John Ferrara
DATA 698 – Capstone Project
DRAFT WORK PRODUCT

Modeled Estimates of Urban Canopy Effects on Building Energy Use

Introduction

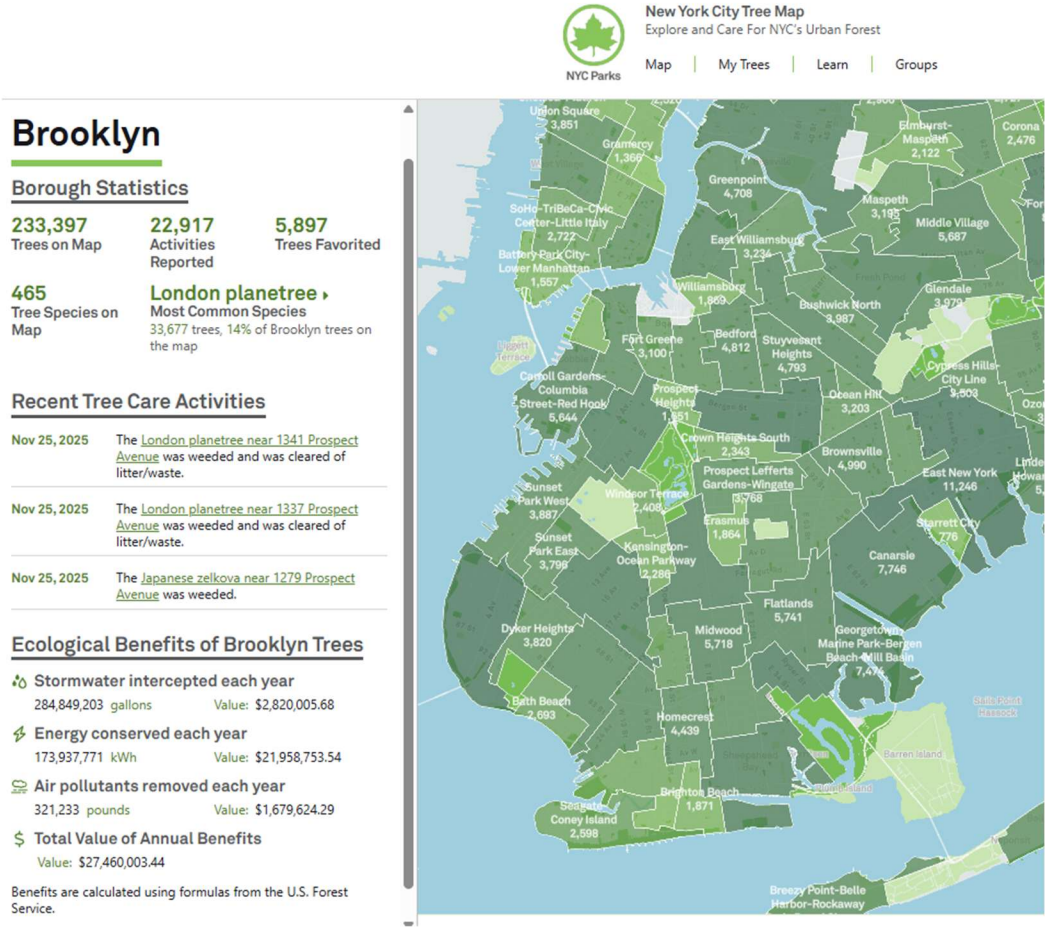
Background

The benefits of urban trees are fairly well known, particularly for their impact on localized temperatures. Urban trees help moderate harsh temperatures via evapotranspiration cooling effects, hedging against winter winds, and even absorbing potential flood water through their root systems during precipitation events.[i] As extreme weather occurrences, such as heat waves or polar vortex air systems, become more common in places like New York City, urban trees are a simple yet effective policy solution to help remedy some of the negative impacts of such weather.

Over the past decade or so, urban planning efforts have focused on increasing the level of urban trees planted, maintained, and otherwise integrated into holistic climate resilience plans by cities across the globe. As a major city, New York City, has multiple projects focused on leveraging urban trees to help enhance quality of life for residents. Projects such as "MillionTreesNYC"[ii] and "NYC Parks' Neighborhood Tree Planting"[iii] are just two examples of initiatives aimed at using trees as a constructive tool for quality of life. As highlighted, urban trees help remove air pollutants from local environment, reduce flash flooding events by slowing the flow of rainwater, lessen the negative impacts of severe heat waves, and can help reduce the overall energy use of near by buildings.

Besides the mentioned programs, New York City has additional tools allowing for users to examine such impacts of urban trees. The city has a Tree Map[iv] that lists the location of trees across the city. The data is tracked by underlying NYC tree data stemming from forestry programs like the NYC decennial Tree Count.[v] The map platform has multiple calculations allowing users to quantify the benefits of trees at different geographical layers and boundaries across the city with numeric estimations. For instance, when examining the platform's calculations for the borough of Brooklyn, a quantified total of $27,460,003 in estimated ecological benefits stemming from the presence of 233,397 trees in the borough (Figure 1).[vi] More specifically, for the purposes of this project, according to the Tree Map a total of 173,937,771 kWh of energy is listed as being conserved each year valuing over $21.9 million dollars.

Figure 1 : New York Tree Map Platform Looking at Brooklyn at the Borough Level[vii]



New York City's Tree Map platform highlights that these calculations originate from formulas provided by the USDA Forest Service's i-Tree software.[viii] i-Tree's website, states that the i-Tree platform is the "worldwide standard when it comes to quantifying the benefits that trees provide"[ix]. Various methodologies and technical papers are used to provide an estimate for these benefits.

Considering that New York City's buildings are the number one carbon pollution source for the city stemming from heating and cooling needs, predominantly fossil fuel burning for heating, its reasonable to examine impacts of urban trees and their presence on building energy use intensity.[x] This project uses a patchwork of publicly available dataset to analyze and model the impact of tree presence on building energy use within certain NYC building types. More specifically, does tree canopy coverage, or the number of trees, within 50 feet of a residential building explain any shift in weather normalized energy use intensity?

Overall, this study aims to answer the question: Do changes in tree presence, using urban tree canopy coverage and tree counts as proxies, produce measurable changes in weather normalized building energy use intensity for multifamily residential buildings in New York City between 2010 and 2017?

Potential Impact

The benefits of successfully quantifying some of the impact of urban trees on energy use intensity in New York City would be multi-pronged. Firstly, it would provide greater insight into this relationship helping to improve the estimates provided by the National Forest Service i-Tree's generalized equations.

Secondly, it would provide developers, city planners, and other decision makers with a nuanced understanding of tree impact on energy use. This would help with future developments, fiscal decisions, and potentially assist with regulatory compliance concerning energy efficiency.

Thirdly, due to this paper examining a specific subset of residential buildings covered by Local Law 84 reporting mandates,[xi] the quantification of energy impacts can yield a nuanced and specialized projection on tree presence on building energy use rather than a less-accurate generalized projection.

Lastly, as alluded to previously, this knowledge benefits New York City, as well as other cities with similar climates, as urban planning decisions for identifying ideal tree planting areas and building designs can be extrapolated using such methodologies.

Background and Literature

The literature review for this project started with the known foundations for the i-Tree platform, as this is the underpinning of NYC's Tree Map projections outlined in the introduction.

i-Tree Benefits & Value

As outlined, the United States Department of Agriculture Forest Service initially built and currently maintains the i-Tree platform, which allows the public to properly estimate the economic benefits of trees and the ecosystem services they provide.[xii] According to a 2007 paper by McPherson, et al., calculating tree value via a cost-focused framework compared to a benefit-focused framework can yield different results. As an example, a single green ash tree planted aged 40 years is values at $5,807 using cost-based approach. However,

with the benefit-focused framework the same tree is valued at $3,102 in Fort Collins, CO and $5,022 in Boulder, CO. The paper highlights how the benefit-focused approach includes, and allows for nuance in estimating value by "explicitly reflect[ing]" the impacts of "tree location on benefits such as energy savings."[xiii] This nuance in value system measurements supports the need for localized study and quantification of tree impact.

When looking specifically at tree presence and the direct impact on building energy usage, another paper stood out in the citations of the i-Tree reference ecosystem, a 1993 paper also by McPherson, et al.[xiv] The paper, *Energy conservation potential of urban tree planting,* estimates the direct and indirect impact of 25 foot trees, aged roughly 15 years, yielding 10-15% savings for cooling effects, with peak-demand estimates being closer to 8-10%. However, it should be noted, this study examined single family, two-story structures for these estimates.

Supplementing the conclusions of the 1993 paper, a 1999 paper, also from McPherson et al.[xv], *Carbon dioxide reduction through urban forestry: Guidelines for professional and volunteer tree planter,* looks at the impact trees have on reducing wind speeds and the subsequent heating costs for buildings. Shading impact on cooling for buildings was also examined. The paper breaks down the building age into three brackets: built before 1950, build between 1950 and 1980, and built post-1980. This categorization allowed for the controlling of various building techniques and materials in the structures. The structures themselves were predominately single family one or two story buildings. Lastly, the paper concluded that in the presence of three mature trees can yield a savings of 25 – 43% reduction in cooling needs and 12 – 23% reduction at peak demand window for these structures.

Modeling Tree Impacts

Expanding beyond the papers cited by the i-Tree ecosystem, more recent papers have also carried out similar analyses as this project. The overall findings of such studies are as follows:

- A 2021 study by He, et al., examined the correlation between the land surface temperature and the tree cover of two different geographic areas: Boston and Baltimore/Washington D.C.. The study targeted atypically hot days over the course of five years to derive their findings. The results indicate that there was a non-linear relationship between the two variables, particularly highlighting that the cooling rate stemming from the presence of trees increased as the temperature increased. In short, trees are dampening the extreme heat by increasing the rate of cooling for hotter temperatures.[xvi]

John Ferrara
DATA 698 – Capstone Project
DRAFT WORK PRODUCT

- A 2022 study by Zhu, et al., looked at low-rise structures in Nanjing, China and the impact that local vegetation had on heating and cooling in buildings. The presence of vegetation helps blunt the cooling and heating needs of such buildings.[xvii]

- A 2021 paper by Olu-Ajayi et al., examined various machine learning modeling techniques for forecasting building energy consumption. Artificial Neural Network (ANN), Gradient Boosting (GB), Deep Neural Network (DNN), Random Forest (RF), Stacking, K Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision tree (DT) and Linear Regression (LR) were looked at. The study concluded that Deep Neural Network (DNN) modeling had the best predictive results for modeling energy use. This study did not look at trees, but just building energy usages.[xviii]

- A 2021 paper by Tsoka et al., analyzed trees in Thessaloniki, Greece and concluded that energy savings of up to 54% have been yielded via continuous shading from tree canopy coverage. However, beyond direct shading of building facades, cooling impacts of street trees are relatively minor.[xix]

Apart from the studies referenced, the most relevant paper found was titled: *Quasi-experimental evidence that the urban tree canopy reduces residential energy consumption*.[xx] The 2025 paper by Ravazdezh et al. examined air conditioning in buildings in Ottawa, Canada. Specifically, the study looked at 2,000 single-family (R1) zoned structures and trees within 40 feet of these structures. Using a difference-in-differences modeling technique the study determined that a 10% increase in the urban tree canopy within 40 feet of a R1 zoning structure yields a 2.9% reduction in electricity use, specifically when the trees have leaves.

Difference-in-Differences Modeling

Based on the literature review, specifically the finding of the *Quasi-experimental evidence that the urban tree canopy reduces residential energy consumption* study, difference-in-differences (DID) modeling seems to be the best technique to use for modeling tree impact on energy consumption in New York City.

DID modeling, is a technique that predates the randomized experiment model[xxi] and is used to find the effect of one variable's impact in instances where multiple different variables may play a role in influencing the dependent variable being examined. Typically, this technique uses data from a period of time before the independent variable change and then a period after.[xxii] In short, the DID methodology allows for the analysis of an outcome shift stemming from a variable change. The technique compares two groups of units, one

impacted and one not impacted by the variable change, to model before and after the change in question.

Ridge & Lasso Regression Modeling

Other than the DID modeling technique, Ridge and Lasso regressions should be used to examine the relationship between energy use in buildings, the presence of trees, and other building-level features. Using these two types of regression methods, will allow for the prediction of impact stemming from trees on energy use in buildings, as well as the influence of the individual building features retained in the heavily processed data.

While one of the literature review studies found a non-linear relationship between vegetation and building energy use, these regression techniques, which are built on top of the foundational ordinary least squares regression model, penalization terms in their formulas to help reduce overfitting and deal with multicollinearity related issues. Having multiple features, like such as canopy change category, tree count, building height, and number of floors, that are likely to show collinearity, these ridge and lasso regression are well suited to attempt prediction modeling for changes in building energy use intensity.

Main Takeaways

Multiple papers support the premise that urban trees and their impact on energy usage is an area of interest that needs more study. Specifically, looking at distinct geographies, building types, and other niche area not fully examined. For instance, most of the studies found looked at single-family homes. Not other building types. In addition, the dense, mixed-used or high rise housing more common to New York City seems to be under examined.

While there are a multitude of models in the i-Tree suite of tools, this project's goal is to identify a tangible relationship between trees and building energy usage in a unique environment like New York City. As a result, the modeling technique most appropriate for this type of analysis is the difference-in-differences model. Building off the 2025 Ottawa study, the impacts of New York City's canopy and its potential impact one residential buildings needs study. Additionally, the ridge and lasso regression modeling techniques will provide complementary insights on tree presence along side a broad set of building characteristics.

Data

Overview

Having established the appropriate modeling techniques and grounded this project within necessary background information for context, the next step is to look at the data that will be used for this analysis. As expected, there's no clear data set already created for New York City concerning this area of study. For the purposes of this analysis, one will be constructed using a multitude of different publicly available data sets from city-level agencies, sourced predominantly from New York Open Data.[xxiii] While a multitude of supplementary data sources were used in this project, there were several foundational data sources.

Data Foundations

NYC Building Energy and Water Data Disclosure for Local Law 84 (2010 – 2024)[xxiv]

With the implementation of Local Law 84 in 2009,[xxv] and subsequent amendments of the law,[xxvi] data on energy use for buildings that meet the criteria outlined in the legislation has been recorded since 2010. This data provides insights into energy and water usage for buildings of various types. The law particularly mandates reporting via the ENERGY STAR Portfolio Manager tool[xxvii] for those buildings over 25,000 square feet, those buildings owned by the city government, those buildings that are two or more structures on the same tax lot that exceed 100,000 square feet, and those under same condominium board that exceed 100,000 square feet. Specifically, for this project, the datapoint most critical for analysis is the weather normalized energy usage intensity (EUI) metric. It "measures the amount of energy use one can expect to see on a building's utility bill given normal weather conditions, i.e., if there is not a significant heat wave in the summer or a milder-than-average winter."[xxviii]

Beyond energy usage, information on the building structures themselves is also contained in this data. Critical contextual data points such as unique identifiers like BBL (Borough, Block, Lot) ids, energy source types, energy meter information, building use (e.g., commercial, residential, etc.), and other information. The raw data messy with many nulls in the peripheral columns.

Tree Canopy Change (2010 – 2017)[xxix]

A dataset derived from Light Detection and Ranging (LiDAR) highlighting the difference in tree canopy coverage in New York City from 2010 through 2017. Using LiDAR, various canopy coverage in 2017 was categorized as either having: No Change, Gain, or Loss when compared to the same scan in 2010. No change indicates no shifts in canopy coverage

from trees in that area, while a gain or loss value would indicate a gain in forestry canopy coverage since 2010, or a loss in coverage, respectively.

Forestry Tree Points[xxx]

This data set is a tree inventory for the city. Its composed of latitude and longitude values for each tree in the dataset. Of the several forestry datasets available on NYC Open Data, this dataset is the core dataset for tree inventory. Addiitonal information can be joined with this dataset, such as Service Requests, Inspections, and Work Orders.

Forestry Work Orders Points[xxxi]

A supplementary dataset for tree point Work Orders, which are used log actions taken by NYC Park staff, contractors or others. This data specifically is used in order to identify any potential planting dates or dates and locations where a street tree was removed.

Supplementary Datasets

NYC Planning Labs API[xxxii]

This data source was used to obtain the Borough, Block, Lot (BBL) number for each of the buildings where it was needed. Where there were null values in the BBL section of the Local Law 84 building data, address-oriented fields were used in order to obtain a BBL from this API to enrich the data.

PLUTO Building Data[xxxiii]

Once the BBL identifier was obtained for those that had null values, the BBLs were used to pull in additional information on those buildings. Such as the number of floors in the building, year the building was build, roof height of the building, and other similar types of city planning metrics.

Google Maps API[xxxiv]

Google Maps API was used for geocoding Local Law 84 properties that did not have latitude and longitude values populated in the raw data. This was done using address oriented columns.

2010 Census Tract Shapefile[xxxv]

The shapefile for Census Tracts within New York City was obtained and used in order to enrich the Local Law 84 buildings data with census tract. For potential mapping or other needs.

Building Footprint Shapefile[xxxvi]

This data set, which is published by the NYC Office of Technology and innovation (OTI), is the geographic boundaries for individual buildings within NYC. This is specifically useful, as geocoding addresses yield a centroid. However, for this project the trees need to be within a specific proximity to a building, which means the structure's perimeter must be obtained in order to have an accurate analysis.

Methodology

To carry out this analysis, as outlined, data was pulled from multiple different sources. Null values were dealt with through additional enrichment through a variety of methods. There were three main scripts used for processing the data into one coherent working dataset for analysis.

Part 1: Buildings Data[xxxvii]

The first part of this process included pulling in the initial Local Law 84 data via the Socrata API on NYC Open Data. This data was ingested for every year of coverage from 2010 through 2024. However, after reviewing what tree data was available, this was eventually shifted to only include data from 2010 and 2017. Data for these years was ingested, and cleaned. For cleaning, columns containing information on dimensions of data irrelevant to the analysis were dropped.

Additional filtering on the remaining columns was completed for further data refining. For instance, a column named "metered_areas_energy" was limited to null values and those indicating that the entire property was metered for energy usage. In short, values for "Whole Building" or "Whole Property" were kept, as well as null instances due to the column having a large number of null values across multiple years. For many columns in this aggregated dataset there were null values were present, particularly for earlier years. It seems that additional columns were added over time, so earlier years just have null values post-dataframe concatenation with later years. With this in mind, when filtering on most peripheral columns, null values were kept due to this exact issue. The filtering was over inclusive in this way as opposed to under inclusive.

Self-enrichment, and additional variables were created using the context of the dataset itself. For instance property Identification numbers, which had completely null values in 2010 and 2011, were enriched via BBL identification numbers from later years. The data was limited to include only residential properties using relevant columns such as "list_of_all_property_use", "primary_property_type", etc.

After ensuring that self-enrichment options were exhausted, traditional enrichment with supplementary datasets was carried out. While many properties in the data had latitude and longitude values for a geographic point, many didn't. Removing those instances where address information could not be salvaged from null values, those with address-focused fields with enough information available were cleaned up and used in order to geocode the data using Google Maps API. Having coordinates for all of the properties in the dataset allows for the use of spatial joins later in the pipeline when needed.

The cleaned addresses generated for the Google Maps geocoding were also used with the NYC Planning Labs API to pull in BBL values for those properties where it was missing. Once this was completed, with all properties having BBLs and geographic coordinates, MapPluto data was used to pull in metrics like the number of floors in the building, year the building was built, roof height of the building, and other similar types of city planning metrics via the BBL number.

Further additions to the running dataset came from a series of joins to enrich the data. A spatial join using the 2010 Census Tract shapefile. The Census Tract was added for each property, using the intersection of Census Tract geography and the geographic point for that respective property. A join via BBL was used in order to pull in geographies for building footprints, essentially removing the limitation of a single fixed point for a building. These building footprints are a better representation of where a building has presence, allowing for more accurate tree information for the area around each building.

Lastly, the Forestry Canopy Coverage dataset was spatially joined into the data. This Lidar-based dataset outlining if tree canopy coverage increased, decreased, or did not change from 2010 through 2017, was added to the data via a nearest neighbor spatial join. The join allow for a 50 foot buffer around building footprints to properly capture the relevant canopy coverage for the area around a building.

Now that most of the data for the properties was enriched, additional filtering, cleaning, and refining was carried out to yield the working building dataset. Most importantly in this section was the filtering using energy meter alerts and records where energy use is estimated to remove unreliable data from the working set. Renaming of columns from different years in order to yield a standardized nomenclature across all years was also carried out.

Ultimately, while all of the years pulled from the API were processed, only building information for both 2010 and 2017 were kept. The BBLs that were present in both years were kept in the data, and the overall dataset was limited to those years due to the canopy data timeframe. The final table was exported to CSV for use in Part3.

Table 1 – Part 1 Yielded Building Data Sample[1]

Original Column Name	Example Value	Column Description
year_ending_year	2010	Calendar/fiscal year of data
property_id	2552023	Internal property identifier
bbl	1022020009	Borough–Block–Lot ID
address_1	420 West 206th Street	Primary street address
city	New York	City name
gfa_building_ft2	77967	Gross floor area of building (ft^2)
site_eui_kbtu_ft	101.2	Site EUI ($kBtu/ft^2$)
site_energy_use_kbtu	7,892,541.1	Total site energy use (kBtu)
source_energy_use_kbtu	11,458,759.9	Total source energy use (kBtu)
postcode	10034	ZIP code
Borough	Manhattan	NYC borough
BldgClass	D4	Building class code
LandUse	3	Land use code
NumFloors	6	Number of floors
UnitsRes	74	Number of residential units
canopy_change_class	No Change	Tree canopy change classification

Part 2: Tree & Forestry Data[xxxviii]

This section pulled in Street Tree and other Forestry data from NYC Open Data. The datasets ingested via Socrata API were Forestry Tree Data points and the Working Orders for forestry trees. The Forestry Tree data points were used as the best, non-static count of trees in the city. While most of these points in this dataset, seemingly were created with the Tree Census of 2015, this dataset had been updated since.

The forestry points were limited by date columns, “createddate” and “planteddate”, in order to keep only the trees that existed from 2010 through 2017, which is the same window as the Canopy Change data joined into the data in Part 1. More specifically, instances where “createddate” values were after the end of 2017 were removed, while “planteddate” values after 2009 are also excluded on the assumption they are too young to impact winds, or provide substantial shade. For both of these columns, there were null values, so these nulls were treated as older trees or unknown values and kept in the data. Furthermore, using “tpcondition” and “tpstructure” columns dead trees or tree stumps were dropped from the data as well.

Forestry workorder data was also folded into this by joining on a unique identifier for tree points. Workorder data was limited to those tickets that were closed, as the analysis looks

[1] Table shows a subset of key variables from the final building dataset. Additional fields (not shown) include building age, stories, census tract IDs, weather-normalized energy variables, quality-control flags, and geometry/footprint and canopy fields used in the spatial analysis.

at previous years. The workorder data was used to highlight any tree removals that took place between 2010 and 2017. For estimating a tree removal and tree count consideration within a year, if a tree was removed in July or later they are included in the year count, if they are removed before that points they are excluded.

These two data sets were used in conjunction with one another to generate a long data table that accounted for each tree estimated to be in existence before by 2010 and then the assumption is made that the tree is still present unless it is removed. New plantings are not added in, as they are not in these data sets and the assumption is that young trees will have negligible impact on building energy. This final table was exported to CSV for use in Part3.

Table 2 – Part 2 Yielded Forestry Data Sample

Original Column Name	Example Value	Column Description
objectid	230120	Internal object / record ID
dbh	18	Diameter at breast height (inches)
tpstructure	Full	Tree structure / crown fullness (e.g., Full, Retired)
tpcondition	Good	Tree condition (e.g., Good, Unknown)
plantingspaceglobalid	B9DDFFE7-7387-4923-91EA-6E9212AA324F	Unique ID for planting space
geometry	POINT(-73.93851920790104 40.60738960999758)	Geometry in WKT (Well-Known Text) format
globalid	FF71F967-C0E7-478E-BD3A-C54A2927A624	Unique ID for the tree record
genusspecies	*Acer - maple*	Botanical genus and common name
createddate	9/4/2015 14:54	Date/time record was created
location	{'type': 'Point', 'coordinates': [-73.9385, 40.6074]}	Geometry as GeoJSON
stumpdiameter	0	Stump diameter in inches (0 if not a stump)
updateddate	*(blank in sample)*	Last updated timestamp (if available)
riskrating	*(blank / not set in sample)*	Tree risk rating
riskratingdate	*(blank in sample)*	Date risk rating was assigned
planteddate	*(blank in sample)*	Date the tree was planted
lat	40.60738961	Latitude (decimal degrees)
lon	-73.93851921	Longitude (decimal degrees)
removal_date_est	*(blank in sample)*	Estimated removal date
removed_before_2018	FALSE	TRUE/FALSE flag if removed before 2018
removal_year	*(blank in sample)*	Year the tree was removed
include_in_year	*(blank in sample)*	Flag to include record in a given reporting year
manual_year	2010	Manually assigned year for reporting / analysis

Part 3: Aggregation and Analysis[xxxix]

The third and final part of the data processing for this project includes reading in both the datasets stemming from Part 1 and Part 2: the building data and the forestry data.

Firstly, the years were limited to 2010 and 2017. Only BBLs that existed in both years were kept for this analysis. See Table 3 below for a break down of unique BBLs left in each year and canopy change category. Additionally, only those buildings that had non-null values for

the "weather_normalized_site_eui" column were kept, as this column is critical for the analysis. In short those BBLs with valid energy data that exist in the data in 2010 and 2017 are kept for analysis.

Table 3 – Unique BBL Count After Initial Year & EUI Filtering and BBL Limiting

Canopy Change	Count 2010	Count 2017
Gain	202	202
Loss	107	107
No Change	1,093	1,093

Starting with this newly limited data, several other cleaning steps are carried out to yield the final working dataset. The "year_built" column was reinforced by coalescing "year_built" and "construction_year" columns, so as to ensure a value was present. Those instances that still remained null, were enriched using the NYC OpenData API to pull in PLUTO data via BBL. Additionally, all instances of null values for the "num_floors" column, which had the number of floors in the building, were dropped too. BBLs that were associated with multiple variations of addresses were manually reviewed, modified, and corrected to ensure a 1:1 relationship between BBLs and address.

Using this refined data, a spatial join between the tree-based latitude/longitude points and the building footprint geographies was executed. This join allowed for a buffer of 50 feet around the building in order to capture nearby trees. Trees that did not yield any BBL within 50ft were dropped from the data, and BBLs with no trees were not carried forward as an inner join was used. The data was aggregated up to remove the tree species data and just have a total number of tree counts for each BBL for each year. To aggregate the data, the following columns were used in the "group by" clause, with the subsequent columns being summed or averaged for each instance.

- Columns Grouped By: year, ct2010, bbl, year_built, BldgClass, NumFloors, canopy_change_class
- Aggregated Columns: tree_count (summed); UnitsRes, ground_elevation, height_roof, weather_normalized_site_eui (averaged)

Table 4 – Unique BBL Count After Building Footprint Spatial Join with Tree Data

Canopy Change	Count 2010	Count 2017
Gain	199	199
Loss	104	104
No Change	1,083	1,083

Finally, an ultimate round of data cuts and feature engineering were carries out to further help control for externally shifting variables. These steps were performed in the following order:

- Extreme outlier values for the "weather_normalized_site_eui" column were identified and removed. Those values that were above the 99th percentile and those that were below the 1st percentiles of column values were identified and dropped. Essentially removing extremes to improve distribution normalcy.
- Building Classification and zoning classes, the values found in the "BldgClass" column, were limited to include those in class "C" or "D", as these classes are residential buildings. Class C are those buildings with no elevator, and Class D are those buildings with an elevator.[xl]
- Using the "BldgClass" column, a flagging feature was engineered. The "commercial_floor_flag" column was created to flag those buildings that have commercial zoning on the ground floor. Classification values 'C7', 'D6', and 'D7' are flagged as having a commercial first floor, while the rest of the values are purely residential.[xli] This allows for the controlling potential business turnover, which depending on the types of businesses, would yield energy use fluctuations independent of the building and surrounding trees.
- The working dataset was then limited to those buildings that were 6 stories or under.
- Buildings in the data that had a "year_built" constructed before 1900 were removed.
- A categorical variable was engineered using the "year_built" column to group the buildings into three categories: Pre-1950, 1951-1980, Post-1980. This was modeled after McPherson et al. (1999)[xlii].

Table 5 – Unique BBL Count After Final Filtering and Control Limitations

Canopy Change	Count 2010	Count 2017
Gain	93	93
Loss	36	36
No Change	386	386

After these last processing steps were completed the data was ready for analysis. Table 5 contains the updated counts by Canopy category and year, While Table 6 contains information on the structure of the final dataset.

Table 5 – Final Dataset for Analysis

Original Column Name	Example Value	Column Description
year	2010	Reporting / analysis year
ct2010	001900	2010 Census tract code for the building
bbl	2023090001	Borough–Block–Lot ID for the building
year_built	1920	Year the building was constructed
year_built_bracket	Pre-1950	Categorical age band derived from year_built
BldgClass	D7	NYC building class code describing the building type / use
NumFloors	5.0	Number of floors in the building
UnitsRes	75.0	Number of residential units
ground_elevation	9.0	Ground elevation (units as in source data, likely feet)
height_roof	40.7	Height of roof above ground.
canopy_change_class	Gain	Tree canopy change classification
weather_normalized_site_eui	92.68	Weather-normalized site EUI (kBtu/ft^2)
tree_count	12	Total number of trees within 50 ft of the building footprint
commercial_floor_flag	1	Commercially zoned first floor flag (1/0)

Analysis & Modeling

Exploratory Analysis

Preliminary examination of the finalized dataset shows where the buildings that fit the filtered framework exist within the city. After multiple rounds of filtering, limiting uncontrollable variables, and data processing to obtain a clean dataset for a subset of uniform buildings within New York City, the buildings used in this analysis seem to be mostly confined to a small subset of Census Tracts in specific regions of the city.

Looking at Figure 2, most of the buildings in the analysis are within northern Manhattan, particularly the Inwood and Washington Heights areas, parts of the Bronx, Queens,

Brooklyn, and then also in select neighborhoods in Staten Island. For the Bronx the buildings seem to be clustered in the Bronx neighborhoods of Kingsbridge, Concourse and Clairmont. In Queens, they're in the neighborhoods of Corona, Jackson Heights and Elmhurst. Within Brooklyn the buildings show up in neighborhoods of Clinton Hill and Bed-Stuy. Lastly, in Staten Island the BBLs show up in the neighborhoods of Fresh Kills and Richmond.

For the Canopy Change data, when mapped as it is in Figure 3, there are many places in the city where the Tree Canopy had no change, however, when looking at the maps of the loss and gain classes, there seems to be fewer instances of extreme shifts. This may not be ideal for the analysis, as the focus is to identify shifts in the influence of canopy coverage and it's impact on energy usage intensity in buildings within the city.

In Figure 4 this becomes more evident. The count of buildings for each canopy change category is uneven across the classes. The number of buildings in the "No Change" category was by far the largest by orders of magnitude. While both the "Gain" and "Loss" categories had substantially lower building counts in the dataset, with the "Loss" category having the lowest number of instances in the data.

To finalize the understanding of the building in the final dataset, Figure 4 and Figure 5 show differing dimensions of insightful building characteristics. Figure 5 showcases the number of buildings that are in the each of the brackets created from the 'year_built' column. The vast majority of the buildings are from those built before 1950, with much smaller counts of buildings falling into the more modern brackets of built post-1980, and built between 1950 and 1980. Furthermore, the canopy change categories are much more spread out in the most populous built pre-1950 category than the other two.

Lastly, Figure 6 shows the frequency of buildings in the data by NYC Building Classification code.[xliii] The most common category in the data is D1, followed by D4 and D7. These categories are defined as:

- D1 Classification: ELEVATOR APARTMENTS - ELEVATOR APT; SEMI-FIREPROOF WITHOUT STORES
- D4 Classification: ELEVATOR APARTMENTS - ELEVATOR COOPERATIVE
- D7 Classification: ELEVATOR APARTMENTS - ELEVATOR APT; SEMI-FIREPROOF WITH STORES

In other words, the most common buildings in the analysis are apartment buildings with elevators.

Figure 2 – Analyzed BBLs by Census Tract

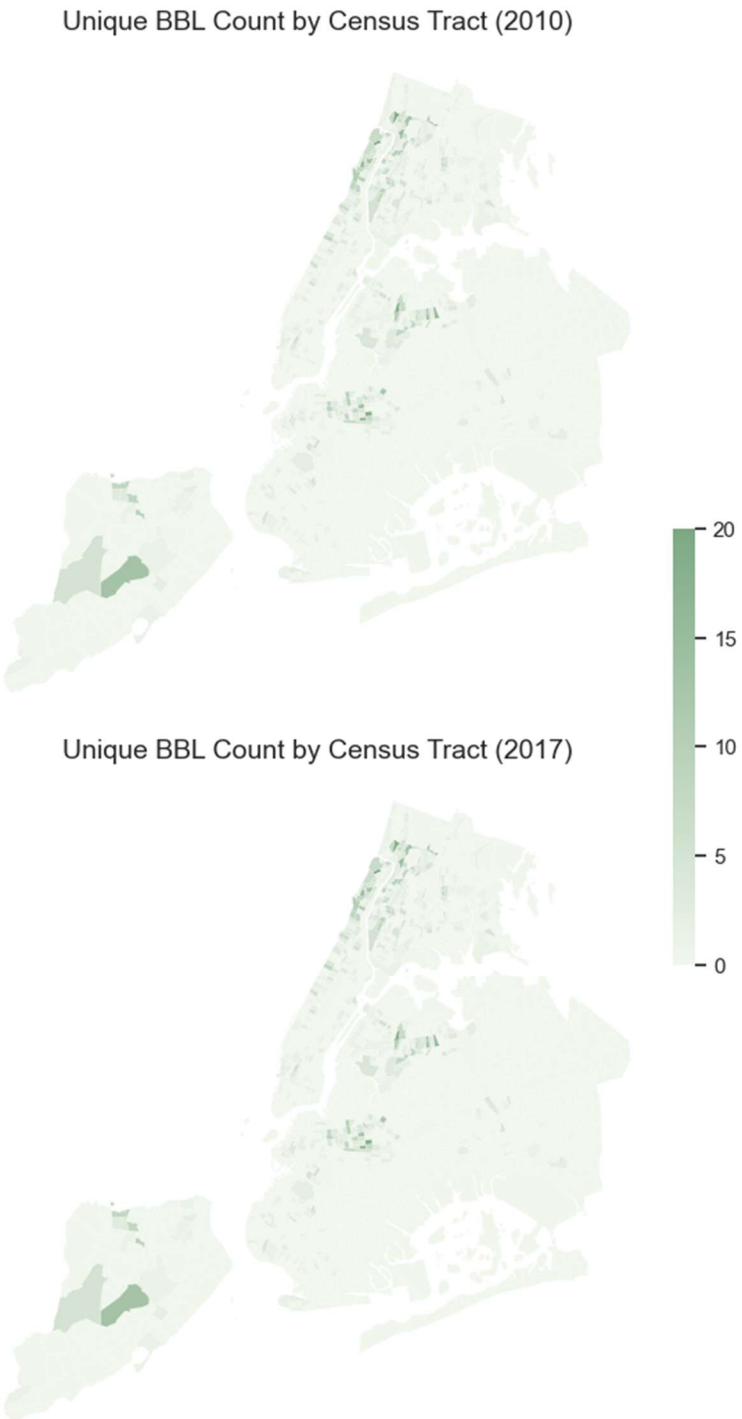


Figure 3 – LiDAR-Based Canopy Change by Category

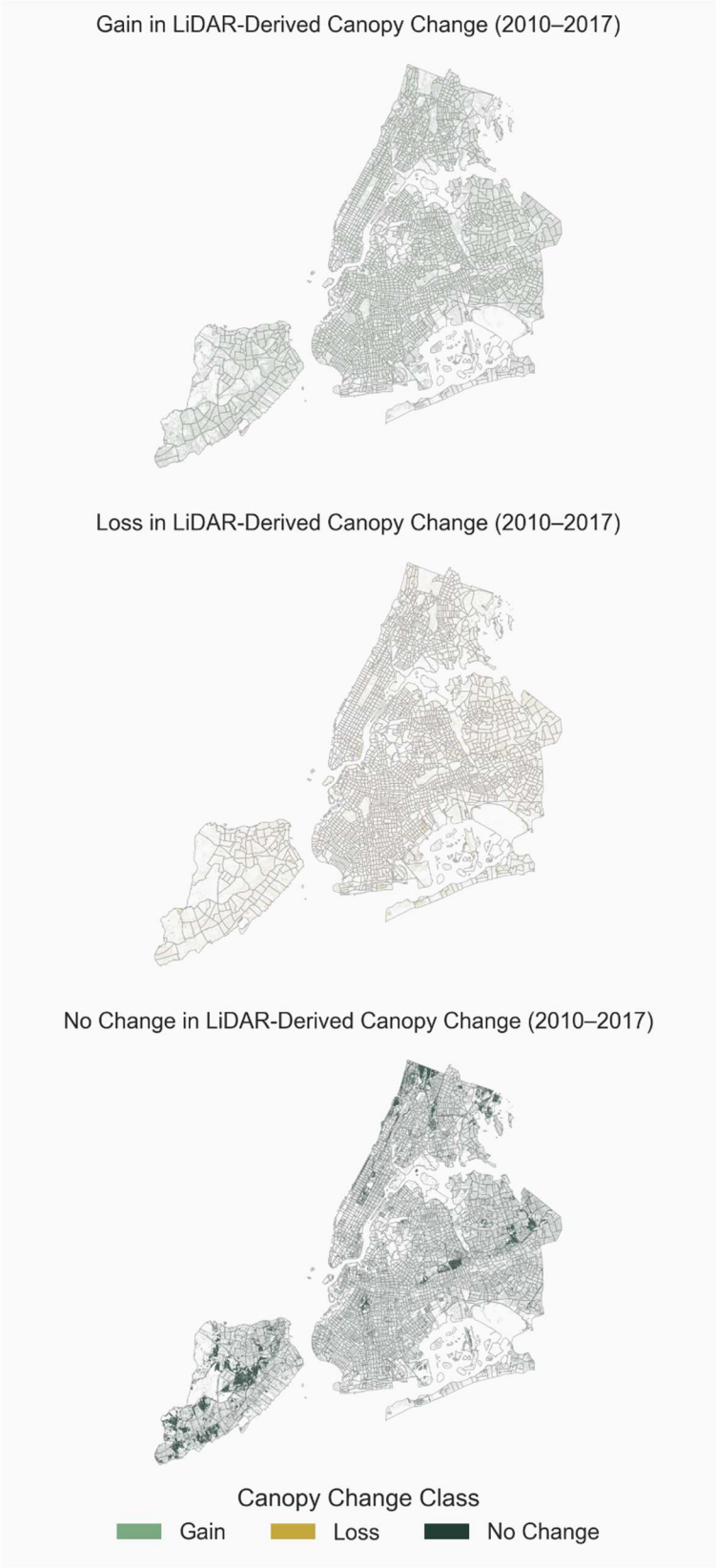


Figure 4 –Building Count by Canopy Change Class

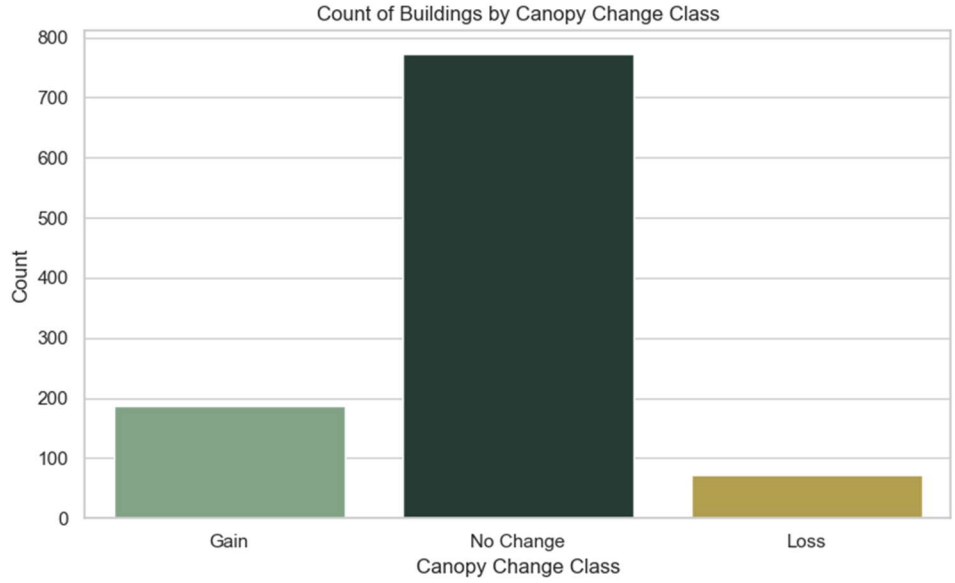


Figure 5 –Building Counts by Year Built Bracket

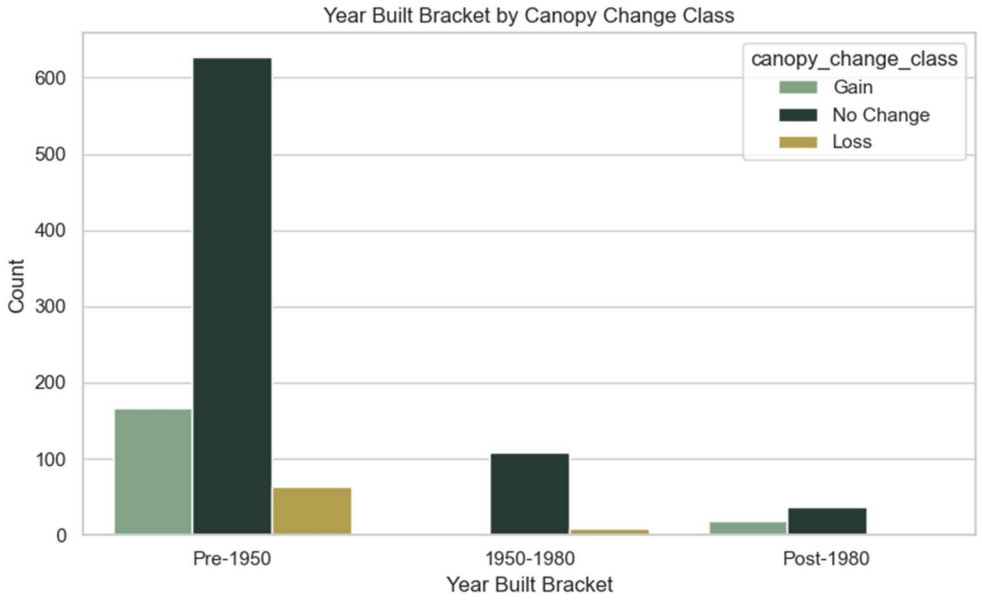
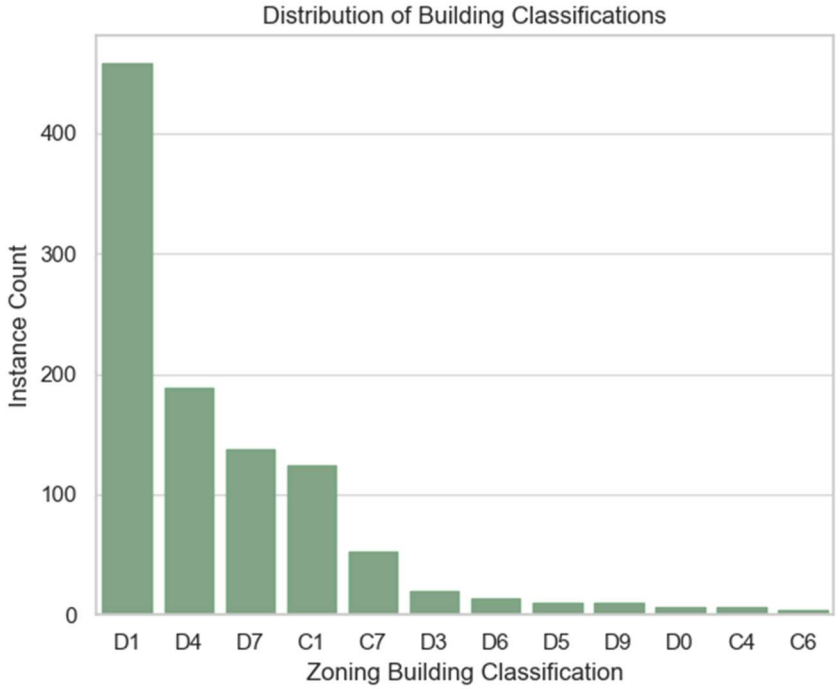


Figure 6 – Building Zoning Class Counts



After placing the data into a geospatial context to obtain a sense of where in the city the examined buildings exist, where in the city canopy coverage has shifted the most, and understanding what kinds of buildings are being examined. The next step is to look at the features that make up the dataset used for this analysis. To start, the numeric features were placed into a pair plot (Figure 7) and a Correlation Matrix (Figure 8) to examine the relationships between these different variables. The most main focus when reviewing these variables is the Weather Normalized Site EUI feature and its relationships to the other features, particularly that of tree count. However, upon examination of the pair plot, there doesn't seem to be any visual indication of a strong correlation between any of the Weather Normalized Site EUI and other key features.

The correlation matrix (Figure 8) reinforces this visual takeaway by quantifying the relationships between each of the numeric features. Of all of the numeric features in the final dataset, the strongest direct correlation, with 29%, in this matrix was the roof height and the number of floors in the building, which is expected and not relevant to the analysis. However, looking the Weather Normalized Site EUI the strongest correlation for this feature was that between the Number of Floors in a building and the year the building was built. Both of these features have a -8% correlation with the Weather Normalized Site EUI, which means that the more recent the building the lower the Weather Normalized Site EUI value

is. This makes sense, as more modern buildings tend to be more energy efficient. However, the number of floors correlation with the Weather Normalized Site EUI doesn't make sense on its face. This correlation may actually have to do with the height of the building being related to the age of the building. As in, for instance, if 6-story buildings were much more common pre-1950 then it is just another way to look at the same insight. There is a 9% direct correlation between the age of the building and then number of floors as well, the higher number of floors could skew towards of a certain construction era.

Figure 7 – Pair Plot Numeric Features With Canopy Change Class

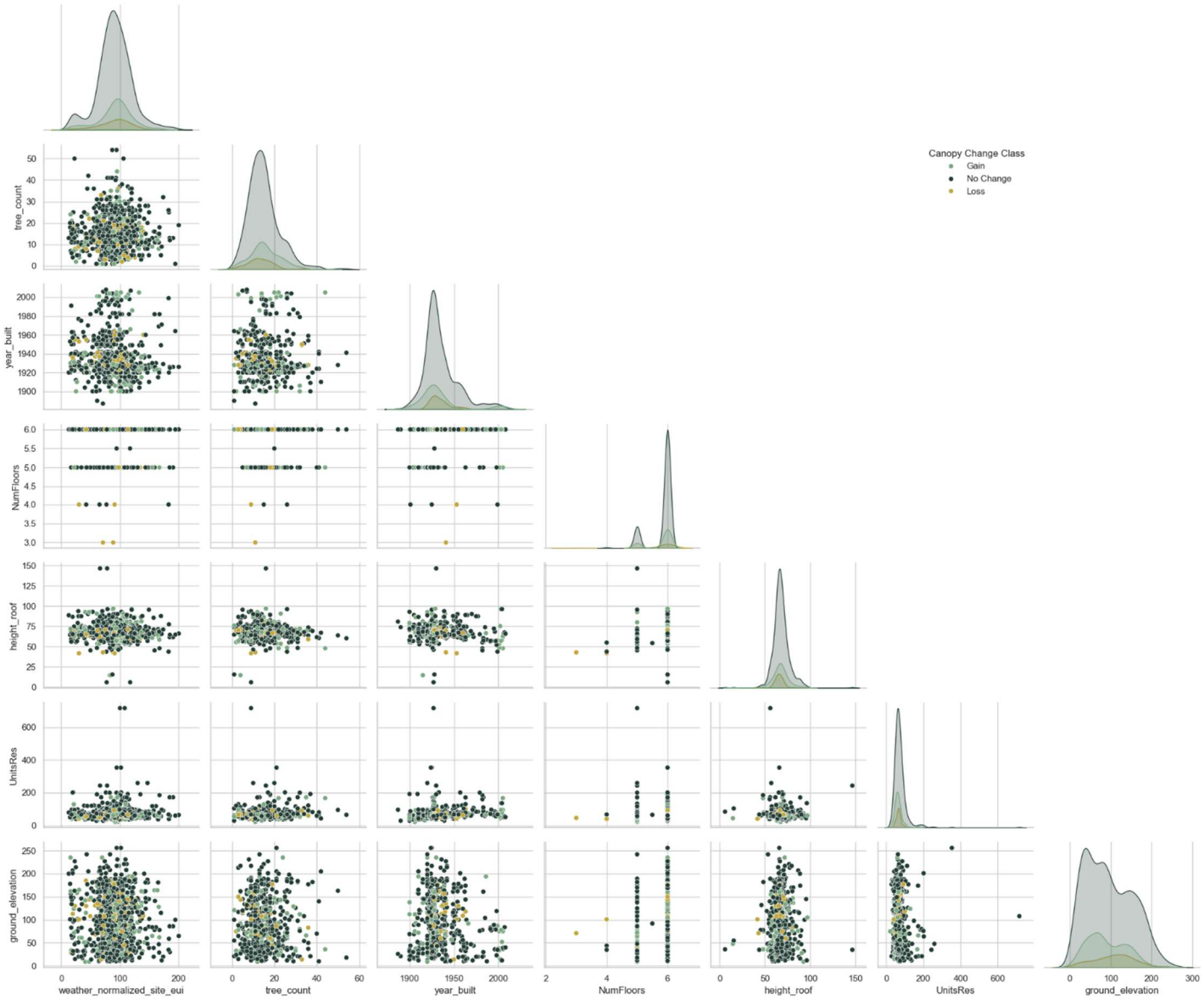
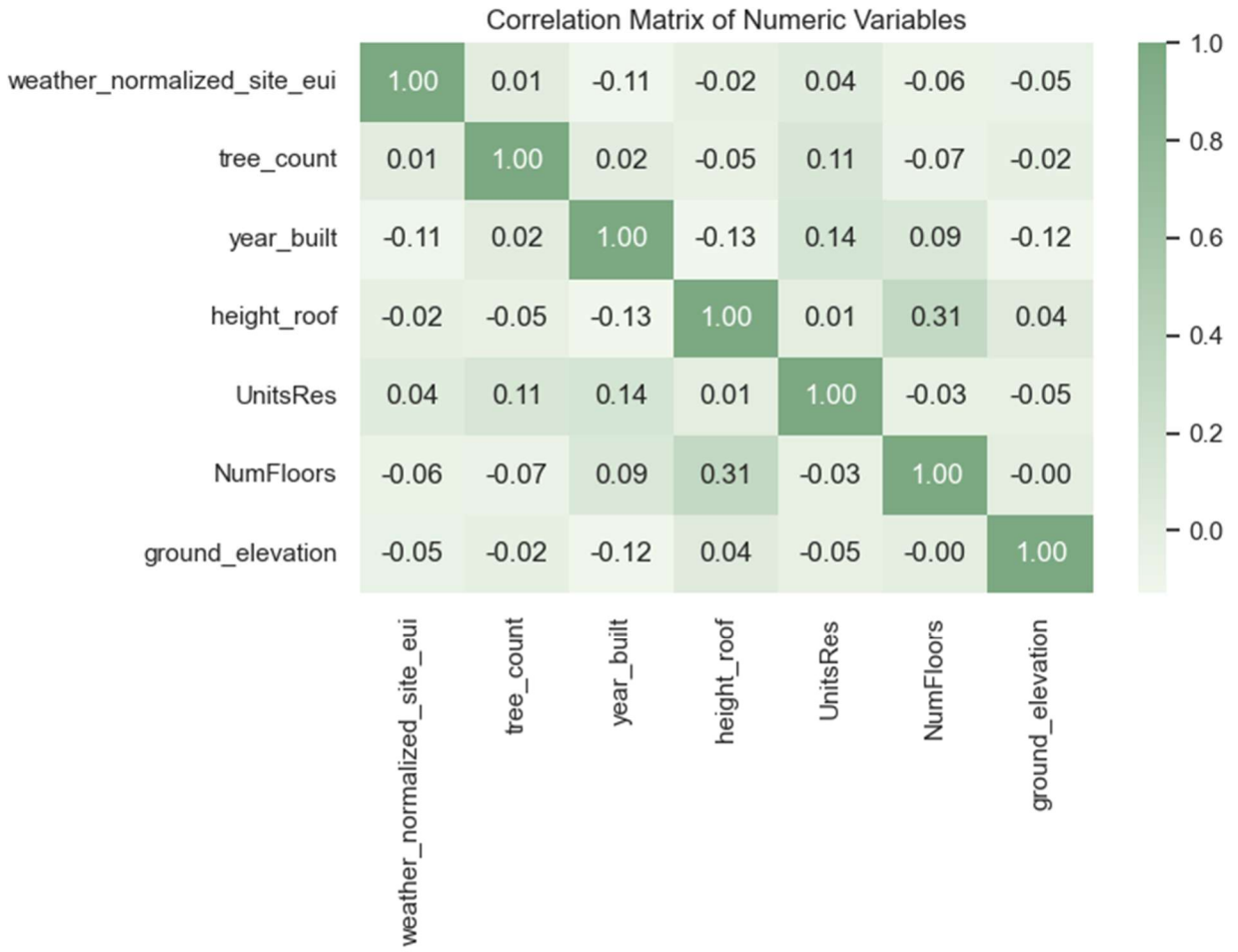


Figure 8 – Correlation Matrix Numeric Features



Beyond preliminary looks at inter-feature relationships, to further gain a proper sense of the data distributions for each of the variables in the data, they were plotted and examined. Most importantly, for the purposes of this analysis, the distributions of Weather Normalized Site EUI (Figure 9) and the tree count (Figure 12). The untransformed weather normalized EUI data has a normal distribution for the most part, there is a slight right skewing in this feature's distribution. Both a log transformation (Figure 10) and a Square Root transformation (Figure 11), don't make the distributions more normal. The distribution of the tree count for each building in the data is a bit less normal, it is more right skewed with some buildings having much higher tree counts than others. Log transformation (Figure 13) on the tree count data, as well as square root transformation (Figure 14) did actually increased the normalcy, at least visually.

Figure 9 – Distribution Histogram of Weather Normalized EUI

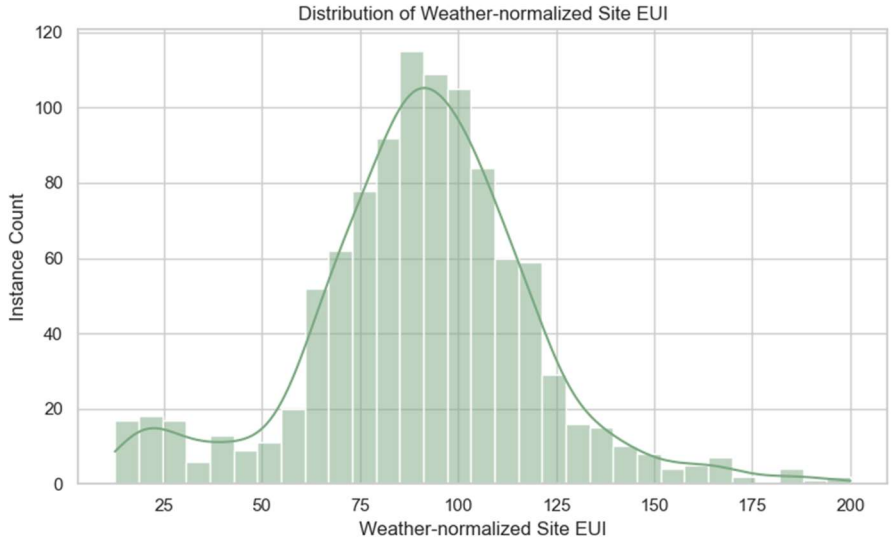


Figure 10 – Distribution Histogram of Log Transformed Weather Normalized EUI

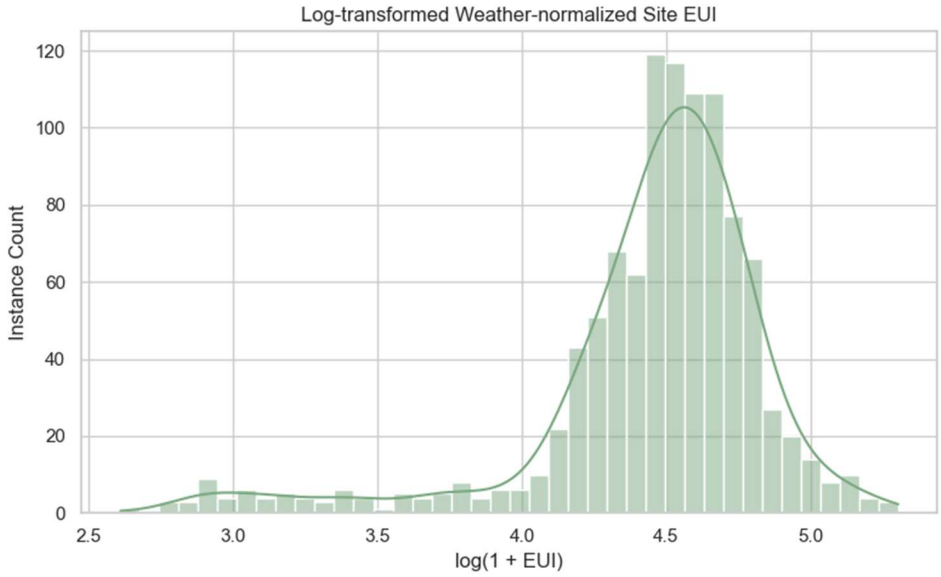


Figure 11 – Distribution Histogram of Square Root Transformed Weather Normalized EUI

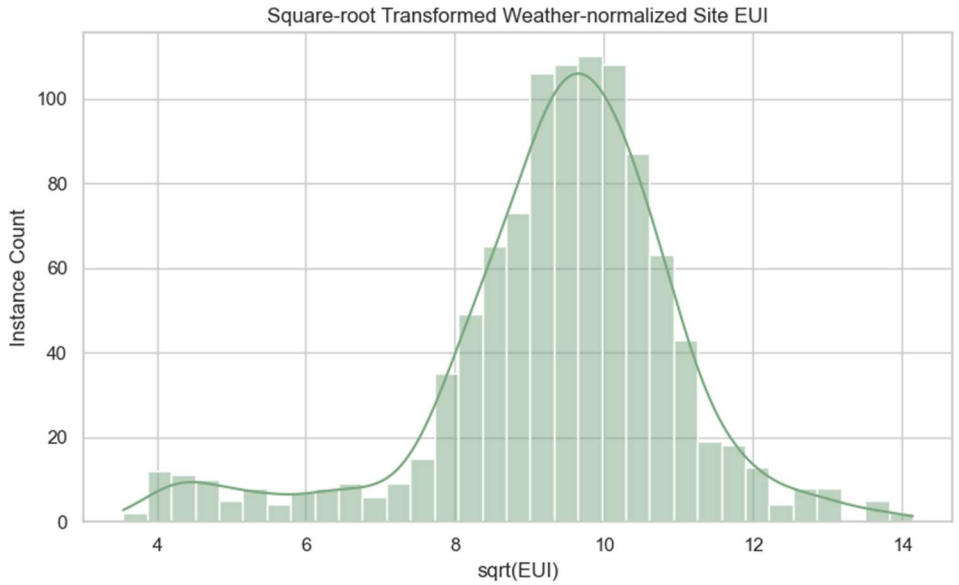


Figure 12 – Distribution Histogram of Tree Count

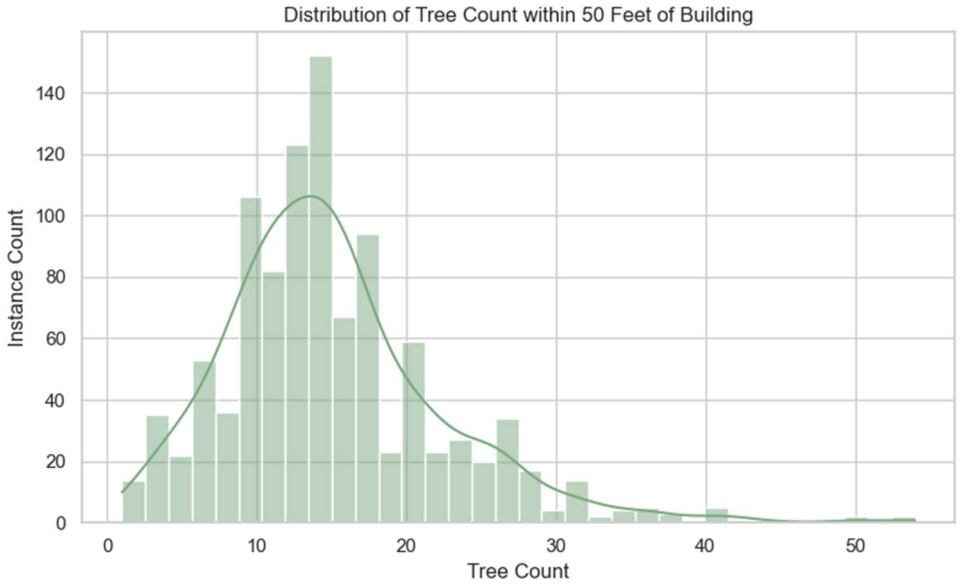


Figure 13 – Distribution Histogram of Log Transformed Tree Count

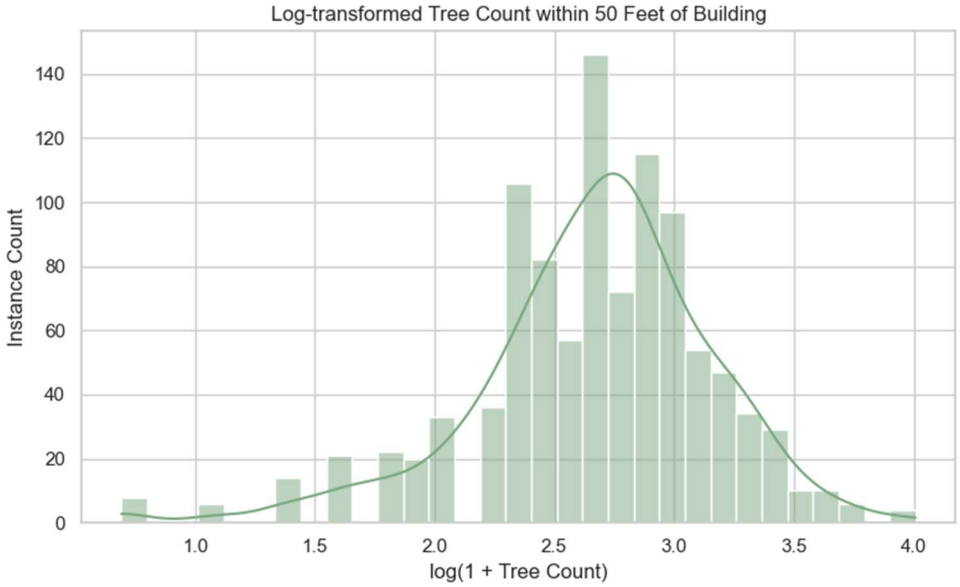
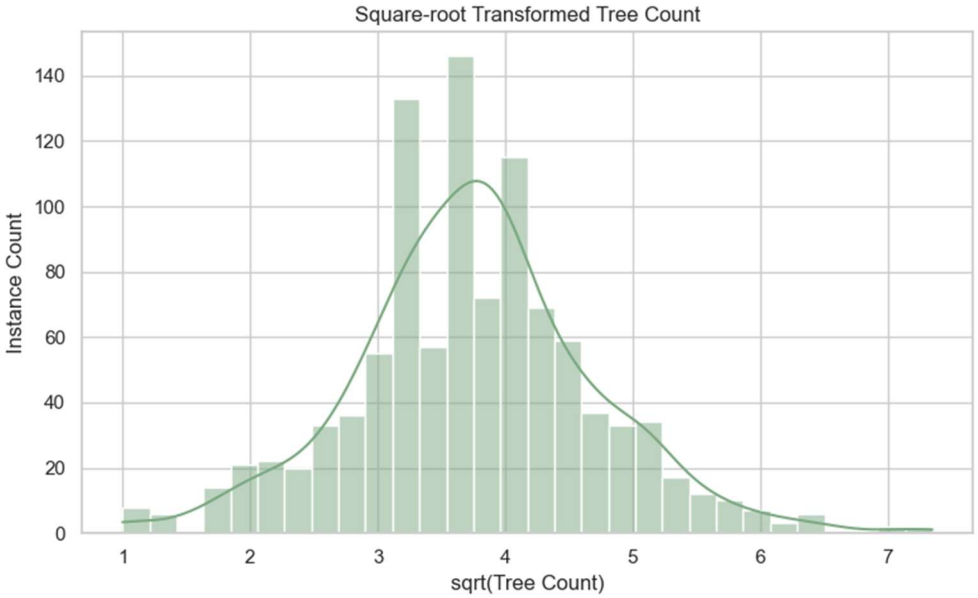


Figure 14 – Distribution Histogram of Square Root Transformed Tree Count



The foundational hypothesis for this project, based on other research papers, is that the greater the numbers of trees around a building, or as a proxy, the greater the canopy coverage around a building the less energy would be used by that building. To gain a sense of what the data shows, before any modeling, its worth looking at how these different features, the tree count, canopy category, and Weather Normalized EUI, relate to each other.

Figure 15 – Tree Counts Box Plot by Canopy Change Class

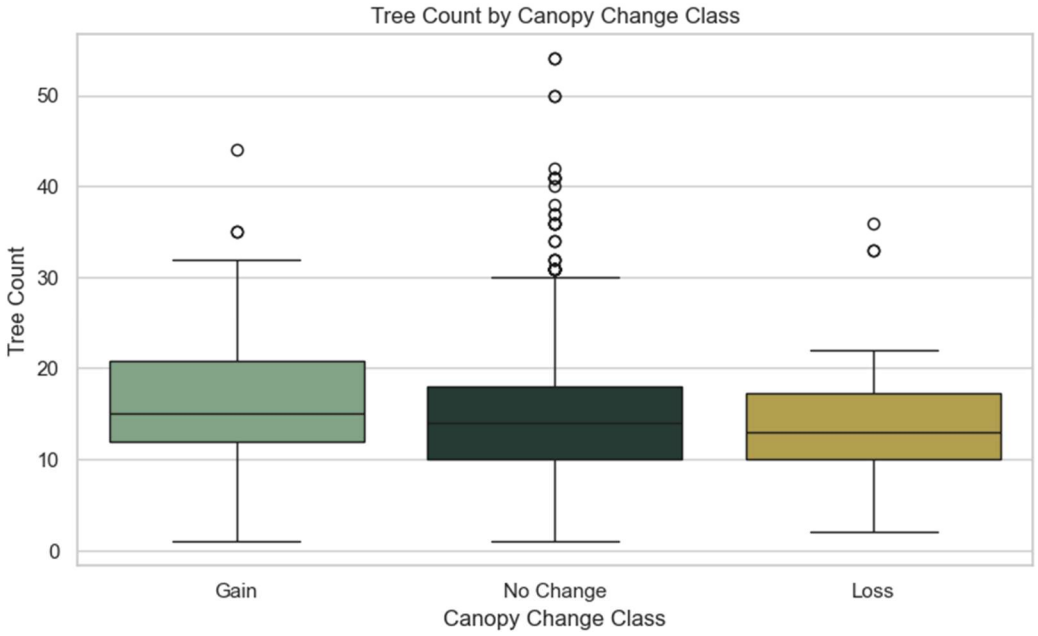
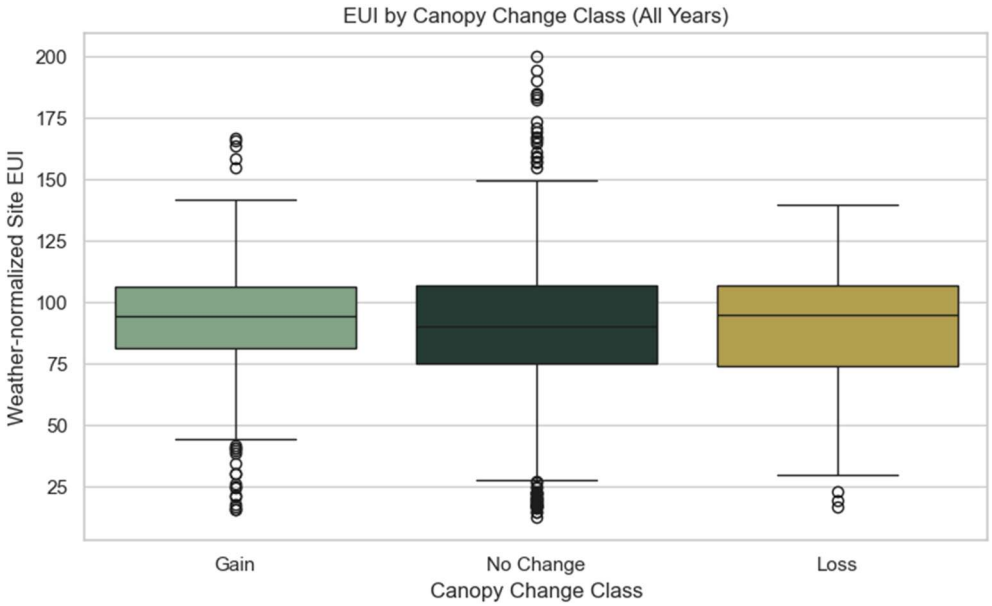


Figure 15 shows the range of tree counts for building in the data set via a box plot, which also categorizes the buildings and their respective tree counts by canopy change category. Looking at this plot, the gain category has a higher median of trees than the no change or loss categories. The loss category has the lowest median value. The no change category has the highest variance in the values with outliers on the higher end of the count,

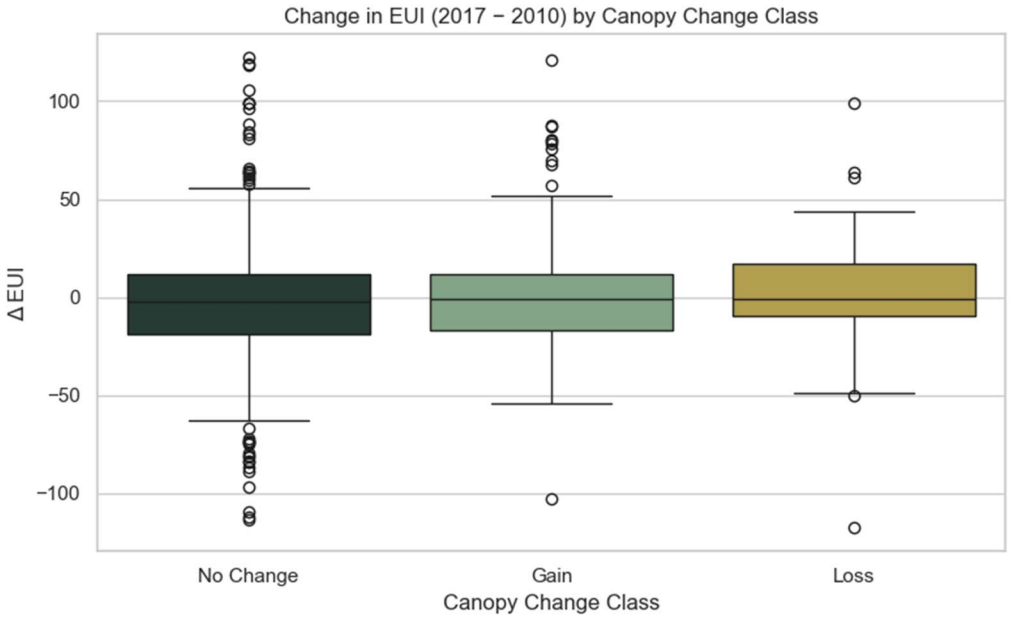
Similarly, Figure 16, which also uses a box plot, highlights the differences in the buildings' weather normalized EUI across each canopy category. The figure shows no overt trend across the categories.

Figure 16 – Weather Normalized EUI Box Plot by Canopy Change Class



A third cut of the data, in Figure 17, using the change in weather normalized EUI for each building between 2017 and 2010, was also placed into the same type of box plot. Even though the temporal dimension was somewhat controlled for, there is still no strong pattern visible across these categories.

Figure 17 – Change in Weather Normalized EUI Box Plot by Canopy Change Class



Modeling the Data – First Technique (Difference in Differences)

For this project there are two main modeling techniques used. The first is using the weather normalized EUI data and the canopy change class feature. Specifically, to look at the change in weather normalized EUI from 2010 through 2017 for these different canopy change categories, identifying any potential impact on energy usage.

Unfortunately, several attempts, using different transformations, and also with and without additional features, were unsuccessful in outlining any impact canopy changes had on the weather normalized EUI numbers. Across all models the r-squares values were poor and the fits were weak. None of the tree features' coefficients were statistically significant, most of the building-focused features were also not significant. However, the one consistently significant covariate was the 1950–1980 year-built bracket. This suggests that the age of buildings may influence the weather normalized eui. However, none of the other construction year categories had statistically relevant numbers.

Overall, these models indicate that, within this dataset and time window, canopy change does not show any noticeable effect on building energy use intensity. The output from each of the attempts executed during the first Difference-In-Differences focused modeling session can be seen in the figures below (Figures 18 through 26), with a summary of all DID results in Table 6.

Figure 18 – Difference in Differences Model 1: Non-Transformed Weather Normalized EUI & Canopy Change while Controlling for Building Features

OLS Regression Results

Dep. Variable:	weather_normalized_site_eui	R-squared:	0.037
Model:	OLS	Adj. R-squared:	0.026
Method:	Least Squares	F-statistic:	2.439
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.00431
Time:	13:16:54	Log-Likelihood:	-4902.4
No. Observations:	1030	AIC:	9831.
Df Residuals:	1017	BIC:	9895.
Df Model:	12		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	112.3566	16.198	6.937	0.000	80.610	144.103
canopy_change_class[T.Gain]	-2.6297	3.784	-0.695	0.487	-10.045	4.786
canopy_change_class[T.Loss]	-3.0275	5.387	-0.562	0.574	-13.586	7.531
C(year_built_bracket)[T.1950-1980]	-14.3528	3.374	-4.254	0.000	-20.965	-7.740
C(year_built_bracket)[T.Post-1980]	-2.3228	3.772	-0.616	0.538	-9.717	5.071
C(commercial_floor_flag)[T.1]	3.7088	2.597	1.428	0.153	-1.381	8.799
post_2017	-1.5478	1.748	-0.885	0.376	-4.975	1.879
post_2017:canopy_change_class[T.Gain]	3.5994	4.156	0.866	0.386	-4.547	11.745
post_2017:canopy_change_class[T.Loss]	4.0271	6.172	0.652	0.514	-8.070	16.125
NumFloors	-3.2950	2.922	-1.128	0.259	-9.021	2.431
UnitsRes	0.0315	0.023	1.373	0.170	-0.013	0.076
ground_elevation	-0.0292	0.019	-1.534	0.125	-0.067	0.008
height_roof	0.0001	0.120	0.001	0.999	-0.234	0.235

Omnibus:	34.226	Durbin-Watson:	1.869
Prob(Omnibus):	0.000	Jarque-Bera (JB):	85.818
Skew:	-0.044	Prob(JB):	2.32e-19
Kurtosis:	4.411	Cond. No.	2.18e+03

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The condition number is large, 2.18e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 19 – Difference in Differences Model 2: Log Transformed Weather Normalized EUI & Canopy Change while Controlling for Building Features

OLS Regression Results

Dep. Variable:	log_eui	R-squared:	0.035
Model:	OLS	Adj. R-squared:	0.024
Method:	Least Squares	F-statistic:	2.090
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.0162
Time:	13:17:59	Log-Likelihood:	-519.69
No. Observations:	1030	AIC:	1065.
Df Residuals:	1017	BIC:	1130.
Df Model:	12		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	4.6818	0.210	22.326	0.000	4.271	5.093
canopy_change_class[T.Gain]	-0.0558	0.057	-0.979	0.328	-0.167	0.056
canopy_change_class[T.Loss]	-0.0461	0.081	-0.570	0.568	-0.204	0.112
C(year_built_bracket)[T.1950-1980]	-0.2022	0.054	-3.763	0.000	-0.308	-0.097
C(year_built_bracket)[T.Post-1980]	-0.0129	0.048	-0.270	0.787	-0.106	0.081
C(commercial_floor_flag)[T.1]	0.0447	0.036	1.242	0.214	-0.026	0.115
post_2017	0.0030	0.026	0.119	0.905	-0.047	0.053
post_2017:canopy_change_class[T.Gain]	0.0657	0.065	1.005	0.315	-0.062	0.194
post_2017:canopy_change_class[T.Loss]	0.0718	0.097	0.743	0.458	-0.118	0.261
NumFloors	-0.0338	0.037	-0.910	0.363	-0.107	0.039
UnitsRes	0.0003	0.000	1.220	0.222	-0.000	0.001
ground_elevation	-0.0005	0.000	-1.757	0.079	-0.001	5.62e-05
height_roof	0.0001	0.002	0.077	0.938	-0.003	0.003

Omnibus:	393.483	Durbin-Watson:	1.868
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1474.741
Skew:	-1.835	Prob(JB):	0.00
Kurtosis:	7.572	Cond. No.	2.18e+03

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The condition number is large, 2.18e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 20 – Difference in Differences Model 3: Square Root Transformed Weather Normalized EUI & Canopy Change while Controlling for Building Features

OLS Regression Results

Dep. Variable:	weather_normalized_site_eui	R-squared:	0.037
Model:	OLS	Adj. R-squared:	0.026
Method:	Least Squares	F-statistic:	2.439
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.00431
Time:	13:19:55	Log-Likelihood:	-4902.4
No. Observations:	1030	AIC:	9831.
Df Residuals:	1017	BIC:	9895.
Df Model:	12		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	112.3566	16.198	6.937	0.000	80.610	144.103
canopy_change_class[T.Gain]	-2.6297	3.784	-0.695	0.487	-10.045	4.786
canopy_change_class[T.Loss]	-3.0275	5.387	-0.562	0.574	-13.586	7.531
C(year_built_bracket)[T.1950-1980]	-14.3528	3.374	-4.254	0.000	-20.965	-7.740
C(year_built_bracket)[T.Post-1980]	-2.3228	3.772	-0.616	0.538	-9.717	5.071
C(commercial_floor_flag)[T.1]	3.7088	2.597	1.428	0.153	-1.381	8.799
post_2017	-1.5478	1.748	-0.885	0.376	-4.975	1.879
post_2017:canopy_change_class[T.Gain]	3.5994	4.156	0.866	0.386	-4.547	11.745
post_2017:canopy_change_class[T.Loss]	4.0271	6.172	0.652	0.514	-8.070	16.125
NumFloors	-3.2950	2.922	-1.128	0.259	-9.021	2.431
UnitsRes	0.0315	0.023	1.373	0.170	-0.013	0.076
ground_elevation	-0.0292	0.019	-1.534	0.125	-0.067	0.008
height_roof	0.0001	0.120	0.001	0.999	-0.234	0.235

Omnibus:	34.226	Durbin-Watson:	1.869
Prob(Omnibus):	0.000	Jarque-Bera (JB):	85.818
Skew:	-0.044	Prob(JB):	2.32e-19
Kurtosis:	4.411	Cond. No.	2.18e+03

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The condition number is large, 2.18e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 21 – Difference in Differences Model 4: Non-Transformed Weather Normalized EUI & Canopy Change while Controlling for Limited Building Features

OLS Regression Results

Dep. Variable:	weather_normalized_site_eui	R-squared:	0.032
Model:	OLS	Adj. R-squared:	0.024
Method:	Least Squares	F-statistic:	2.658
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.00509
Time:	13:22:37	Log-Likelihood:	-4905.2
No. Observations:	1030	AIC:	9830.
Df Residuals:	1020	BIC:	9880.
Df Model:	9		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	112.5411	16.142	6.972	0.000	80.904	144.179
canopy_change_class[T.Gain]	-2.8082	3.764	-0.746	0.456	-10.185	4.569
canopy_change_class[T.Loss]	-3.5127	5.394	-0.651	0.515	-14.085	7.060
C(year_built_bracket)[T.1950-1980]	-13.9276	3.355	-4.151	0.000	-20.504	-7.351
C(year_built_bracket)[T.Post-1980]	-0.6507	3.662	-0.178	0.859	-7.827	6.526
C(commercial_floor_flag)[T.1]	3.5770	2.543	1.406	0.160	-1.408	8.562
post_2017	-1.5674	1.745	-0.898	0.369	-4.988	1.853
post_2017:canopy_change_class[T.Gain]	3.6191	4.150	0.872	0.383	-4.514	11.752
post_2017:canopy_change_class[T.Loss]	4.0349	6.162	0.655	0.513	-8.043	16.112
NumFloors	-3.4181	2.737	-1.249	0.212	-8.783	1.946

Omnibus:	35.210	Durbin-Watson:	1.875
Prob(Omnibus):	0.000	Jarque-Bera (JB):	89.772
Skew:	-0.045	Prob(JB):	3.21e-20
Kurtosis:	4.443	Cond. No.	86.5

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)

Figure 22 – Difference in Differences Model 5: Log Transformed Weather Normalized EUI & Canopy Change while Controlling for Limited Building Features

OLS Regression Results

Dep. Variable:	log_eui	R-squared:	0.030
Model:	OLS	Adj. R-squared:	0.022
Method:	Least Squares	F-statistic:	2.122
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.0262
Time:	13:23:38	Log-Likelihood:	-522.50
No. Observations:	1030	AIC:	1065.
Df Residuals:	1020	BIC:	1114.
Df Model:	9		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	4.6679	0.203	22.981	0.000	4.270	5.066
canopy_change_class[T.Gain]	-0.0567	0.057	-1.002	0.316	-0.167	0.054
canopy_change_class[T.Loss]	-0.0528	0.081	-0.655	0.512	-0.211	0.105
C(year_built_bracket)[T.1950-1980]	-0.1972	0.053	-3.686	0.000	-0.302	-0.092
C(year_built_bracket)[T.Post-1980]	0.0096	0.046	0.208	0.835	-0.081	0.100
C(commercial_floor_flag)[T.1]	0.0414	0.035	1.194	0.232	-0.027	0.109
post_2017	0.0028	0.026	0.109	0.913	-0.047	0.053
post_2017:canopy_change_class[T.Gain]	0.0659	0.065	1.010	0.312	-0.062	0.194
post_2017:canopy_change_class[T.Loss]	0.0719	0.097	0.745	0.456	-0.117	0.261
NumFloors	-0.0341	0.034	-0.991	0.322	-0.102	0.033

Omnibus:	398.775	Durbin-Watson:	1.871
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1529.444
Skew:	-1.852	Prob(JB):	0.00
Kurtosis:	7.682	Cond. No.	86.5

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)

Figure 23 – Difference in Differences Model 6: Square Root Transformed Weather Normalized EUI & Canopy Change while Controlling for Limited Building Features

OLS Regression Results

Dep. Variable:	sqrt_eui	R-squared:	0.032
Model:	OLS	Adj. R-squared:	0.023
Method:	Least Squares	F-statistic:	2.449
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.00978
Time:	13:24:53	Log-Likelihood:	-1961.7
No. Observations:	1030	AIC:	3943.
Df Residuals:	1020	BIC:	3993.
Df Model:	9		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	10.4471	0.876	11.922	0.000	8.730	12.165
canopy_change_class[T.Gain]	-0.1964	0.223	-0.880	0.379	-0.634	0.241
canopy_change_class[T.Loss]	-0.2106	0.320	-0.657	0.511	-0.838	0.417
C(year_built_bracket)[T.1950-1980]	-0.8182	0.204	-4.002	0.000	-1.219	-0.417
C(year_built_bracket)[T.Post-1980]	-0.0032	0.197	-0.016	0.987	-0.390	0.384
C(commercial_floor_flag)[T.1]	0.1888	0.144	1.309	0.191	-0.094	0.471
post_2017	-0.0409	0.101	-0.403	0.687	-0.240	0.158
post_2017:canopy_change_class[T.Gain]	0.2427	0.252	0.964	0.335	-0.251	0.736
post_2017:canopy_change_class[T.Loss]	0.2699	0.375	0.719	0.472	-0.466	1.006
NumFloors	-0.1674	0.149	-1.126	0.260	-0.459	0.124

Omnibus:	167.112	Durbin-Watson:	1.869
Prob(Omnibus):	0.000	Jarque-Bera (JB):	346.296
Skew:	-0.935	Prob(JB):	6.35e-76
Kurtosis:	5.139	Cond. No.	86.5

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)

Figure 24 – Difference in Differences Model 7: Non-Transformed Weather Normalized EUI & Canopy Change while Controlling No Building Features

OLS Regression Results

Dep. Variable:	weather_normalized_site_eui	R-squared:	0.001
Model:	OLS	Adj. R-squared:	-0.003
Method:	Least Squares	F-statistic:	0.4113
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.841
Time:	13:26:40	Log-Likelihood:	-4921.2
No. Observations:	1030	AIC:	9854.
Df Residuals:	1024	BIC:	9884.
Df Model:	5		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	91.3560	1.563	58.458	0.000	88.293	94.419
canopy_change_class[T.Gain]	-0.5710	3.731	-0.153	0.878	-7.883	6.741
canopy_change_class[T.Loss]	-2.8337	5.453	-0.520	0.603	-13.522	7.854
post_2017	-1.7890	1.740	-1.028	0.304	-5.199	1.621
post_2017:canopy_change_class[T.Gain]	3.8406	4.139	0.928	0.353	-4.272	11.953
post_2017:canopy_change_class[T.Loss]	3.8695	6.135	0.631	0.528	-8.155	15.895

Omnibus:	30.448	Durbin-Watson:	1.871
Prob(Omnibus):	0.000	Jarque-Bera (JB):	70.117
Skew:	-0.065	Prob(JB):	5.95e-16
Kurtosis:	4.271	Cond. No.	10.7

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)

Figure 25 – Difference in Differences Model 8: Log Transformed Weather Normalized EUI & Canopy Change while Controlling No Building Features

OLS Regression Results

Dep. Variable:	log_eui	R-squared:	0.002
Model:	OLS	Adj. R-squared:	-0.003
Method:	Least Squares	F-statistic:	0.4514
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.812
Time:	13:27:47	Log-Likelihood:	-537.26
No. Observations:	1030	AIC:	1087.
Df Residuals:	1024	BIC:	1116.
Df Model:	5		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	4.4505	0.022	200.832	0.000	4.407	4.494
canopy_change_class[T.Gain]	-0.0258	0.056	-0.460	0.645	-0.135	0.084
canopy_change_class[T.Loss]	-0.0454	0.082	-0.551	0.582	-0.207	0.116
post_2017	-0.0002	0.025	-0.008	0.994	-0.050	0.050
post_2017:canopy_change_class[T.Gain]	0.0689	0.065	1.059	0.290	-0.059	0.196
post_2017:canopy_change_class[T.Loss]	0.0694	0.096	0.722	0.471	-0.119	0.258

Omnibus:	399.770	Durbin-Watson:	1.875
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1513.184
Skew:	-1.864	Prob(JB):	0.00
Kurtosis:	7.622	Cond. No.	10.7

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)

Figure 26 – Difference in Differences Model 9: Square Root Transformed Weather Normalized EUI & Canopy Change while Controlling No Building Features

OLS Regression Results

Dep. Variable:	sqrt_eui	R-squared:	0.001
Model:	OLS	Adj. R-squared:	-0.003
Method:	Least Squares	F-statistic:	0.3788
Date:	Sat, 06 Dec 2025	Prob (F-statistic):	0.863
Time:	13:28:40	Log-Likelihood:	-1977.7
No. Observations:	1030	AIC:	3967.
Df Residuals:	1024	BIC:	3997.
Df Model:	5		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	9.3953	0.090	104.755	0.000	9.220	9.571
canopy_change_class[T.Gain]	-0.0672	0.221	-0.304	0.761	-0.501	0.366
canopy_change_class[T.Loss]	-0.1755	0.326	-0.538	0.591	-0.815	0.464
post_2017	-0.0537	0.101	-0.531	0.596	-0.252	0.144
post_2017:canopy_change_class[T.Gain]	0.2555	0.251	1.017	0.309	-0.237	0.748
post_2017:canopy_change_class[T.Loss]	0.2599	0.374	0.695	0.487	-0.473	0.993

Omnibus:	166.807	Durbin-Watson:	1.868
Prob(Omnibus):	0.000	Jarque-Bera (JB):	332.981
Skew:	-0.948	Prob(JB):	4.94e-73
Kurtosis:	5.041	Cond. No.	10.7

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)

Table 6 – Difference In Differences (DID) OLS Modeling Results Overview

Model Number	Formula	R^2 Value / Adj. R^2 Value	AIC / BIC	F-Statistic / Prob. (F-Statistic)	Statistically Sig. Coefficients (p-value <0.05) & Notes
1	weather_normalized_site_eui ~ post_2017 * canopy_change_class + NumFloors + UnitsRes+ ground_elevation+ height_roof+ C(year_built_bracket) + C(commercial_floor_flag)	0.037 / 0.026	9831 / 9895	2.439 / 0.00431	C(year_built_bracket)[T.1950-1980]
2	log_eui ~ post_2017 * canopy_change_class + NumFloors + UnitsRes+ ground_elevation+ height_roof+ C(year_built_bracket) + C(commercial_floor_flag)	0.035 / 0.024	1065 / 1130	2.090 / 0.0162	C(year_built_bracket)[T.1950-1980]
3	sqrt_eui ~ ~ post_2017 * canopy_change_class + NumFloors + UnitsRes+ ground_elevation+ height_roof+ C(year_built_bracket) + C(commercial_floor_flag)	0.037 / 0.026	3944 / 4008	2.335 / 0.00645	C(year_built_bracket)[T.1950-1980]
4	weather_normalized_site_eui ~ post_2017 * canopy_change_class + NumFloors + C(year_built_bracket) + C(commercial_floor_flag)	0.032 / 0.024	9830 / 9880	2.658 / 0.00509	C(year_built_bracket)[T.1950-1980]
5	log_eui ~ post_2017 * canopy_change_class + NumFloors + C(year_built_bracket) + C(commercial_floor_flag)	0.030 / 0.022	1065 / 1114	2.122 / 0.262	C(year_built_bracket)[T.1950-1980]
6	sqrt_eui ~ post_2017 * canopy_change_class + NumFloors + C(year_built_bracket) + C(commercial_floor_flag)	0.032 / 0.023	3943 / 3993	2.449 / 0.00978	C(year_built_bracket)[T.1950-1980]
7	weather_normalized_site_eui ~ post_2017 * canopy_change_class	0.001 / -0.003	9854 / 9884	0.4113 / 0.841	N/A
8	log_eui ~ post_2017 * canopy_change_class	0.002 / -0.003	1087 / 1116	0.4514 / 0.812	N/A
9	sqrt_eui ~ post_2017 * canopy_change_class	0.001 / -0.003	3967 / 3997	0.3788 / 0.863	N/A

Modeling the Data – Second Technique (Predictive Ridge & Lasso Regressions)

While the results of the first series of models were lackluster, with no strong findings for canopy change correlations, the second modeling technique was carried out despite the first section's results, It attempta to predict changes in weather normalized EUI with the same building-focused features as well as those focused on tree presence. This section will make use a of bias-variance trade off methodologies like ridge and lasso regressions combined with cross validation in an attempt to identify predictors of weather normalized EUI in the dataset.

Using the finalized dataset, the data was restructured to have one row per BBL with 2010 and 2017 values for weather normalized EUI in each respective row. Delta values between 2010 and 2017 were created for tree_count and weather_normalized_eui columns. Categorical and numeric features were processed differently, with the categorical variables being encoded in specific breakout columns via the *OneHotEncoder*() sklearn tool, and the numeric data was scaled with z-score normalized via the *StandardScaler*() sklearn tool. The working data was split into 80-20 train-test split sets, and a simple Ordinary Least Squares (OLS) model was created as a baseline.

After this base model, two additional models, one being a ridge linear model and the other being a lasso linear model, were built. Both the latter two models used a cross-validation method of 5 different folds of the data. Furthermore, for the alpha values used in the lasso and ridge models, we generate an array of 100 logarithmically spaced values between 0.001 and 1,000 using np.logspace(-3, 3, 100). Each of these alpha values was evaluated to find the value that best balances for the bias–variance trade-off and thus is considered the best regularization strength for the model. Of these models the Ridge method with an alpha value of 572.236 yielded the best model. The Lasso model yielded nearly the same r-squared value, but had higher error values (MSE, RMSE, and MAE). While the Ridge model was the relative best, none of these models are actually good at predicting weather normalized EUI shifts with the features available, the ridge model is essentially the least worst. Detailed results for all of these models can be seen in Table 7 below.

Table 7 – Ridge & Lasso Linear Modeling Results Overview

Model Number	Method	R^2 Value	Found Alpha	RMSE/ MSE / MAE
1	OLS Model	-0.2282	N/A	1230.177 / 35.074 / 24.516
2	Ridge Model (cv = 5)	-0.0305	572.236	1032.155 / 32.127 / 22.089
3	Lasso Model (cv = 5)	-0.0341	1.072	1035.806 / 32.184 / 22.087

Results & Discussion

Findings

Across all model specifications, no statistically significant relationship was found between changes in tree canopy classification and changes in weather normalized site energy use intensity (EUI) for the residential buildings examined.

Difference-in-differences estimates produced a low model fit with no statistically significant interactions between canopy change effects and the features in the data. Similarly, predictive modeling approaches, including OLS, Ridge, and Lasso regression, failed to generate significant values with negative R^2 results on the testing set. These results suggest that, within this dataset and timeframe, changes in local tree canopy coverage do not appear to be a strong driver of building-level energy use outcomes for 4 – 6 story multifamily residential structures in New York City.

While good faith efforts were made at constructing models that detect and predict shifts in a building's weather normalized EUI. Again, neither the DID models or the predictive regression models (OLS, Ridge, and Lasso) yielded significant results. With the DID models, the overall fits were low and none of the features, with the exception of the 1950–1980 year-built bracket, were statistically significant. This suggests that changes in energy use were more explained by the year the building was built more than any tree-focused features. All the predictive regression models produced negative r-squares values with the 20% test set. This implies that truly predictive features are not present in this dataset or inter-feature relationships that are not linear.

Overall, these models don't imply that trees have no impact on weather normalized EUI, but just not any large enough to identify in this data.

Challenges Faces

The main challenges in conducting this analysis stemmed from the limitations of the available data. Local Law 84 has mandatory reporting standards for certain buildings, but most of the properties that fall under these requirements are relatively large structures. As building size increases, the impact of nearby trees on energy use is likely to diminish. Many of the studies referenced in the literature review focus on single-family homes surrounded by trees. One or two-story buildings are much more exposed to shading and wind-blocking effects of surrounding trees than the five or six-story buildings that dominate the Local Law 84 sample used in this analysis.

A second constraint was obtaining reliable and detailed tree information for each building. While attempts were made to derive tree counts within 50 feet of each building examined, important contextual information on the trees was missing. Data on tree size, trunk diameter, plant date, age and other useful information, which would impact how much wind is blocked or how much shade is provided, was not available. This resulted in the primary proxy used for tree presence was change in canopy coverage between 2010 and 2017, supplemented by the constructed tree counts. More granular tree attributes, perhaps with newer tree data in the 2025 Tree Count underway, could highlight the effects of specific tree characteristics on energy use. Many of the trees in the data used here, had blank planted dates, but had other timestamp information to imply that perhaps many of the forestry data was derived from the 2015 NYC Street Tree Census. This would mean the baseline count used in this analysis was actually derived from data generated five years later than 2010.

Finally, integrating all of these datasets was only possible because of the underlying building identifiers and geospatial contexts provided in the city-level data. However, the analysis remains constrained by the fact that many smaller buildings and single-family homes fall outside of the Local Law 84 reporting framework. If comparable energy use data and geolocation information for those smaller city structures were to become available, either from the city or from a third-party source, the analysis could be expanded to look at those structures the way other papers have.

Potential Future Work

For the purposes of this analysis, canopy data was used because it's a solid picture of the shift in tree coverage from 2010 through 2017. While this project made use of the data available, future work could entail leveraging additional datasets if they become available. For instance, every 10 years since 1995, New York City conducts a tree count, essentially a census of the trees within the city, their locations, and additional information on them. The next one is currently in progress and the assumption is the data will be available in 2026. When it becomes available, it could easily be integrated into this analysis. Specifically leveraging years of Local Law 84 data that were initially ingested and processed but not used in the analysis. This would allow for more potentially more granular data to be analyzed.

As outlined previously, Local Law 84 data for every year since inception, 2010 through 2024, was ingested and processed. Using the updated and detailed 2025 Tree census data along with 2015 Tree Census data, further analysis could be carried out for 2015 and 2025 Local Law 84 data. Having an additional set of years to look at, examine, and analyze,

particularly with potentially more granular tree information, may allow for further nuance to be identified in the relationship between tree presence and energy usage.

Furthermore, Local Law 84 has been amended over the years since its initial inception. If the law is amended in the future to have more residential buildings fall under the umbrella of energy usage reporting mandates, this would allow for a further examination of tree presence impact on other residential building types. In short, future analysis can leverage changes in available data to integrate any novel nuance.

References

[i] https://www.epa.gov/heatislands/benefits-trees-and-vegetation
[ii] https://www.nycgovparks.org/trees/milliontreesnyc
[iii] https://www.nycgovparks.org/trees/street-tree-planting/neighborhood-tree-planting-program
[iv] https://tree-map.nycgovparks.org/
[v] https://www.nycgovparks.org/reg/trees-count
[vi] https://tree-map.nycgovparks.org/tree-map/borough/3
[vii] https://tree-map.nycgovparks.org/tree-map/borough/3
[viii] https://www.itreetools.org/
[ix] https://www.itreetools.org/
[x] https://comptroller.nyc.gov/services/for-the-public/nyc-climate-dashboard/emissions/#:~:text=Waste-,Buildings,windows%20that%20cause%20higher%20emissions.
[xi] https://www.nyc.gov/site/buildings/codes/ll84-benchmarking-law.page
[xii] https://www.itreetools.org/about
[xiii] McPherson, E. G. (2007). Benefit-based tree valuation. *Arboriculture & Urban Forestry, 33*(1), 1-11. https://doi.org/10.48044/jauf.2007.001 USFS Research & Development+1
[xiv] McPherson, E. G., & Rowntree, R. A. (1993). Energy conservation potential of urban tree planting. Journal of Arboriculture, 19(6), 321–331. https://doi.org/10.48044/jauf.1993.051 USFS Research & Development
[xv] McPherson, E. G., & Simpson, J. R. (1999). *Carbon dioxide reduction through urban forestry: Guidelines for professional and volunteer tree planters* (Gen. Tech. Rep. PSW-GTR-171). Pacific Southwest Research Station, United States Department of Agriculture, Forest Service. Albany, CA. https://doi.org/10.2737/PSW-GTR-171 USFS Research & Development
[xvi] He, C., Zhou, L., Yao, Y., Ma, W., & Kinney, P. L. (2021). *Cooling effect of urban trees and its spatiotemporal characteristics: A comparative study*. *Building and Environment, 204*, 108103. https://doi.org/10.1016/j.buildenv.2021.108103
[xvii] Zhu, S., Li, Y., Wei, S., Wang, C., Zhang, X., Jin, X., Zhou, X., & Shi, X. (2022). The impact of urban vegetation morphology on urban building energy consumption during summer and winter seasons in Nanjing, China. *Landscape and Urban Planning, 228*, 104576. https://doi.org/10.1016/j.landurbplan.2022.104576 UCL Discovery+1
[xviii] Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering, 45*, 103406. https://doi.org/10.1016/j.jobe.2021.103406 Hertfordshire Research Archive
[xix] Tsoka, S., Leduc, T., & Rodler, A. (2021). Assessing the effects of urban street trees on building cooling energy needs: The role of foliage density and planting pattern. *Sustainable Cities and Society, 65*, 102633. https://doi.org/10.1016/j.scs.2020.102633
[xx] Ravazdezh, F., & Rivers, N. (2025). Quasi-experimental evidence that the urban tree canopy reduces residential energy consumption. *Energy and Buildings*
[xxi] Cunningham, S. (2021). 9 Difference-in-differences. *Causal Inference: The Mixtape*. https://mixtape.scunning.com/09-difference_in_differences

[xxii] Facure Alves, M. (2022). 13 - Difference-in-Differences. *Causal Inference for the Brave and True*. https://matheusfacure.github.io/python-causality-handbook/13-Difference-in-Differences.html
[xxiii] https://opendata.cityofnewyork.us/
[xxiv] LATEST: https://data.cityofnewyork.us/Environment/NYC-Building-Energy-and-Water-Data-Disclosure-for-/5zyy-y8am/about_data
[xxv] https://www.nyc.gov/site/buildings/codes/ll84-benchmarking-law.page
[xxvi] https://www.nyc.gov/site/buildings/codes/benchmarking.page
[xxvii] https://www.energystar.gov/buildings/benchmark
[xxviii] https://www.naco.org/articles/energy-star-keeps-it-cool-counties
[xxix] https://data.cityofnewyork.us/Environment/Tree-Canopy-Change-2010-2017-/by9k-vhck/about_data

[xxx] https://data.cityofnewyork.us/Environment/Forestry-Tree-Points/hn5i-inap

[xxxi] https://data.cityofnewyork.us/Environment/Forestry-Work-Orders/bdjm-n7q4/about_data
[xxxii] https://geosearch.planninglabs.nyc/
[xxxiii] *https://www.nyc.gov/content/planning/pages/resources/datasets/mappluto-pluto-change#mappluto*
[xxxiv] https://mapsplatform.google.com/lp/maps-apis/
[xxxv] https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/bmjq-373p/about_data
[xxxvi] https://github.com/CityOfNewYork/nyc-geo-metadata/blob/main/Metadata/Metadata_BuildingFootprints.md

[xxxvii] https://github.com/jhnboyy/CUNY_SPS_WORK/blob/main/FALL2025/DATA698/BuildingDataWork_Part1.ipynb
[xxxviii] https://github.com/jhnboyy/CUNY_SPS_WORK/blob/main/FALL2025/DATA698/TreeWork_Part2.ipynb
[xxxix] https://github.com/jhnboyy/CUNY_SPS_WORK/blob/main/FALL2025/DATA698/Analysis_Part3.ipynb
[xl] https://www.nyc.gov/assets/finance/jump/hlpbldgcode.html
[xli] https://www.nyc.gov/assets/finance/jump/hlpbldgcode.html
[xlii] McPherson, E. G., & Simpson, J. R. (1999). *Carbon dioxide reduction through urban forestry: Guidelines for professional and volunteer tree planters* (Gen. Tech. Rep. PSW-GTR-171). Pacific Southwest Research Station, United States Department of Agriculture, Forest Service. Albany, CA. https://doi.org/10.2737/PSW-GTR-171 USFS Research & Development
[xliii] https://www.nyc.gov/assets/finance/jump/hlpbldgcode.html