SUMMARY

In this project, I aimed to predict housing prices in California using a **Linear Regression** model trained on the **California Housing dataset**. The target variable was the **median house value (MedHouseVal)**, and we experimented with various combinations of features ranging from 2 to 8.

**->Feature Selection and Model Performance**

We tried multiple sets of features and evaluated model performance using **Mean Squared Error (MSE)** and **R² score** on the test data. Below is a comparison of R² scores across different feature sets:

| **Features Used** | **R² Score** |
| --- | --- |
| MedInc, HouseAge | 0.55 |
| MedInc, AveRooms, AveOccup | 0.60 |
| MedInc, AveRooms, AveOccup, HouseAge | **0.62** |
| All 8 features | 0.59 |

The **best R² score of 0.62** was achieved using **four features**:  
MedInc (Median Income), AveRooms (Average Rooms), AveOccup (Average Occupancy), and HouseAge (Age of the house).

**-> Evaluation Metrics**

* **Mean Squared Error (MSE):** ~0.47 (for best model)
* **R² Score:** ~0.62 (explains 62% of the variance in the target)

**-> Visualization**

We visualized the model predictions using a scatter plot of **actual vs predicted values**, where most points were closely clustered around the ideal prediction line, indicating reasonably good performance.

**-> Improvements and Next Steps**

Although the linear model captured basic relationships, performance can be enhanced by:

* Using **non-linear models** (e.g., Random Forest, XGBoost)
* Applying **feature scaling** or **log transformation**
* Conducting **feature engineering** (creating new interaction variables)
* Using **cross-validation** to reduce overfitting
* Detecting and handling **outliers** more effectively