FEBURARY 20, 2019

PROPERTY FRAUD DETECTION

FINAL REPORT

DEQI WAN, HONGFA HUANG, JINZE LI, MINGXUE LI, XINGYUAN CHEN

USC MARSHALL SCHOOL OF BUSINESS

Table of Contents

I. Executive Summary

The purpose of this project is to find out the possible fraud records in the NY Properties Data.

After filling in the missing values in the dataset, we created 45 new variables based on the original fields. To measure variables on different scales, we converted the values to z-scores to aid comparison. PCA analysis is then utilized to remove correlations and reduce dimensionality to eight. We did another Z-scale to give the 8 variables equal importance.

Following the data clearing, two models were built to develop two fraud scores. One calculates the distance of z-scale from origin. Another calculates the distance of records reproduced by the autoencoder from the original records.

Combing the two scores, we got the final fraud score and sorted the records based on the final fraud score. We then analyzed the top 20 records with highest fraud score. Most of the 20 records have good explanation for their unusualness but we succeeded in finding some frauds.

II. Description of Data

1. Description of Data

- Data represent NYC properties assessments for purpose to calculate Property Tax, Grant eligible properties Exemptions and/or Abatements. Data is provided by Department of Finance, owned by NYC OpenData.
- Data is collected and entered into the system by various City employee, like Property Assessors, Property Exemption specialists, ACRIS reporting, Department of Building reporting, etc.
- Data covered property records in November 2010, with 32 fields and 1,070,994 records.

2. Summary Statistics

Numerical Fields

	# records that have a value	% populated	# unique values	# records with value zero	mean	standard deviation	min	max
LTFRONT	1,070,994	100.00%	1,297	169,108	36.6	74.0	0	9,999
LTDEPTH	1,070,994	100.00%	1,370	170,128	88.9	76.4	0	9,999
STORIES	1,014,730	94.75%	111	0	5.0	8.4	1	119
FULLVAL	1,070,994	100.00%	109,324	13,007	874,264.5	11,582,431.0	0	6,150,000,000
AVLAND	1,070,994	100.00%	70,921	13,009	85,067.9	4,057,260.0	0	2,668,500,000
AVTOT	1,070,994	100.00%	112,914	13,007	227,238.2	6,877,529.0	0	4,668,308,947
EXLAND	1,070,994	100.00%	33,419	491,699	36,423.9	3,981,576.0	0	2,668,500,000
EXTOT	1,070,994	100.00%	64,255	432,572	91,187.0	6,508,403.0	0	4,668,308,947
BLDFRONT	1,070,994	100.00%	612	228,815	23.0	35.6	0	7,575
BLDDEPTH	1,070,994	100.00%	621	228,853	39.9	42.7	0	9,393
AVLAND2	282,726	26.40%	58,591	0	246,235.7	6,178,963.0	3	2,371,005,000
AVTOT2	282,732	26.40%	11,130	0	713,911.4	11,652,529.0	3	4,501,180,002
EXLAND2	87,449	8.17%	22,195	0	351,235.7	10,802,213.0	1	2,371,005,000
EXTOT2	130,828	12.22%	48,348	0	656,768.3	16,072,510.0	7	4,501,180,002

Categorical Fields

	# records that have a value	% populated	# unique values	# records with value zero	most common field value
B	1,070,994	100.00%	5	0	4
BLOCK	1,070,994	100.00%	13,984	0	3,944
LOT	1,070,994	100.00%	6,366	0	1
EASEMENT	1,070,994	100.00%	13	0	SPACE
OWNER	1,039,249	97.0%	863,347	30,804	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100.00%	200	0	R4
TAXCLASS	1,070,994	100.00%	11	0	1

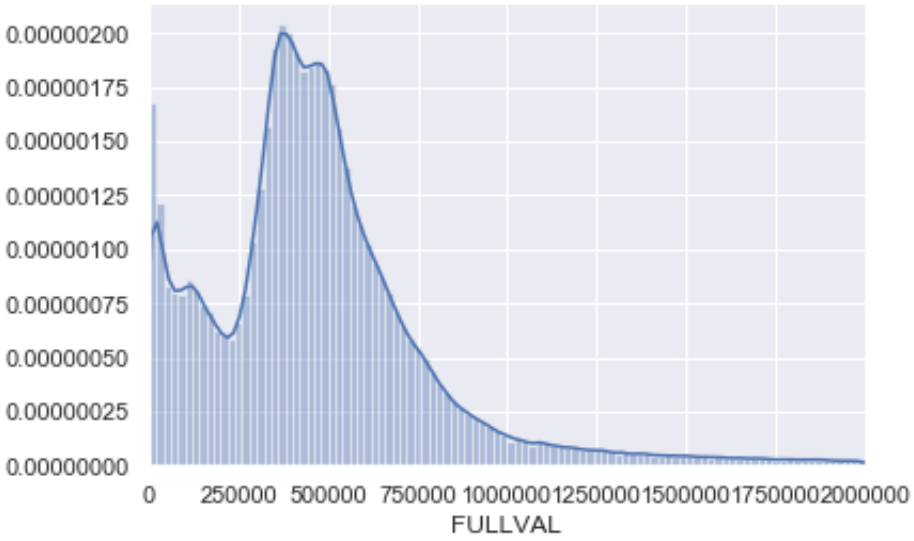
EXT	354,305	33.1%	3	237,094	G
EXCD1	638,488	59.62%	129	0	1,017
STADDR	1,070,318	99.94%	839,280	676	501 SURF AVENUE
ZIP	1,041,113	97.21%	196	0	10,314
EXMPTCL	15,529	1.45%	14	0	X1
EXCD2	92,948	8.68%	60	0	1,017

3. Variable Distributions

Distributions of the 3 important dollar variables (FULLVAL, AVLAND, AVTOT) are showed in this section.

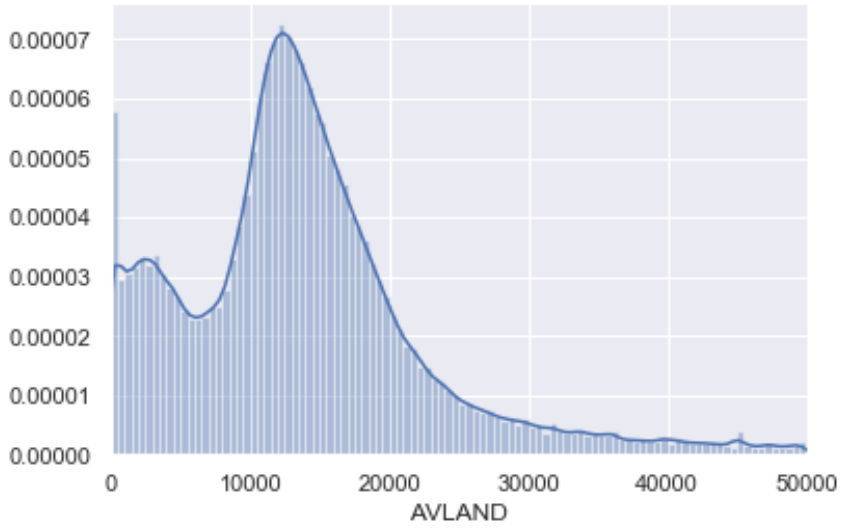
FULLVAL:

FULLVAL is a numeric variable, representing the property's full value. The mean is 874264.5 and the standard deviation is about 11 million. FULLVAL has 109324 unique values and no missing values. A fairly large portion of the records have 0 FULLVAL. Those 0-values will be filled in to detect whether it is a fraud or not. The distribution of this field is shown below:



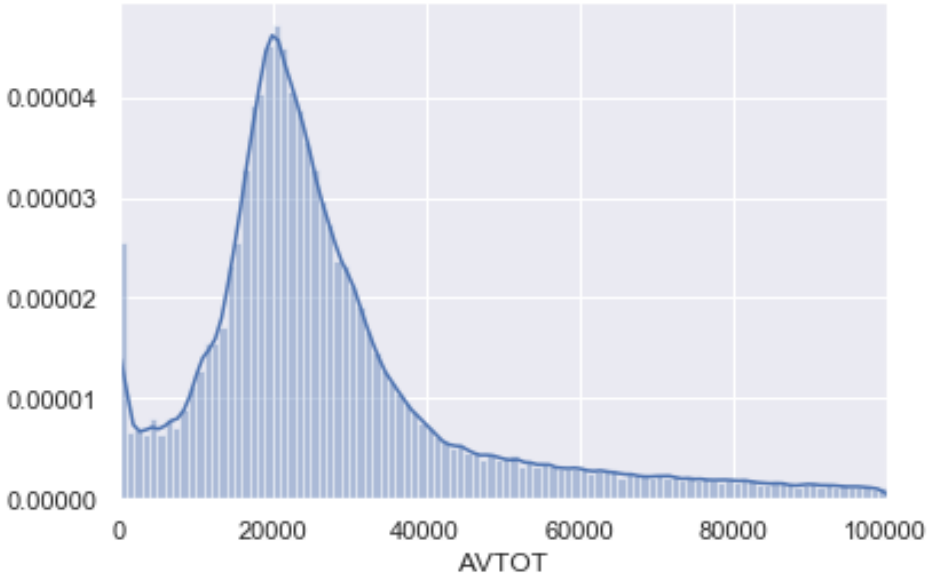
AVLAND:

AVLAND is a numeric variable, representing the property's assessed value. The mean is 85067.9 and the standard deviation is about 4 million. AVLAND has 70921 unique values and no missing value. It has one peak at around 12000. Similarly, a fairly large portion of the data have 0 values. Those 0-values will be filled in later. The distribution of this field is shown below.



AVTOT:

AVTOT is a numeric variable, representing the property's total assessed value. The mean is 227238.2 and the standard deviation is about 6 million. AVTOT has 112914 unique values. No missing values. It has one peak at around 20000. The distribution of this field is shown below:



4. Data Quality Report

Please see the appendix for Data Quality Report.

III. Data Cleaning

We could find many outliers and missing values in the NY Property Dataset. As for the outliers, we used Z scaling algorithms to standardize numerical fields. Here we used the 'scale' function from the 'sklearn.preprocessing' package in Python to process all the numerical fields in the dataset.

Then we created our own variables to compute the unit value of FULLVAL, AVTOT, etc., divided by computed square footage. These will be further introduced in the next part. As for the missing values, firstly we filled all the NA values in ZIP since we would need this field later. Field ZIP contains 29890 NA values and has no Zero values. Since the dataset is ordered by BBLE as well as ZIP codes, and observations with same ZIP code are all next to each other. Even for neighboring different ZIP codes, their BBLEs are similar so they are still in the nearby region. So, for NA Values in field ZIP, we just simply use the previous existing ZIP value to fill in.

In other particular fields which we will analyze later, such as FULLVAL, AVLAND, AVTOT, LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, here is their missing value information:
Field FULLVAL contains 13007 Zero values. No NA values.
Field AVLAND contains 13009 Zero values. No NA values.
Field AVTOT contains 13007 Zero values. No NA values.
Field LTFRONT contains 169108 Zero values. No NA values.
Field LTDEPTH contains 170128 Zero values. No NA values.
Field BLDFRONT contains 228815 Zero values. No NA values.
Field BLDDEPTH contains 228853 Zero values. No NA values.

For these fields, although there are no NA values, zero values make no sense and it's more likely to represent missing values. In order to robustly estimate those observations' possible status without creating potential manmade outliers, or Zero Values in these fields, we can use their corresponding ZIP and TAXCLASS's median value. If a zero value's corresponding ZIP and TAXCLASS have less than 10 records or if that zero value does not have a ZIP record, use its corresponding TAXCLASS's median value.

For field STORIES, it contains 56264 NA values and no Zero values. For NA Values in these fields, we can use their corresponding ZIP and TAXCLASS's median value. If a zero value's corresponding ZIP and TAXCLASS have less than 10 records or if that zero value does not have a ZIP record, use its corresponding TAXCLASS's median value.

IV. Variable Creation

From the dataset, we have 32 fields which are not all useful for our fraud analysis. We select 11 variables (LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, STORIES, ZIP, TAXCLASS, B, FULLVAL, AVLAND, AVTOT) to make the model and do our fraud analysis.

LTFRONT:	property's lot frontage in feet
LTDEPTH:	property's lot depth in feet
BLDFRONT:	building frontage in feet
BLDDEPTH:	building depth in feet
STORIES:	number of floors of the building
ZIP:	property's zip code
TAXCLASS:	current property tax class code
B:	boro Codes
FULLVAL:	full value of the property
AVLAND:	assessed value of Land of the property
AVTOT:	assessed value Total of the property

Then we make lots of variables based on the 11 original variables. Firstly, we create 3 sizes （lotarea, bldarea, bldvol）based on LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, STORIES:

- lotarea = LTFRONT * LTDEPTH
- bldarea = BLDFRONT * BLDDEPTH
- bldvol = bldarea * STORIES

lotarea indicate the total lot size, bldarea is the building size for one floor and bldvol is the total size for the building.

Secondly, based on these three sizes and FULLVAL, AVLAND, AVTOT, we create 9 new variables: r1 – r9:

$$r_1 = \frac{V_1}{S_1} \qquad r_4 = \frac{V_2}{S_1} \qquad r_7 = \frac{V_3}{S_1}$$

$$r_2 = \frac{V_1}{S_2} \qquad r_5 = \frac{V_2}{S_2} \qquad r_8 = \frac{V_3}{S_2}$$

$$r_3 = \frac{V_1}{S_3} \qquad r_6 = \frac{V_2}{S_3} \qquad r_9 = \frac{V_3}{S_3}$$

V_1 = FULLVAL, V_2 = AVLAND, V_3 = AVTOT, S_1 = lotarea, S_2 = bldarea, S_3 = bldvol

We calculate these ratios to find the normalized value of the property. They could help us not influenced by the building/ lot size to analyze the property value.

Thirdly, we create ZIP5 and ZIP3 based on ZIP. ZIP5 is 5 digits of ZIP and ZIP3 is the last 3 digits of ZIP. We prepare these two values for the next steps.

Finally, we create the grouped averages of 9 variables (r1 - r9) we created in the second step. We use ZIP5 (g1), ZIP3 (g2), TAXCLASS (g3), B (g4) to group the 9 variables, and also use all the 9 variables (g5) without group. For ZIP5 (g1), we group the 9 variables by g1 and calculate 9 means of each variable, called <r1>g1, <r2>g1…<r9>g1. Then we define r1g1 – r9g1 using methods below:

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

Then we repeat the same process four times. We eventually have 5 groups of the 9 variables, which are 45 variables (r1g1 – r9g1, r1g2 - r9g2, r1g3 – r9g3, r1g4 – r9g4, r1g5 – r9g5) in total. These variables will help us to get rid of the effect of the location and taxclass on the properties.

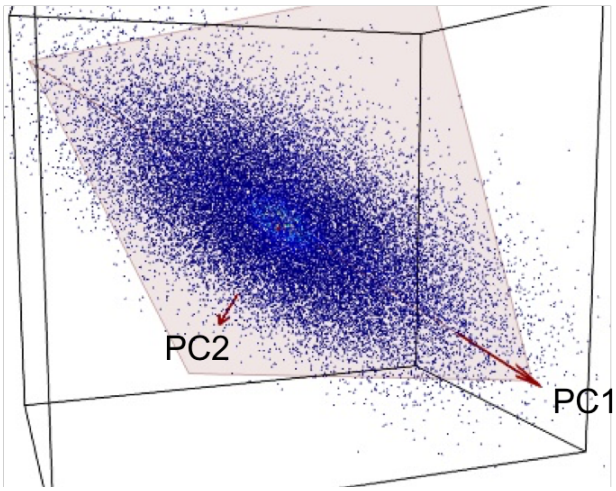
After creating these 45 variables, we are ready for the PCA.

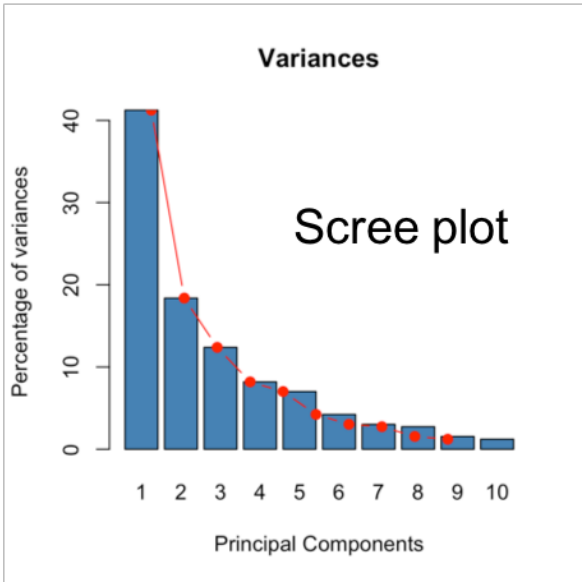
V. Dimensionality Reduction

Since the dataset has too many fields and many of these fields are highly correlated with each other, we decided to reduce the dimensionality before we analyze the data. To reduce the dimensionality and find major factors in the dataset, we used **Principal Components Analysis** (PCA).

PCA can be thought of as fitting an n-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipsoid is small, then the variance along that axis is also small, and by omitting that axis and its corresponding principal component from our representation of the dataset, we lose only a commensurately small amount of information.

To find the axes of the ellipsoid, we must first subtract the mean of each variable from the dataset to center the data around the origin. Then, we compute the covariance matrix of the data, and calculate the eigenvalues and corresponding eigenvectors of this covariance matrix. Then we must normalize each of the orthogonal eigenvectors to become unit vectors. Once this is done, each of the mutually orthogonal, unit eigenvectors can be interpreted as an axis of the ellipsoid fitted to the data. This choice of basis will transform our covariance matrix into a diagonalized form with the diagonal elements representing the variance of each axis. The proportion of the variance that each eigenvector represents can be calculated by dividing the eigenvalue corresponding to that eigenvector by the sum of all eigenvalues.





We chose 8 as the number of our PCA output dimension, so that we had 8 major dimensions left in our final analysis. Here we used the 'PCA' function from the 'sklearn.decomposition' package in Python. After the PCA, we also z scaled the output dataset.

VI. Algorithms

We use two methods to calculate the fraud scores

Method 1: Heuristic Function of the z-scores

1.Second Z Scaling
After the Principal Component Analysis, we want to make our principal component equally important. We do the second z scale.

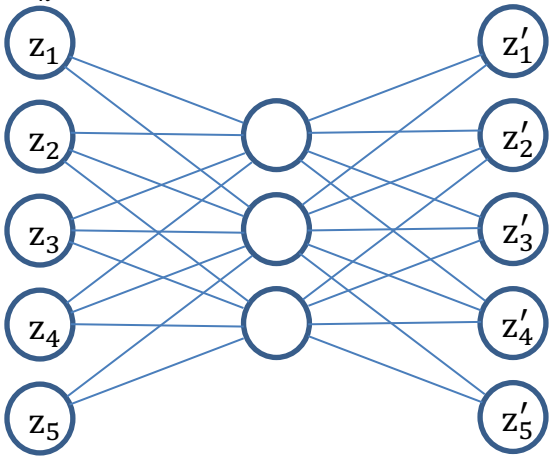
2.Summing the Z-scores
After reducing the dimensions, we do have the principal components uncorrelated. We can sum the Z-scores together by using the absolute value without canceling each other out by using the formulae where i is the record and k is the number order of principal component.

$$S_i = \sum_k |z_k^i|$$

3.Logics behind
If the record has high anomality, the absolute value of z_k^i will be high. The overall score: the sum of the absolute value of z_k^i is a good measurement of evaluating how abnormal the record is. High score might have high propensity leading to fraud.

Method 2: Autoencoder

1.Applying machine learning algorithm to encode new $z_k^{i'}$
Given existing z_k^i, we can also train a model to learn the insights of the data. Then, we can apply the model to generate the new $z_k^{i'}$.



2.Summing the distance

After we generate the new z-score, we can sum the difference between the original ones and the new ones to obtain the new fraud score where i is the record and k is the number order of principal component and $z_k^{i'}$ is the new score.

$$S'_i = \sum_k \left| z_k^i - z_k^{i'} \right|$$

3.Logics behind
Because our model is trained based on former z-scores, the new ones should be close to the original ones. If there is abnormality (possible fraud), the difference of the two will be large.

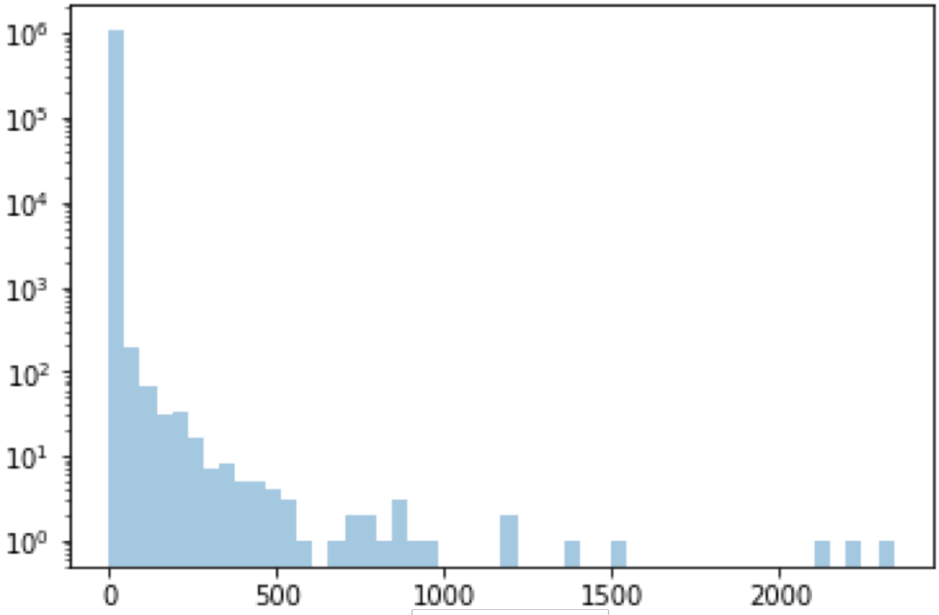
Combining two scores together

Since the score might be quite close, we sum the score together and do the **Extreme Quantile Binning**: Sum two scores and rank them in the descending order. This is the most precise way to do quantile binning.

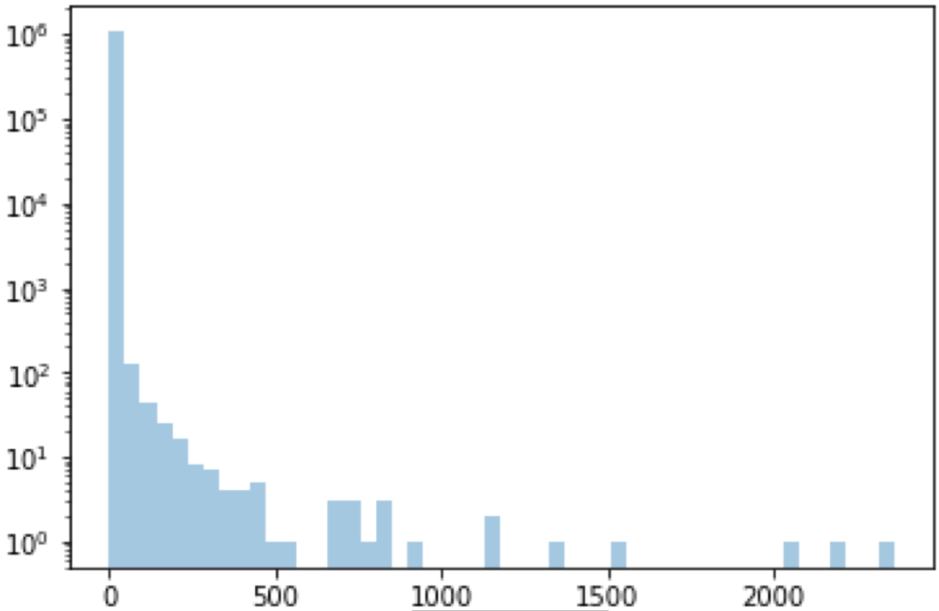
VII. Results

Distributions for 3 Scores

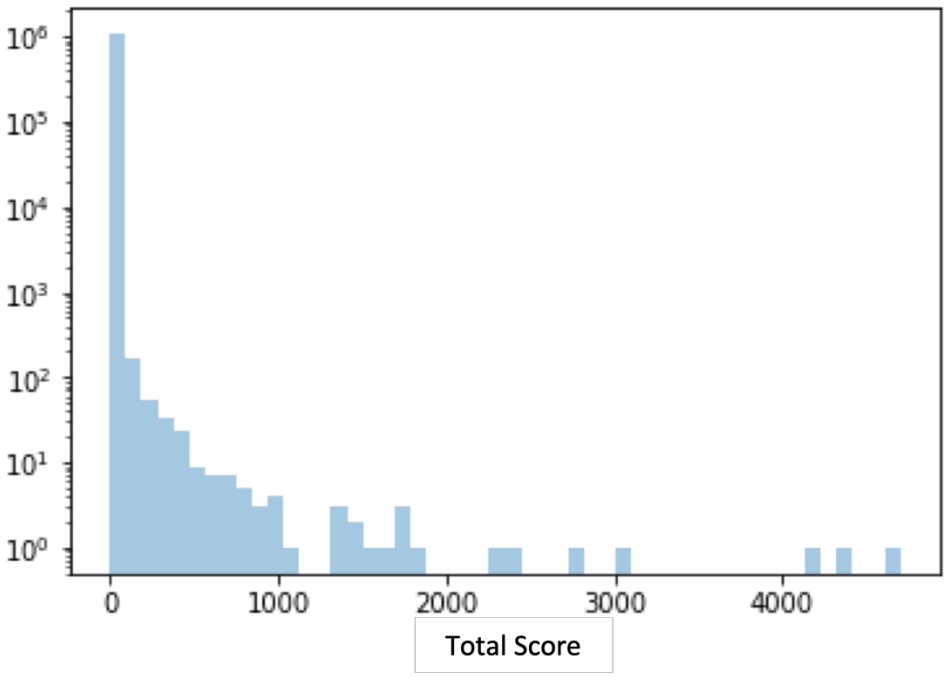
The following graphs are the distributions of our Score 1 (derived from Z-scaling and PCA), Score 2 (derived from Autoencoder) and Total Score (the sum of Score 1 and 2), from which we can further prove the rationality of our two models, since all three distributions are right-skewed:



Score 1



Score 2



Top 20 Fraud Records

We ranked all of our records by executing quantile binning techniques. Below are the top 20 records we found that are highly likely to be fraudulent and we will look more into these 20 records in this section. Basically, we will investigate deeper into the anomalies in their property dollar values, building & lot area, number of missing values and their addresses.

RECORD	SCORE 1	SCORE 2	TOTAL SCORE
632816	1070994	1070994	2141988
565392	1070993	1070993	2141986
1067360	1070992	1070992	2141984
917942	1070991	1070991	2141982
585118	1070990	1070990	2141980
85886	1070989	1070988	2141977
585439	1070988	1070989	2141977
565398	1070987	1070987	2141974
556609	1070985	1070985	2141970
585120	1070984	1070986	2141970
920628	1070986	1070984	2141970
690833	1070983	1070983	2141966
750816	1070982	1070982	2141964
776306	1070981	1070980	2141961

935158	1070980	1070981	2141961
1067001	1070979	1070978	2141957
691879	1070978	1070979	2141957
67129	1070977	1070977	2141954
770594	1070976	1070976	2141952
794105	1070975	1070974	2141949

The following are the investigations we did on every record to see whether it is a fraud or an anomaly.

RECORD 632816:

RECORD	632816	AVLAND	1.32E+06
BBLE	4018420001	AVTOT	1.32E+06
B	4	EXLAND	0
BLOCK	1842	EXTOT	0
LOT	1	EXCD1	NaN
EASEMENT	NaN	STADDR	86-55 BROADWAY
OWNER	864163 REALTY, LLC	ZIP	11373
BLDGCL	D9	EXMPTCL	NaN
TAXCLASS	2	BLDFRONT	1
LTFRONT	157	BLDDEPTH	1
LTDEPTH	95	AVLAND2	1.20E+06
EXT	NaN	AVTOT2	1.20E+06
STORIES	1	EXLAND2	NaN
FULLVAL	2.93E+06	EXTOT2	NaN

This property has 1 for both BLDFRONT and BLDDEPTH while their LTFRONT and LTDEPTH are 157, 95. And when we actually looked it up on the map, we found it is actually a 5-stories building. This inappropriate information provided with regard to building area tremendously increases our variable r2, r5 and r8, which are formulated as FULLVAL or AVLAND or AVTOT divided by (BLDFRONT * BLDDEPTH) respectively.

RECORD 565392:

RECORD	565392	AVLAND	1.95E+09
BBLE	3085900700	AVTOT	1.95E+09
B	3	EXLAND	1.95E+09
BLOCK	8590	EXTOT	1.95E+09
LOT	700	EXCD1	2231

EASEMENT	NaN	STADDR	FLATBUSH AVENUE
OWNER	U S GOVERNMENT OWNRD	ZIP	NaN
BLDGCL	V9	EXMPTCL	X1
TAXCLASS	4	BLDFRONT	0
LTFRONT	117	BLDDEPTH	0
LTDEPTH	108	AVLAND2	8.48E+08
EXT	NaN	AVTOT2	8.48E+08
STORIES	NaN	EXLAND2	8.48E+08
FULLVAL	4.33E+09	EXTOT2	8.48E+08

This property is state owned with a lot of missing values. Despite all that, this property has very high FULLVAL and AVTOT. Therefore, we are guessing that this property might be a vacant lot and that the fraud may come from the imbalance of the filled-in median values for building area and the high assessed total values.

RECORD 1067360:

RECORD	1067360	AVLAND	28800
BBLE	5078530085	AVTOT	50160
B	5	EXLAND	0
BLOCK	7853	EXTOT	0
LOT	85	EXCD1	NaN
EASEMENT	NaN	STADDR	20 EMILY COURT
OWNER	NaN	ZIP	10307
BLDGCL	B2	EXMPTCL	NaN
TAXCLASS	1	BLDFRONT	36
LTFRONT	1	BLDDEPTH	45
LTDEPTH	1	AVLAND2	NaN
EXT	NaN	AVTOT2	NaN
STORIES	2	EXLAND2	NaN
FULLVAL	836000	EXTOT2	NaN

This property's AVTOT is much smaller than its FULLVAL. Also, its LTFRONT and LTDEPTH are both 1. In this case, we think because of the taxable characteristic in AVTOT, the property owner is understating the assessed total value to evade tax. The owner may also be manipulating the lot area to increase the unit value of his or her house in order to take out loans more easily.

RECORD 917942:

RECORD	917942	AVLAND	1.79E+09

BBLE	4142600001	AVTOT	4.67E+09
B	4	EXLAND	1.79E+09
BLOCK	14260	EXTOT	4.67E+09
LOT	1	EXCD1	2198
EASEMENT	NaN	STADDR	154-68 BROOKVILLE BOULEVARD
OWNER	LOGAN PROPERTY, INC.	ZIP	11422
BLDGCL	T1	EXMPTCL	X4
TAXCLASS	4	BLDFRONT	0
LTFRONT	4910	BLDDEPTH	0
LTDEPTH	0	AVLAND2	1.64E+09
EXT	NaN	AVTOT2	4.50E+09
STORIES	3	EXLAND2	1.64E+09
FULLVAL	3.74E+08	EXTOT2	4.50E+09

According to the dataset, the STORIES of this property should be 3. But when we looked it up on the map, it's should actually be 6. Also, this property is used as a hotel. Due to its location in the residential area, when we group by its zip code trying to fill in missing values of its BLDFRONT, BLDDEPTH and LTDEPTH, the filling values tend to be smaller than what's ought to be.

RECORD 585118:

RECORD	585118	AVLAND	1.55E+06
BBLE	4004200001	AVTOT	1.55E+06
B	4	EXLAND	0
BLOCK	420	EXTOT	0
LOT	1	EXCD1	NaN
EASEMENT	NaN	STADDR	28-10 QUEENS PLAZA SOUTH
OWNER	NEW YORK CITY ECONOMI	ZIP	11101
BLDGCL	O3	EXMPTCL	X1
TAXCLASS	4	BLDFRONT	1
LTFRONT	298	BLDDEPTH	1
LTDEPTH	402	AVLAND2	1.59E+06
EXT	NaN	AVTOT2	1.59E+06
STORIES	20	EXLAND2	NaN
FULLVAL	3.44E+06	EXTOT2	NaN

This property is a big vacant lot (see the picture below). its BLDFRONT and BLDDEPTH are both 1, which is normal in this case. But this would cause the variable r2, r5 and r8 to increase greatly just like record 632816.

RECORD 85886:

RECORD	85886	AVLAND	3.15E+07
BBLE	1012540010	AVTOT	3.16E+07
B	1	EXLAND	3.15E+07
BLOCK	1254	EXTOT	3.16E+07
LOT	10	EXCD1	2231
EASEMENT	NaN	STADDR	JOE DIMAGGIO HIGHWAY
OWNER	PARKS AND RECREATION	ZIP	NaN
BLDGCL	Q1	EXMPTCL	X1
TAXCLASS	4	BLDFRONT	8
LTFRONT	4000	BLDDEPTH	8
LTDEPTH	150	AVLAND2	2.81E+07
EXT	NaN	AVTOT2	2.83E+07
STORIES	1	EXLAND2	2.81E+07
FULLVAL	7.02E+07	EXTOT2	2.83E+07

This property is a used as a resort park. its BLDFRONT and BLDDEPTH are both 8, much smaller than its lot area. But this would cause the variable r2, r3, r5, r6 and r8 (FULLVAL or AVLAND or AVTOT divided by (BLDFRONT * BLDDEPTH) or (BLDFRONT * BLDDEPTH * STORIES) respectively) to increase greatly and got caught by our algorithms.

RECORD 585439:

RECORD	585439	AVLAND	252000
BBLE	4004590005	AVTOT	1.67E+06
B	4	EXLAND	0
BLOCK	459	EXTOT	1.42E+06
LOT	5	EXCD1	1986
EASEMENT	NaN	STADDR	11-01 43 AVENUE
OWNER	11-01 43RD AVENUE REA	ZIP	11101
BLDGCL	H9	EXMPTCL	NaN
TAXCLASS	4	BLDFRONT	1
LTFRONT	94	BLDDEPTH	1
LTDEPTH	165	AVLAND2	NaN
EXT	NaN	AVTOT2	NaN
STORIES	10	EXLAND2	NaN
FULLVAL	3.71E+06	EXTOT2	NaN

This property has very small BLDDEPTH and BLDFRONT. This land is used as a hotel. The owner may intend to take on loans by increasing the unit value of this property. The assessed value for this property is way too high, since it's located in Queens and its lot area is not really that big.

RECORD 565398:

RECORD	565398	AVLAND	1.04E+09
BBLE	3085910100	AVTOT	1.04E+09
B	3	EXLAND	1.04E+09
BLOCK	8591	EXTOT	1.04E+09
LOT	100	EXCD1	2191
EASEMENT	NaN	STADDR	FLATBUSH AVENUE
OWNER	DEPT OF GENERAL SERVI	ZIP	NaN
BLDGCL	V9	EXMPTCL	X1
TAXCLASS	4	BLDFRONT	0
LTFRONT	466	BLDDEPTH	0
LTDEPTH	1009	AVLAND2	4.35E+08
EXT	NaN	AVTOT2	4.35E+08
STORIES	NaN	EXLAND2	4.35E+08
FULLVAL	2.31E+09	EXTOT2	4.35E+08

This should be a vacant lot because of the equivalence between its AVTOT and AVLAND. Also, this land's BLDFRONT and BLDDEPTH are both 0, which reinforced our thoughts. The anomaly comes from the huge assessed dollar values and the small building area.

RECORD 556609:

RECORD	556609	AVLAND	6.08E+07
BBLE	3083120001	AVTOT	6.12E+07
B	3	EXLAND	6.08E+07
BLOCK	8312	EXTOT	6.12E+07
LOT	1	EXCD1	2231
EASEMENT	NaN	STADDR	9006 SEAVIEW AVENUE
OWNER	PARKS AND RECREATION	ZIP	11236
BLDGCL	Q1	EXMPTCL	X1
TAXCLASS	4	BLDFRONT	88
LTFRONT	35	BLDDEPTH	62
LTDEPTH	50	AVLAND2	5.86E+07
EXT	NaN	AVTOT2	5.90E+07
STORIES	1	EXLAND2	5.86E+07
FULLVAL	1.36E+08	EXTOT2	5.90E+07

This property's lot area is smaller than its building area which is very unusual. And the FULLVAL, AVLAND, AVTOT are all too high for such small land located in Brooklyn. So, we do consider this case as a fraud and the owner might want to do tax evasion and lend loans at the same time.

RECORD 585120:

RECORD	585120	AVLAND	968220
BBLE	4004200101	AVTOT	968220
B	4	EXLAND	0
BLOCK	420	EXTOT	0
LOT	101	EXCD1	NaN
EASEMENT	NaN	STADDR	28 STREET
OWNER	NaN	ZIP	NaN
BLDGCL	O3	EXMPTCL	NaN
TAXCLASS	4	BLDFRONT	1
LTFRONT	139	BLDDEPTH	1
LTDEPTH	342	AVLAND2	975456
EXT	NaN	AVTOT2	975456
STORIES	20	EXLAND2	NaN
FULLVAL	2.15E+06	EXTOT2	NaN

This property's BLDFRONT and BLDDEPTH are both 1 and its FULLVAL is too high for such small building area with 20 stories. This case should be a deliberate data manipulation.

RECORD 920628:

RECORD	920628	AVLAND	9763
BBLE	4155770029	AVTOT	75763
B	4	EXLAND	0
BLOCK	15577	EXTOT	0
LOT	29	EXCD1	NaN
EASEMENT	NaN	STADDR	7-06 ELVIRA AVENUE
OWNER	PLUCHENIK, YAAKOV	ZIP	11691
BLDGCL	A1	EXMPTCL	NaN
TAXCLASS	1	BLDFRONT	1
LTFRONT	91	BLDDEPTH	1
LTDEPTH	100	AVLAND2	NaN
EXT	NaN	AVTOT2	NaN
STORIES	2	EXLAND2	NaN
FULLVAL	1.90E+06	EXTOT2	NaN

The building area for this property is only 1. The FULLVAL is so much bigger than its AVTOT. Therefore, this could be a tax evasion fraud.

RECORD 690833:

RECORD	690833	AVLAND	1.04E+08
BBLE	4038660070	AVTOT	1.09E+08
B	4	EXLAND	1.04E+08
BLOCK	3866	EXTOT	1.09E+08
LOT	70	EXCD1	2231
EASEMENT	NaN	STADDR	83-98 FOREST PARKWAY
OWNER	PARKS AND RECREATION	ZIP	11385
BLDGCL	Q1	EXMPTCL	X1
TAXCLASS	4	BLDFRONT	20
LTFRONT	610	BLDDEPTH	20
LTDEPTH	534	AVLAND2	9.70E+07
EXT	NaN	AVTOT2	1.02E+08
STORIES	3	EXLAND2	9.70E+07
FULLVAL	2.42E+08	EXTOT2	1.02E+08

This land is used as a golf course. It is reasonable to have small building area. The reason why it has high fraud score is that compared to its FULLVAL, AVLAND, and AVTOT, the building area is too low. And that leads to a super high unit value.

RECORD 750816:

RECORD	750816	AVLAND	0
BBLE	4066610005E	AVTOT	0
B	4	EXLAND	0
BLOCK	6661	EXTOT	0
LOT	5	EXCD1	NaN
EASEMENT	E	STADDR	VLEIGH PLACE
OWNER	M FLAUM	ZIP	NaN
BLDGCL	V0	EXMPTCL	NaN
TAXCLASS	1B	BLDFRONT	0
LTFRONT	1	BLDDEPTH	0
LTDEPTH	1	AVLAND2	NaN
EXT	NaN	AVTOT2	NaN
STORIES	NaN	EXLAND2	NaN
FULLVAL	0	EXTOT2	NaN

This property has 0 as its building area and 1 as its lot area. Also, there is no information about the FULLVAL, AVLAND and AVTOT. This would lead to us filling in a lot of median values based on its zip code and tax class, which tend to be small. And furthermore, this would lead to a very high fraud score.

RECORD 776306:

RECORD	776306	AVLAND	0
BBLE	4080100001	AVTOT	0
B	4	EXLAND	0
BLOCK	8010	EXTOT	0
LOT	1	EXCD1	NaN
EASEMENT	NaN	STADDR	SHORE ROAD
OWNER	TONY CHEN	ZIP	NaN
BLDGCL	Q9	EXMPTCL	NaN
TAXCLASS	4	BLDFRONT	0
LTFRONT	6	BLDDEPTH	0
LTDEPTH	1	AVLAND2	NaN
EXT	NaN	AVTOT2	NaN
STORIES	1	EXLAND2	NaN
FULLVAL	0	EXTOT2	NaN

This property has 0 as its building area and 6 as its lot area. Also, there is no information about the FULLVAL, AVLAND and AVTOT. This would lead to us filling in a lot of median values based on its zip code and tax class, which tend to be small. And furthermore, this would lead to a very high fraud score.

RECORD 935158:

RECORD	935158	AVLAND	236250
BBLE	5000130060	AVTOT	468000
B	5	EXLAND	221378
BLOCK	13	EXTOT	453128
LOT	60	EXCD1	5113
EASEMENT	NaN	STADDR	224 RICHMOND TERRACE
OWNER	RICH-NICH REALTY,LLC	ZIP	10301
BLDGCL	D3	EXMPTCL	NaN
TAXCLASS	2	BLDFRONT	1
LTFRONT	136	BLDDEPTH	1
LTDEPTH	132	AVLAND2	209880
EXT	NaN	AVTOT2	748350
STORIES	8	EXLAND2	195008
FULLVAL	1.04E+06	EXTOT2	733478

This property's BLDFRONT and BLDDEPTH are both 1, which increases our variable r2, r5 and r8, which are formulated as FULLVAL or AVLAND or AVTOT divided by (BLDFRONT * BLDDEPTH) respectively and finally lead to a high fraud score.

RECORD 1067001:

RECORD	1067001	AVLAND	65401
BBLE	5078120132	AVTOT	124910
B	5	EXLAND	0
BLOCK	7812	EXTOT	0
LOT	132	EXCD1	NaN
EASEMENT	NaN	STADDR	238 BEDELL AVENUE
OWNER	DRANOVSKY, VLADIMIR	ZIP	10307
BLDGCL	A3	EXMPTCL	NaN
TAXCLASS	1	BLDFRONT	1
LTFRONT	96	BLDDEPTH	1
LTDEPTH	279	AVLAND2	NaN
EXT	NaN	AVTOT2	NaN
STORIES	3	EXLAND2	NaN
FULLVAL	2.12E+06	EXTOT2	NaN

This property's BLDFRONT and BLDDEPTH are both 1, which increases our variable r2, r5 and r8, which are formulated as FULLVAL or AVLAND or AVTOT divided by (BLDFRONT * BLDDEPTH) respectively and finally lead to a high fraud score.

RECORD 691879:

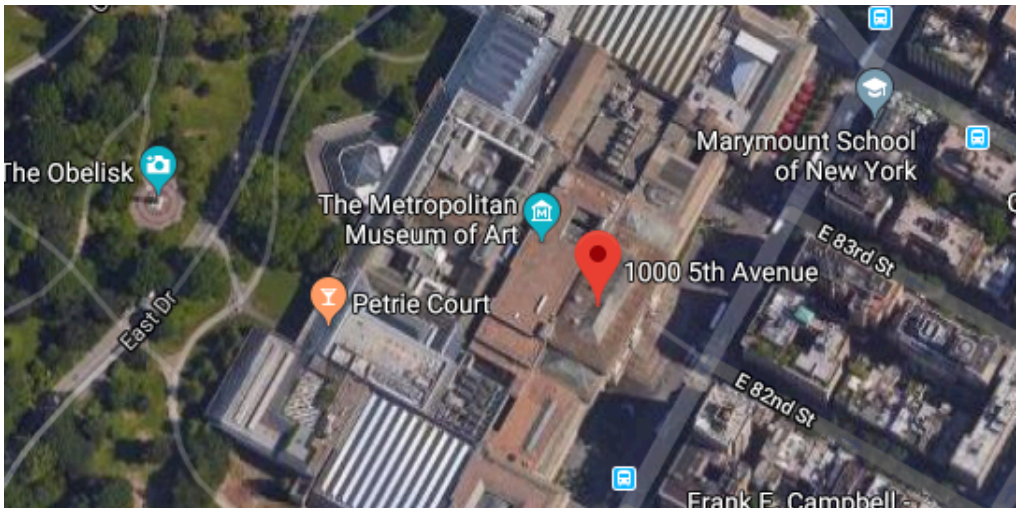
RECORD	691879	AVLAND	4.46E+07
BBLE	4039071425	AVTOT	4.58E+07
B	4	EXLAND	4.46E+07
BLOCK	3907	EXTOT	4.58E+07
LOT	1425	EXCD1	2231
EASEMENT	NaN	STADDR	UNION TURNPIKE
OWNER	PARKS AND RECREATION	ZIP	11421
BLDGCL	Q2	EXMPTCL	X1
TAXCLASS	4	BLDFRONT	27
LTFRONT	0	BLDDEPTH	47
LTDEPTH	47	AVLAND2	4.28E+07
EXT	NaN	AVTOT2	4.41E+07
STORIES	1	EXLAND2	4.28E+07
FULLVAL	1.02E+08	EXTOT2	4.41E+07

This property lacks the information about LTFRONT, making the building area bigger than the lot area. FULLVAL and AVTOT for this property is too high based on its location in Queens and its area.

RECORD 67129:

RECORD	67129	AVLAND	2.67E+09
BBLE	1011110001	AVTOT	2.77E+09
B	1	EXLAND	2.67E+09
BLOCK	1111	EXTOT	2.77E+09
LOT	1	EXCD1	2231
EASEMENT	NaN	STADDR	1000 5 AVENUE
OWNER	CULTURAL AFFAIRS	ZIP	10028
BLDGCL	Q1	EXMPTCL	X1
TAXCLASS	4	BLDFRONT	0
LTFRONT	840	BLDDEPTH	0
LTDEPTH	0	AVLAND2	2.37E+09
EXT	E	AVTOT2	2.47E+09
STORIES	NaN	EXLAND2	2.37E+09
FULLVAL	6.15E+09	EXTOT2	2.47E+09

This property has 0 as its BLDFRONT, BLDDEPTH and LTDEPTH. Also, it has pretty high FULLVAL and AVTOT. But in this case, we do not consider it as a fraud. First of all, it is located in Manhattan and it has 840 as its LTFRONT which is pretty huge. And then when we looked it up on the map, we found out that it is actually The Metropolitan Museum of Arts. Thirdly, the huge dollar values of this property indicate its status. Therefore, this property is just an anomaly rather than a fraud (see the picture below as a proof).



RECORD 770594:

RECORD	770594	AVLAND	1001
BBLE	4076211145	AVTOT	8934
B	4	EXLAND	0
BLOCK	7621	EXTOT	0
LOT	1145	EXCD1	NaN
EASEMENT	NaN	STADDR	220-71 67 AVENUE
OWNER	OH, LAURA E	ZIP	11364
BLDGCL	R3	EXMPTCL	NaN
TAXCLASS	1A	BLDFRONT	0
LTFRONT	1	BLDDEPTH	0
LTDEPTH	1	AVLAND2	NaN
EXT	NaN	AVTOT2	NaN
STORIES	1	EXLAND2	NaN
FULLVAL	251989	EXTOT2	NaN

This property has 0 as its BLDFRONT and BLDDEPTH, 1 as its LTFRONT and LTDEPTH. Also, its FULLVAL is more than 20 times bigger than the AVTOT.

RECORD 794105:

RECORD	794105	AVLAND	15408
BBLE	4089460045	AVTOT	79248
B	4	EXLAND	0
BLOCK	8946	EXTOT	0
LOT	45	EXCD1	NaN
EASEMENT	NaN	STADDR	91-25 75 STREET
OWNER	HAVEN BUILDERS, INC.	ZIP	11421
BLDGCL	B1	EXMPTCL	NaN
TAXCLASS	1	BLDFRONT	1
LTFRONT	37	BLDDEPTH	1
LTDEPTH	100	AVLAND2	NaN
EXT	NaN	AVTOT2	NaN
STORIES	3	EXLAND2	NaN
FULLVAL	1.36E+06	EXTOT2	NaN

This property has 1 as its BLDFRONT and BLDDEPTH. Also, its FULLVAL is much bigger than the AVTOT.

VIII. Conclusions

The purpose of this project is to find out all the possible fraud records in the NY Properties Data. Our method to approach this problem is to create 2 fraud scores and use quantile binning to rank all the scores to find the top high scores. This is how we got the 2 fraud scores: first of all, we explored the dataset and formulated a Data Quality Report, focusing on the main features. Then we set out to filling in the missing values and creating the variables we needed for this project. Secondly, we utilized Z-scaling and PCA to calculate the first fraud score. Then Autoencoder helped us to get the second fraud score. Finally, we applied quantile binning to rank all the scores and created this final dataset that reflects the scores and ranking of each data record.

In the end, we investigated more into the top records to define whether it is a fraud or an anomaly. The results along with the explanations are in the last section. We can see that some of the top fraudulent records are anomalies due to missing values or typos. Others we can clearly see them as frauds.

If we have more time for this project, we will figure out a better way to fill in the missing values. Because we found out that few of our top fraudulent records are fraudulent due to the medians we filled in based on their zip code and tax class. Also, we would build more variables despite of the unit values in each group that we already built.

IX. Appendix

Data Quality Report of NY Property Data

Description of Data

Data represent NYC properties assessments for purpose to calculate Property Tax, Grant eligible properties Exemptions and/or Abatements. Data is provided by Department of Finance, owned by NYC OpenData. Data is collected and entered into the system by various City employee, like Property Assessors, Property Exemption specialists, ACRIS reporting, Department of Building reporting, etc. Data covered property records in year 2010/11, with 32 fields and 1,070,994 observations.

Summary Information of Different Fields

Numerical Fields:

	# records that have a value	% populated	# unique values	# records with value zero	mean	standard deviation	min	max
LTFRONT	1070994	100.00%	1297	169108	36.6	74.0	0	9999
LTDEPTH	1070994	100.00%	1370	170128	88.9	76.4	0	9999
STORIES	1014730	94.75%	111	0	5	8.4	1	119
FULLVAL	1070994	100.00%	109324	13007	874264.5	11582431	0	6.15e+09
AVLAND	1070994	100.00%	70921	13009	85067.9	4057260	0	2668500000
AVTOT	1070994	100.00%	112914	13007	227238.2	6877529	0	4668308947
EXLAND	1070994	100.00%	33419	491699	36423.9	3981576	0	2668500000
EXTOT	1070994	100.00%	64255	432572	91187.0	6508403	0	4668308947
BLDFRONT	1070994	100.00%	612	228815	23.0	35.6	0	7575
BLDDEPTH	1070994	100.00%	621	228853	39.9	42.7	0	9393
AVLAND2	282726	26.40%	58591	0	246235.7	6178963	3	2371005000
AVTOT2	282732	26.40%	11130	0	713911.4	11652529	3	4501180002
EXLAND2	87449	8.17%	22195	0	351235.7	10802213	1	2371005000
EXTOT2	130828	12.22%	48348	0	656768.3	16072510	7	4501180002

Categorical Fields:

	# records that have a value	% populated	# unique values	# records with value zero	most common field value
B	1070994	100.00%	5	0	4
BLOCK	1070994	100.00%	13984	0	3944
LOT	1070994	100.00%	6366	0	1
EASEMENT	1070994	100.00%	13	0	SPACE
OWNER	1039249	97.0%	863347	30804	PARKCHESTER PRESERVAT
BLDGCL	1070994	100.00%	200	0	R4
TAXCLASS	1070994	100.00%	11	0	1
EXT	354305	33.1%	3	237094	G
EXCD1	638488	59.62%	129	0	1017.0
STADDR	1070318	99.94%	839280	676	501 SURF AVENUE
ZIP	1041113	97.21%	196	0	10314
EXMPTCL	15529	1.45%	14	0	X1
EXCD2	92948	8.68%	60	0	1017.0

Other Fields:

RECORD: every observation's number, from 1 to 1,070,994;
BBLE: Concatenation of B, BLOCK, LOT (11 digit numeric);
PERIOD: All observations are same, every observation is 'FINAL';
YEAR: All observations are same, every observation is '2010/11';
VALTYPE: All observations are same, every observation is 'AC-TR'
(These fields will not be considered in the detailed information in the following section)

Detailed information for each field

Field 1

Field Name: RECORD

Description:
RECORD is a categorical variable. It represents each property's unique number. All the numbers are different and there are no missing values in this field.

Unique Values:
RECORD has 1070994 unique values.

Field 2

Field Name: BBLE

Description:
BBLE is a categorical variable. It represents concatenation of B, BLOCK and LOT with 11 digit numbers. All the numbers are different and there are no missing values in this field.

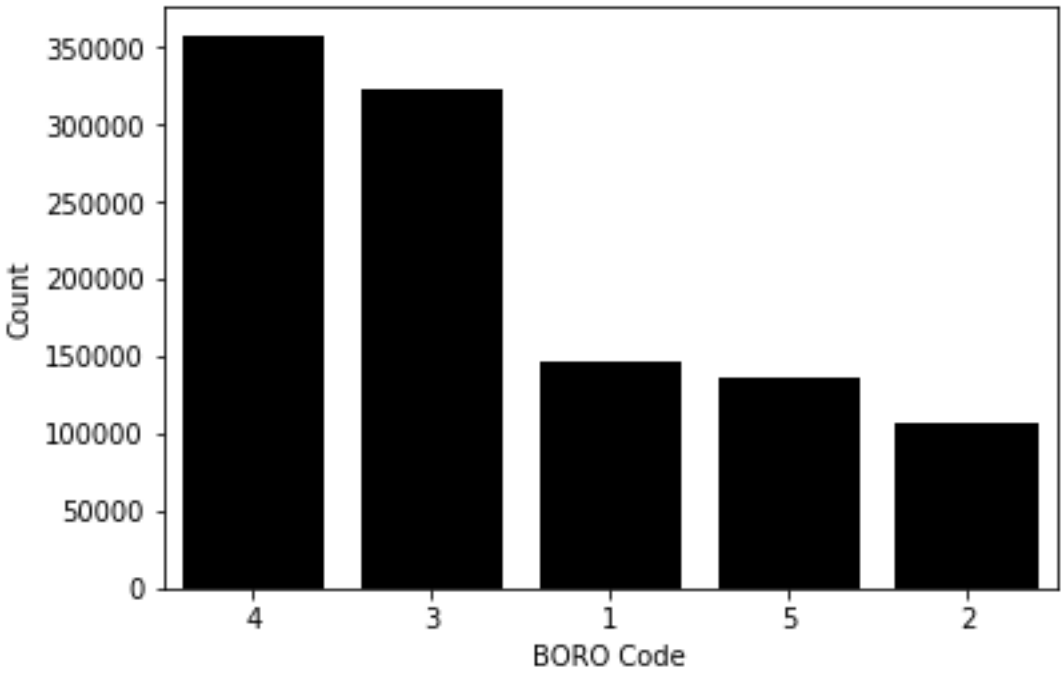
Unique Values:
BBLE has 1070994 unique values.

Field 3

Field Name: B
Description:
B is a categorical variable with labels from 1 to 5. It represents the property's borough code within New York City. The detailed annotations are as follows:

1 = MANHATTAN
2 = BRONX
3 = BROOKLYN
4 = QUEENS
5 = STATEN ISLAND

Unique Values:
B has 5 unique values, ranging from 1 to 5. No missing values. The top 10 most frequently appeared B codes are shown below:



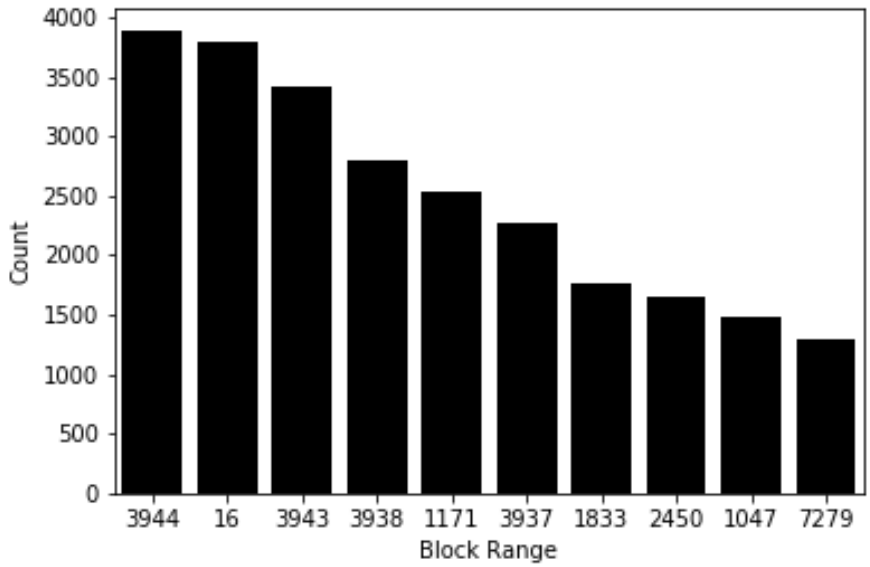
BORO	Percentage (%)
4	33.43%
3	30.18%
1	13.65%
5	12.72%
2	10.02%

Field 4

Field Name: BLOCK

Description:
BLOCK is a categorical variable. It represents the property's valid block range by its borough code within New York City. The detailed annotations are: 1. Manhattan 1-2255 2. Bronx 2260-5958 3. Brooklyn 1-8955 4. Queens 1-16350 5. Staten Island 1-8050

Unique Values:
BLOCK has 13984 unique values, ranging from 1 to 16350. No missing values. The top 10 most frequently appeared BLOCK codes are shown below:



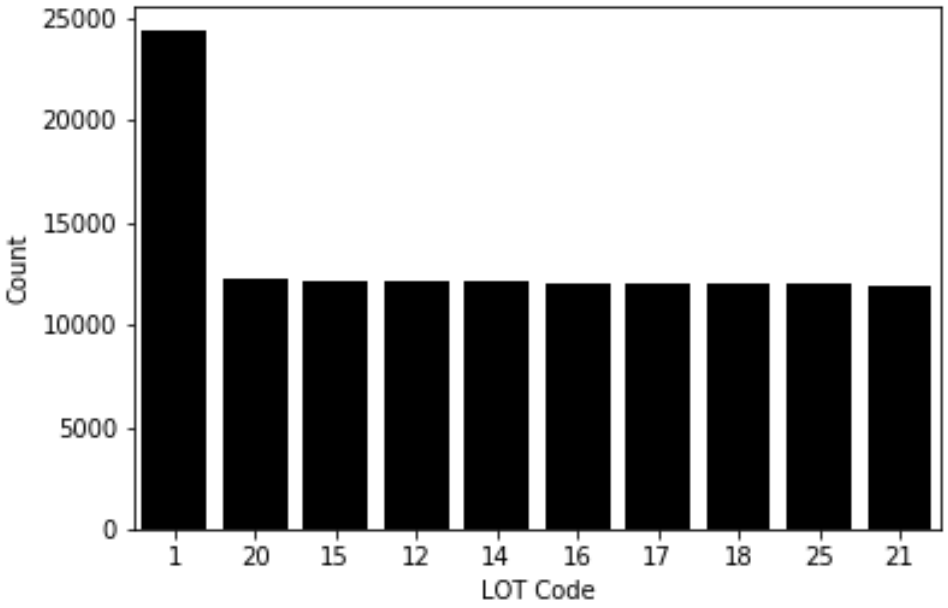
BLOCK	Percentage (%)
3944	0.36%
16	0.35%
3943	0.32%
3938	0.26%
1171	0.24%
3937	0.21%
1833	0.17%
2450	0.15%
1047	0.14%
7279	0.12%

Field 5

Field Name: LOT

Description:
LOT is a categorical variable with 1 to 4 digits. It represents the property's unique lot code within its borough and block.

Unique Values:
LOT has 6366 unique values, ranging from 1 to 9978. No missing values. The top 10 most frequently appeared LOT codes are shown below:



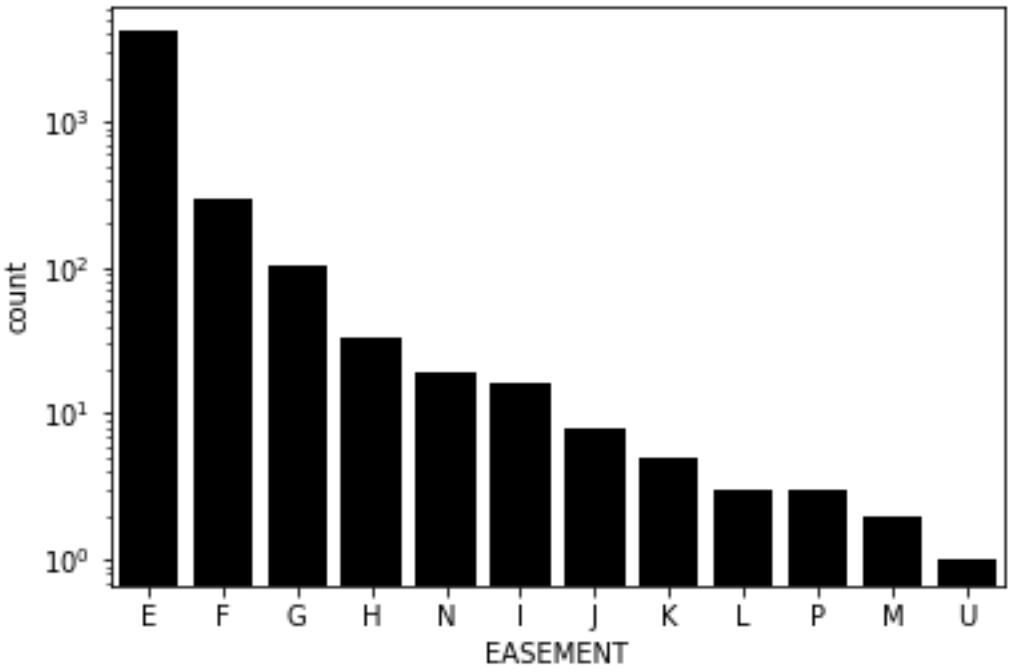
LOT	Percentage (%)
1	2.28%
20	1.15%
15	1.14%
12	1.13%
14	1.13%
16	1.12%
17	1.12%
18	1.12%
25	1.12%
21	1.11%

Field 6

Field Name: EASEMENT

Description:
EASEMENT is a categorical variable with 12 English letters and SPACE. It is a field that is used to describe whether the lot has Easement or not. Here, SPACE means the lot has no Easement.

Unique Values:
EASEMENT has 13 unique values. No missing values and all the blanks mean that these lots have no Easement. The count of each type of EASEMENT column is shown below (Here we ignored the blanks):



EASEMENT	Percentage (%)
E	89.47%
F	6.38%
G	2.20%
H	0.71%
N	0.41%
I	0.35%

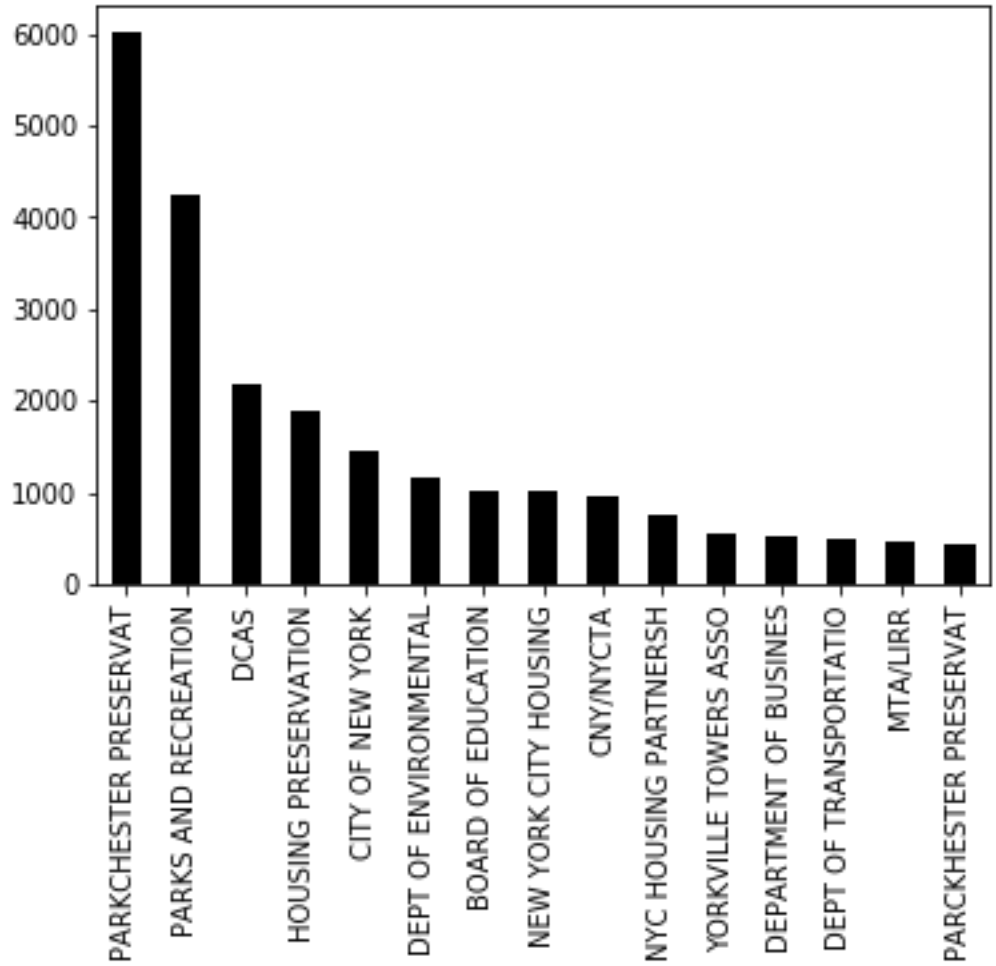
EASEMENT	Percentage (%)
J	0.17%
K	0.11%
L	0.06%
P	0.06%
M	0.04%
U	0.02%

Field 7

Field Name: OWNER

Description:
OWNER is a categorical variable with each property's owner name. The person who owned most properties is called Parkchester Preservat.

Unique Values:
OWNER has 863347 unique values. About 3% values are missing in this column. 30804 records have value zero. The top 10 most frequently appeared owners are shown below:

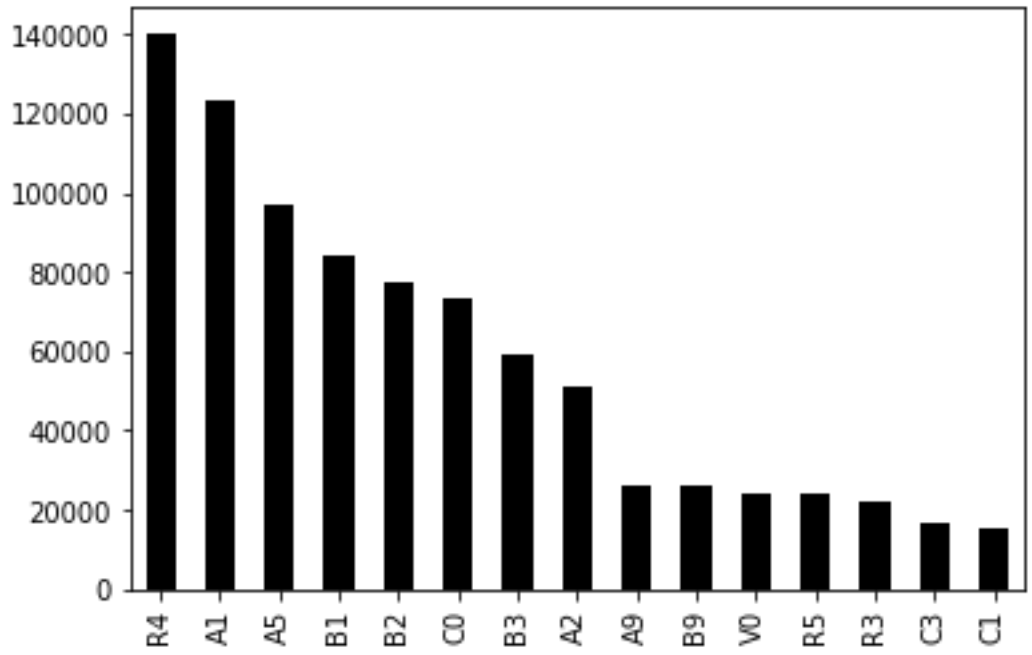


Field 8

Field Name: BLDGCL

Description:
BLDGCL is a categorical variable with Position 1 = ALPHA & Position 2 = NUMERIC. It composes of a letter and a number. It represents the building class. There is a direct correlation between the Building Class and the Tax Class.

Unique Values:
BLDGCL has 200 unique values. No missing values. No records have value zero. The top 10 most frequently appeared building classes are shown below:

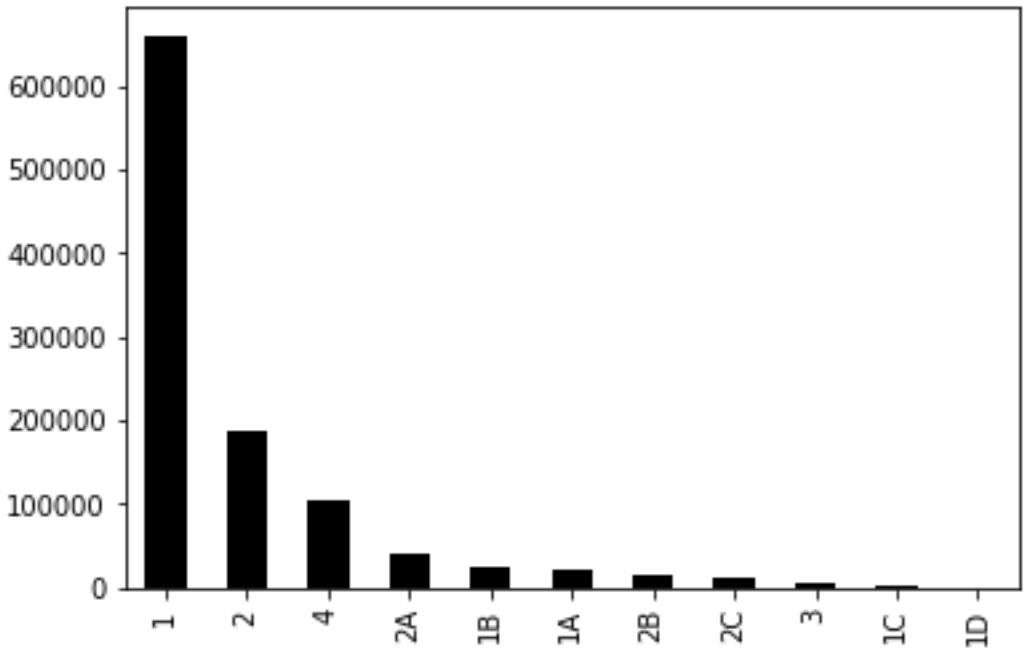


Field 9

Field Name: TAXCLASS

Description:
TAXCLASS is a categorical variable. It represents the property's current property tax class code. The most common field value is 1.

Unique Values:
TAXCLASS has 11 unique values. No missing values. No records have value zero. The count of each type is shown below:

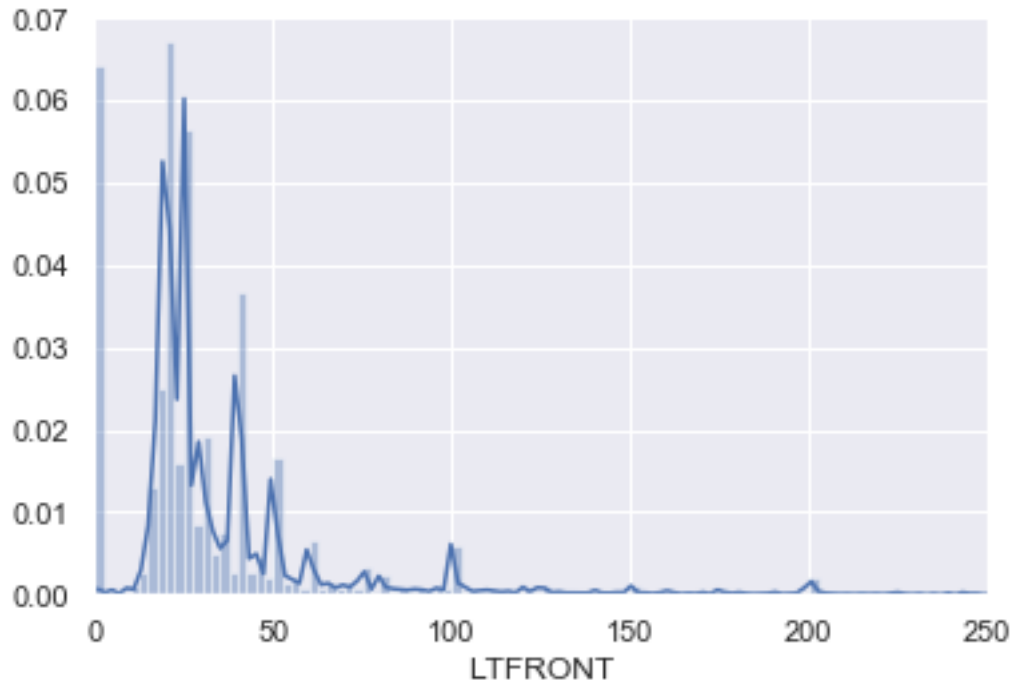


Field 10

Field Name: LTFRONT

Description:
LTFRONT is a numeric variable with range from 0 to 9999. It represents the property's lot frontage in feet. The mean is 36.6 and the standard deviation is 74

Unique Values:
LTFRONT has 1297 unique values. No missing values. The distribution of this field is shown below:

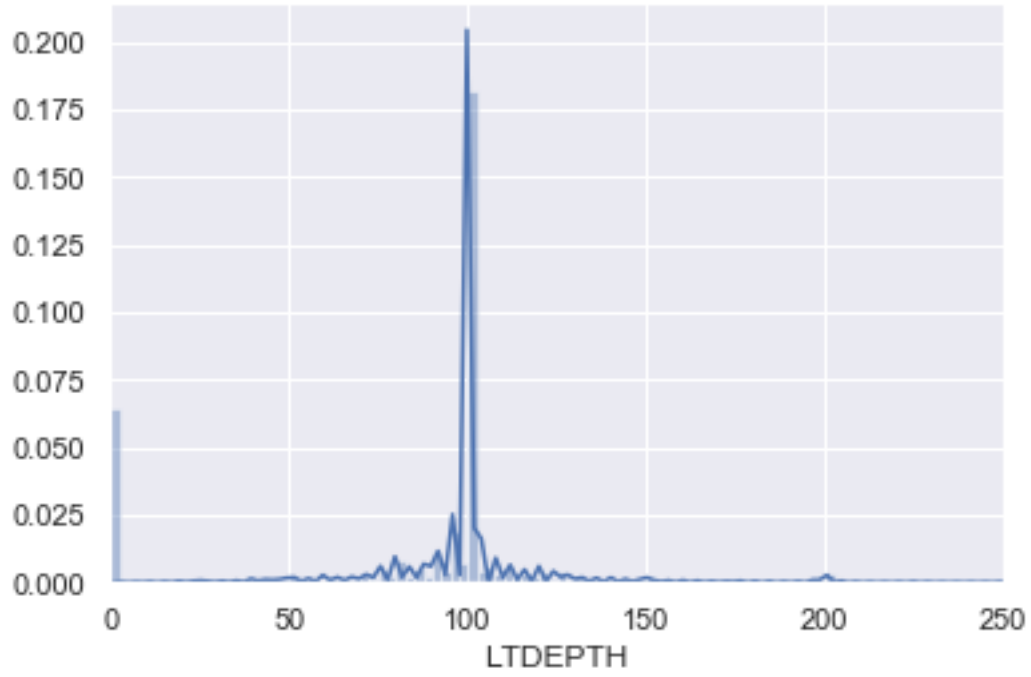


Field 11

Field Name: LTDEPTH

Description:
LTDEPTH is a numeric variable with range from 0 to 9999. It represents the property's lot depth in feet. The mean is 88.9 and the standard deviation is 76.4

Unique Values:
LTDEPTH has 1370 unique values. No missing values. The distribution of this field is shown below:

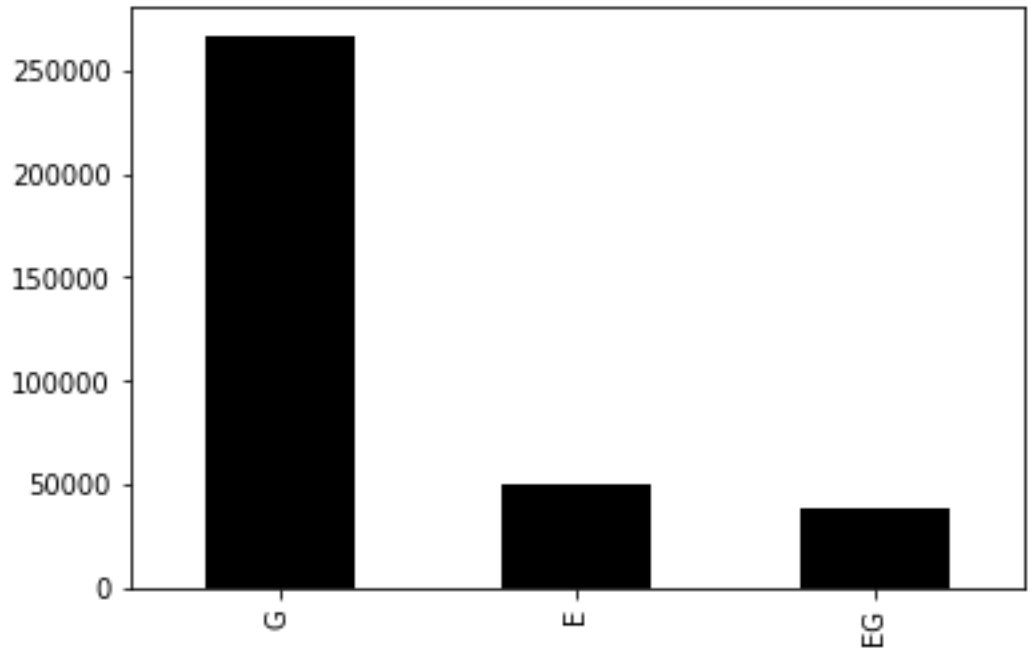


Field 12

Field Name: EXT

Description:
EXT is a categorical variable. It represents the property's extension. 'E': Extension; 'G': Garage; 'EG': Extension and Garage

Unique Values:
EXT has 3 unique values. About 66% data is missing in this field. 237094 records have value zero. The count of each type of EXT is shown below:

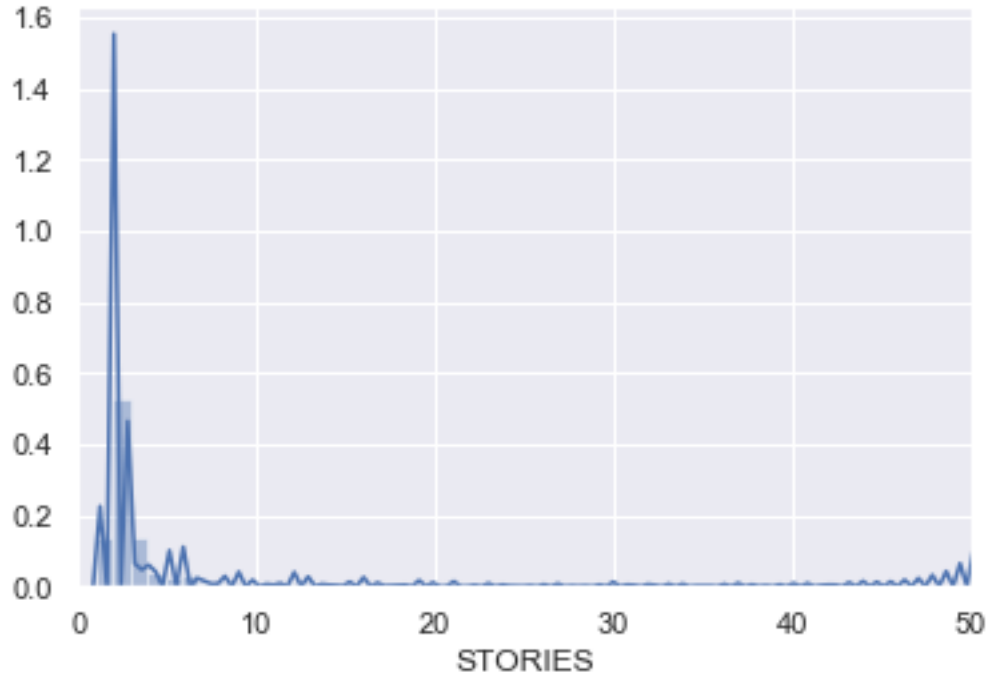


Field 13

Field Name: STORIES

Description:
STORIES is a numeric variable with range from 1 to 119. It represents the property's number of stories. The mean is 5 and the standard deviation is 8.4

Unique Values:
STORIES has 111 unique values. About 5% data is missing in this field. The distribution of this field is shown below:

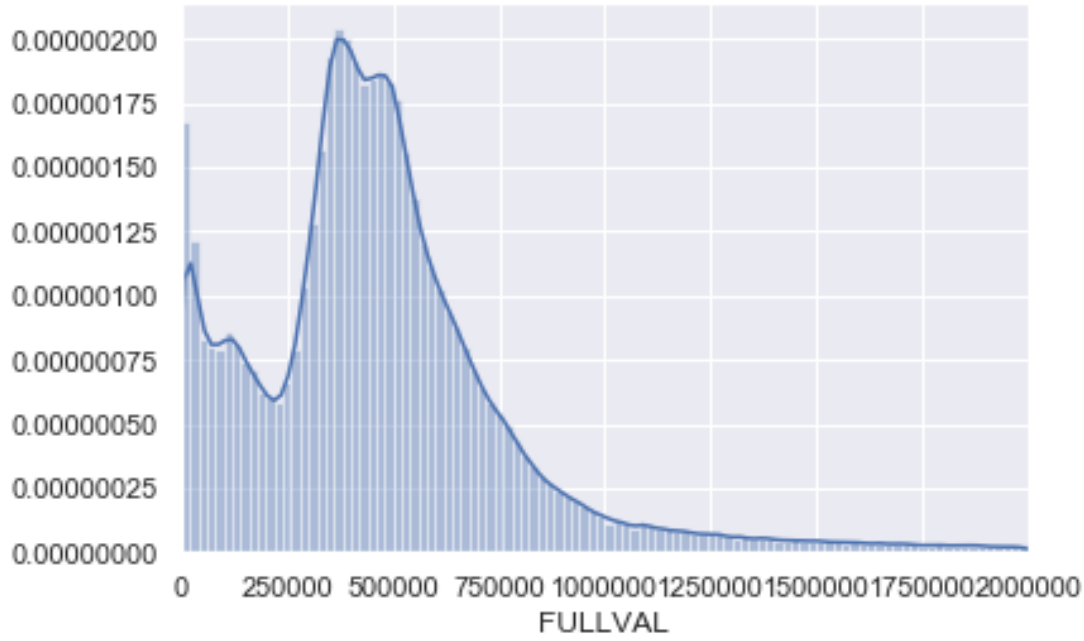


Field 14

Field Name: FULLVAL

Description:
FULLVAL is a numeric variable. It represents the property's full value. The mean is 874264.5 and the standard deviation is very large

Unique Values:
FULLVAL has 109324 unique values. No missing values. The distribution of this field is shown below:

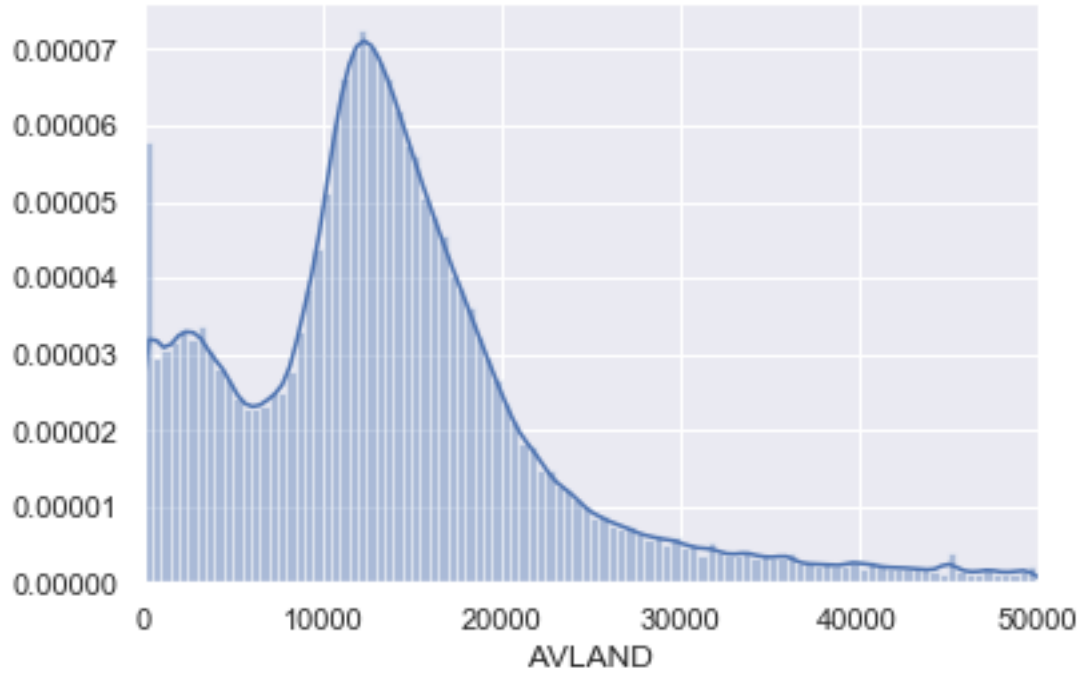


Field 15

Field Name: AVLAND

Description:
AVLAND is a numeric variable. It represents the property's assessed value. The mean is 85067.9 and the standard deviation is very large

Unique Values:
AVLAND has 70921 unique values. No missing values. The distribution of this field is shown below:

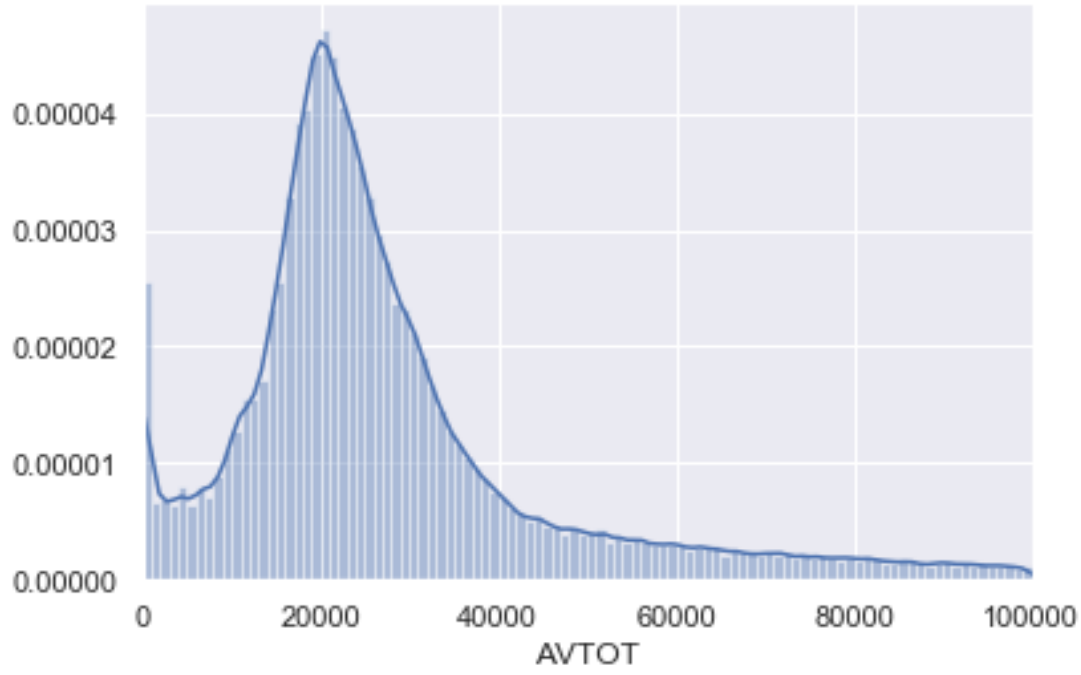


Field 16

Field Name: AVTOT

Description:
AVTOT is a numeric variable. It represents the property's total assessed value. The mean is 227238.2 and the standard deviation is very large

Unique Values:
AVTOT has 112914 unique values. No missing values. The distribution of this field is shown below:

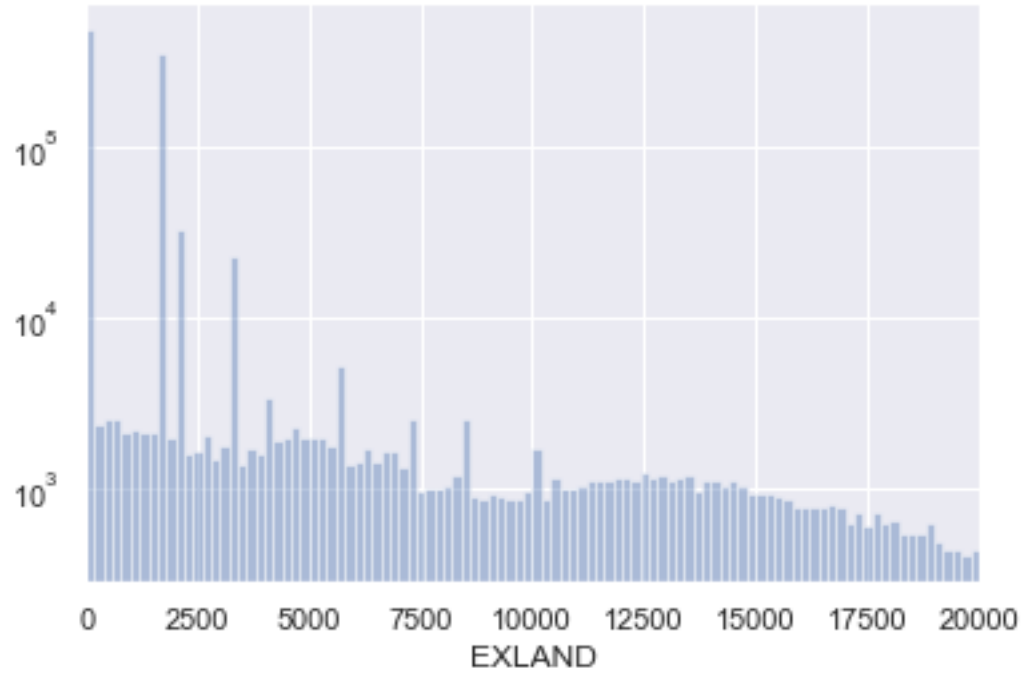


Field 17

Field Name: EXLAND

Description:
EXLAND is a numeric variable. It represents the property's transitional exempt land value. The mean is 36423.9 and the standard deviation is very large

Unique Values:
EXLAND has 33419 unique values. No missing values. The distribution of this field is shown below:

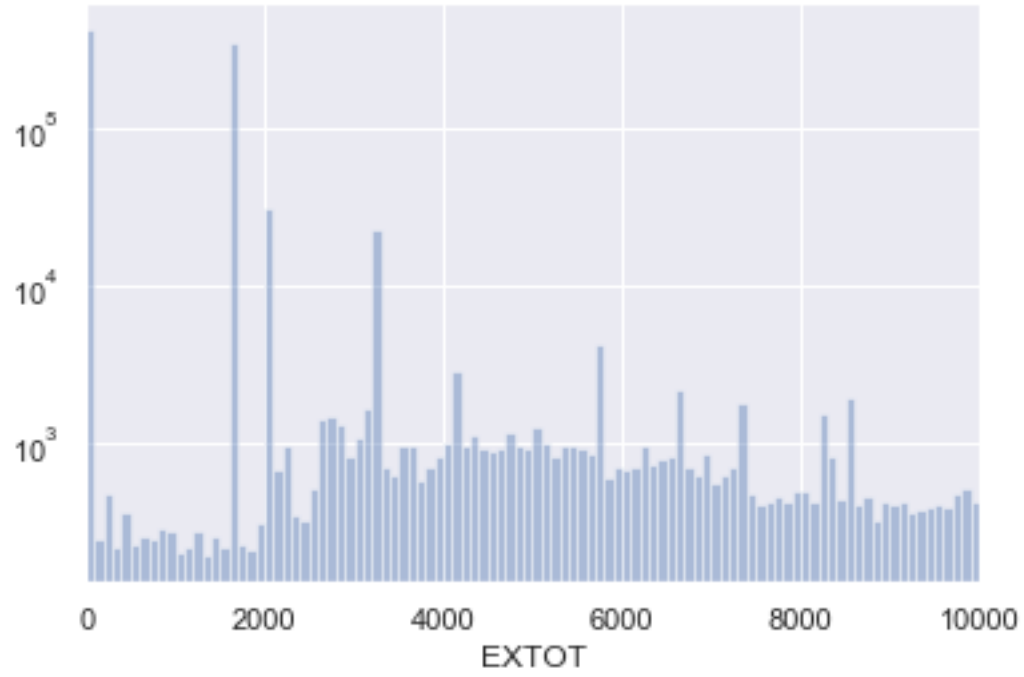


Field 18

Field Name: EXTOT

Description:
EXTOT is a numeric variable. It represents the property's transitional exempt total value. The mean is 91187 and the standard deviation is very large

Unique Values:
EXTOT has 64255 unique values. No missing values. The distribution of this field is shown below:

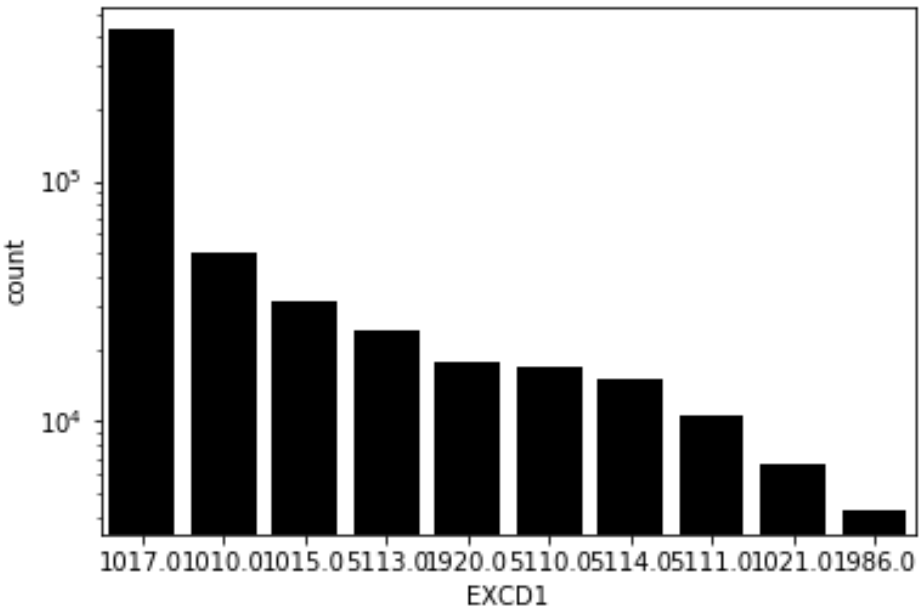


Field 19

Field Name: EXCD1

Description:
EXCD1 is a categorical variable with digit numbers. It is a field that is used to describe the exemption code 1 of the lot.

Unique Values:
EXCD1 has 129 unique values. About 40% values are missing in this column. No records with value zero. The top 10 most frequently appeared EXCD1 code is shown below:



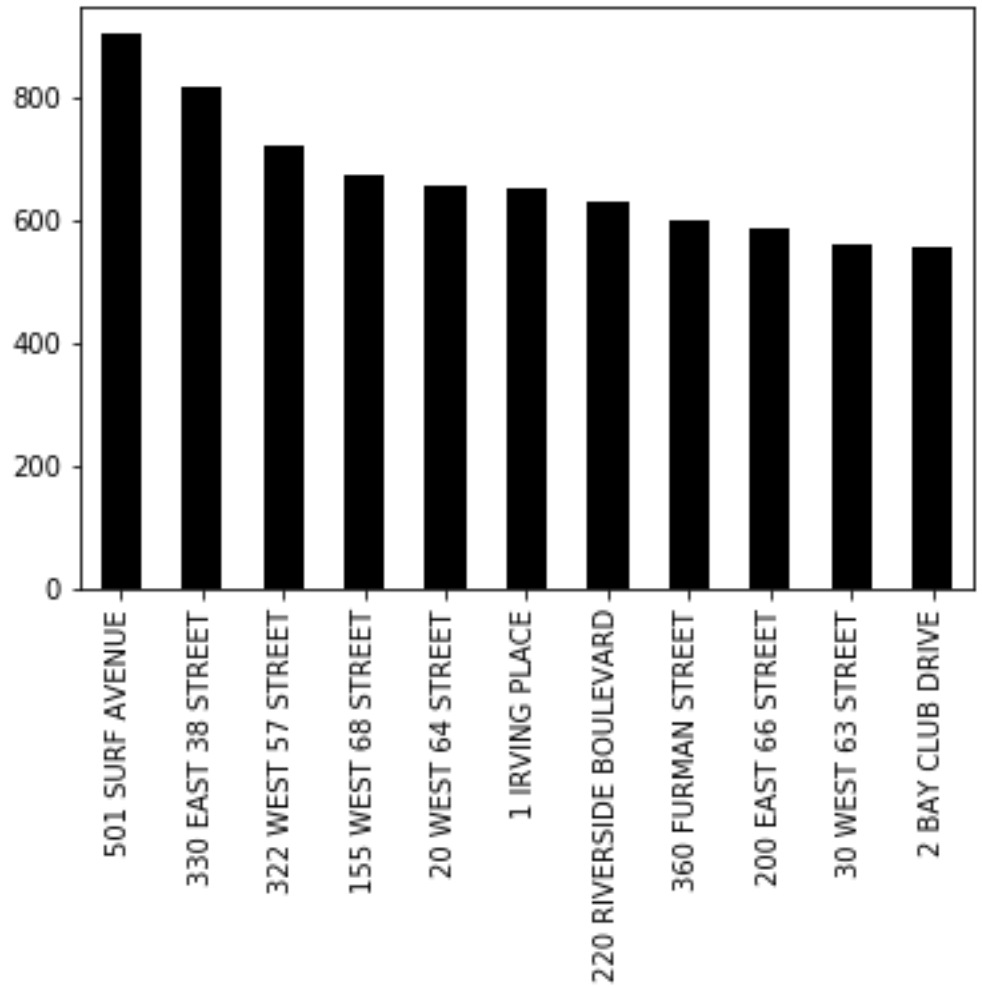
EXCD1	Percentage (%)
1017.0	66.62%
1010.0	7.79%
1015.0	4.91%
5113.0	3.74%
1920.0	2.76%
5110.0	2.64%
5114.0	2.35%
5111.0	1.66%
1021.0	1.04%
1986.0	0.66%

Field 20

Field Name: STADDR

Description:
STADDR is a categorical variable. It is a field that is used to describe the street addresses of the properties.

Unique Values:
STADDR has 839280 unique values. About 0.06% values are missing in this column. There are 676 records with value zero. The top 11 most frequently appeared street addresses are shown below:

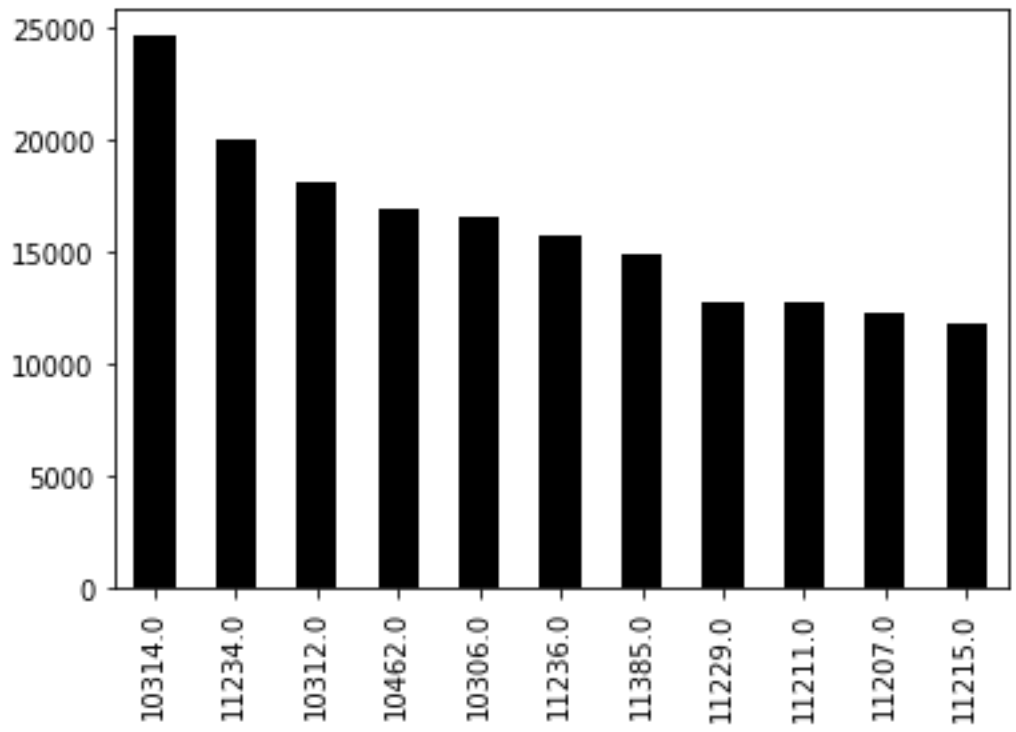


Field 21

Field Name: ZIP

Description:
ZIP is a categorical variable. It is a field that is used to describe the zip codes of the properties.

Unique Values:
ZIP has 196 unique values. About 3% values are missing in this column. There are 0 records with value zero. The top 11 most frequently appeared zip codes are shown below:

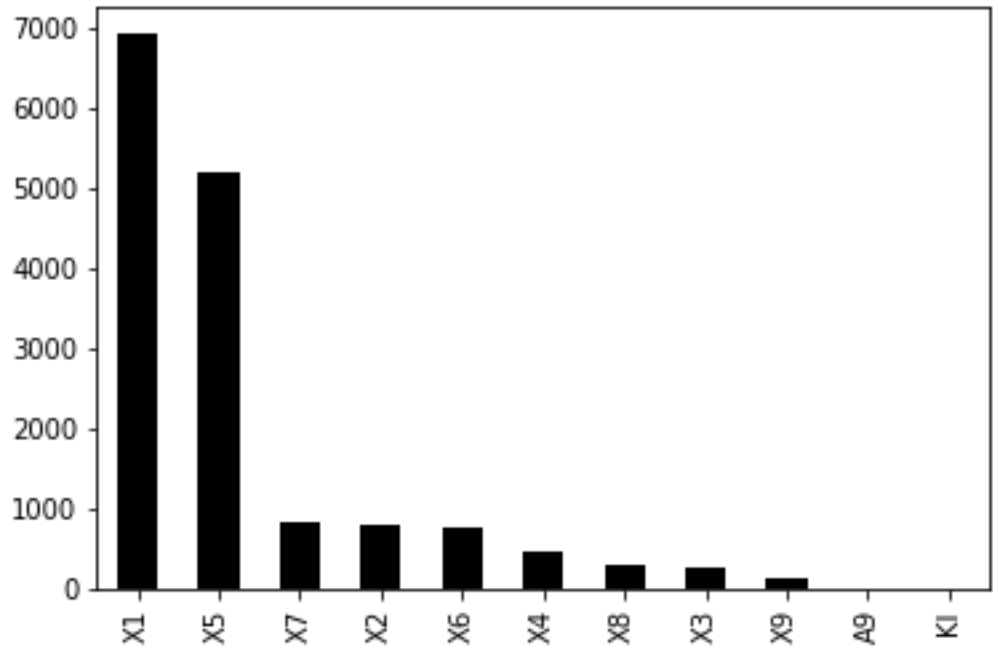


Field 22

Field Name: EXMPTCL

Description:
EXMPTCL is a categorical variable. It is a field that is used to describe the exempt class used for fully exempt properties only

Unique Values:
ZIP has 14 unique values. About 98% values are missing in this column. (This is normal because not a lot of properties are fully exempt). There are 0 records with value zero. The top 11 most frequently appeared zip codes are shown below:

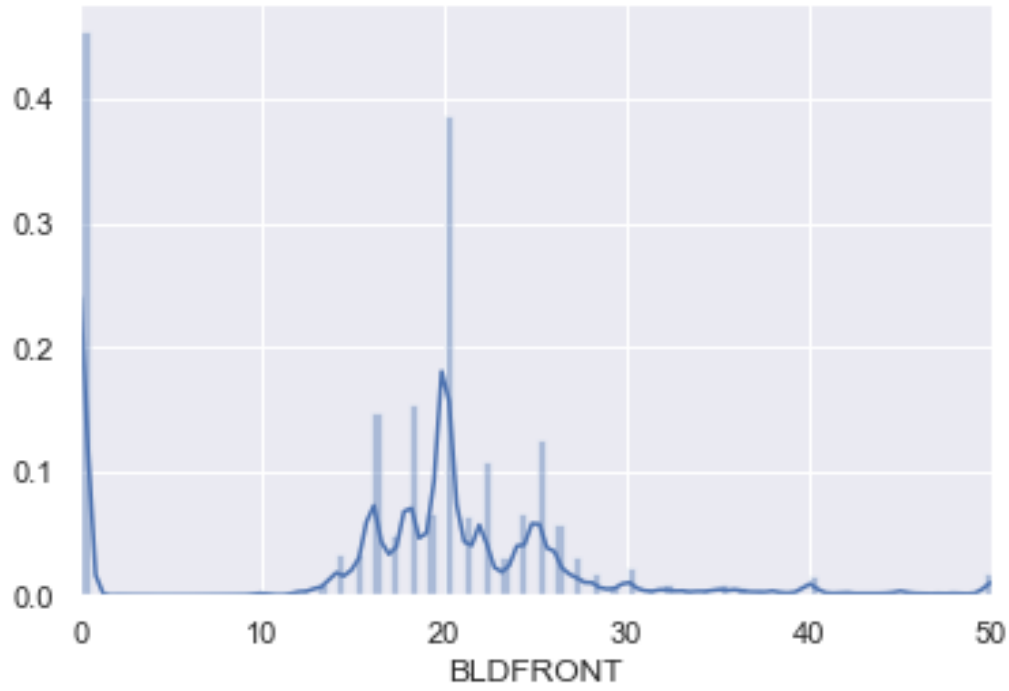


Field 23

Field Name: BLDFRONT

Description:
BLDFRONT is a numeric variable. It represents the property's building frontage in feet. The mean is 23 and the standard deviation is 35.6

Unique Values:
BLDFRONT has 612 unique values. No missing values. The distribution of this field is shown below:

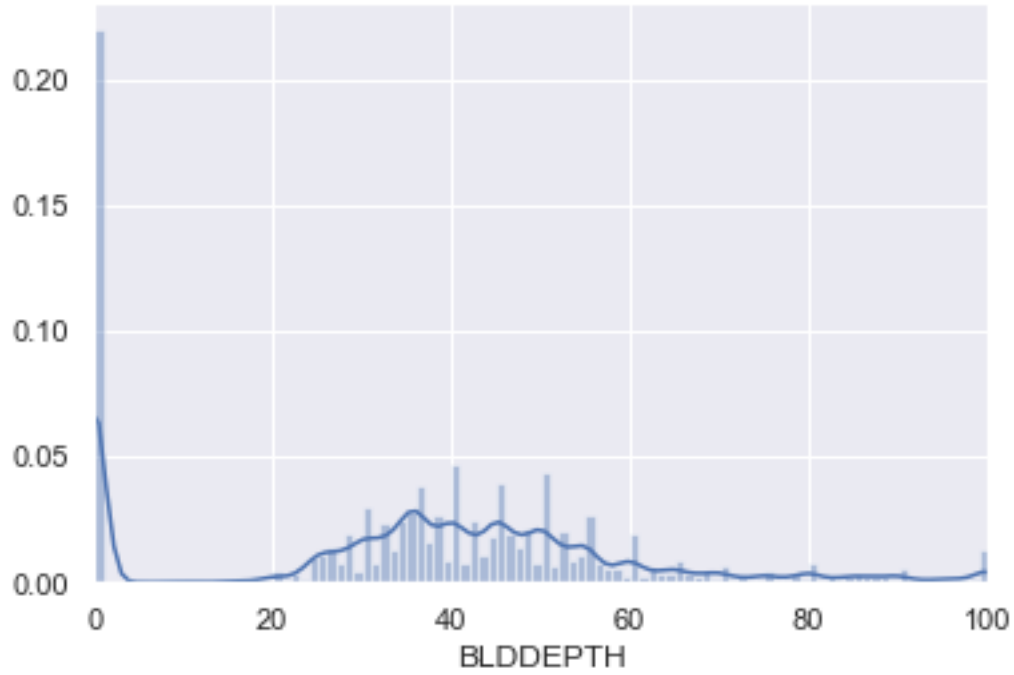


Field 24

Field Name: BLDDEPTH

Description:
BLDDEPTH is a numeric variable. It represents the property's building depth in feet. The mean is 39.9 and the standard deviation is 42.7

Unique Values:
BLDDEPTH has 621 unique values. No missing values. The distribution of this field is shown below:

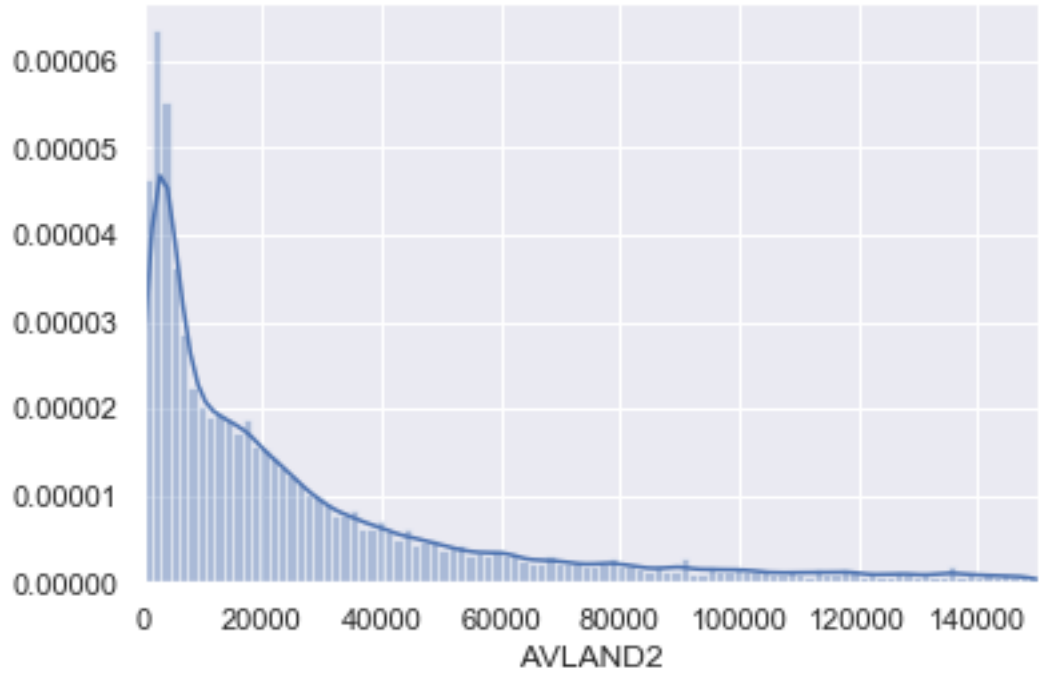


Field 25

Field Name: AVLAND2

Description:
AVLAND2 is a numeric variable. It represents the property's assessed value of land 2. The mean is 246235.7 and the standard deviation is pretty large

Unique Values:
AVLAND2 has 58591 unique values. About 75% data is missing in this field. The distribution of this field is shown below:

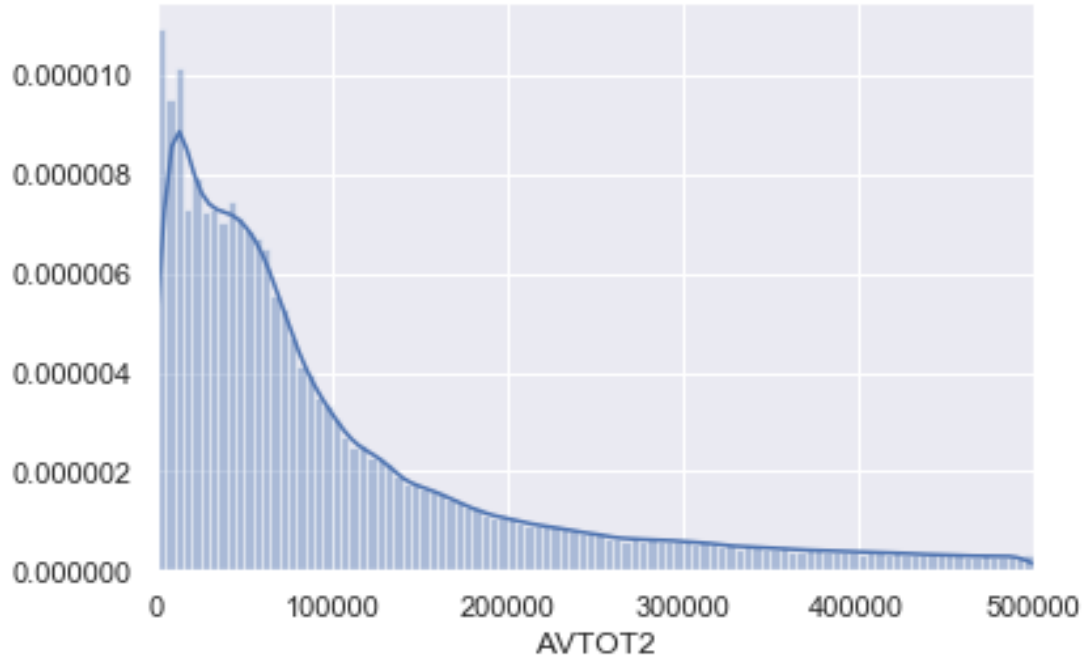


Field 26

Field Name: AVTOT2

Description:
AVTOT2 is a numeric variable. It represents the property's assessed total value of land 2. The mean is 713911.4 and the standard deviation is pretty large

Unique Values:
AVTOT2 has 11130 unique values. About 75% data is missing in this field. The distribution of this field is shown below:

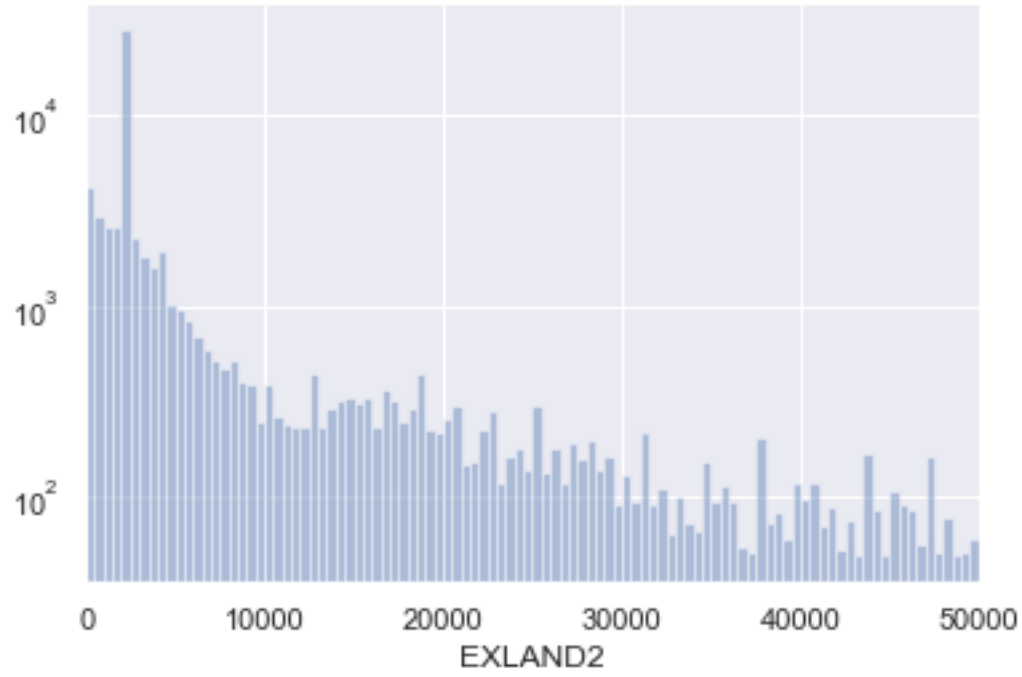


Field 27

Field Name: EXLAND2

Description:
EXLAND2 is a numeric variable. It represents the property's transitional exempt land value 2. The mean is 351235.7 and the standard deviation is pretty large

Unique Values:
EXLAND2 has 22195 unique values. About 92% data is missing in this field. The distribution of this field is shown below:

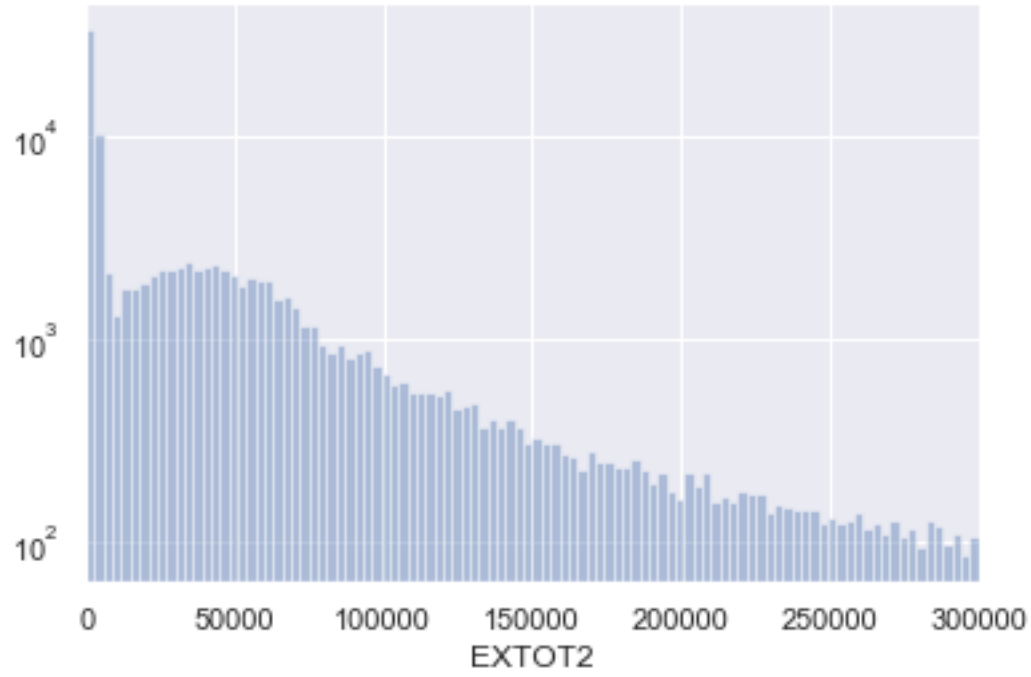


Field 28

Field Name: EXTOT2

Description:
EXTOT2 is a numeric variable. It represents the property's transitional exempt total value 2. The mean is 656768.3 and the standard deviation is pretty large

Unique Values:
EXTOT2 has 48348 unique values. About 88% data is missing in this field. The distribution of this field is shown below:

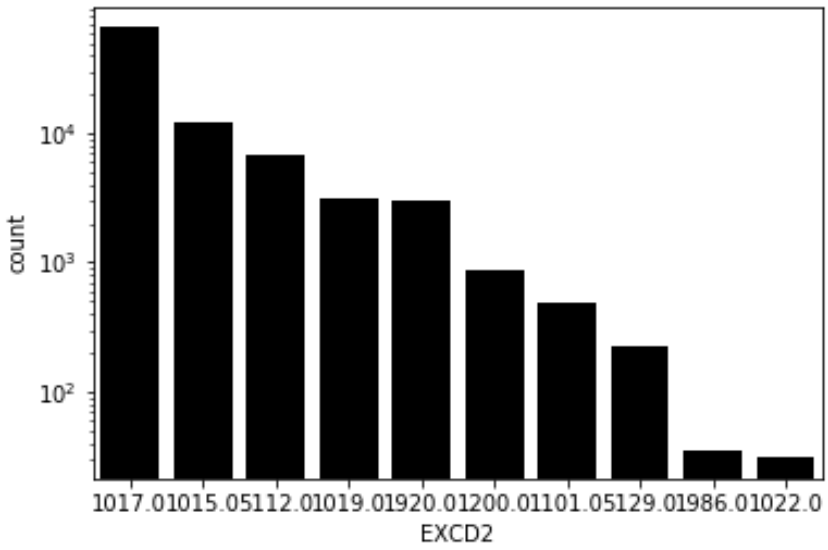


Field 29

Field Name: EXCD2

Description:
EXCD2 is a categorical variable with digit numbers. It is a field that is used to describe the exemption code 2 of the lot.

Unique Values:
EXCD2 has 60 unique values. Over 90% values are missing in this column. No records with value zero. The top 10 most frequently appeared EXCD2 code is shown below:



EXCD2	Percentage (%)
1017.0	70.77%
1015.0	13.27%
5112.0	7.39%
1019.0	3.42%
1920.0	3.19%
1200.0	0.95%
1101.0	0.53%
5129.0	0.24%
1986.0	0.04%
1022.0	0.03%