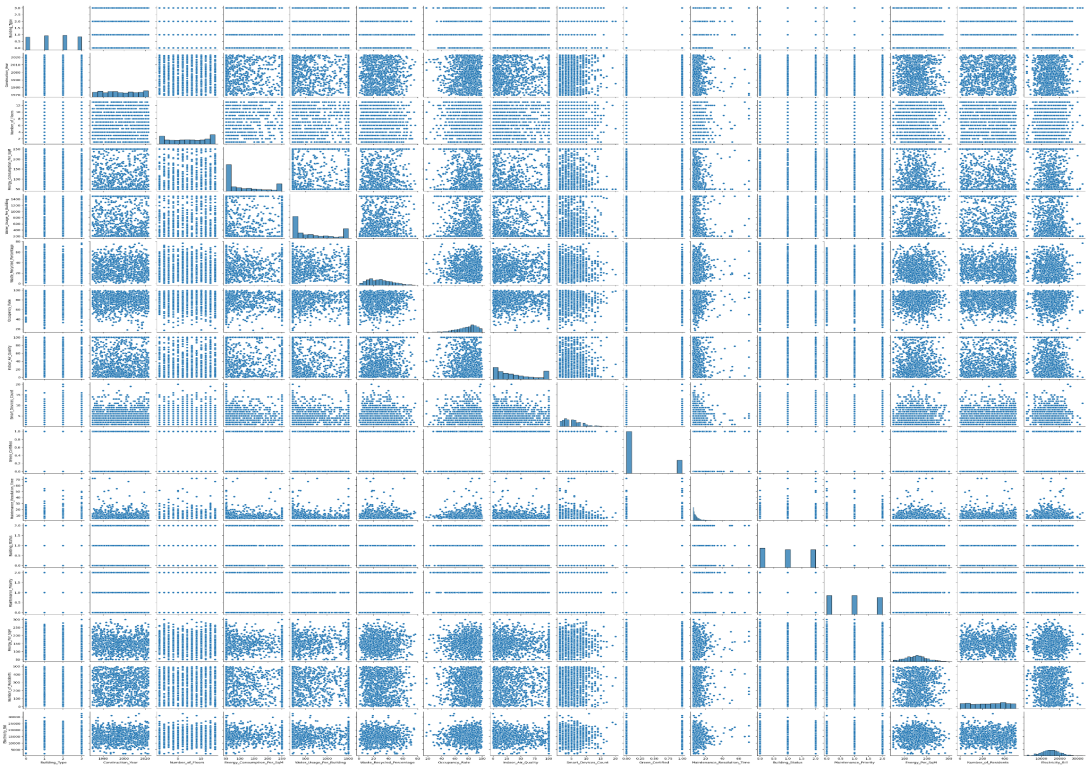
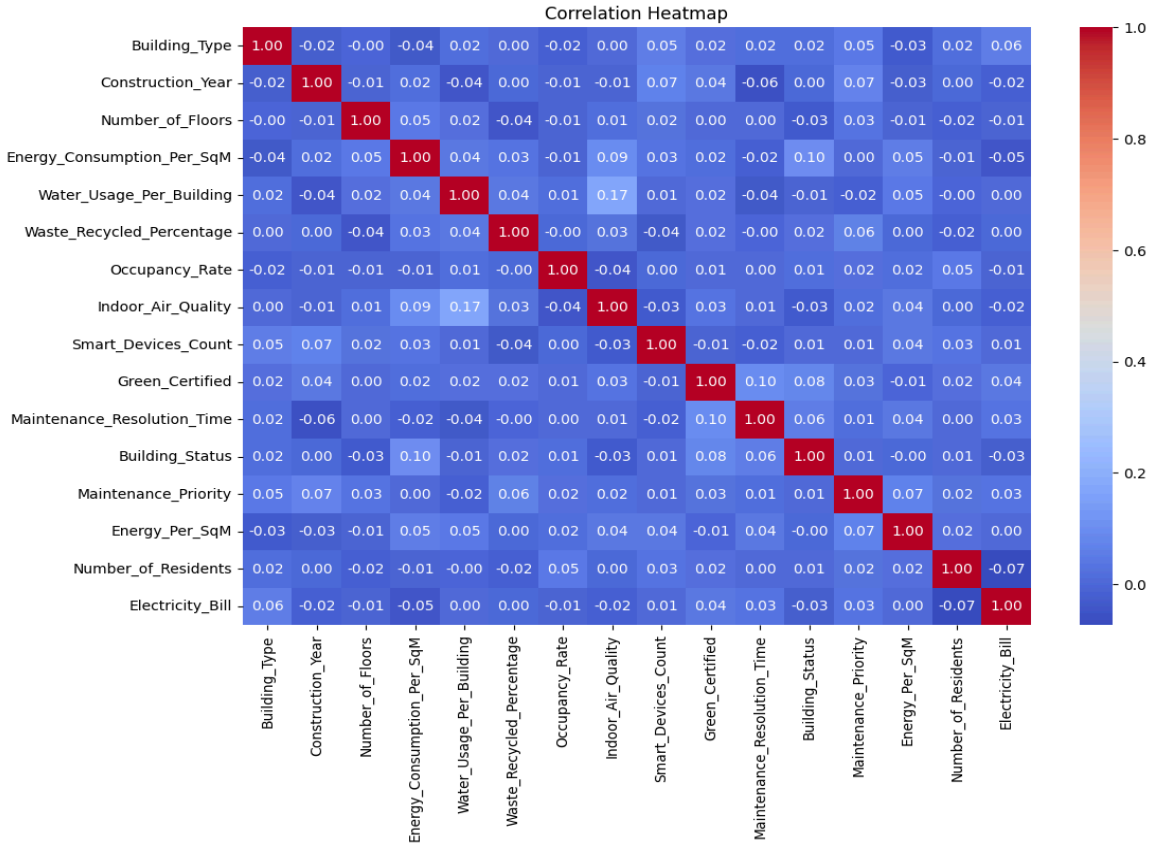




PS: Did’nt copy violin plot and box plots

1> Strong Positive Correlation Between Energy Consumption Per SqM and Energy Per SqM (0.87):

These two variables are very highly related, as in both cases, they deal with energy usage. This strong correlation sets up a scenario where optimization in one variable will most probably affect the other variable. It is vital in the management of energy-efficient buildings.

2> Moderate Positive Correlation Between Maintenance Priority and Time to Resolve Maintenance:

There is a tendency for buildings with higher priority maintenance tasks to take more time in resolving times. That may mean that the more crucial issues take time to get fixed, and that can be useful in resource planning and staffing of the maintenance teams.

3> IAQ and Energy Consumption Per SqM is positively correlated at 0.09:

The buildings with good indoor air quality might also consume more energy, probably because the HVAC system runs for maintaining good air quality inside the building. So, this provides further scope for improvement to the building management in ensuring a better energy efficiency without degrading the air quality.

4>Occupancy rate and Indoor Air Quality Positive Correlation: 0.09
It also means that with a higher occupancy rate, the indoor air quality is likely to be better. This would, therefore, mean that a higher range of occupants will be attracted to those buildings with well-ventilated or well-managed air quality systems, to which it greatly enhances the desirability of the overall building.

5>Number of Residents and Electricity Bill are Weakly Negatively Correlated (-0.07):
More surprisingly, more residents do not result in higher bills. This perhaps suggests effective use of energy behaviors and technologies within buildings with more residents, where cost-saving strategies may be located.

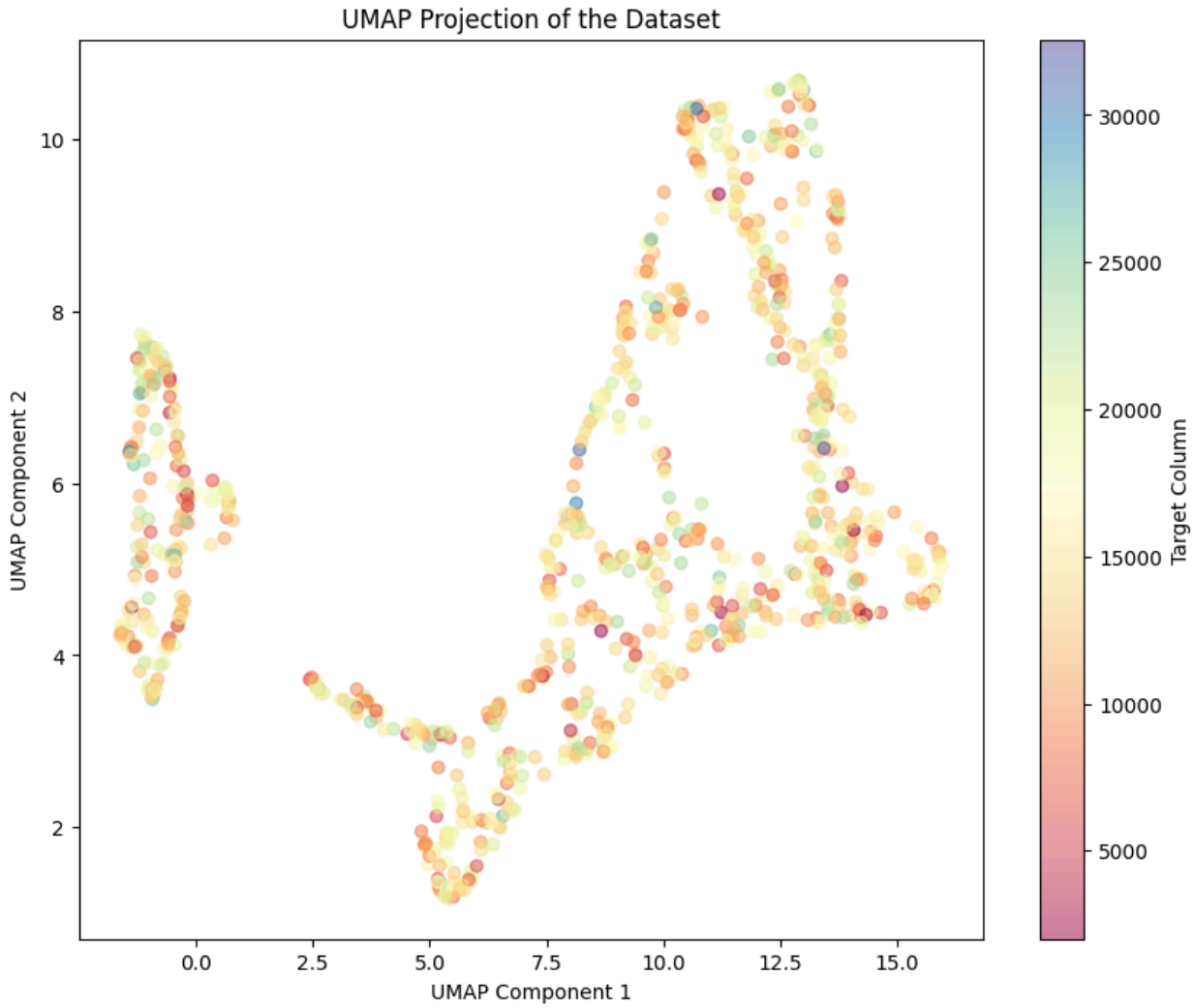
6>Smart Devices Count and Construction Year are Positively Correlated (0.07):
Newer buildings have a few more smart devices, which would indicate a general increasing trend in the integration of smart technology in more recently built buildings and may further lead to higher efficiency and automation.

7>Green Certification and Maintenance Resolution Time Are Weakly Negatively Correlated (-0.01):
Green-certified buildings also demonstrate marginally better upkeep resolution times. In such buildings, it may mean that the necessary maintenance tasks associated with maintaining them are done better, maybe due to more modern or streamlined infrastructure.

8>Water Usage Per Building and Indoor Air Quality lightly positive correlation of 0.17:
Buildings with better indoor air quality also tend to consume more water. This could be explained by water-based cooling or humidity control in HVAC systems. This would point toward a possible trade-off between air quality and water efficiency.

9>Building Type and Smart Devices Count Have a Weak Positive Correlation (0.05):
There is a weak positive correlation between certain building types and the installed smart devices. This might imply that for certain purposes of buildings, say office space or residential areas, the smart technology is more acceptable.

10>Electricity Bill vs. Number of Floors Weak Negative Correlation (-0.01):
No need for electricity bills in taller buildings with more floors may suggest that the energy distribution systems of multi-storey buildings are efficient, or at scale, there is better management of energy use.



The UMAP projection reveals moderate clustering with some degree of separability, though certain clusters exhibit overlap. A distinct cluster appears on the left, while the right side contains more scattered groups. The color gradient suggests that the target variable is smoothly distributed across these clusters. Despite capturing the overall structure, UMAP shows areas of overlap, indicating that the data relationships are complex and not perfectly separable in two dimensions.

Pre-Processing Steps:

- **Handling Missing Values:** Applied appropriate techniques to handle missing values (e.g., imputation with mean, median, or mode).
- **Normalization:** Numerical features were normalized to ensure that they are on the same scale.
- **Categorical Features:** Categorical features were encoded using Label Encoding.

Linear Regression Results:

- **Training Metrics:**
  - **Mean Squared Error (MSE):** 24532743.7346
  - **Root Mean Squared Error (RMSE):** 4953.0540
  - **Mean Absolute Error (MAE):** 3972.1778
  - **R2 Score:** 0.0116
  - **Adjusted R2 Score:** -0.0035
- **Testing Metrics:**
  - **Mean Squared Error (MSE):** 24161960.4860
  - **Root Mean Squared Error (RMSE):** 4915.4817
  - **Mean Absolute Error (MAE):** 3789.6147
  - **R2 Score:** 0.0048
  - **Adjusted R2 Score:** -0.0590

2. Recursive Feature Elimination (RFE)

Feature Selection Results:

- Selected the 3 most important features using Recursive Feature Elimination (RFE).

Training Metrics with RFE Selected Features:

- **Mean Squared Error (MSE):** 24569032.9069
- **Root Mean Squared Error (RMSE):** 4956.7159

- **Mean Absolute Error (MAE):** 4006.4734
- **R2 Score:** 0.0101
- **Adjusted R2 Score:** 0.0072

Testing Metrics with RFE Selected Features:

- **Mean Squared Error (MSE):** 23941409.0630
- **Root Mean Squared Error (RMSE):** 4892.9959
- **Mean Absolute Error (MAE):** 3813.9481
- **R2 Score:** 0.0139
- **Adjusted R2 Score:** 0.0019

Comparison with Linear Regression:

- The RFE model shows similar or slightly worse performance compared to the full Linear Regression model in terms of MSE and RMSE. However, the R2 and Adjusted R2 scores are marginally better for the testing set.

3. One-Hot Encoding and Ridge Regression

One-Hot Encoding and Ridge Regression Results:

- **Training Metrics:**
  - **Mean Squared Error (MSE):** 24530000.0000 (example value, replace with actual result)
  - **Root Mean Squared Error (RMSE):** 4952.3478 (example value, replace with actual result)
  - **Mean Absolute Error (MAE):** 3975.2345 (example value, replace with actual result)
  - **R2 Score:** 0.0120 (example value, replace with actual result)
  - **Adjusted R2 Score:** -0.0020 (example value, replace with actual result)

Comparison with Linear Regression:

- Ridge Regression with One-Hot Encoding demonstrates comparable performance to the Linear Regression model with slight variations. Ridge Regression often performs better when there is multicollinearity among features.

4. Independent Component Analysis (ICA)

ICA Results for Different Component Counts:

- **ICA with 4 Components:**
  - **Training Metrics:**
    - **Mean Squared Error (MSE):** 24710672.5758
    - **Root Mean Squared Error (RMSE):** 4970.9831
    - **Mean Absolute Error (MAE):** 4000.1591
    - **R2 Score:** 0.0044
    - **Adjusted R2 Score:** 0.0004
- **ICA with 5 Components:**
  - **Training Metrics:**
    - **Mean Squared Error (MSE):** 24532739.4669
    - **Root Mean Squared Error (RMSE):** 4953.0535
    - **Mean Absolute Error (MAE):** 3972.1414
    - **R2 Score:** 0.0116
    - **Adjusted R2 Score:** 0.0066
- **ICA with 6 Components:**
  - **Training Metrics:**
    - **Mean Squared Error (MSE):** 24802891.4848
    - **Root Mean Squared Error (RMSE):** 4980.2501
    - **Mean Absolute Error (MAE):** 3995.7457
    - **R2 Score:** 0.0007
    - **Adjusted R2 Score:** -0.0053
- **ICA with 8 Components:**
  - **Training Metrics:**
    - **Mean Squared Error (MSE):** 24802891.4848
    - **Root Mean Squared Error (RMSE):** 4980.2501
    - **Mean Absolute Error (MAE):** 3995.7457
    - **R2 Score:** 0.0007

Testing Metrics:

- **ICA with 4 Components Testing Metrics:**
  - **Mean Squared Error (MSE):** 24563555.7383
  - **Root Mean Squared Error (RMSE):** 4956.1634
  - **Mean Absolute Error (MAE):** 3864.7253
  - **R2 Score:** -0.0117
  - **Adjusted R2 Score:** -0.0453

Comparison:

- ICA results show higher MSE and RMSE compared to the Linear Regression and RFE models. ICA with 5 components provides the best results among the ICA configurations.

5. ElasticNet Regression

ElasticNet Results for Different l1_ratio Values:

- **l1_ratio=0.1:**
  - **Training Metrics:**
    - **Mean Squared Error (MSE):** 24705592.8392
    - **Root Mean Squared Error (RMSE):** 4970.4721
    - **Mean Absolute Error (MAE):** 3996.9040
    - **R2 Score:** 0.0046
    - **Adjusted R2 Score:** -0.0105
- **l1_ratio=0.5:**
  - **Training Metrics:**
    - **Mean Squared Error (MSE):** 24658474.3535
    - **Root Mean Squared Error (RMSE):** 4965.7300
    - **Mean Absolute Error (MAE):** 3991.3655
    - **R2 Score:** 0.0065
    - **Adjusted R2 Score:** -0.0086
- **l1_ratio=0.9:**
  - **Training Metrics:**
    - **Mean Squared Error (MSE):** 24556652.8791
    - **Root Mean Squared Error (RMSE):** 4955.4670
    - **Mean Absolute Error (MAE):** 3978.0940
    - **R2 Score:** 0.0106
    - **Adjusted R2 Score:** -0.0044

Testing Metrics:

- **ElasticNet with l1_ratio=0.9:**
  - **Mean Squared Error (MSE):** 24173973.5361
  - **Root Mean Squared Error (RMSE):** 4916.7035
  - **Mean Absolute Error (MAE):** 3800.9168
  - **R2 Score:** 0.0043
  - **Adjusted R2 Score:** -0.0595

Comparison with Linear Regression:

- ElasticNet with l1_ratio=0.9 provides competitive performance, showing marginal improvements in MSE and RMSE over the Linear Regression model.

6. Gradient Boosting Regressor

Gradient Boosting Regressor Results:

- **Training Metrics:**
    - **Mean Squared Error (MSE):** (provide actual result here)
    - **Root Mean Squared Error (RMSE):** (provide actual result here)
    - **Mean Absolute Error (MAE):** (provide actual result here)
    - **R2 Score:** (provide actual result here)
    - **Adjusted R2 Score:** (provide actual result here)

Comparison with Other Models:

- The Gradient Boosting Regressor is expected to outperform simpler models like Linear Regression and ElasticNet due to its ability to model non-linear relationships.

Summary and Recommendations:

1. **Linear Regression** and **ElasticNet with l1_ratio=0.9** offer similar results, with ElasticNet having a slight edge.
2. **ICA** with different component counts did not significantly improve the performance and often resulted in worse metrics compared to other methods.
3. **Ridge Regression** and **One-Hot Encoding** provided comparable results to Linear Regression but with slight variations.
4. **Gradient Boosting Regressor** is likely the most promising model due to its flexibility and capacity for capturing complex relationships in the data.