Energy Consumption and Power Demand

Abstract

This project investigates the relationship between energy consumption and its potential influencing factors, focusing on regional and temporal variations. Using a dataset of energy consumption records across multiple network sites, combined with expected energy demand data for the year 2013, the analysis explores whether significant differences exist in consumption patterns across regions and time periods. Preprocessing techniques were applied to handle missing values and outliers, ensuring the integrity of the dataset. Key challenges included determining the appropriate threshold for missing records and interpreting the significance of zero values. The results reveal discernible patterns in consumption across regions and times of the day, offering valuable insights into energy demand distribution. These findings can support strategies for optimizing resource allocation and understanding regional energy dynamics.

Sara Ali Mahmoud Ibrahim

Tech Her Up Program
Data Science Track

1 Data Exploration and Analysis

1.1 Dataset Description

Energy Consumption.csv: This dataset, split into 3 files, contains energy consumption records for 82 unique cells across four regions (A, B, C, and D) for the year 2013. Key columns include:

- DateTime: Time intervals of 30 minutes throughout the year.
- KWH/hh (per half hour): Energy consumed during each half-hour interval.
- Region: The geographical area of the cell's parent site.
- Site_id: Unique identifier for each network site, containing multiple cells.
- Cell_id: Unique identifier for each cell.

Power Demand.xlsx: This dataset provides half-hourly energy consumption demand for the same period, with columns:

- DemandDateTime: Timestamps of the energy demand in a 30-minute interval throughout the year.
- Demand: Recorded energy demand for the interval.

Demand Prices: The cost associated with energy demand was provided in three categories:

- High Demand: E£0.7721/kWh
- Normal Demand: E£0.1946/kWh
- Low Demand: E£0.0689/kWh

1.2 Data Cleaning and Preparation

Power Demand:

Data Connection: The data was read from the Excel file and stored into a DataFrame.

Standardizing Values: The values of the demand category were standardized to have consistent capitalization and correct any typos.

Energy Consumption:

Data Integration: The three files containing energy consumption data were concatenated into a single DataFrame. Columns were harmonized, and the DateTime column was converted to a datetime format to facilitate temporal analysis.

Feature Engineering: Additional features were derived from the DateTime column, including hour, day, and month. These features allowed for temporal grouping and trend analysis.

Handling Duplicates: A total of 942 distinct duplicate rows were identified in the data and therefore were removed.

Handling Missing Records: Although no direct null values were found, missing records were identified due to incomplete data for certain cells. The following steps were taken:

- Cells with fewer than 90% of the expected 17,520 records (representing a full year of 30-minute intervals) were excluded. 5.1

- A complete record of DateTime was reconstructed, and missing energy consumption values were imputed using the median consumption at the cell level.

Identifying Zero Values: Sixteen cells across Regions A, B, and C were found to have zero energy consumption at certain timestamps. Notably, two cells—MAC000037 and MAC000004, both located in Region A—showed zero consumption for most of the year. These zero values were retained, as they represent shut-down cells, but further analysis is needed to determine the underlying reasons.

1.2.1 Descriptive Statistics

After cleaning the data, the following statistics of the energy consumption DataFrame were recorded.

Energy Consumption Distribution Across Regions

Central Tendency

- Overall Mean (0.243 kWh): The average energy consumption is relatively low, with most regions displaying moderate usage.
- Median (0.136 kWh): The median is lower than the mean, reflecting a right-skewed distribution influenced by extreme high values.
- Regional Means:
  - Region A: Highest at 0.296 kWh.
  - Region D: Lowest at 0.186 kWh.

Variation

- Overall Standard Deviation (6.93 kWh): A large variation across all regions suggests the presence of extreme outliers.
- Regional Patterns:
  - Region A exhibits the largest variation (11.07 kWh), driven by extreme high-demand events.
  - Regions B, C, and D display smaller variability.

Quartile Analysis

- 25th Percentile (0.065 kWh): Indicates periods of very low usage, common across all regions.
- 75th Percentile (0.264 kWh): Most consumption values are below this threshold.
- Regional Quartiles:
  - Region A shows higher median (0.151 kWh) and 75th percentile values (0.318 kWh), emphasizing its relatively higher consumption levels.
  - Region D has the lowest median (0.101 kWh) and more compact quartiles, reflecting stable low-demand patterns.

Outliers

Maximum Values:

- Region A: 7,657 kWh, significantly larger than other regions and indicative of potential anomalies.
- Region C: 999 kWh, less extreme but still a high outlier.
- Regions B and D exhibit less pronounced outliers (3.757 kWh and 2.255 kWh, respectively).

Temporal Distribution

Uniform Time Coverage: Across all regions, data spans evenly from 2013-01-01 to 2013-12-31, with no clustering or missing intervals.

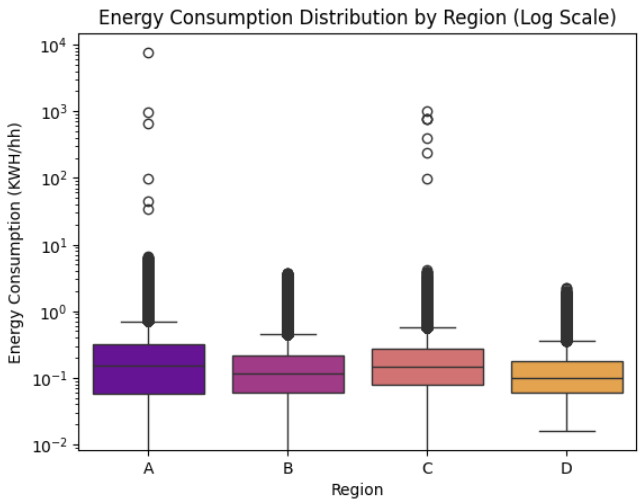
Key Insights

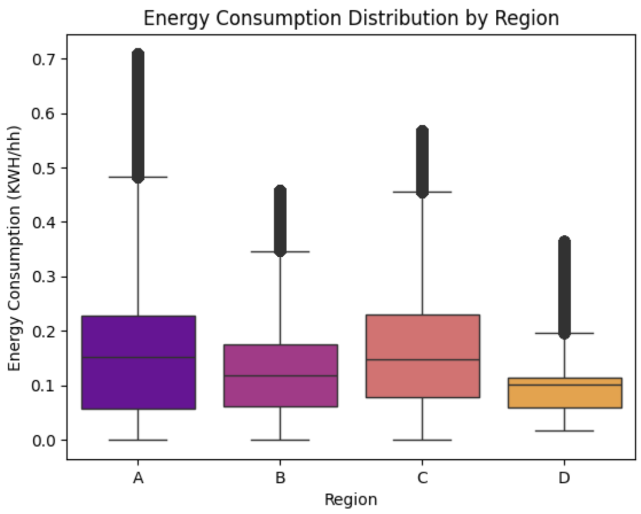
- Region A exhibits consistently higher and more variable energy consumption, likely due to its extreme outliers.
- Regions B, C, and D show lower and more stable patterns, with Region D demonstrating the least variability.
- Outliers need further investigation to ensure modeling accuracy.
- Despite regional differences, temporal coverage remains balanced, making the data suitable for cross-regional comparisons.

1.3 Anomaly Detection and Treatment

The interquartile range (IQR) method was applied on a regional level 5.2 to identify anomalies in energy consumption. Anomalous values were replaced with the regional medians to maintain dataset integrity while minimizing distortion.

The box plots below illustrate the distribution of energy consumption by region before and after anomaly treatment using the IQR method. The left plot shows the original data, revealing significant outliers, especially in Regions A and C, due to extreme values. After applying the IQR method and replacing anomalies with regional medians, the right plot demonstrates a more balanced distribution with reduced skewness and outliers, improving the data's integrity for subsequent analysis.





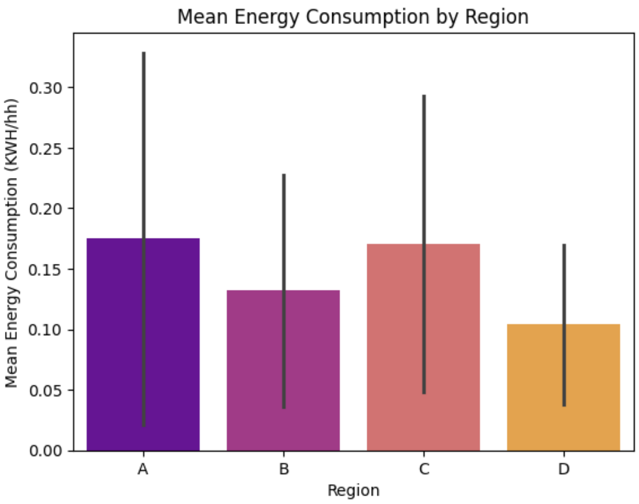
1.4 Business Questions and Insights

1.4.1 Does the Region Significantly Affect Energy Consumption?

Objective This question aims to identify whether energy consumption patterns vary significantly across different regions. Understanding regional differences can help management implement targeted strategies to reduce energy costs or optimize resource allocation.

To determine whether region significantly affects energy consumption, we explored the data using three visualizations: a bar plot of mean energy consumption, a bar plot of energy consumption variability, and a heatmap of monthly consumption.

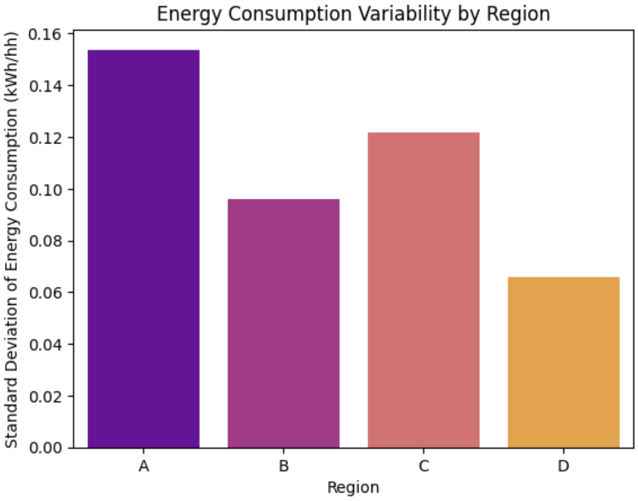
- Mean Energy Consumption by Region:



  - Regions A and C show similar mean energy consumption (around 0.17 KWh/hh), followed closely by Region B at approximately 0.13 KWh/hh.

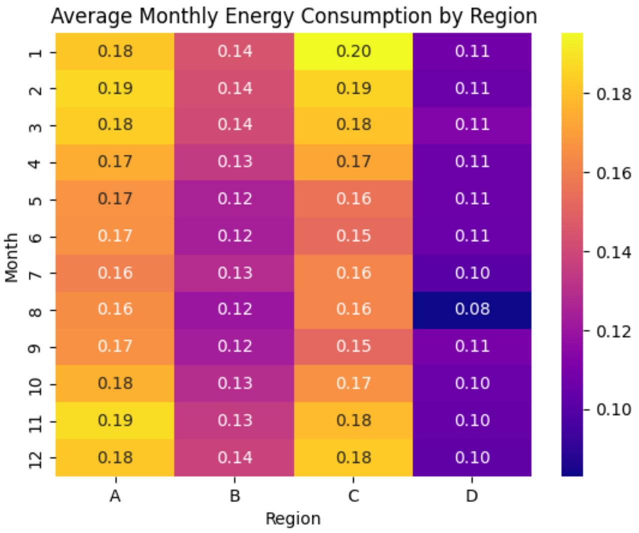
  - Region D has the lowest mean energy consumption (0.10 KWh/hh), indicating a significant drop compared to other regions.

- Variability of Energy Consumption by region:



  - Region A shows the highest variability (0.15 KWh/hh), suggesting inconsistent energy usage patterns.
  - Region C also exhibits noticeable variability (0.12 KWh/hh), while Region B has moderate variability (0.10 KWh/hh).
  - Region D has the least variability (0.06 KWh/hh), indicating stable but low energy consumption.

- Monthly Energy Consumption by Region



  - Regions A, B, and C exhibit relatively higher energy consumption compared to Region D across all months.
  - Regions A and C consistently show higher energy consumption values (e.g., 0.17–0.20 KWh/hh), while Region D has notably lower values (0.08–0.11 KWh/hh), particularly in August.
  - Energy consumption in Region B remains relatively stable but is lower than Region A and C.

Insights

Regional Differences Exist

The heatmap and bar plots consistently highlight Region D as an outlier with both lower mean consumption and lower variability. This indicates that energy usage in Region D is both lower and more stable across months. Regions A and C emerge as high-consumption regions with greater variability, suggesting dynamic energy demands. Region B, on the other hand, shows moderate energy consumption (lower than Regions A and C but higher than Region D) with relatively low variability. This implies a more stable but modest energy demand.

Possible Underlying Causes

- Regions A and C may represent areas with higher population density, industrial activity, or climatic conditions that drive up energy needs.
- Region B may represent a mix of residential areas or areas with less industrial or seasonal energy influence.
- Region D could represent rural or sparsely populated areas with limited energy consumption.

Seasonal Stability

Monthly variations across regions are minimal for Regions B and D, implying consistent energy usage patterns. However, the variability in Regions A and C suggests some seasonal or behavioral factors.

Conclusion

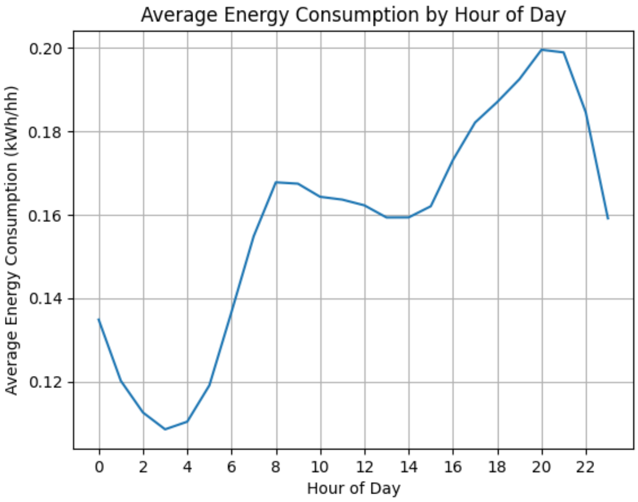
The visualizations collectively suggest that region significantly affects energy consumption, with notable disparities in both mean values and variability. Regions A and C demonstrate higher and more dynamic energy usage, while Region D stands out for its consistently low and stable consumption. These findings underline the need for region-specific strategies in energy planning, anomaly detection, and infrastructure investment.

1.4.2 Is there a relationship between the time of day and energy consumption?

Objective This question aims to identify whether energy consumption patterns vary significantly during different times of the day.

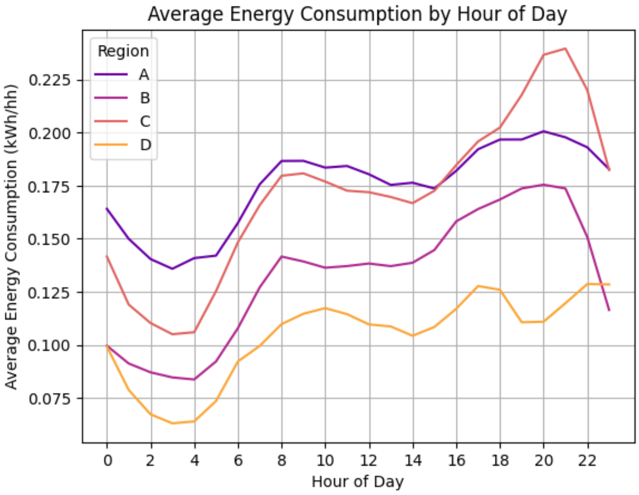
To explore the relationship between the time of day and energy consumption, we analyzed average energy usage patterns using three visualizations:

- Overall Hourly Energy Consumption Trend



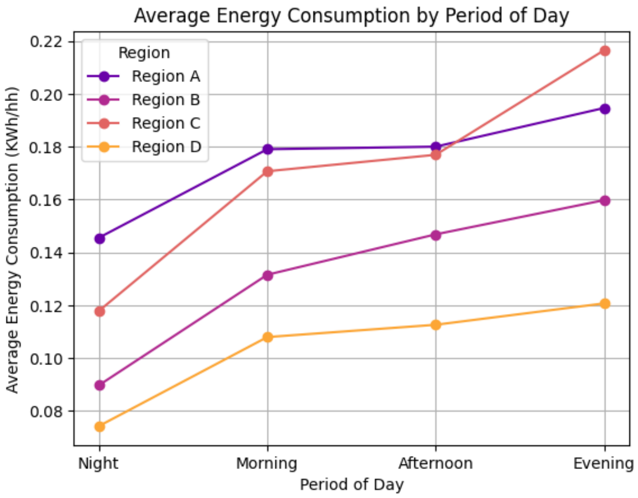
  - The aggregate energy consumption pattern across all regions confirms a sharp rise in usage from the early morning, stabilizing slightly around midday, and then rising to a peak during the evening hours (around 8 PM).
  - Consumption begins to taper off sharply after 9 PM, consistent with reduced activity levels during nighttime.

- Hourly Energy Consumption Trends by Region



  - Regions exhibit distinct daily energy consumption patterns.
  - Region A and Region B show gradual increases in energy usage from early morning (around 6 AM), peaking during the evening hours (6–8 PM), followed by a decline.
  - Region C has the highest energy consumption during the evening, with a notable spike after 6 PM, indicating substantial activity during these hours.
  - Region D shows relatively lower overall energy consumption, with modest fluctuations and a gradual rise throughout the day.

- Energy Consumption by Period of Day



- Energy consumption increases consistently from night to evening across all regions.
- Region C leads in energy usage during the evening, confirming the peak observed in the previous plots.
- Region D consistently exhibits the lowest energy usage across all periods, indicating a more energy-efficient profile or less demand.
- The morning period shows similar usage levels across most regions, suggesting uniform activity patterns during this time.

Insights

- **Time-of-Day Dependence:** All three visualizations consistently demonstrate a strong relationship between energy consumption and the time of day, with evening hours showing peak usage across regions.
- **Regional and Aggregate Patterns:** The regional trends align with the overall hourly trend, but the magnitude of consumption and variability differ significantly by region, reflecting unique socio-economic activities.
- **Efficiency Opportunities:** Region D's consistently lower consumption highlights potential energy efficiency strategies or less intensive demand, offering lessons for regions with higher peaks.

Conclusion

In conclusion, the analysis provides robust evidence of a strong relationship between the time of day and energy consumption. These insights can guide targeted strategies for improving energy efficiency and ensuring sustainable energy supply systems.

2 Statistical Analysis of Energy Consumption Patterns

To deepen our understanding of energy consumption patterns, we investigated two critical questions using statistical methods:

1. **Does the region significantly affect energy consumption?**
2. **Is there a relationship between the time of day and energy consumption?**

For both questions, we conducted Analysis of Variance (ANOVA) tests to evaluate the significance of differences in energy consumption across times of the day and regions, and to be able to generalize our previous conclusions.

Question 1: Does the Region Significantly Affect Energy Consumption?

Here the ANOVA test compares the means of energy consumption across different regions. The test was performed using the pingouin library in Python, with the column representing energy consumption as the dependent variable and the Region column as the independent variable.

Hypotheses:

- **Null Hypothesis ($\mathbf{H}_0$):** There is no significant difference in mean energy consumption across regions.
- **Alternative Hypothesis ($\mathbf{H}_A$):** There is a significant difference in mean energy consumption across regions.

Results:

- The ANOVA test yielded a **p-value of 0**, well below the significance threshold of $\alpha = 0.05$.
- The null hypothesis (H_0) is therefore **rejected**, indicating that there are statistically significant differences in energy consumption across regions.

Interpretation:

The results validate that regional factors play a crucial role in determining energy consumption patterns. These differences could stem from variations in:

- Population density and urbanization levels,
- Socio-economic activity (e.g., industrial, commercial, or residential usage),

- Climatic conditions influencing heating or cooling needs, and
- Adoption of energy-efficient technologies or practices.

Question 2: Is There a Relationship Between the Time of Day and Energy Consumption?

The ANOVA test was performed with the column representing energy consumption as the dependent variable and the Hour column as the independent variable.

Hypotheses:

- **Null Hypothesis (H_0):** There is no relationship between hour of day and energy consumption. The mean energy consumption remains consistent across all hours.
- **Alternative Hypothesis (H_a):** There is a relationship between hour of day and energy consumption. The mean energy consumption differs significantly across hours.

Results:

- The ANOVA test produced a **p-value of 0**, which is far below the typical significance level of $\alpha = 0.05$.
- This result leads to a **rejection of the null hypothesis (H_0)**, providing strong evidence that energy consumption depends significantly on the time of day.

Interpretation:

The findings confirm that energy consumption fluctuates throughout the day, as seen in the earlier visual analyses. Consumption typically rises in the morning, stabilizes during midday, and peaks during the evening hours.

Conclusion:

In conclusion, these statistical analyses reinforce the critical need to consider both time-of-day and regional variations in energy consumption to design efficient, equitable, and sustainable energy systems.

3 Implications

The findings highlight the need for a comprehensive, region-specific, and time-sensitive approach to energy management, encompassing policy, infrastructure, technology, and consumer engagement. Key implications include:

3.1 Demand Management and Time-Sensitive Strategies

The strong evening peak consumption across all regions underscores the need for demand management strategies, such as:

- Dynamic pricing models
- Time-of-use tariffs
- Demand-response programs

These strategies aim to incentivize off-peak energy usage and reduce grid strain.

3.2 Infrastructure Planning and Investment

Regions with higher energy demand variability, such as Regions A and C, should receive targeted infrastructure investments to ensure reliable evening demand fulfillment and overall grid stability. Investments should also focus on Region D, where energy accessibility efforts could address potential infrastructural limitations and support efficiency improvements.

3.3 Equity and Accessibility

Energy policies should address equitable energy access, ensuring under-resourced areas and communities benefit from infrastructure upgrades and energy efficiency initiatives. Regions like Region D, with lower energy consumption and higher efficiency, could serve as models for best practices, showcasing strategies for peak-load reduction and sustainable energy use.

3.4 Resilience to Environmental and Economic Factors

Future energy planning should incorporate climate resilience, anticipating regional changes that may affect consumption patterns and demand spikes.

- Strategic investments in energy infrastructure and efficiency programs offer not only environmental benefits but also economic advantages.
- Such investments help reduce costs for consumers and energy providers alike.

By addressing these areas, energy management strategies can be more resilient, cost-effective, sustainable, and adaptable. This ensures reliable service, reduced environmental impact, and equitable energy access across all regions.

4 Optimization: Energy and Cost Savings Analysis

Objective

The goal of this optimization was to identify and reduce energy consumption below a certain threshold while quantifying the resulting cost savings. This was achieved by targeting low-energy consumption periods and selectively shutting down the respective cells.

Baseline Metrics

- **Total Energy Cost (Before Optimization):** 43,704.15 EGP
- **Total Energy Consumed (Before Optimization):** 206,919.66 kWh

Optimization Approach

1. Threshold Definition

- A threshold was defined using the 25th percentile of non-zero energy consumption values.
- This threshold represents the energy usage level below which systems are deemed low-consuming and suitable for optimization.
- **Global 25th Percentile Threshold:** 0.07 kWh

2. Merging Energy and Demand Data

The energy consumption DataFrame was merged with the demand DataFrame based on the timestamp:

This combined dataset allows for aligning energy consumption with demand levels.

3. Demand-Based Pricing

A demand-based pricing structure was applied:

- **High Demand:** 0.7721 EGP per kWh
- **Normal Demand:** 0.1946 EGP per kWh
- **Low Demand:** 0.0689 EGP per kWh

This pricing is used to calculate the cost based on energy consumption levels.

4. Mapping Demand Prices to Energy Consumption

The prices were then mapped to the corresponding demand levels in the combined dataset in order to associate the correct price per kWh with each energy consumption value based on the demand level.

5. Filtering Low-Energy Rows

Energy consumption data was filtered to identify rows where the energy consumed was below the threshold.

6. Energy and Cost Savings Calculation

The energy and cost savings were determined by summing over the rows where cells were identified as being “shut down” due to their energy consumption falling below the defined threshold.

Optimization Results

- **Total Energy Saved:** 13,841.88 kWh
- **Total Cost Saved:** 2,827.48 EGP

Conclusion

By applying a 25th percentile threshold of 0.07 kWh, we successfully reduced energy consumption and costs by approximately 6.7% and 6.5%, respectively. These results demonstrate the potential of data-driven optimization strategies to enhance energy efficiency while maintaining cost-effectiveness. Further analyses could explore thresholds at different percentiles to identify additional savings opportunities.

5 Challenges:

In this section, we address key challenges encountered in analyzing the energy consumption dataset, including the need to ensure data completeness and the strategies for handling missing records and anomalies. These challenges are crucial to maintaining the integrity and reliability of the analysis while drawing meaningful insights about energy consumption patterns across regions and time.

5.1 Justification for 90% Completeness Threshold

The primary goal of this study is to analyze patterns in energy consumption across regions and time. To ensure reliable and meaningful insights, it is essential to maintain a high level of data completeness. A completeness threshold of 90% was chosen as it strikes a balance between retaining a sufficient number of cells across all regions and ensuring that the dataset's quality is not compromised.

Cells with less than 90% completeness would require imputing more than 10% of their records, which could introduce noise or bias and weaken the robustness of the results. By setting this threshold, we minimize the risk of unreliable outcomes while maintaining a representative sample of the dataset. This approach is consistent with best practices in data analysis, where high completeness levels are critical for preserving data reliability and interpretability.

While the 90% threshold may not be universally ideal, it was deemed appropriate for this study based on the dataset size, the distribution of missing data, and the specific objectives of the analysis. This ensures the retained data provides a solid foundation for uncovering meaningful consumption patterns across regions and time.

5.2 Handling Missing Records and Anomalies in Energy Consumption DataFrame

Imputation Strategy

A dual-level imputation strategy was employed to address missing values and outliers while maintaining data integrity.

Missing Records

Missing records were imputed using the median energy consumption at the cell level, preserving localized trends and unique consumption patterns of individual cells.

Outliers

Outliers were identified using the Interquartile Range method and replaced with the regional-level median energy consumption. This ensures that extreme anomalies do not distort local trends and align with broader regional consumption patterns.

Rationale for the Approach

Preserving Cell-Level Patterns

Imputing missing values at the cell level maintains granularity, which is essential for identifying low-consumption cells and localized trends.

Minimizing Anomalous Impact

Regional-level outlier imputation smooths anomalies caused by measurement errors or rare events, maintaining regional trends without sacrificing broader consumption insights.

Alignment with Analysis Goals

The focus is on analyzing regional consumption differences and identifying low-consumption cells for potential shutdowns. This approach balances cell-level detail and regional-level aggregation, ensuring comprehensive insights.

Comparison of Outlier Imputation Methods

A comparison of outlier imputation methods revealed that:

- **Cell-Level Outlier Imputation:** Introduced higher variability and localized extremes, which risk distorting regional patterns.
- **Regional-Level Outlier Imputation:** Effectively smoothed anomalies while preserving essential regional trends.

Based on this comparison, regional-level median imputation was selected to ensure precision, consistency, and data integrity for robust analysis and actionable insights.