🏠 Summary of Analysis and Findings

🌟 1. Problem Definition

- **Goal**: Predict house prices (`MedHouseVal`) using features like median income (`MedInc`), house age (`HouseAge`), average rooms (`AveRooms`), population, etc.

📊 2. Data Collection

- **Dataset**: California housing data with **20,640 samples**.
- **Key Features**:
  - Independent: `MedInc`, `HouseAge`, `AveRooms`, `AveBedrms`, `Population`, `AveOccup`, `Latitude`, `Longitude`.
  - Dependent: `MedHouseVal` (normalized house value).
- **Data Source**: Fetched from `sklearn.datasets`.

🛠️ 3. Data Preprocessing

- **Scaling**: Features normalized using `StandardScaler` to ensure consistent scaling.
- **Missing Data**: No missing values detected in the dataset.
- **Correlation Highlights**:
  - **Strong Positive**: `MedInc` and `MedHouseVal` (68.8%).
  - **Weak**: Most other features had correlations below 15%.

🎯 4. Feature Selection

- **Methodology**: Variance Inflation Factor (VIF) to address multicollinearity.
- **Retained Features**:
  - `MedInc`, `HouseAge`, `Population`, `AveOccup` (account for **47.8%** variance).
- **Impact**:
  - `MedInc` had the highest predictive influence (VIF: 2.5).

🔍 5. Exploratory Data Analysis (EDA)

- **Heatmap**: Highlighted feature relationships using a coolwarm palette.
- **Scatter Plots**:

    - `AveRooms` vs. `AveBedrms` (84.7% correlation).
    - Geographic clustering observed for `Latitude` vs. `Longitude`.
- **Distribution**: `MedInc` showed skewness, hinting at outliers or data concentration.

🚀 6. Model Development

- **Train-Test Split**:
    - Training: **16,512 samples** (80%).
    - Testing: **4,128 samples** (20%).
- **Algorithm**: Linear Regression:
    - Dependent on `MedInc`, `HouseAge`, `Population`, `AveOccup`.
    - Coefficients confirmed `MedInc` as the strongest predictor.

✅ 7. Assumptions Validation

- **Linearity**: Verified through residual vs. predicted plots.
- **Independence**:
    - Durbin-Watson statistic: **2.01** (ideal).
- **Homoscedasticity**: Residual variance consistent across predictions.
- **Normality**: Histogram and Q-Q plot confirmed normal distribution of residuals.
- **Multicollinearity**: Addressed through feature selection.

📈 8. Model Evaluation

- **Performance Metrics**:
    - **Mean Absolute Error (MAE)**: **0.683** (~13.7%).
    - **Mean Squared Error (MSE)**: **0.844**.
    - **Root Mean Squared Error (RMSE)**: **0.919** (~18.4%).
    - **R-squared**: **37.6%**, showing moderate predictive power.
- **Residual Analysis**:
    - Average prediction error: **0.563**.

💾 9. Model Saving and Loading

- **Saved Models**:
    - **Pickle**: `house_model.pkl`.
    - **Joblib**: `house_model.joblib`.

- **Real-Time Predictions**: Tested successfully with logical trends observed.

🔮 10. Prediction Insights

- **Batch Predictions**: Applied on the test dataset (20% of data).
- **Real-Time Case**: Inputs like `[100, 200, 300, 400]` yielded realistic results.

✨ 11. Key Recommendations

- **Multicollinearity**: Perform VIF before scaling features.
- **Non-Linear Models**: Consider ensemble methods (e.g., Random Forest) for better R-squared scores.
- **Data Enrichment**: Add features like proximity to schools, crime rates, or zoning laws for improved predictions.
