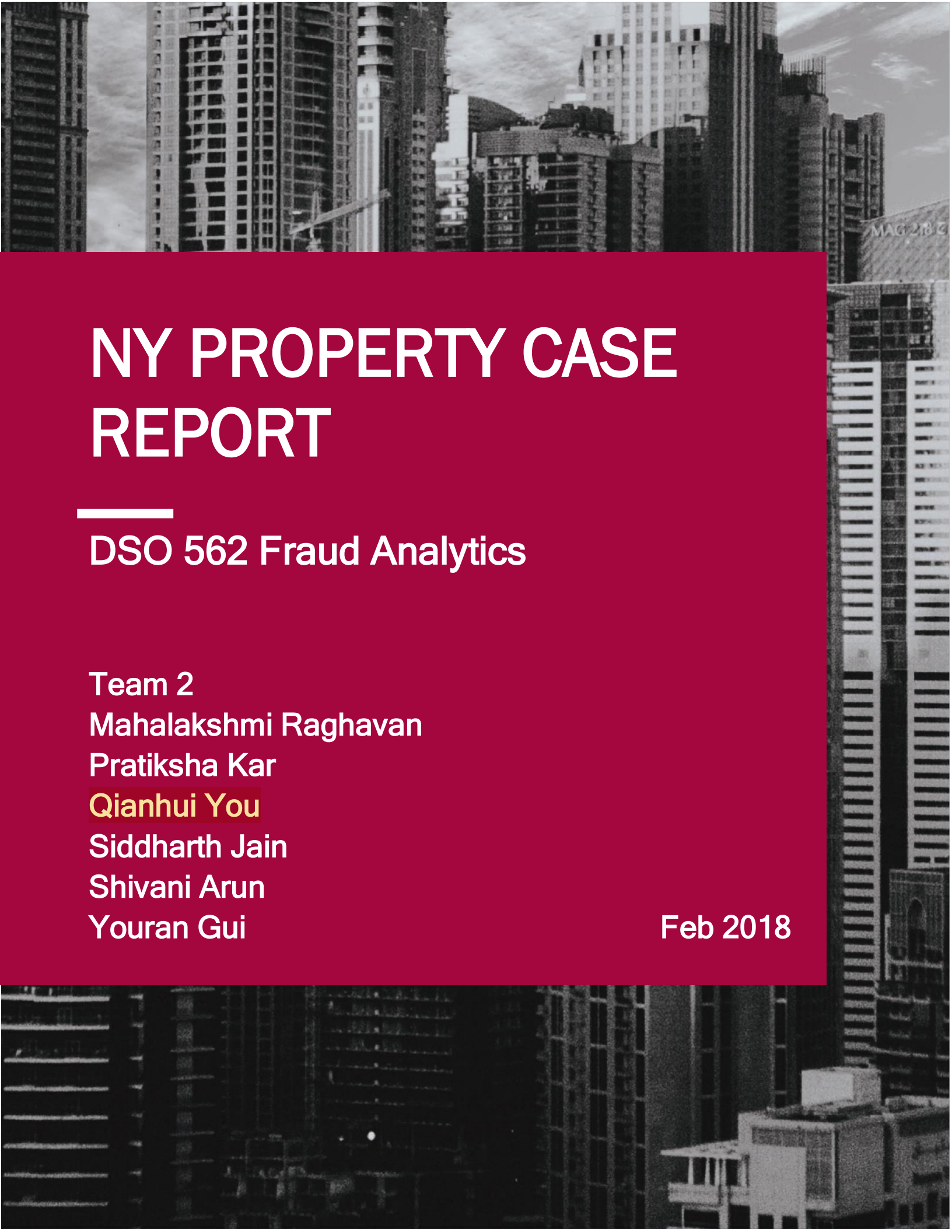


NY PROPERTY CASE REPORT

DSO 562 Fraud Analytics

Team 2
Mahalakshmi Raghavan
Pratiksha Kar
Qianhui You
Siddharth Jain
Shivani Arun
Youran Gui

Feb 2018

Table of Contents

Executive Summary

Housing valuation fraud is conducted when the assessed value greatly deviates from the true worth of the property. In this project, we are looking to build a fraud algorithm that can effectively detect the abnormality in valuation records. The dataset we analyzed is NYC Property Valuation and Assessment Data, which consists of 1,048,575 valuation records created in 2011 by NYC Department of Finance. There are 29 original variables in the dataset. These variables include different valuation and assessment values, spatial measurements, geographical information and regulatory codes. We selected 15 of those variables based on expert knowledge, data exploration and additional research and cleaned the data by replacing null values with grouped average. The detailed data cleaning methodology can be found in the Data Cleaning section.

While many of the variables can be potential fraud signals, our domain expert suggested that the market value, assessed land value and assessed total value of the property are the keys to identify this type of fraud. Furthermore, we adjusted these three variables with different property dimensions and spatial information and compared each value with the average value within similar cohorts. After the process, we created a total of 80 expert variables to be used in our fraud model. The variable creation section documents the creation of each of the variables.

After creating the variables, we proceeded on a z-scaling method in order to set the variables in the same comparable scale. However, with 80 variables in hand, high correlation and dimensionality among variables would significantly affect the accuracy and performance of our algorithm. Therefore, we conducted a principle component analysis (PCA) to reduce the dimensionality. From the PCs generated, we chose the top 10 PCs, which can explain about 90% of the variation. After PCA, we z-scaled the top 10 PCs again to prepare them for our fraud algorithm.

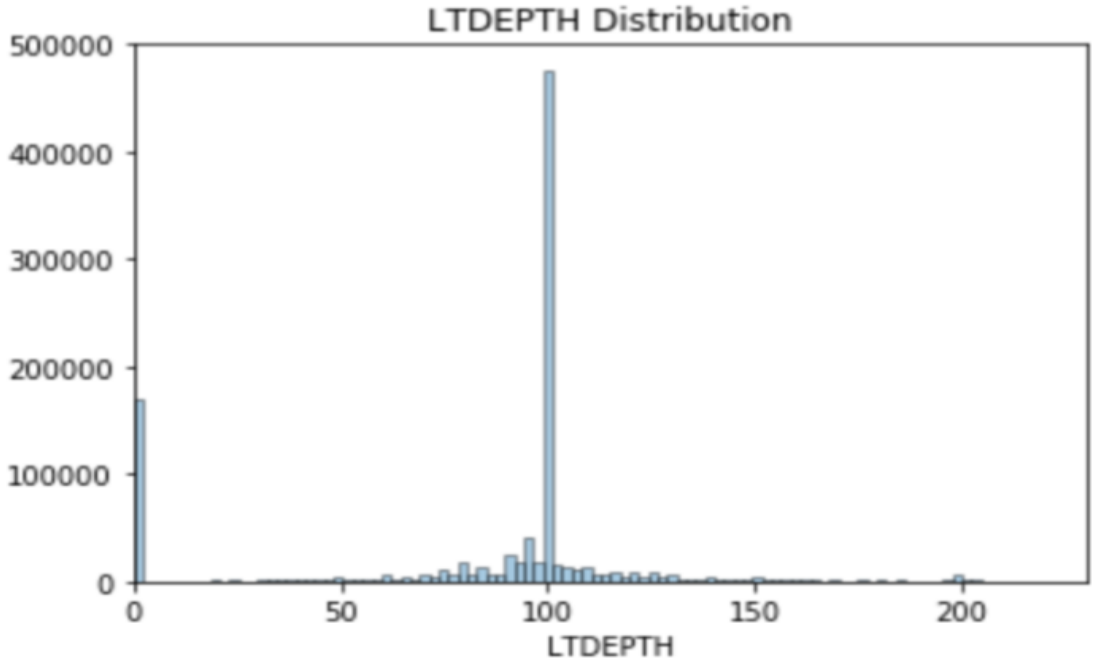
We applied two different algorithms to generate fraud scores. The first one is a heuristic model, in which we summed the squared z-scores to calculate the Euclidean distance for each record. The second technique applied is Autoencoder, which reproduces the input without the need for labels. It learns the general pattern followed by the data and hence highlights any anomalies. We took the sum of the squared differences between the original values and their respective reproduced values. This gave us yet another set of Euclidean distances. In order to make these two different results more comparable, we used quantile binning. With 3,075 bins and 341 records in each bin, we assigned two scores to each record based on its overall ranking obtained after applying the two fraud algorithms. As a final result, we took the average of the two scores to be our final fraud score. Please refer to the Algorithm Section for a complete explanation of our fraud algorithm.

The results of our fraud model demonstrated a successful detection of abnormal records. Our fraud model flagged a number of records, which either had an unusually high valuation compared to their dimensions or a disproportional market value versus assessed value. The distribution of the fraud score and the top 11 abnormal records can be found in the Algorithm and Result sections.

Data Description

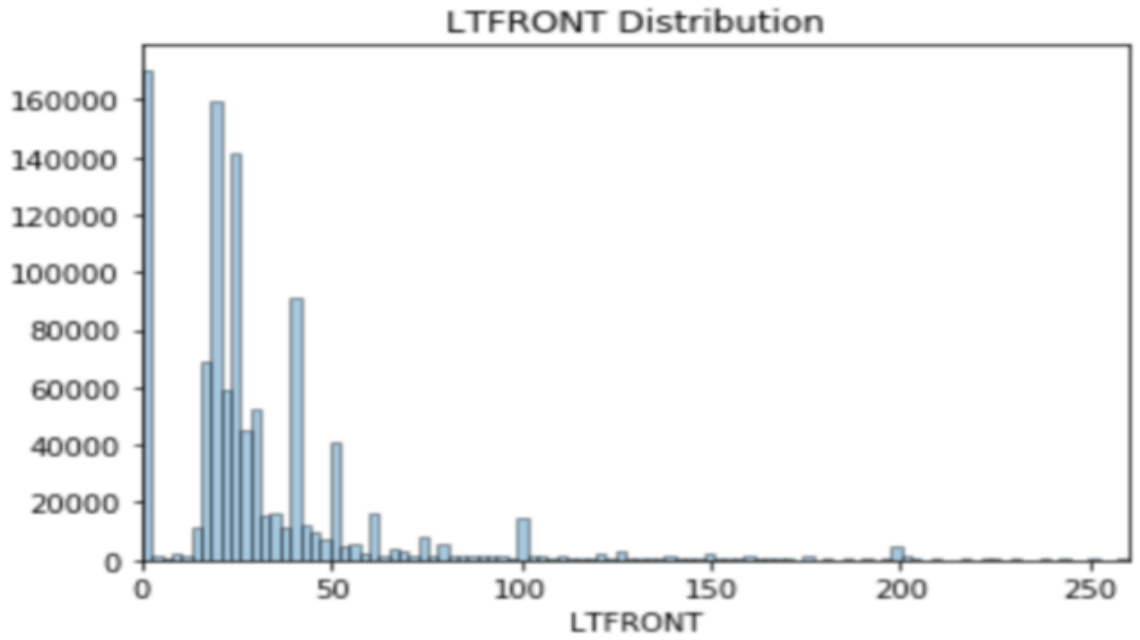
New York data is a property valuation and assessment data. It provides descriptive information, market value, assessed value and other miscellaneous information about the properties and land in New York. This dataset consists of 1,048,575 observations and 30 fields describing these observations. The observations are uniquely identified by the "record" field, which was manually added to the data. Catering to our goal of finding unusual records in the dataset, we majorly concentrate on the following fields:

1. **LTDEPTH:** Measurement of depth of the lot in feet.
   - Minimum: 0
   - Maximum: 9,999
   - Mean: 88.27

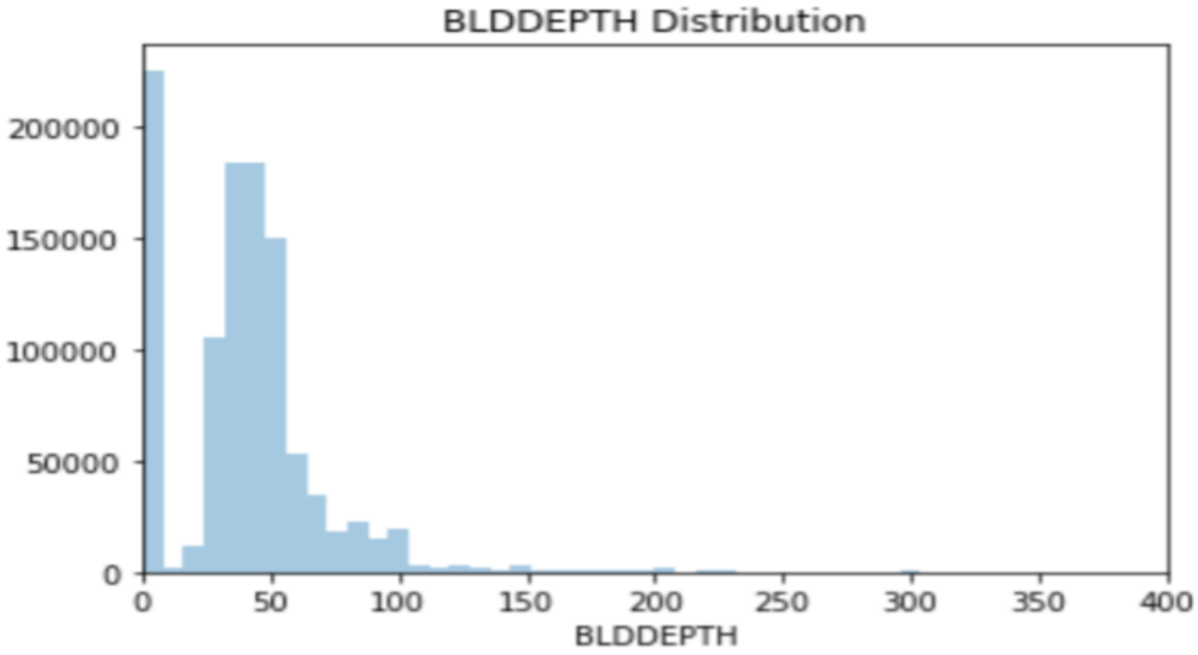


2. **LTFRONT:** Measurement of the front lot.
   - Minimum: 0
   - Maximum: 9999

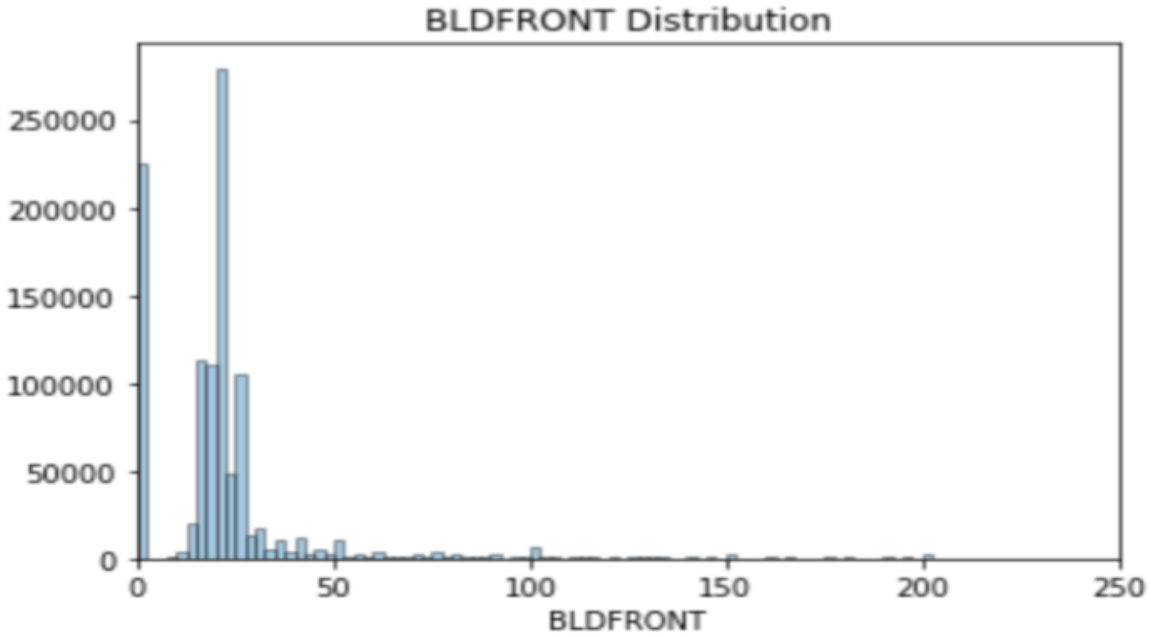
- Mean: 36.17



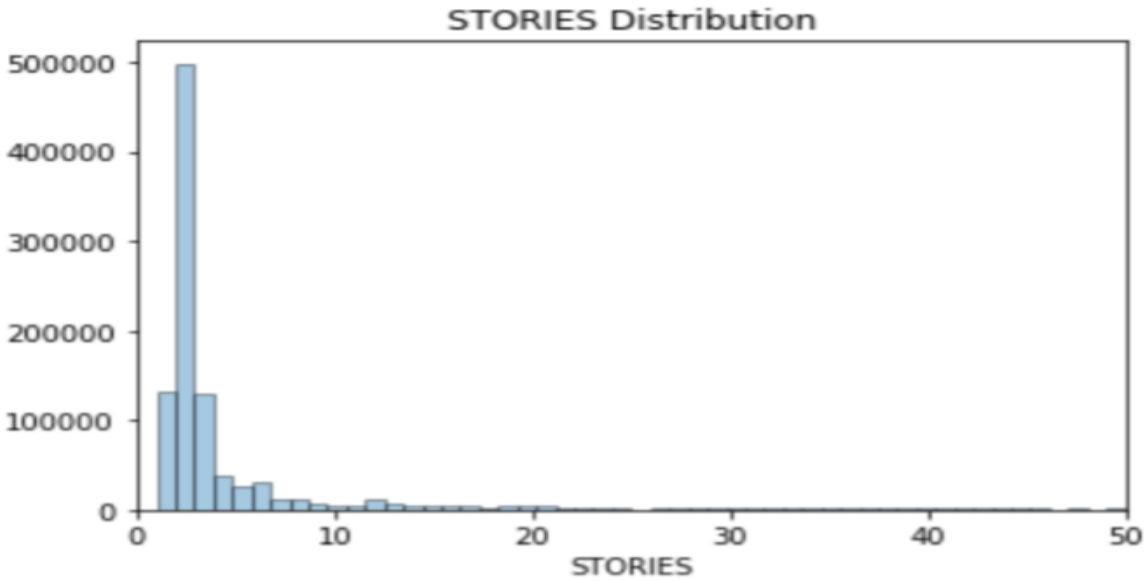
3. **BLDDEPTH:** Measurement of building depth.
   - Minimum: 0
   - Maximum: 9,393
   - Mean: 40.07



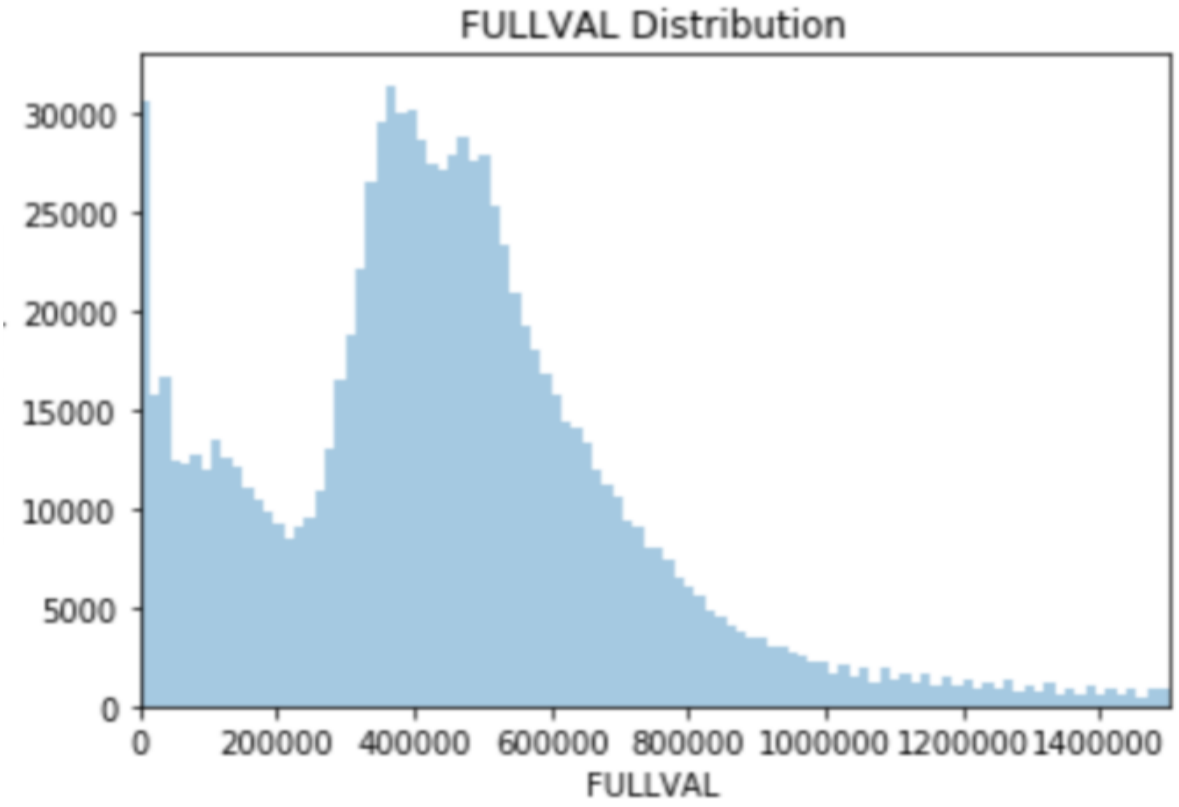
4. **BLDFRONT:** Measurement of building front.
   - Minimum: 0
   - Maximum: 7,575
   - Mean: 23.01



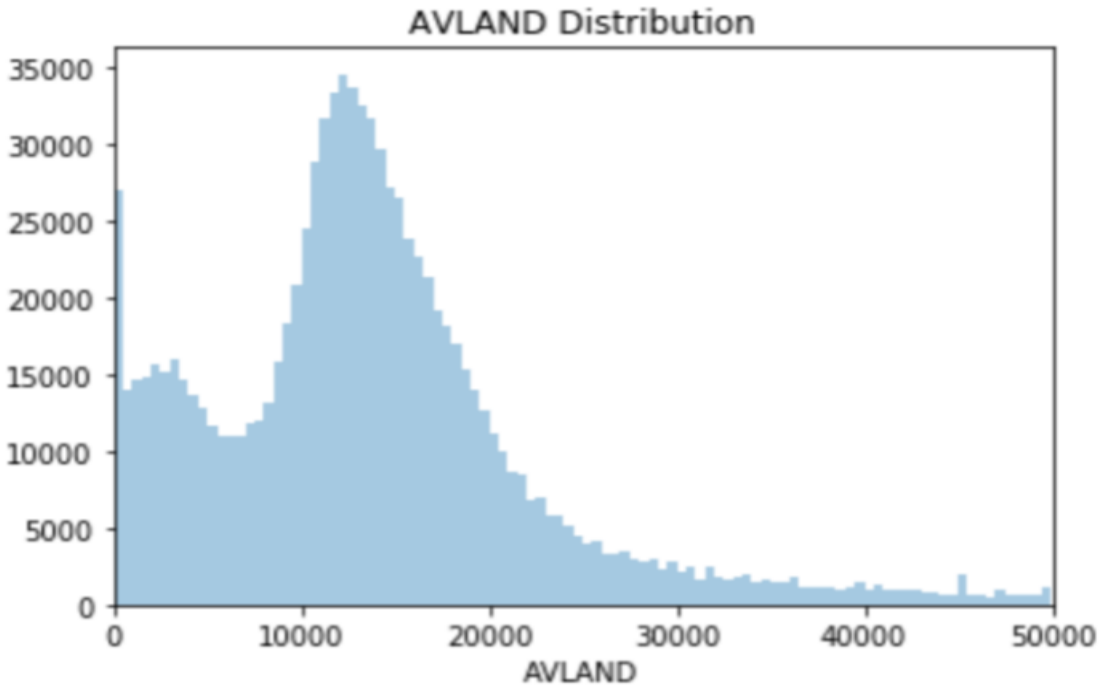
5. **STORIES:** Number of stories a building has.
   - Minimum: 1
   - Maximum: 119



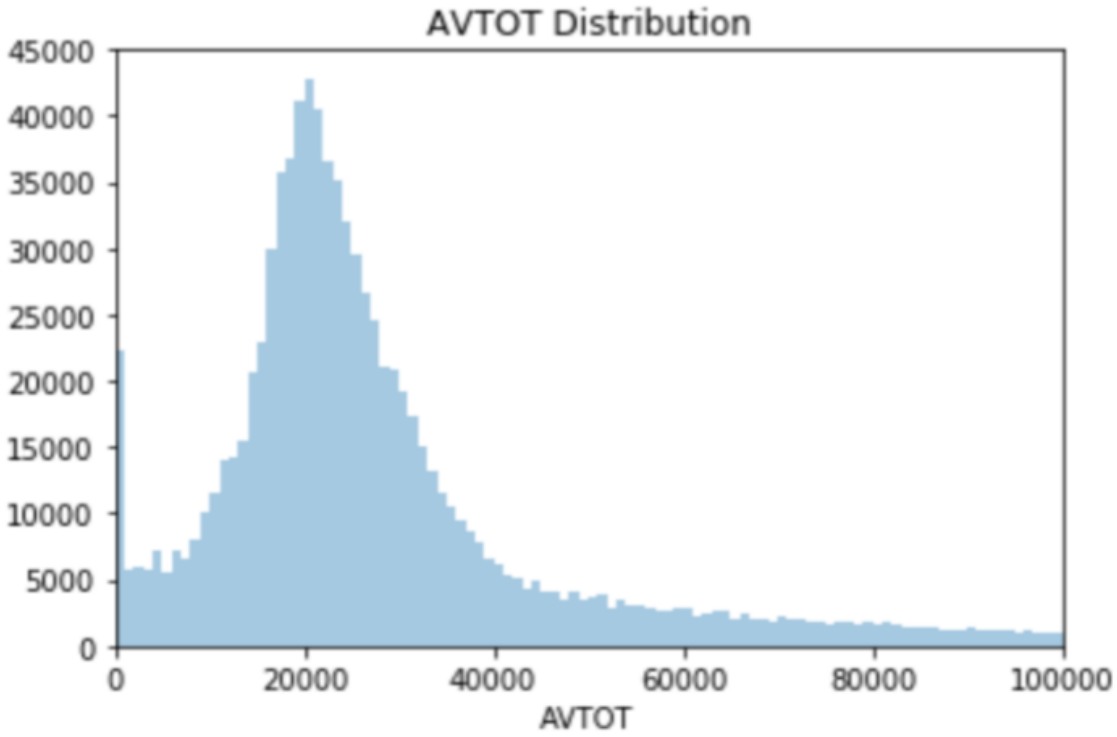
6. **FULLVAL:** Full value of the property.
   - Minimum: 0
   - Maximum: 6,150,000,000
   - Mean: 880,487.66



7. **AVLAND:** Assessed value of land.
   - Minimum: 0
   - Maximum: 2,668,500,000
   - Mean: 85,995.03



8. **AVTOT:** Total assessed value of the property.
   - Minimum: 0
   - Maximum: 4,668,308,947
   - Mean: 230,758.18



Data Cleaning

After deciding the variables for further analysis, we proceeded with data cleaning. Cleaning was done in the following ways:

1. FULLVAL, AVTOT, AVLAND:
   These three fields were populated by zeroes for some observations. To replace zeroes, we calculated averages for all the fields FULLVAL, AVTOT and AVLAND (grouped by tax-class) for non-zero observations and replaced the zeroes in observations with the mean value for the corresponding tax-class.

2. STORIES:
   For two tax-classes, namely 1B and 3, ~100% of the observation had "NA" as stories. To deal with this, we calculated the overall mean for stories and replaced "NA" by that value.

   Note: We ignored tax-class based average for these two tax classes as for class 1B only two records had a numerical entry for stories and for class 3 only three records had a proper value for stories. These values might not be representative of the bigger picture as they are a minority's

   For other tax classes, we calculated the class-wise average and replaced "NA" values with the corresponding mean values.

3. LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH:
   For these fields, we calculated the mean (grouped by tax-class), not considering the records with zero or "NA" values, then replaced the missing values and zeroes with mean values corresponding to the tax-class to which the record belongs.

   We considered grouping by tax-class because of a direct relationship between this field and valuation/size of the property.

Variable Creation

After consulting with a housing valuation expert, we identified three most crucial variables (monetary variables) for our fraud detection:

- **FULLVAL**: Current market value of the property
- **AVLAND**: Assessed land value of current fiscal year
- **AVTOT**: Assessed total property value of current fiscal year

Although the above fields are strong fraud signals, we would not want to compare these parameters directly, as the properties vary dramatically in lot size, building size, and number of stories. Therefore, it is necessary to adjust these three monetary values by property dimensions. The adjustment we made was based on five building measurements (spatial variables), which were either calculated or extracted directly from the dataset:

- **STORIES**: Number of stories in the building
- **LOTAREA**: Lot area calculated from *LTFRONT * LTDEPTH*
- **BLDAREA**: Building area calculated from *BLDFRONT * BLDDEPTH*
- **LOTAREA/BLDAREA**: Ratio between lot area and building area
- **BLDVOL_ADJ**: Building volume is calculated from *BLDAREA * STORIES* and adjusted based on the number of times the same *STADDR (street address)* appears in the dataset

  Note: This adjustment is due to the reason that valuation on large properties might not always represent the value of the entire property. Separate valuations can be made for individual units within the same property. When this is the case, the adjustment divides the building volume by the number of records sharing the same address.

By adjusting each of the three monetary variables with the five spatial variables, we created 15 variables to be used later in our fraud model. The formula for each of these variables can be found in the table at the end of this section.

After spatial adjustment, we also wanted to compare each property with other properties in the same region or same tax class. One way to do so was to first group the properties based on criteria such as zip code and tax class, then calculate the within-group average of each of the 15 variables created above. Finally, we calculated the ratio between each individual value and its respective group average to form a new set of variables. The five variables we used for grouping are:

- **ZIP5**: The five-digit zip code of the property
- **ZIP3**: The first three digits of the zip code
- **BORO**: One-digit code for different regions of New York City
  1 - Manhattan
  2 - Bronx
  3 - Brooklyn
  4 - Queens
  5 - Staten Island
- **TAXCLASS**: The tax class of the property
- **BLDGCL**: The building class of the property

We applied each grouping ratio to some of the 15 variables and created an additional 63 variables.

In total, we created 80 expert variables and the formula of each can be found in the following table.

First Layer of variables

Variable Name	Formula
AVLAND_AVTOT	AVLAND/AVTOT
FULLVAL_AVTOT	FULLVAL/AVTOT
FULLVAL_STORIES	FULLVAL/STORIES
AVLAND_STORIES	AVLAND/STORIES
AVTOT_STORIES	AVTOT/STORIES
FULLVAL_LOTAREA	FULLVAL/LOTAREA
AVLAND_LOTAREA	AVLAND/LOTAREA
AVTOT_LOTAREA	AVTOT/LOTAREA
FULLVAL_BLDAREA	FULLVAL/BLDAREA
AVLAND_BLDAREA	AVLAND/BLDAREA
AVTOT_BLDAREA	AVTOT/BLDAREA
LOTAREA_BLDAREA_FULLVAL	FULLVAL*(LOTAREA/BLDAREA)
LOTAREA_BLDAREA_AVLAND	AVLAND*(LOTAREA/BLDAREA)
LOTAREA_BLDAREA_AVTOT	AVTOT*(LOTAREA/BLDAREA)
FULLVAL_BLDVOL	FULLVAL*(BLDAREA*STORIES/STADDR COUNT)
AVLAND_BLDVOL	AVLAND*(BLDAREA*STORIES/STADDR COUNT)
AVTOT_BLDVOL	AVTOT*(BLDAREA*STORIES/STADDR COUNT)

Second Layer of variables created with the formula x/x̄i

x/i	ZIP5	ZIP3	BORO	TAX CLASS	BLDGCL
FULLVAL_STORIES	FULLVAL_STORIES_ZIP5	FULLVAL_STORIES_ZIP3	FULLVAL_STORIES_BORO	-	-
AVLAND_STORIES	AVLAND_STORIES_ZIP5	AVLAND_STORIES_ZIP3	AVLAND_STORIES_BORO	-	-
AVTOT_STORIES	AVTOT_STORIES_ZIP5	AVTOT_STORIES_ZIP3	AVTOT_STORIES_BORO	-	-
FULLVAL_LOTAREA	FULLVAL_LOTAREA_ZIP5	FULLVAL_LOTAREA_ZIP3	FULLVAL_LOTAREA_BORO	FULLVAL_LOTAREA_TAXCLASS	FULLVAL_LOTAREA_BLDGCL
AVLAND_LOTAREA	AVLAND_LOTAREA_ZIP5	AVLAND_LOTAREA_ZIP3	AVLAND_LOTAREA_BORO	AVLAND_LOTAREA_TAXCLASS	AVLAND_LOTAREA_BLDGCL
AVTOT_LOTAREA	AVTOT_LOTAREA_ZIP5	AVTOT_LOTAREA_ZIP3	AVTOT_LOTAREA_BORO	AVTOT_LOTAREA_TAXCLASS	AVTOT_LOTAREA_BLDGCL
FULLVAL_BLDAREA	FULLVAL_BLDAREA_ZIP5	FULLVAL_BLDAREA_ZIP3	FULLVAL_BLDAREA_BORO	FULLVAL_BLDAREA_TAXCLASS	FULLVAL_BLDAREA_BLDGCL
AVLAND_BLDAREA	AVLAND_BLDAREA_ZIP5	AVLAND_BLDAREA_ZIP3	AVLAND_BLDAREA_BORO	AVLAND_BLDAREA_TAXCLASS	AVLAND_BLDAREA_BLDGCL
AVTOT_BLDAREA	AVTOT_BLDAREA_ZIP5	AVTOT_BLDAREA_ZIP3	AVTOT_BLDAREA_BORO	AVTOT_BLDAREA_TAXCLASS	AVTOT_BLDAREA_BLDGCL
LOTAREA_BLDAREA_FULLVAL	LOTAREA_BLDAREA_FULLVAL_ZIP5	LOTAREA_BLDAREA_FULLVAL_ZIP3	LOTAREA_BLDAREA_FULLVAL_BORO	-	-
LOTAREA_BLDAREA_AVLAND	LOTAREA_BLDAREA_AVLAND_ZIP5	LOTAREA_BLDAREA_AVLAND_ZIP3	LOTAREA_BLDAREA_AVLAND_BORO	-	-
LOTAREA_BLDAREA_AVTOT	LOTAREA_BLDAREA_AVTOT_ZIP5	LOTAREA_BLDAREA_AVTOT_ZIP3	LOTAREA_BLDAREA_AVTOT_BORO	-	-
FULLVAL_BLDVOL	FULLVAL_BLDVOL_ZIP5	FULLVAL_BLDVOL_ZIP3	FULLVAL_BLDVOL_BORO	FULLVAL_BLDVOL_TAXCLASS	FULLVAL_BLDVOL_BLDGCL
AVLAND_BLDVOL	AVLAND_BLDVOL_ZIP5	AVLAND_BLDVOL_ZIP3	AVLAND_BLDVOL_BORO	AVLAND_BLDVOL_TAXCLASS	AVLAND_BLDVOL_BLDGCL
AVTOT_BLDVOL	AVTOT_BLDVOL_ZIP5	AVTOT_BLDVOL_ZIP3	AVTOT_BLDVOL_BORO	AVTOT_BLDVOL_TAXCLASS	AVTOT_BLDVOL_BLDGCL

Algorithms

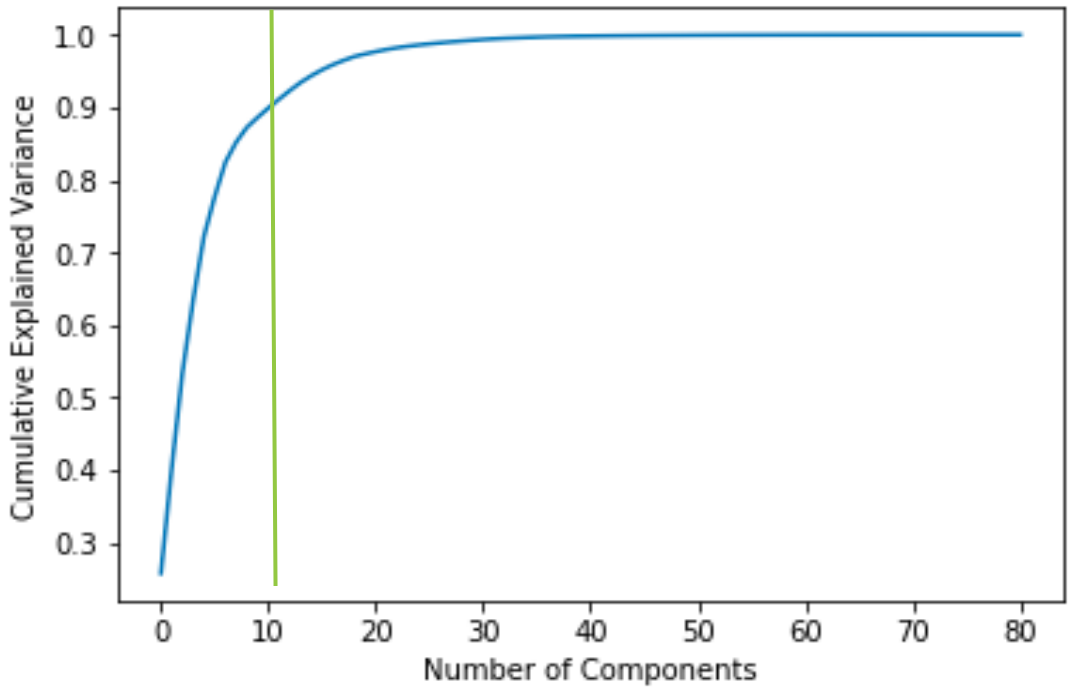
Following the creation of the expert variables that would be used in the algorithms to detect anomalous records, we went through various steps to arrive at the final fraud score. These steps are described below:

Z - Scaling: As different variables in this dataset have different scales (units) of measurement, it was essential to normalize data to get a reasonable covariance analysis. To do so, we used z-scaling to assign equal weights to all the 80 expert variables. This prepared the dataset for Principal Component Analysis.

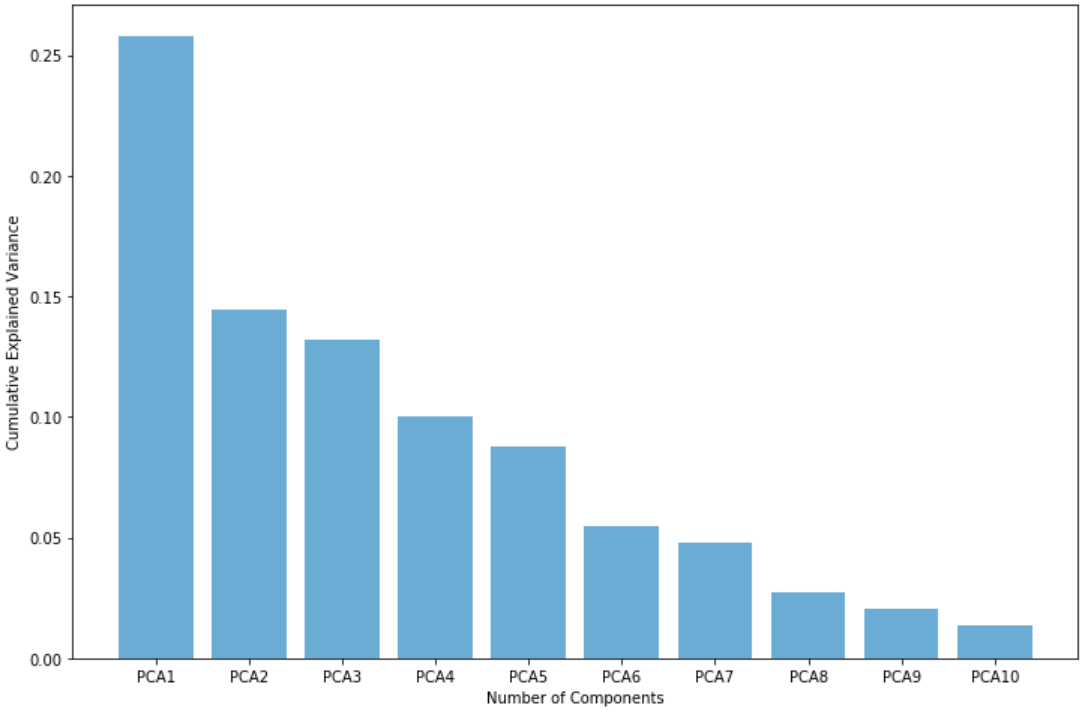
Principal Component Analysis: Next, we performed Principal Component Analysis (PCA) on 80 expert variables for dimensionality reduction. The idea behind PCA was to find a low-dimension set of axes that summarize the data. Since many of the 80 variables were highly correlated, it also helped to reduce the data to a parsimonious set of variables that still contained most of the variables from the larger set. The output variables, so obtained, were a linear combination of the input features. We z-scaled the results of the PCA algorithm once again to normalize the data.

Selection of Principal Components to Include: Each of the principal components explained a diminishing percentage of the variance in the data. We reduced the high dimensionality of our data and selected the first 10 Principle Components (PCs) which accounted for ~90% of the variation in our data.

This is because including additional number of PCs was explaining a very small percentage of variation in the data and was potentially adding significant noise. Further, the curve given below shows that it was appropriate to choose a cut-off point at 88-90% of the cumulative explained variance.



The graph below shows the explained variance for each of the top 10 PCs in a decreasing order.



Fraud Score Calculation

We used two methods to calculate fraud scores for each record:

a. **Heuristic Scoring:**
In this method, for the top 10 z-scored PCs, both Manhattan and Euclidean distance metrics were calculated.

Scoring Methodology:

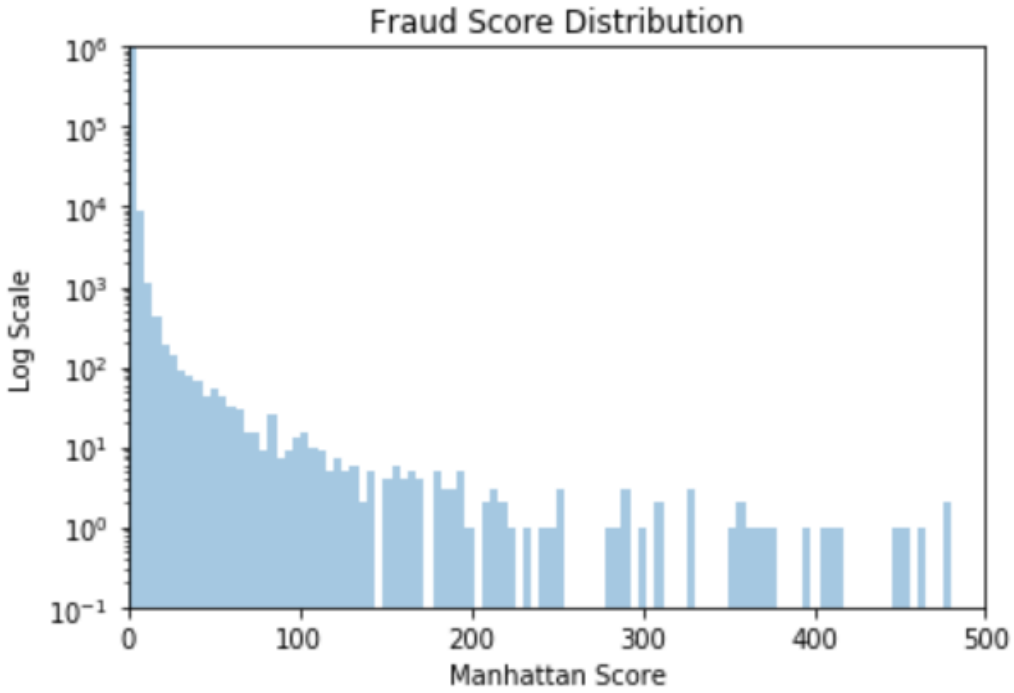
The scoring methodology used for both the distance metrics is given below:

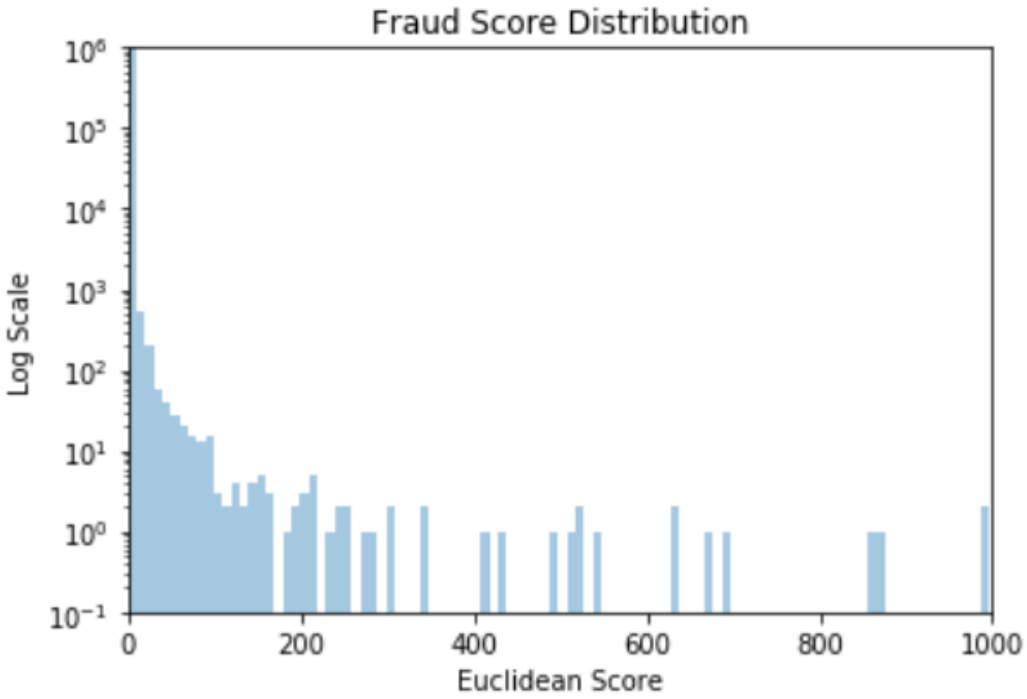
$$S = (\sum_i | z_i |^n)^{1/n}$$

where:
S = Heuristic Score
z_i = 10 Principle Components for $i \in [1,10]$

When n = 1, S represents the Manhattan Distance measure of Heuristic Score and when n = 2, S represents the Euclidean Distance measure of Heuristic Score.

For our analysis, we calculated both the Manhattan and the Euclidean Distance measures/Fraud Scores. The distributions of each of the fraud scores are given in the graphs below.





Both the distributions are right-skewed with a long tail and represent the shape of a regular fraud score distribution.

b. **Scoring using the reconstruction error from a trained Autoencoder:**

An Autoencoder is essentially a feed forward neural network (NN), as well as an unsupervised learning (feature learning) algorithm. An Autoencoder model predicts the input, given the same input, without the need for labels. To do this, it tries to learn an approximation of an "identity" function and applies the property of backpropagation, by setting the target value same as the input.

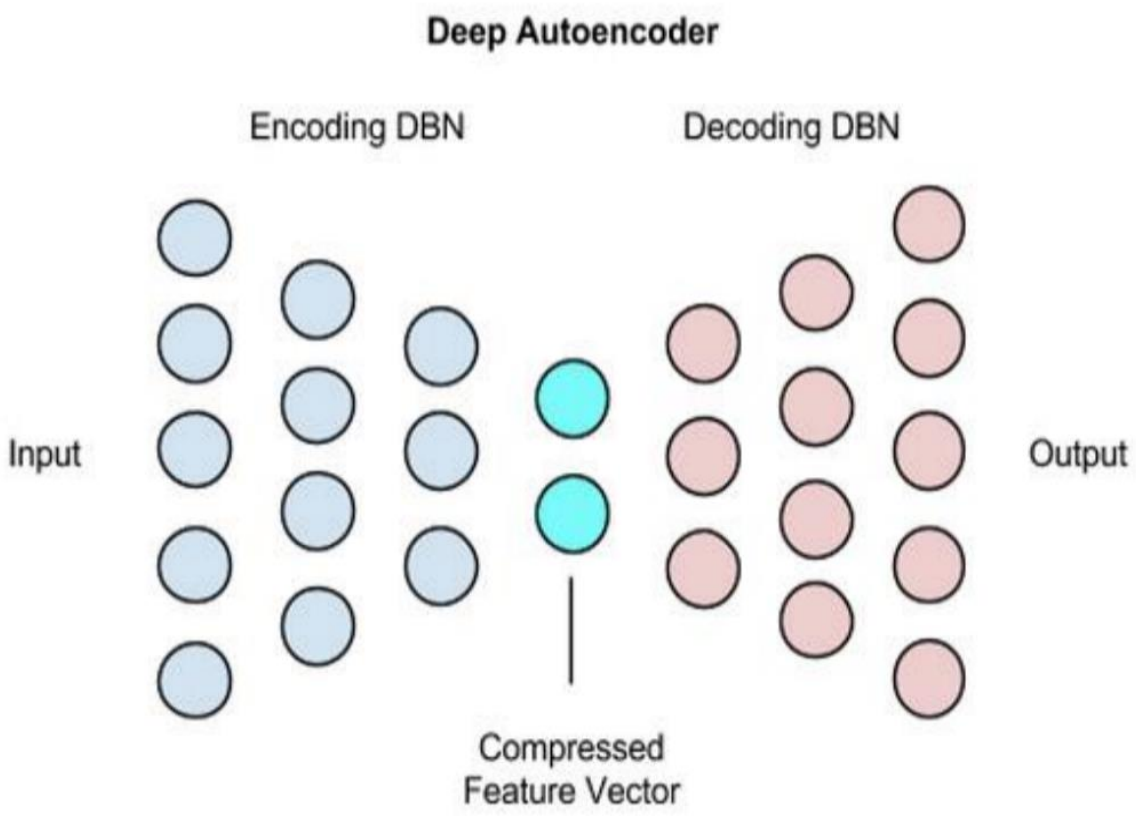
Since an Autoencoder learns from itself gradually, once an input dataset is provided, after several iterations (aka, epochs), it will produce an output, for each record of the input dataset. The normal records would be reconstructed almost as similar to their corresponding input records; however, unusual records would have a different reconstructed output.

This difference between the output and its corresponding input is defined as the reconstruction error. Records that have high reconstruction errors are indicative of unusualness.

In practice, the traditional squared error is often used as the reconstruction error of each output:

$$L(x,x') = |x - x'|^2$$

We used this property of an Autoencoder to identify unusual or fraudulent records in our dataset.



Scoring Methodology:

An Autoencoder was trained to use the 10 x 1,048,575 dataset after dimensionality reduction from the PCA method. Then, this Autoencoder was run on all the records of the same dataset.

To find the fraud score using Autoencoder, we again calculated both the Manhattan and Euclidean distance between the input and output records.

Thus, our second fraud score is the Euclidean distance between the inputs and reconstructed outputs of the Autoencoder for each record.

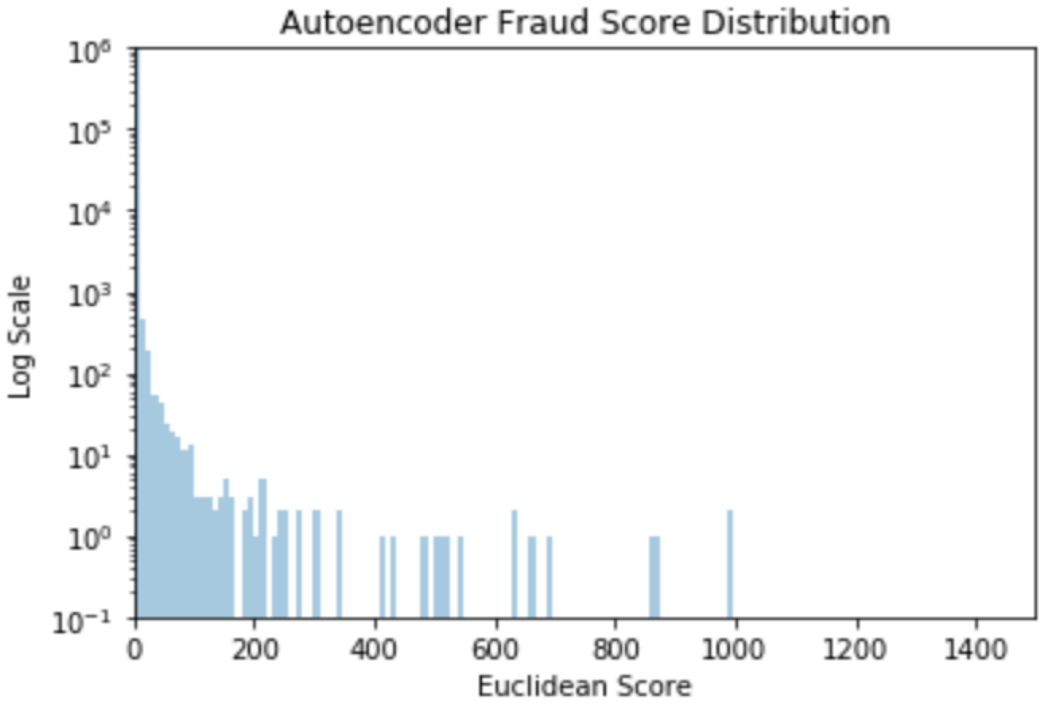
The fraud score for each record is calculated as:

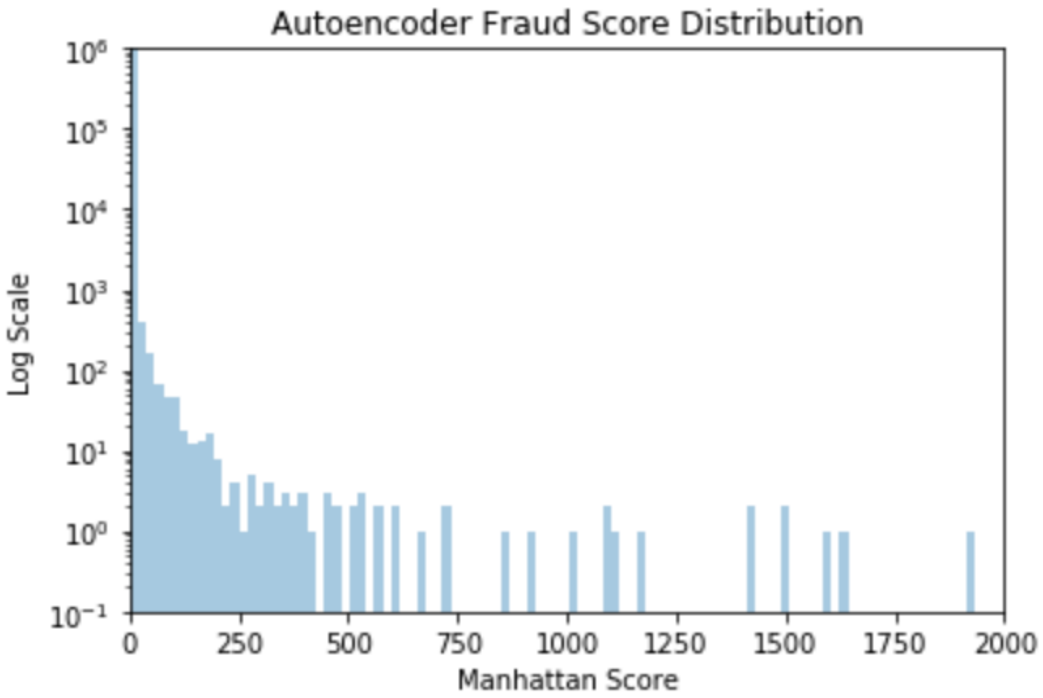
$$S_{ae} = \sum_i (|x_i - x_i'|^n)^{1/n}$$

where,
S_{ae} = Autoencoder Fraud Score
x = Input
x' = Reconstructed output
i ∈ [1,10] for the 10 PCA columns in our 10 x 1,048,575 dataset

When n = 1 then S_{ae} represents the Manhattan Distance measure of Autoencoder Score and when n = 2 then S_{ae} represents the Euclidean Distance measure of Autoencoder Score.

The distribution of the Autoencoder Fraud Scores can be seen in the graphs below.





Both the distributions are right skewed with a long tail and represent the shape of a regular fraud score distribution.

Combining the Fraud Scores: Quantile Binning

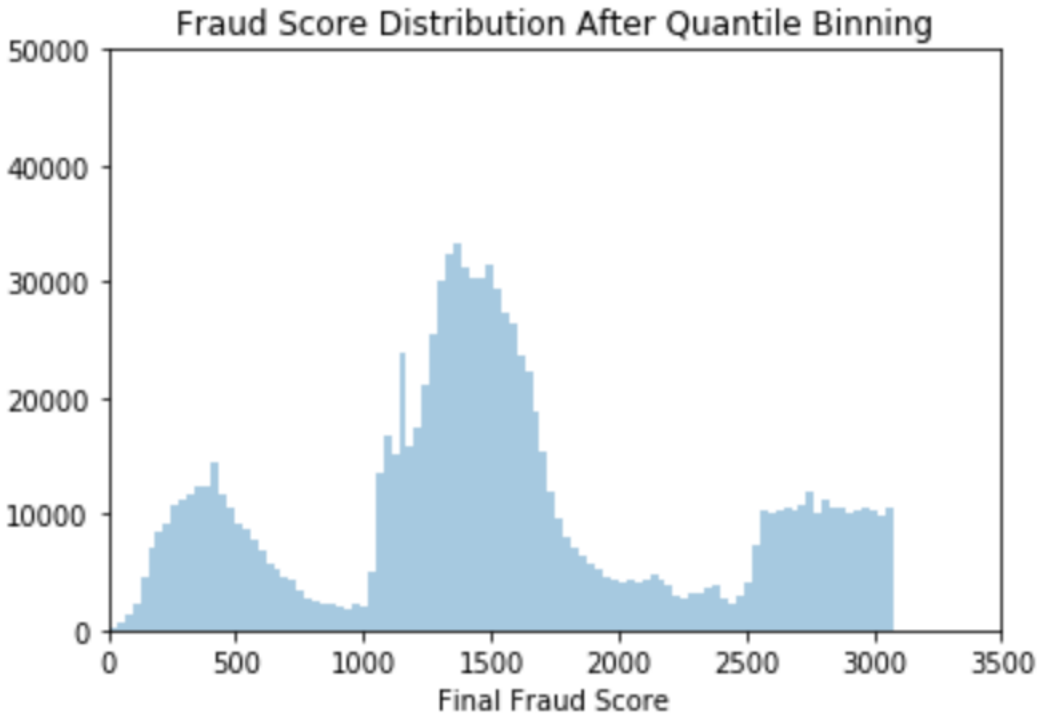
A record predicted as fraud by both our models (Heuristic and Autoencoder) gave us a higher degree of confidence in its anomalous nature.

In order to make these two different results of fraud scores more comparable, we used quantile binning. With 3075 bins and 341 records in each bin, we assigned two scores, each between 1 and 3075, to each record based on its overall rankings in both the fraud algorithms respectively. As a final result, we took the average of the two scores to be our final fraud score.

To look for the top anomalous records, we took our top 341 highest scoring records, and sorted them again by our original Autoencoder scores. We have described the 11 most unusual records with highest fraud scores in the Results section.

Results

Based on our model output, we identified some anomalous property characteristics and valuations among the ~1 million properties that were analyzed. Below is the distribution of the final fraud scores of our records after quantile binning.



Some of the flagged records are as described below.

1. **Record 315453:** This record is found to be unusual because of the high valuation of this property. It is valued at $5.2 billion, for a 1 story building. The Lot Front and the Lot Depth of this property is exceptionally large as well, i.e. 3,030 feet and 5,948 feet respectively. The Building Front and Building Depth is also quite low, i.e. around 380 feet and 240 feet. This property is owned by the New York State Department.

2. **Record 61238**: The Building Front and Building Depth, at 60 feet and 30 feet respectively, are small compared to valuation of $91 million. The Lot Front and Lot Depth, at 735 feet and 103 feet respectively, is also small for this evaluation. The property is owned by Parks and Recreation.

3. Record 970081: This property is valued at $70 million. The Building Front and Building Depth, at 8 feet by 8 feet respectively, is very less compared to the valuation. The Lot Front and Lot Depth of this property is 4,000 feet and 565 feet respectively. The property is owned by Parks and Recreation.

4. Record 623955: This property is valued at $7 million, for a 1 story building. The Building Front and Building Depth at 10 feet by 10 feet respectively, and the Lot Front and Lot Depth at 524 feet by 561 feet respectively, is valued at a very high cost of $7 million. The property is owned by Parks and Recreation.

Note: *The ongoing price for a 1 bed is around $7 million in Park Avenue but the building size is too small in this record*
Source: *https://www.cityrealty.com/nyc/midtown-east/432-park-avenue/54898*

5. Record 977471: This property has a valuation of $3 million for a 20-story building. The property is owned by New York City Economic. It is located at Queens Plaza South. For a 20-story building, the property is under-evaluated.

Note: *On further research, we found that this property actually has 4 stories.*
Source: *https://streeteasy.com/building/28_10-queens-plaza-south-queens*

6. Record 787892: This property has a valuation of $2.15 million, for a 20-story building. The Building Depth and Building Front, at 1 feet by 1 feet respectively, is unusually small for a 20-story building. The Owner field is also missing.

7. Record 24586: This property has a valuation of $3.7 million for a 10-story building. The Building Depth and Building Front, at 1 feet by 1 feet respectively is unusually small for a 10-story building. The Lot Front and Lot Depth is at 94 feet and 165 feet respectively. The street address is 11-01 43 Avenue and the name of the Owner is 11-01 43RD Avenue Rea.

8. Record 5393: This property has an unusually high valuation of $3 million where the Building Front and the Building Depth is at 1 feet by 1 feet and the

Lot Front and Lot Depth is 157 feet and 95 feet respectively. The property is owned by 864163 Realty, LLC.

9. Record 376243: This property has a Full Valuation of $3.8 billion which is quite low compared to the Assessed Valuation of $18 billion. The property has a Lot Front and Lot Depth of 4,910 feet and 133 feet respectively. The Building Front and Building Depth is at 65 feet and 88 feet. The property is owned by Logan Property, INC.

10. Record 78804: This property is valuated at $4.3 billion for a Lot Front of 117 feet and Lot Depth of 108 feet. The Building Front and Building Depth is also very small for the valuation, at 64 feet and 88 feet respectively. This building contains 5.4 stories and is owned by US Government Owned.

11. Record 518829: This 1-story property is valued at $1 million, though its Assessed Value is given as $101. The Lot Front and Lot Depth is 60 feet and 787 feet respectively. The Building Front and Depth is at 134 feet and 680 feet. The property is owned by Port of New York Auth.

Conclusion

Our model is successful in identifying records with unusual characteristics. Based on the results, our algorithm flagged the records that have exceptionally large or small property values. It also flagged properties on the basis of tax-class, as it was indicative of a direct relationship to valuation and size of the property. Even though some additional properties were flagged, they are not necessarily indicative of fraud.

On further consultation with industry experts, we can bring in more improvement to our model, that will reduce the likelihoods of false positives. This will include an improvement on expert variables and training on a larger data set to understand the behavior of fraudulent instances.

Appendix

A: PROPERTY ASSESSMENTS WITH HIGH FRAUD SCORES

RECORD	BLDGCL	TAXCLASS	BORO	FULLVAL_1_	ZIP3	ZIP	STADDR	BBLE	BLOCK
5393	D9	2	4	172.9846	113	11373	86-55 BROADWAY	4018420001	1842
24586	H9	4	4	172.9846	111	11101	11-01 43 AVENUE	4004590005	459
787892	O3	4	4	172.9846			28 STREET	4004200101	420
977471	O3	4	4	172.9846	111	11101	28-10 QUEENS PLAZA SOUTH	4004200001	420
61238	Q1	4	2	141.9504			BROADWAY	2059000150	5900
970081	Q1	4	1	357.9754			JOE DIMAGGIO HIGHWAY	1012540010	1254
623955	Q1	4	3	244.3999			PARK AVENUE	3020250001	2025
315453	T1	4	4	172.9846			GRAND CENTRAL PKWY	4009260001	926
376243	T1	4	4	172.9846	114	11422	154-68 BROOKVILLE BOULEVARD	4142600001	14260
78804	V9	4	3	244.3999			FLATBUSH AVENUE	3085900700	8590
518829	Z9	4	3	244.3999	112	11201	EAST RIVER	3001990126P	199

RECORD	LOT	EASEMENT	OWNER	LTFRONT	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT
5393	1		864163 REALTY, LLC	157	95	1	29,30,000	13,18,500	13,18,500
24586	5		11-01 43RD AVENUE REA	94	165	10	37,12,000	2,52,000	16,70,400
787892	101			139	342	20	21,51,600	9,68,220	9,68,220
977471	1		NEW YORK CITY ECONOMI	298	402	20	34,43,400	15,49,530	15,49,530
61238	150		PARKS AND RECREATION	735	103	1	9,15,00,000	3,84,30,000	4,11,75,000
970081	10		PARKS AND RECREATION	4000	150	1	7,02,14,000	3,14,55,000	3,15,96,300
623955	1		PARKS AND RECREATION	524	561	1	75,14,000	32,13,000	33,81,300
315453	1		NEW YORK STATE DEPART	3030	5948	1	5,27,90,00,000	34,38,00,000	2,37,55,50,000
376243	1		LOGAN PROPERTY, INC.	4910	132.6583	3	37,40,19,883	1,79,28,08,947	4,66,83,08,947
78804	700		U S GOVERNMENT OWNRD	117	108	5.474805	4,32,63,03,700	1,94,68,36,665	1,94,68,36,665
518829	126	P	PORT OF NEW YORK AUTH	60	787	1	10,70,225	101	4,81,601

B: FILE DESCRIPTION

New York data is a Property Valuation and Assessment Data. It provides descriptive information, market value, assessed value and other miscellaneous information about the properties in New York.

- File Name: NY Property Data.xlsx
- Source: *https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and- Assessment-Data/rgy2-tti8*

- Client: New York City Open Data (Property Valuation and Assessment Data)
- Number of records: 1,048,575
- Number of Fields: 30 fields (6 continuous, 13 categorical)
- Time Frame: 2010-2011
- Number of records: 1,048,575

Summary Statistics for Numerical Variables

Below is the tabular version of the 5 number summary, i.e. minimum, first quartile, median, third quartile, and maximum for the variables present in the data set.

The table also contains the standard deviation, number of unique values, and the percentage of populated records for each variable.

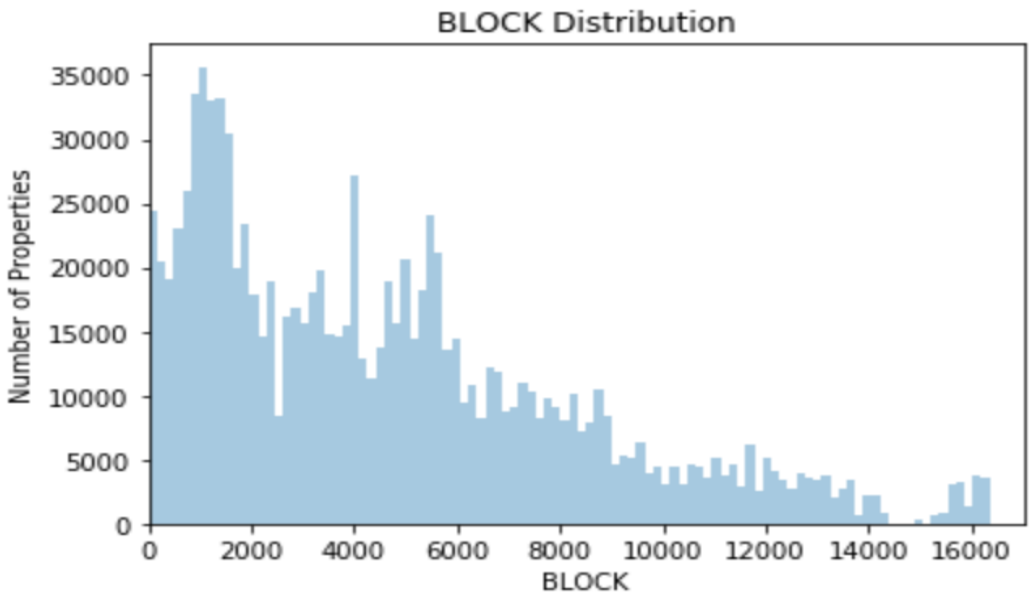
COLUMN:	COUNT	MEAN	STD	MIN	25%	50%	75%	MAX	PERCENT POPULATED	UNIQUE VALUES
RECORD	1048575	524288	302697.7	1	262144.5	524288	786431.5	1048575	100%	1048575
BLOCK	1048575	4708.8867	3699.547	1	1534	3944	6797	16350	100%	13949
LOT	1048575	370.09	860.53	1	23	49	146	9978	100%	6366
LTFRONT	1048575	36.17	73.73	0	19	25	40	9999	100%	1277
LTDEPTH	1048575	88.27	75.47	0	80	100	100	9999	100%	1336
STORIES	996433	5.06	8.43	1	2	2	3	119	95.02%	112
FULLVAL	1048575	880487.7	11702930	0	303000	446000	619000	6.15E+09	100%	108277
AVLAND	1048575	85995.03	4100755	0	9160	13646	19706	2.669E+09	100%	70529
AVTOT	1048575	230758.2	6951206	0	18385	25339	46095	4.668E+09	100%	112294
EXLAND	1048575	36811.79	4024330	0	0	1620	1620	2.669E+09	100%	33186
EXTOT	1048575	92543.81	6578281	0	0	1620	2090	4.668E+09	100%	63805
EXCD1	622642	1604.5	1388.13	1010	1017	1017	1017	7170	59.38%	130
ZIP	1022219	10934.3	526.57	10001	10453	11215	11364	33803	97.48%	197
BLDFRONT	1048575	23.02	35.79	0	15	20	24	7575	100%	610
BLDDEPTH	1048575	40.07	43.04	0	26	39	51	9393	100%	620
AVLAND2	280966	246365.5	6199390	3	5705	20059	62338.75	2.371E+09	26.80%	58170
AVTOT2	280972	716078.7	11690170	3	34013.5	80010	240792	4.501E+09	26.80%	110891
EXLAND2	86675	351802.2	10852480	1	2090	3053	31419	2.371E+09	8.27%	21997
EXTOT2	129933	658114.8	16129810	7	2889	37116	106629	4.501E+09	12.39%	48107
EXCD2	90941	1371.66	1105.49	1011	1017	1017	1017	7160	8.67%	61

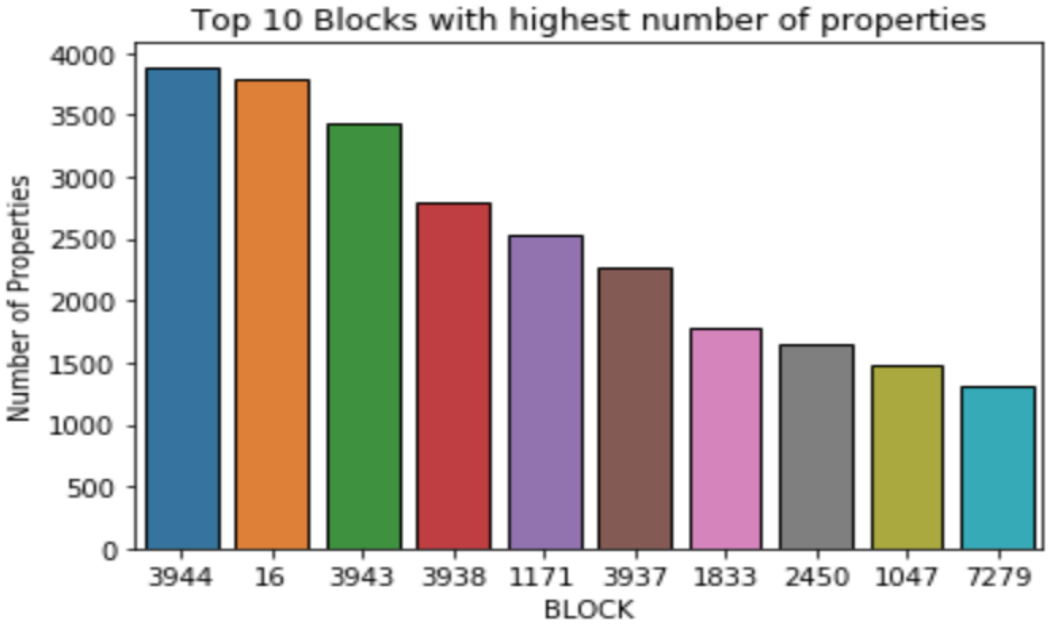
Fields (Both Numerical and Categorical Variables) :

1) RECORD
   - **Description:** Discrete numeric variable for unique identification of records
   - **Number of Blank records:** 0
   - **Number of Unique Values:** 1,048,575
   - **Percentage populated:** 100%

2) BBLE
   - **Description:** Discrete numeric variable for unique identification of properties, formed on concatenation of BORO code, BLOCK and LOT
   - **Number of Blank records:** 0
   - **Number of Unique Values:** 1,048,575
   - **Percentage populated:** 100%

3) BLOCK
   - **Description:** Discrete numeric variable of length 5
   - **Valid Block Ranges by BORO:**
     . MANHATTAN 1 TO 2,255
     . BRONX 2,260 TO 5,958
     . BROOKLYN 1 TO 8,955
     . QUEENS 1 TO 16,350
     . STATEN ISLAND 1 TO 8,050
   - **Number of Blank records:** 0
   - **Number of Unique Values:** 13,949
   - **Percentage populated:** 100%
   - **Minimum:** 1
   - **Maximum:** 16,350
   - **Mode:** 3,944 with 3,888 records
   - **BLOCK Distribution by Number of Properties:**

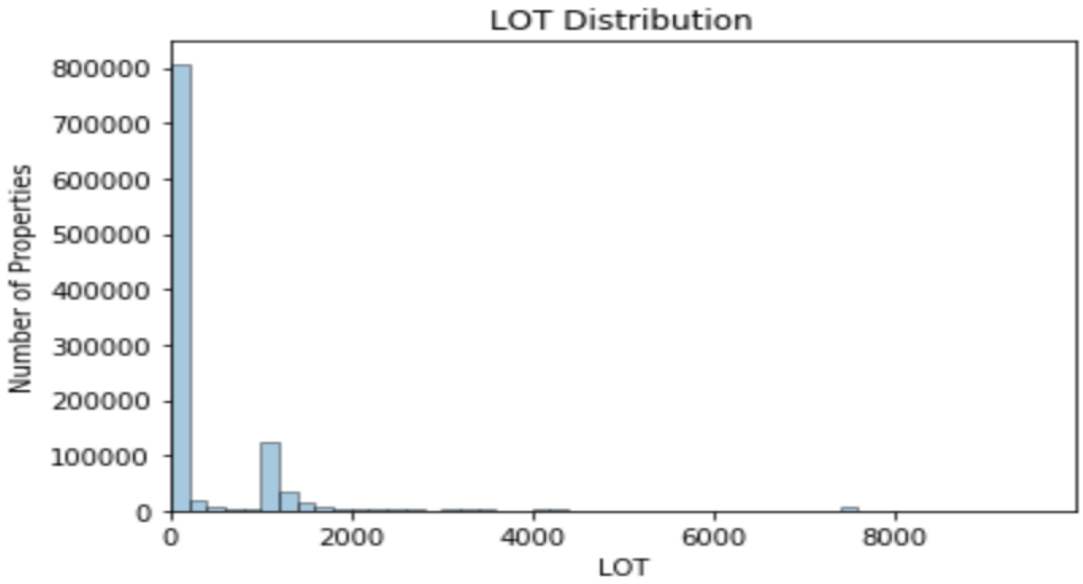


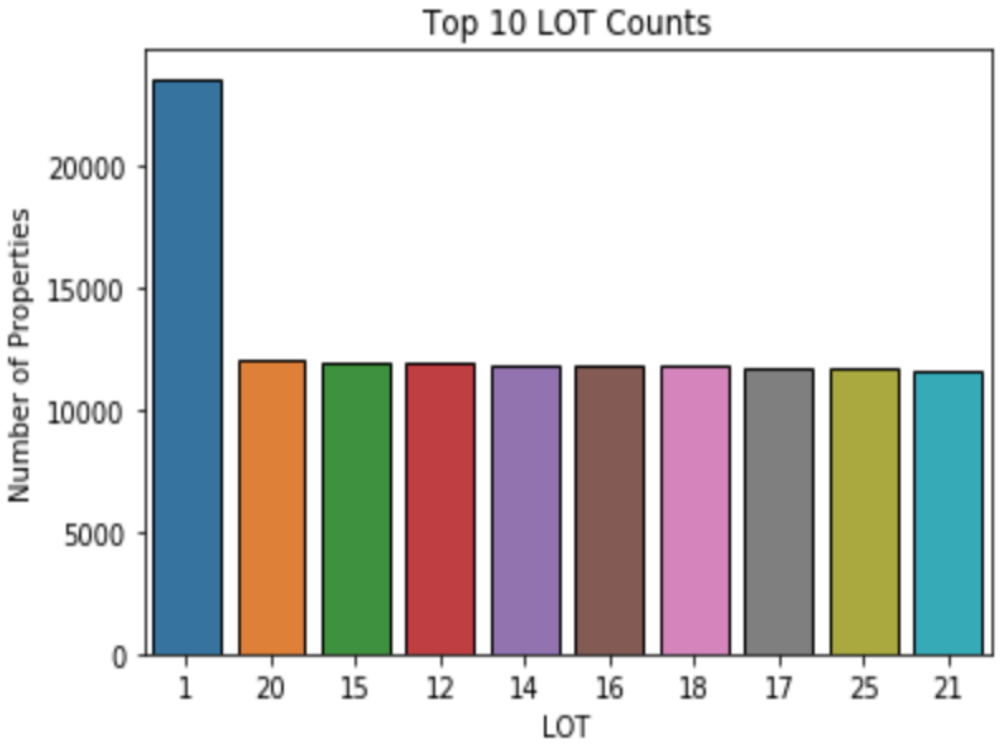


4) LOT

- **Description:** Discrete numeric variable of length 4. Unique within BORO/BLOCK.
- **Number of Blank records:** 0
- **Number of Unique Values:** 6,366
- **Percentage populated:** 100%
- **Minimum:** 1
- **Maximum:** 9,978

- **Mode:** 1 with 23,570 records
- **LOT Distribution by Number of Properties:**



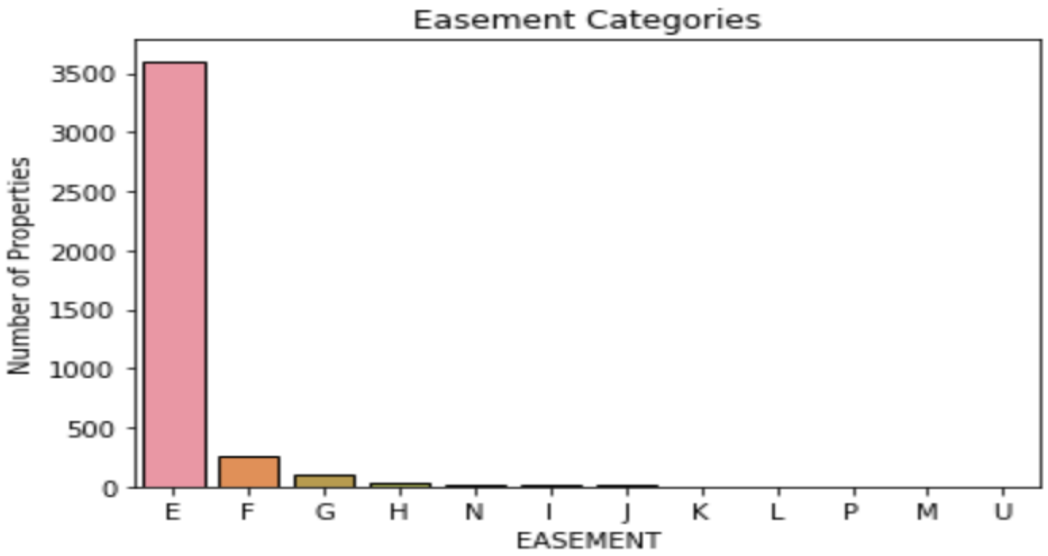


5) EASEMENT

- **Description:** Categorical variable of length 1 alpha.
  - . SPACE Indicates the lot has no Easement
  - . 'A' Indicates the portion of the Lot that has an Air

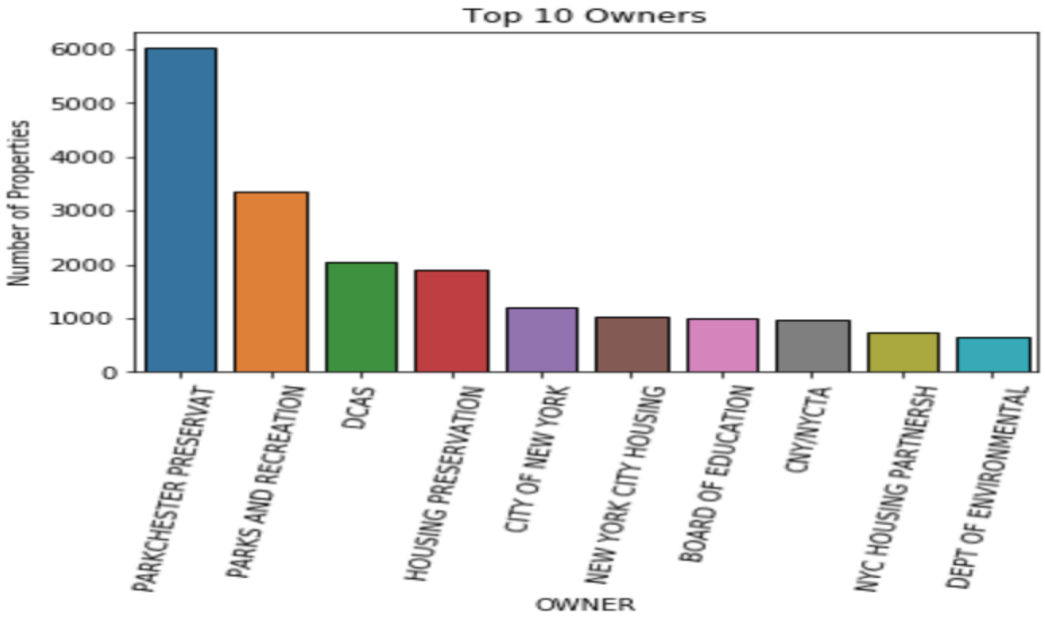
- 'B' Indicates Non-Air Rights
- 'E' Indicates the portion of the lot that has a Land
- 'F' through 'M' are duplicates of 'E'.
- 'N' indicates Non-Transit
- 'P' indicates Piers
- 'R' Indicates Railroads
- 'S' Indicates Street
- 'U' Indicates U.S. Government

- **Number of Blank records:** 1,044,532
- **Number of Unique Values:** 13
- **Mode:** E with 3,603 records
- **Percentage populated:** 38.56%
- **EASEMENT Distribution by Number of Properties:**



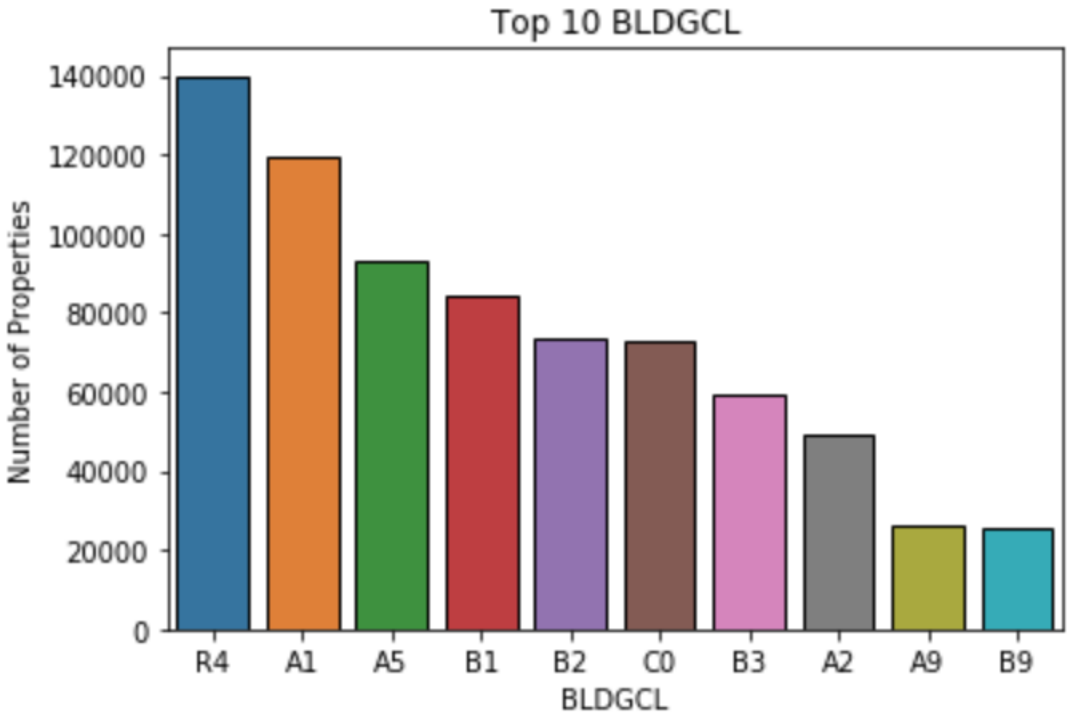
6) OWNER

- **Description:** Categorical Variable of length 21 Characters providing the owner's name
- **Number of Blank records:** 31,081
- **Number of Unique Values:** 847,054
- **Mode:** PARKCHESTER PRESERVAT with 6,021 records
- **Percentage populated:** 97.04%
- **OWNER Distribution by Number of Properties:**
  Some owners own a huge number of properties. Most of them are government agencies but some are private players



7) BLDGCL

- **Description:** Categorical Variable of length 2 Characters. Building Class Position 1 = ALPHA & Position 2 = NUMERIC. There is a direct correlation between the Building Class and the Tax Class
- **Number of Blank records:** 0
- **Number of Unique Values:** 200
- **Mode:** R4 with 139,879 records
- **Percentage populated:** 100
- **BLDGCL Distribution by Number of Properties:**



8) TAXCLASS

- **Description:** Categorical Variable of length 2 Characters. Current Property Tax Class Code (NYS Classification) valid values are:
  - TAX CLASS 1 = 1-3 UNIT RESIDENCES
  - TAX CLASS 1A = 1-3; STORY CONDOMINIUMS ORIGINALLY A CONDO
  - TAX CLASS 1B = RESIDENTIAL VACANT LAND
  - TAX CLASS 1C = 1-3 UNIT CONDOMINUMS ORIGINALLY
  - TAX CLASS 1D = SELECT BUNGALOW COLONIES

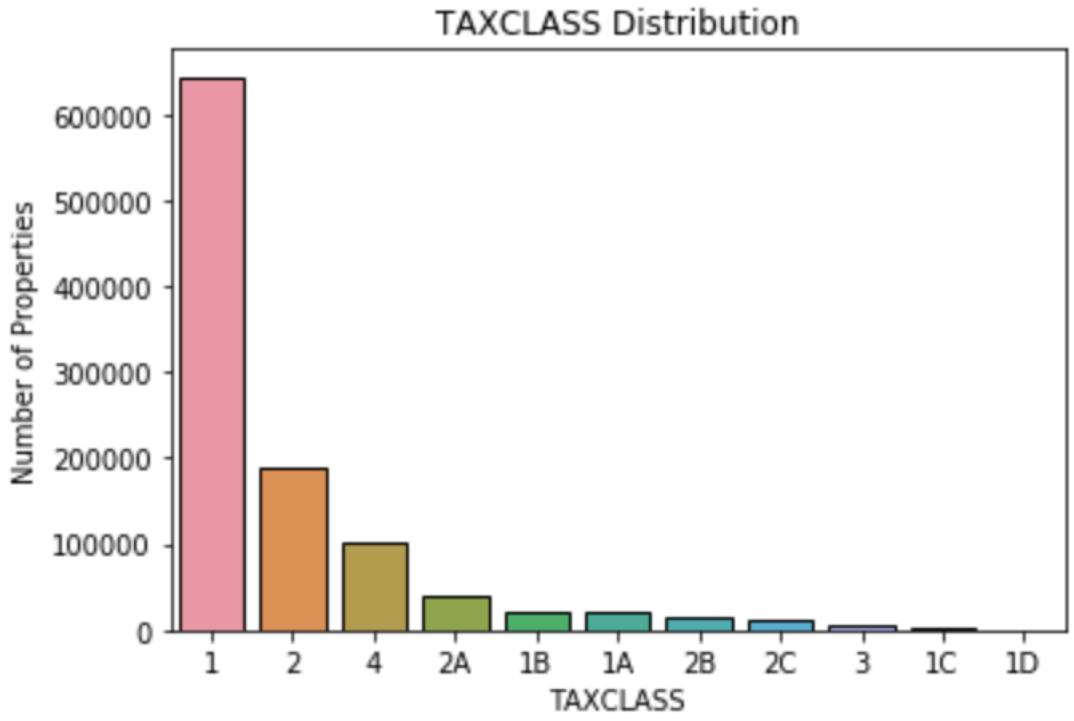
  - TAX CLASS 2 = APARTMENTS
  - TAX CLASS 2A = APARTMENTS WITH 4-6 UNITS
  - TAX CLASS 2B = APARTMENTS WITH 7-10 UNITS
  - TAX CLASS 2C = COOPS/CONDOS WITH 2-10 UNITS
  - TAX CLASS 3 = UTILITIES (EXCEPT CEILING RR)
  - TAX CLASS 4A = UTILITIES - CEILING RAILROADS
  - TAX CLASS 4 = ALL OTHERS

  NOTE - There is a direct correlation between the Building Class and the 1st position of the Tax Class. If the Building Class is known the Tax Class can be generated.

  TAX CLASS BLDG-

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

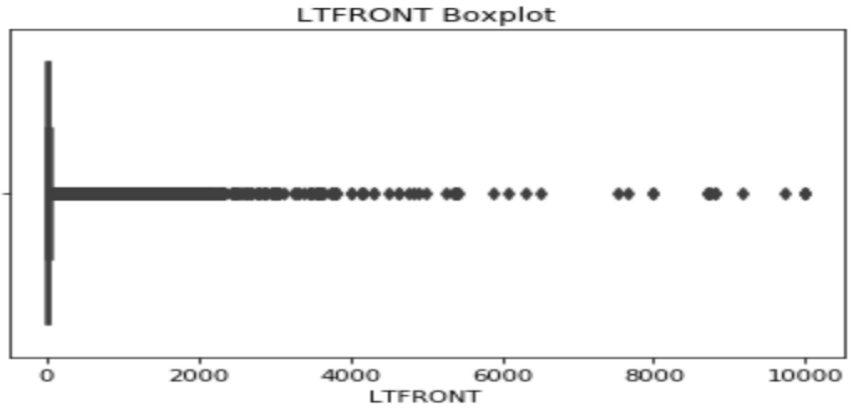
- **Number of Blank records:** 0
- **Number of Unique Values:** 11
- **Percentage populated:** 100%
- **Mode:** 1 with 643,774 records
- **TAXCLASS Distribution by Number of Properties:**



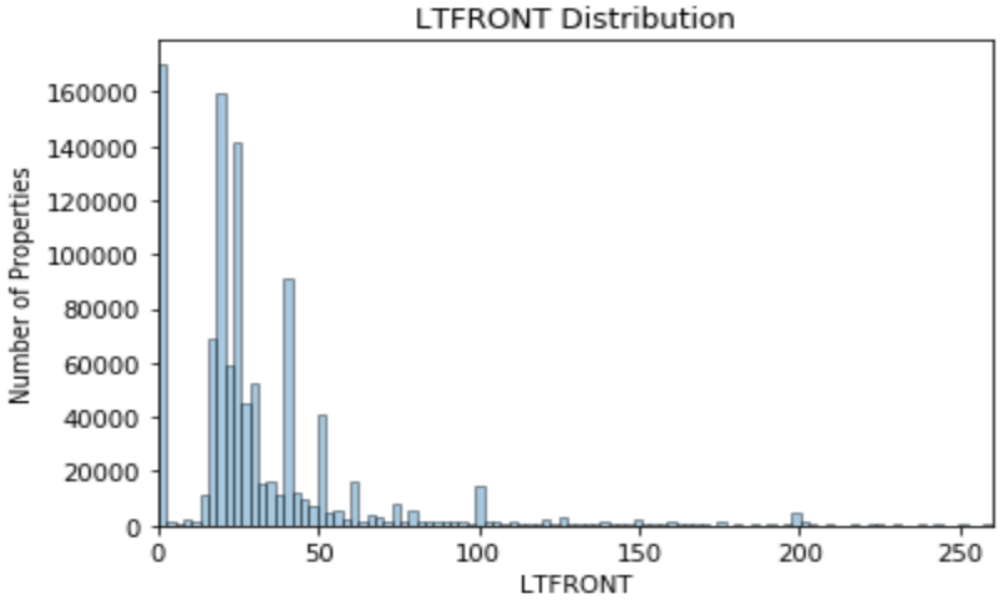
9) LTFRONT

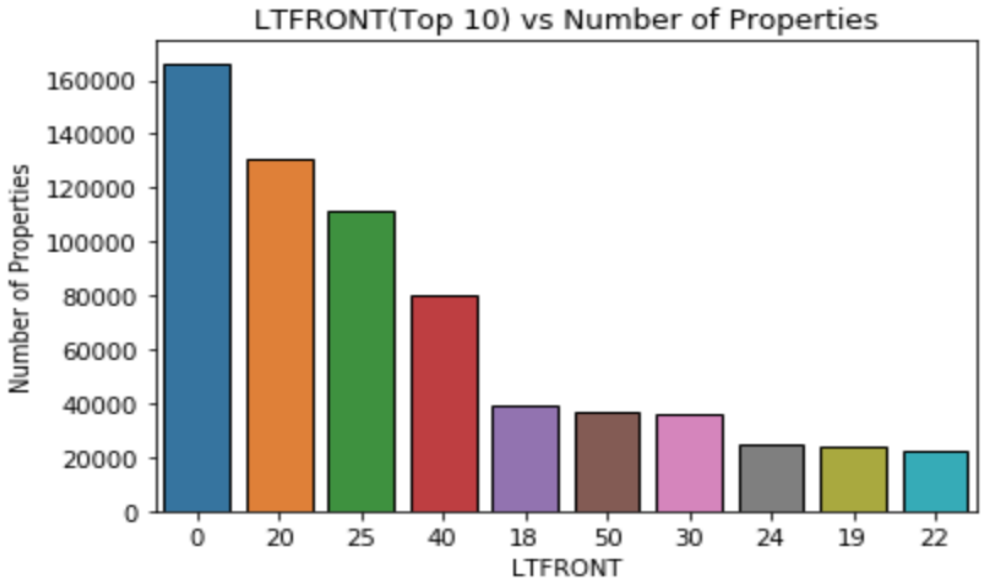
- **Description:** Continuous numeric variable
- **Number of Blank records:** 0
- **Percentage populated:** 100%
- **Minimum:** 0
- **Maximum:** 9999
- **Mean:** 36.17
- **Median:** 25
- **Mode:** 0 with 168,867 records
- **Standard Deviation:** 73.73

- Boxplot (to check for outliers):



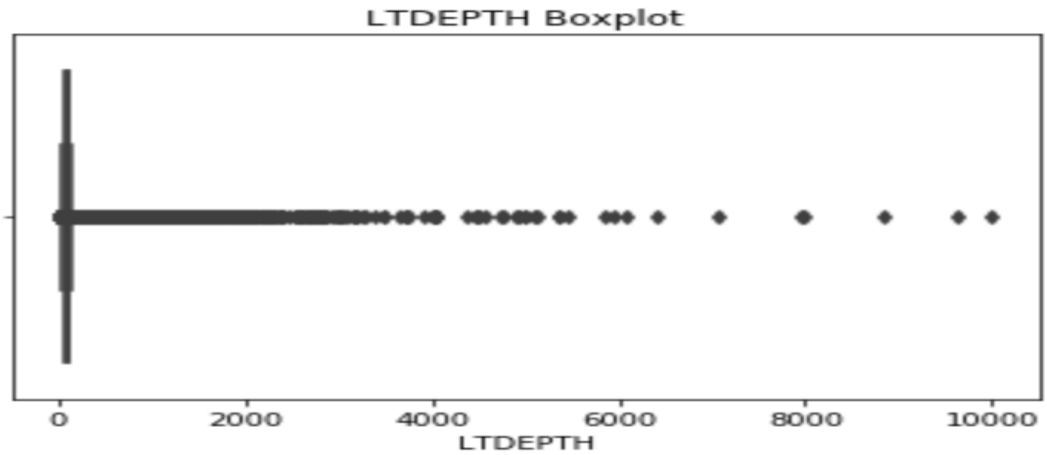
- LTFRONT Distribution by Number of Properties:



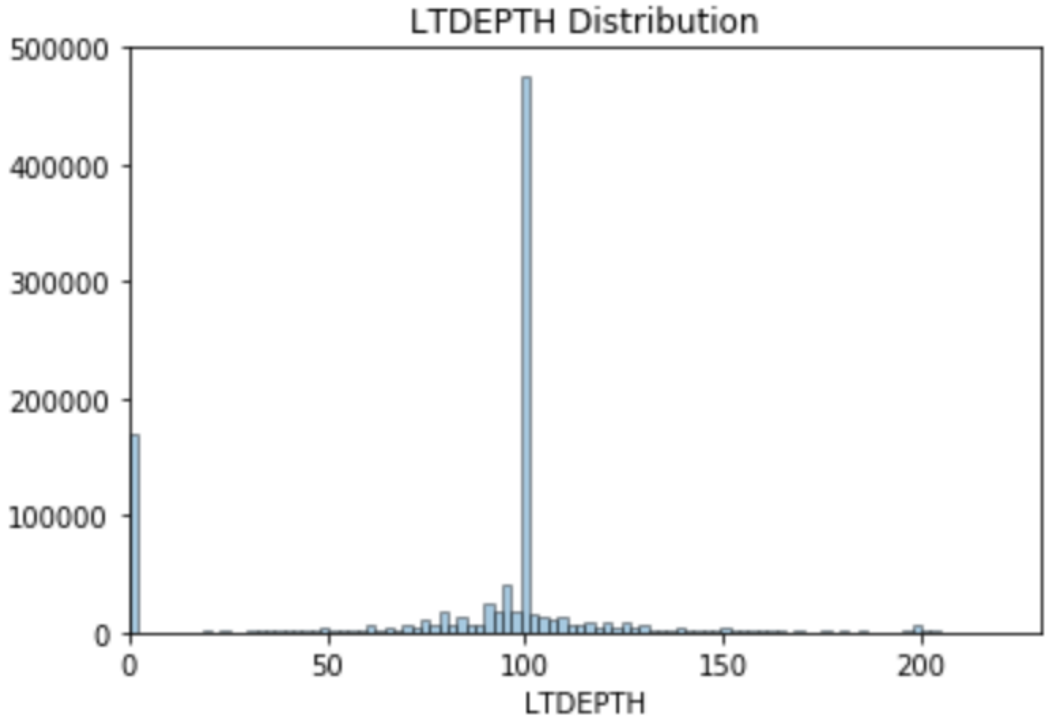


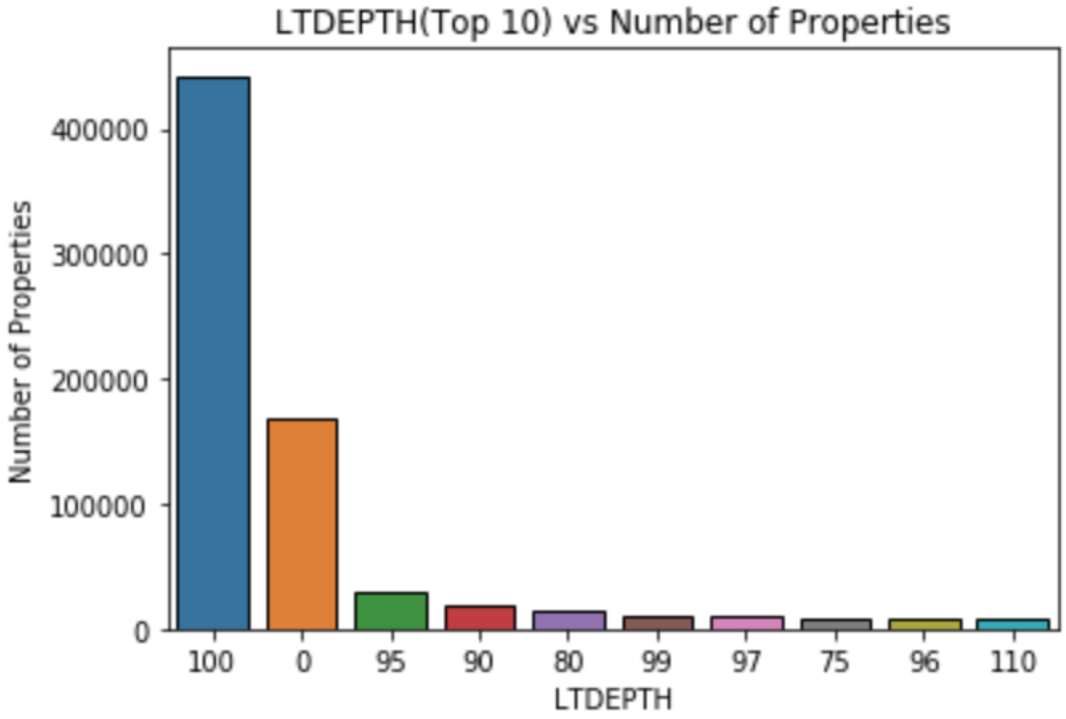
10) LTDEPTH

- **Description:** Continuous numeric variable of length 7 for the LOT Depth in feet.
- **Number of Blank records:** 0
- **Minimum:** 0
- **Percentage populated:** 100%
- **Maximum:** 9,999
- **Mean:** 88.27
- **Median:** 100
- **Mode:** 100 with 457,583 records
- **Standard Deviation:** 75.47
- **Boxplot (to check for outliers):**



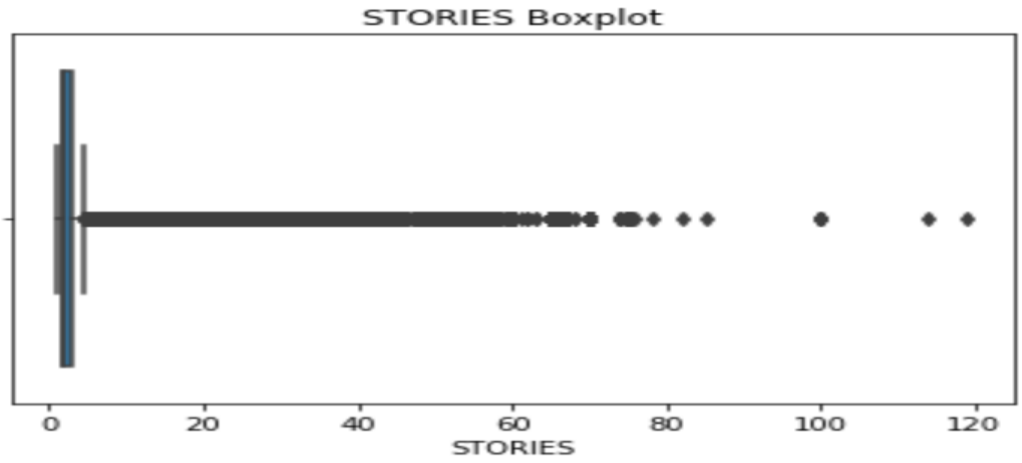
- LTDEPTH Distribution by Number of Properties:



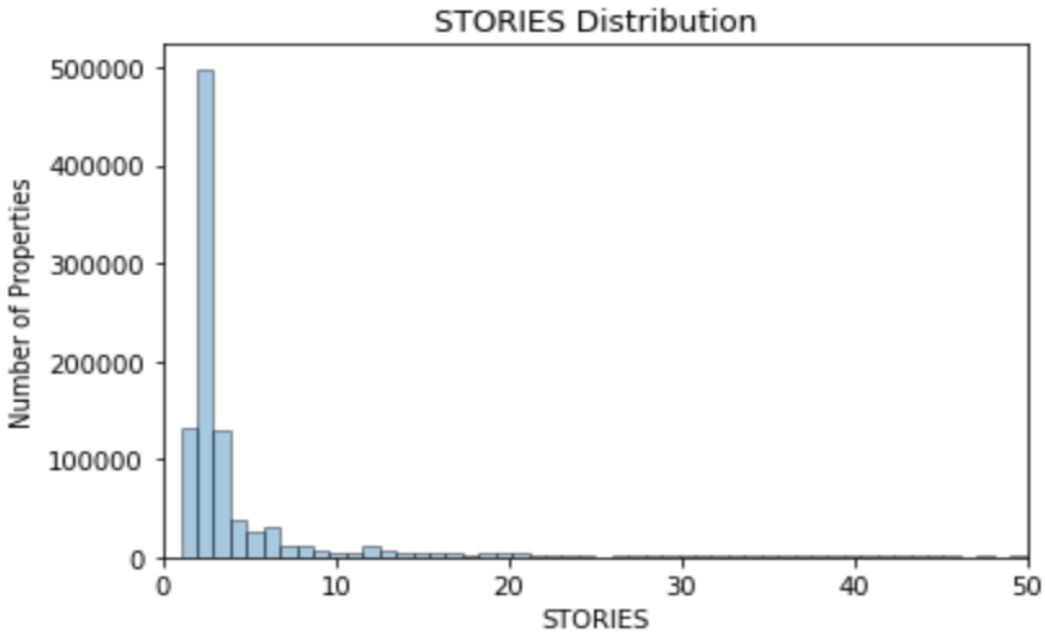


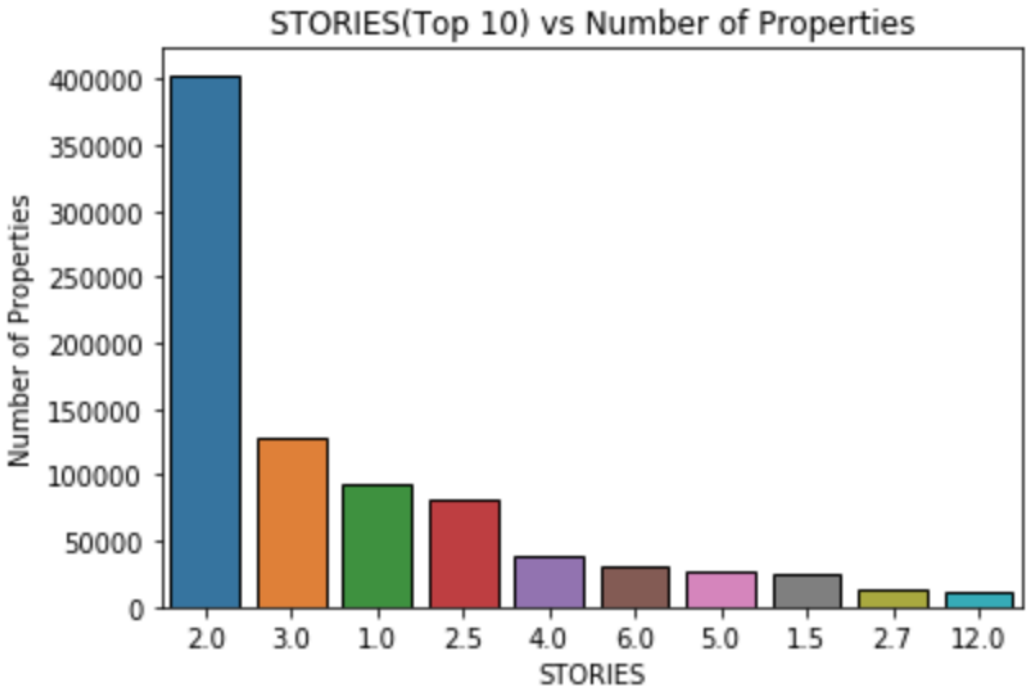
11) STORIES

- **Description:** Discrete Numeric variable of length 6 (999.99) describing the number of stories for the building
- **Number of Blank records:** 52,142
- **Number of Unique Values:** 112
- **Percentage populated:** 95.03%
- **Mode:** 2 with 403,318 records
- **Standard Deviation:** 8.43
- **Boxplot (to check for outliers):**



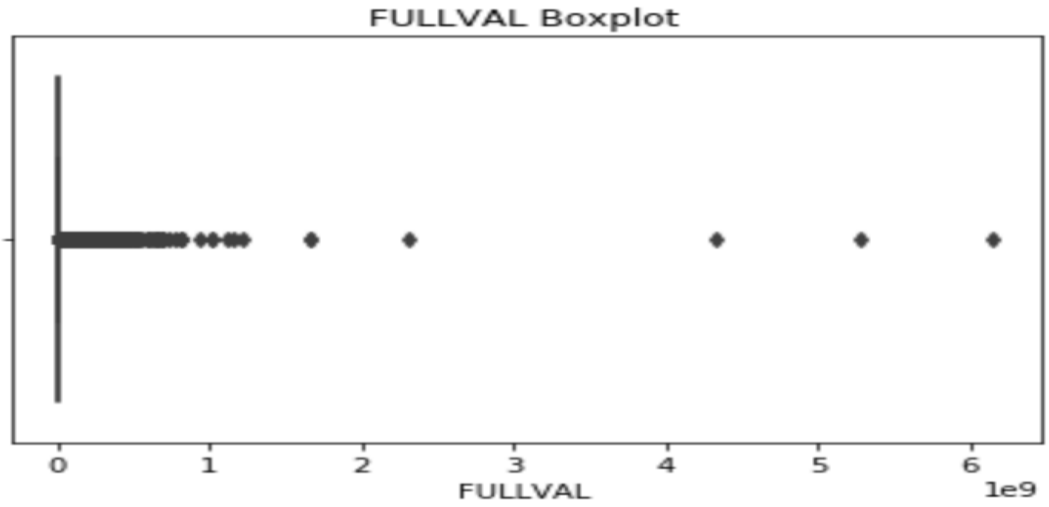
- **STORIES Distribution by Number of Properties:**



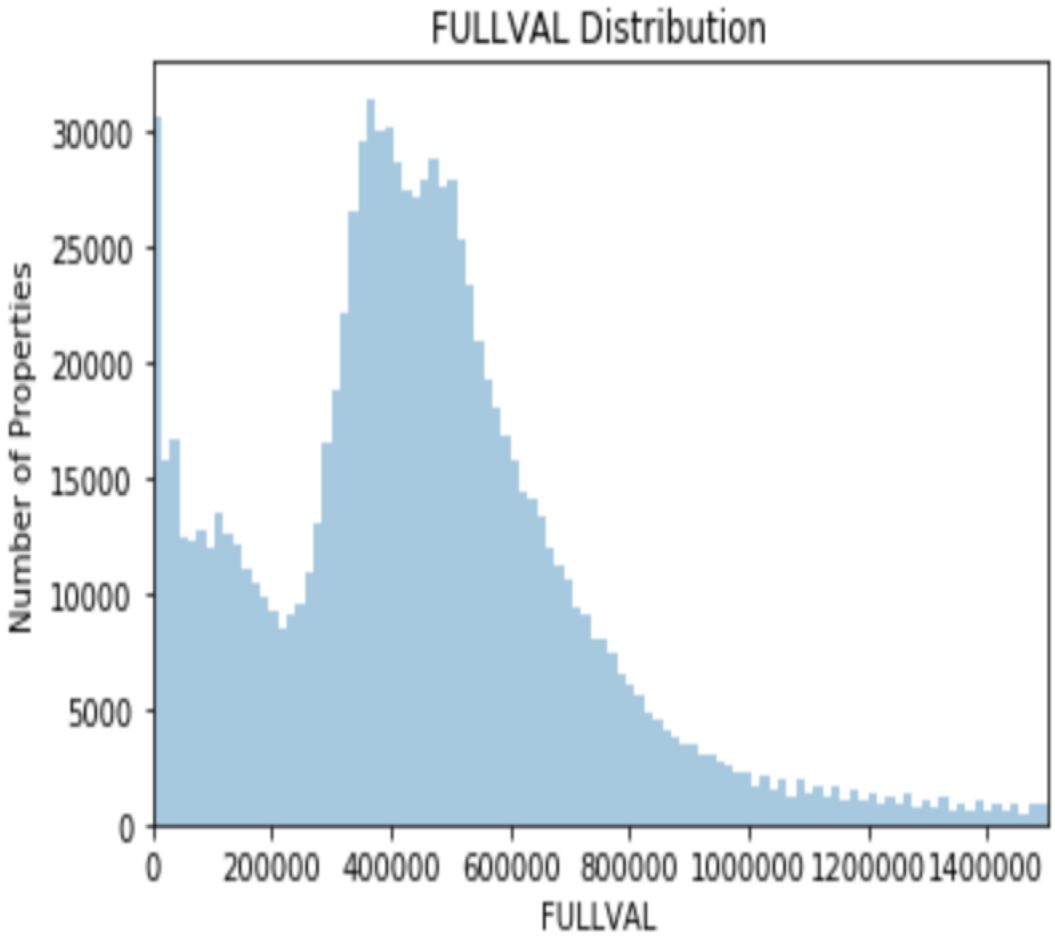


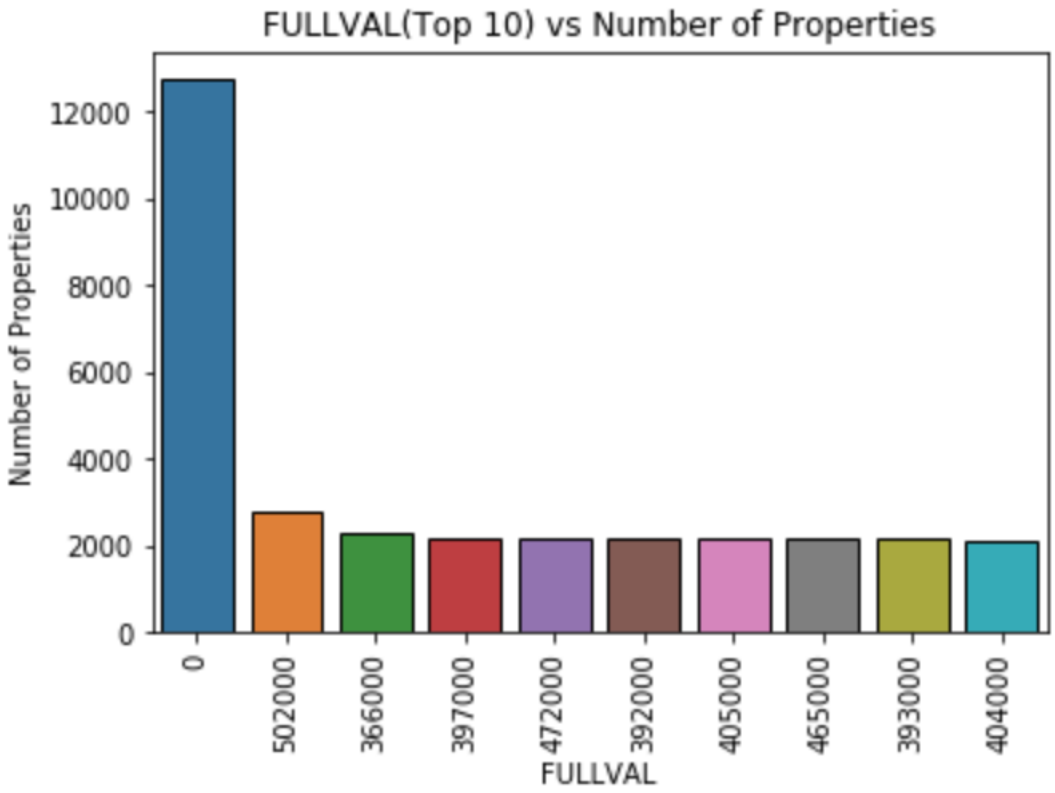
12) FULLVAL

- **Description:** Continuous numeric variable of Length 11 (no decimals). If not zero, it is the current year's total market value of the land
- **Number of Blank records:** 0
- **Percentage populated:** 100%
- **Percentage of zeroes:** 1.22%
- **Minimum:** 0
- **Maximum:** 6,150,000,000
- **Mean:** 880,487.66
- **Median:** 446,000
- **Mode:** 0 with 12,762 records
- **Standard Deviation:** 11,702,927
- **Boxplot (to check for outliers):**



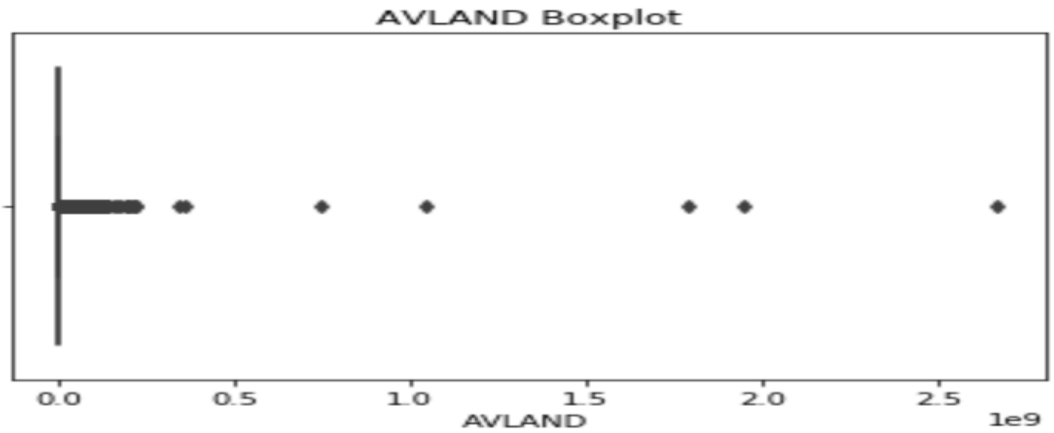
- FULLVAL Distribution by Number of Properties:



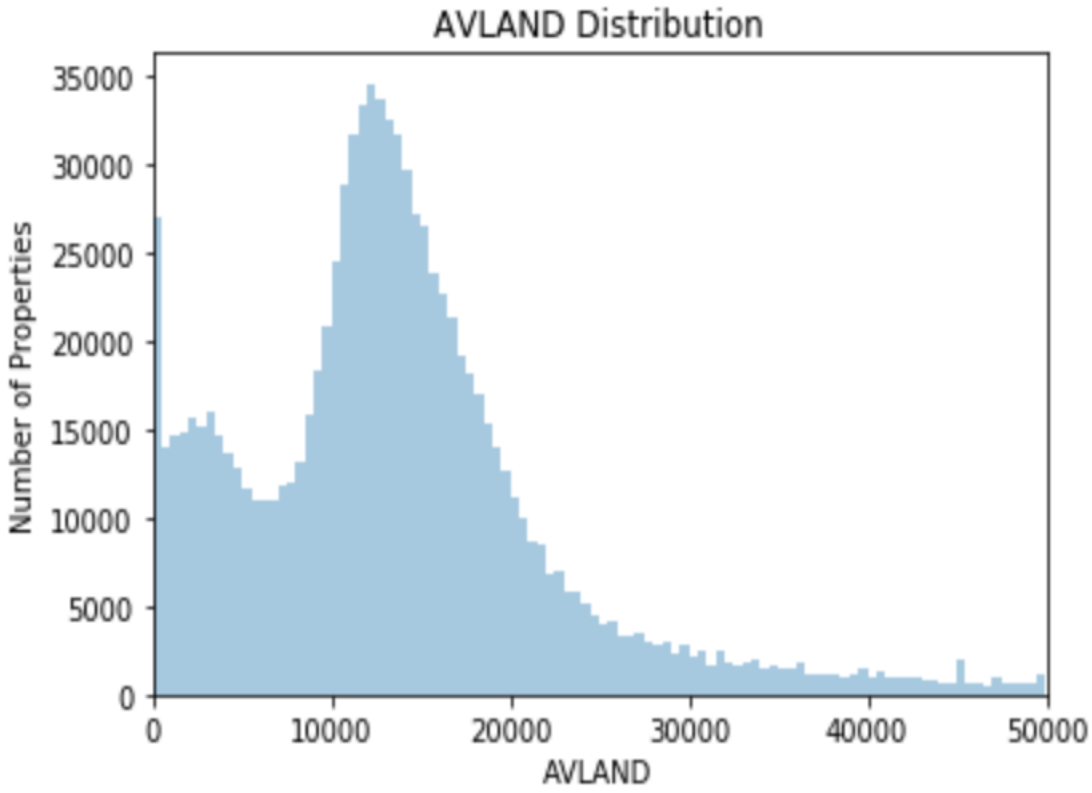


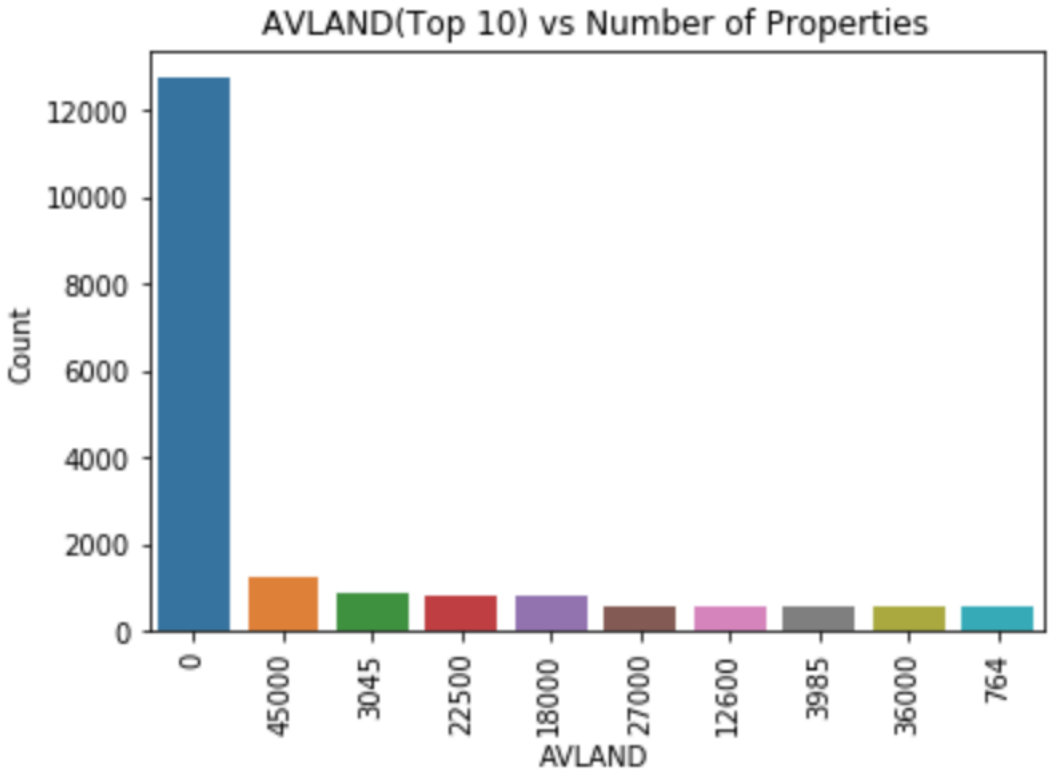
13) AVLAND

- **Description:** Continuous numeric variable of length 9 (no decimals). If not zero, it is the total land area.
- **Number of Blank records:** 0
- **Minimum:** 0
- **Percentage populated:** 100%
- **Percentage of zeroes:** 1.22%
- **Maximum:** 2,668,500,000
- **Mean:** 85,995.03
- **Median:** 13,646
- **Mode:** 0 with 12,764 records
- **Standard Deviation:** 4,100,755.29
- **Boxplot (to check for outliers):**



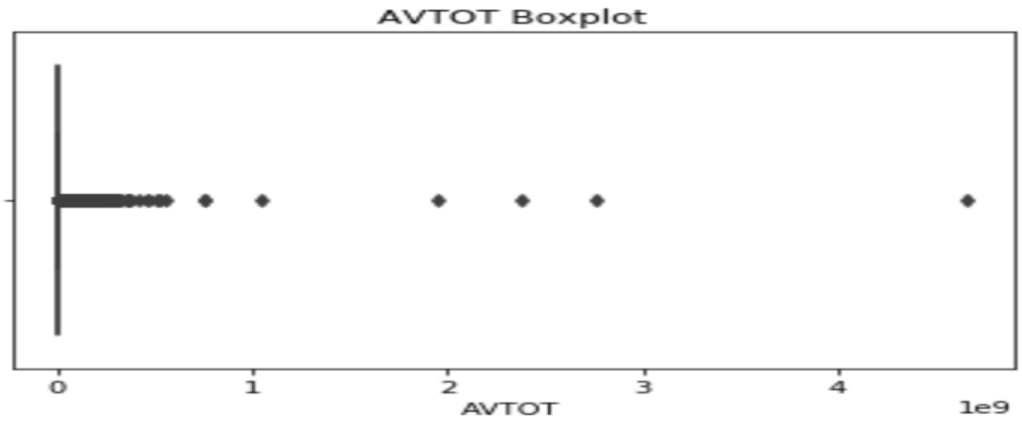
- AVLAND Distribution by Number of Properties:



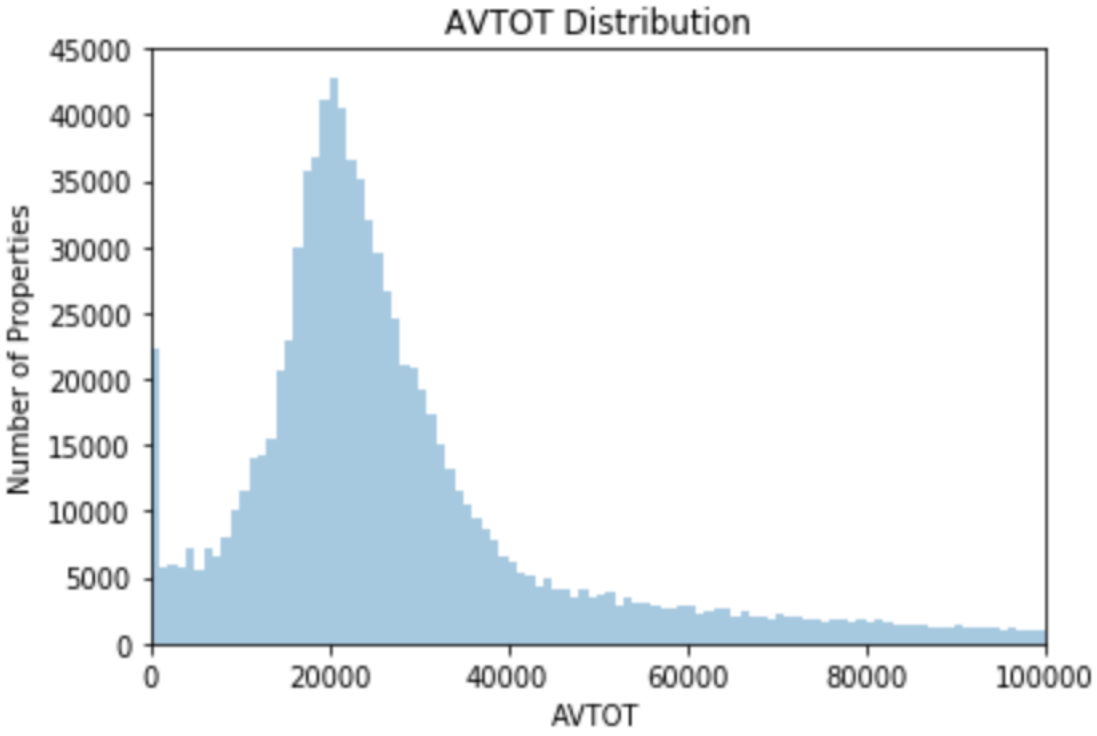


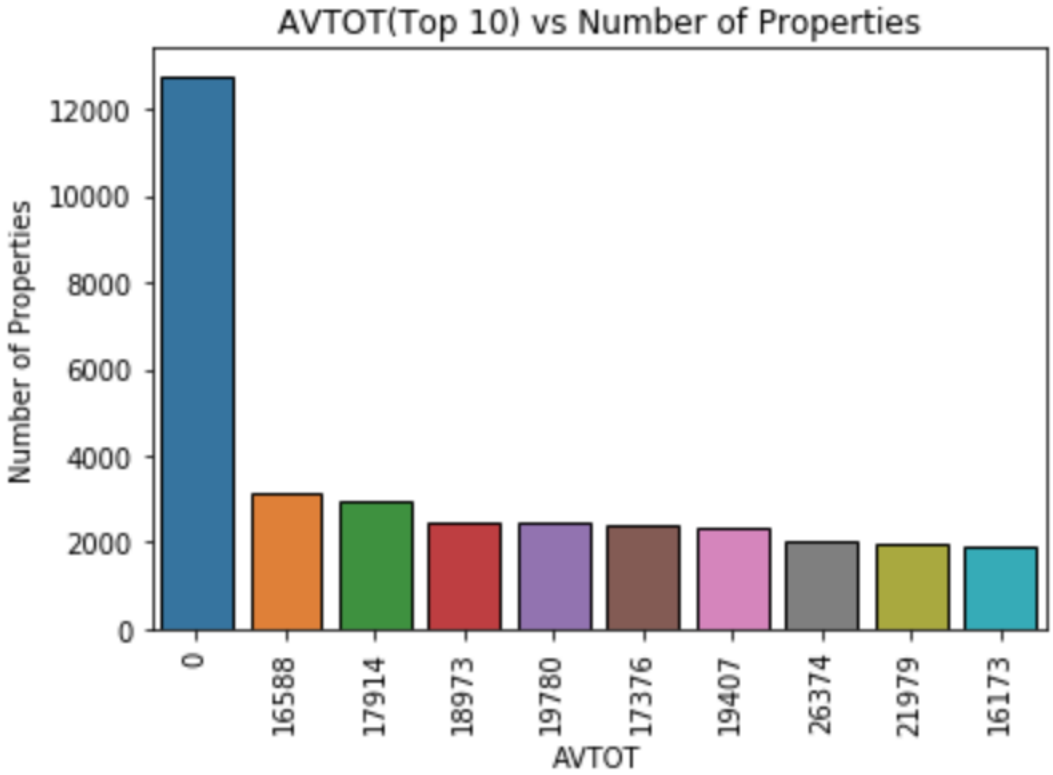
14) AVTOT

- **Description:** Continuous numeric variable of length 5 (no decimals) describing the total number of units in the building
- **Number of Blank records:** 0
- **Minimum:** 0
- **Percentage populated:** 100%
- **Percentage of zeroes:** 1.22%
- **Maximum:** 4,668,308,947
- **Mean:** 230,758.18
- **Median:** 25,339
- **Mode:** 0 with 12,762 records
- **Standard Deviation:** 6,951,205.99
- **Boxplot (to check for outliers):**



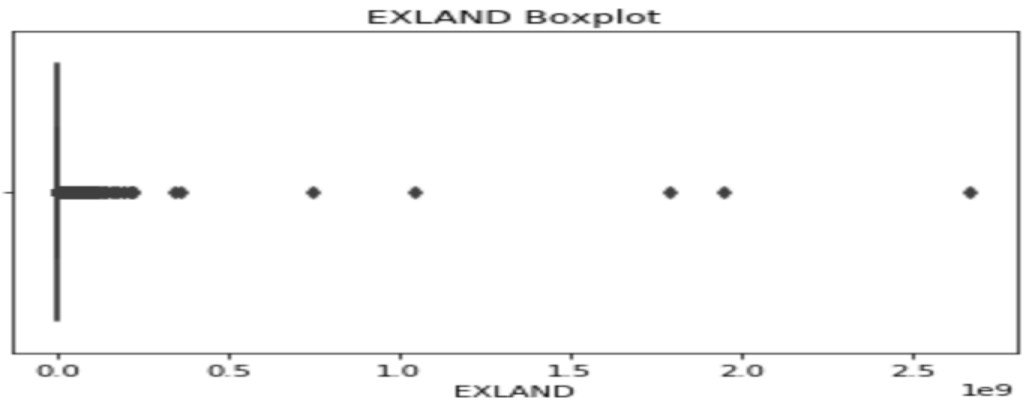
- AVTOT Distribution by Number of Properties:



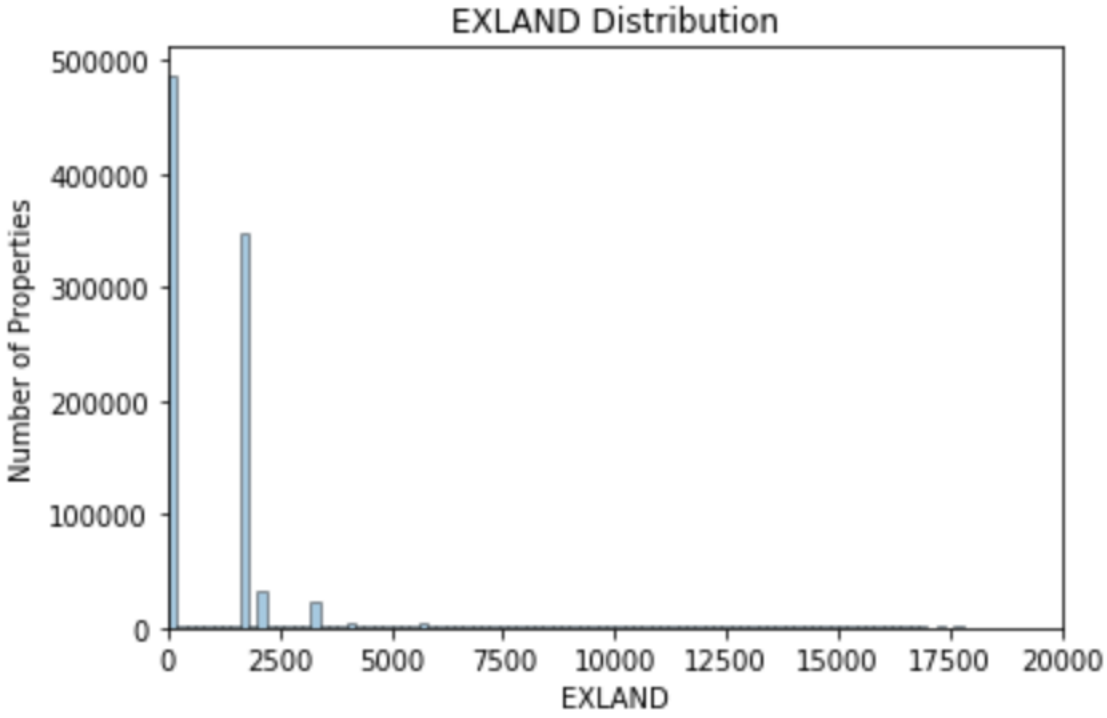


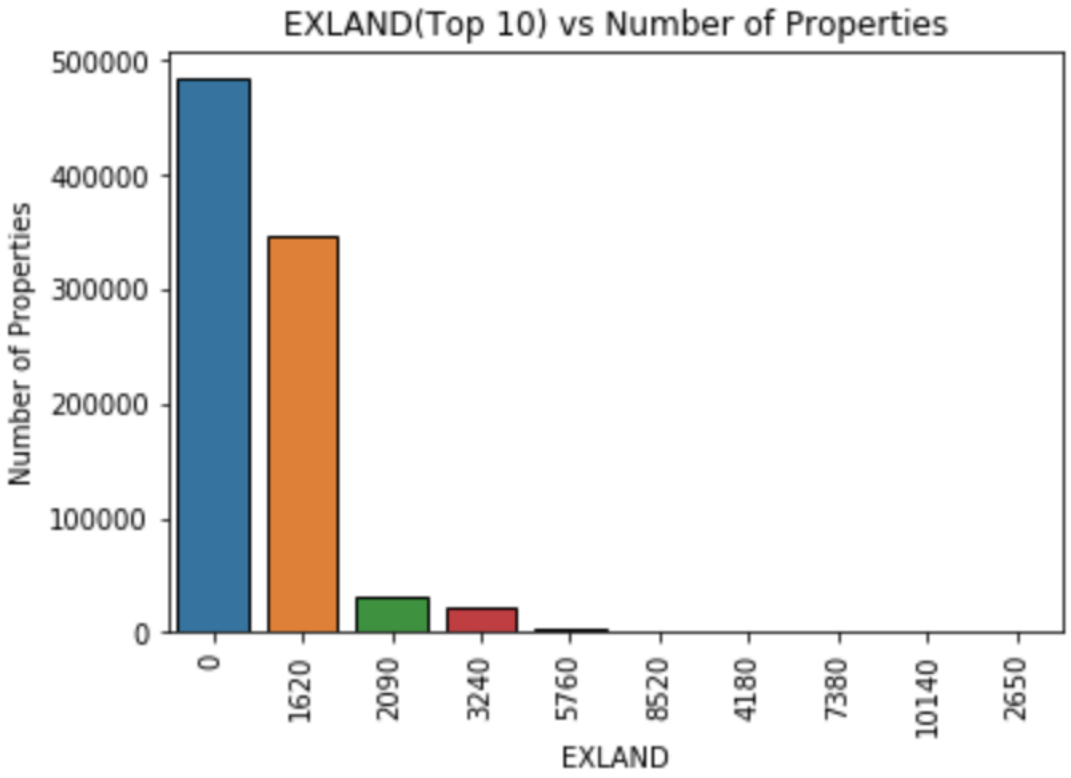
15) EXLAND

- **Description:** Continuous numeric variable
- **Number of Blank records:** 0
- **Minimum:** 0
- **Maximum:** 2668500000
- **Percentage populated:** 26.80%
- **Percentage of zeroes:** 0%
- **Mean:** 36,811.79
- **Median:** 1,620
- **Mode:** 0 with 484,224 records
- **Standard Deviation:** 4,024,329.64
- **Boxplot (to check for outliers):**



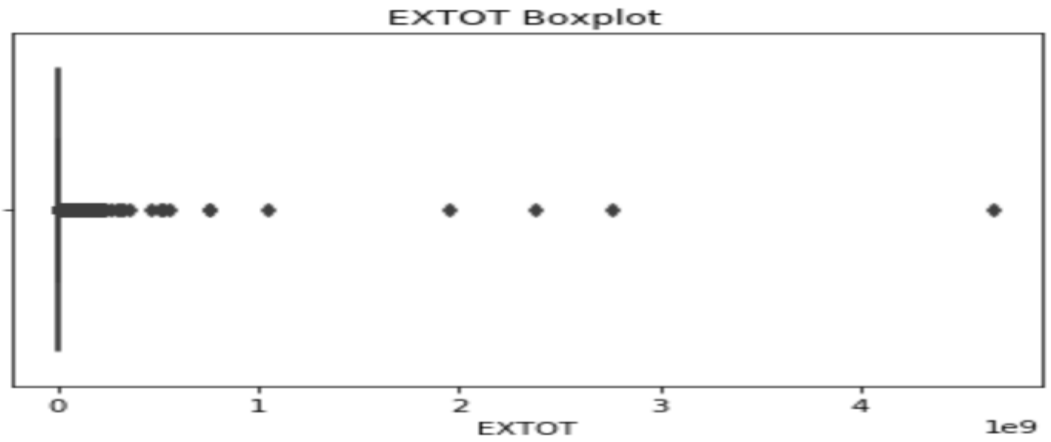
- EXLAND Distribution by Number of Properties:



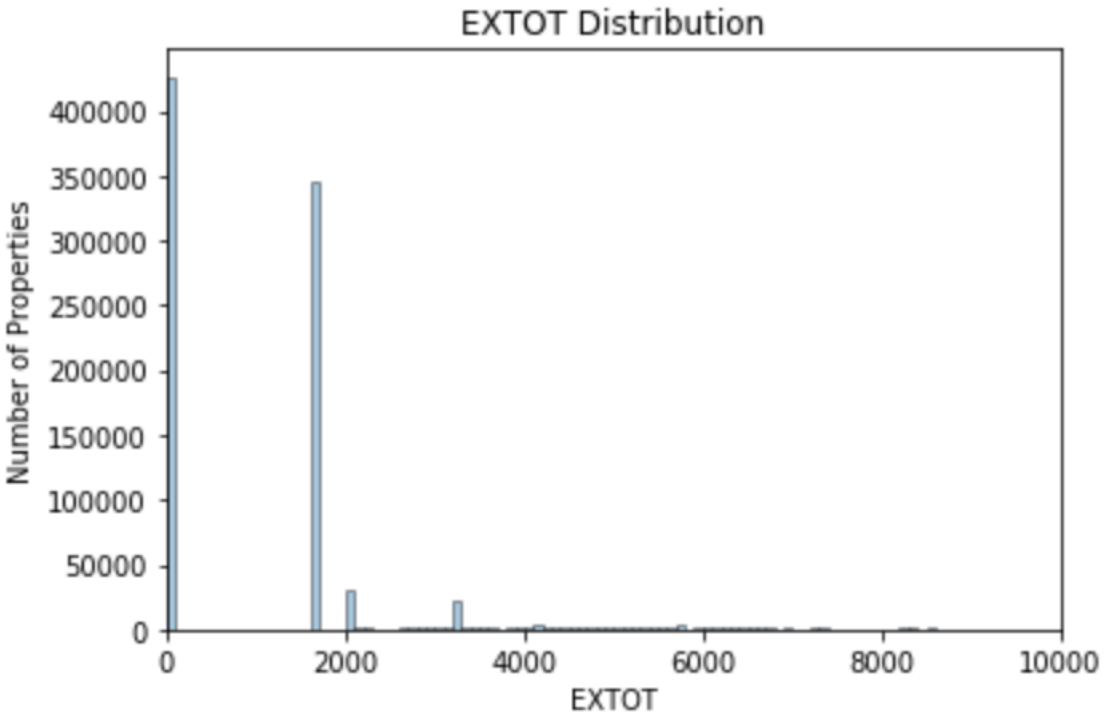


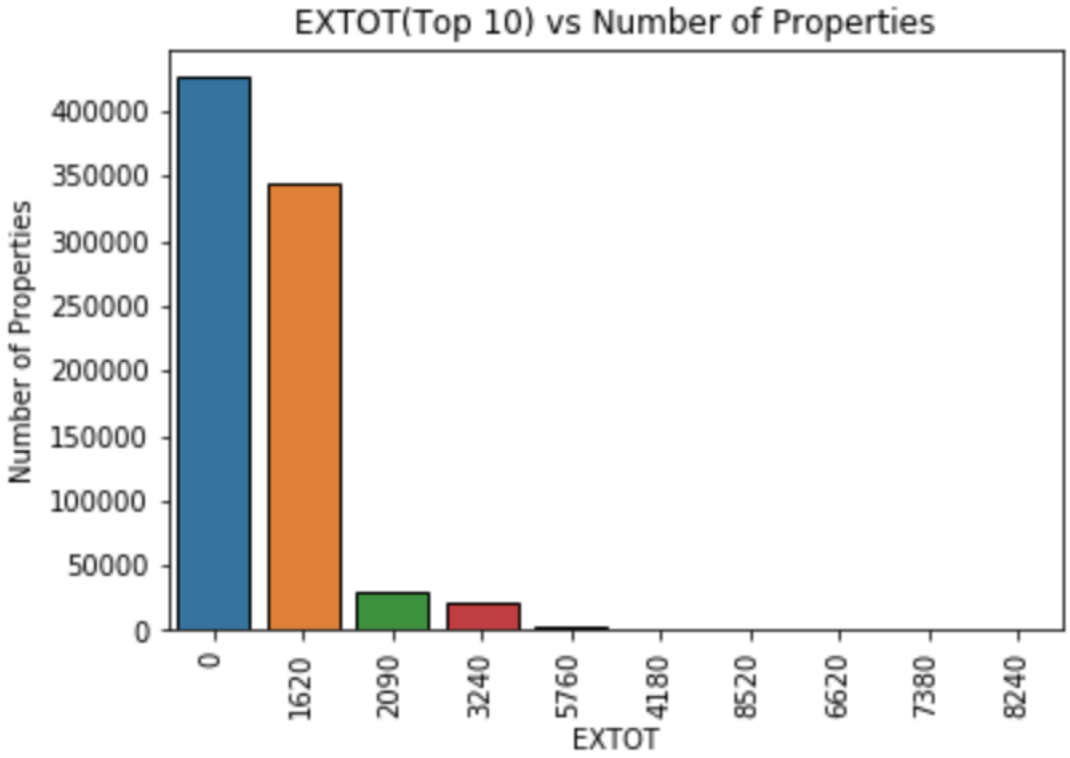
16) EXTOT

- **Description:** Continuous numeric variable
- **Number of Blank records:** 0
- **Minimum:** 0
- **Maximum:** 4,668,308,947
- **Percentage populated:** 100%
- **Percentage of zeroes:** 40.63%
- **Mean:** 92,543.82
- **Median:** 1,620
- **Mode:** 0 with 425,999 records
- **Standard Deviation:** 6,578,281.44
- **Boxplot (to check for outliers):**



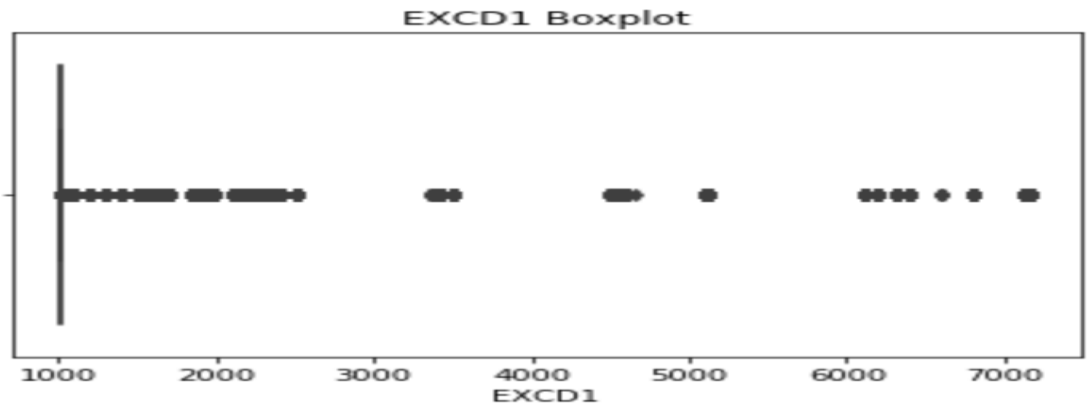
- EXTOT Distribution by Number of Properties:



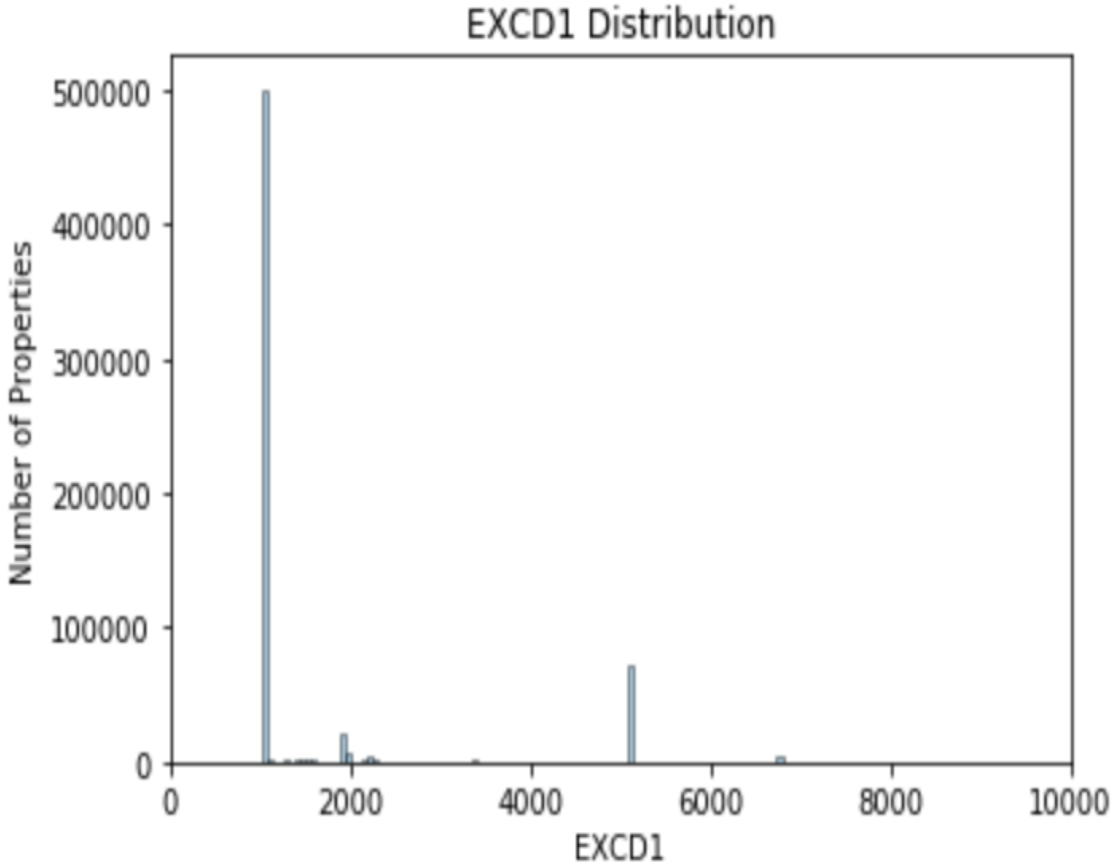


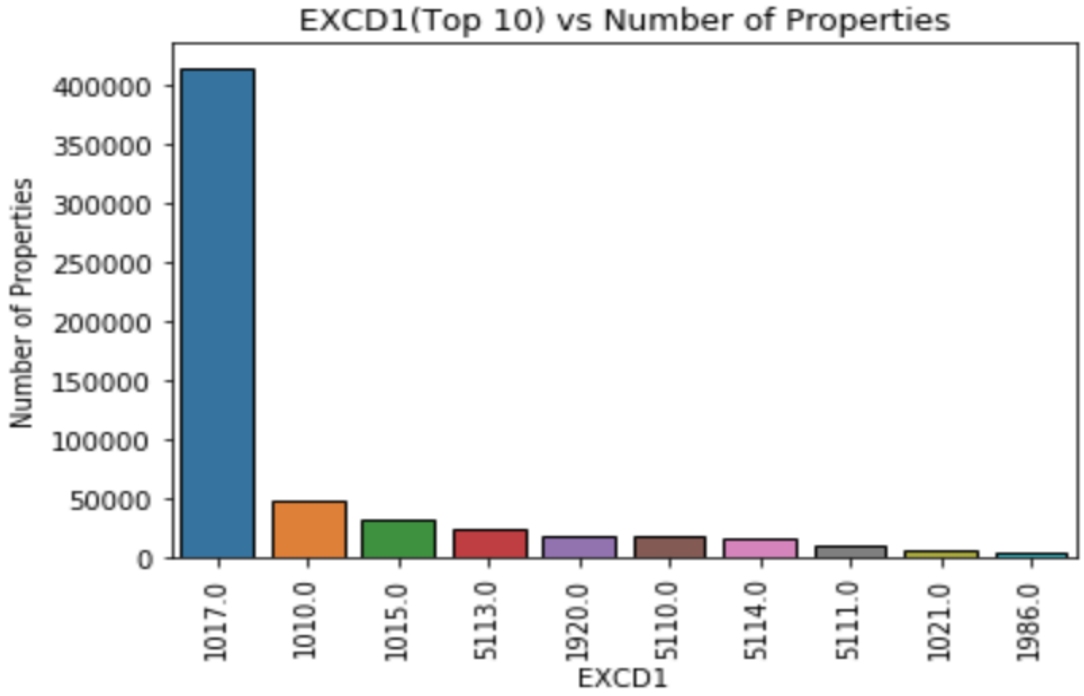
17) EXCD1

- **Description:** Continuous numeric variable
- **Number of Blank records:** 425,933
- **Percentage populated:** 59.38%
- **Minimum:** 0
- **Maximum:** 7,170
- **Mean:** 1,604.50
- **Median:** 1,017
- **Mode:** 1,017 with 414,222 records
- **Standard Deviation:** 1,388.13
- **Boxplot (to check for outliers):**



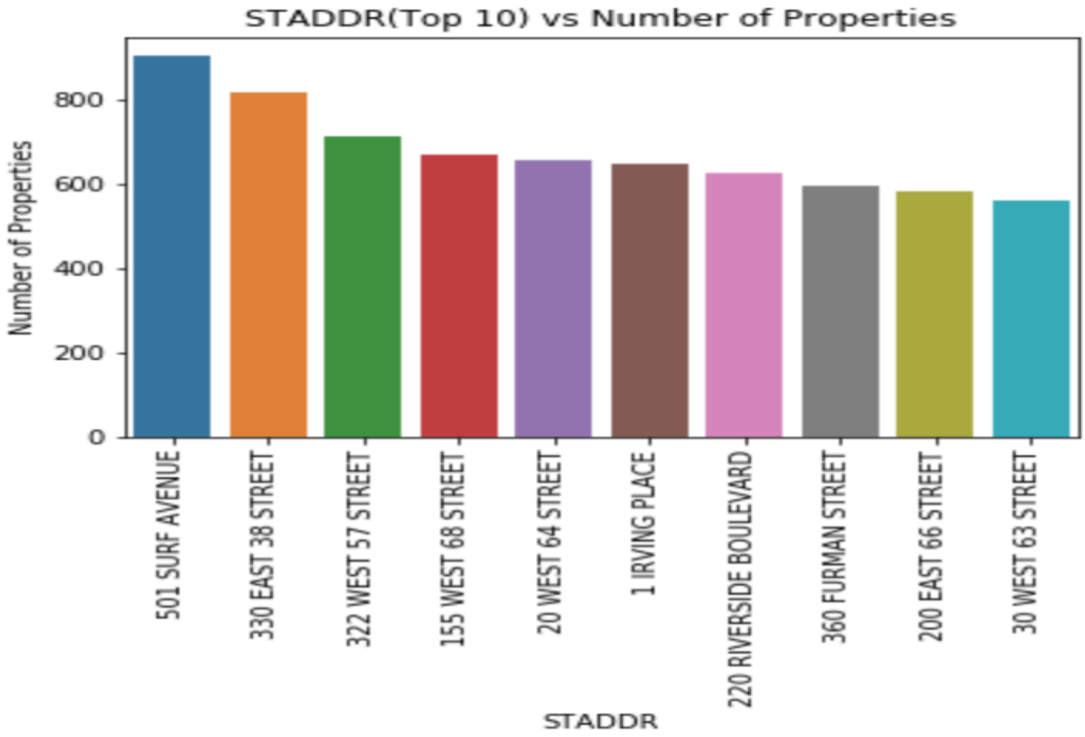
- EXCD1 Distribution by Number of Properties:





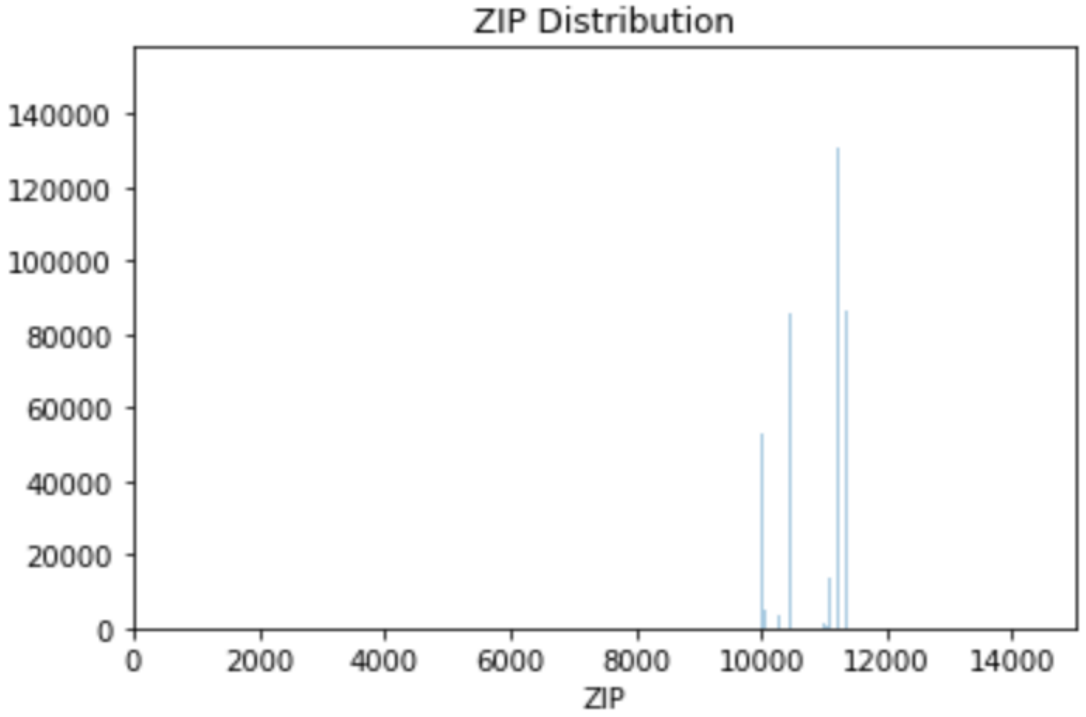
18) STADDR

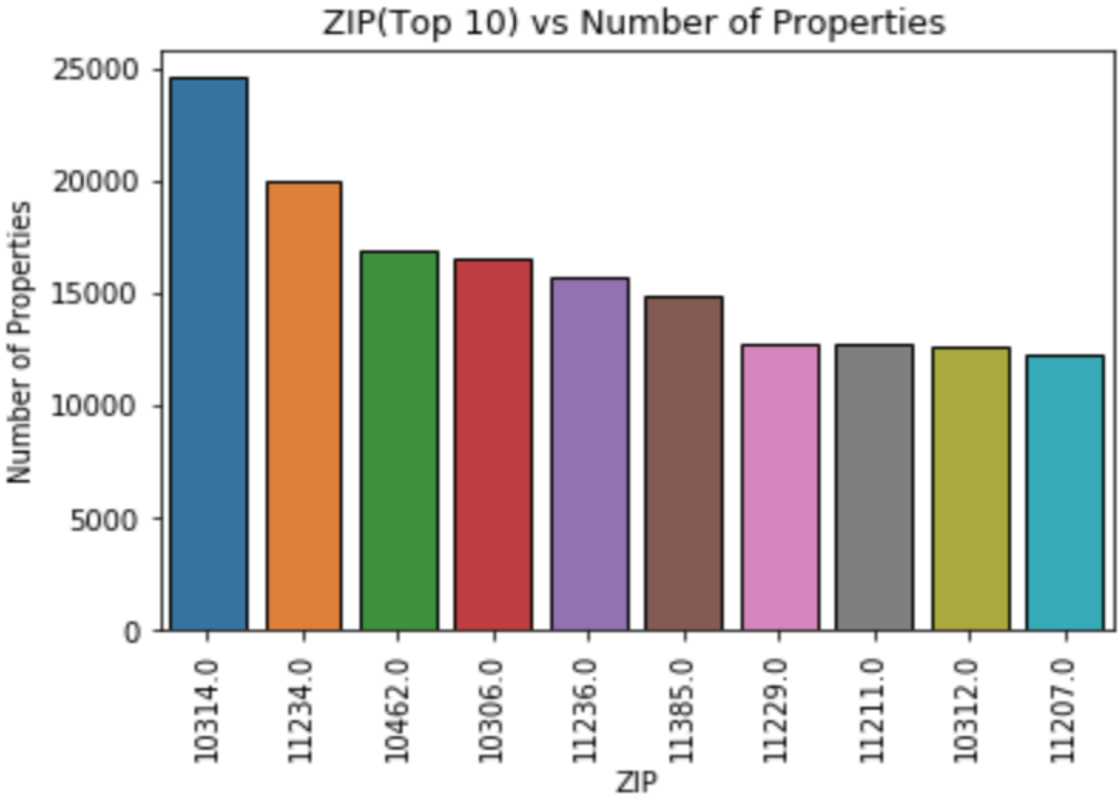
- **Description:** Categorical Variable
- **Number of Blank records:** 641
- **Number of Unique Values:** 820,637
- **Percentage populated:** 99.94%
- **Mode:** 501 SURF AVENUE with 902 records
- **STADDR Distribution by Number of Properties:**



19) ZIP

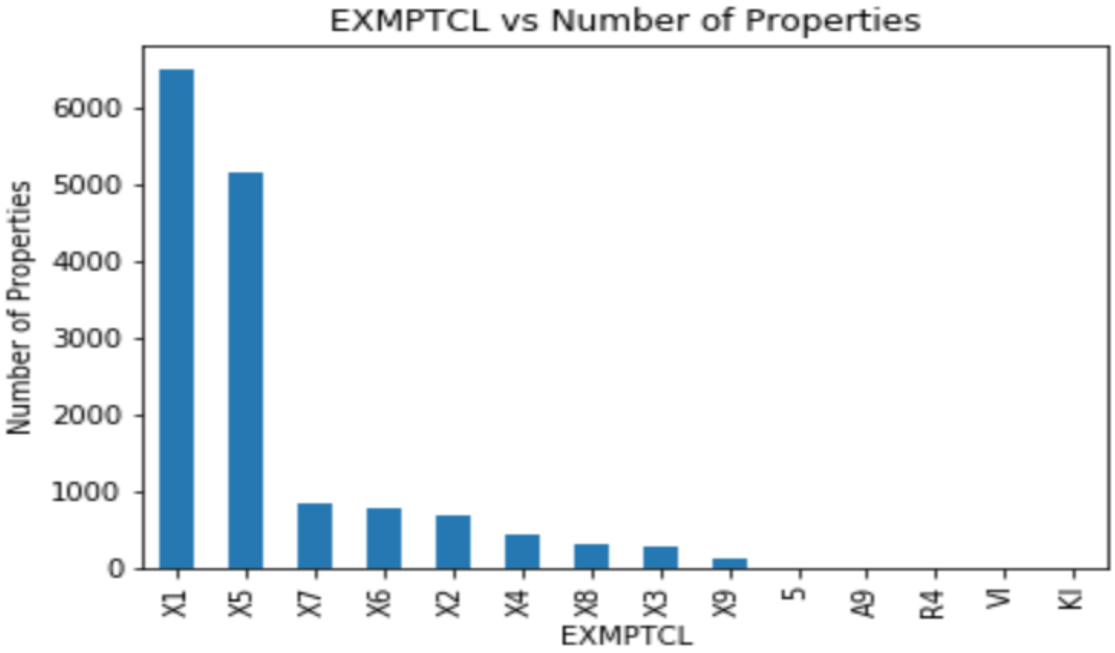
- **Description:** Discrete numeric variable describing the zip code of an area
- **Number of Blank records:** 26,356
- **Number of Unique Values:** 197
- **Percentage populated:** 97.49%
- **Mode:** 10,314 with 24,605 records
- **ZIP Distribution by Number of Properties:**


ZIP Distribution
140000
120000
100000
80000
60000
40000
20000
0
0
2000
4000
6000
8000
10000
12000
14000
ZIP


ZIP(Top 10) vs Number of Properties
Number of Properties
25000
20000
15000
10000
5000
0
10314.0
11234.0
10462.0
10306.0
11236.0
11385.0
11229.0
11211.0
10312.0
11207.0
ZIP

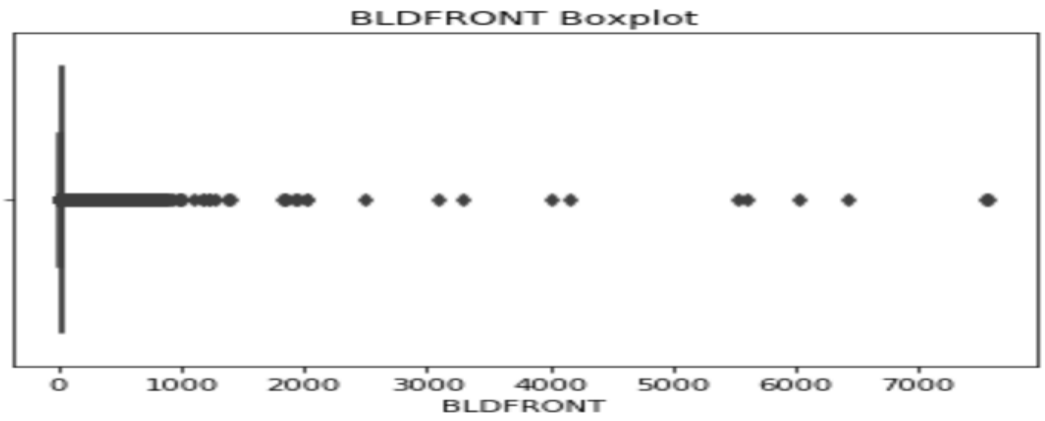
20) EXMPTCL

- **Description:** Categorical Variable of length 2 characters. Exempt Class used for fully exempt properties only, i.e. 'X1 - X9'.
- **Number of Blank records:** 1,033,583
- **Number of Unique Values:** 15
- **Percentage populated:** 1.43%
- **Mode:** X1 with 6,494 records
- **EXMPTCL Distribution by Number of Properties:**

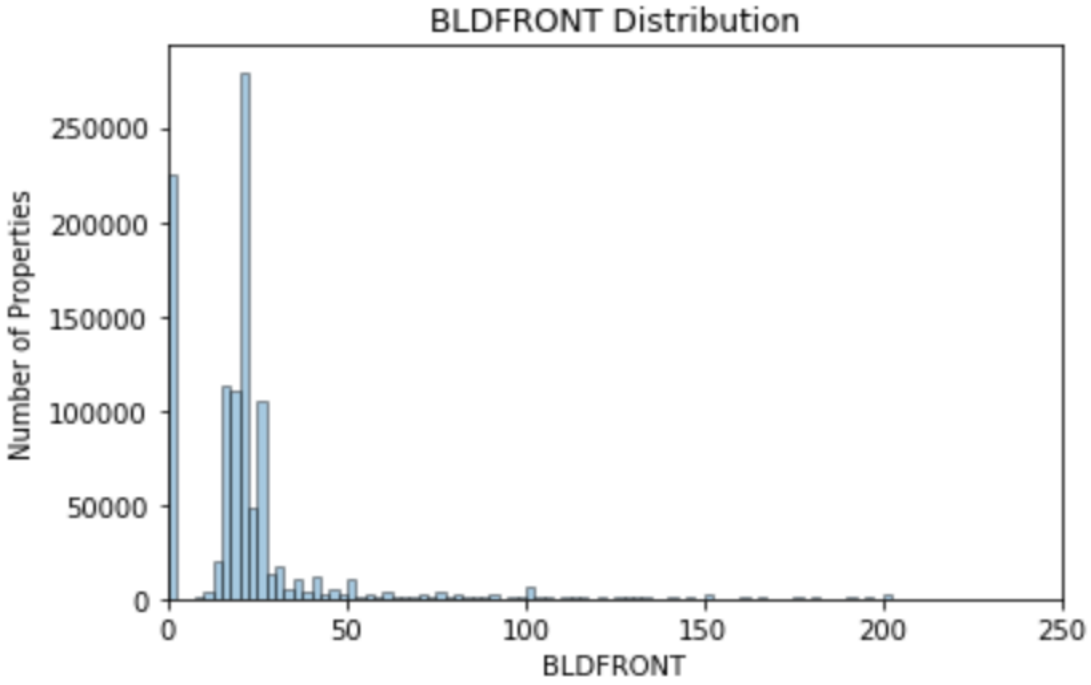


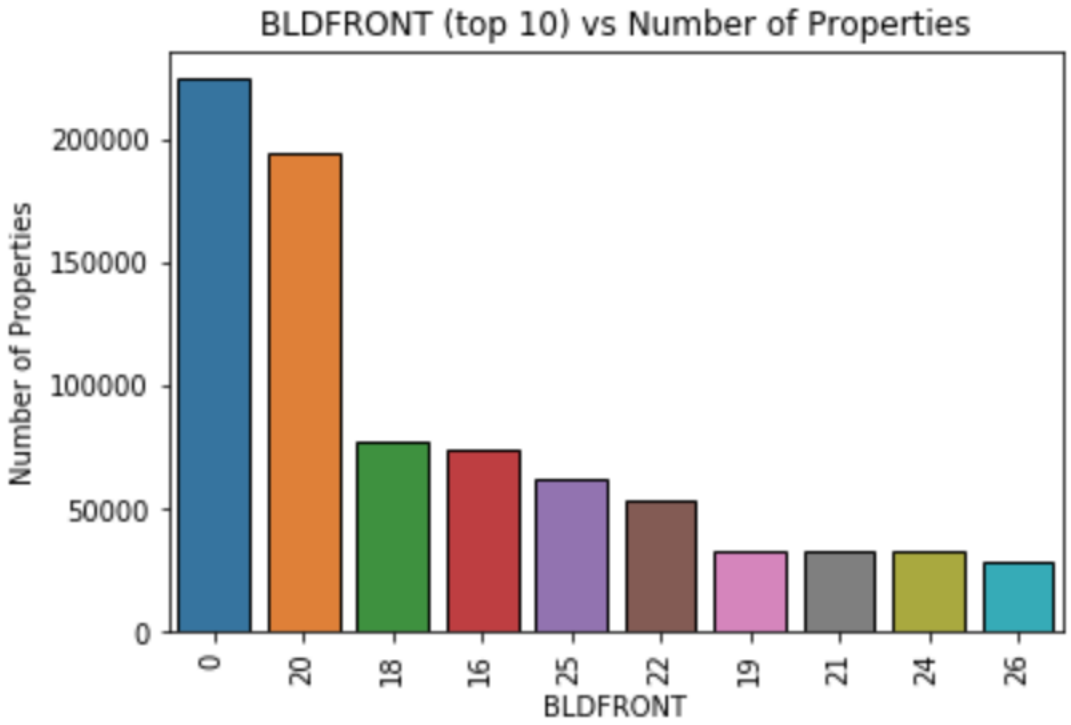
21) BLDFRONT

- **Description:** Continuous numeric variable of length 7 (9999.99) describing the building Frontage in feet
- **Number of Blank records:** 0
- **Percentage populated:** 100%
- **Percentage of zeroes:** 21.43%
- **Minimum:** 0
- **Maximum:** 7,575
- **Mean:** 23.01
- **Median:** 20
- **Mode:** 0 with 224,661 records
- **Standard Deviation:** 35.79
- **Boxplot (to check for outliers):**



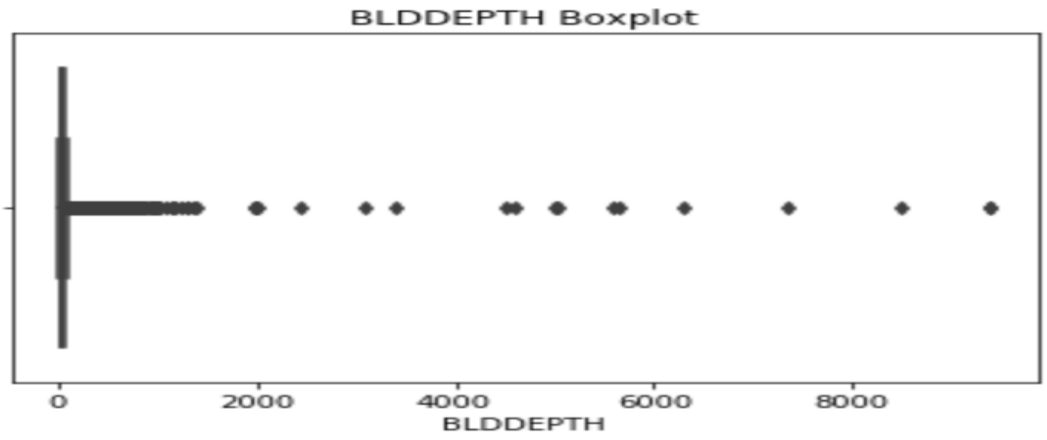
- BLDFRONT Distribution by Number of Properties:



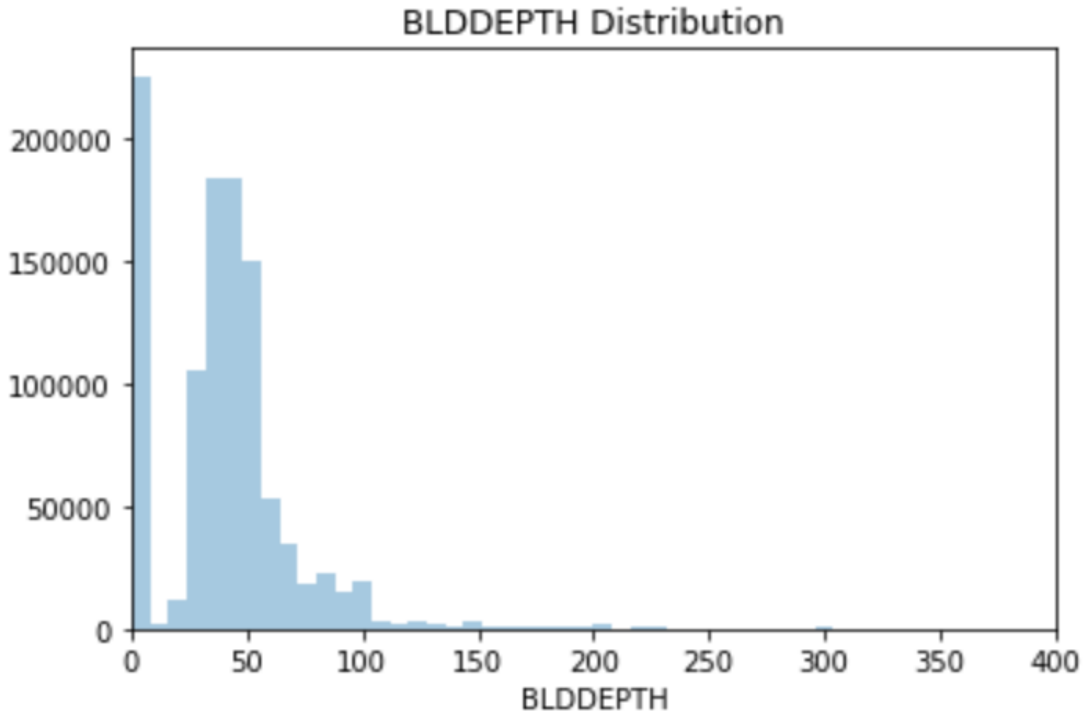


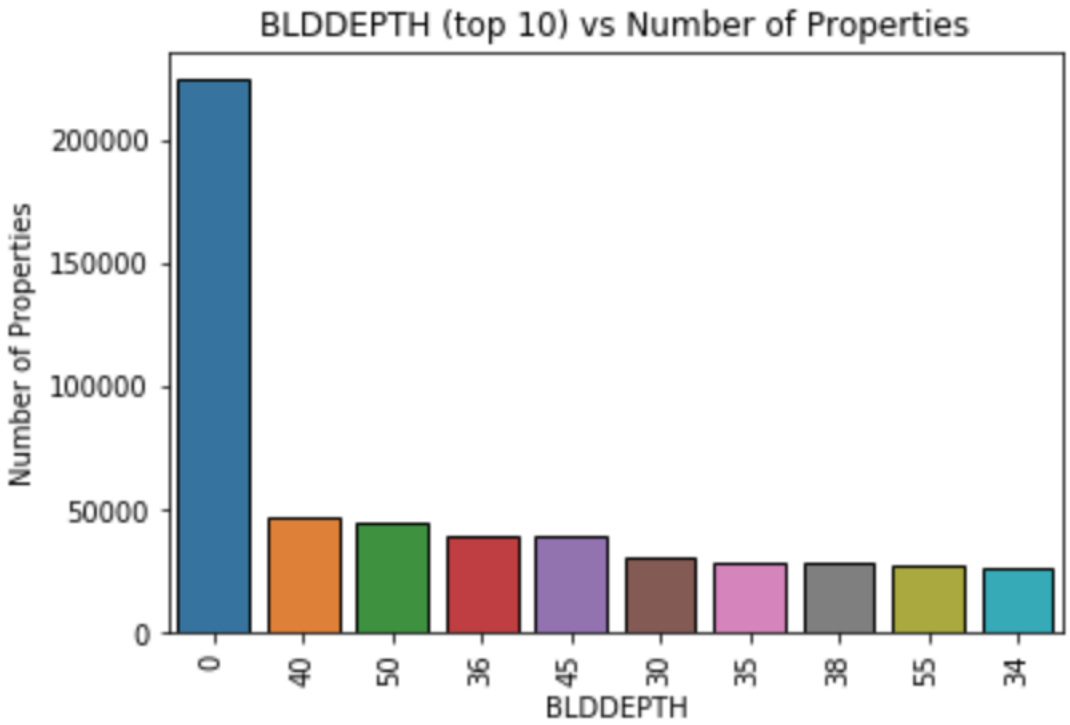
22) BLDDEPTH

- **Description:** Continuous numeric variable of length 7 for building depth in feet.
- **Number of Blank records:** 0
- **Minimum:** 0
- **Maximum:** 9,393
- **Percentage populated:** 100%
- **Percentage of zeroes:** 21.43%
- **Mean:** 40.07
- **Median:** 39
- **Mode:** 0 with 224,699 records
- **Standard Deviation:** 43.04
- **Boxplot (to check for outliers):**



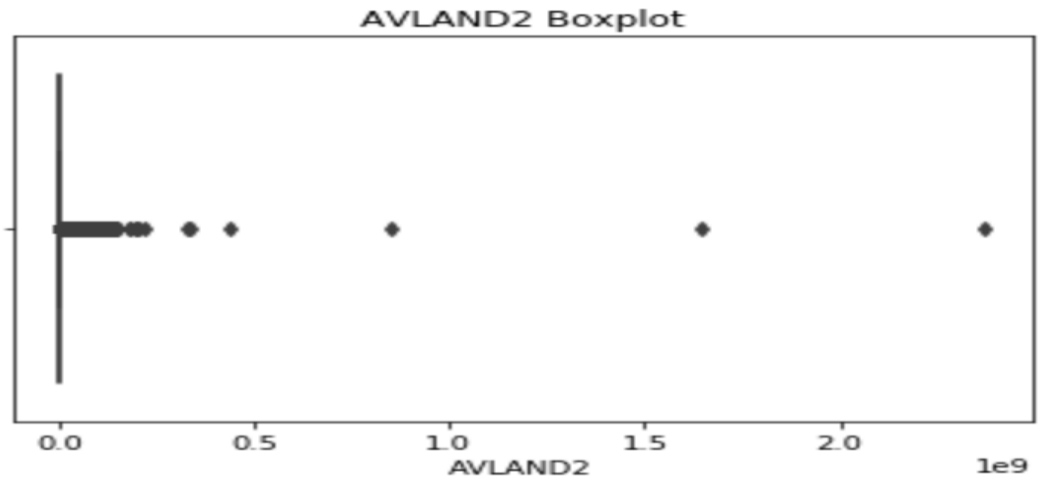
- BLDDEPTH Distribution by Number of Properties:



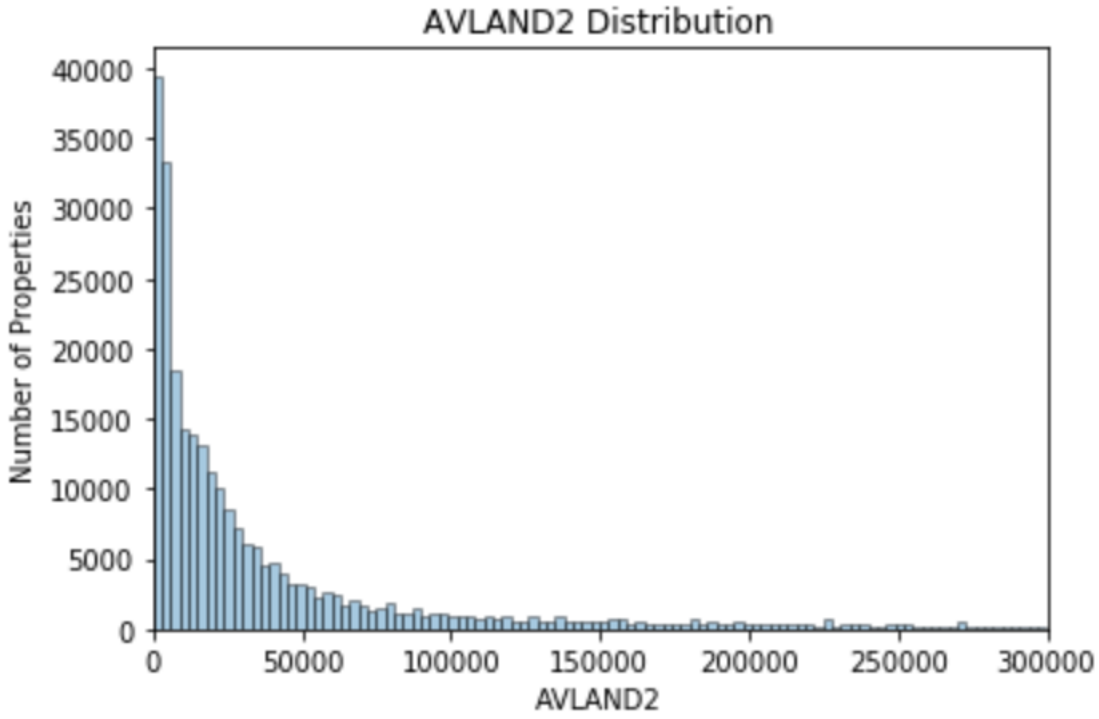


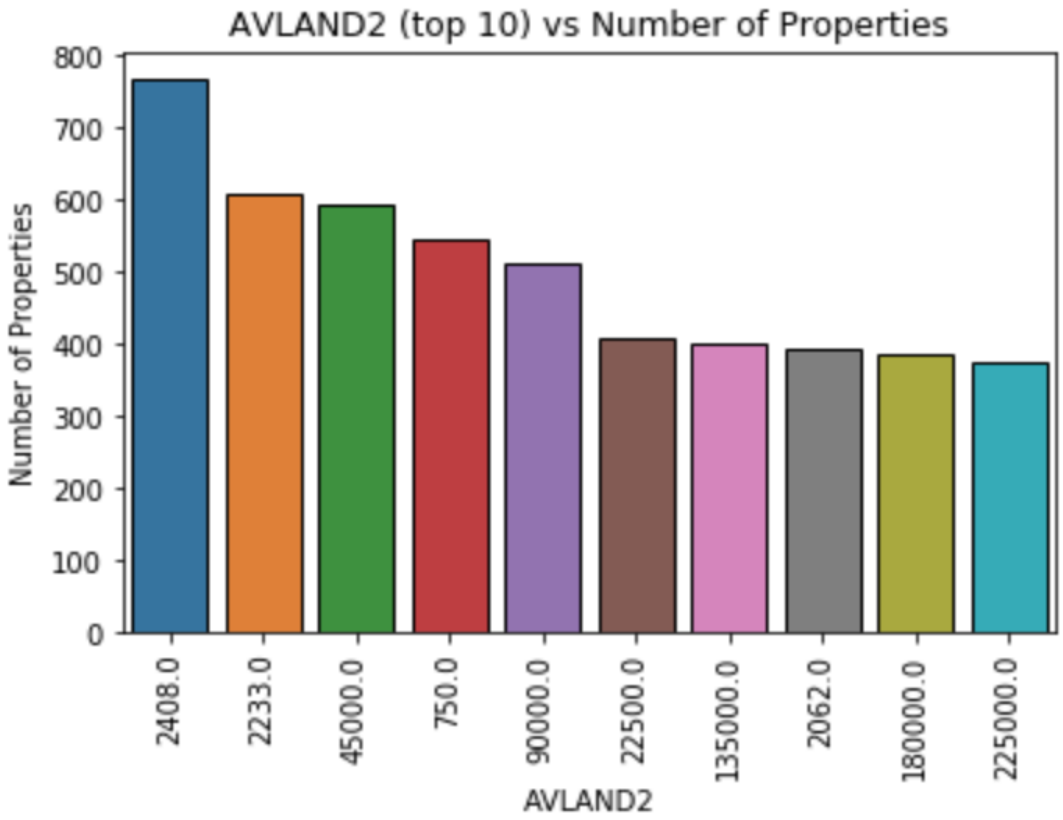
23) AVLAND2

- **Description:** Continuous numeric variable
- **Number of Blank records:** 767,609
- **Minimum:** 3
- **Maximum:** 2,371,005,000
- **Percentage populated:** 26.80%
- **Percentage of zeroes:** 0%
- **Mean:** 246,365.48
- **Median:** 20,059
- **Mode:** 2,408 with 767 records
- **Standard Deviation:** 6,199,389.59
- **Boxplot (to check for outliers):**



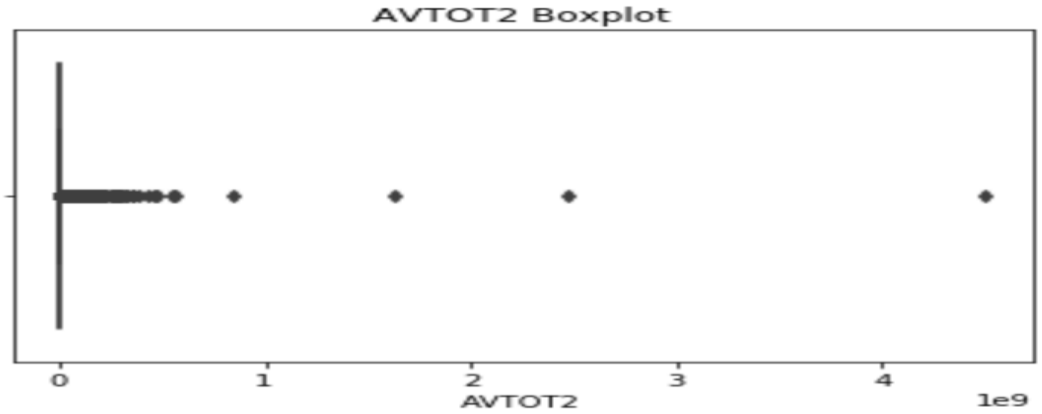
- AVLAND2 Distribution by Number of Properties:



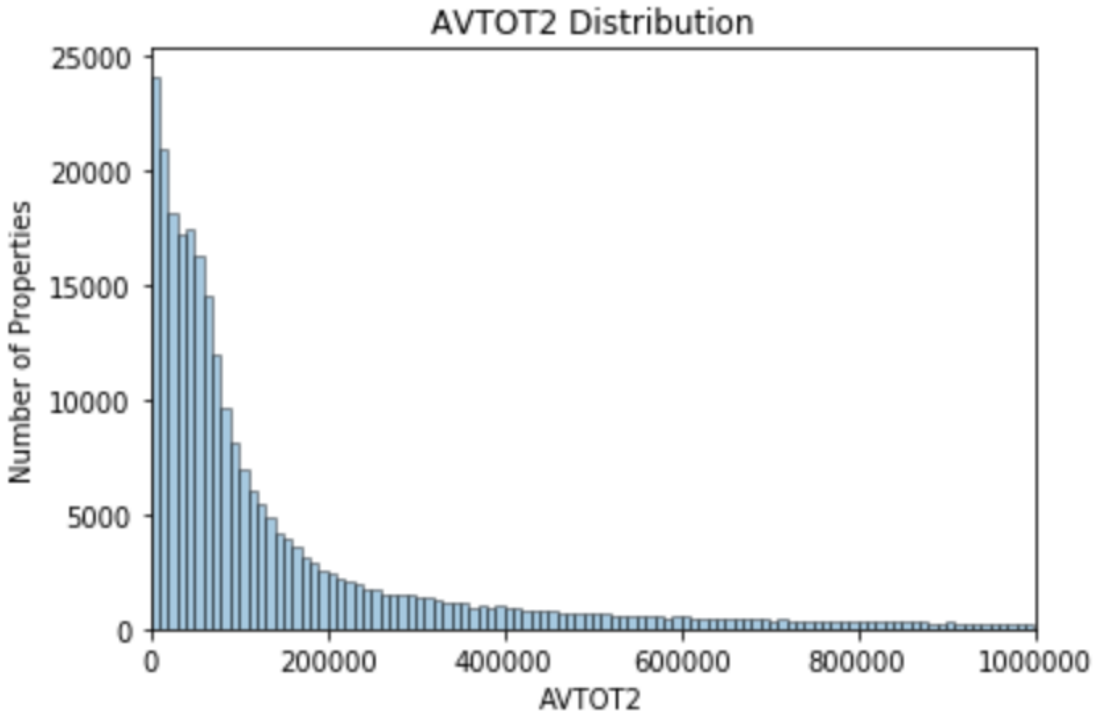


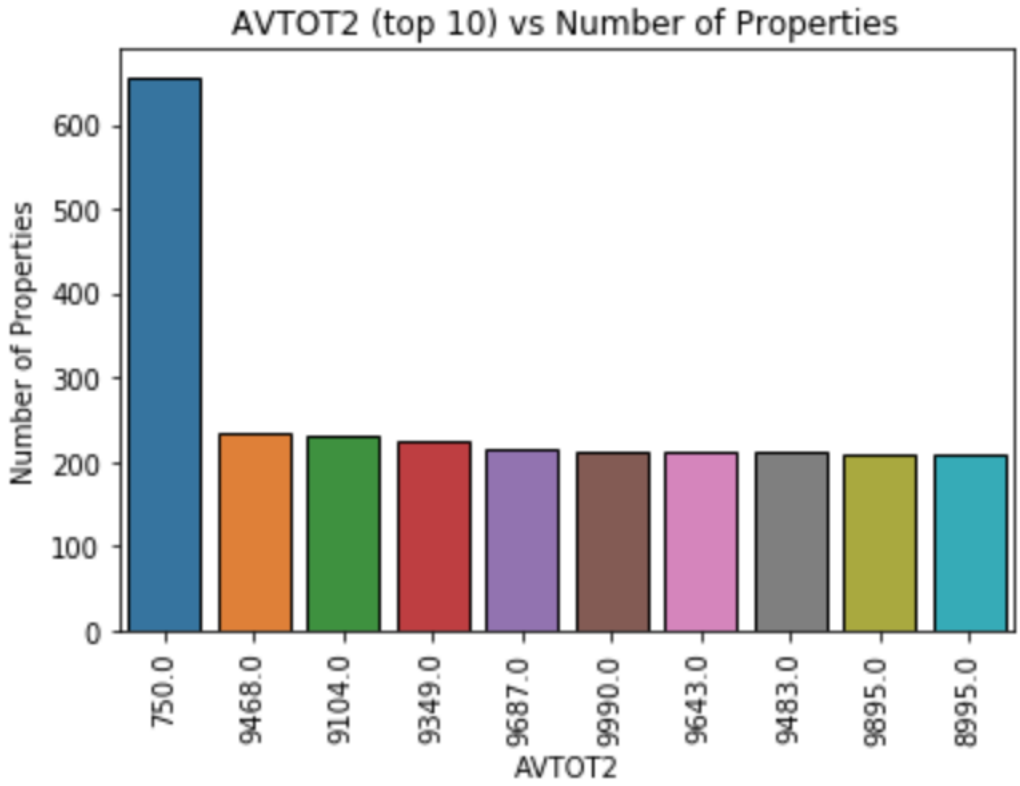
24) AVTOT2

- **Description:** Continuous numeric variable
- **Number of Blank records:** 767,603
- **Percentage populated:** 26.80%
- **Percentage of zeroes:** 0%
- **Minimum:** 3
- **Maximum:** 4,501,180,002
- **Percentage populated:** 26.80%
- **Percentage of zeroes:** 0%
- **Mean:** 716,078.71
- **Median:** 80,010
- **Mode:** 750 with 656 records
- **Standard Deviation:** 11,690,165.49
- **Boxplot (to check for outliers):**



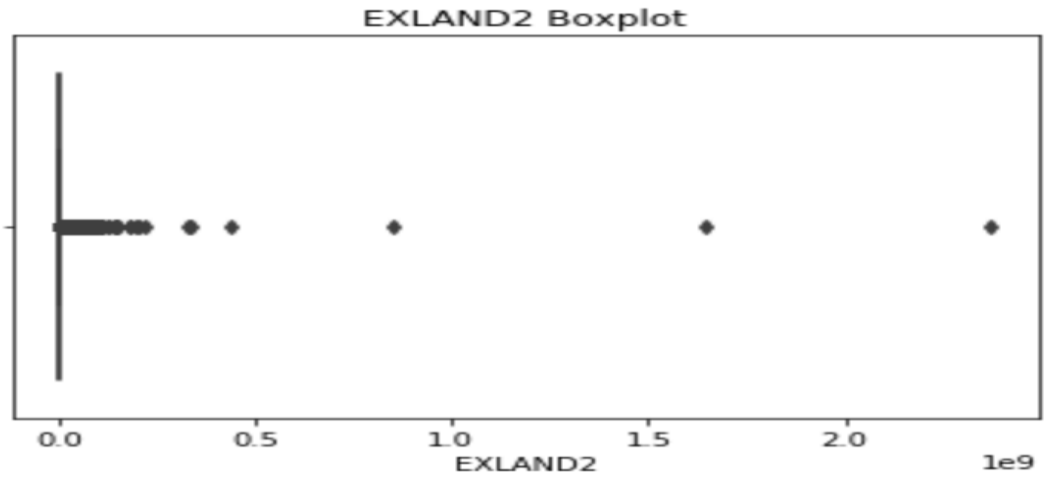
- AVTOT2 Distribution by Number of Properties:



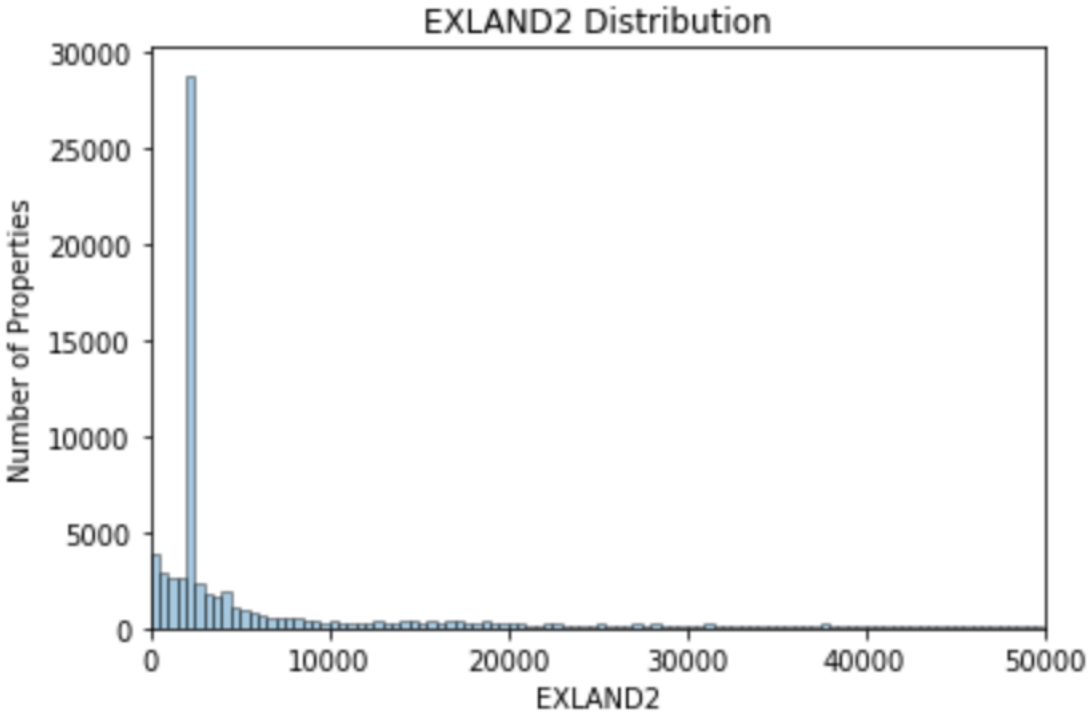


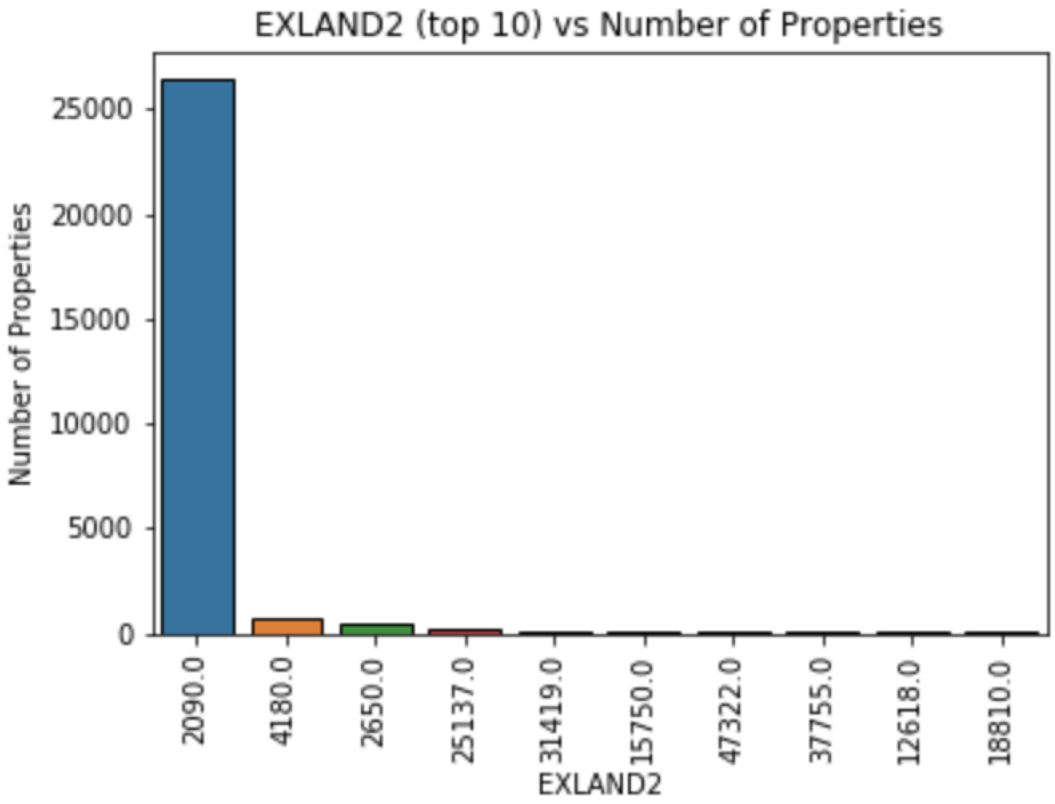
25) EXLAND2

- **Description:** Continuous numeric variable
- **Number of Blank records:** 961,900
- **Percentage populated:** 8.27%
- **Percentage of zeroes:** 0%
- **Minimum:** 1
- **Maximum:** 2,371,005,000
- **Mean:** 351,802.21
- **Median:** 3,053
- **Mode:** 2,090 with 26,393 records
- **Standard Deviation:** 10,852,484.04
- **Boxplot (to check for outliers):**



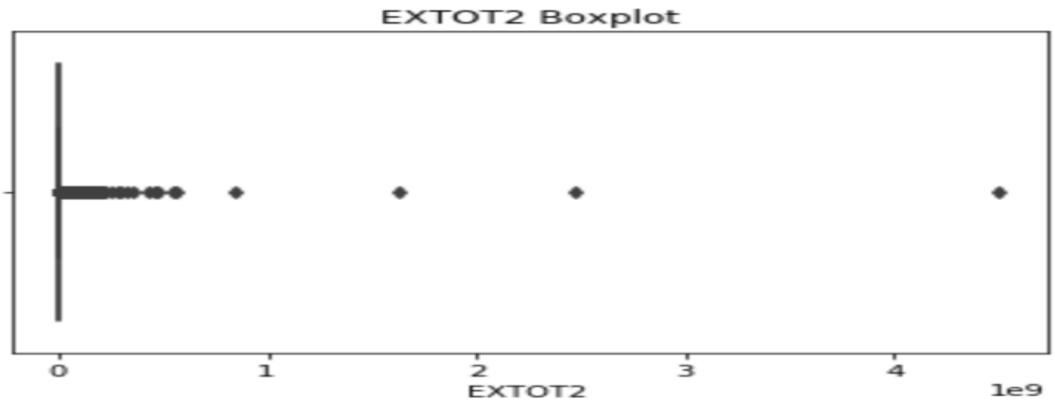
- EXLAND2 Distribution by Number of Properties:



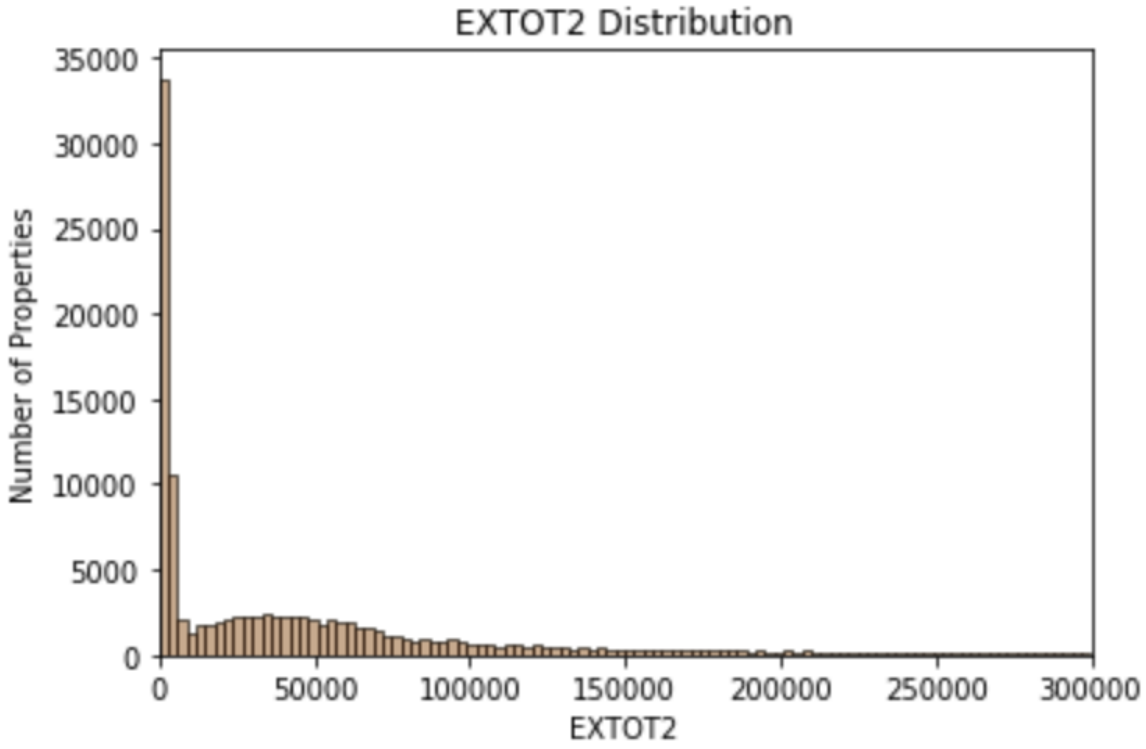


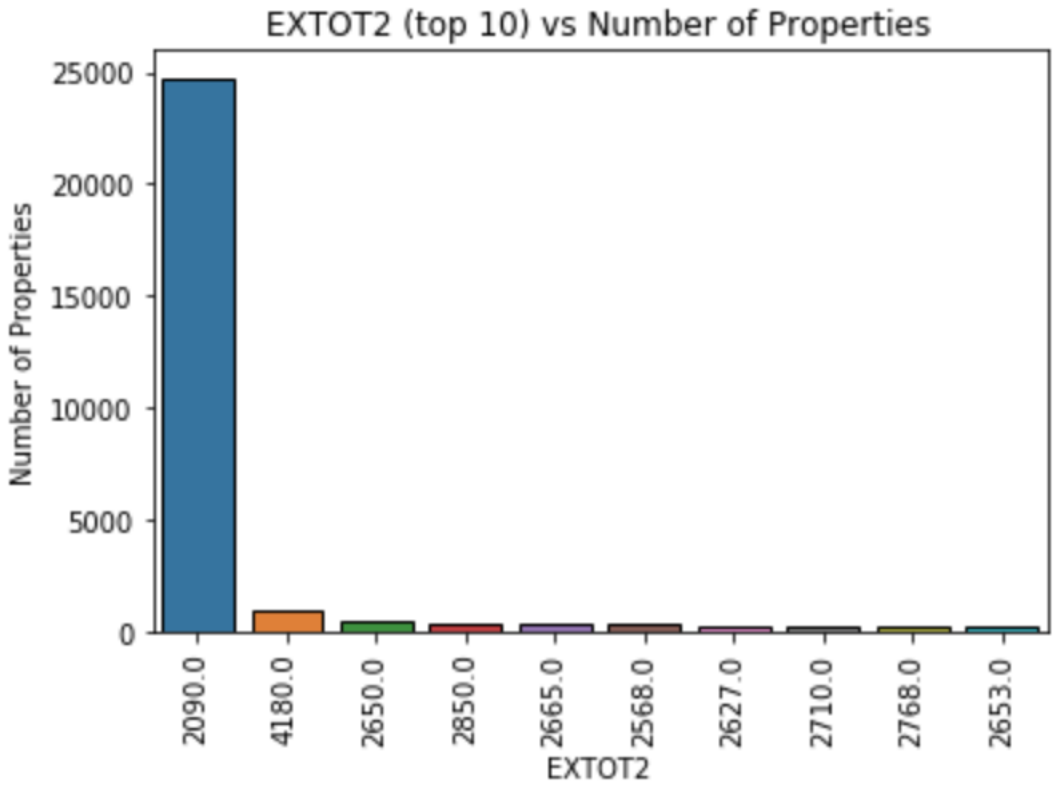
26) EXTOT2

- **Description:** Continuous numeric variable
- **Number of Blank records:** 918,642
- **Percentage populated:** 12.39%
- **Percentage of zeroes:** 0%
- **Minimum:** 7
- **Maximum:** 4,501,180,002
- **Mean:** 658,114.78
- **Median:** 37,116
- **Mode:** 2,090 with 24,739 records
- **Standard Deviation:** 16,129,808
- **Boxplot (to check for outliers):**



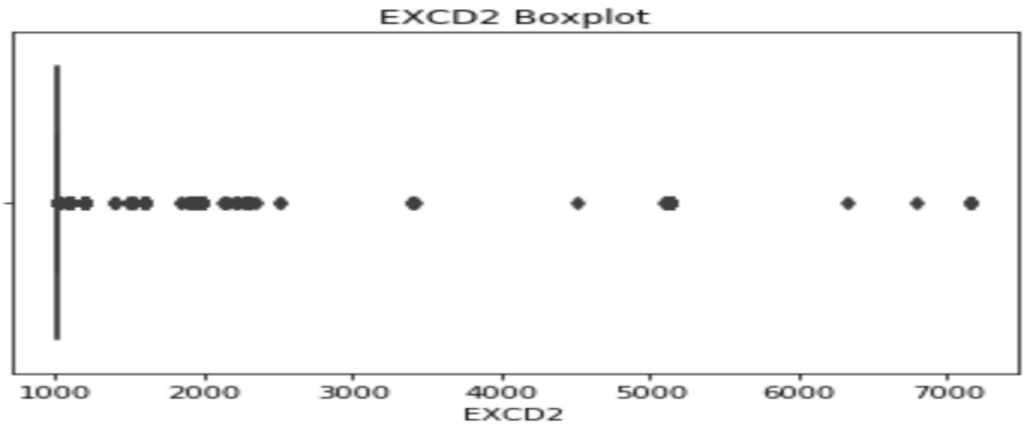
- EXTOT2 Distribution by Number of Properties:



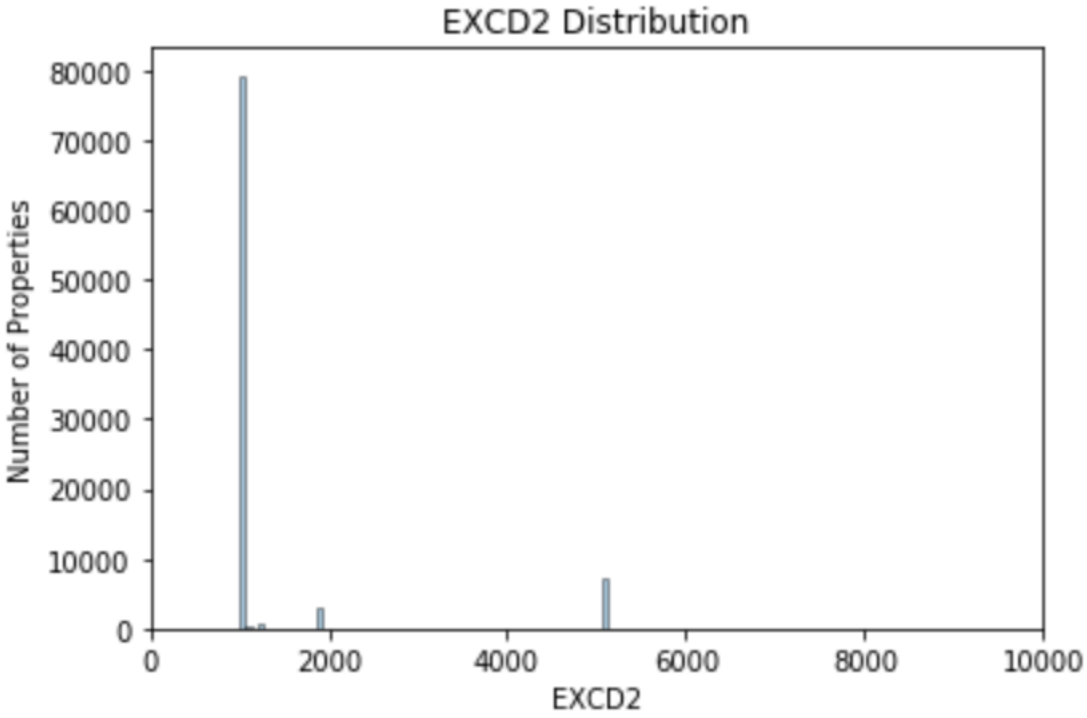


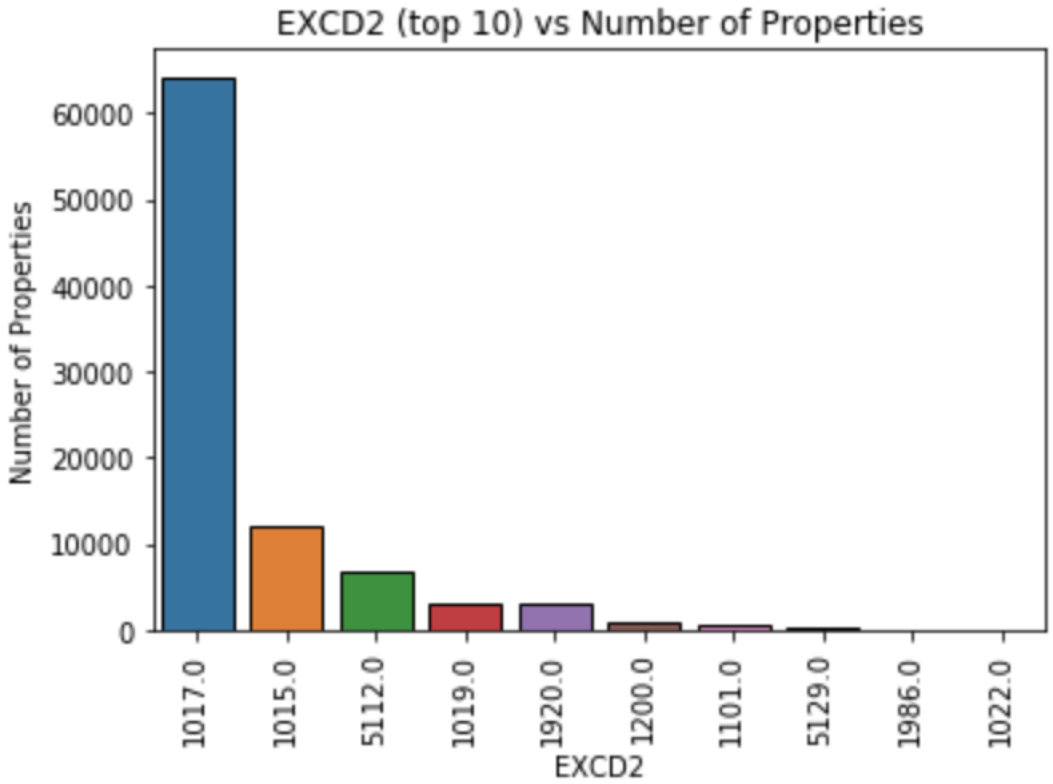
27) EXCD2

- **Description:** Continuous numeric variable
- **Number of Blank records:** 957,634
- **Percentage populated:** 8.67%
- **Percentage of zeroes:** 0%
- **Minimum:** 1,011
- **Maximum:** 7,160
- **Mean:** 1,371.65
- **Median:** 1,017
- **Mode:** 1,017 with 64,223 records
- **Standard Deviation:** 1,105.48
- **Boxplot (to check for outliers):**



- EXCD2 Distribution by Number of Properties:





28) PERIOD

- **Description:** Categorical Variable
- **Number of Blank records:** 0
- **Number of Unique Values:** 1, i.e. FINAL

29) YEAR

- **Description:** Categorical Variable
- **Number of Blank records:** 0
- **Number of Unique Values:** 1, i.e. 2010/11

30) VALTYPE

- **Description:** Categorical Variable
- **Number of Blank records:** 0
- **Number of Unique Values:** 1, i.e. AC-TR