NY PROPERTY TAX FRAUD REPORT

Table of Contents

EXECUTIVE SUMMARY:

The city of New York suspects property tax fraud in the properties. To address this, an algorithm is developed to analyze the data and detect potential anomalies using the NY property dataset of 32 records and about a million rows.
The focus would be identifying instances where property values don't match their corresponding characteristics, suggesting underpayment of taxes through misrepresentation. By checking these unusual or suspicious patterns in dollar values and comparing the property values with their corresponding characteristics, detects the anomalies that can be used to alert the city to potential cases of tax fraud. In the process, we used an average of metrics generated from two algorithms and ranked them in descending order to understand which property tax has a higher chance of fraud because of their strangeness. This approach can save the city significant time and resources in manually reviewing records and help to identify and prevent tax fraud.

DATA DESCRIPTION:

The dataset used for this study is “New York Property Data,” which is property valuation and assessment data for housing and development. It is provided by the Department of Finance (DOF) with NYC OpenData as the dataset owner. The data set has 32 fields (14 numeric, 18 categorical) and 1070994 rows and is in the time period of 2010/11.

FIELD SUMMARY:

Numeric Fields:

Field Name	% Populated	Min	Max	Mean	StdDev	% Zero
LTFRONT	100	0	9999	36.6	74	169108
LTDEPTH	100	0	9999	88.8	76.3	170128
STORIES	94.7	1	119	5	8.4	0
FULLVAL	100	0	6150000000	874264.5	11582431	13007
AVLAND	100	0	2668500000	85068	4057260	13009
AVTOT	100	0	4668308947	227238	6877529.3	13007
EXLAND	100	0	2668500000	36423.9	3981575.7	491699
EXTOT	100	0	4668308947	91187	6508402.8	432572
BLDFRONT	100	0	7575	23	35.5	228815

BLDDEPTH	100	0	9393	39.9	42.7	228853
AVLAND2	26.4	3	2371005000	246235.7	6178962	0
AVTOT2	26.3	3	4501180002	713911.4	11652528.9	0
EXLAND2	8.1	1	2371005000	351235.6	10802212.6	0
EXTOT2	12.2	7	4501180002	656768.2	16072510.1	0

Categorical Fields:

Field Name	**% Populated**	**Unique Values**	**Most Common Value**
RECORD	100	1070994	N/A
BBLE	100	1070994	N/A
BORO	100	5	4
BLOCK	100	13984	3944
LOT	100	6366	1
EASEMENT	0.4	13	E
OWNER	97	863348	PARKCHESTER PRESERVAT
BLDGCL	100	200	R4
TAXCLASS	100	11	1
EXT	33	4	G
EXCD1	59.6	130	1017
STADDR	99.9	839281	501 SURF AVENUE
ZIP	97.2	197	10314
EXMPTCL	1.4	15	X1
EXCD2	8.6	61	1017
PERIOD	100	1	FINAL
VALTYPE	100	1	AC-TR
YEAR	100	1	2010/11

DATA CLEANING:

The data provided has a lot of missing values which need to be cleaned before proceeding anymore. A series of steps have been used to perform this task. Firstly, based on the idea of removing government-owned buildings or public entities that are generally tax-exempt and we are focused on private owners committing tax fraud. To accomplish this, from the 'OWNER' column, the rows which contain the words 'NYC,' 'NEW,' 'YORK,' 'DEPT," STATE OF NEW YORK," PUBLIC SERV,' 'BOARD,' 'GOVT OWNED,' 'CNY,' 'PRESERVATION,' 'PARKS,' 'PARK,' 'GOVERNMENT,' 'NATIONAL,' 'DEPARTMENT,' 'CITY,' 'N.Y.C.,' 'N.Y.,' 'N.Y.C,' 'N.Y,' 'YORK CITY,' 'NYS,'

'NYS DEPT,'' NEW YORK CITY' is removed. This deleted about 34,985 rows. Nextly, the missing values are imputed for which the logic is explained in the next steps.

IMPUTATION LOGIC:

For the algorithm to score all the records, they need to be filled, but the main idea is that the imputation must be in such a way that these imputed values should not cause any unusualness factor; hence we make them as neutral as possible. The fields that are used to make variables for the next steps are filled. The imputed variables and the logic for each are mentioned below.

1. ZIP: Filled in the missing ZIP values with the mode ZIP for the tax, which had 21,772 records as initially missing.
2. STORIES: This initially has 43,968 records missing, for this firstly, I created a new column named FAR; this is the Floor Area Ratio. This is the ratio of building volume to lot size. This was chosen as every county has a set of rules for the floor-to-area ratio. Using non-zero entries, mean FAR is calculated for every zip code. Then this is used to fill in the missing number of stories.

3. FULLVAL: Filled in the missing FULLVAL values by calculating the average FULLVAL for properties with the same building class (TAXCLASS) and in the same zip code (ZIP).

4. AVLAND: Filled in the missing AVLAND values by calculating the average AVLAND for properties with the same building class (TAXCLASS) and in the same zip code (ZIP).
5. AVTOT: Filled in the missing AVTOT values by calculating the average AVTOT for properties with the same building class (TAXCLASS) and in the same zip code (ZIP).
6. LTFRONT: Filled in the missing LTFRONT values by taking the median LTFRONT value for properties with the same building class (TAXCLASS) and on the same block (BLOCK).
7. LTDEPTH: Filled in the missing LTDEPTH values by taking the median LTDEPTH value for properties with the same building class (TAXCLASS) and on the same block (BLOCK).
8. BLDFRONT: Filled in the missing BLDFRONT values by taking the median BLDFRONT value for properties with the same building class (TAXCLASS) and on the same block (BLOCK).

9. BLDDEPTH: Filled in the missing BLDDEPTH values by taking the median BLDDEPTH value for properties with the same building class (TAXCLASS) and on the same block (BLOCK).

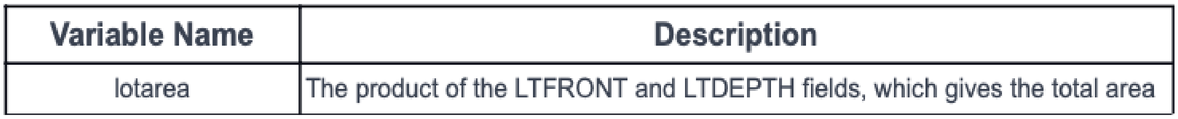
VARIABLES DESCRIPTION:

The variables are generated with an emphasis on the relationship between property value and property characteristics. In order to understand the property correctly, I considered the factors such as the property area of both land and building, property type estimated using the TAXCLASS field, and to get an estimate of the property value, the factors of location through ZIP CODE field and the pricing per square foot of land and building and compared that with the current value through research of the properties. The variables below are made to satisfy these conditions.

The following variable description should provide a brief explanation of all calculated variables, including the logic for why each variable measures some kind of unusualness that is of interest. The description should explain how each variable was calculated and what it measures.

$V1$ = FULLVAL	$S1$ = LTFRONT * LTDEPTH
$V2$ = AVLAND	$S2$ = BLDFRONT * BLDDEPTH
$V3$ = AVTOT	$S3$ = $S2$ * STORIES

r1 = $\frac{V1}{S1}$	r4 = $\frac{V2}{S1}$	r1 = $\frac{V3}{S1}$
r2 = $\frac{V1}{S2}$	r5 = $\frac{V2}{S2}$	r1 = $\frac{V3}{S2}$
r3 = $\frac{V1}{S3}$	r6 = $\frac{V2}{S3}$	r1 = $\frac{V3}{S3}$

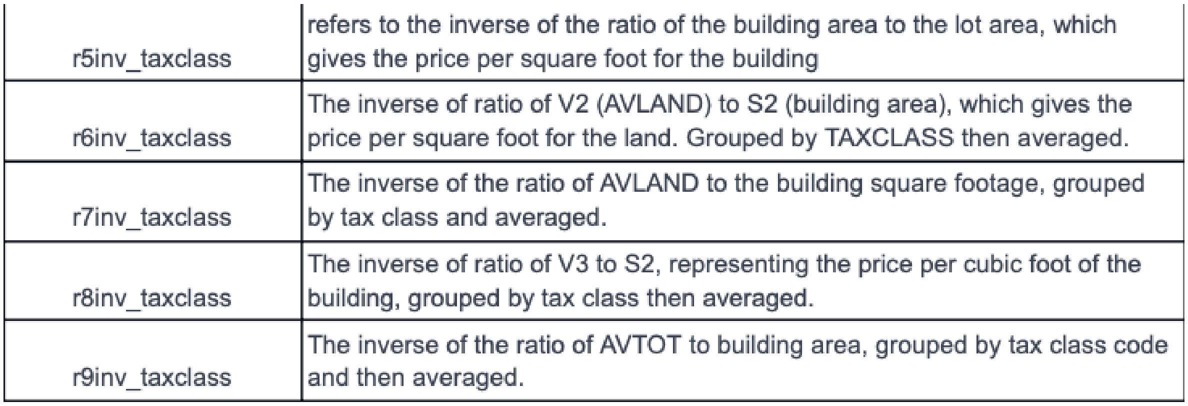


Variable Name	Description
lotarea	The product of the LTFRONT and LTDEPTH fields, which gives the total area

bldarea	The product of the BLDFRONT and BLDDEPTH fields, which gives the total area of the building in square feet.
bldvol	The product of bldarea and STORIES, which gives the total volume of the building.
r1	The ratio of FULLVAL to lotarea, which gives the price per square foot for the land.
r2	The ratio of FULLVAL to bldarea, which gives the price per square foot for the building.
r3	The ratio of FULLVAL to bldvol, which gives the price per cubic foot for the building.
r4	The ratio of AVLAND to lotarea, which gives the assessed value per square foot for the land.
r5	The ratio of AVTOT to lotarea, which gives the assessed value per square foot for the total property.
r6	The ratio of FULLVAL to AVLAND, which gives the relative price of the land compared to the total property value.
r7	The ratio of FULLVAL to AVTOT, which gives the relative price of the total property compared to the assessed value.
r8	the ratio of the total assessed value of the property (V3) to the area of the building on the property
r9	r9 is a variable that represents the ratio of the total property value v3 to the total area of the property s3
r1_inv	The inverse of r1, which flags unusually small values of the land price per square foot.
r2_inv	The inverse of r2, which flags unusually small values of the building price per square foot.
r3_inv	The inverse of r3, which flags unusually small values of the building price per cubic foot.
r4_inv	The inverse of r4, which flags unusually small values of the assessed value per square foot for the land.
r5_inv	The inverse of r5, which flags unusually small values of the assessed value per square foot for the total property.
r6_inv	The inverse of r6, which flags unusually large values of the relative price of the land compared to the total property value.
r7_inv	The inverse of r7, which flags unusually large values of the relative price of the total property compared to the assessed value.
r8_inv	r8_inv refers to the inverse of r8, this variable can be used to identify unusually small $ values as outliers.

r9_inv	r9_inv refers to the inverse of r9, this variable can be used to identify unusually small $ values as outliers.
VRnorm	The normalized version of r8, with the mean value set to 1.
Value_ratio	The maximum value of VRnorm and its inverse, which flags unusually large or small values of the ratio of the total property value to the sum of the assessed values for the land and total property.
FAR	Floor to Area Ratio (bldvol/lotarea)
r1_zip5	r1_zip5: Ratio of FULLVAL to lot area, which gives the price per square foot for the property. Grouped by zip5 class then averaged.
r2_zip5	r2_zip5: Ratio of total building area (BLDAREA) to lot area, which gives the building density per unit land area. Grouped by zip5 class then averaged
r3_zip5	r3_zip5: Ratio of building area of units built before 1940 (BLDAREA_LT40) to lot area, which gives the density of pre-1940 buildings per unit land area. Grouped by zip5 class then averaged.
r4_zip5	r4_zip5: Ratio of land area of property that is zoned for commercial use (LANDAREA_COMR) to lot area, which gives the density of commercial zoned land per unit land area. Grouped by zip5 class then averaged.
r5_zip5	r5_zip5: Ratio of land area of property that is zoned for residential use (LANDAREA_RES) to lot area, which gives the density of residential zoned land per unit land area. Grouped by zip5 class then averaged.
r6_zip5	r6_zip5: Ratio of building area of property that is zoned for commercial use (BLDAREA_COMR) to lot area, which gives the density of commercial buildings per unit land area. Grouped by zip5 class then averaged.
r7_zip5	r7_zip5: Ratio of building area of property that is zoned for residential use (BLDAREA_RES) to lot area, which gives the density of residential buildings per unit land area. Grouped by zip5 class then averaged.
r8_zip5	r8_zip5: Ratio of total property value (FULLVAL) to total building area (BLDAREA), which gives the price per unit building area. Grouped by zip5 class then averaged.
r9_zip5	r9_zip5: Ratio of total property value (FULLVAL) to product of total building area and number of stories (BLDAREA*STORIES), which gives the price per unit of building area and story. Grouped by zip5 class then averaged
r1inv_zip5	The inverse of ratio of FULLVAL to Lotarea, which gives the price per square foot for the land. Grouped by zip5 class then averaged.
r2inv_zip5	The inverse of ratio of r2, grouped by zip5 class then averaged.
r3inv_zip5	r3inv_zip5: Inverse ratio of FULLVAL to AVLAND.

r4inv_zip5	Inverse ratio of FULLVAL to AVTOT.
r5inv_zip5	Inverse ratio of AVLAND to lot area.
r6inv_zip5	Inverse ratio of AVTOT to lot area.
r7inv_zip5	Inverse ratio of AVLAND to AVTOT.
r8inv_zip5	Inverse ratio of FULLVAL to the building area.
r9inv_zip5	Inverse ratio of AVTOT to the building area.
r1_taxclass	The ratio of FULLVAL to lotarea, which gives the price per square foot for the land. Grouped by tax class then averaged.
r2_taxclass	The ratio of FULLVAL to bldarea, which gives the price per square foot for the building.Grouped by tax class then averaged.
r3_taxclass	The ratio of FULLVAL to bldvol, which gives the price per cubic foot for the building. Grouped by tax class then averaged.
r4_taxclass	The ratio of AVLAND to lotarea, which gives the assessed value per square foot for the land.Grouped by tax class then averaged.
r5_taxclass	The ratio of AVTOT to lotarea, which gives the assessed value per square foot for the total property.Grouped by tax class then averaged.
r6_taxclass	The ratio of FULLVAL to AVLAND, which gives the relative price of the land compared to the total property value.Grouped by tax class then averaged.
r7_taxclass	The ratio of FULLVAL to AVTOT, which gives the relative price of the total property compared to the assessed value.Grouped by tax class then averaged.
r8_taxclass	The ratio of the total assessed value of the property (V3) to the area of the building on the property.Grouped by tax class then averaged.
r9_taxclass	It is a variable that represents the ratio of the total property value v3 to the total area of the property s3.Grouped by tax class then averaged.
r1inv_taxclass	r1inv_zip5 is the inverse of the ratio of r1, grouped by tax class and averaged.
r2inv_taxclass	r2inv_zip5 is the inverse of the ratio of FULLVAL to lotarea, indicating price per square foot for land, grouped by tax class and averaged.
r3inv_taxclass	r1inv_zip5 is the inverse of the ratio of FULLVAL to lotarea, indicating price per square foot for land, grouped by tax class and averaged.
r4inv_taxclass	r1inv_zip5 is the inverse of the ratio of FULLVAL to lotarea, indicating price per square foot for land, grouped by tax class and averaged.



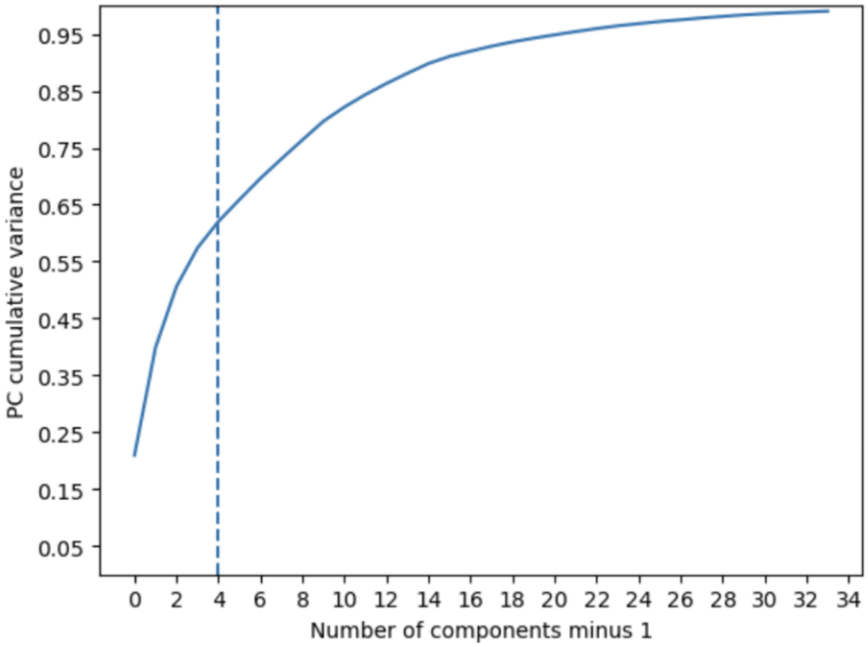
r5inv_taxclass	refers to the inverse of the ratio of the building area to the lot area, which gives the price per square foot for the building
r6inv_taxclass	The inverse of ratio of V2 (AVLAND) to S2 (building area), which gives the price per square foot for the land. Grouped by TAXCLASS then averaged.
r7inv_taxclass	The inverse of the ratio of AVLAND to the building square footage, grouped by tax class and averaged.
r8inv_taxclass	The inverse of ratio of V3 to S2, representing the price per cubic foot of the building, grouped by tax class then averaged.
r9inv_taxclass	The inverse of the ratio of AVTOT to building area, grouped by tax class code and then averaged.

DIMENSIONALITY REDUCTION:

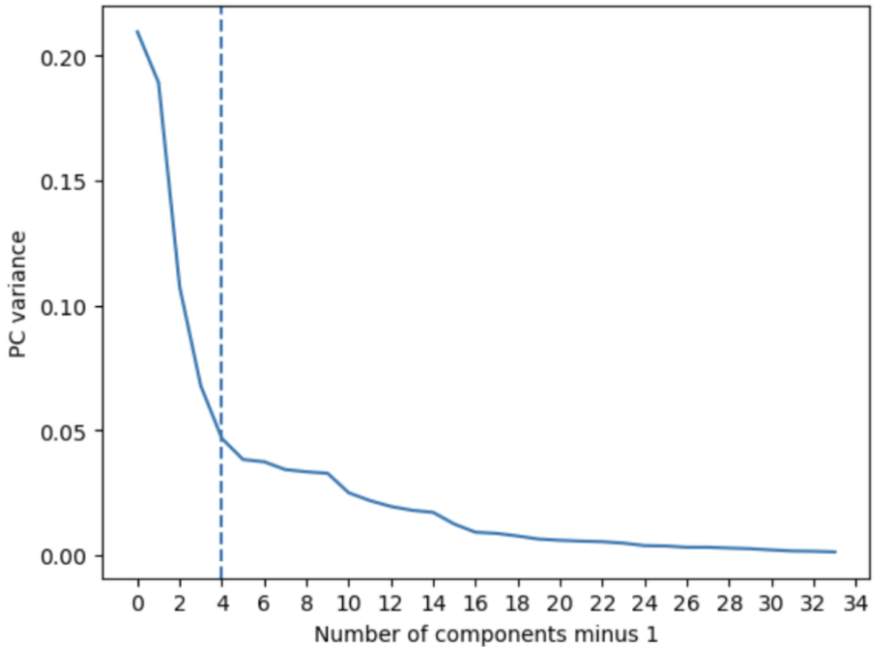
Generally, the process of dimensionality reduction is used in unsupervised models to reduce the risk of overfitting. When the number of features is large relative to the number of observations, unsupervised models can become very complex and may end up capturing noise or idiosyncrasies in the data that are not representative of the problem, which can lead to poor performance on new data. Dimensionality reduction is the technique of reducing the number of dimensions in the data set while preserving as much of the original information as possible. This can help simplify the data and make it more manageable to work with.

In this analysis, the principal component analysis is used to reduce the dimensionality of the data. This works by first centering the data based on predicting the mean of each feature from each observation and then calculating its covariance matrix. The eigenvalues and eigenvectors of the covariance matrix are calculated in which the eigenvectors are the principal components, and the eigenvalues help us understand the amount of variance in the data caused by each component. The principal components are sorted by magnitude. After calculating the principal components, the original data records are rewritten in terms of their principal components. a scree plot is generated to decide how many PCs to be kept.

Firstly, the variables are zscaled and then the PCA is performed to understand the desired amount of components through cumulative variance and scree plots shown below.
The below plot shows the cumulative variance of PCs. it shows that about 80% of all the variance can be kept by considering the first 5 PCs.

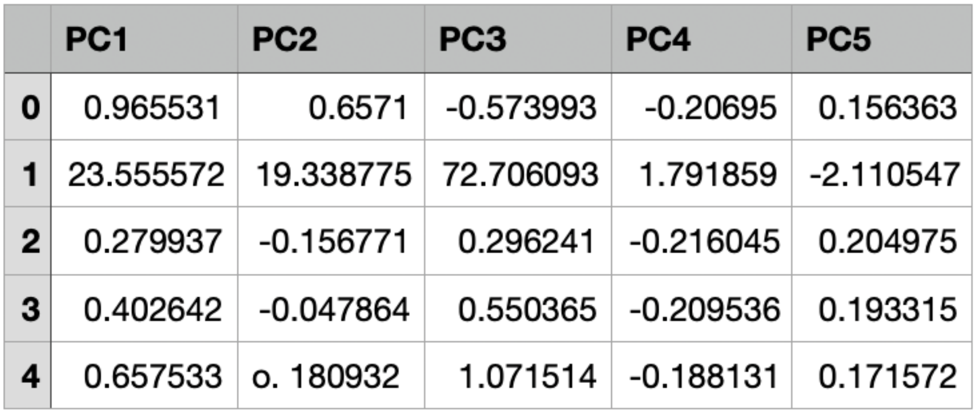


Below is the scree plot showing the decreasing variance of PCs. From the plot, it can be seen that at least 4 PCs must be kept, after which, the variance is low.



Now, the PCA is done again, keeping only the top components. The cumulative sum of explained variance ratios for the PCA model is calculated to understand the running total of explained variance from the first component to the current component. Here, the values are 20%, 39%, 50%, 57%, and 62% approximately for the respective top 5 PCs. This data now with 1,036,009 record and 5 PCs is zscaled again to make all the PCs equally important.

The below list shows the head of data after the above processes.



	PC1	PC2	PC3	PC4	PC5
0	0.965531	0.6571	-0.573993	-0.20695	0.156363
1	23.555572	19.338775	72.706093	1.791859	-2.110547
2	0.279937	-0.156771	0.296241	-0.216045	0.204975
3	0.402642	-0.047864	0.550365	-0.209536	0.193315
4	0.657533	o. 180932	1.071514	-0.188131	0.171572

ANOMALY DETECTION ALGORITHMS:

Two anomaly detection algorithms were used for this analysis.

1. Minkowski distance from origin to check for outliers
2. Autoencoder

After the above processes of data preparation and scaling, the data can now be imagined as a cloud around an origin with the same scaling for all dimensions. In this case, the outlier can be understood by calculating the distance from the origin. To calculate this the general distance measure of Minkowski distance is used. The formula for it is given as

$$s_i = \left(\sum_n |z_n^i|^p \right)^{1/p}$$

Where si gives the Minkowski distance, z = data in absolute, p = p1

For the powers of p (p1 for Minkowski), and p2(for autoencoder) for both the algorithms, a range between 1 and 4 is considered to provide a good balance between sensitivity to outliers and robustness to noise. For p=1 gives the Manhattan distance, which is useful when we wish to ignore the magnitude of difference and focus on direction, p = 2 gives the Euclidean distance, which is useful when the scale of features is important and to penalizes large differences more heavily than small ones and p = infinity give the Chebyshev distance. Having p ranges from 1 to 4 strikes a balance between these extremes.

Using p1 of 1.5 for the minkowski distance gave a max score of 653.71

An autoencoder generally works by first encoding data into a compressed form and then decoding it back to its original form. The difference between this original data and decoded data is considered to be the reproduction error which is used to detect potential fraud.

To calculate this, a neural net model is created and trained on the zscaled and PCA dataset. After that, the model is trained on the input data and output data again to find patterns in input data since this is unsupervised learning.

The predicted output values from the above model object are subtracted from the original PCA and zscaled data. The score is then calculated by taking the absolute value of the error array, raising it to the power p2, summing the values along rows, raising again to the power p2 and considering the maximum value.

The formula for reduction error is
Error = pca_out – data_pca_zs
Score2 = (score2 = ((error.abs()**p2).sum(axis=1))**oop2)

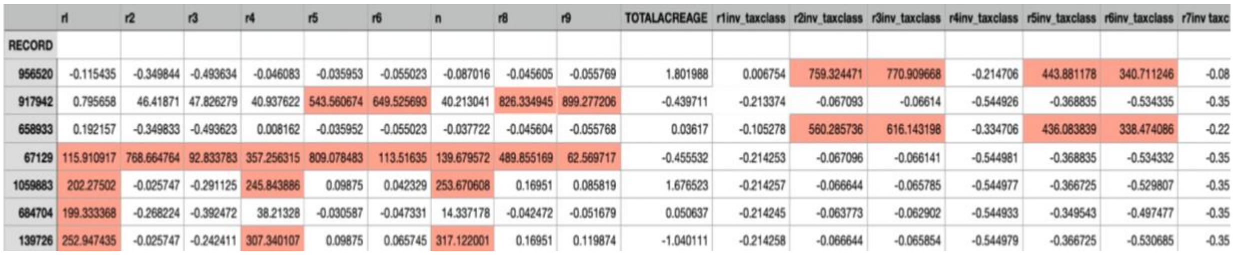
Using a p2 =4 for autoencoder gave a max score of 573.78

The scores generated are then zscaled to make sure that they are on equal scales. These scores are then ordered by rank for the record. The final score to consider is then calculated by averaging both these scores and sorting the records from high to low.

RESULTS:

These sorted records are the records with high to low strangeness based on the scores. For these top records, the zscaled values are considered for the variables, which explain how many standard deviations away is the zscaled value from the mean. This shows which variables have unusual values. In order to aid with the search of understanding the variables, a heatmap can be generated which shows which variables are too large or too small individually and also for the groupings of TAXCLASS and ZIPCODE.

The below table shows some of the extreme zscores for variables highlighted in red.

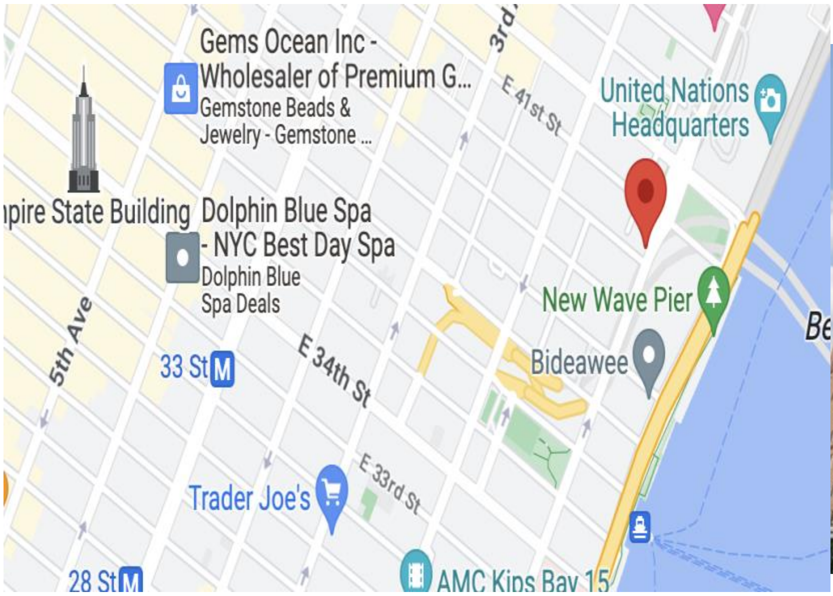


	rl	r2	r3	r4	r5	r6	n	r8	r9	TOTALACREAGE	r1inv_taxclass	r2inv_taxclass	r3inv_taxclass	r4inv_taxclass	r5inv_taxclass	r6inv_taxclass	r7inv taxc
RECORD																	
956520	-0.115435	-0.349844	-0.493634	-0.046083	-0.035953	-0.055023	-0.087016	-0.045605	-0.055769	1.801988	0.006754	759.324471	770.909668	-0.214706	443.881178	340.711246	-0.08
917942	0.795658	46.41871	47.826279	40.937622	543.560674	649.525693	40.213041	826.334945	899.277206	-0.439711	-0.213374	-0.067093	-0.06614	-0.544926	-0.368835	-0.534335	-0.35
658933	0.192157	-0.349833	-0.493623	0.008162	-0.035952	-0.055023	-0.037722	-0.045604	-0.055768	0.03617	-0.105278	560.285736	616.143198	-0.334706	436.083839	338.474086	-0.22
67129	115.910917	768.664764	92.833783	357.256315	809.078483	113.51635	139.679572	489.855169	62.569717	-0.455532	-0.214253	-0.067096	-0.066141	-0.544981	-0.368835	-0.534332	-0.35
1059883	202.27502	-0.025747	-0.291125	245.843886	0.09875	0.042329	253.670608	0.16951	0.085819	1.676523	-0.214257	-0.066644	-0.065785	-0.544977	-0.366725	-0.529807	-0.35
684704	199.333368	-0.268224	-0.392472	38.21328	-0.030587	-0.047331	14.337178	-0.042472	-0.051679	0.050637	-0.214245	-0.063773	-0.062902	-0.544933	-0.349543	-0.497477	-0.35
139726	252.947435	-0.025747	-0.242411	307.340107	0.09875	0.065745	317.122001	0.16951	0.119874	-1.040111	-0.214258	-0.066644	-0.065854	-0.544979	-0.366725	-0.530685	-0.35

UNUSUAL CASES:

1.

RECORD	95995
BBLE	1013530001
STREET ADDRESS	724 1 AVENUE
ZIP CODE	10017
STORIES	1
TAXCLASS	4
OWNER	BERGAMINI, JENNIFER
BLDDEPTH	20
BLDFRONT	15
FULL VAL	17354800
LTFRONT	197
LTDEPTH	378
AVLAND	7785000
AVTOT	7809660

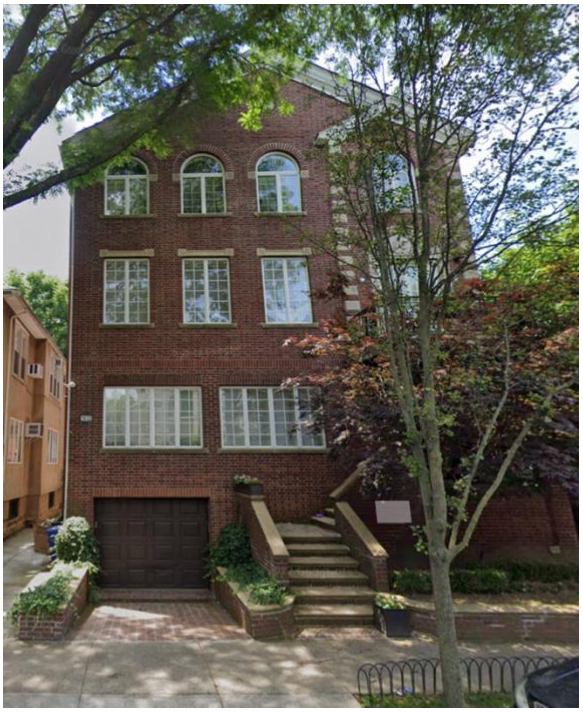

Gems Ocean Inc - Wholesaler of Premium G...
Gemstone Beads & Jewelry - Gemstone ...
3rd
E 41st St
United Nations Headquarters
pire State Building
Dolphin Blue Spa - NYC Best Day Spa
Dolphin Blue Spa Deals
New Wave Pier
Be
5th Ave
33 St M
E 34th St
Bideawee
Trader Joe's
E 33rd St
28 St M
AMC Kips Bay 15

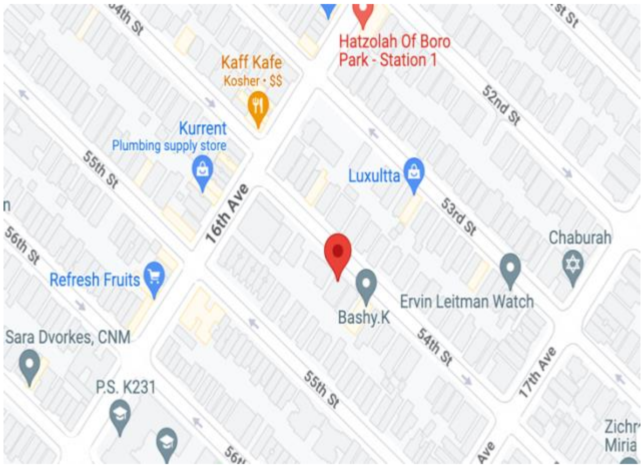
- The property record has very high values above standard deviation for variables r3, r6, r9, and moderately high values for variables r2, r5, r8 which are dependent on BLDFRONT, BLDDEPTH.
- This property shows the stories as 1, which is inaccurate.
- The variables r3,r6,r9 are dependent on stories (as r3 = v1/s3 and s3 = BLDFRONT*BLDDEPTH*STORIES), which could be the reason why they are so high.
- The variables r2, r5, r8 are dependent on BLDFRONT and BLDDEPTH which could be the reason they are moderately high.
- Wrong number of stories resulting in high variable values
- Needs further investigation.

2.

RECORD	435070
BBLE	3054850020
STREET ADDRESS	1628A 54 STREET
ZIP CODE	11204.0
STORIES	
TAXCLASS	1B
OWNER	KLEIN HARRY
BLDDEPTH	0
BLDFRONT	0
FULL VAL	5730
LTFRONT	0

LTDEPTH	2
AVLAND	10
AVTOT	10



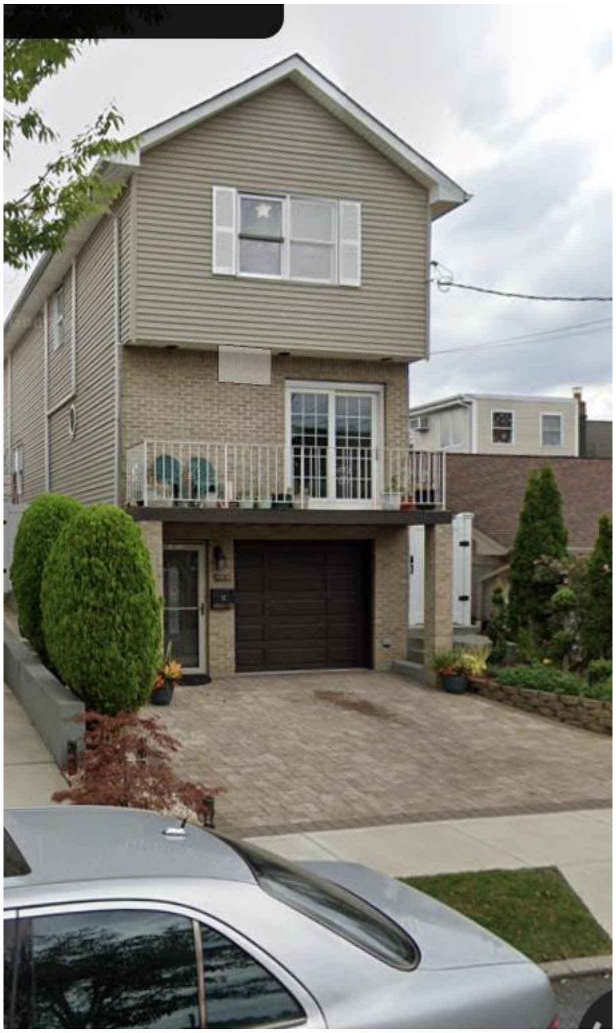


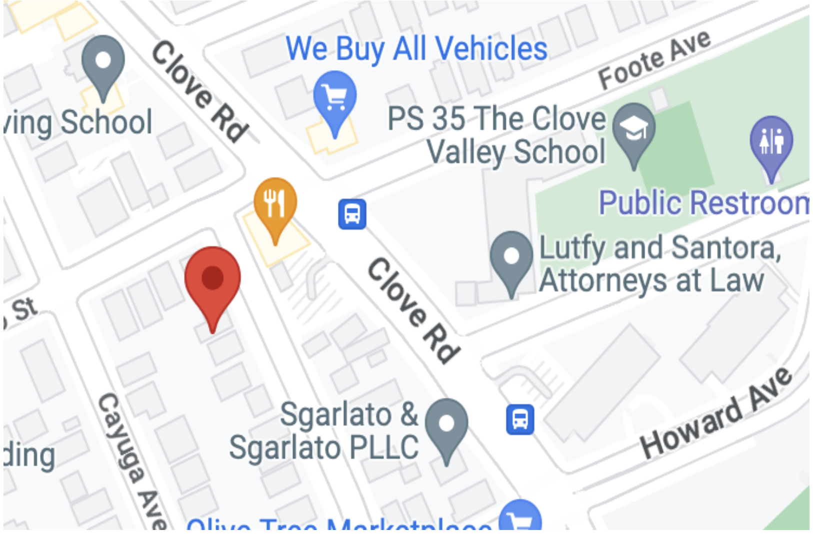
- This record has high values for r8inv_zip5 and r9inv_zip5
- This shows that the property has unusually low values for the corresponding zip code
- The low value of AVTOT could be the reason.
- Needs further investigation

3.

RECORD	**956520**
BBLE	5006590012
STREET ADDRESS	12 ONEIDA AVENUE
ZIP CODE	10301
STORIES	3
TAXCLASS	1
OWNER	TROMPETA RIZALINA
BLDDEPTH	5020
BLDFRONT	1812
FULL VAL	348200

LTFRONT	25
LTDEPTH	91
AVLAND	15600
AVTOT	20892



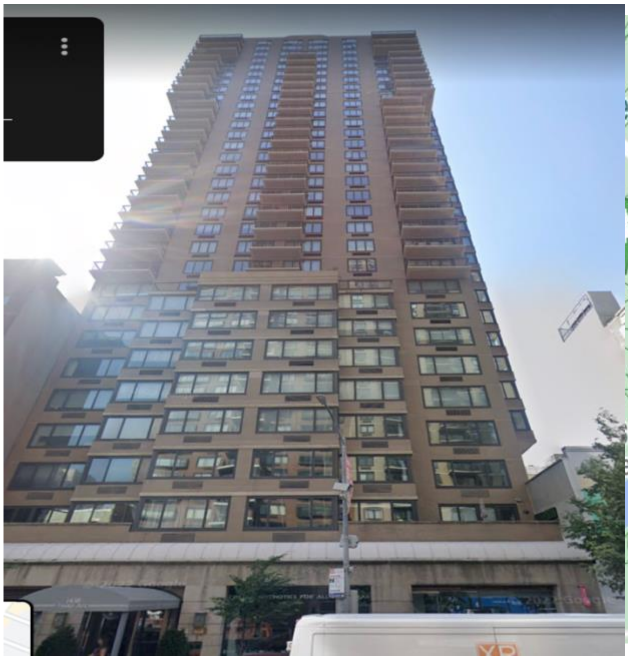


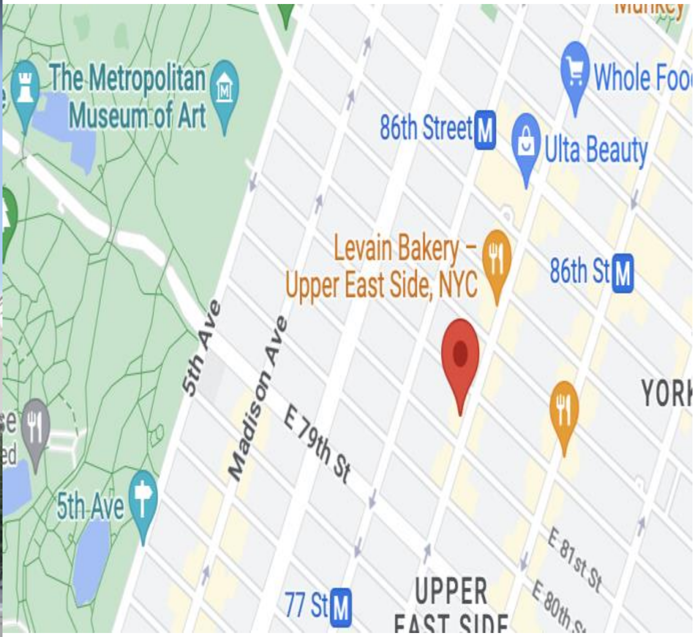
- The building size S2 is very high compared to its V1 – FULLVAL
- This could be why the values of r2inv_zip5, r3inv_zip5, r5inv_zip5, r6inv_zip5, r8inv_zip5, r9inv_zip5 are very high.
- These values show that the property value is too low for that zipcode
- Needs further investigation

4.

RECORD	**111420**
BBLE	1015101092
STREET ADDRESS	1438 3 AVENUE

ZIP CODE	10028
STORIES	31
TAXCLASS	2
OWNER	BOXWOOD FLTD PARNTERS
BLDDEPTH	9393
BLDFRONT	7575
FULL VAL	296508
LTFRONT	75
LTDEPTH	93
AVLAND	22896
AVTOT	133429



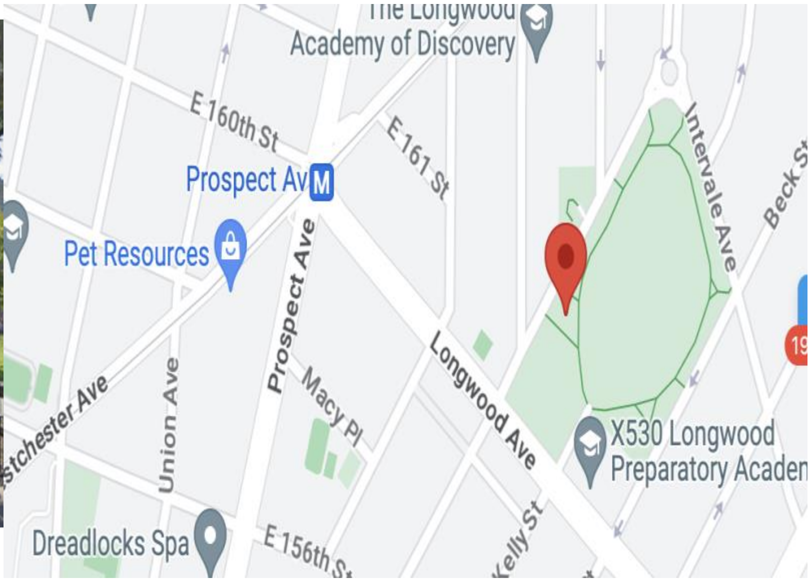


- Values of FULLVAL, AVLAND and AVTOT are really low
- Hence the values of variables r3_inv, r6_inv, r9_inv are unreasonably low.
- These low values show that the property values are too low for that region.
- Needs further investigation

5.

RECORD	**155893**
BBLE	2027020001

STREET ADDRESS	810 DAWSON STREET
ZIP CODE	10459
STORIES	1
TAXCLASS	4
OWNER	ATTRACTIVE HOME, INC
BLDDEPTH	31
BLDFRONT	73
FULL VAL	3080000
LTFRONT	4
LTDEPTH	31
AVLAND	1075500
AVTOT	1386000



- Has high values for r1_zip5, r4_zip5, r7_zip5, r1_TAXCLASS, r4_ TAXCLASS, r7_ TAXCLASS
- This could be why the property has unusually high values for its neighborhood.
- Could be caused by AVLAND, AVTOT, LTFRONT and LTDEPTH being low.
- Needs further investigation

SUMMARY:

The process of data analysis on the NY Property Tax data is very intriguing. I started with a data quality report to understand the high level view of the data issues in the dataset, followed by cleaning the data by excluding values that I thought were not needed and then filling in the missing values, followed by understanding and generating new variables that affect the relationship between the property price and the property characteristics. Then performed, dimensionality reduction using PCA to retain the important patterns in data while reducing dimensions. I then used the average of a maximum of two scores generated from Minkowski distance and Autoencoder reduction error. The sorted records from high to low helped me understand which records were anomalies. I then deep-dived into investigating some anomaly records for a deeper understanding, of which 5 are mentioned above.

However, the algorithm may not capture what the client is looking for. This would need an analyst to adjust the algorithm. This can be done by hyperparameter tuning, creating more new features from the dataset, using methods such as bagging or boosting, changing the ideology of cleaning the data or regularization.

One such idea could be to check if making a new dataset from abnormal values of AVTOT and AVLAND values and perform the data analysis on it to see and compare if the records are even stranger or the same.

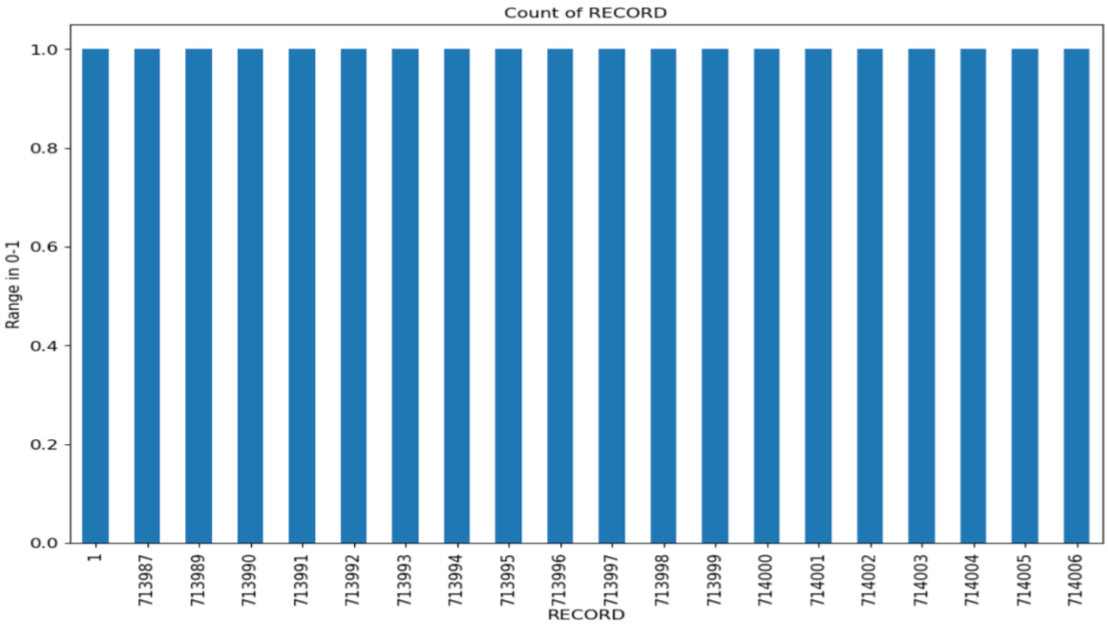
The overall process can be further improved by incorporating new methods.

APPENDIX:

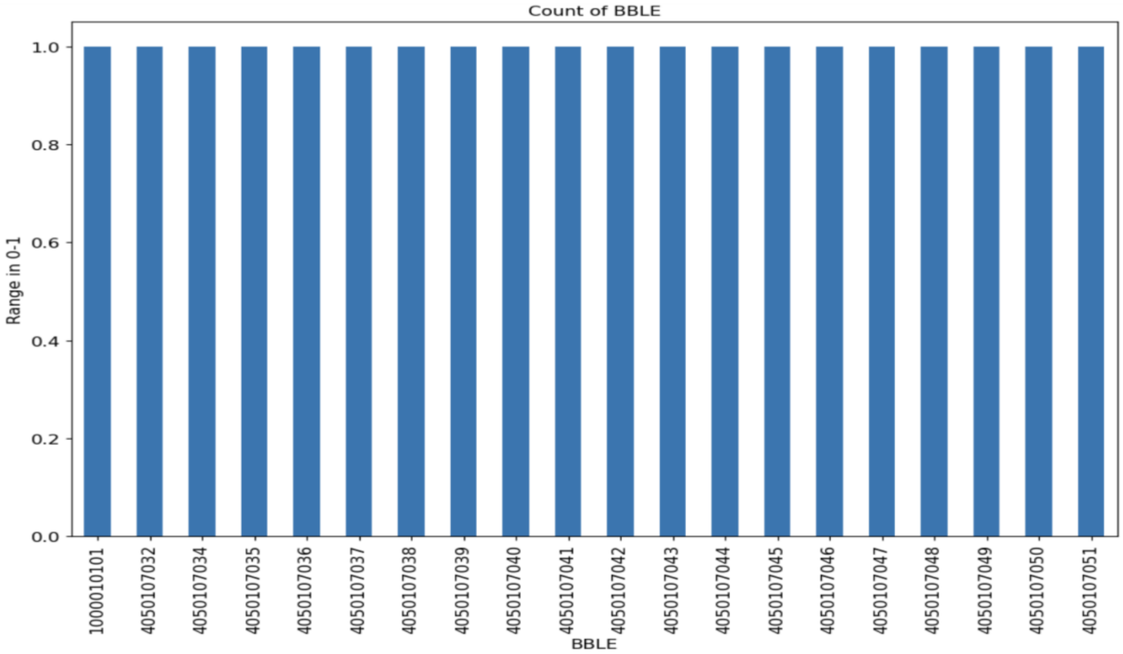
DATA QUALITY REPORT:

Individual field summary:

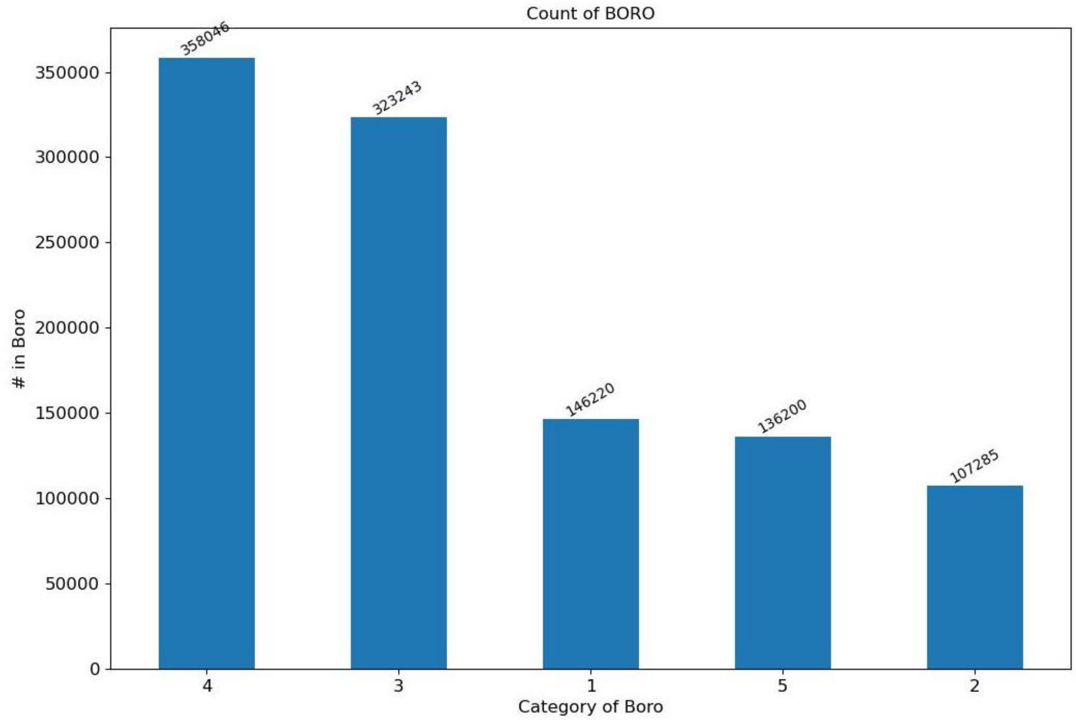
1. RECORD: This is the identification number of the record. Every identification number is unoque.



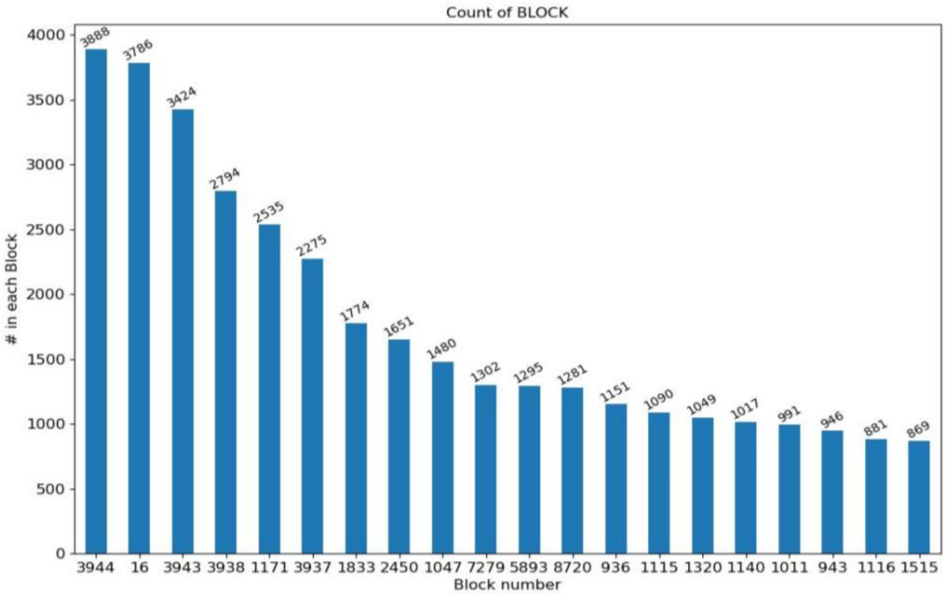
2. BBLE: It is an identification number that signifies Boro, Block, Lot and Easement code.



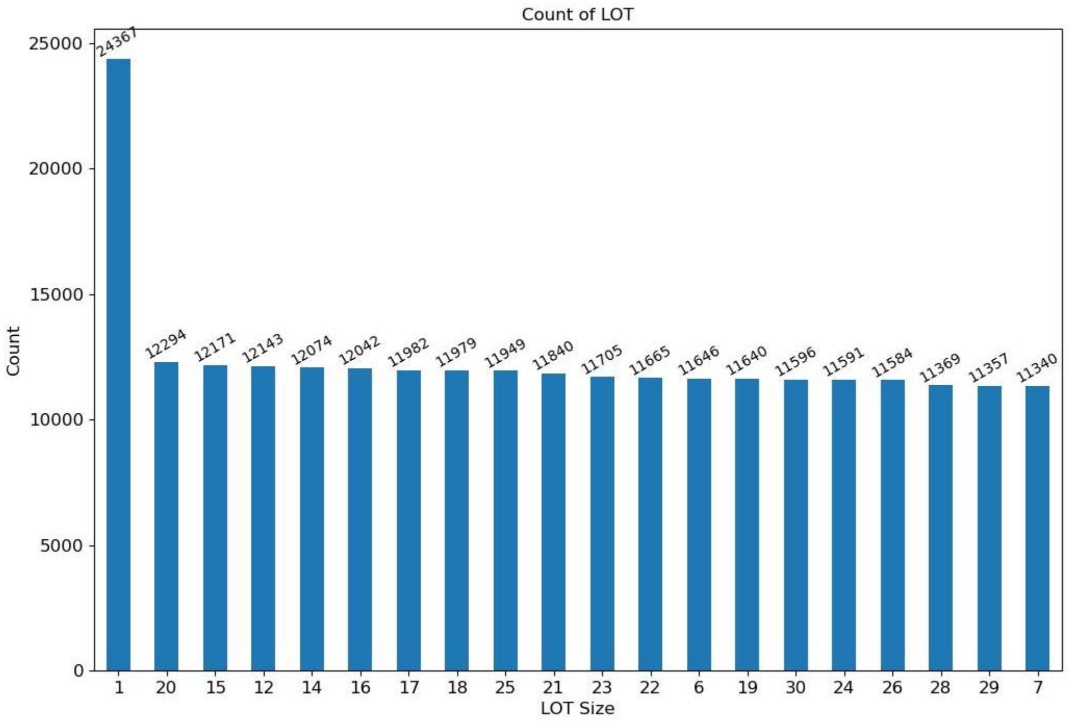
3. BORO: Stands for Borough. It has 5 values. 1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island.



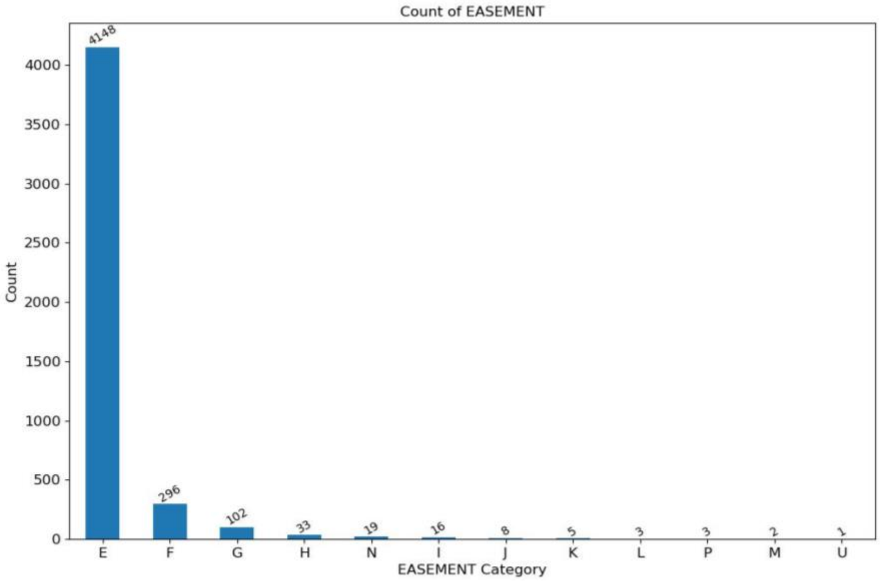
4. BLOCK: These are block numbers taken from the block ranges based on BORO. Manhattan: 1 to 2255, Bronx: 2260 to 5958, Brooklyn: 1 to 8955, Queens: 1 to 16350, Staten Island: 1 to 8050.



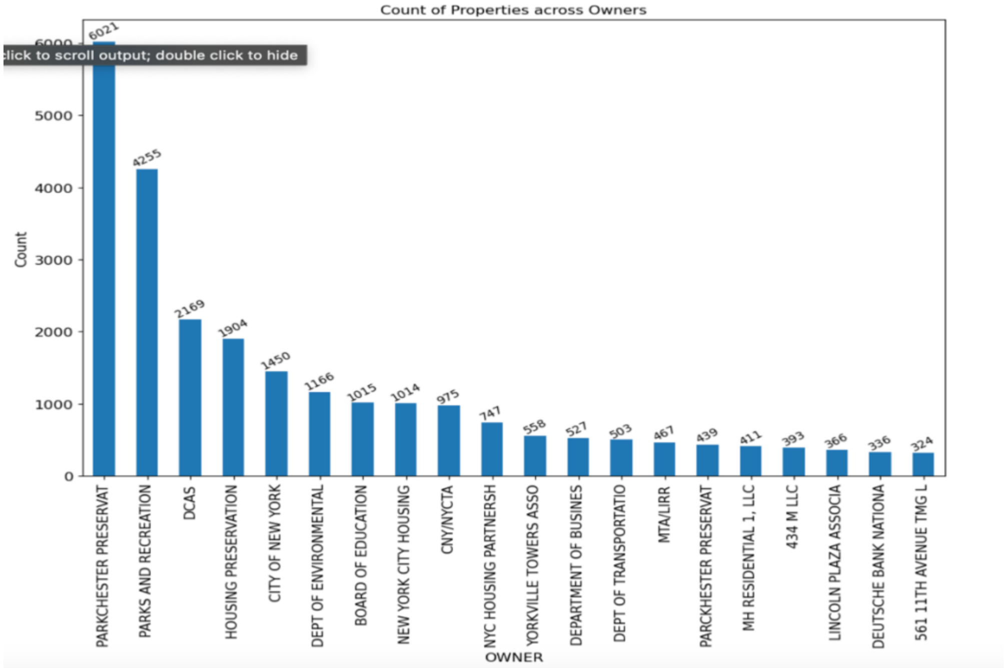
5. LOT: Indicates the LOT size with respective to each property.



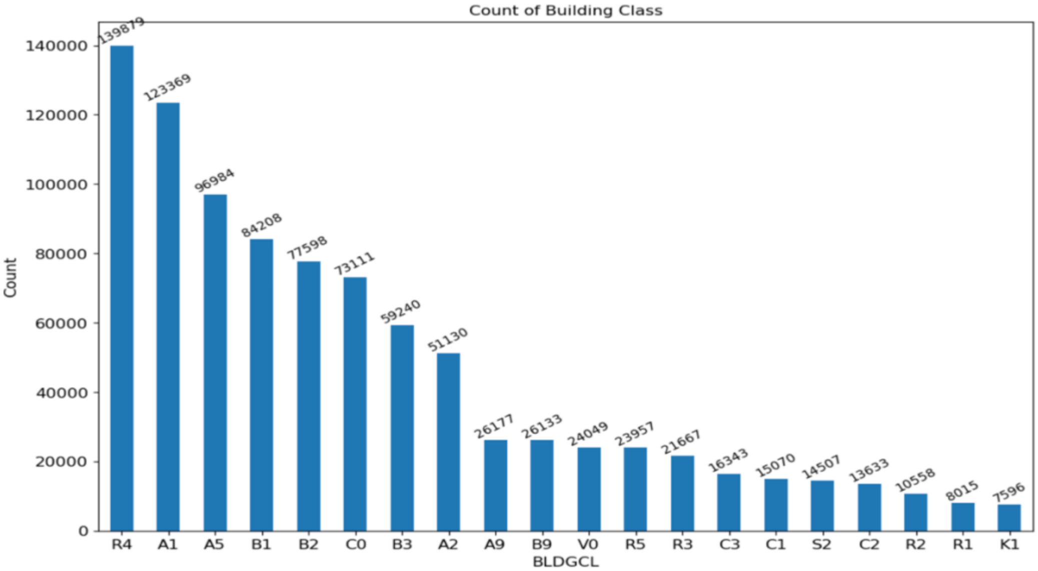
6. EASEMENT: Describes which type of Easement does the property have. Space = No Easement A = Air Easement
B = Non-Air Rights E = Land Easement
F Thru M are duplicates of E N = Non-Transit Easement
P = Pier R = Railroad
S = Street U = U.S. Government



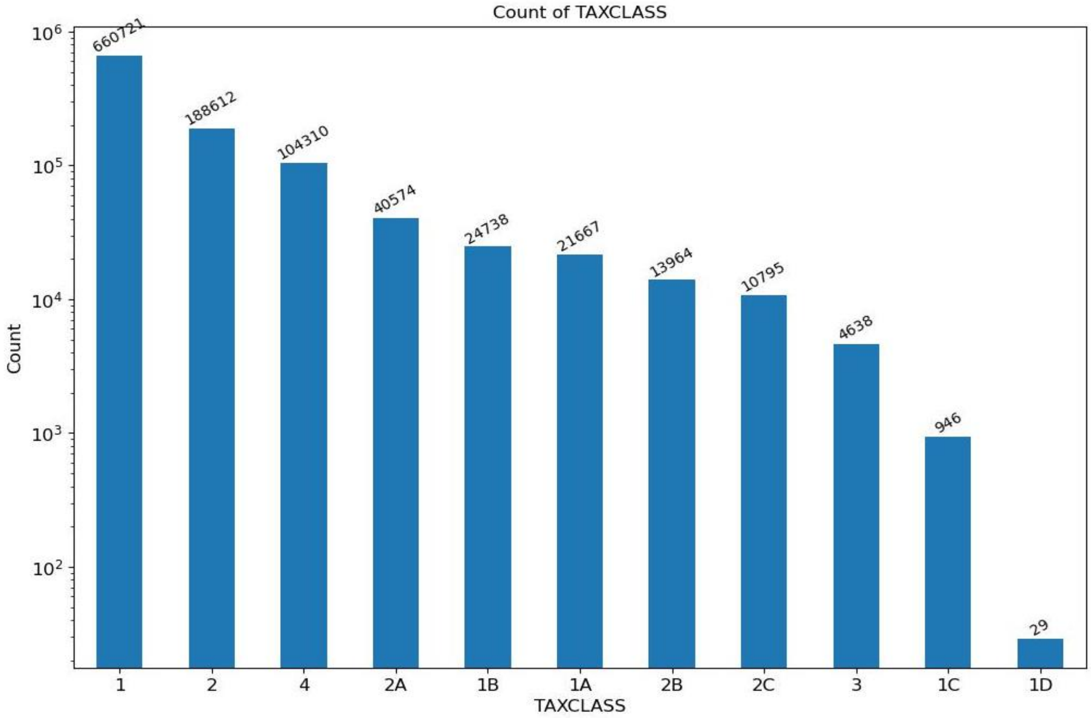
7. OWNER: It indicates the name of the owner.



8. BLDGCL: it denotes the class of the Building.

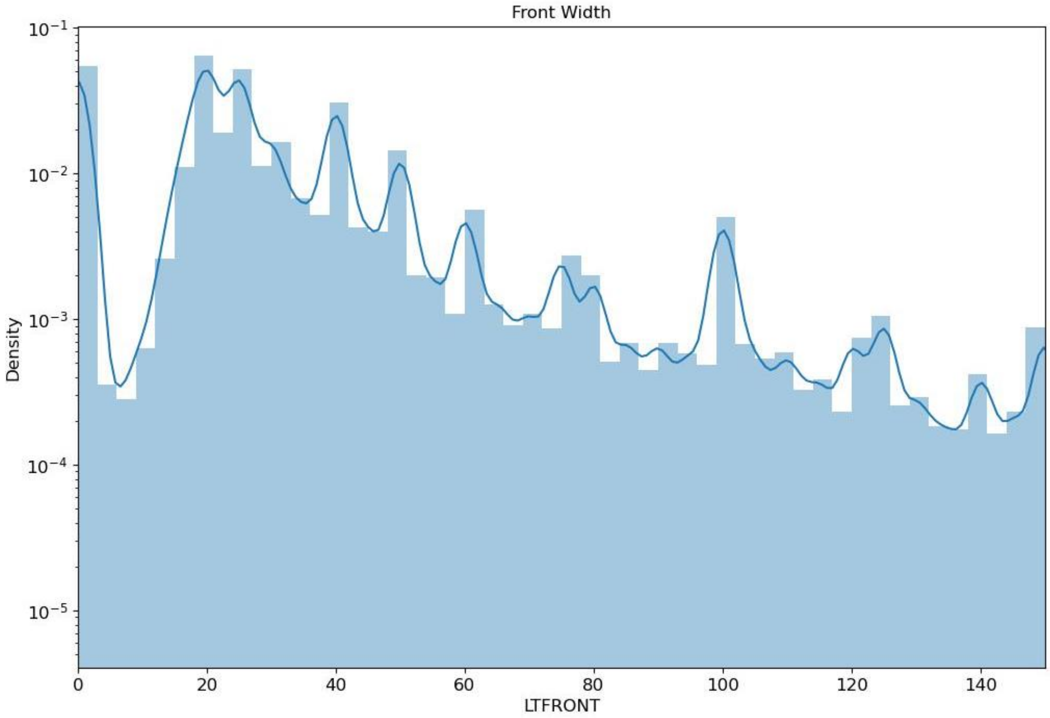


9. TAX CLASS:
1 = 1 - 3 Unit Residence, 2 = Apartments, 2A = 4, 5, or 6 Units, 3 = Utilities, 4 = All Others.



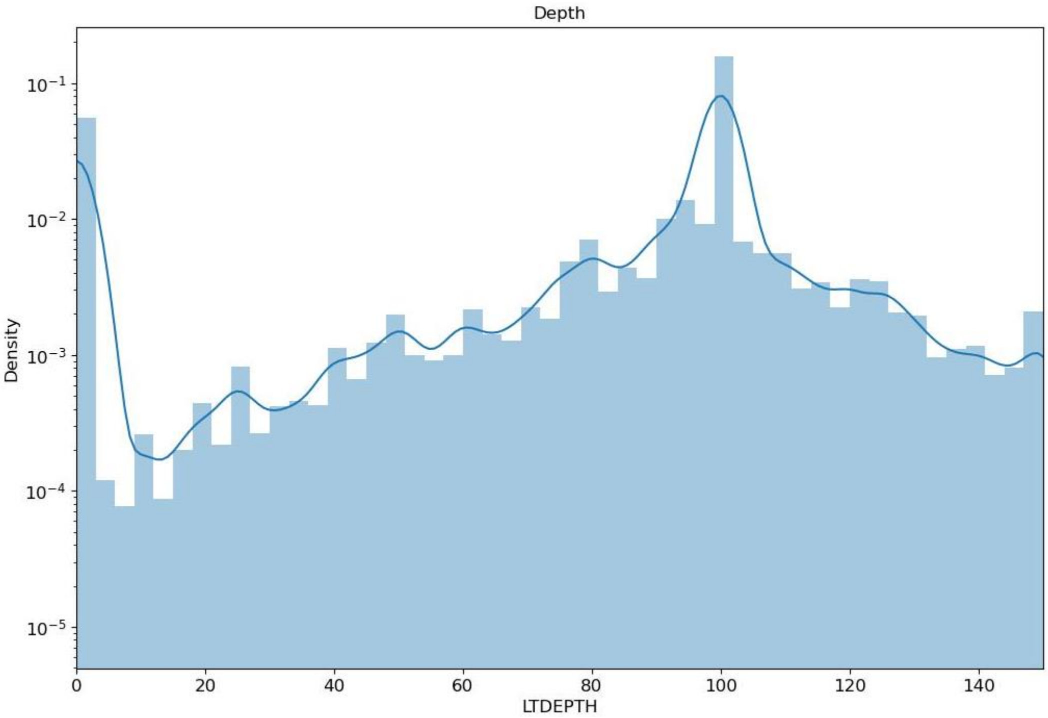
10. LTFRONT: It is the Lot Width.

The below figure uses a logarithmic scale, and the Lot width is limited to 102.

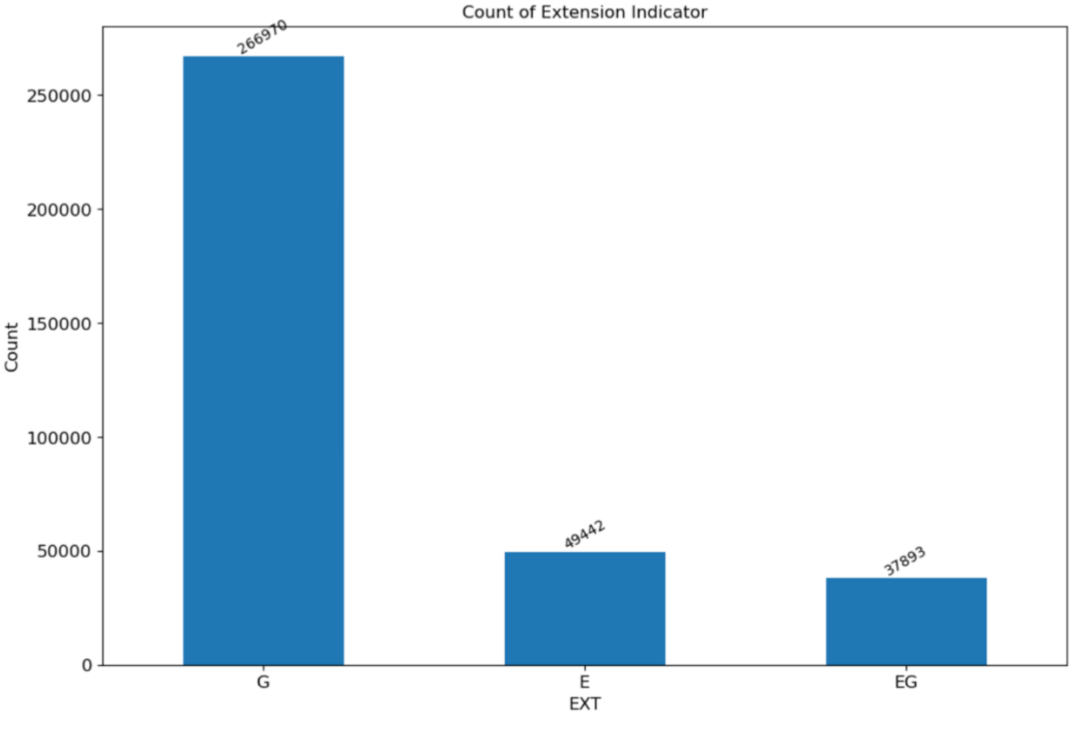


11. LTDEPTH: It is the Lot Depth.

The below figure uses a logarithmic scale and the Lot Depth is limited to 710.

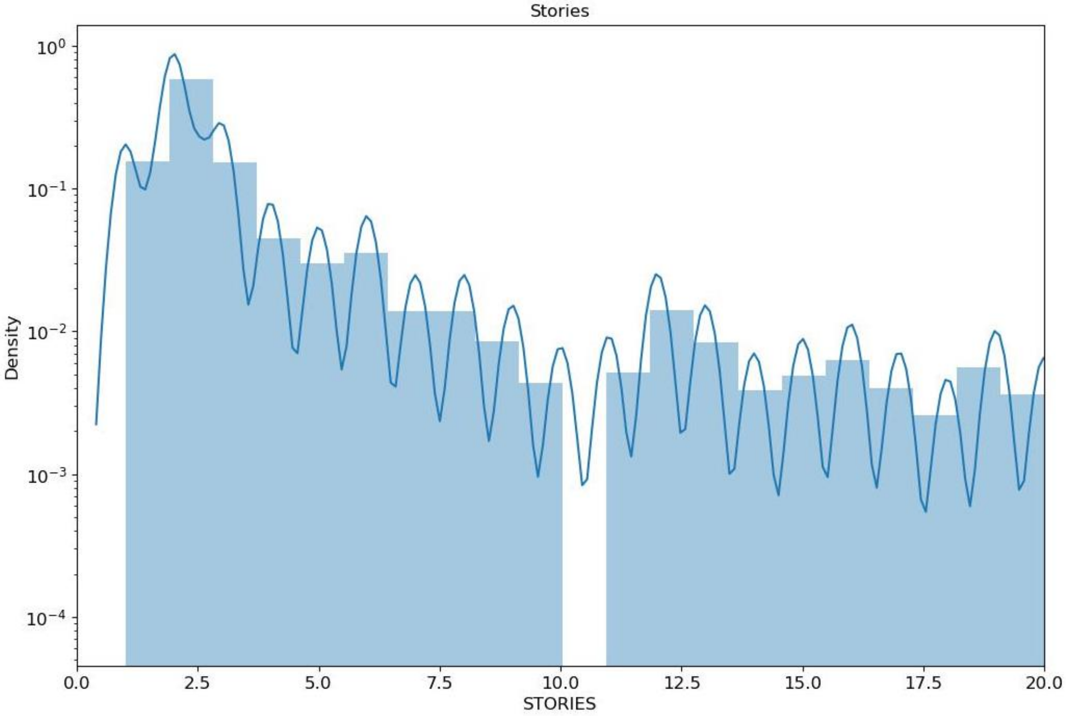


12. EXT: Extension Indicator



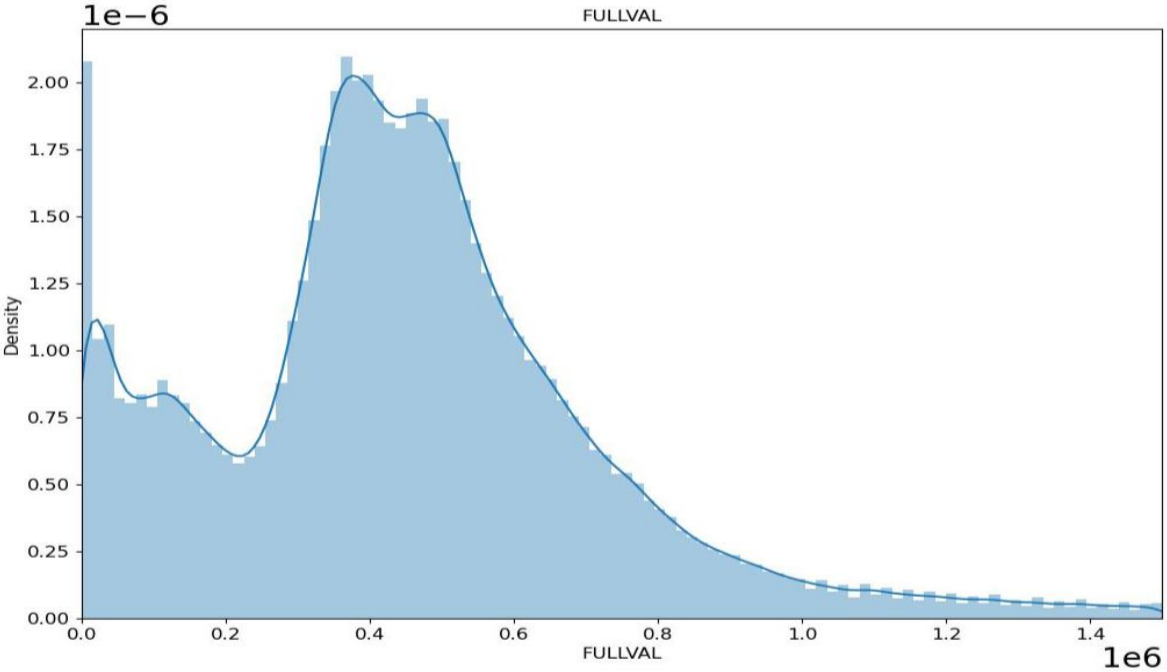
13. STORIES: Number of Stories in Building.

The below figure uses a logarithmic scale and the Stories are limited to 50.

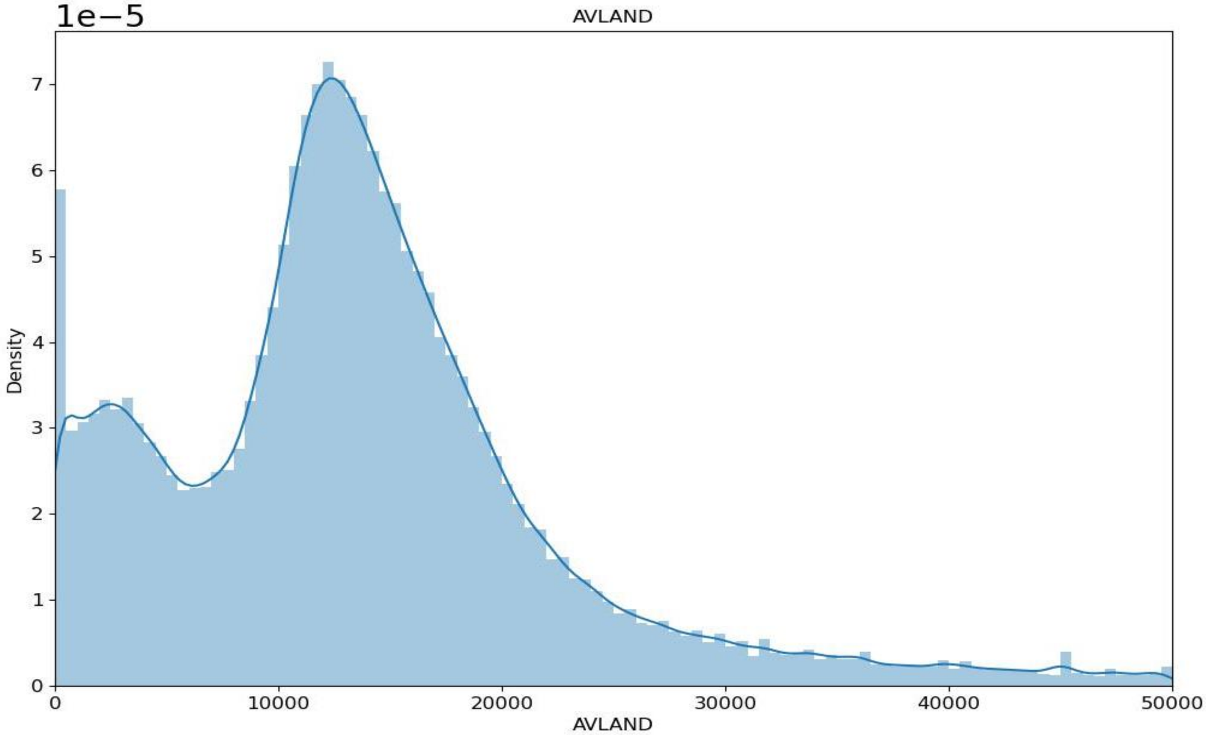


14. FULLVAL: Market Value. 1500000

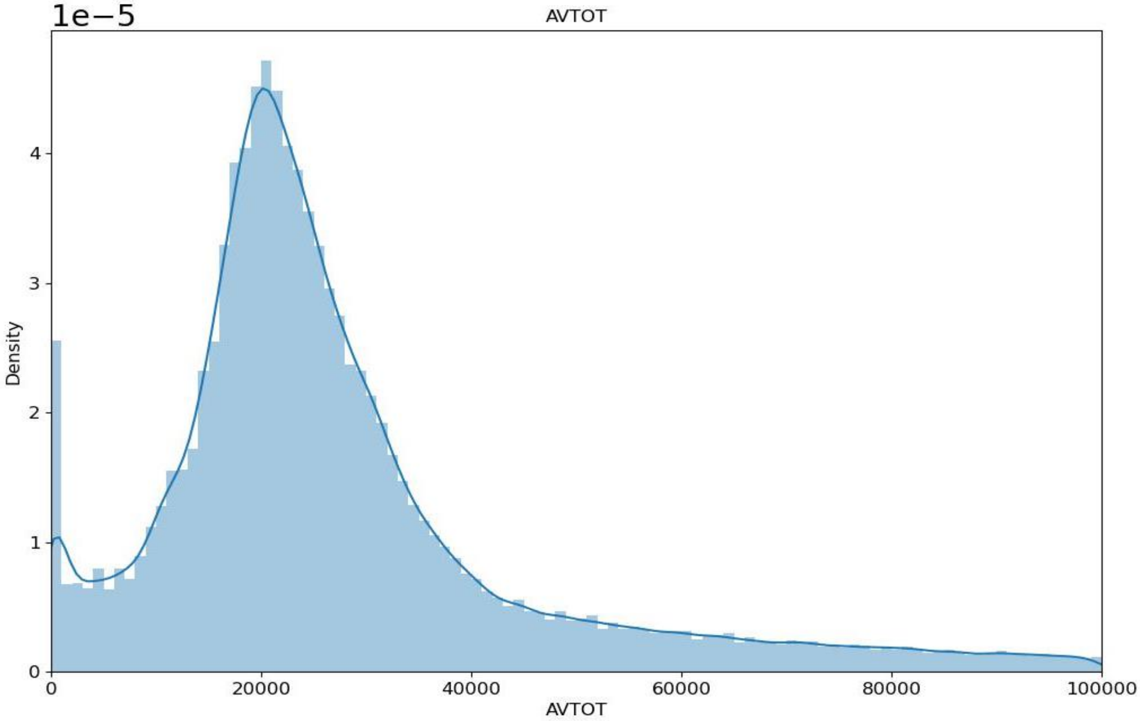
The below figure uses a logarithmic scale and displays up to a market value of 1500000



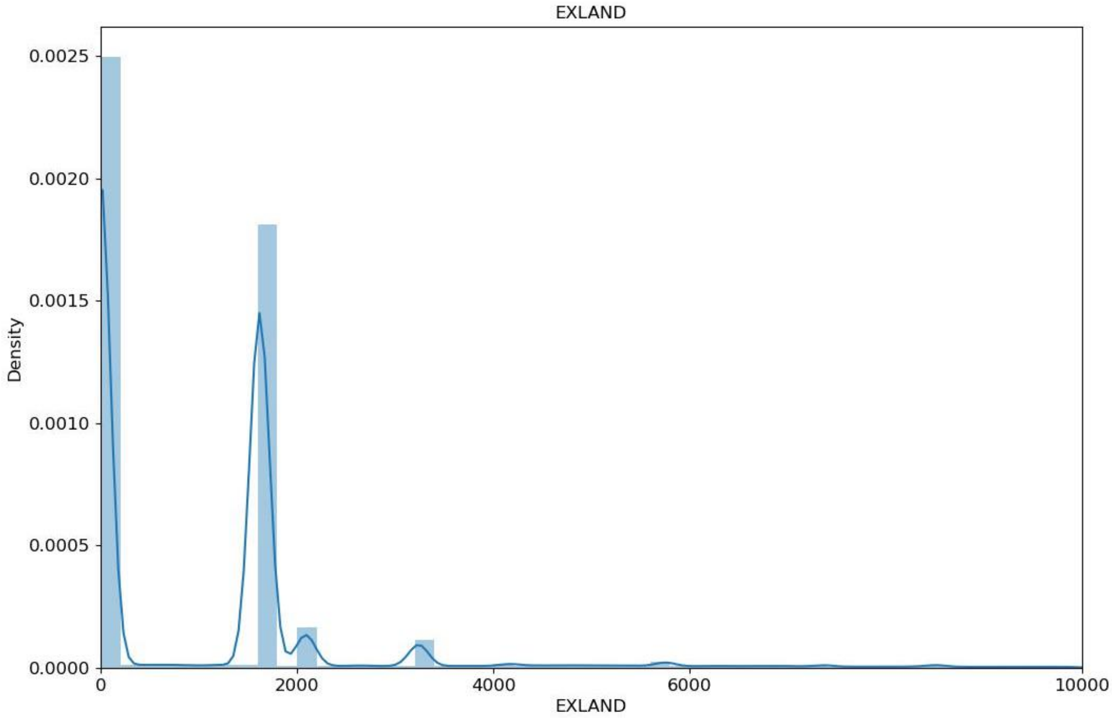
15. AVLAND: Actual Land Value.



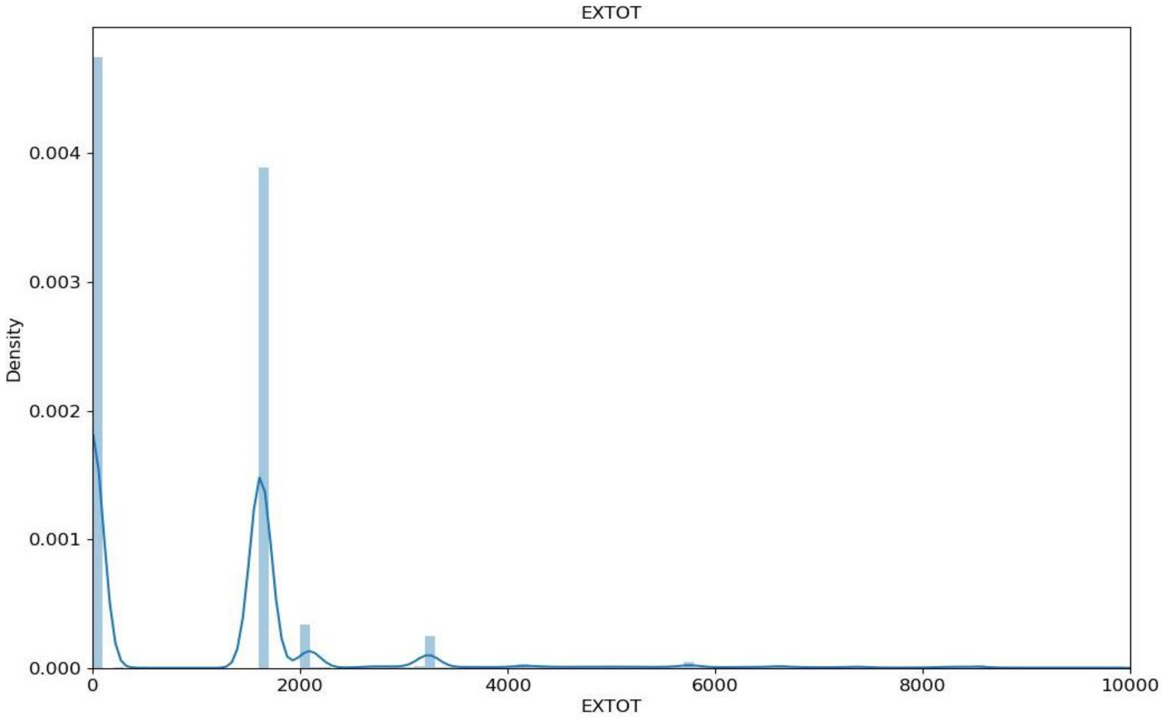
16. AVTOT: Actual Total Value



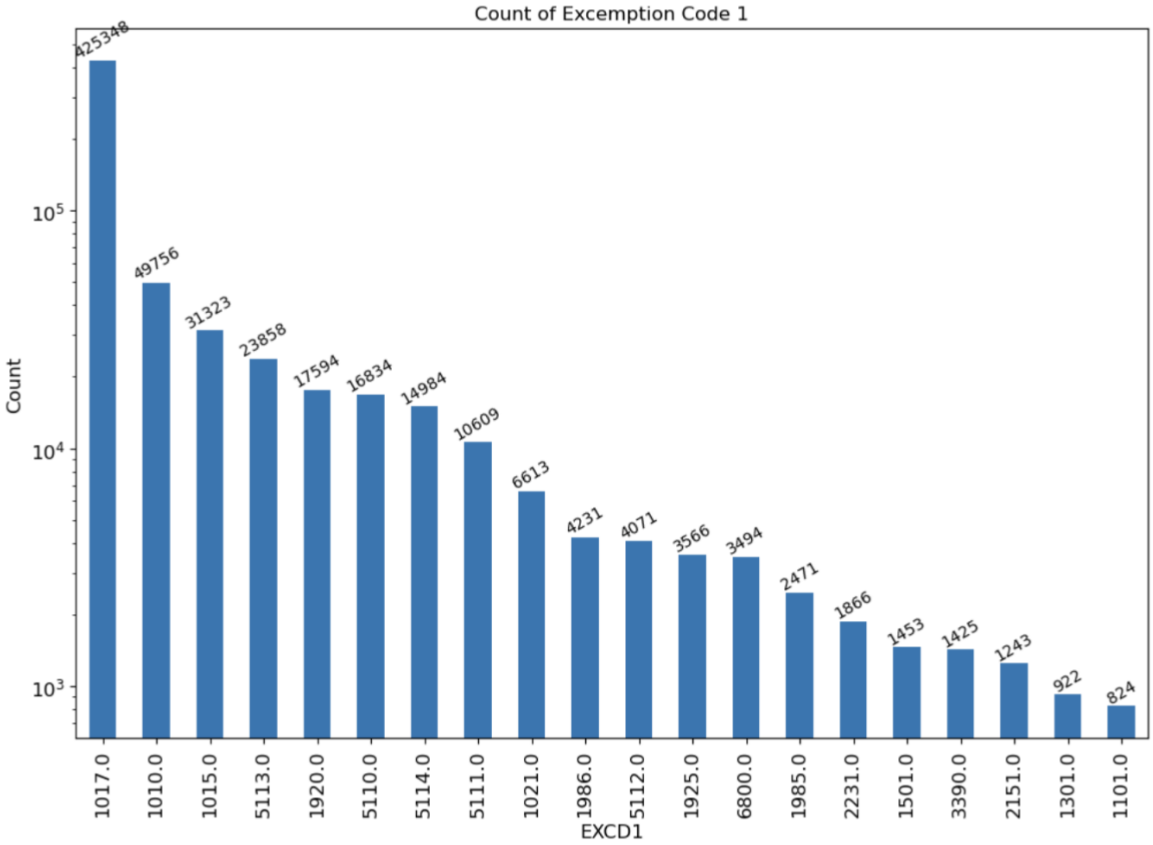
17. EXLAND: Actual Exempt Land Value. The below figure uses a logarithmic scale.



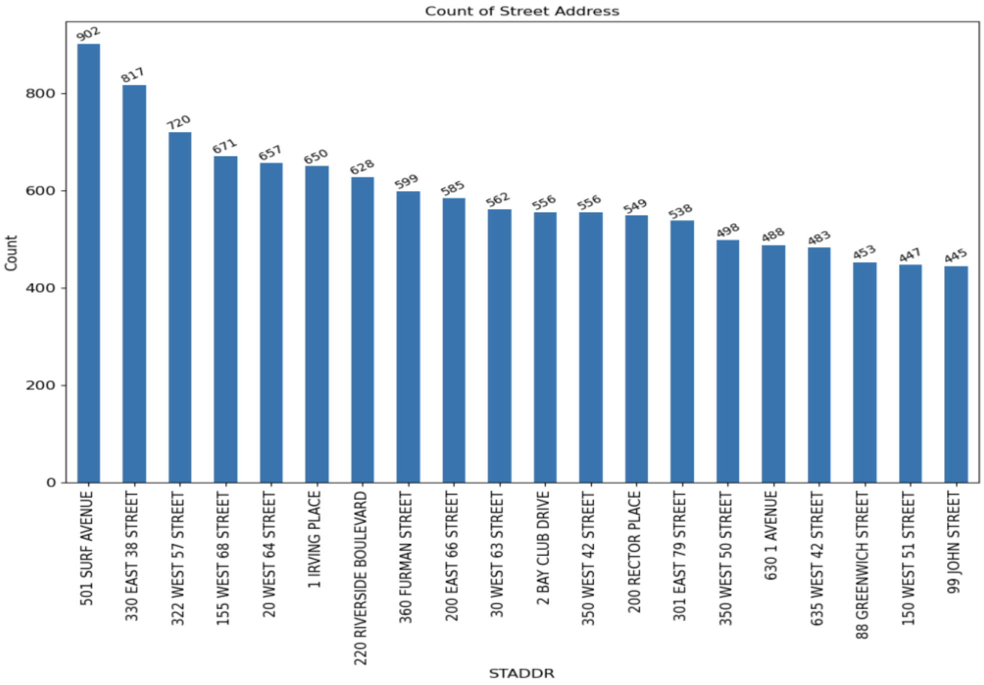
18. EXTOT: Actual Exempt Land Total



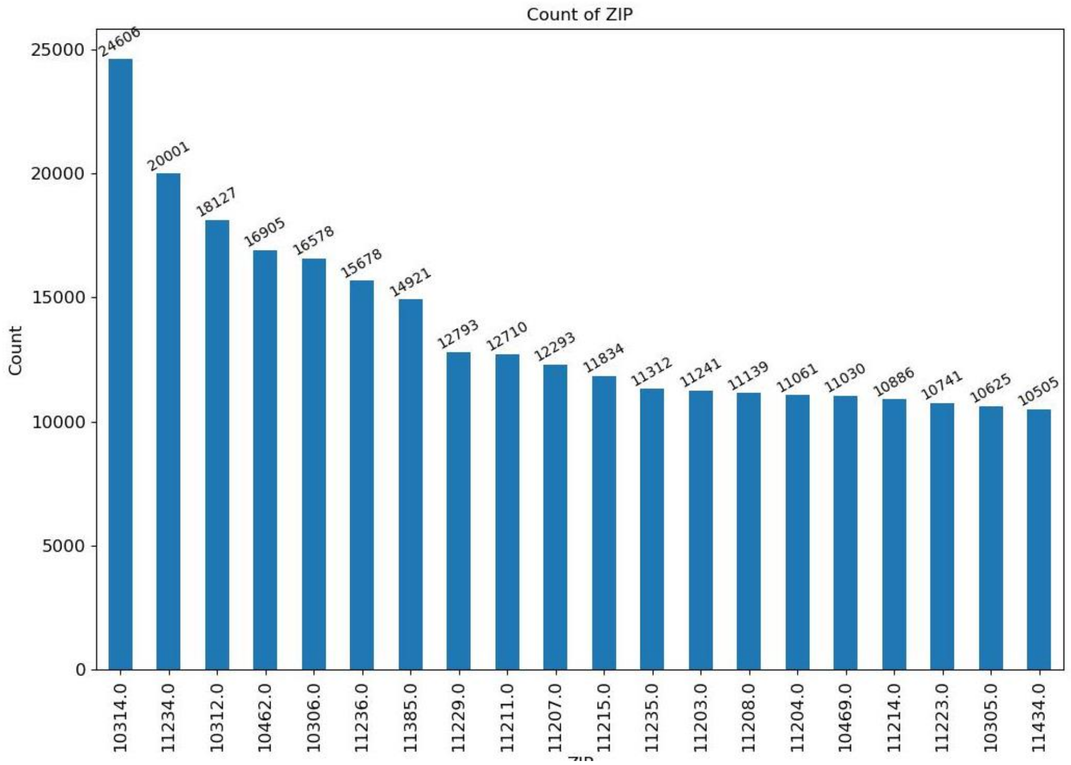
19. EXCD1: Exemption Code 1



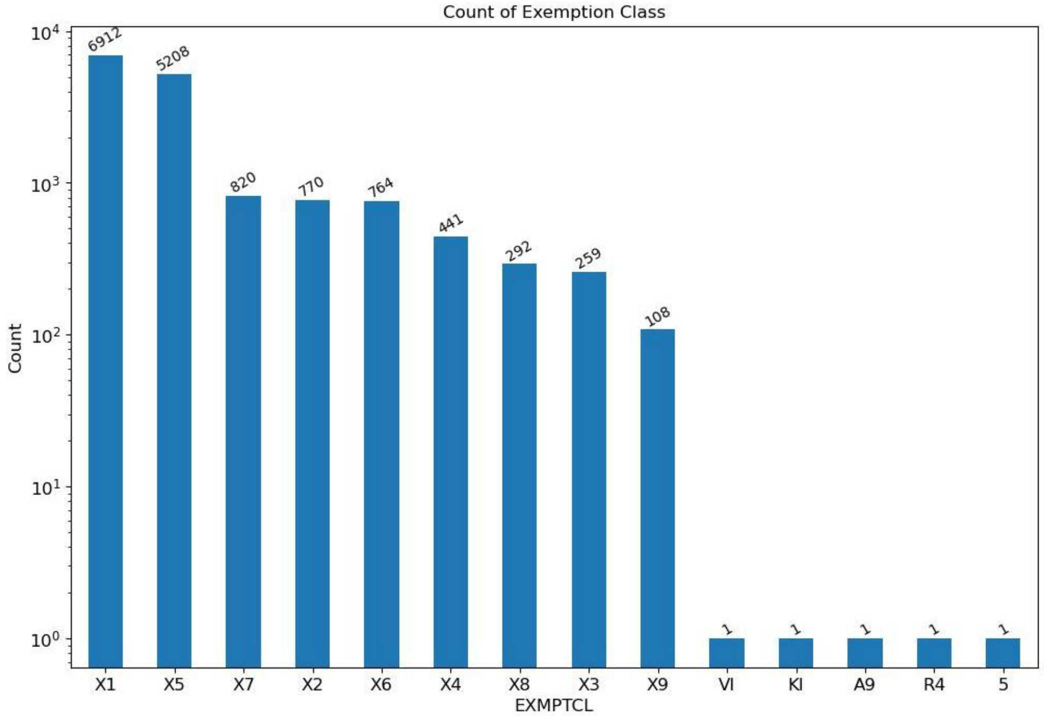
20. STADDR: Street Address.



21. ZIP: ZIP code.

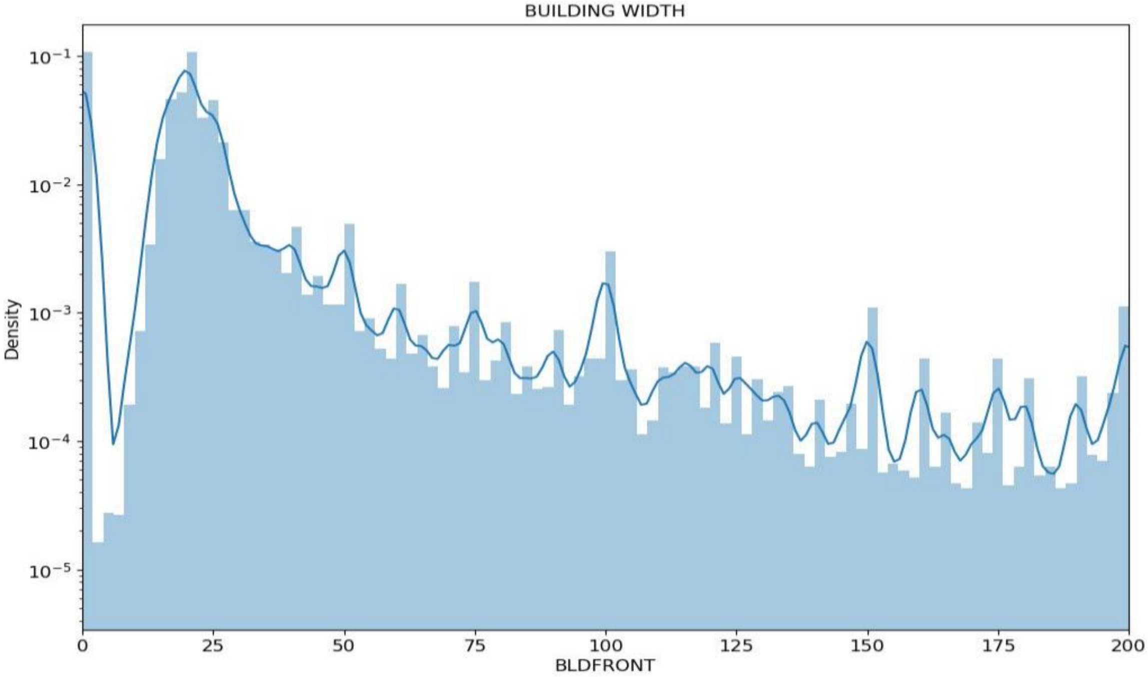


22. EXMPTCL: Exemption Class.



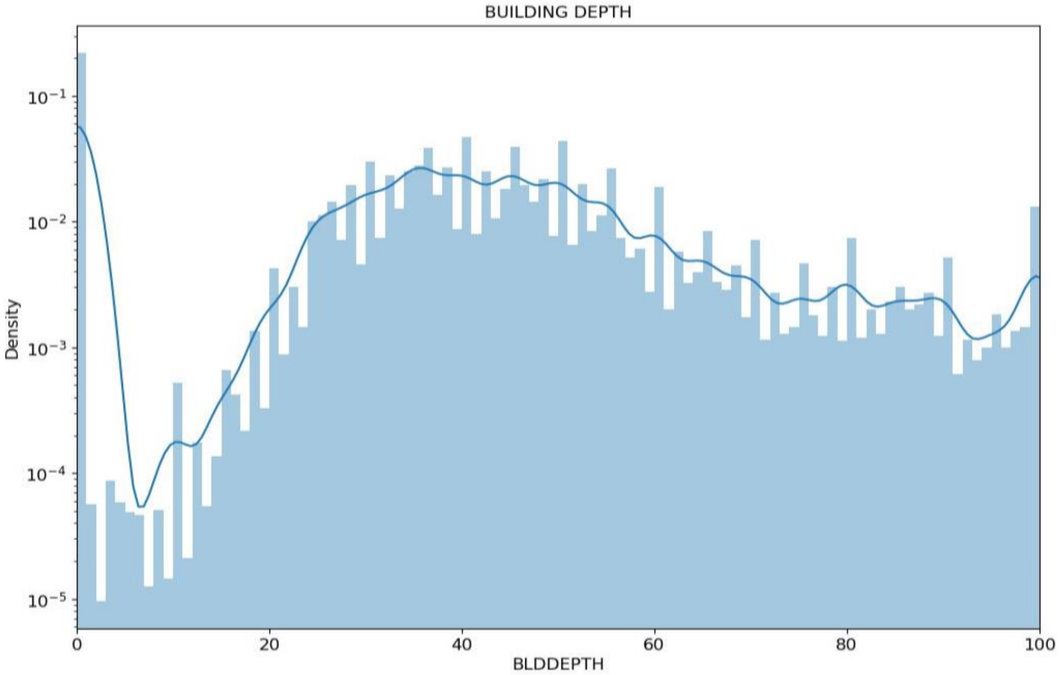
23. BLDFRONT: Building Width.

The below figure uses a logarithmic scale and displays up to a Building Width of 150.



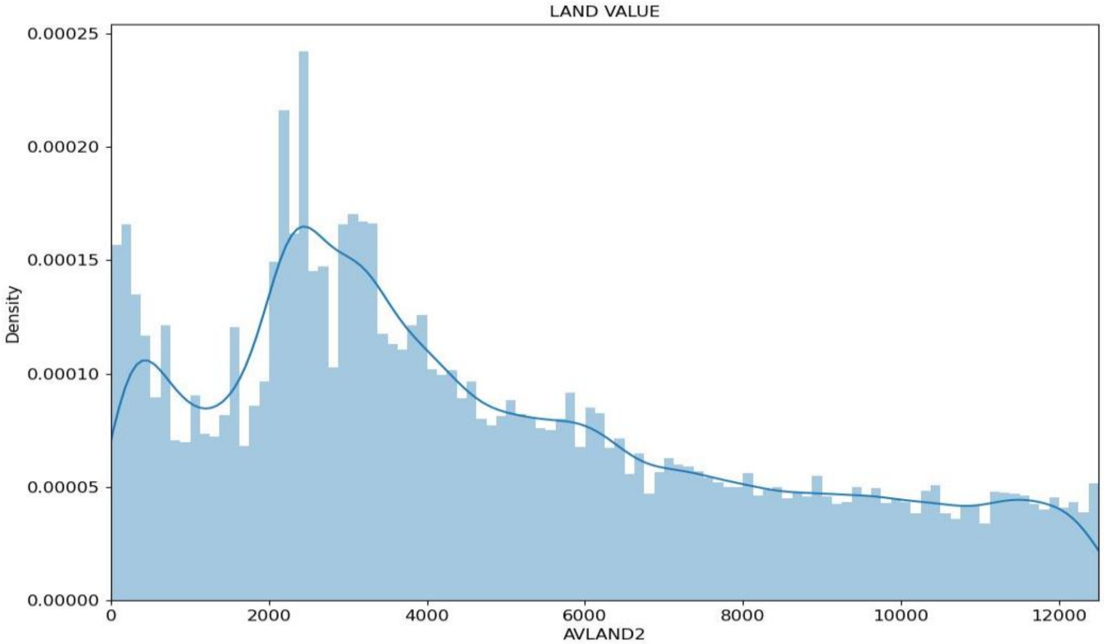
24. BLDDEPTH: Building Depth.

The below figure uses a logarithmic scale and displays Density up to a Building Depth of 100.



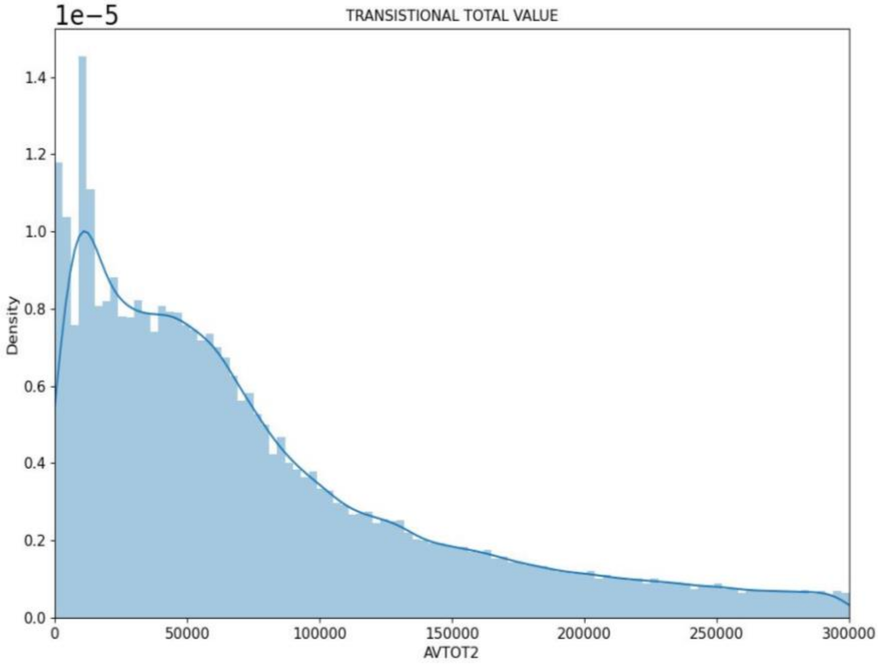
25. AVLAND2: Transitional Land Value.

The below figure displays number of entities up to a 12500 transitional land value.



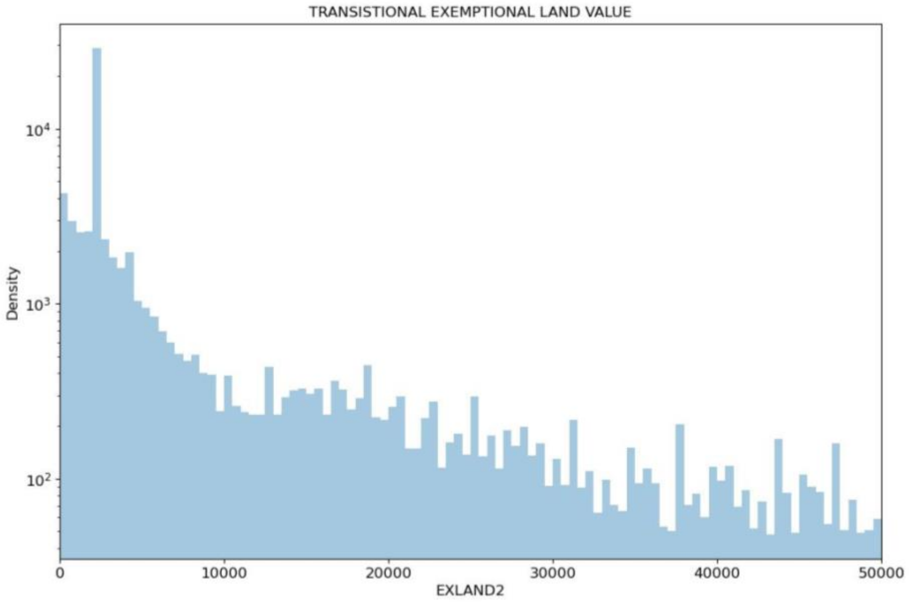
26. AVTOT2: Transitional Total Value.

The below plot shows till a total transitional value of 300000.



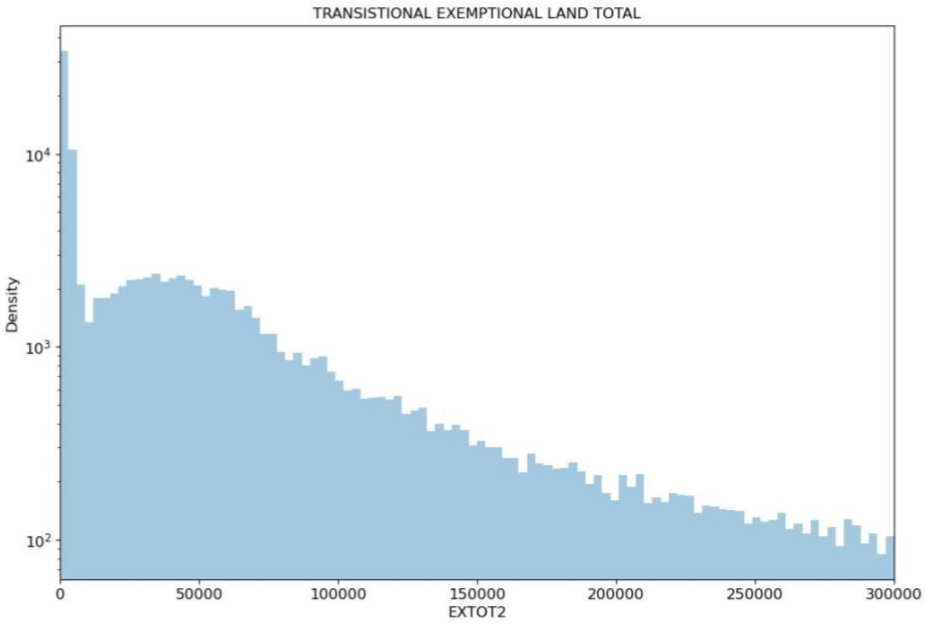
27. EXLAND2: Transitional Exemption Land Value

The below figure uses a logarithmic scale and displays Density up to an Exemption value of 50000.

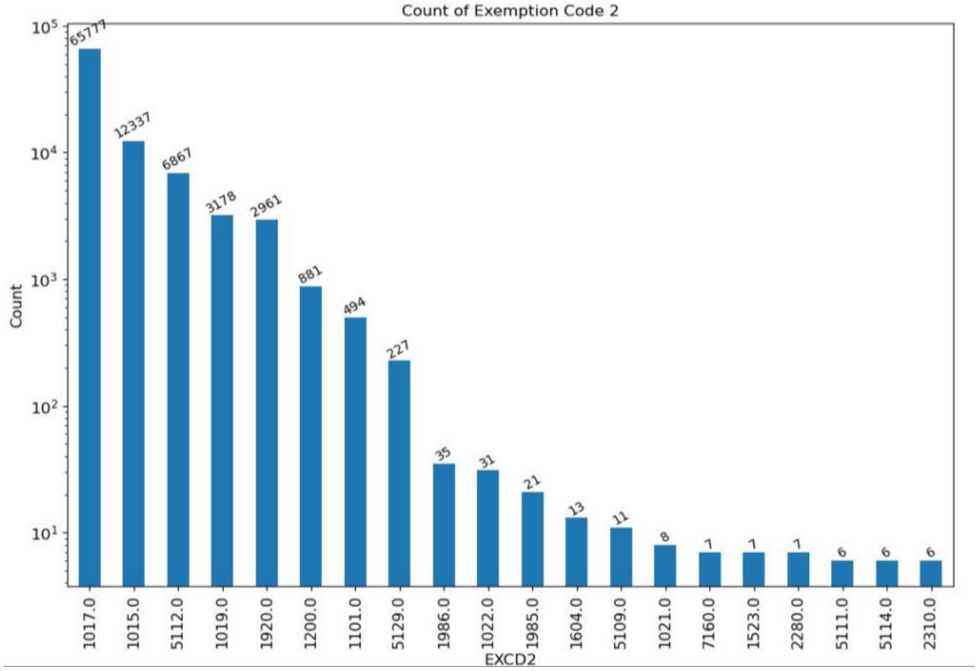


28. EXTOT2: Transitional Exemption Land Total.

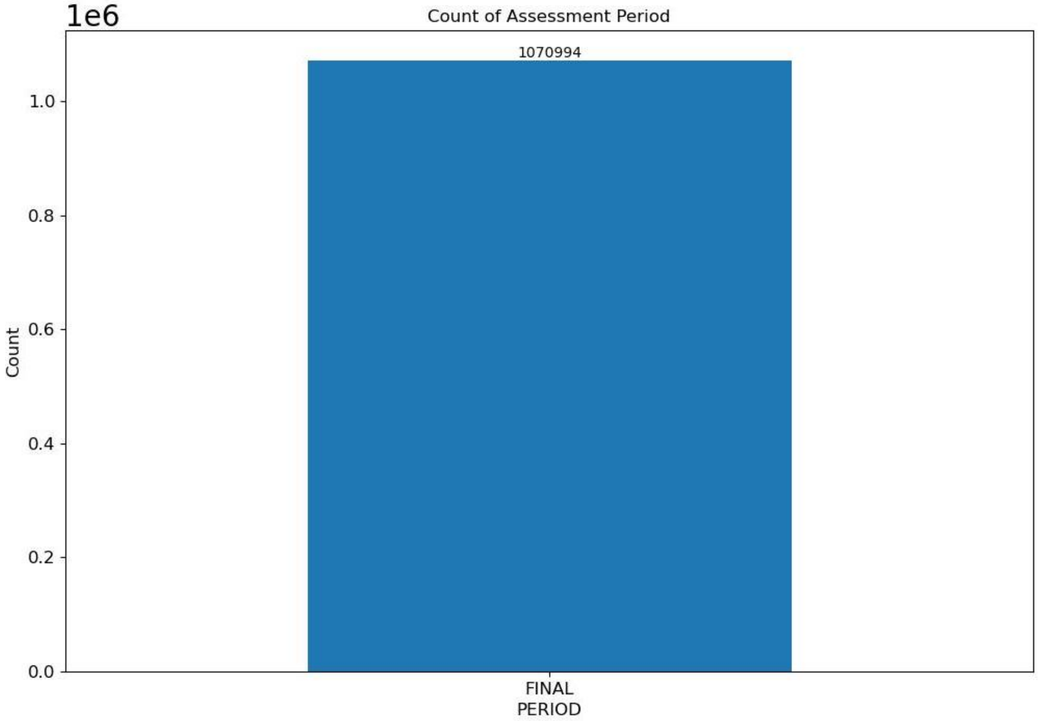
The below figure uses a logarithmic scale and displays Density up to an Exemption value of 300000.



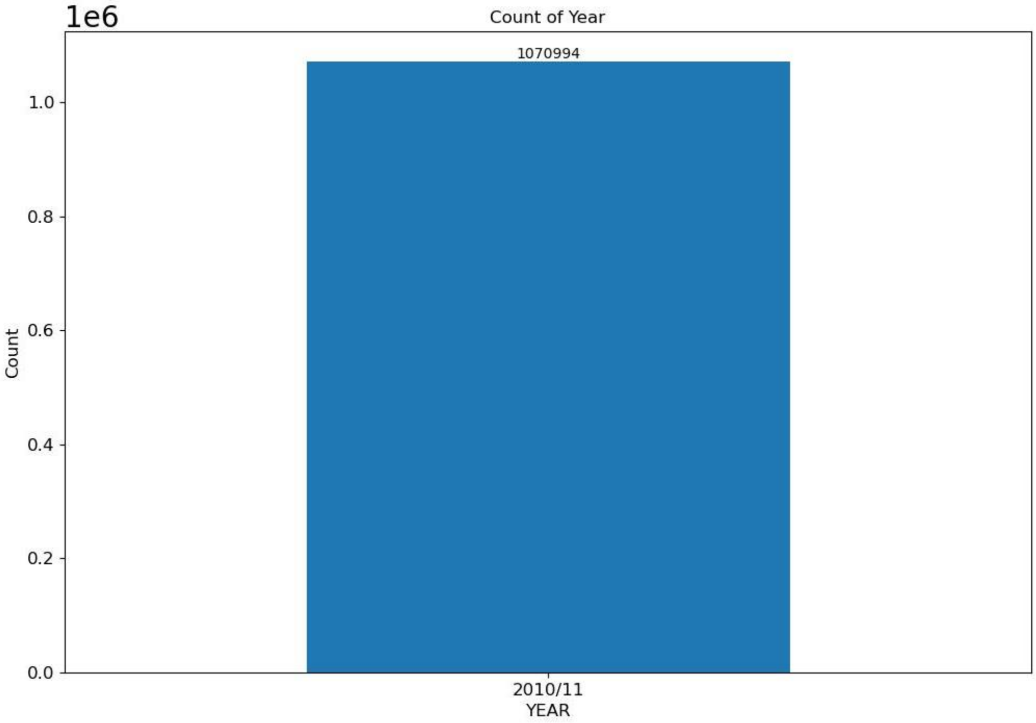
29. EXCD2: Exemption Code 2.
The below figure uses a logarithmic scale.



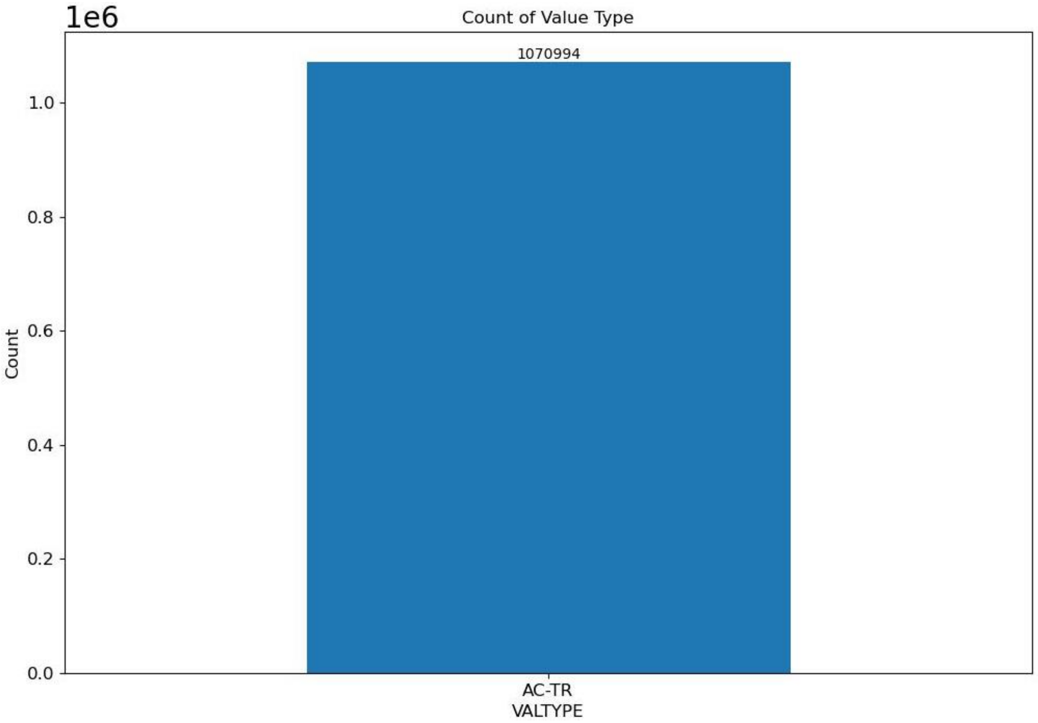
30. PERIOD: Assessment Period. Has only one value.



31. YEAR. Has only one value.



32. VALTYPE: Has only one Value.


1e6
Count of Value Type
Count
AC-TR
VALTYPE