Analysis on the Persisting Effects of Redlining to Chicago Neighborhoods

Joel Meyer, Illinois Institute of Technology[*]

Amirezza Eshraghi, Illinois Institute of Technology[†]

Spring 2022 SoReMo Fellowship Project Final Technical Report

Contents

Executive Summary 1

Abstract 2

Background 3

Technical details 4

Findings 6

Future Work 8

License 8

Acknowledgements 9

Appendix 9

Executive Summary

Redlining was predominately apparent in the everyday lives of black and minority residents of many major cities throughout the 20th century. While no longer a legal practice, the effects of such structural racism have persisted in more than just the racial and ethnic makeup of neighborhoods.

A rather new area of research looks at the relationship between historical redlining practices and current levels of green infrastructure. Using geospatial data, this body of research compares neighborhoods in Chicago based on each neighborhood's percentage of area covered in redlining maps—broken down by grade—with the green infrastructure measures of tree canopy area and public park area. Tree Canopy area is by far the most common measure looked at for green infrastructure in the research already published on the topic. Because Chicago has been lauded by the Environmental Protection Agency (EPA) for their initiatives to bolster the urban tree canopy by a conglomerate of many community partners, this research provides a starting point for quantitative decision-making on strategic planning and funding for such initiatives.

We managed data collection through publicly available geospatial datasets that we manually cleaned and tabulated by neighborhood. Tree canopy information was collected using image processing techniques applied

[*] Contact the author.

[†] Contact the author.

to manually generated images of each neighborhood. These manually generated images are archived and available for future research. We also conducted initial linear regression studies with publicly available datasets on energy and the housing market.

The main challenge in looking at the relationship between redlining practices and green infrastructure is the vast amount of time between which these two variables are collected. For instance, redlining information generally comes from around the 1940's, whereas green infrastructure information was collected from as recently as possible. Using publicly available datasets, no strong linear relationships exist between the two categories of variables studied, however clustering neighborhoods into shared qualities from each category helps to bridge the vast time difference in variables studied. This allows us to see which neighborhoods have characteristics that line up with predictions on green infrastructure based on redlining, which may have vastly better green instructure measures as a result of neighborhood investment over time, and which, if any, may have worse green infrastructure measures as a result of declining investment in the neighborhood.

A major part of the research project was to also create visual representations at the scale of individual neighborhoods overlaying redlining information and green infrastructure measures. These are all archived in the Illinois Tech repository and available for use by the public. Each neighborhood has two maps: one comparing redlining information with tree canopy, and one comparing redlining information with public parks.

Our findings indicate that neighborhoods currently fit into four clusters that share characteristics related to the data collected. In order of lowest rates of green infrastructure to highest:

- **Cluster 1** (N=28) were primarily redlined and have lower canopy coverage, but slighter higher rate of public parks as a percentage of land area compared to the cluster below
- **Cluster 2** (N=33) were a mix of yellowlined and redlined with a canopy in line with the mean for all neighborhoods
- **Cluster 3** (N=19) were primarily yellowlined and have a similar canopy to the cluster below, but a much lower concentration of public parks as a percentage of land area.
- **Cluster 4** (N=18) were primarily not districted in redlining maps and have high canopy coverage and a high concentration of public parks as a percentage of land area.

Findings also indicate the potential for a relationship between redlining information and other factors in the built environment. While further studies should be conducted to strengthen the suggestion of correlated relationships, this broad overview of initial research suggests the potential for such predictive relationships to exist.

While we cannot go back in time to prevent redlining from happening, there is a societal need to help alleviate the disparities structurally erected from past practices that affected home ownership and segregated neighborhoods in such an outright way. As previous research on the subject indicates, neighborhoods with high instances of redlining are often hotter and have less clean air. Improving the urban tree canopy is a first step towards redressing those inequalities. There are already organizations dedicated to such a mission, however, they still face hurdles in funding and maintenance. They need to build a strong case for the allocation of funds into this important piece of green infrastructure. Perhaps the data collected, images produced, and findings outlined in this research could help these organizations get one step closer to securing the financial resources necessary to create lasting changes in the green infrastructure of neighborhoods currently suffering the most.

Abstract

This report looks at "redlining" maps produced by the Home Owners' Loan Corporation (HOLC) to analyze the current green infrastructure levels in various neighborhoods of Chicago. Redlining maps, produced in the first half of the 20th century, essentially graded neighborhoods on their riskiness for mortgage lenders. These maps often followed strict racial lines marking neighborhoods with a majority of black and other minority

occupants in red to indicate they were the riskiest, hence the term redlining.[1] Gathering geospatial data from these HOLC maps and overlaying them with maps showcasing green infrastructure indicators provides a visual representation of the relationship between structural disinvestment in neighborhoods and their current green infrastructure levels. Additionally, some neighborhoods break the pattern, showcasing which areas have changed the most with investment, indicating changes that can be associated with things like gentrification. In the end, k-means clustering best showcases the patterns that have emerged when looking at two variables with such a large gap in time between when the data was collected: redlining information from the 1940's, and green infrastructure indicators from to 2010's.

This research was taken a step further to begin descriptive analytical assessments of the ways in which redlining information relates to a variety of other datasets on the built environment. In this way, trends started to emerge that could suggest redlining as a predictor for other variables. However, the neighborhood scale at which the data was grouped does not produce strong enough indications. Nonetheless, continuing this line of study at a smaller scale, perhaps at the census tract level, could lead to more conclusive findings with a stronger correlation between redlining information and other datasets on the built environment that highlight the need for, and lack thereof, of green infrastructure in disenfranchised areas.

Background

The term "redlining" comes from maps drawn by the HOLC, and later taken up by both the Federal Housing Administration (FHA) and the Veterans Administration. These maps indicated where it was deemed safe, and where it was deemed risky to insure mortgages, which were needed in order to buy property anywhere in a major city.[2] Areas colored red indicated that a neighborhood was deemed too risky to insure mortgages, and these areas were almost always black neighborhoods or neighborhoods adjacent to them. This made home ownership nearly impossible for many black and minority residents in major cities following World War II. The rest of the maps were broken down into yellow districts for declining, blue for still desirable, and green for best. Odds of getting an insured mortgage increased respectively moving through the list from red to green.

Green Infrastructure was defined in the Water Infrastructure Improvement Act, passed by Congress in 2019, as "the range of measures that use plant or soil systems, permeable pavement or other permeable surfaces or substrates, stormwater harvest and reuse, or landscaping to store, infiltrate, or evapotranspirate stormwater and reduce flows to sewer systems or to surface waters."[3] Green infrastructure comes in a variety of scales and when implemented can also help provide cleaner air, provide flood protection, diversify habitats, and beautify green spaces.[4]

Other research has been done on the relationship between redlining and green infrastructure in other cities in the United States. One study claims the disparities in air pollution measurements were more pronounced when looking at HOLC grade rather than race or ethnicity.[5] Another analyzed data from many U.S. cities and found that in 94% of regions studied, formerly redlined areas were hotter than formerly non-redlined areas by an average of 4.7°F.[6] It's not hard to imagine why—less shade from trees means the sun can warm larger areas on the street, composed of cementitious elements that have a high heat retention. Relating this information into visual maps has also been conducted for many cities to showcase the inequalities present in

[1] Camila Domonoske, "Interactive Redlining Map Zooms in on America's History of Discrimination," NPR, NPR, October 19, 2016. https://www.npr.org/sections/thetwo-way/2016/10/19/498536077/interactive-redlining-map-zooms-in-on-americas-history-of-discrimination

[2] Terry Gross, "A 'Forgotten History' of How the U.S. Government Segregated America," NPR, Fresh Air, May 3, 2017. https://www.npr.org/2017/05/03/526655831/a-forgotten-history-of-how-the-u-s-government-segregated-america

[3] Language taken from the Water Infrastructure Improvement Act, Public Law 115-436, January 14, 2019. https://www.congress.gov/115/plaws/publ436/PLAW-115publ436.pdf

[4] "What is Green Infrastructure," EPA, Last modified March 31, 2022. https://www.epa.gov/green-infrastructure/what-green-infrastructure

[5] Haley Lane et al, "Historical Redlining is Associated with Present-Day Air Pollution Disparities in U.S. Cities," ACS Publications, Environmental Science and Technology Letters, March 9, 2022. https://doi.org/10.1021/acs.estlett.1c01012

[6] Jeremy Hoffman, "The Effects of Historical Housing Policies on Resident Exposure to Intra-Urban Heat: A Study of 108 US Urban Areas," ResearchGate, Climate 8, no. 12 (2020): 1. http://dx.doi.org/10.3390/cli8010012

green infrastructure. Clearly, inequalities in green infrastructure affect many areas of life from public health to energy consumption and housing.

Chicago offers a great starting point for a detailed look at this relationship because of its initiatives towards improving the urban tree canopy—a major component of green infrastructure. The EPA lists the Chicago Region Trees Initiative (CRTI) under its examples for improving urban tree canopy as a way to strengthen green infrastructure. CRTI is composed of several community organizations, business partnerships, and government organizations aiming to expand and diversify the urban tree canopy. The largest initiative of its kind in the United States, CRTI cites urban development and a lack of funding for planting and proper care as major hurdles to their mission.[7]

This research focuses on tree canopy as a major indicator of green infrastructure in Chicago neighborhoods for this reason. Because there are publicly available datasets produced by community stakeholders invested in the Chicago area urban tree canopy, we imagine this area of research would be useful for helping these organizations with strategic planning as well as having something to use when seeking resources such as grants in order to improve the urban tree canopy in Chicago. This research can also be potentially utilized when making decisions on the approval of developments within the city.

Technical details

Geographic Imaging Software (GIS) was primarily used for data collection. By isolating layers of publicly available geospatial datasets, we were able to collect information on the area of neighborhoods and redlining districts, as well as park district areas and tree canopy coverage. The table below outlines the layers, sources, and file types.

Layer	Source	File Type
Boundaries - Chicago Neighborhoods	Chicago Data Portal[8]	Vector
Chicago Redlining Maps	Mapping Inequality - Redlining in New Deal America; University of Richmond[9]	Vector
Parks - Chicago Park District Park Boundaries (current)	Chicago Data Portal[10]	Vector
Chicago Regional Land Cover Data Set	Spatial Analysis Laboratory (SAL) at the University of Vermont[11]	Raster

[7] This information was taken from the about section of the CRTI website which can be located at https://chicagorti.org/about

[8] "Boundaries - Neighborhoods," Chicago Data Portal, https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Neighborhoods/bbvz-uum9. The shapefile format was utilized. Neighborhood boundaries in Chicago, as developed by the Office of Tourism. These boundaries are approximate and names are not official. Data provided by City of Chicago.

[9] Robert K. Nelson et al, "Mapping Inequality," American Panorama, ed. Robert K. Nelson and Edward L. Ayers. https://dsl.richmond.edu/panorama/redlining. The shapefile format of the data was utilized. The site also contains scans of the maps from the National Archives as well as Area Description Images. All of the scans of the HOLC maps are in public domain, with the vast majority coming from the National Archives.

[10] "Parks - Chicago Park District Park Boundaries (current)," Chicago Data Portal, https://data.cityofchicago.org/Parks-Recreation/Parks-Chicago-Park-District-Park-Boundaries-curren/ej32-qgdr. The Shapefile format of the data was utilized. Current boundaries of Chicago Park District properties as of November 4, 2016. Data provided by Chicago Park District; Public Domain.

[11] Jarlath O'Neil-Dunne, "Chicago Regional Land Cover Data Set," Letters from the SAL (2016). http://letters-sal.blogspot.com/2016/06/chicago-regional-land-cover-dataset.html. Seven land cover classes were mapped: 1) tree canopy, 2) grass/shrub, 3) bare soil, 4) water, 5) buildings, 6) roads/railroads, and 7) other paved surfaces. The tree canopy layer was isolated and utilized. The dataset is limited to a resolution of one pixel as 2ft. Project partners for the dataset include the USDA Forest Service, American Forests, the Bank of America Charitable Foundation, the Morton Arboretum, the Chicago Metropolitan Agency for Planning, and the Field Museum.

The code we wrote for the report can be found on Github, along with the datasets. A manual for following along is located with Github, as well as listed out in the Appendix

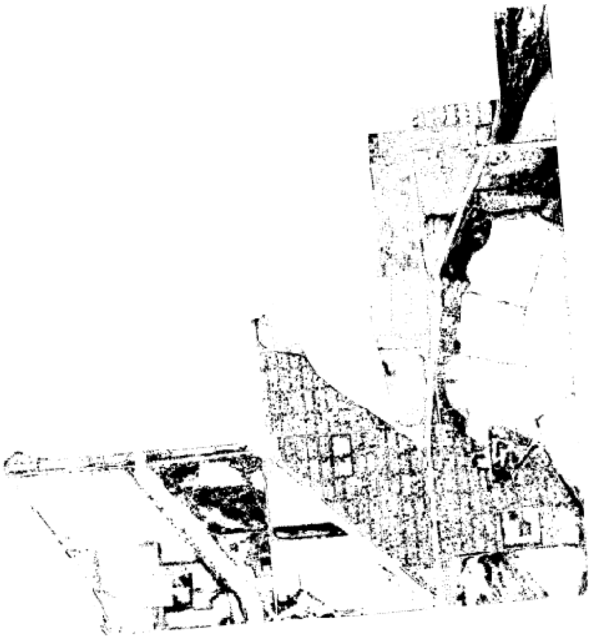
Vector Information

For calculating areas, the redlining and park boundaries needed to be slightly reworked to be bounded by the neighborhood boundaries. Manipulation involved splitting redlining and park boundaries that crossed over any neighborhood boundary to isolate each variable by neighborhood. Areas were tabulated and used to calculate the percentages of a neighborhood that were redlined, yellowlined, blueline, and greenlined as well as the percentage of a neighborhood covered by public park land.

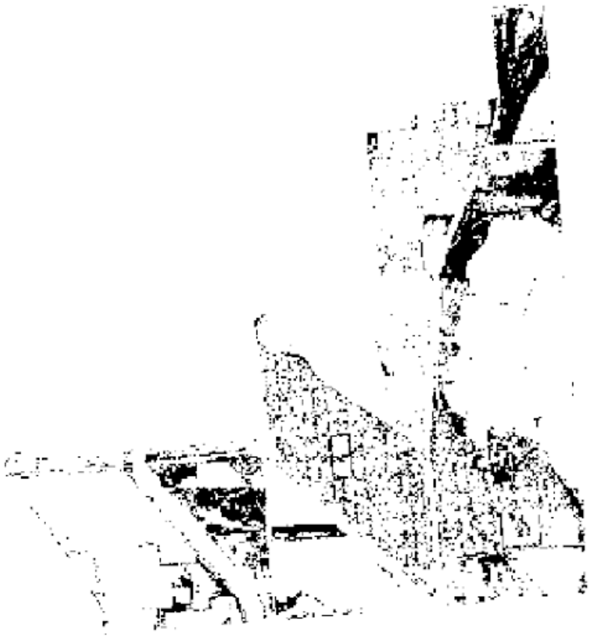
Raster Information

Because the tree canopy was a raster layer, image processing techniques were utilized to calculate the percentage of each neighborhood covered by the tree canopy. For producing images, the raster information for the tree canopy was overlaid with the vector information for the neighborhood boundary manually, which leads to some deviation in accuracy. However, this error is minimal given the resolution achievable within the software utilized, namely QGIS and Adobe Illustrator. A .png image of the tree canopy was eventually exported, which was masked using the neighborhood boundary vectors. Each of these masks was separated to different layers and exported as a new .png for each neighborhood to calculate the percentage of each neighborhood covered by the tree canopy.

Image processing techniques were used to calculate the percentage of each neighborhood covered by the tree canopy. With the .png images exported as outlined above, a count of black pixels (which represent tree canopy) over the count of alpha pixels (opaque pixels as compared to transparent pixels to define the border) gave the percentage. The image was then recreated in Python given the count to check for accuracy and resolution. Had the Python-generated image come back with less granularity and sharp contrast, the count would have been rejected. An example of an original manual .png output and a python generated image of the count is presented below.



Original .png image for Hegewisch created in Adobe Illustrator.



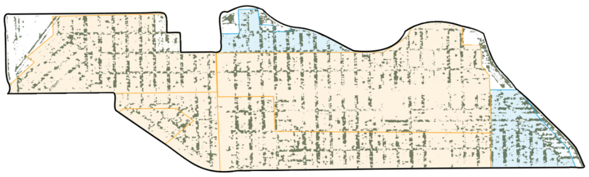
Python generated image for Hegewisch used to check accuracy and resolution.

The .png images of all of the neighborhoods are archived and can be accessed in the Illinois Tech repository for further research.

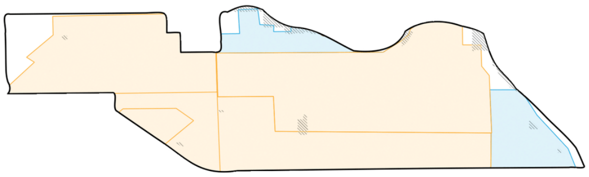
These three sets of information–redlining boundary percentages, public park percentages, and tree canopy percentages–are combined in Table 1 in the appendix. Note the names of the neighborhoods from the Chicago Data Portal are referenced as the neighborhoods for which data was collected.

Visual Representation

Maps of each neighborhood comparing redlining boundaries to public parks, and redlining boundaries to the tree canopy were created in a similar manner, combining layers in QGIS and exporting in Adobe Illustrator to change visual characteristics. All of these maps can be found in the Illinois Tech repository. An example for Albany Park has been provided below. For this neighborhood, you can see the yellow portions that indicate yellowlining and the blue portions that indicate bluelining.



Map of Albany Park comparing redlining and tree canopy



Map of Albany Park comparing redlining and public parks

Findings

Distributions

Foremost, visualizing the distribution of features for each neighborhood helps provide an initial understanding of redlining and green infrastructure compared to a normal distribution. The distribution charts in Table 2 shows how the elements are concentrated across the city of Chicago. Additionally, using QQ-plots found in Table 3 shows the lack of normality in the features that data was collected on.

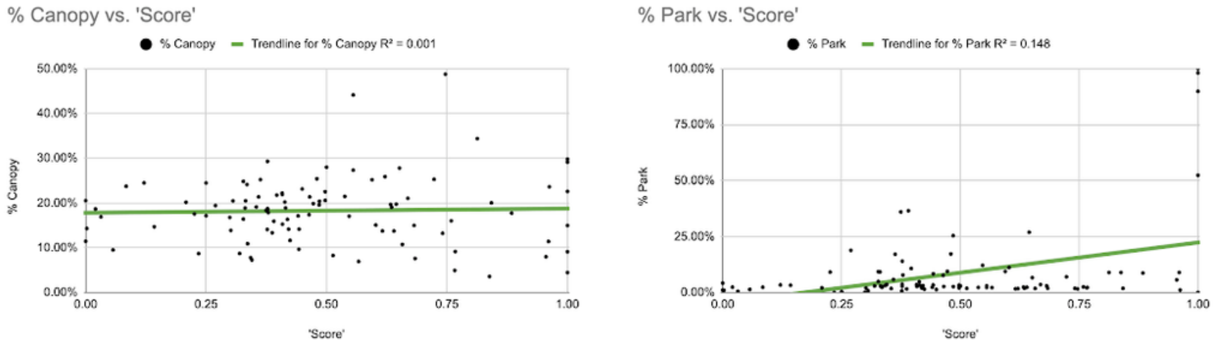
Given the low normality within the dataset, simple trendlines are unlikely to significantly show any correlations between redlining information and green infrastructure characteristics. Scatterplots in Table 4 show the relationships between each of the four redlining categories and both green infrastructure features.

Linear Regression

Another method for extracting findings from the data set would be to combine all of the redlining percentages into a single quantifiable score for each neighborhood. This would allow for comparison between redlining and the green infrastructure variables to find trends with a regression line. This was tested by creating a score for each neighborhood on a scale of 0-1 in which neighborhoods closer to 0 would have been more affected by redlining, and neighborhoods closer to 1 would have been less affected. To calculate the score, the following equation was used:

(%RL*1 + %YL*1.25 + %BL*1.5 + %GL*1.75 + %UL*2) - 1.

This equation weights redlining to skew a neighborhood score closer to zero, and weights areas not districted in redlining maps to skew closer to one. In the equation above, %RL is the % redlined column of Table 1, with other variables named respectively following the table. This method is far from soundproof, but was conducted as an introductory look at the data. For instance, areas not districted in redlining maps were either undeveloped at the time, or slated for industrial use and thus had no housing in the area. These two use-cases shape neighborhoods in vastly different ways. On one hand, undeveloped areas often provided a haven for new development to house the white flight fleeing from their neighborhoods due to blockbusting, which would reflect an influx of wealth into a neighborhood. On the other hand, industrial zones were often concentrated near redlined areas because housing located in areas with high noise and air pollution weren't ideal.



Clearly no strong relationships can be drawn from these rudimentary tests. This is likely due to the large gap in time between the two variables being compared. The redlining percentages are taken from maps produced around the 1940's, whereas the green infrastructure data was taken from information collected within the past decade. With over half a century between datasets, it's safe to assume that neighborhoods have changed in terms of investment and development between the 1940's and now.

Clustering

In order to account for the gap in time between datasets, drawing conclusions from clusters provides better insights into the main ways in which neighborhoods are shaped by redlining and green infrastructure. Utilizing K-means Clustering, similar data points are aggregated together based on the least difference in shared characteristics in the dataset. Looking at 98 neighborhoods, the number of clusters was a vital question to be discerned because too few clusters wouldn't create distinct enough features in the clusters, and too many would over-emphasize small differences in the data set. Using the Elbow Rule, which looks at a Within Cluster Sum of Square (WCSS), the ideal number of clusters was tested between a range of 1 to 11 clusters, and 4 was identified as having the least error. Tables for the clustering can be found in Table 5. Map 1 is a map of the clustering results.

Prediction Power

Clustering the datasets on shared characteristics of redlining, tree canopy, and public park area showed a potential for similarities among our clusters and other features of interest from other datasets on housing, energy usage, and public health which are factors affected by a neighborhood's history of redlining and current levels of green infrastructure. In other words, the ways in which redlining and green infrastructure work together in a neighborhood can shape other factors within the built environment.

To get initial insight, we worked with a variety of publicly available datasets outlined below.

Dataset	Source
% of Households Owner Occupied	Institute of Housing Studies at DePaul University[12]
% of Housing as Single Family Units	Institute of Housing Studies at DePaul University[13]
Sales per 100 Residential Parcels	(Average from 2005-2021) Institute of Housing Studies at DePaul University[14]
% of Households Cost Burdened	Institute of Housing Studies at DePaul University[15]

Testing these new features against the initial percentages found for redlining and green infrastructure to see if they fit a simple linear regression model can indicate where relationships might occur. Because only 66 neighborhoods had all of the information available, the findings are not significant to prove correlation or prediction power. Testing only linear regression models is far from exhaustive. However, they do indicate trends that can provide a jumping-point for further research, outlined in the next section.

The results of the linear regression tests can be found in Table 6. Redlining was significant in the % of Households Cost Burdened and % of Households Owner Occupied. Both redlining and yellowlining were significant in % of Housing as Single Family Units and Sales per 100 Residential Parcels.

Future Work

The groupings of clusterings indicate that most neighborhoods follow a trend in which redlining relates to lower levels of green infrastructure. Visualizing these clusters allowed us to gain insights into how redlining information and green infrastructure might interact in reference to other factors on energy and the housing market. An initial look at clusters indicates that clusters are not bound geographically in concentric bands radiating out from the city center. This means that certain neighborhoods on the outskirts of the city have the same concentration of green infrastructure in relation to the city center which is much more dense. Taking into account factors on energy and the housing market can help explain the impact redlining may have had in these areas. Regression tests for these new datasets compared to redlining and green infrastructure can be found in Table 6.

Better understanding the relationships between our initial datasets on redlining and green infrastructure with the new datasets on energy and the housing market would likely need to be pursued through non-linear relationship tests. In order for these tests to have potential for prediction power, a smaller unit of scale in comparison to the neighborhood level, would likely need to be utilized to have a higher sample size. This could be done at the census tract level for instance, which has almost 900 occurrences in Chicago, as compared to the just under 100 neighborhoods within the city.

License

[12] "Housing Market Indicators Data Table," Institute for Housing Studies. https://www.housingstudies.org/data-portal/. On the IHS Housing Market Indicators Data Portal, users can search for, view, and download 13 indicators of housing market health in the Chicago region. The Housing Market Indicators Data Portal makes available data from five core data sets in the IHS Data Clearinghouse. Data available include indicators related to the composition of the housing stock (Cook County only), characteristics of property sales, mortgage lending activity, foreclosure filings, and completed foreclosure auction activity.

[13] Ibid.

[14] Ibid.

[15] Ibid.

Acknowledgements

The authors would like to thank the SoReMo Fellowship Program at the Illinois Institute of Technology, and in particular the participating faculty from the Applied Mathematics department.

The authors would like to thank the Camras Scholars Program at the Illinois Institute of Technology who provided feedback on an initial draft of the findings.

Lastly, the authors would like to thank Gemma Brizzolara, a student in the College of Architecture at the Illinois Institute of Technology, who provided assistance with the map productions for each neighborhood.

Appendix

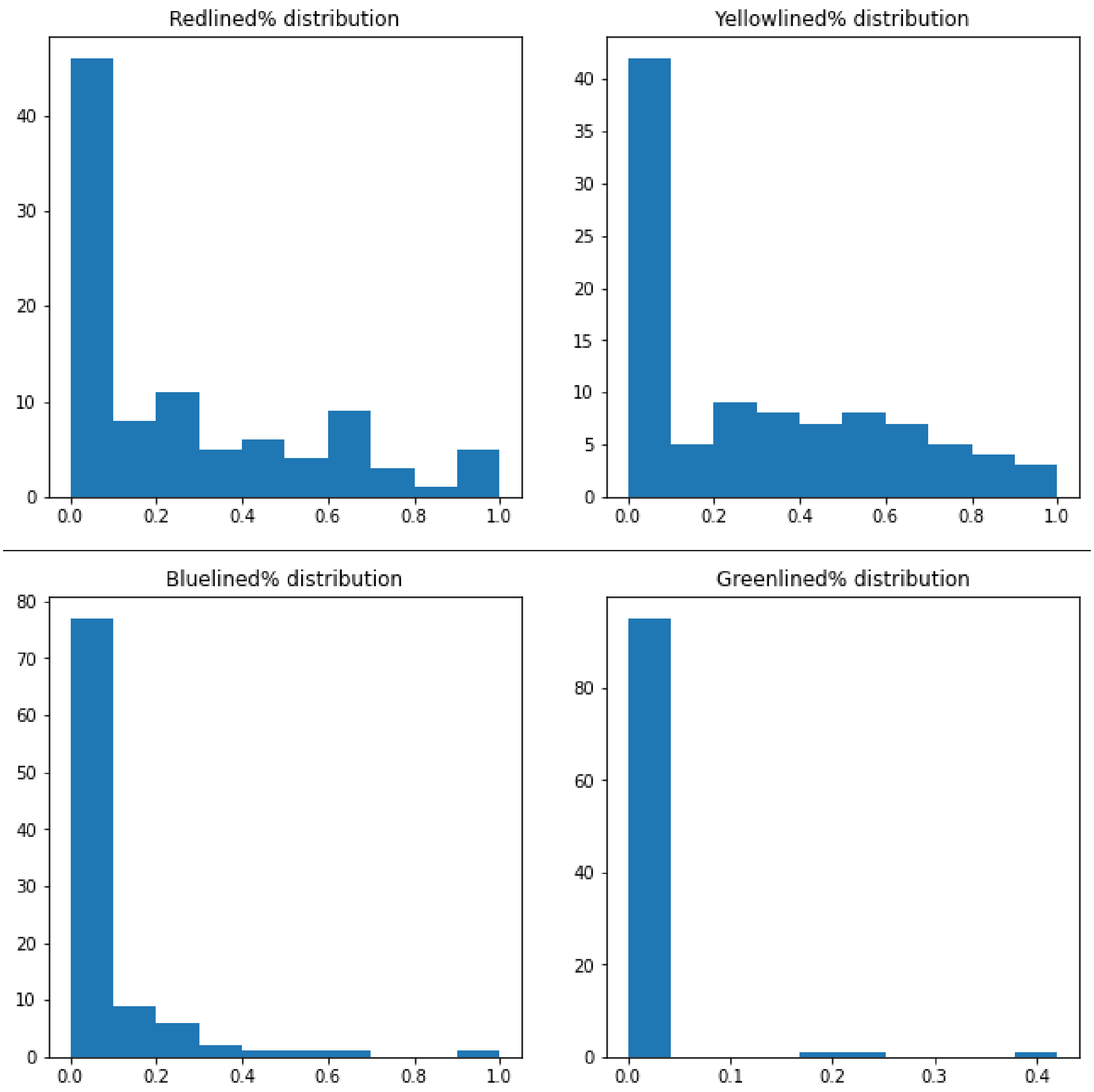
Table 1: Geospatial Data Areas Chart by Neighborhood

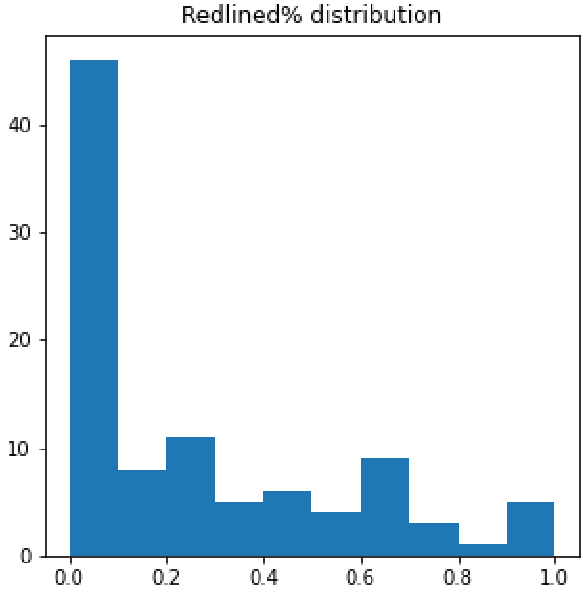
Neighborhood	Redlined	Yellowlined	Bluelined	Greenlined	Park	Canopy
Albany Park	0.00%	80.55	12.18%	0.00%	2.32%	24.13%
Andersonville	0.00%	100.00%	0.00%	0.00%	0.39%	24.47%
Archer Heights	10.28%	28.47%	0.00%	0.00%	1.88%	7.60%
Armour Square	68.05%	0.00%	0.00%	0.00%	2.93%	8.72%
Ashburn	57.99%	0.00%	0.00%	0.54%	2.00%	16.33%
Auburn Gresham	21.25%	44.37%	15.45%	0.00%	2.76%	18.74%
Austin	0.00%	78.21%	0.00%	0.00%	4.77%	20.21%
Avalon Park	0.00%	4.63%	59.33%	0.00%	3.57%	21.04%
Avondale	19.62%	56.92%	0.00%	0.00%	0.76%	14.10%
Belmont Cragin	12.69%	49.12%	19.24%	0.00%	2.72%	15.26%
Beverly	3.58%	29.47%	29.17%	16.78%	2.22%	44.15%
Boystown	0.00%	100.00%	0.00%	0.00%	0.22%	17.15%
Bridgeport	55.74%	0.00%	0.00%	0.00%	3.33%	9.67%
Brighton Park	28.12%	39.26%	0.00%	0.00%	1.49%	11.63%
Bucktown	61.29%	0.00%	0.00%	0.00%	1.61%	13.35%
Burnside	46.22%	0.00%	0.00%	0.00%	2.26%	21.47%
Calumet Heights	7.12%	39.18%	0.00%	0.00%	1.92%	19.07%
Chatham	5.80%	37.18%	35.60%	0.00%	2.60%	20.39%
Chicago Lawn	13.12%	65.48%	0.00%	0.00%	14.02%	18.59%
Chinatown	65.50%	0.00%	0.00%	0.00%	3.57%	7.28%
Clearing	0.00%	31.04%	0.00%	0.00%	2.23%	9.15%
Douglas	73.05%	0.00%	0.00%	0.00%	18.87%	19.42%
Dunning	0.00%	65.85%	1.80%	0.00%	2.06%	20.59%
East Side	28.48%	15.05%	0.00%	0.00%	11.15%	15.06%
East Village	99.76%	0.00%	0.00%	0.00%	1.04%	14.30%
Edgewater	0.00%	85.97%	4.87%	0.00%	9.19%	18.89%
Edison Park	0.00%	0.00%	100.00%	0.00%	2.95%	28.00%
Englewood	71.09%	22.37%	0.00%	0.00%	3.34%	24.48%
Fuller Park	66.38%	0.00%	0.00%	0.00%	2.65%	10.90%
Gage Park	7.27%	67.49%	0.00%	0.00%	3.10%	14.16%
Galewood	0.00%	34.42%	48.94%	0.00%	3.27%	22.52%
Garfield Park	39.18%	50.98%	0.00%	0.00%	9.10%	17.56%
Garfield Ridge	3.92%	29.32%	0.00%	0.00%	1.85%	13.21%
Gold Coast	8.98%	28.63%	17.43%	24.59%	12.16%	17.06%
Grand Boulevard	85.74%	0.00%	0.00%	0.00%	3.24%	14.69%
Grand Crossing	30.93%	44.84%	0.00%	0.00%	1.77%	19.12%
Grant Park	0.00%	0.00%	0.00%	0.00%	89.91%	22.58%
Greektown	76.55%	0.00%	0.00%	0.00%	0.00%	8.70%

Neighborhood	Redlined	Yellowlined	Bluelined	Greenlined	Park	Canopy
Hegewisch	11.58%	0.00%	0.00%	0.00%	8.71%	17.75%
Hermosa	0.00%	61.54%	19.50%	0.00%	2.34%	17.13%
Humboldt Park	13.45%	53.57%	0.00%	0.00%	7.70%	17.37%
Hyde Park	1.17%	60.69%	10.65%	0.00%	17.24%	25.41%
Irving Park	0.00%	88.32%	2.07%	0.00%	4.91%	24.83%
Jackson Park	0.00%	0.00%	0.00%	0.00%	98.12%	29.13%
Jefferson Park	0.00%	70.70%	4.05%	0.00%	1.29%	23.13%
Kenwood	23.66%	49.99%	5.15%	0.00%	17.10%	25.23%
Lake View	0.00%	67.58%	2.86%	2.53%	9.46%	19.86%
Lincoln Park	41.65%	10.91%	3.40%	0.00%	25.41%	19.57%
Lincoln Square	0.00%	32.01%	21.72%	0.00%	6.64%	27.81%
Little Italy, UIC	69.87%	0.00%	0.00%	0.00%	2.03%	13.89%
Little Village	0.00%	51.20%	0.00%	0.00%	1.65%	13.75%
Logan Square	36.14%	57.36%	0.00%	0.00%	2.11%	20.18%
Loop	16.19%	0.00%	0.00%	0.00%	8.96%	3.59%
Lower West Side	43.35%	0.00%	0.00%	0.00%	2.17%	6.92%
Magnificent Mile	0.00%	0.00%	0.00%	0.00%	0.00%	4.46%
Mckinley Park	55.71%	0.00%	0.00%	0.00%	8.34%	14.16%
Millenium Park	0.00%	0.00%	0.00%	0.00%	52.37%	29.80%
Montclare	0.00%	21.04%	0.00%	0.00%	1.87%	20.03%
Morgan Park	36.37%	23.48%	16.55%	0.00%	3.70%	29.31%
Mount Greenwood	58.50%	0.00%	0.00%	0.00%	3.15%	18.90%
Museum Campus	0.00%	0.00%	0.00%	0.00%	100.00%	9.10%
Near South Side	65.75%	0.00%	0.00%	0.00%	3.19%	7.79%
New City	20.02%	15.75%	0.00%	0.00%	2.96%	14.99%
North Center	0.00%	78.91%	0.00%	0.00%	3.18%	21.90%
North Lawndale	29.50%	50.35%	0.00%	0.00%	9.29%	16.41%
North Park	0.00%	0.00%	37.44%	0.00%	8.96%	34.38%
Norwood Park	0.00%	38.17%	18.50%	0.00%	1.72%	25.89%
O'Hare	0.00%	0.00%	0.00%	0.00%	0.02%	14.97%
Oakland	60.94%	0.00%	0.00%	0.00%	36.56%	15.93%
Old Town	100.00%	0.00%	0.00%	0.00%	1.05%	20.53%
Portage Park	0.00%	85.06%	6.00%	0.00%	2.89%	20.50%
Printers Row	27.68%	0.00%	0.00%	0.00%	7.00%	25.30%
Pullman	15.35%	11.70%	0.00%	0.00%	2.51%	16.02%
River North	36.71%	0.00%	0.00%	0.00%	2.42%	19.68%
Riverdale	3.76%	0.00%	0.00%	0.00%	1.15%	23.62%
Rogers Park	1.31%	73.81%	14.87%	0.00%	5.84%	21.36%
Roseland	9.37%	44.09%	22.15%	0.00%	2.66%	21.36%
Rush & Division	100.00%	0.00%	0.00%	0.00%	4.18%	11.46%
Sauganash, Forest Glen	0.00%	6.55%	19.74%	42.23%	1.62%	48.78%
Sheffield & DePaul	65.90%	34.26%	0.00%	0.00%	2.34%	23.74%
South Chicago	48.98%	17.50%	0.00%	0.00%	7.79%	17.86%
South Deering	3.94%	0.00%	0.00%	0.00%	8.98%	11.40%
South Shore	1.27%	59.48%	28.94%	0.00%	10.75%	21.77%
Streeterville	0.00%	0.00%	8.93%	0.00%	5.67%	7.98%
Ukrainian Village	14.61%	73.97%	0.00%	0.00%	0.11%	16.78%
United Center	94.33%	0.00%	0.00%	0.00%	1.38%	9.51%
Uptown	0.00%	47.38%	0.00%	0.00%	26.92%	19.74%
Washington Heights	28.61%	30.10%	16.07%	0.00%	3.45%	22.19%
Washington Park	62.54%	0.00%	0.00%	0.00%	36.03%	18.37%
West Elsdon	0.00%	48.03%	0.00%	0.00%	2.32%	13.73%

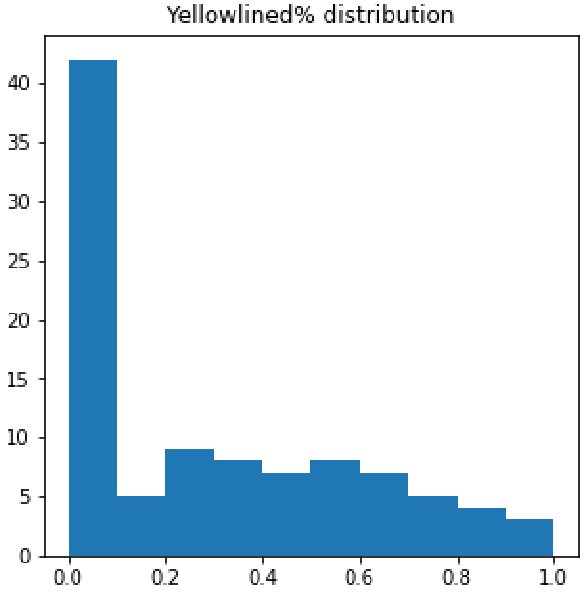
Neighborhood	Redlined	Yellowlined	Bluelined	Greenlined	Park	Canopy
West Lawn	4.01%	40.35%	0.00%	0.00%	1.82%	10.72%
West Loop	23.40%	0.00%	0.00%	0.00%	2.13%	4.95%
West Pullman	28.21%	21.67%	0.00%	0.00%	2.98%	27.35%
West Ridge	0.00%	8.18%	68.82%	0.00%	9.45%	25.19%
West Town	48.65%	0.00%	0.00%	0.00%	1.84%	8.27%
Wicker Park	97.96%	0.00%	0.00%	0.00%	2.36%	18.67%
Woodlawn	47.76%	65.41%	0.00%	0.00%	0.52%	16.90%
Wrigleyville	0.00%	92.55%	0.00%	0.00%	0.76%	20.48%

Table 2: Distribution of Features

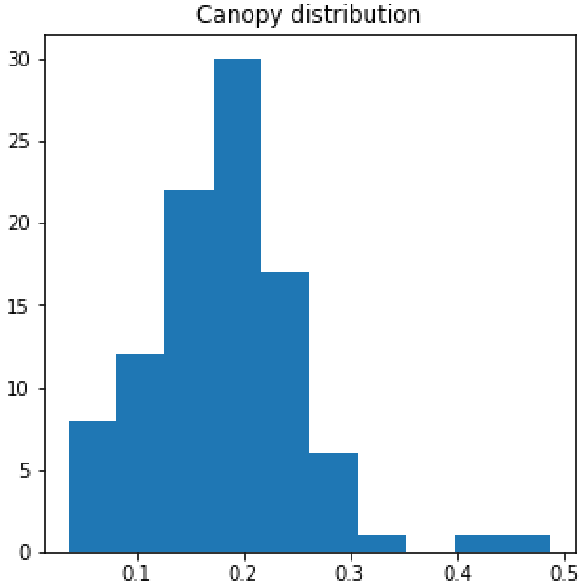




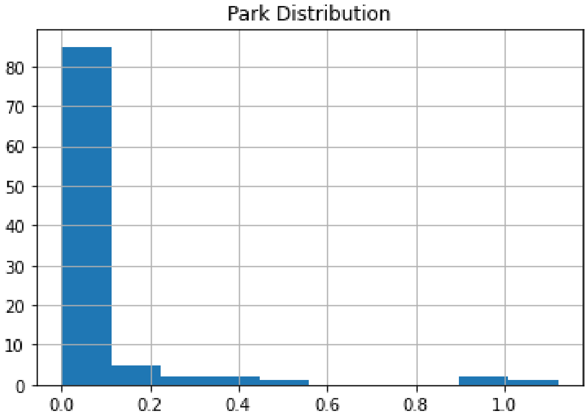
Redlined% distribution



Yellowlined% distribution



Canopy distribution



Park Distribution

Table 3: QQ-Plots

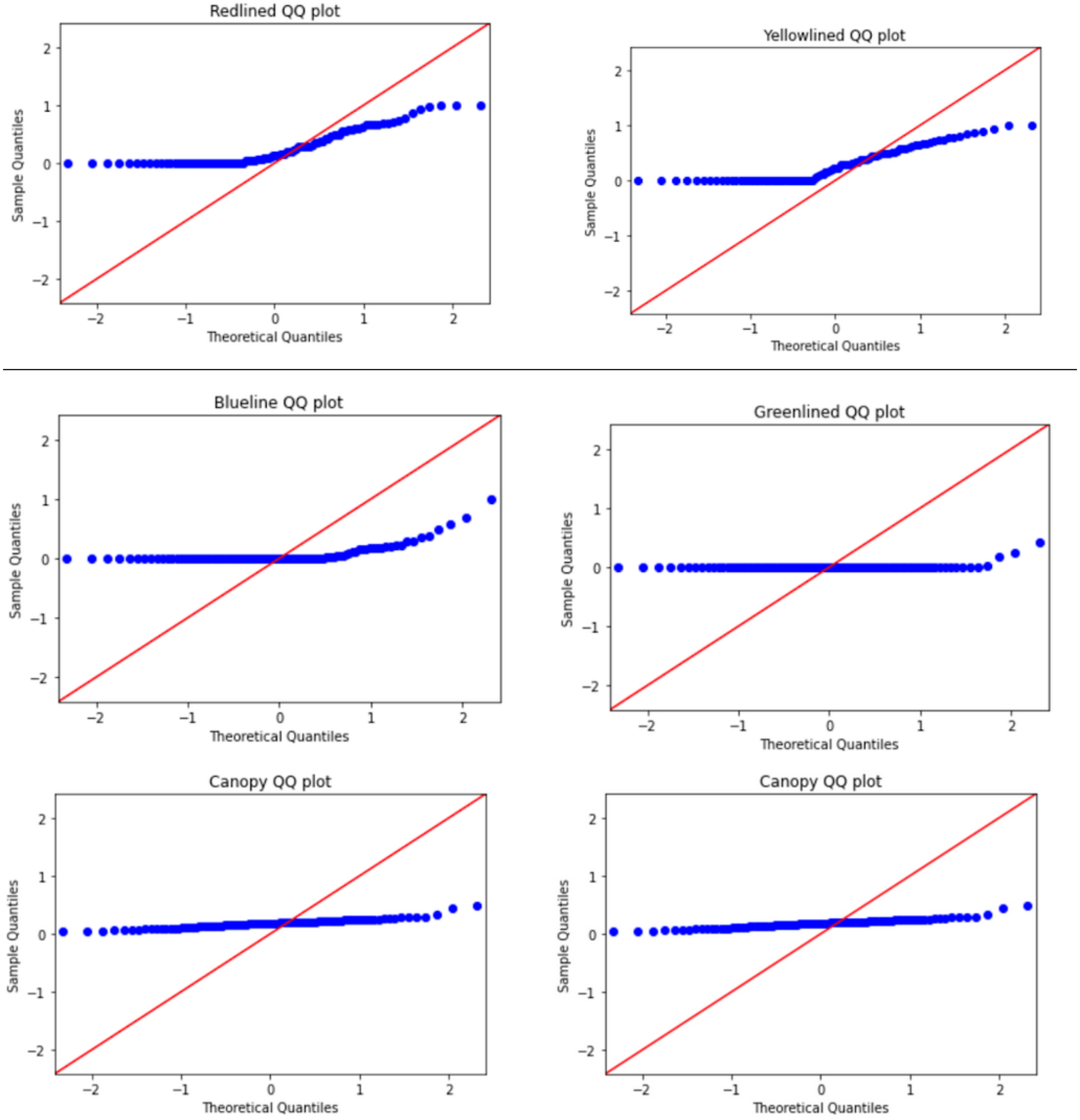


Table 4: Scatterplots Comparing HOLC Map Identification with Green Infrastructure Measures

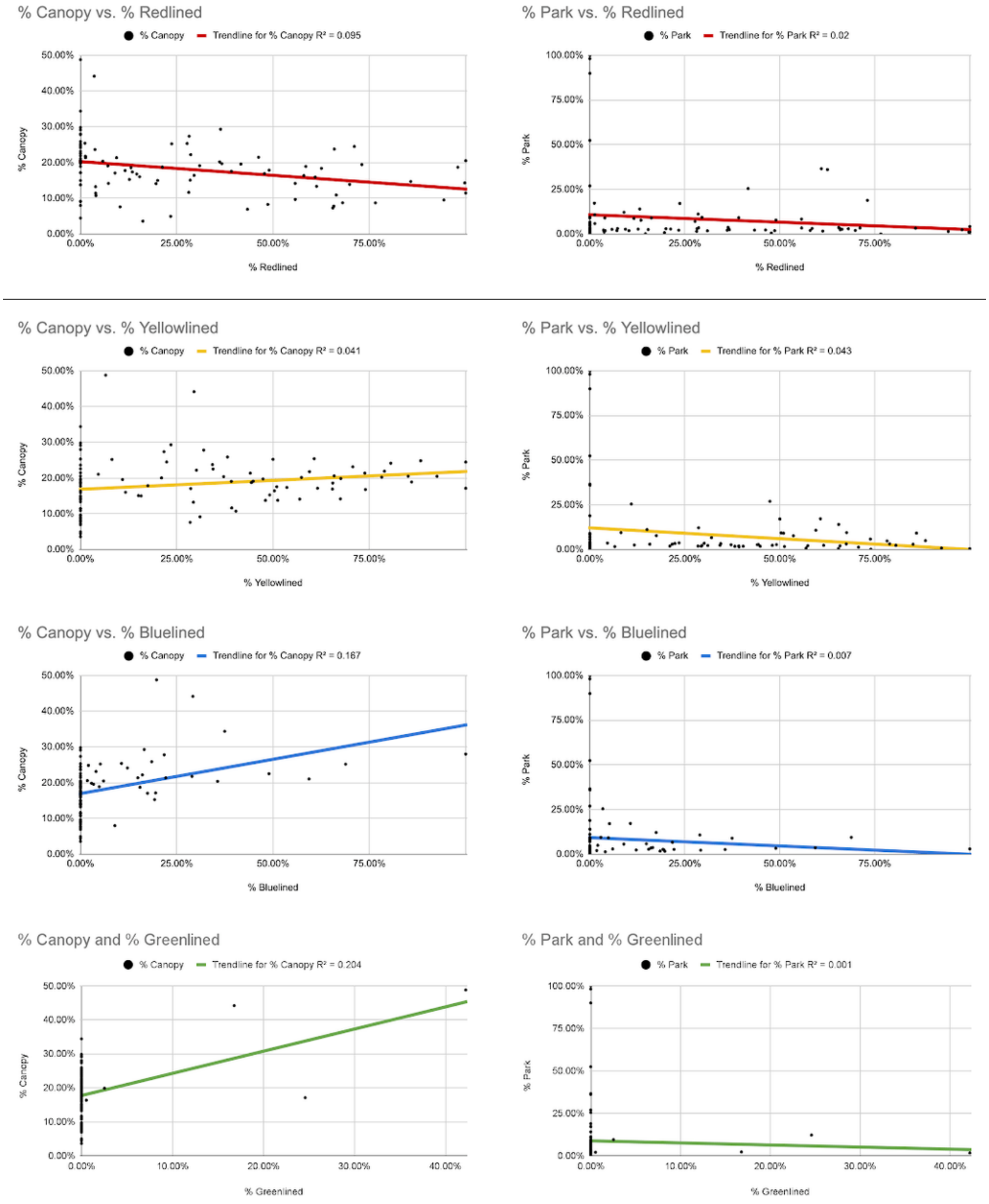


Table 5: K-means Clustering Results

Cluster 1 (28): Armour Square, Ashburn, Bridgeport, Bucktown, Burnside, Chinatown, Douglas, East Village, Englewood, Fuller Park, Grand Boulevard, Greektown, Lincoln Park, Little Italy/UIC, Lower West Side, Mckinley Park, Mount Greenwood, Near South Side, Oakland, Old Town, River North, Rush & Division, Sheffield & Depaul, South Chicago, United Center, Washington Park, West Town, Wicker Park

	% Redlined	% Yellowlined	% Bluelined	% Greenlined	% Park	% Canopy
Mean	67.07	3.03	0.12	0.02	6.61	14.81
STD	18.433	8.26	0.64	0.10	9.98	5.25
Min	36.71	0.00	0.00	0.00	0.00	6.92
Max	100.00	34.26	3.40	0.54	36.56	24.48

Cluster 2 (33): Archer Heights, Auburn Gresham, Avondale, Belmont Cragin, Beverly, Brighton Park, Calumet Heights, Chatham, Clearing, East Side, Galewood, Garfield Park, Garfield Ridge, Gold Cost, Grand Crossing, Humboldt Part, Kenwood, Lincoln Square, Little Village, Logan Square, Montclare, Morgan Park, New City, North Lawndale, Norwood Park, Pullman, Roseland, Uptown, Washington Heights, West Elsdon, West Lawn, West Pullman, Woodlawn

	% Redlined	% Yellowlined	% Bluelined	% Greenlined	% Park	% Canopy
Mean	15.53	38.18	8.06	1.25	4.72	18.90
STD	14.04	13.43	12.68	5.10	5.49	6.94
Min	0.00	11.70	0.00	0.00	0.52	7.60
Max	47.76	65.41	48.94	24.59	26.92	44.15

Cluster 3 (19): Albany Park, Andersonville, Austin, Boystown, Chicago Lawn, Dunning, Edgewater, Gage Park, Hermosa, Hyde Park, Irving Park, Jefferson Park, Lake View, North Center, Portage Park, Rogers Park, South Shore, Ukrainian Village, Wrigleyville

	% Redlined	% Yellowlined	% Bluelined	% Greenlined	% Park	% Canopy
Mean	2.03	76.64	5.67	0.13	4.99	20.60
STD	4.50	12.68	8.10	0.58	4.92	3.04
Min	0.00	59.48	0.00	0.00	0.11	14.16
Max	14.61	100.00	28.94	2.53	17.24	25.41

Cluster 4 (18): Avalon Park, Edison Park, Grant Park, Hegewisch, Jackson Park, Loop, Magnificent Mile, Millennium Park, Museum Campus, North Park, O'Hare, Printers Row, Riverdale, Sauganash/Forest Glen, South Deering, Streeterville, West Loop, West Ridge

	% Redlined	% Yellowlined	% Bluelined	% Greenlined	% Park	% Canopy
Mean	4.80	1.08	16.35	2.35	22.75	20.11
STD	8.82	2.55	30.02	9.95	35.69	12.03
Min	0.00	0.00	0.00	0.00	0.00	0.36
Max	27.68	8.18	100.00	42.23	100.00	48.78

Map 1: Clusters Map

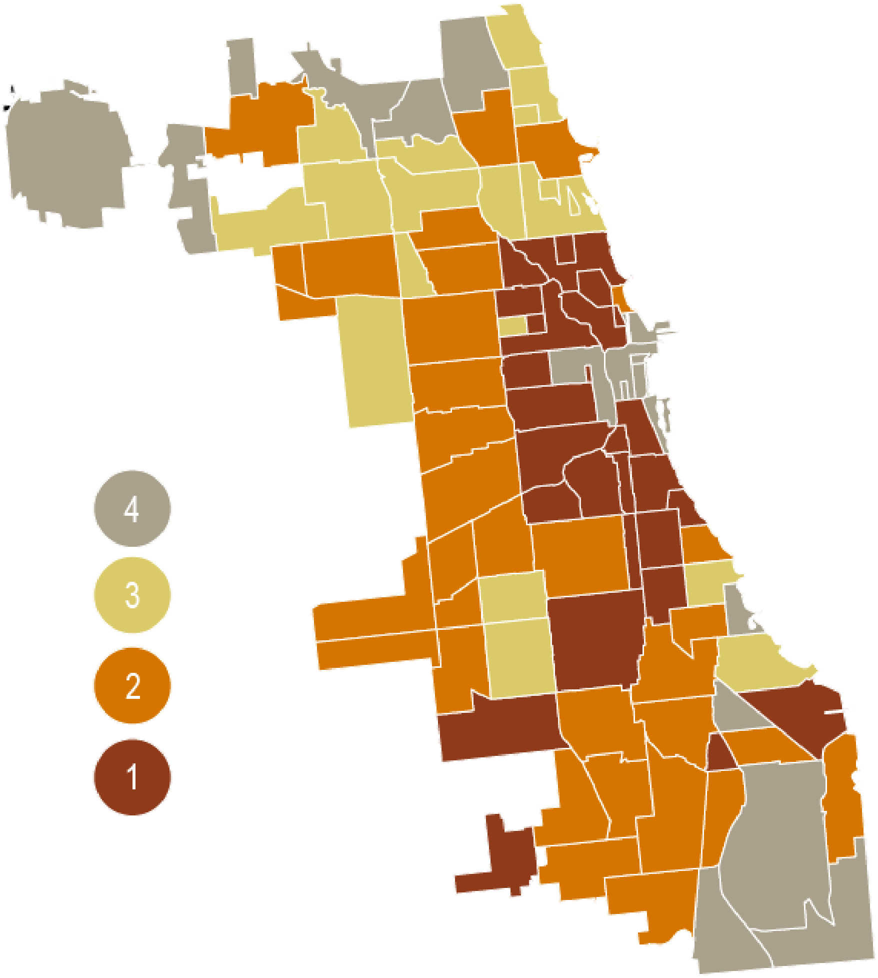
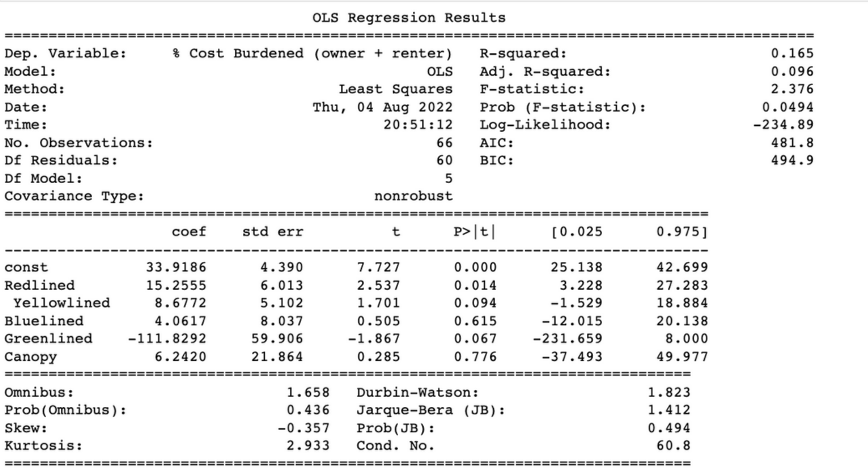


Figure 1: Clustering Map

Table 6: Linear Regression Test with New Datasets

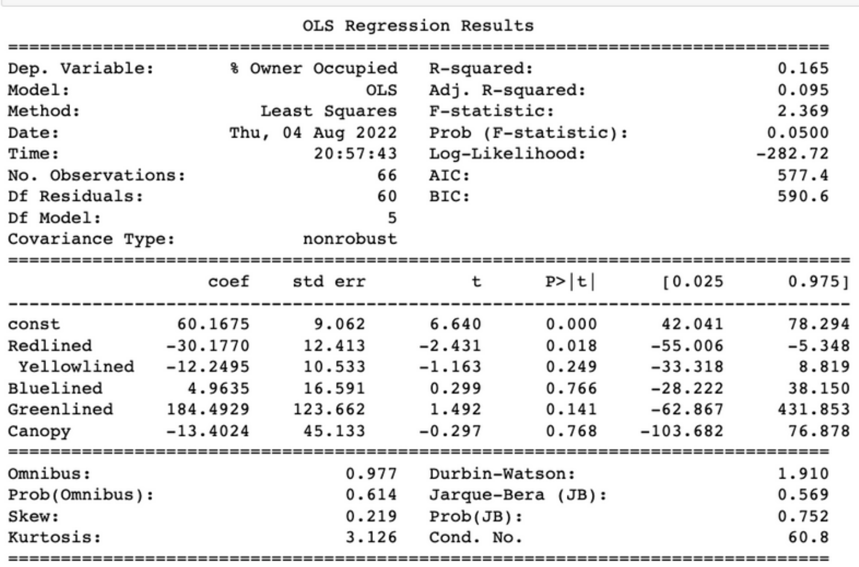
% of Households Cost Burdened



```
                                  OLS Regression Results
===========================================================================================
Dep. Variable:     % Cost Burdened (owner + renter)   R-squared:                      0.165
Model:                                          OLS   Adj. R-squared:                 0.096
Method:                               Least Squares   F-statistic:                    2.376
Date:                              Thu, 04 Aug 2022   Prob (F-statistic):            0.0494
Time:                                      20:51:12   Log-Likelihood:               -234.89
No. Observations:                                66   AIC:                            481.8
Df Residuals:                                    60   BIC:                            494.9
Df Model:                                         5
Covariance Type:                          nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const          33.9186      4.390      7.727      0.000      25.138      42.699
Redlined       15.2555      6.013      2.537      0.014       3.228      27.283
 Yellowlined    8.6772      5.102      1.701      0.094      -1.529      18.884
Bluelined       4.0617      8.037      0.505      0.615     -12.015      20.138
Greenlined   -111.8292     59.906     -1.867      0.067    -231.659       8.000
Canopy          6.2420     21.864      0.285      0.776     -37.493      49.977
==============================================================================
Omnibus:                        1.658   Durbin-Watson:                   1.823
Prob(Omnibus):                  0.436   Jarque-Bera (JB):                1.412
Skew:                          -0.357   Prob(JB):                        0.494
Kurtosis:                       2.933   Cond. No.                         60.8
==============================================================================
```

Figure 2: RT Cost Burden

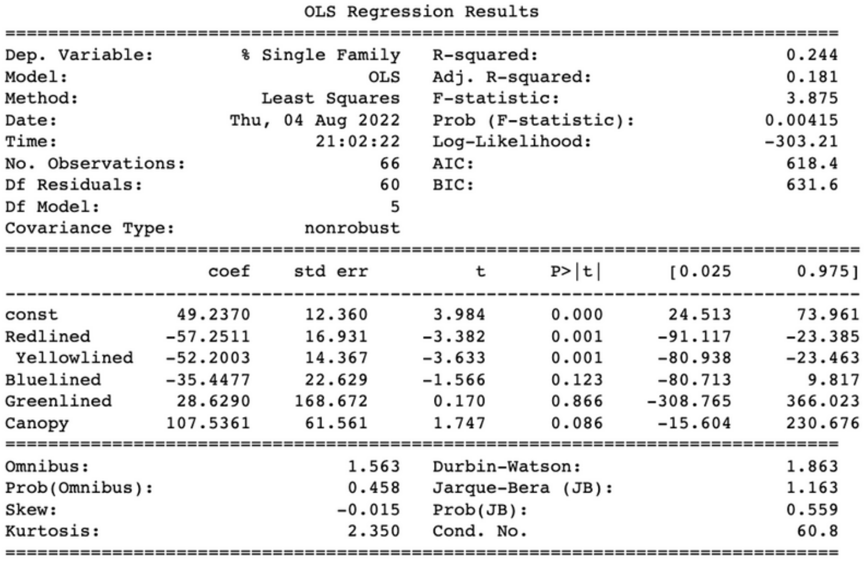
% of Households Owner Occupied



```
                            OLS Regression Results
==============================================================================
Dep. Variable:       % Owner Occupied   R-squared:                       0.165
Model:                            OLS   Adj. R-squared:                  0.095
Method:                 Least Squares   F-statistic:                     2.369
Date:                Thu, 04 Aug 2022   Prob (F-statistic):             0.0500
Time:                        20:57:43   Log-Likelihood:                -282.72
No. Observations:                  66   AIC:                             577.4
Df Residuals:                      60   BIC:                             590.6
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          60.1675      9.062      6.640      0.000      42.041      78.294
Redlined      -30.1770     12.413     -2.431      0.018     -55.006      -5.348
 Yellowlined  -12.2495     10.533     -1.163      0.249     -33.318       8.819
Bluelined       4.9635     16.591      0.299      0.766     -28.222      38.150
Greenlined    184.4929    123.662      1.492      0.141     -62.867     431.853
Canopy        -13.4024     45.133     -0.297      0.768    -103.682      76.878
==============================================================================
Omnibus:                        0.977   Durbin-Watson:                   1.910
Prob(Omnibus):                  0.614   Jarque-Bera (JB):                0.569
Skew:                           0.219   Prob(JB):                        0.752
Kurtosis:                       3.126   Cond. No.                         60.8
==============================================================================
```

Figure 3: RT Owner Occupied

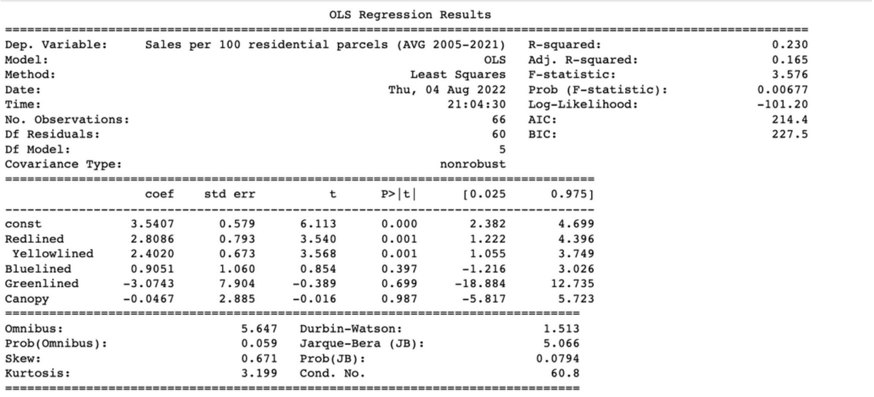
% of Housing as Single Family Units



```
                            OLS Regression Results
==============================================================================
Dep. Variable:        % Single Family   R-squared:                       0.244
Model:                            OLS   Adj. R-squared:                  0.181
Method:                 Least Squares   F-statistic:                     3.875
Date:                Thu, 04 Aug 2022   Prob (F-statistic):            0.00415
Time:                        21:02:22   Log-Likelihood:                -303.21
No. Observations:                  66   AIC:                             618.4
Df Residuals:                      60   BIC:                             631.6
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           49.2370     12.360      3.984      0.000      24.513      73.961
Redlined       -57.2511     16.931     -3.382      0.001     -91.117     -23.385
 Yellowlined   -52.2003     14.367     -3.633      0.001     -80.938     -23.463
Bluelined      -35.4477     22.629     -1.566      0.123     -80.713       9.817
Greenlined      28.6290    168.672      0.170      0.866    -308.765     366.023
Canopy         107.5361     61.561      1.747      0.086     -15.604     230.676
==============================================================================
Omnibus:                        1.563   Durbin-Watson:                   1.863
Prob(Omnibus):                  0.458   Jarque-Bera (JB):                1.163
Skew:                          -0.015   Prob(JB):                        0.559
Kurtosis:                       2.350   Cond. No.                         60.8
==============================================================================
```

Figure 4: RT Single Family

Sales per 100 Residential Parcels



```
                                         OLS Regression Results
=======================================================================================================
Dep. Variable:     Sales per 100 residential parcels (AVG 2005-2021)   R-squared:                 0.230
Model:                                                          OLS   Adj. R-squared:            0.165
Method:                                               Least Squares   F-statistic:               3.576
Date:                                              Thu, 04 Aug 2022   Prob (F-statistic):      0.00677
Time:                                                      21:04:30   Log-Likelihood:          -101.20
No. Observations:                                                66   AIC:                       214.4
Df Residuals:                                                    60   BIC:                       227.5
Df Model:                                                         5
Covariance Type:                                          nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           3.5407      0.579      6.113      0.000       2.382       4.699
Redlined        2.8086      0.793      3.540      0.001       1.222       4.396
 Yellowlined    2.4020      0.673      3.568      0.001       1.055       3.749
Bluelined       0.9051      1.060      0.854      0.397      -1.216       3.026
Greenlined     -3.0743      7.904     -0.389      0.699     -18.884      12.735
Canopy         -0.0467      2.885     -0.016      0.987      -5.817       5.723
==============================================================================
Omnibus:                        5.647   Durbin-Watson:                   1.513
Prob(Omnibus):                  0.059   Jarque-Bera (JB):                5.066
Skew:                           0.671   Prob(JB):                       0.0794
Kurtosis:                       3.199   Cond. No.                         60.8
==============================================================================
```

Figure 5: RT Sales

Guide 1: Github Manual

As uploaded on Github, "data2.csv" corresponds to redlining and canopy information regarding each neighborhood in the map and "dataU.csv" corresponds to other features we gathered from different sources corresponding to the built environment. Each notebook can be run by locally downloading the datasets and changing the path of the file in each notebook to reproduce the results. The clusterings should be applied with the same number of clusters to get the same result as in the report and paper.

Pixels.ipynb: In this notebook you will find the function that takes the .png file as input and computes the area covered by the tree canopy in the given neighborhood divided by the total area of the image, which is the total area of the neighborhood . It's important to work with the .png files because the boundary of the image collides with the white, default blank surface of the canvas in .jpg format, whereas in .png images the boundary of the image taken is separate to the canvas.

CanopyAnalyze.ipynb: First, in order to load and run the notebook locally, remember to change the path to the datasets accordingly. In this notebook, we went through the basic statistics including mean, median, standard deviation and quarter percentiles for each feature and plotted the distribution (histogram) of each feature independently.

The correlation matrix among the dataset features can be found where our main variable of interest (Canopy) had been plotted against the highest correlation ("Yellowlined" and "Redlined") given by the matrix.

We applied the PCA method to the main dataset to reduce the dimensionality for plotting. The elbow rule had been leveraged to determine the best number of clusters; using sum of squared errors, we get k=4. Clustering took place on the dataset after applying PCA using 4 clusters and has been plotted in a 2-D plane as a scatterplot using different colors to represent distinct clusters. After the initial clustering, there are some basic statistics (mean, median, std and percentiles) of each cluster themselves.

Clustering_ML.ipynb: This notebook focuses on three main subjects:

Test of Normality of the features: We plotted and leveraged the QQ plots of each features to have a better sense of the normality of each feature (i.e can we assume that each feature distribution among the different neighborhoods have a normal Gaussian distribution or not).

Visualizing the clusters in a map of Chicago: You can find the different maps corresponding to clustering result. After applying clustering, we colored the neighborhood with the corresponding cluster color. It's worth noting that in our clustering implementation, we didn't use any feature of proximity, only the datasets on redlining and green infrastructure.

Regression/ Prediction Power: After obtaining the clusters, using the clusters label and the features we had already, we evaluated the prediction power of our labels with respect to other features in the built environment related to energy and housing.

In our dataset, since our sample size is too small for a machine learning setup (specially testing the prediction power of our clusters) we didn't use the classic cross-fold validation and used the dataset as a whole. To reproduce the result of linear regressions, testing each of the target variables with respect to the other variables, simply change the target variable to the desired one and fit the regression model created to the data obtained.