Fraud Detection for the NY property data

DSO562 Fraud Analytics

Team 2:
Andrew Huan-Wei Chang
Bei Wang
Gregy Thomas
Xinyu Bao
Yi-Chen Lin
YingHong Lin

Table of Contents

Part I. Executive Summary

This report describes the process where we assessed the New York City property data using unsupervised method. We used a combination of statistical packages in Python and R. We first filled in missing values and created 45 expert variables. Then we performed Principal Component Analysis (PCA) to reduce dimensionality. Later, we utilized machine learning algorithm, including heuristic algorithm and an autoencoder neural network, to calculate fraud scores based on abnormality of every record. At the end, we evaluated why the 10 records with top combined scores are fraudulent.

Part II. Description of Data

-Dataset name: NY Property Data
-Data source: Department of Finance in New York City government
-Dataset Overview

- Number of records: 1,070,994
- Number of fields: 32
- Time period: Year 2010 to 2011

Summary statistics of fields

Field	Field type	Data Type	#Rows populated	%Rows populated	#Unique values	%Unique values	#Zeros in Population
RECORD	Categorical	int64	1070994	100.00%	1070994	100.00%	-
BBLE	Categorical	object	1070994	100.00%	1070994	100.00%	-
B	Categorical	int64	1070994	100.00%	5	0.00%	-
BLOCK	Categorical	int64	1070994	100.00%	13984	1.31%	-
LOT	Categorical	int64	1070994	100.00%	6366	0.59%	-
EASEMENT	Categorical	object	4636	0.43%	12	0.28%	-
OWNER	Categorical	object	1039249	97.04%	863347	83.07%	-
BLDGCL	Categorical	object	1070994	100.00%	200	0.02%	-
TAXCLASS	Categorical	object	1070994	100.00%	11	0.00%	-
LTFRONT	Numerical	int64	1070994	100.00%	1297	0.12%	169108
LTDEPTH	Numerical	int64	1070994	100.00%	1370	0.13%	170128

EXT	Categorical	object	354305	33.08%	3	0.00%	-
STORIES	Categorical	object	1014730	94.75%	111	0.01%	-
FULLVAL	Numerical	float64	1070994	100.00%	109324	10.21%	13007
AVLAND	Numerical	float64	1070994	100.00%	70921	6.62%	13009
AVTOT	Numerical	float64	1070994	100.00%	112914	10.54%	13007
EXLAND	Categorical	float64	1070994	100.00%	33419	3.12%	491699
EXTOT	Categorical	float64	1070994	100.00%	64255	6.00%	432572
EXCD1	Categorical	float64	638488	59.62%	129	0.02%	-
STADDR	Categorical	object	1070318	99.94%	839280	78.41%	-
ZIP	Categorical	float64	1041104	97.21%	196	0.02%	-
EXMPTCL	Categorical	object	15579	1.45%	14	0.10%	-
BLDFRONT	Numerical	int64	1070994	100.00%	612	0.06%	228815
BLDDEPTH	Numerical	int64	1070994	100.00%	621	0.06%	228853
AVLAND2	Numerical	float64	282726	26.40%	58591	20.72%	-
AVTOT2	Numerical	float64	282732	26.40%	111360	39.39%	-
EXLAND2	Categorical	float64	87449	8.17%	22195	25.38%	-
EXTOT2	Categorical	float64	130828	12.22%	48348	36.96%	-
EXCD2	Categorical	float64	92948	8.68%	60	0.07%	-
PERIOD	Categorical	object	1070994	100.00%	1	0.00%	-
YEAR	Categorical	object	1070994	100.00%	1	0.00%	-
VALTYPE	Categorical	object	1070994	100.00%	1	0.00%	-

Important variables for consideration

Name	Type	Description
RECORD	categorical	Unique Identifier of each record
BLOCK	categorical	Block Number Index
TAXCLASS	categorical	Tax class
FULLVAL	numerical	Total market value of the property
AVLAND	numerical	Total Land Area
AVTOT	numerical	Assessed Value of the property
ZIP	categorical	Postal zip code of the property
STORIES	categorical	Number of stories for the building
LTFRONT	numerical	Lot Frontage in feet
LTDEPTH	numerical	Lot Depth in feet
BLDFRONT	numerical	Building Frontage in feet
BLDDEPTH	numerical	Building Depth in feet

Summary of important numerical variables

Field	mean	std	min	25%	50%	75%	max
LOT	364.72	853.2	1	23	49	143	9978
LTFRONT	36.64	74	0	19	25	40	9999
LTDEPTH	88.86	76.4	0	80	100	100	9999
FULLVAL	874264.51	11582430	0	304000	447000	619000	6150000000
AVLAND	85067.92	4057260	0	9180	13678	19740	2668500000
AVTOT	227238.17	6877529	0	18374	25340	45438	4668309000
BLDFRONT	23.04	35.6	0	15	20	24	7575
BLDDEPTH	39.92	42.7	0	26	39	50	9393

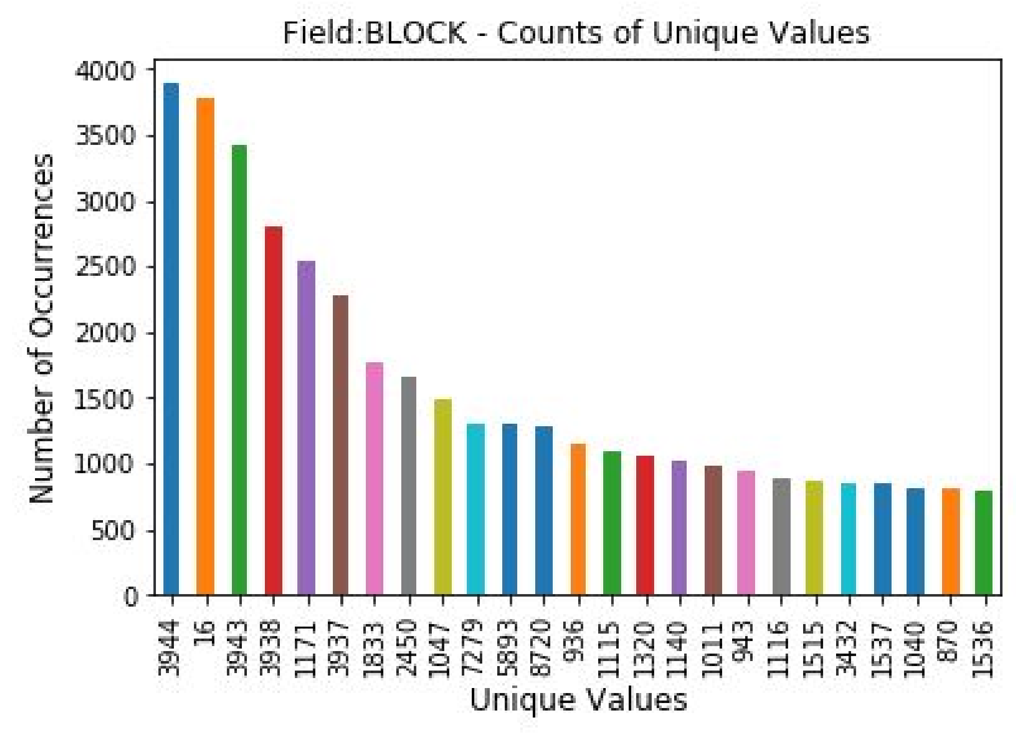
Summary of important categorical variables

Field	Total Count	#Unique values	Most Common Value	Frequency
B	1070994	5	4	358046
BLOCK	1070994	13984	3944	3888
TAXCLASS	1070994	11	1	660721
STORIES	1014730	111	2	415092
ZIP	1041104	196	10314	24606

We scrutinized the important variables by plotting their distribution across the data using Bar charts and density plots. An exhaustive data quality analysis report is attached in the appendix of this report. The data analysis findings of the important variables are as shown below:

Block

Block is a categorical field that represents valid block ranges for various borough codes. The count plot for the top 25 categories is as shown below:



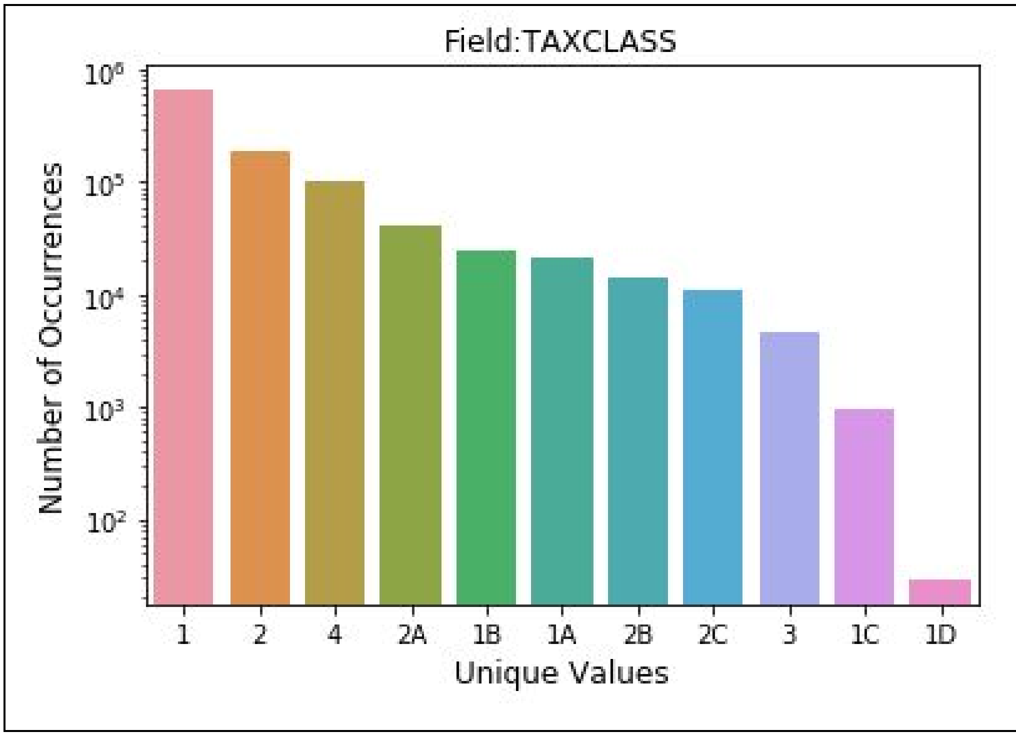
	Value	Count	Percent
0	3944	3888	0.36
1	16	3786	0.35
2	3943	3424	0.32
3	3938	2794	0.26
4	1171	2535	0.24
5	3937	2275	0.21
6	1833	1774	0.17
7	2450	1651	0.15
8	1047	1480	0.14
9	7279	1302	0.12
10	5893	1295	0.12
11	8720	1281	0.12
12	936	1151	0.11
13	1115	1090	0.10
14	1320	1049	0.10
15	1140	1017	0.09
16	1011	991	0.09
17	943	946	0.09
18	1116	881	0.08
19	1515	869	0.08
20	3432	853	0.08
21	1537	842	0.08
22	1040	821	0.08
23	870	809	0.08
24	1536	796	0.07

TAXCLASS

TAXCLASS represents the current property tax code. The various tax classes are:

- 1 = 1-3 UNIT RESIDENCES
    - 1A = 1-3 STORY CONDOMINIUMS ORIGINALLY A CONDO
    - 1B = RESIDENTIAL VACANT LAND
    - 1C = 1-3 UNIT CONDOMINIUMS ORIGINALLY TAX CLASS 1
    - 1D = SELECT BUNGALOW COLONIES
- 2 = APARTMENTS
    - 2A = APARTMENTS WITH 4-6 UNITS
    - 2B = APARTMENTS WITH 7-10 UNITS
    - 2C = COOPS/CONDOS WITH 2-10 UNITS
- 3 = UTILITIES (EXCEPT CEILING RR)
- 4 = ALL OTHERS

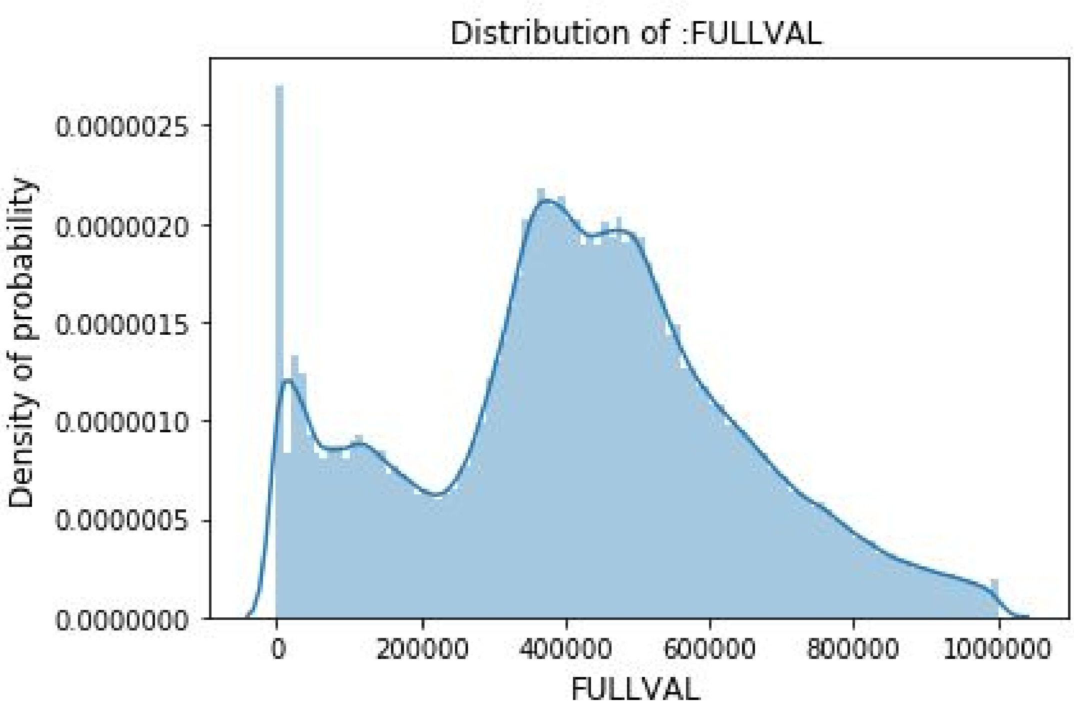
The distribution of TAXCLASS is as shown below:



	Value	Count	Percent(%)
0	1	660721	61.69
1	2	188612	17.61
2	4	104310	9.74
3	2A	40574	3.79
4	1B	24738	2.31
5	1A	21667	2.02
6	2B	13964	1.30
7	2C	10795	1.01
8	3	4638	0.43
9	1C	946	0.09
10	1D	29	0.00

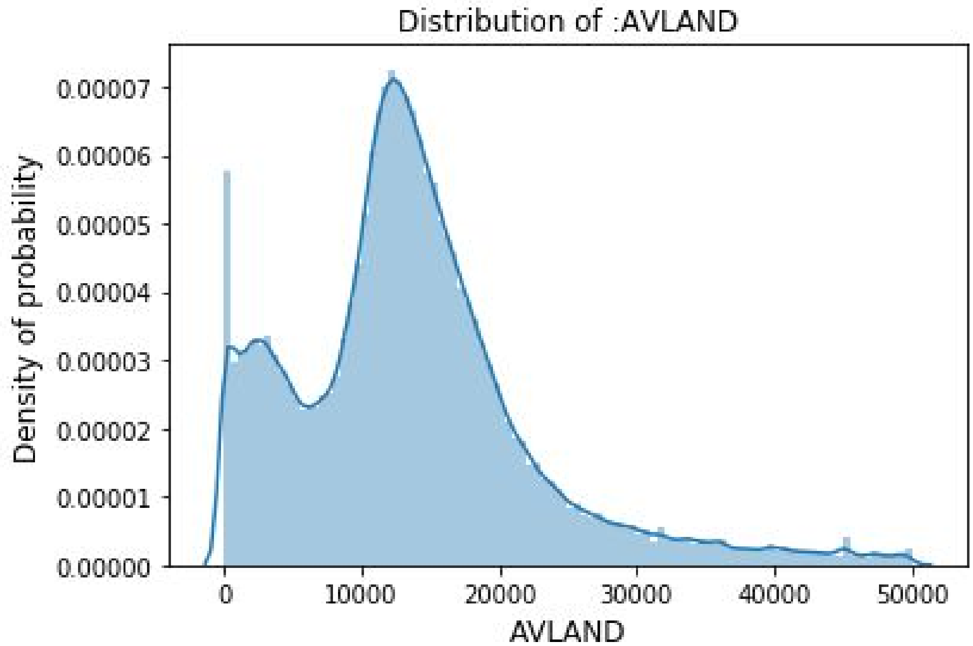
FULLVAL

FULLVAL represents the total market value of the property. The records with value greater than 1,000,000 are treated as outliers, which are omitted in the distribution plot.



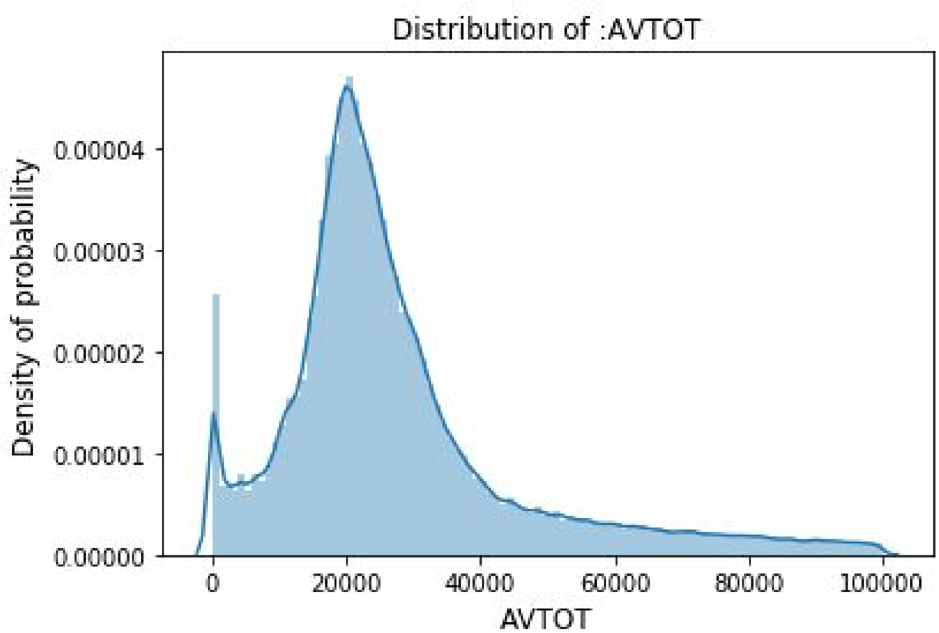
AVLAND

AVLAND stands for the total land value of the property. Records with value greater than 50,000 are treated as outliers, which are excluded from the distribution plot.



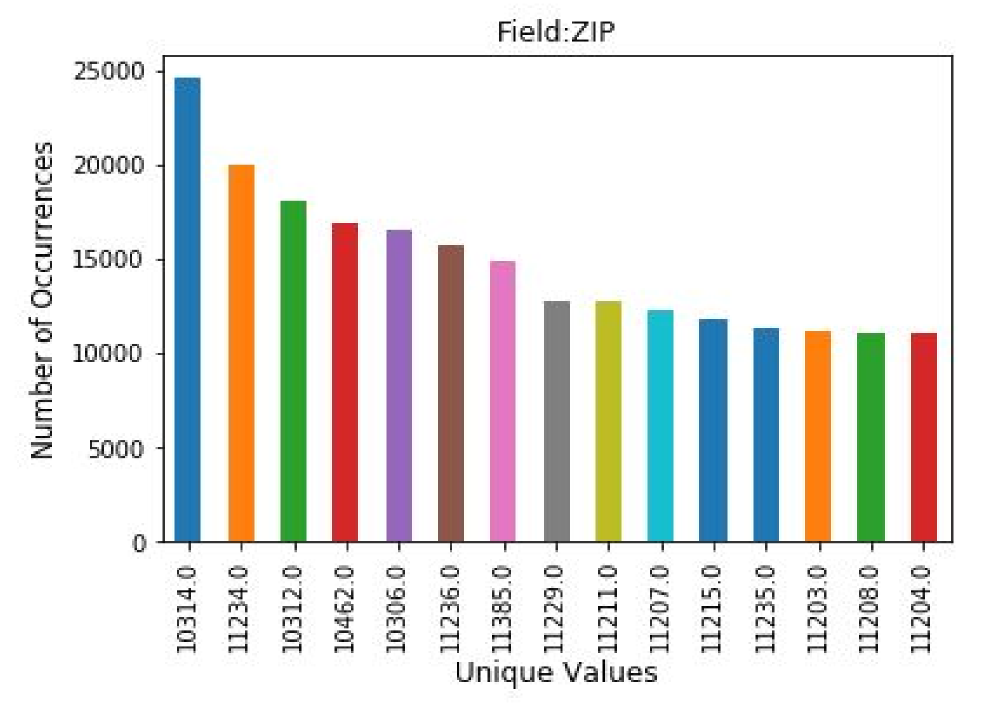
AVTOT

AVTOT stands for the assessed value of the property. The records with values greater than 100,000 are treated as outliers and are omitted in the distribution plot below.



ZIP

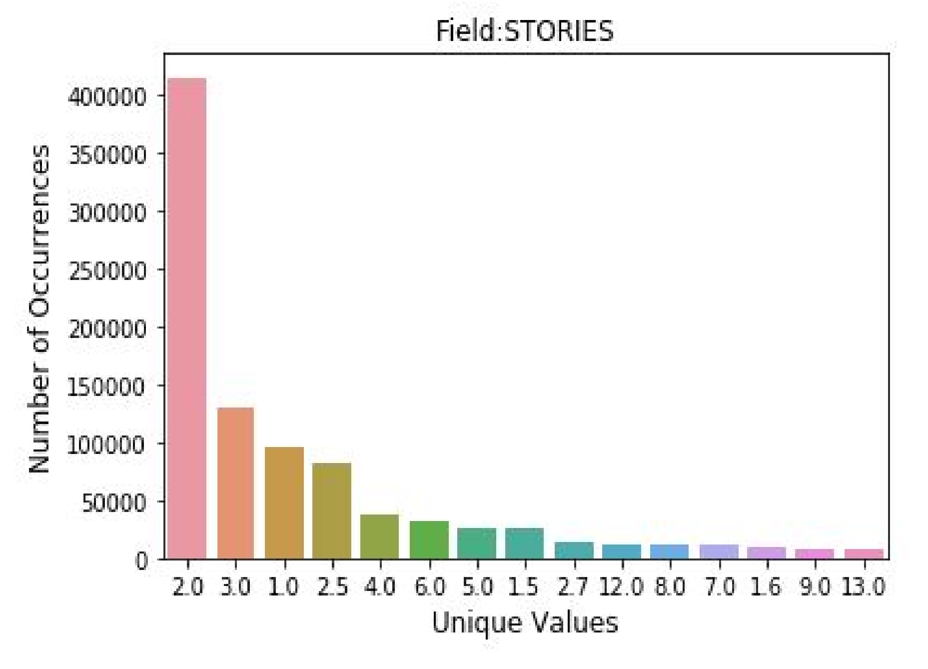
ZIP represents the zip code of the property. The count plot of top 15 values is as shown below:



	Value	Count	Percent(%)
0	10314.0	24606	2.36
1	11234.0	20001	1.92
2	10312.0	18127	1.74
3	10462.0	16905	1.62
4	10306.0	16578	1.59
5	11236.0	15678	1.51
6	11385.0	14921	1.43
7	11229.0	12793	1.23
8	11211.0	12710	1.22
9	11207.0	12293	1.18
10	11215.0	11834	1.14
11	11235.0	11312	1.09
12	11203.0	11241	1.08
13	11208.0	11139	1.07
14	11204.0	11061	1.06

STORIES

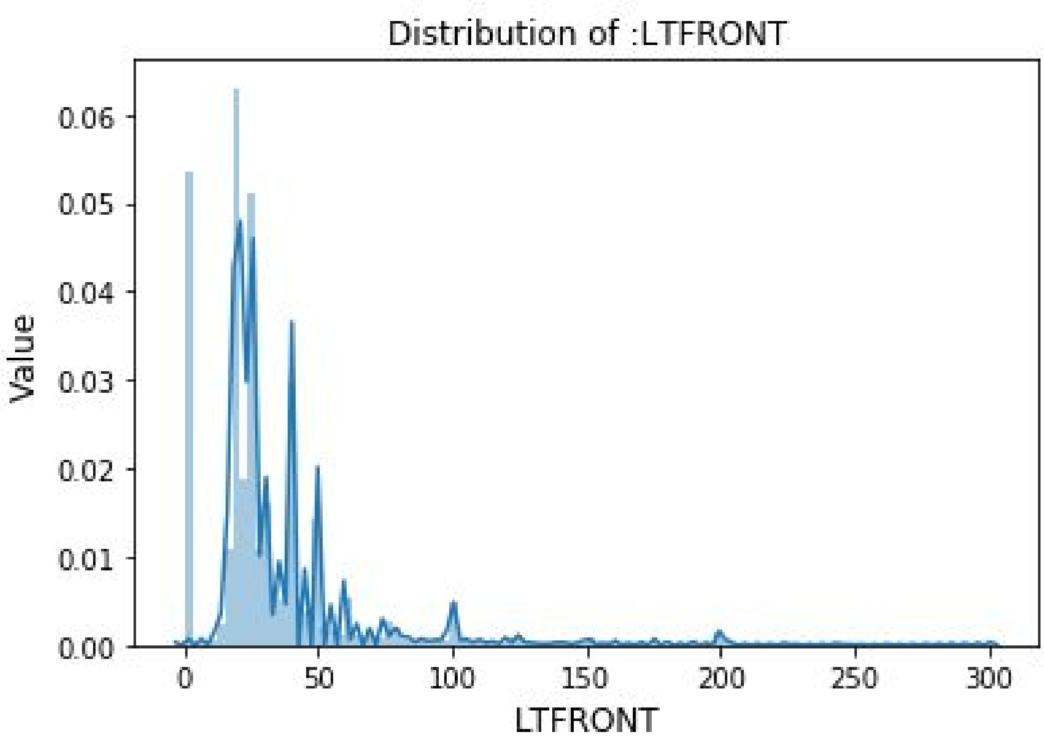
STORIES represents the number of stories in a building. The count plot of top 15 values is as shown below:



	Value	Count	Percent(%)
0	2.0	415092	40.91
1	3.0	130127	12.82
2	1.0	96706	9.53
3	2.5	82292	8.11
4	4.0	38342	3.78
5	6.0	30936	3.05
6	5.0	25971	2.56
7	1.5	24770	2.44
8	2.7	13595	1.34
9	12.0	12198	1.20
10	8.0	11953	1.18
11	7.0	11899	1.17
12	1.6	8954	0.88
13	9.0	7343	0.72
14	13.0	7330	0.72

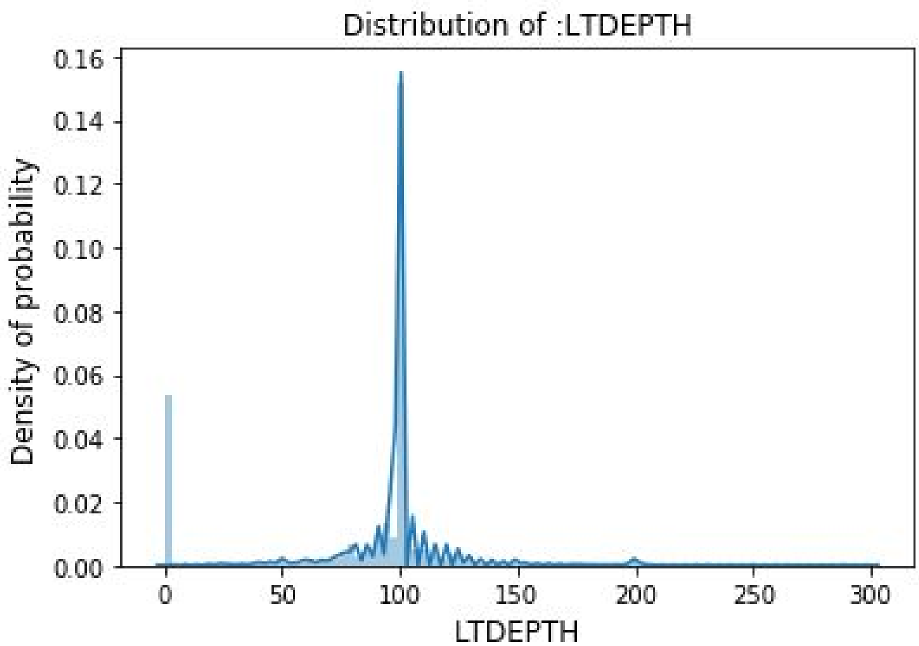
LTFRONT

LTFRONT measures the lot frontage in feet. The records with value greater than 300 are treated as outliers and are excluded from the distribution plot below.



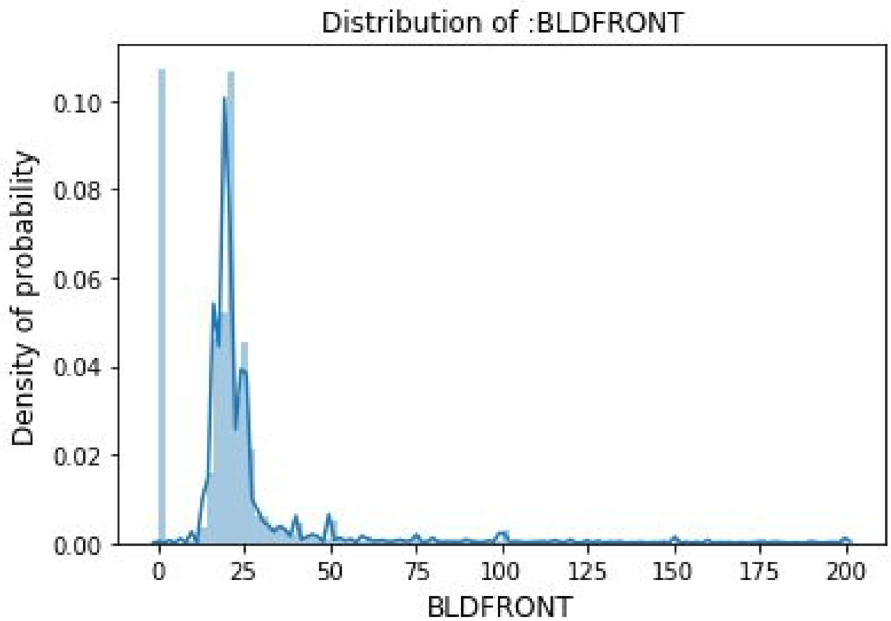
LTDEPTH

LTFRONT measures the lot depth in feet. The records with value greater than 300 are treated as outliers and are omitted from the distribution plot.



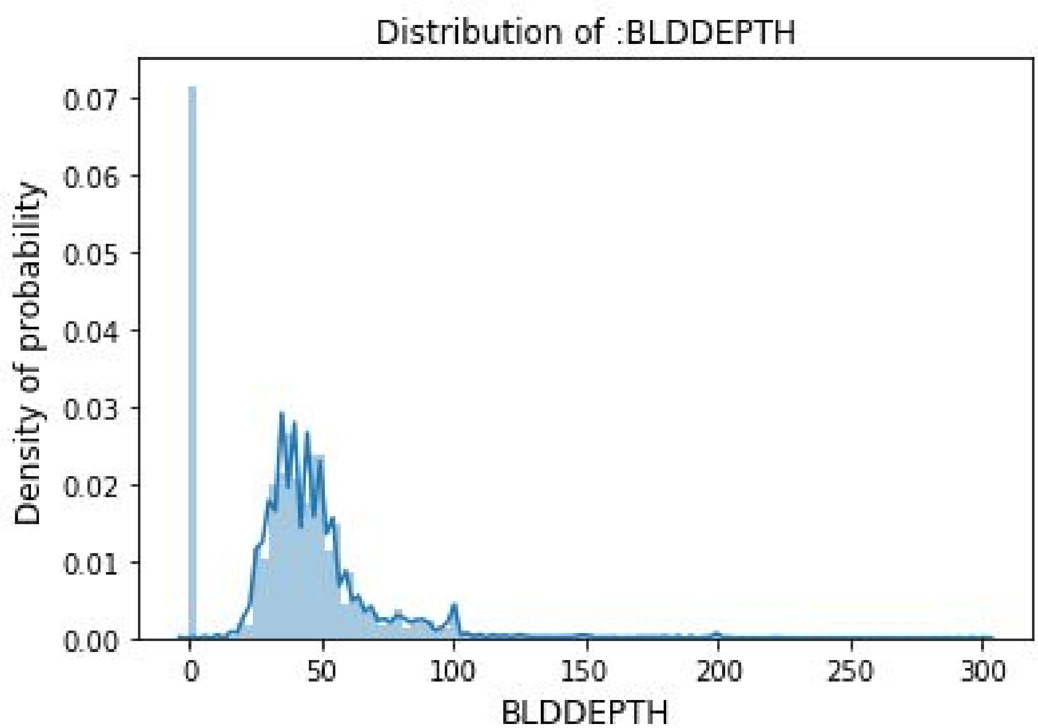
BLDFRONT

BLDFRONT measures the building frontage in feet. The records with value greater than 200 are treated as outliers and are omitted from the distribution plot.



BLDDEPTH

BLDDEPTH measures the building frontage in feet. The records with value greater than 300 are treated as outliers and are omitted from the distribution plot.



From the above plot, we can see that if we eliminate 'Zero's, we get an almost normal distribution around 40.

Part III. Data Cleaning

From the data quality analysis, we observed that there were many missing values, and for further analysis we needed to fill in those missing fields with 'innocuous' values that would not drastically change the distribution of the variable across the data and also would not introduce any anomalies. The approaches we took for data cleaning for each variables are as described below:

ZIP

We first filled in missing values for ZIP because we needed it for categorizing other variables later. Considering that Zip code contains geographical information, we decided to aggregate the data by BLOCK and fill in the most common zip code for each block. This handled the majority of missing values in ZIP. However, there were still about 500 missing values after this step. It appeared that those blocks only contain one valid zip code while all other zip codes are null. As a result, we proceed to a second aggregation by TAXCLASS. It is reasonable to assume that properties with the same characteristics cluster together. For example, commercial properties would be located in shopping district while residential properties are within residential areas. Then, similar properties should have the same zip code. Therefore, we then filled in those 500 missing zip codes with most common field from that TAXCLASS.

STORIES

Similar to the assumption mentioned above, properties with the same characteristics tend to locate nearby each other. Also, given the fact that this variable has a large number of outliers, we decided to use the median instead of mean. Therefore, we first filled in the missing values in stories using median from its BLOCK-aggregated group, which has 13,984 unique values. When certain block only contains stories with null value, this aggregation still returned 814 missing values. Then we turned to ZIP, which has 194 unique values. ZIP compared to BLOCK is more generic, but it should still provide the values for stories with similar accuracy, so we filled in those 814 missing values by taking the median stories from that ZIP based aggregation.

FULLVAL, AVLAND, AVTOT

These three variables would be aggregated using the cleaned ZIP. Properties within the same zip code should not vary in terms of characteristics, stories, and size. Also, since the original dataset is skewed to the right, it is more reasonable to use the median instead of mean. We first converted all the zeros into NA, then replaced NAs with the median from that ZIP block.

LTFRONT, LTDEPTH

If both LTFRONT and LTDEPTH equal 0, we set them to be the median value of LTFRONT, LTDEPTH, which are 25 and 100 respectively. Furthermore, since lot frontage and lot depth are correlated, we referred to the average value of each other to fill in missing value when the value of either column is missing. The steps we took are described as follows.

First, we excluded records whose LTFRONT equals 0 or LTDEPTH equals 0. Then we used aggregation to calculate average LTDEPTH by LTFRONT and average LTDEPTH by LTFRONT.

If original LTDEPTH equals 0 and original LTFRONT not equals 0, we set LTDEPTH to be the average LTDEPTH of the corresponding LTFRONT.

If original LTFRONT equals 0 and original LTDEPTH not equals 0, we set LTFRONT to be the average LTFRONT of the corresponding LTDEPTH.

Note: For LTDEPTH, LTFRONT who cannot be filled out by above logic, we set them to the median value of LTFRONT, LTDEPTH, which are 25 and 100 respectively.

BLDFRONT, BLDDEPTH

If both BLDFRONT and BLDDEPTH equal 0, we set them to be the median value of BLDFRONT, BLDDEPTH, which are 20 and 39 respectively. Moreover, since building frontage and building depth are correlated, we refered to the average value of each other to fill in missing value when the value of either column is missing. The steps we took are described as follows.

First, we excluded records whose BLDFRONT equals 0 or BLDDEPTH equals 0. Then we used aggregation to calculate average BLDDEPTH by BLDFRONT and average BLDFRONT by BLDDEPTH.

If original BLDDEPTH equals 0 and original BLDFRONT not equals 0, set BLDDEPTH to be the average BLDDEPTH of the corresponding BLDFRONT.

If original BLDFRONT equals 0 and original BLDDEPTH not equals 0, set BLDFRONT to be the average BLDFRONT of the corresponding BLDDEPTH.

For BLDFRONT, BLDDEPTH who cannot be filled out by above logic, set them to the median value of BLDFRONT, BLDDEPTH, which are 20 and 39 respectively.

Part IV. Expert Variable Creation

It is vital to create expert variables that can better explain the properties. After cleaning the data, we decided to create expert variables using the existing variables. We did so by creating three layers.

Layer 1 (3 size variables and 3 property value variables)

In the first layer, we chose 3 primary property value variables (FULLVAL, AVLAND, AVTOT). Also, we created three property size variables (lotarea, bldarea, bldvol). The area of lot (lotarea) equals to the product of the lot frontage in feet and the lot depth in feet. The area of building (bldarea) equals to the product of the building frontage in feet and the building depth in feet. The volume of building (bldvol) equals to the product of the building area and the number of building stories.

- lotarea = LTFRONT * LTDEPTH
- bldarea = BLDFRONT * BLDDEPTH
- bldvol = bldarea * STORIES

- FULLVAL: full market value of property
- AVLAND: assessed total value of land
- AVTOT: assessed total value of property

Layer 2 (9 variables)

In the second layer, we created 9 variables, each of the 3 property value variables (FULLVAL, AVLAND, AVTOT) normalized by each of the 3 sizes created in layer 1.

FULLVAL

- FULLVAL_LOTAREA = FULLVAL/LOTAREA
- FULLVAL_BLDAREA = FULLVAL/BLDAREA
- FULLVAL_BLDVOL = FULLVAL/BLDVOL

AVLAND

- AVLAND_LOTAREA = AVLAND/LOTAREA

- AVLAND_BLDAREA = AVLAND/BLDAREA
- AVLAND_BLDVOL = AVLAND/BLDVOL

AVTOT

- AVTOT_LOTAREA = AVTOT/LOTAREA
- AVTOT_BLDAREA = AVTOT/BLDAREA
- AVTOT_BLDVOL = AVTOT/BLDVOL

Layer 3 (5 groups and 45 expert variables)

In the third layer, we first created 5 groups (ZIP, ZIP3, TAXCLASS, BORO, ALL). We created ZIP3 by extracting the first 3 digits (left to right) from the original ZIP, which contains 5 digits. "All" means the entire data set without any grouping, from which we calculate the overall averages or medians.

Next, we took 9 variables we obtained in layer 2 and grouped each of them by 5 groups. After that, we calculated mean in each group and then divided those 9 variables by their corresponding mean in the group. For example FULLVAL_LOTAREA_ZIP5= (FULLVAL/LOTAREA) / (mean of FULLVAL/LOTAREA group by ZIP).

Finally, we created 45 (9 variables*5 groups) new expert variables.

ZIP

Variable Name	**Description**
FULLVAL_LOTAREA_ZIP5	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of properties grouped by ZIP
FULLVAL_BLDAREA_ZIP5	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of properties grouped by ZIP
FULLVAL_BLDVOL_ZIP5	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of properties grouped by ZIP
AVTOT_LOTAREA_ZIP5	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of properties grouped by ZIP

AVTOT_BLDAREA_ZIP5	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of properties grouped by ZIP
AVTOT_BLDVOL_ZIP5	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of properties grouped by ZIP
AVLAND_LOTAREA_ZIP5	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of properties grouped by ZIP
AVLAND_BLDAREA_ZIP5	Ratio of AVLAND/BLDAREA to Average AVLAND/BLDAREA of properties grouped by ZIP
AVLAND_BLDVOL_ZIP5	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of properties grouped by ZIP

ZIP3

Variable Name	**Description**
FULLVAL_LOTAREA_ZIP3	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of properties grouped by ZIP3
FULLVAL_BLDAREA_ZIP3	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of properties grouped by ZIP3
FULLVAL_BLDVOL_ZIP3	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of properties grouped by ZIP3
AVTOT_LOTAREA_ZIP3	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of properties grouped by ZIP3
AVTOT_BLDAREA_ZIP3	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of properties grouped by ZIP3
AVTOT_BLDVOL_ZIP3	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of properties grouped by ZIP3
AVLAND_LOTAREA_ZIP3	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of properties grouped by ZIP3

AVLAND_BLDAREA_ZIP3	Ratio of AVLAND/BLDAREA to Average AVLAND/BLDAREA of properties grouped by ZIP3
AVLAND_BLDVOL_ZIP3	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of properties grouped by ZIP3

TAXCLASS

Variable Name	**Description**
FULLVAL_LOTAREA_TAXCLASS	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of properties grouped by TAXCLASS
FULLVAL_BLDAREA_TAXCLASS	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of properties grouped by TAXCLASS
FULLVAL_BLDVOL_TAXCLASS	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of properties grouped by TAXCLASS
AVTOT_LOTAREA_TAXCLASS	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of properties grouped by TAXCLASS
AVTOT_BLDAREA_TAXCLASS	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of properties grouped by TAXCLASS
AVTOT_BLDVOL_TAXCLASS	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of properties grouped by TAXCLASS
AVLAND_LOTAREA_TAXCLASS	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of properties grouped by TAXCLASS
AVLAND_BLDAREA_TAXCLASS	Ratio of AVLAND/BLDAREA to Average AVLAND/BLDAREA of properties grouped by TAXCLASS

AVLAND_BLDVOL_TAXCLASS	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of properties grouped by TAXCLASS

BORO

Variable Name	**Description**
FULLVAL_LOTAREA_BORO	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of properties grouped by BORO
FULLVAL_BLDAREA_BORO	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of properties grouped by BORO
FULLVAL_BLDVOL_BORO	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of properties grouped by BORO
AVTOT_LOTAREA_BORO	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of properties grouped by BORO
AVTOT_BLDAREA_BORO	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of properties grouped by BORO
AVTOT_BLDVOL_BORO	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of properties grouped by BORO
AVLAND_LOTAREA_BORO	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of properties grouped by BORO
AVLAND_BLDAREA_BORO	Ratio of AVLAND/BLDAREA to Average AVLAND/BLDAREA of properties grouped by BORO
AVLAND_BLDVOL_BORO	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of properties grouped by BORO

All

Variable Name	**Description**
FULLVAL_LOTAREA_ALL	Ratio of FULLVAL/LOTAREA to Average FULLVAL/LOTAREA of properties
FULLVAL_BLDAREA_ALL	Ratio of FULLVAL/BLDAREA to Average FULLVAL/BLDAREA of properties
FULLVAL_BLDVOL_ALL	Ratio of FULLVAL/BLDVOL to Average FULLVAL/BLDVOL of properties
AVTOT_LOTAREA_ALL	Ratio of AVTOT/LOTAREA to Average AVTOT/LOTAREA of properties
AVTOT_BLDAREA_ALL	Ratio of AVTOT/BLDAREA to Average AVTOT/BLDAREA of properties
AVTOT_BLDVOL_ALL	Ratio of AVTOT/BLDVOL to Average AVTOT/BLDVOL of properties
AVLAND_LOTAREA_ALL	Ratio of AVLAND/LOTAREA to Average AVLAND/LOTAREA of properties
AVLAND_BLDAREA_ALL	Ratio of AVLAND/BLDAREA to Average AVLAND/BLDAREA of properties
AVLAND_BLDVOL_ALL	Ratio of AVLAND/BLDVOL to Average AVLAND/BLDVOL of properties

Part V. Principal Component Analysis

Overview

After expert variable creation, the next step is to perform Principal component analysis (PCA). PCA is an advanced technique to extract important independent and uncorrelated principal features from a large set of variables available in a dataset. Mathematically, PCA can be implemented by eigen decomposition or singular value decomposition. Before building fraud scores for each property record, we used PCA to reduce dimension and to remove correlation between features. Since the original variables have different scales and variances, directly performing PCA on them could lead to unreliable results. To avoid it, we used z-score to normalize variables prior to implementing PCA.

Function

Reducing Dimension

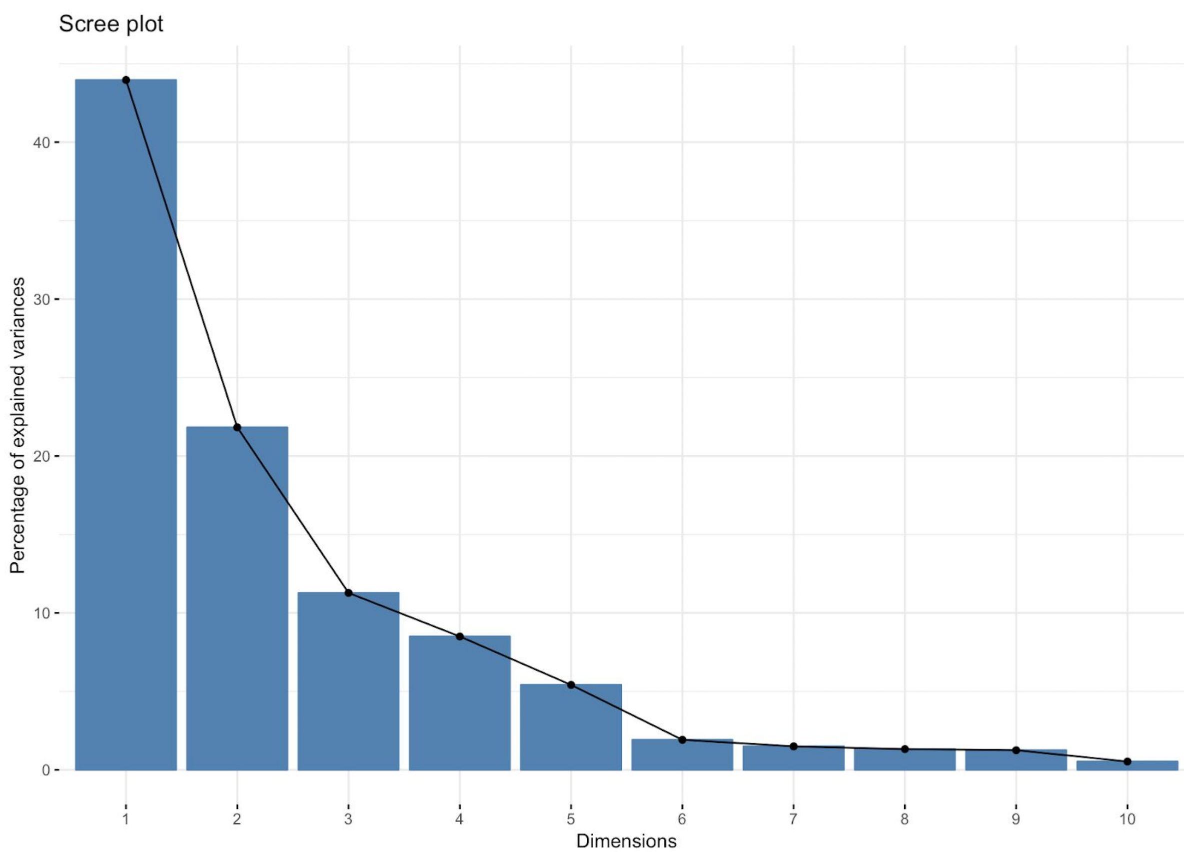
While creating new variables, in order to capture as much valuable information as we can, we created 45 expert variables, which inevitably have overlaps and are correlated with each other. PCA creates independent principal components (PC's) that are a linear combination of the original variables, such that the maximum variance can be extracted from the variables with zero correlation. By doing so, it can help reduce the number of variables and transform a large set of variables into a small set of features that still contains most of the information. PCA orders all PC's in decreasing order of their eigenvalues. This means that the top PC captures the maximum variance in the data, and each succeeding PC accounts for the remaining variability. This way, we can keep the top PC's and drop the 'least important' PC's in the results to retain the most influential part of the data.

Removing Correlation

As we know, the expert variables are not exactly independent of one another. While creating PC's to capture the largest variance, PCA also satisfies the orthogonality condition, that is each PC represents an eigenvector perpendicular to the other PC's in the eigenspace. By defining a new orthogonal coordinate system, PCA helps remove the correlation of features.

Results

As a result, PCA provided a rotation matrix, where each column contains the principal component loading vector. By examining how much variance each PC explains in the *scree plot*, we can decide the number of PC's we should keep. The higher the *explained variance* by a selected number of PC's, the more information was contained in those PC's. Below is the scree plot, where the horizontal axis represents PC's and the vertical axis represents the percentage of explained variance. As we can see, the first PC can explain about 45% variance while the succeeding PC's explanation continuously decrease.



Through calculation, we observed that the cumulative percentage of explained variance is about 95.87% for the top 7 PC's. Therefore, we decided to keep PC1 through PC7 to perform further analysis. Since we wanted all PC's to be equally important, we z-scaled these features before feeding them into the following algorithm.

Part VI. Fraud Algorithms

Once we had the data dimensionality reduced using PCA and z-scaled, we fed the data into our fraud algorithm. We developed two algorithm that calculate fraud scores for each record. Our two different algorithms:

1. Heuristic function
2. Autoencoder

Model 1: Heuristic Model

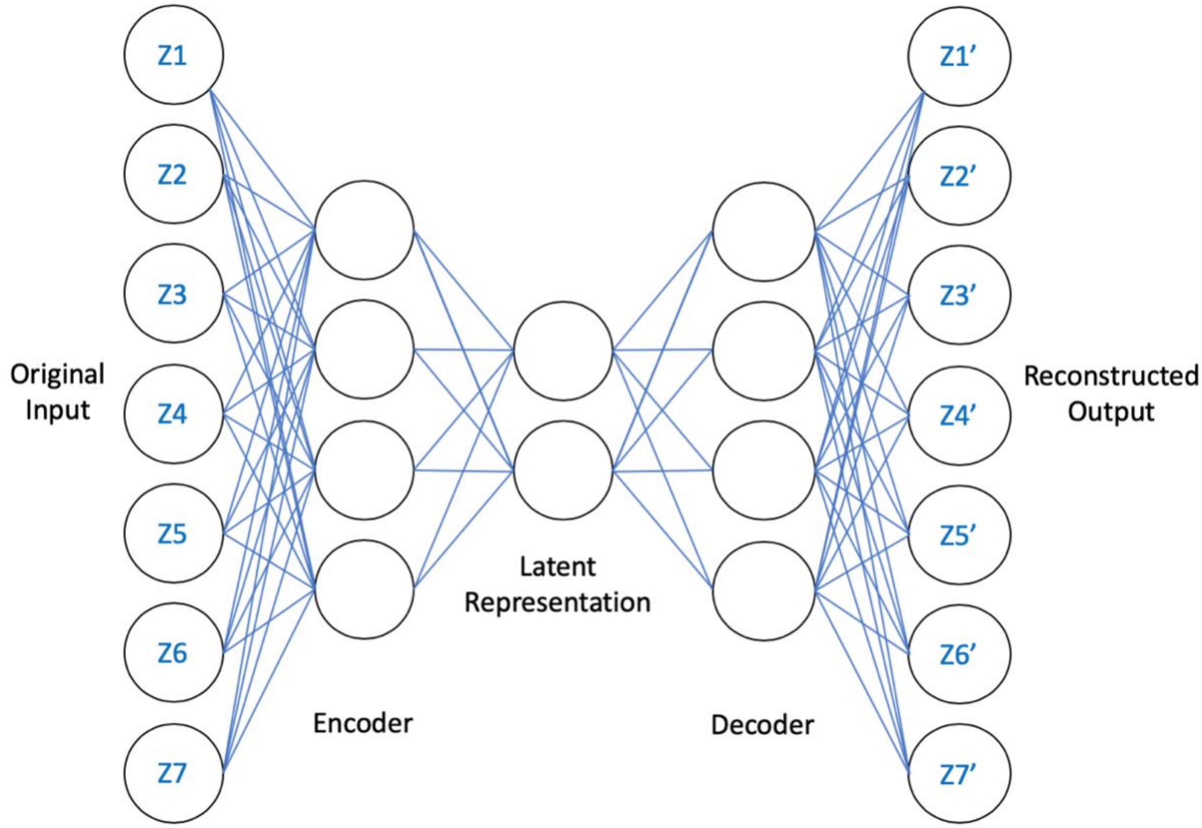
In the Heuristic Function we used Mahalanobis Distance to get the distance of each record from the origin (mean 0) and calculated the fraud scores based on z-scores. The Euclidean distance, a special case of the Mahalanobis distance with equal variances of the variables and zero covariances is used as the heuristic function to calculate the fraud score. We chose Euclidean distance (n=2) as it is rotation invariant. The Heuristic Function formula is as below:

$$S_{HU} = \sqrt[2]{\sum_k |Z_k^i|^2}$$

Here Zi represents the z-scaled value of each features of the i'th record. Mathematically Zi represents how far the record is away from mean 0 of the feature. When Zi is large, the fraud score also would be large. We can detect the potential fraud records (outliers) by plotting the distribution of the fraud score calculated. The distribution plot is shown in the following section of the report.

Model 2: Autoencoder

Autoencoder is an artificial neural network that aims to reconstruct its inputs as outputs. This network includes two parts: Encoder and Decoder. While encoder compresses the input into a latent-space representation, decoder aims to reconstruct the input from the latent-space representation. The neural network representation of a typical encoder is as shown below.



In fraud analysis, we can detect anomaly records with an autoencoder. By calculating the distance between the reconstructed (output) value from the autoencoder and the original input value, we can identify the anomaly records, because the fraudulent behaviors tend to have *longer distance* than other normal behaviors. The autoencoder distance formula is as below:

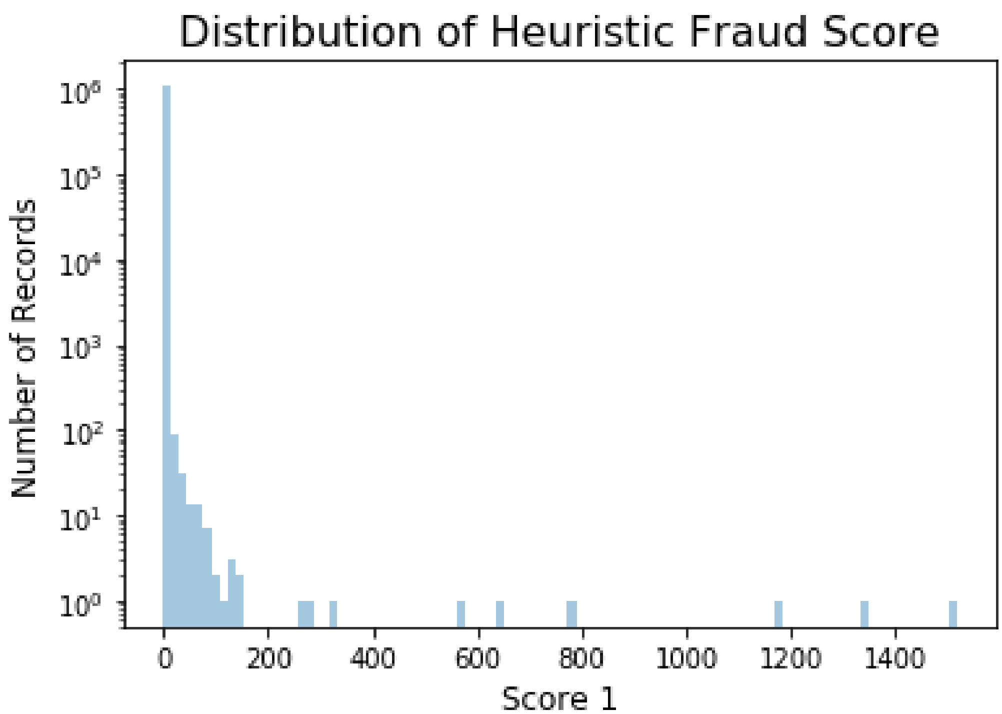
$$S_{AE} = \sqrt[2]{\sum_k |Z'^i_k - Z^i_k|^2}$$

In this project, we trained a Keras based autoencoder unsupervised model with the z-scaled PCA data. This autoencoder function had 3 hidden layers: an encoder, a latent-space representation, and a decoder. Encoder and decoder layer both contains 4 nodes, and latent representation has 2 nodes. After training the model and extracting new reconstructed output, we computed the fraud score using the formula shown above. We can detect the potential fraud records (outliers) by plotting the distribution of the fraud score calculated. The distribution plot is shown in the following section of the report.

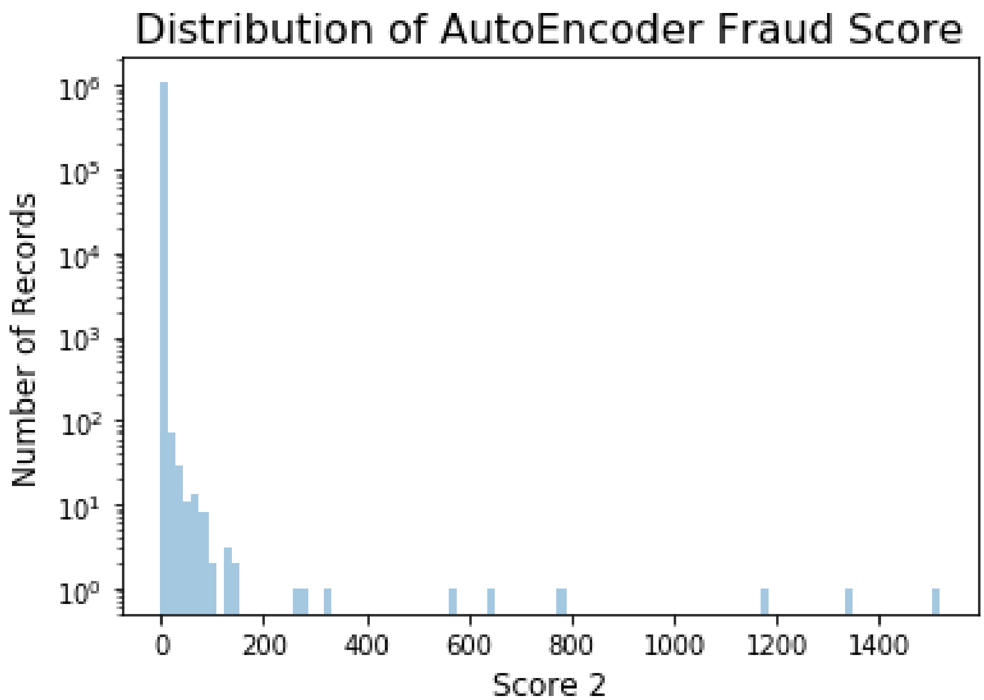
Part VII. Results

Dist Plot

The distribution of fraud scores from the Heuristic model is shown in the plot below. The distribution is right-skewed with a long tail.



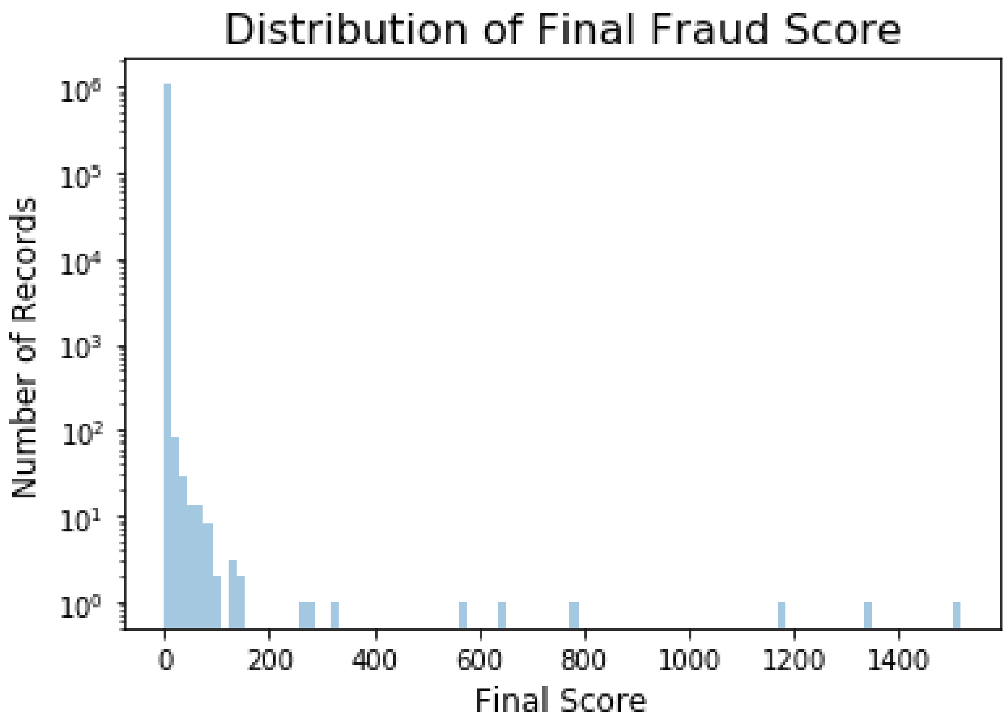
The distribution of fraud scores with Autoencoder shows in the below plot. It is right skewed and with long tail.



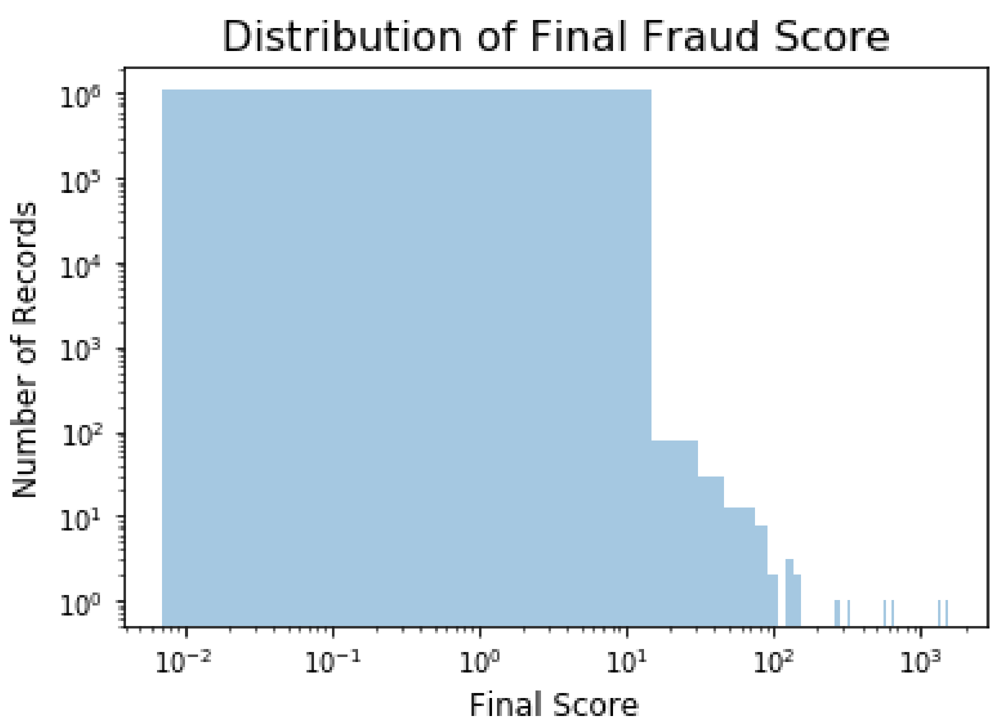
Final Score

After having the fraud score from both the Heuristic model and Autoencoder, we combined these two scores together then divided it by 2 to get the final score.

Plotting on a log scale on the y-axis, we got the following distribution of the final fraud score as below.



We also tried a different way to look at the final fraud score distribution by adjusting the scale.



The distribution plot shows that there are normal records with final fraud scores ranging from 10^{-2} to 10^{1}, and some outliers that have final fraud score greater than 10^{2}. We later ranked the records based on their final fraud score. Rank 1 is given to the record that has the least chance of being fraud (i.e. the least fraud score) and the top potential fraud record has a rank of 1070994.

Fraud & Normal Comparison

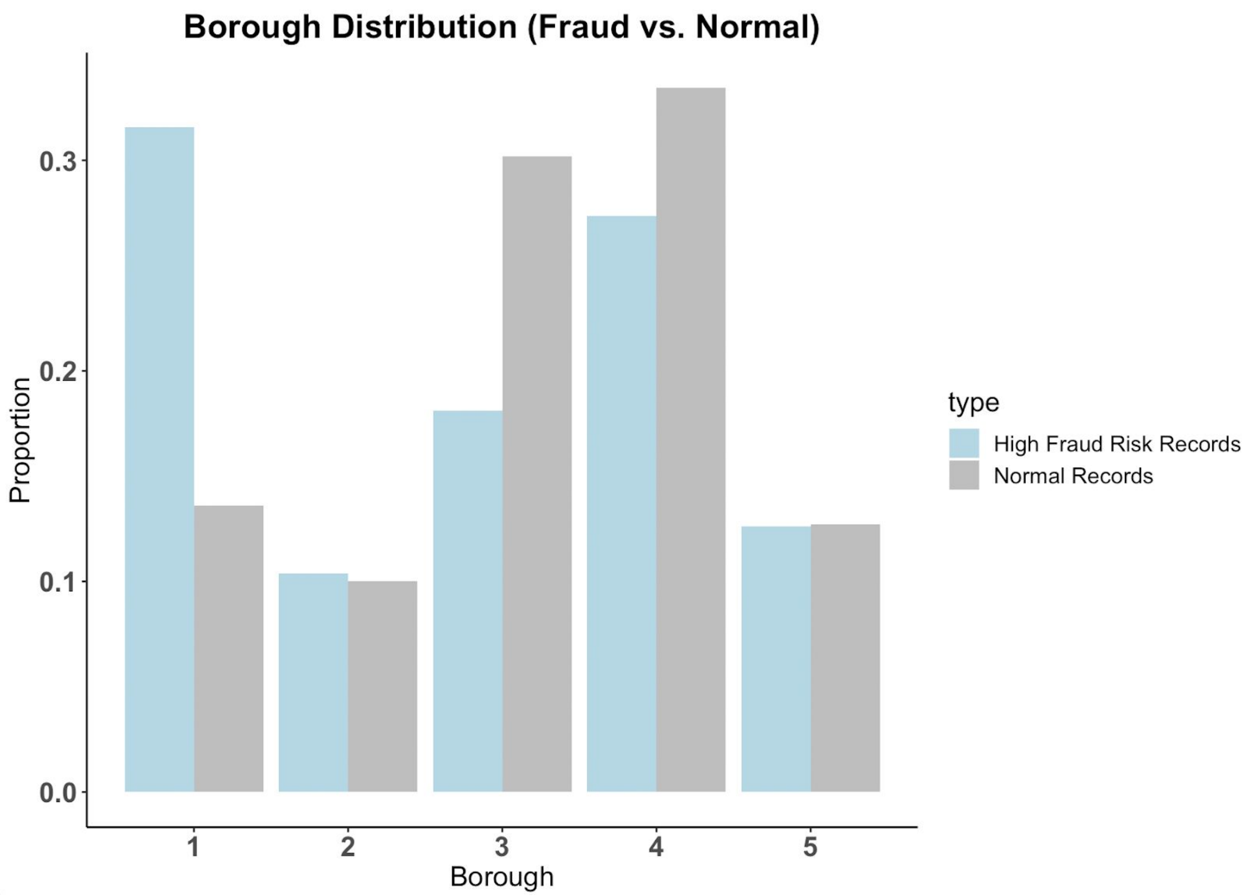
Based on the ranking of fraud score, we recognized top 0.2% records (2142 records) as high fraud risk records and the remaining 99.8% records as normal records. Then we compared the mean, median and SD of key numerical variables of these two groups of records.

	Remaining 99.8% Records			Top 0.2% Records		
Variable	**Mean**	**Median**	**SD**	**Mean**	**Median**	**SD**
FULLVAL	788,546	450,000	5,504,269	46,785,198	12,100,000	223,465,275
AVLAND	61,823	13,775	883,129	11,803,542	2,508,592	87,792,677
AVTOT	184,283	25,574	2,473,736	22,098,792	4,696,380	141,918,407
STORIES	4.93	2	8.28	7.9	2	12.6
LTfront	40	25	58	454	201	853
LTdepth	104	100	57	462	224	664
BLDfront	27.3	20	34	28.4	20	45
BLDdepth	48.3	39	38	41.6	39	37

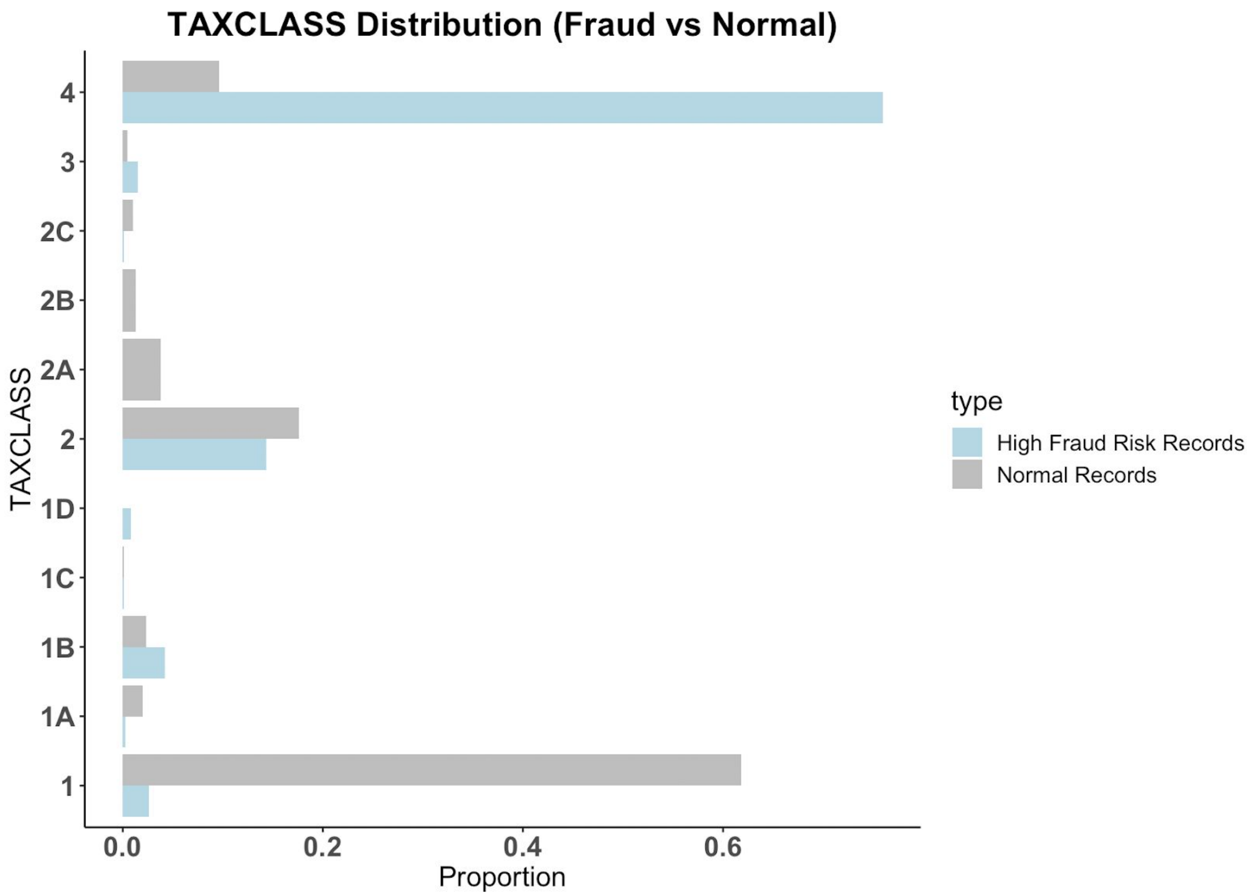
1) Based on the table above, we found that high fraud risk records have especially high FULLVAL, AVLAND, AVTOT. This indicates potential fraud properties tend to have larger land area and be reported with especially high market value and assessed value.

2) Besides the value of the property, we found potential fraud properties have especially high values for STORIES, LTfront, LTdepth. However, the mean and median of BLDfront and BLDdepth of potential fraud properties are not a lot higher than those of normal records, suggesting potential fraud records tend to be reported with normal building front and normal building depth but with especially high stories, lot frontage and lot depth.

Next, we would explore categorical fields and compare the difference between fraud records and normal records.



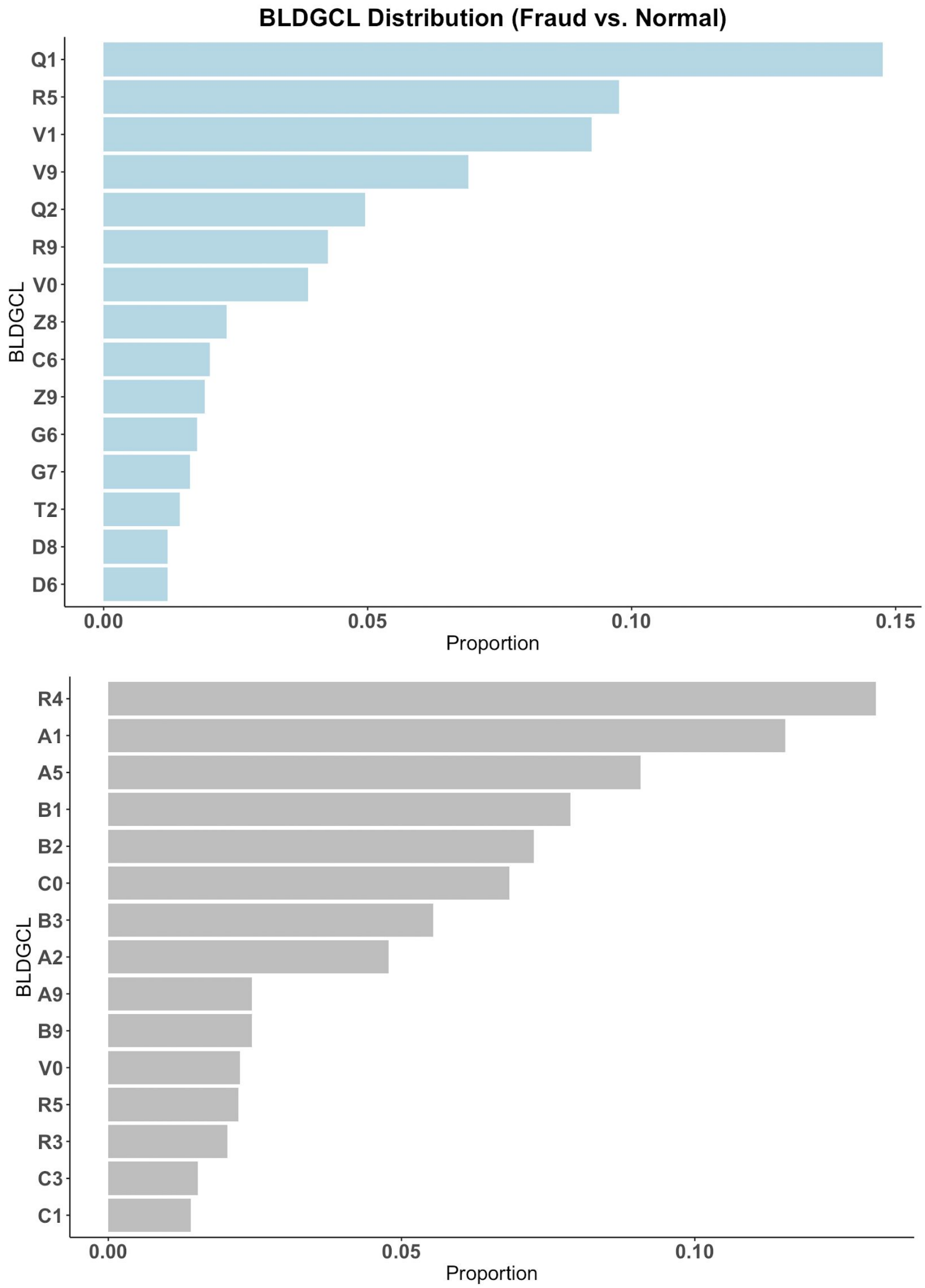
For Borough distribution, as we can see from the above plot, over 30% of fraud records are located in Manhattan. On the other hand, only less than 15% of normal scores are located there. Noticeably, the proportion of normal scores in Brooklyn is significantly higher than that of fraud scores.



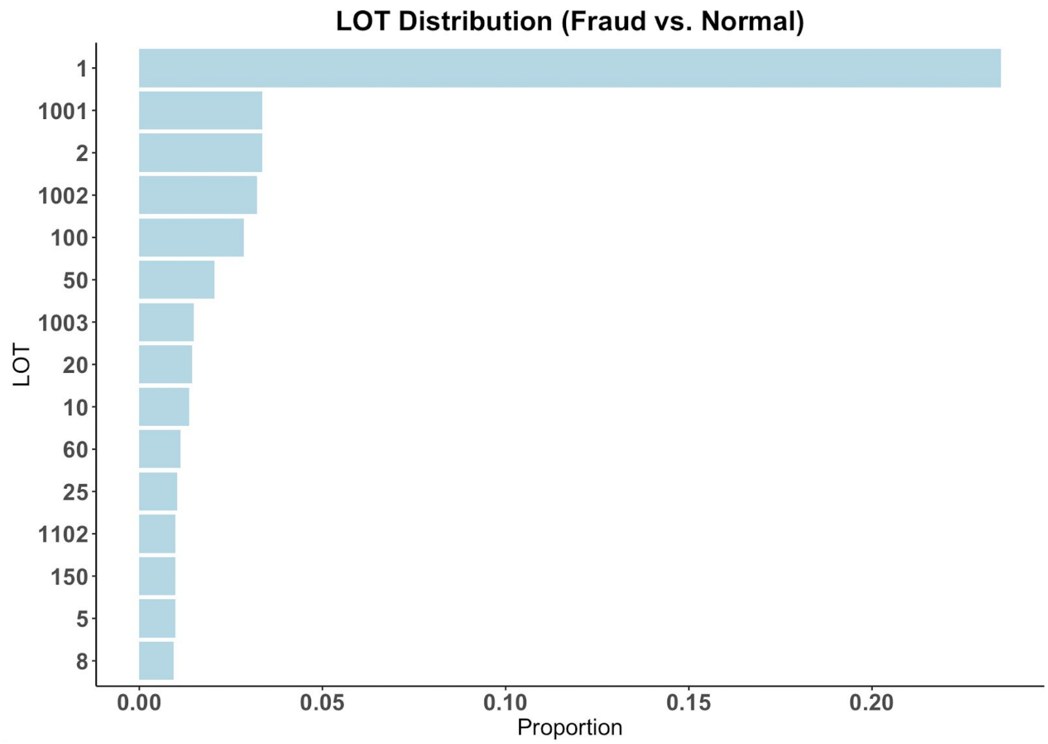
By comparing TAXCLASS of fraud records and normal records, we can see that over 70% of fraud records have a tax class of 4. On the other hand, normal records are much more likely to have a tax class of 1.

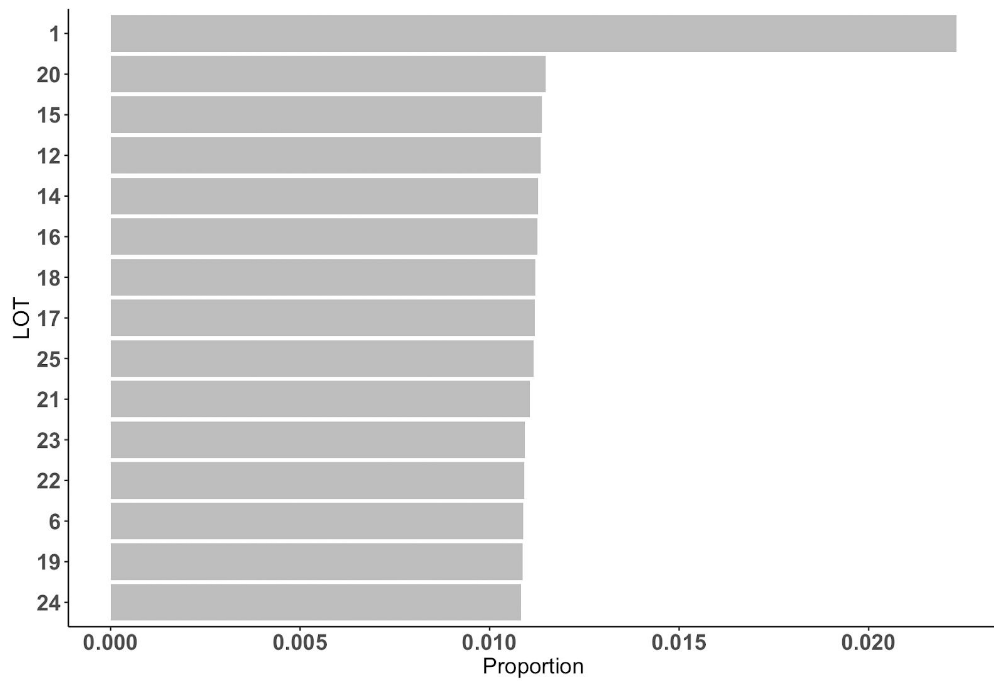
The top 15 common values of BLDGCL are given in the following plot. Since there is a direct correlation between the tax class and building class, the result is consistent with above. For fraud records, there are 10 out of top 15 building classes which correspond to tax class of 4. On the other hand, for normal records, there are 10 out of top 15 building classes which correspond to tax class of 1.

Tax Class	Building Class
1	A0 - A9, B1 - B9, C0, G0, R3, R6, R7, S0 - S2, V0, V2, V3, Z0
2	C1 - C9, D0 - D9, R0, R1, R2, R4, R8, R9, S3, S4, S5, S9
3	U1 - U2, U4 - U9
4	ALL OTHERS

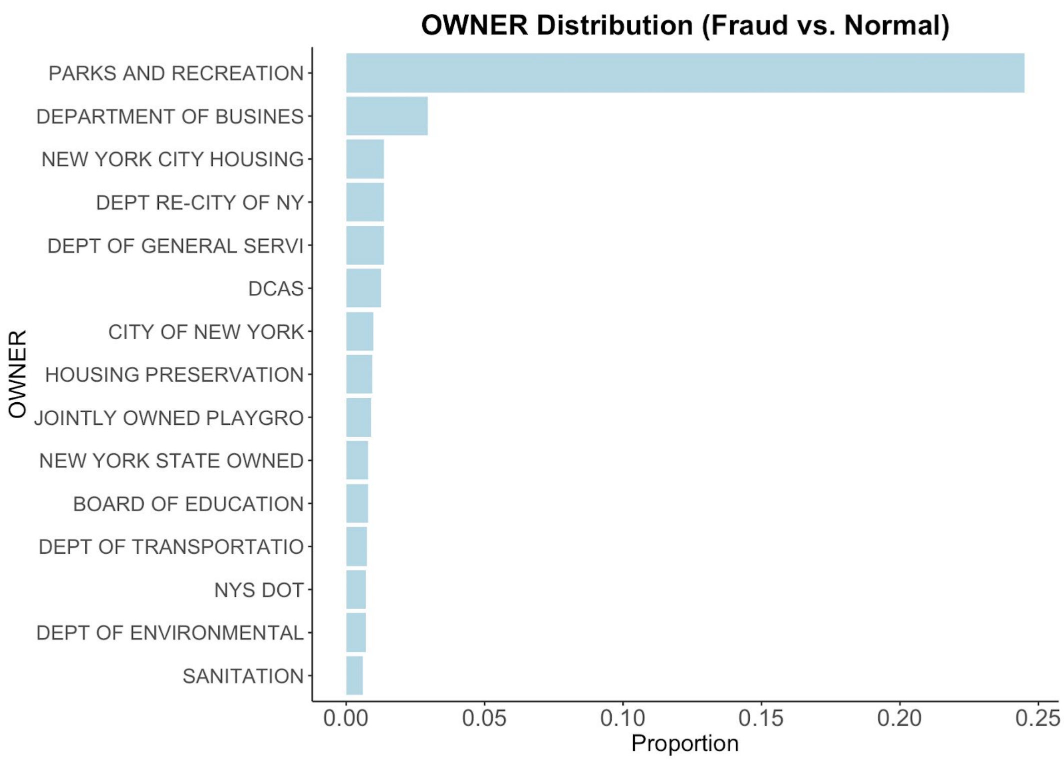


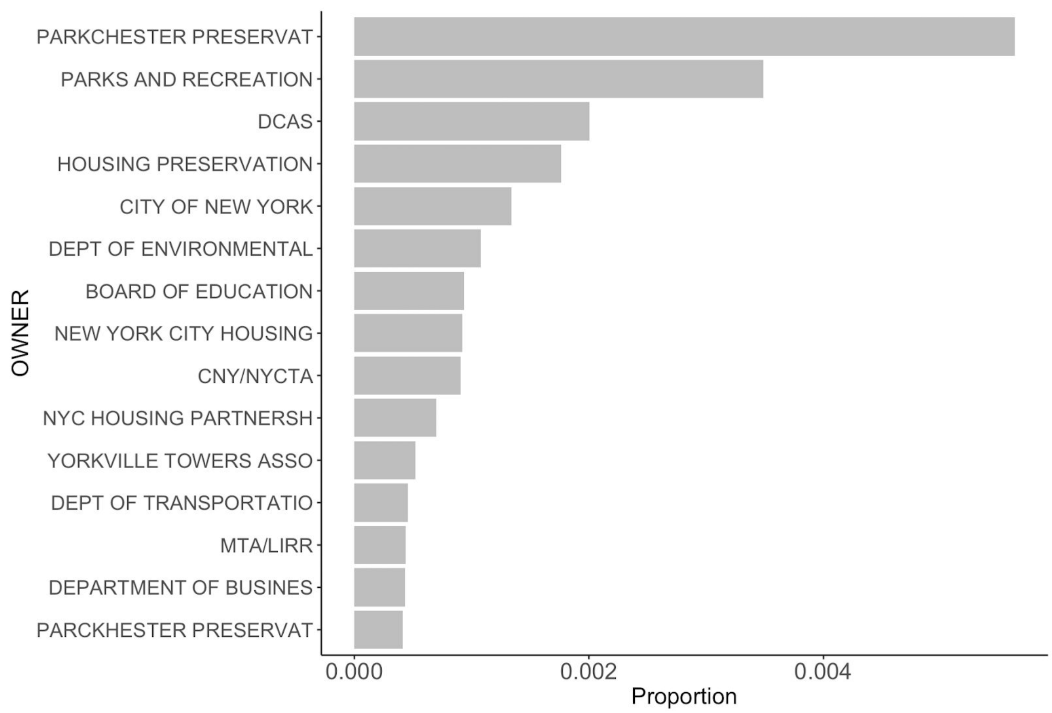
From the distribution of unique number of lots within blocks (LOTS), it is obvious that compared with normal records, there are about 10% of fraud records have a lot value of more than 1000.





The top 15 values of OWNER is given in the following plot. As we can see, about 25% of fraud records which belong to PARKS AND RECREATION, which is worth us investigating further.


OWNER Distribution (Fraud vs. Normal)
PARKS AND RECREATION
DEPARTMENT OF BUSINES
NEW YORK CITY HOUSING
DEPT RE-CITY OF NY
DEPT OF GENERAL SERVI
DCAS
CITY OF NEW YORK
HOUSING PRESERVATION
JOINTLY OWNED PLAYGRO
NEW YORK STATE OWNED
BOARD OF EDUCATION
DEPT OF TRANSPORTATIO
NYS DOT
DEPT OF ENVIRONMENTAL
SANITATION
OWNER
0.00
0.05
0.10
0.15
0.20
0.25
Proportion


PARKCHESTER PRESERVAT
PARKS AND RECREATION
DCAS
HOUSING PRESERVATION
CITY OF NEW YORK
DEPT OF ENVIRONMENTAL
BOARD OF EDUCATION
NEW YORK CITY HOUSING
CNY/NYCTA
NYC HOUSING PARTNERSH
YORKVILLE TOWERS ASSO
DEPT OF TRANSPORTATIO
MTA/LIRR
DEPARTMENT OF BUSINES
PARCKHESTER PRESERVAT
OWNER
0.000
0.002
0.004
Proportion

10 Records with Top Fraud Scores

Now, let's walk through the top 10 fraud scores. The following tables display data after filling in missing values based on fraud scores in descending order:

RECORD	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	FULLVAL	AVLAND	AVTOT
632816	157	95	1	1	2,930,000	1,318,500	1,318,500
917942	25	100	20	39	374,019,900	1,792,809,000	4,668,309,000
565392	117	108	20	39	4,326,304,000	1,946,837,000	1,946,837,000
67129	840	868	20	39	6,150,000,000	2,668,500,000	2,767,500,000
565398	466	1009	20	39	2,310,884,000	1,039,898,000	1,039,898,000
918204	8000	2600	20	39	1,662,400,000	748,080,000	748,080,000
585118	298	402	1	1	3,443,400	1,549,530	1,549,530
85886	4000	150	8	8	70,214,000	31,455,000	31,596,300
585120	139	342	1	1	2,151,600	968,220	968,220
585439	94	165	1	1	3,712,000	252,000	1,670,400

RECORD	B	OWNER	BLDGCL	TAXCLASS	ZIP	STORIES
632816	4	864163 REALTY, LLC	D9	2	11373	1
917942	4	LOGAN PROPERTY, INC.	T1	4	11422	3
565392	3	U S GOVERNMENT OWNRD	V9	4	11426	2
67129	1	CULTURAL AFFAIRS	Q1	4	10028	2
565398	3	DEPT OF GENERAL SERVI	V9	4	11426	2
918204	4	U S GOVERNMENT OWNRD	V9	4	11211	1
585118	4	NEW YORK CITY ECONOMI	O3	4	11101	20
85886	1	PARKS AND RECREATION	Q1	4	10303	1
585120	4	NaN	O3	4	11217	20
585439	4	11-01 43RD AVENUE REA	H9	4	11101	10

We segregated those top 10 records into three groups based on their characteristics.

1) Government Owned Properties

This group includes record owned by U.S government--565392, 585398, 918204, 585118. Even though these three records are top ranked on fraud scores, we do not consider them as fraud incident.

RECORD	OWNER	LTfront	LTdepth	BLDfront	BLDdepth	FULLVAL	AVLAND	AVTOT
565392	U S GOVERNMENT OWNRD	1.06	0.05	-0.22	-0.25	373.26	479.82	282.95
565398	DEPT OF GENERAL SERVI	5.90	13.71	-0.22	-0.25	199.34	256.28	151.12
918204	U S GOVERNMENT OWNRD	110.33	37.82	-0.22	-0.25	143.38	184.36	108.71
585118	NEW YORK CITY ECONOMI	3.57	4.50	-0.78	-1.26	0.22	0.36	0.19

2) Records with Unusual Z-Scores

This group contains three records: 917942, 67129, 85886.

RECORD	OWNER	LTFRONT	LTDEPTH	BLDFRONT	BLDDEPTH	FULLVAL	AVLAND	AVTOT
917942	LOGAN PROPERTY, INC.	-0.22	-0.07	-0.22	-0.25	32.20	441.85	678.54
67129	CULTURAL AFFAIRS	11.08	11.57	-0.22	-0.25	530.64	657.69	402.24
85886	PARKS AND RECREATION	54.88	0.69	-0.57	-1.07	5.98	7.73	4.56

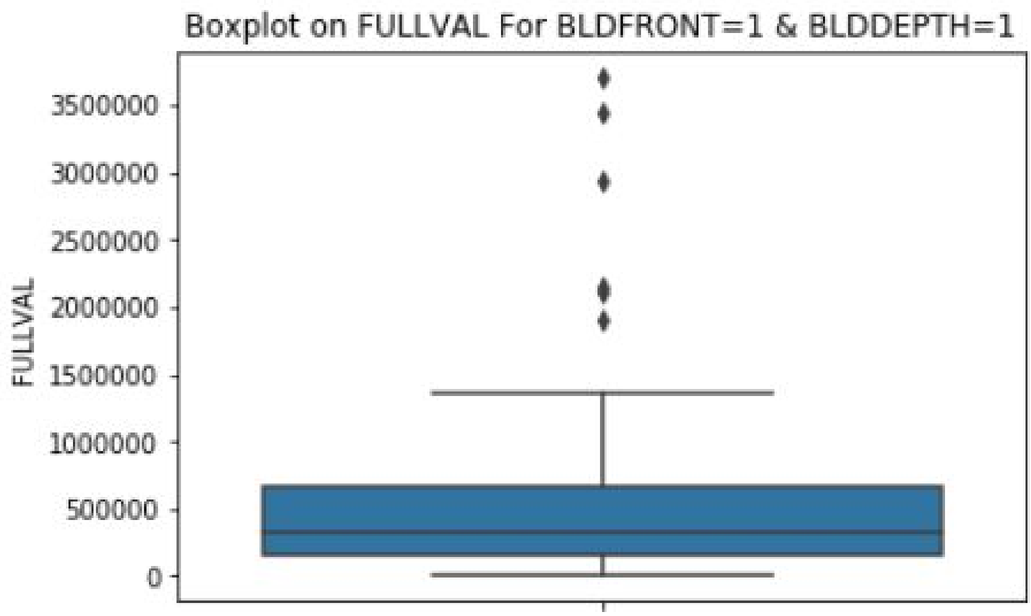
These records all have z scores that are extremely high. For example, AVLAND for record 917942 is 441 standard deviation away from the mean.

3) Records with Usual Z-Scores

This group contains 632816, 585118, 585120, 585439. However, as we mentioned above, 585118 is not counted as fraud because of its owner.

It is not obvious from this table why 632816, 585118, 585120, 585439 made to our top 10 records while all z scores are within a reasonable range. Therefore, we went back to the original data to see if there are any connections among these four records. Then we spot that these four records all share the same BLDFRONT and BLDDEPTH of one. From the initial data, 77 records have a BLDFRONT of one and 59 records have a BLDDEPTH of one. In the same token, 55 records have BLDFRONT and BLDDEPTH value at one, which only occupied 0.005% of the entire sample.

To further explore these four values, we abstract those 55 records.



We can tell from the box plot that the majority of FULLVAL concentrates around $300,000, and there are several outliers. Since AVLAND and AVTOT are similar to FULLVAL, we did the same process and found that this pattern apply to those two fields as well.

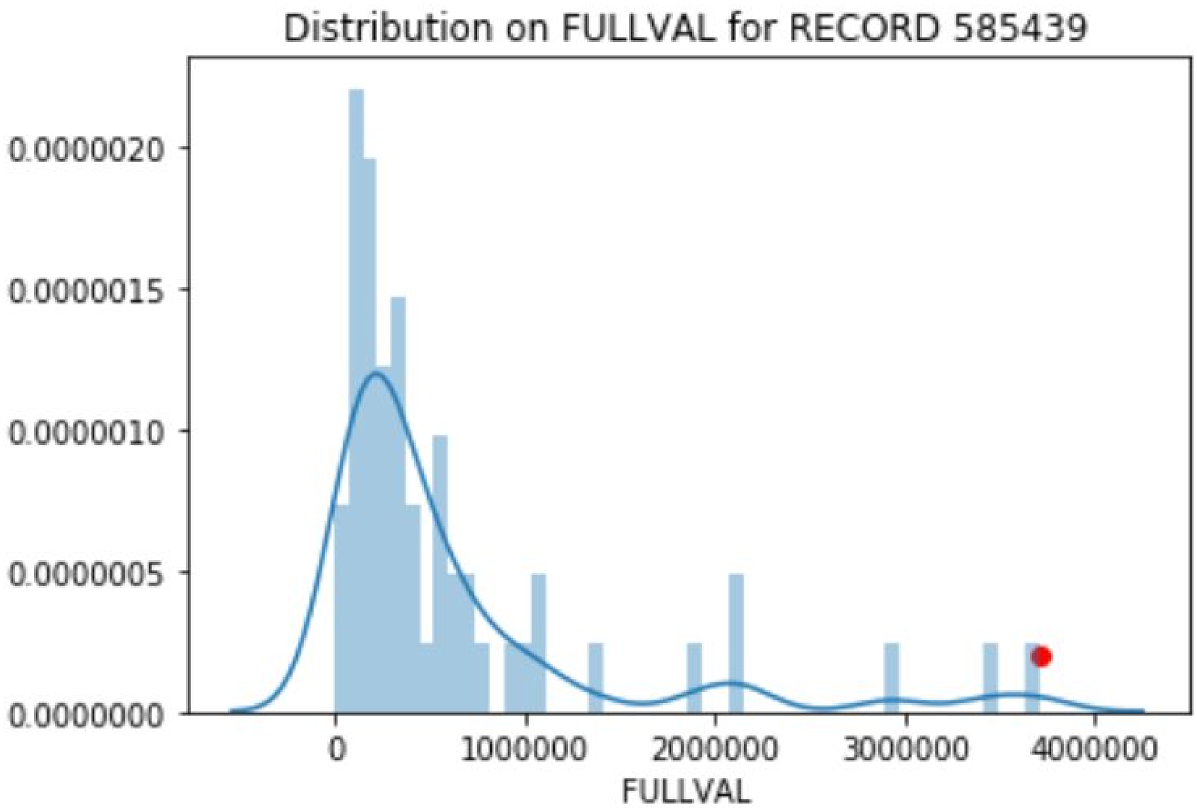
After sorting this subset based on FULLVAL on descending order, we found out that the four records from the top 10 fraud scores are also the four records with the highest FULLVAL, AVLAND, and AVTOT with an building area of 1.

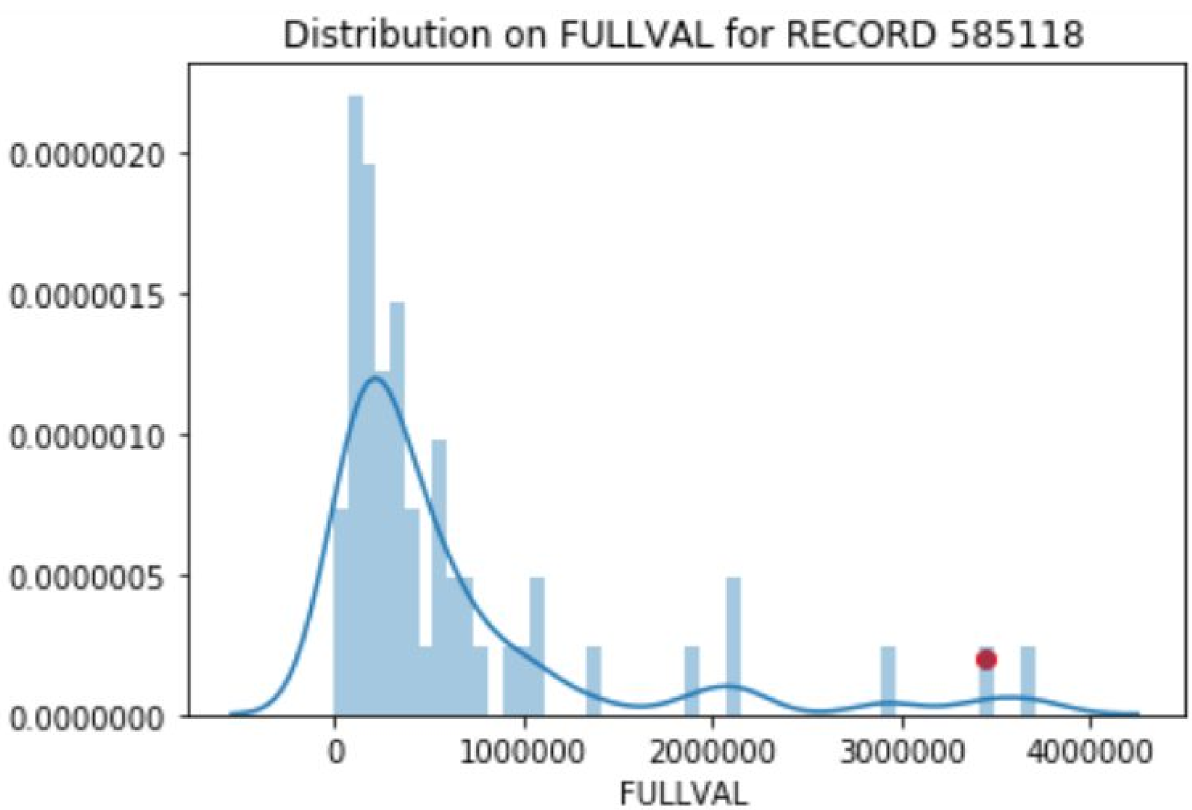
RECORD	FULLVAL
585439	3,712,000.00
585118	3,443,400.00
632816	2,930,000.00
585120	2,151,600.00
1067001	2,120,000.00
920628	1,900,000.00
794105	1,356,000.00
797936	1,088,000.00

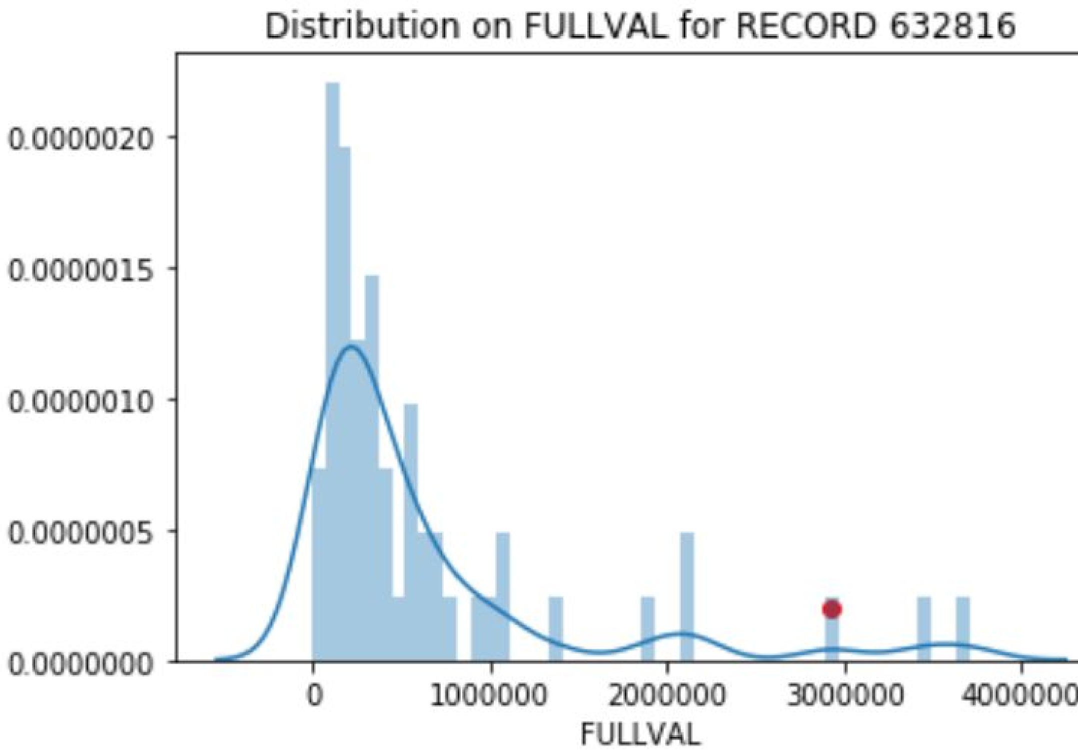
RECORD	AVLAND
585118	1,549,530.00
632816	1,318,500.00
585120	968,220.00
585439	252,000.00
935158	236,250.00
127153	112,500.00
886634	108,450.00
935081	101,250.00

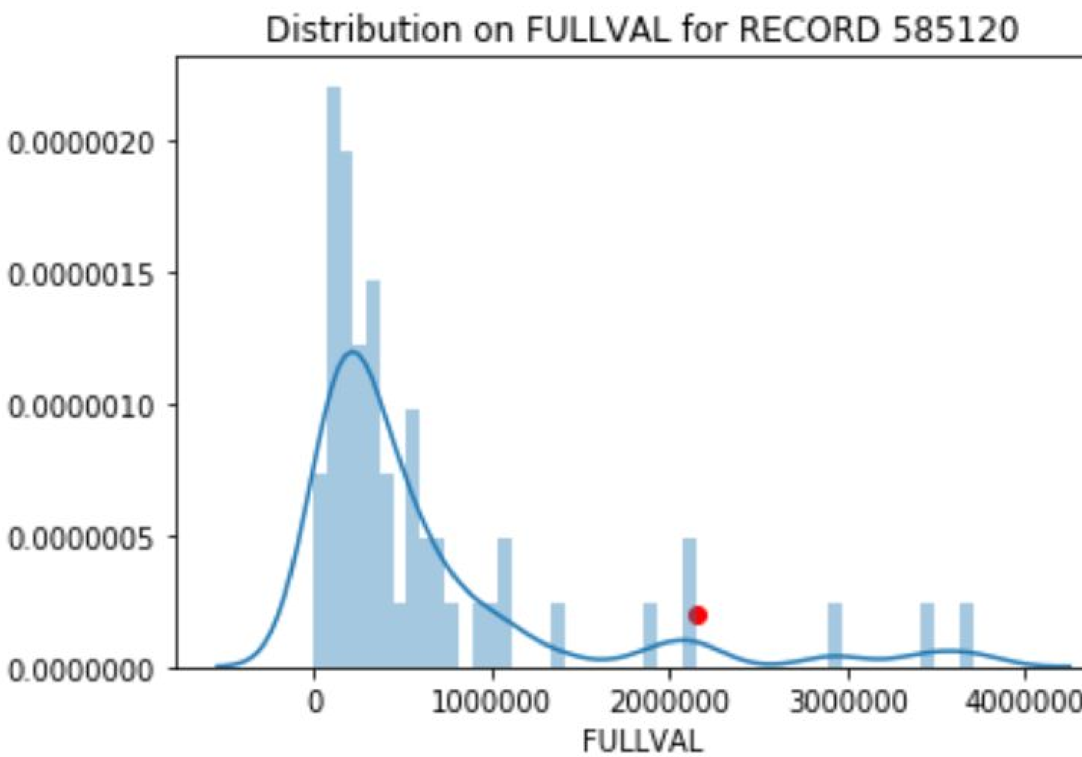
RECORD	AVTOT
585439	1,670,400.00
585118	1,549,530.00
632816	1,318,500.00
585120	968,220.00
935158	468,000.00
823224	281,250.00
127153	160,200.00
835933	129,600.00

Distribution of those four records and the 55 records are as followed:


Distribution on FULLVAL for RECORD 585439
0.0000020
0.0000015
0.0000010
0.0000005
0.0000000
0
1000000
2000000
3000000
4000000
FULLVAL


Distribution on FULLVAL for RECORD 585118
0.0000020
0.0000015
0.0000010
0.0000005
0.0000000
0
1000000
2000000
3000000
4000000
FULLVAL





Distribution of AVLAND and AVTOT follow similar distribution. For simplicity, we elimite those plots. Our assumption is value of a property is correlated with the size. As a result, a building with an area of one should not have a high market value. This explained what these four records appeared on the top 10 list.

Part VIII. Conclusions

We followed a standard process in this project: data preprocessing, dimension reduction, machine learning application. We first filled in missing values for the records we need for further steps: ZIP, STORIES, FULLVAL, AVLAND, AVTOT, LTFRONT, LTDEPTH. Then we calculated the ratios of value against size, for example, FULLVAL was divided by BLDVOL. In addition, we created 45 expert variables, which is calculated as the ratio compared to mean of that group. Before performing PCA to reduce dimensionality, we z-scaled all variables to ensure they have the same unit. After keeping the first seven PCs, we z-scaled those seven PCs again.

As we proceeded for further analysis, we created two models for predicting the fraud scores of the records. The first model that we used was a Heuristic model, where we added up the square of z-scores from all the variables of a record so that they don't cancel each other out. Once the z scores were squared, we added all the z scores per record and took the square root of the sum to give a single 'Euclidean' fraud score per record. The second model we used was an advanced Artificial Neural Network called Autoencoder, where we trained the model to predict the observed data. We then z-scaled the output from model and performed the same 'Euclidean' fraud score calculation on each fraud score. Now that we had two fraud scores i.e. score 1 from Heuristic model and score 2 from AutoEncoder, we took the average of these scores to calculate the final score. Finally we plotted the distribution of these 3 score to identify the outliers. We then ranked all the score based on the final fraud score, here a rank 1 means the record is at the least chance of being fraud and the top potential fraud record has a rank of 1070994.

For a deeper analysis, we identified the top 10 records with the highest fraud scores.Then we evaluated each record to determine if they are true fraudulent records. We first took out properties owned by U.S government, and those records are not likely to be fraud; then we check records with high z-scores compared to the entire dataset. Lastly, we explored more on records with normal z-scores, and we found out those records have extreme value on FULLVAL, AVLAND, AVTOT from the same group.

If we have more time, we would like to create more expert variables beside those 45 variables. Also, we did not pay much attention to fields besides the ones we used to create 45 variables. Some potential fields we can dig more are ZIP and BLOCK. It is possible that certain area has higher percentage of fraudulent records.

Also, we only did 10 epochs for auto encoder since our initial attempt of 100 epochs took more than four hours. As a result, we kept the number of iteration low to shorten processing time. With more available time and processing power, we can increase the number of epochs to see if the result would be different.

Part IX. Appendix