A Socio-economic Analysis of Chicago

The Value of Predicting Socio-economic Status

- Machine learning is being leveraged to analyze the effects of economic and financial changes on the well-being of a population, and vice versa.
- Modeling the relationship can help create and maintain a fair social structure, along with a happy citizenry.
- Study focus:
  - Predict the hardship index of a community based on the interplay of a number of factors, including per capita income, poverty level, and housing situation.
  - Guide city authorities by suggesting locations to open new parks.
- Governments: use it to guide policy formulation; businesses: use it to enter markets that would benefit from their presence; citizens: use it to understand their community better.

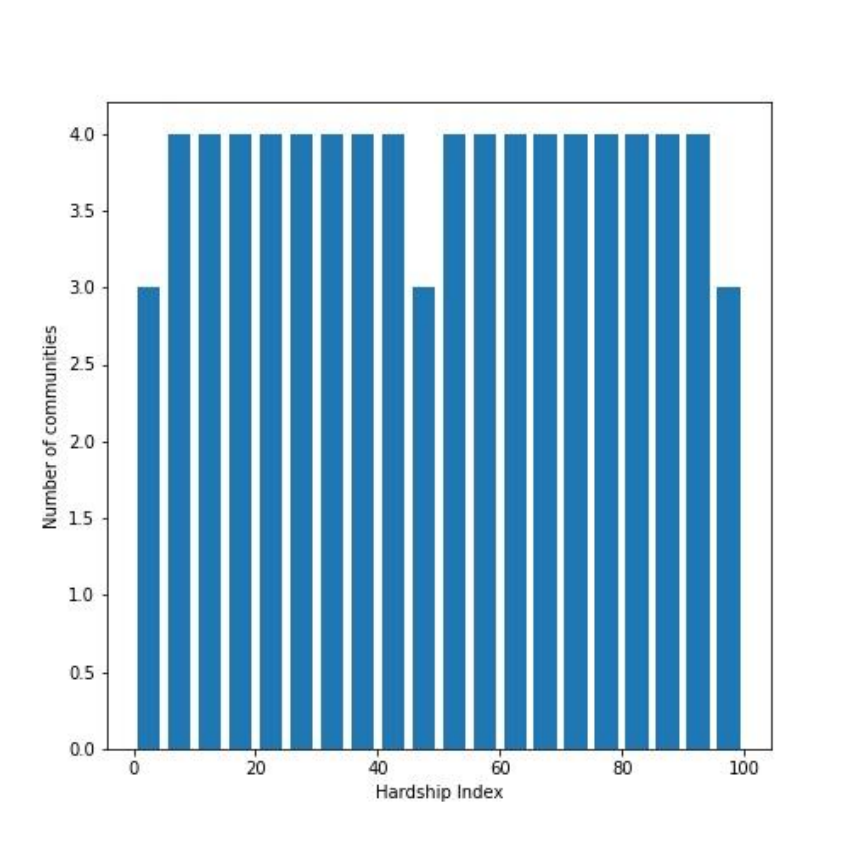
The Relationship Between the Hardship Index of a Community and Socio-economic Factors

Data Acquisition & Cleaning

- Data sources:
  - Census data for Chicago between 2008-2012, obtained from the city's data portal.
  - Crime data for the city between 2008-2012, obtained from Kaggle.
  - Chicago public school data for the academic year 2011-2012, obtained from the city's data portal.
- Each datasets cleaned, relevant features selected, then combined.
- Final dataset contained 10 features.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

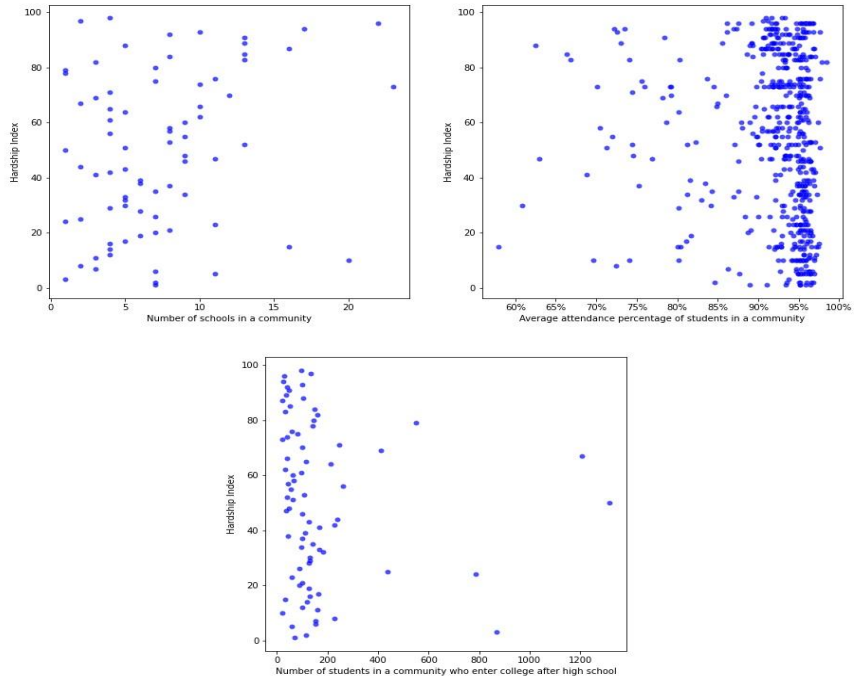
Exploratory Data Analysis & Visualization



- Histogram of hardship index not very informative.
- Even spread of the metric over city's communities.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

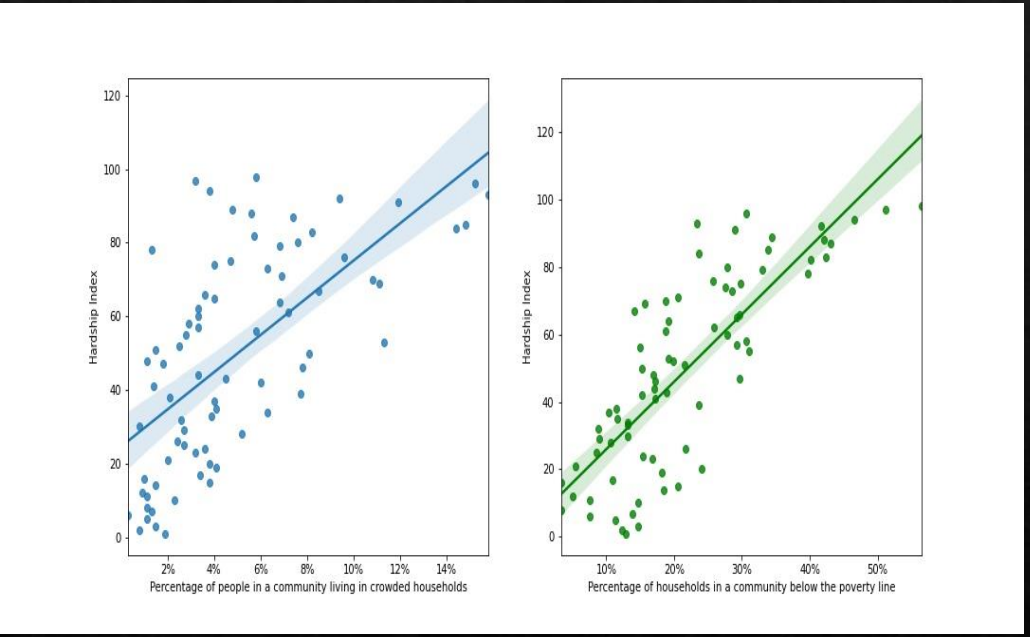
Exploratory Data Analysis & Visualization



- None of the factors from the public school data have an impact on the hardship index.
- Pattern may emerge if population numbers for each community could be factored in, but information not available.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

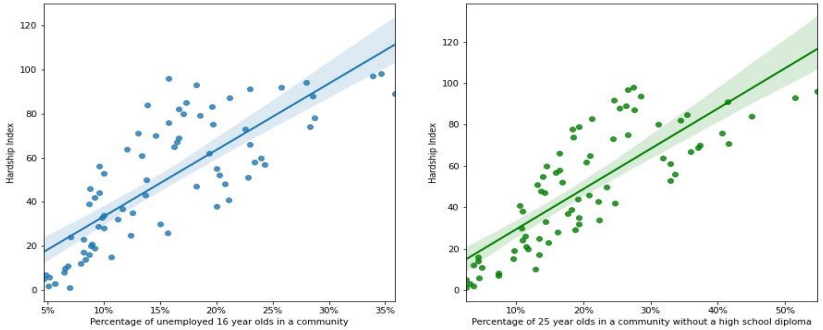
Exploratory Data Analysis & Visualization



- Large variation of hardship index even if a community has a low percentage of people living in cramped houses.
- Communities having a relatively large number of households below the poverty line can still have a hardship index close to zero.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

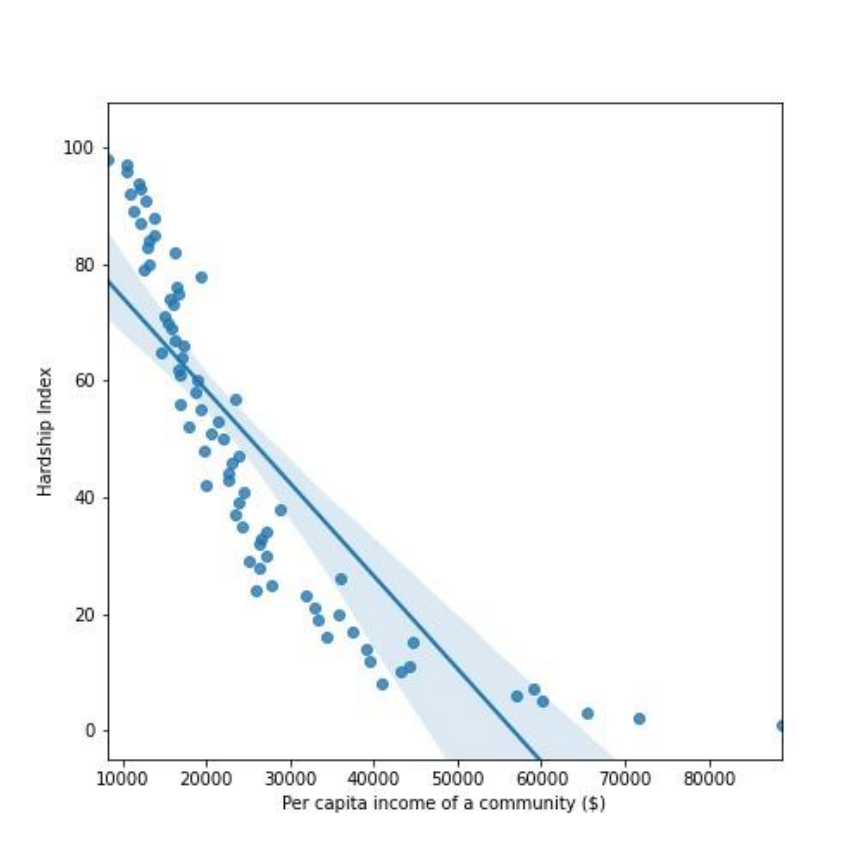
Exploratory Data Analysis & Visualization



- Absence of a job for young adults in high school (16 year olds) can strain a community.
- Importance of a high school-level education, at minimum, is seen in the top-right graph.
- Once the percentage of people below 18 or above 64 years of age goes above 30%, the potential for hardship increases dramatically, showcasing the importance of having a substantial ratio of members in the workforce.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

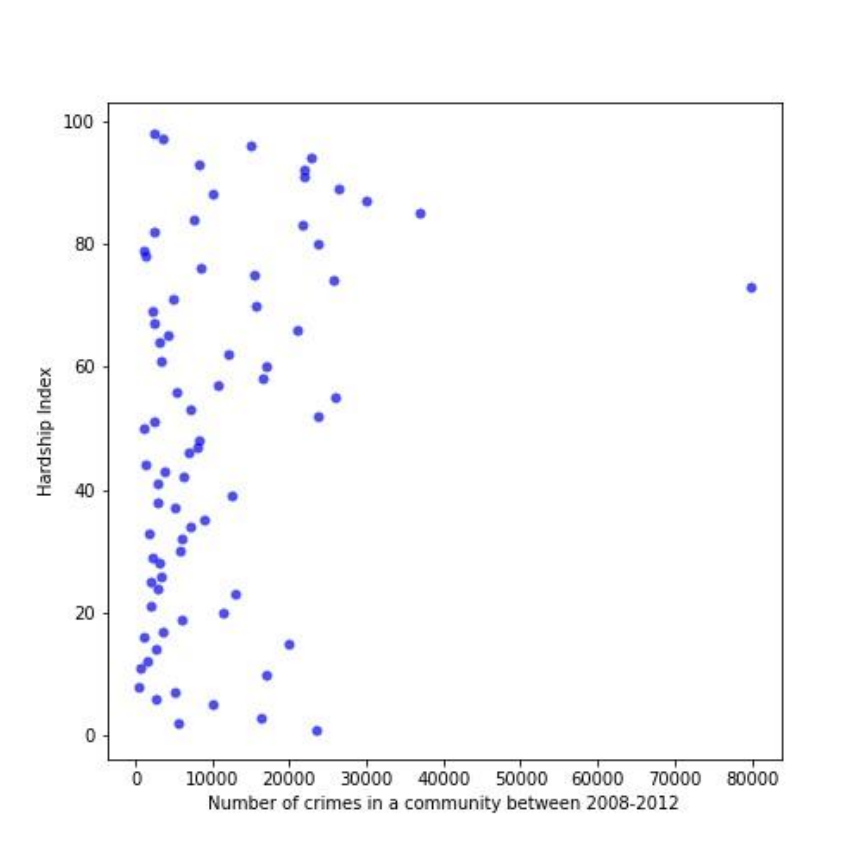
Exploratory Data Analysis & Visualization



- A steep drop of the hardship index is observed with increasing per capita income.
- When it becomes more than about 60,000 $, the index becomes effectively zero.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

Exploratory Data Analysis & Visualization



- Surprisingly, (cumulative) number of crimes does not seem to have an impact on the hardship.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

Predictive Modeling & Analysis

- Variables to use for machine learning picked: percent of people living in crowded households, percent of households living in poverty, percent of unemployed 16 year olds, percent of 25 year olds without a high school diploma, percent of community residents under 18 or over 64 years of age, and per capita income.
- Three models tested: multiple linear regression (with 6 independent variables), polynomial regression of degrees 2 & 3.
- Models evaluated using k-fold cross-validation and R^2 metric.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

Predictive Modeling & Analysis

	Multiple Linear Regression	Polynomial Regression; n=2	Polynomial Regression; n=3
Fold 1	0.9751	0.9566	0.8675
Fold 2	0.9788	0.9171	0.8019
Fold 3	0.9653	0.9280	0.9577
Fold 4	0.9775	0.9434	0.9625
Fold 5	0.9497	0.7903	0.9768
Average	0.9693	0.9071	0.9133

Good score obtained with the multiple linear regression model, whereas R^2 shows quite sizeable fluctuations depending on the slice of data used for a second degree polynomial model. The same problem is observed in the third degree polynomial model.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

Predictive Modeling & Analysis

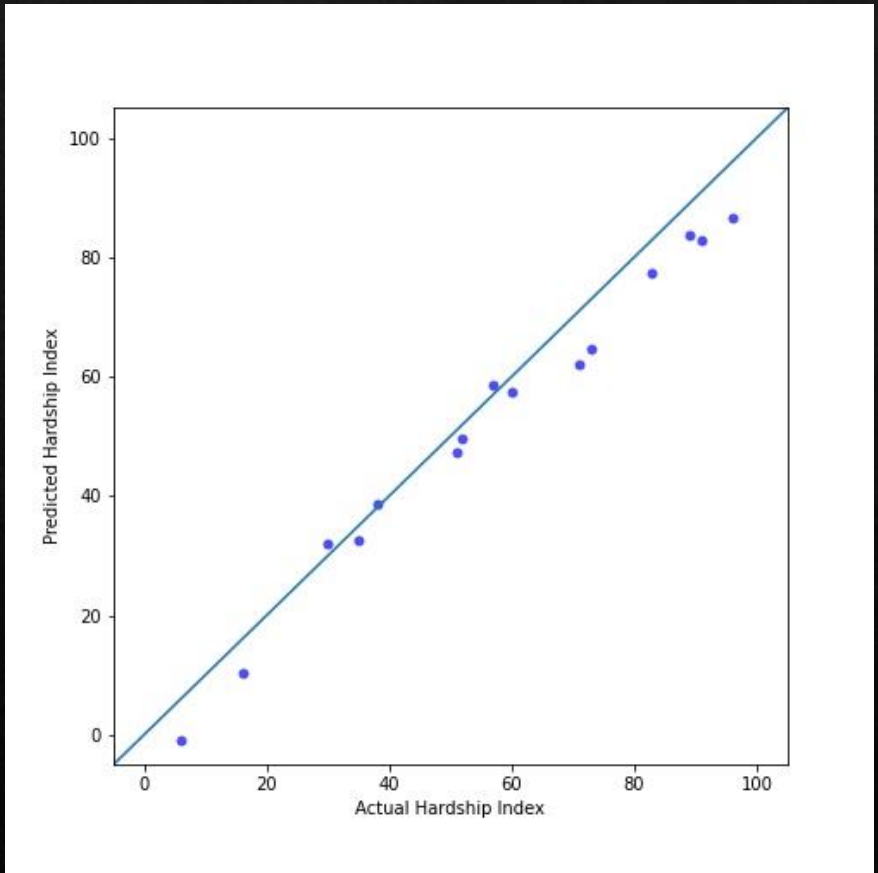
Multiple linear regression model was chosen, and the full training set was fed to the model, which returned the following intercept and coefficients:

Intercept	Percent of crowded housing	Percent of households below poverty	Percent of 16 year olds w/o job	Percent of 25 year olds w/o high school diploma	Percent aged below 18 or above 64	Per capita income
48.5081	2.8012	8.9105	6.4735	10.4676	5.18867	-1.9559

Using the model with these values on the test data gave an R^2 score of **0.95322**.

The Relationship Between the Hardship Index of a Community and Socio-economic Factors

Predictive Modeling & Analysis



- The straight light blue diagonal line represents a perfect fit.
- It seems like the model consistently underpredicts the data. This is likely a consequence of the small number of data points we are working with.

Predicting Useful Locations of Public Parks in the City

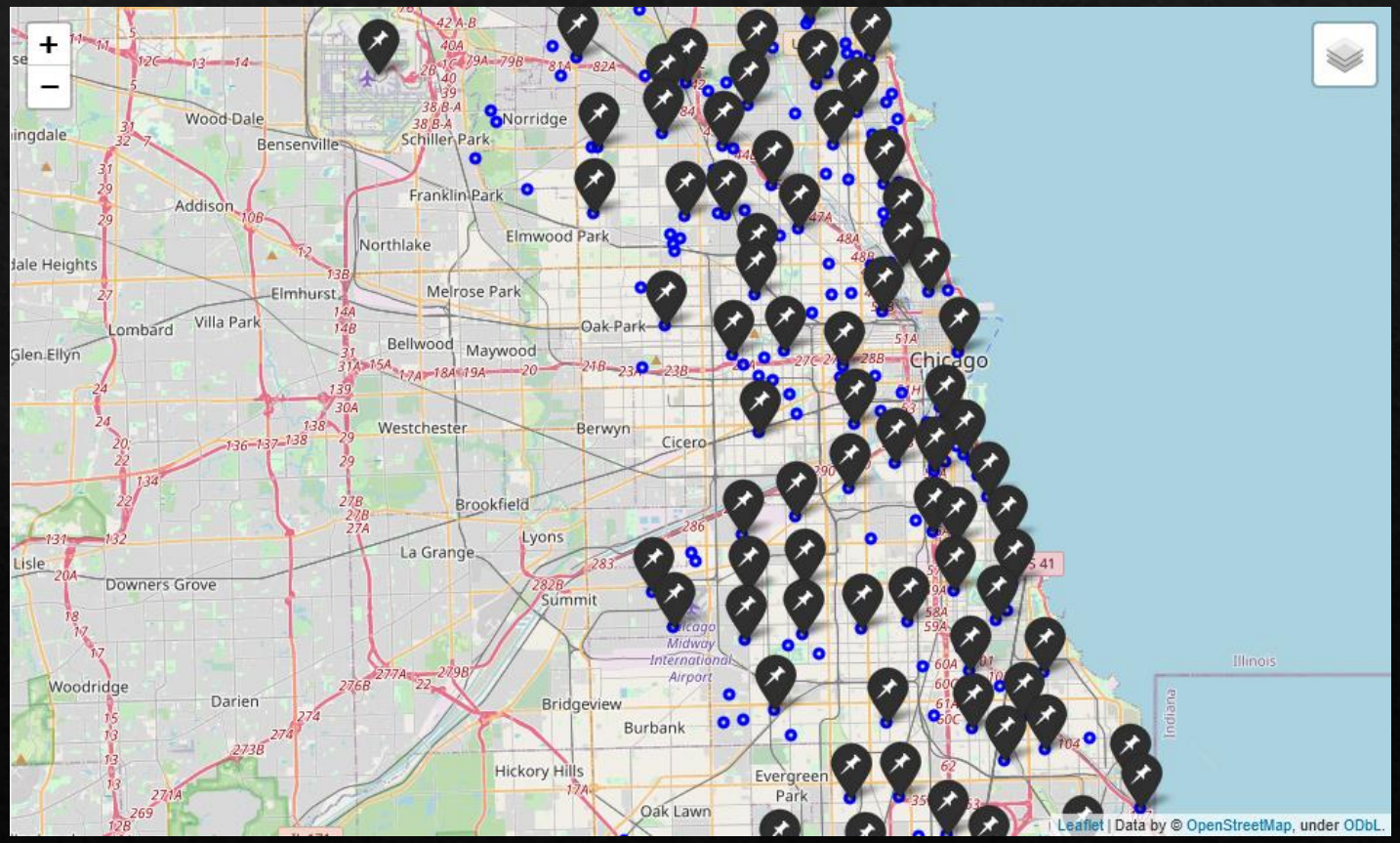
Data Acquisition & Cleaning

- Data sources:
  - The list of neighbourhoods in Chicago by community area was scraped from Wikipedia.
  - The latitude and longitude of each neighbourhood was then retrieved using the `geocoder` module.
  - The next step was using the Foursquare API to retrieve some information about parks located around the neighbourhoods of Chicago.
- As before, the dataset was cleaned, and relevant features selected.

Predicting Useful Locations of Public Parks in the City

Exploratory Data Analysis & Visualization

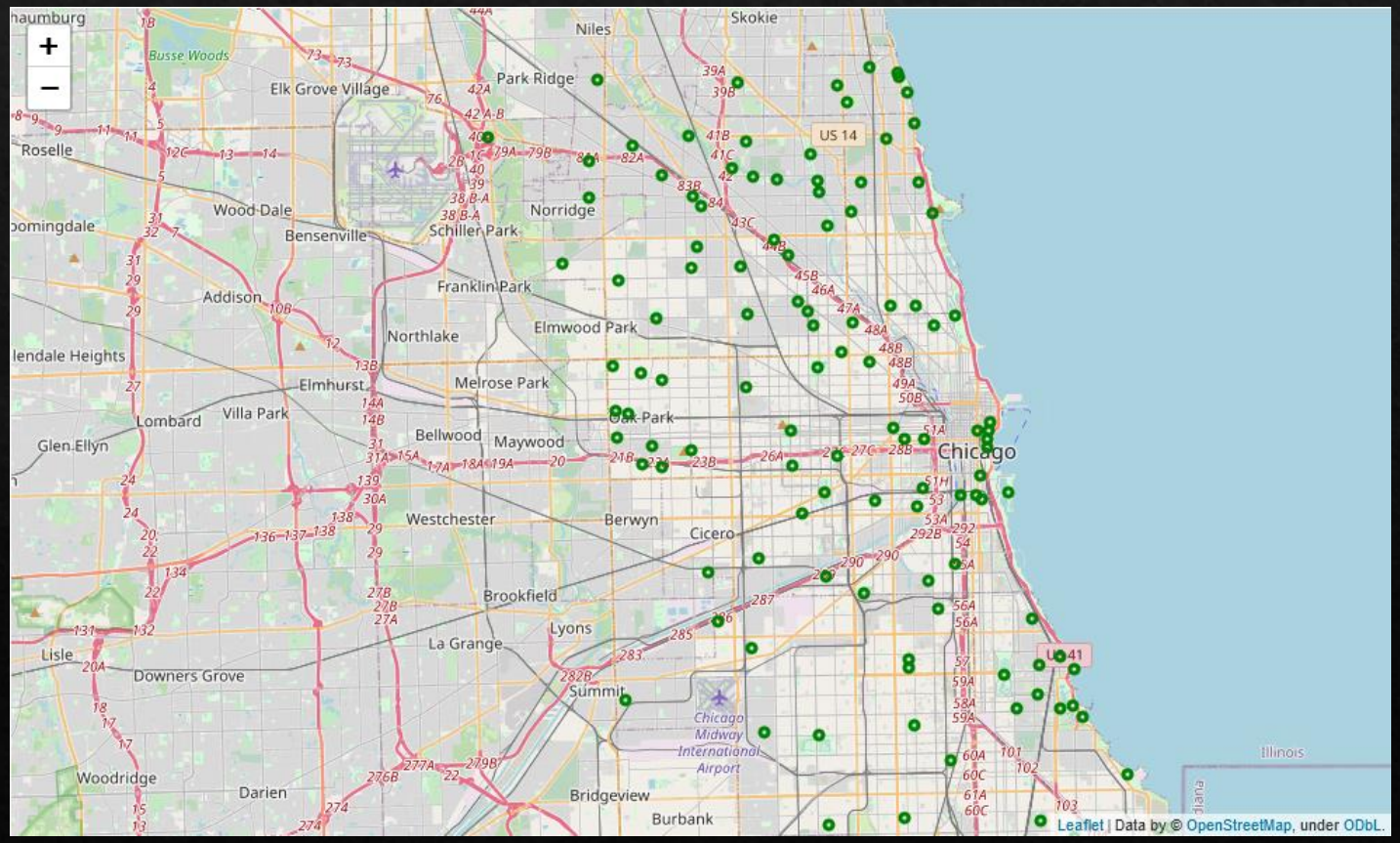
Map of Chicago, its community areas (black pins), and its neighbourhoods (blue circles).



Predicting Useful Locations of Public Parks in the City

Exploratory Data Analysis & Visualization

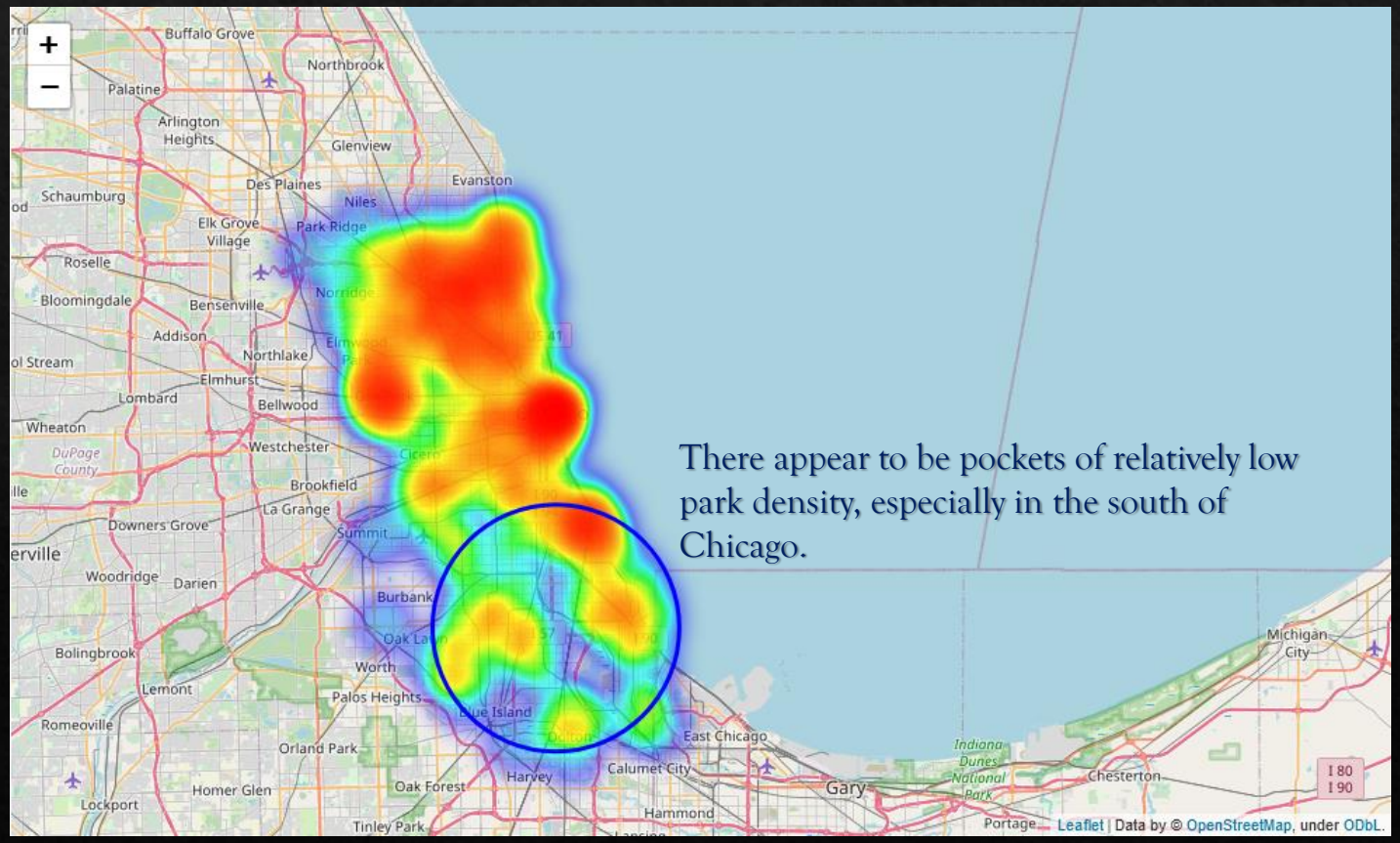
Map of public parks in Chicago (green circles).



Predicting Useful Locations of Public Parks in the City

Exploratory Data Analysis & Visualization

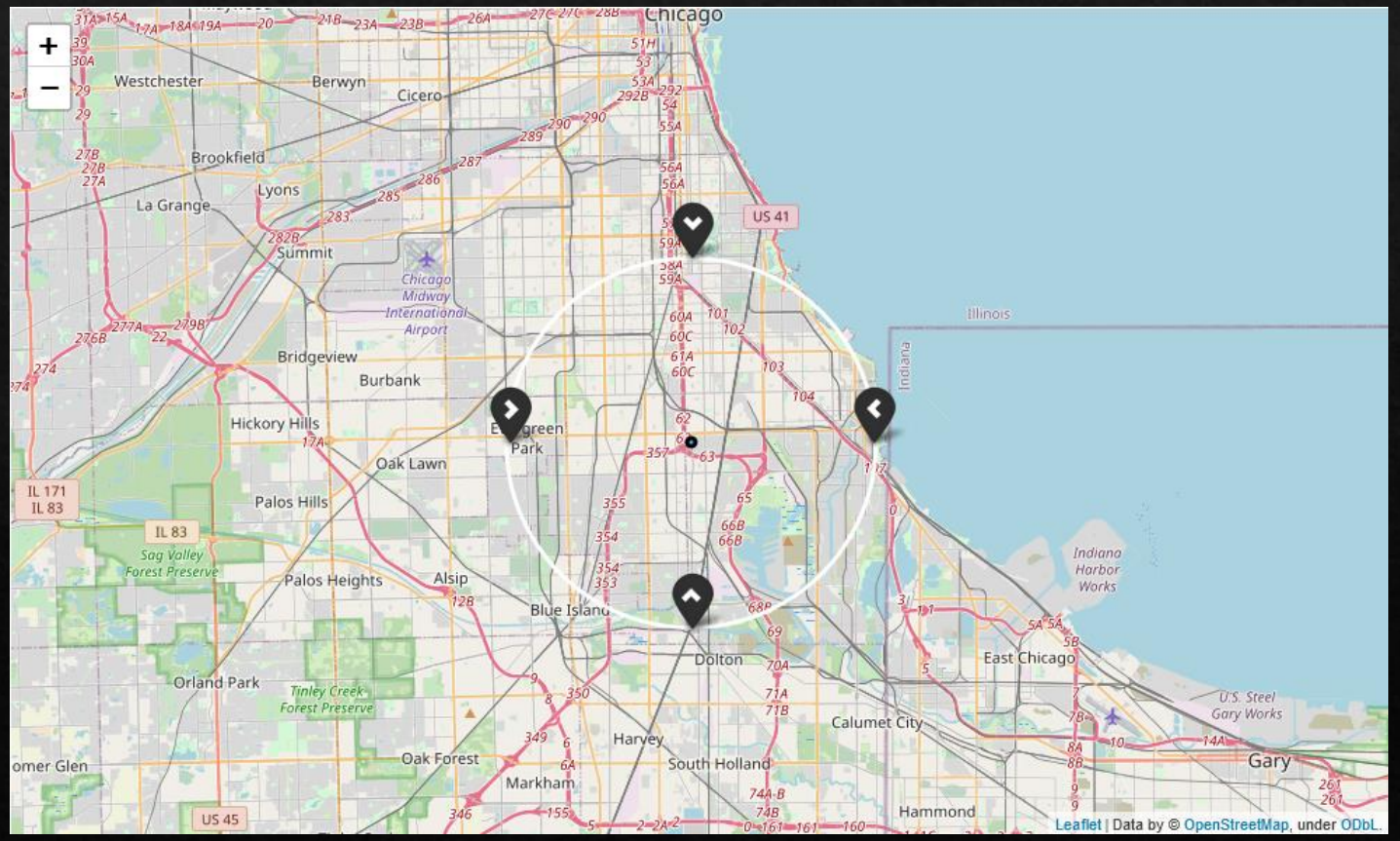
Heatmap of public parks in Chicago.



Predicting Useful Locations of Public Parks in the City

Machine Learning Modeling & Analysis

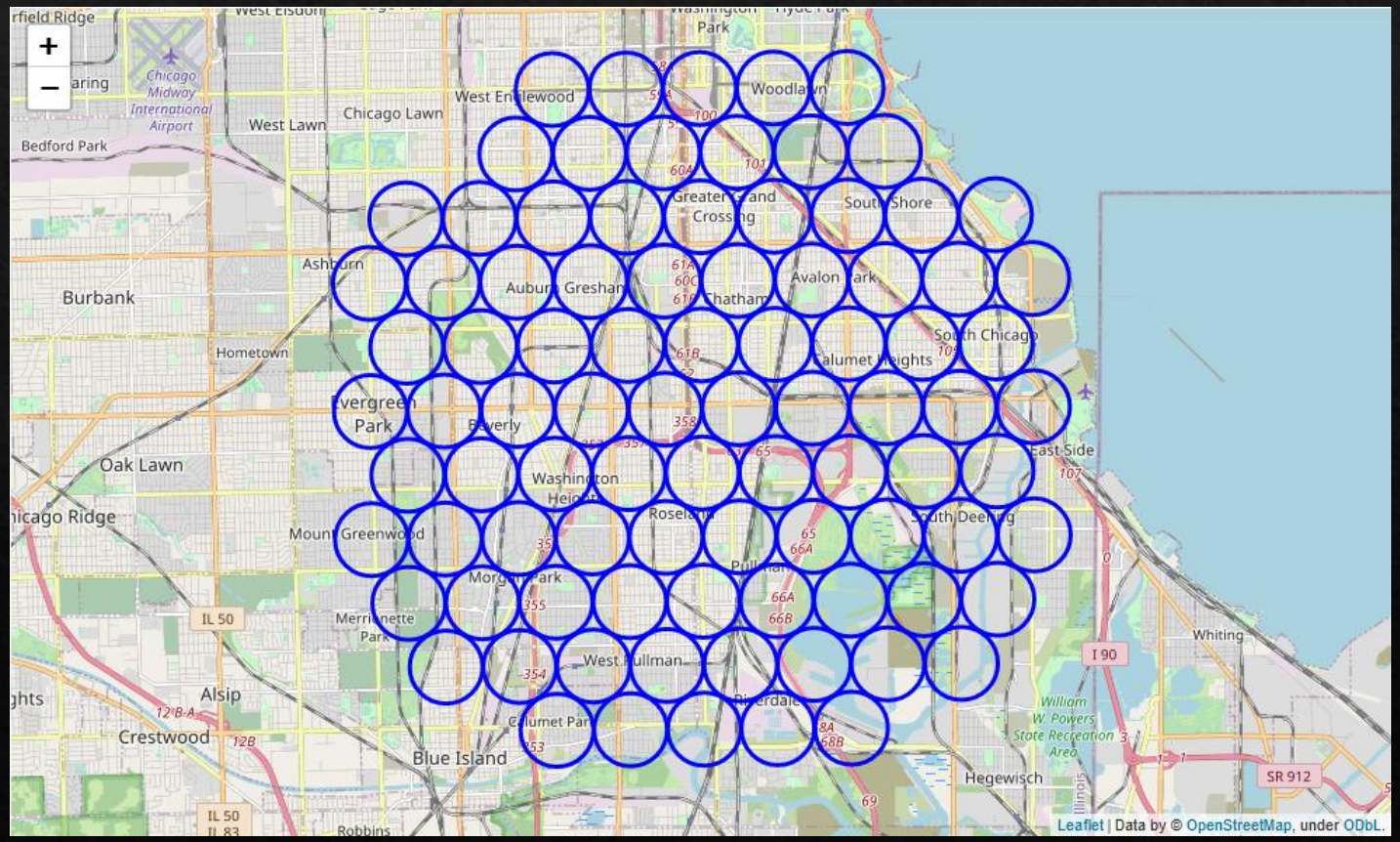
Harlan Community Academy HS was picked as central point of search, and a 7.5 km radius around it was examined.



Predicting Useful Locations of Public Parks in the City

Machine Learning Modeling & Analysis

To obtain candidate locations, a hexagonal grid was created in this region of interest.



Predicting Useful Locations of Public Parks in the City

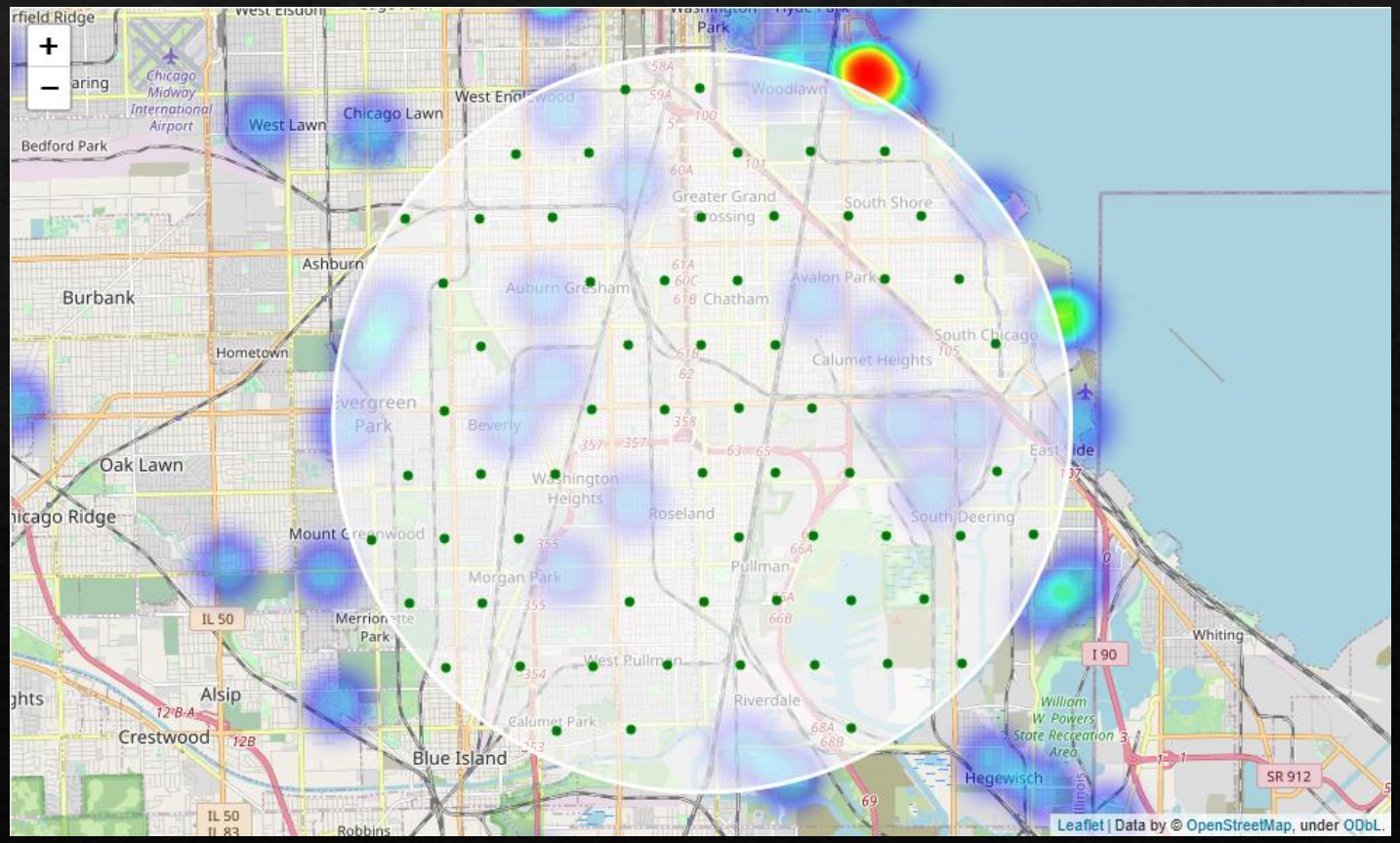
Machine Learning Modeling & Analysis

- The approach taken was to count the number of parks in the vicinity of each grid cell, as well the distance to the closest park, then check for cells containing few parks nearby, with the closest park, well, not so close by.
- The average number of parks (2.42) & average distance to a park (1.18 km) in a community was calculated.
- The filtering criteria were thus set: areas with 2 parks or less in their vicinity, and where the distance to the closest park is more than 1 km.
- 88 areas were found with 2 parks or less in their vicinity, and 63 were found where the nearest park is more than 1 km away. The number of areas satisfying both conditions was also 63; these were the same areas where the parks were more than 1 km away.

Predicting Useful Locations of Public Parks in the City

Machine Learning Modeling & Analysis

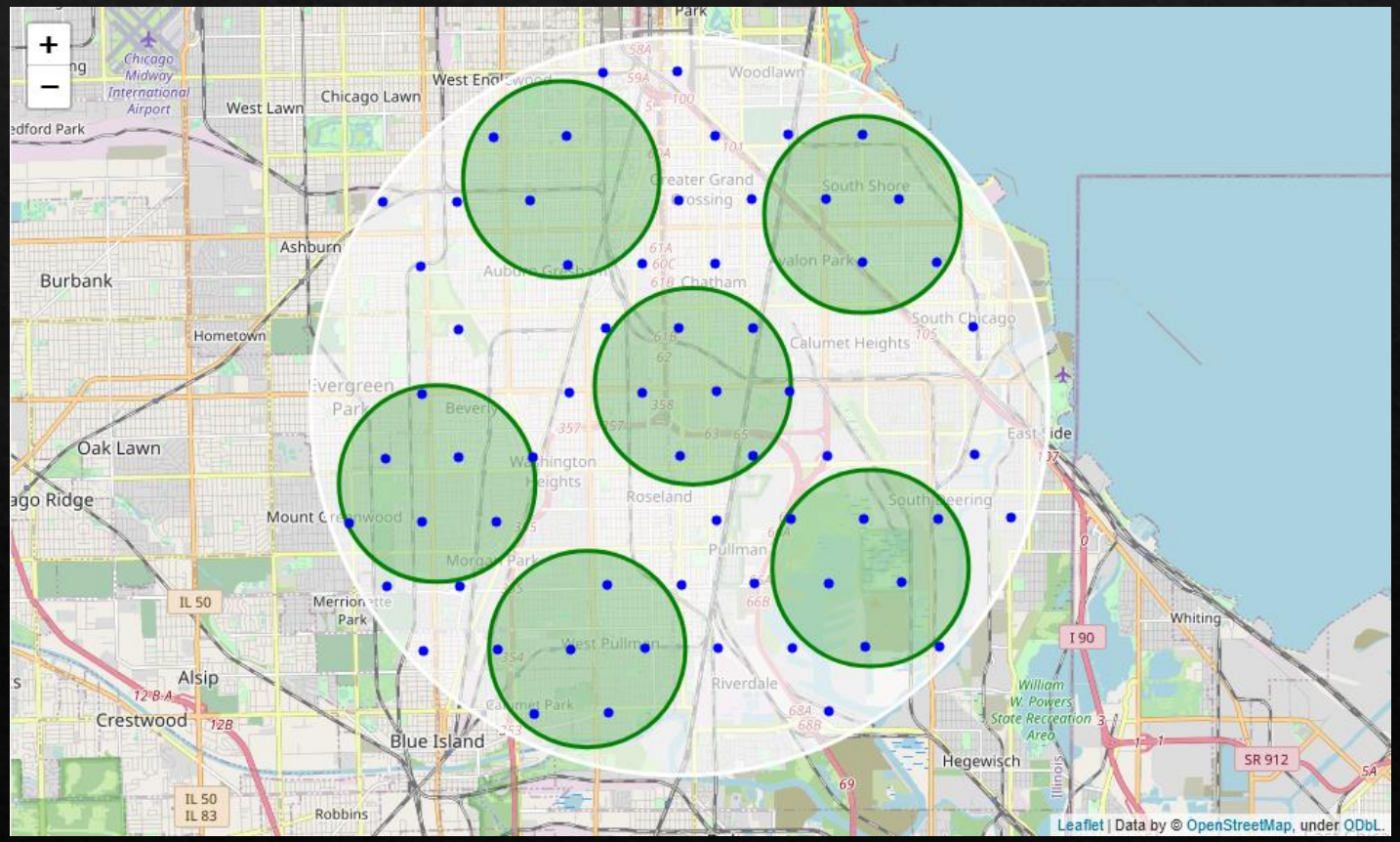
Candidate locations satisfying the criteria set in the previous slide.



Predicting Useful Locations of Public Parks in the City

Machine Learning Modeling & Analysis

k-means clustering was then performed on candidate locations to focus the search; 6 clusters were found.



Predicting Useful Locations of Public Parks in the City

Machine Learning Modeling & Analysis

- Finally, the addresses of the cluster centers were reverse geocoded, giving approximate locations to begin further investigation:
  - 10411 South Claremont Avenue, Beverly, Chicago, Illinois 60643
  - 1949-1951 East 78th Street, South Shore, Chicago, Illinois 60649
  - South Green Street, West Pullman, Chicago, Illinois 60827
  - 234 East 95th Street, Roseland, Chicago, Illinois 60617
  - 7359 South Carpenter Street, Englewood, Chicago, Illinois 60621
  - 11555 South Stony Island Avenue, South Deering, Chicago, Illinois 60617

Conclusions

- This study had two aims:
  - analyze the socio-economic factors that affect the well-being of a community's members, developing a predictive model in the process
  - find areas lacking in public parks, enabling municipal powers to be proactive in providing this proven beneficial resource to their constituencies
- The result of pursuing the first aim, a linear regression model that can predict the hardship index of a community based on socio-economic statistics, can be very useful to government, for example, in order to improve their policies so that communities provide a healthy, nurturing environment for its residents.
- Pursuing the second aim led to an algorithmic way to find areas lacking a particular resource, providing a great starting point for action. Although parks in Chicago were used in this study, the methodology can be utilized for any public good.