Milestone 3

Sukhraj Dulay, Maitri Shah, Paula Zhuang, Josh Zhang

Contents

1 Finalized Research Question

Based on the available data, from categories present to initial analysis done in Milestone 2, we have the following finalized research question:

How do building characteristics, such as size, age, and primary use type, impact energy use and greenhouse gas emissions in Cambridge, MA?

To study this question, we will first understand the data provided to us (going in detail into the relevant columns and categories) and then determine how best to compare various characteristics of buildings to greenhouse gas emissions. We will search for relationships between both single characteristics and gas emissions, as well as the combination of more than one category and how it affects the emissions of greenhouse gases for such buildings in Cambridge, MA. The ultimate goal of this project is to provide a predictive model for the greenhouse gas emissions of some property in Cambridge, which can provide relevant information on future energy emission for such buildings, areas in which decrease can be attempted, and extrapolations for buildings not exactly identified in this dataset.

2 Data Description

Our dataset, provided by the City of Cambridge under the BEUDO program, contains information on energy and water use for various properties within Cambridge, MA, from 2015 to 2022. This data is gathered under an ordinance that requires large buildings to report their energy usage. Property owners submit their data through ENERGY STAR Portfolio Manager, and the City of Cambridge supplements this with parcel-level data.

3 Summary of the Data

After coding through the dataset and analyzing our dataset's shape, columns, and quantity of missing, we find the following information and also categorized our data within the following:

- **Dataset Shape:** (851 rows, 46 columns)
- **Identifiers for Each Building:** Reporting ID, Annual Report Received, Owner
- **Building Category / Characteristics:** BEUDO Category, PD Parcel Living Area, PD Parcel Units, Year Built, Buildings Included Count, Primary Property Type - Self Selected, All Property Uses, Property GFA - Self Reported (ft2), ENERGY STAR Score (¿ 50% null values but kept for additional color)
- **Energy Information:** 'Natural Gas Use (therms)', 'Natural Gas Use (kBtu)', 'Site Energy Use (kBtu)', 'Weather Normalized Site Energy Use (kBtu)', 'Site EUI (kBtu/ft2)', 'Weather Normalized Site EUI (kBtu/ft2)', 'Source Energy Use (kBtu)', 'Weather Normalized Source Energy Use (kBtu)', 'Source EUI (kBtu/ft2)', 'Weather Normalized Source EUI (kBtu/ft2)', 'Total GHG Emissions (Metric Tons CO2e)', 'Total GHG Emissions Intensity (kgCO2e/ft2)'
- **Less Important (either too many null values (>50% of the rows are null) or reasoning described):** Data Year (all are in 2022), MapLot (does not affect property information), Address (same as MapLot), Reported Residential Units, Buildings Included (specific buildings is less useful), Owner Line 2, 'Electricity Use - Grid Purchase (kWh)', 'Electricity Use - Grid Purchase (kBtu)', 'Fuel Oil #1 Use (kBtu)', 'Fuel Oil #2 Use (kBtu)', 'Fuel Oil #4 Use (kBtu)', 'Fuel Oil #5 & 6 Use (kBtu)', 'Diesel #2 Use (kBtu)', 'Kerosene Use (kBtu)', 'District Chilled Water Use (kBtu)',

> ‘District Steam Use (kBtu)‘, ‘Electricity Use - Generated From Onsite Renewable Systems (kWh)‘, ‘Water Use (All Water Sources) (kgal)‘, ‘Water Intensity (All Water Sources) (gal/ft2)‘, Latitude, Longitude, Location (doesn’t affect characteristics)

Given that our research question aims to create a predictive model for greenhouse gas emissions for properties, we categorized our data into four groups: building identification to ensure unique and accurate tracking of each building, building category/characteristics to explore correlations between building traits and emissions, energy information to use in predicting future emissions, and less significant data to exclude from the model due to limited relevance. From this point forward, we have dropped all variables that were classified in the less important category due to limited information or relevance reasons.

Note: descriptive data for each feature (data type, mean, max, etc.) were not included for each feature in this report due to the number of features. However, they are in the Jupyter Lab.

3.1 Data Cleaning

In our initial exploration, we explored the shape of the data and its features, and this informed the approach to EDA in this report.

Additionally, we looked at missing data. Importantly, we found that some features have extensive missing values, and we decided not to use those features in our final research question and analysis (the columns were dropped). The features that we include in this report have minimal missing data.

3.2 Distribution of Categorical Data

Analyzing the most significant categorical variables, it was logical to simply plot the distributions of each respective variable in hopes of focusing specific attributes within these categorical variables for further simplification of our final model. Analyzing the frequency distribtuions of `Annual Report Received`, `BEUDO Category`, `Primary Property Type - Self Selected` we see the following:

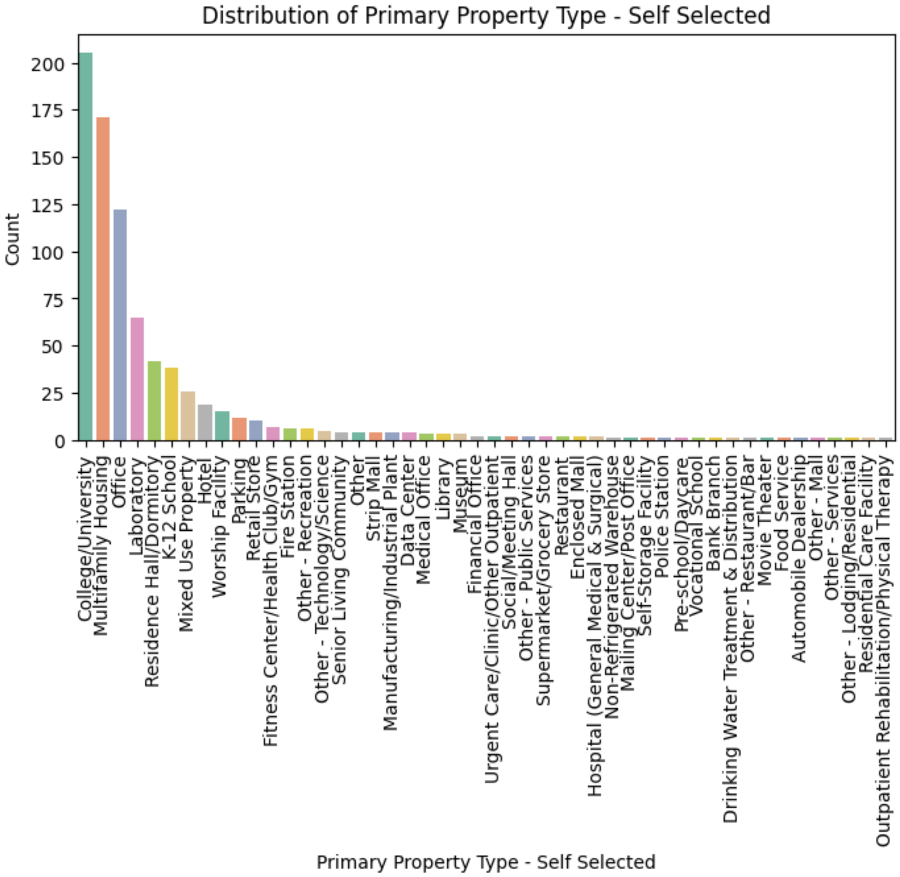


Figure 1: Distribution of Primary Property Type

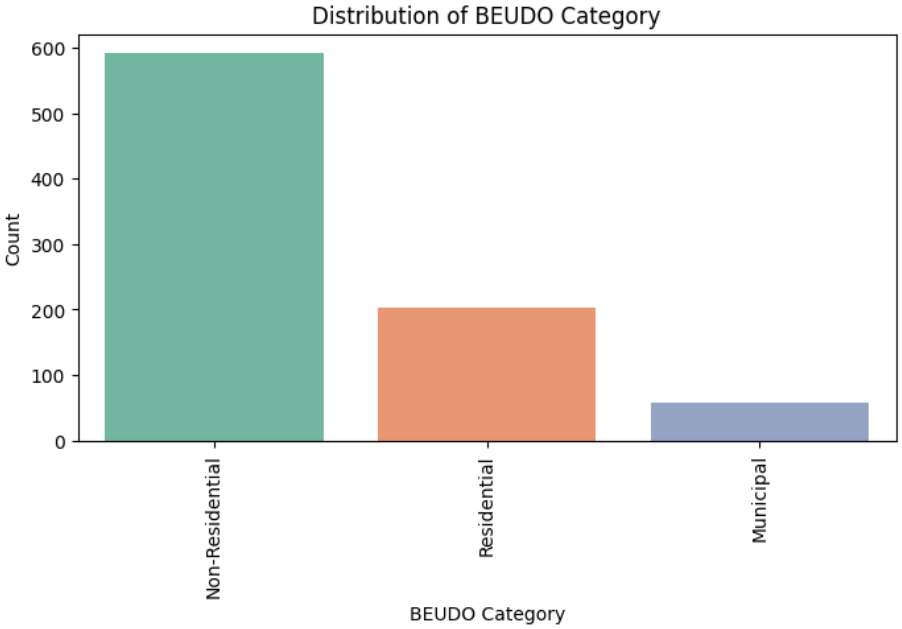


Figure 2: Distribution of BEUDO Category

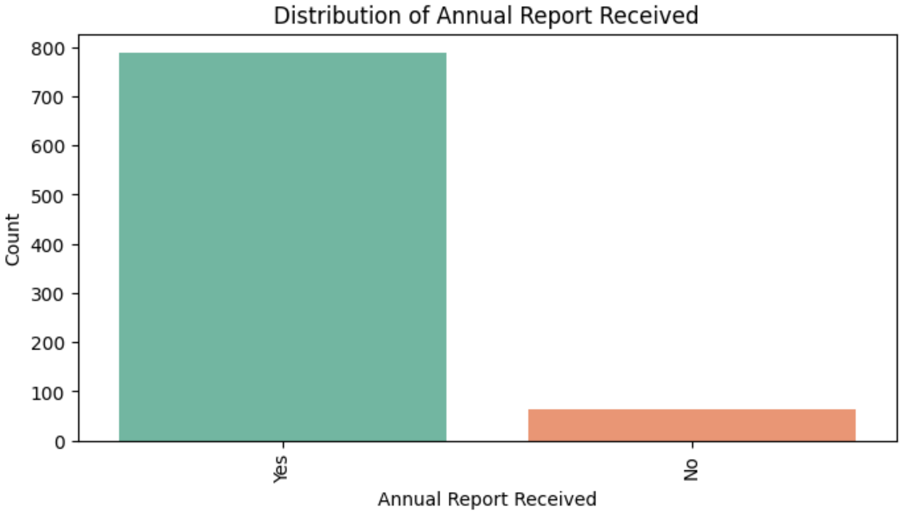


Figure 3: Distribution of Annual Report Received

The categorical visualizations provide a few key takeaways that are important as we further continue to study the relationships between building category

/ characteristics and energy consumption.

As per figure 1, the top three primary property types—College/University, Multifamily Housing, and Office—represent the majority of buildings in Cambridge. By focusing our analysis on these three types, we can simplify the dataset while still accurately reflecting broader trends in the city. Per figure 2 as well, the data shows that the majority of properties are non-residential, with a ratio of 3:1 compared to residential properties and 12:1 compared to municipal properties. Per figure 3, while not as relevant, we find that the majority of residents successfully submitted their annual report.

3.3 Distribution of Building Category Variables

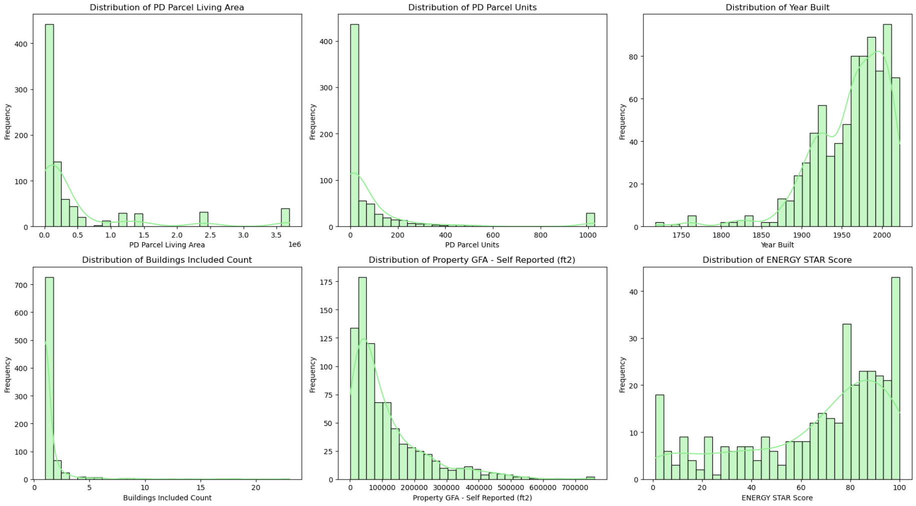


Figure 4: Distribution of Annual Report Received

As per the figure above, these histograms provide a general sense of relative size, year built, and other characteristics of the buildings, but there is a very clear skew for all except for ENERGY STAR Score. It would make more sense to compare these to the property types to see if specific categories of buildings (combination of type with some of these characteristics) can be generated.

We can compare Property GFA, Year Built, and ENERGY STAR Score with the Primary Property Types to see distribution across each property type. These three numerical variables are selected because we have the most distribution in data for them, as seen in the above histograms (the others are very concentrated at one or two values. Let us proceed:

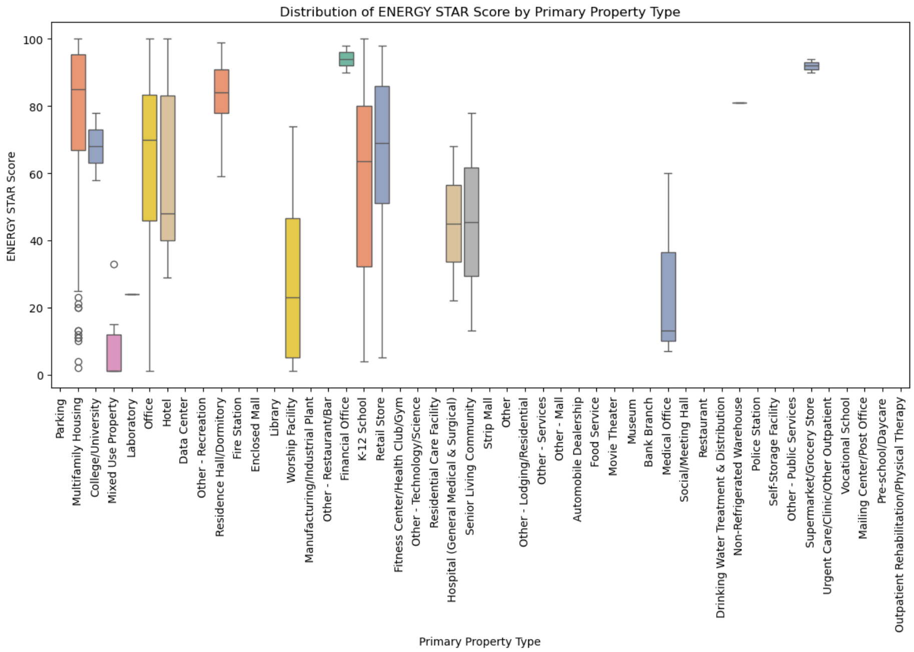


Figure 5: Distribution of ENERGY STAR Score by Primary Property Type

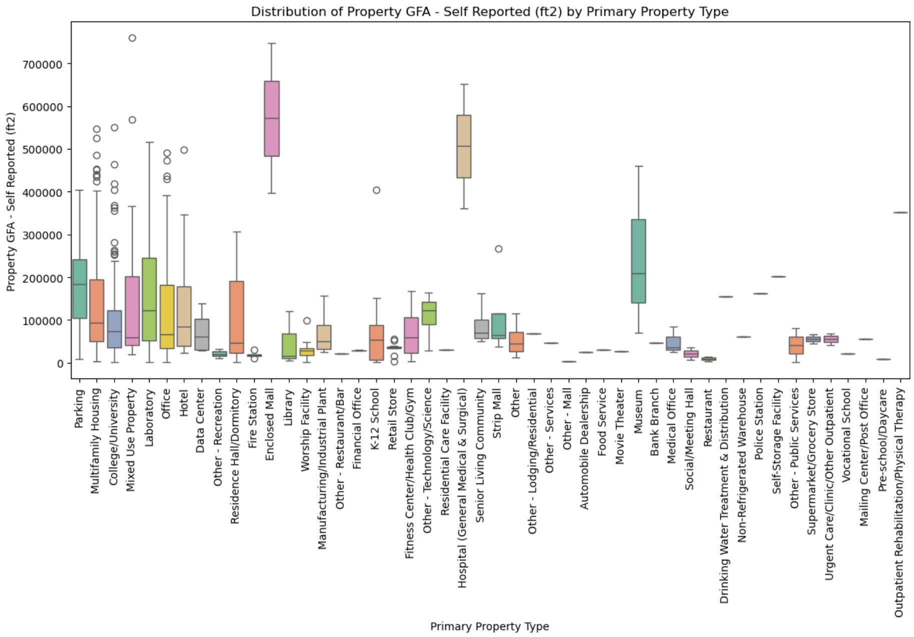


Figure 6: Distribution of Property GFA by Primary Property Type

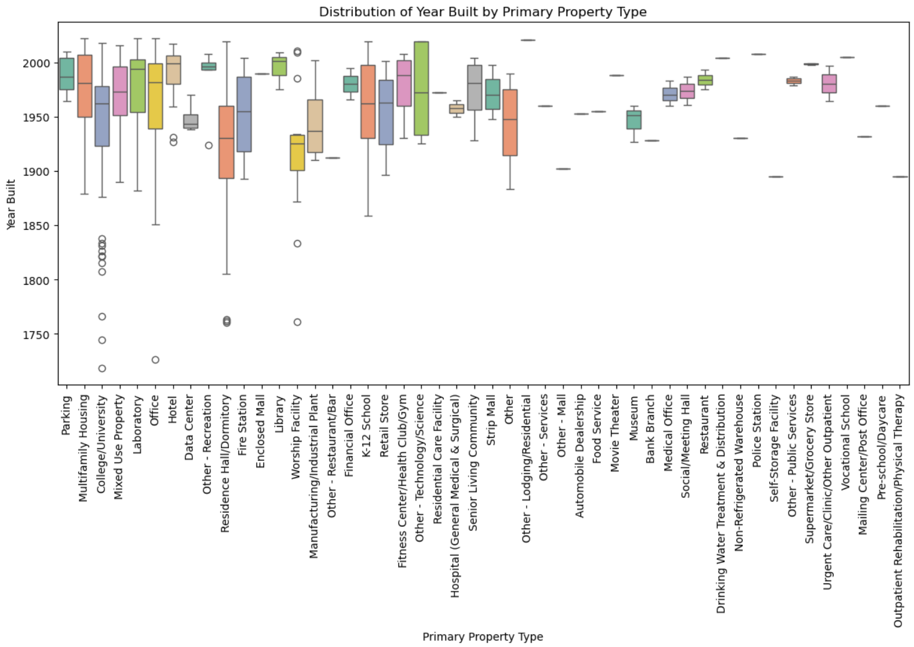


Figure 7: Distribution of Year Built by by Primary Property Type

From this, we come to a few general conclusions about the buildings themselves. While this does not inform much about the correlation with energy usage, consistent and logical findings here support the validity of the data and buildings surveyed.

- **Year Built-** Buildings vary widely in construction year, with some categories showing more recent build trends (hotel, recreation facilities, etc.) while others have much older construction years (religious worship facility, college/university, etc.). A lot of these makes sense, because we see a lot of renovation and new buildings in the office and family home space but less in religious facilities or college campuses, which are often preserved over a period of time.

- **Propery GFA - Self Reported (ft2)-** For this metric, there is significant variation across categories, with some properties (college/university and enclosed mall) having much larger overall areas than others (bank branch, medical office). This makes a lot of sense given the type of property and fits with general assumptions around these buildings, which supports the accuracy of the data. It is also important to note that there is a wide range of overall areas in some categories, especially multifamily housing; this suggests a mix of both large and small buildings that make up this category, which is something to consider when we consider the energy consumption of this group in particular. It is also important to note, however, that it is likely that some categories just don't have enough data to - for instance, grocery store / supermarket is shown to have a very small overall area, which logically does not seem correct when compared to buildings like medical offices and libraries.

- **ENERGY STAR Score-** The ENERGY STAR Score also varies significantly between property types, but seemingly without much reasoning. Some categories, like residence hall / dormitory and financial office, have fairly high scores; on the other hand, buildings like mixed use property and medical offices have pretty low energy scores. It is difficult to determine the validity of these since we do not have an intuitive idea of what energy scores should be for different property types. It is also likely that ENERGY STAR scores vary across property types due to differences in building use, regulations, or retrofitting practices. In such cases, types with consistent scores may reflect standardized energy practices within those categories.

4 Deeper Understanding, Insights, and Visualizations

This section includes understanding the data and its patterns, trends, class imbalances, relationships, and outliers, in relation to our question. It also includes meaningful insights and visualizations.

4.1 Correlations

We start by looking at the correlations between variables that we are interested in, to gain a deeper understanding of how our analysis will proceed.

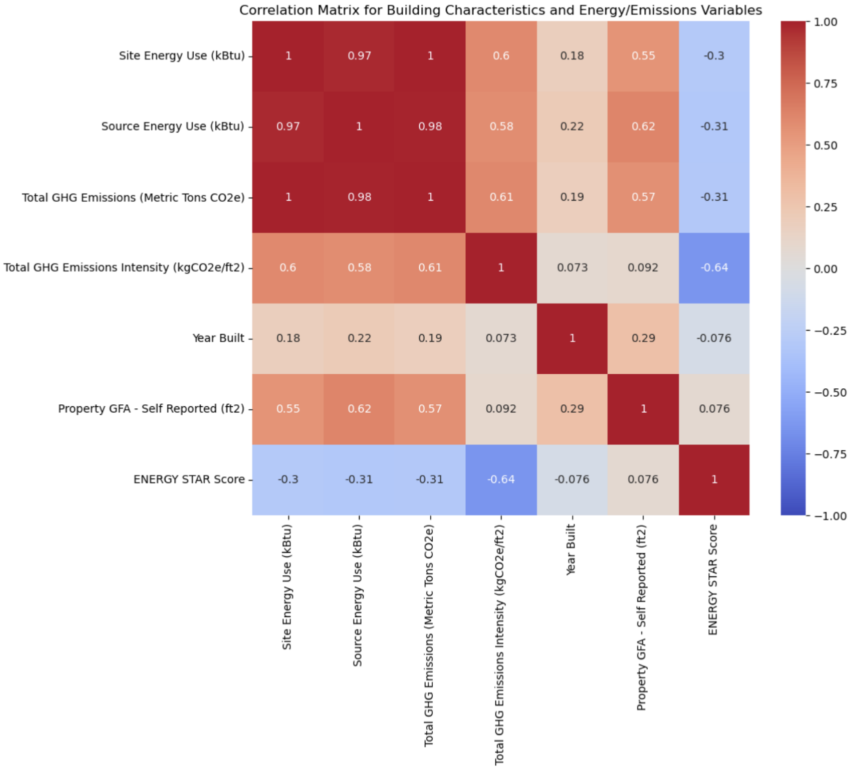


Figure 8: Correlation Matrix for Building Characteristics and Energy/Emissions Variables

After analyzing this heat map, we find a multitude of correlations between variables:

1. First, we see that the Site Energy Use (kBtu), Source Energy Use (kBtu), and Total GHG Emissions (Metric Tons CO2e) variables are all highly correlated with each other, with coeficients bordering 1. Additionally, we see that all three variables also have essentially identical correlations when compared to each of the characteristics variables as well. Being that this is the case, this potentially indicates that these variables are actually capturing similar aspects of data of energy consumption and emissions and that we may be able to utilize only of these variables to encompass the data for all 3 of these variables. This makes intuitive sense as well.

2. GHG Emissions Intensity is fairly correlated with the previous three variables, but not highly. This implies that GHG Emissions Intensity is calculated in a different way rather than a simple division of Total GHG Emissions.

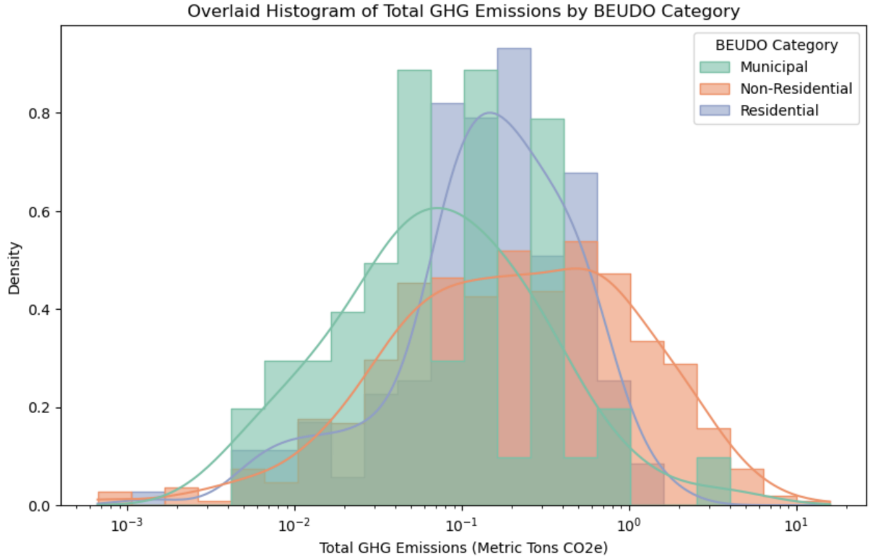
3. We find that there is a moderately high positive correlation between Prop-

erty GFA - Self Reported (ft2) and each of the respective energy/emission variables with a correlation of ≈ 0.57. This suggests that larger buildings tend to utilize more energy and emit more GHG, which falls in lines of the expectation that larger buildings tend to utilize more energy/emissions.

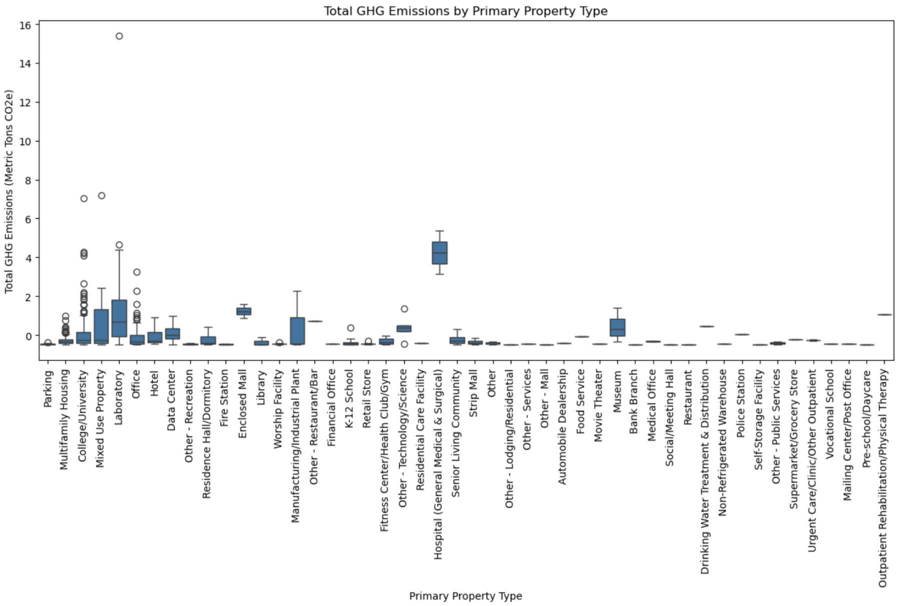
4. The Year Built is found to a weak positive correlation between the energy/emission variables of ≈ 0.20. Through this limited correlation given, no conclusive inference can be made here.

4.2 Exploring the Relationship Between GHG Emissions and Building Characteristics

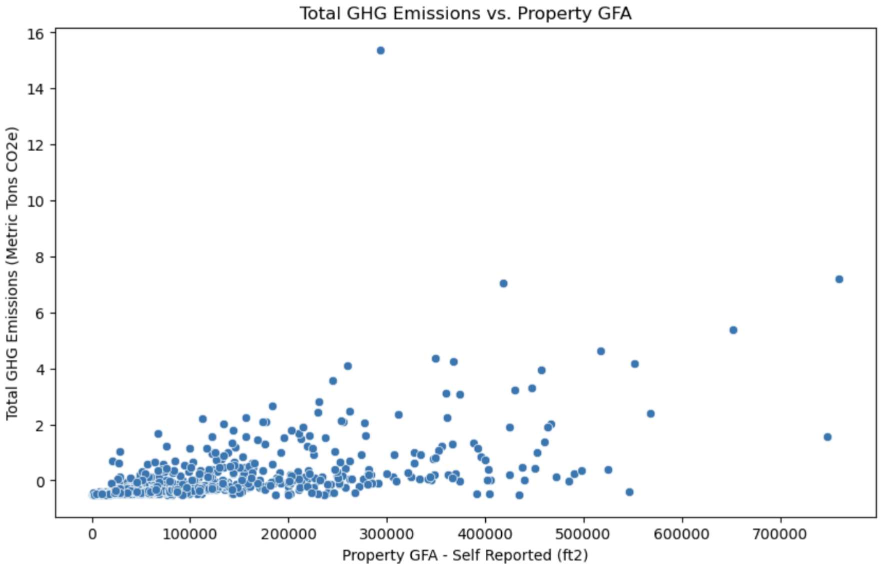
In this section, we explore the relationship between GHG Emissions and various building characteristics. We look at total GHG Emissions here, and not GHG emissions intensity, although their graphs were fairly similar.



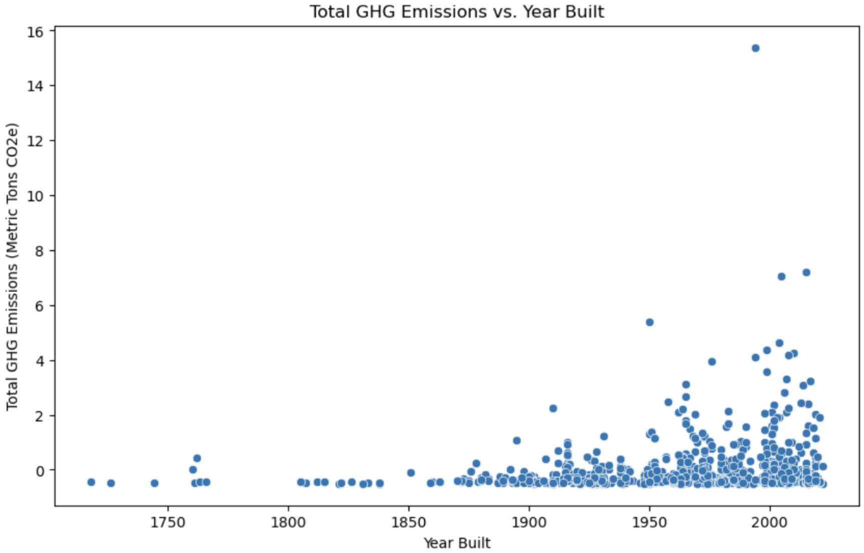
This histogram of ”Total GHG Emissions” categorized by BEUDO Category shows how emissions are distributed differently across the building types. The distributions for each of the categories are non-identical and shifted: Municipal is the most left-shifted, Residential is in the middle, and the distribution for Non-Residential is the most right-shifted. Additionally, the Non-Residential category exhibits a wider range of emissions, with a peak that is lower and broader compared to the other categories, suggesting significant variability in emissions levels. Generally, BEUDO category seems like a fair predictor of GHG emissions, with further potential for accurate prediction when combined with features like property type, GFA, and year built.



The boxplot of ”Total GHG Emissions” by ”Primary Property Type” reveals substantial variability across different property types, indicating that this variable has strong predictive potential for GHG emissions. Certain property types, such as ”Laboratory,” ”Mixed Use Property,” and ”Hospital,” exhibit higher median emissions with a broad range, highlighting the influence of energy-intensive activities or large-scale operations on emissions. In contrast, many other property types show low median emissions with limited variability, suggesting more predictable and uniform energy use. Outliers are present in several categories, indicating specific buildings with exceptionally high emissions, likely due to unique factors such as large size, high occupancy, or specialized energy usage. This variability suggests that incorporating ”Primary Property Type” into the model as a categorical feature can improve prediction accuracy. Additionally, further analysis to understand the drivers of emissions within high-variability categories could enhance model interpretability and predictive power.



The scatter plot of ”Total GHG Emissions” versus ”Property GFA” suggests a positive relationship between the size of a property and its total GHG emissions. As the property GFA increases, there is a general trend toward higher emissions, though with considerable variability, particularly among larger properties. This pattern indicates that larger buildings typically produce more emissions, likely due to increased energy demands. However, the spread of data points, especially for large GFA values, suggests that other factors also influence emissions, such as property type, building age, and energy efficiency measures. Overall, including ”Property GFA” as a predictor in the model seems like it would be useful for estimating GHG emissions.



The scatter plot of ”Total GHG Emissions” versus ”Year Built” reveals some key patterns. Buildings constructed more recently (post-1950) appear to exhibit a wider range of GHG emissions, with some properties showing significantly high emissions. This may be due to the increasing size and complexity of modern buildings or variations in energy efficiency standards over time. Conversely, older buildings tend to cluster around lower emission values, which could reflect smaller sizes, lower energy consumption, or updates and retrofits to improve energy performance. While there is no clear linear relationship, the distribution suggests that ”Year Built” can serve as a valuable predictor.

5 Baseline Model

While still incredibly early in our process, for our baseline model, given that we are attempting to predict emissions given some house, we selected to utilize a Lasso Regression on Polynomial Features. This approach plans on applying a polynomial regression that will capture non-linear relationships between our selected categorical variables and greenhouse gas emissions. From there, albeit that the regression is complex and slower than the Ridge alternative, the Lasso regularization will reduce scenarios of over-fitting by penalizing less important features and will also set certain coefficients to 0 in the event that we misdiagnosed the relevance of some predictor.

Essentially, we chose this model was chosen because it balances the flexibility of polynomial regression to capture non-linear patterns with Lasso's feature selection capability, helping to avoid over-fitting, and catch irrelevant predictors. This approach should provide a strong baseline for accurately predicting greenhouse gas emissions.