Analysis of Data Science Final Project

An in-depth exploration of methodologies, key components, and outcomes for effective project execution in data science practices.

presenter: Basmala Barakat Ahmed
supervisor : Eng.Mohammed Essam Saad

ENERGY PROJECT OVERVIEW

Data Science Final Project: Comprehensive Analysis

Overview of Key Topics

01

Introduction to the Project

An overview of the project's objectives, focusing on analyzing energy consumption patterns across various conditions, including holidays versus non-holidays.

02

Data Cleaning Process

A detailed explanation of the data cleaning methods applied to ensure the accuracy and reliability of the dataset used in the analysis.

03

Analysis Methods

An exploration of the analytical techniques employed, including statistical analysis and machine learning models to identify trends and insights.

04

Insights and Findings

Presentation of key insights derived from the analysis, highlighting significant patterns in energy usage across different regions and time periods.

05

Conclusion

A summary of the project's outcomes, implications for energy consumption policies, and suggestions for future research directions.

ENERGY ANALYSIS

Project Overview

A Comprehensive Overview of Objectives, Data Sources, and Description

Objectives

The project aims to critically analyze energy consumption data to uncover trends and anomalies, assess the influence of various factors such as time, holidays, and regional variances on energy usage, and utilize statistical methods to substantiate the findings.

Data Sources

We will utilize energy consumption data measured in KWH per household for every half hour, alongside demand data categorized into three levels: 'High', 'Normal', and 'Low', providing a comprehensive view of energy usage.

Data Description

The dataset comprises a total of 1,380,252 entries for energy consumption and 17,520 entries for demand data. Key columns include DateTime, KWH/hh (kilowatt-hours per household), site_id, and region, which are

DATA CLEANING

Data Cleaning and Preparation

Essential Steps for Effective Data Management

Loading Data

01 Utilized pandas to import data from various sources including CSV and Excel files, consolidating multiple dataframes into one for streamlined analysis.

Handling Missing Values

02 Conducted a thorough check for NaN values; confirmed that the energy dataset was complete while the demand dataset also had no missing entries.

Removing Duplicates

03 Identified and eliminated 942 duplicate entries from the energy data, a crucial step to ensure the accuracy and reliability of the dataset.

Data Type Conversion

04 Converted the 'DateTime' column into a proper datetime format, enhancing the dataset's suitability for time series analysis and ensuring accurate temporal data handling.

Merging Dataframes

05 Executed a merge operation on the energy and demand dataframes based on 'DateTime', allowing for a more integrated and comprehensive analysis of the datasets.

FEATURE ENGINEERING

Feature Engineering

A Comprehensive Overview of Extracted and Analyzed Features

Is_Holiday Feature

A binary feature 'Is_Holiday' was created to differentiate between holiday and non-holiday periods, aiding in the analysis of consumption patterns during special occasions.

Day Period Classification

The 'DayPeriod' feature was defined to categorize time into segments such as Morning, Afternoon, Evening, and Night, helping to reveal energy usage trends across different times of the day.

impact Day of Week

give a clear view of how energy consumption varies across different days of the week, helping you understand if certain days have higher or lower consumption.

Data after cleaning,merged and adding new column

to help us in visualization

	cell_id	DateTime	KWH/hh (per half hour)	site_id	region	Demand	Date	Time	Year	Month	Day	Day_of_Week	Is_Holiday	Hour	DayPeriod
0	MAC000002	2013-01-01 00:00:00	0.219	A	A	Normal	2013-01-01	00:00:00	2013	1	1	Tuesday	Not Holiday	0	Night
1	MAC000002	2013-01-01 00:30:00	0.241	A	A	Normal	2013-01-01	00:30:00	2013	1	1	Tuesday	Not Holiday	0	Night
2	MAC000002	2013-01-01 01:00:00	0.191	A	A	Normal	2013-01-01	01:00:00	2013	1	1	Tuesday	Not Holiday	1	Night
3	MAC000002	2013-01-01 01:30:00	0.235	A	A	Normal	2013-01-01	01:30:00	2013	1	1	Tuesday	Not Holiday	1	Night
4	MAC000002	2013-01-01 02:00:00	0.182	A	A	Normal	2013-01-01	02:00:00	2013	1	1	Tuesday	Not Holiday	2	Night

CLICK HERE TO EDIT LABEL

Data Insights

Click here to edit subtitle

Outlier Detection

Utilized Boxplots to visualize outliers in energy consumption data, effectively highlighting data points that deviate significantly from the norm and MAD to Handle outlier because of the data non normal distribution .

Assumption 1

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

Assumption 2

Identify regions with higher or lower energy consumption, and determine whether energy usage is higher on specific days, hours, or during holidays

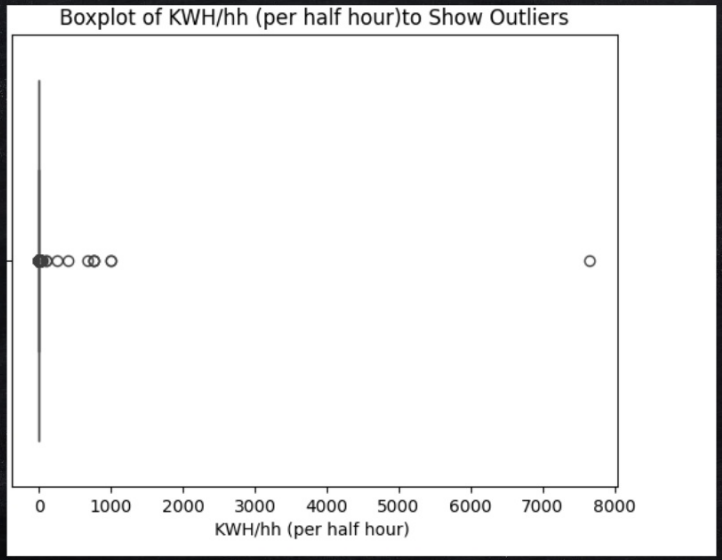
Energy Saving Threshold for Low Consumption Cells

To recommend a threshold for shutting down cells with low consumption

check outlier using Boxblot

That refer we have alot of outlier

and I used MAD to Handle the outlier beacause of the Data is non normal Distribution



ENERGY INSIGHTS

Analysis Exploration

Insights on Regional and Temporal Variations

Regional Analysis

Analyzed energy consumption patterns by region and day of the week. The visualization of average KWH/hh consumption revealed significant variances across different regions, indicating diverse

Holiday vs. Non-Holiday Comparison

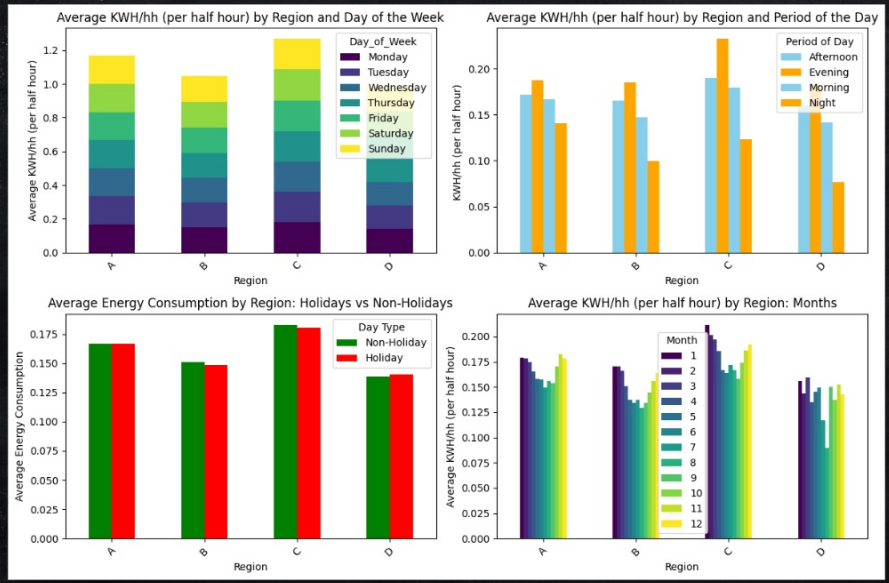
Conducted a detailed comparative analysis of energy consumption during holidays versus non-holidays. The results showed minimal differences in average consumption levels, suggesting consistent

Period of Day Analysis

Investigated energy consumption patterns throughout the day, noting that consumption peaks during the evening hours. This pattern prompted further investigation into user behaviors, with bar chart

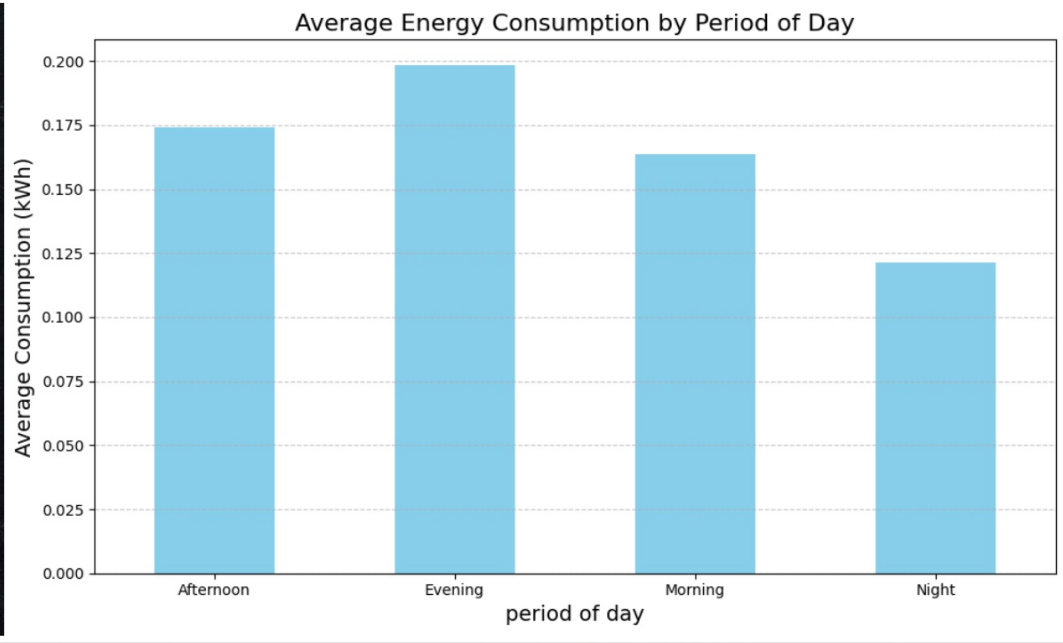
Region(Day,period,month,holiday) with Energy

appear that it’s adifference between region for all categories we use



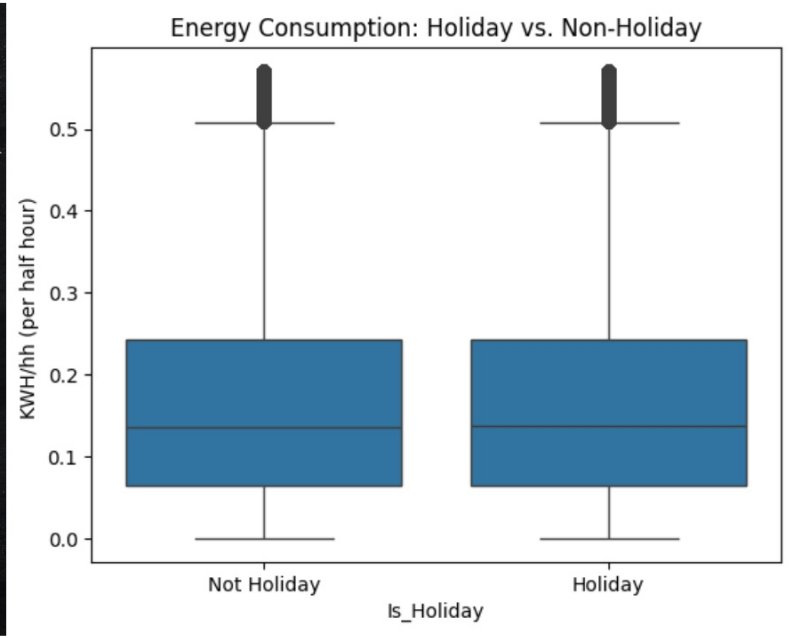
Energy Consumption with period of Day

that appear for me that Energy consumption is difference on period of Day



AVG using of Energy Consumption on Holiday

this image appear that avrage use of Energy Consumption& the difference between them is simple this image



ENERGY CONSUMPTION ANALYSIS

Hypothesis Testing

with Numeric Data and Categorical

Testing the Impact of Holidays

using ttest for Numeric &z-test for proportion

Testing Regional Variance

A Chi-Square test & An ANOVA assessed energy consumption across different regions.

Impact of Time of Day

A Chi-Square test & An ANOVA test was executed to examine the influence of time of day on energy consumption.

Threshold Recommendation and Cost Savings

Threshold Determination

01 **KNN to Divided Data**

The KNN model effectively segments the dataset, providing confidence in the "Low" and "High" energy categories

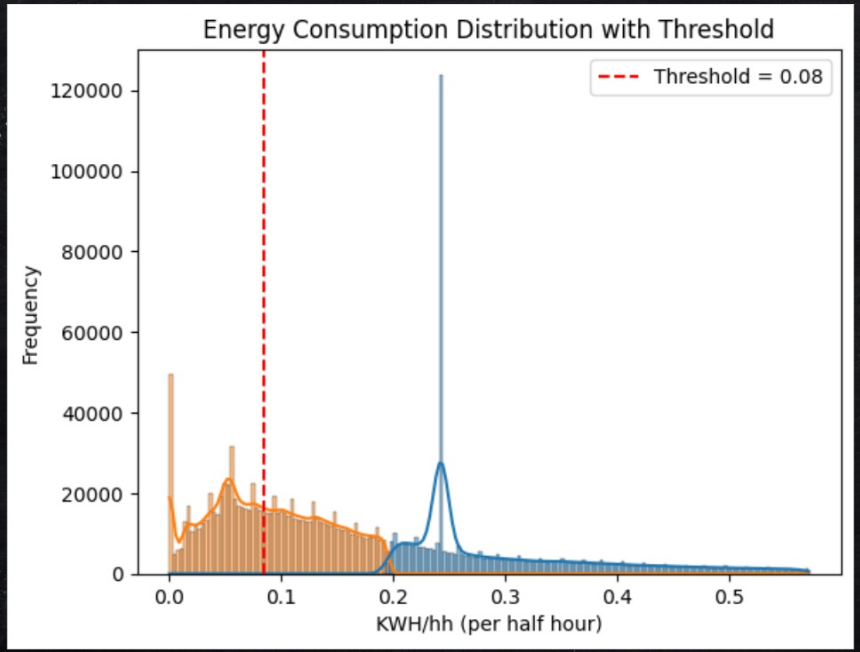
02 **Data-Driven Threshold**

The threshold derived from the median of the "Low" cluster ensures that it is robust and representative of the low-energy cells

03 **Why using this Data Driven threshold**

To Balance Between Savings and Operations Using the median avoids being overly aggressive (like the minimum) or too lenient (like the average).

threshold which use to shutdown cell



THRESHOLD RECOMMENDATION AND COST SAVINGS

Cost Analysis and Savings

Analysis and Recommendations

01

Cost Analysis

calculate the price for each cell based on its energy demand category. By mapping the demand values to

02

Total Cost Calculation

By multiplying the price per unit by the energy consumption, you get the total cost for each cell in pens. This is necessary because the original price is

03

Cost Savings Calculation

If the energy consumption (KWH/hh (per half hour)) is less than the low energy threshold (low_cluster_median), the cell's energy consumption is set to 0 (indicating it's shut down). Otherwise, it keeps the original energy consumption value.

04

Conclusion

total cost of Data before save opption: 29991.72EGP
total cost of Data after save opption: 27548.546EGP
cost saving: 2443.17EGP

thank you!