

# Task 2 Write-up

## Objective

This task involved building a regression model to predict house prices using the **California Housing dataset**. The goal was to apply preprocessing, train a linear regression model, evaluate its performance, and visualize key insights including residuals and model fit.

---

## Dataset Overview

- **Dataset:** `fetch_california_housing()` (20,640 records, 8 features)
  - **Target variable:** PRICE (median house value in \$100,000s)
  - Dataset was clean, with **no missing values**
- 

## Preprocessing Steps

- Performed **Exploratory Data Analysis (EDA)** via distribution plots and a correlation heatmap
  - Split data into training and testing sets (80/20 split)
  - Applied **standardization** (`StandardScaler`) to normalize features
  - Visualized **before and after scaling** (e.g., for `MedInc`) to confirm successful transformation
- 

## Model Training

Used **Linear Regression** from `scikit-learn`.

Trained on the scaled training data and evaluated on the test set.

### Model Parameters:

- `fit_intercept=True`
- `copy_X=True`
- `positive=False`

### Learned Feature Coefficients:

Visualized via horizontal bar plot to show which features most influence predicted prices (e.g., `MedInc` had the highest positive weight).

---

## Evaluation Metrics

- ♦ **R<sup>2</sup> Score:** 0.576
- ♦ **MSE:** 0.556
- ♦ **RMSE:** 0.746

These metrics indicate a moderate level of predictive accuracy for a basic linear model without regularization.

---

## Visualizations

- ♦ **Actual vs Predicted Prices:** Strong linear trend, some under/over-prediction visible
  - ♦ **Distribution of Residuals:** Centered around zero, reasonably symmetric
  - ♦ **Residuals vs Predicted:** No strong heteroscedasticity, supporting model consistency
- 

## Conclusion

Linear Regression performed reasonably well on the California Housing dataset. While not perfect, it captured general price trends. Further improvement could be achieved with more advanced models (e.g., Ridge, Lasso, or Tree-based regressors), but this project provided a solid foundation in regression modeling, evaluation, and interpretability.