

# *NYC Greenhouse Gas Emission - Where and Why?*

*Spotting Energy-Inefficient Buildings*

*By Alex Friedman Pingqiao Wang  
Eva Sharman Geonhee Han*



*TRANSCENDING DISCIPLINES, TRANSFORMING LIVES*



**COLUMBIA ENGINEERING**  
The Fu Foundation School of Engineering and Applied Science

# A Summary

“Six million New York buildings account for  $\frac{1}{3}$  of statewide greenhouse gas emissions (GE).”  
Successful building energy conservation measures are crucial to lower GE.



- Such measures are costly to implement.
- Energy star score is an existing efficiency metric but may inadequately describe building-level performance.
- It is crucial to *accurately* spot energy inefficiency at a local building level in a time- and cost-economical manner.
- We provide a framework for practitioners to easily identify these targets.

# *Motivation*

## *& Initial Data Exploration*

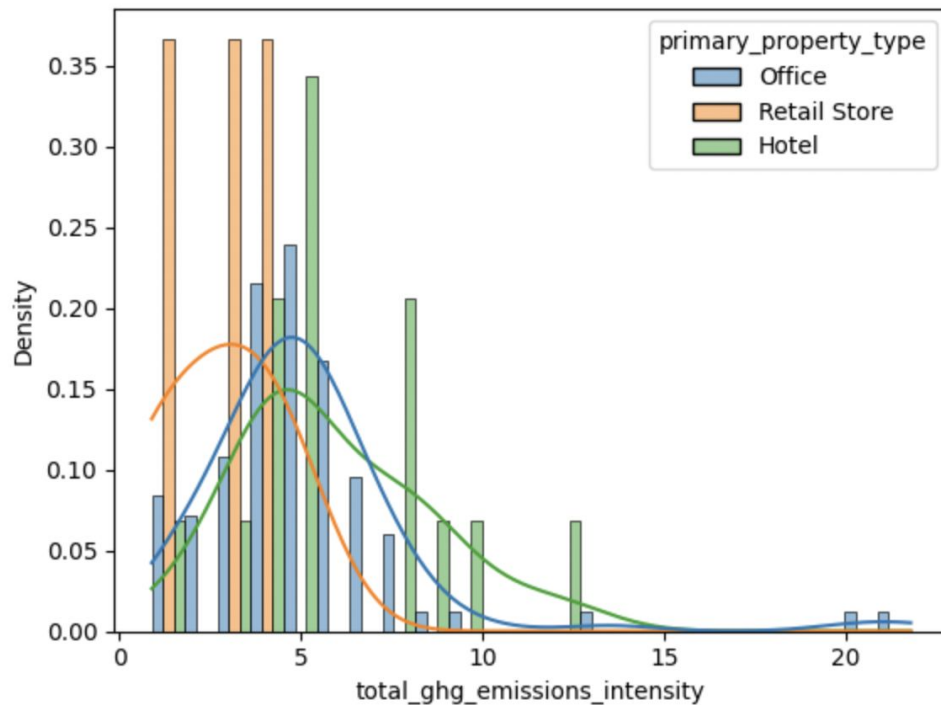
# Exploration and Early Stage Data Visualization

	parent_property_id	parent_property_name	year_ending	automobile_dealership_gross	medical_office_number_of	estimated_data_flag_fuel_3
0	Not Applicable: Standalone Property	Not Applicable: Standalone Property	2021-12-31T00:00:00.000	NaN	NaN	NaN
1	20599688	Stellar - Campus West 93rd Street	2021-12-31T00:00:00.000	NaN	NaN	No
2	Not Applicable: Standalone Property	Not Applicable: Standalone Property	2021-12-31T00:00:00.000	NaN	17.93	No
3	Not Applicable: Standalone Property	Not Applicable: Standalone Property	2021-12-31T00:00:00.000	NaN	NaN	No
4	Not Applicable: Standalone Property	Not Applicable: Standalone Property	2021-12-31T00:00:00.000	NaN	NaN	No
...	...	...	...	...	...	...

## Observation 1:

Very hard to “make sense” of the data.  
What are the key factors?

# Exploration and Early Stage Data Visualization



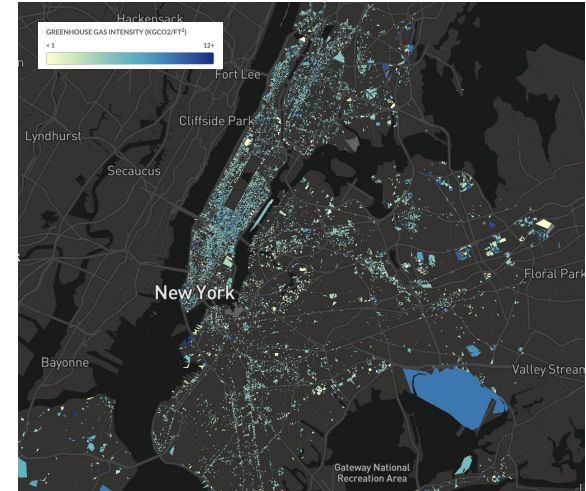
## Observation 2:

Cross-categorical  
heterogeneity in  
outcome distribution.

How can we model this?

# This leads to our Goals/RQ.....

- **What factors** do and don't explain NYC greenhouse gas emission intensity?
- Given characteristics of major buildings, can we
  - **Identify** “local outliers” with poor energy efficiency?
  - **Recommend** non-trivial candidates: buildings with high energy star scores despite poor efficiency?



GHG Intensity, 2017. NYC Mayor's Office of Sustainability.

*How can we understand GHG emission intensity of NYC buildings through analysis of 2021 energy and usage factors?*



# *Data & Model*

# Data Cleaning

- (1) Carefully/manually drop extraneous columns, drop properties without associated row data, replace missing values, and alternate invalid observations.
- (2) Conceptually partition the data column-wise.
- (3) Remove outliers in a distribution-agnostic manner

Energy Use Metrics;  
Data Quality Flags; ...



The diagram illustrates the conceptual partitioning of data columns into three groups:

- Location:** Indicated by a double-headed arrow above the first four columns (property\_id, latitude, longitude, borough).
- Property & Use:** Indicated by a double-headed arrow above the next four columns (primary\_property\_type, largest\_property\_use\_type, largest\_property\_use\_type\_1, year\_built).
- Energy Use Metrics; Data Quality Flags; ...:** Indicated by a single-headed arrow pointing right above the final three columns (number\_of\_buildings, occupancy, ...).

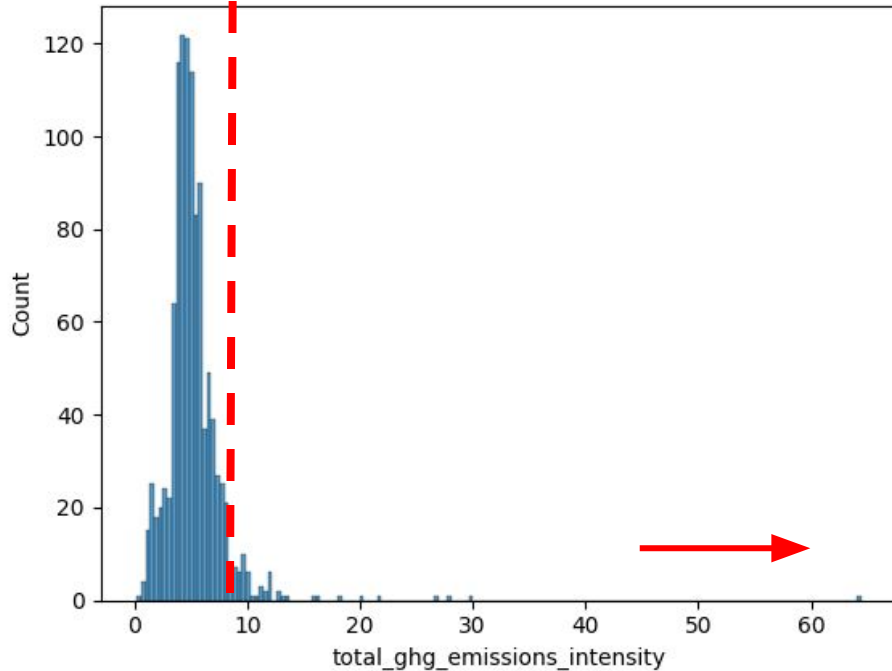
	property_id	latitude	longitude	borough	primary_property_type	largest_property_use_type	largest_property_use_type_1	year_built	number_of_buildings	occupancy	...
0	21205224	40.769272	-73.913633	QUEENS	Multifamily Housing	Multifamily Housing	25000	2010	1	100	...
1	2665352	40.790503	-73.96792	MANHATTAN	Multifamily Housing	Multifamily Housing	260780	1970	1	100	...
2	2665400	40.792758	-73.965171	MANHATTAN	Multifamily Housing	Multifamily Housing	324378	1943	1	100	...
7	2665443	40.837333	-73.94006	MANHATTAN	Multifamily Housing	Multifamily Housing	52428	1958	1	100	...
8	2665447	40.837275	-73.94423	MANHATTAN	Multifamily Housing	Multifamily Housing	70384	1973	1	100	...
...	...	...	...	...	...	...	...	...	...	...	...
1980	4095518	40.670965	-73.862535	BROOKLYN	Senior Living Community	Senior Living Community	42000	1975	1	100	...
1981	22480734	40.692915	-73.98815	BROOKLYN	K-12 School	K-12 School	76200	1927	1	100	...
1983	6275883	40.693963	-73.992561	BROOKLYN	Multifamily Housing	Multifamily Housing	88941	2010	1	100	...
1985	14719028	40.782735	-73.977327	MANHATTAN	Multifamily Housing	Multifamily Housing	94659	1924	1	100	...
1997	21967832	40.754847	-73.987833	MANHATTAN	Hotel	Hotel	179000	2021	1	55	...

1098 rows × 68 columns



# Probabilistic Model: Bayesian Hierarchical Truncated Normal

A hierarchical model allows one to jointly estimate parameters for all subsets and categories of data. It also performs the best for skewed, truncated and multi-levelled data. This model has the advantage over estimating subset parameters individually since it may reduce some noise.



$$\underbrace{y_i}_{\text{Greenhouse Intensity}} = y_i^* \mathbb{I}\{y_i^* \geq 0\},$$

$$y_i^* \sim N(\mu_i, \sigma^2),$$

$$\mu_i = \alpha + \underbrace{[\mathbf{X}\boldsymbol{\beta}]_i}_{\text{Local Variations}}.$$

$$\sigma_i \sim \text{Exponential}(\tau_i), \quad \tau_i \sim \text{Exponential}(1).$$

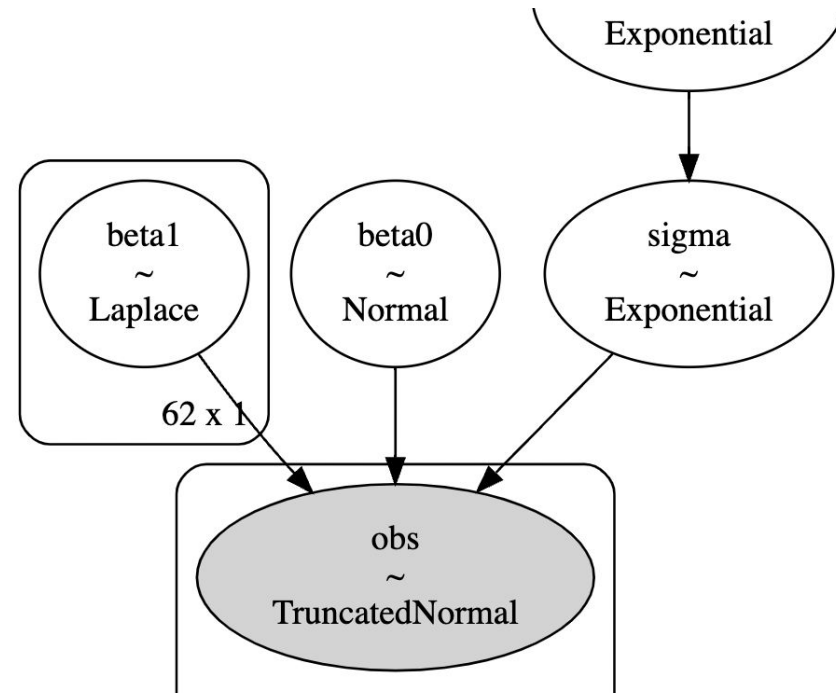
## (1) What factors do/don't explain GHG emission intensity?

“Bayesian Lasso” for Variable Selection.

$$\beta = \begin{bmatrix} \beta_{\text{num}} \\ \beta_{\text{cat}} \end{bmatrix}, \quad \beta_i \sim \text{Laplace}(0, 1).$$

Considerations:

- (Mild) shrinkage effect of coefficients;
- Computational ease over “SOTA” priors.



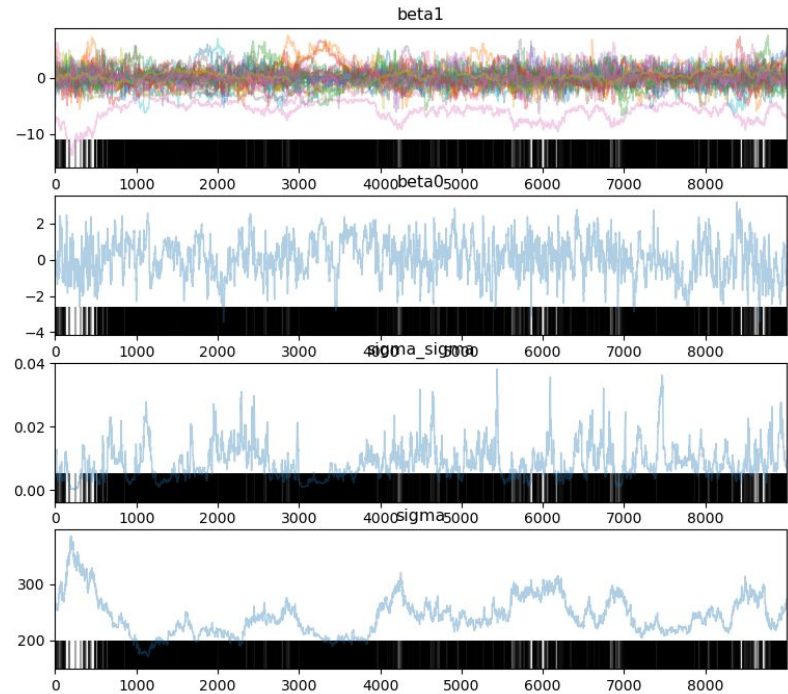
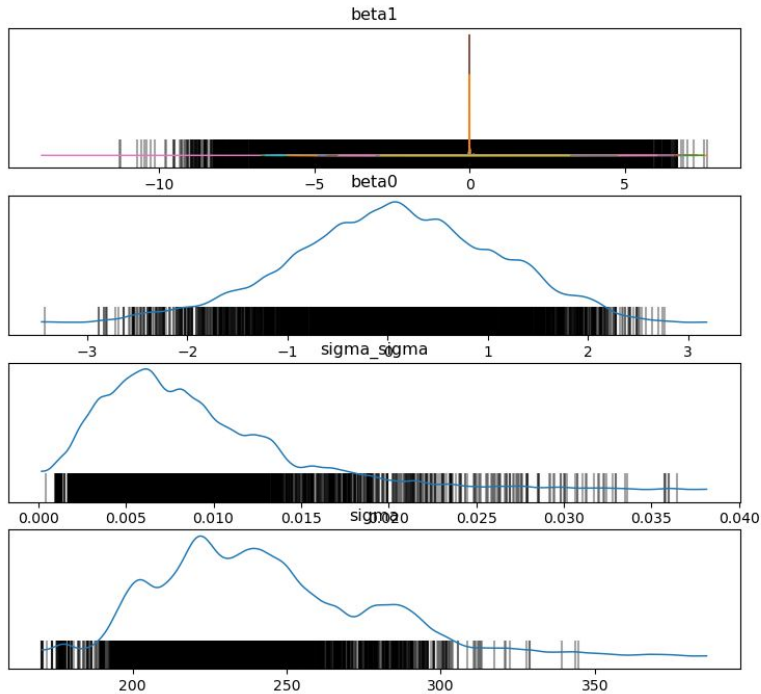
## (2) Identifying “Local Outliers” with a Multilevel Structure

$$\mathbf{X} = [\mathbf{X}_{\text{num}} \quad \mathbf{X}_{\text{cat}}],$$
$$\mathbf{X}_{\text{num}} \in \mathbb{R}^{N \times K_{\text{num}}},$$
$$\mathbf{X}_{\text{cat}} = \begin{bmatrix} \text{(Manhattan)} & \text{(Queens)} & \dots & \text{(Bronx)} & \text{(Multifamily Housing)} & \dots & \text{(K-12 School)} \\ 1 & 0 & \dots & 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 1 \\ \vdots & & & & & & \\ 0 & 0 & \dots & 1 & 1 & \dots & 0 \end{bmatrix}$$

- Examine data relative to predictive distribution.
- Considering location, property & use details, various energy use metrics, and data quality flags, which buildings are “local underperformers”?

# Model Performance

- Mean Acceptance Rate: 0.75
- We visually inspect convergence



# *Major Findings*

*And policy recommendations*

## Notable factors that contribute to GHG emission intensity

	(-)	(~0)	(+)
<b>Property Type / Characteristics</b>	Year Built [-5.91]		# of Bldgs [0.63] Manhattan [0.6] Residence Hall/Dorm [0.26]
<b>Usage Metrics</b>	Natural Gas Use (therms) [-1.44]  Electricity Use - Grid Purchase	Fuel Oil #1 Use (kBtu)  Natural Gas Use (kBtu)  Electricity Use Grid Purchase	Source EUI (kBtu/Ft) [0.85]



## An Example: Identifying “Local Outliers” (High Emission: low pdf, Obs > Pred)

Locally...	Borough	Type	Year Built	Occup.	Star Score	Site EUI (kBtu/ft <sup>2</sup> )	GHG Intensity
Lower	Manhattan	Multifam. Housing	1911	100	1	1207.4	5.7
Lower	Manhattan	Office	2013	60	49	75.8	64.5
Higher	Queens	Multifam. Housing	1961	90	88	86.2	4
Higher	Brooklyn	Multifam. Housing	1964	100	70	106.3	4.4


# Future Directions & Approaches

## Data Suggestions

- Quality: a more systematic Bayesian approach may be taken for missing data.
- Spatio-Temporal Modeling: exploit spatial correlation and temporal variations.

## Policy & Research Recommendations

- Incentivize energy-efficient improvements to tenements, especially in older properties & affordable housing.
- Investigate disparity in energy cost burden through additional demographic data.
- Prioritize clean-energy implementation plans that allow for building-level use of green power.
- Re-evaluate Energy Star scoring in a more holistic manner.



*Thank you !*

TRANSCENDING DISCIPLINES, TRANSFORMING LIVES



COLUMBIA | ENGINEERING  
The Fu Foundation School of Engineering and Applied Science