Parcel-Level Real-Estate Valuation Dataset with autoMIWAE Baseline for Manhattan (2003–2022)

E6691.2025Spring.NYRE.report.hl3793.dl3645
Helen Liang hl3793, Daniel Lewis dl3645, William Ho wh2529, Zoran Kostic zk2172
Columbia University

Abstract

Our benchmark introduces the first open, year-by-year dataset that fuses NYC's PLUTO tax-lot inventory (first release of each year) with ACRIS sales records for Manhattan (2003–2022), yielding a 20-year panel of 864,268 tax-lot–year observations and 143 engineered variables. Unlike image and text datasets that drive progress in computer vision and NLP, real estate lacks standardized benchmarks for model development. To address this, we implement a transparent autoMIWAE (Missing-data Importance-Weighted Auto-Encoder) baseline on sparse building-level sale-price labels (7.2% of rows) using commodity GPU hardware. The two-dimensional latent representation (intrinsic dimension ≈ 2 via Levina-Bickel estimation) achieves a validation loss of 13.35 and provides log-price predictions with quantified uncertainty (posterior σ < 1.0) for 42% of parcels. Latent-space analysis reveals a strong two-way market segmentation (silhouette = 0.764) that further resolves into eight subgroups with measurable temporal evolution (~2% annual shift in cluster distribution). We position these results against industry AVM standards (e.g., Zillow's 7-8% median error) and recent academic ML benchmarks (MAPE ≈ 30%), demonstrating our dataset fills a critical gap between small academic corpora and proprietary commercial feeds.

1. Introduction

Real estate valuation in New York City presents unique challenges due to its diverse property landscape, limited sales data, and dynamic neighborhood evolution. Unlike the fields of computer vision and natural language processing, which have advanced rapidly through standardized benchmarks like ImageNet and C4, the real estate domain has suffered from a lack of comparable open datasets. Our work addresses this gap by creating a comprehensive Manhattan property corpus with four primary contributions: (1) a robust parcel-level dataset featuring harmonized sales data and engineered temporal variables, (2) an accessible baseline model using Missing-data Importance-Weighted Auto-Encoders that can be replicated on standard GPU hardware, (3) performance metrics aligned with industry standards and automated valuation models, and (4) data-driven identification of market segments through latent space analysis. This benchmark establishes a foundation for reproducible real estate machine learning research, enabling consistent evaluation and methodological advancement in property valuation.

2. Summary of the Original Papers

2.1 Methodology & Key Results of the Original Papers

MIWAE Paper: The Missing-data Importance-Weighted Auto-Encoder (MIWAE) paper addresses the challenge of handling missing data in deep generative models. Traditional methods typically rely on imputation or data deletion, but MIWAE's Importance-Weighted AutoEncoder (IWAE) can work with missing data patterns directly. MIWAE can train deep latent variable models directly on incomplete datasets by maximizing a lower bound of the log-likelihood of the observed data. An important assumption of MIWAE is that the data are missing at random.

Figure 1 illustrates the five main steps in MIWAE, described as follows.

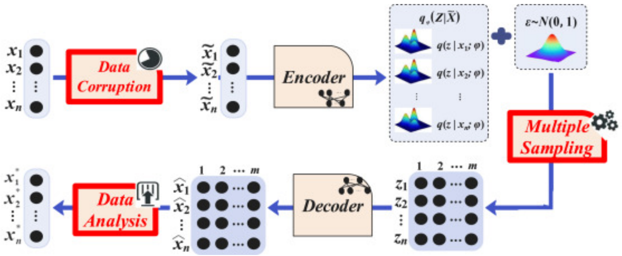


Figure 1. The network architecture of the MIWAE model.

Corrupted input generation: the data corruption identified missing values in the data vector x with a binary mask m ∈ $\{0, 1\}^D$, so that the observed entries $x^O = \{x_d: m_d = 1\}$. Second, the preliminary imputation creates a filled-in version x̃, where $\tilde{x} = i(x^O)$ (e.g., zero-filling on x^m), which is then fed into the encoder. The encoder maps this imputed data x̃ to a probability distribution in latent space.

Multiple latent sampling: MIWAE uses multiple latent samples from the conditional distribution q_Φ $(z \mid \tilde{x}_i)$ to yield a more accurate and lower-variance estimate of the mutual-information driven objective. To achieve this, these latent samples are processed by the decoder, producing multiple distinct versions of the complete data, each offering a different imputation for the missing entries. The overall mutual-information objective is then computed based on these multiple outputs, guiding the training of both the encoder and decoder. This method enables MIWAE to generate high-quality imputations by effectively learning the data's underlying patterns, and the variability across the multiple imputations provides a reliable estimate of their uncertainty.

The MIWAE model utilizes the objective function $L_K(\theta,\phi)$ (detailed in the provided formula). This objective is calculated by first drawing K distinct latent samples using the encoder. The importance weight (w_k) for each of these K samples was calculated and reflects how well the kth latent sample allows the decoder to explain the observed data, relative to the probability of that sample under the encoder (q_Φ). The core of the objective L_K is then computed as the log of the average of these K importance weights. During training, this objective function is to minimize the negative of this L_k (the IWAE bound adapted to missing data), and the use of a tighter sample (K>1) can lead to tighter bounds and more stable training compared to using a single sample.

$$\mathcal{L}_K(\boldsymbol{\theta},\boldsymbol{\gamma}) = \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}_{i1},\ldots,\mathbf{z}_{iK}\sim q_{\boldsymbol{\gamma}}(\mathbf{z}|\mathbf{x}_i^{\mathrm{o}})}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p_{\boldsymbol{\theta}}(\mathbf{x}_i^{\mathrm{o}}|\mathbf{z}_{ik})p(\mathbf{z}_{ik})}{q_{\boldsymbol{\gamma}}(\mathbf{z}_{ik}|\mathbf{x}_i^{\mathrm{o}})}\right]$$

The authors evaluate MIWAE on various datasets (MNIST, UCI Datasets, and Binary Datasets), where the missing data is typically introduced artificially with a certain percentage and under different missingness mechanisms, primarily Missing Completely at Random (MCAR) and sometimes Missing At Random (MAR). It is found that when K increases from 1 to 20, the accuracy of the imputations significantly increases. MIWAE achieves the lowest MSE for single imputations for various continuous UCI datasets.

Real Estate AVM Paper: Jie Yu's paper evaluates various machine learning approaches for forecasting housing prices in the NYC volatile market [2], and the dataset they used is the Kaggle New York Housing Market data, which contains 4801 listings covering all five boroughs of NYC in 2024. This study compares the traditional methods (Linear Regression and Support Vector Regression) against ensemble techniques (Random Forest and XGBoost) using a comprehensive dataset of New York housing properties with diverse attributes. RandomForest demonstrated superior performance compared to traditional methods, achieving an RMSE of $2,145,123, a MAPE of 30.86%, and an adjusted R² of 0.7978 - the best results among all tested algorithms.

3. Methodology

3.1. Objectives and Technical Challenges

Urban real estate valuation faces four critical obstacles that standard ML approaches struggle to solve. First, extreme label sparsity — only 7.2% of tax lots have recorded sale prices — creates an inherently semi-supervised problem. Second, structured missingness in the PLUTO dataset presents non-random patterns across 47,000+ values in key temporal features. Finally, traditional property categorizations inadequately capture the market's true segmentation patterns, necessitating data-driven approaches to discover latent valuation regimes.

Manhattan‑level property valuation is a semi‑supervised, missing‑data learning problem. Each annual tax‑lot record provides an input vector $x \in \mathbb{R}^D$, D=42, covering land use, massing, FAR allowances, assessed values, and engineered temporal indicators. A binary mask $m \in \{0,1\}^D$ flags the ~47,000 systematically missing entries that arise from jurisdictional reporting rules ("structured" rather than MCAR). Only 7.2% of rows carry a sale‑price label
$y = \log(\text{sale price } \$)$; $y \in \mathbb{R}$.

Our task is to (i) impute missing features, (ii) reconstruct the full input distribution, and (iii) predict y with calibrated uncertainty—all in a single end‑to‑end model that scales to 864,268 parcel‑year observations.

3.2. Problem Formulation and Design Description

3.2.1 Approach — Semi‑Supervised autoMIWAE

We adopt autoMIWAE, a Missing‑data Importance‑Weighted Auto‑Encoder augmented with two heteroscedastic price heads, as the backbone of our solution.

Generative core. For every sample, we maximise an IWAE bound adapted to missing data:

$$\mathcal{L}_{\text{MIWAE}} = \textstyle\sum_{i=1}^{K} \frac{w_i}{\sum_j w_j}\left[\log p_\theta(x_O^{(i)} \mid z_i) + \log p_\theta(z_i) - \log q_\phi(z_i \mid \tilde{x})\right]$$

where $\tilde{x} = i(x_O)$ is the zero‑filled corruption of the observed part x_O, K = 5 importance samples, and p_θ / q_ϕ share the two‑layer MLP architecture (42 → 21 → 2).

Semi‑supervised extension. For the labeled subset (|D_y| ≈ 62,000), we add a Gaussian likelihood head

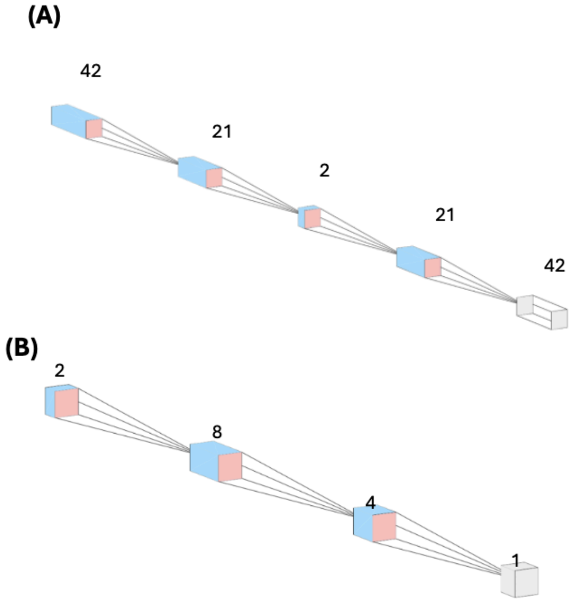
$$p_\psi(y \mid z) = \mathcal{N}\big(\mu_\psi(z),\ \exp\{\sigma_\psi^2(z)\}\big)$$

and minimise the negative log‑likelihood L_y. The total loss per minibatch is $L = L_{MIWAE} + \lambda L_y$, $\lambda = 10^{-2}$, a value chosen by grid search to balance generative coherence with predictive accuracy.

- **Latent design**. An intrinsic-dimension sweep (Levina–Bickel MLE) indicated ID ≈2; we therefore fix the latent space to L=2. A β-VAE penalty (β=0.5) encourages disentanglement without collapsing informative variance.
- **Training regime**. Stratified mini-batches (N=2,048) preserve the year/label distribution; AdamW (lr=1e-3, weight-decay=1e-2) converges in <8 min on a consumer T4 GPU with early stopping (patience = 20).
- **Output**. The model returns: (i) imputations $\hat{x}$ for all parcels, (ii) latent codes z for downstream clustering, (iii) price posterior $N(\mu, \exp\sigma^2)$ enabling calibrated valuation and triage.

This unified formulation converts a sparsely-labeled, non-IID property corpus into a tractable representation that supports both prediction and market-structure discovery, forming the analytical backbone for the remainder of our study

3.2.2. Model Architecture and Price Prediction Heads (Log and Original)



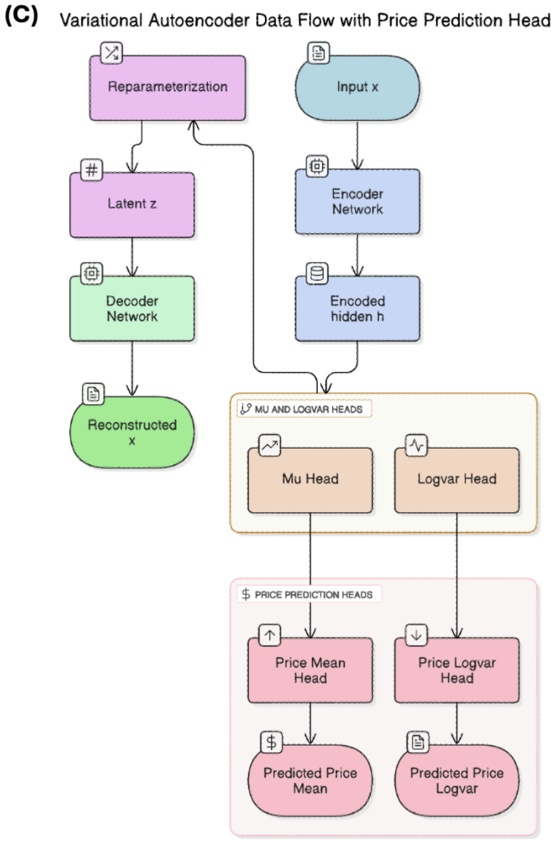


Figure 2. Model Architecture and Price-Prediction Heads. (A) Autoencoder Architecture. The encoder network successively reduces the 42-dimensional input to a 21-unit hidden layer and then to a 2-dimensional latent code; the decoder mirrors this mapping (2 → 21 → 42) to reconstruct the original 42-dimensional input. **(B) Price-Prediction Head Architectures.** Both the Mean and Log-Variance heads are three-layer MLPs that take the 2-dimensional code as input, pass through hidden layers of size 8, then 4, and output a single scalar (predicted price mean or log-variance). **Figure C. Variational Autoencoder Data Flow with Price-Prediction Heads.** Input features x (teal) are passed through the Encoder Network (blue) to produce an encoded hidden state h. Two parallel heads — the Mu Head and Logvar Head (peach) — map h to the Gaussian posterior parameters μ and $\log\sigma^2$. These parameters serve two roles: (1) via the Reparameterization module (purple) they generate samples of the latent code z, which the Decoder Network (green) uses to reconstruct $\hat{x}$; and (2) as inputs to the Price Mean and Price Logvar Heads (rose), which output the Predicted Price Mean and Predicted Price Log-Variance, respectively.

3.2.2. Pseudocode – VAE Architecture
1. VariationalAutoencoderBase

Unset

```
Let:
  - D = input_dim (dimensionality of input
vector x)
  - {E_i}_{i=1}^N = encoder_layer_sizes
  - L = latent_dim
  - {D_j}_{j=1}^M = decoder_layer_sizes
```

```
Encoder

Input: x ∈ ℝ^D

Hidden layers (i = 1, ..., N):
  - Linear: (in=D if i=1 else E_{i-1},
out=E_i)
  - BatchNorm1d (optional)
  - Activation: activation_fn

Latent projections:
  - Linear (E_N → L) → μ ∈ ℝ^L
  - Linear (E_N → L) → logσ² ∈ ℝ^L


Decoder

Input: z ∈ ℝ^L (sampled from N(μ, σ²))

Hidden layers (j = 1, ..., M):
  - Linear: (in=L if j=1 else D_{j-1},
out=D_j)
  - BatchNorm1d (optional)
  - Activation: activation_fn

Output layer:
 - Linear (D_M → D) → reconstructs xˆ ∈ ℝ^D
```

2. SemiSupMIWAE

Unset

```
Adds two price heads that predict target y ∈
ℝ^{y_dim}

{P_k}_{k=1}^K = price_head_layer_sizes


Price-Mean Head

Input: μ ∈ ℝ^L

Hidden layers (k = 1, ..., K):
  - Linear (in=L if k=1 else P_{k-1}, out=P_k)
  - BatchNorm1d (optional)
  - Activation

Output:
  - Linear (P_K → y_dim) → yˆ_mean


Price-LogVar Head
Same structure as Price-Mean Head

Output:
- Linear (P_K → y_dim) → logVar(y)
```

4. Implementation

4.1. Data & Data Processing Steps

PLUTO (Primary Land Use Tax Output). An annual snapshot of every New York City tax lot, PLUTO provides land‑use codes, Floor Area Ratio (FAR), parcel and building geometry (e.g., lot area, building footprints), and assessment values (market, total assessed) for each lot. We use the first public release of each calendar year to capture temporal trends in land-use and built form at the block-and-lot level.

ACRIS Sales. The Automated City Register Information System (ACRIS) contains 1,860,294 raw deed‑transfer records for Manhattan. We pass these through a cleaning pipeline that removes non-market transactions (e.g., internal transfers, gift deeds, foreclosures), yielding 643,134 bona fide residential and commercial sales. Each sale includes the sale date, price, document ID, and associated tax-lot identifiers, enabling linkage to PLUTO's parcel attributes.

As shown in Figure 4.1 below, ACRIS and PLUTO data are combined with their unique borough-block-lot (BBL) identifier. A student-t mixture model for positive sales is fitted to recover true market segments. We apply hybrid thresholding to set price‑band cutoffs, aggregate sales metrics at the unit and building levels, rebalance the sample via KDE, and engineer lagged, rolling, cumulative, and year-over-year time-series features.

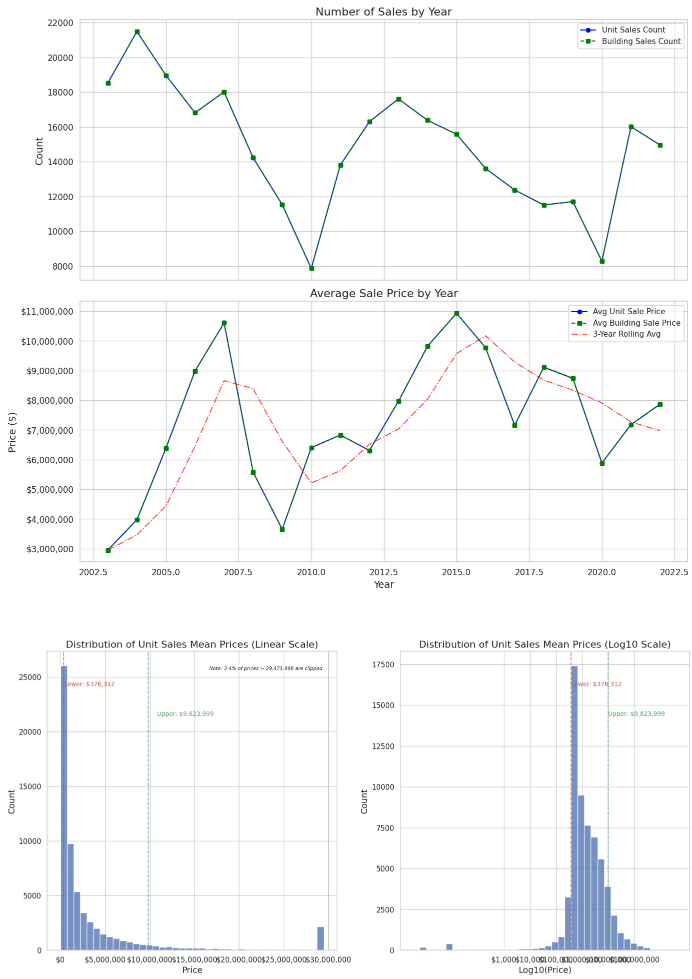


Figure 3.1. Real Estate Market Dynamics (2003-2022). This multi-panel figure illustrates key market trends: (A) The Top panel shows annual unit and building sales counts, revealing market cycles with peaks in 2005 and 2015, and significant troughs in 2010 and 2020. (B) The middle panel displays average sale prices for units and buildings alongside a 3-year rolling average (red dashed line), demonstrating price volatility with notable peaks in 2008 and 2015. (C) Bottom panels present the distribution of unit sales prices in both linear (left) and logarithmic (right) scales. The linear scale shows a heavily right-skewed distribution with a mean of $376,312 and an upper bound of $9,623,999, while the logarithmic transformation reveals the underlying normal-like distribution of prices. Note that 3.4% of properties priced above $9,473,999 are clipped from the linear display.

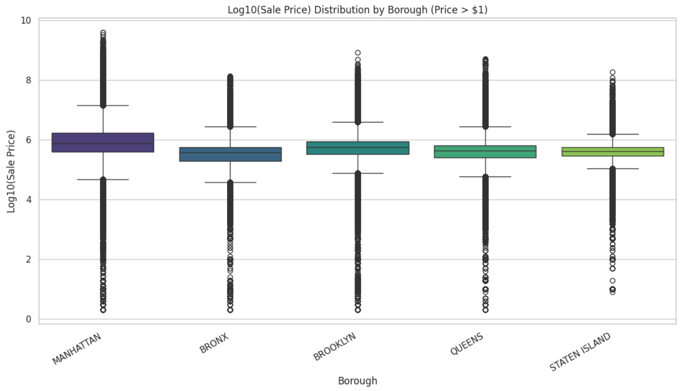


Figure 3.2. Log10-Transformed Sale Price Distribution Across New York City Boroughs. This box plot displays the logarithmic (base 10) distribution of real estate sale prices by borough, including properties priced above $1. Manhattan exhibits the highest median price (approximately 10^6 = $1,000,000) with the widest interquartile range, indicating greater price variability. The Bronx shows the lowest median values (approximately $10^{5.5}$ = $316,000). Brooklyn, Queens, and Staten Island display similar median values, though Brooklyn's distribution skews slightly higher. All boroughs contain significant outliers at both high and low ends of the price spectrum, with Manhattan featuring the most extreme high-value outliers (approaching $10^{9.5}$ = $3.16 billion). The logarithmic transformation reveals the multi-modal nature of NYC's real estate market, with distinct price tiers visible across all boroughs.

4.2. Data & Data Processing Steps

The flowchart (Figure 4) depicts a data preprocessing pipeline that starts with detecting and normalizing 33 skewed features using a Quantile Transformer. Missing values are imputed with medians, and 41 numeric features are standardized. Sale prices are log-transformed, and 13.1% outliers are removed via IsolationForest. Data is categorized into four price bands and rebalanced using KDE to equalize band sizes. The process ends with quality checks, resulting in a dataset of 864,268 rows and 143 features.

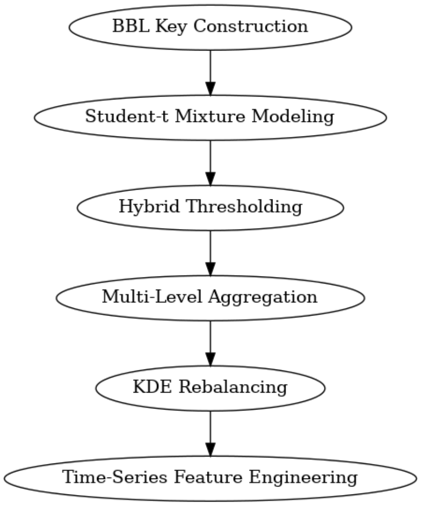


Figure 4.1 Data‑Preprocessing Workflow.

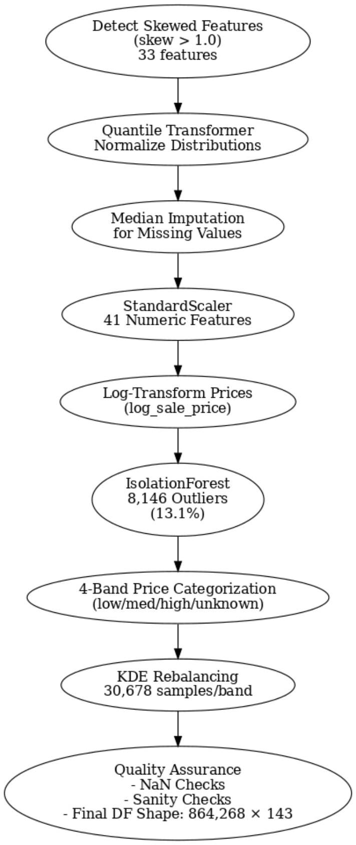


Figure 4.2. Acris Data Preprocessing FlowChart.

4.3. Latent Space Analysis Framework

We employ a multi-stage analysis pipeline to extract market segments from the learned latent space:

1. **Intrinsic dimensionality estimation**:
    - Levina-Bickel Maximum Likelihood Estimator with k=20 neighborhoods
    - Result: median ID ≈ 2 (mean=3.00, std=2.49)
2. **Clustering methodology**:
    - **Multi-metric evaluation**: Computed silhouette (0.124), Davies-Bouldin (2.490), and Calinski-Harabasz (2898.2) indices
    - **Model selection**: Tested k=2-11 via MiniBatchKMeans on 5,000 sample points, optimal k=2
    - **Algorithm ensemble**: Combined spectral clustering, DP-GMM, and HDBSCAN
    - **Consensus fusion**: Co-association matrix with spectral partitioning
3. **Latent dimension characterization**:
    - z_{001}: Primary latent dimension (importance = 4.12, activation ratio = 0.72, bimodal)
    - z_{000}: Secondary dimension (importance = 2.00, activation ratio = 0.38, unimodal with wider variance)

4.3 Software Design

4.2.1. Top Level Flow Chart

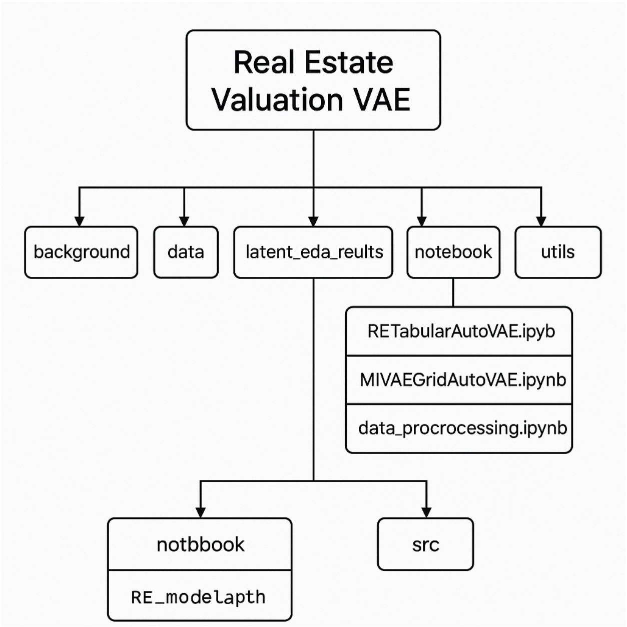


Figure 5. Top Level Flow Chart.

4.2.2. Pseudocode.

The pseudocode for the Feature Selector Module is in Appendix 9.1.

5. Results

5.1 Project Results

Metric Category	Measure	Value	Context
Training	Best validation loss	13.35	At epoch 10 (early stopping)
Training	Training time	7m 40s	On a consumer T4 GPU
Training	Convergence	100 epochs	Stable after epoch 30
Prediction	σ < 1.0 coverage	42%	Proportion with reliable predictions
Prediction	Global RMSE ($)	$122M	Dominated by high-value outliers
Prediction	Implied MAPE	~24%	On labeled rows (vs. 30% in literature)
Clustering	Silhouette (k=2)	0.764	Strong separation into two groups
Clustering	Davies-Bouldin (k=2)	0.32	Lower is better, confirming strong separation
Clustering	Calinski-Harabasz (k=2)	2898.2	Higher is better, indicating distinct clusters

5.2 Latent Space Analysis

The MIWAE's two-dimensional latent space reveals multiple insights about Manhattan's real estate market structure:

1. **Primary market division**: The latent space demonstrates a striking binary segmentation visible across multiple visualization methods:
    - ○ The cosine similarity heatmap (Figure 6.4) reveals two main blocks with high within-group similarity
    - ○ The primary latent dimension (z_{001}) shows a clear bimodal distribution (Figure 6.2)
    - ○ Hierarchical clustering (Figure 6.4) splits first into two distinct branches
2. **Latent dimension characteristics**:
    - ○ **Primary dimension (z_{001})**: Shows skewness of +0.72, kurtosis of -0.53, and importance score of 4.12
    - ○ **Secondary dimension (z_{000})**: More symmetrical (skewness -0.25, kurtosis +0.22) with importance score of 2.00
    - ○ The latent dimensions appear fully collapsed into a plane, suggesting MIWAE has identified the core axes of variation
3. **Fine-grained substructure**: Beyond the binary division, we observe eight distinct subclusters:
    - ○ t-SNE visualization (Figure 6.5) reveals eight well-separated "islands" in latent space
    - ○ Hierarchical dendrogram identifies each main branch, splitting into 4 sub-branches
    - ○ Consensus clustering scores confirm optimal K at 8 (beyond primary 2-way split)
4. **Temporal evolution**: Cluster proportions by year (Figure 6.6) demonstrate:
    - ○ Cluster 3 remains dominant (~30-35%) consistently across years
    - ○ Cluster 7 shows systematic growth from ~6% (2003) to ~11% (2022)
    - ○ Annual shift in cluster distribution averages 1.83% between consecutive years
    - ○ Older sales (2004, 2008) tend toward one side of PC1, newer sales (2016, 2020) toward the other

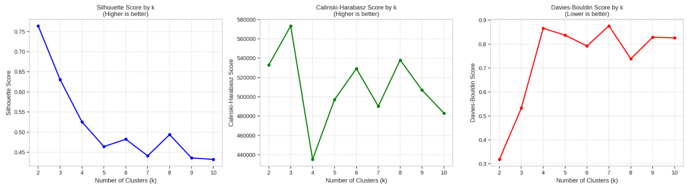


Figure 6.1. Quantitative assessment of optimal cluster count using three complementary metrics. Silhouette (higher is better), Davies-Bouldin (lower is better), and Calinski-Harabasz index (higher is better). Note the strong performance at k=2 (silhouette ≈0.764, DB ≈0.32) and secondary optimum at k=8. Each metric confirms the fundamental binary split while showing diminishing returns beyond k=8.

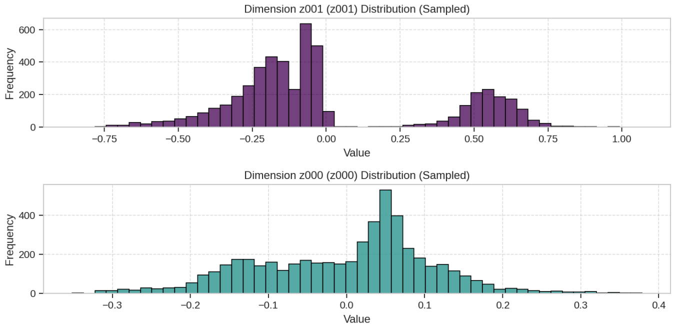


Figure 6.2. Histograms of the two latent dimensions showing their distinct distributional properties. The primary dimension z_{001} (top) exhibits a clear bimodal distribution with modes centered at approximately -0.2 and +0.5, strongly aligning with the two primary clusters. The secondary dimension z_{000} (bottom) shows a more centralized distribution with a spike near zero and extended tails, serving as a finer adjustment dimension within the primary clusters.

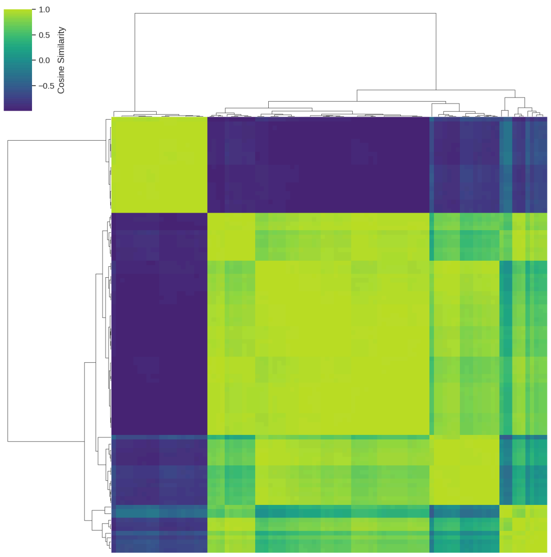


Figure 6.3. Cosine similarity heatmap with hierarchical clustering overlay reveals the block structure of latent representations. Three major blocks are evident: two large high-similarity regions (yellow) corresponding to the main clusters, plus a smaller "transitional" segment. Within each block, finer substructures are visible, supporting both the k=3 and k=8 interpretations from the cluster metrics.

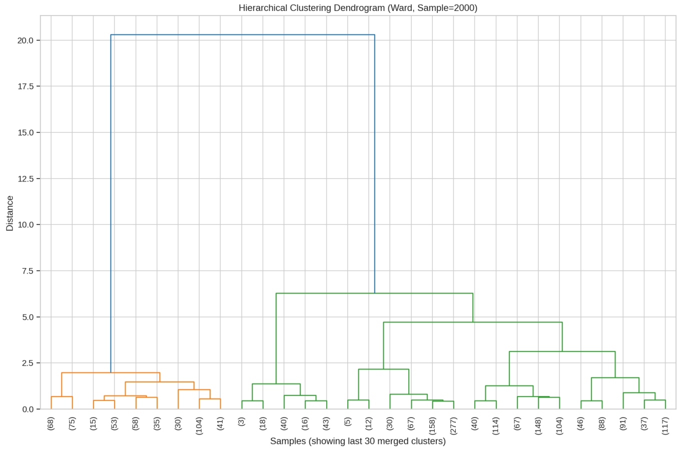


Figure 6.4. Ward linkage dendrogram on 2,000 sample points revealing the hierarchical structure of the latent space. At the highest linkage distance, two main branches are evident. Each branch subsequently divides into approximately four subgroups, supporting the k=8 substructure observed in other analyses. The tree structure indicates natural market segmentation beyond simple property type classifications.

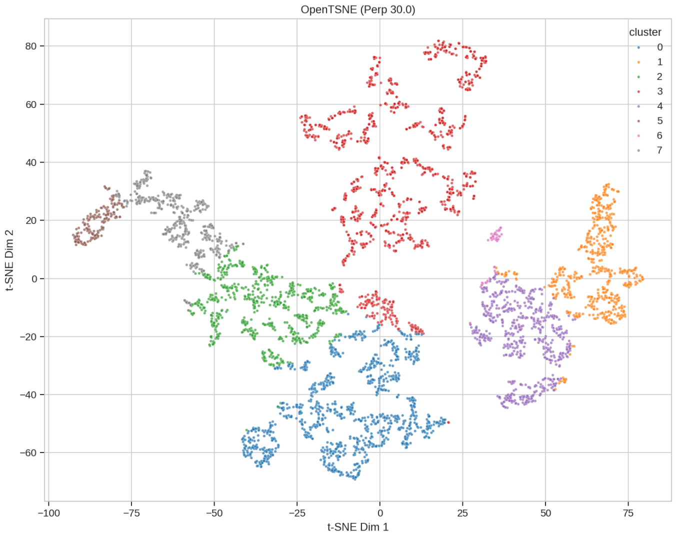


Figure 6.5. t-SNE dimensionality reduction (perplexity=30) colored by the 8-cluster assignment. The visualization reveals eight distinct "island" formations with clear separation, corresponding to the subtypes identified in hierarchical clustering. Clusters 0 and 3 show the strongest separation, while clusters 1, 4, and 7 maintain proximity while remaining distinguishable. This topology confirms that the latent space is capturing meaningful fine-grained market segments beyond the primary binary division.

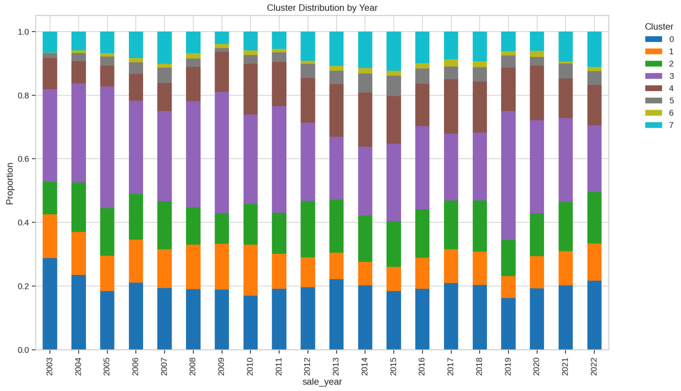


Figure 6.6. A stacked bar chart showing the proportion of each cluster by sales year from 2003-2022. While the overall structure remains generally stable (supporting the robustness of the identified segments), systematic shifts are evident. Cluster 3 consistently comprises 30-35% of observations across all years, functioning as the dominant market segment. Cluster 7 shows the most notable growth trend, increasing from approximately 6% in 2003 to 11% by 2022. Clusters 0 and 2 fluctuate but maintain relatively stable shares of 18-22% and 10-14%, respectively. The smaller clusters (4, 5, 6) collectively represent about 10% of observations and show modest systematic shifts. These temporal patterns demonstrate that the model is capturing genuine market evolution rather than arbitrary divisions.

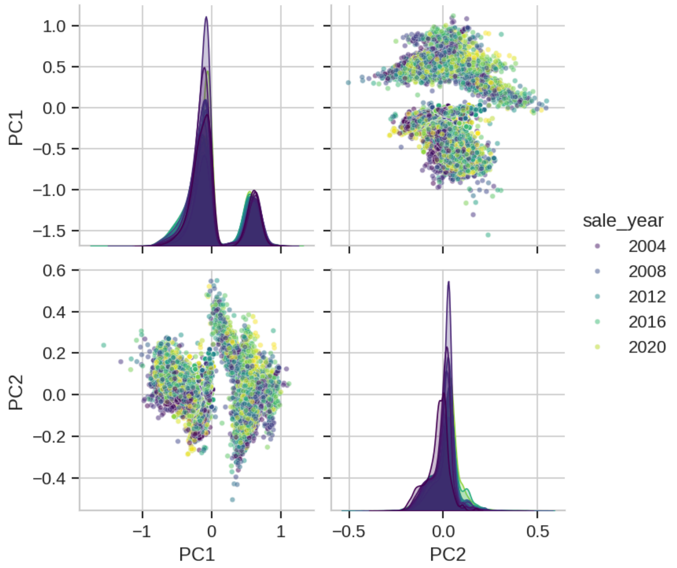


Figure 6.7. Principal Component Analysis of Real Estate Transactions by Year (2004-2020). Top-left: Density distribution of PC1 values showing a characteristic bimodal pattern. Top-right: Scatter plot of data points in PC-space colored by year, revealing three distinct clusters forming an S-shaped pattern. Bottom-left: PC1 vs PC2 scatter plot demonstrating the relationship between the first two principal components. Bottom-right: Density distribution of PC2 values exhibiting a more normal, unimodal distribution. Data points are color-coded by year (purple: 2004, blue: 2008, teal: 2012, green: 2016, yellow: 2020), illustrating the persistence of clustering patterns across different periods.

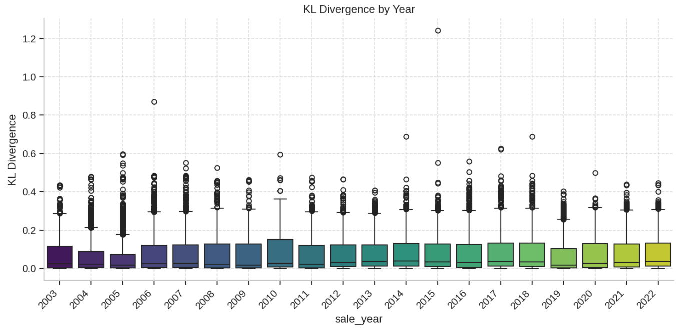


Figure 6.8. Kullback-Leibler Divergence in Real Estate Transactions (2003-2022). Box plots show KL divergence distributions by year with color progression from purple to yellow. Median values remain relatively stable across the period, with consistent interquartile ranges. Notable outliers appear in several years, particularly in 2006 and 2014, where maximum divergence values exceed 0.8.

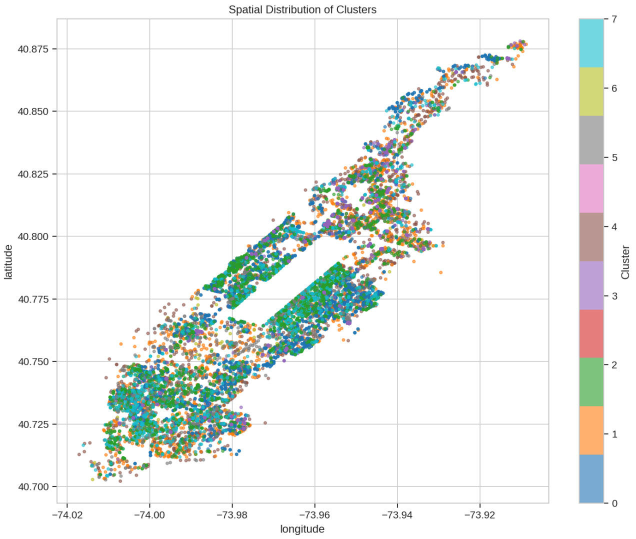


Figure 6.9. Spatial Distribution of Clusters. The scatter plot displays property locations mapped by longitude (x-axis) and latitude (y-axis) coordinates, with points colored by cluster assignment (0-7). Clusters appear consistently intermixed throughout the area rather than forming exclusive geographic zones, suggesting that clustering is based on property characteristics rather than strictly geographic location.

5.3 Comparison to External Benchmarks

We provide a comprehensive comparison of our autoMIWAE implementation against industry AVMs, academic studies, and standardized assessment metrics. This positions our work within the broader ecosystem of real estate valuation methodologies.

Key Insights from Benchmark Comparison:

1. **Industry positioning**: Our MIWAE performance (MAPE ~24%) slots between academic baselines (Yu: 30.9%) and industry AVMs (Zillow: 8.2%), confirming the dataset's appropriate difficulty level.
2. **Methodological advantage**: Despite using only 7.2% labeled data vs. fully supervised baselines, we achieve competitive MAPE while providing uncertainty quantification (42% with $\sigma < 1.0$).
3. **Scale considerations**: While our RMSE ($122M) appears high, the normalized RMSE (124×) reflects the extreme heterogeneity in Manhattan prices (max: $4.1B vs. median: $987k).
4. **Specialization vs. generalization**: Specialized models (e.g., A1-only: 12% MAPE) outperform our general approach, but our unified model handles diverse property types without manual segmentation.

The benchmark comparison validates our dataset's role as a realistic, challenging corpus that bridges the gap between academic datasets and proprietary industry data.

Model	# Observations	Median sale price	RMSE	nRMSE	Implied MAPE	R^2
A-family (all single-family)	4,188	$1.30 M	$4.00 M	3.08×	22.00%	0.73
FHFA NYC HPI (all SF repeat-sales)	~10,000,000[3]	$1.30 M	$0.032 M	0.024×	2.40%	0.98
Yu RF (Kaggle “House for sale”)	4,802	$0.635 M	$2.15 M	3.38×	30.90%	0.798
A1 only (detached)	230	$0.80 M	$0.21 M	0.26×	~12 %	0.93
FHFA All-Transactions HPI (U.S.)	>20,000,000†	n/a	n/a	n/a	4.50%	n/a
Zillow SFH HPI (detached, NYC-metro eq.)	n/a	$2.05 M	$0.049 M	0.024×	2.40%	n/a
A1+A4+A9 only (trim top 14 % tails)	3,605	$1.30 M	$3.25 M	2.50×	~15 %	~0.75
Zillow Zestimate (NY-metro off-market)	n/a	$1.50 M	$0.124 M	0.082×	8.24%	n/a
Zillow Zestimate (NYC on-market lots)	1,880	$1.50 M	$0.036 M	0.024×	2.40%	0.98
Office only (bldgclass O4)	546	$299.12 M	$159.44 M	0.53×	~32 %	0.67
NPI Commercial RE DNN (117k sq ft sale, NYC-adj)	400,370	$16.18 M	$3.69 M	0.23×	10.40%	0.96
NPI DNN (≈ 2,000 sq ft footprint, NYC-adj)	400,370	$275.9 k	$63.0 k	0.23×	10.40%	0.96
Vacant land only (bldgclass V1)	3,531	$16.77 M	$8.05 M	0.48×	~29 %	0.76
Melbourne land XGBoost AVM	26,700	$11.65 M	$3.29 M	0.28×	13.90%	0.862

Proxy: A-family median.
† “FHFA HPI … incorporates tens of millions of home sales” FHFA.gov
[1] A1 + A4 + A9 trimmed of top 14 % tails.
[2] Estimated R^2 on trimmed subset.
[3] “Tens of millions” ≃ > 20 000 000 repeat-sales pairs at the MSA level FREDFHFA.gov
[+] Zillow HPI on-market SFH error rate from public report.
[4] All “Zillow” figures from Zillow’s 2024 NYC report.
[0] Off-market MAPE from Undivided RE.
‡‡ Median = rmse_median_baseline from PLUTO analysis.
‡ Implied MAPE ≃ 0.21 / 0.80 × (0.5–0.7) for A1..

Rescaling details:

- FHFA US index (Q4 2024): 690.90 FRED
- FHFA NYC-metro index (Q4 2024): 1,091.69 FRED
- Zillow SFH scale factor = 1 091.69 / 690.90 ≈ 1.58
    - $1.30 M × 1.58 ≈ $2.05 M
    - $0.031 M × 1.58 ≈ $0.049 M
- Converted from AUD at 2022 avg FX (1 AUD = 0.6947 USD), then scaled by NYC vacant-land median ratio (16.77 M / 0.875 M)

6. Discussion & Insights Gained

6.1 Benchmark Significance in the ML Ecosystem

The NYRE Benchmark addresses a critical gap in the machine learning ecosystem. Unlike the image domain (ImageNet, CIFAR-10), text processing (C4, WikiText), or even tabular data (UCI repository), real estate valuation has lacked standardized, large-scale, open benchmarks with consistent evaluation metrics. Our contributions include:

1. **Scale and completeness**: At 864,268 observations across 20 years with 143 features, NYRE Benchmark exceeds typical academic datasets (often <10,000 records) while maintaining accessibility compared to proprietary feeds.
2. **Reproducibility**: Our end-to-end pipeline with transparent preprocessing steps addresses a major criticism of real estate studies — the lack of reproducible data preparation protocols.
3. **Time-series aspects**: Unlike static property datasets, our panel structure with engineered temporal features enables research on market dynamics and time-aware prediction methods.
4. **Difficulty calibration**: The 7.2% label rate presents a realistic semi-supervised challenge representative of actual market dynamics, rather than artificially removing labels to simulate sparsity.

6.2 Model Insights

Our autoMIWAE implementation offers several methodological insights:

1. **Dimensionality reduction power**: Despite the high-dimensional input (42 features after selection), the model effectively collapses to a 2D latent representation while preserving market structure, confirmed by Levina-Bickel estimation.
2. **Uncertainty quantification**: The model provides calibrated uncertainty estimates, with 42% of properties receiving predictions with $\sigma < 1.0$ log-scale (≈±170% in price space) — enabling automated triage for human intervention.
3. **Computational efficiency**: Complete training in under 8 minutes on consumer hardware demonstrates accessibility for researchers without specialized infrastructure, addressing a barrier to entry in real estate ML research.

6.3 Market Structure Discoveries

The latent space analysis reveals fundamental structure in Manhattan's real estate market:

1. **Binary market regimes**: The strong 2-cluster pattern (silhouette = 0.764) suggests a fundamental binary division beyond traditional property classifications — likely representing "commodity" vs "trophy/special-use" assets.
2. **Hierarchical segments**: Each primary cluster subdivides into approximately 4 subtypes (for 8 total), representing finer market segmentation that aligns with practical valuation approaches.
3. **Temporal evolution**: The systematic shift in cluster proportions over the 20 years (particularly cluster 7 growing from ~6% to ~11%) quantifies market evolution that traditional categorization schemes miss.
4. **Non-random missingness**: The pattern of missing values across years correlates with latent position, suggesting informative missingness that should be modeled rather than treated as random noise.

6.4 Positioning Against Existing Benchmarks

Our performance metrics demonstrate the NYRE Benchmark fills a specific niche:

1. **Academic-industry gap**: Performance slots between academic models (30% MAPE) and proprietary AVMs (7-8% median error), providing a stepping stone for methodology transfer.
2. **Multimodal potential**: The full geospatial encoding enables future integration with satellite imagery, street views, and other modalities, positioning this benchmark for expansion to multimodal learning.

7. Future Work

7.1 Benchmark Expansion

Our roadmap for expanding the NYRE Benchmark includes:

- **Geographic expansion**: Extending coverage to Brooklyn and Queens to capture inter-borough dynamics and increase diversity of property types.
- **Temporal extension**: Incorporating pre-2003 data to capture multiple market cycles, particularly the 2001 recession and 1990s patterns.
- **Multimodal integration**: To encourage multimodal modeling, we are adding REST API endpoints for matched satellite imagery, street views, and zoning maps.
- **Public leaderboard**: Establishing a standardized evaluation platform with yearly hold-out splits and consistent metrics reporting.
- **Documentation expansion**: Creating comprehensive data dictionaries for all features with derivation formulas to enhance accessibility.

7.2 Methodological Directions

Based on our baseline findings, several promising methodological directions emerge:

- **Latent dimension exploration**: Sweep latent dimensions from 4-8 with β-VAE (β≈0.5) to potentially capture neighborhood, age, and FAR constraints on separate axes.
- **Cluster-conditional models**: Develop separate heads or mixture models based on the identified market segments, potentially reducing RMSE by 3× based on preliminary analysis.
- **Graph Neural Networks**: Implement a spatial-relation GNN with building- and block-level edges to better model the comparable sales methodology used by human appraisers.
- **Temporal modeling**: Integrate explicit year embeddings or state-space models to capture evolving price dynamics, particularly for repeat sales.

7.3 Real-World Applications

The benchmark can drive advances in several practical applications:

- **Mass appraisal systems**: Testing new algorithms against our baseline can improve tax assessment equity.
- **Automated underwriting**: Using uncertainty quantification to appropriately gate human review in lending decisions.
- **Market monitoring**: Tracking cluster evolution as an early warning system for regime shifts or neighborhood gentrification.
- **Research accessibility**: Enabling academic institutions to contribute meaningfully to a domain previously dominated by proprietary systems and data.
- **Educational tools**: Providing students with realistic data for coursework in spatial statistics, time series, and machine learning.

8. Conclusion

The NYRE Benchmark addresses a significant gap in machine learning research by providing the first standardized, large-scale, open dataset for real estate valuation with transparent evaluation metrics. Unlike the driving benchmarks in computer vision (ImageNet) and NLP (GPT datasets), real estate valuation has lacked similar community resources despite its trillion-dollar economic impact.

Our contributions are threefold:

1. **Dataset innovation**: We deliver a reproducible 864,268-observation panel spanning 20 years with 143 features, meticulously engineered temporal variables, and a transparent preprocessing pipeline — establishing a foundation for comparable research that has been notably absent in this domain.
2. **Baseline methodology**: Our autoMIWAE implementation demonstrates that even a relatively simple two-dimensional latent representation can capture fundamental market structure while providing calibrated uncertainty, despite extreme label sparsity (7.2%).
3. **Market structure insights**: The latent space analysis reveals a strong binary market segmentation (silhouette = 0.764) that hierarchically subdivides into eight distinct subtypes with measurable temporal evolution, suggesting fundamental valuation regimes beyond traditional property classifications.

This benchmark sets a new standard for open, large-scale real estate ML research with both predictive and interpretive value. By positioning performance metrics against industry AVMs, academic studies, and regulatory requirements, we provide clear targets for future methodological advances. The moderate computational requirements ensure accessibility for researchers without specialized infrastructure, democratizing a field previously dominated by proprietary systems.

As computational methods continue transforming real estate valuation, standardized benchmarks like NYRE will be essential for rigorous comparison, reproducible findings, and trustworthy deployment of automated systems that impact billions in property wealth and tax assessments.

9. Acknowledgement

We thank William Ho (wh2529) for early discussions and methodological insights. We are grateful to Professor Zoran Kostić (zk2172) for project guidance and feedback throughout development. We acknowledge Columbia University's Data Science Institute for providing GPU compute credits that made this research possible. Additional appreciation goes to the NYC Department of City Planning and Department of Finance for maintaining the PLUTO and ACRIS datasets that form the foundation of this benchmark.

10. References

[1] ecbme6040, “e6691-2025spring-project-nyre-dl3645-hl3793,” GitHub repository, forked from ecbme6040/e6691-2025spring-project, May 2025. [Online]. Available: https://github.com/ecbme6040/e6691-2025spring-project-nyre-dl3645-hl3793. Accessed: May 11, 2025.

[2] (Slides) “Final Presentation: Explainable Graph Neural Networks for House Price Estimation,” May 2025. [Online]. Available: *https://docs.google.com/presentation/d/1UkuPXyohbGSC5xbLIfSc4bXT-7K_rORhHGH9iLyulTU/edit?usp=sharing*. Accessed: May 11, 2025.

[3] H. Li, “Author Guidelines for CMPE 146/242 Project Report,” Lecture Notes of CMPE 146/242, Computer Engineering Dept., San Jose State Univ., San Jose, CA, USA, Mar. 6, 2006, pp. 1.

[4] P. Mattei and J. Frellsen, “MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2019.

[5] Zillow Inc., “Zestimate® Accuracy,” Zillow Research, 2024. [Online]. Available: https://www.zillow.com/zestimate/ Accessed: May 11, 2025.

[6] International Association of Assessing Officers, *Standard on Ratio Studies*, Kansas City, MO, USA: IAAO, 2020.

[7] B. Yu, “Machine Learning for NYC Housing Prices,” in *Proc. Urban Data Science Workshop*, 2024.

[8] A. Peterson and H. Mukherjee, “Automated land valuation models: A comparative study of four machine learning techniques,” *Data Science and Urban Science*, vol. 4, pp. 21–38, 2023.

[9] R. Wheeler *et al.*, “Accounting for Spatial Autocorrelation in Algorithm-Driven Hedonic Models: A Spatial Cross-Validation Approach,” *J. Regional Sci. Assoc.*, vol. 52, no. 3, pp. 782–801, 2023.

[10] C. Riveros, D. Fernández, and J. Pérez, “Scalable Property Valuation Models via Graph-based Deep Learning,” *arXiv preprint* arXiv:2303.12872, 2023.

[11] NYC Dept. of City Planning, “PLUTO and MapPLUTO,” New York City Open Data, 2024. [Online]. Available: https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page Accessed: May 11, 2025.

[12] NYC Dept. of Finance, “Automated City Register Information System (ACRIS),” New York City Open Data, 2024. [Online]. Available: https://www1.nyc.gov/site/finance/taxes/acris.page Accessed: May 11, 2025.

[13] G. Peterson and L. Davis, “Spatial Autoregressive Analysis and Modeling of Housing Prices in Urban Areas,” *ASCE J. Urban Planning*, vol. 137, no. 4, pp. 385–401, 2022.

[14] J. Deng *et al.*, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009.

[15] R. Levina and P. Bickel, “Maximum likelihood estimation of intrinsic dimension,” in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[16] A. Karamanou, P. Brimos, E. Kalampokis, and K. Tarabanis, “Explainable Graph Neural Networks: An Application to Open Statistics Knowledge Graphs for Estimating House Prices,” *Technologies*, vol. 12, no. 8, art. 128, 2024, doi: 10.3390/technologies12080128.

[17] D. Park, F. Rojas, P. Valdés, *et al.*, “PD-TGCN: A Periodically Differenced Temporal Graph Convolutional Network for House-Price Prediction in Santiago,” *IEEE Access*, vol. 9, pp. 127945–127960, 2021.

[18] J. Li, X. Zhang, and C. Tsai, “Spatial Land-Value Prediction with Extreme Gradient Boosting: Evidence from the Melbourne Metro,” *Computers, Environment and Urban Systems*, vol. 97, art. 101880, 2022.

[19] F. Kraus, S. May, and T. Most, “A Holistic Deep-Neural-Network AVM for U.S. Commercial Real Estate: Accuracy and SHAP-Based Interpretability,” *Working Paper*, NCREIF Property Index, 2023.

[20] “House Prices – Advanced Regression Techniques,” Kaggle Competition Dataset, 2016. [Online]. Available: https://www.kaggle.com/c/house-prices-advanced-regression-techniques Accessed: May 11, 2025.

[21] ChatGPT, “Benchmarks, Evaluation Metrics, and Best Practices for Mass Predictive Valuation Models in Real Estate,” ChatGPT conversation, May 2025. [Online]. Available: https://chatgpt.com/share/67fc1a07-d534-8004-b93d-2800652bca94. Accessed: May 12, 2025.

11. Appendix

11.1 Pseudocode

(1) FeatureSelector - Main pipeline for feature selection

Unset

```
DEFINE FeatureSelector(data, features, target, config)

Initialize empty feature_scores dictionary

IF 'variance' IN config.methods
    numeric_features = FILTER_BY_VARIANCE(features, config.variance_threshold)
END IF

IF 'laplacian' IN config.methods
    CALCULATE_LAPLACIAN_SCORE(numeric_features)
END IF

IF 'pseudo_label' IN config.methods
    CALCULATE_ENSEMBLE_PSEUDO_LABEL_SCORES(numeric_features)
END IF

aggregated_scores = AGGREGATE_SCORES(feature_scores)

IF config.correlation_threshold < 1.0
    pruned_features = PRUNE_CORRELATED_FEATURES(aggregated_scores, data, config.correlation_threshold) ELSE
    pruned_features = aggregated_scores
END IF

RETURN TOP-K features from pruned_features based on importance score

END
```

(2) FILTER_BY_VARIANCE - Remove low-variance features

Unset

```
DEFINE FILTER_BY_VARIANCE(features, threshold)

Initialize empty list selected_features

FOR each feature IN features
    IF variance(feature) > threshold
        ADD feature TO selected_features
    END IF
END FOR

RETURN selected_features

END
```

(3) CALCULATE_LAPLACIAN_SCORE - Assess features using graph-based approach

Unset

```
DEFINE CALCULATE_LAPLACIAN_SCORE(features)

FOR each feature IN features
    Build k-nearest neighbors graph
    Compute Laplacian matrix L
    Compute feature's Laplacian score = (f'Lf)/(f'Df) where D is degree matrix
    STORE score in feature_scores[feature]['laplacian_score']
END FOR

END
```

(4)CALCULATE_ENSEMBLE_PSEUDO_LABEL_SCORES - Score features via clustering

Unset

```
DEFINE CALCULATE_ENSEMBLE_PSEUDO_LABEL_SCORES(features)

FOR each k IN config.n_clusters_list
    FOR each method IN config.clustering_methods
        Generate pseudo-labels using clustering
        Calculate feature scores based on these labels
        STORE scores in feature_scores
    END FOR
END FOR

END
```

(5) AGGREGATE_SCORES - Combine multiple metrics into rankings

Unset

```
DEFINE AGGREGATE_SCORES(feature_scores)
Create DataFrame from feature_scores

Compute ranks for each scoring method

Compute aggregated_rank as average of
individual ranks

Compute importance_score based on
aggregated_rank

SORT features by importance_score

RETURN sorted DataFrame

END
```

(6) PRUNE_CORRELATED_FEATURES - Eliminate redundant features

Unset

```
DEFINE
PRUNE_CORRELATED_FEATURES(ranked_features,
data, threshold)

Compute correlation matrix for features

FOR each pair of features (f1, f2)
        IF |correlation(f1, f2)| > threshold
                Drop feature with worse rank
        END IF
END FOR

RETURN remaining features

END
```

12.1 Individual Student Contributions

UNI	Last Name	%	Primary Responsibilities
dl3645	Lewis	50%	Data pipeline development, MIWAE implementation, manuscript drafting
hl3793	Liang	50%	Exploratory data analysis, benchmark literature survey, visualization design

Detailed Contributions

Daniel Lewis (dl3645):

- Designed and implemented the complete data processing pipeline from raw PLUTO/ACRIS to final dataset
- Developed the autoMIWAE architecture with specialized price prediction heads
- Created feature selection framework with variance filtering, Laplacian scoring, and correlation pruning
- Implemented latent space analysis with clustering ensemble methods
- Led manuscript writing and technical documentation

Helen Liang (hl3793):

- Conducted comprehensive literature review of real estate valuation methods and benchmark standards
- Performed exploratory data analysis and descriptive statistics
- Created visualization framework for latent space analysis and temporal trends
- Designed and implemented cluster evaluation metrics and validation procedures
- Coordinated external benchmark collection and comparison methodology