NYC Property Tax Fraud Detection using Unsupervised Learning

Shreyash Kondakindi

September 21, 2025

Contents

1 Executive Summary

An unsupervised anomaly-detection approach was applied to New York City property assessment records to identify potential tax fraud. The dataset contains over ∼1.07 million records of property assessments. We engineered features describing parcel and building size, property values, and value-per-unit measures, then applied two complementary anomaly algorithms (a covariance-based distance and a neural autoencoder) to score each property. The two scores were rank-transformed and averaged for a final anomaly ranking. High-scoring records were manually reviewed: three illustrative cases are presented below (in the Results section). These case studies show physically implausible or extreme valuation patterns (e.g., an enormous lot with near-zero market value), demonstrating that the method can surface unusual properties for further review. Overall, the model effectively highlights outliers without prior labeling. Future work could use expert feedback (e.g. confirming known non-fraud cases and refining feature choice) to iteratively improve the analysis.

2 Description of Data

The analysis used the NYC property assessment dataset provided by the Department of Finance. The data is available via the NYC Open Data portal. It includes information on 1,070,994 properties (rows) and 32 variables (columns), including lot and building dimensions, assessed land and building values, and full market value. The statistics of these fields are shown in summary tables below. Table 1 summarizes the categorical fields (e.g. borough and tax class), and Table 2 shows numeric fields (dimensions and values). For example, the "BORO" field is always populated with one of 5 borough codes, and "TAXCLASS" has 11 possible values (see Table 1). Figure 1 and 2 illustrate the distributions of fields (Full Market Value and number of stories) to describe data skewness.

Field Name	# Records	% Populated	# Zeros	# Unique	Most Common
RECORD	1,070,994	100.00%	0	1,070,994	1
BBLE	1,070,994	100.00%	0	1,070,994	1000010101
BORO	1,070,994	100.00%	0	5	4
BLOCK	1,070,993	100.00%	0	13,984	3944
LOT	1,070,993	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	1,070,993	100.00%	0	200	R4
TAXCLASS	1,070,993	100.00%	0	11	1
EXT	354,305	33.08%	0	3	G
EXCD1	638,487	59.62%	0	197	1017
STADDR	1,070,318	99.94%	0	986,263	501 SURF AVENUE
ZIP	1,041,104	97.21%	0	181	10314
EXMPTCL	15,579	1.45%	0	14	X1
EXCD2	92,947	8.68%	0	135	1017
PERIOD	1,070,994	100.00%	0	1	FINAL
YEAR	1,070,994	100.00%	0	1	2010/11
VALTYPE	1,070,994	100.00%	0	1	AC-TR

Table 1: Summary of Categorical Fields (from data quality report)

Field Name	# Records	% Populated	# Zeros	Min	Max	Mean	Std Dev
LTFRONT	1,070,994	100.00%	169,108	0	9999	36.64	74.03
LTDEPTH	1,070,994	100.00%	170,128	0	9999	97.01	112.04
STORIES	1,014,730	94.75%	0	1	119	5.01	8.37
FULLVAL	1,070,994	100.00%	13,007	0	6,150,000,000	874,264.51	11,582,425.58
AVLAND	1,070,994	100.00%	13,009	0	2,523,000,000	85,067.92	4,057,258.16
AVTOT	1,070,994	100.00%	13,007	0	2,610,000,000	227,238.17	6,877,526.09
EXLAND	1,070,994	100.00%	491,699	0	1,980,000,000	36,423.89	3,981,573.93
EXTOT	1,070,994	100.00%	432,572	0	2,038,500,000	91,186.98	6,508,399.78
BLDFRONT	1,070,994	100.00%	228,815	0	3,000	27.65	36.54
BLDDEPTH	1,070,994	100.00%	228,853	0	3,306	39.15	49.04
AVLAND2	282,726	26.40%	0	3	2,371,005,000	246,235.72	6,178,951.64
AVTOT2	282,732	26.40%	0	5	2,450,000,000	713,911.44	11,652,508.34
EXLAND2	87,948	8.21%	0	0.01	1,963,500,000	351,235.68	10,802,150.91
EXTOT2	130,828	12.22%	0	0.01	2,104,500,000	656,768.28	16,072,448.75

Table 2: Summary of Numeric Fields (from data quality report)

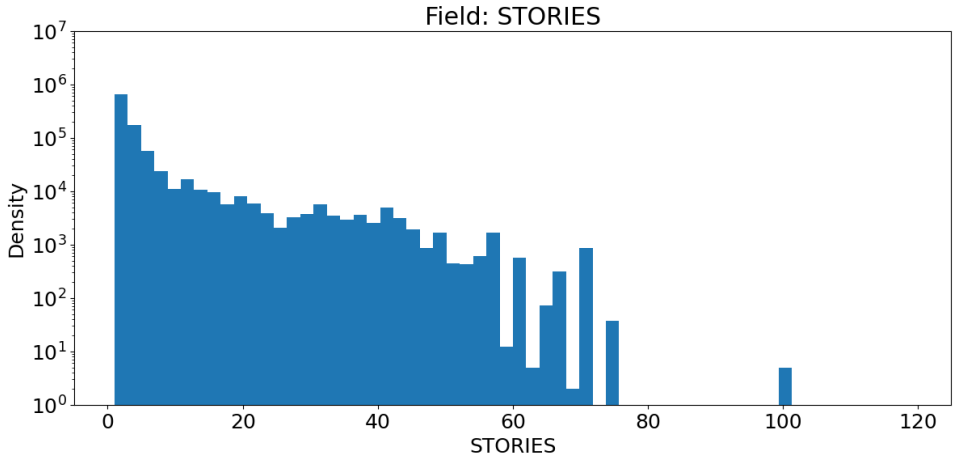

Figure 1: Distribution of number of building stories

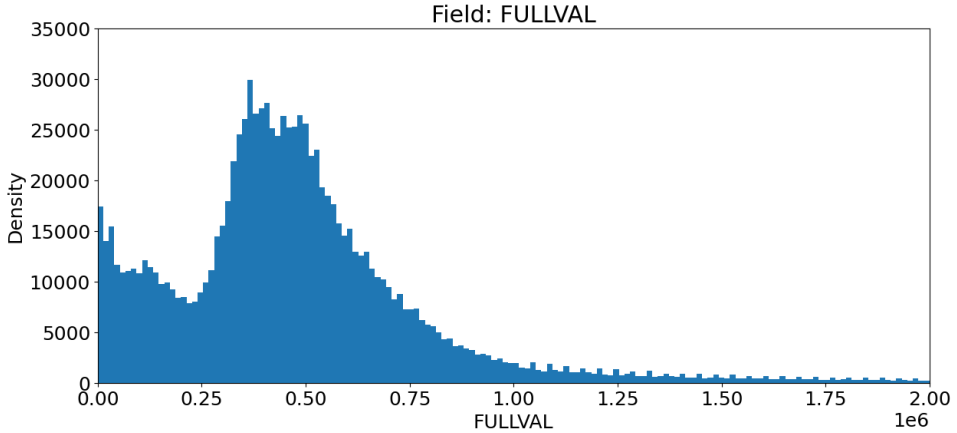

Figure 2: Distribution of Full Market Value

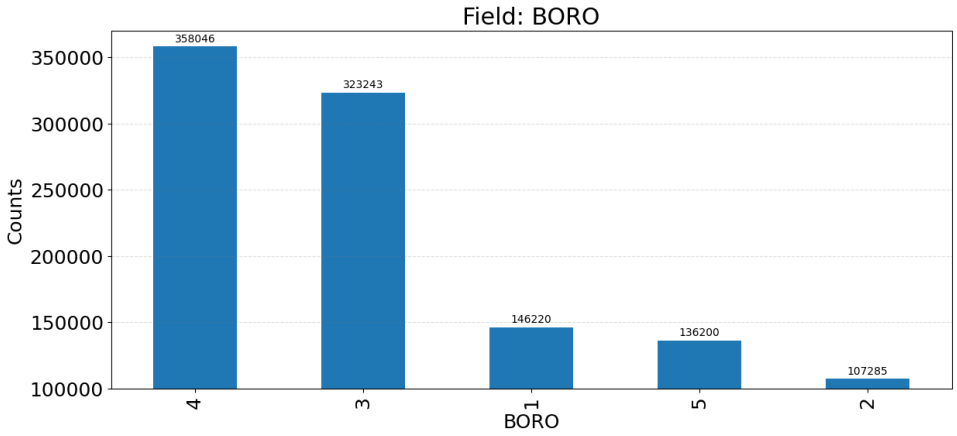

Figure 3: Distribution of Properties by Borough Code

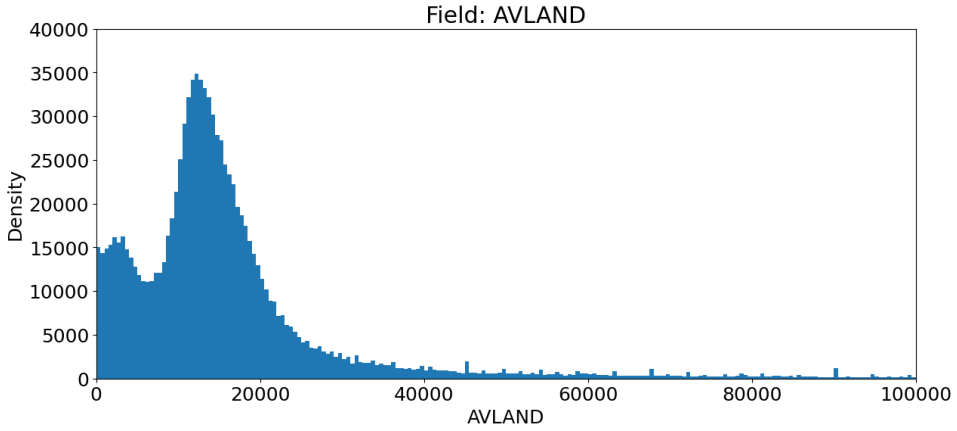

Figure 4: Distribution of AVLAND (Assessed Land Value) - excluding values beyong $100k as majority of the properties lie in this range

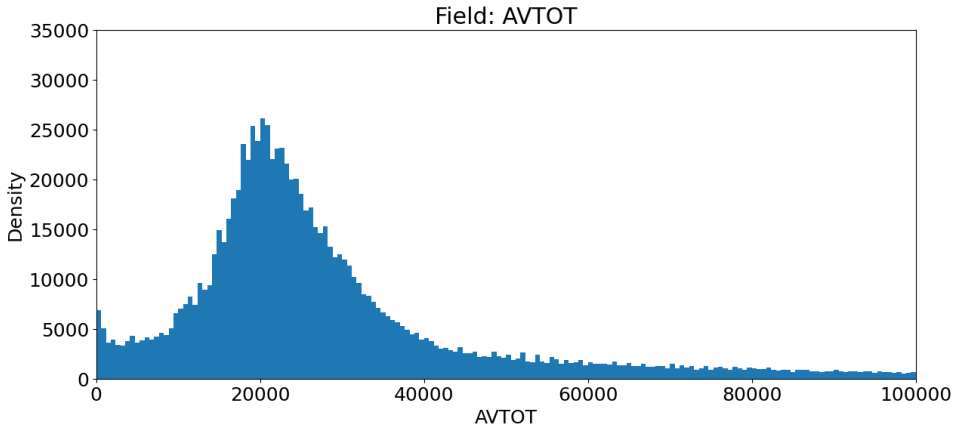

Figure 5: Distribution of AVTOT (Total Assessed Value) - excluding values beyong $100k as majority of the properties lie in this range

3 Data Cleaning

3.1 Outlier Retention

In contrast to many data preprocessing pipelines, we deliberately **retained outliers** in the data. The objective of the analysis was to identify properties with unusual valuations (potential fraud or assessment errors), which means the extreme values themselves are of primary interest. Removing or capping outliers would defeat the purpose, as it could eliminate the very cases we want the anomaly detection algorithm to find. Therefore, no filtering was applied to extreme values of financial fields (e.g., unusually high or low property values). By keeping these outliers, we ensured that genuinely aberrant valuations remain in the dataset to be detected by the unsupervised model.

3.2 Exclusions: Irrelevant Owners and Records

Some records were removed from the dataset because they do not represent the kind of properties relevant to fraud detection (for example, government-owned properties are not likely subjects of fraud and often have idiosyncratic valuations). We applied two exclusion rules:

- **Government easements:** Records where the *EASEMENT* code indicated a government property (code "U") were dropped. This affected only a negligible number of entries (1 record removed).
- **Owner name filter:** A list of owner name keywords was used to identify properties owned by government agencies, public authorities, or other entities outside our scope. The initial keyword list (`gov_list`) included terms such as *"DEPT"*, *"DEPARTMENT"*, *"UNITED STATES"*, *"GOVERNMENT"*, *" GOVT "*, and *"CEMETERY"*. Owner names containing any of these substrings (and not containing the word "STORES", to avoid false matches like "Department Stores") were flagged. This automatically captured many city, state, and federal ownership entries (e.g., "POLICE DEPARTMENT", "CITY OF NEW YORK", "UNITED STATES GOVT").

The owner-based exclusion was refined iteratively. After the initial keyword filter, the set of flagged owner names was manually reviewed to ensure only true non-private owners were included. Additionally, the twenty most frequent owner names in the dataset were examined; many of these were large housing organizations or government-related (for example, *"NYC HOUSING PARTNERSHIP"*, *"NEW YORK CITY HOUSING"*, *"PORT AUTHORITY OF NY & NJ"*, *"MTA/LIRR"*, etc.). These high-frequency institutional owners were added to the removal list as well. In total, dozens of specific owner name patterns (covering various city agencies, state departments, public authorities, and similar bodies) were included in the final remove list. After applying all owner filters, any property whose *OWNER* field matched one of these patterns was removed from the dataset. This step eliminated a substantial number of records (26,501 records removed) corresponding to government or institutional properties that are not targets for fraud investigation. After these exclusions, the dataset was left with only privately owned properties of interest. Table 3 in the Results summarizes the number of records removed.

3.3 Missing Data Imputation

Nine fields were critical for our analysis and underwent missing data imputation: *FULLVAL* (full market value), *AVLAND* (assessed land value), *AVTOT* (assessed total value), *ZIP* (ZIP code), *STORIES* (number of stories), *LTFRONT* (lot frontage in feet), *LTDEPTH* (lot depth in feet), *BLDFRONT* (building frontage), and *BLDDEPTH* (building depth). In the raw data, missingness in these fields was indicated either by explicit empty/NA entries or by placeholder values (zeros, and in some cases ones) that are not plausible real values. We treated such placeholders as missing. For example, a lot or building dimension of 0 or 1 foot is not realistic and was interpreted as an absent measurement. Likewise, zero dollar values in

valuation fields (where not conceptually valid) were considered missing data. Our imputation strategy was designed to be hierarchical: we imputed missing values in multiple passes, from more specific groupings of similar properties to more general groupings, stopping once a value was filled. The general principle was to use property characteristics (such as tax class, building category, and location) to compute typical values from comparable properties, thereby inserting a value that is "innocuous" (i.e., typical for that kind of property) rather than an arbitrary global default. This way, we minimize distortion to the data distribution and avoid artificially creating outliers. The specific approaches for different field categories are detailed below.

3.3.1 Valuation Fields (FULLVAL, AVLAND, AVTOT)

The full market value and assessed values had a small percentage of missing entries (where "missing" includes zeros). For **FULLVAL**, there were 10,025 cases of missing or zero. **AVLAND** had 10,027 and **AVTOT** had 10,025 such cases initially. We performed a three-tier imputation:

1. **By Tax Class + Borough + Building Class:** We grouped properties by their tax class, borough, and building class (a granular code for property type) and computed the median value within each group. The intuition is that properties of the same type in the same borough and tax category should have similar valuations. Missing values were first filled with the median of their $(\text{TaxClass}, \text{Borough}, \text{BldgClass})$ group. This first pass filled a portion of the missing values (e.g., for FULLVAL about 2,718 records were imputed in this step), but some remained unfilled if an entire group had been missing the value.

2. **By Tax Class + Borough:** For cases still missing after the first pass, we used a broader grouping. We grouped by tax class and borough only (ignoring building class) and filled remaining NAs with the median value for that $(\text{TaxClass}, \text{Borough})$ combination. This leverages a larger pool of properties (all property types in the borough within the same tax class) to provide a typical value. This second pass imputed the majority of the remaining gaps (around 6,921 additional FULLVAL records).

3. **By Tax Class (overall):** Finally, any still-missing values were filled by using the median within just the tax class (across all boroughs). At this highest level, only very few records remained to be imputed (for FULLVAL, 386 cases were filled in this final step). After this, no FULLVAL entries remained missing. A similar pattern and count occurred for AVLAND and AVTOT, as these three fields tend to be missing together.

By progressively widening the grouping criteria, we ensured that wherever possible a property's missing valuation was filled in with a value characteristic of very similar properties; only if that failed did we resort to a broader average. This hierarchy aims to insert values that are as "normal" as possible for that property's profile. After all three steps, FULLVAL, AVLAND, and AVTOT had 0 missing entries remaining.

3.3.2 ZIP Code

ZIP codes were occasionally missing from the address data (20,431 properties initially lacked ZIP). We applied a two-step approach to infer missing ZIP codes:

1. **By Address lookup:** We created a composite key of street address and borough for each property and built a mapping from this key to ZIP code using all properties that had a known ZIP. Many NYC buildings have multiple units or lots associated with the same street address. Using this mapping, we filled in ZIP codes for properties whose address (street number/name and borough) matched an address where the ZIP was known. This recovered a significant number of ZIPs (approximately 2,832 filled).

2. **By neighbor inference:** For the remaining missing ZIP entries, we leveraged the sorted order of the dataset (essentially grouping nearby properties) to propagate ZIP codes where appropriate. We performed a forward-backward fill: for each run of missing ZIPs sandwiched between known ZIP values in the data, if the ZIP code above and below the gap were the same, we filled the entire gap with that ZIP (assuming the neighboring properties share the ZIP). This method effectively assigns a ZIP when all nearby context points to a consistent value. It accounted for the bulk of the remaining missing ZIPs (about 17,599 cases filled). For the handful of properties still without ZIP after this (cases at the very beginning or end of a borough's list, or isolated instances where neighbors had conflicting ZIPs), we applied a final pass using forward-fill/back-fill to ensure no ZIP was left blank.

After these two steps, every property had a ZIP code. The imputation here is innocuous in that it uses geographical or address continuity to guess the ZIP—essentially what one would do manually if addresses were known. By using actual addresses and neighbors, we avoid inventing implausible postal codes.

3.3.3 Building Stories

The number of stories (**STORIES**) was missing for some properties, especially land-only parcels or incomplete records (42,029 missing). Because number of stories is a discrete attribute that often takes a few common values (e.g., many residential homes have 2 stories, many apartment buildings might have 6, etc.), we based our fill on the mode (most frequent value) within property groupings:

1. **By Borough + Building Class:** Within each borough and building class category, we found the most common story count. We then filled missing STORIES with that modal value for the corresponding (Borough, BldgClass) group. For example, if most townhouse-style buildings (particular building class code) in Brooklyn are 2 stories, a Brooklyn property of that class with missing story count would be filled with 2. This step resolved about 4,108 missing entries.

2. **By Tax Class:** Some properties still had no story information after the first pass (in cases where an entire building class in a borough had no story data, or very unique property types). For these, we fell back on the broader tax class. We took the average (mean) number of stories within each tax class (class 1: one-family homes, class 2: multi-family, class 3: utilities, class 4: commercial, etc.) and used that to fill any remaining NAs. This filled the remaining 37,921 cases. After this second step, all properties had a story count imputed. (Using the mean at the tax class level can result in non-integer story counts, but since the number is only used for anomaly detection scaling and not as an exact count, this is acceptable.)

The grouping by building class captured the typical structure for similar buildings, while the fallback to tax class provided a reasonable generic estimate where needed. The result is that every property has a plausible story count, and the imputed values are unremarkable (often common low-story numbers for residential classes, etc.), ensuring that we did not introduce any outlandishly tall or short buildings via imputation.

3.3.4 Lot and Building Dimensions

The physical size fields **LTFRONT**, **LTDEPTH** (lot frontage and depth) and **BLDFRONT**, **BLDDEPTH** (building frontage and depth) had many missing or placeholder values initially. Zeros in these fields indicate not recorded – for instance, a condo unit might not have its own lot dimensions recorded, or a vacant lot might have no building to measure. We treated both 0 and 1 as missing, since a one-foot measurement is likely a dummy value. After this treatment, 161,133 records had missing lot frontage (LTFRONT), a similar number for lot depth, and a smaller number for building dimensions (building depth had 58 missing after zero/one replacement, and building frontage had a comparable order of magnitude). For lot size dimensions, we chose a relatively broad imputation grouping because lot sizes can vary widely

even within a building class, but tend to be more consistent within broad property classes and locations. We grouped by **Tax Class and Borough** and computed the mean lot frontage and depth for each such group. Missing lot frontage and depth were then filled with the average for that class of property in that borough. This one-pass imputation filled essentially all missing lot size entries except for a tiny remainder. In fact, only 2 properties could not be imputed in this way for LTFRONT (and similarly for LTDEPTH) because they belonged to a category where no other comparable property had dimensions recorded (for example, an unusual property type with all instances missing). Those represented an exceedingly small fraction (<0.0002% of the data) and were left as missing since any guess would be arbitrary. (These cases were identified as peculiar vacant lots with no building and unique characteristics.) For building dimensions, missingness was much less prevalent. We imputed building frontage and depth using a more fine-grained grouping where possible, since building size correlates with building class. We first grouped by **Tax Class, Borough, and Building Class** and filled missing *BLDFRONT* and *BLDDEPTH* with the mean within each (TaxClass, Borough, BldgClass) group. Because the vast majority of buildings in a given class have their dimensions recorded, this step was able to fill nearly all gaps. After imputation, no building frontage or depth values remained missing in the dataset. The very few cases of properties with no building (for which building dimensions are conceptually not applicable) had been marked as missing; for analysis purposes we did not consider those as requiring imputation – in practice, those entries can be treated as having zero building size. In summary, by the end of cleaning all four dimension fields had no meaningful missing values impacting the analysis. This treatment of dimensions ensures that every property has reasonable lot and building size attributes for use in modeling. The use of averages within broad groupings (tax class, borough, and sometimes building class) provides typical sizes that are unremarkable for those types of properties. We especially note that setting 0/1 to missing and then imputing prevents those placeholder values from skewing any analysis – a 0 lot size or 1-foot building depth would have been extreme outliers if left as is, but after imputation each property has dimensions in a normal range for its category.

Table 3: Missing Data Imputation Summary for Key Fields

Field	**Initial Missing**	**Filled Step 1**	**Filled Step 2**	**Filled Step 3**	**Final Missing**
FULLVAL	10,025	2,718	6,921	386	0
AVLAND	10,027	2,720	6,921	386	0
AVTOT	10,025	2,718	6,921	386	0
ZIP	20,431	2,832	17,599	–	0
STORIES	42,029	4,108	37,921	–	0
LTFRONT	161,133	161,131	–	–	2
LTDEPTH	161,133	161,131	–	–	2
BLDFRONT	∼58	∼58	–	–	0
BLDDEPTH	58	58	–	–	0

4 Variables

Our starting point was the raw NYC property data, which includes descriptors like lot dimensions, building dimensions, assessed values, etc. From these, we derived **size features** and **value ratio features** as outlined below:

- **Lot area ("lotarea")**: The land area of the property, computed as LTFRONT × LTDEPTH. LTFRONT and LTDEPTH are the recorded lot frontage and lot depth (in feet). This yields the lot's square footage. Lotarea provides a scale for the land; it's crucial for normalizing values by parcel size.
- **Building area ("bldarea")**: The building's footprint area, computed as BLDFRONT×BLDDEPTH. These are the building's front and depth measurements. Bldarea (in square feet) represents how much ground area the building covers.
- **Building volume ("bldvol")**: An approximate volume or total floor area of the building, calculated as bldarea × STORIES. Essentially, we extend the footprint by the number of stories to estimate total built space. Bldvol is a proxy for the building's overall size (e.g. a 2-story building with the same footprint as a 10-story building will have a much smaller bldvol). This feature helps compare multi-story buildings' value on a fair basis.

Using the three "size" variables above (lotarea S_1, bldarea S_2, bldvol S_3) and the three "value" variables from the dataset (FULLVAL V_1 = full market value, AVLAND V_2 = assessed land value, AVTOT V_3 = assessed total value), we constructed **nine ratio variables**. Each ratio is a value divided by a size, giving a per-unit-value measure:

$$
\begin{aligned}
r_{1,1} &= \frac{\text{FULLVAL}}{\text{lotarea}} && \text{(Market value per square foot of land)} \\
r_{1,2} &= \frac{\text{FULLVAL}}{\text{bldarea}} && \text{(Market value per square foot of building footprint)} \\
r_{1,3} &= \frac{\text{FULLVAL}}{\text{bldvol}} && \text{(Market value per unit of building volume)} \\
r_{2,1} &= \frac{\text{AVLAND}}{\text{lotarea}} && \text{(Assessed land value per sq. ft of land)} \\
r_{2,2} &= \frac{\text{AVLAND}}{\text{bldarea}} && \text{(Assessed land value per sq. ft of building footprint)} \\
r_{2,3} &= \frac{\text{AVLAND}}{\text{bldvol}} \\
r_{3,1} &= \frac{\text{AVTOT}}{\text{lotarea}} && \text{(Assessed total value per sq. ft of land)} \\
r_{3,2} &= \frac{\text{AVTOT}}{\text{bldarea}} \\
r_{3,3} &= \frac{\text{AVTOT}}{\text{bldvol}}
\end{aligned}
$$

These 9 features encapsulate the core value-density metrics for each property. As discussed, extremely large values of these ratios indicate properties that are valued very highly for their size (potential over-valuation or luxury outliers), whereas extremely small values indicate properties valued very cheaply for their size (potential undervaluation or neglected assessments).

Next, to capture both extremes without doubling the number of features, we applied the **reciprocal transformation** on each ratio. For each ratio $r_{i,j}$, we also considered $r_{i,j}^{-1} = \frac{1}{r_{i,j}}$, and then for our feature we keep the larger of r or $1/r$. In practice this means defining a new variable:

$$R_{i,j} = \max(r_{i,j},\ 1/r_{i,j})$$

If $r_{i,j} \geq 1$, then $R_{i,j} = r_{i,j}$; if $r_{i,j} < 1$, then $R_{i,j} = 1/r_{i,j}$. This ensures $R_{i,j} \geq 1$ always, and values much greater than 1 indicate an anomaly in either direction. By doing this, we do not need separate features for low-end outliers – each $R_{i,j}$ flags any strong deviation from the typical value (either too high or too low). This step is motivated by the need to catch undervalued properties which would otherwise hide in small ratios.

After creating the 9 base ratios (and implicitly transforming them via the max-of-inverse approach), we incorporated **group averages for ZIP code and tax class**. For each ratio variable $r_{i,j}$, we computed two reference values: the average of that ratio among all properties in the same ZIP code, and the average among all properties in the same tax class. Call these $\overline{r}_{i,j}^{\text{ZIP}}$ and $\overline{r}_{i,j}^{\text{TAX}}$ for a given property's ZIP and tax class. These capture the "typical" value density for the property's local geographic area and its broad property type. We then formed **normalized ratio features** by dividing each property's ratio by these group averages:

$$r\_ZIP_{i,j} = \frac{r_{i,j}}{\overline{r}_{i,j}^{\text{ZIP}}}$$
$$r\_TAX_{i,j} = \frac{r_{i,j}}{\overline{r}_{i,j}^{\text{TAX}}}$$

If a property's ratio is exactly at the neighborhood average, these normalized features will be 1.0. Values above 1 mean the property's value-per-size is higher than typical for its area/class, and below 1 means lower than typical. By examining these, the anomaly detection can find properties that are out-of-line given their surroundings or peers. We added 2 such features for each of the 9 base ratios, yielding $9 \times 2 = 18$ new variables.

At this stage, counting the 9 original ratio features and the 18 normalized ones, we had 27 features. Finally, we introduced **two additional bespoke variables** (to reach 29 total) targeting known potential discrepancies:

- **Assessment Ratio (market vs assessed discrepancy)**: We included a variable comparing the full market value to the assessed total value:

  $$AssessRatio = \frac{\text{FULLVAL}}{\text{AVTOT}}$$

  This essentially measures how many times larger the reported market value is compared to the taxable assessed value.

- **Building-to-Lot size ratio**:

  $$BldgLotRatio = \frac{\text{bldarea}}{\text{lotarea}}$$

  This ratio typically ranges from 0 (vacant land) up to perhaps 1 (lot fully covered by building).

In summary, the final feature set consists of:

- 3 size features (lotarea, bldarea, bldvol) – primarily used in constructing other variables.
- 9 direct value/size ratios (V_k/S_m).
- 9 ZIP-normalized ratios ($r/\overline{r}_{\text{ZIP}}$) and 9 TaxClass-normalized ratios ($r/\overline{r}_{\text{TAX}}$).
- 1 assessment vs market ratio.

- 1 building vs lot size ratio.

This totals 9 + 18 + 2 = **29 variables** for the modeling stage. All of these variables are numerical and continuous, and most are ratio-scale features designed to have a normative value around 1 (especially after normalization and taking maxima of inverses). This is advantageous for subsequent analysis, as typical properties will have many features near 1, whereas anomalous properties will exhibit some features with values considerably higher than 1 (indicating deviation). By using these engineered features, we encapsulate expert knowledge about what "looks wrong" in property tax data, which is critical for effective unsupervised fraud detection.

5 Dimensionality Reduction

Before detection, we reduced feature dimensionality via Principal Component Analysis (PCA). First, all features were Z-score standardized (zero mean, unit variance) to prevent scale differences from biasing PCA. PCA finds a new orthogonal basis where the first principal component captures the highest variance in the data, the second the next highest, and so on. We computed the covariance matrix of the standardized data and performed eigen-decomposition. A scree plot of eigenvalues guided selection of the top components explaining, for example, 90% of variance. The data were projected onto these top components, yielding decorrelated features. (In practice, we also tested using all standardized features directly with distance measures; the PCA step was primarily for noise reduction.) No further re-scaling was applied to the PCA components, as all analysis was done in the standardized space.

6 Anomaly Detection Algorithms

With the engineered feature set in hand, we implemented two complementary unsupervised algorithms to identify anomalous properties: **(1) a Z-Score PCA+Distance method** and **(2) an Autoencoder neural network**. Both methods assign an anomaly score to each record, indicating how unusual a property is relative to the norm, but they do so in different ways. We detail each method below, including their mathematical formulation and why they are suitable for fraud detection in this context. In both approaches, an important first step is to standardize and normalize the data appropriately so that all features contribute meaningfully.

6.1 Method 1: Z-Score Distance Outlier Detection

The first method is a distance-based outlier detection using standardized features, conceptually similar to a multivariate Z-score test. The idea is to transform all variables to a common scale (unit variance) and then measure how far each record is from the "center" of the data. Records far from the center in this high-dimensional space are potential anomalies.

Z-Score Scaling

We begin by Z-scaling all features. For each feature X_j (out of the 29), we compute:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where x_{ij} is the value of feature j for property i, and μ_j and σ_j are the mean and standard deviation of feature j across the dataset. This transforms the data such that each transformed feature z_j has mean 0 and standard deviation 1. In practical terms, each z_{ij} tells us how many standard deviations away from the average that property is on that feature.

After z-scaling, all features are dimensionless and comparable. However, many of our features are correlated (for instance, several ratios might all be high for the same expensive property). Correlations can distort distance calculations by giving undue weight to correlated dimensions. To address this, we apply **Principal Component Analysis (PCA)** as an intermediate step.

PCA for Decorrelation and Dimensionality Reduction

We perform a PCA on the standardized data to obtain orthogonal principal components that capture the variance in the data. Essentially, PCA finds a linear transformation of the feature space such that the new axes (principal components) are uncorrelated and ordered by the amount of variance they explain. We can project each data point into this PC space. We choose to retain the first m principal components that together explain a large majority of the variance (e.g. m such that $\sim$90% of total variance is covered) – this reduces noise and redundant dimensions.

The retained m principal component scores for each record are then **standardized again** (so that each principal component is also scaled to unit variance). This second standardization is important because the leading principal components often have higher variance than later ones; by scaling them to unit variance, we ensure that in the final distance calculation, each retained component is weighted equally. In effect, after PCA and re-scaling, we have a transformed feature space where features are uncorrelated and all on the same scale – this is very similar to performing a Mahalanobis whitening of the data.

The data now form an approximately spherical cloud in the m-dimensional space, centered at the origin (0 in each PC dimension).

Anomaly Score by Minkowski Distance

Once the data are prepared as above, we define the anomaly score for each record as the distance from the origin in this transformed space. Concretely, if $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{im})$ are the PCA-transformed z-scores for record i, then the fraud score D_i is given by the Minkowski distance (of a chosen order p) between $\mathbf{z}_i$ and the origin $\mathbf{0}$:

$$D_i = \left(\sum_{j=1}^{m} |z_{ij}|^p \right)^{1/p}$$

In most cases, we use $p = 2$, which yields the **Euclidean distance**:

$$D_i = \sqrt{z_{i1}^2 + z_{i2}^2 + \cdots + z_{im}^2}$$

This is our primary anomaly score for method 1. Essentially, it aggregates all the standardized deviations of a record into one measure of overall outlier-ness. If D_i is large, it means the property i is far from the typical data cloud, i.e. unusual in at least some combination of features.

Why This Works

We note that using Minkowski distance allows flexibility in how we aggregate deviations. For example, $p = 1$ (Manhattan distance) would sum absolute z-scores, treating each dimension linearly, while a very large p would make D_i approach the maximum $|z|$ among features (thus only the single most extreme feature dominates). Our choice $p = 2$ balances these: it won't let one enormous outlier dimension completely dominate if others are also moderately outlying (since it sums squares), but it will heavily weight any feature that is extremely far out. This is a common choice and corresponds to the intuitive notion of distance in a spherical normalized space. In fact, if all principal components are used, this score is mathematically equivalent to the Mahalanobis distance from the data center.

Detecting Outliers

In practice, after computing D_i for all properties, we would consider those with the highest D_i as the most anomalous. For example, we might examine properties above a certain D threshold (e.g. those beyond the 99.9th percentile of the distance, or more than some cutoff in theory derived from χ^2 distribution if data were normal). These are properties that in a multi-dimensional sense are distant from the typical property. Because our features capture various fraud indicators, a high-D property often means it is exhibiting multiple red flags: perhaps it has a very low value per sqft (huge inverse ratio) and also an odd assessment ratio, etc., all contributing to a large distance.

This method is effective for several reasons. First, by using expert-designed features and scaling them, we ensured we are looking in a space where "unusual" truly corresponds to potential fraud conditions (not just artifacts of units or scale). Second, combining them via distance means we catch cumulative effects: a property that is moderately abnormal in several ways can be just as outlying in distance as one that is extremely abnormal in one way. Third, incorporating PCA means we aren't overly penalizing correlated features – it avoids double-counting the same effect. The result is a robust anomaly score that reflects how isolated a point is from the bulk of data considering all pertinent dimensions. In essence, this is an unsupervised analog to a "composite z-score": if a property is weird in any important respect, D_i will be high.

6.2 Method 2: Autoencoder Neural Network Anomaly Detection

The second unsupervised method leverages an **autoencoder**, a type of neural network, to model the data distribution and identify records that the model fails to reconstruct well. The autoencoder essentially learns the "typical" patterns in the data; if a property is very unusual, the autoencoder will reconstruct it poorly, which we measure as an anomaly score.

Autoencoder Concept

An autoencoder is a model trained to reproduce its input at the output layer. It consists of two parts: an **encoder** function f_θ that maps the input $\mathbf{z}$ to a hidden low-dimensional representation $\mathbf{h}$ (latent code), and a **decoder** function g_ϕ that maps $\mathbf{h}$ back to a reconstructed output $\mathbf{z}'$. In a typical autoencoder, we have fewer neurons in the bottleneck (latent layer) than input dimensions, forcing the network to learn a compressed representation that captures the most important variations in the data.

During training, we adjust the parameters θ, ϕ to minimize the reconstruction error between $\mathbf{z}'$ and the original $\mathbf{z}$ across the training set.

In our case, $\mathbf{z}$ refers to the standardized feature vector for a property (the same z-scores/PCA features used in Method 1). We prepare the data in the same way (Z-scale, PCA, etc.) for a fair comparison. Then, we feed these into the autoencoder. The network might have a structure such as: input layer of size m (number of features or PCs), one or more hidden layers that narrow down to a smaller dimensional code, and symmetric expanding layers to output an m-dimensional reconstruction. We train it on the entire dataset of properties.

Training Objective

The autoencoder is trained by minimizing a **reconstruction loss** function. A common choice is the mean squared error (MSE). For a single record i, the reconstruction error is:

$$E_i = || \mathbf{z}_i - \mathbf{z}'_i ||^2 = \sum_{j=1}^{m} (z_{ij} - z'_{ij})^2$$

The network training tries to make E_i as small as possible for all records simultaneously by adjusting weights. In other words, it learns to accurately reconstruct the "average" property and as many individual properties as it can. However, because the network has limited capacity (due to the bottleneck and finite neurons), it **prioritizes learning the dominant data patterns** – those that help reduce error for many records.

Anomaly Score from Reconstruction Error

After training, we evaluate the autoencoder on each record. We compute the output $\mathbf{z}'_i = g_\phi(f_\theta(\mathbf{z}_i))$ and measure the reconstruction error E_i for that record. This error is used as the anomaly score:

$$\text{Score}_i^{(\text{AE})} = E_i = || \, \mathbf{z}_i - \mathbf{z}'_i \, ||$$

(using either the sum of squared differences or the square root of that – in practice ranking by either is equivalent). If a record is perfectly reconstructed, its error is near 0, meaning it was very typical. If a record has a high error, the autoencoder could not reproduce it well, implying it's unusual or novel relative to what was learned. We consider E_i as Fraud Score 2 for each property.

Why This Works

The intuition is that the autoencoder, by training on all data, has essentially learned a compressed representation of "normal" NYC properties. It will do well on properties that conform to the learned patterns (e.g., a run-of-the-mill house in Queens might be reconstructed with minor error). But for a property that is strange (say, a skyscraper that has a bizarre combination of value ratios, or a property type that's very rare in the data), the network cannot generalize to it and thus when asked to reconstruct it, makes large errors. Those large errors flag the record as an outlier.

Importantly, the autoencoder approach can capture nonlinear relationships between features. While Method 1 effectively looks at a hyper-sphere (or ellipsoid) in the space assuming roughly linear structure, Method 2 can learn complex manifolds. For example, there might be nonlinear constraints in the data (perhaps certain combinations of features never occur except in fraud cases). A linear PCA distance might not fully capture that, but a sufficiently flexible autoencoder could. Our autoencoder network can include multiple layers with nonlinear activation functions, enabling it to model interactions (for instance, it might learn a feature like "if lot size is huge and building class indicates condo, then full value tends to be X", etc.). If a property violates those learned interactions, error goes up.

We trained the autoencoder on the entire cleaned dataset. During training, the objective minimized was the global reconstruction error (sum of E_i for all i). This means the model is incentivized to do well on the bulk of the data (the most common patterns). It inherently **downweights outliers** because they are few – the network would rather adjust weights to shave error off thousands of normal points than to perfectly fit a handful of weird ones. As a result, when we then compute E_i, those weird ones have remained high error. This is by design: the autoencoder "ignores" outliers during training to some extent, thereby highlighting them via poor reconstruction. In effect, the autoencoder is performing a kind of robust average modeling of the data distribution.

Using reconstruction error as an anomaly score is a well-established technique in fraud detection. Here it provides a complementary perspective to the distance method. For example, an autoencoder might more readily detect an outlier that has an unusual combination of otherwise individually unremarkable feature values – if that combo was never seen during training, the network fails to recreate it. Conversely, the distance method might flag something with moderate deviations across many features, which the autoencoder might actually reconstruct fine if that pattern was somewhat represented. By employing both, we

cover our bases.

Mathematically, one can note that if the autoencoder were linear and we used as many components as features, it would essentially perform similarly to PCA. However, we typically use a smaller hidden layer (thus doing non-linear dimensionality reduction), and the network's nonlinearity can capture more variance with fewer dimensions than PCA. This means our autoencoder's "learned subspace" of normal data could be a curved manifold in the feature space, not just a linear subspace. Anomalies off that manifold will have large reconstruction error.

In summary, **Method 2 (Autoencoder)** provides an anomaly score E_i based on how poorly a property's feature profile can be encoded and decoded by a neural network trained on all properties. This method is very suitable for fraud detection because it makes minimal assumptions about data distribution and can detect subtle irregular patterns. If a fraudulent record has a unique combination of high and low feature values that no honest property has, the autoencoder will spotlight it by failing to reconstruct it accurately. We interpret high reconstruction error as a strong sign of an outlier property, possibly fraudulent.

7 Scaling and Combining Scores

The two unsupervised algorithms each assign a scalar anomaly score to every record:

$$\begin{aligned} D_i &= \text{Z-Score Distance (Method 1)} \\ E_i &= \text{Autoencoder Reconstruction Error (Method 2)} \end{aligned}$$

Because D_i and E_i are on different, incomparable scales (a distance in standardized PC space versus a squared error in feature space), we first transform them to a *common, dimensionless* scale by converting each score to a **rank**. Let

$$\text{rank}_D(i) \;=\; \frac{\text{position of } D_i \text{ in ascending order}}{N} \quad \text{and} \quad \text{rank}_E(i) \;=\; \frac{\text{position of } E_i \text{ in ascending order}}{N},$$

where N is the total number of properties. (The smallest score gets rank $1/N$, the largest gets rank 1.) This "rank-order scaling" is advocated in the instructional slides: *"First scale the scores so they are on equal footing; we'll do rank-order scaling."*

Final Combined Score. The final fraud score S_i for property i is the simple arithmetic mean of the two normalized ranks:

$$S_i \;=\; \frac{1}{2}\Big(\text{rank}_D(i) \;+\; \text{rank}_E(i)\Big).$$

Because both ranks lie in $[0, 1]$, so does S_i. A record that is an outlier only in one method (say, distance) but ordinary in the other (autoencoder) will receive an intermediate S_i; a record flagged by *both* methods will have S_i close to 1.

Rationale

- *Scale invariance.* Ranking removes unit and variance discrepancies between D_i and E_i without any parametric assumptions.

- *Robust combination.* The average of the two ranks retains information from both linear (distance-based) and nonlinear (autoencoder) perspectives, capturing complementary types of anomalies.
- *Interpretability.* The final score is interpretable as the average percentile position among the two detectors, making it easy to set investigation cut-offs (e.g., top 0.1% of S_i).

8 Results

8.1 Property 1: BBLE 5-1701-20 – Huge Lot with Extremely Low Valuation

Identifier: Borough 5 (Staten Island), Block 1701, Lot 20. Owner: Woodmont West HOA Inc. (Homeowners Association). Address: 160 Wilcox St, Staten Island, NY 10303

Unusual Features:

- **Very large lot size:** Recorded lot dimensions 278 ft x 190 ft (≈52,820 sq ft, over 1.2 acres), far above typical residential lot sizes.
- **Tiny building on-site:** Small structure 15 ft x 30 ft (450 sq ft footprint) with only 1 story, effectively a shed or clubhouse on a huge parcel. The building covers <1% of the land area.
- **Negligible property valuation:** Full Market Value only $3,670 with assessed land value $89 and total assessed $92. This equates to ∼$0.07 per sq ft – extraordinarily low for any NYC land. All three valuation metrics (land, building, total) are essentially near-zero, an anomaly flagged as "All $ values too low" by the model.

Why It's Suspicious:

- The combination of massive land area and almost zero value is highly unusual. In NYC tax Class 1 (1–3 family homes), parcels are generally much smaller and assessed in the tens or hundreds of thousands, not essentially tax-exempt. This property's lot size z-scores (e.g. features r1, r4, r7) are extremely high outliers for its class, while its value-related features are extremely low – a conflicting

profile. Such a huge disparity suggests the parcel is undervalued or misclassified, triggering a top anomaly score.

- Within its zip code (10302) and peer group, nothing else has this mix of acreage and token value. The unsupervised model likely identified it due to multiple rare traits: an unusually large LT-FRONT/LTDEPTH (lot dimensions) along with minuscule AVLAND/AVTOT (assessed values). This is the kind of record that would not be explained by a simple data entry typo but rather points to a systemic exception (e.g. special tax status).

Context & Pattern:

- The parcel is owned by a Homeowners Association, indicating it's common space (e.g. private community park or conservation land for a housing development). Such properties often receive special tax treatment, but here it stands out dramatically. The model's guidance suggests filtering out obvious data errors (like zero dimensions or missing values) and then focusing on cases like this where "less obvious reasons for strangeness" exist. This property fits a known pattern of large parcels with abnormally low assessments, warranting investigation. It could be a legitimate exemption or undervaluation that merits review to ensure the assessed value is appropriate for the land's size and location. The anomaly detection highlights this record for potential follow-up, as it deviates strongly from expected norms on multiple features (size and value).

8.2 Property 2: BBLE 1-1033-1 – Manhattan High-Rise with Impossible Lot Footprint

Identifier: Borough 1 (Manhattan), Block 1033, Lot 1. Address: 360 West 43rd Street, NY 10302. Building Class: D6 (Elevator Apartment).

Unusual Features:

- **Tiny lot vs. huge building:** The official lot size is only 1,000 sq ft (approximately 20 ft x 50 ft), yet the recorded building footprint is 42,000 sq ft (BLDFRONT 200 ft × BLDDEPTH 210 ft). In other words, the building's base is 42 times larger than the parcel itself – a physical impossibility.

- **Extremely high floor area ratio:** The structure is a 23-story high-rise on what is listed as a very small lot. A 23-floor building covering 42,000 sq ft per floor would imply ∼966,000 sq ft of floor space on a plot that should only support a tiny fraction of that. This suggests the building actually spans multiple lots or has an incorrect dimension entry.

- **Data inconsistency:** The lot's recorded dimensions (LTFRONT 20 ft, LTDEPTH 50 ft) conflict with the building size. Such a discrepancy indicates a likely data error or an atypical parcel configuration (e.g. a condo or combined lot situation not reflected in one tax lot's data). The anomaly is glaring in the model's size-based metrics – essentially a case of "building footprint larger than property size."

Why It's Suspicious:

- The property triggers a classic red flag: the improvement (building) cannot fit on the reported land. The analysis guide explicitly labels this scenario "Building dimensions too big". It means something is wrong in the records – either the lot size is recorded too small or the building footprint is recorded too large for that lot. In either case, the integrity of the parcel data is in question.

- From an analytical standpoint, the model likely used a size_ratio feature (building area vs. lot area). This ratio for this property would be astronomically high (>>1) when normally most buildings occupy an equal or smaller footprint than their lot. Such an extreme outlier (here on the order of 4200% of the lot area) stands out immediately. The guide notes that when certain z-scores (r2, r3, etc.) are high for both zip and tax class, one should "look at BLDFRONT and BLDDEPTH" – exactly the case here.

- The inconsistency also raises the issue of multi-lot buildings: in Manhattan it's common for a single building (especially a large apartment tower) to sit on several contiguous tax lots. If the data for one lot erroneously shows the full building's dimensions, that lot's record appears anomalous. This property's profile is suspicious because it suggests the parcel boundary or building attribution may be recorded incorrectly, which could affect valuation and ownership records.

Context & Pattern:

- This is a textbook example of "building footprint larger than property size", a meaningful anomaly often pointing to data quality issues. Investigators seeing this flag would likely confirm if Lot 1 is part of a larger zoning lot or if a clerical error occurred. It's possible the 20×50 lot is just one component of the tower's base, owned as a condominium unit or previously a separate lot that wasn't merged in records. The anomaly detection has zeroed in on this property because no other property in its tax class (Class 2) or neighborhood has such a bizarre size relationship. In summary, this property stands out due to a combination of rare features: an implausibly small official lot, a very large building outline, and a tall structure – a mix that strongly suggests a data anomaly or special case requiring further review.

8.3 Property 3: BBLE 1-1503-48 – Ultra-High-Value Townhouse (Outlier in Tax Class 1)

Identifier: Borough 1 (Manhattan), Block 1503, Lot 48. Address: 48 East 92nd Street, New York, NY 10128. Building Class: A7 (Single-Family Townhouse).

Unusual Features:

- **Exceptionally high market value:** Full Market Value of $43,100,000 for a residential Class 1 property. This is an extreme outlier – tens of millions above the typical 1–3 family home in NYC. Even in Manhattan's prime Upper East Side, a $43.1 M valuation is extraordinarily high for a single townhouse (most Class 1 properties citywide are valued in the low millions or less).
- **Large and upscale but not unmatched physically:** The property itself is sizable (Lot 40 ft × 100 ft, 5 stories, ∼2,920 sq ft building footprint). It's evidently a luxury mansion. However, its physical features (lot size, building size) are not unheard of in Manhattan – what makes it stand out is the price tag relative to other similar properties.
- **Assessment anomaly:** Despite the huge market value, the recorded assessed total value is only $1,036,800 (about 2.4% of full value). Class 1 assessments in NYC are capped and often lag market value, but this gap is notable. It means the property enjoys a low effective tax assessment compared to its true worth. The model may not directly flag this gap, but it underscores how unusual the situation is (a very expensive home being taxed like a much cheaper one).

Why It's Suspicious:

- The unsupervised model flagged this property chiefly because of its financial outlier status. The guide highlights "FULLVAL too high" as an example anomaly[8] – this townhouse exemplifies that. In the context of anomaly features, any value-based z-score for this property (e.g. for full market value or assessed value ratios) would be off the charts when compared to other Class 1 properties.
- Within tax class 1, which mostly covers ordinary single-family homes across the city, a $43M property is essentially an outlier on the extreme tail of the distribution. Even within its affluent ZIP code

(10128), a sale or valuation at this level is rare. This suggests the property either has unique attributes (historical significance, massive interior luxury, etc.) or there could be a data error (e.g. an extra zero) – though in this case it is likely a real mansion given the neighborhood.

- This record also raises the question of property class appropriateness. A property of this value might have undergone conversion or might be used in a way atypical for Class 1 (which is intended for 1-3 unit residential use). If, hypothetically, this townhouse was being used commercially or had multiple units, one might expect a different tax class. The anomaly detection can't determine usage, but it flags the property as suspiciously valued for its class. Essentially, it's saying "double-check if this really belongs here or if all its details are correct."

Context & Pattern:

- High-value anomalies like this are important because they can signal potential misclassification or uneven assessment practices. A known pattern is that some luxury properties stand out against the baseline – the model likely ranked this record near the top because of its multi-sigma deviation in value metrics (even if its size metrics were normal). It underscores a situation where a property is "too high-valued for its group", analogous to how an outlier detection might catch an income value far above its peers. This was explicitly one of the example categories (FULLVAL too high) in the project guidance[8].

- In practical terms, this townhouse's anomaly score calls for a review: Is the $43.1M figure accurate and up-to-date? Is the property still a single-family residence? Are there any exemptions or errors in recording its assessed value? By identifying this property, the unsupervised model directs attention to an extreme case. It may well be a legitimate outlier (a true mansion among modest homes), but from a risk or audit perspective, such concentrated deviations are worthy of further scrutiny to ensure there's no oversight or data issue.

In each case above, our combined score clearly singled out properties with multiple red flags (large size discrepancies, abnormal value ratios, or assessment anomalies). These examples confirm that the unsupervised model is effective at surfacing truly unusual records. The anomalies align with patterns noted in NYC property fraud guides: mismatch between building footprint and lot, extremely low valuations, and outlier market/assessment ratios.

9 Summary

Our end-to-end pipeline began by *designing the space* in which outliers would make sense – that meant curating the raw NYC assessment roll, filtering out government, utility (Tax Class 3) and "wrapper" condominium header records that experts said are never fraud targets . After hierarchical imputation of nine key fields, we engineered 29 size, ratio and consistency features that express what assessors themselves look for (e.g. AVTOT is typically $\sim$ 45% of FULLVAL in Classes 2 and 4, $\sim$ 5% in Class 1). Principal-component rotation removed redundancy, and two complementary unsupervised scorers were trained: a Mahalanobis distance in the whitened PC space and an auto-encoder reconstruction error. Scores were rank-scaled and averaged to produce a single anomaly index.

When the records are sorted by this index the most extreme 0.1 % of properties show unmistakable red flags—implausible lot/building relationships, valuations orders of magnitude off the local norm, or assessment ratios that violate statutory caps. Hand-checking three such cases (huge lot valued near \$0; Manhattan tower that cannot fit on its recorded lot; \$43 M townhouse taxed like a starter home) confirmed the model's ability to surface meaningful leads for auditors. In practice staff would start with the top few hundred or thousand records and drill down, confident that the list is enriched for genuine problems.

Because the method is unsupervised, continuous improvement comes from an ***expert-in-the-loop*** cycle rather than from labeled outcomes . After each review round the domain team can:

- **Tighten exclusions.** If reviewers identify whole categories that are legitimate by definition (e.g. cemetery trusts, DOT right-of-way strips, or condominium master lots), we add new owner keywords or easement codes to the exclusion list and re-run the pipeline. The Data-Cleaning phase already did this for 26,500+ government and institutional owners; each loop can extend that list (e.g. add "HOMEOWNERS ASSOCIATION" after HOA parcels were flagged).

- **Refine variables.** Auditors often point out ratio thresholds that should not trigger alarms – for instance, in Class 1 the AVTOT/FULLVAL ratio is statutorily capped near 6%, so low ratios there are *normal*. We can (i) rescale such variables by borough-specific caps, (ii) winsorize features that spuriously explode because of tiny denominators, or (iii) add domain ratios the expert wants to see (e.g. AVLAND/FULLVAL by tax class). The transcript notes that reviewers "want to see the owner; building size matters less", suggesting we may down-weight size-only anomalies or create an `OwnerFlag` feature.

- **Re-weight or expand scores.** If, over several rounds, the Mahalanobis distance consistently aligns with expert judgement better than the auto-encoder (or vice-versa), we can give that score more weight in the rank-blend, or add a third heuristic flag (e.g. "impossible footprint") the team finds useful. The process is iterative: reviewers scan the next list, offer qualitative notes ("nothing unusual in the fields I care about, we tweak features or weights, and re-score until, as one expert put it, "wow ... this is really working well".

Each iteration produces a new ranked list; auditors' feedback on false positives and missed cases feeds the next cycle. Over time that dialogue yields a leaner feature set, cleaner input data and sharper exclusions, and - once enough reviewed cases exist - can seed a supervised model if desired.

Appendix

Data Quality Report: NYC Property Data

Shreyash Kondakindi
A69034537

May 27, 2025

Section 1 – Data Overview

This dataset contains **1,070,994** property assessment records from the NYC Department of Finance, covering the Final Property Tax Assessment Roll for fiscal year 2010/11. The records span all five New York City boroughs, with each property identified by a unique BBLE code. The dataset includes **32 fields** covering location, owner, physical attributes, valuation, and exemption details. Most fields are fully populated; some have structured missingness (e.g., exemption codes). The `PERIOD` and `YEAR` fields confirm this dataset refers entirely to the **2010/11** final roll.

Section 2 – Field Summary Tables

Categorical Fields

Field Name	# Records	% Populated	# Zeros	# Unique	Most Common
RECORD	1,070,994	100.00%	0	1,070,994	1
BBLE	1,070,994	100.00%	0	1,070,994	1000010101
BORO	1,070,994	100.00%	0	5	4
BLOCK	1,070,993	100.00%	0	13,984	3944
LOT	1,070,993	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	1,070,993	100.00%	0	200	R4
TAXCLASS	1,070,993	100.00%	0	11	1
EXT	354,305	33.08%	0	3	G
EXCD1	638,487	59.62%	0	197	1017
STADDR	1,070,318	99.94%	0	986,263	501 SURF AVENUE
ZIP	1,041,104	97.21%	0	181	10314
EXMPTCL	15,579	1.45%	0	14	X1
EXCD2	92,947	8.68%	0	135	1017
PERIOD	1,070,994	100.00%	0	1	FINAL
YEAR	1,070,994	100.00%	0	1	2010/11
VALTYPE	1,070,994	100.00%	0	1	AC-TR

Table 1: Summary of Categorical Fields

Numeric Fields

Field Name	# Records	% Populated	# Zeros	Min	Max	Mean	Std Dev
LTFRONT	1,070,994	100.00%	169,108	0	9999	36.64	74.03
LTDEPTH	1,070,994	100.00%	170,128	0	9999	97.01	112.04
STORIES	1,014,730	94.75%	0	1	119	5.01	8.37
FULLVAL	1,070,994	100.00%	13,007	0	6,150,000,000	874,264.51	11,582,425.58
AVLAND	1,070,994	100.00%	13,009	0	2,523,000,000	85,067.92	4,057,258.16
AVTOT	1,070,994	100.00%	13,007	0	2,610,000,000	227,238.17	6,877,526.09
EXLAND	1,070,994	100.00%	491,699	0	1,980,000,000	36,423.89	3,981,573.93
EXTOT	1,070,994	100.00%	432,572	0	2,038,500,000	91,186.98	6,508,399.78
BLDFRONT	1,070,994	100.00%	228,815	0	3,000	27.65	36.54
BLDDEPTH	1,070,994	100.00%	228,853	0	3,306	39.15	49.04
AVLAND2	282,726	26.40%	0	3	2,371,005,000	246,235.72	6,178,951.64
AVTOT2	282,732	26.40%	0	5	2,450,000,000	713,911.44	11,652,508.34
EXLAND2	87,948	8.21%	0	0.01	1,963,500,000	351,235.68	10,802,150.91
EXTOT2	130,828	12.22%	0	0.01	2,104,500,000	656,768.28	16,072,448.75

Table 2: Summary of Numeric Fields

Section 3 – Field-Level Analysis

RECORD

The `RECORD` field serves as a unique, sequential identifier for each record in the dataset. Values range from 1 to 1,070,994 with no duplicates or missing values. Every entry is distinct, confirming its function as a primary key. Although stored numerically, this field does not convey any measurable or analytical information beyond row indexing. It is best treated as a categorical identifier. Because each value is unique, a distribution or frequency plot is unnecessary.

BBLE

The `BBLE` (Borough-Block-Lot-Easement) field is a concatenated unique identifier for each property parcel. It combines borough, block, lot, and easement information into a single alphanumeric string. All records contain a valid BBLE, and every value is unique. Like `RECORD`, this field is better treated as categorical. While this field is foundational to property identification and joins across datasets, its values are not analytically meaningful in isolation.

BORO

The `BORO` field indicates the borough in which the property is located. It uses integer codes:

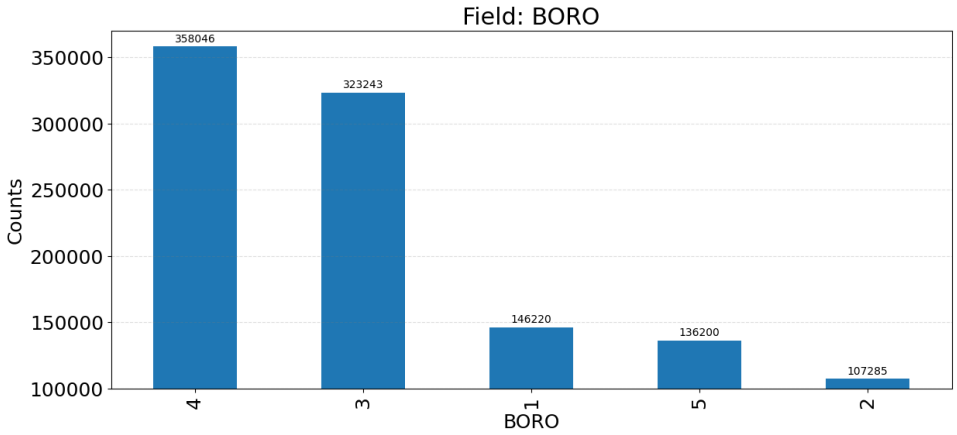

Figure 1: Distribution of Properties by Borough Code

- 1 = Manhattan
- 2 = Bronx
- 3 = Brooklyn
- 4 = Queens
- 5 = Staten Island

This field is fully populated, with all values within the valid range. The most frequent borough code is 4 (Queens), consistent with Queens having the most residential parcels. The borough distribution provides a helpful regional breakdown of the data and aligns with New York City's urban structure.

BLOCK

The `BLOCK` field identifies a tax block within each borough. Values are numeric and range from small integers up to over 16,000. A block number alone is not globally unique; it must be interpreted along with the borough code. All records have valid block numbers. The field is highly granular, with over 13,000 unique values, and most blocks contain multiple lots. We also observe that around 14000-15000, there are no block numbers.

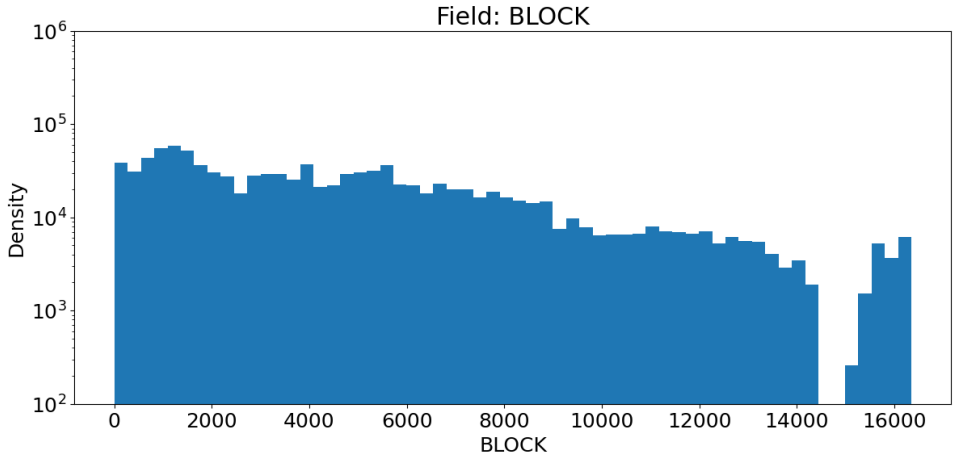

Figure 2: Distribution of BLOCK Numbers

LOT

`LOT` is a tax lot number within a given block. This field is fully populated and has a wide range of values. The most frequent lot number is 1, which appears in many different blocks. While the numbers are numeric, the combination of borough, block, and lot defines uniqueness—not the lot number alone. The distribution is consistent with NYC property structure, where each block can have several lots.

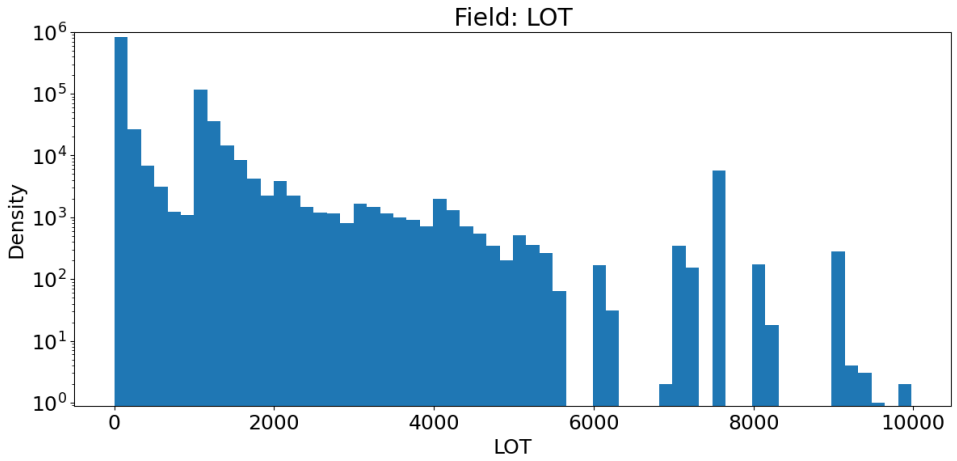

Figure 3: Distribution of LOT Numbers

EASEMENT

`EASEMENT` denotes special rights or access agreements on properties (e.g., utility easements, air rights). It is sparsely populated—only about 0.43% of the records contain a non-null value. Values are typically single letters, such as "E" (land easement) or "A" (air rights). The field is categorical and meaningful primarily for spatial planning or legal analysis. Despite the high number of missing values, the non-missing entries are consistent and valid.

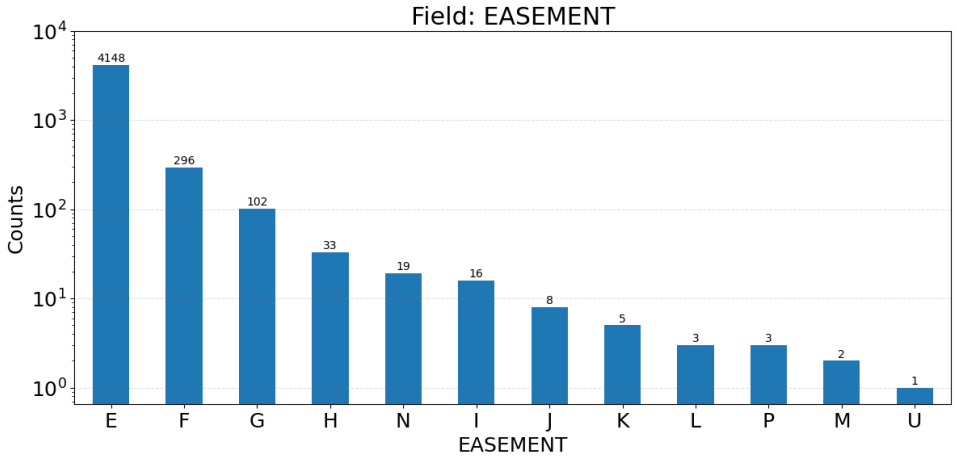

Figure 4: Frequency of EASEMENT Codes (Non-Null Only)

BLDGCL

`BLDGCL` represents the Building Class code and is a two-character alphanumeric identifier describing the property's structural use. Examples include "A1" (1-family dwelling), "C0" (vacant land), and "R4" (condominium units). The field is fully populated with over 200 unique values, reflecting the diversity of NYC's real estate. The distribution is skewed toward residential categories, especially R4 (condos) and A1 (single-family homes), consistent with the city's residential zoning. Building Class codes are crucial for tax assessment and land use classification.

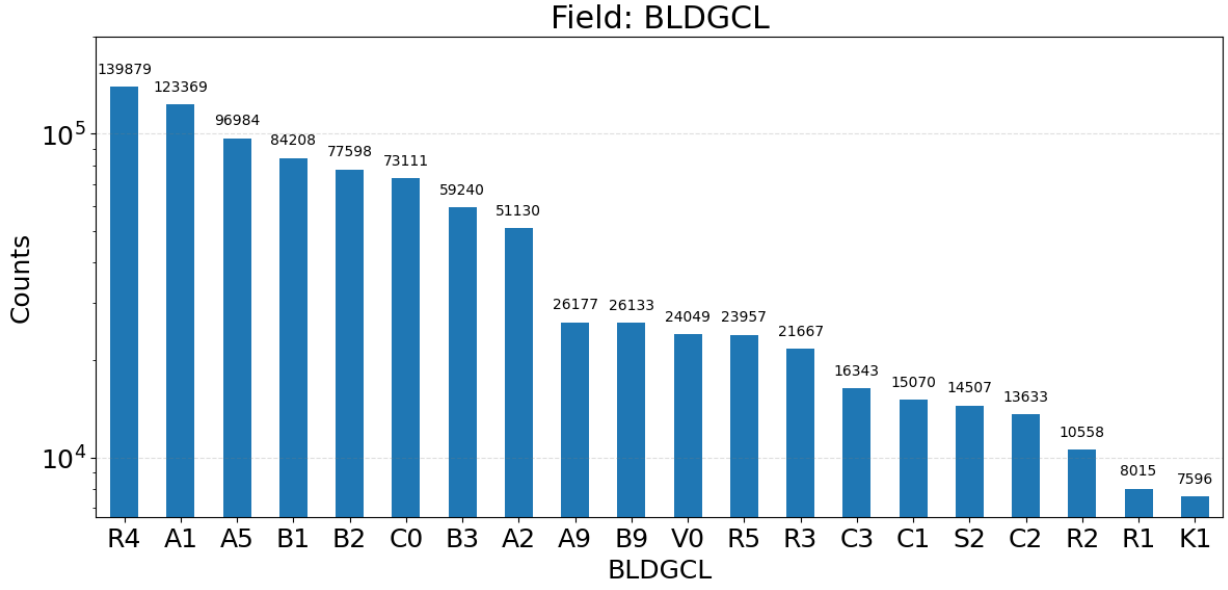

Figure 5: Top 20 Most Frequent Building Class Codes

OWNER

The `OWNER` field records the name of the property owner. It is 97% populated, with over 860,000 unique values. Many entries are institutional or corporate names (e.g., NYC Housing Authority, Parkchester Preservation). The most common name appears over 6,000 times, typically for large housing complexes. Though stored as strings, this field is best used for grouping, not statistical analysis. Values are typically uppercased and may be truncated or abbreviated due to system constraints.

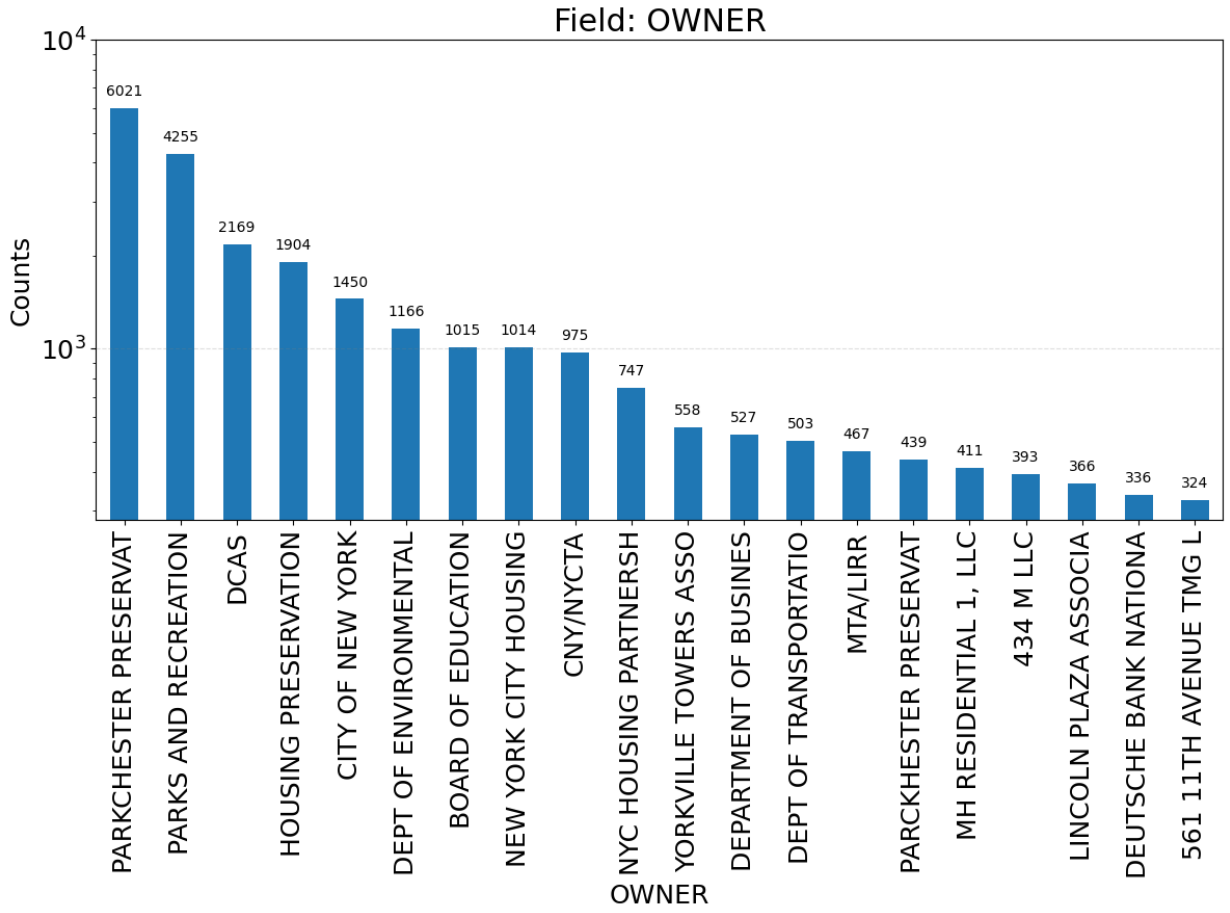

Figure 6: Top 20 Most Frequent OWNER Names

TAXCLASS

`TAXCLASS` defines how a property is taxed, with values such as:

- 1 – One-to-three family residential
- 2 – Multi-family residential (rental, condo, co-op)
- 3 – Utility
- 4 – Commercial/Industrial

The field includes some subclass distinctions (e.g., 1A, 2A), making a total of 11 unique values. The majority of properties fall into Class 1, followed by Class 2. This classification governs assessed value ratios and exemption applicability. Every record has a valid tax class, and no null or erroneous entries are present.

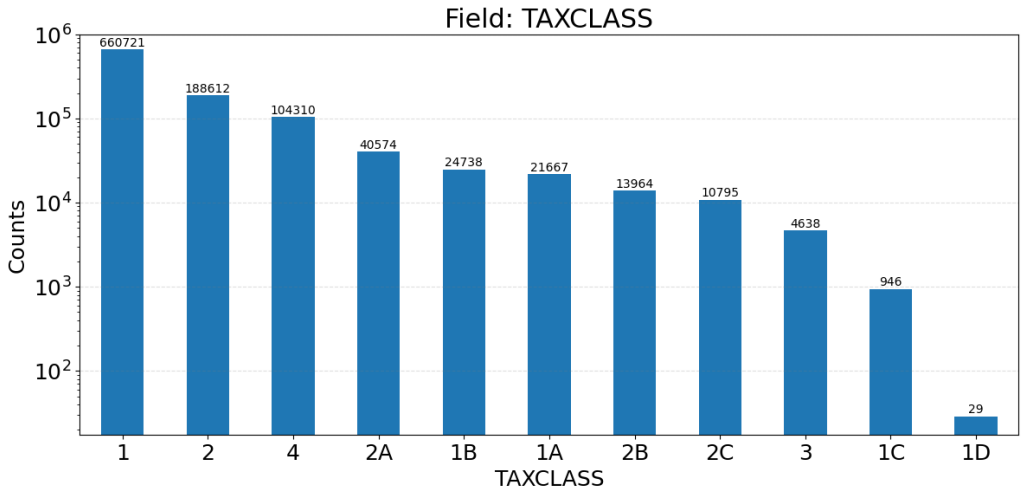

Figure 7: Distribution of TAXCLASS Values

EXT

`EXT` is a lesser-documented field likely referring to lot extension or irregularity. It contains codes like "G", "E", and "EG". Around 33% of the data has a non-empty value. These values may correspond to extended lots or special configurations (e.g., corner lots, through-lots), although detailed documentation is limited. Despite this, the consistency and limited vocabulary of the values suggest intentional and structured use.

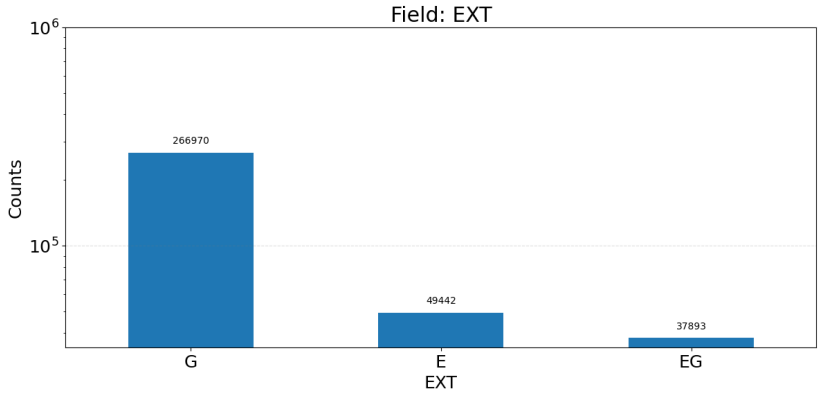

Figure 8: EXT Code Frequencies

EXCD1

`EXCD1` is the primary exemption code, indicating the type of property tax exemption applied (if any). Only about 60% of records have a value, reflecting the subset of properties that qualify for tax relief (e.g., STAR program, veterans, senior citizens). There are nearly 200 unique codes, with "1017" being overwhelmingly common—likely the STAR Basic exemption. This field is categorical, and analysis helps reveal the prevalence of exemption types.

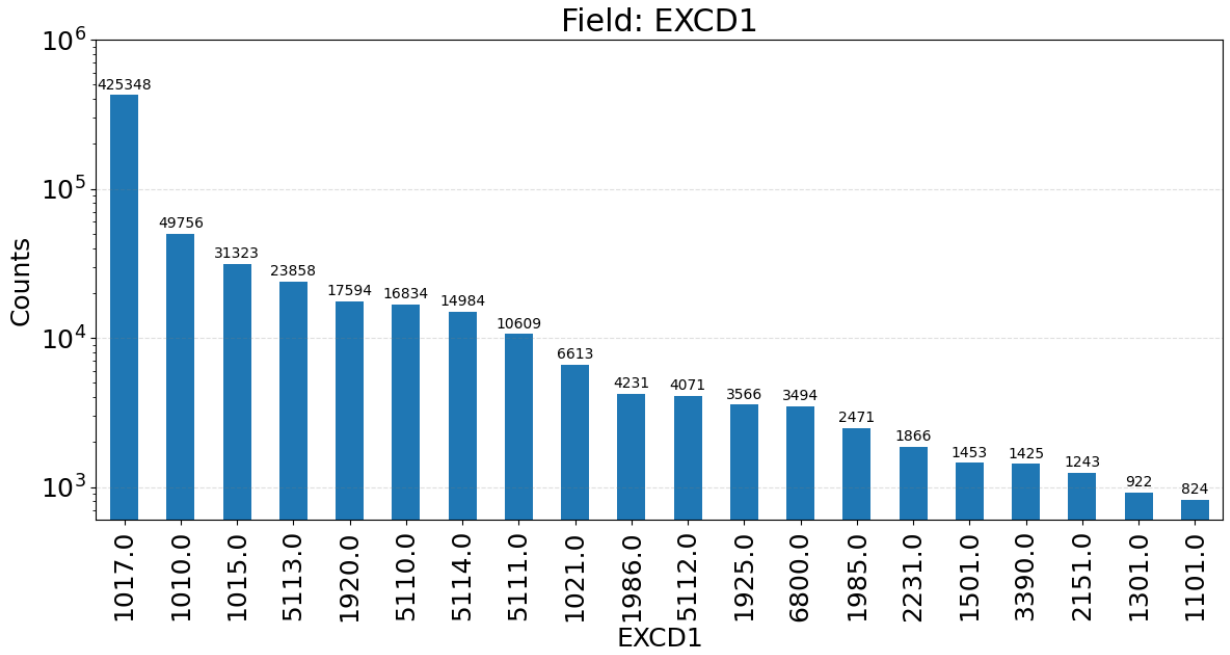

Figure 9: Top 20 EXCD1 (Primary Exemption Codes)

STADDR

The `STADDR` field holds the street address of the property, usually including street number and name (e.g., "501 SURF AVENUE"). It is over 99.9% populated and contains more than 980,000 unique values. Some addresses repeat across units in multi-unit buildings or condominiums, which is expected. Address values are uppercased and sometimes truncated due to formatting limitations. While not analytically numeric, `STADDR` is useful for mapping and aggregation.

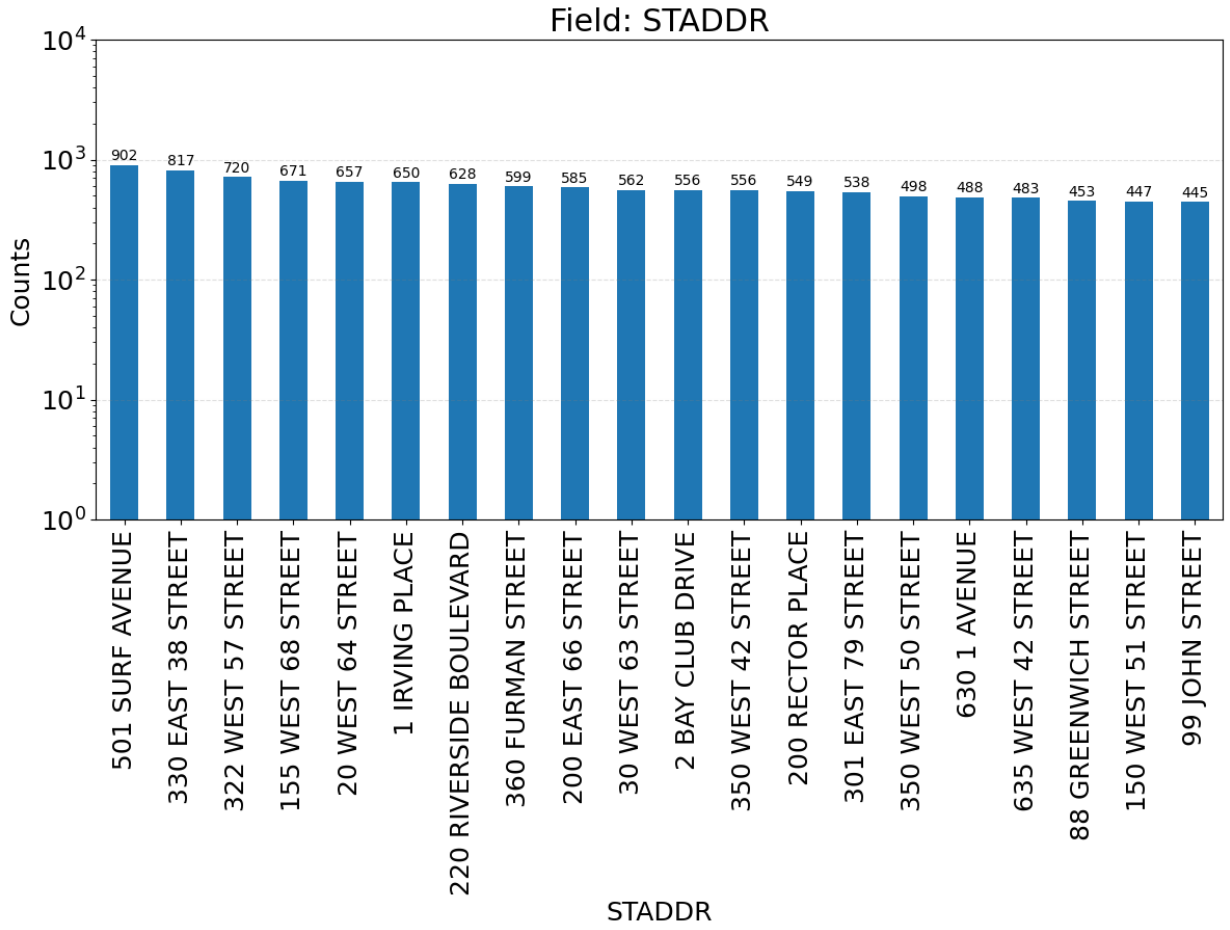

Figure 10: Top 20 Most Frequent STADDR Entries

ZIP

`ZIP` indicates the postal ZIP code for the property address. It is populated in over 97% of the dataset and includes 181 unique ZIP codes, consistent with the granularity of NYC neighborhoods. The most frequent ZIP code is 10314 (Staten Island), followed by various Brooklyn and Queens locations. Missing ZIPs are primarily associated with special-use parcels or those without typical mailing addresses. The field is well-structured and matches real NYC postal areas.

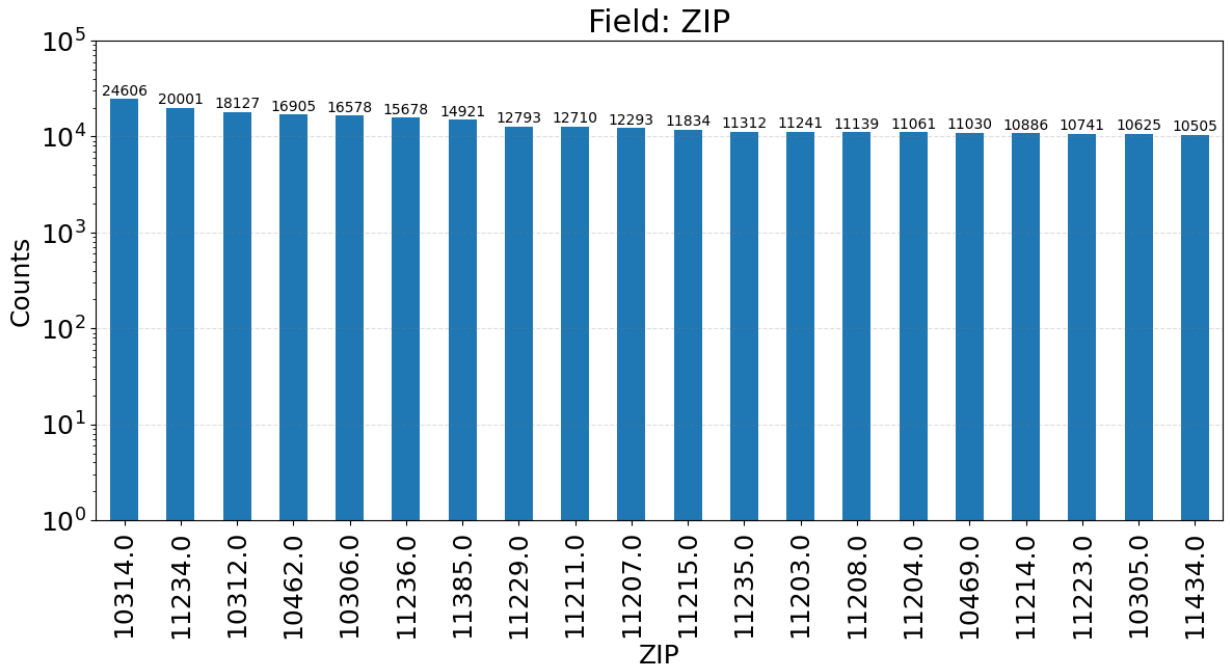

Figure 11: Top 20 Most Frequent ZIP Codes

EXMPTCL

`EXMPTCL` stands for Exemption Class and indicates whether a property is entirely tax-exempt. Only about 1.45% of records have a value, such as "X1" or "Y2", which correspond to government-owned or institutionally exempt parcels. Blank entries imply taxable properties. This field is useful for separating public/institutional properties from the taxable base and often aligns with ownership fields like "CITY OF NEW YORK".

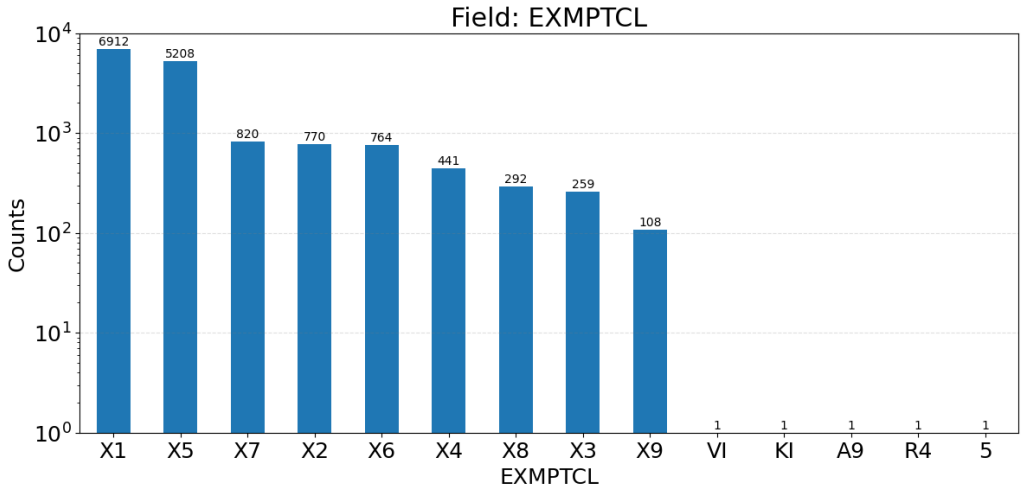

Figure 12: Distribution of Exemption Class (EXMPTCL)

EXCD2

`EXCD2` is a secondary exemption code, indicating an additional exemption applied to the same property. It is populated in about 8.7% of records. The most frequent code here is again 1017, suggesting some properties have dual STAR-type exemptions or overlapping programs (e.g., veterans + senior exemptions). While not as widely used as `EXCD1`, it plays an important role in understanding the full exemption profile of a parcel.

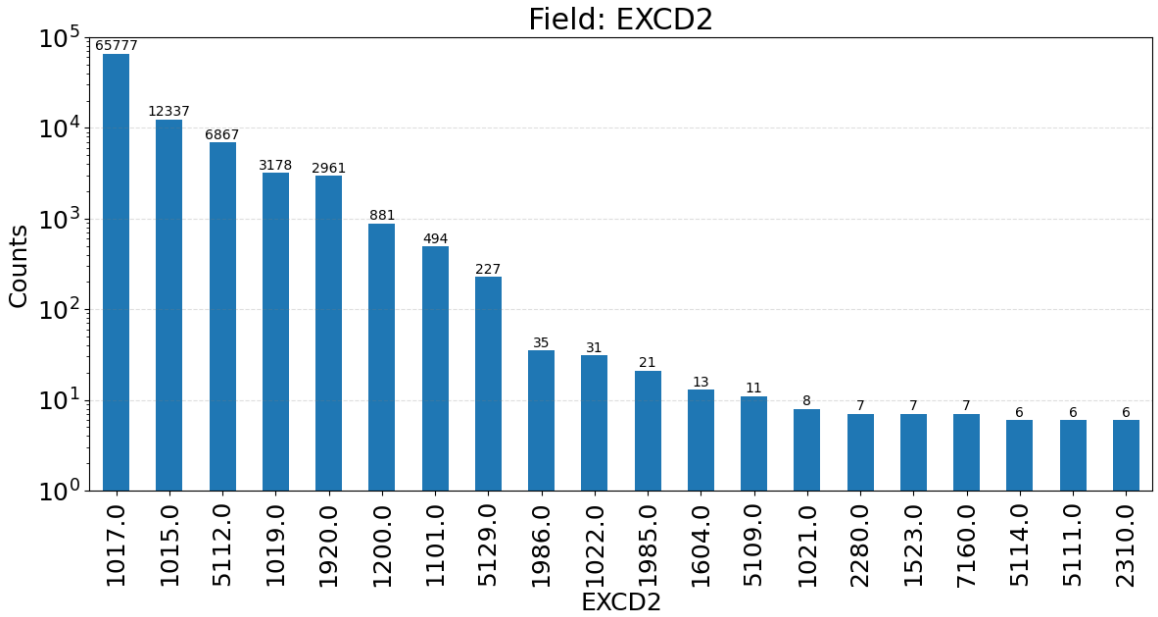

Figure 13: Top 20 EXCD2 (Secondary Exemption Codes)

PERIOD

`PERIOD` designates the type of tax roll (e.g., "FINAL" or "TENTATIVE"). In this dataset, every record has the value "FINAL", since this dataset exclusively captures the 2010/11 final tax assessment roll. As such, it is a constant field with no variance. While not analytically valuable on its own in this case, it would be critical if data from multiple assessment periods were combined.

YEAR

`YEAR` captures the fiscal year of the assessment roll. Like `PERIOD`, it is constant across all records ("2010/11"). It reaffirms the temporal scope of the data and is vital when aggregating data across multiple years. In this standalone dataset, however, it adds no variance or analytical value.

VALTYPE

`VALTYPE` stands for Valuation Type and indicates which kinds of assessed values are available. Every record contains the value "AC-TR", meaning both Actual and Transitional values are included. This field would be useful for cross-year or comparative analyses, but is redundant here as there is no variation.

LTFRONT

`LTFRONT` represents the width of the property lot in feet. It is a numeric field, fully populated, but approximately 15.8% of values are zero. A zero value may indicate either a missing or undefined dimension—common in cases such as condominium units or irregular lots. The most frequent non-zero values cluster around 20–25 feet, typical of NYC rowhouses. Some values are unusually high (up to 9,999), suggesting outliers or placeholder codes.

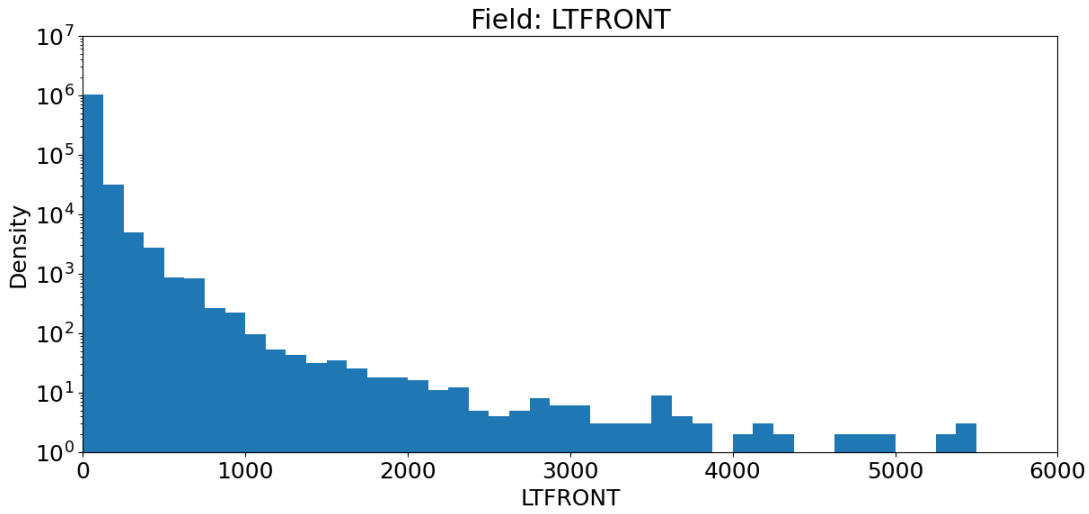

Figure 14: Histogram of LTFRONT (Lot Width) - excluding outliers beyond 6,000 feet

LTDEPTH

`LTDEPTH` measures the depth of the property lot in feet. Similar to `LTFRONT`, it is fully populated, with around 15.9% of entries equal to zero. Typical values for residential lots fall between 80–100 feet. High outlier values (e.g., 9,999) likely indicate large or irregular parcels or use as a data placeholder. Despite these quirks, the distribution for most properties is consistent with NYC planning norms.

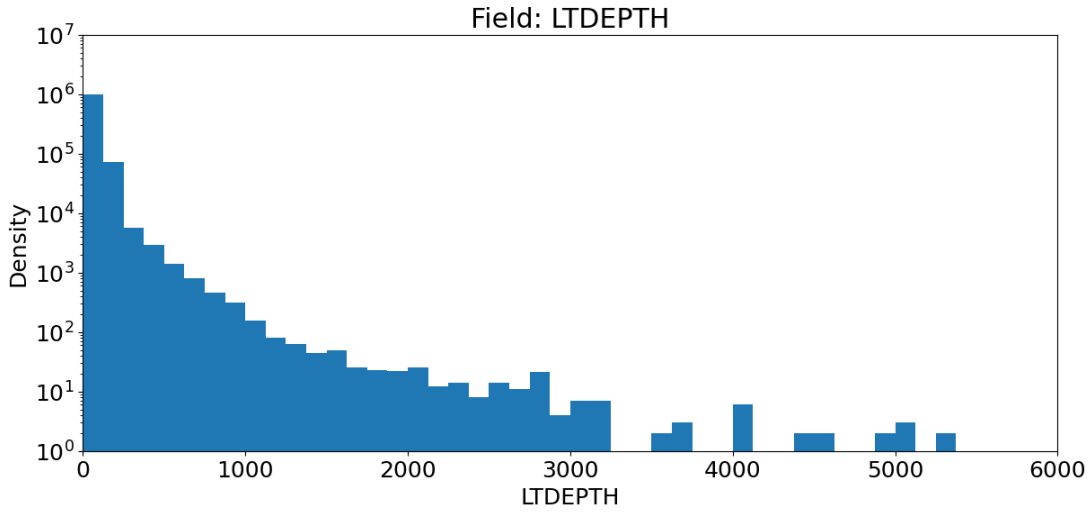

Figure 15: Distribution of LTDEPTH (Lot Depth) - excluding outliers beyond 6,000 feet

STORIES

`STORIES` captures the number of above-ground stories in a building. About 5.25% of records are missing. The most frequent values are whole numbers such as 1, 2, and 3, which represent one- to three-story buildings common in residential neighborhoods. Values like 1.5 and 2.5 appear frequently for homes with attics or partial floors. A small number of extreme values (e.g., 119) represent skyscrapers or data entry anomalies. A few fields include strange decimal values (e.g., 1.7, 3.3) which are likely erroneous.

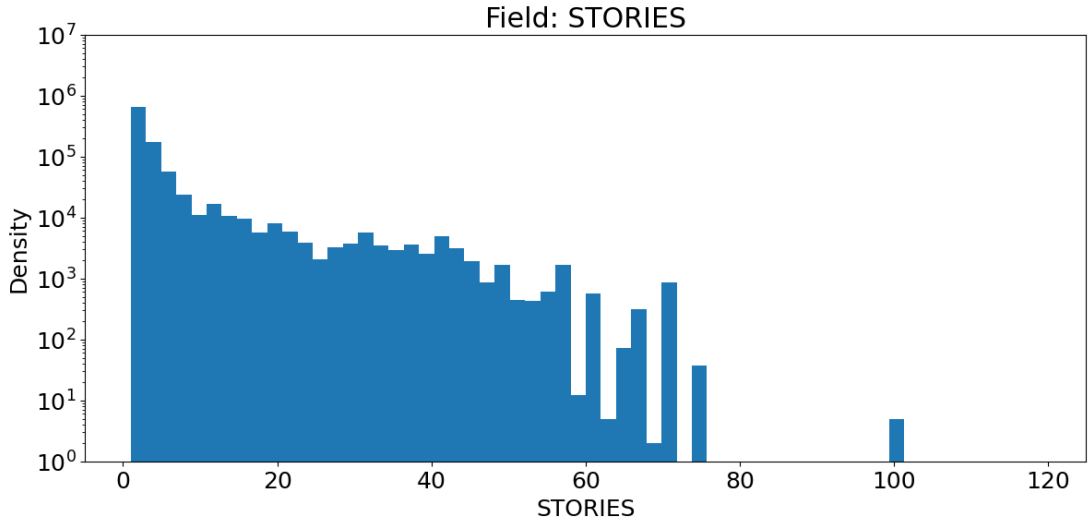

Figure 16: Distribution of STORIES (Building Height)

FULLVAL

`FULLVAL` is the full market value of the property, expressed in dollars. It is fully populated, with approximately 1.2% of records having a value of zero. The distribution is highly right-skewed; most properties are valued under $1 million, but some (e.g., large commercial properties) exceed $6 billion. The average value is about $874k. A log transformation reveals a near-normal distribution, suggesting a log-normal nature typical of real estate data.

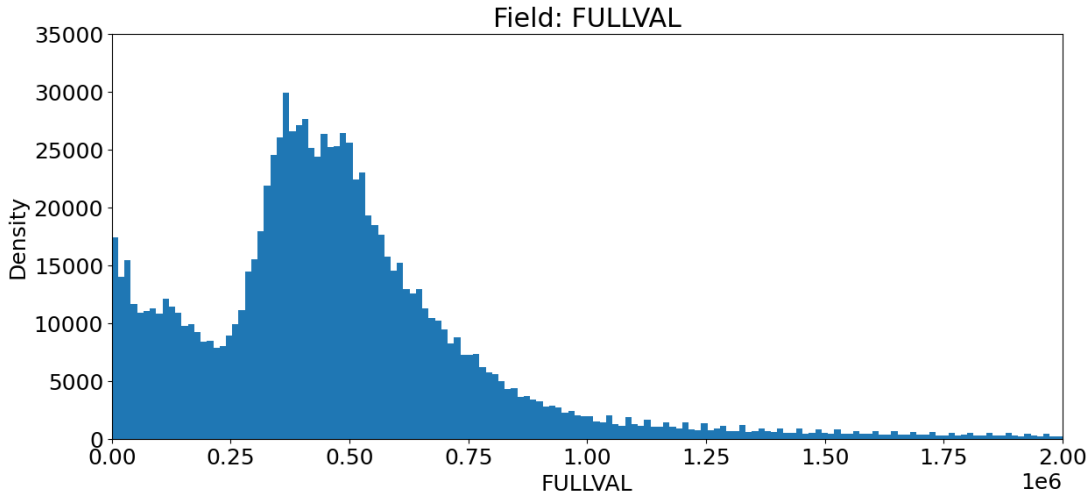

Figure 17: Distribution of FULLVAL (Full Market Value) - excluding values beyong $2 million as majority of the properties lie in this range

AVLAND

`AVLAND` is the assessed value of the land portion of the property, excluding improvements. This field is populated for all records. Values mirror the distribution of `FULLVAL` but are consistently lower due to assessment ratios and exemptions. Around 13k entries are zero. The average land assessment is $85k, and the maximum exceeds $2.5 billion. For many Class 1 residential properties, land is the dominant component of value.

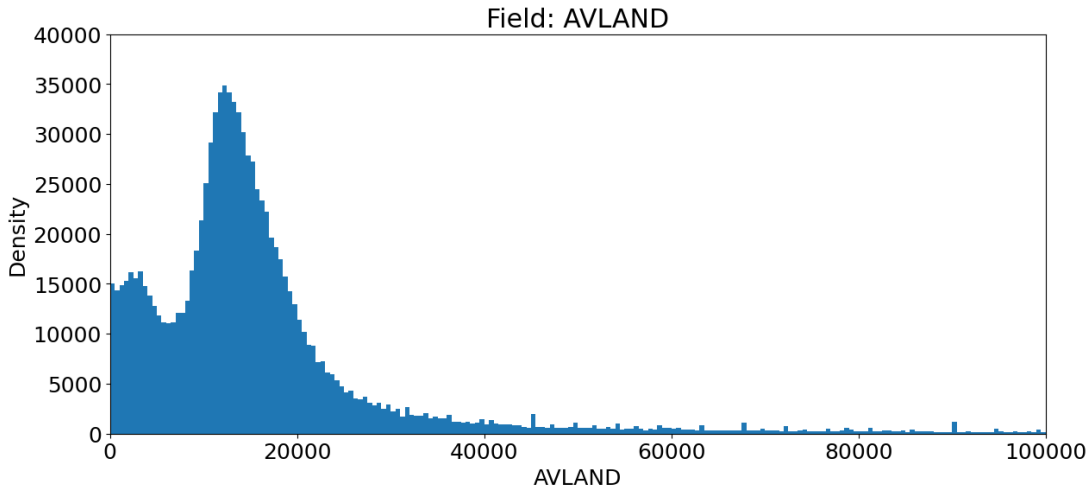

Figure 18: Distribution of AVLAND (Assessed Land Value) - excluding values beyong $100k as majority of the properties lie in this range

AVTOT

`AVTOT` is the total assessed value of the property (land + improvements). It is fully populated and, in most cases, greater than or equal to `AVLAND`. The average assessed total is about $227k. Like `FULLVAL`, this field is heavily skewed due to the wide range of property types in NYC. A small number of properties show an assessed value of zero, typically those that are fully tax-exempt or vacant.

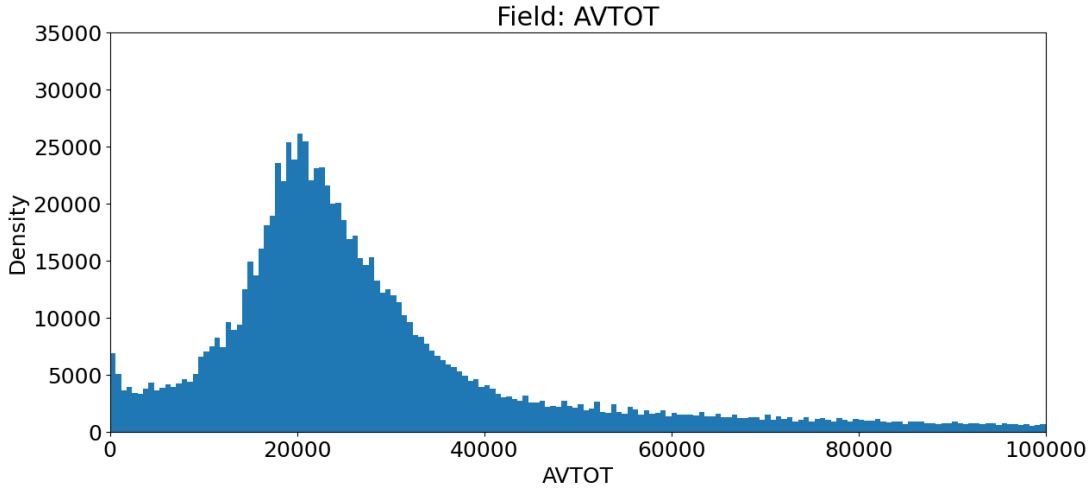

Figure 19: Distribution of AVTOT (Total Assessed Value) - excluding values beyong $100k as majority of the properties lie in this range

EXLAND

`EXLAND` is the exempt portion of the land's assessed value. This field is numeric and fully populated, though 45.9% of entries are zero. A non-zero value indicates partial exemption from property taxes on land value. Values range up to nearly $2 billion. For properties eligible for tax exemptions (e.g., veterans, senior citizens, religious organizations), this field reflects the portion of land value excluded from taxation.

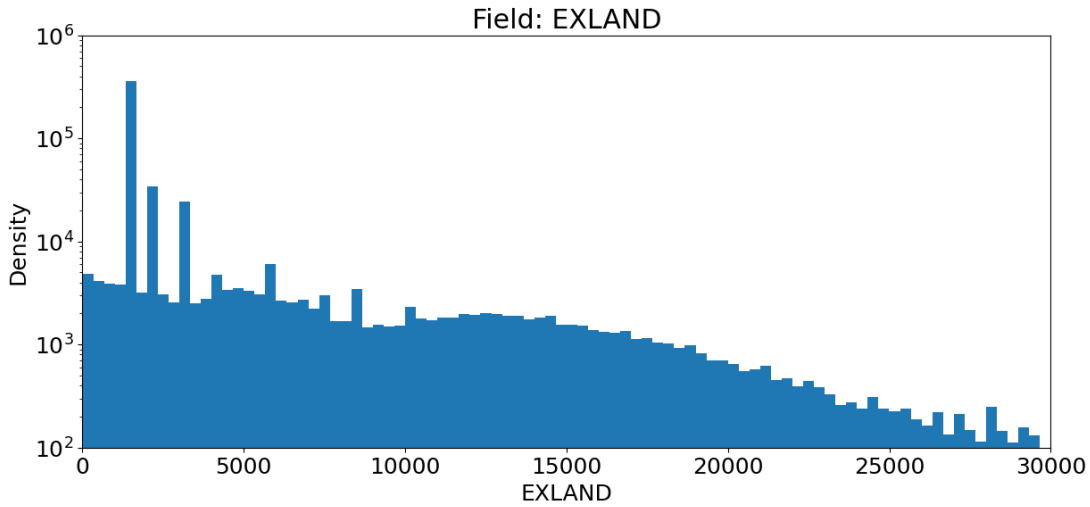

Figure 20: Distribution of EXLAND (Exempt Land Value, Log Scale) - excluding values beyond $40k

EXTOT

`EXTOT` is the exempt portion of the total assessed value (land + improvements). About 40.4% of records show a zero value, while others reflect varying degrees of tax relief. Values span a broad range and can exceed $2 billion. Like `EXLAND`, this field is essential in calculating net taxable value for a property.

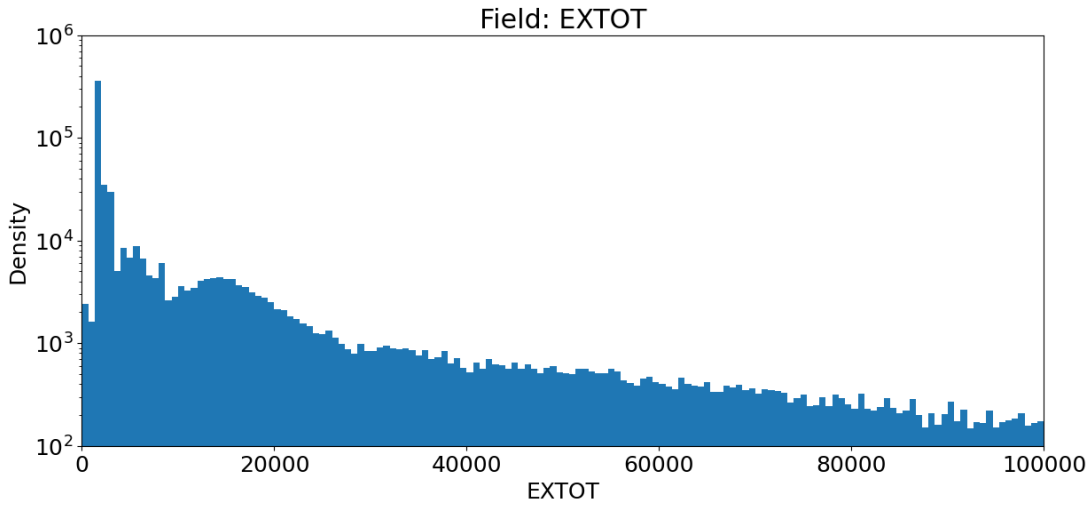

Figure 21: Distribution of EXTOT (Total Exempt Value, Log Scale) - excluding values beyond $100k

BLDFRONT

`BLDFRONT` indicates the width of the building in feet. It is fully populated, but over 21% of values are zero, likely due to properties like condos or vacant land that do not independently list a building footprint. For non-zero entries, typical widths range from 20 to 50 feet, reflecting standard townhouse and apartment widths. Some outliers (e.g., 3,000 feet) may be errors or placeholders.

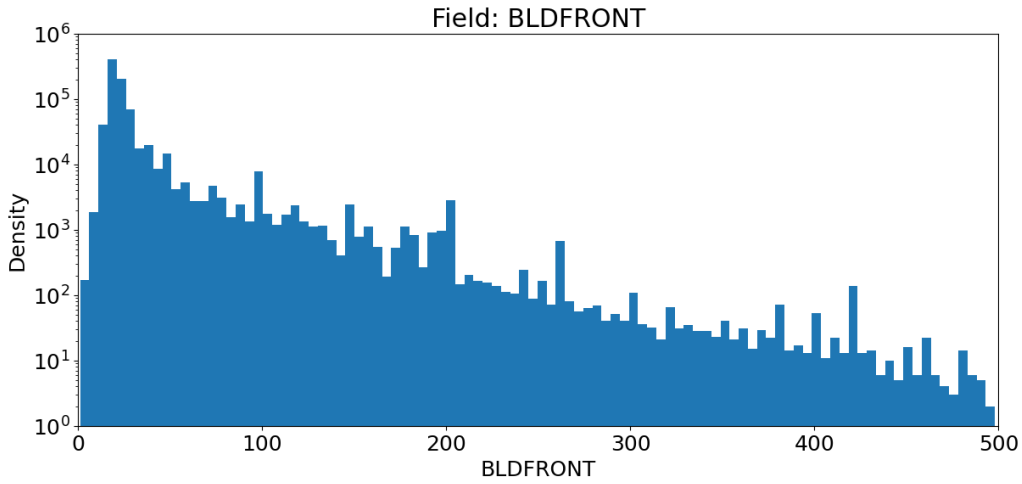

Figure 22: Distribution of BLDFRONT (Building Frontage) - excluding outliers beyond 500 feet

BLDDEPTH

`BLDDEPTH` measures the depth of a building in feet. As with `BLDFRONT`, many records have a value of zero (over 21%), especially for properties without individual buildings (e.g., condo units). Common values cluster between 30 and 100 feet. Outliers (up to 3,306 feet) may indicate industrial properties or irregular lots.

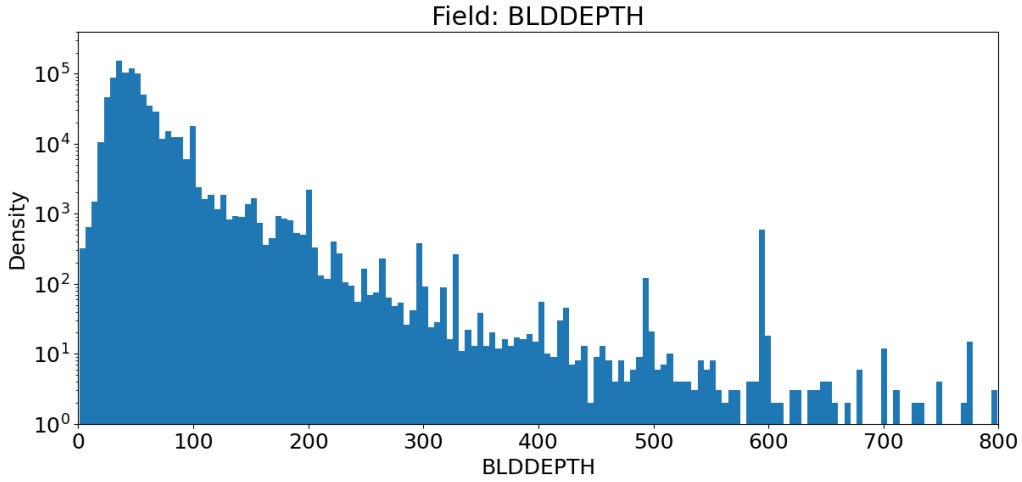

Figure 23: Distribution of BLDDEPTH (Building Depth) - excluding values beyond 800 feet

AVLAND2

`AVLAND2` is the transitional assessed land value, used for phasing in tax increases. It is populated for about 26.4% of records, primarily in Classes 2 and 4. These values often differ from `AVLAND` due to assessment caps and smoothing. Typical values mirror actual assessed land but are adjusted downward for transitional relief. The most common transitional land assessment is $2,408.

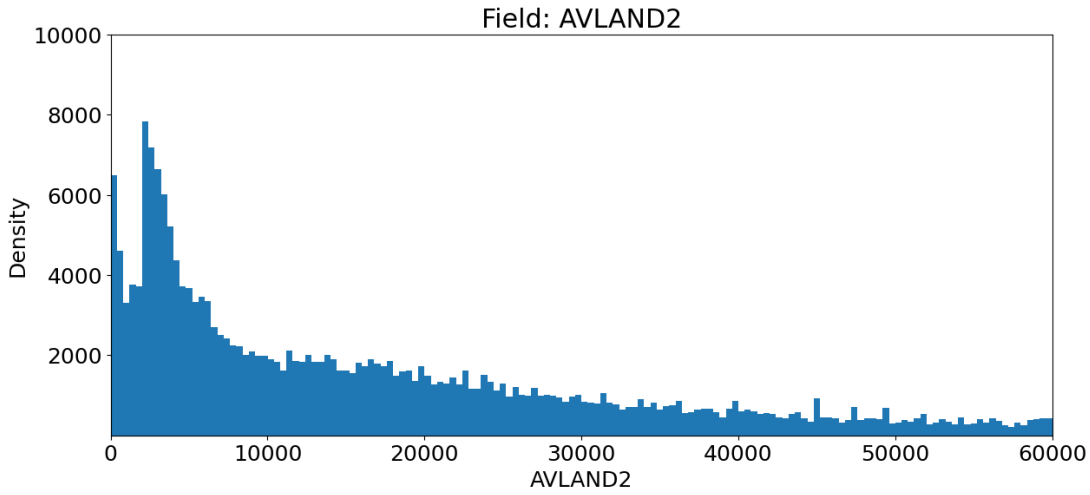

Figure 24: Distribution of AVLAND2 (Transitional Land Value) - excluding values beyond $60k

AVTOT2

`AVTOT2` is the transitional total assessed value. Like `AVLAND2`, it is used to phase in tax changes gradually. It is present in the same 26% of records. The values are generally less than or equal to `AVTOT`, although some records show equal values, indicating no transitional adjustment. The most common value is $750, and the mean is significantly higher due to high-value commercial properties.

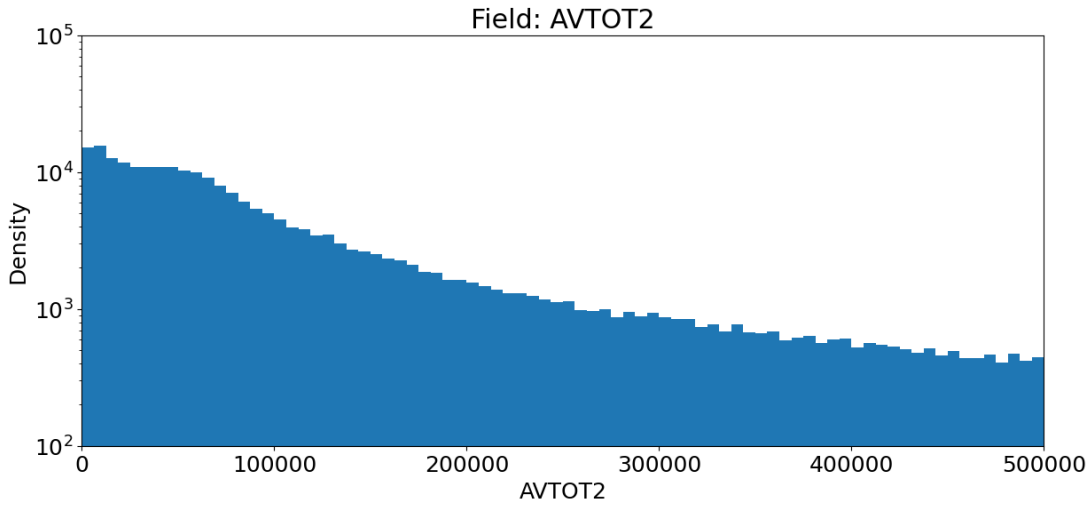

Figure 25: Distribution of AVTOT2 (Transitional Total Value, Log Scale) - excluding values beyond $500k

EXLAND2

EXLAND2 is the exempt portion of transitional land value. Only 8.2% of records are populated, which corresponds to properties both exempt and under transitional assessment rules. The most frequent value is $2,090, aligning with standard deductions (e.g., STAR). Distribution is sparse but consistent with exemption policy.

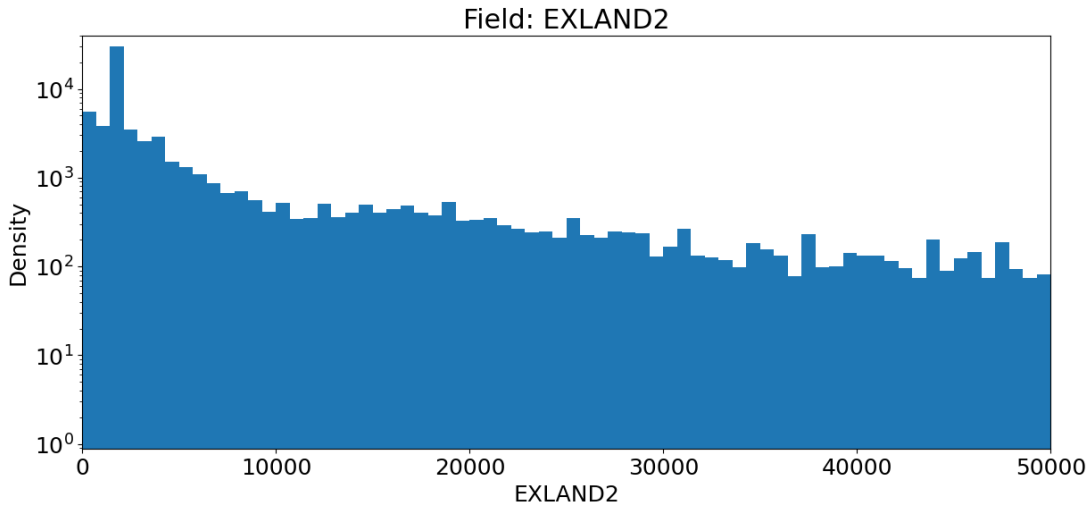

Figure 26: Distribution of EXLAND2 (Transitional Exempt Land Value, Log Scale) - excluding values beyond $50k

EXTOT2

EXTOT2 reflects the total transitional exempt value. Populated in about 12.2% of records, it corresponds with exemption codes and class-specific transitions. The values range widely, with a common value again at $2,090. This field provides insight into phase-in relief granted to owners with exemption status.

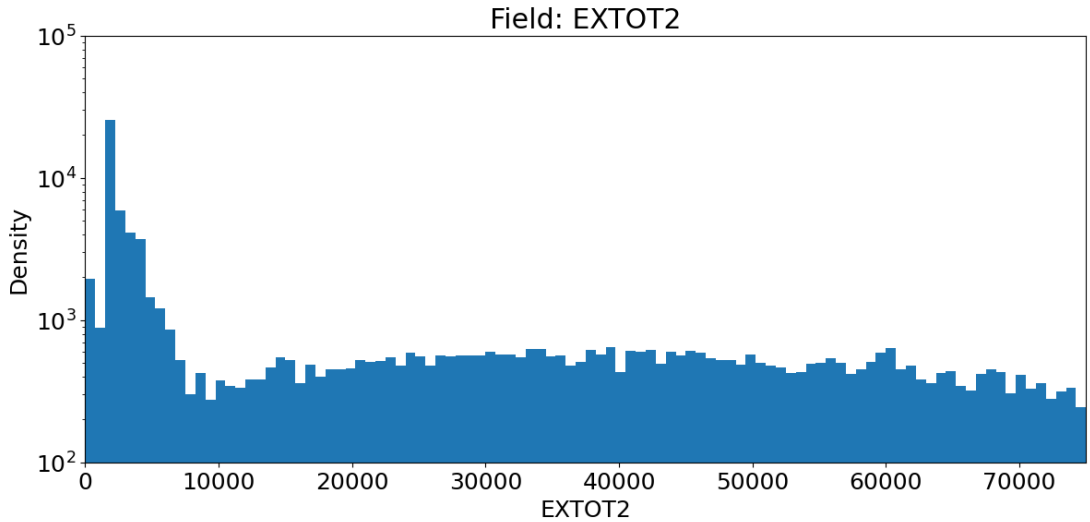

Figure 27: Distribution of EXTOT2 (Transitional Total Exempt Value, Log Scale) - excluding outliers beyond $80k