A screenshot of a computer screen

Description automatically generated  
  
  
**1. Correlation Analysis**

A **correlation heatmap** visually depicts the relationships between all numerical variables. Correlation quantifies the degree to which two variables move together, with values ranging from -1 to 1.

**Key Outputs:**

• **Correlation Matrix**: A table that shows pairwise correlation coefficients between numerical features.

• **Heatmap**: Color-coded visualization of the matrix where darker shades show stronger correlations.

**How to Interpret:**

• **Strong Positive Correlation (close to +1)**: As one variable increases, the other increases.

• **Strong Negative Correlation (close to -1)**: As one variable increases, the other decreases.

• **Weak Correlation (close to 0)**: No linear relationship between the two variables.

**Example Insights:**

• If **Work Hours** and **EUI** have a strong positive correlation (e.g., 0.75), it implies that buildings open for longer hours consume more energy.

• Detect multicollinearity: Variables with high correlations (like **Percent Cooled** and **Energy Use Intensity**) may introduce redundancy in regression models.

A graph of energy use intensity by building activity

Description automatically generated  
**2. Box Plot of EUI by Building Activity**

The **box plot** visualizes the distribution of **Energy Use Intensity (EUI)** for different categories of building activities (e.g., office, residential, retail). This plot reveals key aspects of data distribution, such as median, interquartile range (IQR), and potential outliers.

**Key Outputs:**

• **Median**: The line within the box that shows the central value of the data for each category.

• **Interquartile Range (IQR)**: The range between the 25th and 75th percentiles, representing the spread of most data points.

• **Outliers**: Points that fall outside 1.5 times the IQR, flagged as unusual observations.

**How to Interpret:**

• Taller boxes (larger IQR) indicate higher variability in EUI for that building activity.

• Outliers may reveal buildings that consume significantly more or less energy than expected for a particular activity.

**Example Insights:**

• **Residential buildings** might show a lower median EUI compared to **commercial buildings**, as households typically use less energy than offices.

• Outliers in **office buildings** may indicate inefficient energy consumption, prompting the need for energy audits.

A graph of energy use intensity

Description automatically generated  
  
  
**3. Scatter Plot of EUI vs Work Hours**

A **scatter plot** visually depicts the relationship between **Energy Use Intensity (EUI)** and **Work Hours**, where each point represents an individual building. This step is crucial for identifying clusters, patterns, or outliers that might affect analysis.

**Key Outputs:**

• **Relationship Line**: If a trend exists, a regression line may be added.

• **Clusters**: Groups of points in specific areas suggest homogeneity in the data.

• **Outliers**: Isolated points that deviate from overall patterns.

**How to Interpret:**

• A clear upward trend (positive slope) indicates that as **Work Hours** increase, **EUI** also increases.

• Clusters of points may hint at subgroups (e.g., commercial buildings vs. residential buildings) with distinct energy use patterns.

• Outliers may identify buildings with inefficient energy use or measurement errors.

**Example Insights:**

• Buildings operating 24/7 (like hospitals) may have significantly higher EUI than office spaces that operate only 9-to-5.

• If there is no clear trend, it suggests that **Work Hours** alone may not fully explain variations in **EUI**, prompting the inclusion of other predictors in the model.

A screenshot of a computer

Description automatically generated  
  
  
4. **ANOVA (Analysis of Variance) Results**

ANOVA tests if the mean **EUI** differs significantly between groups (e.g., office, residential, and retail buildings). This step helps determine whether differences in **building activity** significantly affect **EUI**.

**Key Outputs:**

• **F-statistic**: Measures the variance between groups relative to the variance within groups.

• **P-value**: Tests the null hypothesis that all group means are equal. A p-value < 0.05 suggests at least one group has a significantly different mean.

**How to Interpret:**

• If the **p-value < 0.05**, reject the null hypothesis, indicating that **EUI** differs significantly between building activities.

• Post-hoc tests (like Tukey’s test) help pinpoint which groups differ significantly.

**Example Insights:**

• ANOVA might reveal that **office buildings** have significantly higher EUI compared to **residential buildings**, justifying targeted energy-saving strategies for office spaces.

A group of blue dots

Description automatically generated  
  
  
**5. Pairwise Visualization (Pair plot)**

A **pair plot** provides a comprehensive visual exploration of relationships between all numerical features in the dataset. It helps detect clusters, patterns, and potential outliers.

**Key Outputs:**

• **Scatter Plots**: Show relationships between each pair of numerical variables.

• **Histograms**: Display the frequency distribution for each variable.

**How to Interpret:**

• If two variables show a linear pattern in their scatter plot, it suggests a possible correlation.

• Variables with strong linear relationships are potential predictors in regression models.

**Example Insights:**

• Pair plot might reveal that **Percent Cooled** has a linear relationship with **EUI**, suggesting that cooling strategies impact energy use.

• Clusters may indicate subgroups of buildings (e.g., commercial vs. residential) with distinct energy use patterns.  
  
  
**Summary of Results**

The EDA reveals critical insights into the dataset’s structure, relationships, and key drivers of energy consumption.

**1. Descriptive Statistics**: Energy Use Intensity (EUI) is highly variable, with possible outliers suggesting unusually high energy consumption in certain buildings.

**2. Correlation Analysis**: **Work Hours** and **EUI** are positively correlated, indicating longer hours increase energy use.

**3. Box Plot**: **Office buildings** show higher EUI than **residential buildings**, suggesting opportunities for energy-saving measures.

**4. ANOVA**: Confirmed significant differences in EUI across building activity groups.

**5. Multiple Linear Regression**: **Work Hours** and **Percent Cooled** are significant predictors of **EUI**, while some variables may have less impact.

**6. Pair plot**: Clusters in the **Work Hours** vs. **EUI** plot suggest subgroups, like 24/7 operations vs. part-time use.