Detecting Fraud in New York
Feb 2018

DSO 562 – Project 1 Report

Team 8-Fraud Police:
- Babak Biglari
- Raman Kumar
- Ruhan Dong
- Modan Wang
- Zhengli Zhang
- Chun Yang

Contents

1. Executive Summary

This report build a fraud model for the NY Property data and tries to detect fraud using unsupervised machine learning methods. This data includes quantitative analysis for NY property Valuation and Assessment Data and has over 1M records for properties across the city of New York and information on their sizes, values, owner, building classes, tax classes, etc. The analysis includes data cleaning, choosing and building special variables, standardization and dimensions reduction, building fraud algorithm, calculating fraud score, and finally identifying potential fraud cases. The tools used are R and analysis methods include Z-Scaling, Principal Component Analysis (PCA) and Autoencoder.

During data analysis, we decided to group our data using "Tax Class" to serve as our linking intermediary to group the data and to be able to understand the relationship among the data. For missing fields' values, we decided to fill in the empty data by average values. Our scaling process includes Z-scaling and quantile binning. Also, we decided to leave the outliers in the data since in fraud detection algorithm. We used the value variables such as FULLVAL divided by the space variables such as LOTAREA to construct derivative variables.

Using outlier detection via z-scores and 'h2o' package in R to perform autoencoder, we calculated fraud score for each of the one million properties by doing sum of the errors and taking the square root of the sum. Records with high scores are determined to be potentially fraudulent.

Detailed examination of the most suspicious records indicates that potentially fraudulent properties have significantly higher values in a lot of variables compared to the majority of records. Few properties having tax class 1D (Bungalow Colonies), however they belongs empty lot. In another case, property has a small "LTFRONT" and "LTDEPTH", however, the "FULLVAL" is unusually large. Some properties are also significantly undervalued and hence paying lower taxes than they should. Meanwhile, most of the potentially fraudulent properties are located in Manhattan borough and belong to the tax class 4.

Further examination of the top 10 most suspicious records shows that the owners of these properties are mostly real estate agencies and organizations instead of single households. By considering the top 10 results, we noticed few of the records belong to the government. In our analysis, our focus was identifying possible fraud against both local and federal government. Therefore, we excluded those properties from our top 10 list.

2. Data overview and manipulation

After we were informed about this project and provided with the raw data, we started our research by talking to our domain experts. Through our conversation with domain experts, we gained better understanding about variables in data set and learned about important industry metric. These metrics assisted us to create better variables for our algorithm. Our clients are interested in finding possible fraudulent records. By definition fraud analysis is all about finding the odd value between given data. Therefore, our focus and goal in this project is to identify irregularities within the records. We acknowledge that this analysis can be very detailed by considering all the stakeholders and entities involved. However, at this stage, we are presenting minim viable product with simplicity in mind to ensure our algorithm performs the best. Further and far more detail research and data analysis needed to consider all the touchpoints. Below is the process in which we used to analyze the case.

2.1. Data Quality Report

After understanding the client's concerns and needs, we decided to dedicate a generous amount of time in familiarizing ourselves with the data. Data Quality Report(DQR) was conducted to get a better understanding of each variables, and to understand how many values and entries are missing in the data. Missing information is part of any data, however, in the case of fraud detection, missing values could have an added significance in our algorithm. A complete guide regarding DQR can be found in the appendix.

2.1. Missing Fields

In order to address our missing value fields, we started on focusing on the end goal: finding abnormal data. It is important to fill out missing entries by information which are not abnormal themselves. Therefore, we decided to use neutral values to fill in the data. We considered the following methods:

- replacing values by the average of total data
- replacing by most common value in the entire data set
- replacing by average considering different grouping and linkages, such as Borro, Tax Class, etc.

Below is how we decided to fill-in the missing variables:

Variable	Replacement
If STORIES = 0	we filled in the missing values with the average "STORIES" in its "TAXCLASS"
If ZIP = 0	we filled in the missing values with "00000"
If LTFRONT=LTDEPTH= 0	LTFRONT= 30 and LTDEPTH =100
If BLDFRONT=BLDDEPTH= 0	BLDFRONT= 20 and BLDDEPTH = 40
If FULLVAL= 0	We replaced the zero value with the mean of their tax group
If AVLAND= 0	
If AVTOT= 0	
If LTFRONT= 0	
If LTDEPTH= 0	
If BLDFRONT= 0	
If BLDDEPTH= 0	

We decided to group our data using "Tax Class". These variables served as our linking intermediary to group the data and to be able to understand the relationship between the data. This allowed us to aggregate information based on these groups and normalize fields by these groupings. Eventually, this enabled us to identify the anomaly in a set of information in that group. For instance, looking at the average of assessed value of the land for a certain tax class, allows us to look at a value and determine if that observation is abnormal.

2.3. Variable Selection and Variable construction

For the variables BBLE, we extracted its first digit and changed the variable name to "BORO", indicating the borough where the property located. All the values in variables ZIP have 3 digits. Therefore, we change the variable name to "ZIP5". Also, we extracted its first three digits and create a new variable named "ZIP3".

Following in depth variable analysis, we noticed that there are few variables which are not informative. Those are "STADDR", "OWNER", "BLOCK", "LOT", "PERIOD", "YEAR" and "VALTYPE". In this part of our analysis, we try to keep variables which are going to contribute to making a better decision. Name and Address, even though important, cannot be used in algorithm to draw conclusion with the given data. Other variables, such as Year, are not adding new information so we decided to ignore them in this analysis. Based on our DQR, we have also other fields which are missing a lot of data, such as "EXCD1", "EXMPTCL", "AVLAND2", "AVTOT2", "EXLAND2", "EXTOT2" and "EXCD2". Even though missing data in fraud detection is important,

however, we will not be able to create a good algorithm when majority of the data are missing. Therefore, we excluded the above variables as well.

2.3.1 Combining existing variables

We defined the product of the variables LTFRONT and LTDEPTH as a new variable lotarea, indicating the lot size of each property. We multiplied the values of BLDFRONT, BLDDEPTH and defined the output as a new variable bldarea, indicating the area of each building. Finally, we multiplied the values of bldarea and STORIES, and defined the output as a new variable bldvol, indicating the volume of each building.

Variable Reference		
	Name	Equation
Primary	**FULLVAL**	full value of building
	AVLAND	assessed value of land
	AVTOT	assessed value of property
Derived special variables	**lotarea**	lotarea = LTFRONT X LTDEPTH
	bldarea	bldarea = BLDFRONT X BLDEPTH
	bldvol	bldvol = bldarea X STORIES
Derivatives	**fV_la**	$fV_la = \frac{FULLVAL}{lotarea}$
	fV_ba	$fV_ba = \frac{FULLVAL}{bldarea}$
	fV_bv	$fV_bv = \frac{FULLVAL}{bldvol}$
	vl_la	$vl_la = \frac{AVLAND}{lotarea}$
	vl_ba	$vl_ba = \frac{AVLAND}{bldarea}$
	vl_bv	$vl_bv = \frac{AVLAND}{bldvol}$
	vt_la	$vt_la = \frac{AVTOT}{lotarea}$
	vt_ba	$vt_ba = \frac{AVTOT}{bldarea}$
	vt_bv	$vt_bv = \frac{AVTOT}{bldvol}$

2.3.2. The derivative 9 variables

We used the value variables divided by the space variables to construct derivative 9 variables. Since value and space are the attributes that vary from one property to another property, such normalization will make new features that are more comparable among different properties.

2.3.3. Final 45 expert variables

By using zip5, zip3, tax class, borough, all (non-group), we created the grouped averages of these 9 variables and got the final expert 45 variables. For example, if we used FULLAL (numerator) and LOTAREA (denominator) grouped by zip5, the expert variable would be mean of FULLVAL/lotarea grouped by zip5.
In total, we created 45 expert variables.

2.4. Scaling and Normalization

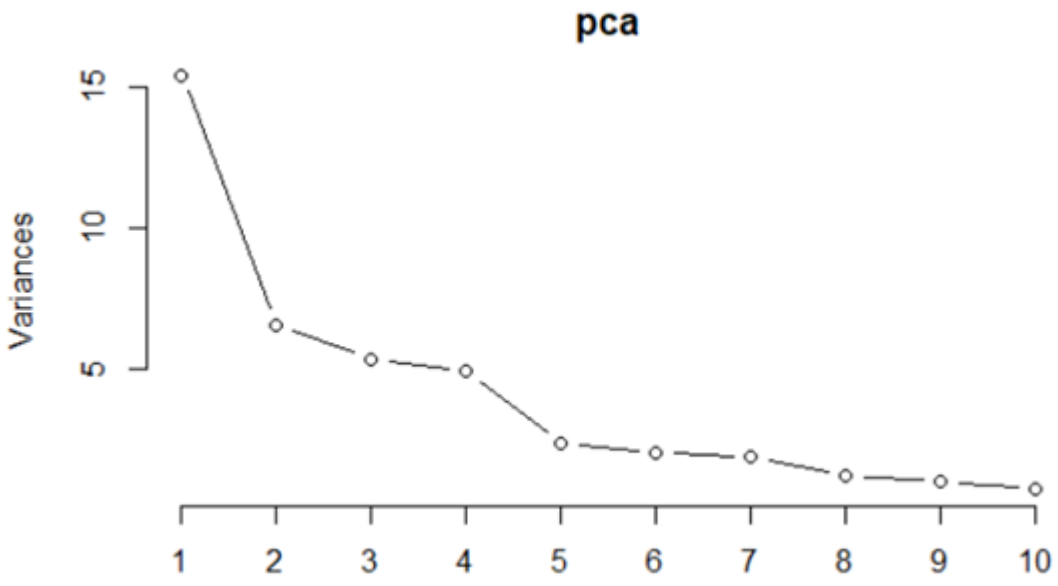
Most machine learning algorithms are very sensitive to data in different scales and would experience difficulties in handling those differences. Therefore, it is logical to put all of our variables into the same scale. We considered the following methods:

	Procedure	Limitation
Z-Scaling	Subtract the mean from each value and divide by standard deviations	- After Z-Scaling, data of that variables have mean of 0 and standard deviation of 1. - More robust in comparison to dividing by range - Outliers need to be identified and removed for better results
Quantile binning	- Transfer variable score into its quantile rankings. Divide each distribution into bins with equal population and then replace original value (score value) by the bin number. For instance, this way we can add the top 1% riskiest scores and obtain a new score. - Used for combining scores and model output - used for combining fraud scores	

First, we used Z-scaling for all the variables which contained no missing information. Than we decided to use Principal Component Analysis(PCA) to reduce the multivariate data to a smaller number of important

dements by creating new coordinate system. PCA finds the direction of maximum variance in cloud of data which creates our first axis called Principle Component 1. PC1 is linear combination of original variables that explains maximum variance in the original data. Then, PAC rotates the axis in a way that PC1 becomes the new x-axis. Furthermore, the next biggest variance creates the second axis called PC2 which would be orthogonal to PC1. We started with original data and transformed them into Principal Component space. We ranked them in order of the power they have in explaining variability. Then, we decided to keep 7 Principal Component since these PCs were able to explain more than 85.636 % of the variability in our data. This also allowed us to reduce dimensionality. We run R's "prcomp()" to perform PCA, using the option "center=TRUE", 'scale = TRUE' to standardize the original 45 variables.

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     3.9336 2.5604 2.3166 2.2221 1.52572 1.42444 1.35885 1.09173 0.99443 0.85397 0.80861 0.7881 0.74836 0.71291
Proportion of Variance 0.3438 0.1457 0.1193 0.1097 0.05173 0.04509 0.04103 0.02649 0.02198 0.01621 0.01453 0.0138 0.01245 0.01129
Cumulative Proportion  0.3438 0.4895 0.6088 0.7185 0.77024 0.81533 0.85636 0.88285 0.90482 0.92103 0.93556 0.9494 0.96181 0.97310
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21    PC22    PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     0.50068 0.48902 0.35819 0.3420 0.29343 0.27253 0.27021 0.25778 0.21281 0.17957 0.17078 0.14062 0.13198 0.09589
Proportion of Variance 0.00557 0.00531 0.00285 0.0026 0.00191 0.00165 0.00162 0.00148 0.00101 0.00072 0.00065 0.00044 0.00039 0.00020
Cumulative Proportion  0.97867 0.98399 0.98684 0.9894 0.99135 0.99300 0.99463 0.99610 0.99711 0.99782 0.99847 0.99891 0.99930 0.99950
                          PC29    PC30    PC31    PC32    PC33    PC34    PC35    PC36    PC37    PC38    PC39    PC40    PC41
Standard deviation     0.09432 0.07935 0.04845 0.04000 0.03556 0.03215 0.02342 0.01165 0.01116 0.008238 0.004056 0.0009201 9.063e-14
Proportion of Variance 0.00020 0.00014 0.00005 0.00004 0.00003 0.00002 0.00001 0.00000 0.00000 0.000000 0.000000 0.0000000 0.000e+00
Cumulative Proportion  0.99970 0.99984 0.99989 0.99993 0.99996 0.99998 0.99999 1.00000 1.00000 1.000000 1.000000 1.0000000 1.000e+00
                            PC42       PC43       PC44       PC45
Standard deviation     3.07e-14 2.452e-14 2.21e-14 1.262e-14
Proportion of Variance 0.00e+00 0.000e+00 0.00e+00 0.000e+00
Cumulative Proportion  1.00e+00 1.000e+00 1.00e+00 1.000e+00
```

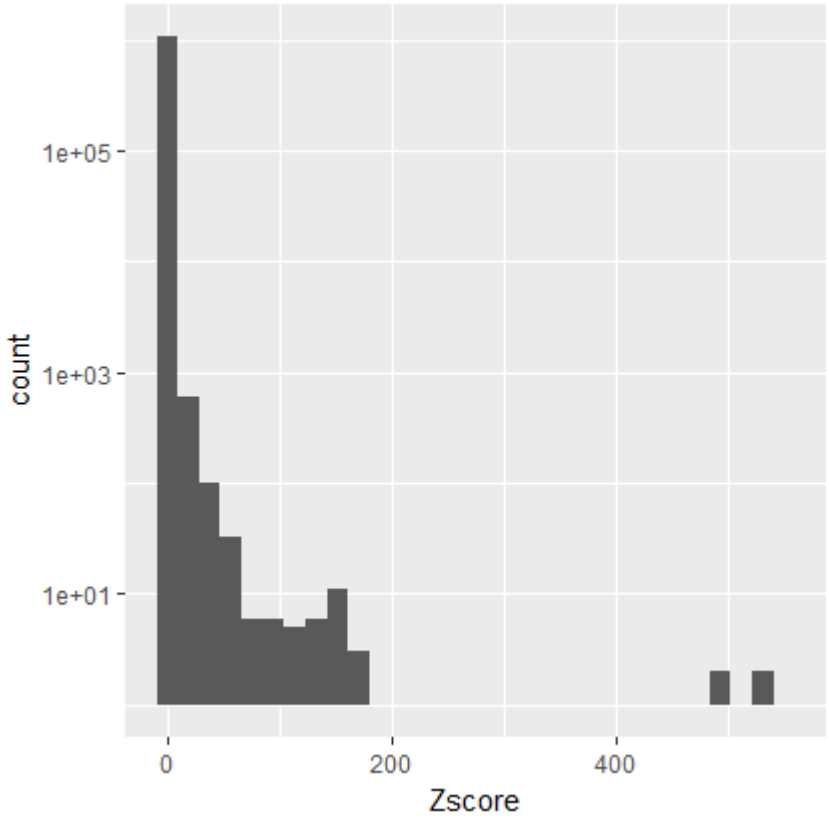


Finally, we Z-scaled each PC variables one more time to reduce the dimension of original data and prepared the final dataset which used to calculate the fraud scores.

3. Fraud Algorithms

3.1 Outlier detection via z-scores

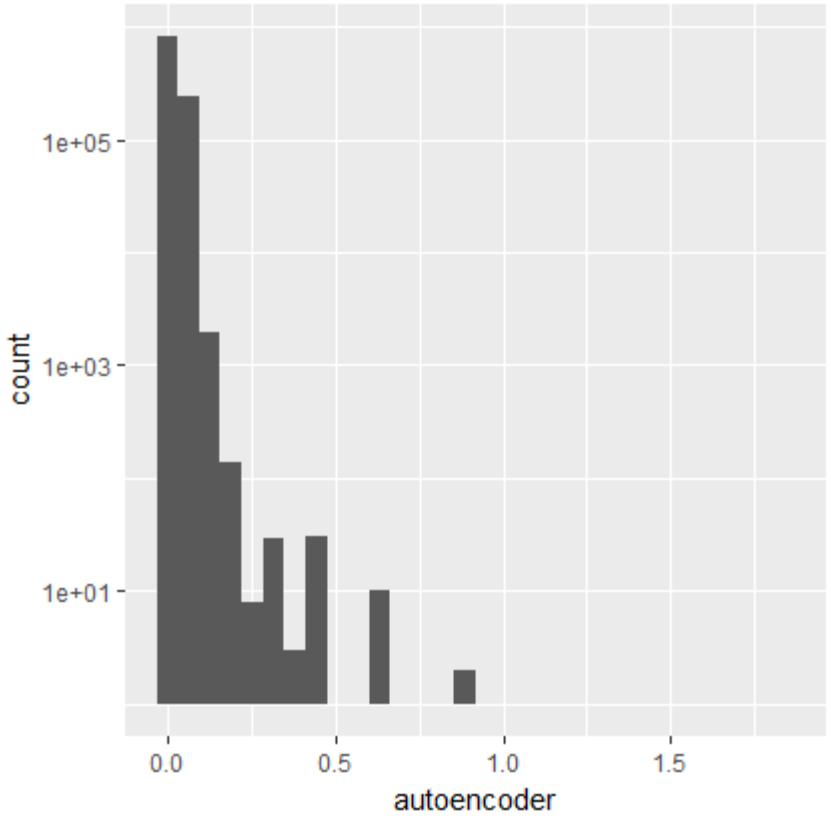
We used the following equation: $S = (\sum_i (|z_i|)^2)^{\frac{1}{2}}$. This equation sums all the variations along each record and it is a good indication of which record deviate more compared to the other records. The histogram of the score after taking the log of count is shown below, which is right skewed:



3.2 Autoencoder:

We used 'h_2o' package in R to perform autoencoder. The neural network has three layers: an input layer, a hidden (encoding) layer, and a decoding layer, where we compressed data to have only 2 dimensions. H_2o. anomaly automatically returns MSE of original data to reconstructed data, which means we used n equals 2 also in the formula: $S = \left(\sum_i |z_i - z_i'|^2\right)^{\frac{1}{2}}$. To calculate score, we sum up the error and took the square root of the sum.

The histogram of the score after taking the log of count is shown below, which is also right skewed:



4. Results

Because the scale of Z-score and Autoencoder-score is different, we cannot compare them directly, instead we rank the scores in each method, one record one rank and took the average. So, the top 10 records which may be fraud are:

Record	Z-score	autoencoder	rank1	rank2	Final score
78804	551.1552152	1.466703692	1048575	1048574	1048574.5
5393	533.6358824	1.243302707	1048573	1048572	1048572.5
22921	468.6588357	1.842785808	1048570	1048575	1048572.5
294061	491.0733563	1.34703791	1048571	1048573	1048572
1046264	540.0646942	0.929608333	1048574	1048570	1048572
376243	495.5440548	1.216999899	1048572	1048571	1048571.5
508286	172.0987928	0.875899953	1048567	1048569	1048568
246251	170.8989952	0.861184832	1048566	1048568	1048567
81046	152.9894223	0.648322185	1048563	1048567	1048565
447396	188.5277558	0.519711006	1048568	1048557	1048562.5

After looking at more detail regarding each entry, we eliminated 3 of them from our list and added the next 3 high score to be included in our ranking:

Record	Z-score	autoencoder	rank1	rank2	Final score
78804	551.1552152	1.466703692	1048575	1048574	1048574.5
5393	533.6358824	1.243302707	1048573	1048572	1048572.5
1046264	540.0646942	0.929608333	1048574	1048570	1048572
376243	495.5440548	1.216999899	1048572	1048571	1048571.5
508286	172.0987928	0.875899953	1048567	1048569	1048568
246251	170.8989952	0.861184832	1048566	1048568	1048567
447396	188.5277558	0.519711006	1048568	1048557	1048562.5
915072	152.4002017	0.639833534	1048561	1048562	1048561.5
946615	152.3800832	0.640620394	1048557	1048565	1048561
456015	152.5919978	0.627004228	1048562	1048559	1048560.5

Also, we divide the total records into 1000 bins after rank. The top 0.1% scores all get score equals 1000. And then we took the average of the two scores and get the final quantile scores. 348 records get the final score equals 1000. Sample Data is shown below:

Record	Zscore	Autoencoder	Rank1	Rank2	Finalscore
1	0.907941	0.0135296	93	496	294.5
2	0.812773	0.01651966	12	629	320.5
3	6.344618	0.05564243	978	909	943.5
4	6.054107	0.0569655	974	915	944.5
5	5.514393	0.08388837	970	997	983.5
6	1.354088	0.01062257	462	95	278.5

4.1 Top 10 Record Based on Final Score Ranking:

1) Record 1250253- WILLIAM R. BRENNAN (202-30 ROCKAWAY POINT BLVD)

William R. Brennan is associated with Cedar Brook Country Club, Inc. The Property has a tax class 1D (Bungalow Colonies), however the picture shows an empty lot. Further investigation is needed to determine appropriate tax codes.

Furthermore, it appears that majority of the property is empty land. However, "AVLAND" is a lot less than "FULLVAL". Also, there are no values for "BLDFRONT" and "BLDDEPTH" in order to come logical conclusion regarding foul play possibilities. We recommend further investigation.

2) Record 78804- U S GOVERNMENT OWNRD (FLATBUSH AVENUE)

This record is missing zip code. This makes it very difficult to find the exact location of the property. Flatbush Ave is a big area and it covers many properties. Further investigation is needed to make better judgement about this record.

Moreover, this record has a small "LTFRONT" and "LTDEPTH", however, the "FULLVAL" is unusually large. We also missing information regarding "STORIES", "BLDFORNT", and "BLDDEPTH" which makes it more complicated to determine the reasoning behind this discrepancy. Further investigation is recommended on this record.

3) Record 5393- 864163 REALTY, LLC (86-55 BROADWAY)

"The ELM EAST" is trademark of 864163 Realty, LLC. This is an apartment complex and has correct tax code assigned. The recorded values seem out of place. The value "BLDFRONT" and "BLDDEPTH" are both 1" which is to be expected given the fact it is an apartment complex. However, "LTDEPTH" and "LTFRONT" are also very low number. Furthermore, the value of this apartment complex is very low, $2.93million. Given the size of this property and its luxury status, the market value is way higher than $2.9mil.

4) Record 37624-Logan Property, Inc (138-68 Brookville Blvd)

By simply searching Logan Property Inc on web, you come across few companies which none operate in Queens or New York State. The tax class indicate "4" which cover industrial and commercial buildings. By searching the address, the result shows a residential house. Houses in that neighborhoods are from $450k-$800k based on Redfin search. However, the "FULLVAL" of the property is recorded as $374million. Furthermore, the value of "AVLAND" is far greater than the value of "FULLVAL". This means that the value of the land is greater than the value of entire property, which is odd. The record is also missing values for "LTDEPTH", "BLDFRONT" "BLDDEPTH" which limits our investigation.

5) Record 447396- DEPT OF GENERAL SERVI (FLATBUSH AVE)

This could be a government organization. However, we are missing the zip code for this address and we don't know the exact location to determine accuracy. It is important to note that "AVLAND" and "FULLVAL" are very large. The market value of the property is recorded as 2.3billion dollar. We recommend a further investigation to identify root cause of this anomaly.

6) Record 508286- BREEZY POINT COOPERAT (217-02 BREEZY POINT BLVD)

The tax class for this record indicates "1D", Bungalow Colonies. It however doesn't look like bungalows when looking at google earth photos. Moreover, the "FULLVAL" is about 24X the "AVLAND" value.

7) Record 1046264- Tony Chen (SHORE ROAD)

This record is missing a lot of fields. Due to absent of zip code, we won't be able to determine the location of this property. Furthermore, "TONY CHEN" is a very general name which makes it easy to blend within thousands of records. Finally, only 7 out of the 25 columns have value for "TONY CHEN" and we are missing a lot of information for this individual. We recommend investigation into this record.

8) Record 915072-RAM MAHENDRA (1403 WARING AVE)

There is a disconnect between "AVLAND" and "FULLVAL" value. For houses in this age range, the real value of property is in the value of the land. However, "FULLVAL" is 30X of "AVLAND".

9) Record 946615-S M R Holding CORP (15 8 Ave.)

There is a disconnect between "AVLAND" and "FULLVAL" value. This house was built in 1920 and the real value of property is in the value of the land. However, "FULLVAL" is 81X of "AVLAND".

10) Record 456015-FONG TSO LEE LINA WON (29 BAYSIDE DRIVE)

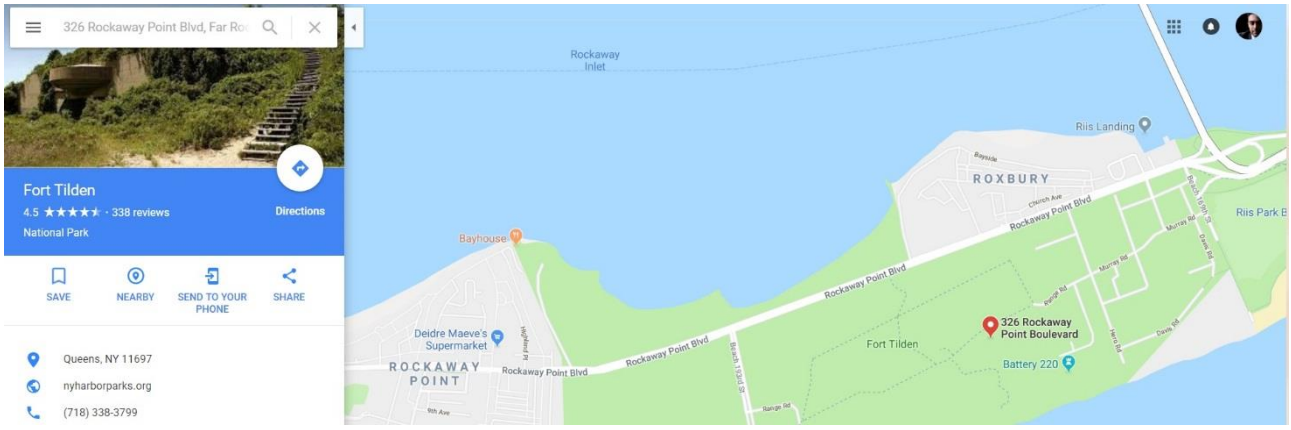
There is a disconnect between "AVLAND" and "FULLVAL" value. For houses in this age range, the real value of property is in the value of the land. However, "FULLVAL" is 37X of "AVLAND".

4.2 Removed Record from top 10 list:

By considering the top 10 results, we noticed few of the records belong to the government. In our analysis, our focus was identifying possible fraud against both local and federal government. Therefore, we will exclude the following from our top 10 list.

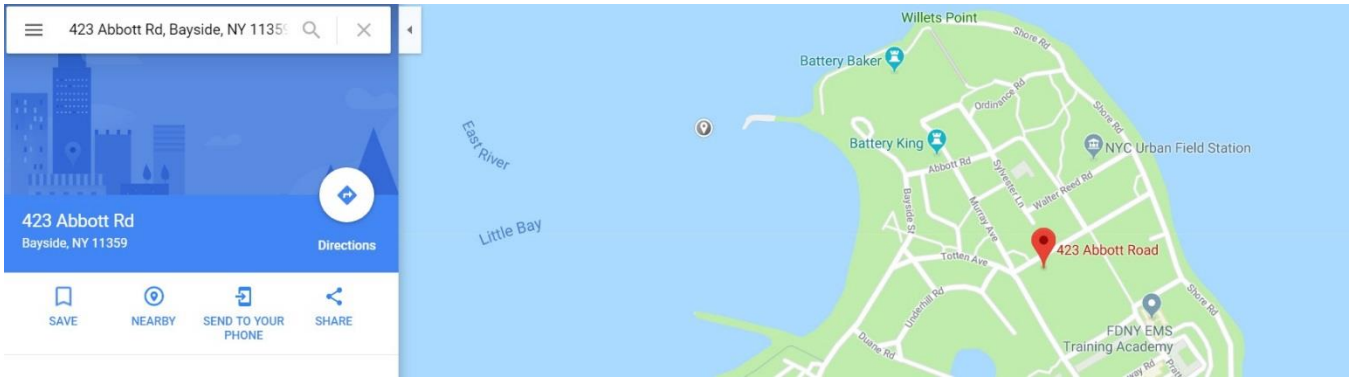
1) Cultural Affairs (1000 5th Ave)

2) DEPT RE-CITY OF NY (326 Rockaway Point Blvd)



This property is Fort Tilden which is part of "Harbor Defenses of Southern New York". The Harbor Defenses of New York is branch of U.S. Army Coast Artillery Corps.

3) The City of New York (423 ABBOT ROAD)



Fort Totten(Queens) is owned by New York City Department of Parks and Recreation

5. Appendix

5.1. Property Value Assessment:

Taxable Value: Actual or transactional asssed value (whichever is less) minus any exemptions.
Transactional Assessed Value: increases to your assessed value are phased in at 20% per year rate (Except for physical changes)
Market Value: Finance's estimate of your property's worth.
The amount used to calculate your property taxes. The formula for calculating Assessed Value is: Market Value X Level of Assessment = Assessed Value.
The % of market value used to calculate your property's assessed value. Also known as the assessment ratio. (Tax Class 1 - 6%, Tax Class 2, 3 and 4 - 45%)

5.2. Tax Class Definition

Property in NYC is divided into 4 classes:
Class 1: Most residential property of up to three units (family homes and small stores or offices with one or two apartments attached), and most condominiums that are not more than three stories.
Class 2: All other property that is not in Class 1 and is primarily residential (rentals, cooperatives and condominiums). Class 2 includes:
Sub-Class 2a (4 - 6 unit rental building);
Sub-Class 2b (7 - 10 unit rental building);
Sub-Class 2c (2 - 10 unit cooperative or condominium); and
Class 2 (11 units or more).
Class 3: Most utility property.
Class 4: All commercial and industrial properties, such as office, retail, factory buildings and all other properties not included in tax classes 1, 2 or 3.

5.3. Reference

Tax Class Definition
http://www1.nyc.gov/site/finance/taxes/property-tax-rates.page

5.4 Data Quality Report (DQR)

Basic Information

Summary: The City of New York Property Valuation and Assessment Data contains 30 fields, including property owners, sizes, values, lot sizes and taxes, etc.
Number of Records: 1048575
Time Period: November, 2010
Source: https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8

Summary Statistics

Numeric Variables:

Numeric Variable	% Populated	# Unique	Mean	Std	Min	1st Qua	Median	3rd Qua	Max
LTFRONT	100.00	1277	36.17	73.73	0.00	19.00	25.00	40.00	9999.00
LTDEPTH	100.00	1336	88.28	75.48	0.00	80.00	100.00	100.00	9999.00
STORIES	95.03	112	5.06	8.43	1.00	2.00	2.00	3.00	119.00
FULLVAL	100.00	108277	880487.70	11702930.00	0.00	303000.00	446000.00	619000.00	6150000000.00
AVLAND	100.00	70529	85995.03	4100755.00	0.00	9160.00	13646.00	19706.00	2668500000.00
AVTOT	100.00	112294	230758.20	6951206.00	0.00	18385.00	25339.00	46095.00	4668309000.00
EXLAND	100.00	33186	36811.79	4024330.00	0.00	0.00	1620.00	1620.00	2668500000.00
EXTOT	100.00	63805	92543.81	6578281.00	0.00	0.00	1620.00	2090.00	4668309000.00
BLDFRONT	100.00	610	23.02	35.79	0.00	15.00	20.00	24.00	7575.00
BLDDEPTH	100.00	620	40.07	43.04	0.00	26.00	39.00	51.00	9393.00
AVLAND2	26.80	58170	246365.50	6199390.00	3.00	5705.00	20059.00	62338.75	2371005000.00
AVTOT2	26.80	110891	716078.70	11690170.00	3.00	34013.50	80010.00	240792.00	4501180000.00
EXLAND2	8.27	21997	351802.20	10852480.00	1.00	2090.00	3053.00	31419.00	2371005000.00
EXTOT2	12.39	48107	658114.80	16129810.00	7.00	2889.00	37116.00	106629.00	4501180000.00

Categorical Variables:

Categorical Variable	% Populated	# Unique
RECORD	100	1048575
BBLE	100	1048575
BLOCK	100	13949
LOT	100	6366
EASEMENT	0.385571	13
OWNER	97.035691	847054
BLDGCL	100	200
TAXCLASS	100	11

Categorical Variable	% Populated	# Unique
EXCD1	59.379825	130
STADDR	99.938869	820638
ZIP	97.486494	197
EXMPTCL	1.42975	15
EXCD2	8.672818	61
PERIOD	100	1
YEAR	100	1
VALTYPE	100	1

Field 1: RECORD

1048575 unique values, ranging from 1 to 1048575. It serves as indexing numbers.

Field 2: BBLE

1048575 unique values. First ten records are:

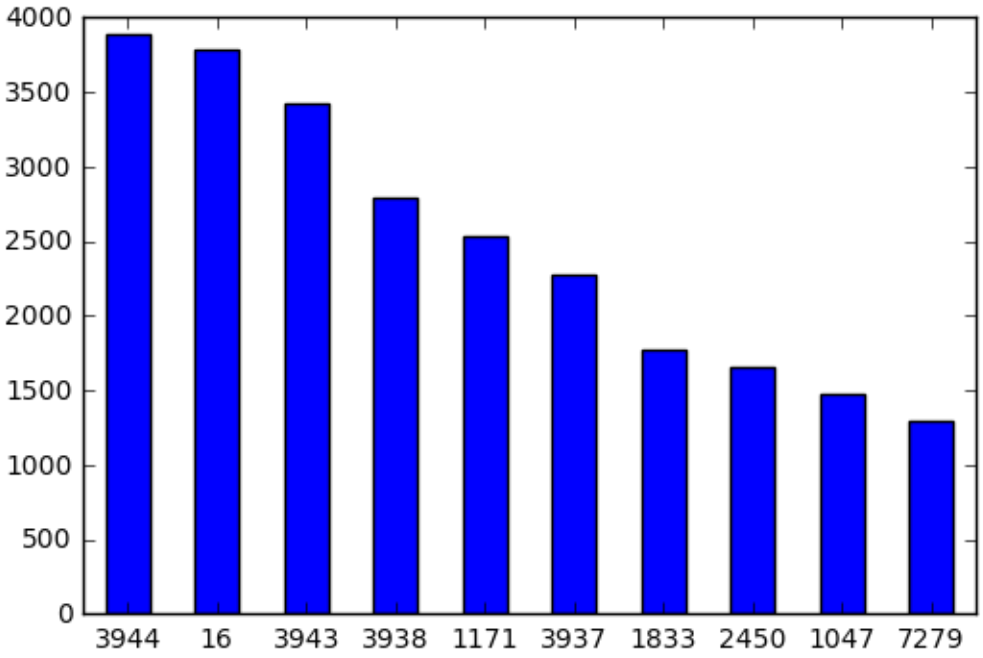
1	3046020035
2	5046820019
3	3074790028
4	4027980132
5	1006950027E
6	4031810007
7	4051861001
8	3082020064
9	4052570008
10	3070780050

It is the combination of BORO code (1 digit), BLOCK code (5 digit), LOT code (4 digit) and EASEMENT code (1 digit if exists)

Field 3: BLOCK

Distribution of Top 10 frequent block codes:

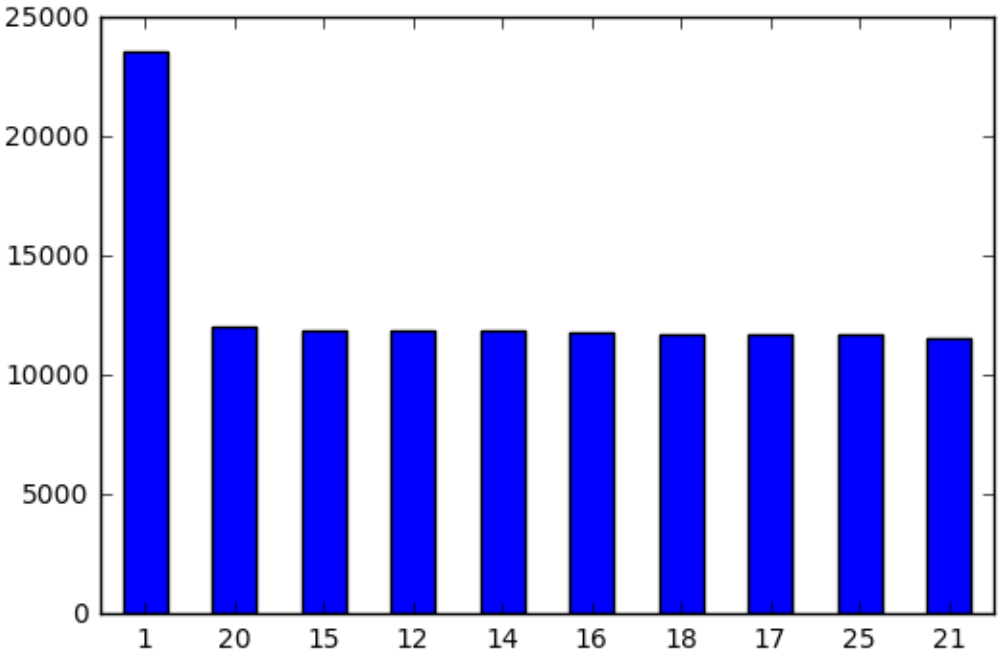
BLOCK	Count	Percentage(%)
3944	3888	0.37
16	3786	0.36
3943	3424	0.32
3938	2794	0.27
1171	2535	0.24
3937	2275	0.21
1833	1774	0.17
2450	1651	0.16
1047	1480	0.14
7279	1302	0.12



Field 4: LOT

Distribution of Top 10 frequent lot codes:

LOT	Count	Percentage(%)
1	23570	2.25
20	12045	1.15
15	11904	1.14
12	11894	1.13
14	11864	1.13
16	11810	1.13
18	11763	1.12
17	11728	0.12
25	11692	1.12
21	11593	1.11



Field 5: EASEMENT

Distribution of 13 easement codes:

EASEMENT	Count	Percentage(%)
E	3603	0.34
F	265	0.03
G	95	0.01
H	30	0.00
N	14	0.00
I	7	0.00
J	4	0.00
K	3	0.00
L	2	0.00
M	17	0.00
P	2	0.00
U	1	0.00
NULL	1044532	99.61

Most of the properties does not have easements.

Field 6: OWNER

Distribution of Top 10 frequent owners including missing values:

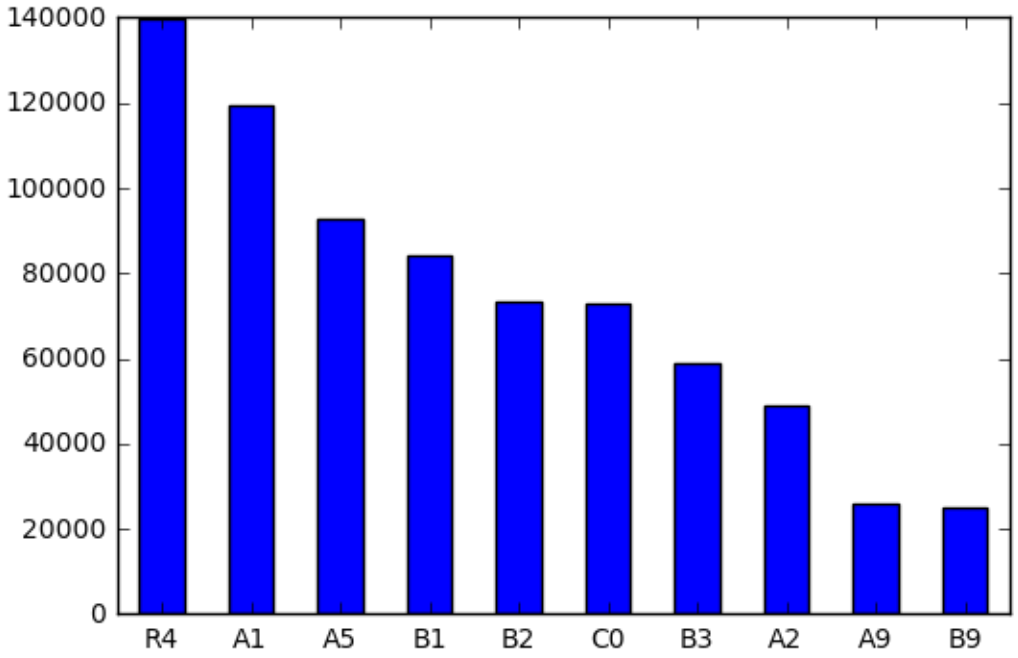
OWNER	Count	Percentage(%)
NULL	31081	2.96
PARKCHESTER PRESERVAT	6021	0.57
PARKS AND RECREATION	3358	0.32
DCAS	2053	0.20
HOUSING PRESERVATION	1900	0.18
CITY OF NEW YORK	1189	0.11
NEW YORK CITY HOUSING	1014	0.10
BOARD OF EDUCATION	1003	0.10
CNY/NYCTA	975	0.09
NYC HOUSING PARTNERSH	747	0.07

Most of the records having missing or unknown owners.

Field 7: BLDGCL

Distribution of Top 10 building classes:

BLDGCL	Count	Percentage(%)
R4	139879	13.3
A1	119340	11.4
A5	92896	8.9
B1	84054	8
B2	73156	7
C0	72077	8
B3	59091	5.6
A2	49085	4.7
A9	25931	2.5
B9	25235	2.4



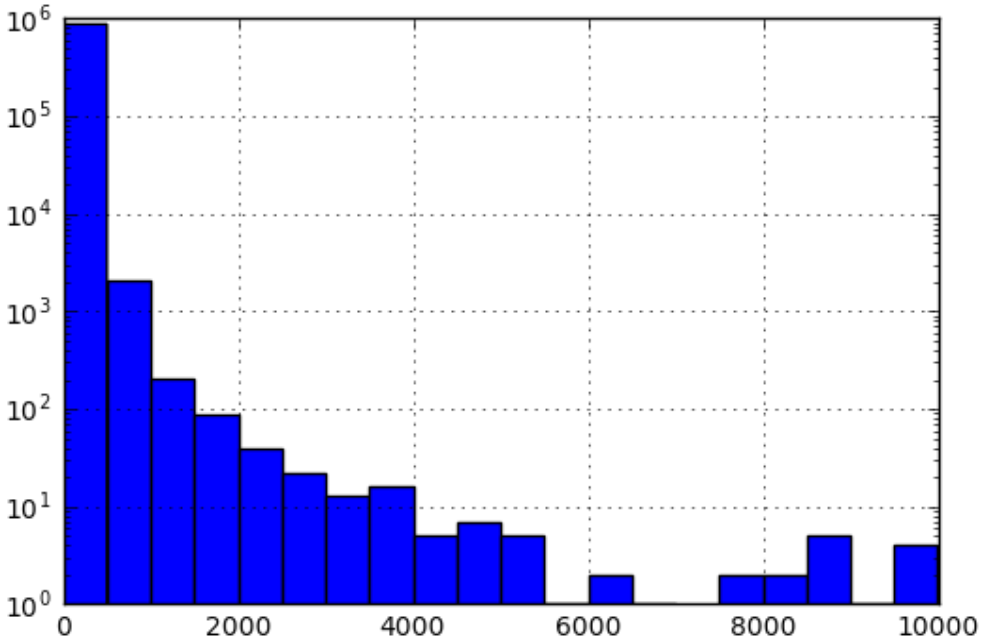
Field 8: TAXCLASS

Distribution of 11 tax classes:

TAXCLASS	Count	Percentage(%)
1	642774	61.4
2	188592	18.0
4	102281	9.8
2A	40558	3.9
1B	22193	2.1
1A	20899	2.0
2B	13962	1.3
2C	10795	1.0
3	4546	0.4
1C	946	0.1
1D	29	0.0

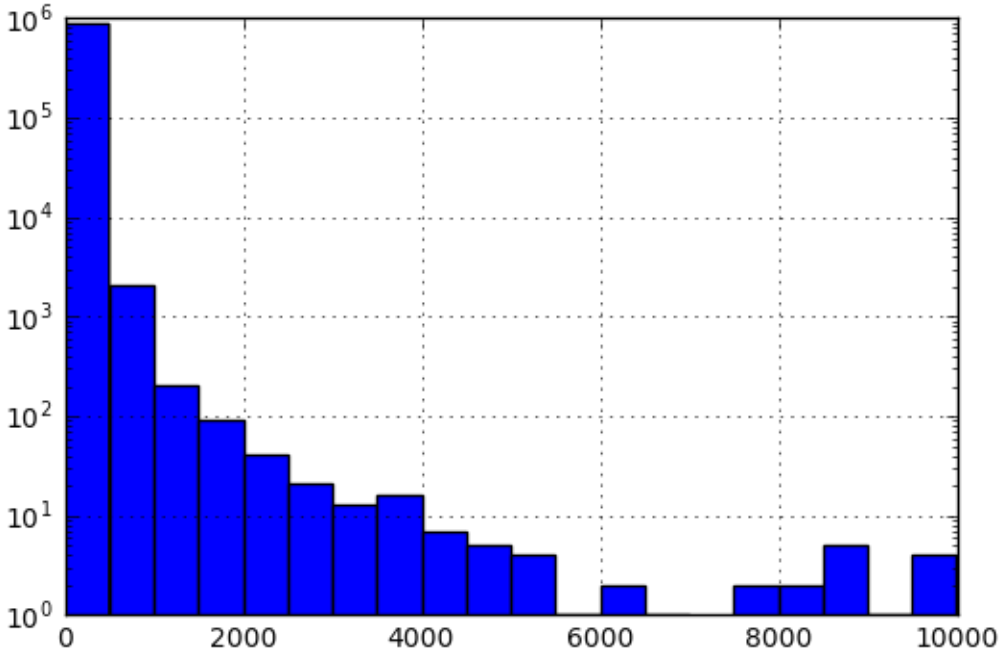
Field 9: LTFRONT

Distribution of **log** of lot front after deleting zero values:



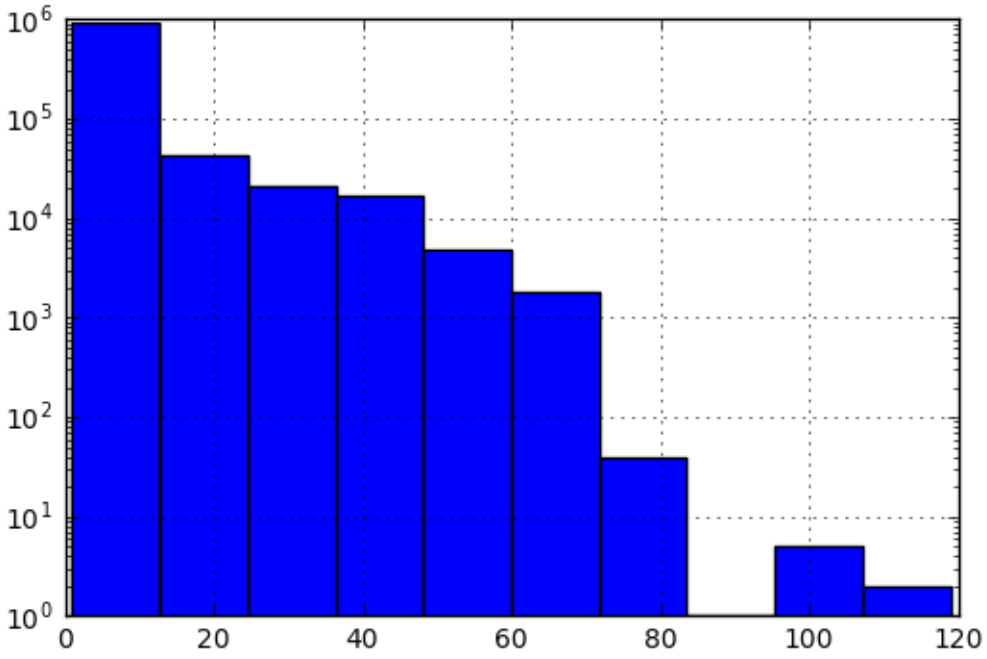
Field 10: LTDEPTH

Distribution of **log** of lot depth after deleting zero values:



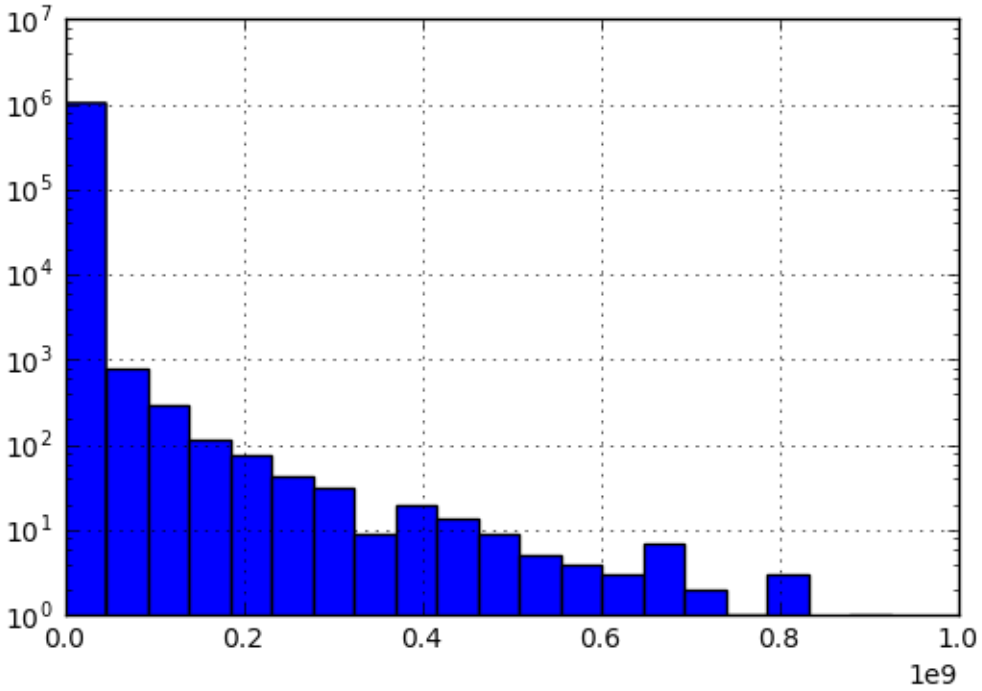
Field 11: STORIES

Distribution of **log** of stories:



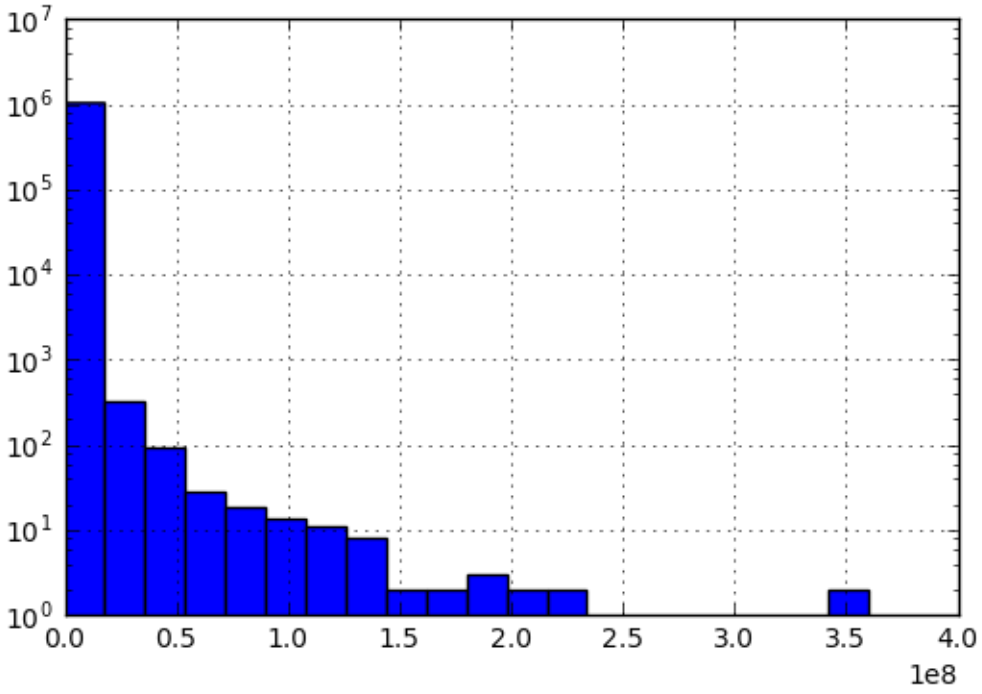
Field 12: FULLVAL

Distribution of **log** of full value whose data range is (0, 1e9):



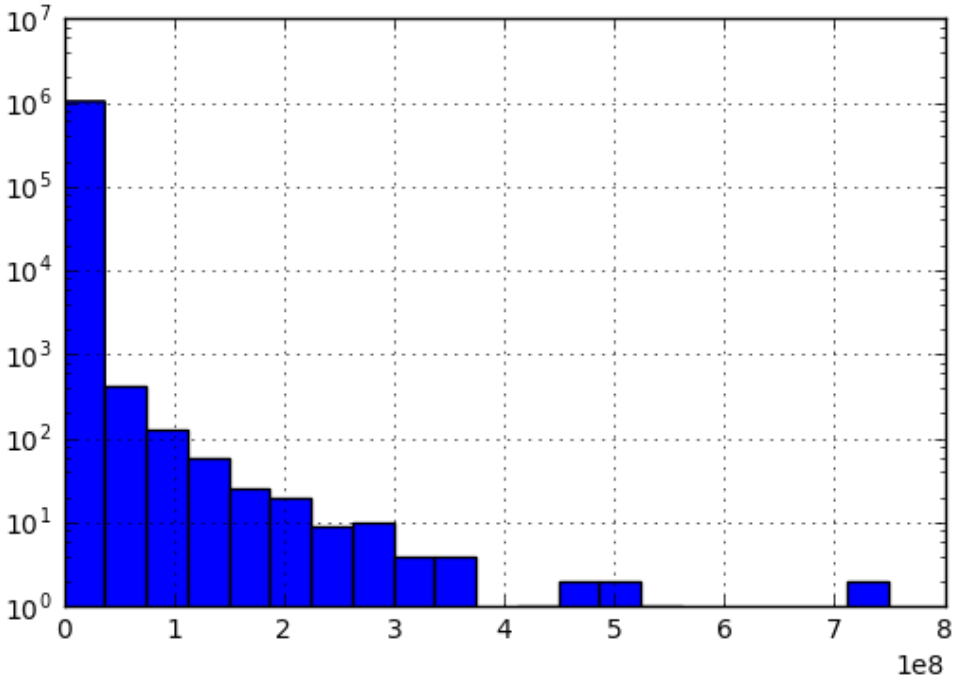
Field 13: AVLAND

Distribution of **log** of assessed land value whose data range is (0, 5e8):



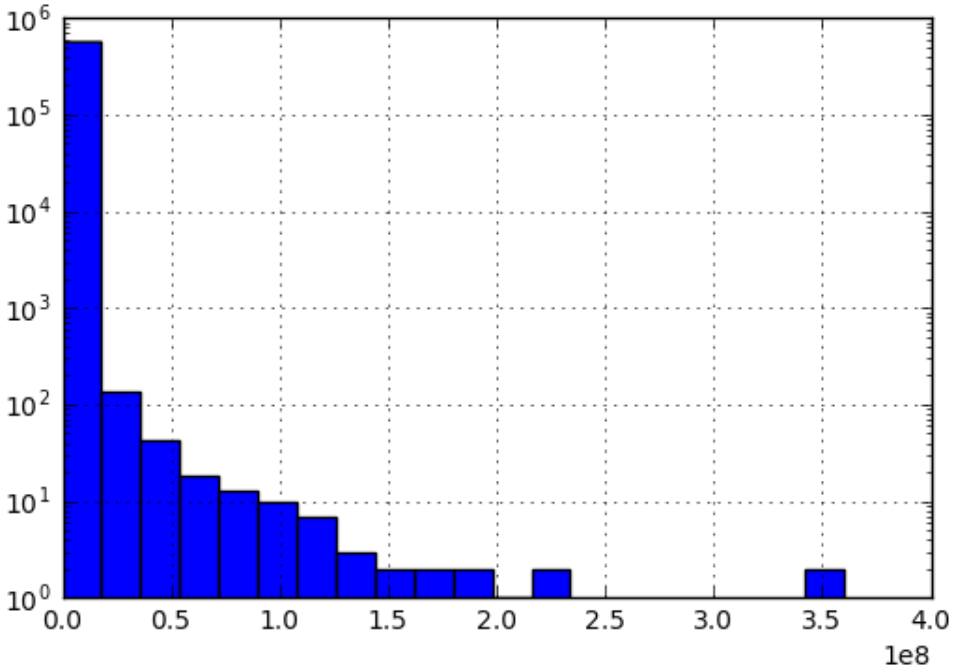
Field 14: AVTOT

Distribution of **log** of assessed total value whose data range is (0, 1e9):



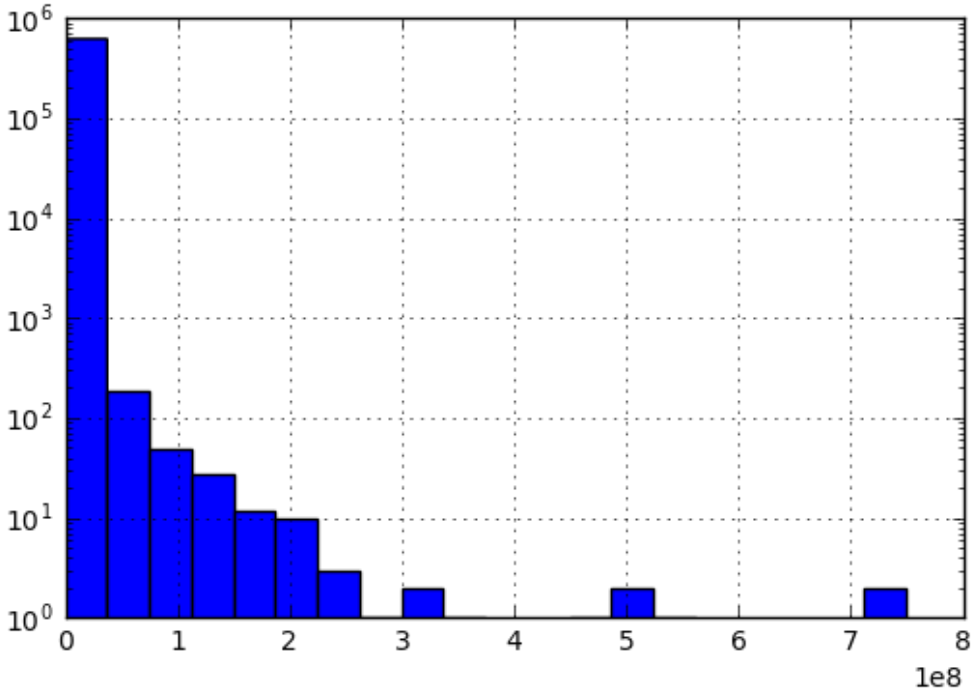
Field 15: EXLAND

Distribution of **log** of exempt land value whose data range is (0, 5e8):



Field 16: EXTOT

Distribution of **log** of exempt total value whose data range is (0, 1e9):



Field 17: EXCD1

Distribution of Top 10 frequent exempt class 1:

EXCD1	Count	Percentage(%)
1017	414222	39.5
1010	48322	4.6
1015	30849	2.9
5113	23842	2.3
1920	17594	1.7
5110	16834	1.6
5114	14984	1.4
5111	9165	8.7
1021	6567	6.3
1986	4212	4.0

Field 18: STADDR

Distribution of Top 10 frequent street address including missing values:

STADDR	Count	Percentage(%)
501 SURF AVENUE	902	0.09
330 EAST 38 STREET	817	0.08
322 WEST 57 STREET	720	0.07
155 WEST 68 STREET	671	0.06
20 WEST 64 STREET	657	0.06
1 IRVING PLACE	650	0.06
NULL	641	0.06
220 RIVERSIDE BOULEVARD	628	0.06
360 FURMAN STREET	599	0.06
200 EAST 66 STREET	585	0.06

Field 19: ZIP

Distribution of Top 10 frequent zip codes:

ZIP	Count	Percentage(%)
10314	24605	2.3
11234	20001	1.9
10462	16905	1.6
10306	16576	1.6
11236	15678	1.5
11385	14921	1.4
11229	12793	1.2
11211	12710	1.2

10312	12634	1.2
11207	12293	1.2

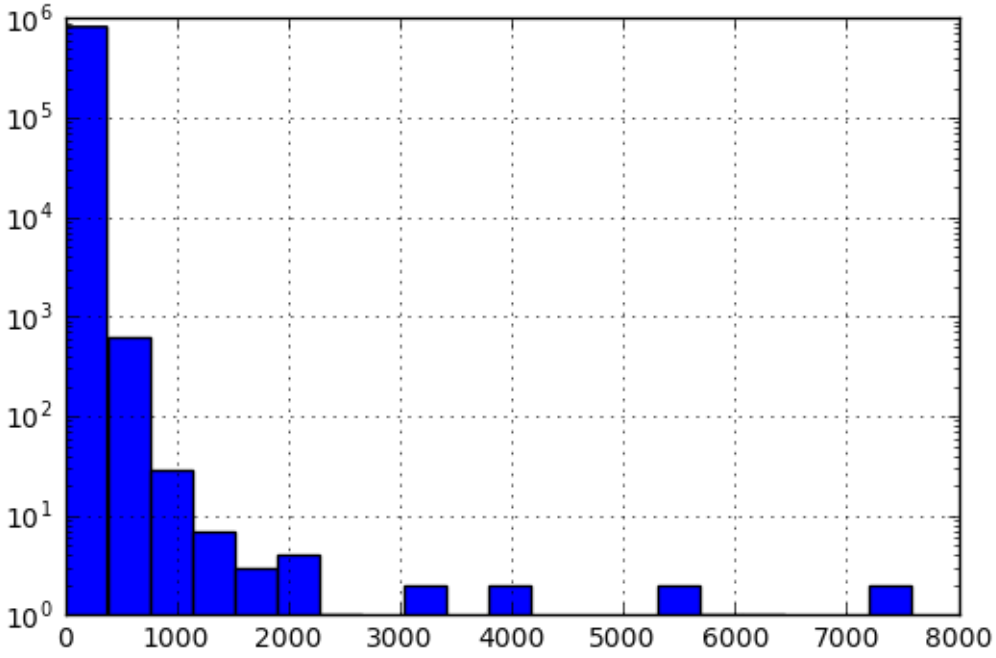
Field 20: EXMPTCL

Distribution of all exempt classes:

EXMPTCL	Count	Percentage(%)
X1	6494	43.41
X5	5158	34.41
X7	818	5.46
X6	76	5.07
X2	665	4.44
X4	438	2.92
X8	289	1.93
X3	260	1.73
X9	105	0.70
KI	1	0.01
5	1	0.01
A9	1	0.01
Vi	1	0.01
R4	1	0.01

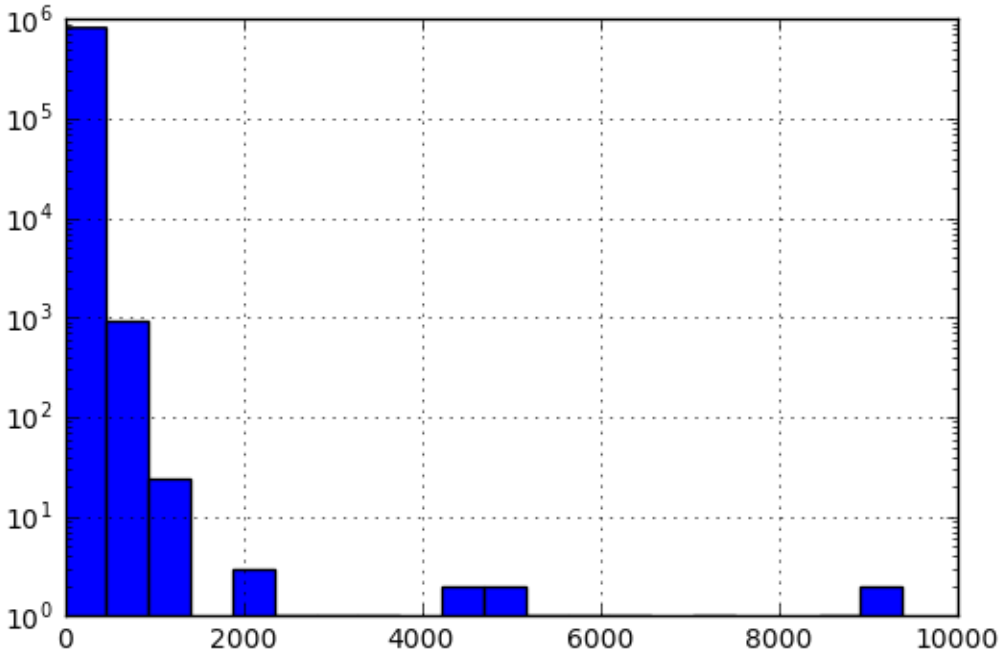
Field 21: BLDFRONT

Distribution of **log** of building front after deleting zero values:



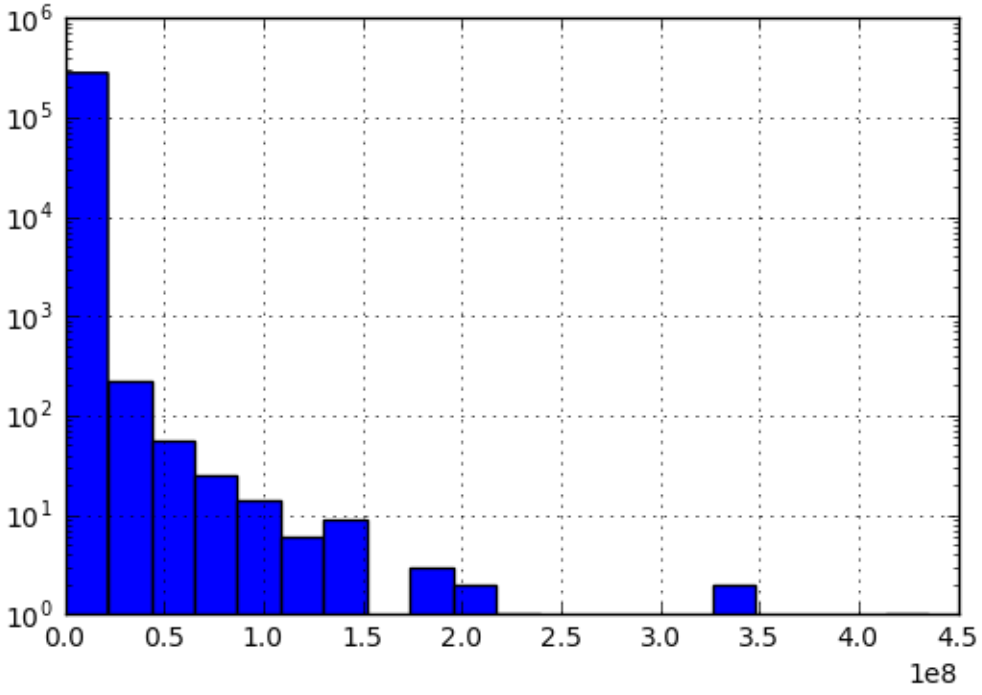
Field 22: BLDDEPTH

Distribution of **log** of building depth after deleting zero values:



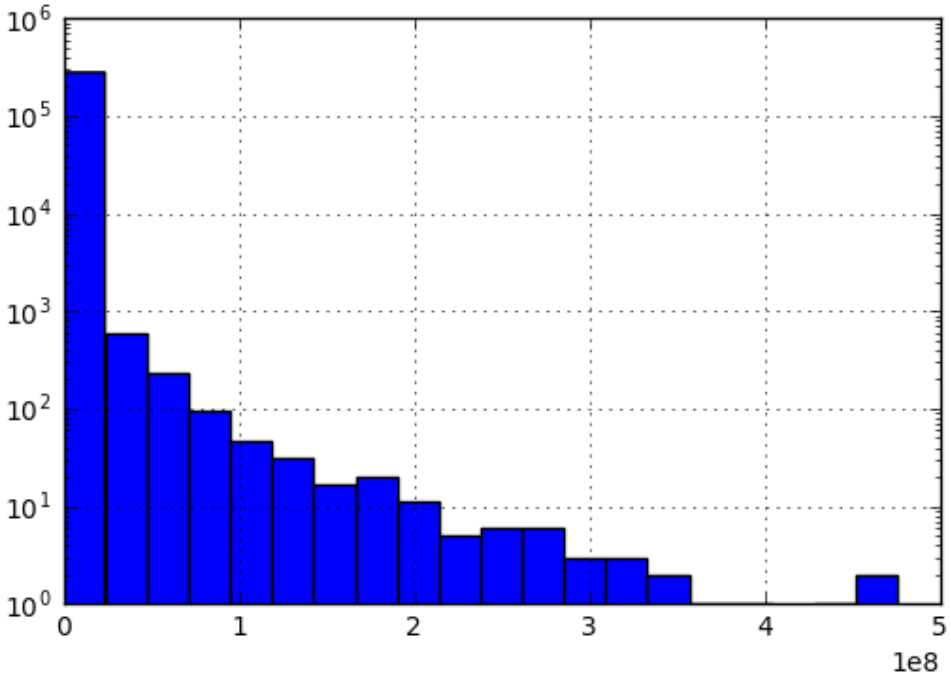
Field 23: AVLAND2

Distribution of **log** of assessed land value 2 whose data range is (0, 5e8):



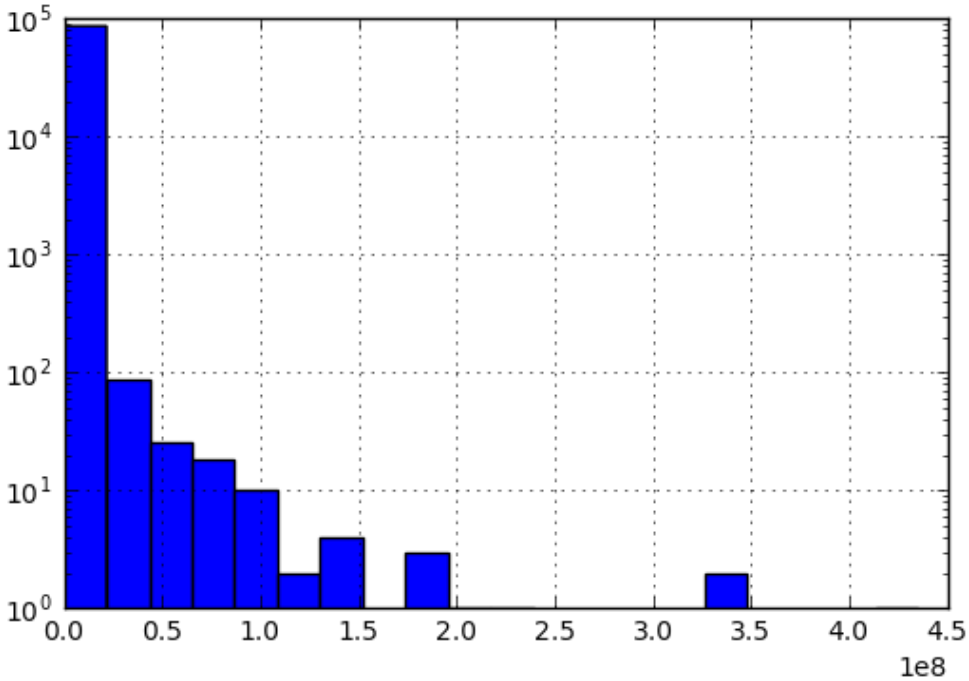
Field 24: AVTOT2

Distribution of **log** of assessed total value 2 whose data range is (0, 5e8):



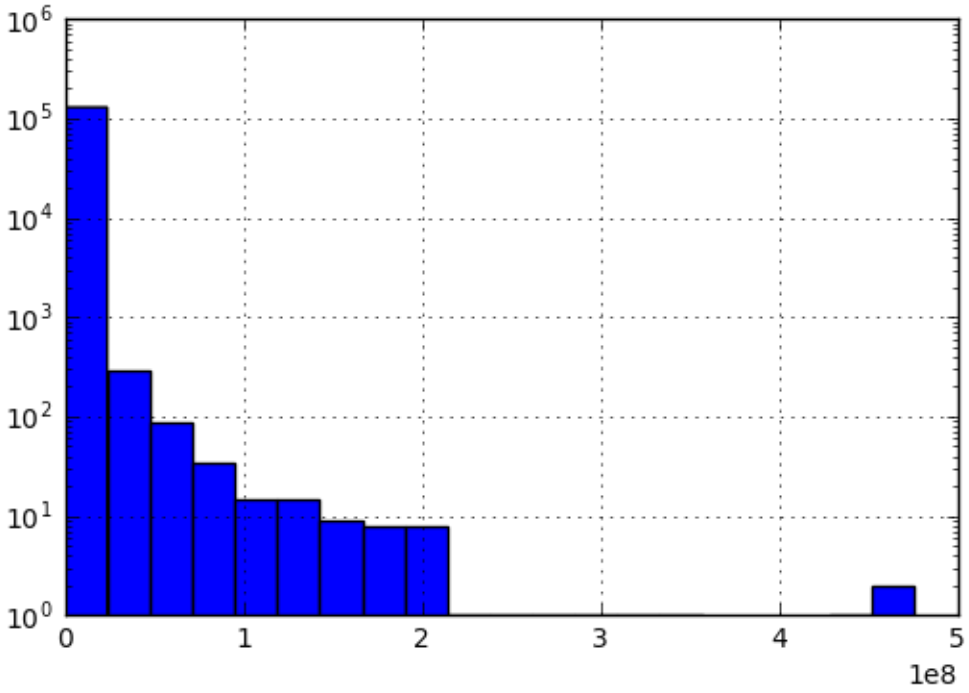
Field 25: EXLAND2

Distribution of **log** of exempt land value 2 whose data range is (0, 5e8):



Field 26: EXTOT2

Distribution of **log** of exempt total value 2 whose data range is (0, 5e8):



Field 27: EXCD2

Distribution of Top 10 frequent exempt class 2:

EXCD2	Count	Percentage(%)
1017	64223	70.6
1015	12038	13.2
5112	6867	7.6
1019	3034	3.3
1920	2961	3.3
1200	875	1.0
1101	493	0.5
5129	227	0.2
1986	34	0.0
1022	31	0.0

Field 28: PERIOD

Same for all records: “FINAL”

Field 29: YEAR
Same for all records: “2010/11”

Field 30: VALTYPE
Same for all records: “AC-TR”