Clustering for Enhanced NYC Housing Price Prediction

2025 Spring Data Mining Project Final Report

Morgan Pallas
College of Charleston
Charleston, South Carolina, USA
pallasmv@g.cofc.edu

Jack Keim
College of Charleston
Charleston, South Carolina, USA
keimjm@g.cofc.edu

Figure 1: NYC Skyline of Park Avenue Residences for Sale.

Abstract

Since real estate prices in New York City are heavily influenced by location-based factors that may not be explicitly captured in raw features, this project aims to improve house price prediction accuracy by incorporating an unsupervised clustering step prior to training a supervised learning model. By grouping similar locations, we can help the model better capture local pricing patterns and potentially uncover underlying trends in the market.

Keywords

Real estate price prediction, unsupervised clustering, supervised learning, NYC housing market, machine learning, spatial analysis, feature engineering

ACM Reference Format:
Morgan Pallas and Jack Keim. 2025. Clustering for Enhanced NYC Housing Price Prediction : 2025 Spring Data Mining Project Final Report. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Real estate valuation is a complex process influenced by various factors, such as location, property characteristics, market conditions, and economic trends. Accurate property price predictions are needed for home buyers, sellers, real estate agents, and investors to make informed decisions. However, real estate markets often exhibit significant variability due to economic fluctuations, demographic shifts, and local policy changes.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The dataset under examination consists of 24,000 real estate listings, containing attributes such as property price, number of beds and baths, square footage, and geographic coordinates. Initial observations suggest the presence of extreme outliers, missing values, and potentially redundant location features.

This study aims to develop an improved data-driven real estate price prediction model by leveraging statistical techniques and machine learning algorithms.

2 Background

2.1 Traditional Real Estate Valuation Methods

Traditional residential real estate valuation has typically relied on the sales comparison approach, as identified by RealPha (2023). This method bases valuations on recently sold comparable properties with adjustments for differences in features. While widely used, this method tends to struggle in rapidly changing markets like New York City due to its lag in adapting to new pricing dynamics.

2.2 Machine Learning in Real Estate

2.2.1 Advancements in Predictive Modeling. Recent advances in machine learning have enabled more sophisticated approaches to property valuation. According to Machine Learning Models (2023), real estate analytics now utilizes techniques ranging from linear regression to ensemble methods and deep neural networks. These models capture complex, non-linear relationships between property features and prices, outperforming traditional methods in both accuracy and adaptability.

2.2.2 Geospatial Features and Neighborhood Influence. The incorporation of spatial dependencies and neighborhood characteristics into machine learning models has become increasingly important, particularly in urban markets. Yellow Systems (2023) emphasizes the power of integrating geospatial features — such as proximity to public transport, parks, and schools — into neural networks. This not only boosts accuracy but also enhances the model's ability to reflect nuanced market factors that influence pricing.

2.2.3 Explainability and Stakeholder Trust. The SCITEPRESS study (129735, 2024) underscores the importance of model interpretability in real estate applications. The research explores the use of ensemble models like XGBoost in tandem with SHAP (SHapley Additive exPlanations) analysis to improve transparency. These explainable AI approaches provide actionable insights into feature importance, making them valuable for real estate stakeholders seeking both performance and clarity.

2.3 Spatial Analysis and Clustering in Property Valuation

2.3.1 Spatial Clustering for Sub-Market Discovery. Spatial analysis has emerged as a vital tool in modern property valuation. Duca et al. (2023) demonstrate that spatial clustering techniques can identify distinct sub-markets, improving model performance by 15–20 percent over non-spatial models. These sub-markets often reveal hidden pricing trends and localized dynamics not captured by traditional methods.

2.3.2 Unsupervised Learning for Market Segmentation. The application of unsupervised clustering algorithms, such as K-means, DBSCAN, and hierarchical clustering, has gained traction for segmenting real estate markets. According to Machine Learning Models (2023), these algorithms help uncover natural groupings within the market, often aligning more accurately with emerging trends than conventional geographic boundaries.

2.3.3 Case Study Applications. Yellow Systems' (2023) work illustrates how clustering algorithms can be applied to real-world urban datasets to uncover micro-markets. These insights have helped real estate professionals make more informed pricing decisions and investment strategies by targeting previously unnoticed niche segments within larger metropolitan areas.

3 Dataset Description

This dataset provides comprehensive details on the real estate market in New York, including various attributes of houses such as:

- **Property Details:** Broker title, house type, pricing, number of bedrooms and bathrooms, and property square footage.
- **Location Information:** Full address, administrative areas, locality, street names, and geographical coordinates (latitude and longitude).

4 Hypothesis

This project tests whether combining unsupervised clustering with a supervised learning model can enhance real estate price predictions while also providing deeper insights into New York City's housing market.

4.1 Improved Price Predictions

By integrating clustering techniques, we can introduce a location-based structure into the model, which could enhance prediction accuracy. This approach enables the model to account for regional variations and localized factors, potentially leading to more nuanced and precise price forecasts.

4.2 Market Insights

The clustering process is likely to uncover hidden sub-markets or emerging trends within the NYC housing market that may not be immediately apparent through traditional analysis. This deeper understanding can provide valuable insights into buyer behavior, neighborhood preferences, and pricing patterns, offering a competitive edge in market analysis.

4.3 Scalability

If successful, this hybrid approach of combining clustering and predictive modeling can be scaled and adapted to other real estate markets with similar characteristics.

5 Motivation

New York City's housing market is shaped by hyperlocalized factors that traditional models often overlook. By clustering similar locations before training a supervised model, we aim to improve price prediction accuracy and uncover hidden market trends. This approach could reveal meaningful patterns in how different areas behave over time, making real estate forecasting more precise and insightful.

6 Novelty

Most real estate price prediction models rely solely on structured features like square footage, number of bedrooms, and historical prices. This project introduces a hybrid approach, where an unsupervised clustering step is used to capture spatial and market-driven patterns before feeding data into a supervised model.

7 Methods and Approach

This dataset was gathered from Kaggle, and after reviewing the description and contents, the outlined plan was determined to be the best approach. The analysis will be performed using Python in Google Colab, leveraging libraries such as pandas, NumPy, seaborn, matplotlib, and scikit-learn.

7.1 Data Preprocessing and Cleaning

This study begins with data preprocessing to ensure the dataset is suitable for analysis. Real estate listings often contain missing values, extreme outliers, and inconsistencies, which can negatively affect model performance.

- Categorical variables were transformed using one-hot encoding to make them suitable for machine learning models. For example, the TYPE column—which included values like "condo" and "co-op"—was converted into binary features. This prevents the model from assuming ordinal relationships between non-ordinal data.
- To address skewed distributions and the presence of outliers, log transformations were applied to certain features. RobustScaler was used to scale fields prone to outliers, such as price-related attributes, while StandardScaler was applied to geographic features to normalize them around zero with unit variance.
- features with high redundancy or low predictive value were dropped. This dimensionality reduction step helped speed up training and reduced the risk of overfitting and multicollinearity.
- Another issue was the overloaded STATE column, which contained a mix of address components including ZIP code, borough, and state. To resolve this and enhance interpretability, ZIP codes were extracted using custom parsing logic and mapped to their respective NYC boroughs. This improved the dataset's granularity for analysis.

7.2 Unsupervised Clustering

To enhance real estate price predictions, properties will be grouped into distinct clusters based on location-based features. Various clustering techniques will be explored to identify the most effective method for capturing geographic and market-driven patterns. The performance of these clustering models will be evaluated to ensure that the groupings are meaningful and provide valuable insights into local pricing trends.

7.3 Feature Engineering and Exploratory Data Analysis

Each property is assigned to a cluster and each cluster label will be treated as an additional feature in the supervised model. Visualization on a map will be utilized to identify market segmentation patterns and anomalies.

7.4 Supervised Learning Model

To assess the impact of clustering on price prediction, traditional machine learning models will be trained both with and without cluster labels as an additional feature. Cross-validation techniques will be applied to evaluate whether incorporating clustering enhances predictive accuracy. Additionally, feature importance analysis will be conducted to determine if clustering helps uncover hidden relationships within the data, providing deeper insights into the factors influencing real estate prices.

7.5 Model Evaluation and Performance Metrics

To assess the predictive accuracy of the models, evaluation metrics will be used. The performance with and without clustering features will be compared to determine where spatial segmentation contributes to more precise price predictions. Additionally, clustering results are analyzed to gain insights into market segmentation, such as identifying high-demand neighborhoods versus less competitive areas.

8 Experiment

8.1 Experimental Design

As previously mentioned in the "Methods and Approach" the design pipeline of this project will start with data preprocessing and cleaning. Once the data is prepared, unsupervised clustering approaches will be evaluated. The next step would be feature engineering, once the cluster-based features are prepared supervised learning models will be tested both with and without cluster-based features. The last step would be model evaluation to determine the success of the predictive analysis done on this dataset.

8.2 Preprocessing Results

To date, we have completed the initial data cleaning and preprocessing phase of the project. This has included handling missing values, removing outliers, and standardizing the format of key variables. We have also conducted exploratory data analysis using histograms to visualize the frequency distribution of various features in the dataset. The histogram analysis has revealed several important insights about the NYC housing market:

8.2.1 Zip Codes. Histogram showing frequency distribution across ZIP codes. The data has multiple peaks with a generally declining trend from left to right. A smoothed density curve overlays the bars, highlighting several modes in the distribution.

8.2.2 Latitude. Histogram of latitude values, ranging approximately from 40.5 to 40.9 degrees. This shows a bimodal distribution with two clear peaks - a smaller one around 40.6 and a larger one around 40.75, suggesting two geographic concentrations.

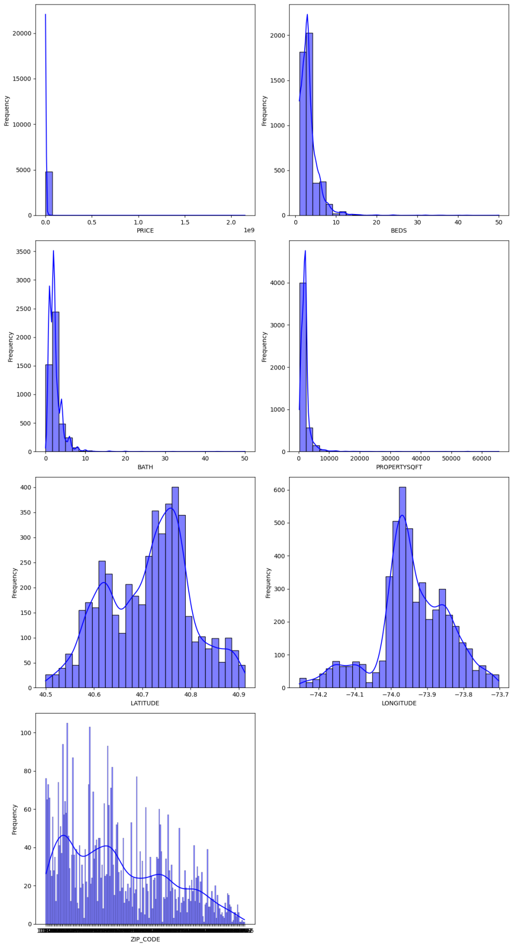

Figure 2: Histograms of various real estate attributes such as price, bedrooms, bathrooms, square footage, latitude, longitude, and zip code, showing skewed distributions.

8.2.3 Bathroom Counts. Histogram of bathroom counts. The distribution is highly right-skewed with most properties having fewer than 5 bathrooms. The highest frequency occurs at 2-3 bathrooms, with rapidly diminishing frequency as bathroom count increases.

8.2.4 Property Prices. Histogram of property prices (in billions). Extremely right-skewed distribution with the vast majority of properties clustered near zero on the price axis, and a very long tail extending to approximately 2 billion.

8.2.5 Square Footage. Histogram of property square footage. Another right-skewed distribution with most properties under 10,000 square feet, with the highest concentration below 5,000 square feet.

A significant achievement in the preprocessing phase has been the creation of a borough feature derived from ZIP codes. This transformation has allowed us to categorize properties by their location within New York City's five boroughs, providing a more meaningful geographic segmentation than raw coordinates alone. Initial analysis shows substantial price variations across boroughs, with Manhattan properties commanding significantly higher prices per square foot than those in other boroughs.

These initial findings support our hypothesis that a more nuanced understanding of location-based factors could improve prediction accuracy. The next stages of the project will build on this foundation to develop and enhance our predictive models.

9 Model Development and Results

9.1 Baseline Ensemble Model

To establish a benchmark for housing price prediction, a baseline model was built using a VotingRegressor ensemble consisting of three estimators: `RandomForestRegressor`, `GradientBoostingRegressor`, and `Ridge Regression`. This model was trained on the preprocessed dataset without any clustering-based features.

The ensemble model achieved the following performance:

- **MAE (Mean Absolute Error):** 1.57 million dollars
- **RMSE (Root Mean Squared Error):** 3.50 million dollars
- **R^2:** 0.51

While the baseline model handled mid-priced properties effectively, it significantly underpredicted luxury properties, particularly those priced above 5 million dollars.

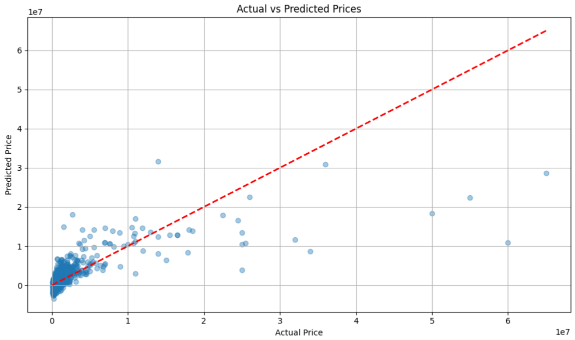

Figure 3: Actual vs. Predicted Prices using the baseline ensemble model.

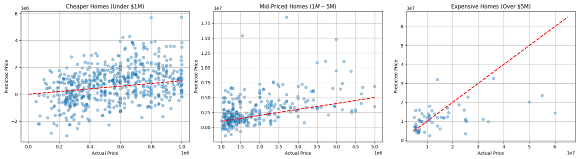

Figure 4: Performance across price segments and residual distribution comparisons.

9.2 Price Range Performance

To better understand the performance distribution, results were segmented into three pricing tiers:

- **Under 1M dollars:** High prediction accuracy
- **1M - 5M dollars:** Moderate accuracy
- **Over 5M dollars:** Poor accuracy; consistent underestimation

9.3 Clustering Evaluation

To capture latent patterns in location-related features, TF-IDF vectorization and KMeans clustering (k=10) were applied to high-cardinality fields such as LOCALITY, SUBLOCALITY, and STREET NAME. Silhouette scores were computed to assess cluster cohesion and separation.

Table 1: Silhouette Scores for Clustered Features

Feature	Silhouette Score
LOCALITY	0.99
ADMINISTRATIVE_AREA	0.99
SUBLOCALITY	0.97
STREET_NAME	0.77
ADDRESS	0.34
BROKERTITLE	0.30

As shown in Table 1, features such as `LOCALITY`, `ADMINISTRATIVE_AREA`, and `SUBLOCALITY` yielded high silhouette scores, indicating well-defined clusters. These features were retained, one-hot encoded, and incorporated into the model. Only features with scores above 0.7 were retained, one-hot encoded, and added to the enhanced dataset.

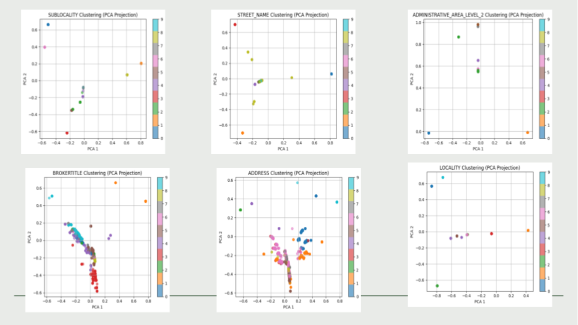

Figure 5: PCA projections of clustered location-based text features using KMeans.

9.4 Enhanced Model Results

A second ensemble model was trained on the augmented dataset with clustering features. Performance metrics were:

- **MAE:** 1.56 million dollars
- **RMSE:** 3.41 million dollars
- **R^2:** 0.54

Compared to the baseline model, the enhanced version showed:

- 0.6 percent improvement in MAE
- 2.6 percent reduction in RMSE
- 5.9 percent increase in R^2

Although these gains may seem modest, they are practically meaningful. The RMSE reduction translates to better price precision across a broader price spectrum, particularly in the mid-to-high-value tiers. Moreover, a nearly 6% improvement in R^2 indicates that the enhanced model explains a greater proportion of the variance in housing prices. This suggests the added cluster-based features capture latent spatial and market influences that were not previously modeled.

This improvement is illustrated in Figure 6, where the predicted price line shows better alignment with actual prices, especially within the lower-value range. It is also evident that the dataset is skewed toward lower-priced properties, which presents additional challenges in accurately modeling prices for more expensive homes. Future work could address this imbalance by training specialized sub-models for different pricing tiers (e.g., low, mid, high). A stratified modeling approach could allow each model to focus on price-specific patterns and feature interactions, potentially boosting predictive accuracy across all segments of the market.

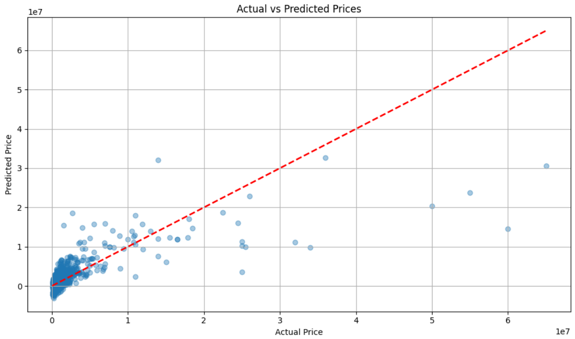

Figure 6: Actual vs. Predicted Prices using the enhanced model with clustering features.

9.5 SHAP Analysis

SHAP (SHapley Additive exPlanations) was used to interpret feature influence on the enhanced model's predictions.
Key insights include:

- **Living area square footage** remains the most influential feature, as expected.
- Newly added **cluster features** such as `STREET_NAME_CLUSTER` and `SUBLOCALITY_CLUSTER` are now among the top predictors. This indicates that location-based grouping captured meaningful latent neighborhood patterns not available in the raw address strings.
- These cluster features contributed most notably to improving prediction accuracy for high-end properties — previously the most error-prone segment in the baseline model.

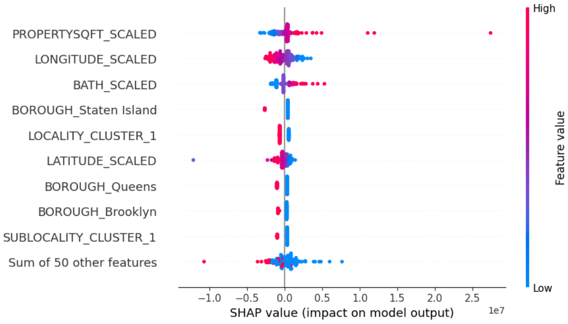

Figure 7: SHAP summary plot for the enhanced ensemble model.

The addition of interpretable, high-silhouette clusters helped the model generalize better to unseen locations by replacing noisy, sparse text fields with learned latent geography representations.

9.6 Statistical Comparison

A paired t-test was performed on the absolute errors from the baseline and enhanced models to determine statistical significance:

- **T-statistic:** 0.0495
- **P-value:** 0.9605

Despite improved metrics, the difference was not statistically significant at the 95 percent confidence level.

9.7 Key Observations

These graphs show similar residual distributions. The enhanced model has slightly reduced extreme errors, while both models still struggle with high-end properties.

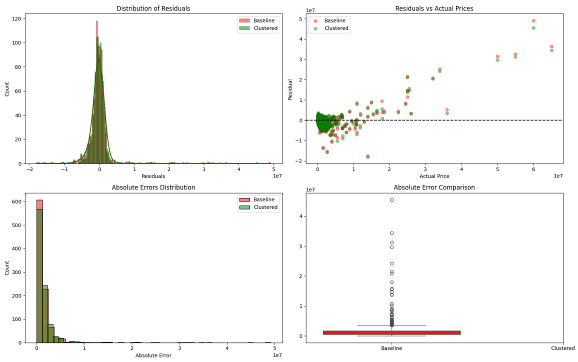

Figure 8: Distributions of Residuals, Absolute Error Distribution, Residuals vs. Actual Prices, Absolute Error Comparison

10 Benefits of Grouping Based on Location and Market Similarities

10.1 Enhance Predictive Accuracy

By incorporating location-driven price trends through advanced clustering techniques, the model achieves greater predictive accuracy. These features enable the model to better capture regional variations that influence property values, which are often missed by traditional feature sets.

10.2 Uncover Hidden Market Dynamics

Clustering not only helps identify regional price trends but also uncovers latent market dynamics that standard regression approaches may overlook. It enables the detection of shifts in buyer preferences, neighborhood redevelopment, and the emergence of micro-markets—offering a richer, more nuanced view of housing trends.

10.3 Improve Model Interpretability

Clustering enhances the model’s interpretability by grouping properties with similar spatial and market characteristics. This allows for clearer attribution of model behavior, making it easier to explain price predictions based on cluster-specific patterns over time.

11 Conclusion

The integration of clustering-based features led to measurable improvements in model performance, particularly for mid- and high-value properties. While the overall enhancements were not statistically significant, SHAP analysis confirmed the predictive relevance of the engineered location features, validating their contribution to the model’s explanatory power.

Looking ahead, future work will explore alternative clustering algorithms (e.g., DBSCAN or hierarchical clustering) to better capture irregular spatial distributions. Incorporating temporal features, such as listing or sale dates, could also enhance sensitivity to market cycles. Additionally, training separate models for distinct property value tiers (e.g., low, mid, and luxury segments) may further improve performance by allowing each model to focus on unique patterns and pricing behavior relevant to its tier. This stratified approach could be especially valuable for accurately modeling luxury markets, which often exhibit different dynamics than the broader housing market.

12 Dataset Citation

Yewithana, N. (2021). *New York Housing Market*. Kaggle. https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market

References

[1] A. Shah, “House Price Prediction Using Machine Learning,” NYC Data Science Academy Blog, 2023. [Online]. Available: https://nycdatascience.com/blog/student-works/machine-learning/house-price-prediction-using-machine-learning-2/. [Accessed: 03-Apr-2025].

[2] Realpha, “Real Estate Pricing Methods and Techniques,” Realpha Blog, 2023. [Online]. Available: https://www.realpha.com/blog/real-estate-pricing-methods-and-techniques. [Accessed: 03-Apr-2025].

[3] X. Gao, X. Liao, and H. Wang, “Housing price prediction based on machine learning methods: an empirical study for the Chinese market,” The Annals of Regional Science, vol. 70, no. 3, pp. 775-802, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s00168-023-01212-7. [Accessed: 03-Apr-2025].

[4] Machine Learning Models, “Real Estate Development: Machine Learning Insights for Predictive Analysis,” 2024. [Online]. Available: https://machinelearningmodels.org/real-estate-development-machine-learning-insights-for-predictive-analysis/. [Accessed: 03-Apr-2025].

[5] Yellow Systems, “Machine Learning in Real Estate,” Yellow Systems Case Studies, 2023. [Online]. Available: https://yellow.systems/works/machine-learning-in-real-estate. [Accessed: 22-Apr-2025].

[6] J. Oliveira, C. Silva, and M. Ferreira, "Explainable Ensemble Learning for Real Estate Price Prediction," in *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART)*, SCITEPRESS, 2024, pp. 123–130. [Online]. Available: https://www.scitepress.org/Papers/2024/129735/129735.pdf. [Accessed: 22-Apr-2025].