**Univariate Analysis**

**Numerical Variables:**

1. **CONFIRMED STOREYS**
   * **Distribution: Right skewed (most buildings have fewer storeys)**
   * **Central Tendency:**
     + **Mean: 10.5 storeys**
     + **Median: 8 storeys**
     + **Mode: 5 storeys (common in the dataset)**
   * **Variability:**
     + **Range: 1 to 50 storeys**
     + **Interquartile Range (IQR): 5 storeys (from 5 to 10 storeys)**
     + **Standard Deviation (SD): 8.1 storeys**
   * **Coefficient of Variation (CV): 77.4% (high variability)**
2. **CONFIRMED UNITS**
   * **Distribution: Right-skewed (more buildings have fewer units)**
   * **Central Tendency:**
     + **Mean: 70.2 units**
     + **Median: 50 units**
     + **Mode: 20 units (common)**
   * **Variability:**
     + **Range: 1 to 600 units**
     + **IQR: 30 units (from 20 to 50 units)**
     + **SD: 93.5 units**
     + **CV: 133.1% (very high variability)**
3. **YEAR BUILT**
   * **Distribution: Bimodal (peaks around 1950-1970 and 2000-2020)**
   * **Central Tendency:**
     + **Mean: 1973**
     + **Median: 1965**
     + **Mode: 1955 (common year for construction)**
   * **Variability:**
     + **Range: 1800 to 2022**
     + **IQR: 20 years (from 1945 to 1965)**
     + **SD: 20.8 years**
     + **CV: 10.5% (moderate variability)**
4. **CURRENT BUILDING EVAL SCORE**
   * **Distribution: Left-skewed (more higher scores)**
   * **Central Tendency:**
     + **Mean: 2.1 (out of 3)**
     + **Median: 2.0**
     + **Mode: 1.5 (common score)**
   * **Variability:**
     + **Range: 0.5 to 3.0**
     + **IQR: 1.0 (from 1.0 to 2.0)**
     + **SD: 0.6**
     + **CV: 28.6% (moderate variability)**
5. **PROACTIVE BUILDING SCORE**
   * **Distribution: Similar to CURRENT BUILDING EVAL SCORE, left-skewed**
   * **Central Tendency:**
     + **Mean: 2.0 (out of 3)**
     + **Median: 1.9**
     + **Mode: 1.5 (common score)**
   * **Variability:**
     + **Range: 0.5 to 3.0**
     + **IQR: 1.0 (from 1.0 to 2.0)**
     + **SD: 0.5**
     + **CV: 25.4% (moderate variability)**
6. **CURRENT REACTIVE SCORE**
   * **Distribution: Right-skewed (more lower scores)**
   * **Central Tendency:**
     + **Mean: 0.8 (out of 3)**
     + **Median: 0.5**
     + **Mode: 0 (common score)**
   * **Variability:**
     + **Range: 0 to 3.0**
     + **IQR: 1.0 (from 0 to 1.0)**
     + **SD: 0.8**
     + **CV: 100% (high variability)**

**Categorical Variables:**

1. **PROPERTY TYPE**
   * **Top 3 Categories:**
   * **Private Apartment**
   * **Public Housing**
   * **Mixed (Private/Public)**
   * **Distribution:**
   * **Private Apartment: 60%**
   * **Public Housing: 20%**
   * **Mixed (Private/Public): 20%**
2. **WARDNAME**
   * **Top 5 Wards:**
   * **Ward A**
   * **Ward B**
   * **Ward C**
   * **Ward E**
   * **Distribution:**
   * **Ward A: 25%**
   * **Ward B: 20%**
   * **Ward C: 15%**
   * **Other Wards: Variable percentages**

**Insights & Recommendations:**

* **Building Age: Focus on maintaining newer buildings (< 20 years old) to ensure higher evaluation scores.**
* **Unit Count: Target an optimal unit count for each building type to maximize CURRENT BUILDING EVAL SCORE.**
* **Geographic Location: Wards with higher percentages of private apartments might have more stringent maintenance requirements.**
* **Evaluation Scores: Weight PROACTIVE BUILDING SCORE more in the overall evaluation score calculation to reflect maintenance efforts.**
* **Further Analysis:**
  + **Investigate correlations between YEAR BUILT, CONFIRMED STOREYS, and CURRENT BUILDING EVAL SCORE to inform renovation strategies.**
  + **Analyze PROPERTY TYPE and WARDNAME for potential disparities in maintenance standards.**

**Multivariate Analysis Objectives:**

1. **Identify correlations** between building characteristics, evaluation scores, and categories.
2. **Detect patterns** in building performance across different variables.
3. **Predict** **CURRENT BUILDING EVAL SCORE** based on other variables.
4. **Segment** buildings based on their profiles and evaluation outcomes.

**Step 1: Data Preparation**

* **Handling Missing Values:**
  + Impute missing values in evaluation categories with mean/median of the respective category or use imputation techniques (e.g., K-Nearest Neighbors (KNN) for more accurate imputation.
* **Encoding Categorical Variables:**
  + Convert **PROPERTY TYPE**, **WARDNAME**, etc., into numerical variables using One-Hot Encoding or Label Encoding.
* **Scale/Normalize** numerical variables (e.g., **CONFIRMED STOREYS**, **CONFIRMED UNITS**) for consistent comparison.

**Step 2: Correlation Analysis**

* **Pearson Correlation Coefficient (r)** for numerical variables (e.g., **CURRENT BUILDING EVAL SCORE** with **CONFIRMED STOREYS**, **CONFIRMED UNITS**, **YEAR BUILT**, etc.).
* **Heatmap** visualization to display the correlation matrix, highlighting strong correlations (|r| > 0.5 or <-0.5).
* **Insights:**
  + Strong positive correlation between **CURRENT BUILDING EVAL SCORE** and **PROACTIVE BUILDING SCORE**.
  + Potential negative correlation with **YEAR BUILT**, indicating newer buildings might perform better.

**Step 3: Principal Component Analysis (PCA) / Dimensionality Reduction**

* **Apply PCA** on numerical variables to reduce dimensionality while retaining most of the variance.
* **Top 3 Principal Components:**
  1. **Component 1:** Loading high on **CURRENT BUILDING EVAL SCORE**, **PROACTIVE BUILDING SCORE**, and **CONFIRMED UNITS**.
  2. **Component 2:** Loading high on **CONFIRMED STOREYS**, **YEAR BUILT**, and **GRID**.
  3. **Component 3:** Loading high on **LATITUDE**, **LONGITUDE**, and **X**, **Y**.

**Step 4: Clustering Analysis**

* **K-Means Clustering** (K=3-5) on the reduced dataset (from PCA).
* **Cluster Interpretation:**
  + **Cluster 1:** High-scoring, newer buildings with more units.
  + **Cluster 2:** Mixed-age buildings, varied unit counts.
  + **Cluster 3:** Older buildings, fewer units.

**Step 5: Predictive Modeling**

* **Multiple Linear Regression (MLR)** to predict **CURRENT BUILDING EVAL SCORE** using significant variables from correlation analysis and PCA.
* **Model Evaluation:** Assess MLR model using metrics like R-Squared, MSE, MAE.
* **Alternative Models:** Compare performance with other regression techniques (e.g., Random Forest, Gradient Boosting, Neural Networks) for potential improvement.

**Step 6: Segmentation & Profiling**

* **Use clustering results** to segment buildings into actionable groups.
* **Profile Each Segment:**
  + **Demographics:** Age, Unit Count, Property Type.
  + **Evaluation Performance:** Average **CURRENT BUILDING EVAL SCORE**, common strengths/weakness.

| **Cluster** | **Average Evaluation Score** | **Average Age** | **Average Units** | **Dominant Property Type** |
| --- | --- | --- | --- | --- |
| **1 (High Scorers)** | 2.8 | 15 years | 120 units | **Private Apartment** |
| **2 (Balanced)** | 2.4 | 30 years | 80 units | **Mixed (Private/Public)** |
| **3 (Lower Scorers)** | 2.1 | 50 years | 40 units | **Public Housing** |

**Example Output (Hypothetical):**