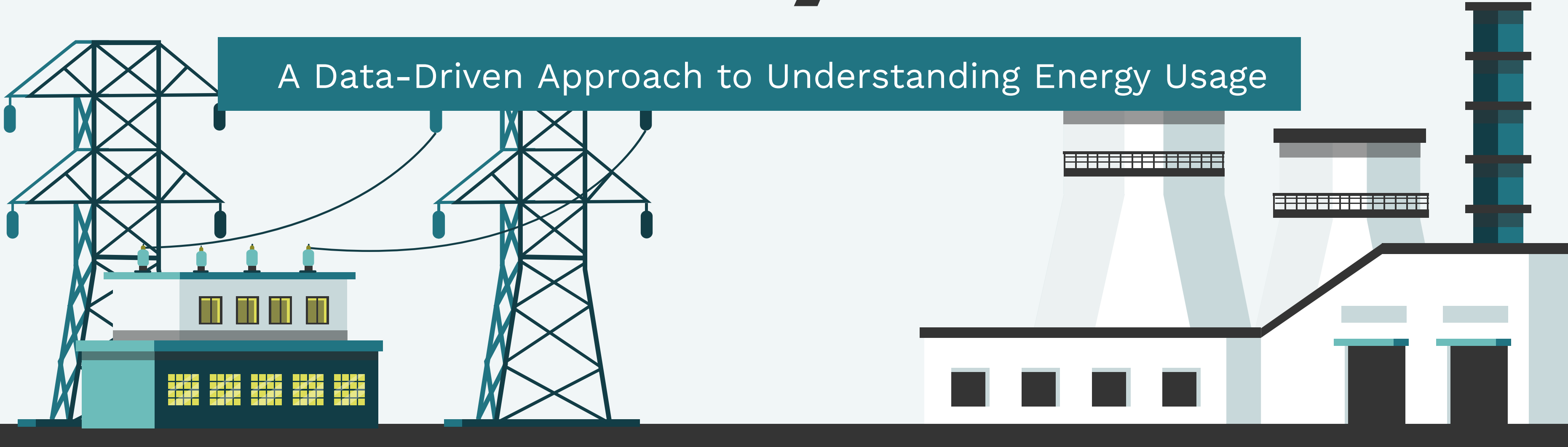
Energy Consumption Analysis

A Data-Driven Approach to Understanding Energy Usage



Introduction

Project Overview

- Title: Energy Consumption Analysis & Optimization
- Goal: To analyze energy consumption patterns and identify opportunities for optimizing energy usage to reduce costs and improve efficiency.

Motivation

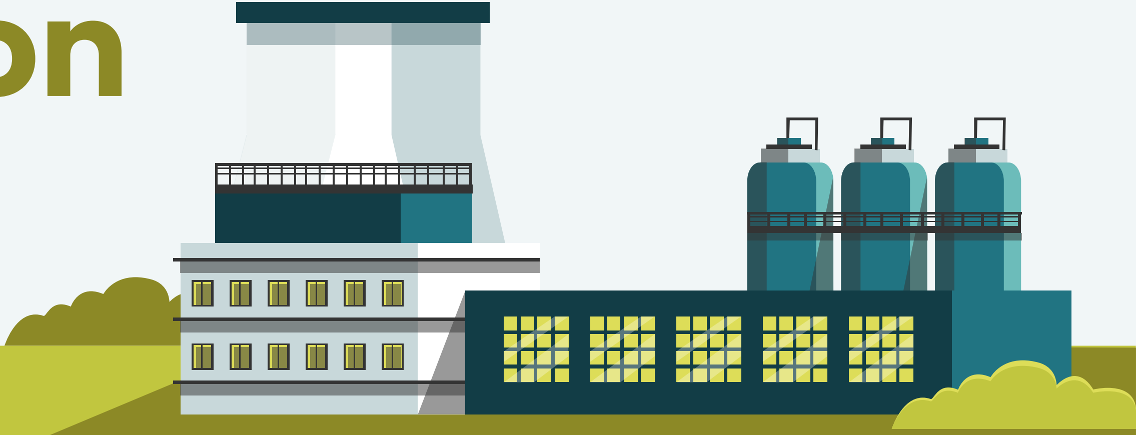
- The growing demand for energy, coupled with the need for sustainable practices, makes it crucial to understand consumption patterns.
- Optimizing energy usage leads to cost savings, energy efficiency, and environmental benefits.

Objectives

- Explore consumption patterns and trends.
- Detect anomalies and outliers in the energy data.
- Calculate potential energy savings and cost reductions based on different demand levels.

Table of content

01
Data Collection



Data Sources

Sites Energy Consumption Data

- DateTime: Timestamp of energy consumption.
- Site_id: unique site identifier contains more than one cell.
- Cell_id: unique cell identifier.
- Region: geographical area that identifies site location.
- KWH/hh (per half hour): Energy consumed in half hour intervals.

Rows: 1,380,252 records
Columns: 5

Demand Data

- DemandDateTime: Timestamp corresponding to the demand data.
- Demand: Categorical field indicating the demand level (High, Normal, High).

Rows: 17,520 records
Columns: 2

02

Data Exploration and Analysis

Exploratory Analysis

Data Overview

Inspecting the first and last few rows to understand the structure

Describe Data

Statistical overview of the dataset

Data Structure

Understand the basic structure of the dataset

Data Distribution

Plotting histograms and box plots to visualize the distribution of data features

Data Cleaning and Preparation

Handle Missing Values

The dataset has no missing values.

Drop Duplicates

There are 942 duplicate rows, which have been removed to maintain data quality and avoid bias

Change Date Format

Convert the 'DateTime' column, from an object (string), into a datetime format for easier time-based analysis and manipulation

Data Consistency

Standardized column names and values for consistency and clarity across the dataset.

Merging

Merged `sites_energy` and `demand` DataFrames on the 'DateTime' column to consolidate data.

Anomaly **Detection**

Objective

To identify and handle anomalies in the KWH/hh (per half hour) column to ensure data quality and consistency.

Steps Applied

1. Calculated IQR (Interquartile Range):
   - Q1 (25th Percentile): Found the lower quartile.
   - Q3 (75th Percentile): Found the upper quartile.
   - IQR: Calculated as $IQR=Q3-Q1$IQR=Q3−Q1.

2. Determined Outlier Boundaries:
   - Upper Bound: $Q3+1.5\times IQR$Q3+1.5×IQR
   - Lower Bound: $Q1-1.5\times IQR$Q1–1.5×IQR

3. Identified Outliers:
   - Detected 120,217 rows with outlier values.

4. Replaced Outliers:
   - Replaced all outlier values with the median value of the KWH/hh (per half hour) column.

Outcome

Successfully handled anomalies, resulting in a cleaner and more reliable dataset.

Demand Prices

Objective

This step involves calculating the energy cost for each entry in the DataFrame based on the demand type and the corresponding price per kWh.

- **Mapping Demand Prices**

The demand column in the DataFrame is mapped to its corresponding price per kWh using the demand_prices dictionary. The prices are assigned to a new column, 'Demand Price'.

- **Calculating Energy Cost**

The 'Energy Cost' is then computed by multiplying the 'KWH/hh (per half hour)' by the 'Demand Price' for each row. This gives the total cost of energy consumption for each half-hour period.

Benefits

This process allows for dynamic energy cost calculation based on demand and consumption data.

Demand	Demand Price
High	77.21
Normal	19.46
Low	6.89

Feature Engineering

Objective
To derive meaningful features from the DateTime column for better temporal analysis.

Steps Applied
1. Extracted Date and Time from the DateTime column.
2. Derived the Hour feature for hourly trend analysis.
3. Dropped the original DateTime column for simplicity.

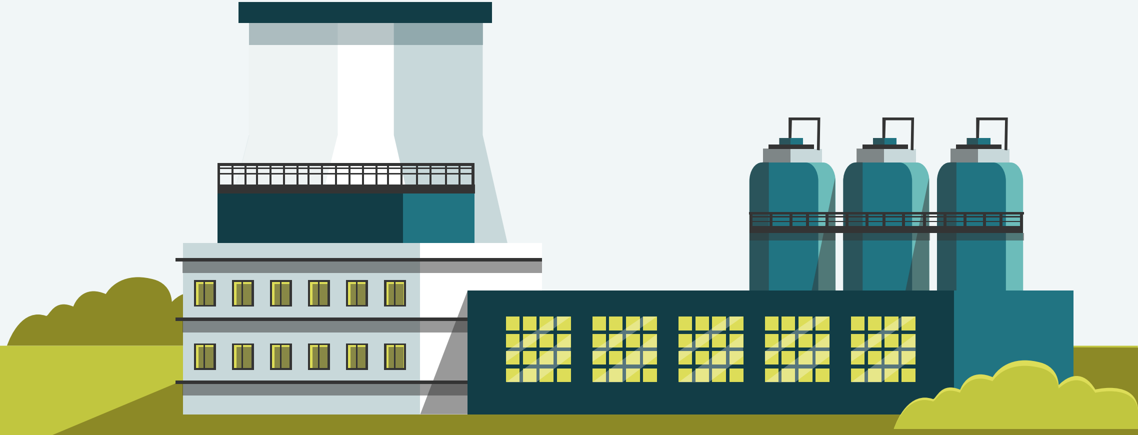
Benefits
- Enables focused analysis of daily, hourly, and time-specific patterns in energy consumption.
- Simplifies data structure for downstream analysis and modeling.

Outcome
The dataset now contains more interpretable and analysis-friendly time-related features.

cell_id	KWH/hh (per half hour)	site_id	region	Demand	Demand Price	Energy Cost	Date	Time	Hour
MAC000002	0.219	A	A	Normal	19.46	4.26174	2013-01-01	00:00:00	0
MAC000002	0.241	A	A	Normal	19.46	4.68986	2013-01-01	00:30:00	0
MAC000002	0.191	A	A	Normal	19.46	3.71686	2013-01-01	01:00:00	1
MAC000002	0.235	A	A	Normal	19.46	4.57310	2013-01-01	01:30:00	1

03

Data Insights



210.13K

Total Energy Consumption (KWh)

4.43M

Total Energy Cost

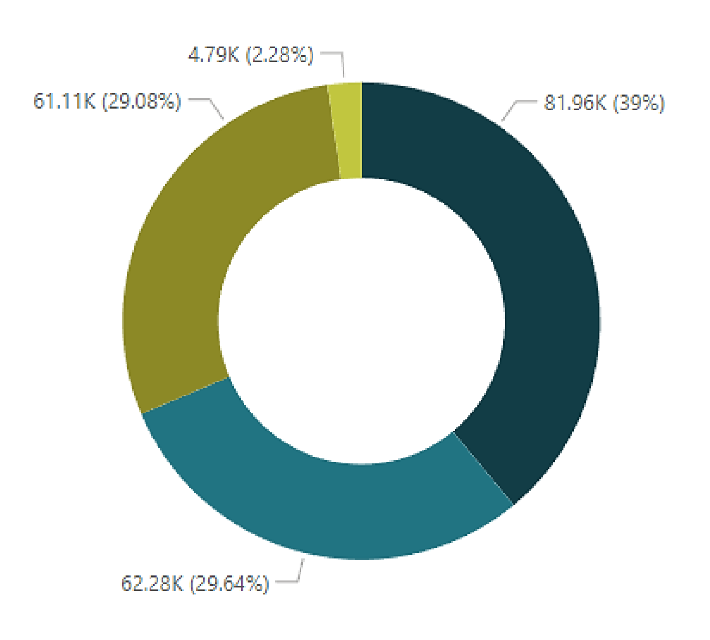
Total Energy Consumption by Region

Region A 39%

Region B 29.64%

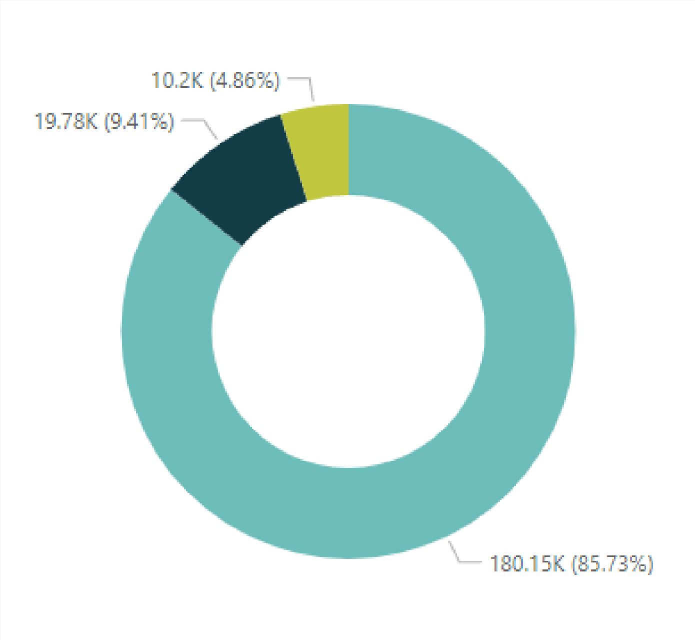
Region C 29.08%

Region D 2.28%



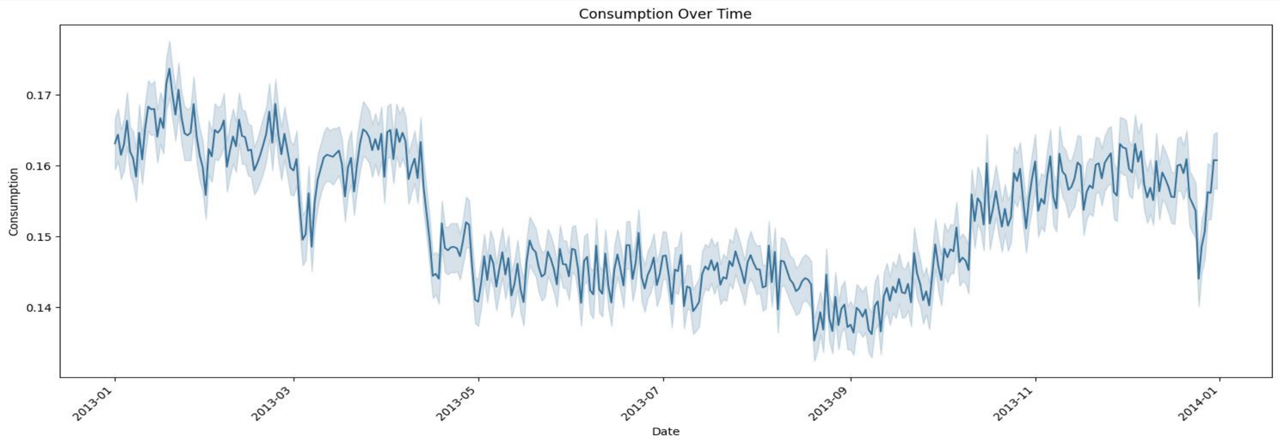
Total Energy Consumption by Demand

Normal Demand	85.73%
Low Demand	9.41%
High Demand	4.86%



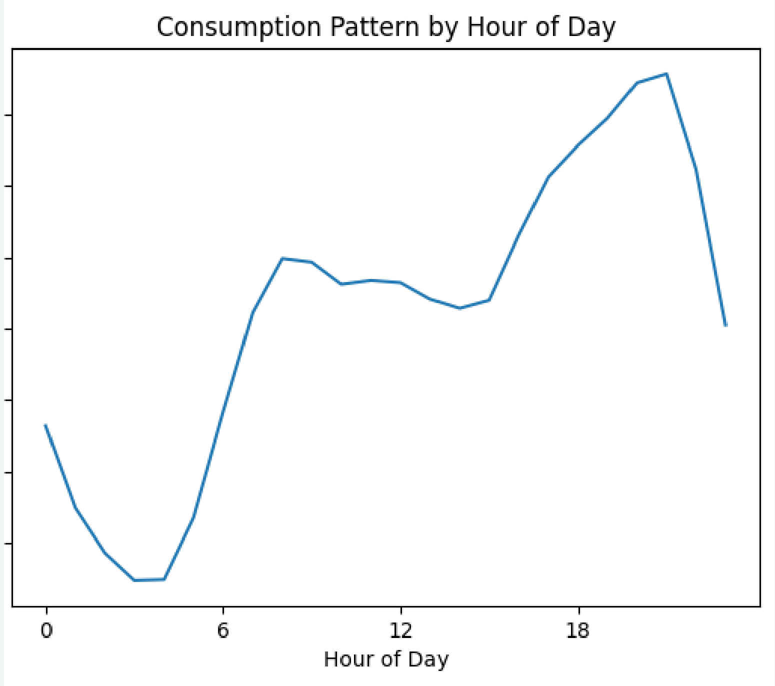
Energy Consumption Over Time

- Fluctuations in consumption are evident, with certain periods showing more significant variation.
- There is a seasonal increase towards the end of 2013, suggesting external factors like temperature changes could be influencing consumption patterns.
- The sharp rise in early 2014 could merit further investigation to understand if it is due to changes in demand, policy, or external influences.

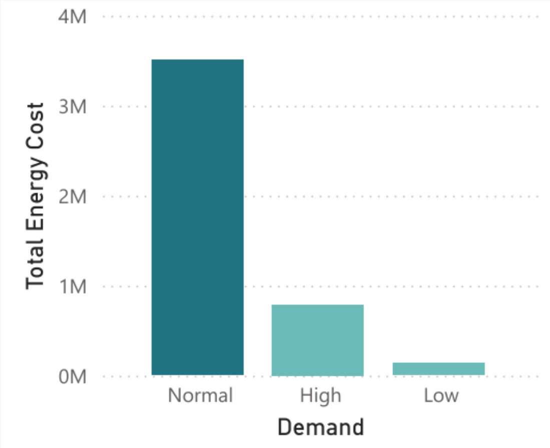


Average Energy Consumption by Hour of Day

- The highest consumption occurs during the evening hours (18:00–24:00), likely related to increased household activities.
- The lowest consumption is observed during the early morning hours (00:00–06:00), which is typical for most residential areas due to fewer activities taking place.
- Steady consumption in the late morning and early afternoon could suggest stable, continuous demand, potentially related to work hours or base usage.



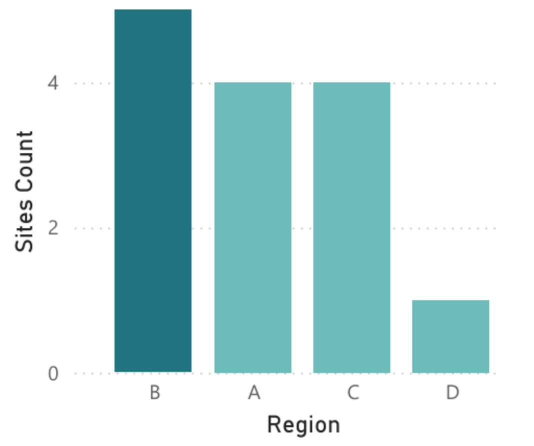
Energy Cost by Demand



Normal demand incurs the highest energy costs, significantly surpassing high and low demand, with a total of 3,505,674.42, indicating that most energy costs are driven by periods of normal demand.

Num of Sites by Region

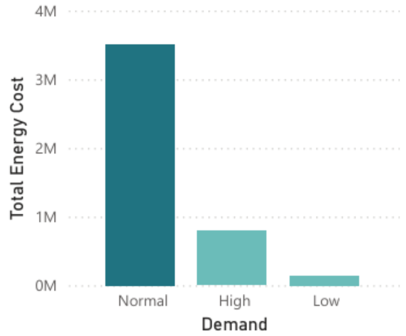
The distribution of sites across regions reveals that Region B has the highest number of sites (5), followed by Regions A and C with 4 sites each, and Region D with the lowest, having just 1 site.



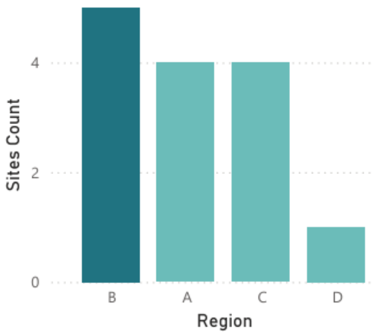
⚡ Energy Consumption Analysis

Num of Sites	Num of Regions	Num of Cells	Total Energy Cost	Total Energy Consumption (KWh)
14	4	82	4.43M	210.13K

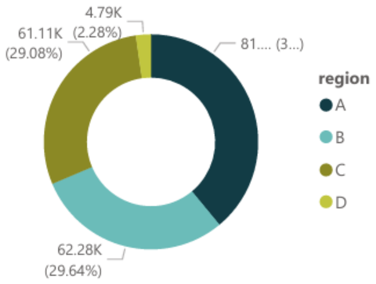
Energy Cost by Demand



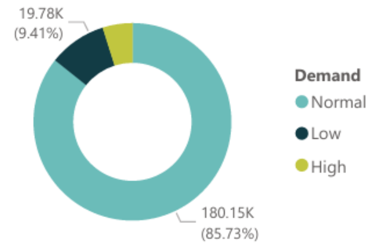
Number of Sites by Region



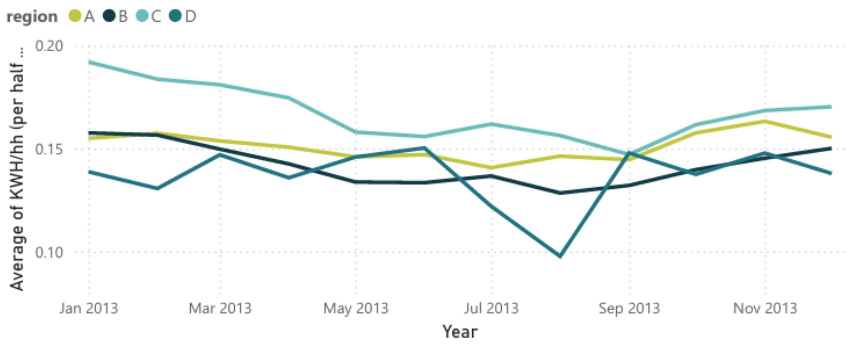
Total Energy Consumption by Region



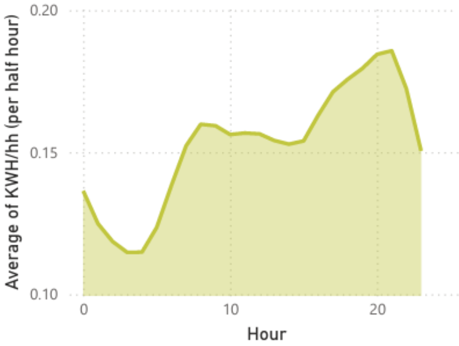
Total Energy Consumption by Demand Category



Average Energy Consumption by Time and Region



Average Energy Consumption by Hour of Day



Energy Pricing

Demand Price	Demand
77.21	High
19.46	Normal
6.89	Low

Energy Costs by Region

Sum of Energy Cost	region
1,715,402.10	A
1,321,746.37	B
1,291,014.05	C
101,643.08	D

04

Statistical Hypothesis Testing

Does the region significantly affect energy consumption?

Objective
Using ANOVA, Test if the region significantly affects energy consumption.
Hypotheses
- Null Hypothesis (H_0): No significant difference in energy consumption between regions.
- Alternative Hypothesis (H_1): Significant difference in energy consumption across regions.

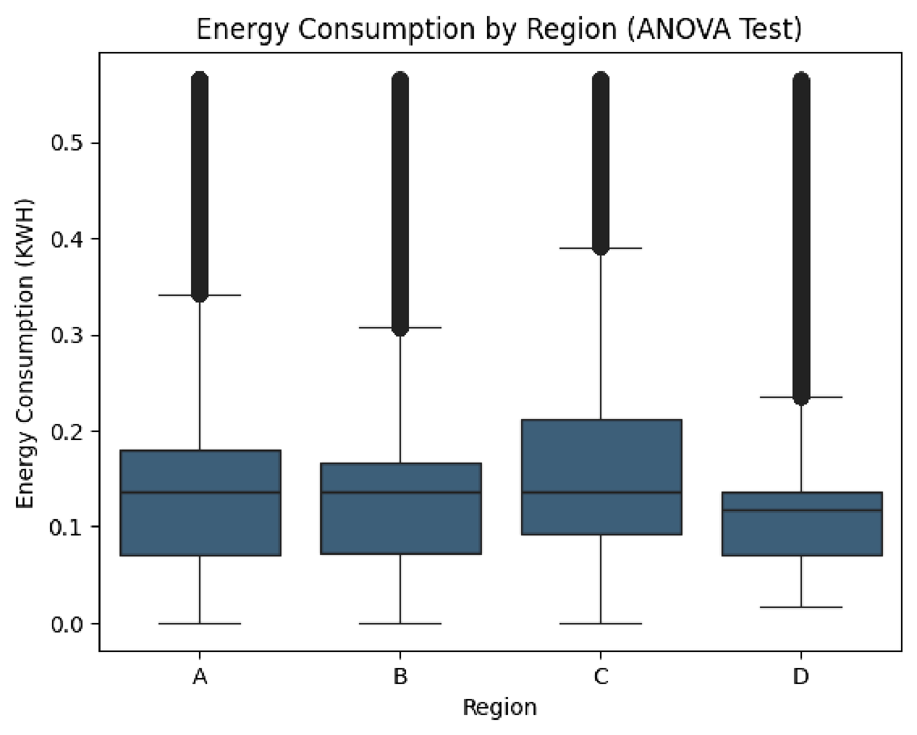
Key Results
- F-value: 3715.53 (Large difference between group means compared to within-group variability)
- P-value: 0.0 (Significant, reject H_0)
- Effect Size (η^2): 0.008 (Small effect size, suggesting the difference is noticeable but not large)

Assumptions Checked
- Normality: Shapiro-Wilk test revealed a violation of normality (p-value < 0.05).
- Homogeneity of Variances: Levene's test showed a violation of equal variances (p-value $= 0.0$).

Limitations
- Non-Normality: Results of ANOVA might not be reliable due to the violation of normality.
- Heteroscedasticity: The violation of homogeneity of variance could affect the robustness of ANOVA.

Conclusion
The results suggest a significant difference in energy consumption across regions, but the violations of normality and homogeneity of variance indicate that alternative methods like Welch's ANOVA could be more appropriate.

Energy Consumption by Region (ANOVA Test)



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

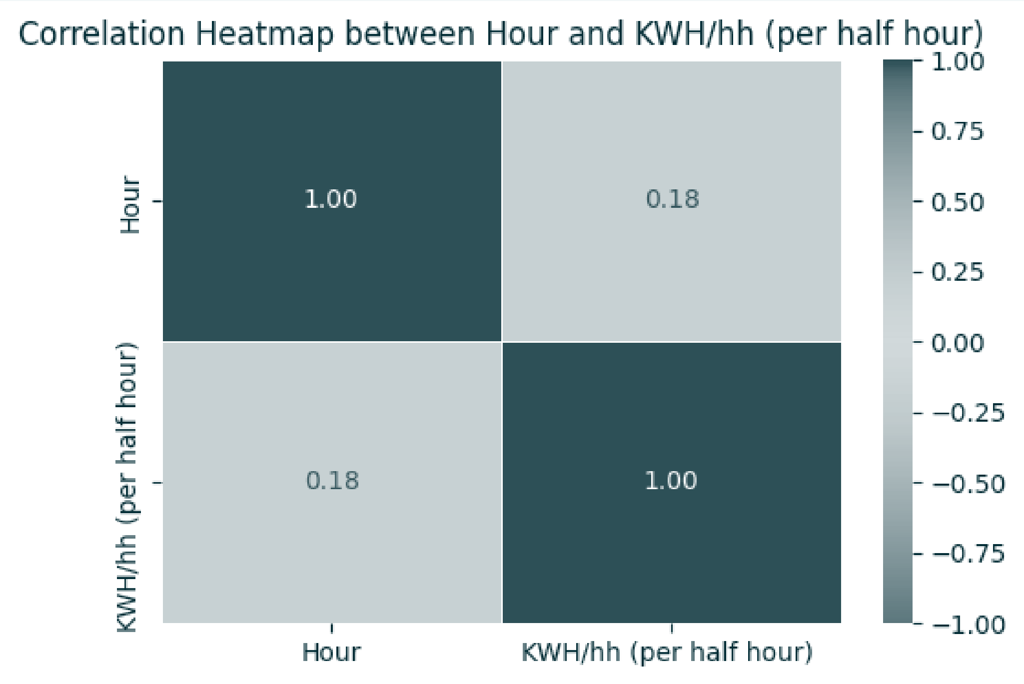
Objective
Using Spearman's Correlation, to Examine if there is a monotonic relationship between the hour of the day and energy consumption.
Key Results
- Spearman Correlation Coefficient (r): 0.175867 (Weak positive correlation)
- P-value: 0.0 (Statistically significant, unlikely to be due to random chance)

Assumptions Checked
- Monotonic Relationship: The weak correlation suggests a positive monotonic relationship, although not strong.
- Data Type: Spearman's rank correlation is appropriate for ordinal and continuous data.

Limitations
- Weak Correlation: The weak correlation suggests that other factors may influence energy consumption, not just the hour of the day.
- Outliers: Sensitivity to outliers, which can distort the correlation results.

Conclusion
The relationship between the hour of the day and energy consumption is statistically significant, though weak.

Correlation Heatmap between Hour and KWH/hh (per half hour)



05

Threshold Recommendation and Cost Savings

Threshold Determination

Goal
Determine a threshold to identify low-energy-consuming cells.

Method Used
- Calculated threshold using the 25th percentile of the energy consumption data (KWH/hh).
- This value represents the point below which the lowest 25% of energy consumption values lie, which corresponds to 344,247 rows × 10 columns of data.

Why the 25th Percentile?
- The 25th percentile highlights low-energy usage, meaning these cells consume minimal energy.
- These cells are candidates for shutdown or optimization to reduce overall energy consumption and costs.

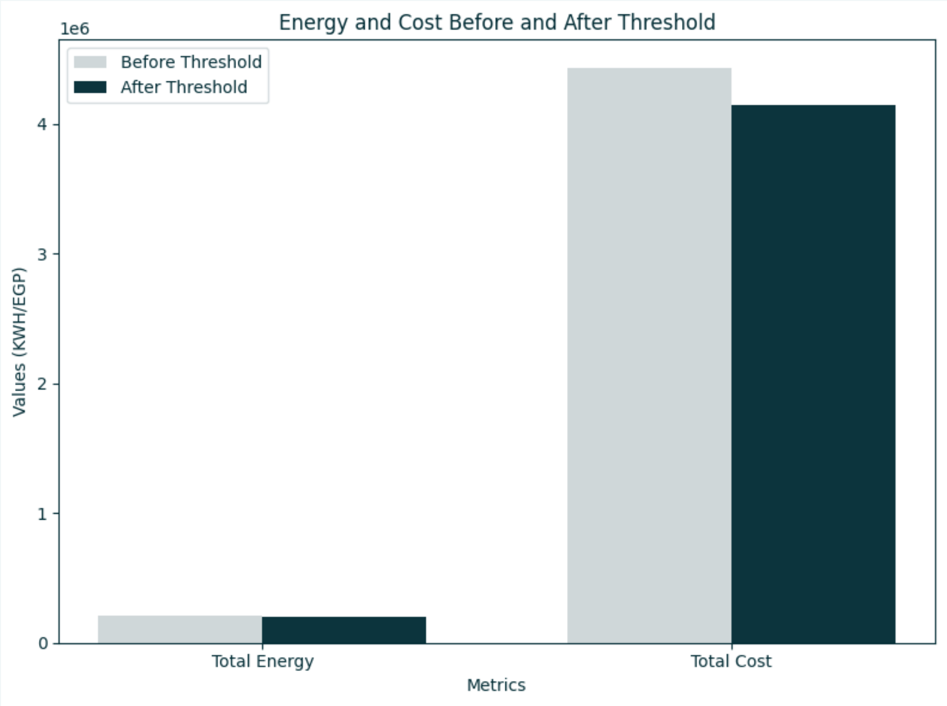
Identifying Low-Consumption Cells
- Cells with energy consumption below the threshold (0.078 KWH/hh).
- These cells are analyzed for potential energy and monetary savings.

Cost Savings Comparison

- **13,868,37 KWH**
  Energy Savings
- **283,216,00 EGP**
  Cost Savings

Before Threshold
Total Energy: 210,131.59 KWH.
Total Cost: 4,429,805.61 EGP.

After Threshold
Total Energy: 196,263.22 KWH.
Total Cost: 4,146,589.61 EGP.

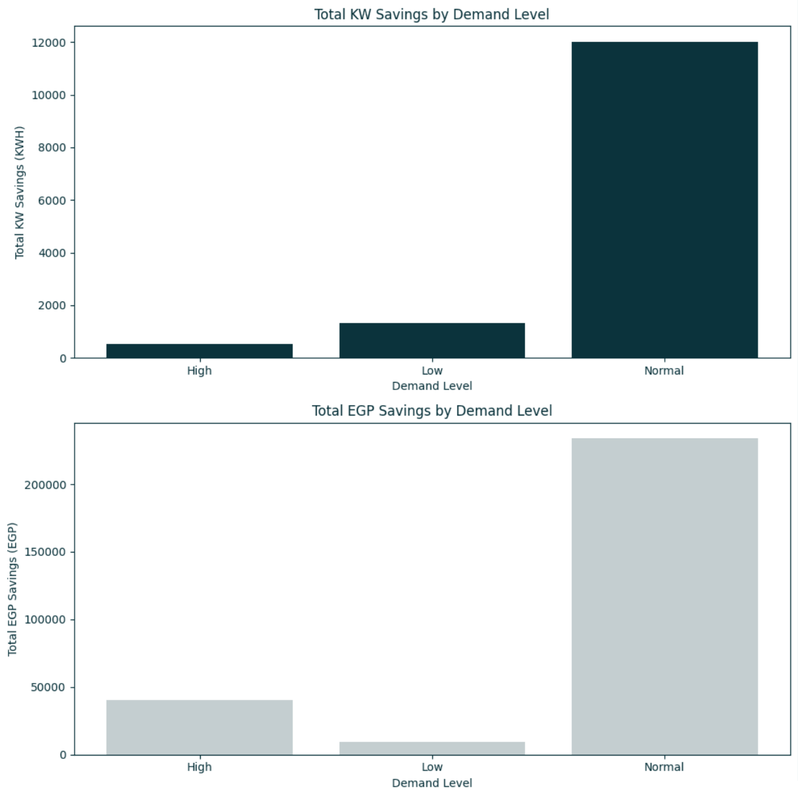


Savings Breakdown by Demand Level

High Demand
Energy Savings: 521.51 KWH.
Cost Savings: 40,265.94 EGP.

Low Demand
Energy Savings: 1,334.90 KWH.
Cost Savings: 9,197.49 EGP.

Normal Demand
Energy Savings: 12,011.95 KWH.
Cost Savings: 233,752.57 EGP.



Recap Key Insights

Throughout this analysis, we identified the 25th percentile as the threshold for low energy consumption, targeting opportunities for cost and energy savings. By focusing on cells below this threshold:

- **Energy savings:** 13,868.37 KWH
- **Cost savings:** 283,216.00 EGP

Additionally, the breakdown by demand levels provided actionable insights, with the Normal demand level contributing the highest savings, emphasizing its significance for optimization.

Highlight the Impact

This approach demonstrates how a data-driven strategy can lead to tangible benefits:

- **Operational Efficiency:** Reducing energy waste while maintaining performance in critical areas.
- **Sustainability Goals:** Aligning with broader efforts to minimize environmental impact.
- **Cost Optimization:** Significant monetary savings, highlighting the financial viability of implementing this strategy.

Thanks!