NYC OpenData

Property Valuation and Assessment Data Fraud Analysis

University of Southern California

MS Business Analytics

DSO 562 Fraud Analytics

Spring 2020

Prepared By:

Aixuan Liu	**aixuanli@usc.edu**
Chao Wang	**wang724@usc.edu**
Cheng Shi	**shi005@usc.edu**
Chengjun Liu	**cliu4428@usc.edu**
Minglu Chi	**mingluch@usc.edu**
Shijie Xiang	**shijiex@usc.edu**
Yasmine Badawy	**ybadawy@usc.edu**

Table of Contents

1. Executive Summary

This report provides an analysis of The City of New York Property Valuation and Assessment Data for fraud detection using unsupervised machine learning methods. The tool used is Python, and the main method applied for analysis was Principal Component Analysis and Autoencoder.

Top 10 records were examined in the end for common patterns of abnormality. Common features of these records were small building/lot sizes and high values. Among the top 10 records, 4 records are government-owned properties with either a high value or a small building front/depth. Another record shows that being close to an airport also indicates extremely high value. The remaining 5 records are identified as possible frauds.

The original dataset was a record set of more than 1 million properties across the city of New York. The data was modeled by the City of New York to provide information on the property sizes, values, owners, building classes, tax classes, etc. The process of analysis included: cleaning data, building expert variables, dimensionality reduction, calculating fraud score, and inspecting potential fraud records.

2. Description of Data

2.1 Data Description

Dataset Name: Property Valuation and Assessment Data

Dataset Source: NYC City Government, Department of Finance

Time Period: 11/2010

Number of Fields: 32

Number of Records: 1,070,994

2.2 Summary

2.2.1 Categorical Variables

Name	# of records that have a value	% populated	# of unique values	Most common field value
B	1070994	100.0	5	4
TAXCLASS	1070994	100.0	11	1
ZIP	1041104	97.2	197	10314

2.2.2 Numeric Variables

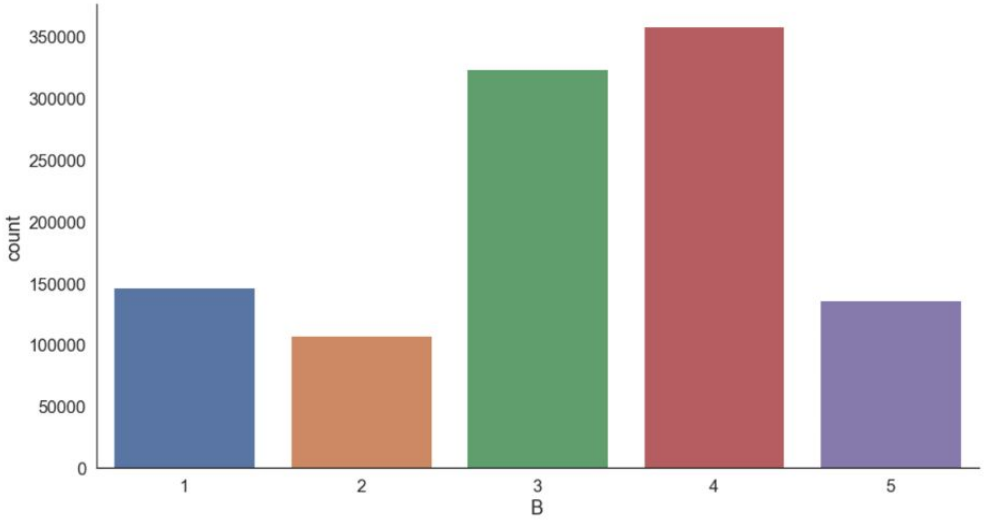
	# records that have a value	% populated	# unique values	# records with value zero	Mean	Standard Deviation	Min	Max
LTFRONT	1070994	100	1297	169108	36.6	74.0	0	9999
LTDEPTH	1070994	100	1370	170128	88.9	76.4	0	9999
STORIES	1014730	94.7	112	0	5.0	8.4	1	119
FULLVAL	1070994	100	109324	13007	874264.5	11582431.0	0	6150000000
AVLAND	1070994	100	70921	13009	85067.9	4057260.1	0	2668500000
AVTOT	1070994	100	112914	13007	227238.2	6877529.3	0	4668308947
BLDFRONT	1070994	100	612	228815	23.0	35.6	0	7575

BLDDEPTH	1070994	100	621	228853	39.9	42.7	0	9393

2.3 Data Exploration

2.3.1 Field name: B

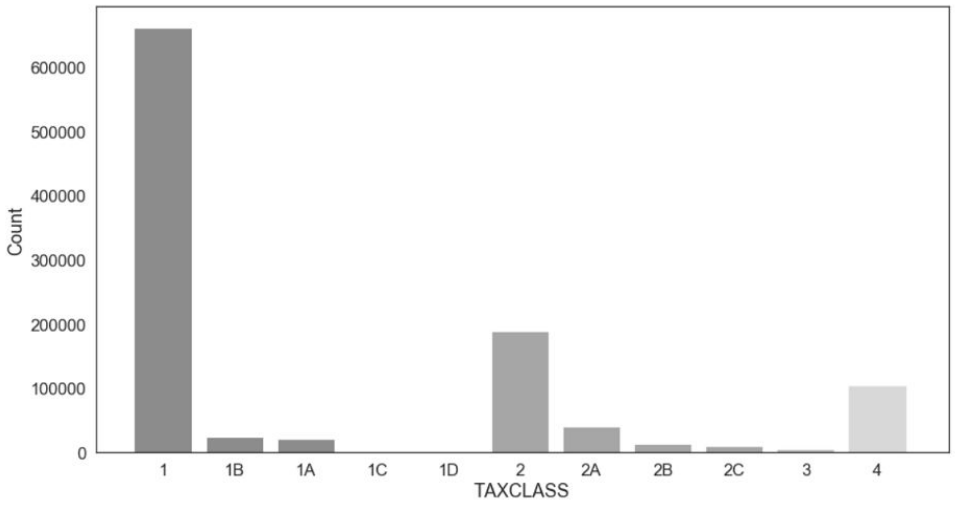
Description: Borough code.



2.3.2 Field name: TAXCLASS

Description: Current Property Tax Class Code (NYS Classification).

There were 11 unique values in this field and the distribution is shown below.



2.3.2 Field name: ZIP

Description: Postal Zip code of the property.

Top 10 Field Values	
ZIP	**COUNT**
10314	24606
11234	20001
10312	18127
10462	16905
10306	16578
11236	15678
11385	14921
11229	12793

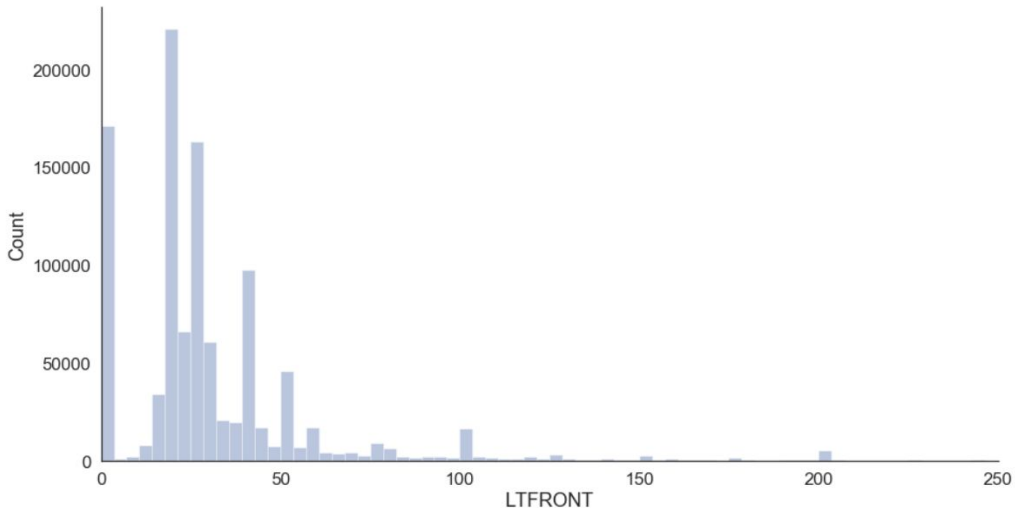
11211	12710
11207	12293

2.3.4 Field name: LTFRONT

Description: Lot frontage in feet.

Outliers: LTFRONT > 250

Histogram: Covering 99.06% of populated records.

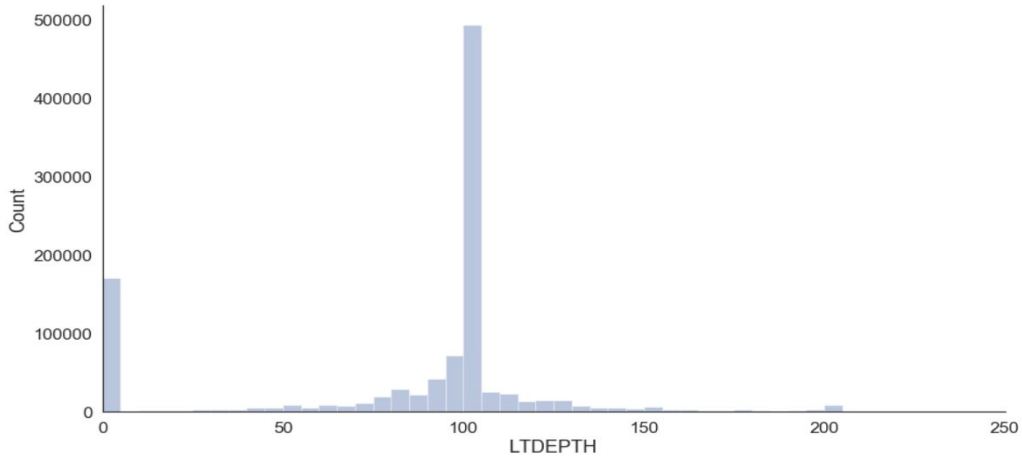


2.3.5 Field name: LTDEPTH

Description: Lot depth in feet.

Outliers: LTDEPTH > 100

Histogram: Covering 97.40 % populated records.

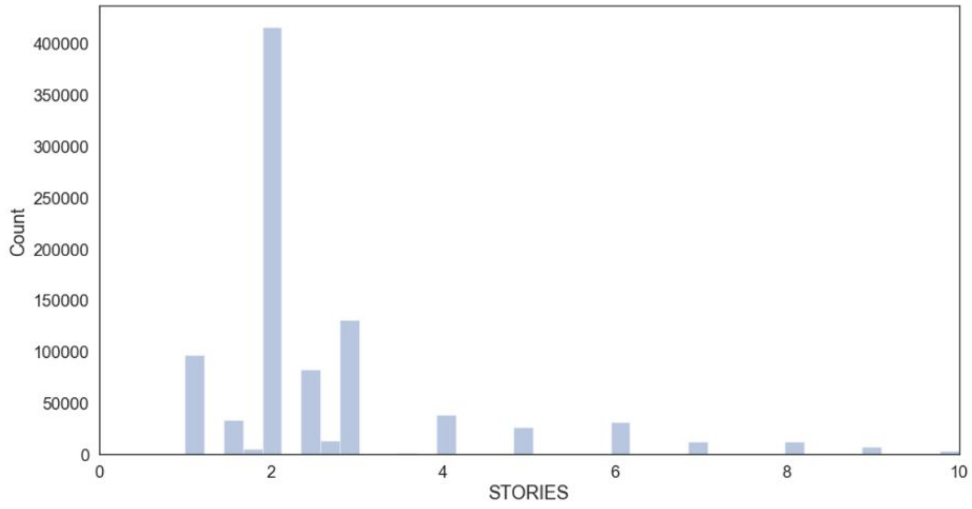


2.3.6 Field name: STORIES

Description: The number of stories for the building.

Outliers: STORIES > 10

Histogram: Covering 89.62% populated records.

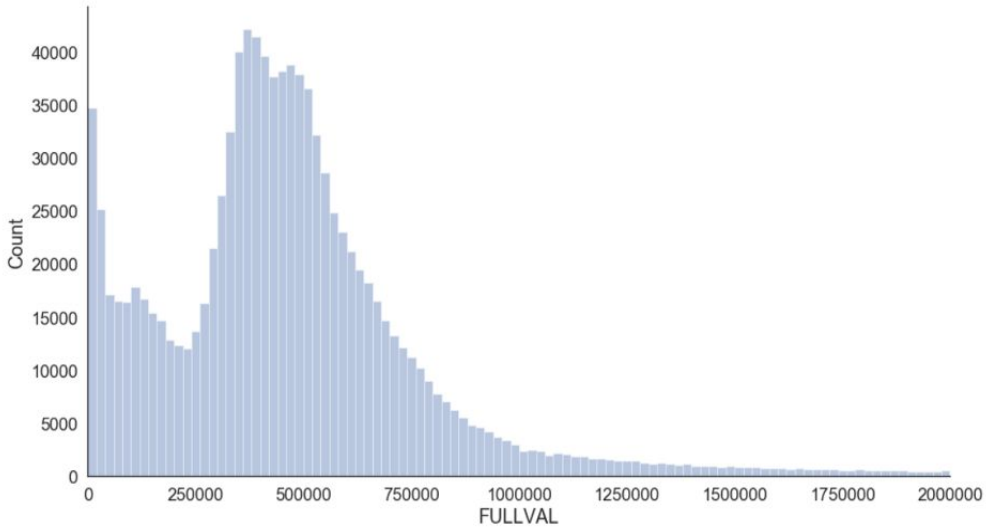


2.3.7 Field name: FULLVAL

Description: Total market value of property.

Outliers: FULLVAL > 2,000,000

Histogram: Covering 96.31% of populated records.

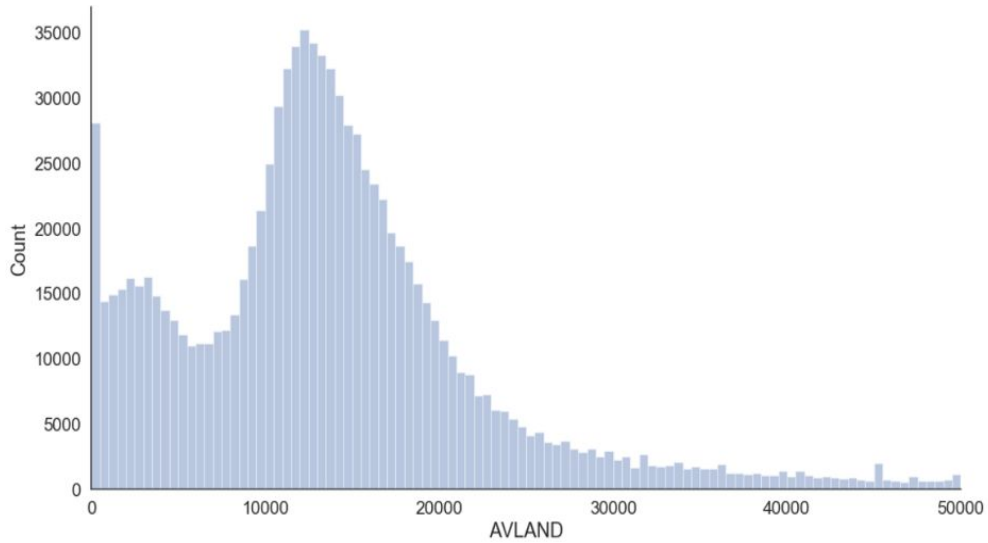


2.3.8 Field name: AVLAND

Description: Actual Market value of the land.

Outliers: AVLAND > 50,000

Histogram: Covering 90.53% of populated records.

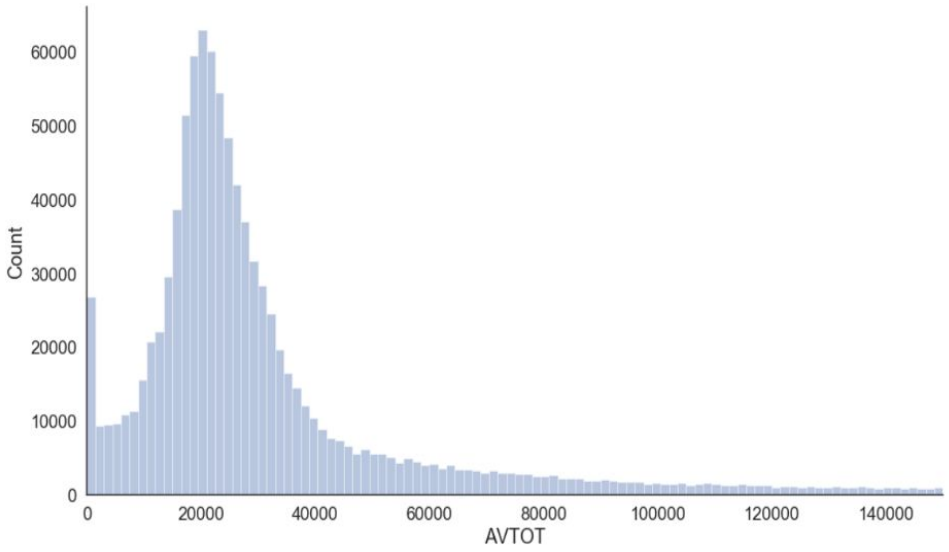


2.3.9 Field name: AVTOT

Description: Actual total market value.

Outliers: AVTOT > 150,000.

Histogram: Covering 89.58% of populated records.

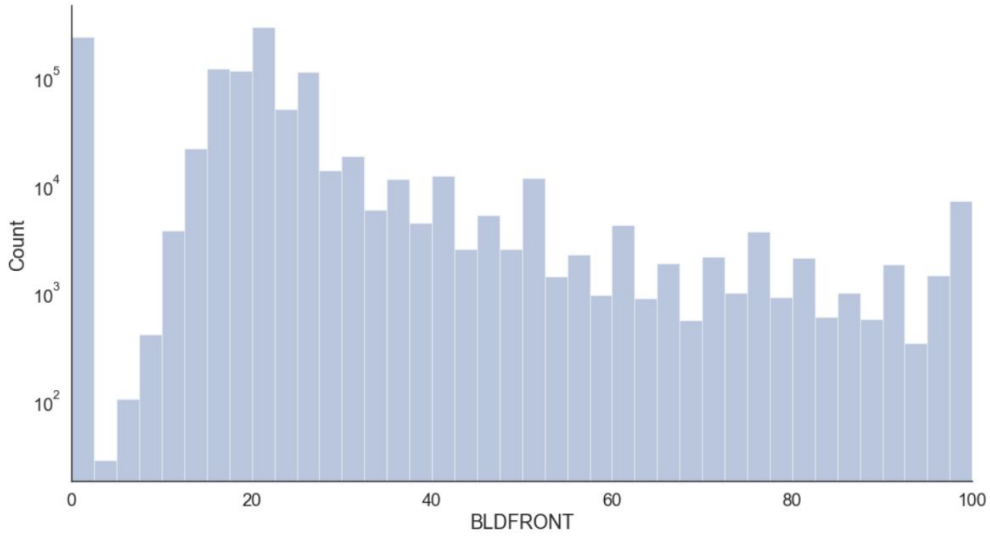


2.3.10 Field name: BLDFRONT

Description: Building Frontage in feet.

Outliers: BLDFRONT > 100

Histogram: Covering 97.37% populated records.

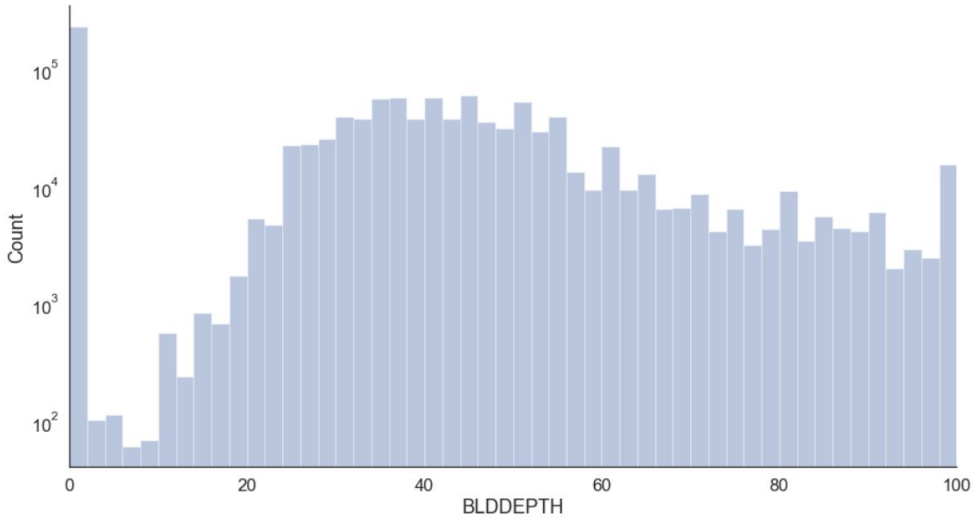


2.3.11 Field name: BLDDEPTH

Description: Building Depth in feet.

Outliers: BLDDEPTH > 100

Histogram: Covering 97.40 % populated records.



3. Data Cleaning

3.1 Process taken for handling missing field values

In order to fill in missing values for a particular record that has missing data, we categorized the record into subsets, and then used the median or the most common value of the subset field over the relevant missing records.

3.2 Itemized fields with empty values

3.2.1 ZIP

The ZIP field was aggregated by the B field. Since ZIP is a categorical field, we filled the missing values in ZIP with the most frequent value of that group.

3.2.2 STORIES:

We went over three conditions of aggregations to fill in the missing values in the STORIES field. In each condition, we filled in the missing values of stories with the median of that group.

- Condition 1: We aggregated by B, TAXCLASS and ZIP.
  If the number of records in that group is smaller than 5, we move on to condition 2.
- Condition 2: We aggregated by B and TAXCLASS.
  If the number of records in that group is smaller than 5, we move on to condition 3.
- Condition 3: We aggregated by TAXCLASS only.

3.2.3 FULLVAL, AVLAND, AVTOT

We took the same steps used above for STORIES to fill in missing values in FULLVAL, AVLAND and AVTOT. For these columns, we treated value 0 as missing values. We

first aggregated the dataset by TAXCLASS and ZIP. Then we replaced zeros with the median of that group in these three fields. If the number of records in a particular group was smaller than 5, we aggregated the data by TAXCLASS only, and replaced zeros with the median of that group.

3.2.4 LTFRONT & LTDEPTH

We took the same steps to fill in missing values in LTFRONT and LTDEPTH. We first aggregated the dataset by ZIP, B and TAXCLASS, and replaced missing values with the median of that group.

If there were groups with less than 5 complete records, we tried filling the missing values in these groups by the median of a broadened group aggregated by ZIP and B. If did not meet the minimum threshold, we aggregate by ZIP only or B only, until every record with missing LTFRONT and/or LTDEPTH fell into a group with more than 5 complete records.

3.2.5 BLDFRONT & BLDDEPTH

We applied the same logic and method as LTFRONT and LTDEPTH to BLDFRONT and BLDDEPTH.

4. Variable Creation

4.1 Create 3 sizes variables

We first created three new variables representing size of properties from five existing variables, which are:

a. lotarea = LTFRONT * LTDEPTH

b. bldarea = BLDFRONT * BLDDEPTH
c. bldvol = bldarea * STORIES

4.2 Calculate 9 building volume variables

We then created 9 variables representing the value per unit area for each property. Since there were not expert opinions on which measure might be the most representative, we tried 9 combinations with each of the 3 values normalized by each of these 3 sizes:

a. r1=FULLVAL/lotarea
b. r2=FULLVAL/bldarea
c. r3=FULLVAL/bldvol
d. r4=AVLAND/lotarea
e. r5=AVLAND/bldarea
f. r6=AVLAND/bldvol
g. r7=AVTOT/lotarea
h. r8=AVTOT/bldarea
i. r9=AVTOT/bldvol

4.3 Group values of each variable to get 45 columns

After creating the 9 variables above, we grouped the values of the 9 variables by these 5 groups respectively to get new 45 columns: ZIP5, ZIP3, TAXCLASS, B, and all. For each column, we calculated the means of the 5 groupings, and divided each value by the means of the record's category, which gave us a value that represented the relative "outlierness" (i.e. number of standard deviations) of each value inside a group.

5. Dimensionality Reduction

After creating the 45 variables of interest, we performed a Principal Component Analysis (PCA) to remove correlations and reduce dimensions, which in turn yielded a better balance between model error and complexity. Specifically, we took the following steps:

5.1 Conduct Z-scaling for all columns to transform model inputs to the same scale

By conducting Z-scaling, we made sure that the PCA model did not confuse differences in scales with differences in variance. After scaling, all data became a cloud around the origin with the same scaling for all dimensions, and we were able to calculate the fraud score as the distance from the origin. In this way we used PCA to identify which components truly brought the most variance to the data.

5.2 Perform PCA with all columns

A PCA model first finds the most dominant direction in the data that has the most variance, and then finds the succeeding dominant directions one at a time on the condition that each one is perpendicular to the previous dominant direction. By tilting the dimensions, columns of high linear correlation were combined by the PCA model to form principal components, which removed linear correlation and yields components of higher independence. We set the PCA model to form 15 components.

5.3 Remove components that only explain a limited amount of variance

By sorting the PCs' magnitude in descending order and checking the explained variance of each PC, we saw that the first five components explained about 90% of the variance. Therefore, we removed the last 10 components to reduce dimensionality.

5.4 Conduct Z-scaling again

We conducted Z-scaling again on the remaining components to ensure that all inputs were of equal importance. Since it was hard to interpret the meaning of each principal component, and we did not hold assumptions about each input's contribution to a record's abnormality, we conducted Z-scaling again to the remaining components to put all components to the same scale for further modeling. This step was also necessary to prepare data for the calculation of the first fraud score, the formula of which is:

$$\left(\sum_{i=1}^{M} |x_i|^n \right)^{1/n}$$

Using this formula to calculate distance, which determined when the choice of n was bigger than 1, we put more weight on PCs that had larger scales. Thus, to make sure that we treated all PCs equally, we Z-scaled again before calculating the fraud scores.

6. Algorithms

6.1 Calculating Fraud Score One: Heuristic Use of Z-Scores of Principal Components

6.1.1 Why can Z-scores be used for anomaly detection?

Z-scores anchor data of different scales to the same standard normal distribution with a mean of zero and standard deviation of one. Outliers usually are extreme in some dimensions. In this way, some dimensions of outliers will be likely to take relatively extreme values on the standard normal curve. To measure the general distance of dimensions from normality, we needed to leverage the average deviation of all

dimensions and highlight extreme deviations in certain dimensions by combining the Z-scores of the principal components that remain in the dataset using distance from origin point in n-dimensional space. When n was large enough, we were able to highlight extreme deviation. Considering leveraging other features, we specified n as 2 to calculate distance under 2-dimensional space. This enabled us to be able to identify records that displayed higher abnormality across the most important dimensions in the data by observing whether records showed large distance from the origin point.

6.1.2 Steps to calculate the fraud score with Z-scores

The formula of the first fraud score was:

$$\left(\sum_{i=1}^{M} |x_i|^n\right)^{1/n}$$

After the second Z-scaling, we took all PCs to the power of 2 to remove negative signs and calculate an average distance, added up all Z-scores of the principal components for each individual record, and then took the square root of the summed Z-scores. When n was high, extreme values carried even higher weight in the summed score. Therefore, we chose a lower n to weigh each dimension equally. The resulting positive number reflected the distance of a record from the origin. The higher the number was, the higher possibility that the record was an outlier.

6.2 Calculating Fraud Score Two: Anomaly Detection by Autoencoder

6.2.1 How is Autoencoder applied to find abnormal records?

Autoencoder uses neural networks to learn the common patterns of the dataset, compress and extract the main features of the dataset into fewer dimensions and nodes in inner layer of neural network, and then reverse and expand main features into initial dimensions and restore the common patterns of records. By going through the process of reducing dimensions and restoring data, an autoencoder keeps the main patterns and

features of the data input and removes the noises that cause deviation from normal estimation.

If a record was abnormal, we were sure it would not follow common patterns, so the restored value of an abnormal record would be quite different from the original record. We calculated how far a restored record was from its original value, and the records with the largest deviation were likely to be outliers.

6.2.2 Steps to calculate the fraud score with Autoencoder

We first trained the Autoencoder model on the entire dataset with original Z-scaled five principal components, and then we used the Autoencoder to reproduce the dataset. Now we had two datasets, the original dataset and the reproduced dataset, and we measured the reproduction error as abnormality for specific records between the two datasets, thus getting desired fraud score. We calculated the distance of each Z-scaled principal component by applying the absolute value of the difference in each Z-score. After that, for each record, we took the difference of each Z-score to the power of one specific number of n, summed them up, and took the square root of the summed score, which resulted in the second fraud score:

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

Usually n is set at 2 or 3 to simulate the measurement of distance in 3-dimension or 4-dimension space. For our measurement of fraud score, here we specified n as 2.
The positive fraud score reflects the distance of one record from the reproduced dataset to the original dataset. The higher the number is, the higher possibility that the record is an outlier.

6.3 Final Fraud Score: Leveraging Two Fraud Scores

6.3.1 Why not use the two original scores?

At this point, we had two fraud scores calculated with two different methods, one being the summed distance of the Z-scaled principal components from the origin, and one being the difference between the original Z-scaled principal components and the output of an Autoencoder. Due to the difference in methods and thus the difference in the meaning of the outputs, we could not directly combine the two scores numerically for the final fraud score. Here we used quantile binning as a way to bring the two scores to the same ground.

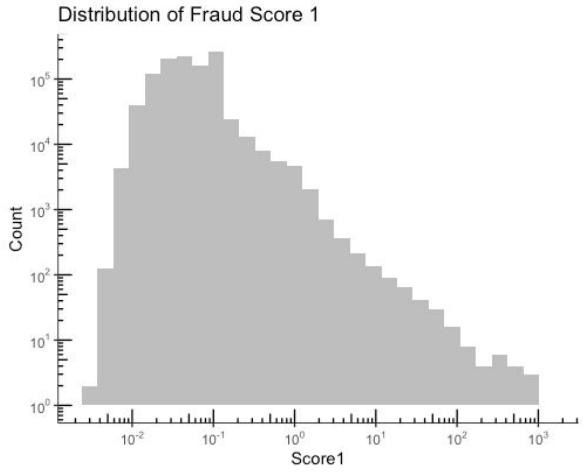
6.3.2 Rank ordering and the final score

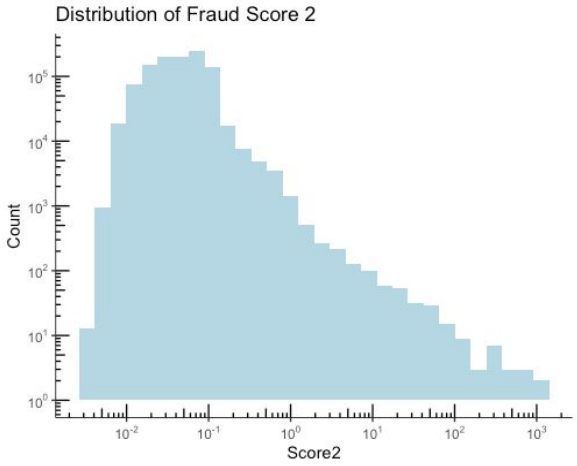
To conduct quantile binning, we chose a bin size of one, which assigned ranking numbers to records. We sorted each record by the two fraud scores in descending order and replaced the scores with respective ranking orders. Then we took an average of the two rankings as the final scores, and took 10 records, ranked 1 to 10, with largest final scores as highly likely to be fraud. The smaller the rank number, the higher the indication of elevated abnormality.

7. Results

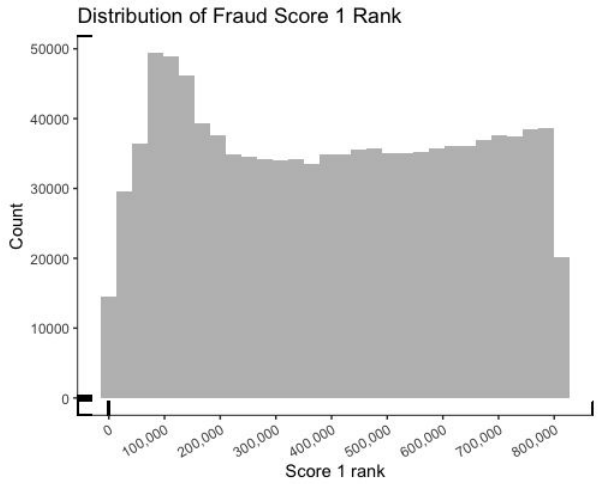
7.1 Score distribution

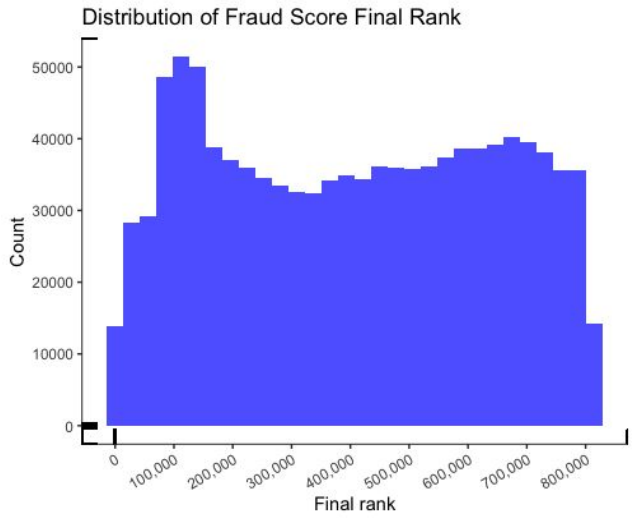
7.1.1 Distribution of Fraud Score 1 and 2





7.1.2 Distribution of the rankings of Fraud Score 1, 2, and the final Score





7.2 Top 10 records

The top 10 abnormal records' RECORD values were:

Ranking	RECORD	Ranking	RECORD
1	632816	6	85886
2	565392	7	585439
3	1067360	8	585120
4	917942	9	565398
5	585118	10	920628

7.2.1 Rank 1 record

The building front, building depth and story of this record with highest fraud rank were all 1 depth and story. While it is not abnormal for the story number to be 1, building front and depth of 1 is abnormal.

According to the street name (86-55 Broadway), Zip code (11373) and building class (D9), we identified the property to be The Elm East, which is an apartment building.[1] The official website of The Elm Rentals also showed the same information.[2] Then, from StreetEasy, a New York real estate website verified by The Elm Rentals on its official website, we found that The Elm East was built in 2015.[3] The trademark of The Elm East was registered in 2012.[4] The owner of the property is 864163 Realty, LLC, which filed in

[1] Nyc Building Classes & Building Classification https://www.propertyshark.com/mason/text/nyc_building_class.html
[2] The Elm East | http://elmrentals.com/the-elm-east/#residences
[3] https://streeteasy.com/building/the-elm-east
[4] the Elm East Trademark Of Jerry Pi Serial Number 85366624

2006.[5] However, the time of raw dataset was 2010, which is after the time when the company was filed but before the time when the building was registered. We could not find any information within this time period. So we did not know if there was another building in the same location in 2010.

In terms of the value, the total market value of the property (FULLVAL) was $2.9 million. It is abnormal for a property to have $ 2.9 million market value when both building front and building depth are 1 feet. We assumed that, no matter whether there is another building in the same spot in 2010, the fact that property value and building size didn't correspond to each other indicated a high possibility of fraud.

7.2.2 Rank 2 record

The property of this record also had very high property value and land value. The lot front and lot depth were both over 100 feet, whereas its building front and building depth were 19 and 42 feet respectively. Besides, the property owner was the US government. The building class also indicated that the building was probably for government or public use, as building class V9 indicating miscellaneous use (Department of Real Estate and Other Public Places).[6]

The property was located on Flatbush Avenue. According to Wikipedia, "Flatbush Avenue is a major avenue in the New York City Borough of Brooklyn".[7] Based on all this information, we came to believe that the property is a government-owned building with a large land in central Brooklyn. The good location probably accounted for the high value of the property.

Corporationwiki - https://trademarks.corporationwiki.com/marks/the-elm-east/85366624/
[5] 864163 Realty, Llc in Woodside Ny - Company Profile
Corporationwiki - https://www.corporationwiki.com/p/2q7wq5/864163-realty-llc
[6] Nyc Building Classes & Building Classification
https://www.propertyshark.com/mason/text/nyc_building_class.html
[7] Flatbush Avenue
https://en.wikipedia.org/wiki/Flatbush_Avenue

7.2.3 Rank 3 record

For the property of this record, both its lot front and lot depth were 1 foot, which are abnormal in reality. It didn't have owner information.

The building class (B2) indicated that it was a frame-built two-family dwelling.[8] The tax class (T1) also showed that it was among most residential properties of up to three units.[9] Using the street name of the property, we found that "20 Emily Ct, Staten Island, NY is a multiple occupancy home that contains 2,850 sq ft and was built in 1999. This home last sold for $234,000 in December 1999." [10] Since the property was last sold in 1999, we assumed that it was still a multifamily house in 2010.

But being a multifamily house still cannot account for the abnormal lot data. We also found several news reports that show some multifamily houses are associated with fraud. WSJ reports on a "Ponzi Scheme-like" scam about multifamily mortgage fraud.[11] Therefore, the property of this record was highly indicated to be fraud.

7.2.4 Rank 4 record

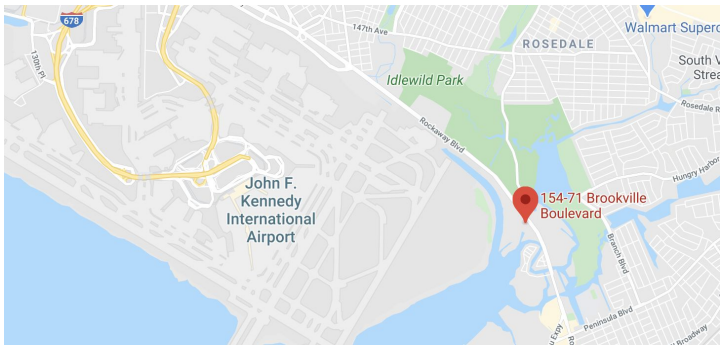
The property was of high land value, as its actual market value of the land (AVLAND) is approximately 5 times of total market value of the property (FULLVAL). The building class (T1) indicated that the property was for airport, airfield, terminal use.[12] Then we

---

[8] Class B Buildings - Two Family Dwellings
https://www.propertyshark.com/info/class-b-buildings-two-family-dwellings/#b2
[9] https://www1.nyc.gov/site/finance/taxes/definitions-of-property-assessment-terms.page
[10] 20 Emily Ct, Staten Island, Ny 10307
Zillow, Inc - https://www.zillow.com/homedetails/20-Emily-Ct-Staten-Island-NY-10307/58579273_zpid/
[11] Sec Accuses Major U.s. Landlord Of Running 'ponzi Scheme-like' Scam
Cezary Podkul -
https://www.wsj.com/articles/sec-accuses-major-u-s-landlord-with-running-ponzi-scheme-like-scam-11558555886
[12] Nyc Building Classes & Building Classification
https://www.propertyshark.com/mason/text/nyc_building_class.html

used Google Maps to search the location by the street name. The property was near the JFK Airport.

According to an article published on Mortgage Introducer, “Homes near airports outside London cost more despite the noise and air pollution, research from free online estate agent OwnerSellers has found.”[13] The record was detected as fraud by our model, but the property location played an important role and accounted for its high value.



7.2.5 Rank 5 record

The owner of the property was “New York City Economi” according to the record, which indicated a probably of reference to an economy-related department. It was located on 28-10 Queens Plaza South. The building class (O3) indicated that the property was an office building with ten stories and over.[14] But the abnormal part of this record was that both building front and building depth were 1 foot.

According to Queens Plaza (Queens)’s wikipedia information, “In 2005, the U.S. Congress approved a measure to demolish the municipal parking lot and turn it into a 1.5 acres (6,100 m2) park.”[15] According to an article published in Cision in 2019, “Tishman Speyer is currently Long Island City's most prolific developer of both office and

[13] House Prices Higher Near Airports
Ryan Bembridge - https://www.mortgageintroducer.com/house-prices-higher-near-airports/
[14] Nyc Building Classes & Building Classification https://www.propertyshark.com/mason/text/nyc_building_class.html
[15] Queens Plaza (queens)
https://en.wikipedia.org/wiki/Queens_Plaza_(Queens)

residential space. Between 2011 and 2019, Tishman Speyer will have completed and fully leased 3.7 million square feet of dynamic mixed-use developments in the Queens Plaza district of Long Island City."[16] Tishman started construction on the project in 2016, according to another article published in The Real Deal.[17]

It was quite clear that in 2010 the property was a land for park use owned by the government, which accounted for why building depth and building front were 1. After 2011 Tishman Speyer bought the land. In 2016 the construction of office building began.

7.2.6 Rank 6 record

The abnormal data in this record was the small building front (8 feet) and building depth (8 feet) as opposed to the big lot front (4000 feet) and lot depth (150 feet). The building class of this property(Q1) indicated it was a park. The property owner is also "PARKS AND RECREATION". A photo we found in Google Maps based on the location (Joe DiMaggio Highway), manifested that the property is a park now.

But was it a park in 2010?

From NYC Parks, we found that "In 2000, seven acres of land stretching from 68th to 72nd Streets was added to Riverside Park, called Riverside Park South. This section of the park, …, was made possible by the construction of new portions of the West Side Highway, now known as the Joe DiMaggio Highway, and Trump/New World (the site's developers)."[18] It was clear that the property has been a park since 2010.

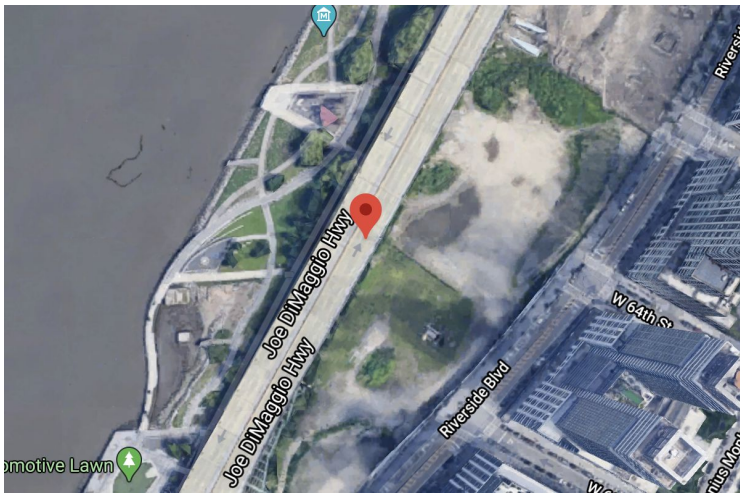
---

[16] The Jacx, Tishman Speyer's 1.2 Million Square Foot Creative Office Development In Long Island City, Reaches Full Occupancy Months Prior To Opening
Tishman Speyer -
https://www.prnewswire.com/news-releases/the-jacx-tishman-speyers-1-2-million-square-foot-creative-office-development-in-long-island-city-reaches-full-occupancy-months-prior-to-opening-300782314.html

[17] Tishman Speyer: 28-10 Queens Plaza South: Macy's Queens
Kevin Rebong -
https://therealdeal.com/2019/01/22/macys-love-of-queens-has-given-rob-speyer-867000-reasons-to-smile/

[18] Riverside Park
https://www.nycgovparks.org/parks/riverside-park/history

It is reasonable for a park to have a large area for land and a small area for building. Therefore, the possibility of this record being fraud was eliminated.



7.2.7 Rank 7 record

Based on the property's street name (11-01 43 AVENUE), we found that it is now a hotel called "Z NYC Hotel". The building class (H9) also indicated that it is a hotel.[19] But the building front and building depth are 1 foot in 2010, which are abnormal data.

According to Booking.com, Z NYC Hotel can be booked via the website since June 2011.[20] The earliest press we found on the official hotel website was published on June 3, 2011 with the title "The Z Hotel Is Almost Ready – It Just Needs Some Booze".[21] Therefore, it was possible that in 2010 the hotel was under construction in 2010, but the data recorder treated property in construction with 1 foot building front and 1 foot building depth. Based on these findings, we determined it is also possible that the record is fraud.

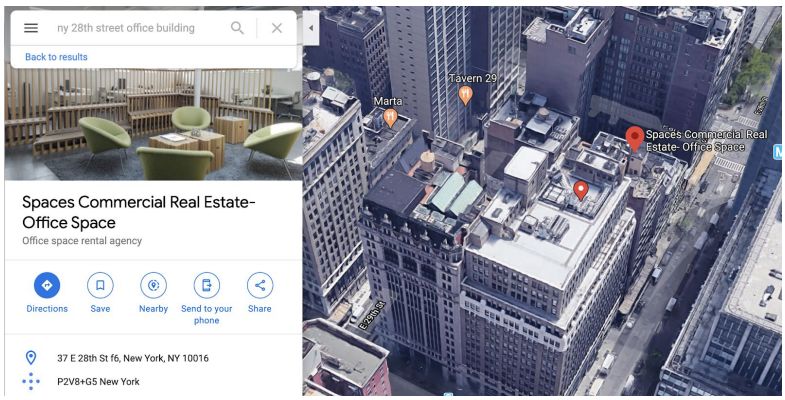
[19] Nyc Building Classes & Building Classification https://www.propertyshark.com/mason/text/nyc_building_class.html
[20] Z NYC Hotel https://www.booking.com/hotel/us/z-new-york.zh-cn.html?aid=318615%3Blabel&#hotelTmpl
[21] News & Press
https://www.zhotelny.com/press.htm

7.2.8 Rank 8 record

The property had no owner information. What was abnormal in this record was that both building front and building depth are 1 foot, as opposed to its 139 feet lot front and 342 feet lot depth.

The property is located on 28th street. The building class (O3) indicated that the property is an office building with 10 stories and over. We used "ny 28th street office building" as keywords to search in Google Maps. We found that on the 6th floor of the property is a WeWork-like office space owned by Spaces Commercial Real Estate. The firm was founded in 2015, whereas the time of raw data is 2010.[22]

Then we found that "37 E 28th" is the full street name of the property, as is indicated by the Google Map picture. We used it as the keyword to search in Google and found that the building was built in 1909.[23] The building had quite high value. There is a high probability that the record is fraud, since the building was built earlier than 2010 but with only 1 foot building front and 1 foot building depth according to the record.



[22] Spaces Commercial Real Estate https://www.crunchbase.com/organization/spaces-commercial-real-estate#section-overview
[23] https://streeteasy.com/building/37-east-28-street-new_york

7.2.9 Rank 9 record

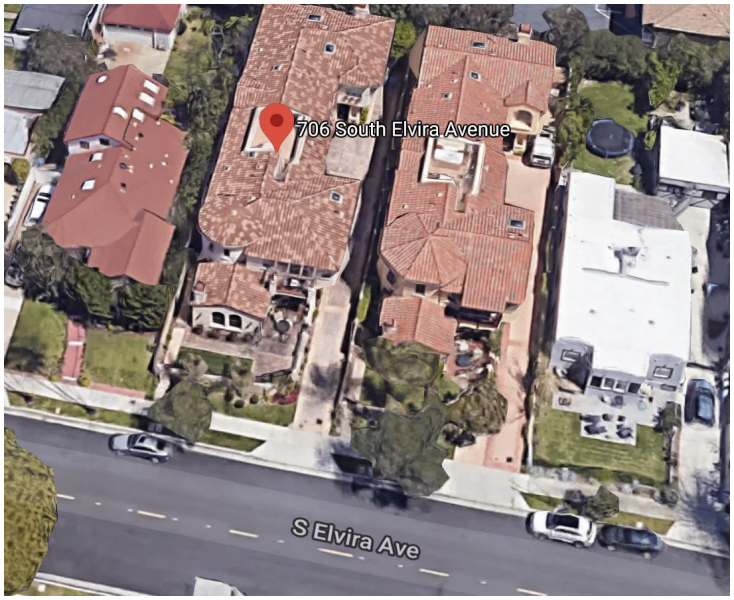
This record is like the Rank 2 record. The property is located on Flatbush Avenue, which is a major avenue in Brooklyn. The property is owned by the Department of General Service. The building class (V9) indicates miscellaneous use (Department of Real Estate and Other Public Places).[24]

The property was of very high property and land value. In terms of size, the lot front and lot depth of the building were 466 and 1009 feet respectively, while the building front and building depth were just 19 feet and 42 feet. This record was detected as fraud due to the high value per building front and value per building depth were abnormal. Nevertheless, the good location probably accounted for the high value of the property.

7.2.10 Rank 10 record

The abnormal part of this record was that both building front and building depth are 1 foot. The building class(A1) indicated that it is two stories detached (small or moderate Size, with or without Attic).[25] The building was built in 2008, according to StreetEasy.[26] It is unusual for a house to have a small building front and building depth in 2010 as it was built two years ago. Therefore the property had a high possibility of fraud.

[24] Nyc Building Classes & Building Classification
https://www.propertyshark.com/mason/text/nyc_building_class.html
[25] Class A Buildings - One Family Dwellings
https://www.propertyshark.com/info/class-buildings-one-family-dwellings/#a1
[26] https://streeteasy.com/building/7_06-elvira-avenue-queens



To sum up, high value and low building/lot size were the common characteristics shared by the top 10 records.

In detail, Rank 2, 5, 6 and 9 records were all government-owned properties which featured high value or low building front/depth. Property in rank 4 record was also of high value because it was near an airport. Other 5 records were owned by a corporation or an individual. All of them had very small building/lot size and some of them had high value, which indicated the possibility of fraud without sound reason.

The explanation above was based on research of online information, and valid information was omitted. The fraud detection can be improved with further adjustment in the future.

8. Appendix

8.1 Data Quality Report

8. 1.1 Data Description

Dataset Name: Property Valuation and Assessment Data
Dataset Source: NYC City Government, Department of Finance
Time Period: 11/2010
# of Fields: 32
# of Records: 1,070,994

8.1.2 Summary

8.1.2.1 Numeric Fields

	# records that have a value	% populated	# unique values	# records with value zero	Mean	Standard Deviation	Min	Max
LTFRONT	1070994	100.0	1297	169108	36.6	74.0	0	9999
LTDEPTH	1070994	100.0	1370	170128	88.9	76.4	0	9999
STORIES	1014730	94.7	112	0	5.0	8.4	1	119
FULLVAL	1070994	100.0	109324	13007	874264.5	11582431.0	0	6150000000
AVLAND	1070994	100.0	70921	13009	85067.9	4057260.1	0	2668500000
AVTOT	1070994	100.0	112914	13007	227238.2	6877529.3	0	4668308947
EXLAND	1070994	100.0	33419	491699	36423.9	3981575.8	0	2668500000

EXTOT	1070994	100.0	64255	432572	91187.0	6508402.8	0	4668308947
BLDFRONT	1070994	100.0	612	228815	23.0	35.6	0	7575
BLDDEPTH	1070994	100.0	621	228853	39.9	42.7	0	9393
AVLAND2	282726	26.4	58592	0	246235.7	6178962.6	3	2371005000
AVTOT2	282732	26.4	111361	0	713911.4	11652528.9	3	4501180002
EXLAND2	87449	8.2	22196	0	351235.7	10802212.7	1	2371005000
EXTOT2	130828	12.2	48349	0	656768.3	16072510.2	7	4501180002

8.1.2.2 Categorical Fields

	# records that have a value	% populated	# unique values	most common field value
RECORD	1070994	100.0	1070994	2047
BBLE	1070994	100.0	1070994	3068950052
B	1070994	100.0	5	4
BLOCK	1070994	100.0	13984	3944
LOT	1070994	100.0	6366	1
EASEMENT	4636	0.4	13	E
OWNER	1039249	97.0	863347	PARKCHESTER PRESERVAT
BLDGCL	1070994	100.0	200	R4
TAXCLASS	1070994	100.0	11	1
EXT	354305	33.1	4	G
EXCD1	638488	59.6	130	1017
STADDR	1070318	99.9	839281	501 SURF AVENUE

ZIP	1041104	97.2	197	10314
EXMPTCL	15579	1.5	15	X1
EXCD2	92948	8.7	61	1017
PERIOD	1070994	100.0	1	FINAL
YEAR	1070994	100.0	1	2010/11
VALTYPE	1070994	100.0	1	AC-TR

8.1.3 Data Exploration

8.1.3.1 RECORD

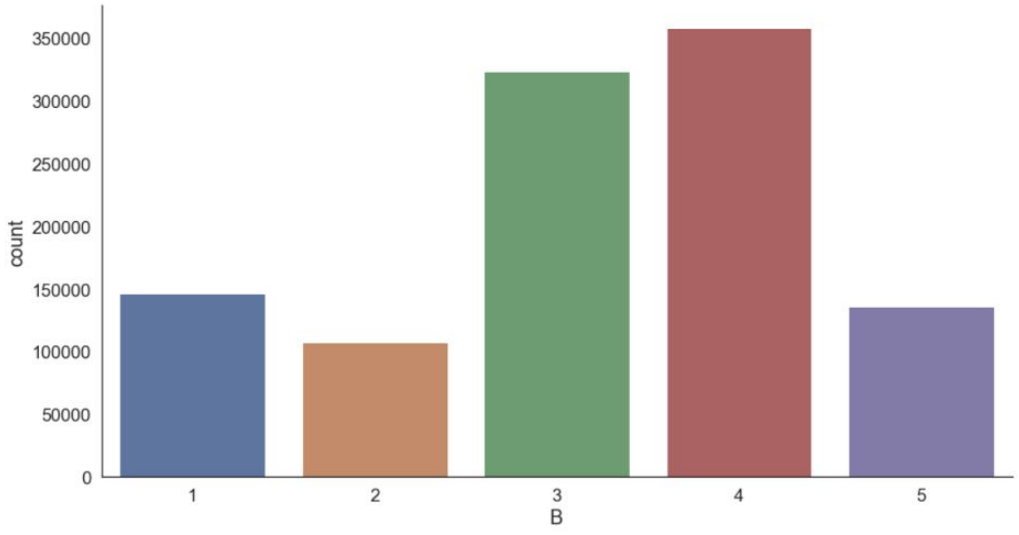
Description: To uniquely identify each record

8.1.3.2 BBLE

Description: Concatenation of borough code, block code, lot code (unique # within borough/block) and easement code

8.1.3.3 B

Description: Borough code



8.1.3.4 BLOCK

Description: Valid block ranges by borough

Top 10 Field Values	
BLOCK	COUNT
3944	3888
16	3786
3943	3424
3938	2794
1171	2535
3937	2275
1833	1774
2450	1651
1047	1480
7279	1302

8.1.3.5 LOT

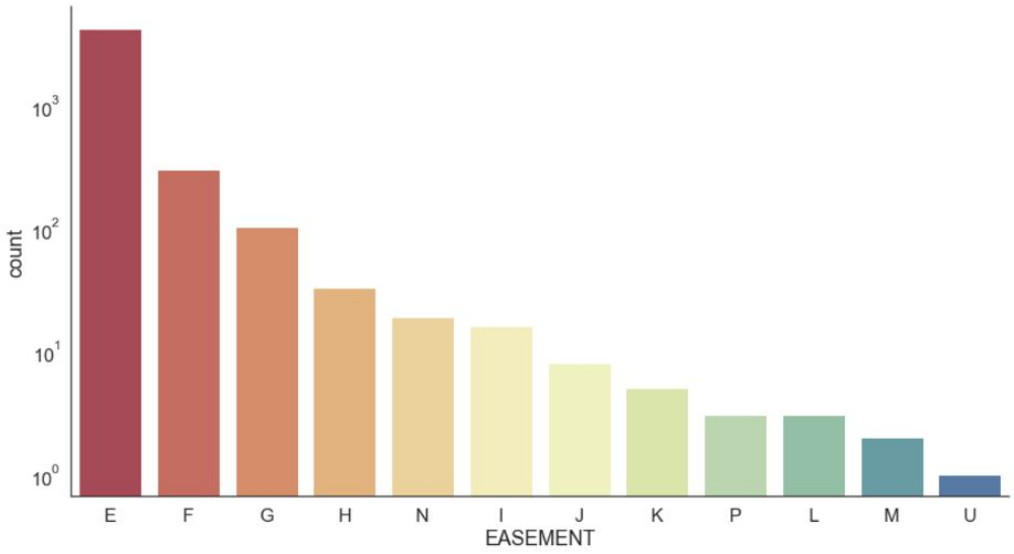
Description: Unique # within borough/block

Top 10 Field Values	
LOT	COUNT
1	24367
20	12294
15	12171

12	12143
14	12074
16	12042
17	11982
18	11979
25	11949
21	11840

8.1.3.6 EASEMENT

Description: Describe easement



8.1.3.7 OWNER

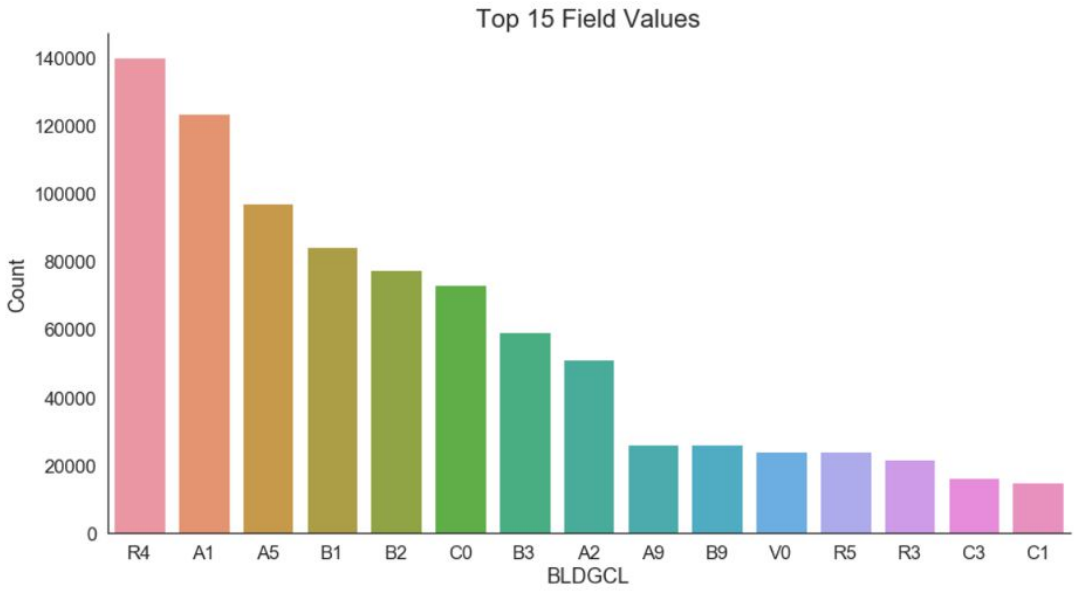
Description: Owner's name

Top 10 Field Values	
OWNER	COUNT

PARKCHESTER PRESERVAT	6020
PARKS AND RECREATION	4255
DCAS	2169
HOUSING PRESERVATION	1904
CITY OF NEW YORK	1450
DEPT OF ENVIRONMENTAL	1166
BOARD OF EDUCATION	1015
NEW YORK CITY HOUSING	1014
CNY/NYCTA	975
NYC HOUSING PARTNERSH	747

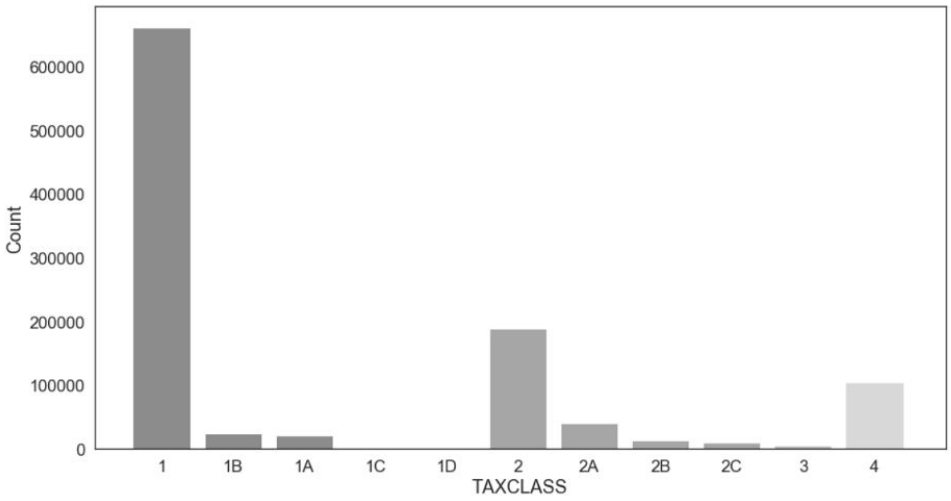
8.1.3.8 BLDGCL

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



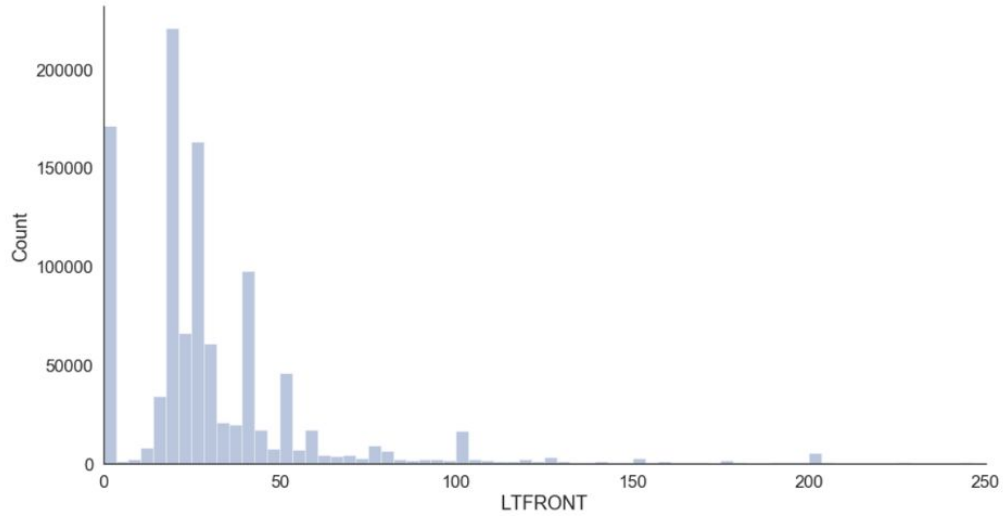
8.1.3.9 TAXCLASS

Description: Current Property Tax Class Code (NYS Classification)



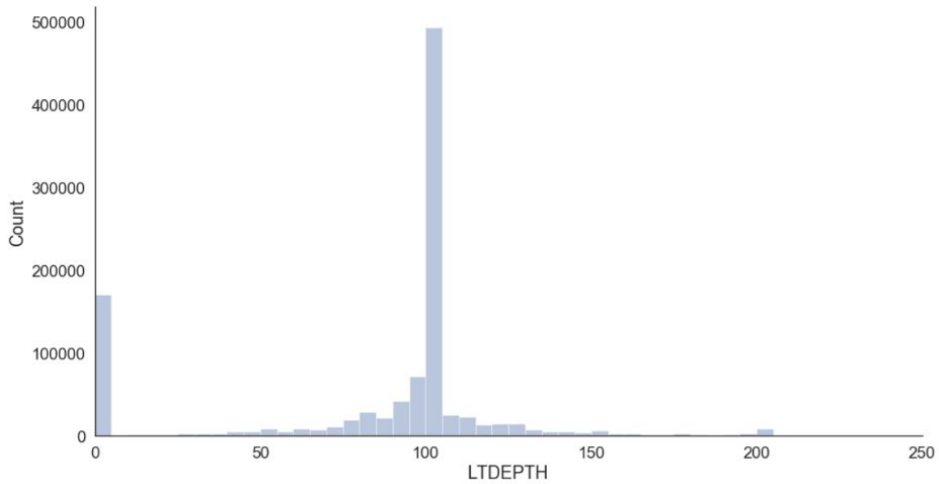
8.1.3.10 LTFRONT

Description: Lot Frontage in feet
Outliers: LTFRONT > 250
Histogram: Covering 99.06% populated records



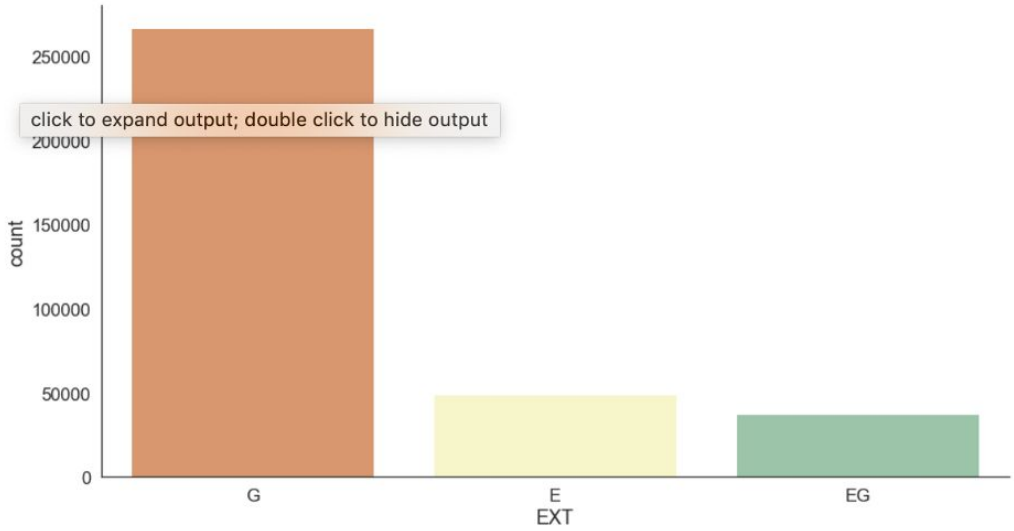
8.1.3.11 LTDEPTH

Description: Lot Depth in feet
Outliers: LTDEPTH > 250
Histogram: Covering 98.89% populated records



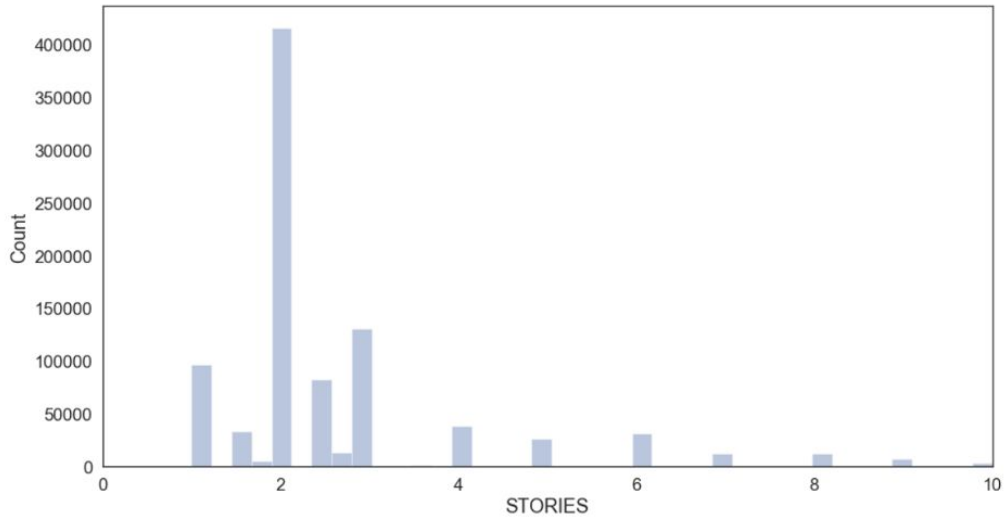
8.1.3.12 EXT

Description: Extension.
'E' = EXTENSION
'G' = GARAGE
'EG' = EXTENSION AND GARAGE



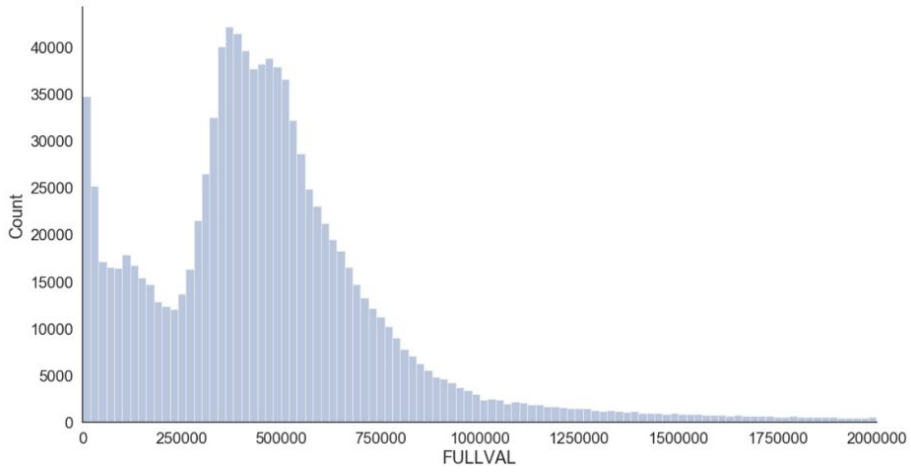
8.1.3.13 STORIES

Description: The number of stories for the building (# of Floors)
Outliers: STORIES > 10
Histogram: Covering 89.62% populated records



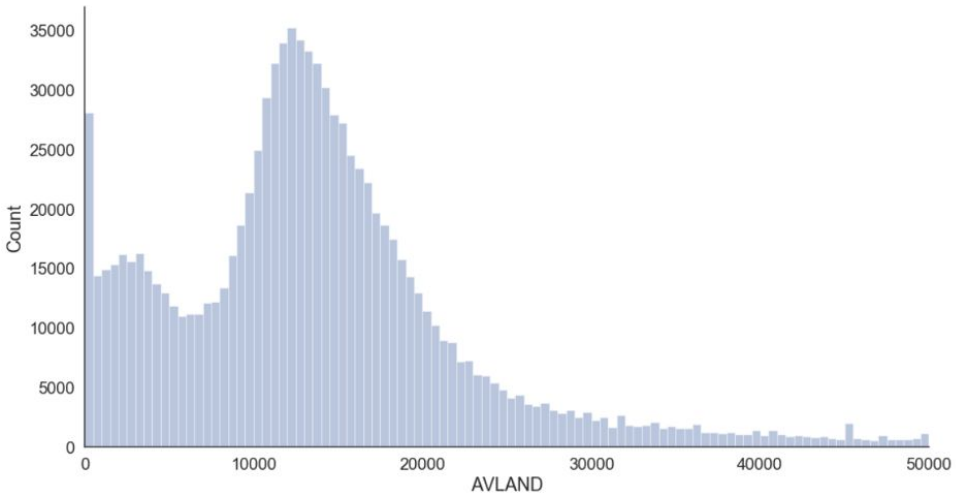
8.1.3.14 FULLVAL

Description: Total Market Value of property
Outliers: FULLVAL > 2000000
Histogram: Covering 96.31% populated records



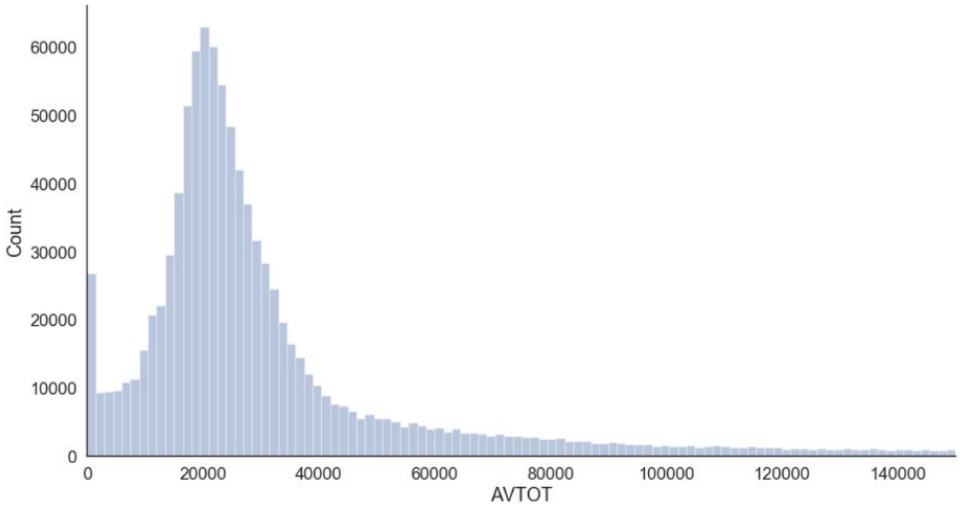
8.1.3.15 AVLAND

Description: Actual Market value of the land
Outliers: AVLAND > 50000
Histogram: Covering 90.53% populated records



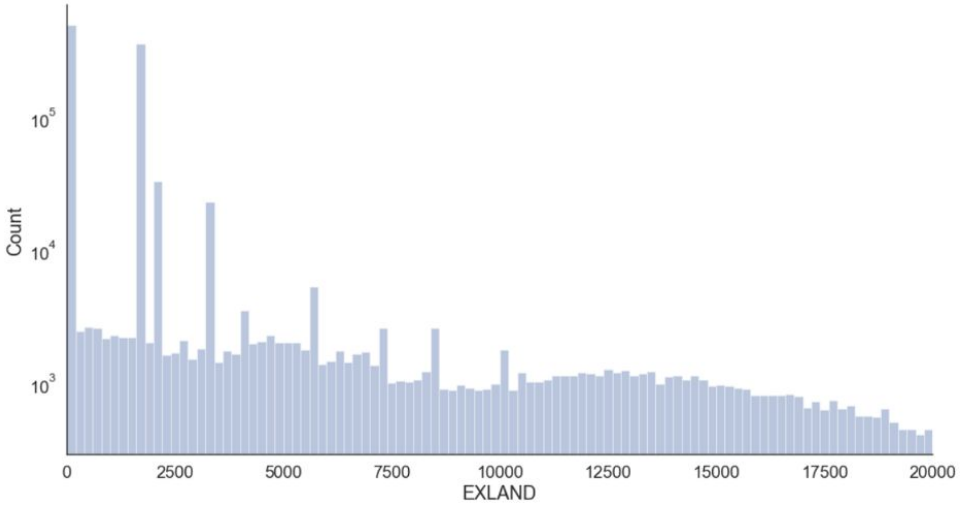
8.1.3.16 AVTOT

Description: Actual Total market value
Outliers: AVTOT > 150000
Histogram: Covering 89.58% populated records



8.1.3.17 EXLAND

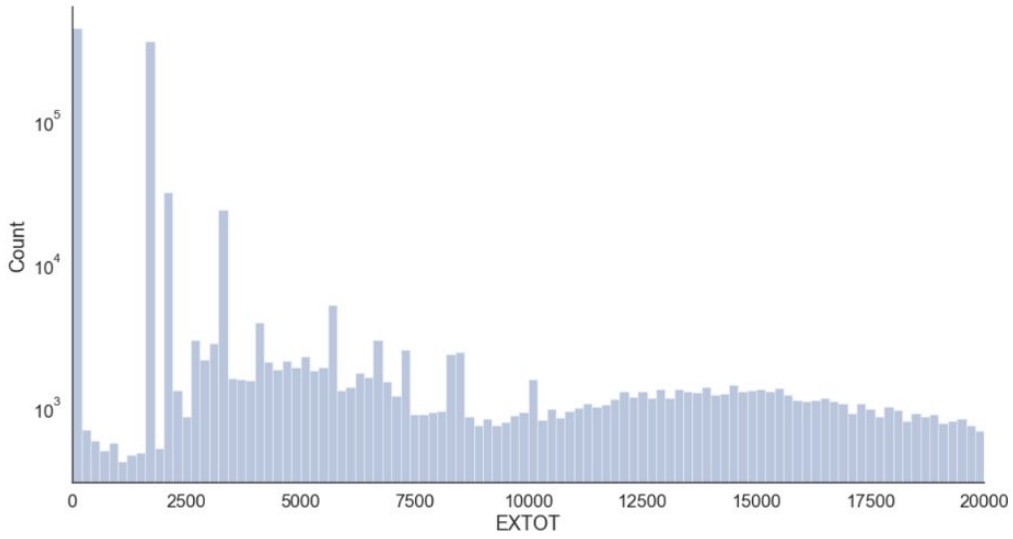
Description: Actual Exempt Land Value
Outliers: EXLAND > 20000
Histogram: Covering 96.82% populated records



8.1.3.18 EXTOT

Description: Actual Exempt Total Value
Outliers: EXTOT > 20000

Histogram: Covering 90.40% populated records



8.1.3.19 EXCD1

Description: Exemption Code 1

Top 10 Field Values	
EXCD1	COUNT
1017	425348
1010	49756
1015	31323
5113	23858
1920	17594
5110	16834
5114	14984
5111	10609

1021	6613
1986	4231

8.1.3.20 STADDR

Description: Street name for the property

Top 10 Field Values	
STADDR	COUNT
501 SURF AVENUE	902
330 EAST 38 STREET	817
322 WEST 57 STREET	720
155 WEST 68 STREET	671
20 WEST 64 STREET	657
1 IRVING PLACE	650
220 RIVERSIDE BOULEVARD	628
360 FURMAN STREET	599
200 EAST 66 STREET	585

8.1.3.21 ZIP

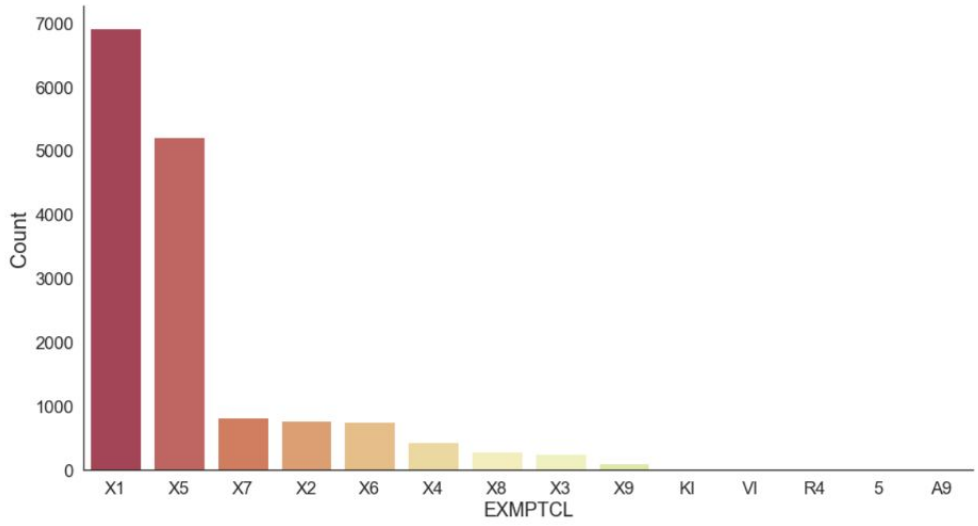
Description: Postal Zip code of the property

Top 10 Field Values	
ZIP	COUNT

10314	24606
11234	20001
10312	18127
10462	16905
10306	16578
11236	15678
11385	14921
11229	12793
11211	12710
11207	12293

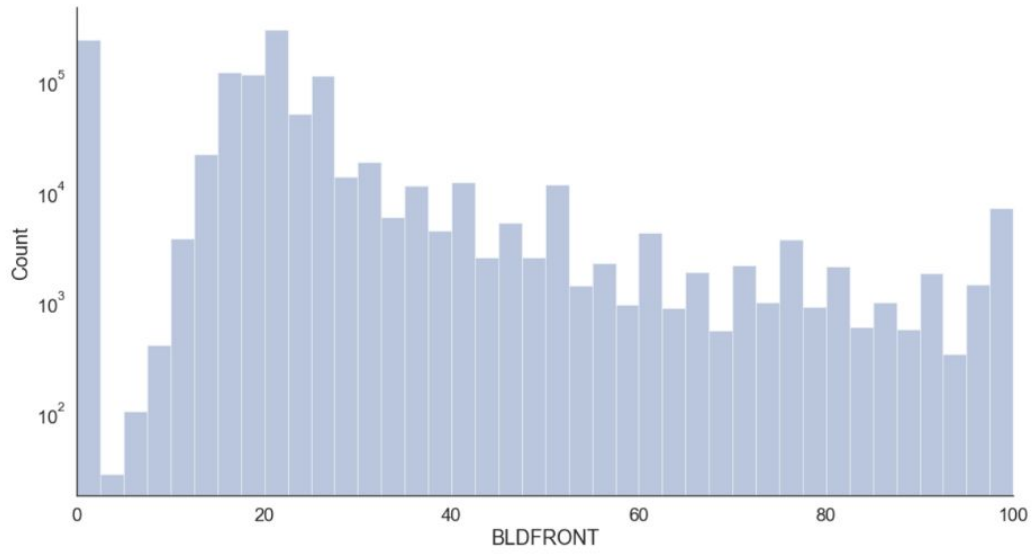
8.1.3.22 EXMPTCL

Description: Exempt Class used for fully exempt properties only



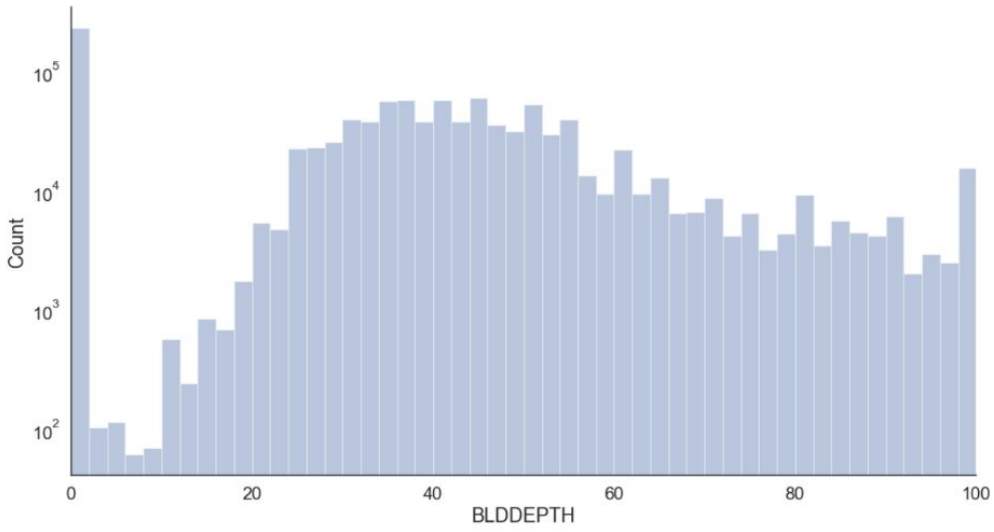
8.1.3.23 BLDFRONT

Description: Building Frontage in feet.
Outliers: BLDFRONT > 100
Histogram: Covering 97.37% populated records



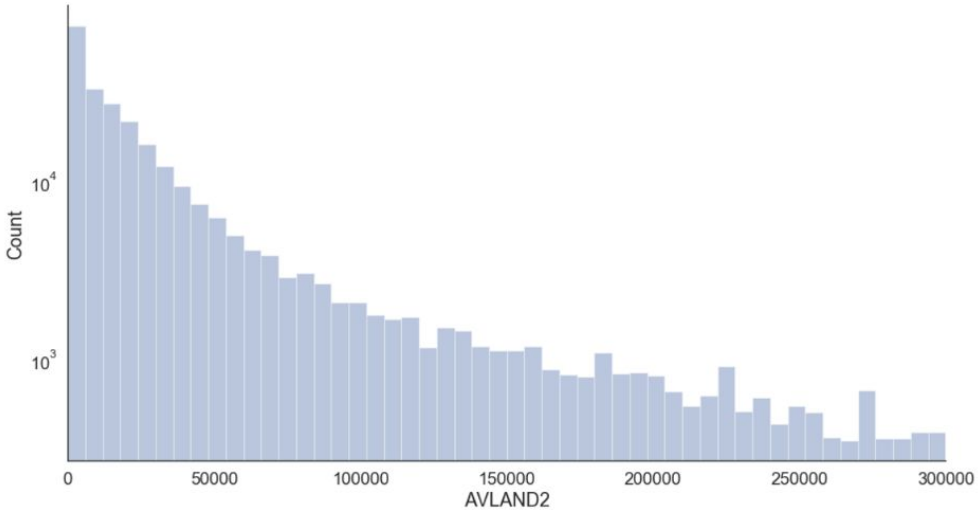
8.1.3.24 BLDDEPTH

Description: Lot Depth in feet.
Outliers: BLDDEPTH > 100
Histogram: Covering 97.40 % populated records



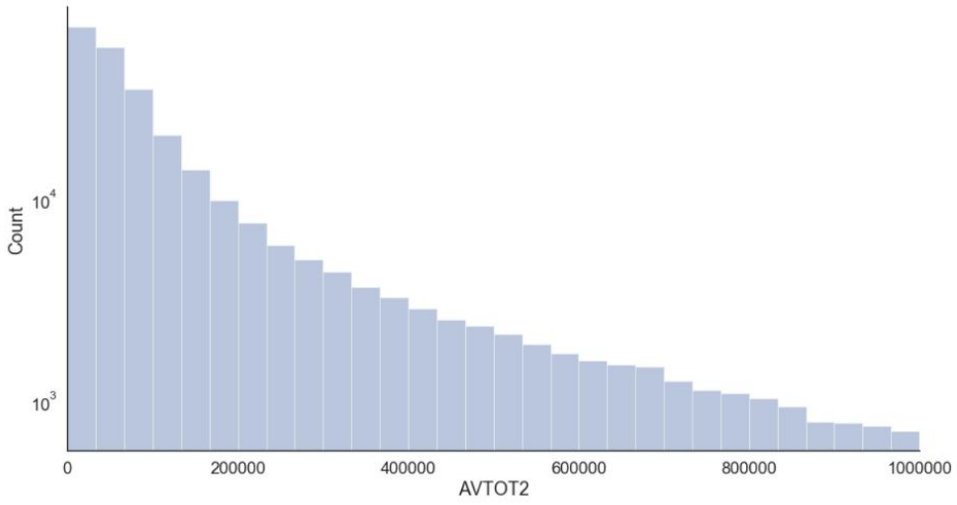
8.1.3.25 AVLAND2

Description: Transitional Land Value
Outliers: AVLAND2 > 300000
Histogram: Covering 91.48% populated records



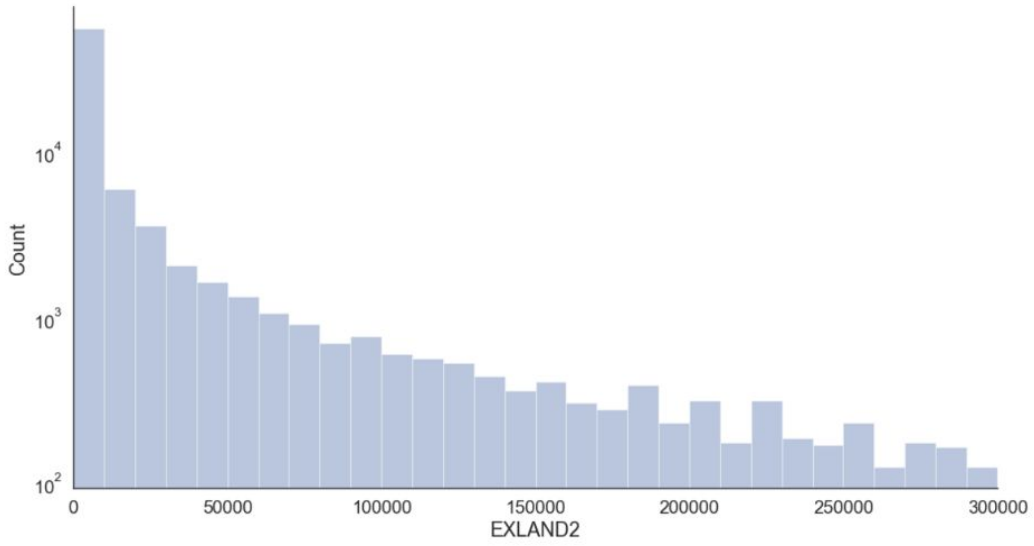
8.1.3.26 AVTOT2

Description: Transitional Total Value
Outliers: AVTOT2> 1000000
Histogram: Covering 91.62% populated records



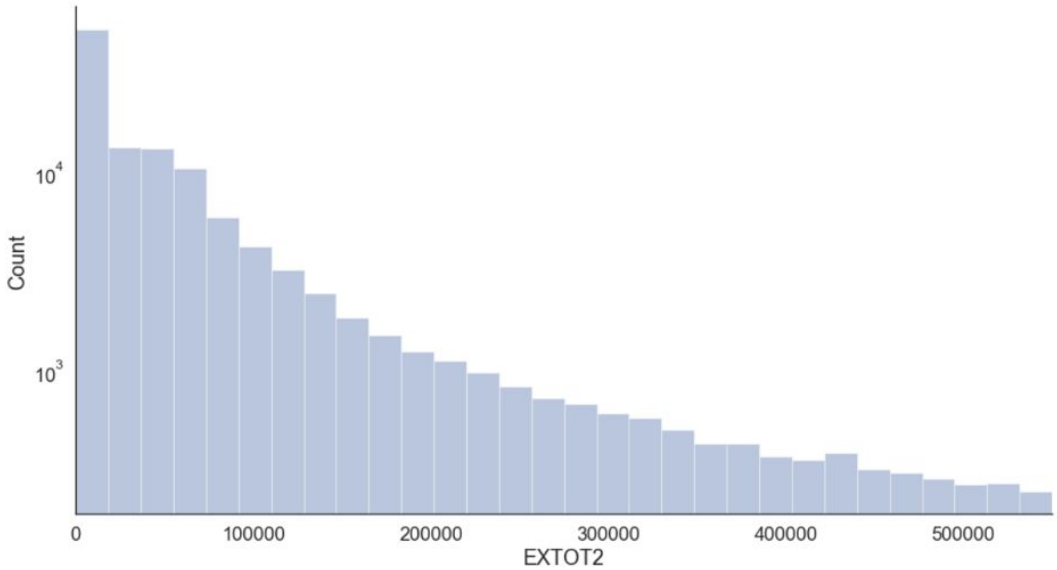
8.1.3.27 EXLAND2

Description: Transitional Exemption Land Value
Outliers: EXLAND2> 300000
Histogram: Covering 90.91% populated records



8.1.3.28 EXTOT2

Description: Transitional Exemption Land Total
Outliers: EXTOT2> 550000
Histogram: Covering 90.53% populated records



8.1.3.29 EXCD2

Description: Exemption Code 2

Top 10 Field Values	
EXCD2	COUNT
1017	65777
1015	12337
5112	6867
1019	3178
1920	2961
1200	881
1101	494
5129	227
1986	35
1022	31

8.1.3.30 PERIOD

Description: Assessment Period (when data was created)
All data is 'FINAL'

8.1.3.31 YEAR

Description: Assessment Year
All data is '2010/11'

8.1.3.32 VALTYPE

Description: Type of the data value
All data is ‘AC-TR’