MEASURE ENERGY CONSUMPTION



Table of contents

Introduction:

Data cleaning

Data clensing

EDA

Analysis of data:

Data visualization.

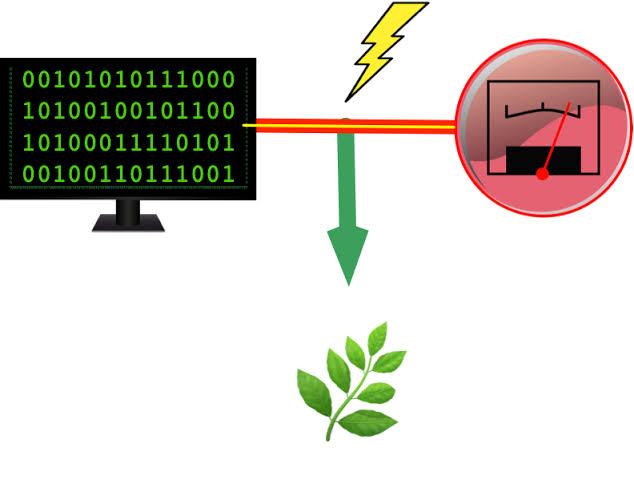
Data modelling.

Histogram representation.

Conclusion.

Introduction:

The efficient utilization of energy resources has become a critical concern in today's world, where environmental sustainability and economic viability are at the forefront of global priorities. Energy consumption, across various sectors such as residential, commercial, and industrial, plays a pivotal role in shaping our future. As we face mounting challenges related to climate change and the depletion of finite energy sources, it is imperative to measure, understand, and ultimately optimize energy consumption.



This study seeks to delve into the multifaceted landscape of energy consumption assessment, providing a foundation for informed decision-making and sustainable practices. By examining the methodologies, technologies, and strategies employed in measuring and managing energy usage, we aim to shed light on the significance of this endeavor. In doing so, we hope to empower individuals, businesses, and governments to make informed choices that not only reduce energy costs but also contribute to a more environmentally responsible and sustainable future. This exploration will encompass a wide spectrum of topics, from data collection techniques to the implementation of cutting-edge technologies, highlighting the relevance of assessing and optimizing energy consumption in our modern world.

A black and red square with a black stripe

Description automatically generated with medium confidence

Data cleansing, also referred to as data cleaning or data scrubbing, is the process of fixing incorrect, incomplete, duplicate or otherwise erroneous data in a data set. It involves identifying data errors and then changing, updating or removing data to correct them. Data cleansing improves data quality and helps provide more accurate, consistent and reliable information for decision-making in an organization.

Data cleansing is a key part of the overall data management process and one of the core components of data preparation work that readies data sets for use in business intelligence (BI) and data science applications. It's typically done by data quality analysts and engineers or other data management professionals. But data scientists, BI analysts and business users may also clean data or take part in the data cleansing process for their own applications.

A hand holding a diagram of data cleansing

Description automatically generated

Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected.

EDA is an important first step in any data analysis. Understanding where outliers occur and how variables are related can help one design statistical analyses that yield meaningful results. In biological monitoring data, sites are likely to be affected by multiple stressors. Thus, initial explorations of stressor correlations are critical before one attempts to relate stressor variables to biological response variables. EDA can provide insights into candidate causes that should included in a causal assessment.

A diagram of data cleaning

Description automatically generated

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time. Data cleaning is the process that removes data that does not belong in your dataset. Data transformation is the process of converting data from one format or structure into another. Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format for warehousing and analyzing. This article focuses on the processes of cleaning that data.

OBJECTIVES AND SCOPES

Energy conservation can be achieved in two different ways that include reducing the amount of primary energy consumed to supply the useful energy requirement (energy efficiency), and reducing the end point use of nonessential energy.

Energy consumption is a key indicator of the environmental impact and operational efficiency of your facility management (FM) activities. Measuring and monitoring your energy use can help you identify opportunities for improvement, reduce costs, and comply with regulations and standards. But how can you measure energy consumption for your FM activities? Here are some tips and tools to help you get started.

Monitoring energy consumption in households allows us to identify which appliances and devices consume the most energy in our homes. By knowing this information, we can use them more efficiently and reduce our electricity consumption, which translates into long-term economic savings.

Economic Savings: Monitoring energy consumption in households allows us to identify which appliances and devices consume the most energy in our homes. By knowing this information, we can use them more efficiently and reduce our electricity consumption, which translates into long-term economic savings.

ENERGY MONITORING :

As the name suggests, energy monitoring is the process of monitoring energy consumption, whether that’s an individual asset or an entire building.

Energy monitoring software provides vital insight that can help to control and conserve energy in the future. Without conducting energy monitoring, you won’t be able to distinguish where inefficiencies lie within your business to effectively rectify the situation.

According to the Carbon Trust, typically 20% of all business’ annual energy costs are wasted through the use of inefficient equipment. Effective energy monitoring on a granular asset level continuously monitors equipment’s performance and energy output and instantly recognises when an asset is not operating efficiently.

The continuous monitoring of energy consumption provides a daily breakdown of energy usage and makes it easier to identify trends and spikes. By reviewing energy trends over the course of the day, energy managers can determine the best solutions to reduce consumption and possible scheduling options to minimise usage during predicted spikes, such as going off-grid. Real-time energy monitoring is completely scalable, meaning you can monitor one building or an entire estate of facilities.

Monitoring real-time energy data also provides energy managers with better insights into cost per unit of energy. This metric helps organisations to reduce costs in demand or power factor

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

RED = "\033[91m"

GREEN = "\033[92m"

YELLOW = "\033[93m"

BLUE = "\033[94m"

RESET = "\033[0m"

df = pd.read\_csv("/kaggle/input/hourly-energy-consumption/AEP\_hourly.csv")

df["Datetime"] = pd.to\_datetime(df["Datetime"])

# DATA CLEANING

print(BLUE + "\nDATA CLEANING" + RESET)

# --- Check for missing values

missing\_values = df.isnull().sum()

print(GREEN + "Missing Values : " + RESET)

print(missing\_values)

# --- Handle missing values

df.dropna(inplace=True)

# --- Check for duplicate values

duplicate\_values = df.duplicated().sum()

print(GREEN + "Duplicate Values : " + RESET)

print(duplicate\_values)

# --- Drop duplicate values

df.drop\_duplicates(inplace=True)

# DATA ANALYSIS

print(BLUE + "\nDATA ANALYSIS" + RESET)

# --- Summary Statistics

summary\_stats = df.describe()

print(GREEN + "Summary Statistics : " + RESET)

print(summary\_stats)

# Data Visualization

# Line plot for energy consumption over time

plt.figure(figsize=(12, 6))

plt.plot(df.index, df["AEP\_MW"], label="Energy Consumption (AEP\_MW)")

plt.xlabel("Datetime")

plt.ylabel("Energy Consumption (MW)")

plt.title("Energy Consumption Over Time")

plt.grid()

plt.legend()

plt.show()

# SAVING THE FILE

df.to\_csv("/kaggle/working/cleaned\_AEP\_hourly.csv", index=False)

print(BLUE + "\nDATA ANALYSIS" + RESET)

print(GREEN + "Data Cleaned and Saved !" + RESET)

DATA CLEANING

Missing Values :

Datetime 0

AEP\_MW 0

dtype: int64

Duplicate Values :

0

DATA ANALYSIS

Summary Statistics :

Datetime AEP\_MW

count 121273 121273.000000

mean 2011-09-02 03:17:01.553025024 15499.513717

min 2004-10-01 01:00:00 9581.000000

25% 2008-03-17 15:00:00 13630.000000

50% 2011-09-02 04:00:00 15310.000000

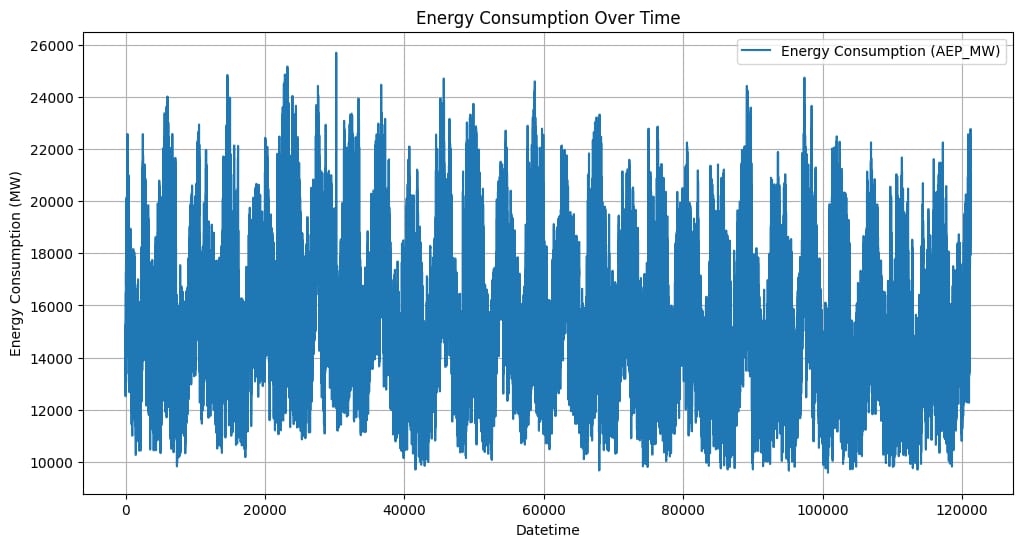
75% 2015-02-16 17:00:00 17200.000000

max 2018-08-03 00:00:00 25695.000000

std NaN 2591.399065

DATA ANALYSIS

Data Cleaned and Saved !



import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings("ignore", category=UserWarning)

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.svm import SVR

from sklearn.metrics import mean\_squared\_error, r2\_score

RED = "\033[91m"

GREEN = "\033[92m"

YELLOW = "\033[93m"

BLUE = "\033[94m"

RESET = "\033[0m"

df = pd.read\_csv("/kaggle/input/hourly-energy-consumption/AEP\_hourly.csv")

df["Datetime"] = pd.to\_datetime(df["Datetime"])

# DATA CLEANING

print(BLUE + "\nDATA CLEANING" + RESET)

# --- Check for missing values

missing\_values = df.isnull().sum()

print(GREEN + "Missing Values : " + RESET)

print(missing\_values)

# --- Handle missing values

df.dropna(inplace=True)

# --- Check for duplicate values

duplicate\_values = df.duplicated().sum()

print(GREEN + "Duplicate Values : " + RESET)

print(duplicate\_values)

# --- Drop duplicate values

df.drop\_duplicates(inplace=True)

# DATA ANALYSIS

print(BLUE + "\nDATA ANALYSIS" + RESET)

# --- Summary Statistics

summary\_stats = df.describe()

print(GREEN + "Summary Statistics : " + RESET)

print(summary\_stats)

# SUPPORT VECTOR MODELLLING

print(BLUE + "\nMODELLING" + RESET)

# Reduce the dataset size for faster training

df = df.sample(frac=0.2, random\_state=42)

# Split the data into features (Datetime) and target (AEP\_MW)

X = df[["Datetime"]]

y = df["AEP\_MW"]

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

X, y, test\_size=0.2, random\_state=42

)

# Preprocess the features (Datetime) to extract the day of the year

X\_train["DayOfYear"] = X\_train["Datetime"].dt.dayofyear

X\_test["DayOfYear"] = X\_test["Datetime"].dt.dayofyear

# Convert X\_train and X\_test to NumPy arrays

X\_train = X\_train["DayOfYear"].values.reshape(-1, 1)

X\_test = X\_test["DayOfYear"].values.reshape(-1, 1)

# Standardize the data

scaler = StandardScaler()

X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

# Create an SVR (Support Vector Regression) model with a linear kernel

svr = SVR(kernel="linear", C=1.0)

# Train the SVR model

svr.fit(X\_train\_scaled, y\_train)

# Predict on the test set

y\_pred = svr.predict(X\_test\_scaled)

# Evaluate the model

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

print(f"Mean Squared Error: {mse}")

print(f"R-squared: {r2}")

# Plot the actual vs. predicted values

plt.figure(figsize=(10, 6))

plt.scatter(X\_test, y\_test, color="b", label="Actual")

plt.scatter(X\_test, y\_pred, color="r", label="Predicted")

plt.xlabel("Day of the Year")

plt.ylabel("Energy Consumption (MW)")

plt.title("SVR Model: Actual vs. Predicted")

plt.legend()

plt.grid()

plt.show()

# DATA VISUALIZATION

print(BLUE + "\nDATA VISUALIZATION" + RESET)

# --- Line plot

print(GREEN + "LinePlot : " + RESET)

plt.figure(figsize=(10, 6))

sns.lineplot(data=df, x="Datetime", y="AEP\_MW")

plt.xlabel("Datetime")

plt.ylabel("Energy Consumption (MW)")

plt.title("Energy Consumption Over Year")

plt.grid()

plt.show()

# --- Histogram

print(GREEN + "Histogram : " + RESET)

plt.figure(figsize=(10, 6))

plt.hist(

df["AEP\_MW"],

bins=100,

histtype="barstacked",

edgecolor="white",

)

plt.xlabel("AEPMW")

plt.ylabel("Frequency")

plt.title("Histogram of MEGAWATT USAGE")

plt.show()

DATA CLEANING

Missing Values :

Datetime 0

AEP\_MW 0

dtype: int64

Duplicate Values :

0

DATA ANALYSIS

Summary Statistics :

Datetime AEP\_MW

count 121273 121273.000000

mean 2011-09-02 03:17:01.553025024 15499.513717

min 2004-10-01 01:00:00 9581.000000

25% 2008-03-17 15:00:00 13630.000000

50% 2011-09-02 04:00:00 15310.000000

75% 2015-02-16 17:00:00 17200.000000

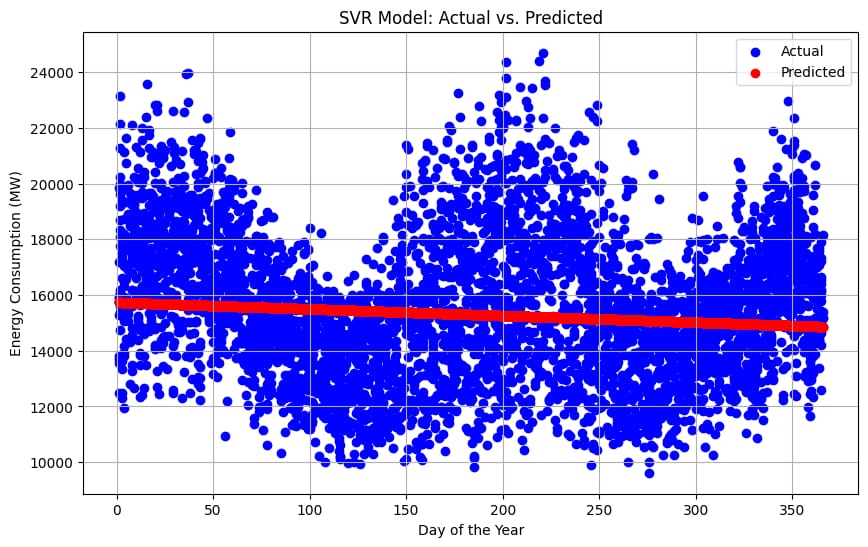
max 2018-08-03 00:00:00 25695.000000

std NaN 2591.399065

MODELLING

Mean Squared Error: 6758395.805638685

R-squared: 0.00270160624748228



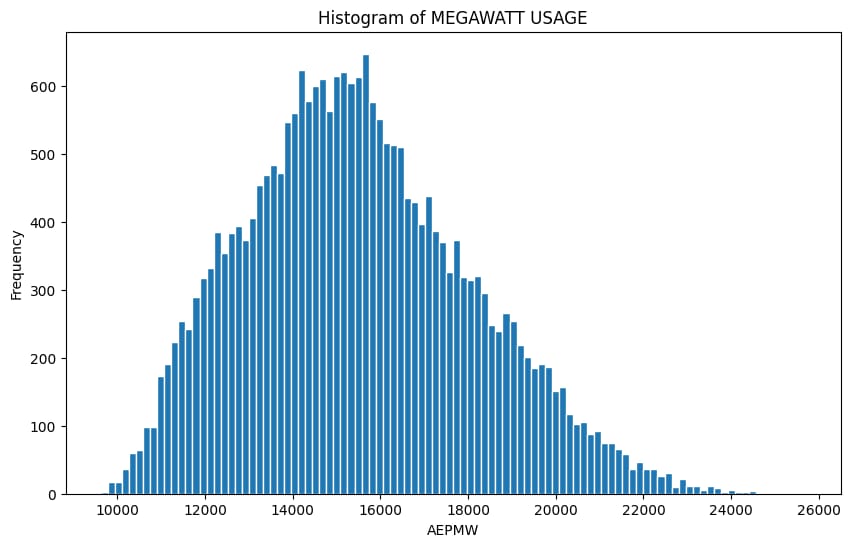
DATA VISUALIZATION

LinePlot

A graph showing the amount of energy consumption

Description automatically generated

Histogram



Software Consumption Measurement Process

The proposal described in this paper is a repeatable process for measuring the consumption of a software application, hereinafter called the Software Under Test (SWUT). The process consists of the following four phases:

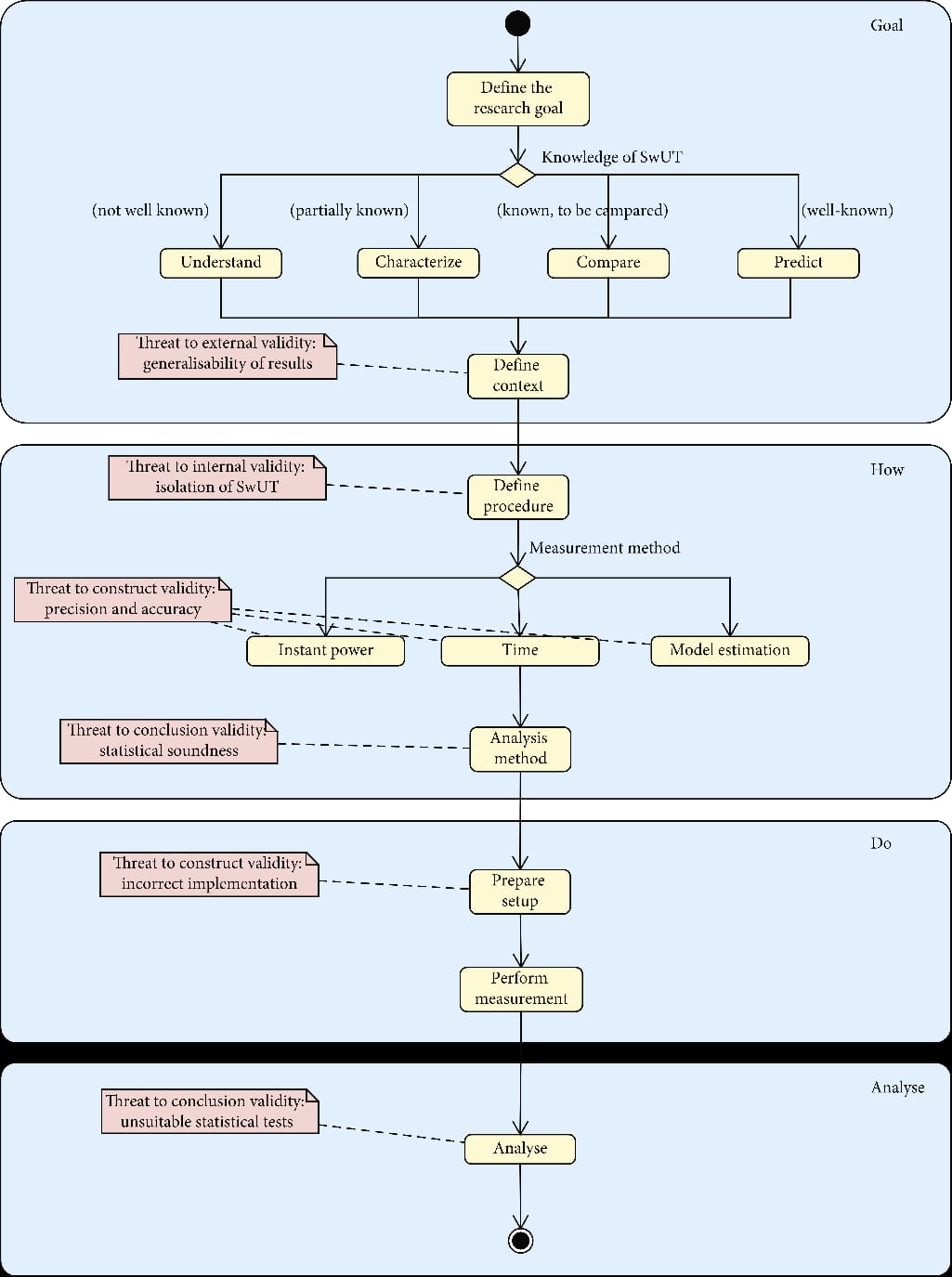
(i) Goal (G): define the research question, the target device(s) on which the measurements will take place, and the context in which the SWUT is executed

(ii) How (H): decide how consumption will be measured and the procedure needed to carry out the measurement

(iii) Do (D): carry out the measurement and collect the data

(iv) Analyse (A): analyse the data and address the research question(s)

The UML activity diagram in Figure 1 summarizes the main activities and decisions encompassed by the process and the relative threats to the validity of the results.



Conclusion: The aim of this project was to identify the variables that influence the generation, the consumption and the price of the electricity.

For the structure of the electricity production, we have seen that the energy mix varies tremendously from one region to another and from one state to another. We cannot determine whether a mix defines the price per KWh or not. However, power generation using coal and hydropower is correlated with low energy costs.

Even if we considered to study the fluctuation of the electricity price for the next years, it turned out impossible to achieve a result. Therefore, it would be interesting to analyse accurate external data such as the weather, the cost price per KWh per energy, the political decisions, etc. Those aspects have a direct influence on the price of energy.

A thank you card with pictures of wind turbines and light bulb

Description automatically generated

..