Sites Energy Consumption

Presented by: Roaa Ehab Ahmed
Date: 15-12-2024
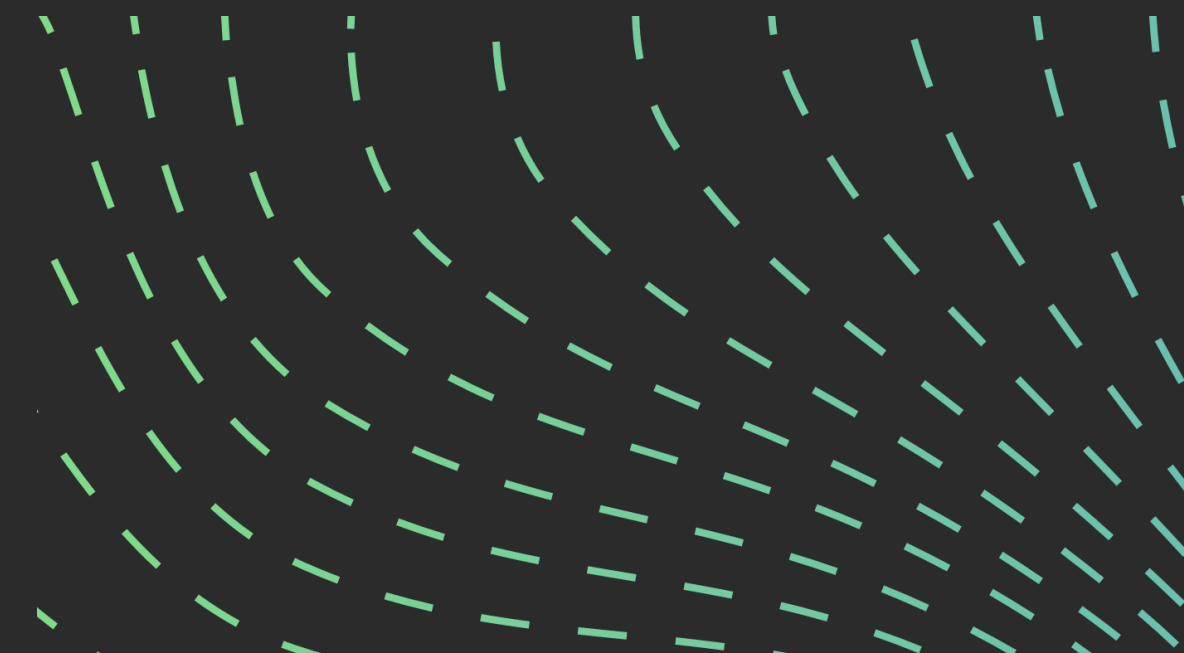
Table of content

Introduction

- The objective of this project was to save energy by Monitoring the energy consumption of our daily usage to help us save energy

- We'll explore the reasons why tracking our energy consumption is essential, the benefits it brings to our households and communities, and practical strategies for monitoring and reducing our energy usage.

>>>

Mission

our mission is to save energy by analyzing the data we can answer the following Questions

- Does the region significantly affect energy consumption?

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

- How to save energy?

Methodology

- **IMPORT NEEDED LIBRARIES**
  - pandas
  - seaborne
  - Matplotlib

- **DATA CLEANING**
  - check datatype of each column
  - check Null values
  - Drop duplicates
  - make sure that there is no missing records
  - Anomaly detection

- **EXPLORATORY ANALYSIS**
  - plot different figures

- **STATISTICAL HYPOTHESIS TESTING**
  - Hypothesis Formulation
  - Test Selection and Application
  - Threshold Recommendation and Cost Savings

Data Exploration and Analysis

Available Data

Data contains four csv files:

 1- three Energy Consumption.csv

contain power consumption of 82 cells from diffrent regions. the reading of the cells was each half our for an entire year (2013)

2- Power Demand.csv

contain the demand of each half our for an entire year (2013)

Tip: Collaboration makes teamwork easier! Click "Share" and invite your teammates to fill this up. Use this whiteboard page for bulletins, brainstorms, and other fun team ideas!

Data Reading

Demand:

 this data contain three sheets. only the first sheet contain data the remaining two are empty

 Conclusion after reading the data:

1- this table contain 17520 recode

2-No Null Values

3- Date Types are correct

Data Cleaning

Demand

- The demand column contains values expressed in various formats that represent the same underlying amount.
- Check and drop duplicates (there is no duplicates in this dataframe)

Data Reading

Energy Consumption:

 this data is divided on three csv file. so we must concat them all first

 Conclusion after reading the data:

1- file is missing some record as the expected no.'1436640' Not '1380252'

2- DateTime column must be Datetime type

3- there are four regions

4- there are 82 cell across all regions combined

Data Cleaning

Energy Consumption

- change DateTime column to datetime64[ns] with the correct formate
- check and drop duplicates
- find the cells which has less than 17520 reading with further analysis there is 70 cells contain missing values
- Find missing date and time by defining **add_miss** function that:

    1- find missing date time value   2-add new record with this value
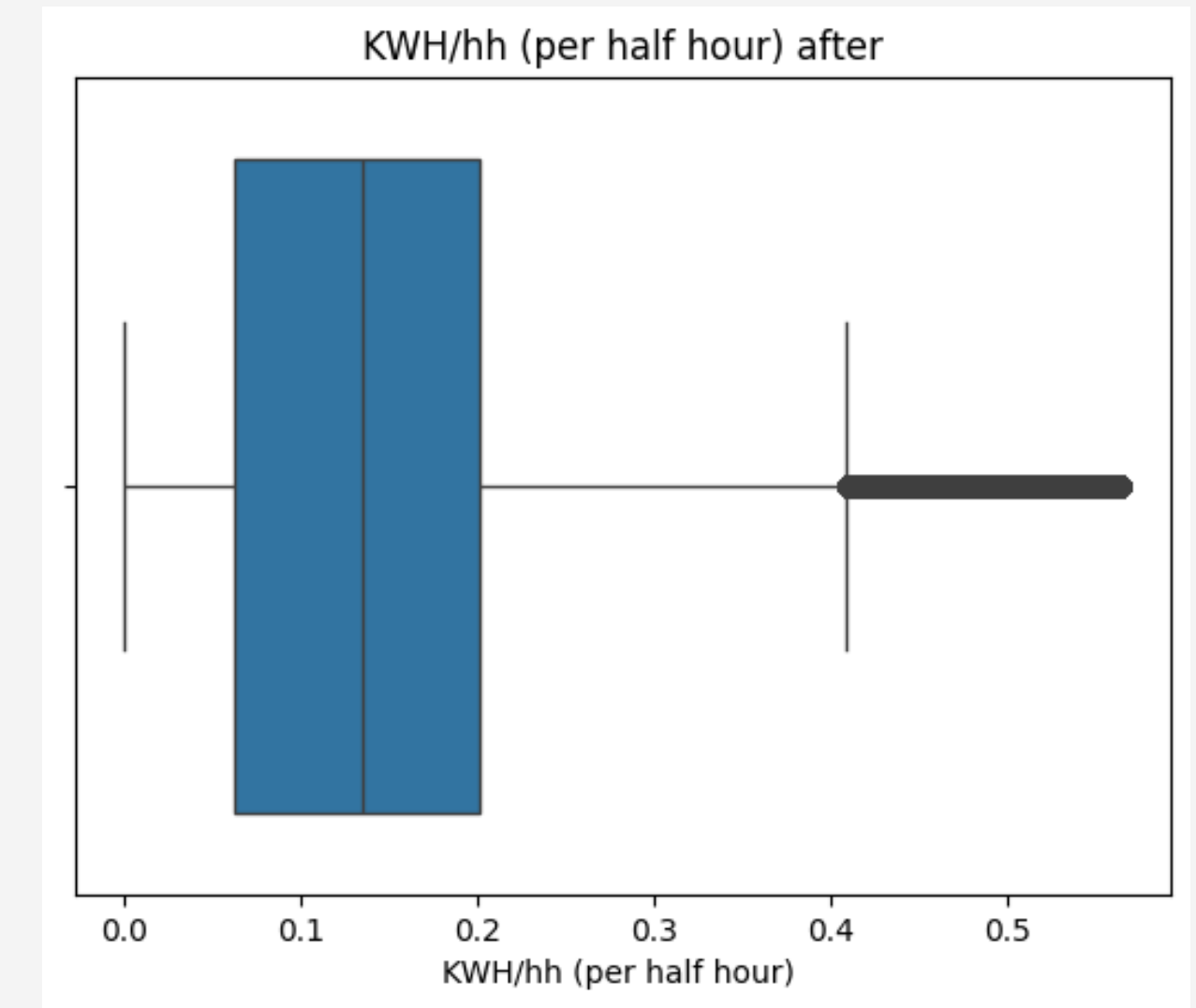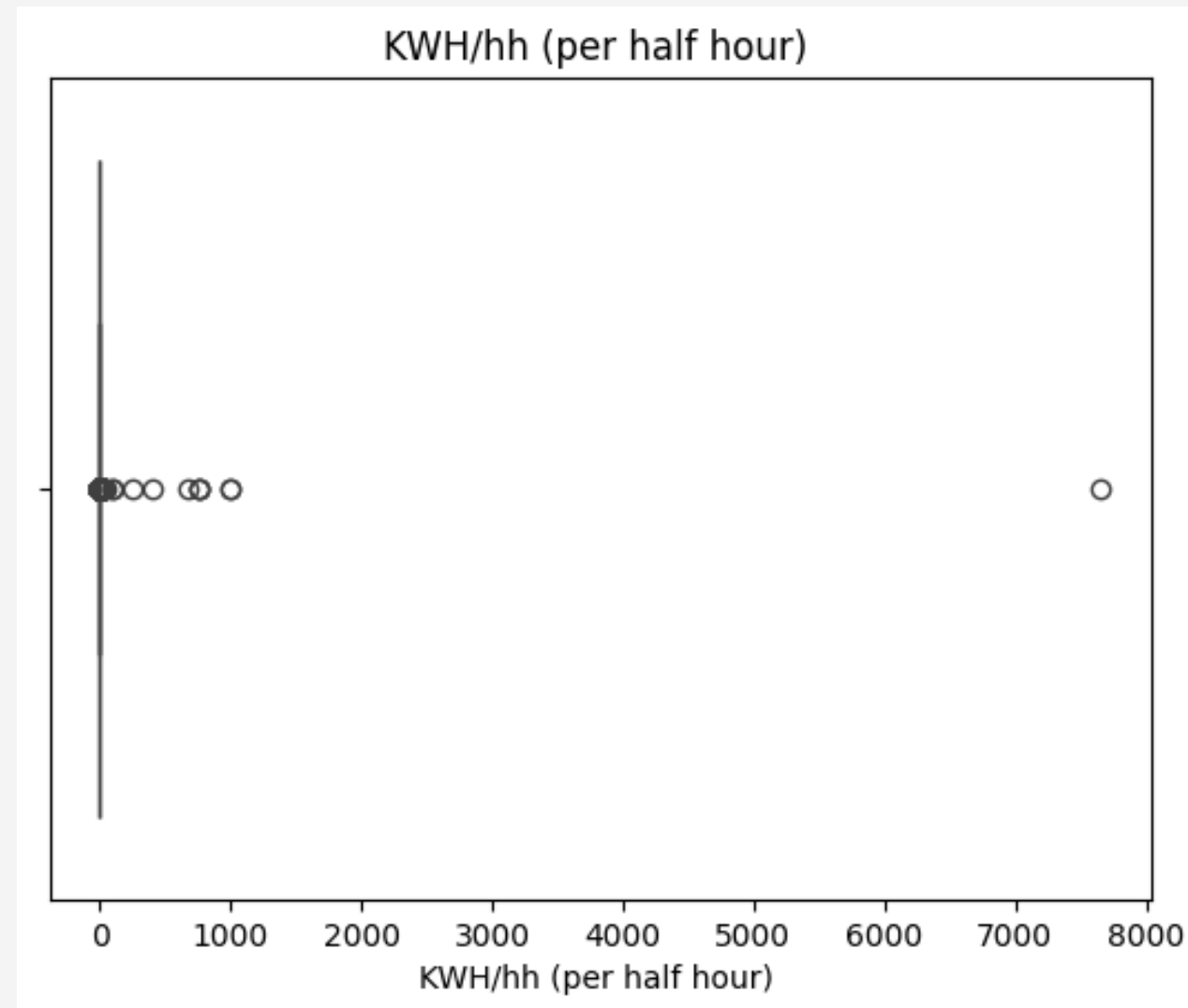
                    3-impute KWhh with median of this cell

Anomaly Detection

- Demand_df:
  after trying to univariate and Mulitvariate no useful anomalies
  insight was detected

- consumption:
  univariate: applying IQR and MAD
  Multivariate: by applying KNN , IForest and LOF (Local outlier factor)

»»
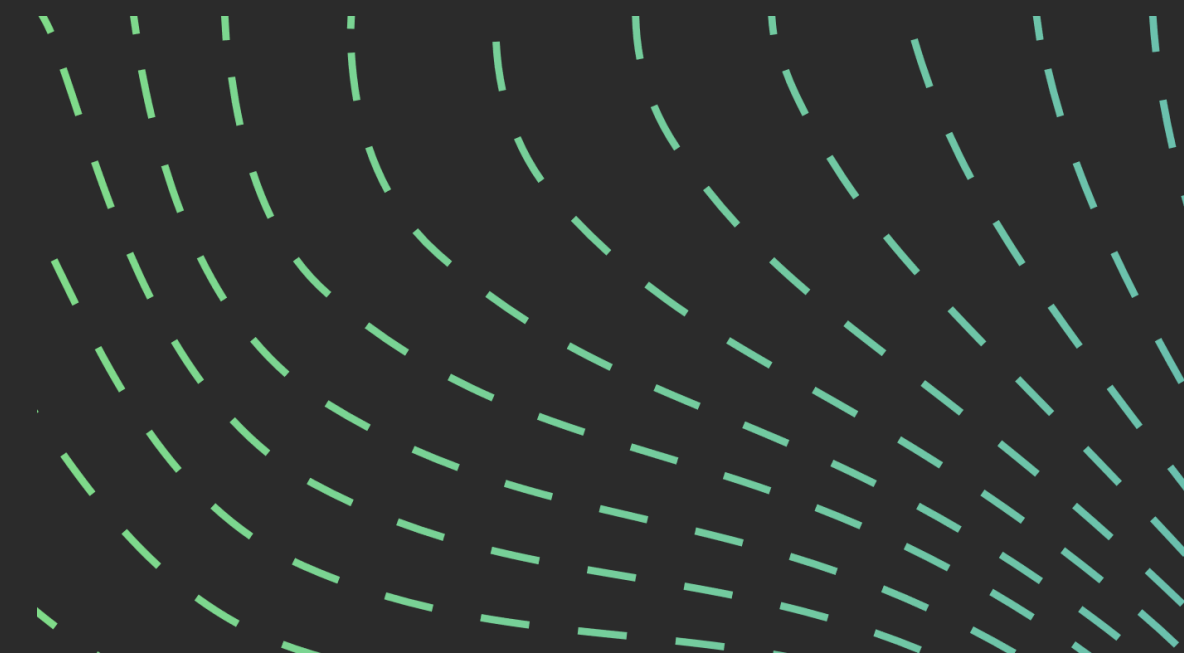
AD of consumption df

univariate

KWH/hh (per half hour)

KWH/hh (per half hour) after

the opposite box plot shows the anomlies in KWH/hh (per half hour) column

after applying both IQR and MAD and replace high outliers with median

Multivariate

Energy Consumption

- applied tests :

  1- KNN   2-IForest   3- LOF

- encoding date time
- Less than 3% of the total data displayed detected anomalies.
- They can simply remain as they are.

Exploratory Analysis

Analysis

Energy Consumption

- there are four main region
- the following pie chart show each region and the consumption
- region A has the most consumption
- Region D has the least consumption
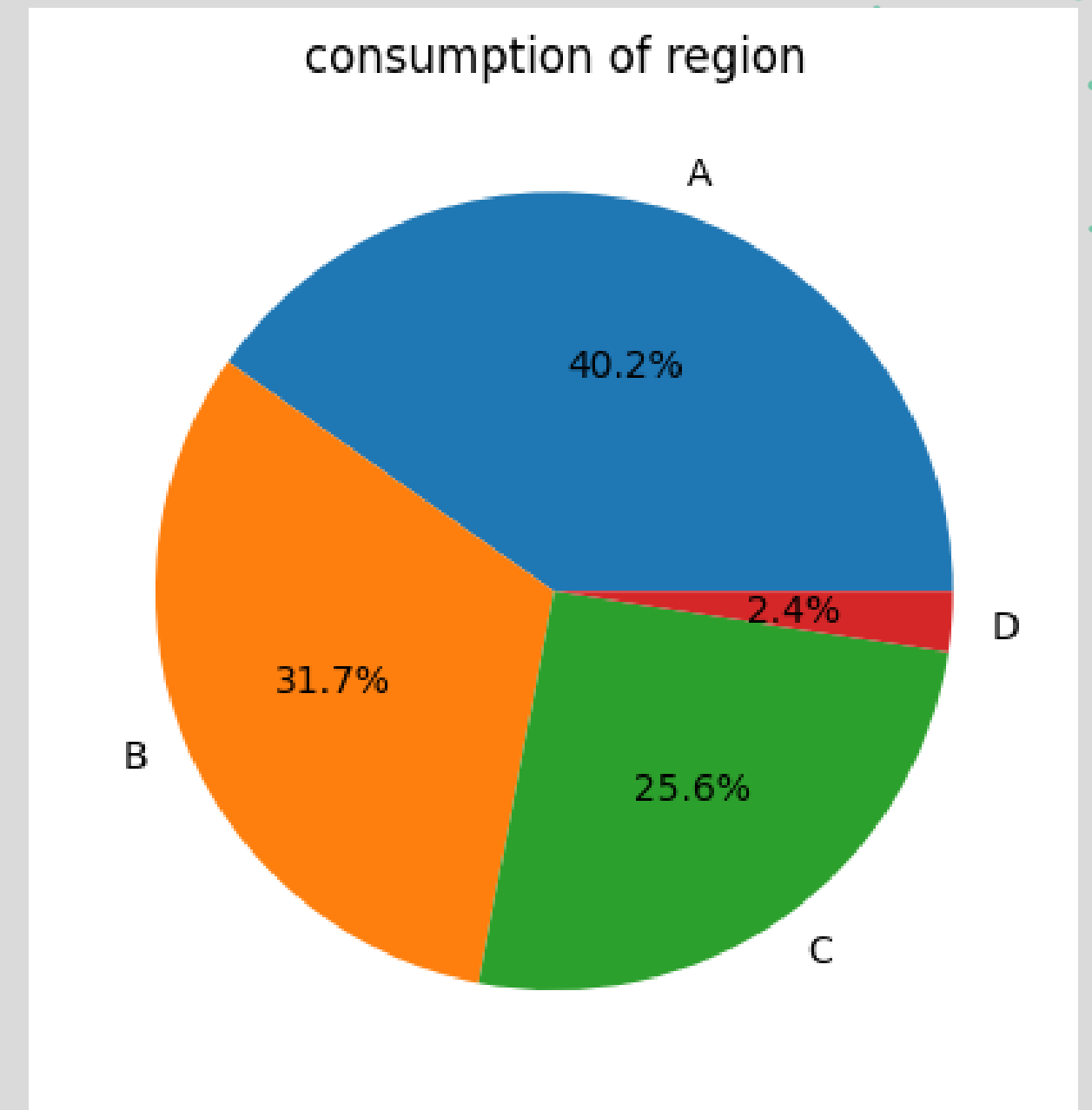
consumption of region

Analysis

Energy Consumption

- with further analysis
- Region A has 33 cell
- Region B has 26 cells
- Region C has 21 cells
- Region D has 2 cells
- this explain why Region A have the most consumption

Cells in each region
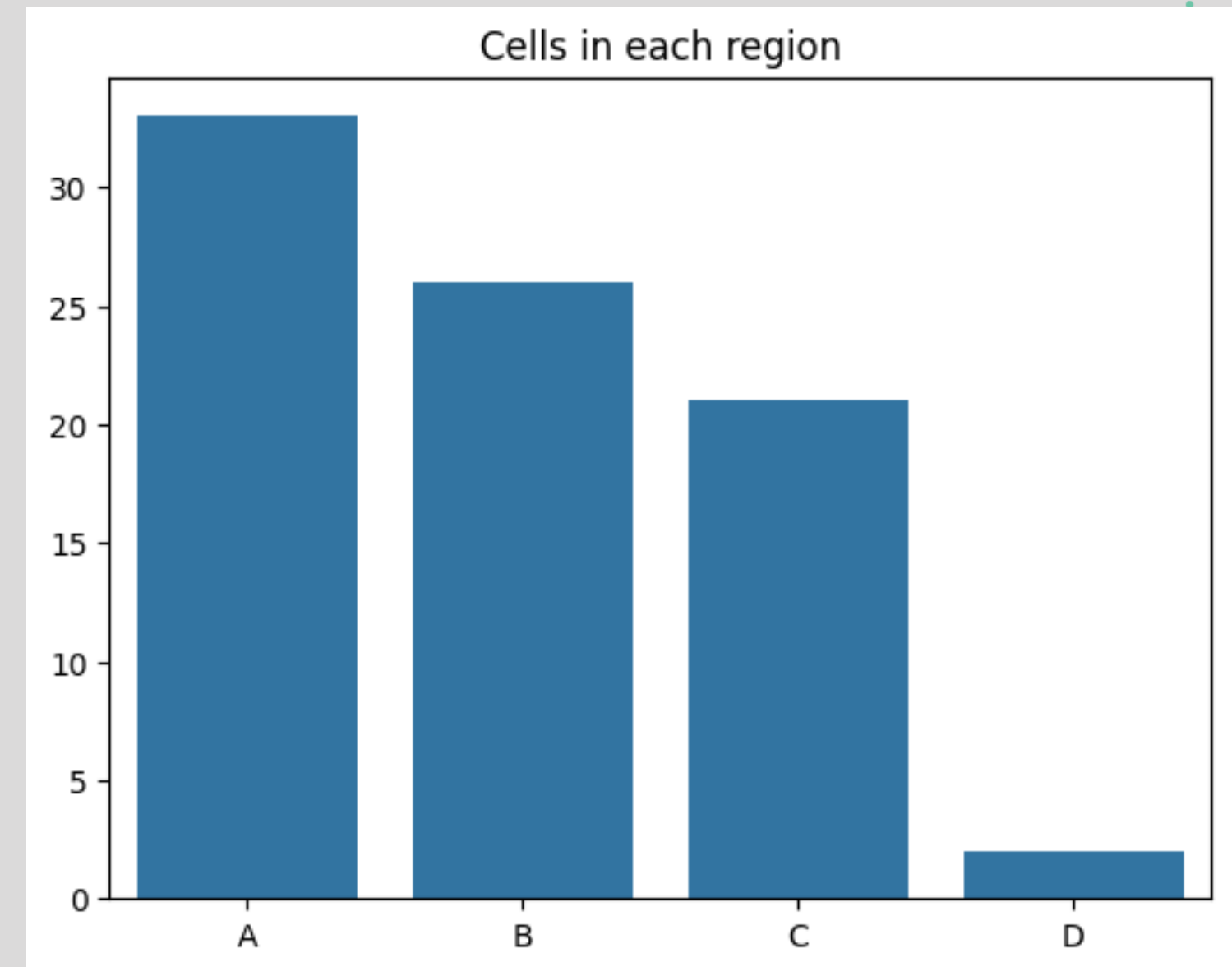
Analysis

Energy Consumption

- Site E has the most reading in the whole this means that it has the most no of cells
- There are 4 site in region A ['A' 'E' 'D' 'C']
- There are 5 site in region B ['P' 'Q' 'L' 'N' 'K']
- There are 4 site in region C ['H' 'J' 'F' 'G']
- There are 1 site in region D ['U']
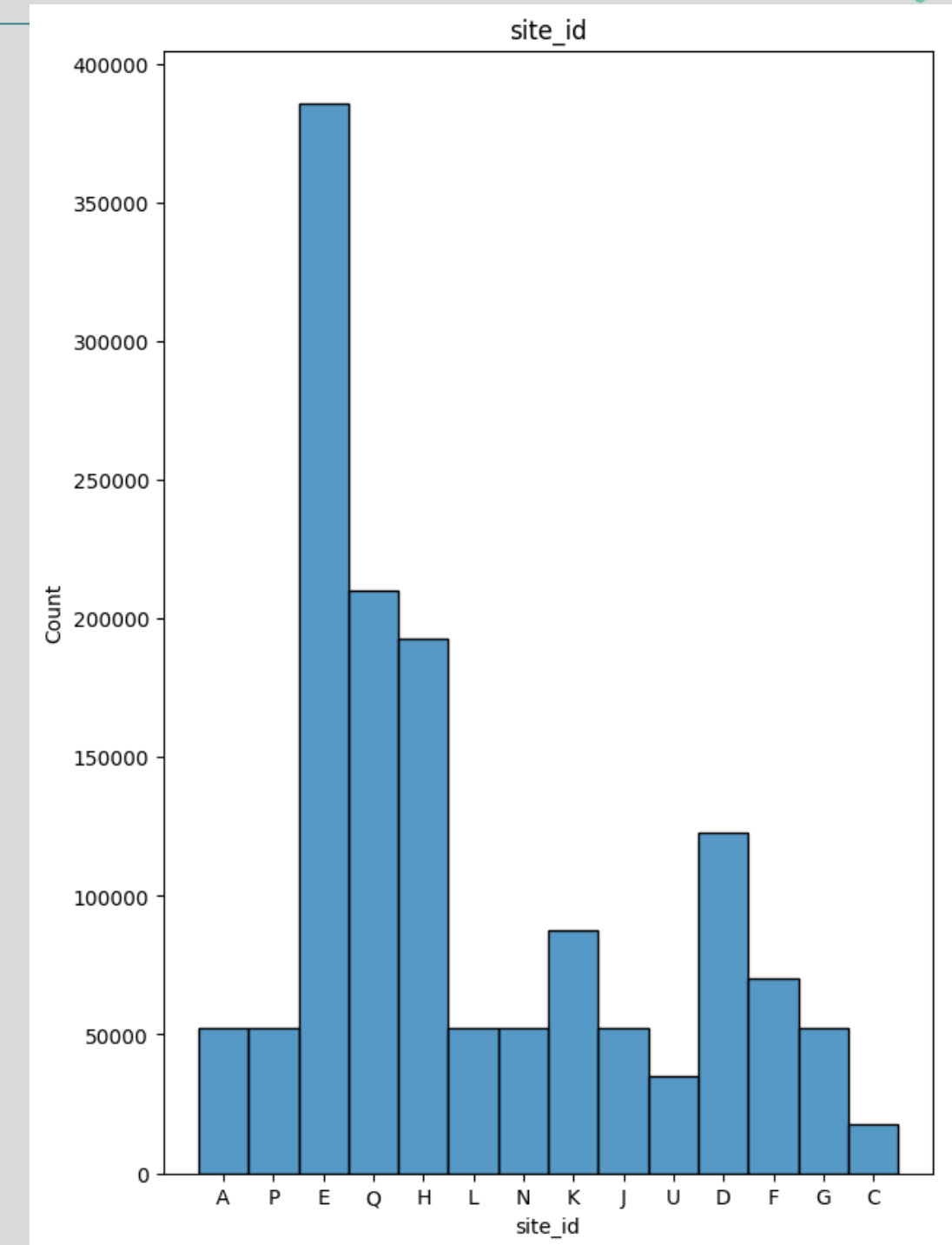
Analysis

Energy Consumption

- while there is great difference in the number of cell in each Region the median of the consumption is slightly different
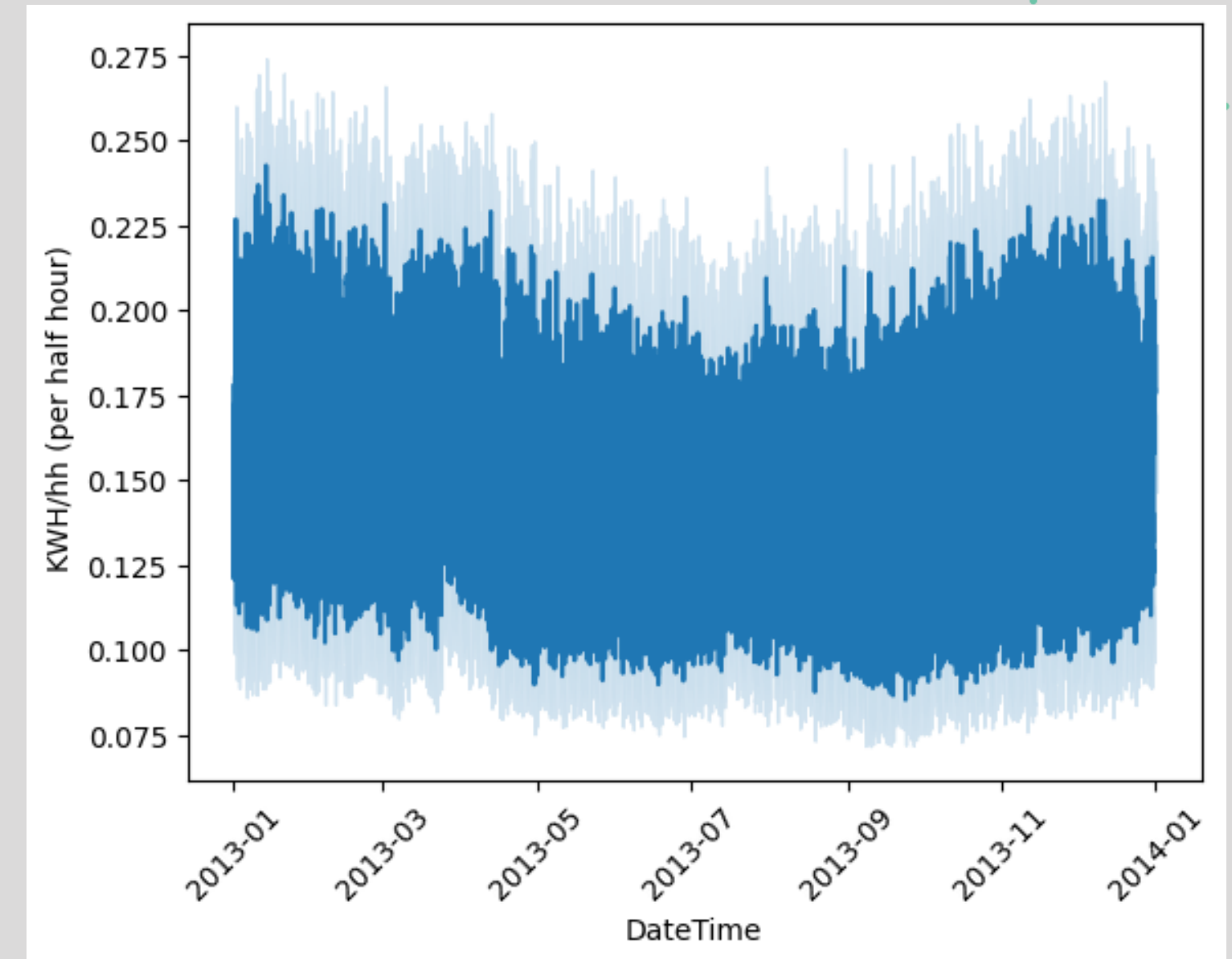

median consumption of each region
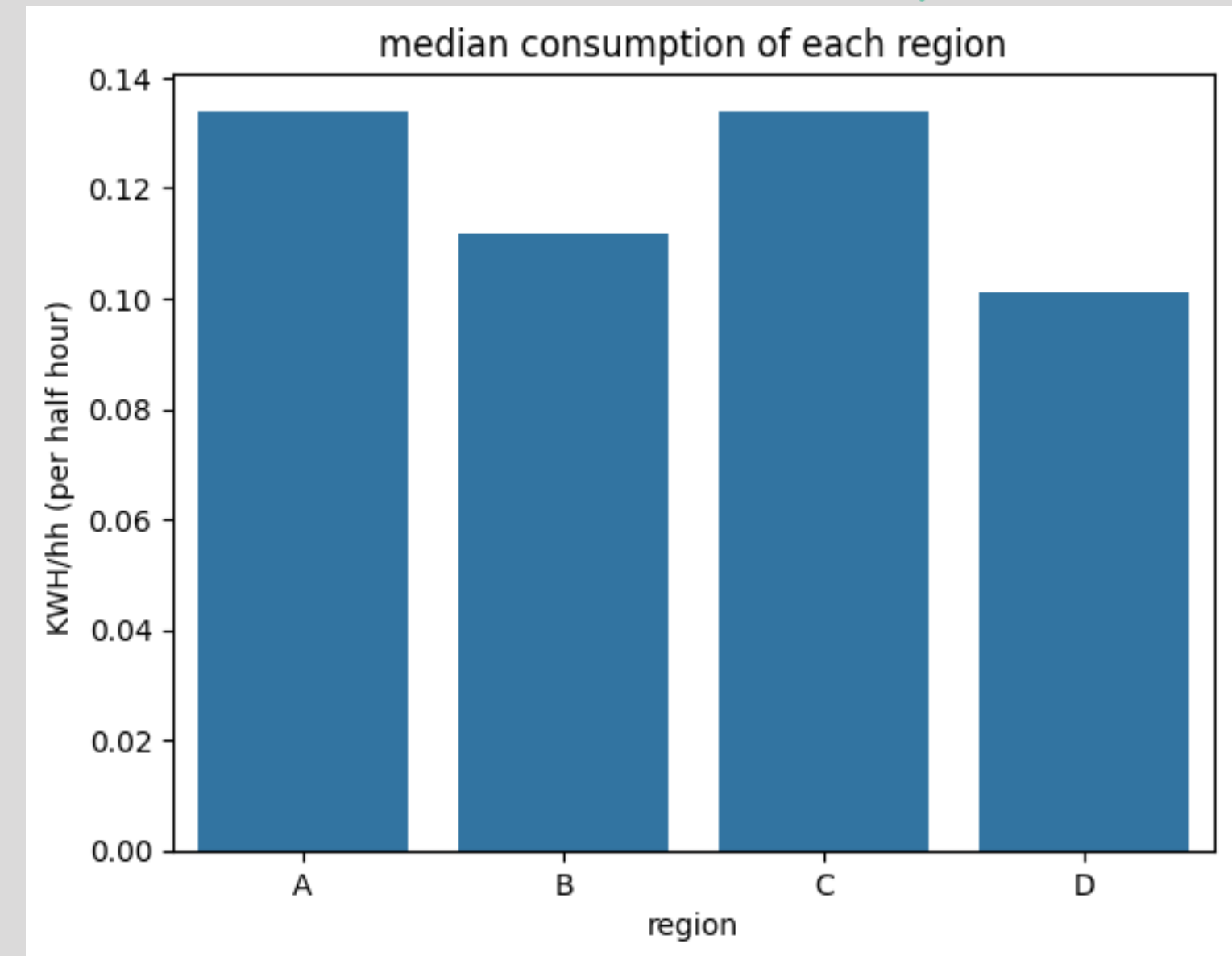
Analysis

Energy Consumption

- this figure explains the consumption across the whole year
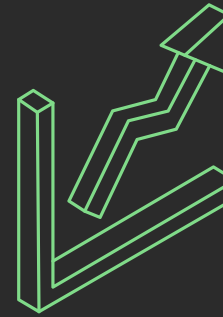- May, June, July and August shows a slight decrease in the consumption

Feature engineering

Add new columns

- cost according to the demand:
  Low= 3.445 p/kWh

  Normal= 9.73 p/kWh

  High= 38.73 p/kWh
- month column in names
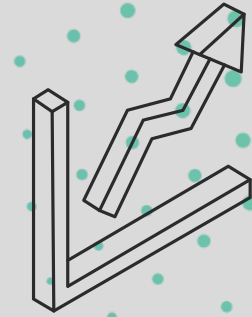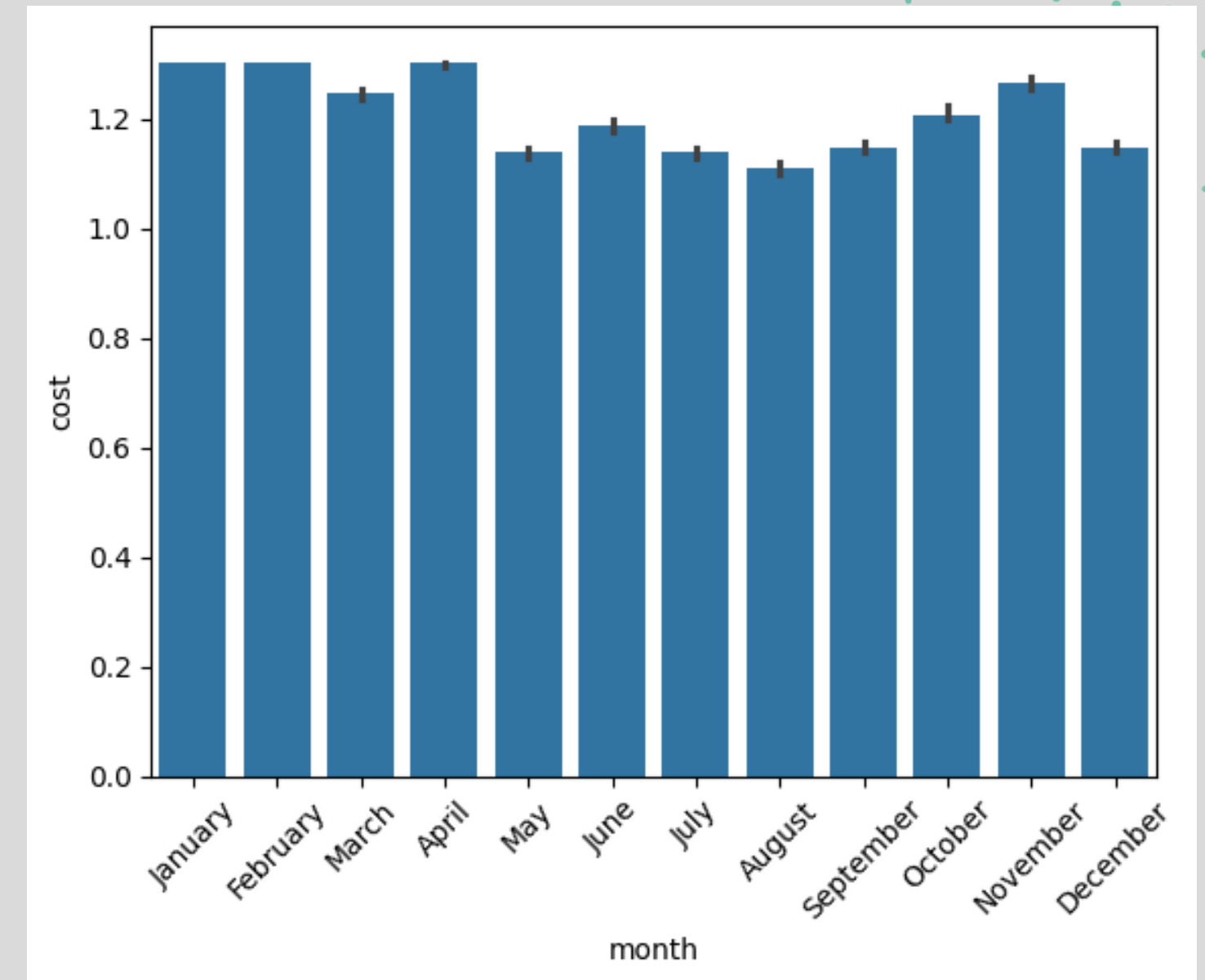- time of the day column night, morning, afternoon, evening

Data Insights
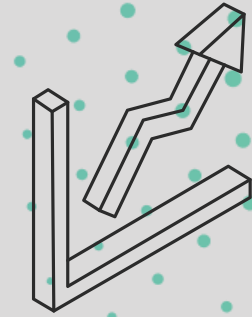
- figure shows the consumption of each month
- conclusion :May, June, July , August and September shows a slight decrease in the consumption
- the cause of this decrease in consumption is might be because of the summer weather
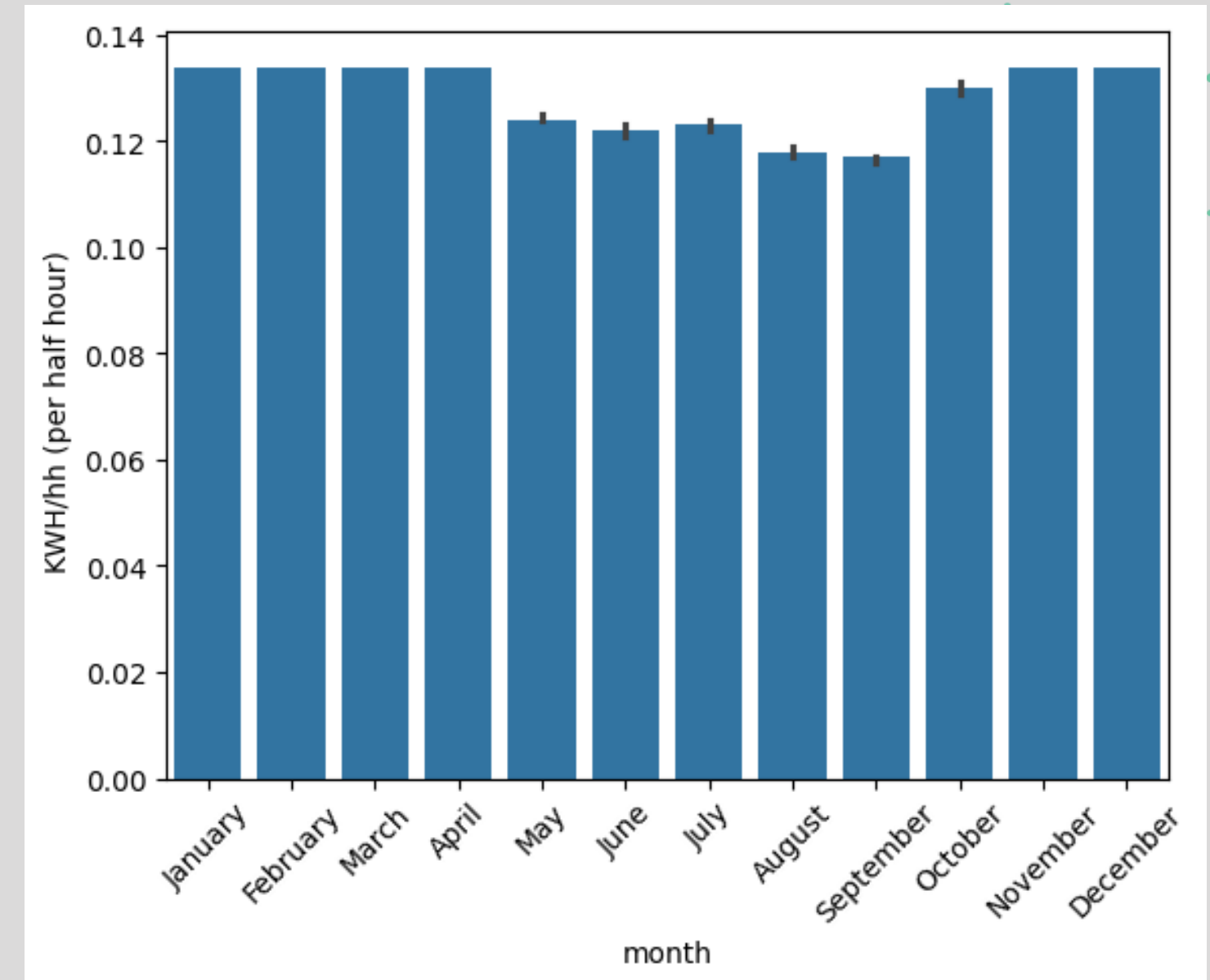
Data Insights

- figure shows the cost of each month

- conclusion :May, June, July and August shows a slight decrease in the cost

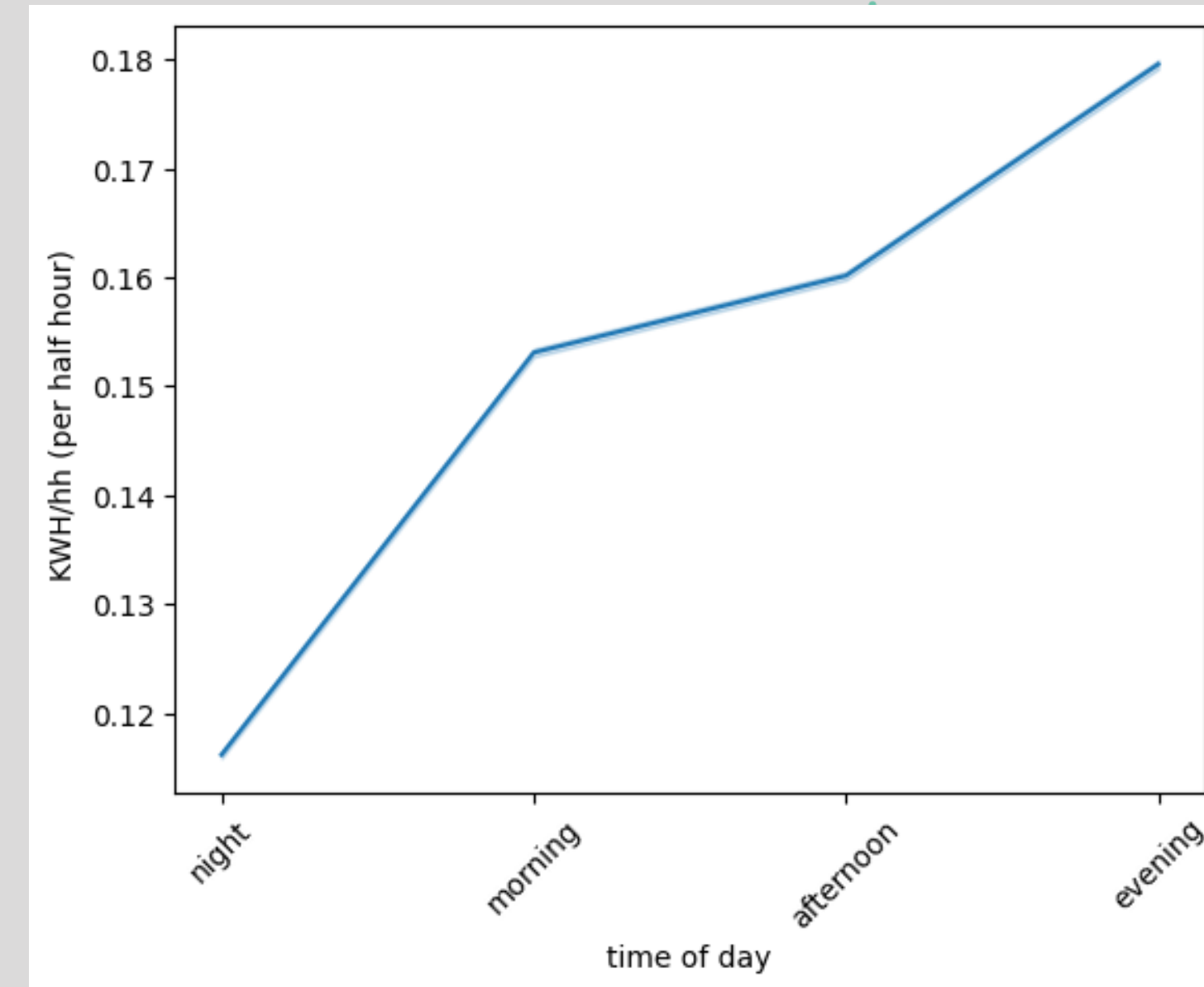- this due to the low consumption at these months as shown in the previous slide

Data Insights

- figure figure show the consumption with time of the day
- the consumption at night is less any other time
- the consumption at morning and afternoon is slightly different
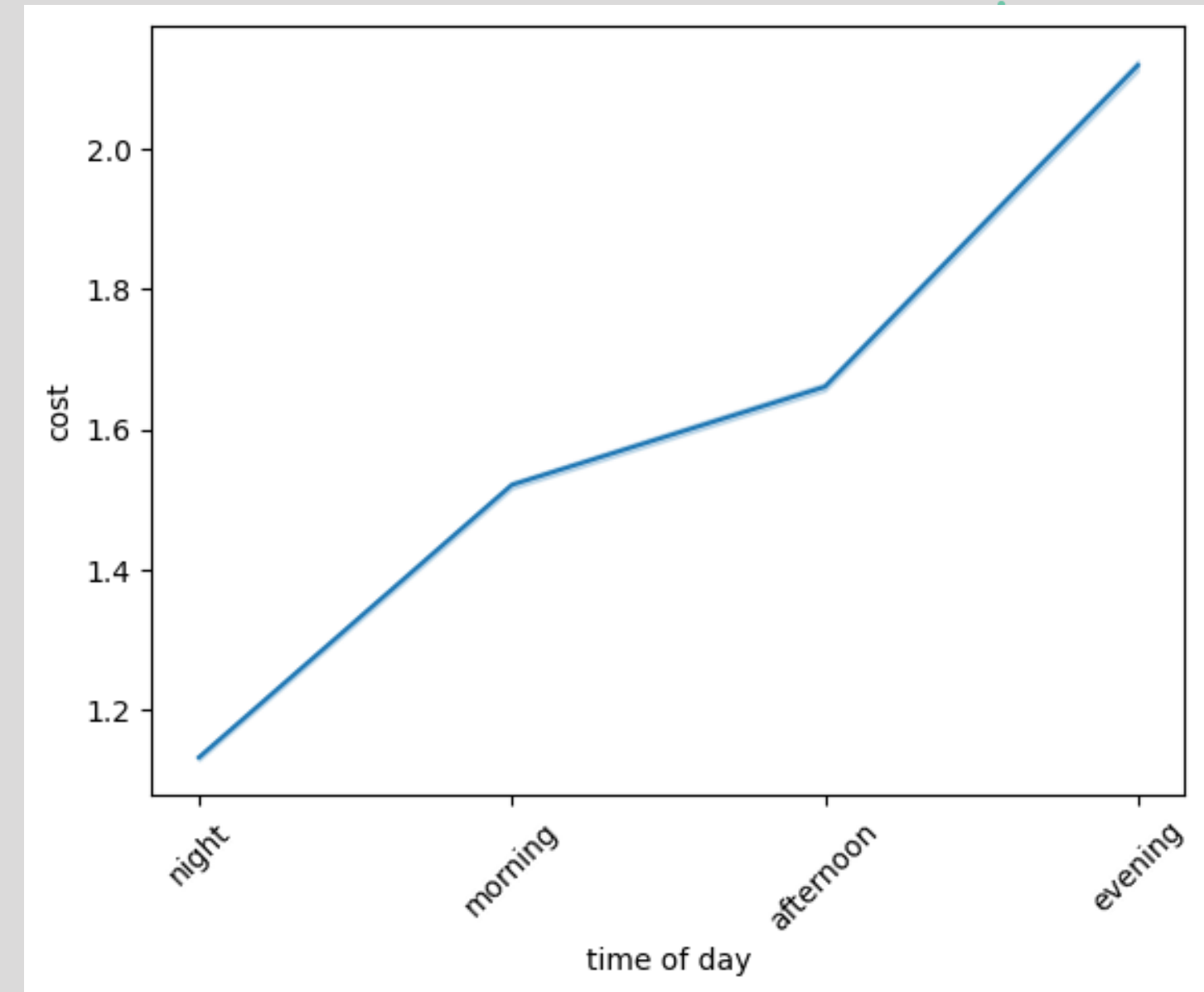- while the most consumption is at evening(6 to 12)

Data Insights

- figure figure show the cost with time of the day

- this graph is fully dependent on the conclusion in the previous slide
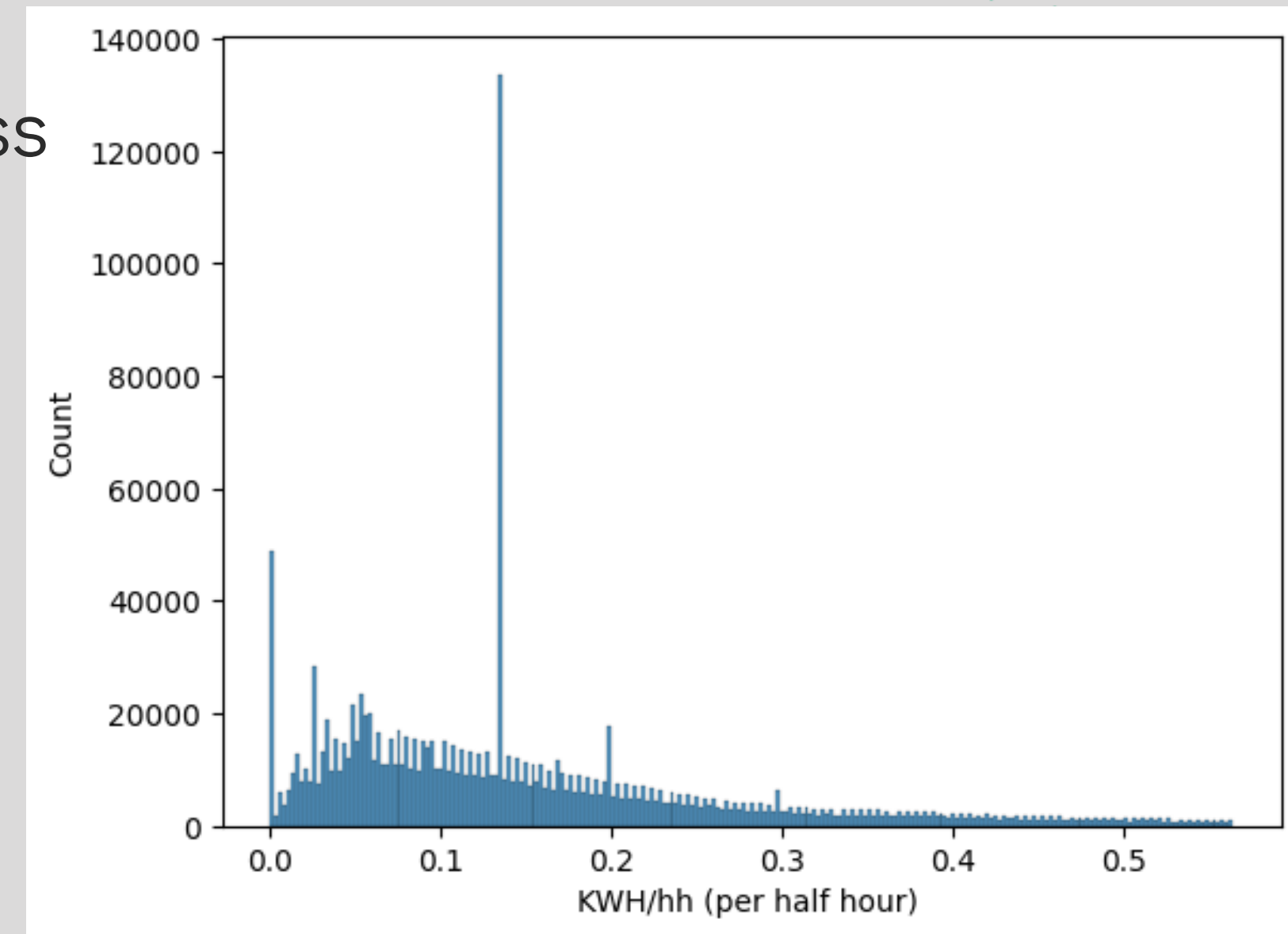
Statistical Hypothesis Testing

Hypothesis

	Does the region significantly affect energy consumption?	Is there a relationship between the time of day and energy consumption?
• Hypothesis Formulation	• H0:there is a diffrence in cost according to region • HA: there is no difference	• H0:there is a relationship between consumption and time of the day • HA: there is no relation
• Test Selection and Application	• ANOVA • PaireWise	• ANOVA • PaireWise
• Results	• p_value=0(Not logic) • reject the H0	• p_value=0(Not logic) • reject the H0

Assumptions and Limitations

- the consumption column show right skewness
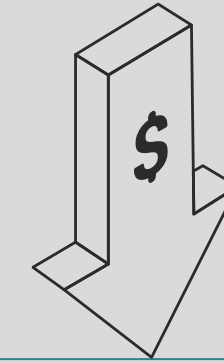- this may cause to the wrong P-value
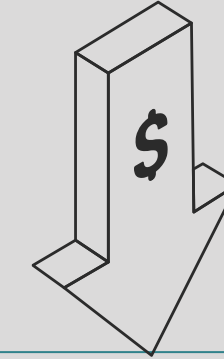
Threshold Recommendation and Cost Savings

Threshold Recommendation

- shut all cells that are below q1
- this method estimates to save about 25% of the cost
- q1=0.062
- total savings = 117953.5

Threshold Recommendation

- shut all cells that are below the mean with 1 standard deviation
- this method can save up to 13.6 % of the cost
- Std = 0.03
- Total save cost = 22927.2

Thank You