Project Report: Predicting California Housing Prices

1. Aim and Objective

The primary goal of this project was to implement and evaluate three different linear regression models to predict housing prices in California. The models used were **Multiple Linear Regression**, **Ridge Regression (L2)**, and **Lasso Regression (L1)**. The objective was to preprocess the data, analyze feature relationships, build the models, tune their hyperparameters, and compare their performance to identify the most effective model for this dataset.

---

2. Data Preprocessing and Exploratory Data Analysis (EDA)

- **Dataset Loading:** The project utilized the California Housing dataset, which was loaded directly from the `sklearn.datasets` library. The dataset consists of 20,640 rows and 9 columns, including 8 feature variables and 1 target variable (`MedHouseVal`).
- **Data Cleaning:** An initial check for missing values confirmed that the dataset was complete, with no null entries requiring imputation.
- **Feature Analysis:**
    - The independent variables (features) include `MedInc`, `HouseAge`, `AveRooms`, `AveBedrms`, `Population`, `AveOccup`, `Latitude`, and `Longitude`
    - Histograms were generated for each feature to understand their distributions. Features like `MedInc`, `AveRooms`, and `Population` showed a right-skewed distribution, indicating the presence of outliers.
    - A correlation matrix was used to examine the relationships between variables. The analysis identified **`MedInc` (Median Income)**, **`AveRooms` (Average Rooms)**, and **`Latitude`** as the top three features with the highest absolute correlation to the target variable, `MedHouseVal`.
- **Data Splitting and Scaling:** The data was split into a training set (80%) and a testing set (20%). The features were then scaled using `StandardScaler` to normalize their range, which is essential for regularization models like Ridge and Lasso.

---

3. Model Building and Hyperparameter Tuning

Three regression models were built and evaluated. For Ridge and Lasso, hyperparameter tuning was performed using `GridSearchCV` with 10-fold cross-validation to find the optimal regularization strength (`alpha`).

1. **Multiple Linear Regression:** This model served as the baseline. The evaluation of its residuals plot indicated the presence of non-linear patterns in the data and violations of homoscedasticity, suggesting a simple linear model might not be the best fit for this dataset.
2. **Ridge Regression (L2 Regularization):** This model is used to prevent overfitting by penalizing large coefficients. Through `GridSearchCV`, the best hyperparameter was found to be `alpha=1.0`.
3. **Lasso Regression (L1 Regularization):** This model can perform feature selection by shrinking irrelevant feature coefficients to exactly zero. The optimal hyperparameter was determined to be `alpha=0.0010`.

4. Results and Model Comparison

The performance of each model was evaluated on the test set using Root Mean Squared Error (RMSE), R-squared (R²), and Mean Absolute Error (MAE). The results are summarized below.

Model	RMSE	R² Score	MAE
Linear Regression	0.745581	0.575788	0.533200
Ridge Regression	0.745557	0.575816	0.533193
Lasso Regression	**0.744642**	**0.576856**	**0.533145**

Analysis of Results:

- All three models produced very similar results, with marginal differences in their performance metrics.
- **Lasso Regression** emerged as the top-performing model, albeit by a very slight margin. It achieved the lowest RMSE and the highest R² score.
- The **R² score** for all models was approximately **0.58**, which means that the features in the models explain about 58% of the variability in the median house value. This indicates a moderate predictive power.
- The similarity in performance between the regularized models (Ridge, Lasso) and the standard Linear Regression model suggests that multicollinearity was not a significant issue and that most features were relevant to the prediction.

5. Conclusion

The project successfully demonstrated the process of building, tuning, and evaluating regression models for price prediction.

Based on the evaluation metrics, **Lasso Regression is the best-performing model** for this dataset, though its improvement over the other models is minimal.

The overall moderate performance (R² ≈ 0.58) and the patterns observed in the residual plots suggest that the linear relationship assumed by these models does not fully capture the complexity of the housing price data. For future work, exploring **non-linear models** such as

Polynomial Regression, Random Forest, or Gradient Boosting could potentially yield more accurate predictions by better modeling the intricate patterns within the data.