🏠 Housing Price Prediction Project Report

# 1. Project Overview

The goal of this project is to build a machine learning model that predicts housing prices in California using structured housing data.   
The project uses the California Housing dataset from Scikit-learn, which includes various features such as median income, housing age, average rooms, and location.   
This end-to-end project involves data preprocessing, visualization, modeling using linear regression, and evaluation of the results.

# 2. Dataset Information

• Source: sklearn.datasets.fetch\_california\_housing()  
• Total Rows: 20,640  
• Target Variable: MedHouseVal – Median house value for households within a block.  
• Features Used: MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude

# 3. Methodology

## 3.1 Data Loading & Exploration

Used pandas and seaborn for exploration. No missing values found. Heatmap revealed that MedInc has the highest correlation with MedHouseVal.

## 3.2 Feature Selection

All original features were retained. MedInc, AveRooms, and HouseAge showed positive correlation with housing price. Latitude and Longitude used as numeric predictors without further geospatial engineering.

## 3.3 Data Splitting

Dataset was split into training and test sets using train\_test\_split(X, y, test\_size=0.2, random\_state=42).

## 3.4 Model Training

Applied Linear Regression using Scikit-learn. Model was trained on 80% of the data.

# 4. Model Evaluation

Metrics:  
• Mean Squared Error (MSE): ~0.5559   
• R² Score: ~0.5758  
  
Interpretation:  
The model captures general pricing trends but lacks high accuracy due to simplicity of the linear model, no feature transformation, and capped target variable at 5.0.  
Visualization:  
Scatter plot between actual vs predicted values shows the model performs well around the mean but struggles on higher-end capped values.

# 5. Key Insights

• Median income is the strongest predictor of house price.  
• Location features could be more powerful if engineered properly.  
• High-density zones do not directly imply higher value.

# 6. Future Improvements

• Apply non-linear models (Random Forest, XGBoost).  
• Use feature engineering (e.g., rooms per household).  
• Remove or model separately capped target values (MedHouseVal = 5.0).  
• Visualize geographic patterns with maps.

# 7. Deliverables

• Python Jupyter Notebook (.ipynb)  
• Trained model (optional .pkl)  
• Visualizations (correlation heatmap, scatter plots)  
• Project report (this document)