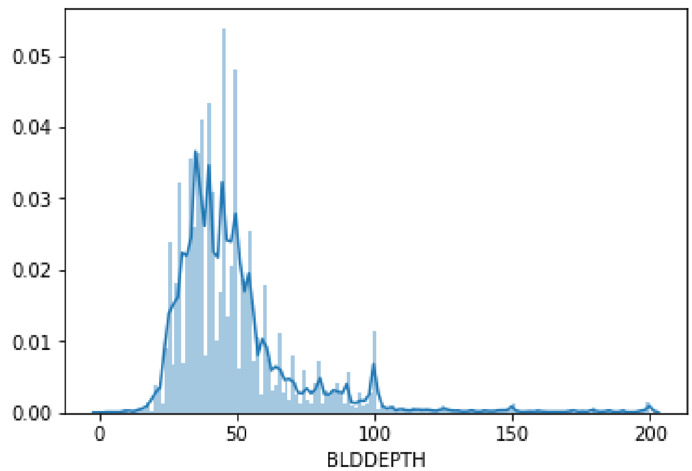




BLDDEPTH – Building Depth

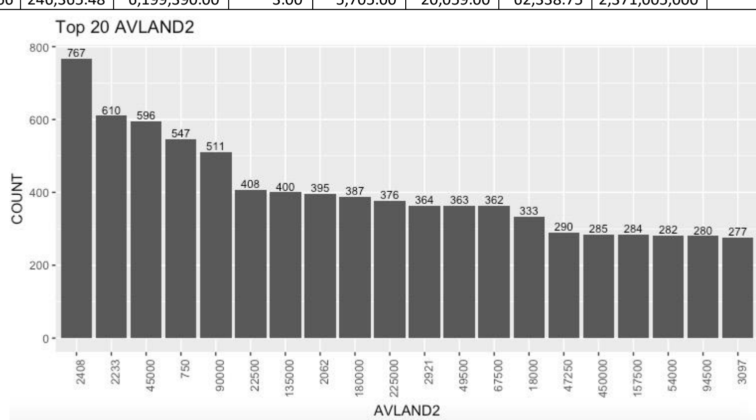
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
BLDDEPTH	int64	1,048,575	40.07	43.04	0.00	26.00	39.00	51.00	9,393	100.00%	620

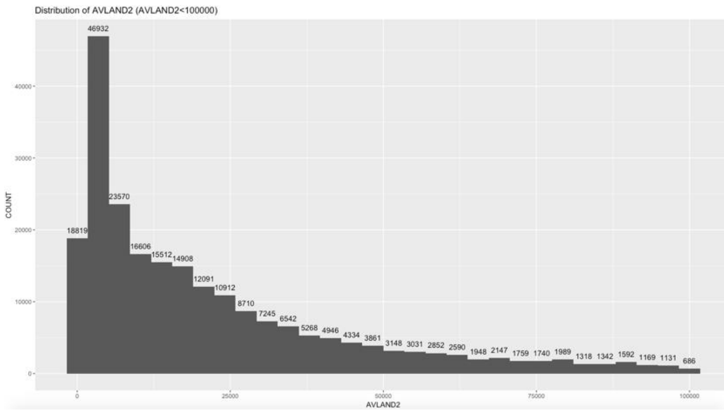




AVLAND2 – Transitional Land Value

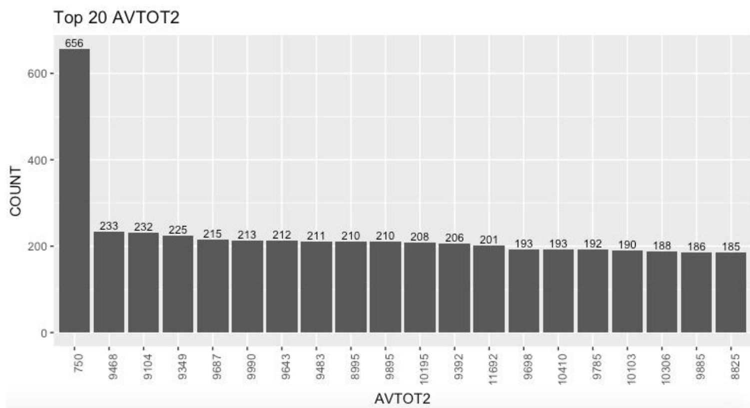
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
AVLAND2	float64	280,966	246,365.48	6,199,390.00	3.00	5,705.00	20,059.00	62,338.75	2,371,005,000	26.80%	58,169

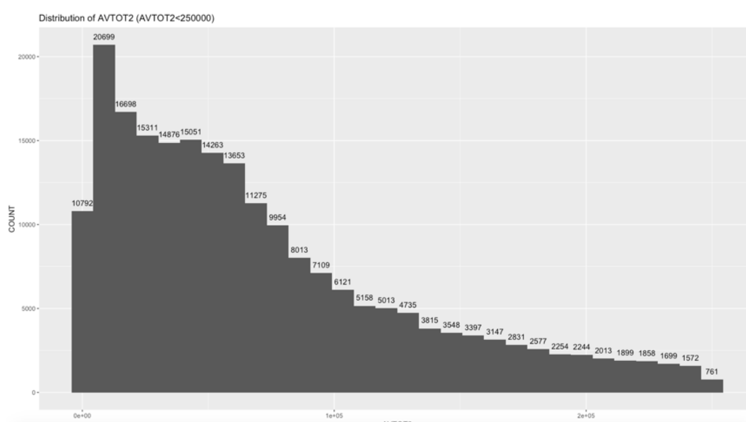




AVTOT2 – Transitional Total Value

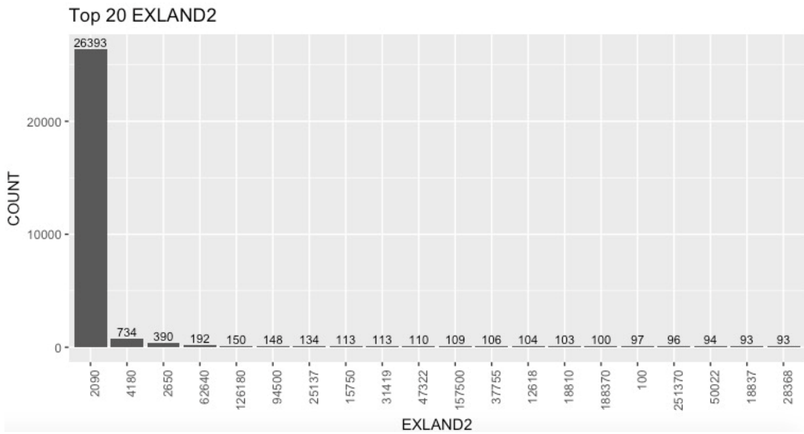
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
AVTOT2	float64	280,972	716,078.71	11,690,170.00	3.00	34,013.50	80,010.00	240,792.00	4,501,180,000	26.80%	110,890

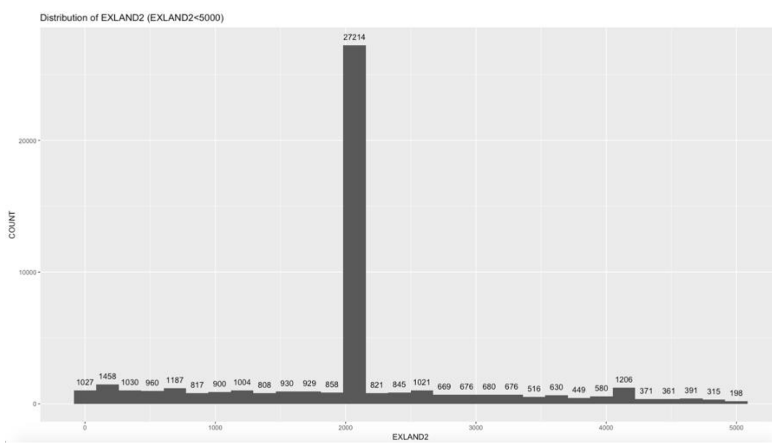




EXLAND2 – Transitional Exempt Land Value

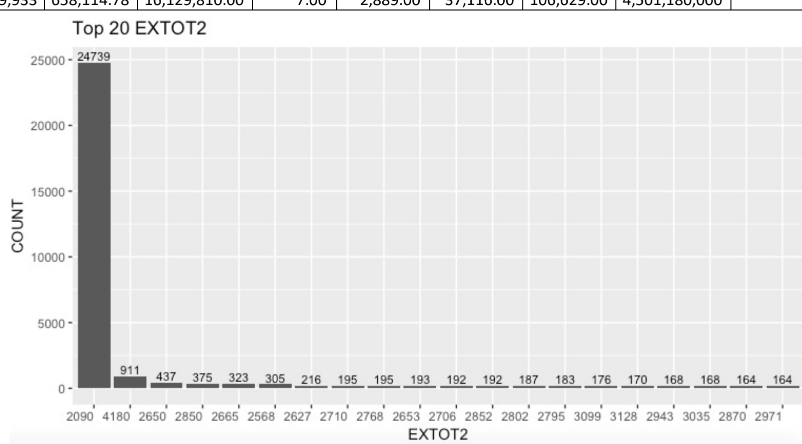
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXLAND2	float64	86,675	351,802.21	10,852,480.00	1.00	2,090.00	3,053.00	31,419.00	2,371,005,000	8.27%	21,996

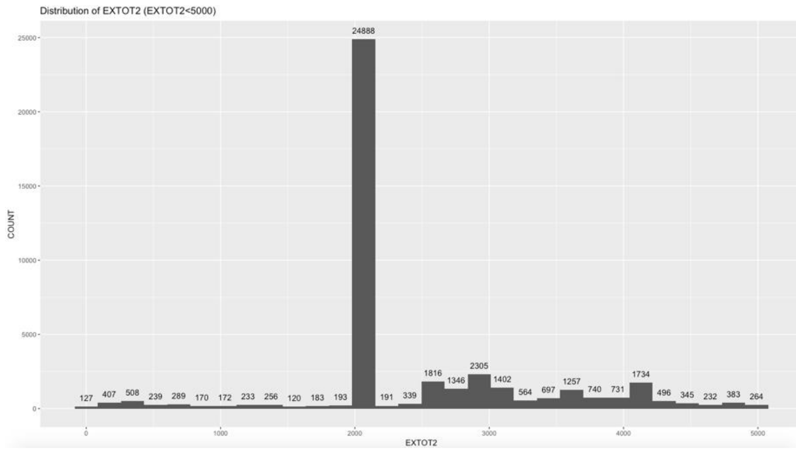




EXTOT2 – Transitional Exempt Land Total

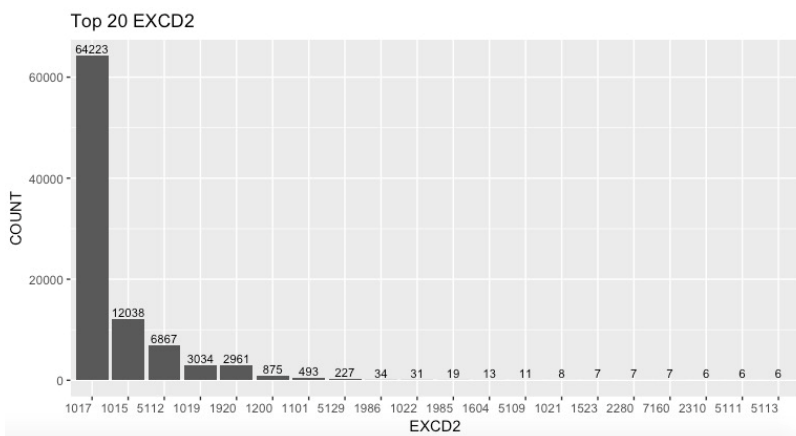
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXTOT2	float64	129,933	658,114.78	16,129,810.00	7.00	2,889.00	37,116.00	106,629.00	4,501,180,000	12.39%	48,106

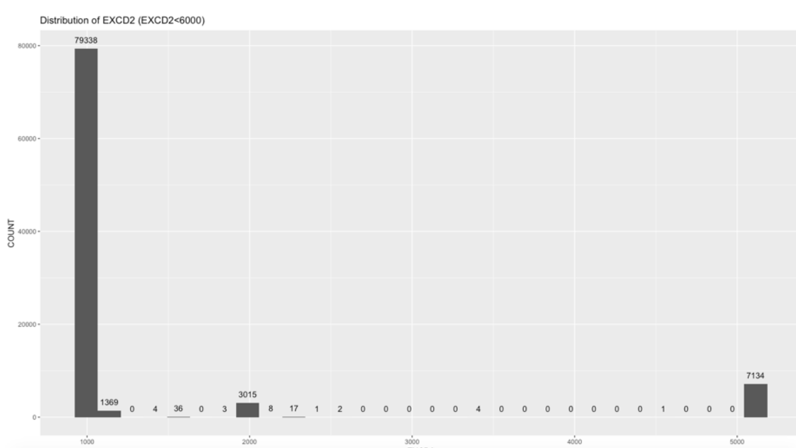




EXCD2 – Exemption Code 2

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXCD2	float64	90,941	1,371.66	1,105.49	1,011.00	1,017.00	1,017.00	1,017.00	7,160	8.67%	60

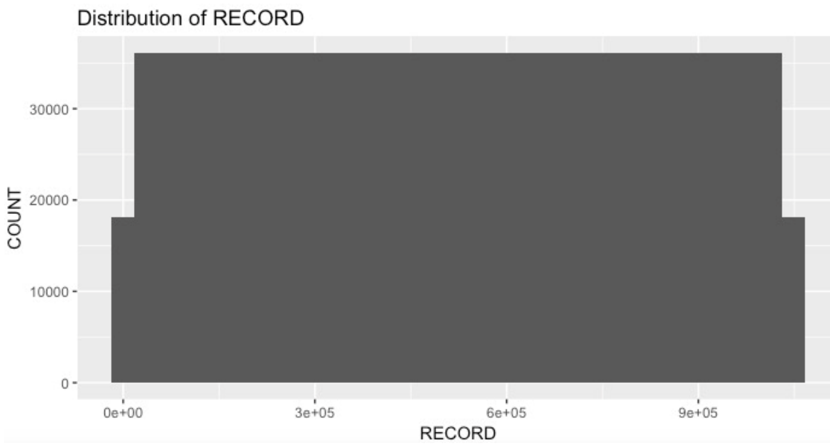




4. Categorical Data Analysis

<u>RECORD</u> – Record ID

There are 1048575 records in this dataset, so the record ID varies from 1 to 1048575.



<u>BBLE</u> – Concatenation of AV_BORO, AV_BLOCK, AV_LOT, AV_EASEMENT

There are 1048575 different records in this dataset, which means every record has a unique BBLE.

<u>EASEMENT</u> – Easement description

Space indicates that 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' through 'M' are duplicates of 'E';

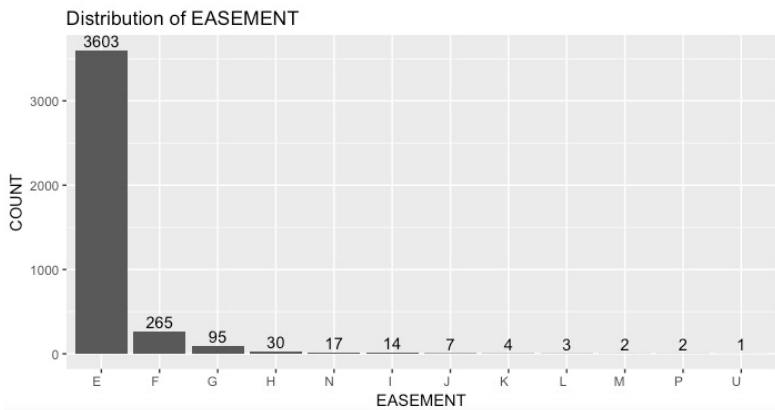
'N' indicates non-transit easement;

'P' indicates piers;

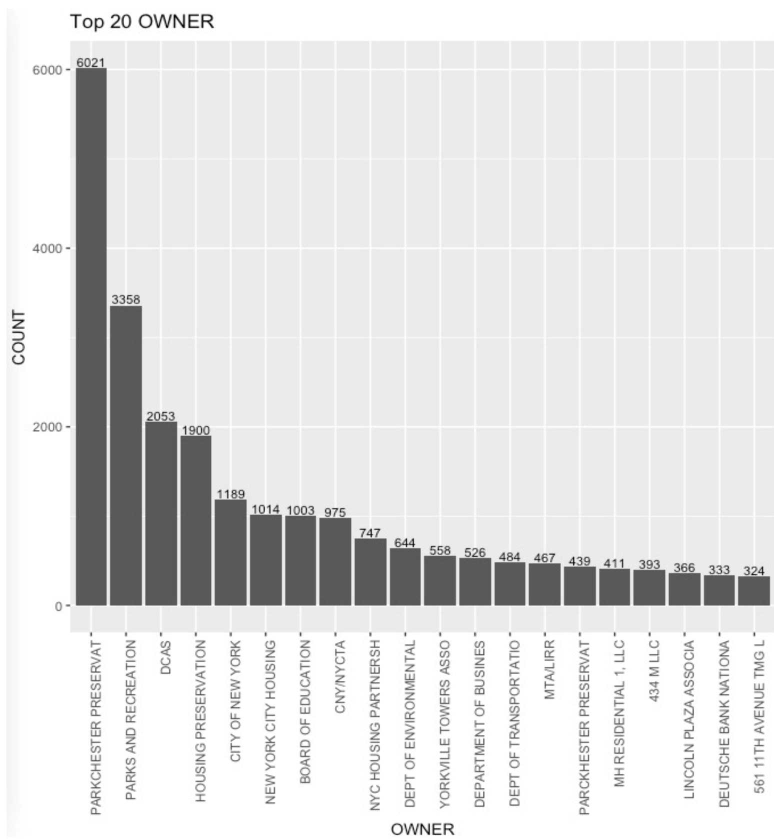
'R' indicates railroads;

'S' indicates street;

'U' indicates U.S. government;



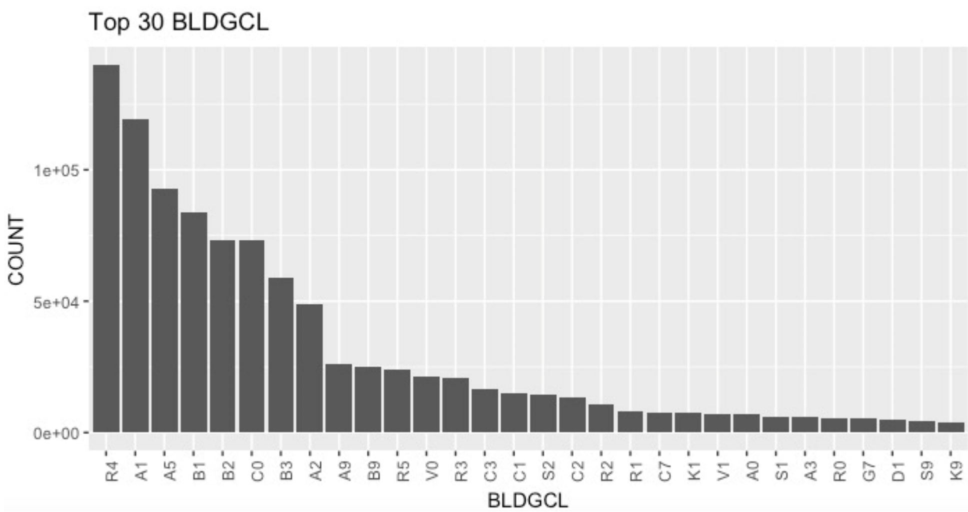
OWNER – Owner of property



BLDGCL – Building Class

Class A represents the highest quality buildings in their market. Buildings of class B are generally a little older, but still have quality management and tenants. Buildings of class C are older buildings (usually more than 20 years), are located in less desirable areas, and are in need of extensive renovation.

Predictor	Data Type	count	Percentage Populated	# of unique values
BLDGCL	object	1,048,575	100.00%	200



TAXCLASS – Tax Class

TAXCLASS 1 = 1-3 unit residences;

TAXCLASS 1A = 1-3 story condominiums originally a condo;

TAXCLASS 1B = Residential vacant land;

TAXCLASS 1C = 1-3 unit condominums originally tax class 1;

TAXCLASS 1D = Select bungalow colonies;

TAXCLASS 2 = Apartments;

TAXCLASS 2A = Apartments with 4-6 units;

TAXCLASS 2B = Apartments with 7-10 units;

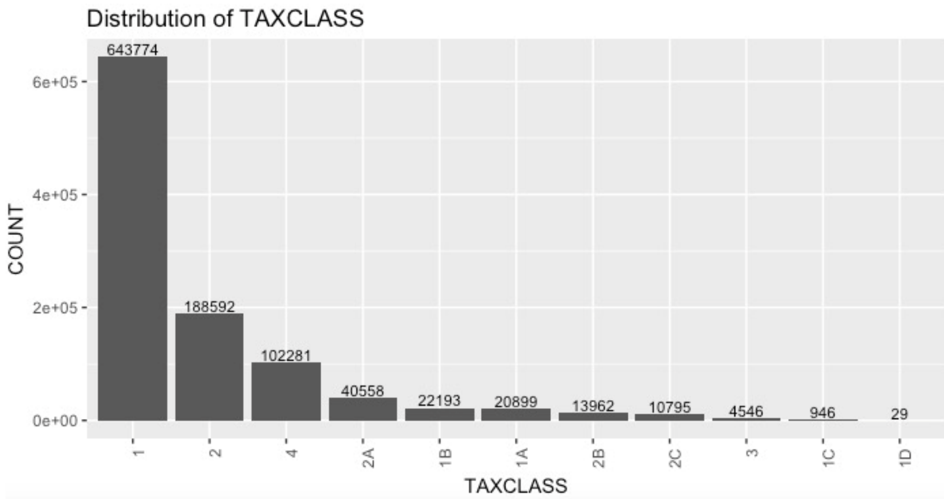
TAXCLASS 2C = Coops/condos with 2-10 units;

TAXCLASS 3 = Utilities (except ceiling rr);

TAXCLASS 4A = Utilities - ceiling railroads;

TAXCLASS 4 = All others

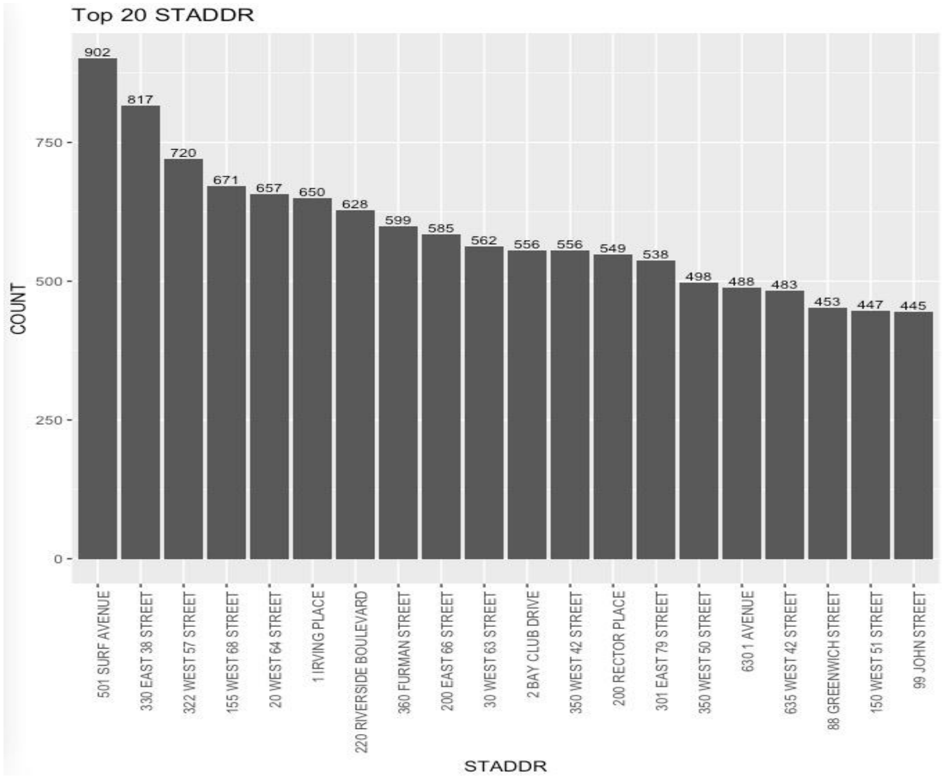
Predictor	Data Type	count	Percentage Populated	# of unique values
TAXCLASS	object	1,048,575	100.00%	11



STADDR – Street Address

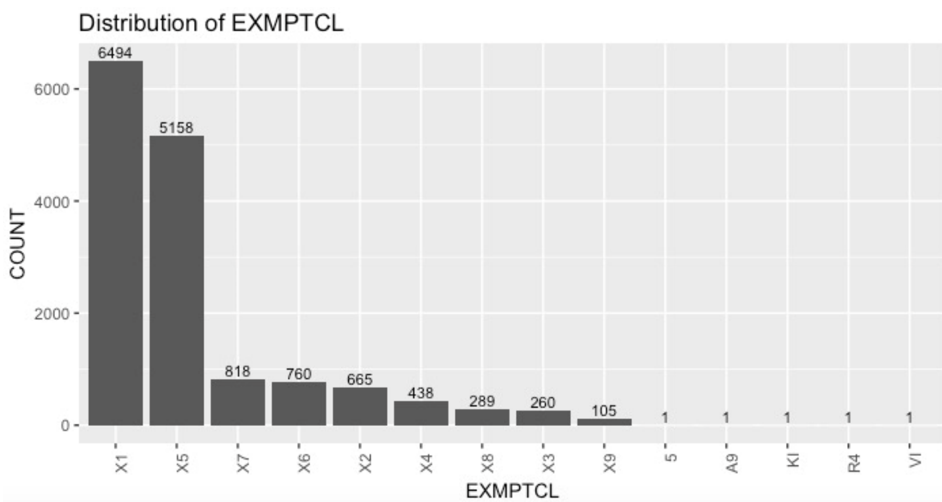
This variable has too many unique values. A good way to visualize it would be to create a heatmap on Google Maps, which is beyond the scope of this course.

Predictor	Data Type	count	Percentage Populated	# of unique values
STADDR	object	1,047,934	99.94%	820,637



EXMPTCL – Exempt Class

Predictor	Data Type	count	Percentage Populated	# of unique values
EXMPTCL	object	14,992	1.43%	14



PERIOD – Assessment Period

Predictor	Data Type	count	Percentage Populated	# of unique values
PERIOD	object	1,048,575	100.00%	1

Single variable: FINAL

YEAR – Assessment Year

Predictor	Data Type	count	Percentage Populated	# of unique values
YEAR	object	1,048,575	100.00%	1

Single variable: 2010/11

VALTYPE

Predictor	Data Type	count	Percentage Populated	# of unique values
VALTYPE	object	1,048,575	100.00%	1

Single Variable: AC-TR

References

[1] TJN Profit Shifting Tax Loss Estimates.
https://www.taxjustice.net/2017/03/22/new-estimates-tax-avoidance-multinationals/

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