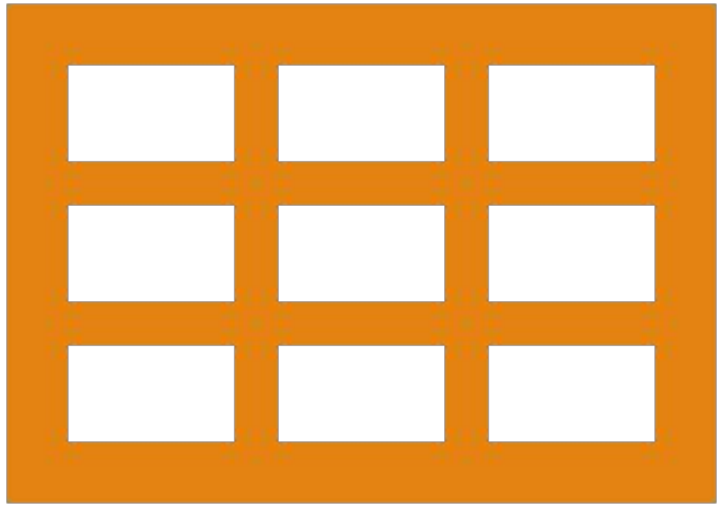
# Cambridge Energy Use

Authors:

Jie Zhao, Michael Assmus, Mohanish Kashiwar



# Agenda

- Problem Statement
- Visualization Analysis
- Data processing
- Modeling and Results
- Conclusion/ Inferences
- Future Work/ Scope of improvement

# Project Overview

This project investigates energy consumption patterns in residential buildings throughout the City of Cambridge. Using publicly available datasets on building characteristics and energy usage, our goal was to:

- Understand how building attributes influence energy efficiency.
- Identify key predictors of high or low energy use.
- Build predictive models for Weather-Normalized Site Energy Use Intensity (EUI).
- Provide actionable insights to improve energy performance.

We focused on **residential properties**, given data complexity and time constraints.

# Problem Statement

Energy consumption in urban residential buildings is influenced by several factors—including building age, size, construction materials, and maintenance conditions. However, these relationships are often non-linear and difficult to identify without robust data analysis.

This study aims to:

- Explore patterns in Cambridge's building and energy datasets.
- Model EUI using a range of statistical and machine learning methods.
- Highlight building characteristics associated with better energy performance.

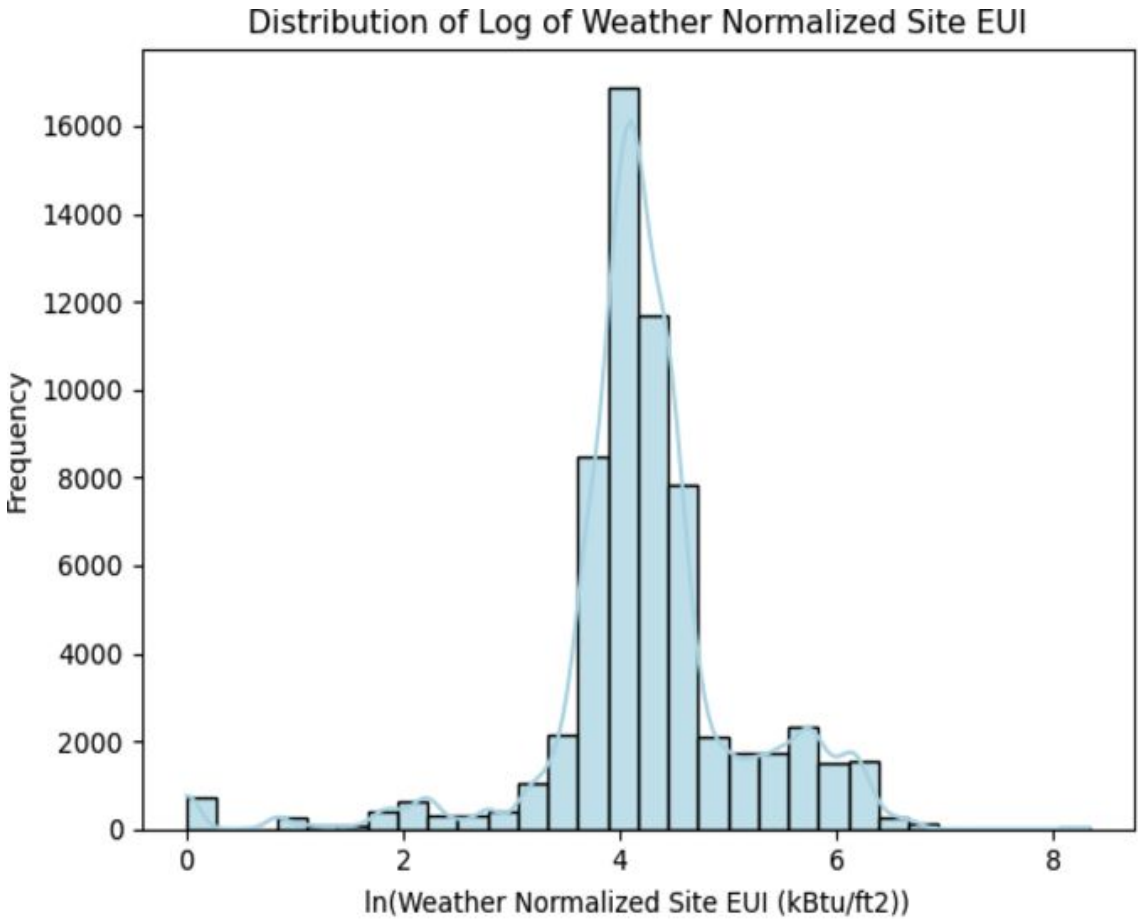
# Visualization/EDA/Analysis

Cambridge **Property Database**: contains physical details of the properties at the building or unit level, including wall types, interior living space,and much more.

Cambridge **Building Energy Use** data**:** contains energy and water use data at the parcel level.

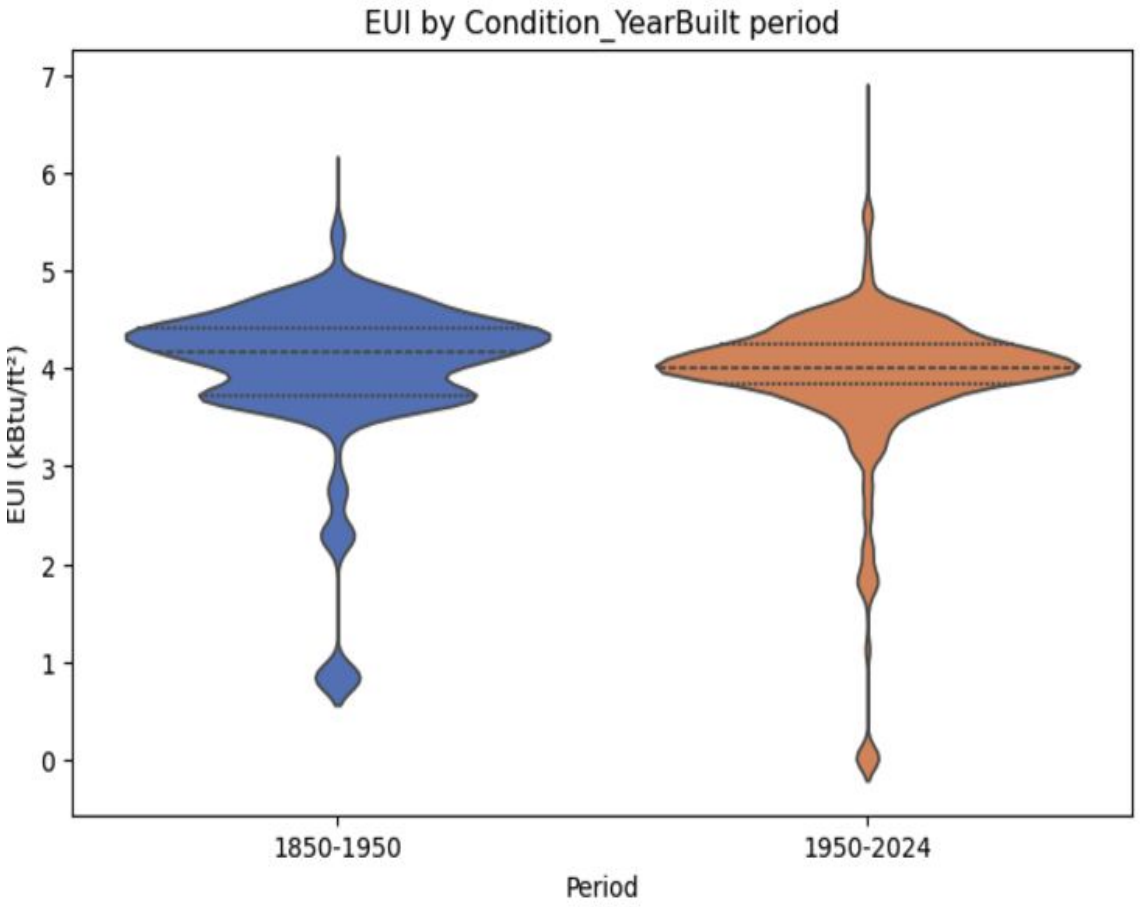
We merged the two datasets by lot and data year for analysis.

# Visualization/EDA/Analysis




- **Response variable**: Y = Weather-Normalized Site Energy Use Intensity (EUI), measured in kBtu/ft²
- Log transformed y closely follows a normal distribution. Majority values are around 4000.

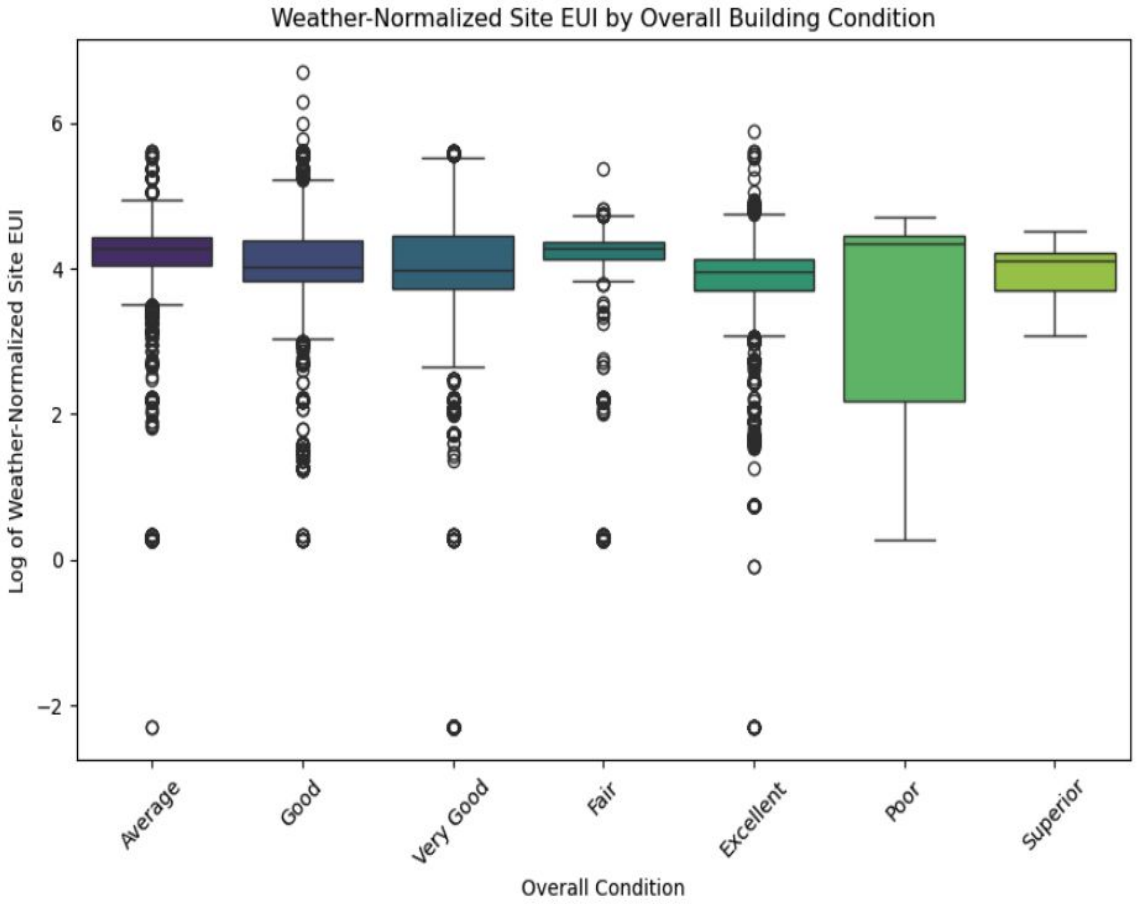
# Visualization/EDA/Analysis




- Strong relationship between **numerical features** and Weather normalized Site EUI:
  - Property GFA self reported (ft)
  - Bedroom number
  - Condition_YearBuilt ( example in chart)
  - PD parcel living area

- An increase energy use from 1850 -1950, and a decrease pattern from 1950 to 2024. Older building has higher energy consumptions.

# Visualization/EDA/Analysis




- Strong relationship between **categorical features** and Weather normalized Site EUI.
  - Primary property type
  - Exterior Wall Type
  - Exterior_roofmaterial
  - Overall Condition ( example in chart)

- Buildings in "Superior" condition appear to have better energy efficiency and less variation in performance. Buildings in "Poor" condition show high variability, which could suggest inconsistent energy management.

# Data processing

**Data Challenges**

- Missing data across many columns.
- Non-standardized categorical features.
- Different levels of aggregation between datasets (building vs lot).

**Solutions Implemented**

- Imputed missing values using within-lot mean/mode or similar-property aggregates.
- Standardized categorical values by manually cleaning inconsistent entries.
- Re-aggregated datasets to the lot level using weighted metrics (e.g., interior living space).

These steps produced a clean, analysis-ready dataset suitable for modeling.

# Modeling and Results

## Baseline

- Goal: Simple linear regression model to set baseline
- Limitations: Weak predictive power, struggles to capture nonlinear relationships.
- Result: Poor fit—unable to capture non-linear relationships.
-

**$R^2$ : 0.125**

**MSE: 0.638**

## Polynomial

- Goal: Capture non-linear relationships suggested by the residual patterns
- Limitation: No meaningful improvement from increased complexity, suggesting a linear model is not appropriate.

**$R^2$ : 0.125**

**MSE: 0.638**

## Gradient boosting

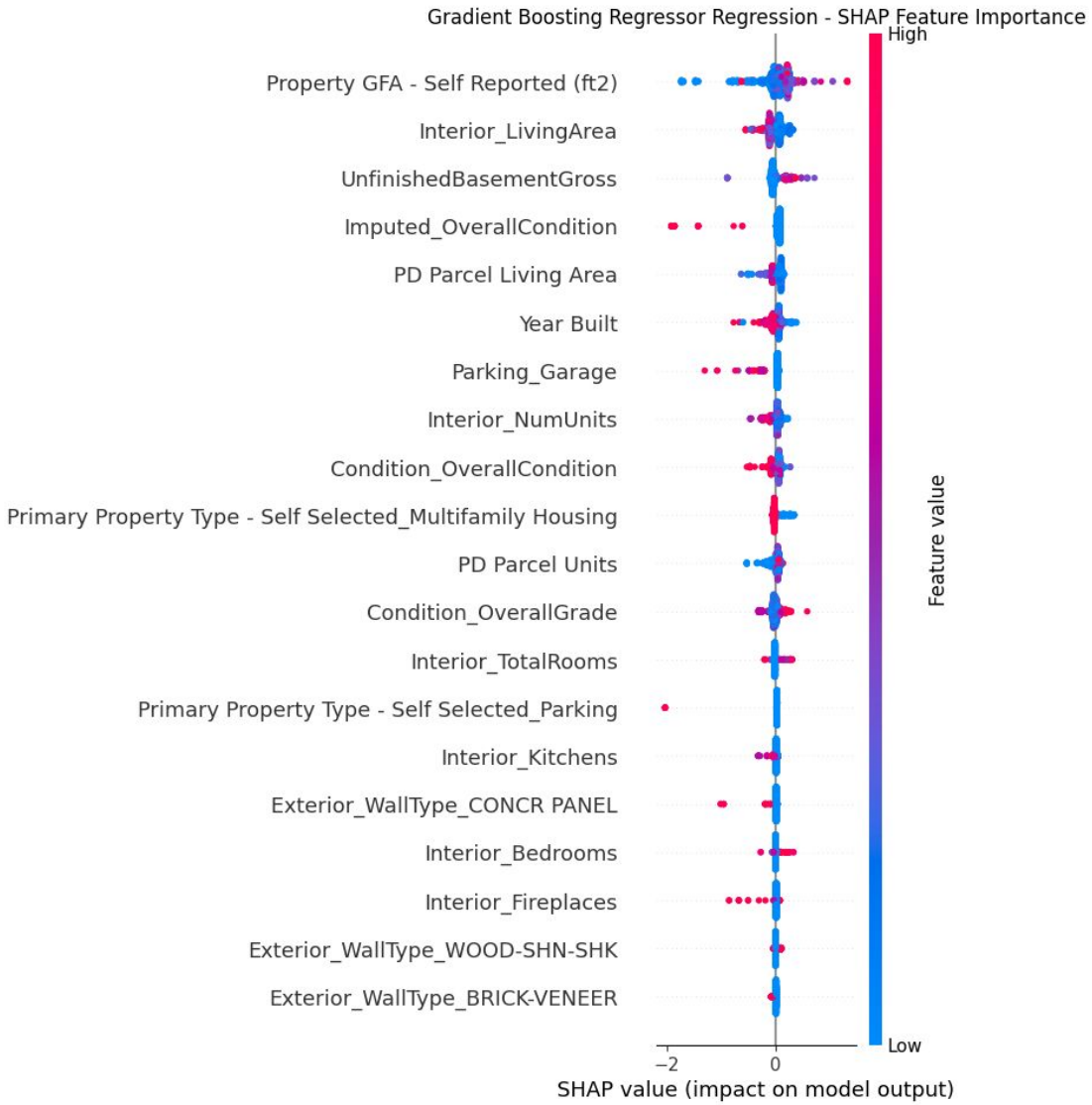
- Goal: Capture complex relationships without assumptions about data distribution.
- Result: Strong predictive performance and robustness.


**$R^2$ : 0.740**

**MSE: 0.190**

**Gradient boosting significantly outperformed traditional models, indicating that energy efficiency is shaped by numerous interacting factors that are not linearly separable.**

# Conclusion and inference




- **Building condition matters more than individual material choices.** Well-maintained buildings are consistently more efficient.
- **Modern buildings (post-1950) tend to have lower EUI**, likely due to better insulation standards and building codes.
- **Unfinished basements correlate with higher energy consumption**, suggesting heat loss or poor insulation.
- **Material types show trends but limited predictive power** when controlling for other factors.

# Future Work / Scope of Improvement

- **Include commercial & mixed-use buildings:** Requires additional feature engineering.
- **Incorporate more predictors:** HVAC types, insulation ratings, renovation history.
- **Improve imputation:** Use model-based or KNN imputation for better accuracy.
- **Explainability:** Use SHAP values to interpret XGBoost model behavior.
- **Data refresh:** Cambridge regularly publishes updated energy data for more recent analysis.

# Technical Deep Dive

**End-to-End Data Pipeline**

- **Ingestion:** Raw CSVs from city-open data sources and PDF audits.
- **Normalization:** Standardized categorical fields using controlled vocabulary mapping.
- **Feature Engineering:**
  - Log-transform of EUI and skewed numeric features.
  - Polynomial interactions for age × condition.
  - Derived features: decade built, maintenance score, finished basement flag.
  - One-hot encoding for materials, roof types, and heating systems.
- **Modeling Pipeline:**
  - Scikit-learn pipeline with transformers + model objects.
  - 10-fold cross-validation with stratification by property type.
  - Hyperparameter tuning via randomized search (learning rate, depth, gamma, subsample).

**XGBoost Model Specification**

- **Objective:** `reg:squarederror`
- **Tree depth:** 6–10
- **Learning rate:** 0.05–0.3
- **Boosting rounds:** ~300
- **Subsample / colsample_bytree:** 0.6–0.9
- **Regularization:** L1 (alpha) = 0.1–1.0, L2 (lambda) = 1–3
- **Evaluation:** RMSE + $R^2$

**Explainability Techniques (Planned & Partial)**

- **Permutation importance** to identify global feature influence.
- **SHAP values** for:
  - Local explanations of individual residential parcels.
  - Global interaction maps (e.g., age × condition).
- **EUI response curves** for monotonicity checks.

# Thank you