Final Report

Anish Shah

Understanding the Relationship between Tree Canopy and Crimes in Buffalo, New York.

Buffalo Sewer Authority

Supervisor: Kevin Meindl
kmeindl@buffalosewer.org

Abstract

The goal of this project is to understand the relationship between the tree canopy and crime in Buffalo, New York using geospatial data analysis. Various recent findings in the literature suggest that neighborhood characteristics such as trees and other vegetation are inversely associated with crime. The extent to which urban tree cover influences crime is a debate in the literature. So, with the help of this research taking advantage of the geocoded crime point data and high-resolution tree canopy data is used to address this problem in Buffalo City. As cities are moving towards 'green' growth plans and must look to incorporate sustainable methods of crime prevention into city planning our research project has implications for the urban planning policy.

Introduction

Having abundant tree canopy in municipalities is very valuable. Many studies have concluded that tree canopy reduces the costs of cooling a home, slows the formation of urban smog, and promotes a general sense of wellness and tranquility. Some literature also suggests that it reduces the feelings of danger among people.

More recent additions to the literature, with the use of remote sensing and GIS to observe larger sample populations with more quantitative methods on the relationship between tree canopy and crime, have drawn conclusions similar to those of Kuo and Sullivan (2001).

The belief that tree canopy correlates negatively with crime has been affirmed to varying degrees in studies covering Austin, Texas (Snelgrove, Michael, Waliczek, and Zajicek, 2004), Baltimore County, Maryland (Troy, Grove, and O'Neil-Dunne, 2012), Philadelphia, Pennsylvania (Wolfe and Mennis, 2012) and Portland, Oregon (Donovan and Prestemon, 2012). The goal of this study was to further explore the relationship within the geography of the city of Buffalo, New York.

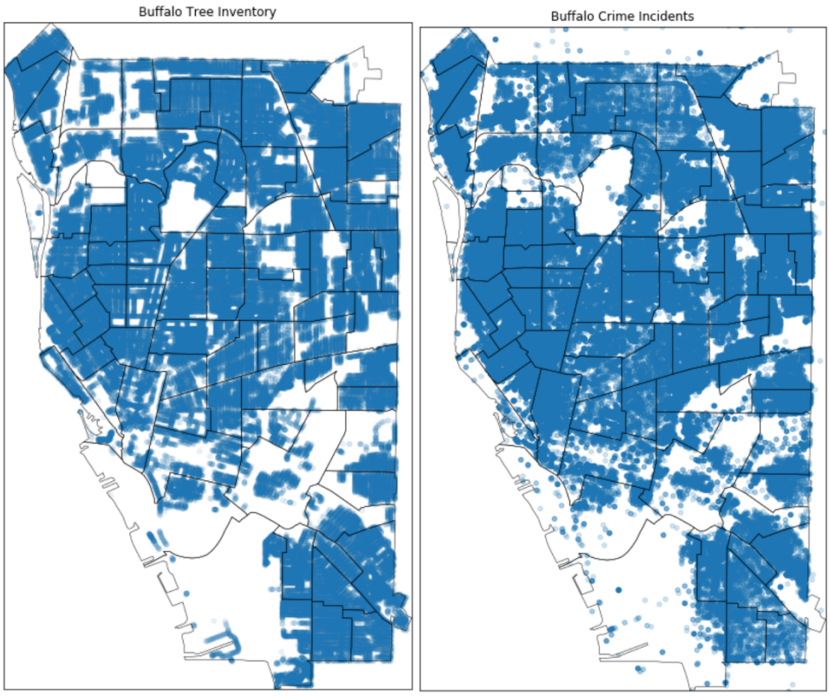

Figure 1. Individual Trees and Crimes in Buffalo

Methods

Study Area

The area and units of observation used in this study were the 35 of the neighborhoods that make up the city of Buffalo.

Figure 2: Buffalo neighborhood boundaries

The 35 neighborhoods occupy an area of 40.84 square miles and are home to a population of 258,989 people.

Data

All of the data was acquired from the Buffalo Open Data portal and ArcGIS spatial statistic tools.

Tree Data

The ‘Tree Inventory’ dataset was used in order to study tree canopy. This data was high dimensional but uncleaned and needed a lot of processing. Total Leaf Surface Area for each neighborhood was calculated. Percentages of canopy cover were calculated for the land area of the neighborhoods.

Tree inventory dataset contains both street trees and park trees (identified by the column 'Editing' where BUFFALO TREE = street tree and 'OLMSTEAD PARKS' is an Olmstead park tree) however not all parks were inventoried and OLMSTEAD PARKS is only a small number of all City of Buffalo Parks (there are only 6 Olmstead Parks and over 40 regular city parks).

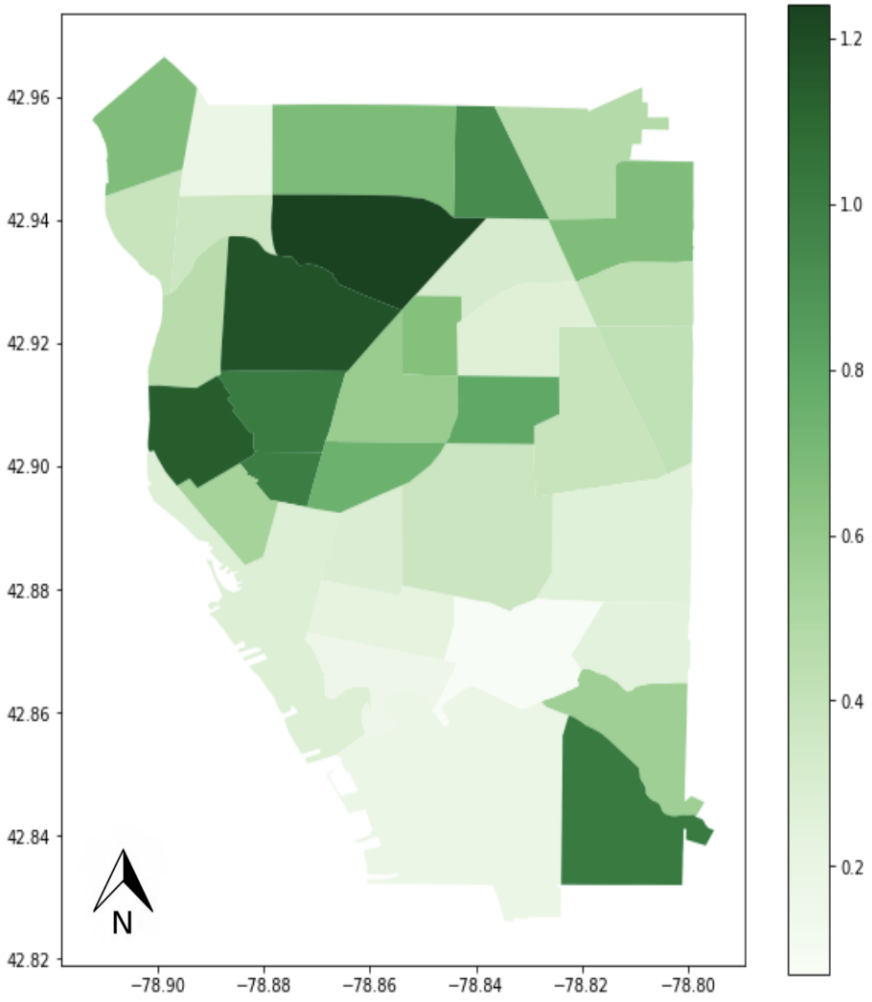

Figure 3: Tree Canopy coverage in Buffalo neighborhoods by percentage

Crime Data

Data on crime occurrences in Buffalo neighborhoods were obtained from 'Crime_Incidents' dataset from the Buffalo city open data portal. Data before 2009 has been termed unreliable on the website so we excluded it from our analysis.

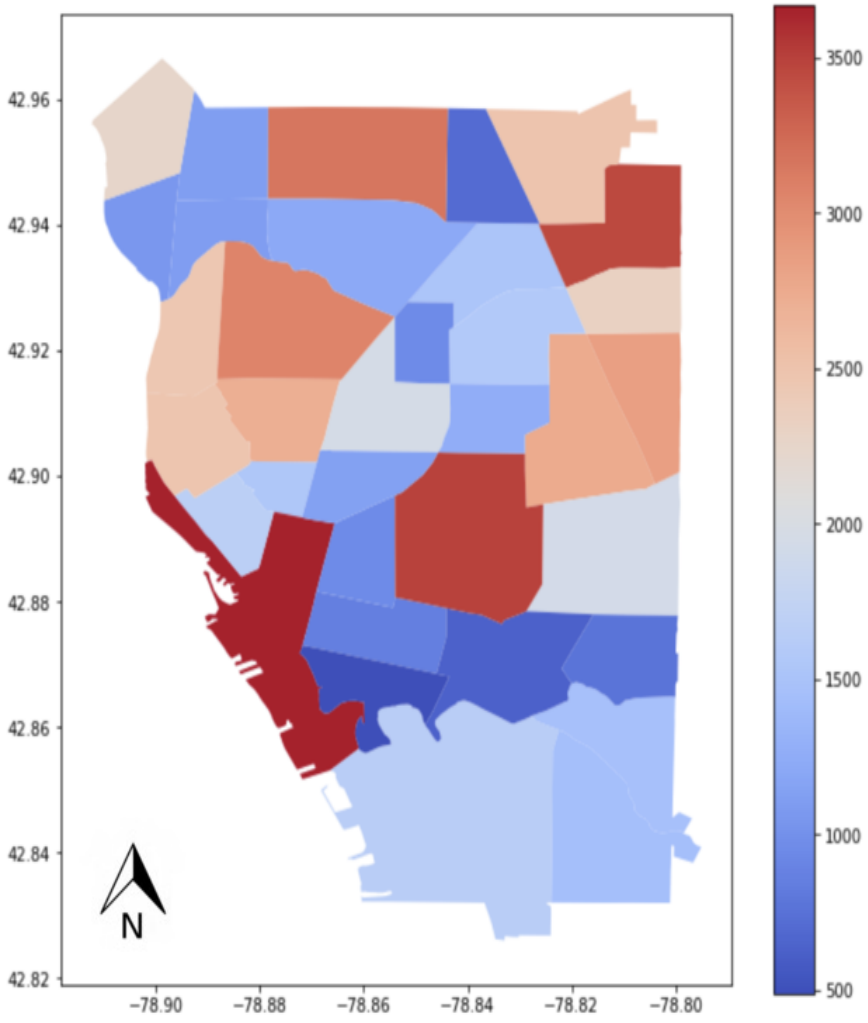

Figure 4. Number of Crimes in Buffalo Neighborhoods

The crime rate for each neighborhood was calculated per 100 people in the neighborhood using the formula:

$$\frac{\text{(Crimes per neighborhood)}}{\text{Total population per neighborhood}} \text{ x } 100$$

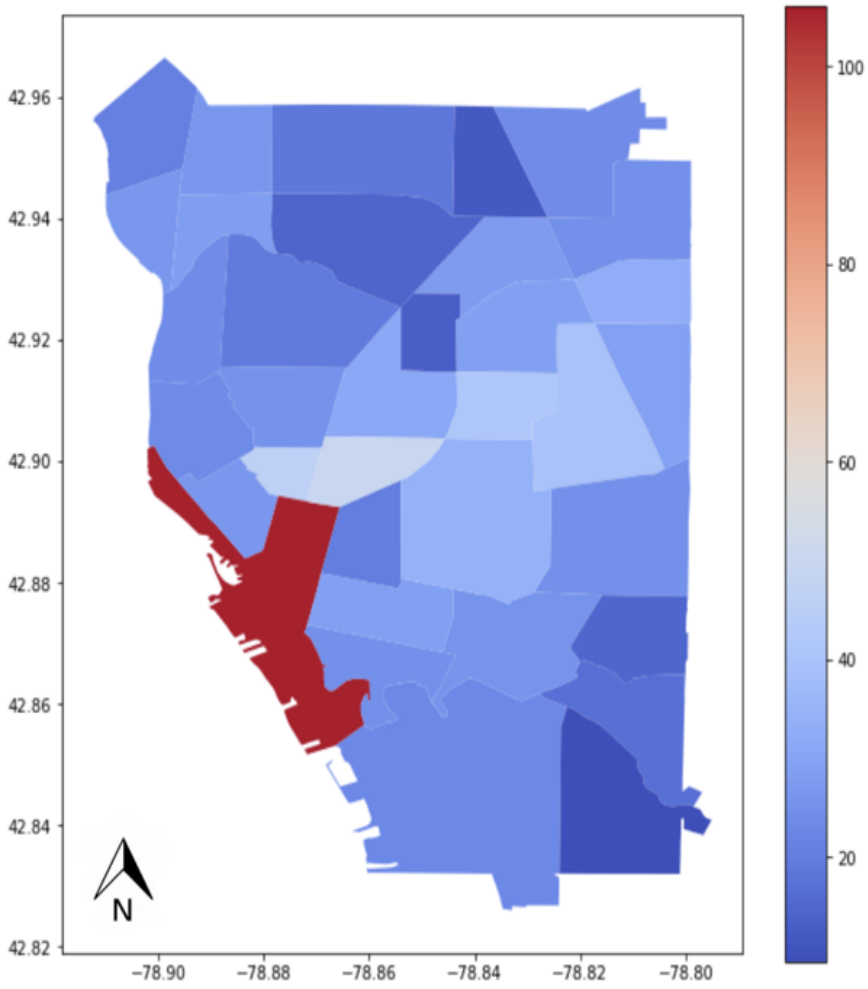

Figure 5. Total indicator crimes per 100 residents in Buffalo neighborhoods

Analysis

Extensive Exploratory Data Analysis (EDA) was carried out to see various aspects related to crime and tree datasets. Some of the analysis couldn't be performed due to wrong data. For example the 'incident time' in crime dataset has just values in 12-hour format but AM or PM isn't specified so while performing analysis all of the crimes occurred in just first 12 hours of a day, which can't be the case.

Some of the results from the EDA are:

Here we can see in Figure 6. Larceny/Theft appear to be the largest type of crimes occurred followed by Burglary, Assault, Robbery, etc.

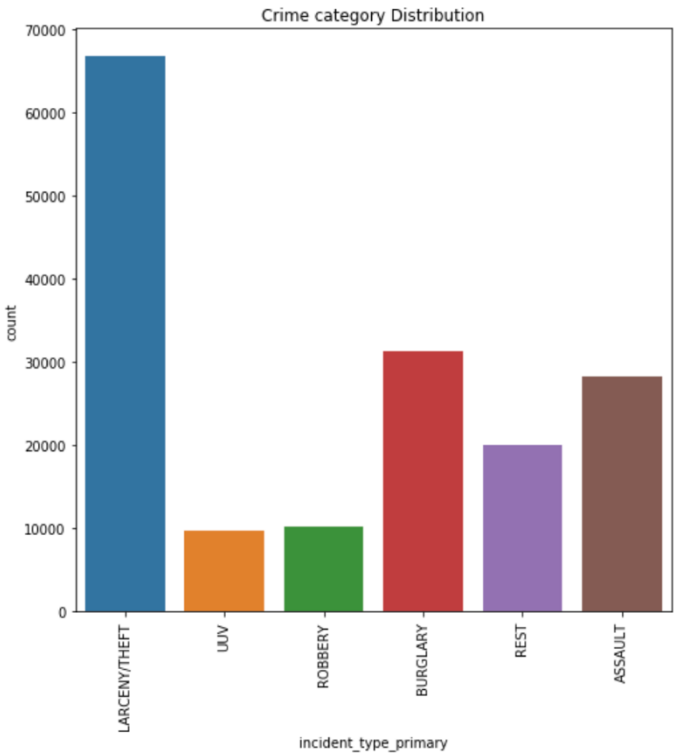

Figure 6. Crime Category Distribution of Buffalo

Time Series graph was plotted to gain some interesting insights:

We can observe here that every year there's a drop-in number of crimes committed at the end of the year and beginning of the new year which is basically the holiday season and also the coldest months here in Buffalo.

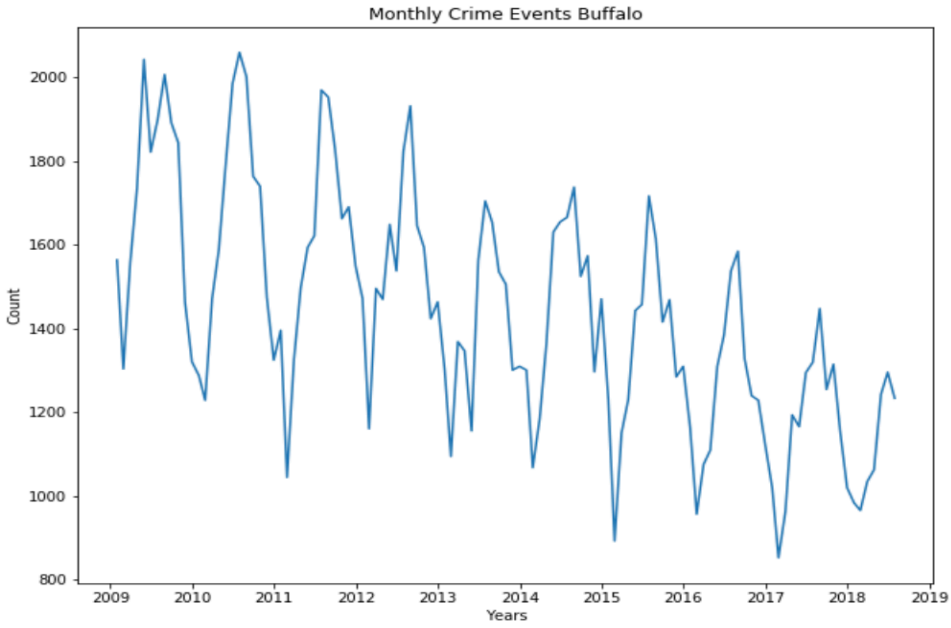

Figure 7. Time Series Graph for Crime incidents from 2009

Statistical Analysis

Following the work of Troy et al.'s (2012) study of Baltimore County, the Ordinary Least Squares Regression (OLS) and Lasso Regression were used to discern the relationship between Buffalo Crime Incidents and street tree canopy coverage.

First, we selected some variables from our human point of view which might lead to the high crime rate in a neighborhood and plotted against each other in a scatter plot. The selected variables were:

- 'crime_rate': Crime rate for each neighborhood.
- 'poverty_rate': Poverty rate for each neighborhood.
- 'hsedu': We took high school education as our education variable.
- 'unemployment_rate': Unemployment Rate for each neighborhood.
- 'nbhd_leafsurf_perc': Leaf Surface Area percentage for each neighborhood compared to the total surface area of that neighborhood.
- 'percent_vacant_units': Percentage of vacant houses in that neighborhood.

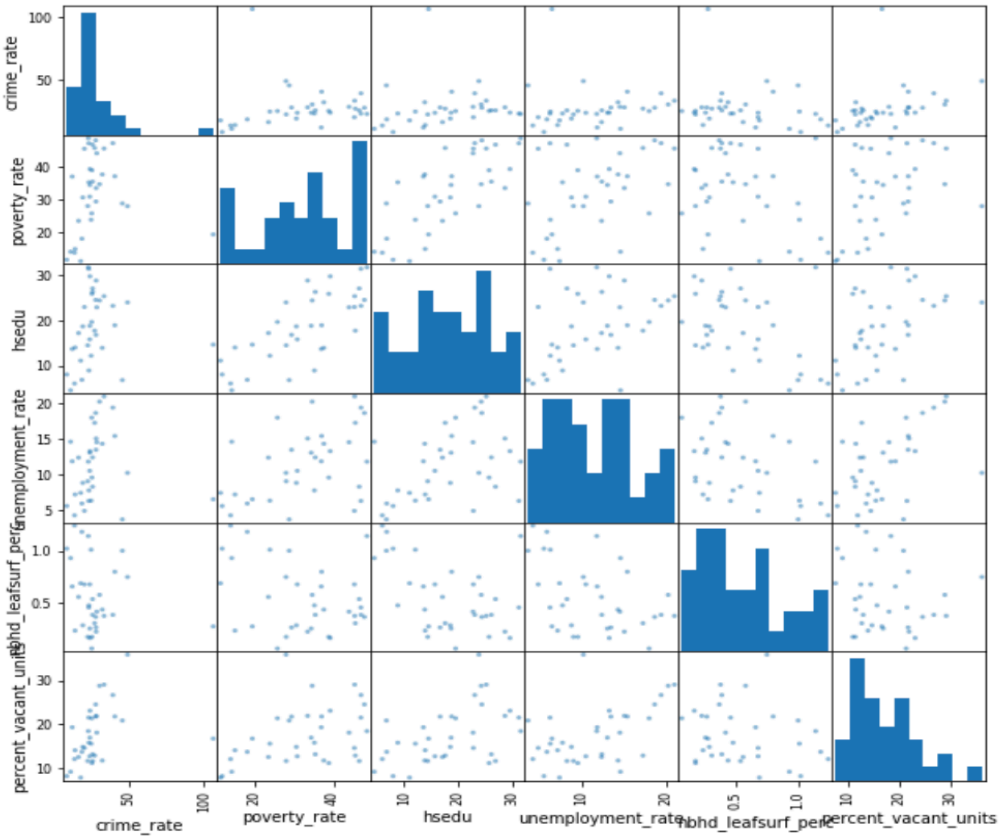

Figure 8. Scatter plot for humanly selected variables

As seen in Figure 8, Poverty and Education, Education and Unemployment have a positive linear correlation. We just took a minimun high school education. Based on the results above, and the knowledge of real-world situations in that higher education tends to result in higher income, and therefore lower poverty levels, we can infer that the census measured Education in such a way that higher education was quantified as a lower value.

Seeing this correlation, we applied OLS regression to predict education from poverty since they appeared correlated in the scatter matrix.

OLS Regression Results

Dep. Variable:	hsedu	R-squared:	0.559
Model:	OLS	Adj. R-squared:	0.546
Method:	Least Squares	F-statistic:	41.89
Date:	Fri, 17 Aug 2018	Prob (F-statistic):	2.42e-07
Time:	02:32:18	Log-Likelihood:	-105.96
No. Observations:	35	AIC:	215.9
Df Residuals:	33	BIC:	219.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.1369	2.660	0.803	0.428	-3.276	7.550
poverty_rate	0.5024	0.078	6.472	0.000	0.344	0.660

Omnibus:	0.494	Durbin-Watson:	1.466
Prob(Omnibus):	0.781	Jarque-Bera (JB):	0.631
Skew:	-0.199	Prob(JB):	0.729
Kurtosis:	2.475	Cond. No.	105.

The model performed somewhat good considering the data is not normalized. Looking at the p-values from the Linear Regression model above comparing Education and Poverty, it has a p-value much less than 0.05 and therefore is statistically significant. We can reject the null hypothesis, and we can see there is a correlation between Education and Poverty

Now heading towards our main goal, tried predicting crime rate from just neighborhood leaf surface area percentage.

OLS Regression Results

Dep. Variable:	crime_rate	R-squared:	0.037
Model:	OLS	Adj. R-squared:	0.008
Method:	Least Squares	F-statistic:	1.273
Date:	Fri, 17 Aug 2018	Prob (F-statistic):	0.267
Time:	02:32:18	Log-Likelihood:	-146.03
No. Observations:	35	AIC:	296.1
Df Residuals:	33	BIC:	299.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	33.4205	5.468	6.112	0.000	22.296	44.545
nbhd_leafsurf_perc	-9.7071	8.602	-1.128	0.267	-27.208	7.794

Omnibus:	51.405	Durbin-Watson:	1.216
Prob(Omnibus):	0.000	Jarque-Bera (JB):	290.506
Skew:	3.242	Prob(JB):	8.27e-64
Kurtosis:	15.537	Cond. No.	4.18

The model performed very bad. To understand the reason behind this we plot individual scatter plot of crime rate vs leaf surface area percentage for each neighborhood.

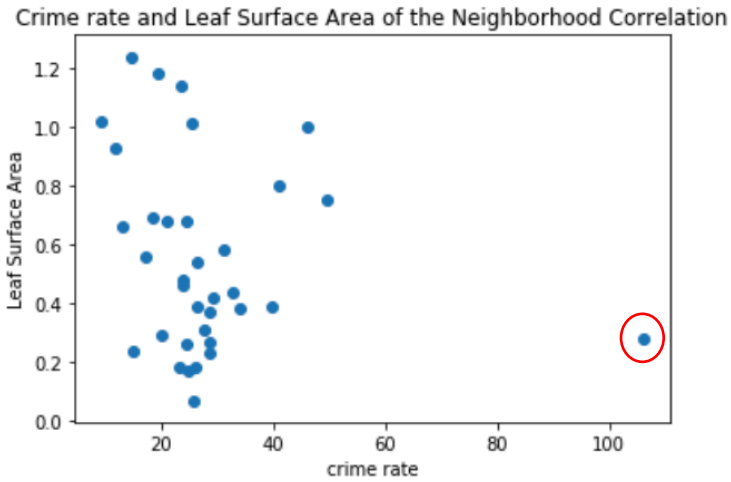

Figure 9. Scatter plot of Crime Rate and leaf surface area of the neighborhoods

From figure 9, we can infer that one outlier which is having a high crime rate is degrading the model's performance. So, we removed that outlier and checked the model again. This time we got an R^2 value of 0.539 which is a good model but not the best fit.

So, to get the features that are important and also to normalize all the subset features we used Lasso Regression technique. Also, because the data is not that big or else we could've used XGBoost classifier technique.

Lasso gave us an R^2 value of 0.73 after identifying the best features. These features were:

- Crimes: Crimes in that neighborhood.
- nbhd_leafsurf_perc: Leaf Surface Area percentage for each neighborhood compared to the total surface area of each neighborhood.
- total_persons: Population of each neighborhood.
- percent_vacant_units: Percentage of vacant houses in that neighborhood.

OLS Regression Results

Dep. Variable:	crime_rate	R-squared:	0.802
Model:	OLS	Adj. R-squared:	0.774
Method:	Least Squares	F-statistic:	29.29
Date:	Fri, 17 Aug 2018	Prob (F-statistic):	8.24e-10
Time:	02:32:19	Log-Likelihood:	-95.133
No. Observations:	34	AIC:	200.3
Df Residuals:	29	BIC:	207.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	18.9856	3.421	5.549	0.000	11.988	25.983
crimes	0.0105	0.002	6.499	0.000	0.007	0.014
nbhd_leafsurf_perc	6.5185	2.716	2.400	0.023	0.964	12.073
total_persons	-0.0028	0.000	-6.612	0.000	-0.004	-0.002
percent_vacant_units	0.3048	0.150	2.032	0.051	-0.002	0.612

Omnibus:	2.439	Durbin-Watson:	2.128
Prob(Omnibus):	0.295	Jarque-Bera (JB):	1.797
Skew:	0.563	Prob(JB):	0.407
Kurtosis:	2.993	Cond. No.	4.07e+04

Figure 10. Results for the model using features selected by lasso regression.

Computed the model again using the new normalized data and removed outlier whether there's any correlation between crimes and tree canopy but the model again didn't fit good and therefore stating that negative relationship can be observed between tree canopy and crimes.

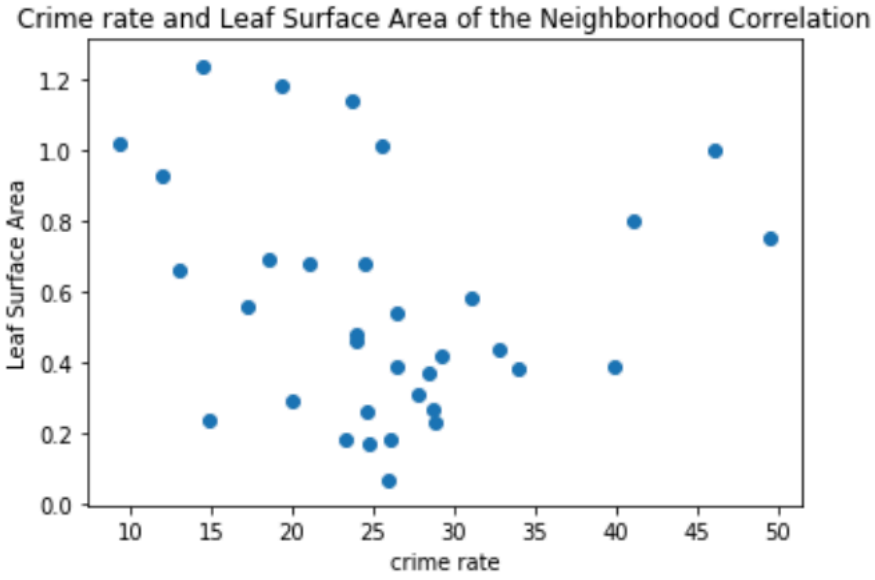

Figure 11. Crime rate and Leaf Surface Area Scatter plot using Normalized data

Modeling for crime as the dependent variable, OLS regression produced a reliable model from the features got from Lasso regression and after normalizing the data.
OLS regression found a statically significant negative relationship ($p < 0.02$) between crime and tree canopy. The adjusted R-squared value, a measure of the model performance, was 0.77, meaning the model explains 77% of the occurrence of the dependent variable of crime.

Challenges overcome:

- One of the biggest challenges in this project was the lack of good quality of data. For example:
  - **i.** The crime data available on the open data portal is only reliable after the year 2009.
  - **ii.** Missing census tracts such as unemployment rate, crime rate needed to be calculated on our own for each individual neighborhood using individual formulae:
    - **Labor Force:** Employed + Unemployed.
    - **Labor Force Participation (LFP) Rate:** Labor Force / Population.
    - **Employment Rate (EPOP):** Employed / Population.
    - **U3 Rate (Official Unemployment):** Unemployed / Labor Force.
  - **iii.** Crimes per neighborhood also were needed to be calculated using external tools.

- Geospatial data is a whole new domain when it comes to data analysis. Handling such data requires the development of precise models, or the results could vary which might lead to incorrect findings and analysis.
- Many census tracts included geospatial data for neighborhoods whereas crime and tree inventory data were for block groups.
- Only Olmstead park trees were included and not all city park trees, since there is no data available.
- GIS tools such as ArcGIS and QGIS was used to acquire accurate geospatial data for neighborhoods as it isn't available on the open data portal.
- These GIS tools consumed a good amount of time to get hold of.
- Importance of getting clean data to work on was realized.
- Data cleaning and pre-processing is an important step in the data science pipeline, and now I know why.

Impact of the project on the organization

The findings from this project are going to be very useful for the Buffalo Sewer Authority in future landscape and urban planning projects. As a negative relationship can be observed, trees would be planted in the vacant spaces nearby streets to reduce the crime occurrences in the city and make Buffalo City safer in the future. Of course, Trees are not the only thing affecting the crime but they might be one of the very useful ways to fight crime and also reduce the pollution on the streets and improve air quality index.

Future Research

A natural direction for further research on this topic regarding the city of Buffalo would be to acquire LiDAR data and Park trees data and to run the same models again on that data, to compare how they fare. Geographically Weighted Regression (GWR) can be applied once we have that data which will be available soon on the open data portal.

References

1. Akbari, H., Pomerantz, M., and Taha, H. 2001. Cool Surfaces and Shade Trees to Reduce Energy Use and Improve Air Quality in Urban Areas. Solar Energy, Vol. 70 No. 3, 2001. pp 295-310. Retrieved February 22, 2013 from Science Direct.
2. Donovan, G., and Prestemon, J. 2012. The Effect of Trees on Crime in Portland, Oregon. Environment and Behavior, Vol. 44 No. 1, 2012. pp 3-30. Retrieved August 7, 2012 from Sage Publications.
3. Michelle C. Kondo, SeungHoon Han, Geoffrey H. Donovan, and John M. MacDonald. 2015. The Effect of Trees on Urban Crime: Evidence from the Spread of the Emerald Ash Borer in Cincinnati, from the Department of Criminology, University of Pennsylvania.
4. Mary K. Wolfe, Jeremy Mennis. 2012. Does Vegetation encourage or suppress urban crime? Evidence from Philadelphia, PA. Landscape and Urban Planning Journal.
5. Andrew W. Eckerson. 2013. Understanding the relationship between tree canopy and crime in Minneapolis, Minnesota using Geographically Weighted Regression. Department of Resource Analysis, Saint Mary's University of Minnesota, Winoma.
6. Austin Troy, J. Morgan Grove, Jarlath O'Neil-Dunne. 2012. The relationship between tree canopy and crime rates across an urban-rural gradient in the greater Baltimore region. Landscape and Urban Planning Journal.