House Price Prediction

Comprehensive Analysis

Prepared By:

- SENHAJI RHAZI, YOUSSEF
- AHMAD ABURAHMA
- SAMEH RAAFAT
- OSAMA MORAD

House Price Prediction: Comprehensive Analysis

Housing prices are a critical reflection of economic health and market trends. This analysis explores the factors that influence property values beyond just location and square footage. By examining 23 key features, we aim to predict housing prices accurately and identify the most significant attributes affecting property valuation.

Our comprehensive approach combines data cleaning, exploratory analysis, and advanced statistical methods to provide actionable insights for buyers, sellers, and real estate professionals.

Project Context & Objectives

Market Understanding

Housing prices reflect economic conditions and are crucial for both buyers and sellers in making informed decisions about property transactions.

Price Prediction

Develop a model to accurately predict housing prices based on multiple property features and locality characteristics.

Feature Identification

Determine which property attributes have the greatest impact on housing prices to guide investment and selling strategies.

Dataset Overview

Structural Details

- Number of bedrooms and bathrooms
- Total floors and area measurements
- Living area and basement size

Property Characteristics

- Year built and renovated
- Condition and quality ratings
- Furnished status (0-No, 1-Yes)

Geographical Attributes

- Zip code and coordinates
- Waterfront view (0-No, 1-Yes)
- Neighborhood features

Data Cleaning & Preprocessing

Missing Value Treatment

Identified missing values after numerical conversion and applied KNN Imputer to treat columns with NANs, resulting in a complete dataset.

Categorical Feature Analysis

Examined unique values in categorical features like bedrooms (0-33), bathrooms (0-8), floors (1-4), and condition ratings (1-5).

Category Consolidation

Merged low-frequency categorical feature values with similar mean price categories to reduce complexity.

Outlier Mitigation

Applied log1p transformation to numerical features to reduce the impact of extreme values and improve data distribution.

Cluster Characteristics Analysis

Cluster 0: Mid-Range Properties

Moderate median prices with slightly larger house sizes than Cluster 1. Houses are relatively newer (median age 36-37 years) with a typical condition rating of 3.

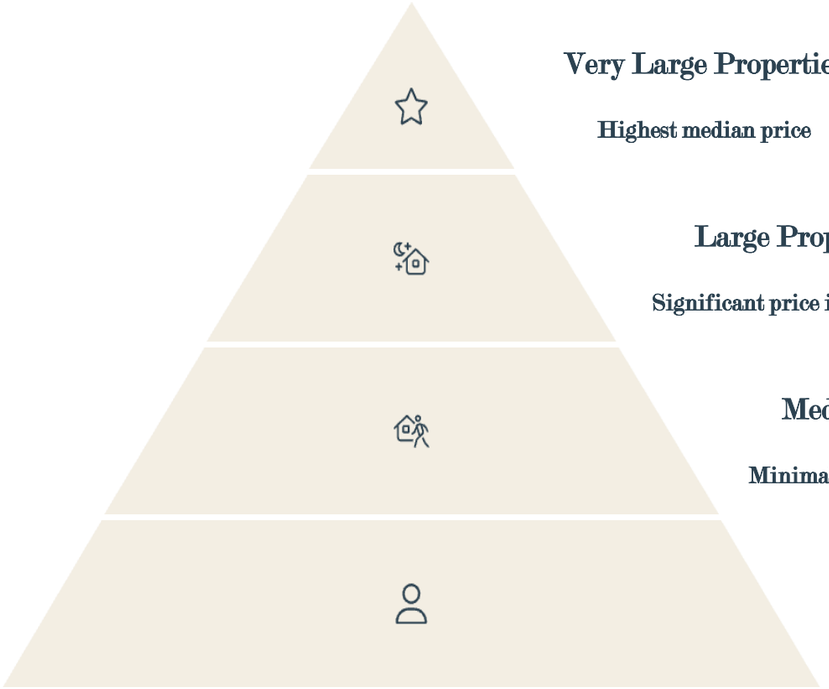
Cluster 1: Older, Established Homes

Moderate to high median prices despite having the oldest houses (median age 53 years). Similar house sizes to Cluster 0 with the same typical condition rating of 3.

Cluster 2: Premium Properties

Highest median prices with the largest median house sizes. Houses are relatively newer (similar age to Cluster 0) with the same typical condition rating of 3.

Property Size Impact on Price



Very Large Properties

Highest median price

Large Properties

Significant price increase from medium

Medium Properties

Minimal price difference from small

Small Properties

Baseline price point

- Our analysis shows that median prices rise as property size increases, which is expected in real estate markets.
- Price increases with property size.
- Incremental differences notable: Small→Medium (minimal), Medium→Large (significant), Large→Very Large (largest).

Property Age Impact on Price

Old Properties

Highest median price, likely due to historical value, prime locations, or unique architectural features

Mature Properties

Similar median price to young properties, showing consistent valuation in middle-aged homes

New Properties

High median price, reflecting modern amenities, updated designs, and contemporary features

Young Properties

Moderate median price, representing the middle market segment

- Old properties have highest median price (prime locations, unique features).
- New properties also highly valued (modern amenities).
- Young and mature properties exhibit similar median prices

Enhancing Real Estate Price Prediction with Advanced Model Engineering

A comprehensive analysis of model performance and feature engineering for accurate property value prediction.

Data Preparation

Data Cleaning & Feature Selection: Target encoding replaces the categorical values with a numerical representation that reflects the average value of the target variable

Feature Categorization: Separating numerical and categorical features

Feature Scaling (Standardization): Standardizing numerical features

Encoding Categorical Features: Applying one-hot encoding to categorical features and Dropping the first category to avoid multicollinearity.

Data Integration: Merging the scaled numerical and encoded categorical data into a single dataset (df_processed).

Target Variable Handling: Removing the target variable from the feature dataset (X) and Storing the target variable separately.

Train-Test Split: Spliting the dataset into training and testing sets: 70% training, 30% testing .

Model Performance Comparison

--- Original Model Performance Comparison (Train vs Test) ---

	Model	Test RMSE	Test R^2 Score	Train RMSE	Train R^2 Score
0	Linear Regression	0.360576	0.868788	0.364812	0.867422
1	Ridge Regression	0.360572	0.868791	0.364816	0.867420
2	Lasso Regression	0.995529	-0.000201	1.001923	0.000000
3	Decision Tree Regression	0.502447	0.745223	0.000345	1.000000
4	Random Forest Regressor	0.354486	0.873183	0.133457	0.982257
5	Ada Boost Regressor	0.479646	0.767822	0.478198	0.772204
6	Gradient Boost Regressor	0.354272	0.873336	0.345049	0.881398
7	XG Boost Regressor	0.335751	0.886233	0.214933	0.953981

Gradient Boost Regressor

Best-performing model with Test R^2: 0.8733, Train R^2: 0.8814.

Shows minimal gap between scores, indicating robust performance without overfitting.

Linear Regression

Very small gap between train and and test scores.

Shows no overfitting but delivers slightly lower test R^2 than boosting boosting models.

Ridge Regression

demonstrates very low overfitting, with similar performance to Linear Regression.
It adds a regularization term to reduce overfitting but does not significantly outperform Linear Regression.

Model Comparison using Hyperparameter Tuning

--- Tuned Model Performance Comparison (Train vs Test) ---

	Model	Test RMSE	Test R^2 Score	Train RMSE	Train R^2 Score
0	Linear Regression	0.360576	0.868788	0.364812	0.867422
1	Ridge Regression	0.360573	0.868790	0.364814	0.867421
2	Lasso Regression	0.467127	0.779783	0.475353	0.774906
3	Decision Tree Regression	0.408639	0.831477	0.342312	0.883272
4	Random Forest Regressor	0.356341	0.871852	0.235350	0.944823
5	Ada Boost Regressor	0.463349	0.783331	0.463978	0.785550
6	Gradient Boost Regressor	0.343675	0.880800	0.128722	0.983494
7	XG Boost Regressor	0.340234	0.883175	0.145966	0.978776

Linear Regression: Test R^2 Score: 0.8688, Train R^2 Score: 0.8674 and it has Minimal gap between train and test scores, indicating stable performance with no overfitting.

Ridge Regression: Test R^2 Score: 0.8688, Train R^2 Score: 0.8674, Similar performance to Linear Regression with added regularization to prevent overfitting.

Decision Tree Regressors tend to overfit because they fit the training data perfectly, especially when there is no pruning or regularization applied.

0.8688

Linear R² Test Score

Train score: 0.8674. Minimal gap indicates stable performance.

0.8688

Ridge R² Test Score

Train score: 0.8674. Regularization prevents overfitting.

High

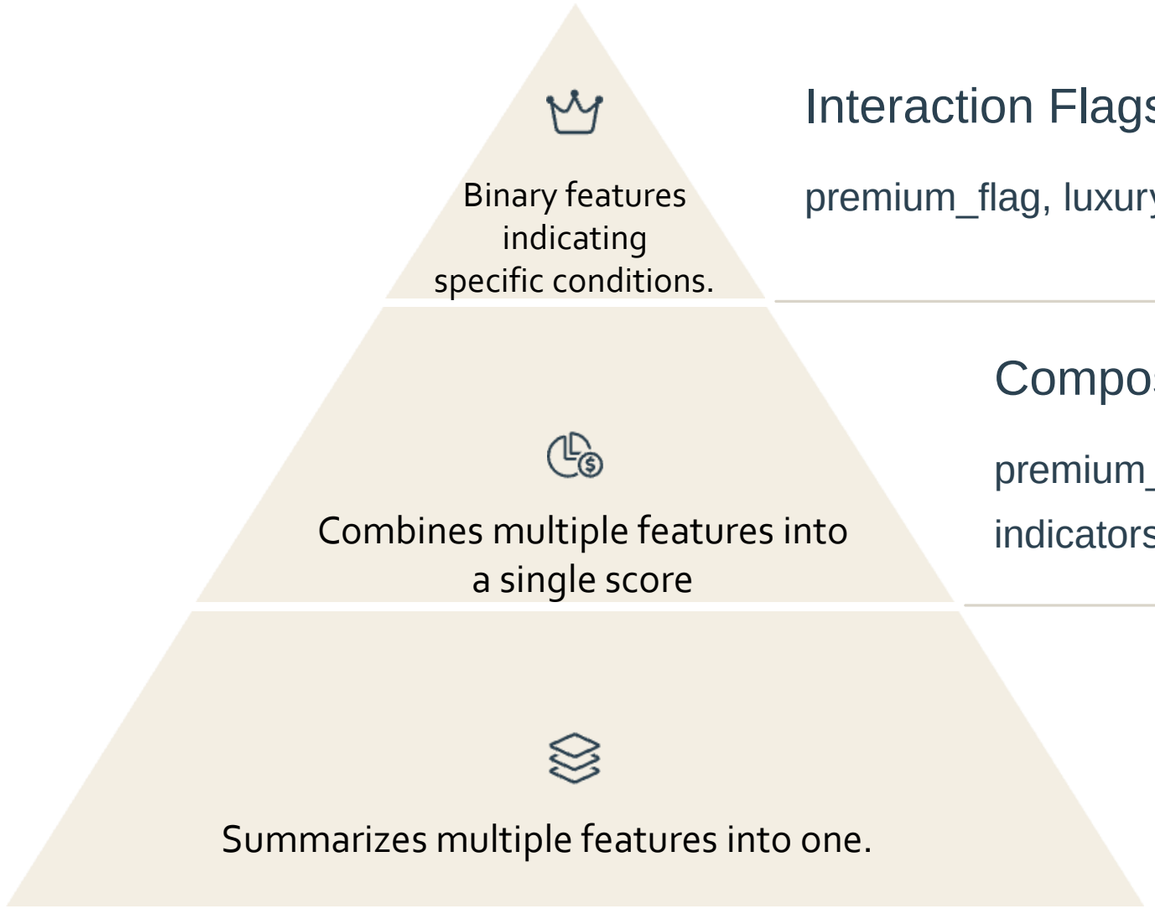
Decision Tree Overfitting

Perfect training fit without pruning leads to poor generalization.

Hyperparameter tuning further optimizes model performance by finding ideal parameters.

Feature Engineering

Feature Engineering is essential in the data preprocessing phase of machine learning, as it involves creating new features from existing data to enhance predictive model performance



Interaction Flags

premium_flag, luxury_flag indicate specific conditions

Composite Features

premium_features_score, luxury_features_score combine multiple indicators

Aggregated Features

premium_features_count summarizes multiple features into one

Model Comparison with Feature Engineering

--- Original Model Performance Comparison on the feature engineered dataframe (Train vs Test)

	Model	Test RMSE	Test R^2 Score	Train RMSE	Train R^2 Score
0	Linear Regression	0.360481	0.868857	0.364809	0.867425
1	Ridge Regression	0.360472	0.868864	0.364812	0.867422
2	Lasso Regression	0.995529	-0.000201	1.001923	0.000000
3	Decision Tree Regression	0.494693	0.753026	0.000345	1.000000
4	Random Forest Regressor	0.353110	0.874165	0.133383	0.982277
5	Ada Boost Regressor	0.475946	0.771390	0.472270	0.777817
6	Gradient Boost Regressor	0.355690	0.872320	0.348007	0.879355
7	XG Boost Regressor	0.336730	0.885569	0.215007	0.953949

- **Gradient Boost Regressor** :Test R^2 Score: 0.87128 , Train R^2 Score: 0.87897 it has Minimal overfitting, better generalization.it is Strong model with low variance between train and test.
- **Linear Regression** shows consistent results with a very small gap between training and testing scores. RMSE is relatively low, indicating decent prediction it has Minimal overfitting indicating the model generalizes well.
- **Ridge Regression,** which applies L2 regularization, performs almost identically to Linear Regression.It also shows minimal overfitting with a balanced performance between training and testing.

Top Models After Evaluation

Gradient Boost: Baseline Model

Strong performance with original features.

Excellent balance between accuracy and generalization.

Gradient Boost: Feature Engineered

Enhanced predictions with custom feature creation.

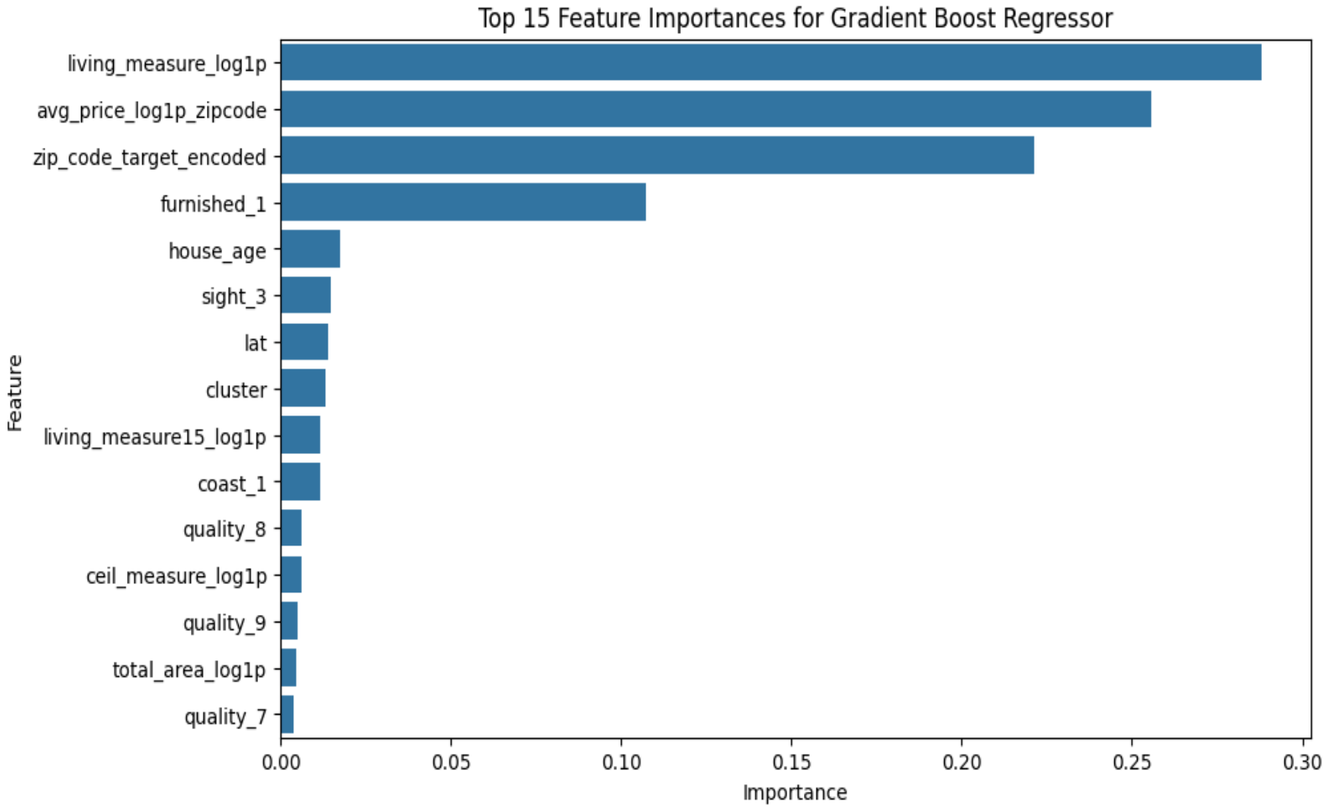
Maintained minimal gap between train and test scores.

Linear Regression: Feature Engineered

Simple yet effective with engineered features.

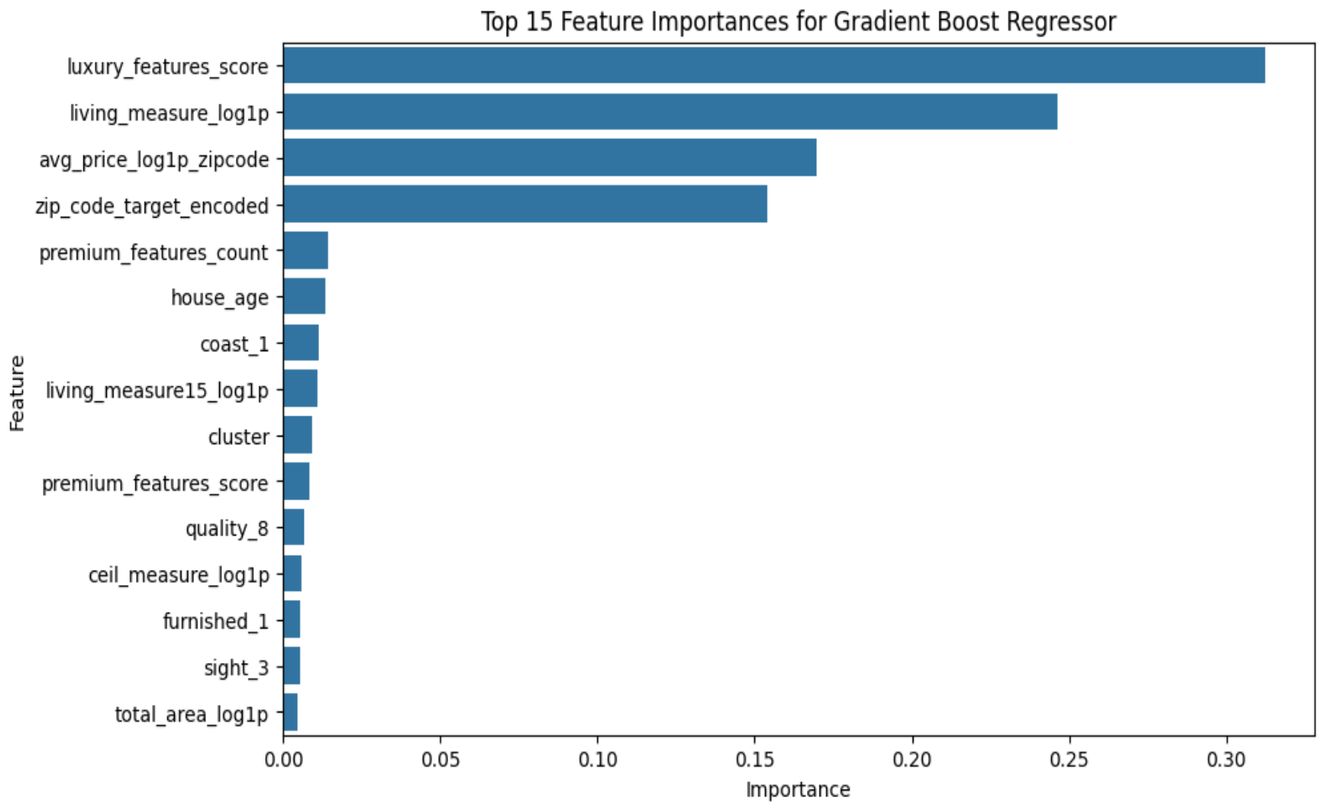
Highly interpretable results with minimal overfitting.

Gradient Boost Regressor: Baseline Model



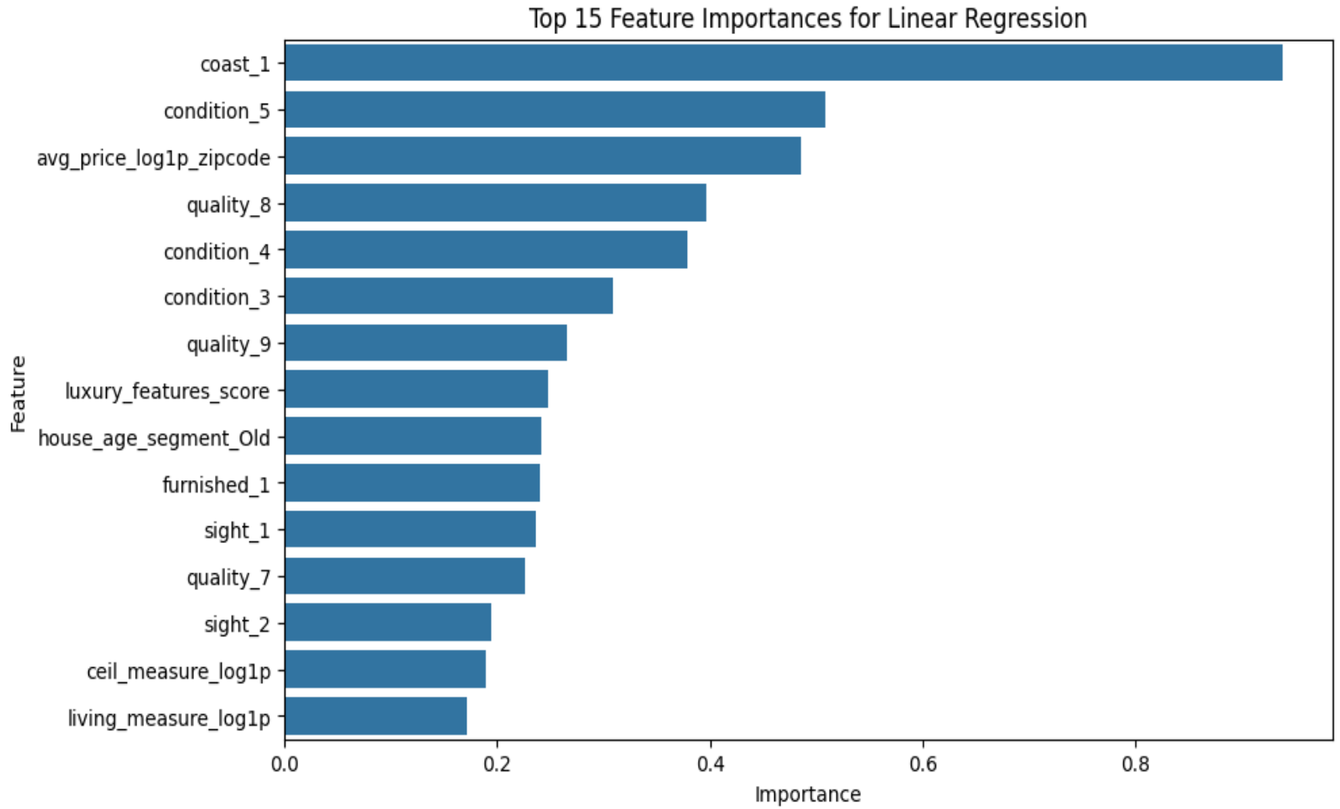
1. Living Area is the most influential feature, Larger living spaces significantly increase property value, making size a crucial factor.
2. Average Price by Zip Code : High importance highlights that property values strongly correlate with historical pricing trends in the area.
3. Encoding zip codes allows the model to capture geographical pricing patterns.
4. Furnished homes have a higher perceived value, buyers often prefer ready-to-move-in homes, leading to higher selling prices.
5. House Age: Older homes tend to have lower values.

Gradient Boost Regressor : Feature Engineered Model



1. Luxury Features Score is the most influential factor in determining house prices.
2. Living Area Size : Larger living spaces directly impact property value.
3. Average Price by Zipcode :Location remains a critical factor influencing property values.
4. Zip Code Target Encoding:Reflects the impact of neighborhood quality and historical price trends.
5. Homes with additional premium features such as swimming pools, large garages, and landscaped gardens enhance overall value.

Linear Regression
Feature Engineered Model



1. Coastal properties are the most influential factor in determining house prices.
2. Homes in excellent condition (rated 5) are highly valuable.
3. verage Price by Zipcode :Location remains a critical factor for pricing.
4. (quality_8):High-quality construction materials and designs add substantial value.

Business Insights & Recommendations

For COO (Operations & Risk Management):

- Focus on acquiring and maintaining high-demand inventory in prime zip codes.
- Implement strategic renovation plans for older homes to enhance their market value.
- Use Gradient Boost to refine inventory management by monitoring demand and price trends.

For CMO (Marketing & Customer Acquisition):

- Target high-value segments looking for luxury and furnished properties.
- Develop seasonal marketing strategies to promote properties during peak seasons.
- Showcase premium features using virtual tours and data-driven pricing insights to attract premium buyers.

For Data Science Manager (AI & Predictive Analytics):

- Utilize a hybrid modeling strategy combining Gradient Boost for capturing non-linear patterns and Linear Regression for ensuring consistent predictions.
- Continuously optimize models by incorporating external data (mortgage rates, economic trends).
- Develop a dynamic pricing engine to adjust pricing based on seasonal and economic factors, improving overall market responsiveness.

Value to the Company & Societal Impact

Strategic Business Benefits

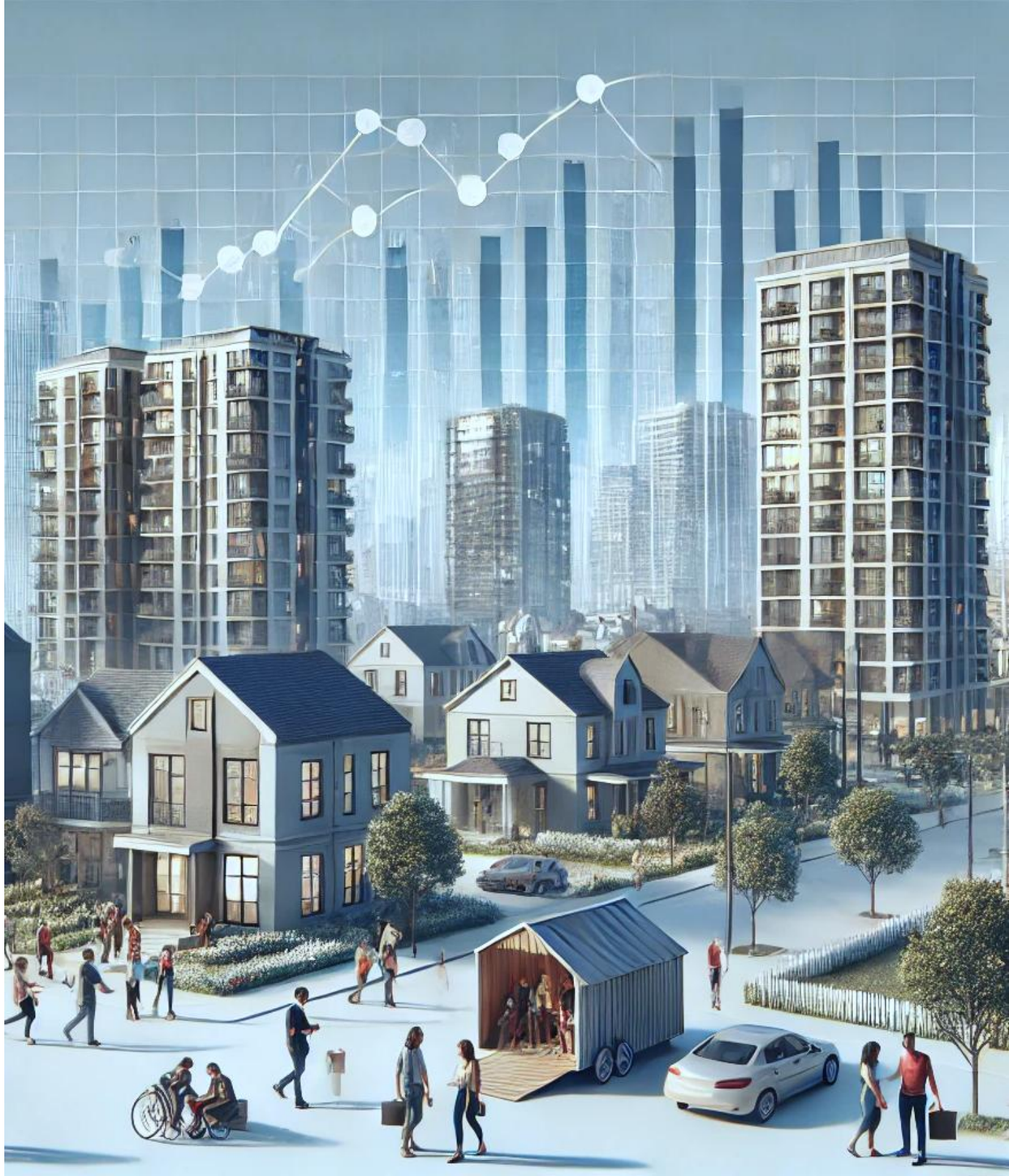
The model aids in dynamic pricing, smarter inventory decisions, and maximized ROI by targeting high-demand areas and optimizing renovation strategies.

Societal Value Creation

By promoting fairer pricing and data transparency, the system can influence responsible urban development, housing equity, and environmental sustainability.

CSR Alignment

The focus on upgrading existing homes and market transparency aligns with corporate social responsibility, supporting smart growth and community wellbeing.



Further Analysis Scope

Sentiment & Behavioral Insights

Future work could include text mining on listing descriptions and buyer sentiment analysis to enrich the prediction model with qualitative insights.

Temporal & Geospatial Expansion

Incorporating time-based trends and spatial clustering can help detect seasonality effects and regional pricing anomalies.

Deployment & Integration

The model can be embedded in business tools, real-time dashboards, or mobile apps to guide on-the-fly pricing and investment decisions.

Thank You