FRAUD ANALYTICS

PROJECT 1

New York Property

ABSTRACT

Design an unsupervised fraud model to find anomalies in New York property data

Team 9

Anni Cai, Suraj Patel, Yuyao Shen, Nanchun Shi, Bingru Xue

CONTENTS

EXECUTIVE SUMMARY

Team 9 proposed to design an unsupervised fraud detection algorithm at the request of the City of New York. The goal for this project is to find unusual valuations: Total Market Value (FULLVAL), Actual Land Value (AVLAND) and Actual Total Value (AVTOT) from "NY property data" dataset which has 1,070,994 records and 32 variables.

After imputation for missing values, we created new variables to best represent these value fields to look for anomalies. For example, we normalized value fields by lot area, building area and building volume and grouped them by five different levels of granularity and after that, we had 45 variables. Then, we implemented Principal Component Analysis to remove correlations and reduce dimensions.

Upon the completion of data preparation, we applied two outlier detection methods (Heuristic Function and Autoencoder) to generate an abnormal score for each record and normalized it by taking its rank order then combined these two scores by taking their average to get final anomaly score for each record.

As a result, we have a list of records sorted by its abnormality and we recommend the City of New York to start investigating from top 10.

Description of data

This dataset is from New York City OpenData, which includes the valuation and assessment of properties in New York City assessed in 2010 and 2011. It has 1070994 records and 32 variables.

Numeric Fields

| Field name	Field type	# of non-NA values	% populated	# unique values	# '0' records	Mean	Std	Min	Max
LTFRONT	numeric	1,070,994	100	1,297	169,108	36.64	74.03	0	9,999
LTDEPTH	numeric	1,070,994	100	1,370	170,128	88.86	76.4	0	9,999
STORIES	numeric	1,014,730	94.75	111	0	5.01	8.37	1	119
FULLVAL	numeric	1,070,994	100	109,324	13,007	874,264.51	11,582,430.99	0	6,150,000,000
AVLAND	numeric	1,070,994	100	70,921	13,009	85,067.92	4,057,260.06	0	2,668,500,000
AVTOT	numeric	1,070,994	100	112,914	13,007	227,238.17	6,877,529.31	0	4,668,308,947
EXLAND	numeric	1,070,994	100	33,419	491,699	36,423.89	3,981,575.79	0	2,668,500,000
EXTOT	numeric	1,070,994	100	64,255	432,572	91,186.98	6,508,402.82	0	4,668,308,947
BLDFRONT	numeric	1,070,994	100	612	228,815	23.04	35.58	0	7,575
BLDDEPTH	numeric	1,070,994	100	621	228,853	39.92	42.71	0	9,393
AVLAND2	numeric	282,726	26.40	58,591	0	246,235.72	6,178,962.56	3	2,371,005,000
AVTOT2	numeric	282,732	26.40	111,360	0	713,911.44	11,652,528.95	3	4,501,180,002
EXLAND2	numeric	87,449	8.17	22,195	0	351,235.68	10,802,212.67	1	2,371,005,000
EXTOT2	numeric	130,828	12.22	48,348	0	656,768.28	16,072,510.17	7	450,1180,002

Categorical Fields

| Field name	Field type	# of non-NA values	% populated	# unique values	Most Common Field
Record	categorical	1,070,994	100	1,070,994	NA
BBLE	categorical	1,070,994	100	1,070,994	NA
B	categorical	1,070,994	100	5	4
BLOCK	categorical	1,070,994	100	13,984	3944
LOT	categorical	1,070,994	100	6,366	1
EASEMENT	categorical	4,636	0.43	12	E
OWNER	categorical	1,039,249	97.04	863,346	PARKCHESTER PRESERVAT
BLDGCL	categorical	1,070,994	100	200	R4
TAXCLASS	categorical	1,070,994	100	11	1
EXT	categorical	354,305	33.08	3	G
EXCD1	categorical	638,488	59.62	129	1017
STADDR	categorical	1,070,318	99.94	839,280	501 SURF AVENUE
ZIP	categorical	1,041,104	97.21	196	10314
EXMPTCL	categorical	15,579	1.45	14	X1
EXCD2	categorical	92,948	8.68	60	1017
PERIOD	categorical	1,070,994	100	1	FINAL
YEAR	date/time	1,070,994	100	1	2010/11
VALTYPE	categorical	1,070,994	100	1	AC-TR

Important Distributions/Histograms:

FULLVAL

Distribution of FULLVAL
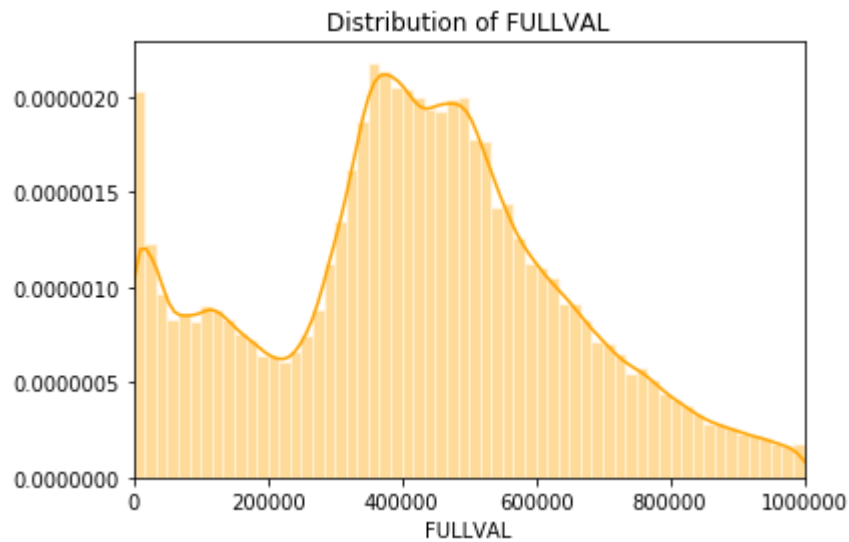
AVTOT
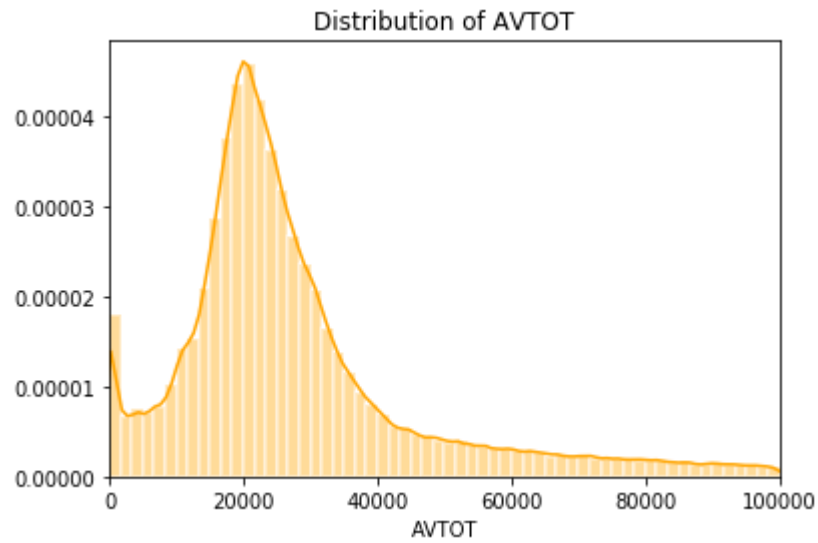
AVLAND
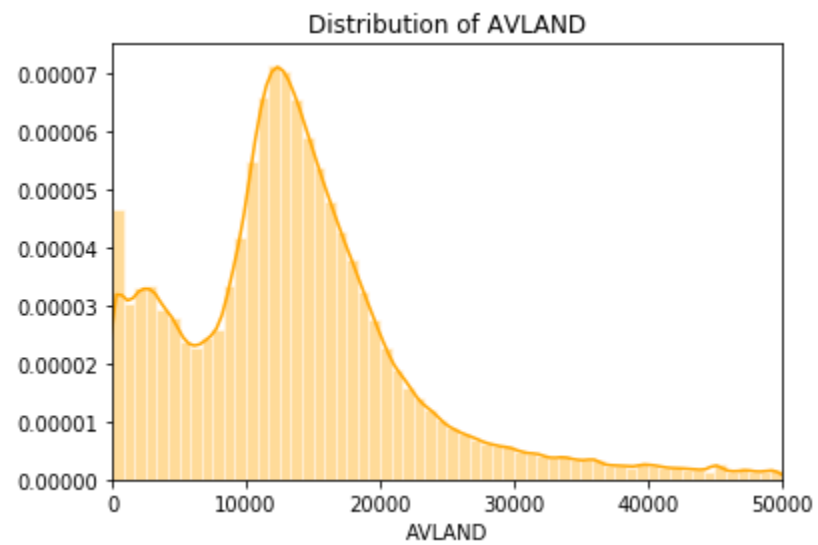
DATA CLEANING

We first imputed **ZIP** because zip code is the prerequisite for imputing other variables. We grouped the ZIP data by borough and block and distinguished two kinds of all groups: one with all null values, and the other with at least one non-null value. For the groups with all null values, we filled in their zip codes based on their street addresses. If the street address doesn't exist as well, we used the minimum zip code in that borough. For the groups with at least one non-null value, we used the most common zip code in the group to fill in.

For **FULLVAL, BLDFRONT, BLDDEPTH, LTFRONT, LTDEPTH** and **STORIES**, we first grouped by zip code and building class and moved along in the following order if there existed a group with less than 5 records:

1. Group by zip code and building class (ZIP, BLDGCL)
2. Group by zip code and tax class (ZIP, TAXCLASS)
3. Group by zip code only (ZIP)
4. Group by borough (B)

We filled in missing fields with the median value of that group.

We believed AVTOT is highly related to building size, therefore we derived the following formula to calculate building value per square foot (BVPSF) for the imputation of AVTOT:

$$BVPSF_i = \frac{AVTOT_i}{BLDFRONT_i * BLDDEPTH_i * STORIES_i}$$

Since we already filled in BLDFRONT, BLDDEPTH and STORIES, we were ready to calculate BVPSF. Again, we aggregated data into groups, and we set different levels of granularity to avoid data scarcity (less than 5 records in a group) as follow:

1. Group by zip code and building class (ZIP, BLDGCL)
2. Group by zip code and tax class (ZIP, TAXCLASS)
3. Group by zip code (ZIP)
4. Group by borough (B)

We took the median BVPSF in each group and multiplied by estimating property's useable building area (BLDFRONT*BLDDEPTH*STORY) to get its estimated AVTOT.

For **AVLAND**, we used the same logic for the hierarchy of grouping as well as for dealing with data scarcity. We first calculated land value per square foot (LVPSF) for each record:

$$LVPSF_i = \frac{AVLAND_i}{LTFRONT_i * LTDEPTH_i}$$

Then for each record with missing AVLAND value, we use the median of its proper group and multiplied by its LTFRONT and LTDEPTH to obtain an estimated land value for the record.

VARIABLE CREATION

Firstly, we created 3 size variables:

1. LOTAREA (S1) = LTFRONT * LTDEPTH
2. BLDAREA (S2) = BLDFRONT * BLDDEPTH
3. BLDVOL (S3) = BLDAREA * STORIES

Then we normalized 3 value variables (V1 = FULLVAL, V2 = AVLAND, V3 = AVTOT) using these 3 size variables above:

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

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

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

After that, we separated records into 5 groups: 3-digit zip code, 5-digit zip code, tax class, borough and all. For each group g, we calculated r_{ig}, the average of each r_i for each group g. And for each record, we calculated and appended 45 variables:

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

DIMENSIONALITY REDUCTION

After we created 45 ratios, we z-scaled the dataframe. Then we conduct Principal Component Analysis.

Below is the scree plot:
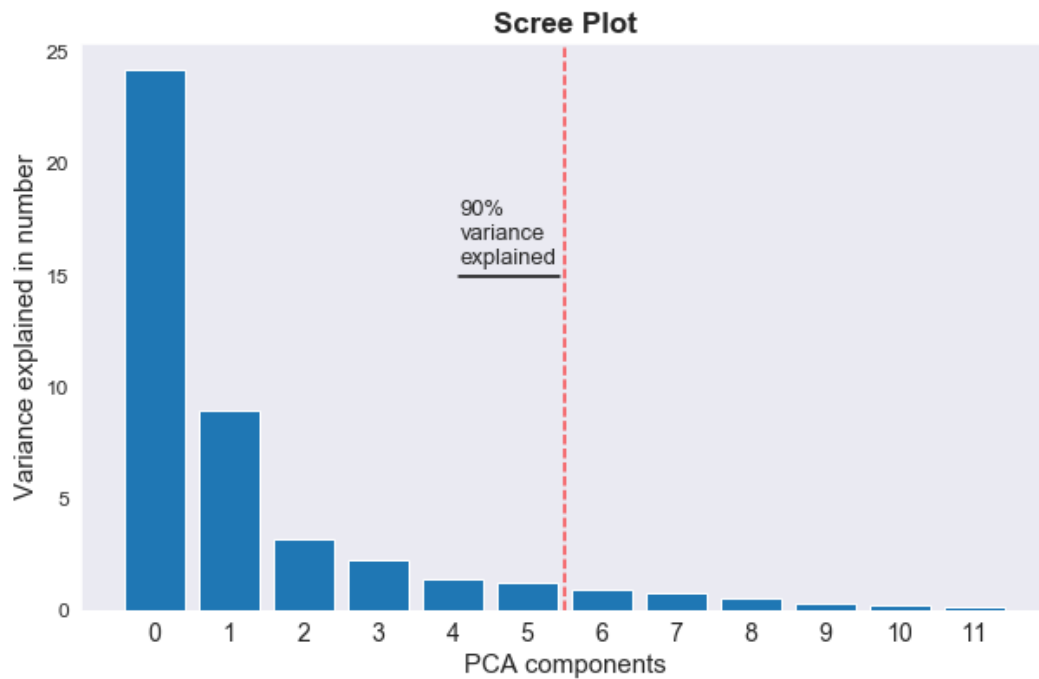
We chose the first 5 components that explained about 90% variance of the original z-scaled data.

Then we z-scale again and obtained the dataset with shape (1070994,5) ready for scoring.

ALGORITHMS

Heuristic Function:

We used the following function to calculate the first score:

$$s_i = \left(\sum_k |z_k^i|^n \right)^{1/n}, \quad n \text{ anything}$$

where z represents PC values for each record and here we chose n to be 2.

Autoencoder:

We built an autoencoder model from keras library, set one hidden layer with two nodes, trying to reproduce z-scaled PCs. Then we combined PCs with encoded PCs using the following function:

$$s_i = \left(\sum_k |z_k'^i - z_k^i|^n \right)^{1/n}, \quad n \text{ anything}$$

where z's represent PC and encoded PC, and here we chose n to be 2.

Combining two scores:

We rank score 1 and score 2 and created two rank order columns, then combined rank orders using the average. We sorted the dataset by final score and picked the top ten records that seem to be abnormal.

RESULTS

Below are the plots of two score distributions:

For the Heuristic Function:

Distribution of Heuristic Function Score

For the Autoencoder:

Distribution of Autoencoder Score
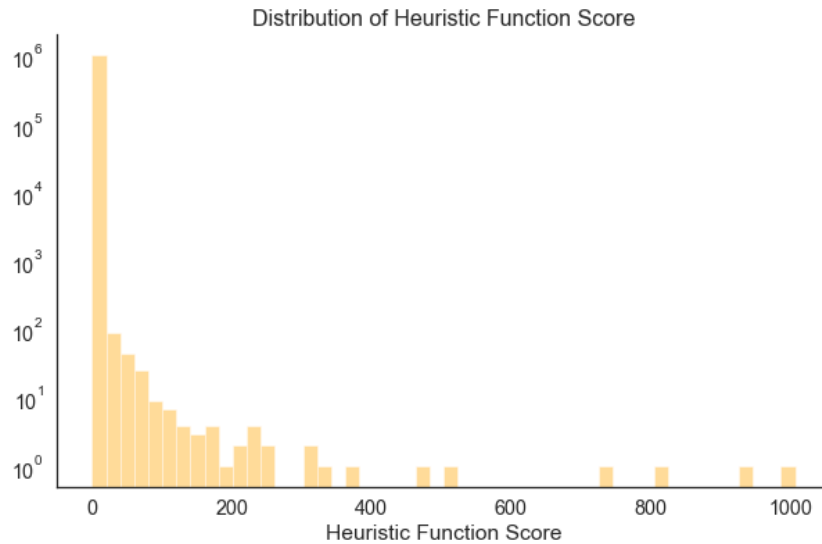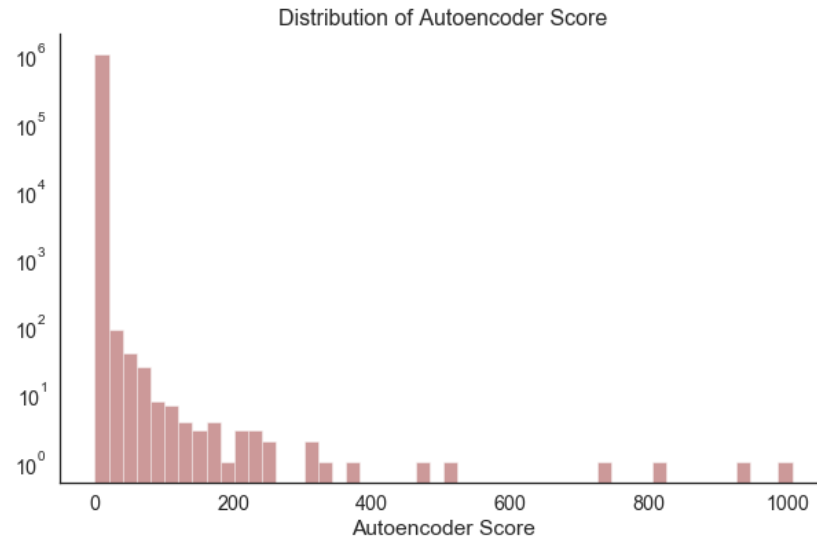
As explained in the previous section, we then combined the two scores for final anomaly detection.

The top 10 presumably abnormal records are as follows. We went back to the z-scaled ratios table as well as the original dataset to track down the potential anomalies:

RECORD	REASON FOR ANOMALIES
#632816	The z-scores of r2, r3, r5, r6, r8 and r9, which are related to building size and building volume, are extremely high and the reason is its BLDFRONT and BLDDEPTH are 1.
#565392	Both BLDFRONT and BLDDEPTH are below 50 percentiles of its group, but its FULLVAL, AVTOT and AVLAND are all above 75 percentiles of its group. We went back to the original data set and found that Its AVTOT, AVLAND, EXLAND and EXTOT are all 1,946,837,000. It might be inappropriate to fill in a median value for this record in the fields of BLDFRONT and BLDDEPTH.
#1067360	For this record, the z-scores of r1 (FULLVAL/ (LOTFRONT * LOTDEPTH)) and r4 (AVLAND/ (LOTFRONT * LOTDEPTH)), and r7 (AVTOT/ (LOTFRONT * LOTDEPTH)) are extremely high. These ratios are all related to lot size. So, we checked the relevant variables in the original dataset and found that LTFRONT =1, LTDEPTH = 1, which are abnormal.
#917942	For this record, the z-scores for r8 are considered high and abnormal because they are above 75 percentiles. Since r8 = v3/s2 = actual total value/ building area, we looked for anomalies in AVTOT, BLDFRONT or BLDDEPTH. We investigated and found that the LTDEPTH (100), BLDFRONT (40) and BLDDEPTH (60) are imputed by us, and their values are relatively small compared to LTFRONT (4910). This record is therefore marked as anomaly because the imputed values are not in consistency with LTFRONT in terms of scale. And this address turned out to be a 7-story Holiday Inn close to JFK.
#585118	For this record, the z-scores of r2 (FULLVAL/ (BLDFRONT * BLDDEPTH)) and r5(AVLAND/ (BLDFRONT * BLDDEPTH)), and r8 (AVTOT/ (BLDFRONT * BLDDEPTH)) are extremely high. These ratios are all related to building size. So, we checked the relevant variables in the original dataset and found that BLDFRONT =1, BLDDEPTH = 1, which are abnormal. Used the street address, to find that it is an empty plot with no building construction. Also, the FULLVAL, AVLAND and AVTOT have values higher than 75% of the dataset.
#585439	The z-scores of r2 (FULLVAL/(BLDFRONT * BLDDEPTH)) in all groups are very big and the reason is its BLDFRONT and BLDDEPTH are all 1's and itself is a 10-story 4-star hotel, which makes the z-score abnormally high.
#920628	Its z-scores for r2 (FULLVAL/ (BLDFRONT * BLDDEPTH)) in all groups are high because its BLDFRONT and BLDDEPTH are all 1's and itself is a nearly 4,000 square foot building size 2-story house.
#585120	The z-scores of r2 (FULLVAL/ (BLDFRONT * BLDDEPTH)) in all groups are very big and the reason is its BLDFRONT and BLDDEPTH are all 1's.
#565398	The z-scores for r2, r3, r5, r6, r8, r9 are all big. Its FULLVAL, AVLAND and AVTOT have values greater than 75% of this dataset. And its LTFRONT and LTDEPTH are also greater than 75% of the dataset. It might be inappropriate to fill in a median value for this record in the fields of BLDFRONT and BLDDEPTH.
#1067001	The z-scores of r2 (FULLVAL/ (BLDFRONT * BLDDEPTH)) in all groups are very big and the reason is its BLDFRONT and BLDDEPTH are all 1's.

CONCLUSIONS

To conclude, our team identified the top 10 anomaly records in "NY property data" dataset with unusual valuations by building an unsupervised fraud detection model. We imputed missing values, built 45 variables to represent values fields, scaled and minimized dimensionality by z-scaling and PCA, and normalized reduced variables again. Finally, we applied both Heuristic algorithm and Autoencoder to get anomaly scores and used rank ordering to scale each score. Combining the results by weighted average rank orders, we obtained top anomaly records.

More precise imputations have been left for the future due to lack of information and time. For example, properties with Tax Class Code "4" have high variations in both size and value variables. We need more detailed classification within this tax code group in order to impute more reasonable values.

Besides, many properties received high anomaly scores due to their extremely low size values. In the future work, we need to verify whether these records were misinputed during the data collection process, or they are attempted frauds.

APPENDIX - DATA QUALITY REPORT

TABLE OF CONTENTS

PART I - INTRODUCTION

The dataset of interest "NY property data", a subset with selected fields and data of the dataset "Property Valuation and Assessment Data", is compiled and updated annually by the Department of Finance, City of New York for purpose to calculate Property Tax, Grant eligible properties Exemptions and/or Abatements. The dataset is obtained from NYC OpenData website.

The dataset contains 32 different feature information of 1,070,994 properties in New York in November, 2010.

PART II - SUMMARY OF DATASET

Below are the summary tables for the 32 fields in the dataset:

I.   Numerical fields:

Field Name	# of Records	% Populated	# Unique Values	# Records w/ value 0	Mean	Standard Deviation	Max	Min
RECORD	1070994	100%	1070994	0	535497.5	309169.5	1070994	1
LTFRONT	1070994	100%	1297	169108	36.6353	74.03284	9999	0
LTDEPTH	1070994	100%	1370	170128	88.86159	76.39628	9999	0
STORIES	1014730	94.75%	111	0	5.006918	8.365707	119	1
FULLVAL	1070994	100%	109324	13007	874264.5	11582430	6150000000	0
AVLAND	1070994	100%	70921	13009	85067.92	4057260	2668500000	0
AVTOT	1070994	100%	112914	13007	227238.169	6877529	4668309000	0
EXLAND	1070994	100%	33419	491699	36423.89	3981576	2668500000	0
EXTOT	1070994	100%	64255	432572	91186.98	6508403	4668309000	0
BLDFRONT	1070994	100%	612	228815	23.042	35.5797	7575	0

Field Name	# of Records	% Populated	# Unique Values	# Records w/ value 0	Mean	Standard Deviation	Max	Min
BLDDEPTH	1070994	100%	621	228853	39.9228	42.707	9393	0
AVLAND2	282726	26.398%	58591	0	246235.719	6178963	2371005000	3
AVTOT2	282732	26.399%	111360	0	713911.436	11652530	4501180000	3
EXLAND2	87449	8.165%	22195	0	351235.684	10902210	2371005000	1
EXTOT2	130828	12.216%	48348	0	656768.282	16072510	4501180000	7

II.   Categorical fields:

Field Name	# of Records	% Populated	# Unique Values	Most Common Value
BBLE	1070994	100%	1070994	-
B	1070994	100%	5	4
BLOCK	1070994	100%	13984	3944
LOT	1070994	100%	6366	1
EASEMENT	4636	0.4329%	12	E
OWNER	1039249	97.0359%	863346	PARKCHESTER PRESERVAT
BLDGCL	1070994	100%	200	R4
TAXCLASS	1070994	100%	11	1
EXT	354305	33.082%	3	G
EXCD1	638488	59.616%	129	1017
STADDR	1070318	99.937%	839280	501 SURF AVENUE
ZIP	1041104	97.209%	196	10314
EXMPTCL	15579	1.455%	14	X1
EXCD2	92948	8.679%	60	1017
PERIOD	1070994	100%	1	FINAL
YEAR	1070994	100%	1	2010/11
VALTYPE	1070994	100%	1	AC-TR

PART III - FIELD DESCRIPTION

1.

Field name: RECORD

Field type: int64

Description: unique index number for each record, ascending from 1 to 1070994.

2.

Field name: BBLE

Field type: object

Description: Concatenation of BORO code, BLOCK code, LOT code, and EASEMENT code. Unique for each record.

3.

Field name: B

Field type: int64

Description: BOROUGH code; 1 = MANHATTAN, 2 = BRONX, 3 = BROOKLYN, 4 = QUEENS, 5 = STATEN ISLAND.
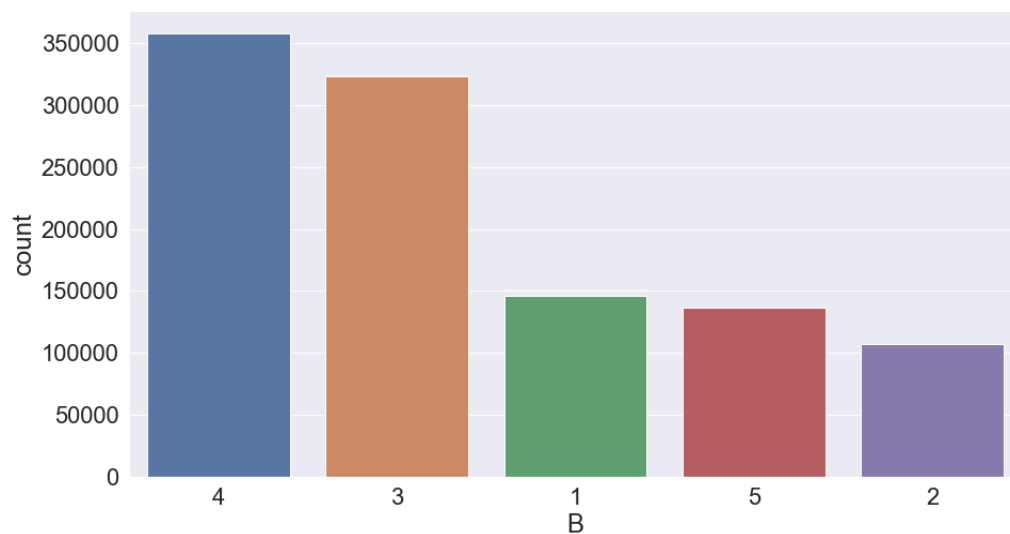
4.

Field name: BLOCK

Field type: int64

Description: valid block ranges by BORO; MANHATTAN 1 to 2,255, BRONX 2,260 to 5,958, BROOKLYN 1 to 8,955, QUEENS 1 to 16,350, STATEN ISLAND 1 to 8,050.

Manhattan:
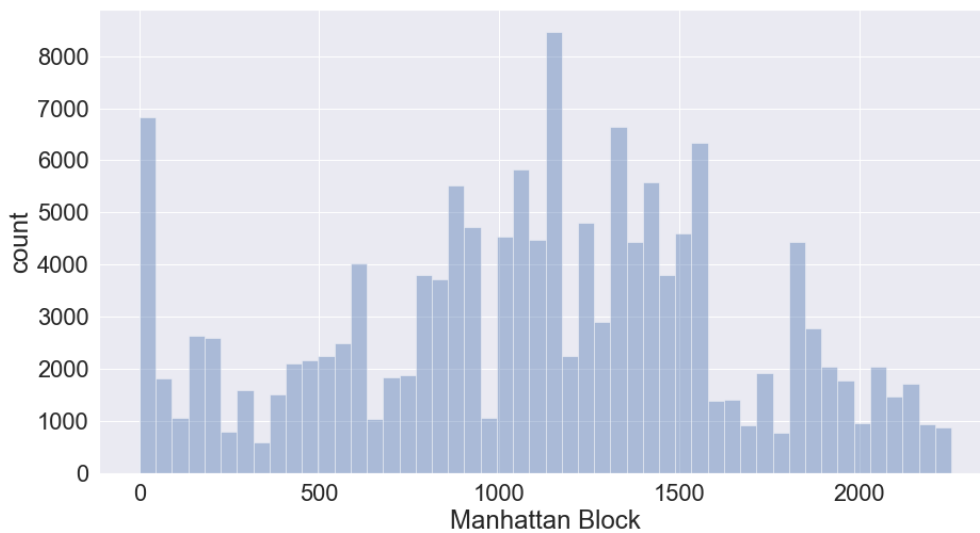
Bronx:
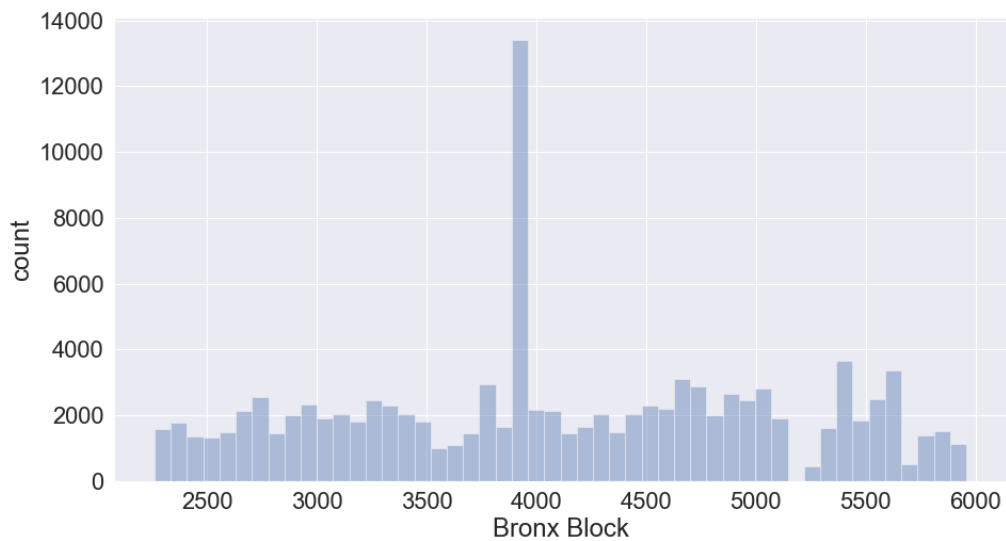
Brooklyn:
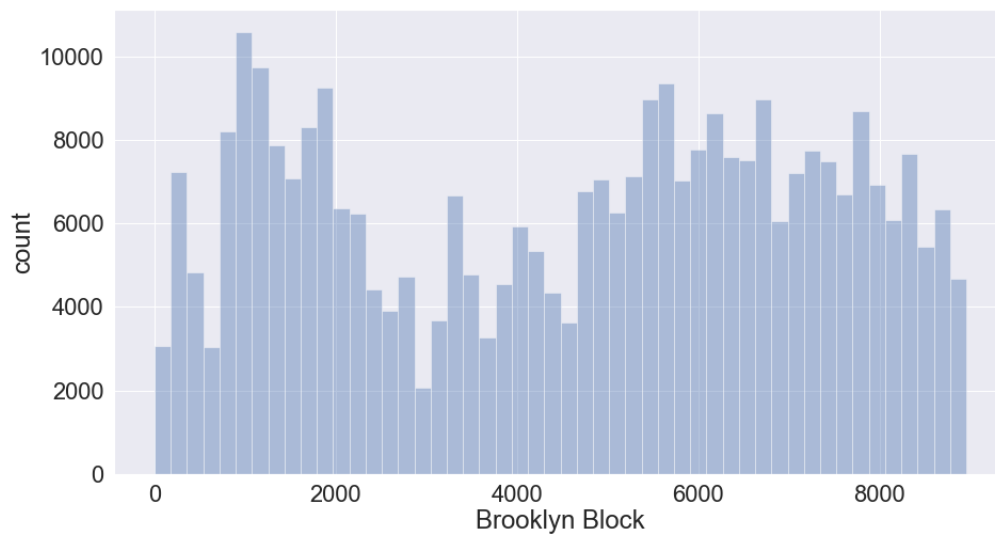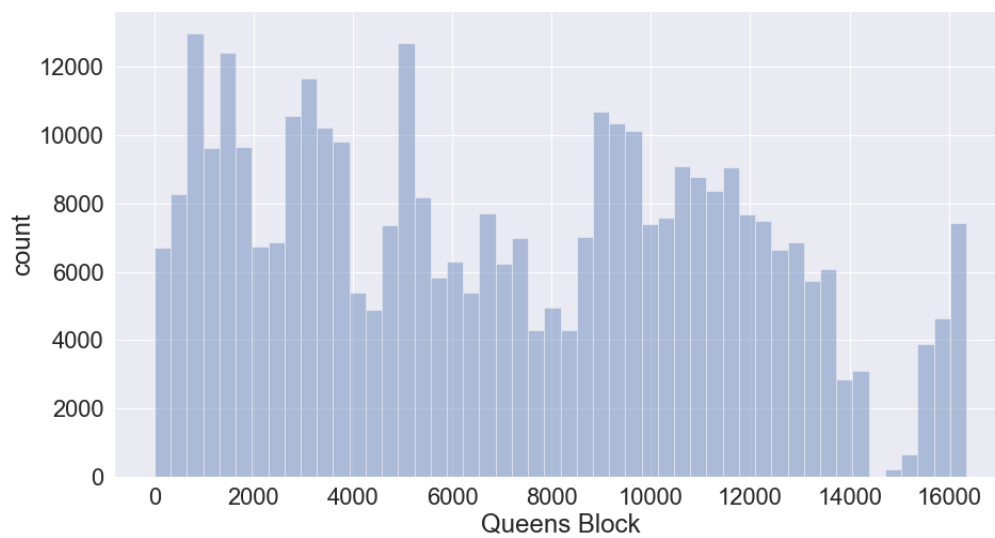
Queens:
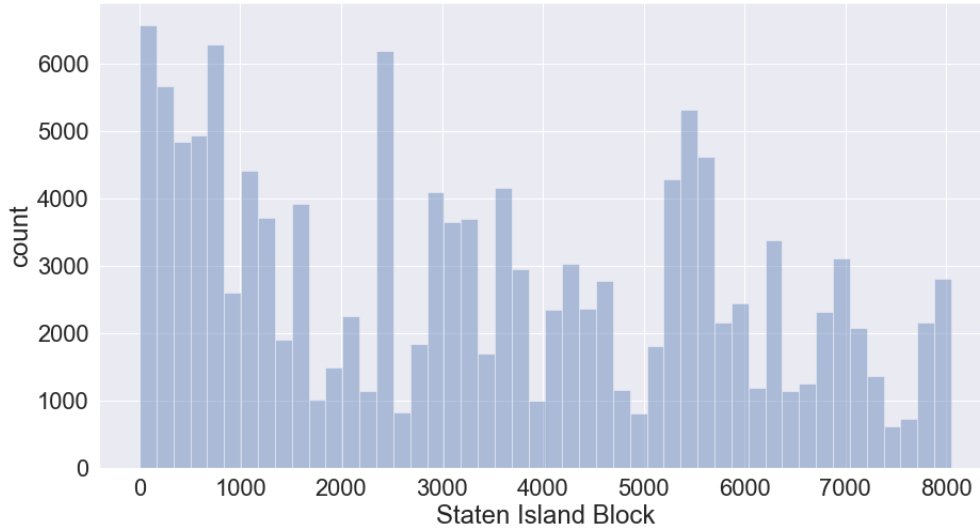
Staten Island:

5.

Field name: LOT

Field type: int64

Description: unique number within BORO/BLOCK

Manhattan:
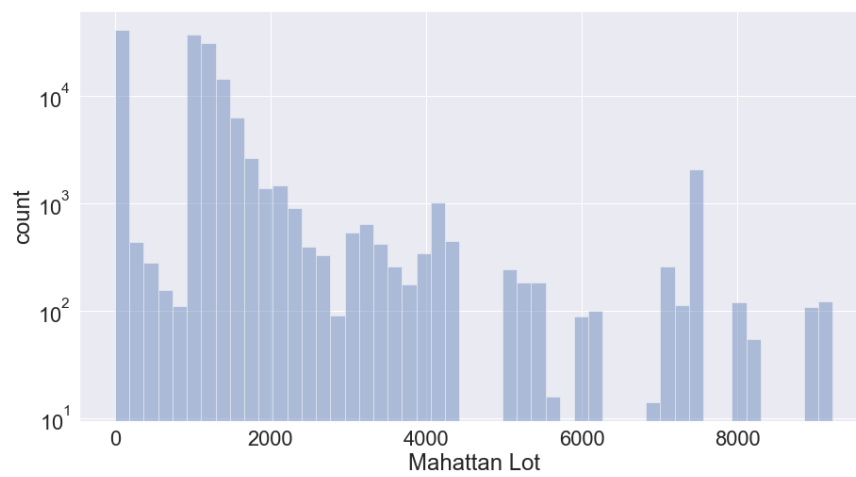
Bronx:
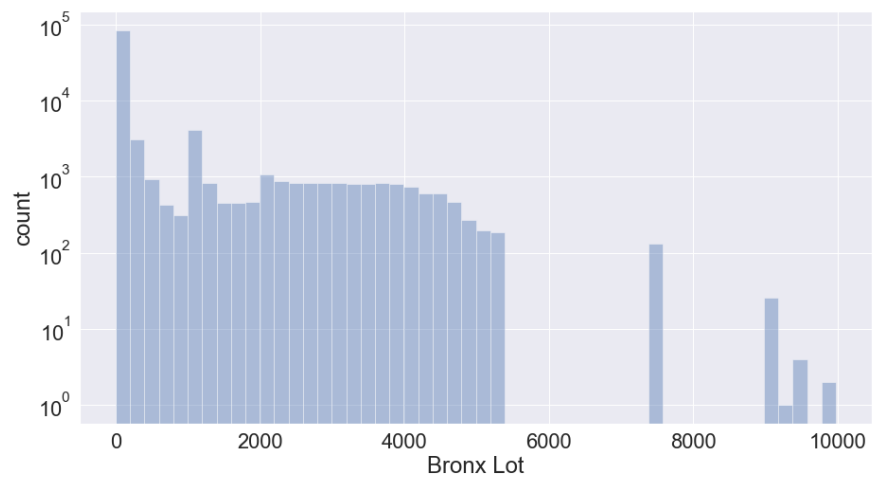
Brooklyn:
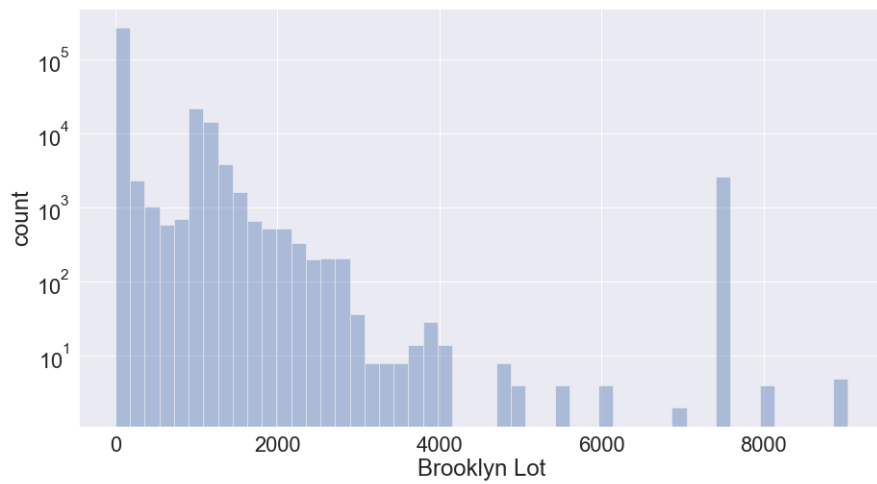
Queens:
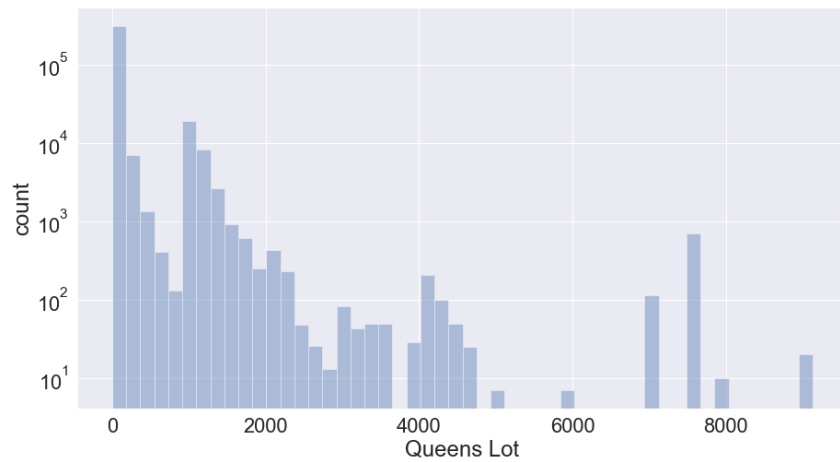
Staten Island:
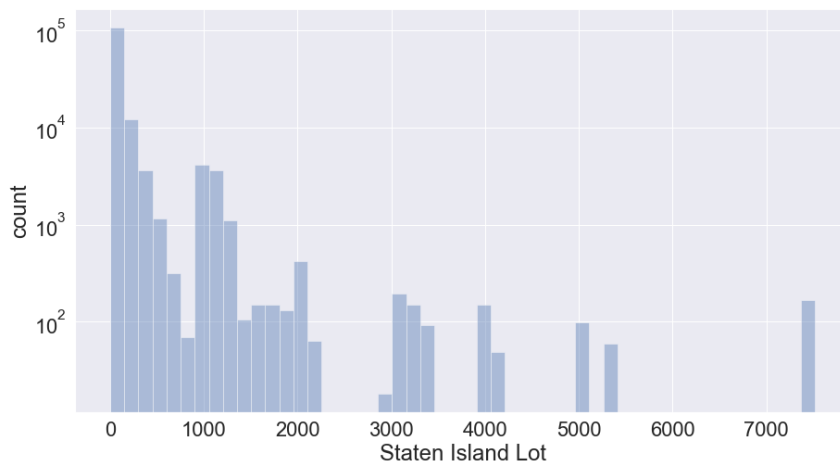
6.

Field name: EASEMENT

Field type: object

Description: SPACE Indicates the lot has no Easement.

'A' Indicates the portion of the Lot that has an Air Easement

'B' Indicates Non-Air Rights.

'E' Indicates the portion of the lot that has a Land Easement

'F' THRU 'M' Are duplicates of 'E'.

'N' Indicates Non-Transit Easement

'P' Indicates Piers.

'R' Indicates Railroads.

'S' Indicates Street

'U' Indicates U.S. Government
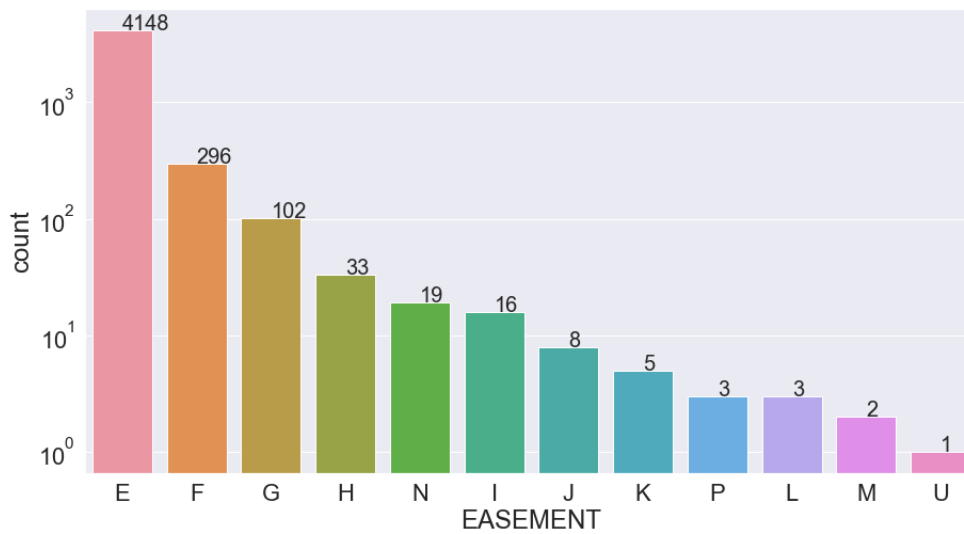
7.

Field name: OWNER

Field type: object

Description: owner names of the properties.

Top 25 owners with most properties:
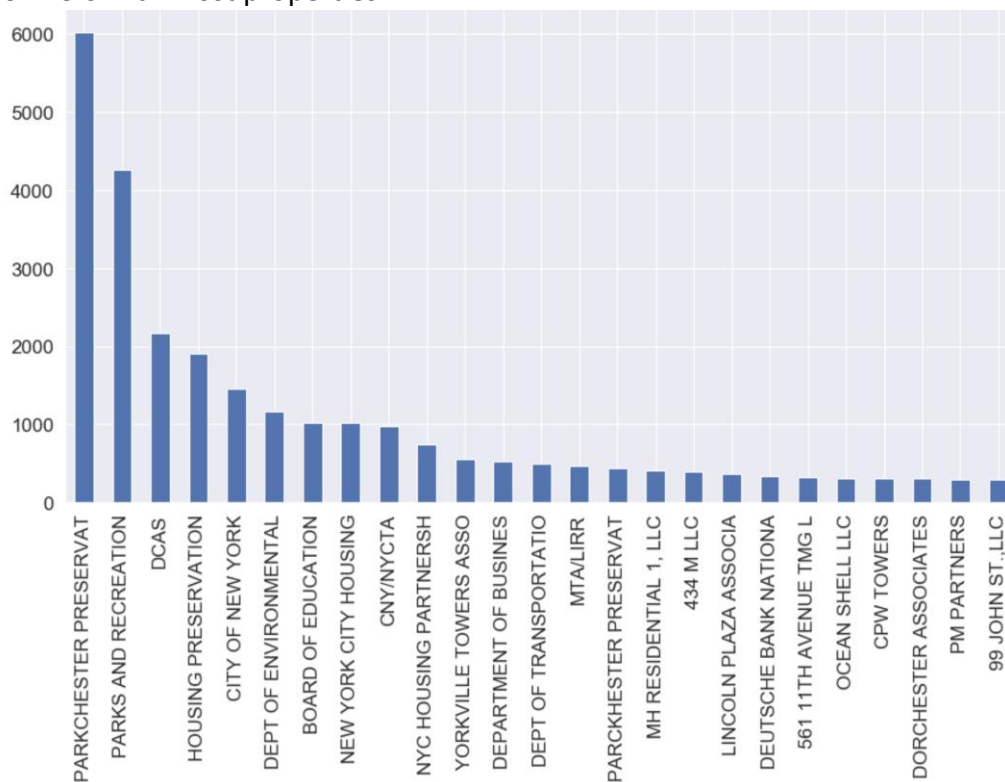
8.

Field name: BLDGCL

Field type: object

Description: building class. There is a direct correlation between the Building Class and the Tax Class.

Top 10 frequent building classes:
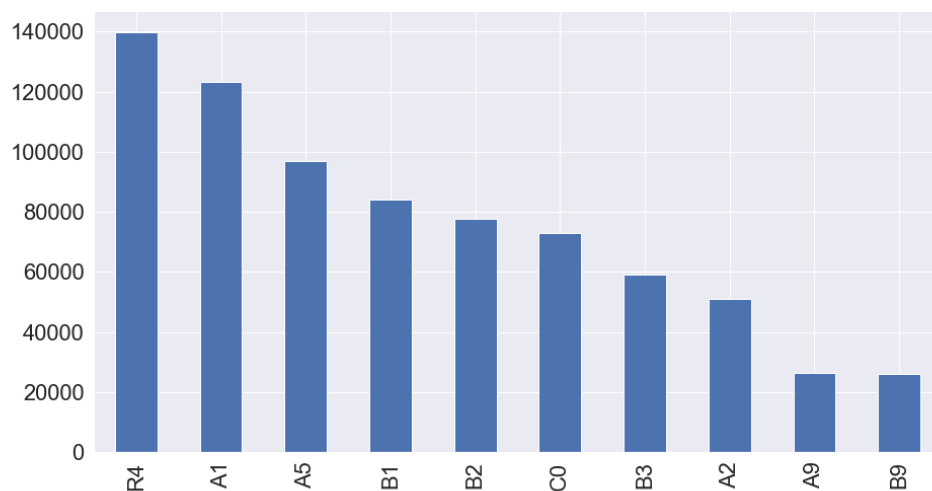
9.

Field name: TAXCLASS

Field type: object

Description: current Property Tax Class Code (NYS Classification);

tax class 1 = 1-3 unit residences

tax class 1a = 1-3 story condominiums originally a condo

tax class 1b = residential vacant land

tax class 1c = 1-3 unit condominiums originally tax class 1

tax class 1d = select bungalow colonies

tax class 2 = apartments

tax class 2a = apartments with 4-6 units

tax class 2b = apartments with 7-10 units

tax class 2c = coops/condos with 2-10 units

tax class 3 = utilities (except ceiling rr)

tax class 4a = utilities - ceiling railroads

tax class 4 = all others
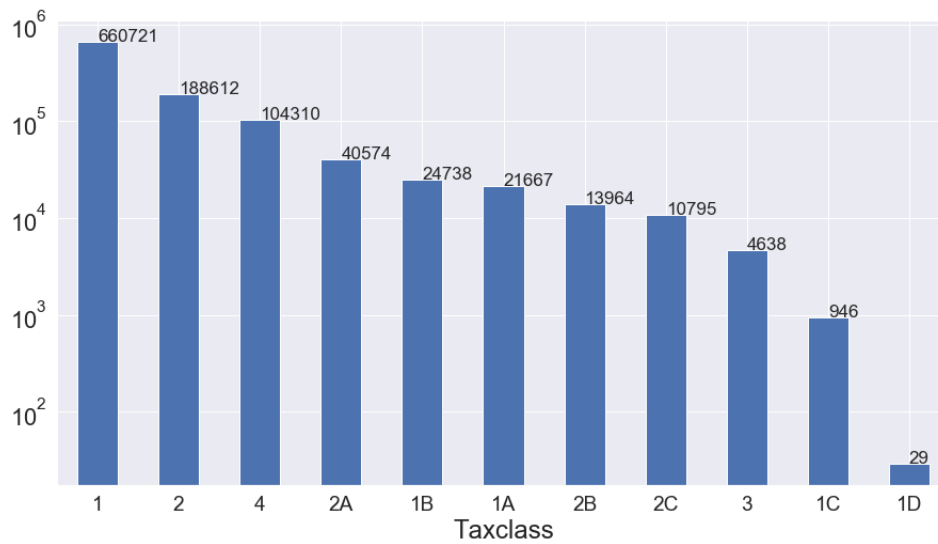
10.

Field name: LTFRONT

Field type: int64

Description: lot frontage in feet.

Properties that have lot frontage < 300:
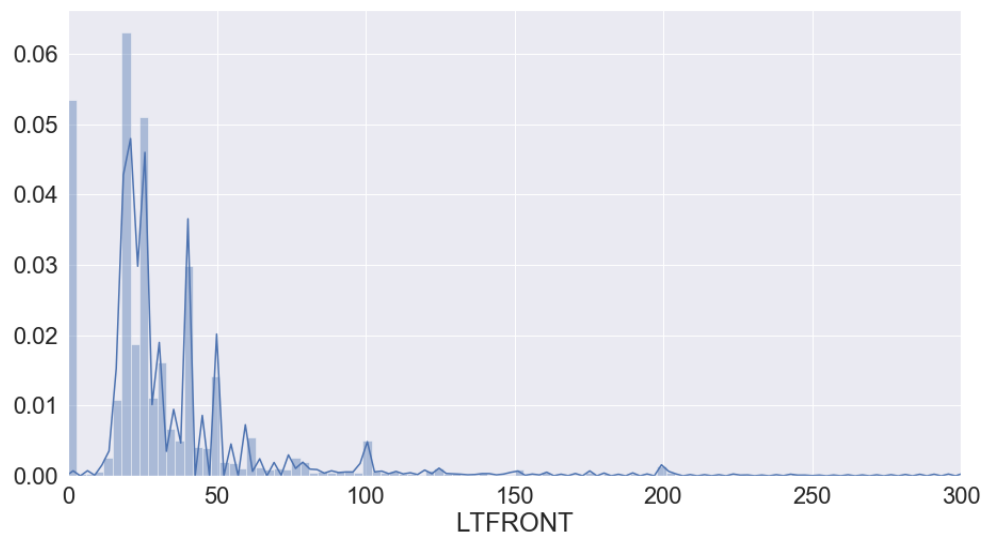
11.

Field name: LTDEPTH

Field type: int64

Description: lot depth in feet.

Properties that have lot depth < 300:
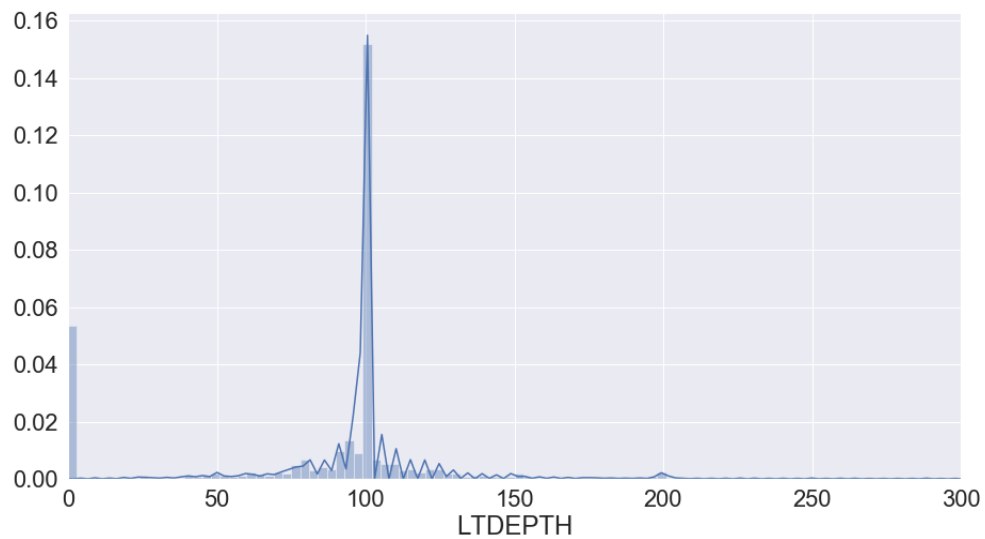
12.

Field name: EXT

Field type: object

Description: extension indicator; 'E' = extension, 'G' = garage, 'EG' = extension and garage
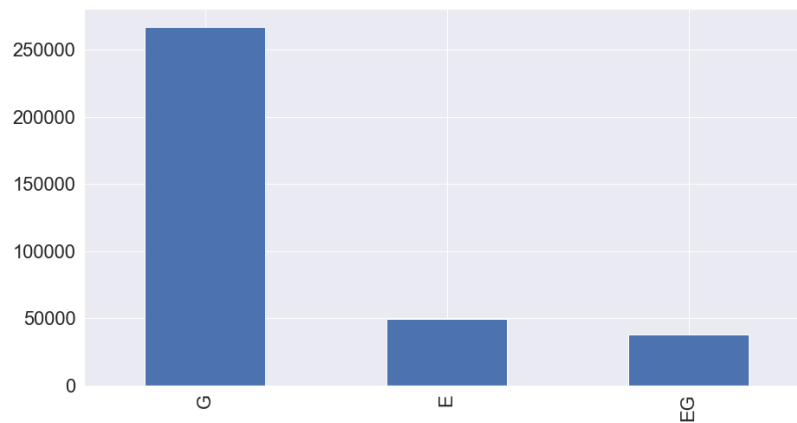
13.
Field name: STORIES
Field type: float64
Description: The number of stories for the building (# of Floors).

Properties that have stories < 50:
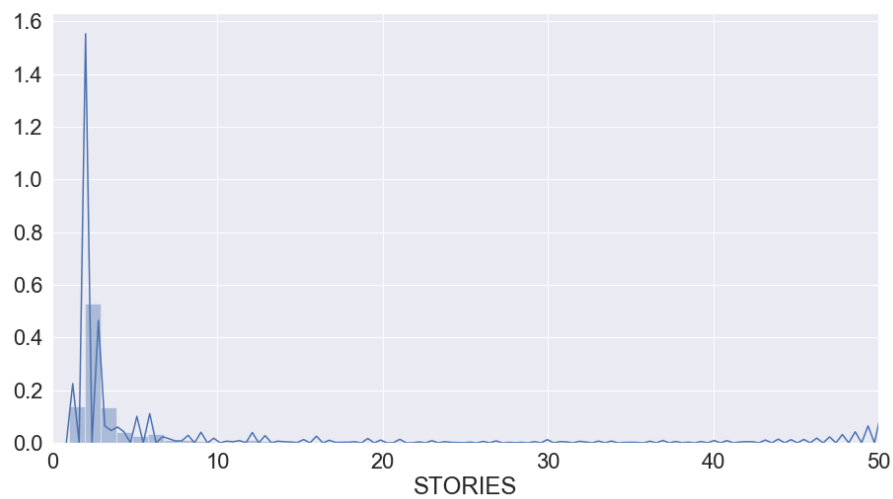
14.
Field name: FULLVAL
Field type: float64
Description: total market value.

Properties that have total market value < 2000000:
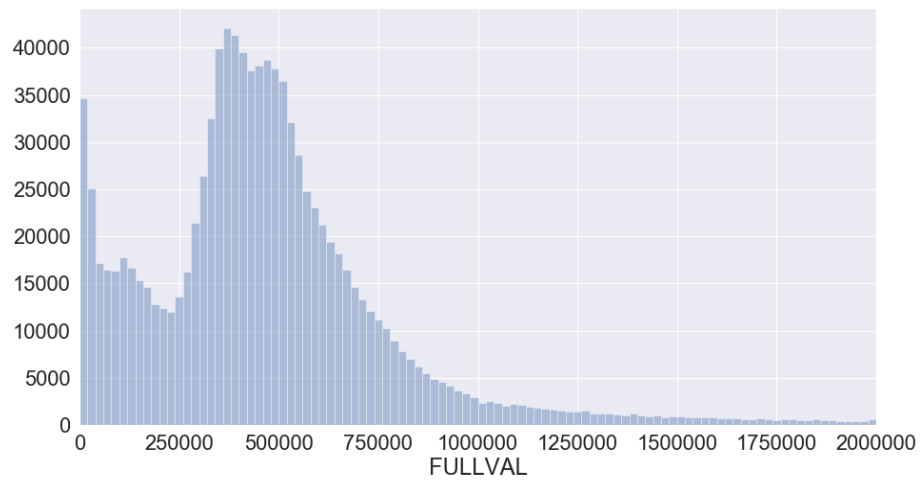
15.

Field name: AVLAND

Field type: float64

Description: actual land value.

Properties that have transitional assessed land value < 50000:
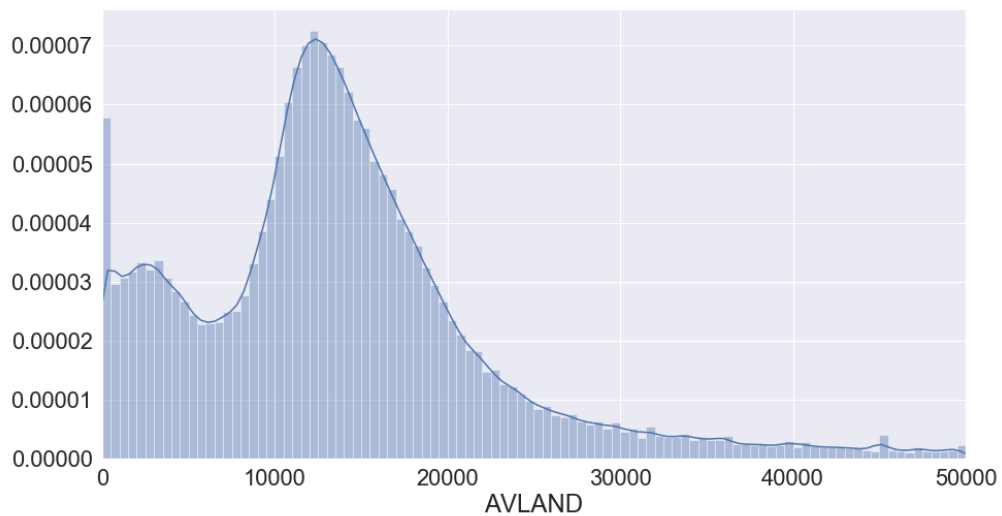
16.

Field name: AVTOT

Field type: float64

Description: actual total value.

Properties that have transitional assessed total value < 100000

AVTOT
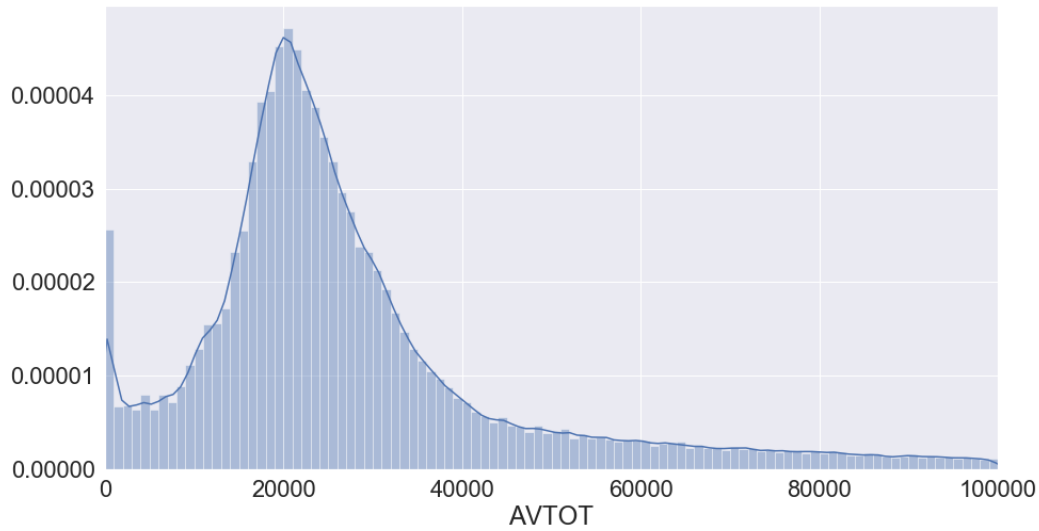
17.

Field name: EXLAND

Field type: float64

Description: actual exempt land value.

Properties that have transitional exempt land value < 20000:

EXLAND
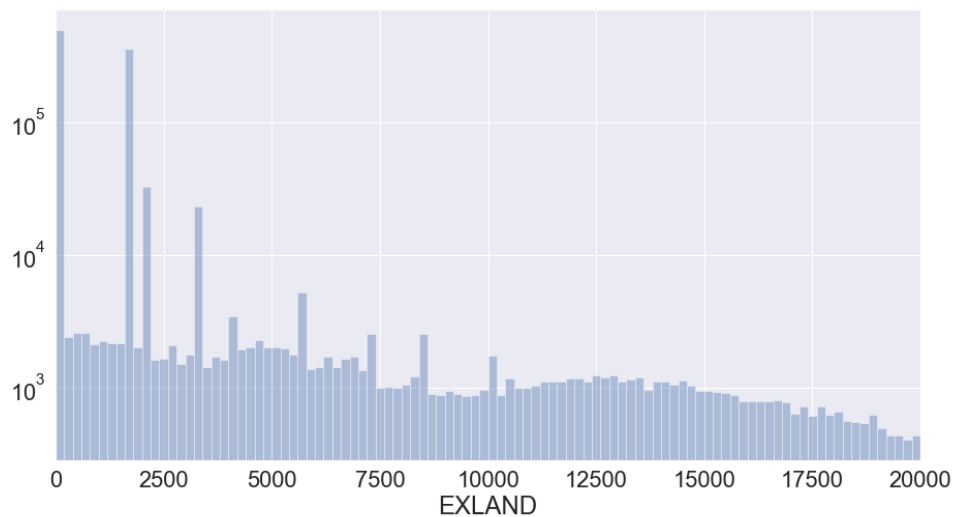
18.

Field name: EXTOT

Field type: float64

Description: actual exempt land total.

Properties that have transitional exempt total value < 20000:
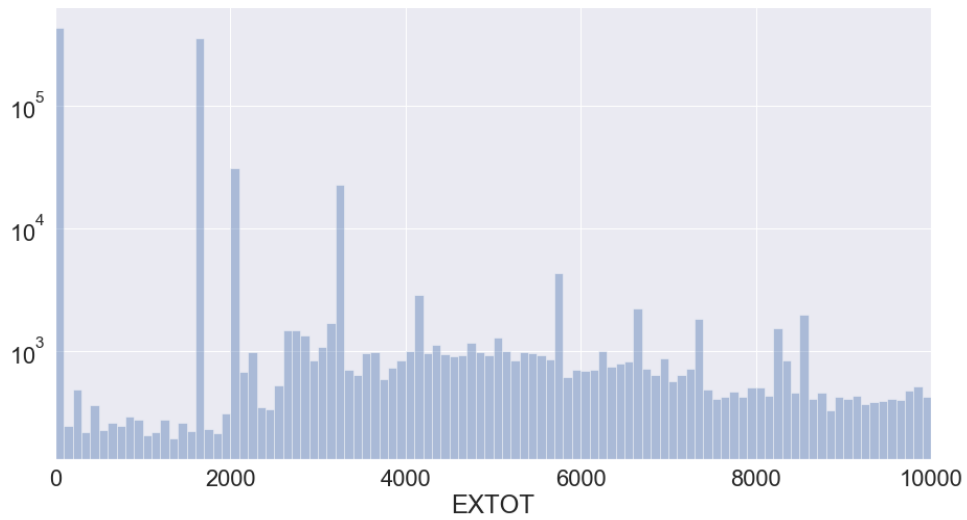
19.

Field name: EXCD1

Field type: float64

Description: exemption code 1.
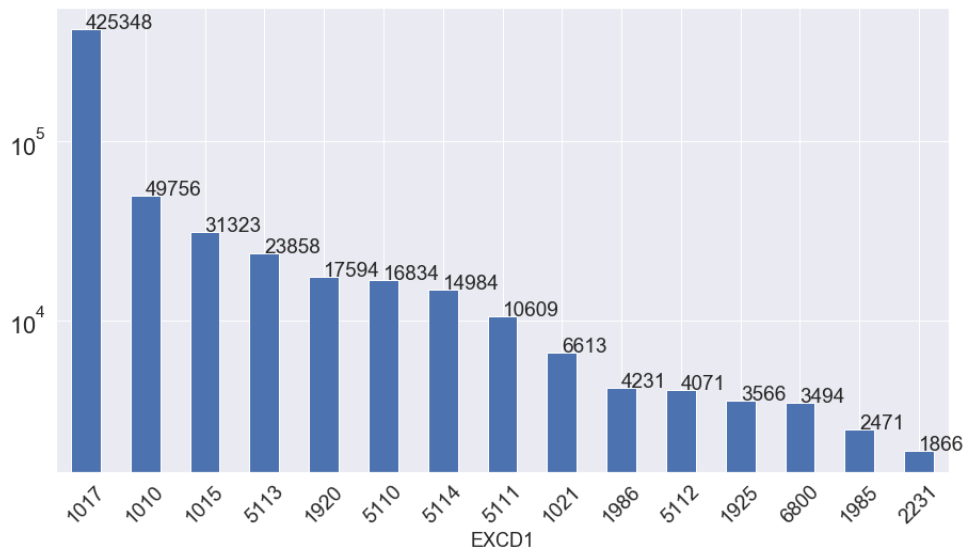
20.

Field name: STADDR

Field type: object

Description: street address. Top 15 addresses:
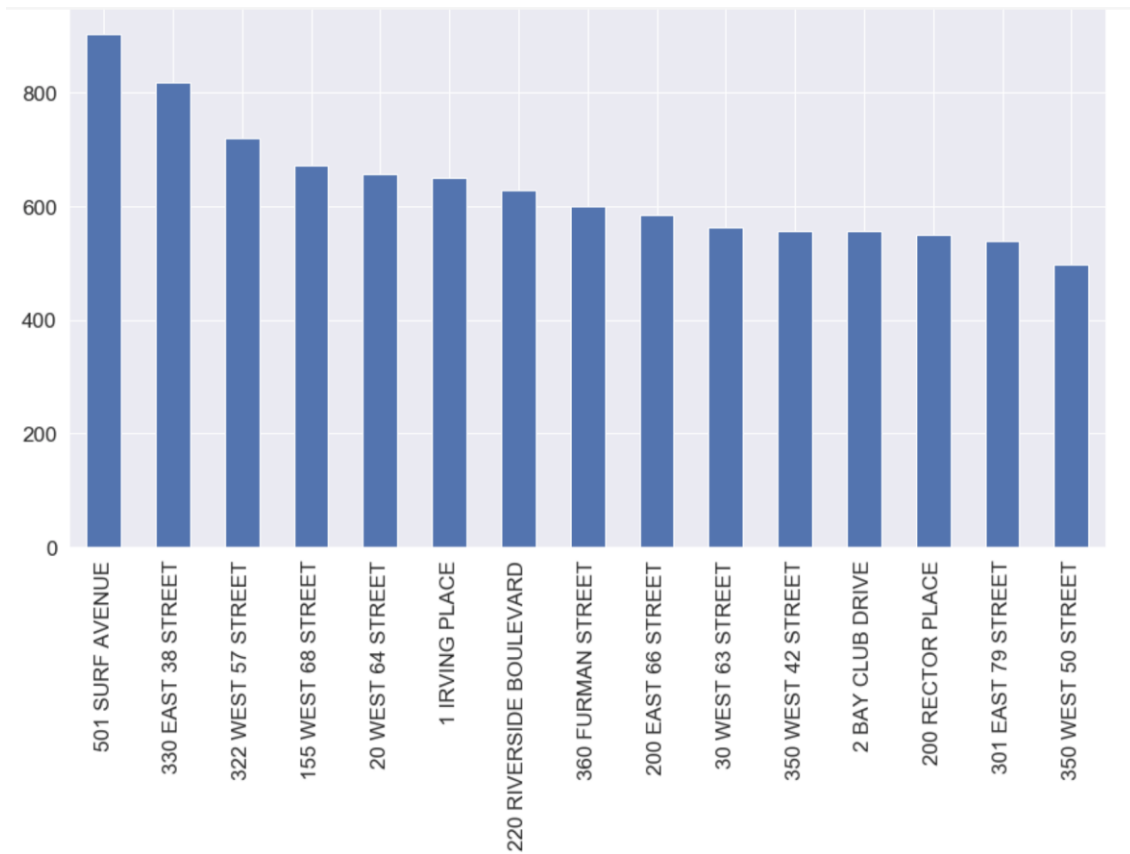
21.
Field name: ZIP
Field type: float64
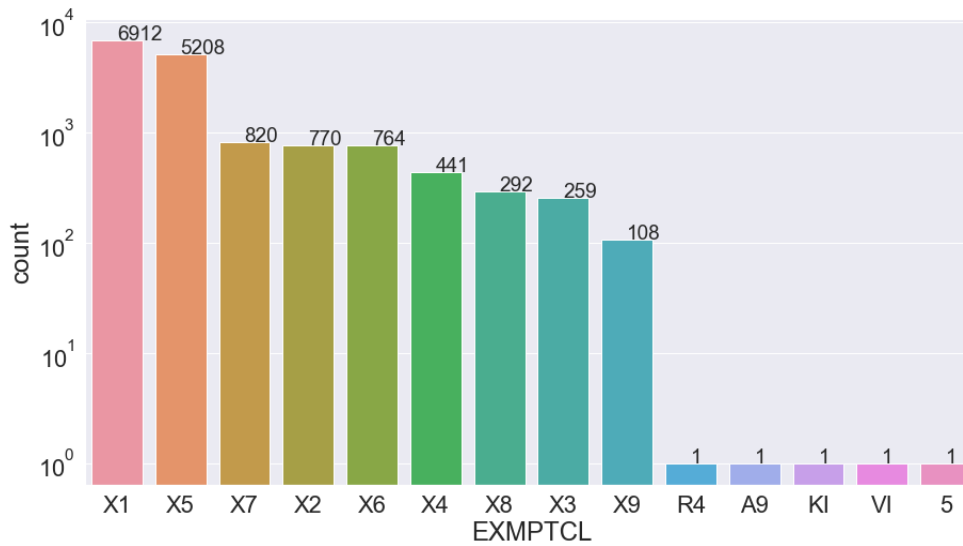**Description:**   zip code.
Top 15 zip codes:
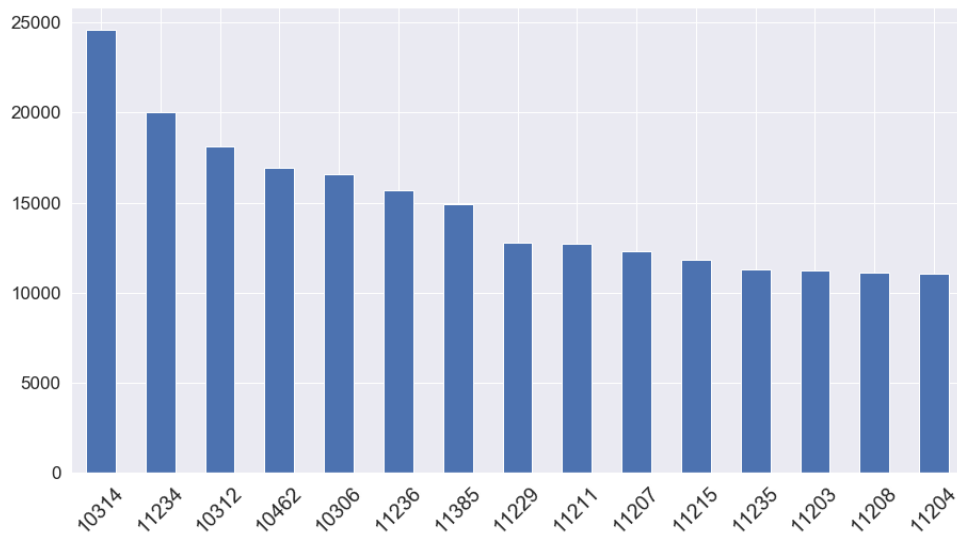
22.

Field name: EXMPTCL

Field type: object

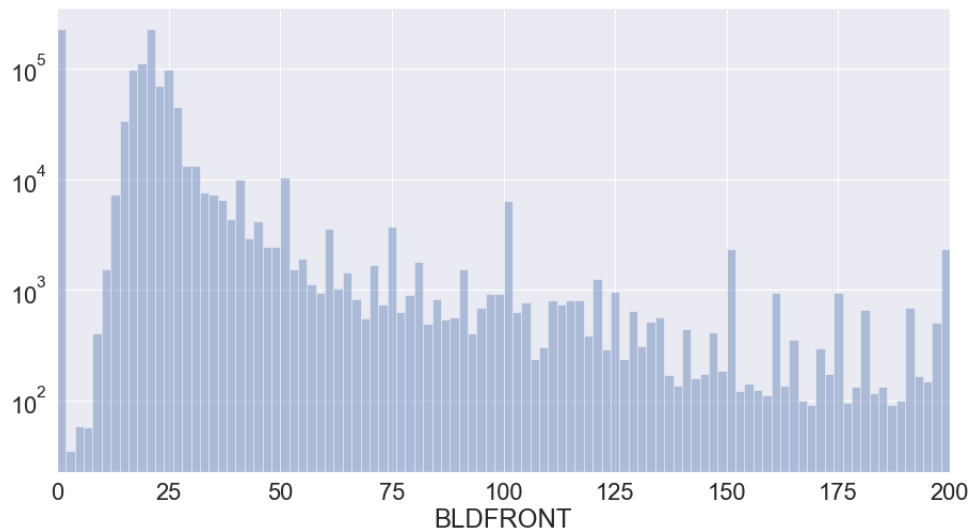**Description:** exempt class used for fully exempt properties only.

23.

Field name: BLDFRONT

Field type: int64
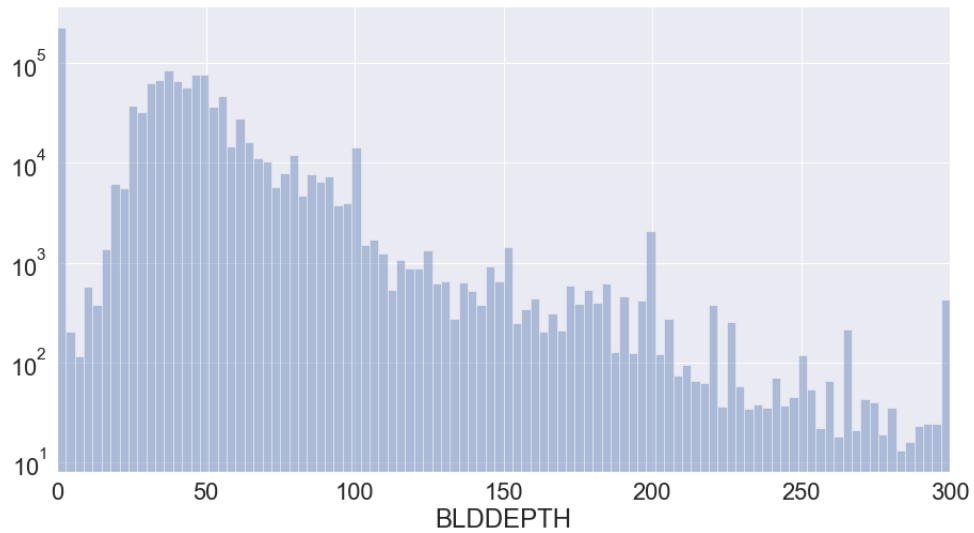
Description: building frontage in feet.

24.

Field name: BLDDEPTH

Field type: int64
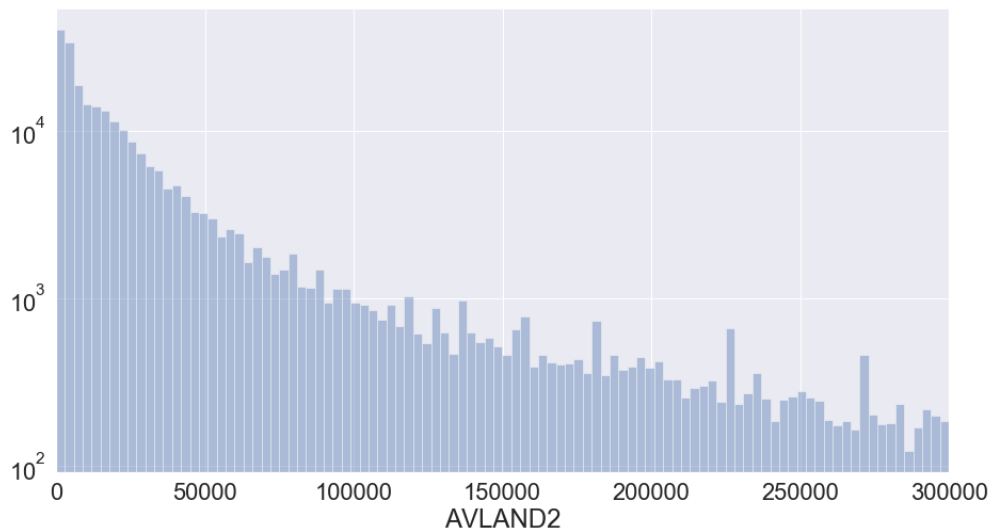
Description: building depth in feet.

25.

Field name: AVLAND2

Field type: float64

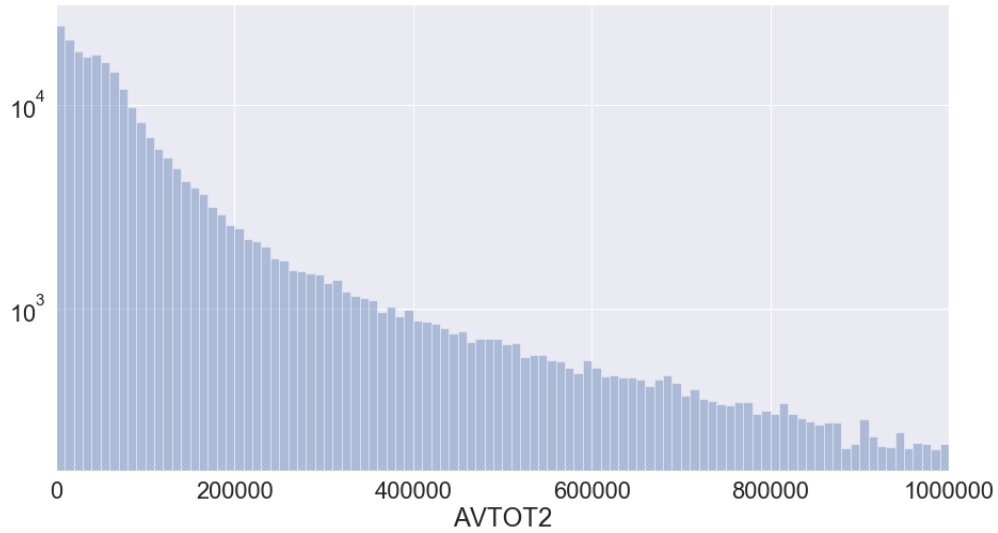**Description:**   transitional land value.

26.

Field name: AVTOT2

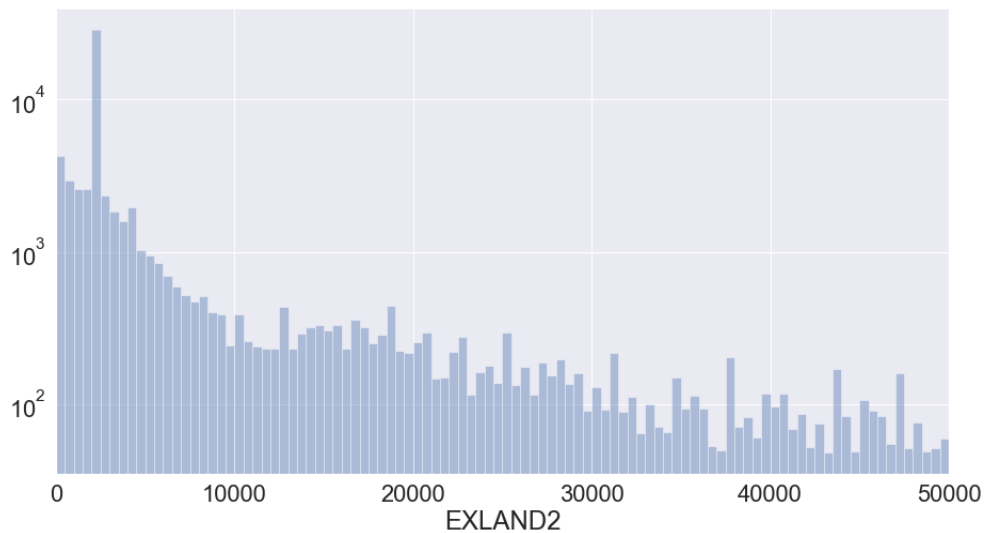**Field type:**   float64

Description: transitional total value.

27.
Field name: EXLAND2
Field type: float64
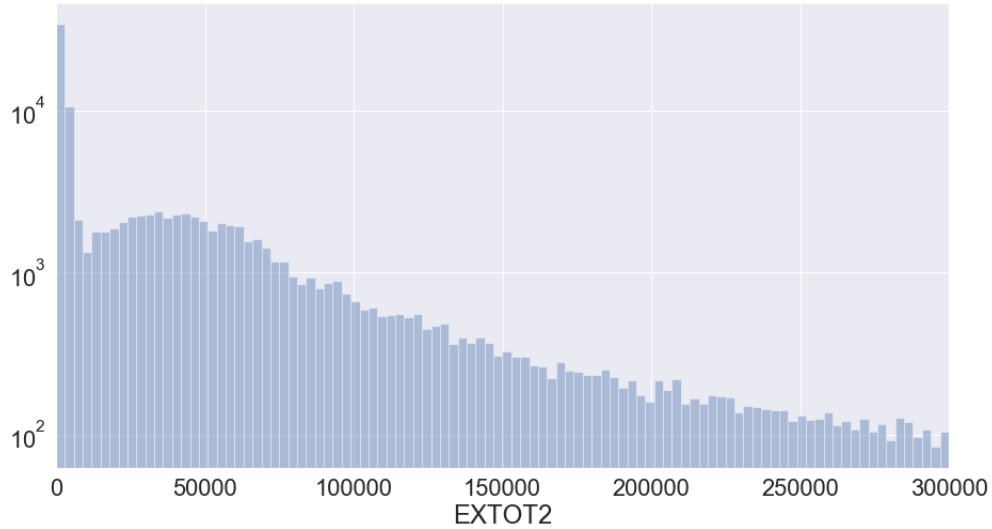**Description:** transitional exempt land value.

28.
Field name: EXTOT2
Field type: float64
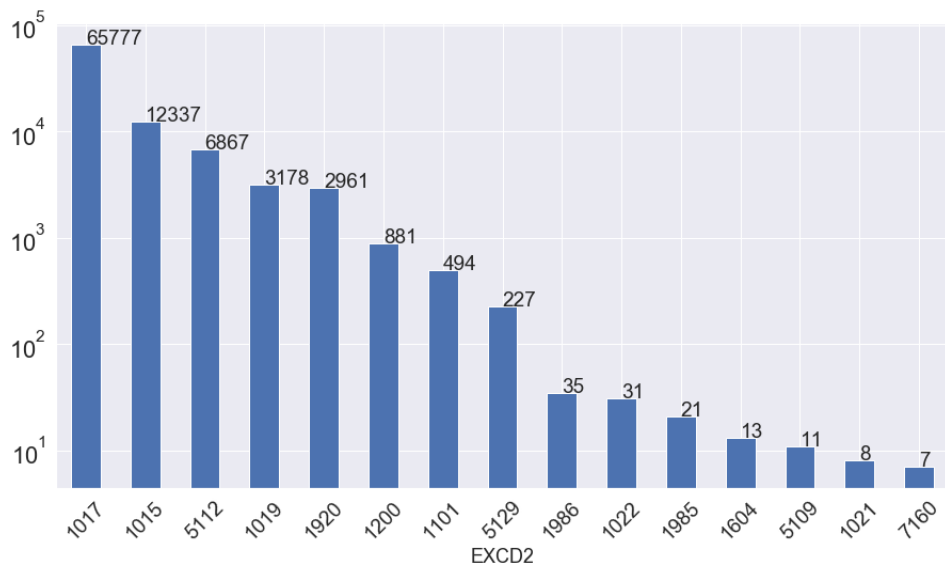**Description:** transitional exempt land total.

EXTOT2

29.

Field name: EXCD2

Field type: float64

Description: exemption code 2.

Top 15 Exemption codes:

EXCD2

30.

Field name: PERIOD

Field type: object

Description: assessment period when file was created; FINAL.

31.
Field name: YEAR
Field type: object
Description: assessment year; 2010.

32.
Field name: VALTYPE
Field type: object
Description: AC-TR.