Problem Description

This project aimed to predict **energy consumption** based on various building characteristics and environmental factors. The primary goal was to develop a predictive model that could estimate how much energy a building would consume given specific inputs. I chose this problem because energy consumption is a critical factor in sustainability, cost management, and environmental impact.

Data Set

- The data used in this project was sourced from **Kaggle**, specifically the **Energy Consumption Dataset for Linear Regression**.
- The following features were selected based on their **logical relevance** and their **correlation with the target variable** (energy_consumption), as visualized using a correlation matrix heatmap:

  (1) square_footage (Numerical) – Indicates building size; larger areas typically require more energy.

  (2) number_of_occupants (Numerical) – More people generally lead to higher energy consumption.

  (3) appliances_used (Numerical) – Reflects the number of active electrical devices.

  (4) building_type_Industrial (from One-Hot Encoding) – Indicates if the building is Industrial, which often has higher energy demands.

Modeling Approach

- For this prediction task, a **Multi-Linear Regression** model was employed. It was chosen for its interpretability and its effectiveness in modeling linear relationships between features and the target variable.
- **Data Preprocessing & Scaling**

  (1) Handling Missing Values**:** The first step involved removing any rows with missing values using df.dropna() to ensure data integrity.

  (2) Categorical features (building_type, day_of_week) were converted into numeric form using One-Hot Encoding with drop_first=True to avoid multicollinearity, and encoded columns were cast to integers.

  (3) Data Splitting: The dataset was then **split** into training (70%) and testing (30%) subsets. A random_state of 42 was used to ensure the split was **reproducible**.

  (4) Standard Scaler was used to standardize numerical features for both training and testing sets, ensuring balanced feature influence and faster convergence in Linear Regression. The scaler was fitted on training data and applied to test data, with the fitted object saved for future use.

Evaluation

- After training the Linear Regression model, I evaluated its performance using several metrics. The results were:
  Mean Absolute Error (MAE): 184.41
  Mean Squared Error (MSE): 45,793.47
  R^2 Score: 0.9452
- These metrics show that my model explains approximately **94.5% of the variance** in energy consumption, which indicates a strong fit, especially for a simple linear regression model with only four features. The MAE and RMSE values reflect relatively low prediction errors in the units of energy used. To ensure the model's robustness and generalizability, I also applied **cross-validation**, which confirmed that the model performs consistently well across different data subsets. For a basic regression project, these results are strong and reliable.

Reflection

- Throughout this project, I learned the importance of carefully framing a prediction problem and selecting relevant features based on domain knowledge. I also gained practical experience in basic data cleaning, using Pandas to handle missing values, and transforming data for modeling.
- One challenge I faced was deciding which features to include. Since I wanted to keep the project simple, I limited myself to four intuitive predictors. However, I realized that real-world energy consumption depends on many other factors, such as weather conditions, insulation quality, and seasonal effects, which I did not consider here. I also learned that for more advanced models, feature scaling and regularization can make a significant difference.
- If I were to redo this project, I would try to collect a larger and more diverse dataset, engineer new features like energy efficiency ratings, and experiment with more complex models. Overall, this project strengthened my understanding of supervised learning and gave me confidence to tackle more advanced regression tasks in the future.