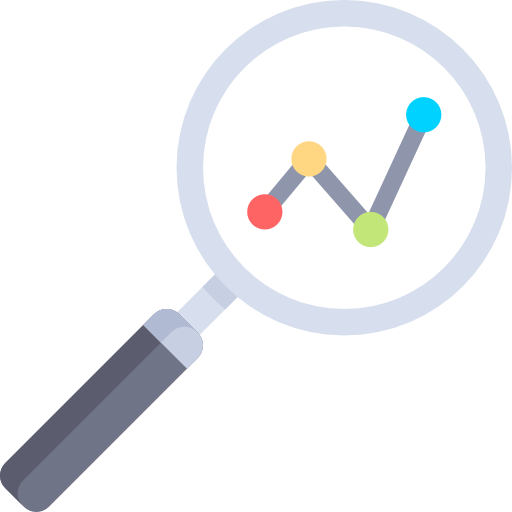
**New York Real Estate Tax Fraud Analytics Model**



Feb 18th, 2018

**Team Members –**

**Gyan Prakash, Yufei Wang, Wei Tang, Alok Abhishek, Zhang (Alex) Weichen, Pratyush Shankar**

Table of Contents

[Executive Summary 3](#_Toc506749977)

[Description of Data 4](#_Toc506749978)

[Data Cleaning method 5](#_Toc506749979)

[Variable Creation 6](#_Toc506749980)

[Algorithms 7](#_Toc506749981)

[Results and observations 8](#_Toc506749982)

[Conclusion 9](#_Toc506749983)

[Appendix 10](#_Toc506749984)

[Data Quality Report: New York Real Estate data set… 11](#_Toc506749985)

# Executive Summary

As per study by Tax justice network the US economy loses about 9% of its GDP in Tax evasion. Studying tax fraud therefore is very important in order to better understand the drivers of tax fraud and how to address them.

This report specifically focuses on property tax fraud analysis using unsupervised machine learning techniques to develop insights into property tax fraud in the NYC real estate market. The report usages NYC tax collection data from 2010, with over 1 million records to analyze anomalies, develop insights and rank each property with a fraud score.

Following diagram illustrates the overview of 11 step process followed to develop fraud score:

As part of this project fraud detection was done using weighted average of outlier detection and autoencoder. A weighted average score was used to identify the properties with high tax fraud potential.

Based on the fraud score some of the parks and government buildings were flagged for potential fraud. These records have been excluded because government buildings are large low stories buildings which are very high in value, these are outliers in New York but not frauds.

Looking at the remaining top fraud candidates – the fraud algorithm has flagged small buildings with very high value and large buildings with very low value as potential fraud. Some of these could have been flagged because some data about the property is missing.

More details of potential fraud candidates are discussed in the result and observation section of the report.

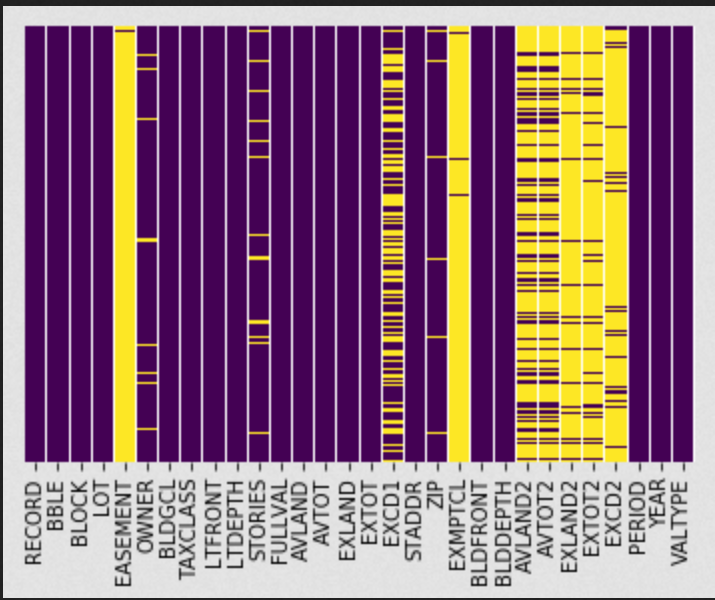
# Description of Data

The New York City department of Finance values properties in NYC every year to calculate the property tax. This report provides property tax data such as market and assessed values, exemptions, and abatements from the assessment year 2010/11. The information is listed by categories, such as borough, tax class, and type of building. There are 1048575 records, 30 columns, which includes 14 categorical variables and 16 numerical variables. The table below shows some basic information of this dataset, like data type, mean, standard deviation, maximum and minimum.

**Dataset Description**

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| LTFRONT | Lot frontage in feet |
| LTDEPTH | Lot depth in feet |
| FULLVAL | Total market value of property |
| AVLAND | Market value of the land |
| AVTOT | Total market value |
| EXLAND | Exempt land value |
| EXTOT | Exempt total value |
| EXCD1 | Exempt condo value |
| BLDFRONT | Building frontage in feet |
| BLDDEPTH | Building depth in feet |
| AVLAND2 | 2nd Market value of the land |
| AVTOT2 | 2nd Total market value |
| EXLAND2 | 2nd Exempt land value |
| EXTOT2 | 2nd Exempt total value |
| EXCD2 | 2nd Exempt condo value |
| BLDGCL | Building class |

Real world data is messy and you can see in the heat map of missing data below that a lot of data is missing from the data set:



Descriptive statistics of the overall data is as shown below:



# Data Cleaning method

**Data Imputation:**

**Step 1 - ZIP**

Fill in 10935 to missing ZIP cells.

**Reason:** 10935 is the average of all non-blank cells ZIP column.

**Step 2 – LTFRONT, LOTDEPTH, BLDFRONT, BLDDEFPTH**

Filling in missing values and replace zero values by 40, 100, 30, 50, respectively.

**Reason:** 40, 100, 30, and 50 are the averages of all non-black and non-zero cells in LTFRONT, LOTDEPTH, BLDFRONT, BLDDEFPTH columns.

**Step 3 – FULLVAL, AVLAND, AVTOT**

Filling in missing values and replace zero values by 880,000, 86,000, 230,000.

**Reason:** 880,487.7, 85,995.0, and 230,758.2 are the averages of all non-blank cells in FULLVAL column.

**Step 4 – EXLAND, EXTOT**

Filling in missing values and replace zero values by 1620.

**Reason:** 1620 is the mode for EXLAND and EXTOT columns. 33.1% EXLAND value is 1620; 32.8% value of EXTOT is 1620.

**Step 5 – STORIES**

If STORIES=0, set it to average stories by zip code.

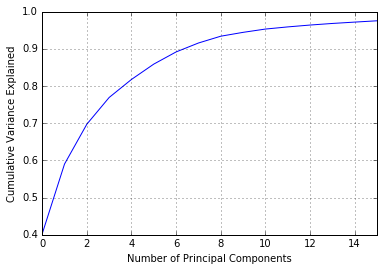
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ZIP** | **STORIES** | **ZIP** | **STORIES** | **ZIP** | **STORIES** | **ZIP** | **STORIES** |
| 10001 | 11 | 10302 | 2 | 11201 | 11 | 11362 | 2 |
| 10002 | 6 | 10303 | 2 | 11203 | 2 | 11363 | 2 |
| 10003 | 10 | 10304 | 2 | 11204 | 2 | 11364 | 2 |
| 10004 | 36 | 10305 | 2 | 11205 | 4 | 11365 | 2 |
| 10005 | 33 | 10306 | 2 | 11206 | 4 | 11366 | 2 |
| 10006 | 32 | 10308 | 2 | 11207 | 3 | 11367 | 3 |
| 10007 | 14 | 10310 | 2 | 11208 | 3 | 11368 | 3 |
| 10009 | 6 | 10312 | 2 | 11209 | 3 | 11369 | 2 |
| 10010 | 21 | 10314 | 2 | 11210 | 3 | 11370 | 2 |
| 10011 | 10 | 10451 | 4 | 11211 | 6 | 11372 | 3 |
| 10012 | 6 | 10452 | 3 | 11212 | 2 | 11373 | 3 |
| 10013 | 8 | 10453 | 3 | 11213 | 3 | 11374 | 6 |
| 10014 | 9 | 10454 | 3 | 11214 | 3 | 11375 | 4 |
| 10016 | 25 | 10455 | 3 | 11215 | 4 | 11377 | 2 |
| 10017 | 30 | 10456 | 3 | 11216 | 3 | 11378 | 2 |
| 10018 | 22 | 10457 | 4 | 11217 | 4 | 11379 | 2 |
| 10019 | 33 | 10458 | 3 | 11218 | 3 | 11385 | 2 |
| 10020 | 48 | 10459 | 3 | 11219 | 3 | 11411 | 2 |
| 10021 | 24 | 10460 | 3 | 11220 | 3 | 11412 | 2 |
| 10022 | 25 | 10461 | 2 | 11221 | 3 | 11413 | 2 |
| 10023 | 26 | 10462 | 8 | 11222 | 3 | 11414 | 2 |
| 10024 | 10 | 10463 | 5 | 11223 | 2 | 11415 | 4 |
| 10025 | 14 | 10464 | 2 | 11224 | 10 | 11416 | 2 |
| 10026 | 10 | 10465 | 2 | 11225 | 3 | 11417 | 2 |
| 10027 | 5 | 10466 | 2 | 11226 | 3 | 11418 | 3 |
| 10028 | 15 | 10467 | 3 | 11228 | 2 | 11419 | 2 |
| 10029 | 6 | 10468 | 4 | 11229 | 2 | 11420 | 2 |
| 10030 | 6 | 10469 | 2 | 11230 | 3 | 11421 | 2 |
| 10031 | 5 | 10470 | 2 | 11231 | 3 | 11422 | 2 |
| 10032 | 6 | 10471 | 3 | 11232 | 3 | 11423 | 2 |
| 10033 | 5 | 10472 | 2 | 11233 | 3 | 11426 | 2 |
| 10034 | 5 | 10473 | 3 | 11234 | 2 | 11427 | 2 |
| 10035 | 7 | 10474 | 2 | 11235 | 4 | 11428 | 2 |
| 10036 | 34 | 10475 | 2 | 11236 | 2 | 11429 | 2 |
| 10037 | 4 | 10803 | 3 | 11237 | 3 | 11432 | 3 |
| 10038 | 19 | 11001 | 2 | 11238 | 4 | 11433 | 2 |
| 10039 | 8 | 11004 | 2 | 11239 | 3 | 11434 | 2 |
| 10040 | 6 | 11040 | 2 | 11243 | 41 | 11435 | 3 |
| 10044 | 16 | 11101 | 6 | 11354 | 5 | 11436 | 2 |
| 10065 | 23 | 11102 | 6 | 11355 | 6 | 11691 | 3 |
| 10069 | 35 | 11103 | 3 | 11356 | 2 | 11692 | 2 |
| 10075 | 22 | 11104 | 3 | 11357 | 2 | 11693 | 3 |
| 10128 | 25 | 11105 | 2 | 11358 | 2 | 11694 | 3 |
| 10280 | 27 | 11106 | 3 | 11360 | 6 | 10935 | 4 |
| 10301 | 3 | 11109 | 18 | 11361 | 2 |  |  |

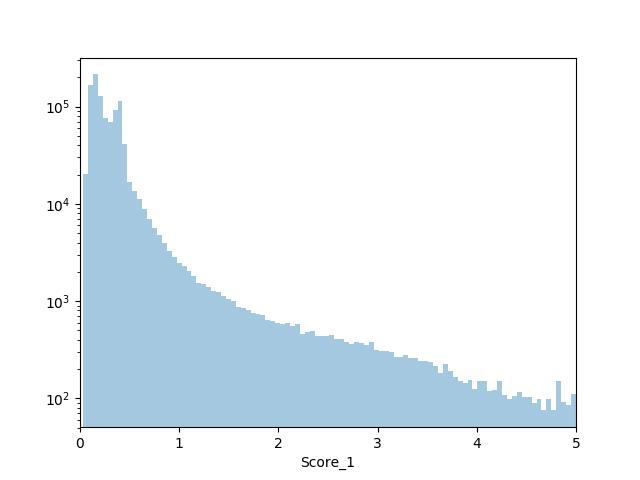
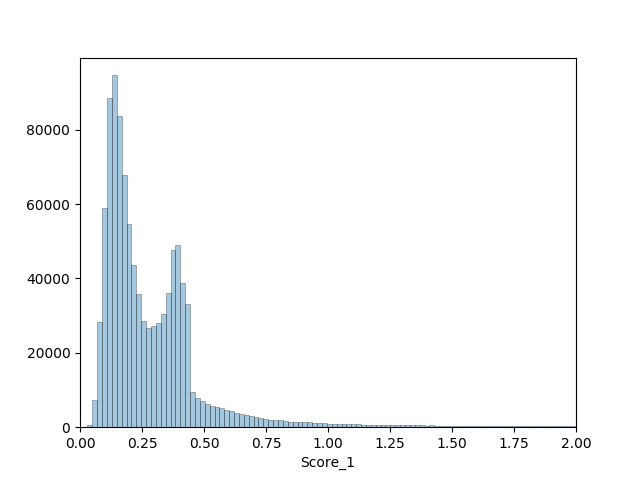
# Variable Creation

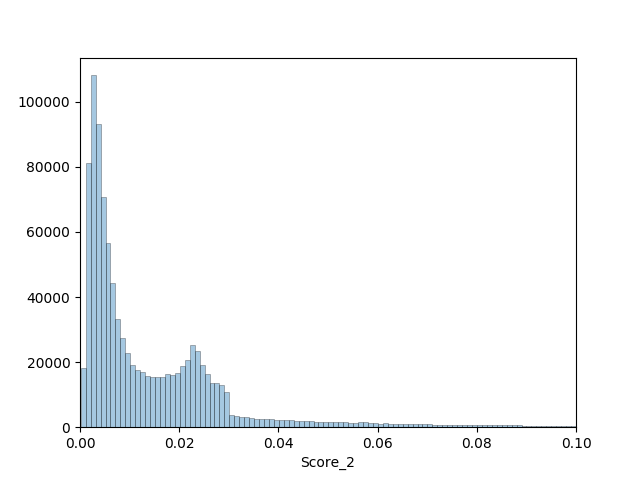
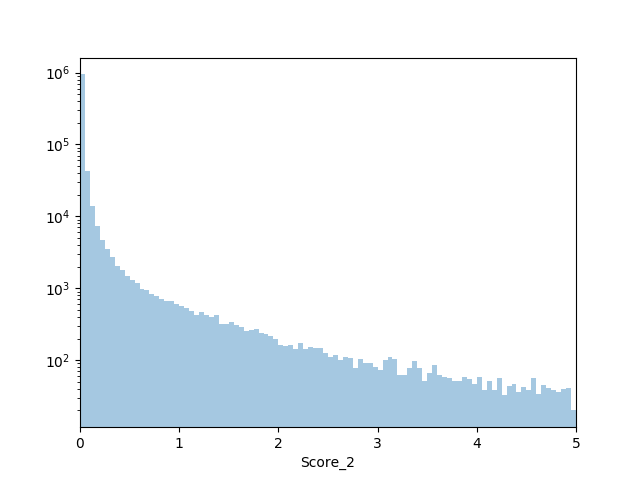
# Algorithms

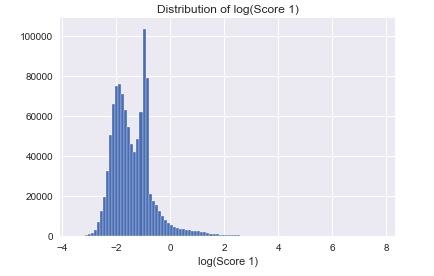
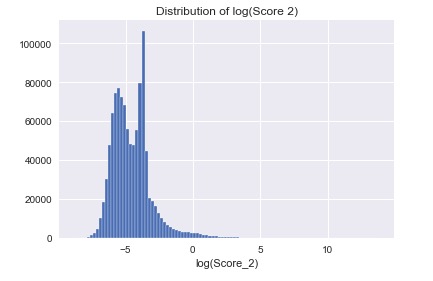
# Results and observations

We started with with reducing the dimensanality of the data. PCA result is shown in diagram below. Starting with over 75 variables, we reduced the number of variabels to 8, which still explained about 93% of the variation in the data.











# Conclusion

Few paragraphs summarizing everything you did and what else you might do with more time.

1. Data analysis
2. Data exploration
3. Variable creation
4. Scaling
5. Supervised learning
6. Result analysis
7. What can we do to improve.

# Appendix

## Data Quality Report: New York Real Estate data set…

**1. Introduction**

The New York City department of Finance values properties in NYC every year to calculate the property tax. This report provides property tax data such as market and assessed values, exemptions, and abatements from the assessment year 2010/11. The information is listed by categories, such as borough, tax class, and type of building. There are 1048575 records, 30 columns, which includes 14 categorical variables and 16 numerical variables. The table below shows some basic information of this dataset, like data type, mean, standard deviation, maximum and minimum.

**2. Dataset Description**

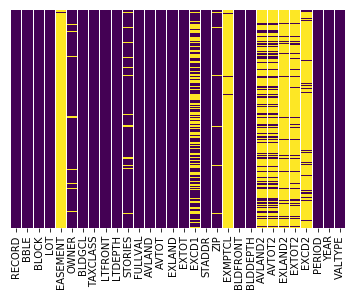
**Acronyms:**

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| LTFRONT | Lot frontage in feet |
| LTDEPTH | Lot depth in feet |
| FULLVAL | Total market value of property |
| AVLAND | Market value of the land |
| AVTOT | Total market value |
| EXLAND | Exempt land value |
| EXTOT | Exempt total value |
| EXCD1 | Exempt condo value |
| BLDFRONT | Building frontage in feet |
| BLDDEPTH | Building depth in feet |
| AVLAND2 | 2nd Market value of the land |
| AVTOT2 | 2nd Total market value |
| EXLAND2 | 2nd Exempt land value |
| EXTOT2 | 2nd Exempt total value |
| EXCD2 | 2nd Exempt condo value |
| BLDGCL | Building class |

**Summary table:**



**Heat map of missing values in the dataset:**



**3. Numerical Data Analysis**

**BLOCK – Block #**

As the block id in different cities may be same, it cannot effectively show the real number of properties in a specific block.

Bock # to area mapping:

Manhattan - 1 to 2,255

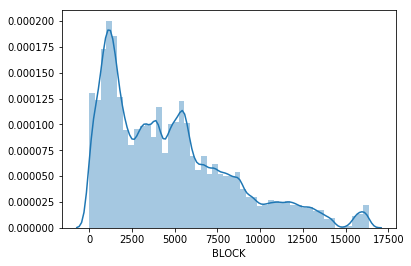
Bronx - 2,260 to 5,958

Brooklyn - 1 to 8,955

Queens - 1 to 16,350

Staten Island - 1 to 8,050

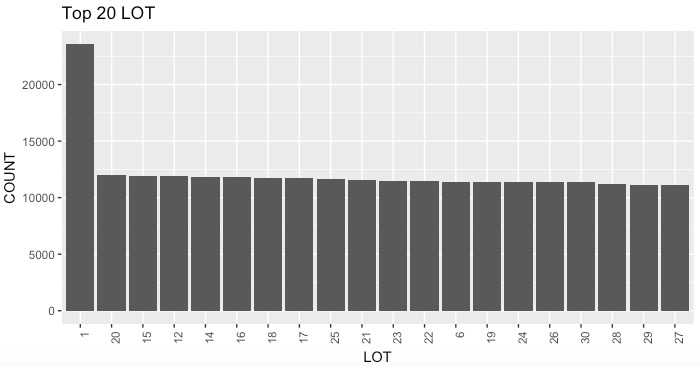


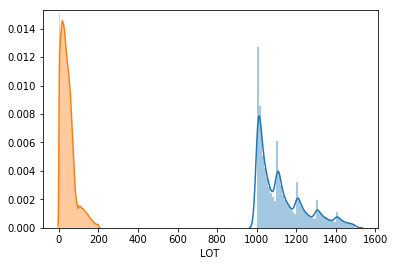
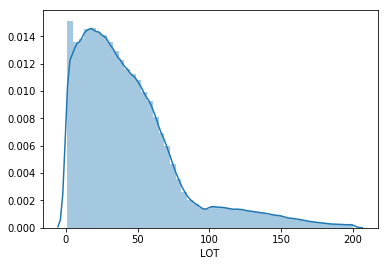


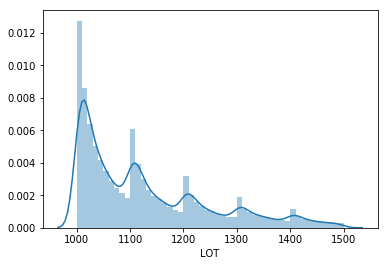
**LOT – Lot # within Block**

Every record has a lot id but some of them share the same lot id. Value 1 has the highest frequency in this dataset, but lot 1 may not be the lot with highest number of properties as lot id is only unique within block.



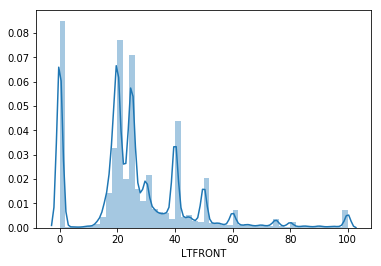
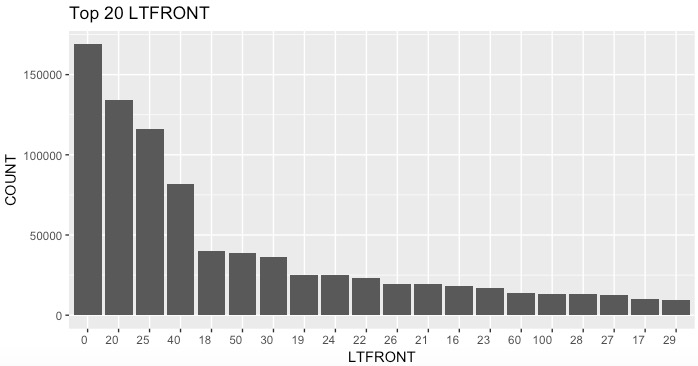






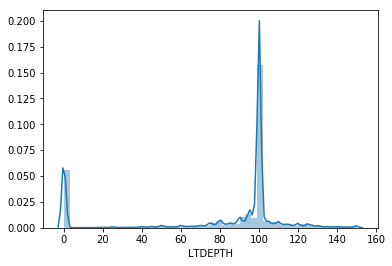
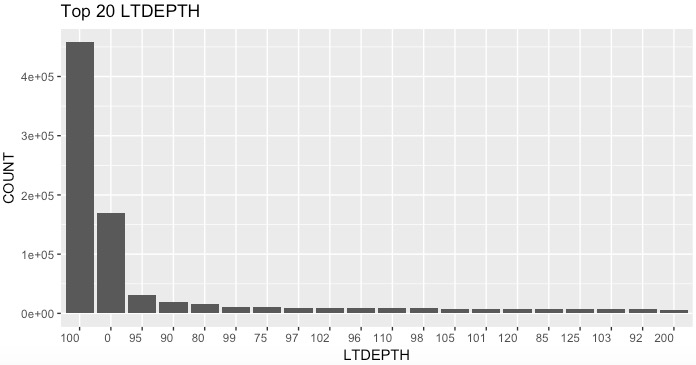
**LTFRONT – Lot Width**





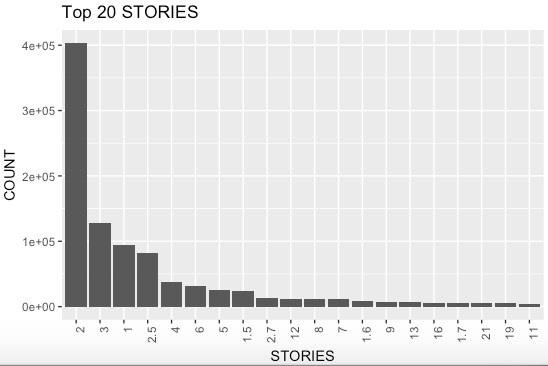
**LTDEPTH – Lot Depth**

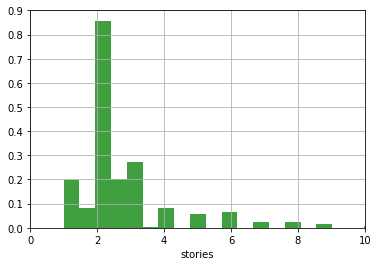
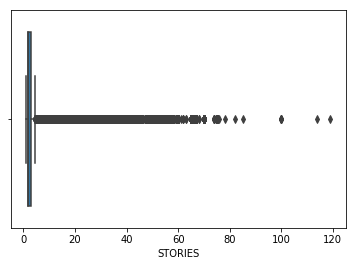


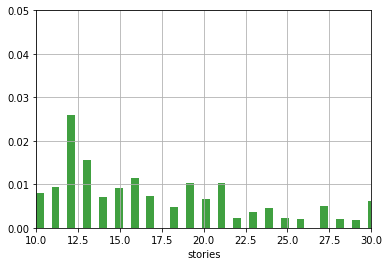
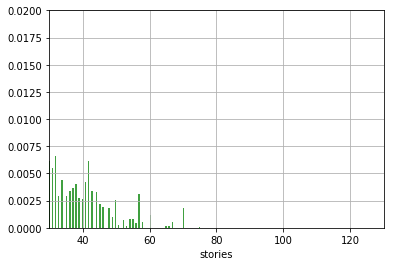


**STORIES – Number of stories in the building**



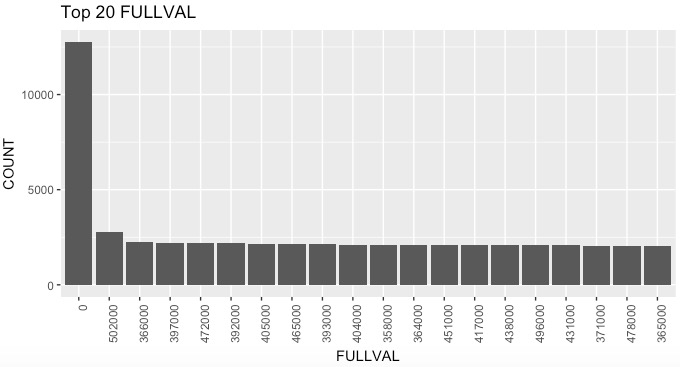


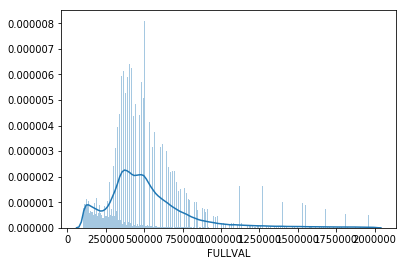




**FULLVAL – Market Value**

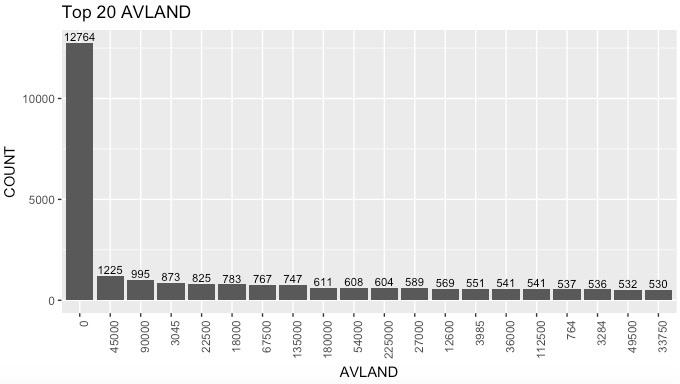


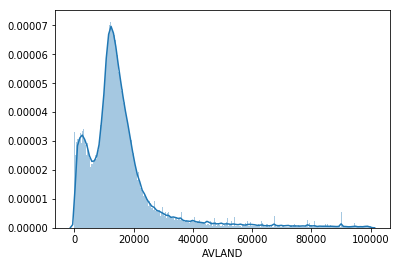




**AVLAND – Actual Land Value**

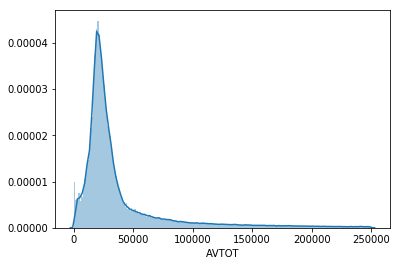
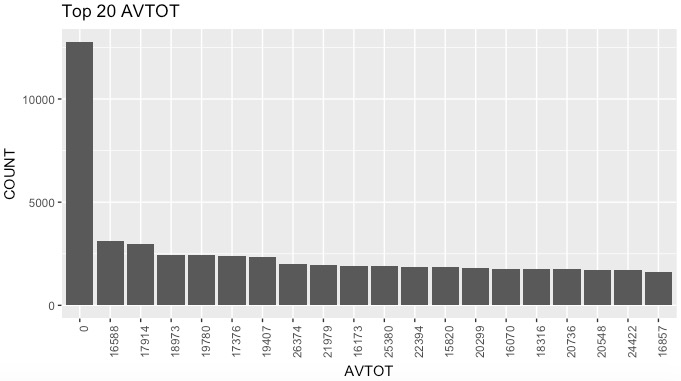






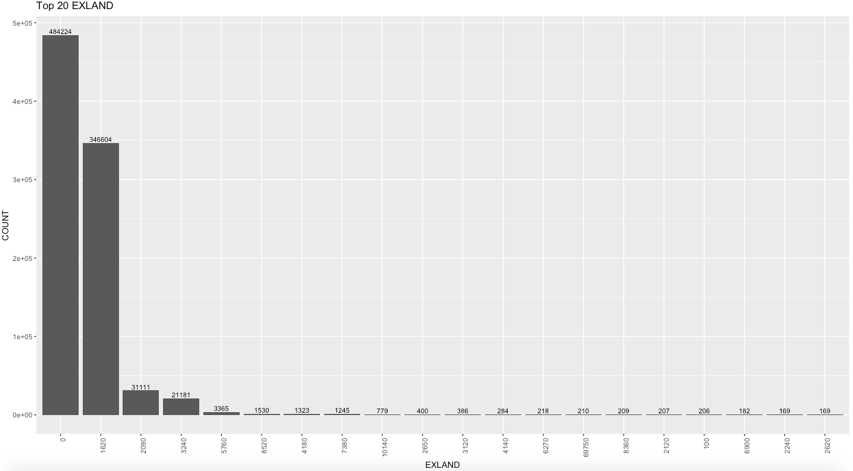
**AVTOT – Actual Total Value**

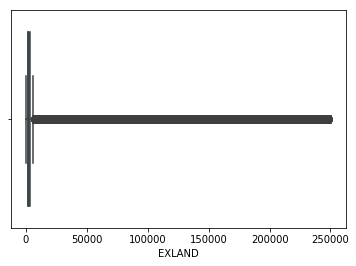


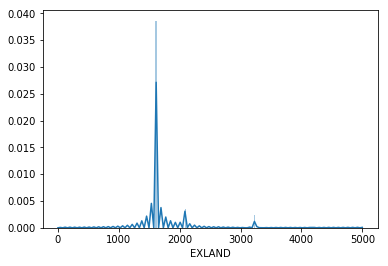


**EXLAND – Actual Exempt Land Value**



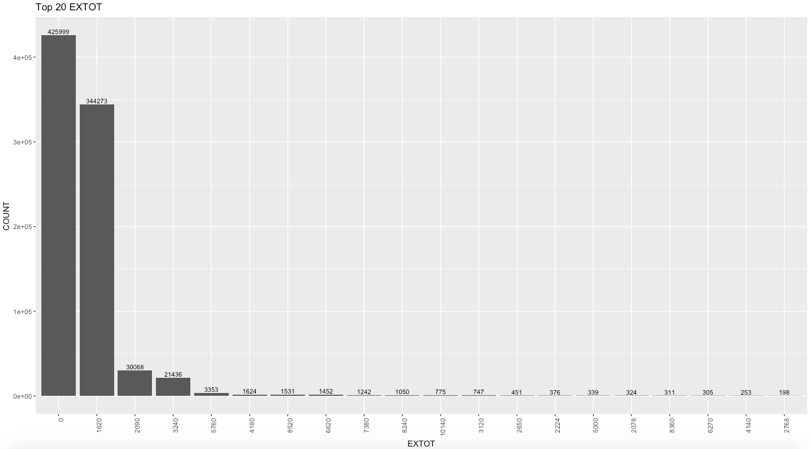


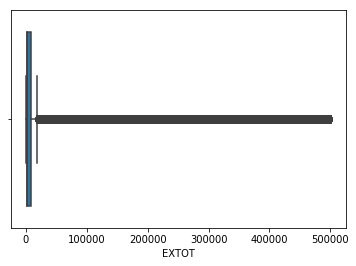
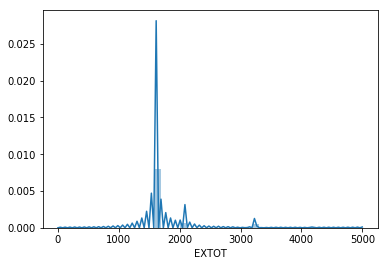




**EXTOT – Actual Exempt Land Total**

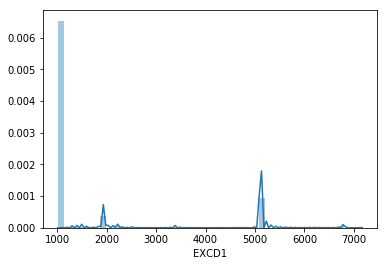
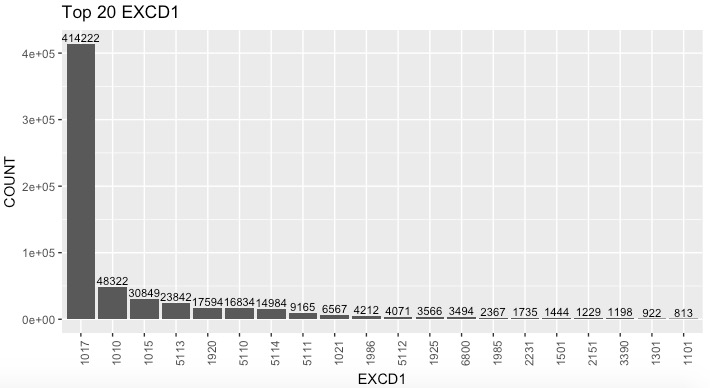




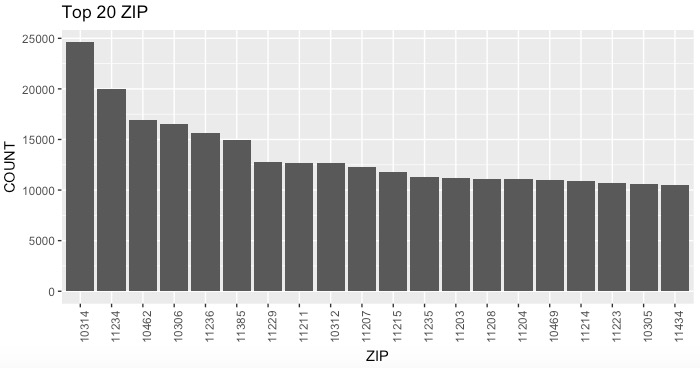
**EXCD1 – Exemption Code 1**

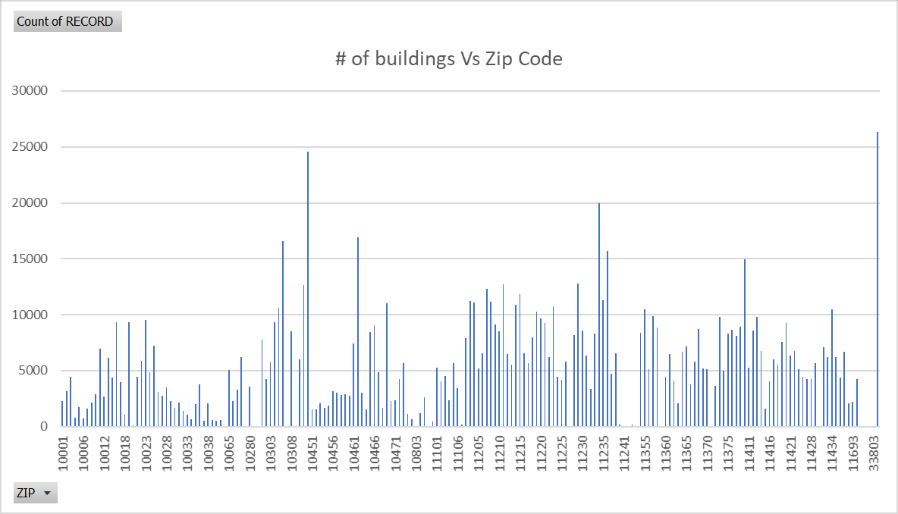




**ZIP – Zip Code**

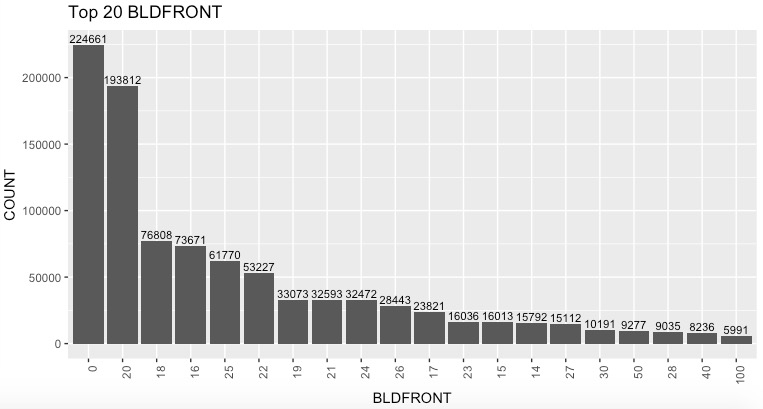


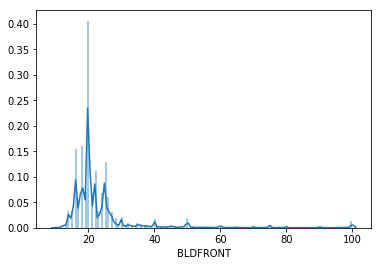




**BLDFRONT – Building Width**

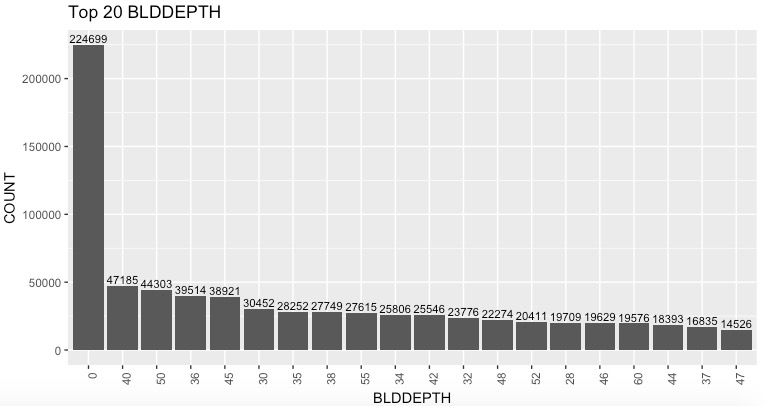


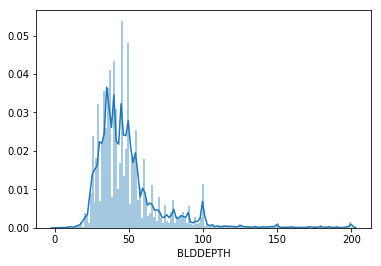




**BLDDEPTH – Building Depth**

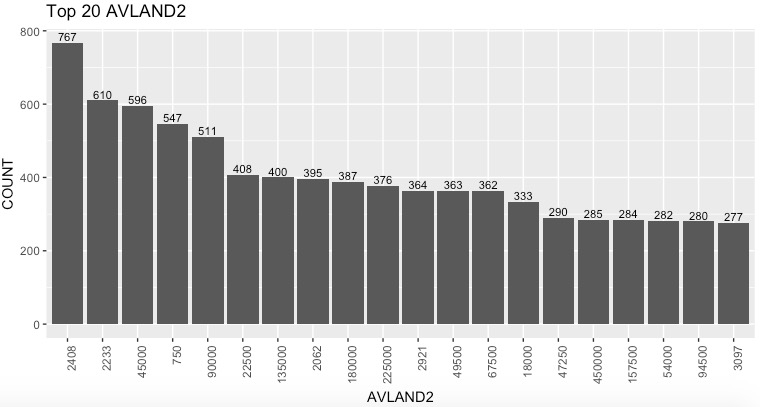


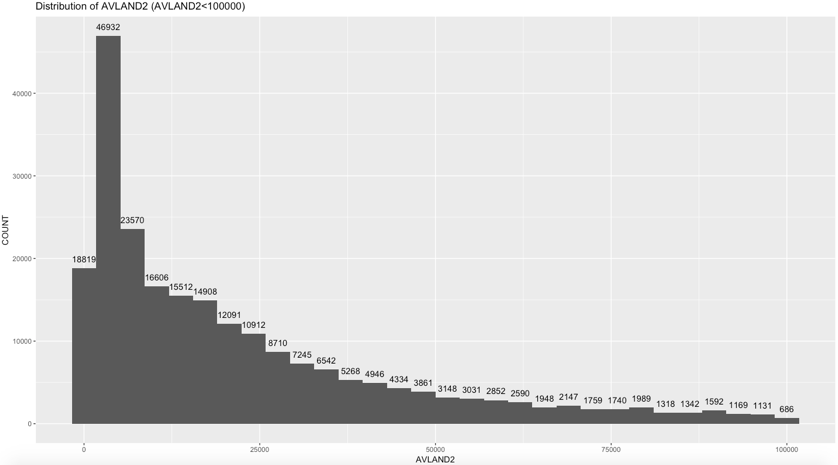




**AVLAND2 – Transitional Land Value**

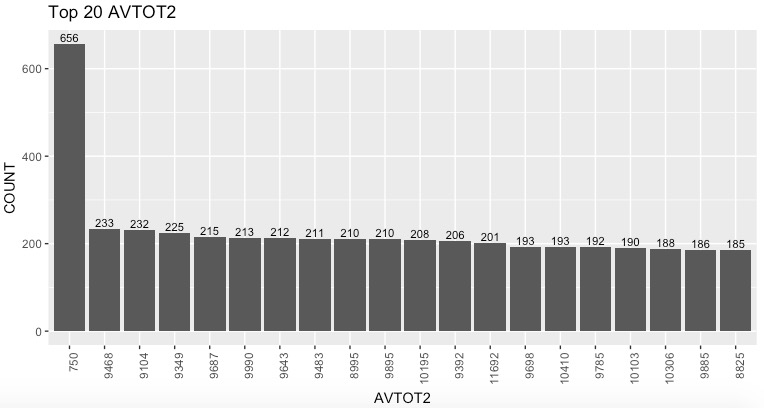


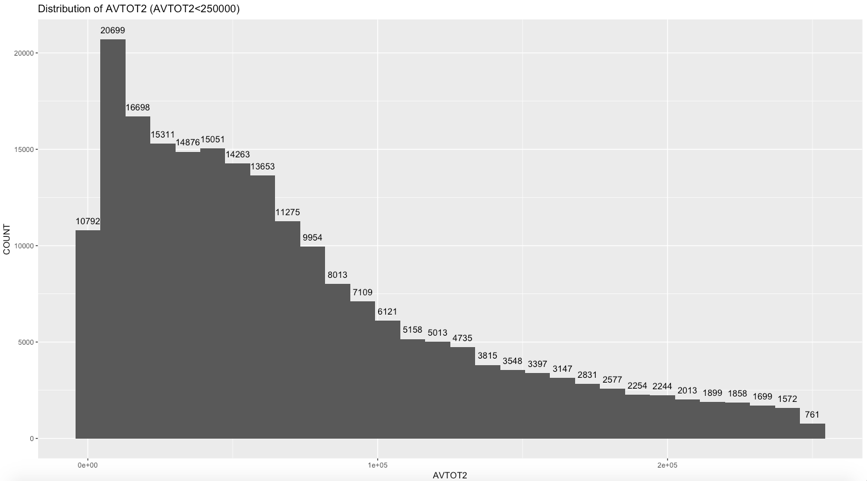




**AVTOT2 – Transitional Total Value**

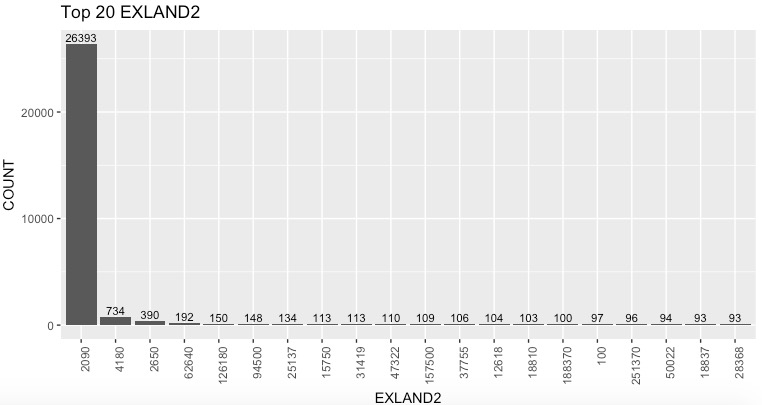


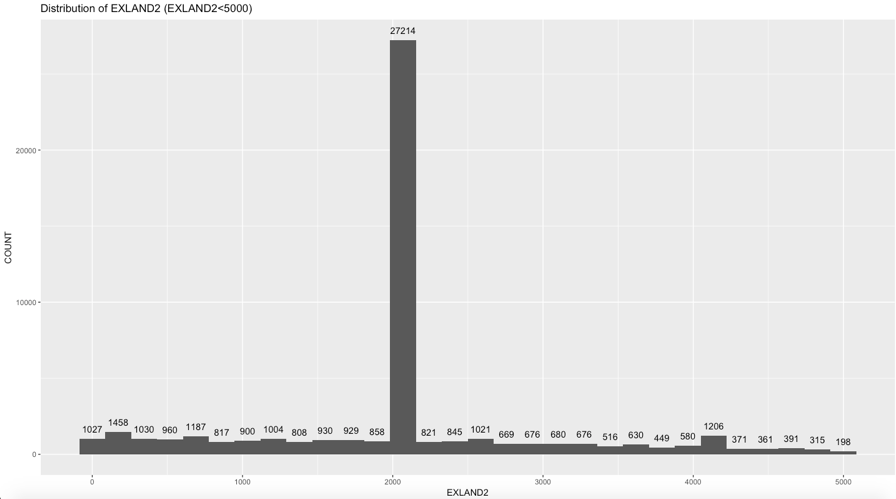




**EXLAND2 – Transitional Exempt Land Value**

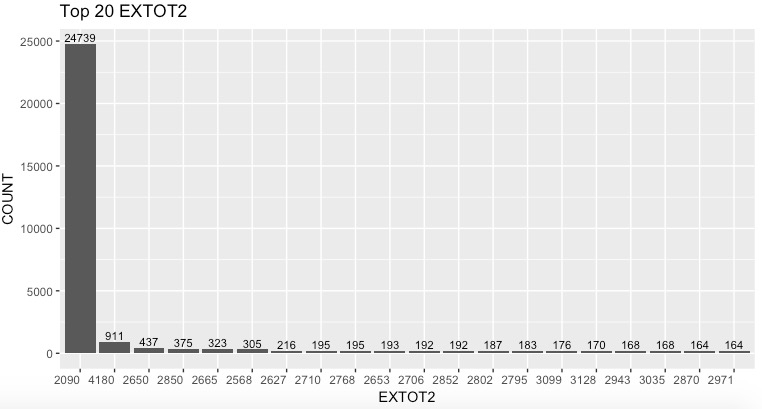


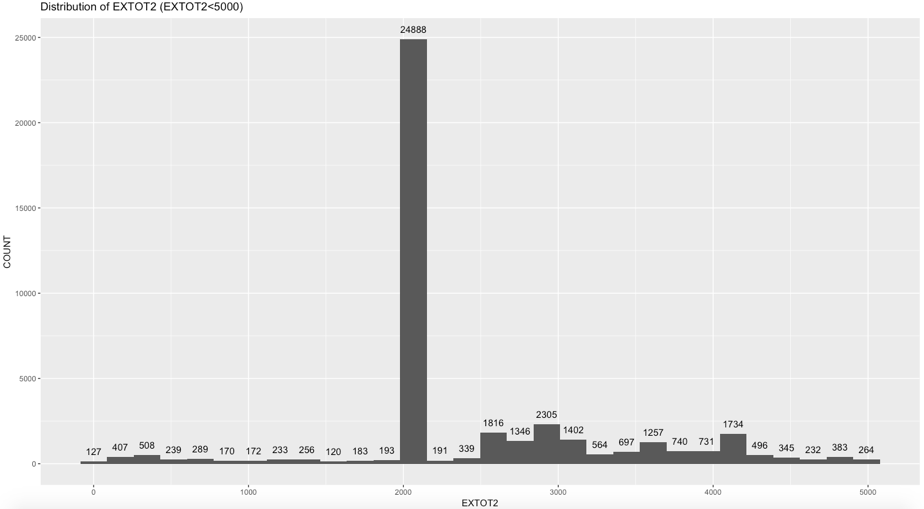




**EXTOT2 – Transitional Exempt Land Total**

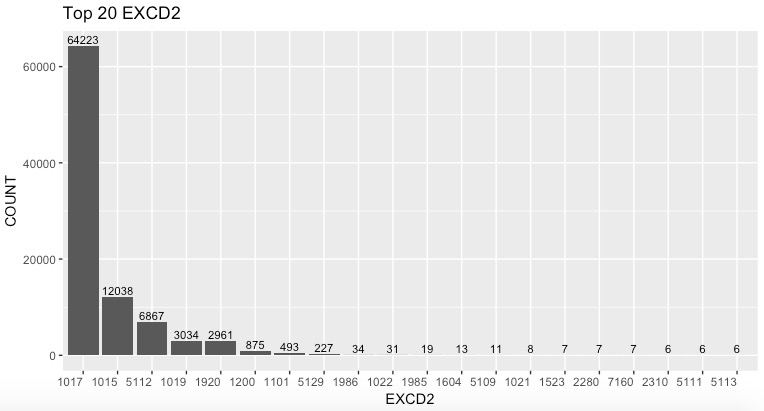


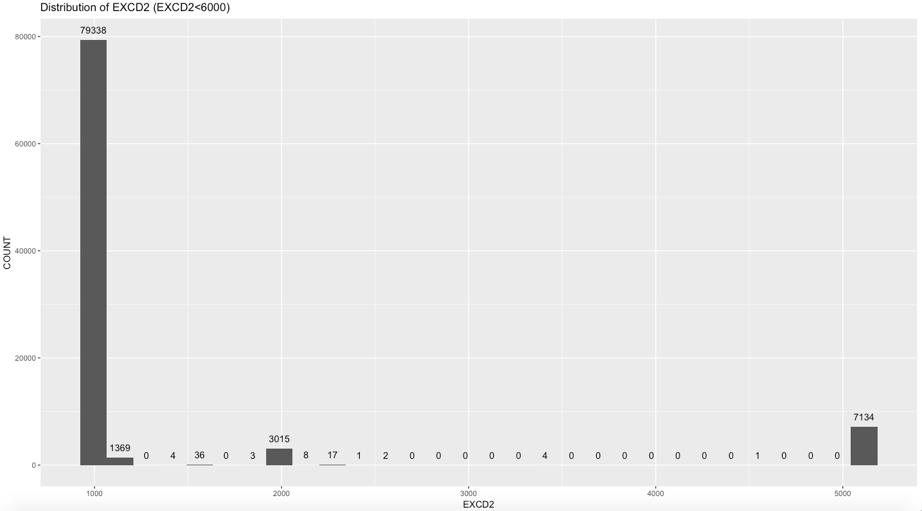




**EXCD2 – Exemption Code 2**



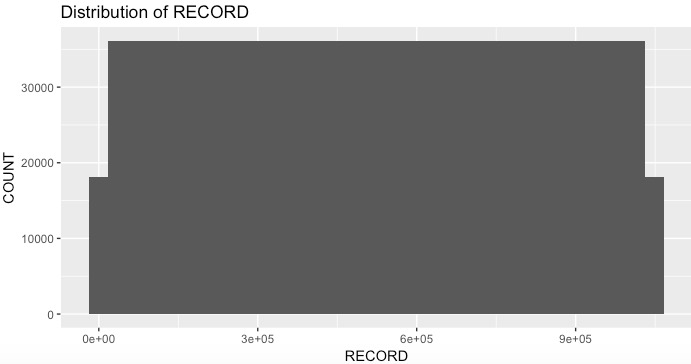




**4. Categorical Data Analysis**

**RECORD – Record ID**

There are 1048575 records in this dataset, so the record id varies from 1 to 1048575.



**BBLE - Concatenation of AV\_BORO, AV\_BLOCK, AV\_LOT, AV\_EASEMENT**

There are 1048575 different records in this dataset, which means every record have a unique BBLE.

**EASEMENT – Easement Description**

SPACE Indicates the lot has no Easement;

'A' Indicates the portion of the Lot that has an Air Easement;

'B' Indicates Non-Air Rights;

'E' Indicates the portion of the lot that has a Land Easement;

'F' THRU 'M' are duplicates of 'E';

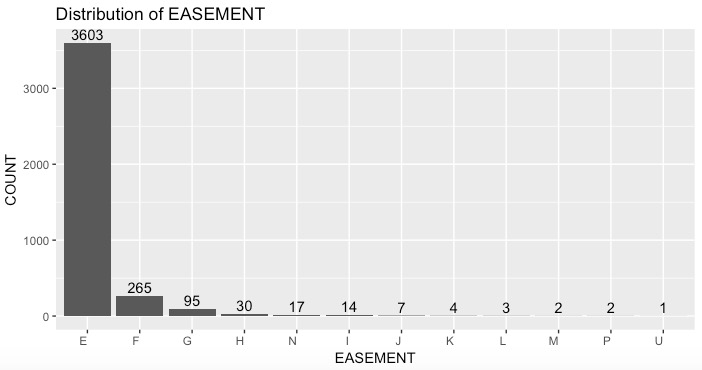
'N' Indicates Non-Transit Easement;

'P' Indicates Piers;

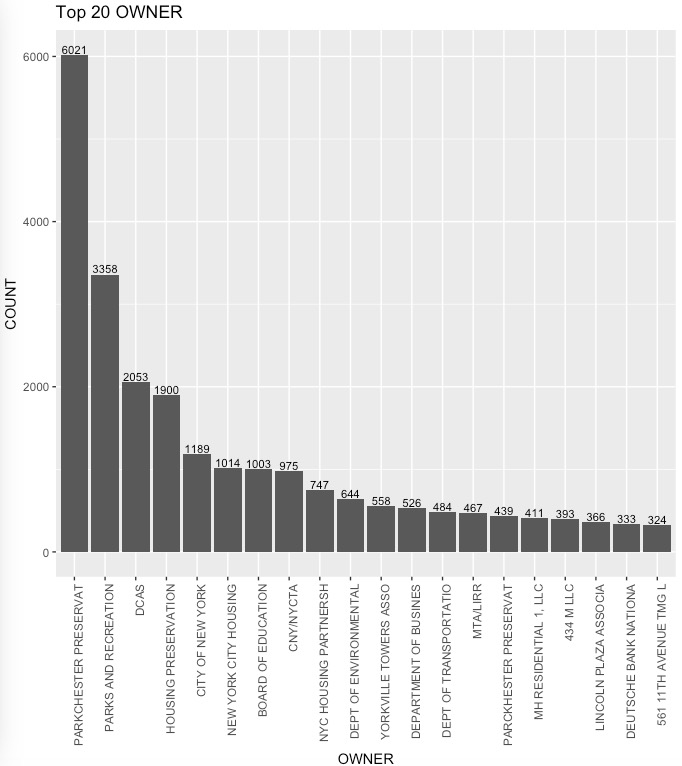
'R' Indicates Railroads;

'S' Indicates Street;

'U' Indicates U.S. Government;



**OWNER – Owner of property**



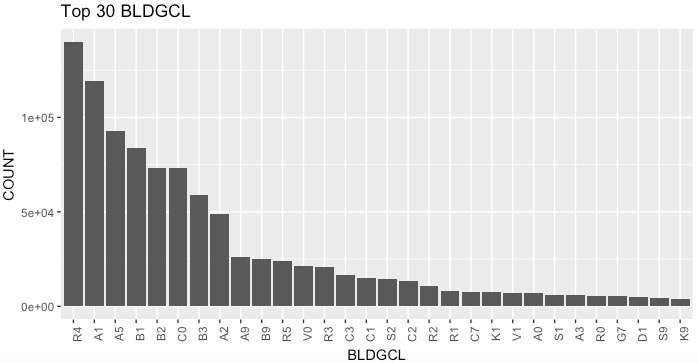
**Class A** represents the highest quality buildings in their market;

Buildings of **Class B** are generally a little older, but still have good quality management and tenants;

Buildings of **Class C** are older buildings (usually more than 20 years), and are located in less desirable areas and are in need of extensive renovation.

**BLDGCL – Building Class**





TAX CLASS 1 = 1-3 UNIT RESIDENCES;

TAX CLASS 1A = 1-3 STORY CONDOMINIUMS

ORIGINALLY A CONDO;

TAX CLASS 1B = RESIDENTIAL VACANT LAND;

TAX CLASS 1C = 1-3 UNIT CONDOMINUMS

ORIGINALLY TAX CLASS 1;

TAX CLASS 1D = SELECT BUNGALOW COLONIES;

TAX CLASS 2 = APARTMENTS;

TAX CLASS 2A = APARTMENTS WITH 4-6 UNITS;

TAX CLASS 2B = APARTMENTS WITH 7-10 UNITS;

TAX CLASS 2C = COOPS/CONDOS WITH 2-10 UNITS;

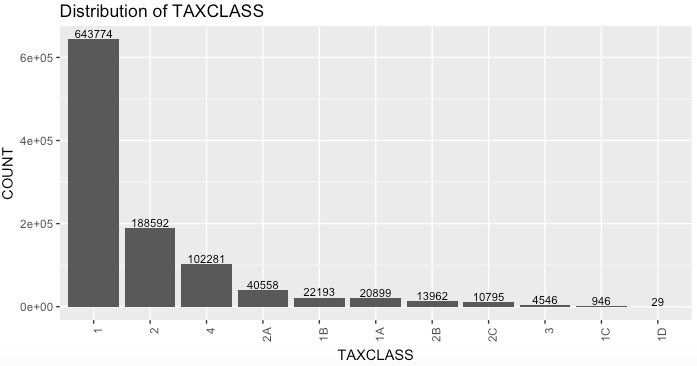
TAX CLASS 3 = UTILITIES (EXCEPT CEILING RR);

TAX CLASS 4A = UTILITIES - CEILING RAILROADS;

TAX CLASS 4 = ALL OTHERS

**TAXCLASS – Tax Class**

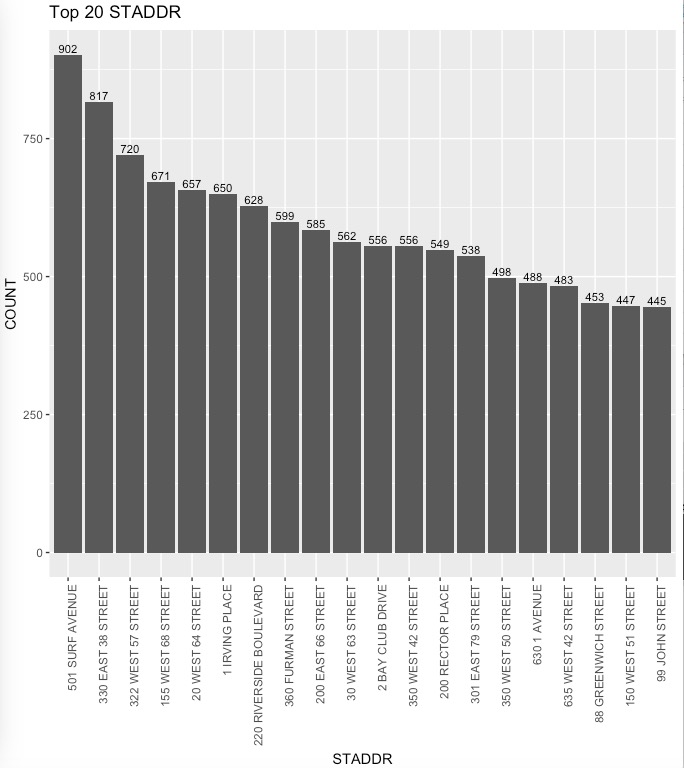




**STADDR – Street Address**

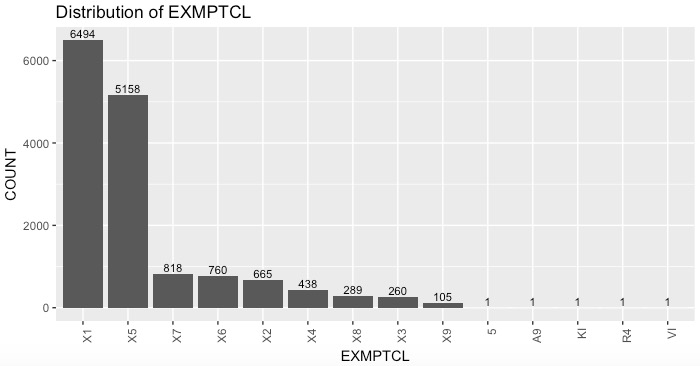
Too many unique value therefore not visualizing this categorical variable. A better way to visualize will be to put it on google map and create heat map visualization to see which areas are more poplar.





**EXMPTCL – Exempt Class**





**PERIOD – Assessment Period**



Single variable – Final.

**YEAR – Assessment Year**



Single Variable – 2010/11

**VALTYPE**



Single Variable – AC-TR