Ways towards A Net-Zero Society, Take NYC as an Example

Background and Introduction

The issues of environmental protection and energy consumption are becoming increasingly important today. One promising solution for creating sustainable cities is implementing a net-zero city, where a city collectively reduces its greenhouse gas emissions to zero and ceases all activities that emit greenhouse gases. However, achieving net-zero status is a complex task due to existing infrastructure, resistance to change, lack of funding, regulatory barriers, and technical difficulties.

Buildings and their construction are responsible for 36% of global energy use and 39% of energy-related CO2 emissions each year, making them a key focus in reducing greenhouse gas emissions [1]. To assist energy-intensive cities like New York City in achieving net-zero emissions goals, we aim to investigate the major influences and trends of property GHG emissions and building energy intensity. By doing so, we can provide recommendations to help cities decrease their energy consumption and become more environmentally friendly. This research can ultimately contribute to the development of sustainable cities that are capable of achieving net-zero status.

Research questions

Our project aims to provide insights and recommendations that can help energy-intensive cities like New York City reduce their energy consumption and move towards a net-zero emissions future. by examining key factors influencing property GHG emissions and building energy intensity. To achieve this, we will utilize the 2015 NYC building energy consumption data set and the 2015 American Community Survey (ACS) Census Data set, which provide valuable insights into the demographics and energy usage patterns of New York City.

---

[1] Budds, Diana. “How Do Buildings Contribute to Climate Change?” *Curbed*, 19 Sept. 2019, archive.curbed.com/2019/9/19/20874234/buildings-carbon-emissions-climate-change.

Firstly, We will focus on Total GHG Emissions as the primary indicator of environmental impact, which will be reported in metric tons of carbon dioxide equivalent (MtCO2e) for the reporting year. We aim to develop a model that examines the correlation between Total GHG Emissions and various variables, such as moving towards net-zero emissions

In addition, there is a strong link between EUI and energy efficiency, when energy efficiency increases, less energy is required to produce output, thus reducing EUI. better energy efficiency also reduces GHG emissions. Improving energy efficiency involves complex measures, such as upgrading mechanical equipment, adjusting HVAC schedules, and implementing other energy-saving strategies requiring detailed treatment [2]. However, the exploration of building energy efficiency requires significant time and resources, and due to practical constraints, we will not be delving into this topic in depth

To further refine our analysis and reduce the potential impact of building energy efficiency on the model, we will introduce a new dependent variable in the second part of our study: building energy usage intensity (EUI). We will explore the relationship between EUI and various building and demographic characteristics, including building use, years since construction, income, age, male rate, and more.

Literature review:

What we already know about the subject

Based on academic research, it has been found that there has been a decrease in energy use and carbon emissions from the largest buildings in New York City over the past 12 years. These metrics are subject to variations from year to year owing to several factors, such as the fuel mix for electricity generation and variable weather conditions. Nonetheless, they have been observed to be on a downward trend due to increased energy efficiency, enhanced regional steam production and delivery, and accelerated fuel switching. Moving forward, it has been proposed

[2] Salmon, Kiernan. "Factors That Affect Building Energy Use." *Facilities Management*, 3 May 2016, facilities.ucdavis.edu/blog/factors-affect-building-energy-use.

that the decarbonization of buildings can be achieved by replacing natural gas with cleaner electricity over the coming decades.

Furthermore, in New York City, greenhouse gas emissions are not evenly distributed across sectors. Large buildings are responsible for approximately 35% of total GHG emissions, and energy use intensities (EUI) also vary across different building sectors. For example, multifamily buildings tend to have higher EUIs compared to other building types, such as offices, hotels, and K-12 schools. Regarding building size, larger buildings tend to have higher EUIs across all building types. Additionally, building age also influences EUI, with older buildings having higher EUIs. This effect is particularly pronounced in multifamily and office buildings, while it is less apparent in hotel buildings. These findings suggest that improving energy efficiency in large and older buildings, especially multifamily and office buildings, is crucial for reducing GHG emissions in New York City.[3]In addition, commercial buildings, hospitals, and laboratories have higher EUIs than offices and K-12 schools. this may be due to energy use preference, equipment loads, etc[4].

In addition, several academic papers have proposed methods for predicting EUI, which can serve as a guide for selecting independent variables in our models based on the strength of their relationship with EUI. For instance, Kontokosta developed a prediction model incorporating geographic and locational variables, building size, building age, and building energy source. This approach can aid in investigating the major influences and trends of property GHG emissions and building energy intensity in energy-intensive cities like New York City[5].

---

[3] Urban Green Council. "Explore NYC's Building Data." *Urban Green Council*, www.urbangreencouncil.org/what-we-do/exploring-nyc-building-data/.

[4] U.S. Energy Information Administration. "Use of Energy in Commercial Buildings - U.S. Energy Information Administration (EIA)." *Eia.gov*, 2016, www.eia.gov/energyexplained/use-of-energy/commercial-buildings.php.

[5] Kontokosta, Constantine E. *Predicting Building Energy Efficiency Using New York City Benchmarking Data*. 1 Jan. 2012. Accessed 30 Apr. 2023.

Additionally, various demographic factors can also significantly impact GHG emissions and EUI. For instance, households with higher incomes tend to consume more energy due to their ability to afford larger homes, more appliances, and devices that require more energy. However, Professor Chen's journal articles analyzing energy use intensity (EUI) across income groups reveal that low-income households have higher EUIs than higher-income households [6].

What we can do for the subject

Because different cities have different demographic factors, these factors will have a large impact on GHG emissions and EUI[7]. Therefore, our study plans to conduct a quantitative analysis of the influence of building characteristics and demographic factors on GHG emissions and EUI using the 2015 New York census data. We aim to validate the trends found in previous academic research in the context of New York City.

It is worth noting that previous studies have mainly focused on the impact of individual variables or homogeneous variables on EUI and GHG emissions. Our study can provide a complementary perspective by incorporating a wider range of variables, such as building characteristics and demographic information, within a narrower geographic scope of New York City.

Additionally, we will test our findings based on the 2015 data using the 2020 New York census data to examine the generalizability of our results and to explore the underlying reasons for similarities and differences in trends across years, particularly with respect to the fuel transition process.

---

[6] Chen, Chien-fei, et al. "Exploring the Factors That Influence Energy Use Intensity across Low-, Middle-, and High-Income Households in the United States." *Energy Policy*, vol. 168, Sept. 2022, p. 113071, https://doi.org/10.1016/j.enpol.2022.113071.

[7] O'Neill, Brian C., et al. "Global Demographic Trends and Future Carbon Emissions." *Proceedings of the National Academy of Sciences*, vol. 107, no. 41, 12 Oct. 2010, pp. 17521–17526, www.pnas.org/content/pnas/107/41/17521.full.pdf, https://doi.org/10.1073/pnas.1004581107.

Data sources

Why we choose these two data sets

Our project is primarily based on two datasets: the Building Energy Consumption and Greenhouse Gas Emissions data and the American Community Survey Data (ACS) collected by the Census Bureau. We further describe these two datasets in detail as follows.

2016 NYC Benchmarking Energy and Water Data Disclosure (Data for Calendar Year 2015) Collected by New York City Department of Finance

This is the most important dataset we need for exploring the subject, which include building energy consumption data. After exploring multiple data sources, we decide to use the 2015 building energy consumption data obtained under NYC Local Law 84/133 Energy Benchmarking, which mandates owners and managers of buildings larger than 50,000 square feet (25,000 after 2016) to report their building's energy usage to the City of New York annually.

The dataset consists of over 50 columns containing information about each property, including property ID and name, geographical location, property use type, building age, energy consumption intensity of different energy types, overall energy use intensity, and greenhouse gas emissions.

American Community Survey Data (ACS) collected by the Census Bureau

To enhance the scope of our research, we incorporated demographic data, which plays a crucial role in the subject. We collected the data for this part from the Census Bureau website [8]. To avoid any discrepancies, we decided to use the 2015 American Community Survey (ACS) 5-Year Estimates data, which is the most comprehensive survey conducted every five years at the block group level.

We explored over 60 datasets covering various topics such as business and economy, education, employment, families and living arrangements, health, housing, income and poverty, populations and people, and race and ethnicity, and selected seven variables to augment our model: average

[8] United States Census Bureau: https://www.census.gov/

household income, the rate of higher education (Bachelor's degree or higher), average commute time, percentage of people living alone, percentage of people who are married, percentage of males, and average age.

To merge the demographic data with the building energy consumption data, we utilized the technique of spatial joining. After examining each column of the NYC 2015 benchmarking dataset, we found a noteworthy field named NYC Borough, Block, and Lot (BBL), which corresponds to NYC PLUTO tax lot information [9]. Using this information, we map each record in the benchmarking dataset to tax lot polygons in ArcGIS Pro, as demonstrated in fig1.

Similarly, we can also map the data collected by the Census Bureau to the block group polygons in NYC, as is shown in fig2.

We then proceeded to perform a spatial join of the demographic data in the red polygons into all the blue polygons inside of it using an ArcGis map with the above two overlaid layers of information, as fig3 shows. The resulting output was a new 2015 NYC benchmarking data table with several more columns from ACS datasets. However, it should be noted that due to the limited precision of ACS data collection, we were only able to obtain data at the block group level, which can contain multiple tax lots. This may result in lower data precision after joining the census data. Nevertheless, with over 500 block groups in our research region, the precision was still deemed acceptable.

Predicting Features and outcomes

The primary focus of our study revolves around the data on greenhouse gas emissions (GHG), which represents the dependent variable in our research model. Prior to examining the relationship between the different variables, we performed some preliminary processing work, including filtering out invalid data rows and obvious outliers and generating a new column called the greenhouse gas emission intensity (GHGI), which divides GHG by property gross floor areas to reduce the impact of building areas on GHG.

[9] BYTES of the BIG APPLE - DCP, PLUTO and MapPLUTO:
https://www.nyc.gov/site/planning/data-maps/open-data.page#pluto

Another crucial field in our dataset is the building energy use intensity (EUI), which is an essential indicator of the problem, especially when we lack data for energy consumption efficiency, which has a significant impact on GHG emission data.

Various types of EUI are available, such as the Weather Normalized Source EUI, which measures energy use intensity in kBtus per gross square foot (kBtu/ft2) at the source of energy generation. However, we opted for the Weather Normalized Site EUI (kBtu/ft2) as our main focus is the property site. This type of EUI measures energy use intensity at the property site for the reporting year, normalized for weather. We also chose this type of EUI as climate significantly affects energy use efficiency. In mild and cool climates, high performance is more achievable than in hot and humid climates, where natural cooling and ventilation may not be as effective. While New York City does not have significant climate differences, we must still consider and minimize the impact of weather. Moreover, we preferred to focus on variables that are easier to observe and predict, and weather is difficult to predict.

Besides these two columns, which represent the y-axis variable in our model, we also identified several fields that can be included in the x-axis variable set in the model, which can be broadly categorized into two groups. The first group comprises columns that may affect the EUI and GHG, such as electricity use intensity (EI), natural gas use intensity (NGI), and water use intensity (WI). The second group consists of more independent variables such as building types, gross floor areas, and building age.

Exploratory data analysis

In the beginning, we try to get a general idea about our data by carrying out exploratory data analysis.Above all, we select valid data for our analysis. We pick up observations with submission status as "In Compliance" and with no missing value, as well as select required columns. Later, we filter out obvious high outliers manually, which typically have an extremely large value.

Furthermore, we classify the buildings by type to get more clear patterns of data distribution. We fill out the data to which kinds of building types number is under 40, which contributes little to

EUI. After that, we have only 6 types of buildings, including hotels, Multifamily Housing, Non-Refrigerated Warehouses, Offices, Residence Hall/Dormitory, and Senior Care Communities.

In the first place, we try to get an understanding of EUI and building features, and the latter consists of energy intensity of different kinds of energies, building age and floor area. To investigate the correlation between the dependent and independent variables, we do computation and plot a figure which reveals that GHGI is highly correlated with EUI, while EUI is highly correlated with the first group of independent variables, especially NGI and EI. The full plot of the correlation matrix of EUI and building features is shown in fig 4.
Additionally, we developed linear models to predict EUI using single variables, and the R-squared values of the model summaries supported this finding. This implies that EI, NGI and WI cannot be included in the model that predicts the GHG or EUI of properties due to high correlation, since they are components making up EUI.
To enhance the visual presentation of our data analysis, we utilized scatter plots to display the correlation between individual variables and EUI, as well as the predictive curves of the single variable models. In order to improve clarity and comprehensibility, we chose to utilize log(EUI) as the y-axis and categorized the data entries by their primary property type. After excluding property types with fewer than 40 records, we retain a total of five property types, with multifamily housing accounting for over 90% of the dataset and having the most data records. This approach allows us to more accurately assess the correlation between variables and EUI for each property type and provides a valuable foundation for our subsequent analysis. The plots of all results are shown in fig 5-9.

variable	EUI	EI	NGI	WI	GFA	BA
R-squared	0.986	0.029	0.869	1.4e-6	0.0005	0.0001

After analyzing the scatter plots and R-squared values in the table above, several insights can be drawn from the dataset. Firstly, it is recommended to use the column weather-normalized site building energy use intensity (WN-EUI) from the raw benchmarking dataset as the y-axis

variable, as it has a high R-squared value of 0.98 in the model predicting greenhouse gas emission intensity (GHGI) using EUI. Moreover, WN-EUI can reduce the negative impact of weather on our data, which is a potential limitation of our dataset.

Secondly, while the R-squared value of the model predicting EUI using gross floor area (GFA) and building age (BA) is small, the scatter plots indicate a positive correlation between EUI and GFA, particularly. Thus, it is plausible to include GFA and BA as variables in our final model to predict WN-EUI using multiple variables.

In the second place, we go to get an understanding of EUI and demographic factors, also combined with selected building features. The demographic factors include household income and higher education rates. We generated a correlation matrix using the selected variables to gain a deeper understanding of the relationships among them. The matrix revealed several noteworthy trends that corroborate the accuracy of our data and indicate which variables should be included in the final model. The full plot of the correlation matrix of EUI and building features is shown in fig 10.

The correlation matrix that we produced using the selected variables offers valuable insights that not only validate the accuracy of the data but also guide the selection of variables to be included in the final model. The matrix reveals many interesting trends, such as a strong positive correlation between the rate of higher education and average household income and the negative correlation between the percentage of people living alone and the percentage of people getting married, which aligns with common sense and intuition. As a result, we only pick higher education rate and live alone rate as a part of the independent variables in our final model and exclude average household income and the percentage of people getting married.

In addition, we created scatter plots and predictive curves for each variable using log(EUI) as the y-axis. The plots of all results are shown in fig 11-17 and the corresponding R-squared values are summarized in the table below.

variable	income	edu	commute	marriage	alone	male	age
R-squared	0.07883	0.08657	0.06911	0.07579	0.06272	0.0636	0.06594

After analyzing models that predicted EUI using individual variables, we incorporated variables with high R-squared values, such as higher education rate, average income, and commuting time, to predict EUI. However, to prevent a high correlation between various demographic variables, we removed the higher education and marriage rates from the model. Additionally, to prevent over-complication, we limited our model to three variables from demographic data, building age, and gross floor area from the benchmarking dataset. Therefore, we chose to include the percentage of people with higher education, average commuting time, living alone rate, and average age in our model.

Analysis

Model fitting

1. Splitting training and testing dataset

   In preparation for the training process, it is necessary to divide the dataset into two sets - a training set and a testing set - for the purposes of cross-validation. In this case, we have used a split ratio of 20 percent, meaning that 80 percent of the data will be utilized for model fitting.

2. Linear model

   Our initial analysis involved fitting our data using a simple model - a multiple linear regression model, which assumes a linear relationship between the dependent variables (GHGI and weather-normalized EUI) and six independent variables: building floor area, building age, rate of higher education, commute time, residents' average age, and percentage of individuals living alone, as well as their interactions. The mathematical expression that represents this linear relationship is as follows.

$$GHGI \sim higherEduRate * commuteTime * area * buildingAge * avgAge$$

$$Weather\_Normalized\_EUI\_ \sim higherEduRate * commuteTime * area * buildingAge * avgAge$$

In our study, we used the mean squared error (MSE) as the performance evaluation metric. This metric measures the average squared differences between the predicted and actual values of the dependent variable. The formula employed to calculate the MSE for each model is presented below.

```
mse <- function(model)
        mean(model$residuals^2)
```

3. Random forest model

   Alongside the basic linear models, we also employed two random forest models to predict the dependent variables using our six independent variables. The mathematical relationship between these variables is expressed through a formula that maps the independent variables to the dependent ones. Building these models necessitated additional R packages, including randomForest, to enable the models to function correctly. Furthermore, a special input variable, "ntree," was required for the model fitting function, determining the number of trees to grow. To ensure that every input row was predicted multiple times, we set the number of trees to 1000.

4. Training result

   The table below shows the training results of our models.

	Linear Model	Random Forest
MSE of GHGI train_set	6.308059e-06	7.140738e-06
MSE of WN_EUI train_set	1658.548	1855.362

The table presented above indicates that the training results of the linear models are marginally superior to those of the random forest models. Specifically, the mean squared error (MSE) of the linear model used to predict WN_EUI is 12% lower than that of the

random forest model used to predict WN_EUI, while the MSE of the linear model used to predict GHGI is 13% lower than that of the random forest model used to predict GHGI.

Model validation

1. Cross-validation

   The table below shows the validation precision of our trained models.

	Linear Model	Random Forest
MSE of GHGI test_set	6.951637e-06	5.822843e-06
MSE of WN_EUI test_set	2148.056	2069.359

   Upon examining the performance of the models on the test dataset, we found that only the random forest model used to predict GHGI values exhibited an improvement in MSE by 18%. The remaining three models demonstrated worse performances, with the linear model used to predict WN_EUI values displaying the poorest results, with an increase in MSE of 29%.

   Overall, the validation results indicate that the models are reasonably effective in predicting GHGI and WN_EUI values using the original dataset entries.

2. Use 2020‘s data set to validate the model performance

   In order to further explore the performance of our models, we chose to utilize them to predict GHGI and WN_EUI values using both the NYC 2020 benchmarking dataset and the 2020 ACS 5-Year Estimate datasets. The same processes, such as pre-processing, spatial joining, and data predicting, were employed on the 2020 datasets as were used on the test dataset in the previous section. The resulting table is presented below.

	Linear Model	Random Forest

MSE of GHGI test_set	7.355317e-06	6.149348e-06
MSE of WN_EUI test_set	2329.013	2201.423

The data table demonstrates that all models produced higher MSE values even when compared with the test dataset. The linear model used to predict WN_EUI values had the worst performance, with a 40% increase in MSE value.

Despite these results, the validation using data collected in 2020 indicates that our models still perform reasonably well in predicting GHGI and WN_EUI values using entirely distinct datasets. This is quite encouraging and suggests that our project has been a small success.

Model tuning

As the linear models exhibited better performance in predicting GHGI and WN_EUI, we opted to use the WN_EUI linear models for fine-tuning. In particular, we examined the residual plot of the WN_EUI linear model (Fig 20), which displayed a normal distribution pattern. We applied a simple Cox-Box transformation by taking the logarithm of WN_EUI to determine if a better residual plot could be generated and if the R-squared value of the model could be increased. Although it is difficult to determine the extent of normal distribution solely from the new plot (Fig 21), the R-squared values in the table below indicate that the model's performance improved at least slightly.

R-squared of WN_EUI linear model	0.07295
R-squared of log(WN_EUI) linear model	0.07444

Interpretation and findings

Since the linear models showed better performance than random forest in predicting GHGI and WN_EUI, we choose the linear models and interpret its coefficients. The results of summary of the model for WN_EUI are shown in fig18.

The intercept coefficient (1.113e+03) represents the estimated value of the dependent variable WN_EUI when all independent variables are zero.

The coefficient for higher education rate (-5.728e+02) is negative, indicating that as higher education rate increases, WN_EUI decreases, and so do all other independent variables.

Floor area's coefficients have small standard errors and p-values, indicating that its effect is statistically most significant. Higher education rate has a p-value greater than 0.5, so our model cannot fully describe its relationship with weather-normalized EUI.

Limitations

We identify some limitations in our analysis work. One kind of limitation is due to the data source itself which cannot be solved during our analysis process, and the other is model limitation which we can improve in our future work.

Data sources limitation

First, the weather normalized EUI used in our data does not distinguish between renewable and non-renewable energy. However, in the process of fuel burning and conversion, non-renewable energy leads to a lot more greenhouse gas emissions than clean energy will increase, while using weather-normalized site EUI cannot intuitively reflect this trend.

Second, the sample size of our dataset is not large enough. Our dataset consists mainly of buildings in NYC and focuses only on 2015, as a result of which our model may not be universally applicable to cities from all over the nation or across the world.

Third, the dataset is neither fully classified nor detailed enough. For one thing, the dataset does not cover further classification of the type of all buildings, such as high-efficiency buildings and

normal ones. For another, when it comes to the demographic factors, only block group level census data is provided.
What's more, there is one significant element not covered: energy efficiency. Weather normalized EUI only illustrates how much energy is consumed in total, while energy efficiency can demonstrate the proportion of energy wasted, which is important to the purpose of our work. If we can get a thorough understanding of energy efficiency, it will be much easier to reduce greenhouse gas emissions by cutting down avoidable energy waste.
Lastly, the demographic data was collected at the block group level, inside which there will be multiple tax lots, meaning some buildings with different EUI data may have the same demographic data, which will significantly reduce the precision of the model.

Model limitations

There exists heteroskedasticity in the regression analysis. Heteroskedasticity arises due to underlying factors that are difficult or impossible to control or measure.
We carry out several methods to mitigate the effects of heteroskedasticity. For example, we transform original data of EUI to its log form to reduce the impact of heteroskedasticity. We also use mean squared errors to get more accurate estimates of the coefficients. Nevertheless, it is impossible to completely eliminate heteroscedasticity in regression analysis and there are possibly more suitable alternative methods we can explore in the future.

Conclusion

Our analysis has revealed a strong correlation between natural gas intensity and weather-normalized energy consumption intensity. This finding confirms that natural gas is the primary energy source for buildings. However, it is worth noting that natural gas consumption is highly sensitive to temperature fluctuations, which implies inequality in energy use. Specifically, higher-income groups have more energy at their disposal and, thus, tend to consume more energy.

This phenomenon could be attributed to several factors, including the size of the property, the age of the building, and the number of occupants. Higher-income households may be more likely to live in larger, older buildings with more amenities and larger living spaces. Additionally, these

households may have a higher number of occupants, leading to increased energy consumption. However, the effect of income on energy consumption cannot be ignored, as higher-income households may have access to more energy-efficient appliances and may be less price-sensitive to energy costs.

Our findings highlight the need for policymakers to consider income inequality and its impact on energy consumption patterns. Incentives for energy-efficient upgrades and policies that encourage more sustainable energy use can help reduce energy consumption and address the inequality in energy use among different income groups [10].

To supplement our analysis, we further examined the 2015 building data and found significant differences in the mean weather-normalized site EUI among different building types. Specifically, the senior care community type had the highest mean WN_EUI, followed by hotels, while multifamily housing had a higher EUI than other types as well. This information may be useful for policymakers and building owners to target specific building types for energy efficiency improvements. The complete results are shown in fig19.

Based on the results of our linear regression model, we found that building age and commute time are the two factors that have a significant influence on EUI. Furthermore, our model shows a positive correlation between these variables and EUI. It is reasonable to assume that older buildings may have lower energy efficiency due to outdated equipment and systems. Therefore, building owners and government entities should prioritize upgrading equipment and improving the building's energy efficiency to reduce energy consumption.

The commute time variable's positive correlation with EUI also makes sense as longer commutes can result in higher energy consumption by associated buildings. Thus, it is imperative to reduce commute time to mitigate building energy intensity. By taking steps to reduce commute time,

---

[10] Tong, Kangkang, et al. "Measuring Social Equity in Urban Energy Use and Interventions Using Fine-Scale Data." *Proceedings of the National Academy of Sciences*, vol. 118, no. 24, 7 June 2021, https://doi.org/10.1073/pnas.2023554118.

such as promoting the use of public transportation or encouraging telecommuting, we can work towards decreasing EUI and achieving more sustainable urban development.

Appendix : (5 page)

Link of our Github repository for this project:

https://github.com/xinyew/CMU_courses_19603_project

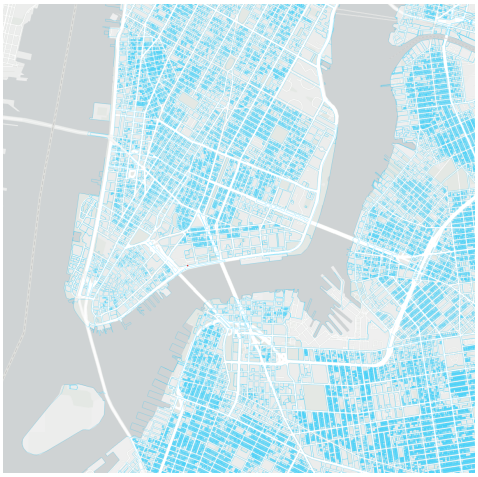

fig 1. Tax Lot from NYC PLUTO Dataset

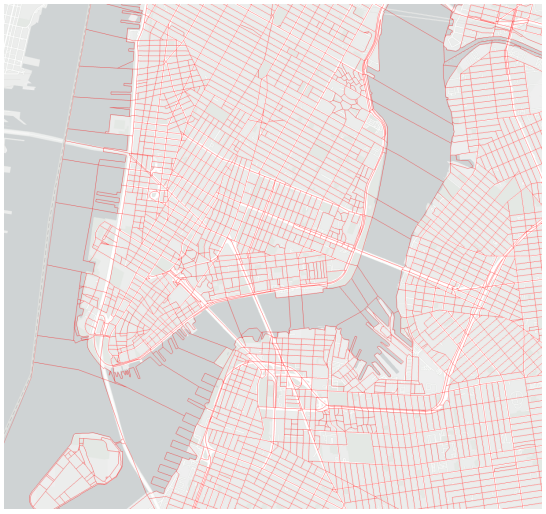

fig 2. Block Groups

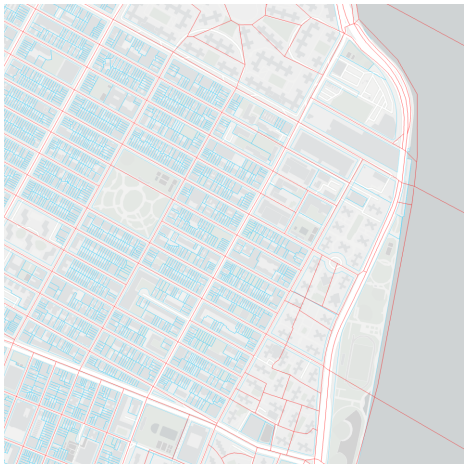

fig 3. Polygons Overlaid of Tax Lord and Block Groups

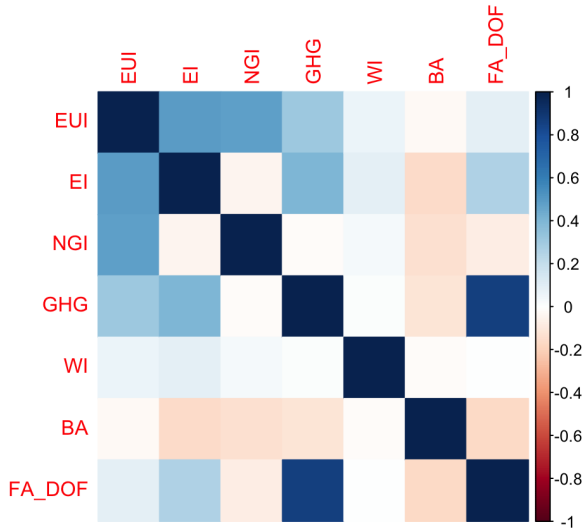

fig 4. Corrplot of Correlation Matrix of EUI and Building Features

fig 5. Linear Relationship of EUI and Energy Intensity, EUI vs log-transformed EUI

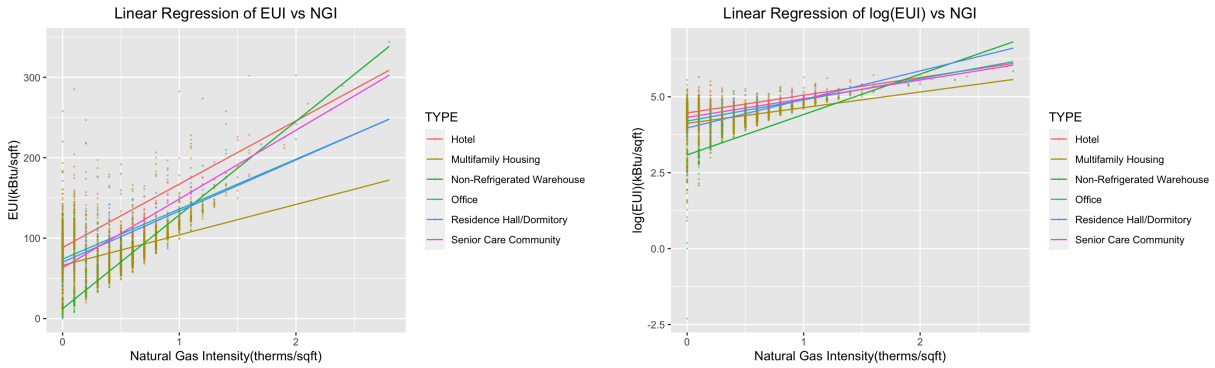

fig 6. Linear Relationship of EUI and Natural Gas Intensity, EUI vs log-transformed EUI

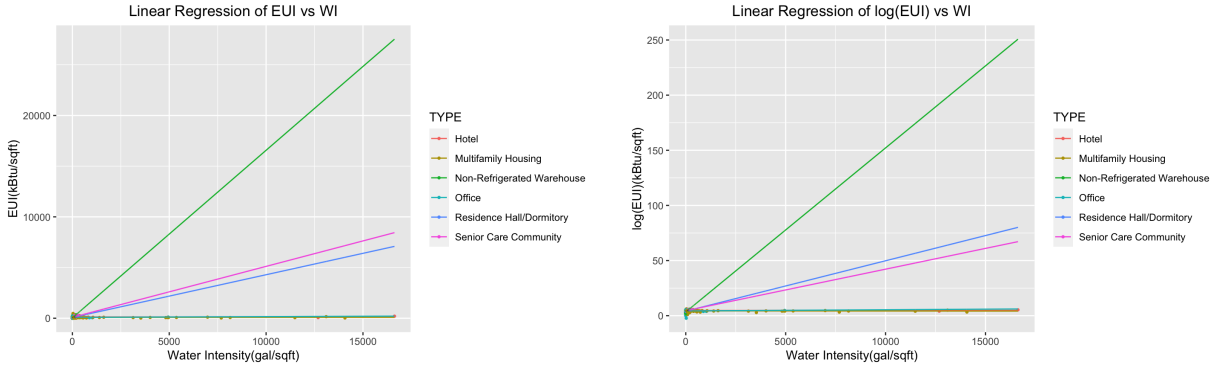

fig 7. Linear Relationship of EUI and Water Intensity, EUI vs log-transformed EUI

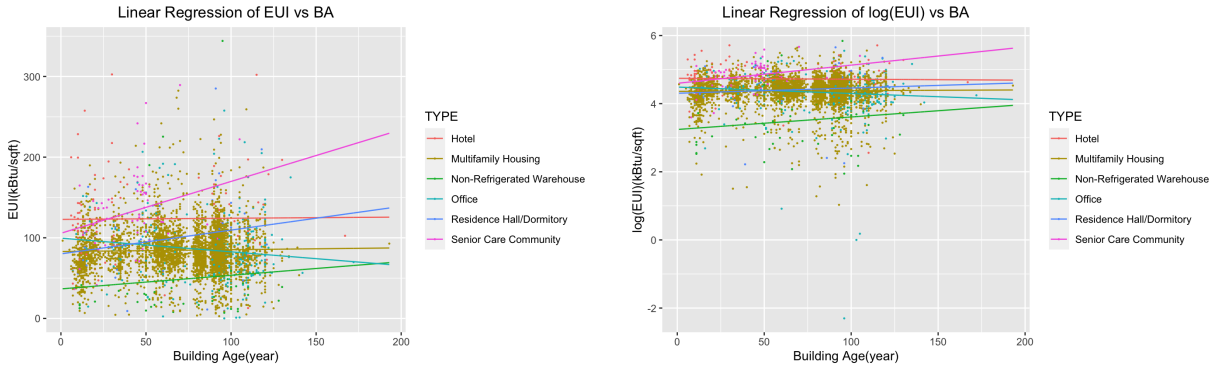

fig 8. Linear Relationship of EUI and Building Age, EUI vs log-transformed EUI

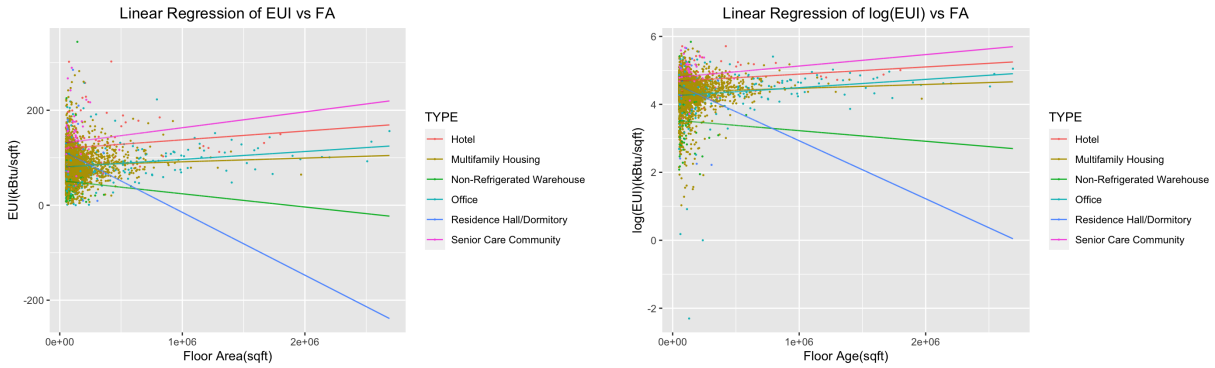

fig 9. Linear Relationship of EUI and Floor Area, EUI vs log-transformed EUI

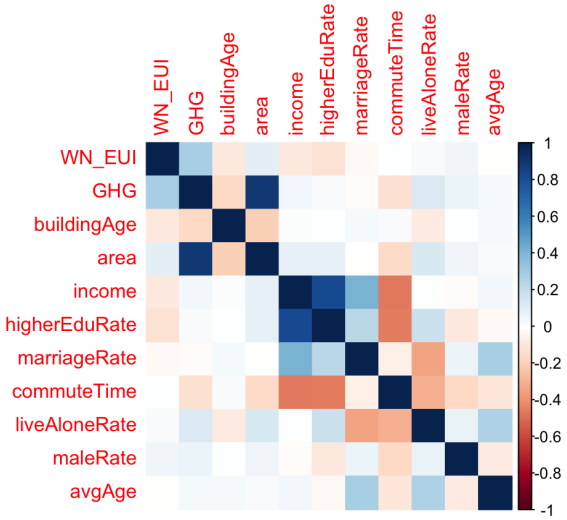

fig 10. Corrplot of Correlation Matrix of EUI and Demographic Factors, with Building Age and Floor Area

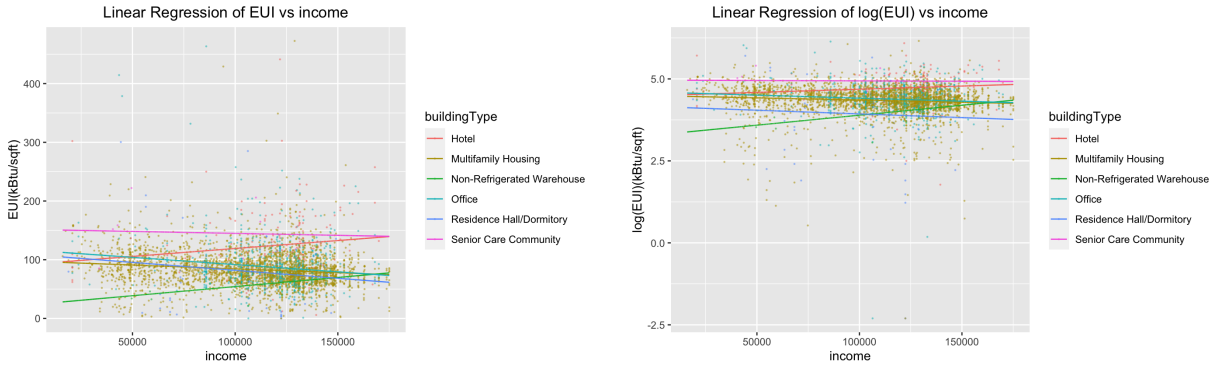

fig 11. Linear Relationship of EUI and Household Income, EUI vs log-transformed EUI

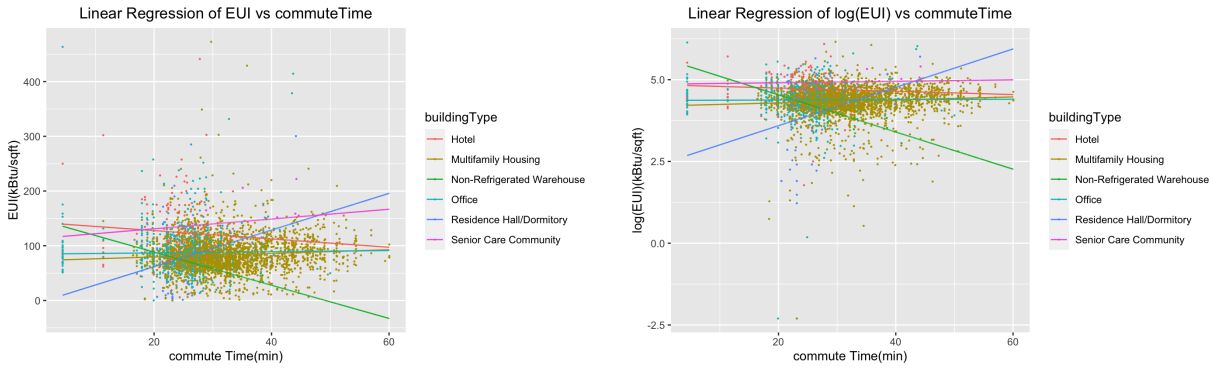

fig 12. Linear Relationship of EUI and Commute Time, EUI vs log-transformed EUI

fig 13. Linear Relationship of EUI and Marriage Rate, EUI vs log-transformed EUI

fig 14. Linear Relationship of EUI and Live Along Rate, EUI vs log-transformed EUI

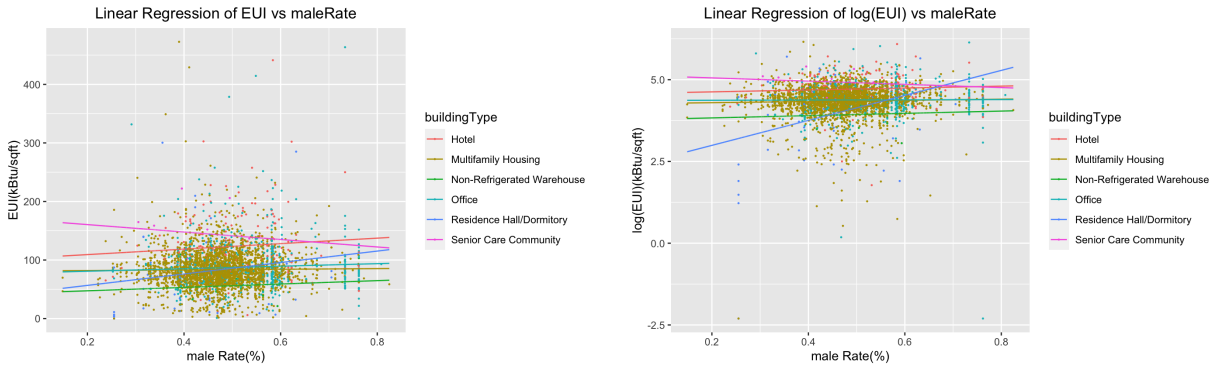

fig 15. Linear Relationship of EUI and Male Rate, EUI vs log-transformed EUI

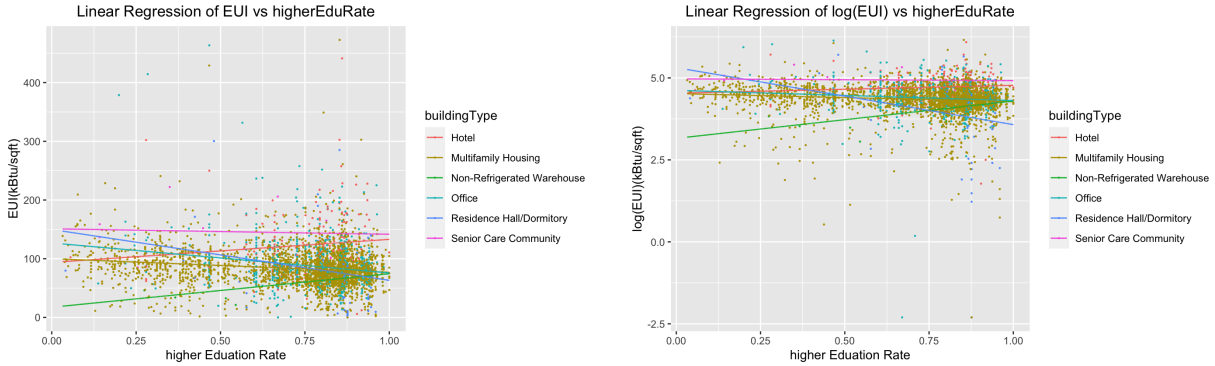

fig 16. Linear Relationship of EUI and Higher Education Rate, EUI vs log-transformed EUI

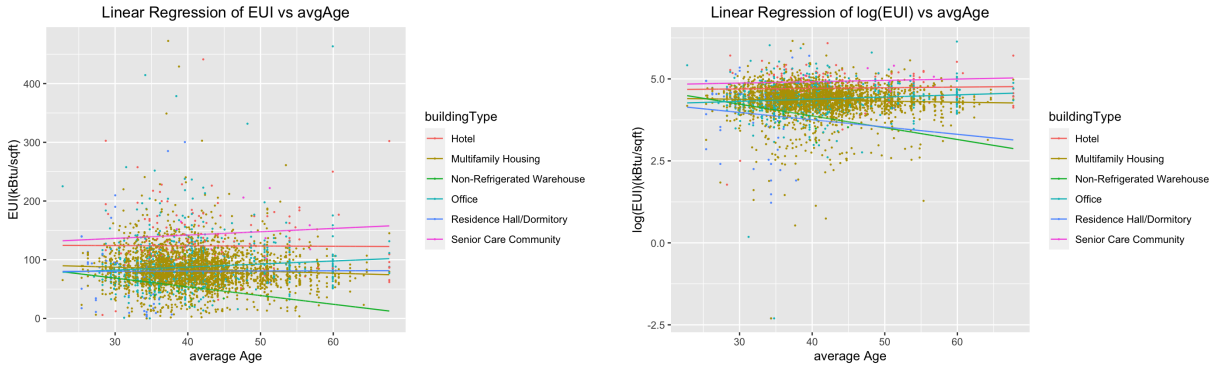

fig 17. Linear Relationship of EUI and Average Age, EUI vs log-transformed EUI

```
Call:
lm(formula = EUI ~ higherEduRate * commuteTime * area * buildingAge *
    avgAge * liveAloneRate, data = train_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-109.67  -17.70    3.14   20.67  357.88

Coefficients:
                                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                     1.113e+03  7.966e+02   1.397  0.16260
higherEduRate                                                  -5.728e+02  1.318e+03  -0.434  0.66399
commuteTime                                                    -2.765e+01  2.496e+01  -1.108  0.26812
area                                                           -6.622e-03  3.420e-03  -1.936  0.05297 .
buildingAge                                                    -2.104e+01  8.894e+00  -2.366  0.01809 *
avgAge                                                         -2.104e+01  1.796e+01  -1.172  0.24145
liveAloneRate                                                  -2.132e+03  1.484e+03  -1.437  0.15097
```

fig 18. Results of Coefficients of Linear Model for WN_EUI

A tibble: 6 × 2

buildingType <chr>	mean(WN_EUI) <dbl>
Hotel	111.54513
Multifamily Housing	72.65088
Non-Refrigerated Warehouse	43.51765
Office	76.01415
Residence Hall/Dormitory	73.08030
Senior Care Community	146.32000

6 rows

fig 19. Mean Weather-normalized Site EUI among Different Building Uses

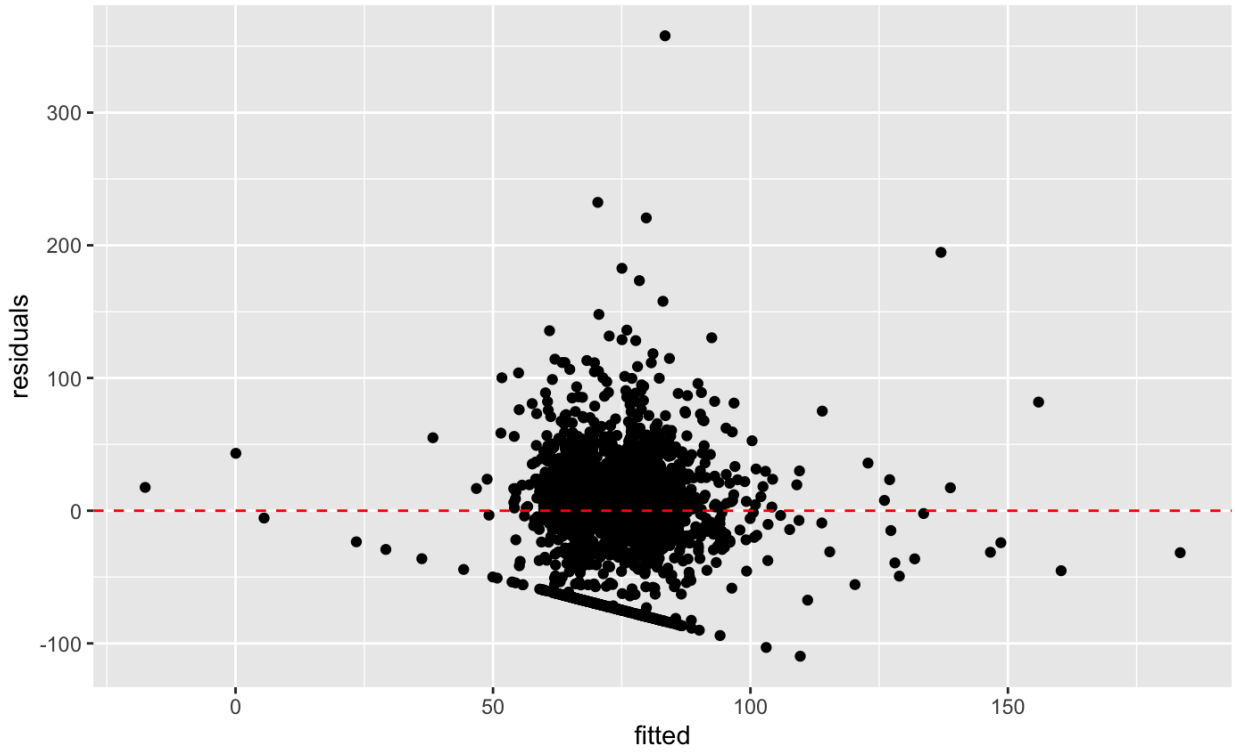

fig 20. Residual Plot of WN_EUI Linear Model

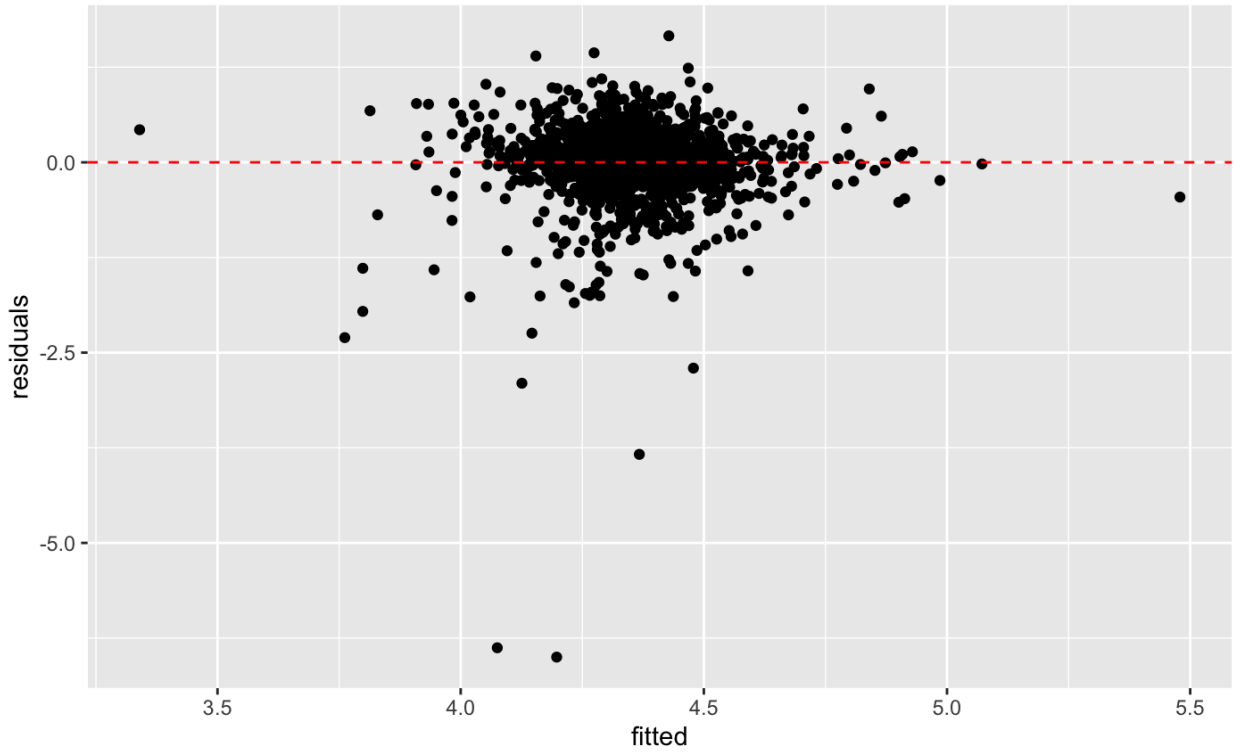

fig 21. Residual Plot of log(WN_EUI) Linear Model