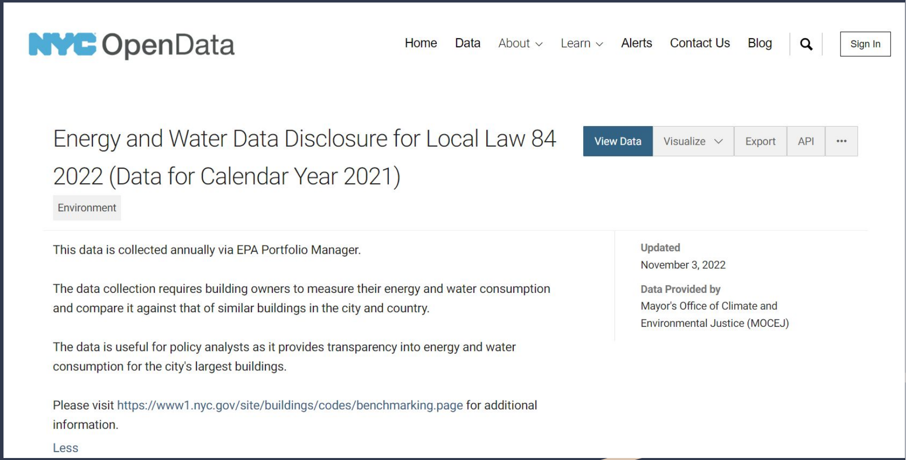
Energy Use Efficiency Prediction for Multi-family Homes in NYC

5291 Project

Group 11:
- tl3184 Tianqi Liu
- yw3946 Yiheng Wu
- yc3210 Siwei Chen

Objective


NYC OpenData

Home Data About Learn Alerts Contact Us Blog Sign In

Energy and Water Data Disclosure for Local Law 84 2022 (Data for Calendar Year 2021)

View Data Visualize Export API

Environment

This data is collected annually via EPA Portfolio Manager.

The data collection requires building owners to measure their energy and water consumption and compare it against that of similar buildings in the city and country.

The data is useful for policy analysts as it provides transparency into energy and water consumption for the city's largest buildings.

Please visit https://www1.nyc.gov/site/buildings/codes/benchmarking.page for additional information.

Less

Updated
November 3, 2022

Data Provided by
Mayor's Office of Climate and Environmental Justice (MOCEJ)

With over 70 percent of the city's emissions stemming from buildings, New York City mandate larger buildings to publicly disclose their energy and water consumption data, committing to reducing greenhouse gas emissions and minimizing its ecological footprint.

We aim to leverage the NYC building data from 2021 to understand what leads to higher or lower energy score, and to provide insights and recommendations that promote sustainability in the city.

Dataset Overview

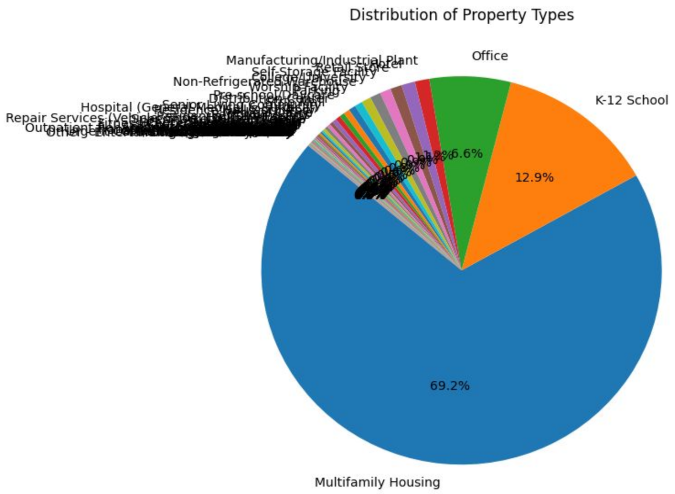
Energy and Water Consumption - Buildings for 2021 (NYC Open Data)

https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/7x5e-2fxh

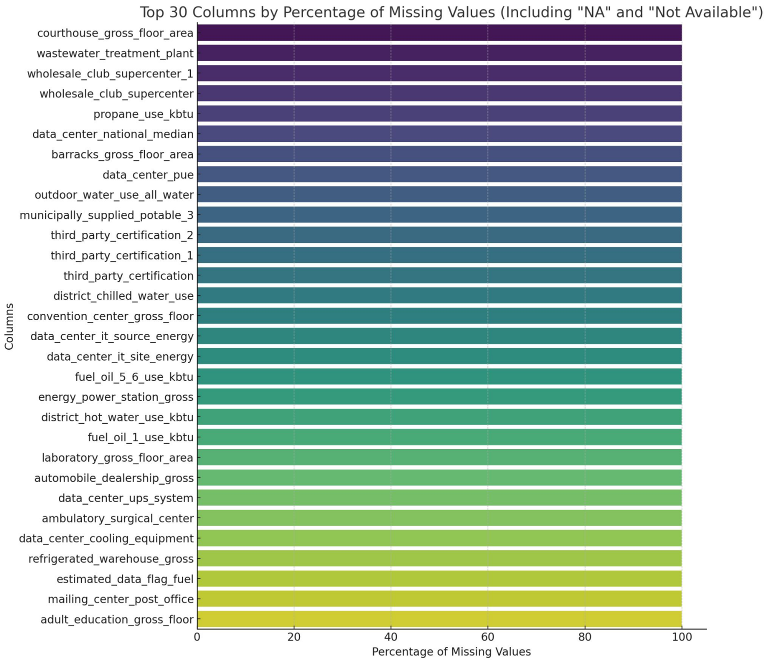
- Extracted 10,000 records from API
- 249 variables
  - Temporal data
  - Geographic data
  - Energy efficiency metrics
  - Building features data

Our Approach:

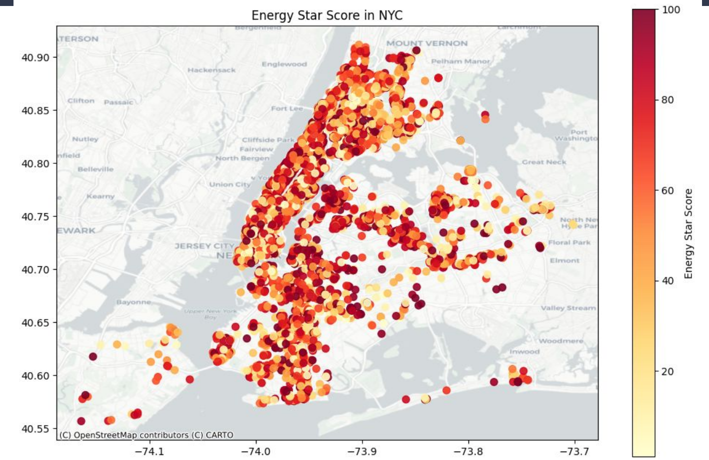
- Use “Energy Star Score” as target variable
- Focusing on multi-family housing (~70% of the data)

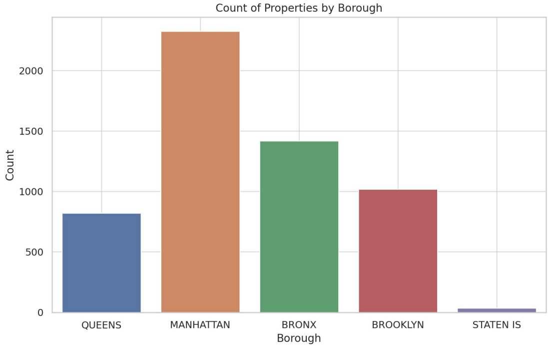


EDA

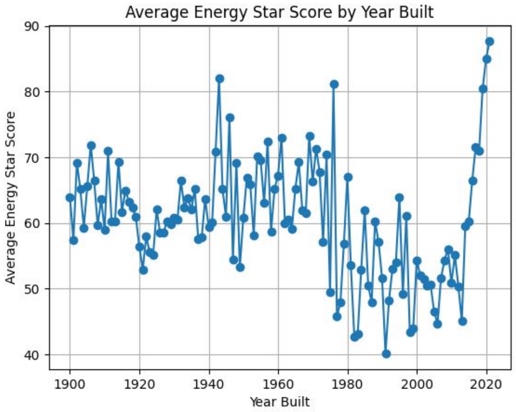


We visualize the proportion of null values for each column in the dataset





Most buildings are located in Manhattan, but Staten Island only have a few



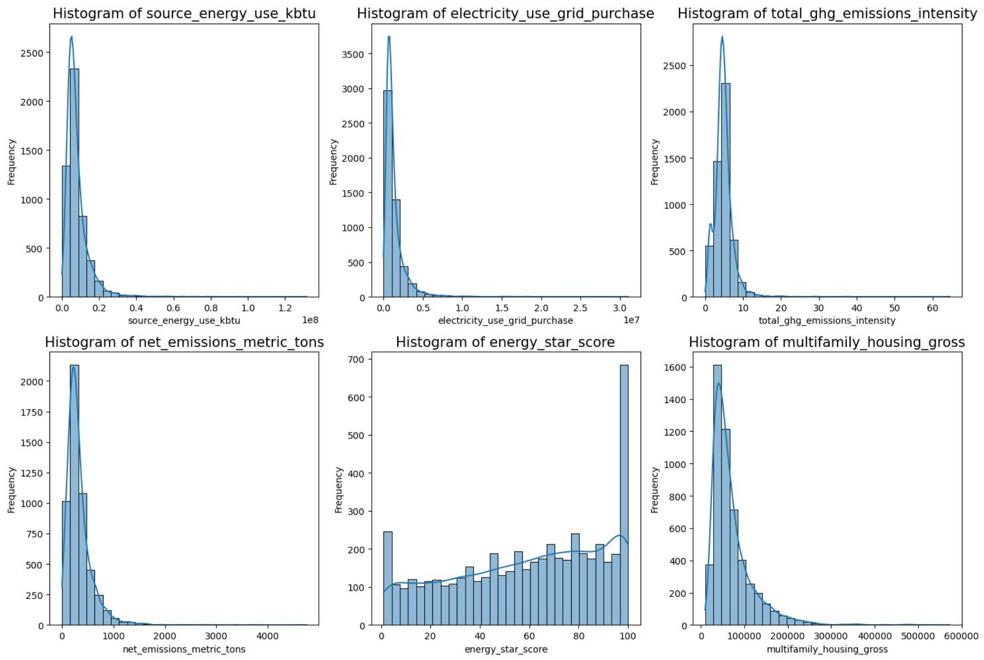
We focus on buildings built after 1900, and found that the energy star scores range from 60 to 70 and then increase steeply in recent years

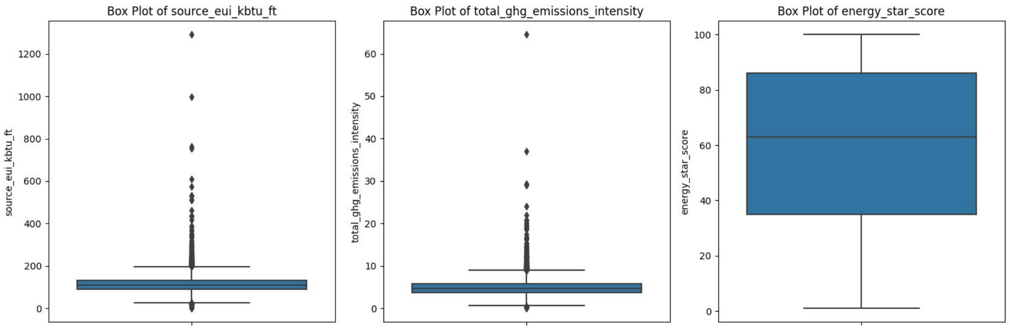
EDA

From the histograms, we can see that many variables are highly right-skewed

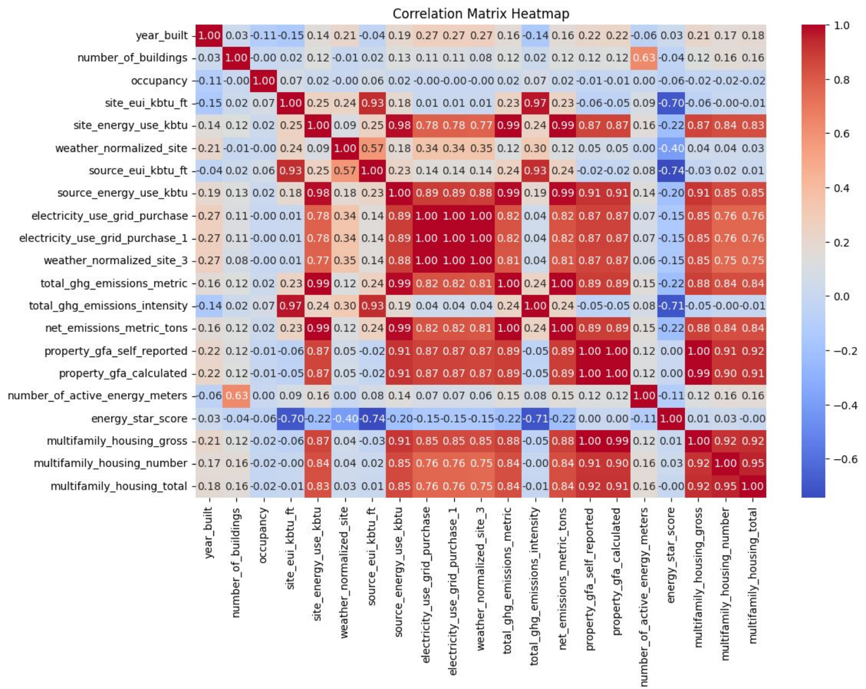
For the box plots, there are data points outside of the lower and upper whiskers, which is considered as outliers

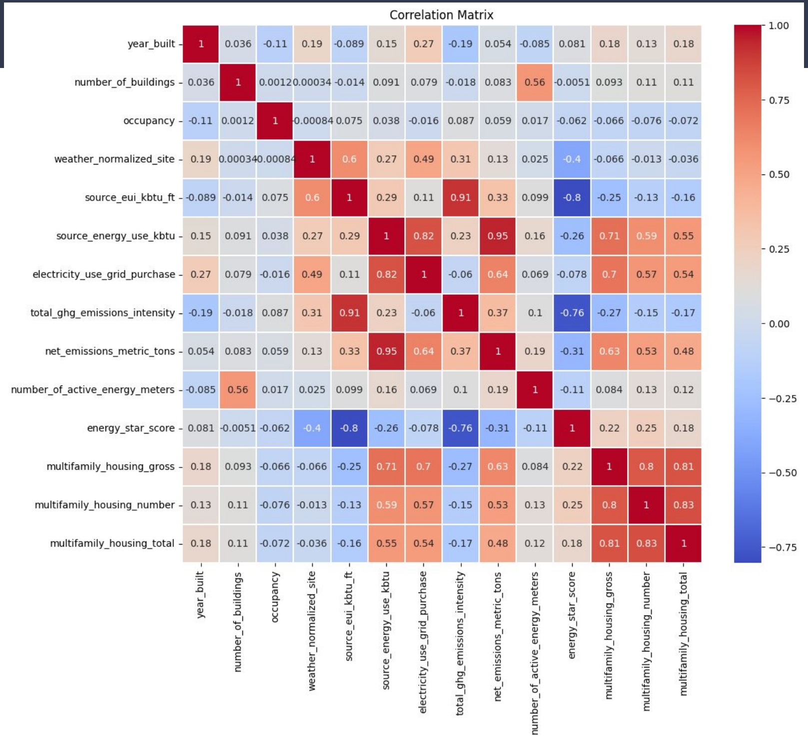
→ Data transformation needed





Correlation Matrix





Before:

Many columns are highly correlated (has a correlation > 0.9) → multicollinearity

After:

Most columns with strong correlations are removed after feature selection

Feature Selection

1. Manually removed irrelevant columns
   - Irrelevant for modelling analyses (ex. ID, names, dates etc.)
   - Don't apply to multi-family housing type
2. Removed highly-correlated columns with similar definitions

3. Filter out columns with mostly missing values

   - Remove columns that have > 50% values that are NAs, "not available" or "not applicable"

→ Reduced to 12 final independent variables (11 numeric & 1 categorical)

	year_built	weather_normalized_site	source_eui_kbtu_ft	source_energy_use_kbtu	electricity_use_grid_purchase	total_ghg_emissions_intensity	net_emissions_metric_tons	number_of_active_energy_meters	energy_star_score
0	2010	1.419371	11.258537	25.719439	19.468447	2.089670	7.271616	8	71
4	1941	1.304555	12.741055	29.673502	21.080516	2.696013	11.204080	11	24
5	1982	1.304555	11.891940	30.103005	21.664608	2.413652	11.563996	11	57
6	1983	1.345407	12.375122	30.078545	21.769757	2.535629	11.507735	14	54
7	1958	1.259551	12.778394	28.719218	20.193318	2.718034	10.322686	8	42

multifamily_housing_gross	multifamily_housing_number	multifamily_housing_total	borough_BRONX	borough_BROOKLYN	borough_MANHATTAN	borough_QUEENS	borough_STATEN IS
2.512112	2.560942	2.379949	0	0	0	1	0
2.528850	3.139992	3.004172	0	0	1	0	0
2.532547	3.381578	2.789633	0	0	1	0	0
2.531342	3.393395	2.731356	0	0	1	0	0
2.524448	3.262529	2.673200	0	0	1	0	0

Data Preprocessing

1. Addressing null values/missing data

   Fill NA's of numerical columns with average values for regression modeling

2. Removing Outliers

   Used the IQR method to remove outliers

3. One-hot encode categorical variable

   We apply one-hot encoding on the "Borough" variable

4. Transforming skewed data

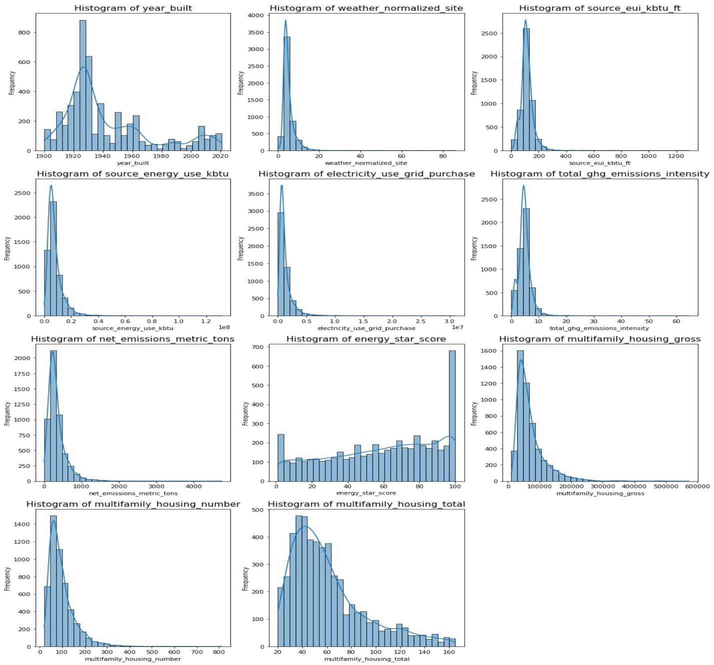
   Data distribution was highly-skewed, therefore, we used Box-Cox transformation for positive variables and Yeo-Johnson transformation for variables that might contain zero or negative values.

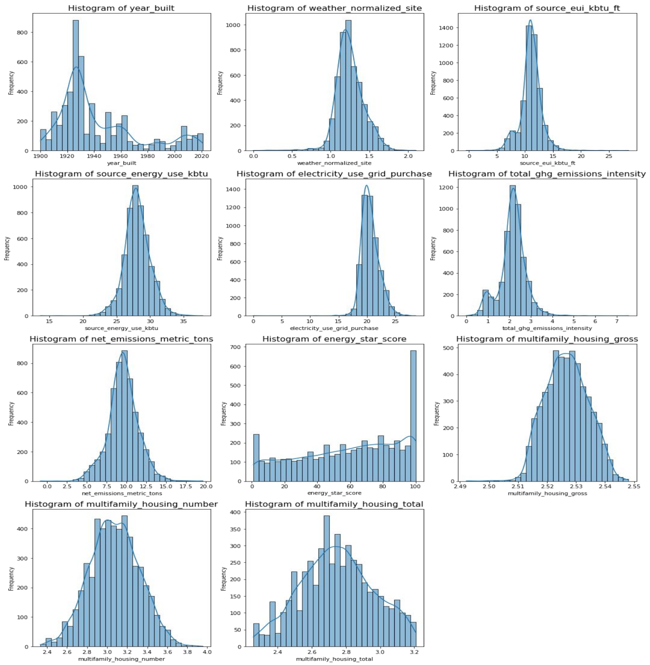
5. Filter data based on property type and year built

   We focus on analyzing the energy usage for multi-family homes built after 1900 as it makes up the majority of the dataset.

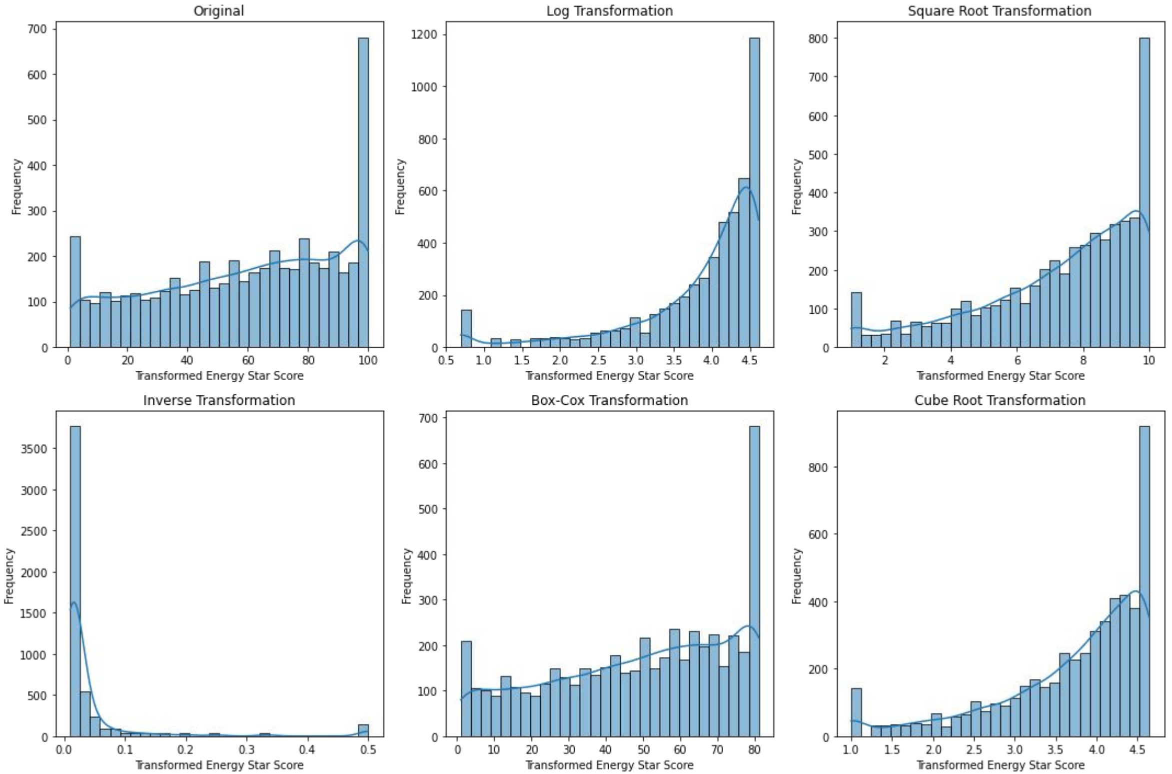
Data Modeling

Checking Distributions for Variables





Transformations for Independent Variable



Transformation	Skewness
original	-0.34049
log	-1.84766
sqrt	-0.94211
inverse	4.310233
boxcox	-0.39152
cbrt	-1.24902

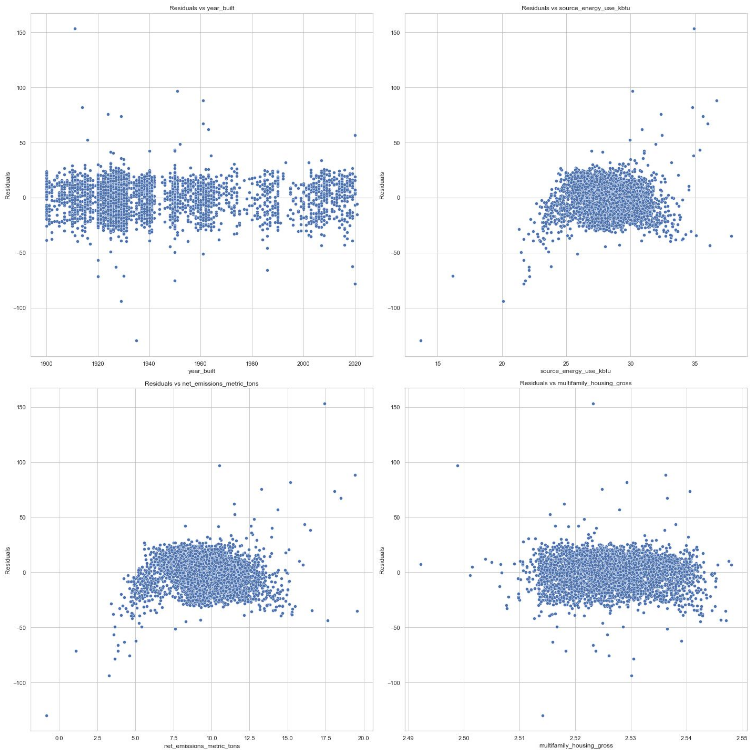
Linear Regression (OLS)

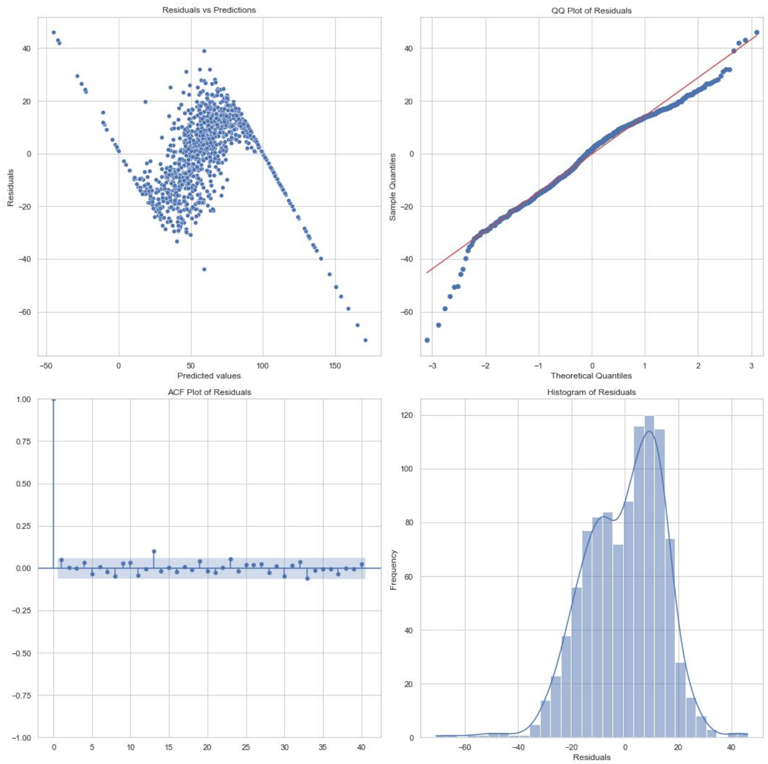
R-squared: 0.7671723306132332
Mean Squared Error: 210.81282319348946

	coef	std err	t	P>|t|	[0.025	0.975]
const	3644.4342	413.238	8.819	0.000	2834.264	4454.605
year_built	0.0154	0.009	1.777	0.076	-0.002	0.032
weather_normalized_site	-35.4729	4.892	-7.251	0.000	-45.064	-25.882
source_eui_kbtu_ft	-1.1937	1.510	-0.791	0.429	-4.153	1.766
source_energy_use_kbtu	-15.5419	2.488	-6.247	0.000	-20.420	-10.664
electricity_use_grid_purchase	4.0989	0.902	4.546	0.000	2.331	5.867
total_ghg_emissions_intensity	-37.3668	4.381	-8.529	0.000	-45.956	-28.777
net_emissions_metric_tons	13.8272	2.102	6.579	0.000	9.707	17.948
number_of_active_energy_meters	-0.2259	0.033	-6.940	0.000	-0.290	-0.162
multifamily_housing_gross	-1350.0961	173.369	-7.787	0.000	-1689.993	-1010.199
multifamily_housing_number	58.2093	1.588	36.650	0.000	55.096	61.323
multifamily_housing_total	-8.8080	1.832	-4.809	0.000	-12.399	-5.217
borough_BRONX	0.5733	1.688	0.340	0.734	-2.736	3.882
borough_BROOKLYN	-1.3302	1.721	-0.773	0.440	-4.704	2.044
borough_MANHATTAN	3.3159	1.672	1.984	0.047	0.039	6.593
borough_QUEENS	1.8862	1.745	1.081	0.280	-1.535	5.308
borough_STATEN IS	-0.9383	3.436	-0.273	0.785	-7.674	5.797

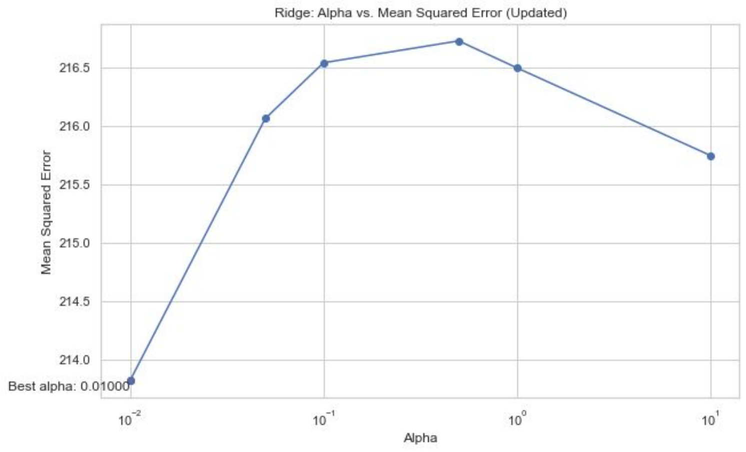
Omnibus:	542.899	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6392.079
Skew:	-0.126	Prob(JB):	0.00
Kurtosis:	9.091	Cond. No.	3.78e+06

Assumption Checks





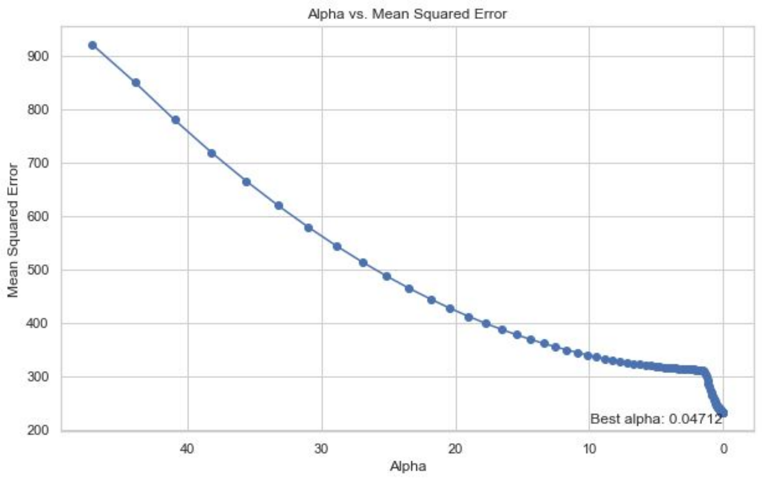
Ridge Regression




```
Best Alpha: 0.01
R-squared_adj: 0.7638531761719999
Mean Squared Error: 213.81813746826862
```

	Ridge_coefficients		Ridge_coefficients
multifamily_housing_gross	-568.857176	borough_MANHATTAN	3.367479
multifamily_housing_number	57.850389	source_eui_kbtu_ft	2.498841
total_ghg_emissions_intensity	-43.818591	borough_QUEENS	1.881426
weather_normalized_site	-32.128679	borough_BROOKLYN	-1.471847
source_energy_use_kbtu	-21.560042	borough_STATEN IS	-0.591514
net_emissions_metric_tons	17.077785	borough_BRONX	0.460186
multifamily_housing_total	-10.190155	number_of_active_energy_meters	-0.228219
electricity_use_grid_purchase	3.740516	year_built	0.015969

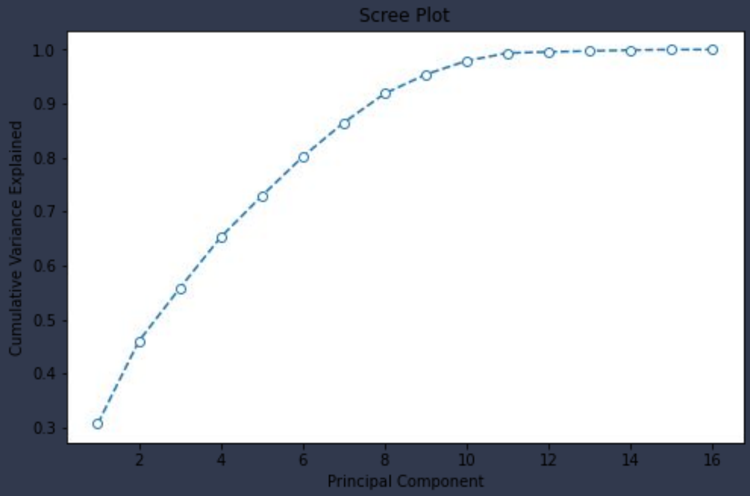
Lasso Regression



Best Alpha: 0.04712096670553455
R-squared_adj: 0.7608761083143419
Mean Squared Error: 216.5137109006037

	Lasso_coefficients		Lasso_coefficients
multifamily_housing_number	53.774012	borough_BROOKLYN	-1.661907
total_ghg_emissions_intensity	-21.955554	borough_QUEENS	0.991773
source_energy_use_kbtu	-11.692631	number_of_active_energy_meters	-0.228087
multifamily_housing_total	-8.262796	electricity_use_grid_purchase	-0.100736
weather_normalized_site	-8.111101	year_built	0.01517
net_emissions_metric_tons	7.177412	multifamily_housing_gross	0
source_eui_kbtu_ft	-3.875526	borough_BRONX	0
borough_MANHATTAN	3.090924	borough_STATEN IS	0

Principal Component Analysis (PCA & PCR)




```
R-squared_adj: 0.7140688219562823
Mean Squared Error: 258.8951692949533
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	59.3295	0.267	222.572	0.000	58.807	59.852
x1	0.0354	0.009	4.035	0.000	0.018	0.053
x2	-0.5541	0.036	-15.335	0.000	-0.625	-0.483
x3	5.3092	0.085	62.539	0.000	5.143	5.476
x4	-10.9861	0.165	-66.670	0.000	-11.309	-10.663
x5	3.8161	0.305	12.516	0.000	3.218	4.414
x6	3.6726	0.487	7.544	0.000	2.718	4.627
x7	-2.6018	0.592	-4.392	0.000	-3.763	-1.441
x8	0.9808	0.704	1.393	0.164	-0.400	2.362
x9	13.9136	1.264	11.007	0.000	11.435	16.392

Omnibus:	396.988	Durbin-Watson:	1.958
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2652.998
Skew:	0.174	Prob(JB):	0.00
Kurtosis:	6.912	Cond. No.	144.

Non-Parametric Model
KNN Regression

Advantage

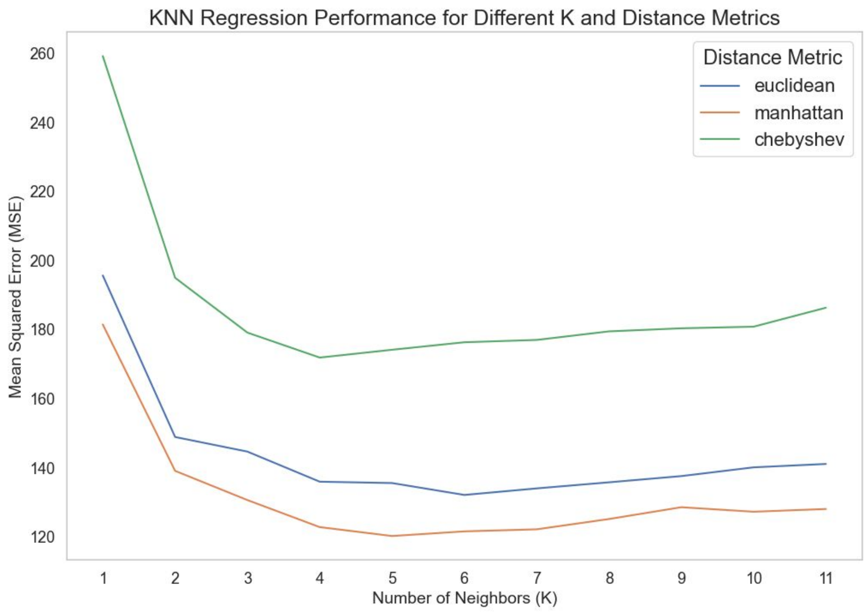
- A local learning algorithm
- Easy to interpret

Performance Comparison

- Number of Nearest Neighbor: K
- Distance Metric: E M C

Limitation

- Sensitive to high-dimension
- Scalability
- Computationally intensive



	Euclidean	Manhattan	Chebyshev
K	6	5	4
r^2	0.85597	0.86893	0.81261
MSE	132.09878	120.21027	171.87336

Decision Tree

Advantage

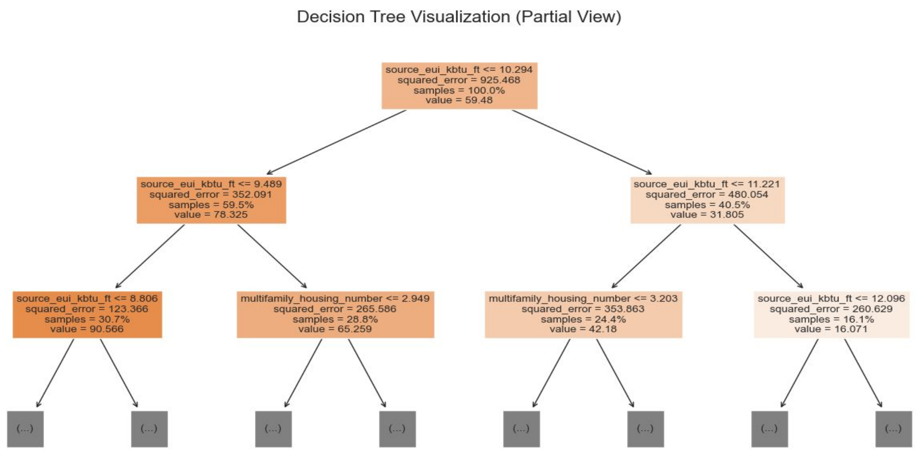
- Highly interpretable models
- Capture nonlinear relationships

Performance Comparison

- Decision Tree Depth

Limitation

- Sensitive to small changes



Random Forest

Advantage

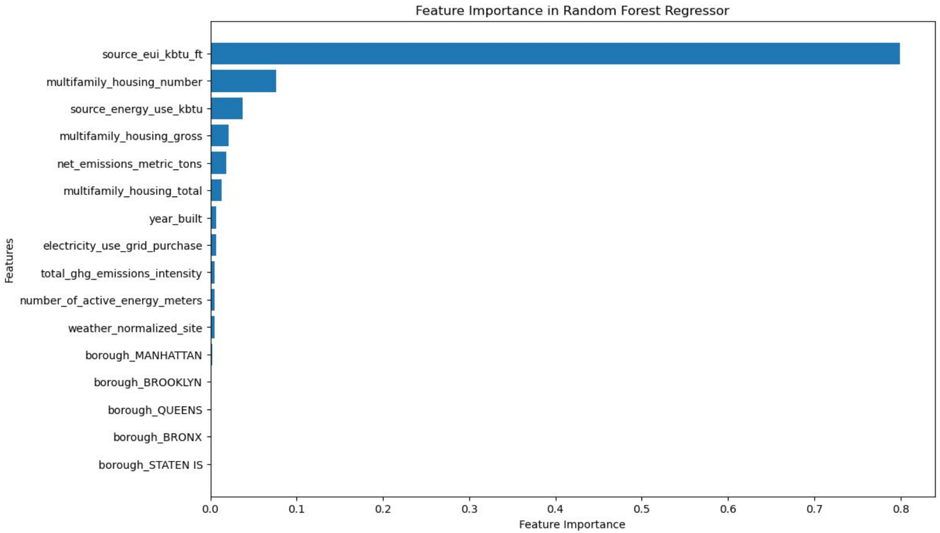
- Better Performance
- Less prone to overfitting

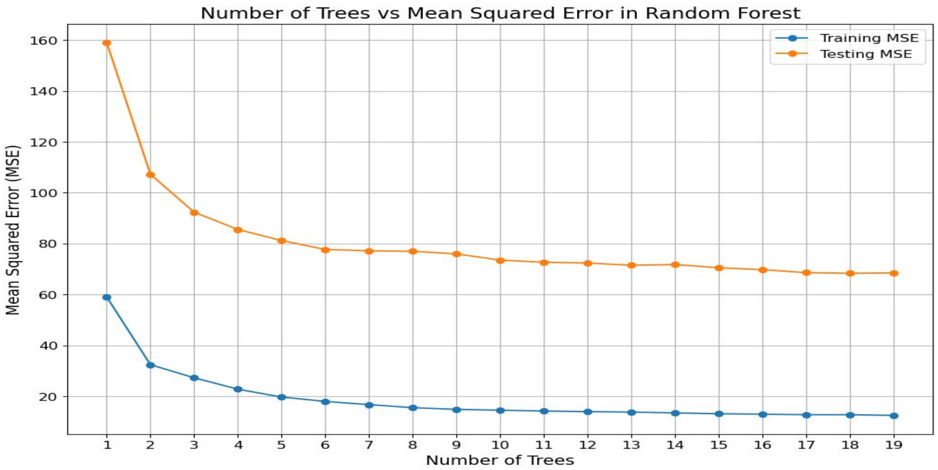
Performance Comparison

- Number of Trees

Important Features

- source eui (kBtu/ft²)
- multifamily housing number
- source energy use (kBtu)





Conclusion

Model Selection

Linear Models:

	OLS	Ridge (α = 0.01)	LASSO (α = 0.046)	PCR
Adjusted r^2	0.76717	0.76385	0.76088	0.71407
MSE	210.81282	213.81814	216.51371	258.89517

Non-parametric Models:

	KNN	Decision Tree	**Random Forest**
Adjusted r^2	0.84442	0.841921	**0.92229**
MSE	143.817842	146.13583	**71.83629**

Outcomes

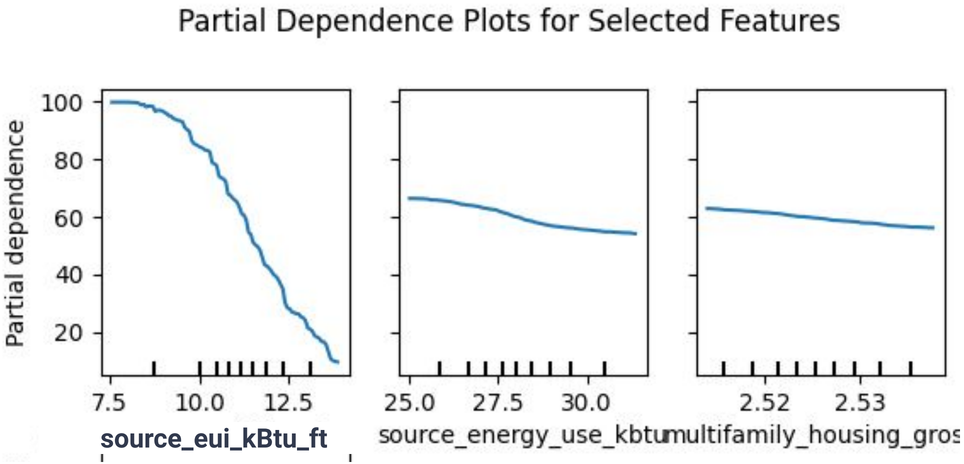
Recall objectives

- predict energy star score of multifamily property

Important Features

- source eui (kBtu/ft^2)
- multifamily housing number
- source energy use (kBtu)

Something from Energy Star



ENERGY STAR

What is Energy Use Intensity (EUI)?

When you benchmark your building in Portfolio Manager, one of the key metrics you'll see is energy use intensity, or EUI. Essentially, EUI expresses a building's energy use as a function of its size or other characteristics.

Future Improvements:

- Include more rows and features, such as: climate, weather and business activities.

ENERGY STAR

ABOUT FOR PARTNERS SEARCH

Find Products Save At Home New Homes Commercial Buildings Industrial Plants

Home » Commercial Buildings » ENERGY STAR Score for Multifamily Housing in the United States

ENERGY STAR Score for Multifamily Housing in the United States

< Back to search results

Last updated: 08-24-2018

The ENERGY STAR Score for Multifamily Housing applies to buildings that contain 20 or more residential living units. The objective of the ENERGY STAR score is to provide a fair assessment of the energy performance of a property relative to its peers, taking into account the climate, weather, and business activities at the property. To identify the aspects of building activity that are significant drivers of energy use and then normalize for those factors, a statistical analysis of the peer building population is performed. The result of this analysis is an equation that will predict the energy use of a property, based on its experienced business activities. The energy use prediction for a building is compared to its actual energy use to yield a 1 to 100 percentile ranking of performance, relative to the national population.

Your building is *not* compared to the other buildings in Portfolio Manager to determine your ENERGY STAR score. Instead, your building is compared to other buildings nationwide that have the same primary use. Where does this peer group come from?

Future Improvements:

ENERGY STAR® Portfolio Manager® Technical Reference

ENERGY STAR Score for Multifamily Housing in the United States

- Try using other columns as target variables (e.g. Total greenhouse gas emissions)

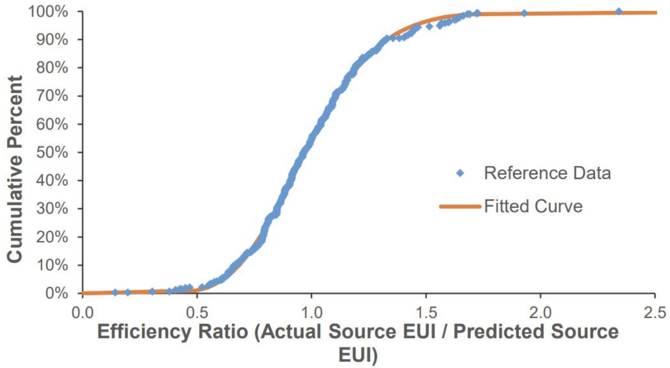
Figure 3 - Final Regression Results

Summary				
Dependent Variable	Source Energy Intensity (kBtu/ft²)			
Number of Observations in Analysis	322			
R^2 Value	0.2298			
Adjusted R^2 value	0.2176			
F Statistic	18.85			
Significance (p-level)	<0.0001			
	Unstandardized Coefficients	**Standard Error**	**T value**	**Significance (p-level)**
Constant	130.7	2.705	48.3	<0.0001
C_Unit Density	48.01	6.416	7.483	<0.0001
C_Bedrooms per Unit	22.64	5.700	3.972	<0.0001
Low Rise	- 19.00	3.976	- 4.777	<0.0001
C_HDD	0.008989	0.001502	5.983	<0.0001
C_CDD	0.01406	0.002494	5.638	<0.0001

Notes:
- *The regression is a weighted ordinary least squares regression*
- *The prefix C_ on each variable indicates that it is centered. The centered variable is equal to difference between the actual value and the observed mean. The observed mean values are presented in Figure 2.*
- *Low Rise is a yes/no variable (1 for yes, 0 for no). A building is defined as low rise (Yes) if it is no taller than 4 stories (e.g., 1-4 stories).*

$$\text{Energy Efficiency Ratio} = \frac{\text{Actual Source EUI}}{\text{Predicted Source EUI}}$$

Figure 4 – Distribution for Multifamily Housing



Thank you!