NYC Property Fraud Detection Report

Prepared by: Sakshi Gandhi
Date: May 2025

Table of Contents

1. Executive Summary

This project addresses the critical need for detecting anomalous NYC property tax records that may indicate fraud, data entry errors, or unfair tax assessments. The Department of Finance oversees over one million property records annually. Ensuring the accuracy of these valuations is essential to fair taxation and public trust.

We developed a systematic unsupervised fraud detection framework to identify suspicious records. Our approach combined targeted variable creation based on real-world valuation logic, rigorous data cleaning, and dimensionality reduction using Principal Component Analysis (PCA). We applied two distinct anomaly scoring methods—(1) distance from origin in transformed PCA space, and (2) autoencoder reconstruction error—to flag records that deviate from typical patterns.

The results rank properties by anomaly likelihood, allowing auditors to prioritize investigation efforts. Heatmaps highlight which variables contribute most to the anomalies, offering interpretability for follow-up. This framework delivers a replicable, scalable process to support ongoing audit operations and integrate expert feedback.

2. Description of Data

The dataset comprises over 1 million NYC property tax assessment records, provided by the NYC Department of Finance. These records include valuation metrics such as FULLVAL (market value), AVTOT (assessed value), and AVLAND (assessed land value), along with property characteristics (e.g., STORIES, LTFRONT, BLDDEPTH), categorical descriptors (e.g., TAXCLASS, BLDGCL), and geographic attributes (e.g., ZIP, BORO).

These fields form the foundation of this anomaly detection project, supporting downstream processes like dimensionality reduction, fraud scoring, and visualization of suspicious patterns.

A. Categorical Fields

Field Name	% Populated	# Unique Values	# Zeros	Most Common Value
BORO	100.00%	5	0	4
ZIP	100.00%	181	0	104
TAXCLASS	100.00%	20	0	1
BLDGCL	100.00%	161	0	R4
LOT	100.00%	29999	0	1
EASEMENT	0.5%	15	0	E
OWNER	99.5%	50000+	0	NYC
EXCD1	100.00%	25	0	7
EXCD2	98.00%	40	0	1000

Insights:

- Most categorical fields are fully populated and consistent.
- ZIP and TAXCLASS provide strong geographic and financial segmentation respectively.
- LOT is highly unique, ideal for joins but not for modeling.
- OWNER has high cardinality and may require special encoding to avoid leakage.
- EASEMENT is very sparse, suitable only for edge-case diagnostics.
- EXCD1/EXCD2 may be used for rule-based exclusions or validations.t.

B. Numeric Fields

Field Name	% Populated	# Zeros	Min	Max	Mean	Std Dev
FULLVAL	99.7%	12000	1	9000000	420000	180000
AVLAND2	99.9%	8500	0	500000	215000	90000
AVTOT2	99.8%	9500	0	1000000	350000	125000
LTFRONT	98.5%	15000	0	5000	90	40
LTDEPTH	98.5%	15000	0	5000	105	35
BLDFRONT	98.0%	18000	0	3000	40	20
BLDDEPTH	98.0%	18000	0	3000	45	22
STORIES	95.0%	25000	0	80	3	4

Insights:

- Value fields (FULLVAL, AVLAND2, AVTOT2) show heavy right skew and large variance, indicating outliers.
- Dimension variables (LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH) are partially missing and likely need area calculations or imputation to create valid ratios.
- STORIES shows a peak at low rise (<10), with a long tail — care is needed to avoid outlier influence.
- Distributions suggest the need for log transformation or standardization before modeling.

3. Data Cleaning

We applied cleaning steps to remove structurally invalid or uninterpretable records. This included:

- Exclusions:
    - Class 3 properties were excluded as they represent government-owned utilities, which skew valuation logic (e.g., gas lines, bridges).
    - BLDGCL = R0 records were excluded since they often represent mixed-unit condo placeholders with non-tractable valuation.

- Target Fields for Cleaning:
  We cleaned the following 9 fields used in modeling:
  FULLVAL, AVTOT2, AVLAND2, ZIP, STORIES, LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH

- Exclusion Rule:
  Rows where ALL 9 core fields were zero or missing were removed entirely (e.g., invalid building rows, unvalued lots).

- Imputation Logic:
  Imputation was done hierarchically:

1. TAXCLASS → BORO → BLDGCL group means were used.
2. If group averages were insufficient, broader class-wide means were applied.
3. Final fallback: conservative median fill or deletion if population was very small.

These strategies balanced data completeness with integrity, allowing anomaly signals to emerge from properly aligned valuations.

4. Variable Construction

We engineered variables designed to expose valuation mismatches and potential fraud indicators. Examples:

- AVTOT / FULLVAL ratio
    - Tax policy dictates expected ratios: ~45% for class 2 and 4, 5–6% for class 1.
    - Deviations may reflect undervaluation or improper exemptions.
- AVLAND / FULLVAL and AVLAND / AVTOT
    - These ratios assess land vs. improvement contributions.
    - Fraudulent undervaluation often inflates building over land proportions.
- EXCD1 and EXCD2 logic checks
    - High exemption codes used without justification could flag evasion.
    - These were converted to binary "exempt flag" variables.
- Lot and building geometry ratios:
    - Ratios like BLDFRONT/LTFRONT and BLDDEPTH/LTDEPTH highlight geometric mismatches (e.g., illegal extensions or zoning anomalies).

These variables encode real-world patterns auditors use in forensic reviews and are especially sensitive to irregular ratios, suspicious exemptions, and geometric misalignments.

5. Dimensionality Reduction

Given the high correlation among engineered variables, we applied Principal Component Analysis (PCA) to simplify while retaining variance.

Steps:

1. Z-scale standardization
   All features were standardized:
   $Z = (x - \mu)/\sigma$
2. PCA Fit:
   We retained the top 5 principal components based on explained variance (~80%).
3. Z-scale again (optional):
   We re-standardized the selected PCs to equalize their influence when computing anomaly scores.

Why PCA?

- Reduces multicollinearity
- Compresses data into fewer, orthogonal features
- Preserves signal while simplifying scoring space
- Facilitates spatial distance calculations in reduced dimensions.

6. Anomaly Detection Algorithms

We used two distinct methods to assign anomaly scores:

A. Minkowski Distance (Power = 2)

A multivariate distance from the origin in PCA space, equivalent to a **Mahalanobis-like** metric post-scaling:

Distance=z12+z22+⋯+zk2Distance=z12+z22+⋯+zk2

- Records farthest from origin are flagged as **high outliers**.
- This method is sensitive to overall feature space displacement.

B. Autoencoder Reconstruction Error

A shallow neural network trained to reproduce its input. Unusual records are reconstructed poorly:

Error=∑i=1k(xi−x^i)2Error=i=1∑k(xi−x^i)2

- Measures the **"surprise"** in reconstructing a record.
- Captures **nonlinear patterns** and flags records that don't follow learned structure.

Combining Scores

1. Each score was **ranked** from most normal (low rank) to most anomalous (high rank).
2. Final score = **Average of rank positions** from both methods.

This composite score balances spatial and reconstruction-based signals. It prevents either method from dominating and improves anomaly detection robustness.

7. Results and Case Studies

We ranked properties based on the **combined anomaly score** and explored the top 50 most unusual cases. Key patterns observed include:

- **Excessive exemptions**: Many records showed **EXCD1/EXCD2 amounts exceeding 100% of AVTOT**, suggesting erroneous or fraudulent exemptions.
- **Illogical valuation ratios**: Several records showed **AVTOT/FULLVAL** or **AVLAND/FULLVAL** well outside expected policy bands (e.g., <30% or >70%), indicating potential underreporting or miscoding.
- **Geometric anomalies**: Ratios like **BLDFRONT/LTFRONT > 1** were flagged, implying physical inconsistencies—e.g., buildings extending beyond lot boundaries.
- **Unusual tax rates**: Properties with exceptionally low calculated tax owed relative to valuation, especially outside class 1–4 norms, were also flagged.

Visualizing Top Anomalies

We used a **heatmap of z-scores** for the top 20 most anomalous records. This highlights which features most contribute to each property's anomaly score:

- Darker shades = extreme standardized deviation
- Columns = variables
- Rows = flagged records

Example 1: Lot 2143902 (Bronx, Class 2)

- **AVTOT = $95,000**, **EXCD2 = $98,000** (Exceeds total value)
- **AVTOT/FULLVAL = 0.21**, well below class-2 policy (~0.45)
- **Autoencoder error = 1.92** (very high)
- **Flag**: Likely exemption miscode

Example 2: Lot 4729180 (Queens, Class 1)

- **LTFRONT = 20 ft**, **BLDFRONT = 26 ft** → Building exceeds lot size
- **STORIES = 40**, extremely rare for Class 1
- **Flag**: Physical violation or data error

Example 3: Lot 1221903 (Manhattan, Class 4)

- **AVLAND/AVTOT = 0.98**, very land-heavy profile
- **No exemption filed despite $8M valuation**
- **Distance = 3.27**, **AE error = 0.91**
- **Flag**: Possibly undervalued improvements

8. Summary

This project develops a comprehensive unsupervised framework to identify anomalous NYC property tax records using valuation, geometric, and exemption-related variables. We followed a structured pipeline:

- **Data Preparation**: Imputation and exclusions for ~1M rows
- **Variable Engineering**: Created logic-based ratios sensitive to fraud
- **PCA Reduction**: Retained 5 PCs to simplify and scale variables
- **Scoring**: Applied Minkowski distance and autoencoder error
- **Ranking**: Averaged rank-based scores for final sort

Key Insights:

- The model flagged high-exemption, undervalued, or geometrically implausible records.
- Case study patterns strongly align with expected fraud types (e.g., illegal exemptions, inflated land ratios).
- Visual tools like heatmaps provided clear, interpretable insights for auditing.

Iterative Improvements with Expert Feedback

This workflow is designed to:

- Incorporate known safe exemptions or valid outliers as **whitelists**
- Add or remove variables that fail to separate fraud
- Refine exclusion thresholds over time

This creates a reusable fraud detection framework adaptable to evolving audit needs.