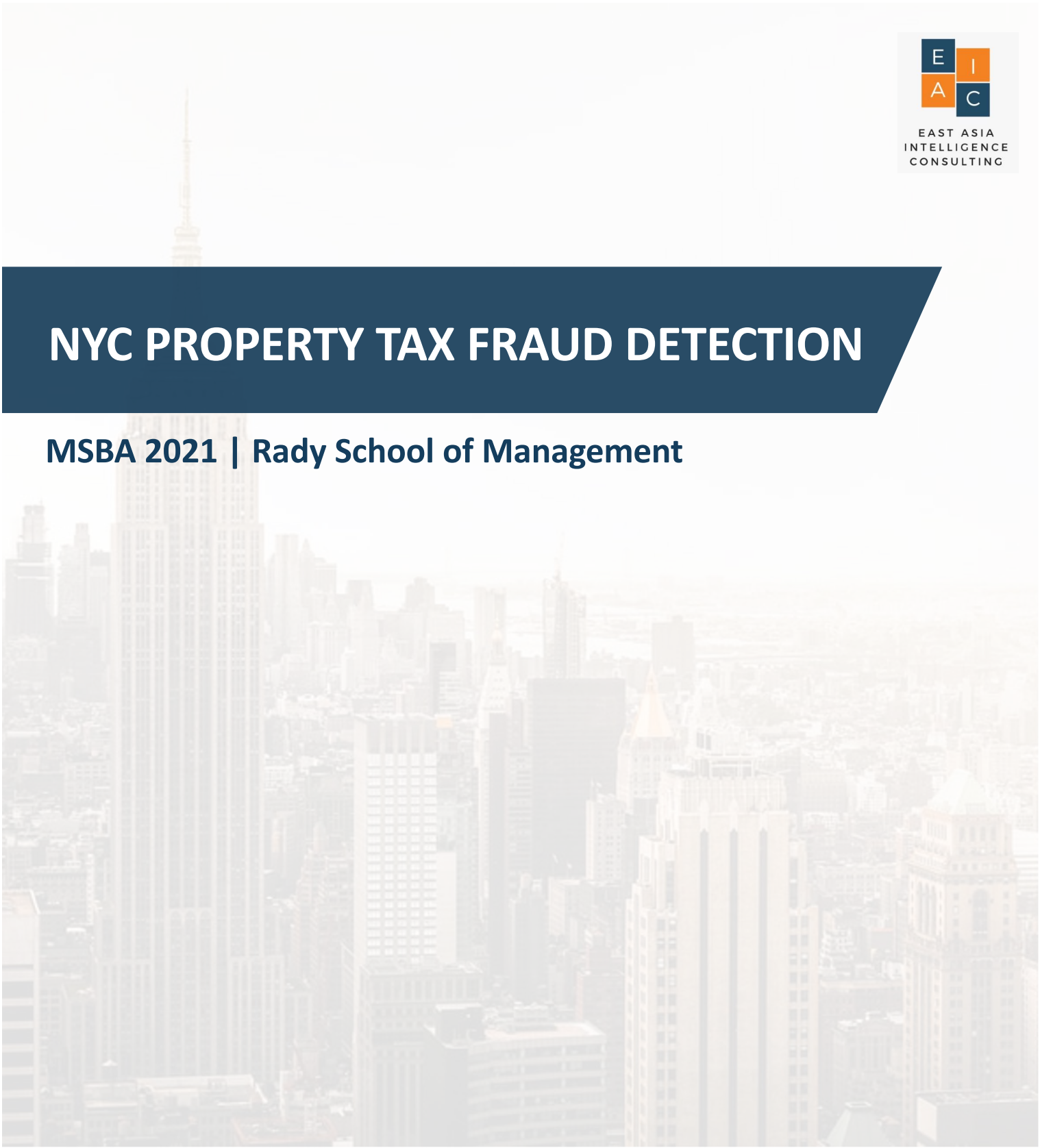
EAST ASIA INTELLIGENCE CONSULTING

NYC PROPERTY TAX FRAUD DETECTION

MSBA 2021 | Rady School of Management



EA Intelligence Consulting

Chia-Hsien Ho, Li Du, Seyoung Ahn
Wil Son Chuah, Yuxiang Zhou, Yuying Liu

Table of Content

Executive Summary

A property tax assessment is an estimation of the market value of real estate. It is conducted by local governments in order to calculate property tax bills and to determine properties eligible for tax exemptions and/or abatements. Data about properties, such as lot size, building size and property use, is collected periodically and entered into the system by various government employees. Since property tax is an important source of income for governments, Incorrect data, regardless of if it's intentional or accidental, can cost them millions of dollars in lost revenues.

The purpose of the project is to identify property tax fraud, which can be defined as intentional manipulation of data in order to receive financial gain. It focuses on building unsupervised fraud algorithms that will detect unusual cases in the property assessment data.

This report presents the findings after conducting data analysis and assessment of unusual properties detected by unsupervised fraud algorithms. The following describes the completion process:

1. Data Preparation - conduct EDA, remove irrelevant records, and impute missing values
2. Feature engineering - create 45 new variables that help detect anomalies
3. Feature selection - reduce dimensionality using principal component analysis (PCA)
4. Fraud detection algorithm - build 2 models using z-score outliers and autoencoder error
5. Anomaly assessment - understand reasons for records with highest fraud scores

Among the top 100 anomalous records that were identified by the fraud detection algorithm, 10 records were selected and were further assessed to discover possible reasons for the unusualness. Our team identified three main reasons: (1) missing fields were imputed by the grouped averages, (2) some fields contain incorrect data, and (3) the dataset is outdated and multiple changes were made over the period.

Limitations exist for unsupervised fraud algorithms as they were developed without knowledge of actual fraudulent records. In addition, domain expertise is required to determine whether or not an anomaly is a fraud. The report also suggests possible ways for improvement of the model.

1. Data Description

1.1 File Description

The “NY_property_data.csv” dataset was prepared by the New York City government in order to calculate property tax and grant eligible properties exemptions and/or abatements. It contains 1,070,994 records of properties and covers 32 fields, including address, owner, property characteristics, etc. It was most recently updated in November 2010.

Dataset Name	Property Valuation and Assessment Data
Dataset Purpose	To calculate property tax and grant eligible properties exemptions and/or abatements
Data Source	NYC Open Data
Time Period	Nov 17, 2010
Number of Fields	32 in total (14 numeric and 16 categorical fields)
Number of Records	1,070,994

Table 1.1: File Description

1.2 Summary Statistics Table

The following tables show basic statistics for each field. There are a total of 14 numeric fields and 18 categorical fields.

1.2.1 Numeric Fields:

Field	# Records	% Populated	# Unique	# Zero	Mean	Std	Min	Max
LTFRONT	1070994	100	1297	169108	36.64	74.03	0	9999
LTDEPTH	1070994	100	1370	170128	88.86	76.4	0	9999
STORIES	1014730	94.75	111	0	5.01	8.37	1	119
FULLVAL	1070994	100	109324	13007	874264.51	11582430.99	0	6150000000
AVLAND	1070994	100	70921	13009	85067.92	4057260.06	0	2668500000
AVTOT	1070994	100	112914	13007	227238.17	6877529.31	0	4668308947
EXLAND	1070994	100	33419	491699	36423.89	3981575.79	0	2668500000
EXTOT	1070994	100	64255	432572	91186.98	6508402.82	0	4668308947
BLDFRONT	1070994	100	612	228815	23.04	35.58	0	7575
BLDDEPTH	1070994	100	621	228853	39.92	42.71	0	9393
AVLAND2	282726	26.4	58591	0	246235.72	6178962.56	3	2371005000
AVTOT2	282732	26.4	111360	0	713911.44	11652528.95	3	4501180002
EXLAND2	87449	8.17	22195	0	351235.68	10802212.67	1	2371005000
EXTOT2	130828	12.22	48348	0	656768.28	16072510.17	7	4501180002

Table 1.2.1: Summary Statistics of Numeric Fields

1.2.2 Categorical Fields:

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

Field	Non-null Records	% Populated	Unique Values	Most Common Value
RECORD	1070994	100.00	1070994	N/A
B	1070994	100.00	5	4
BLOCK	1070994	100.00	13984	3944
LOT	1070994	100.00	6366	1
BBLE	1070994	100.00	1070994	N/A
EASEMENT	4636	0.43	12	E
OWNER	1039249	97.04	863347	PARKCHESTER PRESERVAT
BLDGCL	1070994	100.00	200	R4
TAXCLASS	1070994	100.00	11	1
EXT	354305	33.08	3	G
EXCD1	638488	59.62	129	1017
EXCD2	92948	8.68	60	1017
STADDR	1070318	99.94	839280	501 SURF AVENUE
EXMPTCL	15579	1.45	14	X1
ZIP	1041104	97.21	196	10314
PERIOD	1070994	100.00	1	FINAL
YEAR	1070994	100.00	1	2010/11
VALTYPE	1070994	100.00	1	AC-TR

Table 1.2.2: Summary Statistics of Categorical Fields

1.3 Field Examples

1.3.1 Field 'STORIES'

Description: The number of stories for the building (# of floors)
Type: Numeric
Exclude outliers more than 50. Data in figure 1.3.1 is 99.5% populated.
The plot shows a right skewed distribution. 909398 records in total (89.62% of the entire dataset) have lower than 10 stories.

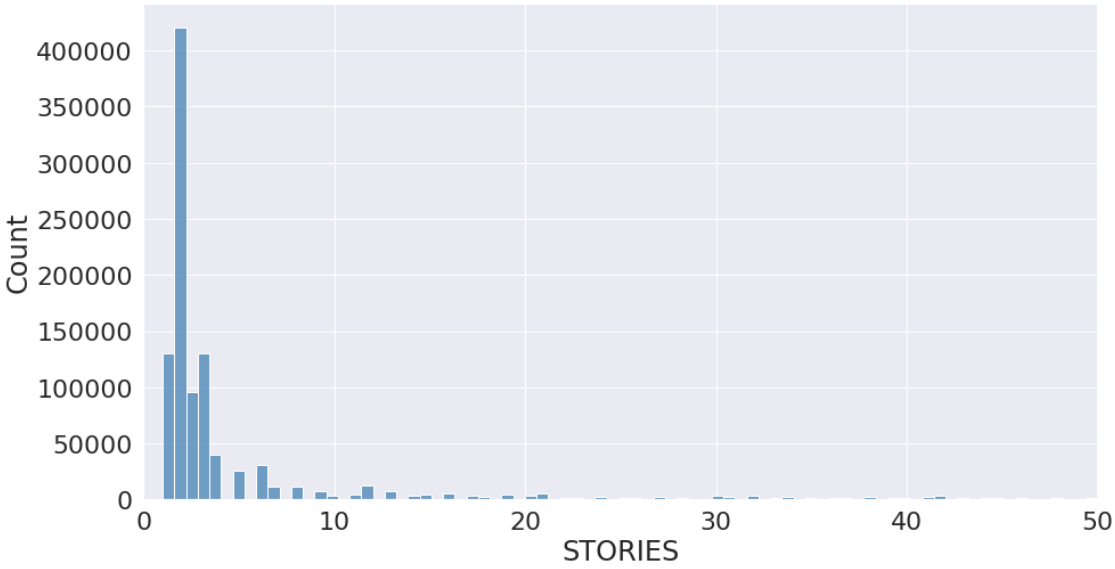


Figure 1.3.1: Frequency distribution of the *'STORIES'* field

1.3.2 Field 'FULLVAL'

Description: total market value of the land
Type: Numeric
Exclude outliers more than 2,000,000. Data in figure 1.3.2 is 96.31% populated.

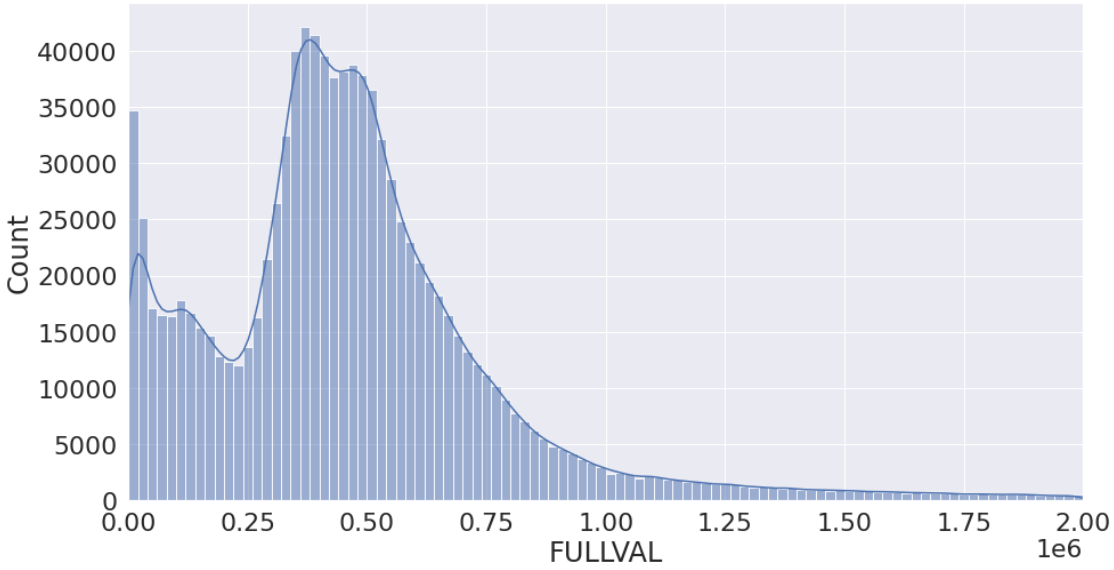


Figure 1.3.2: Frequency distribution of the *'FULLVAL'* field

1.3.3 Field 'AVLAND'

Description: Assessed land value
Type: Numeric
Exclude outliers more than 50,000. Data in figure 1.3.3 is 90.53% populated.

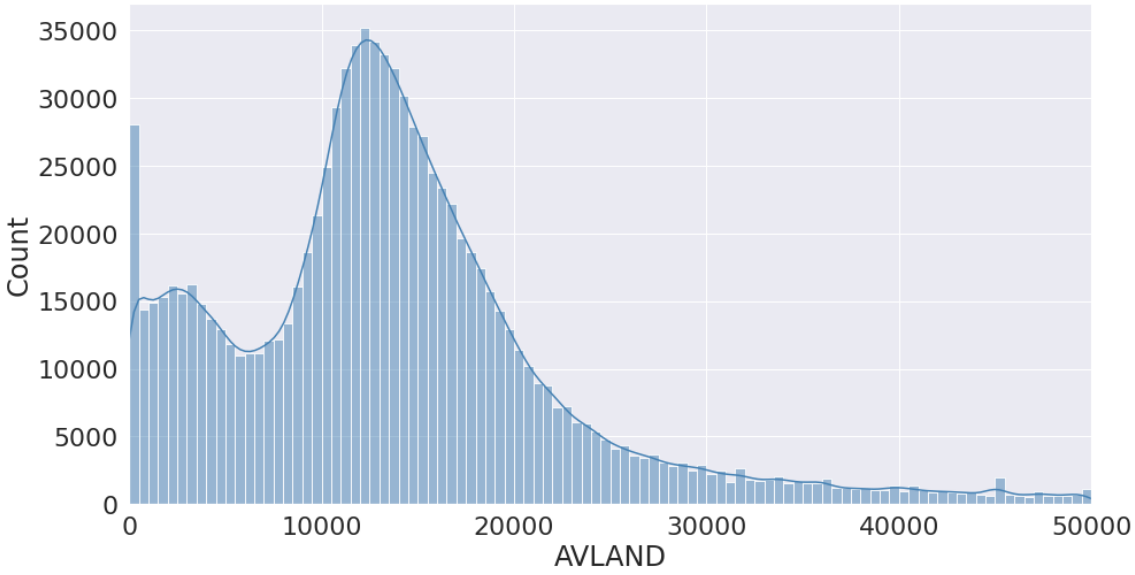


Figure 1.3.3: Frequency distribution of the *'AVLAND'* field

1.3.4 Field 'AVTOT'

Description: Assessed total value
Type: Numeric
Exclude outliers more than 100,000. Data in figure 1.3.4 is 86.05% populated.

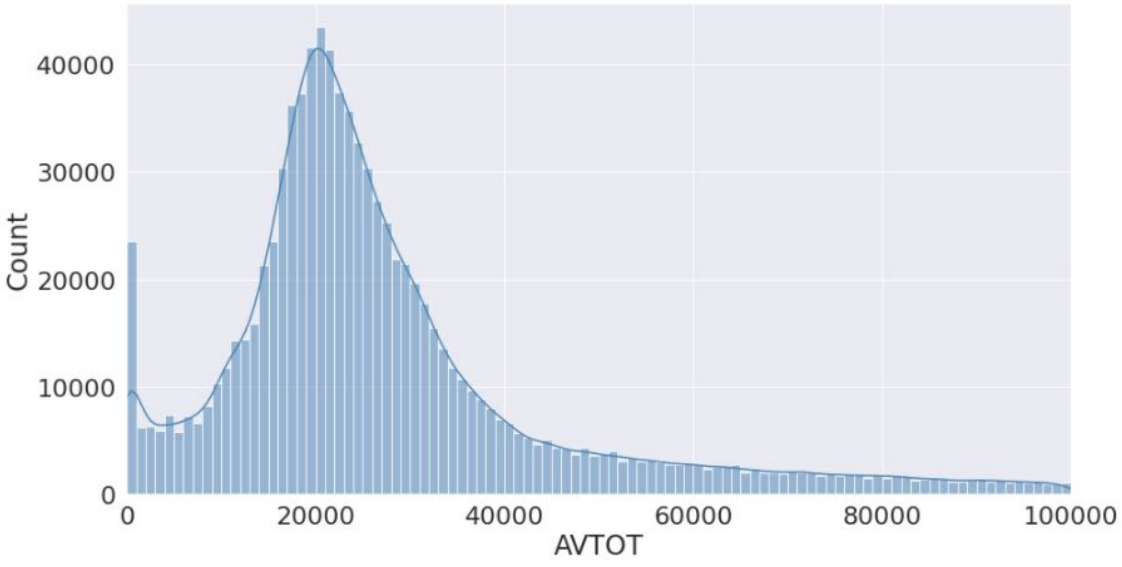


Figure 1.3.4: Frequency distribution of the *'AVTOT'* field

1.3.5 Field 'LTFRONT'

Description: Lot Frontage in feet
Type: Numeric
Exclude outliers more than 300. Data in figure 1.3.5 is 99.3% populated.
169108 records (15.79% of the entire dataset) have a value of 0.

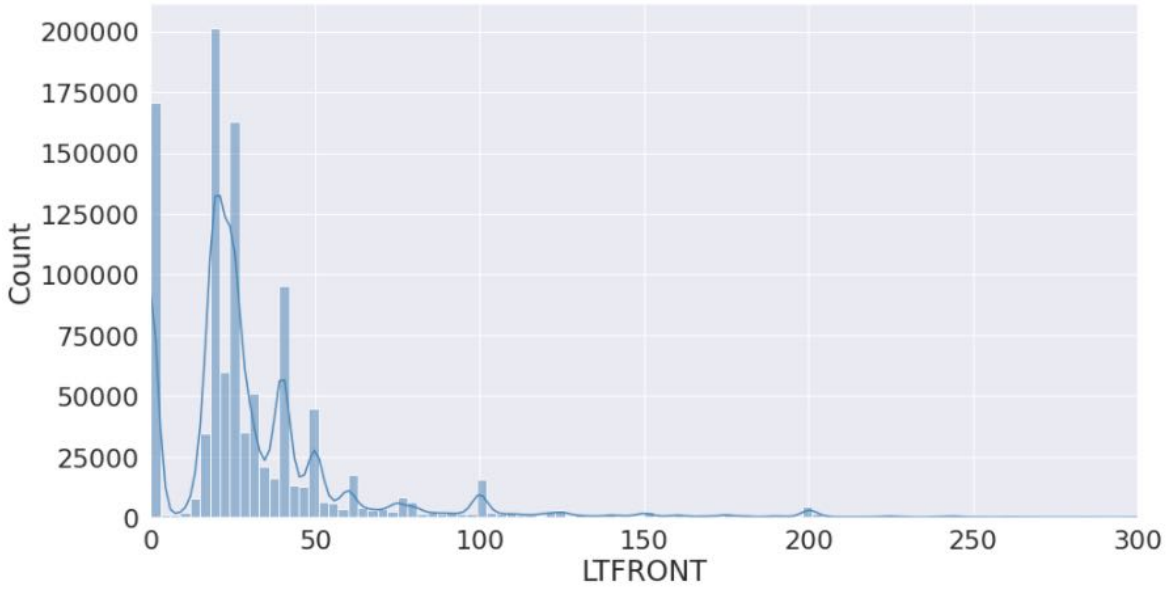


Figure 1.3.5: Frequency distribution of the *'LTFRONT'* field

1.3.6 Field 'LTDEPTH'

Description: Lot depth in feet
Type: Numeric
Exclude outliers more than 300. Data in figure 1.3.6 is 99.17% populated.
170128 records (15.89% of the entire dataset) have a value of 0.
464541 records (43.37% of the entire dataset) have a value of 100.

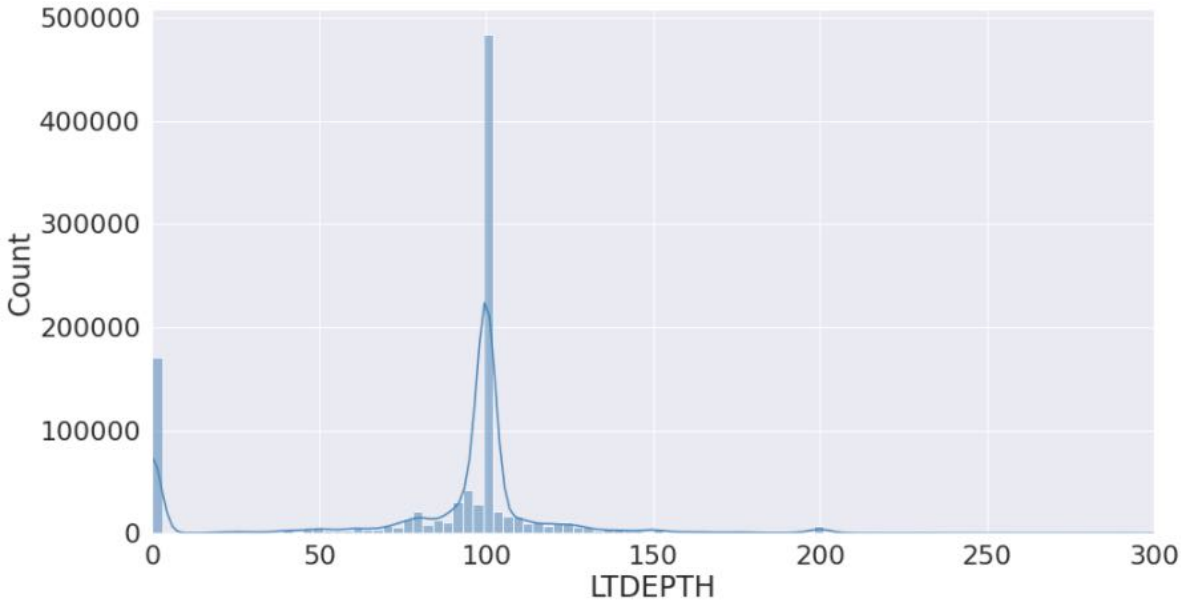


Figure 1.3.6: Frequency distribution of the *'LTDEPTH'* field

1.3.7 Field 'BLDFRONT'

Description: Building frontage in feet
Type: Numeric
There are no values more than 800; therefore data in figure 1.3.7 is 100% populated.

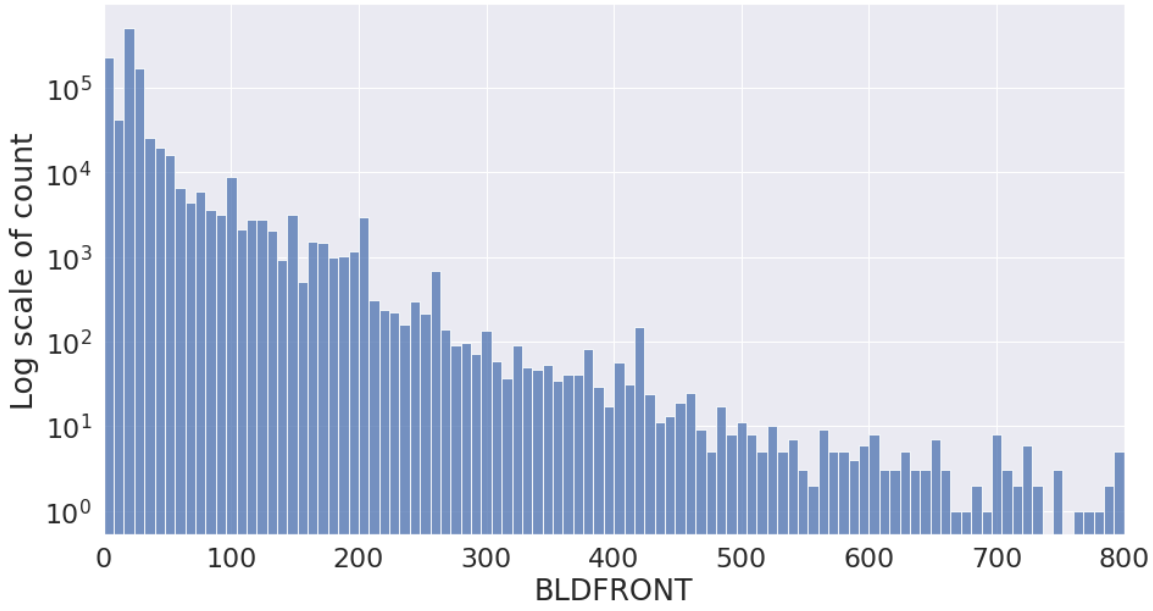


Figure 1.3.7: Frequency distribution of the *'BLDFRONT'* field

1.3.8 Field 'BLDDEPTH'

Description: Lot depth in feet
Type: Numeric
Exclude outliers more than 200. Data in figure 1.3.8 is 99.56% populated.
228815 records (21.36% of the entire dataset) have a value of 0.

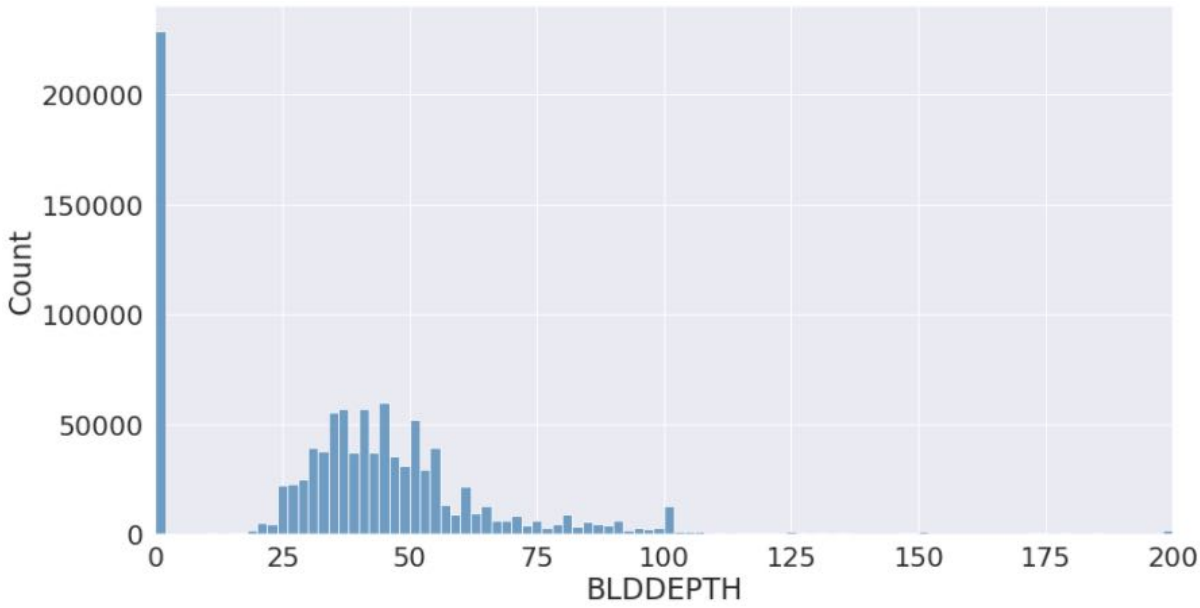


Figure 1.3.8: Frequency distribution of the *'BLDDEPTH'* field

1.3.9 Field ‘ZIP’

Description: Postal Zip code of the property
Type: Categorical

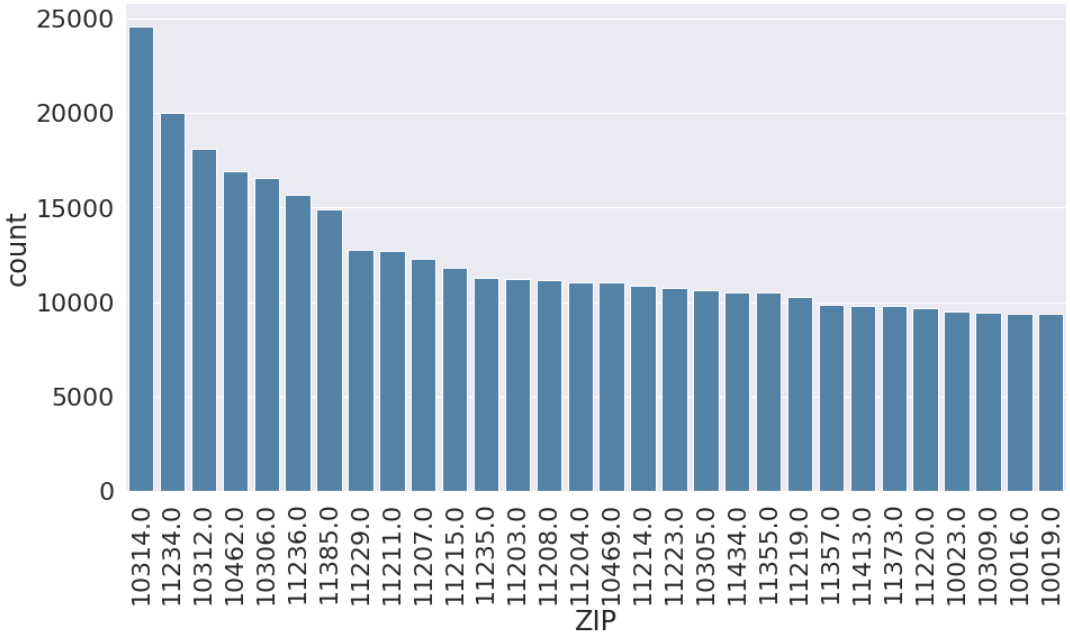


Figure 1.3.9: Frequency Distribution of the ‘*ZIP’* field
(Top 20 most common values)

2. Data Cleaning

2.1 Exclusions

Since we are looking for potential property tax fraud, we would like to remove properties owned by the city, the state and the federal government. In this step, we removed 24,168 records.

PARKCHESTER PRESERVAT	PARKS AND RECREATION	DCAS	HOUSING PRESERVATION
CITY OF NEW YORK	DEPT OF ENVIRONMENTAL	BOARD OF EDUCATION	NEW YORK CITY HOUSING
CNY/NYCTA	NYC HOUSING PARTNERSH	DEPARTMENT OF BUSINES	DEPT OF TRANSPORTATIO
MTA/LIRR	PARCKHESTER PRESERVAT	MH RESIDENTIAL 1, LLC	LINCOLN PLAZA ASSOCIA
UNITED STATES OF AMER	U S GOVERNMENT OWNRD	THE CITY OF NEW YORK	NYS URBAN DEVELOPMENT
NYS DEPT OF ENVIRONME	CULTURAL AFFAIRS	DEPT OF GENERAL SERVI	DEPT RE-CITY OF NY

Table 2.1: Removed Owners List

2.2 Missing Values

During the data exploration stage, we found 9 fields with missing values that need to be imputed using appropriate values. We decided to use the TAXCLASS groupwise average to fill in the missing value because tax class is a good indicator of property type. Table 2.2 below shows how the missing values were imputed.

Field	**Original Value**	**Count**	**Imputation**
ZIP	NA	21772	The record of zip before and after
FULLVAL	NA	13007	Tax Class groupwise average
AVLAND	NA	13009	Tax Class groupwise average
AVTOT	NA	13007	Tax Class groupwise average
STORIES	NA	43968	Tax Class groupwise average
LOTSIZE (LTFRONT & LTDEPTH)	0 & 1	169947 & 170255	1.NA 2.Replace NA with Tax Class groupwise average
BLDSIZE (BLDFRONT & BLDDEPTH)	0 & 1	228892 & 228912	1.NA 2.Replace NA with Tax Class groupwise average

Table 2.2: Missing Values

- Filling missing ZIPs
    - There are 21,772 records that are missing the ZIP field.
    - If the zip code on both the records preceding and following the missing record are the same, we filled in the missing with that zip code. After doing this, 10,400 missing zips were replaced.
    - For the rest of the missing values, we replaced them with the zip code from the record above it.

- Filling missing FULLVAL, AVLAND, and AVTOT
    - There are about 13,000 records missing these 3 property dollar values.
    - Since the missing values could be NA or 0, we first converted all missing values to the same format as NA.
    - Finally, we grouped the data by TAXCLASS and filled in the missing values with the average.

- Filling missing STORIES
    - There are 43,968 records that are missing STORIES(the number of stories in the building)
    - Repeatedly, we grouped the data by TAXCLASS since it's a good indication of the nature of the building and replaced the missing values with the average.

- FIlling missing Lot and Building Size
    - There are about 200,000 records missing some or all of the lot and building sizes.
    - Firstly, we replaced the 0's and 1's by NAs, so they are not counted in calculating mean.
    - Then, for each of these 4 features, we calculated the TAXCLASS groupwise average and imputed the values .

3. Feature Engineering

A total of 48 candidate variables were created. The 3 main variables were created based on lot front, lot depth, building front, building depth, stories. The 9 variables were created based on these 3 main variables with the dollar values (FULLVAL, AVLAND and AVTOT). The remaining 36 variables were created based on the groupwise average of zip5, zip3, tax class and borough. Table 3.1 summarizes the description and the number of variables created in each category.

Category	**Description**	**# Variable Created**
Lot area	LTFRONT x LTDEPTH, area of the land	1
Bld area	BLDFRONT x BLDDEPTH, area of the building	1
Bld vol	BLDAREA x STORIES, volume of the building	1
R1,R2,R3,R4,R5,R6,R7,R8,R9	Mean Market Value of 1 square feet	9
R1,R2,R3,R4,R5,R6,R7,R8,R9 (zip5)	Ratio of Mean Market Value of record to zip 5 groupwise average	9
R1,R2,R3,R4,R5,R6,R7,R8,R9 (zip3)	Ratio of Mean Market Value of record to zip 3 groupwise average	9
R1,R2,R3,R4,R5,R6,R7,R8,R9 (tax class)	Ratio of Mean Market Value of record to tax class groupwise average	9
R1,R2,R3,R4,R5,R6,R7,R8,R9 (borough)	Ratio of Mean Market Value of record to borough groupwise average	9

Table 3.1: Summary of 48 Candidate Variables

- Since the bigger the property the more expensive, we believed that price per square foot, both for land and for a building, was a good standardized metric for the property value.
    - Created 3 sizes
        - lotarea = LTFRONT * LTDEPTH
        - lbdarea = BLDFRONT * BLDDEPTH
        - bldvol = bldarea * STORIES
    - Calculated 9 ratio variables, each of the 3 monetary value fields divided by each of these 3 sizes. (FULLVAL, AVLAND and AVTOT)

- Location has a large influence on a property's value, and zip code and borough are reasonable indicators of location. To learn more about the value, we created 2 new feature zip5 (which is the 5 digits of zip code) and zip3 (which is the first 3 digits of zip code)

- On the other hand, we used TAXCLASS as an indicator for the property type.

- After building these variables, we divided each of the 9 ratio variables by the 4 scale factors from these groups mentioned above, which gave us a total of 45 variables in the end.

- 45 variables
    - r1: total market value of the property / lot size
    - r2: total market value of the property / building size
    - r3: total market value of the property / building volume
    - r4: total market value of the land / lot size
    - r5: total market value of the land / building size
    - r6: total market value of the land / building volume
    - r7: total market value / lot size
    - r8: total market value / building size
    - r9: total market value / building volume
    - r1_zip5: r1 / average r1 grouping by zip5
    - r2_zip5: r2 / average r2 grouping by zip5
    - r3_zip5: r3 / average r3 grouping by zip5
    - r4_zip5: r4 / average r4 grouping by zip5
    - r5_zip5: r5 / average r5 grouping by zip5
    - r6_zip5: r6 / average r6 grouping by zip5
    - r7_zip5: r7 / average r7 grouping by zip5
    - r8_zip5: r8 / average r8 grouping by zip5
    - r9_zip5: r9 / average r9 grouping by zip5
    - r1_zip3: r1 / average r1 grouping by zip3
    - r2_zip3: r2 / average r2 grouping by zip3
    - r3_zip3: r3 / average r3 grouping by zip3
    - r4_zip3: r4 / average r4 grouping by zip3
    - r5_zip3: r5 / average r5 grouping by zip3
    - r6_zip3: r6 / average r6 grouping by zip3
    - r7_zip3: r7 / average r7 grouping by zip3
    - r8_zip3: r8 / average r8 grouping by zip3
    - r9_zip3: r9 / average r9 grouping by zip3
    - r1_taxclass: r1 / average r1 grouping by tax class
    - r2_taxclass: r2 / average r2 grouping by tax class
    - r3_taxclass: r3 / average r3 grouping by tax class
    - r4_taxclass: r4 / average r4 grouping by tax class
    - r5_taxclass: r5 / average r5 grouping by tax class
    - r6_taxclass: r6 / average r6 grouping by tax class
    - r7_taxclass: r7 / average r7 grouping by tax class
    - r8_taxclass: r8 / average r8 grouping by tax class
    - r9_taxclass: r9 / average r9 grouping by tax class
    - r1_boro: r1 / average r1 grouping by borough
    - r2_boro: r2 / average r2 grouping by borough
    - r3_boro: r3 / average r3 grouping by borough
    - r4_boro: r4 / average r4 grouping by borough
    - r5_boro: r5 / average r5 grouping by borough
    - r6_boro: r6 / average r6 grouping by borough
    - r7_boro: r7 / average r7 grouping by borough
    - r8_boro: r8 / average r8 grouping by borough
    - r9_boro: r9 / average r9 grouping by borough

4. Feature Selection

4.1 Feature scaling

For distance-based algorithms, one essential step is to ensure that data is properly scaled across all features.
As the data distributions of dimensions are different, and the measurement among features varies, it is not reasonable to calculate distance based on original scales. One fair solution is to transform data into z-scores, which is the number of standard deviations a point is away from the mean.

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

After z-scaling, for each dimension, the mean is 0, and one unit in distance is one standard deviation of a dimension. Z-score is an efficient and direct indicator for anomalies. A high z-score means an unusual case of the dimension.
Z-scores are further used to calculate distance between points.

4.2 Principal Component Analysis

After feature engineering, there are 45 variables available for analysis. However, more dimensions involved will certainly increase workload while some dimensions cannot work as a powerful indicator of anomalies. Therefore, Principal Component Analysis (PCA) is deployed to select efficient features for fraud analysis and throw away redundant features.

Principal Component Analysis is an unsupervised method that reduces dimensions and removes linear correlations using single value decomposition.

The first step of principal component analysis is to figure out how many variables will be included in further analysis. The ideal situation is that a few selected variables account for 80% to 90% of the total variance. The *PCA* function of *decomposition* module of *sklearn* library will calculate each variable's variance coverage based on z-scores in this case. A cumulative chart can show the total variance covered by top representative features. In the NYC Property case, the top seventeen features have a total variance coverage of 99%, while the top six features together cover about 90% of total variance.

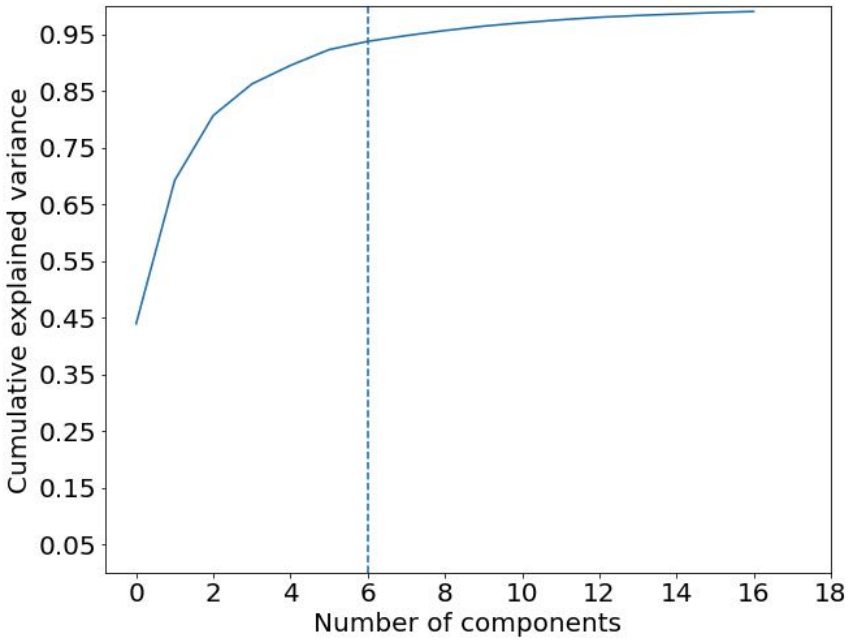


Figure 4.2.1: Cumulative explained variance by number of components

A scree plot can easily illustrate the percentage of explained variance of the first seventeen variables and how the percentage varies among different features. Features that have low explained variance will be dropped, and only the top 6 out of 45 will be used for further analysis.

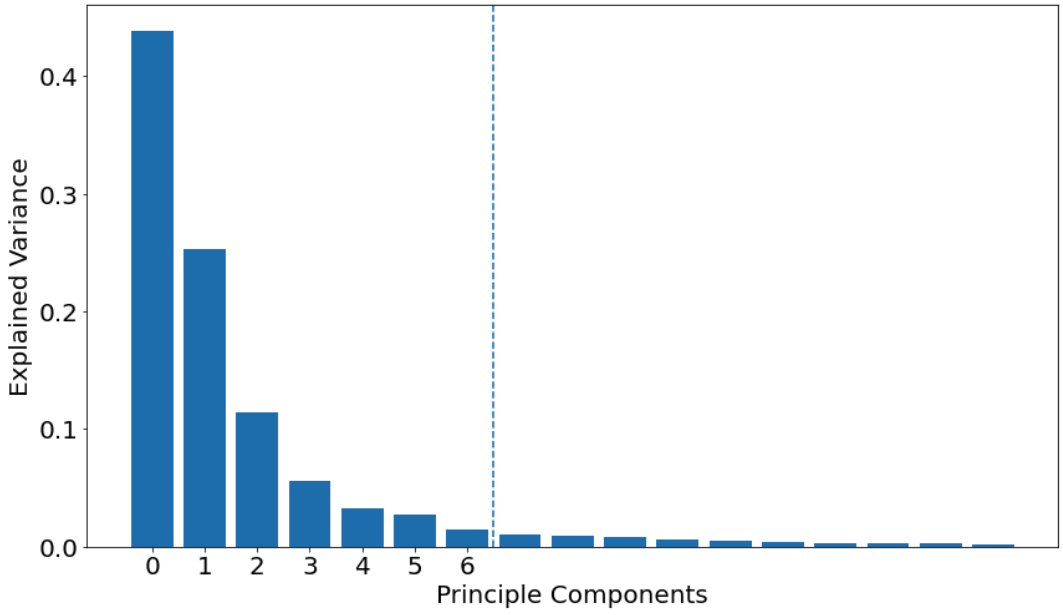


Figure 4.2.2: Scree Plot from the Principal Component Analysis (PCA)

5. Fraud Model Algorithms

5.1 Score 1

Score 1 is a function of six z-scaled principal components (PCs) which separate extreme values or outliers from the dataset. The score is expressed as the Minkowski distance of a vector of z-scaled PCs, written as:

$$s_1^{(i)} = \left(\sum_{k=1}^{6} |PCz_k^{(i)}|^p \right)^{1/p}$$

where the score 1 for record i is the Lp norm of the z-scaled record, consisting of six PCs.

5.2 Score 2

Score 2 is measured by the difference between the original z-scaled records and the z-scaled records reproduced by an autoencoder trained on the dataset. The score is also expressed as the Minkowski between the original and reproduced records, written as:

$$s_2^{(i)} = \left(\sum_{k=1}^{6} |(PCz_k^{(i)})' - PCz_k^{(i)}|^p \right)^{1/p}$$

where the score 2 for record i is the Lp norm of the coordinate-wise difference between the original and reproduced z-scaled records.

5.3 Final Score

The final score is the arithmetic average of score 1 and score 2, written as:

$$s_{final}^{(i)} = \frac{s_1^{(i)} + s_2^{(i)}}{2}$$

6. Results

The top 100 anomalous records were identified based on the final fraud score which is the average of scores 1 and 2. The following presents the assessment result of the top 5 anomalous records as well as 5 other records that stand out from the list.

6.1 Top 5 anomalous records

6.1.1. Record #917942

This property is flagged as the most anomalous record mainly because the lot depth and building dimensions (fields LTDEPTH, BLDFRONT, and BLDDEPTH) were reported as zero on the record, causing the variables r5, r8 and r9 to be extremely high. It was also reported as a 3-story building but was found to be a 7-story one from further investigation. As shown in the figure below, the actual lot and building seem to be very different from what is reported on the record.

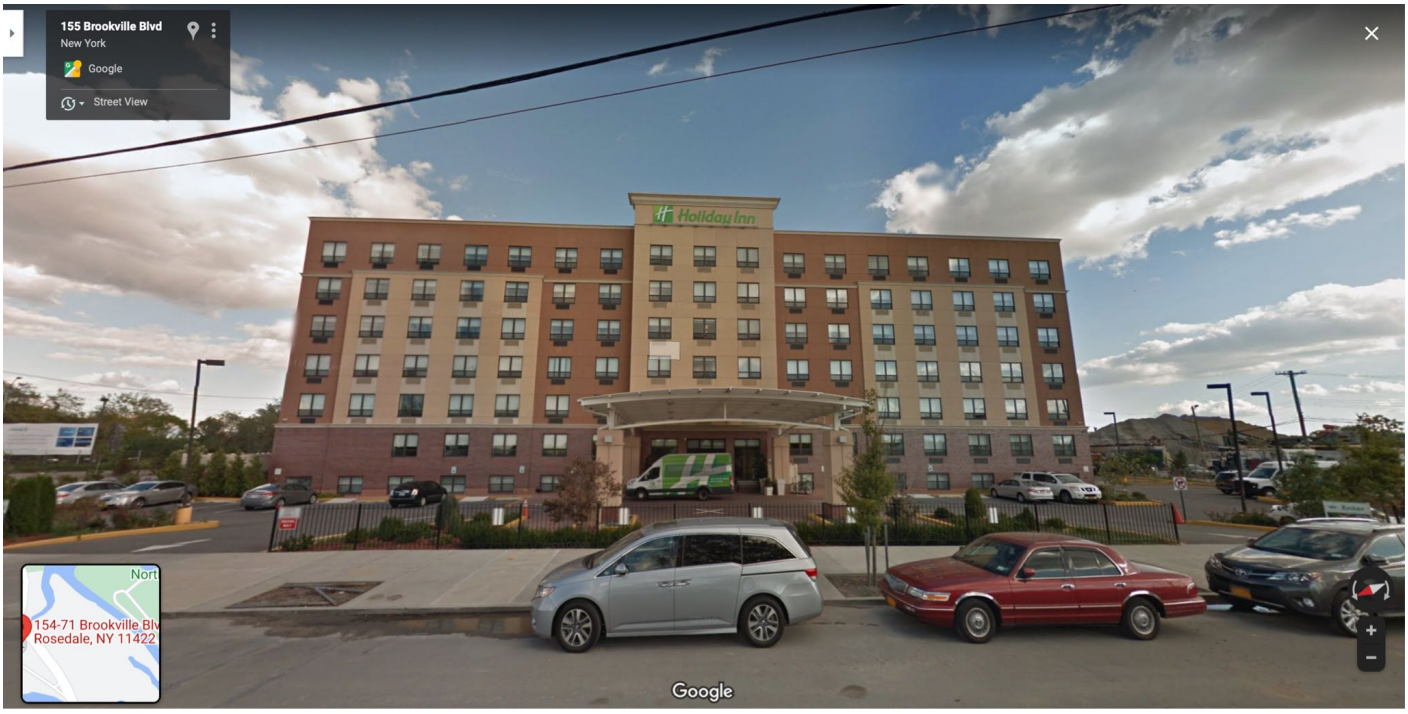


Figure 6.1.1: Screenshot of the property for record #917942

6.1.2. Record #684704

Most of the fields used to calculate the fraud scores are either missing or incomplete for this property. More specifically, the fields FULLVAL, AVLAND, AVTOT and STORIES are missing while the fields BLDFRONT and BLDDEPTH have a value of zero. The only information available is the lot dimensions (2' x 2'), but this is unusually small for the dollar values that were imputed using the field averages, causing the fraud alarm to set off. Its tax class (1B) suggests it is a residential vacant land, but further investigation was not possible due to the incomplete address; only street name was available.

6.1.3. Record #1065870

The main reason for this anomaly is a very high market value ($290 million) in relation to the building dimensions that was reported as zero. With incomplete address information and missing zip code, it is difficult to conduct further investigation of the property. However, based on its tax class (1B), the property is assumed to be vacant land without any structure, and the high market value may be a reflection of its large lot size (2981’ x 1488).

6.1.4. Record #1059883

The record does not show a valid address and owner but only a street name, Sagona Court. After further investigation, there are a total of 4 properties on the street, and we’re not able to determine which property under limited information. Moreover, it is reported in the dataset that the lot area size of the property is 5’x 5’ square feet, which doesn’t make sense by looking from the figure below. Therefore, it was flagged as an anomaly due to high values on variables such as r1, r4, and r7.

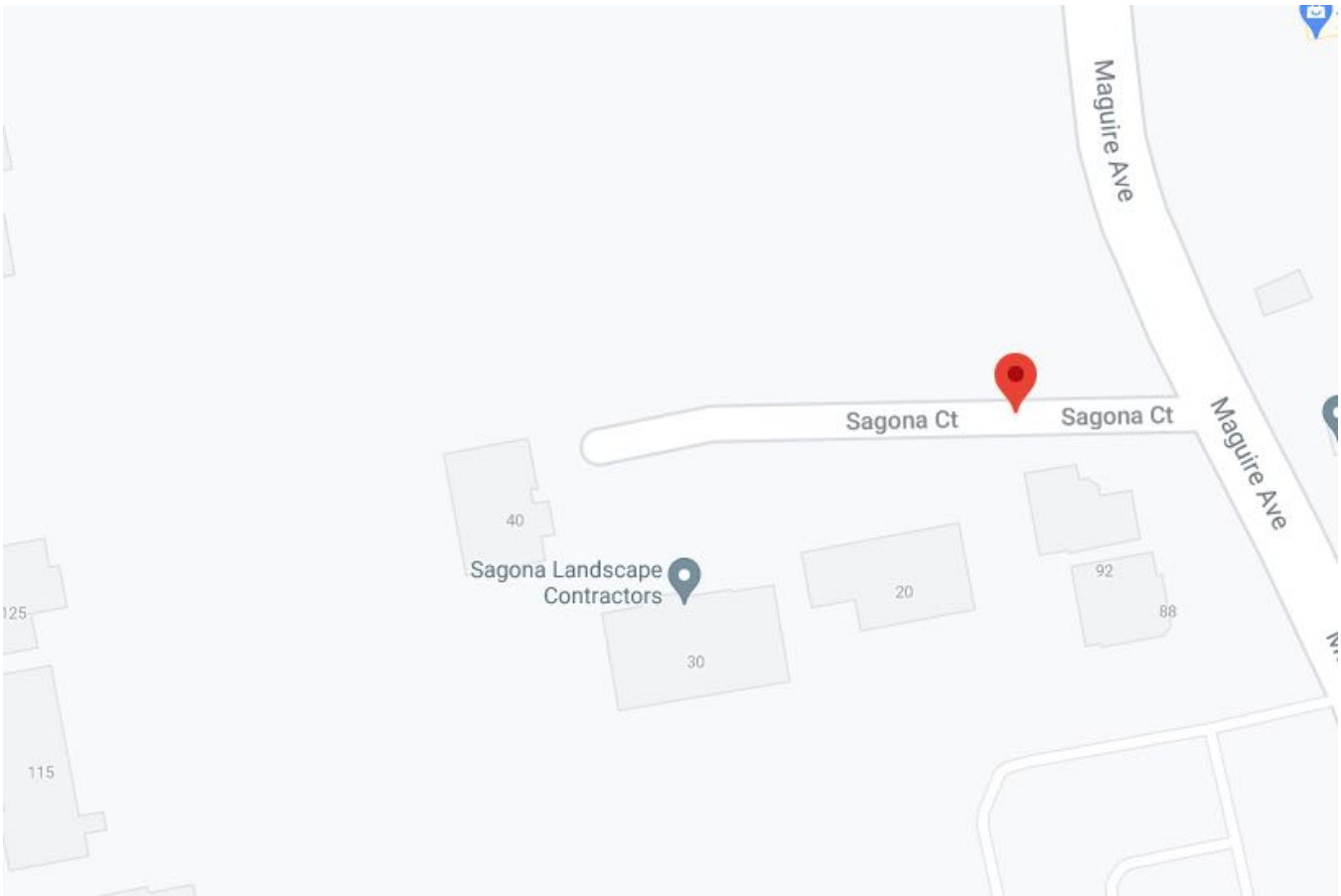


Figure 6.1.4: Screenshot of the property for record #1059883

6.1.5. Record #151044

The street address of this property is “1 East 161 Street”, which turns out to be the Yankee Stadium. Therefore, it is reasonable to have an extremely high total market value of more than 1 billion. However, in the dataset, it was reported that the building size is 0’ x 0’ square feet, which is the main reason that leads to a high value on r2, r5, and r8.

Figure 6.1.5: Screenshot of the property for record #151044

6.2 Other anomalous records within the top 100

6.2.1. Record #41924

The street address for this property is 7 West 21 Street, which is owned by EMB REALTY LLC. This property is a flatiron luxury rental apartment with 20 stories and 231 units, however it is reported as 1 story. Besides its Lot front and Lot depth are reported as 106 and 206 feet respectively while building front and building depth are 10 and 18 feet respectively. Given that this property has a valuation of $7.63 million, it does not make sense for such a large property to have such a small value. This causes all expert variables in comparison to the others substantially high, which is likely why the analytics pipeline gave it a high anomaly score.

(Screenshot of the property shown in the next page)

Figure 6.2.1: Screenshot of the property for record #41924

6.2.2. Record #818084

This property is owned by Mary Immaculate Vacant Lot LLC. Its lot front and lot depth sizes are 112 and 120 feet respectively, while building front and building depth sizes are 4 and 6 feet respectively. Also, it is reported that it is a 1 story building. By searching, however, it is found out that this property is a rental apartment with 8 floors and 57 units. With the external information, it is very anomalous to have such a small valuation as 880 thousands, which leads to expert variables such as r2, r3, r5, r6, r8 and r9 extremely high. In this case, this record can be viewed as anomalous.

Figure 6.2.2: Screenshot of the property for record #818084

6.2.3. Record #330291

This property is a 4 story building located in Pratt Institute at 166 Willoughby Avenue, Brooklyn. It was flagged as anomalous because it has unusually high market value for its building size. The building dimensions on the data are 6’ x 8’ (front x depth), which seem unusually small for its lot size (200’ x 698’) and its market value (FULLVAL) of $13.5 million. With further investigation, it was discovered that the actual building is much larger than 6’ x 8’ as shown in the figure below, indicating incorrect data and possibly fraud.

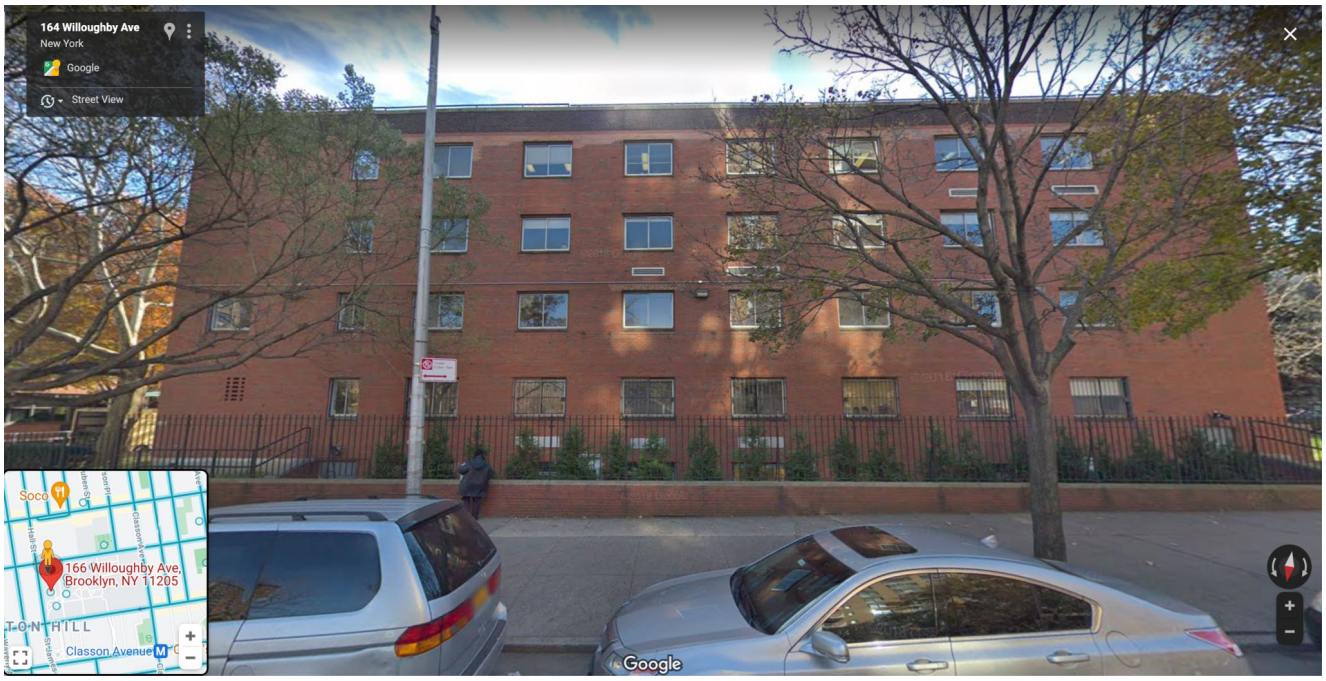


Figure 6.2.3: Screenshot of the property for record #330291

6.2.4. Record #14979

The property is owned by Enjay Associates and located at 69 GRAND STREET, New York. In the dataset, it is reported as a 1-story building with a size of 8’ x 6’ (front x depth) squared feet. However, further investigations show that it is a 8-story building with a much larger actual building size. Therefore, it was flagged as anomalous since it has extremely high value on variables such as r3, r6 and r9, which are mainly related to the building size and the number of stories.

(Screenshot of the property shown in the next page)

Figure 6.2.4: Screenshot of the property for record #14979

6.2.5. Record # 39770

This property is owned by Greenhorn Development, and locates at 142 WEST 23 STREET, Manhattan. Most of the values about this property are true, however, its building front and building depth are both recorded as 8', which makes the building a tiny shed of 64 square feet while indeed it is a 13-story building of normal size. The extremely low value in building size results in high values of r2, r5 and r8, and therefore marks the property as one of the top anomalies.

Figure 6.2.5: Screenshot of the property for record #39770

7. Conclusions

In this project, our main goal was to identify probable fraudulent records with anomalous field values in the NY Property dataset as candidates for fraud investigations.

The project involved the following five steps: data preparation, feature engineering, feature selection, fraud algorithm, and anomaly assessment. To begin with, we prepared a thorough Data Quality Report to assess the quality of the dataset, through which we discovered duplicate owners of records and missing values in some fields. Therefore, we excluded these duplicates and filled in missing field values using adjacent values and grouped averages. Next, we created nine features using size variables and scaled each of the nine features by four grouping variables (e.g. ZIP code and borough code). Since the number of features is large for fraud algorithms, we extracted principal components (PC) from the feature space and decided to keep six PCs as the features for the algorithms. After z-scaling these six PCs, we calculated their L2 norms as score 1, calculated the difference between the original PCs and the reproduced PCs through an autoencoder as score 2, and calculated the arithmetic average of score 1 and score 2 as the final score. Based on the fraud score ranking of all records, we identified the top 100 records with the highest final fraud scores as the most probable fraudulent records.

Among the top 100 records, 10 records were selected and assessed to discover the reasons for high fraud scores. We discovered the three main reasons:

1) Missing values were imputed by the grouped averages, which in some cases result in values that were not relative to other features of a property. For example, for the record #684704, the missing FULLVAL was replaced with the average value which was too high for its lot size (2 feet x 2 feet) reported on the record.
2) Some records contain incorrect data. For example, for the record #917942, the lot depth and building dimensions were reported as zero on the record whereas the actual building is a 7-story large hotel. However, incorrect data is less likely to be intentional data manipulation but more likely to be data input error because the dataset was created by government employees, not individual taxpayers. There is room for data quality improvement.
3) The dataset is outdated, and multiple changes were made over the period. For instance, some properties were recorded to have only one story, but it was sold as a multi-story building when it was fully constructed years later.

Some of the limitations discussed above can be improved in the following ways:

1) Fill in missing values with values that better represent the record. For example, instead of replacing FULLVAL, AVLAND, and AVTOT with the overall average, we can group the dollar values by tax class, borough, stories, and building area. The average value of a property could be the mean price of one unit area (1 square feet) times the total area. Since the total value of property has a high variance as properties vary in many dimensions, many data mismatches induced anomalies.
2) Account for abnormal values in some fields. For instance, there are many records with building front or building depth less than 10, which is highly impossible. This could be considered in the data cleaning step.
3) Try to collect and record original data as much as possible to avoid data mismatches from the data fill-in process.

Appendix: Data Quality Report

Field Description

Dataset Name	Property Valuation and Assessment Data
Dataset Purpose	To calculate property tax and grant eligible properties exemptions and/or abatements
Data Source	NYC Open Data
Time Period	Nov 17, 2010
Number of Fields	32 in total (14 numeric and 16 categorical fields)
Number of Records	1,070,994

Table A.0: File Description

Summary Statistics Table

The following tables show basic statistics for each field. There are a total of 14 numeric fields and 18 categorical fields.

Numeric Fields:

Field	# Records	% Populated	# Unique	# Zero	Mean	Std	Min	Max
LTFRONT	1070994	100	1297	169108	36.64	74.03	0	9999
LTDEPTH	1070994	100	1370	170128	88.86	76.4	0	9999
STORIES	1014730	94.75	111	0	5.01	8.37	1	119
FULLVAL	1070994	100	109324	13007	874264.51	11582430.99	0	6150000000
AVLAND	1070994	100	70921	13009	85067.92	4057260.06	0	2668500000
AVTOT	1070994	100	112914	13007	227238.17	6877529.31	0	4668308947
EXLAND	1070994	100	33419	491699	36423.89	3981575.79	0	2668500000
EXTOT	1070994	100	64255	432572	91186.98	6508402.82	0	4668308947
BLDFRONT	1070994	100	612	228815	23.04	35.58	0	7575
BLDDEPTH	1070994	100	621	228853	39.92	42.71	0	9393
AVLAND2	282726	26.4	58591	0	246235.72	6178962.56	3	2371005000
AVTOT2	282732	26.4	111360	0	713911.44	11652528.95	3	4501180002
EXLAND2	87449	8.17	22195	0	351235.68	10802212.67	1	2371005000
EXTOT2	130828	12.22	48348	0	656768.28	16072510.17	7	4501180002

Table A.1: Summary Statistics of Numeric Fields

Categorical Fields:

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

Field	Non-null Records	% Populated	Unique Values	Most Common Value
RECORD	1070994	100.00	1070994	N/A
B	1070994	100.00	5	4
BLOCK	1070994	100.00	13984	3944
LOT	1070994	100.00	6366	1
BBLE	1070994	100.00	1070994	N/A
EASEMENT	4636	0.43	12	E
OWNER	1039249	97.04	863347	PARKCHESTER PRESERVAT
BLDGCL	1070994	100.00	200	R4
TAXCLASS	1070994	100.00	11	1
EXT	354305	33.08	3	G
EXCD1	638488	59.62	129	1017
EXCD2	92948	8.68	60	1017
STADDR	1070318	99.94	839280	501 SURF AVENUE
EXMPTCL	15579	1.45	14	X1
ZIP	1041104	97.21	196	10314
PERIOD	1070994	100.00	1	FINAL
YEAR	1070994	100.00	1	2010/11
VALTYPE	1070994	100.00	1	AC-TR

Table A.3: Summary Statistics of Categorical Fields

Field Description and Distribution

Field 1: RECORD
Description: uniquely identify each record of data
Type: Categorical

Field 2: BBLE
Description: Concatenation of borough code, block code, lot, and easement
Type: Categorical

Field 3: B
Description: Borough code
Type: Categorical

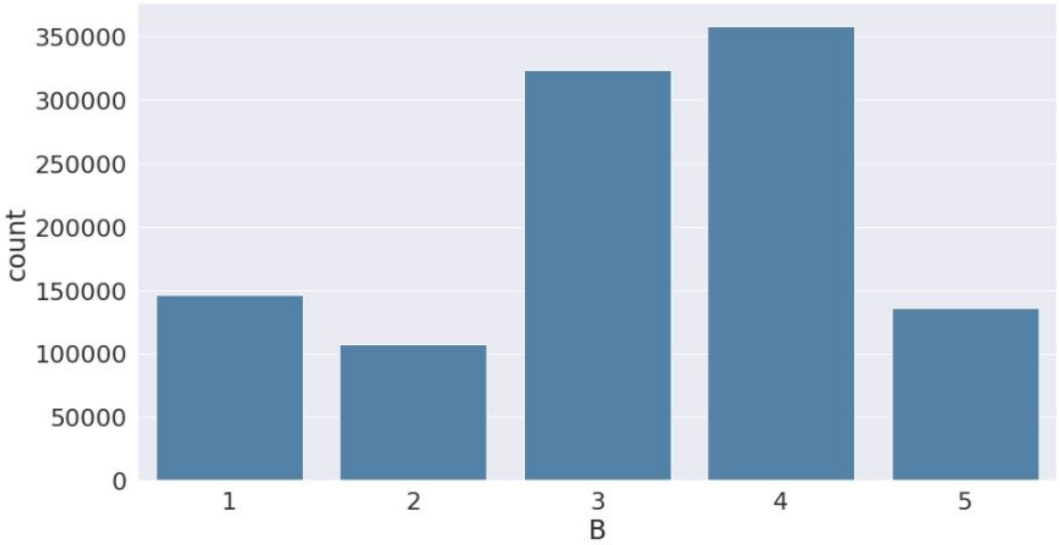


Figure A.1: Frequency Distribution of the *'B'* field

Field 4: BLOCK

Description: Valid block ranges by borough code

- Manhattan 1 to 2,255
- Bronx 2,260 to 5,958
- Queens 1 to 16,350
- Staten Island 1 to 8,050

Type: Categorical

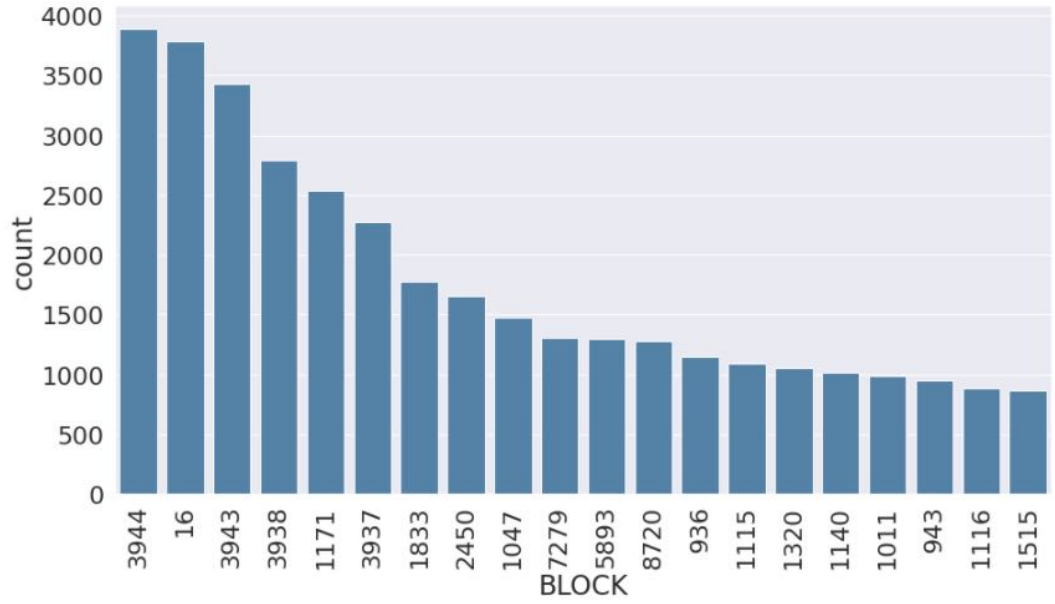


Figure A.2: Frequency Distribution of the *'BLOCK'* field

Field 5: LOT
Description: Unique number within borough/block
Type: Categorical

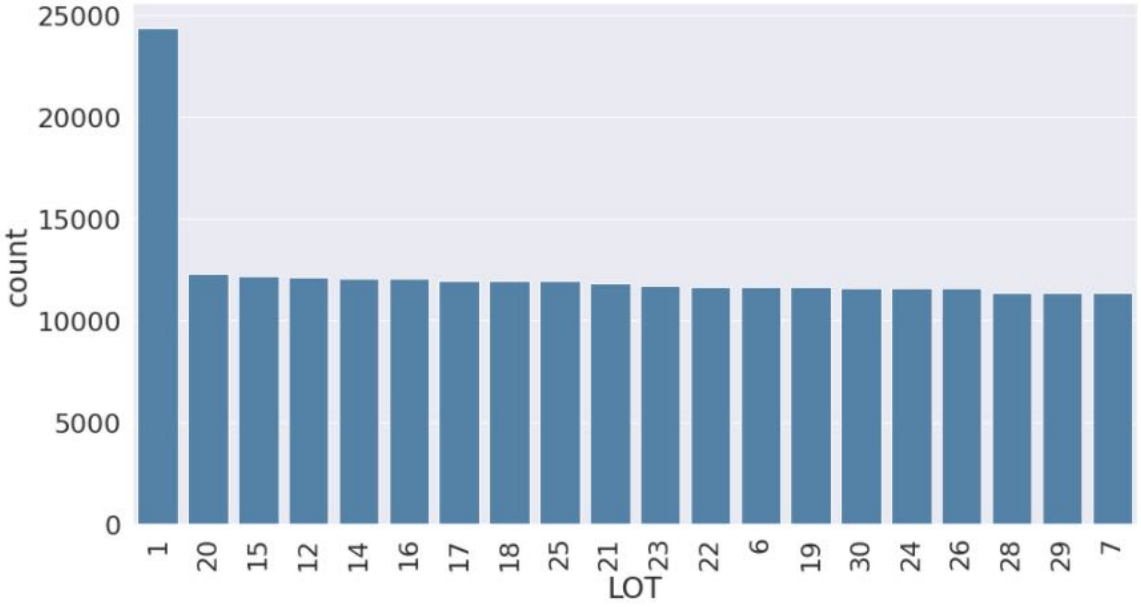


Figure A.3: Frequency Distribution of the *'LOT'* field

Field 6: EASEMENT
Description: A field that is used to describe easement

- 'A' indicates the portion of the lot that has an Air Easement
- 'B' indicates Non-Air Rights
- 'E' indicates the portion of the lot that has a Land Easement
- 'F' through 'M' are duplicates of 'E'
- 'N' indicates Non-Transit Easement
- 'P' indicates Piers
- 'R' indicates Railroads
- 'S' indicates Street
- 'U' indicates U.S. Government

Type: Categorical

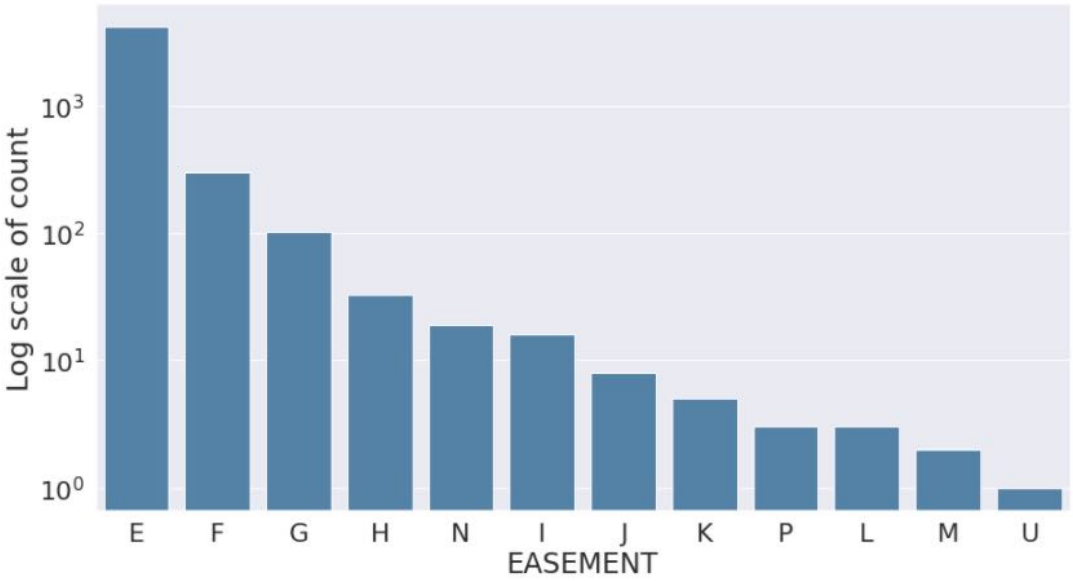


Figure A.4: Frequency Distribution of the *'EASEMENT'* field

Field 7: OWNER
Description: owner's name
Type: Categorical

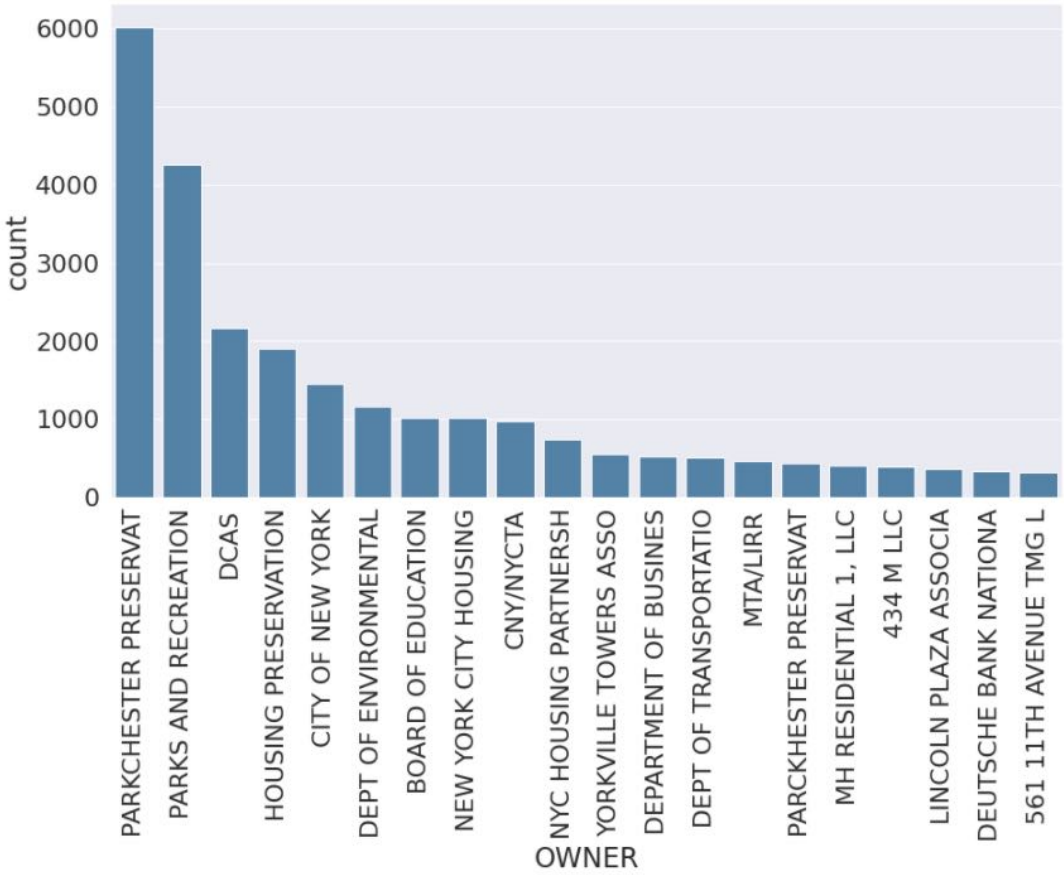


Figure A.5: Frequency Distribution of the *'OWNER'* field

Field 8: BLDGCL
Description: Building Class
Type: Categorical

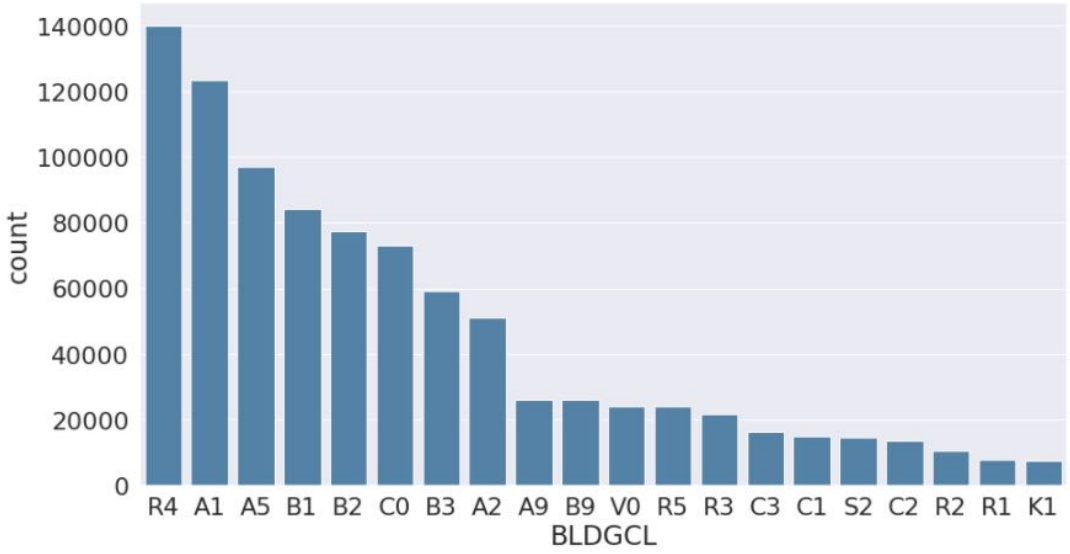


Figure A.6: Frequency Distribution of the *'BLDGCL'* field

Field 9: TAXCLASS
Description: Current Property Tax Class Code (NYS Classification)

- Tax Class 1: 1-3 unit residences
- Tax Class 1A: 1-3 story condominiums
- Tax Class B: residential vacant land
- Tax Class 1C: 1-3 unit condominiums
- Tax Class 1D: select bungalow colonies
- Tax Class 2: apartments
- Tax Class 2A: apartments with 4-6 units
- Tax Class 2B: apartments with 7-10 units
- Tax Class 2C: coops/condos with 2-10 units
- Tax Class 3: utilities (except ceiling RR)
- Tax Class 4A: utilities - ceiling railroads
- Tax Class 4: others

Type: Categorical

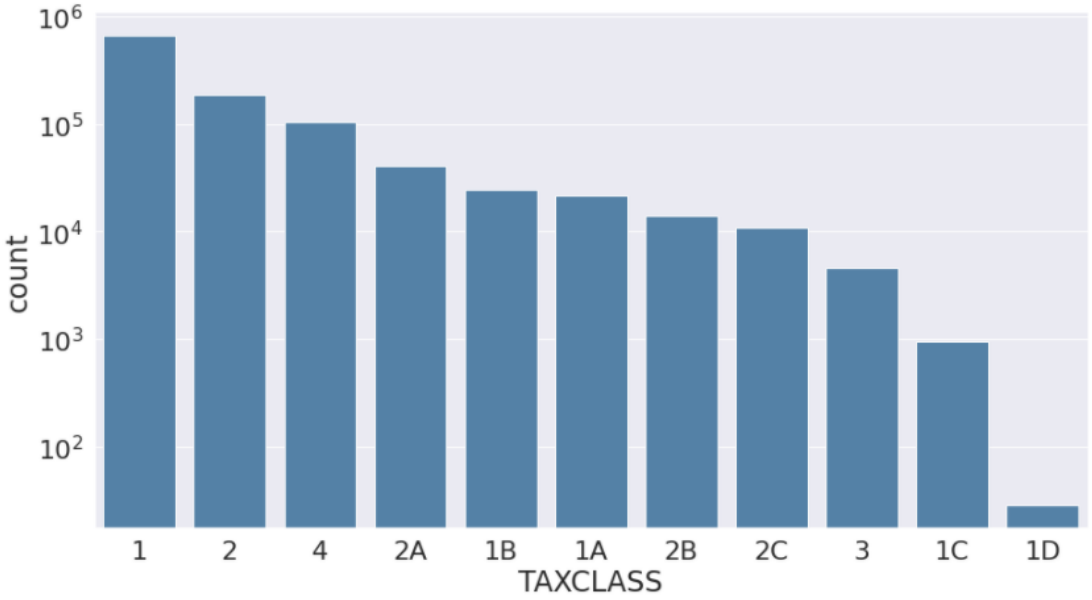


Figure A.7: Frequency Distribution of the *'TAXCLASS'* field

Field 10: LTFRONT

Description: Lot Frontage in feet

Type: Numeric

Exclude outliers more than 300. Data in figure 1.3.5 is 99.3% populated.

169108 records (15.79% of the entire dataset) have a value of 0.

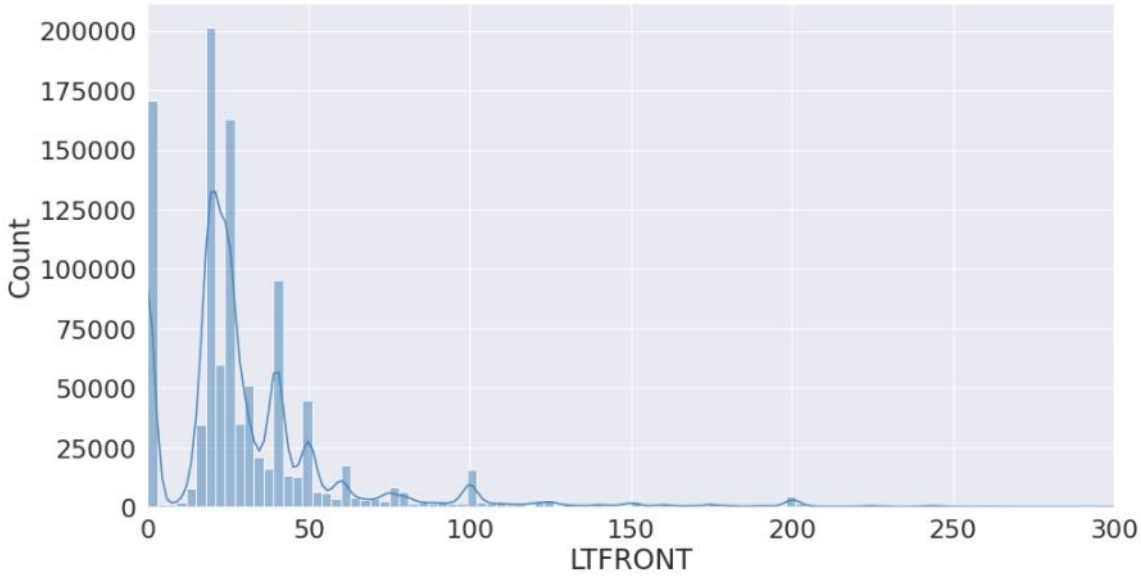


Figure A.8: Frequency Distribution of the *'LTFRONT'* field

Field 11: LTDEPTH
Description: Lot depth in feet
Type: Numeric
Exclude outliers more than 300. Data in figure 1.3.6 is 99.17% populated.
170128 records (15.89% of the entire dataset) have a value of 0.
464541 records (43.37% of the entire dataset) have a value of 100.

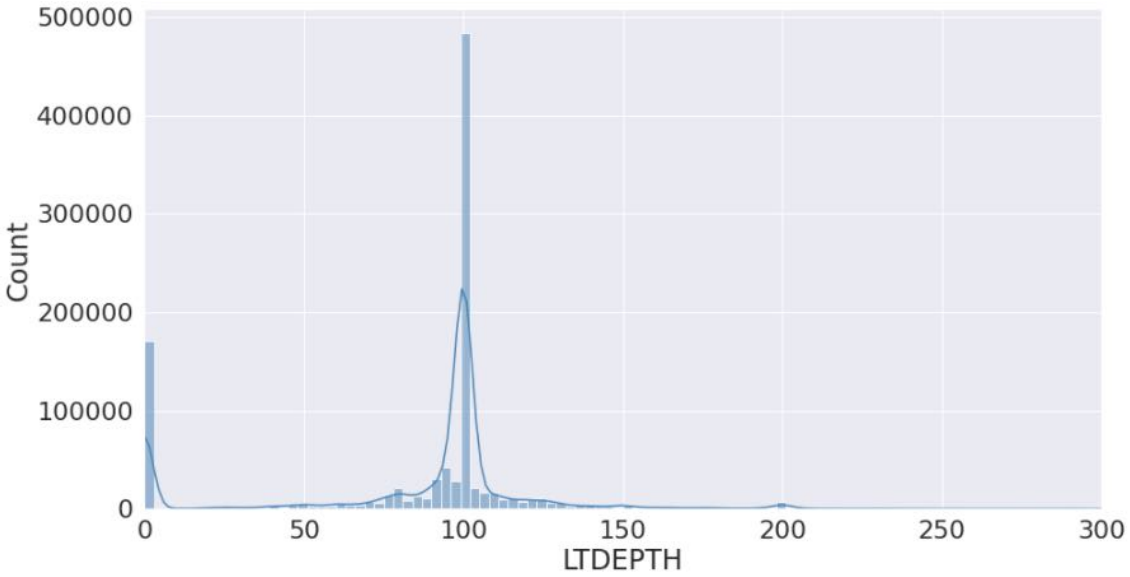


Figure A.9: Frequency Distribution of the *'LTDEPTH'* field

Field 12: EXT
Description: Extension
- 'E': Extension
- 'G': Garage
- 'EG': Extension and Garage

Type: Categorical

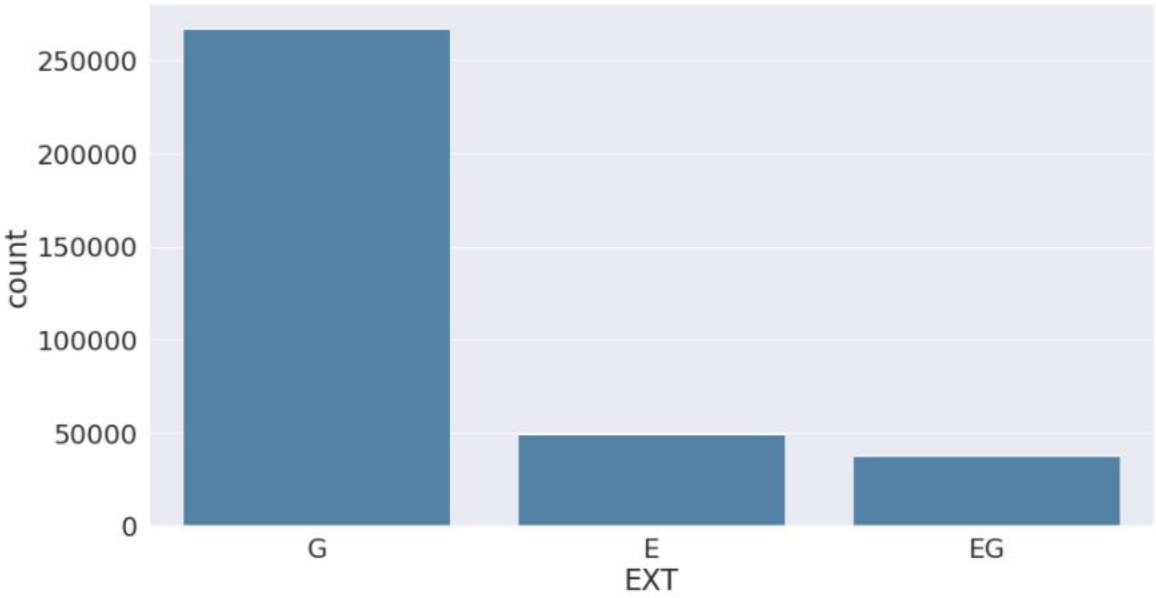


Figure A.10: Frequency Distribution of the *'EXT'* field

Field 13: STORIES
Description: The number of stories for the building (# of floors)
Type: Numeric
Exclude outliers more than 50. Data in figure 1.3.1 is 99.5% populated.
The plot shows a right skewed distribution. 909398 records in total (89.62% of the entire dataset) have lower than 10 stories.

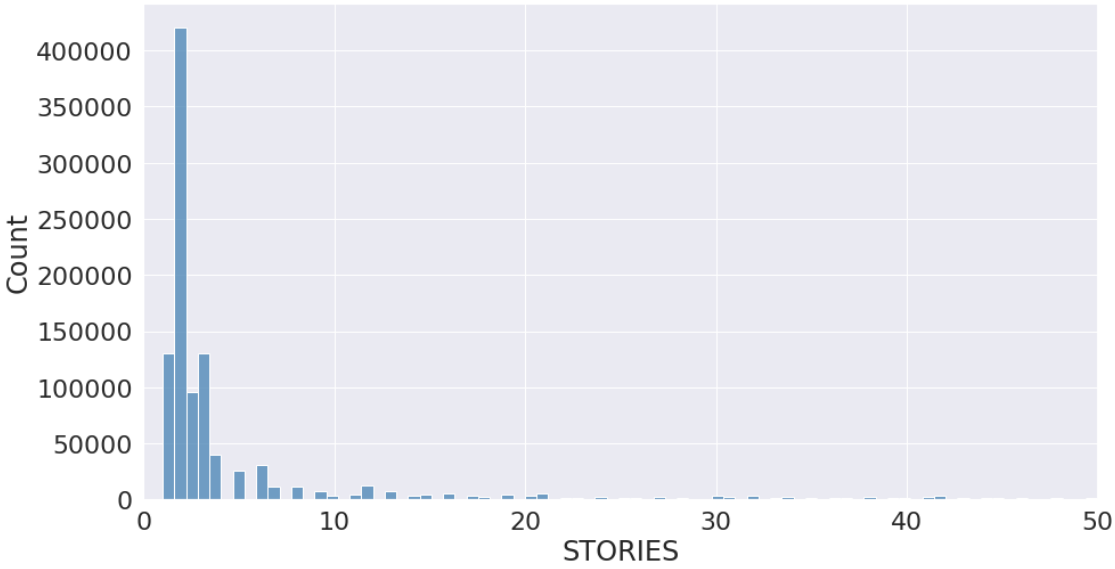


Figure A.11: Frequency Distribution of the *'STORIES'* field

Field 14: FULLVAL
Description: total market value of the land
Type: Numeric
Exclude outliers more than 2,000,000. Data in figure 1.3.2 is 96.31% populated.

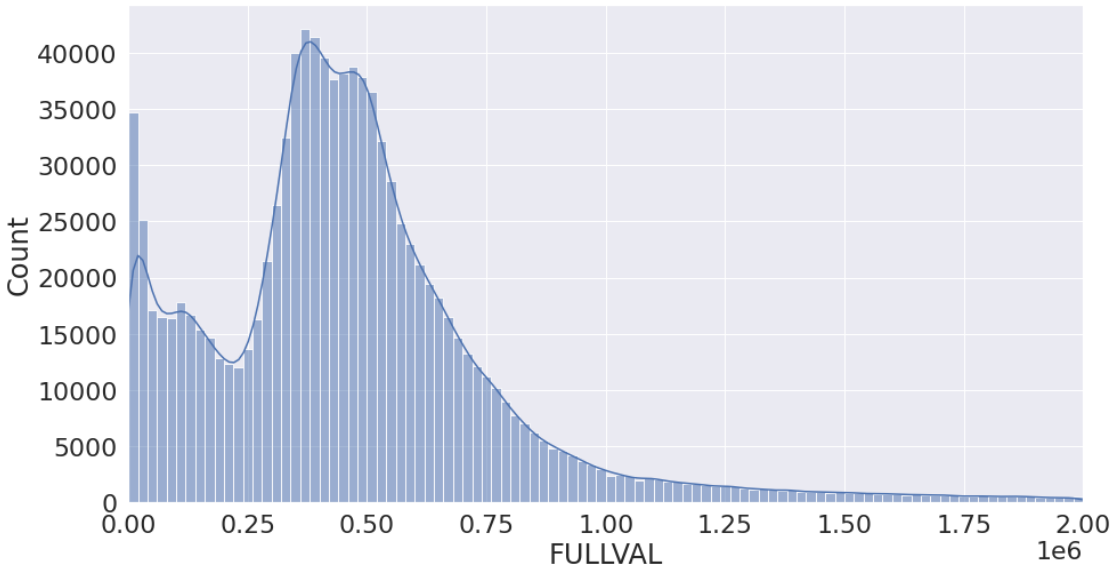


Figure A.12: Frequency Distribution of the *'FULLVAL'* field

Field 15: AVLAND
Description: Assessed land value
Type: Numeric
Exclude outliers more than 50,000. Data in figure 1.3.3 is 90.53% populated.

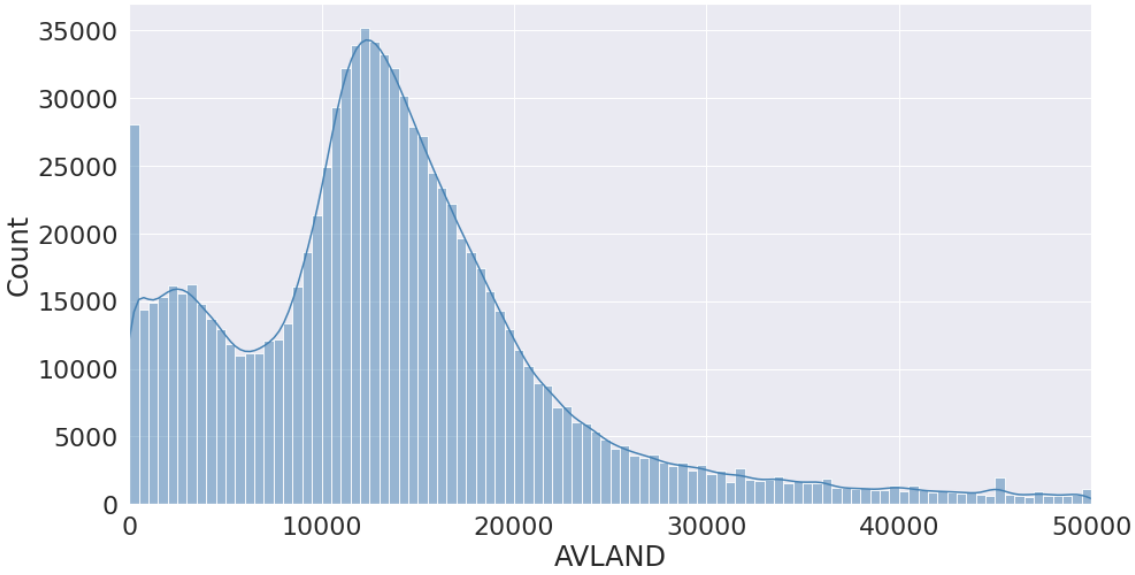


Figure A.13: Frequency Distribution of the *'AVLAND'* field

Field 16: AVTOT
Description: Assessed total value
Type: Numeric
Exclude outliers more than 100,000. Data in figure 1.3.4 is 86.05% populated.

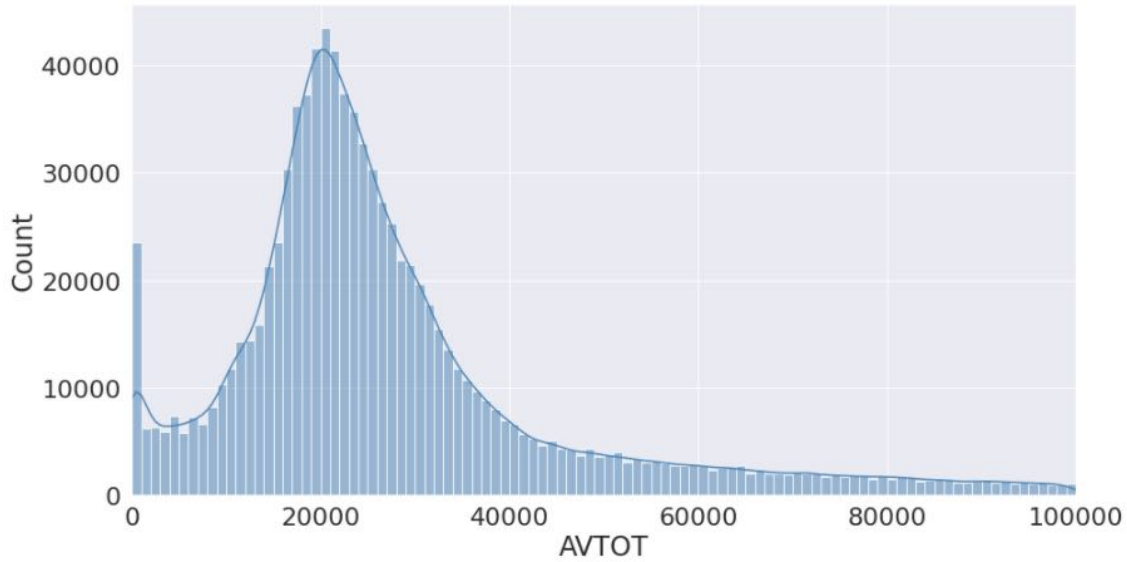


Figure A.14: Frequency Distribution of the *'AVTOT'* field

Field 17: EXLAND
Description: Exempt land value
Type: Numeric
Exclude outliers more than 500,000. Data in the plot below is 99.41% populated.

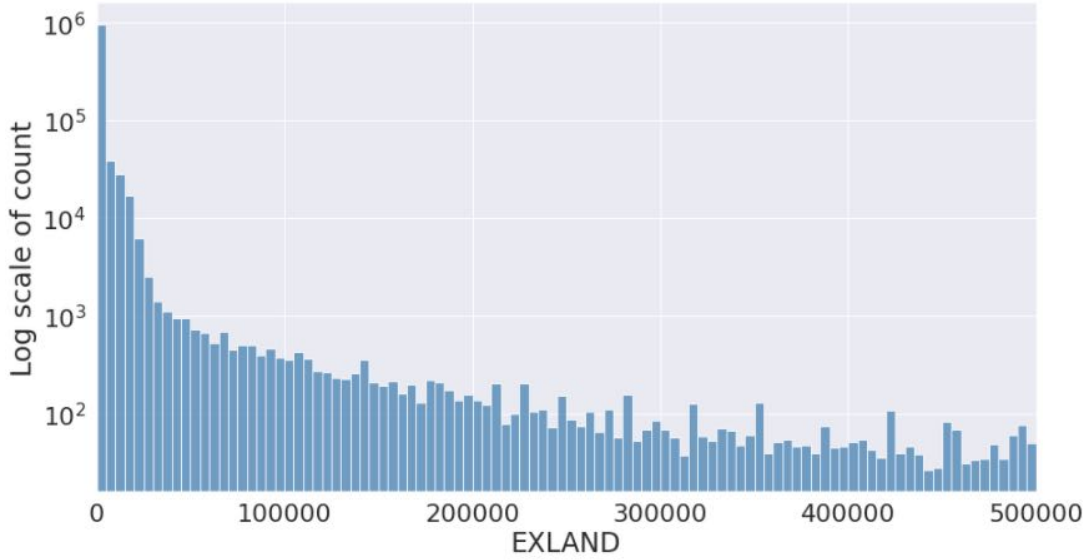


Figure A.15: Frequency Distribution of the *'EXLAND'* field

Field 18: EXTOT
Description: Exempt total value
Type: Numeric
Exclude outliers more than 100,000. Data in the plot below is 96.45% populated.

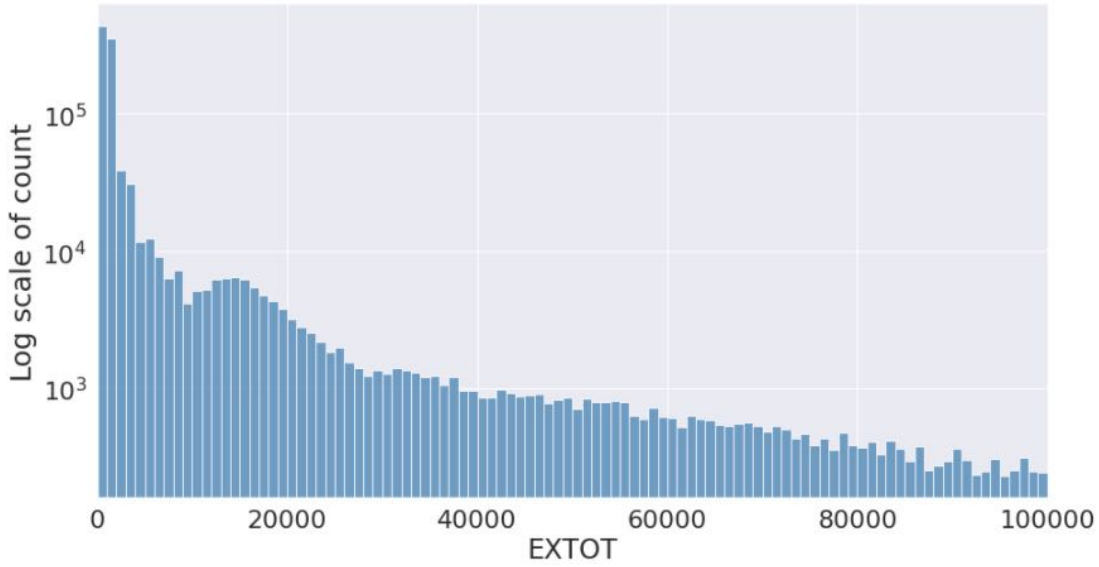


Figure A.16: Frequency Distribution of the *'EXTOT'* field

Field 19: EXCD1
Type: Categorical

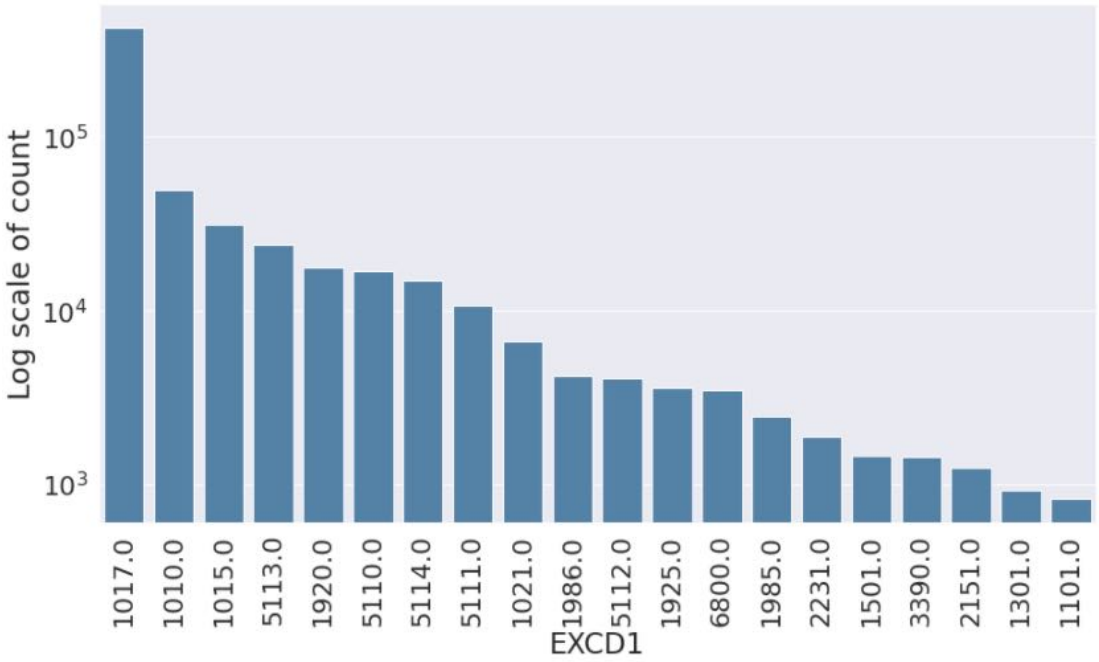


Figure A.17: Frequency Distribution of the *'EXCD1'* field

Field 20: STADDR
Description: Street name of the property
Type: Categorical

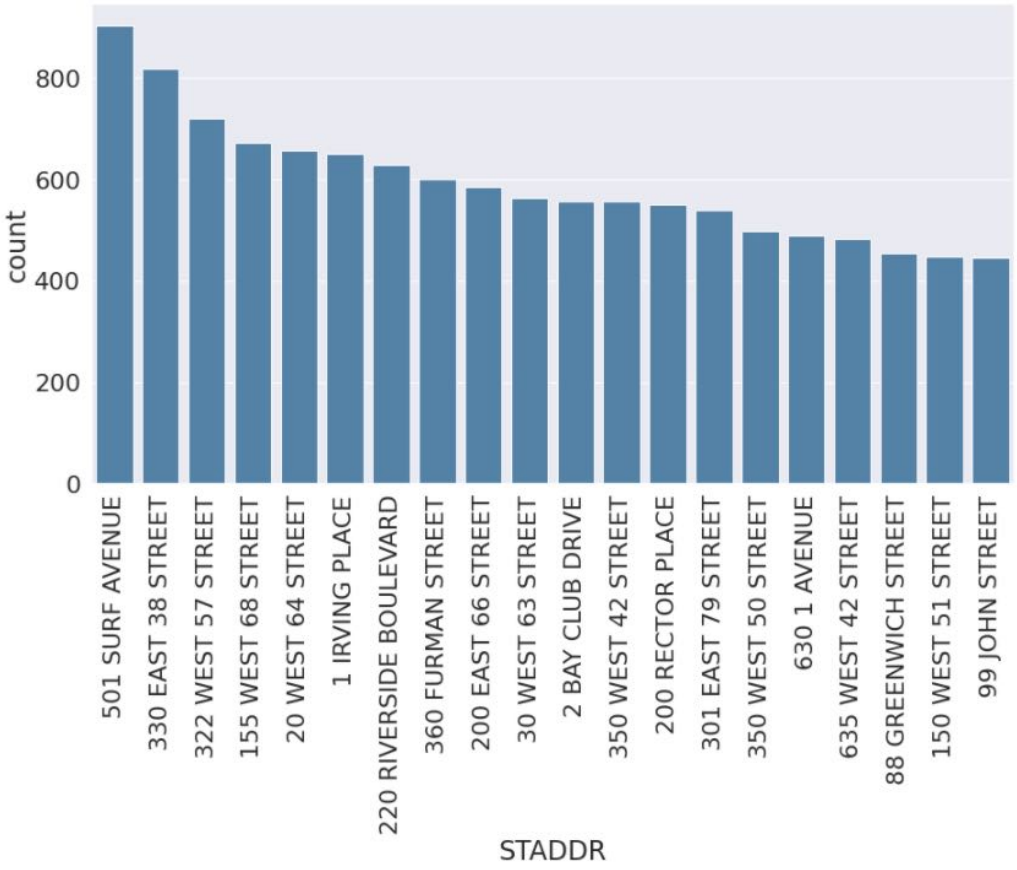


Figure A.18: Frequency Distribution of the *'STADDR'* field

Field 21: ZIP
Description: Postal Zip code of the property
Type: Categorical

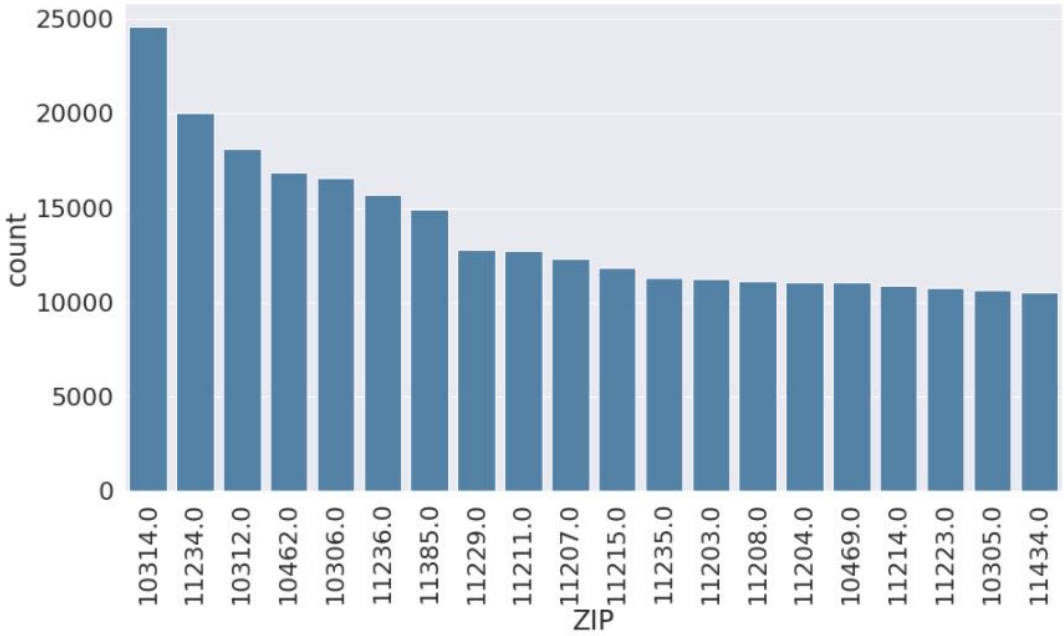


Figure A.19: Frequency Distribution of the *'ZIP'* field

Field 22: EXMPTCL
Description: Exempt Class used for fully exempt properties only
Type: Categorical

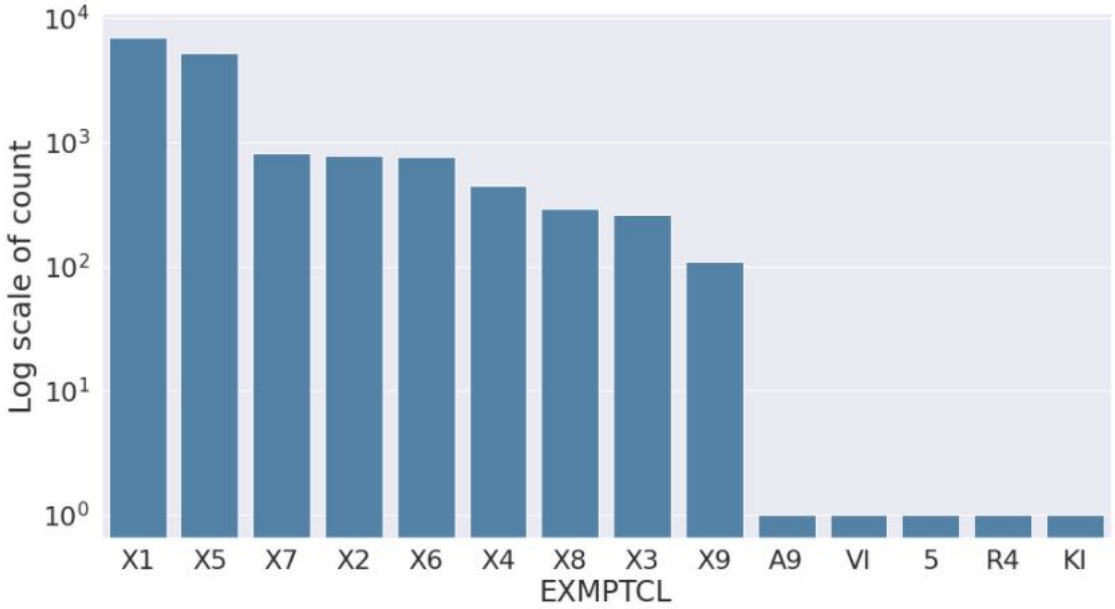


Figure A.20: Frequency Distribution of the *'EXMPTCL'* field

Field 23: BLDFRONT
Description: Building frontage in feet
Type: Numeric
There are no values more than 800; therefore data in figure 1.3.7 is 100% populated.

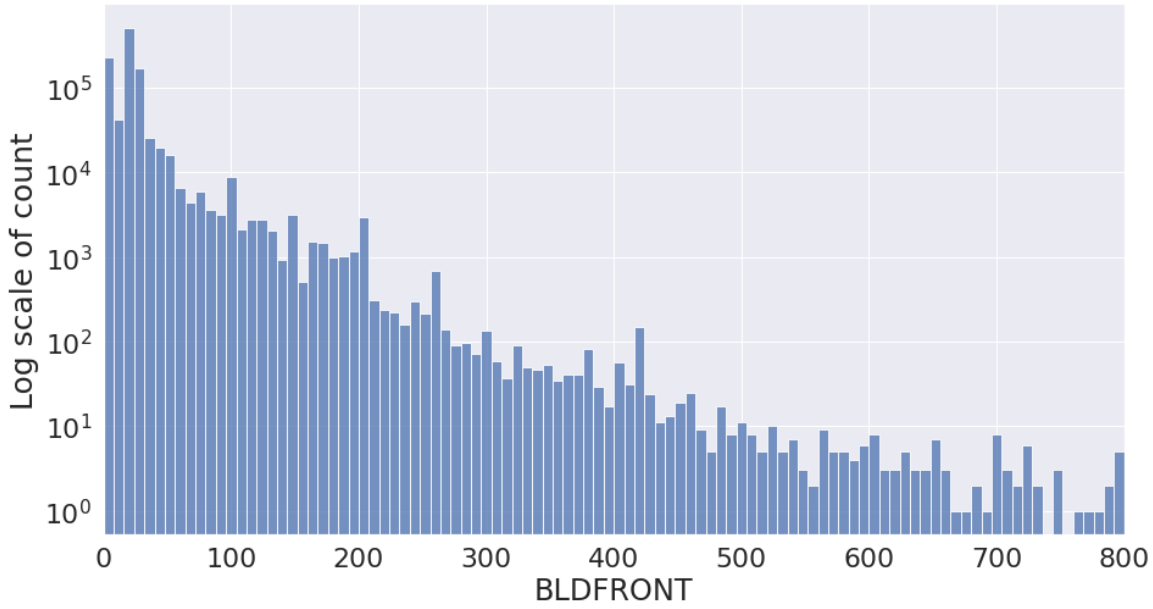


Figure A.21: Frequency Distribution of the *'BLDFRONT'* field

Field 24: BLDDEPTH
Description: Lot depth in feet
Type: Numeric
Exclude outliers more than 200. Data in figure 1.3.8 is 99.56% populated.
228815 records (21.36% of the entire dataset) have a value of 0.

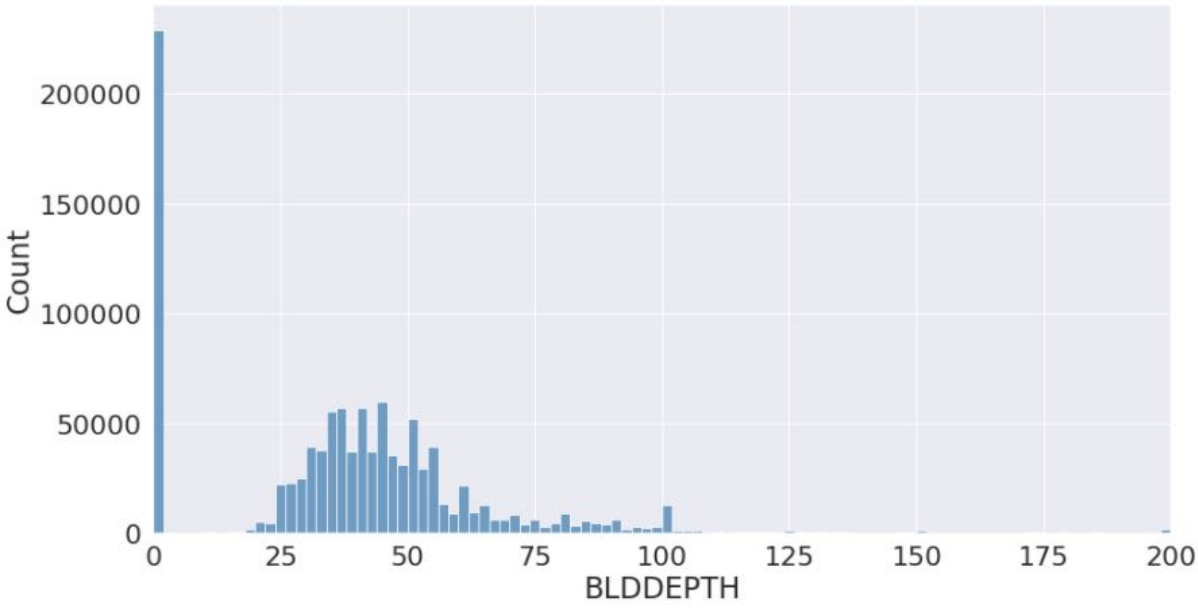


Figure A.22: Frequency Distribution of the *'BLDDEPTH'* field

Field 25: AVLAND2
Description: New market value of land
Type: Numeric
Exclude outliers more than 100,000. Data in the plot below is 81.34% populated.

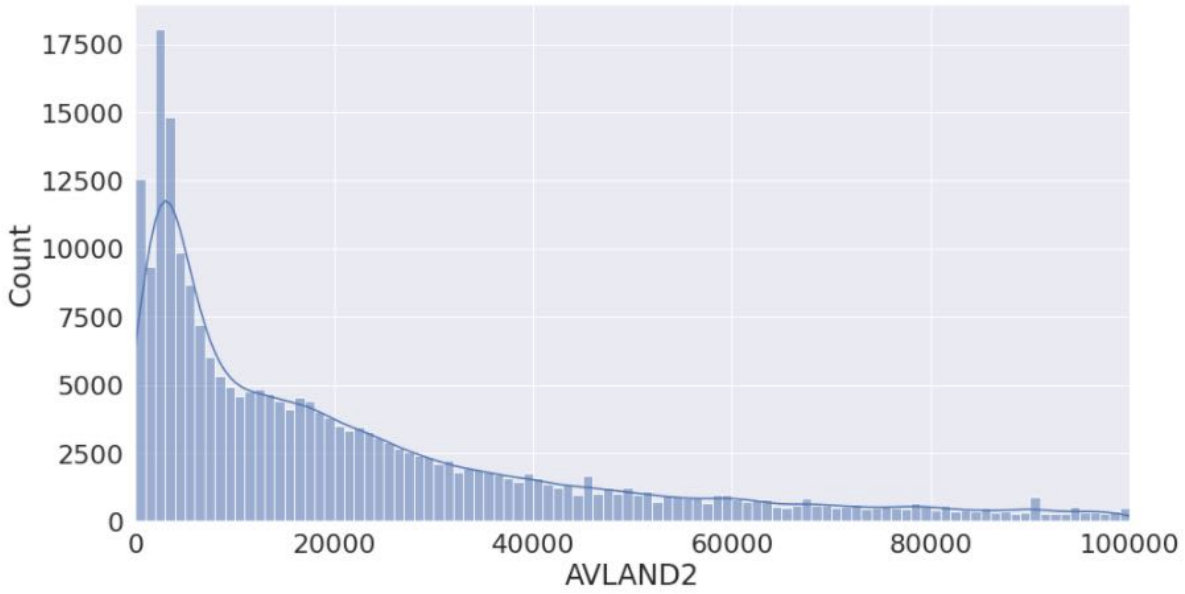


Figure A.23: Frequency Distribution of the *'AVLAND2'* field

Field 26: AVTOT2
Description: New total market value
Type: Numeric
Exclude outliers more than 1,000,000. Data in the plot below is 91.62% populated.

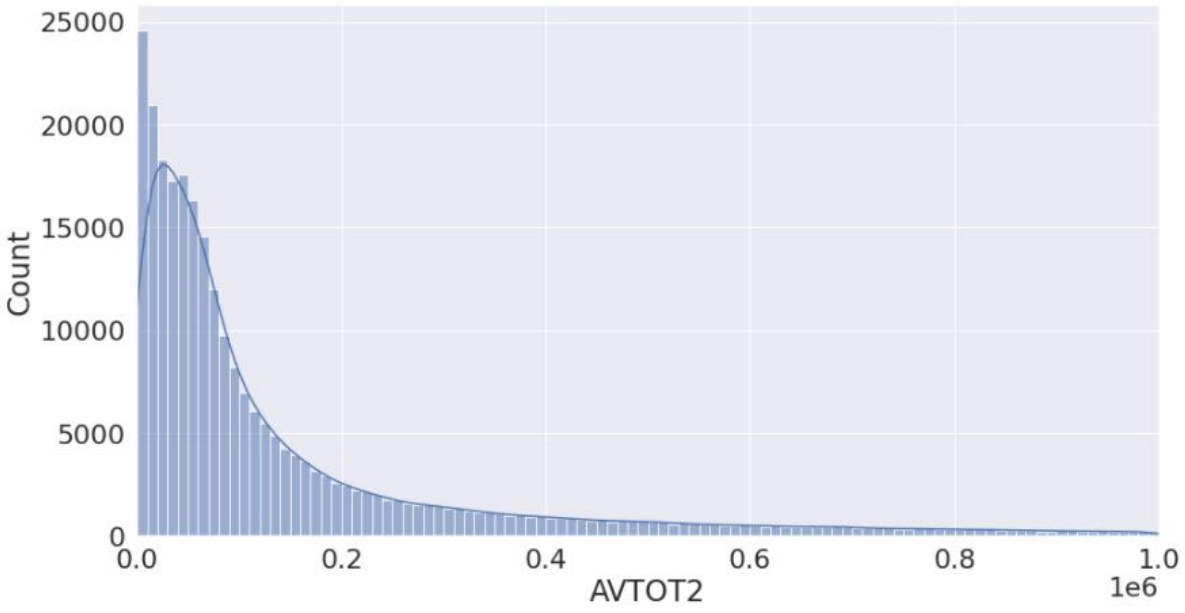


Figure A.24: Frequency Distribution of the *'AVTOT2'* field

Field 27: EXLAND2
Description: New exempt land value
Type: Numeric
Exclude outliers more than 50,000. Data in the plot below is 79.6% populated.

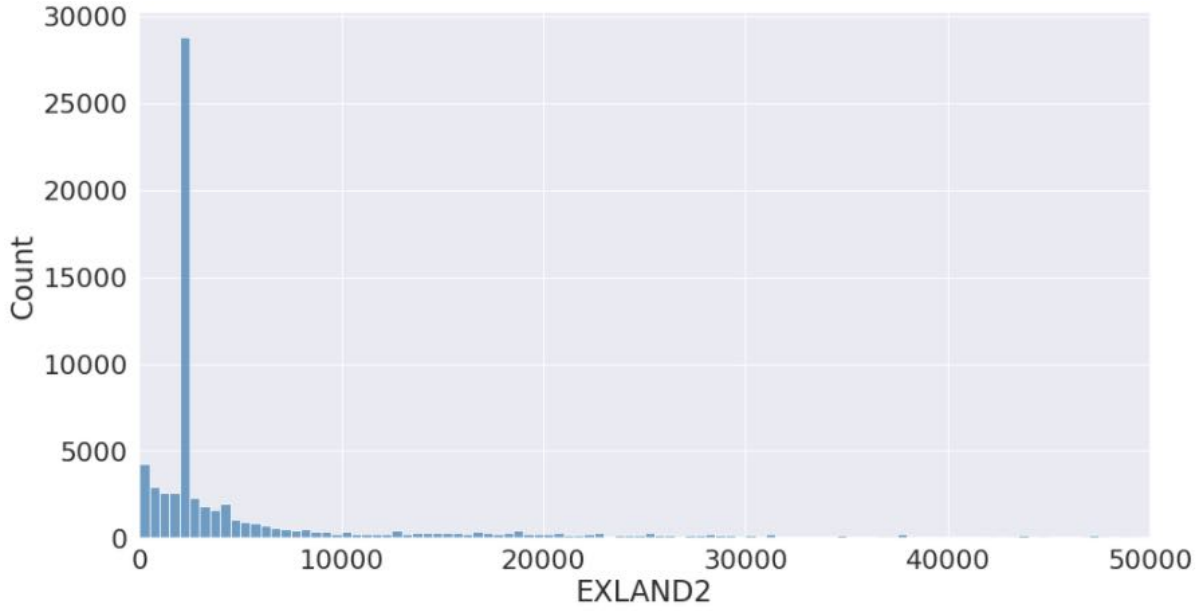


Figure A.25: Frequency Distribution of the *'EXLAND2'* field

Field 28: EXTOT2
Description: New exempt total value
Type: Numeric
Exclude outliers more than 50,000. Data in the plot below is 57.4% populated.

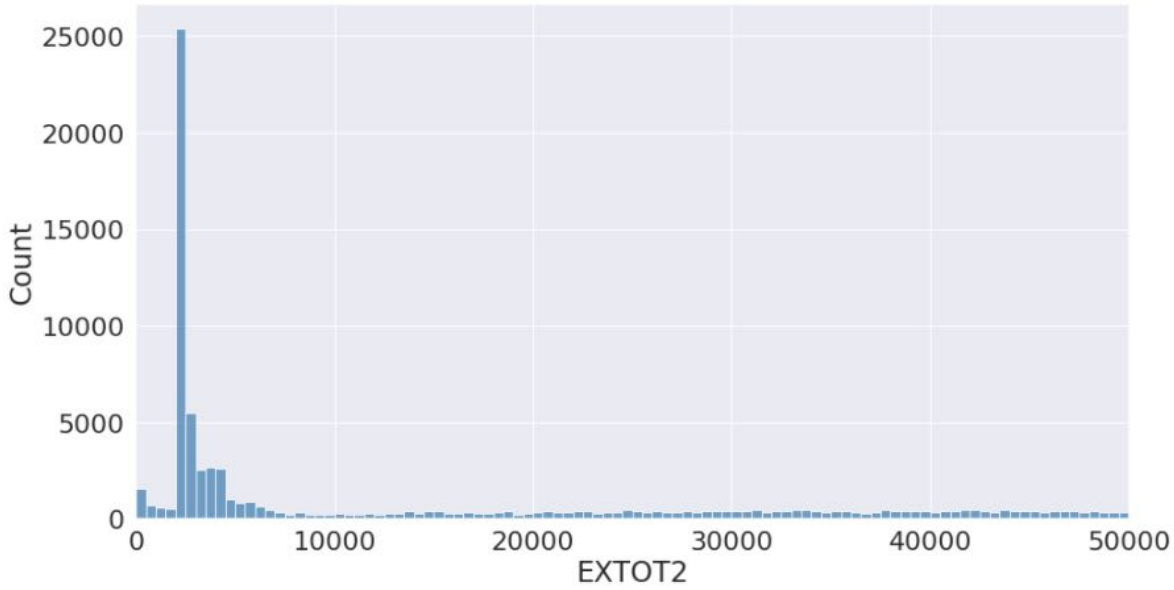


Figure A.26: Frequency Distribution of the *'EXTOT2'* field

Field 29: EXCD2
Type: Categorical

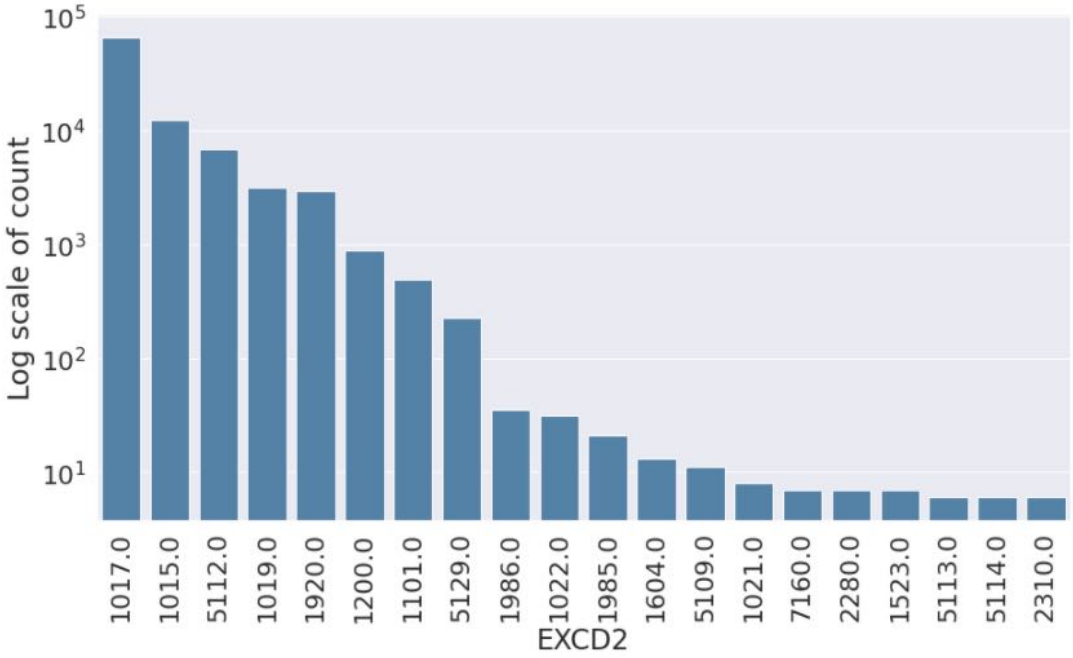


Figure A.27: Frequency Distribution of the *'EXCD2'* field

Field 30: PERIOD
Description: Change of period of file. All records have the same value of 'FINAL'
Type: Categorical

Field 31: YEAR
Description: Year of the file when updated. All records have the same value of '2010/11'
Type: Categorical

Field 32: VALTYPE
Description: The parcel's values reflected in another lot. All records have the same value of 'AC-TR'
Type: Categorical