# Analysis of Principal Component Analysis (PCA) on Housing Price Prediction

September 19, 2025

## Introduction

Predicting real estate prices requires modeling complex relationships between numerous features. This report evaluates the impact of Principal Component Analysis (PCA), a dimensionality reduction technique, on a Linear Regression model's ability to predict median house values using the Boston Housing dataset. The goal is to compare the performance of a standard Linear Regression model against one trained on a PCA-reduced feature set.

## Dataset

This analysis uses the Boston Housing dataset (506 instances, 14 attributes) from boston _housing.csv. The 13 features include crime rate (CRIM), property characteristics (e.g., RM, AGE), and socio-economic factors (e.g., LSTAT,PTRATIO). The target variable is the median home value (MEDV).

## Algorithms

Linear Regression: A supervised learning algorithm that models the linear relationship between features and a target variable. It served as the baseline prediction model.

Principal Component Analysis (PCA): An unsupervised dimensionality reduction technique that transforms correlated features into a smaller set of uncorrelated "principal components." It was used to reduce the 13 input features to 10 principal components.

## Implementation and Results

The data was split into training (80%) and testing (20%) sets, and all features were standardized. PCA was then applied to the scaled training data to generate 10 principal components. Two Linear Regression models were trained: a baseline model on all 13 features and a second on the 10 PCA components. Both were evaluated on the test set using Mean Squared Error (MSE) and R-squared ($R^2$).

Results:

Linear Regression (Without PCA):

Mean Squared Error (MSE): 24.29

R-squared ($R^2$): 0.67

Linear Regression (With PCA):

Mean Squared Error (MSE): 28.69

R-squared ($R^2$): 0.61

The baseline model performed better, explaining 67% of the price variance. The PCA-based model's performance was slightly lower (explaining 61% of variance) due to the expected information loss from dimensionality reduction.