Property Fraud Identification For the City of New York

Tianyi Wang, Jiayuan Zhao, Yinuo Chen, Yating Zhou, Chen Li

MSBA 2021 Cohort

Rady School of Management, UC San Diego

MGTA 463 Fraud Analytics

April 26th, 2021

Professor Stephen Coggeshall

Project Advisor

Table of Content

Executive Summary 1

1 **Data Description** 2

1.1 File Description 2

1.2 Summary Statistics Table 2

1.3 Field Examples 3

1.3.1 Field “LTFRONT” 3

1.3.2 Field “LTDEPTH” 4

1.3.3 Field “BLDFRONT” 5

1.3.4 Field “BLDDEPTH” 6

1.3.5 Field “STORIES” 6

2 **Data Cleaning** 8

3 **Feature Creation** 9

3.1 Unit Value 9

3.2 Ratio of Value 10

4 **Dimensionality Reduction** 12

4.1 Feature Scaling 12

4.2 Applying PCA 12

4.3 Rescaling the Data 13

5 **Fraud Model Algorithms** 14

5.1 Model 1: Z Score Outliers 14

5.2 Model 2: Autoencoder Error 14

5.3 Final Score 14

6 **Results** 15

7 **Summary and Conclusions** 20

Appendix A: Data Quality Report 21

Executive Summary

In the city of New York, property tax fraud is a reoccurring type of action where owners deliberately misrepresent their property characteristics thus underpaying property tax. The city of New York noticed this suspicious activity so they hired us to do this project.

This project uses the dataset provided by the New York City government. The dataset, containing more than a million property records, provides large fields of information on these property records such as owner, address, and assessed tax values. The mission of this project is to create an algorithm that can assess all the records and rank them based on how anomalous they are.

The final report verdict is an overall ranking of property records, with records that are most likely to be fraudulent at the top. To achieve this, our team has gone through the following steps:

1. Exploratory Data Analysis - Analyze and explore characteristics (origin, period, number of records, fields) of the data. We then create graphs like summary tables, graphs to evaluate the quality of the dataset.
2. Data Cleaning – Exclude government-owned properties for this project. Then, we fill in missing values for fields that require calculations (ZIP, STORIES, FULLVAL) to prepare the dataset for the next step.
3. Feature Creation – create 45 new variables based on the original dataset and then calculate statistical properties (mean, max, standard deviation, min) of these variables.
4. Dimensionality Reduction – Represent the high dimensional data in the lower dimensions in order to (1) improve the quality of model inputs and (2) accelerate algorithms when working on data with a large number of features.
5. Fraud Model Algorithms– Establish two models for fraud score calculation: (1) Z score outliers that calculates the Euclidean distance of each record and (2) Autoencoder Error which is the difference between the input and output of the autoencoder vectors. Each record will be used to calculate its two fraud scores.
6. Summarize Results and Provide Conclusion – Combine the two unscaled scores to get the final ranking of the records, sort them in descending order to get the final ranking. We then observe the highly ranked records and investigate why they get labeled as anomalous.

1 Data Description

1.1 File Description

The "Property Valuation and Assessment Data" is collected and entered into the system by various City employee, such as Property Assessors, Property Exemption specialists, ACRIS reporting, Department of Building reporting, etc. It stores 1,070,994 records of real estate assessment property data. And it covers sixteen numeric fields, including lot frontage in feet, lot depth in feet, total market value of the land, assess land value, exempt land value, etc., and sixteen categorical fields, including record, owner's name, building class, tax class, and postal zip code of the property, etc. The dataset came from NYC OpenData and was shared by Professor Stephen Coggeshall in March 2021.

Table 1.1: File Description

Dataset Name	Property Valuation and Assessment Data
Dataset Purpose	Calculate property tax, grant eligible properties exemptions and/or abatements
Data Source	NYC Open Data
Time Period	17-Nov-10
Number of Fields	32
Number of Records	1,070,994

1.2 Summary Statistics Table

1.2.1 Numeric Fields

Ten of sixteen fields are fully populated. Key statistics of these fields are summarized as follows.

Table 1.2.1: Summary Statistics of Numeric Fields

Field Name	# records	# records with value zero	# unique values	% Populated	Mean	Standard Deviation	Minimum Value	Maximum Value
LTFRONT	1070994	169108	1297	100	36.64	74.03	0	9999
LTDEPTH	1070994	170128	1370	100	88.86	76.40	0	9999
STORIES	1014730	0	112	94.75	5.01	8.37	1	119
FULLVAL	1070994	13007	109324	100	874264.51	11582430.99	0	6150000000
AVLAND	1070994	13009	70921	100	85067.92	4057260.06	0	2668500000
AVTOT	1070994	13007	112914	100	227238.17	6877529.31	0	4668308947
EXLAND	1070994	491699	33419	100	36423.89	3981575.79	0	2668500000
EXTOT	1070994	432572	64255	100	91186.98	6508402.82	0	4668308947

EXCD1	638488	0	130	59.62	1602.01	1384.23	1010	7170
BLDFRONT	1070994	228815	612	100	23.04	35.58	0	7575
BLDDEPTH	1070994	228853	621	100	39.92	42.71	0	9393
AVLAND2	282726	0	58592	26.40	246235.72	6178962.56	3	2371005000
AVTOT2	282732	0	111361	26.40	713911.44	11652528.95	3	4501180002
EXLAND2	87449	0	22196	8.17	351235.68	10802212.67	1	2371005000
EXTOT2	130828	0	48349	12.22	656768.28	16072510.17	7	4501180002
EXCD2	92948	0	61	8.68	1364.04	1094.71	1011	7160

1.2.2 Categorical Variables

Ten of sixteen fields are fully populated. Key statistics of these fields are summarized as follows.

Table 1.2.2: Summary Statistics of Categorical Fields

Field Name	# records	# unique values	% Populated	Most Common Field Value
RECORD	1070994	1070994	100	N/A*
BBLE	1070994	1070994	100	N/A*
B	1070994	5	100	4
BLOCK	1070994	13984	100	3944
LOT	1070994	6366	100	1
EASEMENT	4636	13	0.433	E
OWNER	1039249	863348	97.036	PARKCHESTER PRESERVAT
BLDGCL	1070994	200	100	R4
TAXCLASS	1070994	11	100	1
EXT	354305	4	33.082	G
STADDR	1070318	839281	99.937	501 SURF AVENUE
ZIP	1041104	197	97.209	10314
EXMPTCL	15579	15	1.455	X1
PERIOD	1070994	1	100	FINAL
YEAR	1070994	1	100	2010/11
VALTYPE	1070994	1	100	AC-TR

* All values in the RECORD and BBLE field are unique, thus no such a most common field value.

1.3 Field Examples

1.3.1 Field “LTFRONT”

Table 1.3.1: LTFRONT

Description	Lot Frontage in feet
Type	Numeric
Min	0
Max	9999
Mean	36.64
Standard Deviation	74.03

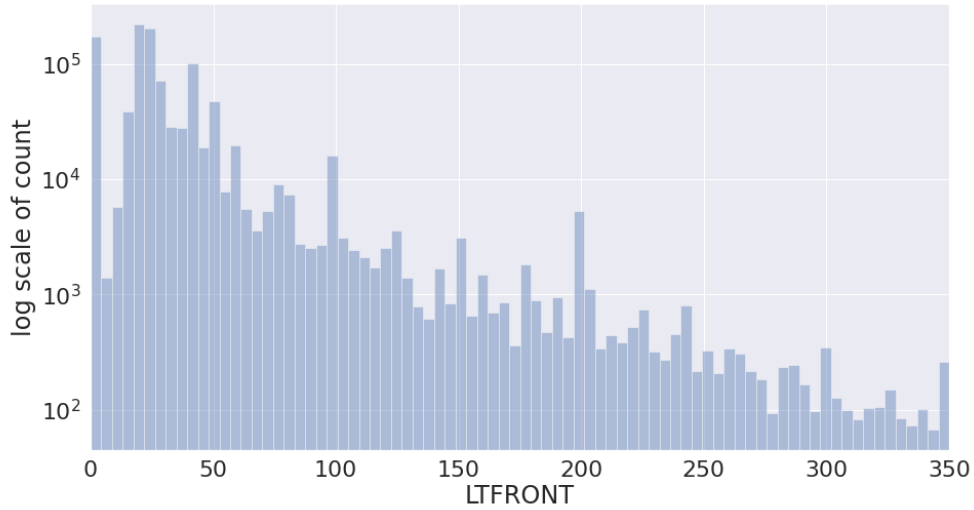


Figure 1.3.1: Frequency Distribution of the *LTFRONT* Field
(Exclude outliers > 350; Data in histogram is 99.42% populated)

1.3.2 Field "LTDEPTH"

Table 1.3.2: LTDEPTH

Description	Lot Depth in feet
Type	Numeric
Min	0
Max	9999
Mean	88.86
Standard Deviation	76.4

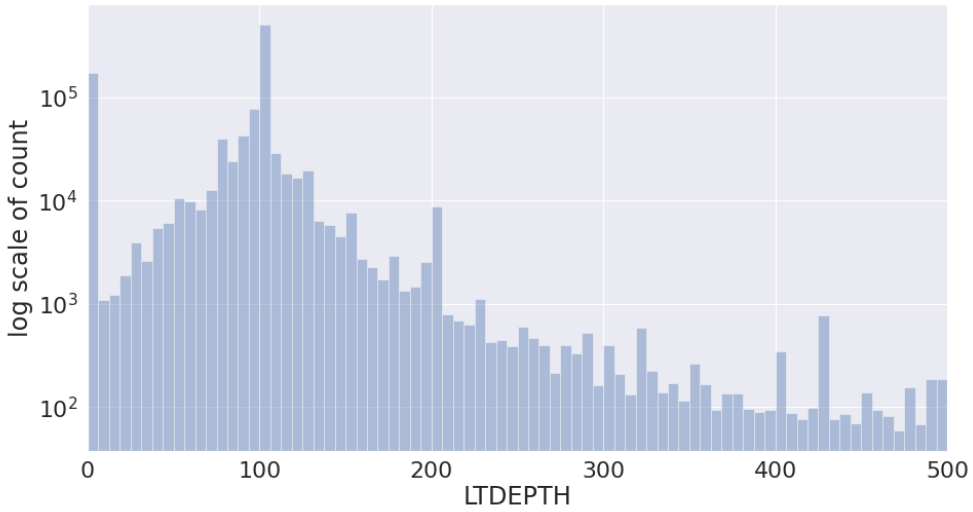


Figure 1.3.2: Frequency Distribution of the LTDEPTH Field
(Exclude outliers > 500; Data in histogram is 99.68% populated)

1.3.3 Field "BLDFRONT"

Table 1.3.3: BLDFRONT

Description	Building frontage in feet
Type	Numeric
MIN	0
MAX	7575
Mean	23.04
Standard Deviation	35.58

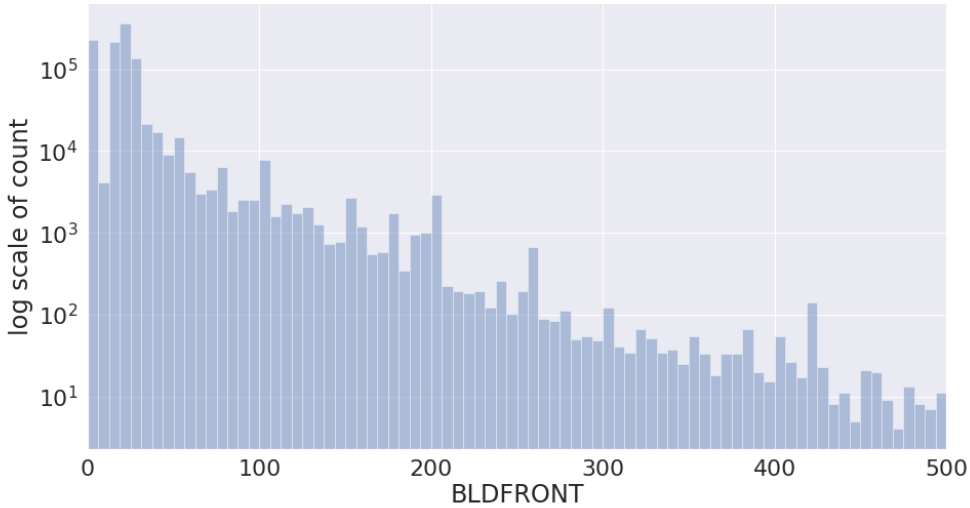


Figure 1.3.3: Frequency Distribution of the BLDFRONT Field
(Exclude outliers > 500; Data in histogram is 99.98%)

1.3.4 Field “BLDDEPTH”

Table 1.3.4: BLDDEPTH

Description	Lot depth in feet
Type	Numeric
Min	0
Max	9393
Mean	39.92
Standard Deviation	42.71

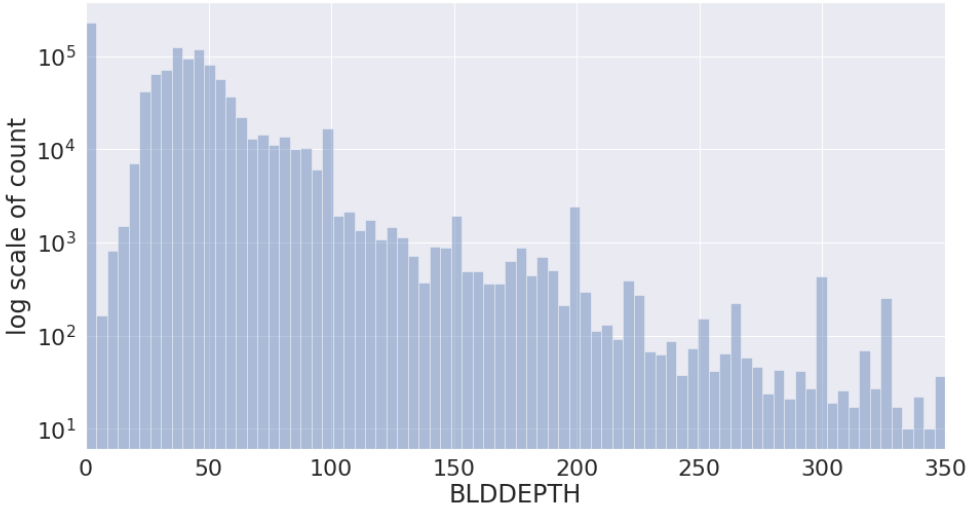


Figure 1.3.4: Frequency Distribution of the *BLDDEPTH* Field
(Exclude outliers > 350; Data in histogram is 99.87%)

1.3.5 Field “STORIES”

Table 1.3.5: STORIES

Description	The number of stories(floors) for the building (number of floors)
Type	Numeric
MIN	1
MAX	119
Mean	5.01
Standard Deviation	8.37

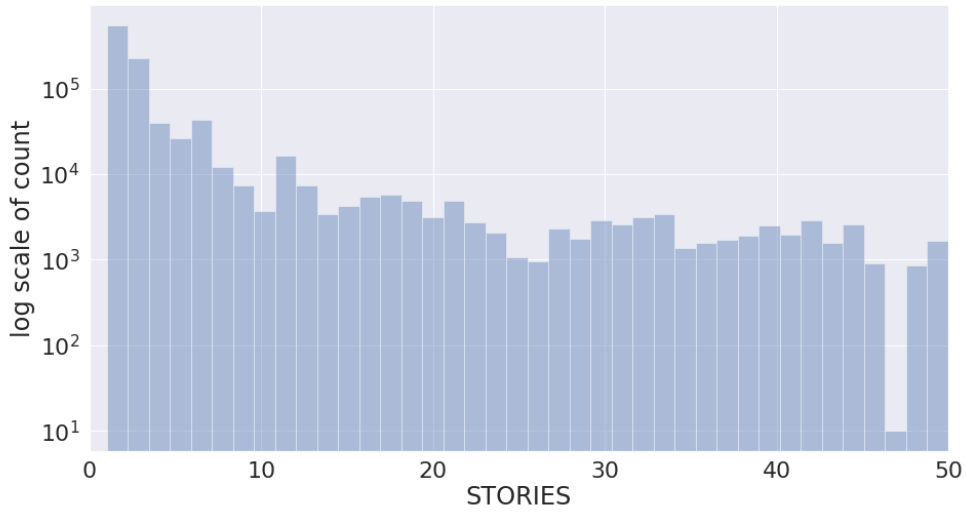


Figure 1.3.5: Frequency Distribution of the *STORIES* Field
(Exclude outliers > 50; Data in histogram is 99.50% populated among 1014730 records)

2 Data Cleaning

The main approach for this property fraud detection project is to find anomalies for property value which might be the potential property tax fraud. During data exploration, we noticed that there are plenty of properties owned by city, state or federal government, which should not be subjects of our investigation. These records would skew the statistics and variable values, thereby affecting our results. To solve this problem, we removed all records owned by city, state or federal government.

The fields about property value are FULLVAL, AVLAND, AVTOT. The property value would be influenced by the location, property type, and property size. The relative fields would be ZIP, TAXCLASS, LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, STORIES. However, we found that there are several missing values in these fields. To deal with this, we need to replace these missing values with innocuous values that won't set off a fraud alarm. Table 2.1 shows how we replaced those missing values.

Table 2.1: Missing Values

TYPE	Field	Missing Value	Replace with
Location	ZIP	NA	First, if the ZIP on both before and after the record with missing ZIP are the same, replace by that ZIP code; Then, replace all missing ZIPs with the ZIP from the record above it.
Value	FULLVAL, AVLAND, AVTOT	NA/0	The average by that record's TAXCLASS
Size	STORIES	NA	The average by that record's TAXCLASS
Size	LOTFRONT, LOTDEPTH, BLDFRONT, BLDDEPTH	NA/ 0/1	The average by that record's TAXCLASS

3 Feature Creation

A total of 45 candidate variables were created based on the existing value fields (FULLVAL, AVTOT, AVLAND) to build metric for property value, thus better quantifying the characteristics of fraud behaviors. Table 3.0 summarizes the description of variables created in each category.

Table 3.0 Summary of 45 Candidate Variables

Category	**Description**	**# of Variables Created**
Unit Value	Unit values for properties, including price per square foot for both building and land, price per building volume (building area * stories)	9
Ratio of Value	Ratio over average value grouped by locations or property types	36

3.1 Unit Value

In this dataset, the property value is represented by the full value of the property (FULLVAL), the assessed land value (AVLAND) and the assessed total value (AVTOT). The general principle is that the bigger the property the more expensive. To better evaluate the property value, Unit Value was employed to represent the standardized value.

The property size can be expressed by both the area and the volume. There are three types of property size, the lot area, the building area and the building volume. Then we calculated the unit value by standardizing the property value by its size, which is done by the logistic function below:

$$\text{Unit Value} = \frac{property\ value}{property\ size},$$

Where the property value and property size denote the variables below.

Table 3.1 Types of Property Value and Property Size

Property Value	**Property Size**
FULLVAL: full value of property	Lot Area = Lot Front * Lot Depth
AVLAND: assessed value of land	Building Area = Building Front * Building Depth

AVTOT: assessed total value	Building Volume = Building Area * Stories

* 3 × 3 = 9 *unit value* variables

3.2 Ratio of Value

The valuation of properties is also highly related with their location and property types. The properties in the same region or having the same building types tend to have similar unit values. Since the goal of this project is to find unusual property records, it is reasonable to categorize properties by their location and property type and then evaluate its abnormality in the given group.

ZIP code is a reasonable indicator of location which can be used to categorize properties. Also, we divided these properties into larger groups by the first three digits of zip code, which represents a sectional center facility or a mail processing facility area. Borough means a town or a district which is an administrative unit and it is also a good category of location. While in terms of property types, since the current property tax class code classifies properties by their building types. It would be a great indicator of the property type. Therefore, we created three types of location categories and one type of property type category as below.

Table 3.2 Types of Location Categories

Category	Description
ZIP	Postal code which divides the area by the delivery addresses
ZIP3	The first 3 digits of ZIP code which represents a sectional center facility or a mail processing facility area
Borough	A town or a district which is an administrative unit
TAXCLASS	The current property tax class code which classifies different buildings by its building types.

To evaluate the abnormality of a property, we calculated the average value of each group and checked the ratio of property's unit value over the group average. A high value of ratio variables indicate that the property's unit value is far from the group average, which could be a reasonable alarm of abnormality. Figure 3.2 shows how a total of 36 ratios of value variables were calculated.

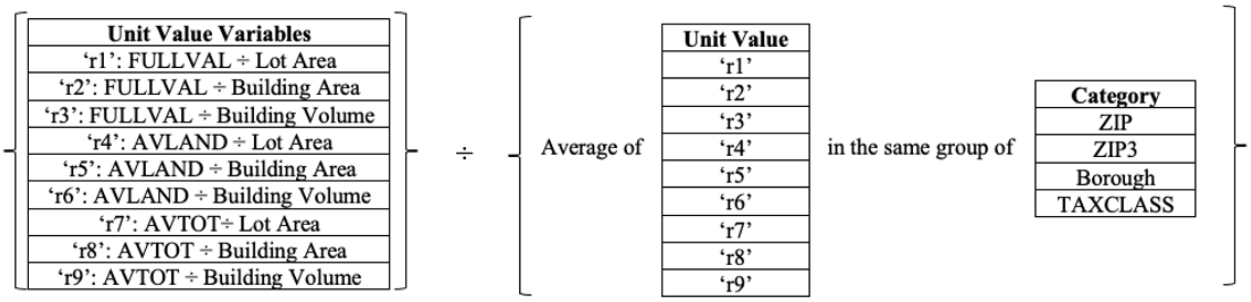


Figure 3.2 Creation of 36 *ratio of value* variables

∗ 9 × 4 = 36 *ratio of value* variables

4 Dimensionality Reduction

Dimensionality reduction is a technique to represent high dimensional data in the lower dimensions without losing much information. It has two benefits:

1. Improve the quality of model inputs by removing the noises, such as linear correlation among features.
2. Speed up algorithms when working on the data with a large number of features

In this project, we will use Principal Component Analysis(PCA) to conduct dimensionality reduction, which compresses data by rotating coordinates and throwing away meaningless dimensions.

4.1 Feature Scaling

Z scaling the data is the prerequisite to perform the PCA because it is only reasonable to compare features on the same scale. For each dimension, we will minus data by its mean (μ_i) and then divide them by standard deviation (σ_i). By doing so, the data center at the origin and are in a comparable range in terms of features.

$$Z_i = \frac{x_i - \mu_i}{\sigma_i}$$

4.2 Applying PCA

On the processed data, the general procedure of PCA is as follows:

1. find the dimensions into which data could map with maximum variance or most widely spread
2. rotate the coordinates to the new direction. Since these new dimensions are important to the representation, they are called principal components.
3. sort the dimensions by their variances
4. use a screen plot to select the least number of dimensions that account for 80% to 90% of the total variance
5. represent the data in the selected dimensions

Based on Figure 4.1, we will keep the first six components and use them to represent data.

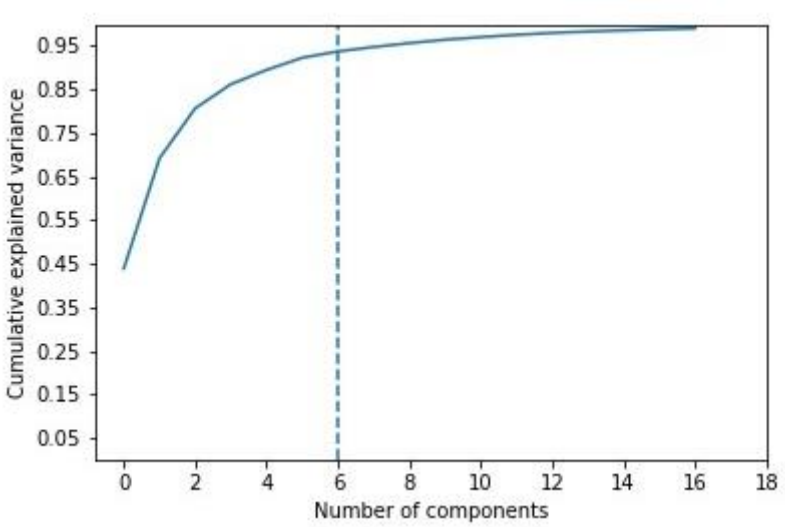


Figure 4.1 Cumulative variance by number of components

4.3 Rescaling the Data

To make data ready for modeling, we need to z scale the result again, as the outliers will be easily captured in the normalized data. The following is a statistical summary of the data.

Table 4.1 Statistical summary of the rescaled data

	PC1	PC2	PC3	PC4	PC5	PC6
Count	1046826	1046826	1046826	1046826	1046826	1046826
Mean	0	0	0	0	0	0
Std	1	1	1	1	1	1
Min	-0.26	-228.03	-199.01	-202.39	-243.89	-200.25
25%	-0.11	-0.15	-0.14	-0.03	-0.12	-0.11
50%	-0.02	-0.06	0	0.01	-0.01	-0.03
75%	0.03	0.03	0.32	0.07	0.05	0.14
Max	690.78	323.98	596.73	250.79	406.89	401.72

5 Fraud Model Algorithms

After reducing dimensions using PCA, we further apply two outlier detection methods and combine the results for a final anomaly score.

5.1 Model 1: Z Score Outliers

The fraud score is a function of these z scaled PCs that looks for extremes. Here we chose the power in the distance formula to be 2, which is the Euclidean distance. The Euclidean distance is the length of a line segment between the two points. It is a common and effective way to measure the closeness of things.

$$s_{1i} = \left(\sum_k |PCz_{ik}|^p \right)^{1/p}$$

5.2 Model 2: Autoencoder Error

An autoencoder was trained on all the data to reproduce the z scaled PC records. Keras was used to train a neural network model to learn efficient data representation in an unsupervised manner. The data records were first compressed from 6 to 3 dimensions and then expanded back to 6 dimensions. The model would have large reconstruction errors on the strange records and smaller errors on the normal records. The fraud score is a measure of difference between the original input record and the autoencoder output record. We choose the power to be 2, which is the Euclidean distance again.

$$s_{1i} = \left(\sum_k |PCz'_{ik} - PCz_{ik}|^p \right)^{1/p}$$

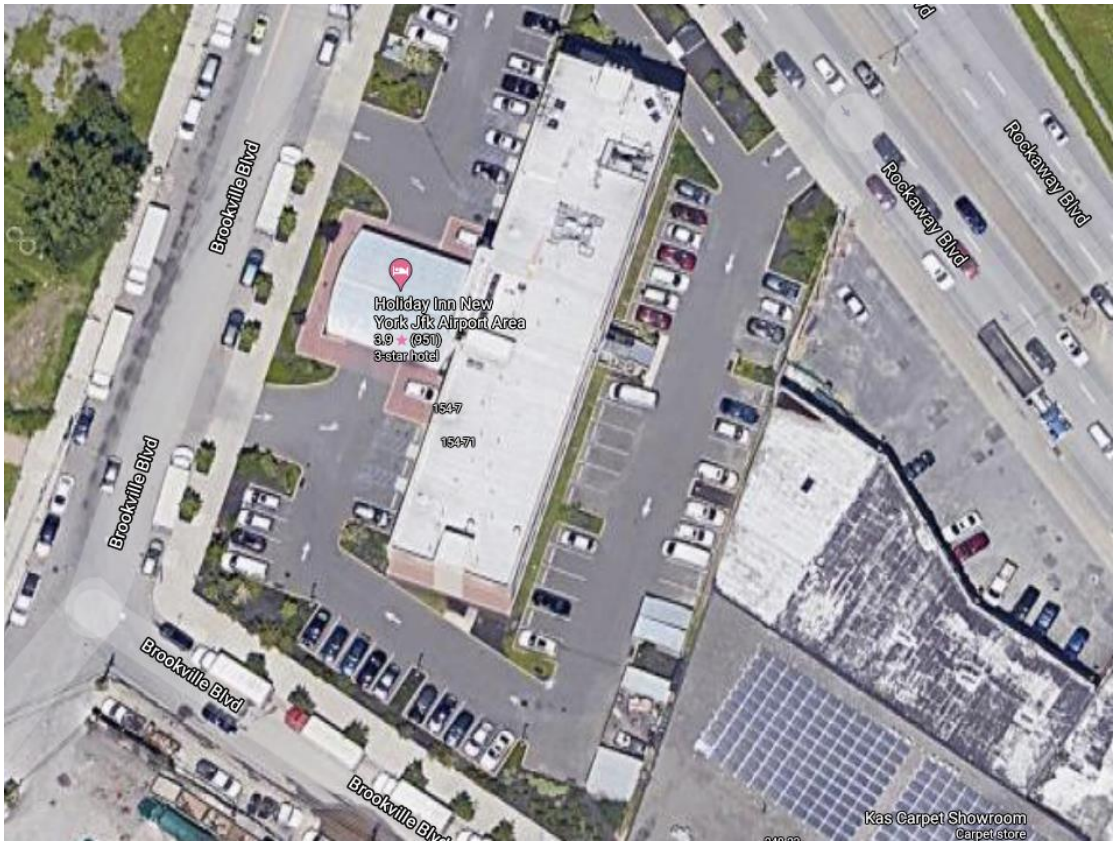
5.3 Final Score

After model 1 and model 2 were constructed, two fraud scores were ranked in descending order separately. The final fraud score is a weighted average ranking of these two score ranks. This final score, from high to low, reflects the likelihood that the property records are fraudulent.

6. Result

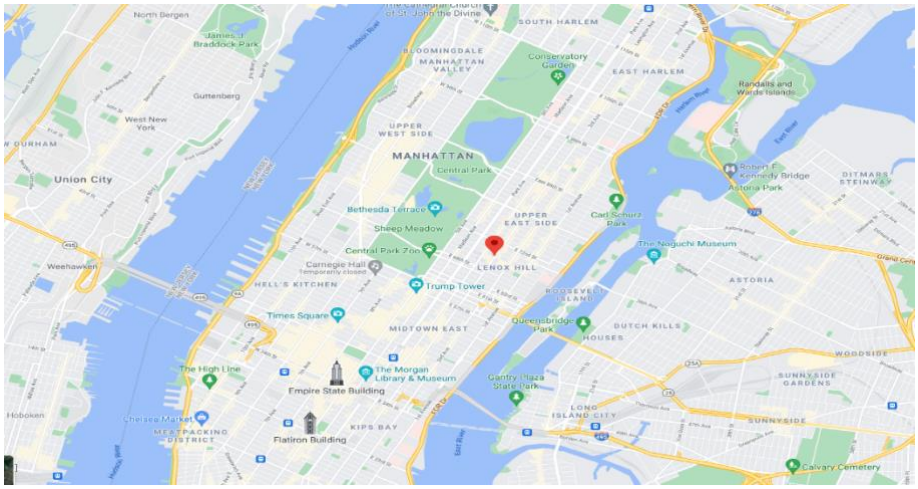
1. Record #917942 (ranked 1st)

This property has a street address of 154-68 Brookville Boulevard in Rosedale. It is a large hotel building with 7 stories. However, It is reported as a 3-stories building with building front and depth 62.08 and 82.3 feet respectively. It does not make sense for a large hotel property to have such a small value for building area measurements. What's more, the lot front is reported as 4910 feet which is not reasonable. These cause the value of the variables 'r2', 'r3', 'r4', 'r5', 'r6', 'r7', 'r8', and 'r9' to be extremely high. The property also has valuation of around $374 million, so it is strange to have a building size that small and this record can hence be viewed as anomalous.



2. Record #684704 (ranked 2nd)

This record ranked as 2nd highest likely to be fraudulent. This registered residential vacant land is located on 69th street in the center Manhattan area which is very valuable and scarce. However, this record has an assessed total value of 0, while having a 2ft x 2ft Lot and located right next to the landmarks like Empire State Building, Central Park, etc.



Also, the zip code of this record is suspiciously missing, while the street name and block are available, making this record highly anomalous.

3. Record #1065870 (ranked 3rd)

This property is located at Hylan Boulevard, which is a residential vacant land. The valuation of this property is $290.17 million. The land front and land depth are 2891 and 1488 feet respectively. While Several fields of this record are suspiciously missing, such as building front, building depth, stories, street number, and zip code. Since the size and valuation of this property is uncommonly large in its tax class group, using the group average to replace the building size makes the ratio of value variables 'r2_taxclass', 'r3_taxcalss', 'r5_taxclass', 'r6_taxclass', 'r8_taxclass', 'r9_taxclass' extremely high, making this record highly anomalous.

4. Record #1059883 (ranked 4th)

This property has few basic information, such as the owner, exempt land value, exempt total value, etc. The reported street name for the property is 'SAGONA COURT', with zip code '10309'. It is reported as an over 5-stories building with lot front and lot depth 5 and 5 feet respectively. However, it seems to be unreasonable that the building front and depth 62.08 and 82.3 feet respectively.

5. Record #151044 (ranked 5th)

The property is located at 1 East 161 Street, Bronx is the Yankee Stadium. Its valuation exceeds the average on the same scale. However, this anomaly is less likely due to fraudulence. As the home field for the famous baseball team-New York Yankee and soccer team-New York City FC, as well as the place holding big events, Yankee Stadium enjoys a higher-than-average valuation.

6. Record #2 (ranked 78th)

Despite our effort to remove as many government owned records as possible, record #2 still slipped through the data cleaning. Record #2 is Ellis Island, a federally-owned land that was once a major immigration inspection center of the U.S.

Although this property record is 100% not fraudulent, it is still interesting to discover why the metrics labeled this as one of the top 100 fraudulent records.

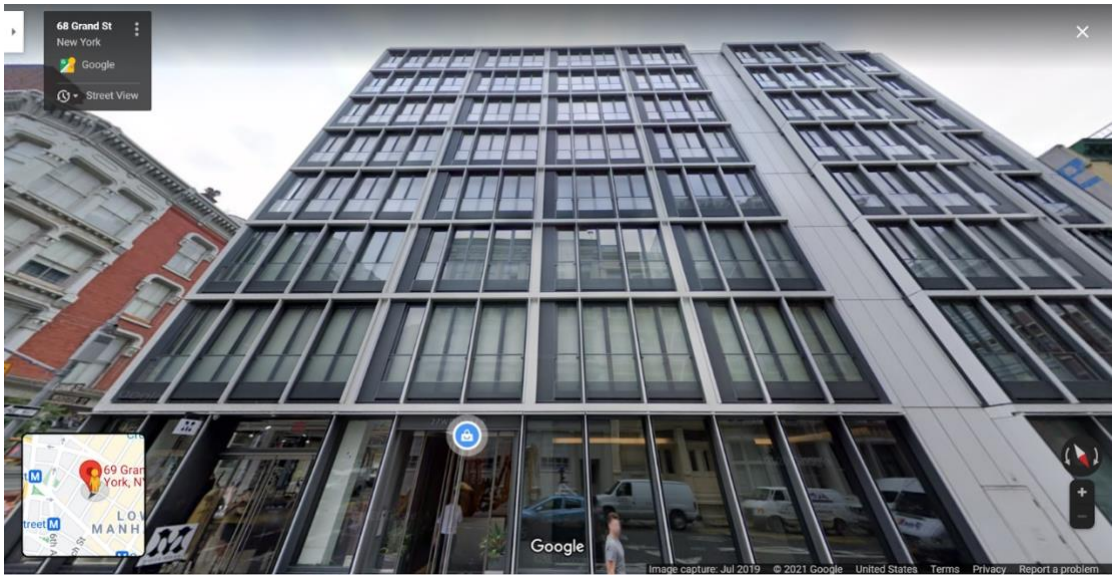
After observation, this is due to the recorded lot front and lot depth size being 27ft x 0 ft, while the assessed full value of this island/property is $193,800,000. This low lot size and staggeringly large value made variables like ‘r1’ through ‘r9’ extremely large, which ultimately made this property ranked very high on the fraudulent chart.

7. Record #95995 (ranked 21st)

The lot front and lot depth are reported as 197 and 378 feet respectively. The valuation of this property is $17.3 million. While the reported building front and building depth are only 15 and 20 feet respectively. The size for building area does not make sense for a property of that valuation, making this record anomalous.

8. Record #14979 (ranked 28th)

This property is owned by ENJAY ASSOCIATES, an Indian Home builder company. The lot front and lot depth are reported as 114 and 80 feet respectively. While the reported building front and building depth are only 8 and 6 feet respectively. Besides, the number of stories for the building is 1, however, the number of floors shown in the map is 8.

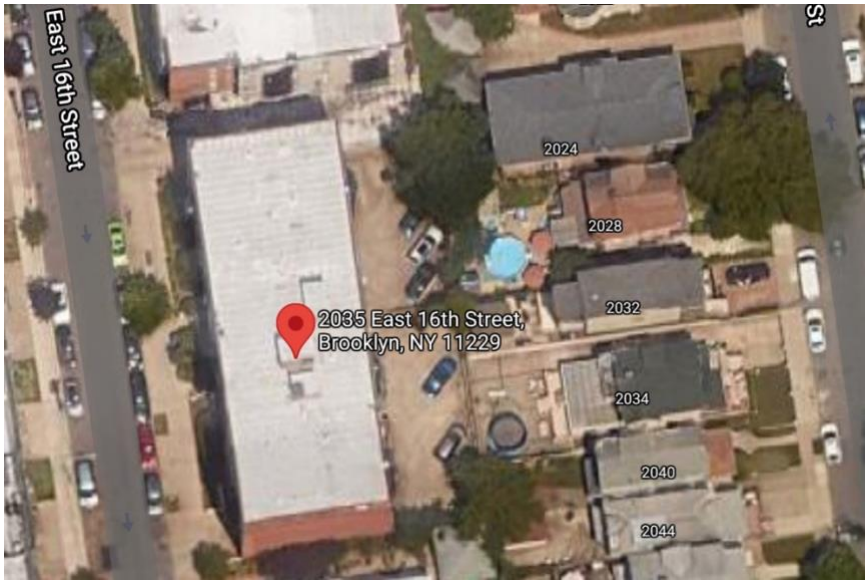


9. Record #116647 (ranked 7th)

Located at 1849 2 AVENUE, this property is owned by MF ASSOCIATES OF NEW, which has lot front and lot depth of 25 and 75. However, its building front and building depth are 70 and 456 feet respectively. The valuation of the property is $161 million and is greatly above the average. It is considered anomalous because its sizes for lot front and lot depth do not match with its valuation and building dimensions.

10. Record #116647 (ranked 71st)

Owned by AVE T REALTY LLC, this property is located at 2035 EAST 16 STREET. It is a six-story building with a lot front and a lot depth of 140 and 100 feet. Its building front and building depth are 14 and 5 feet which is unrealistic. This property has a valuation of $2.31 million, which does not make sense for such a property. Improper size and valuation make this property record anomalous in comparison to the others.



11. Record #39770 (ranked 6th)

Owned by Greenhorn Development, this property is located at 142 West 23 Street, only 0.8 miles away from the Empire State Building. Yet this building with 13 stories values at a suspiciously low value, only 1.02 million dollars. Besides, the reported building front and building depth of 8 feet seem unreasonable.

7. Summary and Conclusions

In the project, we average the results from the Z score outliers model and Autoencoder error to create a fraud score, through which we identify the most likely fraudulent records. We start the project with the Data Quality Report (DQR), a high-level data summary used to verify data. Having ensured that the data is correct, we clean the data, excluding properties own by governments and filling the missing values. Based on the knowledge of property value, we create 45 new variables from three fields- FULLVAl, AVTOT, AVLAND. After that, we normalize the data, perform PCA to reduce data to six dimensions, and rescale them to be ready for modeling. With processed data, we build two fraud scores using Z score outliers and autoencoder error respectively. Averaging the two results for each property, we finalize the fraud scores and rank them in descending order. To cross-validate our result, we investigate 10 properties sampled from the top 100 from the result. Most of their records are highly suspicious.

Finally, we suggest working on the following three directions to further improve this project:

1. collect more data or fill the missing value more finely, as many fields contain missing values.
2. determine the threshold of the fraud score according to further investigation; the fraud score for this project reflects only the possibility
3. evaluate the performance of the model by further investigation; unlike a supervised algorithm, an unsupervised algorithm does not have evaluating metrics.

Appendix A: Data Quality Report

File Description

The "Property Valuation and Assessment Data" is collected and entered into the system by various City employee, such as Property Assessors, Property Exemption specialists, ACRIS reporting, Department of Building reporting, etc. It stores 1,070,994 records of real estate assessment property data. And it covers sixteen numeric fields, including lot frontage in feet, lot depth in feet, total market value of the land, assess land value, exempt land value, etc., and sixteen categorical fields, including record, owner's name, building class, tax class, and postal zip code of the property, etc. The dataset came from NYC OpenData and was shared by Professor Stephen Coggeshall in March 2021.

Table 8.0: File Description

Dataset Name	Property Valuation and Assessment Data
Dataset Purpose	Calculate property tax, grant eligible properties exemptions and/or abatements
Data Source	NYC Open Data
Time Period	17-Nov-10
# of Fields	32
# of Records	1,070,994

Summary Statistics Table

Numeric Fields

Ten of sixteen fields are fully populated. Key statistics of these fields are summarized as follows.

Table 8.1: Summary Statistics of Numeric Fields

Field Name	# records	# records with value zero	# unique values	% Populated	Mean	Standard Deviation	Minimum Value	Maximum Value
LTFRONT	1070994	169108	1297	100	36.64	74.03	0	9999
LTDEPTH	1070994	170128	1370	100	88.86	76.40	0	9999
STORIES	1014730	0	112	94.75	5.01	8.37	1	119
FULLVAL	1070994	13007	109324	100	874264.51	11582430.99	0	6150000000
AVLAND	1070994	13009	70921	100	85067.92	4057260.06	0	2668500000
AVTOT	1070994	13007	112914	100	227238.17	6877529.31	0	4668308947
EXLAND	1070994	491699	33419	100	36423.89	3981575.79	0	2668500000
EXTOT	1070994	432572	64255	100	91186.98	6508402.82	0	4668308947
EXCD1	638488	0	130	59.62	1602.01	1384.23	1010	7170
BLDFRONT	1070994	228815	612	100	23.04	35.58	0	7575
BLDDEPTH	1070994	228853	621	100	39.92	42.71	0	9393

AVLAND2	282726	0	58592	26.40	246235.72	6178962.56	3	2371005000
AVTOT2	282732	0	111361	26.40	713911.44	11652528.95	3	4501180002
EXLAND2	87449	0	22196	8.17	351235.68	10802212.67	1	2371005000
EXTOT2	130828	0	48349	12.22	656768.28	16072510.17	7	4501180002
EXCD2	92948	0	61	8.68	1364.04	1094.71	1011	7160

Categorical Variables

Ten of sixteen fields are fully populated. Key statistics of these fields are summarized as follows.

Table 8.2: Summary Statistics of Categorical Fields

Field Name	# records	# unique values	% Populated	Most Common Field Value
RECORD	1070994	1070994	100	N/A*
BBLE	1070994	1070994	100	N/A*
B	1070994	5	100	4
BLOCK	1070994	13984	100	3944
LOT	1070994	6366	100	1
EASEMENT	4636	13	0.433	E
OWNER	1039249	863348	97.036	PARKCHESTER PRESERVAT
BLDGCL	1070994	200	100	R4
TAXCLASS	1070994	11	100	1
EXT	354305	4	33.082	G
STADDR	1070318	839281	99.937	501 SURF AVENUE
ZIP	1041104	197	97.209	10314
EXMPTCL	15579	15	1.455	X1
PERIOD	1070994	1	100	FINAL
YEAR	1070994	1	100	2010/11
VALTYPE	1070994	1	100	AC-TR

* All values in the RECORD and BBLE field are unique, thus no such a most common field value.

Field Description and Distribution

FIELD 1: RECORD

DESCRIPTION	Unique identifier of each record
TYPE	Categorical
UNIQUE VALUES	1070994

FIELD 2: BBLE

DESCRIPTION	Concatenation of borough code, block code, lot code and easement code
TYPE	Categorical
UNIQUE VALUES	1070994

FIELD 3: B

DESCRIPTION	Borough code
TYPE	Categorical
UNIQUE VALUES	5

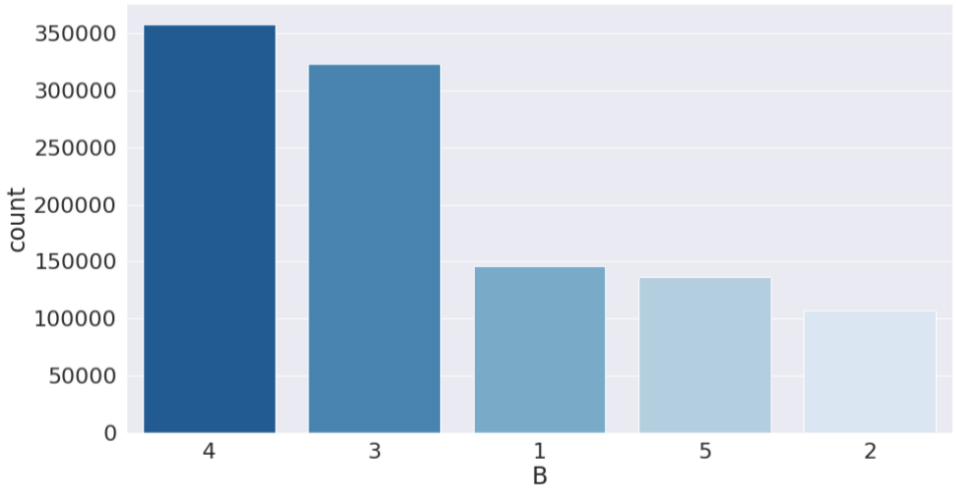


Figure 8.1: Frequency Distribution of the *B* Field

FIELD 4: BLOCK

DESCRIPTION	Valid block ranges by borough
TYPE	Categorical
UNIQUE VALUES	13984

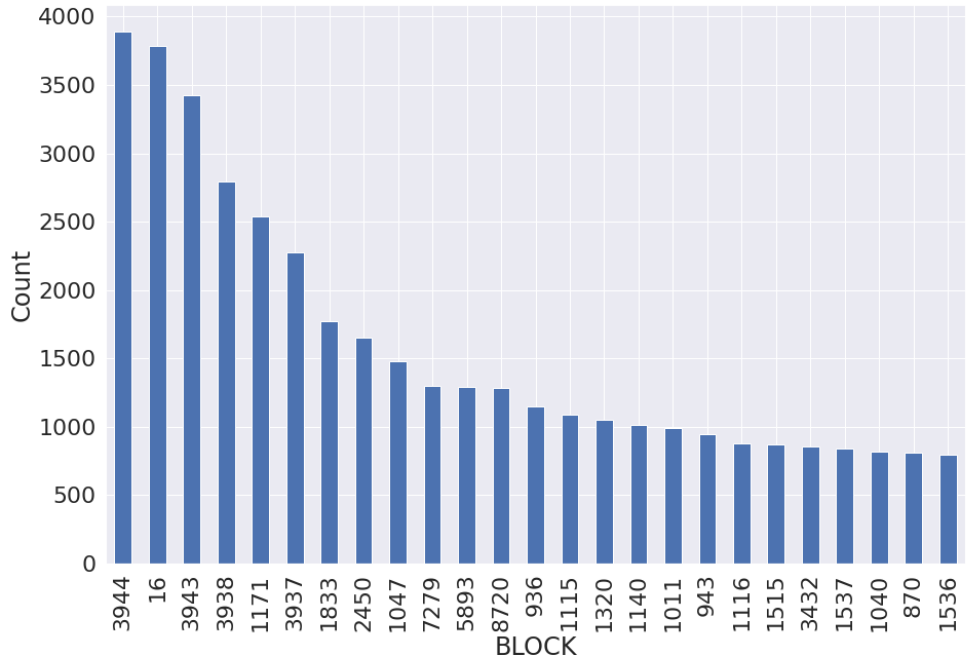


Figure 8.2: Frequency Distribution of the *BLOCK* Field
(Top 25 Most Common Values)

FIELD 5: LOT

DESCRIPTION	Unique number within borough or block
TYPE	Categorical
MOST COMMON FIELD VALUE	"1" occurred the most for 24367 times

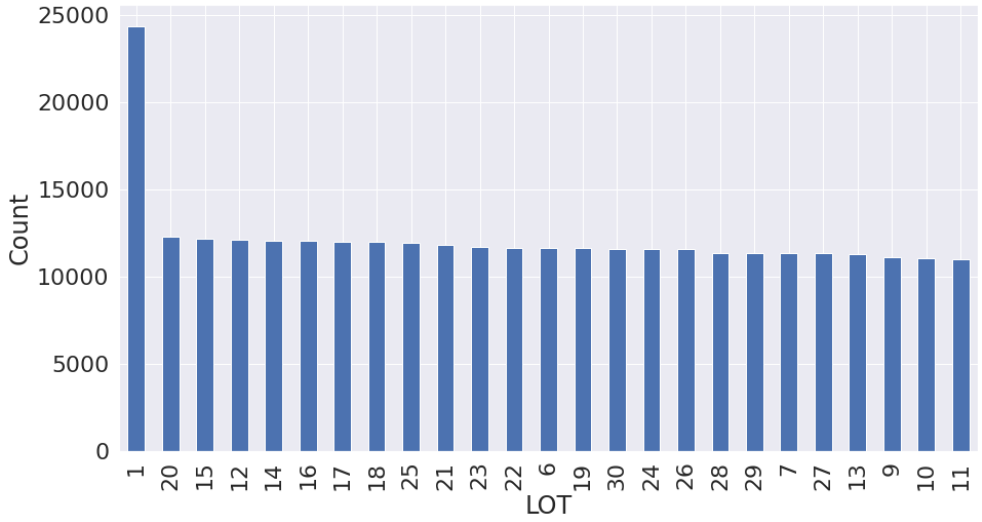


Figure 8.3: Frequency Distribution of the *LOT* Field
(Top 25 Most Common Values)

FIELD 6: EASEMENT

DESCRIPTION	A field that is used to describe easement
TYPE	Categorical
UNIQUE VALUES	13

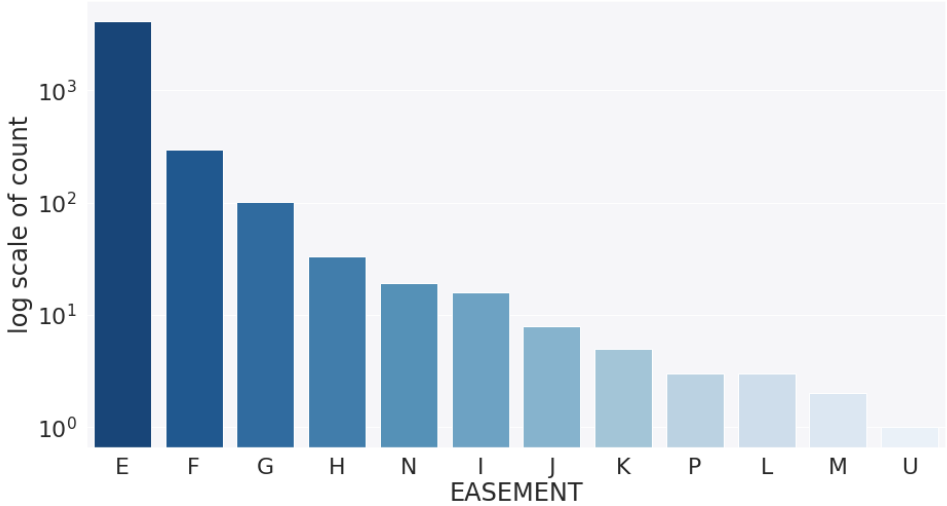


Figure 8.4: Frequency Distribution of the *EASEMENT* Field

FIELD 7: OWNER

DESCRIPTION	Owner's name
TYPE	Categorical
UNIQUE VALUES	863348

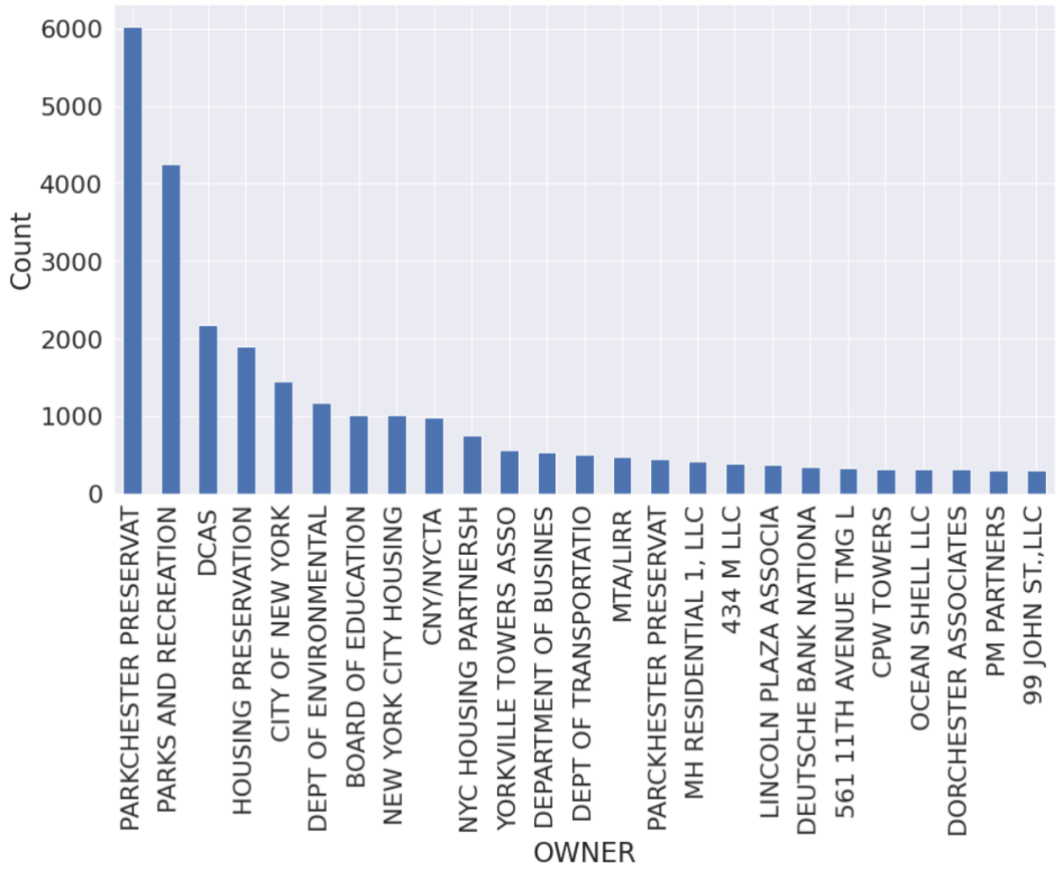


Figure 8.5: Frequency Distribution of the *OWNER* Field
(Top 25 Most Common Values)

FIELD 8: BLDGCL

DESCRIPTION	Building class
TYPE	Categorical
MOST COMMON FIELD VALUE	"R4" occurred the most for 139879 times

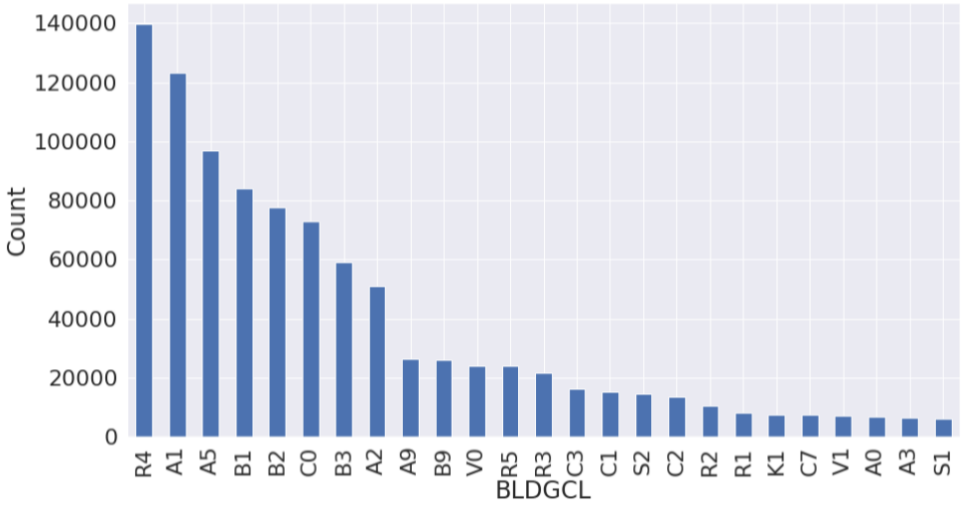


Figure 8.6: Frequency Distribution of the *BLDGCL* Field
(Top 25 Most Common Values)

FIELD 9: TAXCLASS

DESCRIPTION	Current property tax class code (NYS Classification)
TYPE	Categorical
UNIQUE VALUES	11

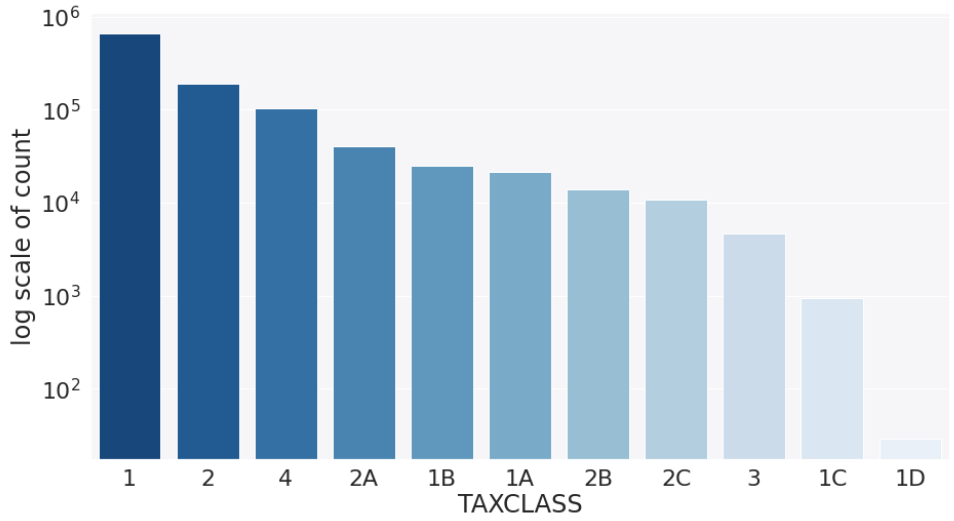


Figure 8.6: Frequency Distribution of the *TAXCLASS* Field

FIELD 10: LTFRONT

DESCRIPTION	Lot Frontage in feet
TYPE	Numeric
UNIQUE VALUES	1297

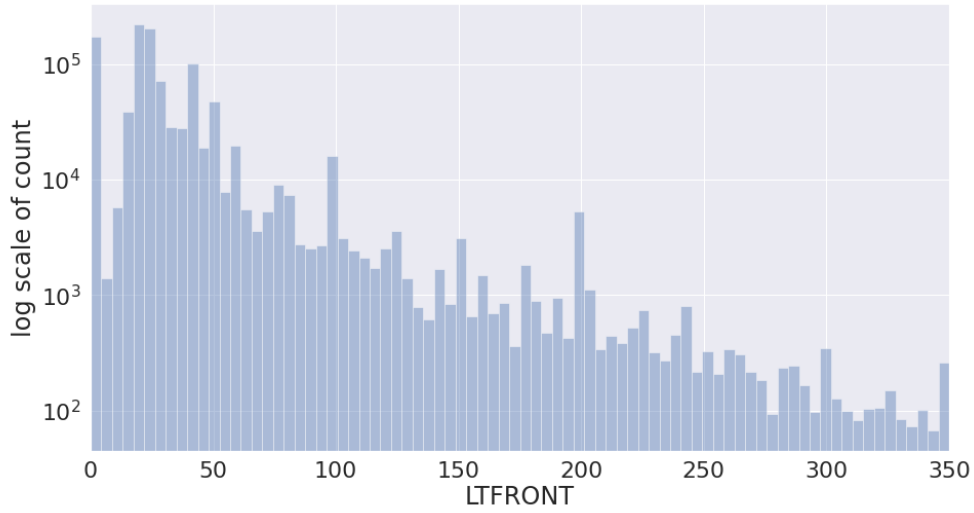


Figure 8.7: Frequency Distribution of the *LTFRONT* Field
(Exclude outliers > 350; Data in histogram is 99.42% populated)

FIELD 11: LTDEPTH

DESCRIPTION	Lot Depth in feet
TYPE	Numeric
UNIQUE VALUES	1370

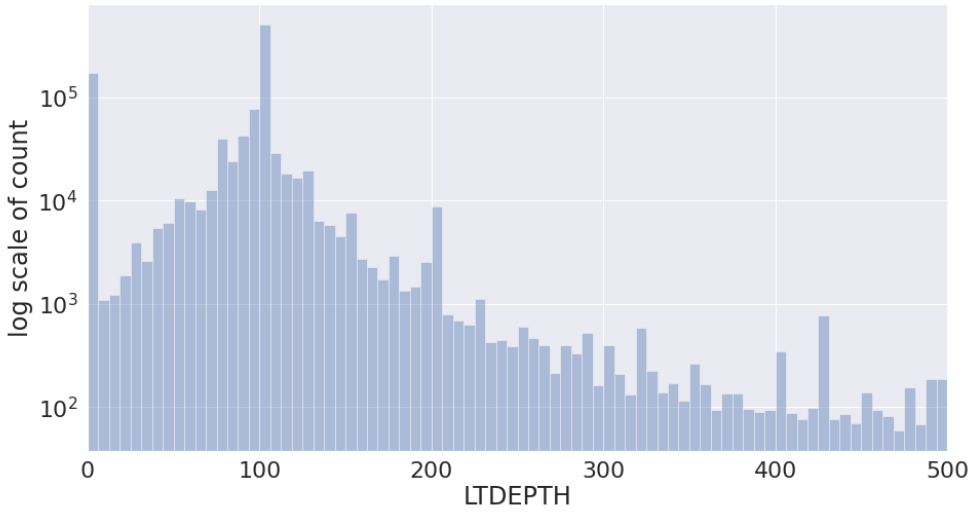


Figure 8.8: Frequency Distribution of the *LTDEPTH* Field
(Exclude outliers > 500; Data in histogram is 99.68% populated)

FIELD 12: EXT

DESCRIPTION	Extension
TYPE	Categorical
UNIQUE VALUES	4

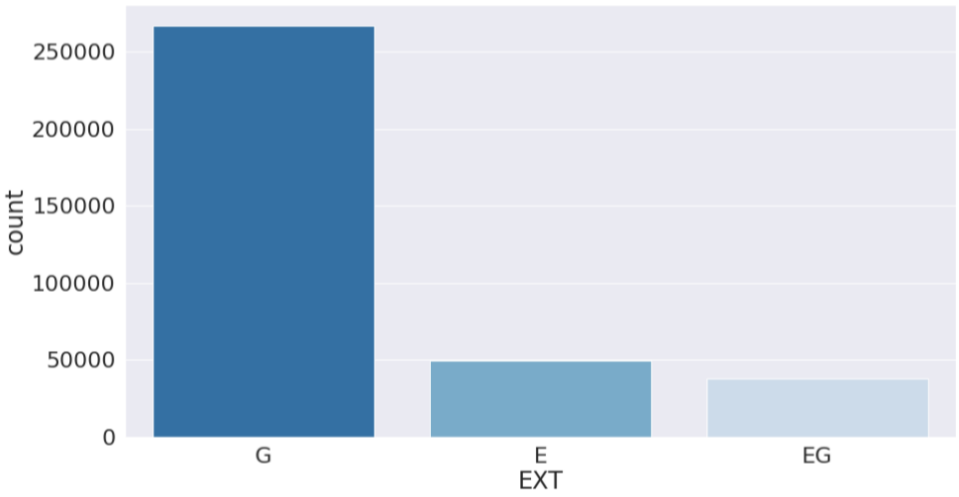


Figure 8.9: Frequency Distribution of the *EXT* Field

FIELD 13: STORIES

DESCRIPTION	The number of stories(floors) for the building (number of floors)
TYPE	Numeric
UNIQUE VALUES	112

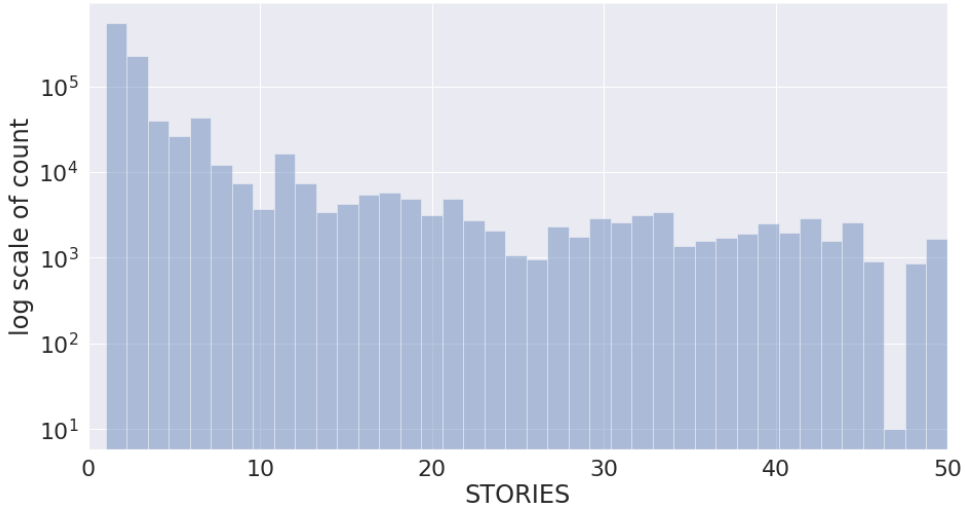


Figure 8.10: Frequency Distribution of the *STORIES* Field
(Exclude outliers > 50; Data in histogram is 99.50% populated among 1014730 records)

FIELD 14: FULLVAL

DESCRIPTION	Total market value of the land
TYPE	Numeric
UNIQUE VALUES	109324

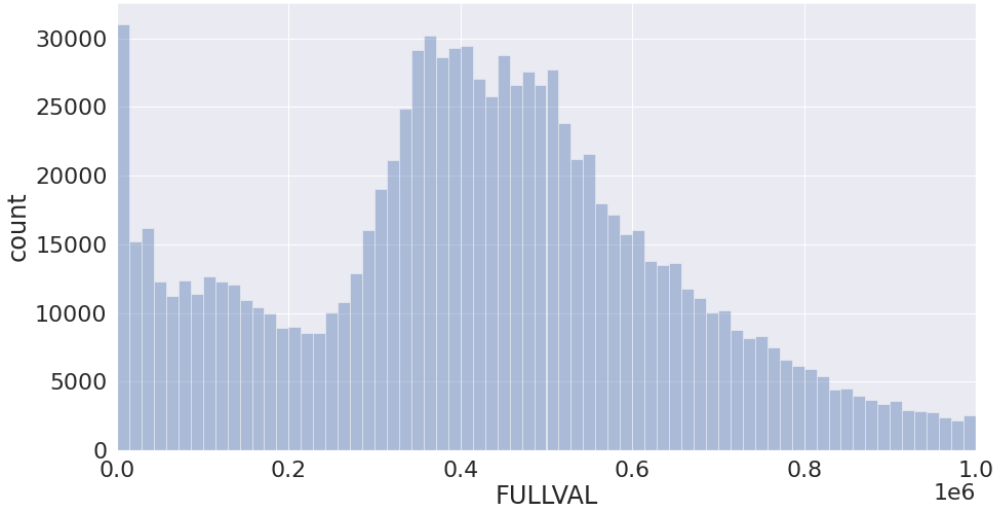


Figure 8.11: Frequency Distribution of the *FULLVAL* Field
(Exclude outliers > 1000000; Data in histogram is 91.35% populated)

FIELD 15: AVLAND

DESCRIPTION	Assess land value
TYPE	Numeric
UNIQUE VALUES	70921

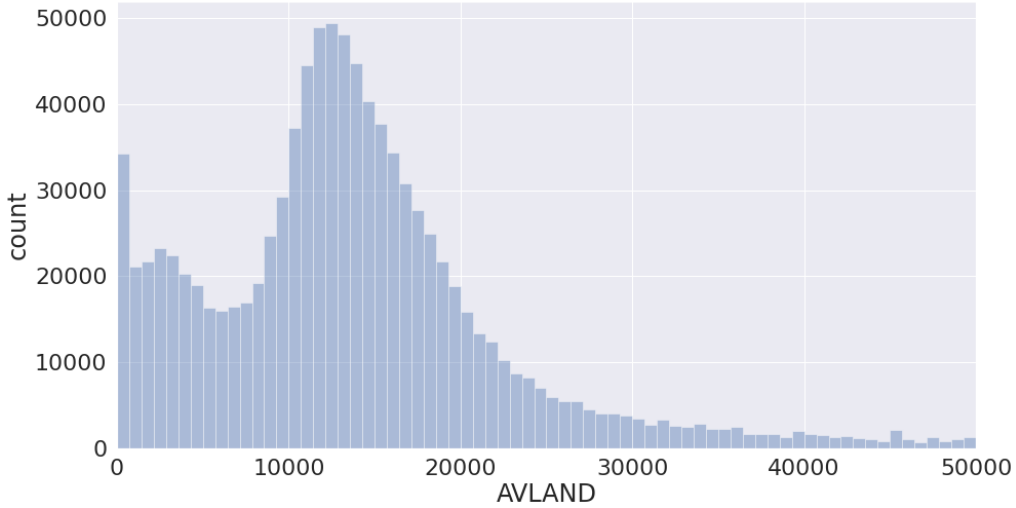


Figure 8.12: Frequency Distribution of the *AVLAND* Field
(Exclude outliers > 50000; Data in histogram is 90.53% populated)

FIELD 16: AVTOT

DESCRIPTION	Assessed total value
TYPE	Numeric
UNIQUE VALUES	112914

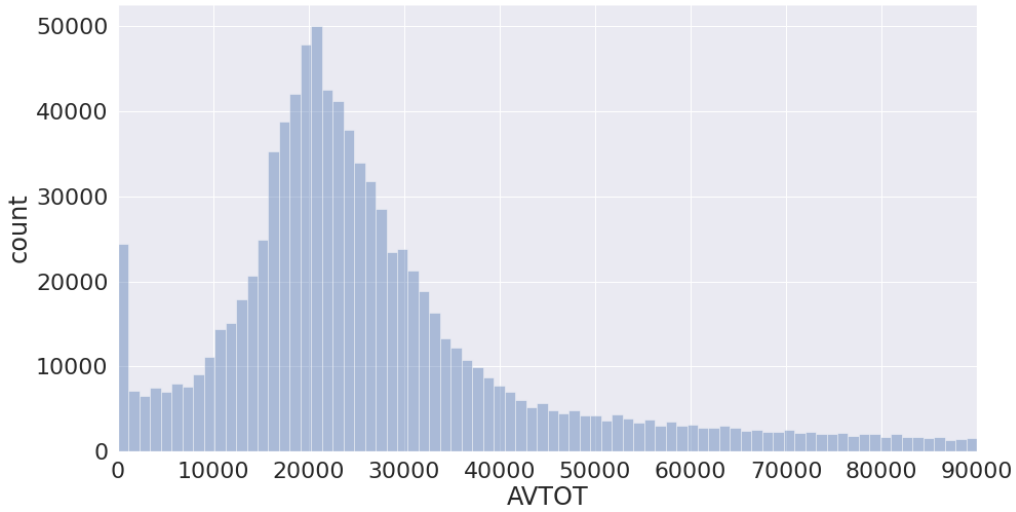


Figure 8.13: Frequency Distribution of the *AVTOT* Field
(Exclude outliers > 90000; Data in histogram is 84.99% populated)

FIELD 17: EXLAND

DESCRIPTION	Exempt land value
TYPE	Numeric
UNIQUE VALUES	33419

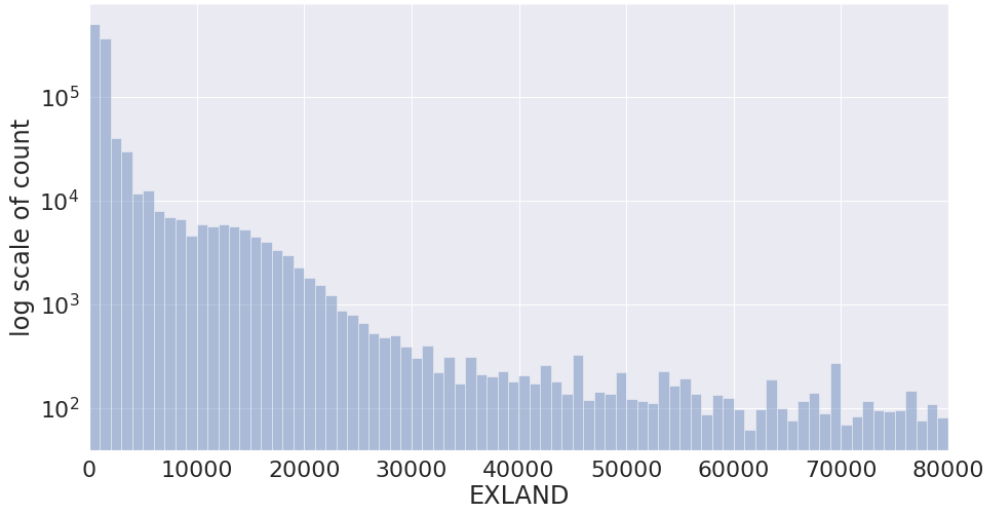


Figure 8.14: Frequency Distribution of the *EXLAND* Field
(Exclude outliers > 80000; Data in histogram is 98.39% populated)

FIELD 18: EXTOT

DESCRIPTION	Exempt total value
TYPE	Numeric
UNIQUE VALUES	64255

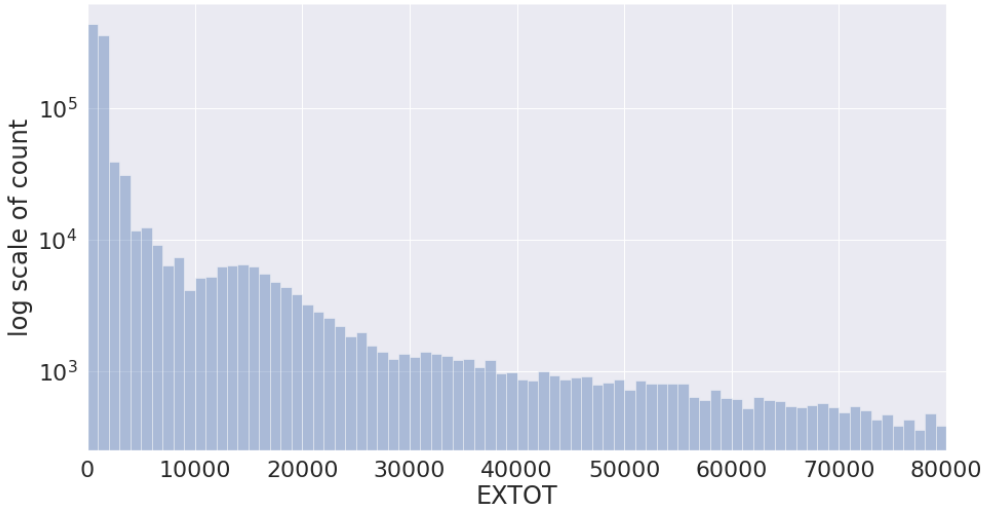


Figure 8.15: Frequency Distribution of the *EXTOT* Field
(Exclude outliers > 80000; Data in histogram is 95.87% populated)

FIELD 19: EXCD1

TYPE	Categorical
UNIQUE VALUES	130

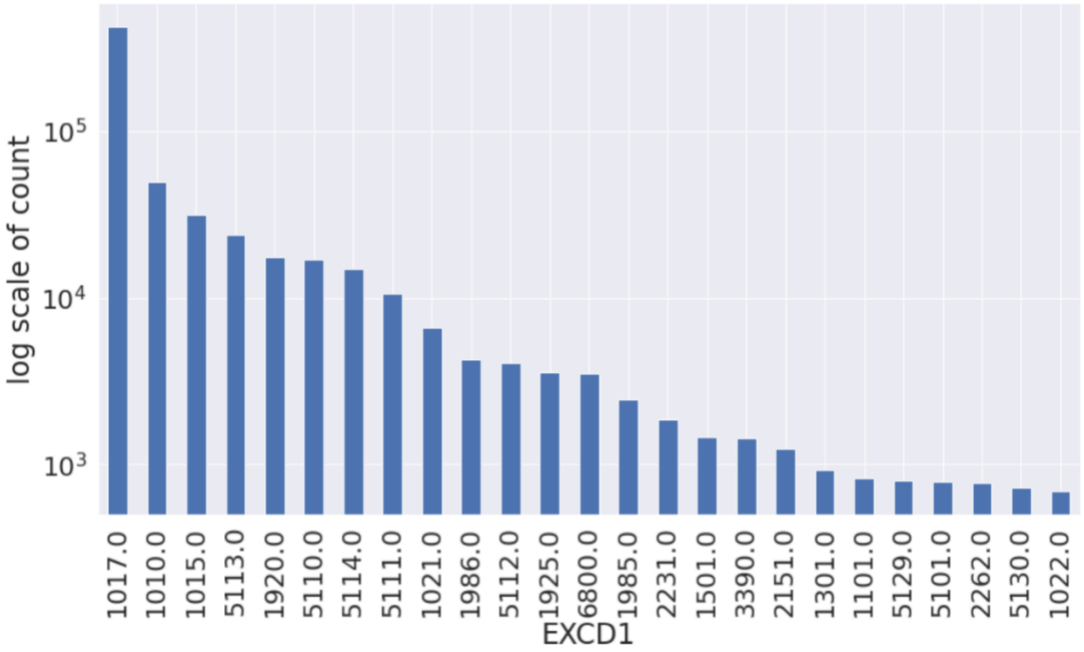


Figure 8.16: Frequency Distribution of the *EXCD1* Field
(Top 25 Most Common Values)

FIELD 20: STADDR

DESCRIPTION	Street name for the property
TYPE	Categorical
UNIQUE VALUES	839281

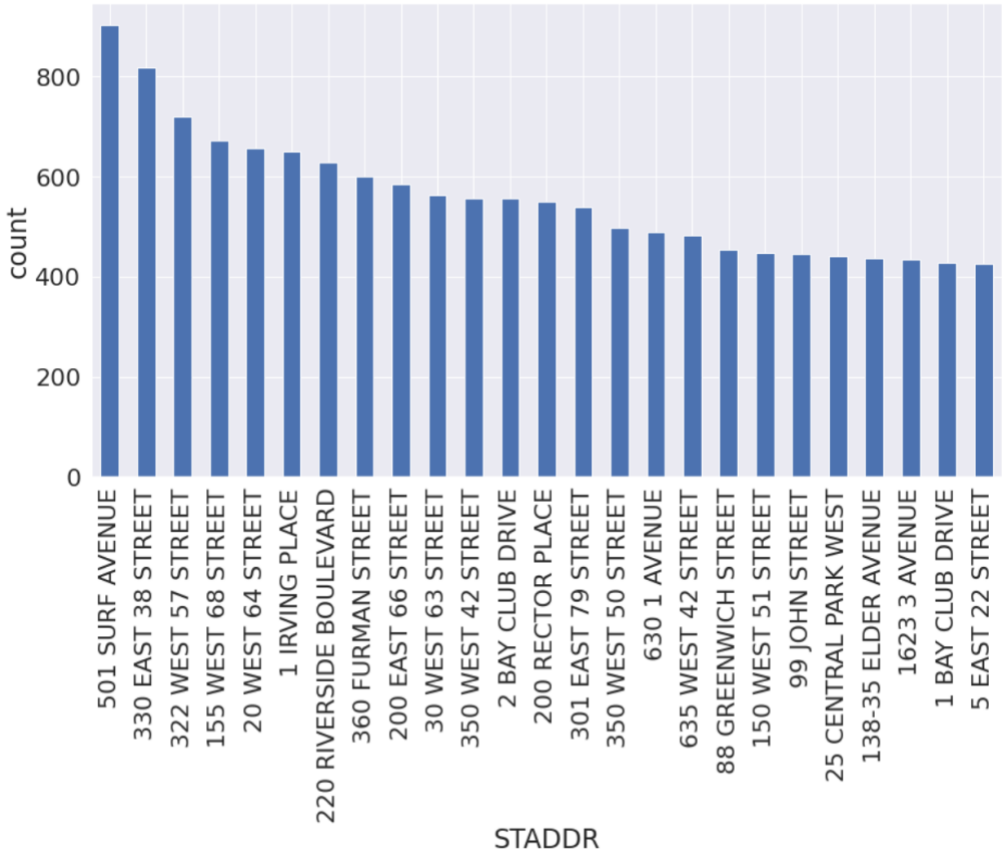


Figure 8.17: Frequency Distribution of the *STADDR* Field
(Top 25 Most Common Values)

FIELD 21: ZIP

DESCRIPTION	Postal zip code of the property
TYPE	Categorical
UNIQUE VALUES	197

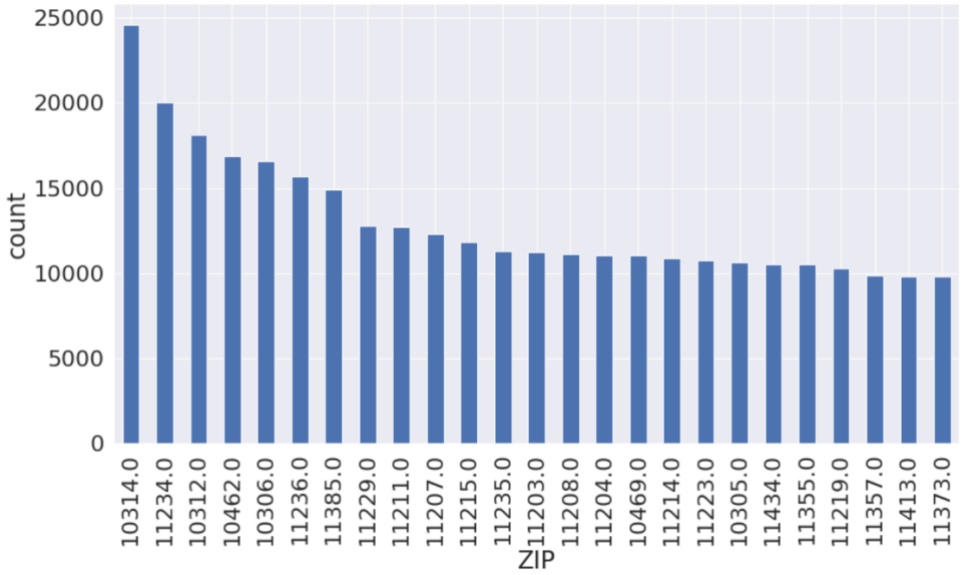


Figure 8.18: Frequency Distribution of the *ZIP* Field
(Top 25 Most Common Values)

FIELD 22: EXMPTCL

DESCRIPTION	Exempt Class used for fully exempt properties only
TYPE	Categorical
UNIQUE VALUES	15

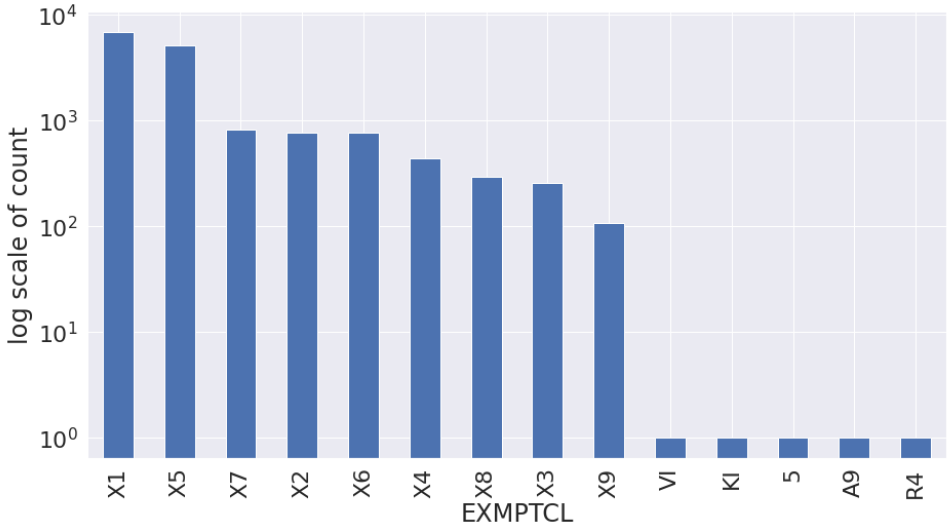


Figure 8.19: Frequency Distribution of the *EXMPTCL* Field
(Top 25 Most Common Values)

FIELD 23: BLDFRONT

DESCRIPTION	Building frontage in feet
TYPE	Numeric
UNIQUE VALUES	612

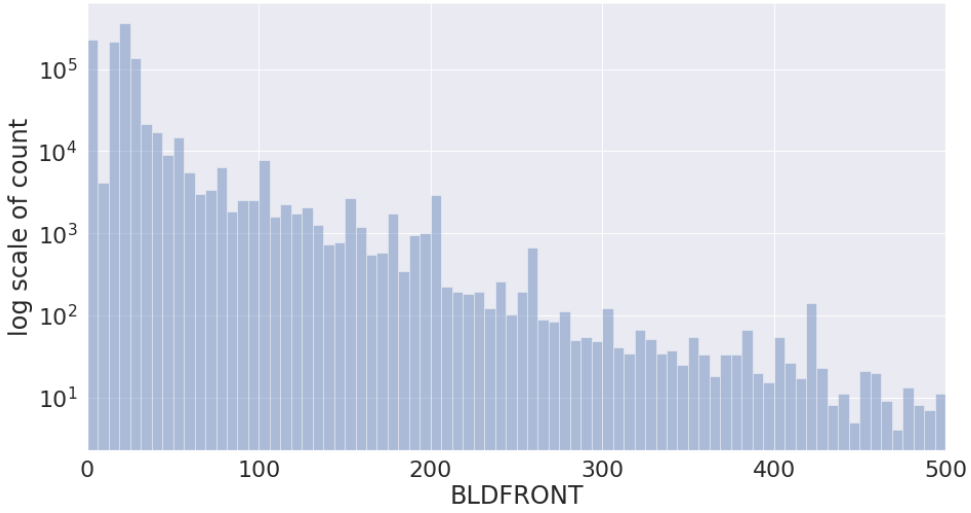


Figure 8.20: Frequency Distribution of the *BLDFRONT* Field
(Exclude outliers > 500; Data in histogram is 99.98%)

FIELD 24: BLDDEPTH

DESCRIPTION	Lot depth in feet
TYPE	Numeric
UNIQUE VALUES	621

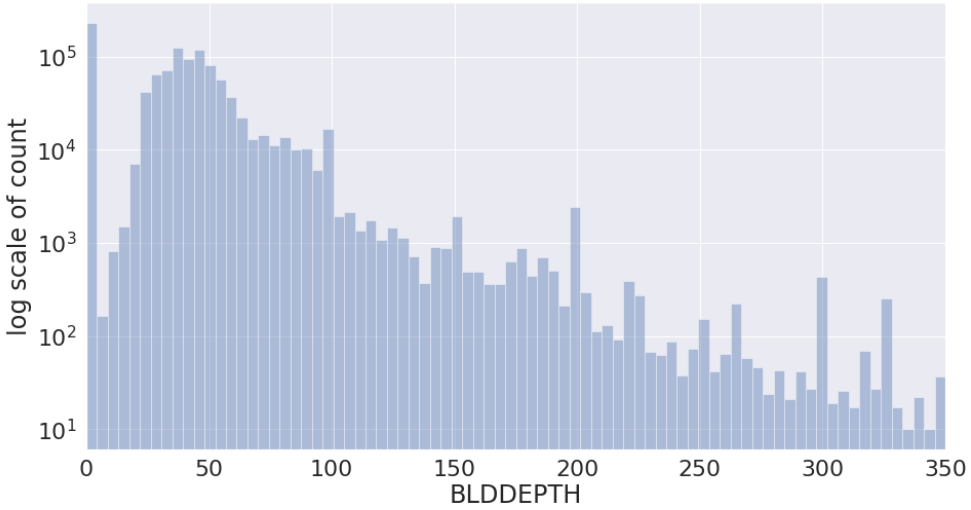


Figure 8.21: Frequency Distribution of the *BLDDEPTH* Field
(Exclude outliers > 350; Data in histogram is 99.87%)

FIELD 25: AVLAND2

DESCRIPTION	New market value of land
TYPE	Numeric
UNIQUE VALUES	58592

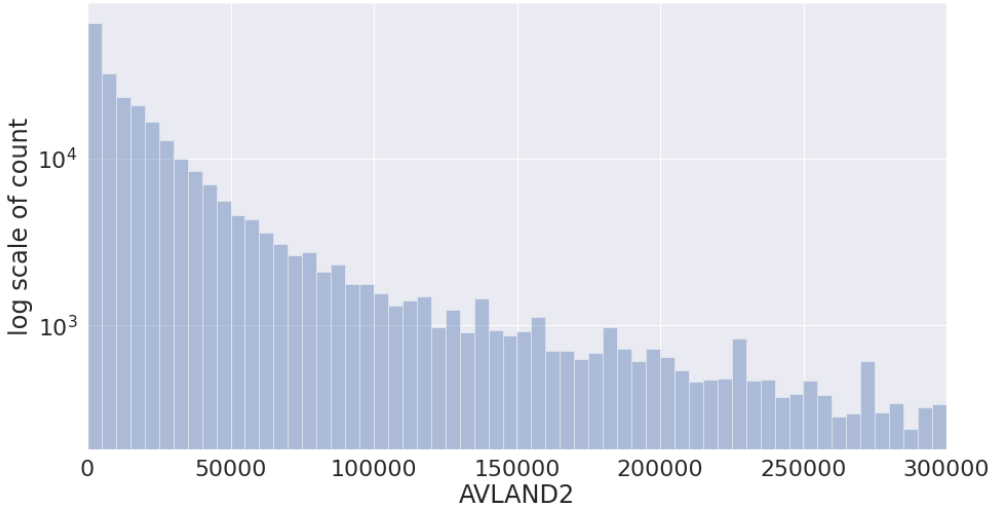


Figure 8.22: Frequency Distribution of the *AVLAND2* Field
(Exclude outliers > 300000; Data in histogram is 91.48% among 282,726 records)

FIELD 26: AVTOT2

DESCRIPTION	New total market value
TYPE	Numeric
UNIQUE VALUES	111361

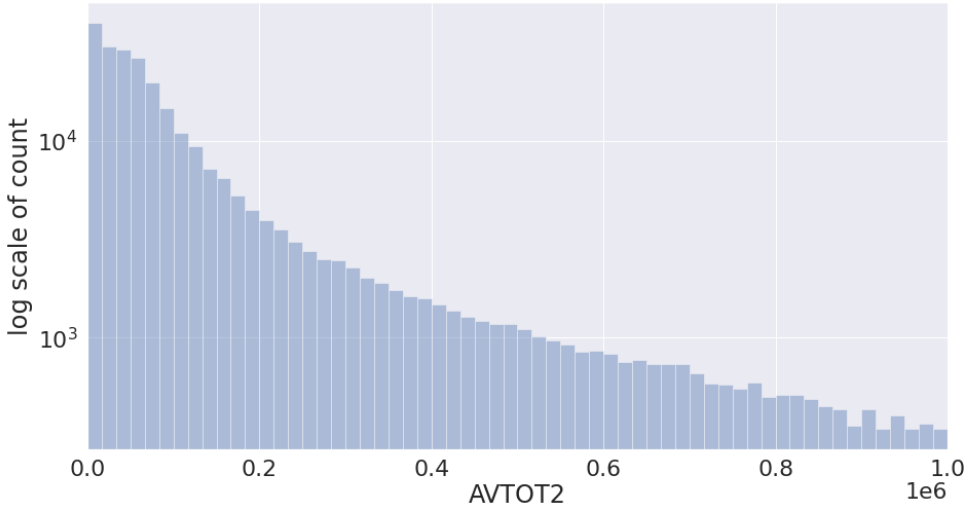


Figure 8.23: Frequency Distribution of the *AVTOT2* Field
(Exclude outliers > 1000000; Data in histogram is 91.62% among 282,732 records)

FIELD 27: EXLAND2

DESCRIPTION	New exempt land value
TYPE	Numeric
UNIQUE VALUES	22196

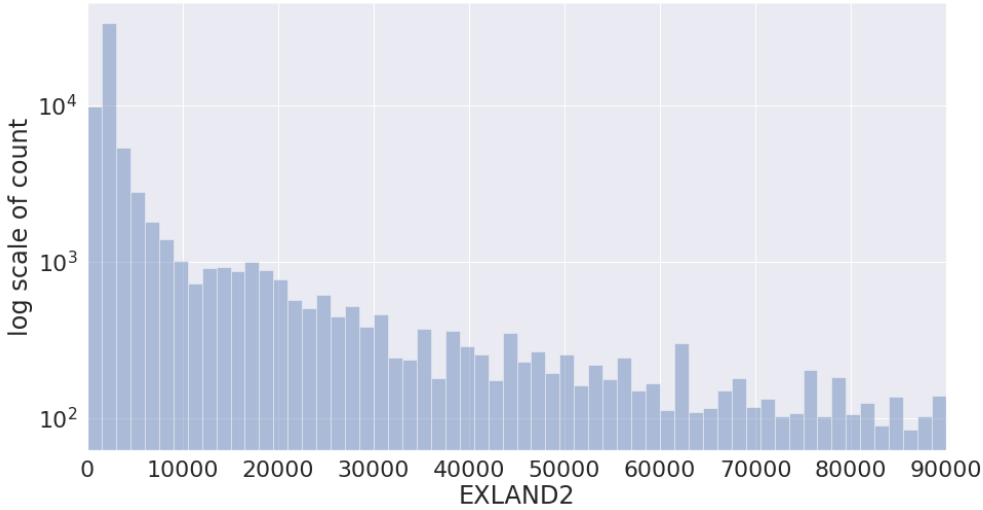


Figure 8.24: Frequency Distribution of the *EXLAND2* Field
(Exclude outliers > 90000; Data in histogram is 83.21% among 87,449 records)

FIELD 27: EXTOT2

DESCRIPTION	New exempt total value
TYPE	Numeric
UNIQUE VALUES	48349

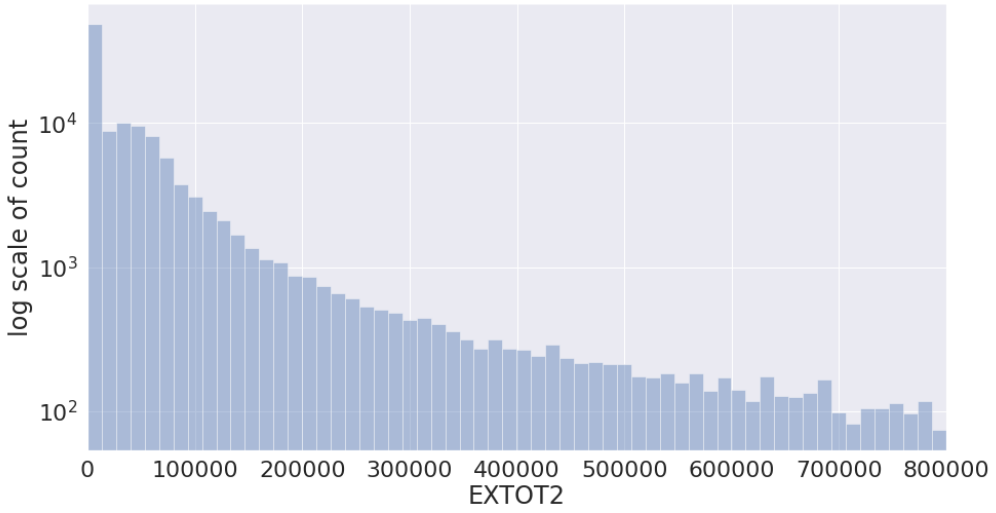


Figure 8.25: Frequency Distribution of the *EXTOT2* Field
(Exclude outliers > 800000; Data in histogram is 92.35% among 130,828 records)

FIELD 28: EXCD2

TYPE	Categorical
UNIQUE VALUES	61

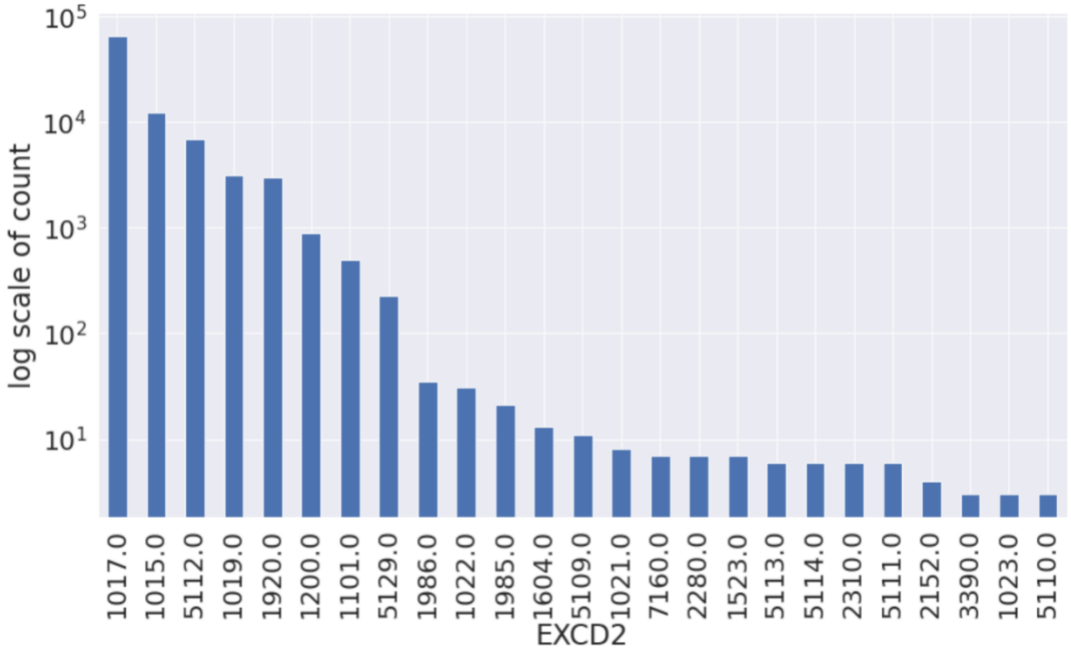


Figure 8.26: Frequency Distribution of the *EXCD2* Field
(Top 25 Most Common Values)

FIELD 29: PERIOD

DESCRIPTION	Change of period of file (all records have the same value)
TYPE	Categorical
UNIQUE VALUES	1

FIELD 29: YEAR

DESCRIPTION	Year when the data was most recently updated (all records have the same value)
TYPE	Categorical
UNIQUE VALUES	1

FIELD 29: VALTYPE

DESCRIPTION	The parcel's values are reflected in another lot (all records have the same value)
TYPE	Categorical
UNIQUE VALUES	1