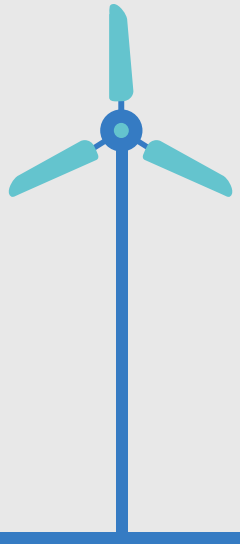
How to use Energy to help our Environment



Contents

Background

Protecting the environment has always been the goal of human's joint effort. Greenhouse gas is one of the major impact on influencing the environment. In this Hackathon, we chose to analyze the greenhouse gas emission using the data set of buildings' energy consumption in New York, finding out how to better help our environment.

Problem Statement

- **How to utilize a standardized Energy Use Intensity (EUI) to compute the relatively electricity and natural gas consumption in order to achieve the greenhouse gas emission goal.**

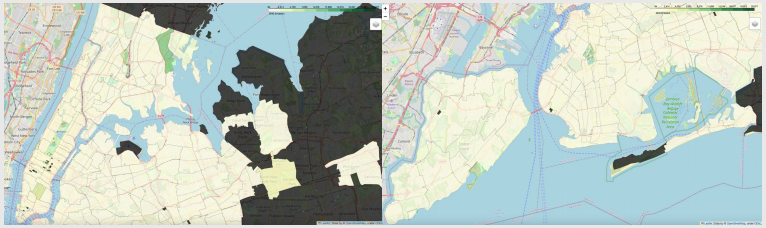
$$EUI \approx \frac{ElectricityUse + NaturalGasUse}{TotalFloorArea}$$

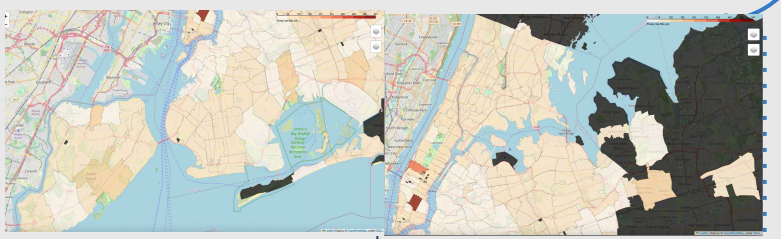
E = A x EF x (1-ER/100)

E = emissions, A = activity rate, EF = emission factor, and ER = overall emission reduction efficiency, in a percentage

Analysis

Geographical View





Total Emission

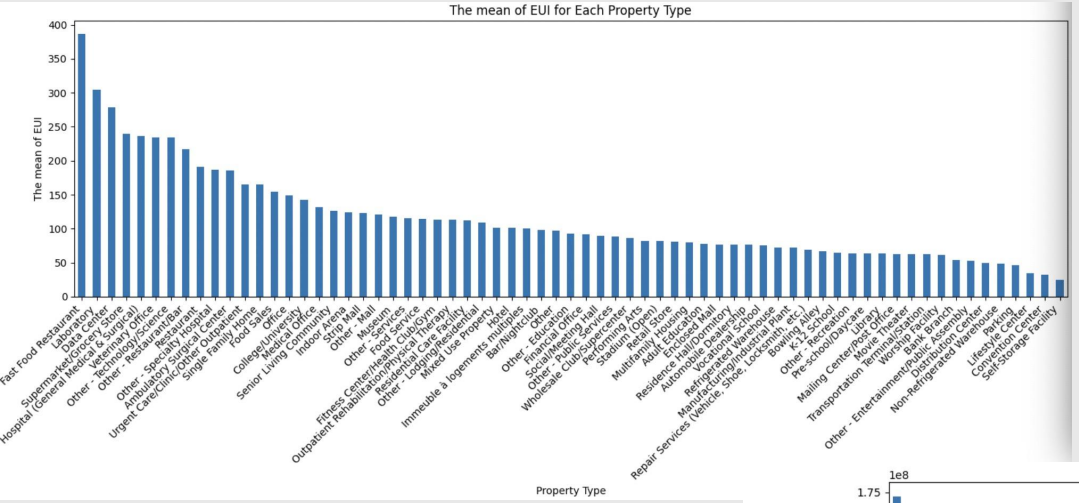
Site EUI

Correlation Matrix between Site EUI and Total GHG Emissions Intensity

	Site EUI (kBtu/ft²)	Total GHG Emissions Intensity (kgCO2e/ft²)
Site EUI (kBtu/ft²)	1.000000	0.999273
Total GHG Emissions Intensity (kgCO2e/ft²)	0.999273	1.000000

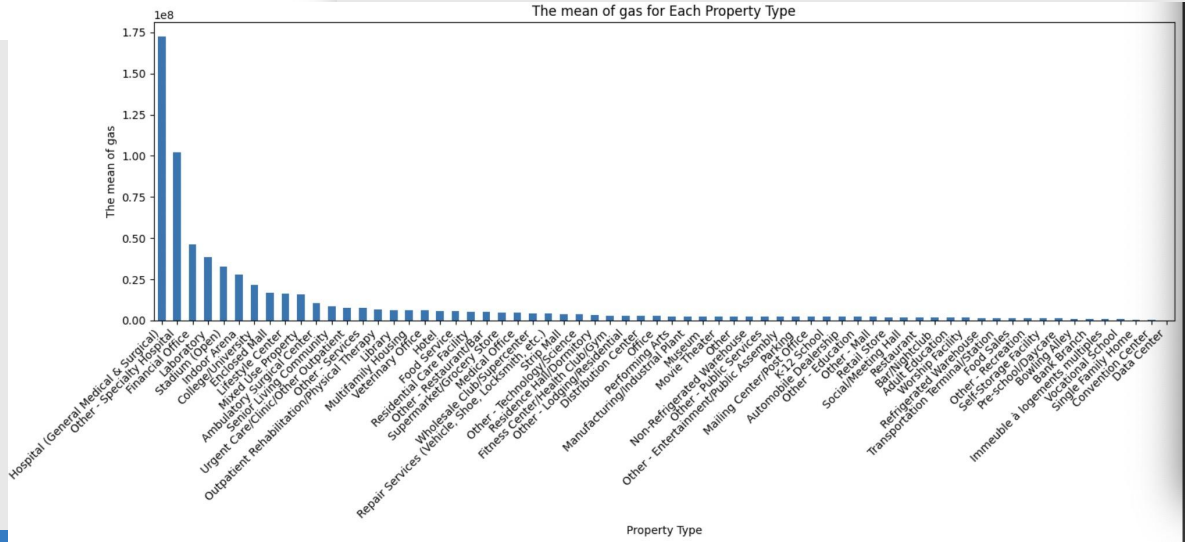
By looking at the Geographical graphs and correlation matrix, we found that they are highly correlated!

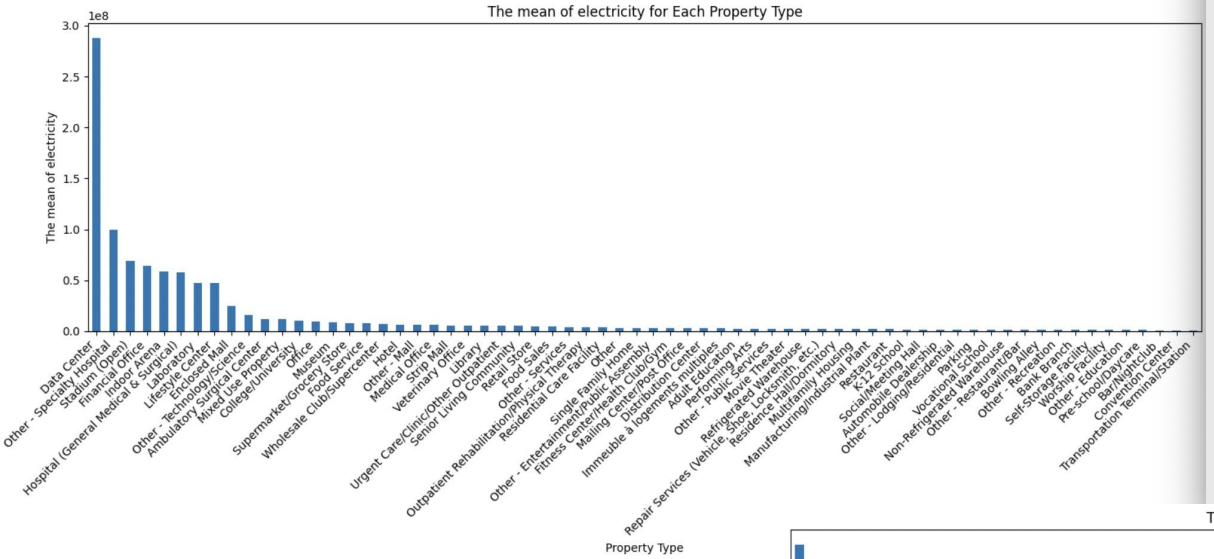
Visualize Average Energy use for different property types



Most of these plots make sense such as data center requires more electricity, but some require more closer inspections such as why hospital use more gas?

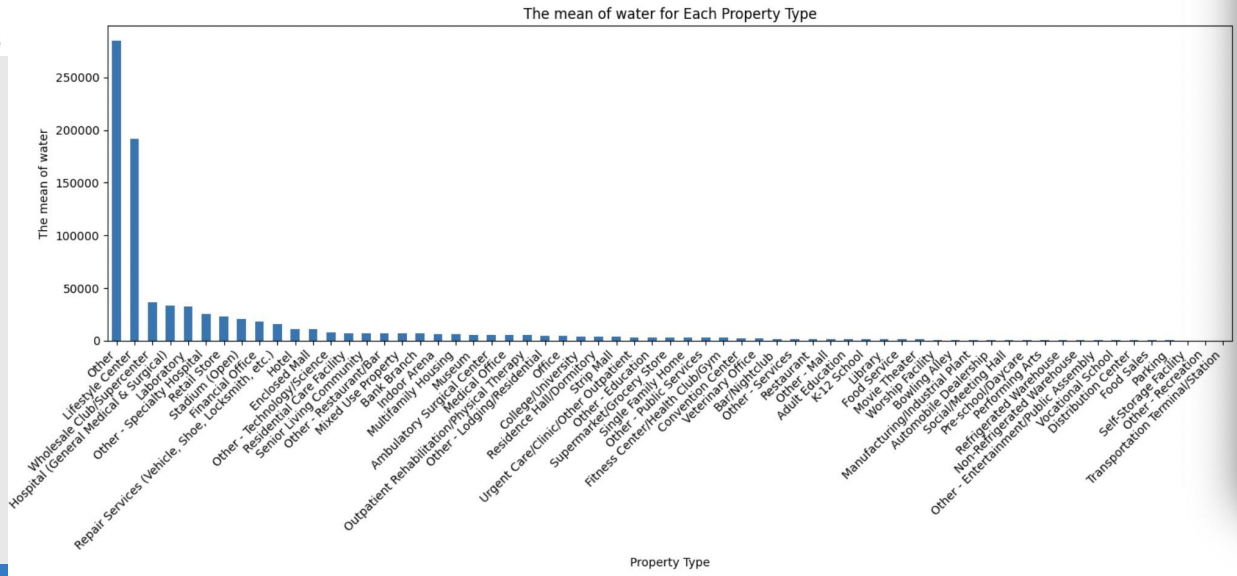
To reduce the eui, maybe we can reduce the number of 24 hours fast food restaurants



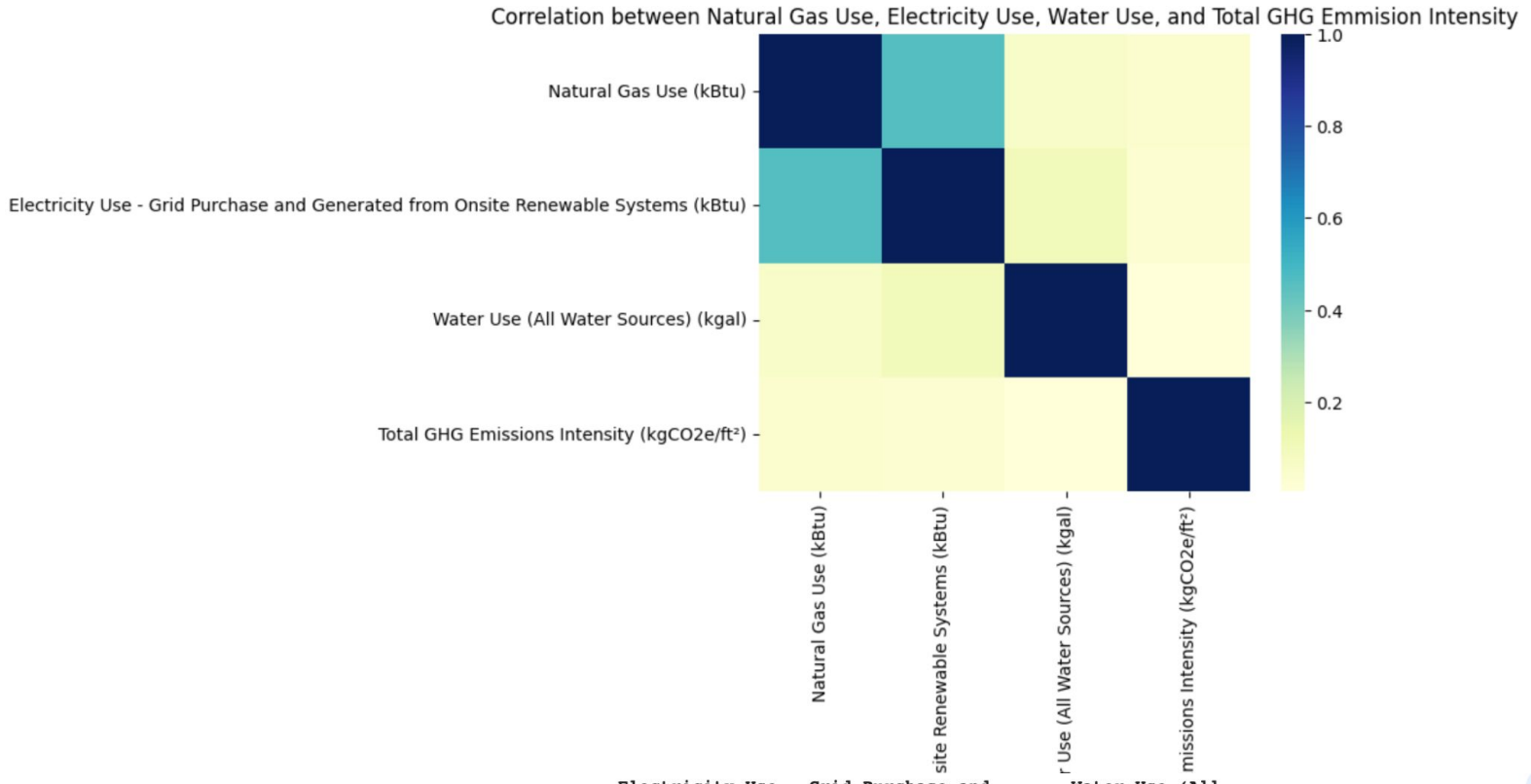


Different property types do have different impacts on energy usage. Maybe add property type as a feature?But we still need to double check later.

Close the landscaping which using water on working days ex.fountain .



Visualizing the correlation matrix again, we found that Energy and water have relatively neutral linear correlations with total GHG emission intensity.
So for this data, we assume that linear regression is inapplicable to build model.



	Natural Gas Use (kBtu)	Electricity Use - Grid Purchase and Generated from Onsite Renewable Systems (kBtu)	Water Use (All Water Sources) (kgal)	Total GHG Emissions Intensity (kgCO2e/ft²)
Natural Gas Use (kBtu)	1.000000	0.461675	0.056475	0.042195
Electricity Use - Grid Purchase and Generated from Onsite Renewable Systems (kBtu)	0.461675	1.000000	0.103719	0.030537
Water Use (All Water Sources) (kgal)	0.056475	0.103719	1.000000	0.005525
Total GHG Emissions Intensity (kgCO2e/ft²)	0.042195	0.030537	0.005525	1.000000

Prediction

The primary reason that we use a prediction model in the project is to find out the combination of electricity and natural gas used to achieve a certain emission goal.

Feature Selection

There are over 150 columns of data so it is crucial to select the most important features for our prediction model. We used univariate feature selection method to do this. And after this step, we are left with columns such as total emission intensity, property types, area codes, and a couple more. But most of them align with our assumption during visualization phase.

What we tried

Lower R^2

Random Forest

Higher R^2

XGBoost is a better model so we chose it!

XGBoost

After performing Hyperparameter Grid Search, we chose the optimal XGBoost model and achieved a R-squared score of 0.95

Outcome

For example, if the government of New York decides to decrease the greenhouse gas by 10% for different property types, we have this suitable data for New York to calculate how much percentage of the Energy Usage Intensity to decrease, leading to the calculation of electricity and natural gas usage.