**Detecting Fraud in the NY Property Dataset**

*DSO 562 Project 1 Report*

***Team 4***

Chong Li

Jie Chen

Raman Deep Singh

Xiaowen Zhang

Yu Dong

February 22, 2017

**Glossary**

Executive summary ……………………………………………………………………………………. 2

Part I. Data Overview ………………………………………………………………………………….. 3

Part II. Data Cleaning ………………………………………………………………………………… 12

Part III. Variable Construction ……………………………………………………………………….. 14

Part IV. Fraud Algorithm ……………………………………………………………………………… 16

Part V. Results ………………………………………………………………………………………….18

Appendix ……………………………………………………………………………………………….. 24

**Executive Summary**

This report provides an analysis and evaluation of The City of New York Property Valuation and Assessment Data for detecting fraud using unsupervised machine learning methods. The tools used are R and Tableau, and methods for analysis include Principal Component Analysis and Autoencoder.

The original data set contains records of more than 1 million properties across the city of New York and information on their sizes, values, owner, building classes, tax classes, etc. The general process of analysis follows data cleaning, building expert variables, standardization and dimensionality reduction, applying fraud algorithm, calculating fraud score, and identifying potential fraud.

Using heuristic fraud algorithm and Autoencoder, a fraud score is calculated for each of the one million properties. Records with high scores are determined to be potentially fraudulent. The report further finds the high score records of two unsupervised machine learning methods partially overlapped, and determines that the overlapped part of the top records from both methods are very likely to be 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. 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.

**Part I. Data Overview**

The City of New York Property Valuation and Assessment Data file is a public available dataset posted by the Department of Finance on the City of New York Open Data website. The dataset contains records of more than 1 million properties across the city of New York and information on their sizes, values, owner, building classes, tax classes, etc. The dataset contains a total of 1,048,575 records (rows) and 30 variables (columns). Among the 30 variables, there are 13 categorical variables, 14 numeric variables, 2 text variables, and 1 date variable. All of the records are taken from November 2011.

Following is description of the variables we consider to be the most important. The complete Data Quality Report can be found in appendix.

Variable Name: **RECORD**

RECORD is a categorical variable. It works as the ordinal reference number for each property Record. It has 1,048,575 unique values, ranging from 1 to 1,048,575. No repeated values or missing values exist.

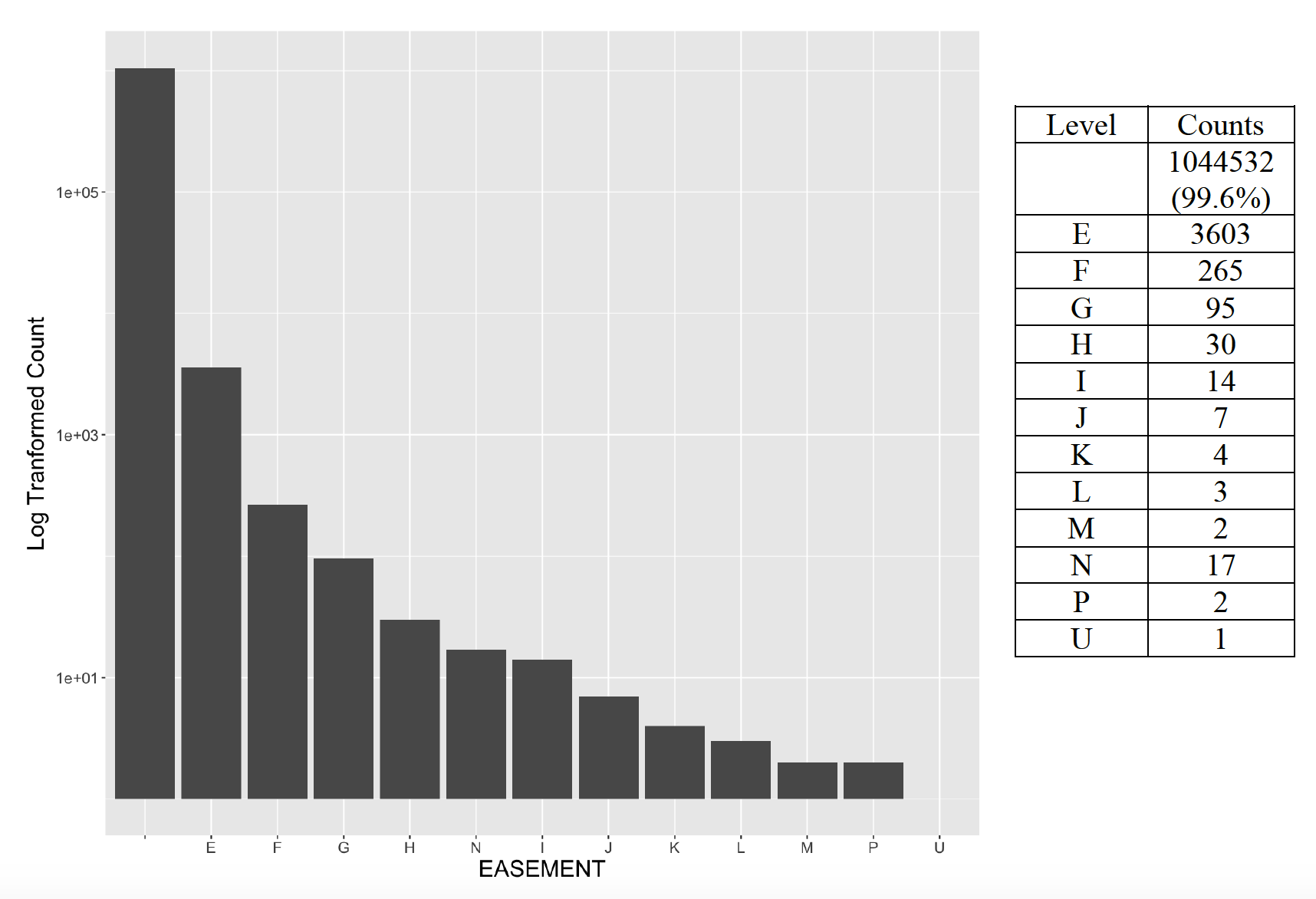
Variable Name: **BBLE**

BBLE is a nominal categorical variable with 10 or 11 digits. It is the concatenation of BORO code (1 digit), BLOCK code (5 digit), LOT code (4 digit) and EASEMENT code (1 digit if

exists). It has 1,048,575 unique values, with no repeated or missing values.

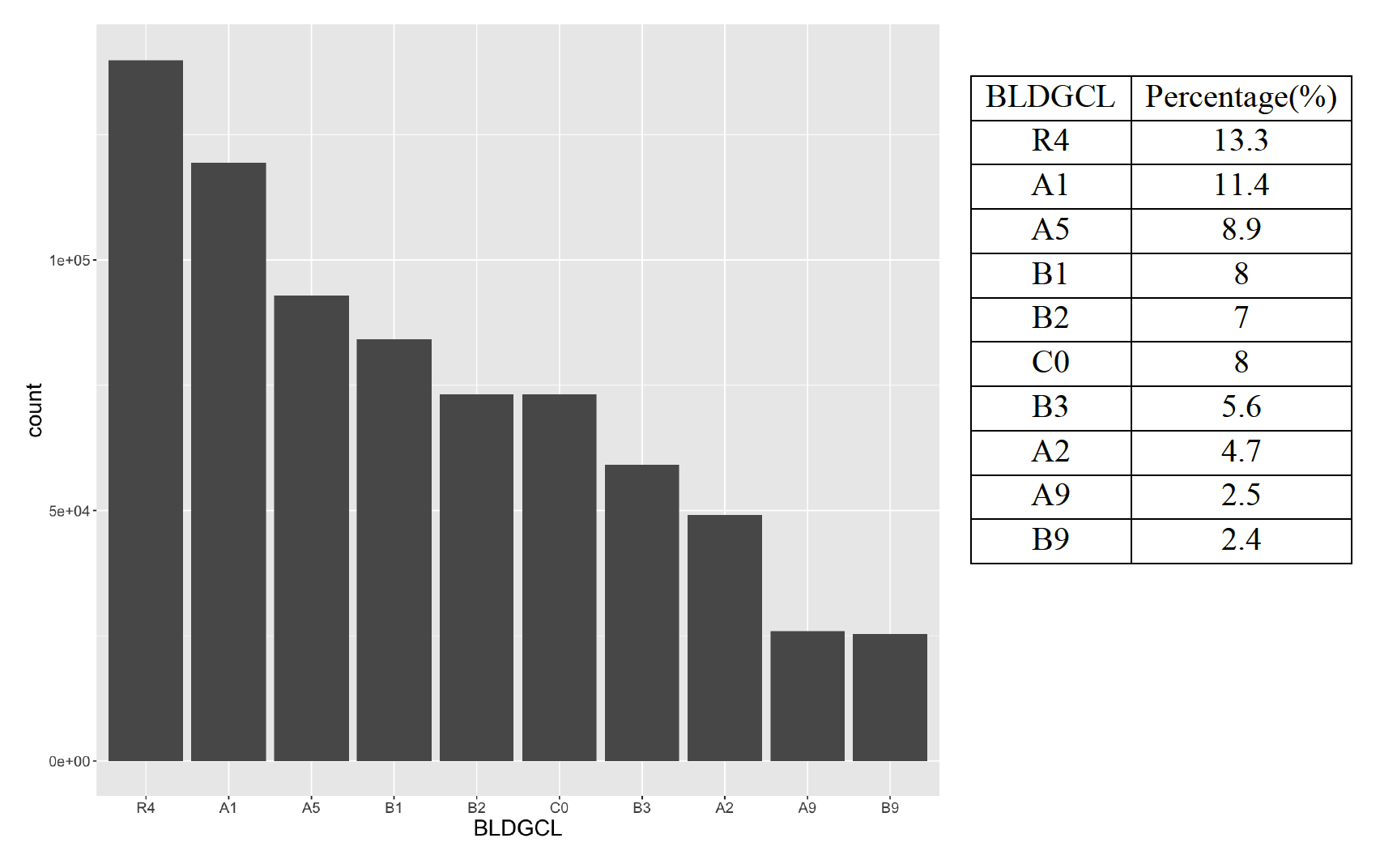
Variable Name: **EASEMENT**

EASEMENT is a nominal categorical variable representing the property’s easement type. It has 13 levels – “”, “E”, “F”, “G”, “H”, “I”, “J”, “K”, “L”, “M”, “N”, “P”, “U”. The null value indicates the property does not have any special easement type. No missing values exist. The sorted bar chart with log transformed y axis is shown below:



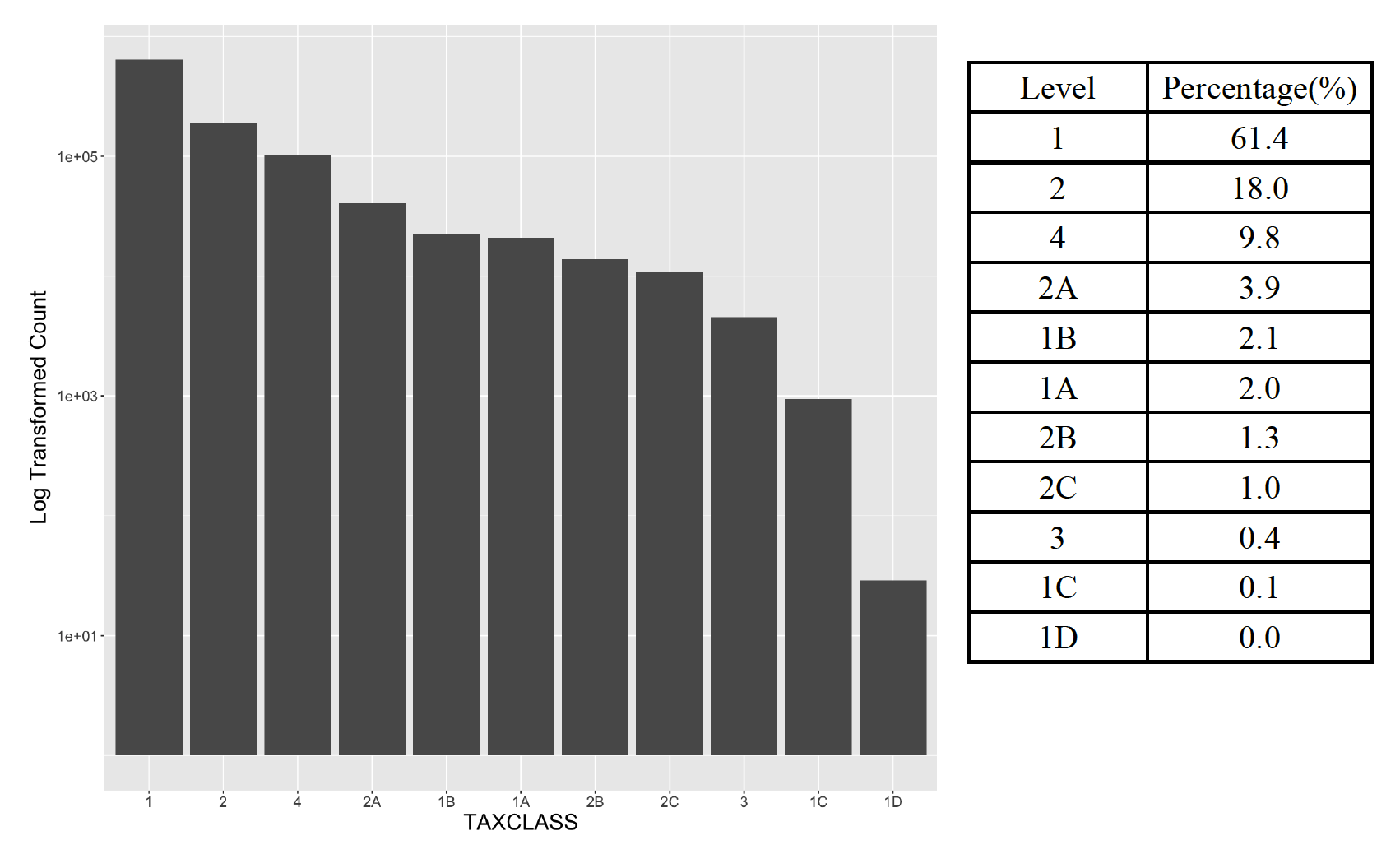
Variable Name: **BLDGCL**

BLDGCL is a nominal categorical variable indicating the building class. It has 200 unique levels. Each level has 2 digits – the first digit is a character from A to Z, the second digit is a number from 0 to 9. No missing values exist. The top 10 most frequently occurred BLDGCL is shown below:



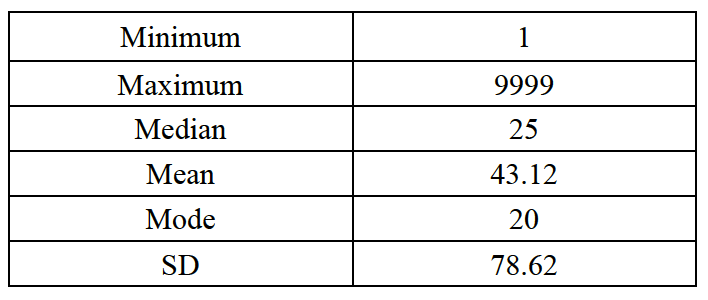
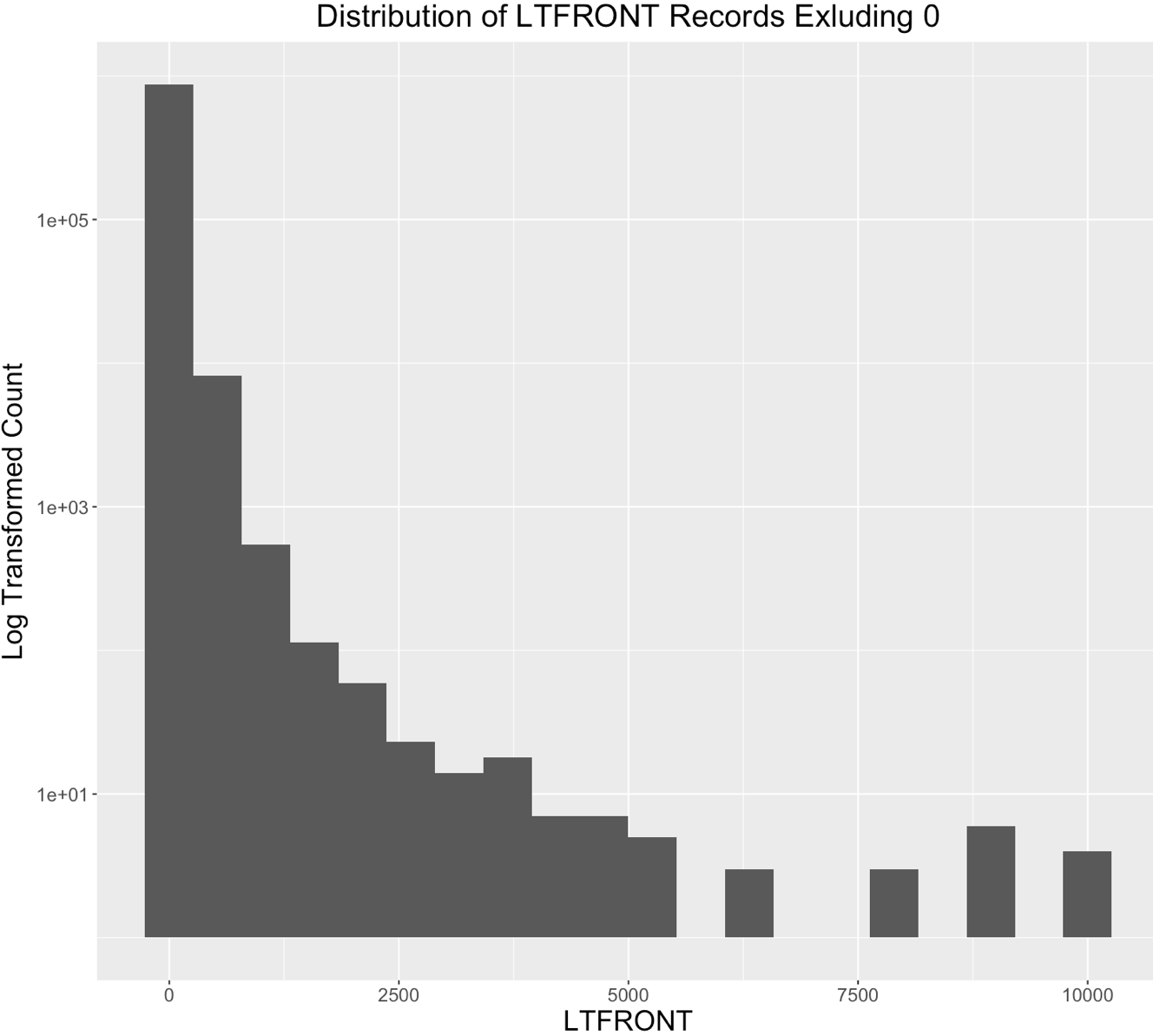
Variable Name: **TAXCLASS**

TAXCLASS is a categorical variable indicating the tax class of the property. It has 11 unique levels – “1”, “1A”, “1B”, “1C”, “1D”, “2”, “2A”, “2B”, “2C”, “3”, and “4”. No missing values exist. Sorted TAXCLASS levels are shown below:



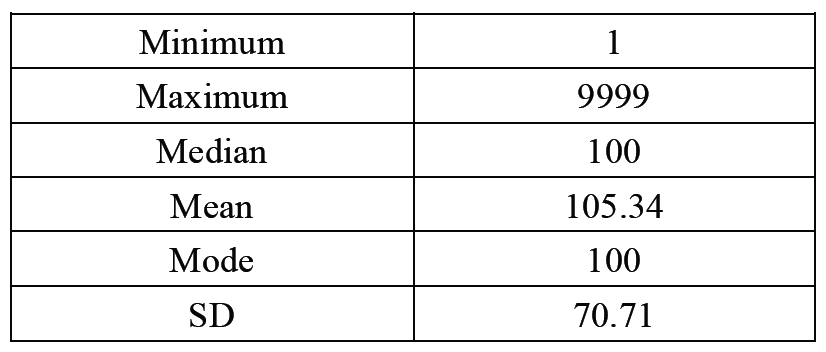
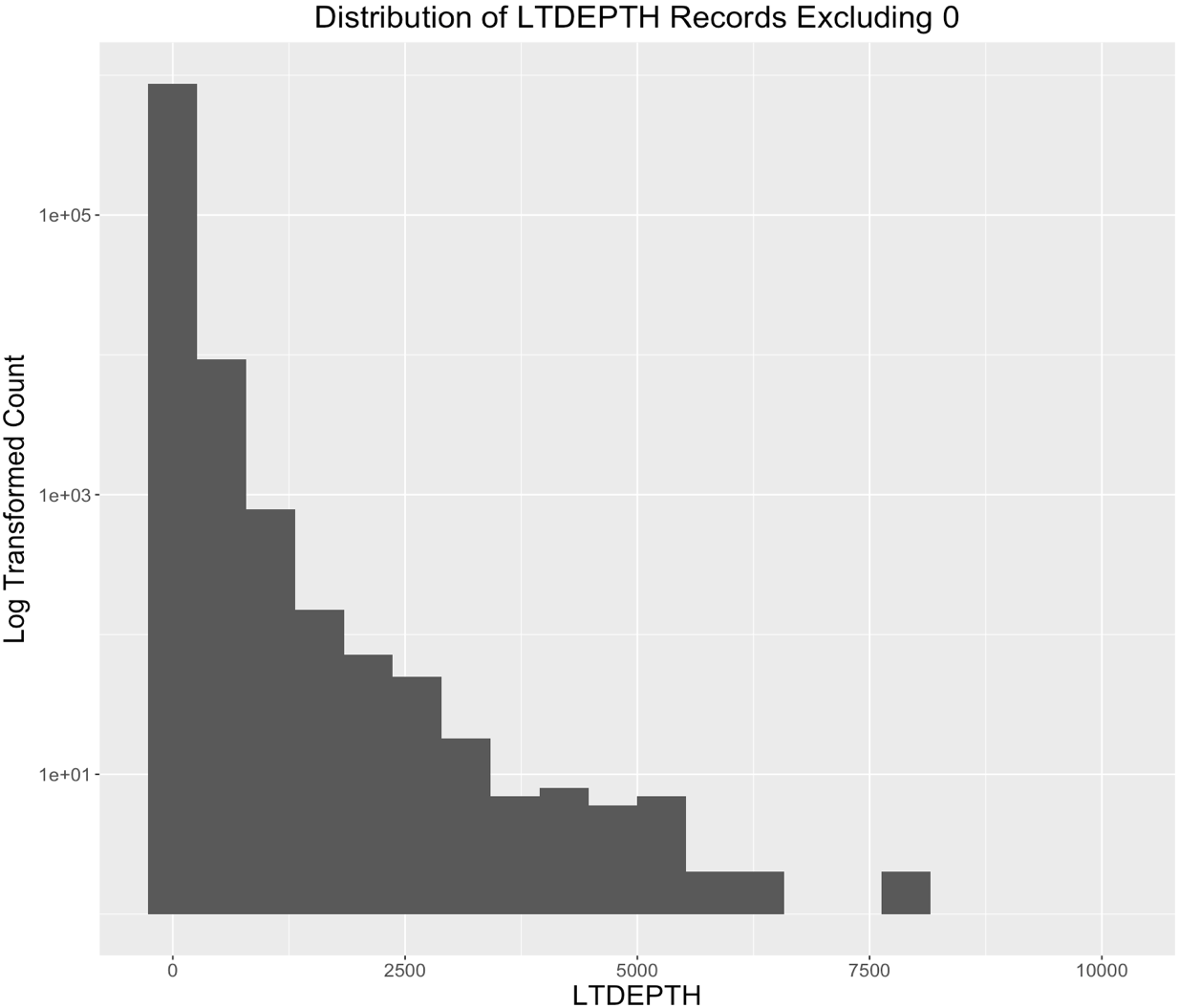
Variable Name: **LTFRONT**

LTFRONT is a numeric variable representing the length of lot frontage in feet. It has 1277 unique values ranging from 0 to 9999. No missing values exist. There are 168,867 records of 0 LTFRONT. A LTFRONT of 0 may indicate missing value. The statistics and distribution excluding 0 records are shown below:



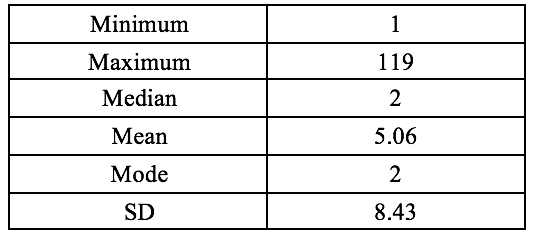
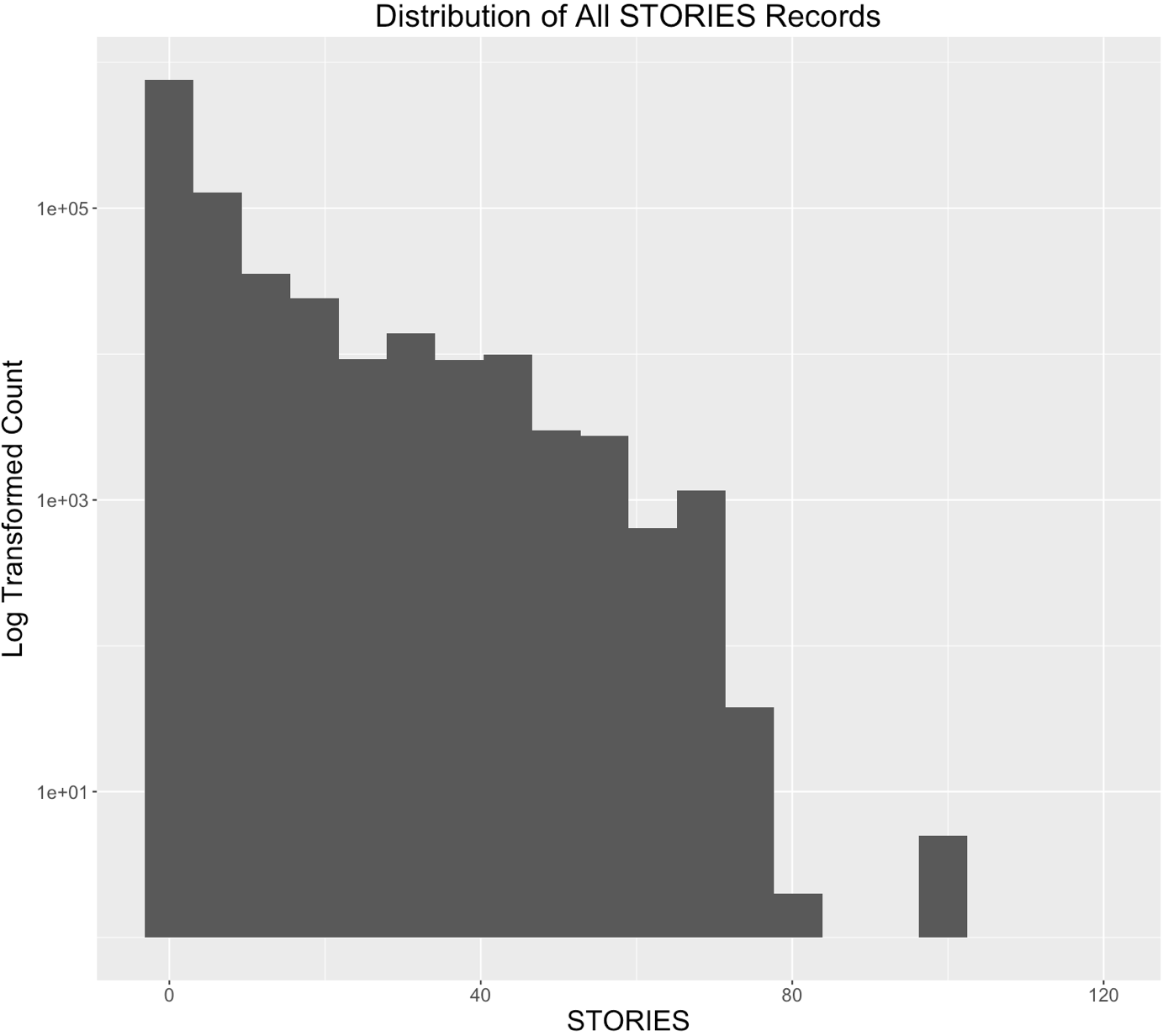
Variable Name: **LTDEPTH**

LTDEPTH is a numeric variable representing the length of lot depth in feet. It has 1336 unique values ranging from 0 to 9999. No missing values exist. There are 169,888 records of 0 LTDEPTH, and a LTDEPTH of 0 may indicate missing value. The statistics and distribution excluding 0 records are shown below:



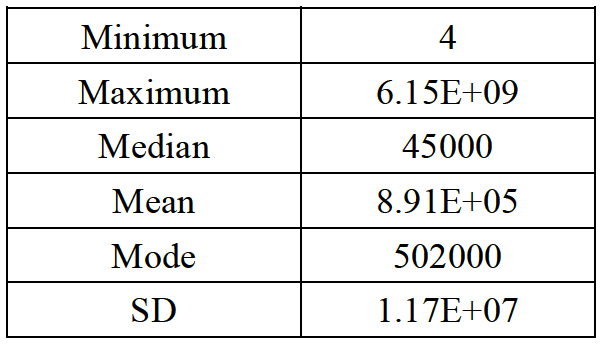
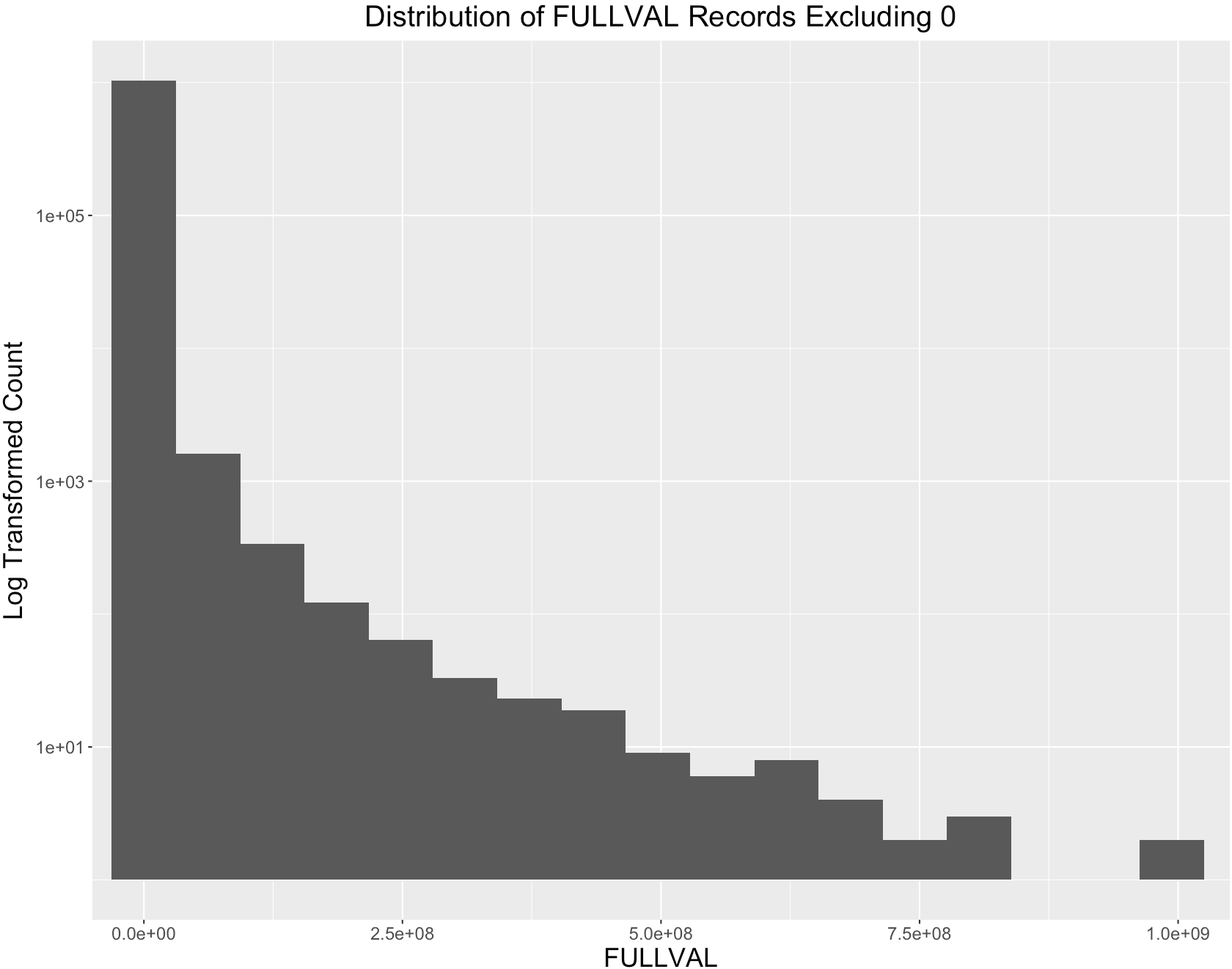
Variable Name: **STORIES**

STORIES is a numeric variable representing the number of stories of the property. It has 112 unique values ranging from 1 to 119. There are 52,142 missing values in the STORIES field. The statistics and distribution are shown below:



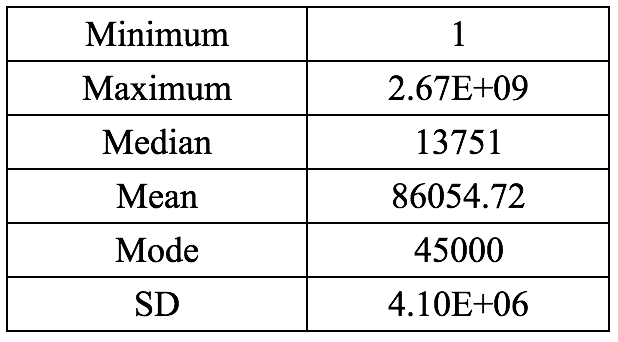
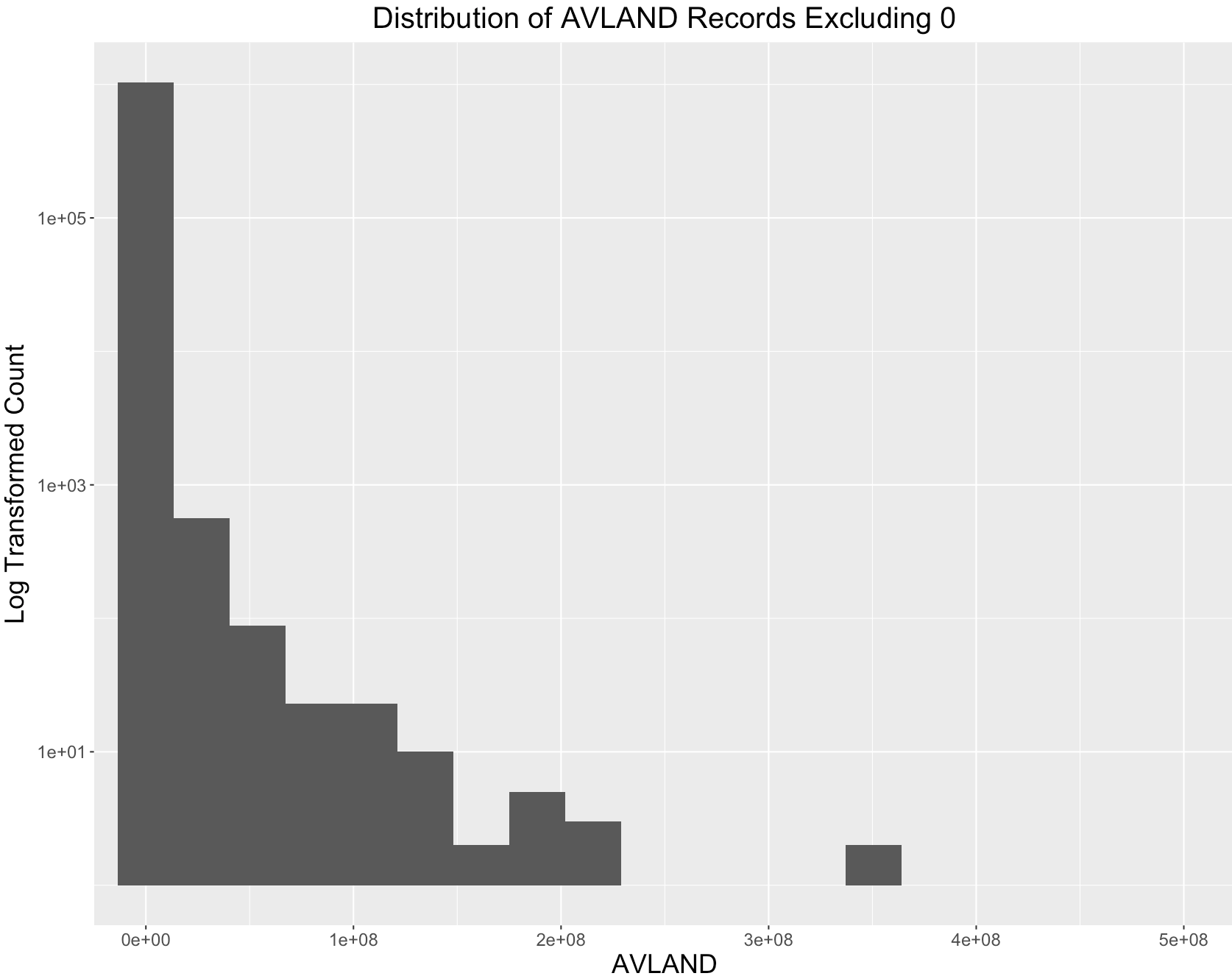
Variable Name: **FULLVAL**

FULLVAL is a numeric variable representing the full value of the property. It has 108277 unique values ranging from 0 to about 6,000,000,000. There are 12,762 properties with the FULLVAL of 0 in the dataset. No missing values exist. The statistics and distribution are shown below:



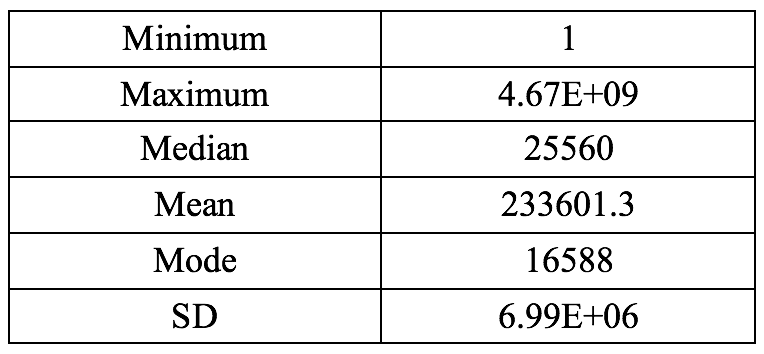
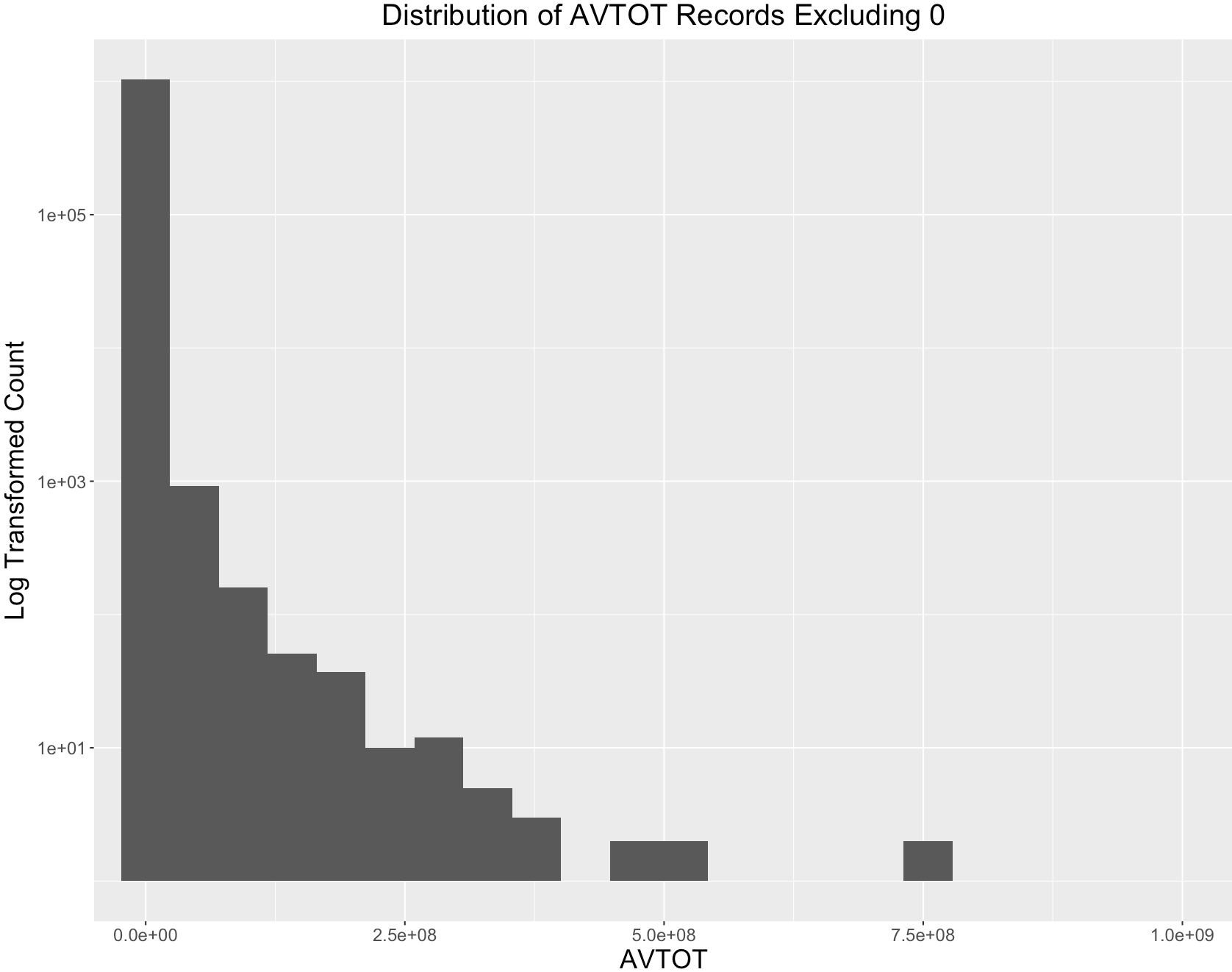
Variable Name: **AVLAND**

AVLAND is a numeric variable representing the assessed value of the land. It has 70,529 unique values ranging from 0 to about 2,700,000,000. There are 12,764 properties with the AVLAND of 0 in the dataset. No missing values exist. The statistics and distribution excluding 0 records are shown below:



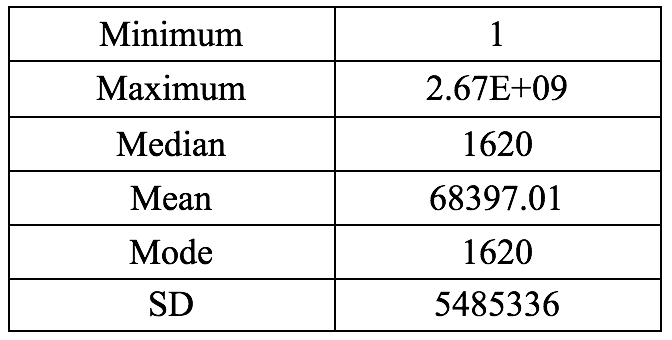
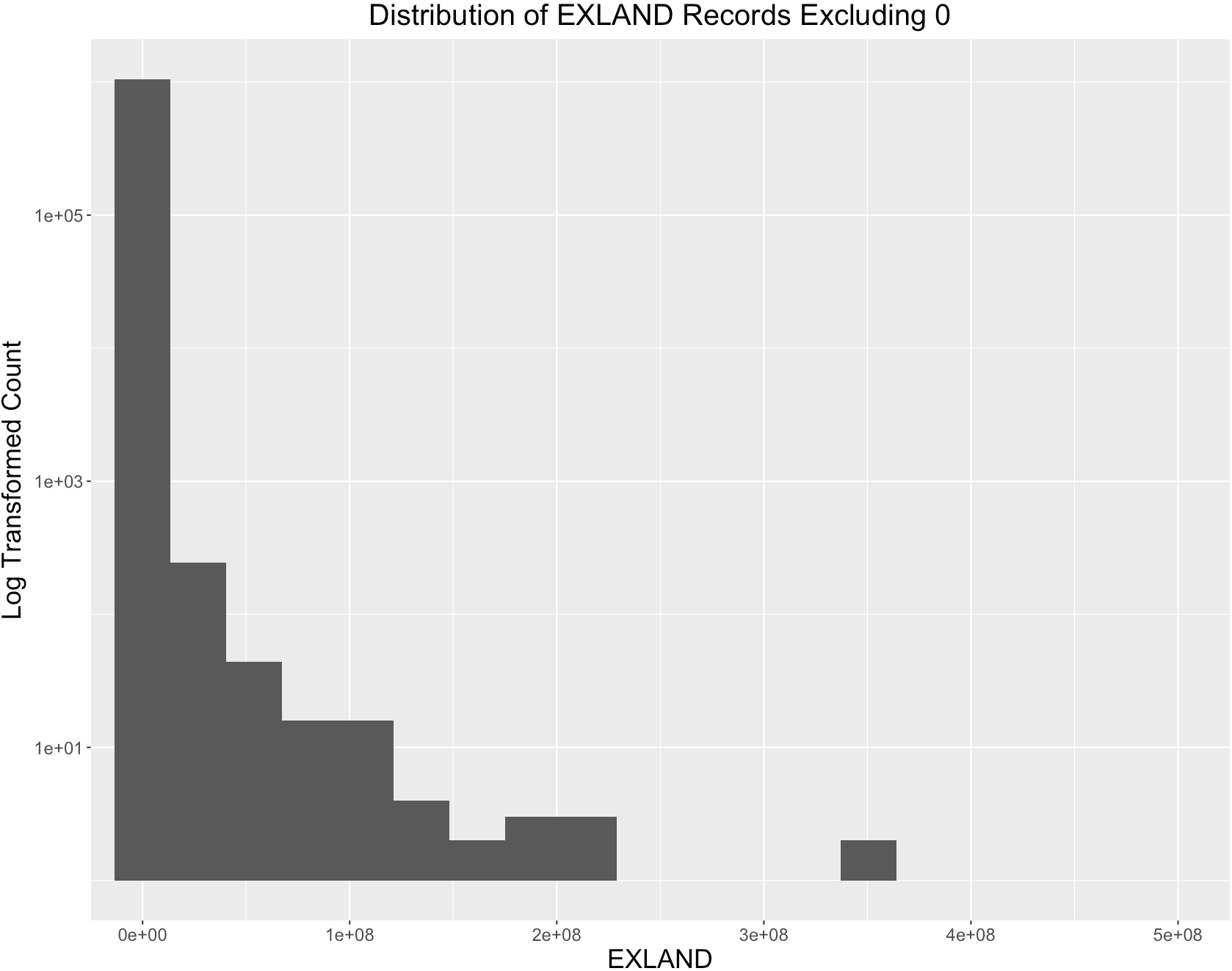
Variable Name: **AVTOT**

AVTOT is a numeric variable representing the assessed total value of the property. It has 112294 unique values ranging from 0 to about 4,700,000,000. There are 12,762 properties with the AVTOT of 0 in the dataset. No missing values exist. The statistics and distribution excluding 0 records are shown below:



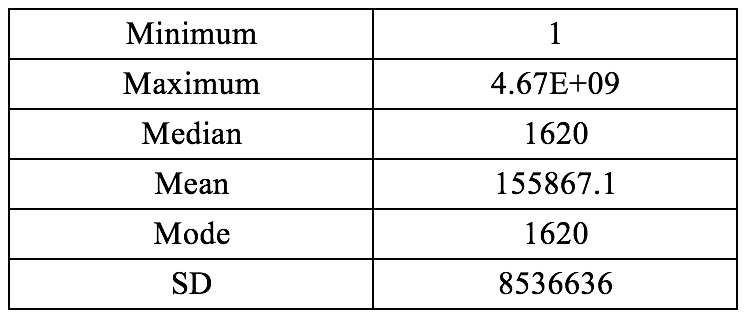
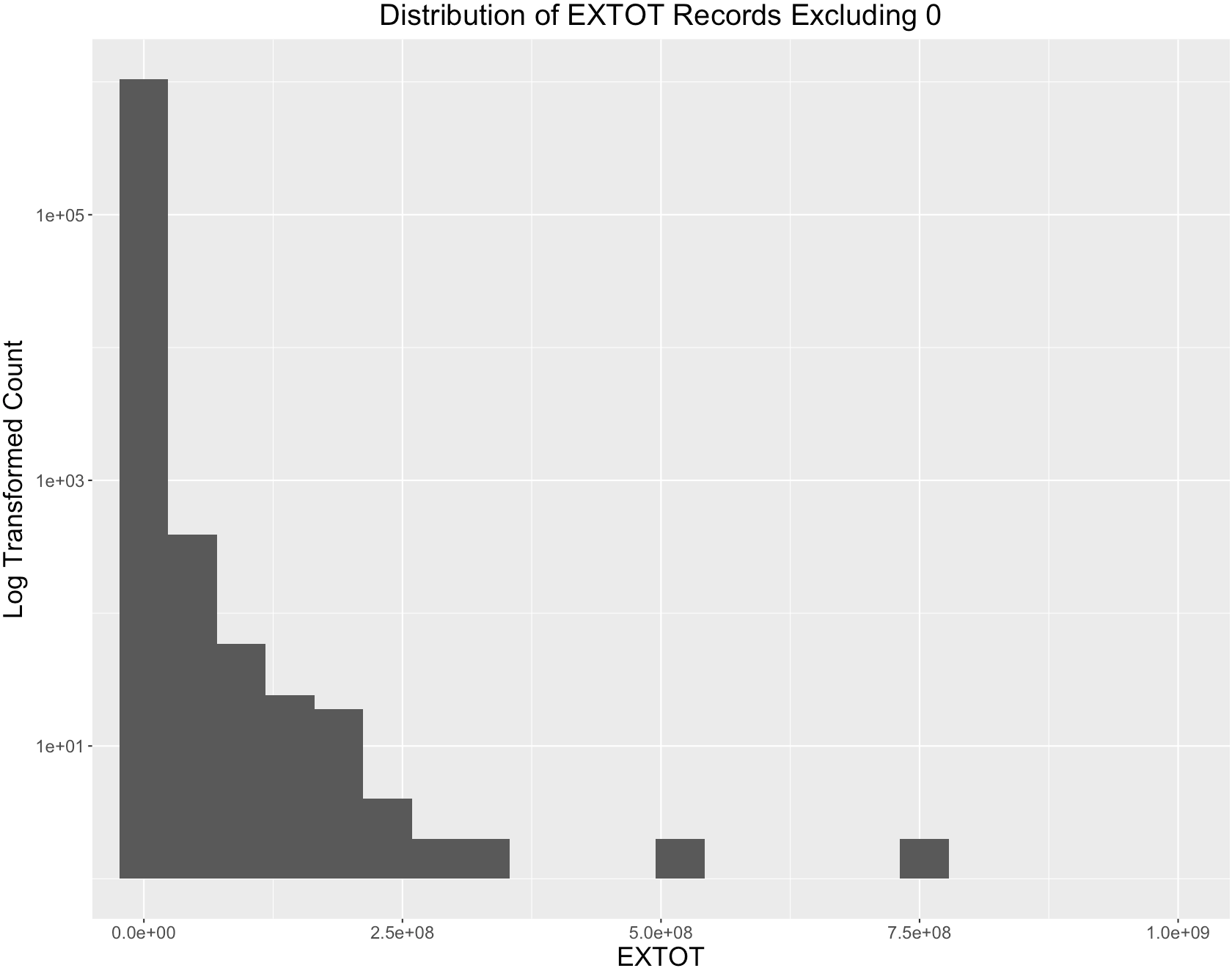
Variable Name: **EXLAND**

EXLAND is a numeric variable representing the value of the exempt land. The value of EXLAND is always smaller or equal to AVLAND. EXLAND has 33186 unique values ranging from 0 to about 2,700,000,000. There are 484,224 properties with the EXLAND of 0 in the dataset. No missing values exist. The statistics and distribution excluding 0 records are shown as below:



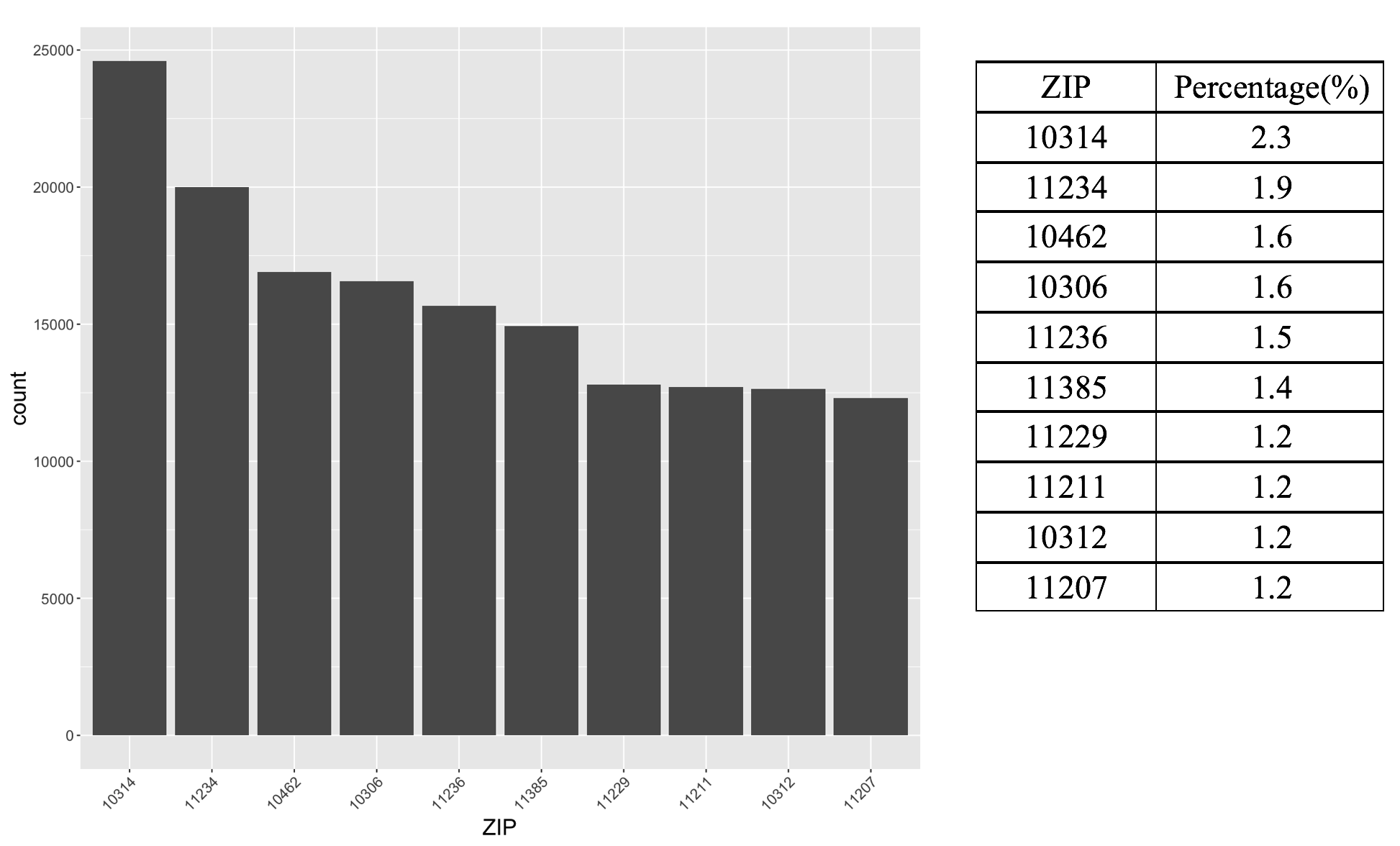
Variable Name: **EXTOT**

EXTOT is a numeric variable representing the total value of the exempt property. The value of EXTOT is always smaller or equal to AVTOT. EXTOT has 63805 unique values ranging from 0 to about 4,700,000,000. There are 425,999 properties with the EXTOT of 0 in the dataset. No missing values exist. . The statistics and distribution excluding 0 records are shown below:



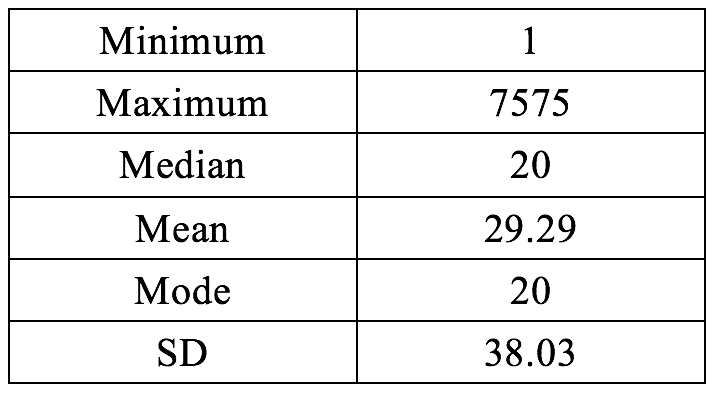
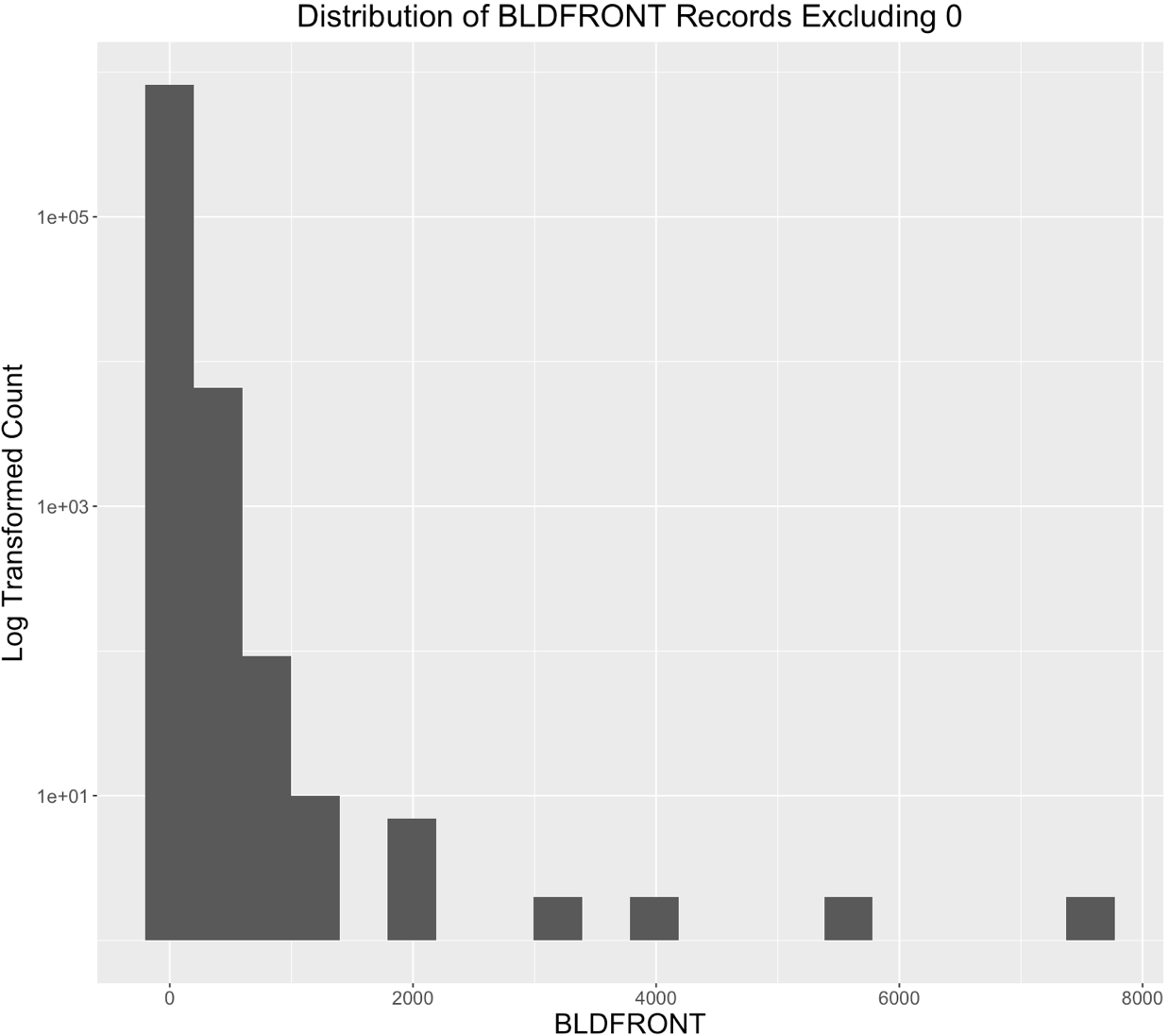
Variable Name: **ZIP**

ZIP is a categorical variable, recording the zipcode of the property. ZIP has 197 unique values and 26,356 missing values. There are three obvious anomalous records with ZIP of 33803, which should be in Florida. The top 20 most frequently occurred ZIP values are:



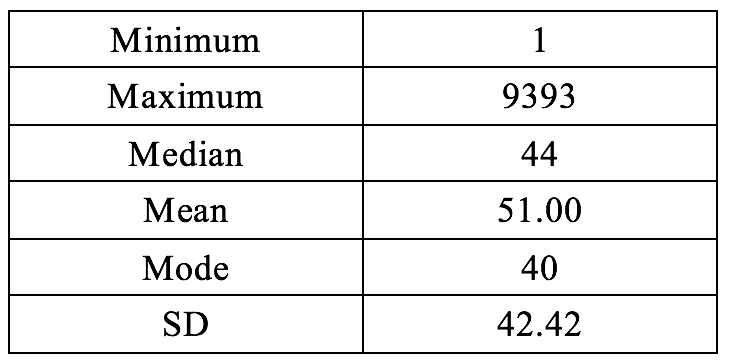
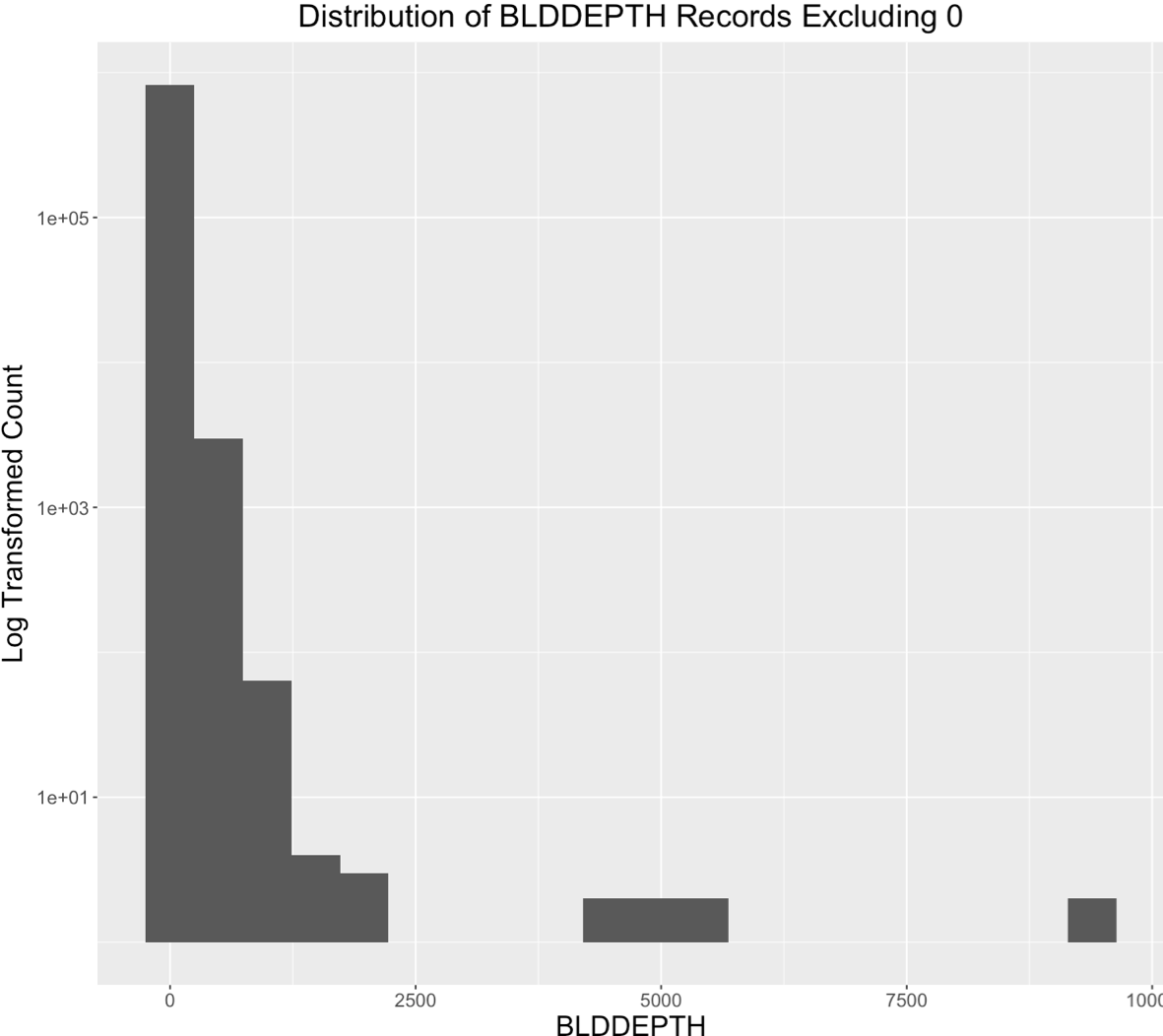
Variable Name: **BLDFRONT**

BLDFRONT is a numeric variable representing the length of building frontage in feet. It has 610 unique values ranging from 0 to 7575. No missing values exist. However, there are 224,661 records with value 0, which could be in fact missing values. The statistics and distribution excluding all records with 0 BLDFRONT are shown below:



Variable Name: **BLDDEPTH**

BLDDEPTH is a numeric variable representing the length of building depth in feet. It has 620 unique values ranging from 0 to 9393. No missing values exist. However, there are 224,699 records with value 0, which could be in fact missing values. The statistics and distribution excluding all records with 0 BLDFRONT are shown below:



**Part II. Data Cleaning**

Before constructing expert variables, we performed data cleaning to prepare the dataset for subsequent analysis.

**1. Adjusting and combining existing variables**

For the variables **BBLE**, we extracted its first digit and changed the variable name to “**BORO**”, indicating the borough where the property located.

For the variable **BLDGCL**, since there used to be 200 unique levels in the form of “[A-Z][0-9]”, and some of the categories had very few records, we only kept the first digit (the character digit) of the BLDGCL variable. Therefore, there are only 26 unique levels after the transformation.

We defined the product of the variables **LTFRONT** and **LTDEPTH** as a new variable **LOT\_AREA**, indicating the lot size of each property.

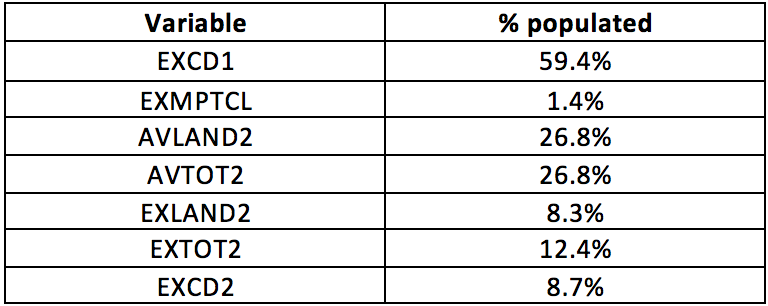
We multiplied the values of **BLDFRONT**, **BLDDEPTH** and **STORIES**, and defined the output as a new variable **BLD\_VOLUME**, indicating the volume of each building.

**2. Removing variables**

We removed three types of variables: less informative variables, less populated variables, and already aggregated variables.

We found 7 less informative variables - **STADDR**, **OWNER**, **BLOCK**, **LOT**, **PERIOD**, **YEAR** and **VALTYPE**. Although the text variables **STADDR** and **OWNER** included important identification information, we regarded them as less informative variables since there are too many levels to feed into our fraud detection model. As for variable **BLOCK** and **LOT**, although they were id numbers for the properties, they were not unique within each **BORO**. Thus they were also determined to be less informative variables. For the variables **PERIOD**, **YEAR** and **VALTYPE**, all of the records took the same value, providing no valuable information to our analysis. During the data cleaning process, we removed all of the above 7 less informative variables.

There are 7 less populated variables **EXCD1**, **EXMPTCL**, **AVLAND2**, **AVTOT2**, **EXLAND2**, **EXTOT2** and **EXCD2**. They could not serve as strong indicators of fraud considering their actual meaning. Therefore, we removed these variables from the dataset. Their percentage populated are shown below.



Since we created the variable **LOT\_AREA** based on **LTFRONT** and **LTDEPTH**, and we created the variable **BLD\_VOLUME** based on **BLDFRONT**, **BLDDEPTH** and **STORIES**, we decided to remove **LTFRONT**, **LTDEPTH**, **BLDFRONT**, **BLDDEPTH** and **STORIES** from our dataset.

**3. Filling in the missing values**

For the variable **EASEMENT**, 99.6% of the properties in this dataset were left blank, indicating they did not have an easement type. Since we considered that **EASEMENT** could be an important indicator for fraud, we filled in the missing values with a newly-created category “NO”.

For the variable **STORIES**, there were 5% records with missing values. We filled in the missing values with the average **STORIES** in their own **TAXCLASS**.

For the variable **ZIP**, there were 2.5% records with missing values. We filled in the missing values with “00000”.

After the data cleaning process, we kept 13 variables in the dataset: **RECORD**, **FULLVAL**, **AVLAND**, **AVTOT**, **EXLAND**, **EXTOT**, **BORO**, **EASEMENT**, **BLDGCL**, **TAXCLASS**, **ZIP**, **LOT\_AREA**, **BLD\_VOLUME**.

**Part III. Variable Construction**

To begin with, we divided the original variables into two sets, 9 numerators and 6 denominators, before constructing expert variables.

The 9 numerators variables are:

1. **FULLVAL**: full value of building
2. **AVLAND**: assessed value of land
3. **AVTOT**: assessed value of property
4. **EXLAND**: exemption value of land
5. **EXTOT**: exemption value of property
6. **FULLVAL / AVTOT**: the ratio of full value of building to assessed value of property
7. **AVTOT / EXTOT**: the ratio of assessed value to exemption value of property
8. **AVLAND / EXLAND**: the ratio of assessed value to exemption value of land
9. **FULLVAL / EXTOT**: the ratio of full value of building to exemption value of property

All the numerators are numeric variables, which closely relate to the monetary value of properties. 1-4 were from the original dataset, while 6-9 were created by us to capture the relationship between full values, assessed values, and exemption values.

When calculating those ratios, we encountered many 0s in some of the numerical variables. Our calculation gave back infinity if we calculated their averages and put them in the denominator position. Value 0s themselves could be signs of fraud, while sometimes they could be valid and reasonable as well. For example, 0 in **EXLAND** meant the property did not have exemptions. Therefore, replacing them with the median value might not be reasonable. Our decision was to substitute these 0s with 1s. Since those values were large enough (usually in thousands), 1s would still be small enough for us to detect anomaly without causing calculation problems.

The 6 denominator variables are:

1. **BORO**: borough code
2. **EASEMENT**: easement is a non-possessory right to use and/or enter onto the real property of another without possessing it.
3. **BLDGCL**(1st): building class
4. **TAXCLASS**: tax class
5. **LOT\_AREA** =LOTFRONT \* LOTDEPTH: measurement of lot area
6. **BLD\_VOLUME** = STORIES \* BLDP \* BLFT: measurement of building volume
7. **ZIP**: zip code

All the denominator variables (except **LOT\_AREA** and **BLD\_VOLUME**) are used to classify numerators. All the denominator variables are used to classify numerators. That is, we divided all those numerators by these denominator variables, calculated median of numerical variables in each group, and divided numerical variables by the median of each group.

For example, if we used **FULLVAL** (numerator) and **TAXCLASS** (denominator), the expert variable would be **FULLVAL/ (median of FULLVAL in the TAXCLASS that the property belongs)**.

For denominators **LOT\_AREA** and **BLD\_VOLUME**, the expert variables were the ratios of the value divided by area (or divided by volume) of a particular property to the average value divided by area (or divided by volume). For example, when we combined **FULLVAL** and **BLD\_VOLUME**, we created a variable **FULLVAL/BLD\_VOLUME/(median of FULLVAL/BLD\_VOLUME)**. Again, we substituted 0s in **BLD\_VOLUME** and **LOT\_AREA** (both usually in thousands) with 1s to avoid calculation error.

We exploited the combinations of each numerators and denominators, except the combinations **AVLAND** and **BLDGVOL**, **EXLAND** and **BLDGVOL**. The reason was that we believed the value of land was not closely related to the building volume.

In total, we created **61** expert variables.

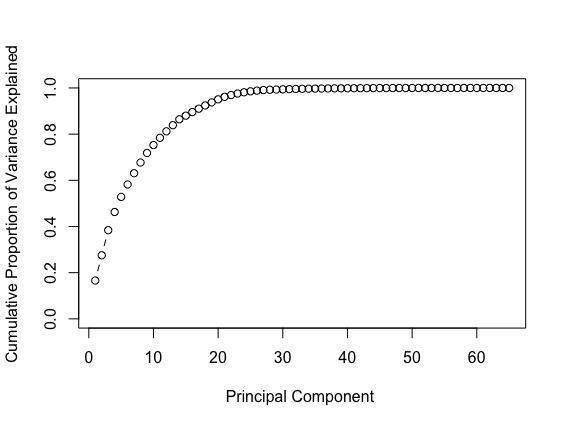
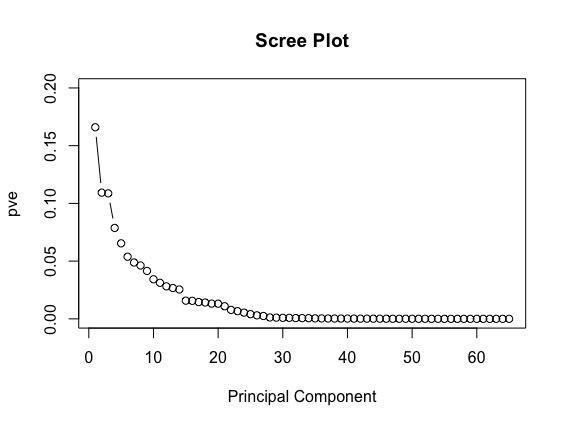
**Part IV. Fraud Algorithm**

After we had all the expert variables and their respective values at hand, we started the process of standardization and dimensionality reduction for further analysis.

We performed principal components analysis using the **prcomp()** function, which was one of several functions in R that could perform PCA. By default, the **prcomp()** function centers the variables to have mean zero. By using the option **scale = TRUE**, we scaled the variables to have standard deviation of 1. The ‘center’ and ‘scale’ components correspond to the means and standard deviations of the variables that were used for standardization prior to implementing PCA.

The **rotation** matrix provided the principal component loadings, each column of **pr.out$rotation** contained the corresponding principal component loading vector.

To compute the proportion of variance explained by each principal component, we simply divided the variance explained by each PC by the total variance explained by all 61 PCs. We made the scree plot and cumulative plot to determine which PCs to keep. We would like to use the smallest number of PCs required to get a good understanding of the data. By examining the scree plot below, we discovered that there is a significant drop between PC14 and PC15. Thus, we decided to keep PC1 through PC14, which explained approximately 90% of the entire dataset.



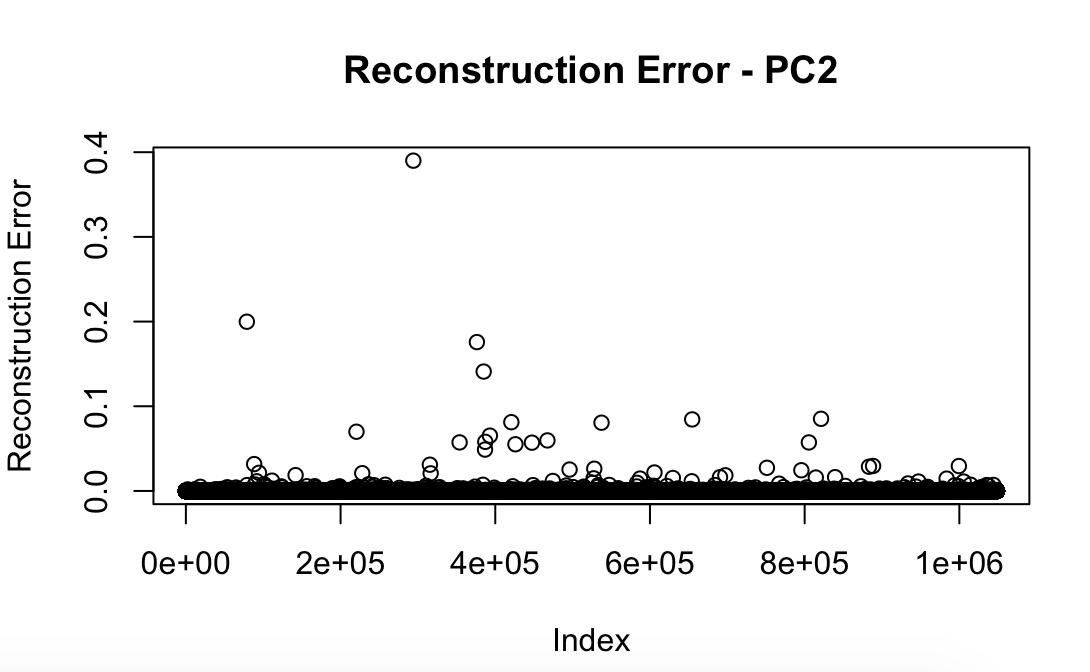
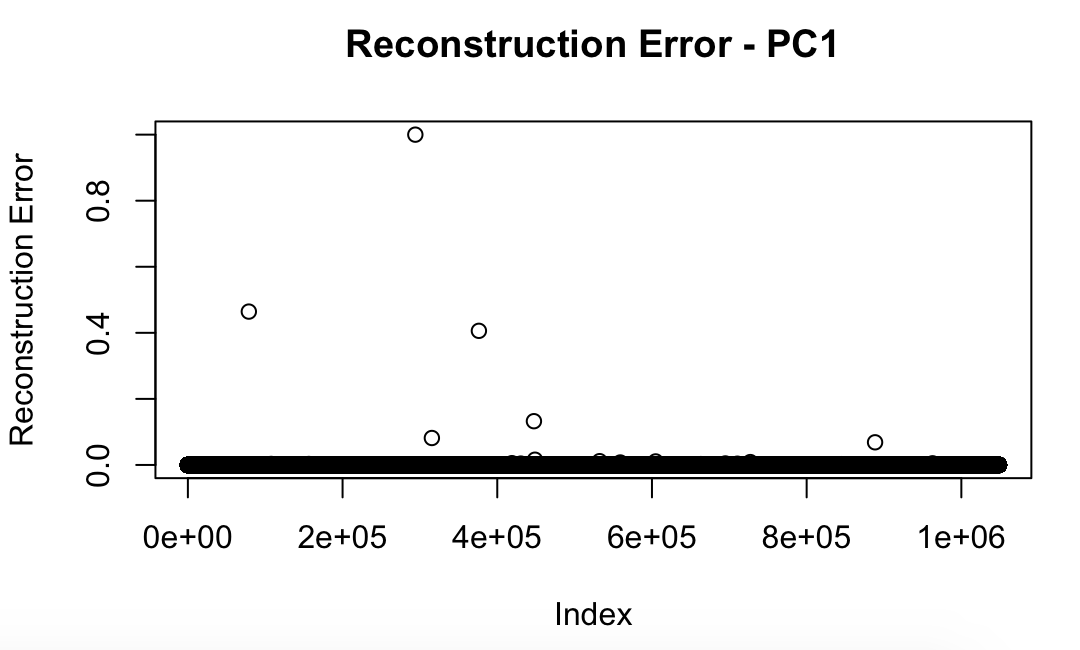
Finally, we reduced the dimension of the original dataset by multiplying the PCA matrix and the original data matrix to get the final dataset for further calculation of the fraud scores.

We used two different ways to calculate fraud scores. The first one being autoencoder, and the second one was a heuristic algorithm.

**Autoencoder:**

First we tried to autoencode our PCs using an R package called “**h2o**”. Then, we called the **deep learning** function with parameter “autoencoder” set to TRUE. This function took the original dataset with all the PCs and autoencoded it. We then called the **h2o.anomaly** function to reconstruct the original dataset using the reduced set of features and calculated a mean squared error between both. We set the “per\_feature” parameter to TRUE because we wanted a reconstruction mean error based on individual features. We saved the reconstruction error in a dataset called “**error**”.

From plotting of the “**error**” dataset, we could see that there were some abnormal values, which might indicate fraud. Below are examples of the distribution of reconstruction errors for PC1 and PC2:



In the end we summed the reconstruction error values of all the PCs to get a single score, which would be our fraud score from autoencoder, for each of the record.

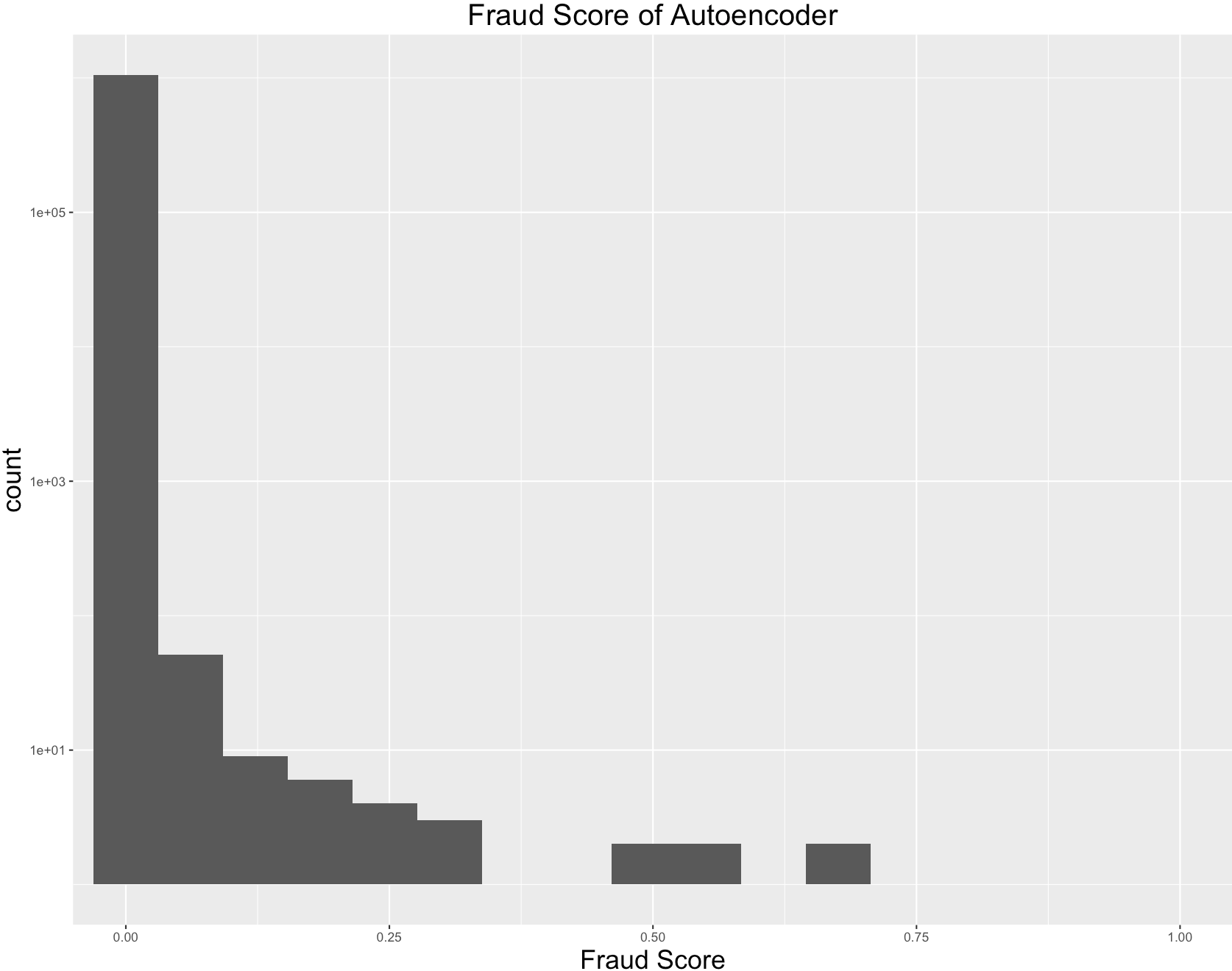
**Heuristic Algorithm:**

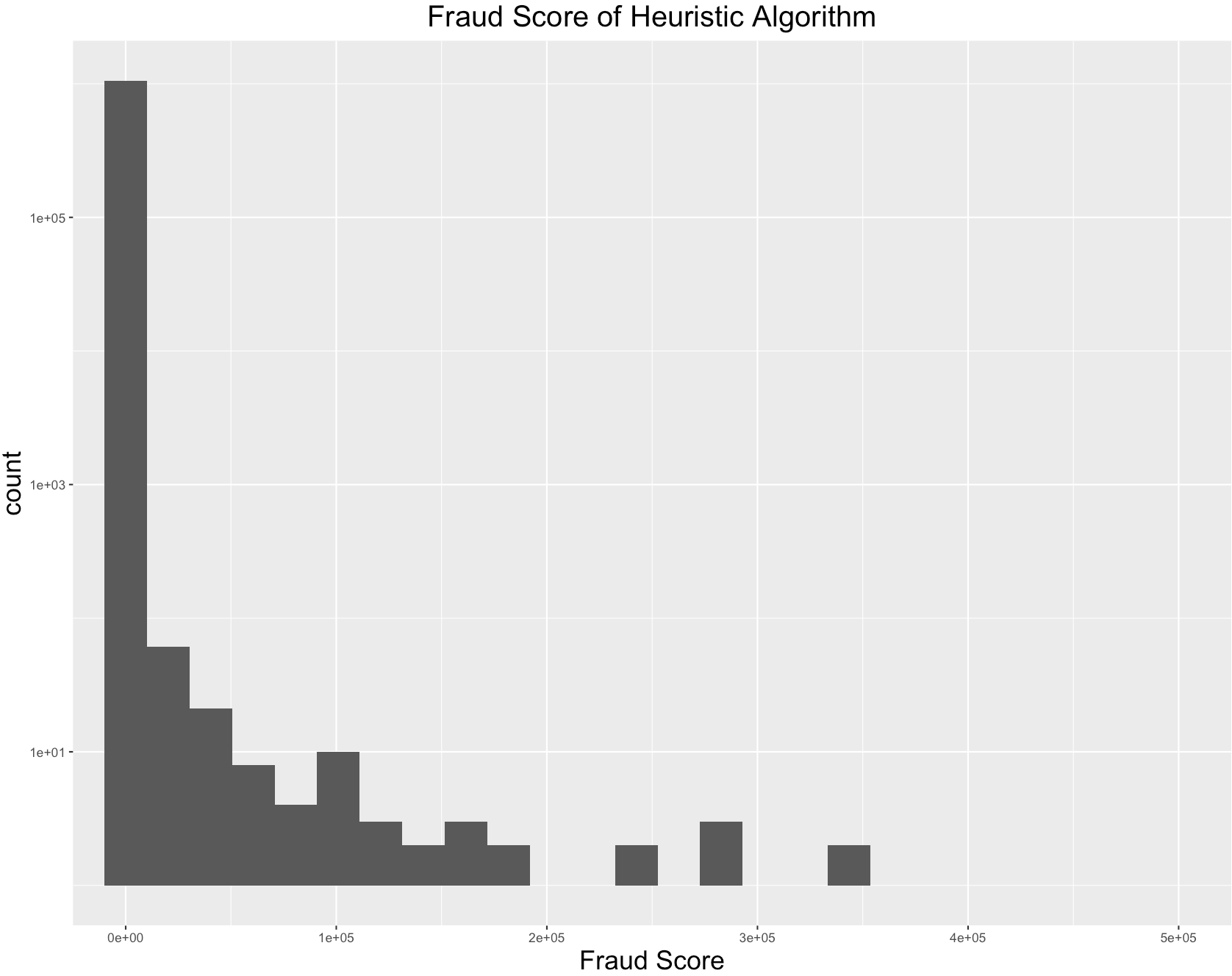
Our algorithm was calculating the **mahalanobis distance** between each record and the (mean,covariance) of records within each particular PC. The **mahalanobis distance** was our fraud score for each record. It was calculated using the function ‘**mahalanobis**’ in R.

**Part V. Results**

Having fraud scores ready, we sorted the records according to fraud scores from both autoencoder and heuristic algorithm outcomes. Not surprisingly, the majority of records had low fraud scores while a small proportion of the records had typically high fraud scores.

Below is an overview of what the distribution of fraud scores from both methods:





We decided to look at the **overlapping** part of the top **1%** high score records from autoencoder output and the top **1%** high score records from heuristic algorithm output. About **70%** of the records from these two algorithms matched, so we selected these overlapped records as the best candidates for potential fraud.

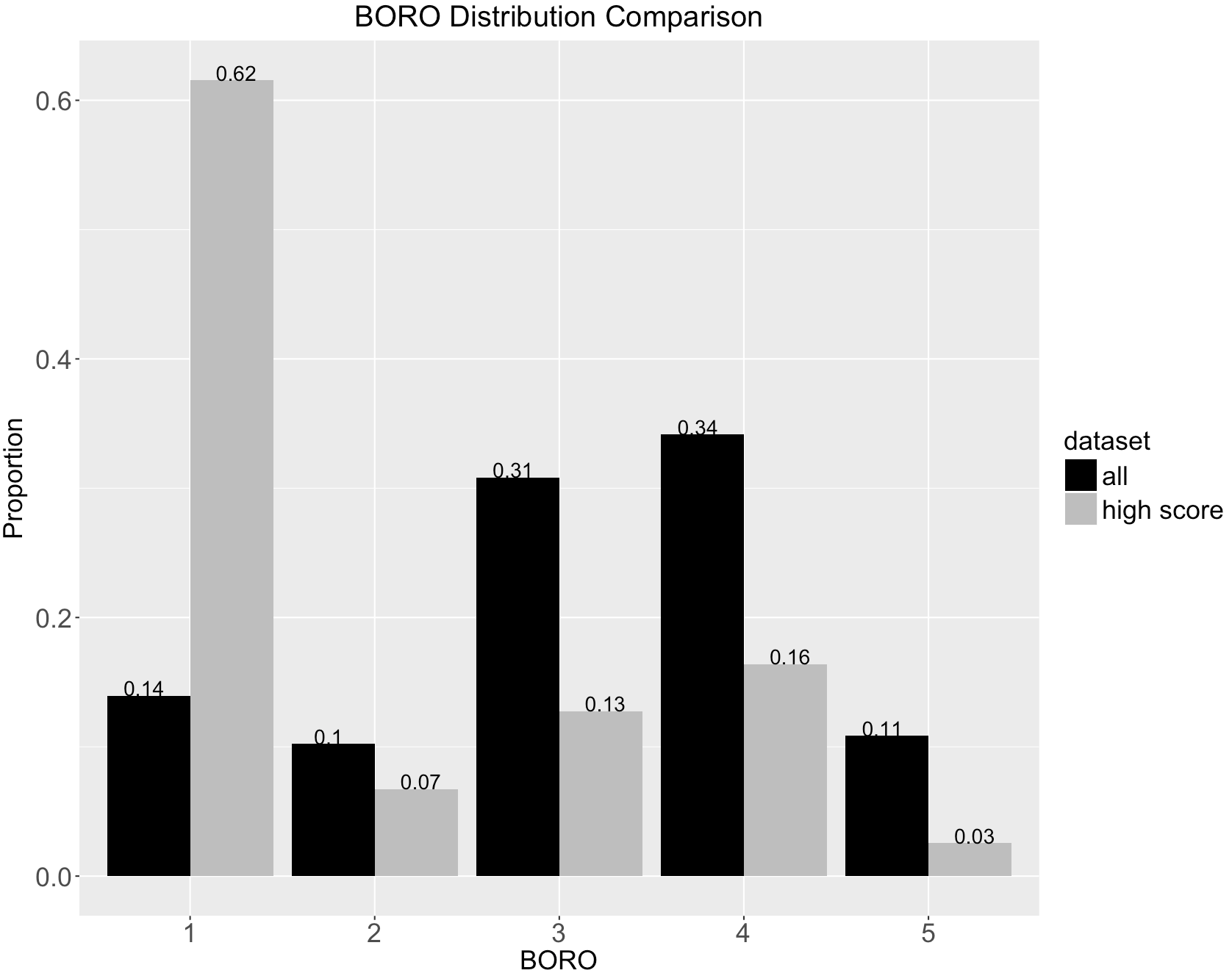
**General Trends:**

We found some general trends on the overlapped part of the top 1% high score records. The following table compares the mean, median and standard deviation of **“Top 1%”** records with the mean, median and standard deviation of complete data.

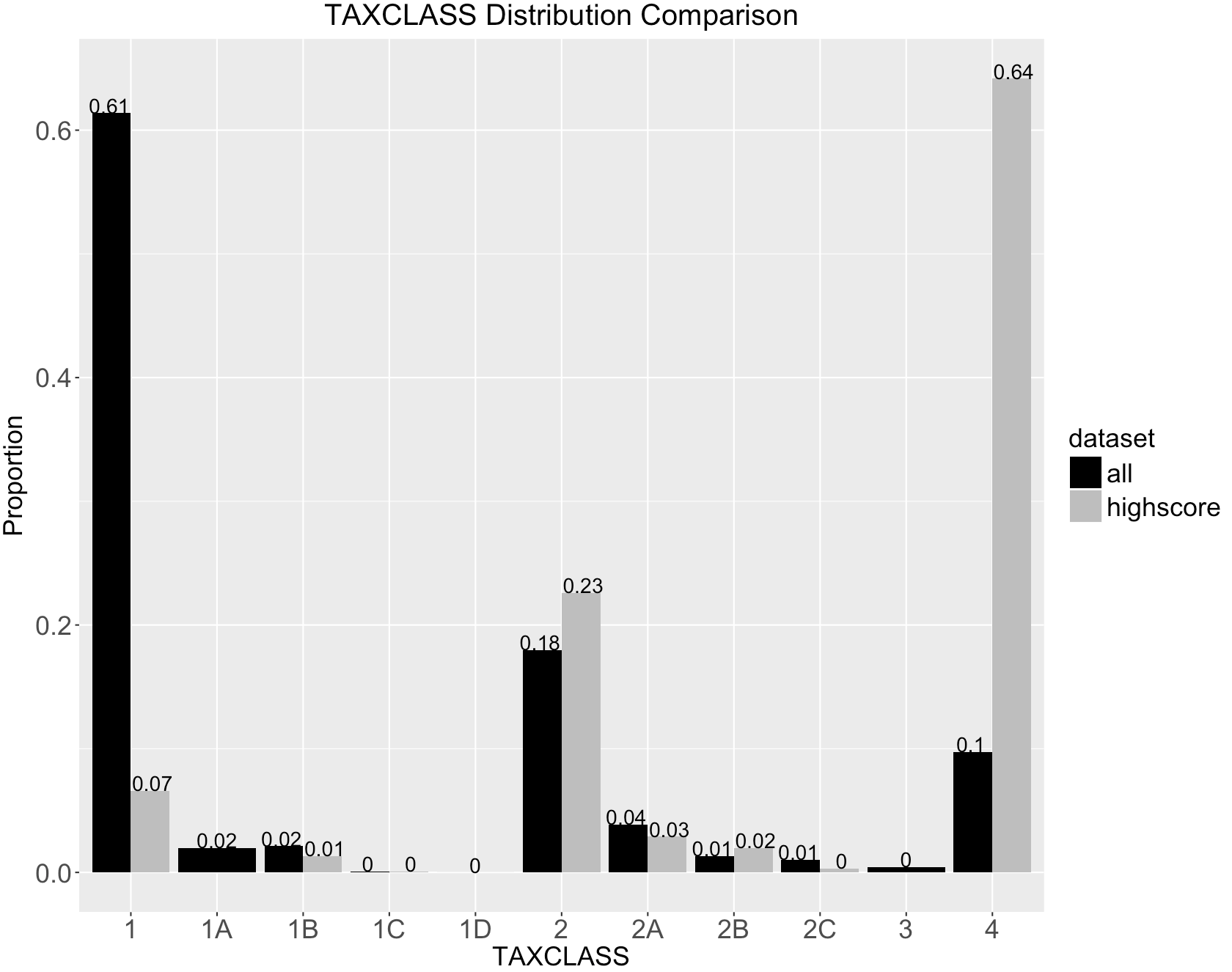
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Complete Data | | | | Top 1% records | | | |
|  | number | mean | stdev | median | number | mean | stdev | median |
| LTFRONT | 1048575 | 36 | 74 | 25 | 7217 | 221 | 457 | 125 |
| LTDEPTH | 1048575 | 88 | 75 | 100 | 7217 | 238 | 379 | 131 |
| STORIES | 996433 | 5 | 8 | 2 | 6754 | 12 | 13 | 6 |
| BLDFRONT | 1048575 | 23 | 36 | 20 | 7217 | 114 | 143 | 92 |
| BLDDEPTH | 1048575 | 40 | 43 | 39 | 7217 | 121 | 114 | 99 |
| LOT\_AREA | 1048575 | 5902 | 154727 | 2400 | 7217 | 153885 | 1754442 | 17574 |
| BLD\_VOLUME | 1048575 | 19043 | 2315821 | 1520 | 7217 | 232034 | 969582 | 60000 |
| FULLVAL | 1048575 | 880488 | 11702927 | 446000 | 7217 | 37590715 | 134865357 | 12230000 |
| AVLAND | 1048575 | 85995 | 4100755 | 13646 | 7217 | 6614340 | 48918875 | 1413000 |
| AVTOT | 1048575 | 230758 | 6951206 | 25339 | 7217 | 16935641 | 81733286 | 5175000 |
| EXLAND | 1048575 | 36812 | 4024330 | 1620 | 7217 | 3565607 | 48339361 | 0 |
| EXTOT | 1048575 | 92544 | 6578281 | 1620 | 7217 | 8156874 | 78753026 | 0 |
| FV\_AT | 1048575 | 22 | 429 | 18 | 7217 | 40 | 1587 | 2 |
| AT\_ET | 1048575 | 95286 | 1942435 | 17 | 7217 | 7532982 | 22081850 | 2169000 |
| FV\_ET | 1048575 | 348064 | 4339495 | 352 | 7217 | 17883478 | 48834099 | 7310000 |
| AL\_EL | 1048575 | 39253 | 777335 | 12 | 7217 | 3022125 | 8811510 | 607500 |

**Insights:**

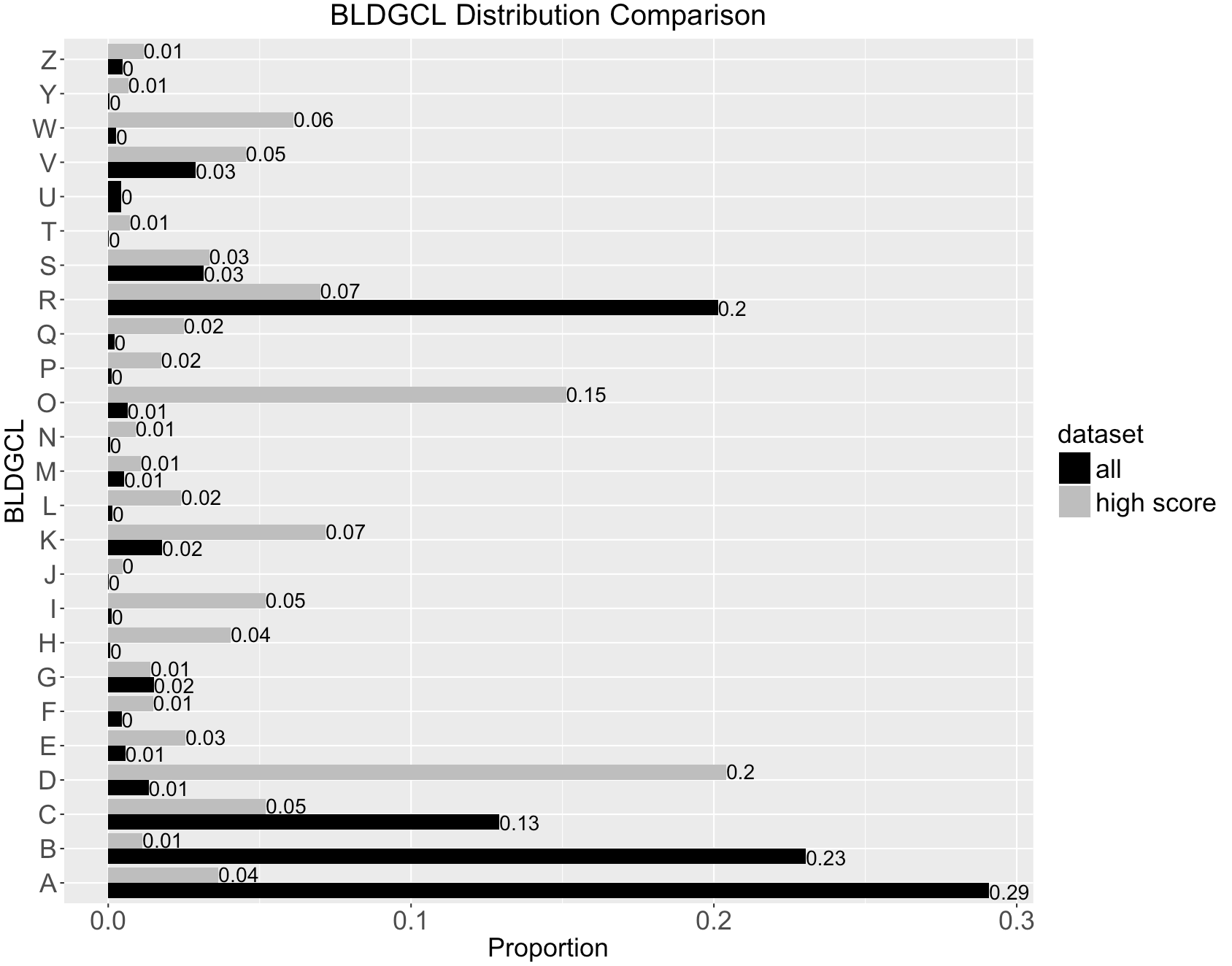
1. We see that the potential fraud properties have significantly higher mean, median and standard deviation values of variables **LTFRONT**, **LTDEPTH**, **STORIES**, **BLDFRONT**, **BLDDEPTH**, **LOT\_AREA**, **BLD\_VOLUME** (all in green in above table) compared to the complete data. This means these are usually the big buildings of the city.
2. **FULLVAL**, **AVLAND**, **AVTOT**, **EXLAND**, **EXTOT**, **FV\_AT**, **AT\_ET**, **FV\_ET**, **AL\_EL** have significantly higher mean, median and standard deviation values of variables in **Top 1%** records when compared to the complete data.
3. The important thing to note is the variable **FV\_AT**. The variable gives us the ratio of **FULLVAL/AVTOTAL.** **FV\_AT** whose mean in complete data is 22 but mean in Top 1% records increases to 40. This tells us that the properties in Top 1% records are being **significantly undervalued** **and hence paying lower taxes** than they should when compared to the complete data.
4. The distribution of **BORO** in the top 1% records is very different from that in the whole dataset. Among the top 1% records that got high scores, over 60% of the properties are located in **BORO** 1, which is Manhattan, while in the original dataset, only 14% properties are in that area. This could be caused by the fact that the house price and land price are much higher in Manhattan than in other areas in New York City.



1. Looking at the **TAXCLASS** of the top 1% high score records, we found that **64%** of the properties belong to the **TAXCLASS** **4**. But in the whole dataset, there are only **10%** properties in **TAXCLASS 4**. This is reasonable because properties in **TAXCLASS 4** represent “UTILITIES - CEILING RAILROADS” and “ALL OTHERS”, which could have a totally different set of values.



1. The top **1%** properties also have higher proportion of **BLDGCL D** and **O**, comparing to that of the whole dataset. Buildings in **BLDGCL D** are elevator apartments, while those in **BLDGCL O** are office buildings. Both tend to have higher values.



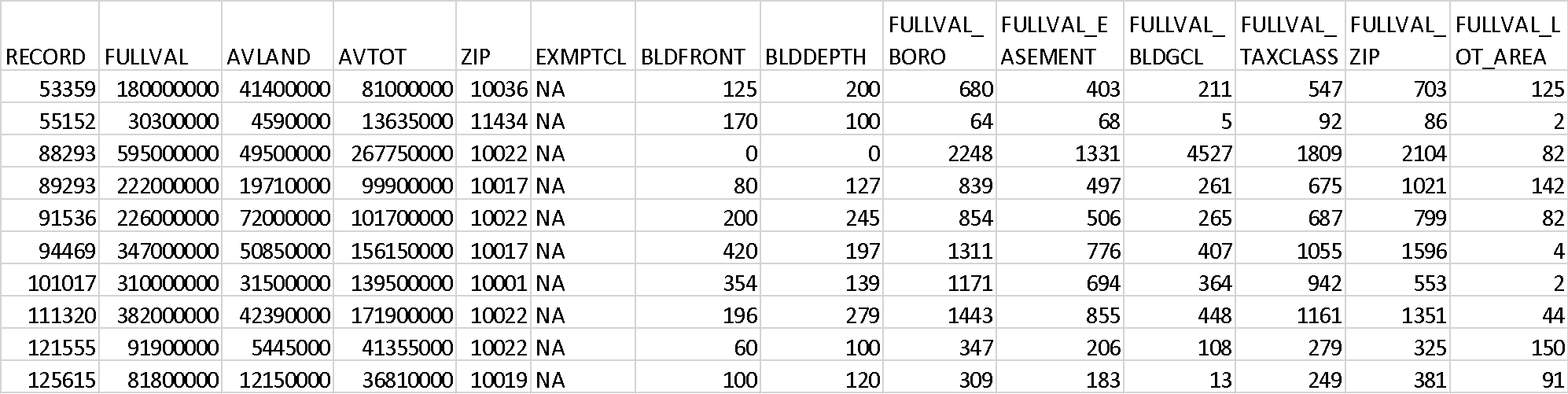
1. OWNER

When we look at the owners of the high score properties, we find that most of the owners are large house agencies or real estate companies. There are also many educational institutions and government entities. Few of these properties belongs to single households.

**Top 10 records:**

The below tables show the top 10 records.





We examined these records one by one:

1. **RECORD No. 53359** - We can see that it has unusual values of FULLVAL with respect to BORO, BASEMENT, BLDGCL, TAXCLASS, ZIP and LOTAREA. The ratios are higher than 100 and surely indicate that there may be some kind of fraud.
2. **RECORD No. 55152** – Although its FULLVAL with respect to LOTAREA and BLDGCL seems fine, it has unusual values of FULLVAL with respect to BORO, BASEMENT, TAXCLASS, ZIP. The ratios are higher than 50 and surely indicate that there may be some kind of fraud.
3. **RECORD No. 88293** –We can see that it has unusual values of FULLVAL with respect to BORO, BASEMENT, BLDGCL, TAXCLASS, ZIP and LOTAREA. The ratios are higher than 1000 and is a must pick record for investigation purposes.
4. **RECORD No. 89293** – We can see that it has unusual values of FULLVAL with respect to BORO, BASEMENT, BLDGCL, TAXCLASS, ZIP and LOTAREA. The ratios are higher than 100 and surely indicate that there may be some kind of fraud.
5. **RECORD No. 91536** – We can see that it has unusual values of FULLVAL with respect to BORO, BASEMENT, BLDGCL, TAXCLASS, ZIP and LOTAREA. The ratios are higher than 100 and surely indicate that there may be some kind of fraud.
6. **RECORD No. 94469** – Although its FULLVAL with respect to LOTAREA seems fine. It has unusual values of FULLVAL with respect to BORO, BASEMENT, TAXCLASS, ZIP and BLDGCL. The ratios are higher than 100 and surely indicate that there may be some kind of fraud.
7. **RECORD No. 101017** – Although its FULLVAL with respect to LOTAREA seems fine. It has unusual values of FULLVAL with respect to BORO, BASEMENT, TAXCLASS, ZIP and BLDGCL. The ratios are higher than 100 and surely indicate that there may be some kind of fraud.
8. **RECORD No. 111320** – We can see that it has unusual values of FULLVAL with respect to BORO, BASEMENT, BLDGCL, TAXCLASS, ZIP and LOTAREA. Most ratios are higher than 100 and surely indicate that there may be some kind of fraud.
9. **RECORD No. 121555** – We can see that it has unusual values of FULLVAL with respect to BORO, BASEMENT, BLDGCL, TAXCLASS, ZIP and LOTAREA. Most ratios are higher than 100 and surely indicate that there may be some kind of fraud.
10. **RECORD No. 125615** – Although its FULLVAL with respect to BLDGCL seems fine. It has unusual values of FULLVAL with respect to BORO, BASEMENT, TAXCLASS, ZIP and LOTAREA. Most ratios are higher than 100 and surely indicate that there may be some kind of fraud.

**NOTE:** The top 10 records that were captured by using the fraud score algorithms had a record which belongs to an exemption category and is owned by US Government so we have removed it from the list of potential frauds.

1.png

2.png

**APPENDIX**

City of New York Property Valuation and Assessment Data

Data Quality Report

**Summary**

**File description:**

The City of New York Property Valuation and Assessment Data file is a public available dataset posted by the Department of Finance on the City of New York Open Data website. The dataset contains the records of more than 1 million properties across the city of New York and information on their sizes, values, owners, building classes, tax classes, etc.

**File Name:**

City of New York Property Valuation and Assessment Data

**Data Source:**

City of New York Open Data Website (<https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8>)

**Number of Records:**

1,048,575 records

**Number of Fields:**

30 variables in total – 13 categorical variables, 14 numeric variables, 2 text variables, 1 date variables

**Time of Records:**

November 2011

**Fields Explanation**

**Field 1**

**Field Name:** RECORD

**Description:**

RECORD is a categorical variable. It works as the ordinal reference number for each property record.

**Unique Values:**

1,048,575 unique values, ranging from 1 to 1,048,575. No repeated values or missing values exist.

**Field 2**

**Field Name:** BBLE

**Description:**

BBLE is a nominal categorical variable with 10 or 11 digits. It is the concatenation of BORO code (1 digit), BLOCK code (5 digit), LOT code (4 digit) and EASEMENT code (1 digit if exists).

**Unique Values:**

1,048,575 unique values. No repeated values or missing values exist.

**Field 3**

**Field Name:** BLOCK

**Description:**

BLOCK is a categorical variable with 1 to 5 digits. It represents the property’s corresponding block code in a certain borough. For each borough, there is a valid block code range:

MANHATTAN 1 TO 2,255

BRONX 2,260 TO 5,958

BROOKLYN 1 TO 8,955

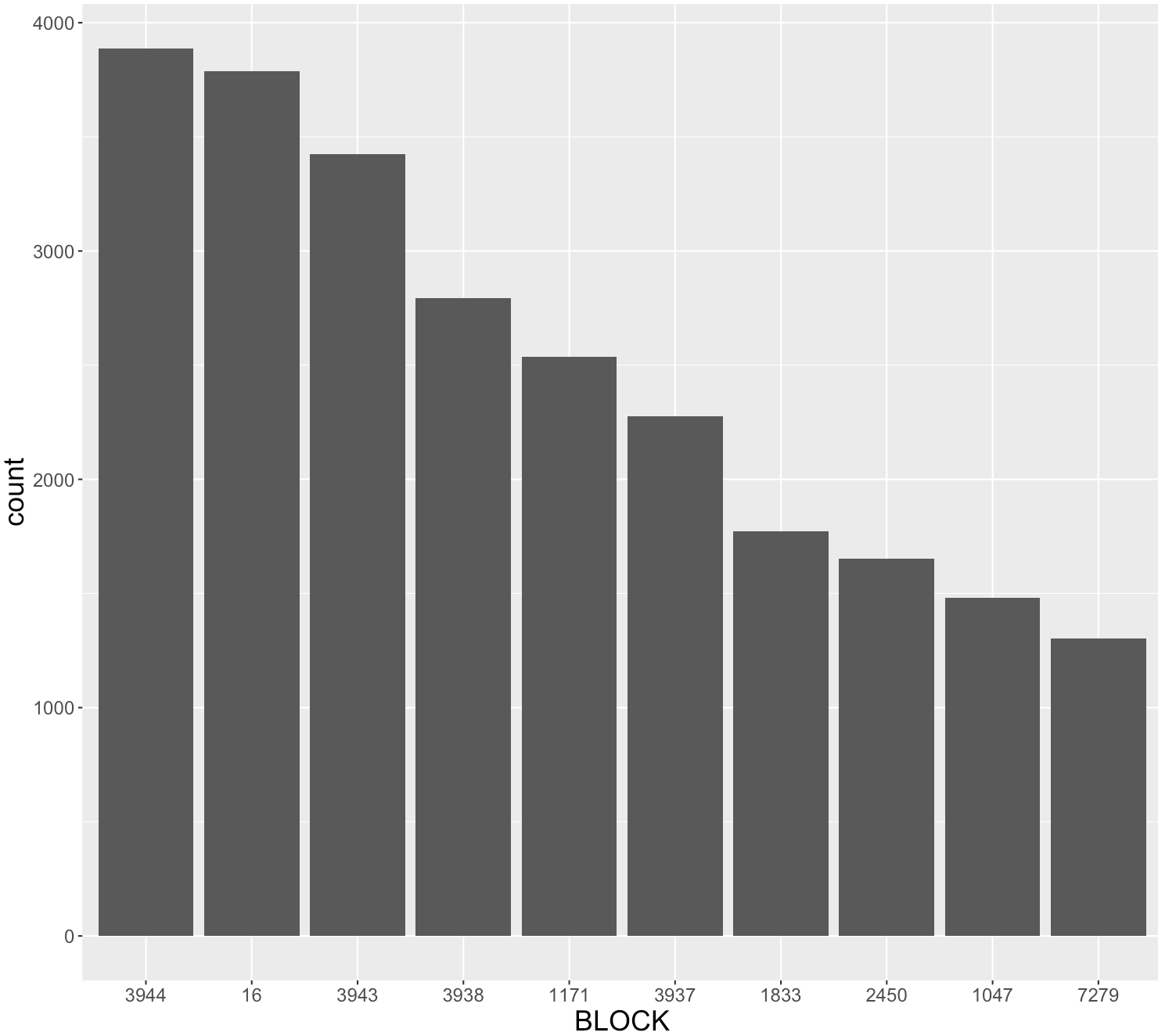
QUEENS 1 TO 16,350

STATEN ISLAND 1 TO 8,050

**Unique Values:**

BLOCK has 13949 unique values, ranging from 1 to 16350. No missing values. The top 10 most frequently appeared BLOCK code is shown below.

|  |  |
| --- | --- |
| BLOCK | Percentage(%) |
| 3944 | 0.37 |
| 16 | 0.36 |
| 3943 | 0.32 |
| 3938 | 0.27 |
| 1171 | 0.24 |
| 3937 | 0.21 |
| 1833 | 0.17 |
| 2450 | 0.16 |
| 1047 | 0.14 |
| 7279 | 0.12 |



**Field 4**

**Field Name:** LOT

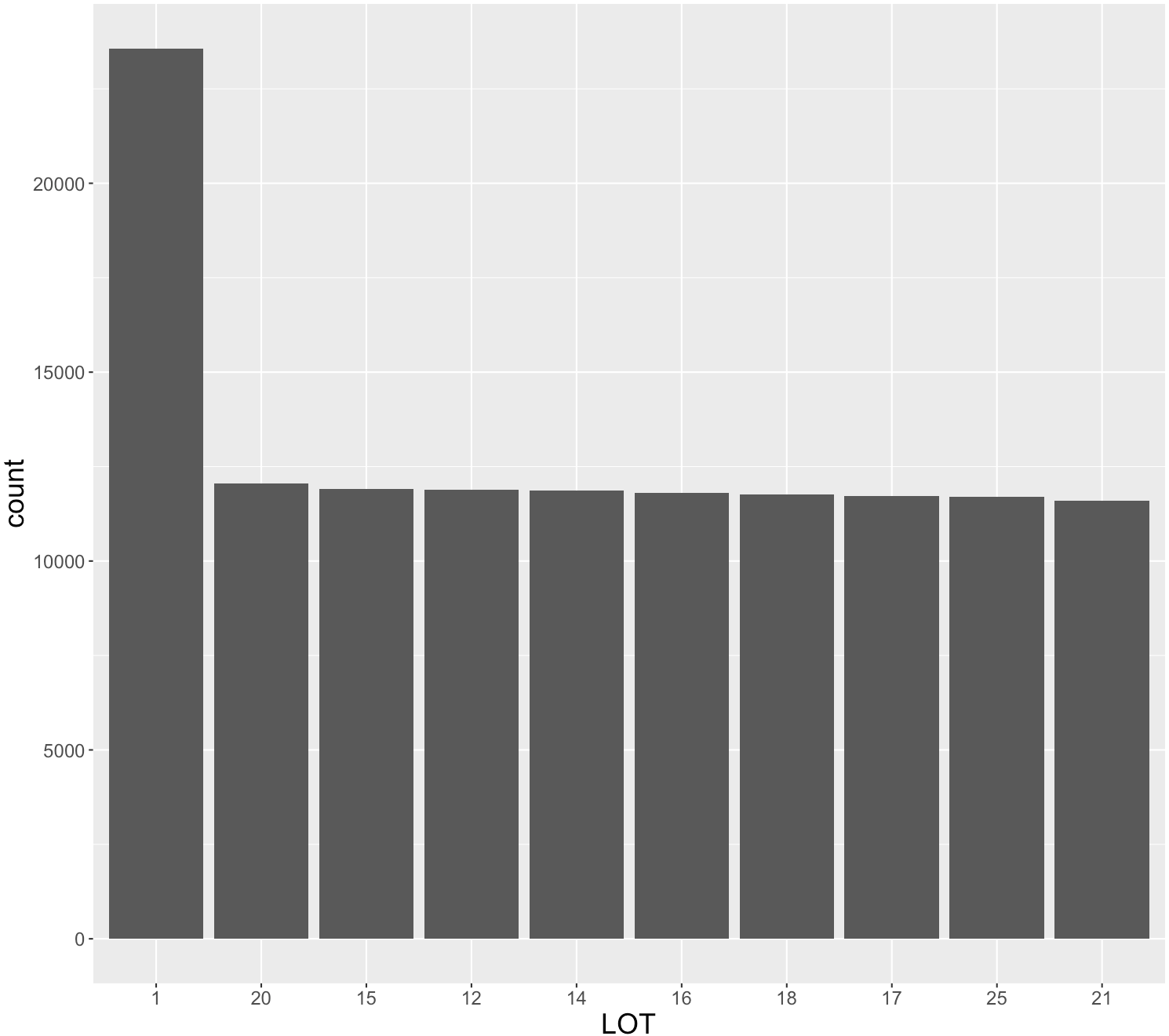
**Description:**

LOT is a categorical variable with 1 to 4 digits. It represents the property’s lot code within its borough and block.

**Unique Values:**

LOT has 6366 unique values, ranging from 1 to 9978. No missing values. The top 10 most frequently appeared LOT code is shown below.

|  |  |
| --- | --- |
| LOT | Percentage(%) |
| 1 | 2.25 |
| 20 | 1.15 |
| 15 | 1.14 |
| 12 | 1.13 |
| 14 | 1.13 |
| 16 | 1.13 |
| 18 | 1.12 |
| 17 | 0.12 |
| 25 | 1.12 |
| 21 | 1.11 |



**Field 5**

**Field Name:** EASEMENT

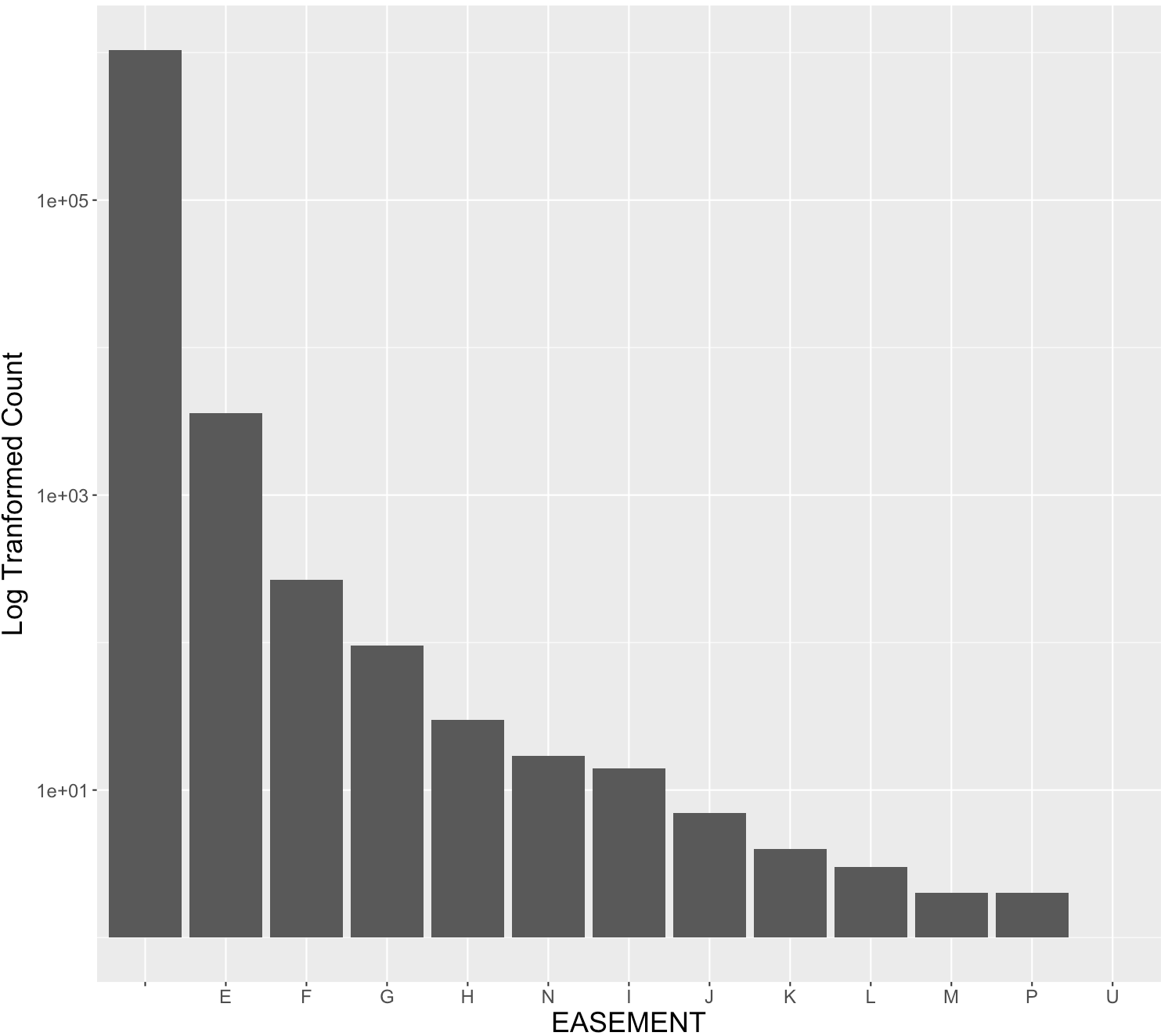
**Description:**

EASEMENT is a nominal categorical variable representing the property’s easement type.

**Unique values:**

EASEMENT has 13 levels – “”, “E”, “F”, “G”, “H”, “I”, “J”, “K”, “L”, “M”, “N”, “P”, “U”. The null value indicates the property does not have any special easement. No missing values exist. The sorted bar chart with log transformed y axis is shown below.

|  |  |
| --- | --- |
| Level | Counts |
|  | 1044532  (99.6%) |
| E | 3603 |
| F | 265 |
| G | 95 |
| H | 30 |
| I | 14 |
| J | 7 |
| K | 4 |
| L | 3 |
| M | 2 |
| N | 17 |
| P | 2 |
| U | 1 |



**Field 6**

**Field Name:** OWNER

**Description:**

OWNER is a text variable indicating the owner of the property.

**Unique Values:**

OWNER has 847055 unique values. In the OWNER field, 31081 properties have the value “”, indicating possible missing values. No missing values exist. The top 10 most frequently occurred OWNER names are:

|  |  |  |
| --- | --- | --- |
| Owner | Count | Percentage(%) |
|  | 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 |

**Field 7**

**Field Name:** BLDGCL

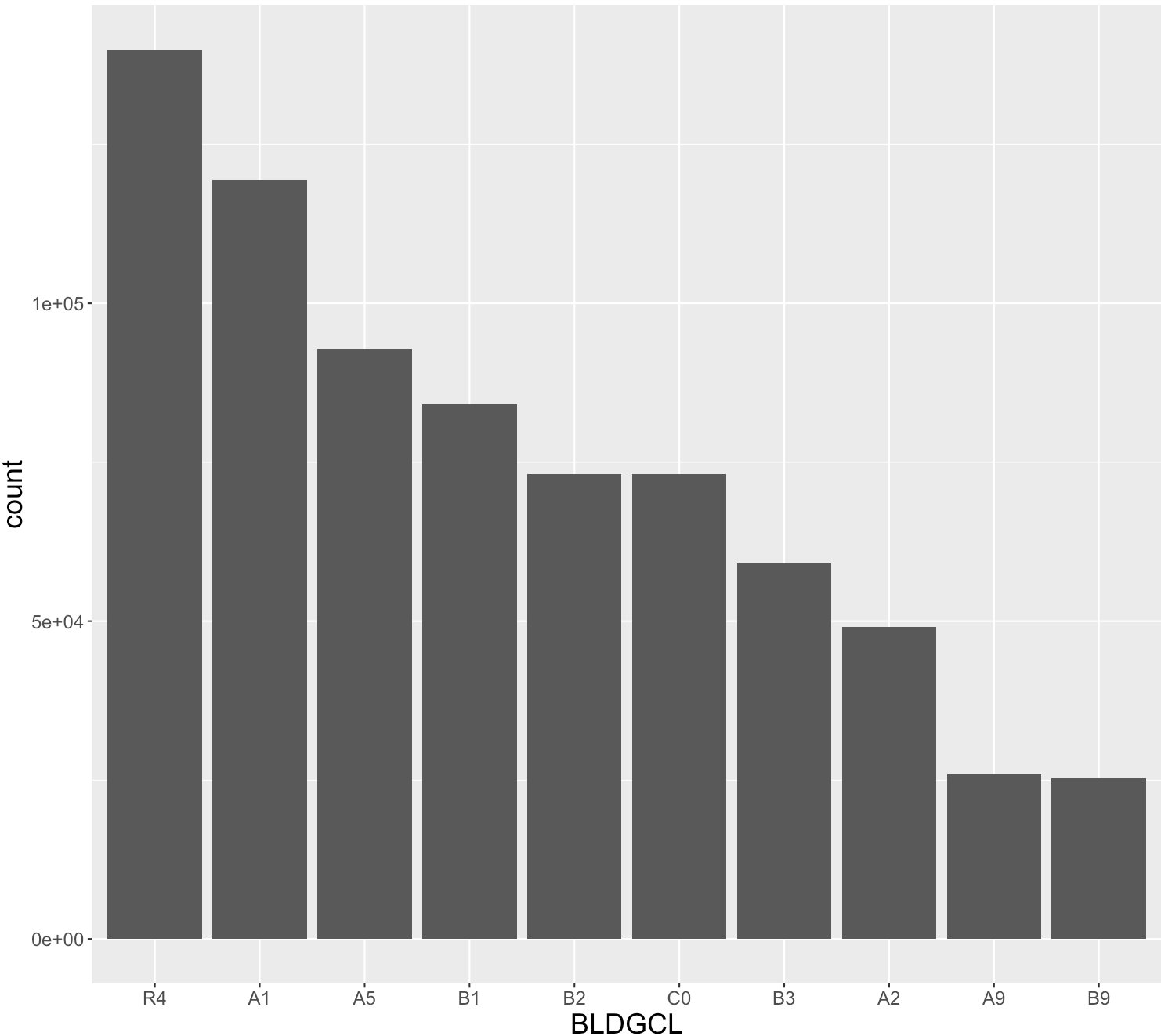
**Description:**

BLDGCL is a nominal categorical variable indicating the building class.

**Unique Values:**

BLDGCL has 200 unique levels. Each level has 2 digits – the first digit is a character from A to Z, the second digit is a number from 0 to 9. No missing values exist. The top 10 most frequently occurred BLDGCL is shown in below:

|  |  |
| --- | --- |
| BLDGCL | Percentage(%) |
| R4 | 13.3 |
| A1 | 11.4 |
| A5 | 8.9 |
| B1 | 8 |
| B2 | 7 |
| C0 | 8 |
| B3 | 5.6 |
| A2 | 4.7 |
| A9 | 2.5 |
| B9 | 2.4 |



**Field 8**

**Field Name:** TAXCLASS

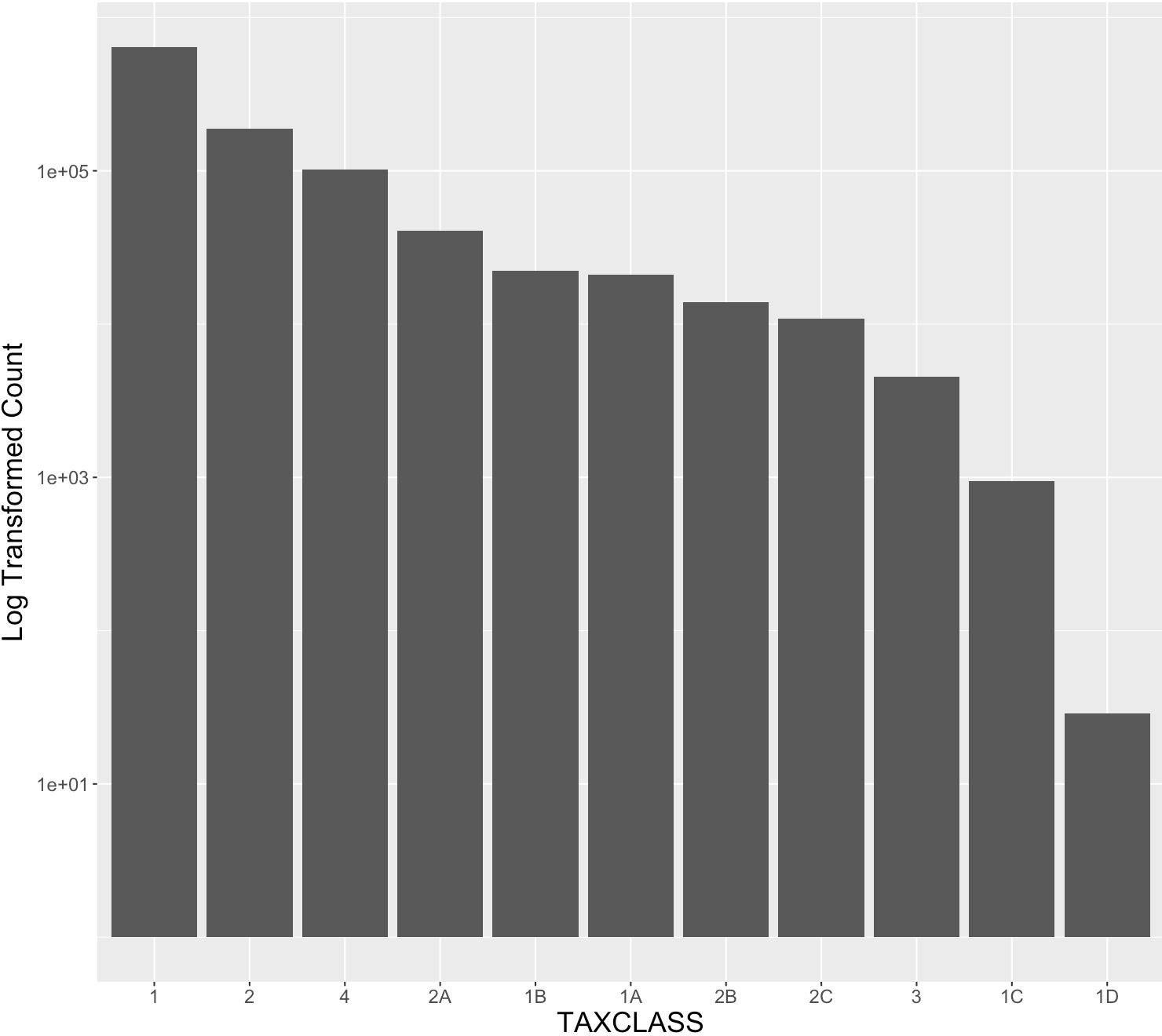
**Description:**

TAXCLASS is a categorical variable indicating the tax class of the property.

**Unique Values:**

TAXCLASS has 11 unique levels – “1”, “1A”, “1B”, “1C”, “1D”, “2”, “2A”, “2B”, “2C”, “3”, and “4”. No missing values exist. Sorted TAXCLASS levels are shown below:

|  |  |
| --- | --- |
| Level | Percentage(%) |
| 1 | 61.4 |
| 2 | 18.0 |
| 4 | 9.8 |
| 2A | 3.9 |
| 1B | 2.1 |
| 1A | 2.0 |
| 2B | 1.3 |
| 2C | 1.0 |
| 3 | 0.4 |
| 1C | 0.1 |
| 1D | 0.0 |



**Field 9**

**Field Name:** LTFRONT

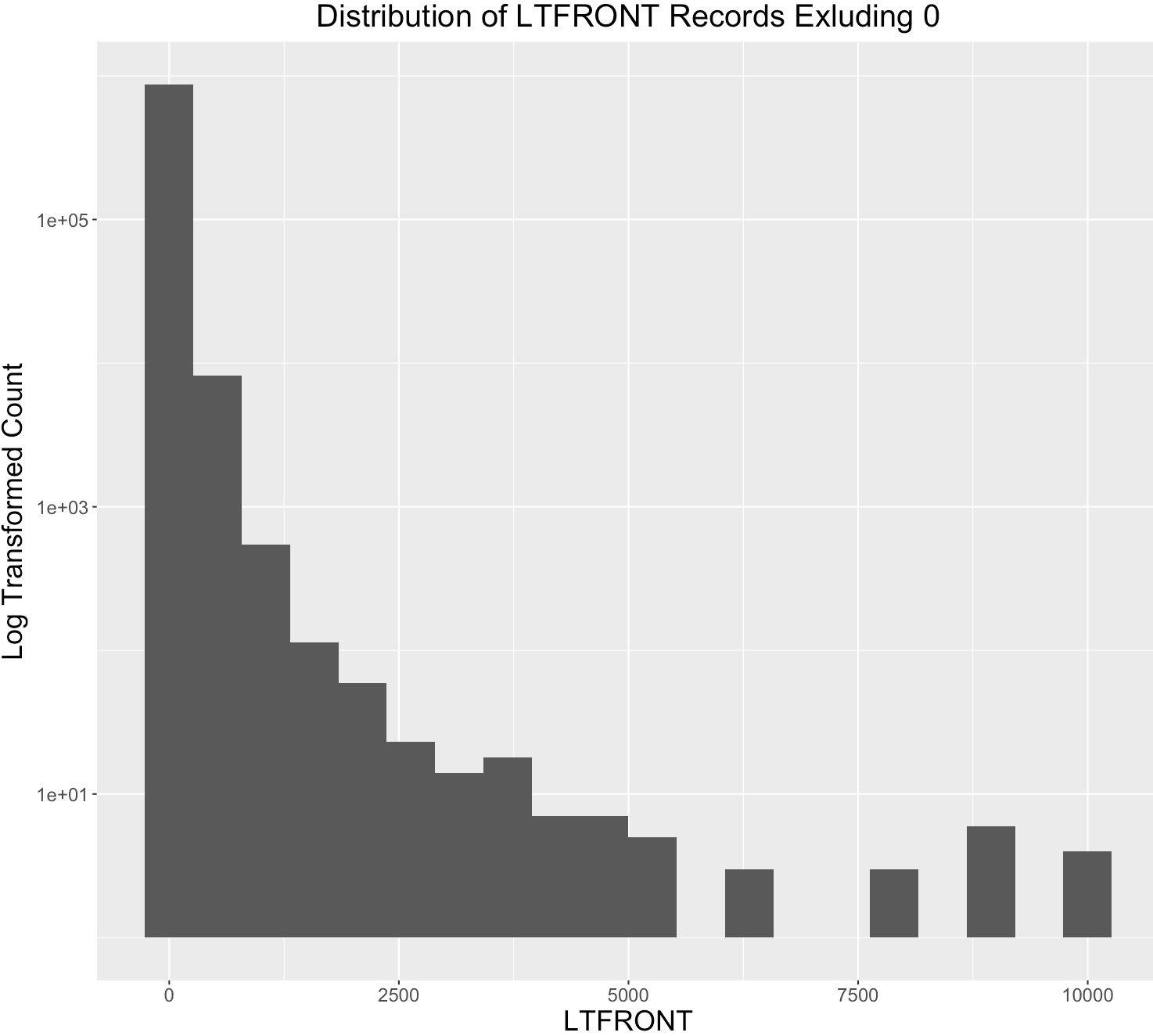
**Description:**

LTFRONT is a numeric variable representing the length of lot frontage in feet.

**Unique Values:**

LTFRONT has 1277 unique values ranging from 0 to 9999. No missing values exist. There are 168,867 records of 0 LTFRONT, and a LTFRONT of 0 may indicate missing value. The statistics and distribution excluding 0 records are shown as below.

|  |  |
| --- | --- |
| Minimum | 1 |
| Maximum | 9999 |
| Median | 25 |
| Mean | 43.12 |
| Mode | 20 |
| SD | 78.62 |



**Field 10**

**Field Name:** LTDEPTH

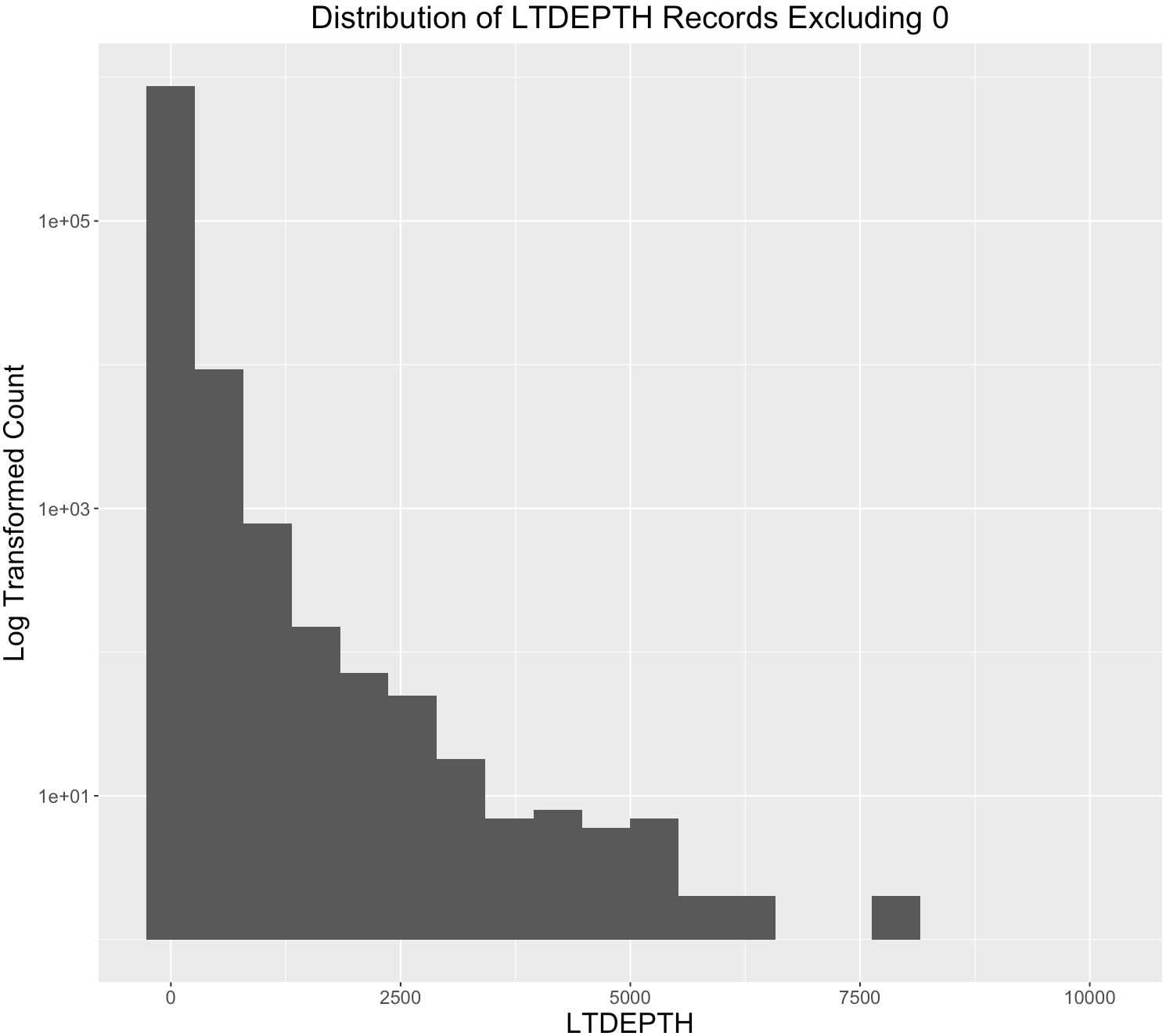
**Description:**

LTDEPTH is a numeric variable representing the length of lot depth in feet.

**Unique Values:**

LTDEPTH has 1336 unique values ranging from 0 to 9999. No missing values exist. There are 169,888 records of 0 LTDEPTH, and a LTDEPTH of 0 may indicate missing value. The statistics and distribution excluding 0 records are shown as below.

|  |  |
| --- | --- |
| Minimum | 1 |
| Maximum | 9999 |
| Median | 100 |
| Mean | 105.34 |
| Mode | 100 |
| SD | 70.71 |



**Field 11**

**Field Name:** STORIES

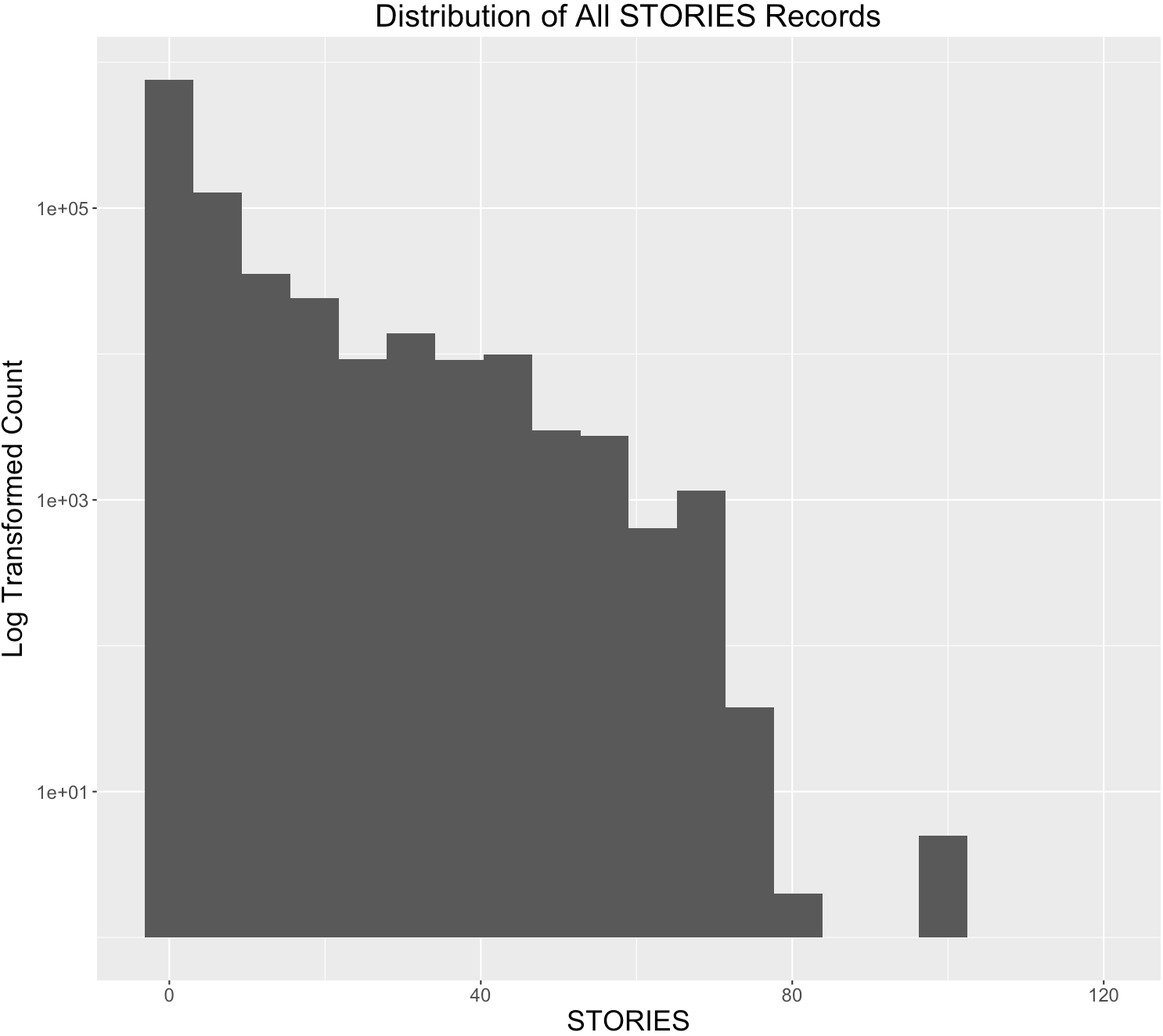
**Description:**

STORIES is a numeric variable representing the number of stories of the property.

**Unique Values:**

STORIES has 112 unique values ranging from 1 to 119. There are 52,142 missing values in the STORIES field. The statistics and distribution are shown as below.

|  |  |
| --- | --- |
| Minimum | 1 |
| Maximum | 119 |
| Median | 2 |
| Mean | 5.06 |
| Mode | 2 |
| SD | 8.43 |



**Field 12**

**Field Name:** FULLVAL

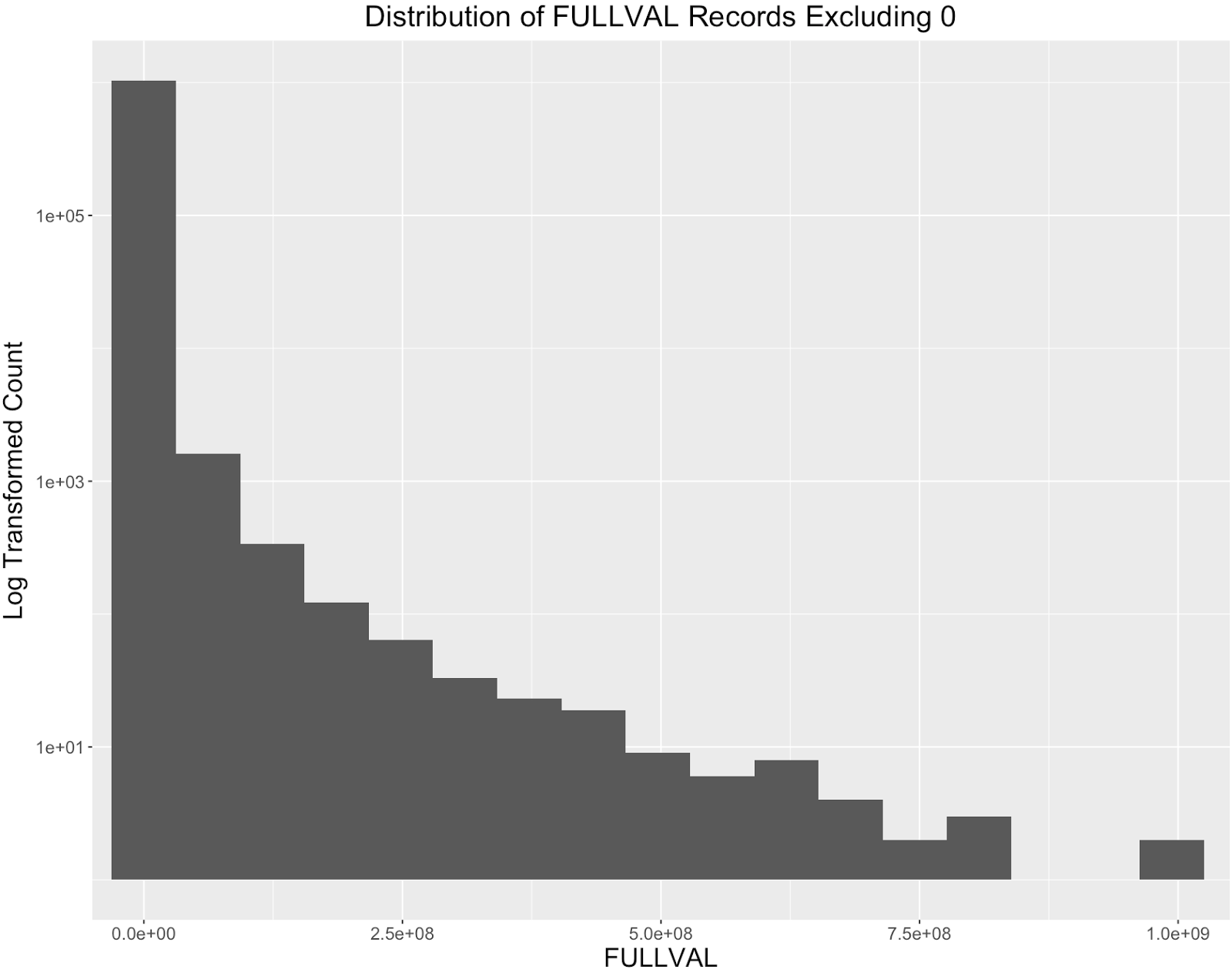
**Description:**

FULLVAL is a numeric variable representing the full value of the property.

**Unique Values:**

FULLVAL has 108277 unique values ranging from 0 to about 6,000,000,000. There are 12,762 properties with the FULLVAL of 0 in the dataset. No missing values exist. The statistics and distribution excluding 0 records are shown as below.

|  |  |
| --- | --- |
| Minimum | 4 |
| Maximum | 6.15E+09 |
| Median | 45000 |
| Mean | 8.91E+05 |
| Mode | 502000 |
| SD | 1.17E+07 |



**Field 13**

**Field Name:** AVLAND

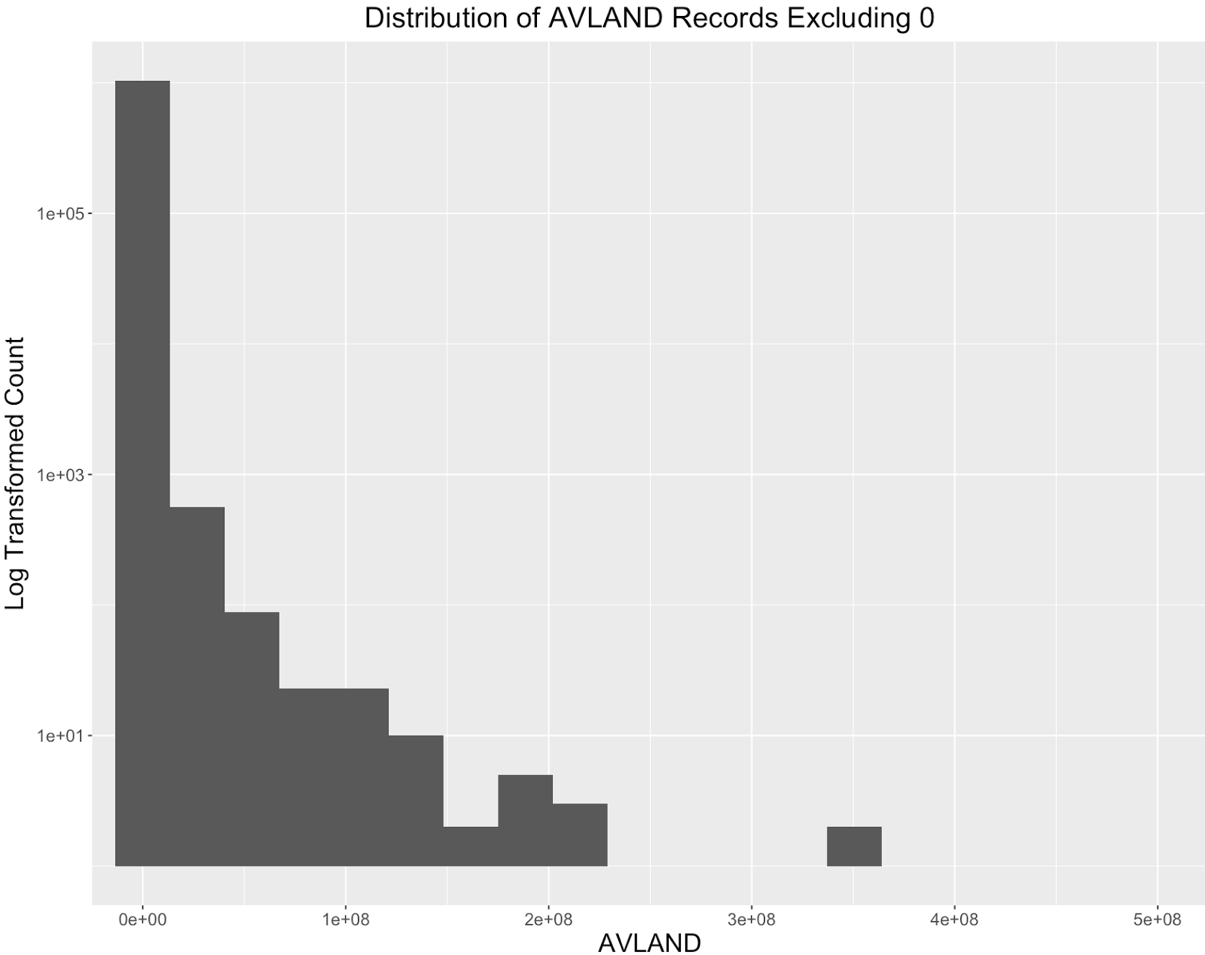
**Description:**

AVLAND is a numeric variable representing the assessed value of the land.

**Unique Values:**

AVLAND has 70,529 unique values ranging from 0 to about 2,700,000,000. There are 12,764 properties with the AVLAND of 0 in the dataset. No missing values exist. The statistics and distribution excluding 0 records are shown as below.

|  |  |
| --- | --- |
| Minimum | 1 |
| Maximum | 2.67E+09 |
| Median | 13751 |
| Mean | 86054.72 |
| Mode | 45000 |
| SD | 4.10E+06 |



**Field 14**

**Field Name:** AVTOT

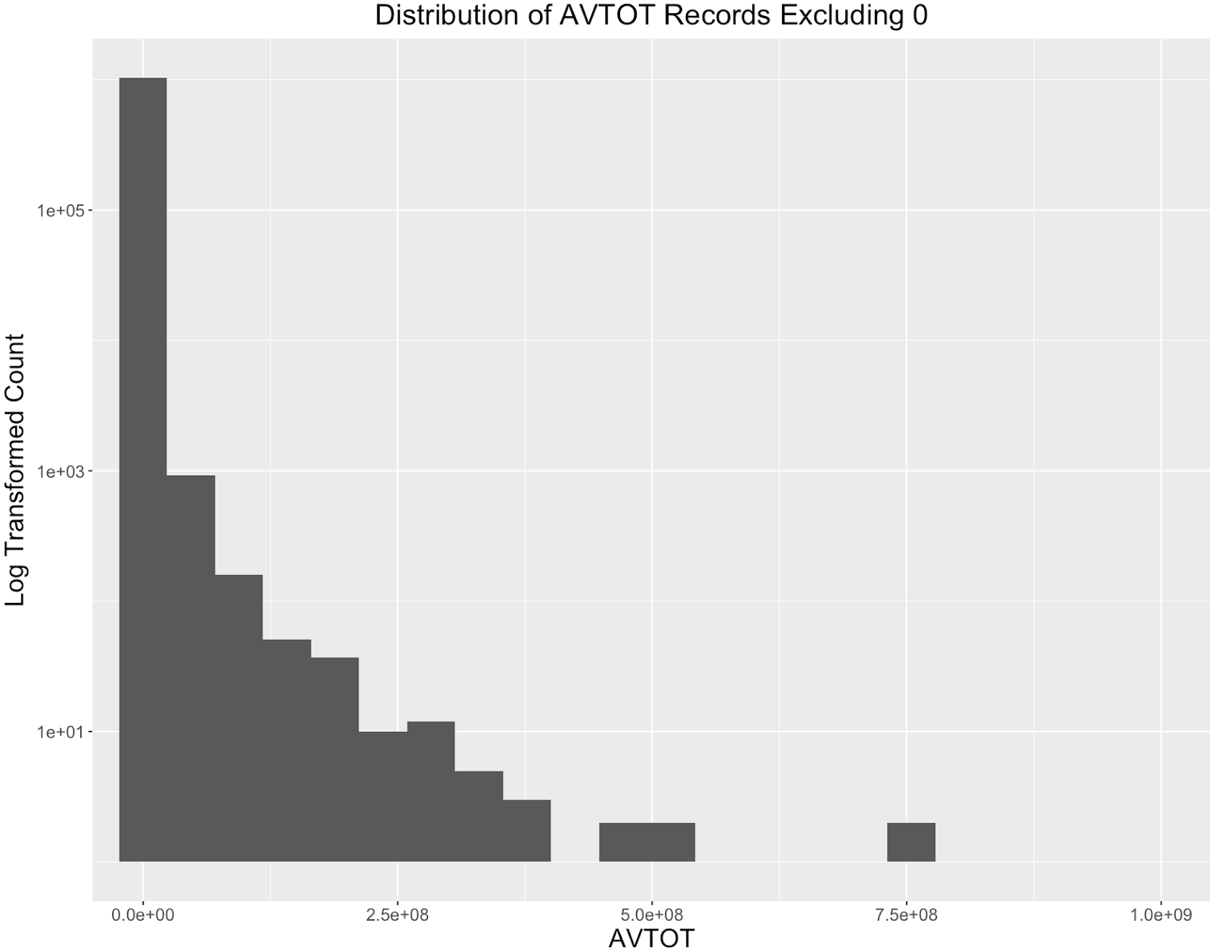
**Description:**

AVTOT is a numeric variable representing the assessed total value of the property.

**Unique Values:**

AVTOT has 112294 unique values ranging from 0 to about 4,700,000,000. There are 12,762 properties with the AVTOT of 0 in the dataset. No missing values exist. The statistics and distribution excluding 0 records are shown as below.

|  |  |
| --- | --- |
| Minimum | 1 |
| Maximum | 4.67E+09 |
| Median | 25560 |
| Mean | 233601.3 |
| Mode | 16588 |
| SD | 6.99E+06 |



**Field 15**

**Field Name:** EXLAND

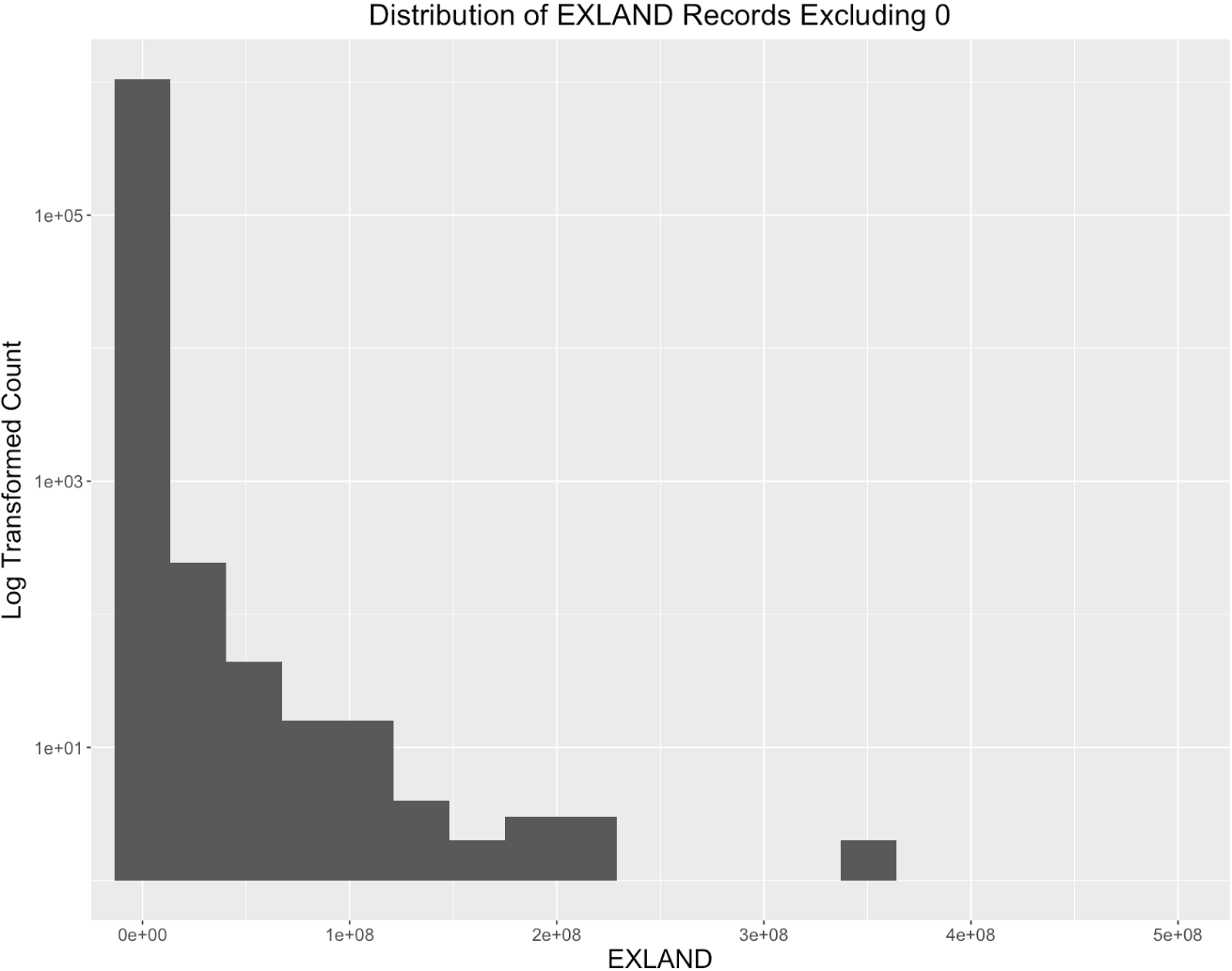
**Description:**

EXLAND is a numeric variable representing the value of the exempt land. The value of EXLAND is always smaller or equal to AVLAND.

**Unique Values:**

EXLAND has 33186 unique values ranging from 0 to about 2,700,000,000. There are 484,224 properties with the EXLAND of 0 in the dataset. No missing values exist. The statistics and distribution excluding 0 records are shown as below.

|  |  |
| --- | --- |
| Minimum | 1 |
| Maximum | 2.67E+09 |
| Median | 1620 |
| Mean | 68397.01 |
| Mode | 1620 |
| SD | 5485336 |



**Field 16**

**Field Name:** EXTOT

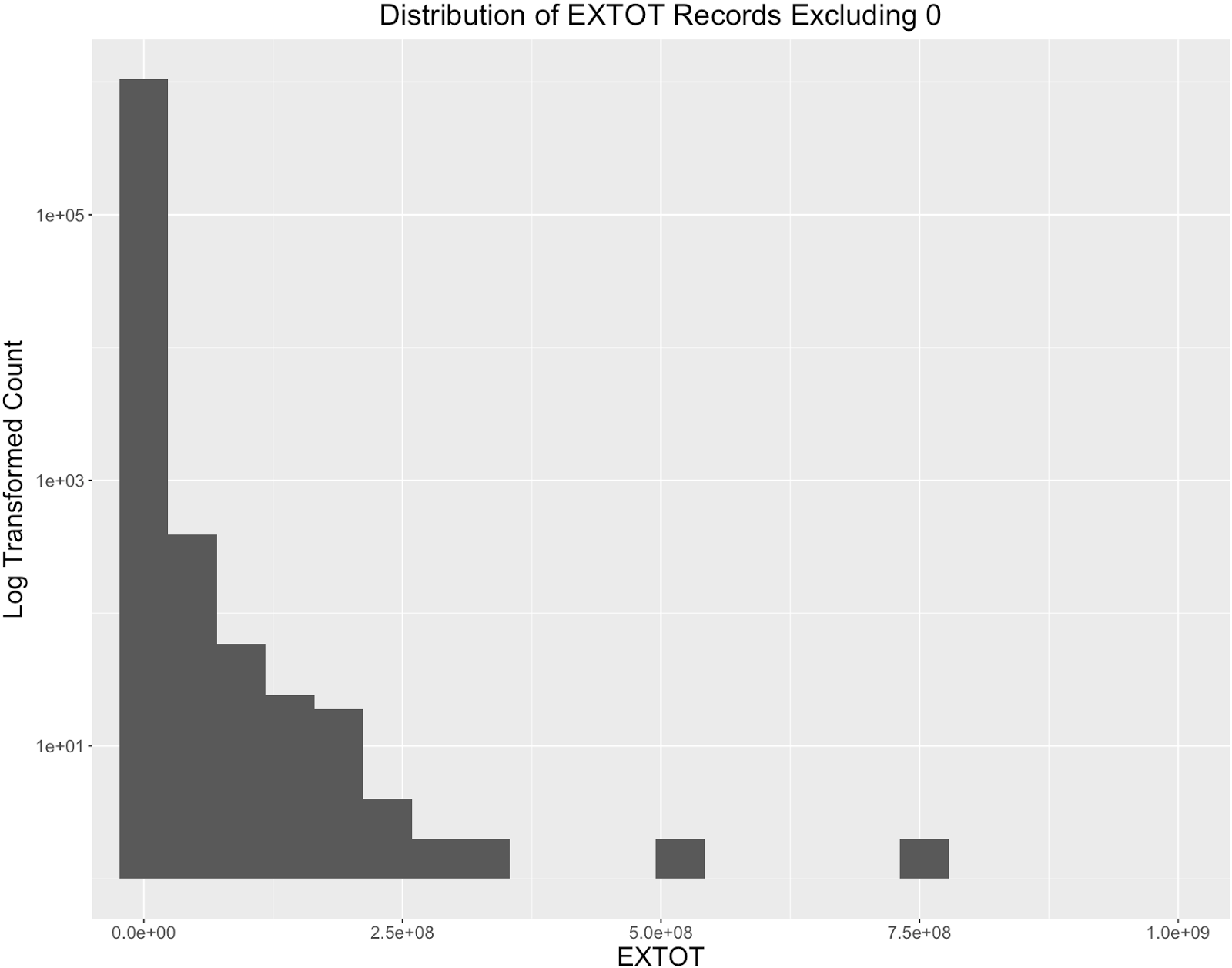
**Description:**

EXTOT is a numeric variable representing the total value of the exempt property. The value of EXTOT is always smaller or equal to AVTOT.

**Unique Values:**

EXTOT has 63805 unique values ranging from 0 to about 4,700,000,000. There are 425,999 properties with the EXTOT of 0 in the dataset. No missing values exist. . The statistics and distribution excluding 0 records are shown as below.

|  |  |
| --- | --- |
| Minimum | 1 |
| Maximum | 4.67E+09 |
| Median | 1620 |
| Mean | 155867.1 |
| Mode | 1620 |
| SD | 8536636 |



**Field 17**

**Field Name:** EXCD1

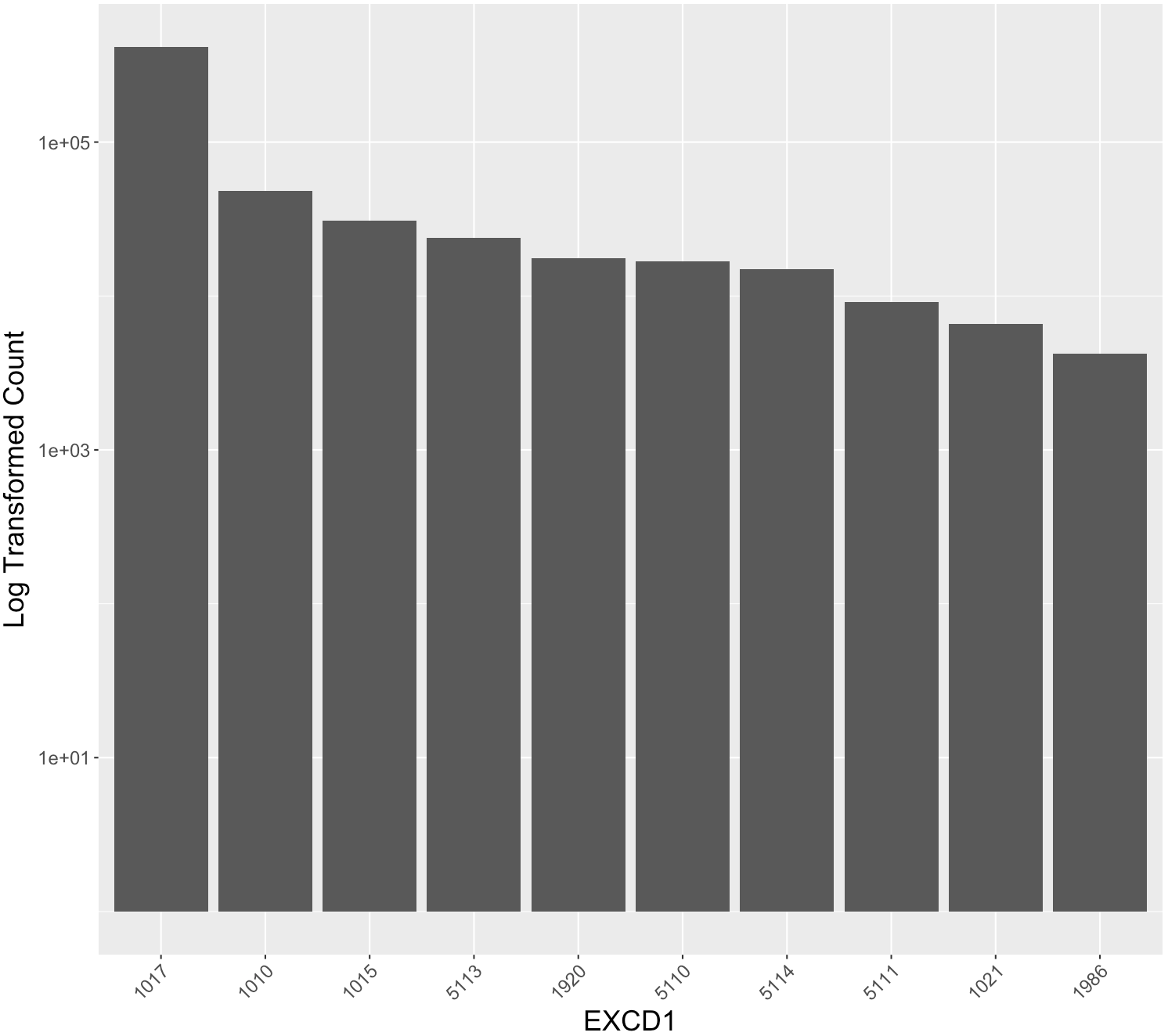
**Description:**

EXCD1 is a categorical variable, possibly representing the code for the exempt reasons.

**Unique Values:**

EXTOT has 130 levels, taking 4-digit numbers from 1010 to 7170. There are 425,933 missing values exist. For all properties that hold missing value in EXCD1, their values in both EXLAND and EXTOT fields are 0. The top 20 most frequently occurred EXCD1 values are shown below.

|  |  |
| --- | --- |
| EXCD1 | Percentage(%) |
| 1017 | 39.5 |
| 1010 | 4.6 |
| 1015 | 2.9 |
| 5113 | 2.3 |
| 1920 | 1.7 |
| 5110 | 1.6 |
| 5114 | 1.4 |
| 5111 | 8.7 |
| 1021 | 6.3 |
| 1986 | 4.0 |



**Field 18**

**Field Name:** STADDR

**Description:**

STADDR is a text variable, representing the street address of the property.

**Unique values:**

STADDR has 820,638 unique values. No missing values exist in this field. However, there are 641 records with the value of “” (null) in the STADDR field, indicating missing values.

|  |  |
| --- | --- |
| Address | Counts |
| 501 SURF AVENUE | 902 |
| 330 EAST 38 STREET | 817 |
| 322 WEST 57 STREET | 720 |
| 155 WEST 68 STREET | 671 |
| 20 WEST 64 STREET | 657 |
| 1 IRVING PLACE | 650 |
|  | 641 |
| 220 RIVERSIDE BOULEVARD | 628 |
| 360 FURMAN STREET | 599 |
| 200 EAST 66 STREET | 585 |

The top 10 most frequently occurred STADDR values are:

**Field 19**

**Field Name:** ZIP

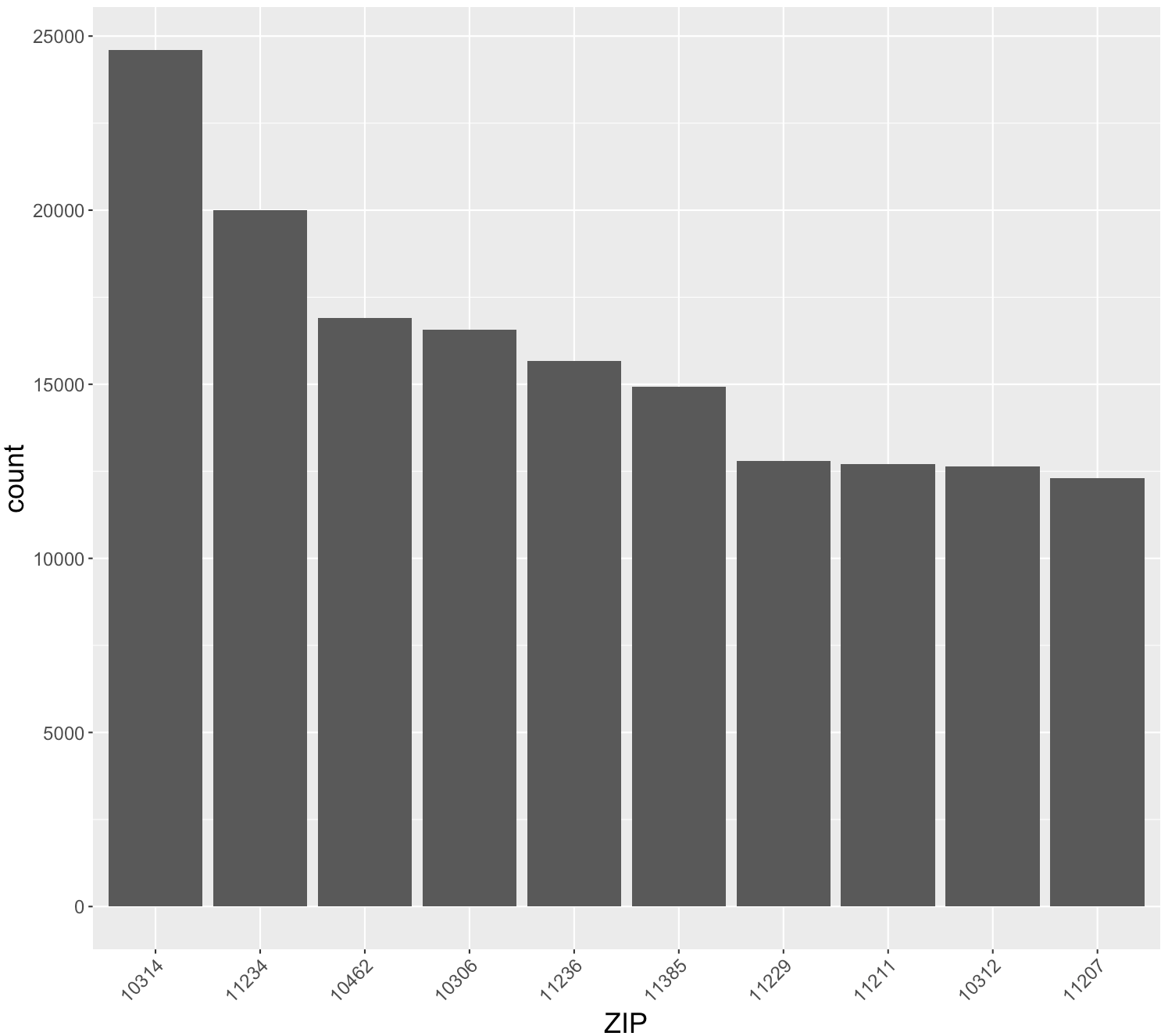
**Description:**

ZIP is a categorical variable, recording the zipcode of the property.

**Unique Values:**

ZIP has 197 unique values and 26,356 missing values. There are three obvious anomaly records with ZIP of 33803, which should be in Florida. The top 20 most frequently occurred ZIP values are:

|  |  |
| --- | --- |
| ZIP | Percentage(%) |
| 10314 | 2.3 |
| 11234 | 1.9 |
| 10462 | 1.6 |
| 10306 | 1.6 |
| 11236 | 1.5 |
| 11385 | 1.4 |
| 11229 | 1.2 |
| 11211 | 1.2 |
| 10312 | 1.2 |
| 11207 | 1.2 |



**Field 20**

**Field Name:** EXMPTCL

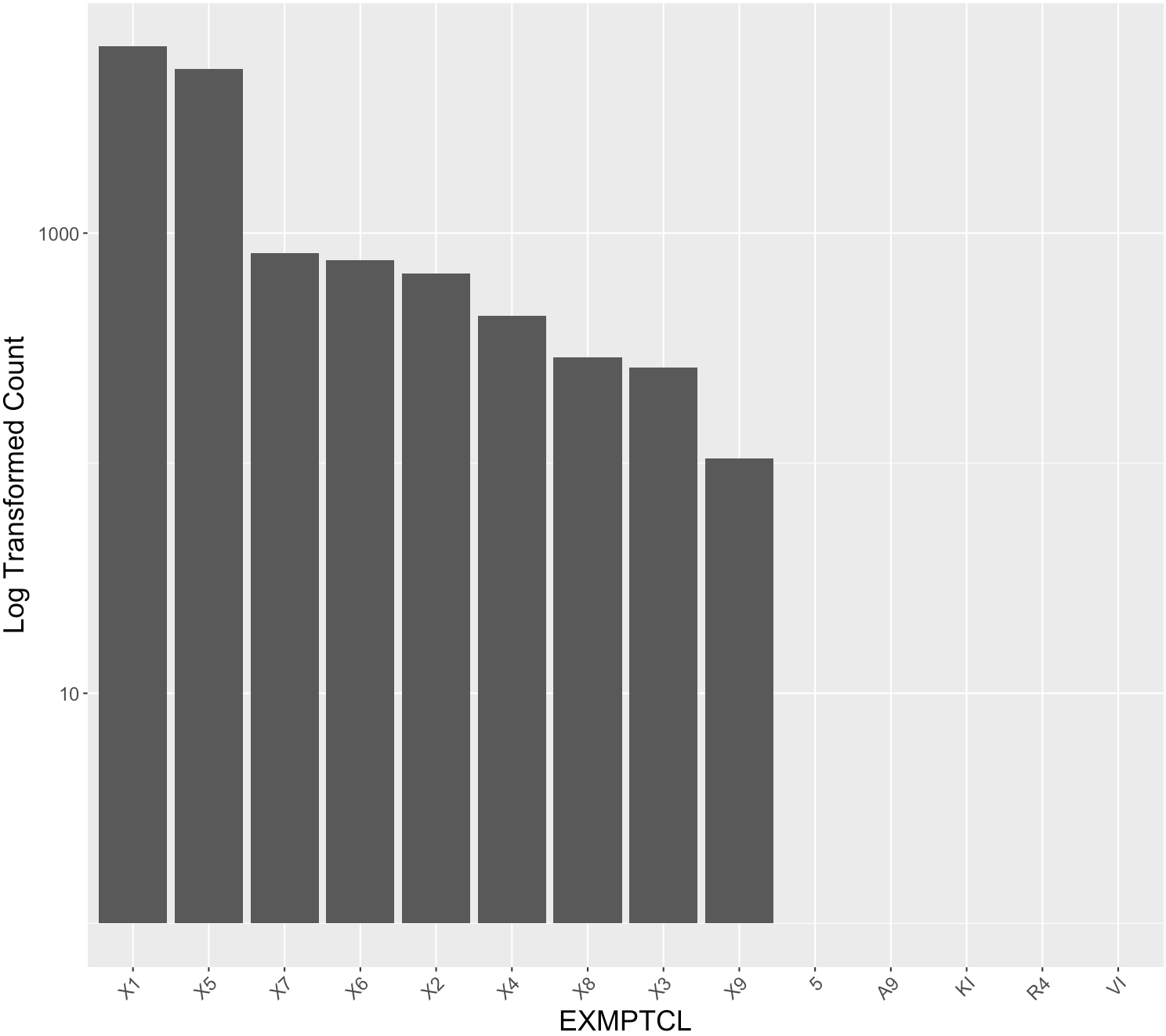
**Description:**

EXMPTCL is a nominal categorical variable, representing the exempt class, which is used for exempt properties only.

**Unique Values:**

EXMPTCL has 15 levels- “”, “5”, “A9”, “KI”, “R4”, “VI”, “X1”, “X2”, “X3”, “X4”, “X5”, “X6”, “X7”, “X8”, and “X9”. 1,033,583 properties take “” value in EXMPTCL. No missing values exist. The sorted bar chart is shown below.

|  |  |
| --- | --- |
| EXMPTCL | Percentage(%) |
| X1 | 43.41 |
| X5 | 34.41 |
| X7 | 5.46 |
| X6 | 5.07 |
| X2 | 4.44 |
| X5 | 2.92 |
| X8 | 1.93 |
| X3 | 1.73 |
| X9 | 0.70 |
| 5 | 0.01 |
| A9 | 0.01 |
| KI | 0.01 |
| R4 | 0.01 |
| VI | 0.01 |



**Field 21**

**Field Name:** BLDFRONT

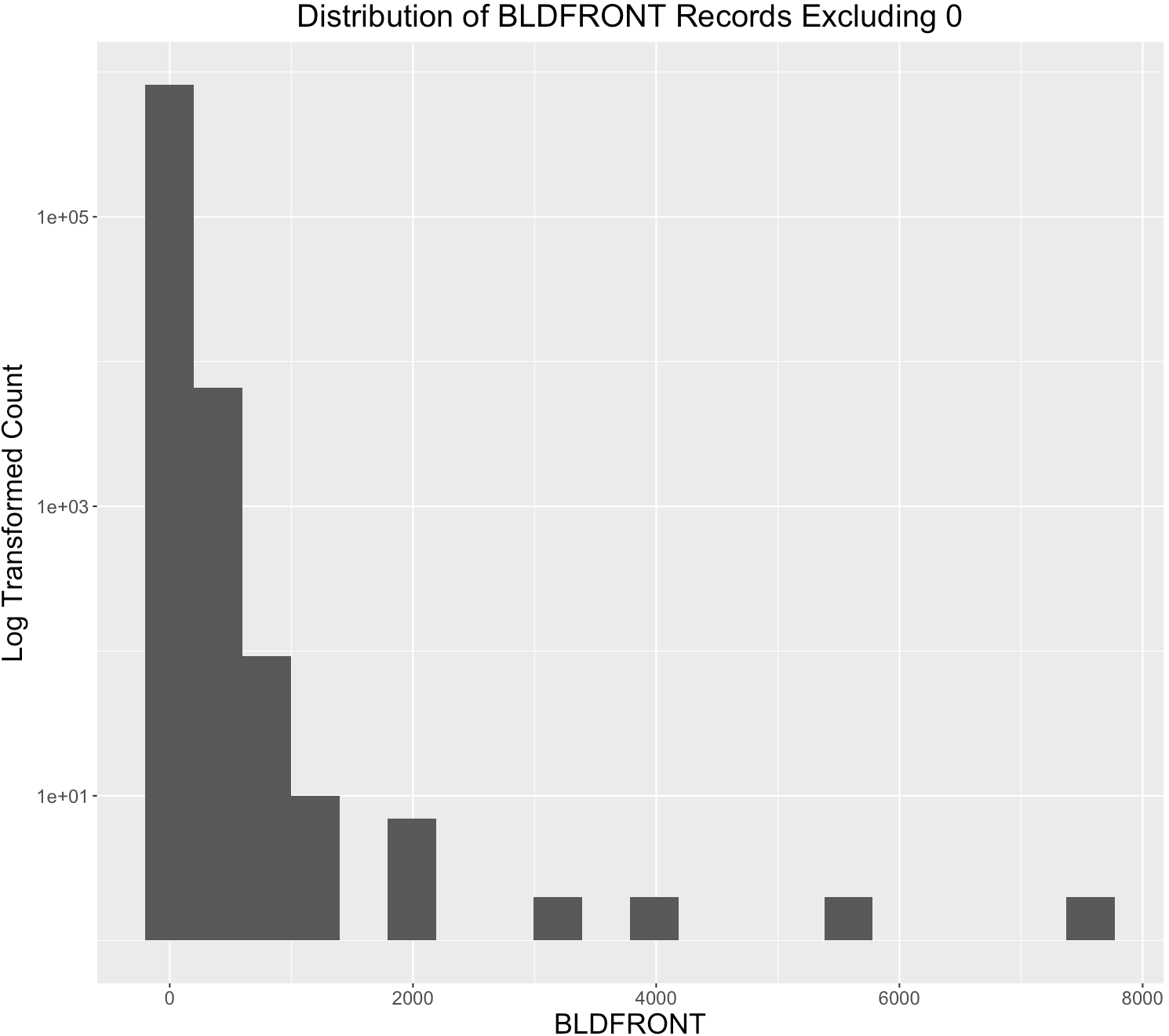
**Description:**

BLDFRONT is a numeric variable representing the length of building frontage in feet.

**Unique Values:**

BLDFRONT has 610 unique values ranging from 0 to 7575. No missing values exist. However, there are 224,661 records with value 0, which could be in fact missing values. The statistics and distribution excluding all records with 0 BLDFRONT are shown as below.

|  |  |
| --- | --- |
| Minimum | 1 |
| Maximum | 7575 |
| Median | 20 |
| Mean | 29.29 |
| Mode | 20 |
| SD | 38.03 |



**Field 22**

**Field Name:** BLDDEPTH

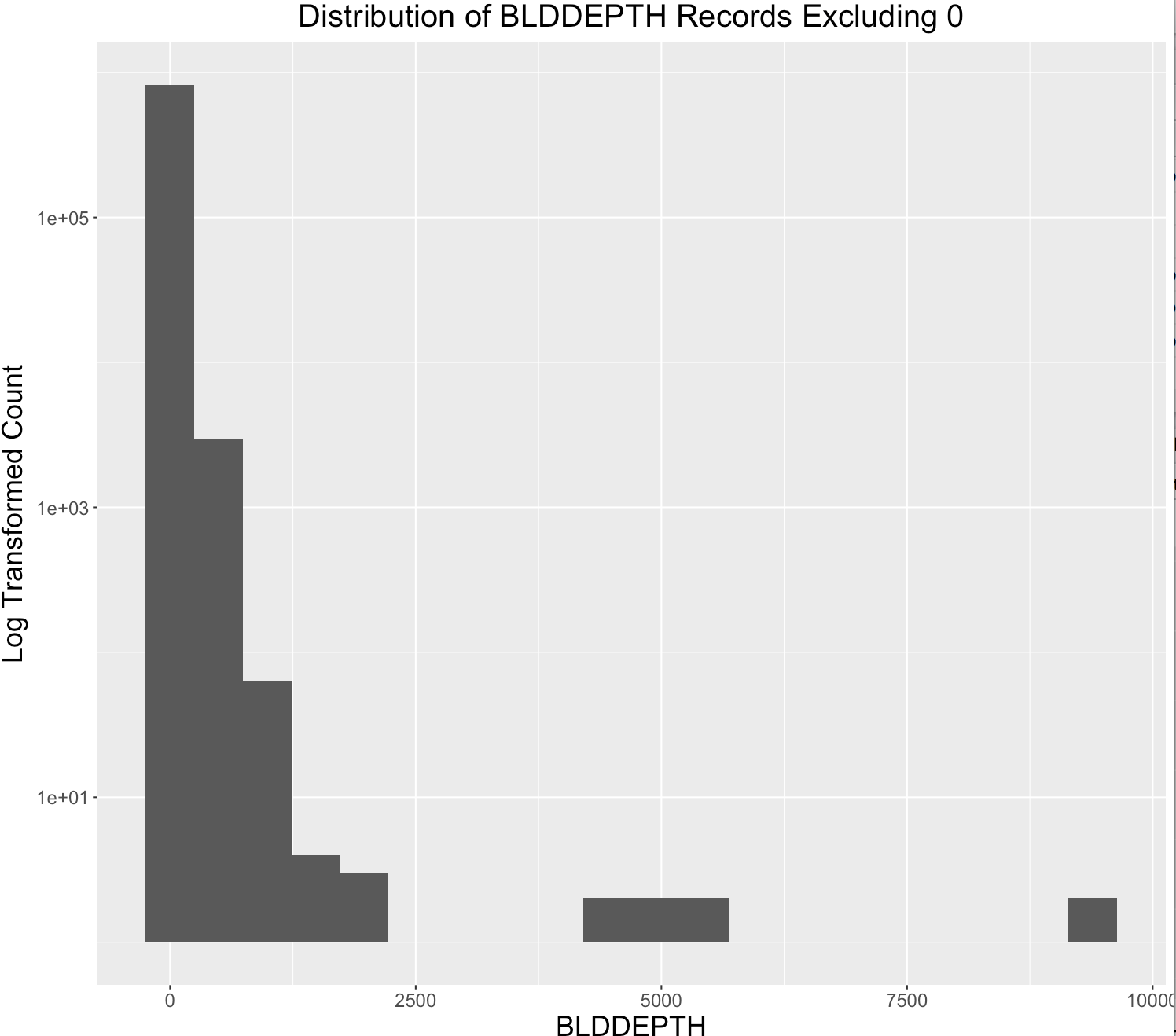
**Description:**

BLDDEPTH is a numeric variable representing the length of building depth in feet.

**Unique Values:**

BLDDEPTH has 620 unique values ranging from 0 to 9393. No missing values exist. However, there are 224,699 records with value 0, which could be in fact missing values. The statistics and distribution excluding all records with 0 BLDFRONT are shown as below.

|  |  |
| --- | --- |
| Minimum | 1 |
| Maximum | 9393 |
| Median | 44 |
| Mean | 51.00 |
| Mode | 40 |
| SD | 42.42 |



**Field 23**

**Field Name:** AVLAND2

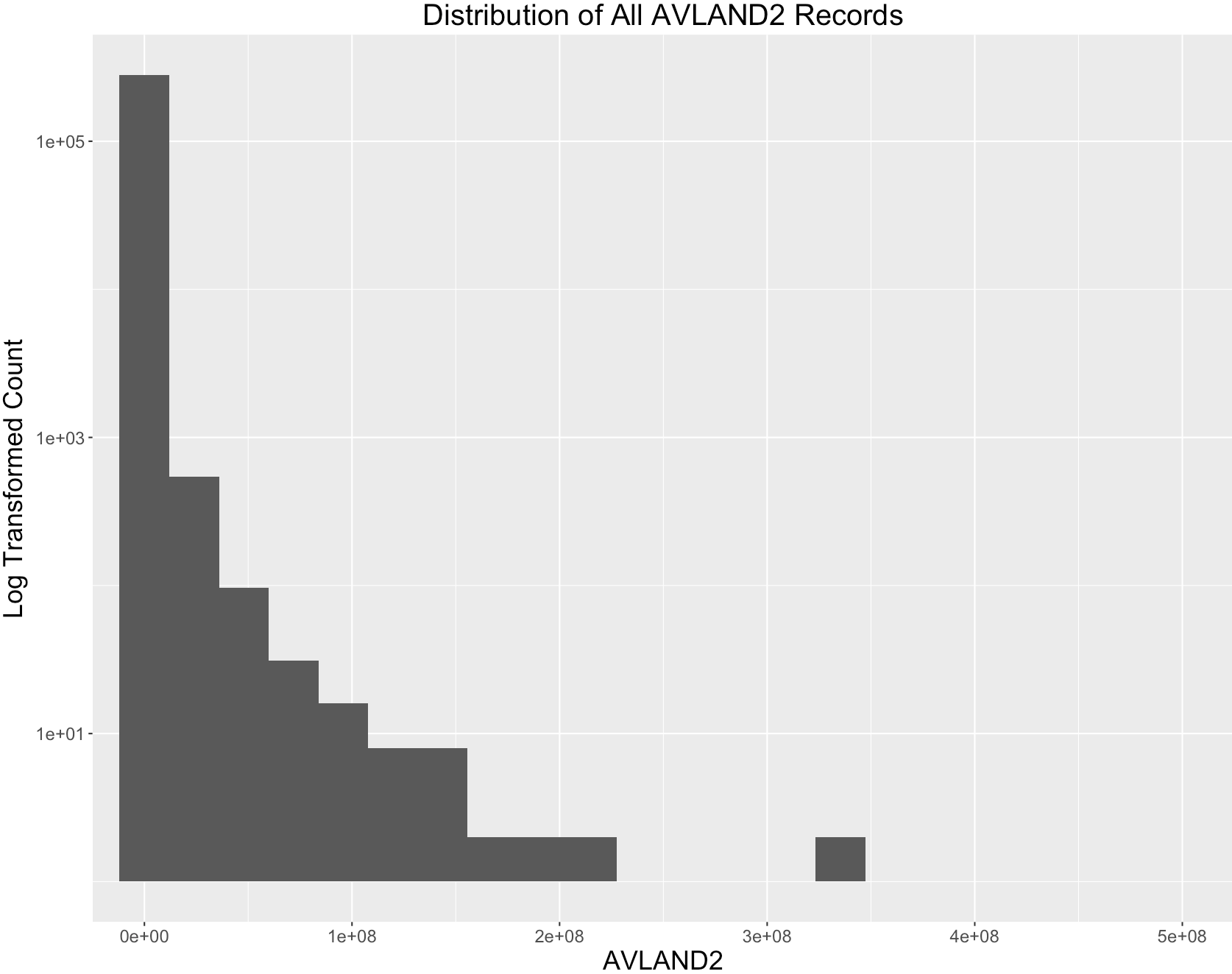
**Description:**

AVLAND2 is a numeric variable representing the assessed value of the land. It could be the updated assessed value compared to AVLAND. Most values of AVLAND2 are lower than their corresponding values of AVLAND.

**Unique Values:**

AVLAND has 58,170 unique values ranging from 3 to about 2,300,000,000. There are 767,609 records of missing values in the AVLAND2 field. The statistics is shown as below.

|  |  |
| --- | --- |
| Minimum | 3 |
| Maximum | 2.37E+09 |
| Median | 20059 |
| Mean | 2.46E+05 |
| Mode | 2408 |
| SD | 6.20E+06 |

****

**Field 24**

**Field Name:** AVTOT2

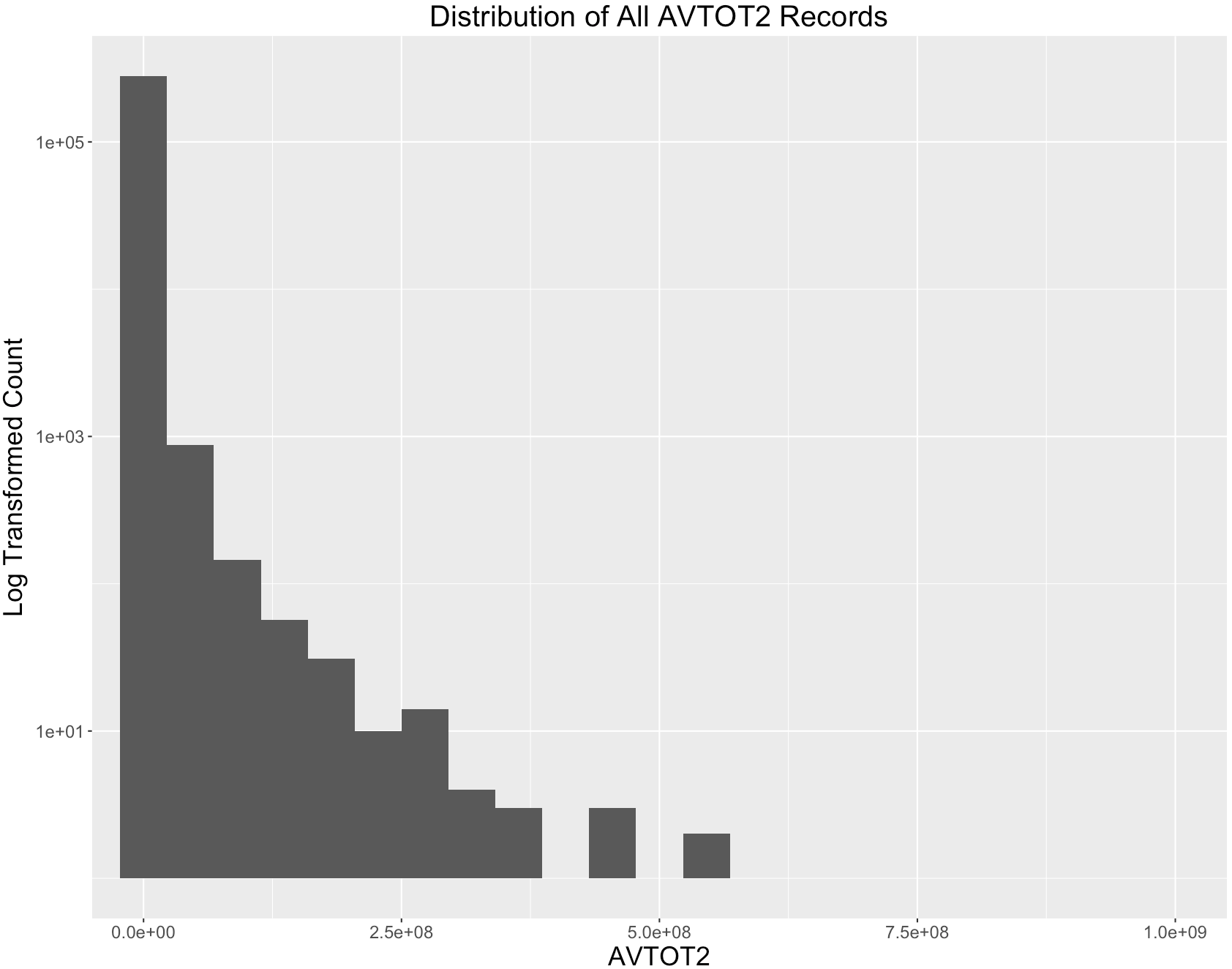
**Description:**

AVTOT2 is a numeric variable representing the assessed total value of the property. It could be the updated assessed value compared to AVTOT. Most AVTOT2 values are smaller than or equal to their corresponding AVTOT value.

**Unique Values:**

AVTOT2 has 110,891 unique values ranging from 3 to about 4,500,000,000. There are 767,603 missing values in the AVTOT2 field. The statistics is shown as below.

|  |  |
| --- | --- |
| Minimum | 3 |
| Maximum | 4.50E+09 |
| Median | 80010 |
| Mean | 7.16E+05 |
| Mode | 750 |
| SD | 1.17E+07 |

****

**Field 25**

**Field Name:** EXLAND2

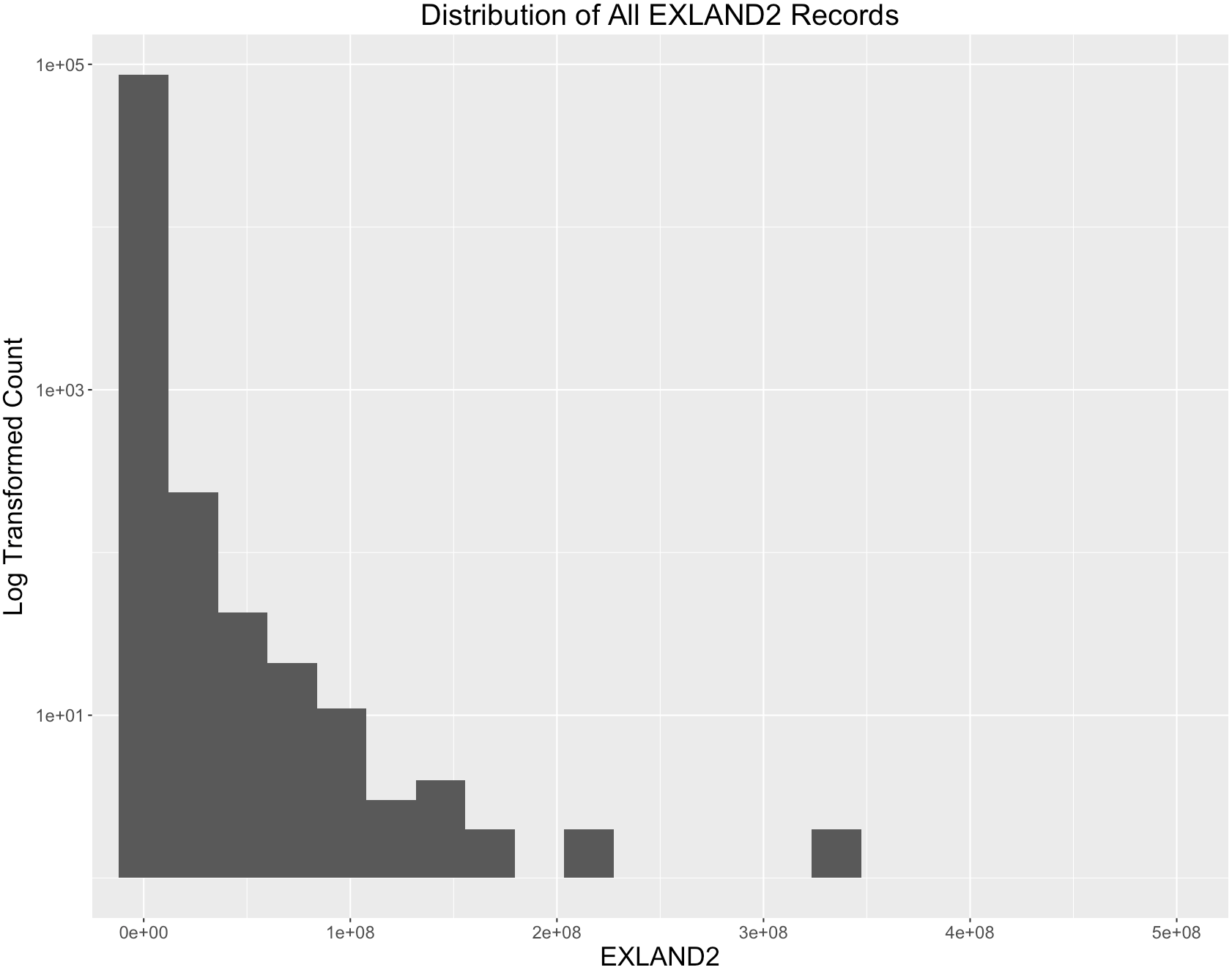
**Description:**

EXLAND2 is a numeric variable representing the value of the exempt land. It could be the updated assessed value compared to EXLAND. Most EXLAND2 values are lower than their corresponding EXLAND values.

**Unique Values:**

EXLAND2 has 21,997 unique values ranging from 7 to about 2,400,000,000. There are 961,900 missing values in the EXLAND2 field. The statistics is shown as below.

|  |  |
| --- | --- |
| Minimum | 7 |
| Maximum | 4.50E+09 |
| Median | 37116 |
| Mean | 6.58E+05 |
| Mode | 2090 |
| SD | 1.61E+07 |

****

**Field 26**

**Field Name:** EXTOT2

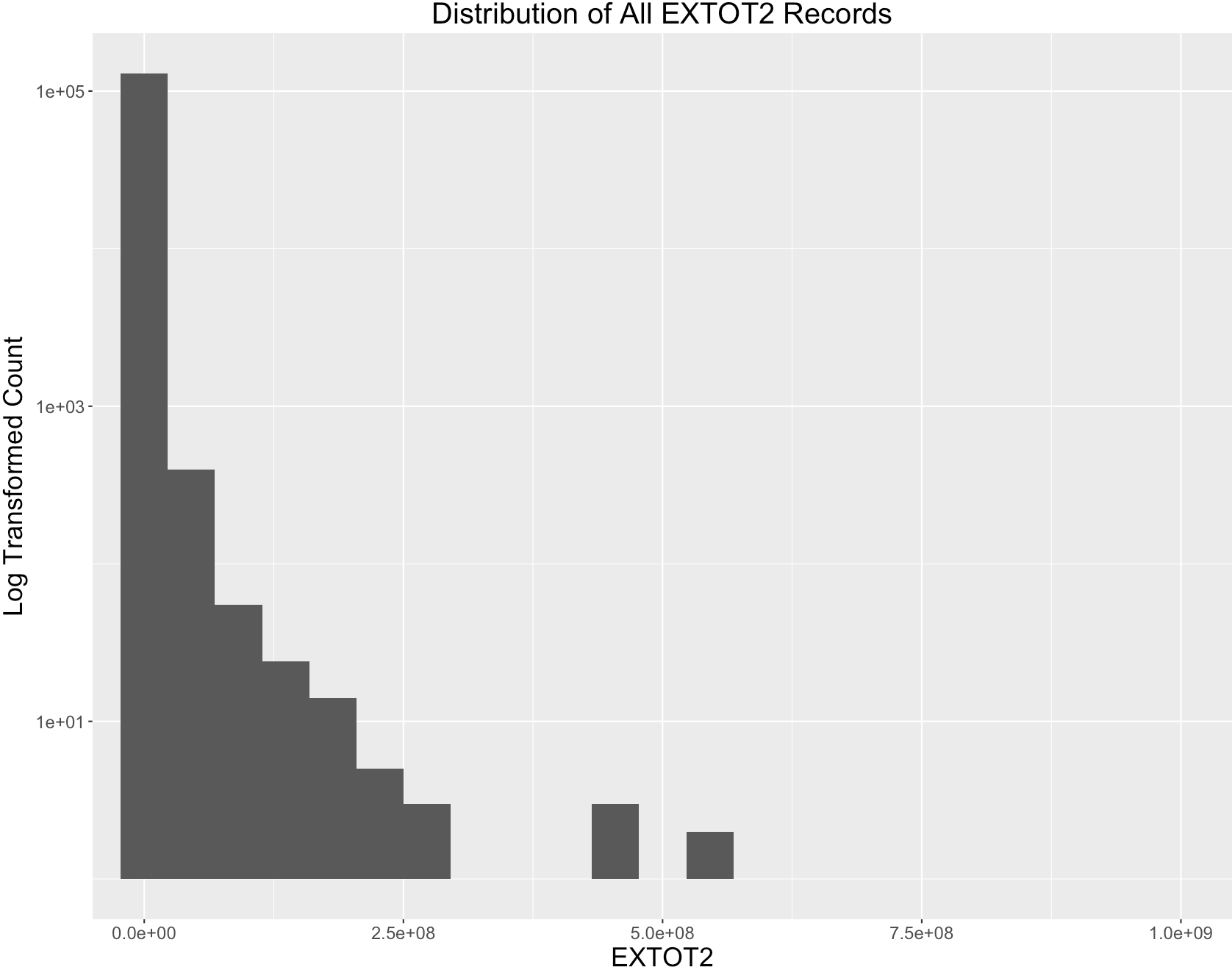
**Description:**

EXTOT2 is a numeric variable representing the total value of the exempt property. It could be the updated assessed value compared to EXTOT. Most EXTOT2 values are lower than their corresponding EXTOT values.

**Unique values:**

EXTOT2 has 48107 unique values ranging from 7 to about 4,500,000,000. There are 918,642 missing values in the EXTOT2 field. The statistics is shown as below.

|  |  |
| --- | --- |
| Minimum | 7 |
| Maximum | 4.50E+09 |
| Median | 37116 |
| Mean | 6.58E+05 |
| Mode | 2090 |
| SD | 1.61E+07 |

****

**Field 27**

**Field Name:** EXCD2

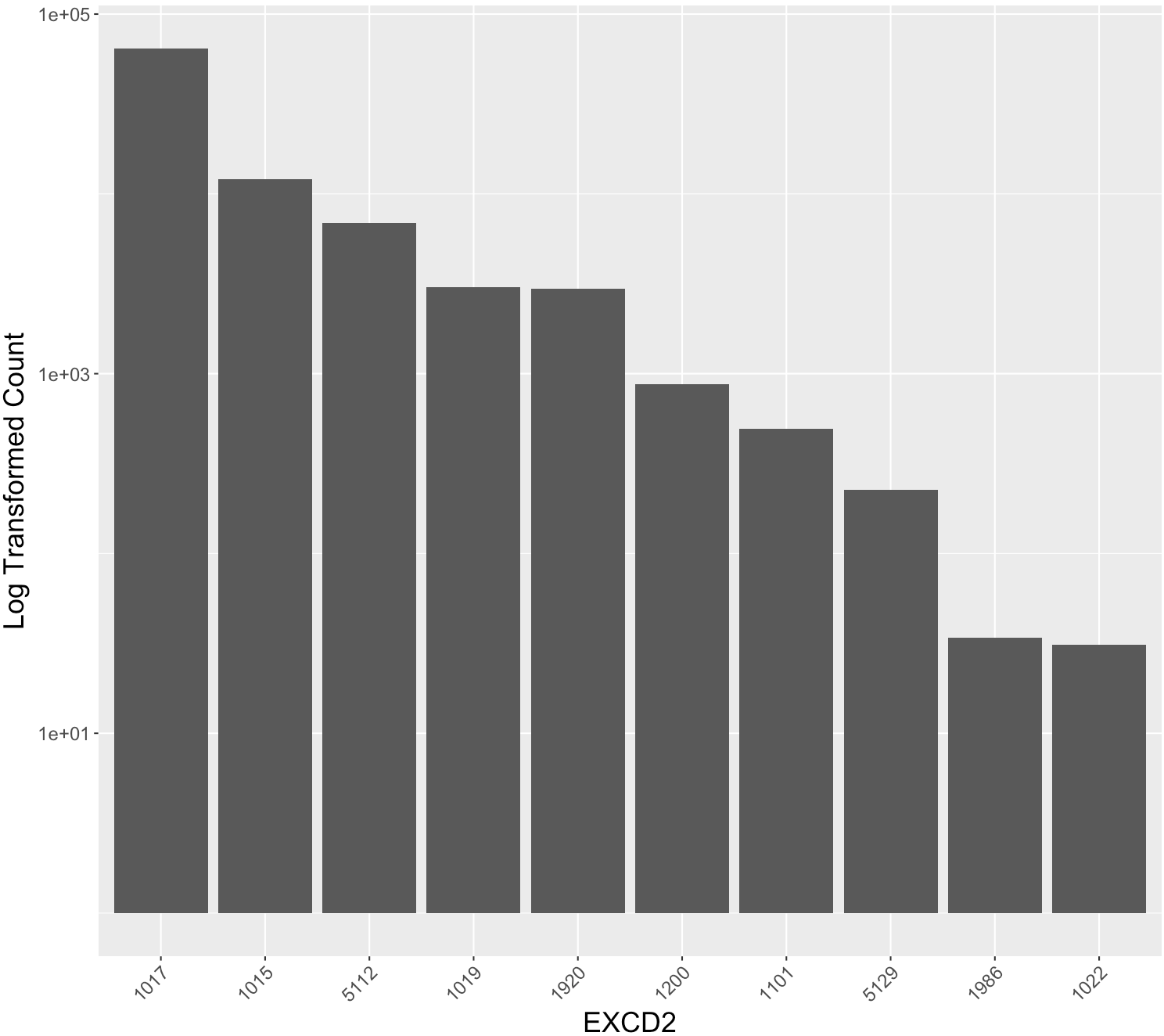
**Description:**

EXCD2 is a categorical variable, possibly representing the code for the exempt reasons for EXLAND2 AND EXTOT2 records.

**Unique Values:**

EXTOT has 61 levels, taking 4-digit numbers from 1011 to 7160. There are 957,634 missing values exist. The top 10 most frequently occurred EXCD2 values are:

|  |  |
| --- | --- |
| EXCD2 | Percentage(%) |
| 1017 | 70.6 |
| 1015 | 13.2 |
| 5112 | 7.6 |
| 1019 | 3.3 |
| 1920 | 3.3 |
| 1200 | 1.0 |
| 1101 | 0.5 |
| 5129 | 0.2 |
| 1986 | 0.0 |
| 1022 | 0.0 |



**Field 28**

**Field Name:** PERIOD

**Description:**

PERIOD is a categorical variable, indicating the change period of the record.

**Unique Values:**

All the records in this dataset take the value of “FINAL” in the PERIOD field.

**Field 29**

**Field Name:** YEAR

**Description:**

YEAR is a date variable, indicating the time that the record is made.

**Unique Values:**

All the records in this dataset take the value of “2010/11” in the YEAR field.

**Field 30**

**Field Name:** VALTYPE

**Description:**

VALTYPE is a categorical variable, indicating the valid type of the record.

**Unique Values:**

All the records in this dataset take the value of “AC-TR” in the VALTYPE field.