📌 Executive Summary

This project focuses on building a predictive model for California house prices using the well-known California Housing Dataset derived from the 1990 U.S. Census. The objective was to explore the dataset, identify key patterns, and implement a machine learning solution capable of estimating median house values.

Key Highlights:

Data Exploration: Conducted statistical analysis and visualizations to understand the relationships between variables such as median income, house age, average rooms, and location.

Preprocessing: Applied normalization, correlation checks, and outlier detection to ensure clean input for the model.

Model Training: Used Linear Regression as a baseline model to capture the relationship between features and house prices.

Evaluation: The model achieved an R^2 score of ~0.59, with performance metrics as follows:

MSE: 0.53

MAE: 0.52

RMSE: 0.72

Deployment Readiness: Saved the trained model using Pickle, enabling future use for predictions on unseen data.

Business Impact:

While the model does not fully capture the complexity of real estate markets, it provides valuable insights into which features drive house prices the most (e.g., income and location). This approach can be extended with advanced algorithms (Random Forest, Gradient Boosting, Neural Networks) for improved accuracy.

In conclusion, the project demonstrates a complete data science workflow — from raw data to a deployable predictive model — and highlights the potential of machine learning in addressing real-world economic challenges.