DETECT ANOMALIES IN NY PROPERTY DATA

Build an algorithmic system to find potential tax fraud

APRIL 29, 2020

MGTA 495 FRAUD ANALYTICS PROJECT 1

JIE CHEN, SHIYI HUA, KEXIN LIU, QIUYI LU, ZHANYI XU, JINGWEN YAN, ZIYUAN YAN

Table of Contents

I. Executive Summary

The purpose of this project is to find potential property tax fraud going on in NYC using the NY property dataset. The dataset consists of financial, geographic, and other information about each property. The city of NY wants an algorithmic system that can look through their one million property records to find potential tax fraud. The kind of fraud they're looking for is people underpaying tax or otherwise misrepresenting their property characteristics. We looked for anomalies in the data by investigating which properties' assessed values are too low or too high and listed the 10 properties with the highest fraud score. For each of the 10 potential fraud records, we gave a detailed explanation of why it is abnormal.

II. Description of Data

The following tables are summary tables of field statistics. The dataset has 1070994 rows and 32 variables. The variables EASEMENT, OWNER, EXT, STORIES, EXCD1, STADDR, ZIP, EXMPTCL, AVLAND2, AVTOT2, EXLAND2, EXTOT2, EXCD2 have null value. The variables LTFRONT, LTDEPTH, EXLAND, EXTOT, BLDFRONT, BLDDEPTH have 0 values. The variables PERIOD, YEAR, VALTYPE have all the same records.

SN	Columns	Data Type	Number populated	% Populated	# Unique Values	# Non-Zeros
1	RECORD	int64	1070994	100	1070994	1070994
2	BBLE	object	1070994	100	1070994	1070994
3	B	int64	1070994	100	5	1070994
4	BLOCK	int64	1070994	100	13984	1070994
5	LOT	int64	1070994	100	6366	1070994
6	EASEMENT	object	4636	0.432869	13	1070994
7	OWNER	object	1039251	97.03612	863349	1070994
8	BLDGCL	object	1070994	100	200	1070994
9	TAXCLASS	object	1070994	100	11	1070994
10	LTFRONT	int64	1070994	100	1297	901886
11	LTDEPTH	int64	1070994	100	1370	900866
12	EXT	object	354305	33.08188	4	1070994
13	STORIES	float64	1014730	94.74656	112	1070994
14	FULLVAL	float64	1070994	100	109324	1057987
15	AVLAND	float64	1070994	100	70921	1057985

16	AVTOT	float64	1070994	100	112914	1057987
17	EXLAND	float64	1070994	100	33419	579295
18	EXTOT	float64	1070994	100	64255	638422
19	EXCD1	float64	638488	59.61639	130	1070994
20	STADDR	object	1070318	99.93688	839281	1070994
21	ZIP	float64	1041104	97.20913	197	1070994
22	EXMPTCL	object	15579	1.45463	15	1070994
23	BLDFRONT	int64	1070994	100	612	842179
24	BLDDEPTH	int64	1070994	100	621	842141
25	AVLAND2	float64	282726	26.39847	58592	1070994
26	AVTOT2	float64	282732	26.39903	111361	1070994
27	EXLAND2	float64	87449	8.165218	22196	1070994
28	EXTOT2	float64	130828	12.21557	48349	1070994
29	EXCD2	float64	92948	8.678667	61	1070994
30	PERIOD	object	1070994	100	1	1070994
31	YEAR	object	1070994	100	1	1070994
32	VALTYPE	object	1070994	100	1	1070994

For numeric fields:

SN	COUNT	MEAN	STD	MIN	25%	50%	75%	MAX
RECORD	1070994	535497.5	3.09E+05	1	267749.25	535497.5	803245.75	1.07E+06
LTFRONT	1070994	36.635301	7.40E+01	0	19	25	40	1.00E+04
LTDEPTH	1070994	88.861594	7.64E+01	0	80	100	100	1.00E+04
STORIES	1014730	5.006918	8.37E+00	1	2	2	3	1.19E+02
FULLVAL	1070994	874264.5054	1.16E+07	0	304000	447000	619000	6.15E+09
AVLAND	1070994	85067.91867	4.06E+06	0	9180	13678	19740	2.67E+09
AVTOT	1070994	227238.1687	6.88E+06	0	18374	25340	45438	4.67E+09
EXLAND	1070994	36423.89069	3.98E+06	0	0	1620	1620	2.67E+09
EXTOT	1070994	91186.98168	6.51E+06	0	0	1620	2090	4.67E+09
EXCD1	638488	1602.014232	1.38E+03	1010	1017	1017	1017	7.17E+03
BLDFRONT	1070994	23.04277	3.56E+01	0	15	20	24	7.58E+03
BLDDEPTH	1070994	39.922836	4.27E+01	0	26	39	50	9.39E+03
AVLAND2	282726	246235.7193	6.18E+06	3	5705	20145	62640	2.37E+09
AVTOT2	282732	713911.4362	1.17E+07	3	33912	79962.5	240551	4.50E+09

EXLAND2	87449	351235.6843	1.08E+07	1	2090	3048	31779	2.37E+09
EXTOT2	130828	656768.2819	1.61E+07	7	2870	37062	106840.75	4.50E+09
EXCD2	92948	1364.041679	1.09E+03	1011	1017	1017	1017	7.16E+03

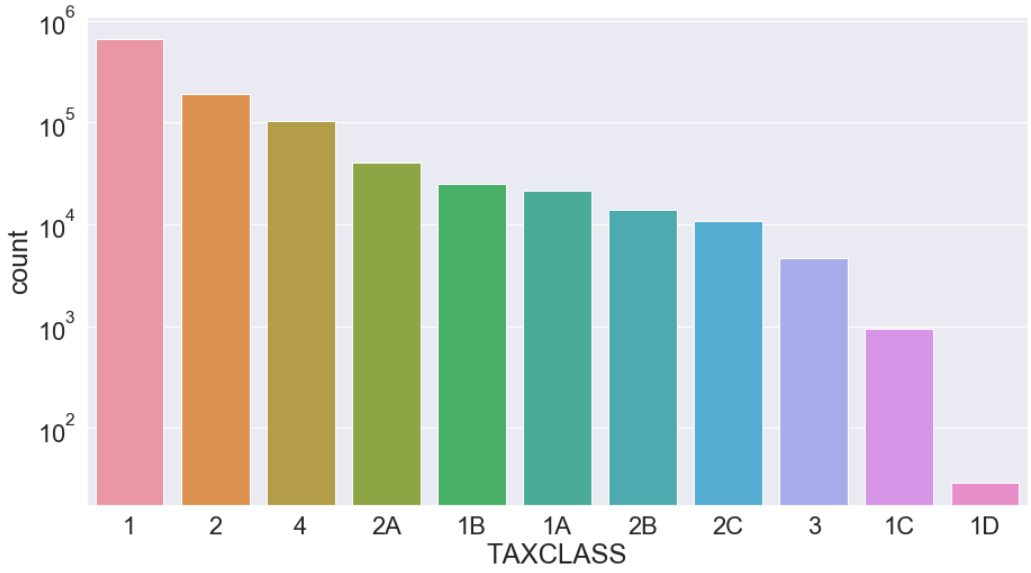
For categorical fields:

	FIELD	COUNT	UNIQUE	TOP	FREQ
1	B	1070994	5	4	358046
2	BLOCK	1070994	13984	3944	3888
3	LOT	1070994	6366	1	24367
4	EASEMENT	4636	13	E	4148
5	OWNER	1039251	863349	PARKCHESTER PRESERVAT	6021
6	BLDGCL	1070994	200	R4	139879
7	TAXCLASS	1070994	11	1	660721
8	EXT	354305	4	G	266970
9	STADDR	1070318	839281	501 SURF AVENUE	902
10	ZIP	1041104	197	10314	24606
11	EXMPTCL	15579	15	X1	6912
12	PERIOD	1070994	1	FINAL	1070994
13	YEAR	1070994	1	2010/11	1070994
14	VALTYPE	1070994	1	AC-TR	1070994

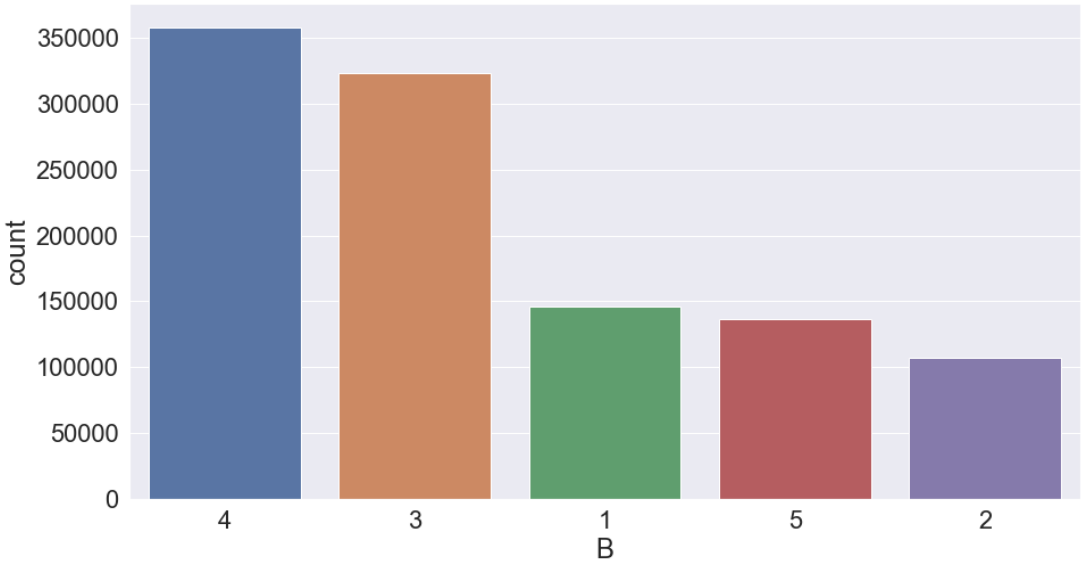
(*NOTE: BBLE is an identifier and has a unique number for each row.)

In consideration of property tax fraud in the city, we consider FULLVAL, AVLAND, AVTOT, ZIP, STORIES, B, LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, TAXCLASS as important variables. The following is the distribution/histograms of those variables.

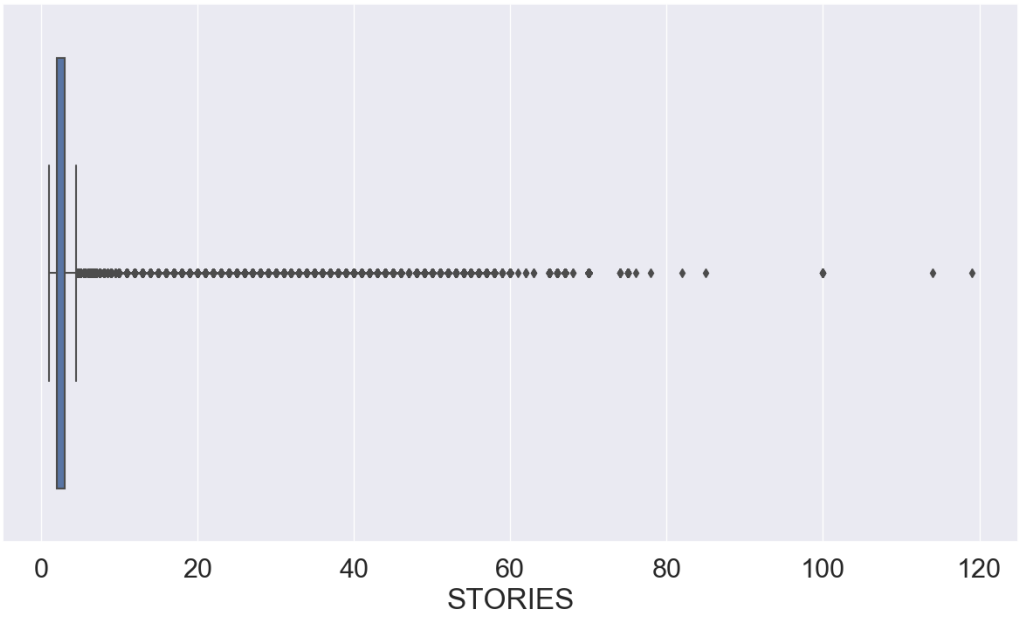
1. **TAXCLASS**: Current Property Tax Class Code (NYS Classification)

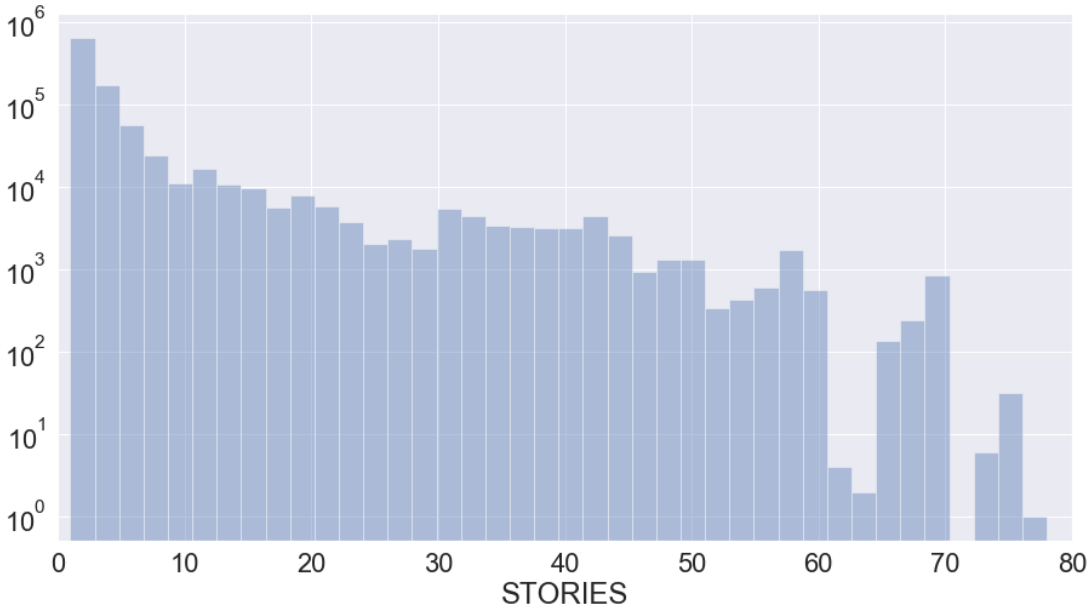


2. **B**: Borough codes. 1 means Manhattan, 2 means Bronx, 3 means Brooklyn, 4 means Queens, and 5 means Staten Island.

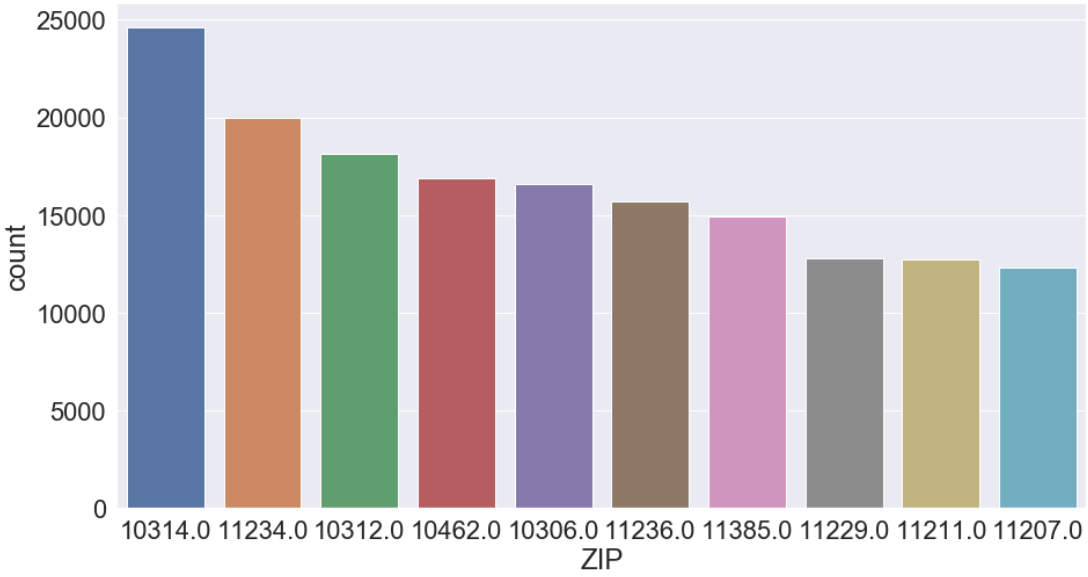


3. **STORIES**: The number of stories for the building (number of Floors)

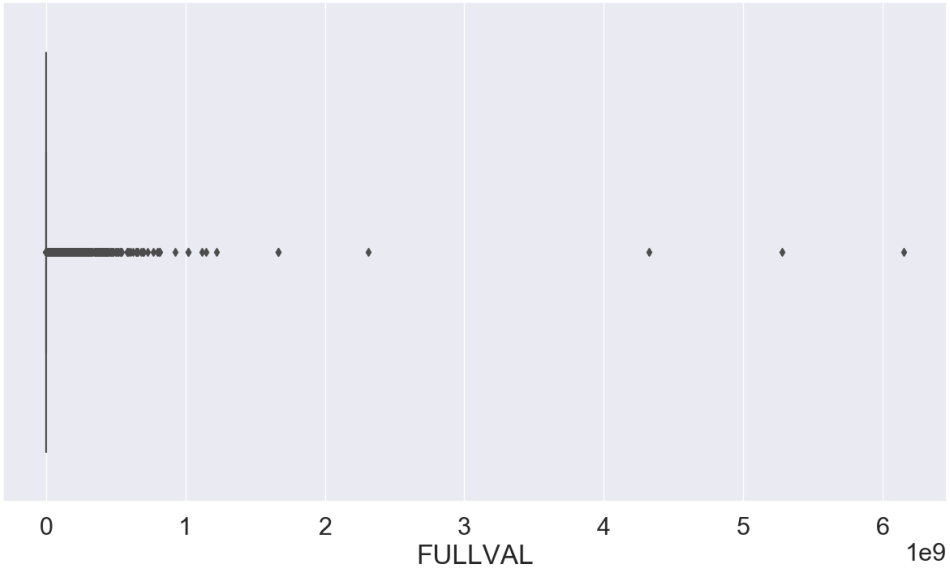


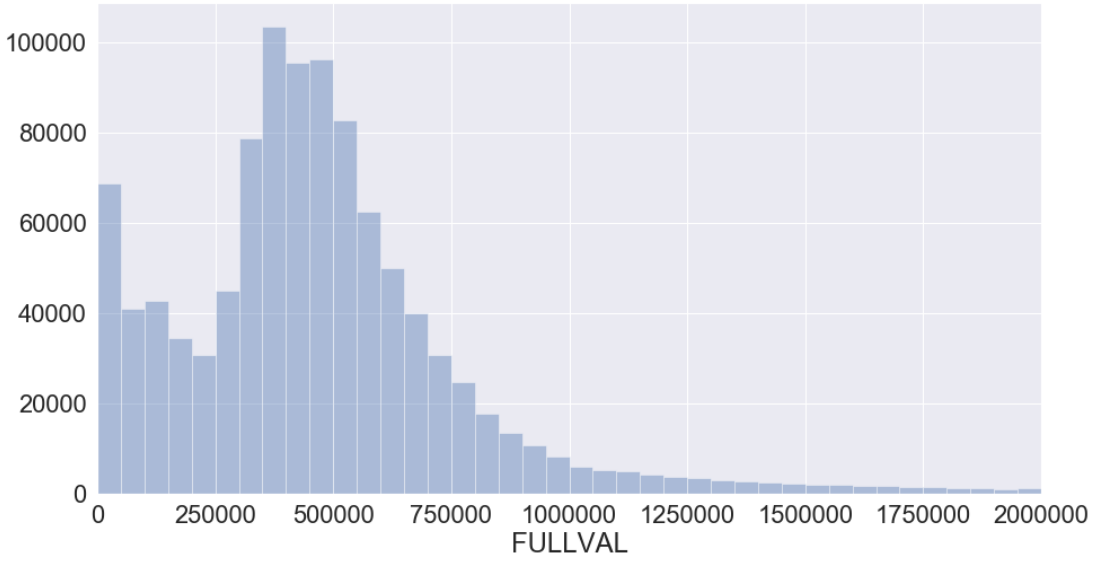


4. **ZIP**: Postal zip code of property

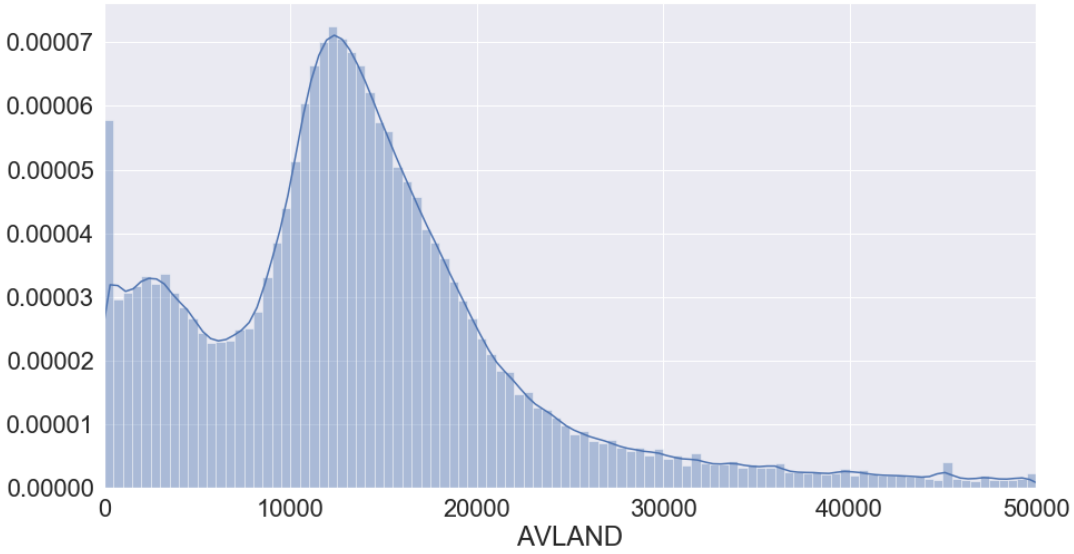


5. **FULLVAL**: Current year's total market value of the land (if not zero)

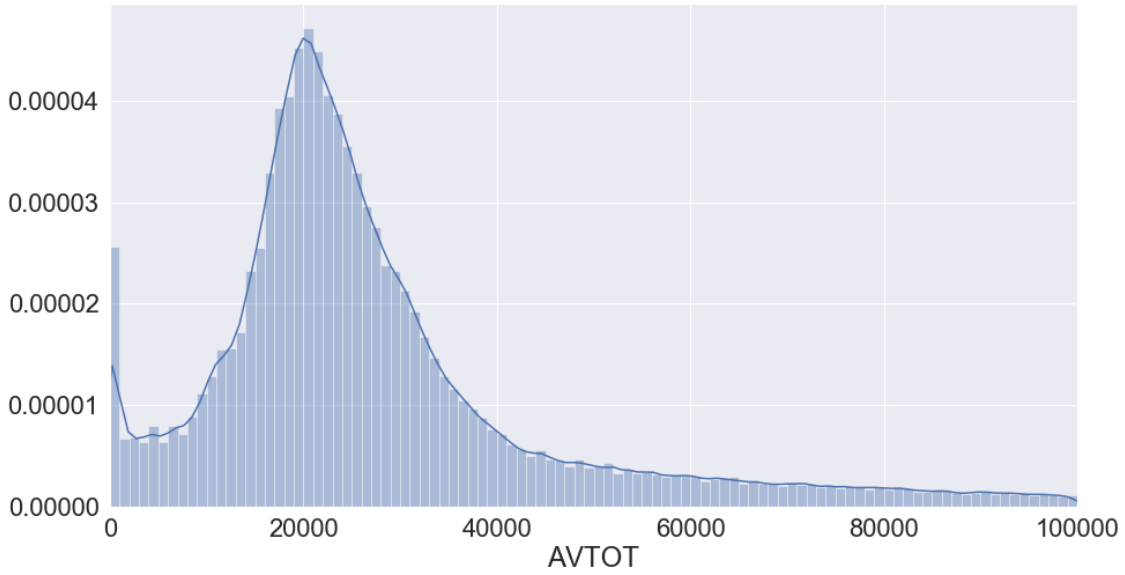




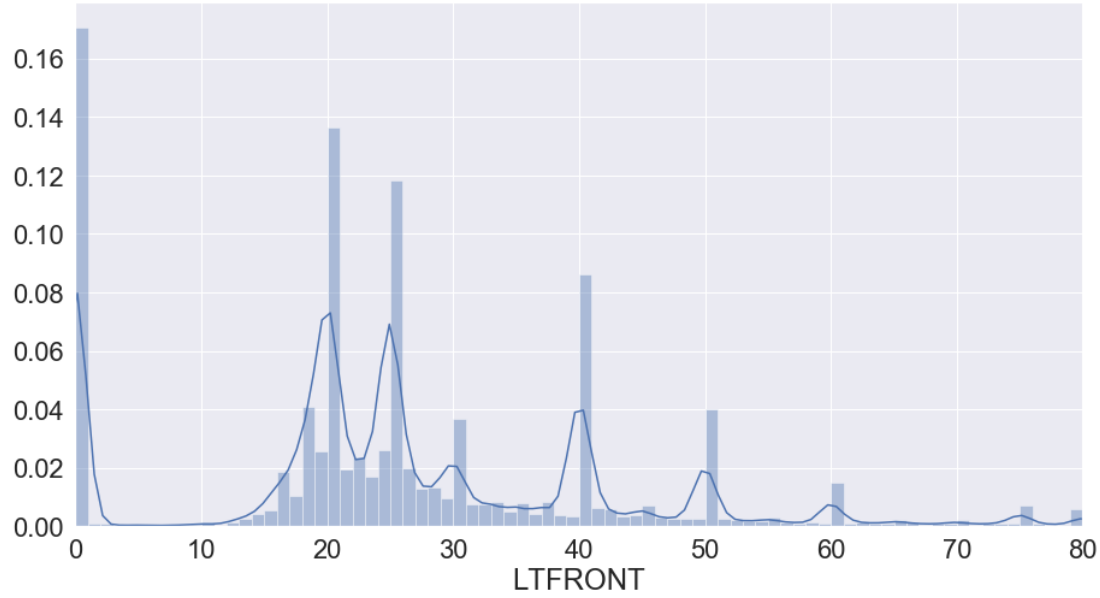
6. **AVLAND**: Assessed land value



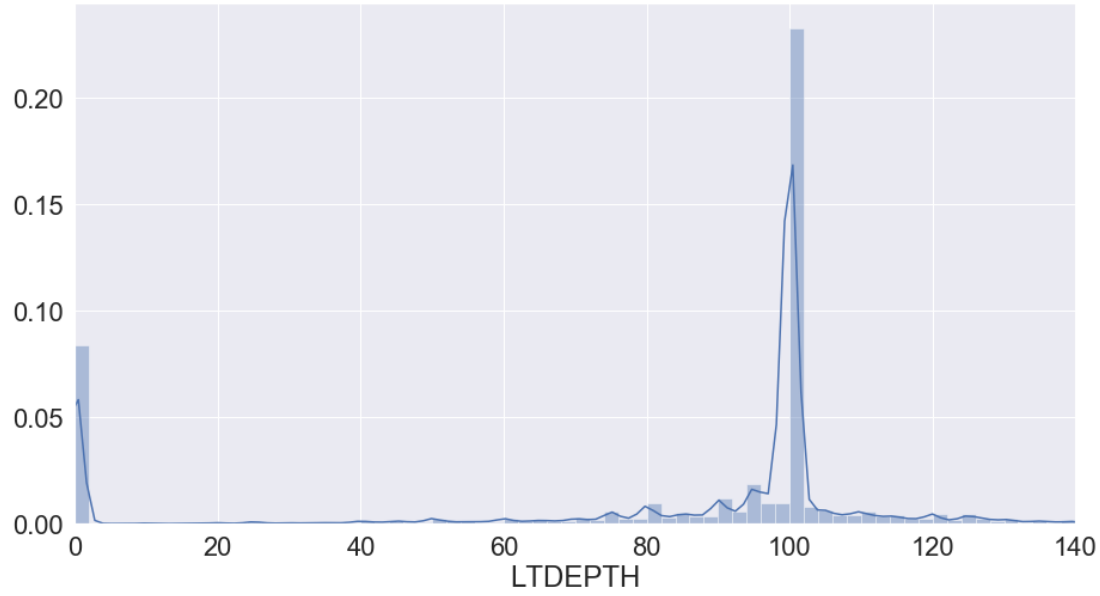
7. **AVTOT**: Assessed total value



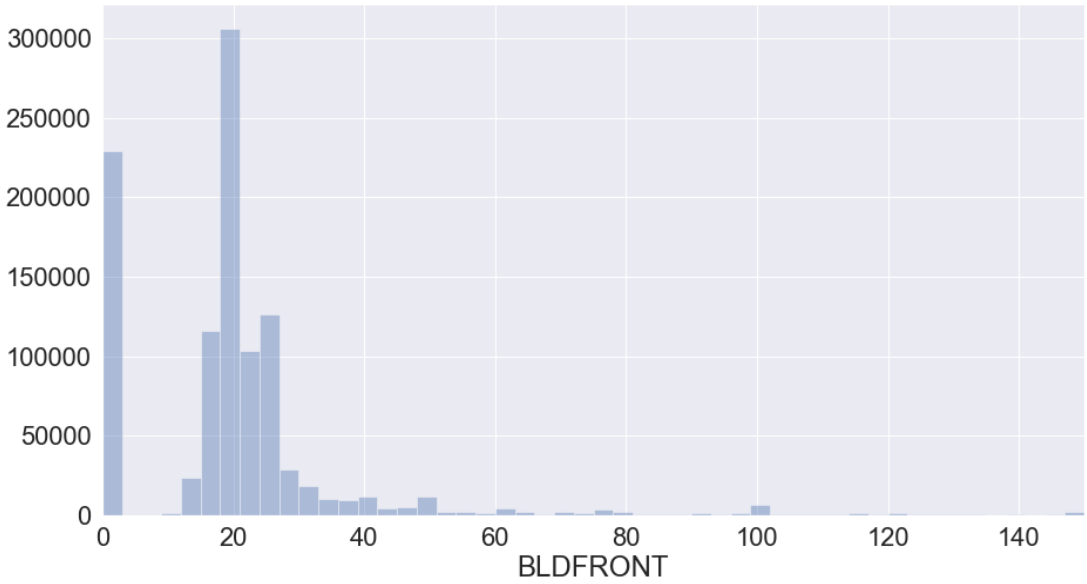
8. **LTFRONT**: Lot frontage in feet



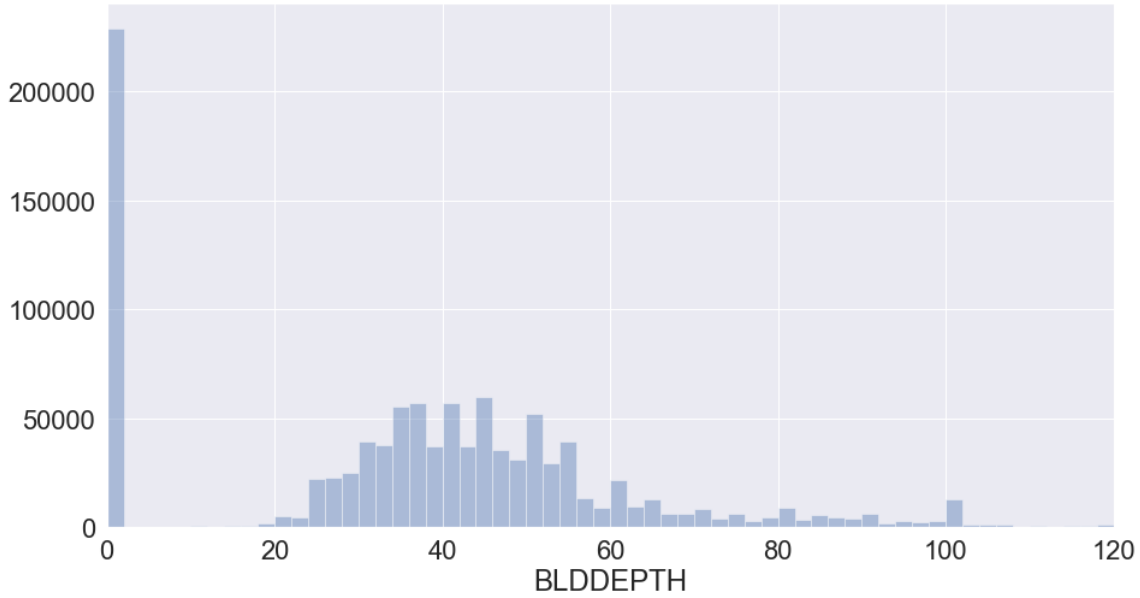
9. **LTDEPTH**: Lot depth in feet



10. **BLDFRONT**: Building frontage in feet



11. **BLDDEPTH**: Building depth in feet



III. Data Cleaning

To get more information, the missing values should not be ignored, thus the missing field should be filled in by some methods. From the important variables, the columns of FULLVAL, AVLAND, AVTOT, ZIP, STORIES, LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH have missing values.

We first fill missing values in the variables of ZIP because we can fill in other variable's missing values through grouping by the variable ZIP. We group by variables B and TAXCLASS to get the most frequent value in each subset and replace the missing zip number for the most frequent records in the same tax class code and borough code. There will be some missing value left because some subsets are too small to have useful zip records. Then I group the dataset only by variable B and replace the missing value by the most frequent records in that subset.

Then we start to fill empty values in STORIES. We group by ZIP and TAXCLASS and then take average to each group to fill in the null value. For the remaining missing data, we replace the null value with the average value of STORIES in the same zip code.

Next, we will exchange the 0 value in LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH. When LTFRONT and LTDEPTH are both 0, we will set LTFRONT to 30 and LTDEPTH to 100. For the remaining missing fields, we will first group by TAXCLASS and ZIP to replace the missing value in LTFRONT by the mean of LTFRONT (except for 0 value) in the

same TAXCLASS and ZIP group. There are still some fields left because of the small subset. We will replace those missing fields with the mean of LTFRONT in the same ZIP number. Missing value in LTDEPTH will be filled in the same way. Moreover, when BLDFRONT and BLDDEPTH are both 0, we will set BLDFRONT to 20 and BLDDEPTH to 40. The other value will follow the same logic as LTFRONT and LTDEPTH. However, for BLDDEPTH, there is still some value missing. We then take one more step that we replace the missing value with the average value of BLDDEPTH in the same tax class code.

Last, we will replace 0 value in FULLVAL, AVLAND, and AVTOT. We will fill these three variables with the average corresponding variables of all the records in the same ZIP and TAXCLASS subset. And then we replace the remaining missing value with the mean value in the same ZIP number.

IV. Variable Creation

Property assessed value is the most important thing we should focus on when we want to detect property tax fraud. Too high or too low the property assessed value is a sign to alert us that there might be something wrong. Therefore, this part will talk about how to build reasonable variables that are used for later models to check property assessed value anomaly.

4.1 Use property value and size to calculate normalized property values for each record

Purely comparing the total value of different properties is not a smart idea, because property values vary quite a lot, depending on how much area and space the property occupied. Therefore, it's better to calculate normalized values by property's area or space for each property, and then compare these normalized values among different properties to see the real property value per square foot or per volume unit.

(1) Find values of properties

In the NY property dataset, there are three fields that related the value of each property (label them v1 to v3):

- v1= FULLVAL
- v2= AVLAND
- v3= AVTOT

"FULLVAL", which means "Current year's total market value of the property"; "AVLAND", which means "Actual land value of the property"; and "AVTOT", which means "Actual total value of the property". This project will use all these three variables to represent the overall value of each property.

(2) Find sizes of properties

Property size can be shown in different ways, and in this project, we create three metrics to represent the size: lot area, building area, and building volume. (Label them s1 to s3)

- s1= lotarea = LTFRONT * LTDEPTH
- s2= bldarea =BLDFRONT * BLDDEPTH
- s3= bldvol = bldarea * STORIES

Lot area can be calculated by lot frontage (width) in feet times lot depth in feet. Building area can be calculated by building frontage (width) in feet times building depth in feet. These are the main metrics representing the area of each property. However, since New York City has so many high-rise buildings that may cost a lot for their constructions, how tall the building is also a very important factor to represent property size. Thus, we also create the building volume metric, bldvol, which is calculated by building area times the number of floors in that building.

(3) Normalize three values by three sizes

For each property, total market value, actual land value, and actual total value will be divided by its lot area, building area, and building volume one by one, to get 9 normalized values (per square foot or per volume unit). Therefore, each record now has 9 ratios below (label them r1 to r9):

- r1=v1/s1=fullval_lotarea=FULLVAL/lotarea
- r2=v1/s2=fullval_bldarea=FULLVAL/ bldarea
- r3=v1/s3=fullval_bldvol=FULLVAL/ bldvol
- r4=v2/s1=avland_lotarea=AVLAND/ lotarea
- r5=v2/s2=avland_bldarea=AVLAND/ bldarea
- r6=v2/s3=avland_bldvol=AVLAND/ bldvol
- r7=v3/s1=avtot_lotarea=AVTOT/ lotarea
- r8=v3/s2=avtot_bldarea=AVTOT/ bldarea
- r9=v3/s3=avtot_bldvol=AVTOT/ bldvol

4.2 Compare property's normalized values with group average values

After we get the normalized property values, it's still hard to compare among properties, because different properties have different geographical locations and tax classes, which lead to

different normalized values. Only within the same group or same class, can we compare those values. So next, we will calculate the average normalized property values by different groups, and then see whether each normalized property value is too high or too low, compared to the average values in the same group.

(1) Identify groups

Groups can be geographic groups, such as zip code, and virtual groups, such as tax classes. In this project, we picked five aspects for grouping: zip5, zip3, borough, taxclass, and all. (Label them g1 to g5)

- g1 =Zip 5: This is the most common factor to divide properties into specific geographic groups.
- g2 =Zip 3: It only keeps the first 3 digits from zip 5, to show broader geographic groups.
- g3 =Taxclass: Tax class is a logical grouping method to group properties that share the same characteristics, like schools, shopping malls, hospitals, and etc.
- g4 =Borough: This is an even broader geographic representation, which only has five types: Manhattan, Bronx, Brooklyn, Queens, and Staten Island.
- g5 =All: The overall normalized property values in this entire dataset.

(2) Scale 9 ratios by 5 group average levels

After defining these five groups, for each ratio (r1 to r9), we calculate the average ratios in each group (g1 to g5), and then we get totally 45 average group ratios and label them $< r1 >_g$, $< r2 >_g$, $< r3 >_g$, $< r4 >_g$, $< r5 >_g$, $< r6 >_g$, $< r7 >_g$, $< r8 >_g$, $< r9 >_g$. (g=1, 2, 3, 4, 5)

Then, for each record, we use its 9 ratios divided by these 45 average group ratios to scale, getting 45 scaled ratios.

$$\frac{r_1}{< r_1 >_g}, \quad \frac{r_2}{< r_2 >_g}, \quad \frac{r_3}{< r_3 >_g}, \quad \cdots \quad \frac{r_9}{< r_9 >_g} \qquad g = 1, \ldots, 5$$

These 45 scaled ratios show the relationship between each record ratios and group average ratios (If the scaled ratio is greater than 1, that means the specific property has more value than group average level; If the scaled ratio is smaller than 1, that means the specific property has less value than group average level; If the scaled ratio equal to 1, that means the specific property has the same value as group average level)

(3) Final 45 variables

The final 45 variables used in our models are those 45 scaled ratios we just calculated above. There is the list. (Noted: these 45 variables or 45 scaled ratios represent the comparison between "individual ratios r1 to r9 for each record" with "group average ratios", so the mean of these 45 variables are always "1". In this way, all the 45 variables will have the same mean, which will facilitate us to compare different variables and do transformation or calculation among them.)

	mean	std	min	max
zip5_fullval_lotarea	1	7.873526	1.58E-06	5363.634
zip5_fullval_bldarea	1	8.722907	2.00E-06	3340.95
zip5_fullval_bldvol	1	10.09991	5.62E-07	5055.745
zip5_avland_lotarea	1	13.33162	1.99E-06	7054.432
zip5_avland_bldarea	1	19.1314	6.58E-07	6980.759
zip5_avland_bldvol	1	21.03732	2.69E-07	6393.941
zip5_avtot_lotarea	1	11.54026	1.03E-06	6291.145
zip5_avtot_bldarea	1	16.24947	2.94E-06	6190.508
zip5_avtot_bldvol	1	17.34815	1.05E-06	5969.527
zip3_fullval_lotarea	1	9.798273	2.01E-06	5774.685
zip3_fullval_bldarea	1	13.28988	3.96E-06	4684.988
zip3_fullval_bldvol	1	14.37989	8.91E-07	8284.668
zip3_avland_lotarea	1	32.25102	8.88E-07	23835.4
zip3_avland_bldarea	1	75.99255	3.50E-06	46119.79
zip3_avland_bldvol	1	71.76642	4.20E-07	44079.66
zip3_avtot_lotarea	1	21.72625	5.85E-07	18824.25
zip3_avtot_bldarea	1	65.13738	8.69E-06	52898.78
zip3_avtot_bldvol	1	59.03888	1.54E-06	39407.37
taxclass_fullval_lotarea	1	8.342609	1.55E-06	5130.76
taxclass_fullval_bldarea	1	18.31379	3.05E-06	14256.41
taxclass_fullval_bldvol	1	68.5831	2.59E-06	70079.92
taxclass_avland_lotarea	1	10.92	2.10E-06	7199.866
taxclass_avland_bldarea	1	47.32591	4.23E-06	46010.66
taxclass_avland_bldvol	1	127.4649	6.66E-07	131394.6
taxclass_avtot_lotarea	1	10.20532	1.55E-06	7148.569
taxclass_avtot_bldarea	1	20.22914	1.13E-05	14256.34
taxclass_avtot_bldvol	1	69.02461	2.82E-06	70079.49
borough_fullval_lotarea	1	9.521606	2.42E-06	5774.685
borough_fullval_bldarea	1	14.033	4.08E-06	5527.622
borough_fullval_bldvol	1	14.88127	9.18E-07	8846.054
borough_avland_lotarea	1	36.02012	8.88E-07	29269.66
borough_avland_bldarea	1	77.15613	3.66E-06	46129.55
borough_avland_bldvol	1	77.36472	4.43E-07	56001.92

borough_avtot_lotarea	1	25.77877	5.85E-07	23749.46
borough_avtot_bldarea	1	63.68487	8.97E-06	44566.14
borough_avtot_bldvol	1	58.14723	1.60E-06	36431.71
all_fullval_lotarea	1	8.02823	1.97E-06	3661.962
all_fullval_bldarea	1	14.61628	4.38E-06	5940.317
all_fullval_bldvol	1	14.51842	4.87E-07	10608.2
all_avland_lotarea	1	25.22862	6.74E-07	17503.32
all_avland_bldarea	1	77.29099	7.28E-06	35066.62
all_avland_bldvol	1	85.02154	5.47E-07	69448.28
all_avtot_lotarea	1	18.03636	2.67E-07	12014.92
all_avtot_bldarea	1	60.45569	7.84E-06	43469.4
all_avtot_bldvol	1	65.45633	2.01E-06	43790

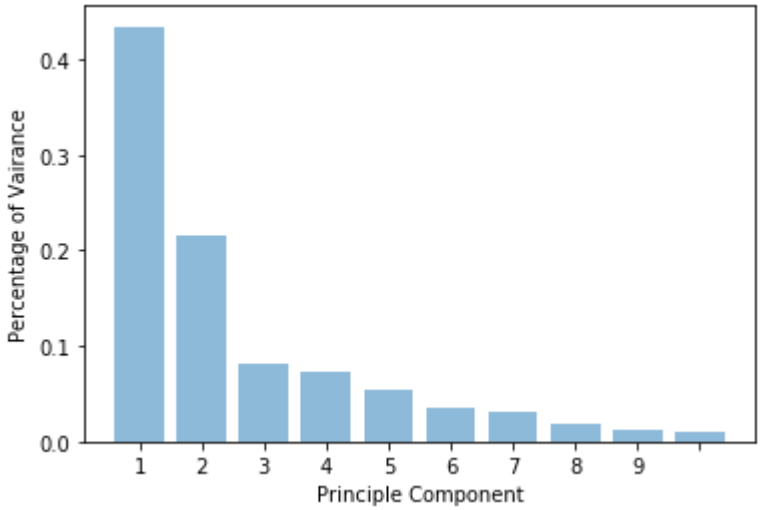
V. Dimensionality Reduction

The advantages of dimensionality reduction are (1) reducing the time and storage space required; (2) removing linear collinearity; (3) limiting the problems that appear when the dimensionality increases; (4) improving efficiency of the machine learning process.

Before reducing the dimensionality, we need to scale the 45 variables to make sure that all the variables have equal importance by using z scaling ($z' = (z - \mu) / \sigma$). Some variables have large differences between their ranges or measured in different measurement units. To get a much more accurate understanding of the influences of each variable, we need to scale the variables.

Based on the correlation matrix (Heat Map in Appendix B) we created above for the 45 variables we created, we can see that many variables are highly correlated. In order to remove those linear correlations among the variables, we applied Principal Component Regression Method (PCR) to reduce dimension but still have all the records in the eigenvalues.

The flowing is the scree plot of the top 10 magnitudes of the PCs, which are proportional to the variance in that PC's directions.



We rewrite the scaled data in terms of the rotated coordinate system (the PCs) after calculating all the PCs. We try to account for about 80% of the total variance. Based on the graphs above, we decide to reduce the dimension to 4. Therefore, we throw away all but the top 4 PCs as they contribute to the vast majority of the system's energy.

After doing the PCR, we need to z scale the new data (with only 4 columns at this time). In this way, we can make sure that each PC has been equally important. After scaling, it's much easier to find the outlier as we only need to calculate their distance to the origin for all dimensions.

VI. Algorithms

After the PCA and the second-time Z scaling, we retain 4 PCs. These reduced dimensions are relatively uncorrelated, similarly scaled, accounting for most of the total variance, and the data points are centred (mean is 0). Now for each record in the data set, the value of each variable explicitly shows how unusual or anomalous that record is in that dimension. We call each of these z scaled variables z score, which is our first type of fraud scores. Without letting them cancel each other out, we add up z scores on each record using the Euclidean distance with the formula of $s_1 = (\sum_{k=1}^{4} |PC_k|^2)^{\frac{1}{2}}$. As a result, we have the first z fraud score (s1) for each of the 1,070,994 records. The z score is a reasonable anomaly detector because it utilizes the Mahalanobis-like distance concept to look for outliers after taking into account scaling and correlations. Like Mahalobonis distance, z score measures the distance of each variable from the origin point.

Then for the second type of fraud scores, we train an autoencoder on the entire data set after the PCA and the second Z scaling. To be specific, an autoencoder is essentially a 2-layer

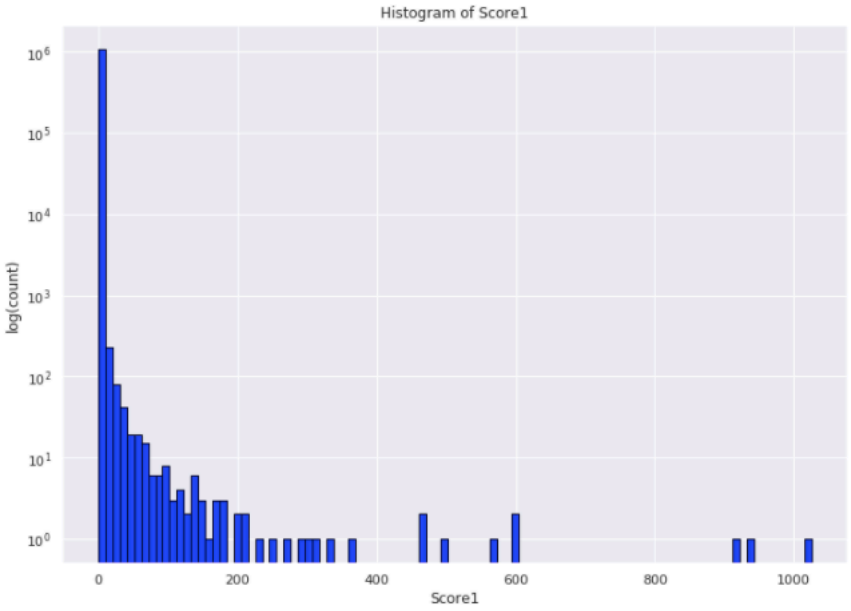
neural network satisfying that (1) the hidden layer is smaller than the size of the input and output layer and (2) the input layer and output layer are the same size. Thus, for our autoencoder model, there is one input layer with an input dimension of 4 due to the reduced 4 PCs, one hidden layer with 3 nodes, and one output layer with the same size as the input layer. After the model is trained, we calculate the distance between the original input vector and the model output vector, which is our second type of fraud scores. We use the Euclidean distance as well with the formula of $s_2 = (\sum_{k=1}^{4} |PC_k - PC_k'|^2)^{\frac{1}{2}}$. As a result, we have the second autoencoder fraud score (s2) for each of the 1,070,994 records. The autoencoder score is a reasonable anomaly detector, because the autoencoder model will learn to reproduce the data records as well as possible and will learn the nature of the bulk of the data. The records that are not reproduced well are unusual records, so a measure of the reproduction error is a measure of unusualness for that record.

To combine the two scores to get the final score, we use extreme quantile binning/rank ordering to scale each score (s1, s2). We sort records by the value of each score in ascending order and replace the original score values with the record's rank order from 1 to 1,070,994. Then we calculate the weighted average of two scaled scores/ranks to get the final fraud score (final fraud score = (scaled s1 + scaled s2) / 2). The top 10 records are those with the top 10 highest final fraud scores.

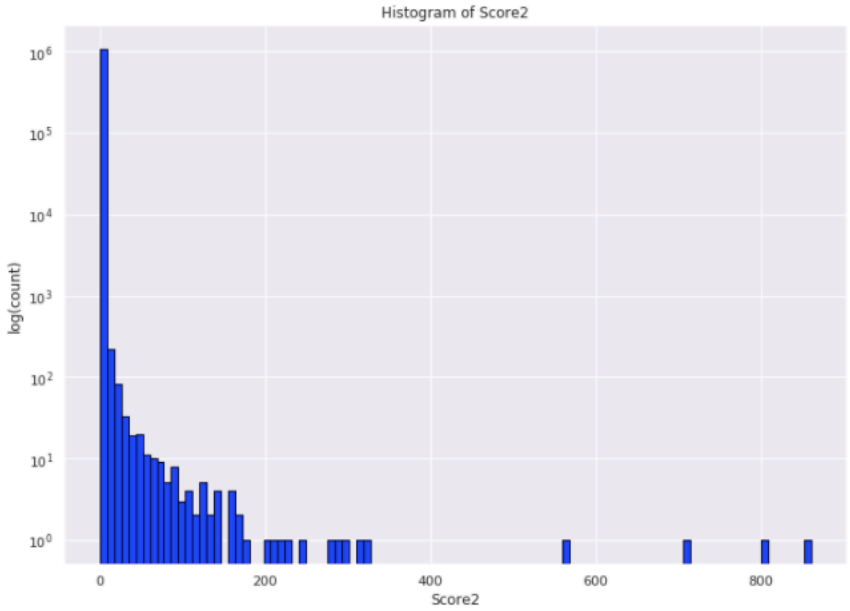
VII. Results

We could see the distribution more clearly by log transforming the count number.

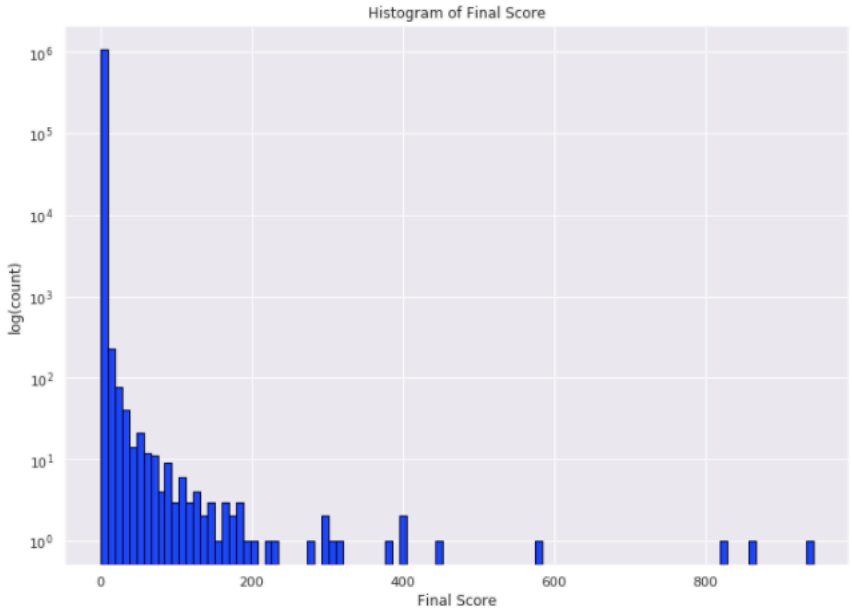
Score 1:



Score2:



Final score:



Top 10 records that look anomalous and explanations:

Anomaly #1 (#632816)

One abnormal variable among the 45 created variables is TAXCLASS_avtot_bldvol, which has a 1008.43 Z-score. As the record 632816 shows below, this means that record 632816 has extremely high AVTOT (1318500) related to BLDVOL (1*1) in TAXCLASS "2" group.

Another abnormal variable among the 45 created variables is TAXCLASS_avland_bldvol, which has a 1008.43 Z-score. As the record 632816 shows below, this means that record 632816 has extremely high AVLAND (1318500) related to BLDVOL (1*1) in TAXCLASS "2" group.

RECORD	B	TAXCLASS	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP	LTFRONT	LTDEPTH
632816	4	2	1	1	1	2930000	1318500	1318500	11373	157	95

Anomaly #2 (#917942)

One abnormal variable among the 45 created variables is ZIP3_avtot_bldarea, which has a 729.15 Z-score. As the record 917942 shows below, this means that record 917942 has extremely high AVTOT (46,7000,0000) related to BLDAREA (29*51) in ZIP3 "114" group.

Another abnormal variable among the 45 created variables is all_avtot_bldvol, which has a 564.96 Z-score. As the record 917942 shows below, this means that record 917942 has extremely high AVTOT (46,7000,0000) related to BLDVOL (29*51*3) in all.

RECORD	B	TAXCLASS	BLOCK	ZIP	STORIES	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH	LTFRONT	LTDEPTH
917942	4	4	14260	11422	3	3.74E+08	1.79E+09	4.67E+09	29.303352	50.77192	4910	105.6431

Anomaly #3 (#1067360)

One abnormal variable among the 45 created variables is ZIP3_fullval_lotarea, which has a 721.82 Z-score. As the record 1067360 shows below, this means that record 1067360 has high FULLVAL (83600) related to LOTAREA (1*1) in ZIP3 "103" group.

Another abnormal variable among the 45 created variables is B_fullval_lotarea, which has a 719.84 Z-score. As the record 1067360 shows below, this means that record 1067360 has high FULLVAL (83600) related to LOTAREA (1*1) in B "5" group.

RECORD	STORIES	B	FULLVAL	AVLAND	AVTOT	ZIP	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH
1067360	2	5	836000	28800	50160	10307	1	1	1	36	45

Anomaly #4 (#750816)

One abnormal variable among the 45 created variables is TAXCLASS_avtot_lotarea, which has a 780.82 Z-score. As the record 750816 shows below, this means that record 750816 has high AVTOT (130298) related to LOTAREA (1*1) in TAXCLASS "IB" group.

Another abnormal variable among the 45 created variables is TAXCLASS_avland_lotarea, which has a 780.82 Z-score. As the record 750816 shows below, this means that record 750816 has high AVLAND (47675) related to LOTAREA (1*1) in TAXCLASS "IB" group.

RECORD	STORIES	B	FULLVAL	AVLAND	AVTOT	ZIP	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH
750816	5	4	658636	47675.27	130298.2	11367	1B	1	1	29.30335	50.77192

Anomaly #5 (#565392)

One abnormal variable among the 45 created variables is B_avland_lotarea, which has a 856.51 Z-score. As the record 565392 shows below, this means that record 565392 has high AVLAND (19,5000,0000) related to LOTAREA (117*108) in B "3" group.

Another abnormal variable among the 45 created variables is ZIP5_avland_lotarea, which has a 626.49 Z-score. As the record 565392 shows below, this means that record 565392 has high AVLAND (19,5000,0000) related to LOTAREA (117*108) in ZIP5 "11229" group.

RECORD	STORIES	B	FULLVAL	AVLAND	AVTOT	ZIP	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH
565392	5	3	4.33E+09	1.95E+09	1.95E+09	11229	4	117	108	29.30335	50.77192

Anomaly #6 (#585118)

One abnormal variable among the 45 created variables is B_avtot_bldarea, which has a 316.1 Z-score. As the record 585118 shows below, this means that record 585118 has high AVTOT (154,9530) related to BLDAREA (1*1) in B "4" group.

Another abnormal variable among the 45 created variables is all_avland_bldarea, which has a 441.86 Z-score. As the record 585118 shows below, this means that record 585118 has high AVTOT (154,9530) related to BLDAREA (1*1) in all.

RECORD	STORIES	B	FULLVAL	AVLAND	AVTOT	ZIP	TAXCLASS	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH
585118	20	4	3443400	1549530	1549530	11101	4	298	402	1	1

Anomaly #7 (#585439)

One abnormal variable among the 45 created variables is B_avtot_bldarea, which has a 340.75 Z-score. As the record 585439 shows below, this means that record 585439 has high AVTOT (1,670,400) related to BLDAREA (1*1) in B "4" group.

Another abnormal variable among the 45 created variables is all_fullval_bldarea, which has a 389.49 Z-score. As the record 585439 shows below, this means that record 585439 has high FULLVAL (371,2000) related to BLDAREA (1*1) in all.

RECORD	STORIES	B	FULLVAL	AVLAND	AVTOT	ZIP	BLDFRONT	TAXCLASS	LTFRONT	LTDEPTH	BLDDEPTH
585439	10	4	3712000	252000	1670400	11101	1	4	94	165	1

Anomaly #8 (#85886)

One abnormal variable among the 45 created variables is ZIP3_fullval_bldvol, which has a 445.69 Z-score. As the record 85886 shows below, this means that record 85886 has high FULLVAL (7021,4000) related to BLDVOL (8*8*1) in ZIP3 “100” group.

Another abnormal variable among the 45 created variables is ZIP5_fullval_bldvol, which has a 436.12 Z-score. As the record 85886 shows below, this means that record 85886 has high FULLVAL (7021,4000) related to BLDVOL (8*8*1) in ZIP5 “10025” group.

RECORD	STORIES	B	FULLVAL	AVLAND	AVTOT	ZIP	BLDFRONT	TAXCLASS	LTFRONT	LTDEPTH	BLDDEPTH
85886	1	1	70214000	31455000	31596300	10025	8	4	4000	150	8

Anomaly #9 (#67129)

One abnormal variable among the 45 created variables is ZIP3_fullval_bldarea, which has a 375.17 Z-score. As the record 67129 shows below, this means that record 67129 has high FULLVAL (61,5000,0000) related to BLDAREA (29*50) in ZIP3 “100” group.

Another abnormal variable among the 45 created variables is all_avland_bldarea, which has a 511.41 Z-score. As the record 67129 shows below, this means that record 67129 has high AVLAND (26,7000,0000) related to BLDAREA (29*50) in all.

RECORD	STORIES	B	FULLVAL	AVLAND	AVTOT	ZIP	BLDFRONT	TAXCLASS	LTFRONT	LTDEPTH	BLDDEPTH
67129	5	1	6.15E+09	2.67E+09	2.77E+09	10028	29.30335	4	840	105.6431	50.77192

Anomaly #10 (#565398)

One abnormal variable among the 45 created variables is B_avland_bldarea, which has a 321.88 Z-score. As the record 565398 shows below, this means that record 565398 has high AVLAND (10,4000,0000) related to BLDAREA (29*50) in B “3” group.

One abnormal variable among the 45 created variables is ZIP3_avland_bldarea, which has a 338.21 Z-score. As the record 565398 shows below, this means that record 565398 has high AVLAND (10,4000,0000) related to BLDAREA (29*50) in ZIP3 “112” group.

RECORD	STORIES	B	AVLAND	AVTOT	EXLAND	ZIP	BLDFRONT	TAXCLASS	LTFRONT	LTDEPTH	BLDDEPTH
565398	5	3	1.04E+09	1.04E+09	1.04E+09	11229	29.30335	4	466	1009	50.77192

VIII. Conclusions

In conclusion, in order to find records which have extremely low or high assessed values, we performed the following steps. We first filled in missing fields (FULLVAL, AVLAND, AVTOT, ZIP, STORIES, LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH) with the most typical value for that field or that record, so that they would not affect the fraud detection process. We then created 3 size variables that account for the area and volume of each property and 3 value variables that represent different measurements for property values, then calculated each property's value normalized by these 3 sizes variables and calculated the average normalized property values by different groups. In this way, we are able to see whether each normalized property value is too high or too low compared to the average values in the same group.

After we had all the necessary variables, we started building the unsupervised learning model. We first z-scaled the 45 variables to make sure that all the variables are on the same scale, then used PCA to choose 4 PCs that accounted for about 80% of the total variance. We then z-scaled the PCs again to make sure that each PC has equal importance. Our final fraud score is a combination of two scaled scores. We calculated score 1 for each of the 1,070,994 records using Euclidean distance. For score 2, we trained an autoencoder on the z-scaled PCs, and the fraud score is the reconstruction error. After using extreme quantile binning to scale each score, we took a mean value of the 2 scores as our final fraud score. Our 10 potential fraud records are the ones with top 10 final fraud scores.

If we had more time, we would do more detailed work when filling the missing fields. Our current method is to group by 2 variables and fill in the missing value using mean/mode or fill in an approximate value by common sense. We could group by more variables to make the filled value more accurate or do some research about NY properties to get a better idea of the approximate values. We would also try different layers and different activation functions for the autoencoder and tune the hyper-parameters to minimize the loss function.

Appendix

Appendix A

Section I: General description of data

Dataset Name: Property Valuation and Assessment Data

Data Description: Data represent NYC properties assessments for purpose to calculate Property Tax, Grant eligible properties Exemptions and/or Abatements. Data collected and entered into the system by various City employee, like Property Assessors, Property Exemption specialists, ACRIS reporting, Department of Building reporting, etc.

Source: NYC Open Data, Department of finance

Time period: 2010

Number of fields: 32

Number of records: 1,070,994

Section II: Summary Table

1. **Summary of all the fields in the dataset**

No.	Name	Data Type	Number populated	% Populated	# of Unique Value	# Zeros
1	RECORD	int64	1070994	100	1070994	0
2	BBLE	object	1070994	100	1070994	0
3	B	int64	1070994	100	5	0
4	BLOCK	int64	1070994	100	13984	0
5	LOT	int64	1070994	100	6366	0
6	EASEMENT	object	4636	0.433	13	0
7	OWNER	object	1039251	97.036	863347	0
8	BLDGCL	object	1070994	100	200	0
9	TAXCLASS	object	1070994	100	11	0
10	LTFRONT	int64	1070994	100	1297	169108
11	LTDEPTH	int64	1070994	100	1370	170128
12	EXT	object	354305	33.082	4	0
13	STORIES	float64	1014730	94.747	112	0
14	FULLVAL	float64	1070994	100	109324	13007
15	AVLAND	float64	1070994	100	70921	13009
16	AVTOT	float64	1070994	100	112914	13007
17	EXLAND	float64	1070994	100	33419	491699
18	EXTOT	float64	1070994	100	64255	432572
19	EXCD1	float64	638488	59.616	130	0
20	STADDR	object	1070318	99.937	839281	0
21	ZIP	float64	1041104	97.209	197	0
22	EXMPTCL	object	15579	1.455	15	0
23	BLDFRONT	int64	1070994	100	612	228815
24	BLDDEPTH	int64	1070994	100	621	228853

No.	Name	Data Type	Number populated	% Populated	# of Unique Value	# Zeros
25	AVLAND2	float64	282726	26.398	58592	0
26	AVTOT2	float64	282732	26.399	111361	0
27	EXLAND2	float64	87449	8.165	22196	0
28	EXTOT2	float64	130828	12.216	48349	0
29	EXCD2	float64	92948	8.679	61	0
30	PERIOD	object	1070994	100	1	0
31	YEAR	object	1070994	100	1	0
32	VALTYPE	object	1070994	100	1	0

2. Summary of numeric fields

Field	count	mean	std	min	p25	median	p75	max
RECORD	1070994	535497.5	309169	1	267749.25	535497.5	803245.75	1070994
LTFRONT	1070994	36.6353014	74	0	19	25	40	9999
LTDEPTH	1070994	88.861594	76	0	80	100	100	9999
STORIES	1014730	5.0069179	8	1	2	2	3	119
FULLVAL	1070994	874264.505	11582430	0	304000	447000	619000	6150000000
AVLAND	1070994	85067.9187	4057260	0	9180	13678	19740	2668500000
AVTOT	1070994	227238.169	6877529	0	18374	25340	45438	4668308947
EXLAND	1070994	36423.8907	3981575	0	0	1620	1620	2668500000
EXTOT	1070994	91186.9817	6508402	0	0	1620	2090	4668308947
EXCD1	638488	1602.01423	1384	1010	1017	1017	1017	7170
BLDFRONT	1070994	23.0427696	35	0	15	20	24	7575
BLDDEPTH	1070994	39.9228362	42	0	26	39	50	9393
AVLAND2	282726	246235.719	6178962	3	5705	20145	62640	2371005000
AVTOT2	282732	713911.436	11652528	3	33912	79962.5	240551	4501180002
EXLAND2	87449	351235.684	10802212	1	2090	3048	31779	2371005000
EXTOT2	130828	656768.282	16072510	7	2870	37062	106840.75	4501180002
EXCD2	92948	1364.04168	1094	1011	1017	1017	1017	7160

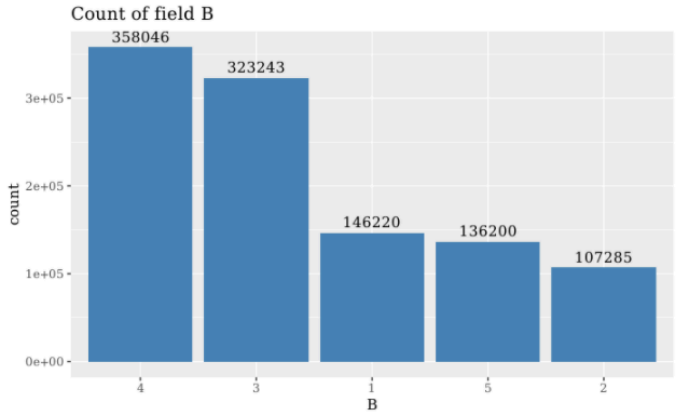
3. Summary of categorical fields

Field	n_obs	n_distinct	mode	freq
BBLE	1070994	1066541	2025390020	10
B	1070994	5	4	358046
BLOCK	1070994	13984	3944	3888
LOT	1070994	6366	1	24367
EASEMENT	4636	12	E	4148
OWNER	1039249	863348	PARKCHESTER PRESERVAT	6020
BLDGCL	1070994	200	R4	139879
TAXCLASS	1070994	11	1	660721
EXT	354305	3	G	266970
STADDR	1070318	839280	501 SURF AVENUE	902
ZIP	1041104	196	10314	24606
EXMPTCL	15579	14	X1	6912
PERIOD	1070994	1	FINAL	1070994
YEAR	1070994	1	2010/11	1070994
VALTYPE	1070994	1	AC-TR	1070994

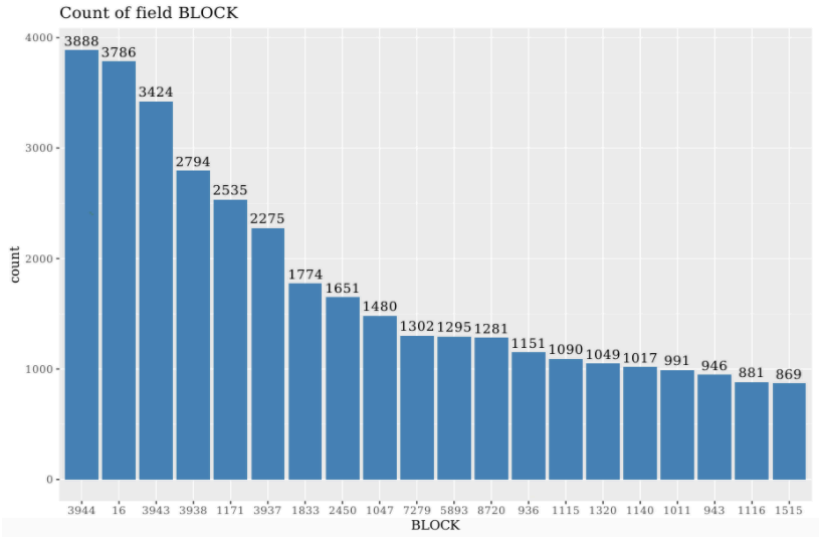
Section III: Detailed description of all the fields

In this section, every field will be explained in detail. For all the numeric fields, the section will show their distributions. In order to present a good pattern of distribution, the scales could be changed by logging y-axis, removing outliers, and limiting ranges. For all the categorical fields, the section will show histograms to count the frequency for each value.

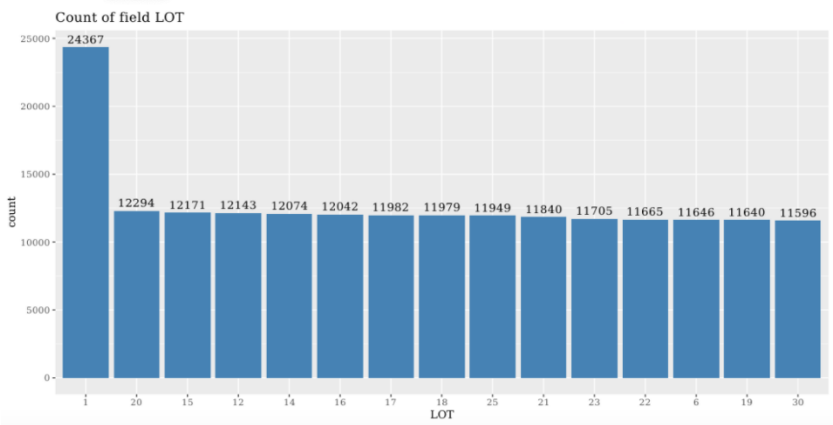
1. **Record:** Index for each record in the dataset
2. **BBLE:** Concatenation of AV_BORO, AV_BLOCK, AV_LOT, AV_EASEMENT. Almost all of the rows have their own BBLE number.
3. **B:** Borough codes, 1 represents Manhattan, 2 represents Bronx, 3 represents Brooklyn, 4 represents Queens, and 5 represents Staten Island.



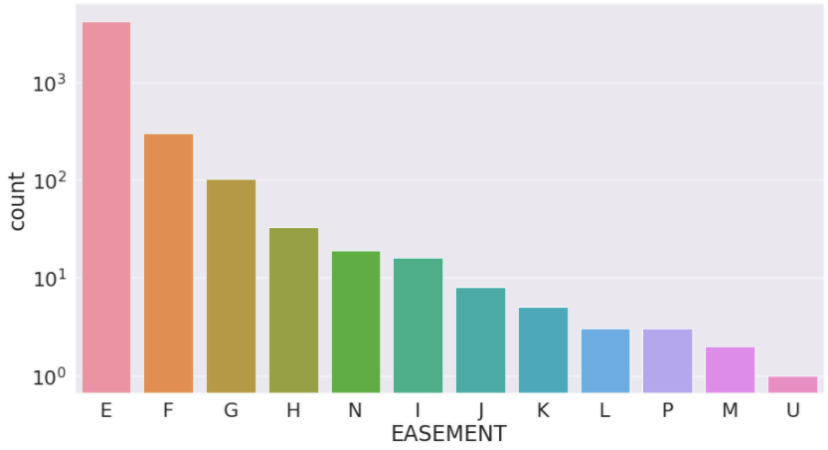
4. **BLOCK:** Valid block ranges by borough.
   Manhattan: 1 to 2,255; Bronx: 2,260 to 5,958; Brooklyn: 1 to 8,955; Queens: 1 to 16,350; Staten Island: 1 to 8,050.



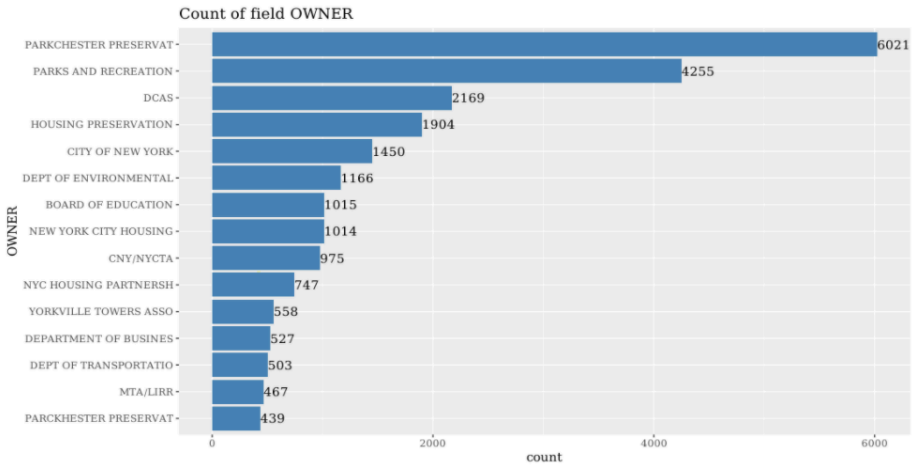
5. **LOT:** Unique number within BORO and BLOCK.



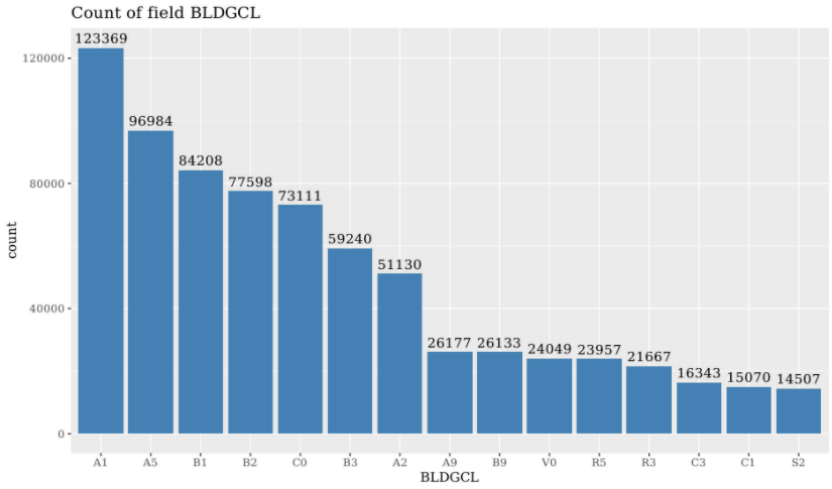
6. **EASEMENT:** Space indicates the lot has no easement. 'A' indicates the portion of the Lot that has an Air Easement, 'B' indicates Non-Air Rights, 'E' indicates the portion of the lot that has a Land Easement, 'F' THRU 'M' are duplicates of 'E', 'N' indicates Non-Transit Easement, 'P' indicates Piers, 'R' indicates Railroads, 'S' indicates Street, 'U' indicates U.S. Government.



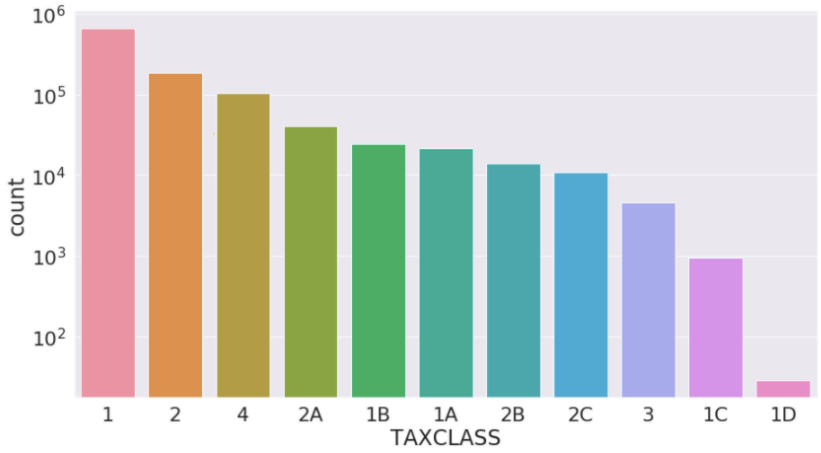
7. **OWNER:** The owner's name.



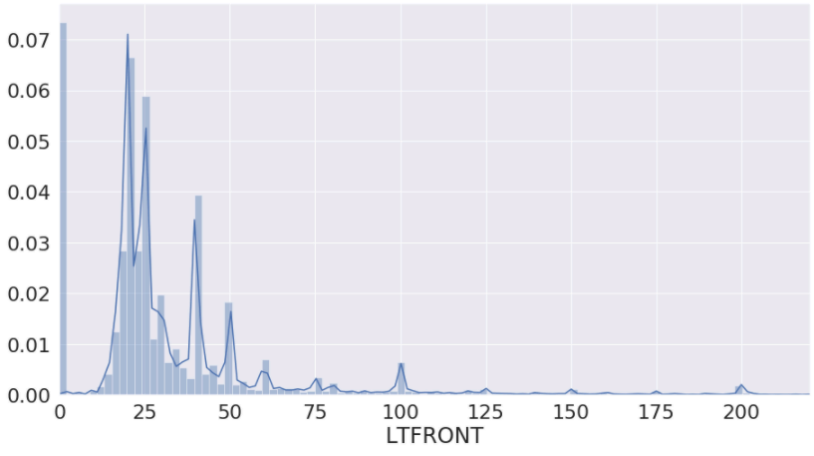
8. **BLDGCL:** Building class. Position 1 = ALPHA & Position 2 = NUMERIC. There is a direct correlation between the Building Class and the Tax Class



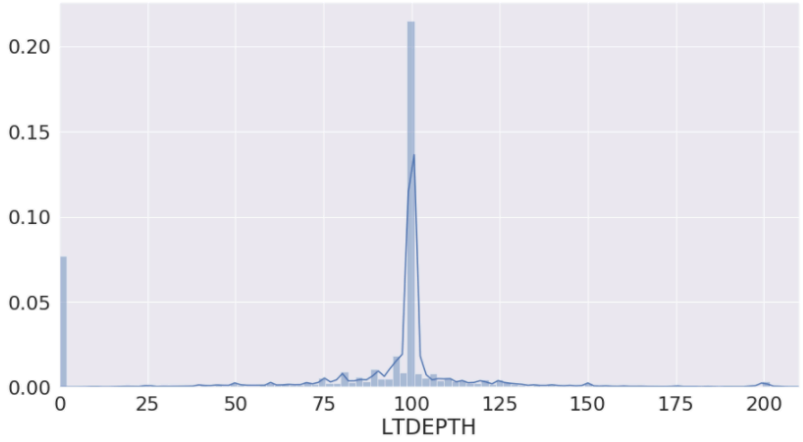
9. **TAXCLASS:** Contains the Tax Class at the Beginning of the Fiscal Year. 1 = 1- 3 Unit Residence, 2 = Apartments, 2A = 4, 5, or 6 Units, 3 = Utilities, 4 = All Others



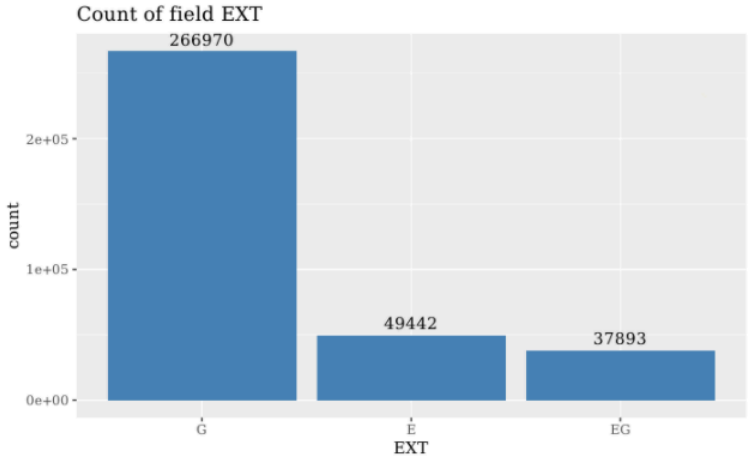
10. **LTFRONT:** Lot Frontage/Width in feet.



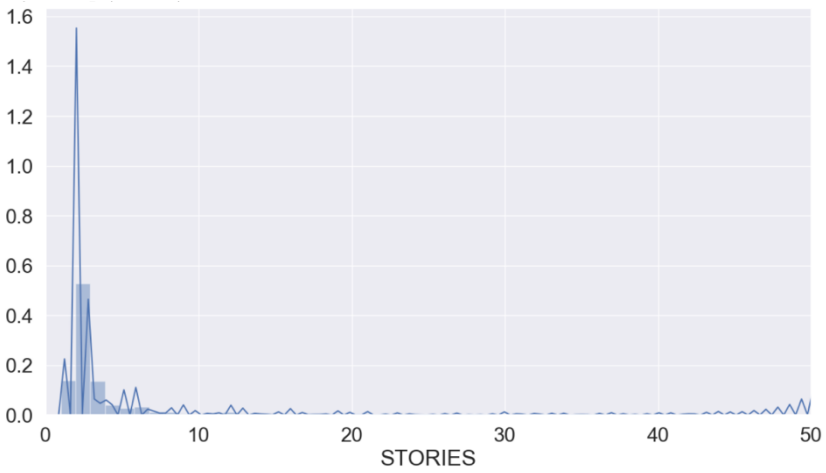
11. **LTDEPTH:** Lot Depth in feet.



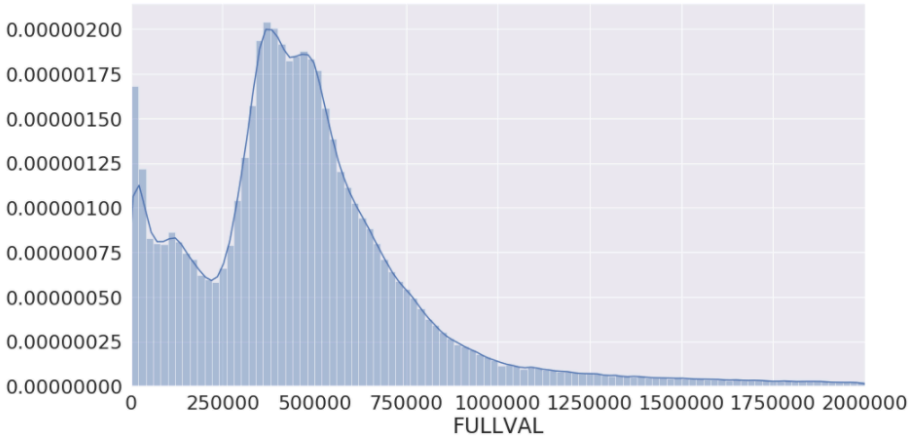
12. **EXT:** Extension Indicator. E=Extension, G=Garage, EG=Extension and garage



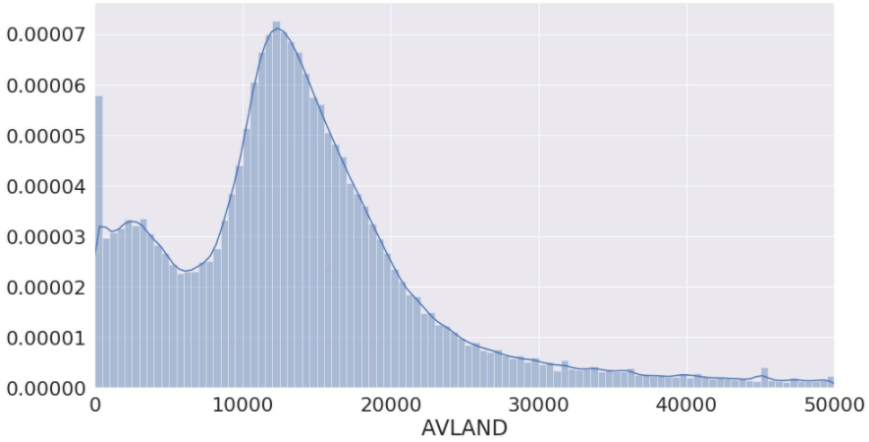
13. **STORIES:** The number of floors for the building



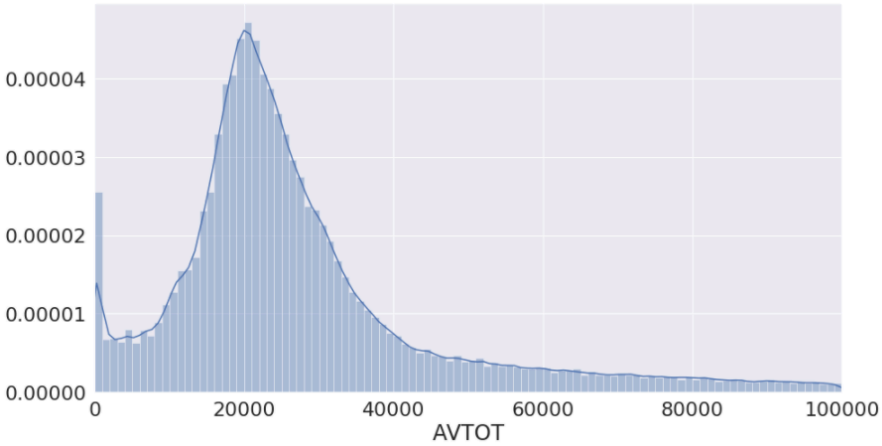
14. **FULLVAL:** Current year's total market value of the property.



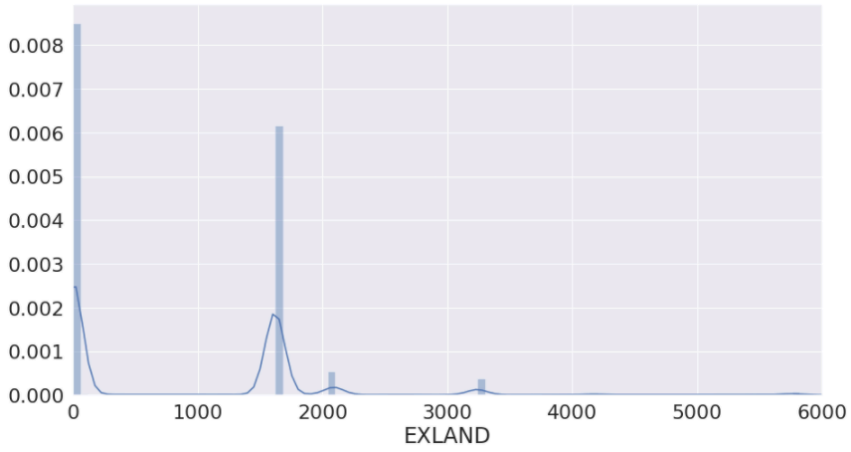
15. **AVLAND:** Actual Land Value.



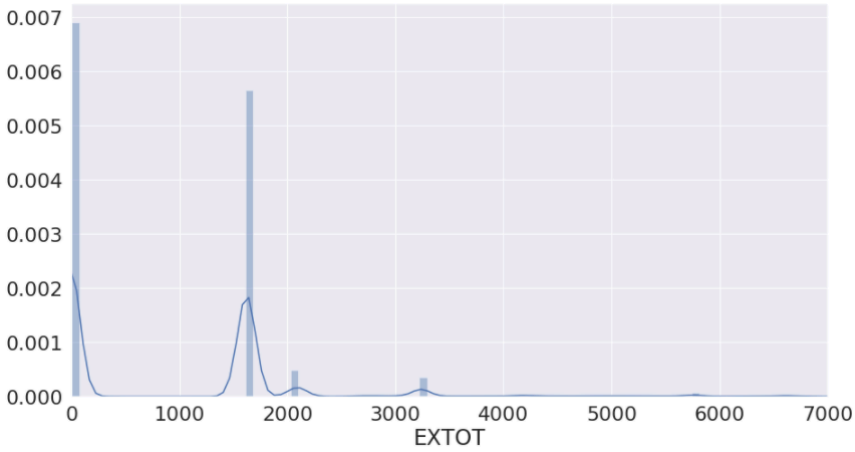
16. AVTOT: Actual Total Value.



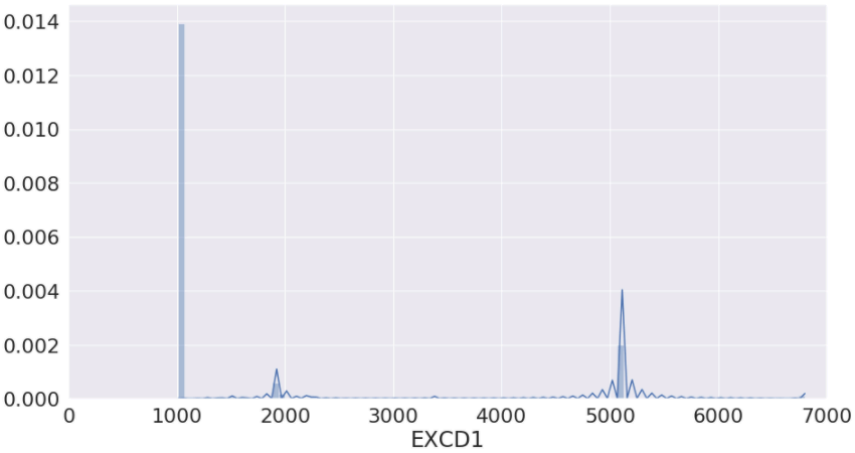
17. **EXLAND:** Actual Exempt Land Value.



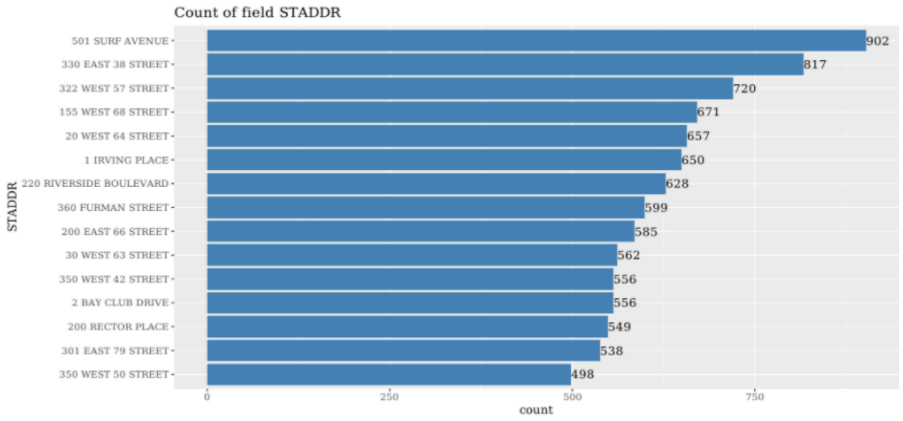
18. **EXTOT:** Actual Exempt Land Total.



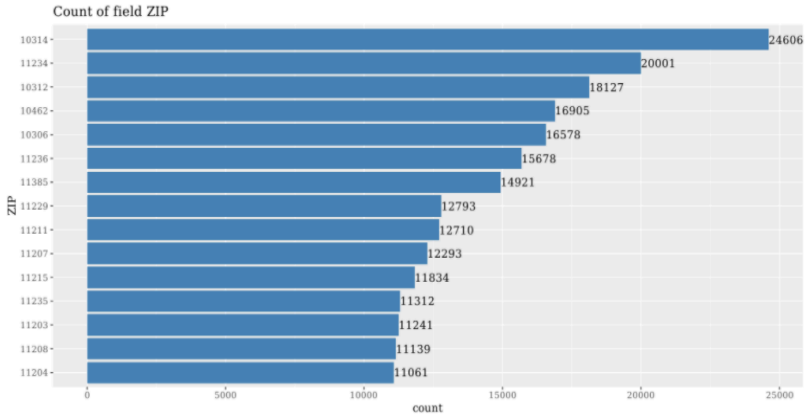
19. **EXCD1:** Exemption Code 1.



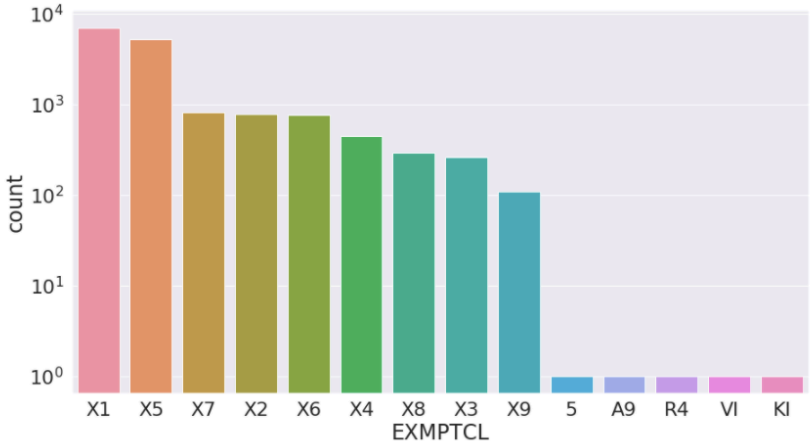
20. **STADDR:** Street name for the property



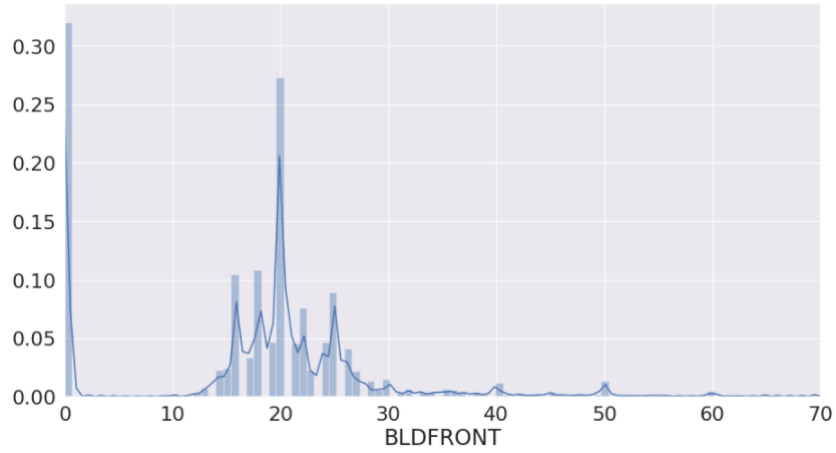
21. **ZIP:** Zip code



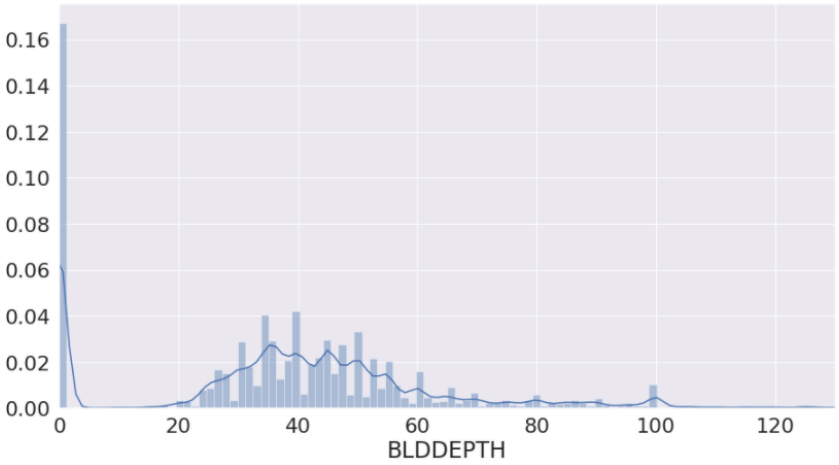
22. **EXMPTCL:** Exempt Class used for fully exempt properties only, including 'X1 - X9'. See Building Class Form for a description of the codes



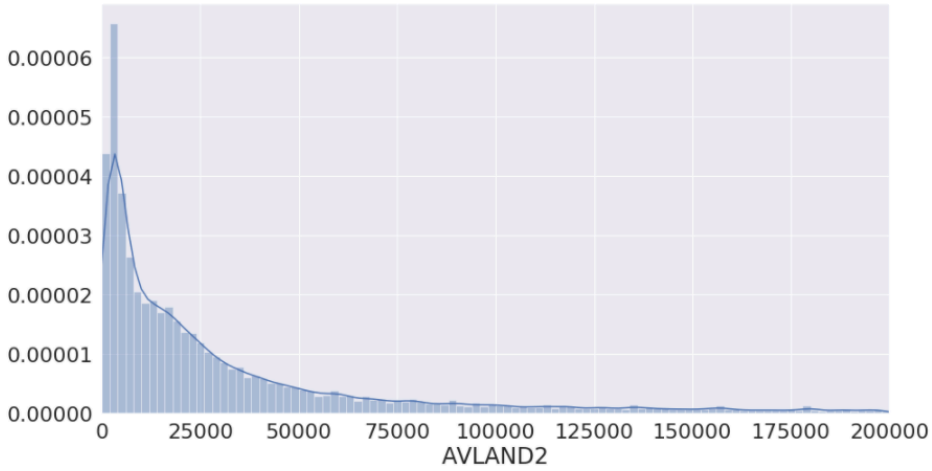
23. **BLDFRONT:** Building Frontage in feet.



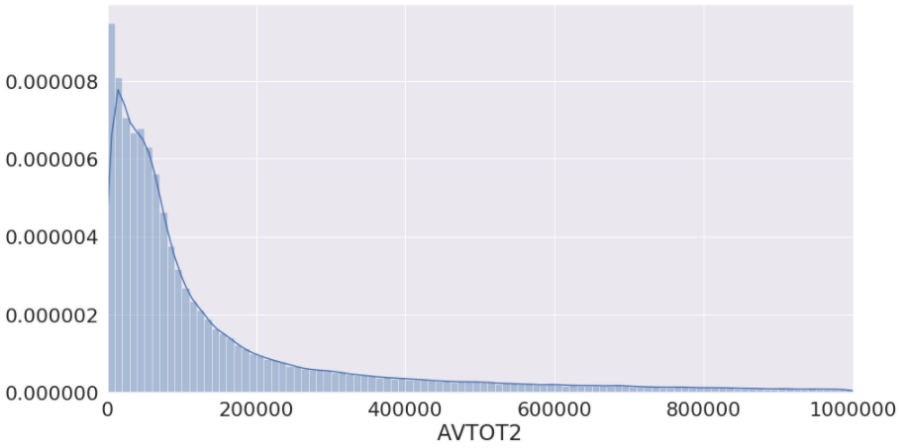
24. **BLDDEPTH:** Building Depth in feet.



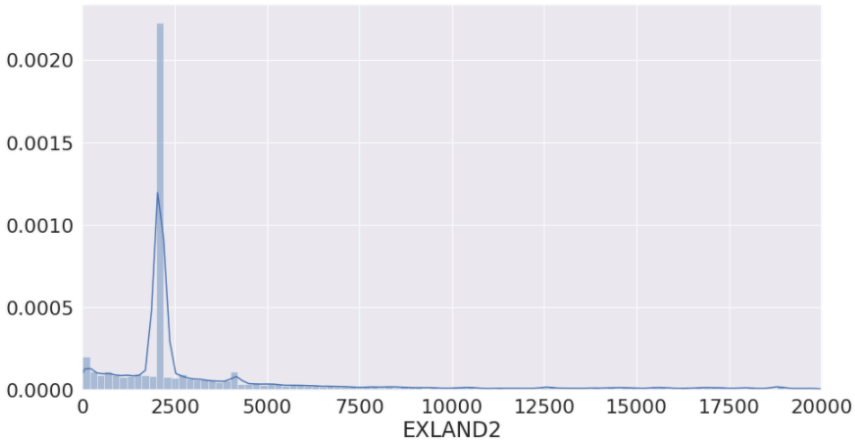
25. **AVLAND2:** Traditional Land Value



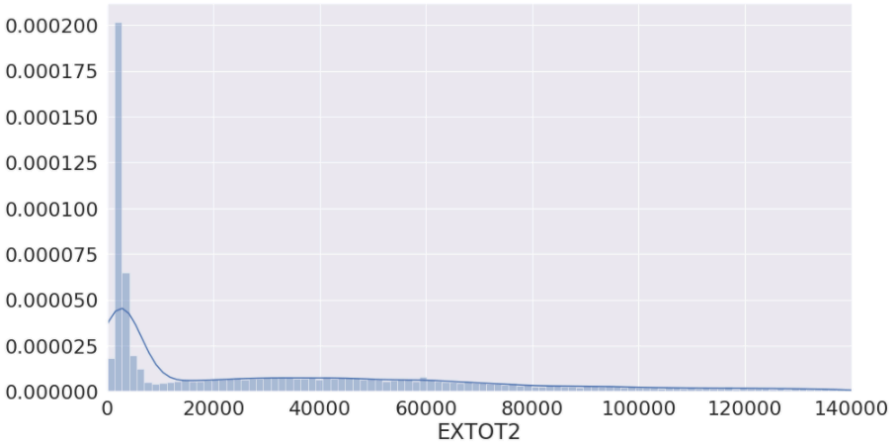
26. **AVTOT2:** Transitional Total Value.



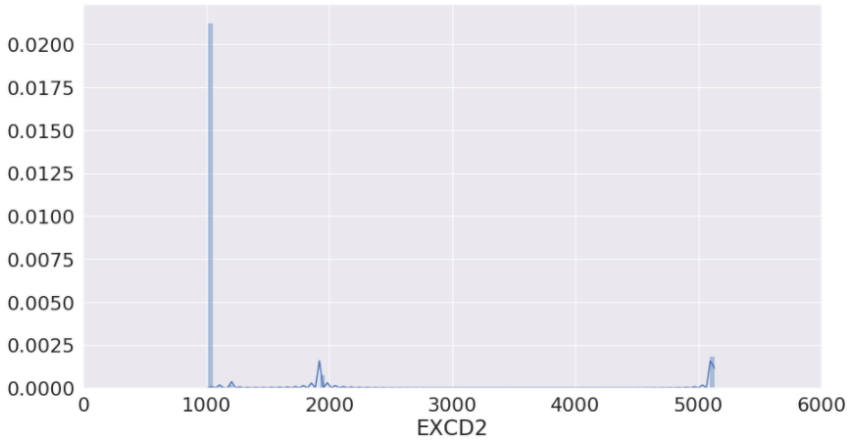
27. **EXLAND2:** Traditional Exemption Land Value



28. **EXTOT2:** Traditional Exemption Land Total



29. **EXCD2:** Exemption Code 2



30. PERIOD: Assessment Period. Only one value: Final

31. YEAR: Assessment Year. Only one value: 2010/11

32. VALTYPE: Only one value: AC-TR.

Appendix B

Correlation Matrix: Heat map