USC Marshall
School of Business

New York Property Data:
Fraud Analytics Project Report

February 2019

6:30-9:30pm Session Team 1:
Justice League Consulting Group

Team Members:
Zongyang Jiao, Chengyin Liu, Jiayi Ma,
Xinyue Niu, Xueyan Gu, Jie Zhao

Table of Contents

Executive Summary

As a team of six, the Justice League Consulting Group has built unsupervised fraud models on the NY Property Data in order to identify fraudulent events. In this report, we began by describing the data and then explained the following processes: data cleaning, variable creation, and dimensionality reduction. We also described the two algorithms used: heuristic function of z-score and autoencoder. In the end, we presented the results and gave our conclusions, where we demonstrated our overall findings and insights.

We first took the time to understand the data and the business problem, which was determining which data records are fraudulent in the NY property dataset. The approach was to find anomalies within the dataset by building unsupervised fraud models. After cleaning the data and filling in missing fields using values that would not cause unwanted dramatic changes in the records, we created 45 expert variables. We then z-scaled the variables to give them equal importance and conducted the Principal Component Analysis to reduce the number of dimensionalities to eight, as well as reducing the correlation between variables. Next, we built a heuristic function of z-score model and an autoencoder model that generated two fraud scores, which we combined into a final fraud score. A higher score indicates a higher probability of fraud. With this score, we then rank-ordered all the entries and found the fraud records.

In the end, the results indicated that our models were effective in identifying fraudulent records for a few reasons. First, the fraud score distributions shared similar trends and displayed reasonable shapes. Second, the top 20 records identified as potential fraud did indeed demonstrate anomalies in some area, upon manual investigation. Interestingly, several of the records appeared to be government properties, and our team believes that some industry expertise can help identify the validity of these records.

Description of Data

The name of the Dataset is Property Valuation and Assessment Data, and it is updated annually. This dataset contains New York City Property valuation and assessment data provided by the Department of Finance (DOF). The data is primarily used to calculate property tax, grant eligible properties exemptions and/or abatements. It covers the time between year 2010 and 2011. In total, the data has 32 fields and 1070994 records.

Field Statistics Summary (Numeric):

Field Name	*Field Type*	*# Records w/ Value*	*% Populated*	*# Unique Values*	*# Records w/ value 0*	*Mean*	*STDEV*	*Min*	*Max*
LTFRONT	Numeric	1.07E+06	100.00%	1297	0	36.64	74.03	0	9999
LTDEPTH	Numeric	1.07E+06	100.00%	1370	0	88.86	76.4	0	9999
STORIES	Numeric	1.01E+06	94.75%	112	5.63E+04	5.01	8.37	1	119
FULLVAL	Numeric	1.07E+06	100.00%	1.09E+05	0	8.74E+05	1.16E+07	0	6.15E+09
AVLAND	Numeric	1.07E+06	100.00%	7.09E+04	0	8.51E+04	4.06E+06	0	2.67E+09
AVTOT	Numeric	1.07E+06	100.00%	1.13E+05	0	2.27E+05	6.88E+06	0	4.67E+09
EXLAND	Numeric	1.07E+06	100.00%	3.34E+04	0	3.64E+04	3.98E+06	0	2.67E+09
EXTOT	Numeric	1.07E+06	100.00%	6.43E+04	0	9.12E+04	6.51E+06	0	4.67E+09
BLDFRONT	Numeric	1.07E+06	100.00%	612	0	23.04	35.58	0	7575
BLDDEPTH	Numeric	1.07E+06	100.00%	621	0	39.92	42.71	0	9393
AVLAND2	Numeric	2.83E+05	26.40%	5.86E+04	7.88E+05	2.46E+05	6.18E+06	3	2.37E+09
AVTOT2	Numeric	2.83E+05	26.40%	1.11E+05	7.88E+05	7.14E+05	1.17E+07	3	4.50E+09
EXLAND2	Numeric	8.74E+04	8.17%	2.22E+04	9.84E+05	3.51E+05	1.08E+07	1	2.37E+09
EXTOT2	Numeric	1.31E+05	12.22%	4.83E+04	9.40E+05	6.57E+05	1.61E+07	7	4.50E+09

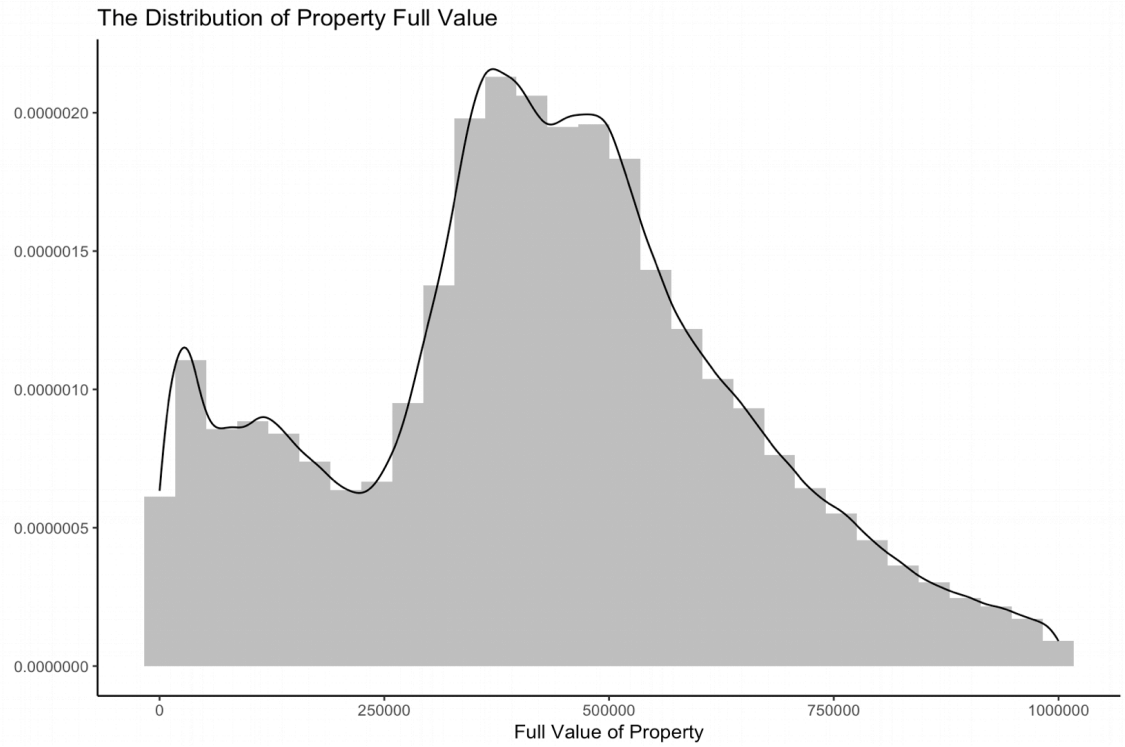
Field Statistics Summary (Categorical):

Field Name	*Field Type*	*# Records w/ Value*	*% Populated*	*# Unique Values*	*# Records w/ value 0*	*Most Common Value*
Record	Categorical	1.07E+06	100.00%	1.07E+06	0	NA
BBLE	Categorical	1.07E+06	100.00%	1.07E+06	0	NA
B	Categorical	1.07E+06	100.00%	5	0	4
Block	Categorical	1.07E+06	100.00%	1.40E+04	0	3944
LOT	Categorical	1.07E+06	100.00%	6366	0	1
EASEMENT	Categorical	4636	0.43%	13	1.07E+06	E

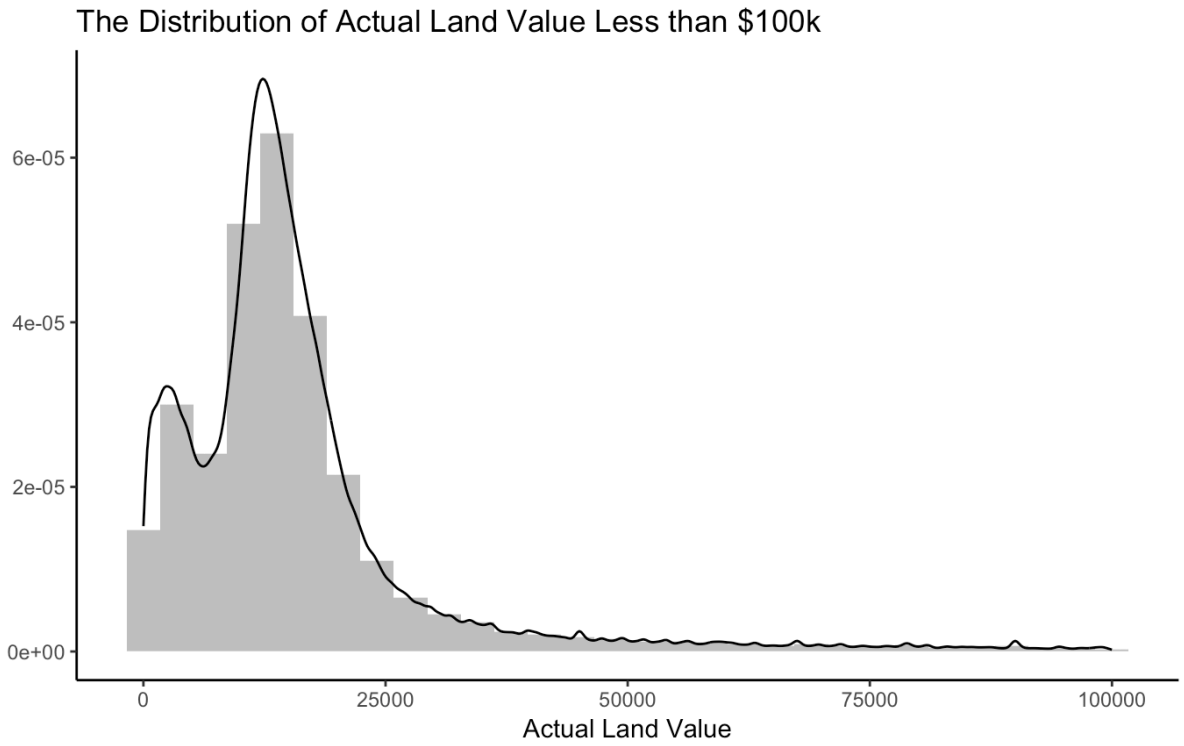
OWNER	Categorical	1.04E+06	97.04%	8.63E+05	3.17E+04	PARKCHESTER PRESERVAT
BLDGCL	Categorical	1.07E+06	100.00%	200	0	R4
TAXCLASS	Categorical	1.07E+06	100.00%	11	0	1
EXT	Categorical	1.07E+06	100.00%	4	0	G
EXCD1	Categorical	6.38E+05	59.62%	130	4.33E+05	1017
STADDR	Categorical	1.07E+06	99.94%	8.39E+05	676	501
ZIP	Categorical	1.04E+06	97.21%	197	2.99E+04	1.03E+04
EXMPTCL	Categorical	1.56E+04	1.45%	15	1.06E+06	X1
EXCD2	Categorical	9.29E+04	8.68%	61	9.78E+05	1017
PERIOD	Categorical	1.07E+06	100.00%	1	0	FINAL
YEAR	Categorical	1.07E+06	100.00%	1	0	2010/11
VALTYPE	Categorical	1.07E+06	100.00%	1	0	AC-TR

Important Distributions (FULLVAL, AVALAND, AVTOT):

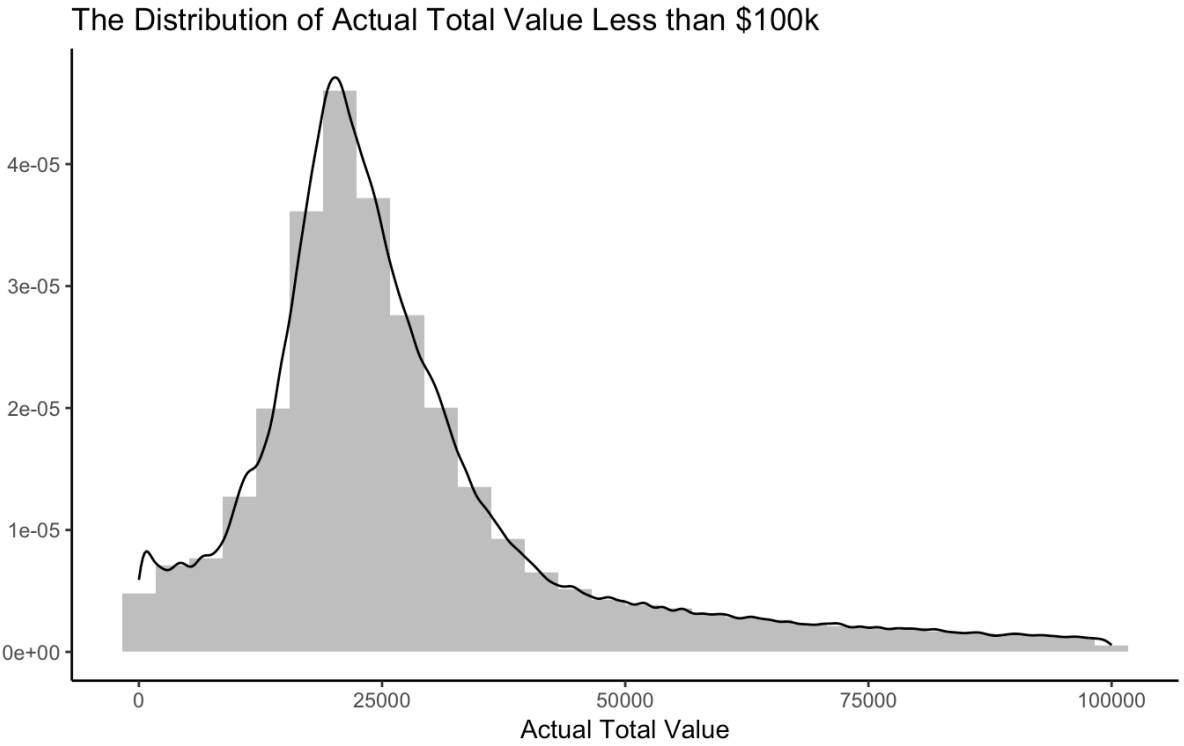
FULLVAL



AVALAND



AVTOT



Data Cleaning

To achieve better results, we used R to clean and manipulate the data before proceeding to next steps in Python. First of all, we extracted 13 variables from the dataset to clean and fill any missing values in applicable fields: [FULLVAL], [AVLAND], [AVTOT], [B], [BLOCK], [BLDGCL], [TAXCLASS], [ZIP], [LTFRONT], [LTDEPTH], [BLDFRONT], [BLDDEPTH], [STORIES].

The general idea of filling missing fields is to use summary statistics of similar groups that the records identify with, while minimizing the likelihood of causing any unwanted abnormal results. The following table summarizes the process of filling missing value:

	1st group	***# NA***	***2nd group***	***# NA***	***3rd group***	***# NA***
ZIP	B, BLOCK	2448	B	0		
FULLVAL	BLDGCL, ZIP	10201	ZIP	0		
AVLAND	BLDGCL, ZIP	10201	ZIP	0		
AVTOT	BLDGCL, ZIP	10201	ZIP	0		
LTFRONT	BLDGCL, ZIP	15868	BLDGCL	0		
LTDEPTH	BLDGCL, ZIP	17605	BLDGCL	0		
BLDFRONT	BLDGCL, ZIP	74663	BLDGCL	4280	ZIP	0
BLDDEPTH	BLDGCL, ZIP	74660	BLDGCL	3281	ZIP	0
STORIES	BLDGCL, ZIP	41742	BLDGCL	4280	ZIP	0

ZIP

Description: there are 29,890 NA values in ZIP field.

Method:

- We filled the null ZIP values based on other properties with the same [B] and [BLOCK]. For example, a property without a ZIP value is in [B] 5, [BLOCK] 8021. Properties in the same B and BLOCK have a ZIP code of 10307. So, we filled this particular NA ZIP using 10307.
- For some combinations of [B] and [BLOCK], all the records have an NA value in ZIP. So, we filled these NAs based only on [B].

FULLVAL, AVLAND, AVTOT

Description: there are no NA values in the fields, but 13,007 records have a value of 0.

Method:

- We first used NA to replace all zero values.
- We then filled the NA values using the field median of properties in the same [ZIP] and [BLDGCL].
- Finally, we filled the remaining NA values using the median of properties in the same [ZIP].

LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH

Description: there are no NA values in the fields, but 169,108 of LTFRONT, 170,128 of LTDEPTH, 228,815 of BLDFRONT, and 228,853 of BLDDEPTH have a value of 0.

Method:

- We first used NA to fill zero values.
- For [LTFRONT] and [LTDEPTH], we filled NA values based on the group median of [ZIP] and [BLDGCL]. Then, we filled the remaining NA values using the group median of [BLDGCL].
- For [BLDFRONT] and [BLDDEPTH], we repeated the above steps. In addition, we filled the last few NA values using the group median of [ZIP].

STORIES

Description: there are 56,264 NA values in STORIES.

Method:

- We filled NA values using the group median of [ZIP] and [BLDGCL].
- Then, we filled the remaining NA values using the group median of [BLDGCL].
- Finally, we filled the last NA values using the group median of [ZIP].

Variable Creation

After cleaning the data, we focused on three variable fields: [FULLVAL], [AVLAND], [AVTOT]. We then built special variables to scale these three value fields. The steps are listed as the following:

(1) Creating three size variables: Lot Area, Building Area, and Building Volume, by the following formulas:

$$\text{LOTAREA} = \text{LTFRONT} \times \text{LTDEPTH}$$
$$\text{BLDAREA} = \text{BLDFRONT} \times \text{BLDDEPTH}$$
$$\text{BLDVOL} = \text{BLDFRONT} \times \text{BLDDEPTH} \times \text{STORIES}$$

(2) Assigning different symbols to value and size variables respectively.

$$\text{V}_1 = \text{FULLVAL} \qquad \text{S}_1 = \text{LOTAREA}$$
$$\text{V}_2 = \text{AVLAND} \qquad \text{S}_2 = \text{BLDAREA}$$
$$\text{V}_3 = \text{AVTOT} \qquad \text{S}_3 = \text{BLDVOL}$$

(3) Calculating nine valuables. Each of the three value variables was normalized by each of the three size variables. We got nine unit-value variables based on Lot Area, Building Area, and Building Volume.

$$r_1 = \frac{V_1}{S_1} \qquad r_4 = \frac{V_2}{S_1} \qquad r_7 = \frac{V_3}{S_1}$$
$$r_2 = \frac{V_1}{S_2} \qquad r_5 = \frac{V_2}{S_2} \qquad r_8 = \frac{V_3}{S_2}$$
$$r_3 = \frac{V_1}{S_3} \qquad r_6 = \frac{V_2}{S_3} \qquad r_9 = \frac{V_3}{S_3}$$

(4) Respectively grouping the records by five scale groups, which are ZIP5, ZIP3, TAXCLASS, BOROUGH, and ALL (all the data).

(5) For each group g, calculating the average of each r_i: we got 45 averages $< r_i >_g$ (i = 1,2, …, 9) for five scale groups (g = 1,2, …,5).

(6) For each record, calculating 45 variables by dividing each of the nine core valuables by the five scale groups. We got 45 variables as below:

$$\frac{r_1}{<r_1>_1} \quad \frac{r_2}{<r_2>_1} \quad \dots \quad \frac{r_8}{<r_8>_1} \quad \frac{r_9}{<r_9>_1}$$
$$\frac{r_1}{<r_1>_2} \quad \frac{r_2}{<r_2>_2} \quad \dots \quad \frac{r_8}{<r_8>_2} \quad \frac{r_9}{<r_9>_2}$$
$$\frac{r_1}{<r_1>_3} \quad \frac{r_2}{<r_2>_3} \quad \dots \quad \frac{r_8}{<r_8>_3} \quad \frac{r_9}{<r_9>_3}$$
$$\frac{r_1}{<r_1>_4} \quad \frac{r_2}{<r_2>_4} \quad \dots \quad \frac{r_8}{<r_8>_4} \quad \frac{r_9}{<r_9>_4}$$
$$\frac{r_1}{<r_1>_5} \quad \frac{r_2}{<r_2>_5} \quad \dots \quad \frac{r_8}{<r_8>_5} \quad \frac{r_9}{<r_9>_5}$$

(7) All of the 45 variables are listed below:

No.	Variable	Formula	No.	Variable	Formula
1	FULLVAL_LTAREA_By_ZIP5	$\frac{r_1}{<r_1>_1}$	24	AVLAND_BLDVOL_By_TAXCLASS	$\frac{r_6}{<r_6>_3}$
2	FULLVAL_BLDAREA_By_ZIP5	$\frac{r_2}{<r_2>_1}$	25	AVTOT_LTAREA_By_TAXCLASS	$\frac{r_7}{<r_7>_3}$
3	FULLVAL_BLDVOL_By_ZIP5	$\frac{r_3}{<r_3>_1}$	26	AVTOT_BLDAREA_By_TAXCLASS	$\frac{r_8}{<r_8>_3}$
4	AVLAND_LTAREA_By_ZIP5	$\frac{r_4}{<r_4>_1}$	27	AVTOT_BLDVOL_By_TAXCLASS	$\frac{r_9}{<r_9>_3}$
5	AVLAND_BLDAREA_By_ZIP5	$\frac{r_5}{<r_5>_1}$	28	FULLVAL_LTAREA_By_B	$\frac{r_1}{<r_1>_4}$
6	AVLAND_BLDVOL_By_ZIP5	$\frac{r_6}{<r_6>_1}$	29	FULLVAL_BLDAREA_By_B	$\frac{r_2}{<r_2>_4}$
7	AVTOT_LTAREA_By_ZIP5	$\frac{r_7}{<r_7>_1}$	30	FULLVAL_BLDVOL_By_B	$\frac{r_3}{<r_3>_4}$
8	AVTOT_BLDAREA_By_ZIP5	$\frac{r_8}{<r_8>_1}$	31	AVLAND_LTAREA_By_B	$\frac{r_4}{<r_4>_4}$
9	AVTOT_BLDVOL_By_ZIP5	$\frac{r_9}{<r_9>_1}$	32	AVLAND_BLDAREA_By_B	$\frac{r_5}{<r_5>_4}$
10	FULLVAL_LTAREA_By_ZIP3	$\frac{r_1}{<r_1>_2}$	33	AVLAND_BLDVOL_By_B	$\frac{r_6}{<r_6>_4}$

11	FULLVAL_BLDAREA_By_ZIP3	$\frac{r_2}{< r_2 >_2}$	34	AVTOT_LTAREA_By_B	$\frac{r_7}{< r_7 >_4}$
12	FULLVAL_BLDVOL_By_ZIP3	$\frac{r_3}{< r_3 >_2}$	35	AVTOT_BLDAREA_By_B	$\frac{r_8}{< r_8 >_4}$
13	AVLAND_LTAREA_By_ZIP3	$\frac{r_4}{< r_4 >_2}$	36	AVTOT_BLDVOL_By_B	$\frac{r_9}{< r_9 >_4}$
14	AVLAND_BLDAREA_By_ZIP3	$\frac{r_5}{< r_5 >_2}$	37	FULLVAL_LTAREA_By_ALL	$\frac{r_1}{< r_1 >_5}$
15	AVLAND_BLDVOL_By_ZIP3	$\frac{r_6}{< r_6 >_2}$	38	FULLVAL_BLDAREA_By_ALL	$\frac{r_2}{< r_2 >_5}$
16	AVTOT_LTAREA_By_ZIP3	$\frac{r_7}{< r_7 >_2}$	39	FULLVAL_BLDVOL_By_ALL	$\frac{r_3}{< r_3 >_5}$
17	AVTOT_BLDAREA_By_ZIP3	$\frac{r_8}{< r_8 >_2}$	40	AVLAND_LTAREA_By_ALL	$\frac{r_4}{< r_4 >_5}$
18	AVTOT_BLDVOL_By_ZIP3	$\frac{r_9}{< r_9 >_2}$	41	AVLAND_BLDAREA_By_ALL	$\frac{r_5}{< r_5 >_5}$
19	FULLVAL_LTAREA_By_TAXCLASS	$\frac{r_1}{< r_1 >_3}$	42	AVLAND_BLDVOL_By_ALL	$\frac{r_6}{< r_6 >_5}$
20	FULLVAL_BLDAREA_By_TAXCLASS	$\frac{r_2}{< r_2 >_3}$	43	AVTOT_LTAREA_By_ALL	$\frac{r_7}{< r_7 >_5}$
21	FULLVAL_BLDVOL_By_TAXCLASS	$\frac{r_3}{< r_3 >_3}$	44	AVTOT_BLDAREA_By_ALL	$\frac{r_8}{< r_8 >_5}$
22	AVLAND_LTAREA_By_TAXCLASS	$\frac{r_4}{< r_4 >_3}$	45	AVTOT_BLDVOL_By_ALL	$\frac{r_9}{< r_9 >_5}$
23	AVLAND_BLDAREA_By_TAXCLASS	$\frac{r_5}{< r_5 >_3}$			

Dimensionality Reduction

After variable creation, we had 45 fields, which were too many for further processing. Therefore, we took the necessary steps in Python to reduce dimensionality. There were three essential steps in the dimensionality reduction process:

(1) Using the z-scale method to normalize 45 variables.
(2) Performing the Principal Component Analysis (PCA) to reduce variables to eight eigenvectors.
(3) Z-scaling again to assign equal weight to all eight eigenvectors.

1. First Z-Scaling

Z-scaling is a common method used for normalization. The standard score of sample x in z scale is calculated as:

$$z = (x - u) / s$$

After the first z-scaling, the dataset looked like the following:

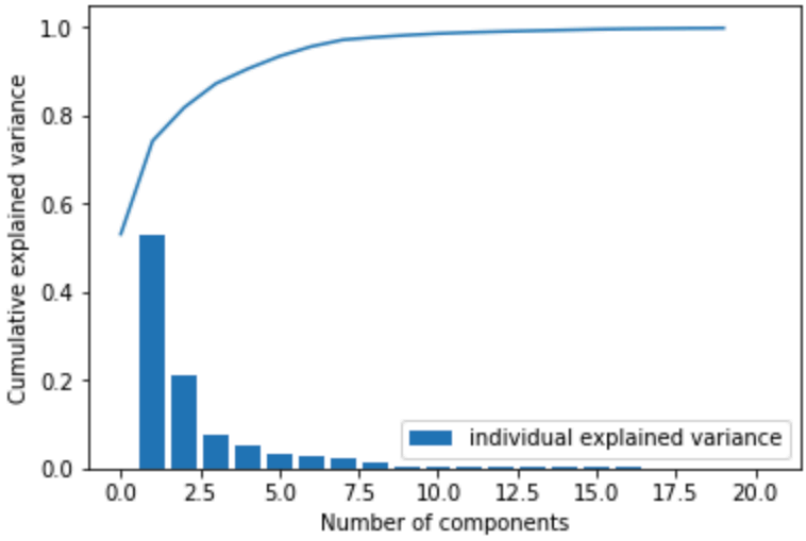
	BBLE	FULLVAL_LTAREA_By_B	FULLVAL_BLDAREA_By_B	...	AVTOT_BLDVOL_By_ALL
0	1000010101	-0.060695	0.900583	...	0.510232
1	1000010201	7.567991	2.510693		1.623317
2	1000020001	-0.019102	-0.035133		0.002156
3	1000020023	-0.051507	0.029057		0.072771
...					

2. Principal Component Analysis (PCA)

The main idea of principal component analysis (PCA) is to reduce the dimensionality of a dataset consisting of many variables correlated with each other, while retaining the variation present in the dataset. PCA transforms the variables to a new set of variables, known as the principal components (or simply, the PCs). The 1st PC retains maximum variation that was present in the original components.

a. Selecting Principal Components

In order to determine exactly how many eigenvectors we should keep, we first selected 20 PCs and graphed the relationship between number of PCs and the cumulative explained variance, which is shown as below:



Cumulative Explained Variance			
PC1	0.5308	PC11	0.9850
PC2	0.7416	PC12	0.9876
PC3	0.8177	PC13	0.9897
PC4	0.8719	PC14	0.9914
PC5	0.9051	PC15	0.9930
PC6	0.9336	PC16	0.9946
PC7	0.9557	PC17	0.9957
PC8	0.9709	PC18	0.9964
PC9	0.9767	PC19	0.9971
PC10	0.9812	PC20	0.9975

Table above demonstrates the cumulative percentage information for PCs from 1 to 20. In figure above, we can see that there is an obvious reduction between the 8th PC and the 9th PC. Therefore, we decided to select the first eight PCs.

b. Eight Principal Components

From the above step, we selected the top eight PCs, and the dataset is demonstrated below:

	Principal component 1	Principal component 2	Principal component 3	Principal component 4	Principal component 5	Principal component 6	Principal component 7	Principal component 8	BBLE
0	2.451661	-0.989880	-0.703086	-0.616673	0.595488	0.792087	-0.878392	0.125770	1000010101
1	21.523377	48.778412	13.456927	-6.158127	1.922825	18.103603	4.171438	0.996754	1000010201
2	-0.014298	0.216899	-0.017475	0.126546	-0.014873	-0.102865	-0.046818	0.052827	1000020001
…									

3. Second Z-Scaling

After PCA, we used the z-scale method again for the selected 8 PCs to make sure that all of them have same weights going into next steps, as demonstrated:

	Principal component 1	Principal component 2	Principal component 3	Principal component 4	Principal component 5	Principal component 6	Principal component 7	Principal component 8	BBLE
0	0.501644	-0.321356	-0.380120	-0.394689	0.487550	0.699028	-0.881291	0.152125	1000010101
1	4.403986	15.835488	7.275422	-3.941386	1.574293	15.976702	4.185205	1.205625	1000010201
2	-0.002926	0.062505	-0.009448	0.080993	-0.012177	-0.090780	-0.046973	0.063896	1000020001
…									

Algorithms

After we reduced variables to the final eight z-scaled fields, we built fraud score1 and score2 using two methods: Heuristic Function and Autoencoder.

Score1. Heuristic Function

With the selected eight principal components, the first fraud score is calculated using this formula:

$$Score_1 = \left(\sum_{k=1}^{8} |PC_k|^2 \right)^{\frac{1}{2}}$$

We can see that this function finds the linear relationship between score1 and all eight z-scaled scores. This way, if one of the principal components is an outlier, its score1 will be very high. This is due to the fact that taking the square of all eight PCs maintains the effect of outliers, which is demonstrated through score1. If score1 is very high, then this record is very likely to be an outlier, or a fraudulent record.

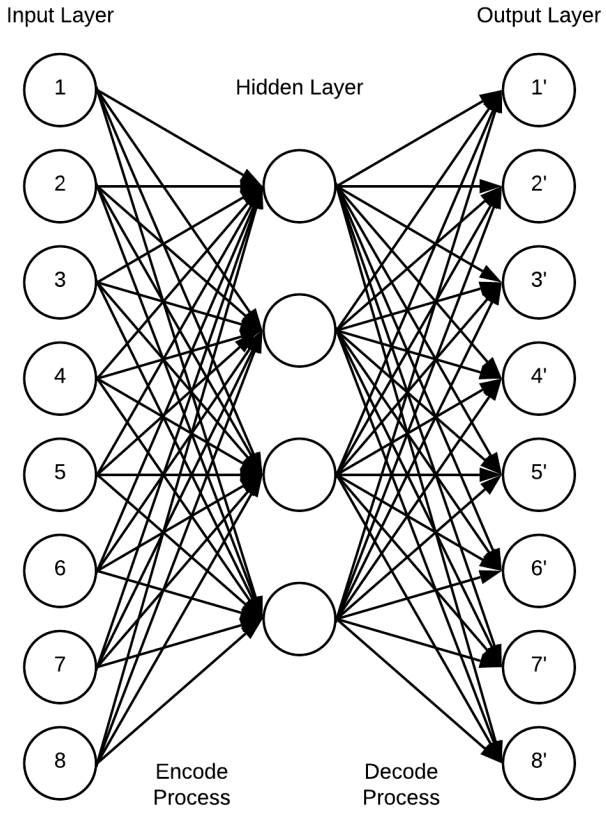
	Principal component 1	Principal component 2	Principal component 3	Principal component 4	Principal component 5	Principal component 6	Principal component 7	Principal component 8	Score1
0	0.501644	-0.321356	-0.380120	-0.394689	0.487550	0.699028	-0.881291	0.152124	1.476939
1	4.403986	15.835488	7.275422	-3.941386	1.574293	15.976702	4.185206	1.205646	24.805785
2	-0.002926	0.070414	-0.009448	0.080993	-0.012177	-0.090780	-0.046973	0.063897	0.162155
3	0.035272	-0.062505	-0.065339	0.002021	-0.037041	-0.070131	-0.107690	-0.016916	0.166126
4	8.014345	-3.546332	-2.735559	-4.058080	-10.132589	-10.132589	-10.622918	-13.945803	24.376021

Score2. Autoencoder

When building the autoencoder model, we aimed to reproduce all data points. Some data points, however, could not be reproduced accurately (similar to the original) based on the given information. These data points are therefore abnormal values, or potential fraudulent records.

To build the autoencoder, we employed a neural network model. Here, we used the preselected eight principal components as the input layer and also as the output layer. During the encoding process, this algorithm reduced the need of variable amount to represent the original data while maintaining key features of the data. We reduced the dimensionalities from eight to four in the hidden layer. In the decoding process, the autoencoder decodes the data in the hidden layer by

using the same encoding function in the opposite way, reproducing each data point to make it as similar to the original data as possible. This process is demonstrated below:



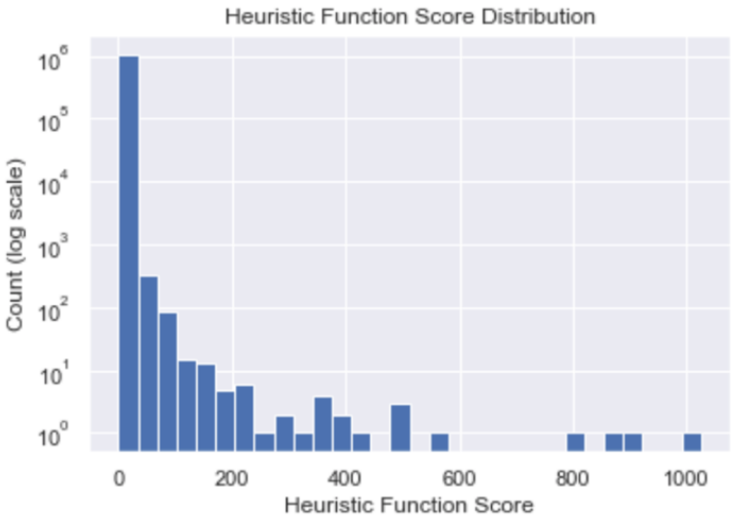
After these two steps, we calculated the differences between the original data and the reproduced data in order to examine and identify which data points failed to be accurately reproduced. The formula used to calculate the differences—or ultimately—the score, is as follows:

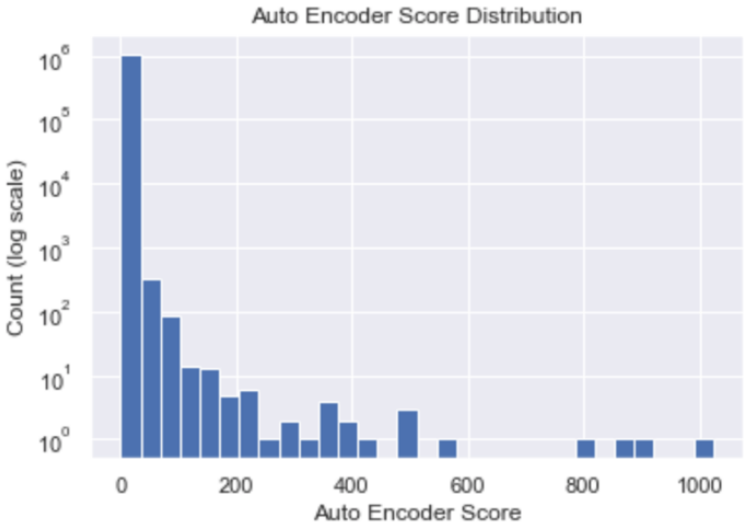
$$Score_2 = \left(\sum_{k=1}^{8} |PC_k - PC'_k|^2 \right)^{\frac{1}{2}}$$

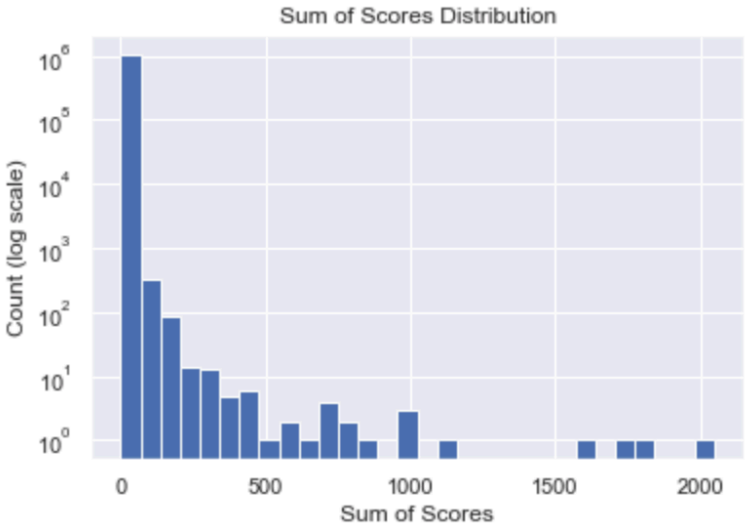
Results

After we calculated two fraud scores using different models, we examined the distributions of those two scores. We plotted the bar charts using a log-scale on the y-axis, since the distribution is highly right-skewed. As demonstrated below, most of the records have small fraud scores, meaning that they are likely to be normal, or non-fraudulent. On the other hand, the records that fall into the tail area are the potential fraud records that we want to examine closer. The distributions generated from two scores are very similar, consistently leading to some highly suspicious records. This also proves the effectiveness of our models.

Fraud Score Distributions:







Top 20 Records with the Highest Fraud Scores

BBLE	OWNER	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	ZIP	BLDFRONT	BLDDEPTH
4018420001	864163 REALTY, LLC	157	95	1	2930000	1318500	1318500	11373	1	1
4080100001	TONY CHEN	6	1	1	0	0	0		0	0
3085900700	U S GOVERNMENT OWNRD	117	108		4326303700	1946836665	1946836665		0	0
5078530085		1	1	2	836000	28800	50160	10307	36	45
4004200001	NEW YORK CITY ECONOMI	298	402	20	3443400	1549530	1549530	11101	1	1
4004590005	11-01 43RD AVENUE REA	94	165	10	3712000	252000	1670400	11101	1	1
3085910100	DEPT OF GENERAL SERVI	466	1009		2310884200	1039897890	1039897890		0	0
1012540010	PARKS AND RECREATION	4000	150	1	70214000	31455000	31596300		8	8
4142600001	LOGAN PROPERTY, INC.	4910	0	3	374019883	1792808947	4668308947	11422	0	0
4066610005E	M FLAUM	1	1		0	0	0		0	0
4155770029	PLUCHENIK, YAAKOV	91	100	2	1900000	9763	75763	11691	1	1
5000130060	RICH-NICH REALTY, LLC	136	132	8	1040000	236250	468000	10301	1	1
2056500001	PARKS AND RECREATION	600	4000	6	190000000	79200000	85500000	10462	0	0
4004200101		139	342	20	2151600	968220	968220		1	1
3011170001	CITY OF NY/PARKS AND	526	250		270500800	115650000	121725360	11215	0	0
5078120132	DRANOVSKY, VLADIMIR	96	279	3	2120000	65401	124910	10307	1	1
3044520002	STARRETT CITY, INC	7536	4356		194000000	86850000	87300000		0	0
1011110001	CULTURAL AFFAIRS	840	0		6150000000	2668500000	2767500000	10028	0	0
2049910126		1	1		0	0	0		0	0
4020180001	DEPT OF GENERAL SERVI	2213	438		432000000	108450000	194400000		0	0

According to the property OWNER field, we divided the potential fraudulent records into three different groups:

- Government-owned
- Individual-owned or NA
- Company-owned

As we could see, several records have U.S. Government or NY City as owners. The characteristics of such property is that it tends to have a large value, zero building frontage and building depth, and almost full exemption of tax. Therefore, it is reasonable that the algorithms selected these records as anomalies, and we may validate these records manually.

Properties owned by individuals or without an owner were also considered as fraudulent records by our models, since their property values are either 0 or very low. They displayed many signs indicating them as fraudulent records. For example, the property whose owner is Tony Chen has 0 market value, land value and total value; the property owned by Pluchenik Yaakov has a land value of 9763. Besides, the building frontage and building depth are usually recorded as 0 or 1.

Company-owned properties are suspicious because they usually have a higher value but very small (0 or 1) building frontage and building depth. For example, Logan Property, Inc. has a property of 1.8 billion dollars land value and 4 billion total value, yet the building frontage and depth are both 0. Therefore, these records are likely to be fraud.

Conclusions

Overall, our goal was to build unsupervised fraud models on the NY Property Data to identify fraudulent events. First, we summarized the data and also cleaned the data by filling in missing fields. Then, we created 45 variables for model-building and z-scaled the variables to put them on the same footing. Next, we performed the Principal Component Analysis to reduce the correlation among variables and reduce the dimensionalities to eight variables. Also, we z-scaled the eight variables again to give them equal importance.

In terms of the model algorithms, we generated two fraud scores using two different methods, including a heuristic function z-score model and an autoencoder model. By using the heuristic model, we could easily look for outliers on a z-based scale. By using the autoencoder model, we reproduced as many records as possible and calculated the differences between any given two records. Then, we combined the two fraud scores into a final fraud score, with a higher score indicating a higher probability of fraud. With this score, we rank-ordered all the entries and found the records that are most likely to be fraud.

Based on our results, the plots of the three fraud scores share similar distributions, which increased the reliability of identifying fraudulent events. We listed top 20 unusual records and identified them as anomalous events in terms of owners' name, market value, actual land value, actual total value, building front, building depth, street address, actual exempt land value, actual exempt total value, and zip code. One thing worthy of pointing out is that for the records with owner names that look like departments of the government, we need more expertise and suggestions to examine those results.

We also raised some further considerations that could be addressed in the project if we had more time. First, when filling missing value, we can first inspect how many missing values each record has. If a record has more than five missing values in the field, then we can filter them out instead of filling them with potentially inappropriate statistics. Besides, when creating variables for the models, we can add additional variables that are helpful to look for the fraud modes. In addition, we can develop several heuristic functions or assign weights to the scores to compare and select the best performing model. Finally, we can consult industry experts to gain more knowledge about the NY property policies and how to deal with government-owned properties, better identifying fraudulent events.

Appendix

Data Quality Report:
NYC Property Data

February 2019

6:30-9:30pm Session Team 1:
Justice League Consulting Group

Team Members:
Zongyang Jiao, Chengyin Liu, Jiayi Ma,
Xinyue Niu, Xueyan Gu, Jie Zhao

1.0 INTRODUCTION

This dataset contains New York City Property valuation and assessment data provided by the Department of Finance (DOF). The data is primarily used to calculate Property Tax, Grant eligible properties Exemptions and/or Abatements. It is updated annually

Source: https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8
Covered Time Period: 2010 - 2011
Number of Fields: 32
Number of Records: 1,070,994

2.0 FILEDS SUMMARY

a. Numeric Fields

Field Name	*Field Type*	*# Records w/ Value*	*% Populated*	*# Unique Values*	*# Records w/ value 0*	*Mean*	*STDEV*	*Min*	*Max*
LTFRONT	Numeric	1.07E+06	100.00%	1297	0	36.64	74.03	0	9999
LTDEPTH	Numeric	1.07E+06	100.00%	1370	0	88.86	76.4	0	9999
STORIES	Numeric	1.01E+06	94.75%	112	5.63E+04	5.01	8.37	1	119
FULLVAL	Numeric	1.07E+06	100.00%	1.09E+05	0	8.74E+05	1.16E+07	0	6.15E+09
AVLAND	Numeric	1.07E+06	100.00%	7.09E+04	0	8.51E+04	4.06E+06	0	2.67E+09
AVTOT	Numeric	1.07E+06	100.00%	1.13E+05	0	2.27E+05	6.88E+06	0	4.67E+09
EXLAND	Numeric	1.07E+06	100.00%	3.34E+04	0	3.64E+04	3.98E+06	0	2.67E+09
EXTOT	Numeric	1.07E+06	100.00%	6.43E+04	0	9.12E+04	6.51E+06	0	4.67E+09
BLDFRONT	Numeric	1.07E+06	100.00%	612	0	23.04	35.58	0	7575
BLDDEPTH	Numeric	1.07E+06	100.00%	621	0	39.92	42.71	0	9393
AVLAND2	Numeric	2.83E+05	26.40%	5.86E+04	7.88E+05	2.46E+05	6.18E+06	3	2.37E+09
AVTOT2	Numeric	2.83E+05	26.40%	1.11E+05	7.88E+05	7.14E+05	1.17E+07	3	4.50E+09
EXLAND2	Numeric	8.74E+04	8.17%	2.22E+04	9.84E+05	3.51E+05	1.08E+07	1	2.37E+09
EXTOT2	Numeric	1.31E+05	12.22%	4.83E+04	9.40E+05	6.57E+05	1.61E+07	7	4.50E+09

b. Categorical Fields

Field Name	*Field Type*	*# Records w/ Value*	*% Populated*	*# Unique Values*	*# Records w/ value 0*	*Most Common Value*
Record	Categorical	1.07E+06	100.00%	1.07E+06	0	NA
BBLE	Categorical	1.07E+06	100.00%	1.07E+06	0	NA
B	Categorical	1.07E+06	100.00%	5	0	4
Block	Categorical	1.07E+06	100.00%	1.40E+04	0	3944
LOT	Categorical	1.07E+06	100.00%	6366	0	1
EASEMENT	Categorical	4636	0.43%	13	1.07E+06	E
OWNER	Categorical	1.04E+06	97.04%	8.63E+05	3.17E+04	PARKCHESTER PRESERVAT
BLDGCL	Categorical	1.07E+06	100.00%	200	0	R4
TAXCLASS	Categorical	1.07E+06	100.00%	11	0	1
EXT	Categorical	1.07E+06	100.00%	4	0	G
EXCD1	Categorical	6.38E+05	59.62%	130	4.33E+05	1017
STADDR	Categorical	1.07E+06	99.94%	8.39E+05	676	501
ZIP	Categorical	1.04E+06	97.21%	197	2.99E+04	1.03E+04
EXMPTCL	Categorical	1.56E+04	1.45%	15	1.06E+06	X1
EXCD2	Categorical	9.29E+04	8.68%	61	9.78E+05	1017
PERIOD	Categorical	1.07E+06	100.00%	1	0	FINAL
YEAR	Categorical	1.07E+06	100.00%	1	0	2010/11
VALTYPE	Categorical	1.07E+06	100.00%	1	0	AC-TR

3.0 DATA QUALITY ASSESSMENT

3.01 RECORD

FILE KEY, to uniquely identify each record of the AV Master, ranging from 1 to 1,070,994

3.02 BBLE

BBLE is an alphanumeric label in the length of 11, generated by concatenating the RECORD, B and BLOCK. Unique for each observation in the dataset.

3.03 B

B stands for borough codes

Table3.1

Key	Value
1	MANHATTAN
2	BRONX
3	BROOKLYN
4	QUEENS
5	STATEN ISLAND

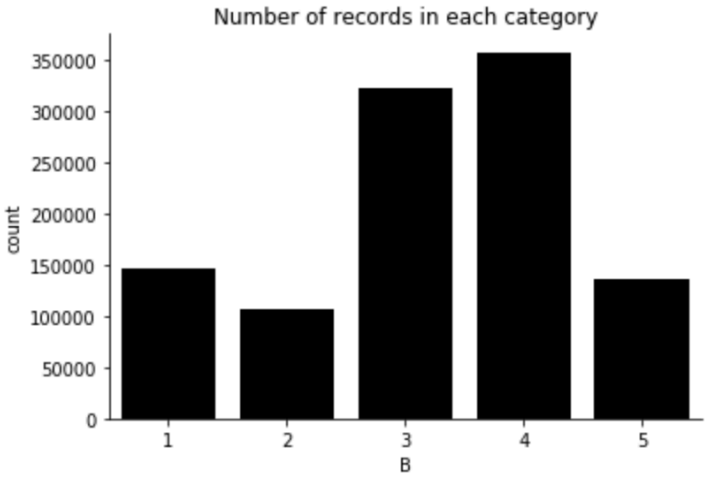


Figure 3.1

3.04 BLOCK

BLOCK stands for valid block ranges by borough

Table 3.2

Borough	Range
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

3.05 LOT

LOT stands for unique number within borough /block

Table 3.3

LOT	Count
1	24367
20	12294
15	12171
12	12143
14	12074
16	12042
17	11982
...	...

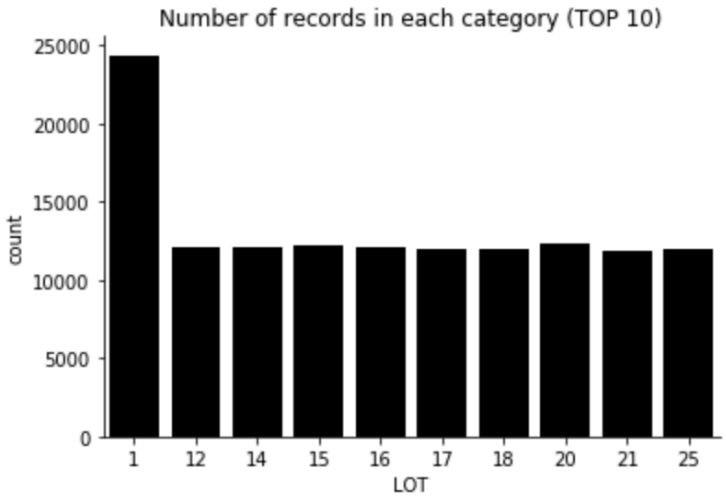


Figure 3.2

3.06 EASEMENT

EASEMENT is a field that is used to describe easement, contains 1,066,358 of NA values

Table 3.4

Filed	Description
Space	the lot has no Easement
A	the portion of the Lot that has an Air Easement
B	Non-Air Rights
E	the portion of the lot that has a Land Easement
F - M	duplicates of 'E'
N	Non-Transit Easement
P	Piers
R	Railroads
S	Street
U	U.S. Government

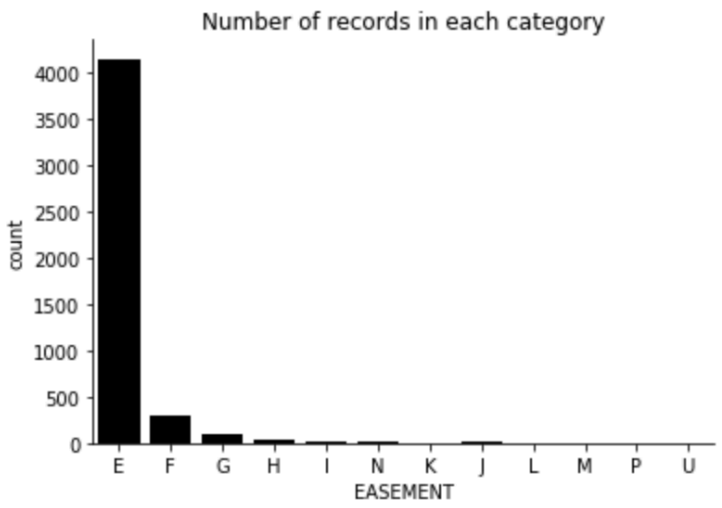


Figure 3.3

3.07 OWNER

WONER stands for owner's name, this field contains 31,745 of NA values.

Table 3.5

Owner	Count
PARKCHESTER PRESERVAT	6021
PARKS AND RECREATION	4255
DCAS	2169
HOUSING PRESERVATION	1904
CITY OF NEW YORK	1450
DEPT OF ENVIRONMENTAL	1166
BOARD OF EDUCATION	1015
...	...

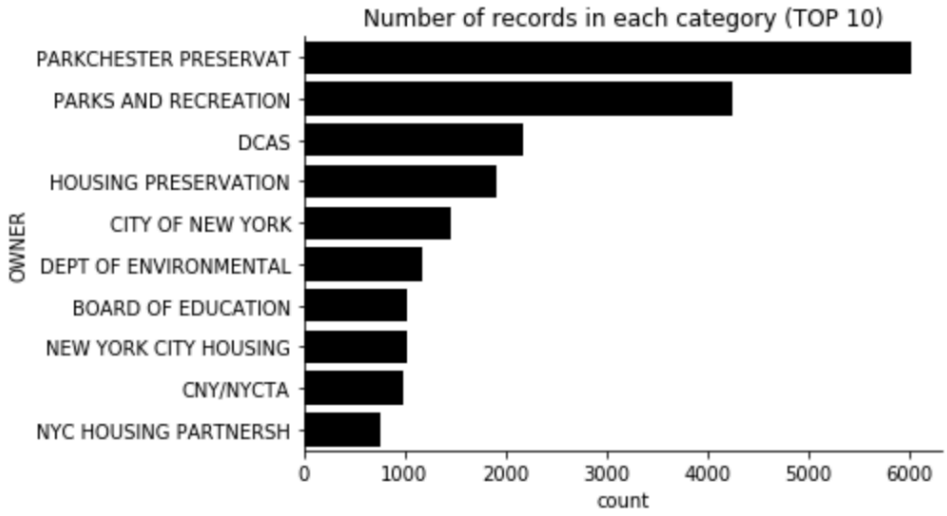


Figure 3.4

3.08 BLDGCL

BLDGCL stands for building class, where the first position is a letter and the second position is a number

Table 3.6

Field	Count
R4	139879
A1	123369
A5	96984
B1	84208
B2	77598
C0	73111
B3	59240
...	...

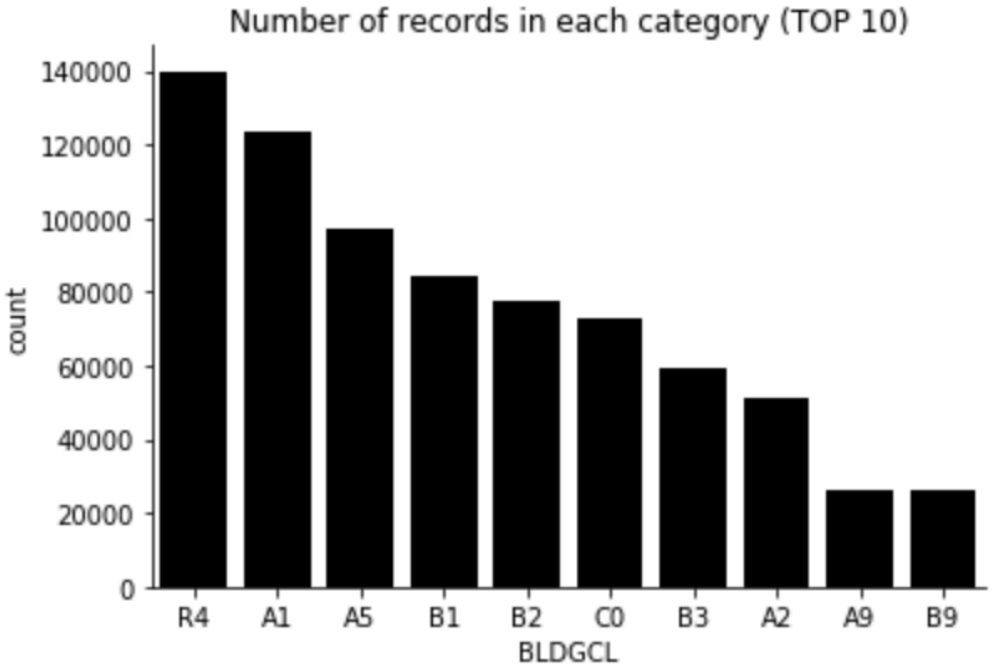


Figure 3.5

3.09 TAXCLASS

TAXCLASS stands for current property tax class code by NYS classification

Table 3.7

TAXCLASS	Count
1	660721
2	188612
4	104310
2A	40574
1B	24738
1A	21667
2B	13964
2C	10795
3	4638
1C	946
1D	29

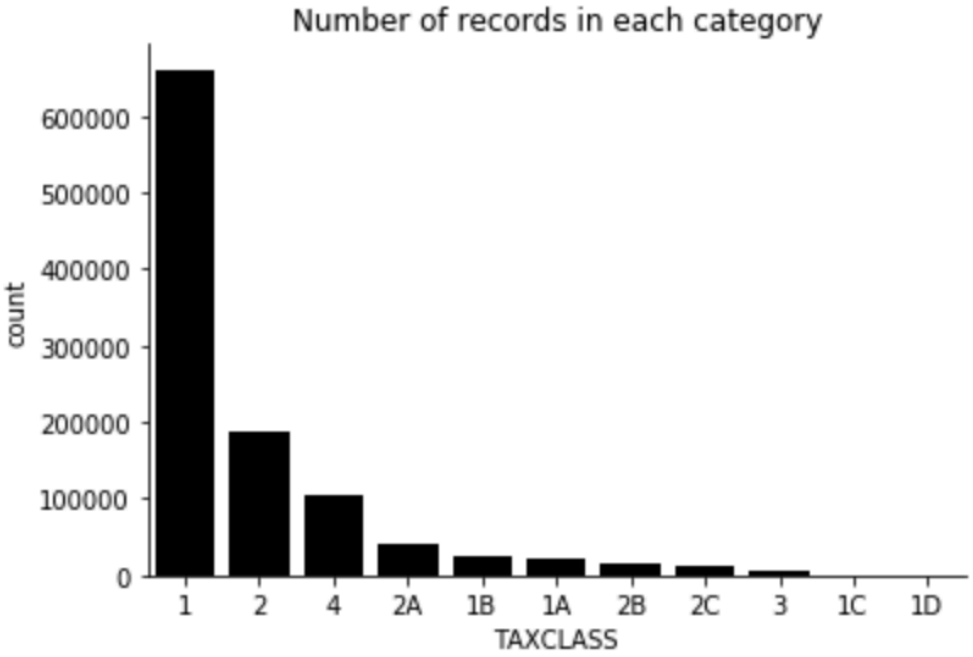


Figure 3.6

3.10 LTFRONT

LTFRONT stands for lot frontage (lot width) in feet.

Table 3.8

Unit	Max	Min	Mean	Std
Feet	0.00	9,999.00	36.63	74.03

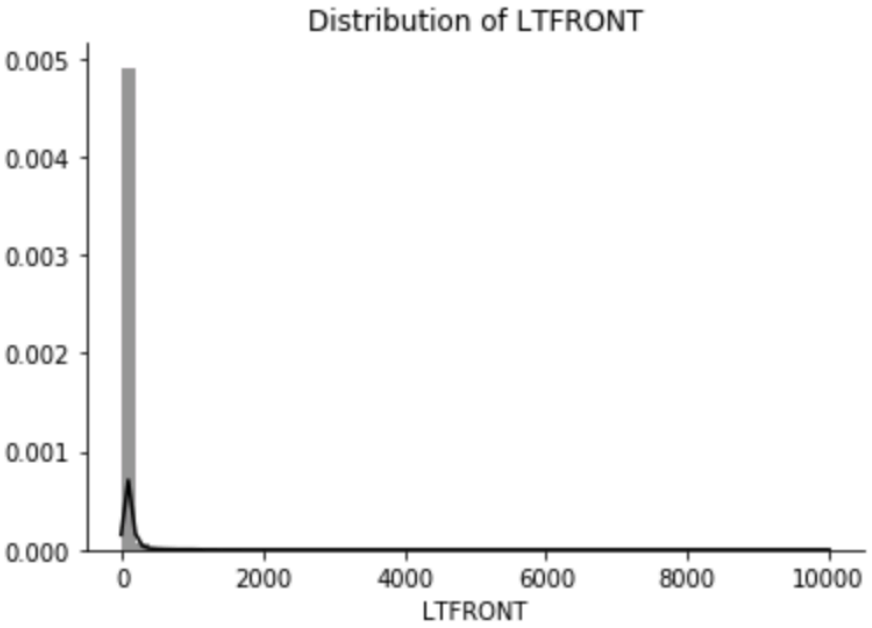


Figure 3.7 (a)

As we can see that there are outliers in the graph above, which make us cannot see the details on the left, so we try to plot the data without the outliers. (LTFRONT <= 250)

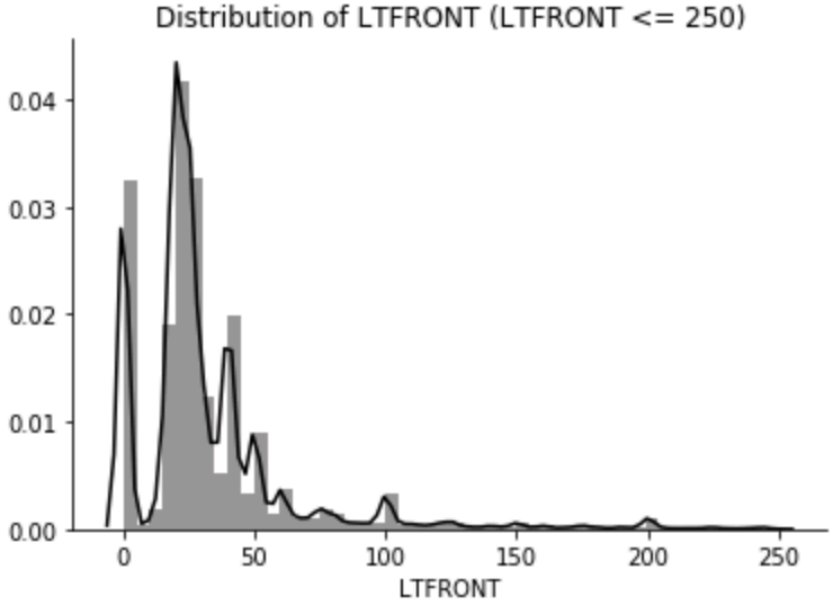


Figure 3.7 (b)

3.11 LTDEPTH

LTDEPTH stands for lot depth in feet.

Table 3.9

Unit	Max	Min	Mean	Std
Feet	0.00	9,999.00	88.86	76.40

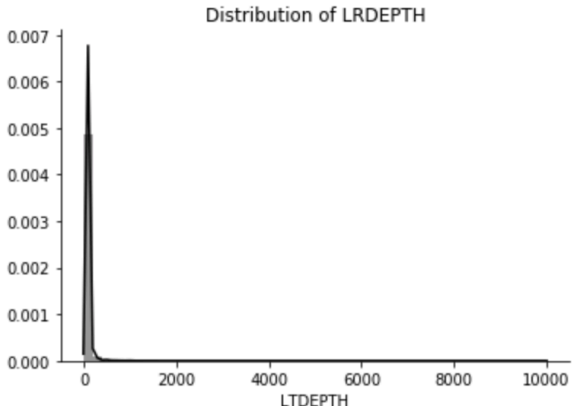


Figure 3.8 (a)

Also, because of the outliers, we cannot see details, so the plot below excludes them. (LTDEPTH <= 250)

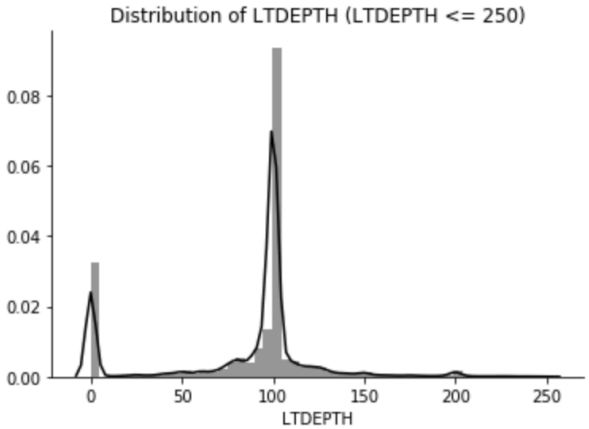


Figure 3.8 (b)

3.12 EXT

EXT stands for extension indicator. It contains NA values of 716,689

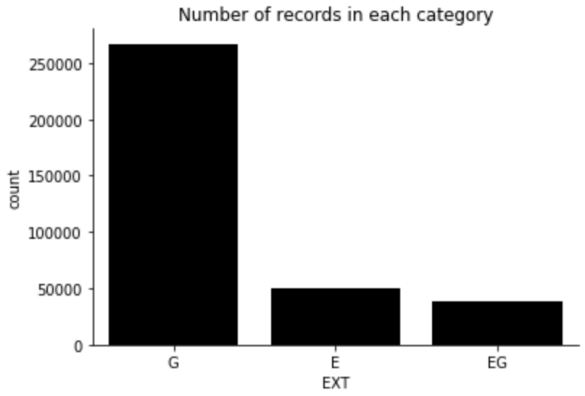


Figure 3.9

3.13 STORIES

STORIES stands for the number of stories (floors) for the building. It contains NA values of 56,264

Table 3.10

Unit	Max	Min	Mean	Std
Floor	119.00	1.00	5.01	8.37

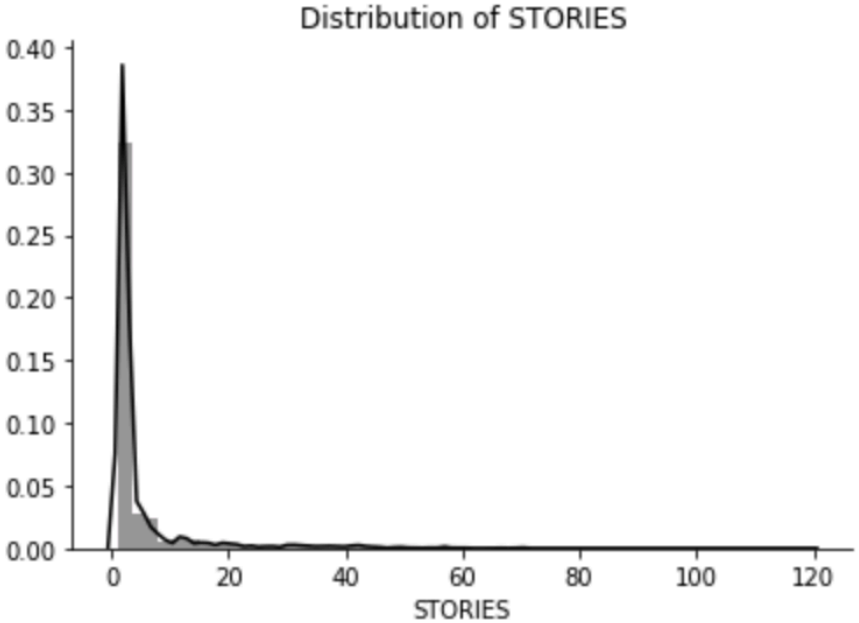


Figure 3.10 (a)

Without outliers (STORIES <= 20):

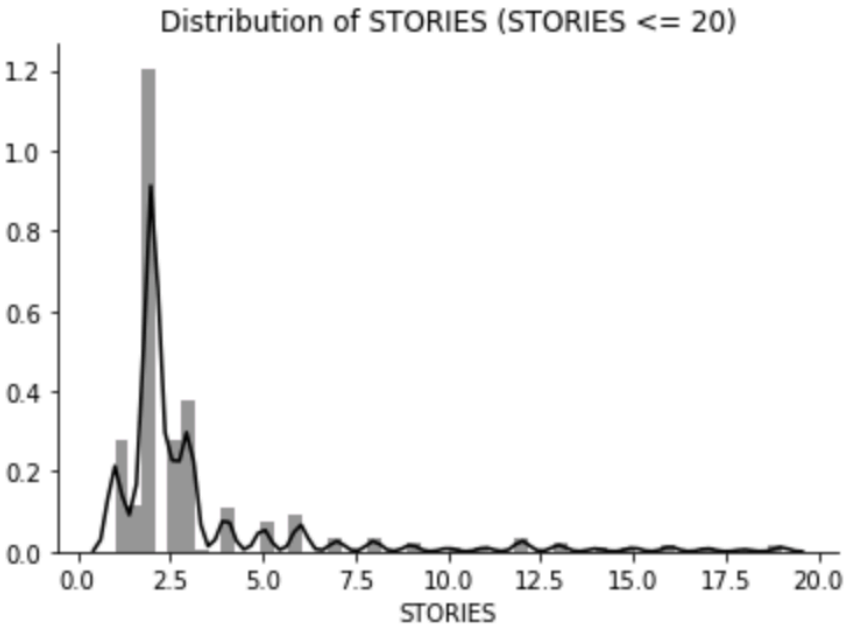


Figure 3.10 (b)

3.14 FULLVAL

FULLVAL stands for market value.

Table 3.11

Unit	Max	Min	Mean	Std
US Dollar	6,150,000,000.00	0.00	874,264.51	11,582,430.99

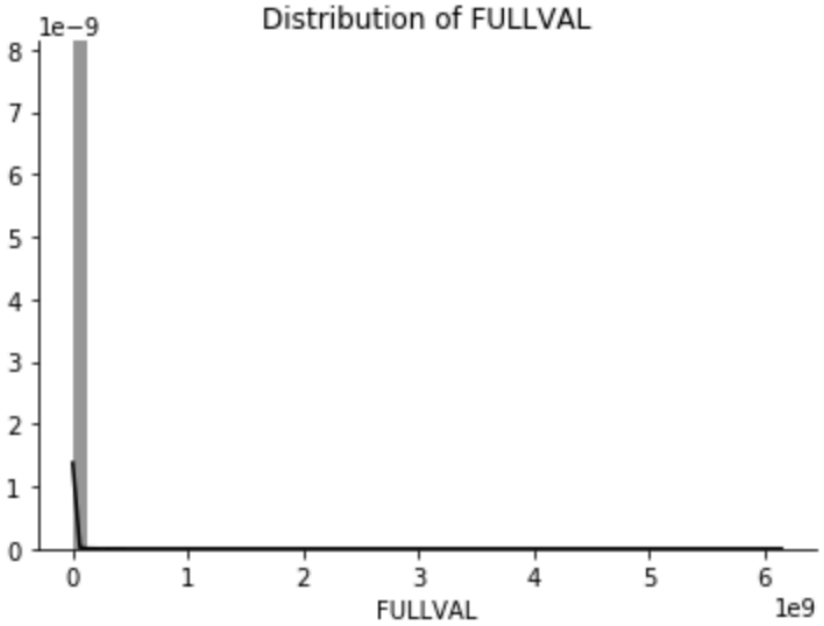


Figure 3.11 (a)

Without outliers (FULLVAL <= 5E+6):

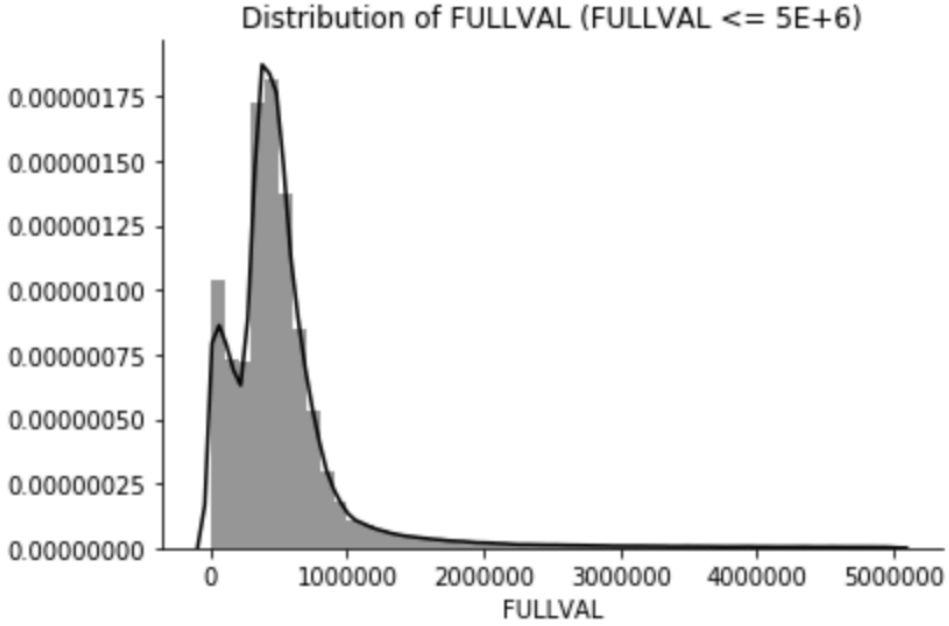


Figure 3.11 (b)

3.15 AVLAND

AVLAND stands for actual land value.

Table 3.12

Unit	Max	Min	Mean	Std
US Dollar	2,668,500,000.00	0.00	85,067.91	4,057,260.06

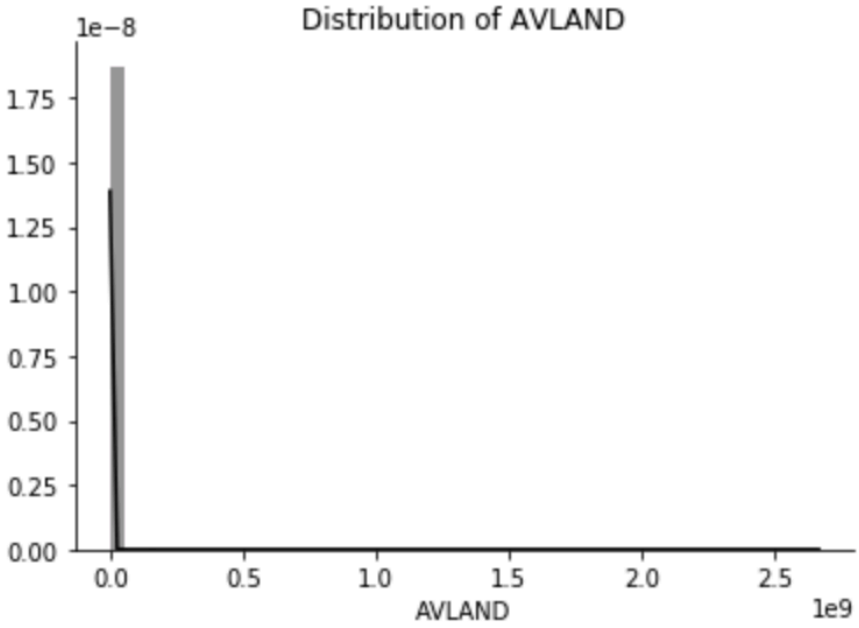


Figure 3.12 (a)

Without outliers (AVLAND <= 1.5E+6):

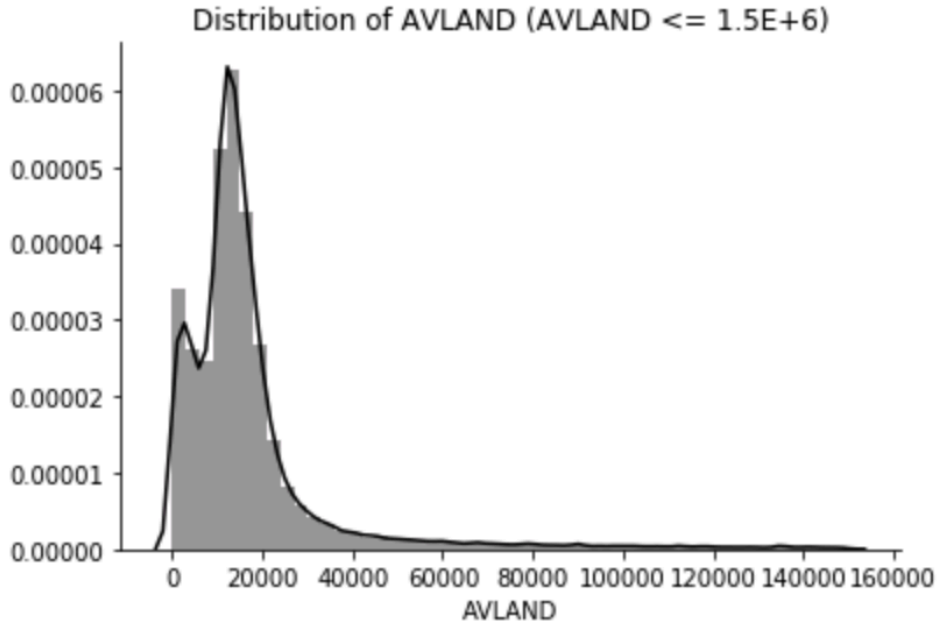


Figure 3.12 (b)

3.16 AVTOT

AVTOT stands for actual total value.

Table 3.13

Unit	Max	Min	Mean	Std
US Dollar	4,668,308,947.00	0.00	227,238.17	6,877,529.31

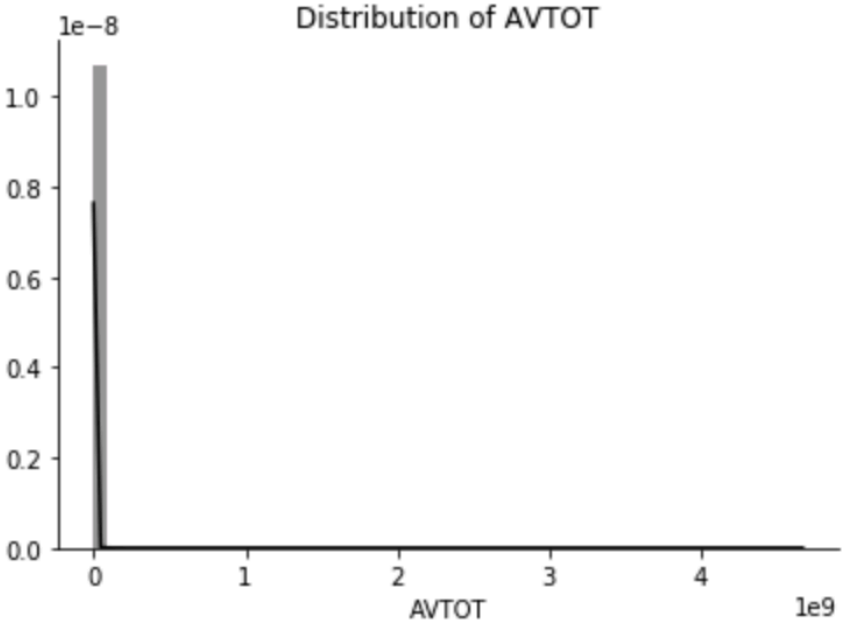


Figure 3.13 (a)

Without outliers (AVTOT <= 1.5E+6):

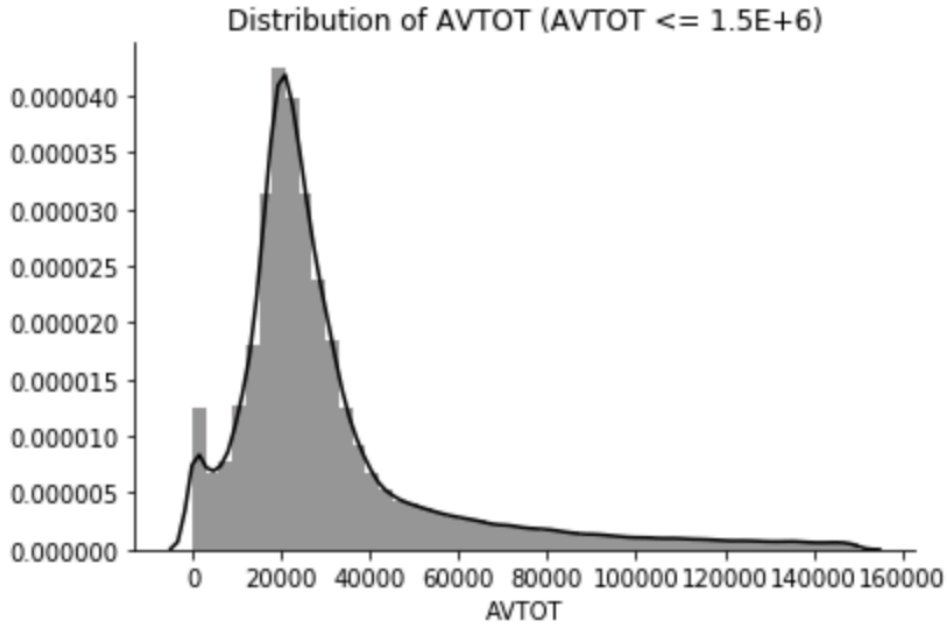


Figure 3.13 (b)

3.17 EXLAND

EXLAND stands for actual exempt land value.

Table 3.14

Unit	Max	Min	Mean	Std
US Dollar	2,668,500,000.00	0.00	36,423.89	3,981,575.79

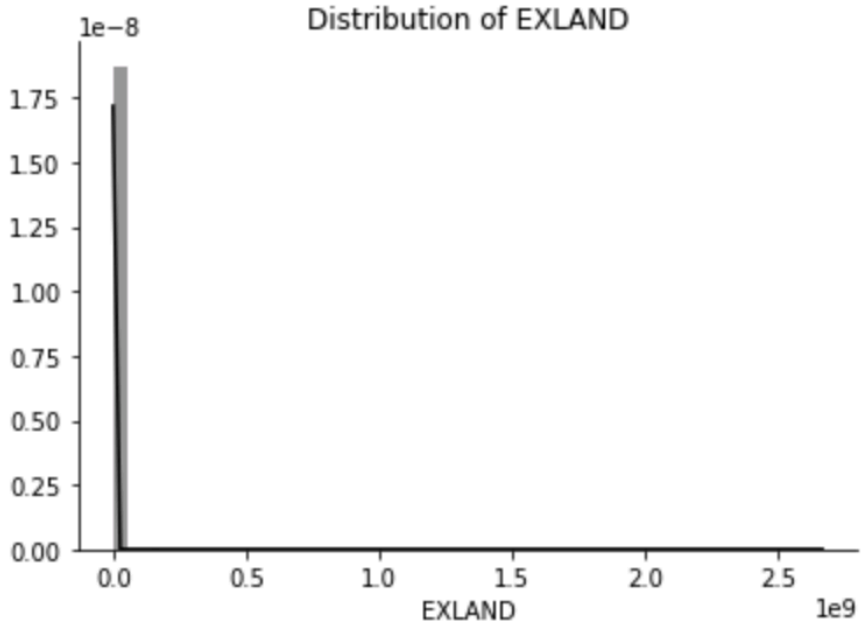


Figure 3.14 (a)

Without outliers (EXLAND <= 5000):

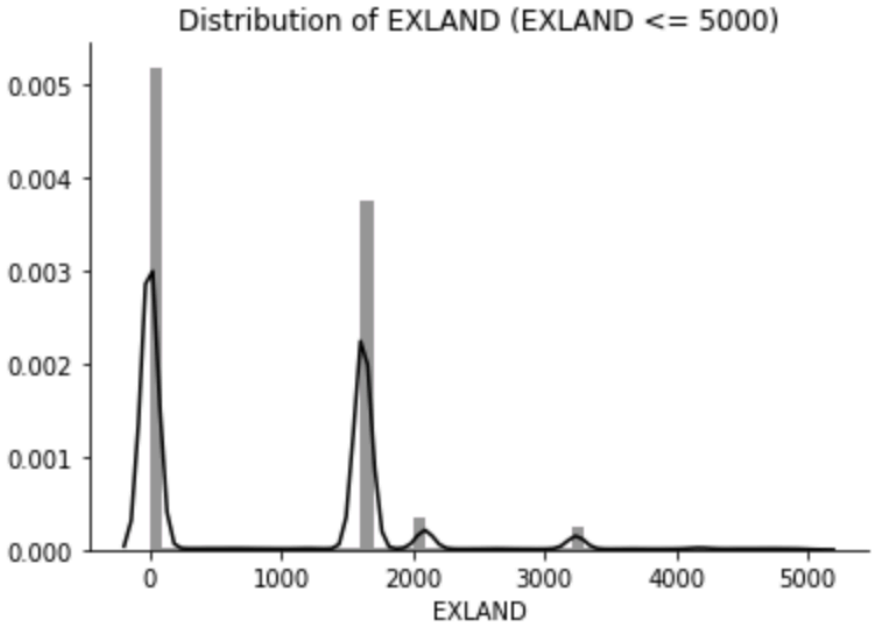


Figure 3.14 (b)

3.18 EXTOT

EXTOT stands for actual exempt land total.

Table 3.15

Unit	Max	Min	Mean	Std
US Dollar	4,668,308,947.00	0.00	91,186.98	6,508,402.82

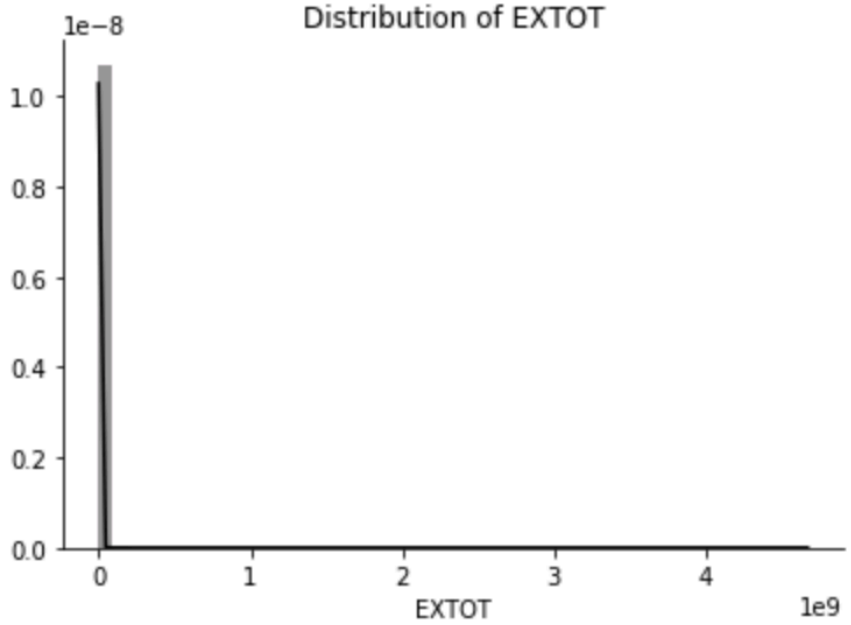


Figure 3.15 (a)

Without outliers (EXTOT <= 5000):

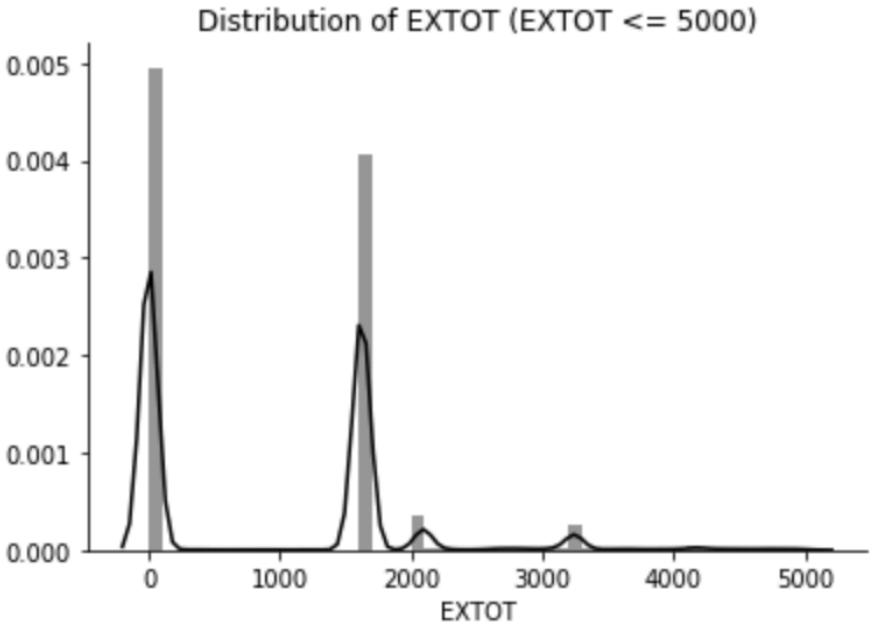


Figure 3.15 (b)

3.19 EXCD1

EXCD1 stands for exemption code 1, it contains NA values of 432,506

Table 3.16

EXCD1	count
1017	425348
1010	49756
1015	31323
5113	23858
1920	17594
5110	16834
5114	14984
…	…

3.20 STADDR

STADDR stands for street address, it contains NA values of 676

Table 3.17

STADDR	count
501 SURF AVENUE	902
330 EAST 38 STREET	817
322 WEST 57 STREET	720
155 WEST 68 STREET	671
20 WEST 64 STREET	657
1 IRVING PLACE	650
220 RIVERSIDE BOULEVARD	628
…	…

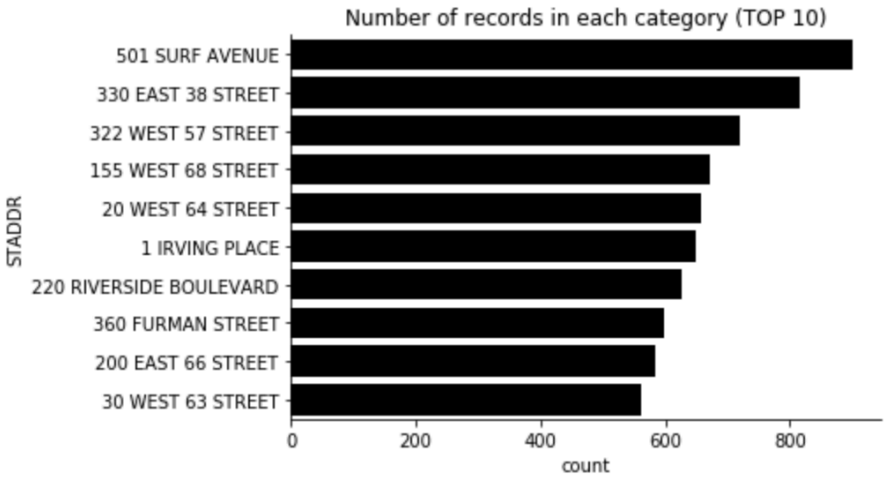


Figure 3.16

3.21 ZIP

ZIP stands for the postal zip code of the property. It contains NA values of 29,890

Table 3.18

ZIP	count
10314	24606
11234	20001
10312	18127
10462	16905
10306	16578
11236	15678
11385	14921
...	...

3.22 EXMPTCL

EXMPTCL stands for the exempt class. It contains NA values of 1,055,415

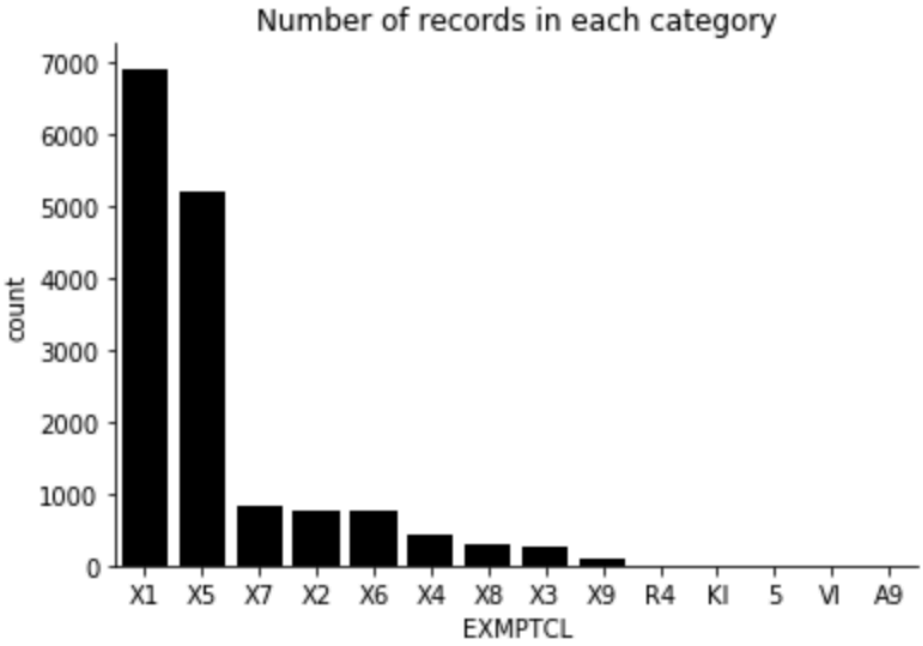


Figure 3.17

3.23 BLDFRONT

BLDFRONT stands for building frontage (width) in feet.

Table 3.19

Unit	Max	Min	Mean	Std
Feet	7,575.00	0.00	23.04	35.58

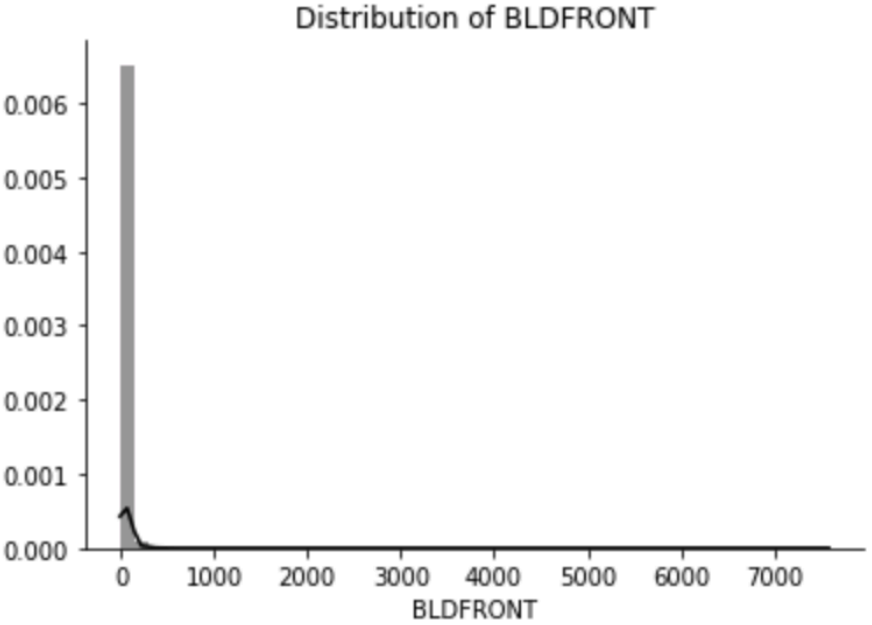


Figure 3.18 (a)

Without outliers (BLDFRONT <= 200):

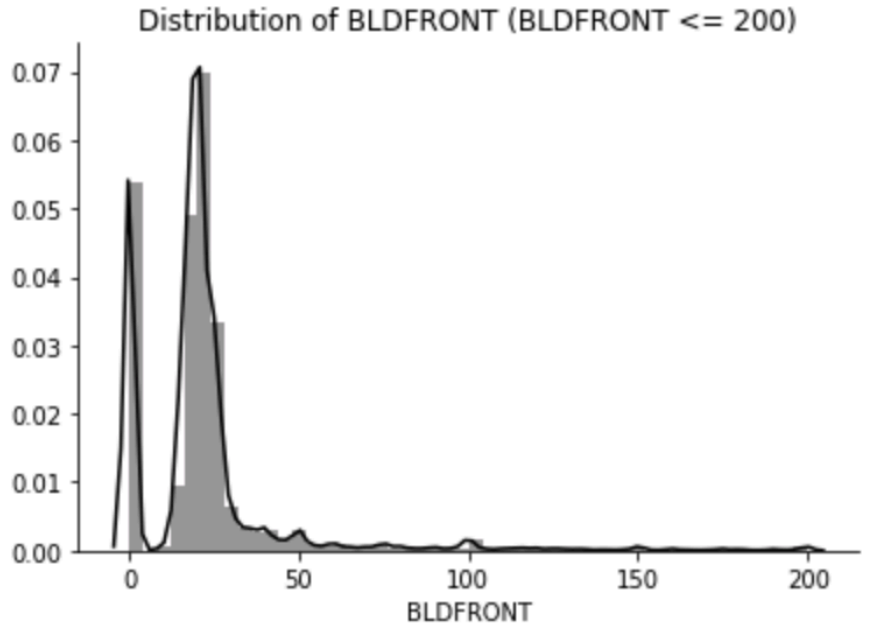


Figure 3.18 (b)

3.24 BLDDEPTH

BLDDEPTH stands for building depth in feet.

Table 3.20

Unit	Max	Min	Mean	Std
Feet	9,393.00	0.00	39.92	42.71

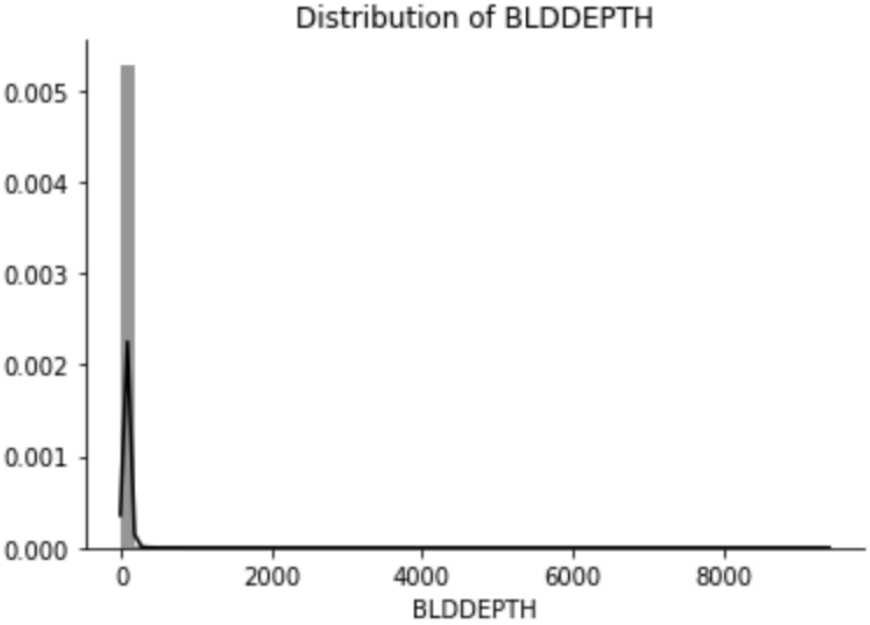


Figure 3.19 (a)

Without outliers (BLDDEPTH <= 200):

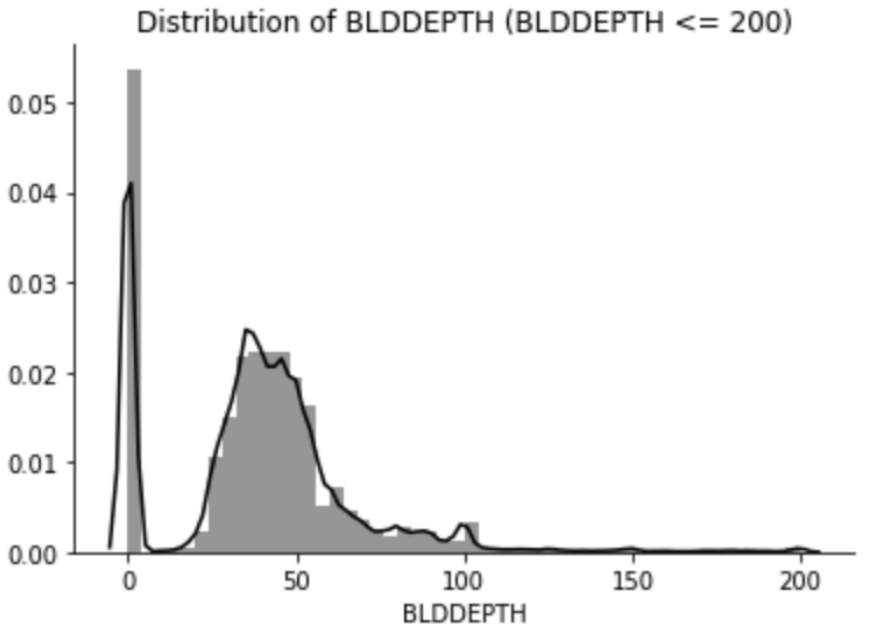


Figure 3.19 (b)

3.25 AVLAND2

AVLAND2 stands for transitional land value, it contains NA values of 788,268

Table 3.21

Unit	Max	Min	Mean	Std
US Dollar	2,371,005,000.00	0.00	246,235.72	6,178,962.56

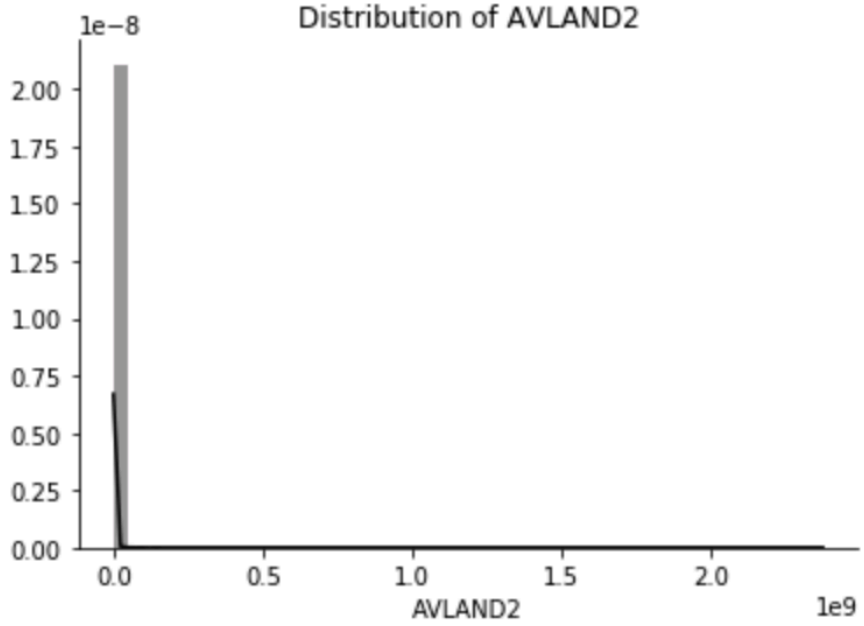


Figure 3.20 (a)

Without outliers (AVLAND2 <= 3E+5):

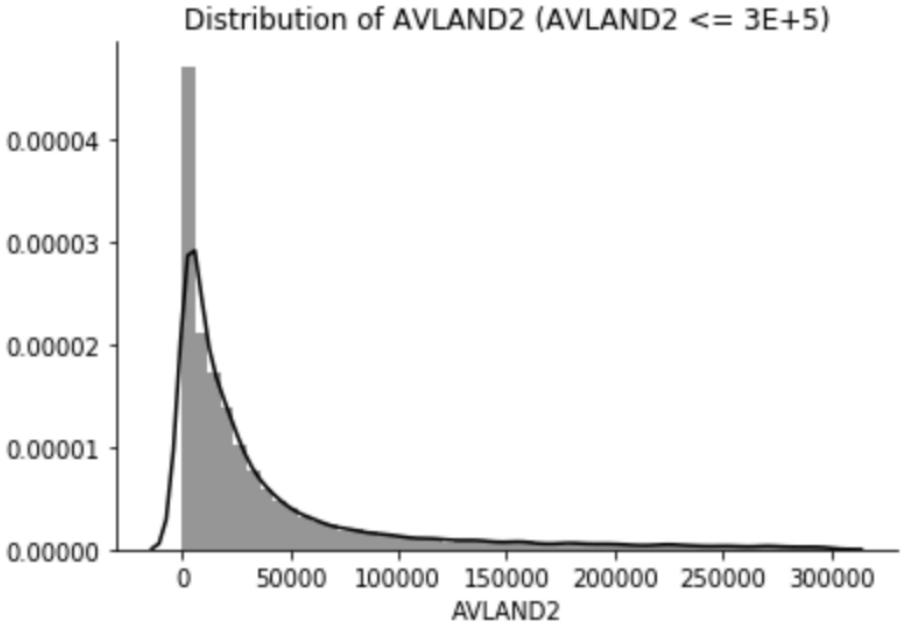


Figure 3.20 (b)

3.26 AVTOT2

AVTOT2 stands for transitional total value, it contains NA values of 788,262.

Table 3.22

Unit	Max	Min	Mean	Std
US Dollar	4,501,180,002.00	0.00	713,911.44	11,652,528.95

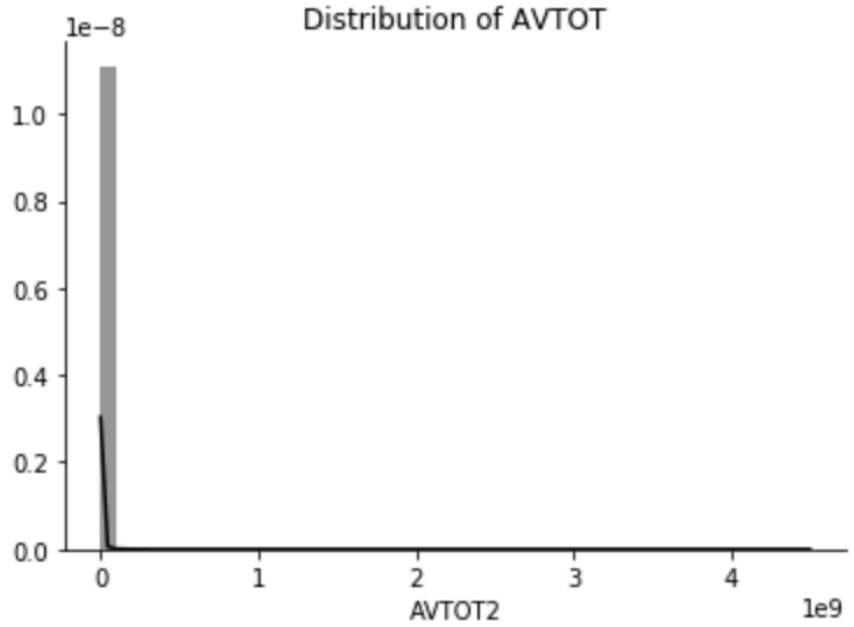


Figure 3.21 (a)

Without outliers (AVTOT2 <= 8E+5):

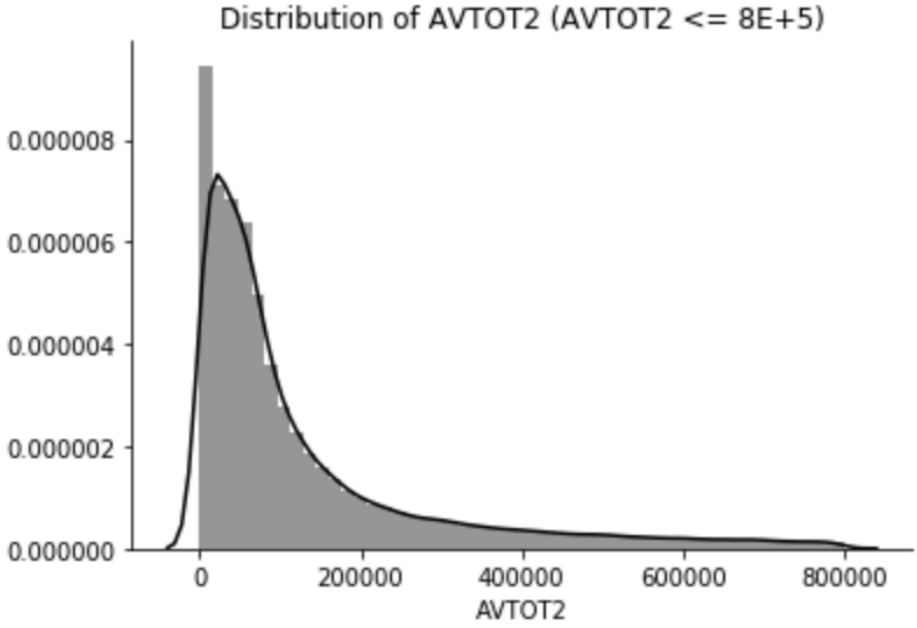


Figure 3.21 (b)

3.27 EXLAND2

EXLAND2 stands for transitional exempt land value, it contains NA values of 983,545.

Table 3.23

Unit	Max	Min	Mean	Std
US Dollar	2,371,005,000.00	0.00	351,235.68	10,802,212.67

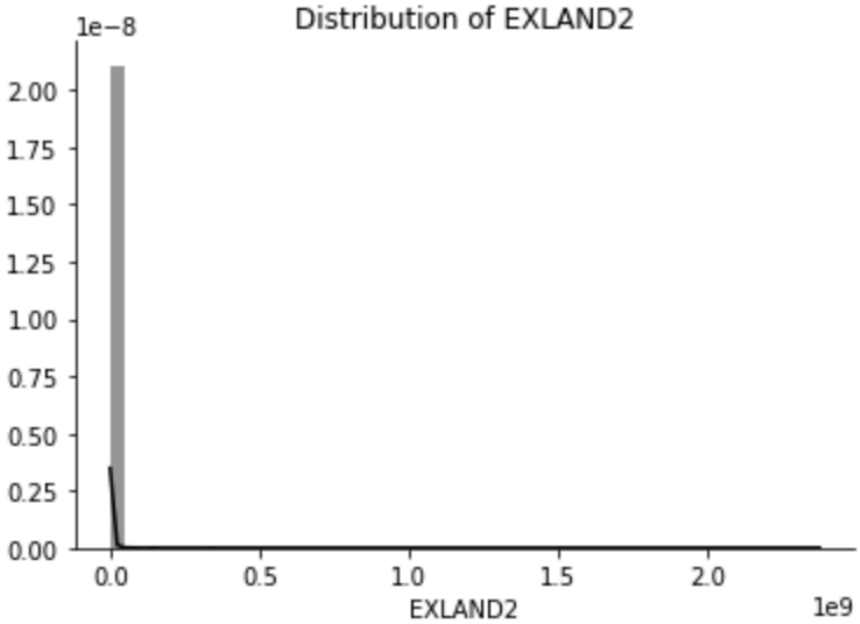


Figure 3.22 (a)

Without outliers (EXLAND2 <= 30,000):

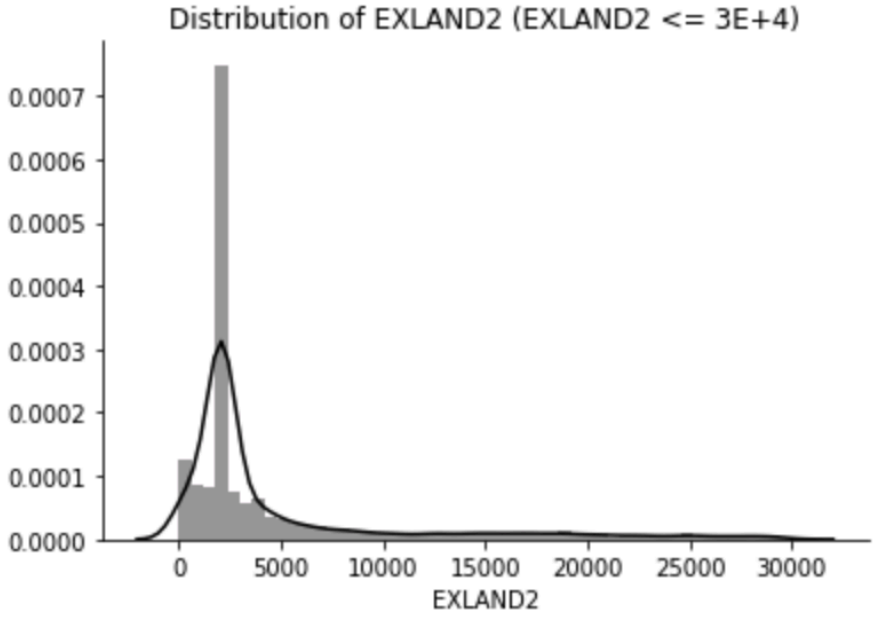


Figure 3.22 (b)

3.28 EXTOT2

EXTOT2 stands for transitional exempt land total, it contains NA values of 940,166.

Table 3.24

Unit	Max	Min	Mean	Std
US Dollar	4,501,180,002.00	0.00	656,768.28	16,072,510.17

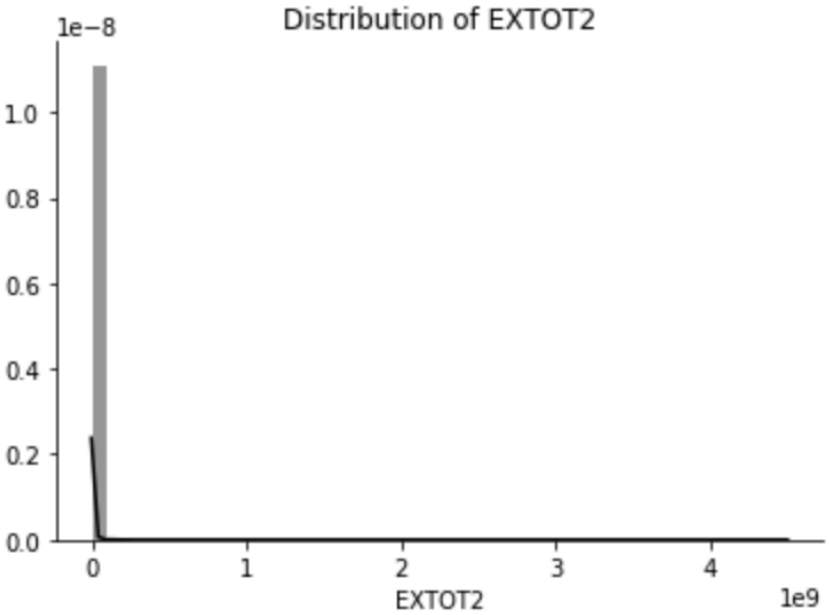


Figure 3.23 (a)

Without outliers (EXTOT2 <= 1E+5):

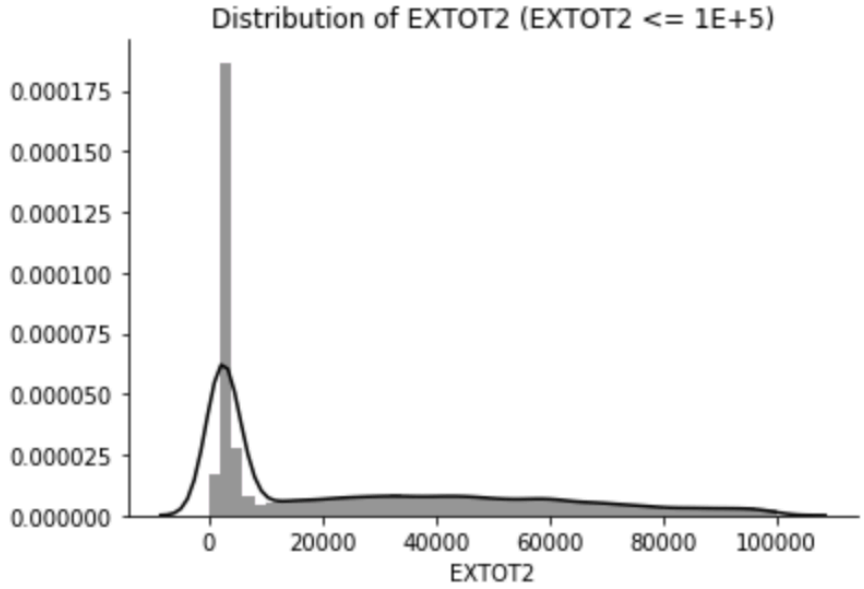


Figure 3.23 (b)

3.29 EXCD2

EXCD2 stands for exemption code 2, it contains NA values of 978,046

Table 3.25

EXCD2	count
1017	65777
1015	12337
5112	6867
1019	3178
1920	2961
1200	881
1101	494
…	…

3.30 PERIOD

Period stands for assessment period when file was created

3.31 YEAR

YEAR stands assessment year. In this dataset, we only have 2010/11 for this field

3.32 VALTYPE

In this dataset, we only have VALTYPE of “AC-TR”