Summary Report for Building Energy Benchmarking Data

Part 1: Data Cleaning and Preprocessing

1.1 Load and Inspect the Dataset

• Load the dataset and display its shape, column names, and data types.
• Identify and list the number of missing values in each column.

Shape, Columns and Data types of Building Emergency Benchmarking

---

1. Shape of the Dataset: (494, 31)
2. Column Names and Data types of each column is given below:

Column Name	Data Type
Property Id	int64
Property Name	object
Address 1	object
City	object
Postal Code	Object
Province	Object
Primary Property Type - Self Selected	object
Number of Buildings	int64
Year Built	int64
Property GFA - Self-Reported (m^2)	Object
ENERGY STAR Score	float64
Site Energy Use (GJ)	Object
Weather Normalized Site Energy Use (GJ)	Object
Site EUI (GJ/m^2)	float64
Weather Normalized Site EUI (GJ/m^2)	float64
Source Energy Use (GJ)	Object
Weather Normalized Source Energy Use (GJ)	Object
Source EUI (GJ/m^2)	float64
Weather Normalized Source EUI (GJ/m^2)	float64
Total GHG Emissions (Metric Tons CO2e)	Object
Total GHG Emissions Intensity ($kgCO2e/m^2$)	float64
Direct GHG Emissions (Metric Tons CO2e)	object
Direct GHG Emissions Intensity ($kgCO2e/m^2$)	float64
Electricity Use - Grid Purchase (kWh)	object
Natural Gas Use (GJ)	object
District Hot Water Use (GJ)	Object
Electricity Use – Generated from Onsite Renewable Systems (kWh)	float64
Green Power - Onsite and Offsite (kWh)	float64
Avoided Emissions - Onsite and Offsite Green Power (Metric Tons CO2e)	float64
Year Ending	int64
Unique ID	object

The Building Emergency Benchmarking has 494 rows and 31 columns in which PropertyId is the indexing colu mn.

Missing Count per column:

	Missing Count	Missing Percentage
Property Id	0	0
Property Name	0	0
Address 1	0	0
City	0	0
Postal Code	0	0
Province	0	0
Primary Property Type – Self Selected	0	0
Number of Buildings	0	0
Year Built	0	0
Property GFA – Self-Reported (m²)	0	0
ENERGY STAR Score	329	66.6
Site Energy Use (GJ)	0	0
Weather Normalized Site Energy Use (GJ)	0	0
Site EUI (GJ/m²)	0	0
Weather Normalized Site EUI (GJ/m²)	0	0
Source Energy Use (GJ)	0	0
Weather Normalized Source Energy Use (GJ)	2	0.4
Source EUI (GJ/m²)	0	0
Weather Normalized Source EUI (GJ/m²)	0	0
Total GHG Emissions (Metric Tons CO2e)	0	0
Total GHG Emissions Intensity (kgCO2e/m²)	0	0
Direct GHG Emissions (Metric Tons CO2e)	0	0
Direct GHG Emissions Intensity (kgCO2e/m²)	0	0
Electricity Use – Grid Purchase (kWh)	0	0
Natural Gas Use (GJ)	10	2.02
District Hot Water Use (GJ)	479	96.96
Electricity Use – Generated from Onsite Renewable Systems (kWh)	450	91.09
Green Power – Onsite and Offsite (kWh)	198	40.08
Avoided Emissions – Onsite and Offsite Green Power (Metric Tons CO2e)	198	40.08
Year Ending	0	0
Unique ID	0	0

1.2 Handling Missing Data

• Drop columns with more than 40% missing values.
• For numerical columns, fill missing values with the median of their respective column.
• For categorical columns, fill missing values with the mode of their respective column.

Column names with missing percentage more than 40% missing values are:

['ENERGY STAR Score',
'District Hot Water Use (GJ)',
'Electricity Use – Generated from Onsite Renewable Systems (kWh)',

'Green Power - Onsite and Offsite (kWh)',
'Avoided Emissions - Onsite and Offsite Green Power (Metric Tons CO2e)']

Columns with more than 40% are dropped. Below are the columns remaining:

Missing Count per column after dropping columns:

	Missing Count	Missing Percentage
Property Id	0	0
Property Name	0	0
Address 1	0	0
City	0	0
Postal Code	0	0
Province	0	0
Primary Property Type – Self Selected	0	0
Number of Buildings	0	0
Year Built	0	0
Property GFA – Self-Reported (m²)	0	0
Site Energy Use (GJ)	0	0
Weather Normalized Site Energy Use (GJ)	0	0
Site EUI (GJ/m²)	0	0
Weather Normalized Site EUI (GJ/m²)	0	0
Source Energy Use (GJ)	0	0
Weather Normalized Source Energy Use (GJ)	2	0.4
Source EUI (GJ/m²)	0	0
Weather Normalized Source EUI (GJ/m²)	0	0
Total GHG Emissions (Metric Tons CO2e)	0	0
Total GHG Emissions Intensity (kgCO2e/m²)	0	0
Direct GHG Emissions (Metric Tons CO2e)	0	0
Direct GHG Emissions Intensity (kgCO2e/m²)	0	0
Electricity Use – Grid Purchase (kWh)	0	0
Natural Gas Use (GJ)	10	2.02
Year Ending	0	0
Unique ID	0	0

Numerical Columns with Median are:

	Numerical Columns	Median
0	Property Id	9.99779e+06
1	Number of Buildings	1
2	Year Built	1978
3	Site EUI (GJ/m²)	1.29
4	Weather Normalized Site EUI (GJ/m²)	1.31
5	Source EUI (GJ/m²)	1.68
6	Weather Normalized Source EUI (GJ/m²)	1.69
7	Total GHG Emissions Intensity (kgCO2e/m²)	117.6
8	Direct GHG Emissions Intensity (kgCO2e/m²)	43.9
9	Year Ending	2021

Categorical Columns with Mode are:

	Categorical Columns	Mode
0	Property Name	Acadia Aquatic & Fitness Centre
1	Address 1	1001 BARLOW TR SE
2	City	Calgary
3	Postal Code	T2G 4K8
4	Province	Alberta
5	Primary Property Type - Self Selected	Fire Station
6	Property GFA - Self-Reported (m²)	1,108.10
7	Site Energy Use (GJ)	1122
8	Weather Normalized Site Energy Use (GJ)	1150
9	Source Energy Use (GJ)	1,010.10
10	Weather Normalized Source Energy Use (GJ)	1012
11	Total GHG Emissions (Metric Tons CO2e)	44
12	Direct GHG Emissions (Metric Tons CO2e)	0
13	Electricity Use - Grid Purchase (kWh)	93,572.40
14	Natural Gas Use (GJ)	1034
15	Unique ID	2019-10002717

1.3 Extracting and Cleaning Data Using Regex

• Use Regex only to:

- Extract numeric values from text-based numeric columns (e.g., Property GFA,
- Energy Use, Emissions).
- Standardize Postal Codes to follow the Canadian format (A1A 1A1).
- Clean and extract meaningful text from Property Names and Addresses.
- Ensure extracted values are properly converted to numerical types for analysis.

Extracted numerical values for the below columns:

['Property GFA - Self-Reported (m^2)',
'Site Energy Use (GJ)',
'Weather Normalized Site Energy Use (GJ)',
'Source Energy Use (GJ)',
'Weather Normalized Source Energy Use (GJ)',
'Total GHG Emissions (Metric Tons CO2e)',
'Direct GHG Emissions (Metric Tons CO2e)',
'Electricity Use - Grid Purchase (kWh)',
'Natural Gas Use (GJ)']

pattern = r'^-?\d+(\.\d+)?$'

Where r means it's a regular expression
^ means pattern must match from the beginning of the string
-? Means the preceding character (-) of ? can occur zero or one times
\d represents digits from 0 to 9
\d+ means digit can occur one or more times
(\.\d+) is an optional match with one or more digits followed by period or dot
    \. Indicated a dot, here \ is used as an escape for dot
    \d represents digits from 0 to 9
    \d+ means digit can occur one or more times
(\.\d+)? Means the preceding characters indicated in () is optional
$ indicates the end of the string

In brief the pattern can be said that it can be a positive or negative digit/(s) followed by an optional decimal part of digit/(s). The entire pattern must be matched from ^ to $.

Standardize Postal Codes to follow the Canadian format (A1A 1A1)

The Canada Post code is in the format of letter, digit, letter, digit, letter, digit. Pattern used to match is pattern = r'^[A-Za-z]\d[A-Za-z][ ]?\d[A-Za-z]\d$'.
Where r means regular expression
^[A-Za-z] means n upper case or lower case letter in the beginning
\d means second is a digit
[]? Means an optional space

In the CSV given we have some of the post codes are not in this format in the given dataset. Examples of such are:

Property ID	Postal code
9481172	T2AoK9
9481172	T2A OK9
9481172	T2A OK9

For the this I have converted to it "T2A 0K9" based on the condition df[df['Property Id'] == 9481172]. Since all the post code for this property has the same postcode.

Clean and extract meaningful text from Property Names and Addresses.

For these the leading and trailing spaces were removed.
The following changes were also made for Address:

Was	Changed to	Pattern used
St/st/ST	Street	r'\b([Ss][Tt])\b'
Av/AV/av/Ave	Avenue	r'\b(AV|av|Av|aV|AVE|ave)\b'
Rd/rd/RD	Road	r'\b([Rr][Dd])\b'
Dr/dr/DR	Drive	r'\b([Dd][Rr])\b'

Part 2: Exploratory Data Analysis (EDA) and Aggregations

2.1 Statistical Summary

The mean and median are measures of central tendency that describe the center of the data distribution. Mean gives the average of all data points. Median is the middle value when the data points are sorted. Standard deviation and variance measure the spread or dispersion of the data or it is the average distance of each data point from the mean. Interpretation of data distribution based on mean and median can be explained as:

- Symmetrical Distribution: Mean and median would be close to each other.
- Right-Skewed Distribution: mean > median, where a few large values pull the mean to the right.
- Left-Skewed Distribution: mean<median, where a few small values pull the mean to the left.

	Column Name	Mean	Median	Distribution	Standard Deviation
0	Number of Buildings	1.06	1.00	Symmetrical	0.278281
1	Property GFA - Self-Reported (m^2)	4752.56	1806.75	Right-skewed	10128.320688
2	Site Energy Use (GJ)	8265.67	2555.65	Right-skewed	19733.748811
3	Weather Normalized Site Energy Use (GJ)	8397.19	2572.50	Right-skewed	19877.336939
4	Site EUI (GJ/m^2)	1.77	1.29	Symmetrical	1.306700
5	Weather Normalized Site EUI (GJ/m^2)	1.81	1.31	Symmetrical	1.331529
6	Source Energy Use (GJ)	10590.53	3238.60	Right-skewed	23438.866328
7	Weather Normalized Source Energy Use (GJ)	10212.80	3127.50	Right-skewed	22869.729340
8	Source EUI (GJ/m^2)	2.28	1.68	Right-skewed	1.597846
9	Weather Normalized Source EUI (GJ/m^2)	2.32	1.69	Right-skewed	1.622897
10	Total GHG Emissions (Metric Tons CO2e)	724.28	228.40	Right-skewed	1490.433401
11	Total GHG Emissions Intensity ($kgCO2e/m^2$)	158.67	117.60	Right-skewed	109.472637
12	Direct GHG Emissions (Metric Tons CO2e)	281.87	78.70	Right-skewed	884.613506
13	Direct GHG Emissions Intensity ($kgCO2e/m^2$)	63.82	43.90	Right-skewed	56.835735
14	Electricity Use - Grid Purchase (kWh)	601102.79	223771.90	Right-skewed	1.184414e+06

	Column Name	Mean	Median	Distribution	Standard Deviation
15	Natural Gas Use (GJ)	5509.85	1532.40	Right-skewed	17221.116059

Right skewed in the energy utilization column means that some of the properties have exceptionally high energy usage.

2.2 Aggregations

• Compute the average Energy Use Intensity (EUI) by Property Type

Heated Swimming pool and Fitness Center/Health Club/Gym has more average energy use.
Below is the Average energy use intensity by property type.

	Primary Property Type - Self Selected	Number of Buildings	Average EUI
3	Heated Swimming Pool	15	4.805333
2	Fitness Center/Health Club/Gym	40	4.385000
0	Distribution Center	5	3.286000
4	Ice/Curling Rink	55	2.182200
12	Other - Recreation	10	2.165000
7	Museum	5	1.584000
16	Social/Meeting Hall	5	1.550000
11	Other - Public Services	25	1.526000
9	Office	120	1.519636
13	Performing Arts	5	1.302000
14	Repair Services (Vehicle, Shoe, Locksmith, etc.)	20	1.248000
1	Fire Station	179	1.208827
15	Self-Storage Facility	5	1.208000
5	Indoor Arena	10	1.106000
8	Non-Refrigerated Warehouse	15	0.768000
6	Mixed Use Property	5	0.458000
10	Other	5	0.070000

• Compute the total Greenhouse Gas (GHG) emissions by year

The highest GHG emission was on the year 2019 (75605.4) and lowest was on 2020(66617.2)

	Year	Total GHG Emissions
1	2020	66617.2
2	2021	68136.9
3	2022	72301.0
4	2023	75132.0
0	2019	75605.4

• Identify the top 5 properties with the highest total energy consumption

Top 5 properties for energy consumption property names and types are listed in the below tables

	Property Name	Site Energy Use (GJ)
92	Stoney Transit Facility	726554.8
74	Municipal Complex	406124.1
95	Village Square Leisure Centre	396268.1
89	Southland Leisure Centre	245215.1
58	Foothills Aquatic Centre and Bauer and Bush Ar...	122386.6

	Primary Property Type - Self Selected	Site Energy Use (GJ)
9	Office	1015336.9
2	Fitness Center/Health Club/Gym	889770.6
0	Distribution Center	726554.8
4	Ice/Curling Rink	567891.7
1	Fire Station	306095.3

2.3 Detecting Outliers Using Regex and IQR

Use Regex only to:
o Identify values that do not conform to expected numeric formats.
o Remove or correct incorrectly formatted numeric values.
This is same as 1.3, but repeated the same steps for the original dataFrame.

• **Apply the Interquartile Range (IQR) method to detect outliers in Total GHG Emissions (Metric Tons CO2e).**

Outliers for each property type is displayed below

	Primary Property Type - Self Selected	Count	LowerBound	UpperBound	Median
0	Distribution Center	5	1042.2500	13648.2500	7768.20
1	Fire Station	179	-16.3000	320.9000	140.50
2	Fitness Center/Health Club/Gym	40	-762.0375	2968.6625	790.45
3	Heated Swimming Pool	15	321.0000	949.0000	693.70
4	Ice/Curling Rink	50	-814.5250	2603.0750	916.55
5	Indoor Arena	5	1192.1500	1647.7500	1372.00
6	Mixed Use Property	5	27.9500	94.7500	65.00
7	Museum	5	250.7500	344.7500	287.10
8	Non-Refrigerated Warehouse	15	-1848.3250	4309.8750	550.90
9	Office	110	-1365.8625	2483.4375	356.20
10	Other	5	26.6000	73.0000	53.10
11	Other - Public Services	25	-2.5500	124.2500	53.00
12	Other - Recreation	10	-430.7125	794.3875	125.90
13	Performing Arts	5	125.5000	145.5000	136.40
14	Repair Services (Vehicle, Shoe, Locksmith, etc.)	10	201.9125	1396.4125	742.85
15	Self-Storage Facility	5	447.4500	598.2500	537.80
16	Social/Meeting Hall	5	286.2500	320.2500	305.50

• Replace outliers with the median value for that property type.

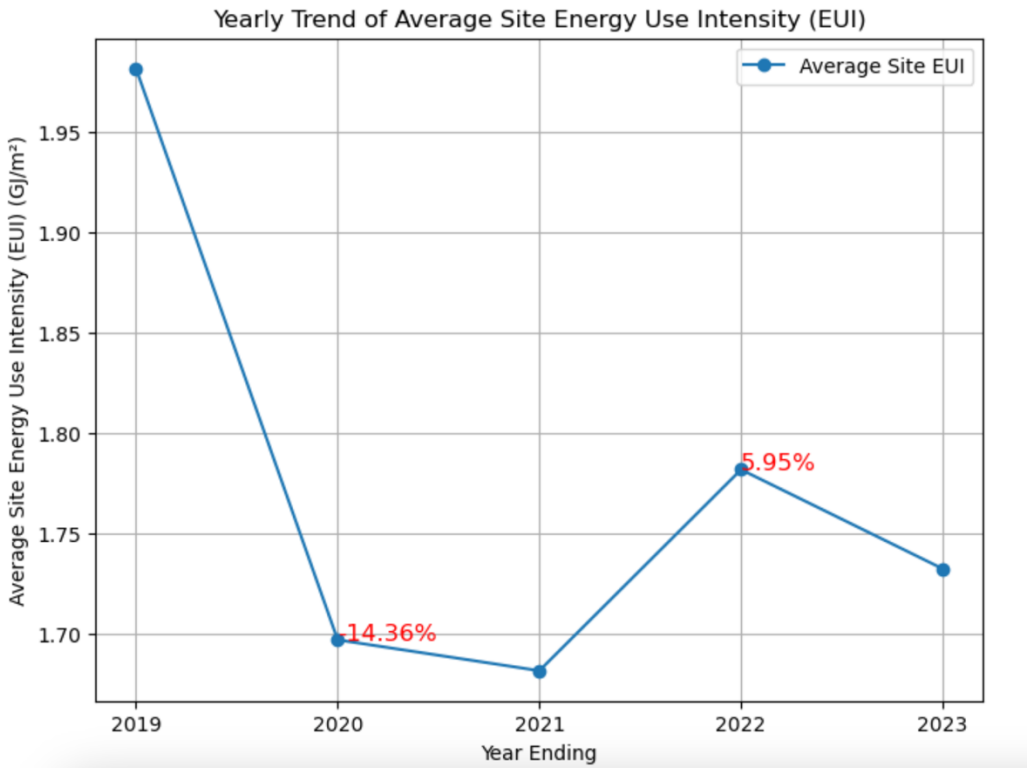
There was a total of 38 records with outliers for each property type. For each property the outlier has checked if "Total GHG Emissions (Metric Tons CO2e)" is not between the lower and upper bound, then it has replaced with the median of corresponding property type. For example: The property type "Distribution Center" has 5 outlier records for "Total GHG Emissions (Metric Tons CO2e)" and those are replaced with median 7768.20. A new column with name "New Total GHG Emissions (Metric Tons CO2e)" is introduced to remove outliers with median value. If the "Total GHG Emissions (Metric Tons CO2e)" column has no outlier then the new column is maintained to have the same value.

Part 3: Data Visualization

3.1 Time-Series Visualization

• Plot the yearly trend of average Site Energy Use Intensity (EUI).

• Highlight any significant increases or decreases in energy usage.



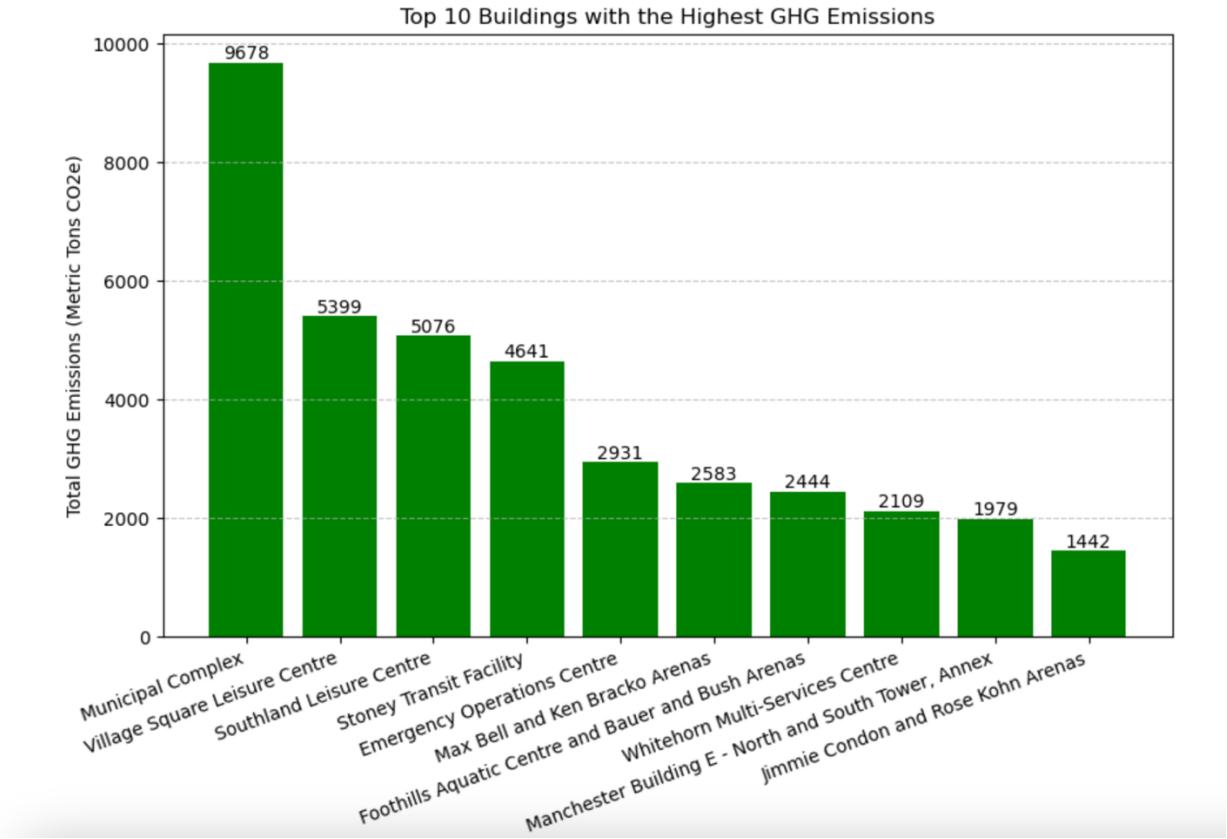
The Average Energy Use Intensity (EUI) trend has analyzed based for each year and if the percentage of change from last year to next is greater than 5% then the trend has been marked on the plot with percentage value in red color. The percentage of average energy use intensity was high on 2019 and on the next year it was lowered with 14.36% and then again decreased on 2021, but the decrease was not significant compared to 2020. For the year 2022 the average has increased to 5.59% from last year.

3.2 Comparative Bar Charts

• Create a bar chart showing the top 10 buildings with the highest GHG emissions.

• Annotate the bar chart with emission values.

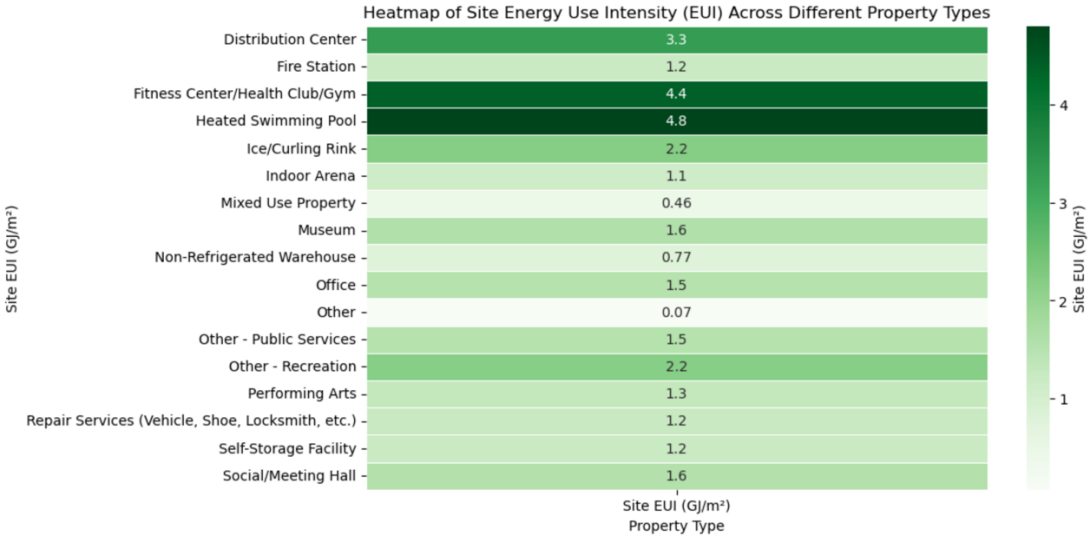
The highest emission buildings are: Municipal Complex(9678), Village square leisure centre(5399), southland leisure center,(5076) Stoney transit facility(4641), etc.



3.3 Heatmap Visualization

• Create a heatmap of energy usage intensity (Site EUI (GJ/m^2)) across different property types.

From the heatmap for energy usage intensity across different property type its clear that the Heated Swimming pools and Fitness centre/Gym/Health Club used more energy than compared to any other properties.



Part 4: Further Analysis

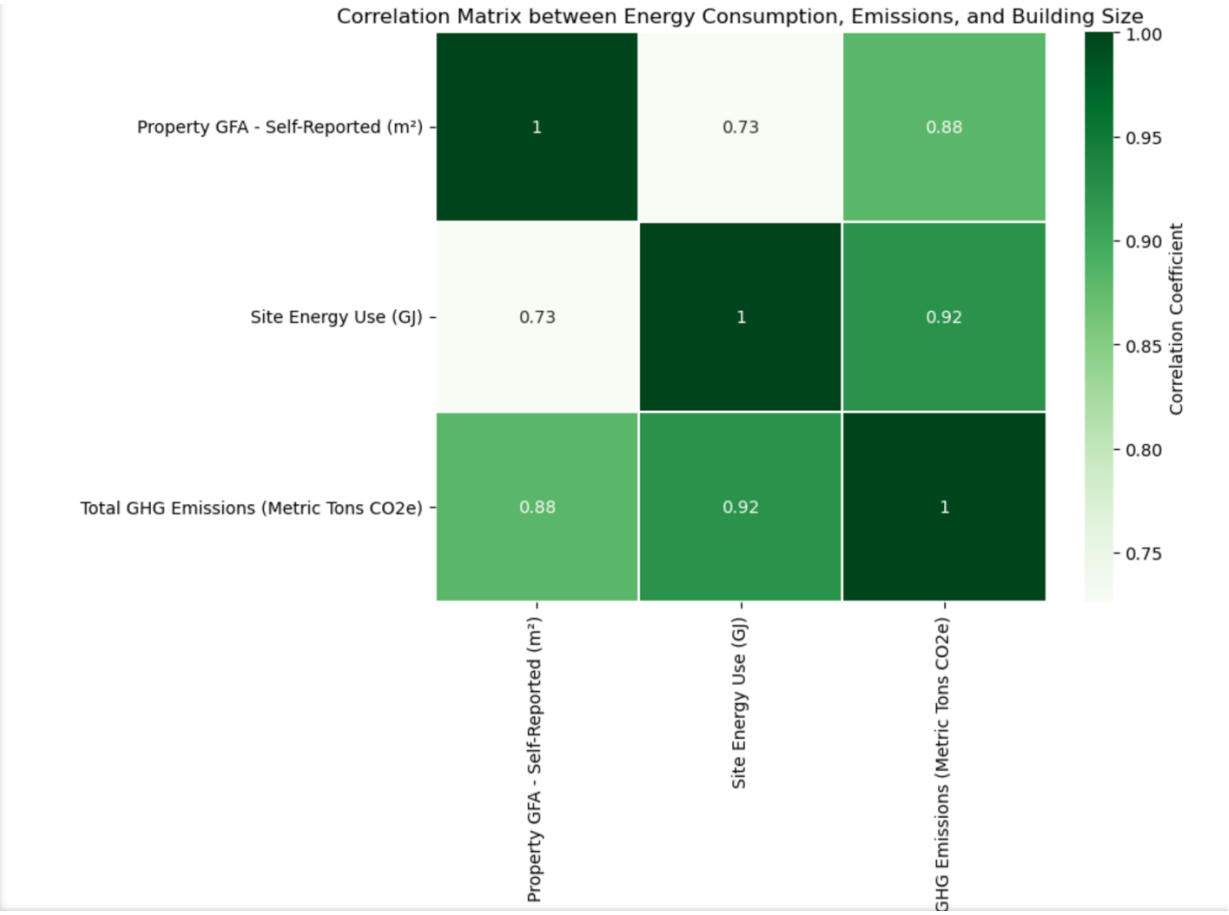
4.1 Correlation Analysis

• Compute and visualize the correlation matrix between energy consumption, emissions, and building size.

• Identify any strong correlations and explain their implications.

	Property GFA - Self-Reported (m²)	Site Energy Use (GJ)	Total GHG Emissions (Metric Tons CO2e)
Property GFA - Self-Reported (m²)	1.000000	0.725977	0.880927
Site Energy Use (GJ)	0.725977	1.000000	0.923281
Total GHG Emissions (Metric Tons CO2e)	0.880927	0.923281	1.000000

The correlation(r) is a value between -1 and 1 which can tell the degree of relationship between two variables. If r = 1 means its strongly positively correlated. 0 means no linear relationship. Negative values indicate weak relationship and if one tends to increase, the other tends to decrease. So, from the above table we can conclude that if the property Gross Floor Area increases the Site energy use and emission also increases. In other words, we can say the Property GFA, Site Energy Consumption and Emission are positively and strongly correlated with correlation between Property GFA and Energy consumption = 0.725977. Property GFA and Emission are correlated with r =0.880927. A strong positive correlation is assumed to have r > 0.7. Also, it is to be noted that if the Site Energy Consumption and Emission are very strongly & positively correlated with r =0.923281. The below heatmap can be interpreted as the more in green indicates the more the correlation is:



4.2 Hypothesis Testing

• Conduct a t-test (t-test is used to compare the means of two groups to determine if they are significantly di6erent from each other. More at Student's t-test - Wikipedia) comparing the average Energy Star Score between two different property types (e.g., Offices vs. Residential buildings).

• Interpret the results and discuss statistical significance

T-statistic measure the difference between group mean relative to variability within the groups whereas the P-value measures the significance for the same, if p value <=0.05 strong evidence for null hypothesis or there exist a significant difference between group means.

	Propety Name1	Propety Name2	T-statistic	P-value
0	Office	Non-Refrigerated Warehouse	-3.399196	9.505373e-04
1	Office	Ice/Curling Rink	3.242393	1.488677e-03
2	Office	Museum	0.609964	5.433735e-01
3	Office	Self-Storage Facility	2.561344	1.196741e-02
4	Office	Distribution Center	3.446742	8.403622e-04
5	Non-Refrigerated Warehouse	Ice/Curling Rink	6.617177	1.284498e-08
6	Non-Refrigerated Warehouse	Museum	2.670944	1.826618e-02
7	Non-Refrigerated Warehouse	Self-Storage Facility	8.201720	1.715875e-07
8	Non-Refrigerated Warehouse	Distribution Center	9.721004	1.377960e-08
9	Ice/Curling Rink	Museum	0.078937	9.374403e-01
10	Ice/Curling Rink	Self-Storage Facility	1.693175	9.690363e-02
11	Ice/Curling Rink	Distribution Center	2.813104	7.089528e-03
12	Museum	Self-Storage Facility	11.342269	3.444883e-04
13	Museum	Distribution Center	6.395890	3.068237e-03
14	Self-Storage Facility	Distribution Center	6.342197	2.224634e-04

Since the tStatistic for Office vs Museum, Ice/Curling Rink vs. Museum and Ice/Curling Rink vs. Self-Storage Facility is nearly 0, that means they do not have significance difference and their means are similar. So, they could show similar characteristics.

For Museum and Self-Storage facility are having high significance difference with each other. So, they have entirely different characteristics. All other statistics are having not much difference than Museum and Self-Storage facility but they do have significant difference in their mean.

Part 5: Reporting and Insights

5.1 Summary Report

• Write a structured report (300-500 words) covering:

o Key trends in energy consumption and efficiency.

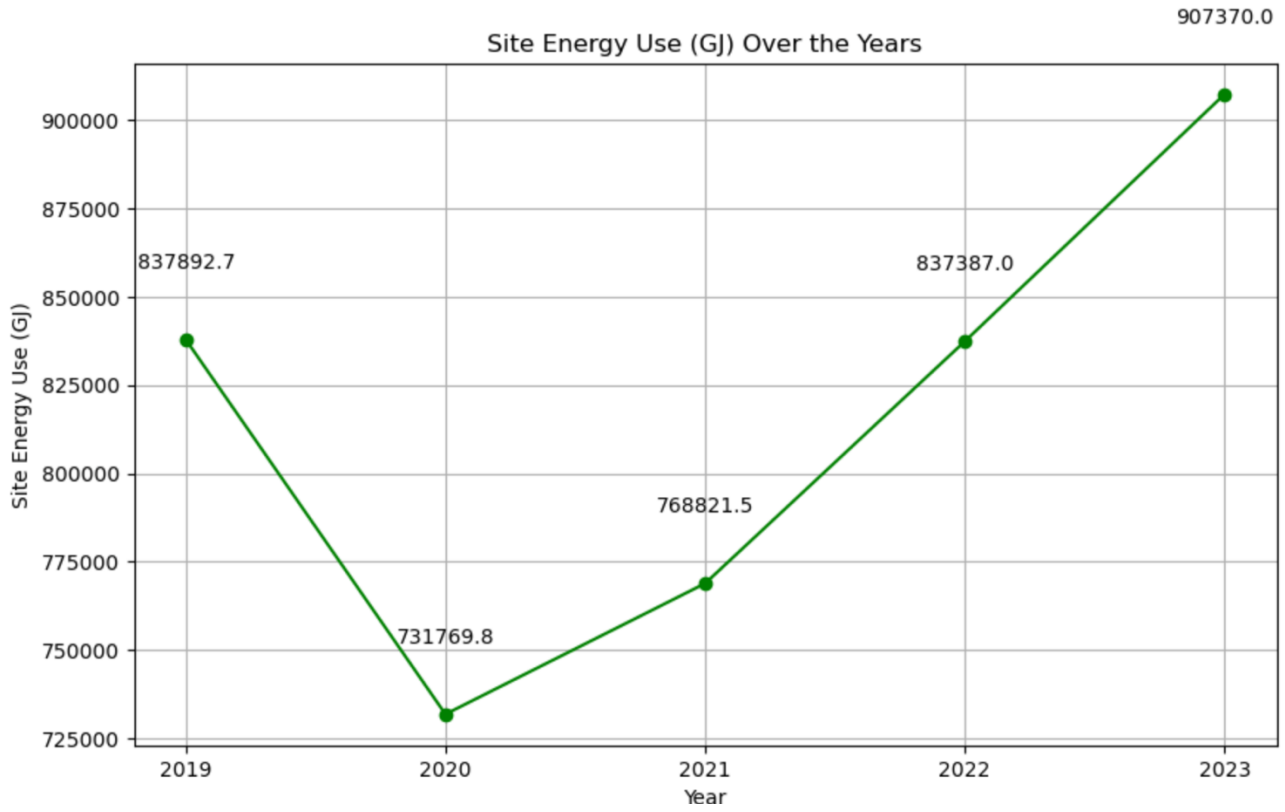
o Seasonal and property type variations.

o Recommendations for improving energy efficiency and reducing emissions.

Summary report for Energy Consumption and efficiency:

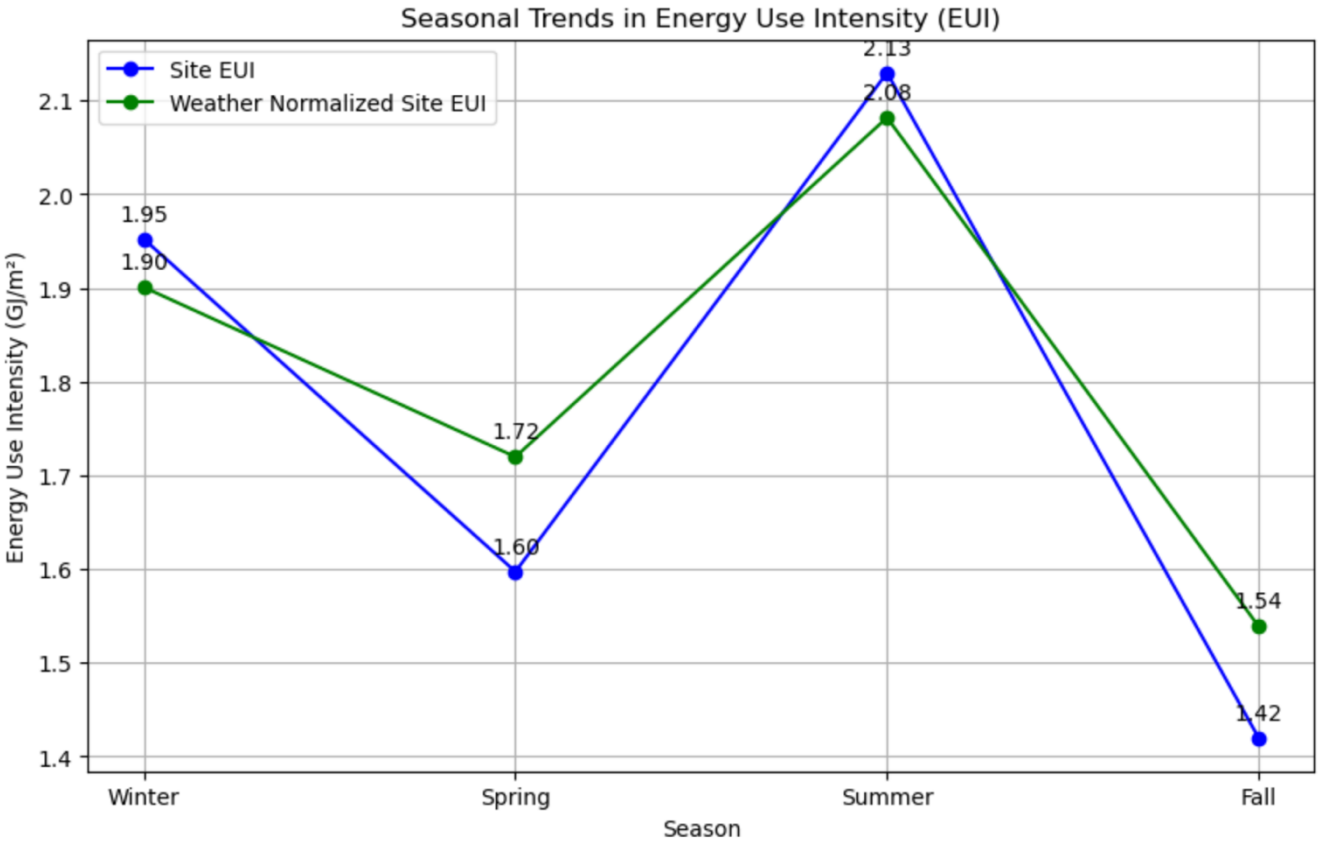
The total energy consumption from the year 2019 to 2023 is 4083241.0GJ, in which Offices are the most energy utilizers. We can say the Stoney Transit facility has the most consumption for all these years. The average utilization per property for the years from 2019 to 2023 is 8265.67004048583GJ. Also, when we check on yearly basis the utilization is keep on increasing. I recommend its better to implement some energy efficient technologies and practices in Office buildings for saving energy. Also, it would be good to have a benchmark utilization for energy and keep track of the performance based on this benchmark

Total Energy Use (GJ): 4083241.0
Average Energy Use per Property (GJ): 8265.67004048583
Year Ending
2019 837892.7
2020 731769.8
2021 768821.5
2022 837387.0
2023 907370.0



Seasonal and property type variations.

Without any weather consideration the average is 1.7747975708502024 and with weather it is 1.810323886639676. By analysing the graph below the extreme weather seasons (winter and summer) the average usage intensity has increased. It's clearly gives us the information that there is a slight increase under weather conditions. This makes it important to analyse and make improvement measures for energy consumption.
The total greenhouse gas emissions from the properties amount to 357,792.5. Average emission is 724.2763157894736. Average emission intensity is 158.67165991902831. This considerably high and some improvement should be done and identify the inefficiencies and potential improvement areas.



Recommendations for improving energy efficiency and reducing emissions

Reducing electricity use can lead to significant cost savings and environmental benefits so I would recommend the following:

1. Conduct energy usage audits
2. Set benchmarks for energy utilization
3. Implement energy efficiency technologies/practices
4. High energy users should be tracking the utilizations
5. Continuous monitoring and implementation of better strategies
6. replacement with less emission energy sources like solar panels and wind energy

Additional Tasks

• Analyze the relationship between building age and energy efficiency.

• Use Regex only to clean and standardize text-based data such as property names and addresses.

• Generate a dashboard-style visualization combining multiple Matplotlib plots for an interactive overview.

• Ensure the GitHub repository follows best practices, including an organized folder structure, detailed README.md, and version-controlled commits with meaningful messages.

Correlation between Building Age and Site EUI: 0.17051313527653905
Correlation between Building Age and Weather Normalized Site EUI: 0.17359003070797616

The correlation between Building Age and Site EUI and Building Age and Weather Normalized Site EUI are seems to be weekly positively correlated. This means that if the building age increases then there would be a slight increase in the EUI but it would be low. So, it's clear that the building age is not an affecting factor for increase in energy consumption. Still, I believe the buildings should be considered to follow some specific strategies to attain the controlled consumption rate. Below is the scattered plot indicating the above-mentioned correlations.