Explanation of the House Price Prediction Model

1. Understanding the Dataset

The dataset contains housing-related features such as the number of rooms, median income, location, and more. These factors influence house prices. Since some values were missing, they were handled by removing incomplete rows to ensure a clean dataset.

2. Feature Engineering and Selection

To improve the predictive power of the model, new features were created:

- **Rooms per household** – Total rooms divided by the number of households.
- **Bedrooms per room** – Ratio of bedrooms to total rooms, which indicates house space.
- **Population per household** – How many people live in each household.
- **Income per room** – Median income divided by total rooms to measure affordability.
- **Location interaction** – Multiplying latitude and longitude to capture geographical effects.

After creating these features, only the most important ones were selected to improve prediction accuracy and reduce unnecessary complexity.

3. Data Preprocessing

Before training, the data was prepared using:

- **Label Encoding:** Converting the categorical `ocean_proximity` feature into numeric values.
- **Scaling:** Applying `StandardScaler()` to normalize feature values so that no single feature dominates the model.

4. Model Training and Evaluation

A **Linear Regression model** was trained using the selected features. The model was evaluated using the **R^2 score**, which measures how well it predicts house prices. The closer the score is to 1, the better the model.

5. Predictions and Testing

A set of **sample houses with different feature values** was tested to check how well the model predicts prices. The results were compared using a **scatter plot of actual vs. predicted prices** to visually analyze accuracy.

<> Final Conclusion

- The model follows the task requirements of predicting house prices using **Linear Regression**.
- It includes **data preprocessing, feature selection, model training, and evaluation**.
- Further improvements can be made by testing different models like Decision Trees or Random Forests.