New York Property
Unsupervised Anomaly Detection

DSO 562 - Group 8

Divya Sripathy
Fabbiha Islam
Hongxuan Wang
Jayant Maheshwari
Jude Alfuraih
Shashank Tiwari
Snehil Saraswat

13 February 2020

Table of Contents

Executive Summary

Group 8 has been hired by the city of New York to detect potential cases of property tax fraud. The following report gives details on the algorithmic system created in order to detect anomalous properties from unlabeled data provided by New York City, consisting of over 1,000,000 property records detailing attributes such as valuations, area, and location.

In order to identify anomalous records in the dataset provided by the city of New York, 2 different fraud scores are calculated for each property using 2 unsupervised learning methods: a heuristic fraud score as a function of z-scores and a trained autoencoder. Both fraud scores have been scaled and combined using a quantile binning/rank ordering method, resulting in an overall fraud score for each record. The specific steps of this analytics pipeline are below:

- **Data Cleaning**: Imputation of Missing Values
- **Variable Creation**: Creation of 45 Expert Variables to Determine Fraud Scores
- **Dimensionality Reduction**: Scaling and Reduction through PCA
- **Algorithms**: Generating Fraud Scores Using 2 Methods and Combining Them
- **Results**: Analysis of Top 10 Anomalous Records

After completing the above analysis, our results show that the following records are flagged as anomalous:

- Records with inconsistency between either lot or building size (front & depth) and their valuation
- Records that show high standard deviations in our expert variables because of a large lot or building size (front & depth)
- Records with extremely high values in our expert variables, compared to the rest of the data
- Records that show inconsistency in the type of building (# of units & stories) and lot size (front & depth)

These results verify that our system successfully detects anomalous records from the NY Property dataset and can be utilized to point at potential cases of property tax fraud.

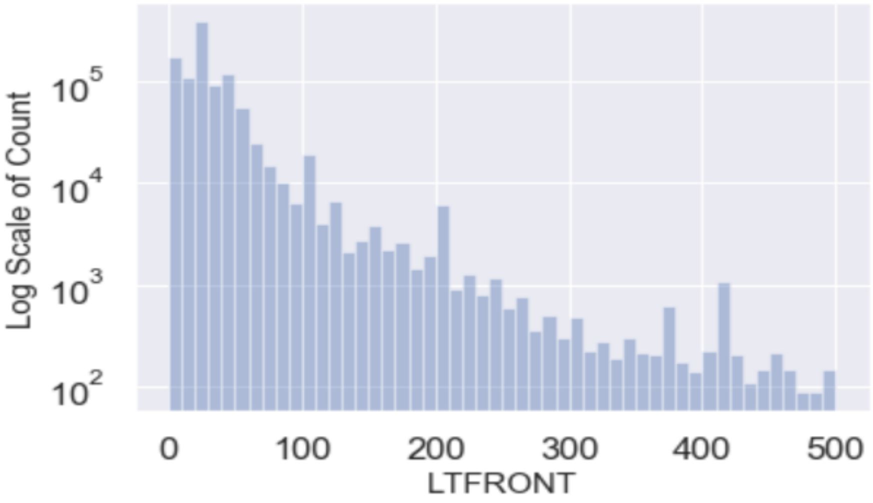
Data Description

The city of NY property dataset is an open data from NYC city government and consists of properties assessments for the purpose of calculating Property Tax, Grant eligible properties, Exemptions and/or Abatements. The data has been collected and entered into the system by various city employees including Property Assessors, Property Exemption specialists, ACRIS reporting, Department of Building reporting, etc. It consists of 1070994 records and 32 columns, 14 of which are numerical and others being categorical. Pertinent fields used in our analysis can be found in the table below.

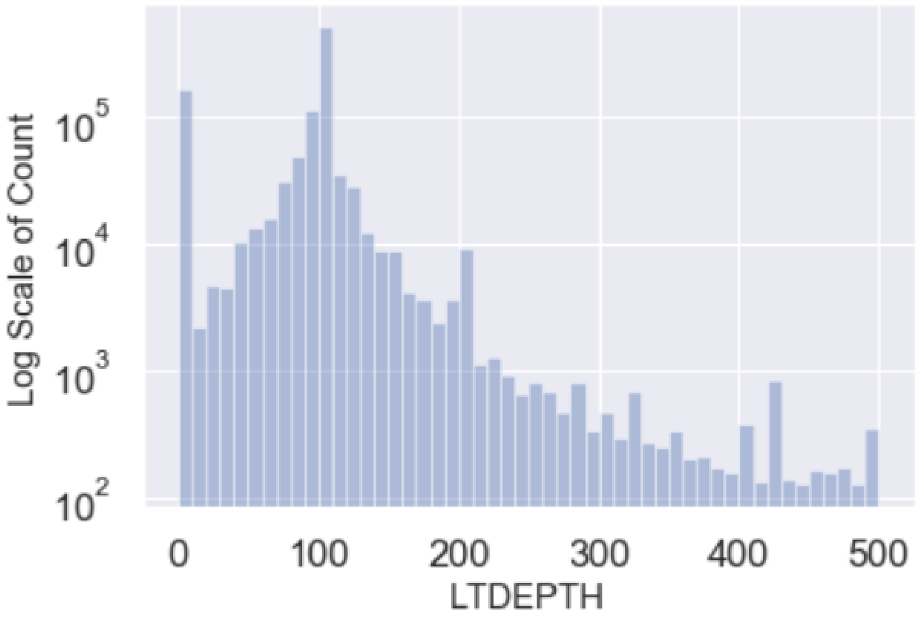
Filed Name	# of records	% Populated	# of 0 value fields	Min	Mean	Median	Max	Standard Deviation
ZIP	1041104	97.2%	0	-	-	-	-	-
STORIES	1014730	94.7%	0	1	5	2	119	8
LTFRONT	1070994	100.0%	169108	0	37	25	9999	74
LTDEPTH	1070994	100.0%	170128	0	89	100	9999	76
BLDFRONT	1070994	100.0%	228815	0	23	20	7575	36
BLDDEPTH	1070994	100.0%	228853	0	40	39	9393	43
FULLVAL	1070994	100.0%	13007	0	874264	447000	6150000000	11582431
AVLAND	1070994	100.0%	13009	0	85068	13678	2668500000	4057260
AVTOT	1070994	100.0%	13007	0	227238	25340	4668308947	6877529

A brief description of each field is given below. These and all other fields are discussed in greater detail in the Data Quality Report provided as Appendix:

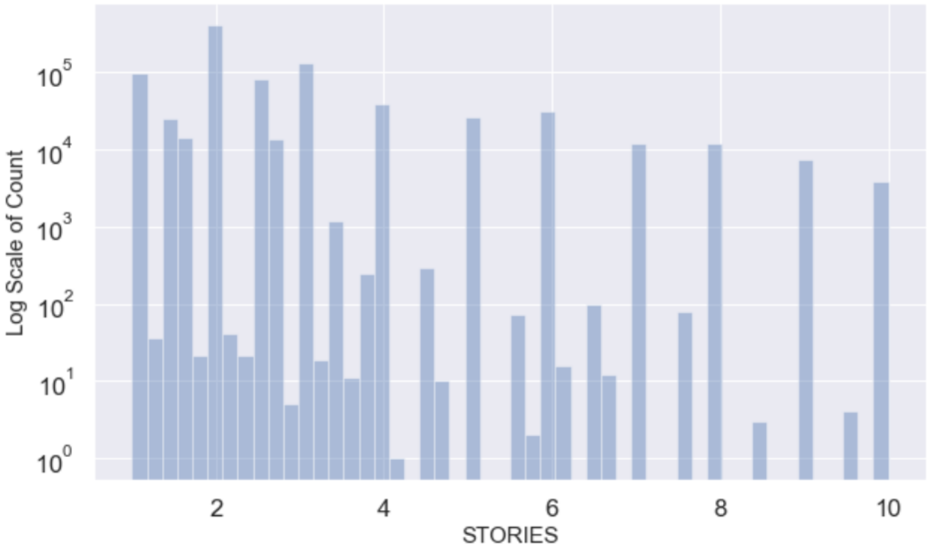
1. Field Name: LTFRONT
   Type: Numerical
   Description: Lot Frontage in feet
   Outliers Removed: > 500
   Data in the histogram is 99.76% populated after removing outliers.



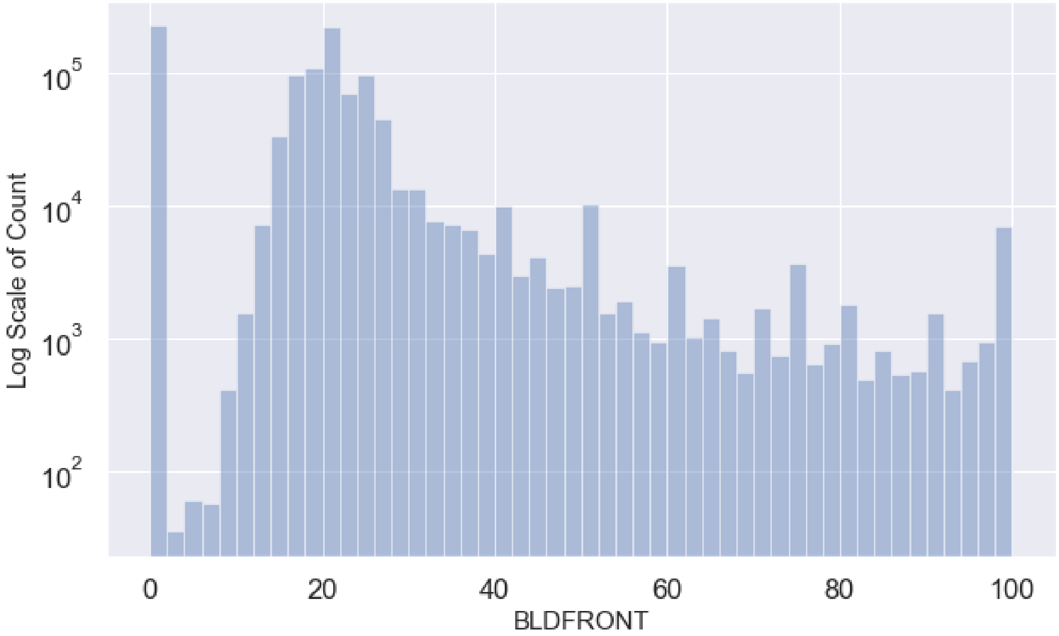
2. Field Name: LTDEPTH
Type: Numerical
Description: Lot Depth in feet
Outliers Removed: > 500
Data in the histogram is 99.68% populated after removing outliers.



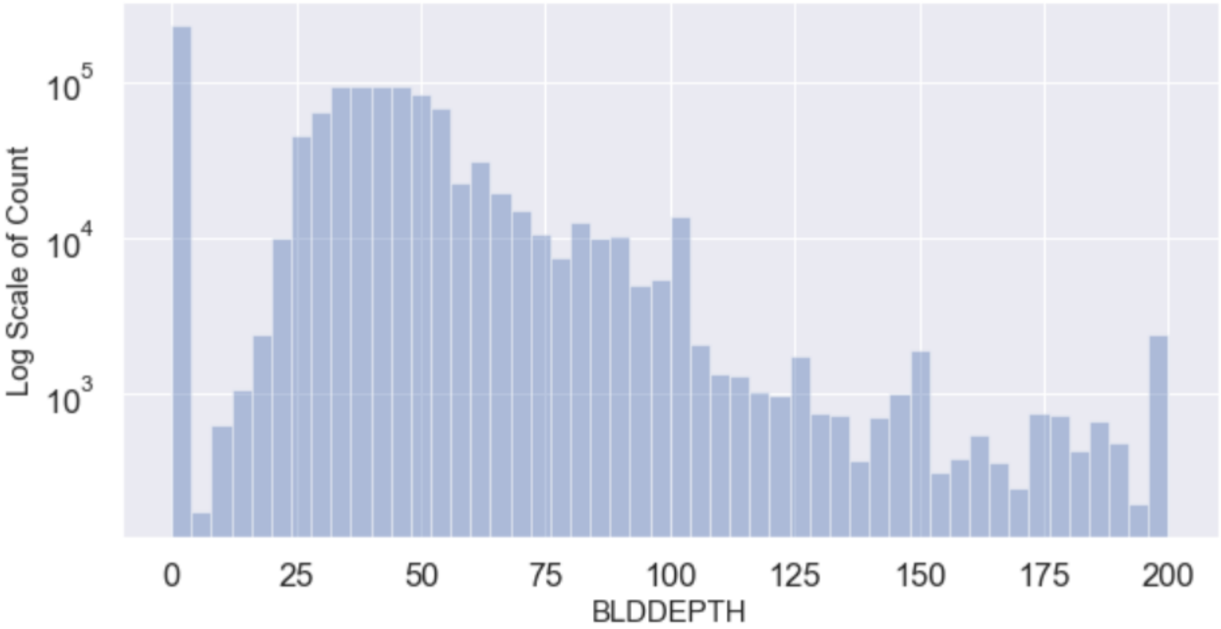
3. Field Name: STORIES
Type: Numerical
Description: The number of stories in the building (# of floors)
Outliers Removed: > 10
Data in the histogram is 89.6% populated after removing outliers.



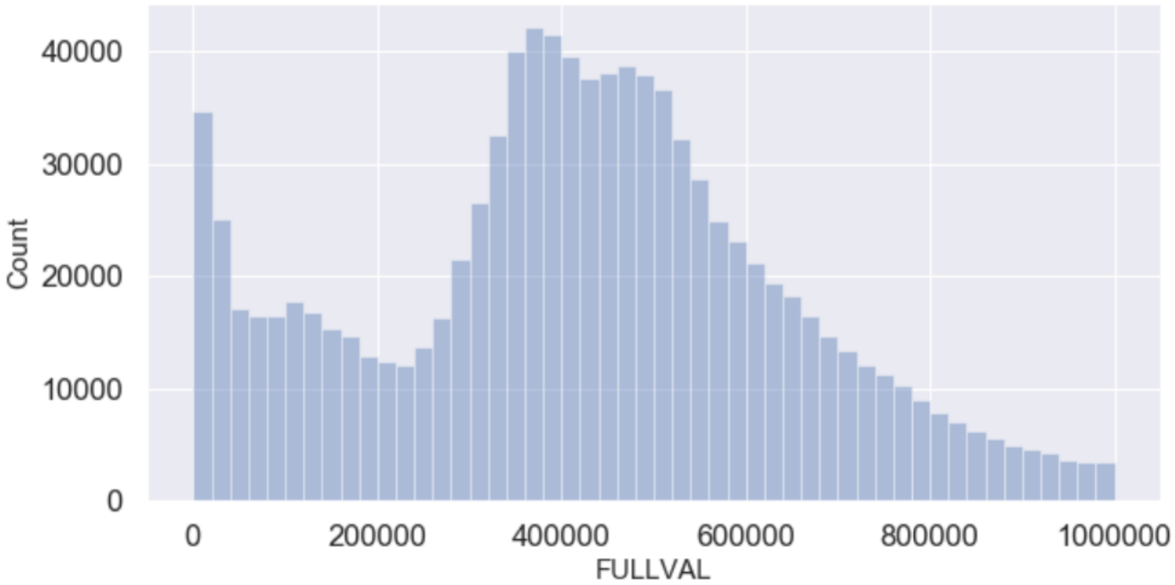
4. Field Name: BLDFRONT
Type: Numerical
Description: Building Frontage in feet
Outliers Removed: > 100
Data in the histogram is 97.3% populated after removing outliers.



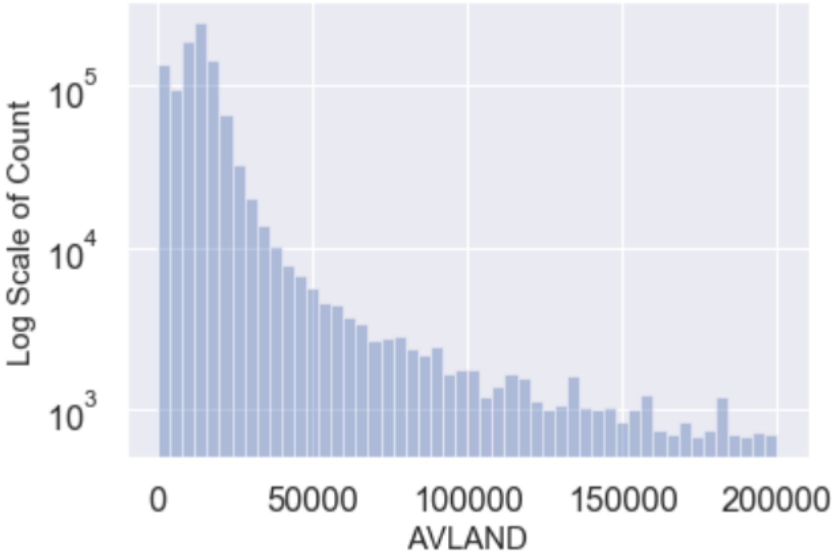
5. Field Name: BLDDEPTH
Type: Numerical
Description: Building Depth in feet
Outliers Removed: > 200
Data in the histogram is 99.6% populated after removing outliers.



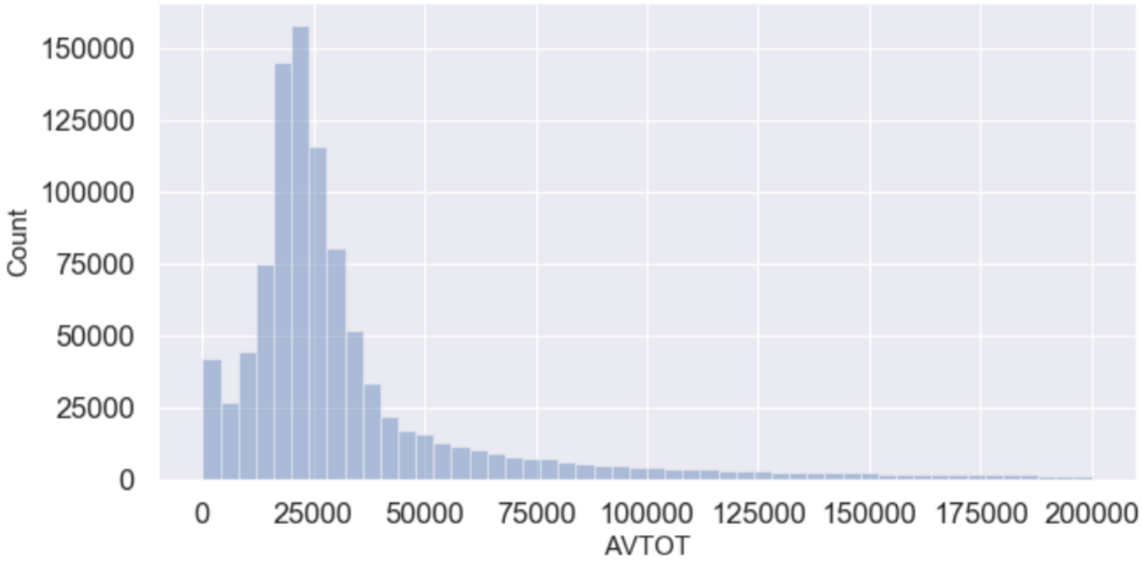
6. Field Name: FULLVAL
Type: Numerical
Description: Total Market Value
Outliers Removed: > 1000000
Data in the histogram is 91.3% populated after removing outliers.



7. Field Name: AVLAND
Type: Numerical
Description: Total Assessed Value of the land
Outliers Removed: > 200000
Data in the histogram is 90.4% populated after removing outliers.



8. Field Name: AVTOT
   Type: Numerical
   Description: Total Assessed Value
   Outliers Removed: > 200000
   Data in the histogram is 91.6% populated after removing outliers.



Apart from ZIP, all other fields are numerical. Only ZIP and STORIES fields have 3-5 % missing data. However, there are many zero values present in most numerical variables that must be treated before further analysis. Fields like STORIES, FULLVAL, AVLAND, and AVTOT have outliers present and thus using the median would be preferred over the mean when imputing missing values.

Data Cleaning

In this section, we impute any missing values in the 9 fields described above.
There are 2 scenarios in which we decided to impute values:

1. The value in the field of interest is null: in this case, we use a logical method determined by the nature of the field to impute the missing value (outlined later in this section).
2. The value in the field of interest has a value of 0: since properties cannot have lot areas or attributes with a measurement of 0, these values must be imputed with the same logical approach as null values.

Imputing these missing values and zero value entries lead to better prediction power down the analytics pipeline as well as the utilization of available data rich in other fields. Our goal when filling missing values is to impute innocuous values that won’t be flagged as anomalous. Please note that any outliers in fields having a value other than 0 have not been imputed to mitigate the risk of unknowingly altering potentially anomalous values.

Methodology for filling missing fields and correcting zero value error:

If values in a field are null or wrongly entered as 0, they are replaced with group mode, median or mean depending on the presence and strength of outliers. If outliers are present, then the median value is preferred over mean as it is more robust to outliers.

1. **Fields with missing values:** Among the 9 fields mentioned in the data description, only the ZIP and STORIES fields have 3-5% missing values where data is not populated. Our method to impute these missing values began with a narrow grouping of similar properties to impute a value and then expanding it to a wider group if the sample size in each group is too small in order to get an accurate value to impute. Our approach is as follows:

   - ***ZIP*:** To populate this field, the missing values are filled with the mode of the zip codes field by grouping BLOCK and BOROUGH code. If the number of records in each group is less than 20, then we group it by BOROUGH only.

   - ***STORIES:*** To populate this field, the missing values are filled with the median of the STORIES field by grouping TAXCLASS, BLOCK, and ZIP. If the number of records in each group is less than 20, then we group by TAXCLASS and BLOCK. If the number of records is still less than 20, then we group by TAXCLASS only.

2. **Fields that have zero values:** The value for fields LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, FULLVAL, AVLAND, and AVTOT consist of zero values which need to be imputed as well.

   - ***LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH:*** To replace the zeros in these fields, we use the median of each by grouping by TAXCLASS, BLOCK, & ZIP. If there are less than 20 records, we group by ZIP & BLOCK. If there are still less than 20 records, we group by ZIP only.

   - ***FULLVAL, AVLAND, AVTOT***: To replace the zeros in these fields, we use the median of each by grouping by TAXCLASS, STORIES, BLOCK & ZIP. If there are less than 20 records, we group by the values by BLOCK, STORIES & ZIP. If there are less than 20 records, we group by the values by BLOCK & ZIP. If there are still less than 20 records, we group by ZIP only.

Our choice for deciding fields to group by depended upon the accuracy of the imputation. For example, to impute the ZIP field, it is most likely that a property in the same block and borough will have the same zip code. For the STORIES field, it is most likely that a property in the same tax class, block, and zip code will have a similar value for the number of stories. We have used a similar logic to determine the grouping of the other fields.

Variable Creation

The feature engineering process from the 9 variables in the previous section is outlined below:

1. Identifying relevant variables based on the business understanding of the problem.
2. Creating meaningful variables to compare them on a level footing, i.e. a normalized measure.

This approach helps in comparing data across other properties on the same scale.

Step 1: Identifying Relevant Variables

The problem at hand deals with identifying properties that have under or over reported the valuation of their land for tax exemptions; therefore, the key variables of interest that determine the land’s value in the dataset are:

1. FULLVAL: The Market Value of the Property (in $).
2. AVLAND: The Assessed Value of the Property’s land (in $).
3. AVTOT: The Total Assessed Value of the Property (in $).

These three variables are pertinent to determining how much tax a property owner pays based on the property’s market value and how much is its assessed value and are therefore important to be considered when detecting anomalous records.

Step 2: Creating Meaningful Variables

The value of every property will be a function of its land area: the lot frontage outside the property. A property is likely to have a higher valuation for a larger land area. Therefore, if these property values are directly used in our model to identify outliers, there is potential for the model to show incorrect bias towards larger properties as outliers since they are not standardized.

Therefore, to transform them to the same scale, the property values are normalized by dividing them by relevant property sizes.

The size metrics available in the dataset are:

1. LTFRONT: The Lot Frontage Length (in feet).
2. LTDEPTH: The Lot Frontage Depth (in feet).
3. BLDFRONT: The Building Frontage Length (in feet).
4. BLDDEPTH: The Building Frontage Depth (in feet).
5. STORIES: The Number of Stories in the Property.

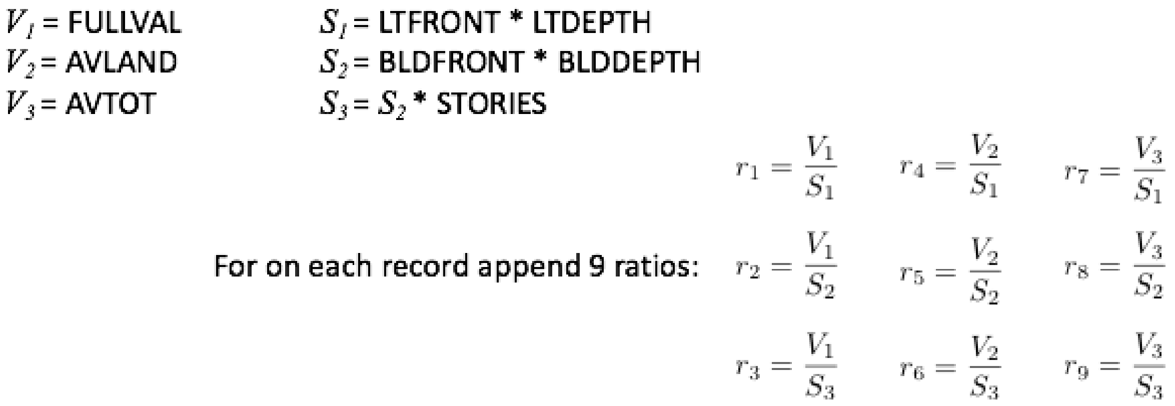
The three relevant size variables of interest are created as follows:

1. Lot Area (LTAREA) = LTFRONT*LTDEPTH
2. Building Area (BLDAREA) = BLDFRONT*BLDDEPTH
3. Building Volume (BLDVOLUME) = STORIES*BLDAREA

Now, by dividing every value measure by the three variables, we create the following nine measures for every entry:

1. r1 – Market Value of the Property per unit lot area.
2. r2 – Market Value of the Property per unit building land area.
3. r3 – Market Value of the Property per unit volume of the building.
4. r4 – Assessed Value of the Land per unit lot area.
5. r5 – Assessed Value of the Land per unit building land area.
6. r6 – Assessed Value of the Land per unit volume of the building.
7. r7 – Total Assessed Value of the Property per unit lot area.
8. r8 – Total Assessed Value of the Property per unit building area.
9. r9 – Total Assessed Value of the Property per unit volume of the building.

A pictorial representation of the same can be found below:



V_1 = FULLVAL S_1 = LTFRONT * LTDEPTH

V_2 = AVLAND S_2 = BLDFRONT * BLDDEPTH

V_3 = AVTOT S_3 = S_2 * STORIES

For on each record append 9 ratios:

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

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

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

Now, to compare how different these normalized valuations are compared to other properties, we separately grouped properties by the 5 groups mentioned below:

1. ZIP5: The zipcode of the property. It is possible for properties in the same zip code to have similar valuations, based on how zip code lines are drawn.
2. ZIP3: The first three digits of the zipcode. This enables the comparison of properties over a larger area but with fairly similar attributes.
3. TAXCLASS: The tax class allocated to the property can help aggregate buildings and structures of similar types together.

4. BOROUGH: Which borough is the property is located in New York. Properties compared within a borough are not likely to have a very high standard deviation within as opposed to among different boroughs.
5. ALL: General comparison across the properties.

For each property now, we calculate 45 ratios by dividing r1-r9 with the group average of r1-r9 across above mentioned 5 groups.

For example, For a property with TAXCLASS A, Borough 2, Zipcode 10020,

We calculate the average values of r1-r9 for all properties with Tax Class A.
We calculate the average values of r1-r9 for all properties with Borough 2.
We calculate the average values of r1-r9 for all properties with Zipcode 10020.
We calculate the average values of r1-r9 for all properties having their Zipcode starting with 100.
We calculate the average values of r1-r9 for all properties.

For each of the variables r1-r9, we now calculate ratios with respect to their corresponding r1-r9s across all of the five groups, thus creating 45 variables for each property.

Aggregators	**avg(rx) = Average Value of rx over ZIP5 = 10020**	**avg(rx) = Average Value of rx over ZIP3 = 100**	**avg(rx) = Average value of rx over TAXCLASS = A**	**avg(rx) = Average value of rx over Borough = 2**	**avg(rx) = Average value of rx across all properties**
r1 Variables	r1_zip5 = r1/avg(r1)	r1_zip3 = r1/avg(r1)	r1_taxclass = r1/avg(r1)	r1_borough = r1/avg(r1)	r1_all = r1/avg(r1)
r2 Variables	r2_zip5 = r2/avg(r2)	r2_zip3 = r2/avg(r2)	r2_taxclass = r2/avg(r2)	r2_borough = r2/avg(r2)	r2_all = r2/avg(r2)
r3 Variables	r3_zip5 = r3/avg(r3)	r3_zip3 = r3/avg(r3)	r3_taxclass = r3/avg(r3)	r3_borough = r3/avg(r3)	r3_all = r3/avg(r3)
r4 Variables	r4_zip5 = r4/avg(r4)	r4_zip3 = r4/avg(r4)	r4_taxclass = r4/avg(r4)	r4_borough = r4/avg(r4)	r4_all = r4/avg(r4)
r5 Variables	r5_zip5 = r5/avg(r5)	r5_zip3 = r5/avg(r5)	r5_taxclass = r5/avg(r5)	r5_borough = r5/avg(r5)	r5_all = r5/avg(r5)
r6 Variables	r6_zip5 = r6/avg(r6)	r6_zip3 = r6/avg(r6)	r6_taxclass = r6/avg(r6)	r6_borough = r6/avg(r6)	r6_all = r6/avg(r6)

r7 Variables	r7_zip5 = r7/avg(r7)	r7_zip3 = r7/avg(r7)	r7_taxclass = r7/avg(r7)	r7_borough = r7/avg(r7)	r7_all = r7/avg(r7)
r8 Variables	r8_zip5 = r8/avg(r8)	r8_zip3 = r8/avg(r8)	r8_taxclass = r8/avg(r8)	r8_borough = r8/avg(r8)	r8_all = r8/avg(r8)
r9 Variables	r9_zip5 = r9/avg(r9)	r9_zip3 = r9/avg(r9)	r9_taxclass = r9/avg(r9)	r9_borough = r9/avg(r9)	r9_all = r9/avg(r9)

Dimensionality Reduction

Before building the unsupervised learning model to identify outliers, it is imperative to ensure that the distributions of all the variables are comparable, there is no strong correlation within different variables and we only have the optimal number of dimensions where the majority of the variability is covered to reduce complexity. Thus, to address this we scale and reduce the number of variables using Principal Component Analysis (PCA).

The procedure applied to the city of New York dataset is as follows:

Step 1: Z-Scaling the variables

At this point in our analysis, the 45 expert variables are on different scales. Therefore, before Z-Scaling is necessary prior to PCA. We use z-scaling, which subtracts the mean of the field column from the field and dividing by the standard deviation of that column. After z-scaling, all fields are centered around a mean of 0 and are on a similar scale.

The Z-Scaling can be done as follows:

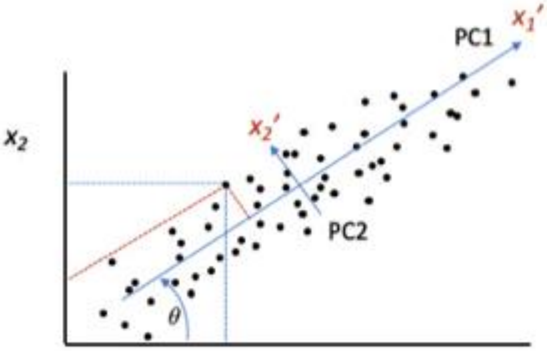
$$z = \frac{x - \mu}{\sigma}$$

where,
x is the value of the variable for a given field
μ is the average of the variable across the column
σ is the standard deviation of the variable across the column

Step 2: Principal Component Analysis

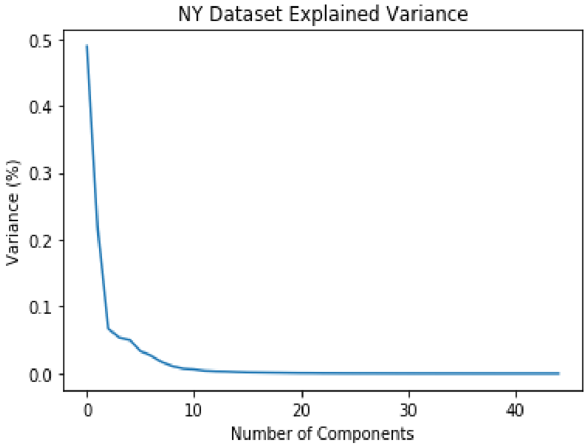
To eliminate the correlation among different variables as well as identifying dimensions across which there is maximum variability, we use the principal component analysis (PCA) technique.

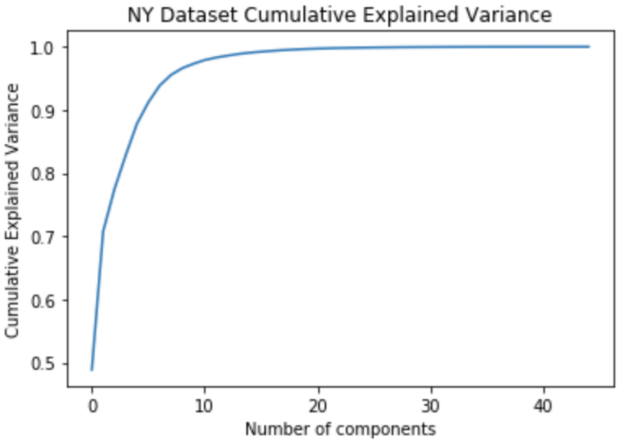
The PCA technique finds the dominant directions in the data and rotates the co-ordinate system along these directions. The components identified are orthogonal. Hence, our 45 expert variables would be converted into a set of linearly uncorrelated variables creating the principal components of the data. The algorithm works by identifying the first principle component (PC1) where there is maximum variability. The next direction (PC2) is orthogonal to PC1 where there is maximum variability with respect to PC1. This approach works so on and so forth until the entire variability is covered across the dataset.



An important advantage of using this technique is that, typically, within 7-10 PCs, almost 80% of the variability is accounted for. We can, therefore, only use those PCs in our unsupervised model as the rest of the dimensions do not explain much of the variation.

By applying PCA to the Z-Scaled NY property dataset, we choose the top 8 principal components that account for ~90% of the variance in our data, while discarding the rest. Below are the scree and cumulative plots depicting the amount of variance accounted for by each PC:





Step 3: Subsequent Z-Scale

After obtaining a linear combination of variables in our PCA, the principle components are not comparable. Therefore, we z-scale the data once more in order to scale it around the origin. This additional transformation allows for an easier distance calculation in our next steps; with z-scaled data (mean of 0 and standard deviation of 1), we can simply measure the distance from the origin when determining whether or not a data point is an anomaly in comparison to the others.

Algorithms

The data after dimensionality reduction has 1070994 records of 8 z-scaled principal components.

We calculate scores for anomaly detection using 2 different methods:

Score 1: Combine the z-scores with a heuristic fraud score formula

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

Using the above generalized distance formula, with a chosen value of n=2 on our data (Euclidean distance), allows us to use the distance of each data point from the origin as a fraud score. The data points that have the farthest distance from the origin are more likely to be anomalous in some manner and scored as high using this metric. Thus, it is a reasonable score to use.

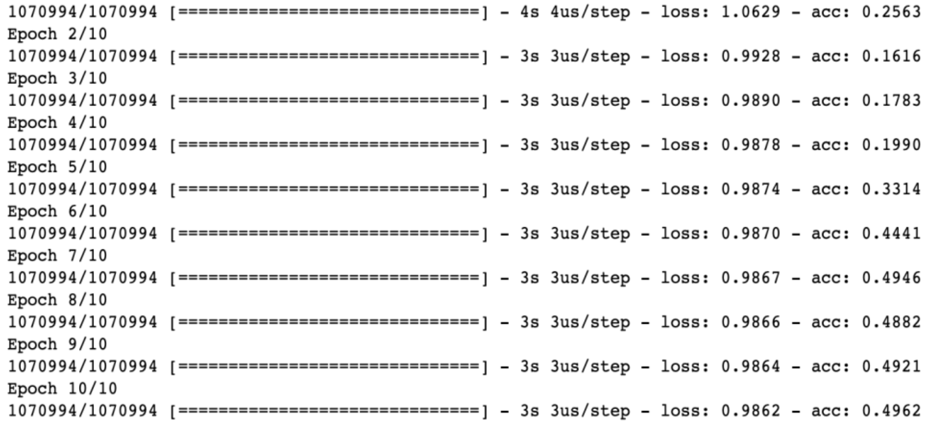
Score 2: Autoencoder

An autoencoder uses a non-linear neural network to weigh and reproduce each record. By weighing every field with the other data points, an autoencoder is able to successfully reproduce it if the field is similar to the bulk of the data.

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

The above formula shows the difference in the input and output layer that helps us in identifying fraud. If the autoencoder is unable to accurately replicate the original record, its replication will have a lower accuracy or a large difference between the original input vector and model output vector and thus a higher Euclidean distance. Such a value indicates that the record in question may be anomalous since the autoencoder has difficulty in replicating it. This measure of reproduction error, therefore, can serve as a reasonable fraud score.

An example of the autoencoder output and accuracy measure for each record can be found below:



```
1070994/1070994 [==============================] - 4s 4us/step - loss: 1.0629 - acc: 0.2563
Epoch 2/10
1070994/1070994 [==============================] - 3s 3us/step - loss: 0.9928 - acc: 0.1616
Epoch 3/10
1070994/1070994 [==============================] - 3s 3us/step - loss: 0.9890 - acc: 0.1783
Epoch 4/10
1070994/1070994 [==============================] - 3s 3us/step - loss: 0.9878 - acc: 0.1990
Epoch 5/10
1070994/1070994 [==============================] - 3s 3us/step - loss: 0.9874 - acc: 0.3314
Epoch 6/10
1070994/1070994 [==============================] - 3s 3us/step - loss: 0.9870 - acc: 0.4441
Epoch 7/10
1070994/1070994 [==============================] - 3s 3us/step - loss: 0.9867 - acc: 0.4946
Epoch 8/10
1070994/1070994 [==============================] - 3s 3us/step - loss: 0.9866 - acc: 0.4882
Epoch 9/10
1070994/1070994 [==============================] - 3s 3us/step - loss: 0.9864 - acc: 0.4921
Epoch 10/10
1070994/1070994 [==============================] - 3s 3us/step - loss: 0.9862 - acc: 0.4962
```

Final score

Now that we have 2 different fraud scores that yield important insights on which records are anomalous, it is necessary to combine them into one all-encompassing fraud score. The method that used to combine the fraud scores generated from the 2 algorithms listed above is quantile binning.

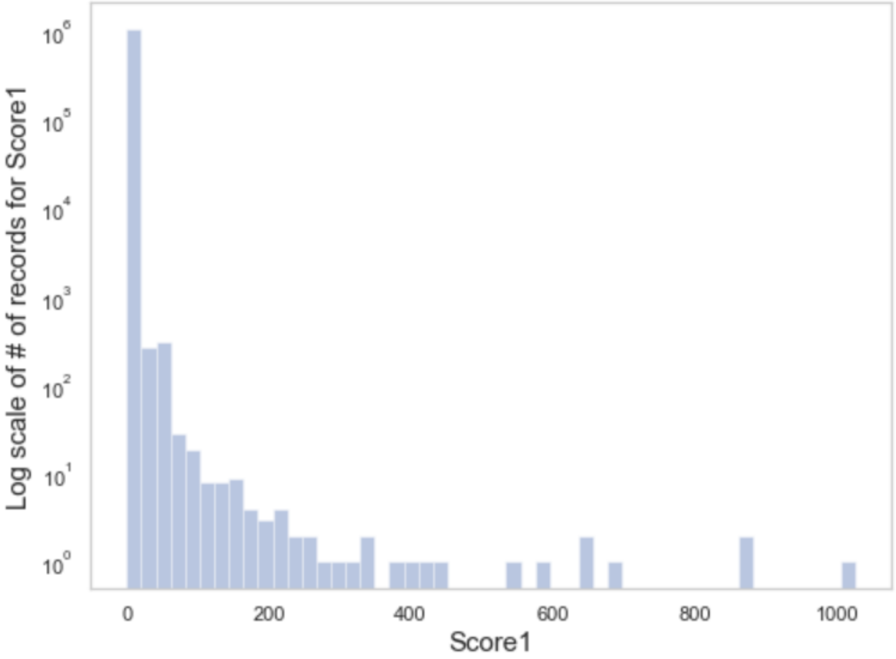
The procedure to obtain the final score is as follows:

1. For each fraud score calculated, we sort the records and bin them so that an equal count of records is in each bin. At this point, the score is replaced with the bin number. When equating the number of bins to the number of records (extreme quantile binning), the sorting will simply replace each fraud score with its rank order.
2. After performing quantile binning on both unsupervised fraud scores, taking a weighted average of the 2 scores for each record provides a scaled manner through which these scores can be combined. Through this step, we are taking both scores into account with appropriate weightage in order to correctly identify anomalous records. In this case, we gave equal weights to both the scores.
3. Once each record has a final score, we sort it to get the top 10 records with the highest scores, as these records have the greatest deviation from the rest and therefore can be viewed as potential anomalies. We then look at extreme z-scale values before the PCA for the individual records to determine what attributes result in this high fraud score. The reason for not using z-scale principle components for analysis is that PCA makes the variables lose their interpretability.

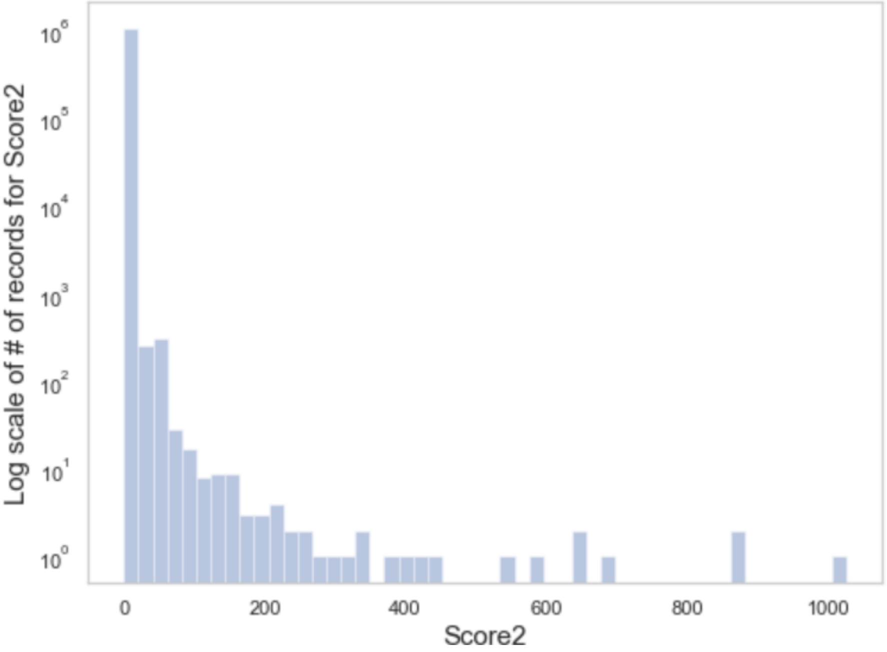
Ultimately, the generation of 2 fraud scores through the heuristic algorithm and the autoencoder as well as subsequent combining of these scores through quantile binning provides an effective pipeline to observe anomalies in the city of NY property dataset and detect potential property tax frauds.

Results

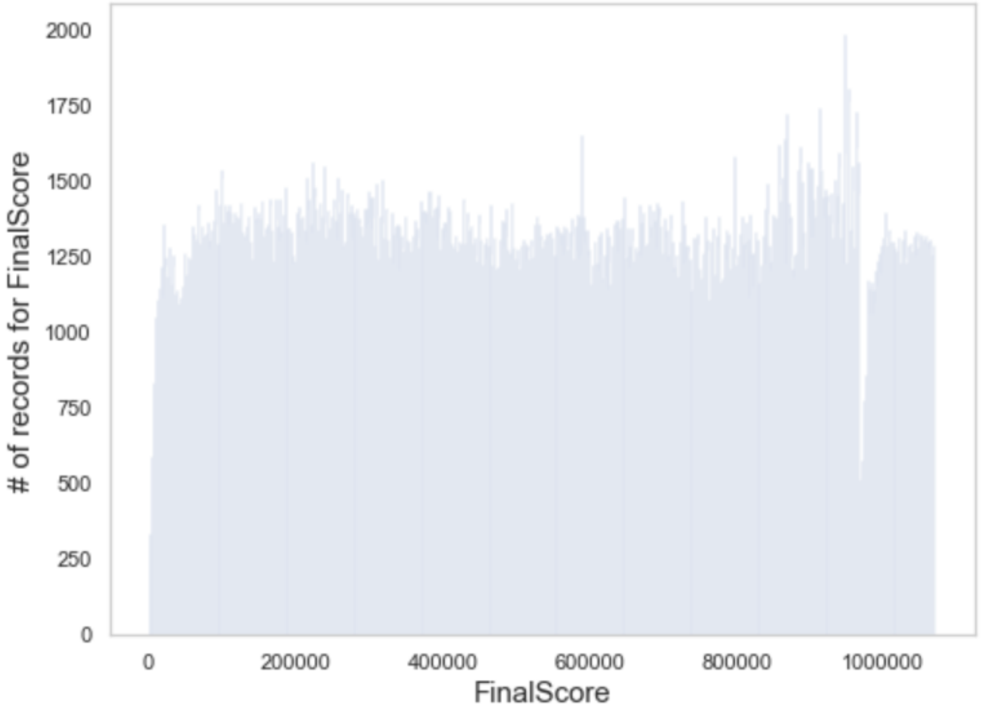
Score 1 Distribution: Combining z-scores with a heuristic fraud score formula



Score 2 Distribution: Autoencoder



Final Score Distribution after Quantile Binning and Weighted Average Combination:



After plotting the distributions of both individual scores and the combined final score, we will demonstrate that our algorithmic anomaly detection system does, in fact, detect anomalous records by subsetting the top 10 most anomalous records, or records that have the highest fraud score. For each record, below is a description of why it is potentially anomalous. All valuations referred to in the below descriptions correspond to the record's value in the FULLVAL field and are rounded for estimation.

Analysis of Top 10 Anomalous Records as Generated by our Fraud System

A table of these records can be found below:

RECORD	PC Z1	PC Z2	PC Z3	PC Z4	PC Z5	PC Z6	PC Z7	PC Z8	Fraud Score1	Fraud Score2	Final Score
632816	592.85	-100.69	718.12	164.98	-369.96	88.07	48.38	-43.31	1070994	1070994	1070994
67129	555.72	-59.96	-588.73	10.68	-206.47	69.66	194.93	-51.71	1070993	1070993	1070993
821853	45.06	706.91	23.46	-387.17	-275.54	-29.12	9.35	128.18	1070992	1070992	1070992
750816	15.04	183.26	-48.88	546.97	203.63	317.40	53.11	65.24	1070991	1070991	1070991
585119	282.53	-38.83	-175.94	-70.85	140.37	-131.23	509.38	59.69	1070990	1070990	1070990
565392	200.34	333.43	81.52	-117.96	391.73	184.74	-2.60	-222.97	1070989	1070989	1070989
1067360	20.70	295.65	-44.58	492.05	103.94	68.29	27.46	30.40	1070988	1070988	1070988
585118	139.91	-24.73	-236.40	5.84	-125.82	81.96	-292.21	-325.38	1070987	1070987	1070987
920628	75.32	-13.11	-47.14	1.05	-9.47	-12.03	-166.66	397.32	1070986	1070986	1070986
917944	22.77	325.88	18.80	-184.35	-40.21	169.95	10.77	-19.96	1070985	1070985	1070985

Record details for the anomalous records are given below:

RECORD	OWNER	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	STADDR	ZIP
632816	864163 REALTY, LLC	157	95	1	1	1	2930000	1318500	1318500	86-55 BROADWAY	11373
67129	CULTURAL AFFAIRS	840	100	17	38	2	6150000000	2668500000	2.77E+09	1000 5 AVENUE	10028
821853	CNY/NYC TA	2	1	29	53	2	593000	17550	31890	87 DRIVE	11432
750816	M FLAUM	1	1	20	36	2	468000	13690	22956	VLEIGH PLACE	11367
585119	CNY/NYC TA	218.5	372	1	1	2	2797500	1258875	1258875		11101
565392	U S GOVERNMENT OWNRD	117	108	54	121	1	4326303700	1946836665	1.95E+09	FLATBUSH AVENUE	11234
1067360		1	1	36	45	2	836000	28800	50160	20 EMILY COURT	10307
585118	NEW YORK CITY ECONOMI	298	402	1	1	20	3443400	1549530	1549530	28-10 QUEENS PLAZA SOUTH	11101
920628	PLUCHENIK, YAAKOV	91	100	1	1	2	1900000	9763	75763	7-06 ELVIRA AVENUE	11691
917944	NYC AGENCY PROPERTIES	5	7	193	226	3	17378000	7820100	7820100	SOUTH CONDUIT AVENUE	11434

Below are the observations using Z-scale values before PCA and the above record details:

Record #1: 632816
The property is owned by 864163 REALTY, LLC and contains 83 luxury rental apartments. It has lot dimensions for front and depth as 157 and 95 feet respectively. However, the building front and building depth are still reported as 1. The property had a valuation of $2.93 million. Having such a small value for building front and building depth, despite having such a large lot area with 83 luxury apartments makes this record different from other records with a similar number of units and therefore anomalous.

Record #2: 67129
The property is located on 5th avenue. It is the Metropolitan Museum of Art owned by Cultural Affairs. It is one of the largest museums in the world, which explains the high standard deviations for all of the 9 expert variables relating to land area and valuation in comparison to the rest of the data.

Record #3: 821853
This property is owned by the NYCTA (New York City Transport Authority). The lot front and lot depth are 2 feet and 1 foot respectively while building front and building depth are 29 and 53 feet respectively. The valuation of the property is $593,000 and is slightly below the mean. The sizes for lot front and lot depth do not make sense for a property of that valuation and building dimensions, making this record anomalous.

Record #4: 750816
The Lot front and Lot depth are reported as 1 foot while the building front and depth dimensions are 20 and 36 feet respectively and the full value of the property is estimated to be $468,000. The street address is Vleigh place. The name of the owner is M. Flaum. Like Record #3, the sizes for lot front and lot depth do not make sense for a property of that valuation and building dimensions, making this record anomalous.

Record #5: 585119
The building front and building depth for this record are also reported as 1. The property has a valuation of $2.79 million. Sizes for lot front and lot depth are 218 and 372 feet respectively that don't make sense for a property of that valuation, the absence of a street address makes this record anomalous in comparison to the others.

Record #6: 565392
The street address for this record is Flatbush Avenue in Brooklyn. The valuation of this property is $4.32 billion; therefore, the values of building front and building depth, 54 and 121 feet respectively, are not reasonable values for a property with such a high valuation. The property is US government owned and has a substantially high value for all of the 9 variables in comparison to the others, which is likely why the analytics pipeline gave it a high anomaly score.

Record #7: 1067360
This property has a street address of 20 Emily Court. It is a 2 unit and 2 story building, but the lot front and lot depth are still reported as 1 foot which leads to expert variables such as r1, r4, and r7 being around 100 standard deviations higher than the origin. Having such low values for lot front and depth despite having 2 units and being multi-story makes this record stand out as anomalous.

Record #8: 585118
This property has a street address of 28-10 Queens Avenue in Brooklyn. It has twin towers with 27-stories. Despite such a large area, the building front and building depth values are still reported as 1 foot; it does not make sense for twin tower properties to have such a small value for building area measurements. This causes the value for the variables r2, r3, r5, r6, r8 and r9 to be extremely high. The property also has a valuation of around $3.4 million, so it is strange to have a building size that small and this record can hence be viewed as anomalous.

Record #9: 920628
7-06 Elvira Avenue is a house located in the Far Rockaway neighborhood in Queens, NY. This property was built in 2008 and has 2 stories and 1 unit. Like the other anomalous records, the building front and building depth have a value of 1 which is suspicious compared to a fairly large

lot area. The property also has a valuation of $1.9 million, which does not make sense given the area measurements of 1 foot each.

Record #10: 917944

This property has a valuation of $17 million. The lot front and lot depth values are 5 and 7 feet, respectively while the building front and depth are 193 and 226 feet respectively. The property is owned by NYC Agency properties and the street address is South Conduit Avenue. A property with such a high valuation and such small dimensions for lot front and depth is strange, making this property stand out as anomalous.

Conclusion:

Given the City of New York Property dataset containing over one million records, our team has followed a systematic process of detecting anomalies through the use of unsupervised machine learning methods. We began this process by understanding the data and then cleaning the data, filling in both Null and 0 values in the dataset, through various grouping and aggregation methods as mentioned in the Data Cleaning section above. After cleaning the data, we moved on to feature engineering using relevant fields from the original dataset. These variables are created based on the following 5 sizes and valuation fields:

BLDFRONT, BLDDEPTH, LTFRONT, LTDEPTH, STORIES

Using these 5 fields, we created 3 new fields:

LOTAREA, BLDAREA, and BLDVOLUME

Each of these 3 new variables is then divided by each of the 3 assessed value fields from the original data (FULLVAL, AVLAND, and AVTOT), resulting in 9 expert variables (labeled as r1 through r9).

Next, these 9 expert variables are grouped 5 times, by:

BOROUGH, ZIP5 (5 digits), ZIP3 (3 digits), TAXCLASS and ALL

and scaled by the average of each group, resulting in 45 total variables per record. These variables are then normalized through Z-Scaling in order to allow for interpretability across all variables. We then conducted Principal Component Analysis in order to identify dimensions with high variability and retained the first 8 principal components as they explained a majority of the variability. Next, we used the Z-scale method once again to standardize across all variables, allowing us to prepare the data for algorithm implementation.

After scaling the data and reducing dimensionality, we are able to implement our algorithms utilizing two methods: first, through a Heuristic Score Model and then with an Autoencoder, as tools to detect anomalies. Each method resulted in a score (Score 1 and Score 2), which are then combined by taking a rank order (quantile binning) and averaging the 2 scores, allowing us to arrive at the Final Score. These Final Scores are then sorted, and the top 10 are explored on a deeper level as anomalous records.

We found that many of the top 10 anomalous records detected by our analytics pipeline reported miniscule values for building front and depth size, while still having valuations that topped millions. Therefore, we could verify that our designed pipeline was indeed successful in detecting strange and anomalous records in the NY property dataset.

Given additional time, the following things could be explored for better results:

1. A deeper level analysis of the anomalous records as identified by our pipeline. Rather than just identifying what makes each record anomalous for the top 10, doing so for the top 100 or 1000 records and grouping them by anomaly "type" can provide more insights into which nature of property tax anomaly is the most prevalent (ie. anomalies with strange area values or strange valuations - these two types of anomalies are different). Such insights could contribute to a greater understanding of the nature of potential property tax fraud that is most likely in New York City.
2. Develop more advanced, detailed logic for the methodology used for filling null and 0 values in the data cleaning portion of this analysis. Doing so would allow for improved accuracy in how these values are being filled, therefore resulting in overall improved accuracy for both unsupervised learning algorithms and the final list of anomalous records.

Ultimately, the algorithmic system created for the city of New York by DSO 562's Group 8 effectively and successfully utilized unsupervised learning methods to detect anomalous property records from the City of New York's Property Dataset.

Appendix: Data Quality Report

The following Data Quality Report (DQR) summarizes and provides a brief description of the NY Property dataset.

1. Data Description

The dataset provided is an open data from NYC city government and consists of property assessments for the purpose of calculating Property Tax, Grant eligible properties, Exemptions and/or Abatements. The data has been collected and entered into the system by various city employees including Property Assessors, Property Exemption specialists, ACRIS reporting, Department of Building reporting, etc.

Dataset Name	NY property data
Data Source	Open data of NYC city government, department of finance
Time Period	November 2010
# of Columns	32
# of Records	1,070,994

2. Summary

The overall dataset has 14 Numerical and 18 Categorical variables (Includes time variables like Period and Year) present. The detailed description is provided below:

2.1 Numerical Summary

Field Name	# of Records with a value	% Populated	# of Unique Values	# of records with 0 value	Mean	Standard Deviation	Min	Max
LTFRONT	1070994	100.0%	1297	169108	36	74	0	9999
LTDEPTH	1070994	100.0%	1370	170128	89	76	0	9999
STORIES	1014730	94.7%	112	0	5	8	1	119
FULLVAL	1070994	100.0%	109324	13007	874264	11582431	0	6150000000
AVLAND	1070994	100.0%	70921	13009	85068	4057260	0	2668500000
AVTOT	1070994	100.0%	112914	13007	227238	6877529	0	4668308947
EXLAND	1070994	100.0%	33419	491699	36424	3981576	0	2668500000
EXTOT	1070994	100.0%	64255	432572	91187	6508403	0	4668308947
BLDFRONT	1070994	100.0%	612	228815	23	36	0	7575
BLDDEPTH	1070994	100.0%	621	228853	40	43	0	9393

AVLAND2	282726	26.4%	58592	0	246236	6178962	3	2371005000
AVTOT2	282732	26.4%	111361	0	713911	11652528	3	4501180002
EXLAND2	87449	8.2%	22196	0	351236	10802212	1	2371005000
EXTOT2	130828	12.2%	48349	0	656768	16072510	7	4501180002

2.2 Categorical Summary

Field Name	# of Records with a value	% Populated	# Unique Values	Most Common Field Value
RECORD	1070994	100.0%	1070994	NA
BBLE	1070994	100.0%	1070994	NA
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

3. Data Field Exploration

Field1

Field Name: RECORD (Type: Categorical)

Description: Unique identifier of each data record

Field2

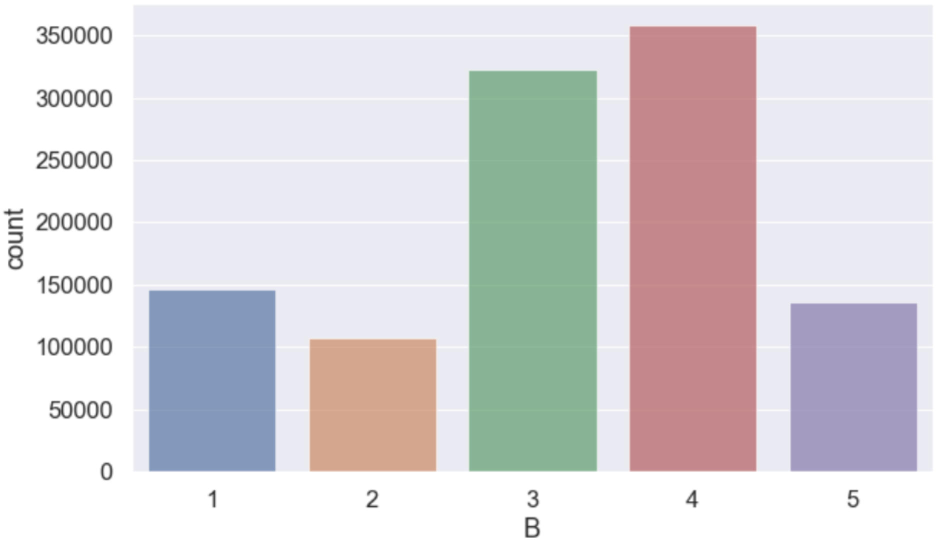
Field Name: BBLE (Type: Categorical)

Description: Concatenation of borough code, block code, Unique # within borough/block and easement. This code is unique to each property.

Field3

Field Name: B (Type: Categorical)

Description: Borough Code



Field4

Field Name: BLOCK (Type: Categorical)

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

Field5

Field Name: LOT (Type: Categorical)

Description: Unique # within borough/block

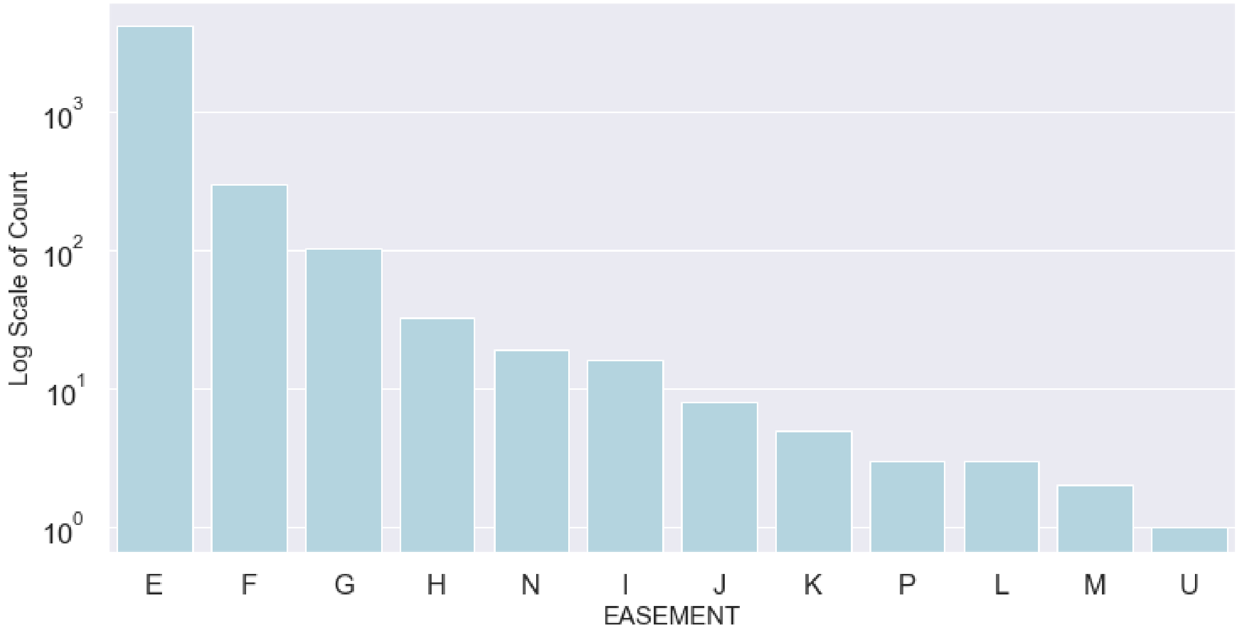
Top 10 Field Values

LOT	Count
1	24367
20	12294
15	12171
12	12143
14	12074
16	12042
17	11982
18	11979
25	11949
21	11840

Field6

Field Name: EASEMENT (Type: Categorical)

Description: Describes Easement



Field7

Field Name: OWNER (Type: Categorical)

Description: Property Owner's Name

Top 10 Field Values

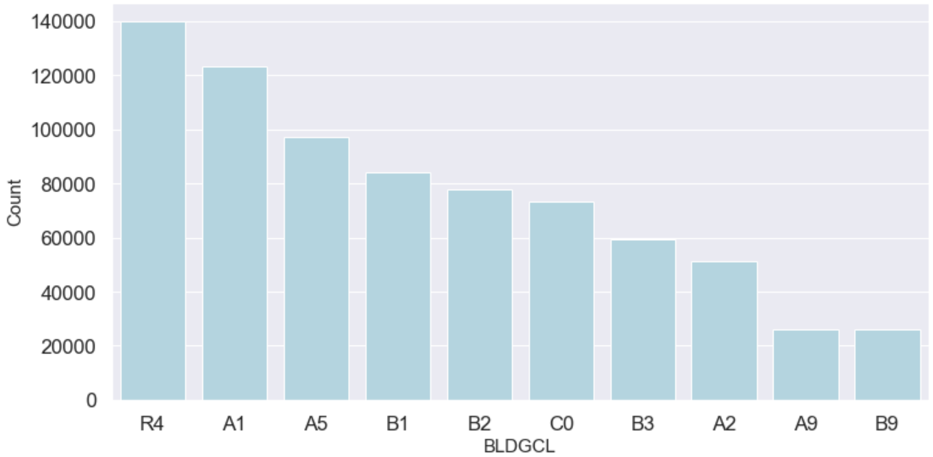
OWNER	Count
PARKCHESTER PRESERVAT	6020
PARKS AND RECREATION	4255
DCAS	2169
HOUSING PRESERVATION	1904
CITY OF NEW YORK	1450
DEPT OF ENVIRONMENTAL	1166
BOARD OF EDUCATION	1015
NEW YORK CITY HOUSING	1014
CNY/NYCTA	975
NYC HOUSING PARTNERSH	747

Field8

Field Name: BLDGCL (Type: Categorical)

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

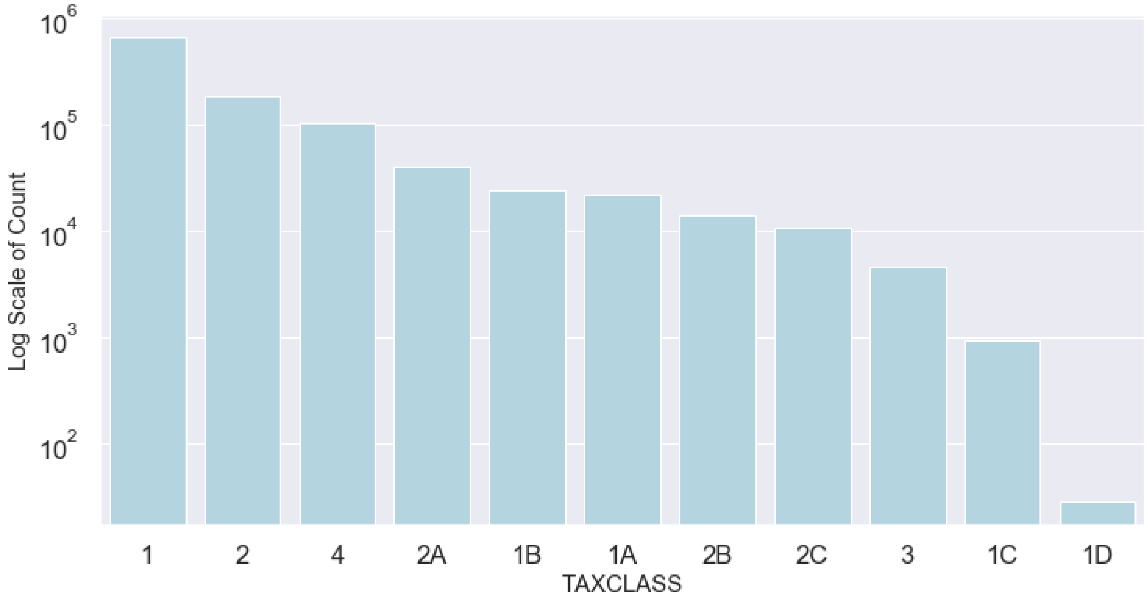
Top 10 Field Values



Field9

Field Name: TAXCLASS (Type: Categorical)

Description: Current Property Tax Class Code (NYS Classification)



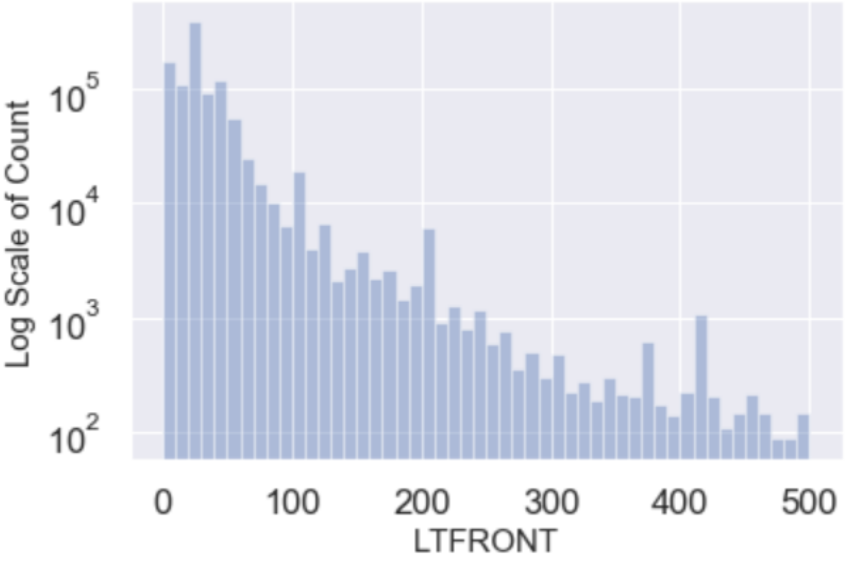
Field10

Field Name: LTFRONT (Type: Numerical)

Description: Lot Frontage in feet

Outliers removed: > 500

Data in the histogram is 99.76% populated.



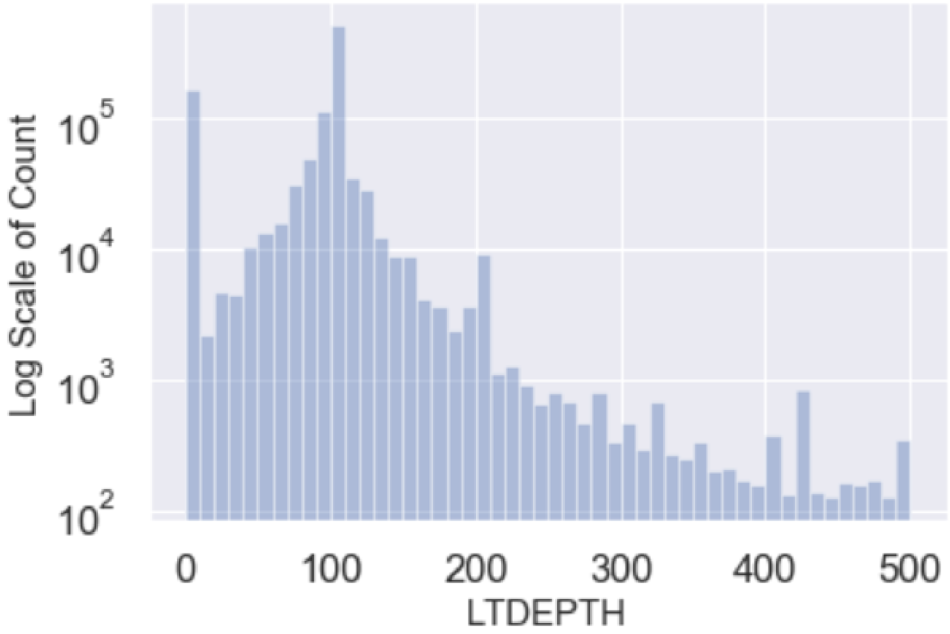
Field11

Field Name: LTDEPTH (Type: Numerical)

Description: Lot Depth in feet

Outliers removed: > 500

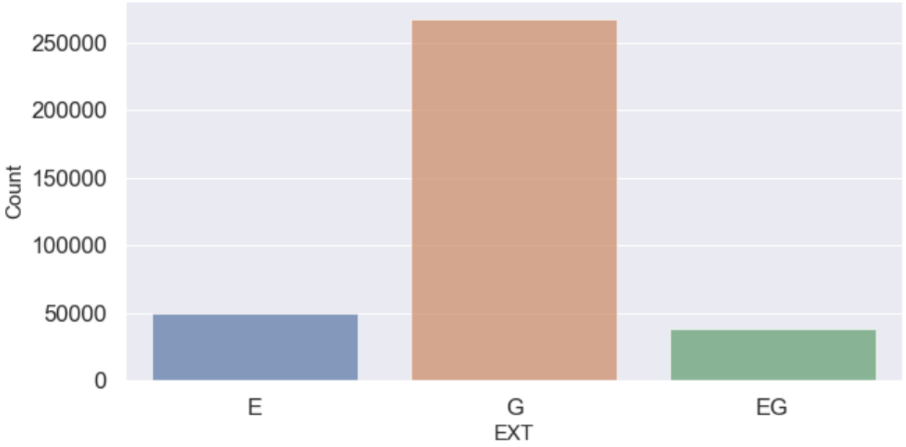
Data in the histogram is 99.68% populated.



Field12

Field Name: EXT (Type: Categorical)

Description: Extension. “E”- extension, ”G”- garage, ”EG” - extension and garage



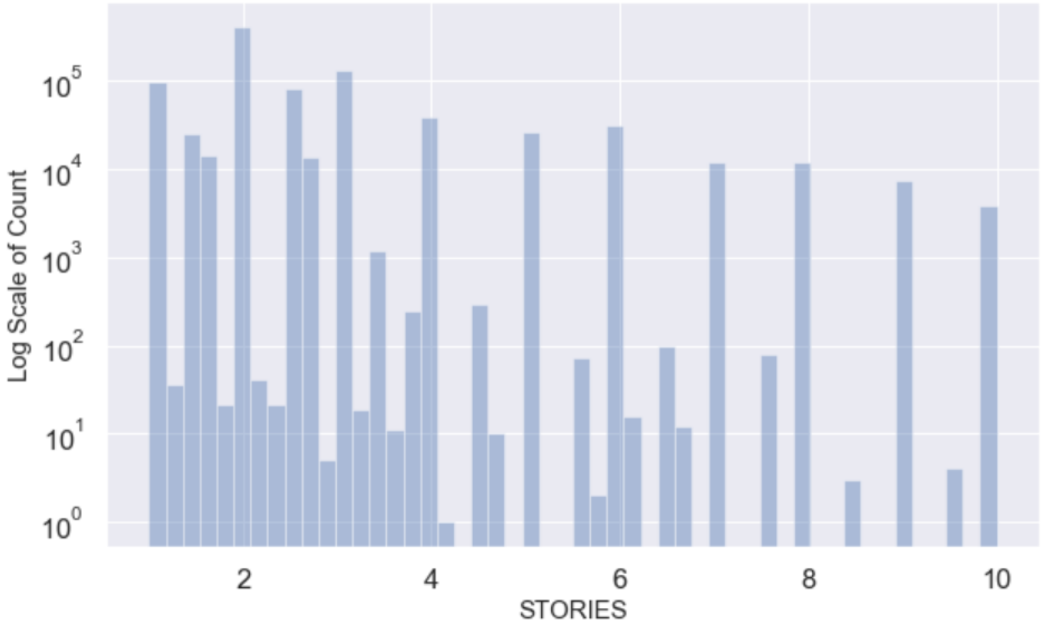
Field13

Field Name: STORIES (Type: Numerical)

Description: The number of stories in the building (# of floors)

Outliers removed: > 10

Data in the histogram is 89.6% populated.



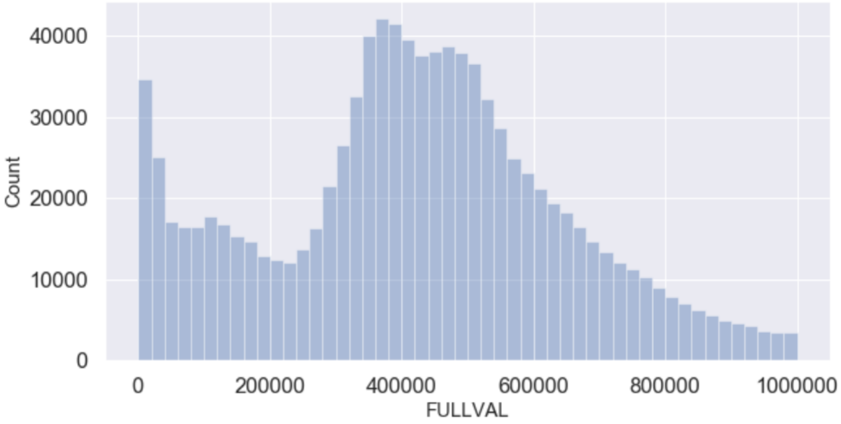
Field14

Field Name: FULLVAL (Type: Numerical)

Description: Total Market Value

Outliers removed: > 1000000

Data in the histogram is 91.3% populated.



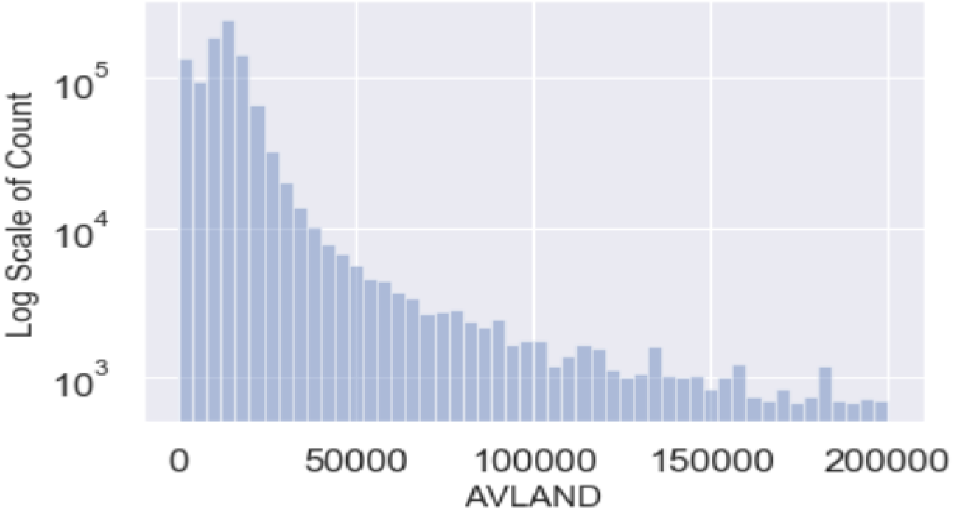
Field15

Field Name: AVLAND (Type: Numerical)

Description: Total Market Value of the land

Outliers Removed: > 200000

Data in the histogram is 90.4% populated.



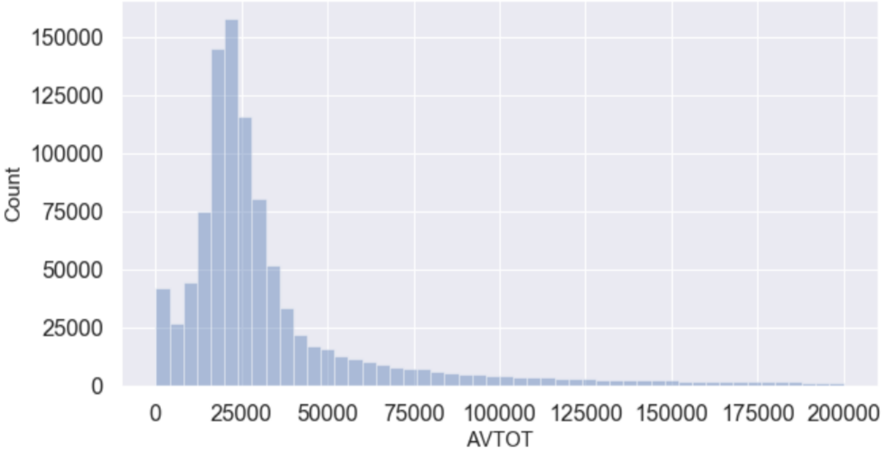
Field16

Field Name: AVTOT (Type: Numerical)

Description: Total Market Value

Outliers Removed: > 200000

Data in the histogram is 91.6% populated.



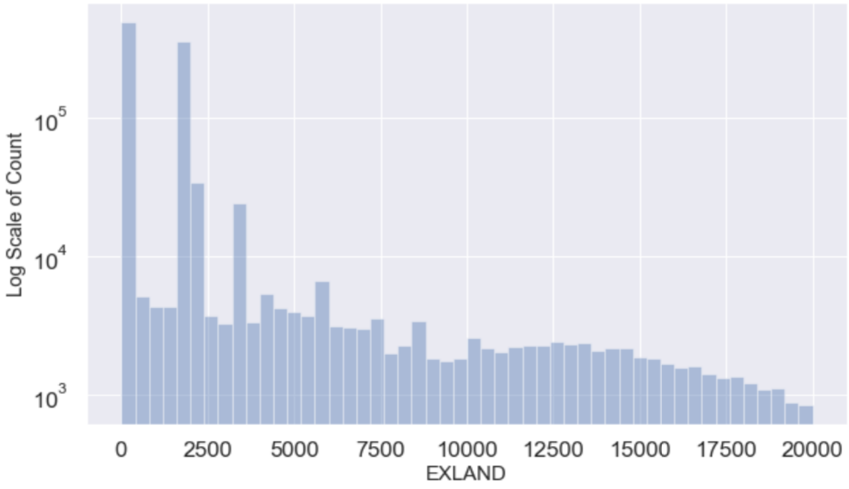
Field17

Field Name: EXLAND (Type: Numerical)

Description: Tentative Transitional Exempt Land Value

Outliers Removed: > 20000

Data in the histogram is 96.8% populated.



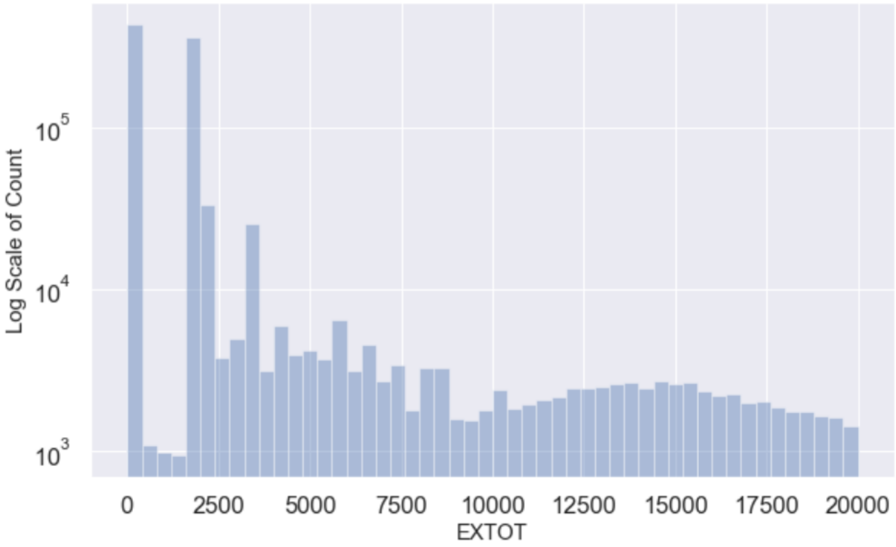
Field18

Field Name: EXTOT (Type: Numerical)

Description: Tentative Transitional Exempt Total Value

Outliers Removed: > 20000

Data in the histogram is 90.4% populated.



Field19

Field Name: EXCD1 (Type: Categorical)

Description: Exempt Code

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

Field20

Field Name: STADDR (Type: Categorical)

Description: Street name of 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
30 WEST 63 STREET	562

Field21

Field Name: ZIP (Type: Categorical)

Description: Postal Zip Code of the property

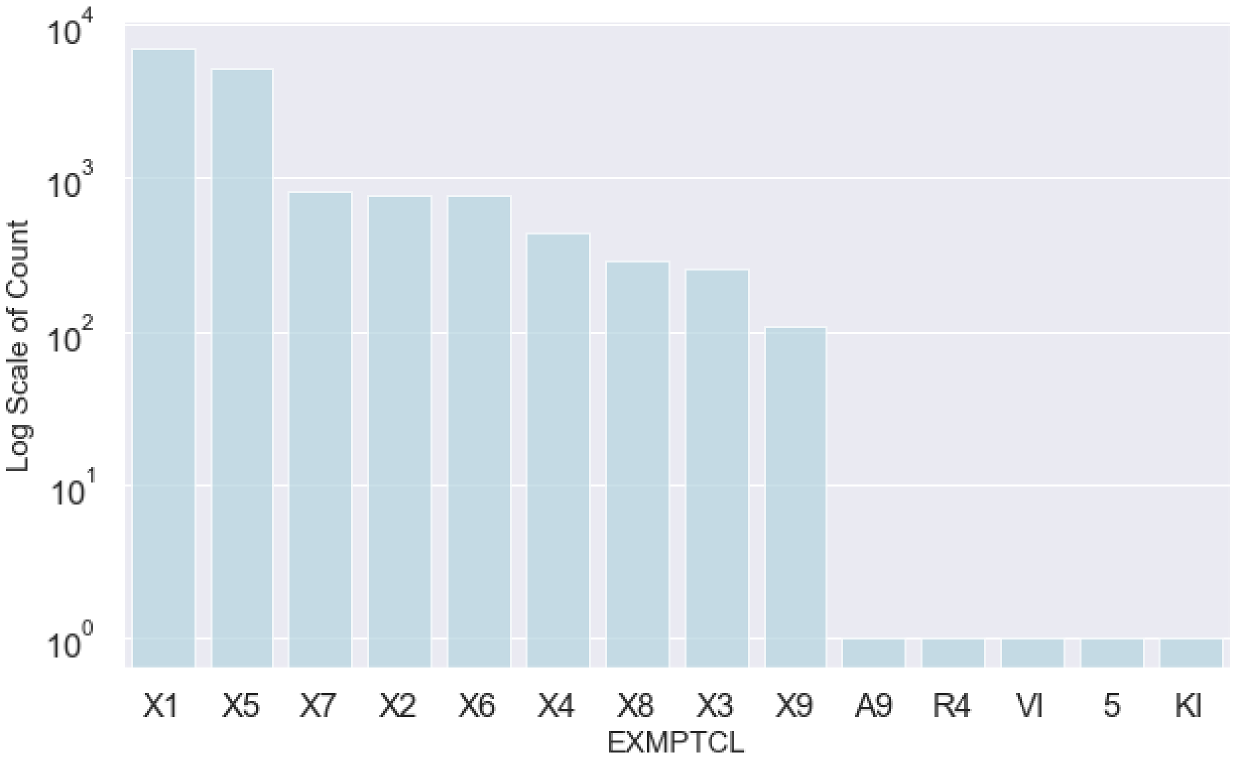
Top 10 Field Values

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

Field22

Field Name: EXMPTCL (Type: Categorical)

Description: Exempt Class used for fully exempt properties only



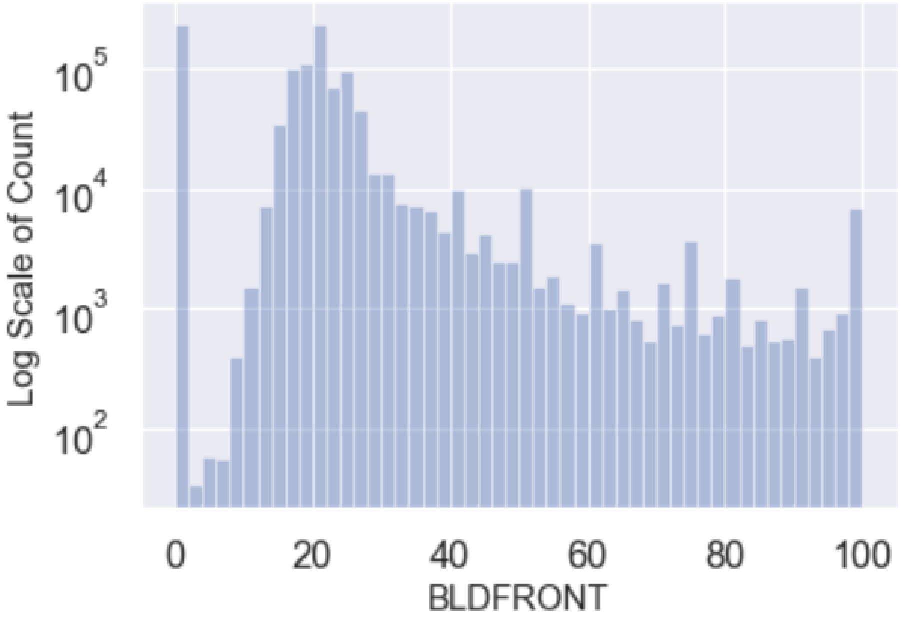
Field23

Field Name: BLDFRONT (Type: Numerical)

Description: Building Frontage in feet

Outliers Removed: > 100

Data in the histogram is 97.3% populated.



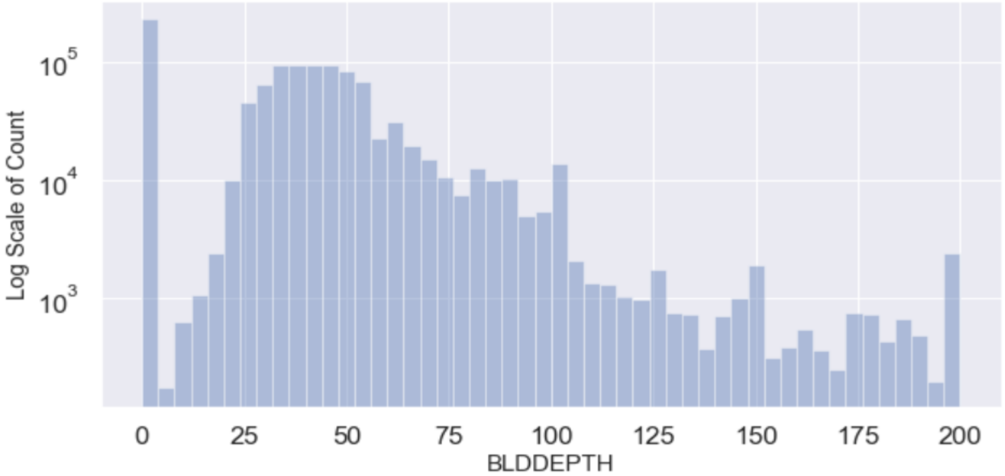
Field24

Field Name: BLDDEPTH (Type: Numerical)

Description: Building Depth in feet

Outliers Removed: > 200

Data in the histogram is 99.6% populated.



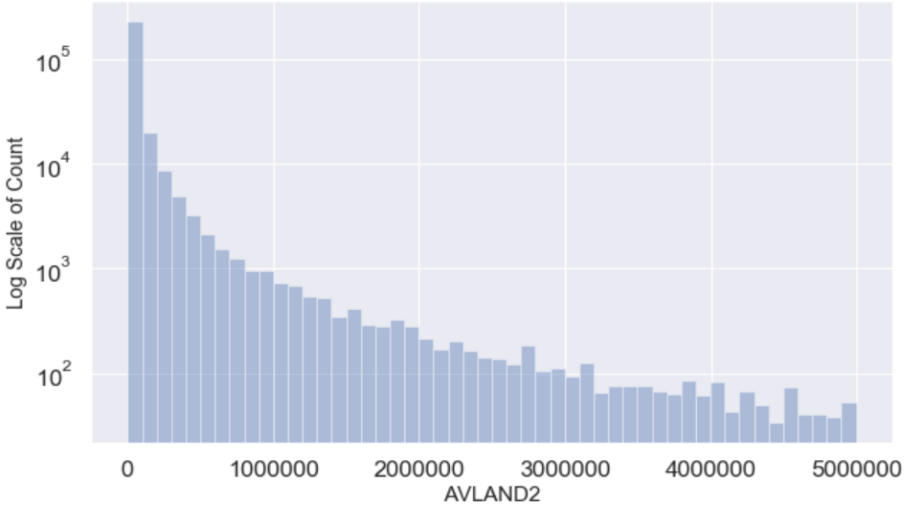
Field25

Field Name: AVLAND2 (Type: Numerical)

Description: New Market Value of the land

Outliers Removed: > 5000000

Data in the histogram is 99.3% populated.



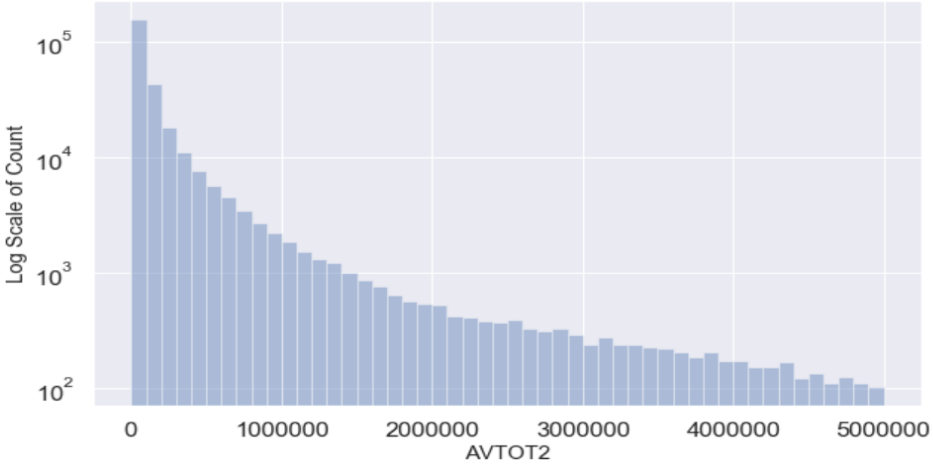
Field26

Field Name: AVTOT2 (Type: Numerical)

Description: New Total Market Value

Outliers Removed: > 5000000

Data in the histogram is 97.9% populated.



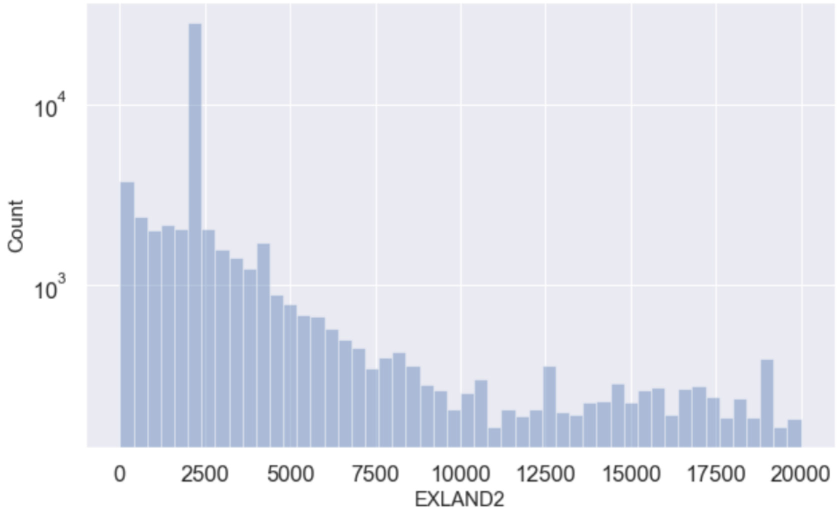
Field27

Field Name: EXLAND2 (Type: Numerical)

Description: New Exempt Land Value

Outliers Removed: > 20000

Data in the histogram is 70.3% populated.



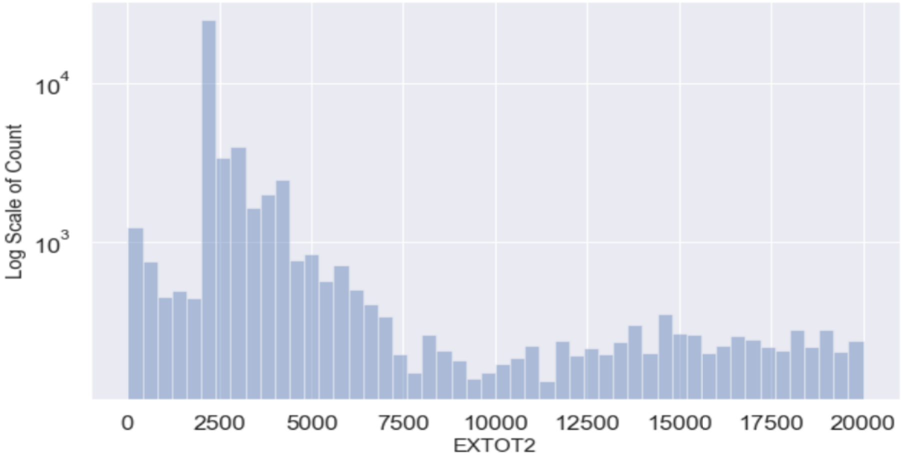
Field28

Field Name: EXTOT2 (Type: Numerical)

Description: New Exempt Total Value

Outliers Removed: > 20000

Data in the histogram is 40.4% populated.



Field29

Field Name: EXCD2 (Type: Categorical)

Description: New Exempt Code

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

Field30

Field Name: PERIOD (Type: Categorical (Date/Time))

Description: All the values are given as ‘Final’

Field31

Field Name: YEAR (Type: Categorical (Date/Time)

Description: All the values are given as 2010/11

Field32

Field Name: VALTYPE (Type: Categorical)

Description: All the values are given as AC-TR