Housing Prices – Regression

1. Problem Statement

The housing price prediction problem is a valuable asset for addressing critical challenges in real estate, urban planning, and financial sectors. By accurately predicting housing prices, businesses can optimize decision-making processes in property investments, market analysis, and economic forecasting. This predictive capability benefits stakeholders such as real estate agencies, homebuyers, investors, and policymakers by providing data-driven insights into market trends, enabling competitive pricing strategies, and promoting informed financial decisions.

2. Proposed Solution

This paper aims to estimate housing prices using multiple machine learning algorithms, leveraging features such as property characteristics, location, and other indicators. By comparing the performance of various models, the project seeks to identify the most effective approach for housing price prediction. The study also evaluates the interpretability and accuracy of the models, contributing to the understanding of data-driven pricing strategies in real estate markets. The findings will provide actionable insights for improving property valuation practices, enhancing market transparency, and fostering evidence-based decisionmaking in the housing sector.

3. Exploratory Data Analysis

3.1 Understanding the Data

Data Source: https://www.kaggle.com/competitions/house-prices-advanced-regressiontechniques/data

The data set has 75 features, and close to 1500 instances of data. This is a good set for most model prediction and building. The number of instances is a bit lesser, therefore EDA becomes a crucial step for this dataset.

3.2 Data Cleaning and Feature Reduction

The data was cleaned for null values. There were multiple null values found. Since many of these values covered large portions of the dataset, they cannot simply be removed (Figure below).

To address this issue, I used the following method for data cleanup:

1. Numeric Values: Replace the Null value with the mean value from the dataset 2. Categorical Values: Replace the Null values with Mode Value from that feature set.
3. Extremely High Null Values: If there is a feature with more than 50% null values, that feature was removed.

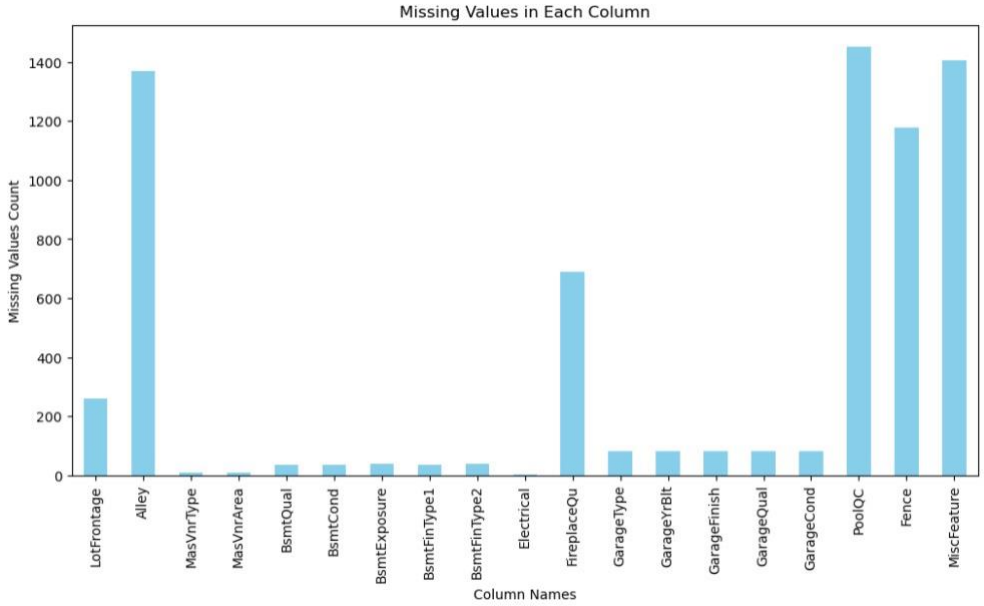


Figure 3 Missing Values in Regression Dataset

3.2.1 Feature Encoding

For the use of model building, I have performed one hot encoding, this resulted in the following dataset to be generated with 217 columns, this is an extremely high number and some type of feature engineering is required to reduce.

3.2.2 Feature Engineering

I have performed Principal Component Analysis on scaled data. Scaling was done using a standard scalar operation, which puts data on a mean of 0 and unit standard deviation of 1. I have made sure to create components which explain 70% of the correlation. This resulted in a dimension reductional from 217 to 70 features. This would make the models simpler and easier for interpretation.

4. Supervised Models Evaluation and Interpretation **4.1**

Building Models

The Models were built by first taking the data and scaling it. Once the data was scaled, the dataset was divided for a train and test split of 80:20. The training set was then further refined with the help of cross validation implementation with 5 folds. This was again paired with a grid-search method to obtain the optimum parameters.

Once the optimum parameters were found for each model, those values were used to create new models final models and used for prediction on the test set. This was followed by appropriate metrics evaluation such as MSE and R^2.

4.1 Comparative Analysis and Model Selection

In this section, we consolidate the performance metrics of all evaluated regression models. The primary focus is on the Mean Squared Error (MSE) as a measure of predictive accuracy—lower values indicate predictions that are, on average, closer to the actual values. We also consider the coefficient of determination (R^2) to quantify how much variance in the target is explained by each model. A higher R^2 suggests that the model captures more of the underlying structure of the data.

Our business context involves forecasting target values (e.g., house prices or similar continuous outcomes) that directly influence important financial or operational decisions. A lower MSE ensures more reliable estimates, reducing costly errors. Meanwhile, a higher R^2 helps identify the key factors driving the target variable, offering valuable insights for improvement strategies, pricing models, or investment considerations.

4.1.1 Comparison of All Models

Summary of Model Performance:

Model	Mean Squared Error (MSE)	R^2 Score
Decision Tree	1.52e+09	0.80
Linear Regression	1.19e+09	0.85
K-Nearest Neighbors (KNN)	7.97e+08	0.74
Random Forest	1.31e+09	0.83

Based on the updated values in the table, the final model selection and the accompanying analysis are as follows:

Analysis:

Decision Tree Model

The Decision Tree model achieves a moderate R^2 score of 0.80, indicating that it explains about 80% of the variance in the target variable. However, its Mean Squared Error (MSE) of approximately 1.52e+09 is relatively high, which suggests significant errors in individual predictions despite capturing the overall trends in the data. This high level of prediction error reduces its reliability for precise forecasting tasks, as the model may struggle to make accurate predictions at a granular level. While the Decision Tree is a robust option for identifying general patterns, it falls short in delivering the accuracy required for more sensitive business applications.

Linear Regression

Linear Regression demonstrates strong performance, with the highest R^2 score of 0.85 among all the models. This means it effectively explains 85% of the variance in the target variable, showcasing its ability to capture relationships between predictors and the outcome. Its MSE of approximately 1.19e+09 is relatively low, ensuring reliable and precise predictions. Additionally, Linear Regression is simple and interpretable, making it easy to understand and communicate its results. This combination of high explanatory power, competitive accuracy, and interpretability makes it a strong candidate for applications that require both actionable insights and dependable predictions.

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) model excels in terms of predictive accuracy, achieving the lowest MSE of approximately 7.97e+08. This indicates that its predictions are, on average, the closest to actual values compared to other models. However, it achieves the lowest R^2 score of 0.74, meaning it explains less of the overall variance in the target variable. This discrepancy suggests that while KNN performs well for individual predictions, it may not fully capture the broader relationships in the data. KNN's reliance on local patterns makes it a suitable choice for tasks where minimizing prediction error is prioritized over understanding the underlying data structure.

Random Forest

The Random Forest model strikes a balance between explanatory power and predictive accuracy, with an R^2 score of 0.83 and an MSE of approximately 1.31e+09. It outperforms the Decision Tree by reducing prediction errors and offers a slightly better R^2 score. However, it does not surpass Linear Regression in terms of accuracy nor KNN in minimizing MSE. While Random Forest effectively captures complex relationships in the data, its performance metrics indicate that it is not the optimal choice for this specific use case. It may still be a viable option for scenarios requiring robust and consistent predictions across varied data distributions.

Choosing the Best Model:
Given the updated metrics, **Linear Regression** is now the best choice. Its combination of the highest **R^2 score (0.85)** and a low **MSE (~1.19e+09)** ensures that it captures substantial variance in the target while maintaining accurate predictions.

Rationale:

1. **High Explanatory Power**: Linear Regression's **R^2 (0.85)** outperforms all other models, providing the clearest understanding of how features influence the target.

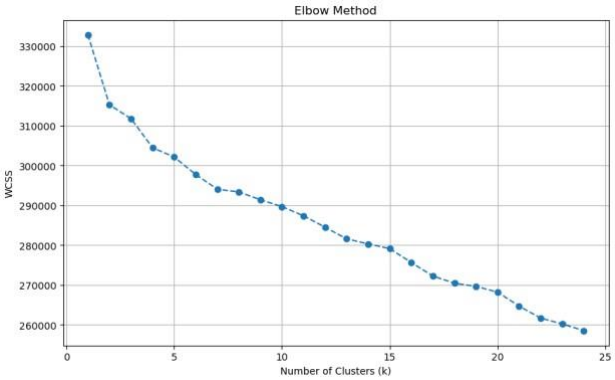
2. **Reliable Accuracy**: Its low **MSE (~1.19e+09)** ensures that predictions closely match actual values, reducing costly misjudgements.

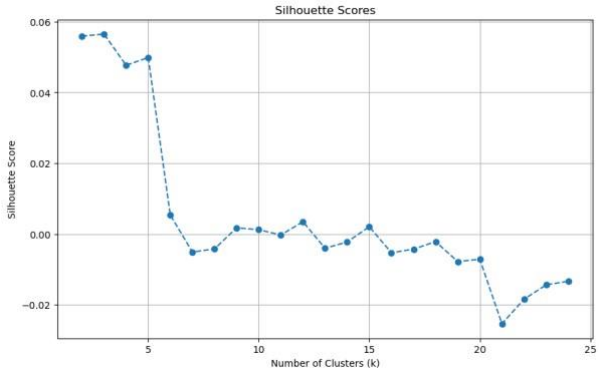
3. **Simplicity and Interpretability**: The straightforward nature of Linear

Regression makes it easier to communicate results to stakeholders while avoiding overfitting risks.

4.2 K-Means Clustering

To perform clustering, I performed K-means clustering and found that the optimal value for k





is 4. This was chosen as it gave one of the highest silhouette scores and the decrease in WCSS is lesser, beyond that point, as seen in the image below.

5 Business Application

A Linear Regression model can be applied for predicting the housing prices. It was has a high explanation power of 0.85. However, the MSE comes out to **1.19e+09**, which will be the mean square error in finding the optimum pricing, is slightly on the higher side. This must be put into account before making the pricing prediction.

5.1 Cluster Descriptions for Housing Prices Analysis

The clustering performed with the help of k-means clustering helped us find the following insights and make the appropriate grouping:

Cluster 0: "Vintage Essentials"

- **Properties**: This cluster represents homes with a blend of vintage charm and functional design. Built predominantly in the early to mid-20th century, these houses have moderate living spaces (GrLivArea ~1360 sq. ft) and simpler layouts with single-story or partial second floors. Key features include smaller garages, basic porches, and functional but unimproved basements. The overall quality and condition are average, making them affordable and ideal for firsttime homebuyers or those seeking budget-friendly options.
- Stakeholder Benefits:
  - **Realtors**: Can market these homes as affordable starter options or for buyers who appreciate older properties with potential for upgrades.
  - **Buyers**: Opportunity to purchase budget-friendly homes in established neighbourhoods, ideal for personalization.

Cluster 1: "Modern Comforts"

- **Properties**: Homes in this cluster are newer (built around 2000) and feature modern layouts and amenities. They have larger living spaces (GrLivArea ~1770 sq ft), well-finished basements, attached garages, and open porch spaces. High-quality construction and contemporary materials dominate, making them suitable for mid-to-upper-income families. These homes are located in suburban areas with spacious lots and well-maintained streets. ○ Stakeholder Benefits:
  - **Realtors**: Can highlight the ready-to-move-in nature and modern features for families looking for convenience and contemporary living.
  - **Buyers**: Ideal for those seeking newer homes with minimal renovation needs and modern aesthetics.

Cluster 2: "Classic Fixer-Uppers"

- **Properties**: Older homes built in the early 20th century (~1915), this cluster represents smaller, functional properties with basic features. Lot areas are moderate, but interiors are dated, with lower quality basements and limited modern amenities. These houses often require significant upgrades and are located in older neighbourhoods, some with historical significance.
- Stakeholder Benefits:
  - **Realtors**: Can position these as investment opportunities or for buyers interested in restoration projects.
  - **Buyers**: Ideal for those seeking affordable entry into the market with the potential to renovate and add value.

Cluster 3: "Suburban Classics"

- **Properties**: These homes are moderately sized (GrLivArea ~1350 sq ft), built in the mid-20th century (~1964), and feature large lots (~11,680 sq ft). They balance traditional layouts with some modern touches like finished basements and attached garages. Located in suburban neighbourhoods, they are well-

  suited for families looking for space and affordability without compromising on functionality.
- Stakeholder Benefits:
  - **Realtors**: Can market these homes as family-friendly options with a balance of affordability and space.
  - **Buyers**: Perfect for families seeking suburban settings with room to grow and upgrade over time.

5.2 Limitation for the model:

The model unlike the decision tree previously plotted for classification task was not able to find the optimal insights. As we had performed feature engineering using PCA. This would add as an extra computation step of expanding what every single PCA component is built of to devise insights from the Decision Tree.

Additionally, even though linear regression offers the advantage of simplicity and explainability, this likely means that, instead of bringing all components down to a 2d space, it would be more helpful to take individual features and then plot linear regression models to gain insights. This can be paired with the best models to obtain the optimal insights.