2025 Spring Final Project
Data Mining CSCI 334

College of Charleston
Charleston, South Carolina

Pallasmv@g.cofc.edu
Keimjm@g.cofc.edu

Clustering for Enhanced NYC Housing Price Prediction

Morgan Pallas

Jack Keim

Introduction

This project started with a dataset of NYC residential prices and property features. Our goal:

- Build a robust model to predict housing prices
- Enhance prediction using clustering on unstructured text features

Background:

Evolving Methods in Real Estate Valuation

Traditional Valuation Methods:

- Sales Comparison Approach: Value homes based on similar properties sold.
-

Rise of ML in Real Estate:

- Enabled more advanced, data-driven valuation methods.
-

Spatial Analysis & Clustering:

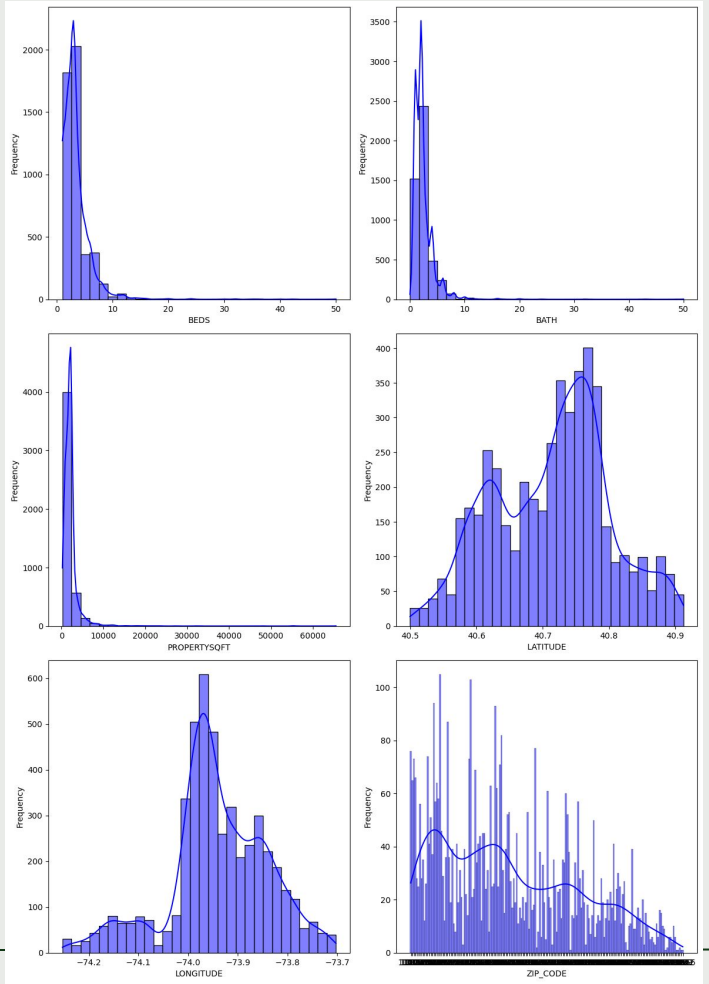
- Spatial clustering: Improves model accuracy by identifying sub-markets with distinct pricing.
- Unsupervised clustering: Reveals natural market segments.

Dataset

Sourced from Kaggle: “New York Housing Market”(~24,000 listings)
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).

Histograms of feature distributions



Cleaning & Preprocessing

Geographic Processing

This dataset had ZIP codes imbedded in the STATE column. (not standard)

Custom parsing logic extracted ZIP codes and mapped each to one of the NYC's five boroughs

This added a high-level geographic feature to help distinguish property value trends across boroughs.

Encoding Categorical Variables

Categorical columns like TYPE (e.g condo, co-op) and the new BOROUGH feature were one-hot encoded

This transformed text labels into binary columns that ML models can process.

Prevents misleading ordinal relationships in non-ordinal data.

Feature Selection & Redundancy Removal

Dropped columns with high redundancy or little predictive value.

Reduced dimensionality to speed up model training and prevent overfitting.

Feature Scaling

Log Transformation: Applied to skewed features to reduce impact of large outlier and normalize distributions

RobustScaler: Used for features susceptible to outliers, such as price-related fields.

StandardScaler: Used to scale Geographical features, centering the data around 0 and scaling by SD.

Baseline Ensemble Model

Model Type

A VotingRegressor ensemble combining:

- RandomForestRegressor
- GradientBoostingRegressor
- Ridge Regression

Performance

MAE: $1.57M
RMSE: $3.50 M
R^2: 0.51

Insights

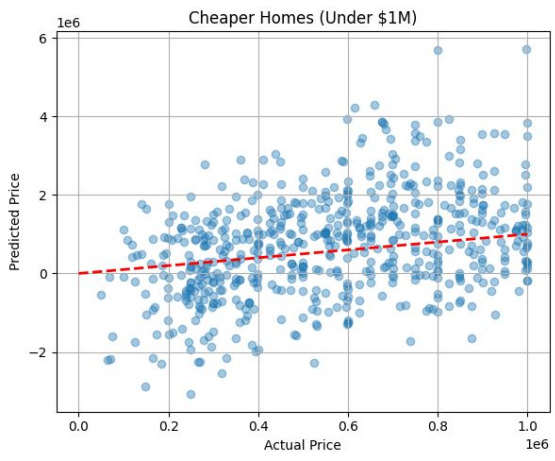
The model performed reasonable well on mid-range properties, but struggled with extremes- particularly underpredicting expensive homes.

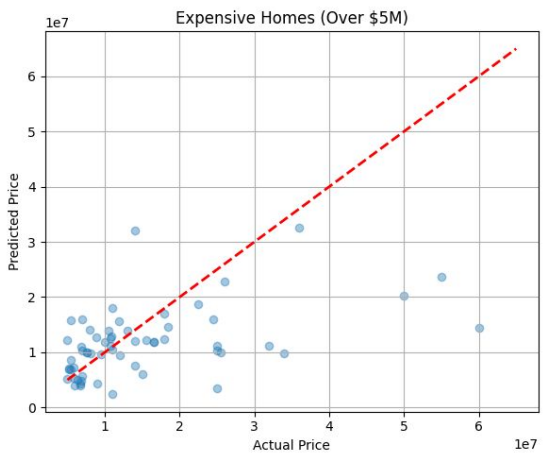
Price Range Performance

Price Range: Under $1M
Performance: Good fit

Price Range: $1M - $5M
Performance: Moderate

Price Range: Over $5M
Performance: Poor fit





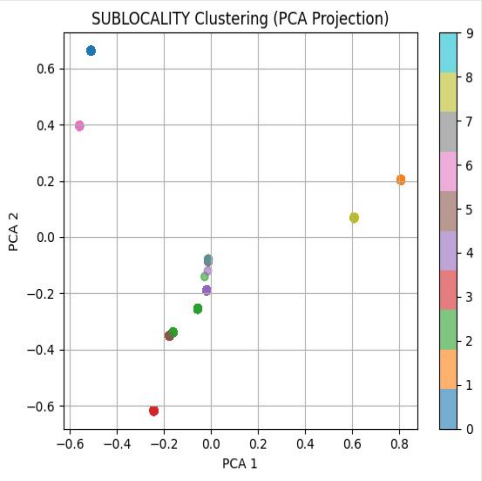
Clustering Evaluation

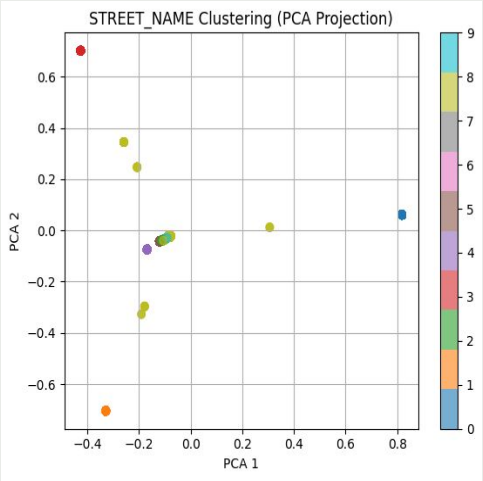
Silhouette Scores:
Higher scores indicate better-defined clusters (score>0.7)

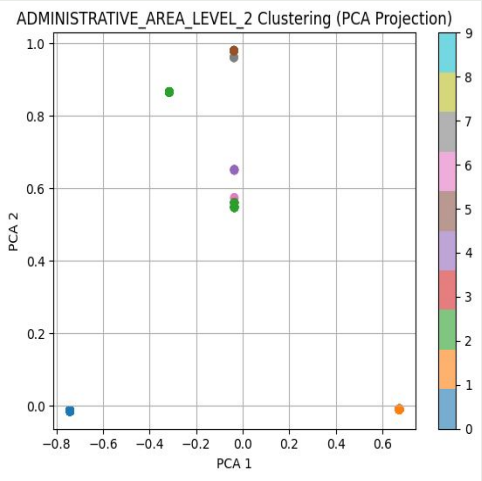
Methods:

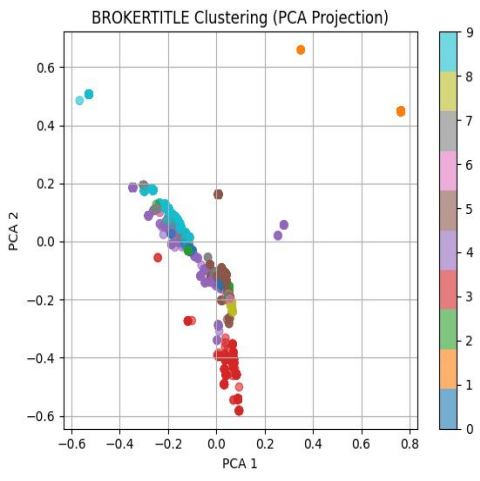
- TF-IDF vectorization
- K-Mean clustering
- PCA for visualization

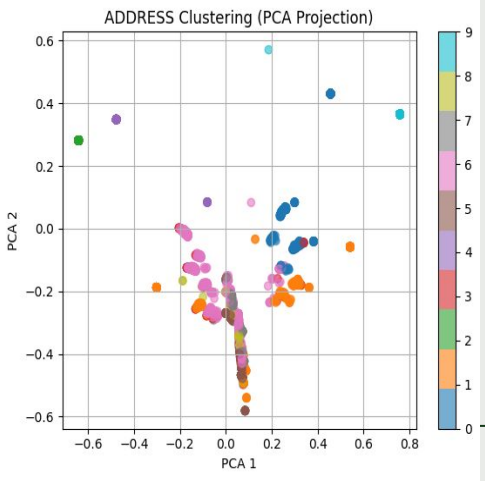
Feature	Score
LOCALITY	0.99
ADMINISTRATIVE_AREA	0.99
SUBLOCALITY	0.97
STREET_NAME	0.77
BROKERTITLE	0.30
ADDRESS	0.34

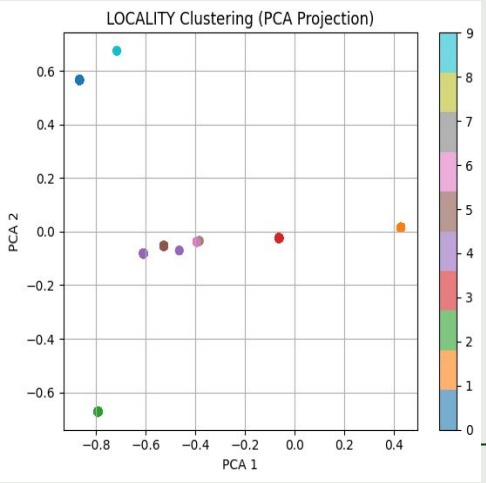

SUBLOCALITY Clustering (PCA Projection)
PCA 1
PCA 2


STREET_NAME Clustering (PCA Projection)
PCA 1
PCA 2


ADMINISTRATIVE_AREA_LEVEL_2 Clustering (PCA Projection)
PCA 1
PCA 2


BROKERTITLE Clustering (PCA Projection)
PCA 1
PCA 2


ADDRESS Clustering (PCA Projection)
PCA 1
PCA 2


LOCALITY Clustering (PCA Projection)
PCA 1
PCA 2

Enhanced Model Results

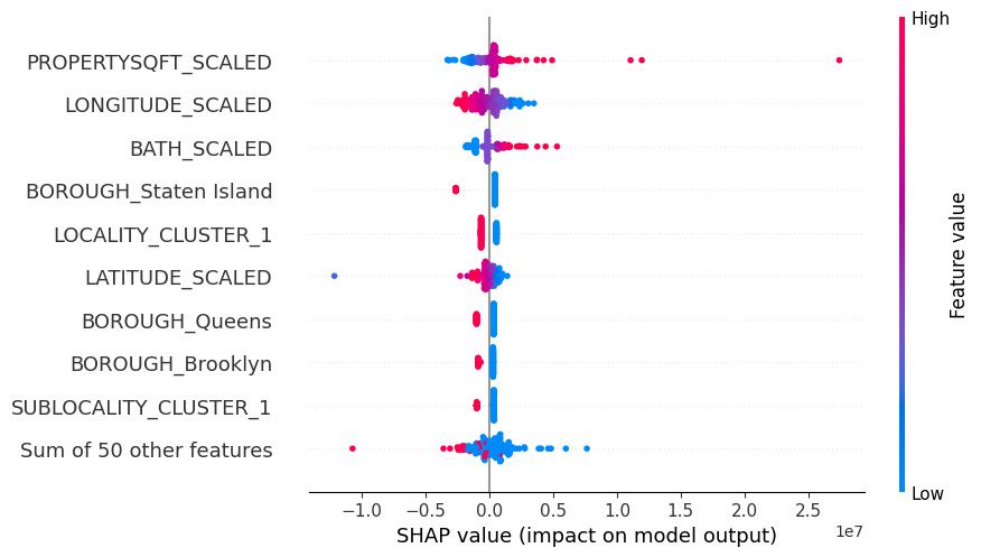
Performance with Cluster Features:

- MAE: $1.56M
- RMSE: $3.41M
- R^2: 0.54

Improvements:

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

SHAP Analysis



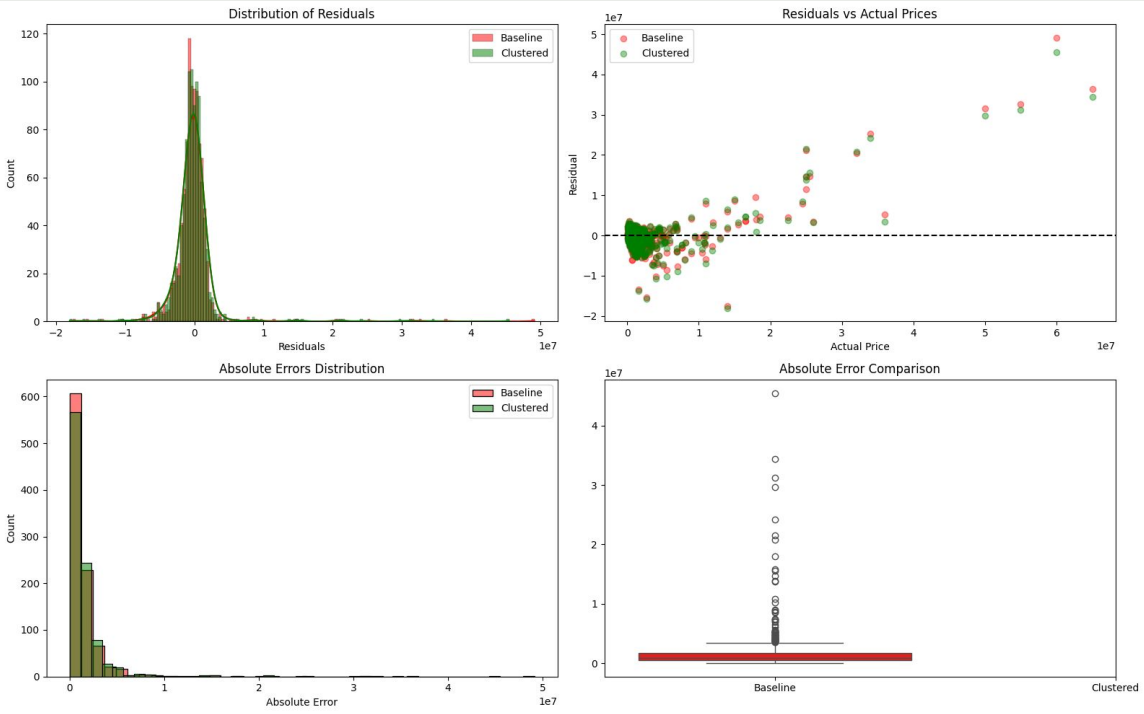
SHAP assigns a unique number to each feature, it helps to measure how important that feature is in predicting outcomes.

Confirmed increased importance of engineered features like STREET_NAME_CLUSTER and LOCALITY_CLUSTER

Statistical Comparison

Paired t-test on absolute errors:
T-Statistic = 0.0495
P-Value = 0.9605

Conclusion: While performance metrics improved, the difference was not statistically significant



Key Observations:

- Similar residual distributions
- Enhanced model slightly reduced extreme errors
- Both models still struggle with high-end properties

Summary

Insights

1. Feature Engineering Impact: Text clustering added predictive value, especially for location-based features
2. Statistical Significance: Improvements from clustering were subtle, not statistically significant
3. Feature Importance: Property size, geographic features, and neighborhood clusters were top predictors

Future Plans

1. Explore alternative clustering methods.
2. Incorporate time-series elements to capture market trends
3. Implement separate models for different property types/price ranges

Impediments

Both models underpredict luxury properties

Questions?