Project 3 Report

Eloise Yu

MS in Business Analytics, Marshall School of Business, University of Southern California

DSO 562: Fraud Analytics

Dr. Stephen Coggeshall

December 14, 2024

Table of Contents

Executive Summary

This project addresses the business problem of analyzing anomalies in the New York City property data to detect potential fraudulent activity. Using a dataset of 1,070,994 records of New York City properties, the goal is to identify anomalies/strangeness in the data to accurately recognize fraudulent activities. This initiative employs advanced unsupervised learning techniques to systematically detect unusual property characteristics and financial discrepancies, ensuring the integrity of property transactions and valuations. Through data preprocessing and feature engineering, we handled missing values, created 29 new variables, and further standardized and normalized them for comparability. Moreover, we employed two unsupervised learning algorithms to compute anomaly scores, Minkowski Distance Scoring and Autoencoder Reconstruction Error. The final anomaly scores ranked properties based on their likelihood of irregularities, signalling possible fraudulent activity.

Data Description

Dataset Overview

This dataset contains 1,070,994 records of New York City properties, featuring 32 fields that include both categorical and numeric data. With 1,070,994 records, this dataset provides a substantial resource for analyzing anomalies/strangeness in the data to detect potential fraudulent activity.

Filed description

1. Record: An ordinal identifier assigned to each property entry.
2. BBLE: A categorical field key representing the Borough, Block, Lot, and Easement code.
3. BORO: A categorical field specifying New York City boroughs, where 1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, and 5 = Staten Island.
4. BLOCK: A categorical field indicating valid block ranges for each borough: Manhattan (1–2255), Bronx (2260–5958), Brooklyn (1–8955), Queens (1–16350), Staten Island (1–8050).
5. LOT: A categorical field denoting the specific plot of land designated for individual ownership.
6. EASEMENT: A categorical field that indicates whether a property has an easement, which is a legal right for a third party to use a portion of the property for a specific purpose. Values include: Space = No Easement, A = Air Easement, B = Non- Air Rights, E = Land Easement, N = Non-Transit Easement, P = Pier, R = Railroad, S = Street, U= U.S. Government, F-M = duplicates of E.
7. OWNER: The owner's name for the property record.
8. BLDGCL: The building class of each property record.

9. TAXCLASS: A categorical field indicating the tax class for each property, with the following classifications: 1 = 1-3 Unit Residence, 2 = Apartments, 2A = 4, 5, or 6 Units, 3 = Utilities, 4 = All Others.
10. LTFRONT: A numerical field denoting the lot width for each property record.
11. LTDEPTH: A numerical field specifying the lot depth for each property record.
12. EXT: A categorical extension indicator for each property record.
13. STORIES: A numerical field indicating the number of stories in the building.
14. FULLVAL: A numerical field representing the market value of each property.
15. AVLAND: A numerical field denoting the actual land value.
16. AVTOT: A numerical field representing the actual total value of the property.
17. EXLAND: A numeric field representing the actual exempt land value of the property.
18. EXTOT: A numeric field indicating the total actual exempt land value.
19. EXCD1: A categorical field denoting the exemption code assigned to the property.
20. STADDR: A categorical field containing the street address of the property.
21. ZIP: A categorical field specifying the property's ZIP code.
22. EXMPTCL: A categorical field identifying the exemption class of the property.
23. BLDFRONT: A numeric field indicating the building's width on the property.
24. BLDDEPTH: A numeric field specifying the building's depth on the property.
25. AVLAND2: A numeric field representing the transitional land value of the property.
26. AVTOT2: A numeric field indicating the transitional total value of the property.
27. EXLAND2: A numeric field describing the property's transitional exempt land value.
28. EXCD2: A numeric field indicating the property's transitional exemption land total.

29. PERIOD: A categorical field denoting the property's assessment period when the data was created. We note that all the information in the property data occurs in the "Final Period".
30. YEAR: A categorical field representing the assessment year. When plotting the distribution, we note that the assessment year across all properties occurs in 2010/2011.
31. VALTYPE: A categorical field specifying the valuation type of the property. Currently, the valuation type for all rows of data are "AC-TR," which refers to Assessed Value per Acre – Total Revenue.

Field Summary Statistics Tables

Table 1

Numeric Field Summary Statistics

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	LTFRONT	numeric	1,070,994	100.0%	169,108	0.00	9,999.00	36.64	74.03	0.00
1	LTDEPTH	numeric	1,070,994	100.0%	170,128	0.00	9,999.00	88.86	76.40	100.00
2	STORIES	numeric	1,014,730	94.7%	0	1.00	119.00	5.01	8.37	2.00
3	FULLVAL	numeric	1,070,994	100.0%	13,007	0.00	6,150,000,000.00	874,264.51	11,582,425.58	0.00
4	AVLAND	numeric	1,070,994	100.0%	13,009	0.00	2,668,500,000.00	85,067.92	4,057,258.16	0.00
5	AVTOT	numeric	1,070,994	100.0%	13,007	0.00	4,668,308,947.00	227,238.17	6,877,526.09	0.00
6	EXLAND	numeric	1,070,994	100.0%	491,699	0.00	2,668,500,000.00	36,423.89	3,981,573.93	0.00
7	EXTOT	numeric	1,070,994	100.0%	432,572	0.00	4,668,308,947.00	91,186.98	6,508,399.78	0.00
8	BLDFRONT	numeric	1,070,994	100.0%	228,815	0.00	7,575.00	23.04	35.58	0.00
9	BLDDEPTH	numeric	1,070,994	100.0%	228,853	0.00	9,393.00	39.92	42.71	0.00
10	AVLAND2	numeric	282,726	26.4%	0	3.00	2,371,005,000.00	246,235.72	6,178,951.64	2,408.00
11	AVTOT2	numeric	282,732	26.4%	0	3.00	4,501,180,002.00	713,911.44	11,652,508.34	750.00
12	EXLAND2	numeric	87,449	8.2%	0	1.00	2,371,005,000.00	351,235.68	10,802,150.91	2,090.00
13	EXTOT2	numeric	130,828	12.2%	0	7.00	4,501,180,002.00	656,768.28	16,072,448.75	2,090.00

Table 2

Categorical Field Summary Statistics

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	RECORD	categorical	1,070,994	100.0%	0	1,070,994	1
1	BBLE	categorical	1,070,994	100.0%	0	1,070,994	1000010101
2	BORO	categorical	1,070,994	100.0%	0	5	4
3	BLOCK	categorical	1,070,994	100.0%	0	13,984	3,944
4	LOT	categorical	1,070,994	100.0%	0	6,366	1
5	EASEMENT	categorical	4,636	0.4%	0	12	E
6	OWNER	categorical	1,039,249	97.0%	0	863,347	PARKCHESTER PRESERVAT
7	BLDGCL	categorical	1,070,994	100.0%	0	200	R4
8	TAXCLASS	categorical	1,070,994	100.0%	0	11	1
9	EXT	categorical	354,305	33.1%	0	3	G
10	EXCD1	categorical	638,488	59.6%	0	129	1017.00
11	STADDR	categorical	1,070,318	99.9%	0	839,280	501 SURF AVENUE
12	ZIP	categorical	1,041,104	97.2%	0	196	10314.00
13	EXMPTCL	categorical	15,579	1.5%	0	14	X1
14	EXCD2	categorical	92,948	8.7%	0	60	1017.00
15	PERIOD	categorical	1,070,994	100.0%	0	1	FINAL
16	YEAR	categorical	1,070,994	100.0%	0	1	2010/11
17	VALTYPE	categorical	1,070,994	100.0%	0	1	AC-TR

Field Distribution

Figure 1

BORO Distribution

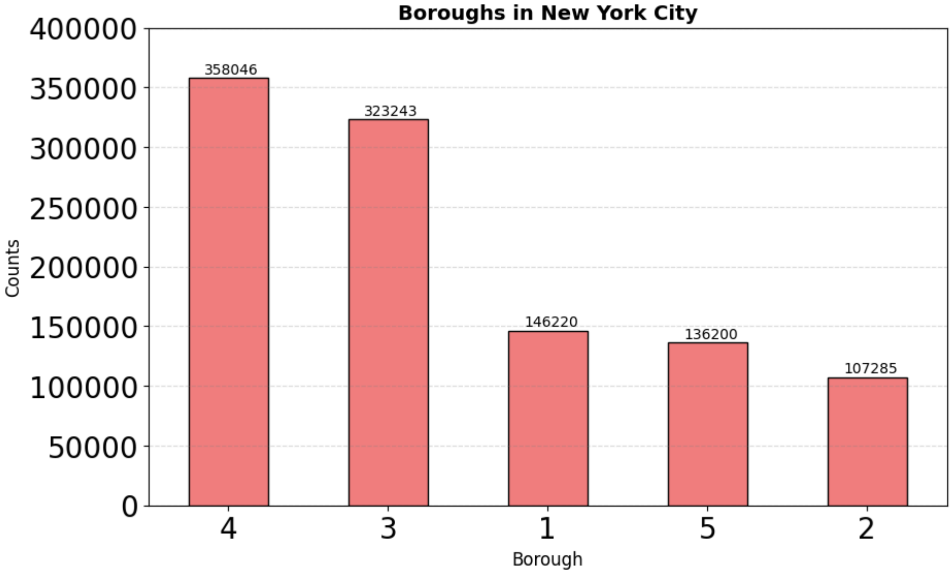


Figure 2

BLOCK Distribution

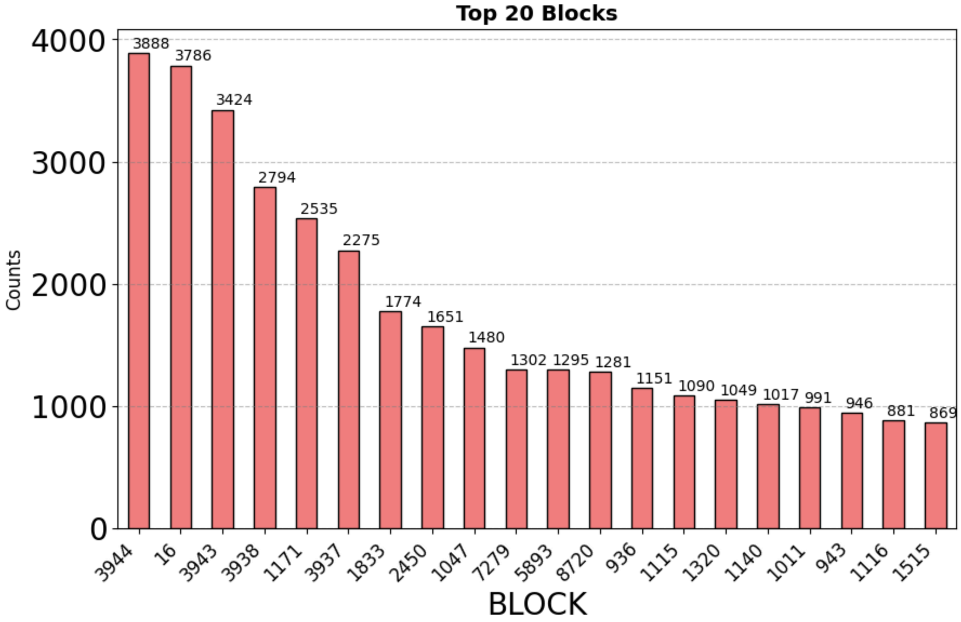


Figure 3

LOT Distribution

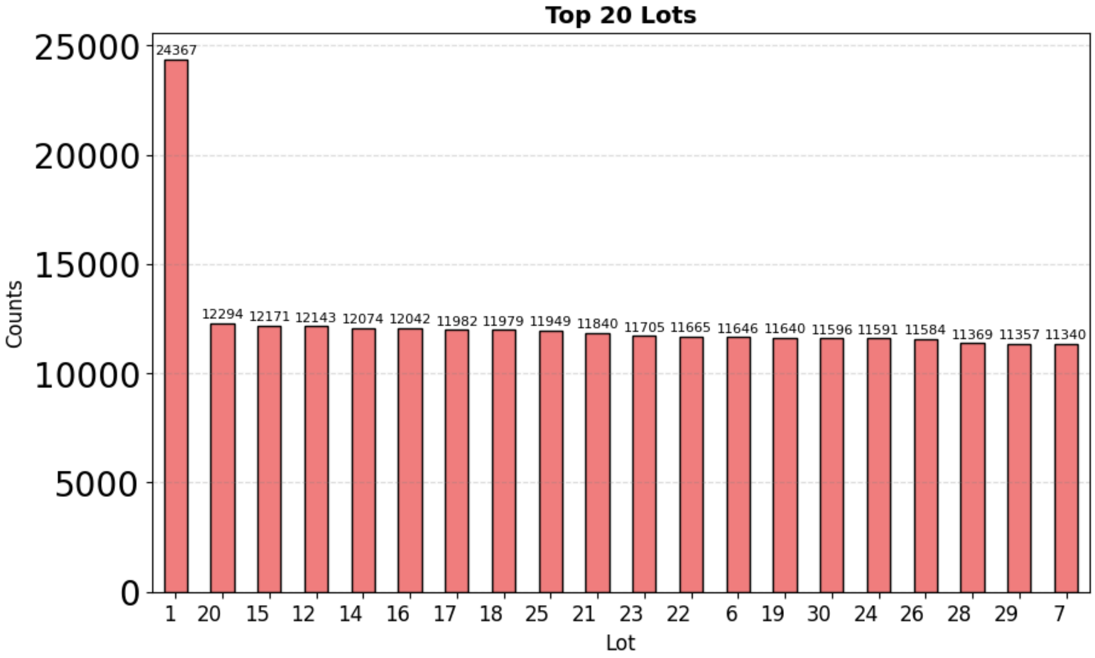
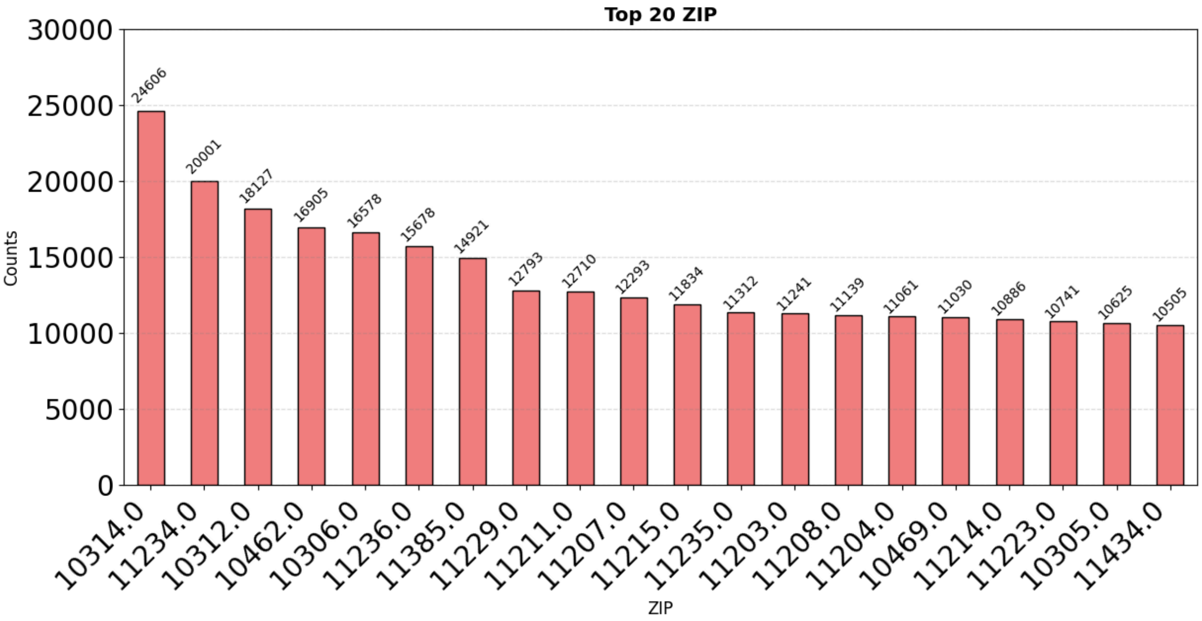


Figure 4

ZIP Distribution



Data Cleaning

Outliers

For an unsupervised fraud problem like this project, we don’t remove outliers. Outliers are what we are looking for.

Exclusions - Removing unwanted properties

Due to the large number of records, we removed properties that are owned by any government agencies. This includes removing properties with Easement ‘U’, and Owner containing 'DEPT ', 'DEPARTMENT', 'UNITED STATES','GOVERNMENT',' GOVT ', and 'CEMETERY'. We also removed city-owned properties which excluded records that have ‘THE CITY OF NEW YORK’ or ‘NYC HIGHWAY DEPT’ in the Owner field. In total, we have removed 26,502 records.

Imputation

ZIP & ZIP3

There are 20,431 missing values in the ZIP column. In order to rectify this, we concatenated the street address and borough code and mapping it back to our missing ZIP values. Assuming that each record is sorted by ZIP, we filled in missing ZIP codes if the previous and the next record have the same ZIP. Finally, the remaining missing values are then entered with the previous ZIP values – reducing the number of nulls in this column to 0. We have also created a zip3 column that converts the current ZIP column into strings.

FULLVALL, AVLAND, AVTOT

The method used to fill in missing values in these columns using a hierarchical approach to maximize the accuracy of the replacements. It first filled NaN values using the mean of these columns within groups defined by TAXCLASS, BORO, and BLDGCL, which provided the most specific context.

For any remaining NaN values, it then used the mean based on the broader groupings TAXCLASS and BORO, and finally just TAXCLASS if necessary. This logic to filling in these missing values is to ensure the most relevant average, preserving more granular patterns before resorting to broader group means where specific data is lacking.

STORIES

Filling in missing values in the STORIES column follows a similar logic to that for FULLVALL and the columns mentioned above. It first used the mode within BORO and BLDGCL, which was more specific to each building type and location. For any remaining gaps, it was filled using the mean within TAXCLASS.

LTFRONT & LTDEPTH

A similar approach was used to impute missing values for both LTFRONT and LTDEPTH. It first used group averages within TAXCLASS and BORO, and then within TAXCLASS alone for remaining gaps.

BLDDEPTH & BLDFRONT

For missing values in both of these columns, we calculated the group averages within TAXCLASS, BORO, and BLDGCL. Without further manipulation, we successfully filled in all missing values.

Variable Creation

The fraud we are looking for in the dataset are the outliers in the property data by focusing on the high and low values which would help us identify in our unsupervised analysis in fraud detection. We first created three variables, ltsize, bldsize, bldvol, to calculate each property's lot size, building size, and building volume by multiplying related numerical fields. To prevent dividing any number by zero, we created epsilon, which is an arbitrary small number.

1. ltsize (lot size) = LTFRONT * LTDEPTH + epsilon
2. bldsize (building size) = BLDFRONT * BLDDEPTH + epsilon
3. bldvol (building volume) = bldsize * STORIES + epsilon

- Ltsize: a numerical variable denoting the lot size by multiplying the lot width by the lot depth for each property record
- Bldsize: a numerical variable representing the building size by multiplying the building frontage and depth
- Bldvol: a numerical variable indication the building volume by multiplying the building size by the number of stories

Then, we created 9 variables by dividing certain fields, FULLVAL, AVLAND, and AVTOT, using the three variables we calculated above.

r1 = FULLVAL/ltsize: Ratio of full value to lot size.

r2 = FULLVAL/bldsize: Ratio of full value to building size.

r3 = FULLVAL/bldvol: Ratio of full value to building volume.

r4 = AVLAND/ltsize: Ratio of assessed land value to lot size.

r5 = AVLAND/bldsize: Ratio of assessed land value to building size.

r6 = AVLAND/bldvol: Ratio of assessed land value to building volume.

r7 = AVTOT/ltsize: Ratio of total assessed value to lot size.

r8 = AVTOT/bldsize: Ratio of total assessed value to building size.

r9 = AVTOT/bldvol: Ratio of total assessed value to building volume.

These variables are normalized by their medians to center their values around 1 for comparability.

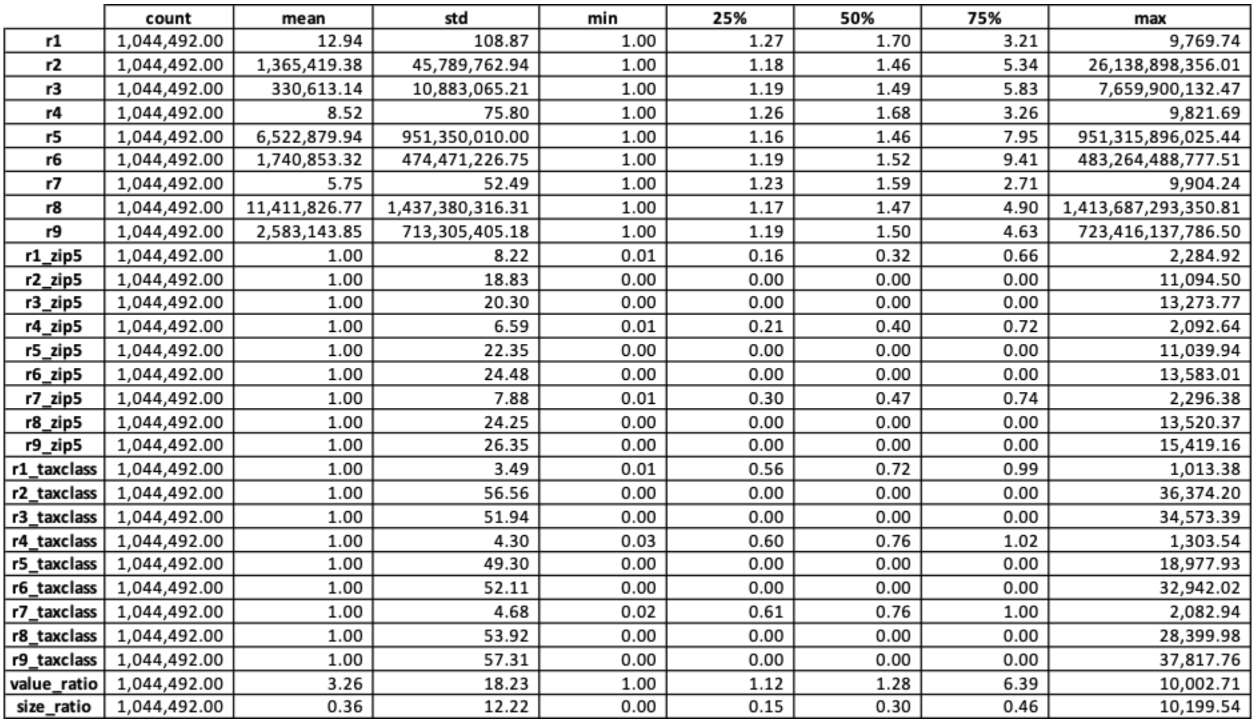
To capture low outliers, the inverse of each variable is calculated, and the maximum value between the original and its inverse is retained to emphasize both high and low extremes. Additional standardization was used by grouping the data by logical categories such as ZIP codes and tax classes, creating variables like r1_zip5 and r1_taxclass that adjust ratios by local averages, highlighting anomalies within specific contexts. Also, more variables are created, such as value_ratio, which compares the full value to the sum of assessed land and total assessed value (standardized to amplify outliers), and size_ratio, which compares building size to lot size to detect unusual configurations. Unnecessary columns, including identifiers and intermediate calculations, are dropped to streamline the dataset.

Dimensionality Reduction

All created variables above are standardized using z-scores to ensure uniform scaling. Furthermore, principal component analysis (PCA) is conducted to reduce dimensionality while retaining 99% of variance. This thorough feature engineering process is designed to detect unusual property patterns indicative of potential fraud or anomalies.

Table 3

Variable Created



	count	mean	std	min	25%	50%	75%	max
r1	1,044,492.00	12.94	108.87	1.00	1.27	1.70	3.21	9,769.74
r2	1,044,492.00	1,365,419.38	45,789,762.94	1.00	1.18	1.46	5.34	26,138,898,356.01
r3	1,044,492.00	330,613.14	10,883,065.21	1.00	1.19	1.49	5.83	7,659,900,132.47
r4	1,044,492.00	8.52	75.80	1.00	1.26	1.68	3.26	9,821.69
r5	1,044,492.00	6,522,879.94	951,350,010.00	1.00	1.16	1.46	7.95	951,315,896,025.44
r6	1,044,492.00	1,740,853.32	474,471,226.75	1.00	1.19	1.52	9.41	483,264,488,777.51
r7	1,044,492.00	5.75	52.49	1.00	1.23	1.59	2.71	9,904.24
r8	1,044,492.00	11,411,826.77	1,437,380,316.31	1.00	1.17	1.47	4.90	1,413,687,293,350.81
r9	1,044,492.00	2,583,143.85	713,305,405.18	1.00	1.19	1.50	4.63	723,416,137,786.50
r1_zip5	1,044,492.00	1.00	8.22	0.01	0.16	0.32	0.66	2,284.92
r2_zip5	1,044,492.00	1.00	18.83	0.00	0.00	0.00	0.00	11,094.50
r3_zip5	1,044,492.00	1.00	20.30	0.00	0.00	0.00	0.00	13,273.77
r4_zip5	1,044,492.00	1.00	6.59	0.01	0.21	0.40	0.72	2,092.64
r5_zip5	1,044,492.00	1.00	22.35	0.00	0.00	0.00	0.00	11,039.94
r6_zip5	1,044,492.00	1.00	24.48	0.00	0.00	0.00	0.00	13,583.01
r7_zip5	1,044,492.00	1.00	7.88	0.01	0.30	0.47	0.74	2,296.38
r8_zip5	1,044,492.00	1.00	24.25	0.00	0.00	0.00	0.00	13,520.37
r9_zip5	1,044,492.00	1.00	26.35	0.00	0.00	0.00	0.00	15,419.16
r1_taxclass	1,044,492.00	1.00	3.49	0.01	0.56	0.72	0.99	1,013.38
r2_taxclass	1,044,492.00	1.00	56.56	0.00	0.00	0.00	0.00	36,374.20
r3_taxclass	1,044,492.00	1.00	51.94	0.00	0.00	0.00	0.00	34,573.39
r4_taxclass	1,044,492.00	1.00	4.30	0.03	0.60	0.76	1.02	1,303.54
r5_taxclass	1,044,492.00	1.00	49.30	0.00	0.00	0.00	0.00	18,977.93
r6_taxclass	1,044,492.00	1.00	52.11	0.00	0.00	0.00	0.00	32,942.02
r7_taxclass	1,044,492.00	1.00	4.68	0.02	0.61	0.76	1.00	2,082.94
r8_taxclass	1,044,492.00	1.00	53.92	0.00	0.00	0.00	0.00	28,399.98
r9_taxclass	1,044,492.00	1.00	57.31	0.00	0.00	0.00	0.00	37,817.76
value_ratio	1,044,492.00	3.26	18.23	1.00	1.12	1.28	6.39	10,002.71
size_ratio	1,044,492.00	0.36	12.22	0.00	0.15	0.30	0.46	10,199.54

Anomaly Detection Algorithms

We used two algorithms for our unsupervised fraud modeling. Prior to using either algorithm, the principal components that we created through PCA are z-scaled to give equal importance to each principal component. Equal importance is achieved as we standardize the PCs to ensure that the means are centered at 0 and variances to 1. The z-scaled data (data_pca_zs) retains only the most meaningful components to measure mostly independent phenomena, making the unsupervised analysis more effective.

The first algorithm we used was the Minkowski Distance Scoring Algorithm. Uses a Minkowski distance formula with power p1 (commonly set to 2, equivalent to Euclidean distance) to summarize the magnitude of deviations in the z-scaled PCs. Large scores highlight records with high-dimensional displacements.

This code calculates a Minkowski Distance Score, referred to as score1, for each record in the dataset based on the z-scaled principal components (data_pca_zs). The Minkowski power p1 we used was set to 2, making this calculation equivalent to the Euclidean distance. The calculation takes the absolute values of all principal component scores, raising them to the power p1, summing these powered values across all dimensions for each record, and then scaling the result back using the reciprocal power oop1=1/p1. This final transformation ensures the computed distances are in the correct magnitude range. The result is a vector of Minkowski distance scores (score1), that represents the degree of deviation of a record from the origin in principal component space. Larger scores indicate stronger deviations that potentially identify anomalies or outliers.

The second algorithm we used is the Autoencoder Reconstruction Error. This code implements an autoencoder-based anomaly detection algorithm to calculate a second anomaly score, score2, which is combined with the Minkowski distance-based score (score1) to create a final anomaly ranking. A simple autoencoder is trained using an MLPRegressor with a single hidden layer of 3 neurons and logistic activation, using z-scaled principal components (data_pca_zs) as both input and target output. The autoencoder compresses and reconstructs the data, with 100 max iterations to focus on identifying patterns rather than perfect reconstruction. After training, the autoencoder predicts reconstructed data, and the reconstruction error is calculated as the difference between input and output. This error is processed using a Minkowski distance with power p2 = 2 (Euclidean distance) to compute score2, which identifies records with poor reconstruction accuracy, signaling potential anomalies.

Both scores from the two algorithms are then rank-order scaled to normalize the scores for comparability. A final anomaly score is computed by taking the weighted average of the rank-ordered scores by equally weighting both methods. Records are then sorted by this final score, with the highest-ranked entries considered the most anomalous. This approach effectively combines the sensitivity of an autoencoder to unusual patterns with the interpretability of a distance-based measure, creating a robust and balanced method for detecting anomalies.

Result

As mentioned above, the final anomaly score is a weighted average of two rank-ordered scores: Minkowski Distance Score (score1) and Autoencoder Reconstruction Error Score (score2). The score reflects the degree of anomaly or unusual behavior associated with a property, with higher scores indicating a greater likelihood of fraud or irregularities. These scores are calculated using unsupervised learning techniques based on features in the dataset. By setting specific thresholds, users can identify top anomalies or flag potential cases for further investigation. Additionally, heatmap visualizations provide insights into the variables that most significantly influence high scores, aiding in interpreting and prioritizing these findings effectively. To further examine the top properties, we can access the top_records dataframe, which contains the highest-scoring records, and compare records with the baseline data.

Case study 1 - Heatmap Analysis

One of the key tools for understanding the drivers behind high anomaly scores is the heatmap. By visualizing the absolute values of the variables contributing to these scores, the heatmap highlights patterns and clusters that reveal which features are most strongly associated with anomalies. For example, significant deviations in property tax values or unusual mortgage amounts might be prominent in highly scored records. In this case, analysts could focus on investigating properties where these financial indicators deviate considerably from the norm. Such insights are crucial for identifying systemic anomalies that could point to fraud or data errors, allowing for targeted interventions.

Case study 2 - Score Sensitivity

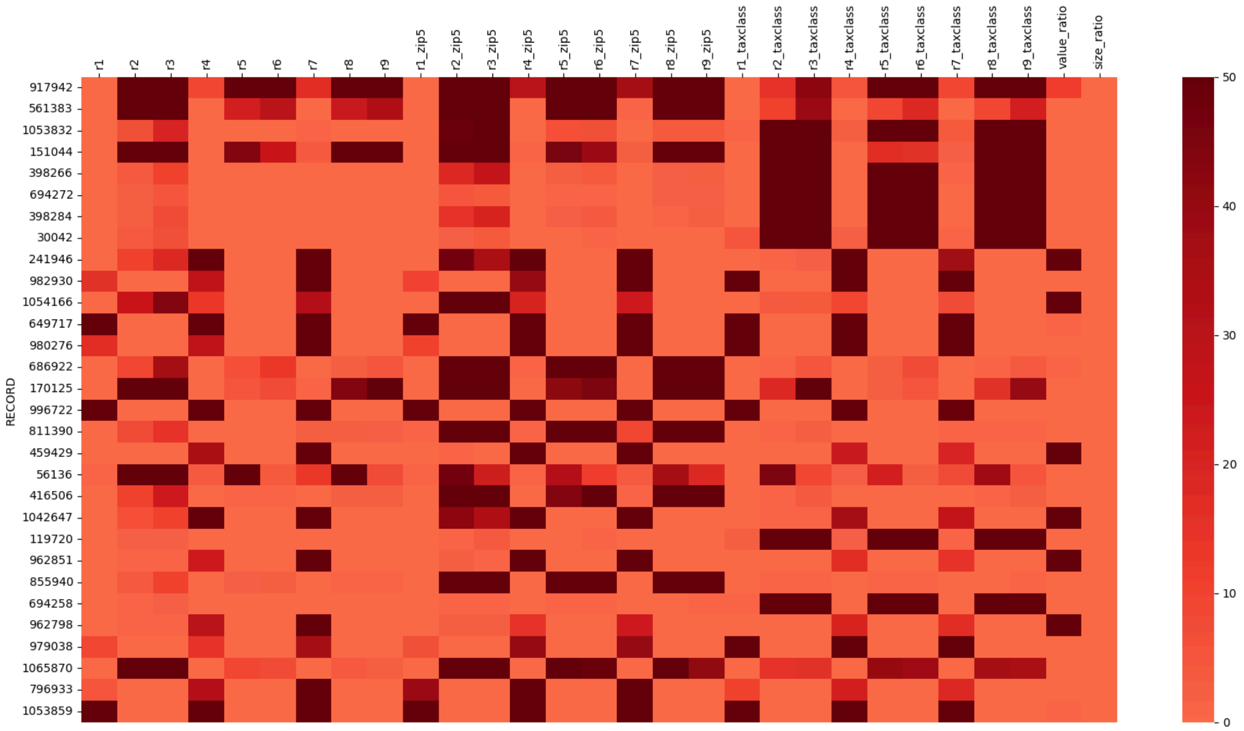
Another valuable approach involves analyzing the sensitivity of the anomaly scores to changes in algorithm parameters, such as the Minkowski distance metric. By varying these parameters and observing the resulting top-scoring properties, analysts can determine which records consistently appear as outliers. Properties that remain in the top anomaly list across different configurations are especially notable, as they may represent true anomalies rather than artifacts of specific model settings. This robustness testing ensures that the identified anomalies are not overly dependent on the choice of parameters, providing a higher degree of confidence in the results.

Case study 3 - Historical Comparison

Comparing current anomaly scores with historical data offers a dynamic perspective on potential issues. This approach can reveal new anomalies that have emerged or detect persistent anomalies that remain unresolved over time. For instance, properties flagged as high anomalies in both current and previous analyses may warrant further investigation to determine if they represent ongoing fraud or systemic issues in the data. Conversely, identifying properties that were previously anomalous but no longer are can indicate successful resolution or correction. This temporal analysis adds a valuable layer of context to the evaluation of anomalies, aiding in trend detection and follow-up prioritization.

Figure 5

Heatmap of Variables & Records



Summary

This project utilized unsupervised learning techniques to identify anomalies in a dataset of New York City properties, aiming to detect potential fraud or unusual behavior. We developed a scoring mechanism, where higher scores indicate a greater likelihood of anomalies. Further, a heatmap was created to visualize all key features driving these scores, enabling a clearer understanding of the variables most strongly associated with anomalies. The top-scoring records were extracted for further investigation, and methods for comparing these results with historical baselines were implemented to assess the stability of the algorithm's performance under different configurations.

The results provided insights into the most anomalous properties, highlighting unusual patterns such as discrepancies in financial indicators like taxes or mortgages. Heatmaps revealed clusters of variables contributing to high scores, enabling targeted analysis of potential outliers. Moreover, sensitivity testing showed the robustness of the scores across different algorithm parameter settings, ensuring that identified anomalies were not artifacts of specific modeling choices. Historical comparisons also allowed for tracking trends over time, distinguishing between persistent and emerging anomalies.

To refine the algorithm with expert feedback, analysts can adjust the weighting of variables or exclude certain features deemed irrelevant or noisy. For example, domain experts might prioritize financial variables or exclude attributes known to correlate weakly with fraudulent behavior. Adjustments to the algorithm's distance metric or threshold for anomaly detection can also be informed by expert judgment, tailoring the model to specific types of anomalies. These changes ensure the results remain relevant and actionable. This process creates a feedback loop where experts iteratively modify the algorithm based on its outputs and their domain knowledge, leading to increasingly accurate and meaningful anomaly detection. The ability to adjust the algorithm and test the impact of these changes on the results ensures that it can adapt to evolving data patterns and expert priorities, making it a powerful tool for proactive fraud detection and prevention.

Appendix

Dataset Overview

This dataset contains 1,070,994 records of New York City properties, featuring 32 fields that include both categorical and numeric data. With 1,070,994 records, this dataset provides a substantial resource for analyzing anomalies/strangeness in the data to detect potential fraudulent activity.

Filed description

1. Record: An ordinal identifier assigned to each property entry.
2. BBLE: A categorical field key representing the Borough, Block, Lot, and Easement code.
3. BORO: A categorical field specifying New York City boroughs, where 1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, and 5 = Staten Island.
4. BLOCK: A categorical field indicating valid block ranges for each borough: Manhattan (1–2255), Bronx (2260–5958), Brooklyn (1–8955), Queens (1–16350), Staten Island (1–8050).
5. LOT: A categorical field denoting the specific plot of land designated for individual ownership.
6. EASEMENT: A categorical field that indicates whether a property has an easement, which is a legal right for a third party to use a portion of the property for a specific purpose. Values include: Space = No Easement, A = Air Easement, B = Non- Air Rights, E = Land Easement, N = Non-Transit Easement, P = Pier, R = Railroad, S = Street, U= U.S. Government, F-M = duplicates of E.
7. OWNER: The owner’s name for the property record.
8. BLDGCL: The building class of each property record.

9. TAXCLASS: A categorical field indicating the tax class for each property, with the following classifications: 1 = 1-3 Unit Residence, 2 = Apartments, 2A = 4, 5, or 6 Units, 3 = Utilities, 4 = All Others.
10. LTFRONT: A numerical field denoting the lot width for each property record.
11. LTDEPTH: A numerical field specifying the lot depth for each property record.
12. EXT: A categorical extension indicator for each property record.
13. STORIES: A numerical field indicating the number of stories in the building.
14. FULLVAL: A numerical field representing the market value of each property.
15. AVLAND: A numerical field denoting the actual land value.
16. AVTOT: A numerical field representing the actual total value of the property.
17. EXLAND: A numeric field representing the actual exempt land value of the property.
18. EXTOT: A numeric field indicating the total actual exempt land value.
19. EXCD1: A categorical field denoting the exemption code assigned to the property.
20. STADDR: A categorical field containing the street address of the property.
21. ZIP: A categorical field specifying the property's ZIP code.
22. EXMPTCL: A categorical field identifying the exemption class of the property.
23. BLDFRONT: A numeric field indicating the building's width on the property.
24. BLDDEPTH: A numeric field specifying the building's depth on the property.
25. AVLAND2: A numeric field representing the transitional land value of the property.
26. AVTOT2: A numeric field indicating the transitional total value of the property.
27. EXLAND2: A numeric field describing the property's transitional exempt land value.
28. EXCD2: A numeric field indicating the property's transitional exemption land total.

29. PERIOD: A categorical field denoting the property’s assessment period when the data was created. We note that all the information in the property data occurs in the “Final Period”.
30. YEAR: A categorical field representing the assessment year. When plotting the distribution, we note that the assessment year across all properties occurs in 2010/2011.
31. VALTYPE: A categorical field specifying the valuation type of the property. Currently, the valuation type for all rows of data are "AC-TR," which refers to Assessed Value per Acre – Total Revenue.

Summary Statistics Tables

Categorical Field Summary Table

	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	RECORD	categorical	1,070,994	100.0%	0	1,070,994	1
1	BBLE	categorical	1,070,994	100.0%	0	1,070,994	1000010101
2	BORO	categorical	1,070,994	100.0%	0	5	4
3	BLOCK	categorical	1,070,994	100.0%	0	13,984	3,944
4	LOT	categorical	1,070,994	100.0%	0	6,366	1
5	EASEMENT	categorical	4,636	0.4%	0	12	E
6	OWNER	categorical	1,039,249	97.0%	0	863,347	PARKCHESTER PRESERVAT
7	BLDGCL	categorical	1,070,994	100.0%	0	200	R4
8	TAXCLASS	categorical	1,070,994	100.0%	0	11	1
9	EXT	categorical	354,305	33.1%	0	3	G
10	EXCD1	categorical	638,488	59.6%	0	129	1017.00
11	STADDR	categorical	1,070,318	99.9%	0	839,280	501 SURF AVENUE

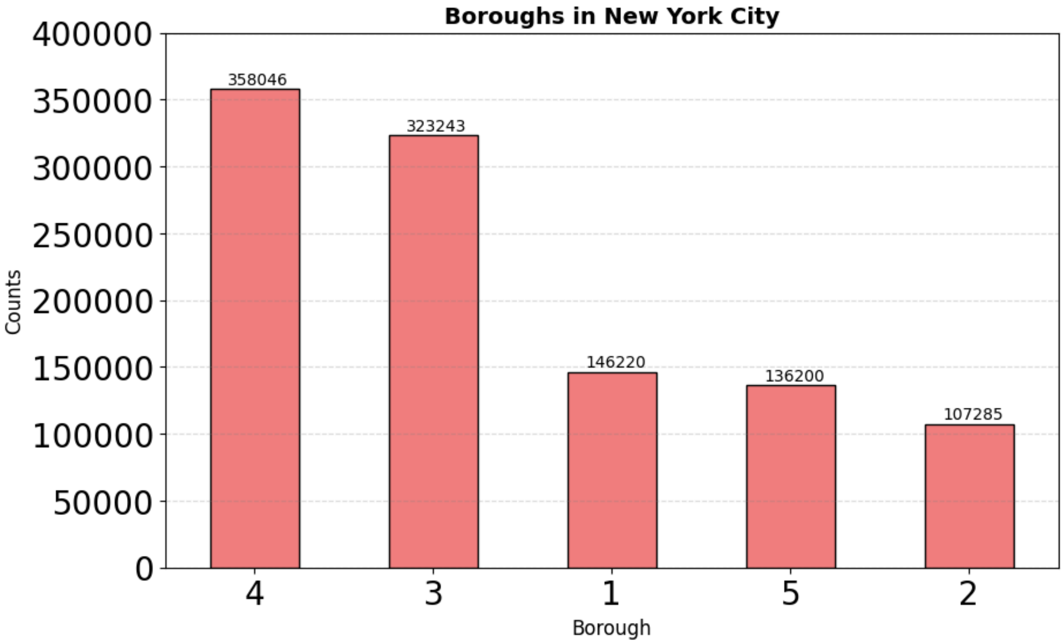
12	ZIP	categorical	1,041,104	97.2%	0	196	10314.00
13	EXMPTCL	categorical	15,579	1.5%	0	14	X1
14	EXCD2	categorical	92,948	8.7%	0	60	1017.00
15	PERIOD	categorical	1,070,994	100.0%	0	1	FINAL
16	YEAR	categorical	1,070,994	100.0%	0	1	2010/11
17	VALTYPE	categorical	1,070,994	100.0%	0	1	AC-TR

Numerical Field Summary Table

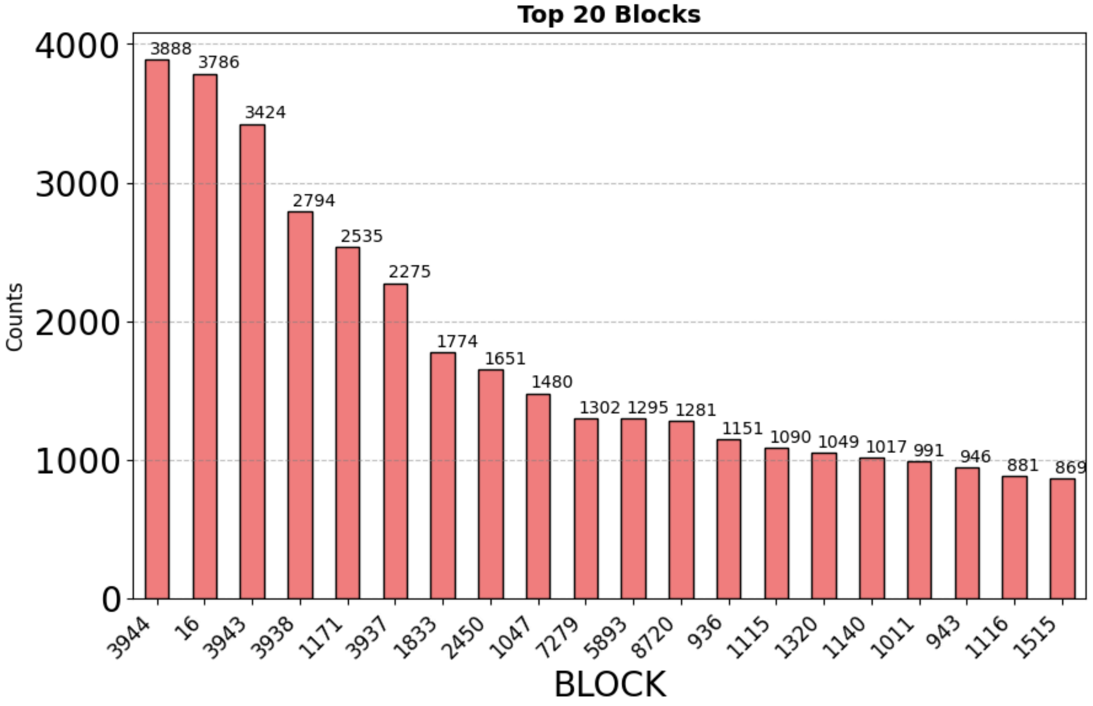
	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
0	LTFRONT	numeric	1,070,994	100.0%	169,108	0.00	9,999.00	36.64	74.03	0.00
1	LTDEPTH	numeric	1,070,994	100.0%	170,128	0.00	9,999.00	88.86	76.40	100.00
2	STORIES	numeric	1,014,730	94.7%	0	1.00	119.00	5.01	8.37	2.00
3	FULLVAL	numeric	1,070,994	100.0%	13,007	0.00	6,150,000,000.00	874,264.51	11,582,425.58	0.00
4	AVLAND	numeric	1,070,994	100.0%	13,009	0.00	2,668,500,000.00	85,067.92	4,057,258.16	0.00
5	AVTOT	numeric	1,070,994	100.0%	13,007	0.00	4,668,308,947.00	227,238.17	6,877,526.09	0.00
6	EXLAND	numeric	1,070,994	100.0%	491,699	0.00	2,668,500,000.00	36,423.89	3,981,573.93	0.00
7	EXTOT	numeric	1,070,994	100.0%	432,572	0.00	4,668,308,947.00	91,186.98	6,508,399.78	0.00
8	BLDFRONT	numeric	1,070,994	100.0%	228,815	0.00	7,575.00	23.04	35.58	0.00
9	BLDDEPTH	numeric	1,070,994	100.0%	228,853	0.00	9,393.00	39.92	42.71	0.00
10	AVLAND2	numeric	282,726	26.4%	0	3.00	2,371,005,000.00	246,235.72	6,178,951.64	2,408.00
11	AVTOT2	numeric	282,732	26.4%	0	3.00	4,501,180,002.00	713,911.44	11,652,508.34	750.00
12	EXLAND2	numeric	87,449	8.2%	0	1.00	2,371,005,000.00	351,235.68	10,802,150.91	2,090.00
13	EXTOT2	numeric	130,828	12.2%	0	7.00	4,501,180,002.00	656,768.28	16,072,448.75	2,090.00

Field Distribution

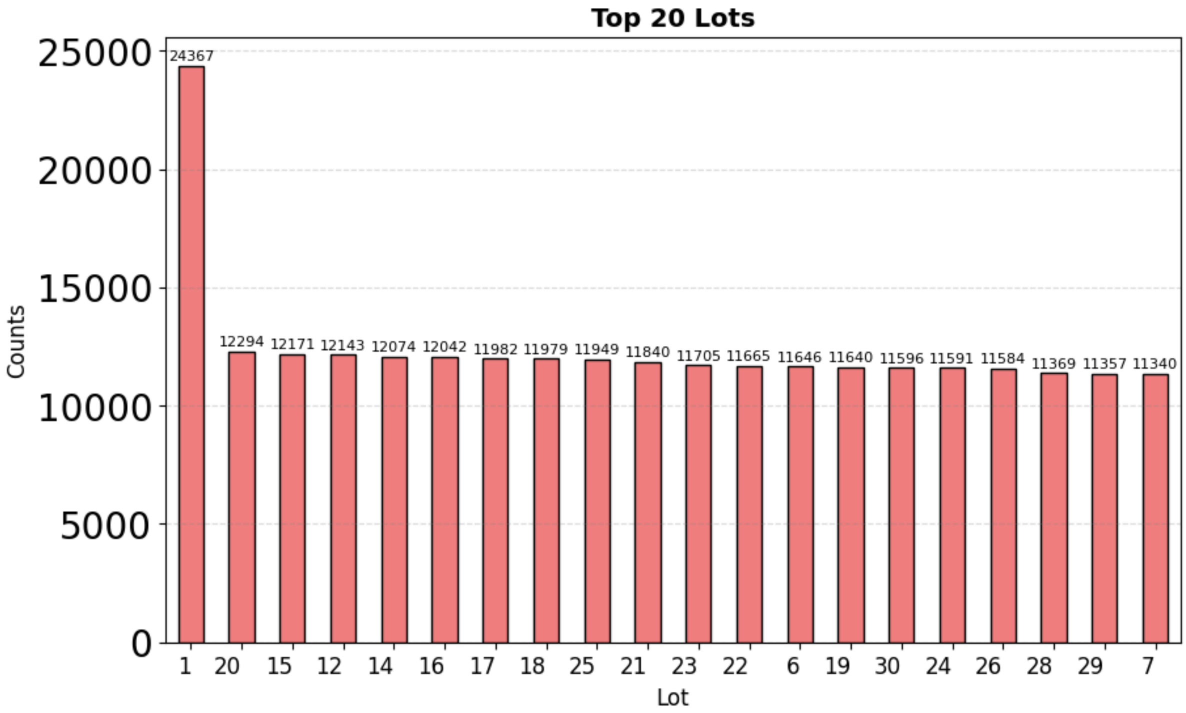
BORO

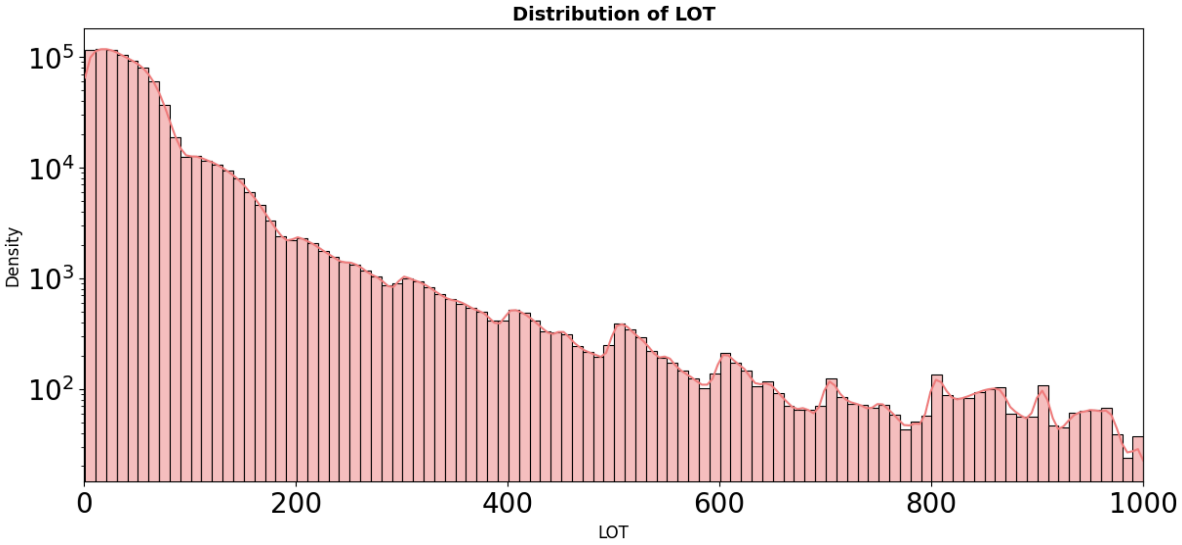


BLOCK

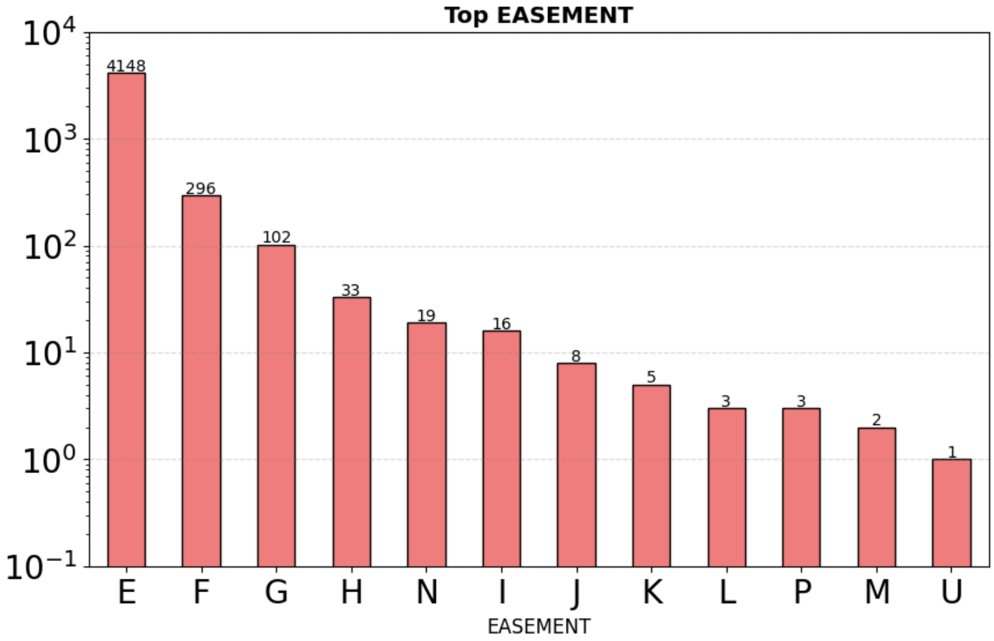


LOT

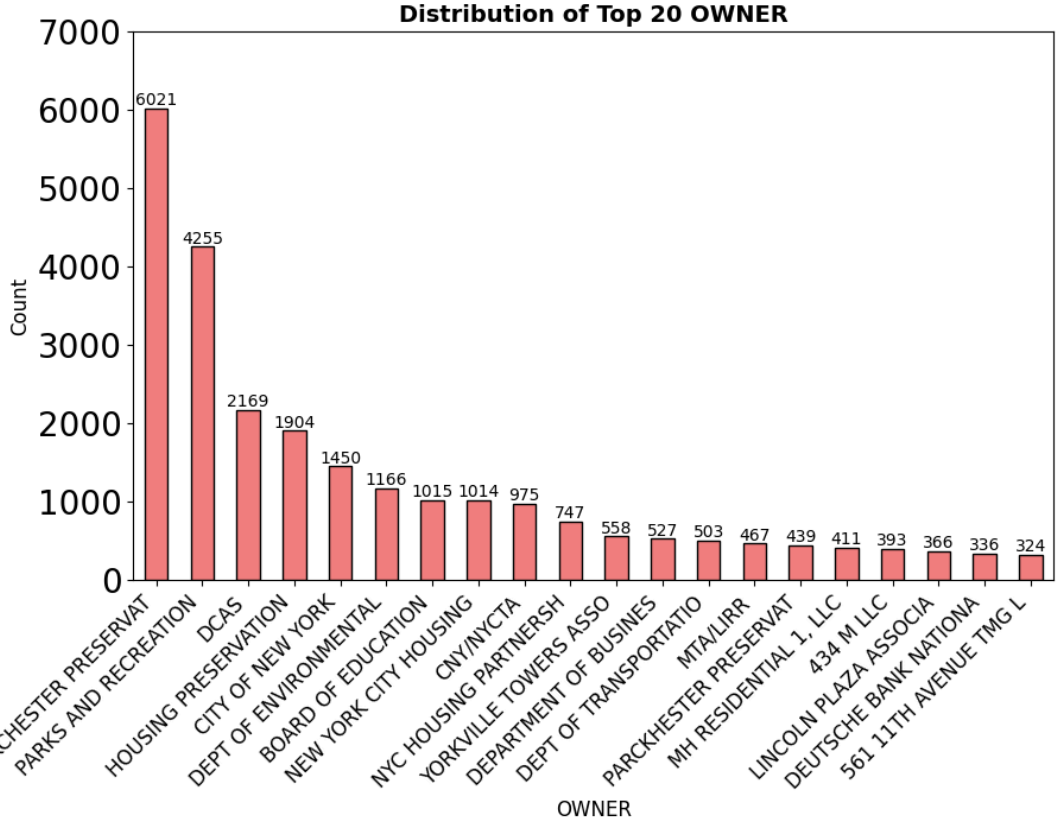

Top 20 Lots
Counts
Lot
25000
20000
15000
10000
5000
0
24367
12294
12171
12143
12074
12042
11982
11979
11949
11840
11705
11665
11646
11640
11596
11591
11584
11369
11357
11340
1
20
15
12
14
16
17
18
25
21
23
22
6
19
30
24
26
28
29
7


Distribution of LOT
Density
LOT
10^5
10^4
10^3
10^2
0
200
400
600
800
1000

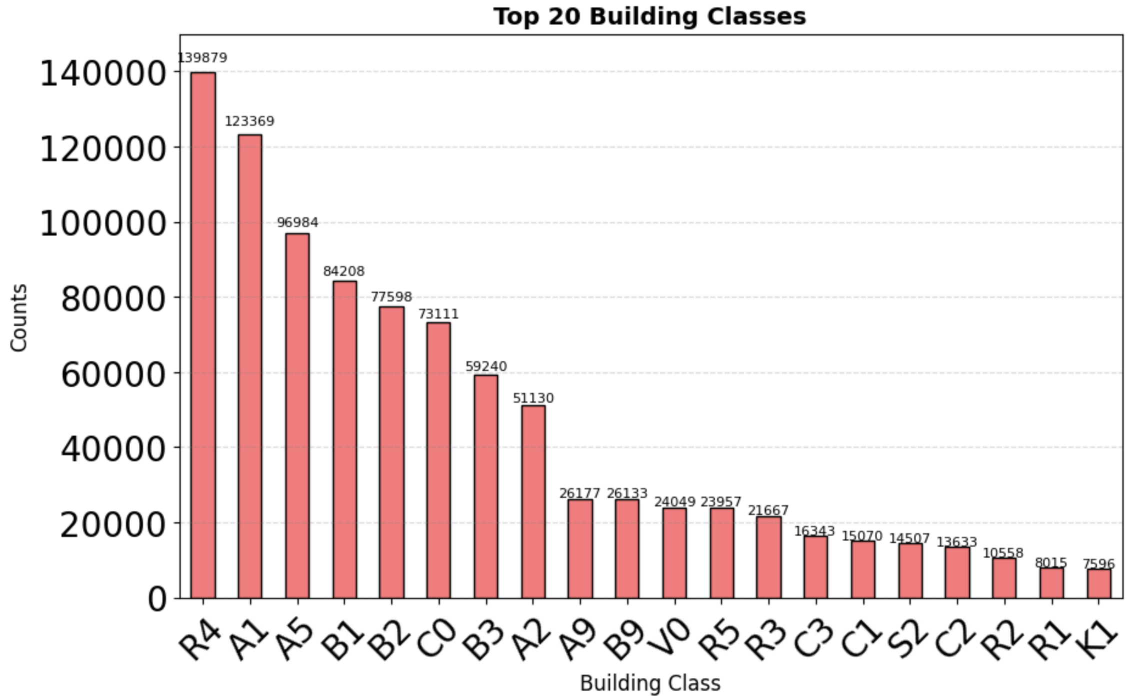
EASEMENT



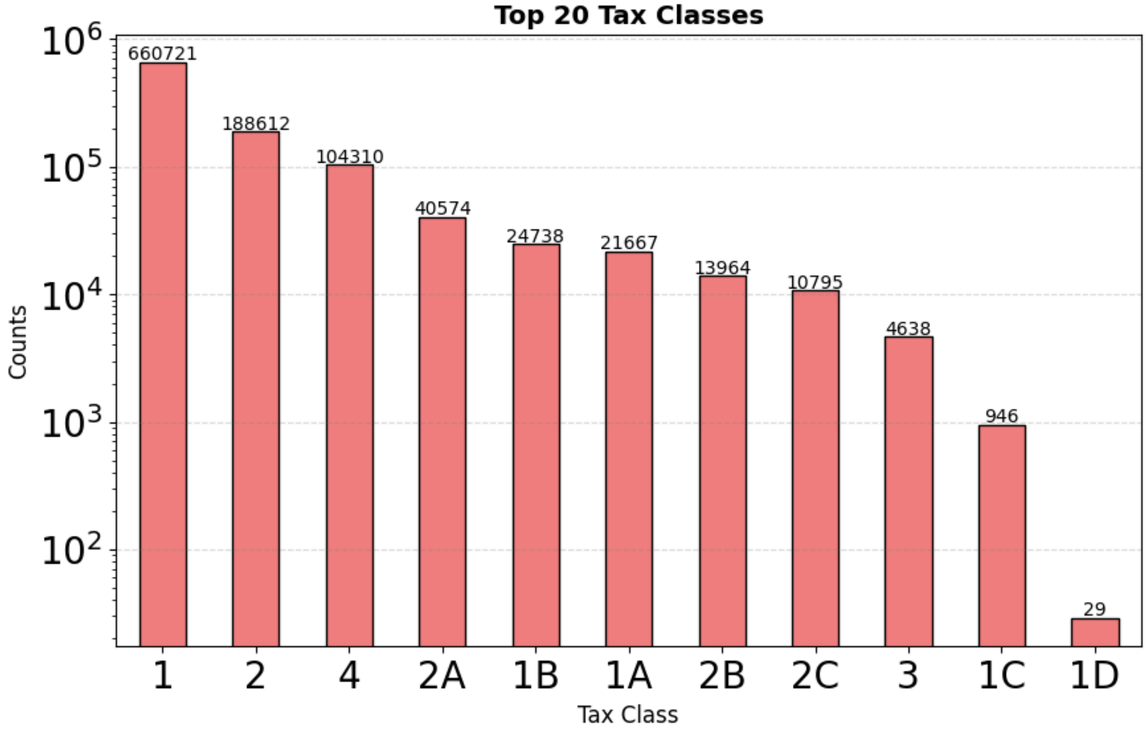
OWNER



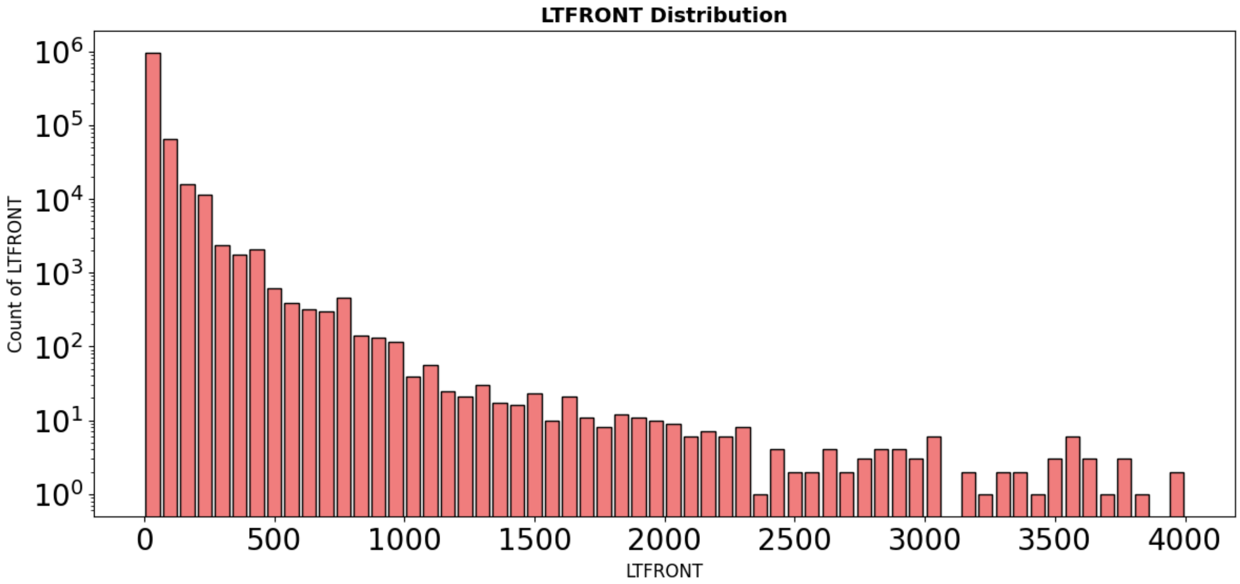
BLDGCL



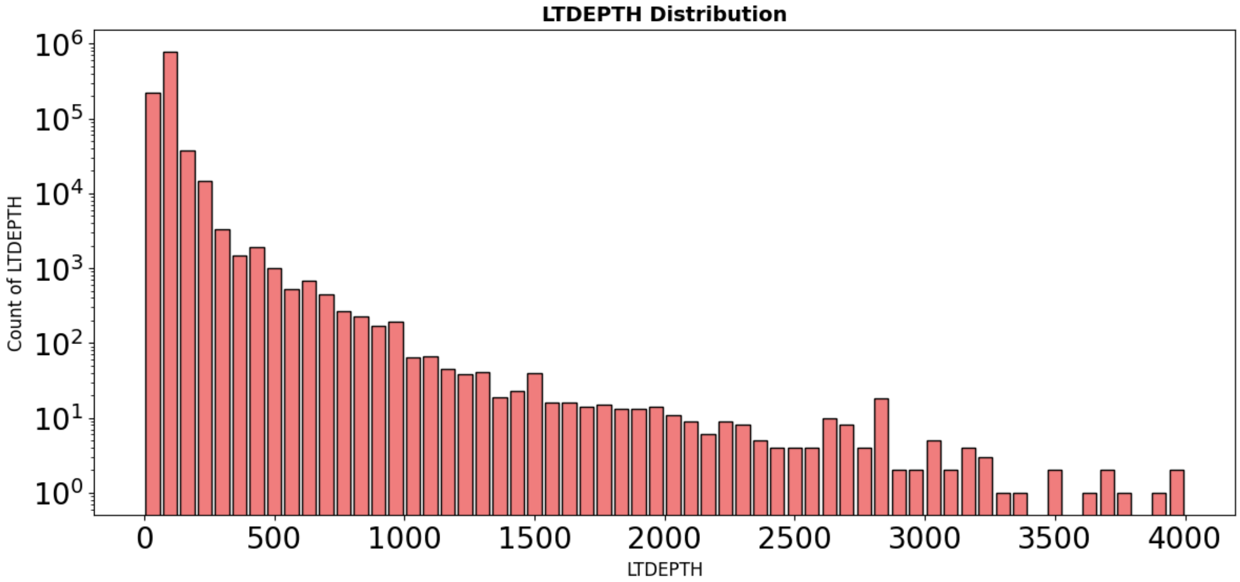
TAXCLASS



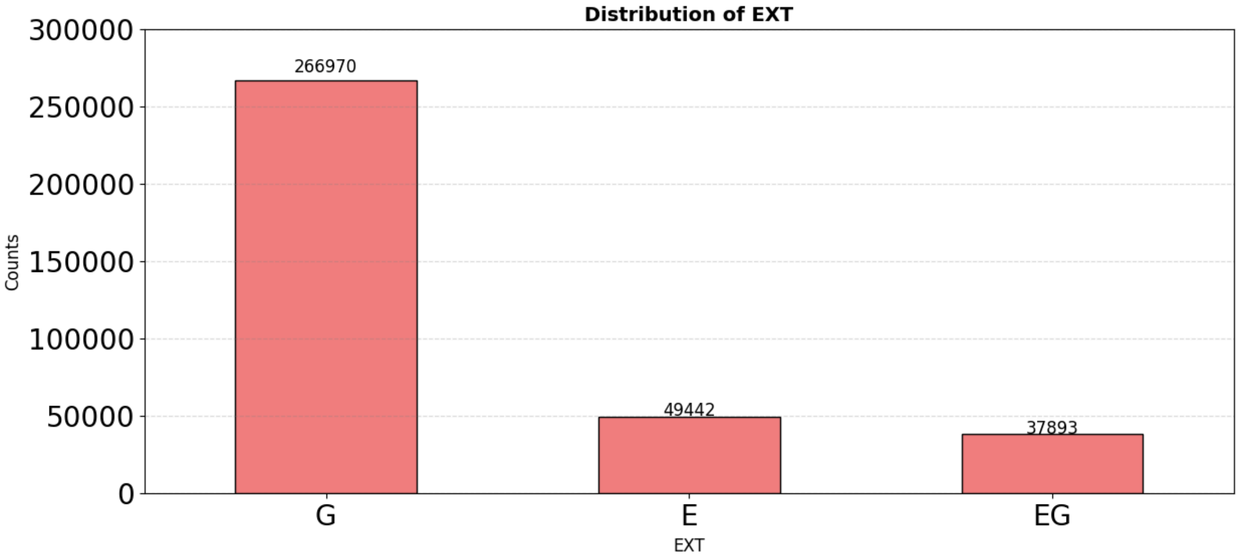
LTFRONT



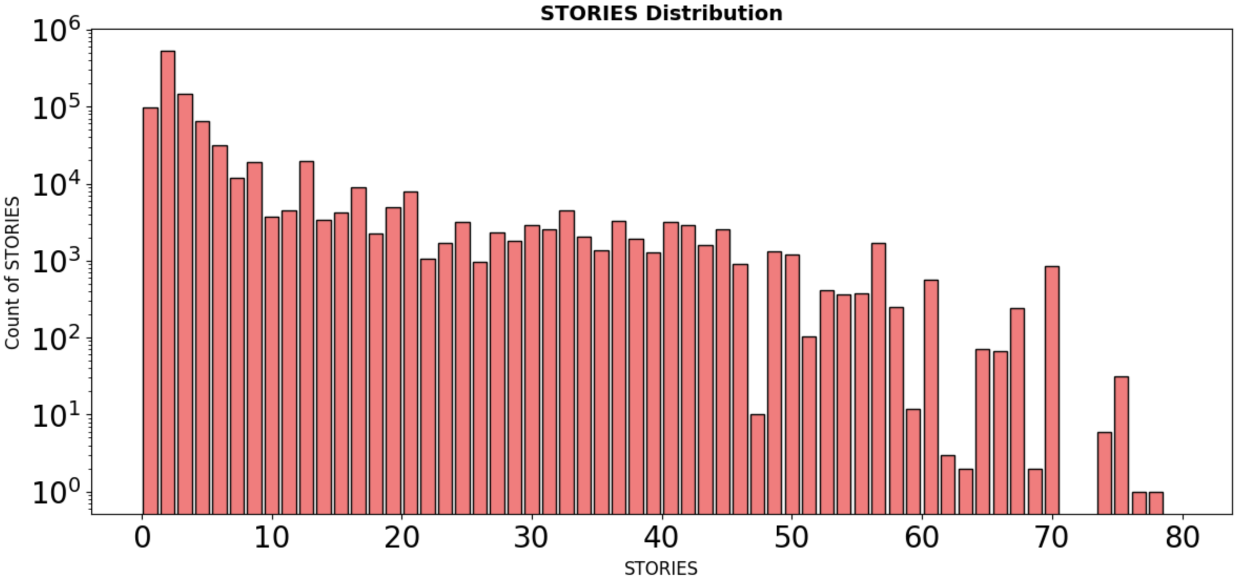
LTDEPTH



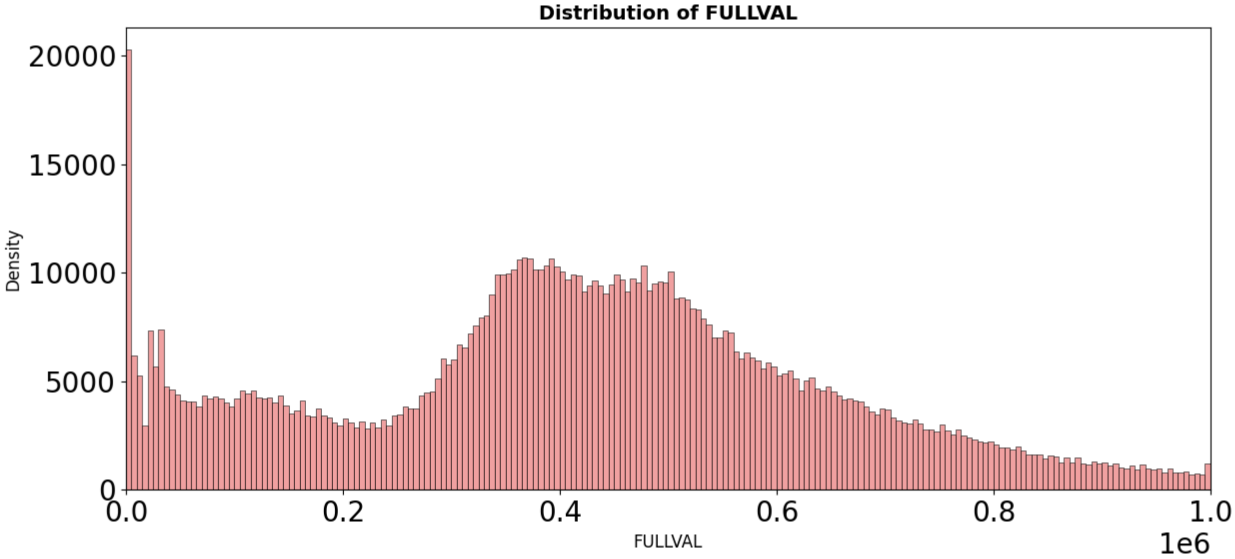
EXT


Distribution of EXT
Counts
300000
250000
200000
150000
100000
50000
0
266970
49442
37893
G
E
EG
EXT

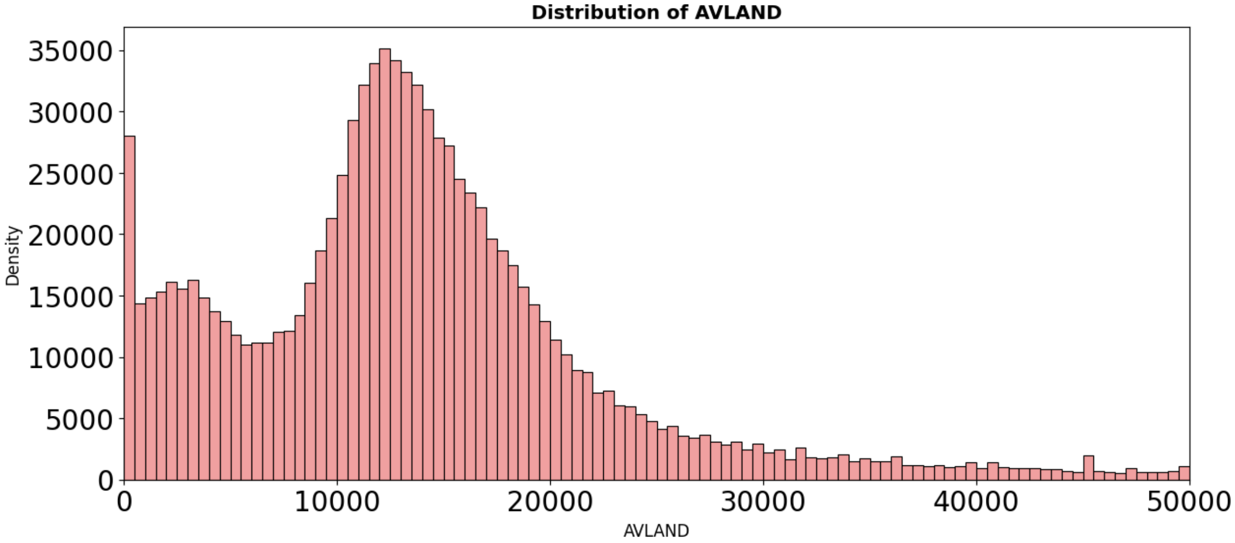
STORIES


STORIES Distribution
Count of STORIES
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
10
20
30
40
50
60
70
80
STORIES

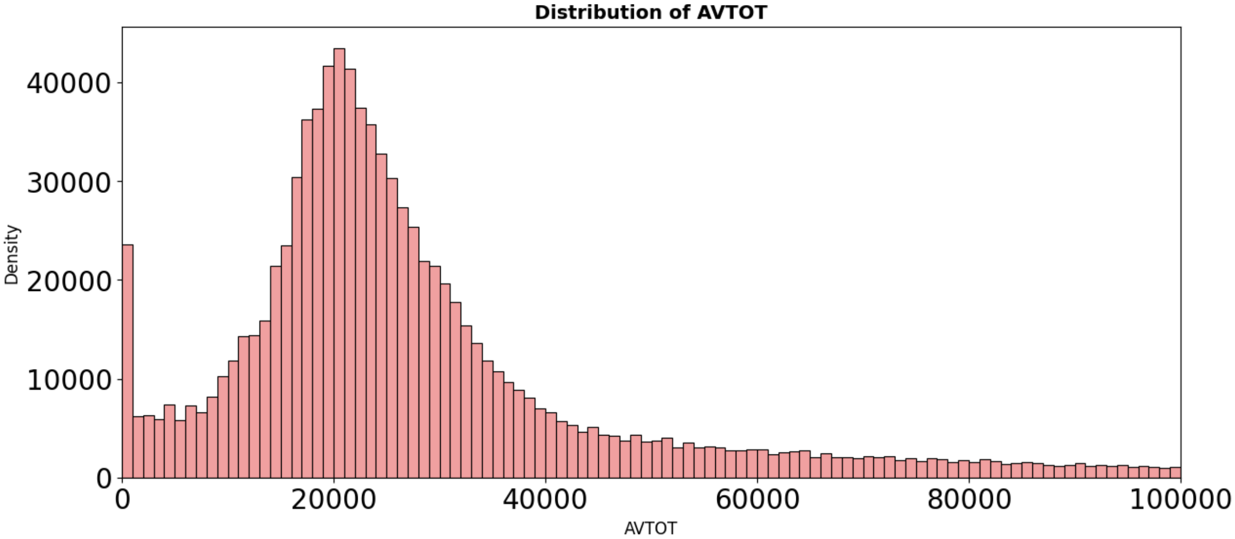
FULLVAL



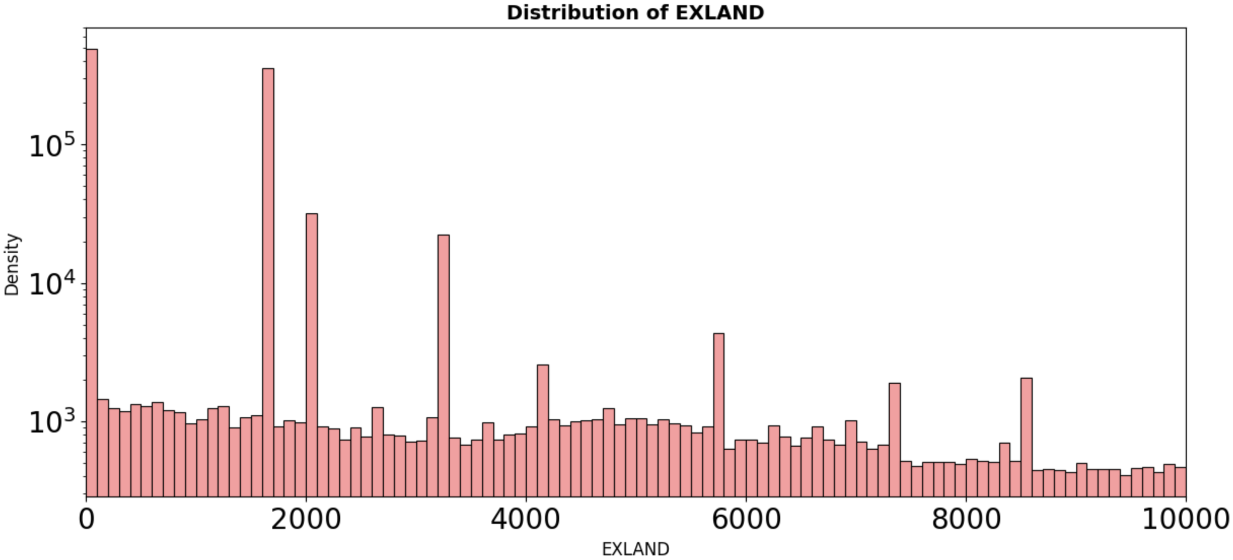
AVLAND



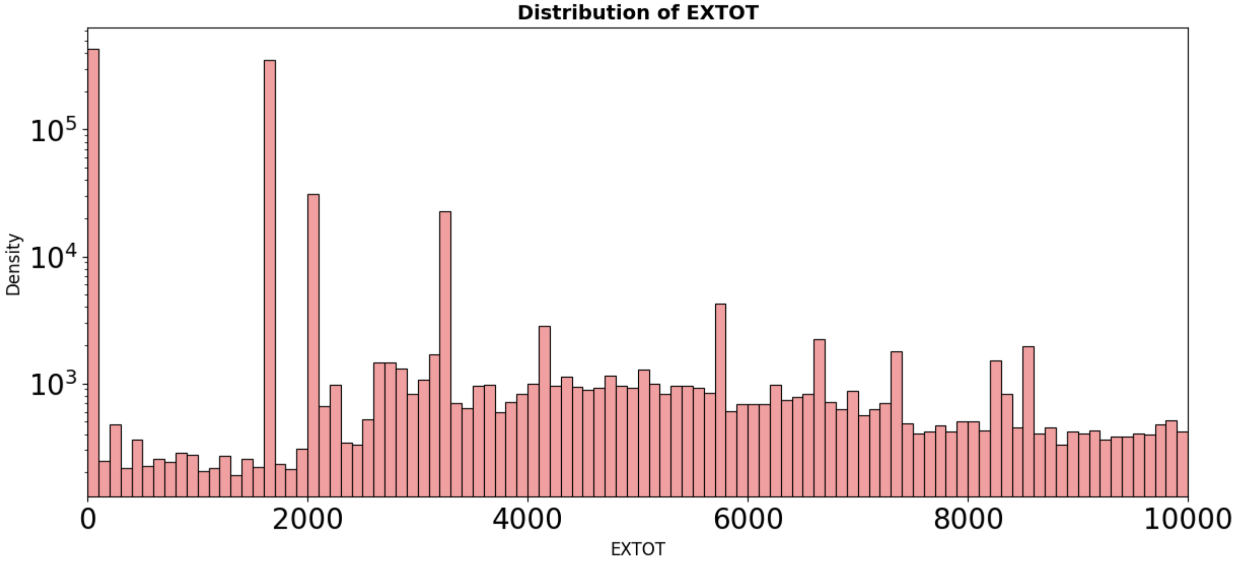
AVTOT



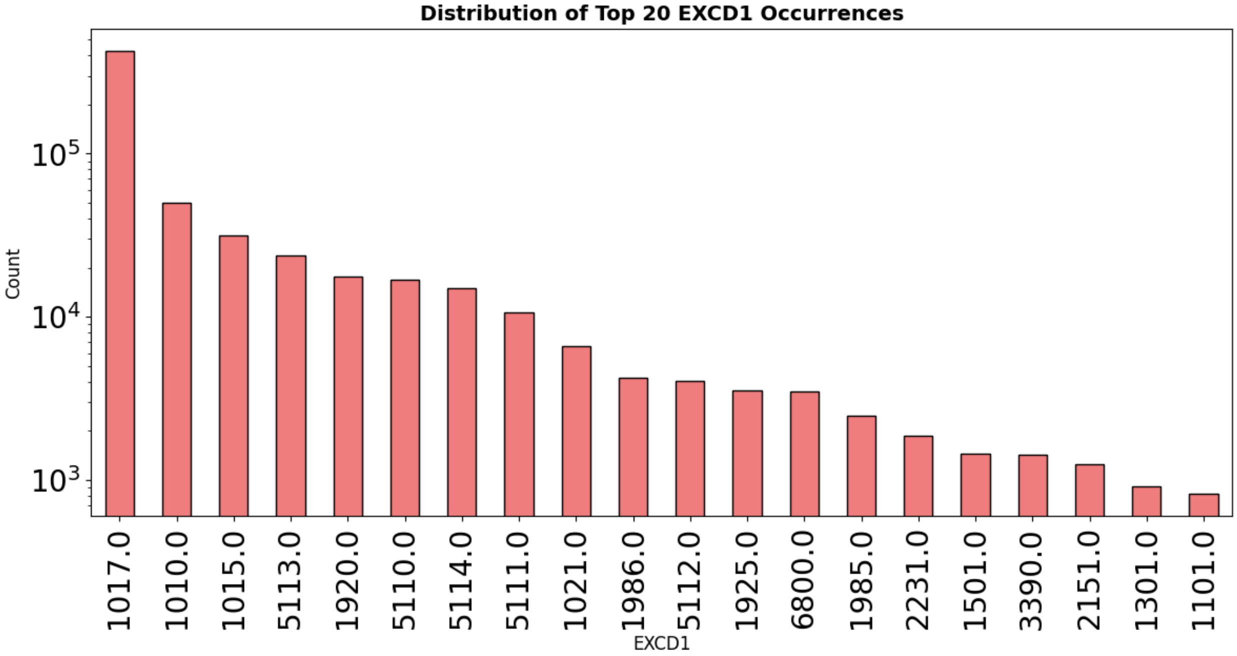
EXLAND



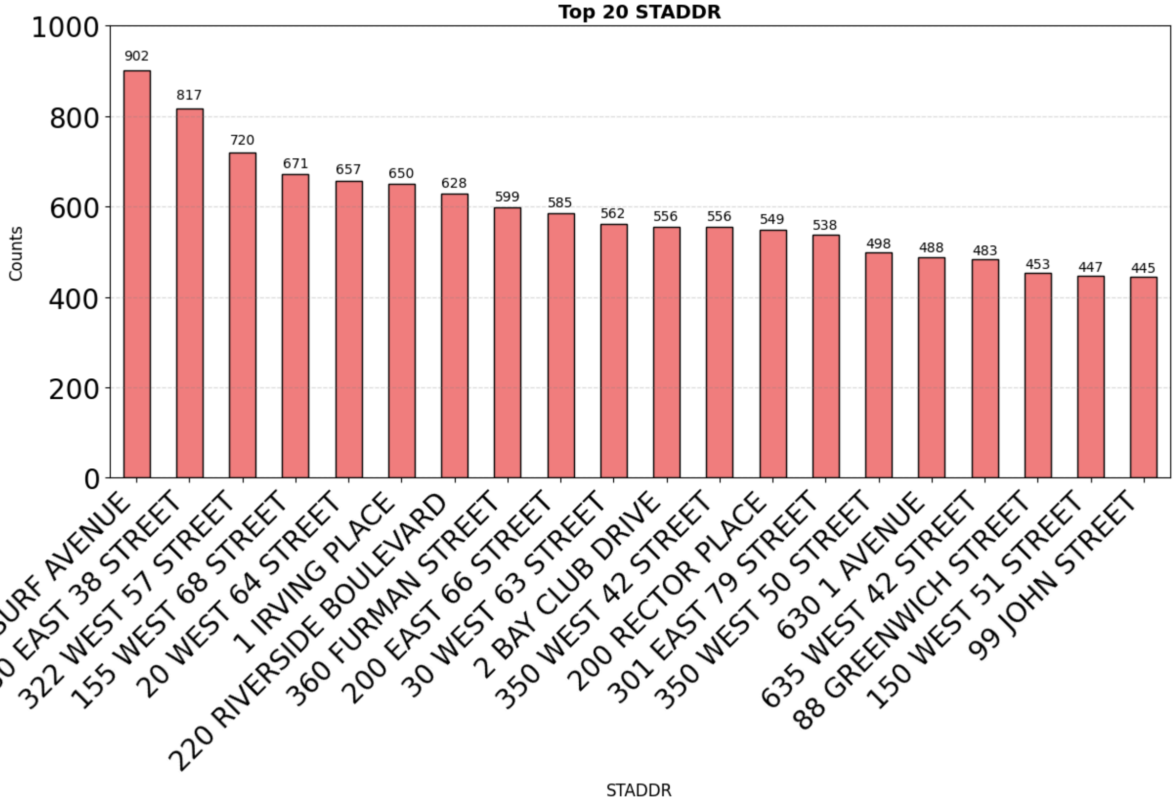
EXTOT


Distribution of EXTOT
Density
10^5
10^4
10^3
0
2000
4000
6000
8000
10000
EXTOT

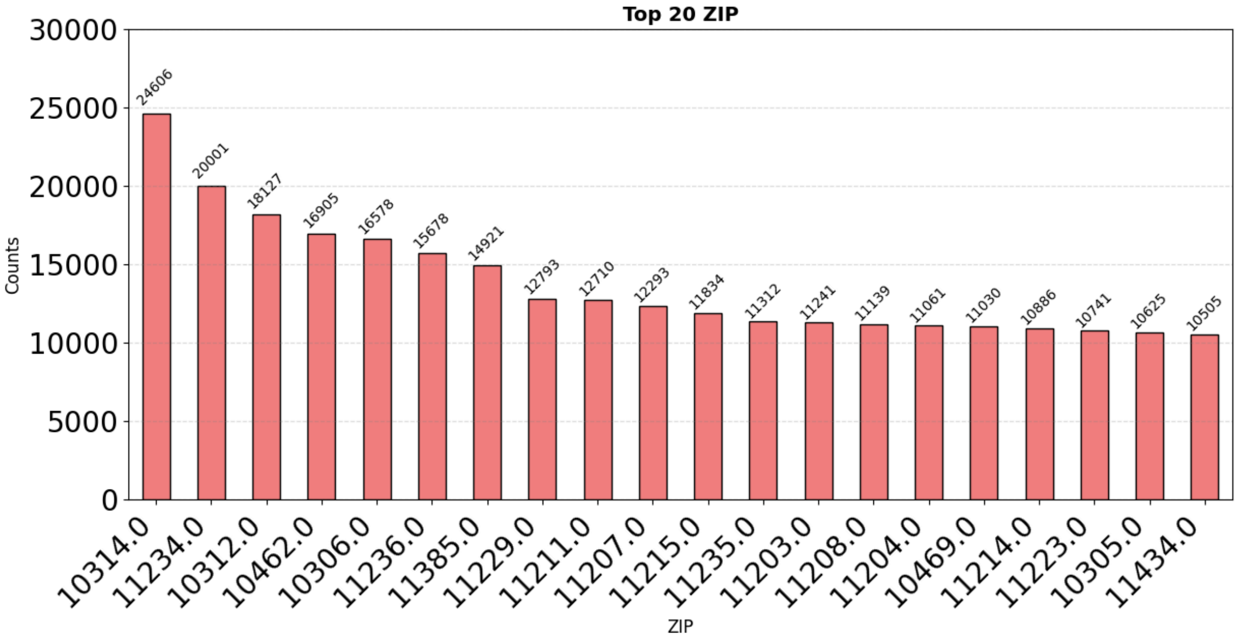
EXCD1


Distribution of Top 20 EXCD1 Occurrences
Count
10^5
10^4
10^3
1017.0
1010.0
1015.0
5113.0
1920.0
5110.0
5114.0
5111.0
1021.0
1986.0
5112.0
1925.0
6800.0
1985.0
2231.0
1501.0
3390.0
2151.0
1301.0
1101.0
EXCD1

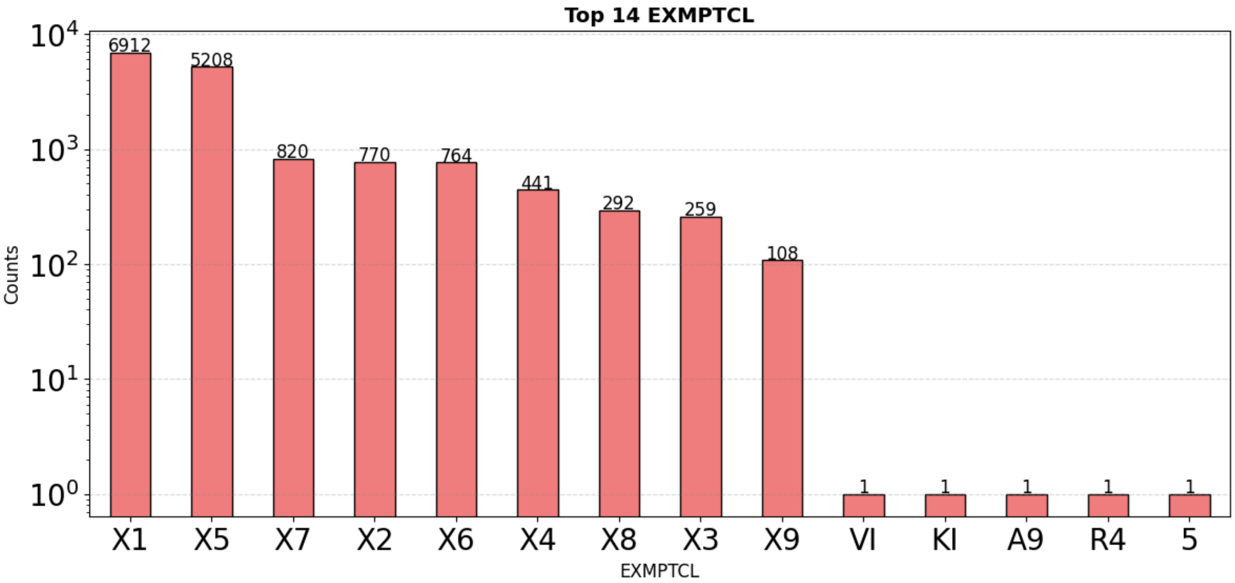
STADDR


Top 20 STADDR
Counts
0
200
400
600
800
1000
902
817
720
671
657
650
628
599
585
562
556
556
549
538
498
488
483
453
447
445
501 SURF AVENUE
330 EAST 38 STREET
322 WEST 57 STREET
155 WEST 68 STREET
20 WEST 64 STREET
1 IRVING PLACE
220 RIVERSIDE BOULEVARD
360 FURMAN STREET
200 EAST 66 STREET
30 WEST 63 STREET
2 BAY CLUB DRIVE
350 WEST 42 STREET
200 RECTOR PLACE
301 EAST 79 STREET
350 WEST 50 STREET
630 1 AVENUE
635 WEST 42 STREET
88 GREENWICH STREET
150 WEST 51 STREET
99 JOHN STREET
STADDR

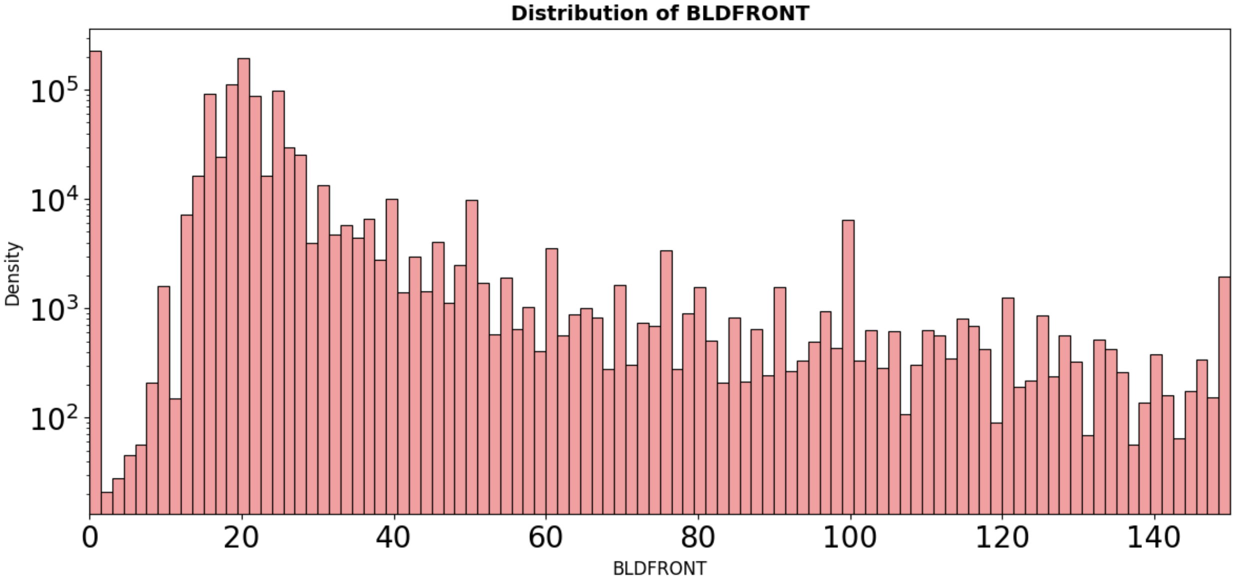
ZIP


Top 20 ZIP
Counts
0
5000
10000
15000
20000
25000
30000
24606
20001
18127
16905
16578
15678
14921
12793
12710
12293
11834
11312
11241
11139
11061
11030
10886
10741
10625
10505
10314.0
11234.0
10312.0
10462.0
10306.0
11236.0
11385.0
11229.0
11211.0
11207.0
11215.0
11235.0
11203.0
11208.0
11204.0
10469.0
11214.0
11223.0
10305.0
11434.0
ZIP

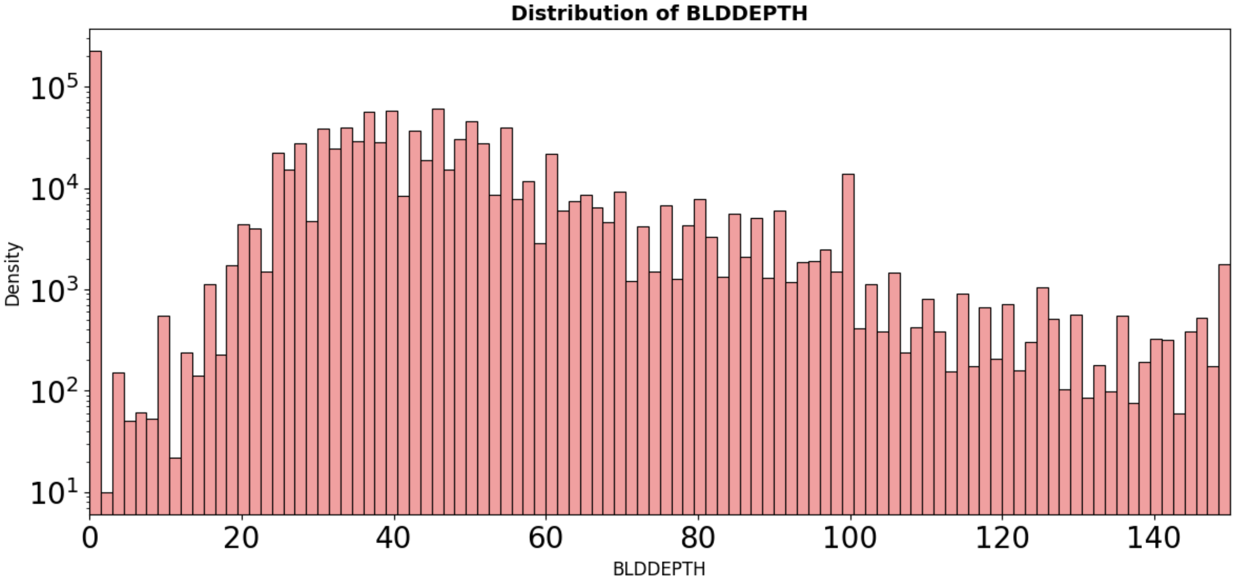
EXMPTCL



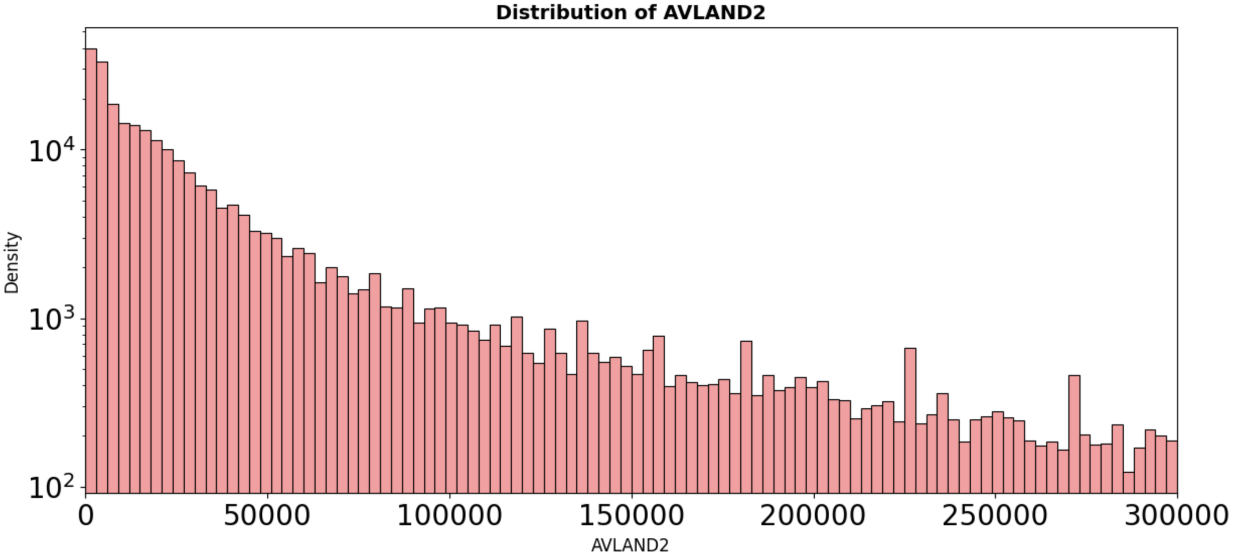
BLDFRONT



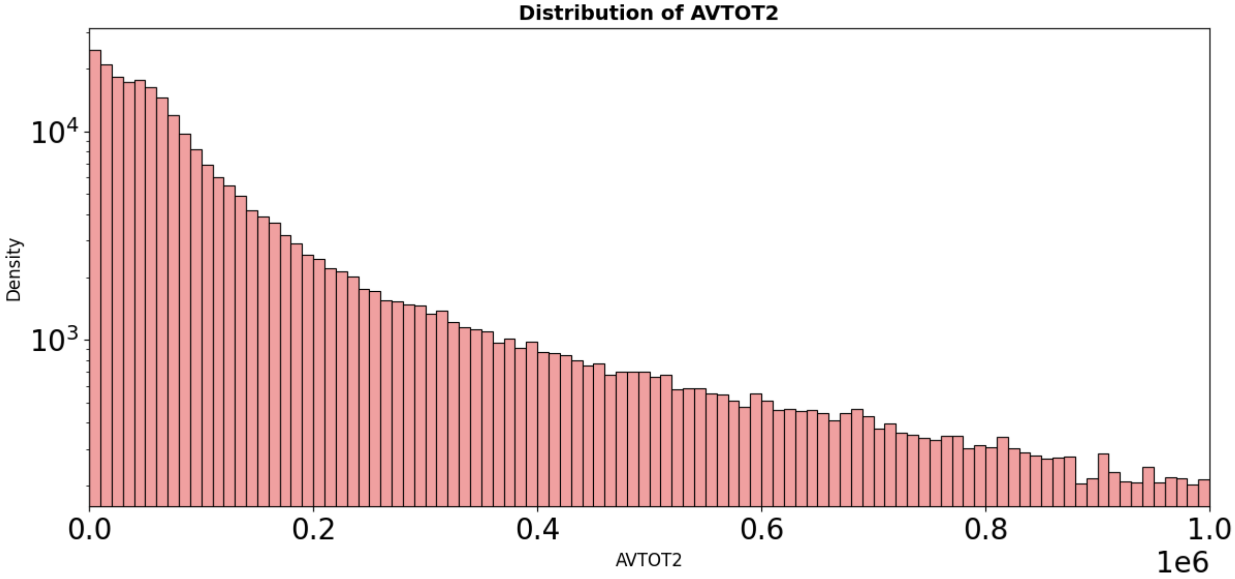
BLDDEPTH



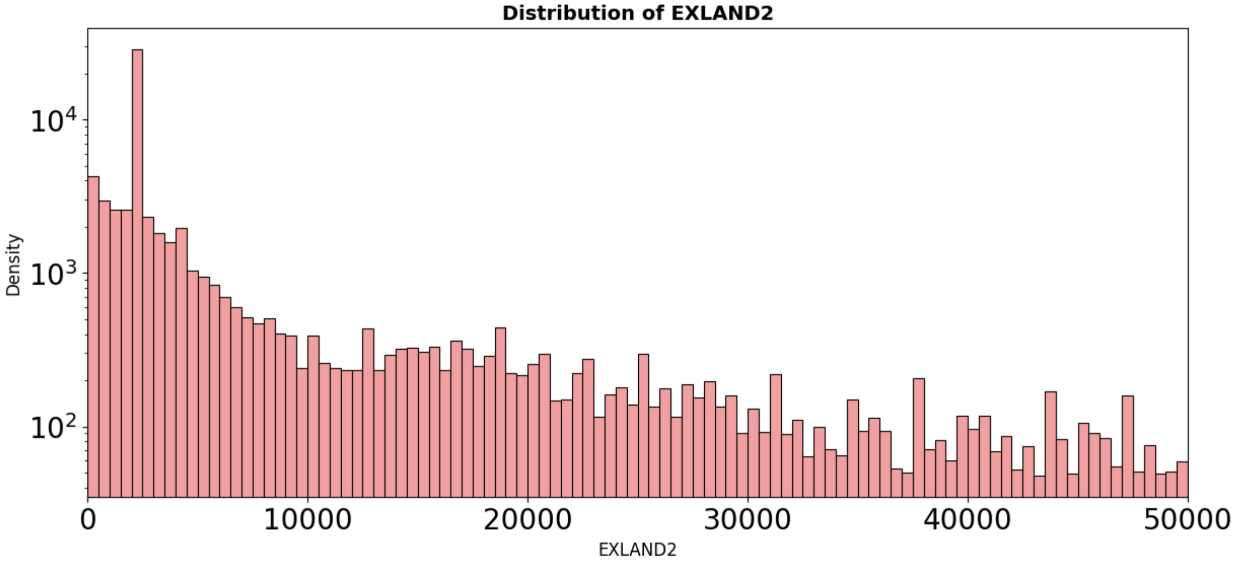
AVLAND2



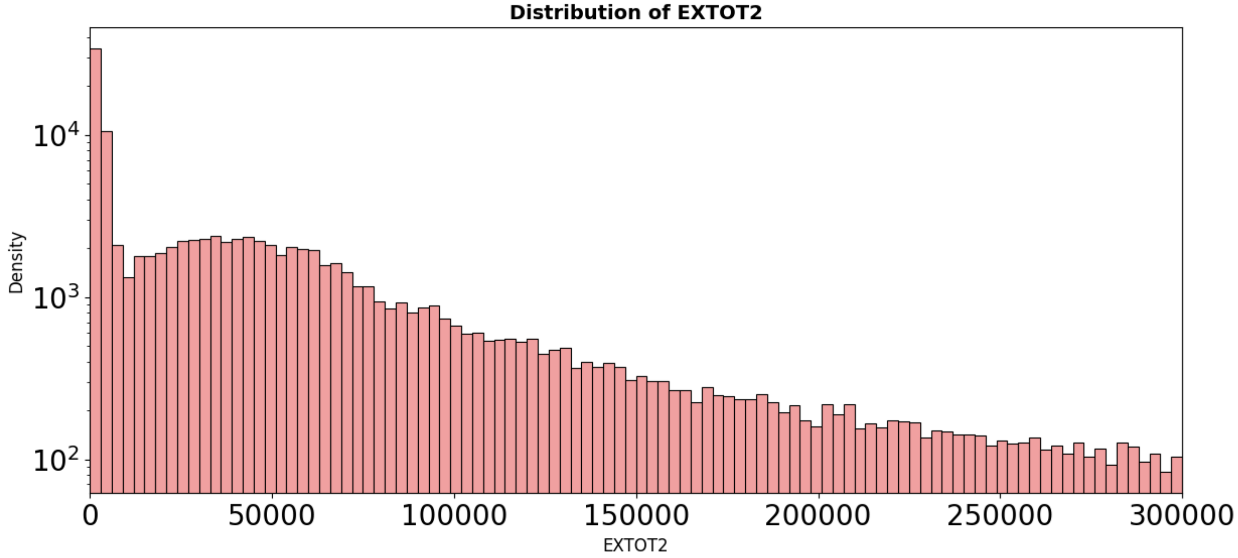
AVTOT2



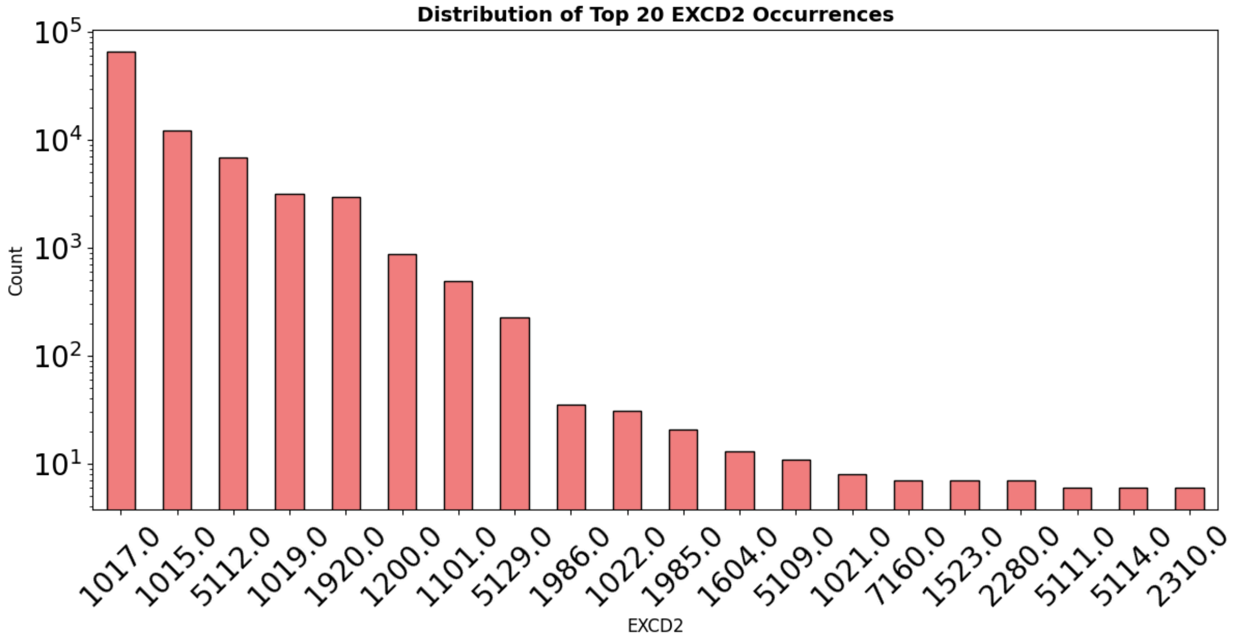
EXLAND2



EXTOT2



EXCD2


Distribution of Top 20 EXCD2 Occurrences
Count
10^5
10^4
10^3
10^2
10^1
EXCD2