**Fraud Analytics Report**

**on NY Property Data**

**Team 2**

**Qianru Zhou**

**Hui Zhu**

**Xingyu Wei**

**Xuan Yang**

**Jingyi Chen**

**Wenting Zhao**

**April 30, 2019**

# **Executive Summary**

In this project, we accepted a request from NY Property to help them detect anomalies in their NY property data records. NY Property Data represents NYC properties assessments for the purpose of calculating property tax, grant eligible properties, exemptions and/or abatements.

Because of the large scale of the dataset, we first examined the data quality and performed exploratory data analysis on some important fields to prepare for the algorithm development in the next step.

Based on the features of this dataset, we conducted feature engineering and dimensionality reduction to develop two unsupervised fraud detection algorithms – Heuristic Algorithm and Autoencoder Algorithm to detect unusual records.

By combining the ranking scores from these two algorithms, we suggested that the final unusual records that have the biggest possibility to be a fraud record should be 10 records which rank top 10 in the final ranking score. Unreasonable full value and lot area are some features of these unusual records.

For future analysis of these unusual records, we suggested NYC properties to conduct further research into whether the abnormalities could have a reasonable explanation because of special property categories such as government building, public park, etc.

**Table of Contents**

[**Executive Summary** 2](#_Toc7543306)

[**I Description of Data** 4](#_Toc7543307)

[**1. Data Description** 4](#_Toc7543308)

[**2. Data Summary** 4](#_Toc7543309)

[2.1 Categorical Variable 4](#_Toc7543310)

[2.2 Numeric Variable 5](#_Toc7543311)

[2.3 Histograms of important fields 6](#_Toc7543312)

[8](#_Toc7543313)

[**II Data Cleaning** 10](#_Toc7543314)

[**III Variable Creation** 11](#_Toc7543315)

[**IV Algorithms** 13](#_Toc7543316)

[**1.** **Heuristic Algorithm** 13](#_Toc7543317)

[**2.** **Autoencoding** 14](#_Toc7543318)

[**3. Score Combination** 15](#_Toc7543319)

[**V Results** 16](#_Toc7543320)

[**1.** **Record: 565392** 19](#_Toc7543321)

[**2.** **Record: 632816** 19](#_Toc7543322)

[**3.** **Record: 1067360** 20](#_Toc7543323)

[**4.** **Record: 917942** 20](#_Toc7543324)

[**5.** **Record: 85886** 21](#_Toc7543325)

[**6.** **Record: 556609** 21](#_Toc7543326)

[**7.** **Record: 821853** 22](#_Toc7543327)

[**8.** **Record: 912501** 22](#_Toc7543328)

[**9.** **Record: 776306** 23](#_Toc7543329)

[**10. Record: 770594** 23](#_Toc7543330)

[**VI** **Conclusion** 24](#_Toc7543331)

[**Appendix** 25](#_Toc7543332)

[**Appendix 1 Data Quality Report** 25](#_Toc7543333)

[**Appendix 2 Top 10 records of Principal Analysis results** 48](#_Toc7543334)

# **I Description of Data**

## **1. Data Description**

**Dataset Name:** NY property data / Property Valuation and Assessment Data

**Data source:** NYC Open Data Website- <https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8>

**Time period:** 2010/11

**Number of columns:** 32

**Number of records:** 1,070,994

## **2. Data Summary**

### **2.1 Categorical Variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Field Names** | **Records that have values** | **% populated** | **# of unique values** | **Most common field value** |
| RECORD | 1,070,994 | 100.0 | 1,070,994 | NA |
| BBLE | 1,070,994 | 100.0 | 1,066,541 | NA |
| B | 1,070,994 | 100.0 | 5 | 4 |
| BLOCK | 1,070,994 | 100.0 | 13,984 | 3944 |
| LOT | 1,070,994 | 100.0 | 6,366 | 1 |
| EASEMENT | 4636 | 0.4 | 13 | E |
| TAXCLASS | 1,070,994 | 100.0 | 11 | 1 |
| EXT | 354305 | 33.1 | 4 | G |
| ZIP | 1,041,104 | 97.2 | 197 | 10,314 |
| EXMPTCL | 15,579 | 1.5 | 15 | X1 |
| PERIOD | 1,070,994 | 100.0 | 1 | FINAL |
| YEAR | 1,070,994 | 100.0 | 1 | 2010/11 |
| VALTYPE | 1,070,994 | 100.0 | 1 | AC-TR |
| OWNER | 1,039,249 | 97.0 | 863,348 | PARKCHESTER PRESERVAT |
| BLDGCL | 1,070,994 | 100.0 | 200 | R4 |
| STADDR | 1,070,318 | 99.9 | 839,281 | 501 SURF AVENUE |
| EXCD1 | 638,488 | 59.6 | 130 | 1017.0 |
| EXCD2 | 92,948 | 8.7 | 61 | 1017.0 |

### **2.2 Numeric Variables**

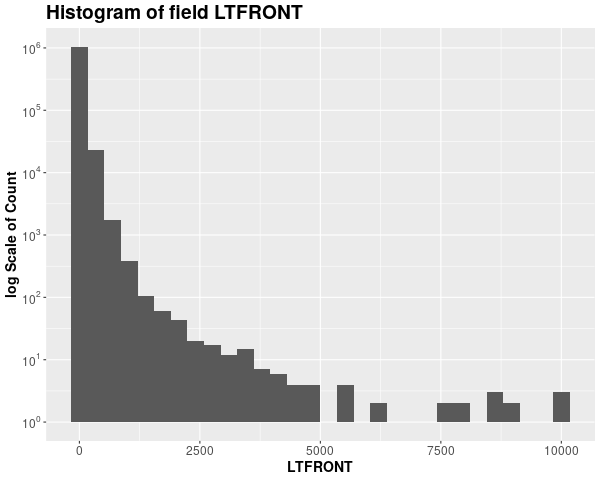
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **# of records** | **% populated** | **#unique values** | **# records with value 0** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
| LTFRONT | 1,070,994 | 100.0 | 1,297 | 169,108 | 36.6 | 74.0 | 0 | 9,999 |
| LTDEPTH | 1,070,994 | 100.0 | 1,370 | 170,128 | 88.9 | 76.4 | 0 | 9,999 |
| STORIES | 1,014,730 | 94.7 | 112 | 0 | 5.0 | 8.4 | 1 | 119 |
| FULLVAL | 1,070,994 | 100.0 | 109,324 | 13,007 | 874,264.5 | 11,582,431 | 0 | 6,150,000,000 |
| AVLAND | 1,070,994 | 100.0 | 70,921 | 13,009 | 85,067.9 | 4,057,260 | 0 | 2,668,500,000 |
| AVTOT | 1,070,994 | 100.0 | 112,914 | 13,007 | 227,238.2 | 6,877,529.3 | 0 | 4,668,308,947 |
| EXLAND | 1,070,994 | 100.0 | 33,419 | 491,699 | 36,423.9 | 3,981,575.8 | 0 | 2,668,500,000 |
| EXTOT | 1,070,994 | 100.0 | 64,255 | 432,572 | 91,187.0 | 6,508,402.8 | 0 | 4,668,308,947 |
| BLDFRONT | 1,070,994 | 100.0 | 612 | 228,815 | 23.0 | 35.6 | 0 | 7,575 |
| BLDDEPTH | 1,070,994 | 100.0 | 621 | 228853 | 39.9 | 42.7 | 0 | 9,393 |
| AVLAND2 | 282,726 | 26.4 | 58,593 | 0 | 246,235.7 | 6,178,962.6 | 3 | 2,371,005,000 |
| AVTOT2 | 282,732 | 26.4 | 111,361 | 0 | 713,911,4 | 11,652,528.9 | 3 | 4,501,180,002 |
| EXLAND2 | 87,449 | 8.2 | 22,196 | 0 | 351,235.7 | 10,802,212.7 | 1 | 2,371,005,000 |
| EXTOT2 | 130,828 | 12.2 | 48,349 | 0 | 656,768.3 | 16,072,510.2 | 7 | 4,501,180,002 |

## **2.3 Histograms of important fields**

**Field 1**

**Field Name: LTFRONT**

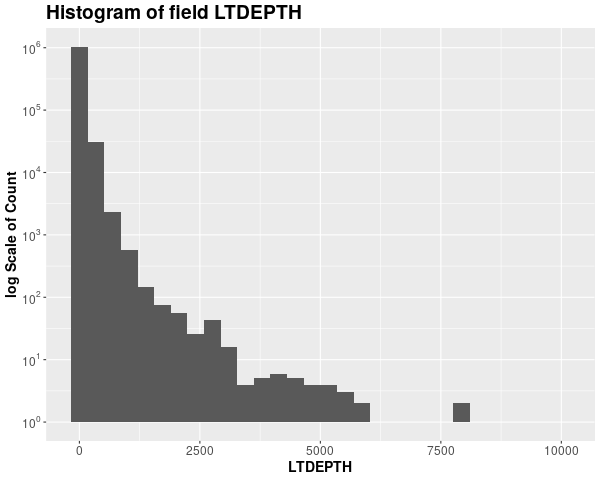
**Description: Lot Frontage in feet.**



**Field 2**

**Field Name: LTDEPTH**

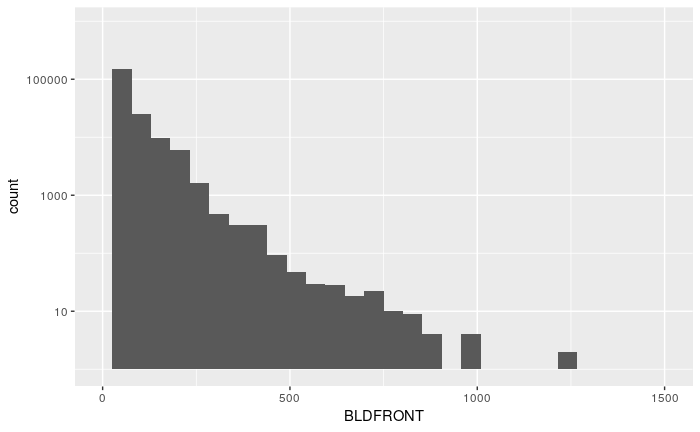
**Description: Lot Depth in feet.**



**Field 3**

**Field Name: BLTFRONT**

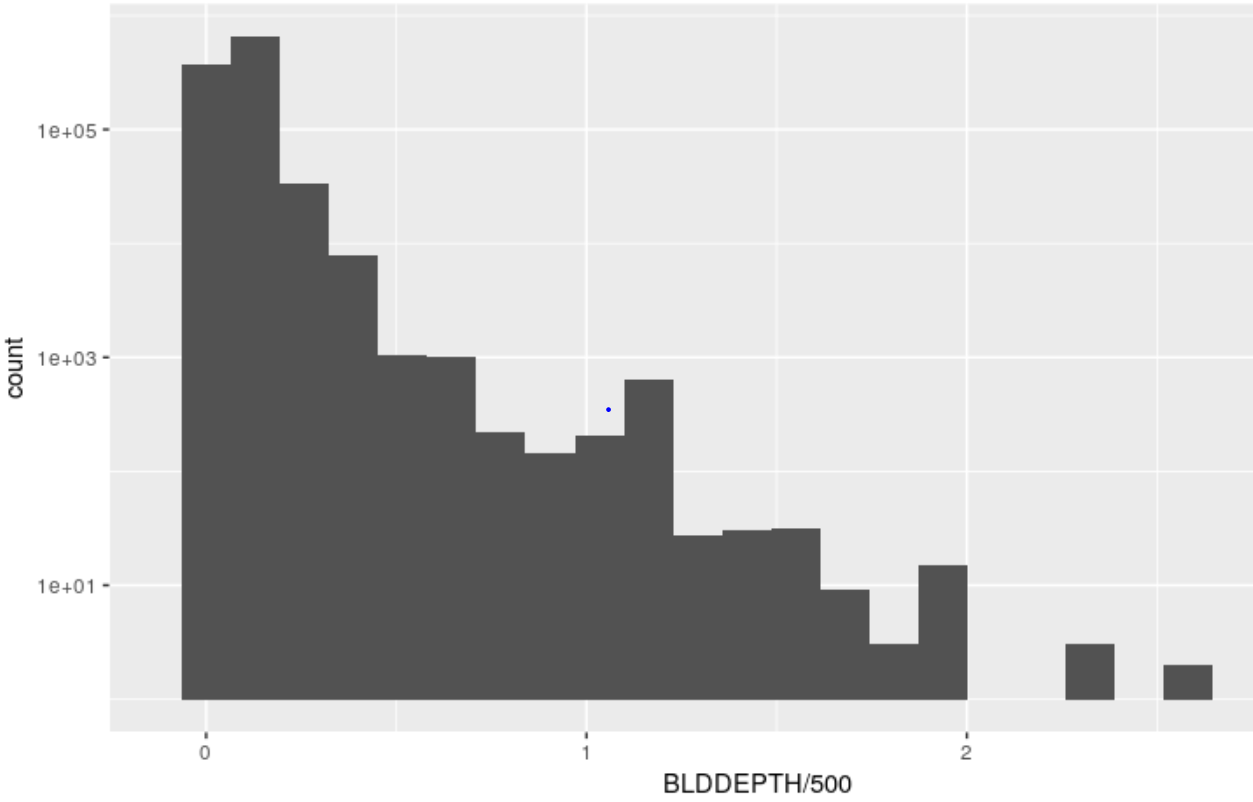
**Description: Building Frontage in feet.**



**Field 4**

**Field Name: BLTDEPTH**

**Description: Building Depth in feet.**



**Field 5**

**Field Name: STORIES**

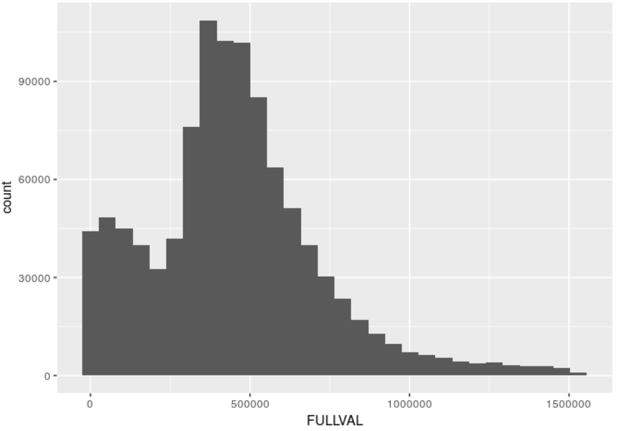
**Description: The number of stories for the building (Number of Floors).**

# 

**Field 6**

**Field Name: FULLVAL**

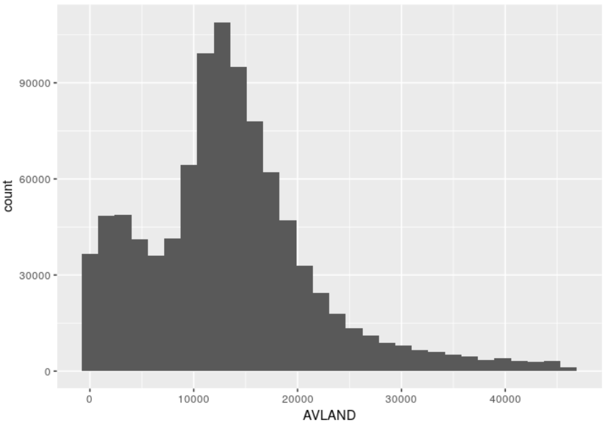
**Description: If not zero, Current year’s total market value of the property.**



**Field 7**

**Field Name: AVLAND**

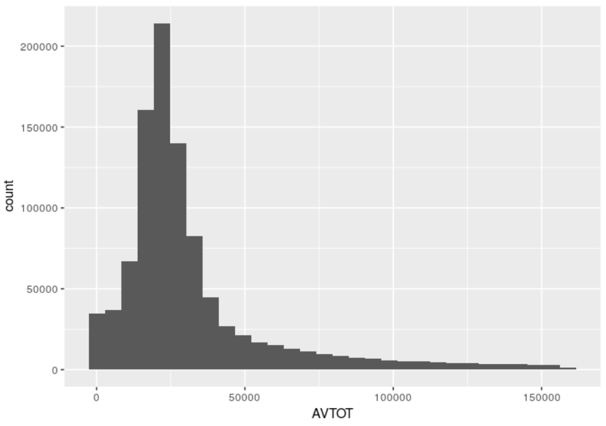
**Description: Assessed land value.**



**Field 8**

**Field Name: AVTOT**

**Description: Assessed total value.**



# **II Data Cleaning**

In this dataset, there are many fields with missing values. We choose 9 fields that are important to algorithm building to fill in missing values.

**Field 1**

**Name: ZIP**

**Method:** Since the distribution of ZIP code in the original dataset has an obvious sequential characteristic, we could fill in a missing ZIP code with the previous one that is available. For example, if a ZIP is missing and the previous ZIP is 11208, we can simply replace the missing value with 11208.

**Field 2**

**Name: STORIES**

**Method:** After checking the original dataset, we noticed that properties at the same street address, in most cases, have the same number of stories. Therefore, we could fill in a missing stories value with the number of stories that a property has at the same street (STADDR)

However, if the number of stories at a certain street is also unknown, we could group by ZIP and BLDGCL, then fill in the missing value with the average number of stories in that ZIP code.

**Field 3 ~ 5**

**Name: FULLVAL, AVLAND, AVTOT**

**Method:** For FULLVAL, AVLAND and AVTOT, since the distributions for the three fields are quite condensed with only a few outliers, hence we could aggregate by ZIP and TAXCLASS and fill in a missing value with the median of that group. If the group size is smaller than 5, we could merely aggregate by TAXCLASS.

**Field 6 ~ 9**

**Name: LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH**

**Method:** For these four fields that are highly correlated with LOT and B (Borough), we could aggregate by LOT and B, and then fill in a missing value with the median of that group. We choose medians in order to minimize the effects of outliers on the projected values. If there are still missing values after filling in the field with group medians, we can continue filling the rest of missing values with group medians gained from aggregating only by B.

# 

# **III Variable Creation**

1. **Critical Variables Selection**

First, we select the following variables from our original dataset

V1= FULLVAL V2 = AVLAND V3 = AVTOT

1. **Variables Group by and Creation**

Then, we created 3 new variables S1, S2, S3. S1 represents the area of lot, S2 represents the one story area of a building, S3 represents the total area of a building

S1 = LTFRONT \* LTDEPTH

S2 = BLDFRONT \* BLDDEPTH

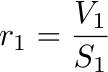
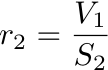
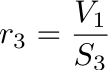
S3 = S2\* STORIES

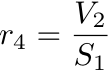
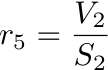
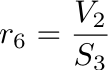
For each record make 9 ratios:

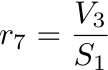
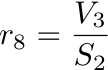
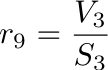
r1, r4, r7 represents market value per lot, per one story area, per total story area

r2, r5, r8 represents land area per lot area, per one story area, per total story area

r3, r6, r9 represents units of building per lot, per one story area, per total story area

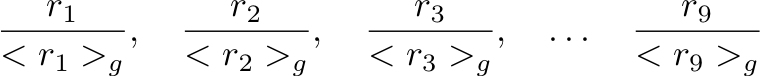
  

Separately group records by the 5 groups: zip5, zip3, TAXCLASS, borough, all because r1-r9 might varies a lot in different area, tax class and borough.

For each group , calculate , the average of each ri for each group g

For each record calculate 45 variables:

 g= 1, …, 5

## 

**The following table shows the list of all 45 variables we created.**

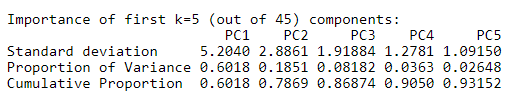
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable Name** | **Formula** | **Variable Name** | **Formula** | **Variable** | **Formula** |
| val\_lft\_zip5 | r1/<r1>1 | land\_ltf\_zip5 | r4/<r4>1 | tol\_lft\_zip5 | r7/<r7>1 |
| val\_lft\_zip3 | r1/<r1>2 | land\_ltf\_zip3 | r4/<r4>2 | tol\_lft\_zip3 | r7/<r7>2 |
| val\_lft\_\_taxclass | r1/<r1>3 | land\_ltf\_taxclass | r4/<r4>3 | tol\_lft\_taxclass | r7/<r7>3 |
| val\_lft\_\_borough | r1/<r1>4 | land\_ltf\_borough | r4/<r4>4 | tol\_lft\_borough | r7/<r7>4 |
| val\_lft\_\_all | r1/<r1>5 | land\_ltf\_all | r4/<r4>5 | tol\_lft\_all | r7/<r7>5 |
| val\_bld\_zip5 | r2/<r2>1 | land\_bld\_zip5 | r5/<r5>1 | tol\_bld\_zip5 | r8/<r8>1 |
| val\_bld\_zip3 | r2/<r2>2 | land\_bld\_zip3 | r5/<r5>2 | tol\_bld\_zip3 | r8/<r8>2 |
| val\_bld\_taxclass | r2/<r2>3 | land\_bld\_taxclass | r5/<r5>3 | tol\_bld\_taxclass | r8/<r8>3 |
| val\_bld\_borough | r2/<r2>4 | land\_bld\_borough | r5/<r5>4 | tol\_bld\_borough | r8/<r8>4 |
| val\_bld\_all | r2/<r2>5 | land\_bld\_all | r5/<r5>5 | tol\_bld\_all | r8/<r8>5 |
| val\_store\_zip5 | r3/<r3>1 | land\_store\_zip5 | r6/<r6>1 | tol\_store\_zip5 | r9/<r9>1 |
| val\_store\_zip3 | r3/<r3>2 | land\_store\_zip3 | r6/<r6>2 | tol\_store\_zip3 | r9/<r9>2 |
| val\_store\_taxclass | r3/<r3>3 | land\_store\_taxclass | r6/<r6>3 | tol\_store\_taxclass | r9/<r9>3 |
| val\_store\_borough | r3/<r3>4 | land\_store\_borough | r6/<r6>4 | tol\_store\_borough | r9/<r9>4 |
| val\_store\_all | r3/<r3>5 | land\_store\_all | r6/<r6>5 | tol\_store\_all | r9/<r9>5 |

# **IV Algorithms**

## **Heuristic Algorithm**

Before performing PCA, the variables are first z-scaled so that the principal components are not dominated by variables of a much larger scale.

After performing PCA, top 5 principal components were selected to cover 93% of the total variance. Summary results are shown below.



After that, the original data was represented in the chosen principal components and the 5 PCs were further z-scaled to be on the same footing scale.

A summary table of minimum, maximum of variables and Euclidean distance were calculated. The records were then arranged with descending order of Euclidean distance and assigned a unique rank number.

For a dataset with many variables, there is a high probability of hidden correlations between variables, PCA could decorrelate the input data. In addition, with such high dimensions, the analysis could be slow and inefficient, PCA captures the most important components without losing much information, thus it is a much more efficient way of analysis.

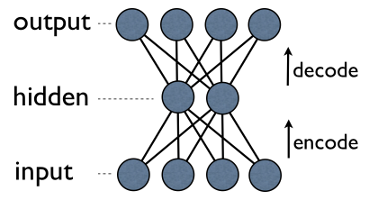
### 

## **Autoencoding**

* **How does an autoencoder work?**

An autoencoder is a neural network that is trained by unsupervised learning, which is trained

to learn reconstructions that are close to its original input. An autoencoder is composed of two parts, an encoder, and a decoder. A neural network with a single hidden layer has an encoder and decoder respectively. There are also weights, transformation function, and bias. The encoder maps an input vector to a hidden representation by an aﬃne mapping following a nonlinearity. The decoder maps the hidden representation back to the original input space as reconstruction by the same transformation as the encoder.



In short, this is training the autoencoder to reproduce the original input x from a noisy input ˆ x. This allows the autoencoder to be robust to data with white noise and capture only meaningful patterns of the data. It uses the reconstruction error as the anomaly score. Data points with high reconstruction are anomalies. After training, the autoencoder will reconstruct normal data very well, while failing to do so with anomaly data which the autoencoder has not encountered.

* **Reasons for using Autoencoder:**

1. Autoencoder is a deep learning model, which can conduct unsupervised learning and produce features in the dataset. This method utilizes deep learning to provide a better method than purely scaling as an autoencoder can study more about different features.
2. The record prediction is more accurate than a normal machine learning model. Also, this is a model using its own records to predict itself, so autoencoder is able to conduct unsupervised learning, unlike the other neural network models.
3. It may be more efficient, in terms of model parameters, to learn several layers with an autoencoder rather than to learn one huge transformation with PCA.
4. An autoencoder can learn non-linear transformations, unlike PCA, with a non-linear activation function and multiple layers.

* **Autoencoder Calculation:**

After z-scaling the cleaned data and conducting Principal Component Analysis, the dataset is z-scaled again to get our final data for autoencoder model. Then, there is an R package called h2o which has an autoencoder deep learning function that allows us to reconstruct the PCA datasets we produced before and study the features to reproduce the same records. The Euclidean Distance between the actual values and the prediction values from autoencoder will be the fraud scores of each record for this case.

**Autoencoder results (Top ten abnormal scores):**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **PCA1** | **PCA2** | **PCA3** | **PCA4** | **PCA5** | **Score** | **RECORD** | **RANK** |
| 835.6818 | -130.911 | 419.6394 | 187.0608 | -1.29663 | 1455.51433 | 565392 | 1 |
| 153.2996 | -673.244 | -451.05 | -159.731 | -185.295 | 955.572566 | 1067360 | 2 |
| 374.6287 | 476.2997 | -737.942 | 357.88 | 45.35209 | 774.638242 | 632816 | 3 |
| 205.6739 | 105.1439 | 125.7514 | -80.3834 | -584.228 | 694.721653 | 917942 | 4 |
| 135.5034 | 121.3464 | -38.0273 | -340.176 | 234.3469 | 379.324027 | 85886 | 5 |
| 61.50492 | -144.653 | 35.10579 | 130.6117 | 275.0408 | 363.198204 | 556609 | 6 |
| 64.30577 | -196.766 | -17.4422 | 107.63 | 180.939 | 350.003329 | 821853 | 7 |
| 53.18943 | -156.76 | 2.742903 | 115.5123 | 199.3594 | 318.413126 | 776306 | 8 |
| 73.60326 | -112.862 | 23.70273 | 38.04658 | 276.9273 | 317.37586 | 912501 | 9 |
| 32.76957 | -172.626 | -154.257 | -94.635 | -119.006 | 307.991139 | 770594 | 10 |

## **3. Score Combination**

**Combine scores from both of the Algorithms:**

After we got the z-score and autoencoder score, we sorted those two scores ascendingly and replaced each score with the rank order. Then each score was on the same footing and could be combined. We calculated the average of the two ranking orders and took it as our final score for each record. From the result table, the smaller the score is, the higher the chance the record is anomalous.

# **V Results**

By using two algorithms and combine scores from each algorithm, finally we got the list of unusual records, and we chose the top 10 unusual records as our examination object. The following table shows the top 10 unusual records with scores from the heuristic algorithm, autoencoding and combination of two ranking scores.

**Final results based on two ranking scores:**

|  |  |  |  |
| --- | --- | --- | --- |
| **RECORD** | **Rank 1** | **Rank 2** | **Final Score** |
| 565392 | 2 | 1 | **1.5** |
| 632816 | 1 | 3 | **2.0** |
| 1067360 | 3 | 2 | **2.5** |
| 917942 | 4 | 4 | **4.0** |
| 85886 | 5 | 5 | **5.0** |
| 556609 | 6 | 6 | **6.0** |
| 821853 | 8 | 7 | **7.5** |
| 912501 | 7 | 9 | **8.0** |
| 776306 | 9 | 8 | **8.5** |
| 770594 | 10 | 10 | **10.0** |

**Heuristic Algorithm Result:**

The following chart shows the distribution of scores from heuristic algorithm. We can see that the distribution is right-skewed.

A screenshot of a cell phone

Description generated with high confidence

**Autoencoder Algorithm Result:**

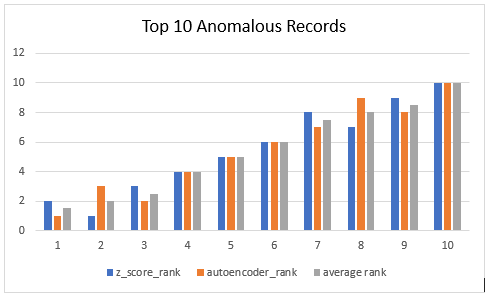
The following chart shows the distribution of scores from the autoencoder algorithm. We can see the distribution is also right-skewed.

A screenshot of a cell phone

Description generated with very high confidence

**Combination Result:**

The following chart shows the comparison of rank among heuristic algorithm rank, autoencoder algorithm rank, and final average rank. We can find that the outcome of the two algorithms are pretty similar so we decide to use average rank as our final score.



**Interpretation from the result:**

In the following paragraphs, we use some tables to describe the reason why these top10 records are unusual.

## **Record: 565392**

For record 565392, there are three similar properties among the whole data set based on ZIP, B, BLOCK, TAXCLASS and LOT size. By comparing them, we can find that the full value, average land value and average total value of Record 565392 are much higher (about 100-1000 times) than the other similar properties.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **B** | **BLOCK** | **TAX**  **CLASS** | **LTFRONT** | **LTDEPTH** | **FULLVAL** | **AVLAND** | **AVTOT** | **ZIP** |
| **565392** | **3** | **8590** | **4** | **117** | **108** | **4,326,303,700** | **1,946,836,665** | **1,946,836,665** | **11229** |
| 565393 | 3 | 8590 | 4 | 324 | 252 | 41,615,000 | 18,096,750 | 18,726,750 | 11229 |
| 565394 | 3 | 8590 | 4 | 155 | 150 | 2,740,000 | 1,138,500 | 1,233,000 | 11229 |
| 565399 | 3 | 8591 | 4 | 200 | 170 | 1,521,450 | 684,653 | 684,653 | 11229 |

## **Record: 632816**

Record 632816 has BLDFRONT and BLDDEPTH values of 1 which seems unusual. In addition, compared to records with same B, TAXCLASS, ZIP and BLDGCL, it has significantly higher values (e.g. FULLVAL, AVLAND, AVTOT).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **BLOCK** | **LOT** | **LTFRONT** | **LTDEPTH** | **FULLVAL** | **AVLAND** | **AVTOT** | **BLDFRONT** | **BLDDEPTH** |
| **632816** | **1,842** | **1** | **157** | **95** | **2,930,000** | **1,318,500** | **1,318,500** | **1** | **1** |
| 624591 | 1,559 | 6 | 34 | 90 | 1,240,000 | 558,000 | 558,000 | 0 | 0 |
| 625256 | 1,584 | 10 | 50 | 193 | 994,000 | 82,350 | 447,300 | 50 | 74 |
| 657545 | 2,870 | 1 | 50 | 100 | 1,250,000 | 54,450 | 562,500 | 44 | 36 |

## **Record: 1067360**

For record 1067360, we found the 4 most similar records based on ZIP, STORIES, TAXCLASS and BLDGCL. As you can see from the following table, the abnormality of this record falls in the LTFRONT and LTDEPTH that are only 1 foot long for each value. As the LTFRONT and LTDEPTH represent the area of the lot, they seem abnormally small compared to other similar records buildings, which indicates the possibility of it being a fraudulent evaluation.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ZIP** | **BLDGCL** | **RECORD** | **LTFRONT** | **LTDEPTH** | **STORIES** | **FULLVAL** | **TAXCLASS** | **AVLAND** |
| **10307** | **B2** | **1067360** | **1** | **1** | **2** | **836000** | **1** | **28800** |
| 10307 | B2 | 1066709 | 40 | 100 | 2 | 674000 | 1 | 23400 |
| 10307 | B2 | 1067202 | 40 | 134 | 2 | 890000 | 1 | 28426 |
| 10307 | B2 | 1068681 | 62 | 100 | 2 | 792000 | 1 | 23307 |
| 10307 | B2 | 1069170 | 51 | 97 | 2 | 651000 | 1 | 19967 |

## 

## **Record: 917942**

By using the same B, BLOCK, BLDGCL, TAXCLASS to filter records, we got 3 similar records. From the following table, we found that the abnormality of Record 917942 is that FULLVAL, AVLAND, EXLAND and EXTOT are unusually larger than the other similar records.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **LTFRONT** | **LTDEPTH** | **STORIES** | **FULL**  **VAL** | **AVLAND** | **EXLAND** | **EXTOT** |
| **917942** | **4910** | **100** | **3** | **374,019,883** | **1,792,808,947** | **1,792,808,947** | **4,668,308,947** |
| 917943 | 1500 | 3000 | 3 | 107,113,000 | 48,150,000 | 48,150,000 | 48,200,850 |
| 917948 | 6500 | 2600 | 1 | 150,000,000 | 67,500,000 | 67,500,000 | 67,500,000 |

## **Record: 85886**

For record 85886, comparing records with same B, TAXCLASS, BLDGEL and similar range of LTFRONT and LTDEPTH values, it has BLDFRONT and BLDDEPT values of 8 which is unusual.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **LTFRONT** | **LTDEPTH** | **FULLVAL** | **AVLAND** | **EXLAND** | **BLDFRONT** | **BLDDEPTH** |
| **85,886** | **4,000** | **150** | **70,214,000** | **31,455,000** | **31,455,000** | **8** | **8** |
| 127,334 | 4,000 | 4,500 | 173,000,000 | 66,600,000 | 66,600,000 | 0 | 0 |
| 131,603 | 3,490 | 500 | 46,968,000 | 20,925,000 | 20,925,000 | 14 | 34 |
| 137,654 | 3,459 | 349 | 49,100,000 | 20,745,000 | 20,745,000 | 40 | 35 |
| 127,322 | 3,000 | 4,000 | 151,000,000 | 13,950,000 | 13,950,000 | 0 | 0 |

## 

## **Record: 556609**

For record 556609, we found the 3 most similar records based on ZIP, STORIES, TAXCLASS and BLDGCL. As you can see from the following table, the abnormality of this record lies in the fact that the property has an unusually high FULLVAL with a relatively small LTFRONT and LTDEPTH, which indicates the possibility of it being a fraudulent evaluation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **LTFRONT** | **LTDEPTH** | **STORIES** | **FULLVAL** | **AVLAND** | **EXLAND** | **EXTOT** |
| **556609** | **35** | **50** | **1** | **136,000,000** | **60,750,000** | **60,750,000** | **61,200,000** |
| 542090 | 201 | 299 | 1 | 1,940,000 | 648,000 | 648,000 | 873,000 |
| 548427 | 200 | 500 | 1 | 3,140,000 | 1,264,500 | 1,264,500 | 1,413,000 |
| 550360 | 179 | 200 | 1 | 1,042,800 | 448,200 | 448,200 | 469,260 |

## 

## **Record: 821853**

For record 821853, there are four similar properties among the whole data set based on ZIP, B, BLDGCL, TAXCLASS and FULLVAL. By comparing them, we can find that the LTFRONT and LTDEPTH of Record 821853 are abnormally smaller than the other similar properties.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **B** | **BLDGCL** | **TAXCLASS** | **LTFRONT** | **LTDEPTH** | **FULLVAL** | **ZIP** |
| **821853** | **4** | **U7** | **3** | **2** | **1** | **138000** | **11432** |
| 821712 | 4 | U7 | 3 | 100 | 30 | 138000 | 11432 |
| 821754 | 4 | U7 | 3 | 200 | 60 | 138000 | 11432 |
| 821845 | 4 | U7 | 3 | 294 | 50 | 138000 | 11432 |
| 821855 | 4 | U7 | 3 | 272 | 50 | 138000 | 11432 |

## **Record: 912501**

For record 912501, there are three similar properties among the whole data set based on ZIP, B, BLDGCL, TAXCLASS and LOT size. By comparing them, we can find that the full value, average land value and average total value of this property are much higher (about 10 times) than the other similar properties with bigger lot size.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **B** | **BLOCK** | **TAX**  **CLASS** | **LT**  **FRONT** | **LT**  **DEPTH** | **FULL**  **VAL** | **AVLAND** | **AVTOT** | **ZIP** |
| **912501** | **4** | **13791** | **4** | **25** | **100** | **128,222,000** | **57,600,000** | **57,699,900** | **11413** |
| 912502 | 4 | 13791 | 4 | 352 | 635 | 12,100,000 | 1,777,500 | 5,445,000 | 11413 |
| 912503 | 4 | 13791 | 4 | 315 | 700 | 13,500,000 | 2,047,500 | 6,075,000 | 11413 |
| 912504 | 4 | 13791 | 4 | 435 | 670 | 28,400,000 | 2,979,000 | 12,780,000 | 11413 |
| 912505 | 4 | 13791 | 4 | 516 | 771 | 17,800,000 | 2,250,000 | 8,010,000 | 11413 |

## **Record: 776306**

By using the same B, BLDGCL, STORIES to filter records, we got 35 similar records. From the following table, we found that the abnormality of Record 776306 is that LTFRONT and LTDEPTH of this record are extremely small but the FULLVAL is unreasonably large.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **LTFRONT** | **LTDEPTH** | **STORIES** | **FULL**  **VAL** | **AVLAND** | **EXLAND** | **EXTOT** |
| **776306** | **6** | **1** | **1** | **524,500** | **95,625** | **0** | **0** |
| 803290 | 93 | 49 | 1 | 432,000 | 19,440 | 0 | 0 |
| 803873 | 45 | 26 | 1 | 13,000 | 5,850 | 0 | 0 |
| 876767 | 31 | 143 | 1 | 26,300 | 11,835 | 11,835 | 11,835 |
| 820414 | 98 | 66 | 1 | 74,000 | 33,300 | 0 | 0 |

## **10. Record: 770594**

By filtering records based the same B, TAXCLASS, ZIP, BLDGCL, Block we got 327 similar records. From the following table, we found that the abnormality of Record 770594 is that BLDFRONT and BLDDEPTH of record 770594 is much higher than other similar records.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **B** | **ZIP** | **TAXCALSS** | **BLDGCL** | **BLOCK** | **BLDFRONT** | **BLDDEPTH** |
| **770594** | **4** | **111364** | **1A** | **R3** | **7621** | **96** | **60** |
| 770450 | 4 | 111364 | 1A | R3 | 7621 | 0 | 0 |
| 770451 | 4 | 111364 | 1A | R3 | 7621 | 0 | 0 |
| 770452 | 4 | 111364 | 1A | R3 | 7621 | 0 | 0 |

## 

## 

# **VI** **Conclusion**

In summary, we first performed exploratory data analysis on the dataset to have a general understanding of the data. After that, we filled empty cells with values we believe to be reasonable, which means they should have minimum impact on the analysis in the later part.

Before building any fraud model, we created 45 new variables which were derived from the variables available.

Two fraud models were used in this project, Principal Component Analysis and Autoencoding. The results were arranged in descending order based on the outlier scores, and then each record was given a unique ranking score, the most unusual record will have a ranking score of 1. These two fraud model produced a very similar result in terms of top 10 unusual records. The final score is an average of the 2 scores produced.

In the end, the top 10 records were examined individually and compared with records with similar features, such as B, TAXCLASS, and ZIP, to determine why these records were marked as outliers by the models.

Further steps like gathering additional information from other sources for these records could be done to examine if these unusual records are reasonable. For example, record 85886 has a pretty large LTFRONT and LTDEPTH but a very small BLDFRONT and BLDDEPTH, thus was marked as usual by the model. However, this property belongs to PARKS AND RECREATION, thus it may be a large park with a small administrative office.

# 

# **Appendix**

## **Appendix 1 Data Quality Report**

1. **Data description**

Dataset Name: NY property data / Property Valuation and Assessment Data

Data source: NYC Open Data Website- <https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8>

Time period: 2010/11

Number of columns: 32

Number of records: 1,070,994

1. **Summary**

**2.1 Numeric Values Table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **# of records** | **% populated** | **#Unique values** | **# Records with value 0** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
| LTFRONT | 1,070,994 | 100.0 | 1,297 | 169,108 | 36.6 | 74.0 | 0 | 9,999 |
| LTDEPTH | 1,070,994 | 100.0 | 1,370 | 170,128 | 88.9 | 76.4 | 0 | 9,999 |
| STORIES | 1,014,730 | 94.7 | 112 | 0 | 5.0 | 8.4 | 1 | 119 |
| FULLVAL | 1,070,994 | 100.0 | 109,324 | 13,007 | 874,264.5 | 11,582,431 | 0 | 6,150,000,000 |
| AVLAND | 1,070,994 | 100.0 | 70,921 | 13,009 | 85,067.9 | 4,057,260 | 0 | 2,668,500,000 |
| AVTOT | 1,070,994 | 100.0 | 112,914 | 13,007 | 227,238.2 | 6,877,529.3 | 0 | 4,668,308,947 |
| EXLAND | 1,070,994 | 100.0 | 33,419 | 491,699 | 36,423.9 | 3,981,575.8 | 0 | 2,668,500,000 |
| EXTOT | 1,070,994 | 100.0 | 64,255 | 432,572 | 91,187.0 | 6,508,402.8 | 0 | 4,668,308,947 |
| BLDFRONT | 1,070,994 | 100.0 | 612 | 228,815 | 23.0 | 35.6 | 0 | 7,575 |
| BLDDEPTH | 1,070,994 | 100.0 | 621 | 228853 | 39.9 | 42.7 | 0 | 9,393 |
| AVLAND2 | 282,726 | 26.4 | 58,593 | 0 | 246,235.7 | 6,178,962.6 | 3 | 2,371,005,000 |
| AVTOT2 | 282,732 | 26.4 | 111,361 | 0 | 713,911,4 | 11,652,528.9 | 3 | 4,501,180,002 |
| EXLAND2 | 87,449 | 8.2 | 22,196 | 0 | 351,235.7 | 10,802,212.7 | 1 | 2,371,005,000 |
| EXTOT2 | 130,828 | 12.2 | 48,349 | 0 | 656,768.3 | 16,072,510.2 | 7 | 4,501,180,002 |

**2.2 Categorical Values Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Field Names** | **Records that has values** | **% Populated** | **# of unique values** | **Most common field value** |
| RECORD | 1,070,994 | 100.0 | 1,070,994 | NA |
| BBLE | 1,070,994 | 100.0 | 1,066,541 | NA |
| B | 1,070,994 | 100.0 | 5 | 4 |
| BLOCK | 1,070,994 | 100.0 | 13,984 | 3944 |
| LOT | 1,070,994 | 100.0 | 6,366 | 1 |
| EASEMENT | 4636 | 0.4 | 13 | E |
| TAXCLASS | 1,070,994 | 100.0 | 11 | 1 |
| EXT | 354305 | 33.1 | 4 | G |
| ZIP | 1,041,104 | 97.2 | 197 | 10,314 |
| EXMPTCL | 15,579 | 1.5 | 15 | X1 |
| PERIOD | 1,070,994 | 100.0 | 1 | FINAL |
| YEAR | 1,070,994 | 100.0 | 1 | 2010/11 |
| VALTYPE | 1,070,994 | 100.0 | 1 | AC-TR |
| OWNER | 1,039,249 | 97.0 | 863,348 | PARKCHESTER PRESERVAT |
| BLDGCL | 1,070,994 | 100.0 | 200 | R4 |
| STADDR | 1,070,318 | 99.9 | 839,281 | 501 SURF AVENUE |
| EXCD1 | 638,488 | 59.6 | 130 | 1017.0 |
| EXCD2 | 92,948 | 8.7 | 61 | 1017.0 |

1. **Data Field Exploration**

Field 1

Field Name: RECORD

Description: Unique identifier of each data record. It is an integer from 1 to 1070994.

Field 2

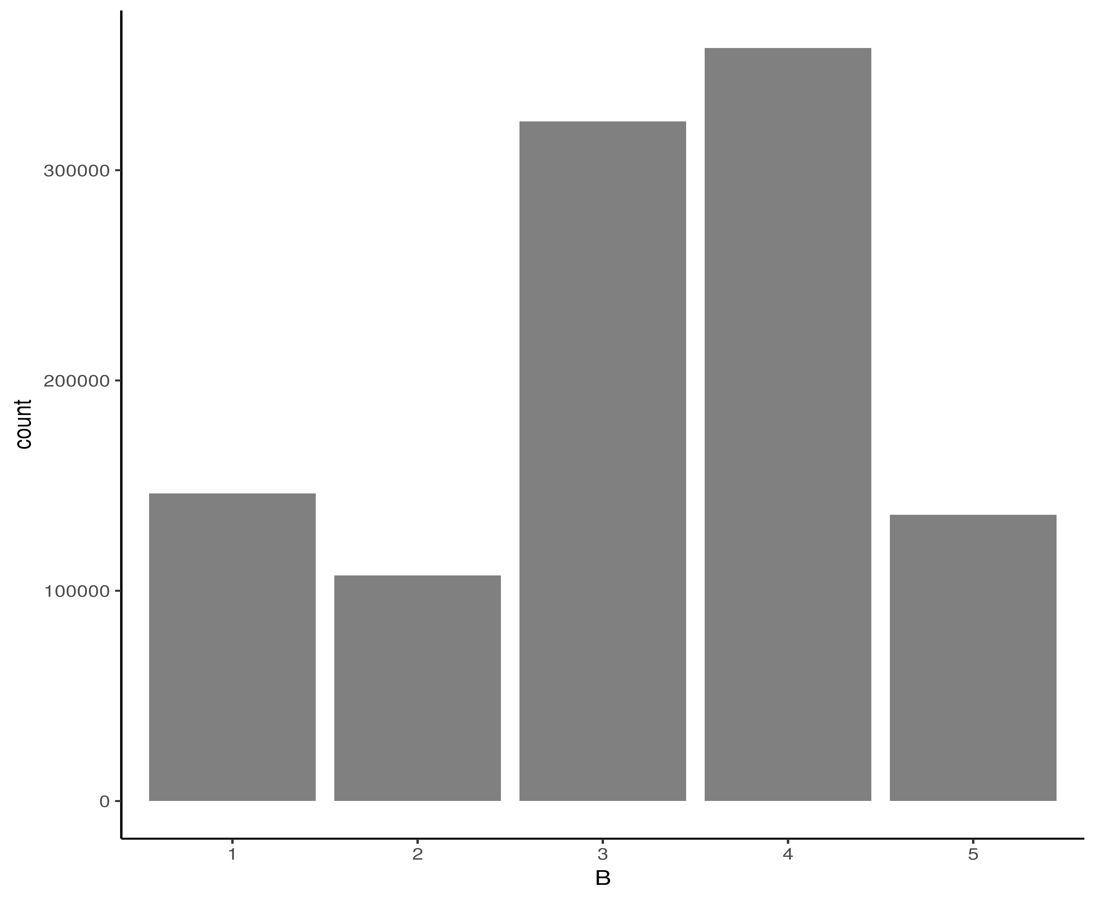
Field Name: BBLE

Description: Concatenation of borough code, block code, Unique # within borough/block, easement. It is a 10-digit code.

Field 3

Field Name: B

Description: Borough codes.



Field 4

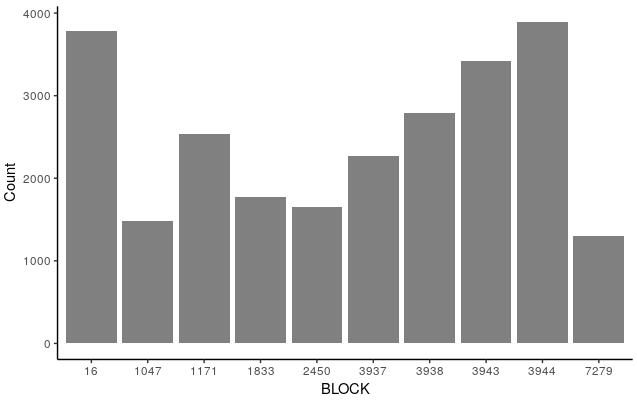
Field Name: BLOCK

Description: Valid block ranges by borough.

Top10 Field Value



Top10 Field Value Plot



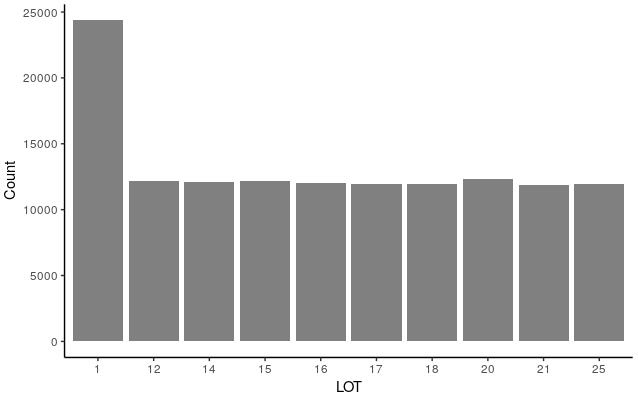
Field 5

Field Name: LOT

Description: Unique # within borough/block.

Top 10 Field Value



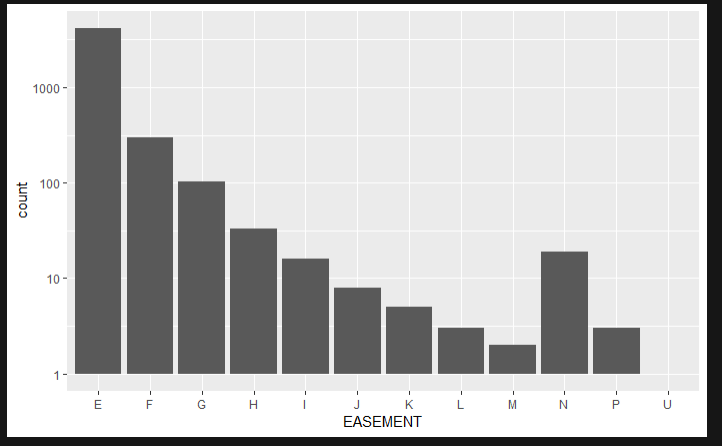
Top10 Field Value Plot

Field 6

Field Name: EASEMENT

Description: Describe easement.

Plot the Easement Field on a Log Scale



Field 7

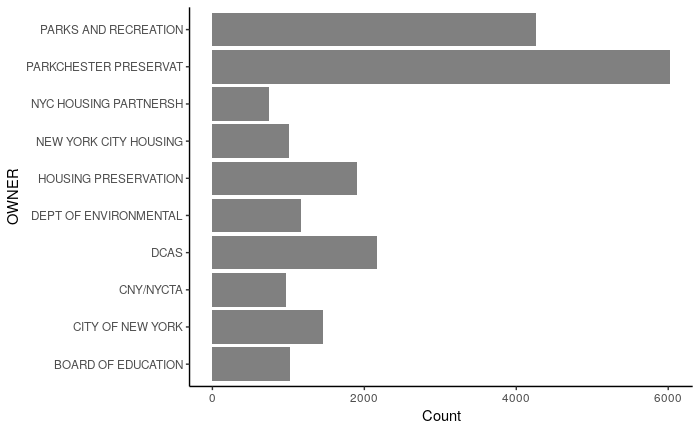
Field Name: OWNER

Description: Owner’s name.

Top 10 Field Value



Top10 Field Value Plot



Field 8

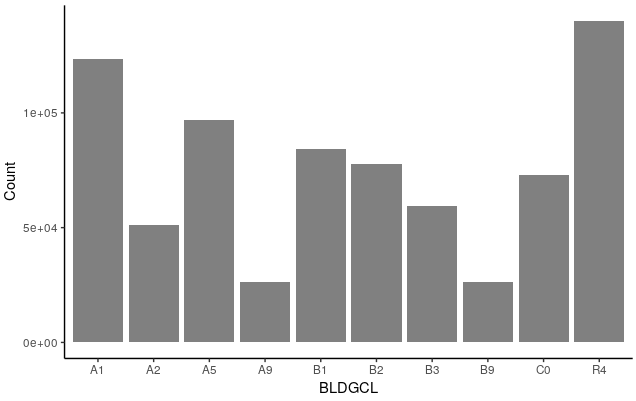
Field Name: BLDGCL

Description: Building class.

Top 10 Field Value



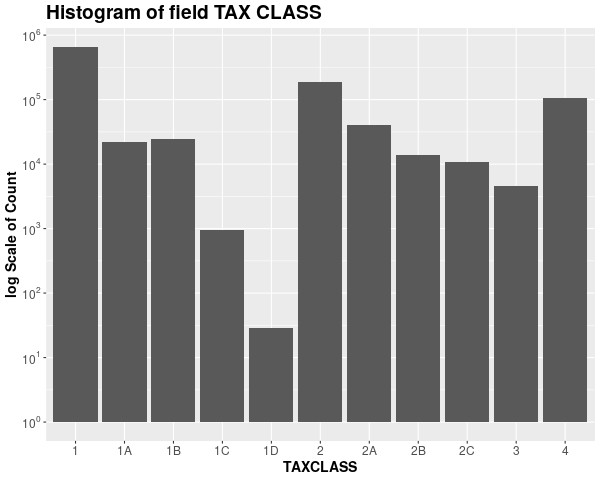
Top10 Field Value Plot



Field 9

Field Name: TAXCLASS

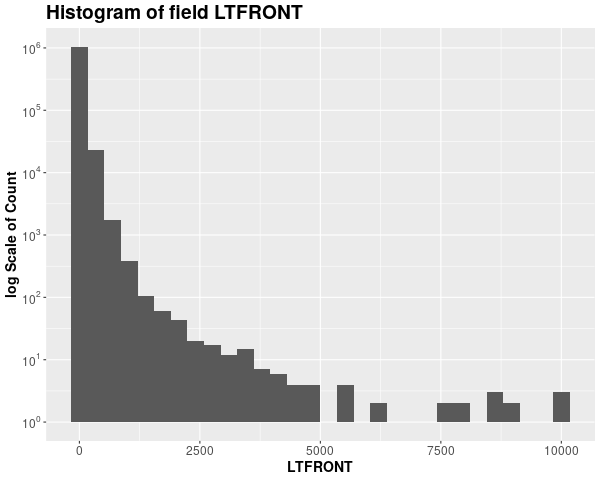
Description: Current Property Tax Class Code (NYS Classification).



Field 10

Field Name: LTFRONT

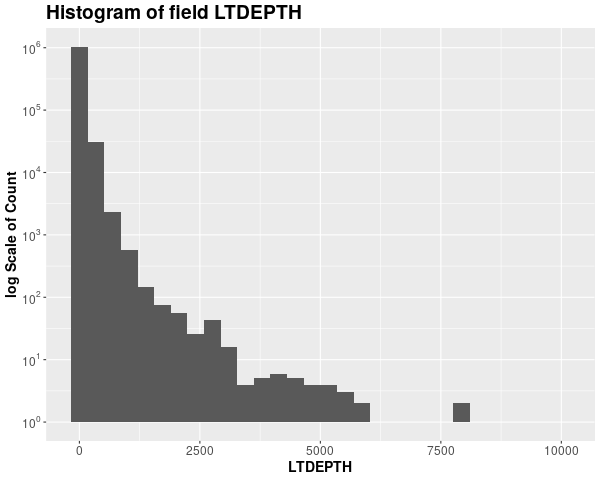
Description: Lot Frontage in feet.



Field 11

Field Name: LTDEPTH

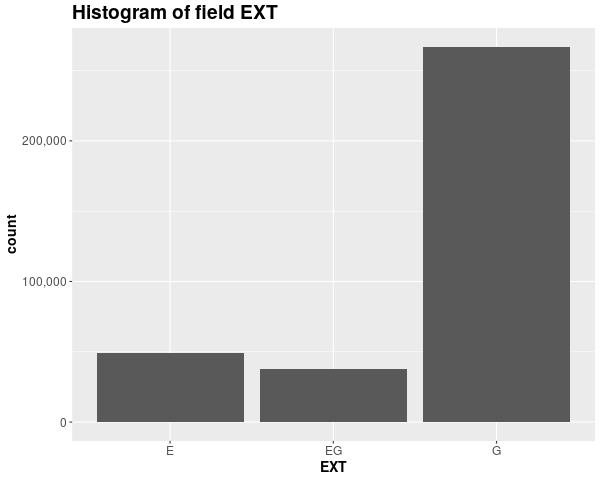
Description: Lot Depth in feet.



Field 12

Field Name: EXT

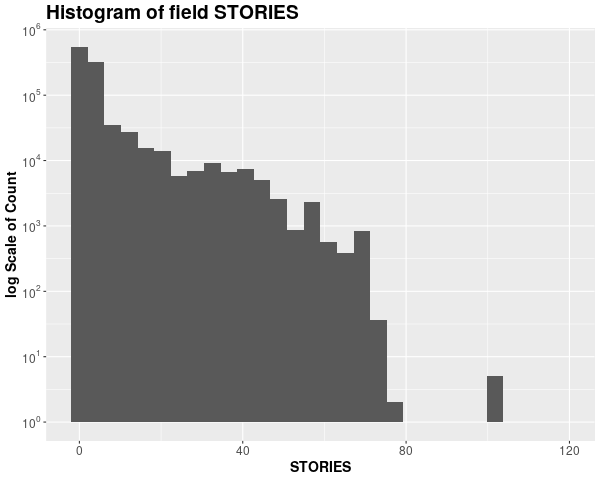
Description: Extension indicator, including garage and property extension.



Field 13

Field Name: STORIES

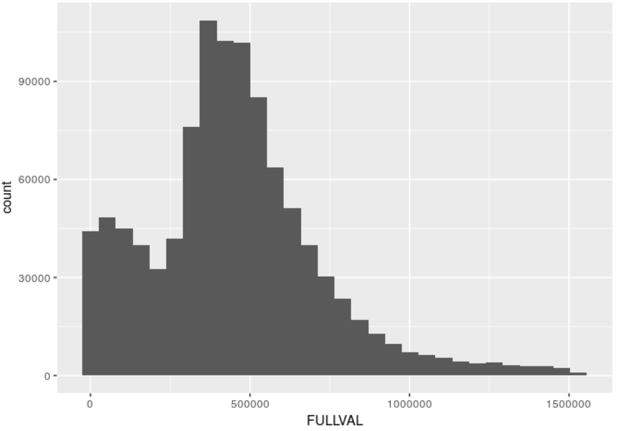
Description: The number of stories for the building (Number of Floors).



Field 14

Field Name: FULLVAL

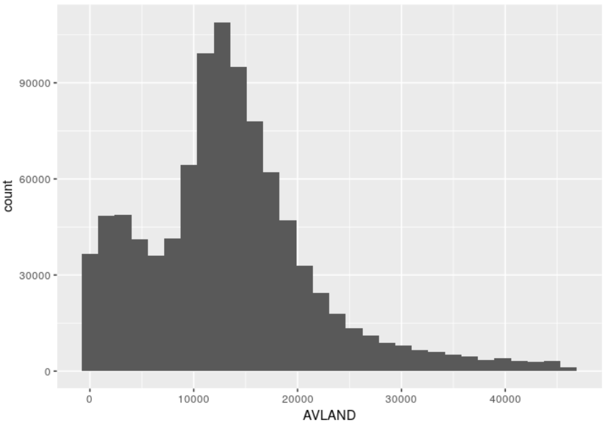
Description: If not zero, Current year’s total market value of the property.



Field 15

Field Name: AVLAND

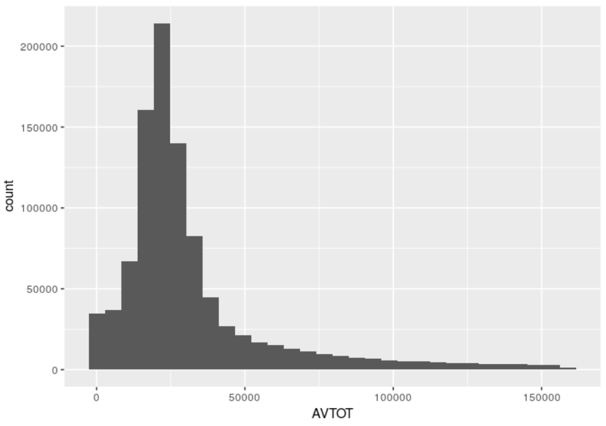
Description: Assessed land value.



Field 16

Field Name: AVTOT

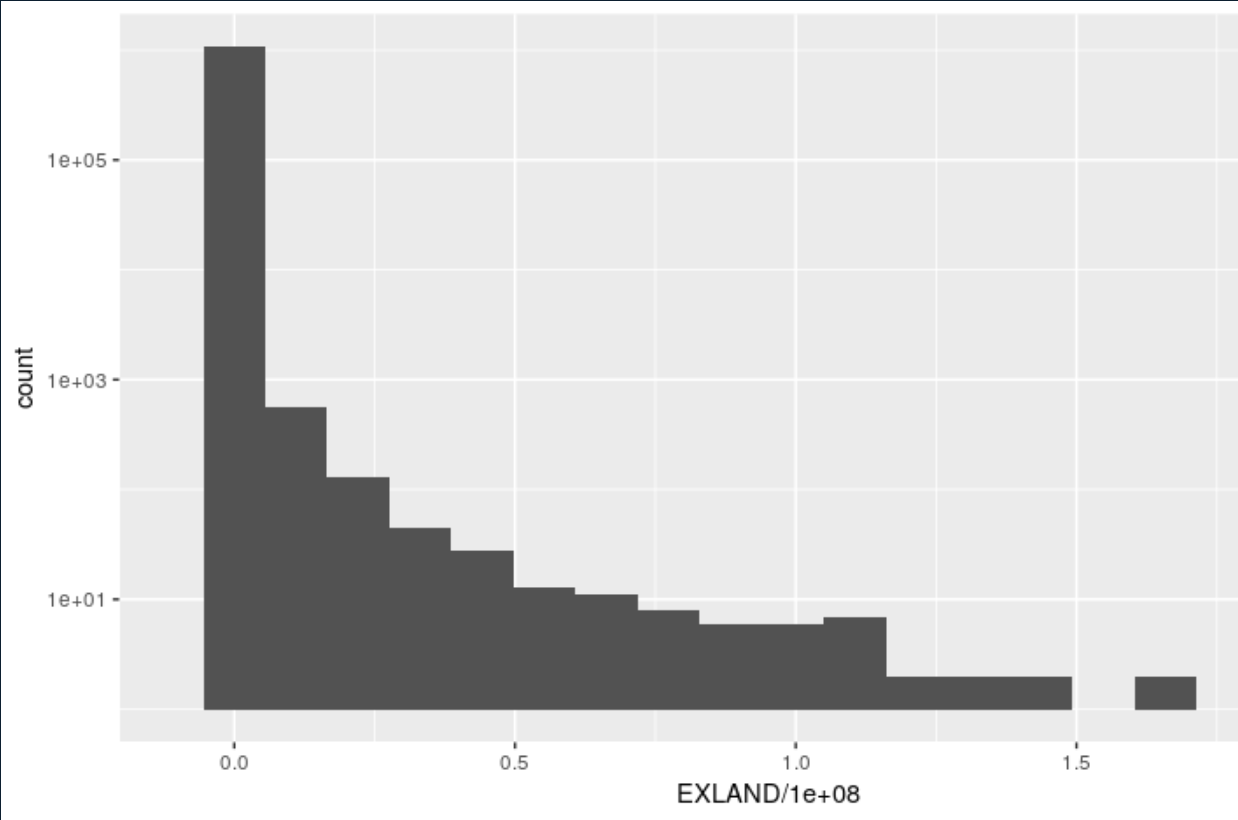
Description: Assessed total value.



Field 17

Field name: EXLAND

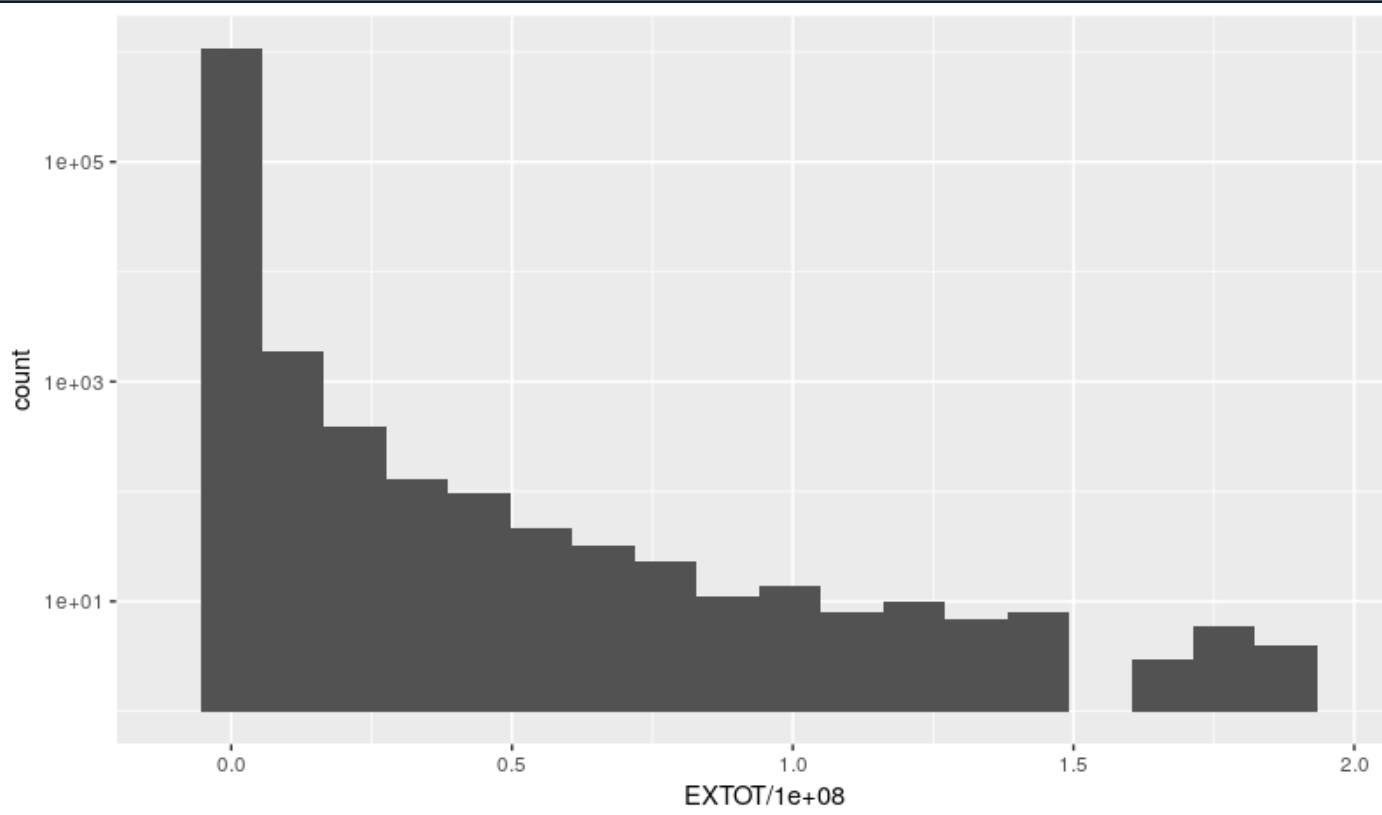
Description: Exempted land value.



Field 18

Field name: EXTOT

Description: Exempted Total Value

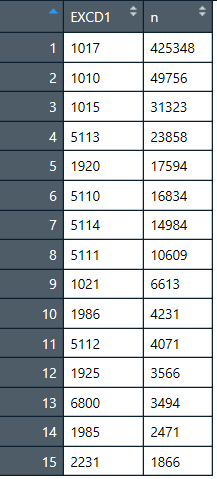


Field 19

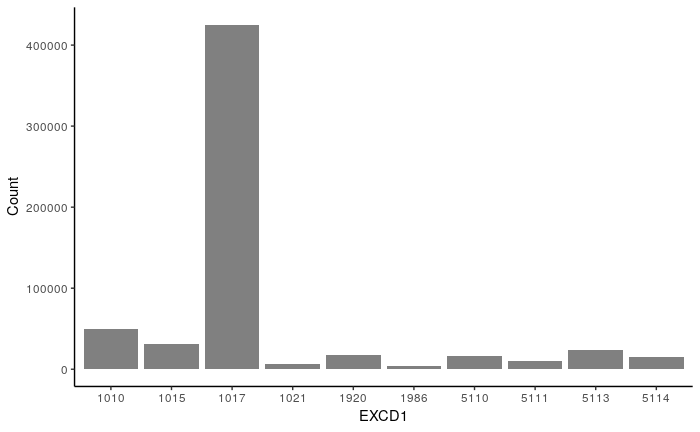
Field name: EXCD1

Description: Exempted Current Dollar Value

Top 10 Field Value



Top 10 Field Value plot

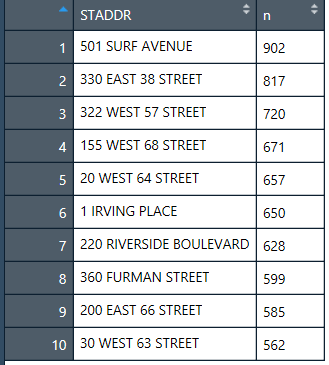


Field 20

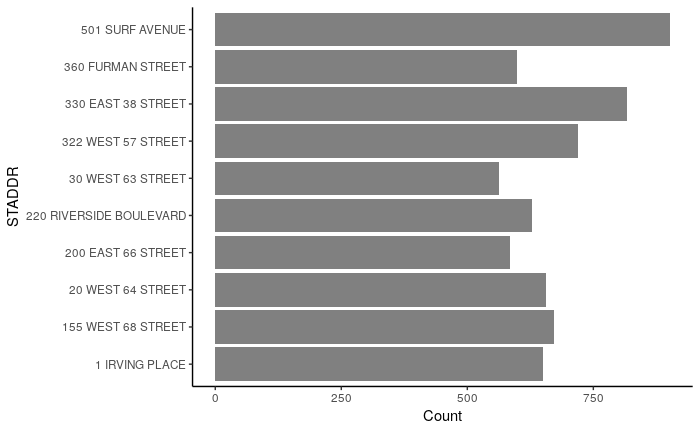
Field name: STADDR

Description: Property Street Address

Top 10 Field Value



Top 10 Field Value plot

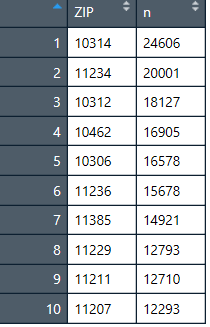


Field 21

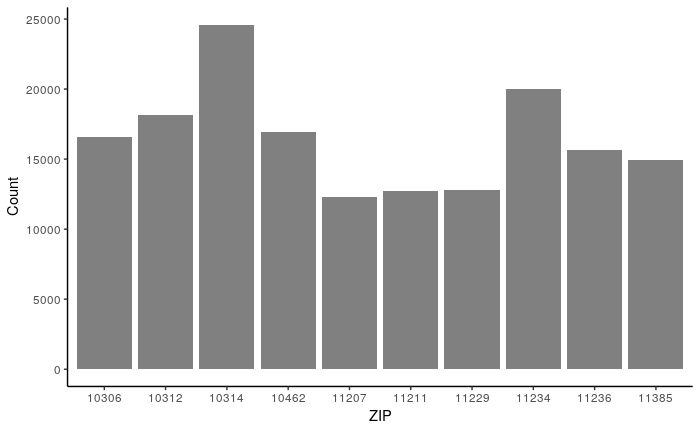
Field name: ZIP

Description: Postal Zip code of the property.

Top 10 Field Value of ZIP Code



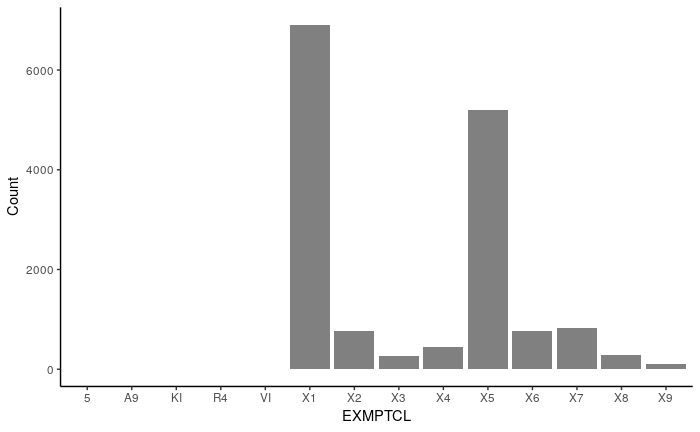
Top 10 Field Value of ZIP Code Plot



Field 22

Field name: EXMPTCL

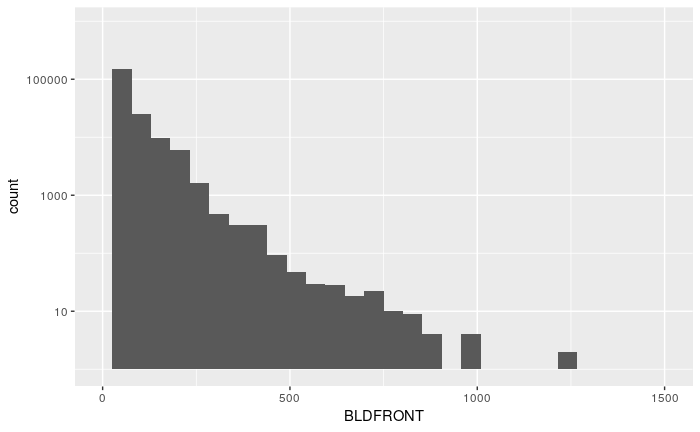
Description: Exempt Class used for fully exempt properties only



Field 23

Field name: BLDFRONT

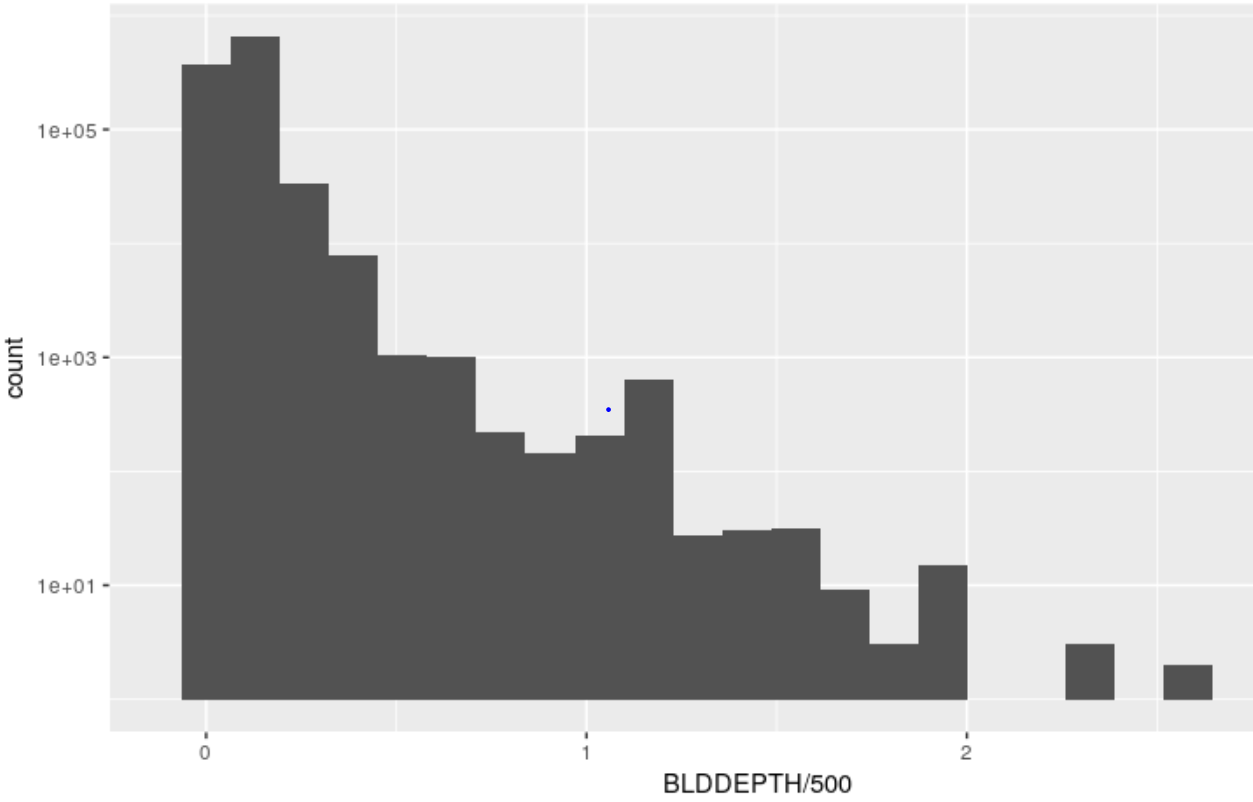
Description: Building Frontage in feet.



Field 24

Field name: BLDDEPTH

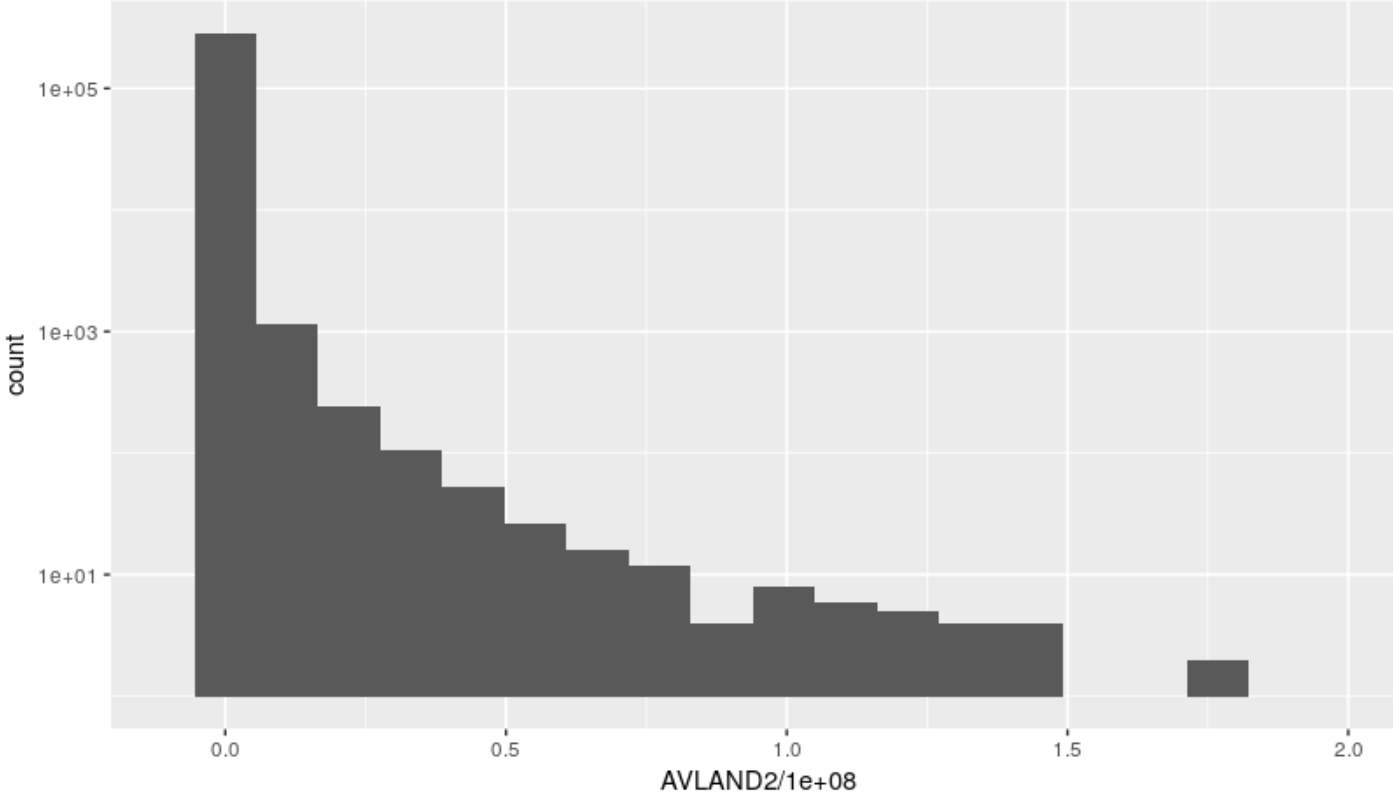
Description: Lot Depth in feet. (With Log Scale)



Field 25

Field name: AVLAND2

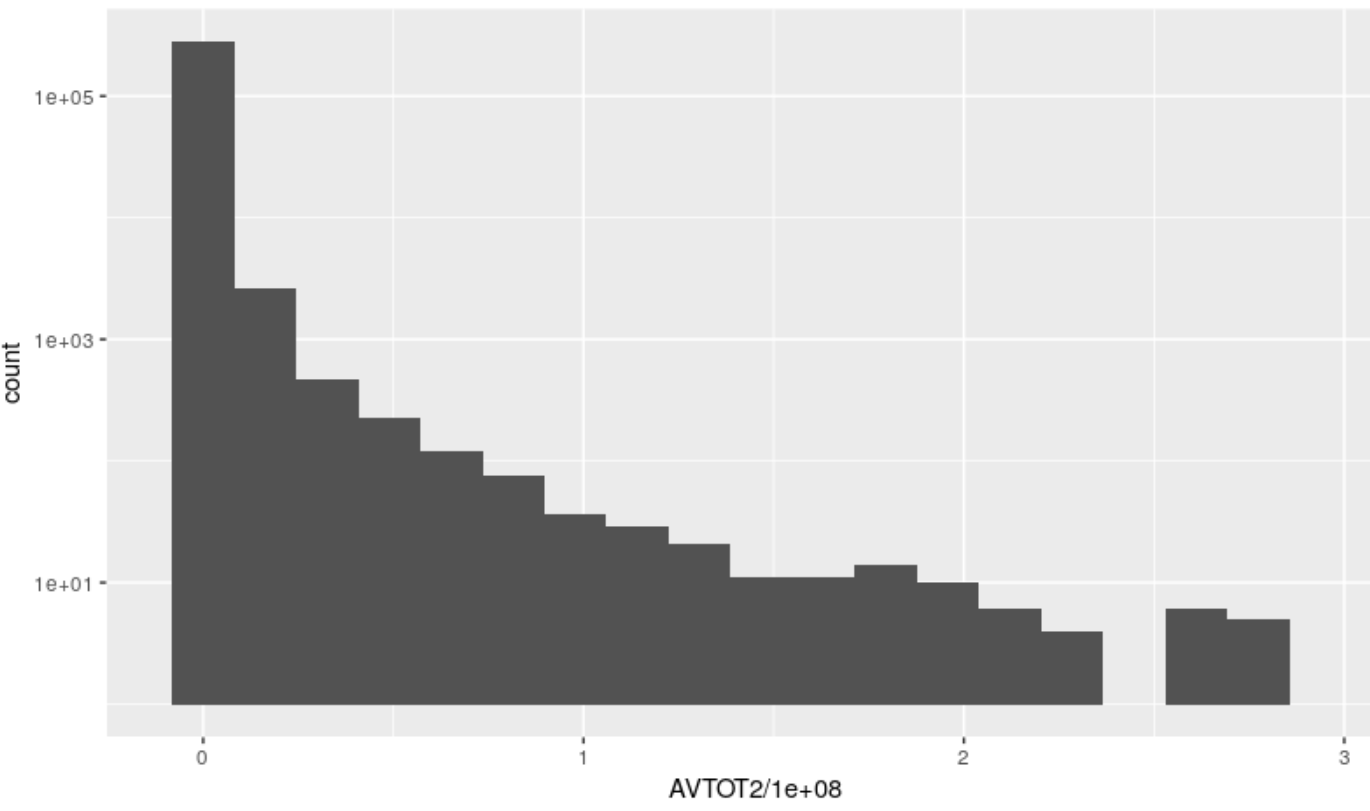
Description: Averaged Value of Land area (With Log Scale)



Field 26

Field name: AVTOT2

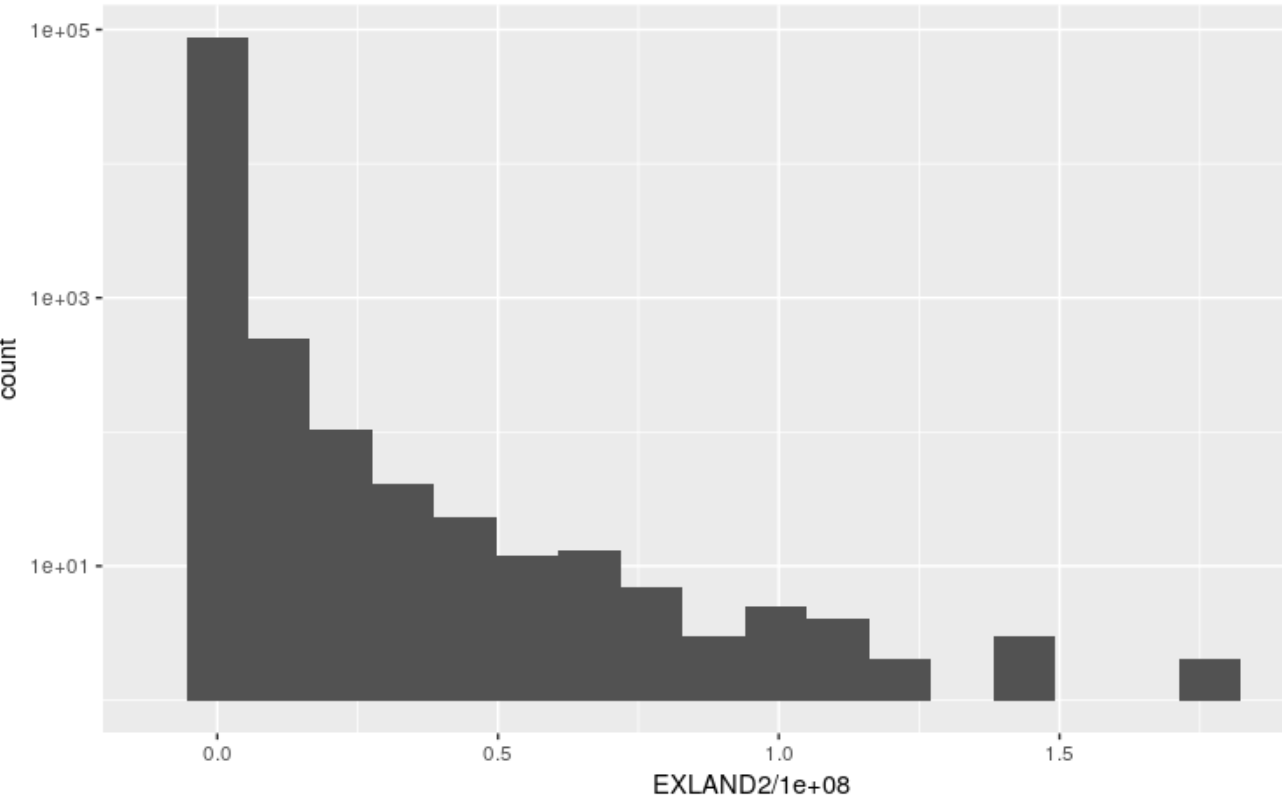
Description: Total Value area (With Log Scale)



Field 27

Field name: EXLAND2

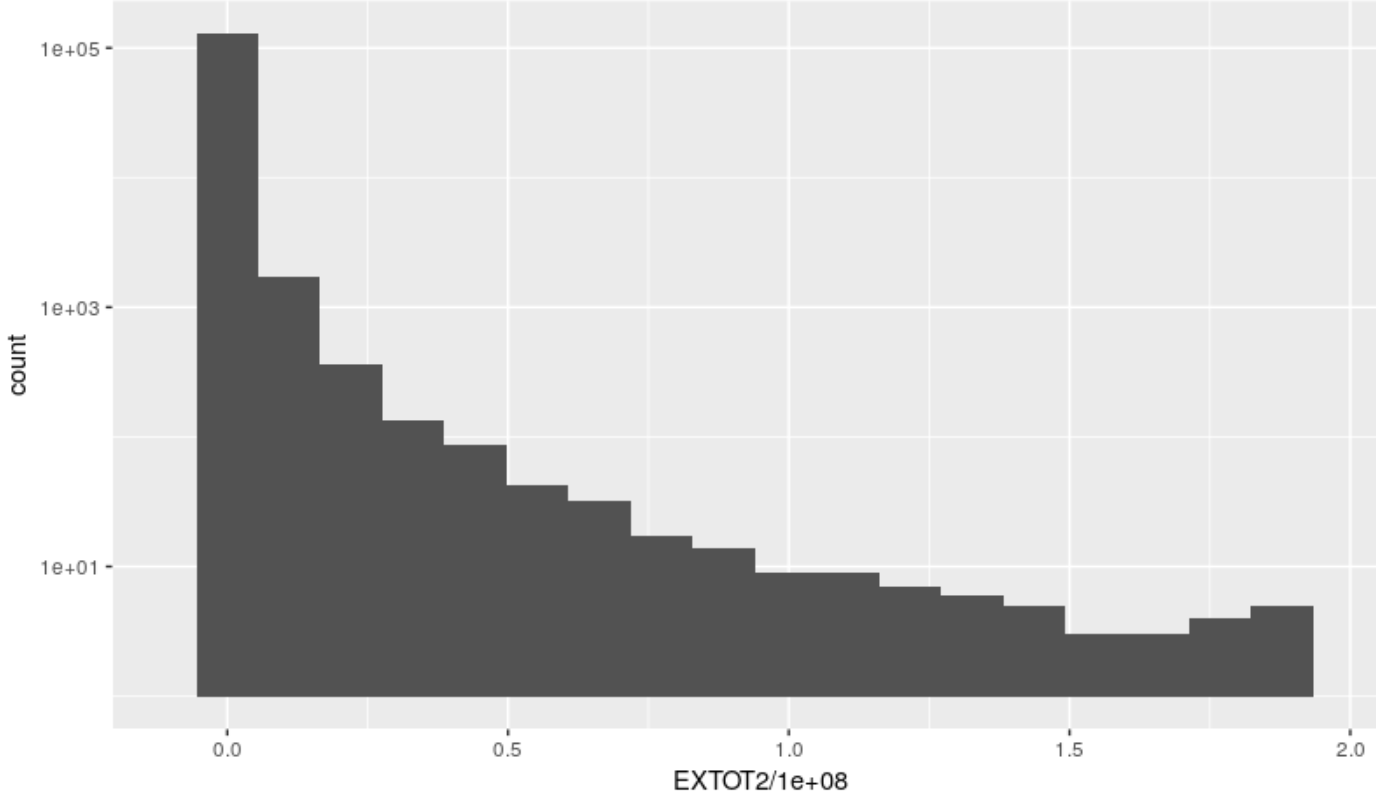
Description: Exempted Land Value area (With Log Scale)



Field 28

Field name: EXTOT2

Description: Exempted Total Value Area (With Scale)

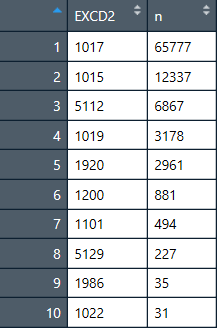


Field 29

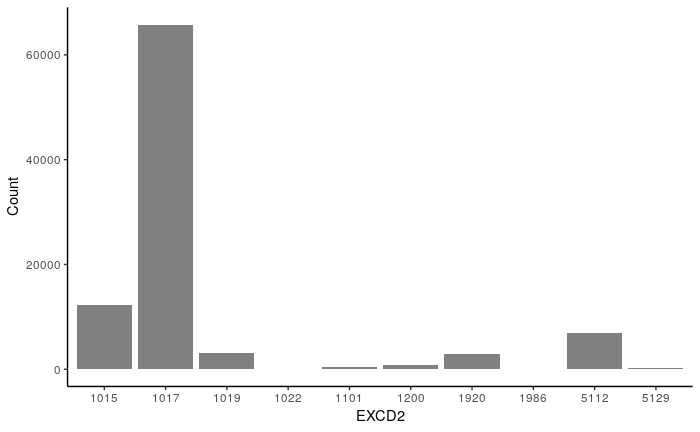
Field name: EXCD2

Description: Exempted Current Area

Top 10 Field Value of Exempted Current Area

****

Top 10 Field Value of Exempted Current area Plot



Field 30

Field name: PERIOD

Description: The Unique Period Value – FINAL Period

Field 31

Field name: YEAR

Description: The Unique Year Value - 2010/11 ( The information is in Nov 2010)

Field 32

Field name: VALTYPE

Description: The Unique Value Type is AC-TR

## **Appendix 2 Top 10 records of Principal Analysis results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **RECORD** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** |
| 632816 | 374.63 | 476.30 | -737.94 | 357.88 | 45.35 |
| 565392 | 835.68 | -130.91 | 419.64 | 187.06 | -1.30 |
| 1067360 | 153.30 | -673.24 | -451.05 | -159.73 | -185.30 |
| 917942 | 205.67 | 105.14 | 125.75 | -80.38 | -584.23 |
| 85886 | 135.50 | 121.35 | -38.03 | -340.18 | 234.35 |
| 556609 | 61.50 | -144.65 | 35.11 | 130.61 | 275.04 |
| 912501 | 73.60 | -112.86 | 23.70 | 38.05 | 276.93 |
| 821853 | 64.31 | -196.77 | -17.44 | 107.63 | 180.94 |
| 776306 | 53.19 | -156.76 | 2.74 | 115.51 | 199.36 |
| 770594 | 32.77 | -172.63 | -154.26 | -94.63 | -119.01 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **RECORD** | **Max** | **Min** | **Sum of PCs** | **Squared sum of PCs** | **Euclidean** | **Rank** |
| 632816 | 476.30 | -737.94 | 1,992.10 | 1,041,901.30 | 1,020.74 | 1 |
| 565392 | 835.68 | -130.91 | 1,574.59 | 926,592.30 | 962.60 | 2 |
| 1067360 | 153.30 | -673.24 | 1,622.62 | 740,052.75 | 860.26 | 3 |
| 917942 | 205.67 | -584.23 | 1,101.18 | 416,954.36 | 645.72 | 4 |
| 85886 | 234.35 | -340.18 | 869.40 | 205,170.37 | 452.96 | 5 |
| 556609 | 275.04 | -144.65 | 646.92 | 118,646.53 | 344.45 | 6 |
| 912501 | 276.93 | -112.86 | 525.14 | 96,853.35 | 311.21 | 7 |
| 821853 | 180.94 | -196.77 | 567.08 | 87,479.63 | 295.77 | 8 |
| 776306 | 199.36 | -156.76 | 527.56 | 80,497.55 | 283.72 | 9 |
| 770594 | 32.77 | -172.63 | 573.29 | 77,787.19 | 278.90 | 10 |