Name: Dimitra Charalampopoulou
Date: May 9th, 2023

*Report collaborated between Zimu Pan and Leo Li (9176889345)

NY Property Project Report

1. Executive Summary:

One of the most common types of fraud is property tax fraud. This project aims to analyze the **NY property data** to find the anomalies property information in New York to catch people underpaying tax by misrepresenting their property characteristics. The original dataset contains a million property records with about 32 fields. Based on the given housing information, our interest is the tax fraud committed by private owners. Therefore, we looked at the "Owner" field and removed the properties that are government-owned by capturing the most common owners. We removed 24478 exclusions from the raw data. Then we described, visualized and filled in the missing values for each remaining variable. After data cleaning, we created 73 new variables based on the existing variables to better assist us obtain a more accurate detection algorithm. Finally, we used two common methods for unsupervised fraud algorithms: Principal Component Analysis and autoencoder to derive two separate fraud scores. The scores were combined and then ranked to get a final fraud score.

The algorithm this project applied is unsupervised machine learning. It doesn't restore any historical data that could indicate the fraud. Therefore, in addition to the final score we obtained, we consulted with the real estate experts to test whether the model is actually efficient in detecting property tax fraud. In order to better visualize it, we color the outliers of ratio variables from each record with z-scaling and find 5 properties that would have potential risk of owners committing tax fraud. With the corresponding address and building information, we are able to analyze the information provided by the owner versus the actual property information to determine if the owner has misclaimed their property tax.

2. Description of Data:

The data is a collection of real New York Property Valuation and Assessment Data provided by the Department of Finance and available on NYC OpenData. There are 14 numeric fields and 18 categorical fields. There are 1,070,994 records for New York Property data in total.

Numeric Fields:

Field Name	# Records With Values	% Populated	# Zeros	Min	Max	Most Common	Mean
LTFRONT	1,070,994	100.00%	169,108	0	9,999	0	37
LTDEPTH	1,070,994	100.00%	170,128	0	9,999	100	89

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

STORIES	1,014,730	94.75%	0	1	119	2	5
FULLVAL	1,070,994	100.00%	13,007	0	6,150,000,000	0	874,265
AVLAND	1,070,994	100.00%	13,009	0	2,668,500,000	0	85,068
AVTOT	1,070,994	100.00%	13,007	0	4,668,308,947	0	227,238
EXLAND	1,070,994	100.00%	491,699	0	2,668,500,000	0	36,424
EXTOT	1,070,994	100.00%	432,572	0	4,668,308,947	0	91,187
BLDFRONT	1,070,994	100.00%	228,815	0	7,575	0	23
BLDDEPTH	1,070,994	100.00%	228,853	0	9,393	0	40
AVLAND2	282,726	26.40%	0	3	2,371,005,000	2,408	246,236
AVTOT2	282,732	26.40%	0	3	4,501,180,002	750	713,911
EXLAND2	87,449	8.17%	0	1	2,371,005,000	2,090	351,236
EXTOT2	130,875	12.22%	0	7	656,768	2,090	656,768

Categorical Fields:

Field Name	# Records With Values	% Populated	# Zeros	# Unique Values	Most Common

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

RECORD	1,070,994	100.00%	0	1,070,994	1
BBLE	1,070,994	100.00%	0	1,070,994	1,000,010,101
BORO	1,070,994	100.00%	0	5	4
BLOCK	1,070,994	100.00%	0	13,984	3,944
LOT	1,070,994	100.00%	0	6,366	1
EASEMENT	4,636	0.43%	0	12	E
OWNER	1,039,249	97.04%	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	1,070,994	100.00%	0	200	R4
TAXCLASS	1,070,994	100.00%	0	11	1
EXT	354,305	33.08%	0	3	G
EXCD1	638,488	59.62%	0	129	1,017
STADDR	1,070,318	99.94%	0	839,280	501 SURF AVENUE
ZIP	1,041,104	97.21%	0	196	10,314
EXMPTCL	15,579	1.45%	0	14	X1
EXCD2	92,948	8.68%	0	60	1,017

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

PERIOD	1,070,994	100.00%	0	1	FINAL
YEAR	1,070,994	100.00%	0	1	Nov-10
VALTYPE	1,070,994	100.00%	0	1	AC-TR

3. **Data Cleaning:**

The main logic behind the data imputation method is to use an alternative value in place of the missing data based on other related fields. In this specific dataset, our interest is the tax fraud committed by private owners. Therefore, we looked at the "Owner" field and removed the properties that are government-owned by capturing the most common owners. We removed 24478 exclusions from the raw data. To fill the missing values in "ZIP" field, we looked the records before and after the missing row, if they are the same, we fill in the zip with that value. We used this logic to fill up 11423 records, and for the remaining 10114 missing records, we fill them with the previous zip value. For the missing values in fields "AVTOT", "AVLAND" and "FULLVAL", we group by "TAXCLASS" and then fill them with the average value within each tax class kind. We use the same logic on filling the missing values in "STORIES" field as well. For the missing values in fields "FULLVAL", "AVLAND" and "AVTOT", we treat both 0 and 1 as missing value and fill them with the average value of grouped by "TAXCLASS".

4. **Variable Creation:**

Variable Summary Table:

Description of Variables	# of Variables
size variables: Lot size, building area and building volume obtained from original variable LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, STORIES	3
Price Ratio Variables: Each of the 3 money value fields normalized by each of the 3 sizes. These variables are useful for indicating the undervalued or overvalued properties and capture the unusually large $ values	9

Inverse Price Ratio Variables: The inverse value of price ratio. These variables demonstrate the relationship between price and building size and capture the unusually small $values	9
Ratio of Zip scale factor Variables: Using zip code as the scale factor, we calculate the average price ratio among zip codes to compare with each price ratio and inverse price ratio. This would give us the property pricing relationship within each particular geographic area.	18
Ratio of Tax Class scale factor Variables: Using tax class as the scale factor, we calculate the average price ratio among all different tax classes to compare with each price ratio and inverse price ratio. This would give us the property pricing relationship within each particular tax class.	18
Value Ratio Variables: The ratio that represents how appropriately the three value fields relate. This is useful to discover the unusually relationship between each value.	1
Ratio of Stories scale factor Variables (New Variables): Using number of story as the scale factor, we calculate the average price ratio among all story number to compare with each price ratio and inverse price ratio. This would give us the property pricing relationship within each particular building structures.	18

In order to find the property value in relationship to size, we first created three size variables. The first variable is to calculate the area of the lot (ltsize) by multiplying the lot frontage size (LTFRONT) with the lot depth size (LTDEPTH). We then focus on the building size especially on building area (bldsize) and building volume (bldvol) by using the variables

"BLDFRONT", "BLDDEPTH" and "STOREIS". Both of these variables can be a good indication on property price change, which means bigger and higher buildings are expected to have higher property valuations. Then we created 9 price ratio variables using "FULLVAL", "AVLAND" and "AVTOT" to divide each size variable to obtain the normalized price per area unit. These variables are useful for indicating the undervalued or overvalued properties and capture the unusually large $ values. We also created 9 inverse price ratio variables corresponding to each price ratio with different property sizes to illustrate the relationship between price and building size. Next, we calculated grouped averages of both price ratio and inverse price ratio variables that were grouped by ZIP3 (first three digits of 5-digit zip code), TAXCLASS and STORIES (customized variables). This would allow us to analyze the property price by location and type of building (embedded into TAXCLASS, STORIES) and help us assess whether a property assessment value is too high or too low. Finally, we created a value ratio variable to capture if any property owners have misclaimed the value by detecting the abnormal correlations between FULLVAL, AVLAND and AVTOT.

5. **Dimensionality Reduction:**

In an unsupervised problem like this, we are unable to use a filter and wrapper to reduce dimension since there's no dependent variable. Therefore, we chose to use Principal Component Analysis. Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform a high-dimensional dataset into a lower-dimensional representation by identifying the directions (principal components) that capture the maximum variance in the data. To perform PCA, we first normalized the dataset ($\frac{x-\underline{x}}{std(x)}$), which involved subtracting the means of each variable and then dividing by the standard deviation. For this problem, we kept the <u>first 5 principal components</u> and subsequently normalized the principal components as well. The reason for normalizing principal components is that we want to equalize the impact of each principal component because otherwise, the first PC will have a larger impact on the algorithm than the others.

6. **Anomaly Detection Algorithm:**

We have used 2 algorithms to find abnormal properties that could indicate fraud. The first algorithm is the sum of normalized principal component analysis values and the second algorithm is the autoencoder. Each algorithm will give each property a score to quantify the strangeness of such property and we combine 2 scores to calculate the final fraud score. Below is the details of both algorithms.

<u>Algorithm 1: Normalized PCA</u>

For the first algorithm, we're looking at the principal component score only. We want to combine all 5 principal components for each property and find out the abnormal ones. To calculate score 1, we use Minkowski distance where p = 2, meaning that we take the absolute value of each variable, raise it to a power of 2, sum all values of each row, and then take a square root of the sum. Through experimentation, we have concluded that the value of p has minimal effect on our algorithms so we use the commonly euclidean distance of p = 2 to calculate score 1. Below is the formula to calculate score 1 for property i.

$$S_i = \sqrt{\sum_n |z_n^i|^2}$$

Algorithm 2: Autoencoder

An autoencoder is a type of neural network that learns to encode and decode data by training on a dataset to minimize the difference between the original input and the reconstructed output. It can be used to detect anomalies by checking the difference between the input data and predicted output data. The larger the difference, the more likely the record is fraud. For the autoencoder algorithm, we build a neural network using 1 hidden layer and 3 neurons. We fit the normalized PCA data into the neural net and calculate the error term as the predicted value - actual value and used similar logic as algorithm 1 to calculate the Euclidean distance of each property. Below is the formula to calculate score 2 for property i.

$$S_i = \sqrt{\sum_n |z_n'^i - z_n^i|^2}$$

Combining 2 Algorithms

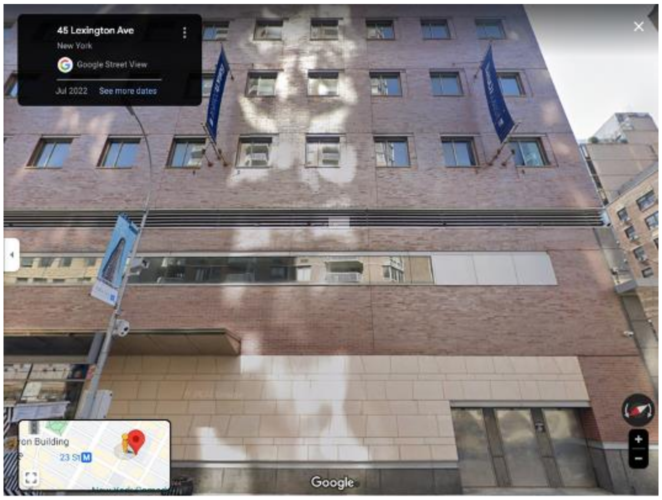
Since the two scores have different scales, to calculate the final score for each property, we ranked the scores of each property. We then used the formula 0.5* score 1 rank + 0.5* score 2 rank to derive the final score.

7. **Results:**

Upon obtaining the final score, we ranked all properties from those with the highest score to the lowest so that we can start to investigate properties with extremely high scores. Similar to any machine learning algorithm, one cannot blindly trust the model. We should combine the result of the algorithm and expertise in real estate to make the final judgment. To properly use the scores, one should start looking at the properties with the highest score and investigate our expert variables to find the reason for the strangeness. A good way is to color code all variables so that we can easily tell which scores are too high or too low. Note that it is important to differentiate wrong values from fraudulent behaviors. Some properties scored high because the values are clearly wrong; therefore, it is imperative to look at all variables holistically to detect potential fraud. Below are 5 properties that we have investigated to be likely fraud.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

Record: 47984 (rank #32)



- LTFRONT: 39
- LTDEPTH: 50
- STORIES: 2
- FULLVAL: 138,000,000
- AVLAND: 11,025,000
- AVLOT: 62,100,000
- BLDFRONT: 39
- BLDDEPTH: 50
- Owner: BERKOWTIZ, ULWT LOUIS
- Address: 45 Lexington Ave

Reasons that it might be fraud:

1) The building has more than 2 stories
2) Lot depth and lot width seem to small for the value of the property
3) r4inv_taxclass and r6inv_taxclass are both extremely low
4) Building size seems to be incorrect
5) BLDFRONT and BLDDPTH are the same as LTFRONT and LTDEPTH. This would be a potential misreport on actually building size, which is indicated by abnormal value of r2 and r3 value.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

Record: 934793 (rank #142)

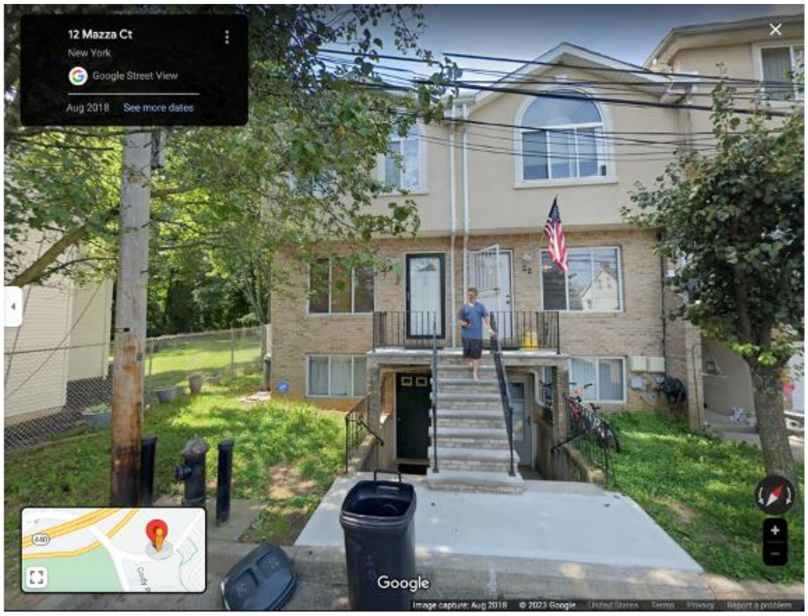
- LTFRONT: 2798
- LTDEPTH: 997
- STORIES: 1
- FULLVAL: 273,000,000
- AVLAND: 10,920,000
- AVLOT: 16,380,000
- BLDFRONT: 30
- BLDDEPTH: 40
- Owner: BREEZY POINT COOPERAT
- Address: 217-02 BREEZY POINT BLVD

Reasons that it might be fraud:

1) Building width and depth seem too small compared to both lot width, lot depth, and the value of the property
2) Very high r1 and r2 scores

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

<u>Record: 1049911 (rank #111)</u>



- LTFRONT: 23
- LTDEPTH: 193
- STORIES: 3
- FULLVAL: 560,000
- AVLAND: 25
- AVLOT: 20,356
- BLDFRONT: 13
- BLDDEPTH: 55
- Owner: KENILWORTH HOLDINGS L
- Address: 1927 ARTHUR KILL ROAD

Reasons that it might be fraud:

1) Actual land value is way too small for this property
2) r4inv_taxclass and r6inv_taxclass are both high

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

Record: 333412 (rank #10)

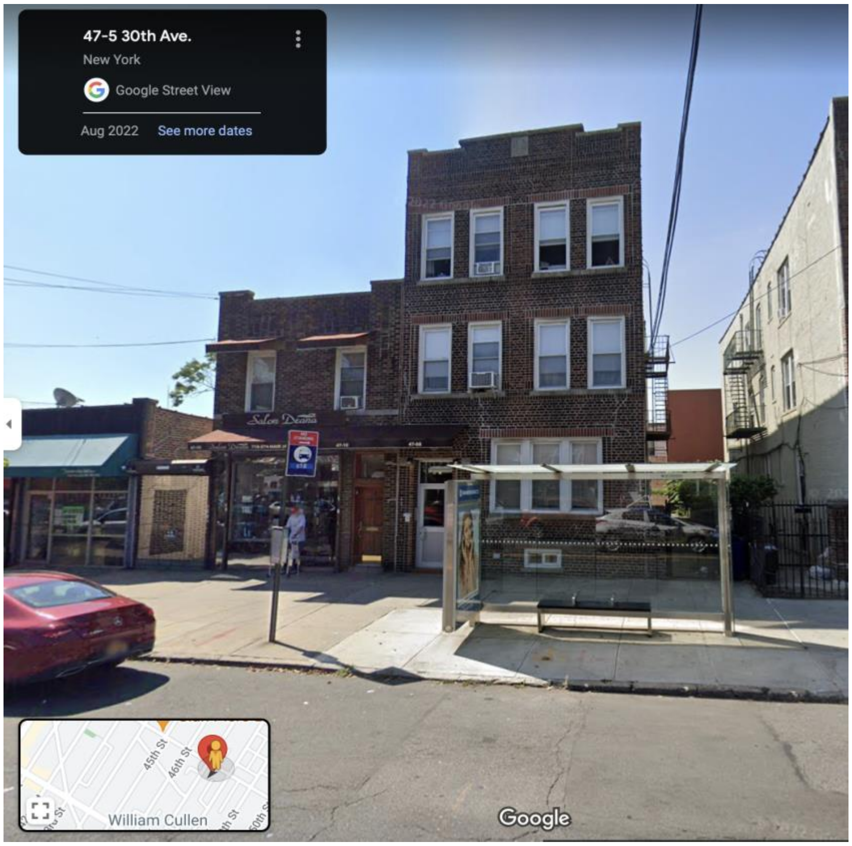
- LTFRONT: 17
- LTDEPTH: 85
- STORIES: 3
- FULLVAL: 9060
- AVLAND: 3874
- AVTOT: 4077
- BLDFRONT: 4017
- BLDDEPTH: 42
- Owner: Spooner Alston
- Address: 37 Monroe Street, 11238 NY

Reasons it might be a fraud:

1) As we see from the property picture above, this property is a segment of the entire street house. It reports an unusual high number for the building front size. This has misrepresented the building size and volume that is indicated by extreme small values of r2 and r3.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

<u>Record: 579512 (ranked #60)</u>



- LTFRONT: 150
- LTDEPTH: 145
- STORIES: 1
- FULLVAL: 775,000
- AVLAND: 344,250
- AVTOT: 348,750
- BLDFRONT: 5
- BLDDEPTH: 5
- Owner: DORMITORY AUTHORITY OF NY
- Address: 47-08 30 STREET

Reasons it might be a fraud:

1) Based on the property picture demonstrated above, the building has multiple stories instead of one as reported.
2) The building front and depth size are not 5 as shown in the picture. This information would misrepresent actual building size which is indicated by the abnormal value of r3.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

8. Summary:

We were given a dataset from New York City's Department of Finance. It contains information on 1,070,994 properties with 32 fields. We were tasked to build machine-learning models to identify potential fraud. This is an unsupervised problem, meaning that we don't know which records are fraudulent; therefore, the goal is to develop an algorithm that can catch the strange and abnormal properties to present to the city of New York to aid them in catching tax fraudsters. To start this project, we first removed 24,478 properties that belong to the government. This dataset contains many missing data. Believing that the dataset is organized by zipcode, we filled in the missing zip code by looking at the records above and below. For the rest of the missing values, we imputed by using the average of the tax class of that property.

We then created 73 variables in total. The logic behind variable creation is to check the price ratio of each property. Since each property has different values(full value, land value, etc.) and different sizes (lot, building, etc.), we calculated each ratio of value/size and the inverse of the value to catch abnormally small values. We then calculated the difference in the value between each property, the average tax class, zip code, and stories. To reduce dimensionality, we employed principal component analysis. We took the first 5 principal components and subsequently normalized all PCs to build our fraud algorithm.

For our model, we built 2 algorithms and combined them to compute the final fraud score of each property. The first algorithm is calculating the Euclidean distance of PCs of each property. For the second algorithm, we built an autoencoder and the result is from the error of the autoencoder prediction and the actual values of the record. We ranked the result of the two algorithms and combine the rank to derive the final fraud score. To use our model, one should investigate the properties with high fraud scores and incorporate real-estate and tax expertise to catch fraud.

Since this is an unsupervised model, we can't quantitatively evaluate the model. The best way to improve it is to show the result to experts and gather feedback from them. For example, we can send the top 100 - 500 records to some experts and ask them to help us evaluate the model. They can tell us if the model is catching any fraud at all or if the model is only catching one type of fraud or catching fraud they're not particularly interested in. With expert feedback, we can adjust the model by adding and subtracting variables and adjusting exclusions. For example we can build variables from AVLAND2 or exclude more properties that our clients are not interested in. This process will likely take 2 - 3 iterations where we present our findings, get feedback, and improve the model. After that, we would likely get a high-performance model that can help catch fraud.

9. Appendix:

Data Quality Report

1. Data Description

This dataset is provided by New York City's Department of Finance. It updates annually and the department uses it to calculate property taxes. It contains **32 fields** and **1,070,994 records**.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

2. Summary Table

(1) Numerical Table

Field Name	% Populated	Min	Max	Mean	Stdev	Most Common
LTFRONT	100%	0	9999	36.64	74.03	0
LTDEPTH	100%	0	9999	88.86	76.40	100
FULLVAL	94.75%	1	1	119	8.37	5.007
AVLAND	100%	0	2,668,500,000	85,067.92	4,057,260	0
AVTOT	100%	0	4,668,308,947	227,238	6,877,529	0
EXLAND	100%	0	2,668,500,000	36,423	3,981,575	0
EXTOT	100%	0	4,668,408,947	91,186	6,508,402	0
BLDFRONT	100%	0	7,575	23.04	35.58	0
BLDDEPTH	100%	0	9,393	39.92	42.71	0
AVLAND2	26.40%	3	2,371,005,000	246,235	6,178,962	2,408
AVTOT2	26.40%	3	4,501,180,002	713,911	11,652,528	750
EXLAND2	8.17%	1	237,100,500	351,235	10,802,212	2,090
EXTOT2	12.22%	7	4501,180,002	656,768	16,072,510	2,090

(2) Categorical Table

Field Name	% Populated	# Unique Values	Most Common Value
Record	100%	1,070,994	N/A
BBLE	100%	1,070,994	N/A
BORO	100%	5	4
BLOCK	100%	13,984	3944
LOT	100%	6,366	1
EASEMENT	0.43%	12	E
OWNER	97.03%	863,347	PARKCHESTER PRE SERVAT
BLDGCL	100%	200	R4
TAXCLASS	100%	11	1
EXT	33.08%	3	G
EXCD1	59.62%	129	1017
STADDR	99.94%	839,280	501 SURF AVENUE
ZIP	97.21%	196	10314
EXMPTCL	1.45%	14	X1
EXCD2	8.68%	60	1017
PERIOD	100%	1	FINAL
YEAR	100%	1	2010/11
VALTYPE	100%	1	AC-TR

3. Visualization of Each Field

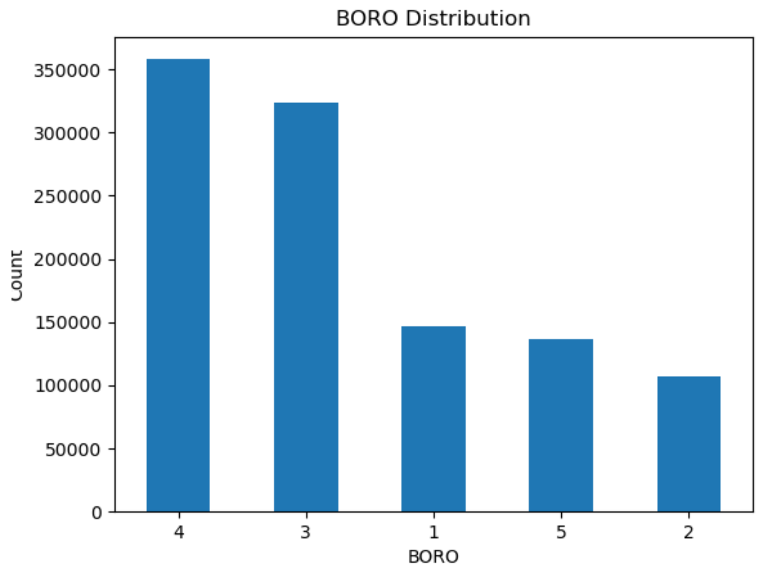
1) Field Name: Record
   Description: a unique number for each property, ranging from 1 to 1070994.

2) Field Name: BBLE
   Description: a unique file key for each property.

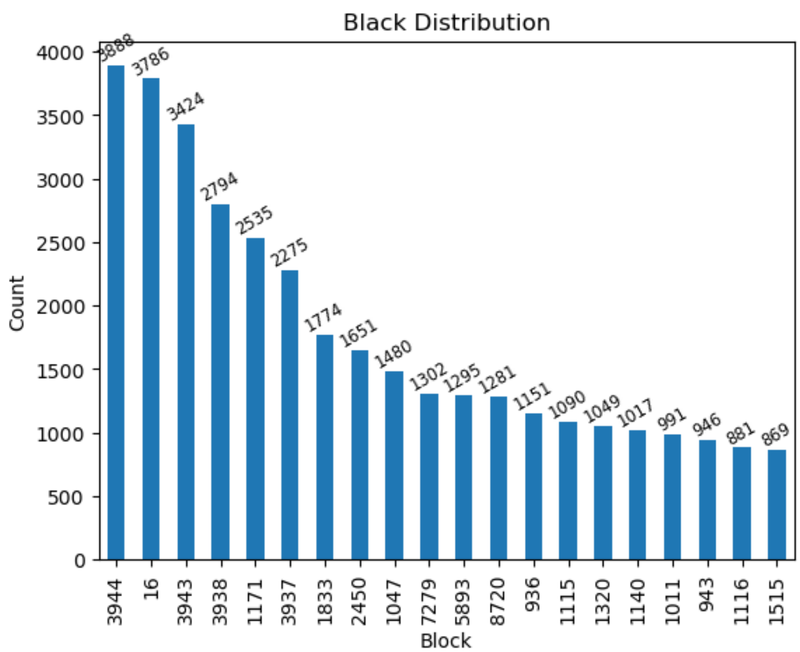
3) Field Name: BORO

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

Description: 5 boroughs of New York City including Manhattan(1), Bronx(2), Brooklyn(3), Queens(4), and Staten Island(5). Below is the distribution. Queens has the most properties in this dataset.

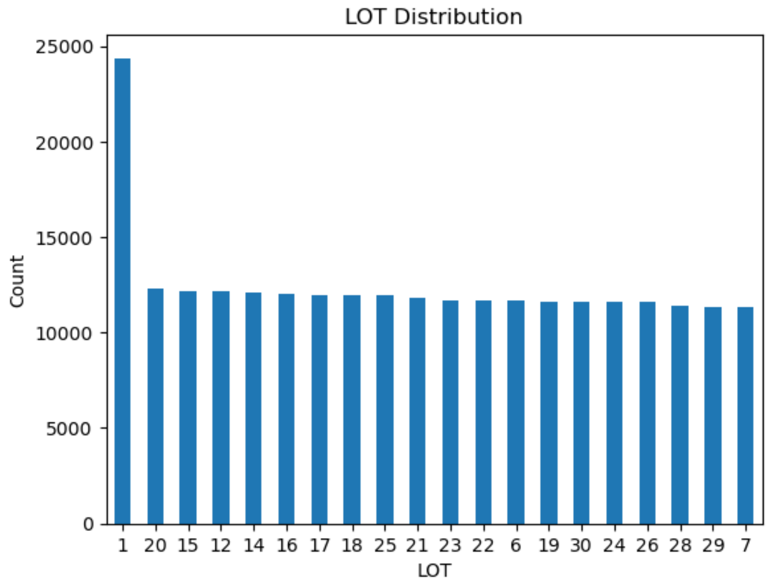


4) Field Name: BLOCK
Description: within each borough, there are lots of blocks. Below is the distribution.

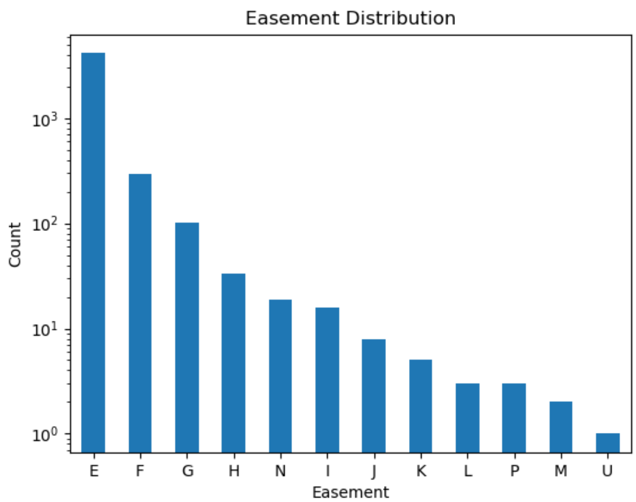


5) Field Name: LOT
Description: A lot is a portion of land that is designated for a specific purpose, such as residential, commercial, etc. Below is the distribution.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

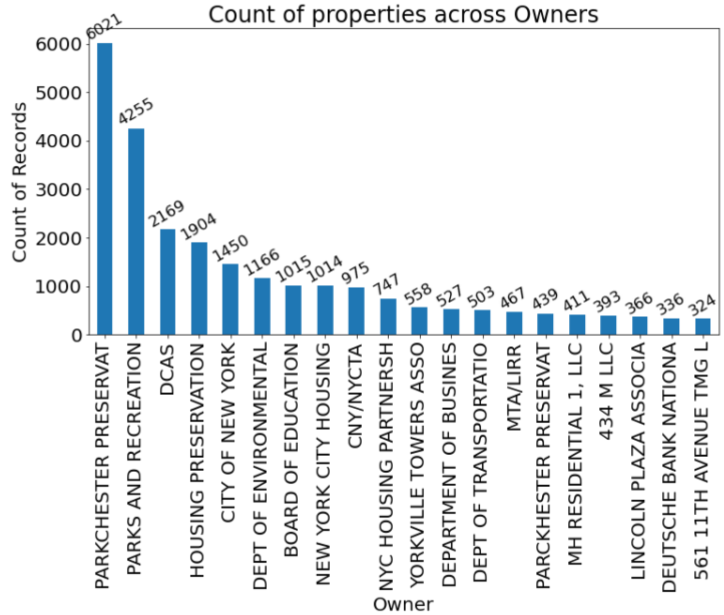


6) Field Name: Easement
Description: Easement means a right to cross or otherwise use someone else's land for a specified purpose. The most common one is land easement.

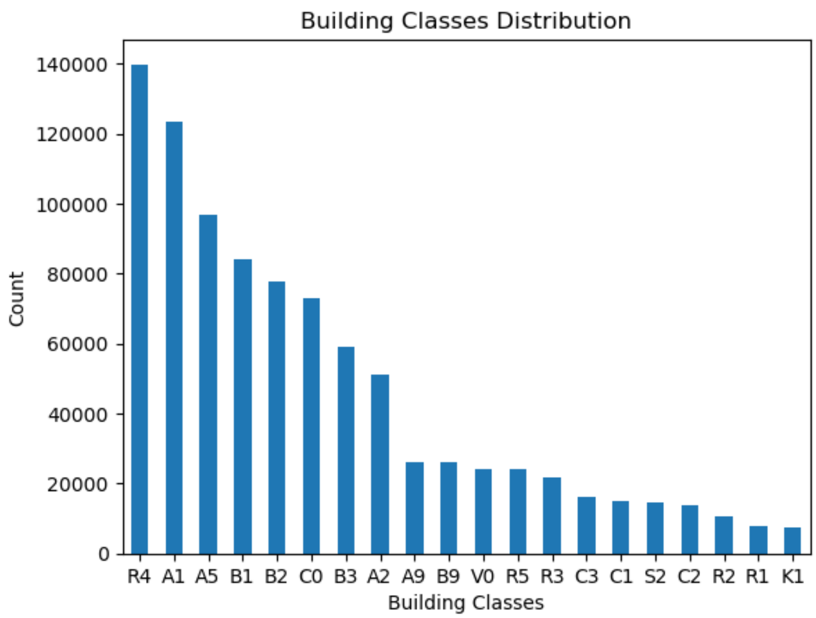


7) Field Name: OWNER
Description: the name of the property owner.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

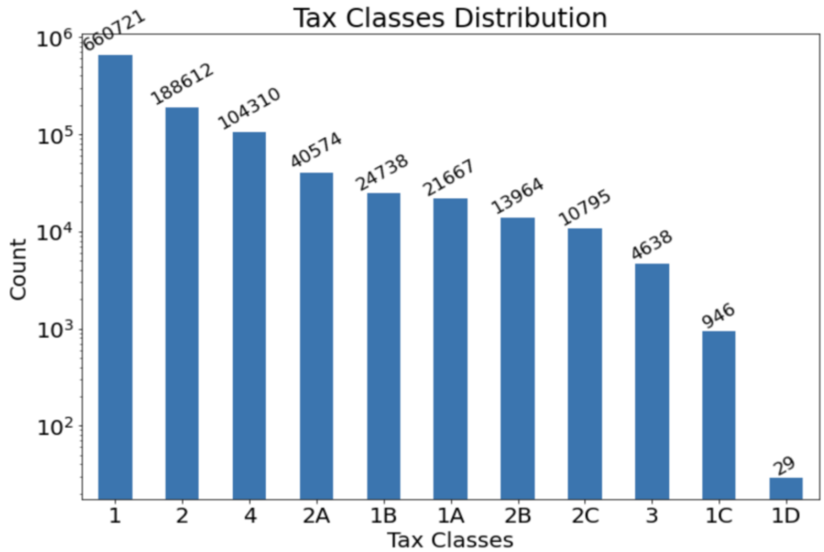


8) Field name: BLDGCL
Description: Building class. There are 200 building classes and the most common ones are R4, A1, A5.

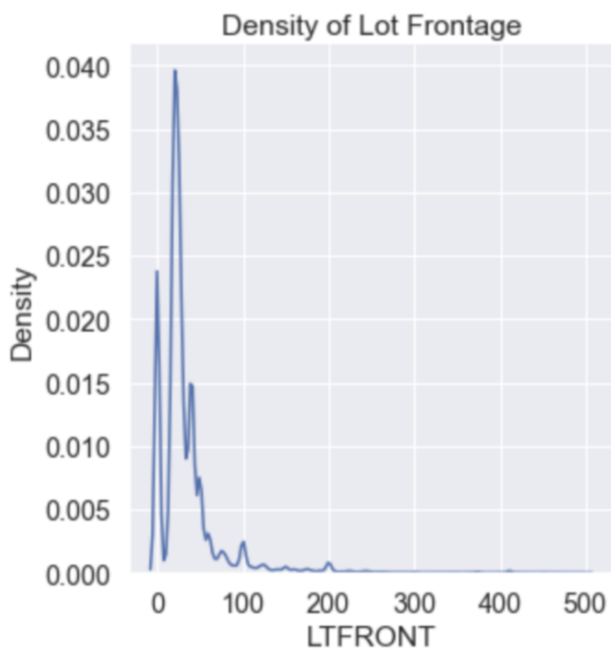


9) Field name: TAXCLASS
Description: Property tax class. There are total 11 classes of property tax. The most common class is 1, appearing more than 660,000 times.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

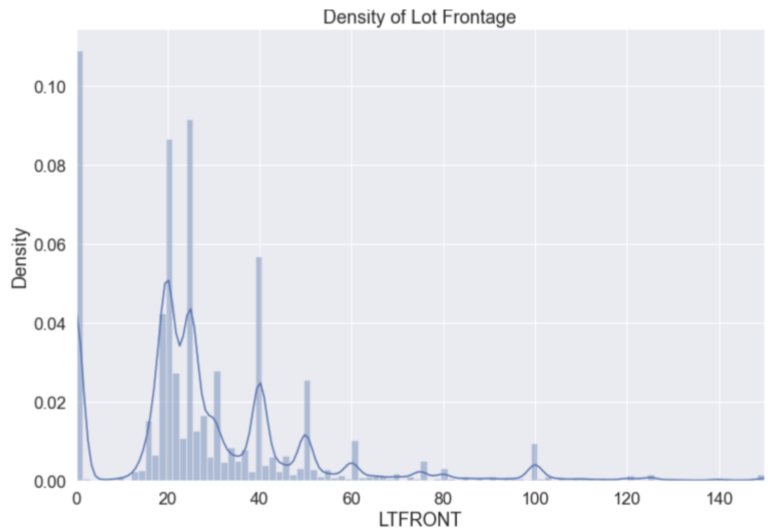


10) Field Name: LTFRONT
Description: The frontage (width) of the lot which is often measured as the average horizontal distance of property lines of a lot. Below is the distribution of density of lot frontage which the width of the lot is under 500 ft.

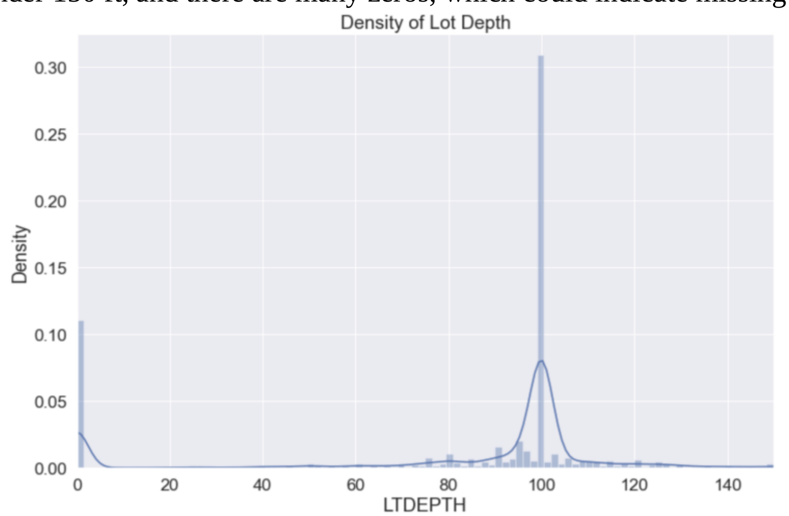


Another distribution of density of lot frontage which the width of lot is under 150 ft. A large number of sizes that are zero means they might be a lot of missing values.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

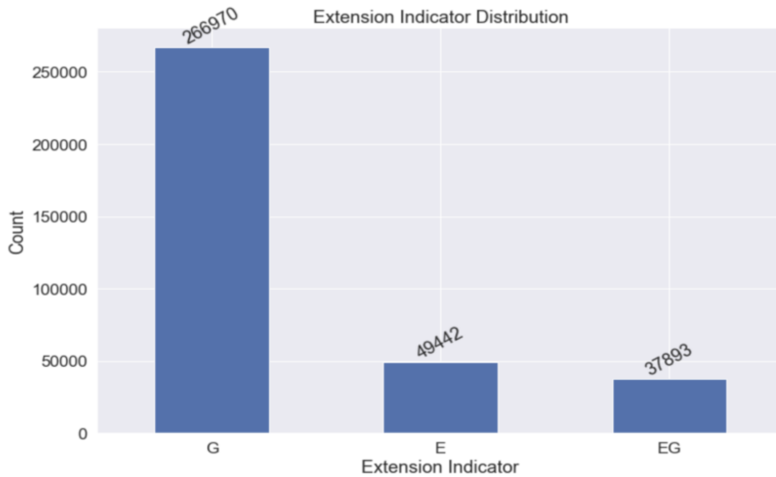


11) Field Name: LTDEPTH
Description: The depth of the lot. The distribution below shows the density of lot depth. Most of the value are under 150 ft, and there are many zeros, which could indicate missing values.

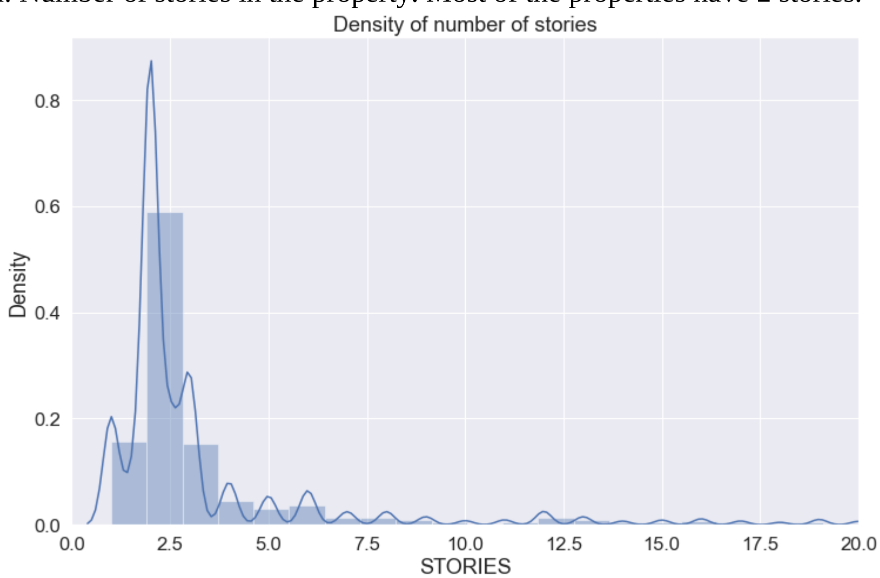


12) Field Name: EXT
Description: The extension Indicator of the property. There are 3 different extension indicators. The most common type is "**G**" with 26,6970 counts.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

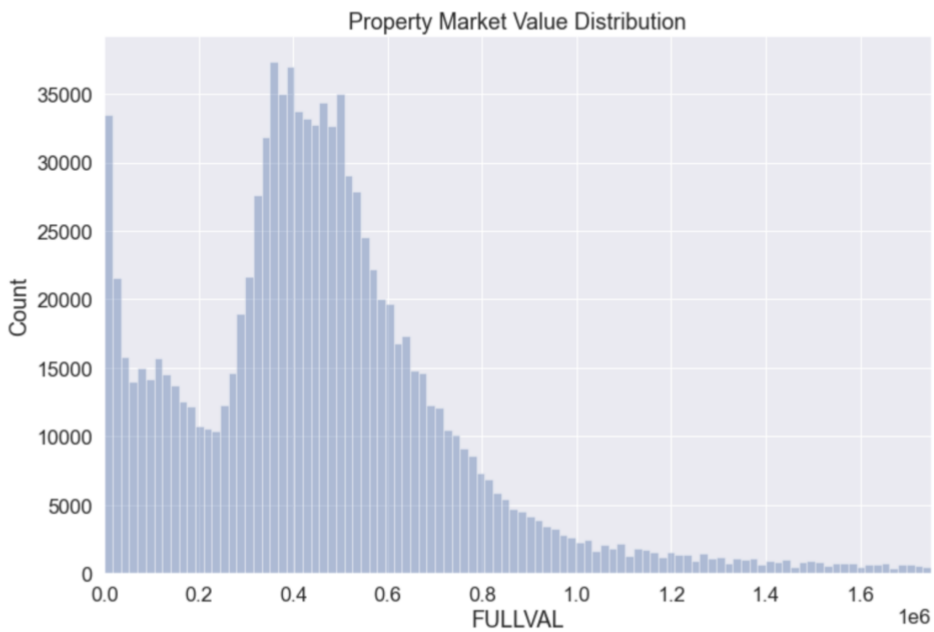


13) Field Name: STORIES
Description: Number of stories in the property. Most of the properties have 2 stories.



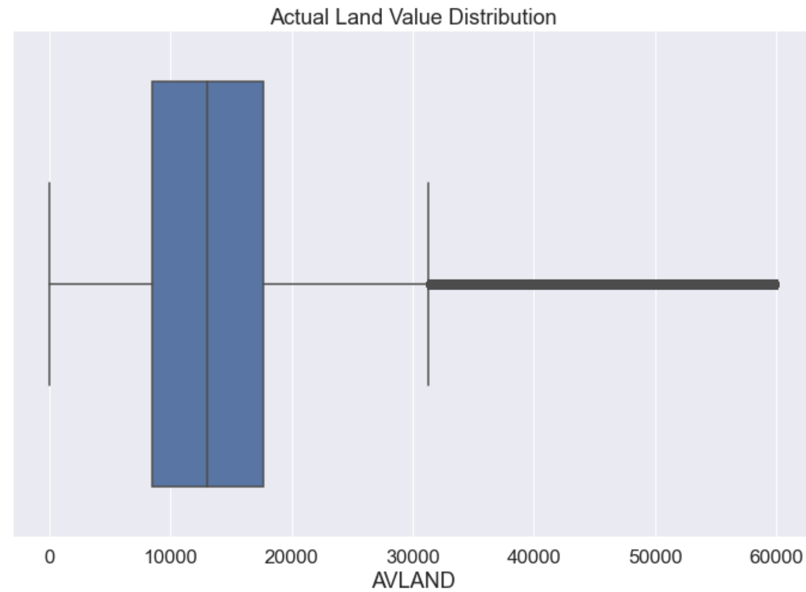
14) Field Name: FULLVAL
Description: The market value of the property. The distribution below shows the market value of each property. The most common market value ranges from 200,000 to 600,000 dollars.

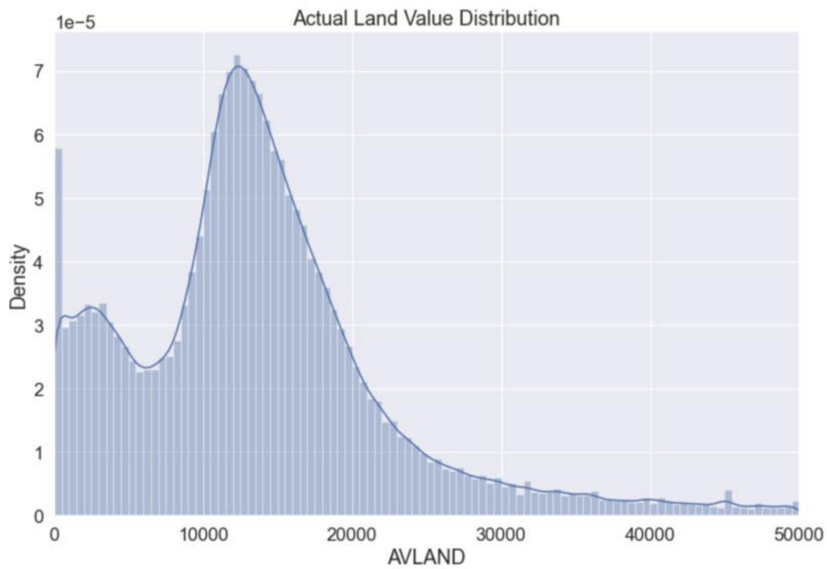
Name: Dimitra Charalampopoulou
Date: May 9th, 2023



15) Field Name: AVLAND
Description: The actual land value of the property. The distribution below shows the actual land value of each property. The most common value ranges from 8,000 to 20,000 dollars.

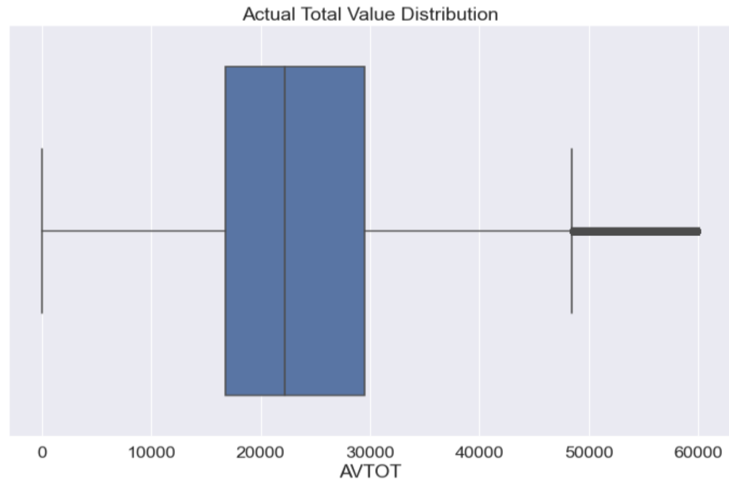
Name: Dimitra Charalampopoulou
Date: May 9th, 2023

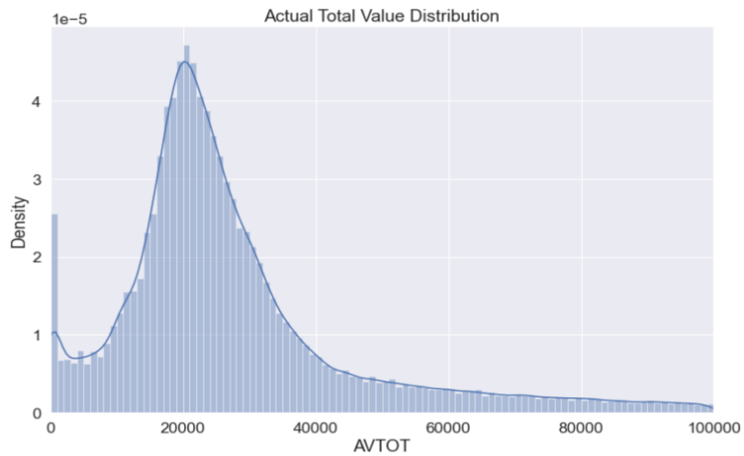




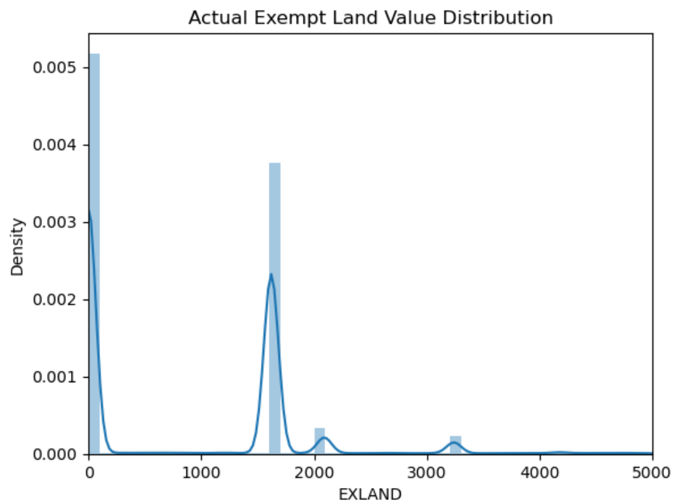
16) Field Name: AVTOT
Description: The actual total value of the property. The distribution below shows the actual total value of each property, and the common value ranges from 17,000 to 30,000 dollars.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023



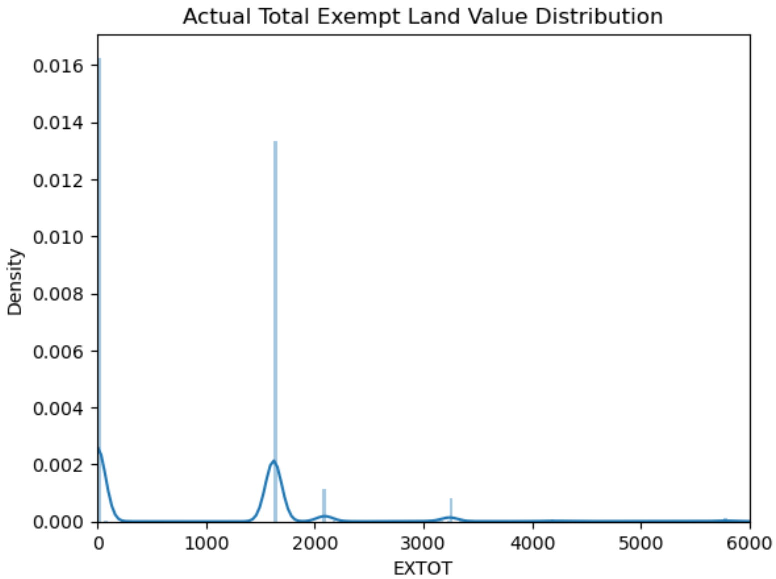


17) Field Name: EXLAND
Description: The actual exempt land value of the property. And most values are 0

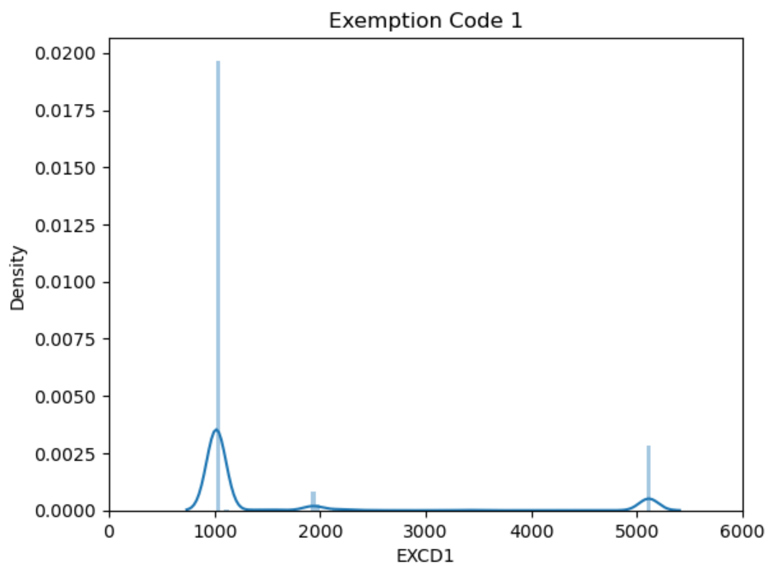


18) Field Name: EXTOT
Description: The actual total exempt land value of the property.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

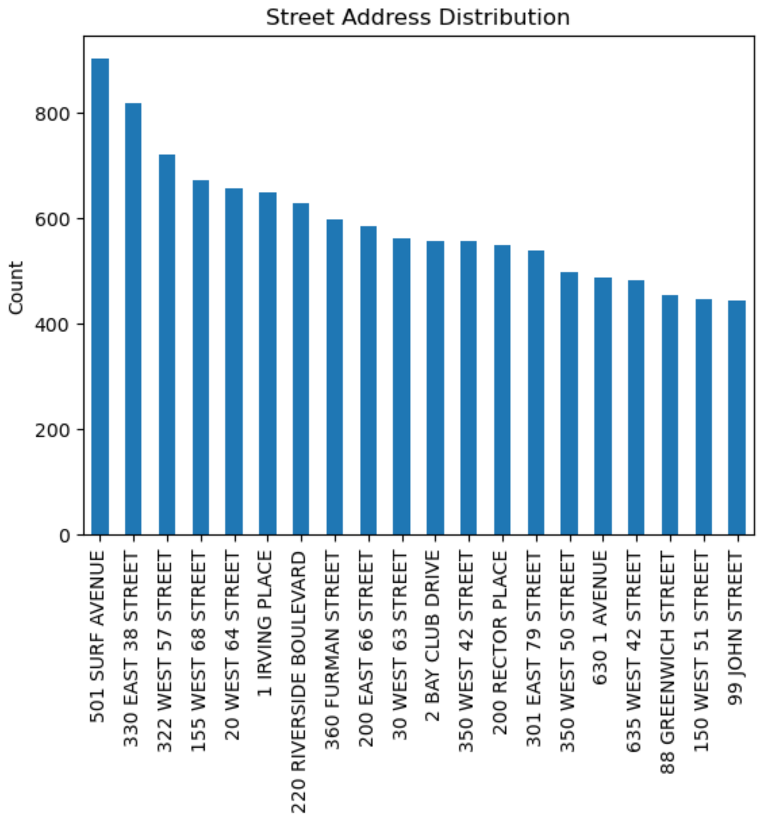


19) Field Name: EXCD1
Description: the exemption code 1 of the property.

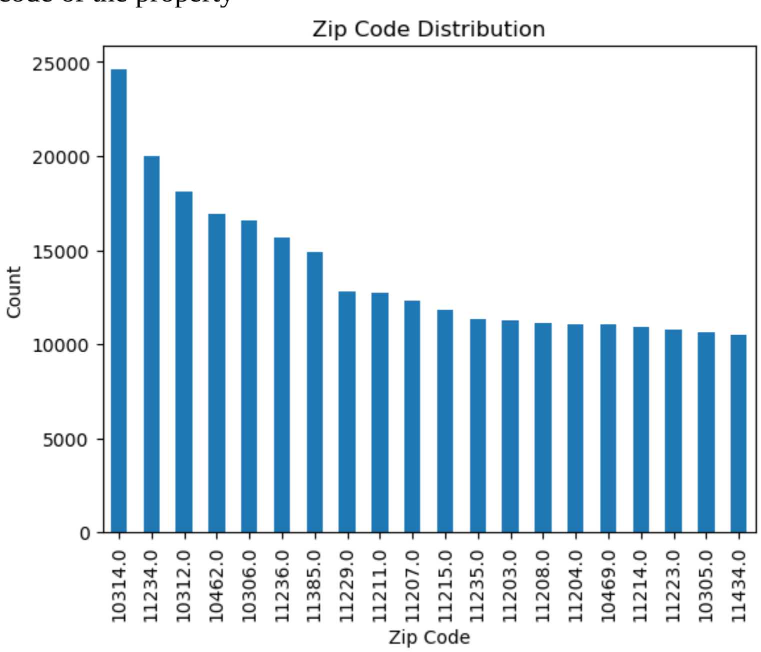


20) Field Name: STADDR
Description: the street address of the property. Surf avenue, 38th, and 37th street have most properties.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

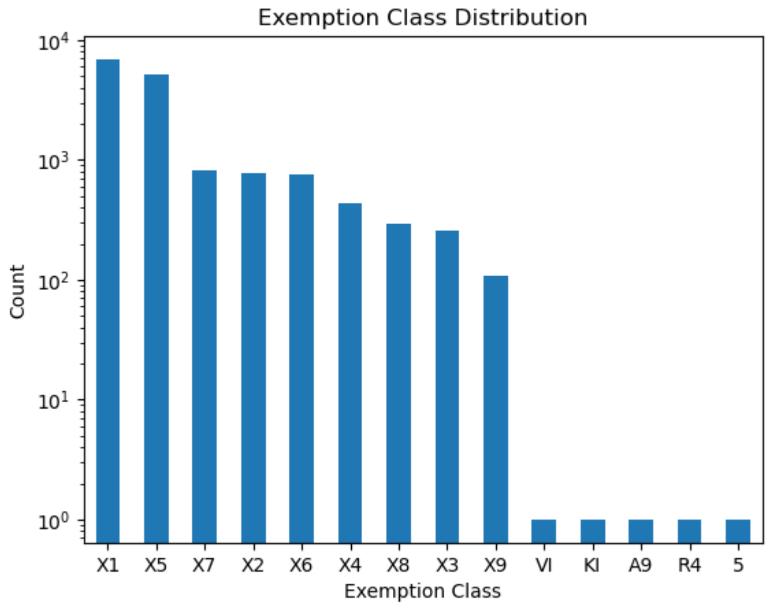


21) Field Name: Zip
Description: Zipcode of the property

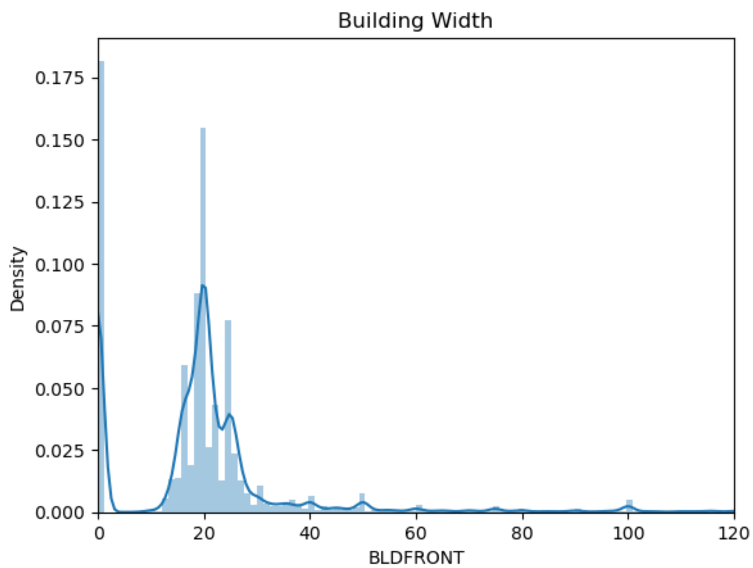


22) Field Name: EXMPTCL
Description: The exemption class of the property. The most common ones are X1 and X5.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

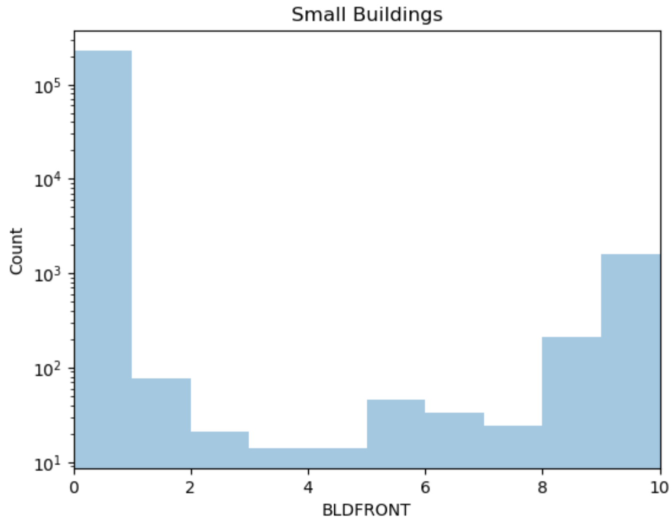


23) Field Name: BLDFRONT
Description: The width of the property.

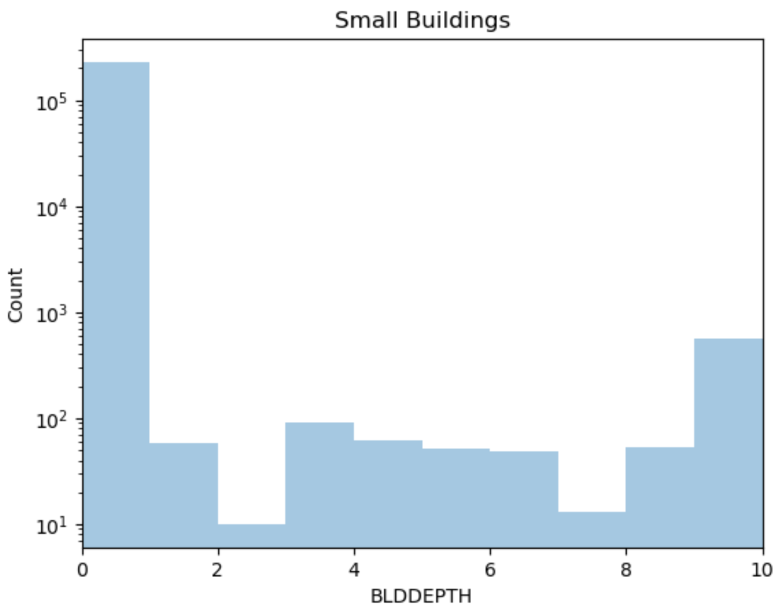


Looking at small buildings including 0:

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

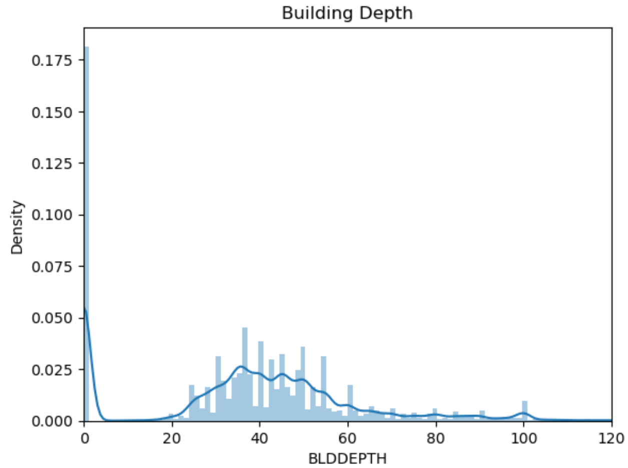


24) Field Name: BLDDEPTH
Description: The depth of the property.

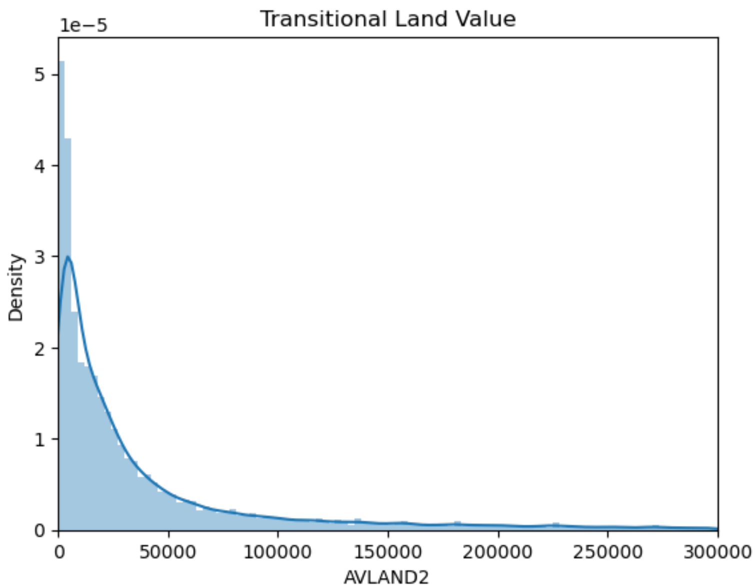


Looking at small buildings including 0:

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

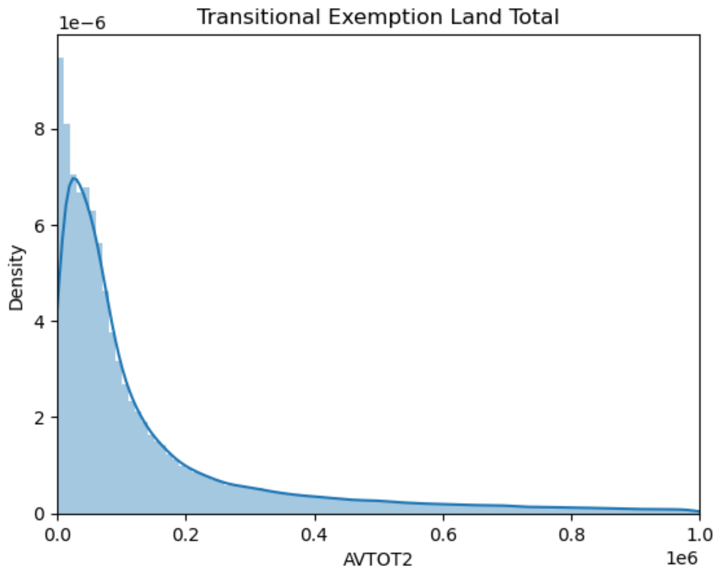


25) Field Name: AVLAND2
Description: The transitional land value of the property.

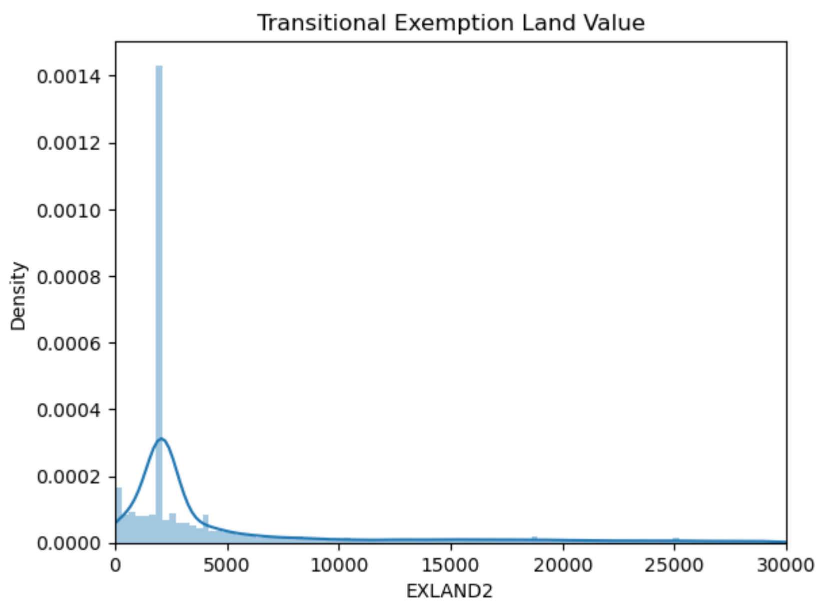


26) Field Name: AVTLOT2
Description: The total transitional exemption land of the property.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

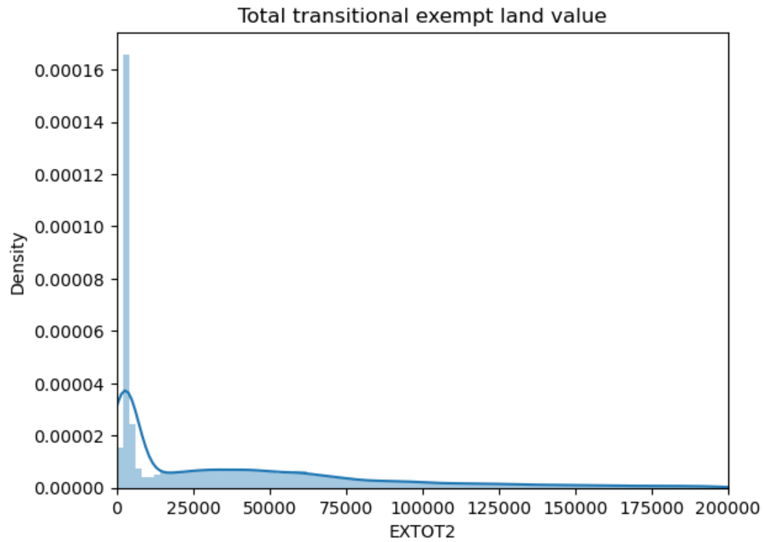


27) Filed Name: EXLAND2
Description: Transitional exempt land value of the property.

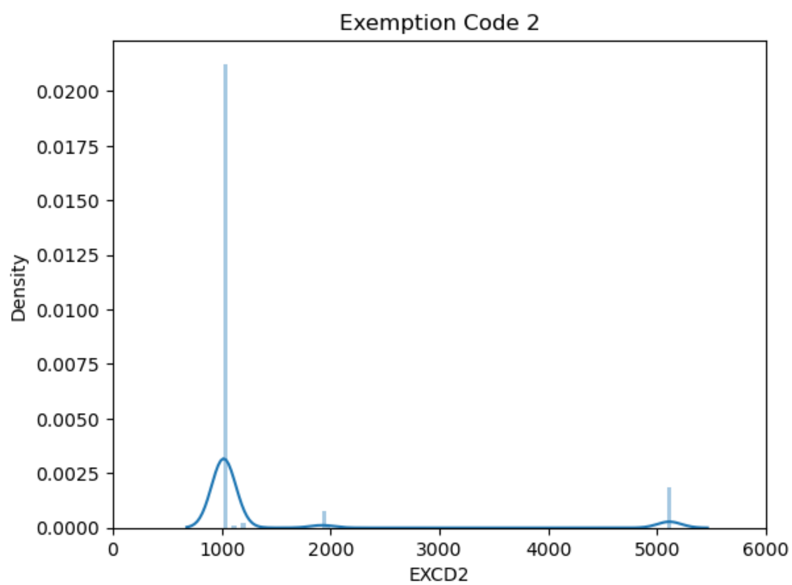


28) Filed Name: EXTOT2
Description: Total transitional exempt land value of the property.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023

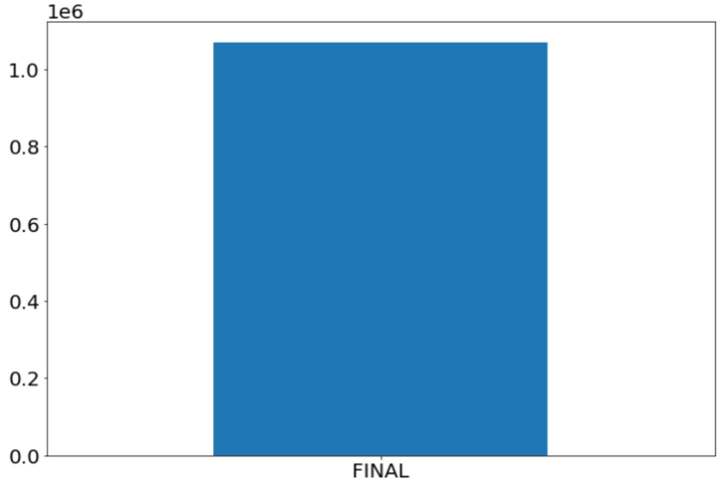


29) Field Name: EXCD2
Description: the exemption code 2 of the property.



30) Field Name: Period
Description: Assessment period when data was created. In this dataset, all property has a period of 'Final'.

Name: Dimitra Charalampopoulou
Date: May 9th, 2023



31) Field Name: YEAR
Description: the year where the information is taken. In this dataset, all properties have the value of ‘2010/11’.

32) File Name: VALTYPE
Description: Potentially mean value type. All properties have VALTYPE of ‘AC-TR’.