

Name: Abdullahi Bala Mohammed

Student ID: 2018673

Course: Machine learning and pattern recognition

## REPORT SUMMARY

### 1. Introduction

The California housing dataset provided with the assignment was used to develop linear regression models for predicting *median\_house\_value*. Preprocessing steps included:

- Dropping the total\_bedrooms feature (as recommended).
- Removing the categorical feature ocean\_proximity.
- Using only numerical features:
  - longitude
  - latitude
  - housing\_median\_age
  - total\_rooms
  - population
  - households
  - median\_income

A 70–30 train-test split was used for evaluation.

### 2. Single-Feature Linear Regression

Each numerical feature was independently used to fit a simple linear regression model. StandardScaler was applied before fitting.

#### Results (Test Set)

**Best-performing feature for all error metrics:**

Error Metric	Best Feature	Value
SSE	median_income	$4.2836 \times 10^{13}$
MSE	median_income	$6.9179 \times 10^9$
MAE	median_income	62,315

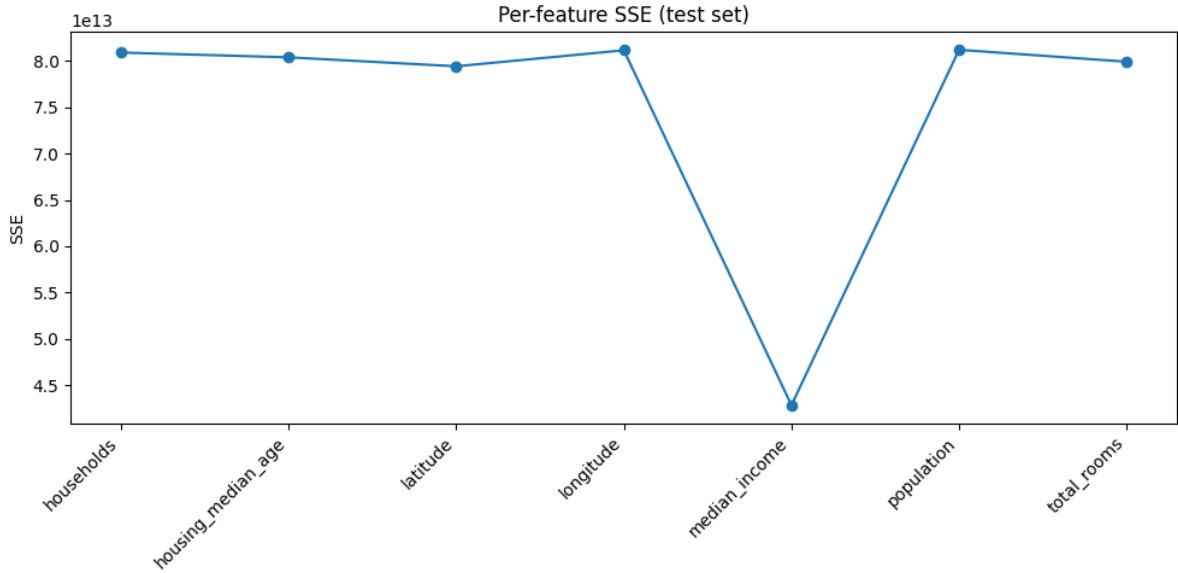
Median income clearly provides the lowest error across the board.

## Interpretation

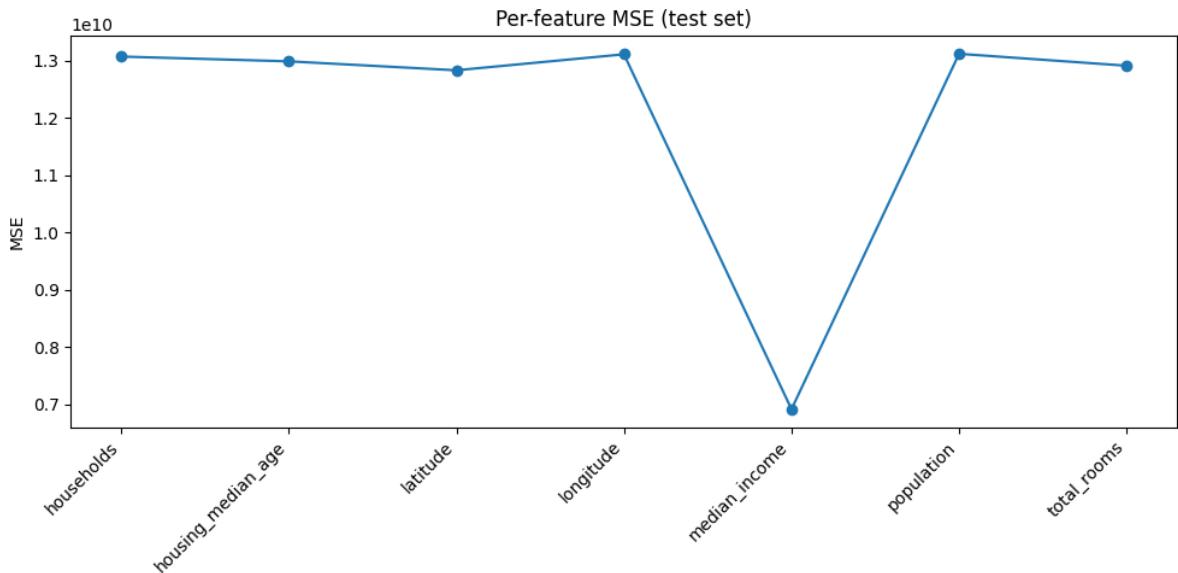
This shows that **median\_income** is the most influential single predictor of housing value. The next best features (total\_rooms, latitude) perform far worse.

## Plots Included

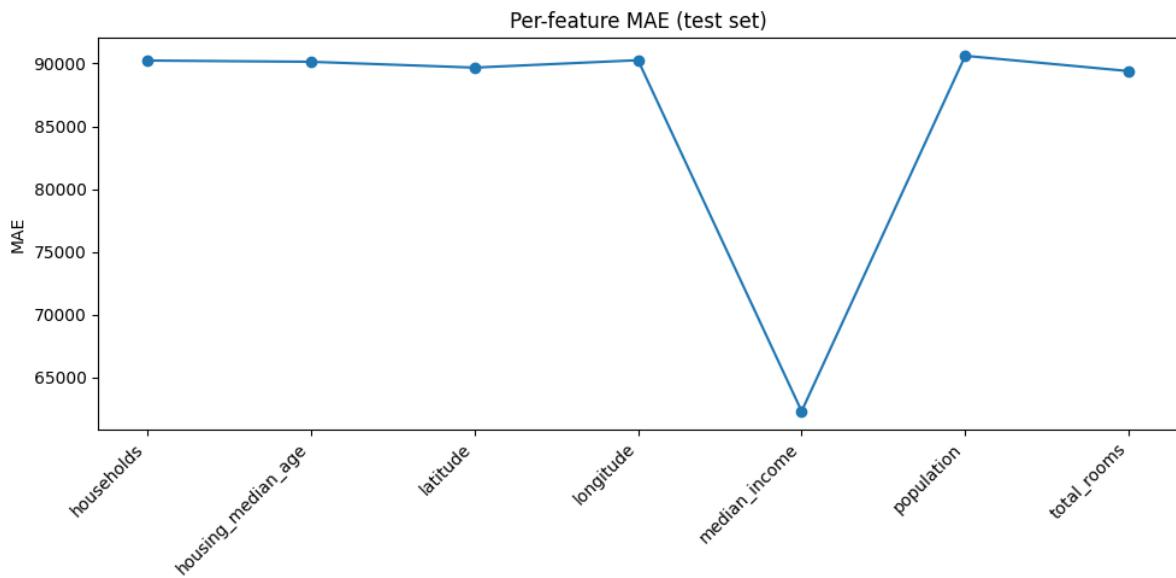
- SSE vs feature index



- MSE vs feature index



- MAE vs feature index



These plots show a dramatic drop for median\_income, confirming its dominance.

### 3. Multivariate Linear Regression (All Features Together)

The model was trained using all 7 numerical features simultaneously.

#### Performance (Test Set):

- **SSE:**  $2.976 \times 10^{13}$
- **MSE:**  $4.806 \times 10^9$
- **MAE:** 51,044

#### Interpretation

Compared to the best single feature (median\_income):

Model	MSE	MAE
Best single feature	$6.917 \times 10^9$	62,315
<b>Multivariate model</b>	<b><math>4.806 \times 10^9</math></b>	<b>51,044</b>

The multivariate model significantly improves prediction accuracy.

## Learned Coefficients

Feature	Coefficient
longitude	-83,676
latitude	-89,564
housing_median_age	+14,487
total_rooms	-3,539
population	-48,079
households	+56,213
median_income	+73,197

Median income still has the strongest positive coefficient, confirming its importance.

## 4. Regularization (5-Fold Cross-Validation)

Ridge and Lasso regression were applied to address potential overfitting and evaluate the effect of regularization strength.

### Ridge Regression

- Best alpha: 11.51
- Mean MSE:  $4.909 \times 10^9$
- Std:  $2.853 \times 10^8$

### Lasso Regression

- Best alpha: 35.56
- Mean MSE:  $4.909 \times 10^9$
- Std:  $2.855 \times 10^8$

### Interpretation

The optimal MSE for both Ridge and Lasso is **slightly higher** than the unregularized multivariate linear regression.

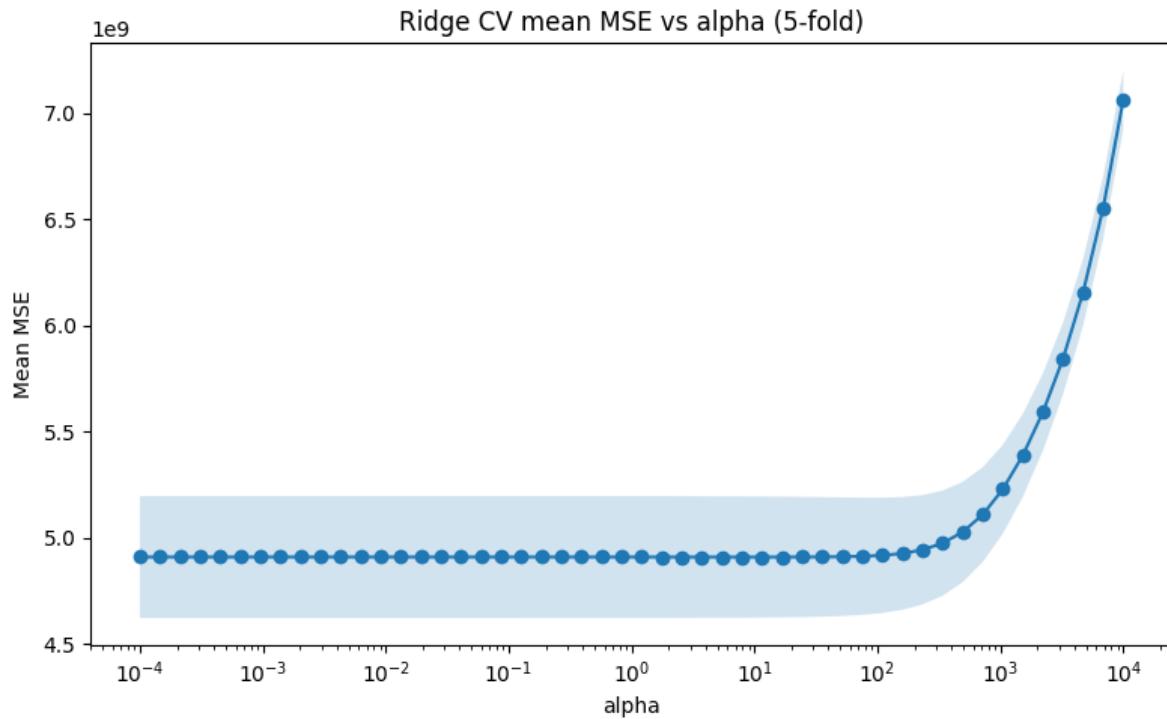
This indicates:

- The model **does not suffer from severe overfitting**.

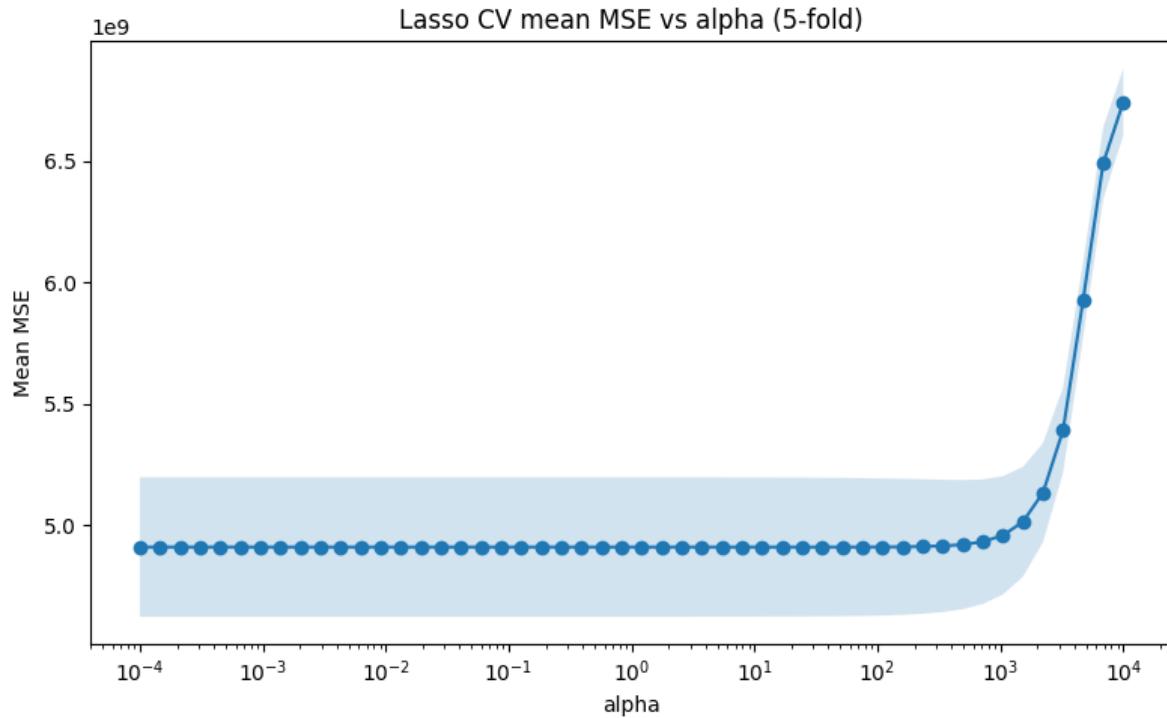
- Regularization does not substantially improve performance on this dataset.
- Ridge performs marginally better than Lasso.

Plots included:

- Ridge CV MSE vs  $\alpha$  (log scale)



- Lasso CV MSE vs  $\alpha$  (log scale)



## 5. Conclusions

### Most Important Feature

Across all experiments, **median\_income** is clearly the most important predictor of housing values.

### Best Model

- The **multivariate linear regression** (without regularization) performed best overall.
- Regularization did not improve performance significantly.

### Final Recommendations

- Include **all numeric features** for best predictive accuracy.
- **Median income** should be highlighted as a primary driver in prediction tasks.
- Regularization can be used for stability, but it does not enhance accuracy for this dataset.