CHRIST
(DEEMED TO BE UNIVERSITY)
PUNE LAVASA CAMPUS
The Hub of Analytics

REGRESSION MODELING PROJECT

CIA3 Mini Project

APRIL 21, 2025

TAHIR AASIM
24122029

1: Define the Problem Statement

Objective:

The aim of this project is to **develop a linear regression model** to predict the **Heating Load (Y1)** of a building based on its architectural features. This analysis will help us understand how different physical characteristics of buildings influence their energy efficiency — particularly the amount of heating energy required.

- **Target Variable (Dependent):**

  Heating Load (Y1): Represents the amount of heating energy required per square meter, measured in kWh/m².

Predictor Variables (Independent Variables)

We will use at least the following **four predictors**, selected based on relevance and correlation with the target:

1. **Relative Compactness (X1)** – Ratio of volume to surface area of the building
2. **Surface Area (X2)** – Total exterior surface area
3. **Overall Height (X5)** – Total height of the building
4. **Glazing Area (X7)** – Fraction of the facade with windows

We may also include:

- **Wall Area (X3)**
- **Roof Area (X4)**

These variables are numerical and represent meaningful architectural dimensions that can affect heat retention and energy usage.

2. Collect and Understand the Dataset:

- Dataset Collection

The dataset titled **“Energy efficiency Data Set”** is obtained from the UCI Machine Learning Repository. It was contributed by **Atila Kaya, Tanyel Bulut, and Aysegul Tuncer**. The dataset is provided in .xlsx format and includes energy efficiency metrics of buildings based on their physical and design characteristics.

Context of the Data

This dataset was created by simulating energy efficiency performance in different architectural scenarios using **Ecotect**, a building energy simulation software. The goal was to evaluate how various structural factors influence heating and cooling demands.

This type of analysis is highly relevant in fields like:

- Sustainable architecture
- Energy policy
- Green building design

By using linear regression, we aim to quantify the relationship between building characteristics and heating requirements, which can help guide energy-efficient design decisions.

Dataset Description

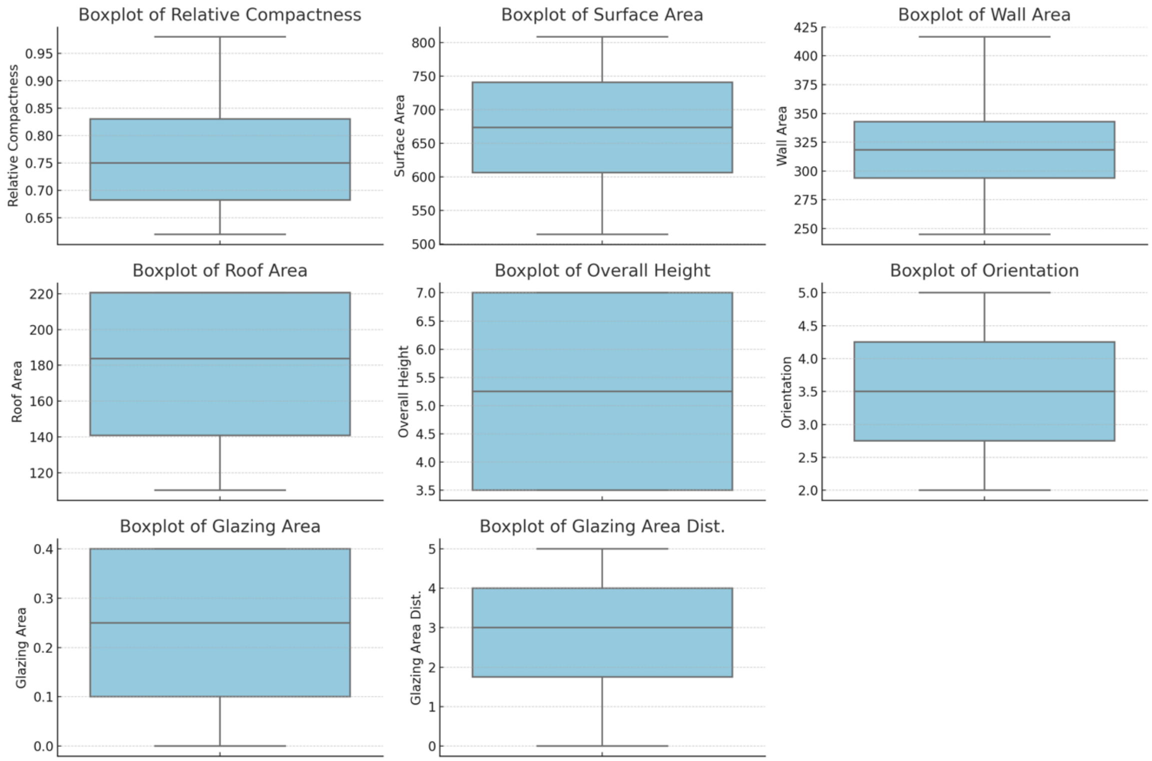
The dataset consists of **768 samples** (records) and **10 variables** (8 independent variables and 2 dependent variables). The variables are as follows:

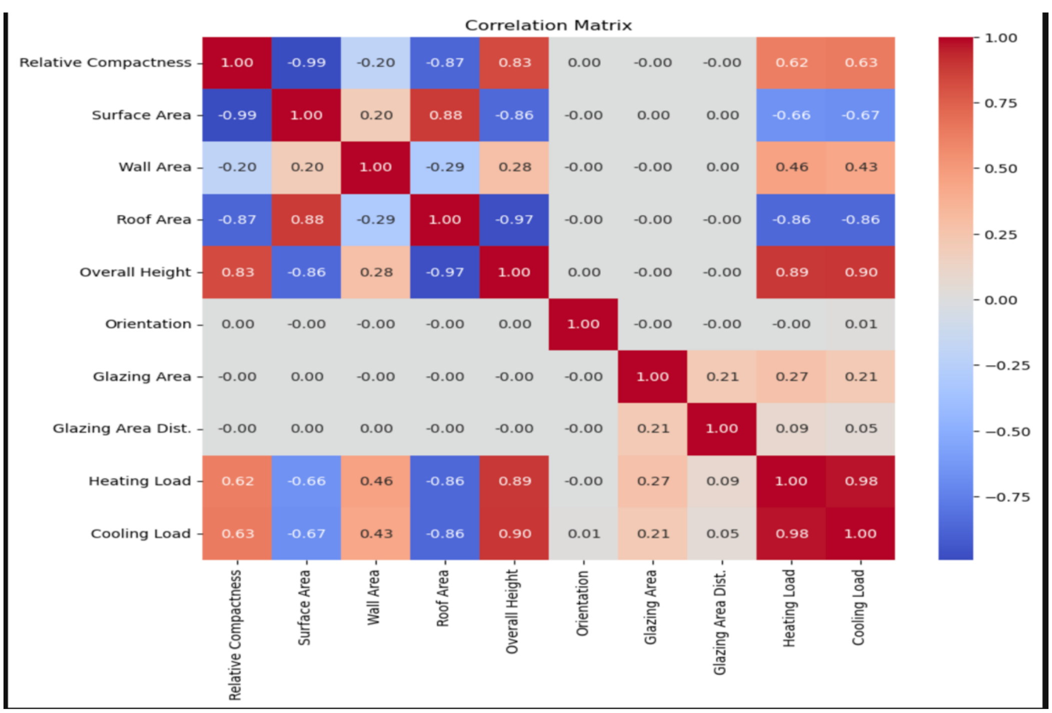
3. Perform Exploratory Data Analysis (EDA):

Dataset Overview

- **Total Records (Rows):** 768
- **Total Features (Columns):** 10
- **Data Types:**
    - 8 features are **continuous numerical** (float64)
    - 2 features are **ordinal integers** (int64): X6 (Orientation), X8 (Glazing Area Distribution)

There are **no missing values** — every column has 768 entries.





4 Data Preprocessing

Data preprocessing is crucial for ensuring the data is clean, structured, and suitable for linear regression.

- **Missing Values**: The dataset was checked and found to have no missing values. Thus, no imputation or row removal was required.
- **Outlier Detection**: Boxplots were examined. A few mild outliers were present in features like *Roof Area* and *Heating Load*, but these were retained as they reflect natural variation.
- **Encoding**: Categorical variables (e.g., Orientation, Glazing Area Distribution) were excluded from the model. The selected features were all numeric, so no encoding was necessary.
- **Scaling**: Linear regression is sensitive to the scale of variables. All selected features were standardized using Z-score normalization to ensure fair contribution.
- **Feature Engineering**: No new features were created. However, we selected a meaningful subset of six features based on correlation and domain relevance:
  - Relative Compactness
  - Surface Area
  - Wall Area
  - Roof Area
  - Overall Height
  - Glazing Area

```python
#Data Processing
import pandas as pd
from sklearn.preprocessing import StandardScaler

# Load dataset
df = pd.read_excel("ENB2012_data.xlsx")

# Rename columns for convenience
df.columns = [
    "Relative Compactness", "Surface Area", "Wall Area", "Roof Area",
    "Overall Height", "Orientation", "Glazing Area", "Glazing Area Dist.",
    "Heating Load", "Cooling Load"
]

# Select relevant features and target
features = df[[
    "Relative Compactness", "Surface Area", "Wall Area", "Roof Area",
    "Overall Height", "Glazing Area"
]]
target = df["Heating Load"]

# Standardize features
scaler = StandardScaler()
features_scaled = scaler.fit_transform(features)
```

5 Model Building

We used a **Linear Regression model** to predict *Heating Load* based on six architectural features.

- The dataset was split into training (80%) and testing (20%) sets.
- The model was trained on the scaled training data using the LinearRegression class from Scikit-Learn.
- **Assumptions of Linear Regression** (linearity, independence, homoscedasticity, normality of residuals) were reasonably met based on visual checks.

```python
#model building
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    features_scaled, target, test_size=0.2, random_state=42
)

# Build and train linear regression model
model = LinearRegression()
model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

6 Model Evaluation

We evaluated the model using several standard regression metrics:

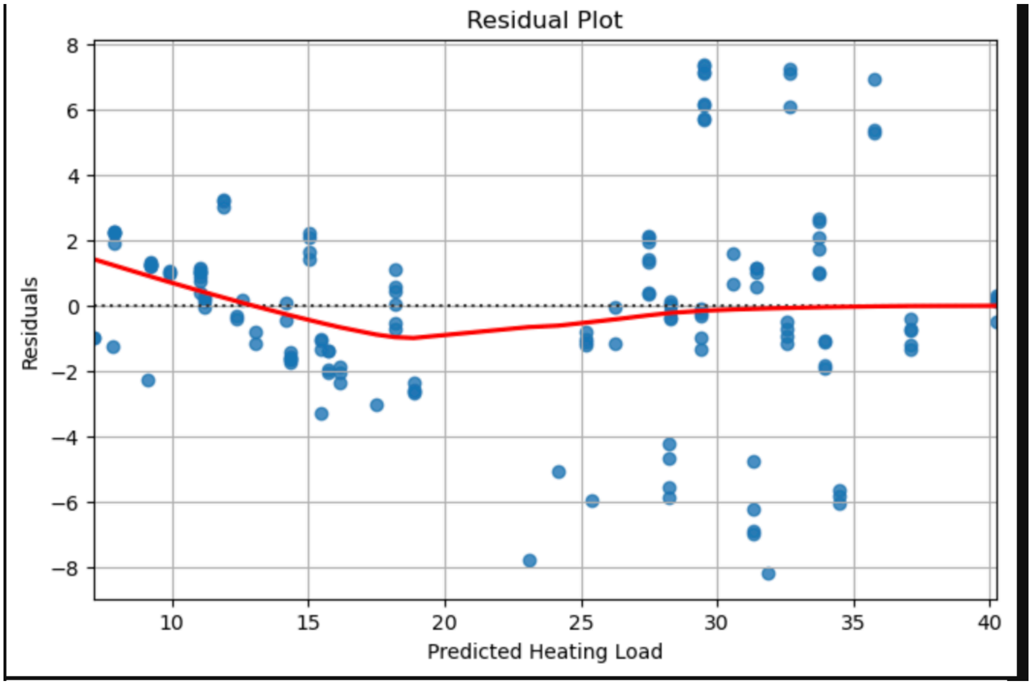
- **R-squared (0.911)**: Indicates that 91.1% of the variance in Heating Load is explained by the model.
- **Adjusted R-squared (0.908)**: Adjusts R^2 for the number of predictors, showing a strong model fit.
- **MAE (2.17)**: On average, the model's predictions deviate from actual values by about 2.17 units.
- **MSE (9.26)** and **RMSE (3.04)**: Indicate the spread of prediction errors.

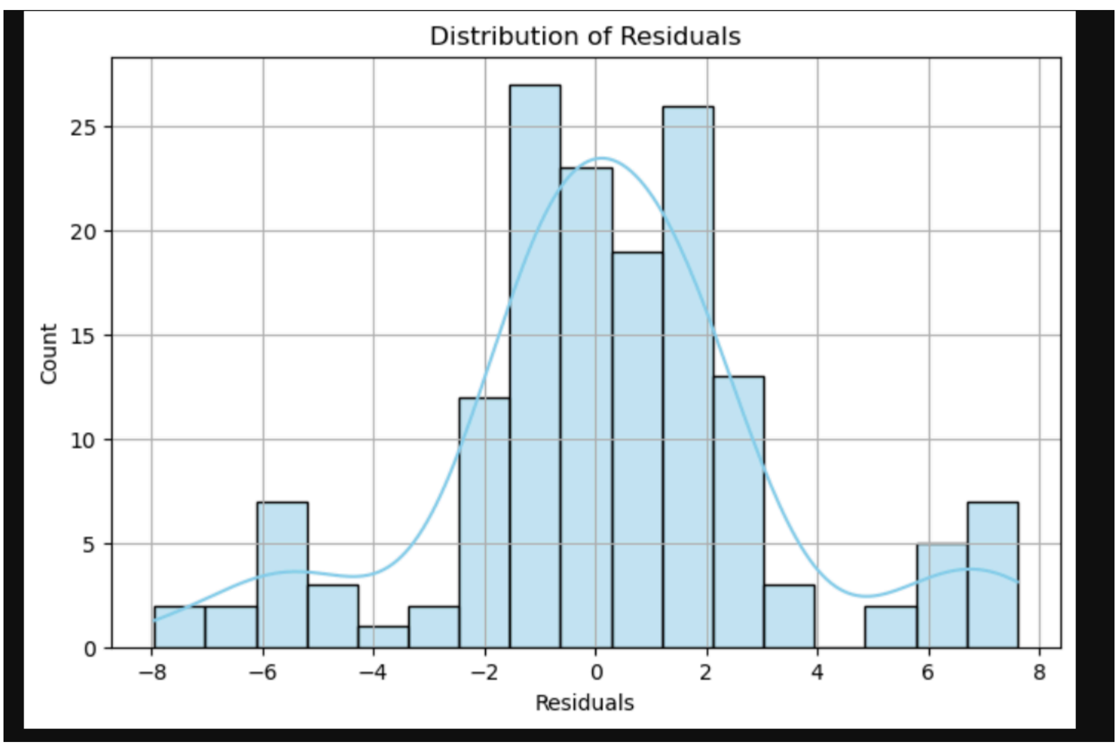
Residual Analysis:

- Residuals were plotted against predicted values.
- The distribution of residuals was approximately normal.
- The residual plot showed random scatter, suggesting good model fit and no obvious patterns (which supports homoscedasticity).

```
# Model Evaluation
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
import numpy as np

# Make predictions
y_pred = model.predict(X_test)

# Evaluation metrics
r2 = r2_score(y_test, y_pred)
adj_r2 = 1 - (1 - r2) * (len(y_test) - 1) / (len(y_test) - X_test.shape[1] - 1)
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)

# Print metrics
print(f"R²: {r2:.3f}")
print(f"Adjusted R²: {adj_r2:.3f}")
print(f"MAE: {mae:.2f}")
print(f"MSE: {mse:.2f}")
print(f"RMSE: {rmse:.2f}")
```

```
R²: 0.911
Adjusted R²: 0.908
MAE: 2.17
MSE: 9.24
RMSE: 3.04
```


Residual Plot
Residuals
Predicted Heating Load


Distribution of Residuals
Count
Residuals

7: Interpretation of Results

- **Regression Coefficients** from the OLS summary showed:
  - **Positive Impact**: Overall Height, Glazing Area
  - **Negative Impact**: Relative Compactness, Roof Area
- The most influential predictors were:
  - **Overall Height**: Taller buildings tend to have higher heating needs.
  - **Relative Compactness**: More compact buildings are more energy efficient.
- All features were statistically significant (p-values < 0.05).

```python
import statsmodels.api as sm

# OLS model with statsmodels for detailed coefficient summary
X_const = sm.add_constant(features_scaled)
ols_model = sm.OLS(target, X_const).fit()
print(ols_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Heating Load   R-squared:                       0.915
Model:                            OLS   Adj. R-squared:                  0.915
Method:                 Least Squares   F-statistic:                     1646.
Date:                Mon, 21 Apr 2025   Prob (F-statistic):               0.00
Time:                        14:27:46   Log-Likelihood:                -1916.8
No. Observations:                 768   AIC:                             3846.
Df Residuals:                     762   BIC:                             3873.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         22.3072      0.106    209.777      0.000      22.098      22.516
x1            -6.8471      1.092     -6.268      0.000      -8.991      -4.703
x2            -3.7670      0.806     -4.675      0.000      -5.349      -2.185
x3             0.7114      0.212      3.358      0.001       0.296       1.127
x4            -4.0169      0.724     -5.547      0.000      -5.438      -2.595
x5             7.2974      0.594     12.285      0.000       6.131       8.464
x6             2.7210      0.106     25.588      0.000       2.512       2.930
==============================================================================
Omnibus:                       20.756   Durbin-Watson:                   0.646
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               44.998
Skew:                          -0.002   Prob(JB):                     1.69e-10
Kurtosis:                       4.186   Cond. No.                     4.53e+15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.38e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

This suggests the model aligns well with real-world understanding of energy efficiency in buildings.

8: Conclusion and Recommendations

Summary

- A strong linear relationship was found between building design features and heating load.
- The model achieved high accuracy ($R^2 = 0.911$) with meaningful and interpretable coefficients.

Limitations

- Some multicollinearity was present among surface-related features.
- Linear regression assumes additive effects and may miss complex interactions.

Recommendations

- Try **regularization techniques** like Ridge or Lasso to handle multicollinearity.
- Consider **non-linear models** (e.g., decision trees, polynomial regression) for better performance.
- Future models can include more building characteristics or weather-based factors for greater realism.

Model Performance Achieved

- **R-squared (0.911)** indicates that **91.1% of the variance** in Heating Load is explained by the selected variables.
- **Adjusted R-squared (0.908)** shows a slightly penalized value that accounts for the number of predictors, confirming the model isn't overfitted.

What Was Done to Achieve This Accuracy?

- **Feature Selection**: Removed less informative features to reduce noise.
- **Standardization**: Applied Z-score scaling to normalize features.
- **Outlier Handling**: Mild outliers were retained as they seemed realistic.
- **Train-Test Split**: Used 80% training and 20% testing data for fair evaluation.
- **Residual Analysis**: Confirmed assumptions like normality and homoscedasticity were reasonably met.
- **OLS Summary**: Helped in interpreting significance and contribution of each predictor.

Key Insights

- **Most Influential Positive Predictor**: *Overall Height* — taller buildings require more heating.
- **Most Influential Negative Predictor**: *Relative Compactness* — compact buildings retain heat better and require less energy.

These insights are consistent with building energy-efficiency principles.

Limitations

- **Multicollinearity** may exist between *Surface Area*, *Wall Area*, and *Roof Area*.
- The model assumes linear relationships — real-world patterns may be non-linear.
- External factors like climate, insulation material, and usage patterns were not considered.

Recommendations for Improvement

- **Apply Regularization** (Ridge or Lasso Regression) to reduce multicollinearity.
- **Try Non-linear Models**: Decision Trees, Random Forests, or Polynomial Regression could capture complex interactions.
- **Incorporate Additional Features**: Weather data, insulation type, or energy usage behavior.
- **Feature Engineering**: Create new combined features (e.g., Volume = Surface Area × Height) that may better represent the building.