University of Southern California

DSO 562: Fraud Analytics

Project 1: Unsupervised Fraud Model on NY Data

Professor: Stephen Coggeshall

Team 9
Yuping He
Kai-Chen Chan
Lin Zhu
Yayi Yang
Shihui Zhang
Hui Zhang

2/21/2018

Table of contents

Executive Summary

The objective of this research project is to gain insights into the fraud dynamics of real estate valuation in New York, New York using unsupervised machine learning techniques. There are incentives to both under-report and over-report property valuation. This study details a methodology to build a robust model that detects unusual properties in terms of valuations and identifies those outliers for close-up examination.

To build this model, we used a public dataset published by City of New York Department of Finance. The dataset lists 1,070,994 property records in New York City with information about owner, location, value, size and as such. The pipeline of our modeling building process includes:

1) project conceptualization and problem framing
2) exploratory data analysis and quality reporting
3) data cleaning
4) feature engineering
5) feature normalization and dimensionality reduction
6) fraud detection modeling
7) Close-up forensic examination.

The overarching idea of this approach is to evaluate properties in terms of extremeness from average population. Those records farthest away from average rating are flagged as potential fraud.

Data Description

The *City of New York Property Valuation and Assessment Data* is published by the City Department of Finance on NYC Open Data website. This dataset is originally purposed to catalogue valuation assessment of all NYC properties to help city agencies calculate *Property Tax*, *Grant Eligible Properties Exemptions* and *Abatements*. In the dataset, a total of 1,070,994 property records across City of New York are provided 32 fields including size, market value, assessed value, owner, building class, tax class. The data was created on September 2, 2011 and last updated on September 10, 2018.

1.1 Summary Table of Fields

1.1.1 Numeric Variables

Variable	*Min*	*Median*	*Max*	*Mean*	*Std*	*# blank*	*# value zero*	*% populated*
LTFRONT	0	25	9999	36.64	74	0	169108	100.00%
LTDEPTH	0	100	9999	88.86	76	0	170128	100.00%
STORIES	1	2	119	5.01	8	56264	0	94.75%
FULLVAL	0	447000	6.1E^9	874264.50	11582430	0	13007	100.00%
AVLAND	0	13678	2.7E^10	85067.92	4057260	0	13009	100.00%
AVTOT	0	25340	4.7E^9	227238.20	6877529	0	13007	100.00%
EXLAND	0	1620	2.7E^9	36423.89	3981576	0	461699	100.00%
EXTOT	0	1620	4.7E^9	91186.98	6508403	0	432572	100.00%
BLDFRONT	0	20	7575	23.04	36	0	0	100.00%
BLDDEPTH	0	39	9393	39.92	43	0	0	100.00%
AVLAND2	3	20145	2.4E^9	246235.70	6178963	788268	0	26.40%
AVTOT2	3	79962.5	4.5E^9	713911.40	11652530	788262	0	26.40%
EXLAND2	1	3048	2.4E^9	351235.70	10802210	983545	0	8.17%
EXTOT2	7	37062	4.6E^9	656768.30	16072510	940166	0	12.22%

1.1.2 Categorical Variables

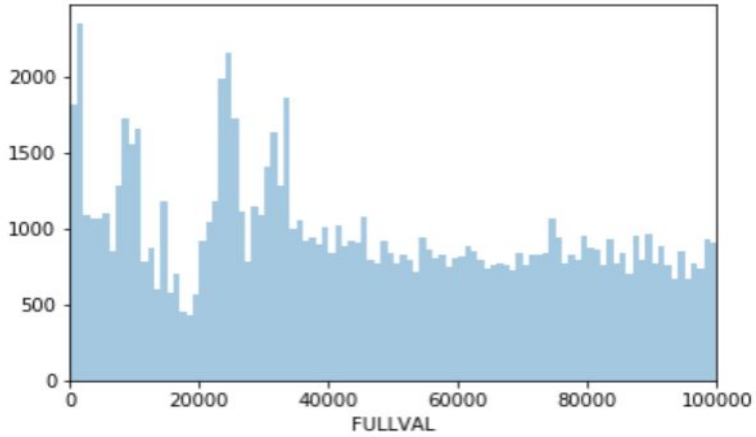
Variable	*# blank*	*% populated*	*# unique*
BBLE	0	100.00%	1070994
B	0	100.00%	5
BLOCK	0	100.00%	13984
LOT	0	100.00%	6366
EASEMENT	1066358	0.43%	13
OWNER	31745	97.04%	863348
BLDGCL	0	100.00%	200
TATCLASS	0	100.00%	11
EXT	716689	33.08%	4
EXCD1	432506	59.62%	130
STADDR	676	99.94%	839281
ZIP	29890	97.21%	197
EXMPTCL	1055415	1.45%	15
EXCD2	978046	8.68%	61
PERIOD	0	100.00%	1
YEAR	0	100.00%	1
VALTYPE	0	100.00%	1

Note: *# bank* is calculated only from empty cells without any value present. It is not equivalent to counting missing values, which are described in the next stage ***Data Cleaning***.

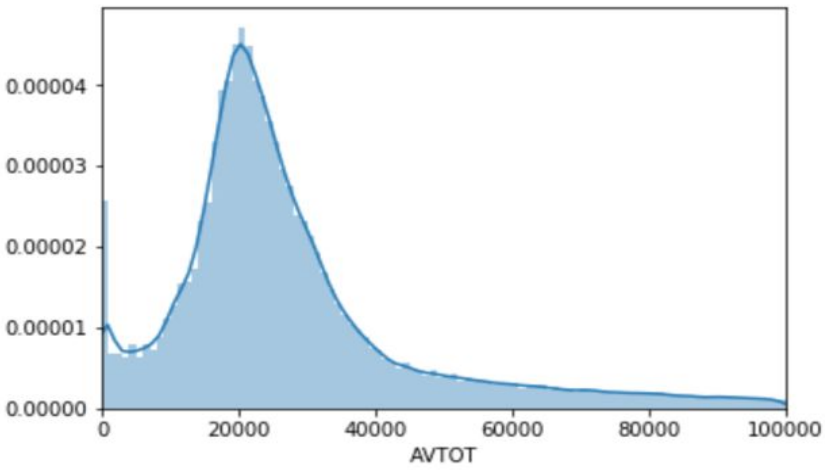
1.2 Key Variables Description

Below is an excerpt of key variables used during model development. All fields are discussed in greater detail in the Data Quality Report attached in appendix.

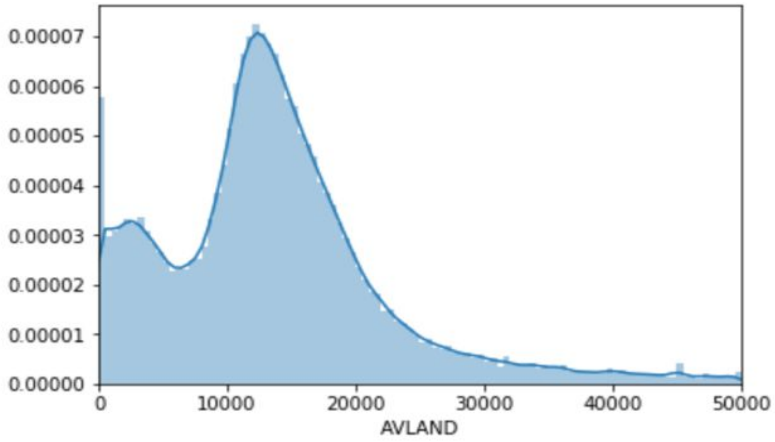
Field Name	*Description*
FULLVAL	Market value of the property in dollars. It is a numeric variable.



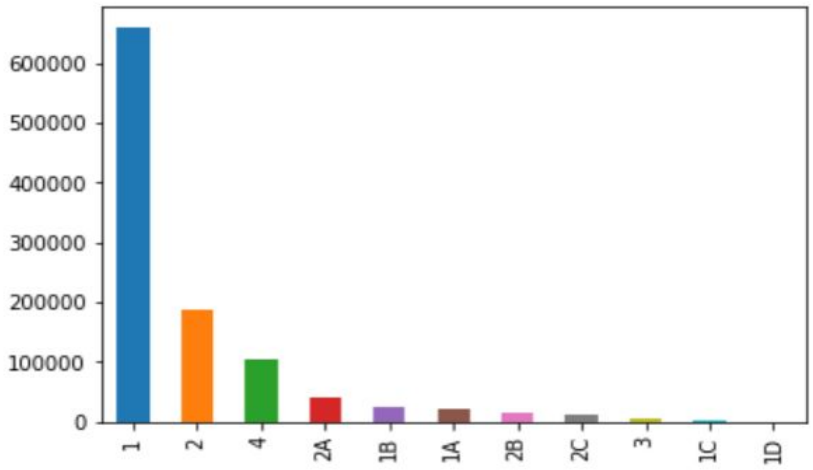
Field Name	*Description*
AVTOT	Actual total value of the property in dollars. It is a numeric variable.



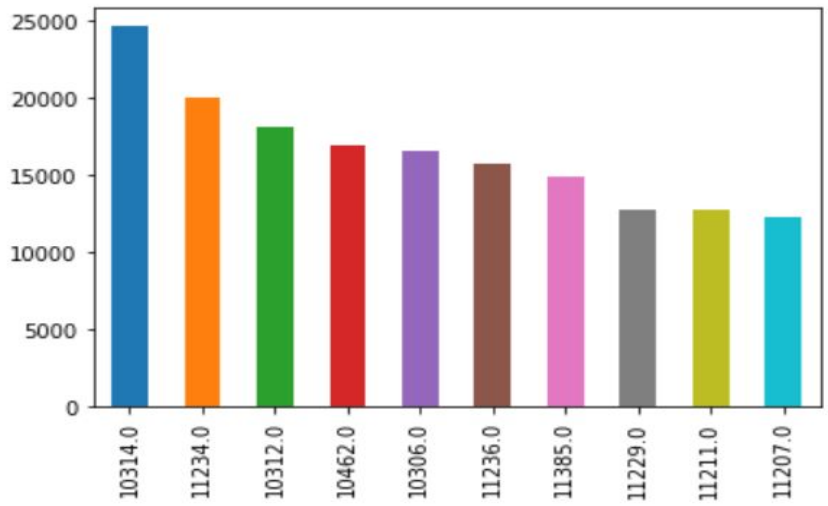
Field Name	*Description*
AVLAND	Actual land value in dollars. It is a numeric variable.



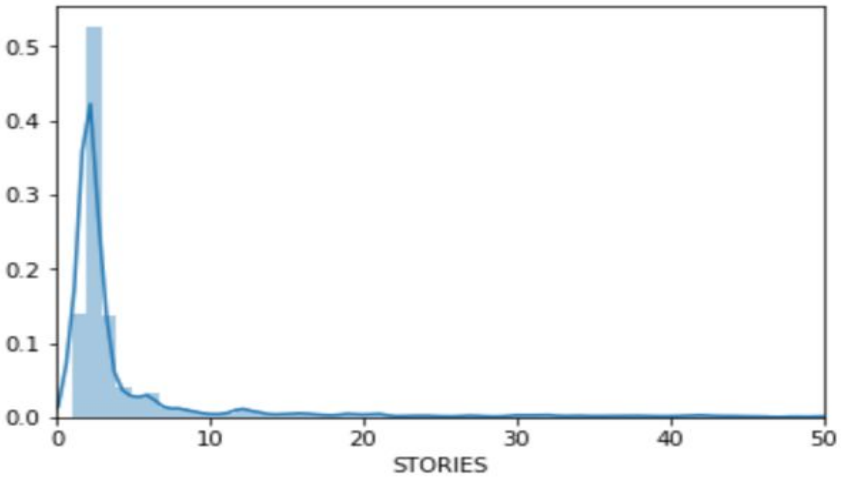
Field Name	*Description*
TAXCLASS	The current property tax class code. It is a categorical variable without missing value.



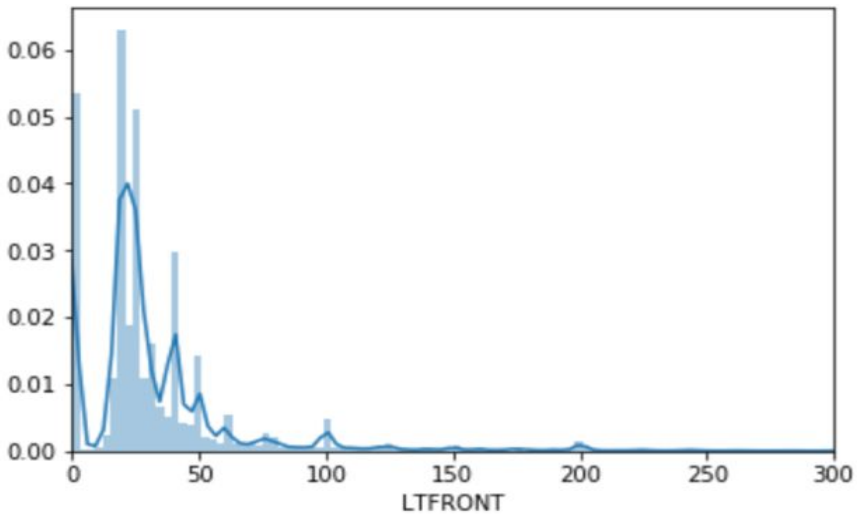
Field Name	*Description*
ZIP	5-digit postal zip code of the property. It is a categorical variable.



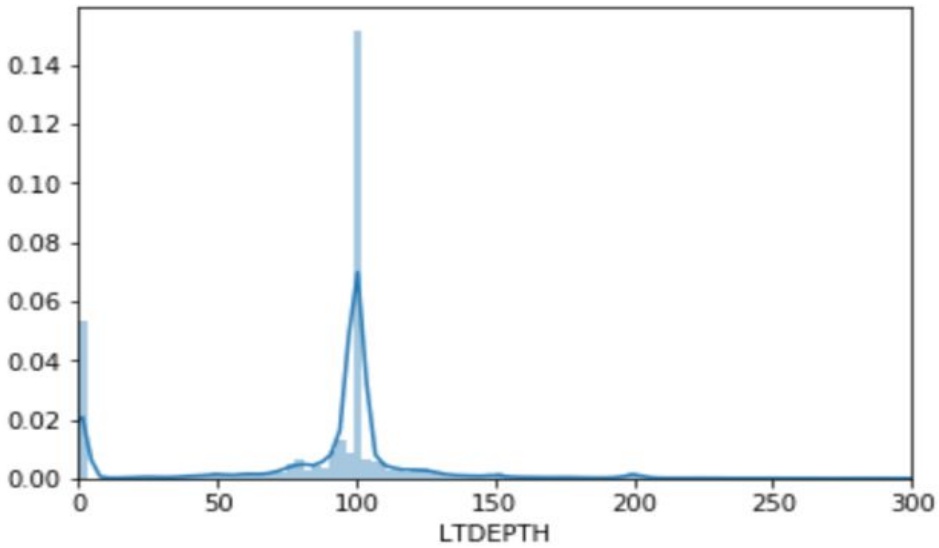
Field Name	*Description*
STORIES	Number of floors of the building. It is a numeric variable.



Field Name	*Description*
LTFRONT	Length of lot frontage. It is a numeric variable.



Field Name	*Description*
LTDEPTH	Horizontal distance between the front and rear property lines of a lot, measured along a line midway between the side property lines. It is a numeric variable.



Data Cleaning

Based on opinion of domain experts in real estate valuation, the following 9 variables were selected to create inputs used to develop the fraud detection model: ZIP, FULLVAL, AVLAND, AVTOT, LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, STORIES. These variables include missing values in the form of either blank or 0. The following table details the severity of missing values for each variable.

Variable name	*Definition*	*Missing records*	*% missing of total*
ZIP	5-digit zip code	29890	2.8%
FULLVAL	Market value	13007	1.2%
AVLAND	Actual land value	13009	1.2%
AVTOT	Actual total value	13007	1.2%
LTFRONT	Lot frontage	169108	15.8%
LTDEPTH	Lot depth	170128	15.9%
BLDFRONT	Building frontage	228815	21.4%
BLDDEPTH	Building depth	228853	21.4%
STORIES	# of floors	56264	5.3%

Imputation methods

For each variable, we provide below the process by which we used to fill in missing values:

- ZIP
  - Use the median of the group aggregate by borough.
- FULLVAL
  - Use the median value of the group aggregated by ZIP code and tax class. If less than 10 records in the that group, aggregate only by ZIP code.
- AVLAND
  - Use the median value of the group aggregated by ZIP code and tax class. If less than 10 records in the that group, aggregate only by ZIP code.
- AVTOT
  - Use the median value of the group aggregated by ZIP code and tax class. If less than 10 records in the that group, aggregate only by ZIP code.
- LTFRONT
  - Use the median value of the group aggregated by building class. If less than 10 records in that group, aggregate by tax class instead.
- LTDEPTH
  - Use the median value of the group aggregate by building class. If less than 10 records in that group, aggregate by tax class instead.

- BLDFRONT
  - Use the median value of the group aggregate by building class. If less than 10 records in that group, aggregate by tax class instead.

- BLDDEPTH
  - Use the median value of the group aggregate by building class. If less than 10 records in that group, aggregate by tax class instead.

- STORIES
  - Use the median value of the group aggregated by building class. If records less than 10 records, aggregate by tax class instead.

Variable Construction

We eventually arrived at 45 expert variables. The following section details the reasoning behind the creation of these variables and exact process the team went through to create them.

Step 1: Create 9 intermediate ratio variables

Valuations - market value, actual land value, and actual total value- alone do not allow one to compare two properties effectively; a larger property might be more valuable in total but less per square feet. Thus, we divided all 3 value measures by 3 sizing measures - namely, lot area, building area and building volume - to place all property on the same footing for comparisons. Please see the resulting 9 ratio variables below.

Ratio variables	*Formula*
r1 - Market value per unit lot area	FULLVAL / (LTFRONT * LTDEPTH)
r2 - Market value per unit building area	FULLVAL / (BLDFRONT * BLDDEPTH)
r3 - Market value per unit building volume	FULLVAL / (BLDFRONT * BLDDEPTH * STORIES)
r4 - Actual land value per unit lot area	AVLAND / (LTFRONT * LTDEPTH)
r5 - Actual land value per unit building area	AVLAND / (BLDFRONT * BLDDEPTH)
r6 - Actual land value per unit building volume	AVLAND / (BLDFRONT * BLDDEPTH * STORIES)
r7 - Actual total value per unit lot area	AVTOT / (LTFRONT * LTDEPTH)
r8 - Actual land value per unit building area	AVTOT / (BLDFRONT * BLDDEPTH)

r9 - Actual total value per unit building	AVTOT / (BLDFRONT * BLDDEPTH * STORIES)

Step 2: Encode 4 categorical variables into the original 9 ratios to form final 9 sets of variables

For each of the 9 ratio variables, we calculated the ratio of its value to the average value of records within the same zip code. The idea is to measure how much a given record's value differs from the group average. For example, a value of 1 in r1_zip5 means that the record has the same market value per unit lot area as the average of all records within the same zip code.

Note: ZIP3 is the first 3 digit of ZIP; ZIP5 is the full 5-digit ZIP

Set 1: Market value per unit lot area	*Variable name*	*Formula*
1	r1_zip5	$\frac{r1}{average\ r1\ of\ all\ records\ in\ the\ same\ ZIP5}$
2	r1_zip3	$\frac{r1}{average\ r1\ of\ all\ records\ in\ the\ same\ ZIP3}$
3	r1_tax	$\frac{r1}{average\ r1\ of\ all\ records\ in\ the\ same\ tax\ class}$
4	r1_b	$\frac{r1}{average\ r1\ of\ all\ records\ in\ the\ same\ borough}$
5	r1_all	$\frac{r1}{average\ r1\ of\ all\ records}$

Set 2: Market value per unit building area	*Variable name*	*Formula*
6	r2_zip5	$\frac{r2}{average\ r2\ of\ all\ records\ in\ the\ same\ ZIP5}$
7	r2_zip3	$\frac{r2}{average\ r2\ of\ all\ records\ in\ the\ same\ ZIP3}$
8	r2_tax	$\frac{r2}{average\ r2\ of\ all\ records\ in\ the\ same\ tax\ class}$

9	r2_b	$\frac{r2}{average\ r2\ of\ all\ records\ in\ the\ same\ borough}$
10	r2_all	$\frac{r2}{average\ r2\ of\ all\ records}$

Set 3: Market value per unit building volume	*Variable name*	*Formula*
11	r3_zip5	$\frac{r3}{average\ r3\ of\ all\ records\ in\ the\ same\ ZIP5}$
12	r3_zip3	$\frac{r3}{average\ r3\ of\ all\ records\ in\ the\ same\ ZIP3}$
13	r3_tax	$\frac{r3}{average\ r3\ of\ all\ records\ in\ the\ same\ tax\ class}$
14	r3_b	$\frac{r3}{average\ r3\ of\ all\ records\ in\ the\ same\ borough}$
15	r3_all	$\frac{r3}{average\ r31\ of\ all\ records}$

Set 4: Actual land value per unit lot area	*Variable name*	*Formula*
16	r4_zip5	$\frac{r4}{average\ r4\ of\ all\ records\ in\ the\ same\ ZIP5}$
17	r4_zip3	$\frac{r4}{average\ r4\ of\ all\ records\ in\ the\ same\ ZIP3}$
18	r4_tax	$\frac{r4}{average\ r4\ of\ all\ records\ in\ the\ same\ tax\ class}$
19	r4_b	$\frac{r4}{average\ r4\ of\ all\ records\ in\ the\ same\ borough}$
20	r4_all	$\frac{r4}{average\ r4\ of\ all\ records}$

Set 5: Actual land value per unit building area	*Variable name*	*Formula*
21	r5_zip5	$\frac{r5}{average\ r5\ of\ all\ records\ in\ the\ same\ ZIP5}$
22	r5_zip3	$\frac{r5}{average\ r5\ of\ all\ records\ in\ the\ same\ ZIP3}$
23	r5_tax	$\frac{r5}{average\ r5\ of\ all\ records\ in\ the\ same\ tax\ class}$
24	r5_b	$\frac{r5}{average\ r5\ of\ all\ records\ in\ the\ same\ borough}$
25	r5_all	$\frac{r5}{average\ r5\ of\ all\ records}$

Set 6: Actual land value per unit building volume	*Variable name*	*Formula*
26	r6_zip5	$\frac{r6}{average\ r6\ of\ all\ records\ in\ the\ same\ ZIP5}$
27	r6_zip3	$\frac{61}{average\ r6\ of\ all\ records\ in\ the\ same\ ZIP3}$
28	r6_tax	$\frac{r6}{average\ r6\ of\ all\ records\ in\ the\ same\ tax\ class}$
29	r6_b	$\frac{r6}{average\ r6\ of\ all\ records\ in\ the\ same\ borough}$
30	r6_all	$\frac{r6}{average\ r6\ of\ all\ records}$

Set 7: Actual total value per unit lot area	*Variable name*	*Formula*
31	r7_zip5	$\frac{r7}{average\ r7\ of\ all\ records\ in\ the\ same\ ZIP5}$

32	r7_zip3	$\frac{r7}{average\ r7\ of\ all\ records\ in\ the\ same\ ZIP3}$
33	r7_tax	$\frac{r7}{average\ r7\ of\ all\ records\ in\ the\ same\ tax\ class}$
34	r7_b	$\frac{r7}{average\ r7\ of\ all\ records\ in\ the\ same\ borough}$
35	r7_all	$\frac{r7}{average\ r7\ of\ all\ records}$

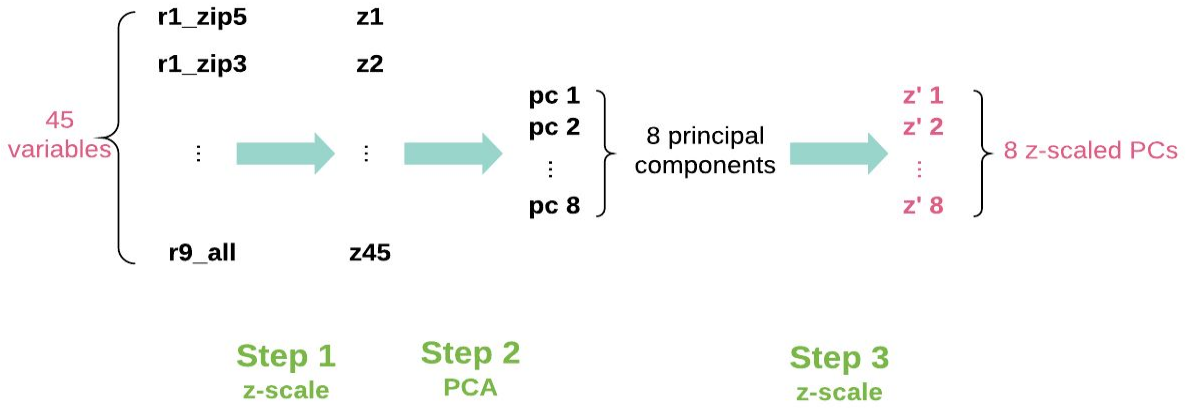
Set 8: Actual land value per unit building area	*Variable name*	*Formula*
36	r8_zip5	$\frac{r8}{average\ r8\ of\ all\ records\ in\ the\ same\ ZIP5}$
37	r8_zip3	$\frac{r8}{average\ r8\ of\ all\ records\ in\ the\ same\ ZIP3}$
38	r8_tax	$\frac{r8}{average\ r8\ of\ all\ records\ in\ the\ same\ tax\ class}$
39	r8_b	$\frac{r8}{average\ r8\ of\ all\ records\ in\ the\ same\ borough}$
40	r8_all	$\frac{r8}{average\ r8\ of\ all\ records}$

Set 9: Actual total value per unit building	*Variable name*	*Formula*
41	r9_zip5	$\frac{r9}{average\ r9\ of\ all\ records\ in\ the\ same\ ZIP5}$
42	r9_zip3	$\frac{r9}{average\ r9\ of\ all\ records\ in\ the\ same\ ZIP3}$
43	r9_tax	$\frac{r9}{average\ r9\ of\ all\ records\ in\ the\ same\ tax\ class}$
44	r9_b	$\frac{r9}{average\ r9\ of\ all\ records\ in\ the\ same\ borough}$

45	r9_all	$\frac{r9}{average\ r9\ of\ all\ records}$

Dimensionality Reduction

Following the feature engineering process came along 45 expert variables. A total of 45 dimensionalities prove ineffective for modeling development. Intuitively, we understood that some of these variables are highly correlated with one another. It was critical that we reduce the dimensionality to a reasonable few while retaining information. We went through the following a 3-step process to extract the most information-rich variables.



Step 1: Scaling
Z-scaling is performed on all 45 final variables to ensure the same weight assignment to each input.

Step 2: Principal Component Analysis
PCA is performed on the z-scaled 45 variables to remove correlation and reduce dimensionality to 8 principal components.

Step 3: Scaling
Z-scaling is performed again on the 8 principal components to ensure same assignment of importance to each. Each of the 8 z-scores is a mathematical encoding of unusualness from the average crowd. The greater the absolute value of a z-score here, the more unusual of a record.

Fraud Detection Model

Two fraud algorithms were used to build the fraud detection model. Each algorithm returns a score that measures unusualness of a property record:

- Mahalanobis + Manhattan distance → **Score 1**
- Auto-encoder error → **Score 2**

After scaling these 2 scores through quantile binning, we computed the final fraud score as a sum of 2 scaled fraud scores:

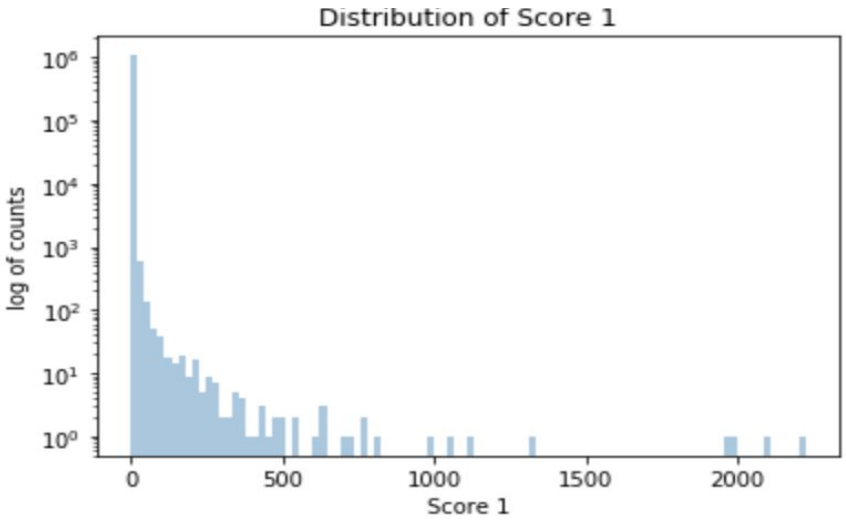
- ❖ **Score Final** ? **Score 1 (scaled)**+**Score 2 (scaled)**

1.1 Mahalanobis + Manhattan Distance

The Manhattan distance is the sum of absolute values of 8 z-scores. The greater the value, the more a record deviates from group average.

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

The distribution of Score 1 is shown as follows.

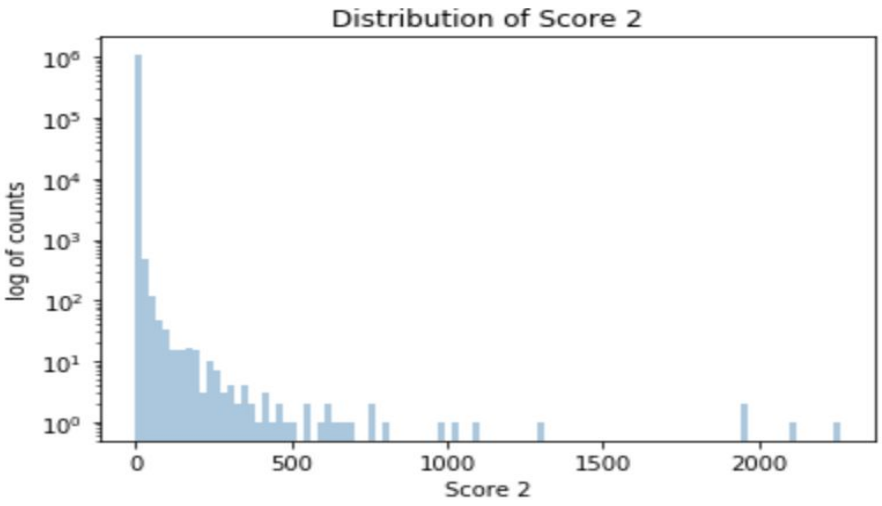


1.2 Autoencoder

Autoencoder is a popular technique that learns from an original dataset and reconstruct a replica that closely matches the original. The error of replication then represents unusualness of a record because the encoder fails to understand the construct of that record based on information learned from the whole dataset. Records that have the largest replication error are considered as outliers.

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

The distribution of Score 2 is shown as follows:



Forensic Examinations & Conclusion

To understand the underlying reasons for outliers, we investigated top 10 outliers from our model. Namely, those records with the highest **Score 1**, **Score2**, and/or **Score Final**.

❖ **Top 10 suspicious records by Score 1**

record	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5	principal component 6	principal component 7	principal component 8	S1	record	bin1
632816	540.636679	-200.7440136	-652.8037416	401.6446538	355.7051485	34.97890969	-36.81075464	0.673416096	2223.997317	632816	1070994
1067360	92.18115565	615.7023299	51.42020404	525.9334891	-290.9715202	-58.42384281	-93.63915012	366.8871429	2095.158835	1067360	1070993
565392	286.0881305	625.0073872	-72.29187258	-511.7832015	358.4642487	77.27977492	19.49101519	36.2430683	1986.648699	565392	1070992
67129	667.660081	-186.3711725	694.6576527	72.20357458	118.5483104	-171.4733739	-46.88203566	-18.80850154	1976.604702	67129	1070991
585118	167.1797274	-42.19181609	25.23688679	-121.8517301	-237.2970149	416.5076083	-281.103935	27.19399053	1318.562709	585118	1070990
585120	130.8600028	-28.25766181	-20.34376888	-129.0274304	-231.7814607	324.8367311	-238.6173247	-8.756842358	1112.481223	585120	1070989
917942	117.3005797	-7.299232849	-92.55460714	-155.7845382	-181.722603	-71.95617853	-383.8899621	-40.11374589	1050.621447	917942	1070988
585439	135.1368129	-39.12148747	30.31871381	-73.24137936	-216.3339647	395.5485993	59.12322615	49.4599689	998.2841526	585439	1070987
920628	71.75132356	-23.30701156	-16.20518048	-13.61839121	-114.5930506	131.9682779	392.7754627	54.21978846	818.4384865	920628	1070986
293330	89.78061673	-13.47312289	-94.58397429	-109.2122499	-160.9867247	-258.1076555	-36.46207187	-7.647582237	770.2539981	293330	1070985

❖ **Top 10 suspicious records by Score 2**

record	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5	principal component 6	principal component 7	principal component 8	s2	bin2
632816	540.636679	215.1563319	652.8037416	401.6446538	355.7051485	31.00005552	59.91810976	0.673416096	2257.538136	1070994
1067360	92.18115565	601.2884648	51.42020404	525.9334891	290.9715202	62.40315332	116.7491946	366.8871429	2107.834325	1070993
67129	650.8121424	186.3711725	694.6576527	72.20357458	118.5483104	171.4733739	46.88203566	19.65310019	1960.601362	1070992
565392	286.0881305	610.2128516	72.29187258	511.7832015	358.4642487	73.18839575	4.279746301	36.2430683	1952.551515	1070991
585118	150.3317888	42.19181609	25.23688679	121.8517301	237.2970149	416.5076083	281.103935	26.34939188	1300.870172	1070990
585120	114.0120643	28.25766181	20.34376888	129.0274304	231.7814607	324.8367311	238.6173247	9.601441008	1096.477883	1070989
917942	100.7409331	7.299232849	92.55460714	155.7845382	181.722603	71.95617853	383.8899621	40.9348441	1034.882899	1070988
585439	118.2888743	39.12148747	30.31871381	73.24137936	216.3339647	395.5485993	59.12322615	48.61537024	980.5916154	1070987
920628	71.75132356	38.15898796	16.20518048	13.61839121	114.5930506	127.8597898	368.9036155	54.21978846	805.3101276	1070986
293330	73.22097203	13.47312289	94.58397429	109.2122499	160.9867247	258.1076555	36.46207187	8.468680207	754.5154514	1070985

We took the top 10 suspicious records back to the original dataset and re-surfaced their other attributes.

❖

RECORD	BBLE	B	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH
67129	1011110001	1	1111	1	NaN	CULTURAL AFFAIRS	Q1	4	840	180
293330	3011170001	3	1117	1	NaN	CITY OF NY/PARKS AND	Q1	4	526	250
565392	3085900700	3	8590	700	NaN	U S GOVERNMENT OWNRD	V9	4	117	108
585118	4004200001	4	420	1	NaN	NEW YORK CITY ECONOMI	O3	4	298	402
585120	4004200101	4	420	101	NaN	NaN	O3	4	139	342
585439	4004590005	4	459	5	NaN	11-01 43RD AVENUE REA	H9	4	94	165
632816	4018420001	4	1842	1	NaN	864163 REALTY, LLC	D9	2	157	95
917942	4142600001	4	14260	1	NaN	LOGAN PROPERTY, INC.	T1	4	4910	332
920628	4155770029	4	15577	29	NaN	PLUCHENIK, YAAKOV	A1	1	91	100
1067360	5078530085	5	7853	85	NaN	NaN	B2	1	1	1

❖

RECORD	STORIES	FULLVAL	AVLAND	AVTOT	EXLAND	EXTOT	EXCD1	STADDR	ZIP
67129	1	6.15E+09	2.67E+09	2.77E+09	2.67E+09	2.77E+09	2231	1000 5 AVENUE	10028
293330	1	2.71E+08	1.16E+08	1.22E+08	1.16E+08	1.22E+08	2231	95 PROSPECT PARK WEST	11215
565392	2	4.33E+09	1.95E+09	1.95E+09	1.95E+09	1.95E+09	2231	FLATBUSH AVENUE	11215
585118	20	3.44E+06	1.55E+06	1.55E+06	0	0	NaN	28-10 QUEENS PLAZA SOUTH	11101
585120	20	2.15E+06	968220	968220	0	0	NaN	28 STREET	11101
585439	10	3.71E+06	252000	1.67E+06	0	1.42E+06	1986	11-01 43 AVENUE	11101
632816	1	2.93E+06	1.32E+06	1.32E+06	0	0	NaN	86-55 BROADWAY	11373
917942	3	3.74E+08	1.79E+09	4.67E+09	1.79E+09	4.67E+09	2198	154-68 BROOKVILLE BOULEVARD	11422
920628	2	1.90E+06	9763	75763	0	0	NaN	7-06 ELVIRA AVENUE	11691
1067360	2	836000	28800	50160	0	0	NaN	20 EMILY COURT	10307

❖

RECORD	EXMPTCL	BLDFRONT	BLDDEPTH	AVLAND2	AVTOT2	EXLAND2	EXTOT2	EXCD2	PERIOD	YEAR	VALTYPE
67129	X1	725	725	2.37E+09	2.47E+09	2.37E+09	2.47E+09	NaN	FINAL	2010/11	AC-TR
293330	X1	25	25	1.06E+08	1.12E+08	1.06E+08	1.12E+08	Na	FINAL	2010/20	AC-TR
565392	X1	44	80	8.48E+08	8.48E+08	8.48E+08	8.48E+08	NaN	FINAL	2010/12	AC-TR
585118	X1	1	1	1.59E+06	1.59E+06	NaN	NaN	NaN	FINA	2010/15	AC-TR
585120	NaN	1	1	975456	975456	NaN	NaN	NaN	FINAL	2010/16	AC-TR
585439	NaN	1	1	NaN	NaN	NaN	NaN	NaN	FINAL	2010/18	AC-TR
632816	NaN	1	1	1.20E+06	1.20E+06	NaN	NaN	NaN	FINAL	2010/11	AC-TR
917942	X4	60	82.5	1.64E+09	4.50E+09	1.64E+09	4.50E+09	NaN	FINAL	2010/17	AC-TR
920628	NaN	1	1	NaN	NaN	NaN	NaN	NaN	FINAL	2010/19	AC-TR
1067360	NaN	36	45	NaN	NaN	NaN	NaN	NaN	FINAL	2010/11	AC-TR

1) **Record 67129 → high fraud possibility**
   - Extremely high value of FULLVAL/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Extremely high value of AVTOT/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Original record missing LTDEPTH, BLDFRONT, and BLDEPTH values
   - Total exemption of the property is high in both EXTOT1 and EXTOT2

2) **Record 293330 → high fraud possibility**
   - Extreme value of FULLVAL/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Extreme value of AVTOT/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Total exemption of the property is high in both EXTOT1 and EXTOT2

3) **Record 565392 → high fraud possibility**
   - Extremely high value of FULLVAL/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Extremely high value of AVTOT/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS

4) **Record 585118 → high fraud possibility**
   - Relatively low value of FULLVAL/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVLAND/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVTOT//(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS

- Relatively low value of FULLVAL /(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
- Relatively low value of AVLAND/(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
- Relatively low value of AVTOT /(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
- 20 stories but BLDDEPTH is only 1

5) **Record 585120 → high fraud possibility**
   - Relatively low value of FULLVAL/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVLAND/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVTOT//(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of FULLVAL /(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVLAND/(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVTOT /(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
   - 20 stories but BLDDEPTH is only 1

6) **Record 585439 → high fraud possibility**
   - Relatively low value of FULLVAL/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVLAND/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVTOT//(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of FULLVAL /(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVLAND/(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
   - Relatively low value of AVTOT /(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
   - 20 stories but BLDDEPTH is only 1

7) **Record 920628 → high fraud possibility**
   - Extremely low value of AVLAND/(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS

- Extremely low value of AVTOT//(LTFRONT* LTDEPTH) with respect to BLOCK and TAX CLASS
- Extremely low value of AVLAND/(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS
- Extremely low value of AVTOT /(BLDFRONT*BLDDEPTH) with respect to BLOCK and TAX CLASS

8) **Record 917942 → low fraud possibility**
   - Further analysis is needed to determine the fraudulent nature of this record. If this is indeed not an outlier, we need to refine our model in the future.

9) **Record 920628 → high fraud possibility**
   - Extremely low value of AVLAND/(LTFRONT* LTDEPTH)
   - Extremely low value of AVTOT/(LTFRONT* LTDEPTH)
   - Extremely low value of AVLAND/(BLDFRONT*BLDDEPTH)
   - Extremely low value of AVTOT /(BLDFRONT*BLDDEPTH)
   - 2 stories but the value of BLDFRONT and BLDDEPTH are equal to 1

10) **Record 1067360 → high fraud possibility**
   - Extremely low value of AVLAND/(LTFRONT* LTDEPTH)
   - Extremely low value of AVTOT/(LTFRONT* LTDEPTH)
   - Extremely low value of AVLAND/(BLDFRONT*BLDDEPTH)
   - Extremely low value of AVTOT /(BLDFRONT*BLDDEPTH)
   - The ratio of Lot area to Building area is too high

Conclusion:

In order to analyze potential fraud records in New York City Properties which contains more than 1 million data. We followed a general process of unsupervised data analysis including data cleaning, filling missing data, variable construction, Z-scale, Principal Component Analysis, fraud algorithms creation, score calculation, and potential fraudulent records identification. Python and R are used during the process since they are effective tools for analyzing.
Z-scale assigned each variables same weight so they can stay at the same level for analyzing. PCA is a tool for eliminating correlation and dimensionality reduction. We repeated those to processes twice in order to get more accurate data for analyzing. Then we applied both Neural Network model for prediction and get a new score. Finally, we applied quantile binning to normalize score1 and score 2 at the same level for comparison.

High scored records are analyzed entirely since they have high suspicious of being fraudulent properties. The top ten scores generated by both methods are same so a detailed analysis are applied to those 10 records. We focused on calculate the value per square of those properties and compare them with others based on to their owners, address, and so on.

There still have many things we can done in the future for making this project more accurate. Firstly, we can fill the missing value more accurately. Instead of using data like ZIP, LTFRONT, LTDEPTH, etc. we can fill up the data according to the property owner or lot area data. Secondly, although we applied the PCA new variables to delete correlations, we can apply this method for the original data at first which can help us find more way to create new variables about find the relationship among AVLAND, AVTOT, FULLVAL to determine the extreme values. Finally, since some areas in New York city belong to government and they may have tax exemption or other advantages. It would be better if we filter the out those privileged properties before analyzing.

Appendix

Data Quality Report

1. Data Description

Data name: Property valuation and assessment data
Available at:
https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8
Data Provided by: Department of Finance (DOF)
Dataset Owner: NYC OpenData
Category: Housing& Development
Data created: September 2, 2011
Metadata Last Updated: September 10, 2018
Number of fields: 32
Number of records: 1070994
Details variables: RECORD, BBLE, B, BLOCK, LOT, EASEMENT, OWNER, BLDGCL, TAXCLASS, LTFRONT, TDEPTH, EXT, STORIES, FULLVAL, AVLAND, AVTOT, EXLAND, EXTOT, EXCD1, STADDR, ZIP, EXMPTCL, BLDFRONT, BLDDEPTH, AVLAND2, AVTOT2, EXLAND2, EXTOT2, EXCD2, PERIOD, YEAR, VALTYPE.

2. List of Info for Each Field

The table below listed the name of the field, the number of records, the percentage of the population, the number of unique values and the number of records with value zero in each field.

	Field name	# records	%populated	#unique values	# records with value zero
1	RECORD	1070994	100	1070994	0
2	BBLE	1070994	100	1070994	0
3	B	1070994	100	5	0
4	BLOCK	1070994	100	13984	0
5	LOT	1070994	100	6366	0
6	EASEMENT	4636	43.29	13	0
7	OWNER	1039249	97.04	863348	0
8	BLDGCL	1070994	100	200	0
9	TAXCLASS	1070994	100	11	0
10	LTFRONT	1070994	100	1297	169108
11	LTDEPTH	1070994	100	1370	170128
12	EXT	354305	33.08	4	0
13	STORIES	1014730	94.75	112	0
14	FULLVAL	1070994	100	109324	13007
15	AVLAND	1070994	100	70921	13009
16	AVTOT	1070994	100	112914	13007
17	EXLAND	1070994	100	33419	491699
18	EXTOT	1070994	100	64255	432572
19	EXCD1	638488	59.62	130	0
20	STADDR	1070318	99.94	839281	0
21	ZIP	1041104	97.21	197	0

22	EXMPTCL	15579	1.46	15	0
23	BLDFRONT	1070994	100	612	228815
24	BLDDEPTH	1070994	100	621	228853
25	AVLAND2	282726	26.40	58592	0
26	AVTOT2	282732	26.40	111361	0
27	EXLAND2	87449	8.16	22196	0
28	EXTOT2	130828	12.22	48349	0
29	EXCD2	92948	8.68	61	0
30	PERIOD	1070994	100	1	0
31	YEAR	1070994	100	1	0
32	VALTYPE	1070994	100	1	0

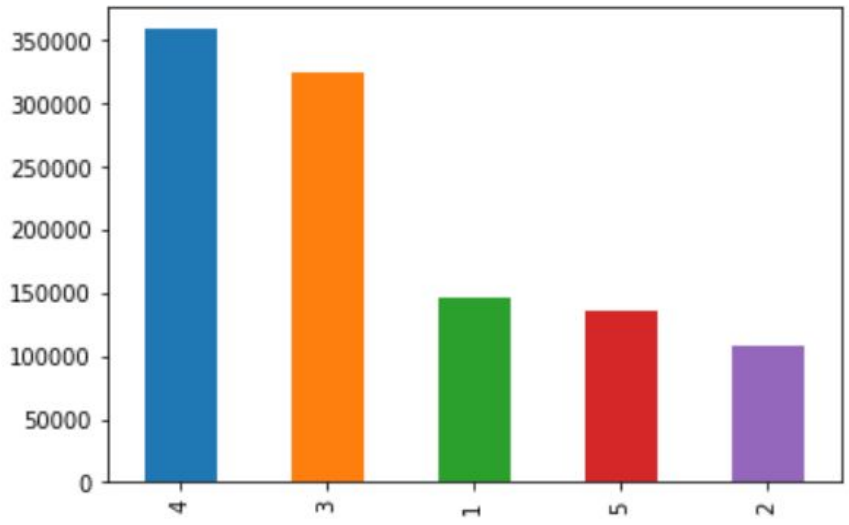
Numeric Fields: LTFRONT, LTDEPTH, STORIES, FULLVAL, AVLAND, AVTOT, EXLAND, EXTOT, EXCD1, BLDFRONT, BLDDEPTH, AVLAND2, AVTOT2, EXLAND2, EXTOT2, EXCD2,
Categorical Fields: BBLE, B, BLOCK, LOT, EASEMENT, OWNER, BLDGCL, TAXCLASS, STADDR, ZIP, EXMPTCL, PERIOD, VALTYPE

3. Description of each field

RECORD:
Name: “RECORD”
Field type: Categorical
Description: Number to track each record

BBLE:
Name: “BBLE
Field type: Categorical
Description: N/A
In this field, all numbers are unique.

B:
Name: “B”
Field type: Categorical
Description: N/A



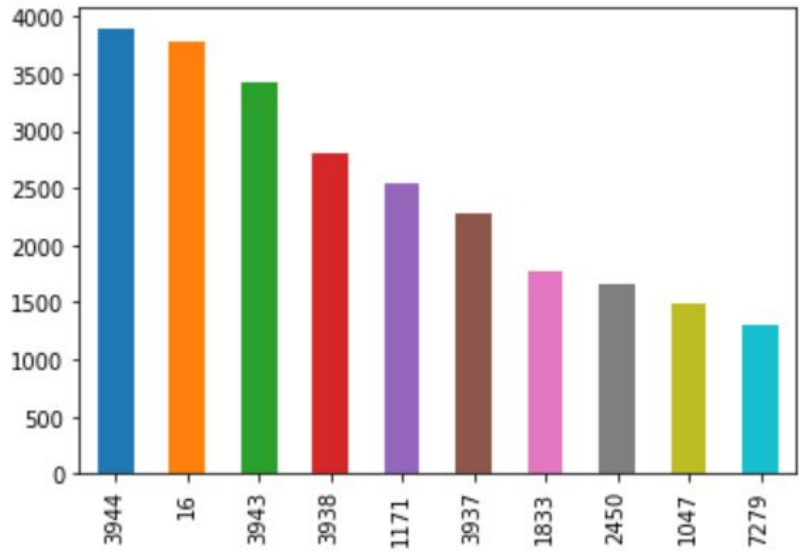
Most common field value: 4

BLOCK:

Name: “BLOCK”

Field type: Categorical

Description: N/A



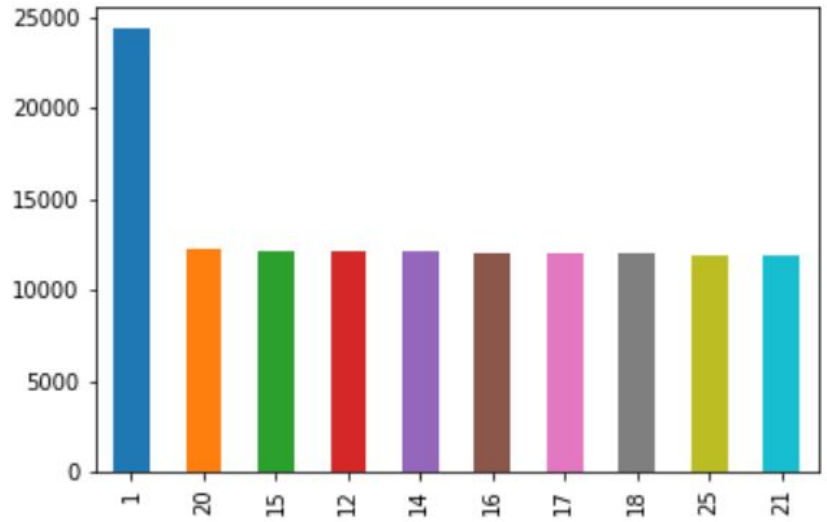
Most common field value: 3944

LOT:

Name: “LOT”

Field type: Categorical

Description: N/A



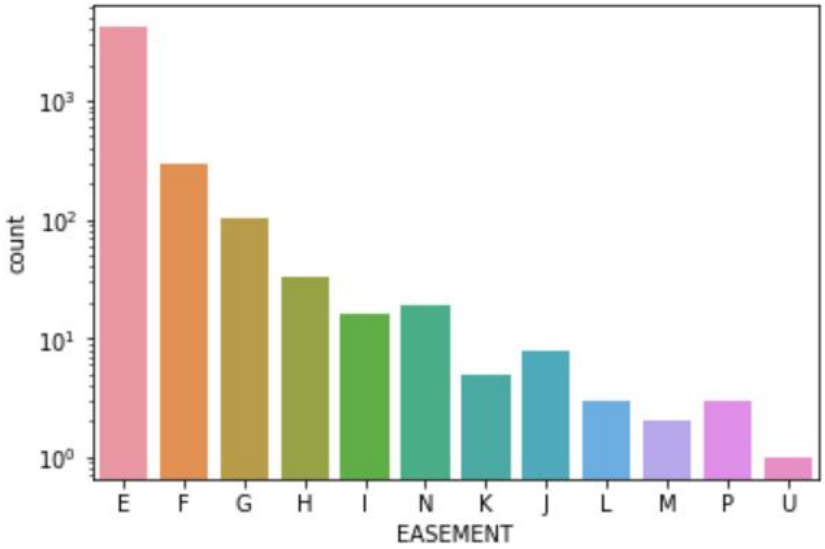
Most common field value: 1

EASEMENT:

Name: “EASEMENT

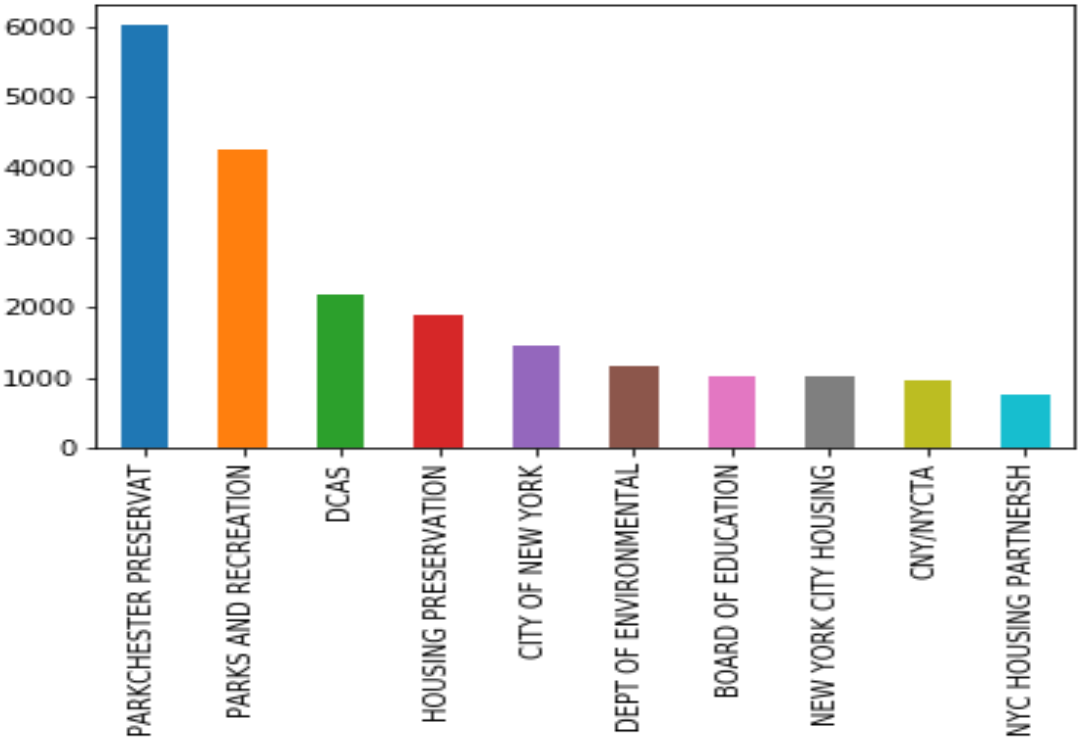
Field type: Categorical

Description: Field that describe easement



Most common field value: E

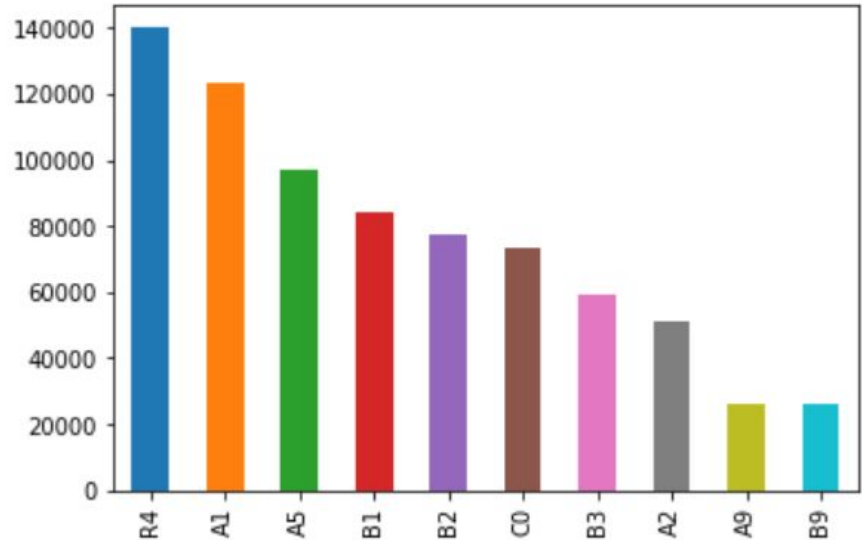
OWNER:
Name: "OWNER
Field type: Categorical
Description: Name and company of that owns the property



Most common field value: PARKCHESTER PRESERVAT

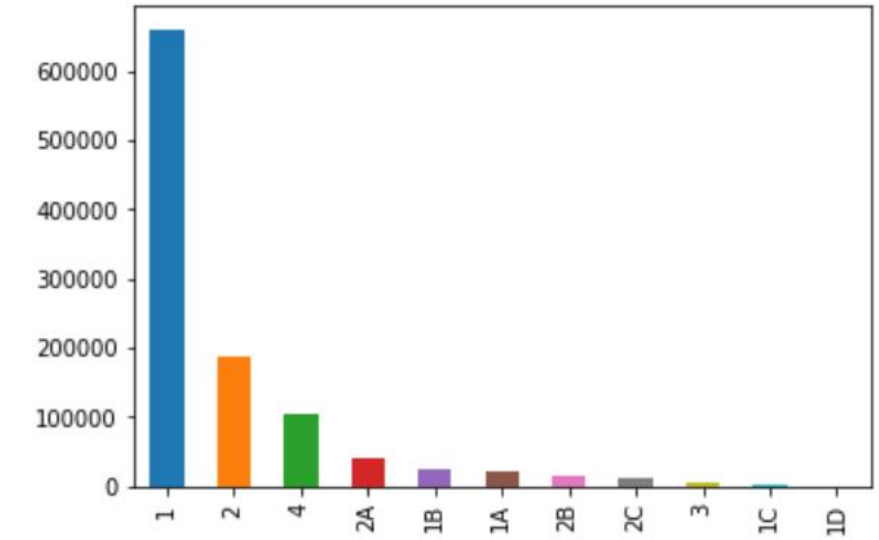
BLDGCL:

Name: “BLDGCL”
Field type: Categorical
Description: N/A



Most common field value: R4

TAXCLASS:
Name: “TAXCLASS”
Field type: Categorical
Description: Code of the property tax class

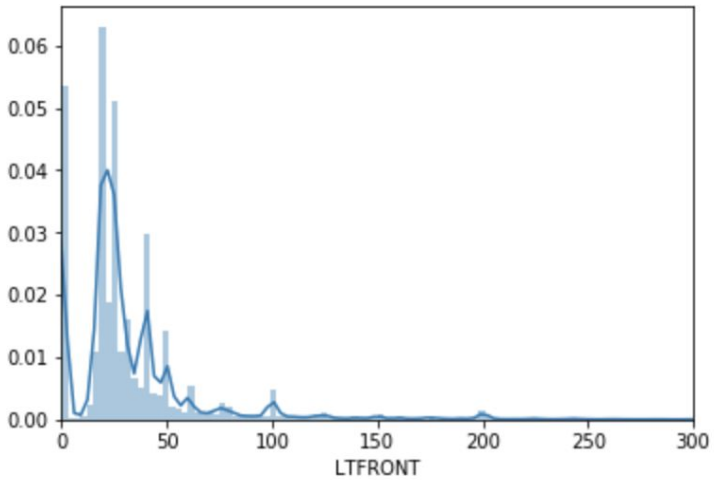


Most common field value: 1

LTFRONT:
Name: “LTFRONT”
Field type: Numeric
Description: Lot Frontage in feet
Statistical facts:

```
count    1.070994e+06
mean     3.663530e+01
std      7.403284e+01
min      0.000000e+00
25%      1.900000e+01
50%      2.500000e+01
75%      4.000000e+01
max      9.999000e+03
```

Distribution:



LTDEPTH:

Name: "LTDEPTH"

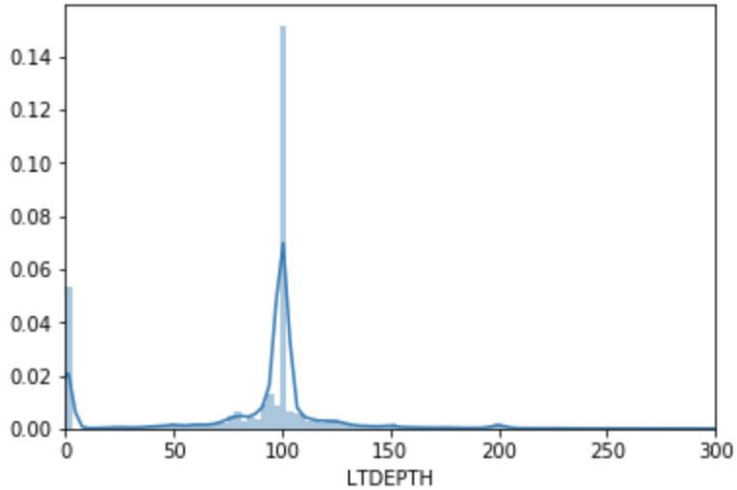
Field type: Numeric

Description: Lot Depth in feet

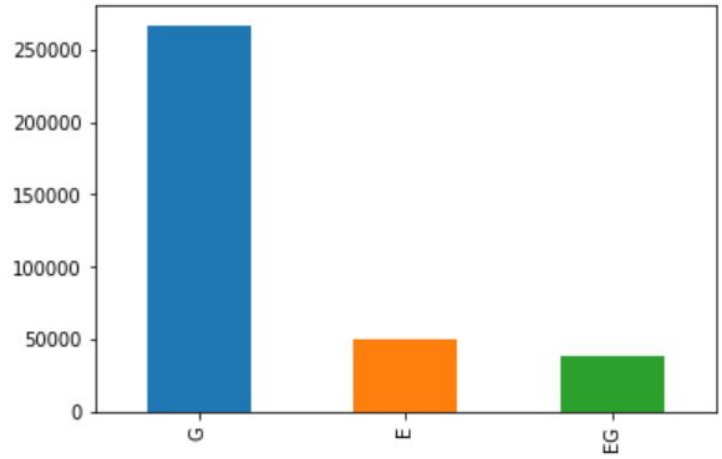
Statistical facts:

```
count    1.070994e+06
mean     8.886159e+01
std      7.639628e+01
min      0.000000e+00
25%      8.000000e+01
50%      1.000000e+02
75%      1.000000e+02
max      9.999000e+03
```

Distribution:



EXT:
Name: “EXT”
Field type: Categorical
Description: N/A

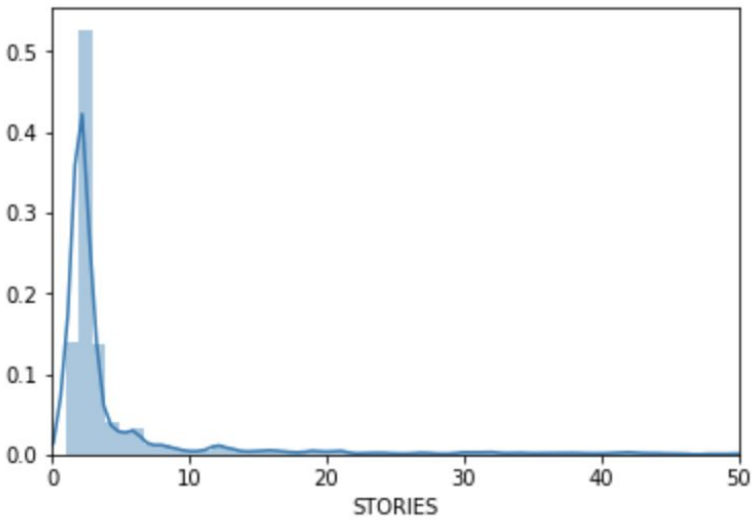


Most common field value: G

STORIES:
Name: “STORIES”
Field type: Numeric
Description: Number of stories in each area
Statistical facts:

count	1.014730e+06
mean	5.006918e+00
std	8.365707e+00
min	1.000000e+00
25%	2.000000e+00
50%	2.000000e+00
75%	3.000000e+00
max	1.190000e+02

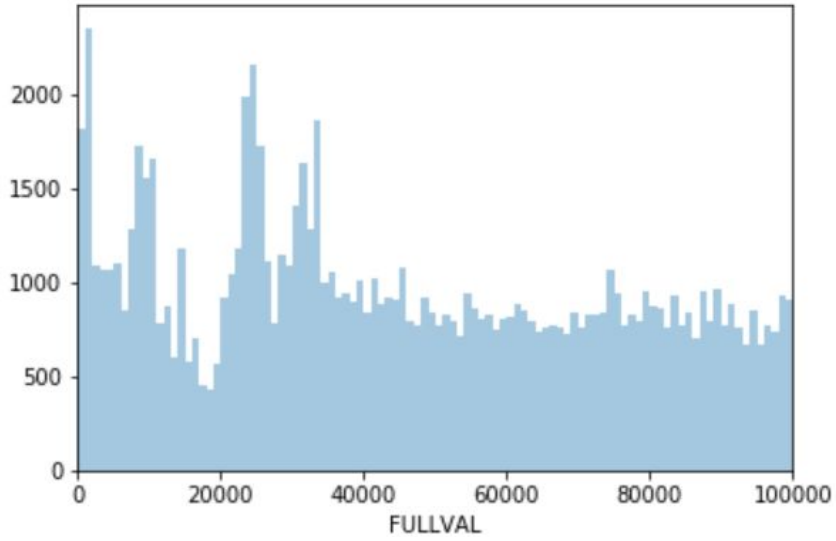
Distribution:



FULLVAL:
Name: “FULLVAL”
Field type: Numeric
Description: Full value of the property
Statistical facts:

count	1.070994e+06
mean	8.742645e+05
std	1.158243e+07
min	0.000000e+00
25%	3.040000e+05
50%	4.470000e+05
75%	6.190000e+05
max	6.150000e+09

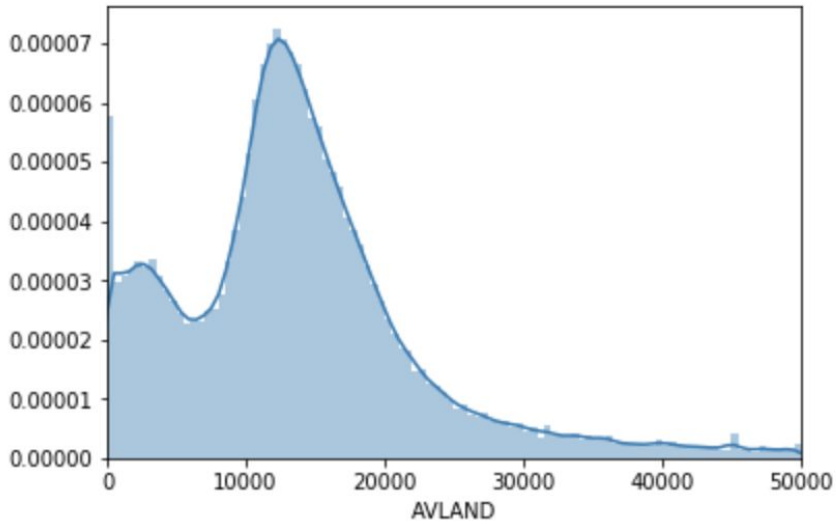
Distribution:



AVLAND:
Name: “AVLAND”
Field type: Numeric
Description:
Statistical facts:

count	1.070994e+06
mean	8.506792e+04
std	4.057260e+06
min	0.000000e+00
25%	9.180000e+03
50%	1.367800e+04
75%	1.974000e+04
max	2.668500e+09

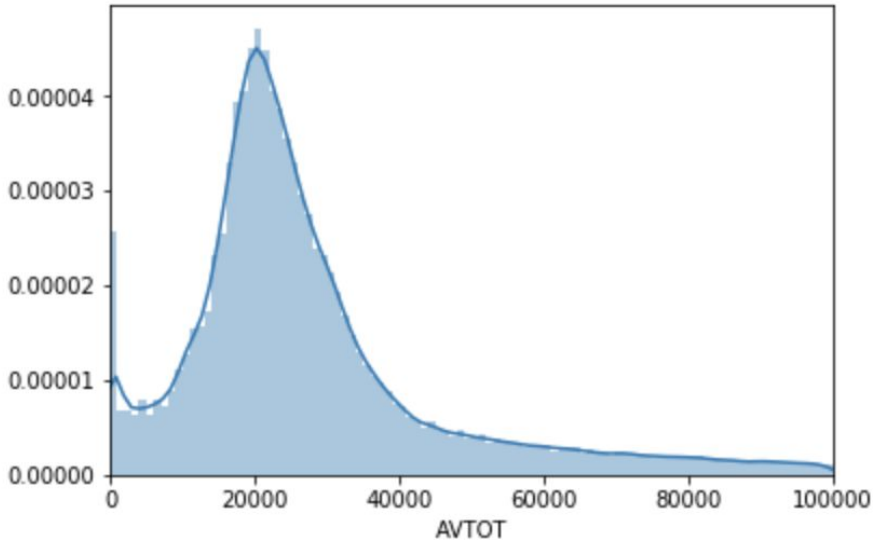
Distribution:



AVTOT:
Name: "AVTOT"
Field type: Numeric
Description: N/A
Statistical facts:

count	1.070994e+06
mean	2.272382e+05
std	6.877529e+06
min	0.000000e+00
25%	1.837400e+04
50%	2.534000e+04
75%	4.543800e+04
max	4.668309e+09

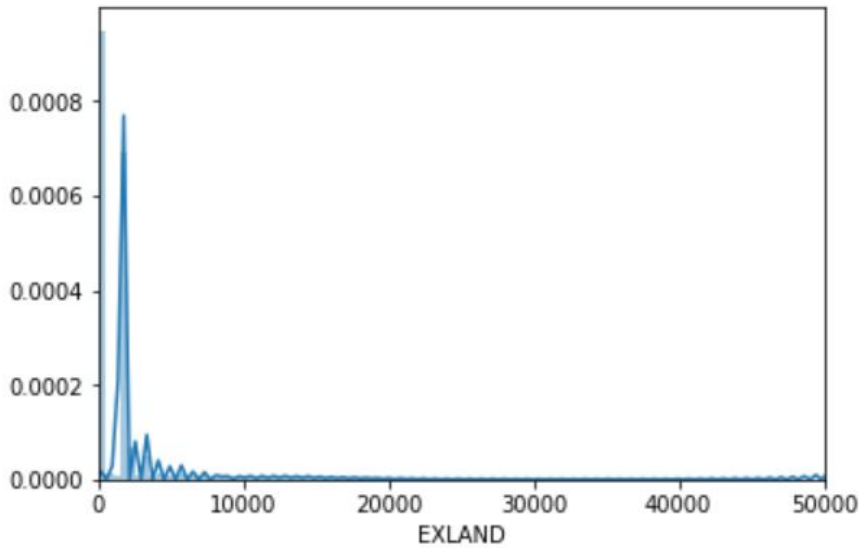
Distribution:



EXLAND:
Name: “EXLAND”
Field type: Numeric
Description: Current Transitional Exempt Land Value
Statistical facts:

count	1.070994e+06
mean	3.642389e+04
std	3.981576e+06
min	0.000000e+00
25%	0.000000e+00
50%	1.620000e+03
75%	1.620000e+03
max	2.668500e+09

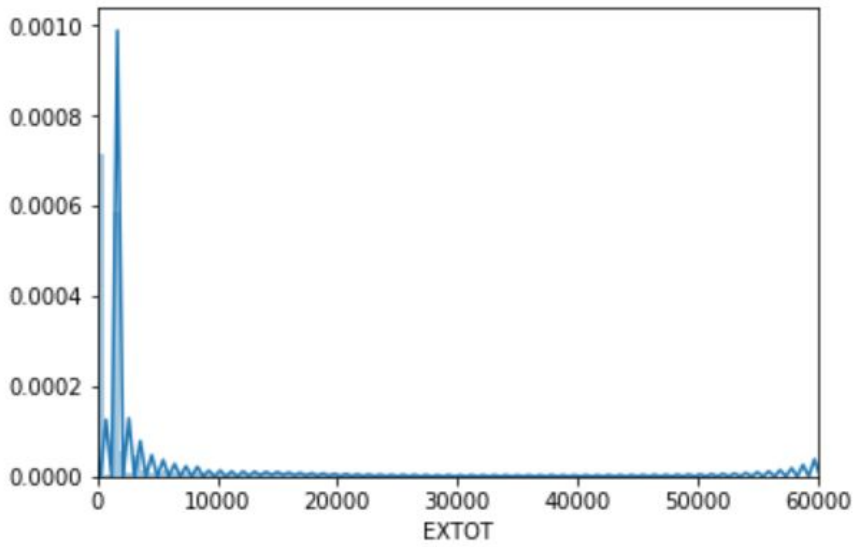
Distribution:



EXTOT:
Name: “EXTOT”
Field type: Numeric
Description: Transitional Exempt Total Value
Statistical facts:

count	1.070994e+06
mean	9.118698e+04
std	6.508403e+06
min	0.000000e+00
25%	0.000000e+00
50%	1.620000e+03
75%	2.090000e+03
max	4.668309e+09

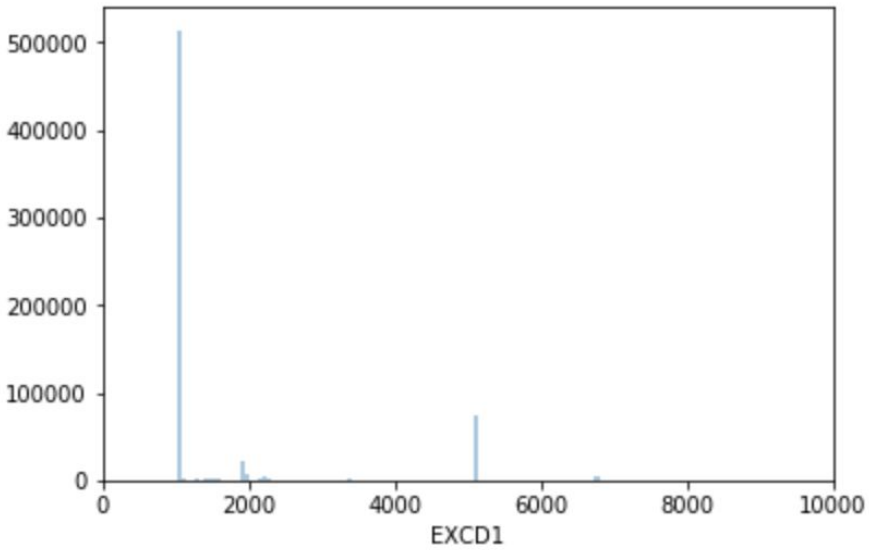
Distribution:



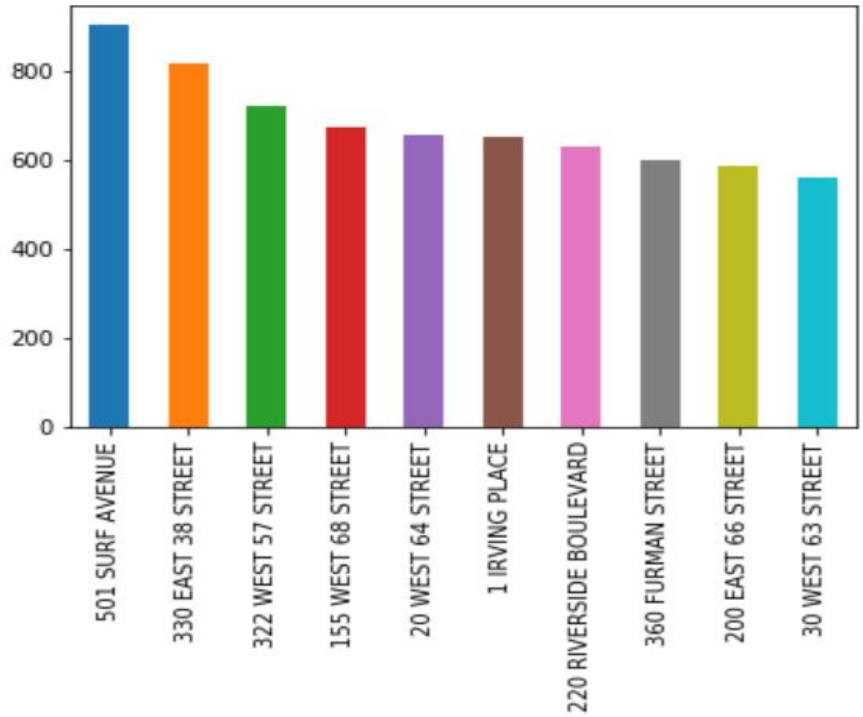
EXCD1:
Name: "EXCD1"
Field type: Numeric
Description:
Statistical facts:

count	638488.000000
mean	1602.014232
std	1384.226741
min	1010.000000
25%	1017.000000
50%	1017.000000
75%	1017.000000
max	7170.000000

Distribution:

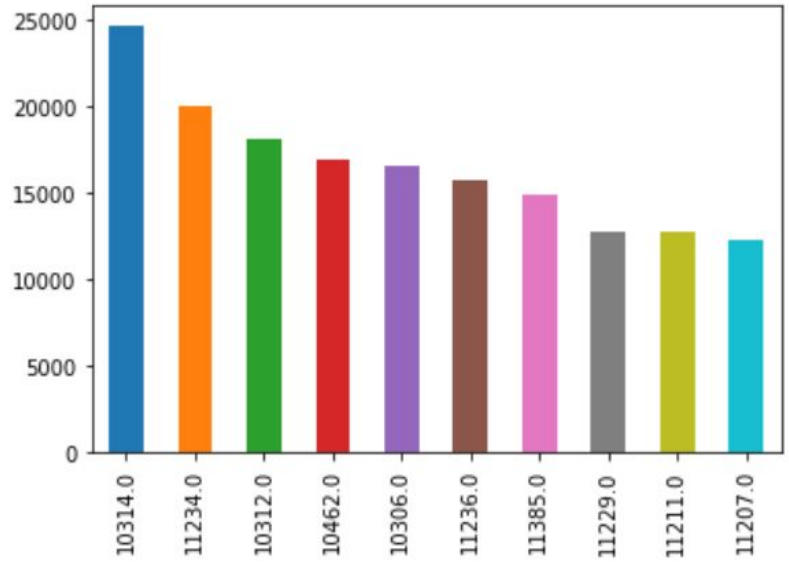


STADDR:
Name: “STADDR”
Field type: Categorical
Description: Street name of the property



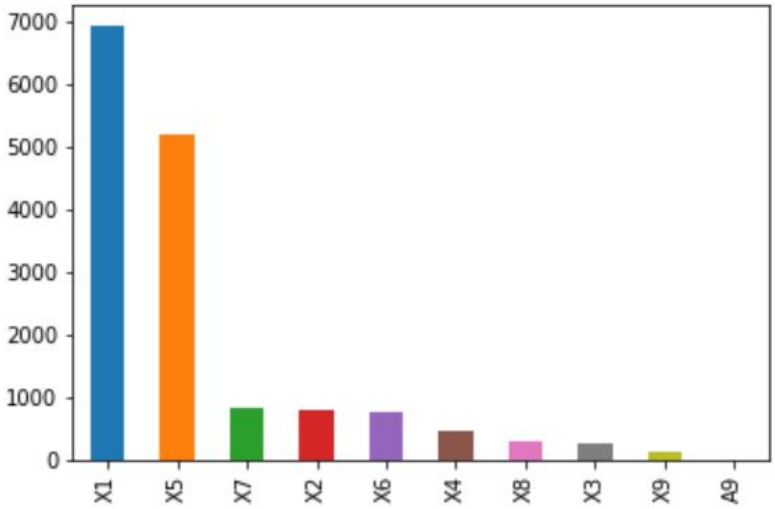
Most common field value: 501 SURF AVENUE

ZIP:
Name: “ZIP”
Field type: Categorical
Description: Postal Zip code of the property



Most common field value: 10314.0

EXMTCL:
Name: “EXMTCL”
Field type: Categorical
Description: Exempt class for fully exempt properties.

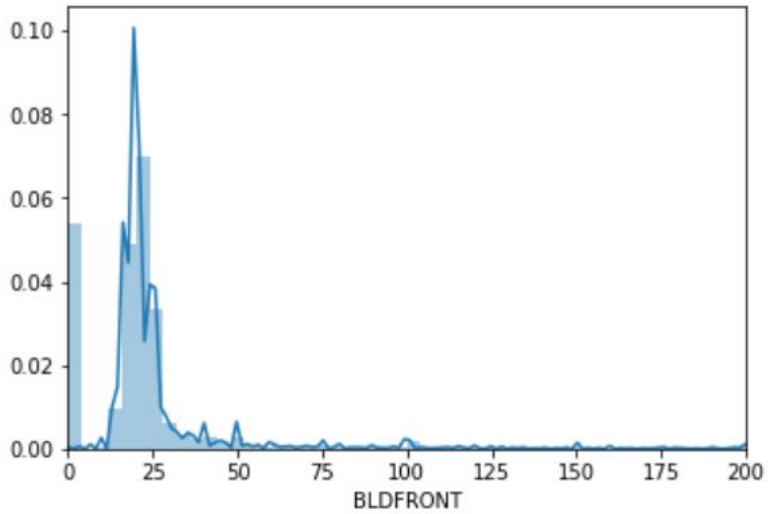


Most common field value: X1

BLDFRONT:
Name: ”BLDFRONT”
Field type: Numeric
Description: N/A
Statistical facts:

count	1.070994e+06
mean	2.304277e+01
std	3.557970e+01
min	0.000000e+00
25%	1.500000e+01
50%	2.000000e+01
75%	2.400000e+01
max	7.575000e+03

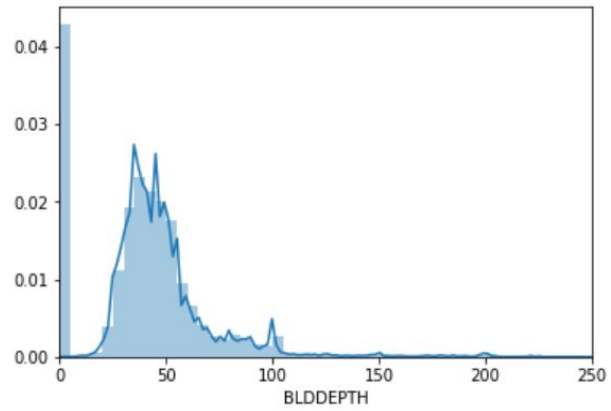
Distribution:



BLDDEPTH:
Name: “BLDDEPTH”
Field type: Numeric
Description: N/A
Statistical facts:

count	1.070994e+06
mean	3.992284e+01
std	4.270715e+01
min	0.000000e+00
25%	2.600000e+01
50%	3.900000e+01
75%	5.000000e+01
max	9.393000e+03

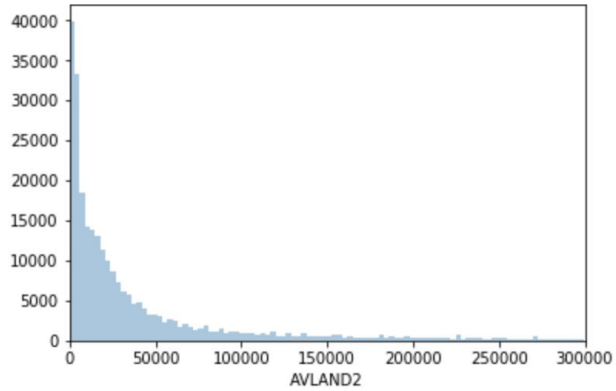
Distribution:



AVLAND2:
Name: “AVLAND2”
Field type: Numeric
Description: N/A
Statistical facts:

count	2.827260e+05
mean	2.462357e+05
std	6.178963e+06
min	3.000000e+00
25%	5.705000e+03
50%	2.014500e+04
75%	6.264000e+04
max	2.371005e+09

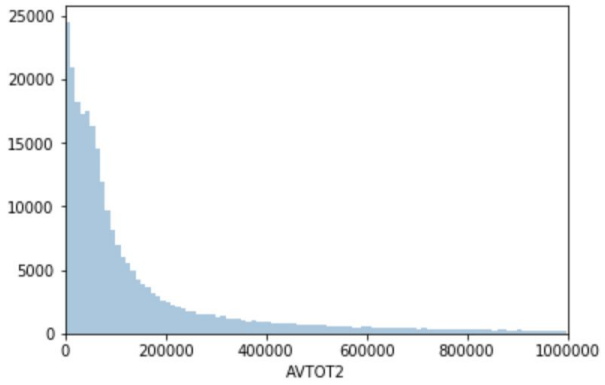
Distribution:



AVTOT2:
Name: “AVTOT2”
Field type: Numeric
Description: N/A
Statistical facts:

count	2.827320e+05
mean	7.139114e+05
std	1.165253e+07
min	3.000000e+00
25%	3.391200e+04
50%	7.996250e+04
75%	2.405510e+05
max	4.501180e+09

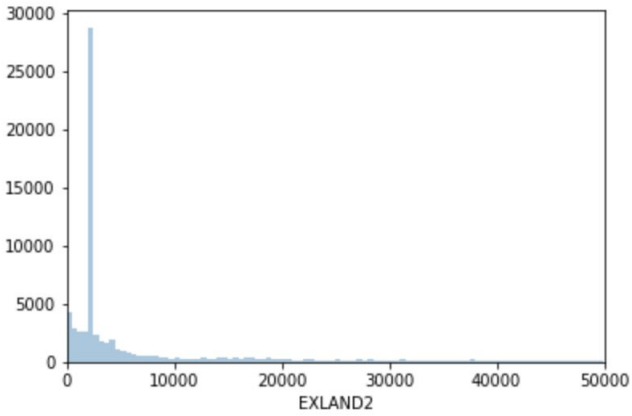
Distribution:



EXLAND2:
Name: “EXLAND2”
Field type: Numeric
Description:
Statistical facts:

count	8.744900e+04
mean	3.512357e+05
std	1.080221e+07
min	1.000000e+00
25%	2.090000e+03
50%	3.048000e+03
75%	3.177900e+04
max	2.371005e+09

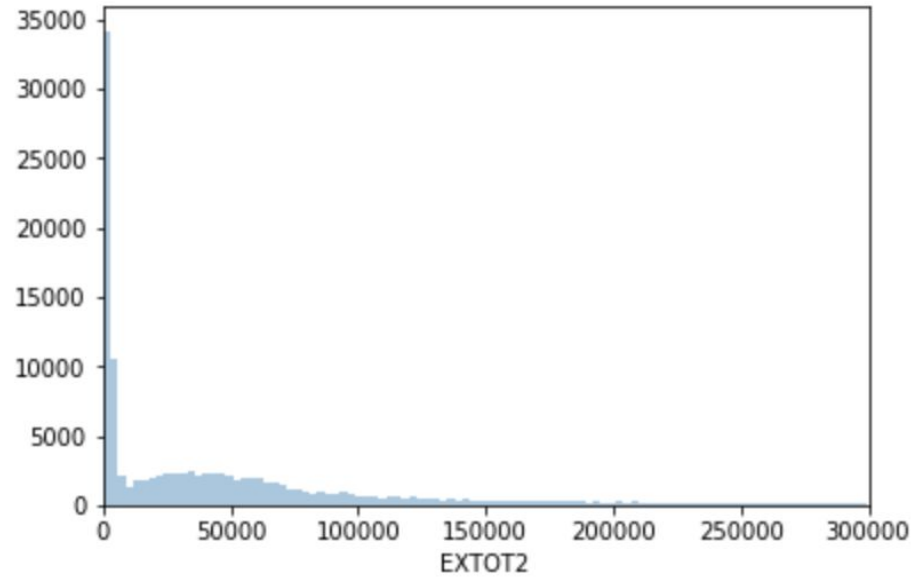
Distribution:



EXTOT2:
Name: “EXTOT2”
Field type: Numeric
Description:
Statistical facts:

```
count     1.308280e+05
mean      6.567683e+05
std       1.607251e+07
min       7.000000e+00
25%       2.870000e+03
50%       3.706200e+04
75%       1.068408e+05
max       4.501180e+09
```

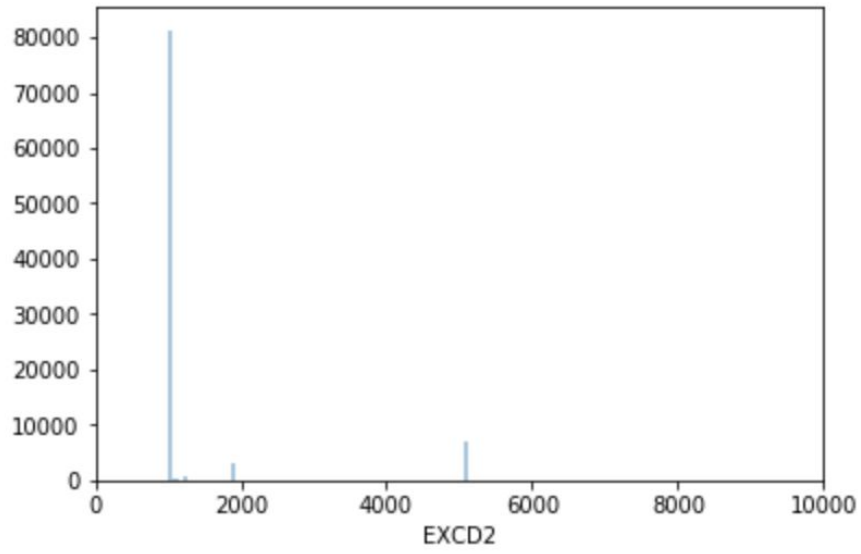
Distribution:



EXCD2:
Name: “EXCD2”
Field type: Numeric
Description:
Statistical facts:

```
count    92948.000000
mean      1364.041679
std       1094.705653
min       1011.000000
25%       1017.000000
50%       1017.000000
75%       1017.000000
max       7160.000000
```

Distribution:



PERIOD:
Name: “PERIOD”
Field type: Categorical
Description: : indicator for the change period of the file
All numbers in this field are same, the table of data in this field is shown as below.
Value= FINAL, no missing data

PERIOD	
FINAL	1070994

YEAR:
Name: “YEAR”
Field type: Date/Time
Description:
Value= 2011/11, no missing data

VALTYPE:
Name: “VALTYPE”
Field type: Categorical
Description:
All numbers in this field are same, the table of data in this field is shown as below.
Value= AC-TR, no missing data

VALTYPE	
AC-TR	1070994