Water Energy Synergy

WATER SUPPLY AND ENERGY USAGE IN AFRICAN REGIONS

BENEDICT OUMA

August 29th, 2025

INTRODUCTION «

Context:
Access to reliable water and energy is a cornerstone of sustainable development. Many African regions face disparities in water availability, energy access, and infrastructure, which affect economic growth, public health, and quality of life. Effective resource planning requires understanding regional needs and potential for renewable energy deployment.

Objective of the Analysis:

1. Classify regions according to solar energy potential.
2. Identify areas with critical water and energy shortages.
3. Provide data-driven recommendations for resource allocation to improve access and sustainability.

PROBLEM STATEMENT

Many regions experience inadequate water supply and energy access, while investments in renewable energy (e.g., solar) are often not optimally targeted. Therefore, the decision-makers need insights to:

1. Predict solar performance tiers for efficient solar panel deployment.
2. Identify high-need regions for water and energy interventions.
3. Allocate resources strategically to maximize impact.

Goal:

Develop a data-driven framework to prioritize regions and guide infrastructure investment decisions based on water, energy, and socio-geospatial indicators.

DATA DESCRIPTION

Dataset:

1. Level: Local Government Area (LGA)
2. Records: 763 LGAs across 37 states
3. Source: Aggregated water and energy statistics, demographic and geospatial data

Data Characteristics

1. 18 original numerical features
2. 2 original categorical features and 2 empty and unnamed columns
3. 10 engineered features
4. No null values or duplicates

Key Features

1. Water Infrastructure: Boreholes, taps, handpumps, rainwater harvesting, unimproved sources
2. Demographics & Geography: Population, area, perimeter, state, LGA
3. Energy Proxy: Average nighttime light (for electricity/energy usage)
4. Spatial: Latitude, longitude, geometry

Feature	Description	Data Type	Feature Type
LGA	Name of the Local Government Area	Categorical	Original
STATE	Name of the state or region	Categorical	Original
Borehole	Count of boreholes available	Numerical	Original
Don't_Know	Count of unknown or unreported water sources	Numerical	Original
Handpump	Count of handpumps present	Numerical	Original
Overhead_Tank	Count of overhead water storage tanks	Numerical	Original
Rainwater_Harvesting_System	Count of rainwater harvesting systems	Numerical	Original
Tap	Count of public taps available	Numerical	Original
Unimproved	Count of unimproved water sources	Numerical	Original
Unimproved_Large_Diameter_Well	Count of unimproved large diameter wells	Numerical	Original
Unimproved_Rainwater_Harvesting_System	Count of unimproved rainwater harvesting systems	Numerical	Original
Unimproved_Well	Count of unimproved wells	Numerical	Original
Unprotected_Spring	Count of unprotected springs	Numerical	Original
Untreated_Surface_Water	Count of untreated surface water sources	Numerical	Original
Average_Nighttime_mean	Proxy for energy access (average night-time light intensity)	Numerical	Original
AREA	Total area of LGA (e.g., in sq. km)	Numerical	Original
Latitude	Latitude coordinate of LGA centroid	Numerical	Original
Longitude	Longitude coordinate of LGA centroid	Numerical	Original
PERIMETER	Perimeter of the LGA boundary	Numerical	Original
Population	Total population of the LGA	Numerical	Original
Distance_to_Center	Distance from LGA centroid to its state centroid (in degrees)	Numerical	Engineered
pop_density	Population density (Population / AREA)	Numerical	Engineered
safe_water	Count of safe water sources (Borehole, Tap, Overhead_Tank, Rainwater_Harvesting_System)	Numerical	Engineered
unsafe_water	Count of unsafe water sources (Unimproved sources, unprotected spring, untreated surface water)	Numerical	Engineered
safe_ratio	Proportion of population with access to safe water	Numerical	Engineered
unsafe_ratio	Proportion of population relying on unsafe water	Numerical	Engineered
water_diversity	Number of distinct water sources (>0)	Numerical	Engineered
climate_stress	Proxy for climate stress: nighttime light / population density	Numerical	Engineered
accessibility	Proxy for accessibility: distance to center normalized by area	Numerical	Engineered
compactness	Shape compactness metric: (PERIMETER^2 / AREA)	Numerical	Engineered

DATA PREPROCESSING

Steps Taken:

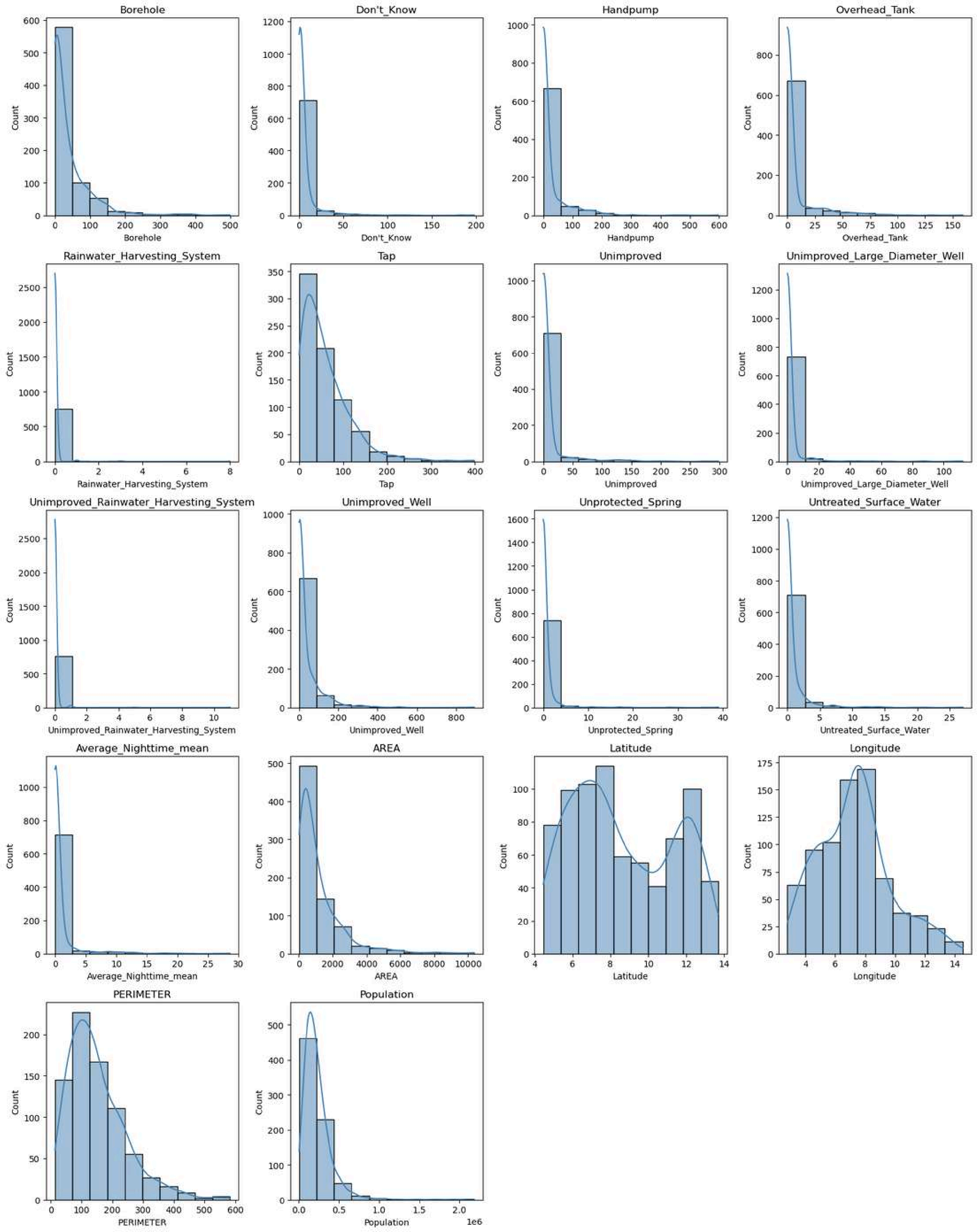
1. Column Cleaning: Dropped irrelevant identifiers and empty columns; standardized column names.
2. Duplicate Handling: Verified and removed duplicates (none found).
3. Outlier Treatment: Applied Winsorization (10% lower/upper) on numerical features to cap extreme values while retaining all records.
4. GeoDataFrame Conversion: Transformed dataset into a GeoDataFrame for spatial analysis and distance-based feature engineering.
5. Feature Engineering: Created new indicators such as:
    - Population density (Population / Area)
    - Safe vs. unsafe water ratios
    - Water source diversity
    - Climate stress (Nighttime mean / pop_density)
    - Accessibility (Distance to state centroid / Area)
    - Compactness (Perimeter^2 / Area)

Outcome: Clean, enriched, and spatially-aware dataset ready for exploratory analysis and modeling

EXPLORATORY DATA ANALYSIS

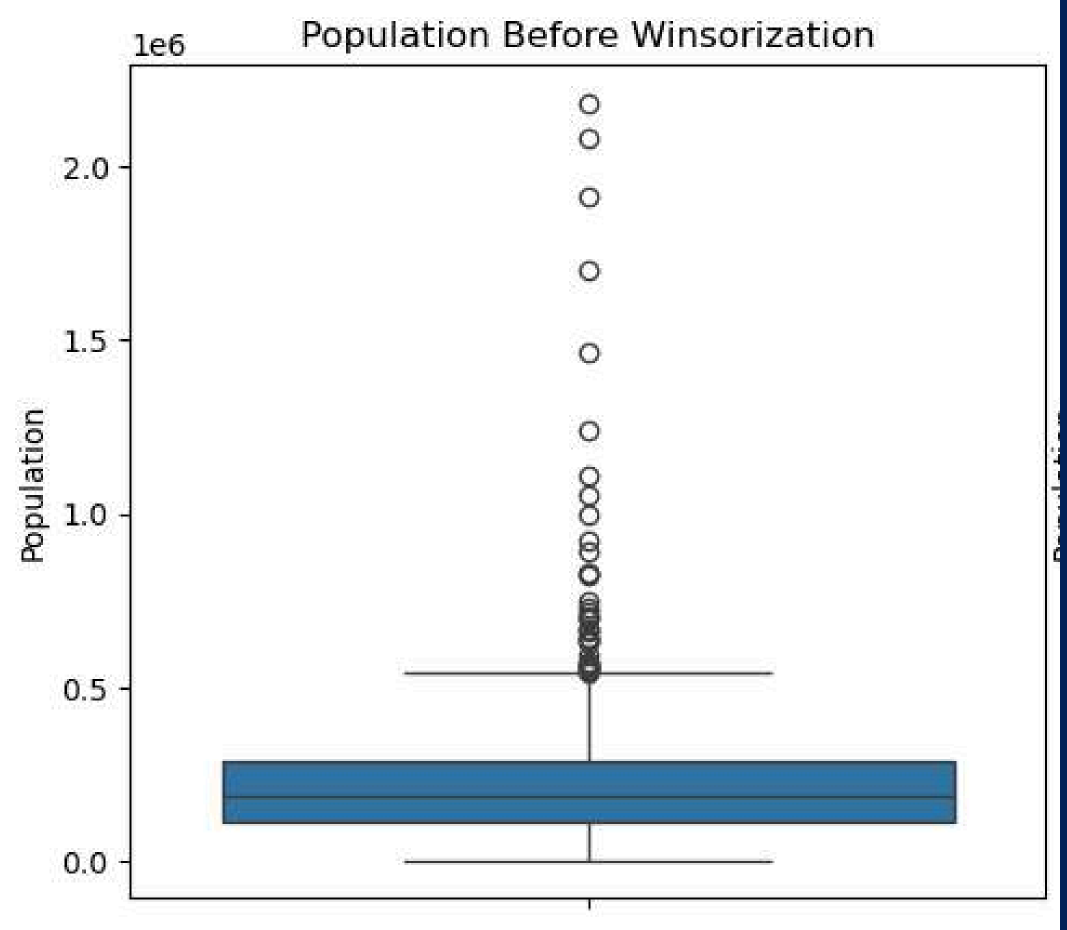
OUTLIER DETECTION FOR ORIGINAL FEATURES: Histogram Subplots

The histogram subplots of the original numerical features reveal a range of distribution patterns indicative of potential outliers. Many features, such as "Don't Know," "Unimproved," "Unprotected Spring," and "Average_Nighttime_mean," exhibit heavy-tailed distributions, showing extreme values that could disproportionately influence the mean. Several variables display multimodal distributions, suggesting the presence of distinct subgroups within the data, while most features show varying degrees of skewness, highlighting asymmetry and potential data imbalance. Additionally, some features, like "Rainwater_Harvesting_System" and "Unimproved_Rainwater_Harvesting_System," have sparse data, which may reflect either low occurrence or collection gaps. Overall, these patterns indicate that a robust outlier handling method is necessary to accurately model the data and account for its complexities.

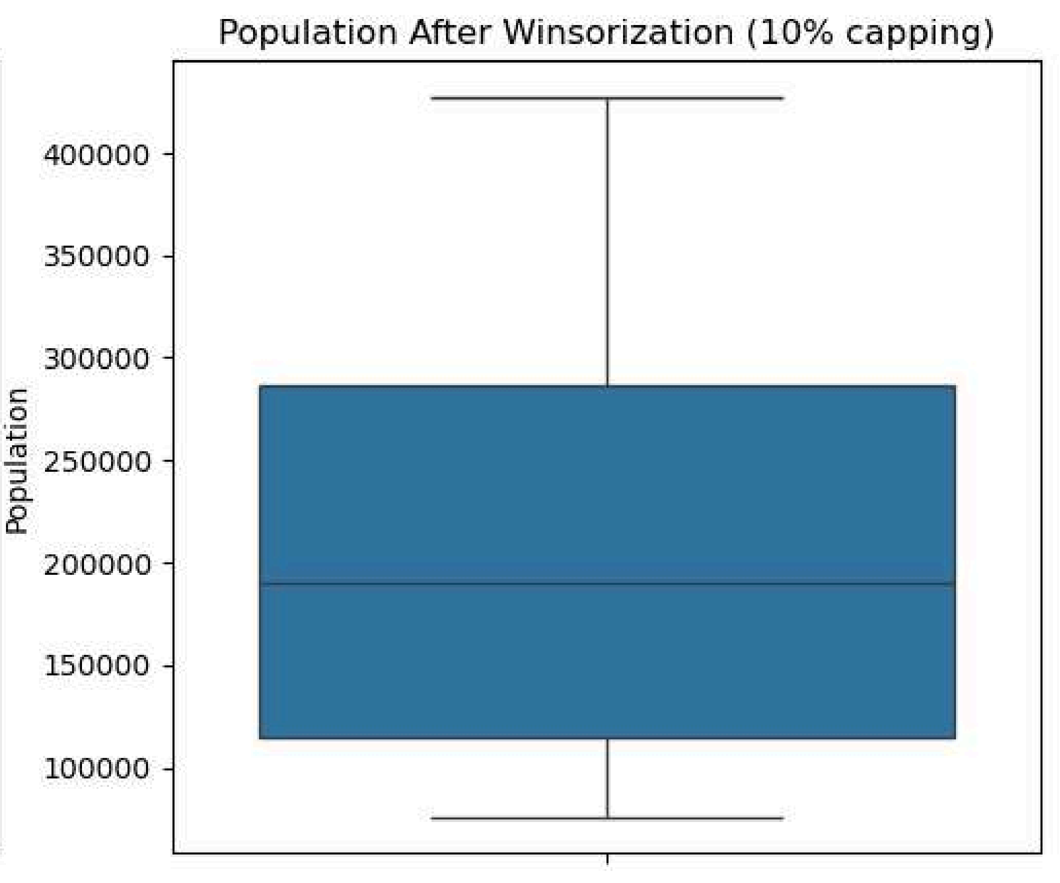


FEATURE COMPARISON (BEFORE AND AFTER OUTLIER HANDLING): Boxplot Subplots

The comparison of population data before and after Winsorization highlights the effectiveness of this technique in mitigating the influence of extreme values. Initially, the distribution was heavily right-skewed, with numerous large outliers pulling the median downward and distorting the overall spread. After applying 10% Winsorization, these extreme values were capped at the 10th and 90th percentiles, substantially reducing visible outliers and producing a more balanced, less skewed distribution. The adjustment also brought the median closer to the box center, reflecting improved symmetry, and lowered the apparent maximum value in line with the capping threshold. Overall, Winsorization transformed the data into a more robust form by limiting the disproportionate impact of extreme values, making it more suitable for statistical analysis and modeling while still preserving the underlying structure of the population distribution.



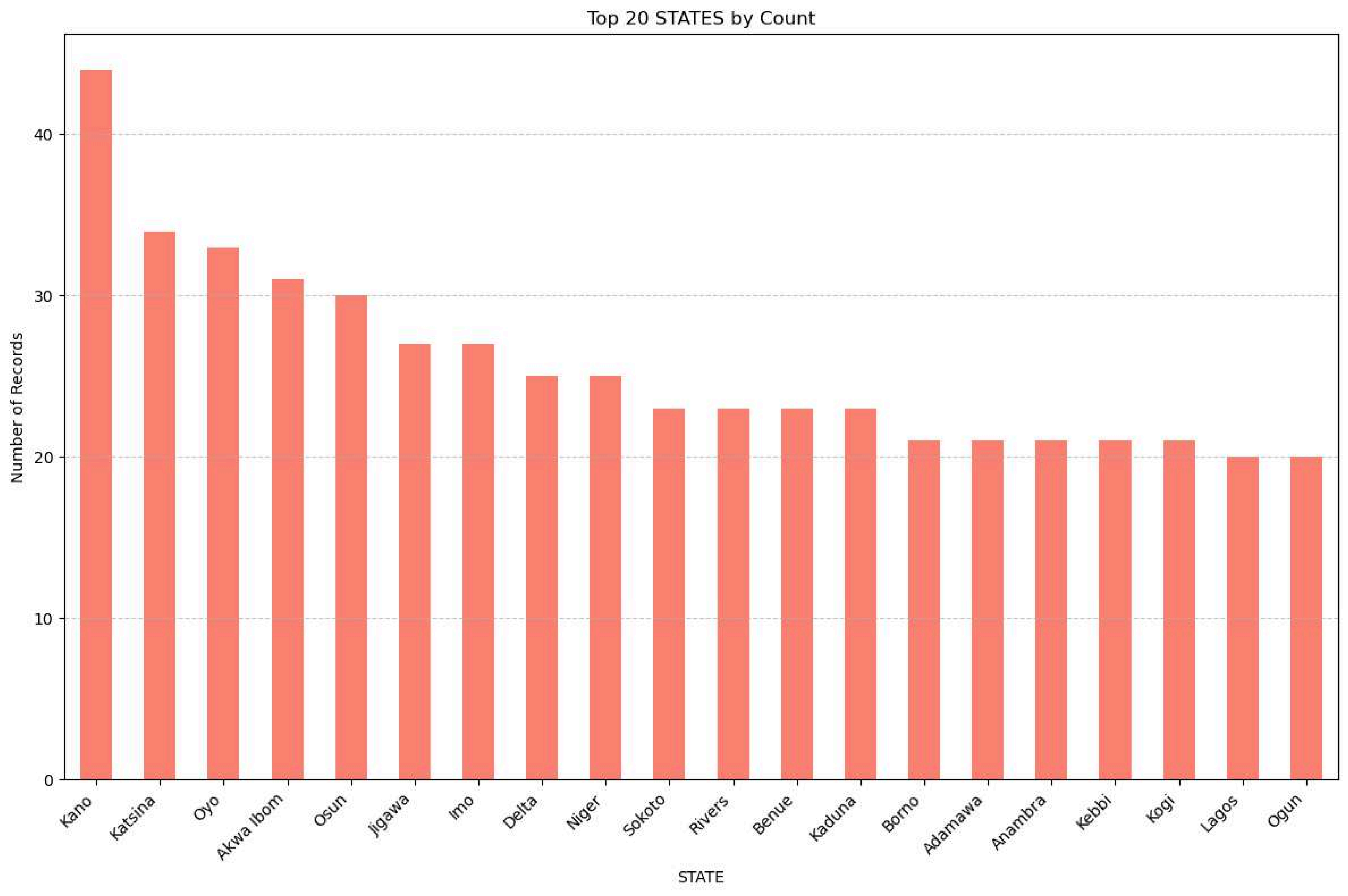
Population Before Winsorization



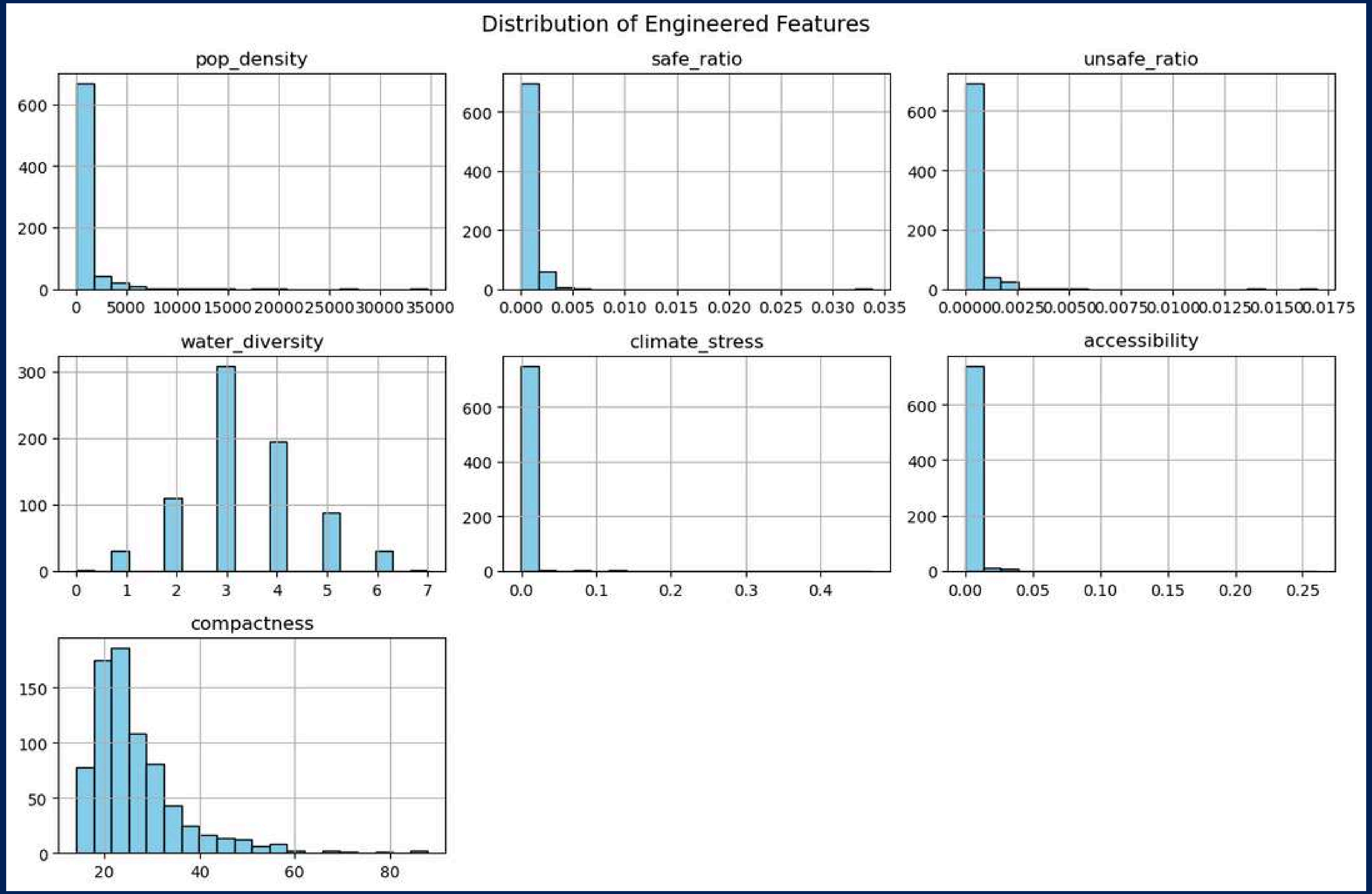
Population After Winsorization (10% capping)

TOP 20 STATES BY COUNT: Bar Chart

This bar chart presents the top 20 Nigerian states by count, offering a clear view of the states with the highest representation. At this stage, all states are shown in their original form. However, to simplify further analyses, states with fewer than 20 counts were later consolidated into a single category labeled Others, creating a new column called grouped_state while dropping the original state column. This post-visualization adjustment ensures that subsequent analysis and modeling are cleaner, more focused, and less affected by sparsely represented states.



ENGINEERED FEATURES DISTRIBUTIONS:

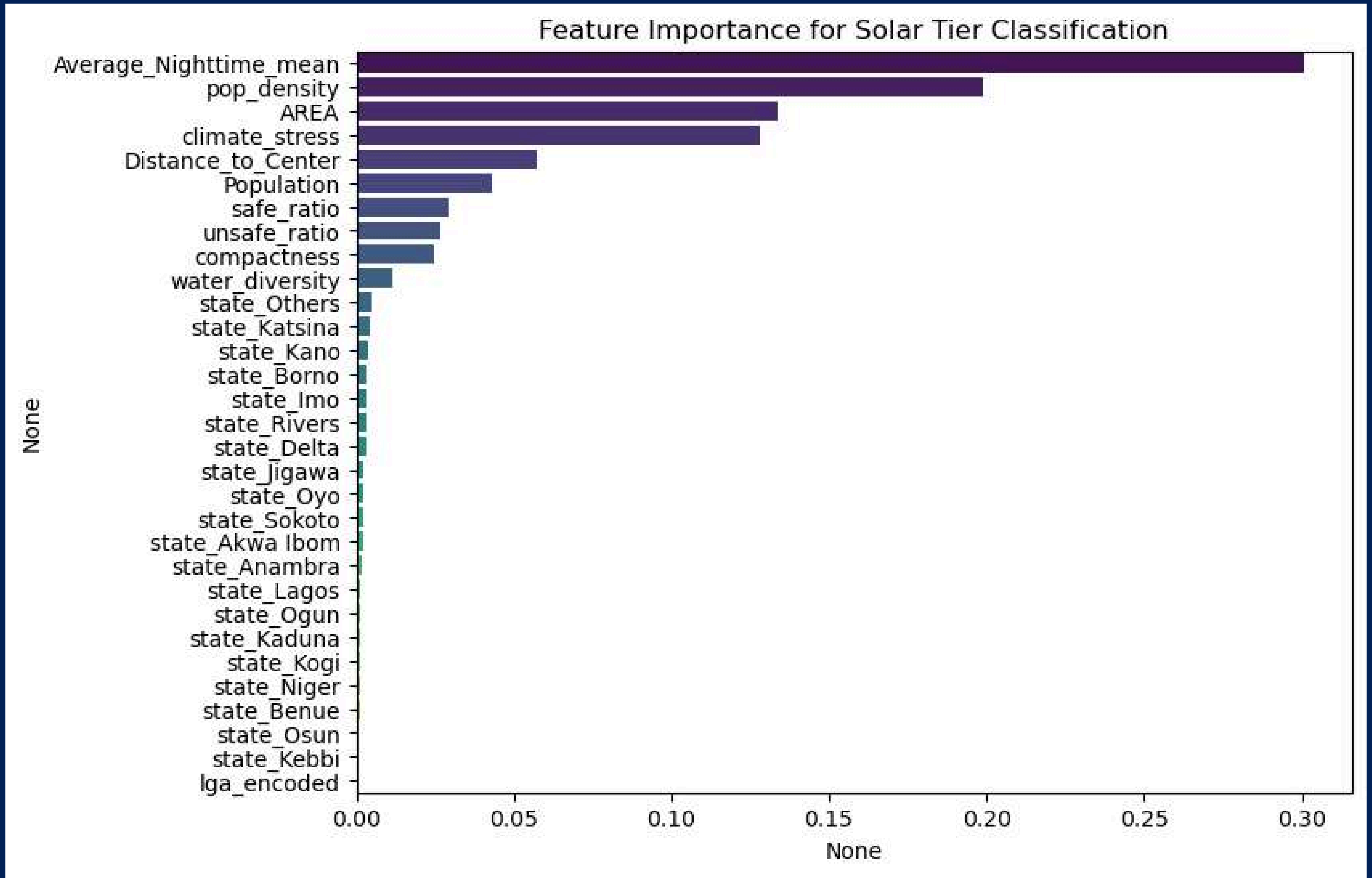


- Population Density: Strongly right-skewed; most LGAs are sparsely populated with few very dense hotspots.
- Safe/Unsafe Ratios: Concentrated near zero, indicating limited safe water coverage relative to population in most LGAs.
- Water Diversity: Peaks at 2–4, showing communities typically rely on multiple water sources rather than one.
- Climate Stress & Accessibility: Right-skewed; most areas face low stress/poor access, with a few extreme outliers.
- Compactness: More normally distributed, but with outliers above 60 indicating irregular boundaries.

Key Takeaway:

Most engineered features are highly skewed, requiring scaling/normalization before modeling. They also highlight disparities - a few regions dominate while many lag behind.

FEATURE IMPORTANCE INSIGHTS



Top Predictors:

- Average Nighttime Mean (0.31) - By far the strongest predictor. Satellite nightlight data acts as a proxy for development and electrification, strongly tied to solar tier classification.
- Population Density (0.20) - Areas with higher or lower densities influence solar feasibility, demand, and infrastructure planning.
- Area (0.13) - Geographic size matters, suggesting land availability impacts solar tiering.
- Climate Stress (0.12) - Environmental and climate conditions influence solar adoption potential.
- Distance to Center (0.10) - Proximity to urban hubs plays a role in accessibility and infrastructure.

Moderate Contributors:

- Population (~0.06) still relevant but less impactful than density.
- Safe/Unsafe Ratios & Compactness (~0.02–0.05) contribute small but meaningful signals.
- Water Diversity has minimal effect but isn’t negligible.

Least Predictive:

- Categorical location variables (state, LGA encodings) contribute almost nothing once richer socio-economic and geographic features are considered.

Key Takeaway:

The model learns primarily from development (nightlights), demographics (density, population), and geography (area, climate, access). Pure administrative boundaries add little predictive power, showing that real-world measurable conditions drive solar tier classification, not location labels.

CHOOSING OPTIMAL K (CLUSTERS)

Inertia (compactness):

- Sharp drop between k=2 - k=4.
- Improvement slows after k=3.

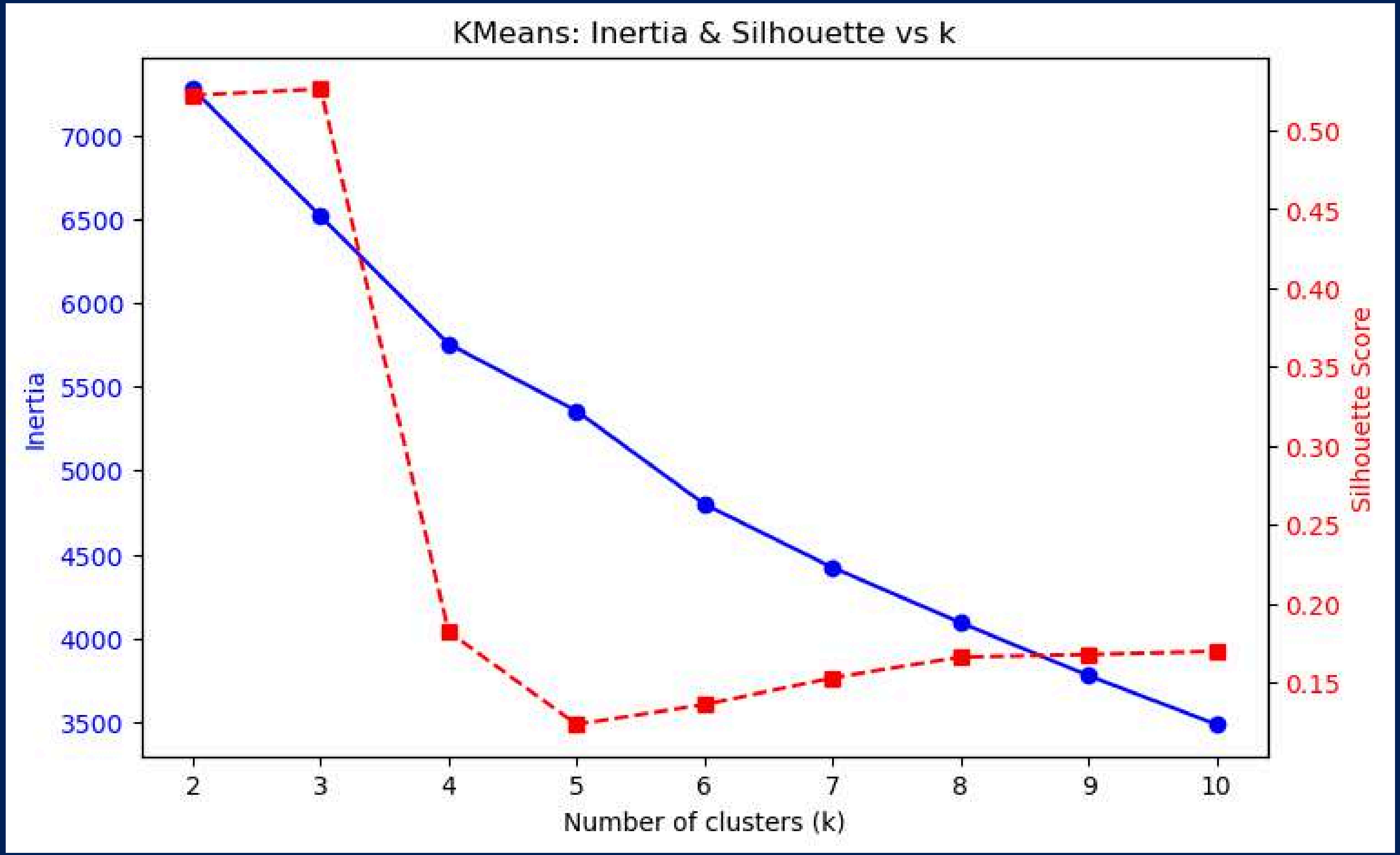
Silhouette Score (separation):

- Highest at k=2 and k=3 (~0.50).
- Sharp decline beyond k=3 (<0.20).

Interpretation:

- k=2 - strong separation but less detailed grouping.
- k=3 - strong separation and good compactness.
- k>3 - weak separation, diminishing returns.

Best choice is k=3 clusters as it shows the most preferrable balance between compactness & separation



CLUSTER-TIER RELATIONSHIP OBSERVATIONS

Cluster 0 (Green – Majority Cluster)

- Dominates across Low, Medium, and High tiers.
- Represents the general population of data points.
- Most concentrated in Low & Medium tiers.

Cluster 2 (Blue – Advanced Group)

- Appears mainly in High Solar Tier, with a small presence in Medium.
- Likely reflects urbanized or higher electrification areas.

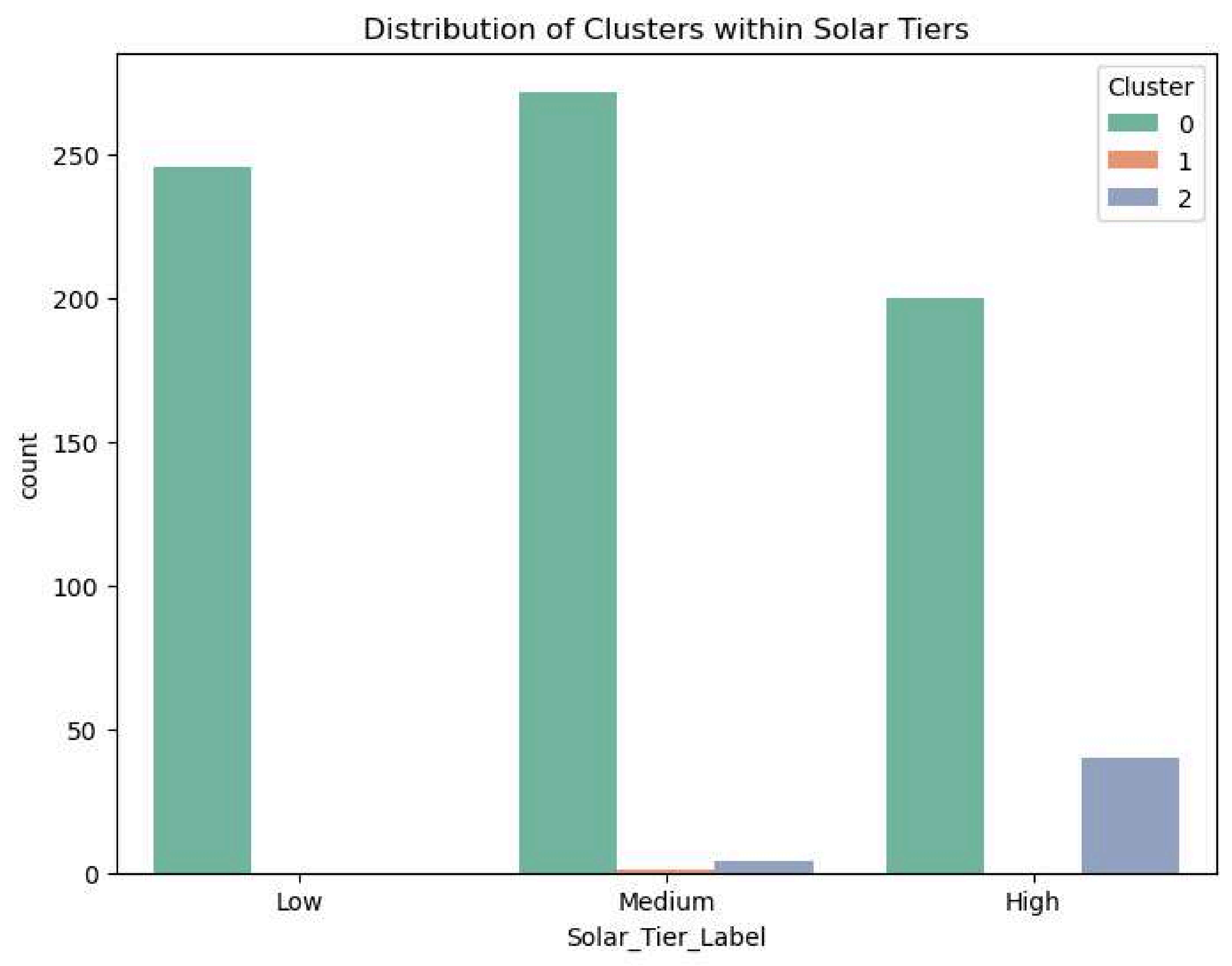
Cluster 1 (Red – Rare/Outliers)

- Very small cluster, nearly absent in all tiers.
- Could represent outliers or special cases.

Interpretation & Insight

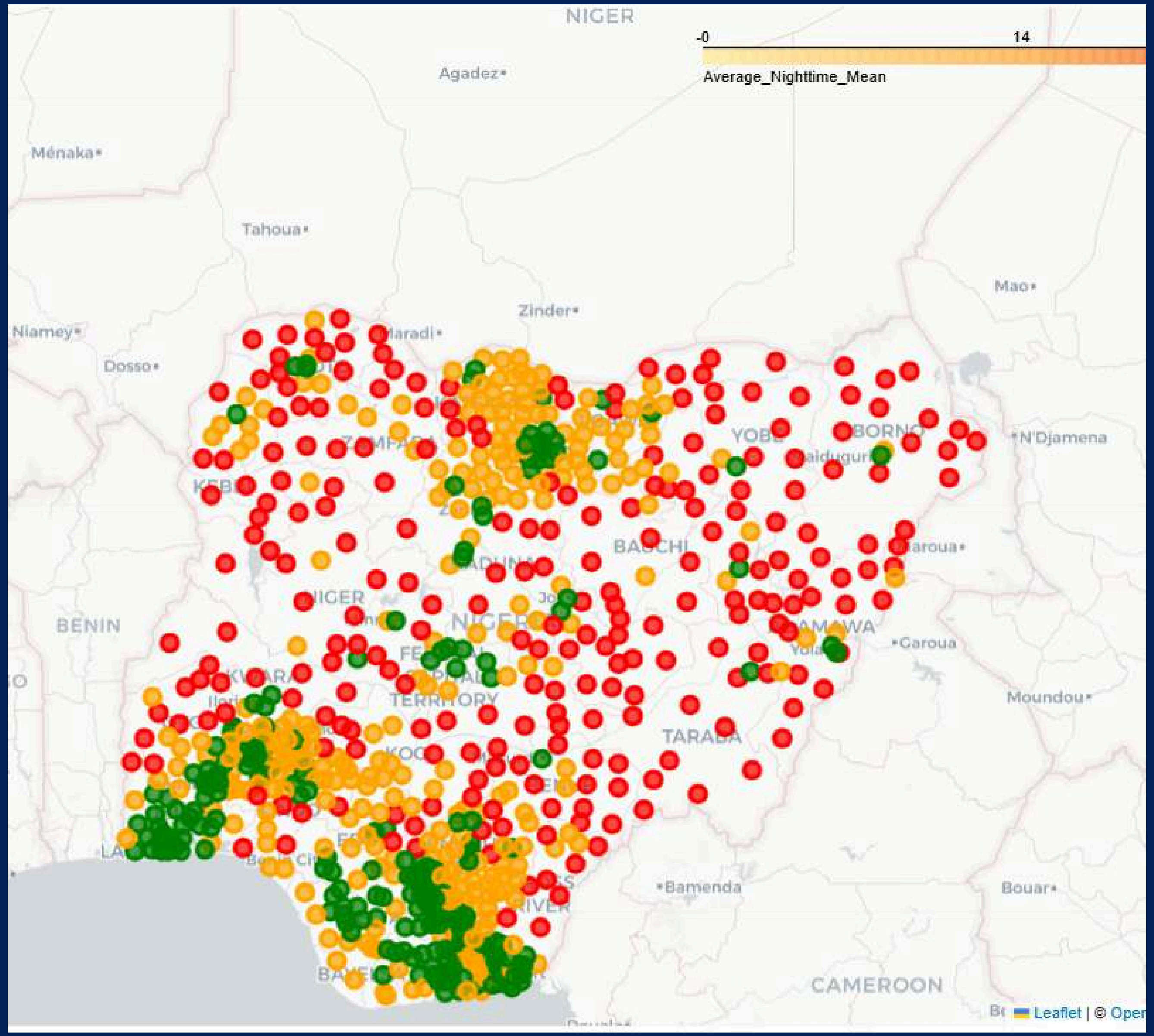
- Clusters do not map perfectly onto solar tiers.
- Partial alignment exists:
  - High Tier - more likely in Cluster 2.
  - Low & Medium Tiers - dominated by Cluster 0.
- Cluster 1 is too small to impact interpretation.

Clustering highlights broad population vs. advanced electrification profiles, but solar tiers cut across multiple clusters.



REGIONAL SOLAR PERFORMANCE CLASSIFICATION

This map illustrates the distribution of solar performance across regions in Nigeria, categorized into Low (red), Medium (yellow), and High (green). Most regions fall into the Low tier, showing widespread underperformance, while Medium-tier areas are scattered as transitional zones. High performance is concentrated in pockets of the southwest, southeast, and parts of north-central Nigeria, often linked to stronger infrastructure, urbanization, or richer solar resources.



Overall, the map highlights clear spatial patterns: urban and economically active areas achieve higher performance, while rural or less-developed zones lag behind. The dominance of Low-tier regions underscores the need for targeted solar investment and policy action, while the clustering of High-tier regions shows how local factors strongly influence solar outcomes.

METHODOLOGY

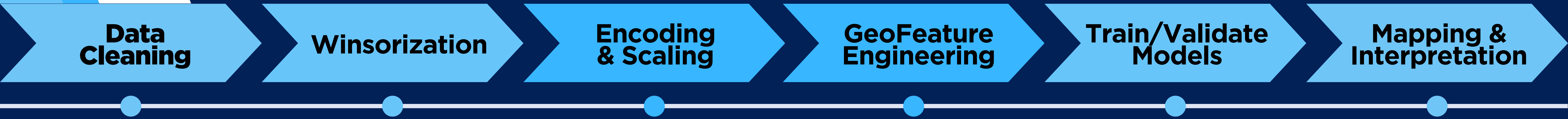
Overview

Approach: Supervised classification of Solar_Tier_Label (Low / Medium / High) + unsupervised clustering for regional segmentation.

Pipeline: Data cleaning → Winsorization (10% cap) → GeoFeature engineering → Encoding/Scaling → Train/validate models → Mapping & interpretation.

Targets & Tasks:

- Classification: Predict Solar_Tier_Label.
- Clustering: Group regions with similar water–energy–climate profiles (supporting targeting & program design).



Features & Engineering

Original features (examples): Water-source counts (Borehole, Tap, Handpump, Unimproved variants), Population, AREA, PERIMETER, Latitude/Longitude, Average_Nighttime_mean.

Engineered features:

- Distance_to_Center: LGA centroid → grouped-state centroid (access proxy).
- pop_density: Population / AREA.
- safe_water / unsafe_water and safe_ratio / unsafe_ratio.
- water_diversity: # of distinct sources present.
- climate_stress: Average_Nighttime_mean / pop_density.
- accessibility: Distance_to_Center / AREA.
- compactness: PERIMETER^2 / AREA.

Encoding & scaling:

- One-hot for GROUPED_STATE (after grouping states with count < 20 into "Others").
- Frequency-encoding for lga.
- StandardScaler for clustering-only numeric inputs.

Modeling Setup

Supervised model:

- RandomForestClassifier (n_estimators=200, class_weight="balanced", random_state=42).

Train/validation:

- Stratified 80/20 train–test split on Solar_Tier_Label.

Clustering:

- KMeans on scaled numeric features; k=3 chosen by inertia “elbow” and peak silhouette (~0.50 at k=2–3).

Model selection rationale:

- RF handles nonlinearity, mixed scales, and feature interactions; robust to outliers (further mitigated by Winsorization).
- KMeans provides complementary unsupervised segmentation to guide program design beyond the label.

MODELING & EVALUATION

Classification

Test performance (RF, 20% holdout):

- Accuracy: 0.93
- Per class:
  - High: Precision 1.00, Recall 0.90, F1 0.95 (support 48)
  - Low: Precision 0.92, Recall 0.98, F1 0.95 (support 49)
  - Medium: Precision 0.90, Recall 0.93, F1 0.91 (support 56)

Baseline context:

- Majority-class accuracy ≈ 36% (predicting “Medium” only) - RF improves accuracy by +57 pts.

Error profile:

- Most residual confusions occur between Medium - High, indicating overlapping boundaries among better-performing regions.

Feature Importance & Clustering

Feature importance (RF):

- Most influential: Average_Nighttime_mean (0.31), pop_density (0.20), AREA (0.13), climate_stress (0.12), Distance_to_Center (0.10).
- Moderate: Population, safe/unsafe ratios, compactness, water_diversity.
- Minimal: Pure location encodings (state dummies, lga_encoded).
- Implication: Quantitative development & geospatial factors drive tiers more than admin labels.

Clustering (KMeans, k=3):

- Cluster 0 (majority): Spans Low-High, concentrated in Low/Medium.
- Cluster 2 (selective): Skews toward High tier (advanced/electrified profile).
- Cluster 1 (tiny): Rare/noisy segment.

RESULTS & INSIGHTS

Determinants of solar performance: Brightness at night (development proxy), density, area/shape, and climate-access factors strongly explain tiers; admin region alone adds little.

Spatial patterns: High-tier clusters concentrate in pockets (southwest/southeast; some north-central); Low-tier dominates broadly, implying clear geographic inequality.

Segmentation signal: Cluster 2 aligns with High-tier regions (likely stronger infrastructure/electrification); Cluster 0 represents the general population; Cluster 1 is negligible.

Water–energy link: Unsafe water burden and climate stress co-occur with lower tiers and lower-density/peripheral geographies, informing integrated interventions.

RECOMMENDATIONS

Prioritize Low-tier, high-need regions:

- Water: Boreholes/taps, treatment, and network reliability upgrades where unsafe_ratio is high.
- Energy: Decentralized solar mini-grids + storage; target areas with high climate_stress and low accessibility.

Scale differentiated strategies by segment:

- Cluster 2 (advanced): Grid-tied/behind-the-meter solar, smart inverters, demand response; accelerate rooftop + C&I.
- Cluster 0 (general): Modular mini-grids, PAYGo SHS, phased capacity additions; bundle with water pumping/treatment.

Resource allocation: Use Priority_Index & combined_need_index to apportion budgets proportionally (you computed state-level allocations already).

Program design:

- Co-target water + energy (solar pumping + purification).
- Focus on peri-urban/rural nodes with long Distance_to_Center and low pop_density.
- Leverage geospatial screening (nighttime lights + density) to pre-rank sites.

M&E & risk: Track tier migration over time; monitor performance in Medium↔High boundary regions; hedge climate risk with storage and efficient end-uses (e.g., solar pumping + chlorination)

CONCLUSION

Recap: Built a geospatial ML pipeline to classify solar tiers, segment regions, and integrate water–energy–climate needs into a prioritization framework.

Key takeaways:

- Nighttime lights, density, area, climate stress, and access shape solar performance far more than admin labels.
- RF achieves strong accuracy (93%) vs. 36% baseline; clustering reveals an advanced segment aligned with High tiers.
- Most regions are Low tier → targeted, integrated water–energy investments are needed.

Next steps:

- Enrich labels with measured solar yield/uptime; add DNI/GHI, grid proximity, market variables.
- Test calibrated probability models, monotonic GBMs, and spatial CV; assess temporal drift.
- Operationalize a site-screening tool (dash/map) to route budgets using the Priority/Need indices.

Water Energy Synergy

THANK YOU

oumaben2000@gmail.com