

Specialization: *Unsupervised Learning*

Business Focus: *Environmental Analytics for Urban Planning*

Tool: *Pandas, Matplotlib, Seaborn, Sklearn, etc.*

Urban Sustainability Indexing Clustering Global Cities for Green Futures

Tools and Technology to be Used



matplotlib

pandas



seaborn

Background

- GreenCityWatch, a global urban sustainability organization, seeks to help governments and development agencies identify cities that are falling behind in achieving sustainability goals such as reduced emissions, waste management, green energy adoption, and livability improvements.

- Their mission is:
 - *To accelerate the transition of cities toward climate resilience, energy efficiency, and sustainable living by enabling data-driven urban policy decisions.*



Problem Statement

- GreenCityWatch has compiled a global dataset of cities with environmental and livability metrics.
- They want to use data science and unsupervised learning to:
 - Cluster cities into groups based on their sustainability indicators.
 - Identify and prioritize the cities in the most critical need of support and funding for sustainability transitions.



Data Description

Column Name	Description
City	Name of the city
region	World region the city belongs to
population	Total population of the city
green_space_pct	% of city area covered in green parks or forest
air_quality_index	Index value (lower is better)
waste_recycled_pct	% of total waste that is recycled
renewable_energy_pct	% of city's energy from renewable sources
avg_commute_time	Average time (minutes) spent commuting per day
carbon_emissions	Annual per capita CO ₂ emissions (metric tons)
water_access_pct	% of population with access to clean water
energy_efficiency_score	City's building energy performance index (0–100)

Objectives

- Use clustering algorithms to segment cities based on sustainability profiles.
- Recommend which cluster(s) should be prioritized for intervention.
- Use EDA and visualizations to interpret patterns, e.g., clusters with high emissions and low recycling/renewables.



Project Workflow

**Data Cleaning:**

- Handle any missing values.
- Remove duplicate records or irrelevant columns.
- Check for and correct anomalies.

Exploratory Data Analysis (EDA):

- Visualize distributions and relationships between features.
- Identify patterns, trends, and potential anomalies.
- Form hypotheses based on data insights.

Data Preprocessing:

- Scale/normalize numerical features and encode categorical data.
- Dimensionality reduction (optional – PCA, t-SNE)

Model Training:

- Select and train machine learning models on the data.
- Elbow Method to find the optimal number of clusters (k).
- Experiment with different algorithms and assess performance.

Model Evaluation (Interpreting Results):

- Visual Inspection: visualize cities in the “critical” cluster (e.g., high carbon emissions, low renewables, poor air quality)

READY TO DELVE IN?

