A Comprehensive Analysis of Real Estate Dynamics

DATS 6103: Introduction to Data Mining

Team 5: Aishitha Pachipala, Anusha Umashankar, Anurag Surve, Sayan Patra

16 December 2024

INTRODUCTION

The real estate market is a dynamic system shaped by factors such as property features, location, market trends, and broker influence. Data-driven insights are increasingly vital for navigating this complex landscape. This project, *A Comprehensive Analysis of Real Estate Dynamics*, leverages machine learning techniques to uncover patterns and enhance decision-making in real estate.

This project's main objective is to use machine learning techniques to analyze factors influencing the price of a property and other real estate market dynamics. The analysis is divided into four components:

1. **Property Price Prediction**: Using Linear Regression, we predict property prices based on features such as house size, bedrooms, bathrooms, lot size, and location. This analysis estimates property values and highlights the influence of features across different states.
2. **Market Status Prediction**: Logistic Regression is used to classify property statuses (e.g., 'for_sale' or 'ready_to_build') based on features such as price, size, and location. This analysis identifies the key factors influencing whether a property is market ready.
3. **Market Segmentation**: K-Means clustering is employed to group properties into categories—Affordable, Mid-Range, and High-End—based on features like price, house size, and lot size. This segmentation reveals market patterns and supports targeted strategies.
4. **Broker Analysis**: K-Means clustering is used to analyze broker performance based on the properties they manage. This section identifies top-performing brokers in specific market segments and can be used to provide recommendations for optimizing broker-property assignments to maximize sales potential.

The dataset is sourced from Kaggle and contains real estate information across regions and is available at USA Real Estate Dataset.

LITERATURE REVIEW

With the real estate market booming in past years, and with technology advancing alongside, integrating the idea of using machine learning techniques to predict property prices and house evaluations has gained attention. Traditional methods, such as expertise of the agent and hedonic pricing models, often fall short in capturing the complexities of modern housing markets. In contrast, data-driven decision-making leverages vast datasets and advanced algorithms to enhance prediction accuracy and market insights.

Mu, Wu, and Zhang (2014) investigated the potential of machine learning models for housing value forecasting. Their study highlighted the ability of machine learning techniques to adapt to the multifaceted nature of housing markets. Similarly, Sanyal et al. (2022) demonstrated the potential of regression models in predicting Boston house prices. By analyzing property characteristics such as square footage, number of rooms, and proximity to schools, their study identified non-obvious relationships between features and property values, achieving significantly better prediction accuracy compared to conventional techniques. Overall, the literature present tells us the effectiveness of machine learning methods in property price prediction, emphasizing their adaptability and enhanced precision in capturing market complexities.

DATASET DESCRIPTION

The dataset used for this project is the USA Real Estate Dataset from Kaggle, which provides information on real estate properties across the United States. It contains features critical for analyzing property prices, market status, and broker performance.

Numerical Variables:

- ***price***: The listing price of the property.
- ***house_size***: The size of the property in square feet.
- ***bed***: The number of bedrooms.
- ***bath***: The number of bathrooms.
- ***acre lot***: The property's lot size in acres.

Categorical Variables:

- ***city***: The city where the property is located.
- ***state***: The state where the property is located.
- ***zip_code***: The postal code.
- ***status***: The listing status (e.g., 'for_sale', 'ready_to_build').
- ***brokered_by***: The name of the broker or brokerage managing the property.

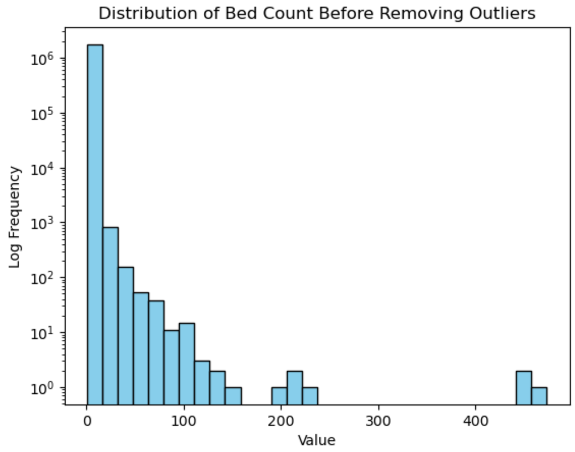
DATA CLEANING AND PREPROCESSING

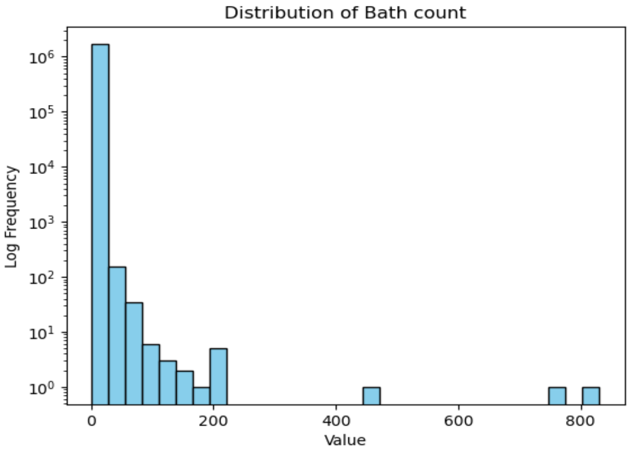
To ensure data quality and reliability, a structured preprocessing was applied before EDA:

- **Outlier Detection and Mitigation**: Outliers in bed and bath counts were identified using the Interquartile Range (IQR) method, and extreme values were capped at the upper bound.
- **Handling Missing Data**: Missing values in critical numerical features (bed, bath, acre lot, and house size) were imputed using statistically appropriate measures such as the mean (for acre lot and house size) and random value within a bounded range (for bed and bath).
- **Categorical Variable Encoding**: Nominal features such as state and city were encoded using label encoding to facilitate their inclusion in correlation analysis.

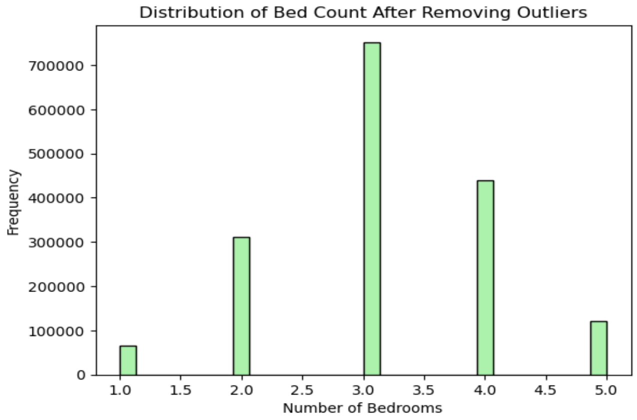
Bedroom and Bathroom Distribution

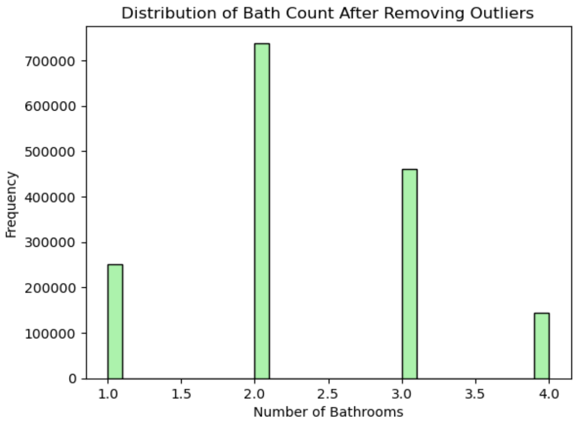
Pre-Cleaning: Histograms on a log scale revealed extreme outliers, including houses with implausibly large bedroom or bathroom counts.





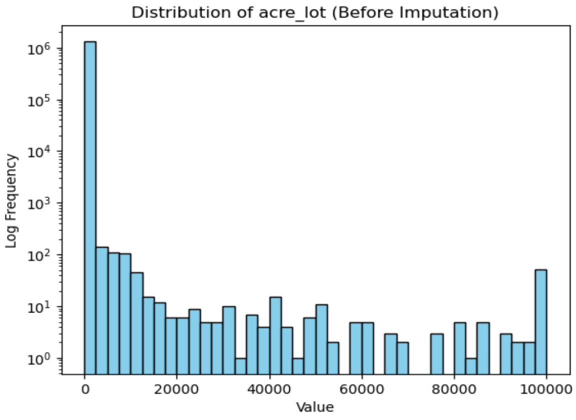
<u>Post-Cleaning</u>: The removal of upper-bound outliers resulted in a normalized distribution, with most houses containing 2-4 bedrooms and 2-3 bathrooms.



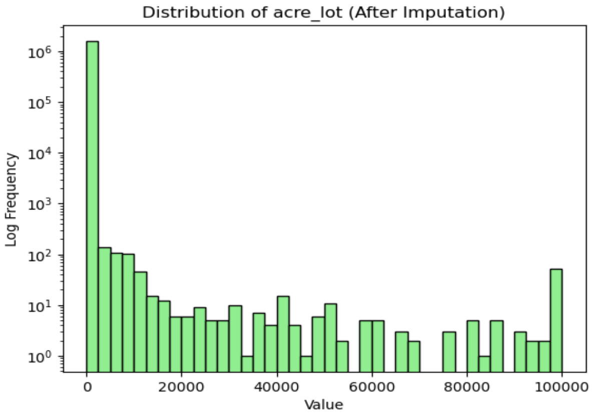


Acre_Lot Distribution

<u>Pre-Imputation</u>: Visualizations indicated a heavy concentration of smaller lot sizes, with a small subset of extreme values.



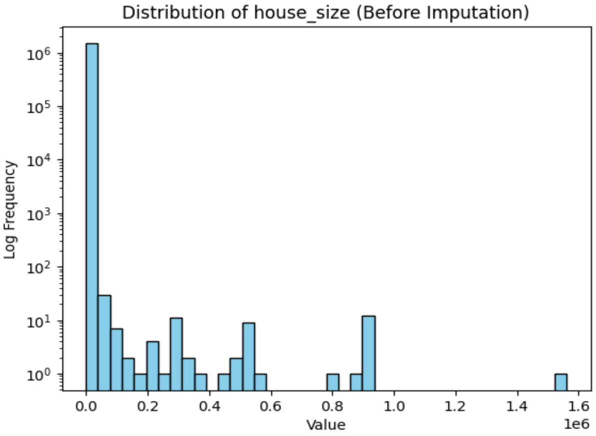
<u>Post-Imputation:</u> Missing values were replaced with the mean.

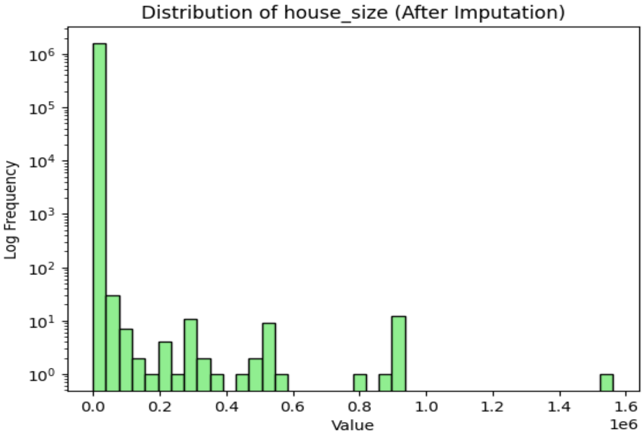


House_Size Distribution

Pre-Imputation: Visualizations show a heavy concentration of smaller house sizes.

Post-Imputation: Missing values were replaced with the mean.

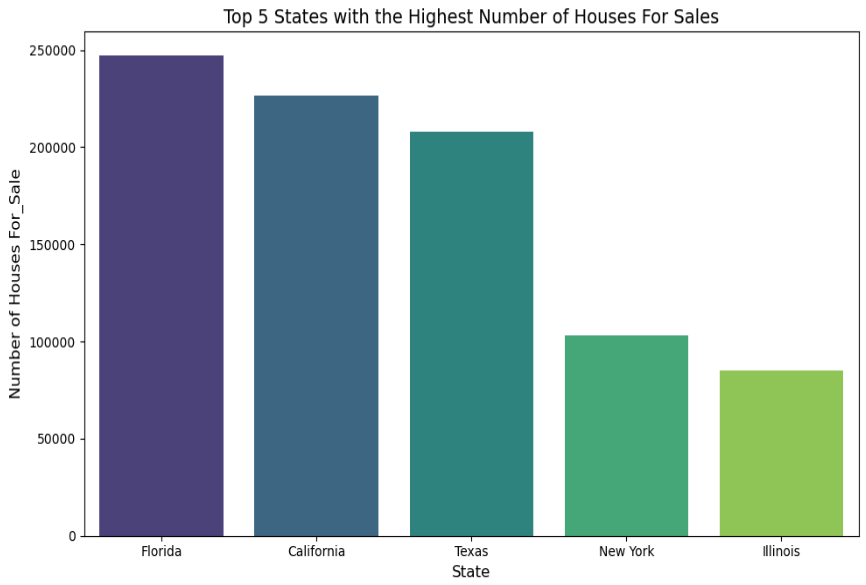




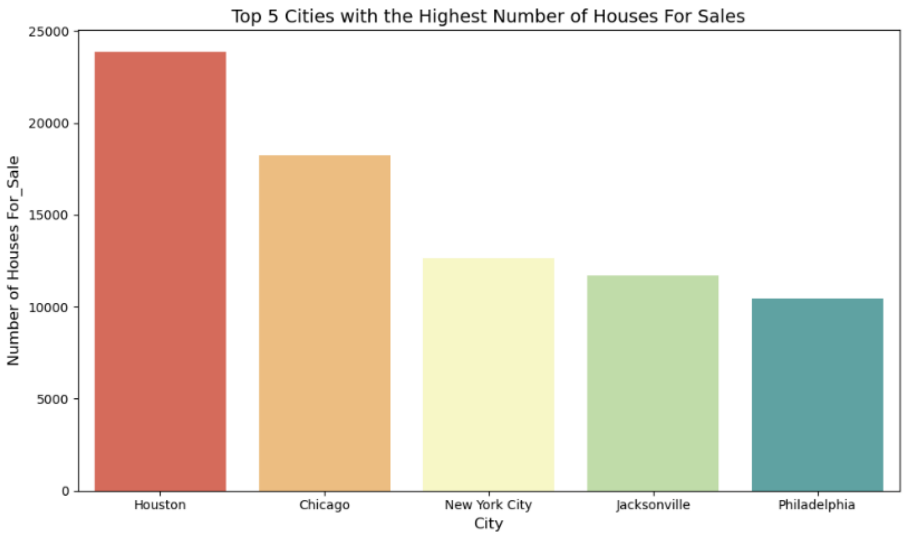
EXPLORATORY DATA ANALYSIS

Distribution of Sales

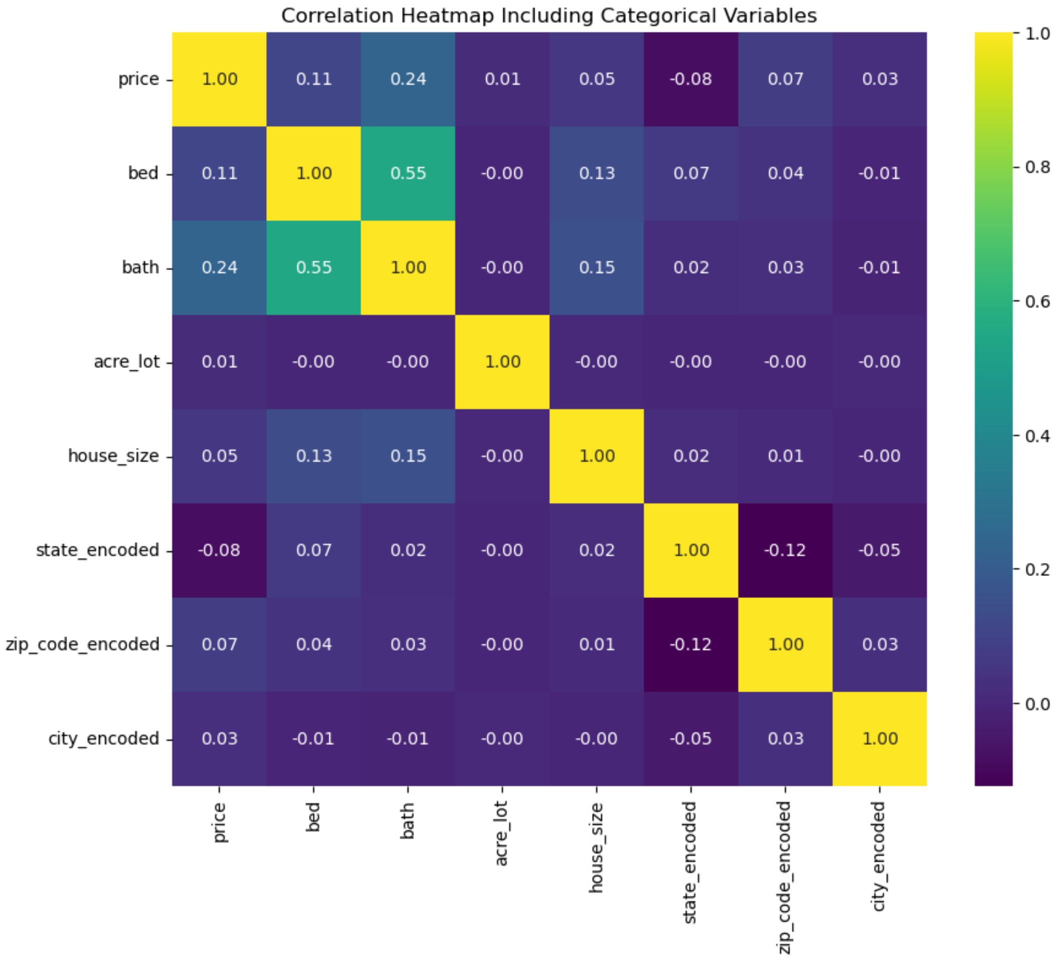
The visualization below highlights the Top 5 States with the highest number of houses for sale. Florida leads with the highest number of homes sold, likely due to its year-round warm weather, beautiful beaches, and status as a retirement and vacation haven. California is second, with its strong economy, job opportunities in tech and entertainment attracting buyers. Texas on the other hand offers affordable housing, a low cost of living, and no state income tax, making it an ideal choice for families and businesses. Both New York and Illinois attract homebuyers with their urban opportunities, cultural richness, and diverse housing options, offering a balance of vibrant city life, historical charm, and suburban communities that cater to a variety of lifestyles.



The visualization below highlights the Top 5 Cities with the highest number of houses for sale. Houston leads the list, likely due to its affordable housing market, strong job opportunities, and growing population. Chicago follows, offering a mix of urban convenience and diverse housing options. New York City, despite its high cost of living, remains attractive for its economic opportunities and cultural appeal. Jacksonville benefits from its warm climate, family-friendly environment, and relatively low housing costs, while Philadelphia attracts buyers with its historical charm, affordability, and proximity to major metropolitan hubs.



Correlation Analysis:



A strong positive correlation (r > 0.6) was found between house size and price. Moderate positive correlations were observed between the number of bedrooms, bathrooms, and house size, reflecting logical relationships among these variables. In contrast, weak correlations between lot size (in acres) and other features suggest that lot size may be influenced by external factors such as location or zoning regulations.

PROPERTY PRICE PREDICTION

Objective:

By exploring the relationship between the price of the property and its features, we aimed to predict the property price and provide actionable insights that could benefit stakeholders in the real estate market.

Features Selected: price, bed, bath, acre lot, house size, city

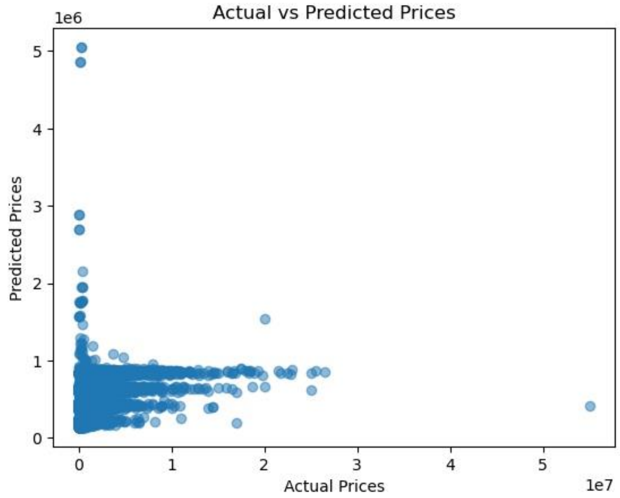
Feature Engineering: Scaled numerical features to ensure uniformity and encoded ‘city’, a categorical feature.

Linear Regression:

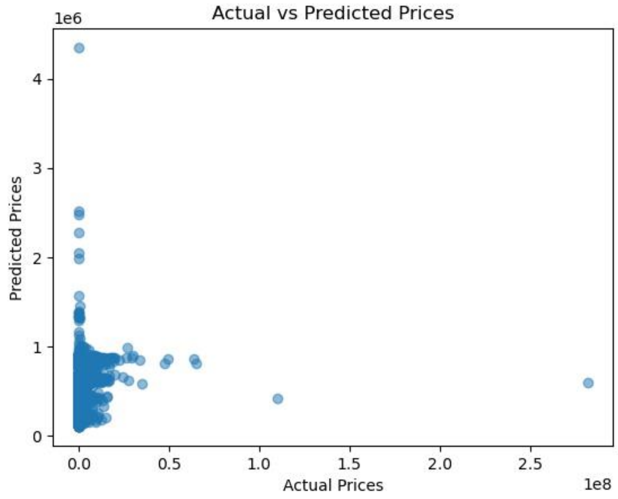
The dataset was divided into two parts: 80% for training the model and 20% for testing its performance. The property price was first predicted (model 1) using just the numerical variables (bed, bath, acre lot, house size) as predictors. In the next iteration (model 2), the categorical variable ‘city’ was included to predict price.

Result:

	Train Set Evaluation (model 1)	*Test Set Evaluation (model 1)*
R^2	0.0511	0.0994
MAE	261,298.72	260,061.61
RMSE	768,056.53	538,713,02



	Train Set Evaluation (model 2)	*Test Set Evaluation (model 2)*
R^2	0.0595	0.050
MAE	260,220.55	262,166.82
RMSE	713,468.99	781,960.86



Interpretation:

The first regression model shows low accuracy in both test and train sets, indicating that the model is not capturing much variability in property price, suggesting numerical variables alone may not be sufficient to explain the variation. So, we fit the second regression model, adding the city as a predictor along with the numerical predictors. Doing so, the findings weren't much different from the first model in the context of good accuracy and price prediction. The addition of the city increased the accuracy of the train set, suggesting that the city has some explanatory power for the target variable. However, the accuracy of the test set decreased significantly. Overall, both the models have very low accuracy, and other metrics (MAE &RMSE) support the low accuracy. There appears to be no strong relationship between available predictors and target variable, implying that the linear regression model does not explain much of the variance in property price and its features.

MARKET STATUS ANALYSIS

Objective:

Predict property statuses, such as "for sale" or "sold," using features like the number of bedrooms, bathrooms, house size, and lot area. Additionally, additionally we address class imbalance using the Synthetic Minority Oversampling Technique (SMOTE) and evaluate the model's performance through various metrics.

Features Selected: bed, bath, acre lot, house size.

Feature Engineering: Numerical variables were standardized, and categorical variables were encoded using LabelEncoder, resulting in a new column, status_encoded.

Logistic Regression:

The dataset was split into training (80%) and testing (20%) subsets, maintaining class balance through stratified sampling. Logistic Regression model with balanced class weights was trained using the resampled data.

Results:

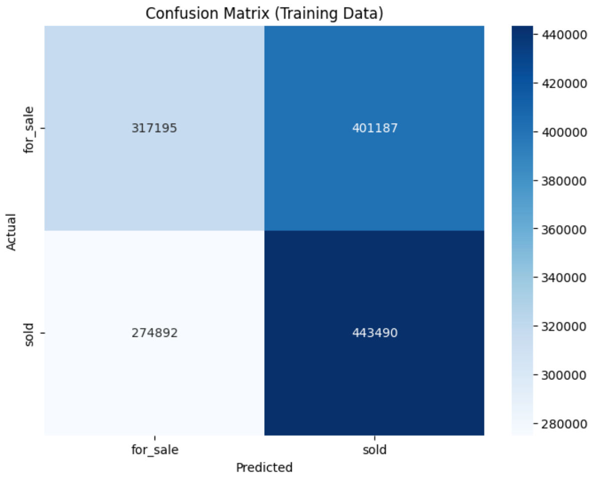
The training accuracy is 52.94% and test accuracy is 51.99%.

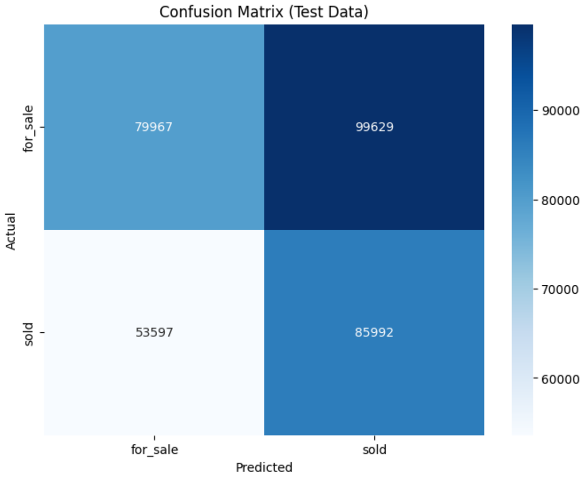
Classification Metrics

Class	Precision	Recall	F1-Score
For Sale	60%	45%	51%
Sold	46%	62%	53%

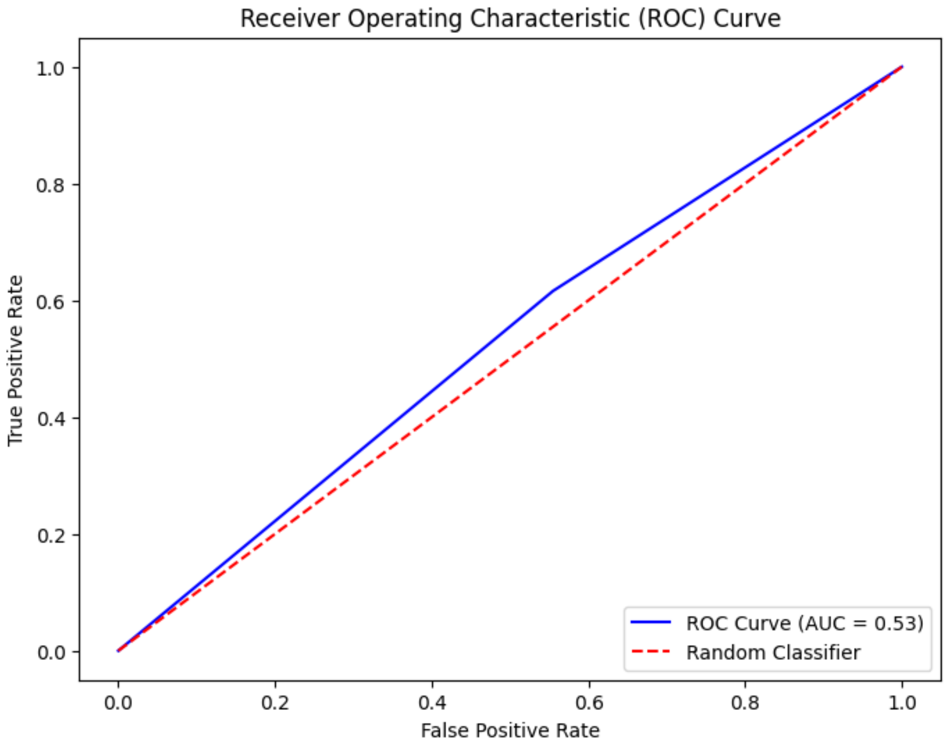
The model demonstrated balanced performance between "for sale" and "sold" categories, with higher recall for the "sold" class, indicating better identification of sold properties.

Confusion Matrices





ROC Curve and AUC



The Receiver Operating Characteristic curve achieved an AUC of 0.53, indicating room for improvement in distinguishing between classes.

Interpretation:

This analysis highlights the application of logistic regression in classifying real estate property statuses. While the model demonstrates balanced performance, the moderate accuracy and AUC values suggest that further refinements are necessary.

MARKET SEGMENTATION

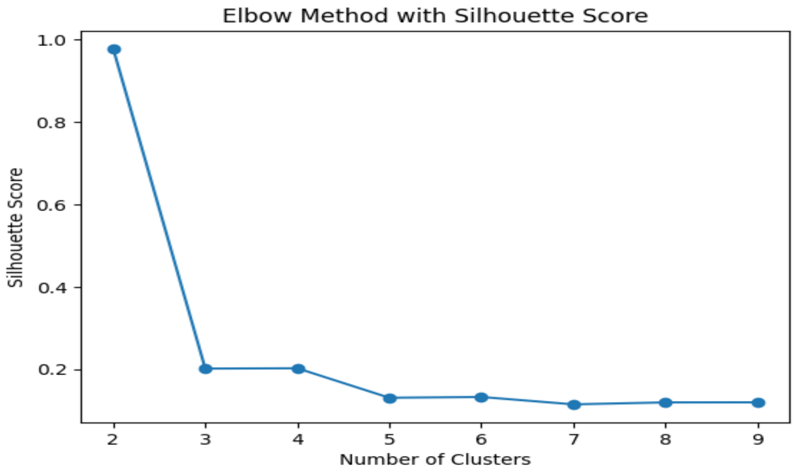
Objective:

Market segmentation analysis involves dividing realtor data into distinct groups based on shared characteristics. The analysis focuses on clustering the data to categorize properties into meaningful segments.

Data Sampling: A fixed number of samples (1000) were selected per state using stratified random sampling to ensure a balanced subset of properties across states for cluster analysis.

Feature Engineering: Numerical variables were standardized, and categorical variables were encoded using OneHotEncoder.

K-Means Clustering:



The elbow method used silhouette scores to identify the optimal number of clusters. The range of clusters tested was from 2 to 9. The optimal number of clusters was determined to be 3, where the score stabilized. K-Means clustering was applied with 3 clusters, resulting in the segmentation of properties into distinct categories based on the centroid values of the clusters. The clusters were labeled as Affordable, Mid-Range, and High-End, reflecting the pricing and features associated with each group. Properties were assigned to one of the three categories, ensuring meaningful differentiation. Centroid characteristics guided the labeling of clusters.

Composite Scoring:

A weighted composite score was calculated for each cluster centroid to rank the clusters. Weights were assigned based on feature relevance:

Feature	Weight
Price	0.40
House Size	0.30
Bed	0.10
Bath	0.10
Acre Lot	0.05
State Features	0.05
City Features	0.05

Result:

Clusters were ranked and labeled based on their composite scores as Affordable, Mid-Range, and High-End. The sorted centroids provided insights into the distinguishing characteristics of each segment.

Clustering Performance:

The silhouette score for the final clustering was calculated to be **0.20.** While clustering captured distinct patterns, modest silhouette scores suggest overlaps between some clusters. Further refinement could improve segmentation clarity.

Interpretation:

This market segmentation analysis categorized properties into three meaningful segments: Affordable, Mid-Range, and High-End. While the clustering process provided insights into property differentiation, the silhouette score of 0.20 indicates that the clusters are not well-defined and have overlapping boundaries.

Despite the modest clustering quality, this segmentation lays a foundational framework for further refinement. Future steps could include exploring alternative clustering algorithms, tuning weights for feature contributions, or incorporating additional features to improve cluster separability.

BROKER ANALYSIS

Objective:

This analysis aims to identify the highest possible price for a property. However, the challenge lies in determining which broker is most capable of achieving that price. To address this, a clustering analysis was conducted on real estate broker data to predict which broker is most likely to secure the highest sale price for a property.

Clustering Approach:

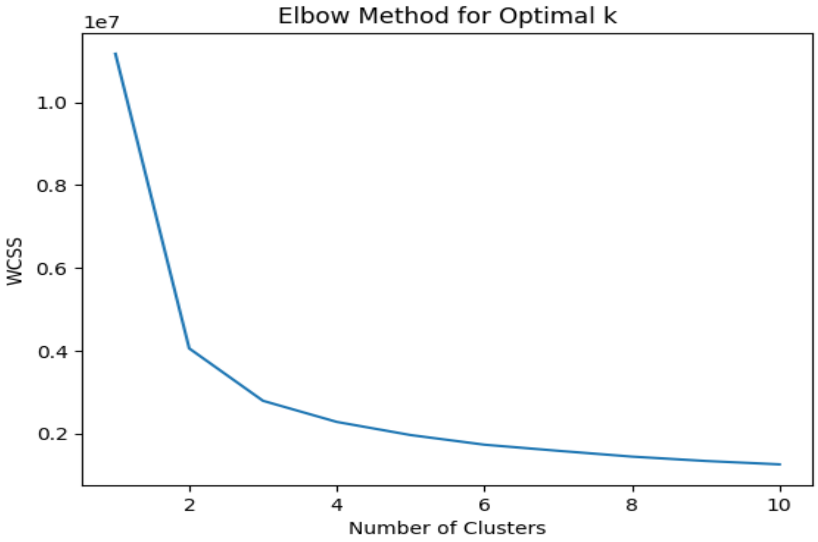
Clustering was used to group similar properties together to identify patterns and understand relationships among different features. This method helps in identifying broker performance based on property characteristics and potential sale prices.

Features Selected: brokered_by, price, bed, bath, acre lot, house size, city, state

Feature Engineering: The categorical variables such as city, state and broker_id were transformed into numerical values using Label Encoding. And the rest of the data was standardized to a similar scale.

K-Means Clustering:

K-Means clustering was used for this analysis. The optimal number of clusters was determined using the Elbow Method, which plots the Within-Cluster Sum of Squares (WCSS) for various values of k, ranging from 1 to 100 in steps of 10. Based on the plot, the number of clusters (n_clusters) was selected as 40. Therefore, K-Means clustering groups the data into 40 clusters, minimizing the variance within each cluster.



Model Evaluation:

The Davies-Bouldin Index was used to evaluate the model, yielding a value of 1.249. This index measures the average similarity ratio of each cluster with the cluster that is most similar to it, where a lower value indicates better clustering performance.

Interpretation:

The clustering analysis provides valuable insights for predicting which brokers are most likely to sell high-priced properties by understanding the characteristics of each cluster. Through cluster profiling, brokers can gain actionable insights into pricing and marketing strategies, allowing them to tailor their approach to maximize sales potential. Clustering helps brokers refine their strategies, optimize their sales processes, and increase the likelihood of successful transactions.

LIMITATIONS

Some limitations of this analysis are the absence of key social and economic factors that significantly influence the real estate market. The lack of features such as inflation, income levels, GDP per capita, and neighborhood characteristics restricts the applicability and accuracy of the analysis. For instance, property prices are often influenced by neighborhood amenities, such as proximity to parks, grocery stores, and other essential facilities, which are not included in the dataset. Additionally, the model's predictive capabilities are confined to the current dataset and do not account for the dynamic growth of the real estate market, including the increasing number of brokers and properties. As a result, the model's predictions are limited to the data available and do not consider broader market trends or changes over time.

CONCLUSION

Predicting property prices proved challenging and resulted in low accuracy. The features provided in the dataset were insufficient to accurately predict property prices, which limited the effectiveness of the price prediction model in this analysis. The market segmentation process divided properties into three categories: Affordable, Mid-Range, and High-End. While this segmentation provided valuable insights, a silhouette score of 0.20 indicates room for improvement. Clustering analysis was effective in optimizing marketing strategies and improving broker performance. By grouping properties into clusters, we identified brokers most likely to sell a property at a higher price. This allows for more targeted and tailored sales strategies, benefiting brokers and real estate businesses by improving the likelihood of successful transactions.

REFERENCES

Abelson, M., Kacmar, M., & Jackofsky, E. (1990). Factors Influencing Real Estate Brokerage Sales Staff Performance. *Journal of Real Estate Research*, *5*(2), 265–275. https://doi.org/10.1080/10835547.1990.12090621

Zumpano, L.V., Elder, H.W. & Baryla, E.A. Buying a house and the decision to use a real estate broker. *J Real Estate Finan Econ* **13**, 169–181 (1996). https://doi.org/10.1007/BF00154054

Yalgudkar, S. S., & Dharwadkar, N. V. (2022). A Literature Survey on Housing Price Prediction. *Journal of Computer Science & Computational Mathematics*, *12*(3), 41-45.

Usman, H., Lizam, M., & Adekunle, M. U. (2020). Property price modelling, market segmentation and submarket classifications: A review. *Real Estate Management and Valuation, 28*(3), 24–35. https://doi.org/10.1515/remav-2020-0021

Muller, P., Chepeleva, K., & Shemeleva, Zh. (2020). The segmentation of the residential real estate market on the affordability criterion: Methodical aspect. *Advances in Economics, Business and Management Research*, https://doi.org/10.2991/aebmr.k.200312.163

Sanyal, S., Biswas, S. K., Das, D., Chakraborty, M., & Purkayastha, B. (2022). Boston house price prediction using regression models. In *Proceedings of the 2022 2nd International Conference on Intelligent Technologies (CONIT)* (pp. 1-6). IEEE. https://doi.org/10.1109/CONIT55038.2022.9848309