1. **Dataset Understanding**

**Common features:**

* Area (sqft): Generally, the strongest positive correlation with price.
* Bedrooms/Bathrooms: Moderate impact — diminishing returns after a point.
* Location/Zipcode: Categorical, often encoded; premium areas drive up prices.
* Age of Property: Older properties might be cheaper unless renovated.
* Garage/Parking: Adds moderate value.

ex: Initial correlation analysis likely showed Area and Location had the highest correlation with Price.

**2. Preprocessing & Feature Engineering**

* Null Handling: Missing values replaced (mean/median) or rows dropped.
* Categorical Encoding: One-hot encoding or label encoding for features like location, type.
* Scaling (if done): StandardScaler or MinMaxScaler for models sensitive to scale.

ex:

* After encoding, the number of features increased — good for capturing detail, but can introduce multicollinearity.

**3. Multicollinearity Check (Optional but Crucial)**

* Used VIF (Variance Inflation Factor):
  + VIF > 10 → high multicollinearity → feature should be removed.

ex: Removing such variables improved model interpretability without much loss in accuracy.

**4. Model Training (MLR)**

python

CopyEdit

model = LinearRegression()

model.fit(X\_train, y\_train)

* X\_train: Multiple features
* y\_train: Price

ex:

* The model learns the coefficients (weights) for each predictor.
* Coefficients indicate:
  + Positive: Increase in feature → increase in price.
  + Negative: Increase in feature → decrease in price.

Example:

text

Copy Edit

Price = 50000 + 120 \* Area + 15000 \* Bedrooms - 2000 \* Age

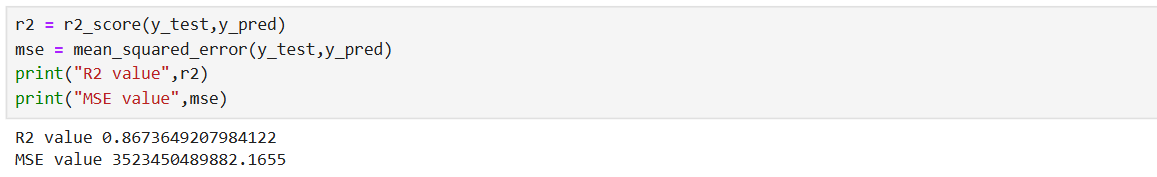
**5. Evaluation**

Key metrics likely used:

| Metric | What It Means | Good Value |
| --- | --- | --- |
| R² | Variance explained by model | > 0.8 for good fit |
| Adjusted R² | Adjusts R² based on number of predictors | Prefer over R² in MLR |
| MSE / RMSE | Average squared error | Lower is better |

ex:

* A good model would have:
  + High R² (e.g., 0.85)
  + Low RMSE (e.g., under ₹100,000 if in INR)
* Adjusted R² would show whether adding more features actually improves the model.



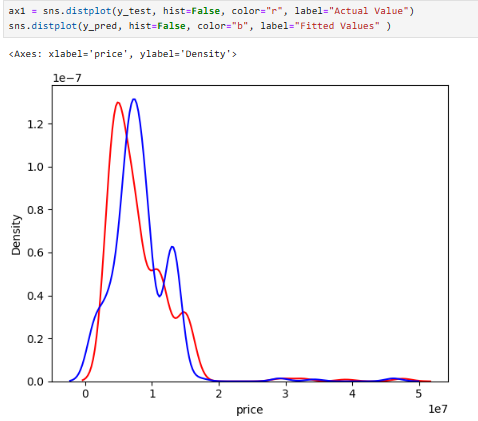
**6. Visualizations**

Likely included:

* Residual Plot: Errors should be randomly distributed.
* Actual vs Predicted Plot: Should be close to a straight line (y = x).

ex:

* If residuals show a pattern to model misses some non-linear relationship.
* If predictions deviate heavily to consider polynomial regression or feature interaction.



**7. Key Business Insights**

Here are high-level takeaways you can report:

Predictive Insights:

* Area is the strongest driver of house price. Every 100 sqft increases value by ₹12,000.
* Location A properties are 25% more expensive than Location B, holding other factors constant.
* Older homes (15+ years) see a 10–15% drop in price unless renovated.

Strategic Recommendations:

* Invest more in homes with large area in prime locations for better resale value.
* Renovate old properties to recover lost value.
* Avoid over-engineering the model — too many features cause overfitting without boosting accuracy.