GHG Emissions in Cambridge Buildings

CS 109A Final Project

Sukhraj Dulay, Maitri Shah, Josh Zhang, Paula Zhuang

Table of Contents

01

Problem Statement

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?"

- **Dataset**: Cambridge Building Energy Use Disclosure Ordinance (BEUDO) Data 2022
  - Data about ***buildings and their properties*** in Cambridge (year built, square footage, type of building, residential vs. non-residential, etc.)
  - Data about ***resource usage and emissions*** (energy usage, water usage, greenhouse gas emissions, etc.)

02

EDA Visualizations

Data Cleaning

Total Records: 851 buildings

Cleaning Steps:

- Removed columns with > 50% missing values.
- Imputed missing values for Year Built and Property Area with means
  - Assuming averages which maintains any correlations within the variables

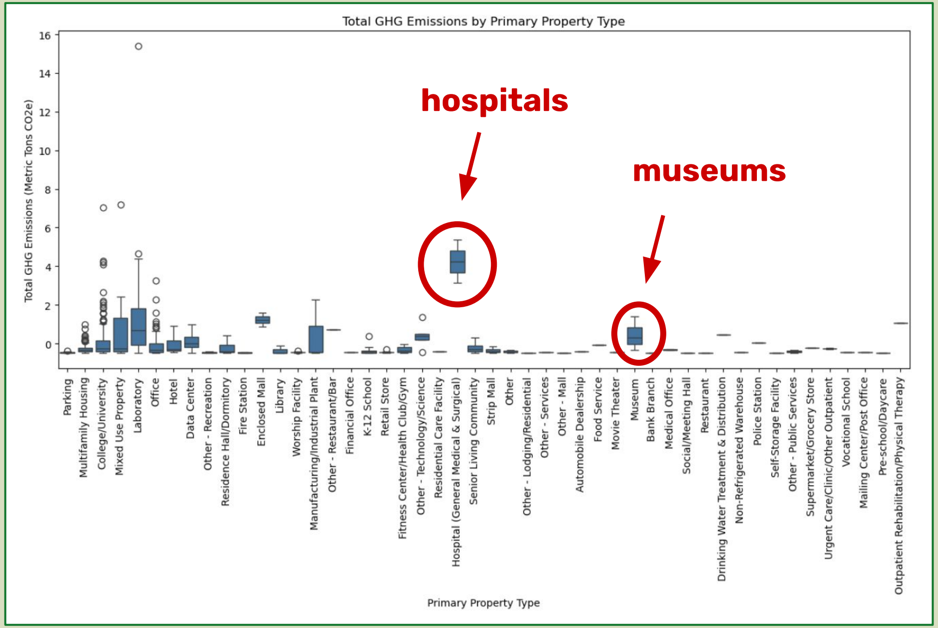
NEXT: EDA to determine good predictors

Total GHG Emissions vs. Property Type

Key Insights:

Total Greenhouse Gas (GHG) Emissions:

- Differs significantly by property type
- Higher emissions: Non-residential categories like hospitals and offices.
- Lower emissions: Residential and smaller properties.

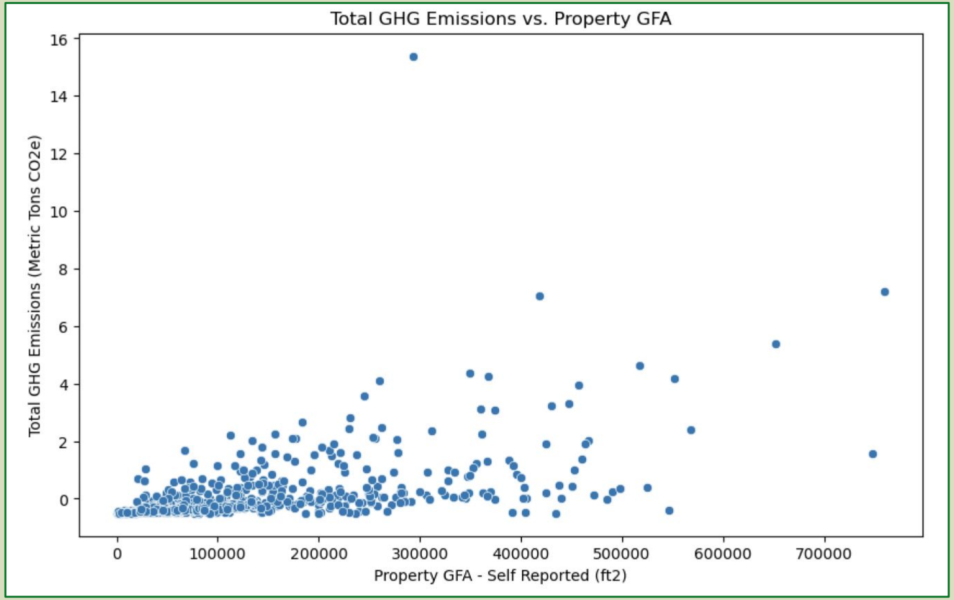


Total GHG Emissions vs. Property Area

Key Insights:

- Larger properties tend to have higher absolute emissions.
- Emissions vary more as property size increases

Outliers: Large buildings with unexpectedly low emissions likely use renewable energy.

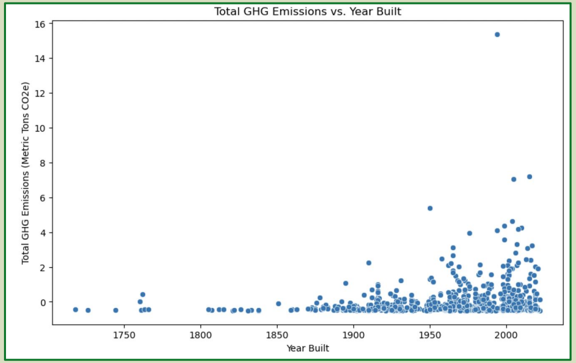


Total GHG Emissions vs. Year Built

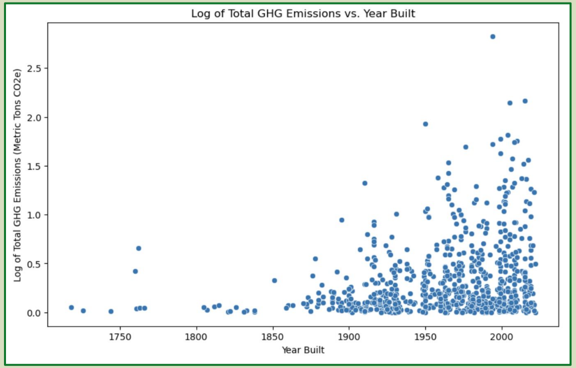
Key Insights:

- **Positive trend** for GHG emissions on average, but **increase in variability** as year built increases
- High concentration of low GHG emissions across all years
- **Weak overall correlation**: Building age alone is not a strong predictor.

Normal scale



Log scale



03

Baseline Model

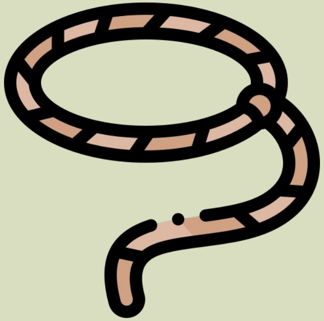
Baseline Model: Lasso Regression

Rationale

- Lasso assumes linear relationship (our relationships have a positive, almost linear trend)
- Shrinks coefficients of less relevant variables to zero → Provides an **interpretable model**

Predictors and Target

- **Predictors**: 'Property Area', 'Year Built', 'Dataset Category'
- **Target**: ‘Total GHG Emissions’



Baseline Model: Pipeline

Baseline Model Pipeline:

1. **Data preprocessing**:
   a. Imputation, scaling, and encoding.
2. **Model training**:
   a. GridSearchCV to find the best alpha value (alpha = 10^{-5})
   b. Train the Lasso algorithm
3. **Evaluation**: Assessing cross-validated RMSE and R^2 metrics

Baseline Model: Results & Analysis

Performance in Context

- **RMSE**: 0.76
    - Smaller than the standard deviation of the target variable (1.000629) and is a small fraction of the range (20.4) → model is performing reasonably well in predicting values
- **R^2**: 0.4
    - Pretty low, only explains 40% of the variability in the target variable
- **Key Predictors**: Property Area and Dataset Category
    - Makes sense: most clear and "linear" relationships to target variable

04

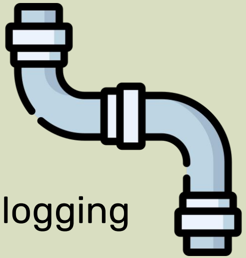
Final Model

Final Model: Random Forest

Steps to Improve Baseline Model

1. **Switch to Random Forest:** Can handle more complex, non-linear relationships; models interactions between predictors; reduces variance
2. **Feature Selection & Engineering**: Experimenting with adding an important feature from our original analysis, **'Primary Property Type'**, and adding interaction terms
3. **Taking Logs:** Experimenting with logging response variables and predictors to convert exponential trends into linear understandings
4. **Optimal Parameters**: Grid search to find best n_estimators (100-500)

Final Pipeline



Data preprocessing

→ Add combination of estimator, interaction terms, extra features, and logging

→ Training, & Evaluation

Final Model Pipeline - Results

Best Model:

450 Estimators, No Interaction Terms, Include Property Type, Log X and y

- **Scores**:
  - **OOB R^2 -** 0.690
  - **RMSE -** 0.373
  - **Train R^2 -** 0.861
- **Key predictors**:
  - Property Area (followed by Laboratory and Year Built)

Combination	Log_X	Log_Y	OOB R^2	RMSE	R^2	Best_Params
{'interaction_terms': False, 'include_property_type': False}	FALSE	FALSE	0.338106719550621	0.2832159947536660	0.9197887003156920	{'random_forest__n_estimators': 100}
{'interaction_terms': False, 'include_property_type': False}	TRUE	FALSE	0.3160021416595370	0.29059178417217900	0.9155564149716300	{'random_forest__n_estimators': 150}
{'interaction_terms': False, 'include_property_type': False}	FALSE	TRUE	0.5437126004395560	0.40248935784183800	0.8380023168240650	{'random_forest__n_estimators': 450}
{'interaction_terms': False, 'include_property_type': False}	TRUE	TRUE	0.5370711806825500	0.4045506177907470	0.8363387976451250	{'random_forest__n_estimators': 300}
{'interaction_terms': True, 'include_property_type': False}	FALSE	FALSE	0.3004761816986500	0.2901329058375100	0.9158228969502830	{'random_forest__n_estimators': 100}
{'interaction_terms': True, 'include_property_type': False}	TRUE	FALSE	0.30748118550667300	0.28923906610001000	0.9163407626415940	{'random_forest__n_estimators': 100}
{'interaction_terms': True, 'include_property_type': False}	FALSE	TRUE	0.520936600079633	0.4211339867580890	0.8226461651972380	{'random_forest__n_estimators': 450}
{'interaction_terms': True, 'include_property_type': False}	TRUE	TRUE	0.5263772679090560	0.408552690303497	0.8330846992457750	{'random_forest__n_estimators': 500}
{'interaction_terms': False, 'include_property_type': True}	FALSE	FALSE	0.44651043083281500	0.2668378385327570	0.9287975679271670	{'random_forest__n_estimators': 450}
{'interaction_terms': False, 'include_property_type': True}	TRUE	FALSE	0.45182379784099300	0.25555472414268900	0.934691782968354	{'random_forest__n_estimators': 150}
{'interaction_terms': False, 'include_property_type': True}	FALSE	TRUE	0.6884901024645230	0.3654923083820490	0.866415372513561	{'random_forest__n_estimators': 500}
{'interaction_terms': False, 'include_property_type': True}	TRUE	TRUE	0.6903692515713370	0.3726355494274600	0.8611427473028950	{'random_forest__n_estimators': 450}
{'interaction_terms': True, 'include_property_type': True}	FALSE	FALSE	0.44801780551072400	0.26222575383646800	0.9312219195796660	{'random_forest__n_estimators': 400}
{'interaction_terms': True, 'include_property_type': True}	TRUE	FALSE	0.4419559488864530	0.26742271662480300	0.9284850906330100	{'random_forest__n_estimators': 450}
{'interaction_terms': True, 'include_property_type': True}	FALSE	TRUE	0.6796364017342340	0.3749816289958010	0.8593887779156550	{'random_forest__n_estimators': 450}
{'interaction_terms': True, 'include_property_type': True}	TRUE	TRUE	0.6752541933591420	0.3755284757102870	0.8589783639307090	{'random_forest__n_estimators': 450}

Final Model Pipeline - Analysis

Scores

- The **OOB R^2** (0.690) recognizes that the model is statistically significant in terms of generalizability against unseen data
- The **RMSE** (0.373) infers our model is interpreting trends with minimal error
- The train **R^2** (0.861) shows that our model is capable of explaining 86.1% of the variability of emissions

05

Results/Inferences

Model Performance Comparison

	Lasso Regression	**Random Forest**
R^2	0.40	0.69
RMSE	0.76	0.37
Key Predictors	Property Area, Dataset Category	Property Area, Laboratory, Year Built

Results and Inferences

- **Larger buildings** are more energy-intensive
- **Non-linear relationships** between predictors and GHG emissions, because:
  - Log-transformation of predictors improved the model
  - Random forest performed much better than Lasso regression
- **Non-residential buildings** are primary contributors to emissions, especially certain energy-intensive property types (like laboratories)
- **Building age** offers information when interactions with other features are considered, indicating that factors like expansion or usage change may be impactful

06

Future Work

Future Work / Next Steps

Actionable Outcomes:

- Targets large non-residential buildings and energy-intensive property types for emission reduction initiatives.
- Focus efforts on a small number of outliers, due to non-linearity of problem

Future Work:

- Individually log or square root every combination of predictor variables to account for exponential or logarithmic variable trends in the final model.
- Compare against other water or energy variables, such as water or electricity usage, in Cambridge buildings.

GHG Emissions in Cambridge Buildings

CS 109A Final Project

Sukhraj Dulay, Maitri Shah, Josh Zhang, Paula Zhuang

Characteristics vs. Property Type

Key Insights:

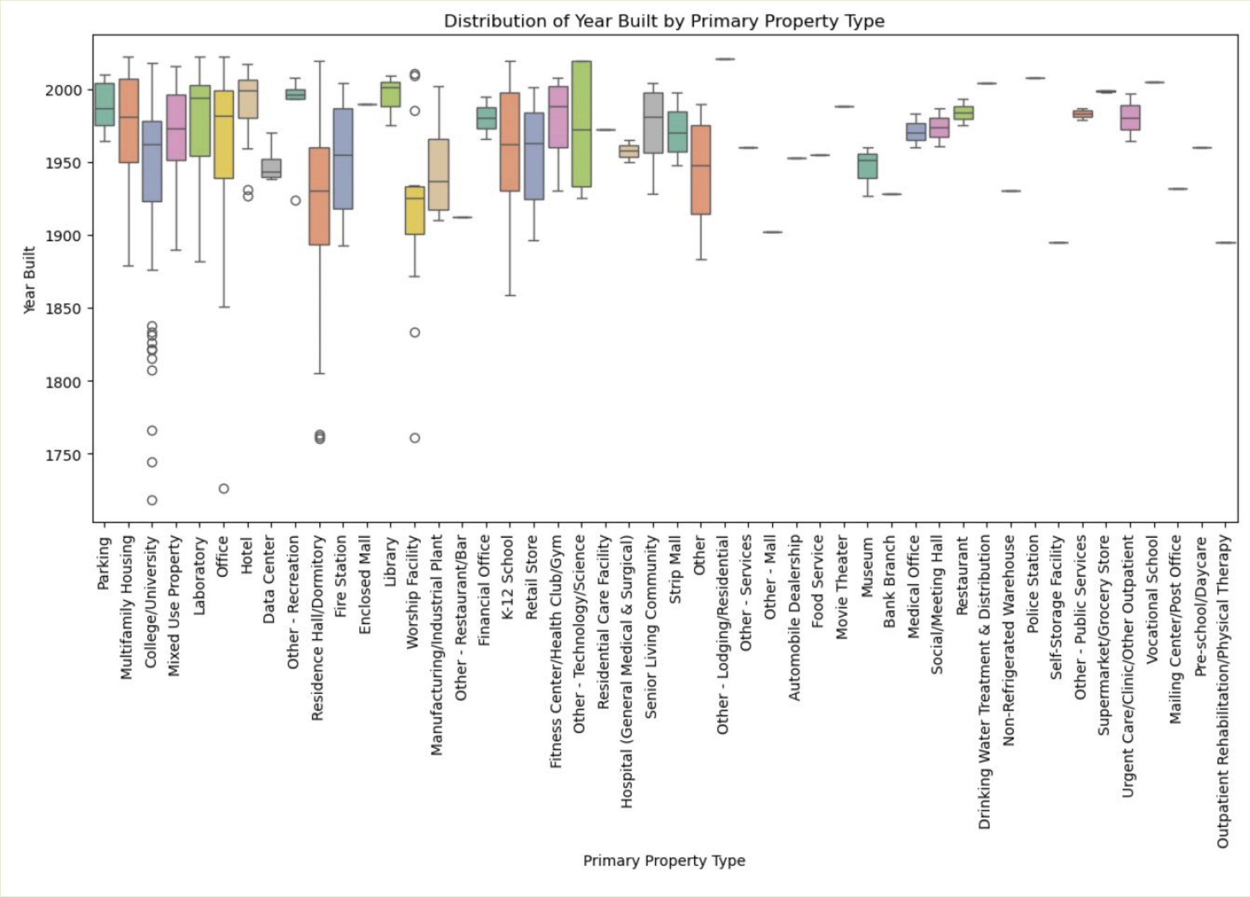
- Property floor area varies widely across property types.
  - Large properties: College/University, Enclosed Mall.
  - Smaller properties: Bank Branch, Medical Offices.

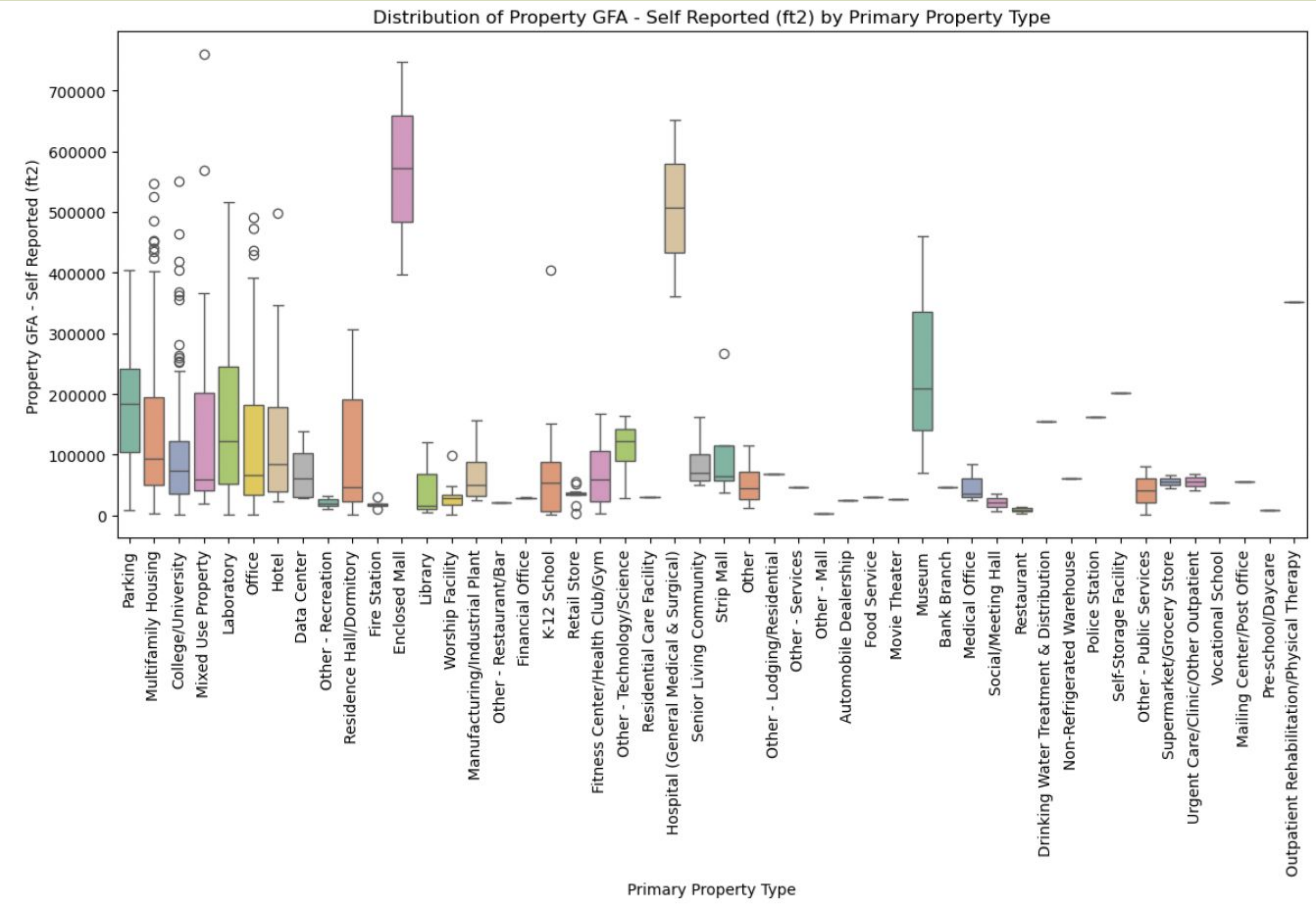
Year Built trends:

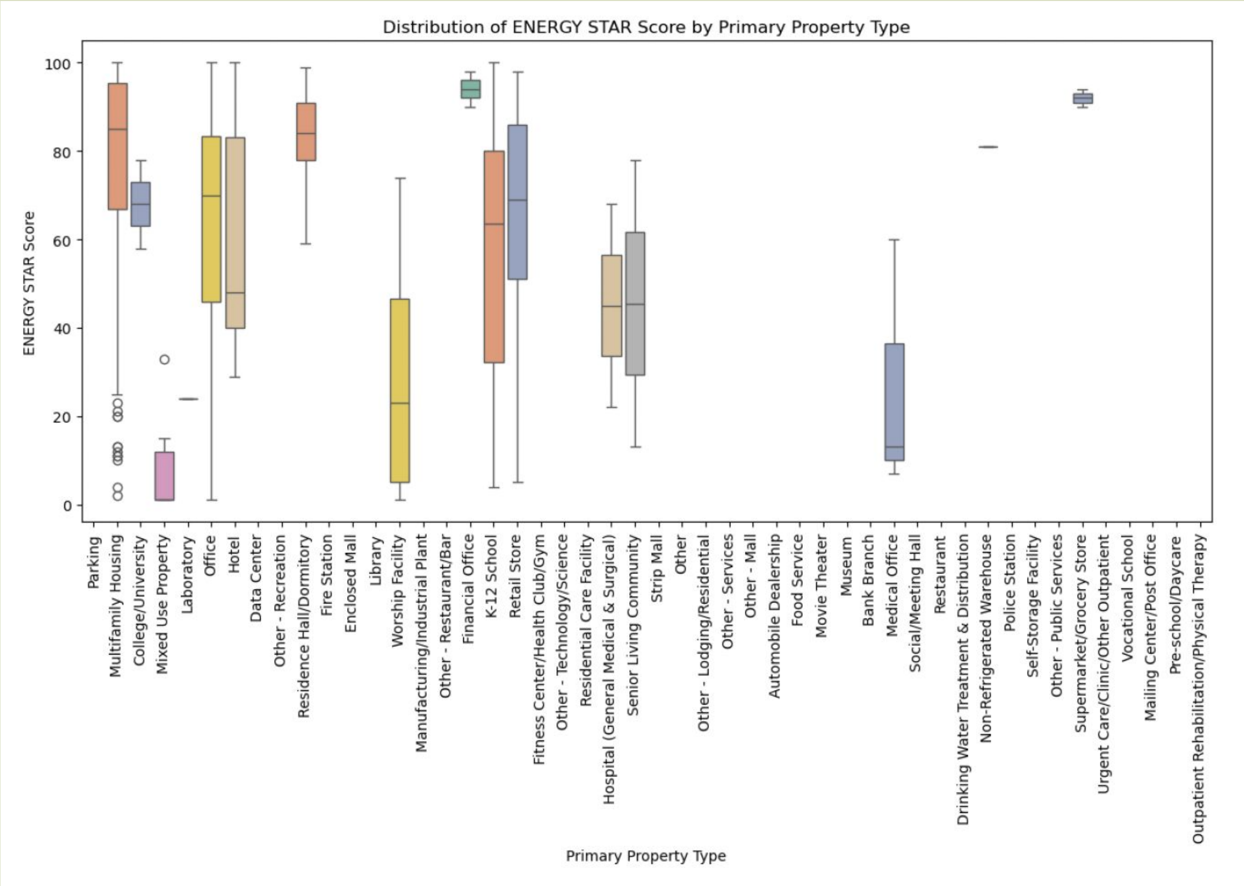
- Older: Religious worship facilities, universities.
- Newer: Hotels, recreation facilities.

ENERGY STAR Scores:

- High: Residence Halls/Dormitories, Financial Offices.
- Low: Mixed-Use Properties, Medical Offices.


Distribution of Year Built by Primary Property Type
Year Built
2000
1950
1900
1850
1800
1750
Primary Property Type
Parking
Multifamily Housing
College/University
Mixed Use Property
Laboratory
Office
Hotel
Data Center
Other - Recreation
Residence Hall/Dormitory
Fire Station
Enclosed Mall
Library
Worship Facility
Manufacturing/Industrial Plant
Other - Restaurant/Bar
Financial Office
K-12 School
Retail Store
Fitness Center/Health Club/Gym
Other - Technology/Science
Residential Care Facility
Hospital (General Medical & Surgical)
Senior Living Community
Strip Mall
Other
Other - Lodging/Residential
Other - Services
Other - Mall
Automobile Dealership
Food Service
Movie Theater
Museum
Bank Branch
Medical Office
Social/Meeting Hall
Restaurant
Drinking Water Treatment & Distribution
Non-Refrigerated Warehouse
Police Station
Self-Storage Facility
Other - Public Services
Supermarket/Grocery Store
Urgent Care/Clinic/Other Outpatient
Vocational School
Mailing Center/Post Office
Pre-school/Daycare
Outpatient Rehabilitation/Physical Therapy


Distribution of Property GFA - Self Reported (ft2) by Primary Property Type
Property GFA - Self Reported (ft2)
0
100000
200000
300000
400000
500000
600000
700000
Parking
Multifamily Housing
College/University
Mixed Use Property
Laboratory
Office
Hotel
Data Center
Other - Recreation
Residence Hall/Dormitory
Fire Station
Enclosed Mall
Library
Worship Facility
Manufacturing/Industrial Plant
Other - Restaurant/Bar
Financial Office
K-12 School
Retail Store
Fitness Center/Health Club/Gym
Other - Technology/Science
Residential Care Facility
Hospital (General Medical & Surgical)
Senior Living Community
Strip Mall
Other
Other - Lodging/Residential
Other - Services
Other - Mall
Automobile Dealership
Food Service
Movie Theater
Museum
Bank Branch
Medical Office
Social/Meeting Hall
Restaurant
Drinking Water Treatment & Distribution
Non-Refrigerated Warehouse
Police Station
Self-Storage Facility
Other - Public Services
Supermarket/Grocery Store
Urgent Care/Clinic/Other Outpatient
Vocational School
Mailing Center/Post Office
Pre-school/Daycare
Outpatient Rehabilitation/Physical Therapy
Primary Property Type


Distribution of ENERGY STAR Score by Primary Property Type
ENERGY STAR Score
100
80
60
40
20
0
Parking
Multifamily Housing
College/University
Mixed Use Property
Laboratory
Office
Hotel
Data Center
Other - Recreation
Residence Hall/Dormitory
Fire Station
Enclosed Mall
Library
Worship Facility
Manufacturing/Industrial Plant
Other - Restaurant/Bar
Financial Office
K-12 School
Retail Store
Fitness Center/Health Club/Gym
Other - Technology/Science
Residential Care Facility
Hospital (General Medical & Surgical)
Senior Living Community
Strip Mall
Other
Other - Lodging/Residential
Other - Services
Other - Mall
Automobile Dealership
Food Service
Movie Theater
Museum
Bank Branch
Medical Office
Social/Meeting Hall
Restaurant
Drinking Water Treatment & Distribution
Non-Refrigerated Warehouse
Police Station
Self-Storage Facility
Other - Public Services
Supermarket/Grocery Store
Urgent Care/Clinic/Other Outpatient
Vocational School
Mailing Center/Post Office
Pre-school/Daycare
Outpatient Rehabilitation/Physical Therapy
Primary Property Type