# Project Report on

# Energy Consumption and Summer Surface Temperature Prediction for the Buildings in New York City

**Course**
Foundation of Data Science (CS-GY 6053)

**Instructor**
Prof Rumi Chunara

**Team Members**
Jugal Pumbhadia (jp6988)
Agnes Park (ap2963)
Jay Desai (jd5558)

## Background and Problem Statement

Urban Heat Island (UHI) effect in cities like New York City causes higher temperatures in densely built areas compared to suburbs. Energy consumption by buildings contributes to this issue. Our project involves creating two machine learning models: one for predicting property energy usage and another for forecasting summer land surface temperatures. These models can help policymakers manage resources and optimize energy in areas with elevated temperatures.

## Prior Research, Studies, Visualizations and Domain Information:

1) https://www.usgs.gov/news/mapping-urban-heat-islands-leads-nyc-council-data-team-landsat
2) https://www.thecity.nyc/2023/07/26/heat-island-hot-map-temperature/
3) https://climate.mit.edu/explainers/urban-heat-islands
4) https://www.urbangreencouncil.org/new-york-citys-2020-energy-and-water-use-report/

## Target Variables

1) Energy Usage (kBtu): A numerical value measured in kBtu.
2) Land Surface Temperature (F): A numerical value measured in Fahrenheit.

## About the data

1) Energy and Water Consumption Data:
    - Source: NYC Open Data
    - Records: 25.2K
    - Features: 253 (Numeric and Categorical)
    - Link: https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/4tys-3tzj

2) Monthly Energy Consumption Data:
    - Source: NYC Open Data
    - Records: 310K
    - Features: 7 (Numeric and Categorical)

    - Link: https://data.cityofnewyork.us/Environment/Local-Law-84-2019-Monthly-Data-for-Calendar-Year-2/njuk-taxk

3) Average Land Surface Temperatures (LST) in Summer:
    - Source: NYC Government’s Environment and Health Data Portal
    - Records: 295
    - Features: 6 (Numeric and Categorical)
    - Link: https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/climate/?id=2141#display=links

## Prediction Problem

We are creating two models i.e. models to predict energy usage and summer land surface temperatures. We want to find if the existing dataset which is used to predict energy usage is enough to predict the summer surface temperatures.

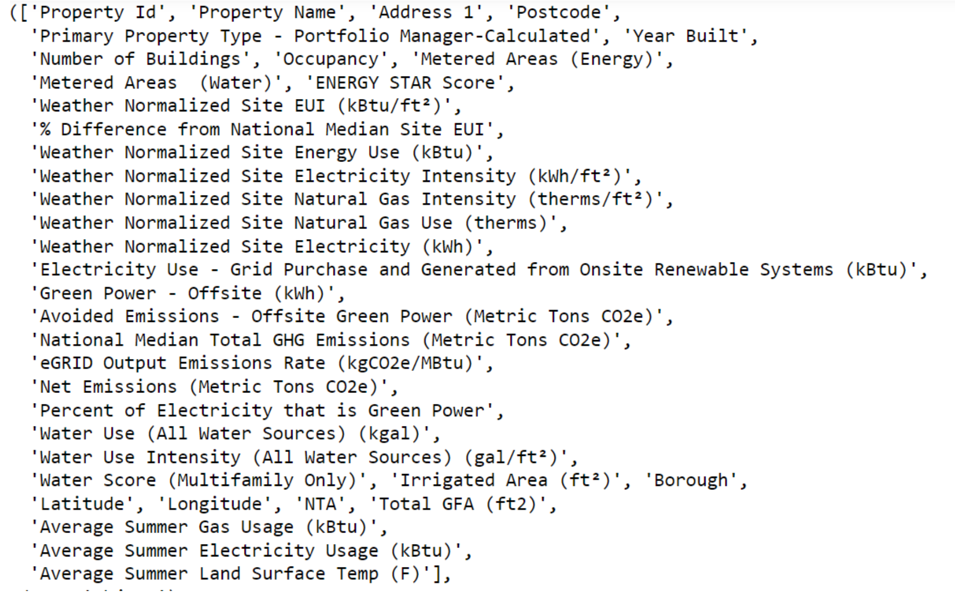
## Analysis Approach

### Data Cleaning and Preprocessing:

- Three datasets were used for the project, providing information on energy consumption, relevant indicators, and land surface temperatures for the calendar year 2018.
- In the first dataset, irrelevant and redundant features were removed, while some features were combined to reduce their number.
- Features with over 80% null records were eliminated, and missing values in the remaining features were filled using median values for numeric features and mode for categorical features.
- Inconsistencies in 'Year Built' and 'Postcodes' were corrected using external sources. Formatting issues in 'Year Built' and 'Property GFA' were rectified, and numeric data types were applied where needed.
- Outliers in numeric features were identified and imputed with median values.
- The second dataset, focused on summer months, was filtered to include only June, July, and August, resulting in 2 numeric and 3 categorical features.
- The third dataset, with no missing values, was filtered to match the 'NTA' standard for consistency with the first dataset.

## Data Merging:

- The first and second datasets were merged using the common 'Property Id' feature, incorporating imputed median values for null entries in the second dataset.
- The resulting combined dataset included a feature named NTA (locality). To align with the third dataset's 'NTA2010,' the datasets were merged based on this common 'NTA' feature, creating a comprehensive dataset for our final analysis and model building.
- After merging the data, we had the following predictor variables:



```
([' Property Id', 'Property Name', 'Address 1', 'Postcode',
  'Primary Property Type - Portfolio Manager-Calculated', 'Year Built',
  'Number of Buildings', 'Occupancy', 'Metered Areas (Energy)',
  'Metered Areas  (Water)', 'ENERGY STAR Score',
  'Weather Normalized Site EUI (kBtu/ft²)',
  '% Difference from National Median Site EUI',
  'Weather Normalized Site Energy Use (kBtu)',
  'Weather Normalized Site Electricity Intensity (kWh/ft²)',
  'Weather Normalized Site Natural Gas Intensity (therms/ft²)',
  'Weather Normalized Site Natural Gas Use (therms)',
  'Weather Normalized Site Electricity (kWh)',
  'Electricity Use - Grid Purchase and Generated from Onsite Renewable Systems (kBtu)',
  'Green Power - Offsite (kWh)',
  'Avoided Emissions - Offsite Green Power (Metric Tons CO2e)',
  'National Median Total GHG Emissions (Metric Tons CO2e)',
  'eGRID Output Emissions Rate (kgCO2e/MBtu)',
  'Net Emissions (Metric Tons CO2e)',
  'Percent of Electricity that is Green Power',
  'Water Use (All Water Sources) (kgal)',
  'Water Use Intensity (All Water Sources) (gal/ft²)',
  'Water Score (Multifamily Only)', 'Irrigated Area (ft²)', 'Borough',
  'Latitude', 'Longitude', 'NTA', 'Total GFA (ft2)',
  'Average Summer Gas Usage (kBtu)',
  'Average Summer Electricity Usage (kBtu)',
  'Average Summer Land Surface Temp (F)'],
```

## Model Building:

- The crucial model-building stage in this project involved algorithm selection, feature engineering, and training and evaluating two regression models for predicting 'Weather Normalized Site Energy Use (kBtu)' and 'Average Summer Land Surface Temp (F).'
- The chosen regression models included Lasso (L1) Regression, Ridge (L2) Regression, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regression (SVR).
- To address the 294 features in the input dataframe, Principal Components Analysis (PCA) was employed for dimensionality reduction. The optimal number of components (n_components = 12) was determined from the Scree Plot.

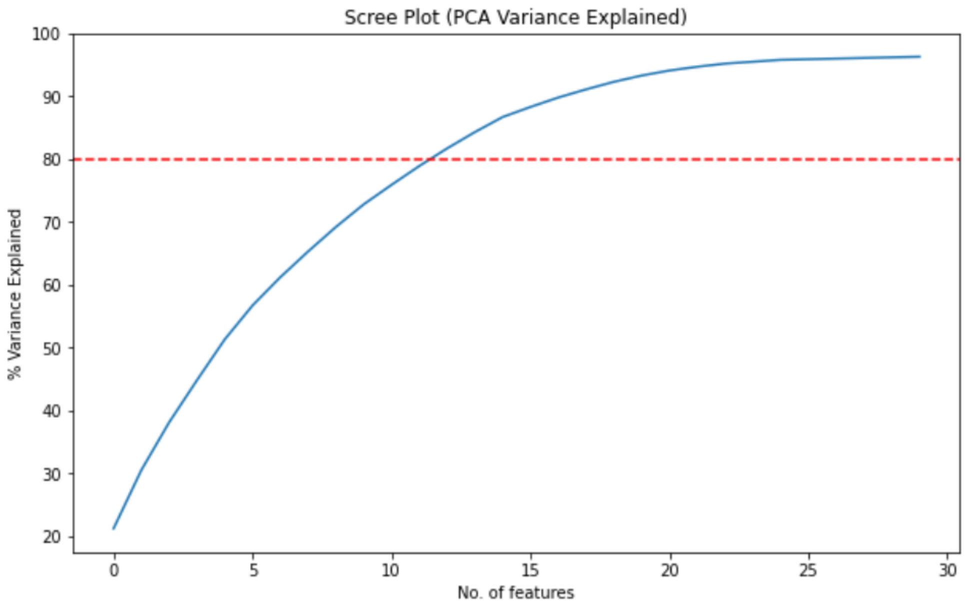


**Figure 8**: Scree Plot for PCA

## Model Hyperparameter Tuning:

- Hyperparameter tuning was performed using GridSearchCV library of sklearn.
- For the first model, Gradient Boosting Regressor with the parameters as learning_rate=0.1, max_depth=5 and n_estimators=50 was selected as the best model hyperparameters.
- For the second model, Decision Tree Regressor with n_estimators=50 was selected as the best model hyperparameter.

## Model Comparison:

We utilized the R2 score as the evaluation metric for both models for the following reasons:

- The R2 score gauges the proportion of variance in the dependent variable predictable from the independent variables, ranging from 0 to 1, where higher values signify a better fit.
- An R2 score of 1 indicates a perfect prediction, while 0 suggests the model fails to explain any variability. Negative R2 values signify a poor fit.

- Higher R2 scores denote models that better explain variability. When comparing models, the one with a higher R2 score is deemed more effective in predicting the target variable. Therefore, we chose R2 scores for model evaluation.

The R2 scores of different models are visualized below:

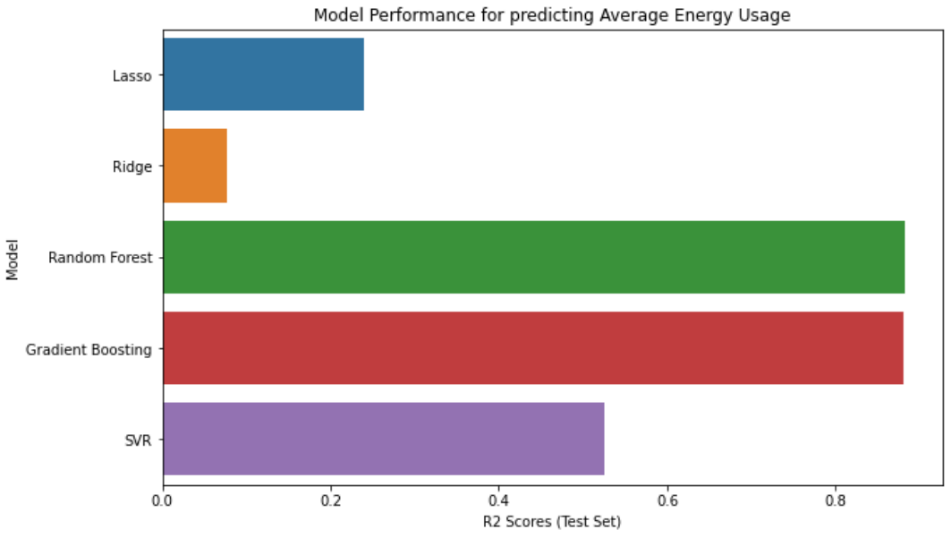


**Figure 9.1**: Random Forest and Gradient Boosting performed well for predicting energy usage

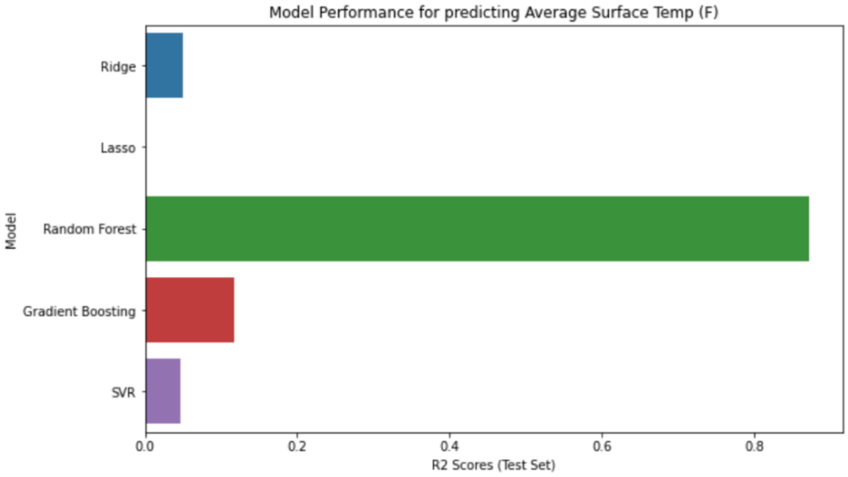


**Figure 9.2**: Random Forest was the best performer in predicting Average Surface Temperatures

# Model Results

## Evaluating the results of Model 1

- Model 1 focused on predicting energy usage in kBtu.
- Random Forest emerged as the top performer with an R2 Score of 0.882 on the test set. Gradient Boosting closely followed with an R2 of 0.881. In contrast, Ridge Regression exhibited poor performance, achieving an R2 Score of 0.07, indicating its limited ability to explain only 7% of the variance in the target variable using predictor features.
- Both Random Forest and Gradient Boosting, being ensemble methods, demonstrated high performance. Their ability to combine multiple weak learners allows them to capture complex relationships and handle noisy data effectively. Hyperparameter tuning likely played a role in optimizing their performance.
- Ridge regression's poor performance can be attributed to its penalty term, discouraging large coefficients. If the data's relationship is not well-described by a linear model, Ridge may struggle to capture underlying patterns. Additionally, Ridge is less effective in feature selection compared to Lasso, potentially resulting in suboptimal performance, especially with irrelevant features in the dataset.

## Evaluating the results of Model 2

- Model 2 focused on predicting average land surface temperatures during the summer season.
- Except for Random Forest, all models exhibited poor performance. Random Forest outperformed with an R2 Score of 0.87, suggesting its superiority in capturing underlying patterns compared to Ridge, Lasso, Gradient Boosting, and SVR.
- Ridge, Lasso, and SVR showed higher errors and lower R2 values, indicating challenges in capturing underlying relationships or potential overfitting.
- Gradient Boosting performed better than Ridge, Lasso, and SVR but lagged behind Random Forest, suggesting a middle-ground fit but still falling short compared to Random Forest.
- Thus, we can say that the existing data is not enough to predict Surface Temperatures for majority of model and more information is required to predict it.

## Assumptions:

The project operated under specific assumptions, primarily focusing on energy indicators based on site rather than source for local analysis. Outlier values were assumed to be genuine, and the impact of sensor and irrelevant data on the analysis was deemed negligible, leading to their removal from the dataset.

## Limitations:

The models lacked consideration for 12-month weather patterns and traffic data, presenting a limitation. Additionally, due to resource and time constraints, the models were trained on a limited set of hyperparameters. The second model, designed for predicting average surface temperature during summer, faced challenges, potentially stemming from an incorrect approach to data cleaning, joining, poor feature engineering, and suboptimal fitting of the training set.

## Conclusion:

Despite limitations, the project holds substantial future potential. Improvements in feature handling, meticulous management of missing values and outliers, and a detailed exploratory analysis can enhance model performance. Further, location-based analysis may unveil new insights. The project's outcomes have the potential to assist government, academia, and citizens of New York City in managing energy resources, emissions, and predicting energy usage and land surface temperatures during summer, ultimately enhancing the quality of life for the city's residents.

**Distribution of Points for team members:**

1) Jugal Pumbhadia (jp6988): 4 points
2) Agnes Park (ap2963): 4 points
3) Jay Desai (jd5558): 4 points