USC Marshall School of Business

Unsupervised Machine Learning Fraud Detection on NYC Property

Group 4 -Juxing Wang, Xiao Xiao, Yu Qin
Tianxing Wang, Yifeng Wang, Di Zhang

Feb 21th, 2019

Table of Content

Executive Summary

This is an unsupervised fraud detection project. With 1070994 properties at hand, we rank all the properties with their anomaly score and try to find the frauds as well as their features. Our work pipeline is data cleaning with R, variable construction, dimensionality reduction with PCA, score calculation and finally result analysis.

A further feature analysis of the top 10 suspicious fraud record gives us a deeper understanding of when to ring the alarm with future property data.

1. Description of Data

1.1 Description of dataset

The data was reported as of 11/17/2010 as the annual report of Property Valuation and Assessment Data by the Department of finance of the City Government of NYC. It has 32 fields, and 1,070,994 number of records in this dataset.

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..

A Data Quality Report in the appendix was attached for further details.

1.2 Summary of variables

The Table1-1 below shows the summary of numerical variables:

Field	#values	%populated	#unique values	#zeros	mean	median	min	max	standard deviation
LTFRONT	1070994	100	1297	169108	36.63	25	0	9999	74.03284
LTDEPTH	1070994	100	1370	170128	88.86	100	0	9999	76.39628
FULLVAL	1070994	100	109324	13007	8.74E+05	4.47E+05	0	6.15E+09	11582431
AVLAND	1070994	100	70921	13009	8.51E+04	1.37E+04	0	2.67E+09	4057260
AVTOT	1070994	100	112914	13007	4.67E+09	2.53E+04	0	4.67E+09	6877529
EXLAND	1070994	100	33419	491699	3.64E+04	1.62E+03	0	2.67E+09	3981576
EXTOT	1070994	100	64255	432572	9.12E+04	1.62E+03	0	4.67E+09	6508403
EXCD1	638388	59.61639	130	0	1602	1017	1010	7170	1384.227
BLDFRONT	1070994	100	612	28815	23.04	20	0	7575	35.5797
BLDDEPTH	1070994	100	621	228853	39.92	39	0	9393	42.70715
AVLAND2	282726	26.39847	58592	0	2.46E+05	2.01E+04	3.00E+00	2.37E+09	6178963
AVTOT2	282732	26.39903	111361	0	4.50E+09	8.00E+04	3.00E+00	4.50E+09	11652529
EXLAND2	87449	8.165218	22196	0	3.51E+05	3.05E+03	1.00E+00	2.37E+09	10802213
EXTOT2	130828	12.21557	48349	0	6.57E+05	4.50E+09	7.00E+00	4.50E+09	16072510
EXCD2	92948	8.678667	61	0	1364	1017	1011	7160	1094.706

Table 1-1 Summary of numerical variables

The Table1-2 below shows the summary of categorical variables:

Field	#values	%pupolated	#unique values
RECORD	1070994	100	1070994
BBLE	1070994	100	1070994
B	1070994	100	5
BLOCK	1070994	100	13984
LOT	1070994	100	6366
EASEMENT	4636	43.28689	13
OWNER	1039249	97.03593	863348
BLDGCL	1070994	100	200
TAXCLASS	1070994	100	11
EXT	354305	33.08188	4
STORIES	1014730	100	112
STADDR	1070318	99.93688	839281
ZIP	1041104	97.20913	197
EXMPTCL	15579	1.45463	15
PERIOD	1070994	100	1
YEAR	1070994	100	1
VALTYPE	1070994	100	1

Table 1-2: Summary of categorical variables

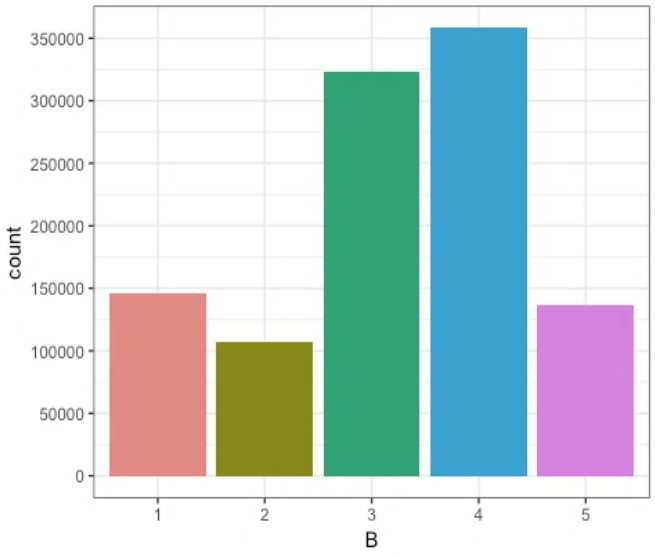
1.3 Important variables:

- **Field B:**

Type: Categorical

1-5 has their meaning that: 1 - MANHATTAN, 2 - BRONX, 3 - BROOKLYN, 4 - QUEENS, 5 - STATEN ISLAND. From the chart we could see, in field “B”, 4(QUEENS) has the highest count.

Number of unique values: 5 %populated: 100

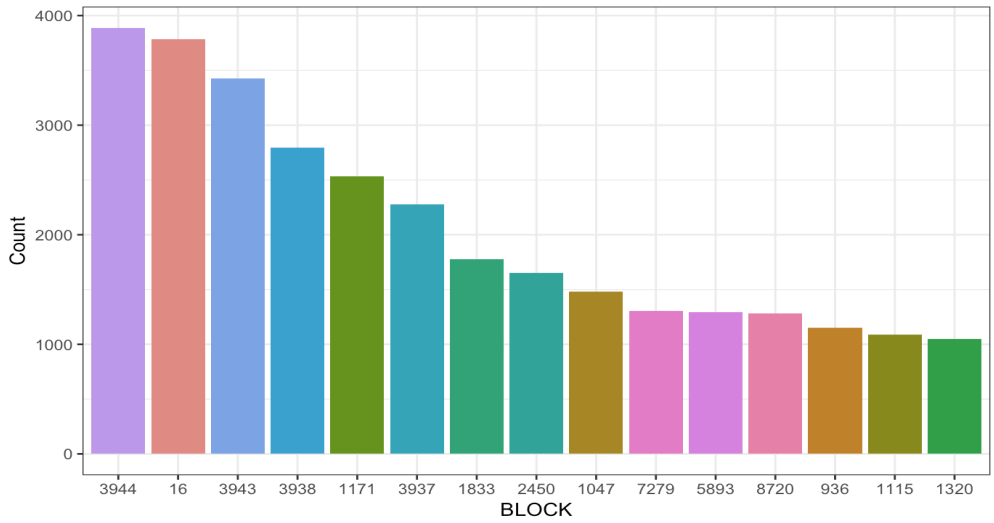


- **Field BLOCK:**

Type: Categorical

Number of unique values: 13984 %populated: 100

And the distribution of Block is as follows:

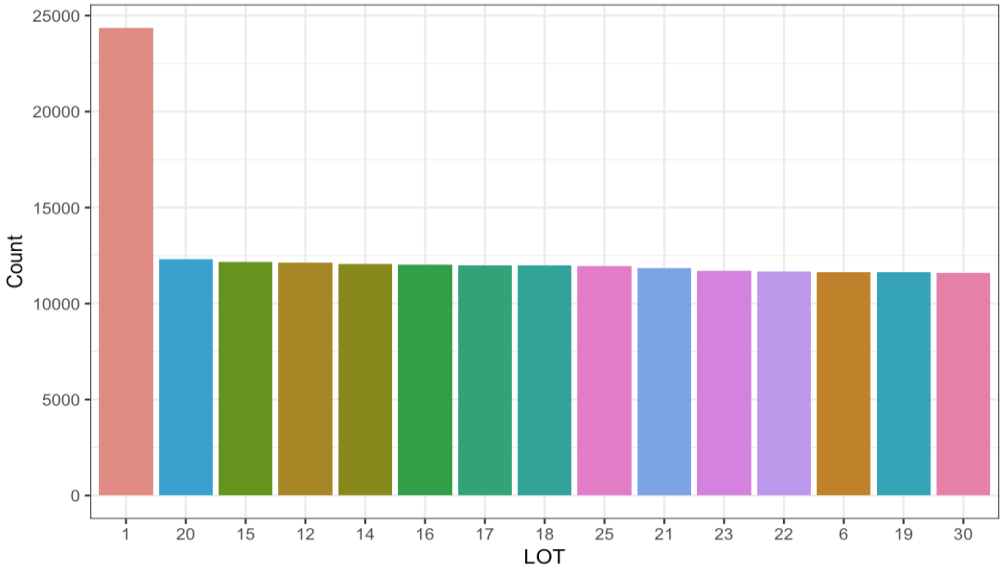


- **Field LOT:**

Type: Categorical

Number of unique values: 6366 %populated: 100

And the distribution of LOT is as follows:



Field TAXCLASS: Current Property Tax Class Code (NYS Classification)

Type: Categorical

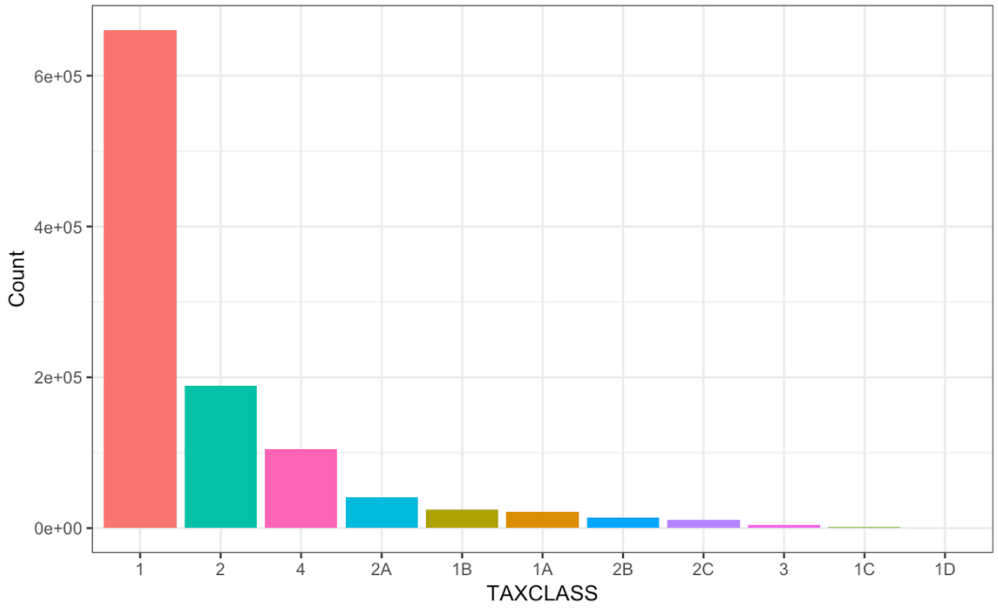
Number of unique values: 11

%populated: 100

Following is the record amount of each TAXCLASS:

##	x	freq
## 1	1	660721
## 2	2	188612
## 3	4	104310
## 4	2A	40574
## 5	1B	24738
## 6	1A	21667
## 7	2B	13964
## 8	2C	10795
## 9	3	4638
## 10	1C	946

And the distribution of TAXCLASS is as follows:



From the chart, we could tell that properties that belongs to the taxclass 1 take the most proportion.

- **Field STORIES：** The number of stories for the property (# of Floors)

Type: Categorical

Number of unique values: 112

%populated: 94.74656

Following are the first 15 highest number of stories:

##	x	freq
## 1	2.0	415092
## 2	3.0	130127
## 3	1.0	96706
## 4	2.5	82292
## 5	4.0	38342
## 6	6.0	30936
## 7	5.0	25971
## 8	1.5	24770
## 9	2.7	13595
## 10	12.0	12198
## 11	8.0	11953
## 12	7.0	11899
## 13	1.6	8954
## 14	9.0	7343
## 15	13.0	7330

We could tell that 2 stories takes great part in all properties in NY.

- **Field FULLVAL:** Total market value of the property

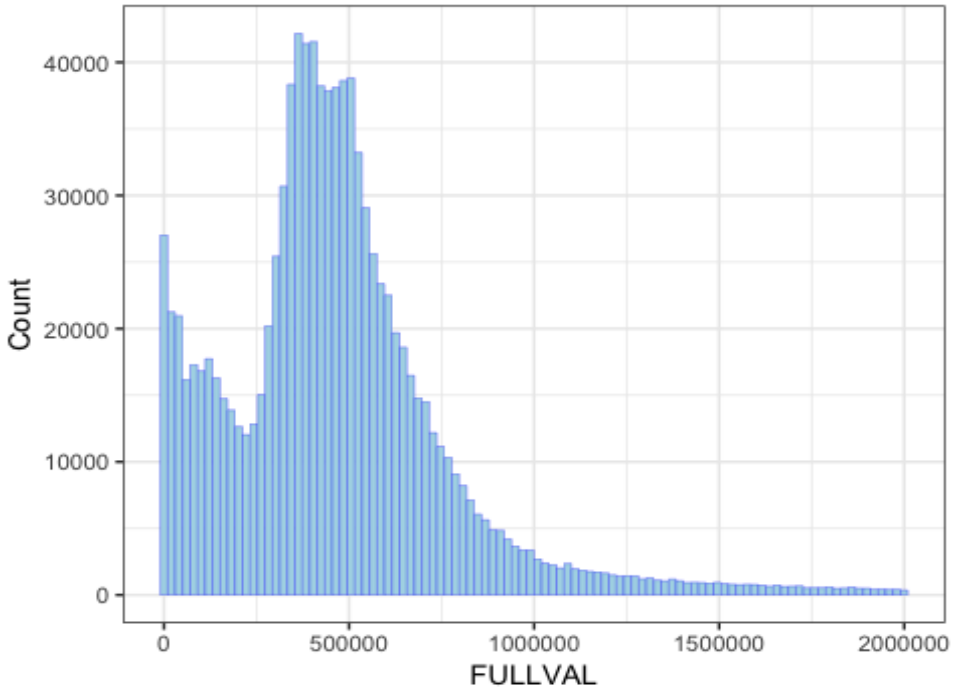
Type: Numerical

Number of unique values: 109324

%populated: 100

##	Min.	1st Qu.	Median	Mean	**3rd Qu.**	**Max.**
##	0.000e+00	3.040e+05	4.470e+05	8.743e+05	6.190e+05	6.150e+09

Standard deviation: 11582431



Percentage of records showed in the chart above is: 96.31221

From this chart, most of the market value of property is about 5000000.

- **Field AVLAND:** Assessed value of land

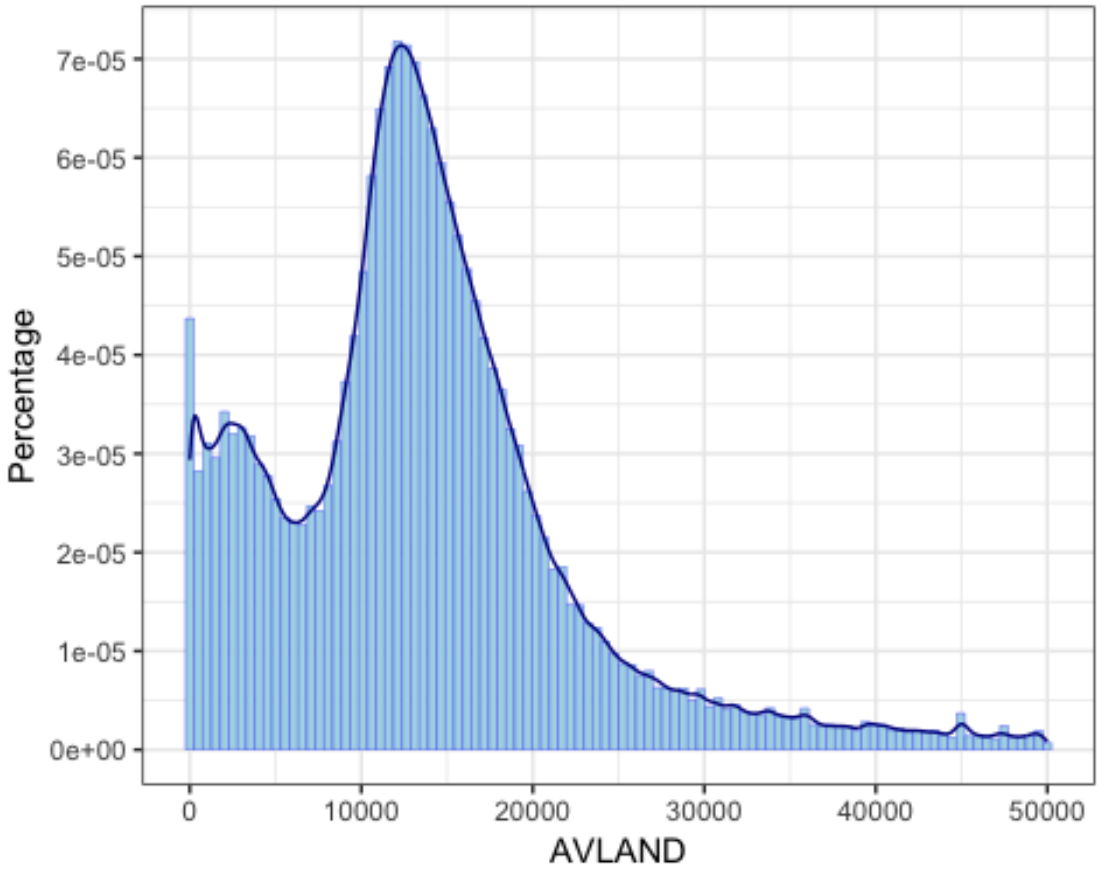
Type: Numerical

Number of unique values: 70921

%populated: 100

##	Min.	1st Qu.	Median	Mean	**3rd Qu.**	**Max.**
##	0.000e+00	9.180e+03	1.368e+04	8.507e+04	1.974e+04	2.668e+09

Standard deviation: 4057260



Percentage of records showed in the chart above is: 90.52721

From the chart above, we could see that most of AVLAND concentrate on between 0 and 20000.

- **Field AVTOT：** Assessed value of total value

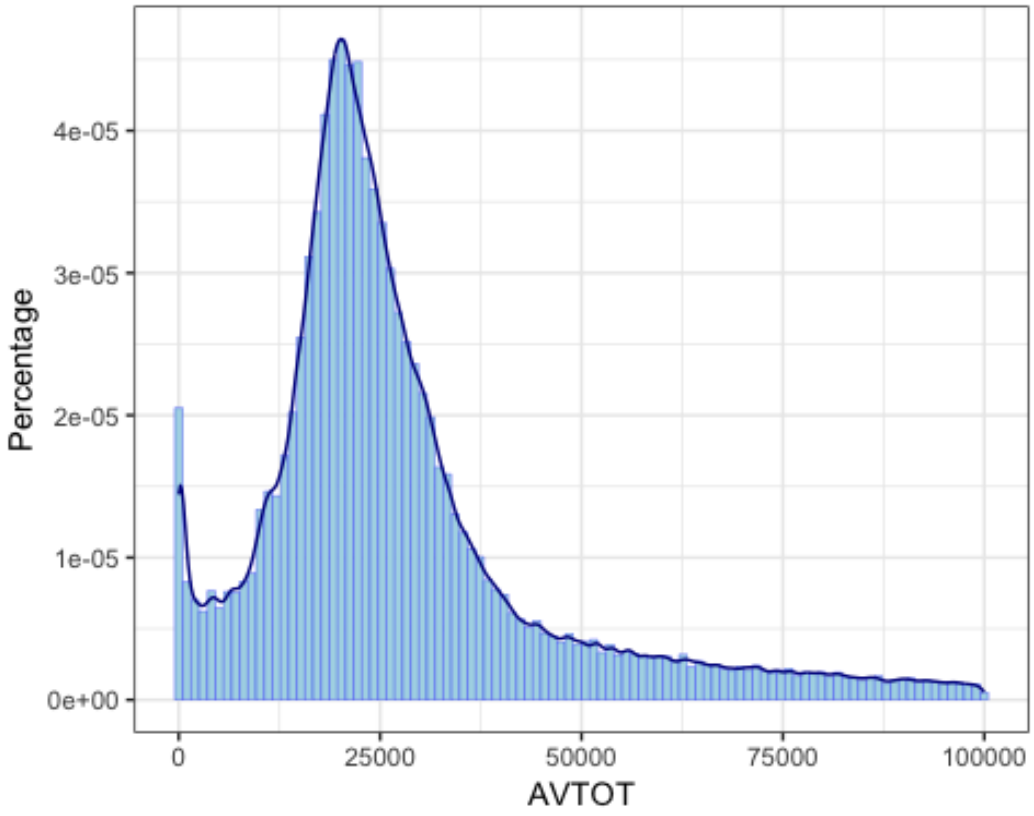
Type: Numerical

Number of unique values: 112914

%populated: 100

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	0.000e+00	1.837e+04	2.534e+04	2.272e+05	4.544e+04	4.668e+09

Standard deviation:



Percentage of records showed in the chart above is: 86.0523

So, the AVTOT around 25000 takes great proportion of all properties in NY.

2. Data Cleaning

We convert all “0” values to NA and use the following rules to fill NA values, for analysis later :

- **ZIP**

Aggregate by B, BLOCK, TAXCLASS. Use the mode of that group. If no records found within a 3-layer (2-layer) aggregation, trace its 2-layer (1- layer) aggregation for mode. For example: if there is no record in a (B, BLOCK, TAXCLASS) group, trace the (B, BLOCK) group it belongs to for mode

- **TAXCLASS**

No missing value

- **B(BOROUGH)**

No missing value

- **STORIES**

Aggregate by B, BLOCK, TAXCLASS. Use the mode of that group. If no records found within a 3-layer (2-layer) aggregation, trace its 2-layer (1- layer) aggregation for mode.

- **LTFRONT/ LTDEPTH**

Aggregate by B, ZIP and STORIES. Use the median of that group.

If no records found within a 3-layer (2-layer) aggregation, trace its 2-layer (1- layer) aggregation for median.

- **BLDFRONT / BLDDEPTH**

Aggregate by B, ZIP and STORIES. Use the median of that group.

If no records found within a 3-layer (2-layer) aggregation, trace its 2-layer (1- layer) aggregation for median.

- **AVTOT**

Aggregate by B, ZIP and STORIES and TAXCLASS. Use the median.

If no records found within an aggregation, trace its higher layer aggregation for median.

- **FULLVAL**

Aggregate by B, ZIP and STORIES and TAXCLASS. Use the median.

If no records found within an aggregation, trace its higher layer aggregation for median.

- **AVLAND**

Aggregate by B, ZIP and STORIES and TAXCLASS. Use the median.

If no records found within an aggregation, trace its higher layer aggregation for median.

3. Variable Creation

Based on original dataset, we extracted 5 variables to help to create additional variables, and there are LTFRON, LTDEPTH, BLDFRONT, BLDDEPTH, and STORIES, as listed below：

Abbreviation	Description
LTFRONT	Lots front length
LTDEPTH	Lot depth length
BLDFRONT	Building front length
BLDDEPTH	Building depth length
STORIES	Number of stories

We calculated additional 3 variables, based on the existing variables. There are LOTAREA, BLDAREA, and BLDVOL, as formula description has shown below:

Abbreviation/formula	Description
LOTAREA = LTFRONT * LTDEPTH	Property lot sizes
BLDAREA = BLDFRONT * BLDDEPTH	Building sizes
BLDVOL = BLDAREA * STORIES	Building volumes

With existing data, FULLVAL, AVALAND, AVTOT, we computed 9 additional variables with previous 3 calculated variables, as formula has shown below:

full_lot = FULLVAL / LOTAREA: a unit of lot area price based on market property value

full_bld = FULLVAL / BLDAREA: a unit of building area price based on market property value

full_bldvol = FULLVAL / BLDVOL: a unit of building volume price based on market property value

avl_lot = AVLAND / LOTAREA: a unit of lot area price based on assessed land value

avl_bld = AVLAND / BLDAREA: a unit of building area price based on assessed land value

avl_bldvol = AVLAND / BLDVOL: a unit of building volume price based on assessed land value

avt_lot = AVTOT / LOTAREA: a unit of lot area price based on property assessed value

avt_bld = AVTOT / BLDAREA: a unit of building area price based on property assessed value

avt_bldvol = AVTOT / BLDVOL: a unit of building volume price based on property assessed value

Then, we used 5 variables as entity to differentiate 9 variables above. There were ZIP5, ZIP3, TAXCLASS, B, and All

Abbreviation	Description
ZIP5	ZIP variable
ZIP3	First 3 digits of ZIP variable
TAXCLASS	Tax class based on each property
B	Borough level
All	Take all the records as one group

The 9 variables were grouped based on the 5 entities variables. There were total 45 expert variables, as definitions have shown below

Record	Variable Name	Description
1	full_lot_zip5	Ratio of full_lot and Average full_lot grouped by ZIP5
2	full_bld_zip5	Ratio of full_bld and Average full_bld grouped by ZIP5
3	full_bldvol_zip5	Ratio of full_bldvol and Average full_bldvol grouped by ZIP5
4	avl_lot_zip5	Ratio of avl_lot and Average avl_lot grouped by ZIP5
5	avl_bld_zip5	Ratio of avl_bld and Average avl_bld grouped by ZIP5
6	avl_bldvol_zip5	Ratio of avl_bldvol and Average avl_bldvol grouped by ZIP5
7	avt_lot_zip5	Ratio of avt_lot and Average avt_lot grouped by ZIP5

8	avt_bld_zip5	Ratio of avt_bld and Average avt_bld grouped by ZIP5
9	avt_bldvol_zip5	Ratio of avt_bldvol and Average avt_bldvol grouped by ZIP5
10	full_lot_zip3	Ratio of full_lot and Average full_lot grouped by ZIP3
11	full_bld_zip3	Ratio of full_bld and Average full_bld grouped by ZIP3
12	full_bldvol_zip3	Ratio of full_bldvol and Average full_bldvol grouped by ZIP3
13	avl_lot_zip3	Ratio of avl_lot and Average avl_lot grouped by ZIP3
14	avl_bld_zip3	Ratio of avl_bld and Average avl_bld grouped by ZIP3
15	avl_bldvol_zip3	Ratio of avl_bldvol and Average avl_bldvol grouped by ZIP3
16	avt_lot_zip3	Ratio of avt_lot and Average avt_lot grouped by ZIP3
17	avt_bld_zip3	Ratio of avt_bld and Average avt_bld grouped by ZIP3
18	avt_bldvol_zip3	Ratio of avt_bldvol and Average avt_bldvol grouped by ZIP3
19	full_lot_taxclass	Ratio of full_lot and Average full_lot grouped by TAXCLASS
20	full_bld_taxclass	Ratio of full_bld and Average full_bld grouped by TAXCLASS
21	full_bldvol_ taxclass	Ratio of full_bldvol and Average full_bldvol grouped by TAXCLASS
22	avl_lot_ taxclass	Ratio of avl_lot and Average avl_lot grouped by TAXCLASS
23	avl_bld_ taxclass	Ratio of avl_bld and Average avl_bld grouped by TAXCLASS
24	avl_bldvol_ taxclass	Ratio of avl_bldvol and Average avl_bldvol grouped by TAXCLASS
25	avt_lot_ taxclass	Ratio of avt_lot and Average avt_lot grouped by TAXCLASS
26	avt_bld_ taxclass	Ratio of avt_bld and Average avt_bld grouped by TAXCLASS
27	avt_bldvol_ taxclass	Ratio of avt_bldvol and Average avt_bldvol grouped by TAXCLASS
28	full_lot_b	Ratio of full_lot and Average full_lot grouped by BOROUGH
29	full_bld_ b	Ratio of full_bld and Average full_bld grouped by BOROUGH
30	full_bldvol_ b	Ratio of full_bldvol and Average full_bldvol grouped by BOROUGH
31	avl_lot_ b	Ratio of avl_lot and Average avl_lot grouped by BOROUGH
32	avl_bld_b	Ratio of avl_bld and Average avl_bld grouped by BOROUGH
33	avl_bldvol_ b	Ratio of avl_bldvol and Average avl_bldvol grouped by BOROUGH

34	avt_lot_ b	Ratio of avt_lot and Average avt_lot grouped by BOROUGH
35	avt_bld_ b	Ratio of avt_bld and Average avt_bld grouped by BOROUGH
36	avt_bldvol_ b	Ratio of avt_bldvol and Average avt_bldvol grouped by BOROUGH
37	full_lot_all	Ratio of full_lot and Average full_lot grouped by all recordings
38	full_bld_ all	Ratio of full_bld and Average full_bld grouped by all recordings
39	full_bldvol_ all	Ratio of full_bldvol and Average full_bldvol of all recordings
40	avl_lot_ all	Ratio of avl_lot and Average avl_lot of all recordings
41	avl_bld_ all	Ratio of avl_bld and Average avl_bld of all recordings
42	avl_bldvol_ all	Ratio of avl_bldvol and Average avl_bldvol of all recordings
43	avt_lot_ all	Ratio of avt_lot and Average avt_lot of all recordings
44	avt_bld_ all	Ratio of avt_bld and Average avt_bld of all recordings
45	avt_bldvol_ all	Ratio of avt_bldvol and Average avt_bldvol of all recordings

4. Dimensionality Reduction

After creating a list of 45 variables, the next step is to reduce the dimensionality of the variables using Principal Component Analysis (PCA).

Principal Component Analysis (PCA)

PCA is mathematically defined as an orthogonal linear transformation, transforming the data to a new coordinate system, by finding the dominant directions in the data and rotating the coordinate system along these directions. Therefore, PCA is a great method for reducing the number of variables needed while retaining necessary variance for analysis.

Steps for dimensionality reduction:

- **Step 1: Z-scale**

  Before implementing PCA, all the variables should be normalized (with the mean of 0 and standard deviation of 1), which is also called z-scale. Z-scale can help to make each original variable have the same scale, avoiding the great difference of number's measurement scale in each variable.

  Scale() function in R is used for the z-scale, and the specific usage is:

  Scale(dataframe, center = TRUE, scale = TRUE)

The “center = TRUE” is to make sure that the mean of variable is 0, and “scale = TRUE” is used to keep each variable having a unit variance.

- **Step 2: PCA**

Since the variables are already normalized, we use the **prcomp()** function from the factoextra package in R to implement PCA for the dataframe. And the summary of importance of components is as follows:

PC	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	4.748	0.501	0.501
PC2	3.1977	0.2272	0.7282
PC3	1.80304	0.07224	0.80045
PC4	1.53102	0.05209	0.85254
PC5	1.40893	0.04411	0.89665
PC6	1.34928	0.04046	0.93771
PC7	0.86392	0.01659	0.95370
PC8	0.66332	0.00978	0.96347
PC9	0.60228	0.00806	0.97153
PC10	0.48004	0.00512	0.97665
PC11	0.4695	0.0049	0.9816
PC12	0.44893	0.00448	0.98603
PC13	0.37784	0.00317	0.98920
PC14	0.32382	0.00233	0.99151
PC15	0.29696	0.00196	0.99349

Table 5-1 Summary of importance of components

And the screen plot pictured by fviz_eig() function in R for the PCA result is:

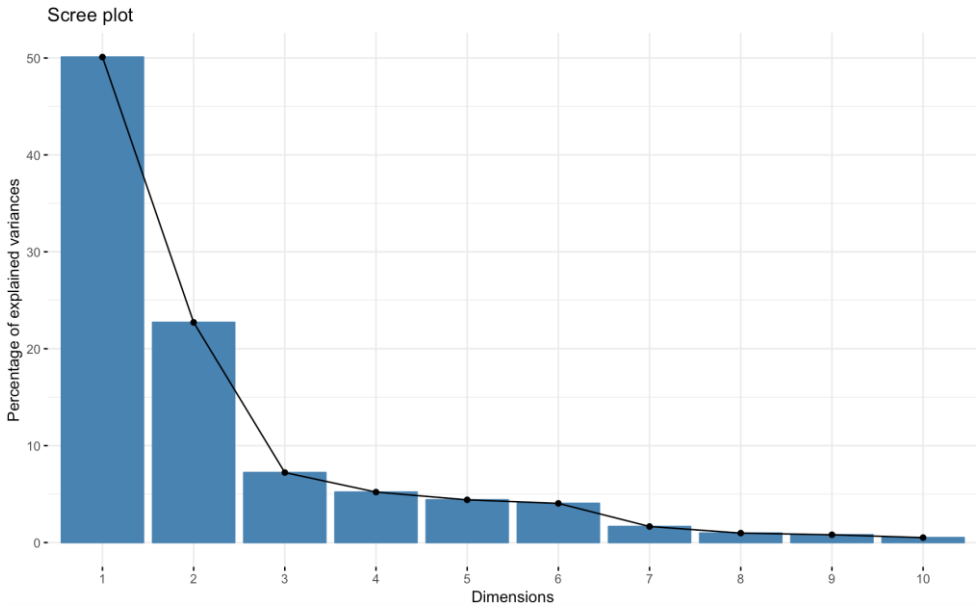


Figure 5-1: Screen plot of PCA result

From the table and chart, we could see that PC1 – PC7 contain **more than 95%** variance of all variables, so PC1- PC7 were chosen as the final outcomes of dimensionality reduction.

- **Step 3: Z-scale**

  After implementing PCA, we have already got the 7 important PCs. Then the PCA result should be normalized, to make each new variable having the mean of 0 and the same standard deviation of 1. And the variables will be used for scoring in the next step.

5. Algorithms

We used two ways to build an unsupervised fraud algorithm. One is used a Mahalanobis-like distance, and the other one is used an autoencoder.

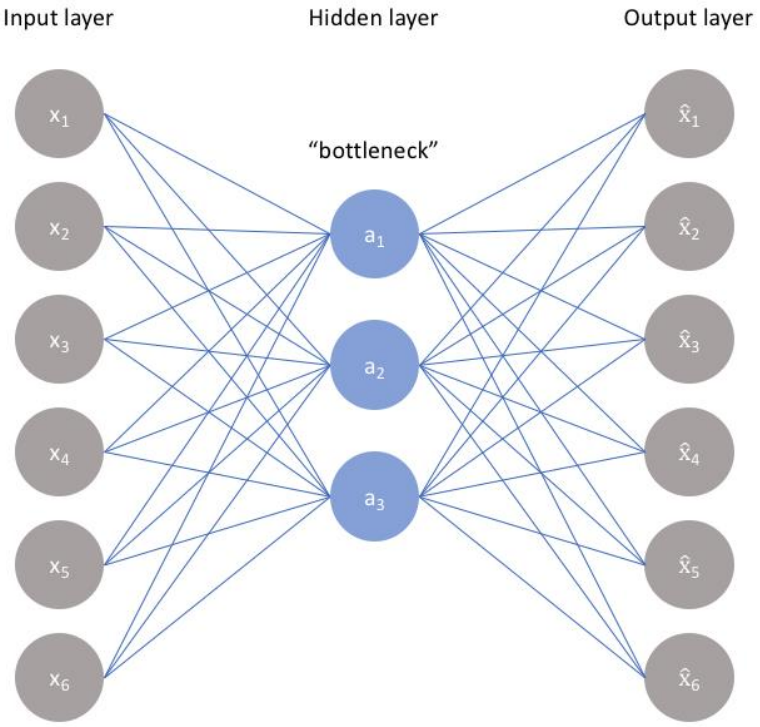
- **Heuristic algorithm**

After second times z-scaling, the variables were applied by heuristic algorithm, and the fraud score was calculated as following formula

$$S = \left(\sum_i |z_i|^n\right)^{\frac{1}{n}}$$

- **Autoencoder**

The purpose of autoencoder is to learn a representation for a set of data by training the network. The autoencoder tries to generate from the reduced encoding a representation as close as possible to its original input, as illustration has been shown below.



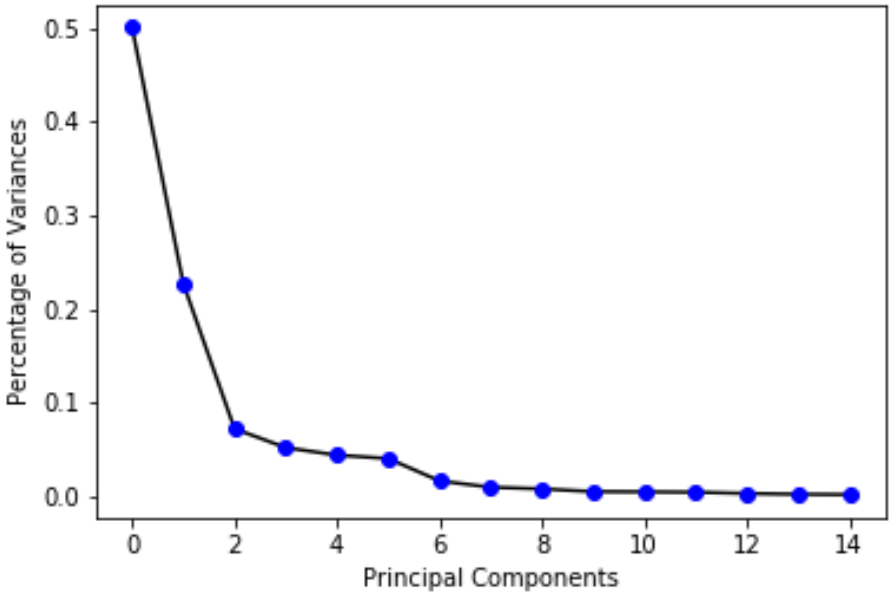
The mathematical expression for calculating autoencoder fraud scores has shown below

$$s = \left(\sum_i |z_i - z_i'|^n\right)^{\frac{1}{n}}$$

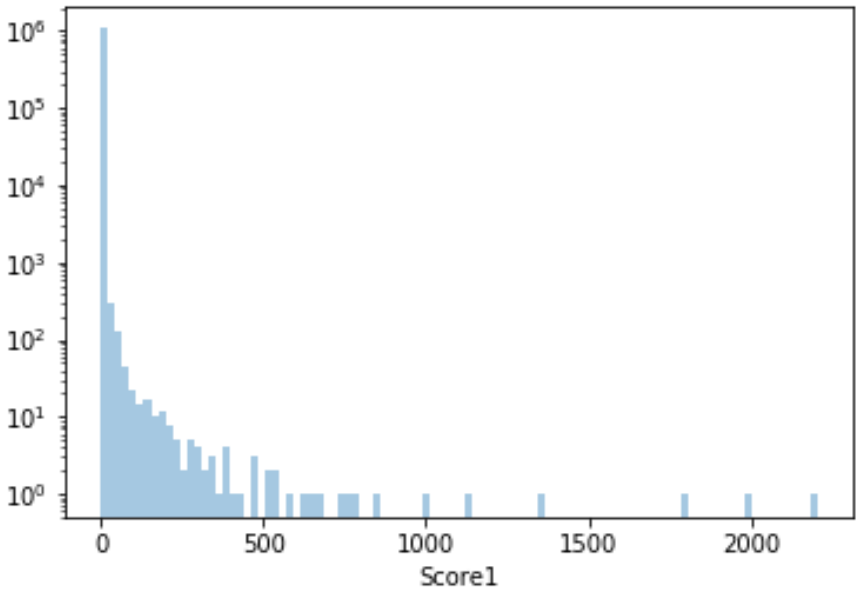
After we finalized the two scores, one was from heuristic algorithms, and the other was from autoencoder, we applied quantile binning on both scores based on those ascending scores of magnitudes. The quantile binning is a way to assign scores to scale equally. We assigned new ranking number based on each score. Because we would want to have more sensitive fraud score prediction, the fraud score was assigned via minimum ranking between heuristic score and autoencoder score.

6. Results and Observations

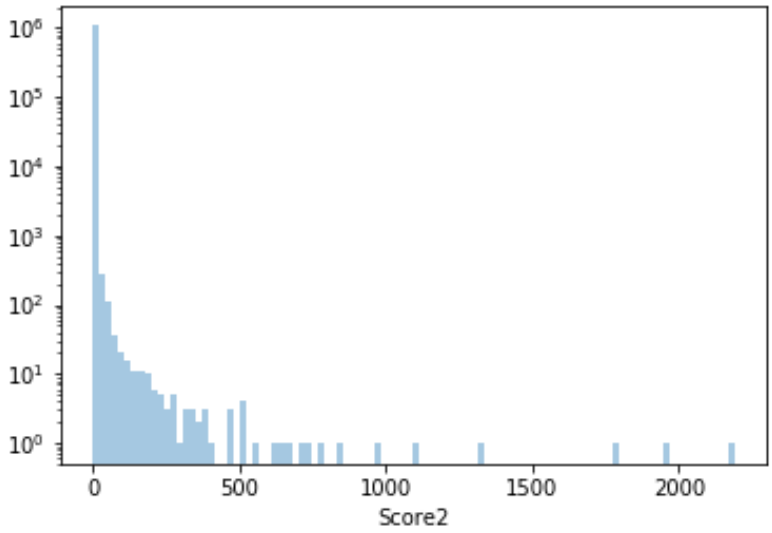
Starting with over 45 variables, we reduced the number of variables to 7 through PCA (dimensionality reduction). These variables explained more than 95.37% of the variance in the data. A plot of number of principal components against cumulative variance explained (courtesy of PCA) has been shown below.



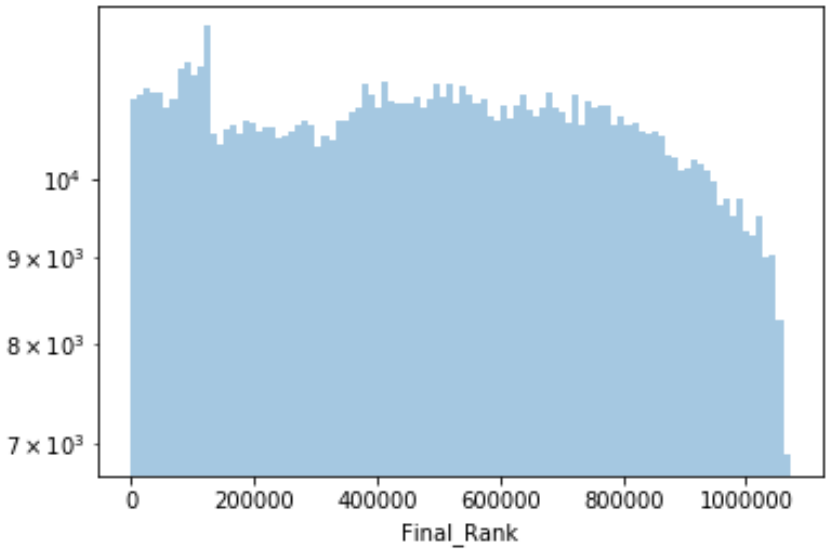
Fraud score calculated based on outlier detection via z-scores algorithm exhibited a right-skewed distribution (refer to the illustrations below). As expected, most of the records had low fraud scores and there were few outliers relative to the number of total records.



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



The final score :



We manually examined the records with high fraud scores. We highlight the government properties and park& recreations and original missing value because although these properties are anomalies on several fronts, we know that government properties, parks and universities have very low risk of tax fraud. As for the missing value, we highlight the instances since the rules we applied is not sophisticated enough and may need further confirmation.

The top candidates for potential tax fraud have been listed below.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
1	632815	864163 REALTY, LLC	1	157	95	2,930,000	1,318,500	1,318,500	1	1
2	565391	U S GOVERNMENT OWNRD	3	117	108	4,326,303,700	1,946,836,665	1,946,836,665	25	50
3	917941	LOGAN PROPERTY, INC.	3	4910	100	374,019,883	1,792,808,947	4,668,308,947	24	36
4	1067359		2	1	1	836,000	28,800	50,160	36	45
5	750815	M FLAUM	2	1	1	472,000	14,922	24,664	20	36
6	230595		2	1	1	380,000	12,938	20,874	20	40
7	585117	NEW YORK CITY ECONOMI	20	298	402	3,443,400	1,549,530	1,549,530	1	1
8	585438	11-01 43RD AVENUE REA	10	94	165	3,712,000	252,000	1,670,400	1	1
9	691878	PARKS AND RECREATION	1	22	47	101,700,000	44,550,000	45,765,000	27	47
10	556608	PARKS AND RECREATION	1	35	50	136,000,000	60,750,000	61,200,000	88	62
11	690832	PARKS AND RECREATION	3	610	534	242,000,000	103,500,000	108,900,000	20	20
12	67128	CULTURAL AFFAIRS	5	840	102	6,150,000,000	2,668,500,000	2,767,500,000	26	80
13	920627	PLUCHENIK, YAAKOV	2	91	100	1,900,000	9,763	75,763	1	1
14	565397	DEPT OF GENERAL SERVI	3	466	1009	2,310,884,200	1,039,897,890	1,039,897,890	25	50
15	585119		20	139	342	2,151,600	968,220	968,220	1	1
16	770593	OH, LAURA E	1	1	1	251,989	1,001	8,934	25	38
17	248664	PARKS AND RECREATION	6	600	4000	190,000,000	79,200,000	85,500,000	21	48
18	1067000	DRANOVSKY, VLADIMIR	3	96	279	2,120,000	65,401	124,910	1	1
19	85885	PARKS AND RECREATION	1	4000	150	70,214,000	31,455,000	31,596,300	8	8
20	153069	CITY OF NEW YORK	6	999	999	1,021,000,000	83,700,000	459,450,000	20	44

Clearly, for some of these records, the full value is exceptionally high. Further scrutiny reveals that such records have no information about lot front, lot depth, etc.

For some other records, the full value of the property per build area is either excessively high or low. Since, a lot of these properties are owned by real estate firms, we can infer that either the properties owned by them are significantly different from an average property or they might be exploiting loopholes (and/or committing potential tax fraud) in the property tax law.

- **Top 10 Candidates for Potential Fraud**

Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
632815	864163 REALTY, LLC	1	157	95	2,930,000	1,318,500	1,318,500	1	1

For this property, the BLDFRONT and the BLDDEPTH are just 1, which is so low and definitely needs further investigation.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
2	565391	U S GOVERNMENT OWNRD	3	117	108	4,326,303,700	1,946,836,665	1,946,836,665	25	50

For this property, the FULLVAL, the AVLAND and the AVTOT are so high, compared to the average. Since the owner is US GOVERNMENT, it may make sense.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
3	917941	LOGAN PROPERTY, INC.	3	4910	100	374,019,883	1,792,808,947	4,668,308,947	24	36

For this property, the FULLVAL, the AVLAND and the AVTOT are so high, compared to the average.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
4	1067359		2	1	1	836,000	28,800	50,160	36	45

For this property, the LTFRONT and the LTDEPTH are just 1, which is so low and definitely needs further investigation.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
5	750815	M FLAUM	2	1	1	472,000	14,922	24,664	20	36

For this property, the LTFRONT and the LTDEPTH are just 1, which is so low and definitely needs further investigation.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
6	230595		2	1	1	380,000	12,938	20,874	20	40

For this property, the LTFRONT and the LTDEPTH are just 1, which is so low and definitely needs further investigation.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
7	585117	NEW YORK CITY ECONOMI	20	298	402	3,443,400	1,549,530	1,549,530	1	1

For this property, the BLDFRONT and the BLDDEPTH are just 1, which is so low and definitely needs further investigation.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
8	585438	11-01 43RD AVENUE REA	10	94	165	3,712,000	252,000	1,670,400	1	1

For this property, the BLDFRONT and the BLDDEPTH are just 1, which is so low and definitely needs further investigation.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
9	691878	PARKS AND RECREATION	1	22	47	101,700,000	44,550,000	45,765,000	27	47

For this property, the FULLVAL, the AVLAND and the AVTOT are so high, compared to the average. Since the owner is PARKS AND RECREATION, it may make sense.

	Index	OWNER	STORIES	LTFRONT	LTDEPTH	FULLVAL	AVLAND	AVTOT	BLDFRONT	BLDDEPTH
10	556608	PARKS AND RECREATION	1	35	50	136,000,000	60,750,000	61,200,000	88	62

For this property, the FULLVAL, the AVLAND and the AVTOT are so high, compared to the average. Since the owner is PARKS AND RECREATION, it may make sense.

7. Conclusion

After filling the missing value by aggregation, we normalize the 3 values by the 3 sizes and group by 5 variables to build 45 expert variables.

The purpose behind building such variables was to identify the ones which are strong predictors of fraud, by narrowing down to a smaller group of variables. Next, we z-scaled all variables for feature selection and dimensionality reduction. Principal Components Analysis (PCA) was then performed on those 45 expert variables and 7 PCs was retained for further analysis. This way, we were able to lose dimensionality significantly without sacrificing variance explained. The 7 PCs were then z-scaled again.

Finally, we used a heuristic algorithm and autoencoder to calculate two fraud scores and combined them to be a weighted fraud score. A further feature analysis of the top 10 suspicious fraud record gives us a deeper understanding of when to ring the alarm with future property data.

- **Scope for Improvement**

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

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

Appendix: Data Quality Report

Section1: Basic information of dataset

This dataset represents key attributes of 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..

Section2: Introduction of dataset

1.dimension of dataset

Rows:1070994

Columns:32

2.names of fields

```
[1] "RECORD"   "BBLE"     "B"        "BLOCK"    "LOT"      "EASEMENT"
[7] "OWNER"    "BLDGCL"   "TAXCLASS" "LTFRONT"  "LTDEPTH"  "EXT"
[13] "STORIES"  "FULLVAL"  "AVLAND"   "AVTOT"    "EXLAND"   "EXTOT"
[19] "EXCD1"    "STADDR"   "ZIP"      "EXMPTCL"  "BLDFRONT" "BLDDEPTH"
[25] "AVLAND2"  "AVTOT2"   "EXLAND2"  "EXTOT2"   "EXCD2"    "PERIOD"
[31] "YEAR"     "VALTYPE"
```

3.brief introduction of each field

```
## 'data.frame':    1070994 obs. of  32 variables:
##  $ RECORD  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ BBLE    : Factor w/ 1070994 levels "1000010101","1000010201",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ B       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ BLOCK   : int  1 1 2 2 3 3 3 3 4 4 ...
##  $ LOT     : int  101 201 1 23 1 2 3 10 1001 1002 ...
##  $ EASEMENT: Factor w/ 13 levels "","E","F","G",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ OWNER   : Factor w/ 863348 levels "","\"54 SATELLITE CO\"",..: 796662 796662 219060 218905…
##  $ BLDGCL  : Factor w/ 200 levels "A0","A1","A2",..: 121 200 188 152 124 124 124 200 138 138 ...
##  $ TAXCLASS: Factor w/ 11 levels "1","1A","1B",..: 11 11 11 11 11 11 11 11 11 11 ...
##  $ LTFRONT : int  500 27 709 793 323 496 180 362 0 0 ...
##  $ LTDEPTH : int  1046 0 564 551 1260 76 370 177 0 0 ...
##  $ EXT     : Factor w/ 4 levels "","E","EG","G": 1 1 2 1 1 1 1 1 1 1 ...
##  $ STORIES : num  NA NA 3 2 1 NA 1 3 50 50 ...
##  $ FULLVAL : num  2.14e+07 1.94e+08 1.05e+08 3.92e+07 2.72e+08 ...
##  $ AVLAND  : num  4.23e+06 1.43e+07 3.90e+07 1.53e+07 1.21e+08 ...
##  $ AVTOT   : num  9.63e+06 8.72e+07 4.71e+07 1.76e+07 1.23e+08 ...
##  $ EXLAND  : num  4.23e+06 1.43e+07 3.90e+07 1.53e+07 1.21e+08 ...
##  $ EXTOT   : num  9.63e+06 8.72e+07 4.71e+07 1.76e+07 1.23e+08 ...
##  $ EXCD1   : num  4600 4600 2191 2191 2231 ...
##  $ STADDR  : Factor w/ 839281 levels "","* SOUTH PORTLAND AVE",..: 346 237 837429 837948 …
```

```
##  $ ZIP     : num  10004 10004 10004 10004 10004 ...
##  $ EXMPTCL : Factor w/ 15 levels "","5","A9","KI",..: 9 9 7 7 7 7 7 9 1 1 ...
##  $ BLDFRONT: int  0 0 709 85 89 0 16 37 0 0 ...
##  $ BLDDEPTH: int  0 0 564 551 57 0 19 227 0 0 ...
##  $ AVLAND2 : num  3.78e+06 1.11e+07 3.23e+07 1.36e+07 1.06e+08 ...
##  $ AVTOT2  : num  8.61e+06 8.07e+07 4.02e+07 1.58e+07 1.08e+08 ...
##  $ EXLAND2 : num  3.78e+06 1.11e+07 3.23e+07 1.36e+07 1.06e+08 ...
##  $ EXTOT2  : num  8.61e+06 8.07e+07 4.02e+07 1.58e+07 1.08e+08 ...
##  $ EXCD2   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ PERIOD  : Factor w/ 1 level "FINAL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ YEAR    : Factor w/ 1 level "2010/11": 1 1 1 1 1 1 1 1 1 1 ...
##  $ VALTYPE : Factor w/ 1 level "AC-TR": 1 1 1 1 1 1 1 1 1 1 ...
```

4.summary of numerical variables

Field	#values	%populated	#unique values	#zeros	mean	median	min	max	standard deviation
LTFRONT	1070994	100	1297	169108	36.63	25	0	9999	74.03284
LTDEPTH	1070994	100	1370	170128	88.86	100	0	9999	76.39628
FULLVAL	1070994	100	109324	13007	8.74E+05	4.47E+05	0	6.15E+09	11582431
AVLAND	1070994	100	70921	13009	8.51E+04	1.37E+04	0	2.67E+09	4057260
AVTOT	1070994	100	112914	13007	4.67E+09	2.53E+04	0	4.67E+09	6877529
EXLAND	1070994	100	33419	491699	3.64E+04	1.62E+03	0	2.67E+09	3981576
EXTOT	1070994	100	64255	432572	9.12E+04	1.62E+03	0	4.67E+09	6508403
EXCD1	638388	59.61639	130	0	1602	1017	1010	7170	1384.227
BLDFRONT	1070994	100	612	28815	23.04	20	0	7575	35.5797
BLDDEPTH	1070994	100	621	228853	39.92	39	0	9393	42.70715
AVLAND2	282726	26.39847	58592	0	2.46E+05	2.01E+04	3.00E+00	2.37E+09	6178963
AVTOT2	282732	26.39903	111361	0	4.50E+09	8.00E+04	3.00E+00	4.50E+09	11652529
EXLAND2	87449	8.165218	22196	0	3.51E+05	3.05E+03	1.00E+00	2.37E+09	10802213
EXTOT2	130828	12.21557	48349	0	6.57E+05	4.50E+09	7.00E+00	4.50E+09	16072510
EXCD2	92948	8.678667	61	0	1364	1017	1011	7160	1094.706

4.summary of categorical variables

Field	#values	%pupolated	#unique values
RECORD	1070994	100	1070994
BBLE	1070994	100	1070994
B	1070994	100	5
BLOCK	1070994	100	13984
LOT	1070994	100	6366
EASEMENT	4636	43.28689	13
OWNER	1039249	97.03593	863348
BLDGCL	1070994	100	200
TAXCLASS	1070994	100	11
EXT	354305	33.08188	4
STORIES	1014730	100	112
STADDR	1070318	99.93688	839281
ZIP	1041104	97.20913	197
EXMPTCL	15579	1.45463	15
PERIOD	1070994	100	1
YEAR	1070994	100	1
VALTYPE	1070994	100	1

Section3: Introduction of each field

Field1: RECORD

RECORD is the key of each record. Type: Categorical

```
##  int [1:1070994] 1 2 3 4 5 6 7 8 9 10 ...
```

Number of values: 1070994

Number of unique values: 1070994

%populated: 100

Field2: BBLE

Type: Categorical

```
##  Factor w/ 1070994 levels "1000010101","1000010201",..: 1 2 3 4 5 6 7 8 9 10 ...
```

Number of values: 1070994

Number of unique values: 1070994

%populated: 100

Field3: B (BORO CODE)

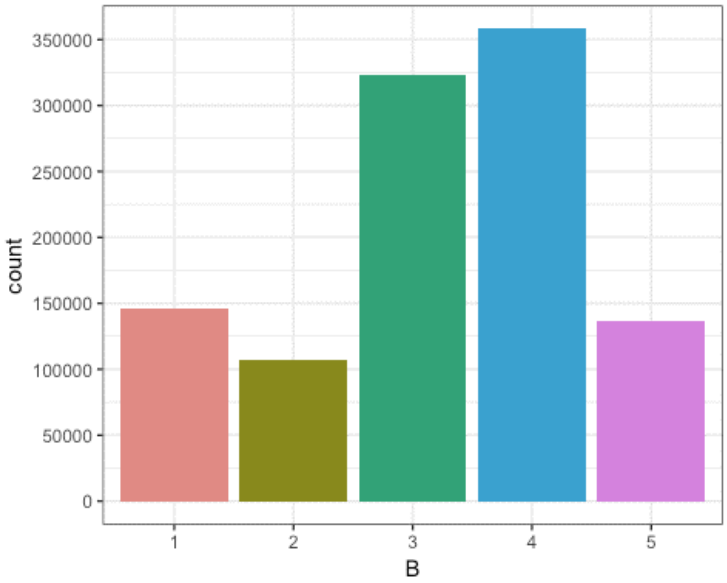
Type: Categorical

```
##  int [1:1070994] 1 1 1 1 1 1 1 1 1 1 ...
```

Since 1-5 has their meaning that: 1 - MANHATTAN, 2 - BRONX, 3 - BROOKLYN, 4 - QUEENS, 5 - STATEN ISLAND. So, the type of B should be categorical.

Number of values: 1070994

Number of unique values: 5 %populated: 100



From the chart we could see, in field “B”, 4(QUEENS) has the highest count.

Field4: Block

Type: Categorical

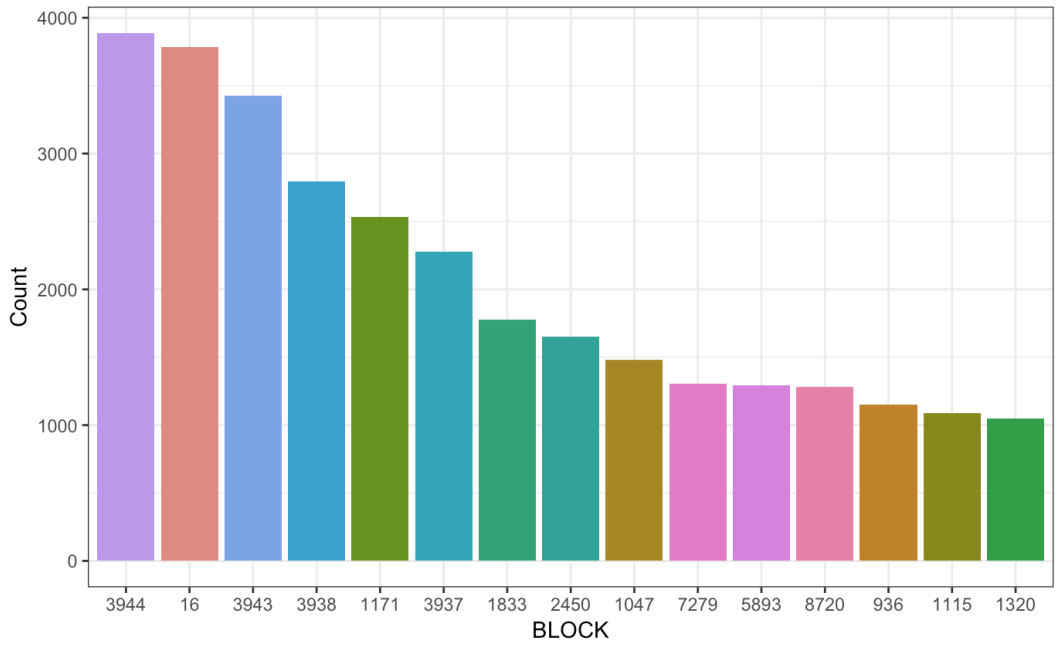
int [1:1070994] 1 1 2 2 3 3 3 3 4 4 ...

Number of values: 1070994

Number of unique values: 13984

%populated: 100

And the distribution of Block is as follows:



Field5: LOT

Type: Categorical

int [1:1070994] 101 201 1 23 1 2 3 10 1001 1002 ...

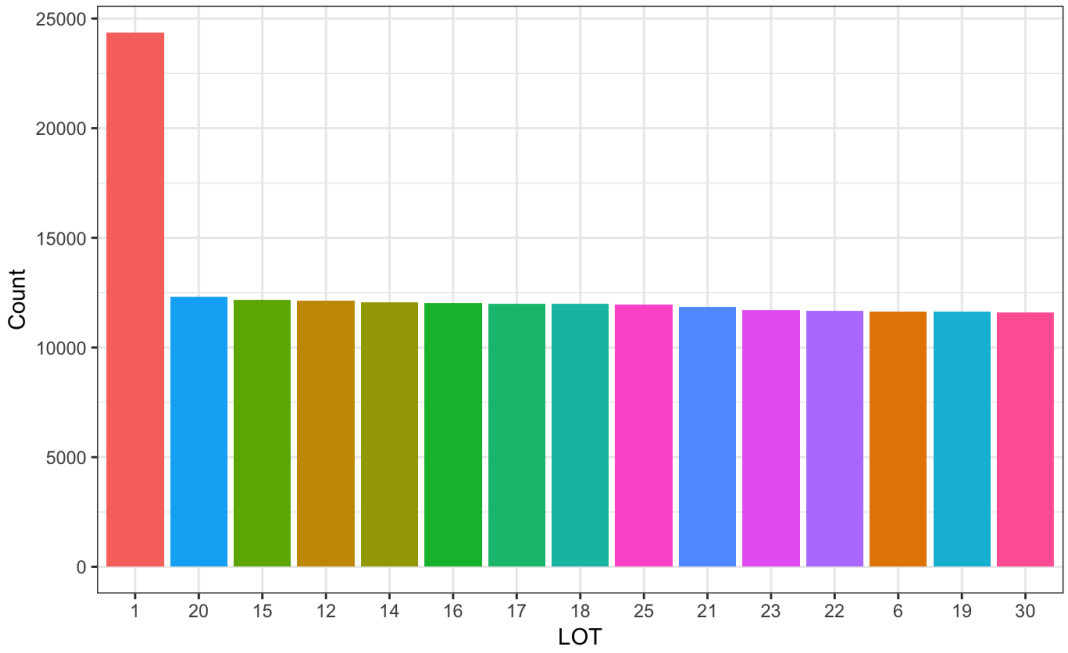
However, the type of LOT should be categorical.

Number of values: 1070994

Number of unique values: 6366

%populated: 100

And the distribution of LOT is as follows:



Field6: EASEMENT

Type: Categorical

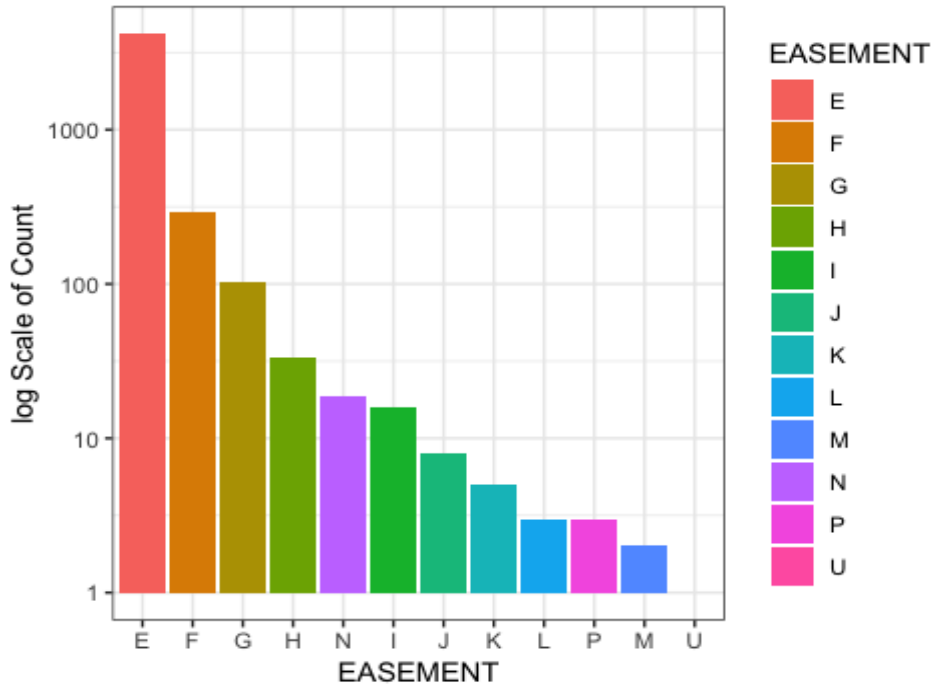
Factor w/ 13 levels "","E","F","G",..: 1 1 1 1 1 1 1 1 1 1 ...

Number of values: 4636

Number of unique values: 13

%populated: 43.28689

```
##    x freq
## 1  E 4148
## 2  F  296
## 3  G  102
## 4  H   33
## 5  I   16
## 6  J    8
## 7  K    5
## 8  L    3
## 9  M    2
## 10 N   19
## 11 P    3
## 12 U    1
```



'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.

From the chart above, "E" has the highest occupation.

Field7: OWNER

Type: Categorical

```
##  Factor w/ 863348 levels "","\"54 SATELLITE CO\"",..: 796662 796662 219060 218905 612642 612642 6
12642 219079 792991 792991 ...
```

Number of values: 1039249

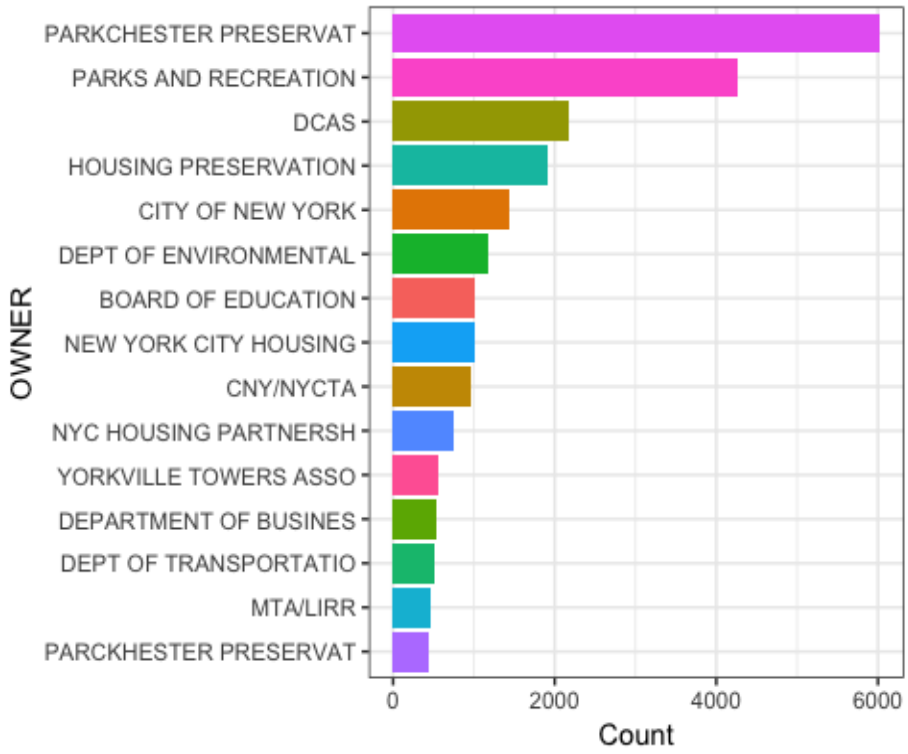
Number of unique values: 863348

%populated: 97.03593

Following is the first 10 owners with highest amount of property in NY:

```
##    OWNER                Count
##    <fct>                <int>
##  1 PARKCHESTER PRESERVAT  6021
##  2 PARKS AND RECREATION   4255
##  3 DCAS                   2169
##  4 HOUSING PRESERVATION   1904
##  5 CITY OF NEW YORK       1450
##  6 DEPT OF ENVIRONMENTAL  1166
##  7 BOARD OF EDUCATION     1015
##  8 NEW YORK CITY HOUSING  1014
##  9 CNY/NYCTA              975
```

10 NYC HOUSING PARTNERSH 747
... with 863,337 more rows



Parkchester Preservat owns the most property in NY.

Field8: BLDGCL

Type: Categorical

Factor w/ 200 levels "A0","A1","A2",..: 121 200 188 152 124 124 124 200 138 138 ...

Number of values: 1070994

Number of unique values: 200

%populated: 100

Following is the highest 10 amount of BLDGCL:

x   freq
1 R4 139879
2 A1 123369
3 A5 96984
4 B1 84208
5 B2 77598
6 C0 73111
7 B3 59240
8 A2 51130
9 A9 26177
10 B9 26133

Field9: TAXCLASS

Type: Categorical

```
##  Factor w/ 11 levels "1","1A","1B",..: 11 11 11 11 11 11 11 11 11 11 ...
```

Number of values: 1070994

Number of unique values: 11

%populated: 100

Following is the record amount of each TAXCLASS:

```
##       x     freq
## 1    1     660721
## 2    2     188612
## 3    4     104310
## 4   2A    40574
## 5   1B    24738
## 6   1A    21667
## 7   2B    13964
## 8   2C    10795
## 9   3      4638
## 10 1C    946
## 11 1D    29
```

We could tell that 1-3 unit residences takes the most occupancy.

Field10: LTFRONT

Type: Numerical

```
##  int [1:1070994] 500 27 709 793 323 496 180 362 0 0 ...
```

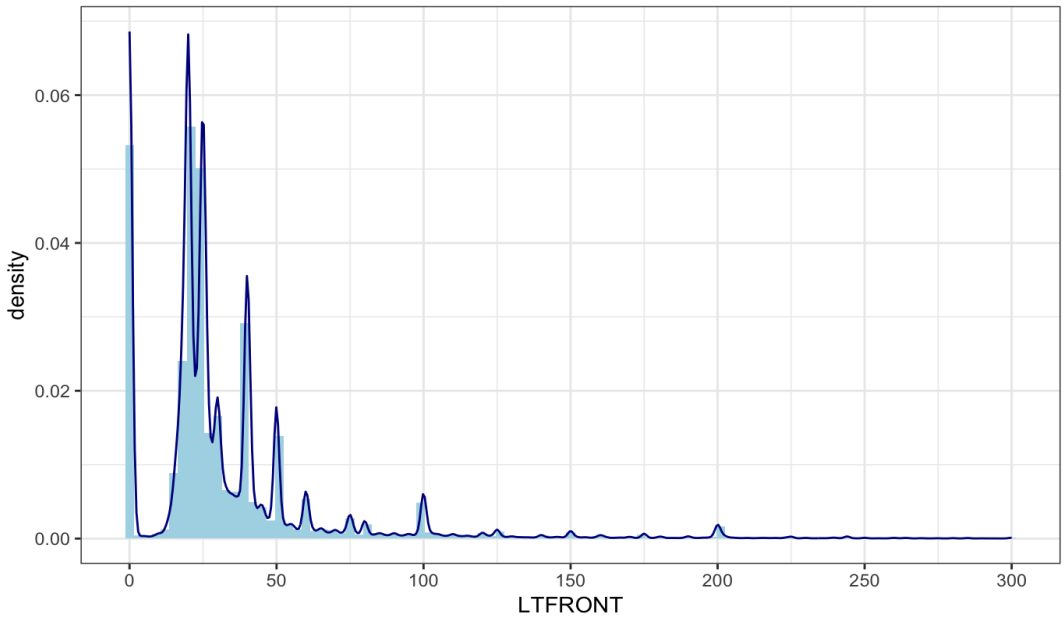
Number of values: 1070994

Number of unique values: 1297

%populated: 100

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   19.00   25.00       36.63       40.00  9999.00
```

Standard deviation: 74.03284



Percentage of records populated in the graph above: 99.29869

Most of the LTFRONT concentrate on 0-50, especially 0 and around 25.

Field11: LTDEPTH

Type: Numerical

```
##  int [1:1070994] 1046 0 564 551 1260 76 370 177 0 0 ...
```

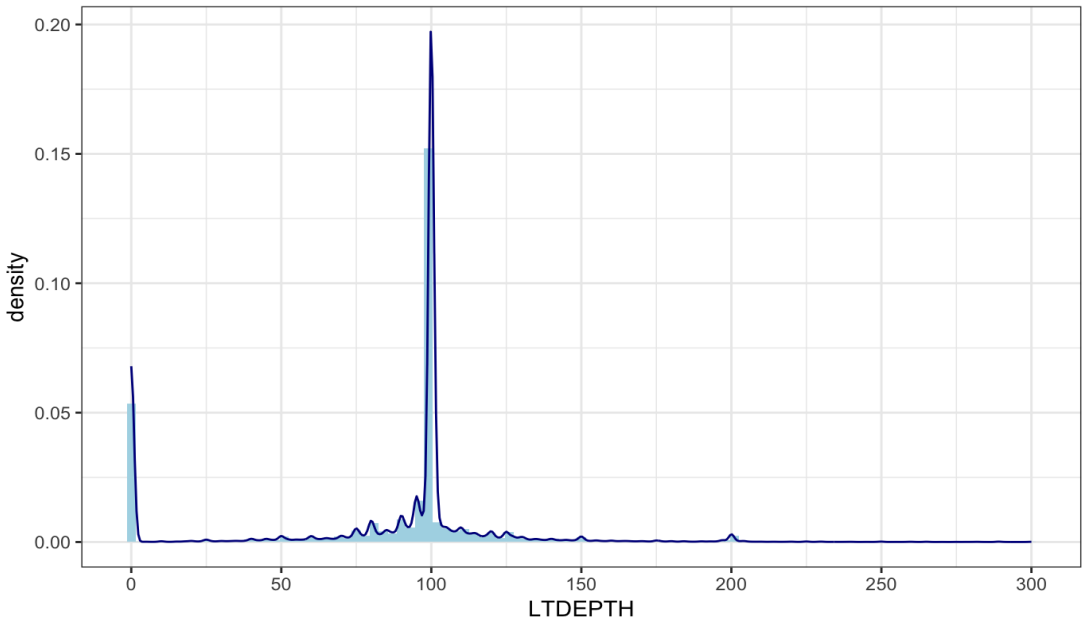
Number of values: 1070994

Number of unique values: 1370

%populated: 100

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   80.00  100.00   88.86  100.00 9999.00
```

Standard deviation: 76.39628



Percentage of records populated in the graph above: 99.1732

The LTPDEPTH of 100 has the highest occupancy.

Field12: EXT

Type: Categorical

Factor w/ 4 levels "","E","EG","G": 1 1 2 1 1 1 1 1 1 1 ...

Number of values: 354305

Number of unique values: 4

%populated: 33.08188

Following is the frequencies of every EXT:

```
##    x   freq
## 1  G 266970
## 2  E  49442
## 3 EG  37893
```

Field13: STORIES

Type: Categorical

num [1:1070994] NA NA 3 2 1 NA 1 3 50 50 ...

Number of values: 1014730

Number of unique values: 112

%populated: 94.74656

Following are the first 15 highest number of stories:

```
##       x   freq
## 1   2.0 415092
## 2   3.0 130127
## 3   1.0  96706
## 4   2.5  82292
## 5   4.0  38342
## 6   6.0  30936
## 7   5.0  25971
## 8   1.5  24770
## 9   2.7  13595
## 10 12.0  12198
## 11  8.0  11953
## 12  7.0  11899
## 13  1.6   8954
## 14  9.0   7343
## 15 13.0   7330
```

We could tell that 2 stories takes great part in all properties in NY.

Field14: FULLVAL

Type: Numerical

```
##  num [1:1070994] 2.14e+07 1.94e+08 1.05e+08 3.92e+07 2.72e+08 ...
```

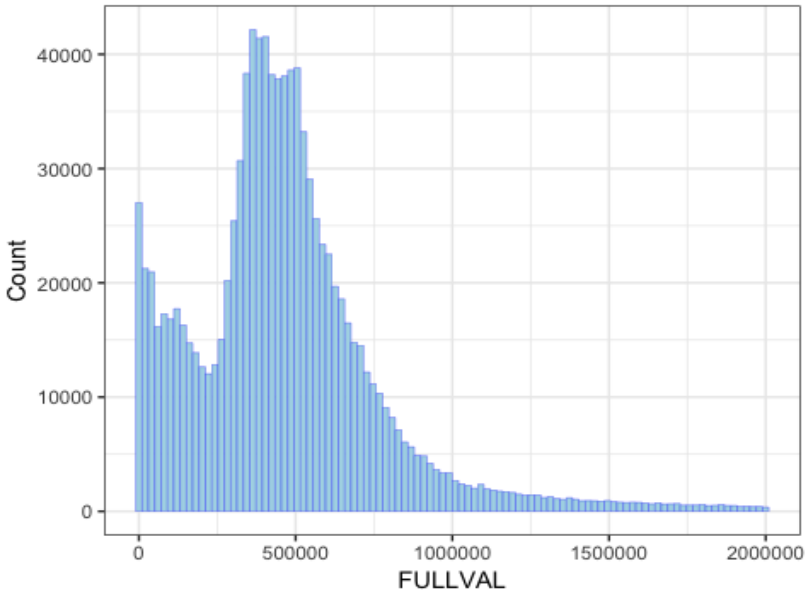
Number of values: 1070994

Number of unique values: 109324

%populated: 100

```
##      Min.       1st Qu.       Median       Mean          3rd Qu.       Max.
## 0.000e+00 3.040e+05 4.470e+05 8.743e+05 6.190e+05 6.150e+09
```

Standard deviation: 11582431



Percentage of records showed in the chart above is: 96.31221

From this chart, most of the market value of property is about 5000000.

Field15: AVLAND

Type: Numerical

num [1:1070994] 4.23e+06 1.43e+07 3.90e+07 1.53e+07 1.21e+08 ...

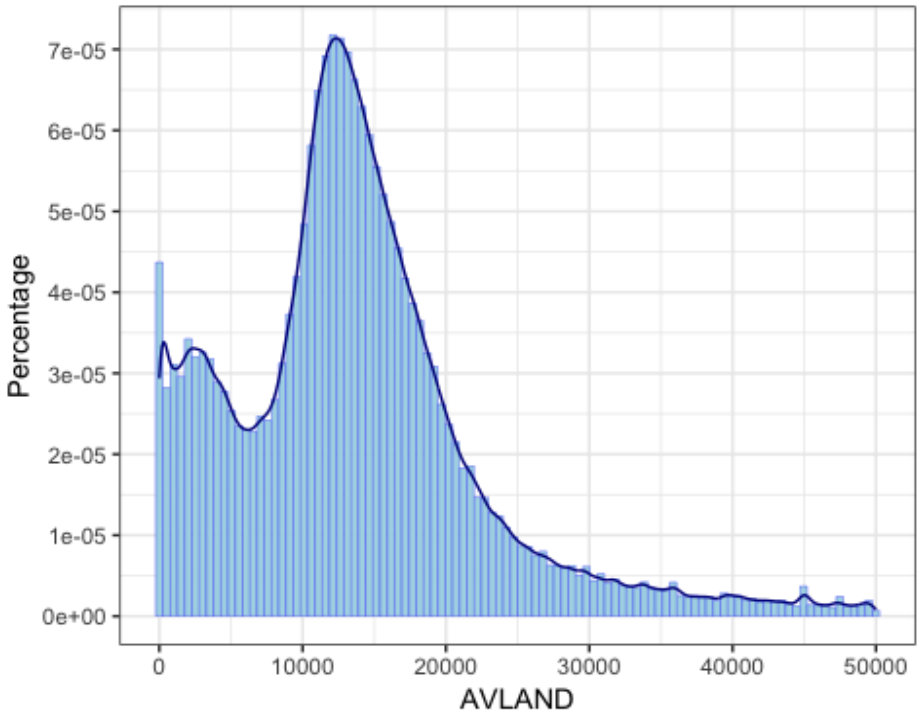
Number of values: 1070994

Number of unique values: 70921

%populated: 100

Min. 1st Qu. Median Mean 3rd Qu. Max.
0.000e+00 9.180e+03 1.368e+04 8.507e+04 1.974e+04 2.668e+09

Standard deviation: 4057260



Percentage of records showed in the chart above is: 90.52721

From the chart above, we could see that most of AVLAND concentrate on between 0 and 20000.

Field16: AVTOT

Type: Numerical

num [1:1070994] 9.63e+06 8.72e+07 4.71e+07 1.76e+07 1.23e+08 ...

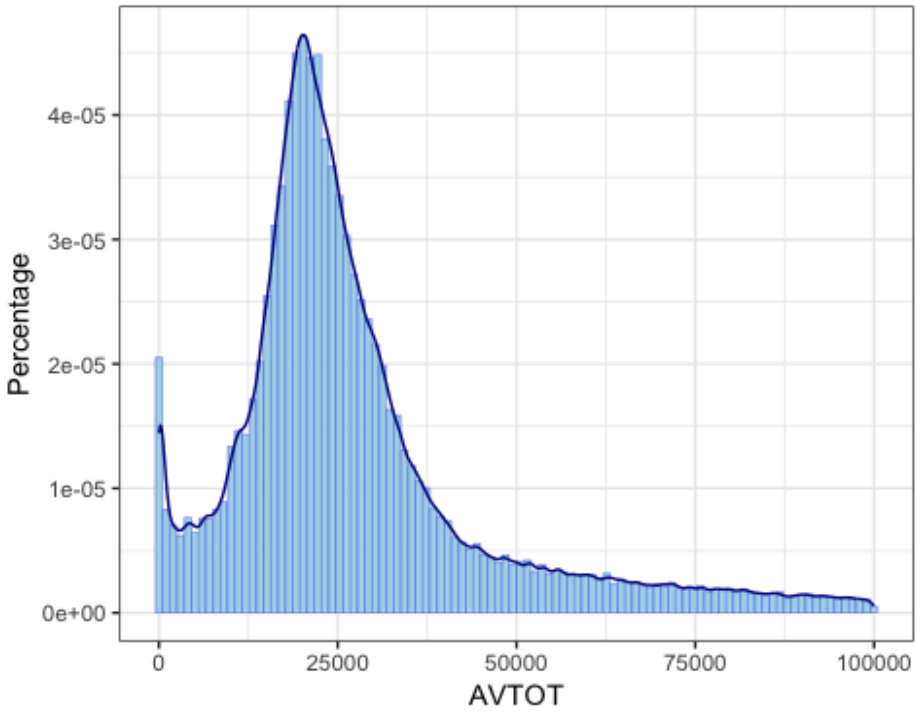
Number of values: 1070994

Number of unique values: 112914

%populated: 100

Min. 1st Qu. Median Mean 3rd Qu. Max.
0.000e+00 1.837e+04 2.534e+04 2.272e+05 4.544e+04 4.668e+09

Standard deviation:



Percentage of records showed in the chart above is: 86.0523

So, the AVTOT around 25000 takes great proportion of all properties in NY.

Field17: EXLAND

Type: Numerical

```
##  num [1:1070994] 4.23e+06 1.43e+07 3.90e+07 1.53e+07 1.21e+08 ...
```

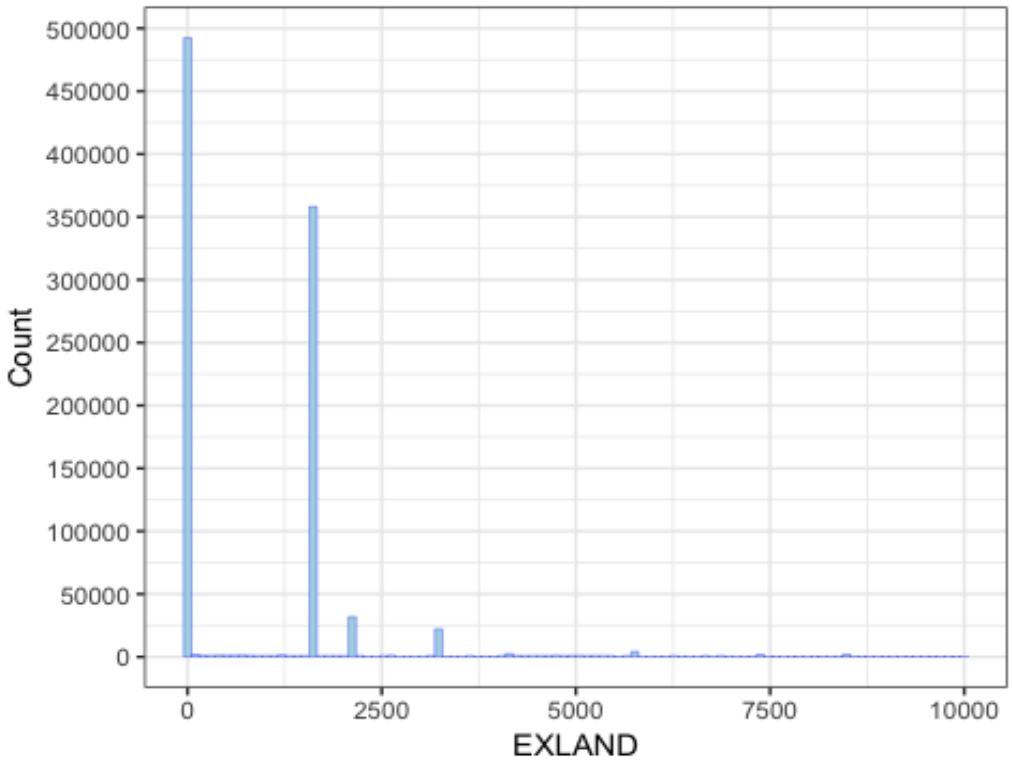
Number of values: 1070994

Number of unique values: 33419

%populated: 100

```
##      Min.       1st Qu.       Median       Mean         3rd Qu.      Max.
## 0.000e+00 0.000e+00 1.620e+03 3.642e+04 1.620e+03 2.668e+09
```

Standard deviation: 3981576



Percentage of records showed in the chart above is: 92.60771

We could tell that the peak of amount of EXLAND appears at 0 and nearly 2000.

Field18: EXTOT

Type: Numerical

num [1:1070994] 9.63e+06 8.72e+07 4.71e+07 1.76e+07 1.23e+08 ...

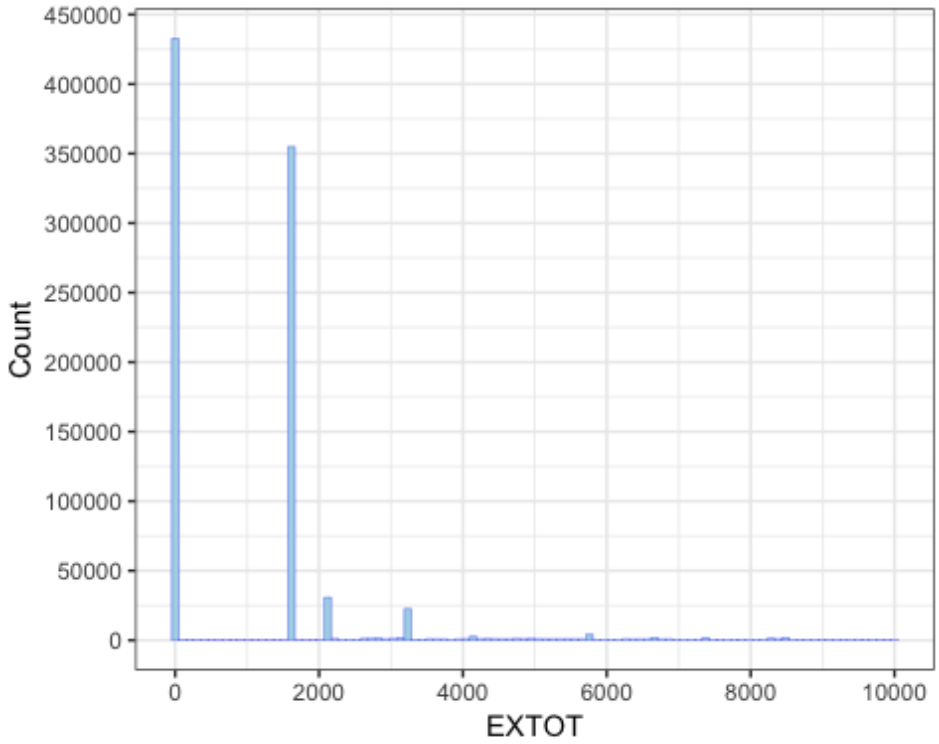
Number of values: 1070994

Number of unique values: 64255

%populated: 100

```
##      Min.      1st Qu.       Median       Mean       3rd Qu.      Max.
## 0.000e+00 0.000e+00 1.620e+03 9.119e+04 2.090e+03 4.668e+09
```

Standard deviation: 6508403



Percentage of records showed in the chart above is:85.33222

From the chart we could see that peak of amount of EXTOT appears at 0 and around 2000 and around 3000.

Field19: EXCD1

Type: Numerical

```
##  num [1:1070994] 4600 4600 2191 2191 2231 ...
```

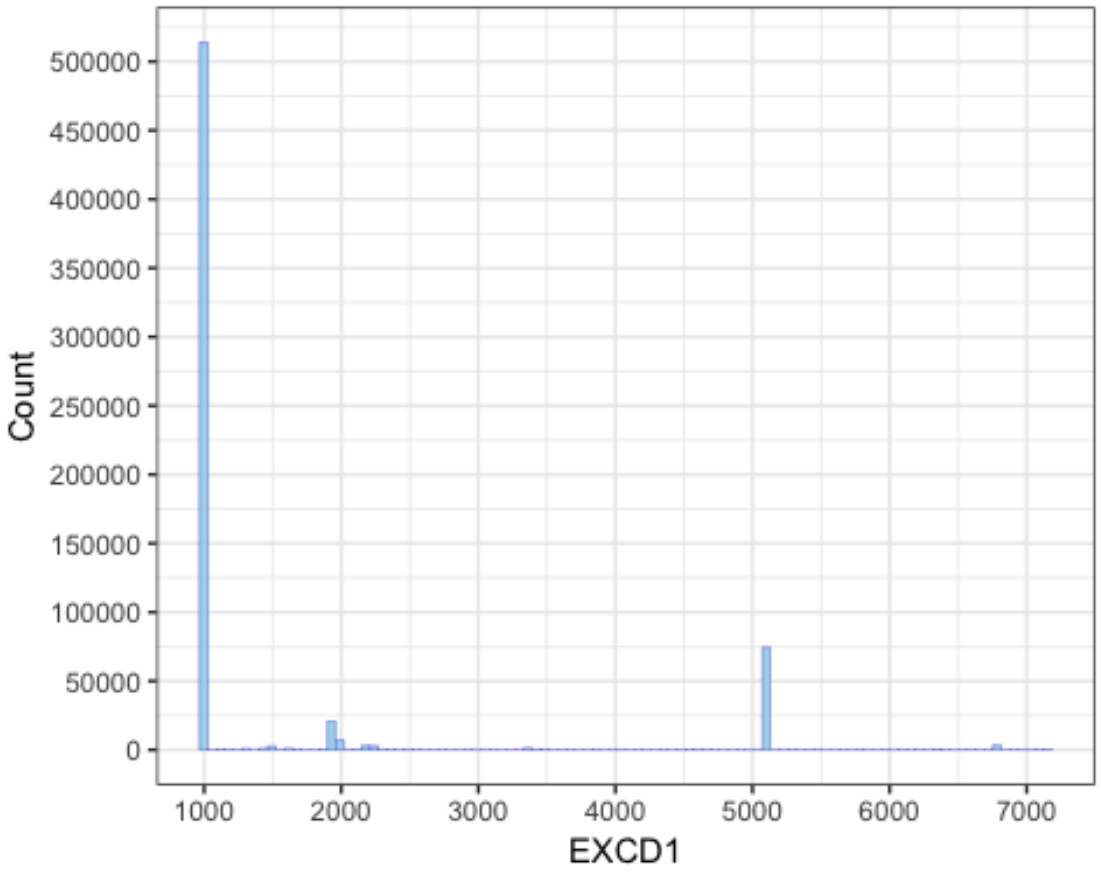
Number of values: 638488

Number of unique values: 130

%populated:.61639

```
##    Min.    1st Qu.  Median    Mean   3rd Qu.    Max.    NA's
##    1010    1017     1017      1602   1017       7170    432506
```

Standard deviation: 1384.227



Percentage of records showed in the chart above is: 59.61639

EXCD1 are most of 1000, nearly 2000 and nearly 5000.

Field20: STADDR

Type: Categorical

Factor w/ 839281 levels "","* SOUTH PORTLAND AVE",..: 346 237 837429 837948 835083 837926 83 7926 528 599 599 ...

Number of values: 1070318

Number of unique values: 839281

%populated: 99.93688

Following are the first 15 of highest amount of STADDR:

##	x	freq
## 1	501 SURF AVENUE	902
## 2	330 EAST 38 STREET	817
## 3	322 WEST 57 STREET	720
## 4	155 WEST 68 STREET	671
## 5	20 WEST 64 STREET	657
## 6	1 IRVING PLACE	650

```
## 7  220 RIVERSIDE BOULEVARD  628
## 8       360 FURMAN STREET       599
## 9      200 EAST 66 STREET.        585
## 10      30 WEST 63 STREET       562
## 11       2 BAY CLUB DRIVE         556
## 12     350 WEST 42 STREET       556
## 13      200 RECTOR PLACE        549
## 14     301 EAST 79 STREET.       538
## 15     350 WEST 50 STREET      498
```

Field21: ZIP

Type: Categorical

```
##  num [1:1070994] 10004 10004 10004 10004 10004 ...
```

Number of values: 1041104

Number of unique values: 197

%populated: 97.20913

This table shows the first 15 of highest amount of zip code:

```
##       x  freq
## 1  10314 24606
## 2  11234 20001
## 3  10312 18127
## 4  10462 16905
## 5  10306 16578
## 6  11236 15678
## 7  11385 14921
## 8  11229 12793
## 9  11211 12710
## 10 11207 12293
## 11 11215 11834
## 12 11235 11312
## 13 11203 11241
## 14 11208 11139
## 15 11204 11061
```

Field22: EXMPTCL

Type: Categorical

```
##  Factor w/ 15 levels "","5","A9","KI",..: 9 9 7 7 7 7 7 9 1 1 ...
```

Number of values: 15579

Number of unique values: 15

%populated: 1.45463

This table states the amount of each exmptcl of property in New York, and X1 and X5 take great part:

```
##     x freq
## 1  X1 6912
## 2  X5 5208
## 3  X7  820
## 4  X2  770
## 5  X6  764
## 6  X4  441
## 7  X8  292
## 8  X3  259
## 9  X9  108
## 10  5    1
## 11 A9    1
## 12 KI    1
## 13 R4    1
## 14 VI    1
```

Field23: BLDFRONT

Type:

```
##  int [1:1070994] 0 0 709 85 89 0 16 37 0 0 ...
```

Number of values:

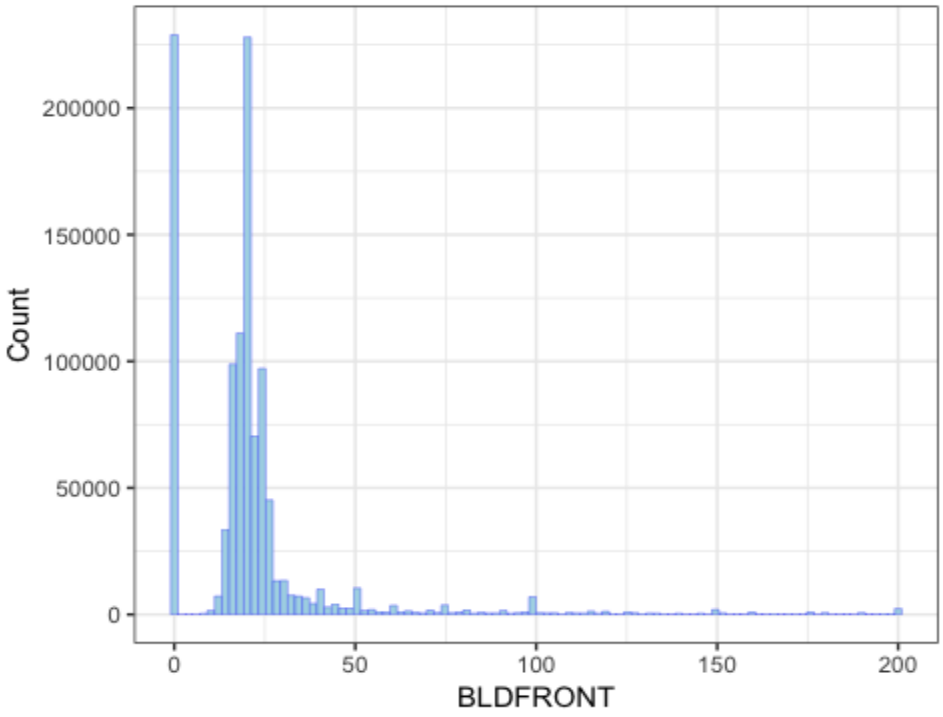
```
##   num_of_values
## 1       1070994
```

Number of unique values: 612

%populated: 100

```
##    Min. 1st Qu.  Median    Mean     3rd Qu.    Max.
##    0.00   15.00   20.00    23.04     24.00     7575.00
```

Standard deviation: 35.5797



Percentage of records showed in the chart above is: 99.55742

The most of BLDFRONT are 0 or around 25.

Field24: BLDDEPTH

Type: Numerical

```
##  int [1:1070994] 0 0 564 551 57 0 19 227 0 0 ...
```

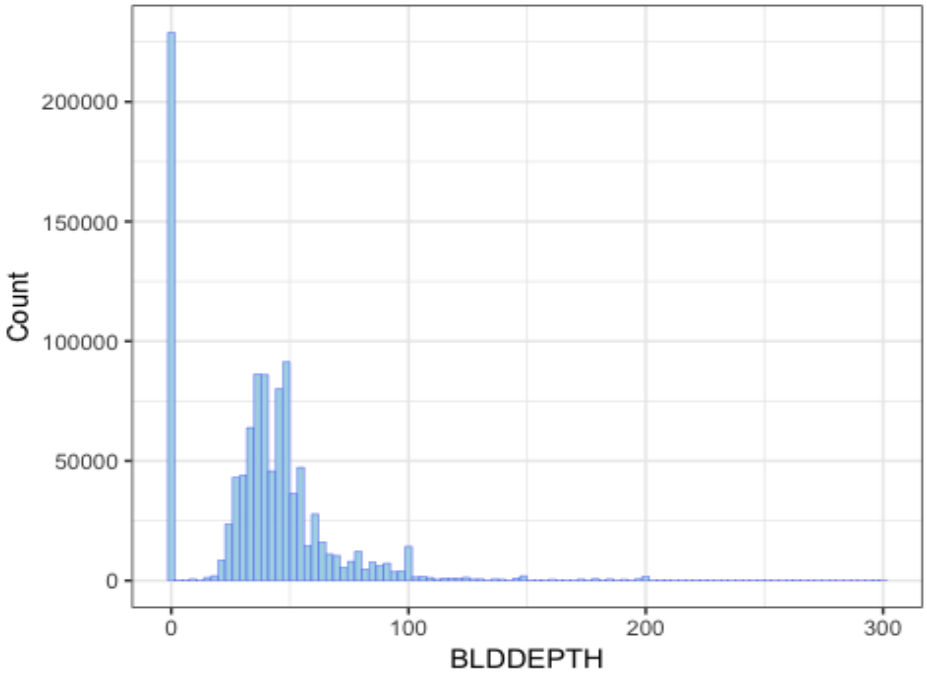
Number of values: 1070994

Number of unique values: 621

%populated: 100

```
##    Min. 1st Qu.  Median    Mean     3rd Qu.    Max.
##    0.00   26.00   39.00      39.92     50.00      9393.00
```

Standard deviation: 42.70715



Percentage of records showed in the chart above is:99.82484

Most of BLDDEPTH are 0 or around 500.

Field25: AVLAND2

Type: Numerical

num [1:1070994] 3.78e+06 1.11e+07 3.23e+07 1.36e+07 1.06e+08 ...

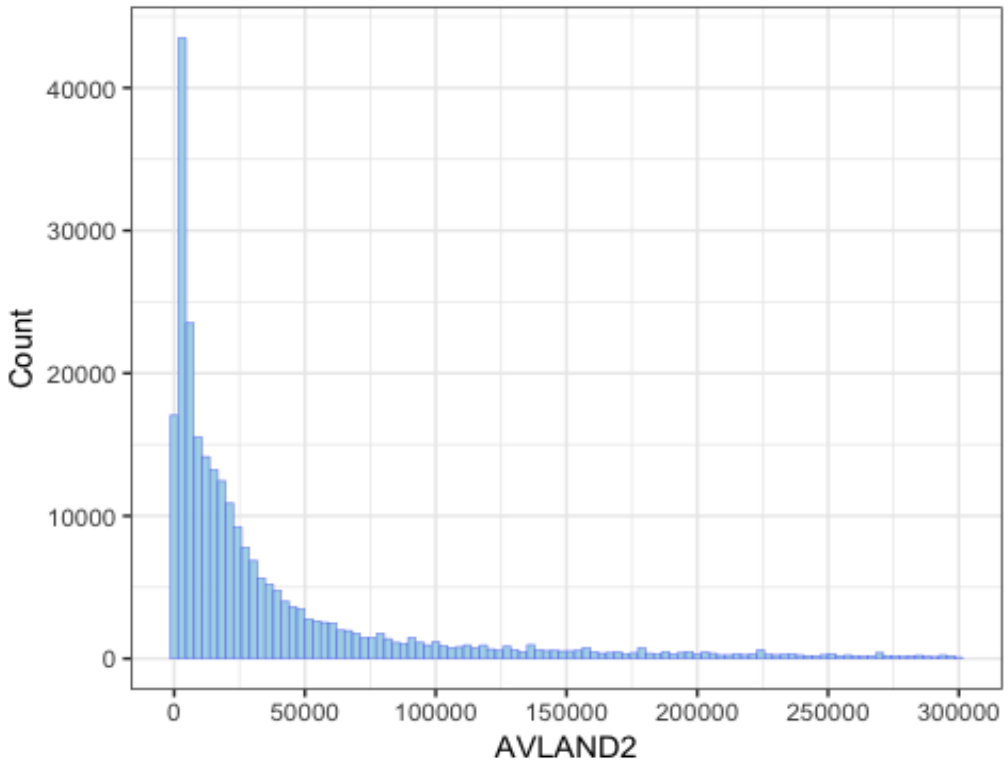
Number of values: 282726

Number of unique values: 58592

%populated: 26.39847

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 3.000e+00 5.705e+03 2.014e+04 2.462e+05 6.264e+04 2.371e+09    788268
```

Standard deviation: 6178963



Percentage of records showed in the chart above is: 24.15037

The chart illustrated AVLAND2 concentrate most between 0 to 50000.

Field26: AVTOT2

Type: Numerical

num [1:1070994] 8.61e+06 8.07e+07 4.02e+07 1.58e+07 1.08e+08 ...

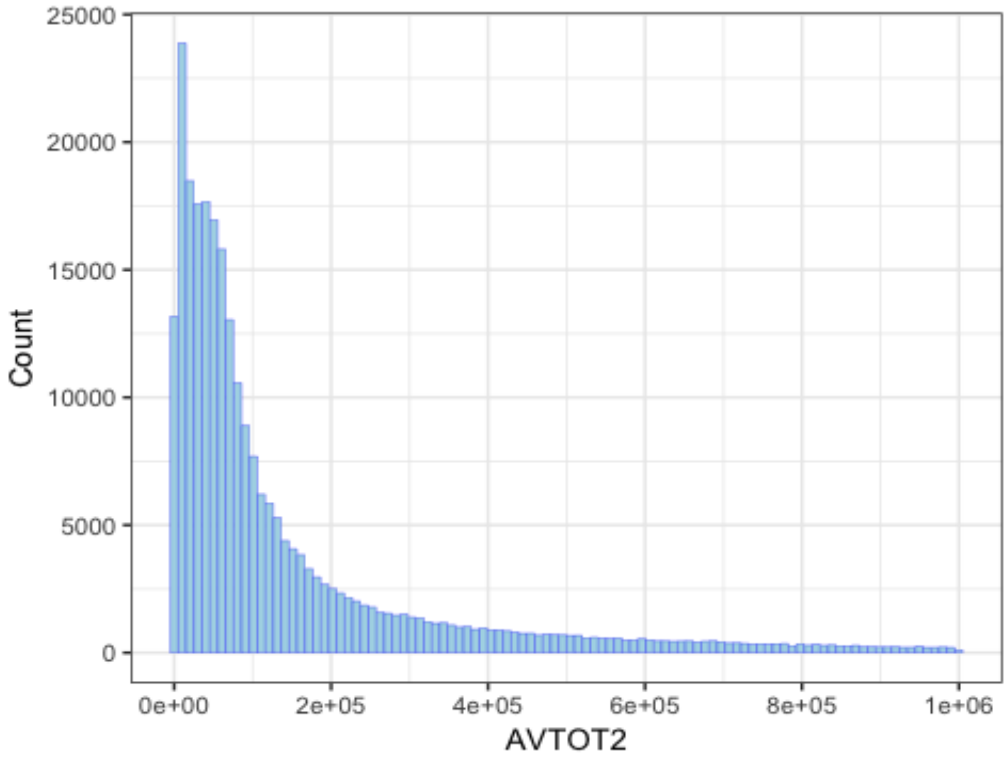
Number of values: 282732

Number of unique values: 111361

%populated: 26.39903

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 3.000e+00 3.391e+04 7.996e+04 7.139e+05 2.406e+05 4.501e+09    788262
```

Standard deviation: 11652529



Percentage of records showed in the chart above is: 26.1219

The chart illustrated AVTOT2 concentrate most between 0 to 200000.

Field27: EXLAND2

Type: Numerical

```
##  num [1:1070994] 3.78e+06 1.11e+07 3.23e+07 1.36e+07 1.06e+08 ...
```

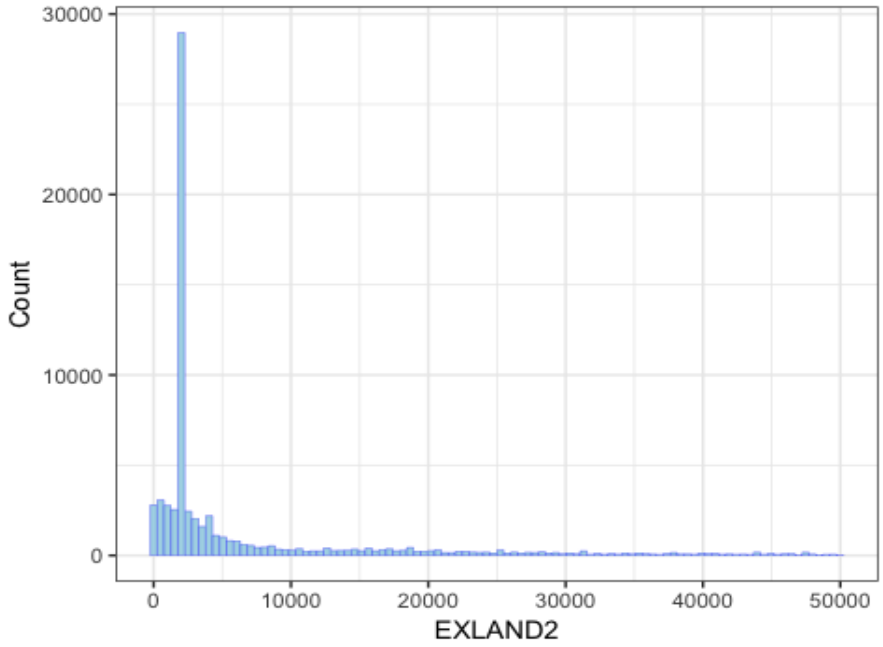
Number of values: 87449

Number of unique values: 22196

%populated: 8.165218

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 1.000e+00 2.090e+03 3.048e+03 3.512e+05 3.178e+04 2.371e+09    983545
```

Standard deviation: 10802213



Percentage of records showed in the chart above is: 6.41815

The highest peak of EXLAND2 appears at nearly 3000.

Field28: EXTOT2

Type: Numerical

num [1:1070994] 8.61e+06 8.07e+07 4.02e+07 1.58e+07 1.08e+08 ...

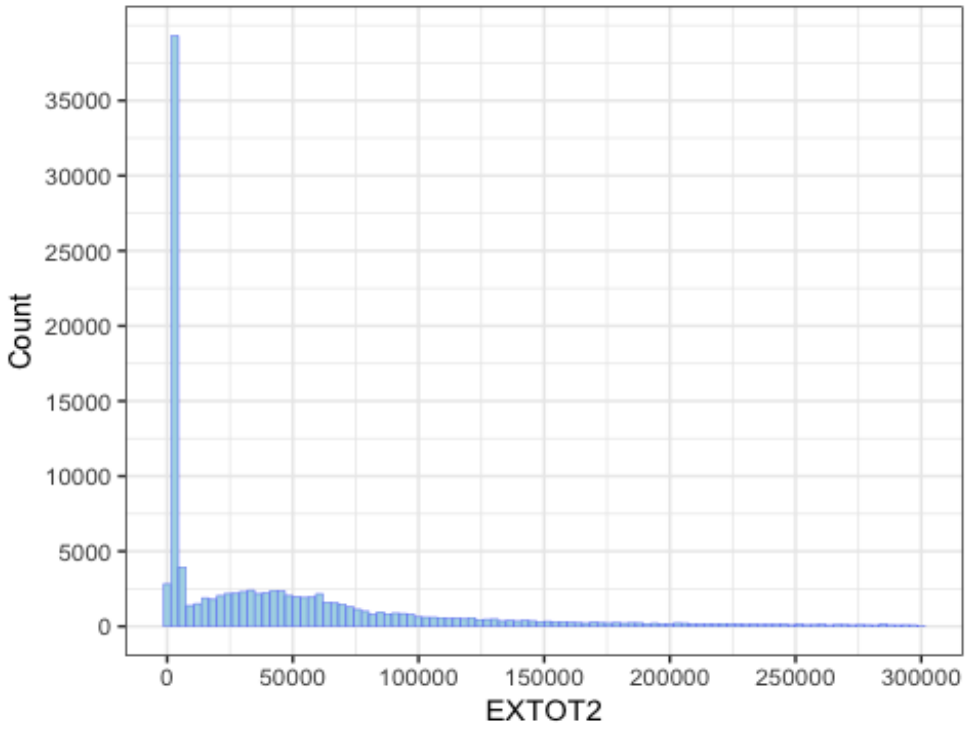
Number of values: 130828

Number of unique values: 48349

%populated: 12.21557

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 7.000e+00 2.870e+03 3.706e+04 6.568e+05 1.068e+05 4.501e+09    940166
```

Standard deviation: 16072510



Percentage of records showed in the chart above is: 10.5854

The highest peak of EXLAND2 appears at nearly 5000.

Field29: EXCD2

Type:

```
##  num [1:1070994] NA NA NA NA NA NA NA NA NA NA ...
```

Number of values: 92948

Number of unique values: 61

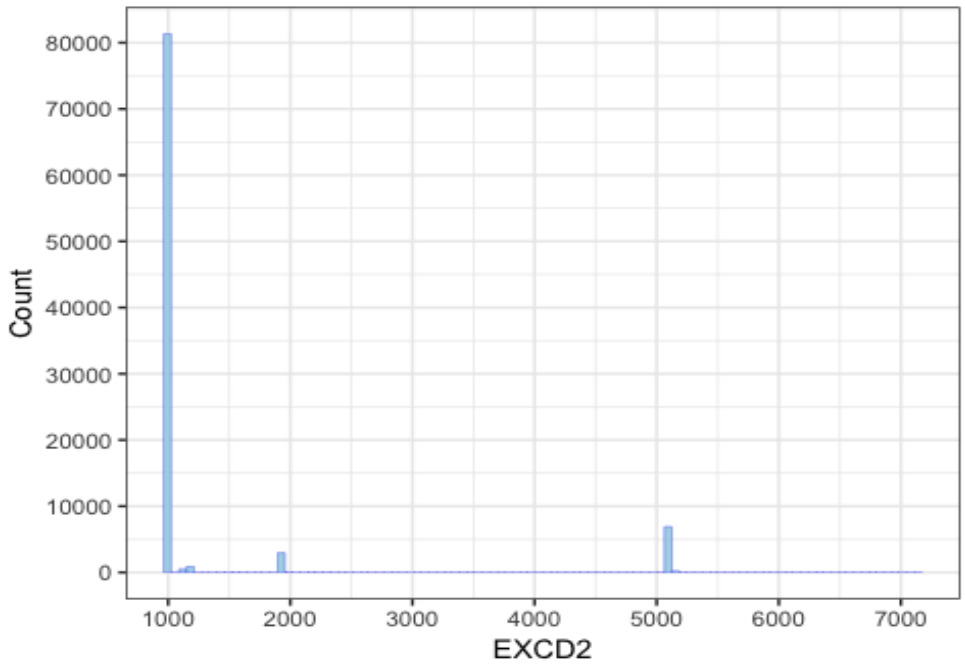
%populated: 8.678667

```
##    Min. 1st Qu.    Median    Mean  3rd Qu.    Max.    NA's
##    1011    1017    1017       1364    1017       7160  978046
```

Standard deviation: 1094.706

Percentage of records showed in the chart above is: 6.78667

The peaks of EXCD2 show at nearly 1000 and nearly 5000.



Field30: PERIOD

Type: Categorical

Factor w/ 1 level "FINAL": 1 1 1 1 1 1 1 1 1 1 ...

So the period is a categorical variable, and just has one value “final”.

Number of values: 1070994

%populated: 100

Field31: YEAR

Type: Categorical

Factor w/ 1 level "2010/11": 1 1 1 1 1 1 1 1 1 1 ...

So the year is a categorical variable, and just has one value “2010/11”.

Number of values: 1070994

Number of unique values: 1

%populated: 100

Field32: VALTYPE

Type:## Factor w/ 1 level "AC-TR": 1 1 1 1 1 1 1 1 1 1 ...

So the valtype is a categorical variable, and just has one value “AC-TR”.

Number of values: 1070994

Number of unique values: 1 %populated: 100