Mini Report:

1.Objective:

The objective of this project is to build and evaluate a Linear Regression model to predict housing prices (MEDV) using key features from the Boston Housing dataset. Feature selection was performed using correlation analysis to identify the most influential variables.

2.Data Summary:

The Boston Housing dataset contains information about housing in Boston suburbs. It includes 506 rows and 14 columns. The target variable is `MEDV` (Median value of owner-occupied homes in $1000s).\

3.Feature Selection (Correlation analysis):

Using a correlation matrix and heatmap, we selected the top 5 features most correlated with `MEDV`:

LSTAT (Negative correlation),RM (Positive correlation),PTRATIO,INDUS,TAX

4.Model Building:

We used a Linear Regression model from `scikit-learn`.

The dataset was split into 80% training and 20% testing using `train\_test\_split()`.

5.Evaluation metrics:

MSE (Mean Squared Error):30.12

RMSE (Root Mean Squared Error): 5.49

R² Score:0.68

The R² score indicates that ~68% of the variance in house prices is explained by the model. The RMSE of 5.49 suggests an average prediction error of $5,490.

6.Insights and Observation:

LSTAT (lower status population) has the strongest negative impact on housing prices.

RM (average rooms per dwelling) positively affects the price.

Although the model performs reasonably well, a nonlinear model or feature engineering might improve results.

7.Conclusion:

The Linear Regression model gives a good starting point for predicting Boston house prices using only 5 features. While the model shows moderate performance, there’s potential to improve accuracy with advanced models like Random Forest or Gradient Boosting, and by addressing outliers or multicollinearity.