ABRILIAN
MAULIDHIA

PINION - JUNIOR DATA ANALYST

TECHNICAL TEST

brilianmaulidhia@gmail.com

Hello,
I’m Brili!

I’m a learner and enthusiast of data with a professional certification as a TensorFlow Developer. Interested in data and new things in technology and try to grow with any experiences.

https://www.linkedin.com/in/abrilianmaulidhia/

Bachelor of Computer Science

International Student
Exchange 2023 Awardee of Universiti
Tun Hussein Onn Malaysia (UTHM)

CONTACT

Malang, Indonesia
brilianmaulidhia@gmail.com
(+62)82233980235

ABRILIAN
MAULIDHIA

INTRODUCTION OF THE CASE

ABC is a building management company responsible for overseeing and maintaining a wide range of assets across Indonesia, with the majority of its portfolio concentrated in Java. Their portfolio includes various types of buildings, ranging from commercial and residential complexes to industrial facilities. The company's operations involve ensuring that these buildings are running efficiently, providing a safe and comfortable environment for occupants, and maintaining the value of the properties. A crucial aspect of this management is the ability to optimize operational expenses, as these costs directly impact profitability and long-term sustainability. One of the most significant contributors to these operational costs is energy consumption, making energy efficiency a top priority for ABC to ensure both cost-effectiveness and environmental responsibility.

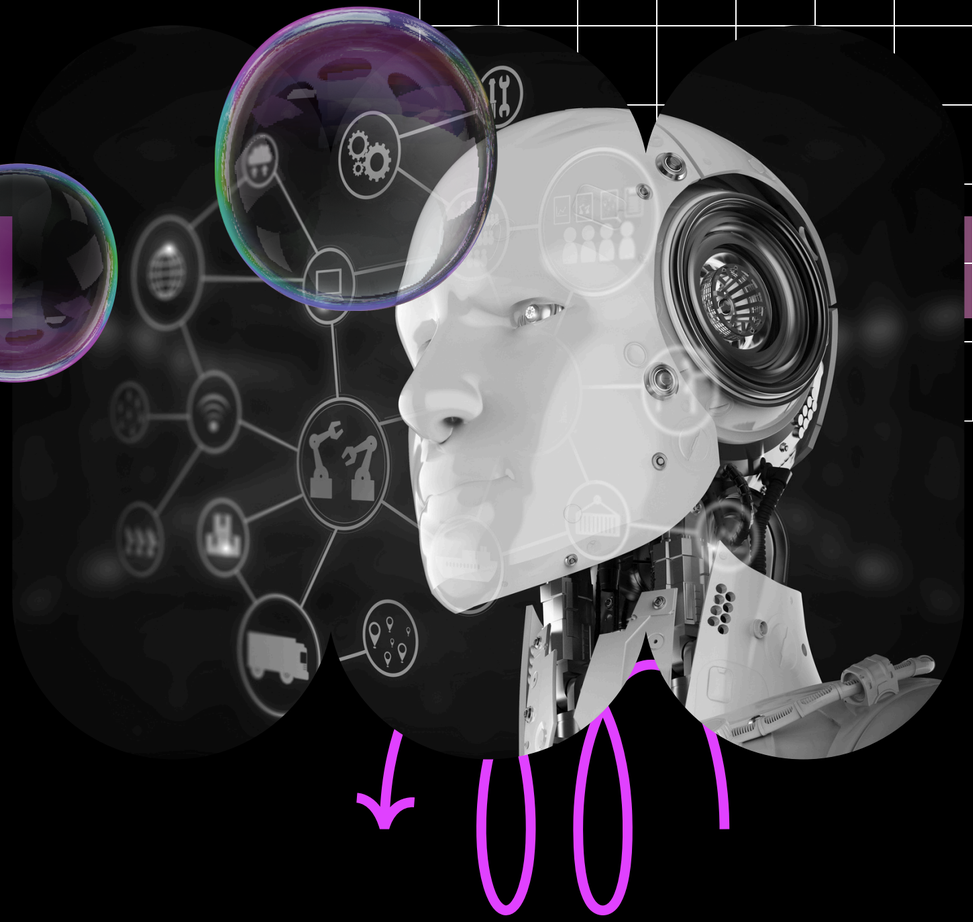
To achieve greater energy efficiency, ABC is committed to making data-driven decisions that canoptimize energy usage across its managed properties. Recognizing the importance of accurate data in understanding energy consumption patterns, ABC has taken the first step by gathering relevant data from its various assets. This data includes metrics such as energy usage rates, peak consumption times, equipment efficiency, and more. The next critical phase involves a detailed analysis of this data to identify areas for improvement, detect inefficiencies, and develop actionable strategies for reducing energy consumption. By leveraging data analysis, ABC aims to implement informed solutions that will not only lower energy costs but also contribute to more sustainable building operations, aligning with both financial and environmental goals.

VARIABLE	DESCRIPTION
Building_ID	Unique identifier for the building
Year	Year of Data Collection
Building_Type	Type of Building (e.g. Residential, Commercial, Industrial)
Floor_Area	Total floor area of the building
Energy_Consumption	Total energy consumption for the building in kilowatt-hours, consists of energy used for AC and lighting within the building.
Energy_Cost	Total cost of energy consumption for the year.
Occupancy_Rate	Percentage of building occupancy.
Region	The geographic region of the building (e.g., Urban, Suburban Rural).
Energy_Score	A calculated score rating the energy efficiency of the building (1 to 100).
Province	The location of the building
Average_Temp	The average temperature of the building which affect the energy consumption
Energy_Consumption_AC	Total energy consumption for Air-conditioning within the building.
Energy_Consumption_Lighting	Total energy consumption for lighting within the building.

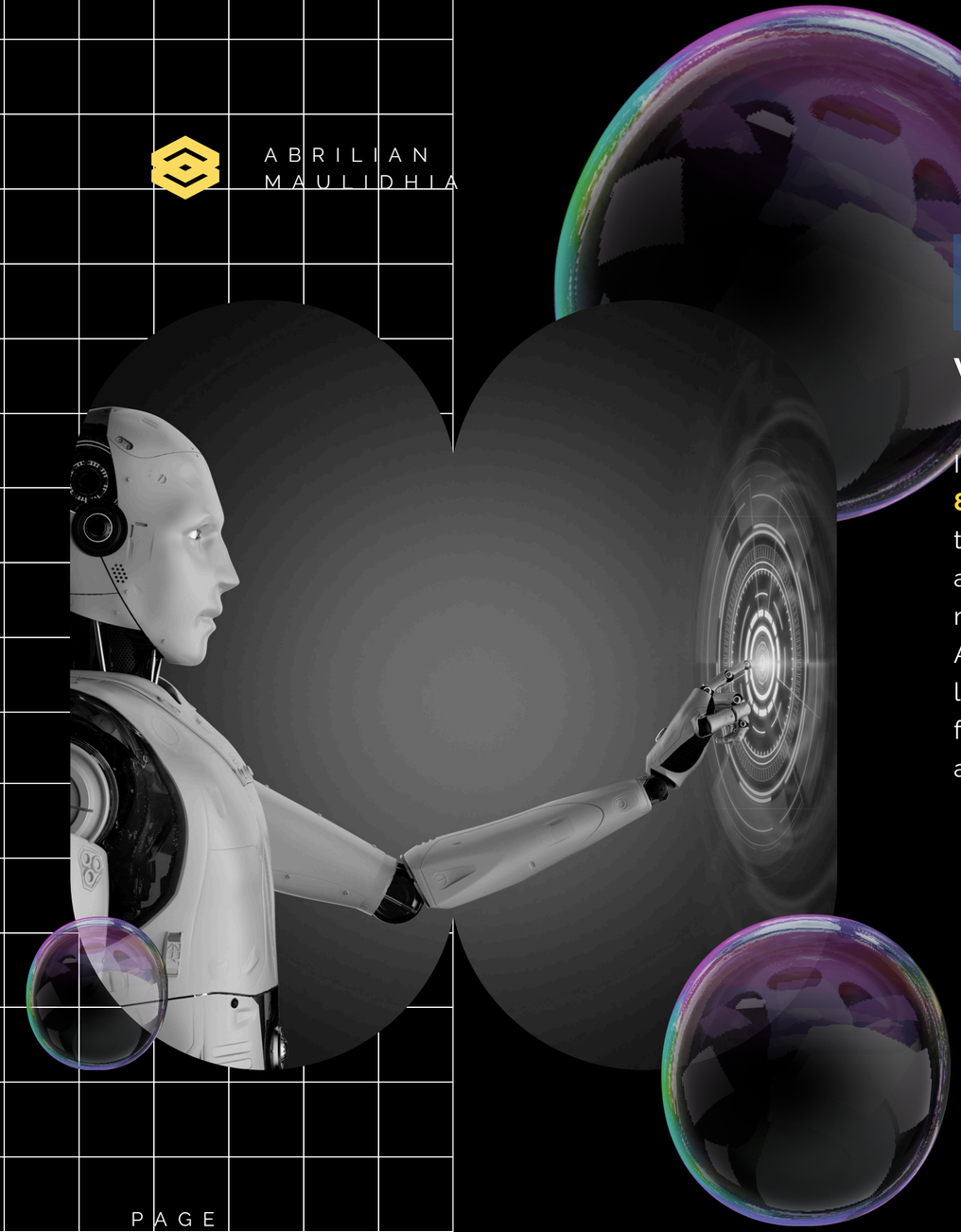
ABRILIAN
MAULIDHIA

#1

We have identified some **missing values** and **outliers** in the dataset. Before proceeding with the analysis, it's essential to ensure the dataset is properly refined. Please create and execute a Python script to preprocess and refine the data accordingly. (Please keep this script with you for the interview session).



ABRILIAN
MAULIDHIA



#1A

What are the missing data points?

In this step, I used **.isnull()** function to identification that the **Occupancy_Rate (%)** column had **8 missing data points**, while all other columns had no missing values. To address this, I filled the missing values in the Occupancy_Rate (%) column with the mean of that column. This approach is effective for maintaining data integrity, especially when the missing data is minimal and the distribution of the data is relatively even.
After filling in the missing values, I rechecked the dataset and confirmed that there are no longer any missing values in any of the columns. With this, the dataset is now fully prepared for further analysis, ensuring that the conclusions drawn will be based on complete and accurate data.

```
Summary of Missing Data Points:
Building_ID                      0
Year                             0
Building_Type                    0
Floor_Area (m²)                  0
Energy_Consumption (kWh)         0
Occupancy_Rate (%)               8
Region                           0
Energy_Score                     0
 Energy_Cost (IDR)               0
Province                         0
Average_Temp                     0
Energy_Consumption_AC (kWh)      0
Energy_Consumption_Lighting (kWh)  0
dtype: int64
```

```
Summary of Missing Data Points:
Building_ID                      0
Year                             0
Building_Type                    0
Floor_Area (m²)                  0
Energy_Consumption (kWh)         0
Occupancy_Rate (%)               0
Region                           0
Energy_Score                     0
 Energy_Cost (IDR)               0
Province                         0
Average_Temp                     0
Energy_Consumption_AC (kWh)      0
Energy_Consumption_Lighting (kWh)  0
dtype: int64
```

ABRILIAN
MAULIDHIA

```
The Number of Duplication is: 0
```

```
Rows with negative values where they shouldn't exist:
Empty DataFrame
Columns: [Building_ID, Year, Building_Type, Floor_Area
Index: []
```

#1B

What is the data that needs to be tidied up?

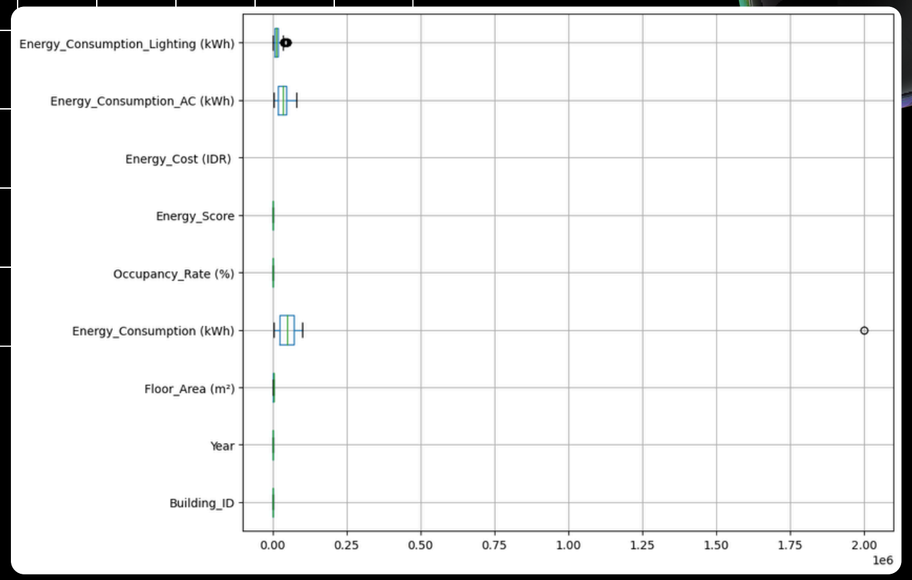
In this step, I identified and addressed potential data issues to ensure the dataset's quality. The key actions taken are as follows,

1. **Missing Data,** Initially, the Occupancy_Rate (%) column had 8 missing values. These were **filled with the mean** value of the column, effectively handling the missing data and ensuring no gaps remained in the dataset.
2. **Duplicate Data,** I checked for duplicated entries in the dataset and confirmed that there were **no duplicates**. This indicates that each record in the dataset is unique, which is essential for accurate analysis.
3. **Negative Values,** Negative values in the **Floor_Area (m^2)** column were **converted to positive values** using the absolute function, ensuring all floor area measurements are valid.
4. **Inconsistent Data,** To address any potential inconsistencies, I **converted** relevant columns (Energy_Consumption (kWh), Energy_Cost (IDR), Energy_Consumption_AC (kWh), and Energy_Consumption_Lighting (kWh)) **to numeric types**. This step ensured that any non-numeric values, which could have been incorrectly entered as strings, were identified and corrected.
5. **Validation of Negative Values,** I then checked for any negative values in these numeric columns, as such values would be logically inconsistent (e.g., negative energy consumption or costs). The results showed that there were **no rows with negative values**, indicating that the data is consistent and accurate in terms of its numeric entries.

The dataset has been thoroughly **cleaned**, with missing values addressed, duplicates checked, and inconsistencies corrected. The absence of negative values confirms the data is now tidy and ready for further analysis.

#1C

Identify any outliers and suggest an approach for managing them



In this step, the outlier analysis was performed on several numeric columns in the dataset, specifically focusing on **Energy_Consumption (kWh)**, **Energy_Cost (IDR)**, **Energy_Consumption_AC (kWh)**, and **Energy_Consumption_Lighting (kWh)**. **Key Findings,**

1. **Energy Consumption (kWh),** Identified one significant outlier 2000000 kWh from Building ID 83. This value is much higher than the typical range, indicating a potential data entry error or an unusual event.
2. **Energy Cost (IDR),** No outliers were found, suggesting that the energy cost values fall within an expected range.
3. **Energy Consumption for AC (kWh),** No outliers were detected, indicating that the values are consistent across the dataset.
4. **Energy Consumption for Lighting (kWh),** Multiple outliers were identified, including values such as 41252 kWh and others. These values may also represent errors or exceptional cases.

Approach for Managing Outliers,

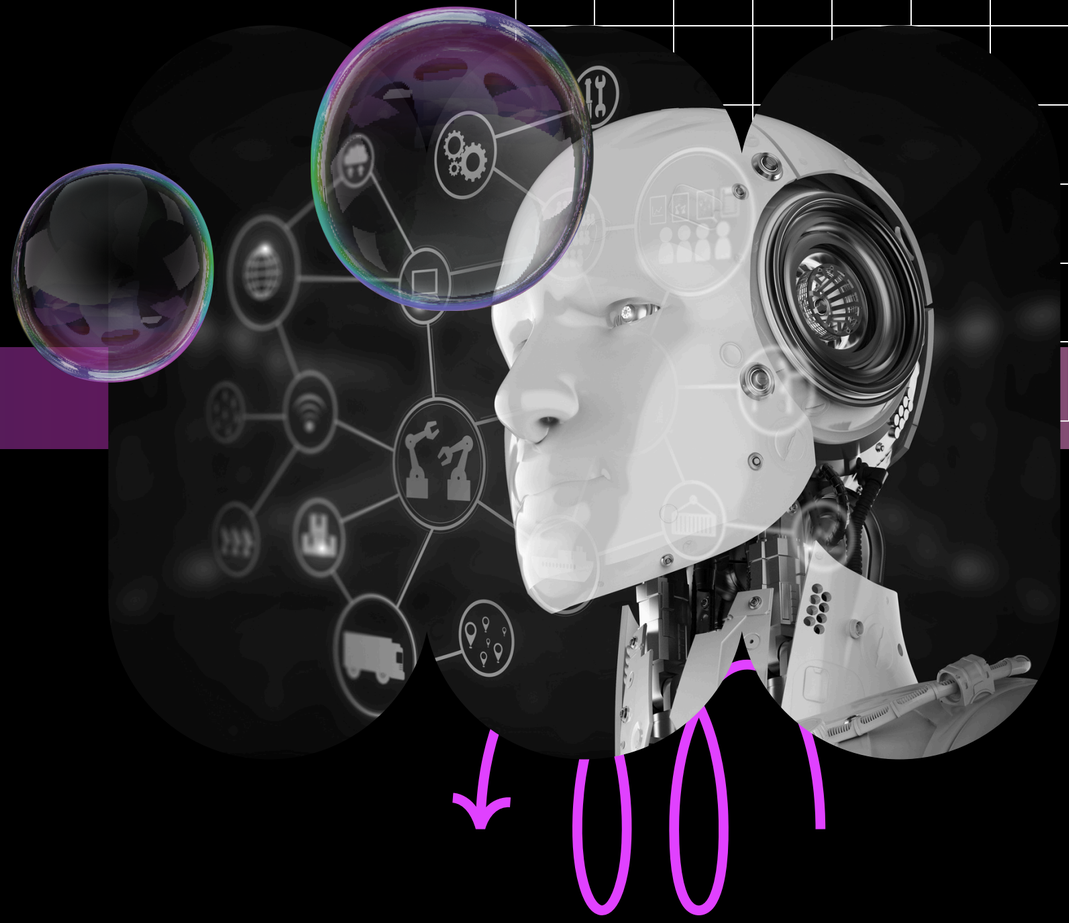
1. **Capping Outliers,** For the significant outlier in the Energy Consumption (kWh) column, it could be capped at the upper bound of the IQR to mitigate its impact on analyses. Similarly, for the Energy Consumption for Lighting, the outliers can be capped to reduce skewness in the dataset.
2. **Removal,** Outliers can be removed if they are deemed erroneous after investigation. The cleaned DataFrame was produced by removing identified outliers from the Energy Consumption for Lighting column.

By identifying and managing outliers, particularly through capping and removal, the dataset has been cleaned to ensure a more accurate and reliable analysis. The cleaned dataset now better represents the central tendencies of the data and minimizes the risk of outlier-driven skewness.

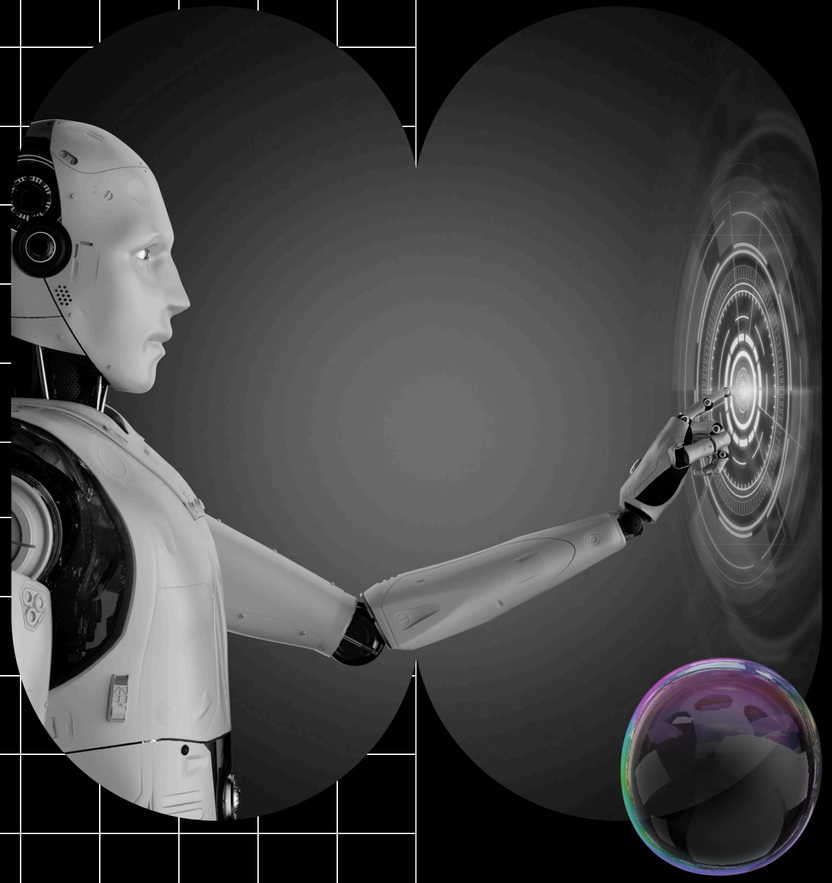
ABRILIAN
MAULIDHIA

#2

Perform **exploratory analysis** on the dataset to give insights on the following queries from ABC



ABRILIAN
MAULIDHIA



#2A

Which province has the highest energy consumption on average?

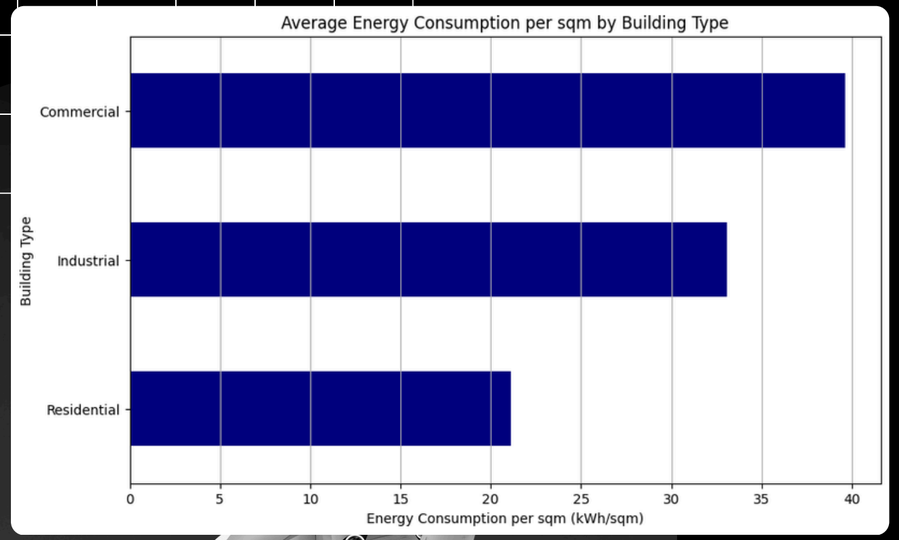
Analyzed the dataset to identify the province with the highest average energy consumption. After ensuring that there were no missing values in the relevant columns, I calculated the **average energy consumption for each province by grouping the data** and computing the mean. I then identified the province with the highest average energy consumption, which turned out to be **Jawa Timur**, with an average of **59,748.00 kWh**. This finding highlights where energy efficiency improvements and resource allocation might be most needed.

```
The province with the highest average energy
consumption is JawaTimur with an average of 59,748.00 kWh.
```

ABRILIAN
MAULIDHIA

#2B

Building_Type	mean	median	min	max	std
Commercial	39.612179	21.405702	3.619253	211.920635	51.434646
Industrial	33.079392	19.252840	1.621597	208.964710	45.401653
Residential	21.110022	13.874369	1.641427	227.900641	43.574752



Describe the energy consumption per sqm profile of each building type and compare them with one another.

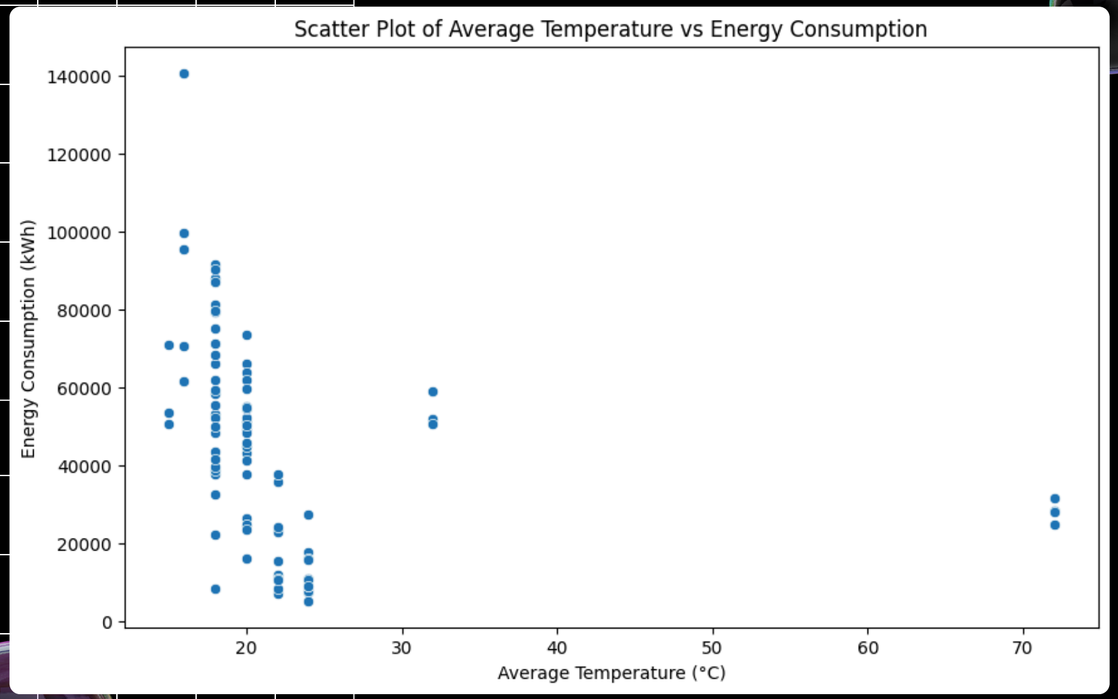
The analysis of energy consumption per square meter (sqm) across different building types yields the following insights,

1. Building Type Profiles

- **Commercial buildings** exhibit the **highest** average energy consumption per sqm at **39.61** kWh/sqm, with a median of 21.41 kWh/sqm. This suggests that commercial spaces have significant energy demands, likely due to higher operational hours and equipment use.
- **Industrial buildings** follow closely with an average of **33.08 kWh/sqm** and a median of 19.25 kWh/sqm. The broad range (from 1.62 kWh/sqm to 208.96 kWh/sqm) indicates variability in energy usage, potentially influenced by the type of industry and machinery.
- **Residential buildings** show the lowest average at **21.11 kWh/sqm**, with a median of 13.87 kWh/sqm. This reflects typical household energy consumption patterns, which are generally less intense compared to commercial and industrial sectors.

2. Comparison Insights

- **Commercial buildings** is the **highest average** energy consumption per sqm compared to others buildings, highlighting their higher operational energy needs.
- The **standard deviation** in energy consumption indicates significant **variability** in all building types, especially in commercial and industrial sectors, suggesting that some buildings may be outliers due to unique operational demands.




```
Correlation between Average Temperature and Energy Consumption: -0.28
```

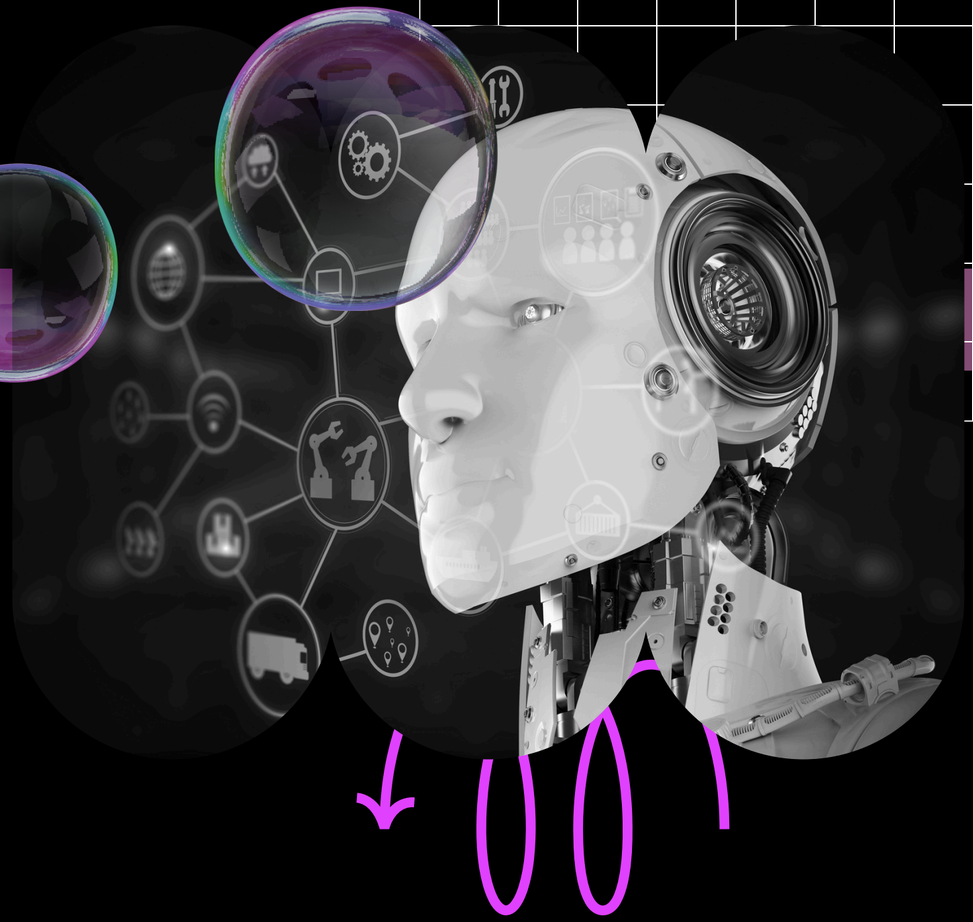
#2C

Is there any correlation between average temperature of a building and its energy consumption? If there is, how do they correlate with each other?

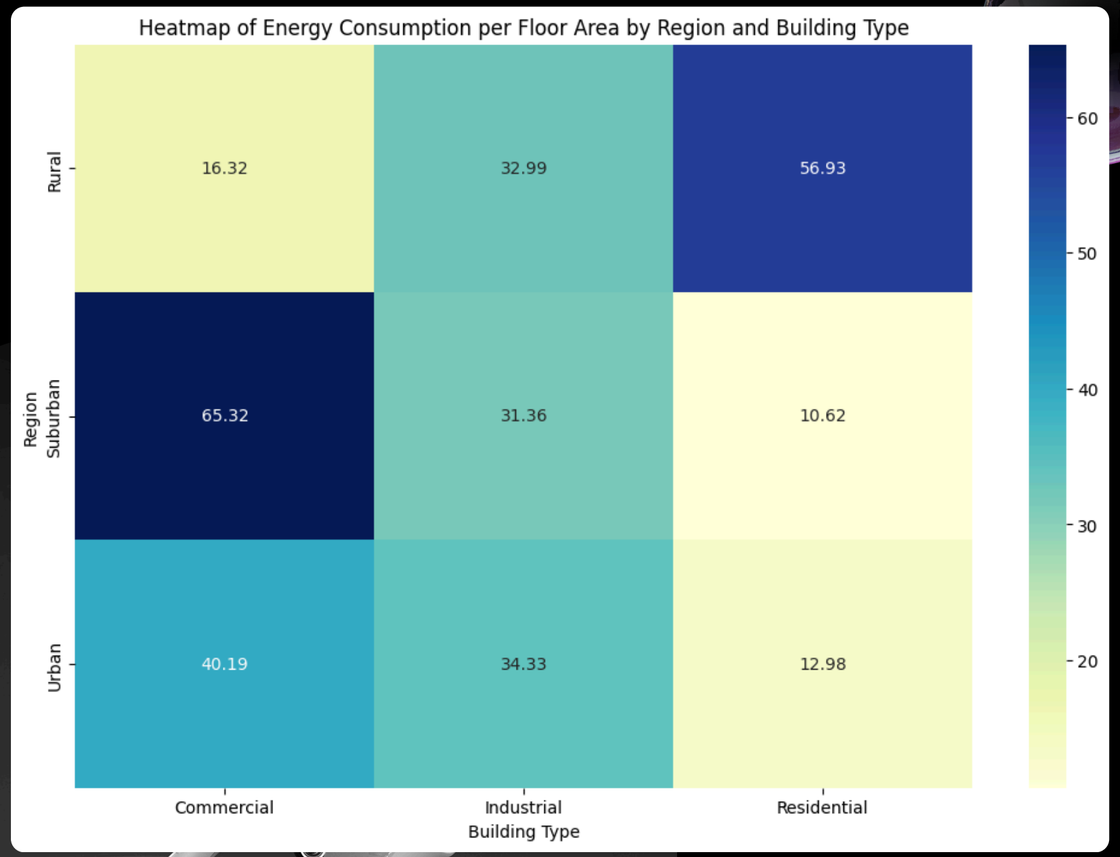
Based on the analysis, there is a negative correlation of **-0.28** between the **average temperature** of a building and its **energy consumption**. This indicates a **weak inverse relationship**, meaning that as the average temperature increases, the energy consumption slightly tends to decrease. However, the correlation is not strong enough to suggest a significant or direct relationship between these two variables. The scatter plot also visualizes this weak negative trend, showing that other factors might play a more prominent role in determining a building's energy consumption.

#3

ABC is **committed** to the **Green Future** initiative, which requires **energy consumption assessments** for the buildings under its management. To facilitate informed decision-making present a clear visualization of the following datasets.



ABRILIAN
MAULIDHIA



#3A

Heatmap (color coded matrix) of Energy Consumption per Floor Area by Region and Building Type. Highlight any patterns that stand out between different types of buildings and regions in terms of their energy consumption efficiency.

1. Energy Efficiency Focus

- **Urban residential buildings are the most energy-efficient**, while suburban residential buildings are the least. Efforts to improve energy efficiency should target suburban residential areas.

2. Commercial Energy Use

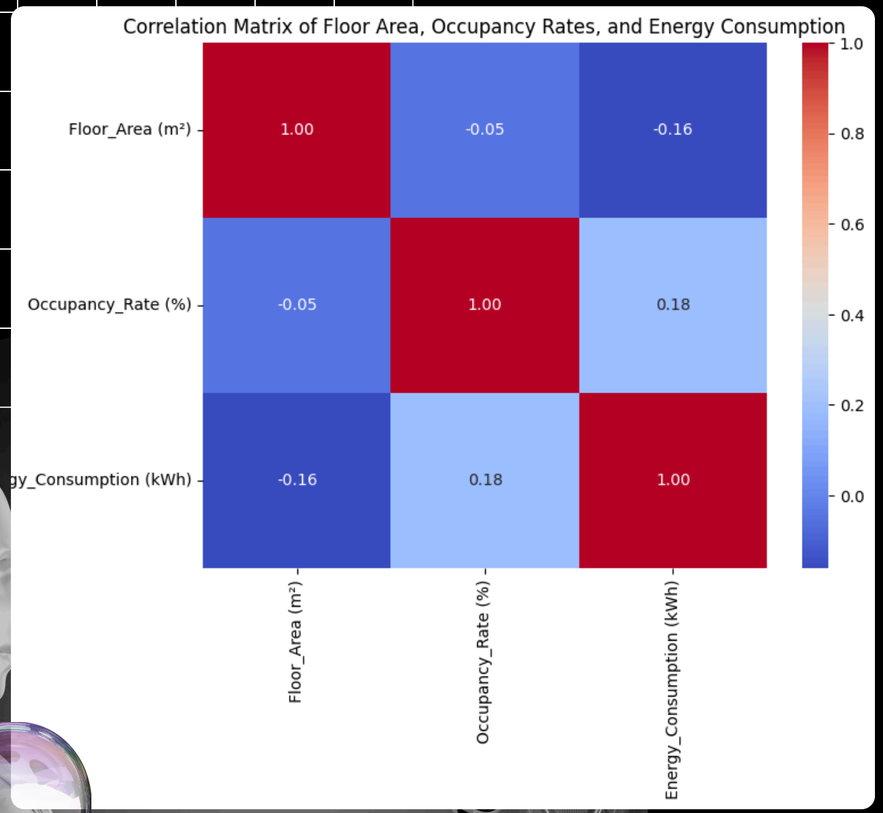
- **Rural commercial buildings have high energy consumption**, indicating a potential area for energy-saving improvements.

3. Industrial Consistency

- The **consistent energy consumption in industrial buildings** suggests that any improvements in energy efficiency could be applied uniformly across all regions.

ABRILIAN
MAULIDHIA

#3B



Plot the correlation of Floor Area, Occupancy Rates, and Energy Consumption. What insights would you provide to ABC based on the data presented in this visualization?

Based on the correlation matrix heatmap of Floor Area, Occupancy Rates, and Energy Consumption, here are the key insights,

1. Weak Correlations, The correlations between Floor Area, Occupancy Rates, and Energy Consumption are generally weak. This indicates that these variables do not strongly influence each other in this dataset.

- **Floor Area and Occupancy Rate,** Slightly negative correlation (-0.05).
- **Floor Area and Energy Consumption,** Negative correlation (-0.16).
- **Occupancy Rate and Energy Consumption,** Slight positive correlation (0.18).

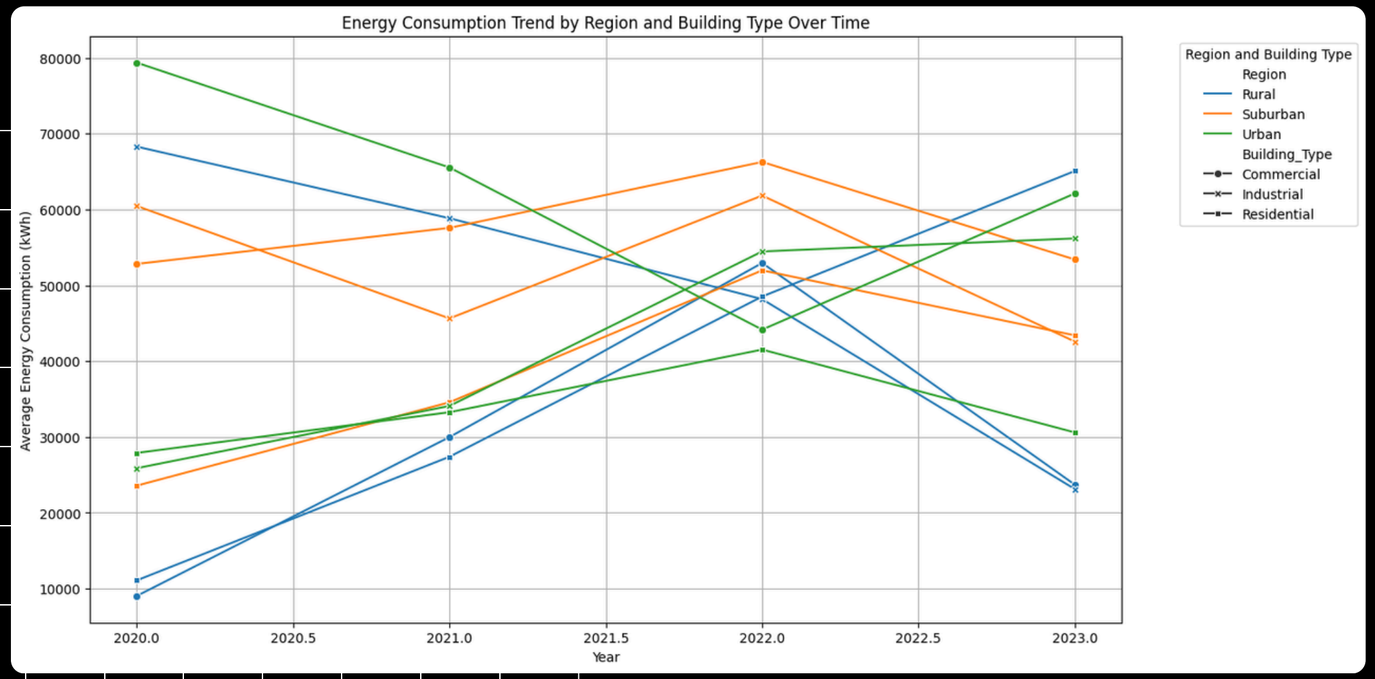
2. Energy Consumption Patterns,

- **Larger floor areas tend to consume slightly less energy**, which might suggest that larger spaces are more energy-efficient per unit area.
- **Higher occupancy rates are associated with a slight increase in energy** consumption, indicating that more people in a space lead to higher energy use.

The **weak correlations** suggest that Floor Area, Occupancy Rates, and Energy Consumption are **not strongly interdependent**. Therefore, we should consider other factors that might have a more significant impact on Energy Consumption, such as building design, insulation, and the efficiency of appliances and systems. Additionally, focusing on optimizing energy use based on occupancy patterns could provide some benefits, but it may not be the primary driver of our energy efficiency improvements.

#3C

Energy Consumption Trend by Region and Building Type Over Time. Elaborate on the pattern of different types of energy consumption against building type and region.



Based on graph, reveals several key patterns,

1. Urban Residential Buildings

- There is a noticeable increase in energy consumption over time, suggesting growing energy demands possibly due to urbanization and increased population density.

2. Suburban Commercial Buildings

- Energy consumption shows a steady rise, indicating expanding commercial activities and possibly less efficient energy use.

3. Rural Industrial Buildings

- A decreasing trend in energy consumption is observed, which might be due to improvements in energy efficiency or a decline in industrial activities in rural areas.

4. Other Trends

- The remaining building types and regions show relatively stable or slightly fluctuating energy consumption patterns, indicating consistent energy use without significant changes.

Urban residential areas are **experiencing increasing energy demands**, while suburban commercial areas also show rising consumption. In contrast, rural industrial areas are becoming more energy-efficient or seeing reduced activity. These trends highlight the need for targeted energy efficiency measures in urban and suburban regions to manage growing energy demands effectively.

#4

Share your thoughts and experiences from working on the exercise above. For example, you could talk about any challenges you faced, like dealing with missing data or outliers, and how you tackled them. We'd love to hear how it went for you!

During this project, I encountered significant challenges with **handling outliers**, particularly in energy consumption data. Initially, I visualized potential outliers using boxplots for numeric columns, which provided a clear view of data distribution and highlighted extreme values. To systematically identify outliers, I implemented the Interquartile Range (IQR) method, calculating the first (Q1) and third (Q3) quartiles to define lower and upper bounds. This approach helped me pinpoint values that fell outside the typical range.

After identifying outliers for key columns, such as energy consumption and energy costs, I decided to **cap extreme values** instead of removing them entirely to retain data integrity. This involved replacing outliers with the calculated bounds, which ensured the dataset remained useful for analysis while mitigating the influence of extreme values. Ultimately, this process of identifying and managing outliers enhanced the reliability of my analyses and allowed me to draw more accurate insights from the data.

THANK YOU!

please click here for more details on gcolab