Unlocking the Market: Predictive Modeling of Housing Prices (Final Report)

MLTeam

View on GitHub

Unlocking the Market: Predictive Modeling of Housing Prices (Final Report)

Introduction/Background

Accurate prediction of housing prices is essential for guiding investment decisions and real estate market analysis, especially given the dynamic nature of property values [1]. Machine learning models have increasingly been applied to improve forecasting accuracy by leveraging large datasets that include property characteristics such as square footage, number of bedrooms, and year built (Pai & Wang, 2020) [2]. Our study utilizes the Kaggle Housing Prices Dataset, which contains 13 features, to explore machine learning approaches for housing price estimation, aiming to enhance the accuracy and efficiency of real estate valuations.

Problem Definition

At the end of 2011, the real estate class was valued at 25 trillion dollars, where more than 50% was residential properties [3]. Because of the vast investments in real estate, as access to shelter is a basic necessity, the housing market, with its fluctuations, is volatile and affects the entire economy [3]. The objective of this study is to leverage ML to estimate housing prices accurately, empowering each stakeholder to make informed decisions. We believe that ML has the potential to help stabilize the economy during the housing crisis as well as provide access to predictions for decision making for all individuals.

Methods

The first step in our methodology is data preprocessing. We attempted to first check for missing data with the isNull command, but our dataset was already clean. If there were missing data, we'd address it by filling it in since we don't want to lose any vital data. Next, we identified our categorical features: [‘mainroad’, ‘guestroom’, ‘basement’, ‘hotwaterheating’, ‘airconditioning’, ‘prefarea’, ‘furnishingstatus’]. We then converted it with one-hot encoding using the get_dummies function from the pandas library. It converted the current data into binary variables, 0 and 1. This way, wach catefory has a unique column that represents its presence. Next, to maintain uniformity across the dataset, we applied a StandardScaler from the sklearn preprocessing library to our numerical features: [‘price’, ‘area’, ‘bedrooms’, ‘bathrooms’, ‘stories’, ‘parking’]. This helps normalize the data and ensures that all features contribute equally to the model's performance. The mean and standard deviation for our dataset is now 0 and 1, respectively. StandardScaler is important to continue with our process. For k-means clustering, scaling is important for measuring the Euclidean distances, so each feature has the same effect.

For exploratory data analysis, we employ K-Means clustering, an unsupervised learning technique. This method groups houses based on key features. By identifying distinct patterns or clusters within the dataset, K-Means provides valuable insights into housing market segmentation, which can inform strategic decision-making and investment strategies.

To predict housing prices, we utilize Linear Regression, a supervised learning model. This model leverages key features such as area and number of bedrooms to estimate property values. We incorporate the Variance Inflation Factor (VIF) to detect and manage multicollinearity by removing highly correlated variables, thereby improving model accuracy.

In addition to Linear Regression, we implement a Random Forest Regressor, another supervised learning model. This decision tree-based approach utilizes all available features. Unlike Linear Regression, Random Forest is adept at handling nonlinear interactions between features, making it a robust choice for complex datasets with intricate relationships among variables.

Results and Discussions

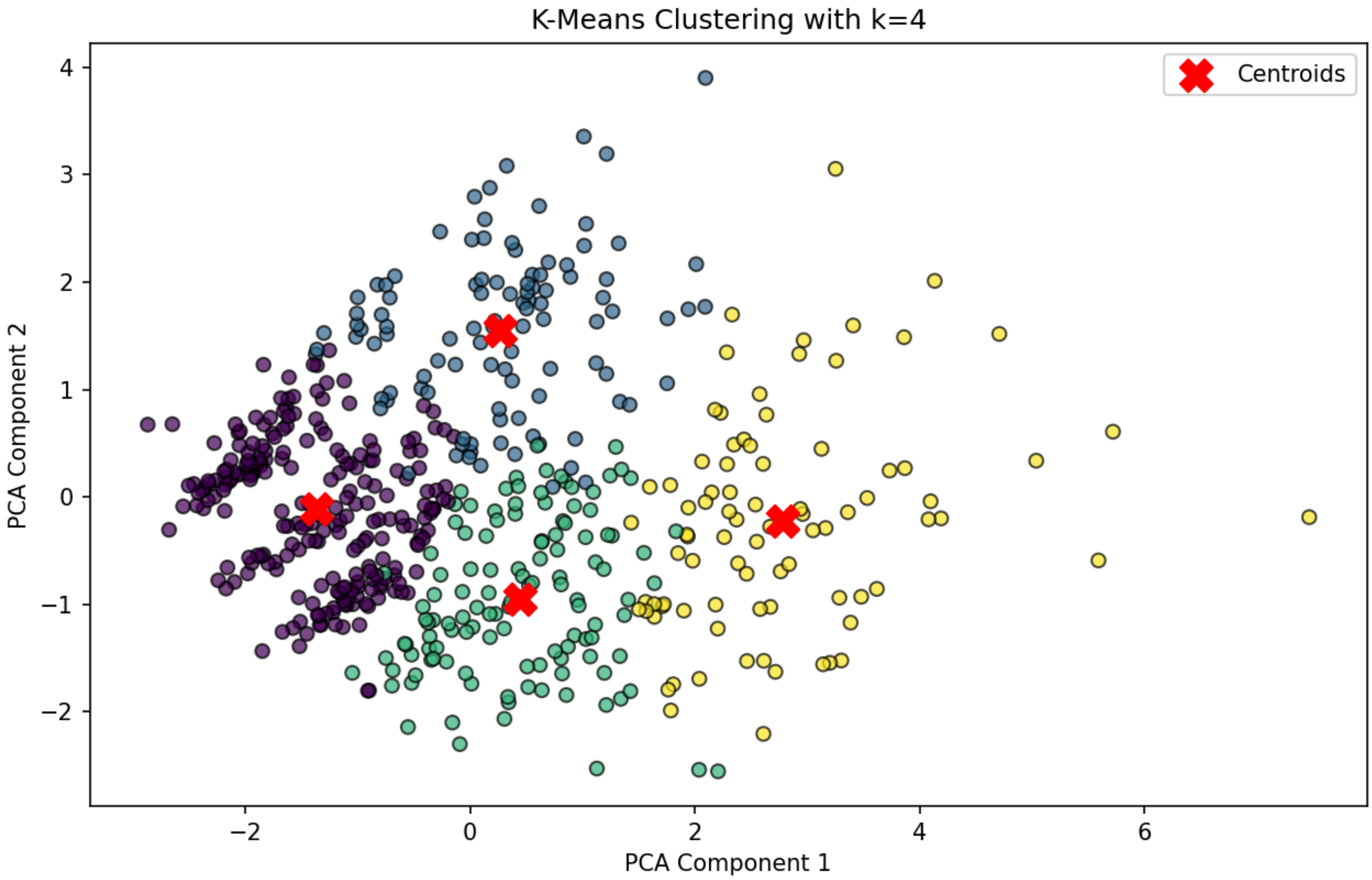
Overview

The analysis aims to explore patterns in housing prices and associated attributes using K-Means clustering and supervised learning models like Linear Regression and Gradient Boosting Regressor. Preprocessing and clustering were performed in a single script (`prices.py`), and backward feature selection was applied to reduce dimensionality. Additionally, K-Means clustering grouped similar houses based on features, while the Elbow Method and silhouette score were used for evaluating cluster quality. To predict housing prices, Linear Regression and Gradient Boosting were implemented, and their performances were evaluated quantitatively and visually.

Visualizations

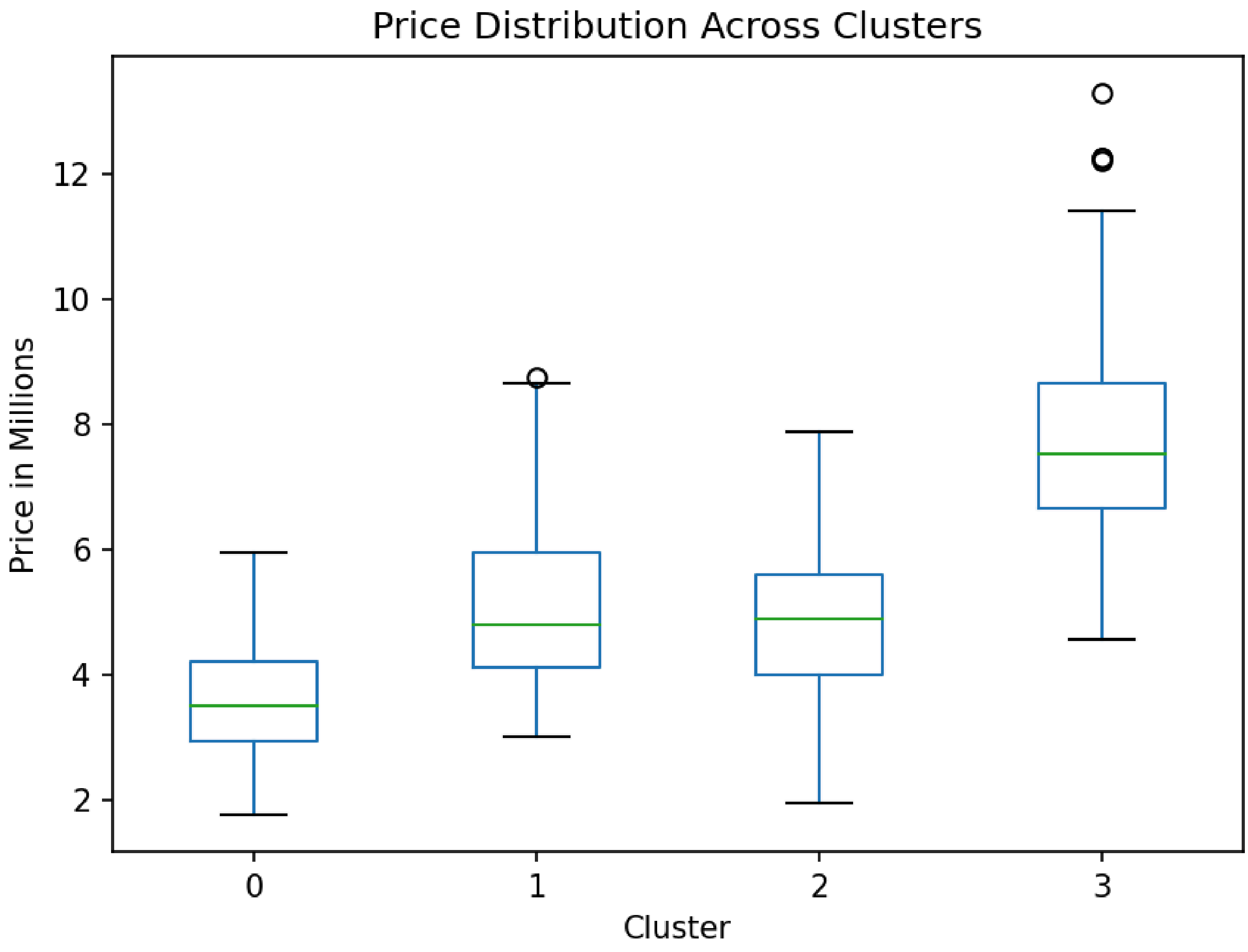
PCA Visualization of Clusters

To visualize the distribution of clusters, we applied PCA to reduce the data to two components. This plot below provides a view of how well-separated the clusters are in two-dimensional space.



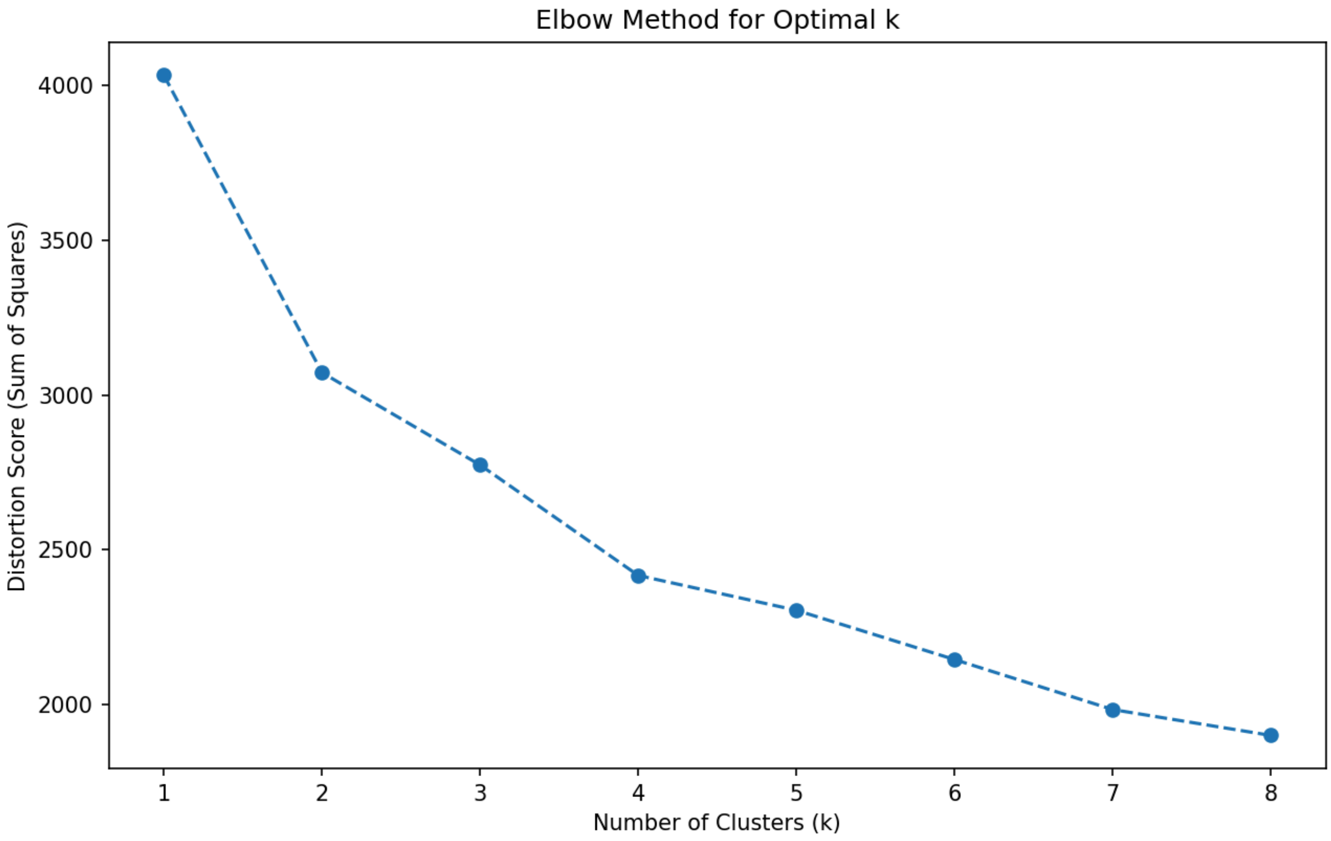
The PCA plot illustrates that each cluster is reasonably well-separated, with centroids marked as red crosses. This separation implies that our clustering model has successfully divided the data into distinct groups. However, some overlap is visible between clusters, suggesting potential areas for further feature refinement or adjustments to cluster parameters.

1. Price Distribution Across Clusters



The boxplot indicates that each cluster corresponds to a distinct price range. This separation implies that the clustering model effectively grouped properties by price, with cluster 3 capturing the highest-priced homes, and cluster 0 representing the lower-priced homes. Outliers in cluster 3 suggest some high-priced properties with unique features, potentially affecting cluster boundaries.

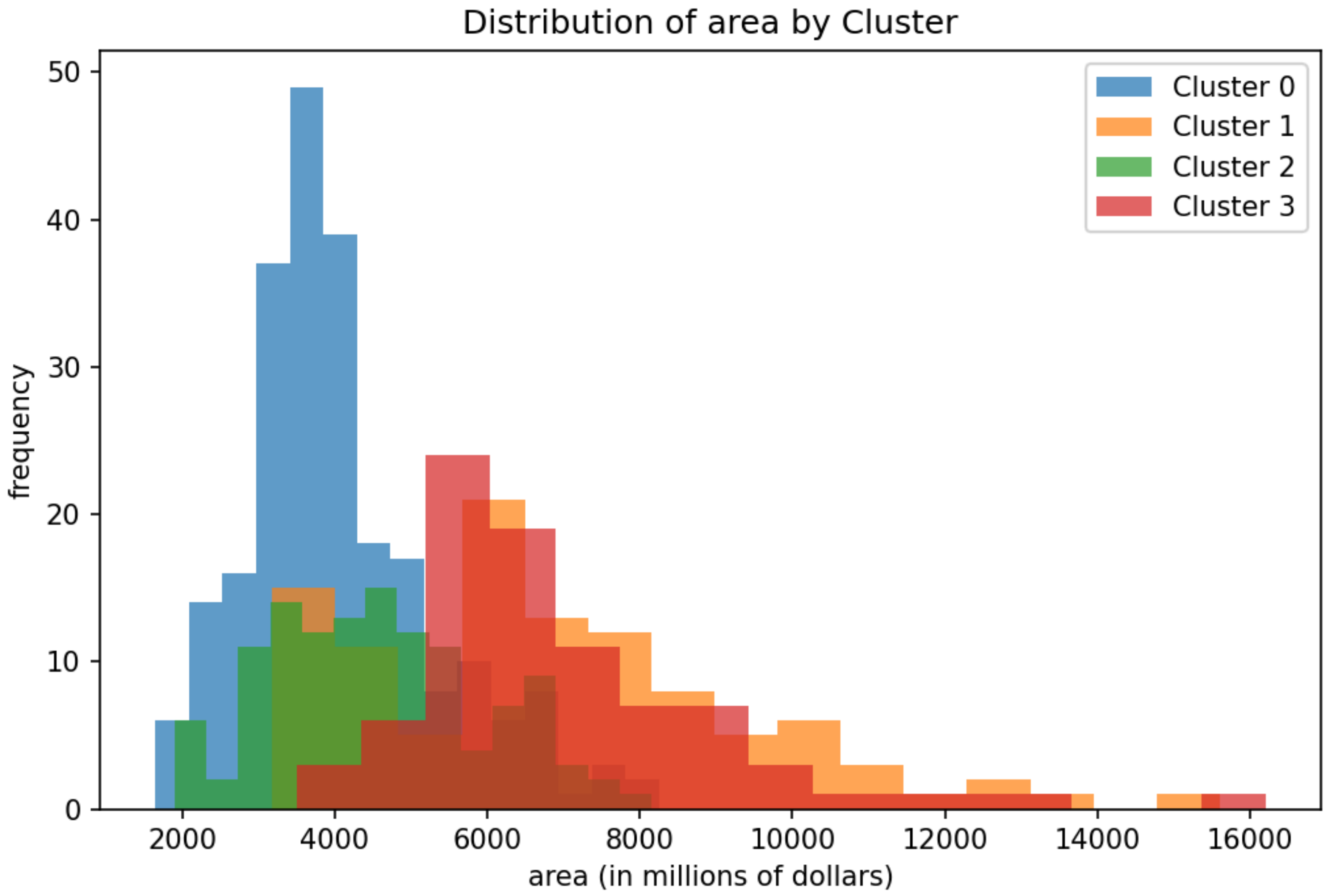
2. Elbow Method for Optimal K



The Elbow Method graph, plotting distortion scores (sum of squares), shows a visible “elbow” around (k = 4), indicating an optimal cluster count. The model performs well with four clusters, as further increases in (k) only marginally improve clustering quality.

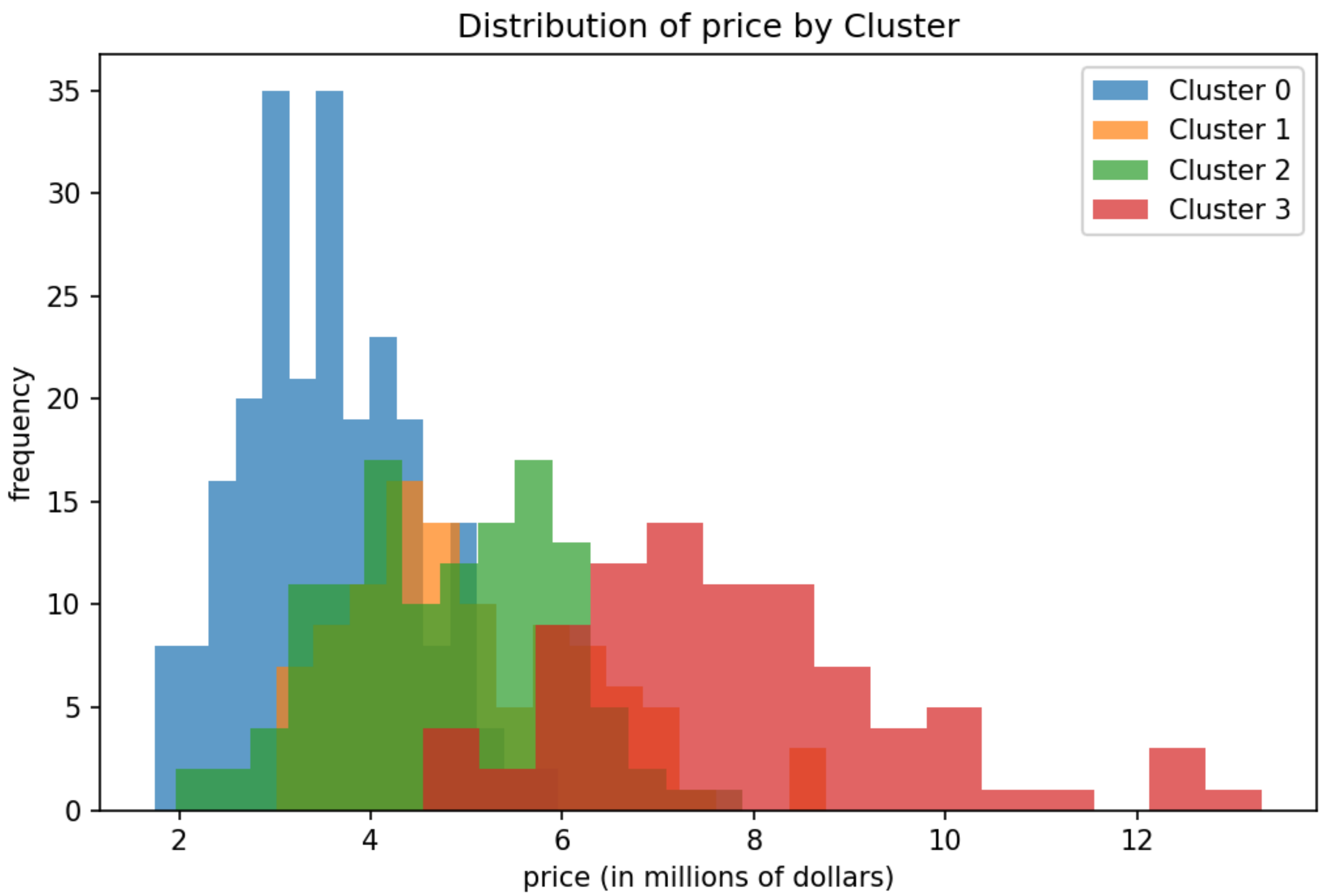
3. Histogram Analysis by Cluster

Area by Cluster



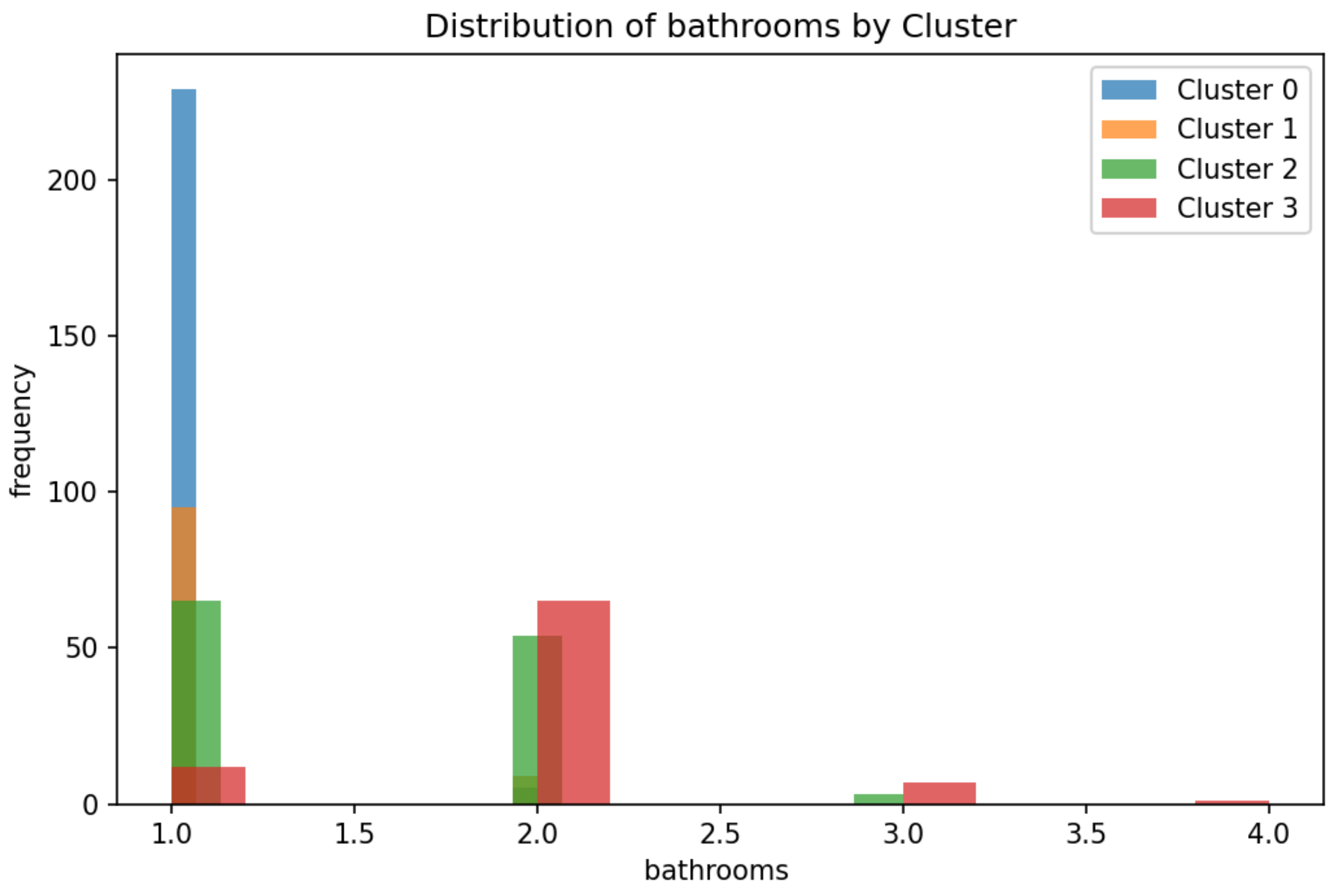
Cluster 0 is mostly associated with smaller areas, while cluster 3 includes larger areas. This grouping confirms that area and price are positively correlated, as clusters with larger areas correspond to higher-priced homes.

Price by Cluster



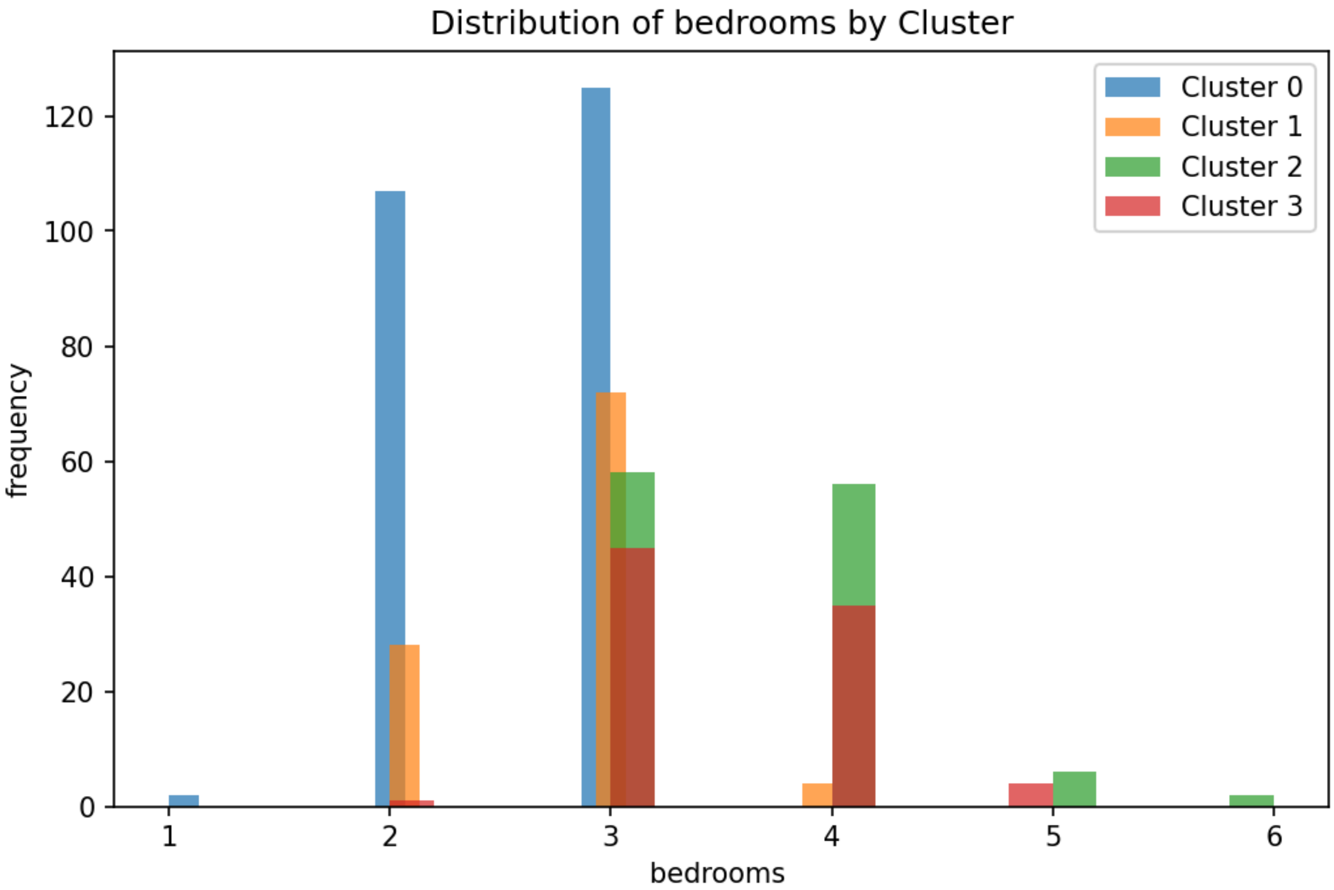
The histogram for price distribution further supports the boxplot findings, showing a clear division among clusters in price ranges. This segmentation indicates effective grouping by price in the clustering model.

Bathrooms by Cluster



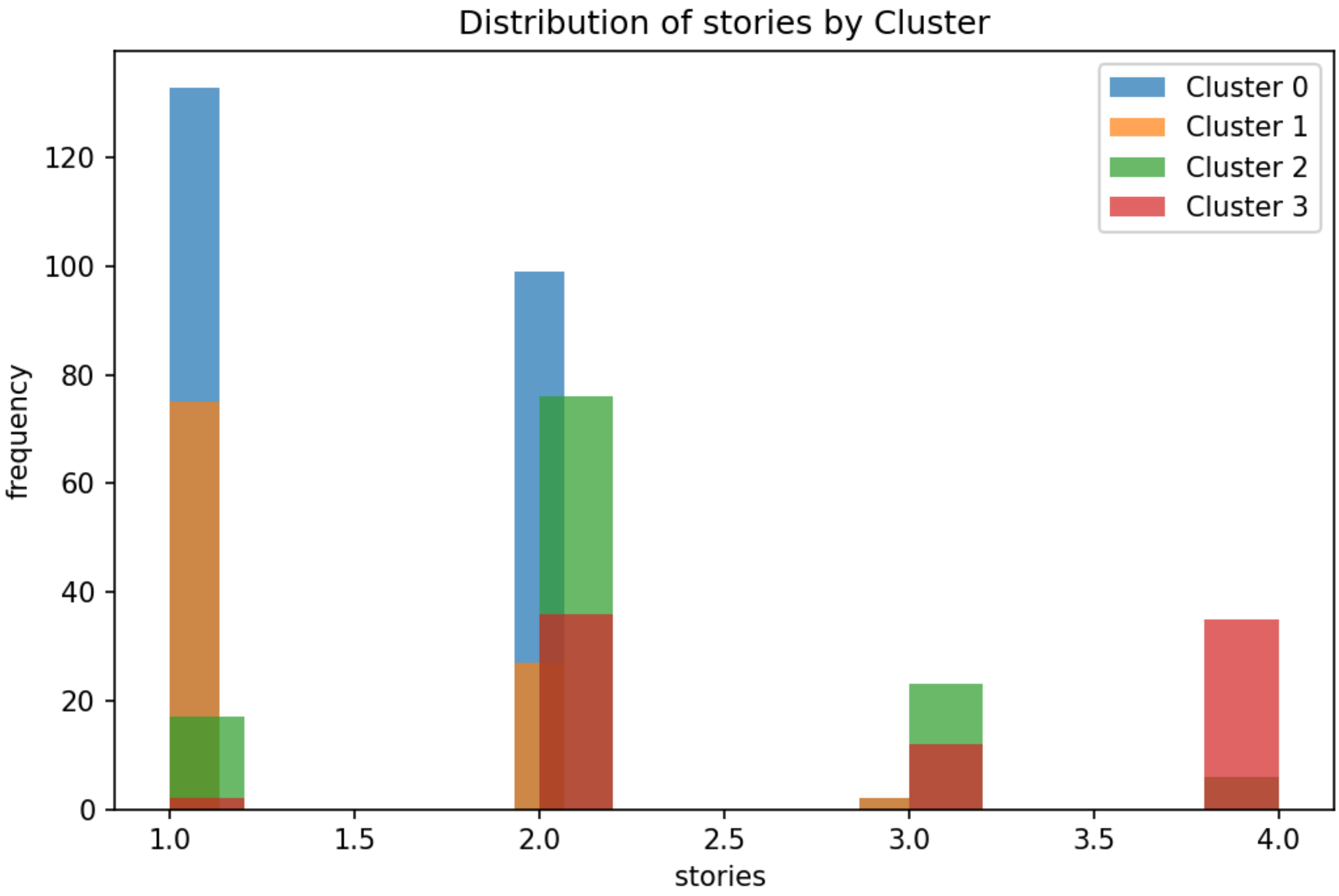
Bathroom counts vary across clusters, with higher counts appearing more frequently in higher-priced clusters. However, the correlation is weaker than for area and price. This may suggest that while bathroom count is a contributing factor, it’s less influential on clustering compared to area and price.

Bedrooms by Cluster



Bedrooms exhibit some overlap among clusters, though lower bedroom counts tend to associate with lower-priced clusters. This pattern, while less distinct than area or price, still provides some separation by bedroom count, implying a weak correlation between bedroom count and pricing tiers.

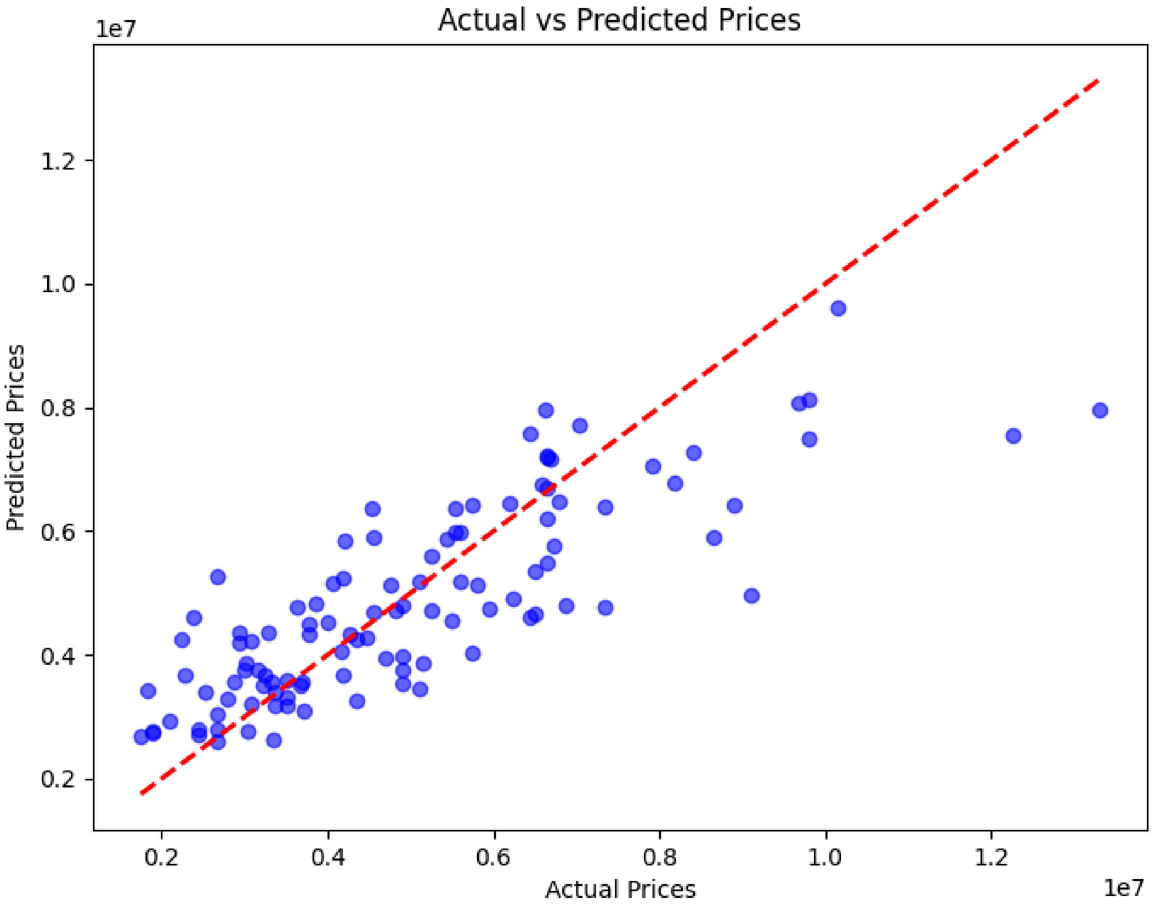
Stories by Cluster



The distribution of property stories across clusters shows that clusters with higher numbers of stories are associated with higher-priced properties, though the overlap suggests that stories alone are not a strong determinant of cluster placement.

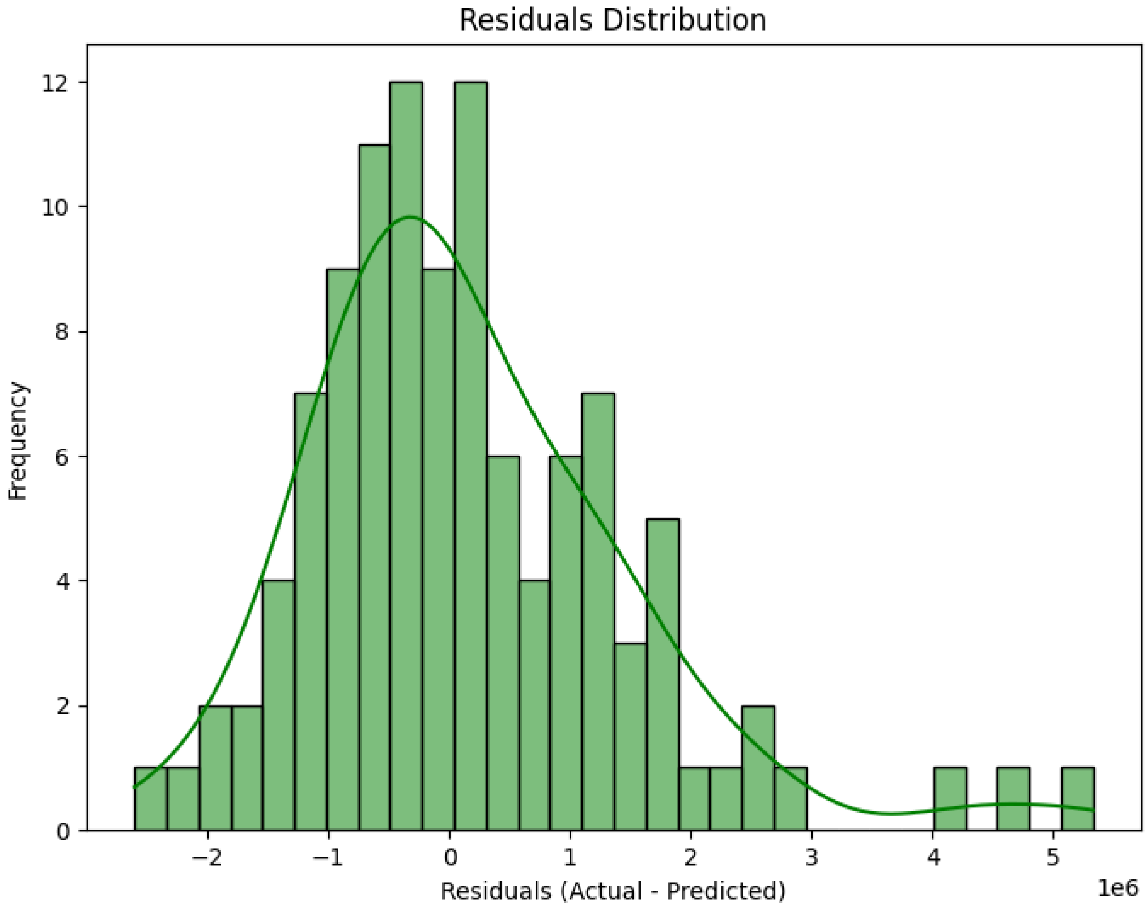
Linear Regression

Actual vs Predicted Prices



The scatter plot of actual vs. predicted prices shows that most points are clustered along the diagonal red line, indicating good agreement between predicted and true values. However, some visible dispersion, particularly at higher price ranges, suggests that the model struggles with higher-priced homes.

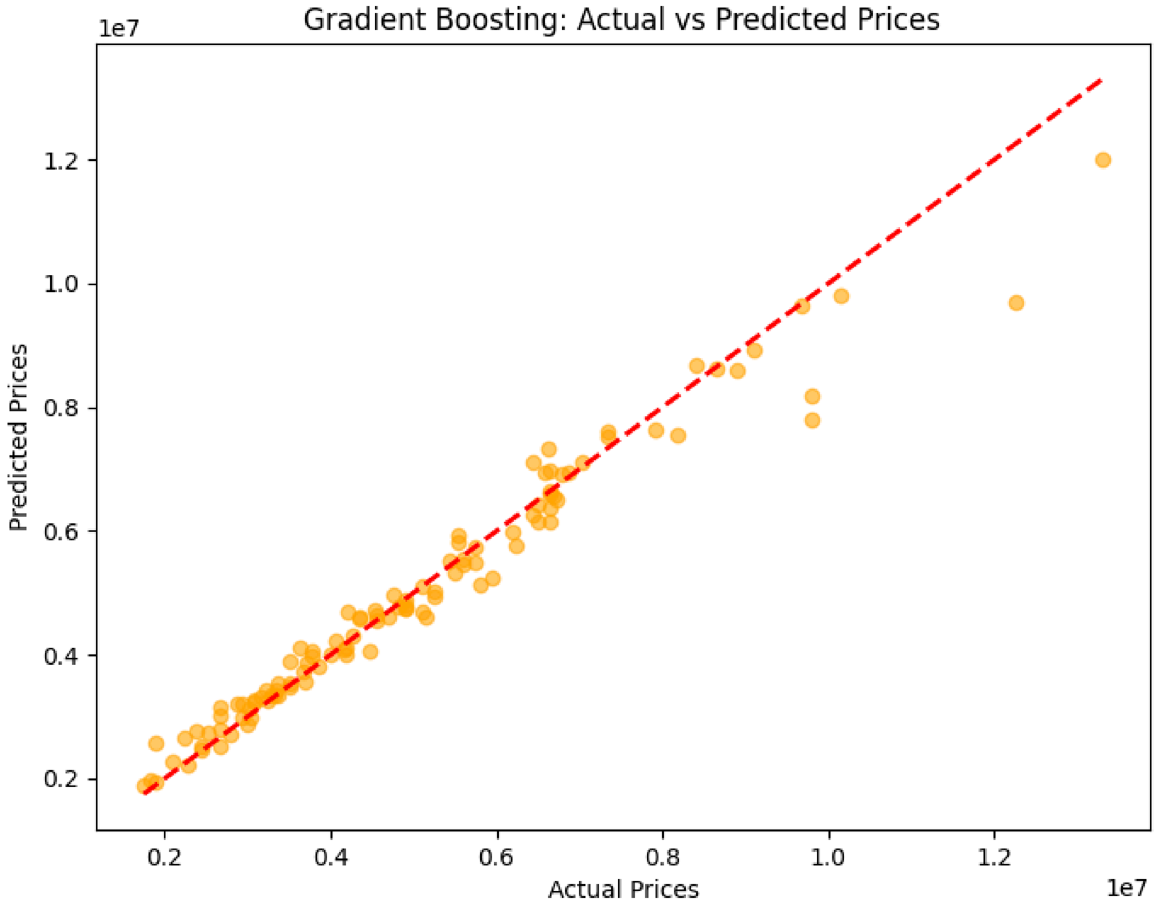
Residuals Distribution



The residuals plot shows a roughly symmetric distribution centered around zero, indicating that the Linear Regression model is unbiased. However, the slight skew and long tail on the right suggest underestimation of some higher-priced homes.

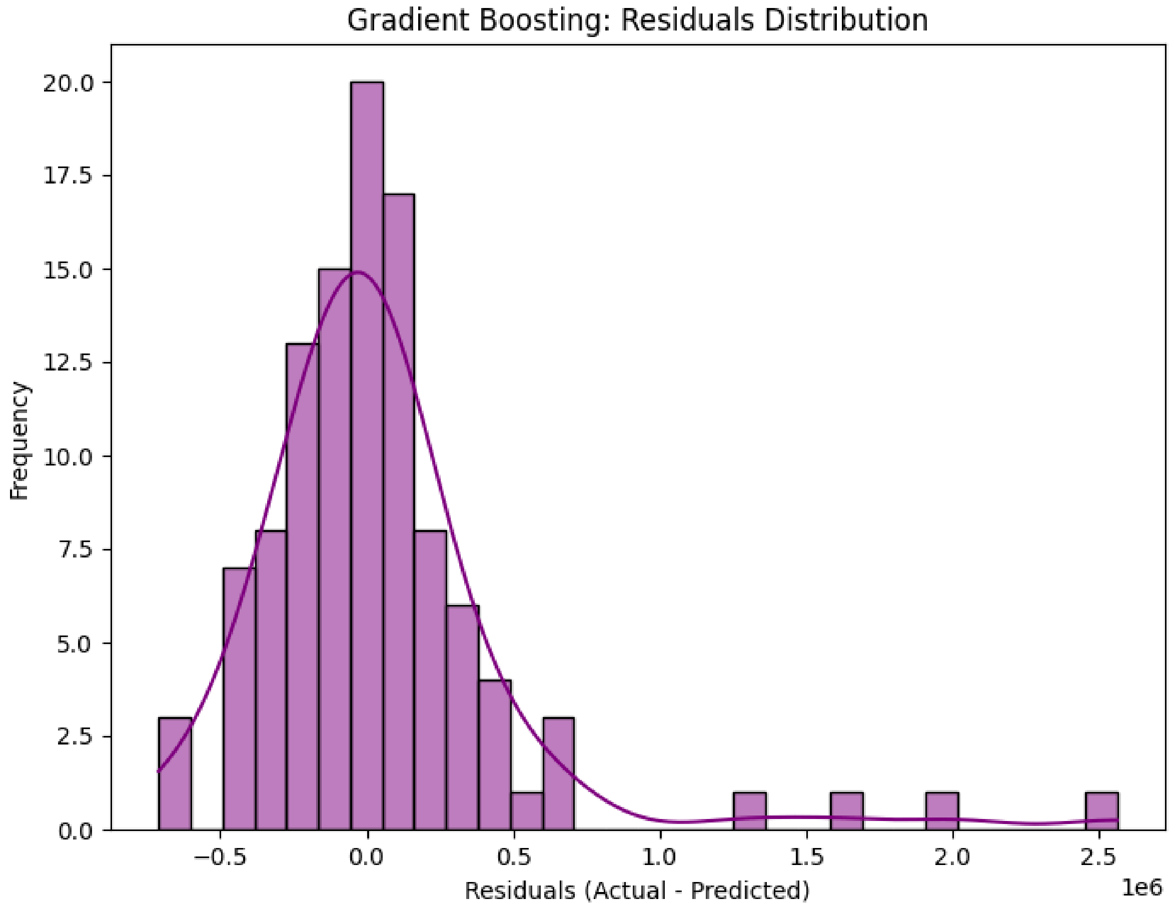
Gradient Boosting Regressor

Actual vs Predicted Prices



The scatter plot of actual vs. predicted prices for Gradient Boosting shows that the points are tightly clustered along the diagonal red line, indicating excellent agreement between predicted and actual values. Compared to Linear Regression, Gradient Boosting demonstrates much less dispersion, particularly at higher price ranges, which highlights its ability to capture more complex relationships.

Residuals Distribution



The residuals plot for Gradient Boosting shows a symmetric distribution centered around zero, indicating that the model is unbiased. The tighter spread of residuals compared to Linear Regression suggests a significant improvement in predictive accuracy.

Quantitative Metrics

- **Elbow Method**: The optimal (k) value is determined to be 4, aligning with the number of clusters visualized.
- **Silhouette Score**: A silhouette score of 0.5 indicates moderately good clustering performance. This score reflects a balance, where clusters are reasonably well-defined but might benefit from further refinement or feature adjustments.
- **Linear Regression Metrics**:
  - **Mean Absolute Error (MAE)**: The MAE of 970,043.40 indicates that, on average, the model's predictions deviate from the actual housing prices by approximately 970,043. While this value provides a useful baseline for accuracy, it is significantly higher than the acceptable range of 5-10% of the average price (250,376.83 to 500,753.67). This indicates that the model may not be sufficiently accurate for practical use in predicting housing prices.
  - **Root Mean Squared Error (RMSE)**: The RMSE of 1,324,506.96 reflects the square root of the mean squared prediction error, giving more weight to larger errors. Compared to the threshold of half the standard deviation of the actual prices (1,129,310.83), the RMSE exceeds this limit, further highlighting the model's challenges in capturing the variability of housing prices accurately.
  - **R^2 Score**: The R^2 score of 0.6529 indicates that the Linear Regression model explains 65.29% of the variance in housing prices. While this suggests that the model captures a substantial portion of the price

variability, approximately 34.71% of the variability remains unexplained. This implies that there may be nonlinear relationships or additional interactions between features that the model fails to capture.

- **Gradient Boosting Regressor Metrics**:
    - **Mean Absolute Error (MAE)**: 271,180.71
        - This value falls within the acceptable range of 5-10% of the average price 250,376.83 to 500,753.67.
        - It reflects a significant improvement over Linear Regression's MAE 970,043.40.
    - **Root Mean Squared Error (RMSE)**: 457,385.07
        - This value is well below half the standard deviation of the actual prices 1,129,310.83, indicating excellent predictive performance.
        - Compared to Linear Regression's RMSE 1,324,506.96, Gradient Boosting drastically reduces errors.
    - **R^2 Score**: 0.9586
        - This score indicates that Gradient Boosting explains 95.86% of the variance in housing prices, compared to Linear Regression's 65.29\%. It demonstrates the model's ability to capture the complex relationships between features and housing prices.

Analysis of Each Model

The clustering model successfully grouped properties based on price, area, and related attributes. The strongest factors influencing clustering were area and price, with these features aligning well across clusters. Features like bathroom count, bedrooms, and stories showed weaker correlations with clusters, suggesting they are secondary in determining property groupings. Overall, the model performs well in segmenting properties by price tiers, though some outliers suggest that further refinement could improve cluster definition.

The performance metrics of the Linear Regression model indicate that, while it provides a baseline understanding of housing prices, its errors (both MAE and RMSE) exceed acceptable thresholds. Additionally, the R^2 score highlights that a significant portion of price variability remains unexplained, which could result from:

- **Nonlinear relationships** between features (e.g., interaction effects between area, bedrooms, and bathrooms).
- **High variance** in certain price ranges, as suggested by the residuals distribution.
- **Outliers** in the dataset, which can disproportionately affect Linear Regression. Given these limitations, **Gradient Boosting Regressor** is a logical next step. Gradient Boosting outperformed both Linear Regression and Random Forest across all evaluation metrics:
- MAE and RMSE both fell within the acceptable thresholds.
- The R^2 score showed a significant improvement, indicating that Gradient Boosting captures nearly all the variance in housing prices.

Compared to Linear Regression, Gradient Boosting effectively handles nonlinear interactions between features, as evidenced by the tighter clustering of predicted vs. actual prices and the more symmetric residuals distribution. Additionally, its robust performance highlights its potential for practical use in housing price predictions.

Changes from initial proposal

We initially proposed using Random Forest for housing price prediction. However, during testing, Random Forest performed worse than Linear Regression, with higher RMSE and lower R^2 values. This was attributed to the limited dataset size and Random Forest's susceptibility to overfitting when trained on small datasets. Consequently, we transitioned to Gradient Boosting, which uses a sequential learning approach to correct prediction errors iteratively. This change resulted in significant improvements across all evaluation metrics.

Next Steps

1. **Feature Importance Analysis**: Use feature importance scores to identify key contributors to the model's predictions and refine the feature set further.
2. **Advanced Models**: Explore more advanced boosting techniques like XGBoost or LightGBM to see if further performance gains can be achieved.

Contribution Table

Name	Contribution
Scott	Gradient Boosting Regressor Implementation, Linear Regression Implementation, Preprocessing, Results & Discussions, Methods, GitHub Pages, README.md
Erik	Linear Regression Implementation, Methods, Results & Discussions
My Duyen Nguyen	Gradient Boosting Regressor Implementation, Preprocessing, Methods
Sanjana	K-Means Algorithm Implementation, Methods, Video Script Linear Regression
Kruthik	K-Means Algorithm Implementation, Methods, Video Script Gradient Boosting

Gantt Chart

- Gantt Chart
- Download the Gantt Chart Excel Sheet

References

[1] C. Ma, "The Investigation and Prediction of Influencing Factors of Australian Housing Price Based on Machine Learning Models," Semantic Scholar, Sep. 2024. [Online]. Available: https://www.semanticscholar.org/paper/The-Investigation-and-Prediction-of-Influencing-of-Ma/77d50f4073bf17848214c21c7e7821574a67783c

[2] P.-F. Pai and W.-C. Wang, "Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices," Applied Sciences, vol. 10, no. 17, p. 5832, Aug. 2020, doi: https://doi.org/10.3390/app10175832.

[3] E. Ghysels, A. Plazzi, R. Valkanov, and W. Torous, "Chapter 9 - Forecasting Real Estate Prices," Science Direct, https://www.sciencedirect.com/science/article/pii/B9780444536839000098 (accessed 2024).

mlteam is maintained by scottwatanuki.
This page was generated by GitHub Pages.