Data 606: Final Report

Unveiling the Urban Heat Island Effect in Boston, Massachusetts

Aysha Afreen Althaf

Table of Contents

Introduction

1.1 Purpose & Motivation

As a metropolis grows, so does the heat. Boston, one of the largest metropolitan areas in the United States, is comprised of many streets, buildings, and parks, each contributing to the city's evolving climate narrative — an effect which is referred to as the Urban Heat Island (UHI) effect. Expansion of urban land coverage comes with a series of consequences which contribute to rising temperatures in urban landscapes.

The building density, land coverage, and lack of vegetation—particularly natural canopies which offer shade—are all characteristics of heavily populated urban environments. These heat-retaining properties lay the foundation for the manifestation of the UHI effect. Concurrently, heat emissions, pollution, and high levels of energy consumption within the city contribute to the rising temperatures, exacerbating the disparity in temperature between urban and rural areas (O'Malley et al, 2014).

To build upon previous works on this topic, we will focus on employing advanced statistical and machine learning techniques designed to reveal nonlinear relationships between land surface temperature and city features. Our investigation serves a dual purpose: to display the intricate relationship between urbanization and temperature rise, and to provide valuable insights for sustainable city development. By understanding the nuances of the UHI effect and identifying key contributors to temperature escalation, we allow for informed decision-making and targeted interventions aimed at mitigating the impact of urbanization on climate.

1.2 Dataset

The dataset employed in this project focuses on the Urban Heat Island (UHI) Effect in Boston and originates from the Harvard Dataverse. It is presented as a CSV file, containing 24,719 records distributed across 14 columns. As an open-source database, it is accessible to the public, facilitating broad usage for analysis and research.

The subsequent section provides a comprehensive overview of the data layers present within the raster stack file. Notably, each layer's corresponding data table variable mnemonic is indicated in parentheses following the layer's name. The construction of some layers involved standard geospatial data sources such as Landsat, while others were directly obtained from publicly accessible repositories like MassGIS. Detailed information about the specific sources for each layer is provided within their respective descriptions.

The variables of interest are given below:

Column Name	Description
Land surface temperature (lst)	Measures the land surface temperature in degrees Celsius (°C) per 30m raster cell. Values range from 6.8°C to 57.68°C.
Albedo (alb)	Measures surface reflectivity with darker surfaces reporting lower albedo scores. Values range from 0 - 1.
Tree canopy fraction (can)	Measures the percent of tree canopy coverage for 30m raster cells. Values range from 0 - 100.
Impervious surface fraction (isa)	Measures the percent of impervious surface coverage per 30m raster cell. Value range from 0 - 100.
Population density (pop)	Measures the number of persons per squared kilometer (using 1km raster cells). Ranges from 0 - 40,729.
Total Population (TotalPop)	Total population of each area (ranges from 11-8971).
Zoning	Indicates whether zoning area is residential or commercial.
ByAuto	The ratio of people in the specified zone who use cars as primary transportation
ByPubTrans	The ratio of people in the specified zone who use public transportation as primary transportation
ByBike	The ratio of people in the specified zone who use bikes as primary transportation
ByWalk	The ratio of people in the specified zone who walk as primary transportation
TotalHouseUnits	Number of houses the specified zone

VacantUnitPer	Number of vacant houses in the specified zone
MedYrBuilt	Median year of houses built in the specified zone

1.3 Methodology

The analysis begins with data cleaning and preparation in which null values are removed for all columns. We also generated dummy variables for categorical variables such as zoning.

Linear regression is used initially to explore these factors' influence on the UHI effect, using significant predictors to understand relationships. Model performance is evaluated using metrics like RMSE and R-squared. However, the homoskedasticity and normality assumptions cannot be met and therefore, alternative models will need to be explored.

To visualize the magnitude of the effects of the predictors on land surface temperature, random forest and gradient boosting are employed. Both models will be compared using RMSE and adjusted R-squared to determine the better fit.

Following this, support vector regression (SVR) will be employed in place of linear regression to determine the direction of the relationships in our data. SVR does not follow the same assumptions as conventional linear regression and only assumes linearity (Smola & Scholkopf, 2004). Insights on urban planning and mitigation strategies will be derived from the findings, emphasizing the importance of addressing the UHI effect in urban development policies.

1.4 Exploratory Data Analysis

Population Density vs Modes of Transportation

Figure 1.1

The scatter plot illustrating the relationship between population density and the use of different transportation modes—automobiles, public transport, biking, and walking—reveals distinct patterns that align with urban planning principles. Notably, there's a visible trend where higher population densities correlate with increased reliance on public transportation and potentially reduced use of personal automobiles. This pattern suggests that in densely populated areas, likely urban centers, efficient public transportation systems are both necessary and more frequently utilized, possibly due to the convenience they offer amidst congested traffic conditions and shorter intra-city travel distances.

However, the scatter plot also indicates considerable variability in the use of bikes and walking across different population densities, hinting that these transportation choices might be influenced by a variety of factors beyond just population density, such as city infrastructure, cultural norms, and geography. The spread of data points suggests that individual and community preferences, along with the availability of safe and accessible infrastructure, play significant roles in determining whether people choose to walk or bike.

Land Surface Temperature Distribution by Zoning

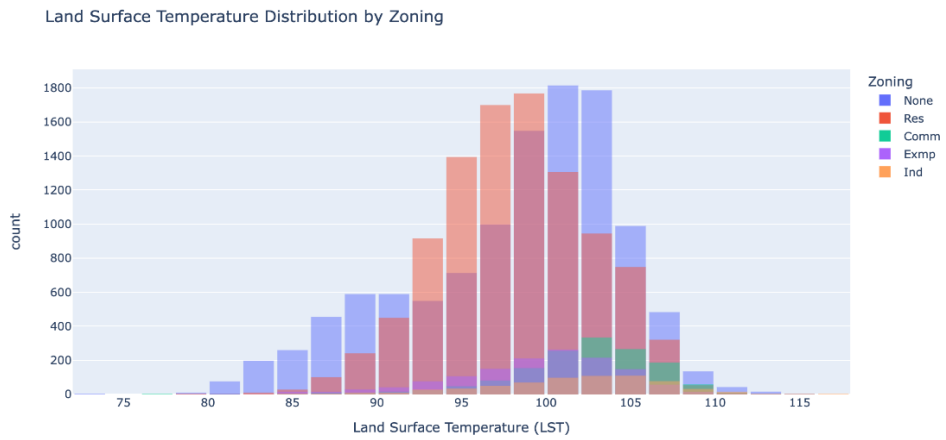


Figure 1.2

The above visual depicting the Land Surface Temperature (LST) Distribution by Zoning highlights the variability in surface temperatures across different zoning categories within the dataset. Each zoning type, represented by a distinct color, shows a range of LST values, indicating how land use and zoning can influence local environmental conditions. Urban zones, characterized by higher density and impervious surfaces, might exhibit higher LST values due to the urban heat island effect, whereas areas with more vegetation or open spaces could show lower temperatures. The overlap between different zoning types in the visual suggests some commonality in LST values across diverse land uses, but the spread within each category underscores the influence of specific zoning characteristics on local temperatures. This visualization serves as a valuable tool for understanding the environmental implications of land use planning and the need for sustainable urban development practices that consider the thermal impacts on the environment.

Vacant Housing Units Percentage by Zoning

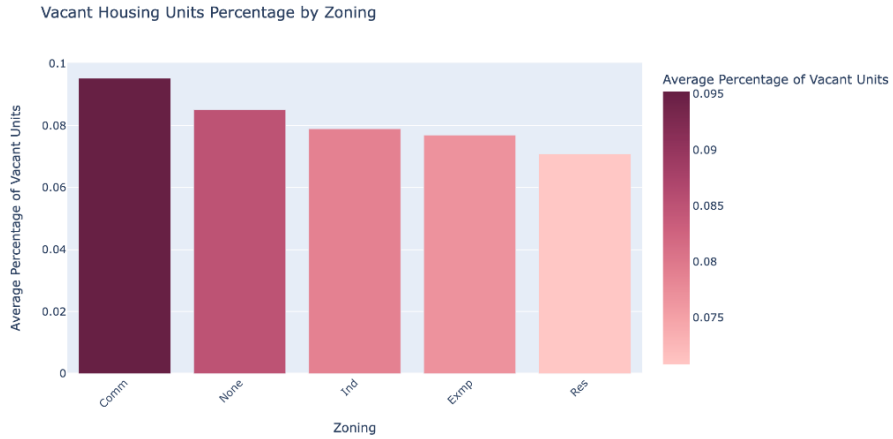


Figure 1.3

The bar chart showcasing the percentage of vacant housing units across different zoning categories offers insights into the occupancy patterns and potentially the development stages of various urban

and suburban areas. Higher percentages of vacant units in certain zoning categories might indicate regions undergoing transition, either due to new developments awaiting occupancy or older areas facing decline. These vacancy rates can have implications for the urban heat island (UHI) effect, as areas with high vacancy rates might lack the maintenance and greenery associated with inhabited spaces, potentially leading to higher local temperatures. Conversely, well-inhabited zones with active land use and green infrastructure could mitigate UHI effects through shade and evapotranspiration.

Understanding these patterns is crucial for urban planning and sustainability efforts. For instance, strategic redevelopment of areas with high vacancy rates could incorporate green spaces and cooling infrastructure to combat UHI effects. Moreover, analyzing vacancy rates alongside zoning provides a nuanced perspective on how land use diversity and occupancy levels contribute to the thermal landscape of urban areas, informing targeted interventions to enhance urban resilience to heat and promote more sustainable and livable environments.

Distribution of Zoning Categories

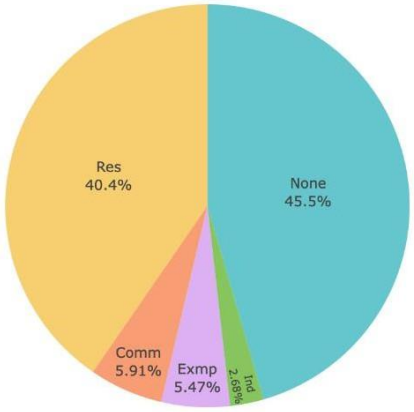


Figure 1.4

The pie chart illustrating Zoning Distribution provides a visual breakdown of how land is allocated across various zoning categories within the dataset, offering a snapshot of land use diversity. The proportions of different zoning types—residential, commercial, industrial, green spaces, etc.—can significantly influence the urban heat island (UHI) effect. Zones with high proportions of impervious surfaces, such as commercial and industrial areas, tend to absorb and retain more heat, exacerbating the UHI effect. Conversely, areas zoned for green spaces or agricultural use can help mitigate UHI through natural cooling processes like shading and evapotranspiration.

This visualization is relevant to understanding and addressing UHI because it highlights the potential areas of concern and opportunities for mitigation based on land use patterns.

Environmental Factors by Zoning

The series of box plots for Environmental Factors by Zoning, covering land surface temperature (LST), canopy cover (CAN), albedo (ALB), and impervious surface area (ISA), each tells a story about how different land uses impact and interact with the urban environment:

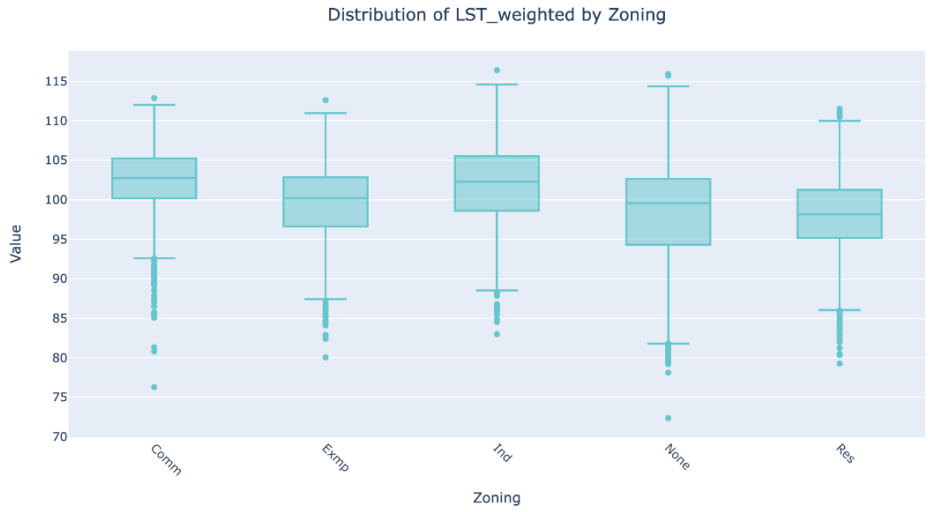


Figure 1.5

The LST box plot likely shows higher median temperatures for zones with dense construction and less vegetation, such as industrial or commercial areas, illustrating the urban heat island effect. Residential zones, especially those with more greenery, might display lower temperatures. The range and outliers in each category indicate microclimate variations within the same zoning type.

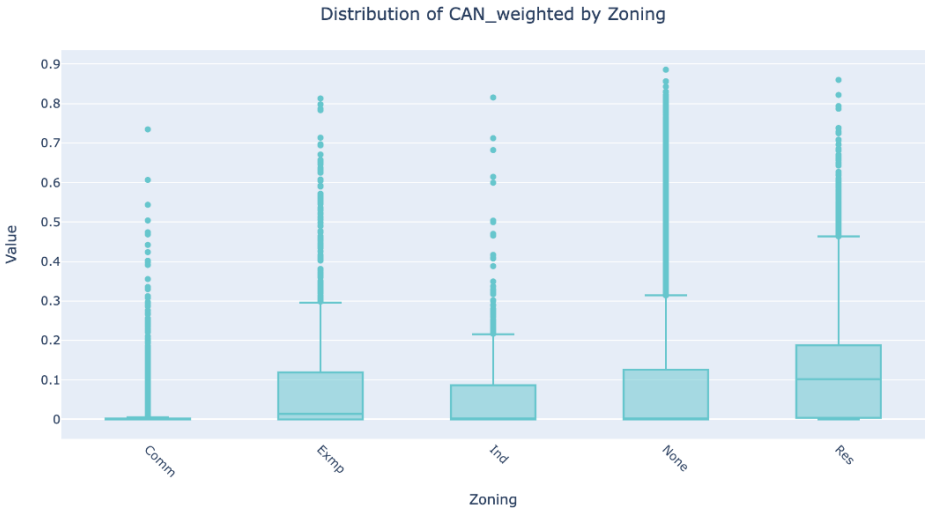


Figure 1.6

This plot probably reveals higher canopy coverage in residential or specially designated green zones, which helps mitigate heat through shading and evapotranspiration. Industrial and commercial zones might exhibit lower canopy cover, contributing to higher local temperatures.

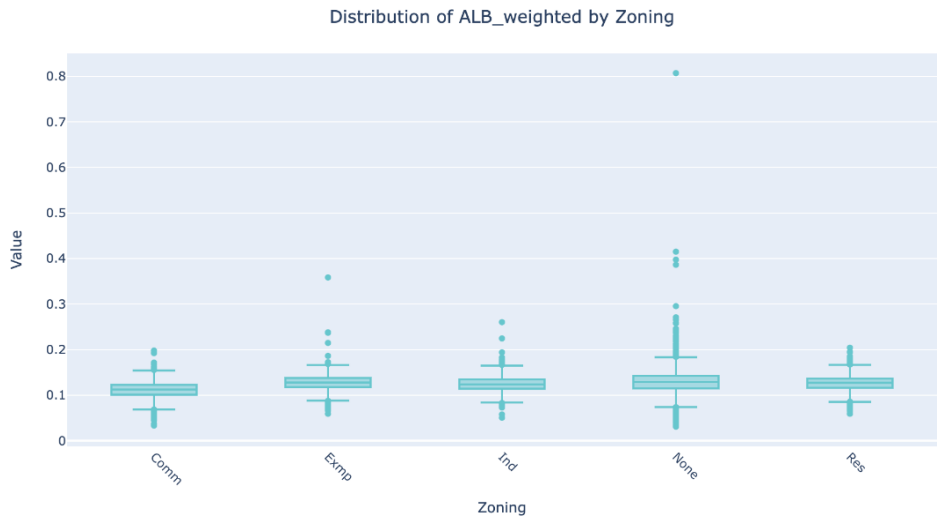


Figure 1.7

The albedo plot could highlight variations in surface reflectivity, with higher albedo expected in areas with more reflective surfaces like commercial zones. Lower albedo in zones with darker surfaces, such as industrial areas or dense residential zones with asphalt roofing, can absorb more heat, amplifying UHI effects.

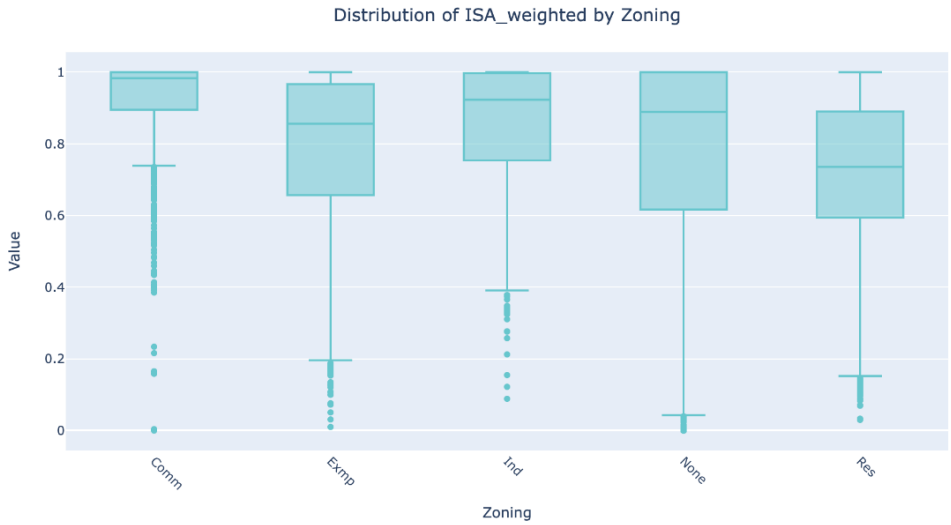


Figure 1.8

This plot likely shows higher impervious surface percentages in commercial and industrial zones, where buildings, roads, and other infrastructure limit the ground's natural cooling. Residential zones, especially those with yards and open spaces, might have lower impervious surface percentages, offering more potential for cooling and water infiltration.

Each of these box plots is relevant for understanding and managing the urban heat island effect, as they collectively demonstrate the environmental impact of different zoning decisions. Urban planners and policymakers can use such insights to promote zoning practices and urban designs that enhance environmental sustainability, such as incorporating green infrastructure in highly impervious zones or optimizing land use for better thermal comfort in urban areas.

Analysis

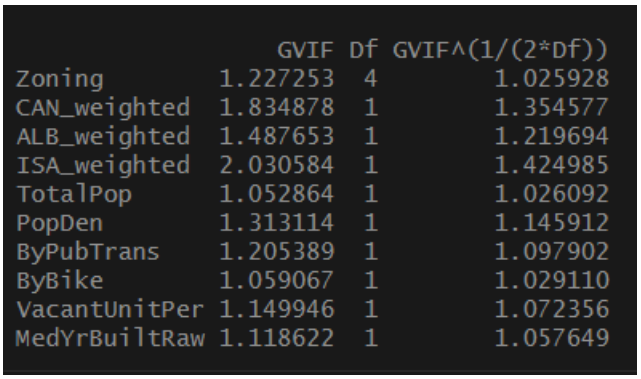
2.1 Linear Regression

To explore relationships between various urban features on the UHI effect, we generated an initial linear regression model. The aim was to ultimately quantify the impact of the urban features on observed temperature patterns. With the preprocessed dataset, we formulated a model with land surface temperature (LST) as the dependent variable and a subset of urban features as independent variables.

In our project, we streamlined the predictive model by removing variables that were statistically insignificant based on their p-values and those that caused multicollinearity as indicated by high Variance Inflation Factor (VIF) scores. The resulting final model includes only the most relevant predictors for a robust and interpretable analysis.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.663e+01  2.684e-01 360.064  < 2e-16 ***
ZoningExmp    -7.615e-01  1.124e-01  -6.778 1.26e-11 ***
ZoningInd      6.003e-01  1.379e-01   4.352 1.35e-05 ***
ZoningNone    -1.052e+00  8.617e-02 -12.212  < 2e-16 ***
ZoningRes     -7.936e-01  8.507e-02  -9.329  < 2e-16 ***
CAN_weighted  -2.010e+01  2.255e-01 -89.113  < 2e-16 ***
ALB_weighted  -9.716e+00  1.401e+00  -6.935 4.20e-12 ***
ISA_weighted   6.754e+00  1.454e-01  46.445  < 2e-16 ***
TotalPop      -2.053e-04  1.318e-05 -15.584  < 2e-16 ***
PopDen         1.112e-05  1.565e-06   7.105 1.24e-12 ***
ByPubTrans     5.779e+00  1.952e-01  29.611  < 2e-16 ***
ByBike         1.219e+01  9.552e-01  12.765  < 2e-16 ***
VacantUnitPer -4.472e+00  4.961e-01  -9.015  < 2e-16 ***
MedYrBuiltRaw  5.376e-09  7.367e-10   7.297 3.05e-13 ***
---
```

Figure 2.1.0 shows the Statistical summary of Linear Regression model

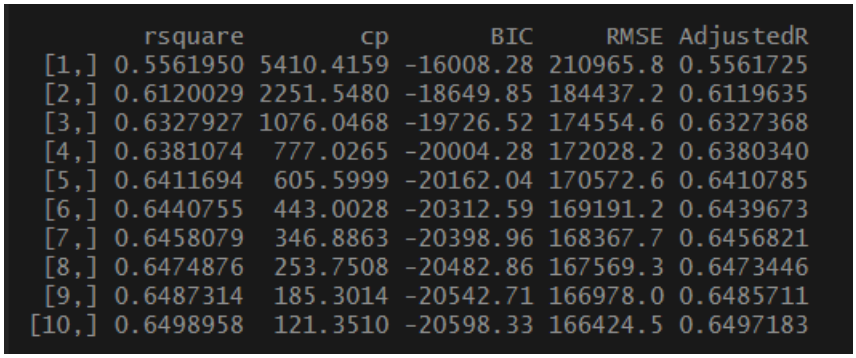


	GVIF	Df	GVIF^(1/(2*Df))
Zoning	1.227253	4	1.025928
CAN_weighted	1.834878	1	1.354577
ALB_weighted	1.487653	1	1.219694
ISA_weighted	2.030584	1	1.424985
TotalPop	1.052864	1	1.026092
PopDen	1.313114	1	1.145912
ByPubTrans	1.205389	1	1.097902
ByBike	1.059067	1	1.029110
VacantUnitPer	1.149946	1	1.072356
MedYrBuiltRaw	1.118622	1	1.057649

Figure 2.1.1 shows the VIF of our variables for the Linear Regression model

After removing variables with high VIF, we are left with the ones whose VIF value is less than 1.5. Hence, we can conclude that our new model does not have multicollinearity.

In our project, we refined our model using the best subset selection approach. This helped us identify the optimal number of predictor variables, balancing model complexity with predictive power, as evidenced by metrics such as CP, R-squared, BIC, and RMSE.



	rsquare	cp	BIC	RMSE	AdjustedR
[1,]	0.5561950	5410.4159	-16008.28	210965.8	0.5561725
[2,]	0.6120029	2251.5480	-18649.85	184437.2	0.6119635
[3,]	0.6327927	1076.0468	-19726.52	174554.6	0.6327368
[4,]	0.6381074	777.0265	-20004.28	172028.2	0.6380340
[5,]	0.6411694	605.5999	-20162.04	170572.6	0.6410785
[6,]	0.6440755	443.0028	-20312.59	169191.2	0.6439673
[7,]	0.6458079	346.8863	-20398.96	168367.7	0.6456821
[8,]	0.6474876	253.7508	-20482.86	167569.3	0.6473446
[9,]	0.6487314	185.3014	-20542.71	166978.0	0.6485711
[10,]	0.6498958	121.3510	-20598.33	166424.5	0.6497183

Figure 2.1.2 shows the value of CP, RMSE and BIC for each number of Variables

From Figure 2.1.2 we decided to choose the lowest value of CP, BIC, RMSE and highest value of adjusted R^2 for our model as this would give us a better goodness of fit and better model considering the penalty for complexity. Therefore, we decided to include all the variables (10) in our model.

Following model specification, we conducted a series of diagnostic checks to assess the validity of the linear regression assumptions. Specifically, we examined the residuals for homoscedasticity, normality, linearity, and independence to ensure the reliability and robustness of the model. However, our analysis revealed significant deviations from the homoscedasticity and normality assumptions, indicating potential limitations and shortcomings of the linear regression approach.

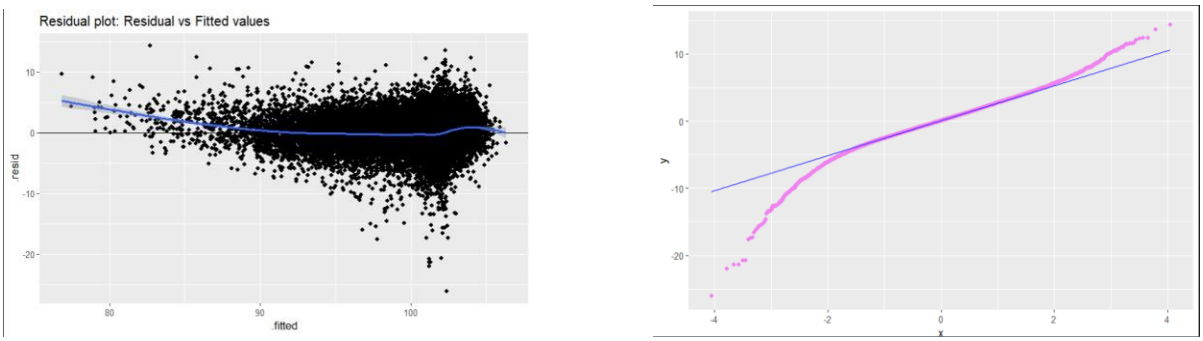


Figure 2.1.3 shows the Homoscedasticity plot and Normality plot for our model

According to Figure 2.1.3, we can conclude that our linear regression model fails the Homoscedasticity and Normality test. This is further made evident when we ran the Breusch-pagan test and Shapiro-Wilk test on our model, where both returned a P value of less than 0.05 which indicates that it does fail the tests mentioned above.

As such, it became evident that alternative modeling approaches capable of accommodating nonlinear relationships and addressing the violated assumptions were warranted. In the subsequent section, we discuss our exploration of Support Vector Regression (SVR) as a promising alternative, offering enhanced flexibility and robustness for analyzing the UHI effect in Boston.

2.2 Random Forest

Permutation feature importance measures a feature's significance in a predictive model. The method shuffles each feature's values randomly. This breaks the connection between the feature and the target. The model's performance usually drops if the feature is important. The size of the drop shows the feature's importance. This calculation happens after the model is trained.

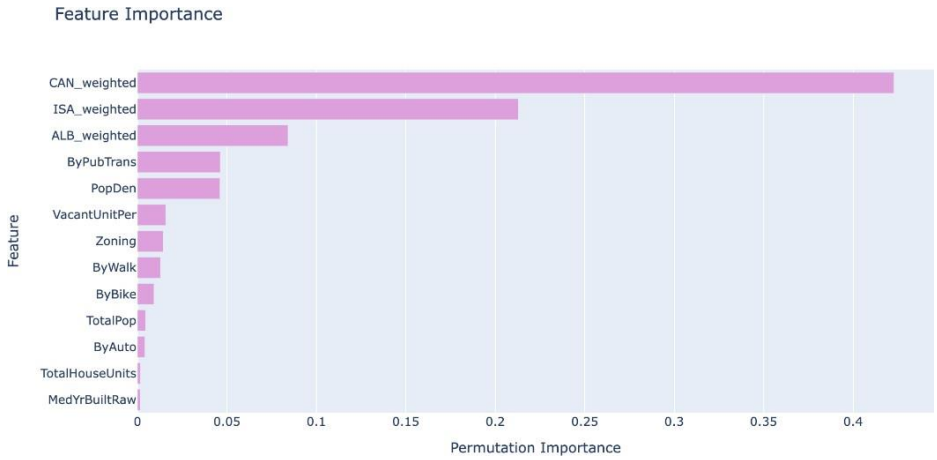


Figure 2.2.0 shows the plot for Random Forest Feature Importance

From the plot above we can see that CAN_weighted has the highest importance among all the features included in the model. This suggests that the canopy cover (weighted) has the most significant impact on the model's predictions for LST. The impervious surface area (ISA) is the second most important feature. It also seems to be quite influential in the model's predictions. The albedo (weighted) comes next in terms of importance, indicating it's also a strong predictor for LST. ByPubTrans represents the mode of public transportation usage and is the fourth in importance, but it's significantly less important than the top three features. Population density has some importance but is less significant compared to the top features. VacantUnitPer and others lower on the chart have progressively lesser importance.

According to the adjusted R-squared, the variables in our model explain approximately 79% of the variation in land surface temperature. The RMSE is approximately 2.42. These metrics will be used for comparison with gradient boosting in the following section.

2.3 Gradient Boosting

Gradient boosting adds new models to an ensemble of models sequentially. In each iteration, a weak base-learner model is trained to address the errors of the ensemble learned thus far (Natekin & Knoll, 2013). This iterative approach often yields more accurate results compared to traditional random forests.

To compare the feature importance produced by the random forest with that which is produced by gradient boosting, we implemented a method of fitting a gradient boosting regressor to predict the land surface temperature (LST). In this process, the feature “ByAuto” is excluded from the predictors as it is low in feature importance and produces multicollinearity. The remaining features are normalized using a standard scaler. The results of feature importance are presented in Figure 2.3.0.

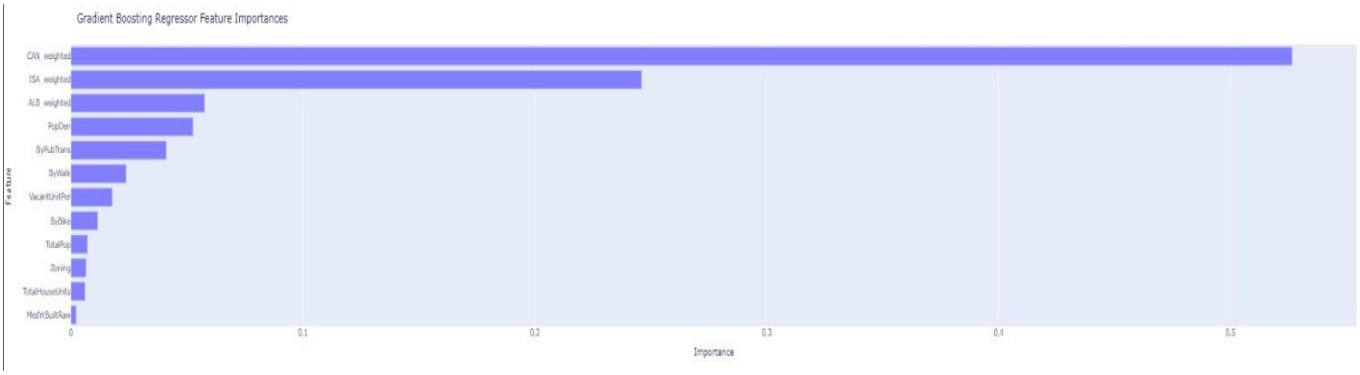


Figure 2.3.0 shows plot for feature importance using gradient boosting

The order of variables is similar for only the first three variables, but CAN_weighted is weighted much higher in terms of importance for the gradient boosting model.

After training, the model yields an R-squared value of approximately 0.7716, signifying that about 77.16% of the variance in LST is explained by the model. The adjusted R-squared score is approximately 0.7715, which adjusts for the number of predictors in the model, reaffirming that the model explains a significant portion of the variance in LST after accounting for the number of features. Additionally, RMSE is approximately 2.56. However, according to these metrics, the conventional random forest model performs better.

2.5 Support Vector Regression

The results from the Support Vector Regression (SVR) model provide insight into how different features affect the Land Surface Temperature (LST). A higher zoning value correlates with a decrease in LST, as indicated by a coefficient of approximately -3.2384. The canopy cover, or CAN_weighted, with a large negative coefficient, suggests a substantial cooling effect on the LST. In contrast, the positive coefficient for impervious surface area, ISA_weighted, implies an increase in LST with more impervious surfaces. Population size and density both show positive links to LST, hinting that areas with more people tend to have higher temperatures.

Interestingly, the use of automobiles shows a negative relationship with LST, while using public transportation, biking, and walking are positively associated with higher LST, albeit to varying degrees. The total number of housing units shows a slight positive effect on LST, suggesting denser housing contributes to warmer temperatures. A higher percentage of vacant units also shows a positive correlation with LST. The age of buildings, captured by MedYrBuiltRaw, shows that newer constructions are linked to lower LST.

The model's adjusted R^2 score is about 0.7715, indicating that roughly 77.15% of the variation in LST is explained by the SVR model. This high value suggests the model fits the data well.

Conclusion

In summary, our analysis has revealed valuable insights into the intricate relationships between urban features and the Urban Heat Island (UHI) effect in Boston. While linear regression provided initial insights, it became evident that its limitations in handling the complexities of our dataset hindered its predictive performance. The assumptions of linearity, equal variance, and independence of errors were not met, indicating the need for alternative modeling approaches.

Random forest and gradient boosting techniques emerged as robust alternatives, offering the ability to capture nonlinear relationships and interactions within the data. Random forest and gradient boosting, through permutation feature importance, identified significant predictors for land surface temperature (LST), highlighting the importance of canopy cover, impervious surface area, and albedo. Based on our evaluation metrics, the random forest model demonstrated superior performance in displaying the most important features for determining land surface temperature.

Moreover, our exploration of Support Vector Regression (SVR) as an alternative modeling approach further enriches our understanding of the UHI effect. SVR's ability to handle non-linear data and robustness against overfitting make it a promising choice for analyzing complex datasets. By leveraging SVR, we can better understand the impact of urban features on observed temperature patterns through the model's ability to display the direction of impact for each predictor.

Particularly, we know that an expansive tree canopy is vital in suppressing the UHI effect due to its ability to provide shade. Also, using city design methods like lighter colored roofs to contribute to higher albedo can significantly reduce land surface temperatures in urban areas (Pearce, 2018).

Overall, our findings emphasize the importance of employing advanced modeling techniques to effectively analyze and predict the UHI effect. By harnessing the capabilities of random forest, gradient boosting, and SVR, we can uncover hidden insights, identify key contributors to temperature variations, and inform evidence-based strategies for mitigating the impacts of urbanization on climate and promoting sustainable urban development.

References

An Investigation into Minimizing Urban Heat Island (UHI) Effects: A UK Perspective, Energy Procedia, Volume 62, 2014, Pages 72-80, ISSN 1876-6102, https://doi.org/10.1016/j.egypro.2014.12.368.

Andrew Trlica, 2017, "Urban Land Cover and Urban Heat Island Effect Database", https://doi.org/10.7910/DVN/GLOJVA, Harvard Dataverse, V3, UNF:6:qjzK93y6ryfXkn0MLfUH3g== [fileUNF]

Christopher O'Malley, Poorang A.E. Piroozfarb, Eric R.P. Farr, Jonathan Gates,

Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013 Dec 4;7:21. doi: 10.3389/fnbot.2013.00021. PMID: 24409142; PMCID: PMC3885826.

Pearce, F. (2018, May 21). Urban Heat: Can White Roofs Help Cool the World's Warming Cities? Yale Environment 360. https://e360.yale.edu/features/urban-heat-can-white-roofs-help-cool-the-worlds-warming-cities

Smola, A., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199-222. Retrieved from https://alex.smola.org/papers/2004/SmoSch04.pdf