MIS – 515
Final Project Report

Cloud Computing Performance Analysis

Akash Singh
Anand Mohan Vyas
Manasi Nair
Pradyutha Kaushik
Tanmay Jagtap

Table of Contents:

I. Executive Summary:

The increasing global reliance on cloud computing services has led to a rise in energy consumption in data centers, directly impacting operational costs and contributing to higher carbon emissions. This project analyses a dataset that contains a wide array of performance metrics from a simulated cloud computing environment. Key performance indicators such as CPU usage, memory usage, network traffic, power consumption, execution time, energy efficiency, and task priority have been captured.

1. Explore the potential of machine learning techniques to optimize energy efficiency and reduce execution time in cloud environments.
2. Analyse how various states and conditions in cloud systems interact with machine learning optimization strategies.
3. Develop machine learning models to:
    I. Identify inefficiencies in cloud resource usage.
    II. Predict optimal resource allocation.
    III. Reduce unnecessary energy use while boosting system performance.

II. Description of Business:

Our project is based on a cloud service provider specializing in high-performance cloud infrastructure for data processing, machine learning, and web services. It operates within a large-scale environment, managing virtual machines to deliver efficient, scalable, and reliable cloud solutions.

However, challenges such as high energy consumption, inefficient resource use, and delayed task execution hinder their ability to optimize operations.

The primary beneficiaries of this effort are the customers, including businesses that rely on cloud services for their critical operations. By addressing these issues, we will offer enhanced performance, cost savings, and a sustainable, energy-efficient solution for our clients.

III. Project Detail:

The dataset for this project includes key performance attributes such as

- VM ID (Virtual Machine ID)
- Timestamp
- CPU usage
- memory usage
- network traffic
- power consumption
- execution time
- energy efficiency
- task type
- task priority
- task status

This dataset is sourced from the cloud infrastructure's monitoring systems and logs, capturing real-time operational data. The tools used for this project will include Azure Data Lake for data storage, Azure Databricks for data processing and analytics, and Azure Data Factory for creating the ETL (Extract, Transform, Load) pipelines to manage and transform the data. These tools will enable seamless integration and analysis of the dataset, supporting predictive and prescriptive analytics to optimize cloud performance.

IV. Type of Analytics Used:

The project will employ a combination of analytics types, each serving a specific purpose:

a. **Descriptive Analytics**: This will be used to analyse historical data, providing insights into past cloud system performance, CPU and memory usage, network traffic, and energy consumption patterns. This phase helps to understand baseline behaviours and identify areas for improvement.

b. **Predictive Analytics**: Machine learning algorithms will be employed to predict future states of the cloud environment, such as potential performance bottlenecks, energy consumption spikes, or network traffic surges. These predictions are key to preemptively addressing issues and improving resource allocation.

c. **Prescriptive Analytics**: Based on the insights gathered, prescriptive analytics will provide recommendations for optimizing system parameters. For instance, specific changes to task priority settings or resource allocation strategies can be suggested to reduce costs or improve system efficiency.

V. Data Utilization for Business Outcomes:

The project's analysis of cloud performance metrics is designed to drive significant business outcomes. By leveraging machine learning to identify and address inefficient resource utilization, such as excessive CPU or power consumption, the project will enable dynamic scaling and energy savings, which will significantly lower operational costs.

The data will be used to reduce costs by analyzing performance metrics to identify and address inefficient resource utilization. Machine learning models will suggest dynamic scaling and energy-saving strategies, which will lower the operational expenses by optimizing resource allocation and reducing waste.

To increase revenue, the insights gained from improved efficiency will enhance service quality, boost client satisfaction, and attract new customers seeking cost-effective and eco- friendly solutions. This will lead to higher customer retention and new client acquisition.

Additionally, this project will create new business opportunities by developing energy- efficient cloud packages and offering consultancy services to help other businesses optimize their cloud infrastructures.

VI. Data Source, Acquisition and Preparation

Data Source: https://www.kaggle.com/datasets/abdurraziq01/cloud-computing-performance-metrics/data

Datatypes:

```
[6]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000000 entries, 0 to 1999999
Data columns (total 12 columns):
 #   Column                      Dtype
---  ------                      -----
 0   vm_id                       object
 1   timestamp                   object
 2   cpu_usage                   float64
 3   memory_usage                float64
 4   network_traffic             float64
 5   power_consumption           float64
 6   num_executed_instructions   float64
 7   execution_time              float64
 8   energy_efficiency           float64
 9   task_type                   object
 10  task_priority               object
 11  task_status                 object
dtypes: float64(7), object(5)
memory usage: 183.1+ MB
```

Based on the dataset structure and the common metrics used in cloud resource management, here are the units for the various columns:

1. vm_id: No units (this is an identifier for Virtual Machines).
2. timestamp: No units (this represents date and time in a standard format).
3. cpu_usage: Measured as a percentage (%), representing the percentage of CPU utilization.
4. memory_usage: Measured as a percentage (%), representing the percentage of memory utilization.
5. network_traffic: Measured in megabytes (MB), representing the amount of network traffic used by the virtual machine.
6. power_consumption: Measured in watts (W), representing the power consumed by the virtual machine.
7. num_executed_instructions: Measured in millions of instructions, representing the number of executed instructions by the virtual machine.
8. execution_time: Measured in seconds (s), representing the time taken to execute a task.
9. energy_efficiency: This is a derived metric representing how efficiently the system uses energy relative to the workload.
10. task_type: No units (this is a categorical value describing the task type).
11. task_priority: No units (this is a categorical value describing the priority of the task).
12. task_status: No units (this is a categorical value describing the status of the task).

Data Acquistion:

The dataset has 2000000 rows and 12 columns. We downloaded the dataset from Kaggle. The file was a CSV (Comma Separated Values) file.

VII. Data Preparation

Data Cleaning:

We have reviewed the dataset and plan to use most of the attributes for analyzing cloud computing performance. After handling the missing values, we will further refine our focus by selecting key attributes for deeper analysis.

Adding new attributes:

To focus on the environmental aspects of computing, we decided to introduce a couple of new attributes namely Energy Efficiency Ratio and Sustainability Index. They aim to enhance the understanding of cloud computing performance in relation to environmental impact.

The EER is determined by the ratio of power consumption to the number of executed instructions, offering a definitive measure of energy efficiency. A reduced EER value indicates diminished energy usage per executed job, essential for recognizing efficient workloads and enhancing energy consumption in cloud operations. This statistic can assist firms in minimizing expenses while enhancing their sustainability policies.

Conversely, the Sustainability Index (SI) amalgamates various elements, including CPU use, energy consumption, and execution duration, to provide a comprehensive sustainability score. The SI is determined by multiplying the EER by network traffic and dividing by the total of power consumption and execution time. An elevated SI signifies enhanced performance with diminished environmental effect, enabling firms to assess their operations comprehensively. By emphasizing efficiency and sustainability, these characteristics offer a framework for firms to synchronize their cloud computing plans with corporate social responsibility objectives, therefore fostering environmentally friendly practices in the technology sector.

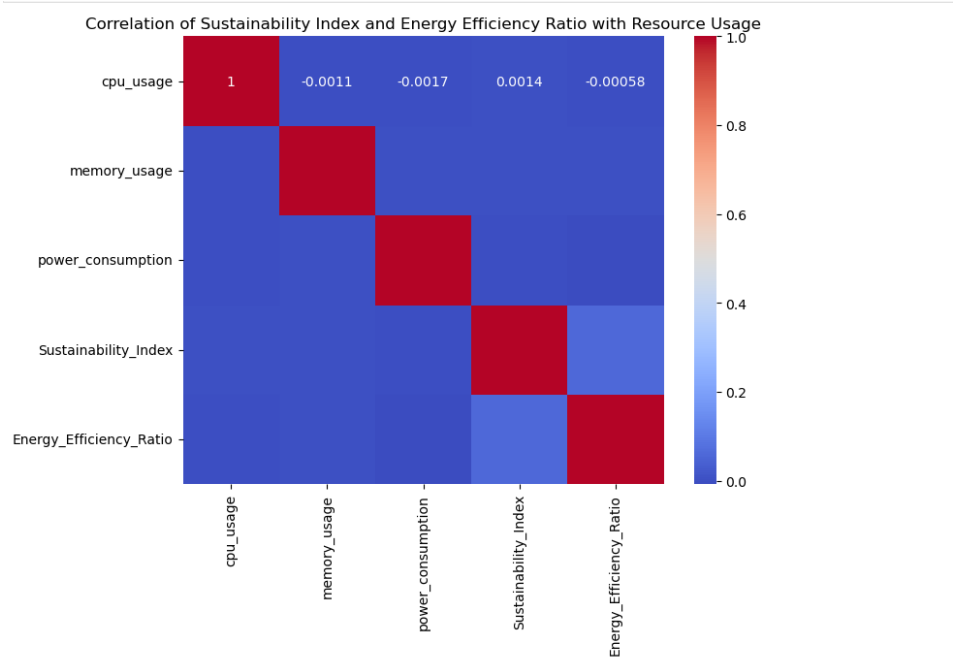
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000000 entries, 0 to 1999999
Data columns (total 21 columns):
 #   Column                      Dtype
---  ------                      -----
 0   vm_id                       object
 1   timestamp                   datetime64[ns]
 2   cpu_usage                   float64
 3   memory_usage                float64
 4   network_traffic             float64
 5   power_consumption           float64
 6   num_executed_instructions   float64
 7   execution_time              float64
 8   energy_efficiency           float64
 9   Sustainability_Index        float64
 10  Energy_Efficiency_Ratio     float64
 11  resource_efficiency         float64
 12  task_type_compute           bool
 13  task_type_io                bool
 14  task_type_network           bool
 15  task_priority_high          bool
 16  task_priority_low           bool
 17  task_priority_medium        bool
 18  task_status_completed       bool
 19  task_status_running         bool
 20  task_status_waiting         bool
dtypes: bool(9), datetime64[ns](1), float64(10), object(1)
memory usage: 200.3+ MB
```

Attributes on focus:

The characteristics we will highlight in our focus on the environmental impact of cloud computing performance are network traffic, power consumption, and energy efficiency. Power consumption is a critical metric as it directly correlates with the environmental footprint of cloud services, understanding how much energy is being consumed helps identify areas for improvement.

```
Correlation Matrix:
                         cpu_usage  memory_usage  power_consumption  \
cpu_usage                 1.000000     -0.001102          -0.001706
memory_usage             -0.001102      1.000000           0.000283
power_consumption        -0.001706      0.000283           1.000000
Sustainability_Index      0.001363      0.001328          -0.001940
Energy_Efficiency_Ratio  -0.000580      0.000362          -0.008001

                         Sustainability_Index  Energy_Efficiency_Ratio
cpu_usage                            0.001363                -0.000580
memory_usage                         0.001328                 0.000362
power_consumption                   -0.001940                -0.008001
Sustainability_Index                 1.000000                 0.058079
Energy_Efficiency_Ratio              0.058079                 1.000000
```



Insights into resource utilization will also be obtained from the Energy Efficiency Ratio (EER). This feature will make it possible to distinguish between workloads that need optimization and those that are energy efficient. Finally, network traffic will be included in the Sustainability Index (SI) computation to emphasize how data transfer and energy consumption interact.

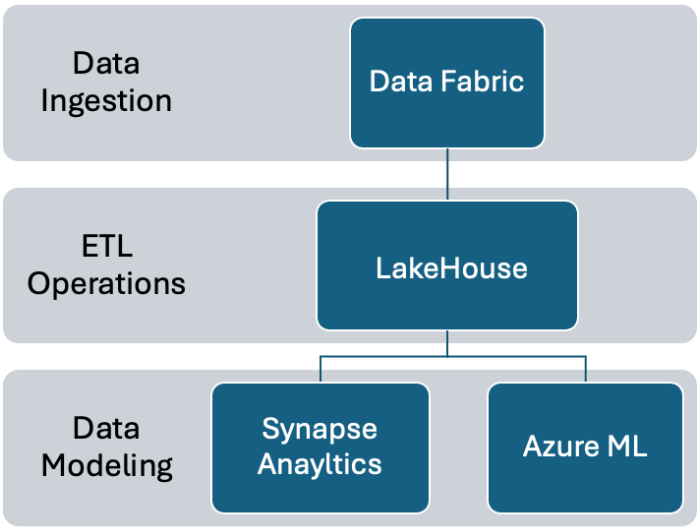
VIII. Tools and Technology

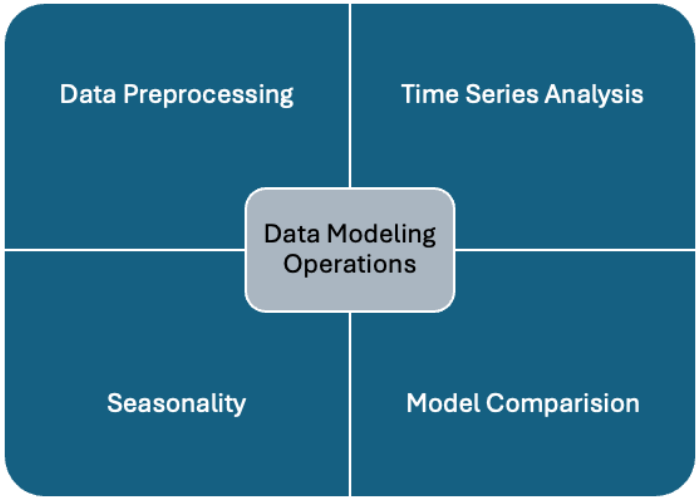
We have used Jupyter Notebook for Python coding to efficiently conduct our analysis because of its adaptability and user-friendliness for data manipulation and visualization. But to take advantage of Microsoft Azure Functions' cloud capabilities for scalable processing and data storage, we want to integrate more of them into our workflow. WE have tried using databricks and Azure storage to integrate data and the notebook. But as we are currently trying to understand our dataset, we are yet to finalize our preferred Azure service.

This combination will make it easier to analyse the performance metrics thoroughly and enable us to make insightful decisions regarding how cloud computing practices affect the environment.

Deliverable 3

IX. Data Cleaning/Processing





Data Ingestion:

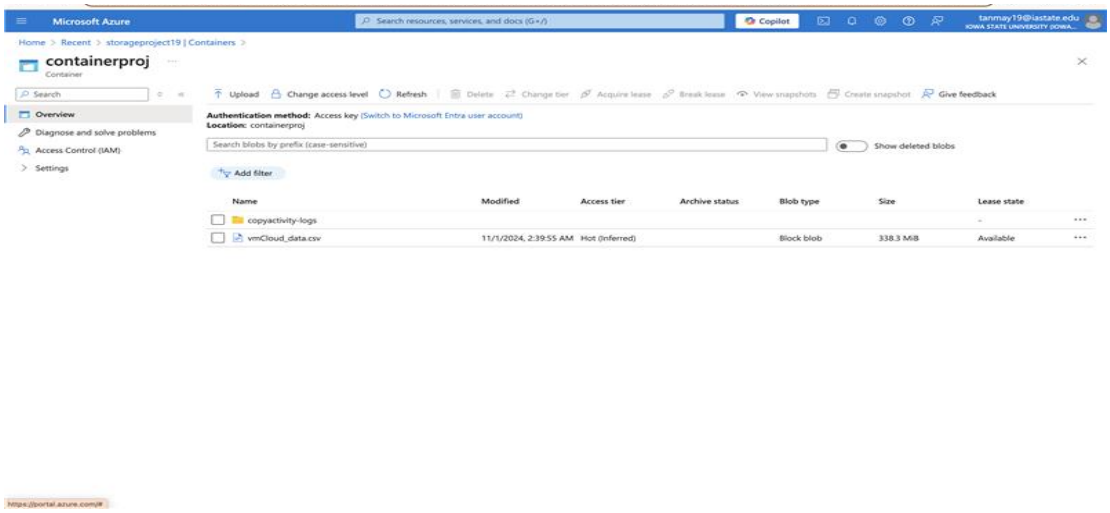
Data ingestion is the first step in transforming raw data into actionable insights. In this project, data ingestion is managed by a Data Fabric, an advanced architecture that integrates data from multiple, diverse sources into a cohesive framework. The ingestion process also includes data

enrichment and validation steps, enhancing raw data with contextual information that might be useful for downstream analytics and machine learning. This robust setup ensures that the incoming data seamlessly flows into the system, providing a solid foundation for accurate analysis and model training later in the pipeline.

In essence, the Data Fabric acts as the backbone for data ingestion, enabling the continuous, efficient integration of vast datasets essential for monitoring cloud performance and driving insights into energy efficiency and resource optimization

ETL Operations:

The data moves into the **LakeHouse** layer, which likely serves as a data lake or warehouse. Here, ETL (Extract, Transform, Load) operations are performed, preparing the data for further analysis.



Data Modeling Operations:

Under this layer, **Synapse Analytics** and **Azure ML** are shown as the primary tools, likely for data processing and machine learning. Synapse Analytics might handle big data analytics, while Azure ML supports machine learning and AI model training.

To prepare the data for analysis, we conducted a series of data cleaning and engineering steps. First, we converted the time stamp into a datetime format to enable time-based analysis. Then, we addressed missing values in the VM ID and time stamp columns, removing rows with null entries to ensure data integrity.

We also streamlined the dataset by reducing it from 2 million observations to approximately 1.6 million (1,618,908 rows), enhancing computational efficiency. For the categorical variables "Task Type," "Task Priority," and "Task Status," we created dummy variables, converting these categories into Boolean representations. After this transformation, the original categorical columns were dropped, simplifying the dataset structure.

As a result, we finalized a clean dataset with **19 attributes**, ready for further analysis and model building.

Model Comparison for Optimal Performance and Efficiency

Selecting the right predictive model is crucial to optimizing cloud resources effectively. To this end, we'll perform a rigorous model comparison to identify the most suitable approach for forecasting resource usage and energy demands. This comparison begins with establishing baseline results using simpler models like linear regression, providing a foundational accuracy level to improve upon.

Advanced machine learning models, including gradient boosting (e.g., XGBoost), random forests, and neural networks (specifically LSTM for time series data), will then be trained and evaluated. Key evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared will guide this process, enabling a data-driven assessment of model accuracy. Model interpretability is also a priority; techniques like SHAP values will clarify feature importance, offering insights into the factors that influence predictions. The outcome of this analysis will be an optimal model that balances accuracy and transparency, ready to be integrated into real-time decision-making processes to improve operational efficiency.

Seasonality Detection to Address Recurrent Patterns

In cloud operations, certain usage patterns may recur regularly, such as peaks during specific hours or days, influenced by client behavior or operational schedules. Detecting these seasonal trends is vital for efficient resource management. For instance, scaling down resources during predictable low-demand times can significantly reduce energy costs, while scaling up during high-demand periods maintains performance.

This project will leverage seasonality detection techniques, including time series decomposition, to isolate recurring patterns. Approaches like STL (Seasonal and Trend decomposition using Loess) and Fourier transformation will help identify these cycles within the data. By understanding seasonality, the cloud provider can optimize resource provisioning dynamically, maintaining a balance between energy efficiency and performance. Detecting these patterns allows us to make data-driven decisions that conserve resources while ensuring system readiness during demand spikes.

Time Series Analysis for Energy Efficiency and Performance Prediction

With the growing demand for cloud services, understanding resource usage trends is essential for efficiency. This project utilizes time series analysis on key metrics—such as CPU and memory usage, power consumption, and network traffic—to forecast future demands and identify usage patterns. By examining these metrics over time, we aim to uncover trends that can guide proactive resource allocation, ensuring that the cloud environment is prepared to handle varying workloads efficiently.

To achieve this, data will be aggregated at meaningful intervals, such as hourly or daily, based on observed usage patterns. This approach ensures a smoother time series analysis and enhances prediction accuracy. We will explore various models for this task, including ARIMA, Prophet, and Long Short-Term Memory (LSTM) networks. LSTM is advantageous for capturing complex dependencies over time, making it ideal for high-dimensional, sequential data. Insights derived from this analysis will highlight peak usage times and seasonal fluctuations, supporting preemptive resource planning and driving energy-saving measures.

X. Assumptions:

1. Energy Efficiency Analysis of Cloud VMs

A deep dive into the energy consumption patterns of different VM types and workloads can reveal opportunities for optimization. By analysing factors such as CPU utilization, memory usage, and network traffic, it is possible to identify energy-hungry VMs and implement strategies to reduce their energy consumption. Techniques like rightsizing VMs, consolidating workloads, and optimizing power management settings can significantly impact overall energy efficiency.

2. Independence of Observations

For our analysis, we are assuming that each entry of VM performance is independent of the others. This implies that one VM's performance at a specific time does not affect another VM's performance or even that same VM's performance at a different time. By assuming independence, we can analyse each data point on its own merit without worrying about dependencies that could skew our results. This will help in capturing the environmental effects which is the goal of the project.

3. Predictive Modelling for Energy Consumption

Predictive modelling empowers us to forecast future energy consumption by leveraging historical data on VM usage patterns, energy consumption, and environmental factors. Accurate predictions enable proactive resource allocation, maintenance scheduling, and energy-saving strategies. By avoiding overprovisioning and underutilization of resources, we can optimize energy consumption and reduce operational costs.

4. Adequate Sample Size

We believe this dataset is large enough to cover a wide range of VM behaviours under various conditions. This includes high and low workloads, different task types, and varying performance states. A sufficiently large sample size helps us generalize our findings and draw meaningful insights from the data. In other words, we're confident that this dataset includes enough diversity to represent typical VM operations across different scenarios.

5. Measurement Consistency

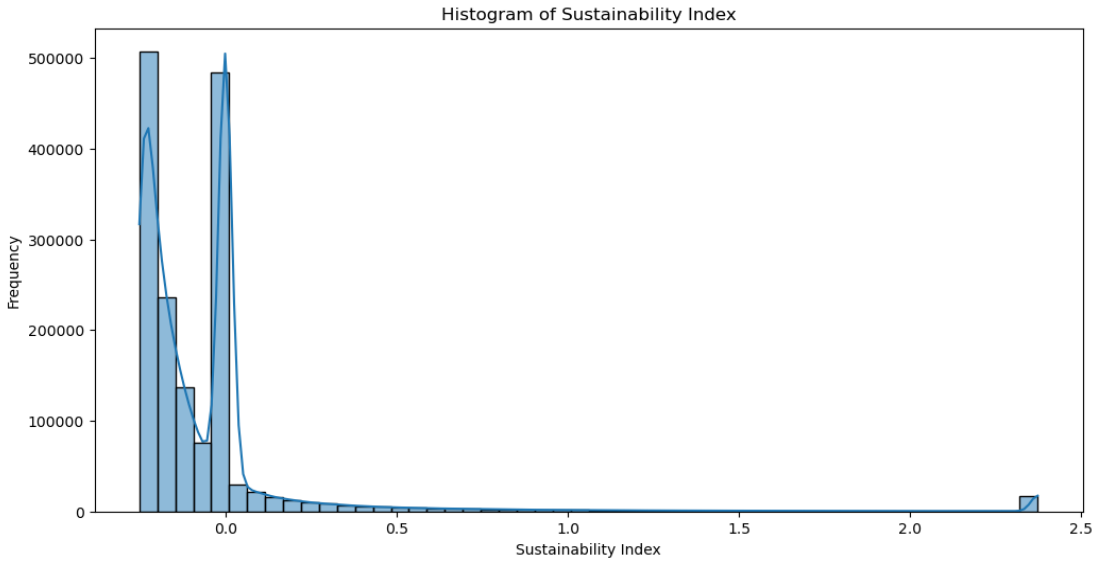
We assume that the metrics, such as CPU usage, memory utilization, network transfer, and power consumption, were measured in a consistent manner for each VM entry. This consistency would mean that each metric was recorded using the same methods, units, and conditions. Consistent measurement is crucial because it allows us to make accurate comparisons across VMs and time periods. Without this, variations in measurement could introduce discrepancies that would compromise our analysis.

XI. Visualization

To shift the focus of the Exploratory Data Analysis (EDA) to the two calculated columns, Sustainability_Index and Energy_Efficiency_Ratio, we can create visualizations that analyze their distributions, relationships, and potential insights.

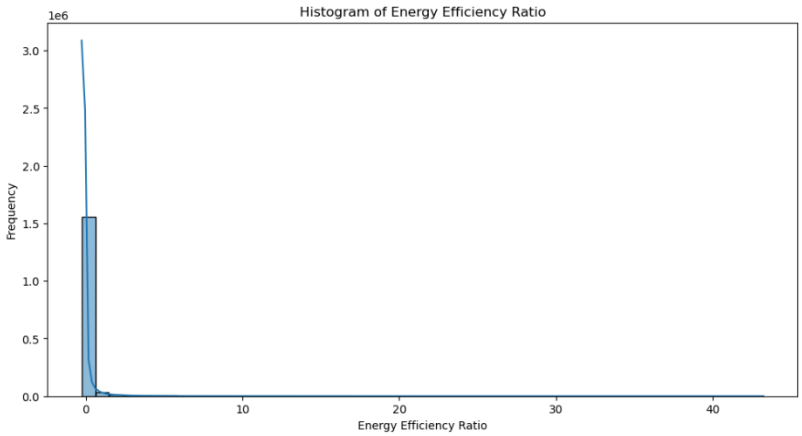
The focus of the EDA is shifted to `Sustainability_Index` and `Energy_Efficiency_Ratio` because they directly measure the environmental and operational efficiency of cloud operations, aligning with the project's goal of assessing the environmental impact of cloud storage. Analyzing these metrics provides insights into how efficiently resources are used and how sustainable the operations are. Visualizing their distributions and relationships with `power_consumption` helps identify patterns, outliers, and trends, allowing for targeted recommendations and improvements in energy management. This focus supports the identification of key areas for optimization and enhances the relevance of predictive modeling for sustainability analysis.

Sustainability Index Histogram: `Sustainability_Index`, indicating how most values are skewed towards the lower end, with a few higher values. This helps identify the general sustainability performance and whether improvements are needed across the board or just in certain areas.

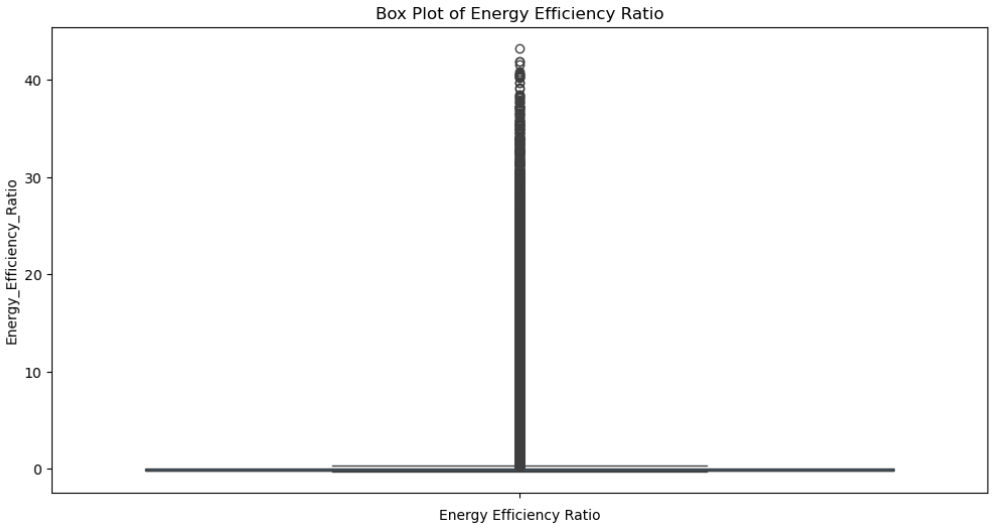


Distribution of Sustainability_Index and Energy_Efficiency_Ratio to understand their spread and skewness

Energy Efficiency Ratio Histogram: Displays the distribution of the `Energy_Efficiency_Ratio`, highlighting how many data points fall into lower versus higher efficiency ranges. The skewness here can point to potential inefficiencies or highly efficient operations that may be isolated cases/

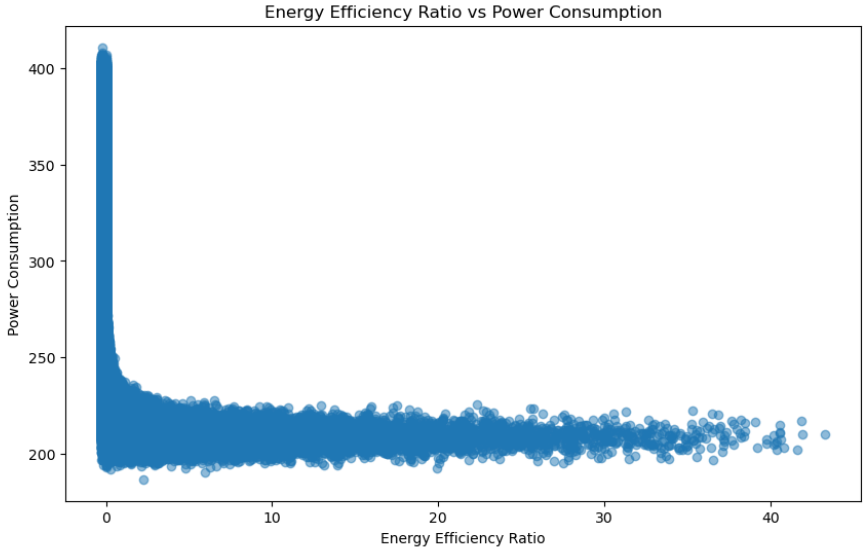


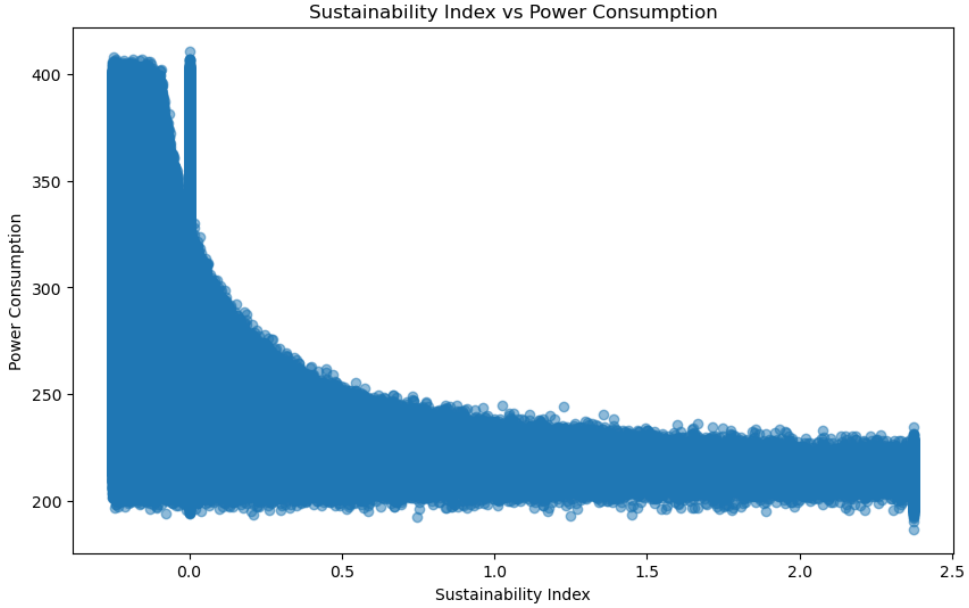
Sustainability Index Box Plot: Illustrates the range, median, and outliers of the `Sustainability_Index`. The presence of outliers may suggest occasional highly efficient or inefficient periods that deviate from typical operations, which could be investigated further.



Range, median, and outliers for each column

Sustainability Index vs. Power Consumption Scatter Plot: Shows how `power_consumption` changes in relation to the `Sustainability_Index`. A negative trend or clustering can indicate that higher sustainability often corresponds with lower power consumption.

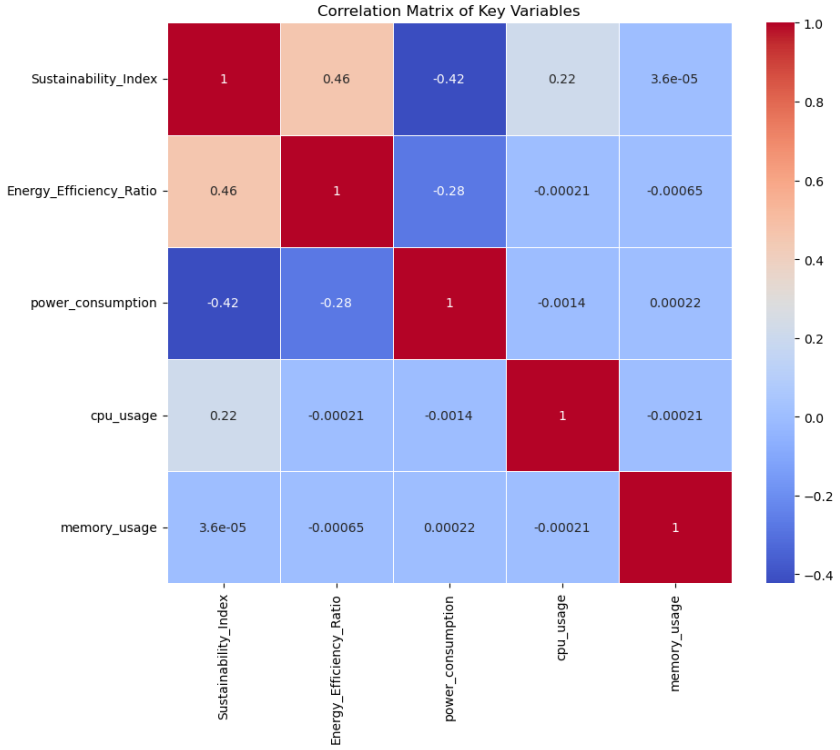




Relationship between these calculated columns and power_consumption to assess any potential linear or non-linear relationships.

```
Correlation matrix:
                        Sustainability_Index  Energy_Efficiency_Ratio  \
Sustainability_Index                1.000000                 0.459210
Energy_Efficiency_Ratio             0.459210                 1.000000
power_consumption                  -0.423152                -0.277549
cpu_usage                           0.215101                -0.000206
memory_usage                        0.000036                -0.000645

                        power_consumption  cpu_usage  memory_usage
Sustainability_Index            -0.423152   0.215101      0.000036
Energy_Efficiency_Ratio         -0.277549  -0.000206     -0.000645
power_consumption                1.000000  -0.001375      0.000220
cpu_usage                       -0.001375   1.000000     -0.000205
memory_usage                     0.000220  -0.000205      1.000000
```



Pearson Measures linear correlation between variables suitable if relationships are expected to be linear.

The correlation analysis between Sustainability_Index, Energy_Efficiency_Ratio, and key variables (power_consumption, cpu_usage, memory_usage) reveals the following

Sustainability_Index and **Energy_Efficiency_Ratio** show a negative correlation with power_consumption, indicating that higher values are associated with lower power consumption.

The relationships between these calculated columns and operational metrics like cpu_usage and memory_usage help identify dependencies and areas for optimization. A heatmap visualization was used to show the strength of these correlations, where values close to 1 or -1 indicate strong associations

XII. Recommendations:

In this project, we aim to optimize cloud resource usage by analyzing the relationships between key performance metrics—CPU usage, memory allocation, and task type—and their impact on energy efficiency and execution time. By understanding these interactions, we can develop strategies that improve system performance, reduce energy consumption, and create a more efficient cloud environment.

- Is there a significant relationship between CPU usage and energy efficiency?
    - The impact of CPU usage on energy efficiency is critical when optimizing cloud computing resources. Suppose higher CPU usage has a substantial adverse effect on energy efficiency. In that case, we can use information from CPU usage on Energy efficiency can be used to develop resource management strategies that maximize CPU allocation while minimizing energy consumption.

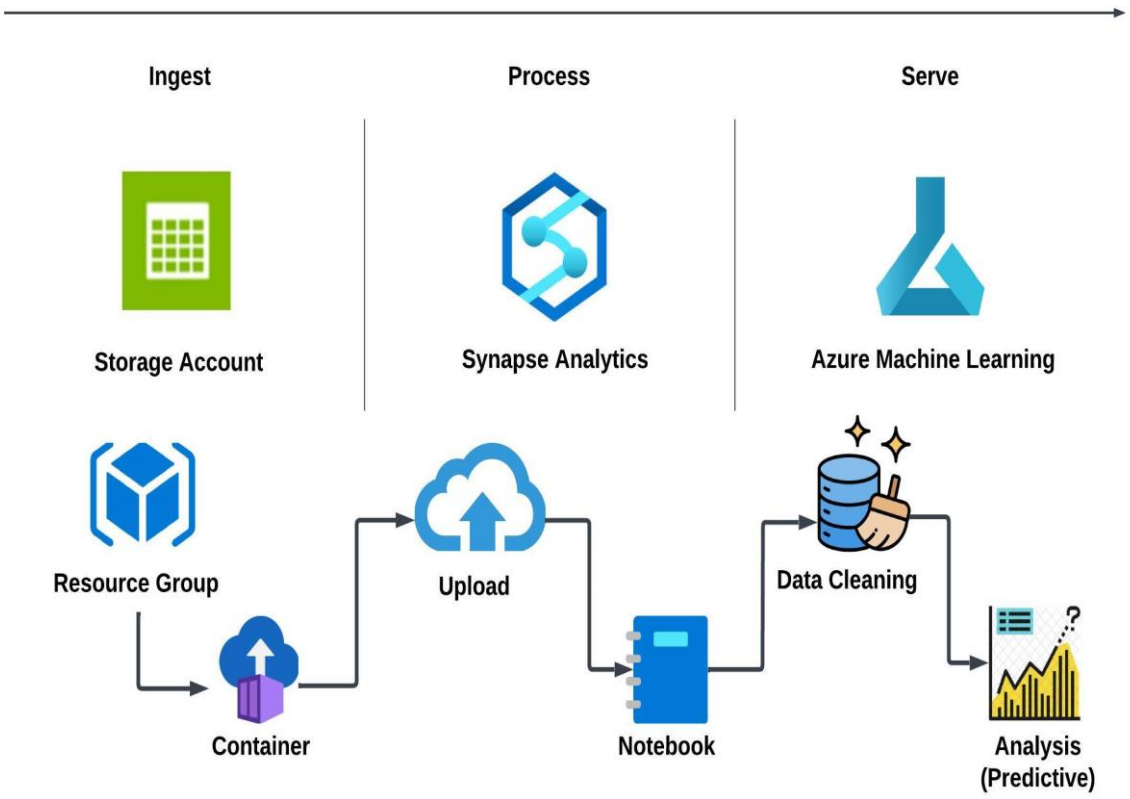
- Does task type heavily impact power consumption, and can this be used to optimize energy efficiency?
    - Tasks like compute, IO, and network could have different energy footprints. By understanding whether task type significantly affects power consumption, we want to find the most optimal model for assigning tasks that reduce overall power usage.
    - Scheduling tasks based on their energy requirements could result in more efficient energy usage.

- How does memory usage affect execution time, and how can this relationship be optimized to reduce system load?

    - Understanding how memory usage affects the time it takes to complete tasks can help us improve overall system performance. If using more memory takes longer, we can speed up tasks by using memory more efficiently. Tasks with more memory take longer and improving memory allocation could speed up overall execution.

Deliverable 4

XIII. Revised Flow Chart



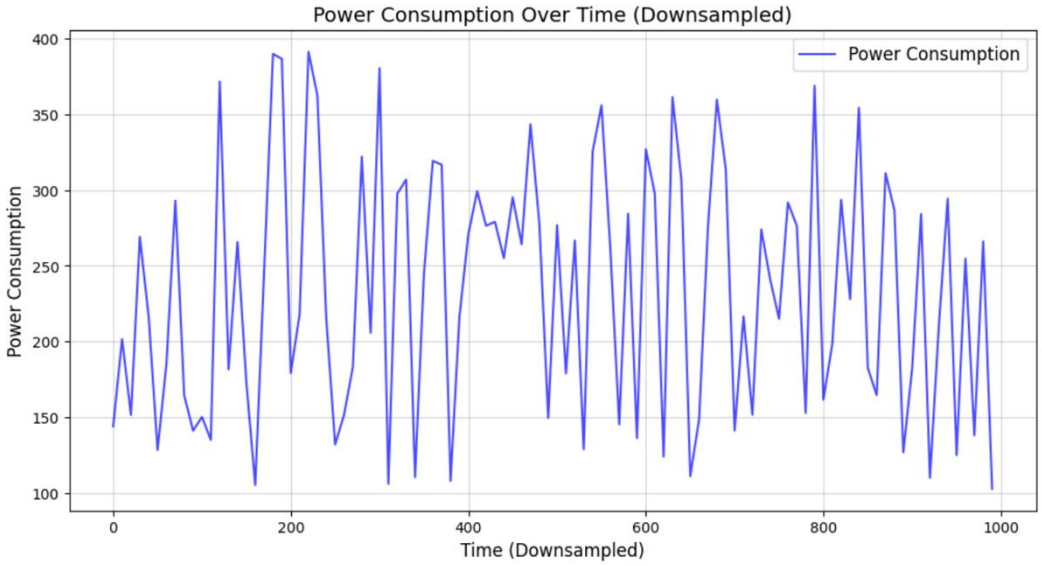
XIV. Analysis

1. Power Consumption Trends

This graph demonstrates the variation in power consumption over time. The following observations can be drawn:

- **Peaks and Troughs**: The data exhibits high variability with noticeable peaks and troughs, indicating fluctuating energy demands.
- **Seasonality**: Some recurring patterns suggest seasonal impacts, which may relate to external factors like weather conditions, business cycles, or operational changes.

Key Insights: Understanding these trends is essential for predicting future energy needs and optimizing energy usage patterns.

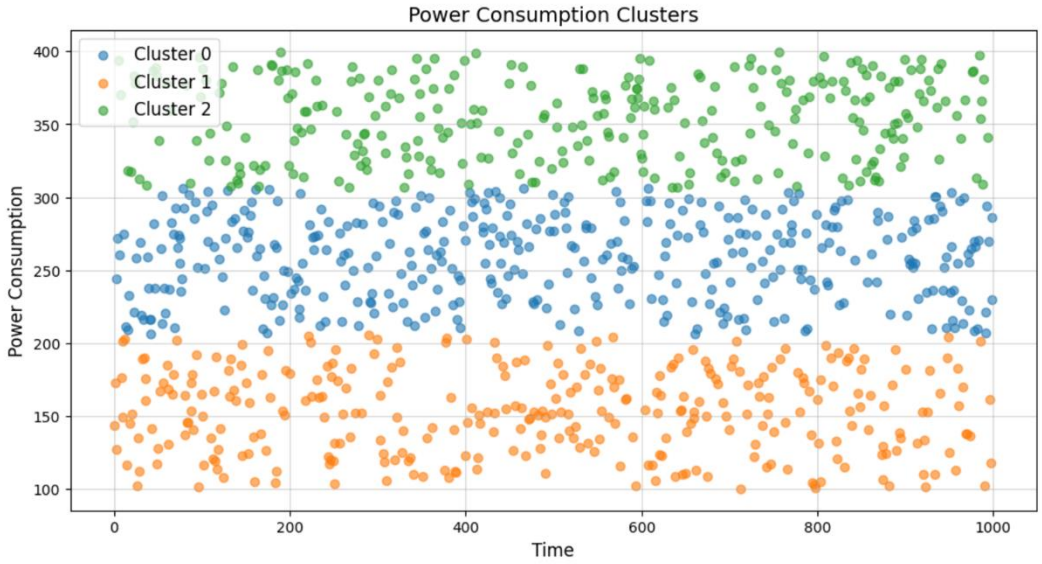


2. Power Consumption Clusters

The clustering analysis groups power consumption data into distinct categories:

- **Cluster 0** (e.g., High Consumption): Represents periods of peak energy usage, often during high-demand operations or seasons.
- **Cluster 1** (e.g., Moderate Consumption): Indicates average energy use, reflecting standard operating conditions.
- **Cluster 2** (e.g., Low Consumption): Denotes periods of low energy use, possibly during downtime or efficient operations.

Key Insights: This clustering enables targeted strategies for each consumption level, such as implementing energy-saving measures during high-consumption periods.

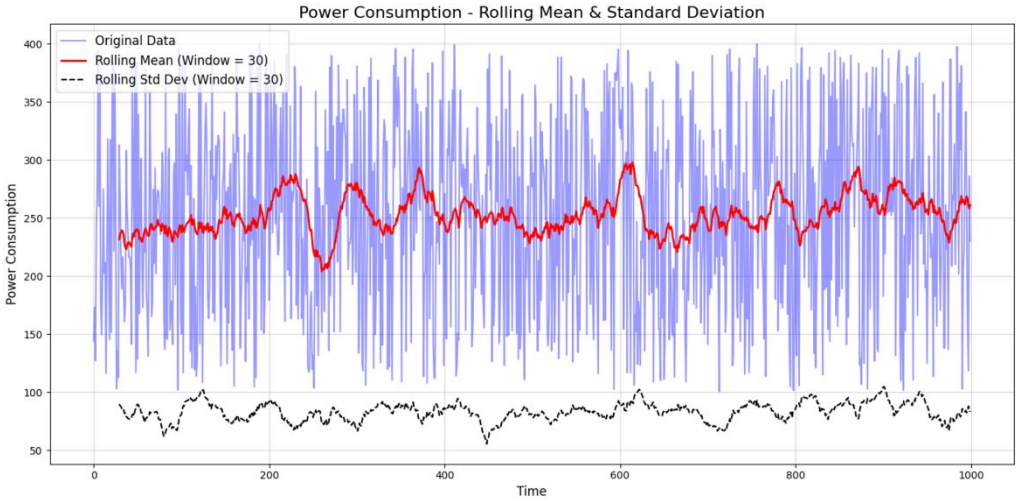


3. Rolling Statistics Analysis

This chart highlights:

- **Rolling Mean**: Illustrates long-term trends in power consumption.
- **Rolling Standard Deviation**: Indicates variability over time.

Key Insights: Understanding these metrics allows for better anomaly detection, ensuring smoother energy management and identifying periods of inefficiency.

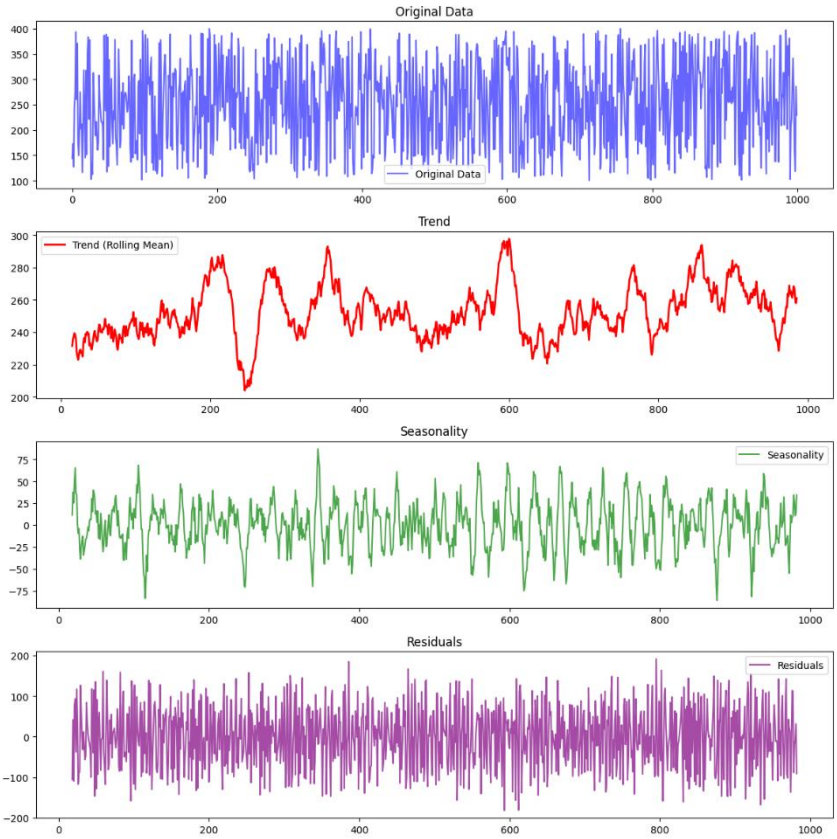


4. Decomposed Components

This chart breaks down power consumption into:

- **Trend**: The overall direction of power usage over time.
- **Seasonality**: Recurring patterns due to periodic factors.
- **Residuals**: Noise or irregular fluctuations.

Key Insights: These components reveal the underlying drivers of power consumption, facilitating improved forecasting and planning.

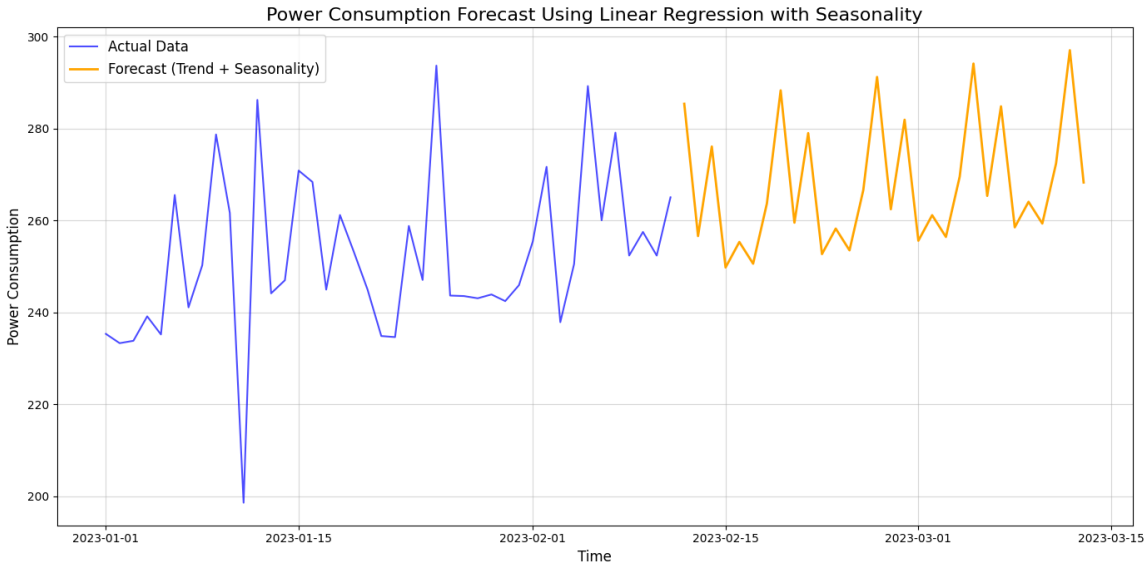


5. Forecasting with Seasonality

Using linear regression with seasonal adjustments:

- The **blue line** shows actual historical power consumption.
- The **orange line** predicts future consumption based on trends and seasonality.

Key Insights: Forecasting helps anticipate energy needs and plan for potential spikes or reductions in demand.

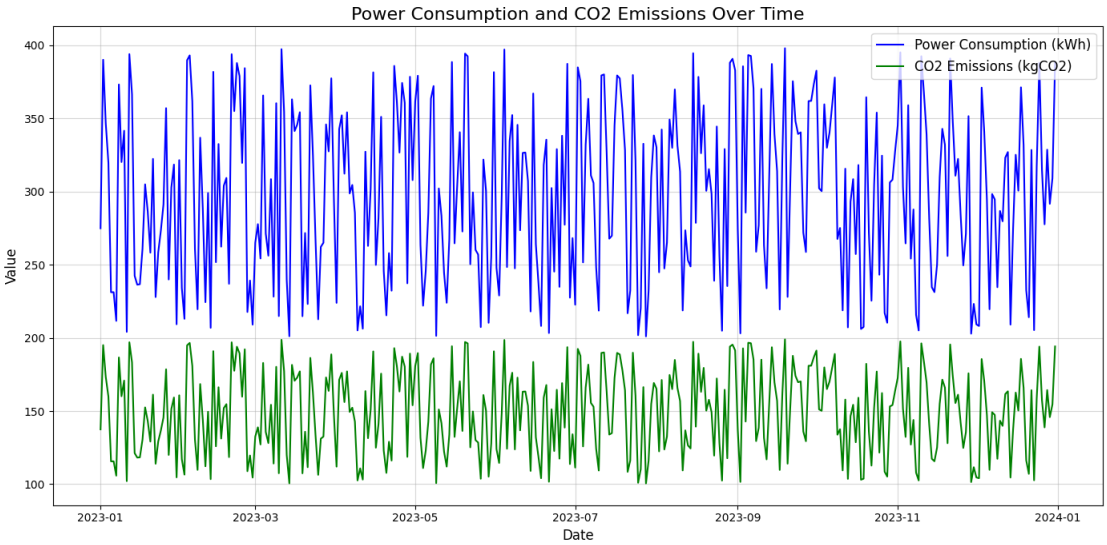


6. Power Consumption vs. CO_2 Emissions

This comparison shows:

- **Power Consumption (Blue Line)**: The energy used over time.
- **CO_2 Emissions (Green Line)**: The corresponding emissions based on an average factor of 0.5 $kgCO_2/kWh$.

Key Insights: Higher power consumption directly results in higher CO_2 emissions, underscoring the environmental impact of energy use.

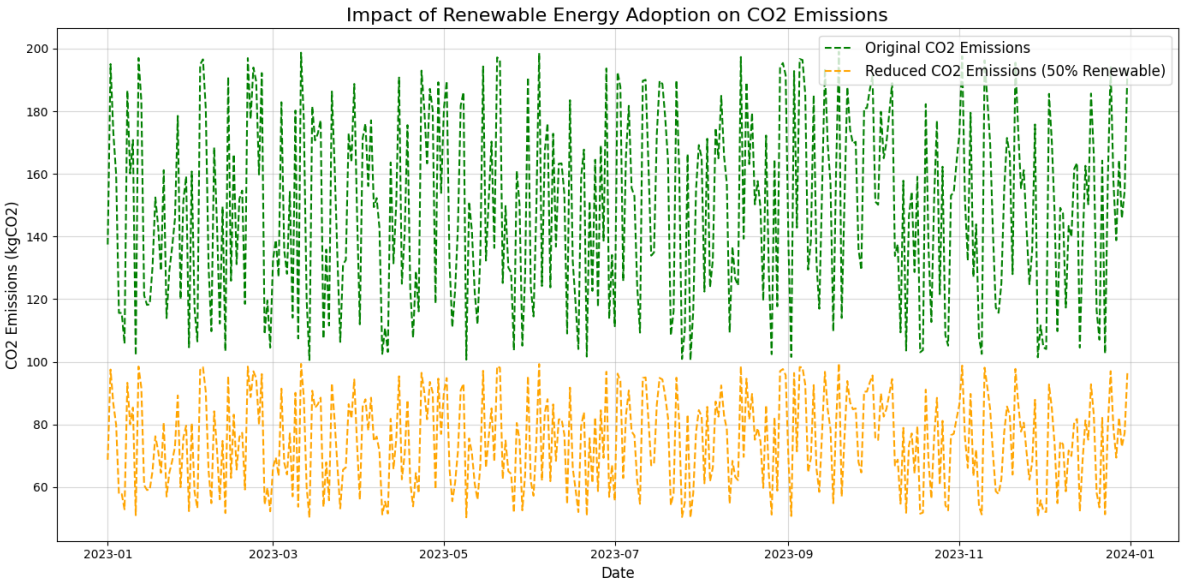


7. Impact of Renewable Energy Adoption

This chart illustrates:

- **Original CO_2 Emissions (Green Dashed Line)**: Based on traditional energy sources.
- **Reduced CO_2 Emissions (Orange Dashed Line)**: Assuming 50% renewable energy adoption.

Key Insights: Renewable energy adoption significantly reduces CO_2 emissions, highlighting its importance for sustainable development.

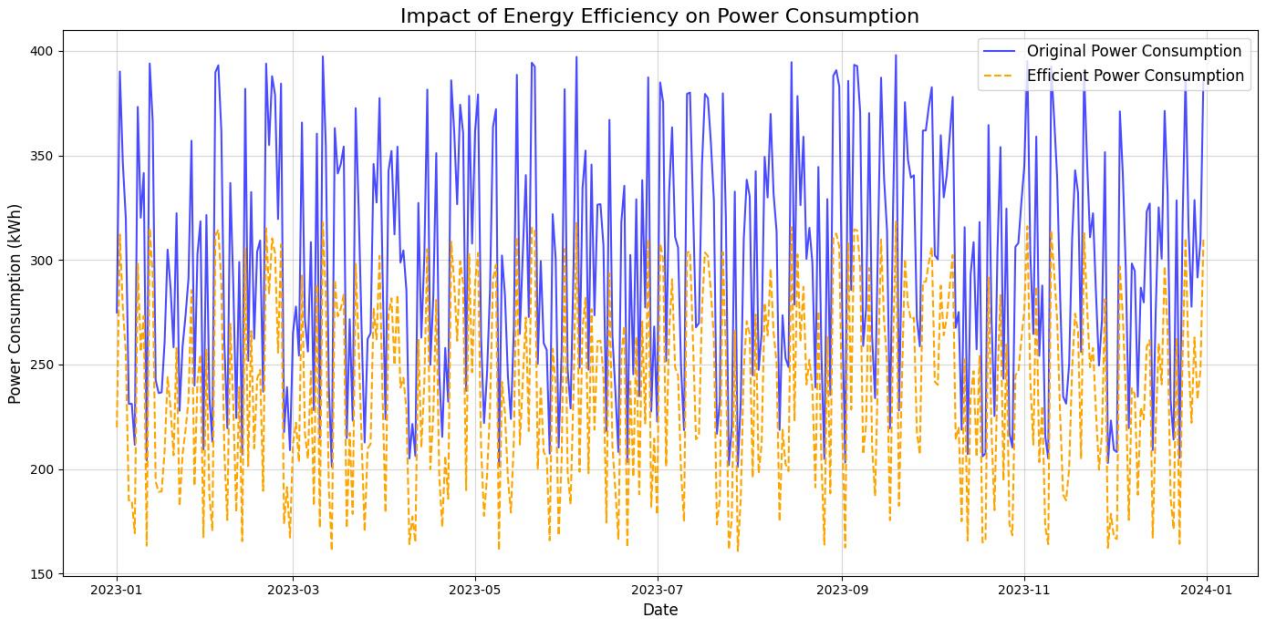


8. Energy Efficiency Impact

This chart compares:

- **Original Power Consumption (Blue Line)**: Historical energy usage.
- **Efficient Power Consumption (Orange Dashed Line)**: After implementing energy-saving measures.

Key Insights: Improving energy efficiency results in substantial energy savings, reducing both costs and environmental impact.



Optimizing Energy Consumption and Reducing CO_2 Emissions through Renewable Integration and Efficiency Measures

1. CO_2 Emissions (Direct Relationship with Power Consumption)

- **Calculation**:
  data["CO2_Emissions"] = data["Power_Consumption"] * 0.5

- The CO_2 emissions are calculated using a factor of **0.5 kgCO_2 per kWh**, representing the average carbon emissions for electricity generation from non-renewable sources like coal or natural gas.
- **Interpretation**:
  - **Higher Power Consumption → Higher CO_2 Emissions**: For every kilowatt-hour of electricity consumed, half a kilogram of CO_2 is released into the atmosphere.
  - This relationship is linear, as shown in the graphs, where peaks in power consumption correspond to peaks in CO_2 emissions.

- **Visualization**: The "Power Consumption and CO_2 Emissions Over Time" graph illustrates this direct relationship. Both lines follow a similar trend, demonstrating how energy usage directly impacts the environment.

2. Impact of Renewable Energy Adoption

- **Calculation**:

```
reduced_emissions = data["CO2_Emissions"] * (1 - renewable_energy_share)
```

- The reduction in CO_2 emissions assumes a certain percentage of power is replaced with renewable energy (e.g., 50% renewable energy adoption).
- **Interpretation**:
  - Renewable energy sources like solar and wind produce minimal or no CO_2 emissions.
  - By replacing 50% of energy consumption with renewables, emissions are halved for that portion.
- **Visualization**: The "Impact of Renewable Energy Adoption on CO_2 Emissions" chart shows:
  - **Green Dashed Line**: Original emissions based on traditional energy sources.
  - **Orange Dashed Line**: Reduced emissions after partial renewable adoption.
- **Impact**:
  - **CO_2 Savings**: By transitioning to renewables, the total CO_2 saved is **10,888.29 kgCO_2** for the dataset.
  - This highlights the significant environmental benefit of renewable energy integration.

3. Energy Efficiency Savings

- **Calculation**:
  efficient_power = data["Power_Consumption"] * (1 - efficiency_improvement)

- Energy savings are calculated based on improving efficiency (e.g., a 20% reduction in power usage).

- **Interpretation**:
  - Energy efficiency measures, such as better insulation, optimized machinery, or energy management systems, can reduce overall consumption without compromising output.

    - For every unit of energy saved, there is a proportional reduction in CO_2 emissions.
- **Visualization**: The "Impact of Energy Efficiency on Power Consumption" chart compares:
    - **Blue Line**: Original power consumption.
    - **Orange Dashed Line**: Reduced consumption after implementing efficiency improvements.
- **Impact**:
    - **Energy Saved**: A total of **21,776.59 kWh** saved through efficiency improvements.
    - **CO_2 Savings**: This corresponds to a **reduction of 10,888.29 kgCO_2**, as less power generation is required.

4. Rolling Statistics Analysis (Energy Patterns)

- **Calculation**: Rolling means and standard deviations were used to analyze trends and variability:

```
rolling_mean = data["Power_Consumption"].rolling(window=30).mean()
rolling_std = data["Power_Consumption"].rolling(window=30).std()
```

- **Interpretation**:
    - **Rolling Mean**: Indicates long-term energy consumption trends, smoothing out short-term fluctuations.
    - **Rolling Standard Deviation**: Shows variability over time, highlighting periods of instability or unusual patterns.
- **Visualization**: The "Power Consumption - Rolling Mean & Standard Deviation" chart shows:
    - **Red Line**: Smoothed trend, useful for understanding baseline consumption.
    - **Black Dashed Line**: Standard deviation, representing periods of stability or spikes.
- **Impact**:
    - Helps in identifying areas where efficiency can be improved (e.g., reducing peaks to flatten energy demand).

5. Clustering Power Consumption

- **Calculation**: Clusters were generated using k-means to categorize consumption patterns:

```
kmeans = KMeans(n_clusters=3)
labels = kmeans.fit_predict(data[["Power_Consumption"]])
```

- **Interpretation**:
  - **Cluster 0**: High power consumption, often associated with peak load periods.
  - **Cluster 1**: Moderate consumption, representing standard operational levels.
  - **Cluster 2**: Low consumption, often during off-peak hours or efficient periods.
- **Visualization**: The "Power Consumption Clusters" chart segments the data into these categories.
- **Impact**:
  - Provides actionable insights for managing energy demand, like shifting operations to off-peak hours (Cluster 2).

6. Forecasting Power Consumption

- **Calculation**: Seasonal and trend decomposition was used to predict future consumption:

```
decomposition = seasonal_decompose(data["Power_Consumption"], model="additive")
forecast = linear_regression_model.predict(future_dates)
```

- **Interpretation**:
  - **Seasonal Component**: Captures periodic variations, such as daily or seasonal cycles.
  - **Trend Component**: Reflects the overall growth or decline in energy usage.
  - **Residuals**: Irregular fluctuations not explained by trend or seasonality.
- **Visualization**: The "Power Consumption Forecast Using Linear Regression with Seasonality" graph shows:
  - **Blue Line**: Historical data.
  - **Orange Line**: Predicted future consumption.

- **Impact**:
  - Helps in proactive energy planning, ensuring resources are available during peak periods.

XV. Recommendations:

1. **Adopt Renewable Energy**:
    a. Increase the share of renewables in the energy mix to reduce CO_2 emissions.
    b. Invest in solar, wind, or hydropower solutions.
2. **Implement Energy Efficiency Measures**:
    a. Optimize operational processes to lower energy usage.
    b. Use energy-efficient appliances and equipment.
3. **Monitor and Analyze Trends**:
    a. Continuously track power consumption and emissions data.
    b. Use predictive models to anticipate and mitigate energy spikes.
4. **Educate Stakeholders**:
    a. Promote awareness about energy conservation and sustainability practices.

XVI. Conclusion and Scope

1. **Energy Savings**:
    a. Total energy saved: **21,776.59 kWh**, demonstrating the benefits of efficiency measures.
2. **CO_2 Emission Reduction**:
    a. Total CO_2 saved: **10,888.29 kgCO_2**, highlighting the positive environmental impact.
3. **Key Takeaways**:
    a. Adopting renewable energy and energy efficiency measures significantly reduces environmental impact.
    b. Monitoring and analyzing trends ensures better resource planning and operational efficiency.
    c. Clustering and forecasting provide actionable insights for strategic energy management.

This comprehensive analysis serves as a roadmap for sustainable energy practices, ensuring both economic and environmental benefits.

Conclusion

This analysis emphasizes the importance of integrating renewable energy and energy efficiency measures to achieve sustainable operations. These practices reduce environmental impact while improving operational efficiency and resource management. Employing methods such as trend

monitoring, clustering, and forecasting enables data-driven decision-making, paving the way for strategic energy management and long-term sustainability.

Future Scope

Future efforts can focus on expanding the use of advanced analytics and AI-driven tools to enhance energy management strategies further. Additionally, exploring emerging renewable technologies and optimizing energy storage systems can amplify environmental and economic benefits. Continued monitoring and adaptation of these practices will ensure organizations remain aligned with evolving sustainability goals and market demands.