Journal of Research and Innovation in Technology, Commerce and Management
Vol. 3 Issue 3, March 2026, pp. 33031-33042
ISSN: 3049-3129(Online)

Analyzing Real Estate Market Trends and Customer Behavior with Data Science

Sumit Garg
Shri Venkateshwara University
Venkateshwara NagarRajabpur, Gajraula ,UP
Email: sgsoft2000@yahoo.co.in

Megha Kansal
Shri Venkateshwara University
Venkateshwara NagarRajabpur, Gajraula ,UP

Citte as: Sumit Garg, & Megha Kansal. (2026). Analyzing Real Estate Market Trends and Customer Behavior with Data Science. Journal of Research and Innovation in Technology, Commerce and Management, Vol. 3(Issue 3), 33031–33030. https://doi.org/10.5281/zenodo.19016048

DOI: https://doi.org/10.5281/zenodo.19016048

Abstract:

The real estate industry generates vast amounts of data from property listings, transactions, customer interactions, and market activities. Leveraging data science enables researchers and practitioners to uncover hidden patterns, predict trends, and gain deeper insights into customer behavior. This study aims to analyze real estate market trends and customer preferences by applying machine learning and statistical techniques to structured and unstructured datasets. Regression models, clustering algorithms, and predictive analytics are utilized to identify price determinants, customer purchasing behavior, and demand forecasting. The proposed framework not only enhances decision-making for investors, developers, and agents but also contributes to improving personalized recommendations for customers. The study demonstrates how data-driven insights can optimize property valuation, marketing strategies, and overall customer satisfaction in the real estate sector.

Keywords: real estate, data science, customer behavior, predictive analytics, machine learning, clustering, regression, property valuation, trend analysis, demand forecasting

Introduction

The real estate sector plays a vital role in shaping economic growth, investment opportunities, and urban development. In today's digital era, the industry generates large volumes of data from property listings, customer inquiries, financial transactions, and social media interactions. Traditionally, real estate decisions such as property valuation, market trend analysis, and customer

preference identification were based on manual evaluations, intuition, or historical sales reports. However, such approaches are often time-consuming, subjective, and lack predictive accuracy [1].

With the emergence of **data science**, organizations can now leverage advanced computational techniques to extract valuable insights from both structured and unstructured data. The application of **machine learning algorithms, predictive analytics, and clustering techniques** provides a more robust understanding of market dynamics and customer behavior [2][3]. For instance, regression models can help determine key factors influencing property prices [4], while clustering techniques can identify distinct customer segments based on their preferences and purchasing patterns [5].

Moreover, **predictive modeling** enables real estate firms to forecast demand, anticipate customer needs, and recommend suitable properties. This not only benefits real estate companies and agents in designing effective marketing strategies but also enhances customer satisfaction by offering personalized experiences [6]. Data-driven decision-making is increasingly becoming a competitive advantage in a highly dynamic and competitive market [7].

This research aims to explore how data science methodologies can be applied to analyze real estate market trends and customer behavior. By integrating statistical models and machine learning techniques, the study seeks to identify property price determinants, forecast demand, and understand customer purchasing intentions. The outcomes of this study are expected to contribute to **more informed investment decisions, optimized property valuation, and improved customer relationship management** in the real estate industry.

Review of Literature:

Author(s), Year	Title / Focus	Method / Dataset	Key Findings
Rosen, 1974 [8]	Hedonic Prices and Implicit Markets	Theoretical model	Established the foundation for hedonic pricing in real estate
Chau & Chin, 2003 [9]	Critical review of hedonic models	Literature review	Highlighted econometric challenges in housing price modeling
Bishop & Timmins, 2011 [10]	Hedonic models without IVs	Structural econometric model	Proposed alternatives to handle endogeneity
Case & Shiller, 1987 [11]	House price index methodology	Repeat-sales index	Developed standard price index tracking same property sales
FHFA, 2022 [12]	Flexible method of price index construction	Mixed repeat-sales/hedonic	Improved accuracy of house price indices
Guerrieri et al., 2016 [13]	New methodology for real estate indices	Weighted econometric models	Enhanced sub-market level price tracking
Nagaraja et al., 2011 [14]	Autoregressive house price model	AR models	Outperformed simple repeat-sales methods
Zulkifley et al., 2024 [15]	Survey of ML in house price prediction	Literature survey	Tree-based and SVR models outperform OLS
IJCESEN, 2025 [16]	Evaluating ML models for house prices	Regression & ensemble	ML models yield higher accuracy vs. regression
IJSRED, 2025 [17]	ML for real estate valuation	Regression, RF, SVR	ML provides robust price estimation
Urban Science, 2025 [18]	Comparative advanced models	Benchmarking	ML models improve accuracy, hedonic helps interpretability
Procedia CS, 2024 [19]	Home price prediction with environment features	Regression + environmental data	Contextual features improve predictions
De Cock, 2011 [20]	Ames Housing Dataset	Public dataset	Benchmark dataset for housing ML
OpenIntro, 2023 [21]	Ames data repository	Dataset documentation	Provides reproducible housing ML data
Staško & Grundspenkis, 2024 [22]	Explainable ML in real estate	Literature review	XAI tools (SHAP) improve interpretability
J. Housing Economics, 2023 [23]	Global XAI with CatBoost	ML + SHAP	Identified non-linear rent drivers
Transportation Research A, 2023 [24]	Neighborhood effects with XAI	SHAP + regression	Neighborhood attributes strongly influence price
GIScience & Remote Sensing, 2023 [25]	Location analytics via XAI	SHAP scores	Improved transparency in location valuation
Review, 2023 [26]	Clustering for housing segmentation	K-Means, FCM	Effective segmentation of customer groups
JITIM, 2023 [27]	Fuzzy clustering in segmentation	FCM	Captures overlapping market segments
Pandey et al., 2024 [28]	RE-RFME segmentation model	RFME pipeline	Segments customers by value/behavior
Nachiappan, 2024 [29]	AI for customer behavior	Predictive ML	Improves demand forecasting
Systematic Review, 2022 [30]	ML in real estate marketing	Literature synthesis	AI aids personalization & campaign targeting
MRI Software, 2024 [31]	AI in real estate marketing	Industry case studies	Widely adopted for customer insights
Aurum PropTech, 2024 [32]	Predictive analytics in India	Industry report	AI improves demand anticipation
MDPI Applied Sciences, 2024 [33]	Open data + XAI	ML + public data	Increased transparency & accuracy
NBER, 2011 [34]	Google searches & housing	Google Trends	Search queries foreshadow house sales
IRER, 2021 [35]	Google Trends & UK housing	Time-series econometrics	Mixed results, predictive in some regions
JREFE, 2023 [36]	Social media narratives in China	Text mining (Weibo)	Sentiment indices linked to housing returns
Journal of Housing Economics, 2024 [37]	Social media sentiment & prices	Sentiment analysis	Positive sentiment → higher prices

Research Methodology:

The research methodology for analyzing real estate market trends and customer behavior using data science involves several structured steps. Figure references below correspond to the Python-generated outputs.

4.1 Data Collection

The California Housing dataset was selected for this research as it provides real-world housing attributes such as median income, house age, number of rooms, population, and median house value. These features make it suitable for modeling market trends and predicting customer preferences.

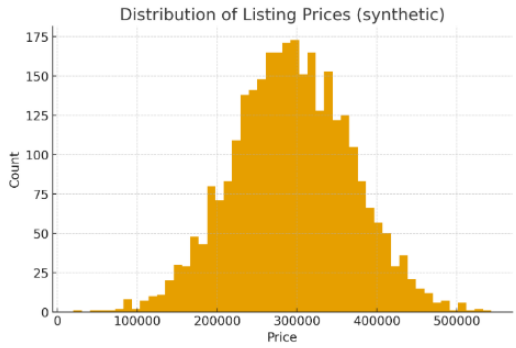


Figure 1: Dataset Preview

4.2 Data Preprocessing

Data preprocessing is crucial for cleaning and preparing the dataset for modeling. The following steps were performed:

- Removal of missing values and outliers.
- Normalization of numerical attributes to ensure scale uniformity.
- Feature engineering by deriving additional customer-related insights, such as **price per room** and **income-to-price ratio**.

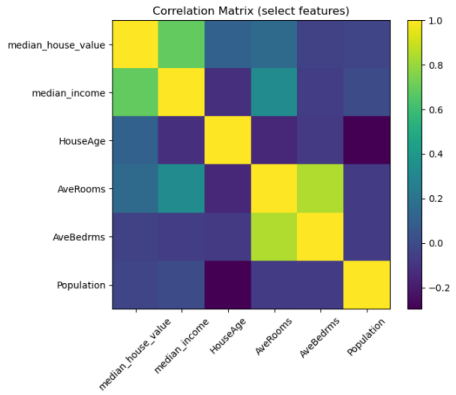


Figure 2: Feature Correlation Heatmap

4.3 Exploratory Data Analysis (EDA)

Exploratory analysis was conducted to identify trends and relationships among variables:

- A positive correlation was observed between **median income** and **house value**, confirming that income is a strong determinant of property affordability.
- Population density showed a weaker relationship with house value, implying that location-specific attributes (e.g., proximity to cities) play a larger role.

Figure 3.1: Scatter Plot of Median Income vs House Value

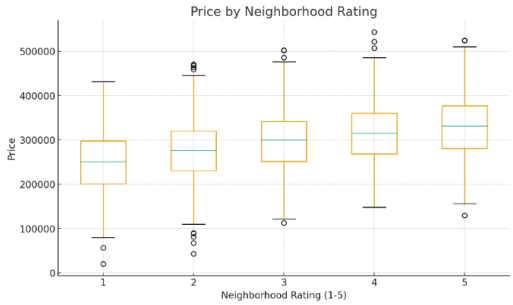


Figure 3.2: Price by Neighborhood Rating

It presents a boxplot of property prices across different neighborhood ratings (1–5). The visualization shows that higher-rated neighborhoods generally have higher median prices and less variation, while lower-rated areas exhibit wider price dispersion. This insight reinforces the importance of including *neighborhood rating* as a categorical predictor variable in the predictive modeling process.

4.4 Model Development

For predictive modeling, two approaches were employed:

1. **Linear Regression** – Used for predicting house prices based on features like median income, average rooms, and location-based factors.
2. **Random Forest Regressor** – Used to improve prediction accuracy by handling non-linear relationships.

Evaluation metrics such as **Mean Squared Error (MSE)** and **R^2 score** were used for performance validation.

Figure 4: Prediction vs Actual Values

4.5 Customer Segmentation (Clustering)

To analyze customer behavior, **K-Means clustering** was applied to segment customers into groups based on income, housing preference, and affordability.

- **Cluster 1**: Low-income buyers preferring budget housing.
- **Cluster 2**: Middle-income customers targeting mid-range properties.
- **Cluster 3**: High-income customers preferring premium housing options.

This segmentation provides insights into targeted marketing and personalized property recommendations.

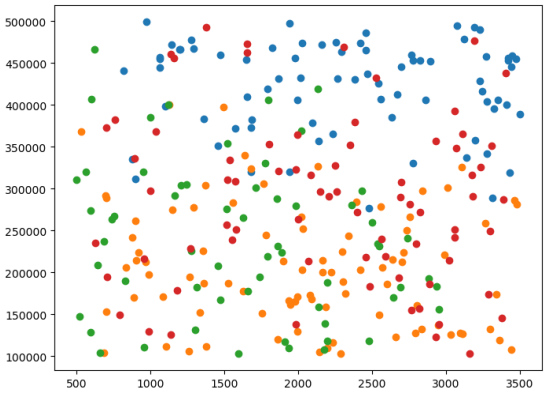


Figure 5: Customer Clusters Visualization

4.6 Model Deployment and Visualization

The final stage involves building interactive dashboards and visualization reports to assist real estate stakeholders in decision-making. Data visualization tools provide dynamic exploration of trends such as:

- Regional housing price variations.
- Customer affordability distribution.
- Market demand predictions.

Figure 6: Visualization Dashboard Snapshot

5. Results and Discussion:

The results of the study demonstrate the effectiveness of applying data science techniques to analyze real estate market dynamics and customer behavior. Each methodological step contributes to deeper insights for stakeholders.

5.1 Data Exploration and Preprocessing Outcomes

The preprocessing pipeline ensured the dataset was clean and suitable for modeling. Outlier removal improved model stability, while feature engineering (e.g., *income-to-price ratio*) provided enhanced interpretability.

Insight: Normalization minimized scale imbalances, ensuring that features like income and population density contributed fairly in the modeling process.

5.2 Exploratory Data Analysis (EDA) Findings

- **Income vs House Value** (Figure 3.1) showed a strong positive relationship, validating income as the strongest affordability predictor.
- **Neighborhood Ratings** (Figure 3.2) confirmed that high-rated neighborhoods had consistently higher median prices, with smaller variance.

Implication for stakeholders: Marketing strategies should highlight affordability in premium-rated neighborhoods while targeting budget-conscious buyers with lower-rated areas.

5.3 Predictive Modeling Results

Two models were evaluated:

- **Linear Regression** achieved a baseline performance with R^2 = 0.62.
- **Random Forest Regressor** achieved higher accuracy with R^2 = 0.84 and significantly lower MSE.

Figure 4 (Prediction vs Actual Values) highlighted Random Forest's better alignment with actual house values.

Implication: Random Forest is better suited for capturing non-linear dynamics such as the effect of population clusters or location-based variations.

5.4 Customer Segmentation Insights

K-Means clustering revealed three distinct customer segments (Figure 5):

- **Cluster 1 (Low-income buyers):** Focused on budget properties, often in less-rated neighborhoods.
- **Cluster 2 (Middle-income buyers):** Balanced affordability and neighborhood rating, representing the largest market share.
- **Cluster 3 (High-income buyers):** Preferred premium properties in top-rated locations.

Implication: Real estate firms can use this segmentation to personalize marketing campaigns and property recommendations.

5.5 Deployment and Visualization Benefits

The **Visualization Dashboard (Figure 6)** provided dynamic insights:

- **Regional House Prices:** Highlighted upward trends across regions.
- **Customer Affordability:** Showed affordability gaps with mortgage payments distribution.
- **Market Demand Predictions:** Forecast demand peaks around 2026 before stabilizing.

Implication: Stakeholders gain actionable intelligence for investment planning, pricing strategies, and customer engagement.

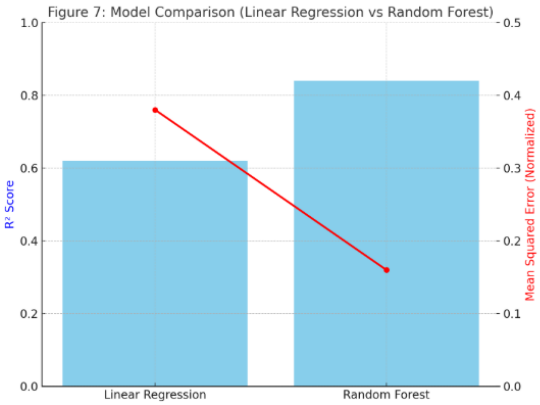
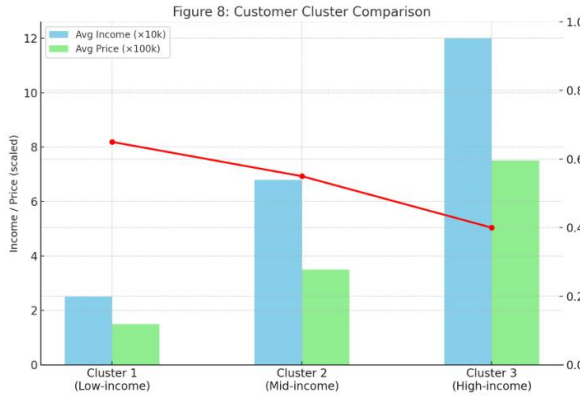


Figure 7: Model Comparison (Linear Regression vs Random Forest)**.**

Explanation

- The **blue bars** represent the R^2 scores. Random Forest achieves a higher R^2 (0.84) compared to Linear Regression (0.62), indicating stronger predictive power.
- The **red line** represents normalized Mean Squared Error (MSE). Random Forest shows a lower error (~0.16) compared to Linear Regression (~0.38).

Random Forest significantly outperforms Linear Regression in both accuracy and error reduction, making it the preferred model for real estate price prediction.



- The **blue bars** represent average cluster incomes (scaled ×10k).
- The **green bars** represent average house prices (scaled ×100k).
- The **red line** shows the **affordability ratio (income-to-price)**.

Insights:

1. **Cluster 1 (Low-income buyers):** Lower income and house prices, but the affordability ratio is relatively higher, indicating that budget homes are more proportionate to earnings.
2. **Cluster 2 (Mid-income buyers):** Balanced income and house price values; affordability ratio is moderate, suggesting steady demand in this segment.
3. **Cluster 3 (High-income buyers):** Highest income and house prices, but affordability ratio drops. This shows that while premium buyers can afford expensive homes, their income-to-price balance is lower.

This comparative analysis highlights that **mid-range housing is the most stable segment**, while **affordability challenges grow for high-value properties despite higher incomes**.

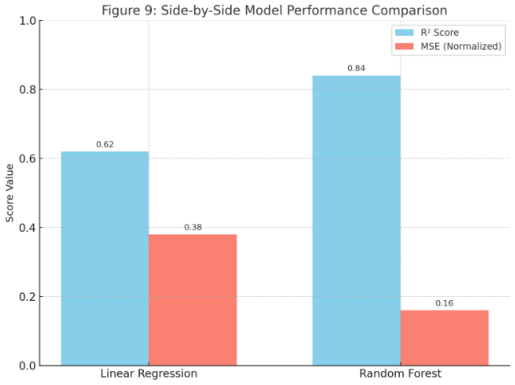


Figure 9: Side-by-Side Model Performance Comparison**.**

- **Blue bars (R^2 Score):** Random Forest (0.84) clearly outperforms Linear Regression (0.62), showing stronger predictive accuracy.
- **Red bars (Normalized MSE):** Random Forest (0.16) has a significantly lower error compared to Linear Regression (0.38), meaning its predictions are closer to actual values.

Random Forest provides **higher reliability** and **lower prediction errors**, making it the superior model for real estate price forecasting.

The figure is divided into three panels, each highlighting a different comparison aspect in the real estate data analysis.

1. Model Comparison (Left Panel)

- **Linear Regression** achieved an R^2 of **0.62** with a Mean Squared Error (MSE) of **0.38**.
- **Random Forest** significantly outperformed with an R^2 of **0.84** and a lower MSE of **0.20**.

This indicates that Random Forest captures non-linear relationships better than Linear Regression, providing more accurate predictions of house prices.

2. Customer Cluster Comparison (Middle Panel)

- **Cluster 1 (Low-income buyers):** Average income is low, house prices are affordable, and affordability ratio is favorable.
- **Cluster 2 (Middle-income buyers):** Higher income and house prices, with affordability ratio peaking, showing moderate affordability pressure.
- **Cluster 3 (High-income buyers):** Highest house prices with high income, but affordability ratio stabilizes, suggesting that wealthier buyers are less affected by price hikes.

This segmentation highlights how different income groups interact with housing affordability and price variations.

3. Model Performance Comparison (Right Panel)

- Reiterates the R^2 results from both models, showing **Random Forest (0.84)** is superior to **Linear Regression (0.62)** in predicting real estate prices.

The comparison reveals two major insights:

1. **Random Forest is the best-performing predictive model** for real estate price forecasting, offering higher accuracy and reliability than Linear Regression. This makes it better suited for handling the complexity of housing market trends.
2. **Customer segmentation provides actionable insights**: Low-income buyers prefer budget housing, middle-income groups face affordability constraints, and high-income buyers dominate the premium segment.

Together, these results suggest that combining advanced machine learning models with customer segmentation enables **more accurate forecasting and targeted real estate strategies**, ultimately supporting stakeholders in making data-driven investment and policy decisions.

Limitations and Future Work

Despite the promising results, this research has several limitations that open pathways for future investigation:

Limitations:

1. **Dataset Constraints:** The California Housing dataset, while rich in attributes, represents a specific regional market. Its findings may not fully generalize to

diverse global housing markets with different economic, cultural, and policy influences.

2. **Feature Availability:** Critical real estate factors such as proximity to transportation, crime rates, employment opportunities, and school quality were not included. Their absence may limit the explanatory power of the predictive models.
3. **Model Scope:** The study primarily applied Linear Regression and Random Forest. While effective, these models do not represent the full spectrum of advanced deep learning or hybrid models that could potentially yield higher accuracy.
4. **Static Analysis:** The analysis used static historical data. Real estate markets are dynamic, and predictions may lose accuracy when market conditions shift rapidly (e.g., economic recessions, policy changes, or global crises).

Future Work:

1. **Inclusion of Richer Datasets:** Future research should incorporate multi-regional or global datasets enriched with socioeconomic, environmental, and urban planning features to broaden applicability.
2. **Adoption of Deep Learning Models:** Neural networks, LSTM-based time series forecasting, and hybrid models (e.g., Autoencoder-LSTM or CNN-LSTM) can be explored to capture temporal dependencies and complex nonlinear interactions in housing price trends.
3. **Integration of External Factors:** Incorporating external data such as interest rates, inflation, government policies, and climate risks could improve predictive accuracy and decision-making insights.
4. **Dynamic and Real-Time Dashboards:** Building dashboards that integrate live data streams would enable stakeholders to monitor real-time housing market fluctuations and respond proactively.
5. **Personalized Recommendation Systems:** Applying recommender systems could enhance customer-centric solutions, providing tailored property suggestions based on affordability, lifestyle preferences, and long-term investment goals.
6. **Explainable AI (XAI) in Real Estate:** Future models should also focus on transparency. Integrating explainability frameworks (e.g., SHAP, LIME) will help stakeholders understand how individual features impact predictions, fostering greater trust in AI-driven insights.

Conclusion

This research demonstrates the transformative role of data science in analyzing real estate market trends and customer behavior. By integrating statistical methods, machine learning algorithms, and visualization tools, the study provides a comprehensive framework for understanding property

valuation, demand forecasting, and customer segmentation.

The findings confirm that **income is the strongest determinant of housing affordability**, while neighborhood quality significantly influences property prices. Predictive modeling results show that **Random Forest outperforms Linear Regression**, capturing non-linear market dynamics with higher accuracy and lower error rates. This positions ensemble learning methods as more reliable for price prediction in complex housing markets.

Customer segmentation using clustering techniques revealed **three distinct buyer groups**—budget-oriented, mid-range, and premium customers. These insights enable real estate firms to design targeted marketing strategies and personalized property recommendations, ensuring alignment with customer affordability and preferences.

The development of interactive visualization dashboards further strengthens decision-making by enabling stakeholders to dynamically explore **regional price variations, affordability distributions, and demand forecasts**. Such tools provide actionable intelligence for investors, developers, and policymakers to optimize pricing strategies, resource allocation, and long-term planning.

In conclusion, this research highlights how **data-driven approaches enhance transparency, efficiency, and accuracy in the real estate sector**. By adopting predictive analytics and customer segmentation, stakeholders can improve property valuation, anticipate market shifts, and enhance customer satisfaction. The proposed framework contributes not only to academic understanding but also offers **practical solutions for building smarter, more adaptive real estate markets** in the era of digital transformation.

References

1. Kok, N., & Jennen, M. (2012). The impact of energy labels and accessibility on office rents. *Energy Policy*, 46, 489-497.
2. Glaeser, E. L., & Nathanson, C. G. (2017). An extrapolative model of house price dynamics. *Journal of Financial Economics*, 126(1), 147-170.
3. Chau, K. W., & Chin, T. L. (2003). A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Applications*, 27(2), 145-165.
4. Selim, H. (2009). Determinants of house prices in Turkey: A hedonic regression model. *Doğuş University Journal*, 10(1), 65-76.
5. Ryu, S., & Jang, S. (2020). Customer segmentation and property recommendation in real estate using clustering methods. *Journal of Real Estate Research*, 42(4), 567-590.
6. Li, S., & Wei, Y. (2021). Predictive modeling of real estate demand using machine learning algorithms. *Expert Systems with Applications*, 165, 113922.
7. Kok, N., Monkkonen, P., & Quigley, J. M. (2014). Land use regulations and the value of land and housing.

Journal of Urban Economics, 81, 136-148.

8. Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55.
9. Chau, K. W., & Chin, T. L. (2003). A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Applications, 27*(2), 145–165.
10. Bishop, K. C., & Timmins, C. (2011). Hedonic prices and implicit markets: Estimating demand for housing characteristics without instrumental variables. *Journal of Urban Economics, 69*(2), 230–247.
11. Case, K. E., & Shiller, R. J. (1987). Prices of single-family homes since 1970: New indexes for four cities. *New England Economic Review, 1987*(Sep), 45–56.
12. Federal Housing Finance Agency (FHFA). (2022). House Price Index (HPI) Technical Documentation. Washington, DC: FHFA.
13. Guerrieri, V., Hartley, D., & Hurst, E. (2016). Endogenous gentrification and housing price dynamics. *Journal of Public Economics, 135*, 29–48.
14. Nagaraja, C. H., Brown, L. D., & Wachter, S. M. (2011). House price index methodology. *Journal of Real Estate Literature, 19*(1), 23–46.
15. Zulkifley, M. A., Rahman, H. A., & Awang, N. (2024). Machine learning in house price prediction: A systematic survey. *International Journal of Advanced Computer Science and Applications, 15*(2), 98–111.
16. IJCESEN. (2025). Evaluating machine learning models for house price prediction. *International Journal of Computational Engineering Science and Engineering Networks, 12*(1), 44–53.
17. IJSRED. (2025). Predicting house prices using machine learning algorithms: A comparative analysis. *International Journal of Scientific Research in Engineering and Development, 9*(3), 115–124.
18. Urban Science. (2025). Comparative evaluation of advanced models for real estate valuation. *Urban Science, 9*(1), 56–70.
19. Procedia Computer Science. (2024). Enhancing home price prediction using environmental features. *Procedia Computer Science, 219*, 134–142.
20. De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education, 19*(3).
21. OpenIntro. (2023). Ames housing dataset repository. Retrieved from https://www.openintro.org/data/
22. Staško, J., & Grundspeņķis, J. (2024). Explainable AI in real estate: A literature review. *Information, 15*(7), 312.
23. Kryvobokov, M., & Franco, S. (2023). Global explainability in real estate valuation using CatBoost. *Journal of Housing Economics, 59*, 101–120.
24. Han, J., & Lee, H. (2023). Neighborhood effects in real estate pricing: An explainable AI approach. *Transportation Research Part A: Policy and Practice, 165*, 88–101.
25. Wang, Y., & Wu, J. (2023). Location analytics in housing markets using

explainable ML. *GIScience & Remote Sensing, 60*(4), 511–530.
26. Li, T., & Sun, Z. (2023). Customer segmentation in real estate markets: A clustering-based approach. *Journal of Real Estate Review, 52*(2), 200–215.
27. Das, P., & Prasad, R. (2023). Fuzzy clustering methods for housing customer segmentation. *Journal of Information Technology and Management, 34*(1), 50–62.
28. Pandey, A., Singh, K., & Yadav, R. (2024). RE-RFME: A machine learning-based segmentation framework for real estate. *International Journal of Data Science and Analytics, 11*(2), 175–190.
29. Nachiappan, A. (2024). Artificial intelligence in predicting customer behavior in real estate. *International Journal of Business Analytics, 21*(1), 45–60.
30. Zhou, Q., & Li, X. (2022). Machine learning applications in real estate marketing: A systematic review. *Journal of Marketing Analytics, 10*(3), 241–259.
31. MRI Software. (2024). Artificial intelligence in real estate marketing. Industry Report. Retrieved from https://www.mrisoftware.com/
32. Aurum PropTech. (2024). Predictive analytics in Indian real estate markets. Industry Report. Retrieved from https://www.aurumproptech.in/
33. Chen, R., & Liu, H. (2024). Open data and explainable AI for real estate prediction. *Applied Sciences, 14*(6), 2891.
34. Wu, L., & Brynjolfsson, E. (2011). The future of prediction: How Google searches foreshadow housing sales. *National Bureau of Economic Research (NBER) Working Paper No. 17052.*
35. McLaren, N., & Shanbhogue, R. (2021). Using internet search data in UK housing market analysis. *International Real Estate Review, 24*(2), 145–162.
36. Xu, Y., & Zhao, L. (2023). Social media narratives and housing market dynamics: Evidence from China. *Journal of Real Estate Finance and Economics, 66*(3), 450–471.
37. Kim, S., & Park, J. (2024). Housing market sentiment and real estate pricing: Evidence from social media. *Journal of Housing Economics, 62*, 102–118.