Project Report

DATA1030 Hands-on Data Science

Junrong Wang
Rhode Island School of Design
Master of Landscape Architecture Candidate
Github repository: https://github.com/JRONGW/DATA1030-Greenhouse-Gas-Emission-of-Buildings-in-NYC

1. Introduction

Buildings in urban areas like New York City are major contributors to greenhouse gas emissions due to energy consumption. Accurate prediction of emissions is critical for sustainable urban planning, but comprehensive data is often lacking. The NYC Building Energy and Water Data Disclosure dataset (2022–Present) offers valuable insights into building features and energy use, providing an opportunity to train machine learning (ML) regression models to estimate emissions for buildings with missing data.

Various models have been used to predict emissions. Traditional approaches like ARIMA and GM (1,1) have shown moderate accuracy, with hybrid models like GM-ARIMA improving predictive performance[2].The average relative errors for the ARIMA and GM (1,1) models were 6.309% and 44.38% respectively[3]. Neural networks, such as BP and LSTM, have demonstrated high accuracy (R^2 up to 0.987), capturing nonlinear relationships between building features and emissions[4][5][6]. Research indicates that the PSO-SVM model significantly reduced MAE and RMSE while increasing R^2 by an average of 0.1 compared to the optimized SVR, BP, and ELM models.[7]. However, these methods often face challenges like long training times and parameter sensitivity. Advances in preprocessing and dimensionality reduction have further enhanced model efficiency and accuracy.

2. EDA analysis

2.1 Summary of Statistics

Data type	**Column names**
Continuous data	Latitude', 'Longitude', 'Total (Location-Based) GHG Emissions (Metric Tons CO2e)', 'Net Emissions (Metric Tons CO2e)', 'sum floor area' 'Year Built'
Ordinal data	ENERGY STAR Score'
Categorical data	NYC Building Identification Number (BIN)', 'Property Name', 'Borough', 'Neighborhood Tabulation Area (NTA) (2020)', 'Calendar Year','major use type'

Table 1. Summary of Statistics

index	Latitude	Longitude	Total (Location-Based) GHG Emissions (Metric Tons CO2e)	Net Emissions (Metric Tons CO2e)	sum floor area	Year Built
count	61790	61790	62712	63222	64169	64169
mean	40.74787855	-73.9360543	914.1075328	810.5927051	127715.6599	1951.318814
std	0.078341916	0.064505648	20571.73738	9198.873058	321985.031	35.26277469
min	40.509037	-74.244118	-43.9	-8385.1	0	1051
25%	40.697285	-73.978987	160.2	181.9	36000	1925
50%	40.75005	-73.946978	292.2	316.5	60883	1941
75%	40.81022675	-73.9009565	558.5	608.975	116897	1974
max	40.912869	-73.700935	3139711.6	2206862.8	22042704	2088

Table 2. Summary of Continuous Data

NYC_Building_Identification_Number_BIN	Frequency
2123911	190
2999999	93
4999999	50
0	49
2109476	32

Property_Name	Frequency
University Center	6
Main Building	6
Cross Bronx Pres LLC - Baychester (2166bay)	6
Alumni Hall	6
29-07 31st Ave	6

Neighborhood_Tabulation_Area_NTA_2020	Frequency
MN0502	1844
MN0802	1422
MN0501	1341
MN0702	1287
MN0401	1158

Calendar_Year	Frequency
2023	33684
2022	30485

major_use_type	Frequency
Multifamily Housing	42030
Office	5118
K-12 School	3780
Data Center	2903
Hotel	1299

Borough	Frequency
MANHATTAN	22493
BROOKLYN	15362
BRONX	11998
QUEENS	10619
STATEN IS	1321

Table 3. Summary of Categorical Data

ENERGY STAR Score	Count
100	3804
1	1078
99	823
86	760
80	708
...	...
8	237
9	236
13	233
7	225
16	225

Table 4. Summary of Ordinal Data

2.2 Correlations

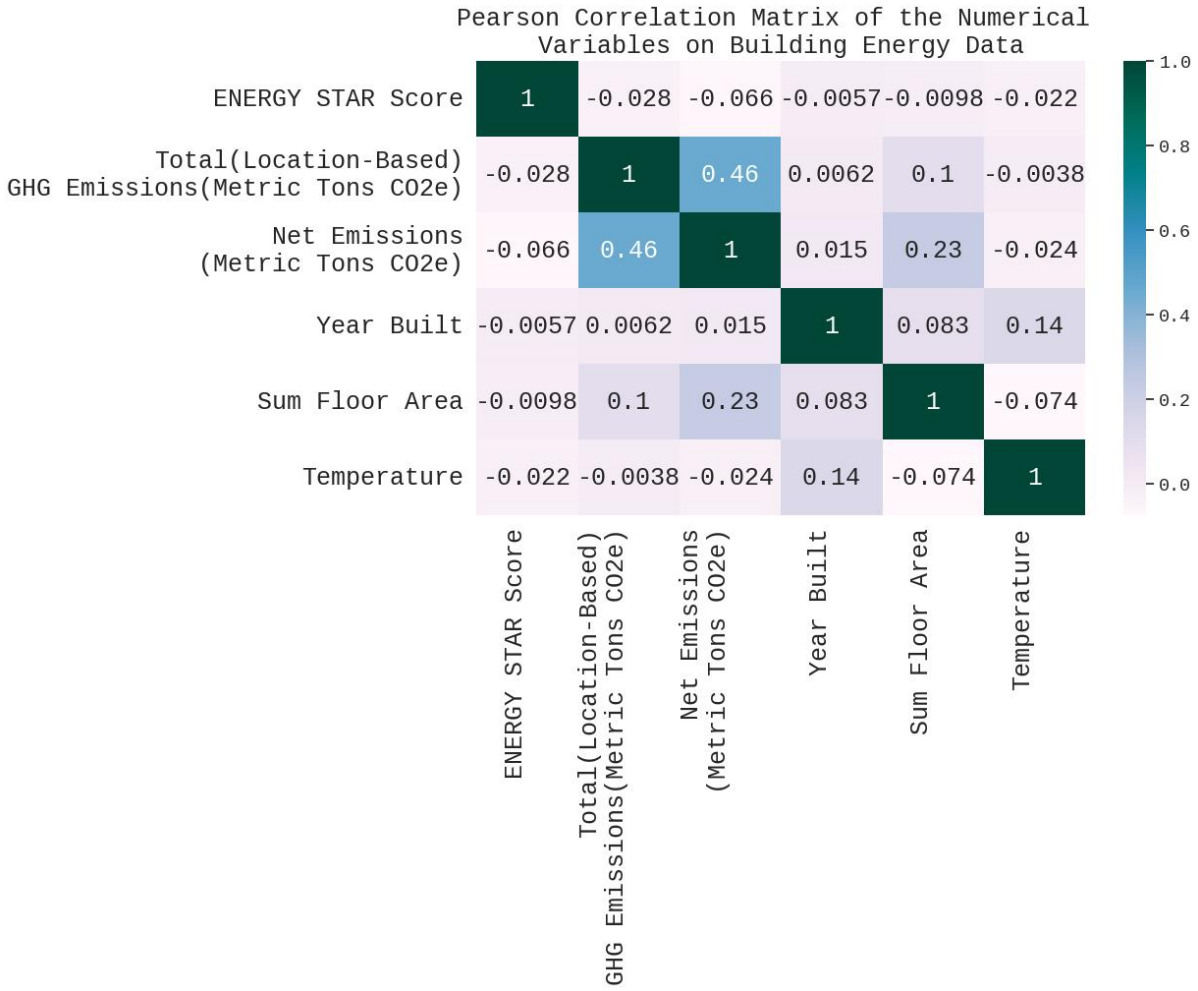

Figure 1. Pearson Correlation Matrix of the Numerical Variables on Building Energy Data

There are little linear relationship between each variables. except for Total GHG Emissions and Net Emissions. It might be better to explore non-linear relationship.

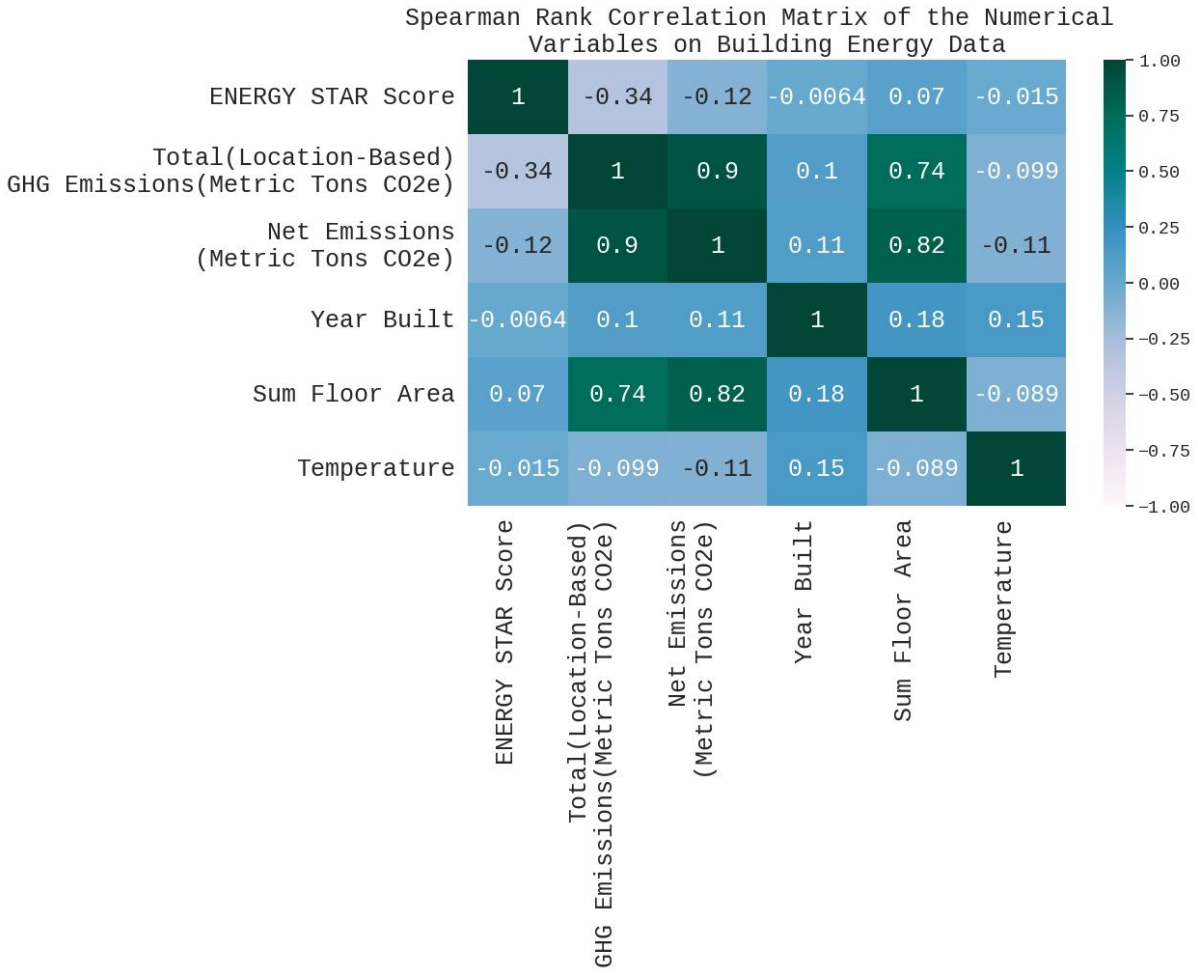

Figure 2. Spearman Correlation Matrix of the Numerical Variables on Building Energy Data

Spearman's rank correlation reveals notable monotonic relationships between sum floor area and Total GHG Emissions.

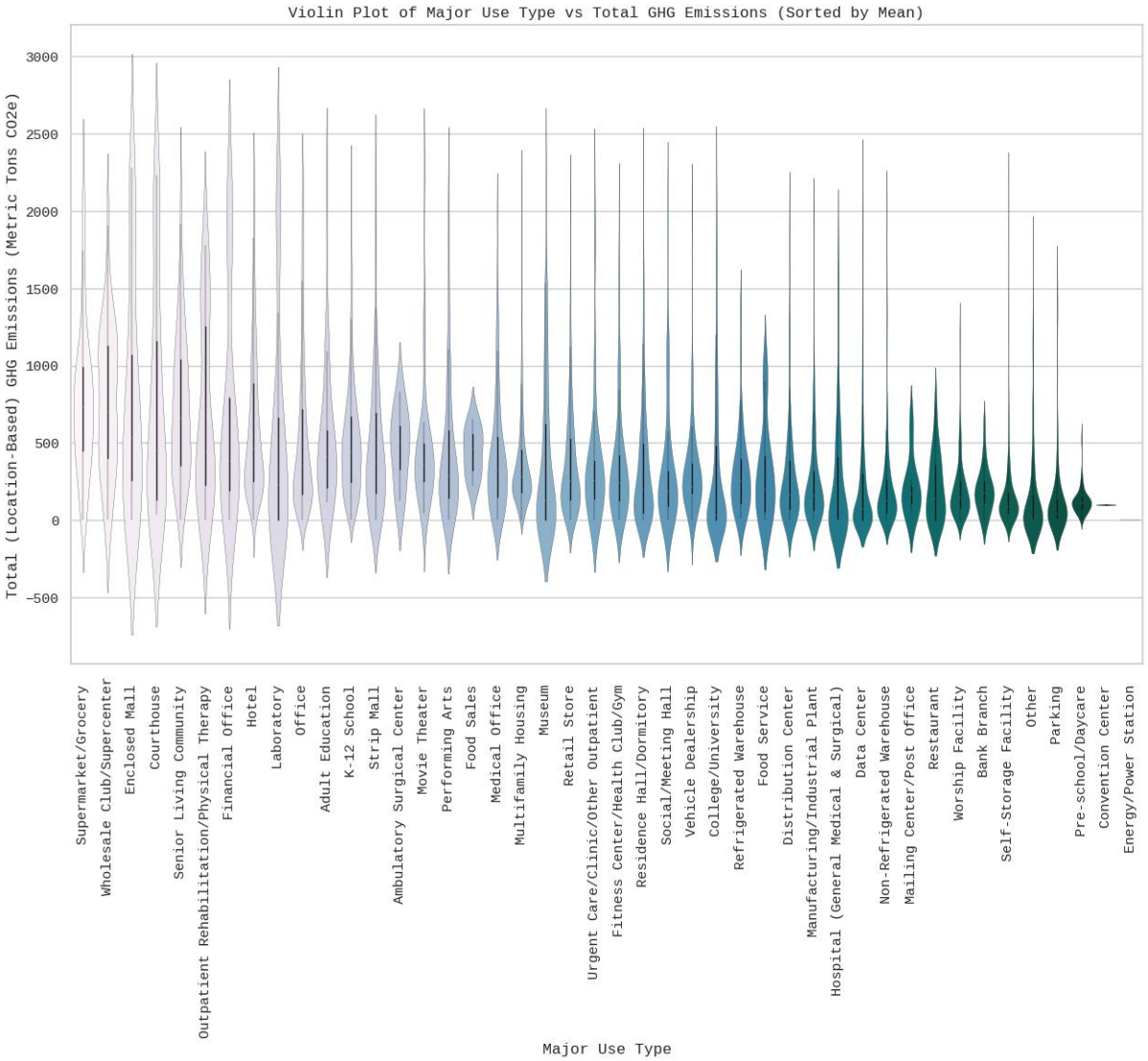

Figure 3. Violin Plot of Major Use Type vs Total Greenhouse Gas Emissions(Sorted by Mean)

We can see that some major use type do contribute to higher GHG emissions

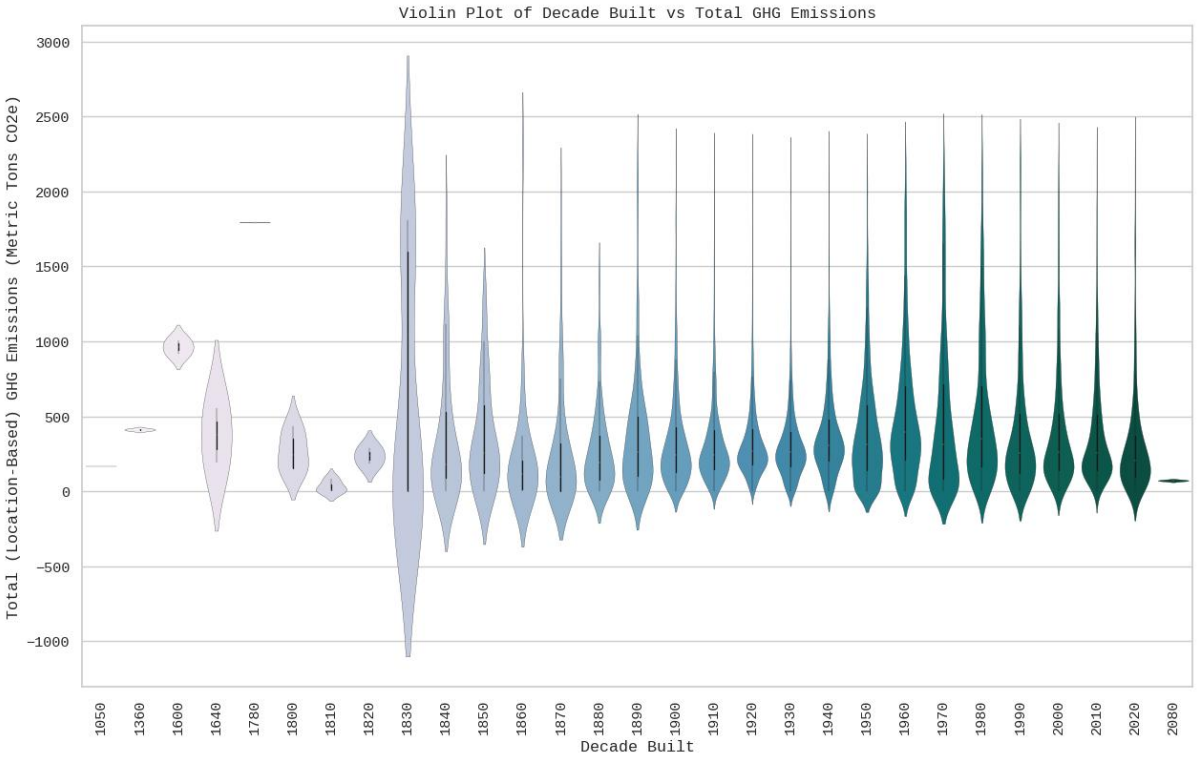

Figure 4. Violin Plot of Decade Built vs Total Greenhouse Gas Emissions

Carbon emissions peak for buildings from the 1830s (due to materials/techniques) and the 1970s (large developments/inefficient energy use), then decline post-1970s with improved regulations and efficiency.

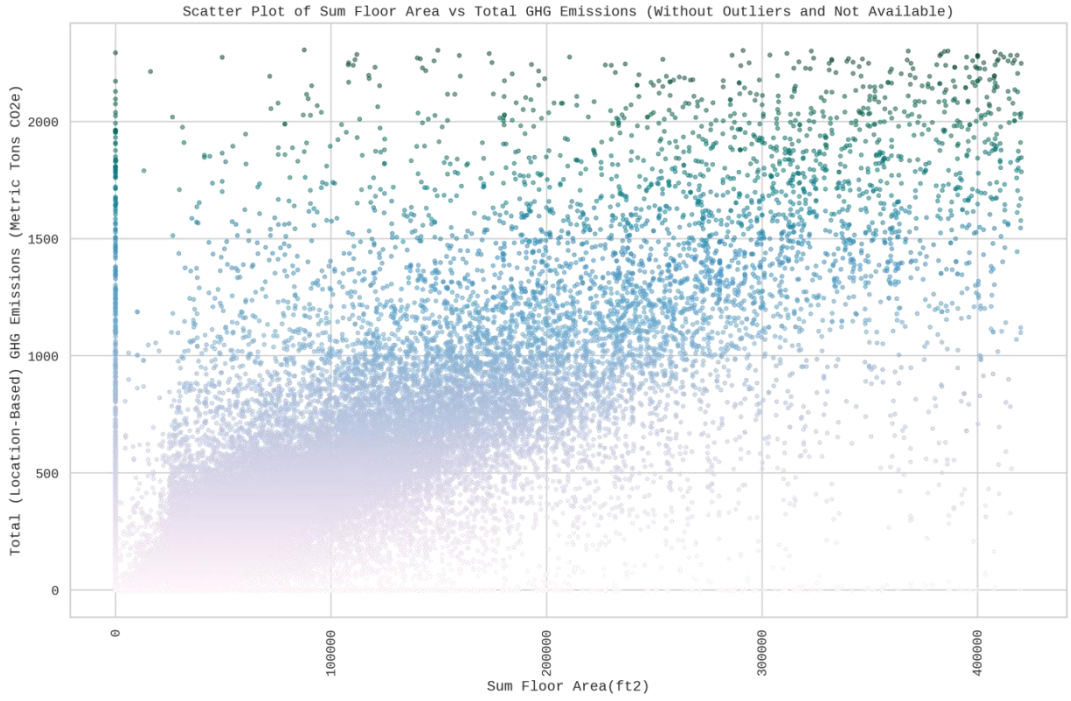

Figure 5. Scatter Plot of Sum Floor Area vs Total Greenhouse Gas Emissions(Without Outliers)

We can observe the general trend where the sum floor area increases alongside the GHG emissions, with some clustering around lower values.

2.3 Other features

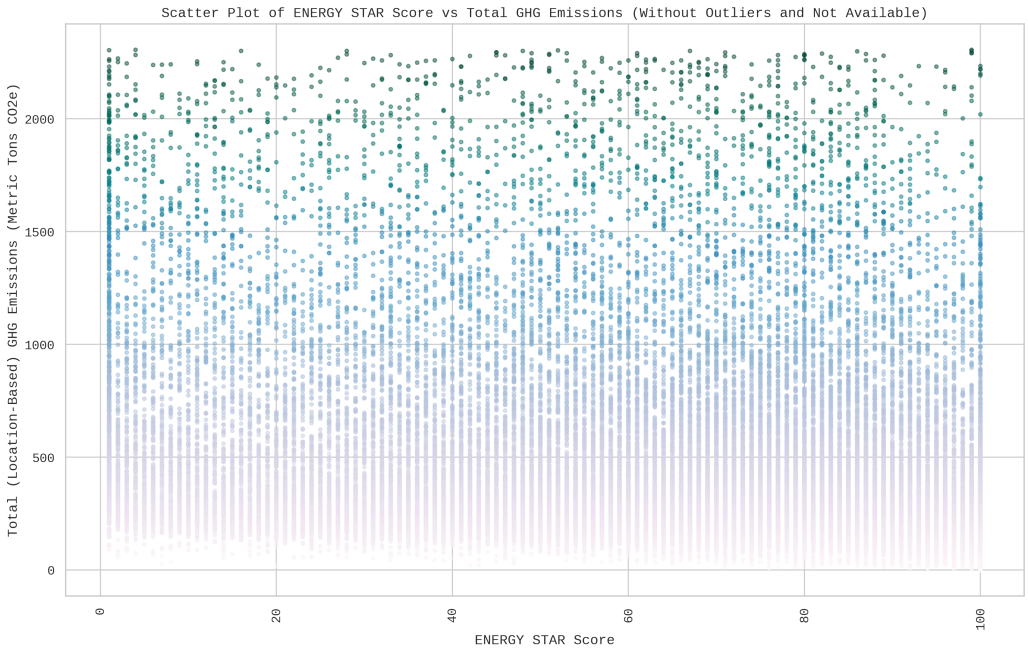

Figure 6. Energy Star Score vs Total Greenhouse Gas Emissions(Without Outliers)

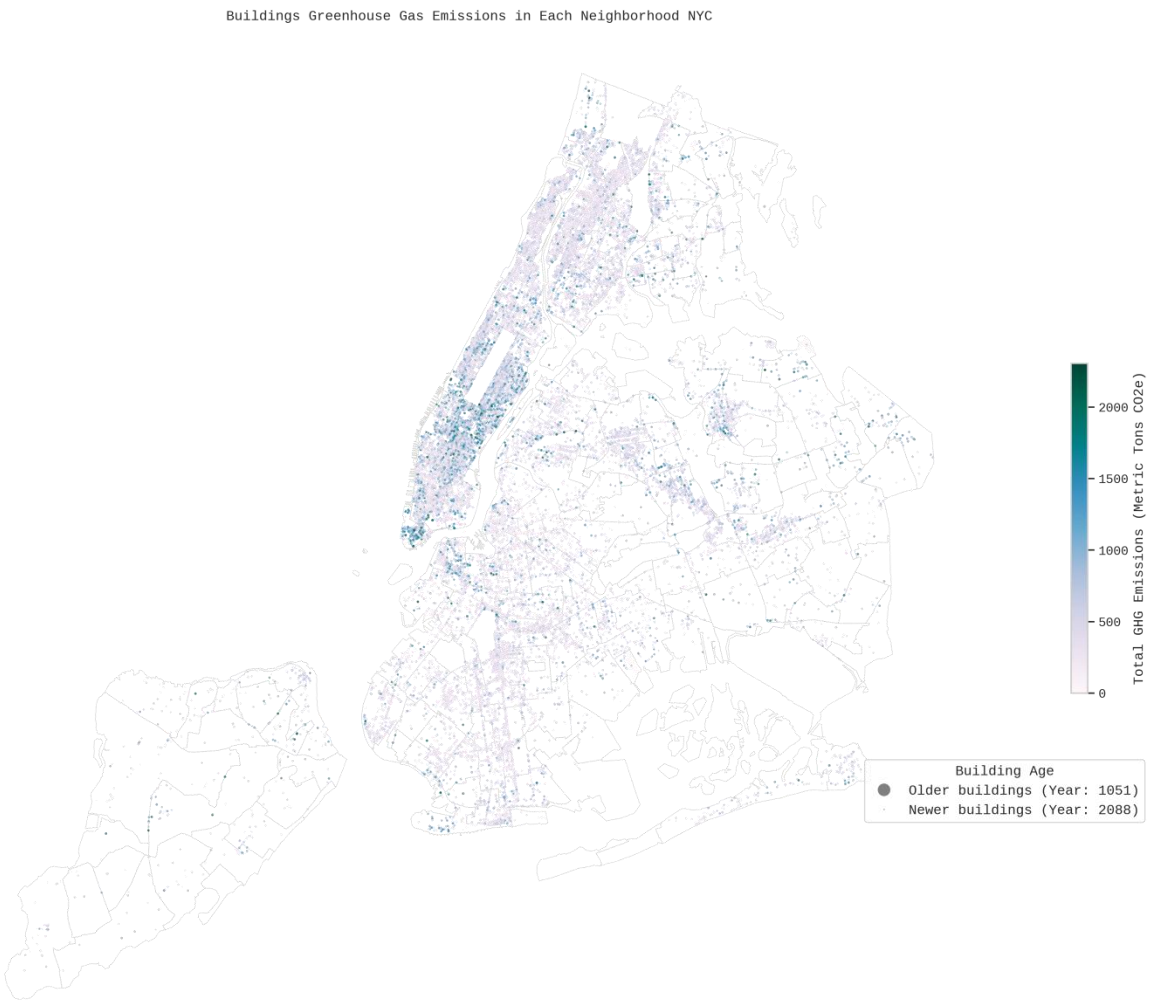

Figure 7. Map of Building Total Greenhouse Gas Emissions in each Neighborhood of NYC

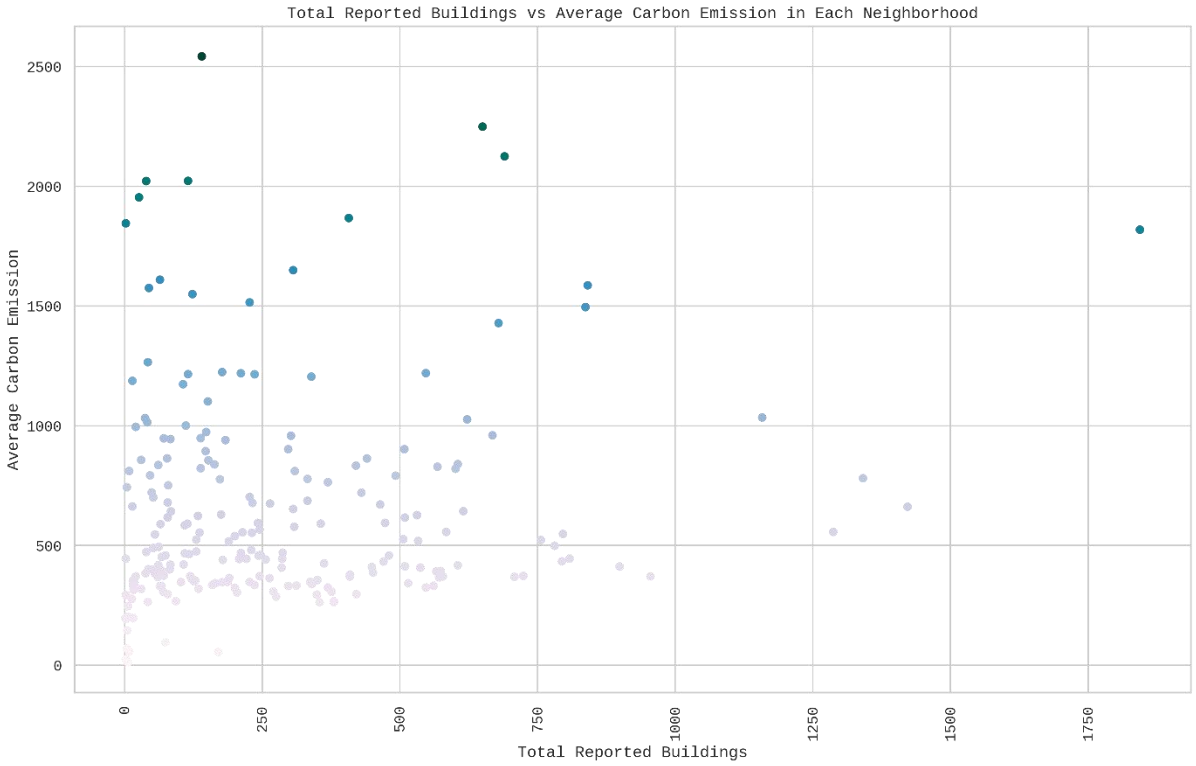

Figure 8. Total Reported Buildings vs Average Carbon Emissions in Each Neighborhood

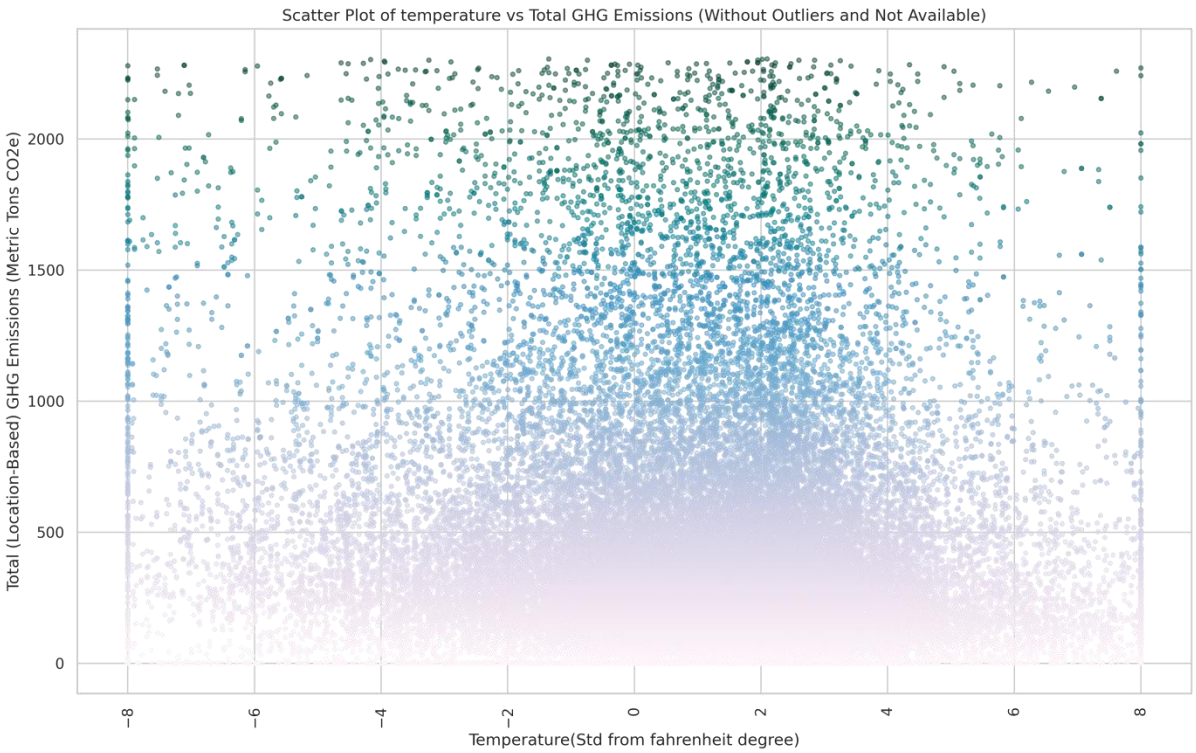

Figure 9. Scatter Plot of temperature vs Total GHG Emissions

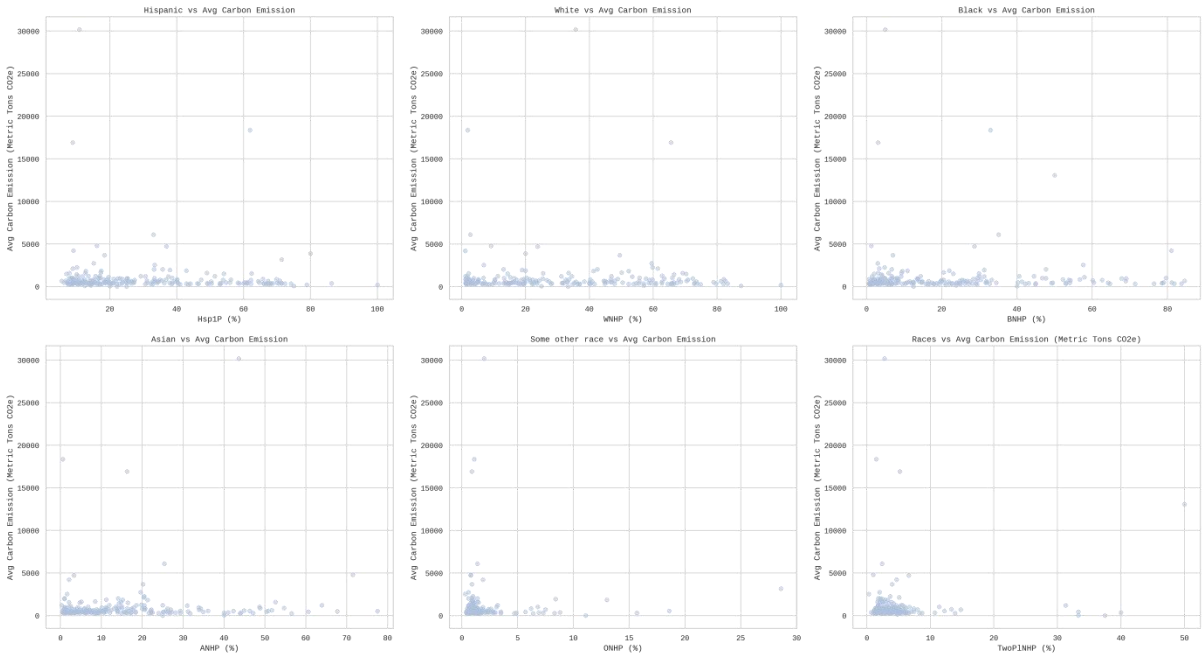

Figure 10. Scatter Plot of Percentage of Races vs Average Carbon Emission in Neighborhoods

The original feature Energy Star score, and feature engineered new features Total reported buildings in neighborhood, temperature, and race percentage has very little correlation to the greenhouse gas emission.

3. Methods

3.1 Missing data treatment

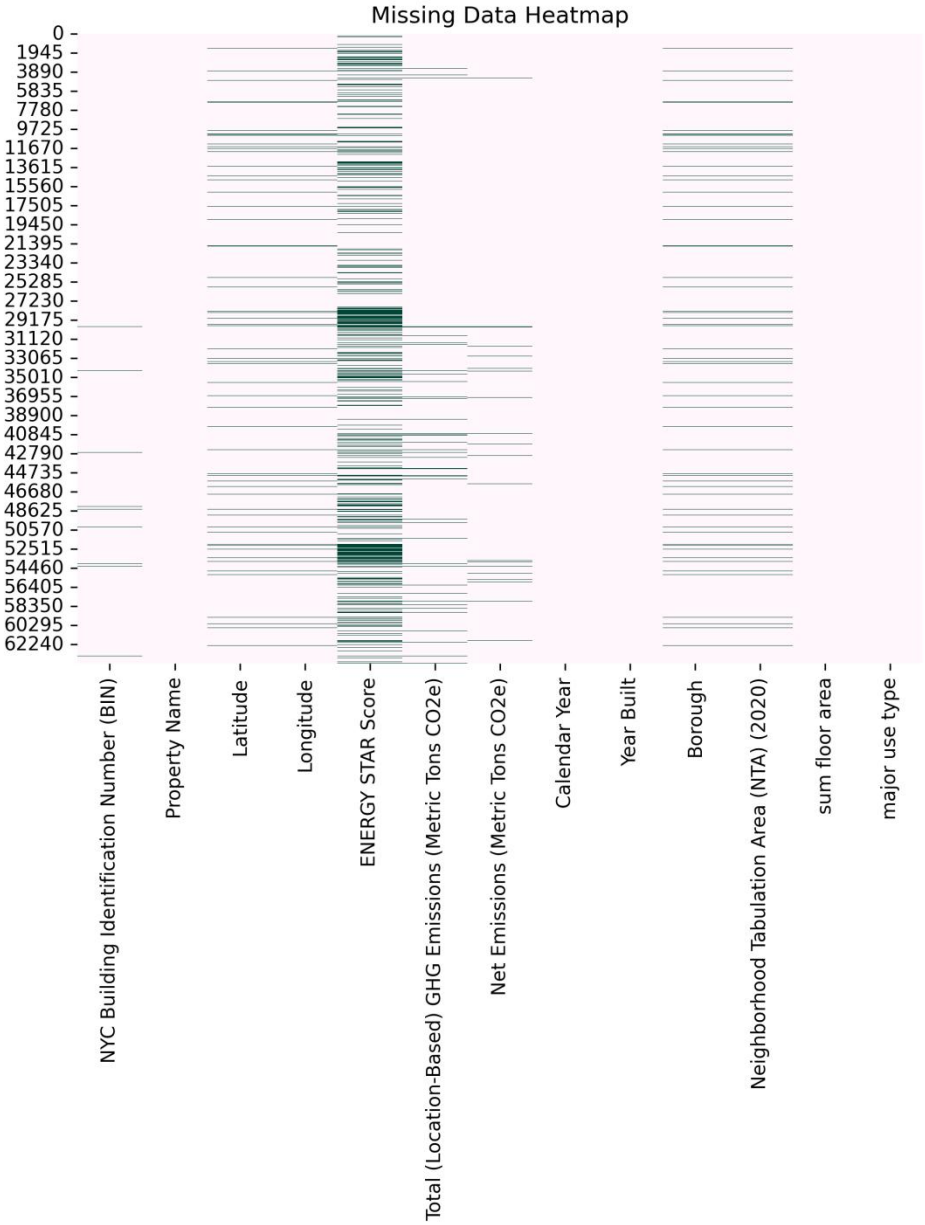

Figure 11. Scatter Plot of Races vs Avg Carbon Emission

Column	Fraction of Null Values
NYC Building Identification Number (BIN)	0.009771
Property Name	0
Latitude	0.037074
Longitude	0.037074
ENERGY STAR Score	0.235784
Total (Location-Based) GHG Emissions (Metric Tons CO2e)	0.022706
Net Emissions (Metric Tons CO2e)	0.014758
Calendar Year	0
Year Built	0
Borough	0.037027
Neighborhood Tabulation Area (NTA) (2020)	0.037058
sum floor area	0
major use type	0

Table 5. Summary of Missing Data

Most missing data is minimal, so dropping affected rows has little impact. Rows with missing Latitude and Longitude are dropped, as imputation is unfeasible and complicates modeling. Rows with missing values, except for 'ENERGY STAR Score,' are removed to ensure data completeness. 'ENERGY STAR Score' is retained but excluded from training due to weak correlation with the target variable, avoiding bias from imputation.

3.2 Outliers treatment

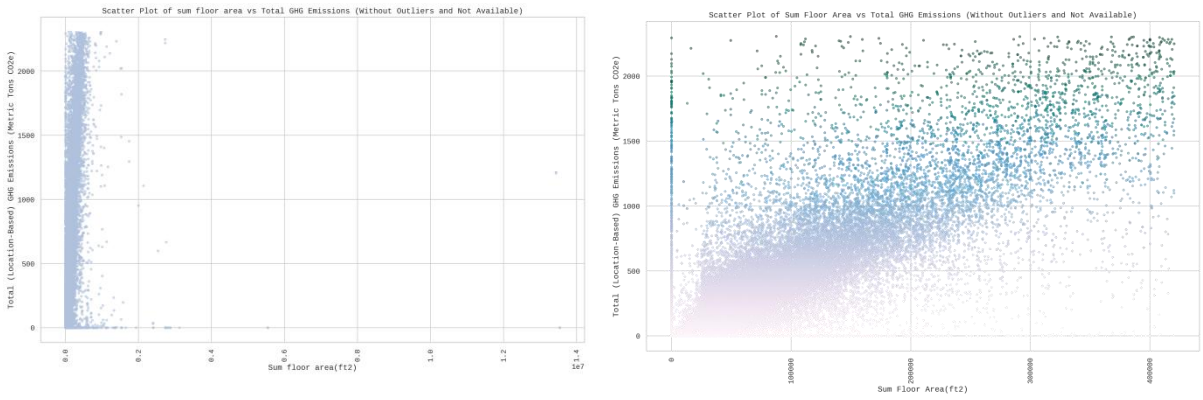

Figure 12. Sum Floor Area Without Outliers and With Outliers vs Total GHG Emissions

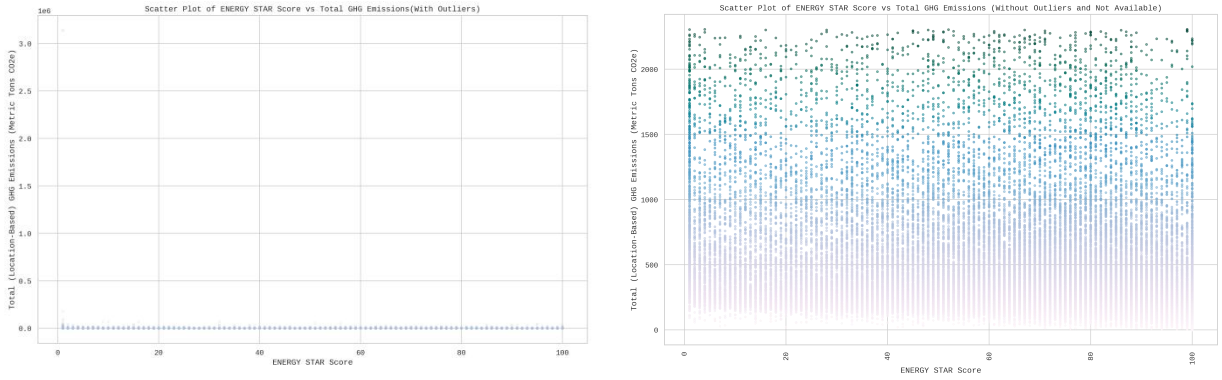

Figure 13. Energy Star Score Without Outliers and With Outliers vs Total GHG Emissions

With outliers the plots are not balanced and the data would yield high MSE in the training. So I remove the outliers. I set the 95th percentile as threshold for filtering out the outliers.

3.3 Preprocessing

3.3.1 Feature Selection

For model training, the features are ‘year built,’ ‘sum floor area,’ and ‘major use type’ as X, and ‘Total (Location-Based) GHG Emissions (Metric Tons CO2e)’ as y, selected for their strong correlation with the target variable during EDA. The ‘Building Identification Number (BIN)’ is used only for splitting to prevent data leakage, ensuring buildings don’t appear in both training and validation/test sets. BIN is dropped before training to ensure the model generalizes to predict emissions for unseen or new buildings, including those outside NYC

3.3.2 Categorical Data - One-Hot Encoding

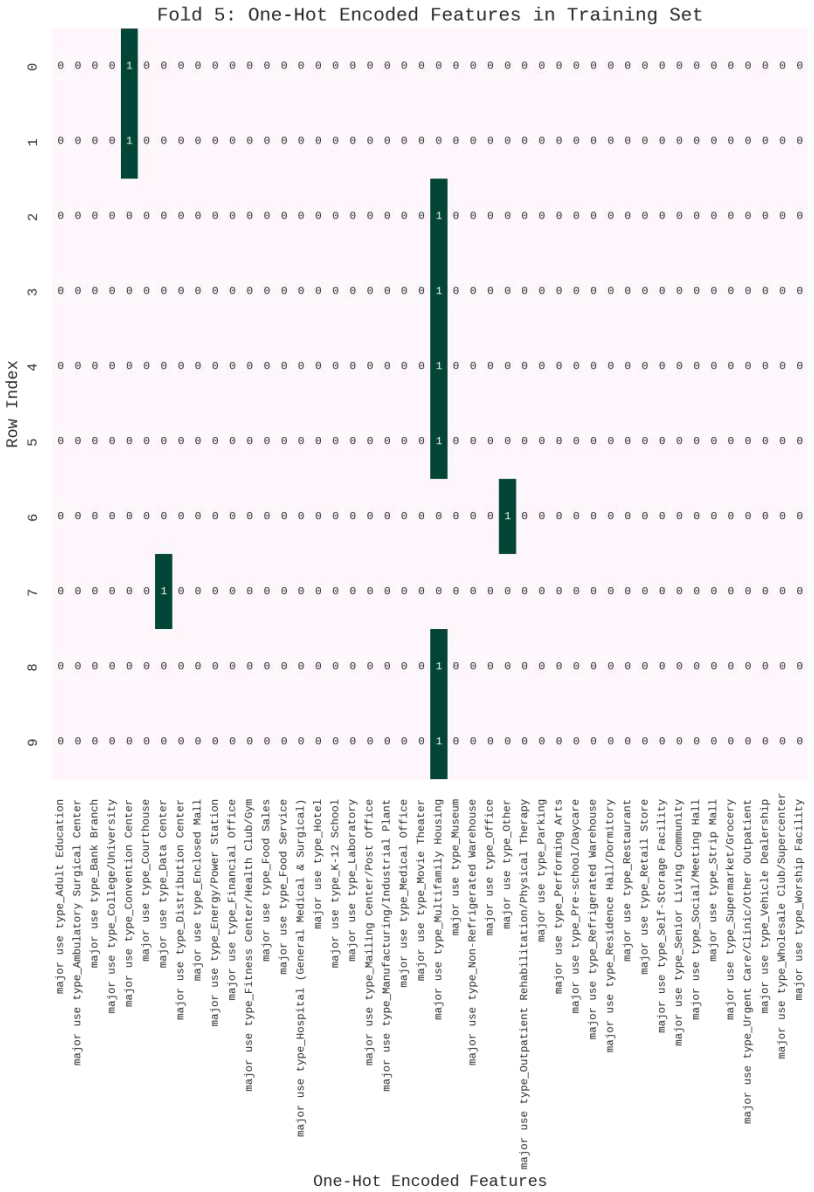

Figure 14. One-Hot Encoded Features in Training Set Fold 5

The categorical feature is processed with One-Hot Encoding.

3.3.3 Continuous Data - Standard Scaler

Before Scaling	
Year Built	sum floor area
1921	31680
1921	31680
1926	206009
1929	53460
1928	76500

After Scaling	
Year Built	sum floor area
0.8389585342	0.002339714595
0.8389585342	0.002339714595
0.843780135	0.01521471793
0.8466730955	0.003948268379
0.8457087753	0.005649878993

Table 6. Standard Scaled Features in Training Set Fold 5

The continuous features are processed with Standard Scaler.

3.4 Cross Validation

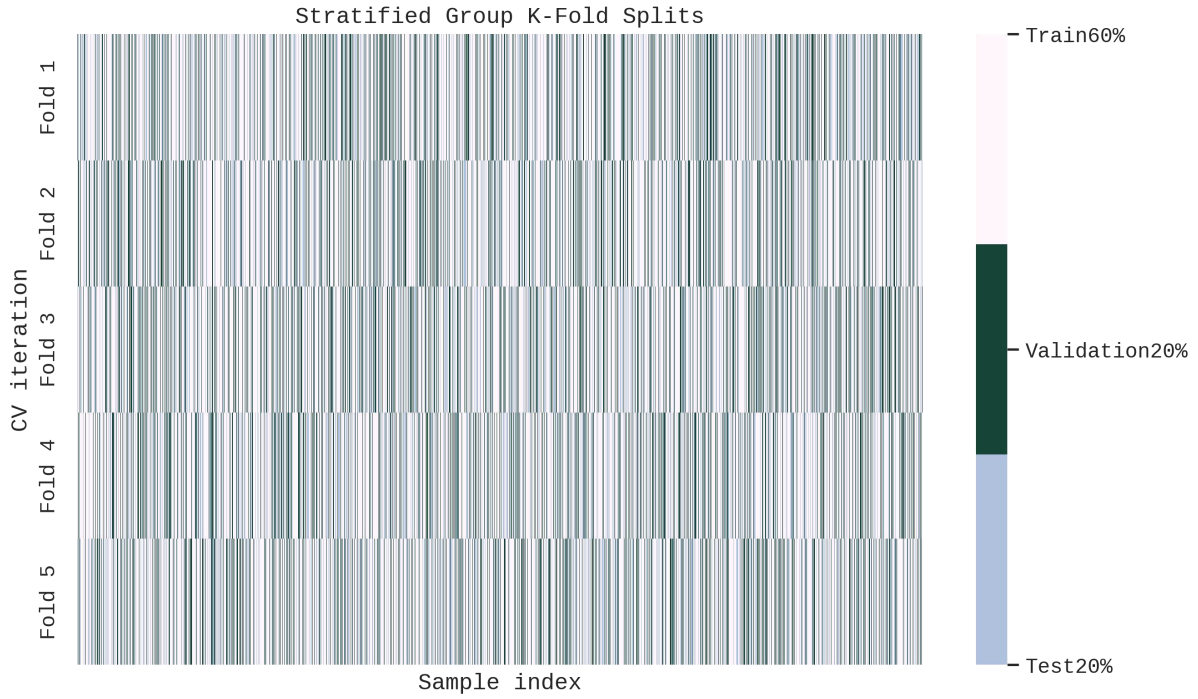

Figure 15. Stratified Group K-fold Splits

3.4.1 Filter Types and IDs

The filter_types function extracts valid major use type values and corresponding building IDs, creating a mask to handle missing values.

3.4.2 Consistency Check

A function identifies cases where a major use type corresponds to only one NYC Building Identification Number (BIN), labeled as consistent. Treating these separately prevents patterns from appearing in validation or test sets that are absent in training.

3.4.3 Split Data into Consistent and Inconsistent Groups

The dataset is divided into consistent (one BIN per major use type) and inconsistent (multiple BINs per major use type) groups.

3.4.4 Stratified Group K-Fold for Inconsistent Data

StratifiedGroupKFold splits inconsistent data, ensuring no building appears in both training and validation/test sets while maintaining major use type stratification across splits.

3.4.5 Further Splitting of Training Data

A secondary split using train_test_split creates training and validation sets from the training portion of the inconsistent data.

3.4.6 Combine Consistent Data

Consistent data is added to the training set, ensuring comprehensive representation of major use types in training.

3.5 Algorithms

Algorithms	**Parameters**
XGBRegressor	"reg_alpha": [0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2],"reg_lambda": [0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2],"max_depth": [3, 10]
Lasso	'alpha': [0.01, 0.1, 1, 10],'fit_intercept': [True, False]
Ridge	'alpha': [0.01, 0.1, 1, 10],'fit_intercept': [True, False]
ElasticNet	'alpha': [0.01, 0.1, 1, 10],'l1_ratio': [0,0.2,0.4,0.8,1]
RandomForestRegressor	'max_depth': [1, 3, 10, 30, 100],'max_features': [1, 3, 10, 30, 100]
SVR	'gamma': [1e-3, 1e-1, 1e1, 1e3],'C': [1e-2, 1e-1, 1e0, 1e1, 1e2]
KNeighborsRegressor	'n_neighbors': [1, 3, 10, 30, 100]

Table 7. Algorithms Used for Training and Hyperparameters Tuned

3.4 Feature Importance

Algorithm	Global or Local	Feature Importance Type
XGB Regressor	Global	Weight
		Gain
		Cover
		Total gain
		Total cover
		SHAP global
	Local	Shap local_best val score
		Shap local_best test score
Random Forest Regressor	Global	Gini importance
	Local	Shap local_best val score
		Shap local_best test score

Table 8. Feature Importance Types

4. Results

4.1 Hyper-parameter Tuning

Algorithms	Best Model Parameters	Average Test R2	Std of Test R2
XGBRegressor	{'xgbregressor__subsample': 0.66, 'xgbregressor__reg_lambda': 0.1, 'xgbregressor__reg_alpha': 0.0, 'xgbregressor__n_estimators': 10000, 'xgbregressor__max_depth': 3, 'xgbregressor__learning_rate': 0.03, 'xgbregressor__colsample_bytree': 0.9}	0.604332176518392	0.0199929735126074
Lasso	{'lasso__fit_intercept': False, 'lasso__alpha': 0.1}	0.1509717817464513	0.1206312934200583
Ridge	{'ridge__fit_intercept': False, 'ridge__alpha': 0.1}	0.15133187110086063	0.12068393053052374
ElasticNet	{'elasticnet__l1_ratio': 0.8, 'elasticnet__alpha': 0.01}	0.1749305573492118	0.08154238704919459
RandomForestRegressor	{'randomforestregressor__max_features': 10, 'randomforestregressor__max_depth': 1}	0.5985147226648648	0.014402540685798779
SVR	{'svr__gamma': 10.0, 'svr__C': 0.01}	0.5990450207198715	0.00969966522538898
KNeighborsRegressor	{'kneighborsregressor__n_neighbors': 100}	0.5936800199059649	0.008758829340313018

Table 9. Algorithms Used for Training and Hyperparameters Tuned

4.2 Model Selection

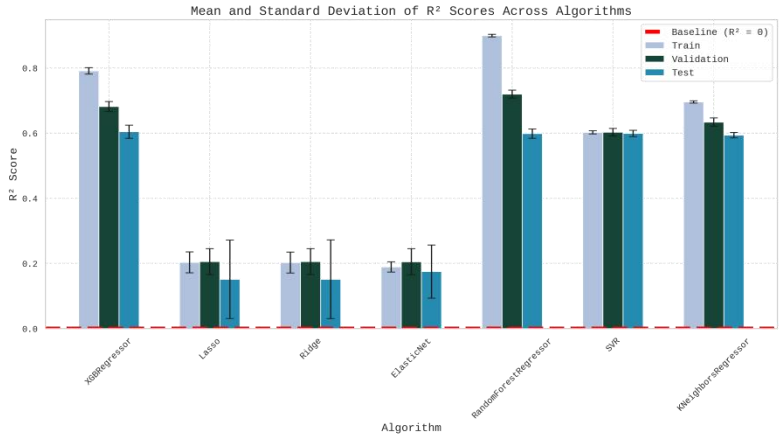

Figure 16. Mean and Standard Deviation of R² Scores Across Algorithms

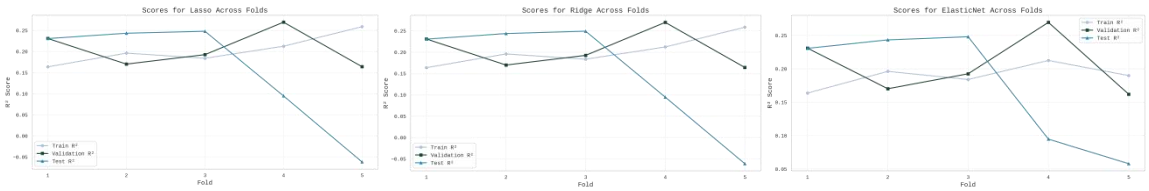

Figure 17. Mean and Standard Deviation of R² Scores for 3 Linear Models(Low R²)

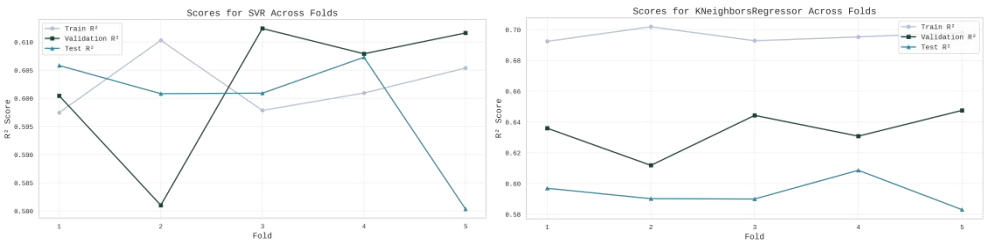

Figure 18. Mean and Standard Deviation of R² Scores for SVR and KNN(Medium R²)

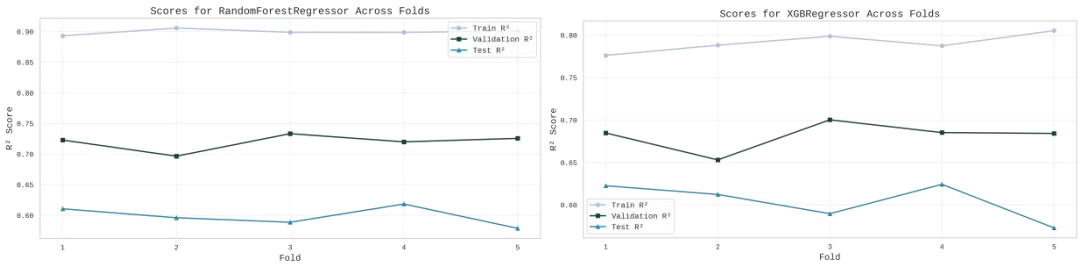

Figure 19. Mean and Standard Deviation of R² Scores for RF and XGB(High R²)

XGBoost and Random Forest emerged as the top-performing models with the lowest validation error.

Linear models underperformed due to their inability to capture non-linear relationships.

Baseline R^2 was 0.0, indicating substantial improvement across all models.

4.3 Feature Importance

Global Feature Importance: Sum Floor Area and Year Built were the most predictive features.

Local Feature Analysis (SHAP): Highlighted nuanced interactions for specific buildings (e.g., high emissions driven by specific use types).

Key Insights: Multifamily Housing significantly influenced emissions, while engineered features like “Race Percentage” had negligible impact.

Surprises: Supermarket/Grocery and Self-Storage Facilities exhibited unexpectedly high impact despite relatively lower frequencies.

Validation: Feature importance rankings aligned with domain knowledge, bolstering model credibility.

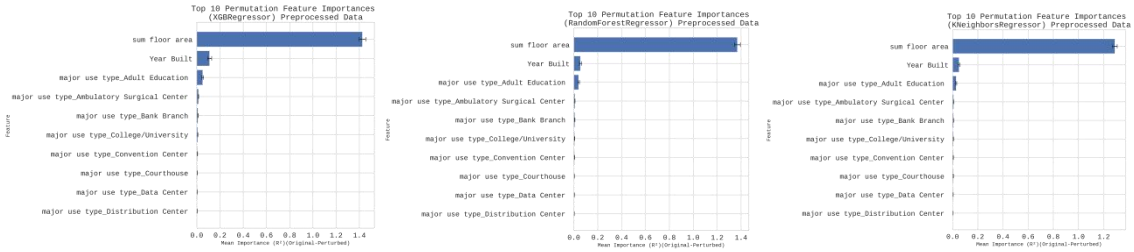

Figure 20. Top 10 Permutation Feature Importances of 3 Algorithms

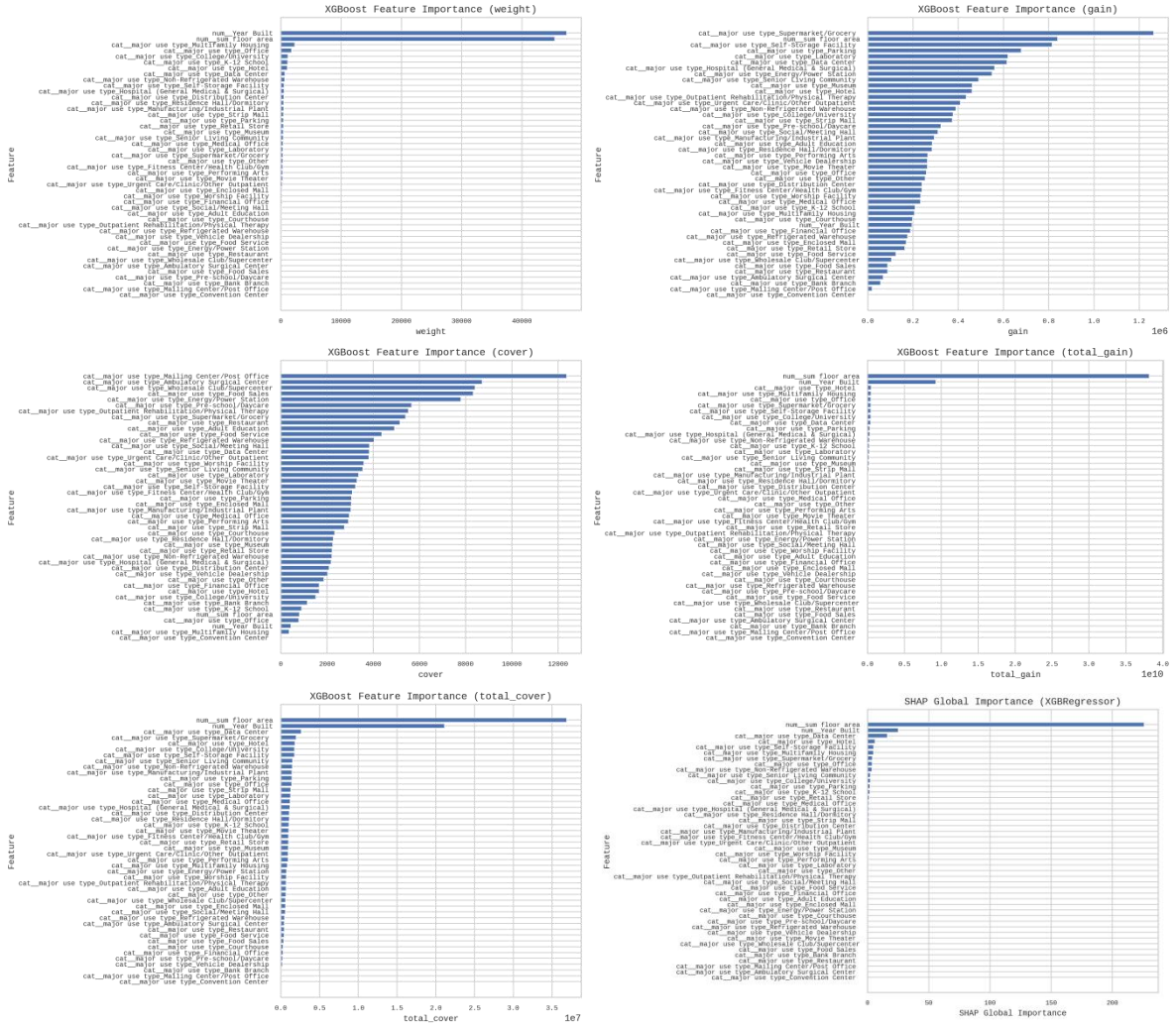

Figure 21. 6 Feature Importance Types of XGBoost

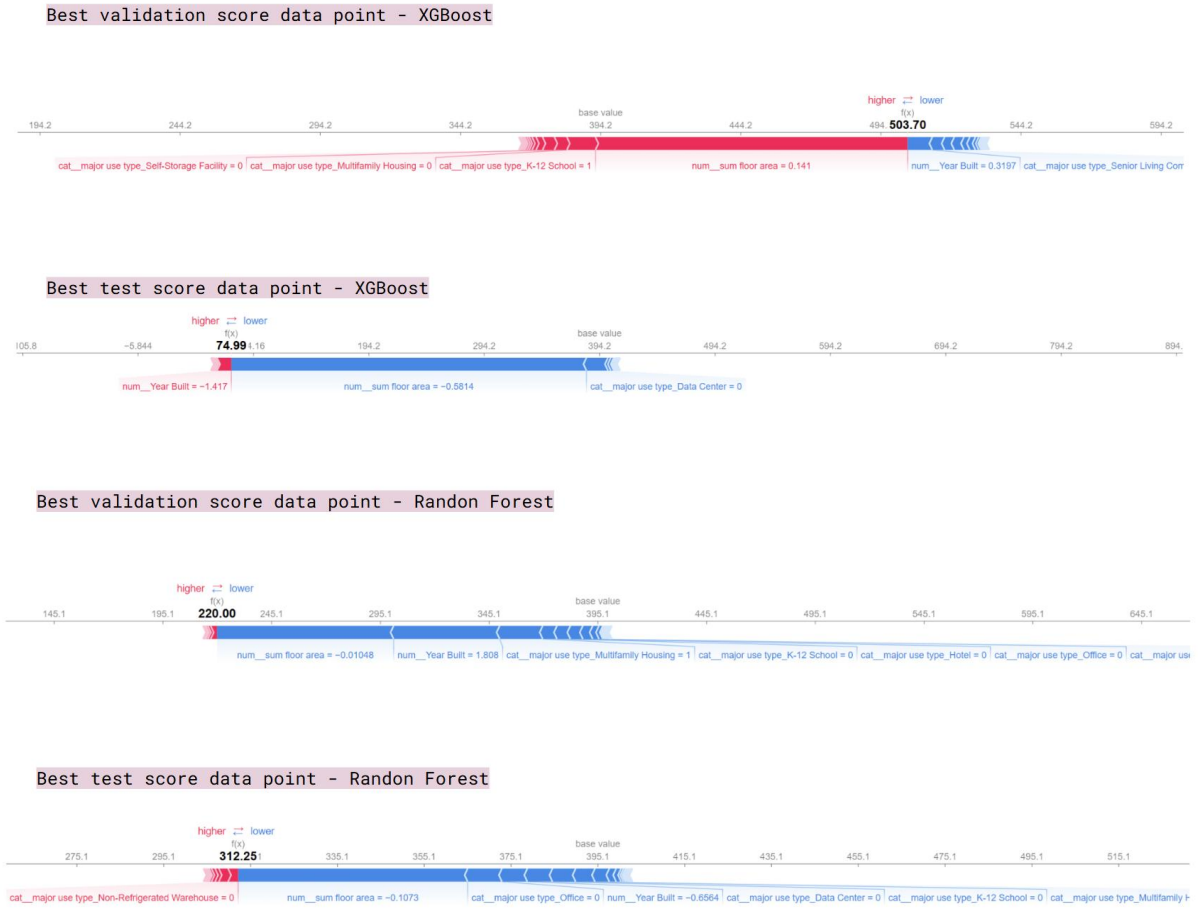

Figure 22. SHAP Local Feature Importance of XGBoost and RandomForest

Algorithm	Global or Local	Feature Importance Type	Top1	Top2	Top3
XGB Regressor	Global	Permutation Importance	num__sum floor area	num__Year Built	
		Weight	num__Year Built	num__sum floor area	cat__major use type_Multifamily Housing
		Gain	cat__major use type_Supermarket/Grocery	cat__major use type_Self-Storage Facility	num__sum floor area
		Cover	cat__major use type_Mailing Center/Post Office	cat__major use type_Ambulatory Surgical Center	cat__major use type_Wholesale Club/Supercenter
		Total gain	num__sum floor area	num__Year Built	cat__major use type_Hotel
		Total cover	num__sum floor area	num__Year Built	cat__major use type_Data Center
		SHAP global	num__sum floor area	num__Year Built	cat__major use type_Data Center
	Local	Shap local_best val score	num__sum floor area	major use type_K-12 School	num__Year Built
		Shap local_best test score	num__sum floor area	cat__major use type_Data Center	num__Year Built
Random Forest Regressor	Global	Permutation Importance	num__sum floor area	num__Year Built	major use type_Adult Education
		Gini importance	num__sum floor area	num__Year Built	major use type_Adult Education
	Local	Shap local_best val score	num__sum floor area	num__Year Built	cat__major use type_Multifamily Housing
		Shap local_best test score	num__sum floor area	cat__major use type_Office	num__Year Built
KNN	Global	Permutation Importance	num__sum floor area	num__Year Built	major use type_Adult Education

Table 10. Feature Importance Results Comparison

5. Outlook

5.1 Limitations

Data Quality: Some validation and test scores are better than training score for some algorithms even after scientific splits, which shows some features are strongly correlated to the target variable and over-fit the validation sets.

5.2 Future Directions

Enhanced Feature Engineering: Incorporate external datasets for energy efficiency, weather patterns, and socio-economic factors.

Model Robustness: Experiment with ensemble techniques (e.g., stacking) to combine strengths of linear and non-linear models.

Uncertainty Quantification: Apply Bayesian methods for more robust predictions under uncertainty.

Interpretability Tools: Expand SHAP analysis to investigate temporal trends in emissions and interactions between features.

Scalability: Adapt the pipeline for larger datasets, optimizing for computational efficiency.

6. Reference

[1] “NYC Building Energy and Water Data Disclosure for Local Law 84 (2023-Present) | NYC Open Data.” n.d. Data.cityofnewyork.us. https://data.cityofnewyork.us/Environment/NYC-Building-Energy-and-Water-Data-Disclosure-for-/5zyy-y8am/about_data.

[2] Gao, Huan, Xinke Wang, Kang Wu, Yarong Zheng, Qize Wang, Wei Shi, and Meng He. 2023. “A Review of Building Carbon Emission Accounting and Prediction Models.” Buildings 13 (7): 1617. https://doi.org/10.3390/buildings13071617.

[3] “Comparison of Forecasting Energy Consumption in Shandong, China Using the ARIMA Model, GM Model, and ARIMA-GM Model.” 2017. Sustainability 9 (7): 1181. https://doi.org/10.3390/su9071181.

[4] Feng, Wentao, Tailong Chen, Longsheng Li, Le Zhang, Bingyan Deng, Wei Liu, Jian Li, and Dongsheng Cai. 2024. “Application of Neural Networks on Carbon Emission Prediction: A Systematic Review and Comparison.” Energies 17 (7): 1628–28. https://doi.org/10.3390/en17071628.

[5] Zhao, Jinhui, Li Jingshun, Wang Panle, et al. "Study on Carbon Peaking Path in Henan Province Based on Lasso-BP Neural Network Model." Environmental Engineering 40, no. 12 (2022): 151–156, 164. https://doi.org/10.13205/j.hjgc.202212020.

[6] Hao, Jiaying, and Gao Jian. "Prediction Model of Building Carbon Emissions-Carbon Reduction Based on NSGA-II Improved BP Neural Network." Building Energy Conservation 44, no. 9 (2016): 122–124. (In Chinese). https://doi.org/10.3969/j.issn.1673-7237.2016.09.029.

[7] Wei, Yang, Zhengwei Chang, Pengchao Hu, Hongli Liu, Fuxin Li, and Yumin Chen. 2024. “Rapid Carbon Emission Measurement during the Building Operation Phase Based on PSO–SVM: Electric Big Data Perspective.” Frontiers in Energy Research 12 (April). https://doi.org/10.3389/fenrg.2024.1329942.