Simple Linear Regression Report

1. Dataset Overview

Name: Krunal Dhapodkar PRN: 22070521006 Semester: 7 Section: A
Dataset Explanation

The dataset, loaded from `ENB2012_data[1].xlsx`, consists of 768 entries and 10 columns. All columns are non-null, indicating no missing values that would require imputation.

Column Descriptions:

- X1: Relative Compactness (float64)
- X2: Surface Area (float64)
- X3: Wall Area (float64)
- X4: Roof Area (float64)
- X5: Overall Height (float64)
- X6: Orientation (int64)
- X7: Glazing Area (float64)
- X8: Glazing Area Distribution (int64)

These `X` columns represent various building characteristics and are likely the independent variables (features) for our regression analysis.

- Y1: Heating Load (float64)
- Y2: Cooling Load (float64)

These `Y` columns represent the building's energy performance and are likely the dependent variables (targets) for our regression analysis. For the simple linear regression, we will typically choose one of these as the target variable.

Key Observations from Descriptive Statistics:

- Numerical Data Types: Most columns are `float64`, with `X6` (Orientation) and `X8` (Glazing Area Distribution) being `int64`. These integer columns might represent categorical features or discrete numerical values.
- Range of Values: The `min` and `max` values for each `X` column show a varied range, indicating different scales for each feature. For instance, `X1` (Relative Compactness) ranges from 0.62 to 0.98, while `X2` (Surface Area) ranges from 514.5 to 808.5. This diversity suggests that feature scaling might be beneficial if we were to use certain types of models, though it's less critical for simple linear regression with a single predictor.
- Target Variables (Y1, Y2): Both `Y1` (Heating Load) and `Y2` (Cooling Load) also show a significant range, from approximately 6 to 43 for Y1, and 10 to 48 for Y2. This wide range indicates variability in the energy loads, which our regression model will attempt to predict.

This initial exploration confirms that the dataset is clean (no missing values) and provides a good understanding of the variables available for simple linear regression.

2. Model Findings and Performance

Findings from Simple Linear Regression

Simple Linear Regression Report

Model Parameters:

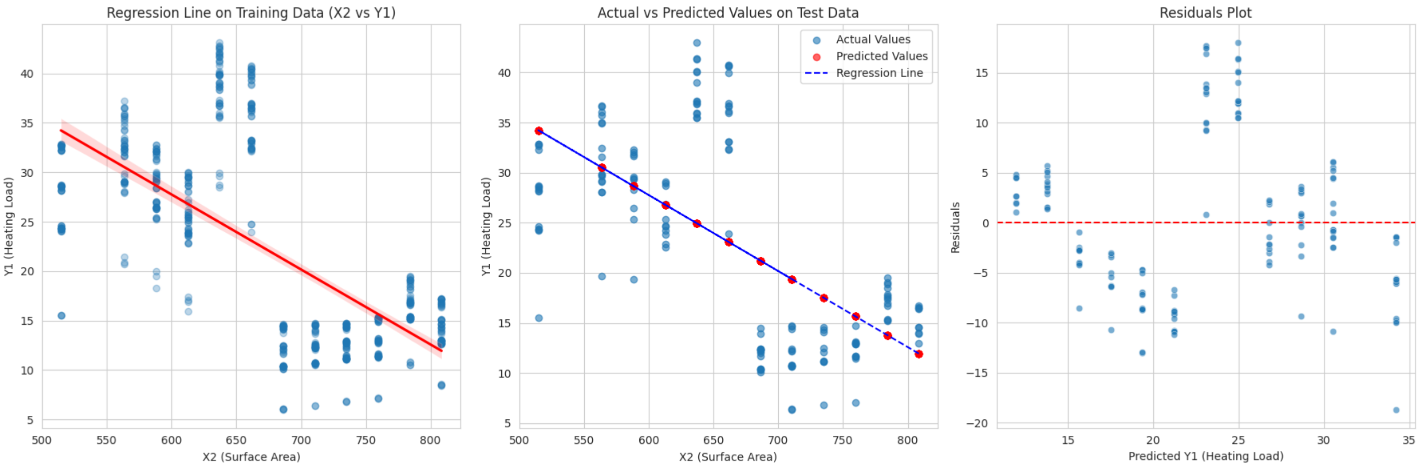
- Intercept: The intercept of the linear regression model is approximately `73.23`. This represents the predicted Heating Load (Y1) when the Surface Area (X2) is zero. In practical terms, this might not be directly interpretable as a physical building cannot have zero surface area, but it's a necessary component of the linear equation.
- Coefficient (Slope): The coefficient for Surface Area (X2) is approximately `-0.076`. This indicates that for every one-unit increase in Surface Area, the Heating Load (Y1) is predicted to decrease by `0.076` units, assuming all other factors are constant. This inverse relationship suggests that larger surface areas, in the context of this dataset's characteristics, tend to be associated with slightly lower heating loads.

Model Performance on Test Data:

The model was evaluated on the test set, and the following metrics were obtained:

- Mean Absolute Error (MAE): `6.46`. This means, on average, the absolute difference between the model's predictions and the actual Heating Load values is 6.46 units.
- Mean Squared Error (MSE): `62.85`. This metric penalizes larger errors more heavily than MAE. It is useful for understanding the variance of the errors.
- Root Mean Squared Error (RMSE): `7.93`. RMSE is the square root of MSE and is in the same units as the target variable (Heating Load). It gives an idea of the typical magnitude of the errors.
- R-squared (R2): `0.40`. The R-squared value indicates that approximately 40% of the variance in the Heating Load (Y1) can be explained by the Surface Area (X2) in this simple linear regression model. This suggests that while Surface Area has some predictive power, a significant portion of the variance in Heating Load is not explained by this single feature, implying that other factors (or more complex models) might be needed for a more accurate prediction.

3. Visualizations


Regression Line on Training Data (X2 vs Y1)
Y1 (Heating Load)
X2 (Surface Area)
Actual vs Predicted Values on Test Data
Actual Values
Predicted Values
Regression Line
Y1 (Heating Load)
X2 (Surface Area)
Residuals Plot
Residuals
Predicted Y1 (Heating Load)