Simple Linear Regression Analysis Report

Project Supervisor:

Dr. Amna Zafar

Project Members

Gul-e-Zahra 2022-CS-75

Department of Computer Science
University of Engineering and Technology
Lahore, Pakistan

1. Dataset Preparation

For this analysis, we used the California Housing Prices dataset, which is suitable for simple linear regression. The dataset was loaded using Pandas, and the first five rows were displayed to understand its structure and values. Below is a preview of the dataset:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms
0	-122.23	37.88	41.0	880.0	129.0
1	-122.22	37.86	21.0	7099.0	1106.0
2	-122.24	37.85	52.0	1467.0	190.0
3	-122.25	37.85	52.0	1274.0	235.0
4	-122.25	37.85	52.0	1627.0	280.0

2. Data Visualization

To visualize the relationship between the independent variable (`median_income`) and the dependent variable (`median_house_value`), a scatter plot was created using Matplotlib. The plot revealed a positive correlation, indicating that higher median incomes are associated with higher median house values.

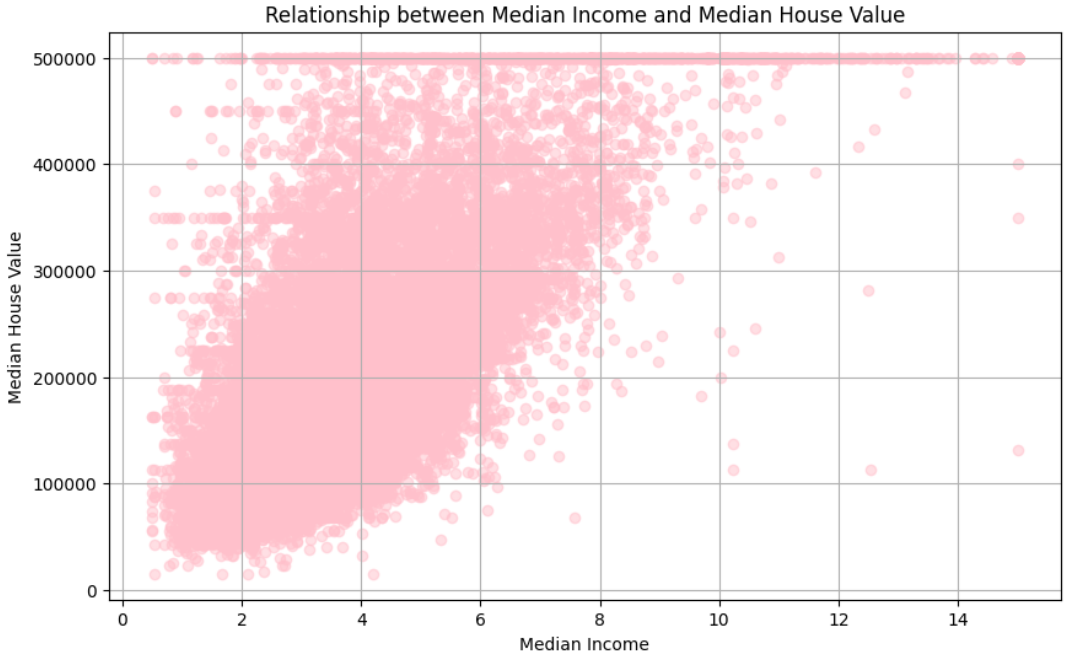

Figure 1: Scatter Plot of Median Income vs. Median House Value

3. Implementing Simple Linear Regression

Using Scikit-learn, we implemented a simple linear regression model. The dataset was split into training (80%) and testing (20%) sets. The model was trained on the training set, and predictions were made on the test set.

4. Loss Function Calculation

The Mean Squared Error (MSE) was used as the loss function to evaluate model performance. The MSE was calculated for both the training and testing sets:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Training MSE:** 6991447170.182823
- **Testing MSE:** 7091157771.76555

The loss function represents the average squared error between the actual and predicted values, giving an indication of the model's accuracy.

5. Cost Function Analysis

Difference Between Loss and Cost Functions

- The loss function measures the error for a single data point.
- The cost function aggregates the errors across the entire dataset, providing a global assessment of the model's performance.

Cost Function Calculation

The cost function, calculated as the Mean Squared Error (MSE) across the entire dataset, was computed for both training and testing datasets. This helps to evaluate the model's generalization performance.

How Cost Function Helps Understand Model Performance

- A low MSE indicates accurate predictions for both training and testing datasets.
- A high MSE suggests underfitting, while a significant difference between training and testing MSE may indicate overfitting.

6. Model Evaluation

To evaluate the model, we calculated the R-squared value and Mean Absolute Error (MAE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Mean Squared Error (Cost Function) on Training Set:** 6991447170.182823
- **Mean Squared Error (Cost Function) on Testing Set:** 7091157771.76555

These metrics indicate the accuracy of the model and how well it fits the data.

7. Linear Regression: Actual vs Predicted

The following plot visualizes the relationship between the actual and predicted values of the dependent variable (median house value) from the test set:

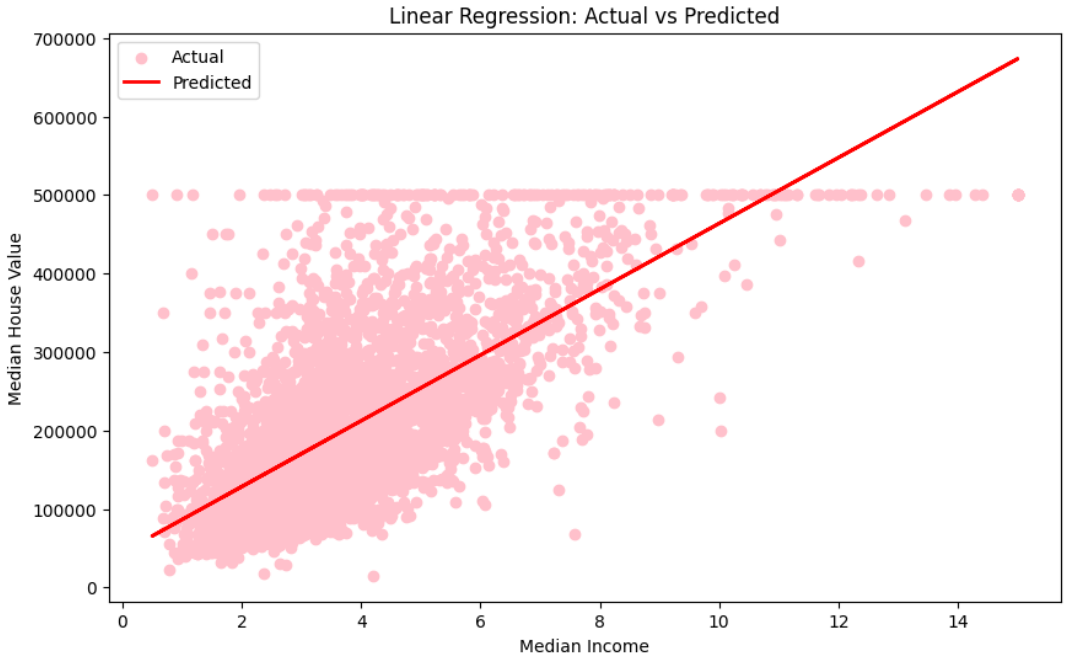

Figure 2: Linear Regression: Actual vs Predicted Values

This plot helps to visually assess the performance of the regression model by comparing the predicted values with the actual values.

Conclusion

The simple linear regression model showed a positive relationship between median income and median house value. The metrics (MSE, R-squared, and MAE) demonstrated that the model could explain a significant portion of the variance in the data. However, further refinement and feature engineering could enhance the model's performance.