# 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.