Property Tax Frauds Project

1. Executive Summary

Aiming at detecting property tax fraud, where individuals intentionally misrepresent their property characteristics to underpay taxes, we examined **1,070,994 records** and **32 fields** of **New York property valuation and assessment information** provided from the city of New York. We developed two models with **58 variables** to identify **anomalies** in the relationship between property characteristics and values. Through two iterations, our final model successfully identified **50 properties with unusual characteristics**, such as incorrect number of stories, building sizes larger than lot sizes, market values significantly higher or lower than expected for the same tax class or location, and high market values compared to actual values.

2. Data Description

- **Overview of Data**

  The data is a collection of New York property valuation and assessment information provided from the city of New York to detect property tax frauds. The data from 2010 covers the properties information with total **1,070,994 records and 32 fields**. Since there is **no fraud label** provided, we will build an unsupervised model to detect unusualness.

- **Observations of Data**

  The followings are summary of statistics for numeric and categorical fields. We can observe that there are **null values** and **high frequency of 0 or 1 values** in numeric fields such as STORIES and FULLVAL. We also noticed some outliers, including a FULLVAL of $6,150,000,000. We fixed these values in the data cleaning process.

  - **Numeric Table**

    Please note that we round all results into two decimal places.

Field Name	% Populated	Min	Max	Mean	Stdev	% Zero
LTFRONT	100.00	0	9,999	36.64	74.03	15.79
LTDEPTH	100.00	0	9,999	88.86	76.40	15.89
STORIES	94.75	1	119	5.01	8.37	0.00
FULLVAL	100.00	0	6,150,000,000	874,264.51	11,582,430.99	1.21
AVLAND	100.00	0	2,668,500,000	85,067.92	4,057,260.06	1.21
AVTOT	100.00	0	4,668,308,947	227,238.17	6,877,529.31	1.21
EXLAND	100.00	0	2,668,500,000	36,423.89	3,981,575.79	45.91
EXTOT	100.00	0	4,668,308,947	91,186.98	6,508,402.82	40.39
BLDFRONT	100.00	0	7,575	23.04	35.58	21.36
BLDDEPTH	100.00	0	9,393	39.92	42.71	21.37
AVLAND2	26.40	3	2,371,005,000	246,235.72	6,178,962.56	0.00
AVTOT2	26.40	3	4,501,180,002	713,911.44	11,652,528.95	0.00
EXLAND2	8.17	1	2,371,005,000	351,235.68	10,802,212.67	0.00
EXTOT2	12.22	7	4,501,180,002	656,768.28	16,072,510.17	0.00

- **Categorical Table**

Please note that we exclude null values when counting # unique values.

Field Name	% Populated	# Unique Values	Most Common
RECORD	100.00	1,070,994	N/A
BBLE	100.00	1,070,994	N/A
BORO	100.00	5	4
BLOCK	100.00	13,984	3944
LOT	100.00	6,366	1
EASEMENT	0.43	12	E
OWNER	97.04	863,347	PARKCHESTER PRESERVAT
BLDGCL	100.00	200	R4
TAXCLASS	100.00	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	100.00	1	FINAL
YEAR	100.00	1	2010/11
VALTYPE	100.00	1	AC-TR

3. Data Cleaning

- **Data Exclusions**
  - **Motivation:** Since our clients were only interested in identifying tax fraud committed by private owners, we **filtered out records of government-owned properties**.

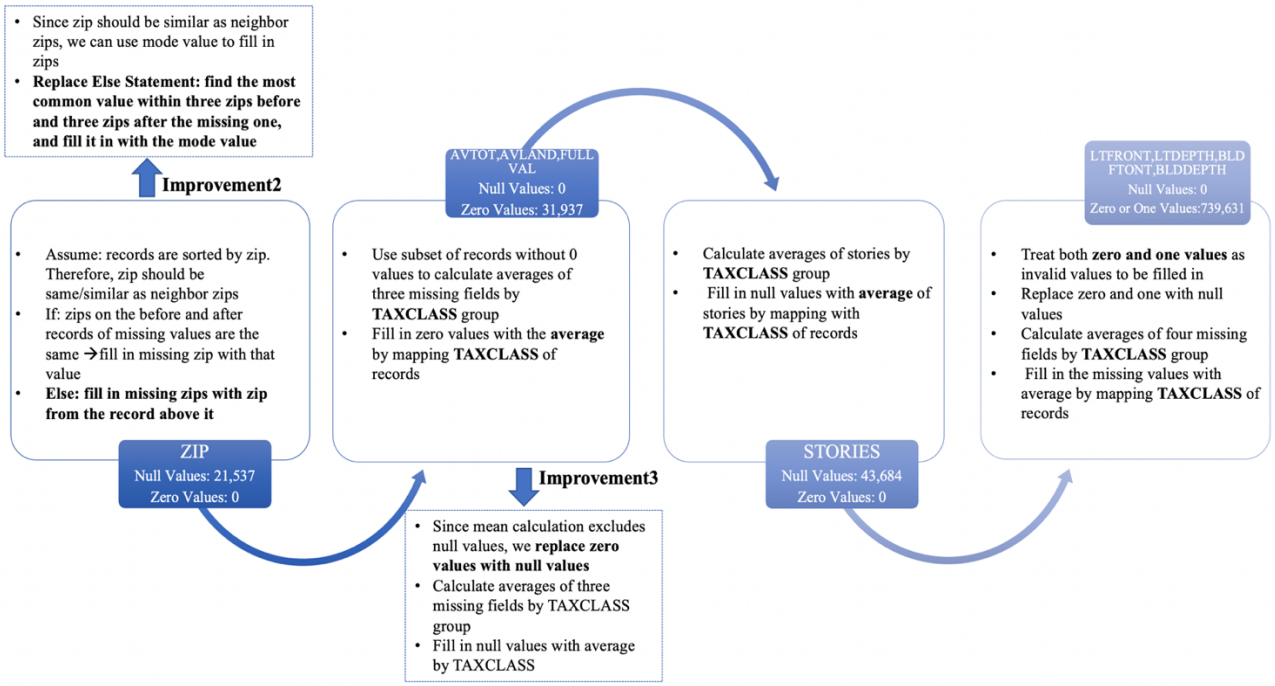
  - **Description of Steps:** Created "remove list" based on most common owners →Manually classified private and government owners on the list and removed private owners→Iterated the above two steps several times to complete the "remove list"(government list) →Excluded records with owner names on the "remove list".

- **Missing Field Values Imputation**
  - **Motivation:** We aimed to detect anomalies between property characteristics and values by utilizing **9 relevant fields out of 32 available** for data imputation, namely: **FULLVAL, AVLAND, AVTOT, ZIP, STORIES, LTFRONT, LTDEPTH, BLDFRONT, and BLDDEPTH**. To avoid triggering unusualness alarms, we filled in missing values with **the most typical values** for these fields.

  - **Missing Values:** We treat both **null and zero values as missing**/invalid values. Especially for lot and building size fields, values equal to 1 will were also treated as invalid values to be filled in.

- **Data Imputation Logic (Two Improvements Included)**

The imputation logic is described below.



4. Variable Creation

- **Identity Fraud Modes/Motivation of Variables**

Some people underpay property taxes by intentionally misrepresenting their property characteristics. Therefore, we focused on **anomalies/unusualness** in the relationship between **property values and property sizes**.

- **Fields Created by Calculating Size:**

Three temporary fields are created to calculate property size.

  - Lot area = ltsize= LTFRONT * LTDEPTH
  - Building area = bldsize = BLDFRONT* BLDDEPTH
  - Building volume = bldvol = bldsize * STORIES

- **Description of Variables**

We generated **58** independent variables to detect **outliers** in the relationship between property values and characteristics. There are **three main kinds of variables**: the ratio of price to size, the inverse of ratio, and the comparison with the average ratio for each property's **ZIP or TAXCLASS**. We also introduced **three new variables** to compare actual and market values, as well as building and lot sizes, to identify anomalies in value and size. Please see the table below for more information on each category of independent variables.

- **Table of Variables**

Family of Variables	Description of Variables/Logic	How to Detect Anomalies from Variables		# Variables
		Normal	**Fraud Signal**	
Price Per Square Foot Variables	Relationship between property value and property size, representing by ratios of **$/size**. Property Value Fields include FULLVAL, AVLAND, and AVTOT while Property Size Fields include ltsize, bldsize, bldvol.	Property value is changing in similar proportion to property size: The larger size, the higher property value.	Outliers from distribution with **unusually large $/size ratio** (eg. high value but small size).	9
Inverse of Price Per Square Foot Variables	Inverse Relationship between property value and property size, representing by ratios of **1/(Price Per Square Foot Variables + 0.01)**.	Property value is changing in similar proportion to property size: The larger size, the higher property value.	Outliers from distribution with **unusually small $/size ratio** (eg. low value but large size).	9
Relative Ratio Variables	Compare the above 18 ratios of each record with average ratios for its **TAXCLASS/ZIP** group. Relative Ratio = Price Per Square Foot Variables or Inverse Variables / Average ratios by TAXCLASS or ZIP.	18 ratios should be similar to average ratios in the same TAXCLASS/ZIP group. Therefore, the **usually relative ratios should be close to 1**.	**Relative ratios that are far from 1** shows that property's value is influenced by its size in a different manner that differs from its geographical or logical neighbors.	36
New Variable: Value Ratio Variable	Compare market value (FULLVAL) with actual value (AVLAND + AVTOT). **Improvement4: Add another relationship between three values. Compare AVTOT with FULLVAL + AVLAND.**	Market value is changing in similar proportion to actual value: The higher actual value, the higher market value.	After normalizing and taking inverse of value ratios, We can identify **unusually large and small value ratios**.	1 (Existing) + 1 (New)
New Variable: Value Change Ratio	Compare **FULLVAL/AVLAND** (value change ratio) for each record with **average FULLVAL/AVLAND for its ZIP** group. The FULLVAL/AVLAND shows that other factors beyond land values can also affect property market values such as city development.	Usually other factors beyond land values that affect market values are similar in same ZIP group. The **value change ratio/average ratio by ZIP should be close to 1**.	Relative ratios that are **far from 1.**	1
New Variable: Size Ratio	Size ratio = building size (bldsize)/lot size (ltsize). Compare size ratio with 1.	Usually, building size <= lot size. Therefore, the **size ratio <= 1**.	If someone misrepresents property size to underpay taxes, he might not follow the general rule that buidling size should be smaller than or equal to lot size. Therefore, if we identify any **size ratio** that **exceeds 1 limit and is much bigger,** there is a fraudulent signal.	1
	Total Independent Variables			**58**

5. Dimensionality Reduction

- **Motivation of PCA**

Since dimensionality is high, data becomes sparse quickly and all points become outliers, causing a **curse of dimensionality**. To overcome the issue for this **unsupervised problem**, we utilized Principal Component Analysis (PCA) to **reduce dimensions and eliminate linear correlations** between variables.

- **Description of PCA**
  - **Definition of PCA:** PCA is a technique that identifies the **dominant direction** in the data by analyzing the spread of data (variance) and **rotates the coordinate system** accordingly. By calculating the rotation matrix, the original data can be expressed in terms of this new coordinate system. PCA can also be used to reconstruct data.

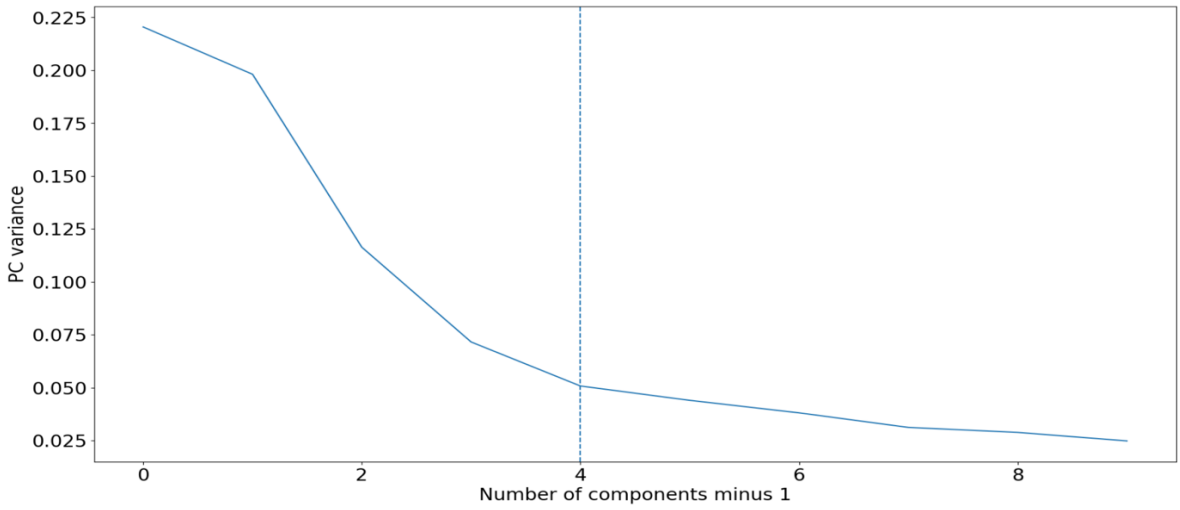
  - **Formula of PCA**
    We obtained eigenvectors for each element and combined them into a matrix called the rotation matrix E. Using this rotation matrix, we transformed the original data. The formula to calculate eigenvectors and transformed data is as follows.

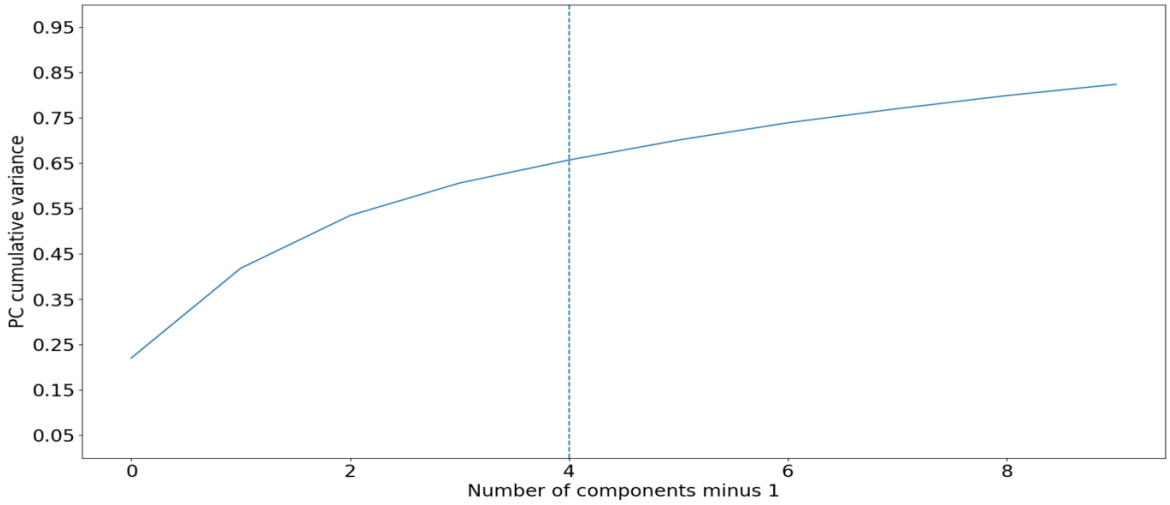
$$p_{il} = \sum_j x_{ij} e_j^l$$

$$\underset{m \text{ rows}}{P} = \underset{m \text{ rows}}{X} \underset{n \text{ rows}}{E} \quad (n \text{ columns})$$

- **Description of Variables Reduction and Preparation Process**
  - **Step 1 Z-Scale Variables**: Z-scaled variables to get rid of the highly squashed nature due to the substantially different variables ranges. Z-scaling is a normalization technique that involves subtracting the mean value of each variable and then dividing it by its standard deviation.

  - **Step 2 PCA**: After examining the natural break in the scree plots, we decided to keep **5 PCs** to account for 65%-70% of the total variance.





- **Step 3 Data Transformation:** Represented each record in the space of the PCs by using rotation matrix E.

- **Step 4 Z-Scale Again:** Z-scaled variables on new PC coordinates to ensure that all dimensions were equally important for the calculation of Minkowski distance. These z-scaled variables were commonly referred to as "z-scores," with a mean of 0 and a standard deviation of 1.

6. Anomaly Detection Algorithms

- **Score 1 Z Scores Outlier**

We applied the Mahalanobis-like distance concept to calculate the **Minkowski distance** with a **power of 2** (also known as **Euclidean distance**) to measure the **difference between z-scores and the origin 0** on each PC. This helped us identify outliers by determining the records with larger distances from the origin, indicating a greater degree of unusualness.

- Below is the formula to calculate z-score outlier using Minkowski distance:

$$s_i^1 = \left(\sum_k |PCz_k^i|^p \right)^{1/p}$$ **where p = 2**

- Another formula to Represent Euclidean Distance to Origin on each PC:

$$D = (x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2)^{1/2}$$

- **Score 2 Autoencoder**

An autoencoder is a process that functionally maps an original vector back to itself. The first step was to train a **neural net** model on the **entire dataset** and used it to predict the output from the original vector input. After that, we used **Minkowski distance with a power of 2** to calculate Score 2, which measures the **difference (error) between the original input vector and the model's output vector**. A record with higher error is considered more unusual.

- Below is the formula to calculate error using Minkowski distance: **where p = 2**

$$s_i^2 = \left(\sum_k |PCz_k'^i - PCz_k^i|^p \right)^{1/p}$$

Autoencoder output

- **Scores Scaling and Combination**

To obtain a final score, we utilized **Rank Order Scaling** by replacing Score 1 and Score 2 with the record's rank order after sorting each score. The two score ranks on the same scale were then **combined using equal weights of 50/50**. The resulting final score was appended back to the records, and the records were sorted in descending order based on the final score for further investigation.

7. Results

- **Use Final Scores and Z-scaled Variables**

We combined scores from two algorithm models to obtain final scores, and then **sorted** the records in **descending order** based on **final scores**. Following this, we generated a **heatmap** of the **z-scaled variables** (z-scores) for the top 200 records, where the darker color represents how much the variable deviates from the mean in terms of standard deviation. By **observing darker regions** in the heatmap, we can detect unusual variable values and investigate more on the records containing those unusual values.

- **Records Examination and Investigation**

Upon examining records with unusual variables, we **reviewed the original fields**, such as size, value, tax class, and address, to **verify the information and identify any anomalies** in the records. To ensure the accuracy of the information, we referred to images of the properties available online, specifically to examine the number of stories. Additionally, we searched real-estate marketplace websites like Zillow or Redfin to access public information on properties, which helped us verify size and market value information. We also examined any anomalies in the records, such as properties with buildings sizes exceeding lot sizes or properties with unusually low or high market values in the same tax class or location. We **provided a list** of records with inconsistencies or anomalies **to domain experts** for further investigation.

- **Seven Examples of Unusual Properties**

We presented seven examples of different cases, including properties with incorrect number of stories, buildings sizes exceeding lot sizes, unusually low or high market values in the same tax class or location, and unusually high market values compared to actual value.

a. Case 1 Incorrect Number of Stories: Record Number 33104

- **Potential Fraud: Incorrect number of stories - Lower number of stories**
- **Information**

Find the Unusualness					
Record: 33104	**r3**	**r6**	**r9**		
z-scores	170.87	216.09	115.96		
Property Original Information					
RECORD	33104	**FULLVAL**	$450,549.00	**LTFRONT**	160
OWNER	WINSTON, MICHAEL EVAN	**AVLAND**	$ 59,821.00	**LTDEPTH**	278
ADDRESS	400 WEST 12 STREET	**AVTOT**	$202,747.00	**BLDFRONT**	90
ZIP	10014			**BLDDEPTH**	160
TAXCLASS	2			**STORIES**	15

- **Property Picture**

Description of Superior Ink at 400 West 12th Street

Superior Ink is a 17-story, LEED-certified, Robert A.M. Stern-designed condominium with interiors by Yabu Pushelberg. A 24-hour concierge, doorman, and live-in superintendent are on staff, and residential amenities include in-building valet parking, sunny fitness center and yoga/Pilates room, a residents' lounge, a screening room, a children's playroom, and a bike room. The building's address at 400 West 12th Street puts it close to Hudson River Park, the Meatpacking District, the Whitney Museum, and popular shopping, dining, and nightlife.

Building Facts

Year Built:	2010	Total Apartments:	62
Building Type:	Condo	Total Floors:	17
Neighborhood:	West Village (Manhattan)	Doorman:	FT Doorman
Minimum Down:	20%	Pets:	Allowed
		Total Floorplans:	40

- **Investigation**

For record 33104, all variables related to r3, r6, and r9, including those grouped by tax class and zip5, are high. We found that r3, r6, and r9 contain the same value **S3 (S2*STORIES)**, and there is no unusualness in S2, since r2, r5, and r8 are normal. This high value is likely due to an incorrect number of stories. To verify this, we checked an online image of the building, which revealed that it is tall with several floors. We also looked up the property's address and discovered that it is a condo built in 2010 with a total of **17 floors**, which contradicts the information provided in the record, stating that the building has only **15 floors**. Therefore, we need to conduct further investigations into the **actual number of stories** for this property.

b. Case 2 Incorrect Number of Stories: Record Number 14875

- **Potential Fraud: Incorrect number of stories - Lower number of stories**
- **Information**

Find the Unusualness						
Record: 14875	**r3**	**r6**	**r9**	**r3_zip5**	**r6_zip5**	**r9_zip5**
z-scores	114.24	144.22	77.63	189.54	131.12	102.86

Property Original Information						
RECORD	14875	**FULLVAL**	$1,150,000.00	**LTFRONT**	43	
OWNER	PONTE EQUITIES INC	**AVLAND**	$ 354,592.00	**LTDEPTH**	82	
ADDRESS	440 WASHINGTON STREET	**AVTOT**	$ 445,662.00	**BLDFRONT**	22	
ZIP	10013			**BLDDEPTH**	82	
TAXCLASS	2B			**STORIES**	6	

- **Property Picture**

Building Facts

Facts	49 units 11 stories
District	Community District 101
	Police Precinct 1
Hurricane Evacuation Zone	Zone 1
Floor Plans	65 floor plans available
Documents and Permits	13 documents and perm
Rentals Listings	1 active rental ($7,800 av
	71 previous rentals ($7,7
Architect	OCV Architects
Developer	Ponte Equities
Interiors	Dimension NY
Leasing and Marketing	Douglas Elliman Develop
Manager	Ponte Equities

- **Investigation**

For record 14875, all variables related to r3, r6, and r9, including those grouped by tax class are high. We found that r3, r6, and r9 contain the same value S3 (S2*STORIES), and there is no unusualness in S2, since r2, r5, and r8 are normal. This high value is likely due to an incorrect number of stories. To verify this, we checked an online **image** of the building, which revealed that it has **more than 6 floors** as reported. We also looked up the property's address and discovered that it is a residential apartment with a total of **11 floors**, which contradicts the information provided in the record, stating that the building has only **6 floors**. Therefore, we need to conduct further investigations into the actual number of stories for this property.

c. Case 3 Building Size Exceeds Lot Size: Record Number 41415

- **Potential Fraud: Wrong Building Size – Too big**
- **Information**

Find the Unusualness					
Record: 41415	**r3**	**r6**	**r3_zip5**	**r6_zip5**	**size_ratio**
z-scores	76.34	96.78	152.02	118.83	-0.72
Property Original Information					
RECORD	41415	**FULLVAL**	$12,190,000.00	**LTFRONT**	86
OWNER	CHURCH ST FRANCIS XAV	**AVLAND**	$ 3,235,500.00	**LTDEPTH**	108
ADDRESS	46 WEST 16 STREET	**AVTOT**	$ 5,485,500.00	**BLDFRONT**	87
ZIP	10011			**BLDDEPTH**	184
TAXCLASS	4			**STORIES**	1

- **Property Picture**

- **Investigation**

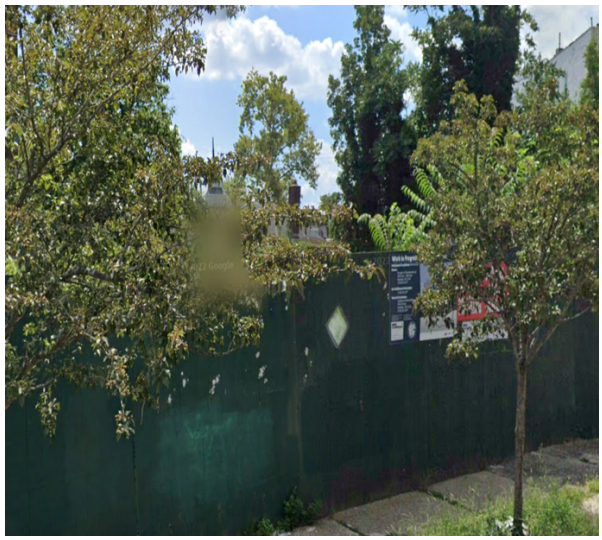
Record 41415 shows high values for all variables related to r3 and r6, including those grouped by tax class and zip5. Upon further investigation, we found that r3 and r6 contain the same value S3 (**S2***STORIES). To confirm this, we checked an online image of the property and found that it is a church with only one story, which matches the story number provided in the record. However, we also calculated the **size ratio**, which is 1-(Building Size/Lot Size), and found that it is negative, indicating an anomaly. Specifically, the **BLDDEPTH value of 184 is much larger than the LTDEPTH value of 108, suggesting that the building size is too big for the lot size.** Therefore, we need to conduct further investigations into the building dimensions for this property.

d. Case 4 Building Size Exceeds Lot Size: Record Number 952083

- **Potential Fraud: Wrong Building Size and Wrong Stories**

Find the Unusualness					
Record: 952083	**size_ratio**				
z-scores	-0.23				
Property Original Information					
RECORD	952083	**FULLVAL**	$ 300,000.00	**LTFRONT**	36
OWNER	CHINA, YU P	**AVLAND**	$ 9,660.00	**LTDEPTH**	100
ADDRESS	88 PROSPECT STREET	**AVTOT**	$ 18,000.00	**BLDFRONT**	44
ZIP	10304			**BLDDEPTH**	100
TAXCLASS	1			**STORIES**	2

- **Property Picture**



Facts and features

Type: VacantLand

Heating: No Data

Year built: 1899

Cooling: No Data

Parking: 0 spaces

Community and Neighborhood Details

Location

Region: Staten Island

- **Investigation**

Record 952083 indicates a negative value for the size ratio, indicating the building size exceeds the lot size. The **BLDFRONT value of 44 is larger than the LTDFRONT value of 36**, which is unusual. However, upon checking an online image of the property, we found it to be **a vacant land without any building or stories**. Further online research revealed that the land has been vacant since 1899. The values of BLDFRONT and STORIES, which are both 44 and 2, respectively, require further investigation by domain experts.

e. Case 5 Lower Market Value in Same Location: Record 827641

- **Potential Fraud: FULLVAL too low** compared to others in the same zip code area and same tax class.

- **Information**

Find the Unusualness						
Record: 827641	**r1inv_zip5**	**r2inv_zip5**	**r3inv_zip5**	**r4inv_zip5**	**r5inv_zip5**	**r7inv_zip5**
z-scores	185.56	224.32	198.83	66.43	57.97	73.69

Property Original Information					
RECORD	827641	**FULLVAL**	$357,000.00	**LTFRONT**	25
OWNER	DERAL HOME LOAN M	**AVLAND**	$ 6,511.00	**LTDEPTH**	97
ADDRESS	145-05 LAKEWOOD AVENUE	**AVTOT**	$ 11,859.00	**BLDFRONT**	18
ZIP	11435			**BLDDEPTH**	28
TAXCLASS	1			**STORIES**	2

Property Information for Same ZIP CODE Same TAX CLASS											
RECORD	**OWNER**	**TAXCLASS**	**LTFRONT**	**LTDEPTH**	**STORIES**	**FULLVAL**	**AVLAND**	**AVTOT**	**ZIP**	**BLDFRONT**	**BLDDEPTH**
827641	DERAL HOME LOAN M	**1**	25	97	2	357000	6511	11859	**11435**	18	28
825698	TARACHAND DEOLALI	**1**	25	100	2	410000	12517	22809	**11435**	20	50
817151	O'HARA, LUCY	**1**	25	100	2	545000	13971	25380	**11435**	20	66
874644	CHASE, NEAL	**1**	23	100	2.5	427000	7737	12855	**11435**	16	36
874385	MARINO A. VARGAS	**1**	30	92	2.7	477000	11131	18895	**11435**	22	40

- **Property Picture**

- **Investigation**

For record 827641, the r1inv_zip5, r2inv_zip5, r3inv_zip5 are high, indicating r1_zip5, r2_zip5, r3_zip5 is extremely small. This shows that the **$/size ratio of this property is much smaller than average ratio in this area** (same zip code). When comparing size properties (LTFRONT, LTDEPTH, BLDFRONT, BLDFRONT, STORIES) to the average properties with the same zip code, we don't find any unnormal values. (LTFRONT>BLDFRONT, LTDEPTH>BLDDEPTH) Also the number of stories 2 matches with the stories on the picture. Therefore, the size seems normal. The strange value is V1, **FULLVAL, which is much lower than the average FULLVAL in this area**. To conclude, we need to investigate more about the FULLVAL of this property.

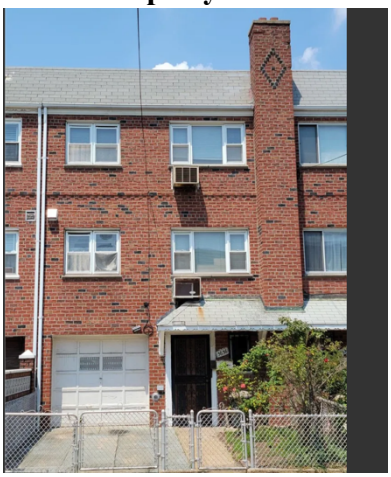
f. Case 6 Lower Market Value in Same Location: Record 583853

- **Potential Fraud: FULLVAL too low** compared to others in the same zip code area and same tax class.

- **Information**

Find the Unusualness											
Record: 583853	r3inv_zip5	r6inv_zip5	r9inv_zip5								
z-scores	186.16	61.64	100.70								
Property Original Information											
RECORD	583853	FULLVAL	$ 701,000.00	LTFRONT	20						
OWNER	KONG YEE LEE	AVLAND	$ 17,323.00	LTDEPTH	100						
ADDRESS	36-30 24 STREET	AVTOT	$ 31,137.00	BLDFRONT	20						
ZIP	11106			BLDDEPTH	48						
TAXCLASS	1			STORIES	3						
Property Information for Same ZIP CODE Same TAX CLASS											
RECORD	OWNER	TAXCLASS	ZIP	LTFRONT	LTDEPTH	STORIES	BLDFRONT	BLDDEPTH	Building Size	FULLVAL	
583853	KONG YEE LEE	1	11106	20	100	3	20	48	2880	701000	
589195	MICHAEL ARAPIS	1	11106	20	100	2	19	55	2090	706000	

- **Property Picture**



Building Information for 36-36 24th Street

Stories	3	Matches with Record	Residences		2
Year Built	1960		Building Size		20' x 48'
Building Sq. Ft	2,880 SF	Matches with Record	Lot Size	Matches with Record	2,017 SF / 20' x 101'
Building Age	-		Building Type		-

- **Investigation**

Record 583853 exhibits high values for r3inv_zip5, r6inv_zip5, and r9inv_zip5, indicating a significantly lower $/size ratio than the average ratio in the same zip code. The property's online address search revealed it to be a townhouse with a lot size of 2000, building size of 2880, and two floors, consistent with size information on the record. Therefore, size is normal. However, a comparison of its FULLVAL to other properties in the same tax class and area showed that its **FULLVAL/Building Size is much lower than average**. As we can see from the table, despite having a **similar FULLVAL** of around $700,000, the property has a significantly larger building size (**800 sqrt ft bigger** than others). Therefore, further investigation into the **FULLVAL** of this property is warranted

g. Case 7 FULLVAL Too High: Record 50917

- **Potential Fraud: FULLVAL Too High.**
- **Information**

Find the Unusualness					
Record: 50917	**Value_Ratio**				
z-scores	16.78				
Property Original Information					
RECORD	50917	**FULLVAL**	$ 5,090,000.00	**LTFRONT**	18
OWNER	ISOLINA GERONA	**AVLAND**	$ 24,117.00	**LTDEPTH**	98
ADDRESS	218 EAST 32 STREET	**AVTOT**	$ 40,248.00	**BLDFRONT**	18
ZIP	10016			**BLDDEPTH**	50
TAXCLASS	1			**STORIES**	4

- **Property Picture**

5 bd | 5 ba | 3,984 sqft

218 E 32nd St, New York, NY 10016

Sold: **$4,400,000** | Sold on 06/25/21 | Zestimate®: **$4,492,500**

Est. refi payment: $28,180/mo | Refinance your loan

2011	--	$42,662 +6%	
2010	--	$40,248	Matches with Record

- **Investigation**

Record 50917 displays high value only for the **value_ratio** variable, which represents the ratio of **FULLVAL (market value) over (AVLAND+AVTOT) (actual value)**. The high value suggests that the **FULLVAL is significantly higher than the actual value**. After conducting an online search for the actual land value and the market value, we discovered that the actual land value of $40,248 matches the value on the record. However, the FULLVAL in 2010 is even higher than the market value sold in 2021 of $4,400,000, which is not normal and indicates a significant decrease in the property's market value. Therefore, the **FULLVAL** of this property in 2010 may be **too high** and requires further investigation by domain experts.

- **Conclusion**

After conducting our initial investigation of records with the highest final scores, we have compiled a list of **50 properties with a high risk of fraud**. However, to prevent false positive cases, we will provide this list to domain experts for their review. Their feedback helped us determine the legitimacy of these properties and enabled us to adjust our model.

8. Summary

In summary, we finished the whole pipeline to build an unsupervised fraud model including data observation, data cleaning, variables creation, dimensionality reduction, model exploration, primary investigation of results, and model improvement. We will describe the process in the following:

- **Data Observation and Data Cleaning**

Upon analyzing the New York property valuation and assessment information from 2010, which consisted of **1,070,994 records and 32 fields**, we identified several issues such as **null values** and a **high frequency of 0 or 1 values in numeric fields** such as STORIES and FULLVAL. We also noticed **outliers**, including a FULLVAL of $6,150,000,000.

At the request of clients, we focused solely on detecting tax fraud committed by private property owners and **excluded records of government-owned properties**. Our objective was to identify anomalies between property characteristics and their corresponding values in the records. To achieve this, we utilized **9 fields** that were relevant to property values and characteristics out of the 32 fields available to create variables. We treated null and zero values as missing values and replaced them with the **most typical value**, which was the average value of the same tax class group. We also **retained outliers** to aid in the detection of anomalies.

- **Variable Creation and Dimensionality Reduction**

To address the clients' concern of detecting fraud where scammers intentionally misrepresent property characteristics to underpay taxes, we created 3 temporary fields for determining size, and **54 variables** to assess the **relationship between value and size**. Furthermore, we included **4 variables** to **compare actual and market values**, as well as to **compare building and lot sizes**.

In order to reduce dimensionality and combine correlated variables, we began by standardizing the original variables with different scales using **z-scaling**. Next, we utilized **PCA** to identify the dominant directions of variance in the data and transformed the original data points into points in new coordinates. We retained the **top 5 principal components** and **applied z-scaling again** to ensure equal significance across all dimensions for distance calculation purposes. These variables after transformation are called z-scores.

- **Two Score Algorithms**

To determine unusualness, two scoring methods are utilized: Score 1, computed through z-score outlier analysis, and Score 2, using an autoencoder. For Score 1, the Minkowski distance between z-scores and the origin 0 is summed for each record on every PC with **p of 2**. On the other hand, Score 2 is computed by measuring the difference between the original input vector and the output vector from a trained neural net model, with the difference determined by

Minkowski distance with **p of 2**. Ultimately, the final scores are obtained by **averaging the rank order** of the two scores.

- **Results Investigation and Findings**

We sorted the records by final scores in descending order and created a **heatmap** of z-scores for the **top 200 records**. We identified unusual variable values in darker regions of the heatmap and **reviewed the original fields**, such as size and value, for those records. To ensure accuracy, we verified the information using images of the properties available online and data from real-estate marketplaces such as Zillow.

After our preliminary investigation, we identified **50 properties with unusual characteristics**, such as incorrect number of stories, building sizes larger than lot sizes, market values significantly higher or lower than expected for the same tax class or location, and high market values compared to actual values. We compiled a list of these anomalous records and forwarded it to domain experts for further evaluation and feedback.

- **Model Improvement**

After receiving feedback from clients, we made modifications to our model in three areas: data exclusion, data imputation, and code improvement.

The client indicated that some public properties were still included in the anomalous records list, prompting us to enhance the **accuracy of classification between private and public properties** during the data exclusion process. To achieve this, we used two methods: creating a government_list by filtering owner names with keywords such as 'DEPT, CITY, STATES, etc.' and by using EXMPTCL to identify properties that didn't pay taxes. These modifications resulted in an increased level of accuracy and efficiency during the classification process.

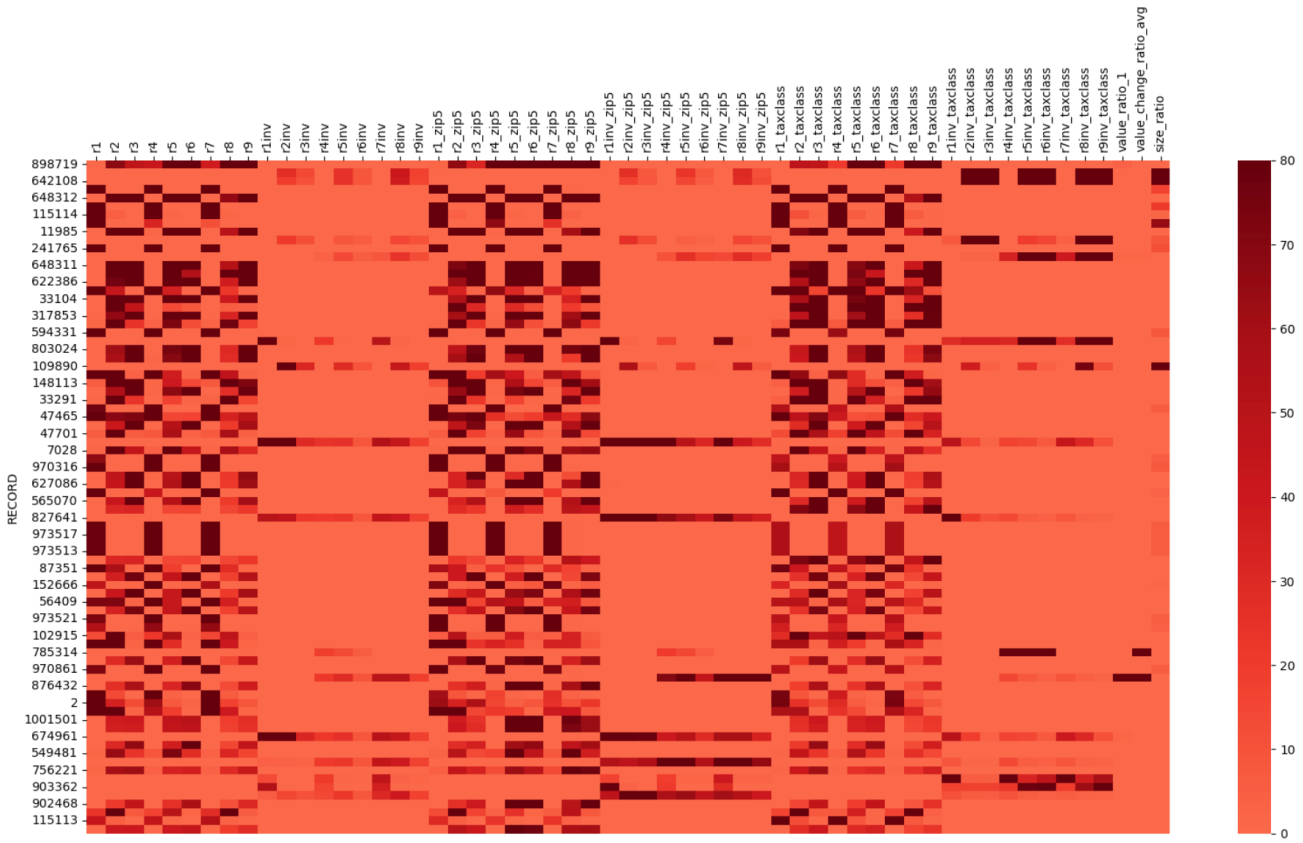
Clients also expressed their **disinterest in fields with a value of 1**, which they believed should be treated as missing values. Therefore, we changed our imputation method to replace both 0 and 1 values in size and value fields with missing values. Furthermore, clients noted that some zip codes did not match the boro information, and thus, we **mapped the boro of records to impute zip** codes accordingly.

To enhance the efficiency of our model process for larger volumes of records, we replaced several manual calculation or imputation codes with for loops. After two iterations of model modification, clients were satisfied with our models and findings.

Appendix

a. Appendix A – Heatmaps of Z scores for Top Records

Below is the heatmaps of Z scores for top records to detect the unusualness.



b. Appendix B – Data Quality Report

Data Quality Report

1. Data Description

The data is a collection of **New York property valuation and assessment information** provided from the city of New York to detect **property tax frauds**. The data covers the properties information with total **1,070,994 records** and **32 fields**. Since there is **no fraud label** provided, we will build an unsupervised model to detect property values that are **unusual** for that property's characteristics.

2. Summary Tables

The summary tables of numeric and categorical data are included in the report.

3. Visualization of Each Field – Distribution

(1) Field Name: RECORD

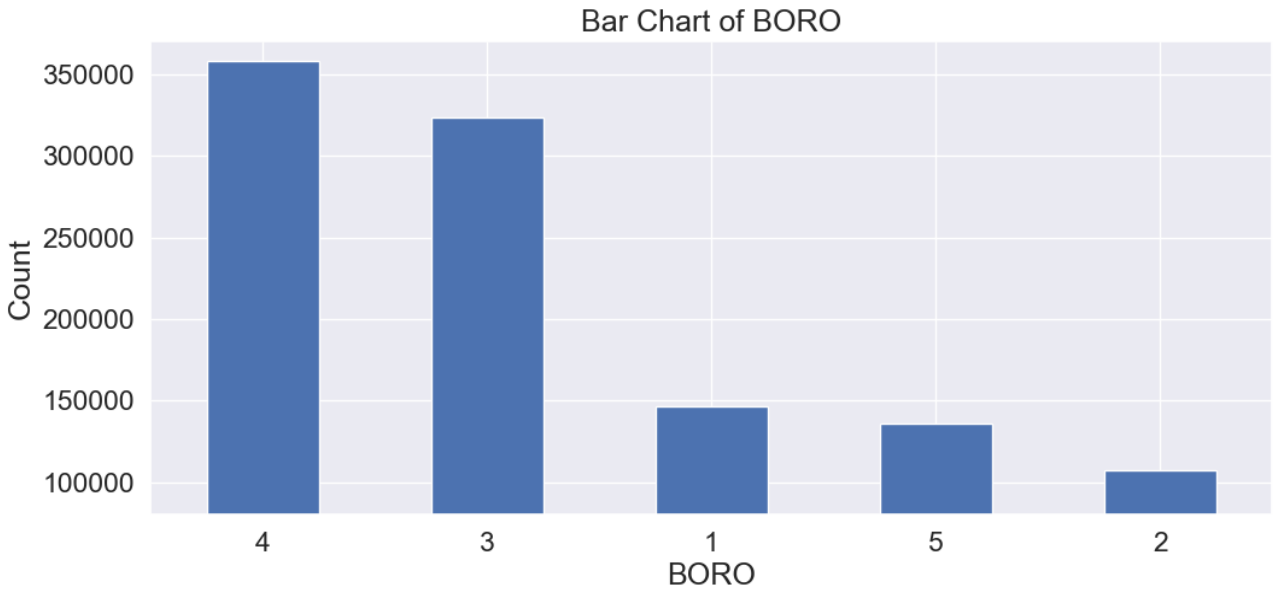
- **Visualization:** This record field has all unique values. Therefore, we don't need a histogram/distribution for this field.
- **Description:** this field is about record number of property information with ordinal unique positive integer from 1 to 1,070,994.

(2) Field Name: BBLE

- **Visualization:** This record field has all unique values. Therefore, we don't need a histogram/distribution for this field.
- **Description:** Series of number containing boro, block, lot, and easement code of each property. This field represent each property with unique positive integer from 1000010101 to 5080500094.

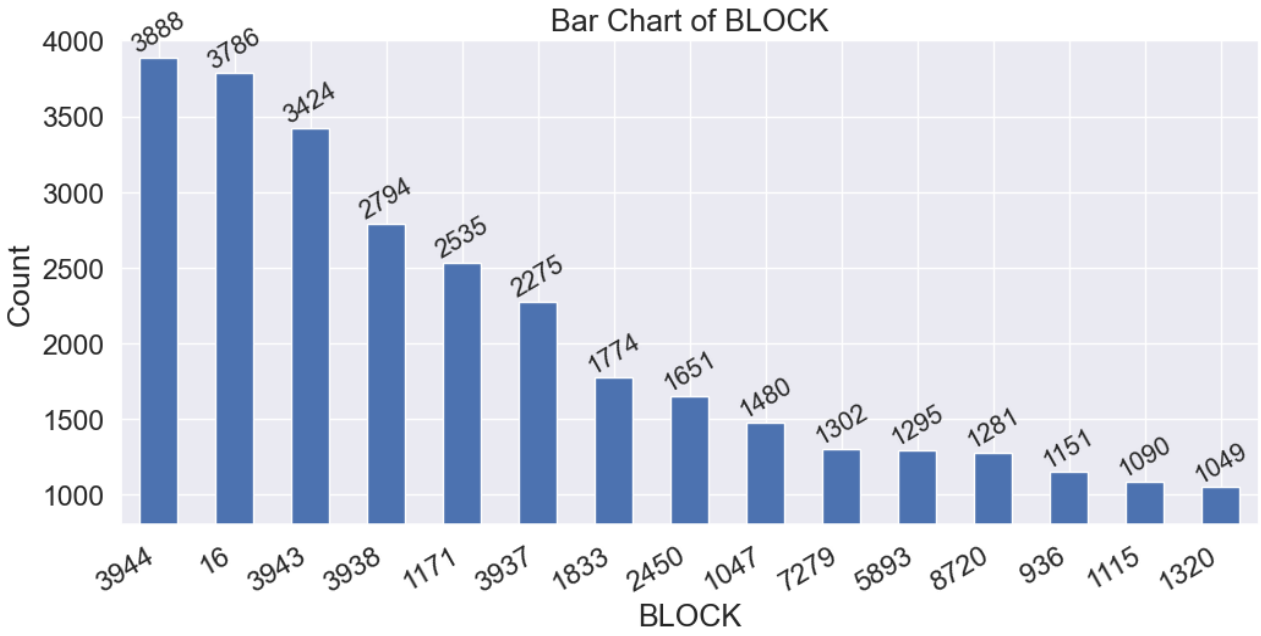
(3) Field Name: BORO

- **Visualization:** Bar Chart of BORO. The chart contains all **5** borough categories. For a better visualization of the shape, y axis starts from 80,000.
- **Description:** Borough in each record/property. The most common borough division shown in records is **4 = Queens**, with total amount of 358,046.



(4) Field Name: BLOCK

- **Visualization:** Bar Chart of BLOCK. The chart selects top **15** field values of block. For a better visualization of the shape, y axis starts from 800.
- **Description:** Valid block ranges by borough in each record/property. The most common block shown in records is **3944**, with total amount of 3,888.

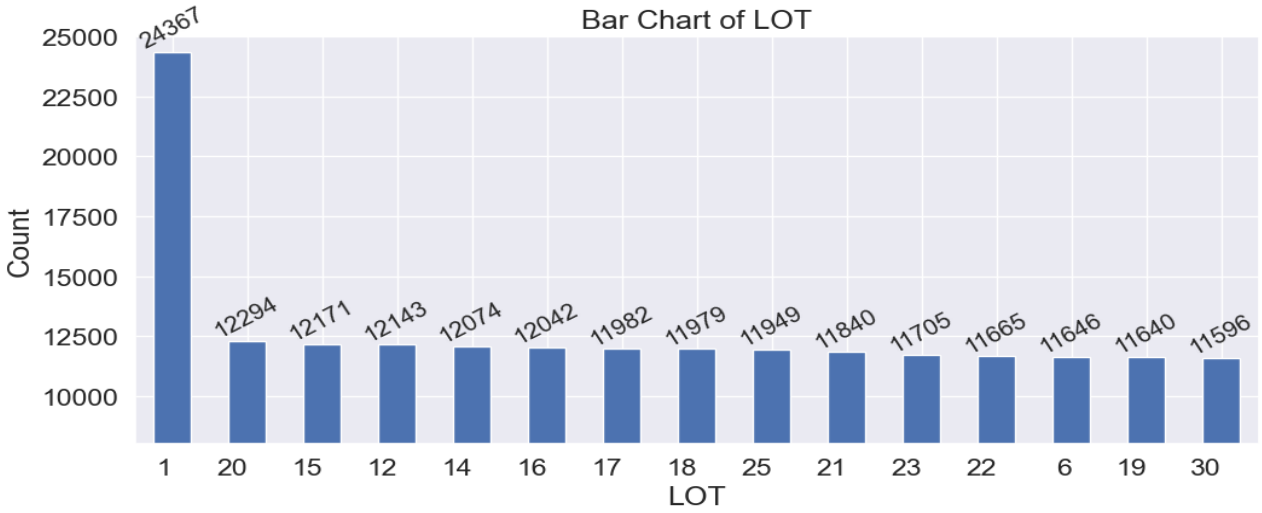


(5) Field Name: LOT

Since we want to find unusual property values, we will not only find the most common lot number in records, but also group LOT in histogram to find the common lot ranges.

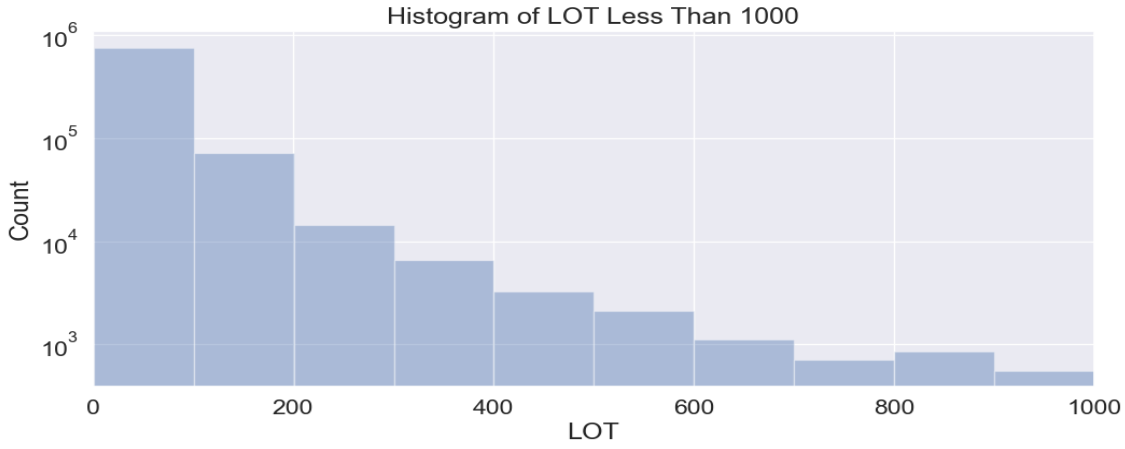
a. Visualization: Bar Chart of LOT. The chart selects top **15** field values of lot. For a better visualization of the shape, y axis starts from 8,000.

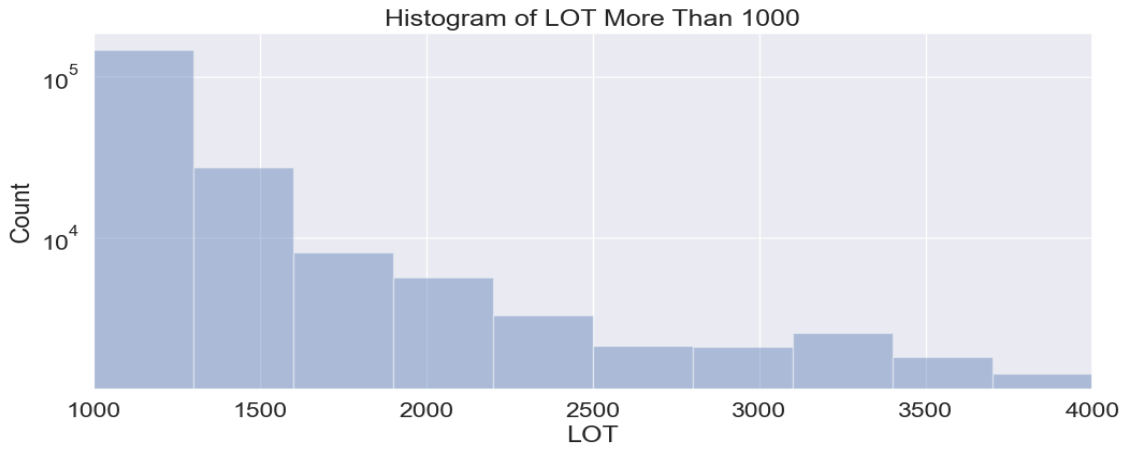
- **Description:** Lot number in each record/property. The most common lot shown in records is **1**, with total amount of 24,367.



b. Visualization: Histograms of LOT. The two charts select lot values from range [0,999] and [1000,3999] to find **most common lot ranges**. For a better visualization of the shape, y axis starts from 1,000.

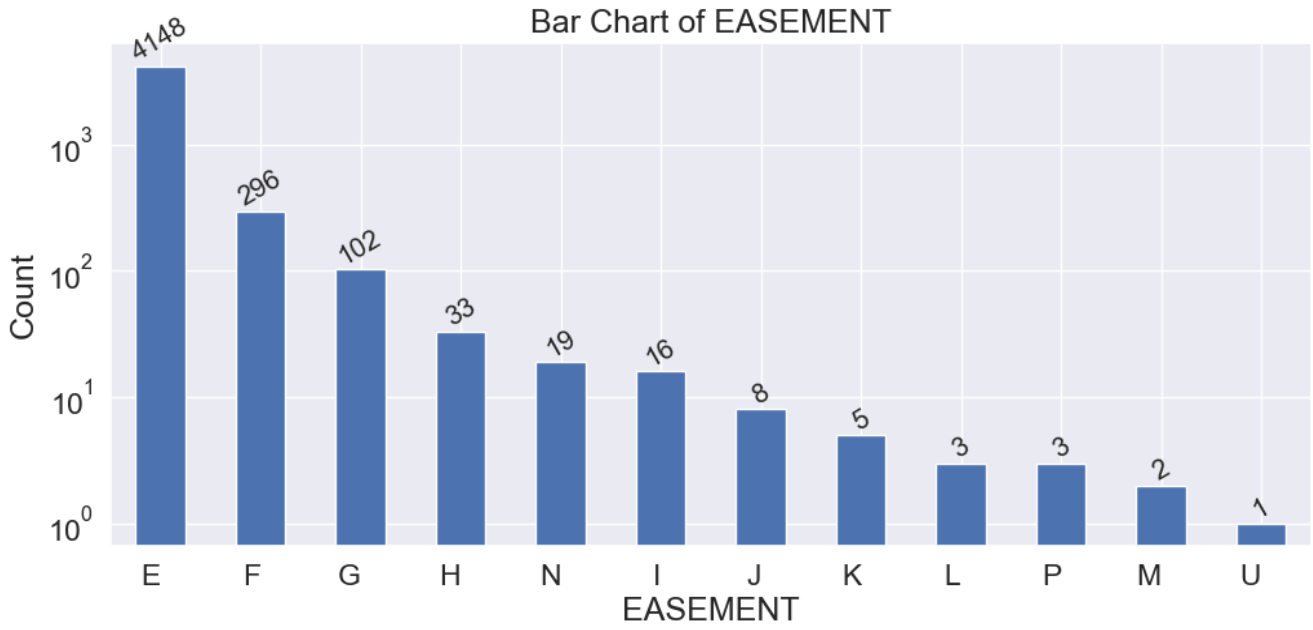
- **Description:** The most common lot ranges with over 100,000 records are: **[0,100] and [1000,1300]**.





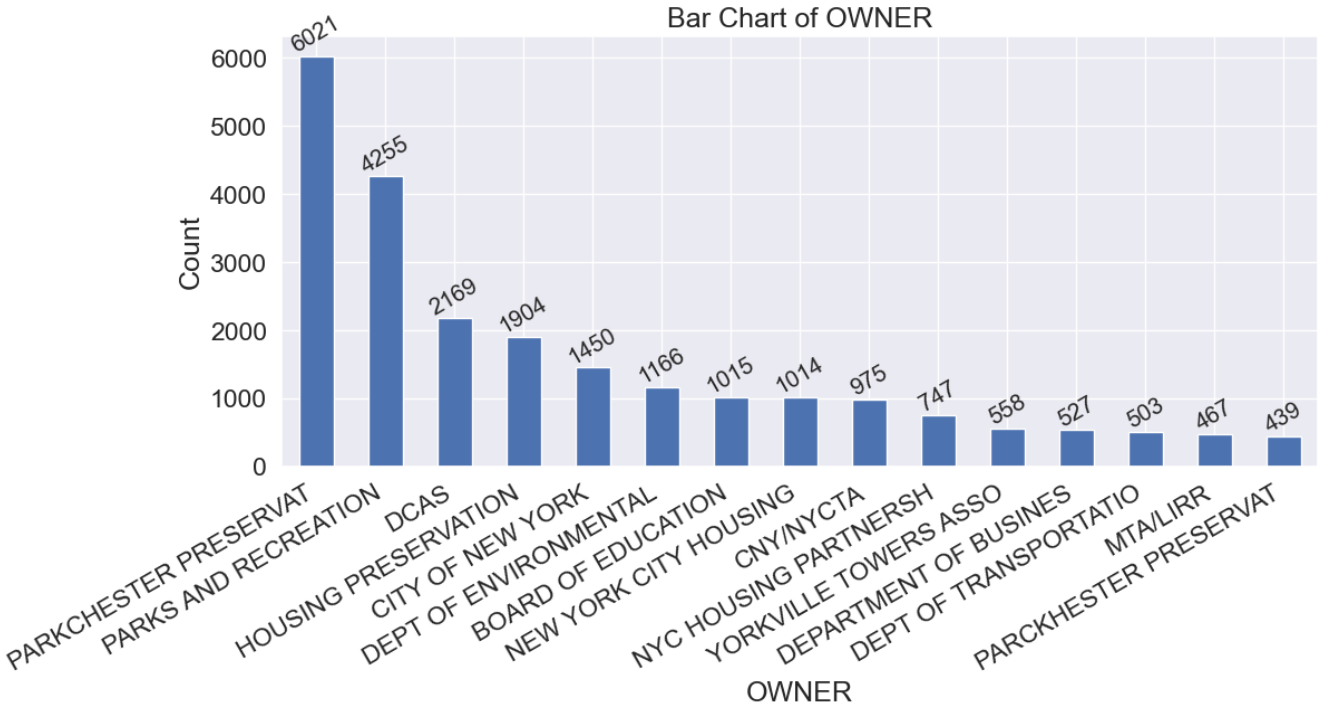
(6) Field Name: EASEMENT

- **Visualization:** Bar Chart of EASEMENT. The chart contains all **12** field values of easement.
- **Description:** Easement information/category in each record/property. The most common easement category shown in records is **E = Land Easement**, with total amount of 4,148.



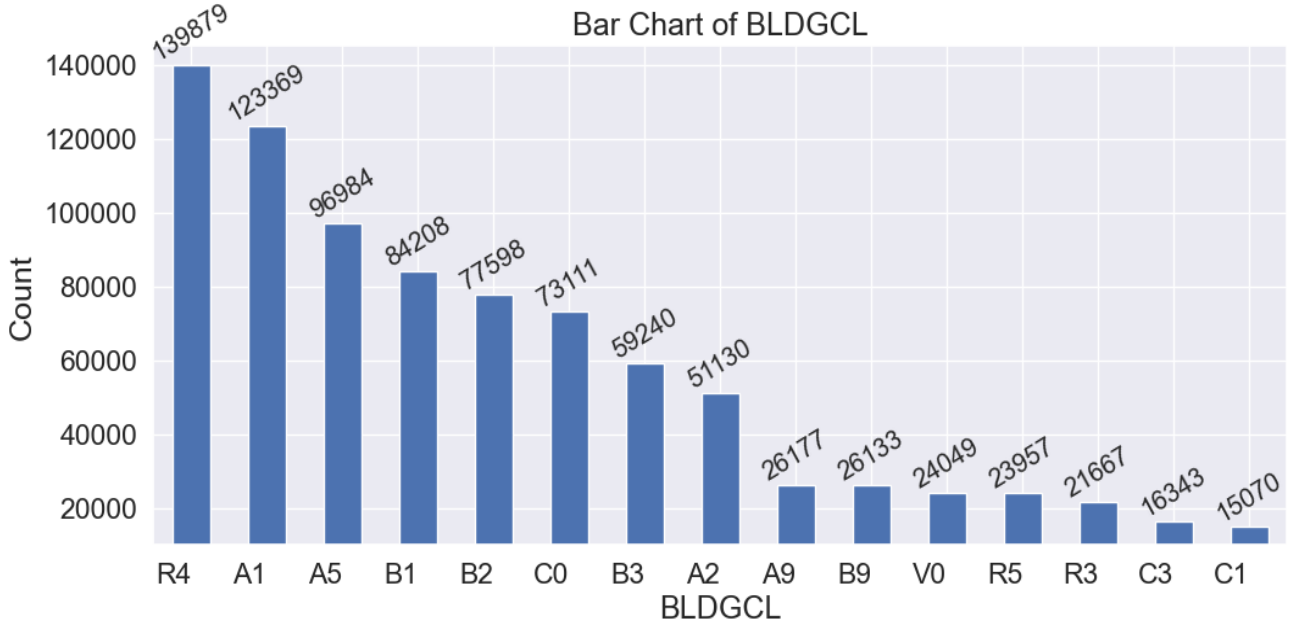
(7) Field Name: OWNER

- **Visualization:** Bar Chart of OWNER. The chart selects top **15** field values of owner.
- **Description:** Owner Name in each record/property. The most common owner name shown in records is **PARKCHESTER PRESERVAT**, with total amount of 6,021. However, another owner PARCKHESTER PRESERVAT with count of 439 is very similar to the most common value and we should investigate more into records with this owner.



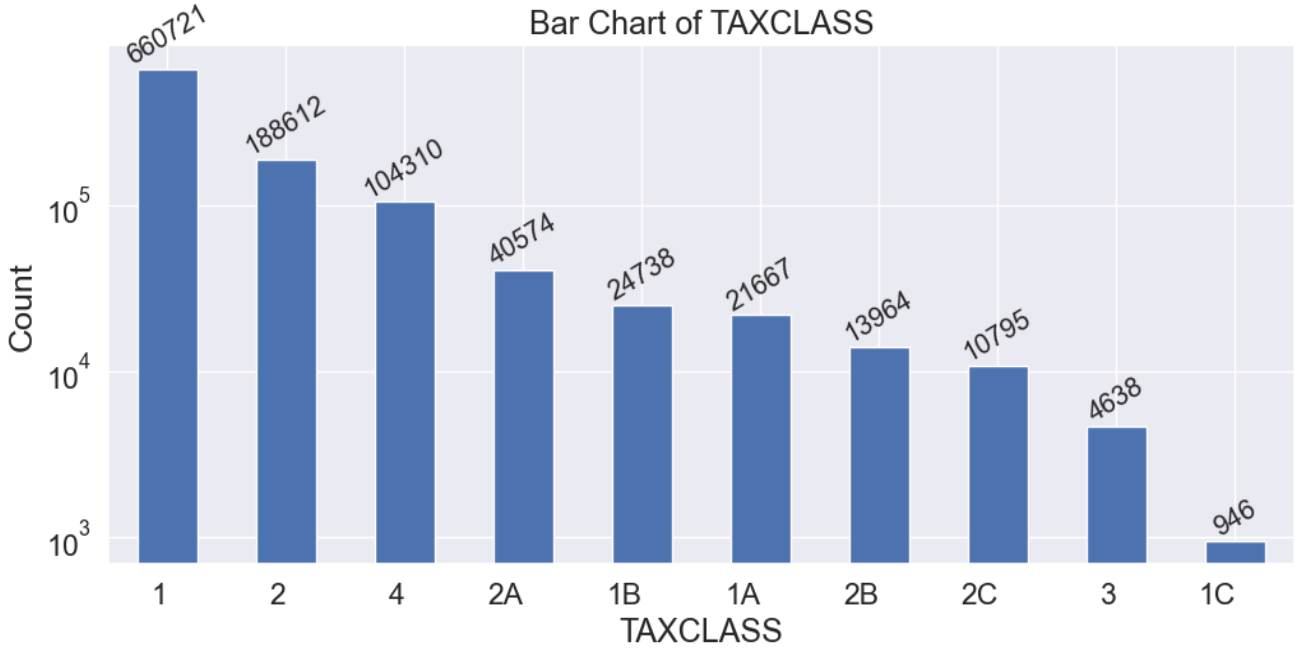
(8) Field Name: BLDGCL

- **Visualization:** Bar Chart of BLDGCL. The chart selects top **15** field values of BLDGCL. For a better visualization of the shape, y axis starts from 10,000.
- **Description:** Building class category in each record/property. The most common building class category shown in records is **R4**, with total amount of 139,879.



(9) Field Name: TAXCLASS

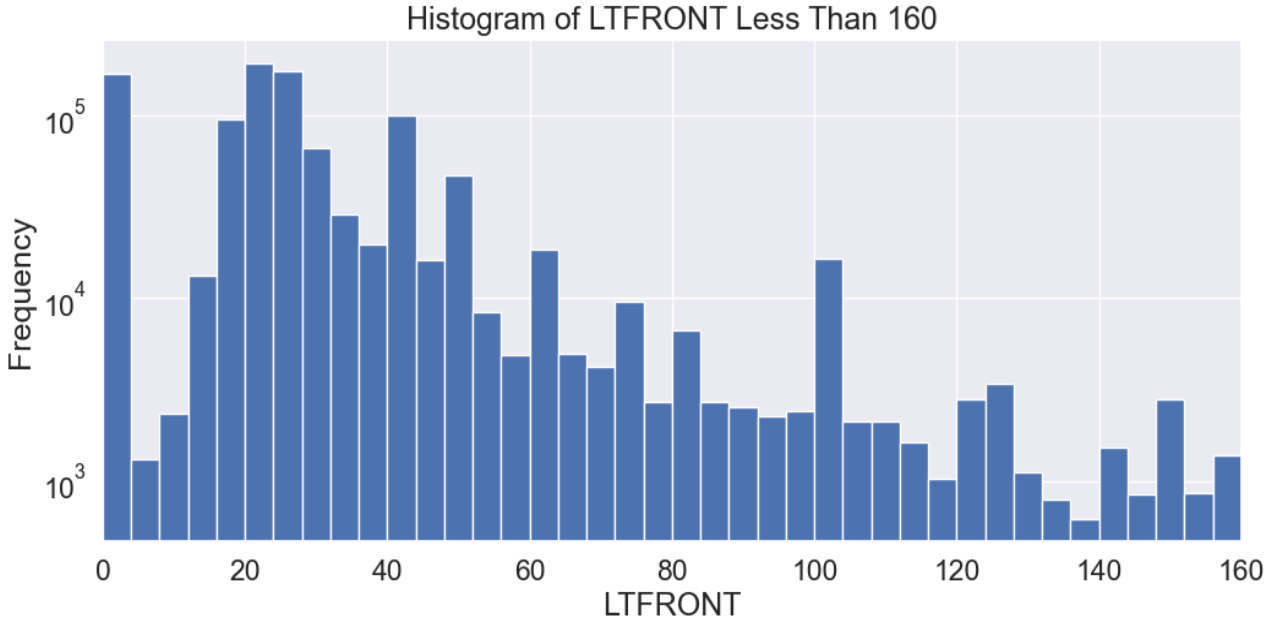
- **Visualization:** Bar Chart of TAXCLASS. The chart selects top **10** field values of TAXCLASS. For a better visualization of the shape, y axis starts from 900.
- **Description:** Tax class of property in each record. The most common tax class shown in records is **1**, with total amount of 660,721. The least common value is 1D with total amount of 29 (not in plot).



(10) Field Name: LTFRONT

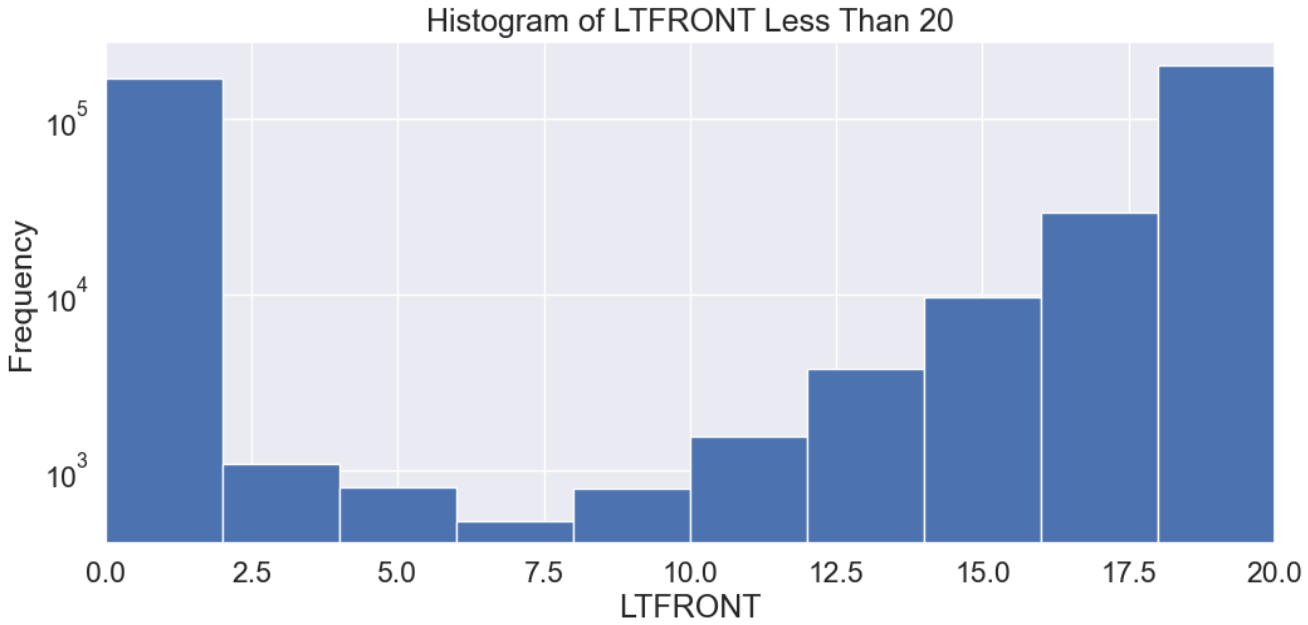
a. **Visualization:** Histogram of LTFRONT with a **range of x in [0, 160],** which covers **97.38%** of the property records.

- **Description:** Lot width of properties. We can observe a **high amount of frequency around 0** and a big drop after that, showing there is missing data in this field. There is another high amount of frequency when lot width is around 20, the frequency begins to drop from 20 to 160.



b. **Visualization:** Histogram of LTFRONT with a **range of x in [0, 20].**

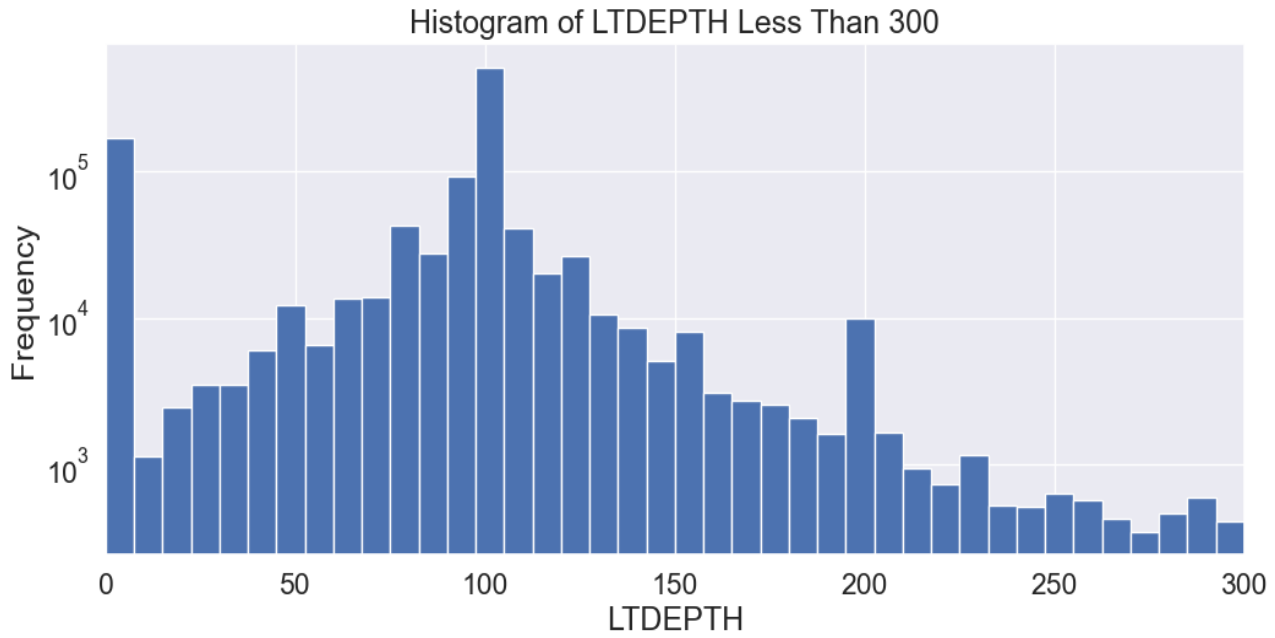
- **Description:** To better understand trend of frequency before lot width of 20, we have another histogram showing the trend of frequency of lot width in [0,20]. The frequency **drops from 0 to 7.5 lot width** and increases from 10 to 20 lot width.



(11) Field Name: LTDEPTH

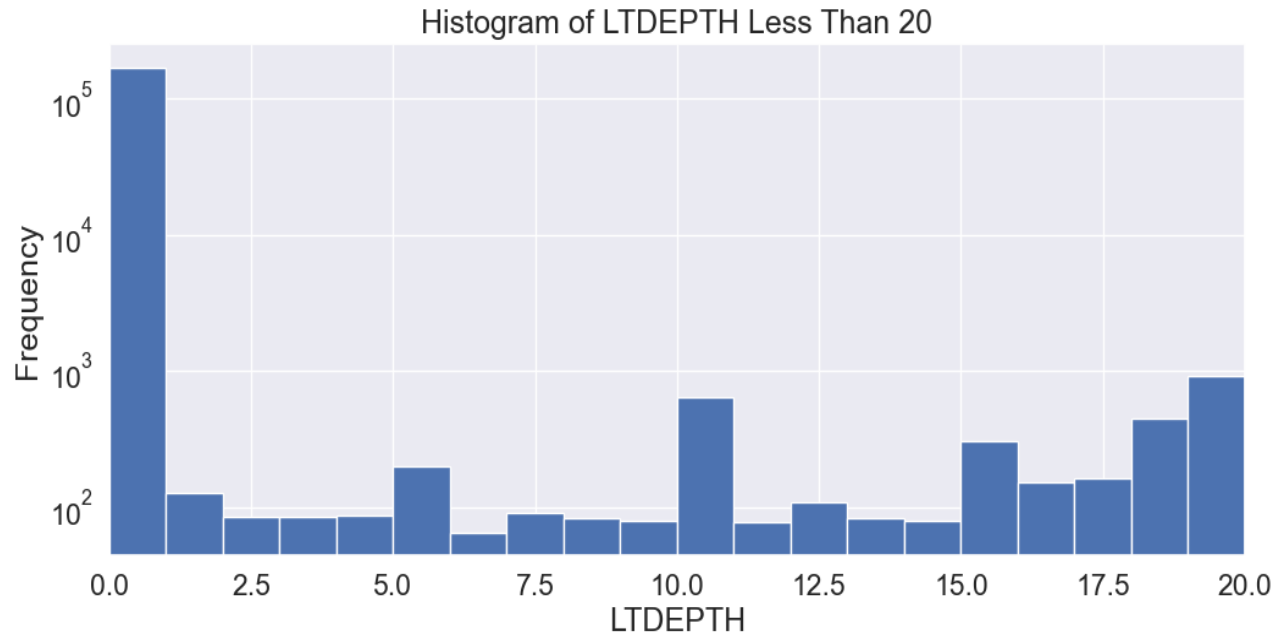
a. **Visualization:** Histogram of LTDEPTH with a **range of x in [0, 300],** which covers **99.17%** of the property records.

- **Description:** Lot depth of properties. We can observe a high amount of frequency around 0 and a big drop after that, showing there is missing data in this field. The frequency increases from 20 to 100 lot depth and begins to drop after lot depth is 100.



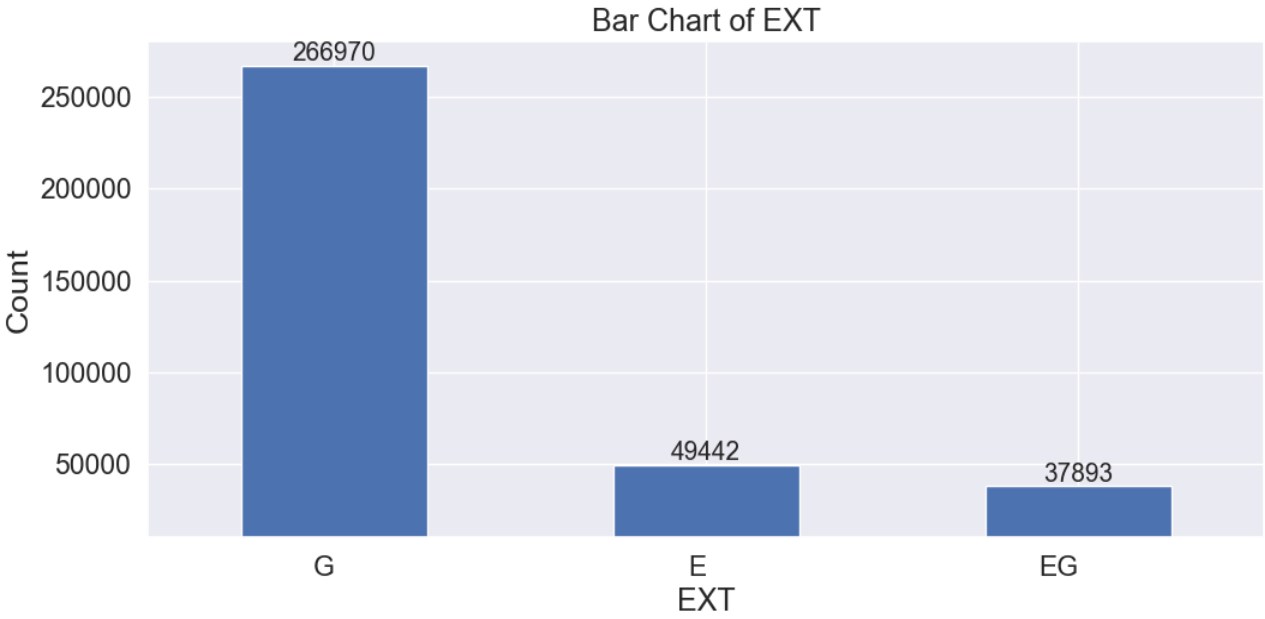
b. **Visualization:** Histogram of LTDEPTH with a **range of x in [0, 20].**

- **Description:** To better understand trend of frequency before lot depth of 20, we have another histogram showing the trend of frequency of lot depth in [0,20]. The frequency **drops from 0 to 2.5 lot depth** and increases from 17.5 to 20 lot depth.



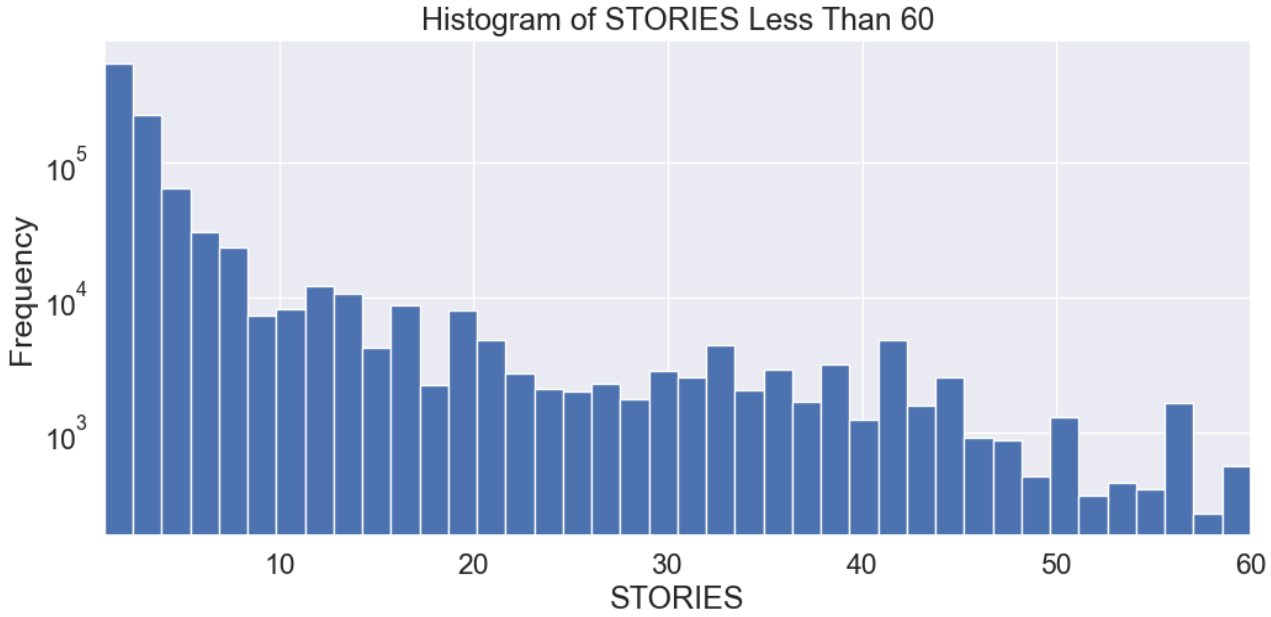
(12) Field Name: EXT

- **Visualization:** Bar Chart of EXT. The chart contains all **3** field values of EXT. For a better visualization of the shape, y axis starts from 10,000.
- **Description:** Extension indicator of property in each record. The most common EXT shown in records is **G**, with total amount of 266,970.



(13) Field Name: STORIES

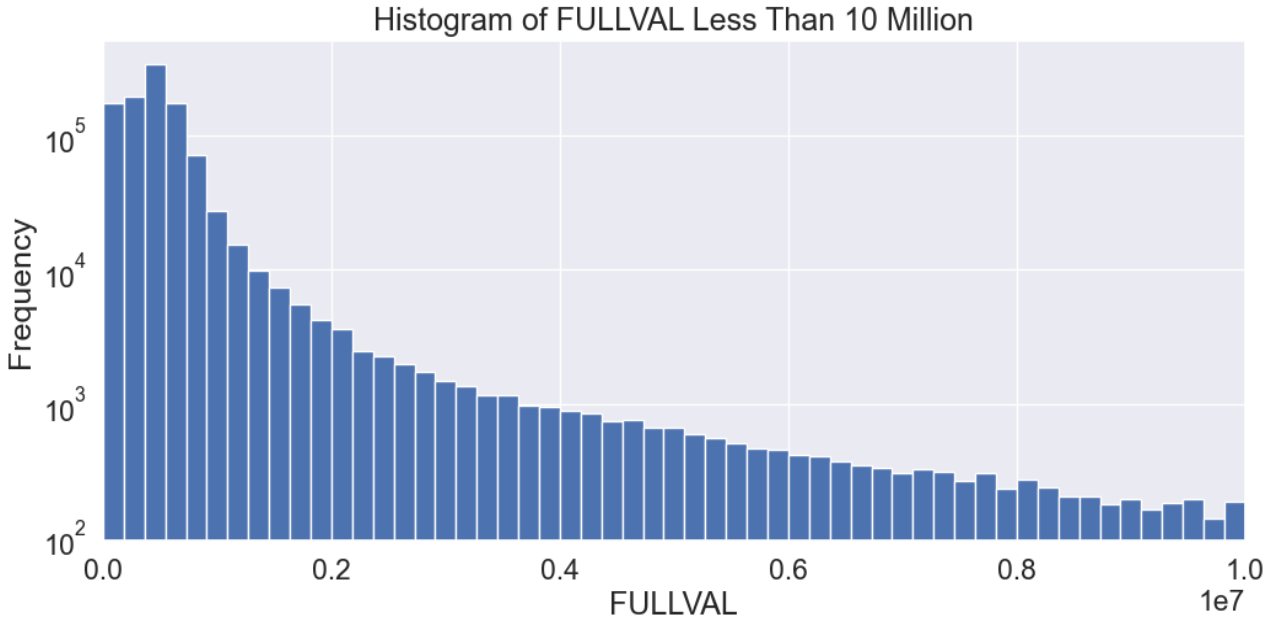
- **Visualization:** Histogram of STORIES with a **range of x in [0, 60],** which covers **94.63%** of the property records.
- **Description:** Number of stories in building of properties. There is no zero value for STORIES. We can observe a trend of decreasing from 1 to 60 stories.



(14) Field Name: FULLVAL

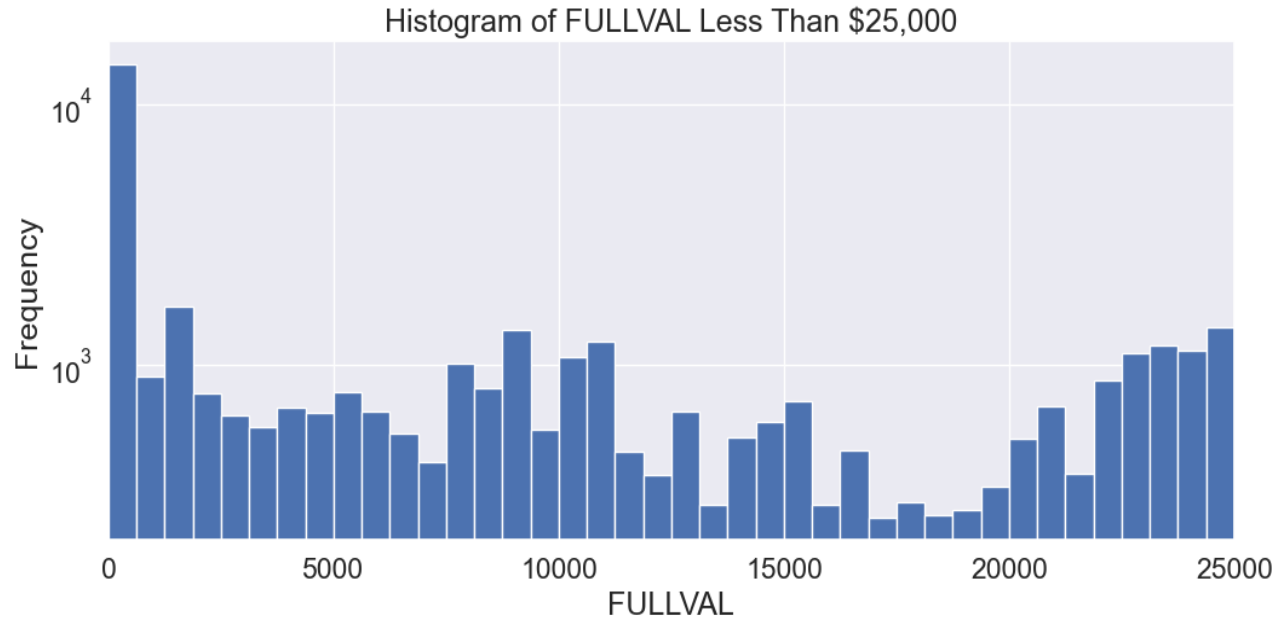
a. **Visualization:** Histogram of FULLVAL with a **range of x in [0, $10,000,000],** which covers **99.28%** of the property records.

- **Description:** Market value of properties. We can observe a **high amount of frequency around 0**, showing there is missing data in this field. There is a trend of decreasing from $1,000,000 to $10,000,000.



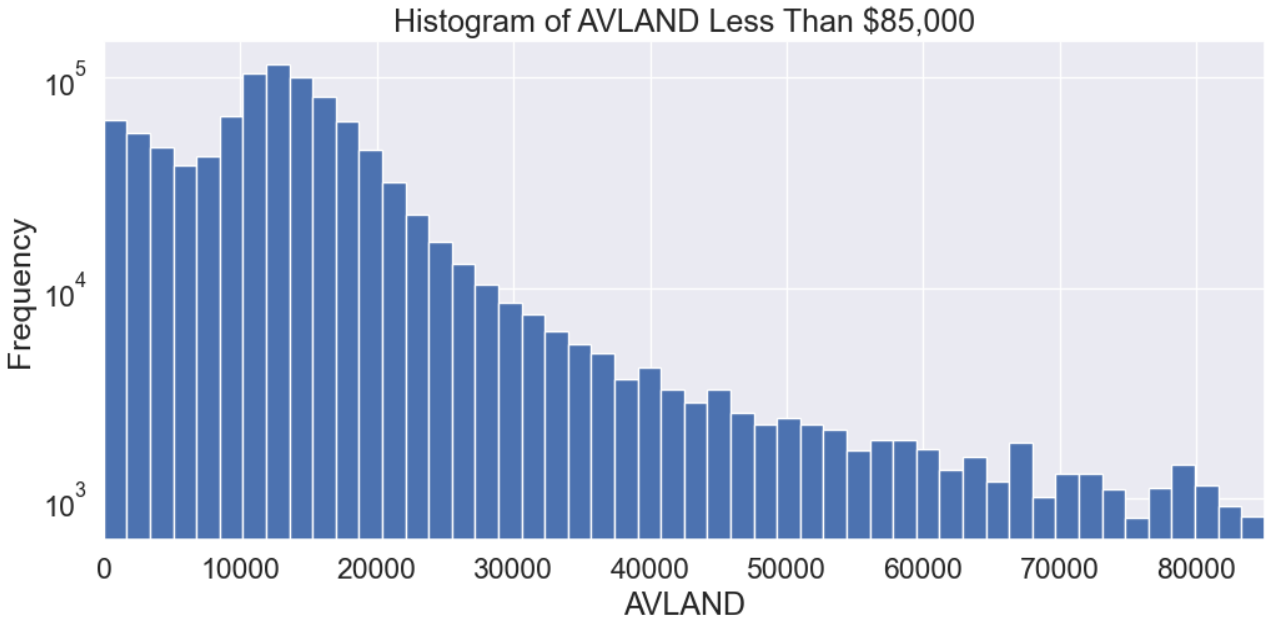
b. **Visualization:** Histogram of FULLVAL with a **range of x in [0, $25,000].**

- **Description:** To better understand trend of frequency before Market value of $25,000, we have another histogram showing the trend of frequency of FULLVAL in [0, $25,000]. The frequency **drops from 0 to $20,000** and increases with market value from $20,000 to $25,000.



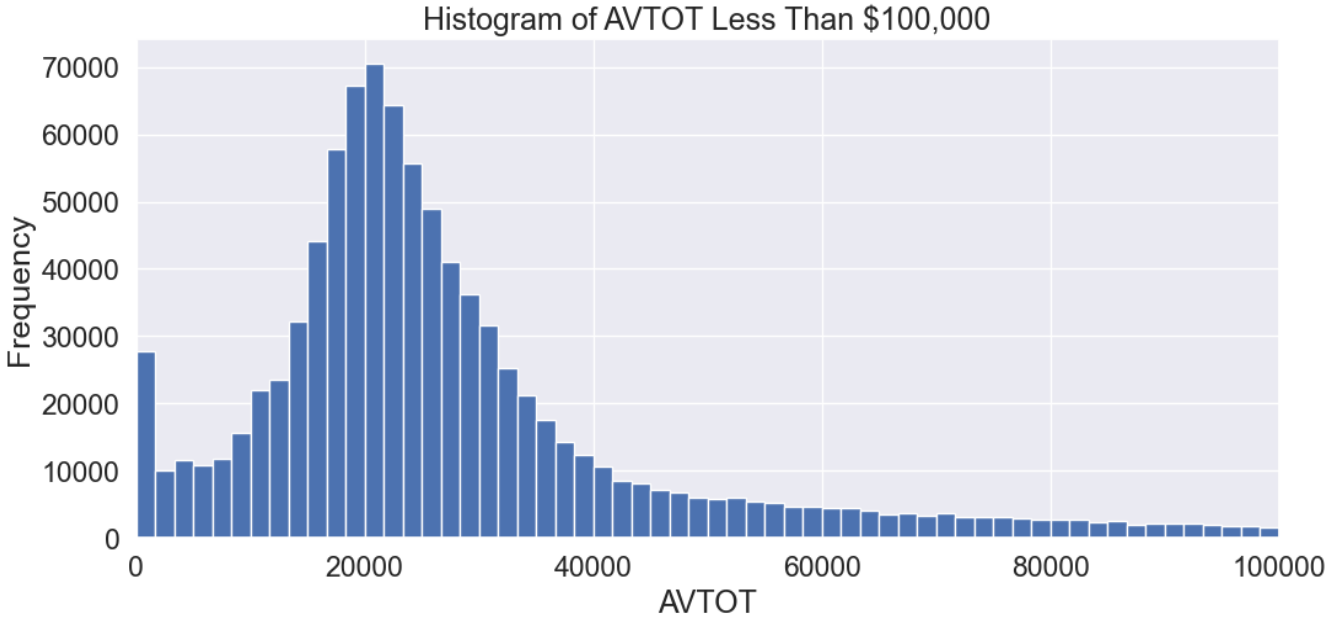
(15) Field Name: AVLAND

- **Visualization:** Histogram of AVLAND with a **range of x in [0, $85,000],** which covers **93.32%** of the property records.
- **Description:** Actual land value of properties. We can observe a **decrease and increase before $10,000**. There is another **trend of decreasing from $10,000 to $85,000**.



(16) Field Name: AVTOT

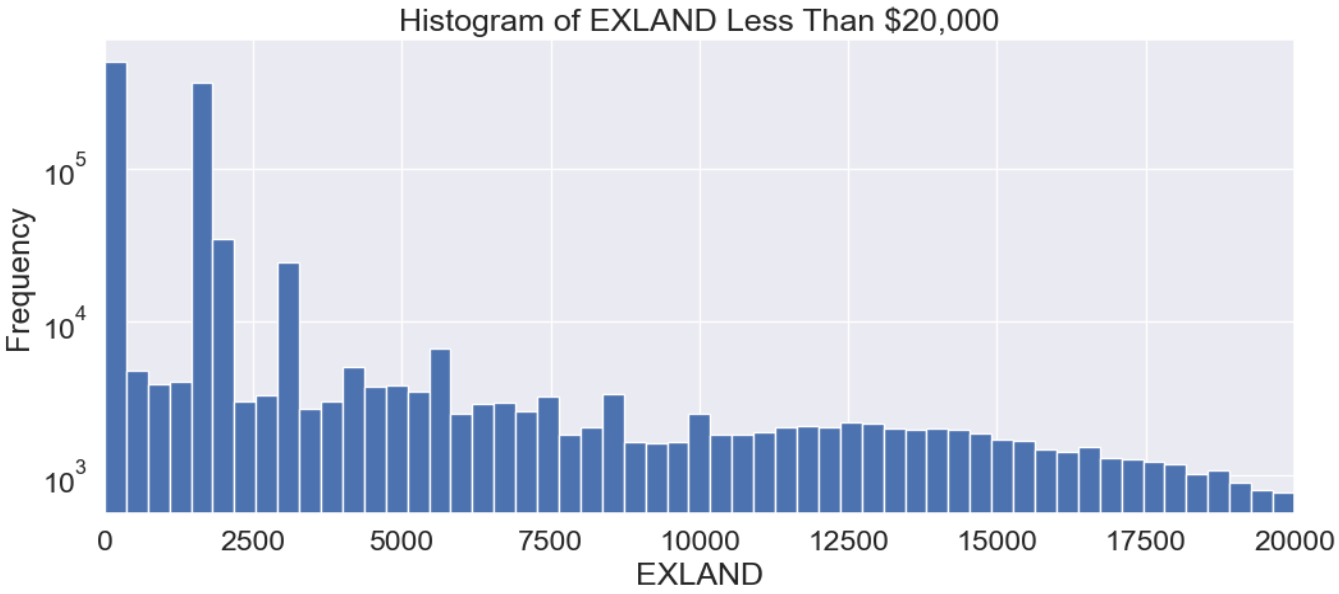
- **Visualization:** Histogram of AVTOT with a **range of x in [0, $100,000],** which covers **86.05%** of the property records.
- **Description:** Actual total value of properties. We can observe a **high frequency of amount around 0, showing there is missing or zero values. The frequency starts to increase till $20,000 and decreases from $20,000 to $100,000**.



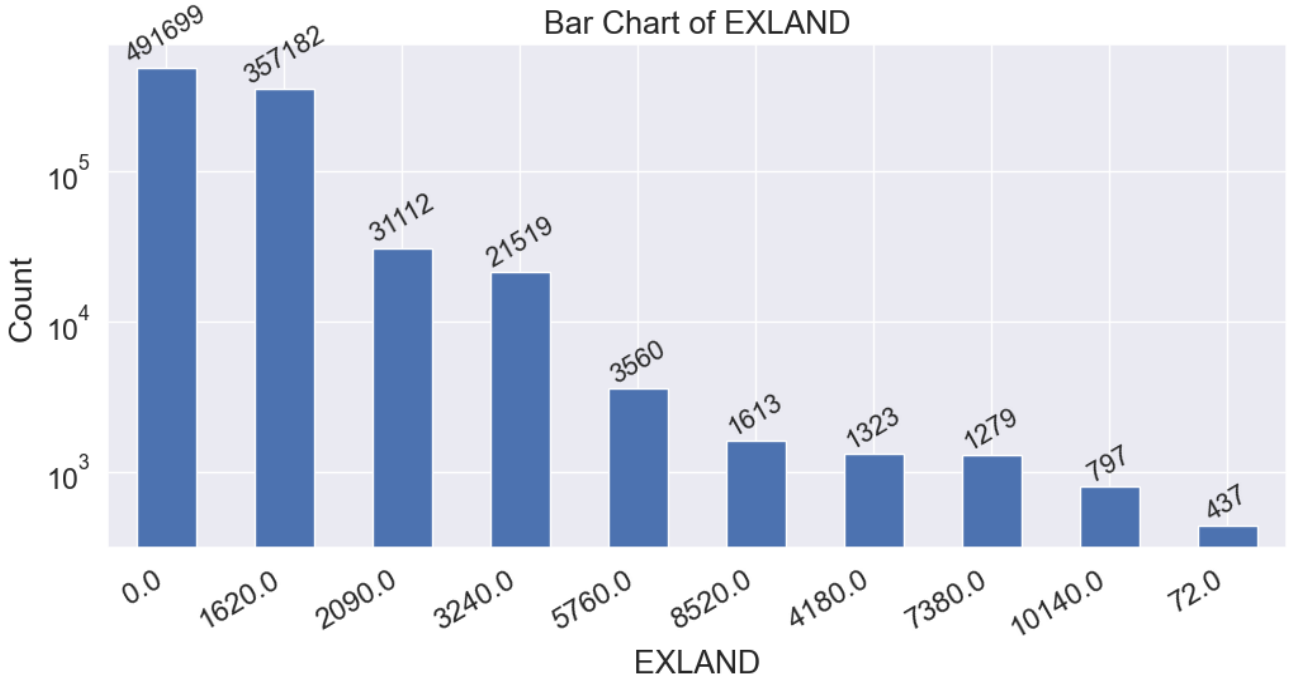
(17) Field Name: EXLAND

a. **Visualization:** Histogram of EXLAND with a **range of x in [0, $20,000],** which covers **96.82%** of the property records.

- **Description:** Actual exemption land value of properties. We can observe a **high frequency of amount around 0, showing there is missing or zero values.** There are **some high frequencies of amount** with specific exemption land value which needs further investigation. Overall, the trend of frequency is stable.



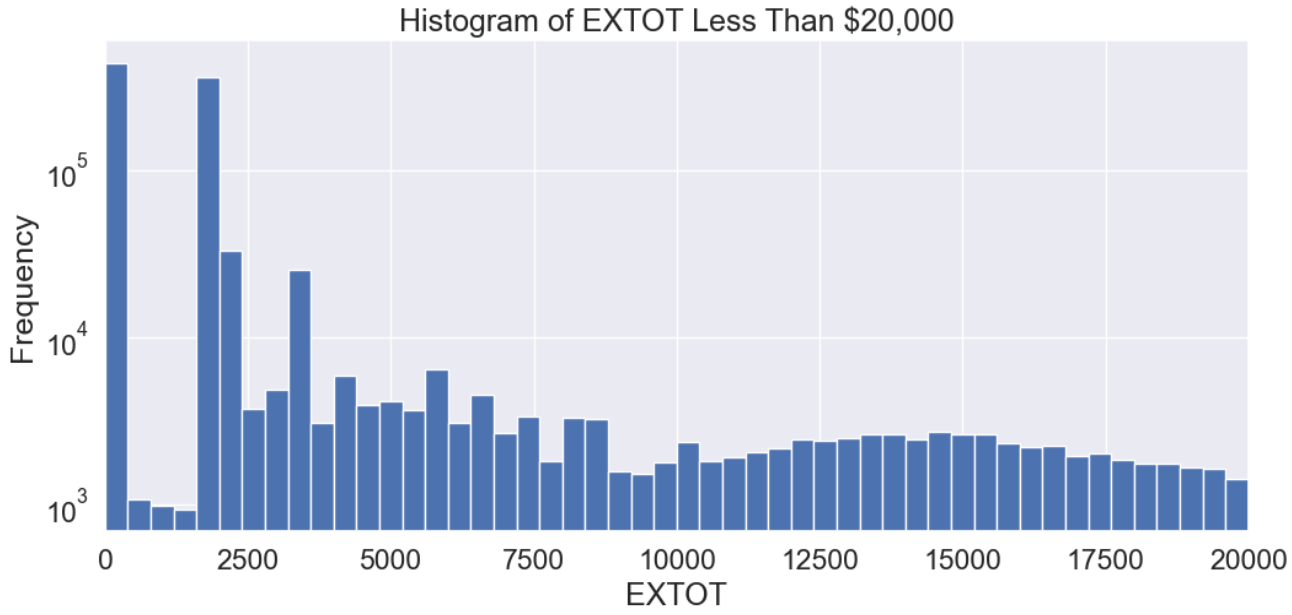
- **Visualization:** Bar chart of EXLAND. We select the top 10 field values of EXLAND.
- **Description:** From the distribution above, we know that there are some certain EXLAND with high frequencies. Therefore, we draw the bar chart to find out **these common values**: **$0, $1,620, $2,090, $3,240** with total counts of over 10,000.



(18) Field Name: EXTOT

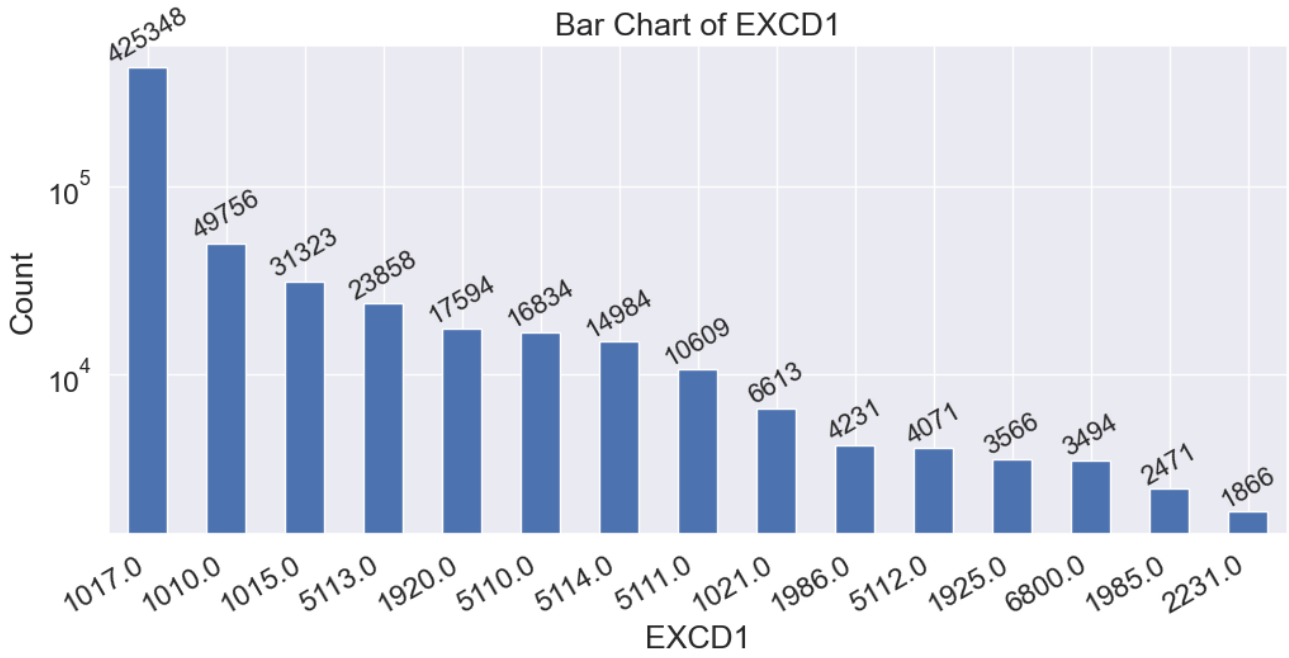
a. **Visualization:** Histogram of EXTOT with a **range of x in [0, $20,000],** which covers **90.40%** of the property records.

- **Description:** Actual exemption land total of properties. We can observe a **high frequency of amount around 0, showing there is zero values.** There is a trend of **decreasing** from $2,500 to $20,000.



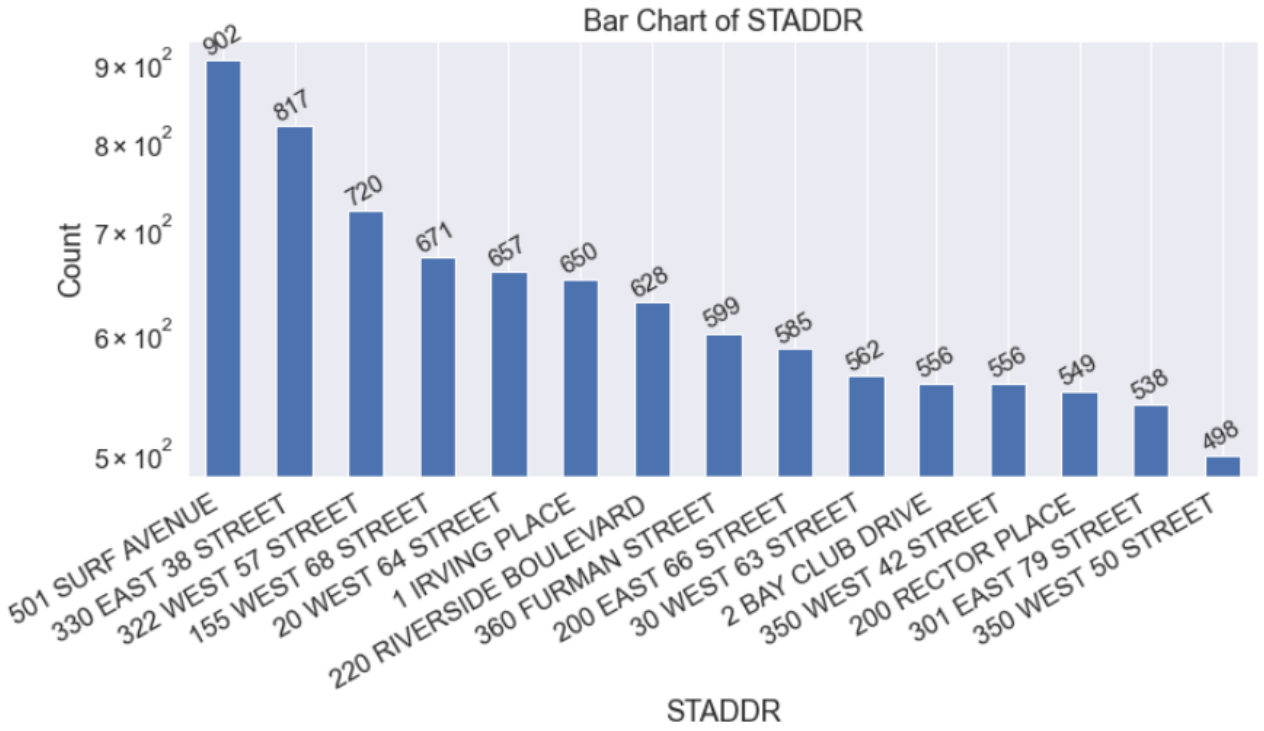
(19) Field Name: EXCD1

- **Visualization:** Bar Chart of EXCD1. The chart selects top **15** field values of EXCD1.
- **Description:** Exemption code 1 in each record/property. The most common exemption code 1 category shown in records is **1017**, with total amount of 425,348.



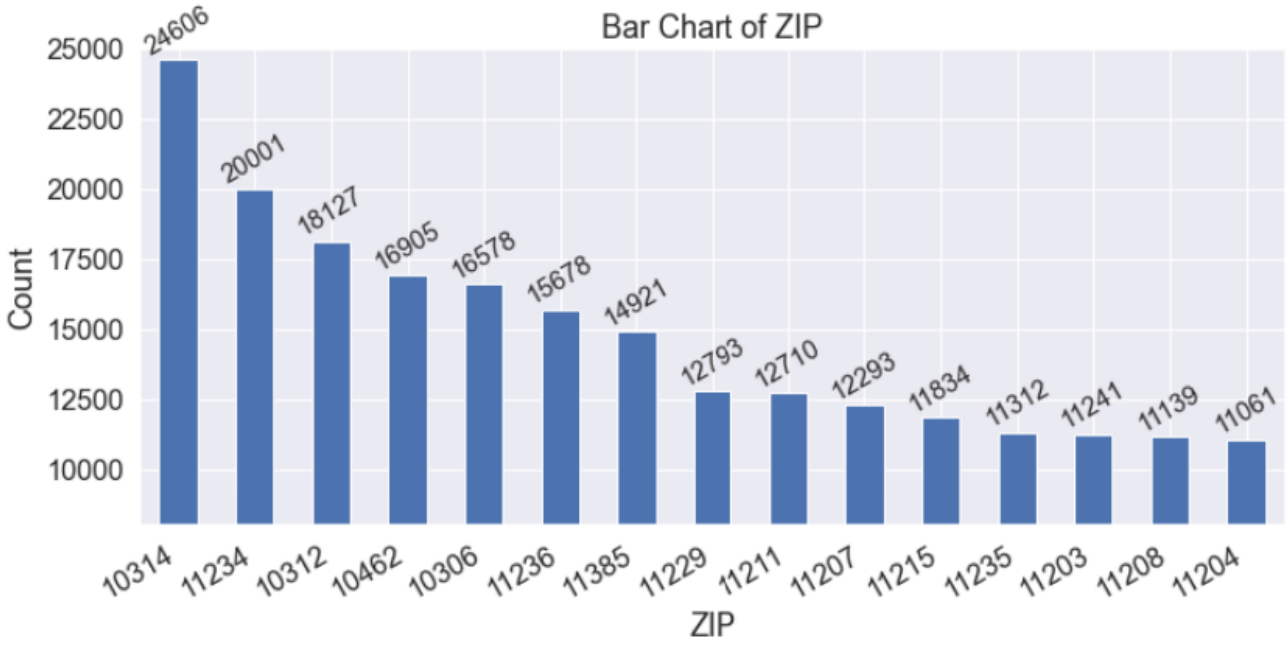
(20) Field Name: STADDR

- **Visualization:** Bar Chart of STADDR. The chart selects top **15** field values of STADDR.
- **Description:** Street address in each record/property. The most common street address shown in records is **501 SURF AVENUE**, with total amount of 902.



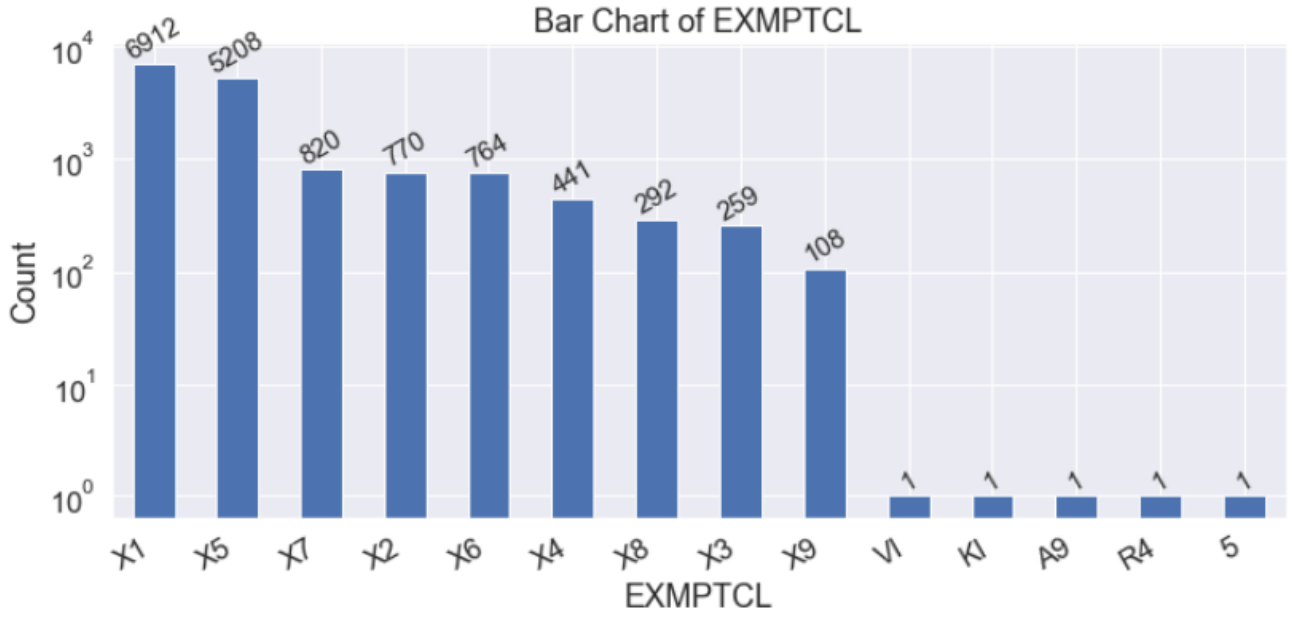
(21) Field Name: ZIP

- **Visualization:** Bar Chart of ZIP. The chart selects top **15** field values of ZIP.
- **Description:** Zip code in each record/property. The most common zip code shown in records is **10134**, with total amount of 24,606.



(22) Field Name: EXMPTCL

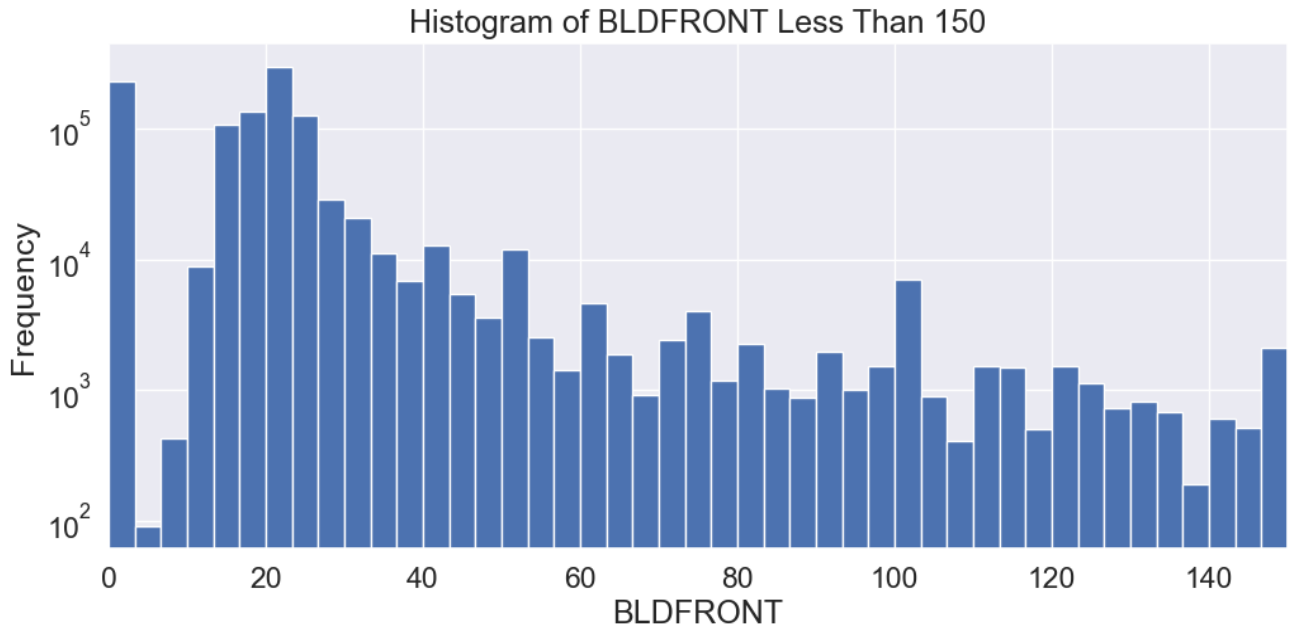
- **Visualization:** Bar Chart of EXMPTCL. The chart selects top **15** field values of EXMPTCL.
- **Description:** Exemption class in each record/property. The most common exemption class shown in records is **X1**, with total amount of 6,912.



(23) Field Name: BLDFRONT

a. **Visualization:** Histogram of BLFRONT with a **range of x in [0, 150],** which covers **98.69%** of the property records.

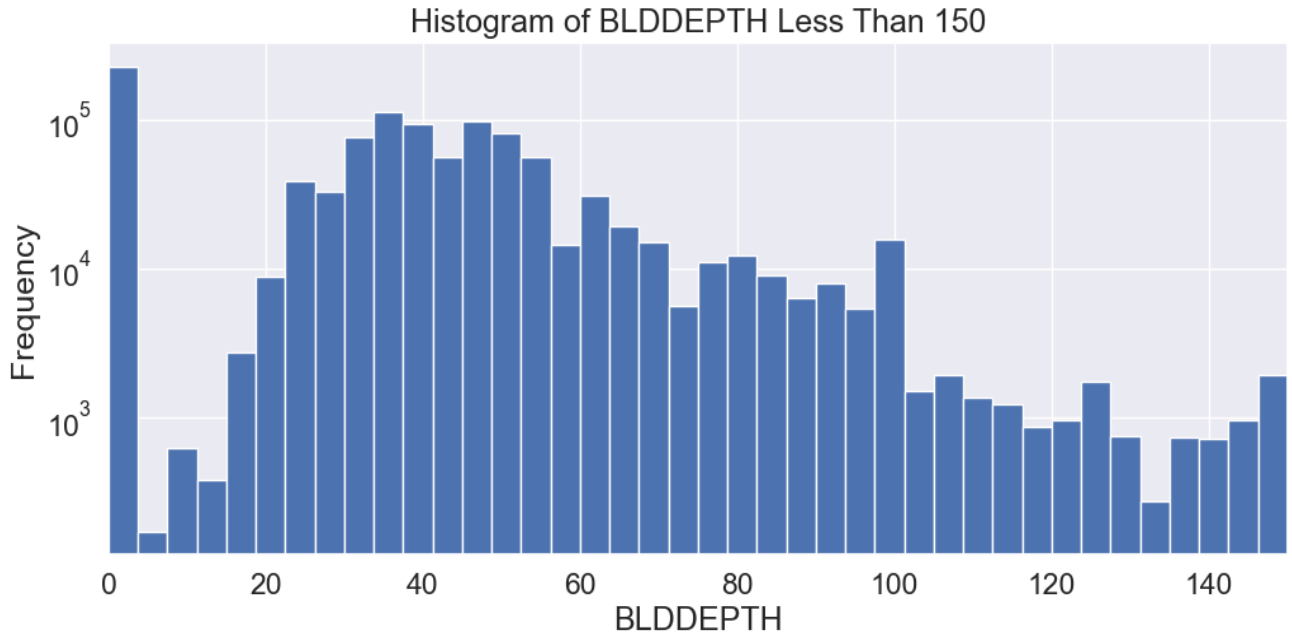
- **Description:** Building width of properties. We can observe a **high amount of frequency around 0** and a big drop after that, showing there is missing data in this field. There is another high amount of frequency when building width is around 20, the frequency begins to drop from 20 to 150.



(24) Field Name: BLDDEPTH

a. **Visualization:** Histogram of BLDDEPTH with a **range of x in [0, 150],** which covers **98.84%** of the property records.

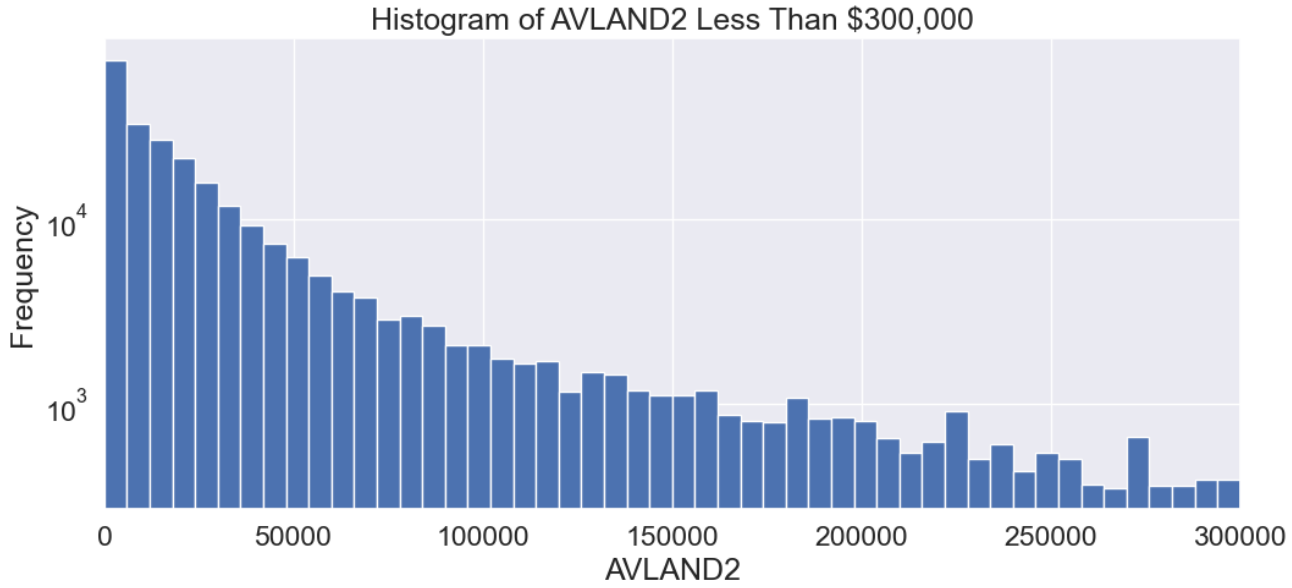
- **Description:** Building depth of properties. We can observe a **high amount of frequency around 0** and a big drop after that, showing there is missing data or zero value in this field. There is another high amount of frequency when building depth is around 40, the frequency begins to drop from 40 to 150.



(25) Field Name: AVLAND2

a. **Visualization:** Histogram of AVLAND2 with a **range of x in [0, $300,000],** which covers **24.15%** of the property records.

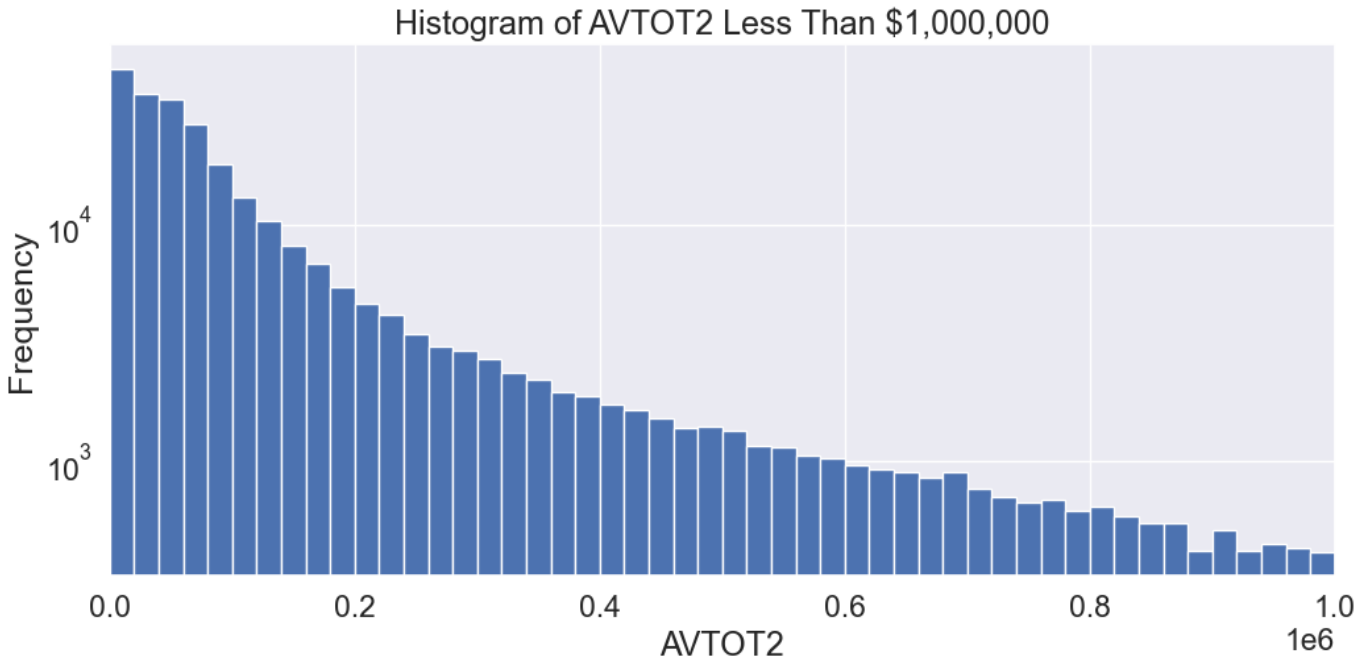
- **Description:** Transitional land value of properties. We can observe a drop from 0 to $300,000.



(26) Field Name: AVTOT2

a. **Visualization:** Histogram of AVTOT2 with a **range of x in [0, $1,000,000],** which covers **24.19%** of the property records.

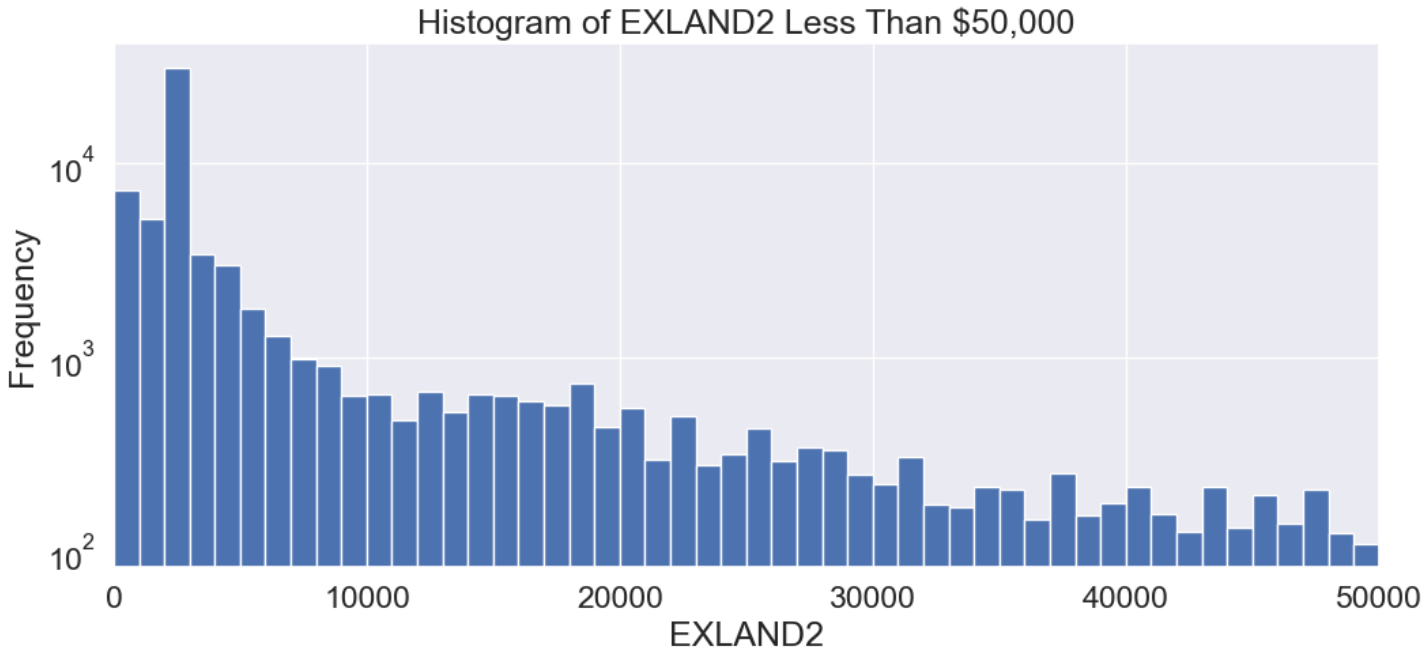
- **Description:** Transitional total value of properties. We can observe a drop from 0 to $1,000,000.



(27) Field Name: EXLAND2

a. **Visualization:** Histogram of EXLAND2 with a **range of x in [0, $50,000],** which covers **6.42%** of the property records.

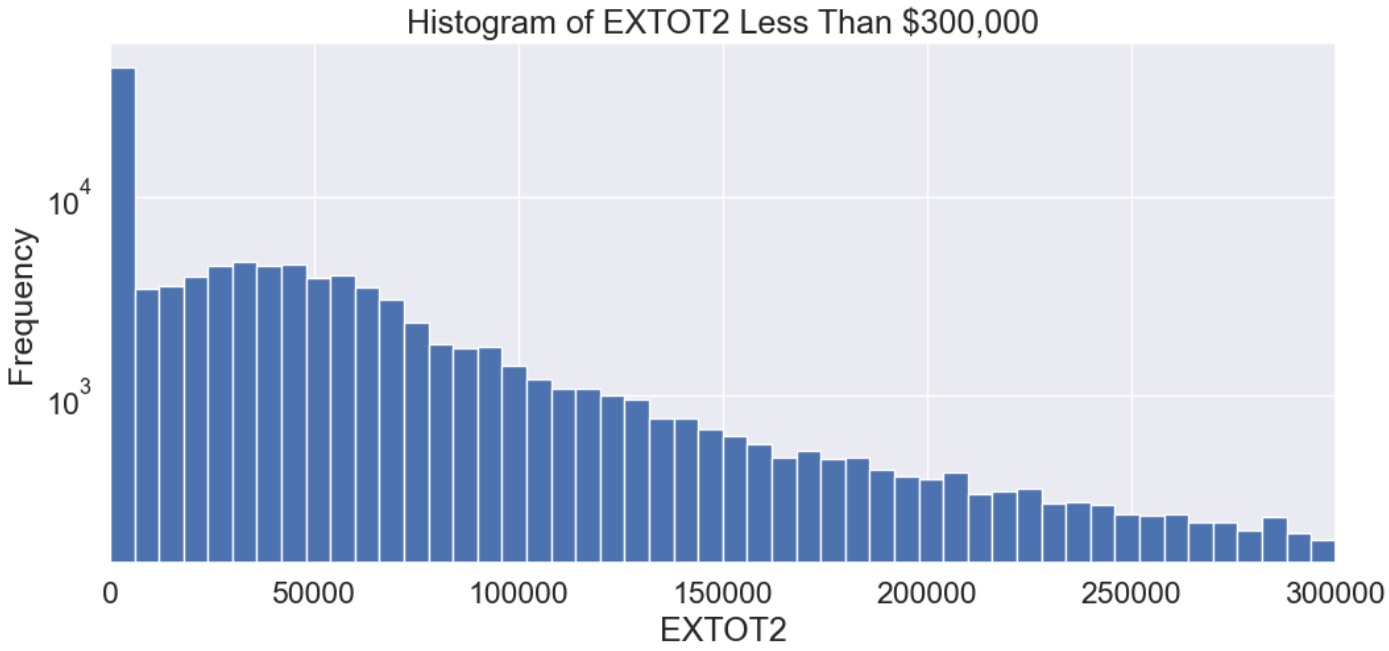
- **Description:** Transitional exemption land value of properties. We can observe a drop from 0 to $50,000.



(28) Field Name: EXTOT2

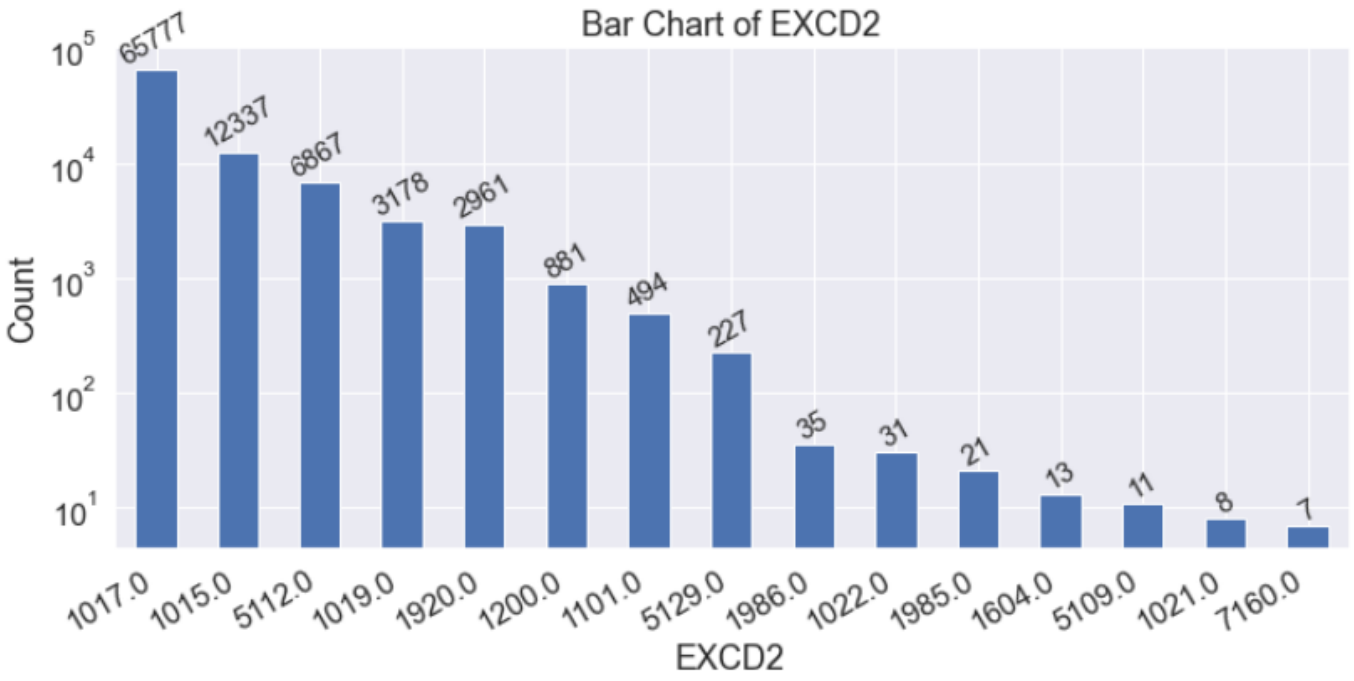
a. **Visualization:** Histogram of EXTOT2 with a **range of x in [0, $300,000],** which covers **10.58%** of the property records.

- **Description:** Transitional exemption total value of properties. We can observe a drop from 0 to $300,000.



(29) Field Name: EXCD2

- **Visualization:** Bar Chart of EXCD2. The chart selects top **15** field values of EXCD2.
- **Description:** Exemption code 2 in each record/property. The most common building class category shown in records is **1017**, with total amount of 65,777.



(30) Field Name: PERIOD

- **Visualization:** This record field has ONLY ONE VALUE. Therefore, we don't need a histogram/distribution for this field.
- **Description:** this field is about assessment period of property information with only one VALUE 'FINAL'.

(31) Field Name: YEAR

- **Visualization:** This record field has ONLY ONE VALUE. Therefore, we don't need a histogram/distribution for this field.
- **Description:** this field is about assessment year of property information with only one VALUE '2010/11'.

(32) Field Name: VALTYPE

- **Visualization:** This record field has ONLY ONE VALUE. Therefore, we don't need a histogram/distribution for this field.
- **Description:** this field is about value type of property information with only one VALUE 'AC-TR'.